fferentiates the species physicalism from the genus naturalism is the condition that any minimal physical duplicate of our world is a duplicate simpliciter of our world.

There is, though, a second problem case that is cited in the literature to show that the condition that any minimal physical duplicate of our world is a duplicate simpliciter of our world fails to state a sufficient condition for physicalism. And the second problem case is compatible with naturalism. The problem is the possibility of 'blockers' (Hawthorne, 2002). As noted above, the minimality thesis allows that a physical duplicate of our world could include 'extras', including extra properties and laws. Suppose, then, that in our world and indeed in any possible world that is a minimal physical duplicate of our world, whenever something has a certain physical property P, it has a certain mental property M. Suppose further that the relationship between having P and M is weaker than metaphysical necessitation, so that there are possible worlds in which when something has P and also has a certain property B, a property that nothing in our world has (and indeed that does not even exist in our world, if properties are contingent existents), then that something lacks M. In that world, having B blocks things that have P from having M. The possibility of such blockers is compatible with the thesis that any minimal physical duplicate of our world is a duplicate simpliciter of our world and, indeed, as well, with naturalism being true of our world. But it is claimed that it should be incompatible with physicalism.

There is no consensus in the literature about how, exactly, the possibility of blocker cases should be handled by physicalists or even whether physicalism should rule out all blocker cases. But one option is to maintain that physicalism is the thesis that everything in our world is either physical or metaphysically necessitated by the physical. (That doctrine faces the problem of a necessarily existing God, since such a God would trivially be necessitated by anything. But, again, that problem might be handled by also requiring for physicalism that naturalism be true.) That thesis is stronger than the thesis that any minimal physical duplicate of our world is a duplicate simpliciter of our world in that it implies that thesis but is not implied by it. Blocker cases do not arise for this stronger thesis since if it is true, then M (which by hypothesis is in our world) will be metaphysically necessitated by the physical, and so any physical duplicate of our world would include it, regardless of what else it included.

The thesis that everything in our world is either physical or necessitated by the physical may be too strong, however, to be required for physicalism. On the assumption that laws of nature are metaphysically contingent, for a physical property to metaphysically necessitate a functional property F, it will have to be a conjunctive physical property that includes as a conjunct not only a physical property that occupies the relevant functional role, but that also includes as conjuncts the physical laws in virtue of which the physical property plays the role. And even that may be insufficient to guarantee that the functional properties that physicalism should capture are metaphysically necessitated by physical properties. The conjunctive property may also have to include as a conjunct the property of being in a world such that any minimal physical duplicate of the world is a duplicate simpliciter of it, which is not a property that would be posited by the completed physics (McLaughlin, 2007).

Yet another strong formulation of physicalism concerns relationships between truths (or facts). Thus, physicalism is sometimes taken to be the thesis that all truths are physical truths or else hold in virtue of physical truths. This may seem too strong since, for instance, mathematical truths do not hold in virtue of physical truths. Indeed, given the extensive use of classical mathematics in physics, arguably, certain physical truths hold, in part at least, in virtue of certain truths of pure mathematics, namely the axioms of Zermelo–Fraenkel set

theory with the axiom of choice, which provide, in conjunction with classical logic, the foundations of the classical mathematics. The in virtue of or grounding claim could be weakened to the thesis that all contingent truths are physical truths or hold in virtue of physical truths. But then it would be compatible with the claim that there is a necessary God, something that would be necessarily true if it were true. Instead of so restricting it, however, it might be claimed that the truths that figure in the foundations of the mathematics deployed in physics should, because they so figure, count as physical truths.

Elimination or Location?

Although there is no consensus as to how, exactly, physicalism should be formulated, it is, as noted above, fairly widely held that physicalism requires that any minimal physical duplicate of our world is a duplicate simpliciter of our world. Even though that does not suffice for physicalism, it is a very strong requirement indeed. Let us call the thesis that any minimal physical duplicate of our world is a duplicate simpliciter of our world 'the physical minimality thesis'. Physicalists aim to defend the physical minimality thesis on empirical grounds and grounds of overall coherence and simplicity of total theory.

Consider the kinds of objects and properties posited by the various special sciences. Those kinds are typically ones that no recognizable descendant of current physics will explicitly posit as such. Given their commitment to the physical minimality thesis, physicalists thus incur the dialectical obligation of either defending an eliminativist position as concerns them (denying their existence Stich, 1999) or else justifying the claim that any minimal physical duplicate of our world will include those kinds and the exact same worldwide pattern of instantiation of them as in our world. Consider also that we ordinarily take there to be persons, beliefs, desires, intentions, emotions, aches, pains, itches, cities, countries, shopping malls, yard sales, wars, and morally right or morally wrong actions, just to name a few of the sorts of things that we ordinarily take there to be. As concerns each of these, the physicalists incurs the dialectical obligation of either defending an eliminativist position or else justifying the claim that any minimal physical duplicate of our world will include them and the exact same worldwide pattern of instantiation of them as in our world. Physicalists are typically (though not invariably) realists about such matters; they typically eschew eliminativism as concerns them. They aim to discharge the dialectical obligation of justifying the claim that any minimal physical duplicate of our world will include these things and have the same worldwide distribution of them as in our world. That is a truly formidable task.

A leading view is that discharging that dialectical obligation requires solving location problems (Jackson, 1998), problems of saying how the things in question, or at least things that metaphysically necessitate them, could, in principle at least, be located in the description that an ideally completed physics would provide of our world, on the assumption that that physics is a recognizable descendant of current physics. Thus, the location problem for an apparently nonphysical property is to see how it or a property that metaphysically necessitates it could be located in such a description. There are two basic strategies that typically get invoked here; which is invoked depends on the kind of apparently nonphysical property in question. (Analogous claims will be true for apparently nonphysical kinds of objects or stuff.) One strategy is to argue that a certain apparently nonphysical property is indeed nonphysical, but is a functional property that must exist and be instantiated, exactly as it is in our world, in any world that is a minimal physical duplicate of our world. The other strategy is to argue that the apparently nonphysical property is in fact a certain physical property, and so not in fact a nonphysical property. That strategy is typically implemented in either of two ways. Sometimes it is argued that there is a functional description that expresses a contingent, a priori reference fixing description for the relevant nonphysical term or concept that we use to denote the property, and that there is a recognizably physical property that meets that description, that is, that fills the functional role described (Jackson, 1998; Kim, 2005). This is a kind of filler (or realization) functionalist account of the property. Also, some physicalists argue that an apparent nonphysical property, for instance, a mental property M, is identical with a certain physical property P, but that the relevant identity claim is a posteriori and M as such bears no a priori link to physical or functional descriptions, and so admits of no functional analysis. Whether the location problems can be solved in all cases of in which eliminativism should be rejected is a highly controversial matter. (It should be mentioned here that physicalist theories of qualia (the what-it-is-like for the subject aspects of experience) often attempt to locate mental properties among physical properties very broadly construed (e.g., among neurobiological properties), rather than among properties of completed microphysics (Smart, 1959; Hill, 1991; Loar, 1997; McLaughlin, 2011). Such a view does not entail the general doctrine of physicalism. But whether qualia can be so located is also a very controversial matter.)

Nonphysicalists that embrace a kind of property dualism maintain that the location problem cannot be solved for certain properties (even if the physical is very broadly construed). For example, nonphysicalists typically argue that the location problem cannot be solved for qualia (Jackson, 1982; Chalmers, 1996; Kim, 2005). Indeed, some maintain that a minimal physical duplicate of our world would be, unlike our world, devoid of consciousness, and so be a world in which our physical duplicates lack any consciousness at all, and so are 'zombies' (Chalmers, 1996; Kirk, 2012).

It has also been argued that the location problem cannot be solved for intentional properties (e.g., being a belief that P) and/or for normative properties (those that fall on the ought side of the is/ought divide). Such property dualists maintain that a minimal physical duplicate of our world would differ from our world either in not including such properties or instantiations of them, and so that physicalism is false.

Suffice it to repeat that physicalism is a controversial doctrine. Its truth remains a matter of dispute.

See also: Causation: Physical, Mental, and Social; Physical Sciences: History and Sociology.

Bibliography

Baker, L.R., 2000. Persons and Their Bodies: A Constitution View. Cambridge University Press, Cambridge, UK.

Bedau, M., Humphreys, P (Eds.), 2006. Emergence: Contemporary Readings in Science and Philosophy. MIT Press, Cambridge, MA.
Broad, C.D., 1925. The Mind and Its Place in Nature. Harcourt Brace, New York.
Chalmers, D., 1996. The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press, New York.
Crane, T., Mellor, D.H., 1990. There is no question of physicalism. Mind 99, 185–200.
Davidson, D., 1970. Mental events. Reprinted in Davidson (Ed.), Essays on Actions and Events. Oxford University Press, Oxford, UK.
Haugeland, J., 1984. Ontological supervenience. The Southern Journal of Philosophy XXII (Suppl.), 1–12.
Hawthorne, J., 2002. Blocking definitions of materialism. Philosophical Studies 110 (2), 103–113.
Hellman, G.P., Thompson, F.W., 1975. Physicalism: ontology, determination, and reduction. Journal of Philosophy 72, 551–564.
Hempel, C., 1969. Reduction: ontological and linguistics facets. In: Morgenbesser, P. Suppes, White, M. (Eds.), Philosophy, Science and Method. St. Martin, New York, pp. 179–199.
Hill, C.S., 1991. Sensations: A Defense of Type Materialism. Cambridge University Press, Cambridge, UK.
Jackson, F., 1982. Epiphenomenal qualia. Philosophical Quarterly 32, 127–136. Reprinted in Lycan, W. (Ed.), 1995. Mind and Cognition. Blackwell, Oxford, UK, pp. 69–77.
Jackson, F., 1998. From Metaphysics to Ethics: A Defense of Conceptual Analysis. Clarendon Press, Oxford, UK.
Jackson, F., 2007. A priori physicalism. In: McLaughlin, B.P., Cohen, J. (Eds.), Contemporary Debates in Philosophy of Mind. Basil Blackwell, Oxford, UK, pp. 185–199.
Kim, J., 2005. Physicalism or Something Near Enough. Princeton University Press, Princeton, NJ.
Kirk, R., 2012. Zombies. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer, 2012 Edition). http://plato.stanford.edu/archives/sum2012/entries/zombies/.
Lemhoff, R., 2013. Intuitionism in the Philosophy of Mathematics. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall, 2013 Edition). http://plato.stanford.edu/archives/fall2013/entries/intuitionism/.
Lewis, D.K., 1983. New work for a theory of universals. Australasian Journal of Philosophy 61 (4), 343–373.
Lewis, D.K., 1994. Reduction of mind. In: Guttenplan, S. (Ed.), A Companion to the Philosophy of Mind. Blackwell Reference, Oxford, UK, pp. 412–431.
Linnebo, Øystein, 2013. Platonism in the Philosophy of Mathematics. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter, 2013 Edition). http://plato,stanford.edu/archives/win2013/entries.platonism-mathematics/.
Loar, B., 1997. Phenomenal properties. In: Block, N., Flanagan, O., Guzeldere, G. (Eds.), The Nature of Consciousness: Philosophical Debates. MIT Press, Cambridge, MA, pp. 597–616.
Loewer, B., 2001. From physics to physicalism. In: Gillett, C., Loewer, B. (Eds.), Physicalism and Its Discontents. Cambridge University Press, Cambridge, UK, pp. 57–73.
McLaughlin, B.P., 1992. The rise and fall of British emergentism. In: Berckermann, A., Kim, J., Flohr, H. (Eds.), Emergence or Reduction?, pp. 49–93.
McLaughlin, B.P., 1999. The philosophy of mind. In: Audi, R. (Ed.), The Cambridge Dictionary of Philosophy. Cambridge University Press, Cambridge, UK, pp. 684–694.
McLaughlin, B.P., 2007. On the limits of a priori physicalism. In: McLaughlin, B.P., Cohen, J. (Eds.), Contemporary Debates in Philosophy of Mind. Basil Blackwell, Oxford, UK, pp. 200–224.
McLaughlin, B.P., 2010. Consciousness, type physicalism, and inference to the best explanation. Philosophical Issues 20, 266–304.
McLaughlin, B.P., 2011. On justifying neurobiologicalism for consciousness. In: Gozzano, S., Hill, C. (Eds.), New Perspectives on Type Identity. Cambridge University Press, Cambridge, UK, pp. 206–228.
McLaughlin, B., Bennett, K., 2011. Supervenience. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter, 2011 Edition). http://plato.stanford.edu/archives/win2011/entries/supervenience/.
Melnyk, A., 2003. A Physicalist Manifesto: Thoroughly Modern Materialism. Cambridge University Press, Cambridge, UK.
Montero, B., 2009. What is the physical? In: McLaughlin, B.P., Beckermann, A., Walter, S. (Eds.), The Oxford Handbook of Philosophy of Mind. Oxford University Press, Oxford, UK, pp. 95–108.
Nagel, T., 1979. Panpsychism. In: Nagel, T. (Ed.), Mortal Questions. Cambridge University Press, Cambridge, UK, Chapter 13.
Papineau, D., 1993. Philosophical Naturalism. Blackwell, Oxford, UK.
Pettit, P., 1993. A definition of physicalism. Analysis 53, 213–223.
Poland, J.S., 1994. Physicalism: The Philosophical Foundations. Oxford University Press, Oxford, UK.
Robinson, H., 2009. Idealism. In: McLaughlin, B.P., Beckermann, A., Walter, S. (Eds.), The Oxford Handbook of Philosophy of Mind. Oxford University Press, Oxford, UK, pp. 189–205.
Rodriguez-Pereyro, G., 2011. Nominalism in Metaphysics. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall, 2011 Edition). http://plato.stanford.edu/archieves/Fall 2011/entries.norminalism-metaphysics/.
Seager, W., 2009. Panpsychism. In: McLaughlin, B.P., Beckermann, A., Walter, S. (Eds.), The Oxford Handbook of Philosophy of Mind. Oxford University Press, Oxford, UK, pp. 206–222.
Shoemaker, S., 2007. Physical Realization. Oxford University Press, Oxford, UK.
Smart, J.J.C., 1959. Sensations and brain processes. Philosophical Review 68, 141–146.
Stich, S., 1999. Eliminative materialism. In: Wilson, R.A., Keil, F.C. (Eds.), The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge, MA, pp. 265–267.
Stoljar, D., 2009. Physicalism. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall, 2009 Edition). http://plato.stanford.edu/archives/fall2009/entries/physicalism/.
Strawson, G., 2008. Realistic monism: why physicalism entails panpsychism. In: Strawson, G. (Ed.), Real Materialism and Other Essays. Clarendon Press, Oxford, UK, pp. 19–52.
Stubenberg, L., 2013. Neutral Monism. In: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter, 2013 Edition). http://plato.stanford.edu/archives/win2011/entries/neutral-monism/.
Van Gulick, R., 2009. Functionalism. In: McLaughlin, B.P., Beckermann, A., Walter, S. (Eds.), The Oxford Handbook of Philosophy of Mind. Oxford University Press, Oxford, UK, pp. 128–151.
Witmer, G., 2001. Sufficiency claims and physicalism: a formulation. In: Gillett, C., Loewer, B. (Eds.), Physicalism and Its Discontents. Cambridge University Press, Cambridge, UK, pp. 57–73.
Yablo, S., 1992. Mental causation. Philosophical Review 101 (2), 245–280.

Piaget, Jean (1896–1980)

Jacques Montangero, University of Geneva, Geneva, Switzerland

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11427–11433, © 2001, Elsevier Ltd.

Abstract

Passionately interested in biology and philosophy in his adolescence, Jean Piaget completed studies in natural sciences in his hometown (Neuchâtel, Switzerland). He then studied psychology and history of sciences in Zurich and Paris. From 1921, when he was appointed researcher at the J.-J. Rousseau Institute in Geneva, until his death in 1980, he conducted research in this institution, with the help of a team of coworkers. His early psychological research on the specificity of children's thought brought Piaget almost immediate international fame. In the 1930s he became professor at the University of Geneva and codirector of the J.-J. Rousseau Institute, and published books on infant cognition with an original model inspired by biology. The numerous books published in the following decades used a new formal model to explain children's (and later adolescents') reasoning abilities. Prodigiously productive, Piaget continued to research and write until the end of his life, studying various fields beside psychology and epistemology and developing constantly new concepts to explain cognitive development. His comprehensive 'constructivist' theory, based on innumerable experimental data, had a considerable influence on child psychology and on the development of cognitive developmental psychology. By its multidisciplinary approach it is a precursor of today's cognitive science.

A most productive scholar in the area of children's intellectual development, Jean Piaget was primarily interested in epistemology or theory of knowledge. His genetic psychology (psychology of the genesis of thought) was intended to provide a scientific basis for epistemology, with the main concern to explain how knowledge can become more rigorous and more objective. Piaget's psychological explanations were inspired by several scientific disciplines and beside his studies in psychology and epistemology he published in other fields such as biology, logic, and sociology. His work derived from evolutionist approaches rooted in the nineteenth century yet it is still in line with scientific trends at the end of the twentieth century.

The core of Piaget's theory is the conception of mind as an active system that, like an organism, assimilates external data into its structures and can adjust to these data. Special emphasis, in the knowledge process, is placed on the subject's activity and the structures through which he or she perceives and understands things. Thus, the development of knowledge is viewed as the construction of new structures (rather than the accumulation of new information) triggered and guided by internal processes during the subject's encounter with reality. Each new stage in this process opens up new possibilities. In this interactionist conception, structures are neither predetermined nor found in the environment and there are striking similarities between biological evolution, cognitive development in children, and the development of scientific thought.

Seven Decades of Scientific Productivity

Childhood and Adolescence (1896–1918)

Jean Piaget was born in Neuchâtel, a small city of Switzerland, on 9 August 1896. He was the son of an historian and professor of medieval literature at the university. Of his mother, he wrote in his autobiography that she was an intelligent and energetic woman, whose psychological imbalance created problems in the family. At the age of 11 years, Jean Piaget started classifying and collecting mollusks with the help of the director of the natural history museum of his hometown. In his teens, he was already an acknowledged specialist in the field and started publishing papers in scientific journals (as of 1911). This activity led him to decide to study natural sciences at the university between 1915 and 1918. At the same time, as of the age of 16 he became passionately interested in philosophy, and particularly in the theory of knowledge or epistemology. He soon decided to devote his life to the elaboration of a biological theory of knowledge, whose goal would be to explain the evolution of ideas both in individuals and in the scientific community. This project was influenced by evolutionary approaches in the field of biology (Lamarck, Darwin) and in philosophical works on issues of intelligence (Spencer, Le Dantec, Bergson). In the following years, Piaget's professor of philosophy A. Reymond convinced him of the importance of a rationalist approach making use of logic and mathematics.

In 1918 Piaget published a philosophical and autobiographical novel (*Recherche*) describing the crisis he had gone through because of the conflict he felt between religion and science, or, in other words, between value and rationality. Eventually, the search for truth, which is the goal of science, became the most important value for Piaget. *Recherche* also contained the outline of a theory of the evolution of living and cognitive organizations based on the idea of improving forms of equilibrium between the whole and its parts. Piaget's twofold interest on the one hand in biology and the associated problems of the organization and evolution of living beings, and on the other hand in epistemology and related issues concerning the validity of knowledge and its development, forms the basis of all his work.

Complementary Studies (1918–21)

Piaget was convinced that a theory of knowledge should be validated by psychological research, and decided to get training in scientific psychology as soon as he graduated from

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61107-8

university with his PhD thesis on mollusks in 1918. He then spent a term in Zurich, attending courses in psychopathology – where he encountered psychoanalysis – and in experimental psychology. He then studied in Paris from 1919 to 1921, where he attended lectures in psychology (specifically by Janet), philosophy (with Lalande), and history of science (by Brunschvicg who had great influence on Piaget's ideas). Courses in psychopathology provided him with a clinical interview method, which he soon used in his discussions with children. In effect, in parallel with his studies, he worked in the laboratory of Binet and Simon, standardizing intelligence tests for children. Piaget was fascinated by the reasoning of the children who did not yet answer 'correctly,' and he discovered that the study of children's thinking could be a means to validate a theory of knowledge. This program, which was to be the main focus of his work for 60 years, was very similar to the program designed by James Mark Baldwin (1861–1934), from whom Piaget borrowed several concepts and the expression 'genetic psychology.'

The Early Psychological Research (1921–29)

In 1921, Claparede offered a position in his Institute for educational sciences (known as the Jean-Jacques Rousseau Institute) in Geneva to Piaget, who was to conduct most of his professional life in this institution. What he found there were stimulating intellectual atmosphere and students willing to assist him in his work. Together with them the young researcher started a series of studies of children's reasoning and soon published his first books: *The Language and Thought of the Child* (1923); *Judgment and Reasoning* (1924); *The Child's Conception of the World* (1926); *The Child's Conception of Physical Causality* (1927) (the titles given here and in the remainder of this text are the English translations of the French originals; the years are those of the original publications). The striking facts he discovered about the use of language, the nondifferentiation of concepts and the explanations of physical phenomena in children, as well as the detailed analyses and the synthetic categorization of these behaviors, brought Piaget almost immediate international fame. These studies showed that young children's thinking is structurally and functionally different from that of older children and adults. Young children lack a true logic of classes and relations, they have difficulties viewing reality from an objective viewpoint, and they lack moral norms, which command compliance on the basis of a feeling of necessity. Children's thinking tends to be egocentric, that is, they have difficulties in looking at things in perspective. In these early works, the passage from egocentric thought to a logical and more objective form of thinking was explained by a social factor: cooperation, or exchange between individuals who consider themselves as equals.

During this period, Piaget was involved in many activities beside his research. He taught child psychology at the J.-J. Rousseau Institute, philosophy and history of science at the University of Neuchâtel (1925–29), was for a few months psychoanalyzed by Sabina Spielrein, a pupil of Freud, and continued to study the evolution of mollusks. In 1923 he married a student at the Geneva Institute. Piaget and his wife Valentine spent much time minutely observing their three children who were born between 1925 and 1931. These observations were the basis of several books published in the following years.

Administrative Responsibilities, Professorship and New Directions of Research (the 1930s)

In 1929, Piaget was appointed professor of history of science at the University of Geneva, where he subsequently became professor of sociology (1939) and of experimental psychology (after the death of Claparede, 1940). He also played an important administrative role at the J.-J. Rousseau Institute as associate director (1929), then as codirector (since 1932), reorganizing the Institute and establishing links between the Institute and the University. At the same time, and until the 1960s, he was director of the *Bureau International de l'education* and managed to convince the governments of many countries to join that organization.

Piaget's most important activities continued to be research and writing. During these years, he accumulated data on children's reasoning. However, apart from his book on moral judgment (1932) – simultaneously in French and in English – which was the sequel to the psychology books published in the 1920s, his major publications dealt with infant cognitive development. The choice of this topic stemmed from Piaget's conviction that logic has its roots in actions and also from his deep interest in the issue of the origins of knowledge. *The Origins of Intelligence* (1936) and *The Child's Construction of Reality* (1937) contained extremely original observations and explanations of infant behaviors, and departed radically from Piaget's earlier writings as far as the explanatory model was concerned. Piaget discovered behaviors that had previously been overlooked, for example, the failure of babies younger than 9 months of age to look for an object hidden in front of them. Replications of these observations and experiments many years later around the world have confirmed these findings that had resulted from the observation of his three children.

In view of theory building, Piaget analyzed the elements of knowledge, the 'schemes' of actions, which become more differentiated and coordinated as development progresses. They evolve into increasingly complex and better-adapted forms of behavior, from automatic 'reflexes' to intelligent activities. This development is a special case of the relationship between the organism and the environment, in other words of biological adaptation. It results from the interaction between the subject of knowledge, who assimilates reality to his or her structures, and the object of knowledge, which may elicit an accommodation of these structures. Cognitive development is viewed as an evolving equilibrium between assimilation and accommodation. In *The Child's Construction of Reality* (1937), Piaget studied the growth of the practical knowledge of object permanence, space, time, and causality in infants. He introduced a new and abstract explanatory notion: the mathematical concept of 'group,' which several authors in the social sciences considered as an interesting model at the time. Piaget applied it to the organization of spatial knowledge in infants and was convinced that it could be used to explain the logic of school-age children. To this end, from 1937 onwards, he worked for 3 years on this model, expressing it in algebraic formulae.

Although Piaget was now a famous author who had received an honorary degree from Harvard University in 1936, his two books on infant cognition were not translated in English until the 1950s. They contained too many abstract considerations and microanalyses and were definitely ahead of their time. A third book, also based on the observation of Piaget's own children, was written during this period and published later: *Play, Dream and Imitation* (1945). It dealt with imitation in infants and the origins of symbolic thought in young children. Intellectual activity as well as play and imitation were defined in terms of equilibrium or disequilibrium between assimilation and accommodation.

The Mature Years: The Advent of Structural Analysis (1940s)

During this and the following period, Jean Piaget was prodigiously productive, thanks to a strict and very regular work regime. In the mornings he would teach, discuss research with his large team of coworkers, and perform his administrative duties. Piaget would then spend the rest of the day mostly by himself. In the early afternoon, while thinking about the topics of his current writing, he would stroll or ride a bicycle in the countryside near Geneva. During the rest of the afternoon he would write, which was the best way for him to elaborate his ideas. Even when traveling or attending conferences he would not let a day go by without writing a few pages. Although he loved classical music he almost never went out in the evenings to movies, theaters or concerts. Both in summer and in winter, his love of nature led him to repeated sojourns in the mountains of the Valais, where he both hiked and worked.

The logical model of 'grouping' of mental operations, which he had recently developed, and which accounted for the organization of inferences in overall structures, inspired much research and writing. It permitted Piaget not only to describe the limitations of young children's reasoning as he had done in his first books, but also to explain the sudden, stable, and rather generalized intellectual progress due to the advent of the elementary logic, of 'concrete operations.' Between 1941 and 1948, his work resulted in six books: *The Child's Construction of Quantities* (with B. Inhelder, 1941), *The Child's Conception of Number* (with A. Szeminska, 1941), *The Child's Conception of Time* (1946), *The Child's Conception of Movement and Speed* (1946), *The Child's Conception of Geometry* (with B. Inhelder and A. Szeminska, 1948), *The Child's Conception of Space* (with B. Inhelder, 1948). These books study the elementary mastery of basic scientific concepts. This mastery, which usually emerges between the ages of 6 and 8 or 9 years, was explained in terms of the underlying 'grouping' structures that organize mental operations. This type of structure entails the reversibility of thinking, for Piaget an essential feature of natural logic. In contrast, younger children think that the number of a set of objects or the quantity of a substance change when spatial transformations occur, and that the duration of a movement corresponds to the distance covered, without taking into account the relative speed of the moving objects. These young children have difficulties comparing the number of elements in a subclass with the number in the superordinate class. As to space, geometrical capacities like inverting a spatial order, constructing a straight line with vertical sticks using a sighting method, and drawing the horizontal line of the water level in a tilted glass also usually emerge after the age of 7 years. The basic concept studied in each book was analyzed and defined in the light of children's answers. For example, number was analyzed in terms of a synthesis between logical classification (class inclusion) and the seriation of asymmetrical relations (ordering by increasing or decreasing size); time was defined as the coordination of movements, in which the ability to consider speed played a fundamental role, and so forth.

Piaget also published a *Treaty of Logic* (1949) and with his coworkers he studied perception in children in the laboratory of experimental psychology. These studies, which attracted Piaget's attention to the phenomenon of regulation, were to provide the subject matter of a much later book. The same holds true for the papers in sociology which were published during this decade and collected in a book called *Sociological Studies* in 1965.

The second World War rendered contact with foreign colleagues difficult, but Piaget was invited in 1942 to give a series of lectures in Paris at the '*College de France*' (published in *The Psychology of Intelligence*, 1946). The abstract nature of his explanations struck many listeners as much as the brilliant analyses inspired by various scientific disciplines. After the war, UNESCO entrusted Piaget with several missions.

Achieving the Ultimate Goal of His Work: The Elaboration of a Genetic Epistemology (1950s)

In the 1950s, while continuing his work in genetic psychology with the help of his coworkers, Piaget devoted much of his effort to the foundation of a genetic epistemology, i.e., the study of the basic concepts and processes of the development of scientific knowledge. This used to be a topic of philosophical study but now became an object of scientific analysis. Actually, Piaget had indeed pursued epistemological issues throughout his entire previous work while ostentatiously focusing on the explanation of children's thinking. Now he wished to discuss questions of knowledge per se as they were raised in the sciences (but continued to use the study of children's cognitive development for this purpose). In 1950, Piaget published three volumes of an *Introduction to Genetic Epistemology*. The book begins with a general introduction defining genetic epistemology and its method and stating the main issues. Piaget develops an idea he had already expressed in earlier writings, that of a 'circle of sciences': physical reality is explained by mathematics, mathematics is the product of mind, which is a result of biological evolution, and organisms are closely connected with physical reality. The book deals in great detail with the epistemology of mathematics and physics, and in lesser detail with the epistemology of biology and the social sciences.

To make progress in the domain of genetic epistemology, Piaget founded the International Center for Genetic Epistemology, which was a working group comprising researchers in psychology and scientists from various disciplines (mathematics, physics, logic, etc.). The Center was launched in 1955 after obtaining, with difficulty, the necessary funding. A new series of books, which contained the results of the studies conducted at the Center for genetic epistemology, appeared from 1957 onwards. About 40 volumes were published in the series until the death of Piaget in 1980.

During the 1950s, when Piaget held a position as professor at the University of Paris (1952–63), his psychological research in Geneva continued to progress and resulted in three books with B. Inhelder: *The Development of Elementary Logical Structures* (1959), a sequel to the studies of concrete operations; *The Origin of the Idea of Chance* (1951) and *The Growth of Logical Thinking from Childhood to Adolescence* (1955). The latter book provided a new description and definition of the highest stage of logical development: the formal operational stage achieved between the ages of 12 and 15 years. The formal model of the INRC group, derived from Klein's mathematical group, provided an account of the new possibilities of reasoning which were available to adolescents. It involved a greater number of combinations of operations than the grouping structures (in particular a 'double reversibility' of thought). Moreover, the experiments by Inhelder showed how the understanding of physical phenomena developed in children and how, through the use of an adequate interview procedure, adolescents could rediscover fundamental scientific laws on their own.

The Last Two Decades (1960 to 1980): The Pinnacle of Fame

Only in the late 1950s was Piaget's work rediscovered by English and American psychologists, although he already had been a most respected professor and researcher in the French-speaking countries, and a well-known author in many other countries. During the following two decades he was among the most cited psychologists and best known scientists in the world. Most of his books were translated from French into foreign languages (a total of 24 different languages). From the end of the 1950s onwards, he received an increasing number of honorary degrees each year. When he died, he had received over 30 honorary degrees and several prestigious awards. However, these honors and the travels they involved did not prevent Piaget from continuing his work.

Until the beginning of the 1970s, studies of genetic epistemology mainly focused on problems related to logic, language, space, time, and causality. The latter topic – strangely neglected for several decades – proved difficult to study and remained on the program of the Center for Genetic Epistemology for several years in succession. On the whole, these books, while discussing issues specific to each science (e.g., the relative importance of time and speed in judgments about duration, or the possibility of geometrical intuition) validated the constructivist conception of Piaget's epistemology. Effective forms of knowledge are neither innate nor drawn from the observation of the environment. They derive from the interaction of active individuals with their environment. A consequence of this position (summarized in the short book *Genetic Epistemology*, 1970a) is that the rate of intellectual development can be only moderately accelerated by education. The goal of education, according to the few texts devoted by Piaget to this topic (for example a long article in 1965) is to favor children's intellectual activity and their discovery ofrelationships and abstract concepts.

During the 1960s Piaget returned to his earliest field of interest: biology. He published a paper on the evolution of mollusks and about the reproduction of the 'sedum' plants, which he grew in his garden and observed carefully. In his book *Biology and Knowledge* (1967) he developed his ideas on biology and the parallel between biological and cognitive development. Simultaneously, he published three books on psychology. They dealt with the 'figurative' aspects of knowledge, whose function is to provide a kind of copy of things: *The Mechanisms of Perception* (1961), *The Child's Mental Imagery* (1966), and *Memory and Intelligence* (1968). These books showed that figurative aspects are progressively subordinated to the subject's operations: some optical illusions decrease with perceptual activity, and with the advent of logical thinking (concrete operations), mental images can adequately represent transformations, while memory of organized material improves. Intelligence is not captured by representation, i.e., a system of images or linguistic symbols; rather intelligence consists in operating on elements of reality or symbols and thus transforming them.

In the 1970s, after Piaget's retirement as professor at the University of Geneva (1971), he continued his work with a large team of researchers at the Center for Genetic Epistemology. Most of his publications dealt with processes of cognitive development. For over 20 years, his main aim had been to demonstrate that efficient forms of knowledge result from structures that organize mental operations, and that these structures, which become increasingly complex over time, are rooted in biological organization. The time had come to deal with the processes involved in the construction of these structures, a topic that Piaget had neglected since the 1930s, except for a few epistemological papers written in the 1950s. During the last decade of his life, between the ages of 75 and 84 years, Piaget developed concepts that had been introduced earlier but needed fuller consideration and a precise definition: equilibration, reflexive abstraction, generalization, the opening up of new possibilities, and so on. These concepts are highly interesting and very original ideas, difficult to operationalize, but with deep and general explanatory value. Unfortunately, Piaget was so eager to continue his work that he changed the focus of his research and his writing nearly every year, although these complex concepts would deserve more sustained treatment.

Reflexive abstraction is the process through which something new is abstracted from experience. What is abstracted is not a property of the external world, but derives from the organization of the subject's actions or operations. Equilibration is a self-regulation process, which is specific to living organizations. At the cognitive level, it is a reaction to an intellectual perturbation (such as a contradiction) which regulates the cognitive system and, in an attempt to compensate the perturbation, creates improved forms of knowledge. For his eightieth birthday in 1976, a kind of dissertation defense was organized: Piaget defended the theses developed in *The Equilibration of Cognitive Structures* (1975) before a 'jury' composed of famous scientists who represented different disciplines. In this last period, Piaget's work also dealt with the grasp of consciousness, contradiction, and dialectics. He explored the relevance to the explanation of cognition of the mathematical concept of morphism and of the idea of logic of meaning. The last work was conducted in collaboration with R. Garcia who was also coauthor of *Psychogenesis and History of Science* (published posthumously in 1983). Piaget also dealt with problems of biology. His post-Lamarckian

ideas attributed some importance to aspects of experience in the evolution of species (for example in *Behavior, a Motor of Evolution*, 1976).

Factors of Scientific Creativity

Piaget died on 16 September 1980, leaving behind a life's work of unusual scope, in terms of depth, originality, interdisciplinary relevance, and volume. The explanation of his exceptional scientific productivity is not only the conjunction of a brilliant mind with a great capacity for work, but also the stimulation of reflection by the results of constantly renewed research. It was also due to Piaget's sociability, which enabled him to easily convince others to work with him; and to a strong sense of a mission to which everything else was subordinate. According to Piaget, having a theory to contradict was a powerful stimulus. The dominant empiricist conceptions in psychology – neoassociationism – and logical positivism in philosophy played this role well into the 1960s. Since the mid-1960s, Piaget was busy developing arguments against the relatively widespread idea, prevalent since Chomsky published his work, that important aspects of knowledge are innate.

The Importance of Piaget's Work for the Social Sciences

Piaget's theory is exceptionally comprehensive, as it gives a detailed account of cognition from birth to adolescence, and also deals with scientific knowledge in adults. It contains an incredible wealth of experimental data covering these periods in life, especially the age range from 5 to 9 years. It deals with fundamental problems often ignored in other psychological theories, such as the relationship between biological activity and cognition, the relative role of the subject and the environment, and the connections between different aspects of cognition such as perception, reasoning, and memory. In some respects Piaget's work was ahead of its time and it could therefore easily establish connections with emerging disciplines such as cybernetics (since the mid-1940s) and, later, artificial intelligence. The concepts of self-regulated systems and of feedback specific to cybernetics were involved, less explicitly, in the Piagetian ideas of reversible structures (present in his work at least since 1937) and of the regulation of cognitive activities (frequently referred to by him in the early 1940s). The artificial intelligence programs which expressed in symbolic language the expertise involved in the mastery of a task bear some resemblance to the Piagetian concept of repeatable action schemes and recall Piaget's attempts to formalize mental structures. Moreover, the multidisciplinary approach known as cognitive science, which emerged in the 1980s, is completely in line with Piaget's approach, with its links between biology, psychology, logic, and physics.

One aspect of Piaget's work, genetic epistemology, which is focused on scientific knowledge as investigated through studies of children, has had little impact outside of Geneva. By contrast, Piaget's psychological work has definitely influenced subsequent orientations and research in the social sciences.

Since the 1920s, Piaget's findings about the reasoning of children had considerable influence on child psychology. The influence was strong in the 1960 and 1970s, in particular regarding the four successive stages of cognitive development, (sensorimotor, preoperational, concrete operational, and formal operational). In fact, Piaget's idea was not to define general stages of child development, but necessary steps in the integration of new forms of knowledge. His detailed findings concerning the emergence of new capacities, such as the search for hidden objects in babies, conservation of quantity in school-age children, and combinatory operations in adolescents remain useful landmarks of psychological development although they cannot be used as psychometric tests. During the same two decades, various experiments in education were directly inspired by aspects of the theory, namely the role of the child's activity and the importance attributed to general structuring abilities rather than specific learning. However, mainstream psychology and education, especially in English-speaking countries, often remained impervious to Piaget's influence, except for the growth of cognitive developmental psychology.

Cognitive developmental psychology became a productive subdiscipline of psychology, especially since the beginning of the 1970s. Practically every developmental psychologist in the world at that time had at least some knowledge of Piaget's work and the very idea of studying cognitive development using a cross-sectional method was inspired by the work previously undertaken in Geneva. Piaget's writing often played the same role of a 'theory to contradict' that associationism or nativist conceptions had played for Piaget. Many authors in developmental psychology launched their careers by conducting experiments that were designed to prove Piaget wrong. Because his constructivist conceptions grant a limited role to social influence, and in particular to education, they were unacceptable to many. Much research has been performed especially in the USA to show that cognitive competencies existed in children well before the ages indicated by Piaget. In order to demonstrate this, the problems presented to the children were modified and compared to those presented by Piaget himself. The race for early competencies in the 1980 and 1990s gave rise to the idea that many cognitive capacities were innate. In the end, Piaget's conceptions were not followed, but thanks to his work a whole field of psychology developed.

Piaget's work paved the way for the cognitivist revolution in psychology, from the mid-1960s onwards. Psychologists had become interested in mental states, a taboo for many decades due to the predominant behaviorist conception. The new generation of psychologists could read in Piaget's work detailed studies of mental states that had been conducted over a period of several decades.

At the end of the twentieth century, Piaget's theory had indirect descendants in the fields of child psychology, cognitive science, and developmental psychology. Ideas of specialists in educational science are more directly inspired by Piaget's work, especially in French-speaking countries and in Latin America. Studies in moral development also derive directly from Piaget's research in this field. More generally, constructivist approaches which attribute importance to the activity of the subject, in applied psychology as well as in other disciplines, usually refer, albeit vaguely at times, to the work of the Swiss psychologist and epistemologist.

See also: Binet, Alfred (1857–1911); Cognitive Development: Child Education; Janet, Pierre (1859–1947); Kant, Immanuel (1724–1804); Kohlberg, Lawrence (1927–87); Köhler, Wolfgang (1887–1967); Piaget's Theory of Human Development and Education; Scientific Concepts: Development in Children; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Bringuier, J.-C., 1989. Conversations with Jean Piaget. University of Chicago Press, Chicago.

Ducret, J.-J., 1990. Jean Piaget: biographie et parcours intellectuel. Delachaux Nestlé. Neuchâtel, Switzerland.

Gruber, H., Vonéche, J. (Eds.), 1995. The Essential Piaget. Aronson, London.

Inhelder, B., Piaget, J., 1955. De la logique de l'enfant a la logique de l'adolescent. P.U.F., Paris [English translation, 1972. The Growth of Logical Thinking From Childhood to Adolescence. Routledge, London].

Montangero, J., Maurice-Naville, D., 1992. Piaget ou l'intelligence en marche. Vocabulaire et aperçu chronologique. Mardaga, Liége, Belgium [English translation, 1997. Piaget or the Advance of Knowledge. Glossary and Chronological Overview. Erlbaum, Mahwah, NJ].

Piaget, J., 1926. La représentation du monde chez l'enfant. Alcan, Paris [English translation, 1972. The Child's Conception of the World. Littlefield Adams, Totowa, NJ].

Piaget, J., 1932. Le Jugement Moral Chez L'enfant. Alcan, Paris [English translation, 1977. The Moral Judgment of the Child. Penguin, Harmondsworth, UK].

Piaget, J., 1936. La naissance de l'intelligence chez l'enfant. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation, 1970. The Origins of Intelligence. Routledge, London].

Piaget, J., 1945. La formation du symbole chez l'enfant. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation, 1962. Play, Dreams and Imitation. Norton, New York].

Piaget, J., 1967. Biologie et connaissance. Gallimard, Paris [English translation, 1971. Biology and Knowledge. The University of Chicago Press, Chicago].

Piaget, J., 1970a. L'épistémologié Génétique. P.U.F., Paris [English translation, 1972. The Principles of Genetic Epistemology. Routledge, London].

Piaget, J., 1970b. Reedited Psychologie et épistémologie. Denoël, Paris [English translation, 1971. Psychology and Epistemology. Grossman, New York].

Piaget, J., 1972a. Reedited Problémes de psychologie génétique. Denoël, Paris [English translation, 1973. The Child and Reality: Problems of Genetic Psychology. Grossman, New York].

Piaget, J., 1972b. Oú Va L'éducation. Denoël-Gonthier, Paris [English translation, 1973. To Understand is to Invent: The Future of Education. Grossman, New York].

Piaget, J., 1975. L'equilibration des structures cognitives. P.U.F., Paris [English translation, 1985. The Equilibration of Cognitive Structures. The University of Chicago Press, Chicago].

Piaget, J., 1981. Le possible et le nécessaire, 2 vols. P.U.F., Paris [English translation, 1987. Possibility and Necessity. University of Minneapolis Press, Minneapolis, MN].

Piaget, J., Inhelder, B., 1941. Le développement des quantités chez l'enfant. Delachaux and Nestlé, Neuchâtel, Switzerland [English translation, 1974. The Child's Construction of Quantities. Routledge, London].

Piaget, J., Inhelder, B., 1948. La représentation de l'espace. P.U.F., Paris [English translation, 1967. The Child's Conception of Space. Norton, New York].

Piaget, J., Inhelder, B., 1968. Memoire et intelligence. P.U.F., Paris [English translation, 1973. Memory and Intelligence. Basic Books, New York].

Vidal, F., 1994. Piaget before Piaget. Harvard University Press, Cambridge, MA.

Piaget's Theory of Cognitive Development

Anik de Ribaupierre, University of Geneva, Genève, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Piaget's theory is one of the major and enduring contributions of the twentieth century to developmental psychology and education, and is firmly grounded in biology and epistemology. This article first reviews some key epistemological concepts, since Piaget's project was epistemological before being psychological. It then briefly summarizes Piaget's theory of the development of intelligence (stages and factors of development). A few major criticisms brought by developmental psychologists are reported, and the major legacy of Piaget's theory is summarized. Present and future directions of research are finally reviewed.

Introduction

It might seem a little bit strange to offer in 2014 an update of a notice on Piaget's theory. By very definition, the notice will adopt a historical connotation; after all no revision to the theory itself has been introduced. Yet, Piaget's theory has been and still is so influential that it has to be included in the revision of the encyclopedia. Moreover, it remains a surprisingly modern theory in view of the later advances in developmental psychology. The update of the notice will evidently concern more closely studies inspired by the Piagetian theory, while the summary of the theory itself will lean very heavily on the notice published in the first edition of this encyclopedia.

Jean Piaget (1896–1980) has been qualified as one of the two giants of psychology of the twentieth century (e.g., Papert, 1999; see also Hopkins, 2011). His theory counts as one of the most influential and enduring theories of cognitive development, and is firmly grounded in biology and epistemology. Although the focus of this article is on the cognitive developmental aspect of his work, it will first review some of the key epistemological concepts that Piaget adopted (such as constructivism, adaptation, equilibration, abstraction), because genetic epistemology and genetic psychology were closely interrelated. Piaget's theory of the development of intelligence (stages and factors of development) will be briefly summarized. After having mentioned a few major criticisms raised in developmental psychology, the major legacy of Piaget's theory is summarized. Unsolved issues as well as some fundamental misunderstandings of the theory by developmental psychologists are discussed. Present and future directions of research are finally reviewed.

The Epistemology of Piaget

Piaget, who trained as a biologist, had as main objective to understand the development of knowledge in the human species, and particularly the development of the scientific and logical thinking; it was not to offer a psychological theory of how children develop. His first originality was to address philosophical questions (such as the origin or the development of knowledge) by empirical means, in particular by relying on the study of child development (supposed to be, at the beginning of his career, a mere 'detour' of some years). He established the discipline of genetic epistemology (the term 'genetic' referring to the concept of genesis or evolution, as proposed by Baldwin), with the aim "to study the roots of the various sorts of knowledge from their most elementary forms on and to follow their development in ulterior levels including scientific thinking," and by grounding it in both historicocritical and psychogenetic approaches. Piaget's epistemology is constructivist, in the sense that knowledge is neither a mere reflection of the external world (realism) nor a projection of preexisting structures of the mind (nativism); it is to be actively constructed; the origins of knowledge are to be found in the practical and cognitive activity of the subject. Piaget's epistemology is also fundamentally interdisciplinary. Note in passing that he created in 1955, with the support of the Rockefeller Foundation, a very successful interdisciplinary center, the Center of Genetic Epistemology, which attracted numerous scientists from the entire world. Many of the ingenious experimental situations were proposed in that interdisciplinary group.

Piaget's primary hypothesis was that intelligence is a form of biological adaptation (e.g., Piaget, 1967), that is, intellectual or cognitive behaviors are the products of an organism, in interaction with its environment, and the end point of biological evolution. This is not to say that psychological phenomena can be reduced to biological ones. Piaget always opposed biological reductionism; intelligence is an organizing activity whose functioning transcends biological organization, by elaborating new structures. The main characteristic common to knowledge and to living organisms is their adaptive character. Adaptation itself, whether biological or cognitive, relies on two mechanisms, defined very early in Piaget's work, namely, assimilation (incorporation of new information into an existing system – see also Baldwin) and accommodation (modification of existing schemes or structures by newly assimilated elements, so as to respond to the demands of the environment). By so doing, actions and operations become coordinated with one another; such coordinations define a new entity, which in turn constitutes a new object for thought and action. This is the process that Piaget labeled equilibration. Contrarily to frequent but too simplistic interpretations, equilibration does not correspond to a stable (even if only temporary) state of equilibrium, but is the process of coordination in itself.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23093-6

As a consequence of the functional continuity that Piaget postulated between biological and psychological entities, a given level of knowledge is always considered to result from a reorganization of the preceding one. Reconstruction not only means more than a mere addition of elements of a lower type in order to attain a higher level, but also implies a total reorganization and a change in scale. One direct consequence of a constructivist perspective is that an epistemologist or a psychologist studying any level of development will ask what characterized the previous level and will stress developmental progression toward higher levels. Piaget always declared that the functioning of a given cognitive system cannot be understood without also understanding how it developed. An essential mechanism for this reconstruction is reflecting or reflexive abstraction, that is, a process by which the subject extracts characteristics or coordinations from actions or operations; it is called 'reflecting' both because it consists of a mental reorganization and because it implies a reflection in a physical sense, or a transfer onto a higher level of what was established at a lower level. Reflecting abstraction is contrasted with empirical or simple abstraction, which bears mainly on observable characteristics of objects or classes of objects.

The Psychology of Jean Piaget

Piaget's psychology of intelligence is the most universally known part of his work, even though he himself did not consider it as the major part (e.g., Piaget, 1947, 1970; Piaget and Inhelder, 1966a). His theory transformed the field of developmental psychology by providing it with a new vision of the development of the child, to the extent that almost every developmental study has some connection with a question Piaget raised. Intelligence is defined as the most general form of coordination of the actions and operations that characterize the various developmental levels, and not as a mental faculty or an entity in itself; it develops through a succession of general stages, defined by overall structures (*structures d'ensemble*). Piaget was, once again, not interested in individual or task-specific performances, but in the generic behavior of subjects in classes of situations; this is what he called the *epistemic subject*.

The successive stages or levels are general structures, which appear in a fixed order of succession, but are not necessarily tied to a given age, and which are such that a given structure n integrates, as a substructure, the preceding structure $n-1$. The number of stages defined in Piaget's publications varied from three to five. Most often, he referred to four stages: (1) the sensorimotor stage (approximately the first 2 years), itself consisting of six substages, or three periods. This stage is characterized by the development of practical schemes, among which is the scheme of object permanence (operationalized by the study of the search for hidden objects), one of the most studied domains in modern infancy research; (2) the preoperational stage from approximately 2–6 years of age, also labeled representative (prerelational and intuitive, or prelogical) period, made possible by the emergence and the development of the symbolic function; (3) the concrete operational stage (7–12 years) during which invariants (conservation), logical classes and relations, the concept of number, as well as different concepts of space develop, all of which are marked by reversibility; and (4) the formal operational stage, during which adolescents develop hypotheticodeductive reasoning (i.e., beyond concrete objects and actions) and experimental schemata. Piaget used abstract algebra to formalize the underlying structures: a group of displacements for the sensorimotor stage, groupings for the concrete operational stage, and a group of four operations (Identity, Negation, Reciprocal, Correlative (INRC) group) for the formal operational stage. He also attempted to formalize the preoperational stage by means of the logic of functions, but this effort was not pursued. Other cognitive activities such as mental imagery, perception, memory, and language were always dealt with by Piaget in relation to the development of operational structures. They were termed 'figurative' because they attempt to represent reality as it appears. Figurative activities are contrasted with (and subordinated to) 'operative' activities that attempt to transform reality (e.g., Piaget and Inhelder, 1966b).

Although Piaget grew less and less interested in the description of these general stages, this is probably the feature of his theory that was most often retained and criticized. In particular, the generality of the stages was questioned by the existence of *decalages*, that is, temporal lags between acquisition of concepts supposed to pertain to the same level. Readers of Piaget have often overestimated the necessity of synchronism in acquisitions across situations as implied by the theory. It should be stressed that Piaget did not postulate that the generality of structures meant strict synchronism, even though some of his writings were admittedly ambiguous with respect to this issue (e.g., Chapman, 1988; de Ribaupierre, 1993).

Factors of Development

Piaget described four factors of development: (1) organic growth, and maturation of the nervous and endocrine systems; (2) experience, which Piaget further decomposed into a factor of physical experience (actions exerted upon objects in order to extract their properties by simple abstraction) and a factor of logicomathematical experience (extraction of knowledge by reflecting upon one's own actions, whether effective or mental); (3) social interactions and transmission; and (4) equilibration. This last factor was undoubtedly the most important for Piaget; he often paid little more than lip-service to the first three factors, whereas he spent considerable time in demonstrating the necessity of equilibration, which he considered to account for the 'vection' or direction of development (Piaget, 1975). Majoring equilibration (*équilibration majorante*) is seen as the causal mechanism underlying the progression of cognitive development. It is a dynamic sequence of states of disequilibrium–equilibrium and embodies the organism's strive toward a more general equilibrium. Equilibration is the only mechanism that, according to Piaget, can account for the emergence of novel behavior, even though more recent work has disputed the idea that Piaget's theory can really account for the emergence of novel behaviors. Cognitive structures are put into disequilibrium by perturbations that resist the assimilatory process and lead to errors or failures of action. Reequilibration is made possible by regulations, which entail the modification of an action as a function of

feedback regarding its previous outcomes, ultimately leading to compensatory reconstructions. Equilibration is thus necessary when contradictions and cognitive conflicts arise, hence the importance given by Piaget to cognitive conflict as a source of cognitive development.

Unsolved Issues in the Piagetian Theory

Piagetian theory has been subjected to numerous criticisms (perhaps proportional to his production) on the part of developmental psychologists, only some of which are summarized here. Most critics should be placed in the context of what Piaget's work was used for by other researchers (psychologists, educators, etc.), rather than directed toward Piaget's project. As mentioned above, Piaget's project was epistemological more than psychological, while most critics addressed the developmental theory of the child. First, numerous decalages, which could not be predicted by the theory, have been observed between concepts supposed to belong to the same structural level; between-task correlations are relatively low, implying that all tasks do not measure the same construct. This led a number of researchers to refute the general character of cognitive development; instead, they claim that development is domain specific. Moreover, the concept of overall logicomathematical structure proved very difficult to operationalize. Second, Piaget's portrait of the child appeared too universal and focused on the subject independently of his/her cultural environment; the biological origins of development were also felt to be too much emphasized. Third, the theory placed too much emphasis on the development of logicomathematical thinking, as contrasted with other forms of thinking. Fourth, Piagetian theory did not provide sufficient information on learning mechanisms and on the effects of training. Consistent with his epistemological project, Piaget was indeed interested in demonstrating the commonalities, in terms of general mechanisms, between learning and development rather than in focusing on the specificities of learning. Linked to this is the difficulty for the Piagetian theory to account for the appearance of novel behaviors. This is not to say that later theories have a better account for the emergence of novelty. Finally, like many other, general psychological theories, Piaget did not take into consideration individual differences; however, it has been shown that the variability in performance, whether intra- or interindividual, is much larger than could at first be suspected and constitutes a fundamental phenomenon rather than mere noise (e.g., de Ribaupierre, 1993; de Ribaupierre and Rieben, 1995).

As an illustration, Hopkins (2011) enumerates three criticisms of the Piagetian theory in the context of the early twenty-first century: (1) Piaget never gave any details about his samples, and provided almost no statistics while sometimes using very few participants; (2) Piagetian stages are too general, whereas present-day developmental theories consider the mind to be modular; (3) Piaget considered a too early age as the end of development (12–15 years of age), and did not consider that major change could take place in late adolescence or adulthood. Once again, these criticisms concern the psychological or educational use that can be made of the Piagetian theory, rather than the epistemological framework. Piaget was indeed not interested in statistics; the most advanced statistics that he provided was descriptive, such as percentages of response categories like in the book *Early Growth of Logic in the Child* (Piaget and Inhelder, 1959/1964). It was sufficient, from the perspective of his epistemological project, to empirically demonstrate the existence of a developmental sequence; it was not necessary to assess whether all children followed the same sequence. Obviously, however, from the standpoint of a psychological theory of development, it is crucial to determine whether all children develop in the same way, and at approximately what age.

The Heritage of Piaget

Despite the criticisms addressed to the theory, Piaget's theory leaves an enormous heritage, some of which is no longer even acknowledged, simply because it has been fully integrated into developmental psychology and education. First, the constructivist view of development that Piaget defended is presently widely held by cognitive developmentalists, as well as by educational and cognitive psychologists. Likewise, the concepts of assimilation and accommodation, even though probably too general, are adopted by a large number of theoreticians. The view of children construed as active thinkers has also been generally endorsed. Second, although there was a period in the 1960s and 1970s during which a large number of developmental researchers argued for domain specificity in development and against the existence of general stages, there is a growing consensus that the characterization of human cognitive development needs something more general than a long list of specific concepts, even if less general than the cognitive structures defined by Piaget. In any case, most developmentalists see the need for structural analyses that transcend any particular cognitive task. Similarly, the concept of scheme (defined as an invariant abstracted across several actions, to which new actions or objects can be assimilated and which allows for generalizations) has been retained and is close to other concepts such as schemas or scripts proposed in cognitive psychology. Third, although the construct of equilibration has been criticized for being too fuzzy, similar conceptualizations have been proposed, whether by dynamic systems theoreticians, or by neo-Piagetians.

Last, but not least, Piaget contributed enormously to the research methods used in developmental psychology, even though he was criticized for not using a standard, quantifiable methodology. He pioneered the use of a clinical method (later to be called critical method), in which the researcher probes for the child's understanding through repeated questioning. Piaget insisted that the child's incorrect responses reflect a complex, underlying cognitive system, and should not be merely recorded as failures. Also, Piaget's work was and still is a source of very ingenious empirical techniques, in many different fields; Piagetian measures such as conservation, class inclusion, spatial representation, perspective taking, object permanence, moral judgment, transitive inference, and many others have each produced countless studies by researchers around the world. Most of these tasks were very productive, because they

generated data that were both counterintuitive and generative of hypotheses about the way children think.

Bond and Tryphon (2007) examined the 1998 edition of the *Handbook of Child Psychology*, one of the most authoritative books in developmental psychology, to assess the impact of Piaget's theory, and compared it with the 1983 edition. Their analysis shows clearly that, as mentioned above, only a restricted period and, more importantly, a restricted set of Piaget's enormous production was retained. Although the number of references shows Piaget to be the most cited author in the field, citations are essentially restricted to his books, particularly those published between 1930 and 1975; most articles (the paper lists 523 articles written) are ignored, except the article that Piaget wrote on his own theory for the 1970 edition of that handbook (chapter reprinted in the 1983 edition). Also, the books most frequently cited are those on the sensorimotor period, on the one hand, and those dealing with the psychological development of the child. As mentioned above, Piaget's epistemological perspective was largely ignored, at least in the Anglophone world, despite the translations and the major importance that Piaget himself granted to this facet of his contribution. This lies in direct continuation of how Piaget's oeuvre was introduced to the North American continent by John Flavell in 1963. It would be interesting to now compare the sixth edition of this handbook, published in 2006; it is probable that the number of references to Piaget's theory will have considerably diminished. As an example, the keyword Piaget does not gather a very high number of references in the index of the recent *Oxford Handbook of Developmental Psychology* (Zelazo, 2013). It is also to be noted that the editor of this impressive handbook, in his introduction, does not quote Piaget, except for the emergence of new representational formats described in the *Construction of Reality in the Child* (Piaget, 1937/1954).

Beyond Piaget's Theory

The Piagetian theory is certainly still very active in developmental psychology and in education; many researchers around the world continue working with Piagetian tasks, even though they do not always acknowledge it, or even know it. From a theoretical point of view, the core part of the theory has been retained while giving rise to new developments, as should be the case with any sound theory. Only some concepts were really abandoned, the best example of which is the logicization of overall structures.

At the time the present notice was first written (first edition of the present encyclopedia), two directions appeared promising for the future while very close to Piaget's proposals, namely, the neo-Piagetian models and the application of dynamic systems modeling to developmental psychology. The objective of the neo-Piagetian models, initiated by Pascual-Leone and exemplified by scholars such as Robbie Case, Kurt Fischer, Graeme Halford, and Robert Siegler in the 1980s and 1990s (e.g., Case, 1992; Case and Edelstein, 1993; Pascual-Leone, 1987; de Ribaupierre, 1997, 2007), was to provide a response to some of the questions left open in the Piagetian theory by combining it with the cognitivist and the learning traditions. These models have in common the search for isomorphisms or invariants across different classes of situations. They maintained or rehabilitated (after the emphasis on domain specificity strong in the 1970s) the notion of general stage; however, the structuralist characteristics are quite different from those postulated by Piaget. First, the hypothesis of a unique, general, logical structure underlying the organization of all behaviors at a given moment of development was abandoned, and replaced by more local ones, thereby granting more importance to context and to specific processes. Second, isomorphisms across domains are considered to result from common constraints or 'strictures' rather than from a common structure: general limits in processing resources or attentional capacity impose a ceiling on the complexity of cognitive performances while allowing for situational and individual variability. Third, the importance given to context has as corollary a much stronger emphasis on task analysis, that is, on an analysis of the characteristics of external situations together with a fine-grained analysis of the strategies that children adopt when confronted with these situations and of the processing load that these strategies entail. Fourth, some neo-Piagetians also assign a more fundamental status to individual differences; consequently, they claim that there may exist different developmental paths for different types of subjects. This implies that development is no longer led by a single, dominant general mechanism such as equilibration, but that several general mechanisms may interact to produce development, leading to a postulate of multidimensionality and multidirectionality in development (Reuchlin et al., 1990; Rieben et al., 1990). Such a postulate has long been adopted by life-span developmentalists (e.g., Baltes et al., 1998), but took longer to be envisioned by child developmentalists.

Another promising way of theorization is provided by nonlinear dynamic systems theories (e.g., Fischer and Bidell, 2006). Such models can be used to detect, and then model, developmental transitions, that is, changes from one stable mode to another. The emphasis is on self-organization, which can be linked to Piaget's concept of equilibration. Although some scholars have begun to use nonlinear dynamics to build specific models of processes of development, particularly in the field of motor development or in conservation tasks (e.g., Thelen and Smith, 1994, 2006; Van Geert, 1994, 2003), these are not yet fully implemented nor operational; moreover, they require heavy methodological designs. Nevertheless, such attempts, which may help uncover discontinuities, i.e., stages, under what appears to be continuous change, could help bring explicit modeling into developmental psychology.

A number of new directions and emphases have also taken place, without explicit reference to Piaget; yet they seem very compatible. To quote only a few of those: theory of mind, role of executive functions, stress on a biological approach. A flourishing direction, that of the 'theory of mind,' was introduced without any reference to Piaget. Yet it is very compatible with the Piagetian theory, even though Piaget and his collaborators did not develop experimental situations in which children would be investigated with respect to their theory on humans in contrast with physical objects, except perhaps in Piaget's very early work (e.g., Piaget, 1932). On the one hand, the type of

paradigm used is close, taking the form of an interview; on the other hand, interest has been brought to the type of errors committed and not only to the age at which theory of mind is 'achieved.' Piaget never used the term 'executive tasks' or 'executive functions,' which constitute a significant part of the present day developmental psychology (e.g., Carlson et al., 2013). Yet, most Piagetian tasks would be considered as executive tasks. The close link established between executive functions and the functioning of the cortex, as well as the rapid growth of the cognitive developmental neuroscience would certainly have pleased Piaget who was criticized for having granted too important a role to biological factors.

See also: Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Cognitive Neuroscience; Cognitive Psychology: History; Executive Functions During Childhood, Development of; Human Development, Theories of; Logical and Hypothetical Reasoning in Adolescence, Development of; Mind, Theories of; Neo-Piagetian Theories of Cognitive Development; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980); Scientific Concepts During Childhood, Development of.

Bibliography

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 1998. Life-span theory in developmental psychology. In: Damon, W. (Ed.), Handbook of Child Psychology, Theoretical Models of Human Development, fifth ed., vol. 1. J. Wiley, New York, pp. 1027–1143.

Bond, T., Tryphon, A., 2007. Piaget's legacy as reflected. In: The Handbook of Child Psychology, 1998 ed. Published online at: http://www.piaget.org/news/docs/Bond-Tryphon-2007.pdf

Carlson, S.M., Zelazo, P.D., Faja, S., 2013. Executive function. In: Zelazo, P.D. (Ed.), The Oxford Handbook of Developmental Psychology, Body and mind, vol. 1. Oxford University Press, Oxford, pp. 706–743.

Case, R., 1992. Neo-Piagetian theories of intellectual development. In: Beilin, H., Pufall, P. (Eds.), Piaget's Theory: Prospects and Possibilities. Erlbaum Associates, Hillsdale, NJ, pp. 61–104.

Case, R., Edelstein, W. (Eds.), 1993. The New Structuralism in Cognitive Development: Theory and Research on Individual Pathways. Karger, Basel.

Chapman, M., 1988. Constructive Evolution: Origins and Development of Piaget's Thought. Cambridge University Press, New York.

de Ribaupierre, A., 1993. Structural invariants and individual differences: on the difficulty of dissociating developmental and differential processes. In: Case, R., Edelstein, W. (Eds.), The New Structuralism in Cognitive Development: Theory and Research on Individual Pathways. Karger, Basel, pp. 11–32.

de Ribaupierre, A., 1997. Les modèles néo-piagétiens: Quoi de nouveau? Psychologie Française 42, 9–21.

de Ribaupierre, A., 2007. Modèles néo-piagétiens du développement cognitif et perspective psychométrique de l'intelligence: y a-t-il convergence? L'Année Psychologique 107, 257–302.

de Ribaupierre, A., Rieben, L., 1995. Individual and situational variability in cognitive development. Educational Psychologist 30, 5–14.

Fischer, K.W., Bidell, T.R., 2006. Dynamic development of action and thought. In: Lerner, R.M. (Ed.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. John Wiley, Hoboken, NJ, pp. 313–399.

Flavell, J.H., 2004. Theory-of-mind development: retrospect and prospect. Merrill-Palmer Quarterly 30 (3), 274–290.

Hopkins, J.R., 2011. The enduring influence of jean piaget. The Observer 24, 10. http://www.psychologicalscience.org/index.php/publications/observer/2011/december-11/jean-piaget.html#.UuaLPQhjyf0.email.

Papert, S., 29 March 1999. Papert on Piaget. Time Magazine Special Issue on "The Century's Greatest Minds", p. 105.

Pascual-Leone, J., 1987. Organismic processes for neo-Piagetian theories: a dialectical causal account of cognitive development. International Journal of Psychology 22, 531–570.

Pascual-Leone, J., Johnson, J., 2011. A developmental theory of mental attention. In: Barrouillet, P., Gaillard, V. (Eds.), Cognitive Development and Working Memory. A Dialogue between Neo-Piagetian Theories and Cognitive Approaches. Psychology Press, Hove, pp. 13–46.

Piaget, J., 1932. Le Jugement Moral Chez L'enfant. F. Alcan, Paris [1932. The Moral Judgment of the Child. Harcourt, Brace & Co, New York].

Piaget, J., 1937. La construction du réel chez l'enfant. Delachaux et Niestlé, Neuchâtel [1954. The Construction of Reality in the Child. Basic Books, New York].

Piaget, J., 1947. La Psychologie de l'intelligence. Colin, Paris [1950. The Psychology of Intelligence. Routledge & Paul, London].

Piaget, J., 1967. Biologie et Connaissance. Gallimard, Paris [1971. Biology and Knowledge. University of Chicago Press, Chicago].

Piaget, J., 1970. Piaget's theory. In: Mussen, P.H. (Ed.), Carmichael's Manual of Child Psychology, third ed., vol. 1. Wiley, New York [Reprinted in Mussen, P.H., 1983. Handbook of Child Psychology, vol. I. Wiley, New York].

Piaget, J., 1975. L'équilibration des structures cognitives. Presses universitaires de France, , Paris [1977. The Development of Thought: Equilibration of Cognitive Structures. Viking Press, New York].

Piaget, J., Inhelder, B., 1959. La genèse des structures logiques elementaires. Delachaux & Niestlé, Neuchâtel [1964. The Early Growth of Logic in the Child. Routledge and Kegan Paul, London].

Piaget, J., Inhelder, B., 1966a. La psychologie de l'enfant, second ed. Presses universitaires de France, Paris. [1969. The Psychology of the Child. Basic Books, New York].

Piaget, J., Inhelder, B., 1966b. L'image Mentale Chez L'enfant. Presses universitaires de France, Paris [1971. Mental Imagery in the Child. Routledge & Paul, London].

Reuchlin, M., Lautrey, J., Marendaz, C., Ohlmann, T. (Eds.), 1990. Cognition: l'individuel et l'universel. Presses universitaires de France, Paris.

Rieben, L., de Ribaupierre, A., Lautrey, J., 1990. Structural invariants and individual modes of processing: on the necessity of a minimally structuralist approach of development for education. Archives de Psychologie 58, 29–53.

Thelen, E., Smith, L.B., 1994. A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge.

Thelen, E., Smith, L.B., 2006. Dynamic systems theories. In: Lerner, R.M. (Ed.), Handbook of Child Psychology, Theoretical Models of Human Development, sixth ed., vol. 1. John Wiley, Hoboken, pp. 258–312.

Van Geert, P., 1994. Dynamic Systems of Development. Harvester Wheatsheaf, New York.

Van Geert, P., 2003. Dynamic systems approaches and modeling of developmental processes. In: Valsiner, J., Conolly, K.J. (Eds.), Handbook of Developmental Psychology. Sage, London, pp. 640–672.

Zelazo, P.D. (Ed.), 2013. The Oxford Handbook of Developmental Psychology. Body and mind, vol. 1. Oxford University Press, Oxford.

Piaget's Theory of Human Development and Education

Azwihangwisi E Muthivhi, University of Cape Town, Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by J.J. Vonèche, volume 17, pp. 11437–11440, © 2001, Elsevier Ltd.

Abstract

The article examines and reinterprets Piaget's central theoretical tenets and their implications for contemporary educational practice. The author argues that Piaget's theory posits human activity as the basis of human knowledge and human development. For Piaget, human activity drives and generates development, at each stage, creating qualitatively new ways of knowing and relating to the world. This interpretation emphasizes the connectedness of mental processes to the world in which, and through which, development takes place and the agentive role played by subjects who transform their world through active engagement. The implications that the theory has for education and the applications that derive from Piaget's theoretical formulation is examined.

Introduction

Born in Neuchatel, Switzerland, in 1896 – and having a solid background in the biological sciences and obtaining a PhD at the age of 22, after studying the behavior of molluscs, Piaget's theory of human development grew out of the contradictions of his early to mid-twentieth century society, contradictions that characterized the apparent tranquil of the society in which young Piaget grew up. Perret-Clermont (1997) provides an examination of the sociocultural context in which Piaget developed his theory to illuminate the meanings Piaget placed on his system of theorizing, most of which may remain implicit in contemporary debates of this theory. The social context of Piaget's Switzerland, as Perret-Clermont suggests, was characterized by conflict between religion and science as well as forms of traditional authority structures. Meanwhile, scientific Darwinism was in its ascendance and provided supposedly rigorous scientific account of biological and social processes (see Stetsenko, 2008; Perret-Clermont, 1997).

Piaget formulated a theory of human development that derived its explanatory principles from biological processes without, however, reducing complex human psychological phenomena to purely biological adaptation (Piaget, 2001). Therefore, Piaget sought to provide an account of the origin and development of knowledge that emphasizes relational processes between the individual and his or her environment or social situations. The relational processes are based on activities carried out simultaneously in the social and material world in which, and through which, development occurs; leading to transformation of self and – at the same time and within every stage – acquisition of qualitatively new modes of relating to the world; as well as new modes of knowing.

This article provides a general outline of Piaget's theory of human development and examines the implications this theory has for education. The discussion begins with a brief outline of the behavioral accounts of learning and development to which Piaget's theory stands opposed. Piaget's theory of human development is examined in more detail, with the main tenets of the theory and its major conceptual grounding elaborated. Lastly, the application of the theory to education is examined with a view to demonstrating the continuing relevance; and enduring significance, of Piaget's theory of human development to education.

Piaget versus Behavioral Explanations

Piaget's greatest impact in the social sciences lies – arguably, in his critique of the behaviorist explanatory framework and its depiction of human psychological processes as comprising habit formation produced through reactions to the effects of environmental stimuli (see Piaget, 2001, 1964). Behaviorism, as psychological tradition, sought to explain human psychological functioning using the same explanatory principles as those used in the natural sciences; thus seeking to establish the causal relationships between environmental conditions on the one hand and human behavior on the other – with the former conceived in terms of stimuli and the later in terms of responses. The stimulus–response framework, as the behaviorist tradition was alternatively called, posited that human actions, including learning, were the result of conditioning or habit formation as a result of positively reinforced responses to stimulus effects. Thus, the human mind was conceived of in terms of external associative relations between responses or reactions to the effects of environmental forces termed stimuli. Behaviorism ruled out the existence of mind as it assumed this to be not amenable to scientific verification through experimental methods similar to those employed in the natural sciences. Only studying the growth of behavior as resulting from the process of habit formation; occurring through conditioning or habituation of positive experiences, could, according to behaviorists, lead to a valid, scientifically rigorous form of knowledge of the human condition (Piaget, 2001, 1964).

While Piaget's theory resonates with social learning theorists such as Bandura and Mischel, the position of the social learning theorists further supports the view of human subjects as purposive, goal directed, and self-regulative rather than driven by environmental influences. As in Piaget's assertion about the centrality of human self-regulative processes and, hence, active self-directedness, Bandura (2001) posits a view of human agency premised on the essential conditions of human intentionality, purposeful pursuits, and self-reflectiveness. This view, as in Piaget's system, contradicts the mechanistic view of

human nature; proposing a picture of human condition as essentially subordinate to, and arising as mere responses to, external forces of environmental stimulation. For Bandura (2001) human activity is essentially agentic and driven; not by nature itself, but by a new functional structure created through and out of human agentic conditions.

In this system, humans exercise self-control over their thoughts and actions rather than being mere reactionaries to external forces. Their actions and thoughts are agentic and therefore involve qualities of intentionality, purposefulness, goal-directedness, forethought, planful proaction, aspiration, self-appraisal, etc. These agentic qualities function within fundamentally active (rather than reactive) and self-regulative human essence.

In a similar vein, Mischel (1973) states that self-regulatory systems involve self-control through self-imposed goals and self-produced consequences for action. That is, persons have the capacity to set performance goals for themselves and react to these with self-criticism or self-satisfaction, depending on whether they deem their performance consequences as meeting their expectations and criteria. Achievement of performance goals will most likely lead to positive self-appraisal, self-gratification, and self-reinforcement, while perception of failure may result in self-condemnation, anxiety, and desperation. The essence of self-regulatory system, according to Mischel (1973), is the subject's adoption of contingency rules, which are important to guide action and specify which behavior is appropriate under specific conditions. Social learning theories resonate with Piaget's system and could be employed to expand on its explanatory power, especially where personality differences in cognitive performance are to be accounted for. Piaget's unparalleled contribution to developmental psychology resides in his account for the genetic origin of active; and self-regulative, essence of human condition through his theory of human development.

Piaget's Theory of Human Development

Piaget posited an all-important construct of equilibration, on the basis of which internal mental processes can be comprehended. For Piaget, the child did not acquire knowledge through reacting to the effects of environmental forces and, as a result, building habits that become crystallized into forms of behavior. The construct of equilibration suggests that the subject is fundamentally active in his or her interaction with the social and physical environment and organizes his or her actions into schemas and structures, on the basis of which further construction of knowledge becomes progressively possible. That is, the subject regulates his or her own actions in the world and, in turn, these actions and the consequences they have, internally, into means for controlling own thinking; by establishing internal anticipatory schemas and structures on the basis of which further acquisition and development of knowledge becomes possible. The concept of equilibration, therefore, unlike the stimulus–response concept of behavior, resuscitates the idea of mind – an abstraction which defines complex forms of human consciousness. From Piaget's point of view, therefore, a rigorous, scientific methodology (scientific – though not in positivistic sense) that employs observation and clinical experimentation could be employed to study the origin, development and functioning of the human mind; without need to resort to the causal explanation of the behaviorist framework.

Equilibration involves two complementary processes of assimilation and accommodation. Assimilation involves an active integration of knowledge into preexistent mental structures. Accommodation involves the internal balancing processes whereby modifications within the existent structure or mode of knowing are carried out, so as to comprehend new knowledge effectively and hence, retain a state of balance or equilibrium to the hitherto disequilibrated structures (see, Piaget, 1964, 2001). Accommodation happens when a state of necessity for change arises on the part of the subject or when the subject becomes aware of discrepancies between his or her state of knowing and new knowledge to be assimilated. Therefore, the two processes of assimilation and accommodation are complementary, the former accounting for the process of learning while the latter accounting for development. That is, development is accounted for as involving qualitative changes that happen in existing structures, necessitated by the realization and the internal need for change, brought about by the awareness of a state of imbalance or state of disequilibrium between new knowledge to be acquired and existing forms of knowledge on the part of the subject.

Piaget argued that learning is explained through the process of assimilation – which, although it involves an active; rather than passive, integration of knowledge into preexistent structures, does not, however, change the structures or existing modes of knowing. What changes, rather, is the knowledge being assimilated, as this is transformed so as to fit into existing mental structures. Thus, knowledge of the world is not imprinted into the subject and the environmental forces do not determine knowledge to be acquired, in the sense assumed within stimulus–response systems. Rather, the subject is active – not in an aimless, purposeless, and trial-and-error fashion characteristic of the operant conditioning models of later stimulus–response frameworks; that sought to explain human learning in ways similar to habit formation processes established through animal behavior-change or conditioning experiments. Piaget's subject, or a child, is active in a purposeful manner where he or she acts on the world to master and comprehend it, and this action leads to internal mental coordinations; a process explained through the concept of equilibration.

Equilibration is therefore a motor for human development and this self-regulatory process emphasizes an active role the subject has in the acquisition of knowledge; which, although arises externally through the relational processes involved during the subject's interaction with the world, is, however, coordinated from within, since it is the subject, ultimately, that acts and comprehends the world in terms consistent with his or her existing structural organization. There is a sense of a biological adaptation involved in this process. However, Piaget termed this functional adaptation, in that it is not adaptation to the external world per se, but to the evolving structural organization or internal coordination of knowledge (Piaget, 2001). That is, the subject does not adapt to fit his or

her existing knowledge to the form of knowledge encountered from the outside, in his or her context of learning and development. Rather, the subject adapts to his or her existing; and yet, evolving conception of the world that he or she continuously modifies as and when the need – or a feeling of necessity, arises because the present state of knowledge is now deemed inadequate and creates dissonance that needs to be compensated. Piaget (1964), using the language of the behaviorists to critic their model of learning, argued that the response was present long before the stimulus effects – in the sense that it is not the external environmental forces that determine; in an associative form, how the subject reacts, but the subject's own, already constructed modes of knowing or knowledge structures.

Therefore, Piaget provided a solid, genetic account of human intellectual development grounded on human activity. First, the explanation emphasized genesis or developmental origin and, therefore, solving the problem of nativists or philosophical frameworks that subscribed to the notion that knowledge was inborn. Second, the account emphasized developmental progression – conceived, not as a quantitative accumulation of knowledge; as the behaviorists have contended, but as involving qualitative transformation of knowledge from its simpler forms to its more complex forms; characteristic of adult forms that – ideally, enables elaboration of logical and scientific knowledge often privileged in social institutions such as formal schooling.

From a genetic point of view, action; according to Piaget, comprises the basis of human knowledge. Unlike the behaviorists, action – for Piaget, did not only involve the physical aspect of human behavior. Action is both concrete and physical, as well as mental and abstract. This account of action simultaneously elaborates on the essential nature of experience, which in the stimulus–response system has been assumed to constitute only the external aspect of human interaction with the physical and social environment. Piaget, on the contrary, argued that action comprises both the external experience of the world involving; first, the subject's interaction with concrete objects in the world and; second, the consequences of actions initially effected to objects in the world. That is, the subject experiences the world through his or her direct interactions with it but eventually turns these experiences inward into his or her internal plane, where coordinations are subsequently conducted mentally. The former type of experience is considered to lead to operative forms of knowledge – that is, knowledge of how to relate to the world externally while the later type of experience is considered to lead to figurative forms of knowledge, which enables the knowing of the world through mental abstractions. Therefore, experience of the former type is externally directed toward the world while the experience of the later type is internally directed, toward mastery of self. It is, therefore, this mental action that leads to mental coordinations. The coordination of a series of actions leads to what is termed a mental structure. A series of these structures combine to form an operation. Piaget defined an operation as a series of actions that are reversible; that is, actions that have been organized into a reversible structure. These operations become the basis of meaningful, reflective activity through which the human mind is able to engage in interpretive, comparative, evaluative, and analytic activities (Piaget, 1964, 2001).

Sensory-Motor Stage

The genetic approach traces the origin of operational knowledge from its developmental origin in the earliest forms of human activity – the sensory-motor actions, which serve as the early forms of mental coordinations. The coordination of sensory and motor actions leads to the construction of schemas or early structural organization of thought; leading to the acquisition of the concepts of space, time, causality, and object permanency. For example, Piaget found that children acquired the notion of object permanence around 18 months. That is, the child becomes aware of the fact that objects continue to exist even when they have been removed from its spatial field – and do not, therefore, cease to exist. The maturational processes of the nervous system are connected to the development of sensory and motor activities and therefore make possible; and are transformed into, developmental processes. The child progressively masters the world through interacting with it – assisted primarily by its sensory and motor activities and as a result begins to construct – mentally, the functional knowledge of the world around it.

Piaget (1964, 2001) provides an example of a sensory-motor child attaining the concept of object permanence through the coordination of experiences and hence the acquisition of a notion that objects do not disappear and cease to exist when removed from visual field. As evidence to the acquisition of this concept is the child's changed behavior with regard to his or her interaction with objects; where by 18 months the child begins to search for objects outside of its visual field, with the goal of finding them located somewhere. With this acquisition is, of course, also the acquisition of the related concept of space, which arises out of the special coordinations that characterizes sensory-motor activities.

The end of the sensory-motor stage heralds the dawn of preoperational thought around the age of two. By this stage, one begins to see the emergence of structural organization of thinking in the form of operations or systems of actions that are internally connected. However, these early forms of thought structures are characterized by the lack of 'mobility' or backward-and-forward movements – a fluidity that progressively resembles the logical and mathematical forms of thinking considered; within this theory, to constitute the apex of knowledge development. While the sensory-motor stage enables mastery of the physical world through coordination of experiences; a process that weans the subject from greater dependence on the effects of environmental forces characteristic of perceptual processes, the preoperational stage increases this distance from reliance on the effects of the environmental forces. This process is primarily facilitated by the development of symbolic forms of relations that characterize the social world and; specifically, the use of language (Piaget, 2001, 1964).

Preoperational Stage

Piaget posited that the preoperational stage emerges around the age of 2 years and concludes at around the age of 4/5 years.

There has been much debate of the actual age range at which this stage concludes. This is not surprising considering that Piaget's experimental work emphasized children's deficient performances on tasks that required reversible forms of mental operations for successful completion – an approach which, perhaps, was unduly influenced by his early work with children at Binet's laboratory. At this stage, Piaget found that children could not perform conservation tasks competently.

Piaget described the preoperational stage predominantly in terms of the child's inability to carry out operational thought or to understand conservation tasks – involving the ability to understand forms of knowledge that require reversible or reflective thought processes. That is, preoperational children were found to lack the ability to comprehend situations where two or more categories are to be held constant – in perspective, in order to arrive at valid conceptualization of a situation.

For example, Piaget presented a number of experiments in which preoperational children were not able to conserve or comprehend quantity, weight, volume, etc. One of the most common of these conservation experiments involved the conservation of quantity that is transferred from a shorter and wider glass container to a taller and narrower glass, thus showing a changed appearance in the taller container; falsely suggesting an increase in quantity. The transfer of quantity is performed in the presence of the child and the child is asked if we now have more or less quantity in the second container as a result of the transfer process. Preoperational children were found to lack full comprehension of the situation, focusing exclusively on the changed appearance of quantity in the taller glass and, as a result, insisting on the notion that there is an increase in quantity (Piaget, 2001, 1964).

Piaget argued that a preoperational child lacks the ability to hold two categories constant and to use full information to arrive at a solution to the problem. That is, the preoperational child is inclined to focus only on one aspect of the problem and to use only this to arrive at a solution of the problem because he or she does not as yet have cognitive ability to hold two or more perspectives constant and to use these simultaneously as cognitive resources for getting to a solution for the problem. This lack is necessitated by the nature of the developmental stage that characterizes the child's present cognitive functioning. The preoperational child must consolidate on his or her knowledge of the world that the sensory-motor activities have enabled him to construct and progressively integrate this knowledge with the form of knowledge that derives predominantly from his or her internal, mental coordinations – a form of knowledge that should lead, not just to mastery of the external world, but also to the mastery of self.

During the preoperation stage, the child's cognitive development is dominated by the need to master self and actions are predominantly directed towards self-mastery and, as a result, the child lacks the ability to 'decenter' (Piaget, 2001). That is, the preoperational child lacks the ability to see things from other people's perspective as he or she insists on own point of view. Therefore, the dominant mode of thinking during this stage is egocentric, making it difficult for the preoperational child to coordinate thinking to arrive; through analysis, at valid solution to the problem. This inability could indeed have negative impact on the acquisition of numeracy and literacy skills should the preoperational child be placed within formal learning situations in school. This difficulty may arise, for example, from the inherent nature of formal learning such as the requirement in reading acquisition; simultaneously, to focus both on the phonemic (requiring the understanding of relations between speech sounds and letters of the alphabet) and the semantic (or meaning-making) aspects of the reading process.

However, an undue emphasis on the cognitive deficiencies of the preoperation stage, as was characteristic of the dominant approaches in Piagetian cross-cultural research, was unfortunate. Positive developmental milestones are achieved during the preoperational period, and this stage serves as good preparation for acquisition of operational forms of thinking. For example, preoperational children acquire the capacity for symbolic representation, of which language development is a crucial component. The preoperational stage is crucial for children to gain mastery of themselves in relation to their growing participation in the world of adults. The preoperational stage is characterized, among other things, by the development of language. Language development during this stage, however, takes a particular form that is consistent with the character of the stage itself. The child acquires language and its symbolic function, but the entire acquisition is oriented toward self-mastery. For example, the preoperational child uses same words acquired during interaction with adults but attaches idiosyncratic meanings to these. Also, the preoperational child may verbalize his or her thoughts as a way of controlling own actions, such as during situations when the child needs to commit something into memory. Meanwhile, the preoperational child plays with toys or objects in his natural environment in creative ways. For example, a brick may be turned into a bus; a piece of wood may become an airplane, etc. In this way, the preoperational child acquires the ability for symbolic representation, important both for language development and social functioning; as well as for learning the various subject matter during formal schooling later on when the child has reached school-going, concrete-operational stage. The significance of egocentric forms of language in particular; and thinking in general which characterize the preoperational stage, has – interestingly, been elaborated outside of the Piagetian framework (see Vygotsky, 1986 for this elaboration).

Concrete Operational Stage

Fortunately, for the school-going child, the forms of thinking required for successful learning of the formal knowledge characterizing the subject matter has already appeared in the child's mental capacities. The cognitive capacities for reflecting on own and other people's perspectives; for comparison and analysis, has already made their first emergence into the cognitive repertoire of the child. That is, all the deficiencies that characterize the preoperational stage are overcome at the concrete operational stage. However, the child still needs the support of concrete activities to carry out operational processes comprehensively. The concrete operational child can, for example, make the necessary analysis of the situation; taking into account all the information available to him or her, and arrive at valid conclusions about the quantity transferred from one glass

container to the other as described earlier. The concrete operational child would insist that the same quantity remains after transfer to the second container because the activity of transferring from one container to the other did not comprise adding to; or taking out from, the available quantity. Therefore, quantity is conserved because all available information constitutive of the situation is taken into account in arriving at a solution to the problem. According to Piaget, this knowledge is not attained through simple accumulation of experience. It is the result of the internal self-regulatory activities of the human mind and comes about as a result of qualitative changes that comprise human mental development.

However, while the concrete operational child is able to understand conservation tasks such as the one involving the constancy of liquid quantity, his or her thought processes are still reliant on concrete activities. That is, thought coordinations that involve reflective and reversible systems of ideas and concepts still rely, predominantly, on concrete situations and objects in the subject's experiences. This idea of concrete operations has mostly been misinterpreted to mean learning through concrete objects. However, while concrete objects could be crucial in getting the preoperational child understand complex ideas, they should always be employed as a means to an end and not as an end in themselves; the end being to generate capacity for analysis, comparison, and reflective activities carried out mentally without dependence on external props. For example, teaching children to carry out basic arithmetic operations of addition, subtraction, division, and multiplication through use of pebbles or abacus (which is a form of employment of concrete objects) should always be aimed at getting them to master these skills 'internally' or inside their mind as abstract cognitive activites. That is, the learning that happens here should be oriented toward the interiorization or internalization of these concrete activities into the internal plane of cognitive functioning.

In terms of children learning in school, the concrete operational stage is critical for laying good foundations for successful learning. Many things could go wrong during this period in cases of poor teaching and learning provision. For example, Muthivhi (2010) and Muthivhi and Broom (2009) studied concrete operational and formal operational learners whose learning was characterized by the forms of teaching that did not prepare them adequately for; and elaborate on, the reflective processes necessary for successful learning of the formal subject matter knowledge. These studies found that, although these learners demonstrated levels of performance consistent with normal developmental stages, they tended to be overreliant on concrete manifestations of the tasks and often derived conclusions to problems from their perceptions of the concrete situation rather than from their conceptualizations of the situation; taking full account of all available information to arrive at valid conclusion.

Formal Operational Stage

During the concrete operational stage, the child is consolidating a particular form of thinking that characterizes the social world of adults – the world of reflective, interpretative, comparative, and analytic activities; all of which are activities of the human mind, which, of course – as Vygotsky (1978) has argued, are dialectically related to the activities of society and culture. Piaget termed this stage in human development, the formal operational stage and focused predominantly on the acquisition and development of the formal activities of the human mind, which he alternatively referred to as the 'logicomathematical' processes (Piaget, 2001, 1964). We can understand this specific emphasis to be influenced by the idea of the 'true forms' of human thought, unencumbered by social traditions and adult impositions, which; of course, had its history in the philosophical traditions of Emmanuel Kant's proposition of the basic categories of thought as involving 'number, space, causality, and time' (Piaget, 2001, 1964; see also Shayer, 1997).

Therefore, Piaget, sought to understand how human individuals – what is termed ontogenetic development – came to acquire the basic categories of the human mind that underlie the formal processes of human thinking; characteristic of the formal activities of scientific and mathematical practices. As a result, logical processes of thought; their origin and development from early adolescent to adulthood, became the main emphasis of Piaget's research. Where he included psychological processes other than those related to the logical processes; such as language development, the role of the social and emotional development, the focus was on how these processes were related to, and could be explained from the vantage point of the functioning of logicomathematical processes.

For Piaget, the formal operational thought, the onset of which begins in early adolescence, marks children's first entry into the world of adults; in the sense that children begin to acquire the forms of thinking and attitudes to the words that are primarily informed – not by how things appear in the world – but by how the world makes sense to them from the point of view of their own interpretations and analysis of situations. That is, they enter the world of thoughtful existence; the world of abstraction, of logical and propositional forms of thinking where the world – and knowledge – is; essentially, a product of mental coordinations. Formal operational children, or – more precisely, adolescence, may; for example, begin to question the status quo and existing social traditions as well as the taken-for-granted world of adults, even as they seek to enter and have meaningful existence in it – hence the radical changes often experienced in the behavior of adolescence.

Clearly, this stage presents positive challenges and opportunities for teachers as these learners are naturally very curious and inquisitive. Piaget's framework, therefore, points to a model of education that is explorative, investigative; as well as directed by the curiosity and inquisitiveness of learners, and curriculum development initiatives that could exploit this golden opportunity to engage learners and involve them in an ongoing quest for, and creation of, knowledge. However, young adolescents are often put off by the traditional forms of education, still widespread all over the world, which focuses on drilling of stale, factual knowledge, often putting pressure on learners to learn most of the subject matter by heart or through quasi-meaningful processes of committing rules and procedures to memory so they could remember them for use as tools to work out solutions to given problems (see Arievitch and

Table 1 Major developmental stages, their approximate age ranges, and the dominant activities that characterize them

Major developmental stages	*Approximate age ranges*	*Dominant cognitive activities*
Sensory-motor	0–18 months	Maturation, sensory and motor coordination
Preoperational	18 months to 5 years	Self-mastery through egocentricism, symbolic representation and early language acquisition
Concrete operations	5–11 years	Decentration, self-monitoring, reflection, analysis, and comparison necessary for logical–operational activities applied to concrete situations and concrete objects.
Formal operations	12 years to adulthood	Logical thought processes proceeding from abstract categories and less dependent on concrete objects and situations. Self-monitoring, self-reflection, analytic and comparative activities, and increasing abstraction of thought processes.

Stetsenko, 2000 for a detailed discussion of traditional methods and their epistemologies). It was for this reason, therefore, that Piaget opposed and doubted the capacity of educational processes to contribute to children's development and their desirability as a means of accelerating development (Table 1, summary of the stages of development).

Application of Piaget's Theory to Education

Piaget's theory has important implications for educational practice. Its influence on modern education is unparalleled. Importantly, Piaget's theory has contributed immensely in transforming the way in which children are conceptualized and hence, how they could more effectively be educated. For Piaget, as could be seen from the foregoing discussion, children are essentially active and do not therefore learn by passive reaction to stimulation around them. That is, they do not learn by passively receiving knowledge from the world around them but actively; through their own agency, transform the world in which they live in order to know it. That does not mean, as others assumed, that children should be left to their own devises or that adults have no role to play in the life of the developing child. However, Piaget cautioned against schooling, and related social systems that provided a child no room to come to the knowledge of the world through his or her own intellectual means; denying children their inherent disposition to know the world in ways that make genuine sense to them.

Piaget posited a child who is actively constructing his or her knowledge of the world through transforming it and thereby also transforming himself or herself. That is, the child knows the world through his or her existing but yet transforming (rather than just static) intellectual capacities. These capacities – as the discussion of the inherent connectedness of the stages to the dominant forms of the child's activity in the world has shown above – are not reducible to individual or social moments of their manifestations. Therefore, learning and development in Piaget's framework, although foregrounding the primacy of the individual in ontogeny, does not suppose a child or childhood as completely autonomous of, and discontinuous with, the society and culture in which children's learning and development take place.

Therefore, teaching informed by the Piagetian theoretical system would proceed from the assumption that children at specific stages of development; such as the age of 6 years, would be able to carry out reflective processes, provided that these are supported by concrete phenomena with which they engage and on which their cognitive functions are applied. The concrete nature of thought processes for the concrete operational child is, however, not limited to the use of concrete objects. Although the use of concrete objects is important and should be encouraged at this stage, this should, however, be employed as a means for getting the child master complex demands of tasks such as carrying out number of operations that may require the support of counters or abacus, for example. That is, the employment of concrete objects becomes a means to an end, the end beings the transformation of the child's thinking towards abstract forms of thinking; where thought processes are increasingly carried out without the need for reliance on concrete support. That is, learners at elementary school levels such as those in the age ranges of 5.5–7.5 years could benefit from instruction that employs concrete objects and examples to illustrate complex ideas and models, as well as arithmetic principles and operations. However, such employment of concrete objects and examples should be employed as a means for helping learners to acquire an idea or principle that they would later employ on their own in solving related problems; and base their future learning of complex knowledge on the foundational ideas and concepts initially acquired through the help of concrete objects and empirical illustrations.

Symbolic or semiotic systems that underpin the learning of reading and writing as well as basic number operations could be acquired in this way. An understanding of the principles for a pedagogy that derives from Piaget's system may be crucial for educators to relate knowledge to early school learners more effectively in ways that contribute deep level learning and development. For example, Piaget's system posits that children during early formal school years are capable of employing complex mental capacities, involving holding two or more perspectives simultaneously. Contemporary research on school children's use of cognitive capacities; comprising the ability to hold two or more perspectives simultaneously during problem solving, suggests that while these capacities may owe their development in children's spontaneous

activities (everyday-life activities that do not necessarily involve teaching), their elaboration and conscious application to problem situations may be constrained if classroom instruction does not foster their conscious application to problem situations (see, Muthivhi, 2010; Muthivhi and Broom, 2009; Shayer, 1997).

For Piaget, the acquisition for this cognitive capacity demonstrates an important shift from self-centered, egocentric thought processes – where the child is not able to consider two or more relevant aspects of a situation to arrive at a valid conclusion – to a 'decentered', self-regulative and reflective form of thinking that characterizes operational thinking of school-going children. Piaget (1964, 2001) has posited symmetrical social relations as necessary for the development of self-regulative thought processes. Symmetrical relations involve relations of mutual epistemic respect and epistemic equality. Epistemic, in the sense that the equality and mutuality involved is necessarily in respect of intellectual engagement, exchange of ideas, adherence to the rules of engagement in pursuit of knowledge and novel ideas, as well as models of knowing the world and self. This view contrasts the assumption that the social in Piaget is solely, and exclusively, in respect of children's relations to – and collaboration with, each other (see, for example, the discussion of the social in Piaget's theory in Shayer, 1997; Perret-Clermont, 1997; Labouvie-Vief, 1996). Contrarily, the notion of mutuality can be employed by teachers to model forms of classroom engagement that foster relations of epistemic equity and mutuality that generate pedagogic situations conducive for genuine learning and cognitive development (see; for example, Ginsburg and Allardice's, 1984 Piaget's inspired model). Meanwhile, the notion of epistemic equity and mutuality can also be employed to organize group-based, collaborative learning activities that contribute to learners' self-regulative, reflective engagement with the world and with novel ideas that come from the subject matter of their classroom learning.

Vygotsky's Framework as a Competing Paradigm

It was, ironically, in its capacity for application to educational settings; to provide solutions to most pressing challenges of educational underperformance, conceptual understanding, and task mastery, that Piaget's theory of the origin and growth of knowledge has failed to make significant advances. Piaget's theory was particularly critiqued for not taking adequate account of the social and cultural factors in its account of cognitive development and functioning and therefore as limited in its possibilities for application to classroom teaching and learning situations. Vygotsky's theoretical system was, on the contrary, credited for doing precisely what Piaget's theory was considered to have failed to achieve, that is, foregrounding the role that sociocultural processes play in human development (e.g., Matusov and Hayes, 2000).

The most powerful among current critiques of Piaget's system probably relate to his view of development as proceeding by way of adaptation – as a form of biological adaptational process. Vianna and Stetsenko (2006), for example, argue that Piaget, by positing a view of children as developing and learning in the spirit of adapting to existing conditions, and mind as evolving out of actions through which people adapt to the world, presents a limiting picture of the human condition. Vygotsky, on the contrary, positing a view of children developing and learning as they actively change the world they live in – simultaneously changing themselves and gaining knowledge of themselves and of the world, presents a position that transcends the constraining picture of humans as changing to fit in with the world into which they contribute nothing towards its transformation.

Conclusion

Piaget's theory of human development has made an enormous contribution to contemporary educational practice. By revealing that children think differently from adults; that children are not little adults in the making, and that their thinking differs from that of adults not only in quantitative terms but in the quality of thought processes involved, Piaget made a lasting contribution to our knowledge of how teachers could appropriately relate to learners, as well as how knowledge could effectively be organized at different stages during children's development.

The theory inspired many educational models such as learner-centered approach, discovery-based learning, activity-based learning, collaborative learning, constructivism, exploratory learning, and active enquiry-based learning models. These, and related educational models that derived from Piaget's theoretical systems, had an immense influence on many educational systems across the world. All of these models, inspired by Piaget's framework, proceed from a common ontology, a worldview that learning is a process that is fundamentally driven, first and foremost, by the activity of the learner – and not necessarily by the actions of teachers or needs of society, with existing social structures serving as unalterable *telio* or end-point of human development.

See also: Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Self-Regulated Learning: Theories, Measures, and Outcomes; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Arievitch, I.M., Stetsenko, A., 2000. The quality of cultural tools and cognitive development: Galperin's perspective and its implications. Human Development 43, 69–92.

Bandura, A., 2001. Social cognitive theory: an agentic perspective. Annual Reviews Psychology 52, 1–26. www.annualreviews.org (accessed 28.01.14.).

Fontana, D., 1981. Early social development. In: Fontana, D. (Ed.), Psychology for Teachers. British Psychological Society & Macmillam Publishers Ltd., Leicester.

Ginsburg, H.P., Allardice, B.S., 1984. Children's difficulties with school mathematics. In: Rogoff, B., Lave, J. (Eds.), Everyday Cognition: Its Development in Social Contexts. Harvard University Press, Cambridge, MA, pp. 184–219.

Labouvie-Vief, G., 1996. Knowing and relating: the lost dimension of knowledge in Piaget's theory. Culture and Psychology 2, 323–332.

Mischel, W., 1973. Toward a cognitive social learning reconceptualization of personality. Psychological Review 80 (4), 252–283.

Matusov, E., Hayes, R., 2000. Sociocultural critique of Piaget and Vygotsky. New Ideas in Psychology 18, 215–239.

Muthivhi, A.E., 2010. Piaget, vygotsky, and the cultural development of the notions of Possibility and necessity: an experimental study among rural South African learners. South African Journal of Psychology 7 (128).

Muthivhi, A., Broom, Y., 2009. School as cultural practice: Piaget and Vygotsky on learning and concept development in post-apartheid South Africa. Journal of Education (47), 55–72.

Perret-Clermont, A., 1997. Revisiting young Piaget in Neuchatel among his partners in learning. In: Smith, L., Dockrell, J., Tomlinson, P. (Eds.), Piaget, Vygotsky and beyond: Future Issues for Developmental Psychology and Education. Routledge, New York, pp. 91–120.

Piaget, J., 2001. The Psychology of Intelligence (Chapter 6: Social Factors in Intellectual development). Routledge Classics, London.

Piaget, J., 1964. Development and learning. In: Ripple, R., Rockcastle, V. (Eds.), Piaget Rediscovered. Cornell University, Ithaca, NY.

Shayer, S., 1997. Piaget and Vygotsky: a necessary marriage for effective educational intervention. In: Smith, L., Dockrell, J., Tomlinson, P. (Eds.), Piaget, Vygotsky and beyond: Future Issues for Developmental Psychology and Education. Routledge, New York, pp. 36–59.

Stetsenko, A., 2008. From relational ontology to transformative activist stance on development and learning: expanding Vygotsky's (CHAT) project. Cultural Studies in Science Education.

Valsiner, J., 1988. Developmental Psychology in the Soviet Union. Harvester Press, Brighton.

Vianna, E., Stetsenko, A., 2006. Embracing history through transforming it: contrasting Piagetian versus Vygotskian (Activity) theories of learning and development to expand constructivism within a dialectical view of history. Theory and Psychology 16 (1), 51–80.

Vygotsky, L.S., 1986. Thought and Language. M.I.T.Press, Cambridge, MA.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Weiten, W., 2001. Psychology Themes and Variations. Wadsworths, Belmont, CA.

Wood, D., 1988. Images of childhood and their reflection in teaching. In: Wood, D. (Ed.), How Children Think and Learn: The Social Contexts of Cognitive Development. Basil Blackwell, Oxford.

Pidgin and Creole Languages

Salikoko S Mufwene, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of creoles and pidgins has been marked by controversy about how they emerged, whether they can be identified by their structural features, and how they stand genetically in relation to their lexifiers. There have also been disagreements about what contact-induced varieties count as creoles, whether expanded pidgins should be lumped together with them, otherwise what distinguishes both kinds of vernaculars from each other, and how other contact-induced language varieties can be distinguished from all the above. Another important question is what they contribute to the understanding of language from the phylogenetic, typological, and sociolinguistic perspectives.

Introduction

The neutral word order *Creole and Pidgin* is used deliberately in the title of this article, because the traditional position that creoles evolved from erstwhile pidgins has been questioned, on the grounds that it is not supported by the history of colonization, which produced both kinds of language varieties (Chaudenson, 1992, 2001, 2003; Mufwene, 2001, 2005, 2008). Much earlier, Alleyne (1971) had disputed the 'baby talk' hypothesis, according to which creoles (and pidgins) had evolved from simplified varieties of European languages in the contact settings that produced them; fossils of inflections still evident particularly in Haitian Creole and even in such others as Saramaccan and Sranan (e.g., *broko* 'break', *dede* 'died, dead') speak otherwise. Below, I will also connect these issues to the role of interpreters in the earlier stages of the contacts between Europeans and non-Europeans in both trade and settlement colonies. The reader will thus be served an account more grounded in the history of colonization than traditional a-historical ones such as Bakker (2009).

The label 'language varieties,' rather than 'languages,' is also used to draw attention to the fact that in many places where these varieties coexist with their lexifiers, their speakers do not think they speak a separate language, only a different variety (Mühlhäusler, 1985; Mufwene, 1988). Lexifier – the language from which a creole has evolved, having selected most of its vocabulary from it and, according to this article, even a great deal of its grammar, often with some modification. It has not been proven that creoles and pidgins may not be considered as new nonstandard dialects of the relevant European languages (see below), on a par with others that emerged in populations that are predominantly of European descent, worse yet, that they are not genetically related to their lexifiers, *pace* Thomason and Kaufman (1988) and Thomason (2001). Mufwene (2001, 2005, 2008) and DeGraff (2003, 2005) expose social biases in the ways linguists and other scholars have singled out creoles and pidgins as outcomes of abnormal or less natural evolutions.

The term *lexifier* is used in this article, as in much of the creolistics literature, as a label of convenience for the language (variety) from which a creole or pidgin has evolved and has typically inherited most or the overwhelming part of its vocabulary. The claim that a creole has typically inherited its grammar from sources other than its lexifier (Holm, 1988; Thomason and Kaufman, 1988; Thomason, 2001) is not backed by the diachronic evidence. Although a creole or a pidgin, often (mis)identified with the basilect, has a grammar very different from the standard variety of its lexifier (called *acrolect*), it has often been shown that much of its grammar can be traced, at least partly, to some of the nonstandard varieties to which its 'creators' were exposed. This is in fact one of the strongest contributions of Sylvain (1936), ironically invoked by some creolists as a forerunner of the relexification hypothesis (see below).

Although creoles and, to some extent, expanded pidgins (see below) have received a great deal of attention since the 1990s from a typological perspective (see in particular McWhorter, 1998, 2001; and the response by DeGraff, 2001), linguists have traditionally investigated them from the point of view of their 'genesis.' Space limitations make it difficult to focus equally on both research dimensions; this article focuses more on determining what kinds of language varieties they are historically, what peculiarities appear to distinguish them from other natural languages, and what they actually teach us about the role of contact in language evolution and about the architecture of language.

As much as the term *creole* has been extrapolated to various contact language varieties, including those with a non-European lexifier (cf Kouwenberg and Singler, 2009), this article is restricted to those lexified by a European language in those colonies where a European language prevailed as the vernacular to which all nonindigenous populations shifted sooner or later. The reasons for this position are articulated below.

What Are Creoles and Pidgins?

Strictly speaking, creoles and pidgins are new language varieties that developed out of contacts between colonial nonstandard varieties of a European language and several non-European languages around the Atlantic and in the Indian and Pacific Oceans during the seventeenth to nineteenth centuries. *Pidgins* typically emerged in trade colonies that developed around trade forts, such as on the coast of West Africa, and on whaling ships, as in the South Pacific. They have reduced structures and restricted functions (typically trade and whaling activities); and initially they served as

nonnative lingua francas to users who maintained their native vernaculars in their intraethnic interactions. Some pidgins have expanded into regular vernaculars, especially in urban settings, and are called 'expanded pidgins.' Examples include Bislama and Tok Pisin (in Melanesia) and Nigerian and Cameroon Pidgin Englishes. Structurally, these are as complex as 'creoles' (Féral, 1989; Jourdan, 1991, 2009), though their evolutionary trajectories are different (see below).

Creoles emerged in settlement colonies whose primary industry consisted typically of sugar cane or rice cultivation, for which non-European slaves or contract laborers were employed who constituted the overwhelming majority of the plantation populations. Examples include Haitian, Mauritian, and Seychellois (lexified by French); Jamaican, Guyanese, and Hawaiian Creole, as well as Gullah in the United States (all lexified by English); and Saramaccan and Sranan in Surinam (lexified by English, with the former heavily influenced by Portuguese and the latter by Dutch). Creoles have also been singled out in Australia, although there are no history Creole populations there. (Interestingly, as explained below, the initial Creole populations in the relevant colonies spoke not creoles but koiné varieties or closer approximations of these. (Koinéization, traditionally ill-defined as 'dialect-leveling,' is the compromise variety that emerges by competition and selection from the contact of dialects of the same language.) Nonetheless, Creole people are mentioned because the term for the colonial vernaculars is a historical derivative of that for people.) Vernaculars such as Cape Verdian Crioulo (lexified by Portuguese) and Papiamentu in the Netherlands Antilles (apparently Portuguese-based but influenced by Spanish) suggest that the plantation industry is not as significant a factor as population growth (including rate of population replacement) and population structure (having to do with early segregation) in the identification of a colonial vernacular as a creole. These considerations help explain why Brazil, which engaged in sugar cane cultivation a century earlier than the Caribbean colonies but had a different population structure, did not produce varieties identified as creoles. To be sure, it is also disputable whether these vernaculars can be singled out as a typological class (*pace* McWhorter, 1998, 2001), let alone a genetic one (see below).

Note that although Melanesian expanded pidgins are associated with sugar cane plantations, they need not be considered as creoles. As pointed out by Keesing (1988), they apparently originated in trade and whaling settings and were adopted as lingua francas on the plantations, in what Chaudenson (1979ff) considers as 'endogenous colonies,' before they evolved into urban vernaculars and expanded their functions and structures. Given that the complexity of their grammars makes them comparable to creoles (see also Siegel, 2008), they raise the question of whether only one evolutionary trajectory need produce this kind of restructuring away from the lexifier.

According to Chaudenson (1992ff), creoles have evolved by basilectalization (structural divergence away from the lexifier, leading to the emergence of a basilect in a creole continuum), from closer approximations of their lexifiers spoken by the earliest slaves. As a matter of fact, linguists have typically considered as creoles the basilectal varieties, which evolved later in history and are structurally the most divergent from the acrolects. These should not be confused with the protovariety identified as *lexifier*, which was typically nonstandard, possibly a colonial *koiné* of the relevant European language that prevailed. Thus one may argue that, evolutionarily, creoles stand to expanded pidgins in a way that a half-empty bottle stands to a half-full bottle: different histories but similar outcomes. Jourdan (2009) provides interesting evidence of very recent complexification of Bislama grammar from less complex structures and using materials from the English lexifier under the influence of substrate languages.

It has been assumed since the nineteenth century that creoles evolved from pidgin ancestors by the acquisition of native speakers and the concurrent expansion of the pidgins' structures to meet more communicative functions. However, history shows that creoles actually emerged in settings where contacts between Europeans and native speakers of the lexifiers could not have been as sporadic as in the trade settings that produced pidgins, certainly not during the homestead phase when the non-European component of the settlement population was the minority and the populations were racially integrated though not necessarily equal socially (Chaudenson, 1992ff; Mufwene, 1997). Geographically, creoles and pidgins developed in separate places, in which fluent speakers of the local vernaculars (who need not have been Europeans in settlement colonies) and the learners had differing interaction patterns, sporadic in trade colonies but regular and apparently also intimate in the settlement colonies (Mufwene, 2005, 2008). Moreover, the term 'pidgin' was first used in print in 1807 (Baker and Mühlhäusler, 1990), much later than the term 'creole,' which was coined in Latin America for locally born people of nonindigenous stock in the late sixteenth century and used in reference to a 'corrupted' variety of Portuguese spoken in Senegal in the late seventeenth century (La Courbe's *Premier voyage* 1688: 192, cited by Arveiller, 1963). Its later, exclusionary extension to other colonial nonstandard varieties spoken primarily by descendants of non-Europeans may have been initiated by locally born European colonists, who were proud to be identified as Creole (given some rights they could invoke to administer the colonies), claimed to have maintained the European language intact, but dissociated themselves from the non-European Creoles (Stewart, 2007).

The etymology of the term 'pidgin' points indeed to 'Pidgin English,' apparently a distortion of 'business English' (though it is probably also due to partial congruence with Cantonese *bei chin* 'give money' or 'pay'; Comrie et al., 1996, p. 146), in Canton, an important trade colony where no colonial plantation industry developed and no variety has been identified as creole.

More doubt on the traditional position that creoles evolved from pidgins is cast by the role that interpreters are shown to have played, in historical accounts, in the early contacts between Europeans and non-Europeans (Bolton, 2000, 2002; Fayer, 2003; Mufwene, 2005, 2014). This is best documented about China, where the interpreters (also identified as 'linguists'), from the indigenous ruling class,

were also required power brokers in trade (Van Dyke, 2005). The evidence also comes from the history of the colonization of Hawaii (Beechert, 1985), where interpreters from the monarchy are reported to have played a role in the spread of English, and that of Africa, where interpreters facilitated its colonization (Fortbath, 1977; Reader, 1997; Lawrance et al., 2006). There is also similar evidence about the colonization of the Americas, where interpreters were used in trade with Native Americans (Karttunen, 1994; Curtin, 1984; Gray and Fiering, 2000; Metcalf, 2005).

There are thus plenty of reasons to question the traditional, a-historical view that derives creoles from pidgins, still espoused by Siegel (2008), Bakker (2009), and a number of other creolists. Although the Pacific illustrates the pidgin-to-expanded pidgin evolutionary trajectory, this is not the evolutionary trajectory of creoles around the Atlantic and in the Indian Ocean, or even in Hawaii. One must always remember that, barring some questionable characterizations of some South-to-East Asian vernaculars as creoles, there is a neat ecological complementary division between creoles and pidgins: the creoles that have informed our theorizing typically emerged in plantation settlement colonies, whereas pidgins emerged in trade colonies or on whaling ships (Keesing, 1988). In Hawaii, Creole emerged in the city but Pidgin on the plantations (Roberts, 1998), which precludes extrapolating hastily from Hawaii to the Caribbean or vice versa, *pace* Bickerton (1981, 1984).

History also suggests that even pidgins may have evolved by gradual basilectalization, not at all abruptly, as the number of people using the trade language increased and most of the new users hardly had enough exposure to the varieties spoken by the earlier interpreters. Expanded pidgins of course arose by recomplexification, under the ecological pressure of the increased and more diversified communicative needs of those who would use pidgins as vernaculars. The often-invoked jargon or prepidgin stage has not been documented. Siegel (2008) associates it with isolated individuals who do not form a community where the new language works as an auxiliary lingua franca. The claim, which confuses interlanguages with the emergence of a new language variety, is also at odds with the role played historically by interpreters in the earlier stages of trade between Europeans and non-Europeans. We will learn more about the emergence of pidgins in doing more research about the conditions under which the interpreters were formed, undoubtedly by immersion, as suggested by the history of the discovery of the rest of the world by Europeans.

Last but not least, it is also noteworthy that no French pidgin emerged on the African coast that can be related to French creoles of the Caribbean and Indian Ocean; and English pidgin in coastal Nigeria (which spread into Cameroon after the Germans lost the colony) may be a late eighteenth to nineteenth century phenomenon (Mufwene 2014). As reported by Huber (1999) and confirmed by Ostler (2005), Portuguese had then functioned as the convenient lingua franca along the trade route all the way to the Far East (and the language of diplomacy in the same world) until after the Dutch, French, and English maritime powers confiscated some of the Portuguese colonies. The slave trade on the African coast appears to have been conducted in Portuguese, spoken by the *grumetes* and the children of the Portuguese 'factors' or *lançados*, who acted both as brokers/intermediaries and as interpreters (Berlin, 1998). Nigerian Pidgin may thus have arisen for reasons that had nothing to do with the slave trade. Tay Boy, the only French pidgin that emerged in colonial French Indo-China, may be a phenomenon similar to the *le français tirailleur* in West Africa (especially Senegal), thus recent (in the twentieth century), long after French creoles had emerged, and was short-lived. The present *Abidjannais*, the French lingua franca spoken by poorly educated Ivoiriens in the southern part of Côte d'Ivoire, is also a recent twentieth century phenomenon.

The terms 'creole' and 'pidgin' have also been extended to some other varieties that developed during the same period out of contacts among primarily non-European languages. Examples of the latter denotational extension include Delaware Pidgin, Chinook Jargon, and Mobilian in North America; Sango, (Kikongo-)Kituba, and Lingala in Central Africa, Kinubi in Southern Sudan and in Uganda; and Hiri Motu in Papua New Guinea (Holm, 1989; Smith, 1995). In the original, lay people's naming practice, the term 'jargon' may have been an alternate to 'pidgin,' though, like the term *patois* in French colonies, it had a much longer, precolonial tradition of being used for a language considered 'barbarous' or just 'unintelligible' to speakers of a language considered 'superior.'

However, for reasons that remain elusive, Hall (1966) and Mühlhäusler (1986) stipulate that pidgins are more stable and jargons are an earlier stage in the 'life-cycle' that putatively progresses from Jargon, to Pidgin, to Creole, to post-Creole by progressive structural expansion, stabilization, and closer approximations of the lexifier. As noted above, history provides no particular information that supports this position, though every second-language learner goes through an interlanguage stage as they develop their competence in the target language, which naturally need not be expected to match that of native speakers nor to be fluent. Although the interlanguage may be unintelligible to native speakers, learning a second language imperfectly (a phenomenon that varies in degrees and varies from one speaker to another) is not tantamount to being stuck in an interlanguage, *pace* Plag (2009), as pointed in Mufwene (2010).

The extension of the term 'creole' is more problematic largely because of its original association with nonindigenous people born in the American settlement colonies and later applications to plants, animals, and customs considered particularly typical of the same colonies (Valkhoff, 1966). It is thus that it was extended to some of the language varieties, while European Creoles did not want this characterization to apply to their colonial dialects (Chaudenson, 1992, 2001; Mufwene, 1997). This has in fact raised issues about whether or not Afrikaans as spoken by the Afrikaners is a creole. Hesseling (1897) saw it in a different light, as not a creole (in contrast with Negerhollands, produced by African slaves in the Virgin Islands), than Valkhoff, who called it a creole.

There are also varieties spoken by descendants of non-Europeans in pockets of small endogenous Portuguese settlement colonies from South Asia (e.g., Korlai in India) to Southeast and East Asia (e.g., Papia Kristang in Malaysia and Mecanese in Macau) that are called creoles. They are

not associated with the plantation industry, not even remotely (like Cape Verdean Crioulo), and do not match the epistemic prototypical stereotype of creoles developed from those of the Atlantic Indian Oceans, but they share with the latter the fact of being a nonstandard vernacular of European origin spoken by a non-European population. Less controversial is perhaps the case of varieties that developed in Northeastern Australia, which are products of largely endogenous plantation settlement colonization (as in Queensland), although this is only a partial characterization of Australia as a settlement colony.

In the absence of any conclusive evidence that defines creoles structurally, *pace* McWhorter (1998), it is not evident that the term *creole* can be extended to just any language that is claimed to have started as a pidgin and has been 'nativized' (see below). Thus, it should not be extended to expanded pidgins such as Melanesian pidgins and Nigerian and Cameroon Pidgin Englishes, as noted above. An important lesson from all this confusion is that population movement and language contact appear to have played a catalyst role in normal language change and language speciation, even in varieties not considered as creoles or pidgins (Trudgill, 1983, 2004; Mufwene, 2001, 2003, 2005, 2008; DeGraff, 2003, 2005, 2009).

The Putative 'Creole Life-Cycle' in Historical Perspective

The myth of this evolutionary trajectory was partly formulated first by Schuchardt (1914), when he also claimed that 'Black English,' now identified as African American Vernacular English (AAVE) or Ebonics, had gone through all the evolutionary stages: pidgin > creole > post-creole. One of the clearest statements of the position is to be found in Bloomfield (1933, p. 474): "when the jargon [which seems to have meant 'pidgin' to him] has become the only language of the subject group, it is a *creolized language*." Hall (1962, 1966) espoused this position, associating the vernacular function of creoles with nativization. Since then, creoles have been defined inaccurately as 'nativized pidgins,' i.e., pidgins that have acquired native speakers and have therefore expanded both their structures and functions (owing to their usage as a vernacular) and have stabilized. In fact, Hall (1966) alternates nativization with indigenization as becoming indigenous to the territory. He also introduced the pidgin-creole 'life-cycle' to which DeCamp (1971) added a 'post-creole' stage, whereby a creole could shed off its 'creole features' (misnomer for features associated with, or typically attested in, creoles lexified by European languages (none of them is exclusive to creoles)), precisely what Schuchardt (1914) thought had happened to AAVE.

The first creolist to have disputed the above evolutionary scenario is Alleyne (1971). As noted above, he argues that fossilized inflectional morphology in Haitian Creole (HC) and the like suggests that the European colonists did not communicate with the Africans in foreigner or baby talk (see below). Supporting this position, Chaudenson (1979ff) argues that plantation communities were preceded by homesteads on which mesolectal approximations of European lexifiers, rather than pidgins, were spoken by earlier slaves. This scenario is supported by the economic history of the relevant territories, as plantations typically took a long time to develop and land-owning families often took generations before accumulating enough capital to shift from a small farm to a large plantation and to gradually buy more and more slaves to work on it. Also, like some economic historians, Berlin (1998) observes that, in North American colonies, Black Creoles spoke the lexifier fluently. It also appears from ads on runaway slaves in English North American colonies that bad or poor English is typically associated with slaves imported as adults from Africa who had arrived recently. In addition, diachronic evidence of creoles suggests that the basilects developed during the peak growth of plantations (in the eighteenth century for most colonies!), when infant mortality was high, life expectancy was short, the plantation populations increased primarily by massive importation of slave laborers, and the proportion of fluent speakers of the earlier colonial varieties kept decreasing (Baker and Corne, 1986; Chaudenson, 1992ff, Mufwene, 2001).

The life-cycle model claims that as a creole continues to coexist with its lexifier, the latter exerts pressure on it to shed some of its 'creole features.' Thus, as claimed by Schuchardt (1914), AAVE would have evolved structurally closer to North American White English than Saramaccan putatively because of increased interactions of African Americans with White Americans since the abolition of slavery. Schuchardt did not factor in the influence that Portuguese exerted on the emergence of Saramaccan's structure, the fact that the substrate populations did not consist of identical proportions of speakers of the same African languages, nor the fact that the demographic disproportions where AAVE emerged, on the cotton and tobacco plantations, were in favor of the European populations, and racial segregation within the relevant North American populations started in the late nineteenth century, as opposed to the late seventeenth or early eighteenth century on the Surinamese plantations.

Jespersen (1921) and Bloomfield (1933) anticipated DeCamp (1971), Bickerton (1973), and Rickford (1987) in invoking 'decreolization' ('loss of "creole" features') – what should more accurately be called 'debasilectalization' (loss of basilectal features; alternative to and more accurate term than *decreolization*) – to account for speech continua in creole communities. Alleyne's (1980) position that social speech continua (consisting of a basilect, a mesolect, and an acrolect) and geographical continua (owing to the fact that creoles' structures have never been identical from Surinam to the United States) is consistent with the history of colonization. (The terms were coined by William Stewart (1965) with the following meanings: *basilect* 'the variety whose structures differ the most from the local standard variety of the lexifier', *acrolect* 'the local standard variety' (often misidentified as the lexifier, which was nonstandard instead), *mesolect* 'the broad range of intermediate varieties between the former two.' There arises of course the logical question of whether such continua should be considered unique to creole-speaking territories.) The population

structures were not identical from one colony to another and not every individual learned the target colonial language from identical speakers or was equally apt at second-language learning. Even within the same colony, such as Jamaica, the situation varied according to whether one lived on a small farm (where there was no segregation), on a large plantation (which was segregated), or in the city (which was not rigidly segregated, Dunn, 1972). Around the Atlantic and in the Indian Ocean, no creole emerged in the city; it was brought there after the abolition of slavery, when many former slaves refused to continue working on the plantations, were evicted from their cabins, and migrated to poor or working class neighborhoods in the emergent cities.

Just the opposite development occurred in Hawaii, where creole evolved in the city (where the Asian populations were more mixed) rather than on the plantations (where the Asian contract laborers were ethnically segregated, lived among themselves, received instructions for work from the overseer, who was their interpreter, and hardly felt the pressure to shift to English as their vernacular). (Unlike the descendants of Africans around the Atlantic and in the Indian Ocean, Asians in Hawaii have not lost their traditional ethnic identities as Chinese (Cantonese, Hokkien, etc.), Japanese, Korean, Filipino, etc. This ecological difference says a lot about the conditions of vernacular usage, which fostered the emergence of creoles.) In the United States, Gullah, which is structurally closer to North American English varieties than Jamaican Creole, emerged on the rice fields, which were demographically smaller than large sugar cane plantations of the Caribbean colonies. On the other hand, AAVE is structurally similar to American White Southern English, with which it shares ancestry on the cotton and tobacco plantations. Bajan (Barbadian Creole) is less divergent from the Caribbean English acrolect than Jamaican Patwa, simply because the island is smaller, had more indentured servants, and the African demographic majority over the European population was not as high as in Jamaica.

DeCamp (1971) did not factor in the above historical facts when he characterized language variation in Jamaica as a 'post-creole continuum.' As a matter of fact he adduced no diachronic evidence in his invocation of 'decreolization' to account for the continuum. He just documented the variation from the basilect to the acrolect so adequately that he could posit an informative 'implicational scale' in the way the variants can be used. However, this need not be correlated with how the language had evolved ultimately from its lexifier.

On the other hand, there is no particular reason why the structures of a creole should be expected to be monolithic. If a variety is characterized as creole just because of the particular sociohistorical ecology of its development (see below), rather than because of its structural peculiarities, it cannot stop being a creole even after some of the features have changed. Besides, basilectal and mesolectal features continue to coexist in these communities, suggesting that Creole has not died yet. Lalla and D'Costa (1990) present copious data against debasilectalization in Caribbean English creoles (CEC), just as Mufwene (1994) adduces linguistic and nonlinguistic arguments against the same process in Gullah. Although Rickford and Handler (1994) show that Barbados may have had a basilect comparable to that Jamaican Creole in the late eighteenth century, no evidence of such debasilectalization (different from the change without loss of the basilect documented by Rickford, 1987 regarding Guyanese Creole) has been documented about any other place. How the basilect may be claimed to have been lost in Barbados but not elsewhere in the Caribbean calls for an explanation. Winford (1997) offers an account of why Creole emerged not only later but not identically in Trinidad and Guyana. Only the latter has a basilect comparable to that of the Jamaica.

Are Creoles Separate Languages from Their Lexifiers?

Another contentious issue about creoles is the common stipulation by linguists that creoles are separate languages from their lexifiers and related excolonial varieties spoken by descendants of Europeans. Thus, the nonstandard French varieties spoken in Quebec and Louisiana, as well as on the Caribbean islands of St. Barths and St. Thomas, are considered dialects of French rather than creoles. Likewise New World nonstandard varieties of Spanish and Portuguese are not considered creoles (with the exception of Palenquero, spoken by a population of primarily African descent!), despite structural similarities that they exhibit with creoles of the same lexifiers, such as São Tomense, Principense (both spoken in the Bight of Biafra), Cape Verdiano Crioulo for Portuguese.

Interestingly, contact-based varieties spoken by descendants of Europeans or populations in which people of European descent are the majority are not considered as creoles. AAVE is not because, as noted above, it shares its origins with American (White) Southern English. Gullah, spoken by a majority-Black population has been stipulated to be a creole (its speakers think Creole is spoken only in Louisiana, which as has also produced 'creole cuisine'!), but Amish English is not, although it is nonstandard and spoken by descendants of people that are primarily of non-English, German-speaking Swiss descent. Ignoring Hjelmslev's (1938) and Posner's (1985) position that creoles are new dialects of European languages, creolists have adopted uncritically this socially based naming tradition in former European settlement colonies, identifying as creoles those varieties of European languages that have been appropriated as vernaculars by non-European majorities.

There is yet no yardstick for measuring structural divergence from the lexifier, nor was the latter the same in every contact setting. Contact of dialects and/or languages was indeed a factor in all colonial settings, especially in places such as the United States and Australia, where descendants of the English constitute a small proportion of the people of European descent. Granted that the varieties spoken by these colonial populations are not as divergent as those labeled creoles in the same nations, what are the particular factors within or other than contact that justify the distinction in ways different from those that account for variation among creoles? Could the fact that majority Europeans of non-English descent shifted to English after its formative phases

in the colonies explain why the influence of the new speakers has remained limited (Mufwene, 2009a)?

Do Creoles Form a Typological or Genetic Class of Their Own?

It has also been claimed that creoles have more or less the same structural design (Bickerton, 1981, 1984; Markey, 1982). This position is as disputable as the counterclaim that they are more similar in the sociohistorical ecologies of their developments (Mufwene, 1986a) or even the more recent claim that there are creole prototypes from which others deviate in various ways (Thomason, 1997; McWhorter, 1998). (According to John H. McWhorter, a creole that exhibits the most "creole features," especially one that also has no tones and no derivational or inflectional affixes. This is different from the notion of "epistemic prototype" used in this article for the creoles (of the Atlantic and Indian Ocean, all lexified by Western European languages) that linguists first investigated and have informed the scholarship on creoles and pidgins. The very fact of resorting to a handful of prototypes for the general creole structural category suggests that the vast majority of them do not share the putative set of defining features, hence that the features cannot be used to single them out as a unique type of language. (Typological classifications are typically of the classical-category kind, with no exceptions, rather than of the prototype-category type, relying on family resemblance with a tiny core of best exemplars!) It is tantamount to saying that some ergative or SOV languages belong less in the lot than others after claiming that they all fit in the same category. As underscored by Meyerhoff (2009), many of the same features and combinations thereof are attested in languages that are not considered as creoles.

On the other hand, variation in the structural features of creoles (lexified by the same language) is correlated with variation in the linguistic and sociohistorical ecologies of their developments (Mufwene, 1997, 2001). The notion of 'ecology' includes, among other things, the nature of the lexifier, structural features of the substrate languages, changes in the ethnolinguistic makeups of the populations that came in contact, the kinds of interactions between speakers of the lexifier and those of other languages, and rates and modes of population growth.

It has also been claimed that creoles are structurally simpler than other languages. In the lead of this claim are McWhorter (1998, 2001), which have generated numerous responses in support or against the position. Leading the latter are DeGraff (2001, 2005, 2009), which also raise the question of how variably complexity is conceived of in the first place. This raises also the question of whether it is even possible to address the question fruitfully without presupposing a bias toward what the architecture of language is expected to be like (Meyerhoff, 2009; Mufwene, 2009b). Generally lacking from claims of the reduction of complexity in creoles are considerations of the interpretation principles that underlie the putatively simpler structures or the interactions between rules and between modules. It is also noteworthy that creoles lexified by European languages lack tones (one of the criteria stipulated by McWhorter, 1998), because their lexifiers are nontonal. It is not obvious that they lack derivations, as made obvious by DeGraff (2001). Regarding inflections, they seem to have followed the tendency observable in the evolution of the same European languages, as is made obvious in Chaudenson (2001, 2003).

To date the best-known creoles have been lexified by English, French, Portuguese, and Dutch. Those of the Atlantic and Indian Ocean are, along with Hawaiian Creole, those that have informed most theorizing on the development of creoles, which I identify as the epistemic prototypes. The reader should remember that the colonization of Hawaii started after the abolition of the slave trade and the population structure on Hawaiian plantations was unlike that instituted around the Atlantic and in the Indian Ocean, where pressure to operate in the European colonial language was experienced since the homestead phase of colonization. Whoever wishes to generalize over the formation of creoles should beware of extrapolating too hastily from one part of the world to the other (Mufwene, 2005, 2008).

While the terms 'creole' and 'creolization' have been applied often uncritically to various contact-induced language varieties, several distinctions, which are not clearly articulated have also been proposed, for instance, between pidgin, creole, koiné, semicreole, intertwined varieties, foreign workers' varieties of European languages (e.g., Gastarbeiter Deutsch), and, we may add, indigenized varieties of European languages (e.g., Nigerian and Singaporean Englishes). The denotations and importance of these terms deserve reexamining (Holm, 1988, 2004; Arends et al., 1995; Mufwene, 1997, 2005; Bakker, 2009).

Confusing expanded pidgins such as Tok Pisin with creoles (Thomason, 2001; cf Siegel, 2008) depends conceptually on whether one subscribes to the position that creoles are nativized pidgins. Discussions that lump them together are informative in showing the extent to which different evolutionary trajectories can nonetheless produce similar structural outcomes. On the other hand, the question arises of whether vernacularization (not to be confused with urbanization, *pace* Bakker, 2009), rather than nativization, is not a more critical factor driving the structural expansion of pidgins into expanded pidgins and keeping creoles from simplifying their morphosyntax to the same extent as pidgins.

Equally problematic is the unrestrained extension of the term 'creole' to a number of 'contact languages' associated indeed with (European) colonization but have not been lexified by European languages or associated with Creole people, based on the disputable assumption that they started as pidgins and have been nativized. Examples include Kikongo-Kituba, Lingala, Sango, and Kinubi in Africa, as well as Sri Lanka Malay, Bazaar Malay, and Ambonese Malay in Asia.

Creole 'Genesis'

The central question here is: how did creoles emerge? The following hypotheses are the major ones competing

today: the substrate, the superstrate, and the universalist hypotheses.

Substratist positions are historically related to the 'baby talk hypothesis,' which can be traced back to late nineteenth-century French creolists: Bertrand-Bocandé (1849), Baissac (1880), Adam (1883), and Vinson (1882). According to them, the languages previously spoken by the Africans enslaved on New World and Indian Ocean plantations were the primary reason why the European lexifiers, which they appropriated, were restructured into creoles. These early creolists assumed African languages to be 'primitive,' 'instinctive,' in 'natural' state, and simpler than the 'cultivated' European languages with which they came in contact. Creoles' systems were considered to be reflections of those of non-European languages, which allegedly reflected the mental inferiority of those who produced and spoke them. The African slaves were thus considered as incapable of learning the putatively more evolved structures of the European languages. The 'baby-talk' connection is that, in order to be understood, the Europeans supposedly had to speak to the Africans like to babies. More or less the same idea is to be found in the 'foreigner talk' hypothesis, according to which Europeans reproduced the non-Europeans' inaccurate approximations of their languages.

The revival of the substrate hypothesis (without its racist component) has been attributed to Sylvain (1936). Although she recognizes significant influence from French dialects, she concluded her book, surprisingly, with the statement that Haitian Creole is Ewe spoken with a French vocabulary. Over two decades later, Turner (1949), disputing American dialectologists' claim that there was virtually no trace of African languages in 'Black English,' highlighted some morphosyntactic similarities between 'the Gullah dialect' and some West-African (especially Kwa) languages. He then concluded that "Gullah is indebted to African sources" (p. 254), which stimulated more research on African substrate influence on African-American English (e.g., Dillard, 1972) and on CECs (e.g., Alleyne, 1980).

Mufwene (1990, 2010) identifies three main schools of the substrate hypothesis today. The first, led by Alleyne (1980, 1996) and Holm (1988), is closer to Turner's approach and is marked by what is also its main weakness: invocation of influence from diverse African languages without explaining what kinds of selection principles account for this seemingly random invocation of sources. This criticism is not *ipso facto* an invalidation of substrate influence. It is both a call for a more principled account, one that articulates the particular ecological factors that appear to have favored various individual influences (thus legitimating what was dubbed the 'Cafeteria Principle'), and a reminder that the nature of such influence must be reassessed.

The second school, identified by its practitioners as the relexification hypothesis (RH), is fully articulated by Lefebvre (1998), who argues that HC consists largely of French lexical entries spoken with the grammar of languages of the Ewe-Fon (or Fongbe) group. Extended to other creoles, the position, has been repeated in some of the contributions to Lefebvre et al. (2006), though some others (see especially Aboh, 2006; Siegel, 2006) dispute it, and almost all of those to Lefebvre (2011) are more cautious. Objections to this phylogenetic hypothesis include the following: (1) RH's 'comparative' approach has not taken into account several features that HC (also) shares with nonstandard French; (2) RH downplays features that HC shares also with several other African languages, which were represented in Haiti during the critical stages of its development, thus it is not obvious why the exclusive focus on the Ewe-Fon languages; (3) studies of naturalistic second language acquisition provide no evidence in support of RH, even if the emergence of creoles could at all be associated exclusively with adult L2-learners (Chaudenson, 2001, 2003); and (4) RH does not account for those cases where HC has selected structural options, which are not consistent with those of Ewe-Fon. Moreover, relexificationists assume, disputably, that languages of the Ewe-Fon group are structurally identical in all respects and that no competition of influence was involved among them. (The competition here refers to unequal ranking by speakers of variants that serve more or less the same function in a system (in a particular ecology or setting of interactions). Selection is the process that resolves the competition, in favoring a particular variant.) The most elaborate critique of RH is DeGraff (2002), which is complemented by various refined analyses of hybridized structures in Haitian by Aboh (2006, 2009, 2015, among many others). For contrary evidence from other creole-like languages (in the sense of this article), see especially Siegel (2006).

The least disputed version of the substrate hypothesis is Keesing's (1988), which shows that substrate languages may impose their structural features almost intact on the new, contact-induced varieties if they are typologically homogeneous, with most of them sharing the relevant features. Thus Melanesian pidgins are like (most of) their substrates in having DUAL/PLURAL and INCLUSIVE/EXCLUSIVE distinctions and in having a transitive marker on the verb. Sankoff and Brown (1976) had shown similar influence with the bracketing of relative clauses with *ia*, as does Sankoff (1993) with the focus marker in Melanesian pidgins and Jourdan (2009) with prepositional-verb markers in Solomon Island Pidgin. Keesing argued that even the subject-object-verb order of Melanesian pidgins' sentences reflects the order of the argument affixes in the verb complex, which is often used elliptically, although the dominant order of free arguments is verb-subject-object in the relevant substrate languages.

This evolution is facilitated by what Singler (1988) calls 'homogeneity of the substrate' (argued for earlier in Mufwene, 1986b), if it is not downright its consequence. However, the pidgins have not inherited all the peculiarities of Melanesian languages. For instance, they do not have their VSO major constituent order, nor do they have much of a numeral classifying system in the combination of *pela* with quantifiers. For an extensive discussion of substrate influence in Atlantic and Indian Ocean creoles, see Muysken and Smith (1986) and Mufwene (1993). For similar discussions about creoles and the like in the Pacific, see Lefebvre et al. (2006) and Lefebvre (2011).

Competing with the above genetic views has been the dialectologist, or superstrate, hypothesis, according to which the primary, if not the exclusive, sources of creoles' structural features are nonstandard varieties of their lexifiers. Regarding African American English, Krapp (1924) and Kurath (1928), for example, claimed that this variety was an archaic retention of the nonstandard speech of low-class European colonists with whom the African slaves had been in contact. According to these dialectologists, African substrate influence was limited to some isolated lexical items such as *goober* 'peanut', *gumbo*, and *okra*. Although substrate influence need not be dismissed offhand (see below), the position is worth considering critically, especially if one factors in the presence of nonnative speakers of English among the European indentured servants with whom the slaves interacted regularly, especially during the early stages of colonization. This is even more significant in the case of Surinam, where the Dutch colonists adopted English as the language of communication with their slaves.

It would take until McDavid (1950) and McDavid and McDavid (1951) before dialectologists made allowance for some African grammatical contributions to AAE. Otherwise, D'Eloia (1973) and Schneider (1989) invoke several dialectal English models to rebut Dillard's (1972) thesis that AAVE developed from an erstwhile West-African Pidgin English (WAPE) brought over by slaves. Had they known then the language of the slave trade was probably Portuguese, they could have pointed out that WAPE varieties probably emerged later than AAE in the first place. The scant evidence presented by Dillard dates from the early eighteenth century and is closer to nonstandard English than to Pidgin English. Hancock's (1986) 'Guinea Coast Creole English,' spoken primarily in the mixed households of English factors and indigenous women need not have been as widespread as he claims and certainly not a pidgin, since the interactions in these settings were not sporadic.

Since the late 1980s, Shana Poplack and her associates have shown that AAVE shares many features with white nonstandard vernaculars in North America and England, thus it has not developed from an erstwhile creole (Poplack and Tagliamonte, 2001; Poplack, 2000). Because some of the same features are also attested in creoles (Rickford, 1998), we come back to the question of whether many, if not most, features of creoles did not originate in their lexifiers in the first place. The other question is also whether African substrate influence on AAVE must of necessity have come through creoles. History suggests that Gullah and AAVE evolved in geographic complementary distribution and apparently concurrently. During the eighteenth century, when the plantation economy was the most prosperous in English North America, only 15% of the slaves were imported from the Caribbean (Rawley, 1991). The oldest CECs would be just emerging then.

Regarding French creoles, the dialectologist position was first defended by Faine (1937), according to whom HC was essentially Norman French. This position was espoused later by Hall (1958), who argues that "the 'basic' relationship of Creole is with seventeenth-century French, with heavy carry-overs or survivals of African linguistic structure (on a more superficial structural level) from the previous language(s) of the earliest speakers of Negro Pidgin French [which can be situated nowhere in colonial history!]; its 'lexical' relationship is with nineteenth- and twentieth-century French" (1958, p. 372). Chaudenson (1989ff) is more accommodating to substrate influence as a factor that accounts for the more extensive structural divergence of creoles from their lexifiers compared to their noncreole colonial kin.

Chaudenson's allowance for substrate influence is fleshed out especially by Corne (1999), who articulates the most explicitly how feature selection can be driven by congruence, even if only partial, between the languages in contact. Although, unlike Pacific pidgins, the Atlantic and Indian Ocean French creoles did not typically emerge in settings that satisfied Singler's (1988) 'homogeneity of the substrate' condition, partial structural congruence between their substrates and nonstandard French favored the selection of the particular features they have, for instance, in the domain of time reference. Aboh (2006ff) has carried this approach farther with detail analyses that show how structural traits can be hybridized in ways similar to biological gene recombination, more specifically with serial verb constructions and number delimitation. One must then determine whether such substrate influence, which does not boil down to mere introduction of features from substrate languages (identified by Allsopp, 1977 as 'apports'), was facilitated by the numerical proportion of speakers of the relevant languages and/or by the time of the insertion of these in the linguistic feature pool. See Singler (1996, 2009) for such considerations regarding HC.

The 'universalist hypotheses,' which stood as strong contenders to substrate hypotheses in the 1980s and 1990s, have forerunners in the nineteenth century. For instance, Adolfo Coelho (1880–1886) partly anticipated Bickerton's (1981ff) 'language bioprogram hypothesis' in stating that creoles "owe their origin to the operation of psychological or physiological laws that are everywhere the same, and not to the influence of the former languages of the people among whom these dialects are found." Bickerton pushed things further in claiming that children, not adults, made creoles by fixing the parameters of these new language varieties to their unmarked, or default, settings as specified in Universal Grammar. To account for cross-creole structural differences, Bickerton (1984, p. 176–177) invokes a 'pidginization index' (PI) that includes the following factors: the proportion of the native to nonnative speakers during the initial stages of colonization, the duration of the early stage, the rate of increase of the slave population after that initial stage, the kind of social contacts between the native speakers of the lexifier and the learners, and whether or not the contact between the two groups continued after the formation of the new language variety. These factors, which one may recognize in Mufwene's ecological (1996ff) approach, were simply not anchored in the actual history of the colonization of the different creole-speaking territories.

Some nagging questions with Bickerton's position include the following: Do structures really support the claim that they were produced primarily by children (cf DeGraff, 1999a; Roberts, 1998)? Is his intuitively sound PI consistent with his creolization as abrupt pidgin-nativization

hypothesis (Mufwene, 1999)? Is the abrupt 'creolization' cum basilectalization hypothesis consistent with the social histories of the territories where classic creoles developed? How can we explain similarities between abrupt creoles and expanded pidgins when the stabilization and structural expansion of the latter is not necessarily associated with restructuring by children (Meyerhoff, 2009), *pace* some claims to be found in references such as Holm (1988) and Thomason (2001)? Is there convincing evidence for assuming that adult speech is less controlled by Universal Grammar than child language is? If so, how can we then account for similarities between abrupt 'creolization' and naturalistic second-language acquisition?

Not all creolists who have invoked universalist explanations have made children critical to the emergence of creoles. For instance, Sankoff (1979) and Mühlhäusler (1981) make allowance for Universal Grammar to operate in adults, too.

Few creolists nowadays subscribe to one exclusive genetic account, as evidenced by the contributions to Mufwene (1993) and implicitly those to Lefebvre (2011). The 'complementary hypothesis' (Baker and Corne 1986; Corne (1999), DeGraff (2009), Hancock 1986; and Mufwene 1986b, 2001) seems to be an adequate alternative, provided we can articulate the ecological conditions under which the competing influences (between the substrate and lexifier languages and within each group) may converge or prevail upon each other. This position was well anticipated by Schuchardt (1909, 1914) in his accounts of the genesis of the Mediterranean Lingua Franca and of Saramaccan. More and more research is now underway uncovering the sociohistorical conditions under which different creoles have developed, for instance, Arends (1989ff), Baker (1982ff), Chaudenson (1979ff), Corne (1999), Mufwene (2001ff), and Arends (1995). Aboh's (2006ff) hypothesis of hybridization of features even within properties of lexical items is an improvement not only over Corne's (1999) congruence model but also on Mufwene's (2001ff) ecology-specific restructuring.

In connection with the above discussion, note that the traditional claim that creoles emerged within one generation, thus abruptly, from a pidgin ancestor, has increasingly been questioned by, e.g., Chaudenson (1979ff), Arends (1986ff), Singler (1996ff), and Mufwene (1996ff). Baker (1995) provides part of the evidence in the context of pidgins, pointing out that their features did not all emerge at the same time. The strongest evidence lies in the fact that the oldest documentary evidence shows more similarity to European nonstandard varieties than the later or present-day materials. The evolutionary scenario is also consistent with the gradual way in which plantation colonies evolved, having started from homestead settings settled by small integrated groups in which the slaves were in the minority. Shortage of money also made it difficult to import more slaves and the colonial populations then grew more by birth than by importation (Mufwene, 2001ff).

Still, the future of research on the development of creoles has problems to overcome. To date, knowledge of the nonstandard lexifiers spoken by the European colonists remains limited, though more research is now underway and much of the scholarship on the dialectology of the European lexifiers is becoming handy. There are few comprehensive and integrated descriptions of creoles' structures, especially from a diachronic perspective, which makes it difficult to determine globally how the competing influences interacted among themselves and how the features selected from diverse sources became integrated into new systems. Few structural facts have been correlated with the conclusions suggested by the sociohistorical backgrounds of individual creoles, especially in studies that make strong claims for the typological singularity of creoles. Other issues remain up in the air, for instance, regarding the markedness model that is the most adequate to account for the selection of features into creoles' systems. Can there really be an ecology-independent, universal scale of markedness that can account for the selection of particular features into the structures of particular creoles or pidgins? For developmental issues on PCs, the following edited collections are highly recommended: Hymes (1971), Valdman (1977), Hill (1979), Muysken and Smith (1986), Mufwene (1993), Arends et al. (1995), DeGraff (1999b), Lefebvre et al. (2006), Ansaldo et al. (2007), Kouwenberg and Singler (2009), Lefebvre (2011), and several edited volumes of the Creole Language Library (Benjamins). One should of course also check *Amsterdam Creole Studies*, the *Journal of Pidgin and Creole Languages*, and *Etudes Créoles*. Several issues of *Pacific Linguistics* and *Te Reo* also include publications on Melanesian pidgins.

Creolistics and General Linguistics

There is much more literature on the emergence, sociology, and morphosyntax of PCs than on their phonologies, semantics, and pragmatics. Faraclas and Klein (2009) are exceptional in devoting almost half the volume to phonetics/phonology. More unusual is Bhatt and Plag (2007), devoted entirely to prosodic features of creoles and other 'contact languages.' The volume of publications is also now increasing on ethnographic and other social aspects of creoles and pidgins, with a special section devoted to such topics in Kouwenberg and Singler (2009).

With the exception of time reference (e.g., Singler, 1990; Michaelis, 1993; Schlupp, 1997) and noun phrase structure and nominal number (for references, see Tagliamonte and Poplack, 1993; Baptista and Gueron, 2007), studies in semantics and pragmatics are in the minority. On the other hand, the development of quantitative sociolinguistics owes a great deal to research on AAE since the mid-1960s (see, e.g., Labov, 1972; Rickford, 1999) and on CECs (e.g., Rickford, 1987). Numerous publications in *American Speech*, *Language in Society*, and *Language Variation and Change* reflect this. There are also several monographs on creolistics today that provide an overview of the scholarship on creoles and pidgins today. They include Romaine (1988), Holm (1988), Manessy (1994), Arends et al. (1995), Mühlhäusler (1986), Sebba (1997), Corne (1999), Chaudenson (2001, 2003), and Siegel (2008). They vary in geographical areas of focus, adequacy, and kinds of lexifiers.

Studies of the morphosyntax of creoles have yet to inform general linguistics beyond the subject matters of time reference, serial verb constructions, and grammaticization. For

instance, studies of lectal continua (e.g., Escure, 1997) have had this potential, but little has been done by creolists to show how their findings may apply to other languages. The nonmonolithic nature of 'mesolectal structures' as the intermediate sociological zone where acrolectal and basilectal features naturally mix into one system allowing more variation should have informed general linguistics against the fallacy of monolithic grammatical systems (Mufwene, 1992). (This is a different position from Labov's (1998) 'coexistent systems' hypothesis, which suggests that each of the systems may still be monolithic.) More work could have been done to eradicate the misconception that continua are typical of creole-speaking communities only; these phenomena can be observed in any population where the standard/nonstandard distinction applies and colloquial speech is accepted as normal. Based on observations such as Rickford (1990) that mesolectal speech is the norm in creole-speaking communities, more work could have been done about whether the notions BASILECT and ACROLECT are not in fact elusive constructs of convenience that are associated with analyses intended to amplify structural differences between creoles and the European languages they have evolved and diverged from.

Likewise, the debate on 'creole genesis' could have informed genetic/historical linguistics on the relevance of varying external-ecology conditions to the actuation question , to language change and speciation, and even to language vitality (Mufwene, 2001, 2005, 2008). (The actuation question is that of identifying the ecological motivation of change, i.e., the particular factors outside language that triggered a particular change at a particular point in time, owing to particular interactional dynamics.)

Note also that a concomitant of the emergence of creoles around the Atlantic and in the Indian Oceans was the loss of heritage languages among the (descendants of) the enslaved Africans. The fact that language loss as a concomitant of language shift was experienced only later by European colonists who had different heritage languages than the dominant one may very well account for why the divergence of other colonial offspring of the same European languages was not as extensive among descendants of most Europeans. It appears that they were acculturated relatively late and did not participate as early as the slaves and European indentured servants in shaping the new colonial vernaculars (Mufwene, 2009a). Their gradual assimilation to the population that had been speaking the dominant European language as a heritage vernacular is another important factor that filtered out much of the adstrate element that had once justified speaking of, for instance, German and Italian Englishes in the United States.

Although lack of consensus among creolists may be invoked as a general reason for this failure to influence general linguistics, alarming indifference from theoretical linguists, especially those engaged in theories of typology and universals, is a more important reason. Granted, the recent publication of the 'Atlas of Pidgins and Creoles' (Michaelis et al., 2013) can make some difference; the question remains of whether the typological values assigned to characteristics of some creoles are inspired from within (which would be more promising) or from without (which is no more than intellectual colonization).

Consensus cannot be expected of creolistics any more than it can of other subfields of linguistics or any other scientific discipline. Nonetheless, in the broader context of language contact (including second-language acquisition), studies of especially the emergence of creoles have been inspiring. For instance, Thomason and Kaufman (1988) is widely cited in studies of indigenized Englishes. Schumann (1978) and Andersen (1983) were an important step in consolidating common interests between second-language acquisition and the emergence of creoles. The journal *Studies in Second Language Acquisition* has sometimes featured lead articles on creoles and SLA. Lefebvre et al. (2006) is indicative of the kind of mutual stimulation that can obtain between creolistics and the scholarship on SLA (Mufwene, 2010). More cross-fertilization might be expected between studies of creole genesis and those of (child) language development (DeGraff, 1999b), as among diverse subfields of linguistics.

No less significant is the contribution that creolists have made to the scholarship on grammaticization, owing largely to the reception that the leading practitioner, Bernd Heine, himself one of the first students of language contact in Africa, has extended to such research. The first milestones are to be found in Bruyn (1995) and Baker and Syea (1996).

There is also hope in the recent growing number of publications addressing the question of whether creoles are simpler than other natural languages, including Dahl (2004), Gil (2007), Faraclas and Klein (2009), DeGraff (2001, 2009), Aboh and Smith (2009), McWhorter (2001, 2012), and Mufwene (2009b, 2013). Creole studies are definitely moving away from the periphery of the linguistics profession. They have also claimed center stage in research on the phylogenetic emergence of language (Bickerton, 1990, 1995, 2010; Botha, 2006; Mufwene, 2008, 2010; Hurford, 2012), competing with the scholarship on gestures (McNeill, 2012) and sign language. Perhaps the time has finally come when creoles can influence the way linguists and other scholars study language more conspicuously.

See also: Dialectology; Language Contact; Sociolinguistics.

Bibliography

Aboh, E.O., 2006. The role of the syntax-semantics interface in language transfer. In: Lefebvre, C., White, L., Jourdan, C. (Eds.), L2 Acquisition and Creole Genesis: Dialogues. Benjamins, Amsterdam, pp. 221–252.

Aboh, E.O., 2009. In: Aboh, E., Smith, N. (Eds.), Competition and Selection: That's all!, 317–344.

Aboh, E., Smith, N. (Eds.), 2009. Complex Processes in New Languages. Benjamins, Amsterdam.

Aboh, Enoch O., 2015. The Emergence of Hybrid Grammars: Language Contact and Change. Cambridge University Press, Cambidge, UK.

Adam, L., 1883. Les idiomes négro-aryens et malayo-aryens: essai d'hybridologie linguistique. Maisonneuve, Paris.

Alleyne, M.C., 1971. Acculturation and the cultural matrix of creolisation. In: Hymes, D. (Ed.), Pidginization and Creolization of Languages. Cambridge University Press, Cambridge, pp. 169–186.

Alleyne, M.C., 1980. Comparative Afro-American: An Historical-Comparative Study of English-Based Afro-American Dialects of the New World. Karoma, Ann Arbor.

Alleyne, M.C., 1996. Syntaxe Historique Créole. Karthala, Paris.

Allsopp, R., 1977. Africanisms in the idioms of Caribbean English. In: Kotey, P.F., Der-Houssikian, H. (Eds.), Language and Linguistics Problems in Africa. Hornbeam Press, Inc., Columbia, pp. 429–441.

Andersen, R. (Ed.), 1983. Pidginization and Creolization as Language Acquisition. Newbury House Publishers, Inc., Rowley, MA.
Ansaldo, U., Matthews, S., Lim, L. (Eds.), 2007. Deconstructing Creole. Benjamins, Amsterdam.
Arends, J., 1986. Genesis and development of the equative copula in Sranan. In: Muysken, P., Smith, N. (Eds.), Substrata versus Universals in Creole Genesis. John Benjamins, Amsterdam, pp. 57–70.
Arends, J., 1989. Syntactic Developments in Sranan: Creolization as a Gradual Process. Doctoral thesis. University of Nijmegen.
Arends, J. (Ed.), 1995. The Early Stages of Creolization. Benjamins, Amsterdam.
Arends, J., Muysken, P., Smith, N. (Eds.), 1995. Pidgins and Creoles: An Introduction. Benjamins, Amsterdam.
Arveiller, R., 1963. Contribution à l'étude des termes de voyage en français (1505–1722). D'Artrey, Paris.
Baissac, C., 1880. Etude sur le patois créole mauricien. Imprimerie Berger-Levrault, Nancy.
Baker, P., 1982. The Contribution of Non-Francophone Immigrants to the Lexicon of Mauritian Creole. Ph.D. thesis. School of Oriental and African Studies, University of London.
Baker, P., 1995. Some developmental inferences from the historical studies of pidgins and creoles. In: Arends, J. (Ed.), The Early Stages of Creolization [Creole Language Library 13]. John Benjamins, Amsterdam, pp. 1–24.
Baker, P., Corne, C., 1986. In: Muysken, P., Smith, N. (Eds.), Universals, Substrata and the Indian Ocean Creoles, pp. 163–183.
Baker, P., Mühlhäusler, P., 1990. From business to pidgin. Journal of Asian Pacific Communication 1, 87–115.
Baker, P., Syea, A. (Eds.), 1996. Changing Meanings, Changing Functions: Papers Relating to Grammaticalization in Language Contact. University of Westminster Press, London.
Bakker, P., 2008. Pidgins versus Creoles and pidgincreoles. In: Kouwenberg, S., Singler, J.V. (Eds.), Handbook of Pidgin and Creole Studies. Blackwell, Oxford, pp. 130–157.
Baptista, M., Gueron, J. (Eds.), 2007. Noun Phrases in Creole Languages: A Multi-faceted Approach. Benjamins, Amsterdam.
Beechert, E.D., 1985. Working in Hawaii: A Labor History. University of Hawaii Press, Honolulu.
Berlin, I., 1998. Many Thousands Gone: The First Two Centuries of Slavery in North America. Harvard University Press, Cambridge, MA.
Bertrand-Bocandé, E., 1849. Notes sur la Guinée portugaise ou Sénégambie méridionale. Bulletin de la Société de Géographie 12, 57–93.
Bhatt, P., Plag, I. (Eds.), 2006. Stress, tone and intonation in creoles and contact languages. Special issue of Sprachtypologie und Universalienforschung/Language Typology and Universals, 59 (2), 131–218.
Bickerton, D., 1973. The nature of a creole continuum. Language 49, 640–669.
Bickerton, D., 1981. Roots of Language. Karoma, Ann Arbor.
Bickerton, D., 1984. The language bioprogram hypothesis. Behavioral and Brain Sciences 7, 173–221.
Bickerton, D., 1990. Language and Species. University of Chicago Press, Chicago.
Bickerton, D., 1995. Language and Human Behavior. University of Washington Press, Seattle.
Bickerton, D., 1999. How to acquire language without positive evidence: what acquisitionists can learn from creoles. In: DeGraff, M. (Ed.), Language Creation and Language Change: Creolization, Diachrony, and Development. MIT Press, Cambridge, MA, pp. 49–74.
Bickerton, D., 2010. Adam's Tongue: How Humans Made Language, How Language Made Humans. Hill and Wang, New York.
Bloomfield, L., 1933. Language. Holt, Rinehart and Winston, New York.
Bolton, K., 2000. Language and hybridization: Pidgin tales from the China coast. Interventions 5, 35–52.
Bolton, K., 2002. Chinese Englishes: from Canton jargon to global English. World Englishes 21, 181–199.
Botha, R., 2006. Pidgin languages as a putative window on language evolution. Language & Communication 26, 1–14.
Bruyn, A., 1995. Grammaticalization in Creoles: The Development of Determiners and Relative Clauses in Sranan. Ph.D. dissertation. University of Amsterdam.
Chaudenson, R., 1979. Les Créoles Français. Fernand Nathan, Paris.
Chaudenson, R., 1989. Créoles et enseignement du français. L'Harmattan, Paris.
Chaudenson, R., 1992. Des îles, des hommes, des langues: essais sur la créolisation linguistique et culturelle. L'Harmattan, Paris.
Chaudenson, R., 2001. Creolization of Language and Culture. Routledge, London.
Chaudenson, R., 2003. La Créolisation: Théorie, Applications, Implications. L'Harmattan, Paris.
Coelho, F.A., 1880–1886. Os dialectos românicos ou neolatinos na Africa, Asia, ae America. Bolletim da Sociedade de Geografia de Lisboa 2, 129–196 (1880–1881); 3, 451–478 (1882), 6, 705–755 (1886).
Comrie, B., Matthews, S., Polinsky, M. (Eds.), 1996. The Atlas of Languages: The Origin and Development of Languages throughout the World. Facts on File, Inc., New York.
Corne, C., 1999. From French to Creole: The Development of New Vernaculars in the French Colonial World. University of Westminster Press, London.
Curtin, P.D., 1984. Curtin, Cross-Cultural Trade in World History. Cambridge University Press, Cambridge.
Dahl, Ö., 2004. The Growth and Maintenance of Linguistic Complexity. Benjamins, Amsterdam.
DeCamp, D., 1971. Toward a generative analysis of a post-creole speech continuum. In: Hymes, D. (Ed.), Pidginization and Creolization of Languages. CUP, Cambridge, pp. 349–370.
DeGraff, M., 1999a. In: DeGraff, M. (Ed.), Creolization, Language Change, and Language Acquisition: A Prolegomenon, pp. 1–46.
DeGraff, M. (Ed.), 1999b. Language Creation and Language Change: Creolization, Dia-Chrony, and Development. MIT Press, Cambridge, MA.
DeGraff, M., 2001. On the origin of creoles: a Cartesian critique of neo-Darwinian linguistics. Linguistic Typology 5, 213–310.
DeGraff, M., 2002. Relexification: a reevaluation. Anthropological Linguistics 44, 321–414.
DeGraff, M., 2003. Against creole exceptionalism: discussion note. Language 79, 391–410.
DeGraff, M., 2005. Linguists' most dangerous myth: the fallacy of creole exceptionalism. Language in Society 34, 533–591.
DeGraff, M., 2009. Language acquisition in creolization and, thus, language change: some Cartesian-uniformitarian boundary conditions. Language and Linguistics Compass 3/4, 888–971.
D'Eloia, S.G., 1973. Issues in the analysis of Negro nonstandard English. Review of Dillard (1972). Journal of English Linguistics 7, 87–106.
Dillard, J.L., 1972. Black English: Its History and Usage in the United States. Random House, New York.
Dunn, R.S., 1972. Sugar and Slaves: The Rise of the Planter Class in the English West Indies, 1624–1713. The University of North Carolina Press, Chapel Hill.
Escure, G., 1997. Creole and Dialect Continua. Benjamins, Amsterdam.
Faine, J., 1937. Philologie créole: études historiques et étymologiques sur la langue créole d'Haïti. Imprimerie de l'Etat, Port-au-Prince.
Faraclas, N., Klein, T.B. (Eds.), 2009. Simplicity and Complexity in Creoles and Pidgins. Battlebridge, London.
Fayer, J., 2003. African interpreters in the Atlantic slave trade. American Anthropologist 45, 281–295.
Féral, C.de, 1989. Pidgin-English du Cameroun. Peters/SELAF, Paris.
Fortbath, P., 1977. The River Congo: The Discovery, Exploration, and Exploitation of the World's Most Dramatic River. Houghton Mifflin Co., Boston.
Gil, D., 2007. Creoles, complexity and associational semantics. Deconstructing creole: new horizons in language creation. In: Ansaldo, U., Matthews, S., Lim, L. (Eds.), John Benjamins, Amsterdam, Philadelphia, pp. 67–108.
Gray, E.G., Fiering, N. (Eds.), 2000. The Language Encounter in the Americas 1492–1800. Berghahn Books, New York.
Hall Jr., R.A., 1958. Creole languages and genetic relationships. Word 14, 367–373.
Hall Jr., R.A., 1962. The life-cycle of pidgin languages. Lingua 11, 151–156.
Hall Jr., R.A., 1966. Pidgin and Creole Languages. Cornell University Press, Ithaca, NY.
Hancock, I.F., 1986. The Domestic Hypothesis, Diffusion and Componentiality: An Account of Atlantic Anglophone Creole Origins. In: Muysken, P., Smith, N. (Eds.), pp. 71–102.
Hill, K.C. (Ed.), 1979. The Genesis of Language. Karoma, Ann Arbor.
Hesseling, D.C., 1897. Het Hollands in Zuid Afrika. De Gids 61, 148–162. Translated as "Dutch in South Africa". In: Markey, T.L., Roberge, P. (Eds.), On the Origin and Formation of Creoles: A Miscellany of Articles. Karoma, Ann Arbor, pp. 1–22.
Hjelmslev, L., 1938. Etudes sur la notion de parenté linguistique. Première étude: relations de parenté de langues créoles. Revue des Etudes Indo-Européennes 1, 271–286.
Holm, J., 1988. Pidgins and creoles. In: Theory and Structure, vol. 1. Cambridge University Press, Cambridge.
Holm, J., 1989. Pidgins and creoles. In: Reference Survey, vol. 2. Cambridge University Press, Cambridge.
Holm, J., 2004. Languages in Contact: The Partial Restructuring of Vernaculars. Cambridge University Press, Cambridge.

Huber, M., 1999. Atlantic creoles and the Lower Guinea Coast: a case against Afrogenesis. In: Huber, M., Parkvall, M. (Eds.), Spreading the Word: The Issue of Diffusion Among the Atlantic Creoles. University of Westminster Press, London, pp. 81–110.

Hurford, J.R., 2012. The Origins of Grammar: Language in the Light of Evolution. Cambridge University Press, Cambridge.

Hymes, D. (Ed.), 1971. Pidginization and Creolization of Languages. Cambridge University Press, Cambridge.

Jespersen, O., 1921. Language: Its Nature, Development and Origin. W.W. Norton & Co., Inc., New York.

Jourdan, C., 1991. Pidgins and creoles: the blurring of categories. Annual Review of Anthropology 20, 187–209.

Jourdan, C., 2009. Complexification or regularization of paradigms: the case of prepositional verbs in Solomon Islands Pijin. In: Aboh, E., Smith, N. (Eds.), Complex Processes in New Languages. Benjamins, Amsterdam, pp. 159–170.

Karttunen, F., 1994. Between Worlds: Interpreters, Guides, and Survivors. Rutgers University Press, New Brunswick, NJ.

Keesing, R.M., 1988. Melanesian Pidgin and the Oceanic Substrate. Stanford University Press, Stanford.

Kouwenberg, S., Singler, J.V. (Eds.), 2009. The Handbook of Pidgin and Creole Studies. Wiley-Blackwell, Malden, MA.

Krapp, G.P., 1924. The English of the Negro. The American Mercury 2, 190–195.

Kurath, H., 1928. The origin of dialectal differences in spoken American English. Modern Philology 25, 385–395.

Labov, W., 1972. Language in the Inner City: Studies in Black English Vernacular. University of Pennsylvania Press, Philadelphia.

Labov, W., 1998. Co-existent systems in African-American vernacular English. In: Mufwene, S.S., Rickford, J.R., Bailey, G., Baugh, J. (Eds.), African-American English: Structure, History, and Use. Routledge, London, pp. 110–153.

Lalla, B., D'Costa, J., 1990. Language in Exile: Three Hundred Years of Jamaican Creole. University of Alabama Press, Tuscaloosa, AL.

Lawrance, B., Osborn, E.L., Roberts, R.L. (Eds.), 2006. Intermediaries, Interpreters, and Clerks: African Employees in the Making of Colonial Africa. University of Wisconsin Press, Madison.

Lefebvre, C., 1998. Creole Genesis and the Acquisition of Grammar: The Case of Haitian Creole. Cambridge University Press, Cambridge.

Lefebvre, C., White, L., Jourdan, C. (Eds.), 2006. L2 Acquisition and Creole Genesis. Benjamins, Amsterdam.

Lefebvre, C. (Ed.), 2011. Creoles, Their Substrates, and Language Typology. Benjamins, Amsterdam.

Manessy, G., 1994. Créoles, pidgins, variétés véhiculaires: procès et genèse. CNRS Editions, Paris.

Markey, T.L., 1982. Afrikaans: creole or non-creole? Zeitschrift fur Dialektologie und Linguistik 2, 169–207.

McDavid Jr., R., 1950. Review of Lorenzo Dow Turner's Africanisms in the Gullah dialect. Language 26, 323–333.

McDavid Jr., R., McDavid, V., 1951. The relationship of the speech of the American Negroes to the speech of whites. American Speech 26, 3–17.

McNeill, D., 2012. How Language Began: Gesture and Speech in Human Evolution. Cambridge University Press, Cambridge.

McWhorter, J.H., 1998. Identifying the creole prototype: vindicating a typological class. Language 74, 788–818.

McWhorter, J.H., 2001. The world's simplest grammars are creole grammars. Linguistic Typology 5, 125–166.

McWhorter, J.H., 2012. Linguistic Simplicity and Complexity: Why Do Languages Undress? DeGruyter, Mouton.

Metcalf, A.E., 2005. Go-Betweens and the Colonization of Brazil, 1500–1600. University of Texas Press, Austin.

Meyerhoff, M., 2009. In: Kouwenberg, S., Singler, J.V. (Eds.), Forging Pacific Pidgin and Creole Syntax: Substrate, Discourse, and Inherent Variability, pp. 48–73.

Michaelis, S., 1993. Temps et aspect en créole seychellois: valeurs et interférences. Helmut Buske, Hamburg.

Michaelis, S., Maurer, P., Haspelmath, M., Huber, M. (Eds.), 2013. The Atlas of Pidgin and Creole Language Structures. Oxford University Press, Oxford, UK.

Mufwene, S.S., 1986a. Les langues créoles peuvent-elles être définies sans allusion à leur histoire? Etudes Créoles 9, 135–150.

Mufwene, S.S., 1986b. The universalist and substrate hypotheses complement one another. In: Muysken, P., Smith, N. (Eds.), Substrata versus Universals in Creole Genesis. Benjamins, Amsterdam, pp. 129–162.

Mufwene, S.S., 1988. Why study pidgins and creoles? Column. Journal of Pidgin and Creole Languages 3, 265–276.

Mufwene, S.S., 1990. Transfer and the substrate hypothesis in creolistics. Studies in Second Language Acquisition 12, 1–23.

Mufwene, S.S., 1992. Why grammars are not monolithic. In: Brentari, D., Larson, G.N., McLeod, L.A. (Eds.), The Joy of Grammar: A Festschrift in Honor of James D. McCawley. Benjamins, Amsterdam, pp. 225–250.

Mufwene, S.S. (Ed.), 1993. Africanisms in Afro-American Language Varieties. University of Georgia Press, Athens.

Mufwene, S.S., 1994. On decreolization: the case of Gullah. In: Morgan, M. (Ed.), Language, Loyalty, and Identity in Creole Situations. Center for Afro-American Studies, Los Angeles, pp. 63–99.

Mufwene, S.S., 1996. The founder principle in creole genesis. Diachronica 13, 83–134.

Mufwene, S.S., 1997. Jargons, pidgins, creoles, and koinés: what are they? In: Spears, A.K., Winford, D. (Eds.), Pidgins and Creoles: Structure and Status. Benjamins, Amsterdam, pp. 35–70.

Mufwene, S.S., 1999. In: DeGraff, M. (Ed.), The Language Bioprogram Hypothesis: Hints from Tazie, pp. 95–127.

Mufwene, S.S., 2001. The Ecology of Language Evolution. Cambridge University Press, Cambridge.

Mufwene, S.S., 2003. Genetic linguistics and genetic creolistics. Journal of Pidgin and Creole Languages 18, 273–288.

Mufwene, S.S., 2005. Créoles, Écologie Sociale, Évolution Linguistique. L'Harmattan, Paris.

Mufwene, S.S., 2008. Language Evolution: Contact, Competition and Change. Continuum, London.

Mufwene, S.S., 2009a. The indigenization of English in North America. In: Hoffmann, T., Siebers, L. (Eds.), World Englishes: Problems, Properties, Prospects. Benjamins, Amsterdam, pp. 353–368.

Mufwene, S.S., 2009b. In: Aboh, E., Smith, N. (Eds.), Postscript: Restructuring, Hybridization, and Complexity in Language Evolution, pp. 367–400.

Mufwene, S.S., 2010. SLA and the emergence of creoles. Studies in Second Language Acquisition 32, 359–400.

Mufwene, S.S., 2013. Simplicity and complexity in creoles and pidgins: what's the metric? A response to Faraclas, N. In: Klein, T.B. (Ed.), Simplicity and Complexity in Creoles and Pidgins, Journal of Language Contact, 6, pp. 161–179.

Mufwene, S.S., 2014. Globalisation économique mondiale des XVIIe–XVIIIe siècles, émergence des créoles, et vitalité langagière. In: Carpooran, Arnaud (Ed.), Langues créoles, mondialisation et éducation. Creole Speaking Unit, St. Louis, Mauritius, pp. 23–79.

Mühlhäusler, P., 1981. The development of the category of number in Tok Pisin. In: Muysken, P. (Ed.), Generative Studies on Creole Languages. Foris, Dordrecht, pp. 35–84.

Mühlhäusler, P., 1985. The number of pidgin Englishes in the Pacific. Papers in Pidgin and Creole Linguistics No.1 Pacific Linguistics A-72, 25–51.

Mühlhäusler, P., 1986. Pidgin and Creole Linguistics. Basil Blackwell, Inc., New York.

Muysken, P., Smith, N. (Eds.), 1986. Substrata versus Universals in Creole Genesis. Benjamins, Amsterdam.

Ostler, N., 2005. Empires of the Word: A Language History of the World. Harper Collins, New York.

Plag, I., 2009. Creoles as interlanguages: phonology. Journal of Pidgin and Creole Languages 24, 119–138.

Poplack, S. (Ed.), 2000. The English History of African-American English. Blackwell, Oxford.

Poplack, S., Tagliamonte, S., 2001. African American English in the Diaspora. Blackwell, Malden, MA.

Posner, R., 1985. Creolization as typological change: some examples from Romance syntax. Diachronica 2, 167–188.

Rawley, J.A., 1991. Slave trade. In: Foner, E., Garraty, J.A. (Eds.), The Reader's Companion to American History. Houghton Mifflin, Boston, pp. 994–995.

Reader, J., 1997. Africa: Biography of a Continent. Alfred Knopf, New York.

Rickford, J.R., 1987. Dimensions of a Creole Continuum: History, Texts, and Linguistic Analysis of Guyanese Creole. Stanford University Press, Stanford.

Rickford, J.R., 1990. Number delimitation in Gullah: a response to Mufwene. American Speech 65, 148–163.

Rickford, J.R., 1998. The creole origins of African-American Vernacular English: evidence from copular absence. In: Mufwene, S.S., Rickford, J.R., Bailey, G., Baugh, J. (Eds.), African-American English: Structure, History, and Use. Routledge, London, pp. 154–200.

Rickford, J.R., 1999. African American Vernacular English: Features, Evolution, Educational Implications. Blackwell, Malden, MA.

Rickford, J.R., Handler, J.S., 1994. Textual evidence on the nature of early Barbadian speech, 1676–1835. Journal of Pidgin and Creole Languages 9, 221–255.

Roberts, S.J., 1998. The role of diffusion in the genesis of Hawaiian Creole. Language 74, 1–39.

Romaine, S., 1988. Pidgin and Creole Languages. Longman, London.
Sankoff, G., 1979. The genesis of a language. In: Hill, K.C. (Ed.), The Genesis of Language. Karoma, Ann Arbor, pp. 23–47.
Sankoff, G., Brown, P., 1976. The origins of syntax in discourse: a case study of Tok Pisin relatives. Language 52, 631–666.
Sankoff, G., 1993. Focus in Tok Pisin. In: Byrne, F., Winford, D. (Eds.), Focus and Grammatical Relations in Creole Languages. Benjamins, Amsterdam, pp. 117–140.
Schlupp, D., 1997. Modalités prédicatives, modalités aspectuelles, et auxiliaires en créole. Max Niemeyer, Tübingen.
Schneider, E.W., 1989. American Earlier Black English: Morphological and Syntactic Variables. University of Alabama Press, Tuscaloosa.
Schuchardt, H., 1909. Die lingua franca. Zeitschrift fur Romanische Philologie 33, 441–461.
Schuchardt, H., 1914. Die Sprache der Saramakkaneger in Surinam. Johannes Muller, Amsterdam.
Schumann, J., 1978. The Pidginization Process: A Model for Language Acquisition. Newbury House, Rowley, MA.
Sebba, M., 1997. Contact Languages: Pidgins and Creoles. St. Martin's Press, New York.
Siegel, J., 2006. Links between SLA and Creole Studies: Past and Present. In: Lefebvre, C., White, L., Jourdan, C. (Eds.), L2 Acquisition and Creole Genesis. Benjamins, Amsterdam, pp. 15–46.
Siegel, J., 2008. The Emergence of Pidgin and Creole Languages. Oxford University Press, Oxford.
Singler, J.V., 1988. The homogeneity of the substrate as a factor in pidgin/creole genesis. Language 64, 27–51.
Singler, J.V. (Ed.), 1990. Pidgin and Creole Tense-Mood-Aspect Systems. John Benjamins, Amsterdam.
Singler, J.V., 1996. Theories of creole genesis, sociohistorical considerations, and the evaluation of evidence: the case of Haitian Creole and the relexification hypothesis. Journal of Pidgin and Creole Languages 11, 185–230.
Singler, J.V., 2009. The sociohistorical context of creole genesis. In: Kouwenberg, S., Singler, J.V. (Eds.), pp. 332–358.
Smith, N., 1995. An annotated list of pidgins, creoles, and mixed languages. In: Arends, J., Muysken, P., Smith, N. (Eds.), Pidgins and Creoles: An Introduction. Benjamins, Amsterdam, pp. 331–374.
Stewart, C. (Ed.), 2007. Creolization: History, Ethnography, Theory. Left Coast Press, Walnut Creek, CA.
Stewart, W., 1965. Urban Negro speech: sociolinguistic factors affecting English teaching. In: Shuy, R. (Ed.), Social Dialects and Language Learning. The National Council of Teachers of English, Champaign, IL, pp. 10–18.
Sylvain, S., 1936. Le créole haïtien: morphologie et syntaxe. Imprimerie De Meester, Wettern, Belgium.
Tagliamonte, S., Poplack, S., 1993. The zero-marked verb: testing the creole hypothesis. Journal of Pidgin and Creole Languages 8, 171–206.
Thomason, S.G., 1997. Contact Languages: A Wider Perspective. John Benjamins, Amsterdam.
Thomason, S.G., 2001. Language Contact: An Introduction. Georgetown University Press, Washington, DC.
Thomason, S.G., Kaufman, T., 1988. Language Contact, Creolization, and Genetic Linguistics. University of California Press, Berkeley.
Trudgill, P., 1983. On Dialect: Social and Geographic Perspectives. New York University Press, New York.
Trudgill, P., 2004. New-Dialect Formation: The Inevitability of Colonial Englishes. Oxford University Press, Oxford.
Turner, L.D., 1949. Africanisms in the Gullah Dialect. University of Chicago Press, Chicago.
Valdman, A. (Ed.), 1977. Pidgin and Creole Linguistics. Indiana University Press, Bloomington.
Valkhoff, M.F., 1966. Studies in Portuguese and Creole – with Special Reference to South Africa. Witwatersrand University Press, Johannesburg.
Van Dyke, P.A., 2005. The Canton Trade: Life and Enterprise on the China Coast, 1700–1845. Hong Kong University Press, Hong Kong.
Vinson, J., 1882. Créole. Dictionnaire des sciences anthropologiques et ethnologiques. Paris.
Winford, D., 1997. Re-examining Caribbean English Creole continua. World Englishes 16, 233–279.

Pilgrimage

Simon Coleman, University of Durham, Durham, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11445–11448, © 2001, Elsevier Ltd.

Abstract

Pilgrimage involves the movement of the traveler or group of travelers away from home, usually with a specific, sacred goal in mind. The institution is of interest within the social sciences because it can entail the cross-cutting of conventional geographical, cultural, and religious boundaries. Older theories of pilgrimage stressed its capacity to encourage moral regeneration and the social unification of otherwise dispersed populations. Many anthropologists have compared pilgrimage with rites of passage within tribal peoples and argued that it allows people to strip away their normal, everyday identities during the restricted space and time of the journey. More recent work in the social sciences emphasizes the continuity between pilgrimage activities and the social, religious, and political fragmentation of daily life. Scholars note that pilgrimage has maintained or even increased its popularity even as the boundaries between religiously motivated travel and mass tourism have sometimes become blurred. Contemporary pilgrimage has therefore been examined in the broader context of the expanded mobility of people, cultures, and objects that is evident in many parts of the world.

Pilgrimage involves the movement of a traveler or group of travelers away from home, usually with a specific, sacred goal in mind. In most cases, the term implies a journey to a place that has been sanctified by the presence or actions of a divinity or holy figure. Certain variations on this pattern are possible. Pilgrims sometimes wander continuously without a fixed destination, or the journey may be considered to occur in a purely metaphorical sense, involving spiritual development and transformation within the person. This article focuses, however, on the most common understanding of pilgrimage: literal movement across space that may also result in the crossing of conventional cultural, social, and religious boundaries.

Themes in the Study of Pilgrimage

Social scientific studies of pilgrimage were relatively sparse until the 1970s, although in the 1950s Eric Wolf had published a short but influential article on pilgrimage to the Virgin of Guadalupe in Mexico (Wolf, 1958). Since the early 1970s, research has expanded and covered all of the world religions. Such academic engagement reflects a continuing and even increasing interest in pilgrimage that has become evident among participants who have better access than ever before to safe and effective modes of travel.

Many earlier studies adopted Durkheimian, functionalist perspectives in analyzing the significance and effects of pilgrimage. In other words, they interpreted the process of traveling to and gathering at a sacred place as providing ritualized opportunities for social unification and moral regeneration within populations that were normally widely dispersed. For instance, Wolf (1958) argued that the shrine of the Virgin of Guadalupe in Mexico City represented a dominant symbol for the whole nation, condensing various historical, mythical, and cultural themes and focusing them on the activity of pilgrimage to a single place.

The most influential social scientific approach to the subject has come from the anthropologist Victor Turner (see, for instance, Turner and Turner, 1978). He notes that, after an original founding event such as a vision, journey, or martyrdom, pilgrimages most frequently involve the gradual organization and routinization of the sacred journey. More speculatively, Turner argues that pilgrimage in the world religions shares certain characteristics with rites of passage in tribal societies. In both forms of ritual, the person is removed from the space and time of the everyday social environment and undergoes an emotionally and physically testing experience that can have lasting effects on perceptions of self-identity. Pilgrims thus enter the realm of what Turner calls social antistructure or *communitas*. He means by these terms that pilgrimage can involve the stripping away of everyday social statuses among participants, resulting in liberation from conformity to conventional norms and hierarchies. Pilgrims ideally cultivate direct and egalitarian bonds of comradeship during the temporary, set-apart time of their journey and dwelling at a shrine. *Communitas* therefore has the potential to contain an element of social critique since it puts the taken-for-granted quality of conventional institutions in question. Representatives of orthodoxy in many of the salvation religions often feel ambivalent towards pilgrimage and attempt either to suppress or control its tendencies to foment lay frustration with religious hierarchies.

Turner notes further that pilgrimage is frequently associated with images of death and rebirth, mirroring the symbolic destruction and reconstruction of the religious traveler's identity. For instance, many who go on the major pilgrimage (or *hajj*) to Mecca will choose to be buried in the white robes they wore on the pilgrimage. It is believed that these people will thus present themselves at the Last Judgment in the garb appropriate to a rejection of worldly, profane actions.

Turner's work has proved to be so significant partly because it appears to provide a vocabulary and theoretical framework that can be generalized and applied to varied cultural contexts. However, numerous subsequent works have criticized aspects of his thesis that pilgrimage is inherently antistructural in form

and significance. Authors have pointed out that such a paradigm is possibly more reflective of a Christian ideal of human fellowship than of social and political realities in religious systems in general. Many examples of pilgrimage appear actually to reinforce situations of conflict between participants. For instance, Messerschmidt and Sharma (1981) examine Hindu pilgrimage in Nepal and find that caste and class divisions are not transcended through common engagement in the ritual. They argue further that conventional Christian ideals of egalitarianism and homogeneity contradict certain Hindu assumptions concerning the inevitable persistence of religious hierarchy.

Powerful and widely cited criticisms of the Turnerian paradigm are contained in a book that is ostensibly devoted to studies of Christian pilgrimage but presents arguments that have much wider application. Eade and Sallnow (1991) characterize pilgrimage sites as ideal locations for contesting, rather than consolidating, the sacred among religious participants. As centers of power that can provide religious legitimation for political aspirations, shrines often present powerful incentives for a particular interest group to attempt to control the behavior and beliefs of religiously motivated visitors (the point is illustrated clearly by the politically charged and often violent history of Jerusalem). Pilgrimages, according to this view, provide the material and ideological focus for competing discourses and agendas, not only between sacred and secular authorities, but also between lay pilgrims and clerics, locals and incomers, members of different denominations, and so on. Sacred sites are described by Eade and Sallnow as the equivalent of empty symbolic vessels waiting to be filled by the possibly mutually contradictory assumptions of the various people who visit them. The authors also suggest that to regard pilgrimage as a unitary institution is highly problematic, even within a single religious system, since it can contain so many historically specific behaviors and meanings. They add, however, that Christian and perhaps other forms of pilgrimage can be conceptualized as involving a basic triad of 'person,' 'place,' and 'text.'

In fact, Turner did anticipate at least some of the later criticisms that have been leveled at his model of pilgrimage. He was aware of the potential for pilgrimage sites to be co-opted by political elites, the fragility of states of *communitas*, and the capacity of pilgrimage sites to encourage widely differing interpretations from pilgrims. However, the notion that pilgrimage should be seen as an arena for contestation as much as cooperation has proved popular in recent years. Its adoption by many scholars is not surprising since it resonates with contemporary social scientific perspectives that have been influenced by postmodern understandings of culture as unstable, fragmentary, and open to multiple appropriation.

A less controversial set of issues, discussed by Turner and others, is raised by the contention that pilgrimage can encourage participants to consider themselves part of translocal religious and cultural systems. Among Hindus, for instance, a sense of participation in a sacred geography that covers the whole of the Indian subcontinent can be cultivated by travel to shrines. Some devotees choose to circumambulate the entire country, visiting sacred spots along the way. Holy places themselves may be officially or unofficially ranked in importance, thus constituting an internally heterogeneous but widely dispersed framework of sites. Among Christians, Jerusalem (located within 'the Holy Land' itself) has often been regarded as the pre-eminent location for pilgrimage, containing as it does so many sites associated with biblical events. It is followed in importance by other locations of transnational importance, such as Rome or Compostela, and then by pilgrimage locations that have significance in national or more narrowly defined regional contexts.

Methods of Study

Pilgrimage raises unusual problems in relation to conventional social scientific methodologies. Any given site is located within a much wider political, economic, and religious framework that must be taken into account in analysis. As a religious practice, pilgrimage has often complemented (and incorporated) other translocal activities like trade, tourism, and exploration. Pilgrims themselves may move across geographical, social, and religious boundaries, so that the scholar is faced with a diffuse and literally shifting object of study.

Various methods can be used to encapsulate the most important aspects of the phenomenon. Pilgrims can be followed as they journey to a shrine (Gold, 1988; Frey, 1998). More commonly, a single pilgrimage center or shrine is focused upon, and the activities of various pilgrim groups observed as they pass through the sacred center (Dubisch, 1995). At a wider, macro level of analysis, whole systems of pilgrimage within a religious tradition may be studied (Turner and Turner, 1978), or comparisons made between pilgrimage systems in the various world religions (Coleman and Elsner, 1995).

The fact that pilgrimage sites tend to persist over time provides a good opportunity to combine historical with sociological interests (Turner and Turner, 1978), and much work in this area demonstrates the ways in which a given site can retain its identity as a sacred place despite the fact that it is appropriated by different and sometimes competing religious systems over perhaps hundreds of years. Guadalupe, for instance, provides a good example of a shrine that has probable pre-Christian as well as Christian roots. The mixture of religious associations increases its significance as a site whose symbolism and mythology can be invoked in current debates concerning the relationship between precolonial, colonial, and post-colonial identities in Mexico.

Future Themes of Research

Pilgrimage has the potential to be at the center of much future research that could be of importance to the social sciences in general. As a social phenomenon, it encourages the researcher to engage with some key theoretical debates in a variety of disciplines. The current popularity of pilgrimage as an activity raises the question of its significance for predictions of secularization and/or religious revitalization in various parts of the world. Some scholars have also speculated over the extent to which apparently secular activities such as going on holiday or visiting the graves of pop stars share the same structural

features as pilgrimage. Certainly, it seems likely that tourism and pilgrimage will become increasingly intertwined in many cultural contexts in the future, given the popularity and flexibility of both activities. In addition, pilgrimage studies are relevant to ongoing discussions of the significance of processes of globalization in the modern world. Studies of sacralized forms of travel, involving the transference of people, ideas, and objects across national borders, can complement other research that is concerned to understand how social, cultural, and political units are being transformed by developments in technologies of mass travel and communication. Furthermore, in a world that is characterized by the displacement of populations from traditional homelands, pilgrimage may emerge as an increasingly important institutional means through which to unite diasporic communities. Tweed (1997) has recently argued, for instance, that pilgrimage allows Cuban exiles in Florida to map the landscape and history of their country of birth on to a new environment. Finally, pilgrimage provides many possible avenues of research in the light of increased social scientific interest in the body and embodiment, ranging from a focus on beliefs in miraculous healing to analysis of forms of movement associated with travel to and around sacred shrines.

See also: Migration, Theory of; Monasticism: West and East; Tradition: Social.

Bibliography

Coleman, S., Elsner, J., 1995. Pilgrimage: Past and Present in the World Religions. Harvard University Press, Cambridge, MA.

Dubisch, J., 1995. In a Different Place: Pilgrimage, Gender, and Politics of a Greek Island Shrine. Princeton University Press, Princeton, NJ.

Eade, J., Sallnow, M.J. (Eds.), 1991. Contesting the Sacred: The Anthropology of Christian Pilgrimage. Routledge, London.

Frey, N.L., 1998. Pilgrim Stories: On and Off the Road to Santiago. University of California Press, Berkeley, CA.

Gold, A.G., 1988. Fruitful Journeys: The Ways of Rajasthani Pilgrims. University of California Press, Berkeley, CA.

Morinis, A. (Ed.), 1992. Sacred Journeys: The Anthropology of Pilgrimage. Greenwood Press, Westport, CT.

Messerschmidt, D.A., Sharma, J., 1981. Hindu pilgrimage in the Nepal Himalayas. Current Anthropology 22, 571–572.

Turner, V., Turner, E., 1978. Image and Pilgrimage in Christian Culture. Columbia University Press, New York.

Tweed, T.A., 1997. Our Lady of the Exile: Diasporic Religion at a Cuban Catholic Shrine in Miami. Oxford University Press, Oxford, UK.

Wolf, E., 1958. The Virgin of Guadalupe: A Mexican national symbol. Journal of American Folklore 7 (1), 34–39.

Place in Geography

Andrew Williams, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Relph, volume 17, pp. 11448–11451, © 2001, Elsevier Ltd.

Abstract

Place is a foundational concept in human geography but is often regarded as synonymous with terms such as location, locale, region, space, territory, and landscape. It is important to illustrate the diverse ways in which place has been conceptualized through the evolving strands of humanist, phenomenological, radical (including Feminist, Marxist, and Postcolonial), and more recent posthumanist and nonrepresentational thought. This article provides an overview of the evolving theorization of place in geography, and reflects on how assemblage thinking may refine geographical conceptions of place.

Place is a foundational concept in human geography, but, when taken uncritically, is used interchangeably with terms of location, locale, region, space, territory, and landscape. Throughout the history of the discipline, place alongside space has provided the building blocks of geographical investigation, and continue to undergo theoretical refinement and debate. This article provides an overview of the evolving theorization of place in human geography, and reflects on how the concept of place might be developed through the notion of assemblage.

Place

The concept of place has played an important role throughout Western philosophy from Plato to the present (see Casey, 1997; Livingstone, 1992). For the most part place was thought of as an unproblematic category, viewed as a passive and empty container (McKeon, 1941) or a human-wrought occupation of preexisting, undifferentiated space. During the second half of the twentieth century, new philosophical approaches, coupled with the changing character of actual 'places' under globalization, brought new questions about ideas of 'place.' These developments can be seen in Agnew's (1987) three basic attributes of place. In its elementary meaning, place refers to 'location,' the 'where of something' in an absolute or relative sense (through agreed measures of latitude or longitude marks, or 5 km south from x). This conception underpinned the spatial science turn in the 1960s and the tenets of central place theory (Christaller, 1966), that considered towns and villages in terms of measurable properties and relative locations, and in doing so largely emptied places of subjective content. The new scientific method sought to replace the overly descriptive tendency that thus far characterized traditional regional geography and its chorological approach to collecting and classifying objective attributes about places (for instance, population size, natural resources, economic productivity) (see Hartshorne, 1959). The second of Agnew's aspect of place is 'locale,' which denotes the material things that structure social and cultural life (buildings, roads, rivers, marketplaces, signposts, for example). The third place-characteristic in Agnew's typology is 'sense of place,' the ways in which places are made meaningful through personal experience and through social mediations and constructions of place.

Humanistic Geography

Sense of place was a driving theme of humanistic geography during the 1970s. Geographers such as Yi-Fu Tuan, Anne Buttimer, Edward Relph, and David Seamon, dissatisfied with the lack of attention spatial science tradition gave to the meanings given to places, drew extensively on the phenomenology of Martin Heidegger and Maurice Merleau-Ponty (Tuan, 1971; Seamon, 1979; Heidegger, 1971; Merleau-Ponty, 1962). This approach sought to clarify the subjectivity of place experience, investigating individuals' attachments to particular places, and the symbolic role of places in fostering individual senses of belonging and dwelling in different cultures.

Place and Placelessness

The modernism of global capital flows, according to Relph (1976), erodes diversity and place distinctiveness by fabricating, uprooting, commodifying 'authentic' place-identities into internationally standardized shopping spaces, modernist planning, and tourist landscapes. This has culminated in scholars describing these homogenous sites as 'nonplaces' (Augé, 1995) or suffering a condition of 'placelessness' (Relph, 1976), where the loss of place diversity equates loss of 'authentic' cultures that produce senses of place. However, Harvey (1989), Entrikin (1991), and Massey (1997) warn of the reactionary impulse entailed in nostalgia for single and stable place-based attachments, and instead point to the opportunities to establish new progressive senses of place that are translocal in character (also see Featherstone et al., 2012). Despite these criticisms, the contributions of humanist geography fundamentally put place, sense of place, and placelessness at the center of a human geography agenda for years to come.

Radical Geography and the Cultural Turn

During the 1980s and 1990s the rise of new critical geographies (feminist, Marxist, postcolonial) and the 'cultural turn' saw place remain a key concern of geographers, albeit with a fundamental distinction. Place became understood less through the notion of the intentional human subject, instead scholars emphasized the role of power and politics, the ways

places are (re)made, and the role place plays in constituting social relationships in society-space (Cresswell, 2011). David Harvey (1973), who had been one of the first to challenge the positivist conception of space as an empty container or pregiven entity, developed a Marxist approach that emphasized how social practices and processes embedded in particular places and spaces are integral to the construction, and overthrowal, of capitalism. Doreen Massey (1984, 1991, 1993), also inspired by critical social theory, Feminism, and Marxist political-economy, set out a more *relational* conception of place through her analysis of political-economic and social processes that produce distinct urban and regional inequalities. Her work has been developed into at least four propositions for thinking about place (Cresswell, 2011). Firstly, place is not static but is produced through mobilities, flows, and differential power-relations. Place is a unique point of interaction within networks of social and economic relations. Secondly, place does not have boundaries, as in the superficial binary between local–global, rather place can be defined by its 'outside linkages.' Here global processes constitute 'local' senses of place and global economic flows are not disembodied ethereal entities but grounded in particular places, practices, and performances. Thirdly, places do not have single unique identities but place identity is a site of contestation by different groups in society, and by multiple, often contradictory, political-economic and social processes. Fourthly, and related, places are specific but are continually reproduced and open to contestation. This latter point led Massey and others to close the discursive distance between 'local' and 'general' struggles, emphasizing how place-based politics need not be parochial but can be networked into more egalitarian relations with global processes (Massey, 1991).

With the advent of the cultural turn in the 1990s (*see* Cultural Geography) the focus shifted from the humanistic accounts of 'sense of place' or Marxist perspectives on capitalist political-economy. Instead the politics of identity, difference and representation (Barnes and Duncan, 1992; Duncan and Ley, 1993) led geographers to develop conceptualizations of place that sought to unravel and challenge the spatial (re) production of inequalities along lines of gender, class, disability and sexuality. Cresswell (1996) examines how different identities are assigned to 'belong' to particular places. These contributions have led to politics of uncovering the explicit and implicit exclusions inherent in place construction, for instance, how homelessness is considered 'out of place' in rural areas through social constructions of the rural idyll (see Cloke et al., 2000). These discourses are reiteratively performed to produce dominant 'orderings of places.' Yet geographers working in a number of different arenas have emphasized the ways place-orderings are open for transgression and resistance. (Gill Valentine's, (1997; Skelton and Valentine, 1998) work on children's geographies and youth cultures illustrates how new meanings, senses of belonging, and experiences of place can be carved out from adult notions of public/private spaces. Geographical research on race has exposed how racisms are (re)produced through legislative spatial segregation (Smith, 1989), through dominant representations in film, television, music, and literature (Keith and Pile, 1993) as well as how places are socially constructed to maintain the 'neutrality' of 'white privilege' (Anderson, 1991; Kobayashi and Peake, 2000; Holloway, 2005). Feminist contributions (Rose, 1993) have challenged the humanistic assumptions of homeplace as stable and secure by highlighting constructions of homeplaces as sites of gendered labor and domestic violence (*see* Gender and Place). Geographers of sexuality have helped deconstruct everyday and institutional spaces of heteronormativity, and emphasize the role space plays in the construction and exclusion of a diverse range of homosexual and heterosexual performances, practices and identities. (Valentine, 1993; Hubbard, 2000). Geographies of religion has drawn attention to the ways places are made sacred (Yorgason and Della Dora, 2009), the construction of religious subjectivity (Hopkins et al 2013) and the ways meaning and experience of the 'divine' is mediated through identity, belief and embodiment (Dewsbury and Cloke, 2009)

Feminist and postcolonial has helped critique the construction and reproduction of place in systems of power/knowledge. In particular, contributions have stressed how place is integral to shaping academic positionalities and situated knowledges (Rose, 1997) – guiding what, how, and why we research and write exactly what we do. Unsurprisingly then, place in academia – country, region, networks, and labor production (see Sidaway, 2000) – is a contested site of social reproduction, which imbricates the academic not only into existing lines of inclusion/exclusion of knowledge but also into webs of proximate or distant relations to otherness, scripting the propensity for responsibility (Cloke, 2004).

'New' Cultural Geography 2000s–Present

The contribution of the cultural turn in informing geographic thinking on place has been widespread, even to the extent that Kirsch (2013: 1) remarks "we are all cultural geographers now," by which, he refers to the normalization of the philosophical, conceptual, and methodological underpinnings of cultural geography across human geography's subdisciplines. Although some identify this as diluting the analytical rigor of cultural geography proper (Cresswell, 2010; Wylie, 2010). If the cultural turn in the 1990s was guilty of overly textual accounts of place, representation, and meaning, recent contributions have sought to reemphasis lived practice, performance, and the materiality of place (Kirsch, 2013; *see* Materiality and Culture), alongside vitalist, post-phenomenological, and nonrepresentational approaches that emphasize the affective, embodied, and performative aspects of place (Dewsbury, 2000; Anderson and Harrison, 2010; *see* Emotion and Affect in Anthropology). These trends do not signal a renaissance of humanistic geography, but rather the inverse, decentering the human subject to focus on the materiality and liveliness of actions that constitute fleeting, but nonetheless powerful, senses of place, meaning, and experience. Meaning is not understood through preestablished systems and structures, or through the intentional consciousness of the bounded individual, instead, the focus is the *production* of meaning through ongoing habitual interaction with the world. These recent developments in sociospatial theorizing have led geographers to increasingly conceptualize place as an assemblage.

Place as Assemblage

Assemblage theory conceives places as constituted through the interplay of the material, the immaterial, and the representational (see Cresswell, 2011 on place and assemblage theory). Based on the post-structural writings of Deleuze and Guattari (1987) and De Landa (2006), geographers have begun to utilize assemblage theory to break down divisions between social–material, near–far, structure–agency to appreciate the contingency of place. That is, how places – sense of places, dwelling, locale, and even economic practices – are composed of heterogeneous elements that may be human and nonhuman, organic and inorganic, technical and natural; yet, gathered and tied together into some form of provisional sociospatial formation (Anderson and McFarlane, 2011). Anderson and McFarlane (2011) note three distinct contributions assemblage theory brings to geographical work on place. Firstly, as an analytical tool, assemblage thinking emphasizes the labor of territorialization (coherence) and deterritorialization (destabilize), the work of coalescing sociomaterial practices that may be otherwise unrelated. This accentuates the emergence rather than the resultant formation, accounting for the quality of flows that hold things together and crucially, the cracks where assemblages can be reterritorialized. The political purchase of understanding place in terms of assemblage lies in the unraveling of flows and connections that entangle heterogeneous elements to produce durable impressions of power (Connolly, 2008). Second, the notion of assemblage seeks to steer a third course between the representational and nonrepresentational, conceptualizing place as a complex and dynamic collage of things (materiality), representations, and performative practices (Cresswell, 2011). The strength of such an inquiry will be in its application, in the fine detail of connecting the embodied, fleeting ephemeral aspects of place with the discursive and expressive meanings and wider processes at work in constituting social worlds. Thirdly, assemblage is an ethos of engaging with the messiness of the world. Places are not reducible to any preexisting framework used to study them. Assemblage thinking encourages a grounded experimentation with the processes of coherence in a way that opens up seemingly dominant place-assemblages to the possibility of resistance and reconfiguration.

See also: Affect and Emotion, Anthropology of; Cultural Geography; Everyday Life, Anthropology of; Gender and Place; Landscape; Materiality and Culture; Scale in Geography; Space and Social Theory in Geography.

Bibliography

Agnew, J., 1987. Place and Politics: The Geographical Mediation of State and Society. Allen and Unwin, Boston.
Anderson, B., Harrison, P., 2010. Taking-Place: Non-representational Theories and Geography. Ashgate, London.
Anderson, B., McFarlane, C., 2011. Assemblage and geography. Area 43 (2), 124–127.
Anderson, K., 1991. Vancouver's Chinatown: Racial Discourse in Canada, 1875–1980. McGill-Queen's University Press, Montreal.
Augé, M., 1995. Non-places: Introduction to an Anthropology of Supermodernity. Verso, New York.
Barnes, T., Duncan, J. (Eds.), 1992. Writing Worlds, Discourse, Text and Metaphor in the Representation of Landscape. Routledge, London.
Casey, E.S., 1997. The Fate of Place: A Philosophical History. University of California Press, Berkley.
Christaller, W., 1966. Central places in Southern Germany. Prentice-Hall, New Jersey.
Cloke, P., 2004. Exploring boundaries of professional/personnel/personal practice and action: being and becoming in Khayelitsha Township, Cape Town. In: Fuller, D., Kitchin, R. (Eds.), Radical Theory/critical Praxis: Making a Difference beyond the Academy. Praxis Press, Kelowna, pp. 92–102.
Cloke, P., Milbourne, P., Widdowfield, R., 2000. Homelessness and rurality: "out-of-place" in purified space? Environment and Planning D: Society and Space 18, 715–735.
Connolly, W., 2008. Capitalism and Christianity, American Style. NC Duke University Press, Durham.
Cresswell, T., 1996. In: Place/Out of Place: Geography, Ideology and Transgression. University of Minnesota Press, Minneapolis.
Cresswell, T., 2010. New cultural geography – an unfinished project? Cultural Geographies 17, 169–174.
Cresswell, T., 2011. Place – part 1. In: Agnew, J., Duncan, J. (Eds.), The Wiley-Blackwell Companion to Human Geography. Oxford, Blackwell, pp. 236–244.
De Landa, M., 2006. A New Philosophy of Society: Assemblage Theory and Social Complexity. Continuum, London.
Deleuze, G., Guattari, F., 1987. A Thousand Plateaus: Capitalism and Schizophrenia. University of Minnesota Press, Minneapolis.
Dewsbury, J.D., 2000. Performativity and the event: enacting a philosophy of difference. Environment and Planning D: Society and Space 18, 473–496.
Dewsbury, J.D., Cloke, P., 2009. Spiritual landscapes: existence, performance and immanence. Social & Cultural Geography 10, 695–711.
Doreen, M., 1984. Spatial Divisions of Labour. Macmillan, London.
Duncan, J., Ley, D., 1993. Place/Culture/Representation. Routledge, London.
Entrikin, J.N., 1991. The Betweenness of Place: Towards a Geography of Modernity. John Hopkins Press, Baltimore.
Featherstone, D., Ince, A., Mackinnon, D., Strauss, K., Cumbers, A., 2012. Progressive localism and the construction of political alternatives. Transactions of the Institute of British Geographers 37, 177–182.
Hartshorne, R., 1959. Perspectives on the Nature of Geography. Rand McNally, Chicago.
Harvey, D., 1973. Social Justice and the City. Arnold, London.
Harvey, D., 1989. The Condition of Postmodernity. Oxford, Blackwell.
Harvey, D., 1993. From space to place and back again. In: Bird, J., Curtis, B., Putnam, T., Robertson, G., Tickner, L. (Eds.), Mapping the Futures. Routledge, London.
Heidegger, M., 1971. Poetry, Language, Thought. Harper and Row, New York.
Holloway, S., 2005. Articulating otherness? White rural residents talk about Gypsy-Travellers. Transactions of the Institute of British Geographers 30 (3), 351–367.
Hopkins, P., Kong, L., Olson, E. (Eds.), 2013. Religion and Place: Landscape, Politics and Piety. Springer, London.
Hubbard, P., 2000. Desire/disgust: mapping the moral contours of heterosexuality. Progress in Human Geography 24 (2), 191–217.
Keith, M., Pile, S. (Eds.), 1993. Place and the Politics of Identity. Routledge, London.
Kirsch, S., 2013. Cultural geography I: materialist turns. Progress in Human Geography 37, 433–441.
Kobayashi, A., Peake, L., 2000. Racism out of place: thoughts on whiteness and an anti-racist geography in the new millennium. Annals of the Association of American Geographers 90, 651–669.
Livingstone, D., 1992. The Geographical Tradition: Episodes in the History of a Contested Enterprise. Oxford, Blackwell.
Massey, D., 1991. The political place of locality studies. Environment and Planning A 23 (2), 267–281.
Massey, D., 1993. Power – geometry and progressive sense of place. In: Bird, J., Curtis, B., Putnam, T., Robertson, G., Tickner, L. (Eds.), Mapping the Futures. Routledge, London.
Massey, D., 1997. A global sense of place. In: Barnes, T., Gregory, D. (Eds.), Reading Human Geography: The Poetics and Politics of Inquiry. Arnold, London.
McKeon, R. (Ed.), 1941. Aristotle: Basic Works. Random House, New York.
Merleau-Ponty, M., 1962. The Phenomenology of Perception. Routledge and Kegan Paul, London.
Relph, E., 1976. Place and Placelessness. Pion, London.
Rose, G., 1993. Feminism & Geography: The Limits of Geographical Knowledge. University of Minnesota Press, Minnesota.
Rose, G., 1997. Situating knowledges: positionality, reflexivities and other tactics. Progress in Human Geography 21 (3), 305–320.

Seamon, D., 1979. A Geography of the Lifeworld: Movement, Rest, and Encounter. St. Martin's, New York.
Sidaway, J.D., 2000. Recontextualising positionality: geographical research and academic fields of power. Antipode 32 (3), 260–270.
Skelton, T., Valentine, G., 1998. Cool Places: Geographies of Youth Cultures. Routledge, New York.
Smith, S., 1989. The Politics of 'Race' and 'Residence'. Polity Press, Cambridge.
Tuan, Y-f, 1971. Geography, phenomenology, and the study of human nature. Canadian Geographer 15, 181–192.
Valentine, G., 1993. (Hetero)sexing space: lesbian perspectives and experiences of everyday spaces. Environment and Planning D: Society and Space 11, 395–413.
Valentine, G., 1997. Angels and devils: moral landscapes of childhood. Environment and Planning D: Society and Space 14 (5), 581–599.
Wylie, J., 2010. Cultural geographies of the future, or looking rosy and feeling blue. Cultural Geographies 17, 211–217.
Yorgason, E., Della D., V., 2009. Editorial: geography, religion, and emerging paradigms: problematizing the dialogue. Social and Cultural Geography 10 (6), 629–637.

Placebo Effect

Fabrizio Benedetti, University of Turin Medical School, and National Institute of Neuroscience, Turin, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The placebo effect is a context effect that is due to the psychosocial context around the therapy and the patient. Many social stimuli, which constitute the ritual of the therapeutic act, can activate a number of mechanisms in the patient's brain. Pain is the condition where most of our knowledge about placebo mechanisms comes from. Four neurotransmitters have been identified in placebo analgesia: endogenous opioids, endocannabinoids, cholecystokinin, and dopamine. Pain is not a special case, and in other conditions, such as Parkinson's disease, placebos induce similar effects.

Definition

The placebo effect, or response, is a psychobiological phenomenon occurring in the patient's brain following the administration of an inert treatment, be it an inert substance or a sham physical treatment such as sham surgery, along with verbal suggestions, or any other cue, of clinical improvement (Benedetti, 2008). It is important to point out that saline solutions or sugar pills will never acquire therapeutic properties. What matters is the psychosocial context that surrounds the inert treatment and the patient. In this sense, to the person conducting clinical trials and to the neuroscientist, the terms 'placebo effect' and 'placebo response' have different meanings. Whereas the former is interested in any improvement that may occur in the group of patients who take the inert treatment, regardless of its origin, the latter is only interested in the improvement that derives from active processes occurring in the patient's brain. In fact, the improvement in patients who are given a placebo can be ascribed to a vast array of factors, such as spontaneous remission of the disease (the so-called natural history), regression to the mean (a statistical phenomenon due to selection biases), patient's and doctor's biases, and unidentified effects of co-interventions. In pragmatic clinical trials, the researcher is interested in the improvement irrespective of its cause, because he or she only needs to establish whether the patients who take the true treatment are better off than those who take the placebo. However, if we are interested in understanding what placebo effects are and how they work, we need to separate them from spontaneous remissions, regression to the mean, biases, and the like.

Therefore, the placebo is not the inert substance alone, but rather its administration within a set of social stimuli that tell the patient that a beneficial treatment is being given (Benedetti, 2008, 2010). It is possible to rule out other phenomena by using the appropriate methodological approach. For example, to rule out spontaneous remission, the placebo group must be compared with a no-treatment group, which gives us information on the natural history of the disease. Likewise, in order to rule out biases, such as those that may occur in subjective symptoms like pain, objective outcome measures must be assessed.

Psychological Mechanisms of the Placebo Response

There is not a single placebo effect, but many, with different mechanisms across different diseases, systems, and therapeutic interventions. Sometimes it is anxiety that is modulated, at some other times reward mechanisms are involved, and in some other circumstances different types of learning, or even genetic variants, may take place in placebo responsiveness.

Most of the research on placebos has focused on expectations as the main factor involved in placebo responsiveness. In the many studies in the literature where expectations are analyzed, the terms 'effects of placebos' and 'effects of expectations' are frequently used interchangeably. In general, expectation is aimed at preparing the body to anticipate an event in order to better cope with it. There are several mechanisms through which expectation of a future event may affect different physiological functions. For example, the expectation of a negative outcome is aimed at anticipating a possible threat, thus increasing anxiety, whereas the expectation of a forthcoming positive outcome may reduce anxiety and/or activate the neuronal networks of reward mechanisms. Anxiety has been found to be reduced after placebo administration in some studies. If one expects a distressing symptom to subside shortly, anxiety tends to decrease (Price et al., 2008; Finniss et al., 2010).

The best evidence that anxiety takes part in placebo responses is shown by the nocebo effect, which is opposite to the placebo effect. In order to induce a nocebo effect, an inert treatment is administered along with negative verbal suggestions of clinical worsening, such as pain increase. Overall, expectations of a negative outcome, such as pain increase, lead to anticipatory anxiety which in turn may result in the amplification of pain (Enck et al., 2008).

Expectations may also induce physiological changes through reward mechanisms. These mechanisms are mediated by specific neuronal circuits linking cognitive, emotional, and motor responses, and are traditionally studied in the context of the pursuit of natural (e.g., food), monetary, and drug rewards. The reward mesolimbic dopaminergic system may be activated in some circumstances when a subject expects clinical improvement after placebo administration, for example, in pain, depression, and Parkinson's disease (Benedetti, 2008; Finniss et al., 2010).

Learning is another mechanism that is central to placebo responsiveness. Subjects who suffer from a painful condition, such as headache, and who regularly consume aspirin, can associate the shape, color, and taste of the pill with pain decrease. After repeated associations, if they are given a sugar pill resembling aspirin in all respects, they will experience pain decrease. Not only can shape, color, and taste of pills be

associated with clinical improvement, but countless other stimuli as well, such as hospitals, diagnostic and therapeutic equipment, and medical personnel features. The mechanism that underlies this effect is conditioning, whereby a conditioned (neutral) stimulus, such as the color and shape of a pill, can become effective in inducing the reduction of a symptom if repeatedly associated with an unconditioned stimulus, that is, the active principle contained in the pill. This type of associative learning may represent the basis of many placebo effects, where the placebo is the conditioned (neutral) stimulus itself. Placebo effects may involve social learning as well, whereby individuals learn from one another by observation (Enck et al., 2008).

There is also some experimental evidence supporting the role of genes in some types of placebo responses, including a genetically controlled serotoninergic modulation of amygdala activity, which is linked to placebo-induced anxiety relief, and a genetically controlled monoaminergic tone, which is related to the degree of placebo responsiveness in major depressive disorder. In fact, some genetic variants have been found that are particularly responsive to placebo treatment in patients suffering from social anxiety disorder and major depressive disorder (Benedetti, 2010).

The Neuroscience of Placebos

The opioid antagonist, naloxone, is capable of blocking placebo analgesia, thus indicating that the endogenous opioid system is involved. This system is a top-down regulatory system extending from cognitive and affective cortical brain regions to the brainstem and spinal cord dorsal horns, with the ability to negatively modulate the incoming nociceptive signals (Basbaum and Fields, 1984). A number of neuroimaging studies have provided important information on the location and timing of endogenous opioid activation (Tracey, 2010). For example, *in vivo* receptor binding with positron emission tomography has shown opioid activation in the pregenual rostral anterior cingulate cortex, insula, nucleus accumbens, and dorsolateral prefrontal cortex (Wager et al., 2007). This activation is abolished by naloxone administration. Data have been provided regarding also the distal part of the antinocicepive system, namely the spinal cord, whereby activity in the dorsal horn is reduced during placebo analgesia (Eippert et al., 2009b).

There are some conditions in which the endogenous opioids are not involved in placebo analgesia, such as when the subjects had previously been exposed to non-opioid drugs. Recent experimental evidence indicates that whenever the opioid system does not take part in the placebo analgesic response, the endocannabinoid system may play a crucial role. In fact, the cannabinoid antagonist, rimonabant, has been found to block non-opioid placebo analgesic responses whereas it has no effect on opioid placebo responses (Benedetti et al., 2011a).

Another neurotransmitter identified in placebo analgesia is cholecystokinin (CCK). In fact, proglumide, a CCK antagonist, potentiates the placebo analgesic response and pentagastrin, a CCK agonist, blocks placebo analgesia, consistent with the anti-opioid action of CCK. Thus, CCK appears to play an inhibitory role in placebo analgesia that is likely to be exerted in the same areas involved in opioid transmission, although no imaging data are at present available. Interestingly, CCK also modulates the counterpart of placebo analgesia – nocebo hyperalgesia – in which an increase in pain occurs following expectations of a negative outcome. In this case, by antagonizing the pronociceptive effect of CCK, proglumide produces the inhibition of the nocebo response (Enck et al., 2008).

Dopamine has been found to be involved in the placebo effect as well. It was first implicated in placebo research in investigations on Parkinson's disease (de la Fuente-Fernandez et al., 2001). It was noted that the motor placebo response in Parkinsonian patients was associated not only with dopamine release in the dorsal striatum, consistent with the role of this structure in motor control, but also in the ventral striatum (nucleus accumbens), which is part of the reward circuit. A placebo-related activation of dopamine in the reward network has also been demonstrated in pain.

While for placebo analgesia a complex and varied neurochemical picture has been gradually outlined, the running portrayal of placebo effects in Parkinson's disease is still quite straightforward, involving only dopamine. In the simulation of a classic clinical trial, patients in the placebo arm, aware of a 50% chance of receiving the dopamine agonist apomorphine, exhibited an extracellular dopamine increase of more than 200%, comparable to the response to amphetamine in healthy subjects (de la Fuente-Fernandez et al., 2001). Placebo administration has also been found to decrease the firing rate of the subthalamic nucleus neurons and change their pattern of activity from bursting to nonbursting. In addition, placebos decrease the firing rate in the substantia nigra pars reticulata and increase neuronal activity in the ventral anterior and ventral anterior lateral thalamus, suggesting that these neuronal changes are due to dopamine release (Benedetti, 2008, 2010).

Little is known about the underlying mechanisms of the placebo response in depression. The main problem is represented by the fact that, unlike single-dose trials of an intervention, such as oral or intravenous analgesia or anti-Parkinson acute therapy studies, antidepressants do not work acutely, requiring on average a minimum of two to three weeks to see clinical effects. Therefore, investigating placebo effects in depression is difficult from both ethical and methodological viewpoints. For example, if one wants to compare a placebo group with a no-treatment group to rule out spontaneous remission, this requires that some patients need not to be treated for a long period of time. This is one of the main reasons why depression, albeit an interesting model for placebo studies, has not been investigated in detail so far. Changes in brain glucose metabolism were measured by means of positron emission tomography in patients with unipolar depression who were treated with either placebo or fluoxetine for 6 weeks, and common and unique responses were described. Both placebo and fluoxetine induced regional metabolic increases in the prefrontal, anterior cingulate, premotor, parietal, posterior insula, and posterior cingulate, and metabolic decreases in the subgenual, para-hippocampus, and thalamus, with larger responses to fluoxetine compared to placebo. However, fluoxetine responses were associated with additional subcortical and limbic changes in the brainstem, striatum, anterior insula, and hippocampus. By contrast, there were no regional changes unique to placebo at 6 weeks.

Interestingly, there were unique ventral striatal (nucleus accumbens) and orbital frontal changes in both placebo and drug responders at 1 week of treatment, that is, well before clinical benefit, which suggests that they were related to expectation and anticipation of the clinical benefit. Such changes were seen neither in the eventual drug nonresponders nor at 6 weeks when the antidepressant response was well established, consistent with an expectation pattern of response (Mayberg et al., 2002). It should be pointed out that the same involvement of the ventral striatum (nucleus accumbens) was found after placebo administration in Parkinson's disease and in pain, which emphasizes once again the important role of reward mechanisms in some placebo responses.

Unconscious conditioning seems to be the sole mechanism involved in the generation of the placebo response in the case of those physiological functions that are only subconsciously controlled, like those involving the immune or the endocrine systems. Thus, placebo modulation of growth hormone (GH) and cortisol secretion could be obtained after conditioning with sumatriptan (a selective 5-$HT_{1B/1D}$ receptor agonist stimulating GH and inhibiting cortisol secretion). However, it was impossible to induce the same modulation by suggestion alone, in the absence of conditioning. It was also not possible to reverse the modulation achieved by conditioning by inducing in the subjects the opposite expectation by verbal instruction. Notably, it was indeed possible to exert this counteracting effect with the induction of opposite expectations in the case of placebo analgesia/hyperalgesia, where cognitive conscious processes play an important role (Benedetti et al., 2003).

Immune functions have also been repeatedly conditioned, both in animals and in humans. For instance, in multiple sclerosis patients, decrements in peripheral leukocyte counts induced by cyclophosphamide could be conditioned by pairing the drug with a strongly flavored beverage, and more recently, in healthy subjects, a similar conditioning with cyclosporine A produced immunosuppression after placebo, as assessed by depressed mRNA expression of IL-2 and interferon-gamma, and lymphocyte proliferation, mediated (at least in rodents) by calcineurin inhibition. Recently, much work has been devoted to the identification of the neural structures involved centrally (amygdala, insular cortex, hypothalamus) and peripherally (cholinergic and catecholaminergic output in the autonomic nervous system) in conditioned immunosuppression (Enck et al., 2008). Reflex control of immunity has been proposed to be part of a neocortical–immune axis, raising the interesting possibility of a role for cognitive factors such as expectations and beliefs in placebo immunological responses.

The Importance of the Prefrontal Lobes in Placebo Responsiveness

A functional disconnection of the prefrontal lobes from the rest of the brain is associated with a loss of placebo responsiveness. This has been investigated in patients suffering from Alzheimer's disease, in which the placebo component of an analgesic therapy is correlated with both cognitive status and functional connectivity among different brain regions (Benedetti et al., 2006). In fact, Alzheimer patients with reduced cognitive scores show reduced placebo component of an analgesic treatment and, moreover, the disruption of the placebo component occurs just when reduced connectivity of the prefrontal lobes with the rest of the brain is present. The loss of these placebo-related mechanisms reduces the overall effectiveness of the treatment, and indeed a dose increase is necessary to make up for this loss in order to produce adequate analgesia.

These data in Alzheimer's patients are supported by the correlation between white matter integrity and the magnitude of placebo responses in healthy subjects. The individual placebo analgesic effect has been found to be correlated with white matter integrity indexed by fractional anisotropy, particularly in the right dorsolateral prefrontal cortex, left rostral anterior cingulate cortex, and the periaqueductal gray. Probabilistic tractography seeded in these regions showed that stronger placebo analgesic responses were associated with increased mean fractional anisotropy values within white matter tracts connecting the periaqueductal gray with the rostral anterior cingulate cortex and the dorsolateral prefrontal cortex (Stein et al., 2012).

The important role of the prefrontal lobes in the placebo effect has been supported by the deactivation of the prefrontal cortex in the experimental setting. By combining naloxone administration with functional magnetic resonance imaging, naloxone has been found to reduce placebo effects as well as placebo-induced responses in pain-modulatory cortical structures, such as the dorsolateral prefrontal cortex and the rostral anterior cingulate cortex. In a brainstem-specific analysis, a similar naloxone modulation of placebo-induced responses in key structures of the descending pain control system, including the hypothalamus, the periaqueductal gray, and the rostral ventromedial medulla, was also found. Naloxone also abolished the increase in coupling between the rostral anterior cingulate cortex and the periaqueductal gray that was induced by the placebo (Eippert et al., 2009a). Therefore, as it occurs for prefrontal degeneration in Alzheimer's disease, placebo analgesic responses are disrupted by the pharmacological blockade of prefrontal opioidergic functioning in the experimental setting.

Prefrontal degeneration in Alzheimer's disease, reduced integrity of prefrontal white matter, and pharmacological blockade of prefrontal opioidergic transmission are not the only conditions in which placebo responses are disrupted. Repetitive transcranial magnetic stimulation has been used to inactivate the prefrontal cortex, particularly the dorsolateral prefrontal cortex, during a placebo analgesic response, and this stimulation of the prefrontal areas blocks placebo analgesia (Krummenacher et al., 2010). Therefore, the inactivation of prefrontal regions by transcranial magnetic stimulation has the same effects as those induced by prefrontal degeneration in Alzheimer's disease, reduced integrity of prefrontal white matter, and opioidergic blockade of prefrontal areas. On the basis of all of these studies, a normal functioning of prefrontal areas, as well as of the descending pain modulating network with which it is connected, appears to be critical for placebo responsiveness. In the presence of a loss of prefrontal control, we also witness a loss of placebo response.

Delivering Treatments Unknown to the Patient

In recent years, a radically different approach to the analysis of placebo effects has been implemented, in which placebo effects are assessed without placebo groups. This experimental approach consists in eliminating the placebo component and maintaining the specific effects of the treatment. In order to eliminate the placebo component, the patient is made completely unaware that a medical therapy is being carried out. Then hidden therapies are compared with open ones. The difference between the hidden and open treatment represents the placebo component, even though no placebo has been given. The importance of this approach resides in the fact that open therapies are expected whereas hidden therapies are unexpected. Therefore, the study of the open–hidden (expected–unexpected) paradigm gives us information on the role of expectations in the therapeutic outcome (Benedetti et al., 2011b).

It is possible to perform a hidden infusion of a drug by means of a computer-controlled infusion pump that is preprogrammed to deliver the drug at the desired time. The essential point is that the patient does not know that any drug is being injected. This hidden procedure is relatively easy to carry out in the postoperative phase, in which the computer-controlled infusion pump can deliver a painkiller automatically, without any doctor or nurse in the room, and with the patient being completely unaware that an analgesic treatment has been started. In general, the analgesic dose needed to reduce the pain by 50% is much higher with hidden infusions than with open ones, indicating that a hidden administration is less effective than an open one. Likewise, the time course of postsurgical pain is significantly different between open and hidden injections. In fact, during the first hour after the injection, pain ratings are much higher with a hidden injection than with an open one. Pain is not a special case, and the same results have been obtained for anxiety and Parkinson's disease (Benedetti et al., 2011b).

From the Clinical Setting to Physical Performance

As for the clinic, also in the sport world there is a zone where placebos (and nocebos) can exert their influence. Here too, chemicals such as vitamins, ergogenic aids, or diet supplements are handed out, or physical treatments and manipulations of different kinds are delivered, and expectations about their effects are set in motion in the athlete's brain. In spite of very different experimental conditions, ranging from short anaerobic sprints to long aerobic endurance cycling, and across many different outcome measures, such as mean power output, time, speed, weight lifted, or perceived exertion, all of the data indicate that athletes' expectations are important factors in physical performance, to be taken into account in training strategies (Pollo et al., 2011).

Contrary to other areas of placebo research, in physical performance knowledge about underlying mechanisms is still scant. Apart from a role for endogenous opioids in placebo pain endurance, not much is known at present. In many studies, athletes are asked to perform at their limit, in an all-out effort. Placebos apparently act by pushing this limit forward, possibly impacting on a central governor of fatigue which, although not identified, has been proposed as a brain center integrating peripheral signals (such as heart and respiratory rate, lactate, carbohydrate availability, and mechanical strain) and central control processes, so as to continuously regulate exercise performance that avoids reaching maximal physiological capacity. This would provide protection against damage on one side and constant availability of a reserve capacity on the other. By altering expectations, placebos could then represent a psychological means to signal the central governor to release the brake, allowing an increase in performance.

See also: Cerebral Cortex; Context Dependent Memory; Empathy: A Social Neuroscience Perspective; Functional Brain Imaging of Language Processes; Prefrontal Cortex; Social Neuroscience of the Twenty-First Century.

Bibliography

Basbaum, A.I., Fields, H.L., 1984. Endogenous pain control systems: brainstem spinal pathways and endorphin circuitry. Annual Review of Neuroscience 7, 309–338.

Benedetti, F., 2008. Placebo Effects: Understanding the Mechanisms in Health and Disease. Oxford University Press, Oxford.

Benedetti, F., 2010. The Patient's Brain: the Neuroscience behind the Doctor-Patient Relationship. Oxford University Press, Oxford.

Benedetti, F., Amanzio, M., Rosato, R., Blanchard, C., 2011a. Non-opioid placebo analgesia in mediated by CB1 cannabinoid receptors. Nature Medicine 17, 1228–1230.

Benedetti, F., Arduino, C., Costa, S., et al., 2006. Loss of expectation-related mechanisms in Alzheimer's disease makes analgesic therapies less effective. Pain 121, 133–144.

Benedetti, F., Carlino, E., Pollo, A., 2011b. The hidden administration of drugs. Clinical Pharmacology & Therapeutics 90, 651–661.

Benedetti, F., Pollo, A., Lopiano, L., et al., 2003. Conscious expectation and unconscious conditioning in analgesic, motor and hormonal placebo/nocebo responses. Journal of Neuroscience 23, 4315–4323.

de la Fuente-Fernandez, R., et al., 2001. Expectation and dopamine release: mechanisms of the placebo effect in Parkinson's disease. Science 293, 1164–1166.

Eippert, F., Bingel, U., Schoell, E.D., et al., 2009a. Activation of the opioidergic descending pain control system underlies placebo analgesia. Neuron 63, 533–543.

Eippert, F., Finsterbusch, J., Bingel, U., Büchel, C., 2009b. Direct evidence for spinal cord involvement in placebo analgesia. Science 326, 404.

Enck, P., Benedetti, F., Schedlowski, M., 2008. New insights into the placebo and nocebo responses. Neuron 59, 195–206.

Finniss, D.G., Kaptchuk, T.J., Miller, F., Benedetti, F., 2010. Biological, clinical, and ethical advances of placebo effects. Lancet 375, 686–695.

Krummenacher, P., Candia, V., Folkers, G., Schedlowski, M., Schönbächler, G., 2010. Prefrontal cortex modulates placebo analgesia. Pain 148, 368–374.

Mayberg, H.S., Silva, J.A., Brannan, S.K., et al., 2002. The functional neuroanatomy of the placebo effect. American Journal of Psychiatry 159, 728–737.

Pollo, A., Carlino, E., Benedetti, F., 2011. Placebo mechanisms across different conditions: from the clinical setting to physical performance. Philosophical Transactions of the Royal Society B: Biological Sciences 366, 1790–1798.

Price, D.D., Finniss, D.G., Benedetti, F., 2008. A comprehensive review of the placebo effect: recent advances and current thought. Annual Review of Psychology 59, 565–590.

Stein, N., Sprenger, C., Scholz, J., Wiech, K., Bingel, U., 2012. White matter integrity of the descending pain modulatory system is associated with interindividual differences in placebo analgesia. Pain 153, 2210–2217.

Tracey, I., 2010. Getting the pain you expect: mechanisms of placebo, nocebo and reappraisal effects in humans. Nature Medicine 16, 1277–1283.

Wager, T.D., Scott, D.J., Zubieta, J.K., 2007. Placebo effects on human μ–opioid activity during pain. Proceedings of the National Academy of Science USA 104, 11056–11061.

Relevant Websites

www.placebo-vw.unito.it – Understanding the Mechanisms of the Placebo Response. A Research Project Sponsored by the Volkswagen Foundation.

http://programinplacebostudies.org/ – Program in Placebo Studies and the Therapeutic Encounter at Harvard Medical School. Hosted at Beth Israel Deaconess Medical Center.

Placebo Studies (Double-Blind Studies)

Harald Walach, European University Viadrina, Frankfurt (Oder), Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Shapiro, volume 17, pp. 11455–11460, © 2001, Elsevier Ltd.

Abstract

Placebo-controlled, double-blind studies are the most common research method to evaluate new, pharmacological interventions in medicine. It is important to separate this usage of the concept from the employment of substances and interventions in medicine whose causal-mechanistic effects are unknown, and yet are applied to patients' benefit. This therapeutic usage of placebo-effects has recently been termed 'the meaning response' or 'the self-healing response.' While clinical, placebo-controlled studies are designed to minimize such effects, therapeutic systems of all kinds intend to maximize them. This article reviews some historical and empirical data pertinent to these discussions, and some problems are highlighted.

Placebo-controlled, double-blind studies are the most common research method to evaluate new pharmacological interventions in medicine. Demanded by regulatory agencies all around the world, they have become standard practice. The history of the placebo-controlled trial is comparatively short: inaugurated as the method of choice by the 'Cornell Conferences on Therapy' in 1946, they became universally accepted in the 1950s. Conceptually speaking, the use of placebos in research is necessary if patients and doctors are to be kept blind to the treatment provided. Thus, the placebo as an inert substance given to patients in clinical trials as a control procedure is a methodological requirement to allow for an unbiased assessment of the effect of a pharmacological intervention. It is important to separate – historically, conceptually, and practically – this usage of the concept from the employment of substances and interventions in medicine and elsewhere whose causal-mechanistic effects are either unknown, unlikely, or too weak, and yet are and have been applied to patients' benefit. This effect probably occurs because human beings are not mechanistic biological automatons, but meaning-making, complex systems with an individual history, with expectations and elaborate cognitive processes that interfere with any type of intervention to create a very complex interaction of effects. This therapeutic usage of placebo-effects has recently been termed 'the meaning response' or 'the self-healing response.' While clinical, placebo-controlled studies are designed to minimize such effects, therapeutic systems of all kinds intend to maximize them. In this article, we will review some historical and empirical data that are pertinent to these discussions, and highlight some of the problems.

Some Historical Notes

The term 'placebo' is Latin and means 'I shall please.' Its usage is derived from Psalm 116.8, which in its medieval Latin version reads "placebo Domino in regione vivorum – I shall please the Lord in the land of the living" (Shapiro and Shapiro, 1997). This psalm was part of the Christian funeral rites and was typically sung at the deathbed. In the late middle ages it became customary in England to have those rites sung by paid mourners. From this tradition, the notion of a placebo as a surrogate for the real thing, in that case the real prayer, developed and was imported into medicine, for the first time in 1785 in a medical dictionary, to signify an intervention that is used to please the patient or to substitute for the real intervention (Shapiro and Shapiro, 1997). Typically this was done either if no good treatment was known, or if the treatment was deemed too invasive, as was often the case with interventions of earlier days, such as the poisonous mercury treatments that were the only real bactericidal treatments available before the advent of antibiotics, colchicine treatment for gout, or phlebotomy as a treatment for everything of an unknown cause. Often such 'placebos' were prescribed by doctors to patients when they saw no reason to intervene but the patient wanted to be treated. So, in order to 'please' the patient, the doctor gave some cordial that the patient might deem effective but which in fact was harmless, comparable to vitamin prescriptions of our times (Kaptchuk, 1998). Some authors even call all treatments before the advent of modern day pharmacology placebo-treatments (Shapiro and Shapiro, 1997). Clearly, such a statement is predicated on the conviction that only efficacious pharmacological treatments can be 'real' treatments. We will shortly scrutinize whether this presupposition is true. Ironically, it is the research around the 'placebo' that undermines this presupposition, as it turns out that a lot of treatments that are 'placebos' from a pharmacological point of view can be very powerful because they harness the 'placebo-effect,' shorthand for the complex self-healing reaction the organism seems to be capable of in response to a complex intervention that instills meaning.

Definitions

Depending on whether a definition of a 'placebo' was attempted from the perspective of research or clinical practice, we have to deal with two completely different families of definitions. The older, widely used definition is from Shapiro and Morris (1978: p. 440):

> A placebo is defined as any therapy, or that component of any therapy, that is deliberately used for its nonspecific, psychologic, or psychophysiologic effect, or that is used for its presumed specific

> effect on a patient, symptom, or illness, but which, unknown to patent *(sic)* and therapist, is without specific activity for the condition being treated… The Placebo effect is defined as the nonspecific, psychologic or psychophysiologic effect produced by placebos.

Note that with this definition a placebo-effect is implicitly defined vis-à-vis a specific effect that is implicitly assumed to be more worthwhile. From such a vantage point any treatment that has no specific activity, or no known specific activity, or perhaps a specific activity, but not for the condition being treated, is considered a placebo. We see immediately that such a definition is problematic for various reasons. First, what is specific for treating a particular disease is very much dependent on our knowledge, our perspective, and the context. For instance, thalidomide was used for its 'specific' tranquillizing and sleep-inducing effect during pregnancy. Only later it became known that it also had very specific teratogenic effects, namely malformations of limbs – arms, fingers, legs – that were being formed during the weeks of gestation when the substance was typically used. Today, it has completely different 'specific' applications as a very specialized medication for leprosy. A second example – bloodletting, phlebotomy, was used widely in the old medicine for its purported 'specific' efficacy against a surplus of vital substances such as blood, bile, or phlegma – mucus, which were thought to be causative of diseases in the ancient medical model of Galenos, Celsus, and others. It probably helped as often as it did damage. From a modern point of view most of these indications were likely 'nonspecific.' However, modern research shows that in specific cases, for instance of high blood pressure, it can be a very specific and effective treatment, but for completely different reasons (Houschyar et al., 2012).

Thus, what counts as 'specific' is a function not only of the 'true' nature of any intervention but also of the state of knowledge or the context of usage. This can also be nicely illustrated by a singular trial that was conducted at a time when 'informed consent,' that is, the willing and knowing affirmation of a patient to be part of a trial, was not ethically required everywhere, and that was conducted to study such an effect (Bergmann et al., 1994). In this trial, 49 in-patients who suffered from pain due to cancer were randomized to either receive the information that they were part of a trial or not to be informed of this fact. Half of them received either naproxen, a pain medication that had been independently shown to have specific efficacy against pain, or placebo. Thus, a quarter of the patients received naproxen with the information that they were part of a trial, and a quarter of patients received naproxen with the assumption that it was their normal bedside medication. Another quarter of the patients received placebo within the context of a trial, expecting either naproxen or placebo, while a quarter of the patients received placebo, although they expected their active bedside medication. The interesting finding here is that naproxen given under the conditions of a trial was twice as effective as the same substance given under the condition of a bedside medication. Even more interesting is the fact that the placebo that patients received under trial conditions, that is, blind and unknowing, but expecting either naproxen or placebo, was more effective than naproxen, the specifically active substance, received under bedside conditions. This example teaches, among many others reviewed in Walach (2011) and Walach et al. (2006), that what counts as a specific effect can be highly variable.

Thus, definitions that start from the implicit, positivist assumption that there is such a thing as a 'true' 'specific' effect that science has to uncover vis-à-vis a more or less stable, nonspecific effect of placebo are hampered right from the beginning. The definition by Grünbaum (1985) tried to solve the conundrum by introducing the differentiation between 'intended' effects and 'unintended' or 'incidental' effects, the latter being the placebo-effects observed in trials. But a closer scrutiny reveals quickly that, again, this attempt at defining the placebo effect is heavily dependent on theory. What counts as 'intended' today, may be 'unintended' tomorrow, and vice versa, as the example of thalidomide above shows.

This problem occurs because of the underlying assumptions of most authors that are related to the dominant research model within behavioral science of the 1970s that is still rampant in medical research today. This assumption assumes that any intervention will have to depend on a 'true' specific effect that is 'out there' independent of context, practitioner, and historical and social situations. And this 'true specific effect' can be made visible through employing, as a methodological trick, a placebo-control that implicitly controls for all other unwanted effects such as expectation; hope; relaxation; reduction of anxiety; statistical regression to the mean; natural course of the disease, including natural healing and improvement; and all other potential psychological effects. Ironically, it was the research around placebos itself that demonstrated that this assumption is at best a very crude abstraction that can be used for registration purposes, but is not an adequate picture of reality.

For this reason, in 1995 a workshop was convened at the National Institutes of Health where specialists in placebo-research such as Dan Moerman, Howard Brody, Irving Kirsch, and others discussed the potential and scope of a new definition. This new definition developed from the realization that placebo-effects are highly individual, and more importantly, an active response of an individual agent to a complex therapeutic stimulus. Thus the placebo was defined as a 'meaning response,' that is, a response of an organism that is a complex cognitive system capable of active engagement and interaction with the environment, to a therapeutic intervention that creates meaning for that individual or fails to do so (Moerman and Jonas, 2002). For instance, if someone has a phobic response to syringes, then an intervention that requires puncturing the skin, such as injections, infusions, or even acupuncture, might create an aversive reaction because syringes are imbued with negative meaning, while for someone who entertains the concept that injections are 'more potent' than oral prescriptions, or that what is potent needs to be felt, is likely to find the same intervention more meaningful. Thus, this new definition is explicitly semiotic and individual. 'Semiotic' hereby refers to the long tradition of semiotic studies that was introduced by Charles Sanders Peirce in the United States and Ferdinand de Saussure in Europe, which aims to understand how meaning is generated by various processes of signification, not only within language systems but also within extralinguistic modes, such as actions or cultural practices. This definition is clearly most

relevant for all therapeutic applications because it recognizes the individual nature of the response of a person to any type of intervention.

The History of Placebo-Controls

Blinding of researchers, patients, and providers is necessary if we want to know about the isolated effect of a particular intervention. This reasoning can be traced back to the middle ages, where Roger Bacon (1214/15–92) was one of the first to advocate 'experimentum – experience, experiment' as a systematic way of gaining knowledge about nature, also in medicine. 'Experiment' is derived from the Latin verb '*experiri* – to seek an experience, to travel,' and an '*expertus* – expert' is one who has ample experience. In those times an experiment was typically a replicable observation, mostly with optical instruments. It was the later Bacon, Francis Bacon, Chancellor to the Queen (1561–1626), who defined an experiment as a replicable manipulation, a question put to nature by a human action or intervention, whose result is nature's answer. It was this experimental thinking, understood as systematic, replicable manipulations of isolated systems that paved the way for the rise of the scientific model, first in physics and chemistry, and finally also in medicine.

Controlled experiments were rare in medicine until the middle of the nineteenth century. A pivotal experiment was the sequence of studies conducted by the French Academy of Science on the theories of Franz Anton Mesmer in Paris in 1784 (Bailly, 1784; Crabtree, 1993). Mesmer was a German doctor who claimed that a special type of magnetism exists, animal magnetism, which could be used to heal patients as well as to influence healthy people. To prove the case the Academy appointed a commission that studied a prominent student of Mesmer's, the dean of the medical school, as he 'magnetized' a 'hysteric' patient by stroking over the patient at a distance. The patient would fall into all sorts of fits, but only if she could actually see the magnetist and his movements. When the commissioners placed a curtain between the magnetist and his patients the effect was gone. This is the first known historical instance of 'blinding,' being represented here by the curtain (Kaptchuk, 1998). After this experience and the further history of hypnotism and magnetism, the power of human relationships, suggestions, and expectations in both provoking and curing symptoms became common knowledge in the medical community. Therefore, it was logical to start thinking about excluding this powerful agent if one wanted to study the 'raw' effect of interventions, bare of any human, psychological, or interactional factors. Blinding was the answer.

Thus with the rise of the pharmacological model in medicine it became ever more important to exclude these human factors of suggestion, relationship, expectation, and hope for cure. How can we blind patients and doctors when researching medical interventions? Only by dispensing medications or performing sham actions that are similar in each respect to the 'true' medication except for the important ingredients. This was the birth of the medical placebo as a control procedure in research. Note that here a therapeutic tradition of handing out comparatively harmless substances specifically for the psychological effects they might produce met the experimental research tradition that required such substances to instantiate control.

Historically, the first researchers that actually used placebos for research purposes were French homeopaths who tested homeopathic medicines, that is, medications that were produced through serial agitations and dilutions according to the teachings of the German doctor and pharmacologist Samuel Hahnemann, against identical placebos in 1832, and German homeopaths who carried out similar experiments in Nuremberg in 1834 (Stolberg, 1996). The first blinded experiment in psychology was conducted by Peirce – the introducer of semiotics – and Jastrow in 1884 in Harvard (Peirce and Jastrow, 1884). In this experiment, volunteers were blindfolded and had to distinguish the smallest difference in weight sensation they could actually perceive. Medicine started to gradually improve on these developments. There were various developments that came together:

In 1935 Sir Ronald Fisher published the first edition of his book 'The Design of Experiments' (Fisher, 1971(orig. 1935)). Here, Fisher, originally an agricultural researcher, introduced the crucial concepts of randomization, first used by John E. Coover in 1917 (Shapiro and Shapiro, 1997), and his statistical method of a frequentist approach to hypothesis testing that allowed for statistical inference and quantifying the probability of making a mistake when making an inference, given the same type of data in another experiment. Parallel developments ensued in the United States and in Germany, making it difficult to decide who has to be given precedence. In Germany, the clinical pharmacologist Paul Martini of Bonn University started to advocate double-blind testing procedures in his 1932 textbook 'Methodology of Therapeutic Trials' (Martini, 1932), which is, of course, based on earlier experiments and insights. In the United States in 1930, Torald Sollmann, head of the AMA's 'Council on Pharmacy and Chemistry,' advocated clinical experimentation in accordance with other scientific disciplines, using the terminology of 'comparative method,' 'double-blind procedure,' and referring to the 'blind test' (Shapiro and Shapiro, 1997).

In 1932 Harry Gold tested xanthine against placebo (lactose) in a single-blind trial against cardiac pain, a milestone, and in 1947 the first individually randomized and controlled clinical trial was carried out: the joint British-US streptomycin trial in tuberculosis patients published in 1948 by Sir Bradford Hill (Doll et al., 1999). It was blinded in that the assessors who judged the X-rays of the patients were unaware of which patient's X-ray they were rating. However, the control patients did not receive placebo, but bed rest only (Shapiro and Shapiro, 1997).

In the wake of these experiences a series of consensus conferences, known as the 'Cornell Conferences on Therapy' (Conferences on Therapy, 1946, 1954) decreed that research in medicine should test new substances against placebos in blinded, randomized studies as of 1946. Randomization was understood as a further essential ingredient in clinical trial methodology because it was the only way to ethically provide treatment in a blinded, placebo-controlled trial. By randomization we mean that the allocation to treatment groups is made by chance. 'Chance' or 'random choice' was performed by using a table of random numbers and allocating, for example, the next even number to treatment and the next uneven number to

control, or by throwing dice. Today, it is performed using computer algorithms, ideally in a way that does not allow the researcher or the recruiting physician to know which group the next patient will be in. This is called 'allocation concealment,' another type of blinding – in that case a blinding as to the outcome of the randomization procedure. If this allocation concealment or allocation blinding is not performed properly, then there might be a danger of bias, if, for instance, the intake-physician has a tendency to allocate more seriously ill patients to the 'real' treatment, especially if the study cannot be completely blinded as to the treatment arms. Only if randomization is performed would everyone have an equal chance. Methodologically, randomization was the only known method that could distribute all known and unknown confounders equally across groups, making unbiased statistical inference possible as advocated by Sir Ronald Fisher. Today there are other methods, such as minimization procedures that use regression algorithms to balance patients, that can in fact be more effective than randomization, but randomization still is the preferred method to avoid bias and create groups within clinical studies, especially within placebo-controlled double-blind studies.

Regulatory agencies, such as the Federal Drug Agency (FDA), the German Federal Institute of Drugs (Bundesinstitut für Arzneimittel, BfArm), and in more recent times, the European Medicines Agency, have now all adopted placebo-controlled trials as the gold standard for the approval of new pharmaceutical agents, except where a good standard already exists. Here, active comparator studies can also be conducted, comparing a new agent to an already approved standard for noninferiority, equality, or superiority. Nowadays, such placebo-controlled clinical trials are elaborate rituals that are conducted, in the majority of cases, by industrial stakeholders and pharmaceutical companies that seek approval for new drugs they have patented. Because such trials are very expensive to conduct, regulated as they are by legal frameworks such as the guideline of Good Clinical Practice, the public sector often conducts trials only some time after new substances have been licensed to market, are in broad usage, and when doubts about their efficacy or safety are being raised. A good and classical example from psychotherapy research was the study of antidepressant treatment by National Institute of Mental Health that compared a well-known and licensed pharmaceutical agent, the tricyclic imipramine, with placebo and two types of psychotherapy (Elkin et al., 1989). The most impressive result was the lack of significant differences and the comparative equality of outcomes across groups. This was mainly due to the fact that the control treatment, placebo, was quite effective.

New pharmacological agents are only required to show two independently significant results in double-blind, placebo-controlled studies for regulatory purposes to gain approval. The number of negative or undecided studies is not weighted as evidence against efficacy, although all material has to be submitted to the regulator.

This whole situation has led one astute observer and insider to conclude that "Randomized trials were introduced in order to protect us from the many useless treatments on the market, but oddly enough, they have given the ultimate power of knowledge production to big pharma that now use them for getting approval for treatments of little or no value and which are often too harmful" (Gøtzsche, 2013: p. 59).

Thus the dialectics of history have led to an interesting situation: What was originally introduced as an instrument of control, enlightenment, and insight, has become an instrument of power and economic dominance. What was conceptualized as an instrument of revealing the truth has become a source of confusion. This is related to the neglected conceptual assumptions that underpin the double-blind research methods.

Tacit Presuppositions, Conundrums, and Critical Aspects

Placebo-effects in clinical trials are a mix of – at least – the following components:

- Measurement artifacts
- Statistical regression to the mean
- Natural course and fluctuations of the disease
- Concomitant treatments
- Other factors of the therapeutic setting, such as generic lifestyle advice
- Genuine psychological therapeutic effects that represent a patient's response to being treated within a trial: expectation, hope, relaxation, a reduction in anxiety, gratitude for being well cared for, conditioning effects of receiving treatment that represent a patient's learning and history with the medical system, among others.

While researchers want to discover the pharmaceutical mechanisms, patients want to know about the general chance of being healed or improved by a treatment, irrespective of the component or method. Patients do not care so much whether a treatment makes use of pharmaceutical or psychological mechanisms to improve their state, as long as it happens safely and reliably (Segar, 2012). Placebos can actually be clinically effective even when patients know that they are receiving placebo. This is elegantly demonstrated by two recent studies. They show that patients with irritable bowel syndrome and migraine patients during an attack can improve clinically and statistically significantly when given open placebo, that is, even when they know they are receiving placebo (Kam-Hansen et al., 2014; Kaptchuk et al., 2010). Such a finding calls into question the implicit assumption that only pharmacological effects whose mechanisms are known can be beneficial and should be used. In fact, purchasers have meanwhile adopted a pragmatic stance that is interested only in effectiveness. Generic clinical effectiveness can also be demonstrated by comparing a treatment against a standard treatment in an open comparative effectiveness trial. Such trials have been called for recently and are part of the new national agenda in the United States (Tinetti and Studenski, 2011).

The assumption underlying the placebo-controlled trial is the paradigm, in the sense of the historian and theoretician of science Thomas Kuhn, that the human organism can be viewed as a passive biological automaton. This was originally conceived by René Descartes in the seventeenth century in his book 'Traité de'l homme – Treatise about Man' (Descartes, 2003(orig. 1664)). Here Descartes postulated that all

animals, including the human organism, can be best understood if they are seen as biological machines, driven by forces and operating on principles that can be understood by scientific analysis and experimentation. Following this lead, the German physician Rudolf Virchow paved the way for the mechanistic turn within medicine in the 1850s. This abstraction was extremely successful in understanding physiology and pathology, coming to grips with acute and epidemic diseases, improving surgery and emergency medicine, and for all acute types of treatments.

Thomas Kuhn (1955) observed that such paradigms are operative for a large group of scientists by providing them with standard experiments, methods, and expectations. In that sense, the double-blind, placebo-controlled trial is not only a method but also a paradigm that makes the stipulation that only specific – pharmacological or psychological – effects are worthwhile and scientifically interesting. Placebo research itself is gradually calling this view into question. This research demonstrates that nonspecific psychological effects of treatments – the 'meaning response,' or the 'self-healing response' (Walach, 2013) – can be very powerful, are associated with well-known causal mechanisms in the brain and thus have a clear physiological basis, and can be therapeutically as strong as the effect of 'real' treatments (Meissner et al., 2011; Colloca et al., 2013). It points to a new research paradigm that understands human beings not as biological automatons, but as complex systems that have agency, can make decisions, respond to complex signs, generate meaning, and do not react to causes only. The complexity of the system and its cognitive processes precludes a clear prediction of a potential reaction to a pharmacological or to any stimulus, at least in the individual case and in the practical situation. And because this is so, we find many conundrums in the research literature.

The Paradox of Small Specific and Strong Generic Effects: The Example of Antidepressants

Precisely because most of the regulatory trials that have led to the licensing of antidepressant agents have been conducted by the industry, about a third of the trials have remained unpublished, most of them negative. Thus a powerful myth of specific, scientifically vindicated efficacy could build up, fostered by the publicity created through early research results. This was so because the pharmaceutical companies are the owners of their data. Although they also have to submit negative studies to the regulator, they are not obliged to publish them. But the regulator requires only two positive studies. Thus, theoretically speaking, a substance that has two studies in its favor might have as many studies not showing superiority over placebo and it would still be licensed. In fact, only 48% of all placebo-controlled studies of anxiolytic or antidepressant treatments submitted to the FDA have shown independently significant results (Khan et al., 2002). This situation allows a large marketing machinery to trumpet out overly optimistic messages about new treatments, creating huge expectations in the public and in practitioners, thus creating large therapeutic effects with only comparatively small specific effects. For instance, the 'true' effect size of antidepressant treatments, estimated including all unpublished negative studies, is $d = 0.32$, that is, only a third of a standard deviation better than placebo (Turner et al., 2008), and yet the substances are highly effective in practice with overall effect-sizes from pre- to post-treatment of 2.5 standard deviations (Rief et al., 2009), precisely because most of the treatment effect is due to a strong placebo-response (Kirsch et al., 2008).

The Efficacy Paradox

One of the tacit presuppositions of the double-blind paradigm is that the 'noise' that is being controlled for in placebo-groups is similar to statistical error variance, and hence uncorrelated and unsystematic, averaging out across treatments. This has been shown to be wrong. Treatment effects in placebo and treatment groups across various types of placebo-controlled trials of a duration of more than 12 weeks are highly correlated at $r = 0.78$ (Walach et al., 2005). The conventional wisdom, however, is that it is sufficient to look at the difference between the control treatment – placebo – and the real treatment, assuming that the variations in the control treatment can be neglected. But what if it cannot be neglected? Then we may find what has been termed the 'efficacy-paradox' (Walach, 2001). This denotes a situation where a nonspecific treatment or a treatment with a very small or negligible specific efficacy can be more effective than a treatment with a very strong specific efficacy, precisely because the nonspecific or placebo-component of the treatment is so strong (**Figure 1**). This is likely to be the case in all situations where patients and providers are very satisfied because they experience strong therapeutic effects. And yet, when probed for specific efficacy in placebo-controlled clinical

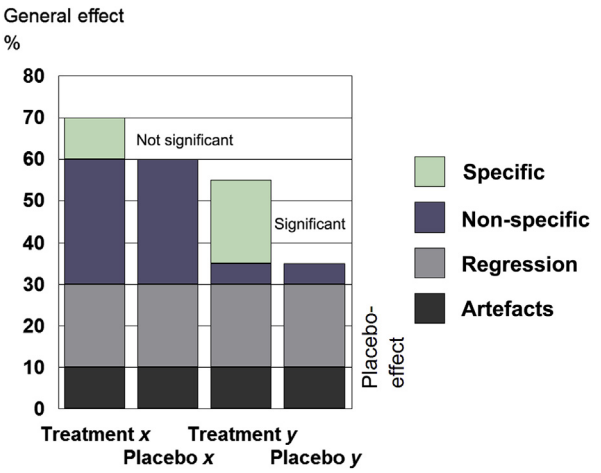

Figure 1 Schematic illustration of the efficacy paradox – Two different treatments, *x* and *y*, are tested against their respective placebo-controls. While treatment *x* is found to be nonsignificantly different from placebo and hence termed 'not efficacious,' treatment *y*, due to a larger specific effect is found efficacious. However, the effect of the placebo in treatment *x* can be larger than the total treatment effect of the specific treatment *y*. Adapted from Walach, H., 2001. The efficacy paradox in randomized controlled trials of CAM and elsewhere: beware of the placebo trap. Journal of Alternative & Complementary Medicine 7, 213–218.

trials, the purported specificity is not there, or is much smaller than expected, or much too small to explain the huge effects, as in many cases of acupuncture or other unconventional treatments. Does this mean that the treatments are worthless? Probably not. The situation rather points to the limitation of the double-blind research paradigm as an instrument of evaluation.

The Inadequacy of the Medical Model for Complex and Behavioral Interventions

The medical pharmacological model has been very powerful, probably because it was a major driver of economic growth since the decades after the Second World War. A lot of researchers have implicitly tried to conform to this model by finding out 'specific' effects of behavioral interventions, such as psychotherapies, relaxation, or meditation interventions. The implicit assumption of linear additivity of those effects seems to be quite naïve. Research shows that this additivity assumption is wrong (Kleijnen et al., 1994). Rather, nonspecific and specific effects interact in a multiplicative fashion, that is, they enhance each other in a synergistic way, or counteract each other. In a synergistic system two components together can do things that neither one single element alone nor both will be able to. An example of a synergistic system is a skilled child rider who can make a powerful horse do things that the horse would not do by itself and the child by itself could not perform either. Precisely because most therapeutic systems are implicitly designed as synergistic systems, it does not make sense to pose the typically pharmacological question of the 'specific' efficacious component of a behavioral intervention.

Some Practical Issues

To create placebos for research is not trivial, as they have to be indistinguishable from the active treatment. Thus, coloring and filling agents, size, packing, and so forth have to be matched, requiring the cooperation of manufacturers. Since most true treatments have quite distinct side-effects, patients, or doctors, or both, can sometimes tell from the lack of side-effects when they are in the placebo-group. Such unblinding seems to happen in about 30% of all trials according to our survey (Walach et al., 2005). The success of blinding is rarely checked in studies, and the extent to which it biases results is unknown. Because of this situation, active placebos are sometimes used, that is, placebos that mimic the side-effects of medications, and analyses show that this procedure increases the placebo-response and makes it more difficult to find differences between groups (Kirsch and Sapirstein, 1999).

If studies have more than two arms, comparing, for instance, two different drugs with a control condition, then it is necessary to construct two different placebos: one group will receive treatment A and placebo B, another group treatment B and a placebo for A, and a third group will receive two placebos, for A and B. Such a design is often called 'double-dummy' because two different placebos are necessary.

The ethical codes that have been developed in the wake of the Nuremberg trials in 1945, after the horrendous crimes committed by physicians on concentration camp inmates became known, require doctors to carefully consider whether placebo-controls are necessary or whether active comparator trials can be done. They also require doctors to fully inform patients about the potential harm through side-effects and benefits they might incur if they enter a trial, which they are likely to receive a placebo, and what the chances of this likelihood are. Doctors also have to consult ethical committees, usually institutional review boards, that scrutinize the protocol and the rationale for it, including the statistical planning and the power analysis. Thus, for ethical reasons, studies need to be as large as necessary to demonstrate an expected effect with a statistically sufficient probability, the so called 'statistical power,' but not larger, so as to not to submit too many patients to potentially inert placebo-treatments. It has also happened quite often that ethical committees demanded unequal randomization ratios that allow for more patients to receive true treatments. Since such ratios have to be disclosed to patients, this also alters their expectation and has made it difficult for some studies to show superiority of otherwise effective drugs, notably in the area of migraine research.

Another practical issue that arises is the fact that placebo-controlled trials are normally designed for proving efficacy, and hence are also only conducted for periods of time necessary to prove this. Very often, the drugs licensed after these trials are then used for much longer periods, with no adequate data to bolster such usage. It is therefore necessary to also look into long-term observational studies, which are typically larger and longer term and can also give information about side-effects and usefulness in the long run.

Finally, placebo-controlled trials are designed to increase internal validity. They achieve this at the expense of external validity: for instance, by introducing many exclusion criteria to homogenize the patient population, or because only a fraction of all potential patients are willing to enter a study. It is therefore necessary when gauging the usefulness of information from a placebo-controlled trial to consider inclusion criteria and the percentage of patients included and to consider the results of such trials against data from other studies, such as cohort studies in large cohorts of patients that have chosen the treatment.

See also: Causal Inference: Overview; Clinical Psychology in Europe, History of; Common Rule: US Regulatory Protections for Human Research Participants; Experimentation in Psychology, History of; External Validity; Fisher, Ronald A (1890–1962); Healing Practices: Alternative and Complementary; Health Outcomes, Assessment of; Hypothesis Testing in Statistics; Informed Consent of Research Subjects; Internal Validity; Logical Positivism and Logical Empiricism; Motivation in Clinical Interventions; Objectivity of Research: Ethical Aspects; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Placebo Effect; Psychological Treatments: Randomized Controlled Trials; Public Health as Social Science; Semiotics; Therapist–Patient Relationship.

Bibliography

Bailly, J.S., 1784. Rapport des Commissaires chargés par le Roi de l'examen du Magnétisme Animal – Imprimé par ordre du Roi. Imprimerie Royale, Paris.

Bergmann, J.-F., et al., 1994. A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer patients. Clinical Trials and Meta-Analysis 29, 41–47.
Colloca, L., Flaten, M.A., Meissner, K. (Eds.), 2013. Placebo and Pain: From Bench to Bedside. Elsevier-Academic Press, Amsterdam.
Conferences on Therapy, 1946. The use of placebos in therapy. New York Journal of Medicine 46, 1718–1727.
Conferences on Therapy, 1954. How to evaluate a new drug. American Journal of Medicine 17, 722–727.
Crabtree, A., 1993. From Mesmer to Freud: Magnetic Sleep and the Roots of Psychological Healing. Yale University Press, New Haven, CT.
Descartes, R., 2003(orig. 1664). Treatise of Man (Traité de l'Homme) (Thomas S. Hall, Trans. and Comment). Prometheus Books, Amherst, NY.
Doll, R., Peto, R., Clarke, M., 1999. First publication of an individually randomized trial (letter). Controlled Clinical Trials 20, 367–368.
Elkin, I., et al., 1989. National Institute of Mental Health treatment of depression collaborative research program: general effectiveness of treatments. Archives of General Psychiatry 46, 971–982.
Fisher, R.A., 1971(orig. 1935). The Design of Experiments. Hafner, New York.
Gøtzsche, P.C., 2013. Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Health Care. Radcliff, London.
Grünbaum, A., 1985. Explication and implications of the placebo concept. In: White, L., Tursky, B., Schwartz, G.E. (Eds.), Placebo – Theory, Research, Mechanisms. Guilford Press, New York, pp. 9–36.
Houschyar, K.S., et al., 2012. Effects of phlebotomy-induced reduction of body iron stores on metabolic syndrome: results from a randomized clinical trial. BMC Medicine 10, 54.
Kam-Hansen, S., et al., 2014. Altered placebo and drug labeling changes the outcome of episodic migraine attacks. Science Translational Medicine 6, 218ra5.
Kaptchuk, T.J., 1998. Intentional ignorance: a history of blind assessment and placebo controls in medicine. Bulletin of the History of Medicine 72, 389–433.
Kaptchuk, T.J., et al., 2010. Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PLoS One 5, e15591.
Khan, A., Khan, S., Brown, W.A., 2002. Are placebo controls necessary to test new antidepressants and anxiolytics? International Journal of Neuro-psychopharmacology 5, 193–197.
Kirsch, I., et al., 2008. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. PLoS Medicine 5, e45.
Kirsch, I., Sapirstein, G., 1999. Listening to Prozac but hearing placebo: a meta-analysis of antidepressant medications. In: Kirsch, I. (Ed.), Expectancy, Experience, and Behavior. American Psychological Association, Washington, DC, pp. 303–320.
Kleijnen, J., et al., 1994. Placebo effect in double-blind clinical trials: a review of interactions with medications. Lancet 344, 1347–1349.
Kuhn, T., 1955. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Martini, P., 1932. Methodenlehre der therapeutischen Untersuchung. Springer, Berlin.
Meissner, K., Kohls, N., Colloca, L. (Eds.), 2011. Placebo Effects in Medicine: Mechanisms and Clinical Implications. Royal Society, London.
Moerman, D.E., Jonas, W.B., 2002. Deconstructing the placebo effect and finding the meaning response. Annals of Internal Medicine 136, 471–476.
Peirce, C.S., Jastrow, J., 1884. On small differences of sensation. Messages of the National Academy of Science 3, 73–83.
Rief, W., et al., 2009. Meta-analysis of the placebo response in antidepressant trials. Journal of Affective Disorders 118, 1–8.
Segar, J., 2012. Complementary and alternative medicine: exploring the gap between evidence and usage. Health 16, 366–381.
Shapiro, A.K., Morris, L.A., 1978. The placebo effect in medicine and psychological therapies. In: Garfield, S.L., Bergin, A. (Eds.), Handbook of Psychotherapy and Behavior Change: An Empirical Analysis, second ed. Wiley, New York.
Shapiro, A.K., Shapiro, E., 1997. The Powerful Placebo: From Ancient Priest to Modern Physician. Johns Hopkins University Press, Baltimore.
Stolberg, M., 1996. Die Homöopathie auf dem Prüfstein. Der erste Doppelblindversuch der Medizingeschichte im Jahr 1835. Münchner Medizinische Wochenschrift 138, 364–366.
Tinetti, M.E., Studenski, S.A., 2011. Comparative effectiveness research and patients with multiple chronic conditions. New England Journal of Medicine 364, 2478–2481.
Turner, E.H., et al., 2008. Selective publication of antidepressant trials and its influence on apparent efficacy. New England Journal of Medicine 358, 252–260.
Walach, H., 2001. The efficacy paradox in randomized controlled trials of CAM and elsewhere: beware of the placebo trap. Journal of Alternative & Complementary Medicine 7, 213–218.
Walach, H., 2011. Placebo controls: historical, methodological and general aspects. Philosophical Transactions of the Royal Society Biological Sciences 366, 1870–1878.
Walach, H., 2013. Placebo effects in complementary and alternative Medicine: the self-healing response. In: Colloca, L., Flaten, M.A., Meissner, K. (Eds.), Placebo and Pain: From Bench to Bedside. Elsevier-Academic Press, Amsterdam, pp. 189–202.
Walach, H., et al., 2006. Circular instead of hierarchical – methodological principles for the evaluation of complex interventions. BMC Medical Research Methodology 6.
Walach, H., et al., 2005. The therapeutic effect of clinical trials: understanding placebo response rates in clinical trials – a secondary analysis. BMC Medical Research Methodology 5, 26.

Placebo-Controlled Trials, Ethics of

Rieke van der Graaf, University Medical Center Utrecht, Utrecht, Netherlands
Annette Rid, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are often good scientific and ethical reasons for using placebo controls in clinical trials. At the same time placebo use is controversial, especially when an established effective treatment is being withheld from the control group. This article gives an overview of the key ethical positions in the controversy around placebo-controlled trials. While some argue that placebo controls can only be used when withholding or delaying an established effective treatment poses no or negligible risks to participants, others hold that the risks should be low or acceptable in light of the social value of the knowledge to be gained from the research. The article also describes different positions on placebo use in trials in low- and middle-income countries where study participants may not have had access to an established effective treatment outside the trial.

Introduction

Before new medical interventions, such as drugs or biologics, are introduced to clinical practice they are normally tested in preclinical and clinical trials. In the preclinical phase, potential new interventions are tested in the laboratory and on animals. Preclinical testing is a prerequisite for clinical testing in humans, but eventually the safety and efficacy of potential new interventions must be tested in humans.

The clinical testing of investigational drugs – the focus of this article – generally proceeds in four phases. In phase I, the maximum tolerated dose of an investigational drug is tested in a small number of study participants (normally 20–50). In phase II, preliminary data on the drug's safety and efficacy are gathered in a larger group of participants (up to 100–300). Phase III trials substantiate the data on safety and efficacy in a significant number of participants (up to several thousands). Typically, the investigational drug is tested in comparison to standard treatment or a 'placebo' – an inert substance, such as a 'sugar pill' or a saline injection that is administered in the same way as the investigational drug, but contains no active substance. Placebo controls are also used in phase I and II trials, although controlled designs are less common in these early phases of clinical testing. The investigational agent is expected to have therapeutic effect in phase 3 trials. If a drug has been proven efficacious in such trials, it can be licensed for regular clinical use by regulatory authorities (e.g., the European Medicines Agency in Europe or the Food and Drug Administration (FDA) in the U.S.).

To evaluate the risk–benefit profile of an investigational drug, it is useful to compare it against placebo. In a randomized, placebo-controlled trial, the study group and the control arm are treated exactly the same, except that the study group receives an active substance. Any random events that happen during the trial (e.g., spontaneous remission), or any clinical effects that are due to positive or negative expectations regarding a treatment (e.g., placebo effect), are therefore expected to be distributed evenly between both groups. In this way, the use of a placebo control allows investigators to determine whether any differences in clinical outcomes between the study arm of a trial (receiving the investigational drug) and the placebo arm can be attributed to the drug under study. These methodological reasons for using placebo controls are also ethically relevant, given that studies without scientific and social value expose participants to risks for no good reason (Emanuel et al., 2000).

However, the use of placebo can be ethically controversial when a proven or established effective treatment exists. Some research ethicists argue that a placebo control should not be used in this situation, as this implies unjustifiably withholding an established effective treatment from participants in the control group (Freedman et al., 1996a,b; Weijer and Miller, 2004). Others consider this stance as too strict. For instance, some argue that the use of placebo controls is acceptable when withholding or delaying an established effective treatment poses low or acceptable risks to participants (WMA, 2013; Emanuel and Miller, 2001). Placebo use is more controversial still in the context of research in low- or middle-income countries (LMICs) when it is used in the presence of an established effective intervention that is not available in the community in which the trial is conducted (e.g., because it is too costly or complex to implement) (London, 2001; Ellenberg and Temple, 2000). This illustrates that the use of placebo controls in clinical trials is both ethically complex and controversial.

This article gives an overview of the ethical controversy surrounding the use of placebos as a comparator in randomized, placebo-controlled clinical trials. Importantly, the article focuses on the use of placebos as it relates to clinical trial design and does not explore other questions regarding placebo use. First, the article does not discuss the placebo effect and its potential implications for informed consent to clinical research. It is well documented that positive expectations about an intervention can improve clinical outcomes, even when patients or study participants receive an inactive treatment (placebo effect) (Kam-Hansen et al., 2014). Conversely, negative expectations may evoke a negative clinical response with patients or participants experiencing real side effects (nocebo effect) (Colloca and Finniss, 2012). These findings raise the possibility that the risk–benefit profile of a clinical trial can be influenced by manipulating participants' expectations in the informed

 http://dx.doi.org/10.1016/B978-0-08-097086-8.11011-6

consent process. While placebo and nocebo effects are relevant for evaluating the risks and potential benefits of study participation, the article does not address to what extent these effects can justify modifications of informed consent (Brim and Miller, 2013).

Second, in research on the placebo and nocebo effect itself it can be methodologically necessary to deceive study participants about the study purpose and the use of placebo interventions. For example, disclosure that participants will receive an inert substance in order to study its impact on their pain will influence their expectations and therefore influence study results. The use of deception in this and other research is discussed in a separate article in this encyclopedia.

Third, the focus of this article is on the use of placebo controls in drug trials. Trials of surgical or anesthetic interventions can involve sham ('fake') procedures that are similar to placebo controls, in that they help investigators to distinguish the effects of the intervention under study from placebo effects or random events. The risks of sham procedures can be considerable, and they can sometimes involve deliberately harming participants (e.g., surgical incision under general anesthesia). Because these features warrant separate discussion, sham procedures are not considered in this article (London and Kadane, 2002; Miller, 2003).

Finally, the present article is only concerned with the use of placebo interventions in the research context. Placebos are also used in the clinical setting as a treatment offered to patients. The ethical issues regarding the clinical use of placebos are discussed elsewhere in this encyclopedia (*see* Placebo Effect).

Rationale for Using Placebo Controls in Clinical Research

From a methodological point of view, randomized, placebo-controlled clinical trials are often considered the gold standard for testing the safety and efficacy of a potential new intervention. In such trials, participants are allocated to two (or more) arms of the trial. In the study arm, participants receive the intervention under investigation, while participants in the control arm receive a placebo. Randomization – assigning participants to the arm of a trial by chance – is the preferred method for preventing confounding and bias in controlled clinical trials. Randomization helps investigators to ensure that study groups will be similar with respect to known (e.g., age) and unknown prognostic factors (e.g., spontaneous remission). If investigators cannot predict in advance to which arm participants will be assigned, this will reduce allocation bias (Friedman et al., 2010). When combined with the use of an inert placebo in the control arm, randomization enables investigators to isolate the efficacy of an investigational drug in order to determine whether different clinical outcomes between the trial arms can be attributed to the drug under study. It is for this reason that randomized, placebo-controlled trials are often considered the gold standard in clinical trial design. Placebo-controlled trials also have a higher 'assay sensitivity' than active controlled trial designs – i.e., they have a better ability to distinguish an effective treatment from a less effective intervention (Temple and Ellenberg, 2000). In an active controlled trial, in which a new drug is compared to an existing drug, data of previous (placebo-controlled) trials are needed in order to know whether the drug in the intervention arm is indeed effective.

Importantly, these methodological considerations matter not only scientifically but also ethically, as research needs to have scientific and social value to be ethically acceptable (Emanuel et al., 2000). If the research question cannot be answered without the use of placebo controls, the use of another trial design is not scientifically valid.

Placebo controls can also help to ensure the scientific validity of a study in light of expected recruitment problems, given that placebo-controlled trials normally require smaller numbers of participants. Trial completion is a continuing challenge in clinical research generally, and it is a particular challenge in research on rare diseases where studies 'compete' for a very limited number of eligible participants. Data cannot be used when trials are not completed, hence the possibility of underrecruitment can compromise the scientific and social value of research studies. By requiring smaller numbers of participants, placebo-controlled trials reduce the possibility of underrecruitment while also exposing fewer individuals to the risks of an unproven intervention.

In addition, because placebo-controlled trials require a smaller number of participants than trials that compare existing treatments, they can be useful for consolidating the evidence to support larger and more expensive active-comparator trials. The reduced costs of placebo-controlled trials may also be a consideration when trials aim to develop low-cost interventions primarily for use in LMICs (see later). In this situation, research budgets can be so tight that alternative trial designs may not be feasible, while inclusion of an active control arm is not always necessary to decide whether or not the study intervention should be adopted by the local healthcare system (provided that some level of efficacy has been proven in the trial). For example, when an established effective treatment is not affordable locally, and likely to remain unaffordable in the future, the costs of an active-comparator trial may not be justified: if the active comparator turns out to be superior it cannot be implemented.

Moreover, placebo controls may enhance the social value of studies since randomized, placebo-controlled trials are often perceived to produce the highest level of evidence in clinical research (Barton, 2000). Although this perception of placebo-controlled trials is contested, the commonly assumed superior quality of evidence produced by placebo-controlled trials still affects the practical uptake of new interventions – or the discontinuation of routine but unproven interventions – among clinicians and policy-makers.

Finally, placebo-controlled trials are frequently preferred – or perceived to be preferred – by regulatory authorities who are responsible for approving drugs for the market. For example, the FDA in the U.S. states that "adequate and well-controlled" studies are essential to test the efficacy of drugs and mentions placebo controls as the first control option (DHHS, 2008). Many therefore believe that the FDA *requires* the use of placebo-controlled trial designs. However, the FDA

acknowledges that the use of placebo controls can be ethically problematic if there is a "known effective therapy," and "placebo or no treatment would be contrary to the interest of the patient" (DHHS, 2008). Thus, while the FDA has a methodological preference for placebo-controlled trials, it does not require such trials in order to make licensing decisions about potential new treatments (Miller, 2008).

In summary, there are often good scientific, ethical, and regulatory reasons to support placebo-controlled trials. However, the use of placebo controls is not always ethically justified and often controversial, especially when it involves withholding or delaying an established effective treatment from participants in the control arm. Before we discuss the key ethical issues surrounding placebo use in this situation, the following section briefly reviews the controversy around placebo-controlled trials from a historical perspective.

Controversy Around Placebo-Controlled Trials

Clinical research expanded enormously in the twentieth century, and randomized-controlled trials became standard practice from the 1950s and 1960s onward (Kaptchuk, 1998). The evidence-based medicine (EBM) movement in the 1980s and 1990s then declared such trials, and especially those involving a placebo control, as the methodological 'gold standard' in clinical research (Sacks et al., 1982). However, scholars in the philosophy of science and other disciplines quickly pointed out that high-quality observational studies and well-designed randomized-controlled trials may produce data of comparable quality. It is for this reason that the strict hierarchy of evidence promoted by the EBM movement, with the randomized-controlled design at the top, has been questioned (Barton, 2000; Cartwright, 2011).

These and other issues sparked a first peak of sustained ethical controversy around randomized, placebo-controlled trials in the 1980s. At the core of this debate were the fundamental ethical obligations of clinician-investigators: Should their conduct be guided by the norms of clinical care or by the norms of research? Some argued that clinician-investigators maintain therapeutic obligations toward patients who participate in clinical trials (Marquis, 1983). Given that treatment in randomized-controlled trials is determined by chance, not by what the clinician-investigator believes is best for the participant, clinician-investigators were seen to violate their therapeutic obligations when enrolling patients in such trials. Enrollment in randomized, placebo-controlled trials in the presence of established effective treatment was seen as a particularly stark violation of clinician-investigators' therapeutic obligations, as it implied withholding an effective treatment (Rothman and Michels, 1994).

To resolve the conflict between the therapeutic obligations of clinician-investigators and the need to conduct rigorous clinical research, scholars introduced the principle of clinical equipoise. In its most influential formulation, this principle holds that patients should only be enrolled in randomized-controlled trials when there is "genuine uncertainty in the medical expert community about the preferred treatment" (i.e., the community of experts is 'equally poised' between a potential new intervention and any established effective treatments; Freedman, 1987). In a state of clinical equipoise, clinicians do not have sufficient information to know which of the available treatment options (if any) is preferable, either because there is no information on what is best or because they disagree about the relative merits of this information where some prefer treatment A and others B (van der Graaf and van Delden, 2012). Therefore, clinicians as investigators do not violate their therapeutic obligations when they randomize patients to the available treatments or placebo, provided that a state of clinical equipoise with regard to these treatments and placebo obtains.

Some commentators have argued in response to these developments that the attempt to align research with the therapeutic obligations of clinicians is fundamentally flawed (Miller and Brody, 2002). On this view, research is an activity that is distinct from clinical care, and therapeutic obligations therefore do not apply in the research context. We discuss this controversy in more detail.

In addition to ethical issues on the level of the individual participant and investigator, the controversy around placebo-controlled trials also touched on issues at the population level. For instance, clinicians argued that comparative effectiveness data are often more valuable for patient care than data from randomized, placebo-controlled trials. For clinicians and patients, it is generally more interesting to know how a potential new intervention compares to an established effective treatment (if any) than whether it is better than 'nothing.' Another issue at the population level is that many expensive treatments are licensed, used, and paid for on the basis of placebo-controlled trials, when it is not clear how they compare to any existing interventions. Although consideration of these issues is essential for determining the overall acceptability of placebo-controlled trials, this article focuses largely on the ethical issues raised on the participant and investigator level.

A second peak in the ethical controversy surrounding randomized, placebo-controlled trials arose in the 1990s in the context of research in LMICs. A series of high-profile articles in *The New England Journal of Medicine* questioned the ethics of the then ongoing AZT (zidovudine) trials in LMICs (Angell, 1997; Lurie and Wolfe, 1997). Researchers were testing a short course of AZT to reduce the risk of maternal–infant HIV transmission, given that the complex and costly standard regimen in use in high-income countries (HICs) was neither feasible nor affordable in resource-poor settings. Most trials tested the short course of AZT against placebo, a trial design that – it was widely agreed – would not have been acceptable in HICs. In HICs, a placebo-controlled trial would have deprived participants in the control arm of an established effective intervention that they would have otherwise received outside the trial. Moreover, withholding this intervention was associated with significant risk, as the standard AZT regimen was proven to reduce perinatal HIV transmission from 25 to 8%. Commentators therefore argued that the AZT trials demonstrated the unjustifiable use of 'double standards' in research ethics, with study participants in LMICs being afforded lower protection than participants in HICs

(Lurie and Wolfe, 1997). This criticism was especially cutting because the trials were sponsored by HIC governments or international organizations, which arguably had the resources to provide participants in the control arm with the standard AZT regimen. However, there was deep controversy surrounding the AZT trials. Many argued that the trials were ethically justifiable because they addressed the health needs of populations in LMICs, and participants in the placebo arm were not made worse off than they would have been outside the trial (Varmus and Satcher, 1997). The ethics of AZT-type trials – which aim to develop a 'substandard' treatment specifically for use in LMICs by testing it against placebo – remains debated to this day.

Even more controversial were placebo-controlled trials in LMICs that tested potential new treatments primarily intended for use in HICs. These trials form part of a larger controversy about the 'outsourcing' of clinical research to LMICs, which is motivated by a range of factors, including the lower cost of conducting research in these countries, a generally large and ready pool of potential study participants for whom trial enrollment may be the only way of accessing medical care, and the oftentimes less onerous research regulations in LMICs (Petryna, 2009). However, outsourcing is particularly controversial when it involves the use of placebo controls that would not be considered acceptable in HICs. A prominent example for these types of trials is the 'Surfaxin trial' (Hawkins and Emanuel, 2008). In this trial, a pharmaceutical company from the U.S. proposed to test a potential new treatment to prevent respiratory distress syndrome in premature infants in Bolivia. Because four established effective treatments were on the market in the U.S. and highly effective in preventing a potentially fatal condition, a placebo-controlled trial design would not have been acceptable in the U.S. However, the company did not think it would succeed in obtaining regulatory approval for the potential new treatment based on a trial comparing their investigational agent against one of the established effective treatments. By contrast, approval based on a placebo-controlled trial was deemed feasible. The company therefore proposed to conduct the trial in Bolivia, although it did not intend to license and market the potential new treatment in Bolivia or other LMICs – the U.S. and other HICs were always the target market. Although the Surfaxin trial never took place, it became a textbook example in the ethical controversy around research in LMICs (Hawkins and Emanuel, 2008).

While AZT-type trials primarily raised questions about investigators' obligations toward individual study participants in LMICs, given that these trials aim to address local health needs, Surfaxin-type trials also raise questions about exploitation on the population level. After all, Surfaxin-type trials expose populations in LMICs to research risks even though they aim to develop interventions that will benefit populations in HICs. This article discusses issues around placebo use only insofar as they concern investigator obligations toward individual participants; readers who wish to explore questions about population-level exploitation are directed elsewhere (Hawkins and Emanuel, 2008).

The controversy around placebo-controlled trials had important ramifications for the World Medical Association's *Declaration of Helsinki*, one of the most prominent documents in research ethics. Up to the 1990s, the *Declaration* required that research participants be assured of the best proven therapeutic method, thereby effectively blocking the use of placebo controls when an established effective treatment exists. The controversy around placebo-controlled trials prompted the World Medical Association to relax its stance on placebo use – but this in turn caused much debate and initiated a rapid succession of revisions of the *Declaration* in the 2000s and 2010s (the current version dates from 2013). Many commentators believe that the *Declaration* has lost much of its authority as a result, in particular regarding the use of placebos in randomized-controlled trials (Lie; FDA, Brazil).

Key Ethical Positions

It is widely agreed that the use of placebo controls raises no ethical concerns when there is no established effective intervention for the condition under study. In this situation, a placebo-controlled trial answers the clinically relevant question whether the potential new treatment is more effective than 'nothing,' and participants are not deprived of the clinical benefits of an established effective intervention. Similarly, placebo use is uncontroversial when an investigational agent is tested as an add-on to standard care (e.g., in oncology where new chemotherapeutic agents are often used in combination with established treatments). In these trials, all participants receive standard care while also being treated with either the investigational agent or a placebo. However, the ethical positions on whether placebo use is acceptable in the presence of an established effective intervention – and, if so, under which conditions – vary widely. The key issue of contention is whether it is acceptable for investigators to expose participants to risks – understood as the likelihood of experiencing physical, psychological, or other harm as result of undergoing a study intervention or research procedure – by deliberately withholding or delaying an established effective treatment. This is also based on the assumption that the placebo interventions themselves typically pose negligible risks (e.g., ingestion of a 'sugar pill,' injection of saline solution through an existing intravenous line).

No or Negligible Risk from Withholding Treatment

Some commentators maintain that the use of placebo control is acceptable only when there is no proven or established effective intervention for the condition under study. This 'no risk' position effectively excludes the use of placebo controls in the presence of an established effective intervention. The closely related 'negligible risk' position allows the use of placebo controls in this situation, but requires that withholding or delaying an established effective intervention poses no more than negligible risks. For example, on the 'negligible risk' position, placebo use is acceptable in the case of self-limiting diseases with mild symptoms where foregoing treatment is consistent with sound medical care (e.g., allergic rhinitis).

The 'no or negligible risk' position is grounded in the idea that clinical investigators have therapeutic obligations toward research participants (Marquis, 1983) and that clinical equipoise, which is grounded in these obligations, is a necessary requirement for clinical research (see previous section; Freedman et al., 1996a,b; Weijer and Miller, 2004). According to proponents of the 'no or negligible risk' position, it is impossible for clinicians to shed their therapeutic obligations simply by taking on the role of investigator. Investigators may therefore not withhold or delay an established effective treatment from participants in the placebo arm, even if doing so only requires a 'small sacrifice' from participants and there are strong methodological reasons for using a placebo control. The 'no or negligible risk' position thus places strict limitations on clinical trial design.

Low Risks from Withholding Treatment

Many believe that the 'no or negligible risk' position is overly strict, as it disallows many valuable studies that are – in their view – ethically justifiable. These commentators frequently adopt a 'low-risk' position on the use of placebo controls when an established effective intervention exists, arguing that placebo controls may be used when (1) the risks of withholding or delaying the established effective intervention are low and (2) there are compelling methodological reasons for using a placebo control (reference). Examples for 'compelling methodological reasons' include situations where the clinical response to the established effective intervention is highly variable; the symptoms of the condition under study fluctuate and/or there is a high rate of spontaneous remission; and the condition under study is known to have a high placebo response (Miller, 2008; Millum and Grady, 2013). In these situations, it can be difficult to determine without a placebo control whether the experimental intervention is efficacious, since the condition may be improving on its own or the observed clinical response may be due to a placebo effect (Temple and Ellenberg, 2000). For example, many trials of antidepressants use placebo controls because patients with depression often have waxing and waning symptoms, and depressive symptoms are known to have a high placebo response (Stone et al., 2009).

Importantly, proponents of the 'low-risk' position have different views on what constitutes low risk. Some commentators or guidelines require that the risks of withholding or delaying an established effective intervention may only lead to temporary discomfort or fleeting harm (CIOMS, 2002). Others are more permissive and interpret low risk as constituting no risk of serious harm (WMA, 2013). Depending on which conception of low risk they endorse, proponents of the 'low-risk' position may consider the same trial as ethically acceptable or unacceptable.

Like the 'no or negligible' risk position, the 'low-risk' position is often grounded in the idea that clinician-investigators have therapeutic obligations toward research participants and that investigators must, accordingly, be in a state of clinical equipoise regarding the interventions under study (Marquis, 1983; Freedman, 1987). However, unlike those who defend a 'no or negligible risk' position, proponents of the 'low-risk' position do not regard the principle of clinical equipoise as an absolute norm that can never be overridden by other moral norms and hence excludes small sacrifices to investigators' therapeutic obligations (van der Graaf and van Delden, 2009; Beauchamp and Childress, 2013). In other words, if there is a standard of care for patients with a certain condition, but the risks of withholding this standard of care are low and placebo control may be needed for compelling reasons, violation of equipoise need not be considered as ethically problematic.

'Acceptable' Risk from Withholding Treatment

A few commentators believe that the 'low-risk' position on placebo-controlled trials still disallows many valuable and – in their view – ethically justifiable studies. These commentators agree with proponents of the 'low-risk' position that placebo use may be justified in the presence of an established effective intervention, but they reject its strict limitation on risk. According to proponents of the 'acceptable risk' position, the use of placebo controls is justified when (1) there are compelling methodological reasons for a placebo-controlled trial design and (2) the risks of withholding or delaying treatment are minimized and reasonable in relation to the knowledge to be gained from the placebo-controlled trial design (Emanuel and Miller, 2001; Miller, 2008).

Proponents of the 'acceptable risk' position typically justify their view based on role-based obligations of investigators, which are seen to be distinct from the role-based obligations of clinicians (Miller and Brody, 2002). As mentioned earlier, according to these commentators, the fundamental ethical obligation of investigators is not to promote the best clinical interests of study participants. Instead, investigators have an obligation to promote the social good by conducting scientifically and socially valuable research, while ensuring that research participants are not exposed to excessive risks for the benefit of others (principle of nonexploitation). The use of placebo controls must therefore be scrutinized like most other research procedures. This implies – consistent with the earlier-mentioned criteria for acceptable placebo use – that the risks of withholding or delaying an established effective intervention must be minimized and justified in relation to the potential clinical benefits to participants and/or the scientific and social value of the knowledge to be gained from the research (Miller and Brody, 2003, 2007; Rid and Wendler, 2011).

'Acceptable' Risk from Withholding Treatment in Research in LMICs

Questions around 'acceptable' risk from withholding or delaying an established effective treatment acquire a further dimension when trials are conducted in LMICs. Many people in LMICs do not have access to established effective interventions that are commonly used in HICs, typically because these interventions are too expensive and/or the healthcare infrastructure for their safe and effective delivery is not available (e.g., services for intravenous treatment or cooling chain for certain vaccines). This lamentable state of affairs can be attributed, at least in part, to background injustices resulting

from the corruption of local governments, international treaties that disadvantage LMIC economies, and many other factors (Pogge, 2002). A key question around the use of placebo controls in LMICs therefore is this: To what extent are the obligations of investigators influenced by background injustices?

Few will argue that background injustices have no role to play – that is, few will argue that it is acceptable for investigators to withhold or delay an established effective intervention simply because it is not available locally (London, 2001). However, commentators disagree profoundly on how background injustices shape the obligations of investigators (and research sponsors). The key issue of contention is which level of care investigators ought to provide participants in the control arm of their trials (van der Graaf and van Delden, 2009; Hawkins and Emanuel, 2008).

Some of those who believe that investigators have therapeutic obligations argue that study participants in the control arm are owed the worldwide best established intervention, independent of where they live (Angell, 1997; Nuffield, 2002; Macklin, 2004). For example, in the AZT trials described earlier, these commentators argued that the use of a placebo control was unacceptable and participants in the control arm should have received the long course of AZT, which was the best treatment available worldwide at the time. Others hold that investigators can discharge their therapeutic obligations by providing less than the worldwide best established intervention – for example, the best treatment that they can reasonably provide under the circumstances (Resnik, 1998), or the treatment that participants are entitled to outside the study according to local health care norms (London, 2001; Kukla, 2007).

Questions about the appropriate standard of care are also relevant for those who believe that investigators' primary obligation is to promote the social good while not exposing participants to excessive risks for the benefit of others (principle of nonexploitation). For these commentators, to determine whether placebo use poses excessive risk, it is necessary to compare the use of a placebo control to the relevant standard of care (i.e., it is necessary to evaluate the risks of withholding or delaying the relevant baseline treatment). Yet views on what constitutes the relevant standard of care again vary. For some, the relevant standard is the treatment that participants would have normally received, provided that the research addresses a question that is relevant to local health needs (Lie et al., 2004). For others, the relevant baseline is the treatment that participants *should* have received outside that trial as a matter of global distributive justice (Emanuel, 2012).

Finally, some commentators argue that investigators have natural obligations – obligations they possess simply because they are persons, not because they occupy a particular professional role. This line of argument becomes especially relevant in LMIC settings where background injustices and limited resources frequently lead to situations of dire need. For example, some commentators argue that investigators should sometimes provide participants in the control arm with effective treatment because all people, including investigators, have a duty to rescue those whose life or health is seriously threatened (Hawkins, 2008). On this view, this general duty of rescue requires investigators to provide an established effective treatment, if one exists, when they can prevent or treat an immediate and serious threat to a patient's life or health at reasonable cost to themselves. This position does not necessarily imply, however, that placebo use can never be justified when it involves withholding or delaying an established effective treatment for a serious condition. These commentators require that placebo be provided if the research has important social value, using a placebo control is the only way of addressing the research question, and the population that the control group represents is reasonably likely to benefit from the research. The latter condition implies that it is ethically problematic to conduct a placebo-controlled trial in an LMIC that will never benefit from the results when the investigators can conduct the trial in an LMIC where the population is highly likely to benefit (Hawkins, 2008).

Key Ethical Guidance

Most ethical guidelines for research involving human subjects adopt a version of the 'low-risk' position on placebo use in the presence of an established effective treatment, although there are exceptions. For example, Brazilian authorities adopt the 'no risk' position by endorsing the *Declaration of Helsinki* in its 2000 version, which allows placebo controls only "in studies where no proven prophylactic, diagnostic or therapeutic method exists" (Brazil, 2008). By contrast, the South African research ethics guidelines adopt an 'acceptable risk' position by allowing the use of placebo controls in the presence of an established treatment, provided that participants will not be harmed – implying, arguably, that participants will not be made worse off than they would have been outside the trial – and the risks to participants are justified by the potential benefits of the research (South Africa, 2004). **Table 1** summarizes prominent ethical guidance on the use of placebo controls in clinical trials.

Concluding Remarks

There are often good scientific, regulatory, and ethical reasons for using a placebo control in randomized-controlled clinical trials. At the same time, placebo-controlled trials are deeply controversial, in particular when the use of a placebo control involves withholding or delaying an established effective treatment from study participants. The ethical controversy goes to the very heart of research ethics, as it revolves around the question what fundamentally justifies the obligations of investigators. Most research ethics guidelines require that the risks of using placebo controls – including the risks of foregoing any established effective interventions – be minimal, and that there be compelling scientific reasons for placebo use. At the same time, the minimal risk threshold is being criticized, and the notion of compelling scientific reasons remains to be specified. Therefore, it seems safe to say that the use of placebo controls in randomized controlled trials will continue to be controversial.

Table 1 Prominent ethical guidance on placebo use in clinical research

World Medical Association Declaration of Helsinki (2013)	The benefits, risks, burdens, and effectiveness of a new intervention must be tested against those of the best proven intervention(s), except in the following circumstances: Where *no proven intervention exists*, the use of placebo, or no intervention, is acceptable; or Where for *compelling and scientifically sound methodological reasons*, the use of any intervention less effective than the best proven one, the use of placebo, or no intervention is necessary to determine the efficacy or safety of an intervention and the patients who receive any intervention less effective than the best proven one, placebo, or no intervention will *not be subject to additional risks of serious or irreversible harm* as a result of not receiving the best proven intervention. Extreme care must be taken to avoid abuse of this option.
Tri-Council Policy Statement 2 (2010)	Article 11.2 (a) A new therapy or intervention should generally be tested against an established effective therapy. (b) As with all alternative choices of a control, a placebo control is ethically acceptable in a randomized controlled clinical trial only if: its use is *scientifically and methodologically sound* in establishing the efficacy or safety of the test therapy or intervention; and *it does not compromise the safety or health of participants*; and the researcher articulates to the REB a *compelling scientific justification for the use of the placebo control*.
Brazil, Conselho Nacional de Saúde, RESOLUÇÃO CNS N° 404 (2008)	Propose the withdrawal of notes for clarification of items related to health care to be provided to volunteers and the use of placebo, since they restrict the rights of volunteers to health care, keeping the following texts of the 2000 version of the Declaration of Helsinki: (…) b) Use of placebo: The benefits, risks, burdens, and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where *no proven prophylactic, diagnostic, or therapeutic method exists*.
Uganda, UNCST, Guidelines for Research Involving Humans (2007)	7.2 *Standard of care for research participants in control groups* (…) in general, placebo-controlled trials may be conducted provided that: 1. Based on knowledge that is available at the commencement of the trial, the new drug or device to be tested does not confer any significant benefit compared to the placebo, and a proven prophylactic, diagnostic, or therapeutic method or established effective intervention does not exist. 2. Withholding an established effective intervention would expose the research participant to at most temporary discomfort or delay in relief of symptoms; 3. Use of an established effective intervention as comparator would not yield scientifically reliable results and the use of a placebo would not add any significant risk or irreversible harm to the research participants. (…)
Council of Europe, Additional protocol concerning biomedical research (2005)	Article 23 2. In research associated with prevention, diagnosis, or treatment, participants assigned to control groups shall be assured of proven methods of prevention, diagnosis, or treatment. 3. The use of placebo is permissible where there are no methods of proven effectiveness, or where *withdrawal or withholding of such methods does not present an unacceptable risk or burden.*
CIOMS International Ethical Guidelines for Biomedical Research Involving Human Subjects (2002)	As a general rule, research subjects in the control group of a trial of a diagnostic, therapeutic, or preventive intervention should receive an established effective intervention. In some circumstances, it may be ethically acceptable to use an alternative comparator, such as placebo or 'no treatment.' Placebo may be used: ● when there is no established effective intervention; ● when withholding an established effective intervention would expose subjects to, at most, temporary discomfort or delay in relief of symptoms; ● when use of an established effective intervention as comparator would not yield scientifically reliable results and use of placebo would not add any risk of serious or irreversible harm to the subjects.

South Africa, Department of Health, Research Ethics Guidelines (2004)	2.15 The choice of intervention (placebo or some treatment) to administer to participants in the control arm of a study may require the balancing of many factors, but the welfare of participants must be paramount. (…) Ethics review committees should verify that the control is appropriate, does not impose risks that are unreasonable in relation to the anticipated benefits, and that placebo controls are not employed without compelling justification. (…) Justifications for using a placebo in the control arm are that: ● No treatment or intervention is accepted as being effective for the condition; ● Treatments or interventions are accepted as being effective for the condition, but the use of a placebo will not result in more than minimal adverse effects that are entirely reversible; ● Treatments or interventions are accepted as effective for the condition, but no scientifically justifiable control option, other than a placebo, meets the objective of the research, and the anticipated benefits of the research substantially outweigh the risks to participants. (…) Ethical guidelines that apply to controlled therapeutic trials are generally adequate to protect the rights of HIV-infected persons. A special case involves the use of placebo after an intervention has already been shown to be effective. The general principle is that the use of placebo in these circumstances is unethical. However, with increasing disparities in health care between wealthy and poor countries, therapy that has been shown to be effective is often unaffordable in resource-poor settings. This is particularly true of therapeutic advances in HIV infection, which is a far bigger health care problem in poor countries in sub-Saharan Africa than in industrialized countries. *It may on occasion be justifiable to use placebo in communities that do not have access to interventions that are the standard care in resource-rich settings*. However, placebos may only be used when the *anticipated benefits outweigh the risks* to participants, and participants will *not be harmed*, and *full justification* must be provided for use of placebo.
ICH, guideline on Choice of control group E10 (2000)	2.1.3 *Ethical issues [regarding placebo control].* When a new treatment is tested for a condition for which *no effective treatment is known*, there is usually no ethical problem with a study comparing the new treatment to placebo. Use of a placebo control may raise problems of ethics, acceptability, and feasibility, however, when an effective treatment is available for the condition under study in a proposed trial. In cases where an *available treatment is known to prevent serious harm*, such as death or irreversible morbidity in the study population, it is generally *inappropriate to use a placebo control*. There are occasional exceptions, however, such as cases in which standard therapy has toxicity so severe that many patients have refused to receive it. In other situations, when there is no serious harm, it is generally considered ethical to ask patients to participate in a placebo-controlled trial, even if they may experience discomfort as a result, provided the setting is noncoercive and patients are fully informed about available therapies and the consequences of delaying treatment. (…) Whether a particular placebo-controlled trial of a new agent will be acceptable to subjects and investigators when there is known effective therapy is a matter of investigator, patient, and institutional review board (IRB)/independent ethics committee (IEC) judgment, and acceptability may differ among ICH regions. (…).

Acknowledgments

This article has benefited from our discussions with the Council of International Organizations of Medical Sciences (CIOMS) Working Group to revise the 2001 CIOMS International Ethical Guidelines for Biomedical Research Involving Human Subjects. Ingeborg van der Tweel and Hans van Delden provided helpful comments on an earlier draft of this article. Annette Rid received funding from the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program (FP7/2007-2013; REA grant no. 301816).

See also: Ethics for Biomedical Research Involving Humans: International Guidelines; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Placebo Effect; Placebo Studies (Double-Blind Studies); Psychological Treatments: Randomized Controlled Trials.

Bibliography

Angell, M., 1997. The ethics of clinical research in the Third World. New England Journal of Medicine 337 (12), 847–849.

Barton, S., 2000. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. British Medical Journal 321 (7256), 255–256.

Beauchamp, T.L., Childress, J.F., 2013. Principles of Biomedical Ethics, seventh ed. Oxford University Press, Oxford.

Benatar, S.R., Brock, G. (Eds.), 2011. Global Health and Global Health Ethics. Cambridge University Press, Cambridge.

Brazil, Conselho Nacional de Saúde, RESOLUÇÃO CNS N° 404 (2008): http://conselho.saude.gov.br/web_comissoes/conep/aquivos/resolucoes/resolucoes.htm

Brim, R.L., Miller, F.G., 2013. The potential benefit of the placebo effect in sham-controlled trials: implications for risk–benefit assessments and informed consent. Journal of Medical Ethics 39 (11), 703–707.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, 2010. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. Interagency Advisory Panel on Research Ethics, Government of Canada, Ontario, Canada.

Cartwright, N., 2011. A philosopher's view of the long road from RCTs to effectiveness. Lancet 377 (9775), 1400–1401.

Colloca, L., Finniss, D., 2012. Nocebo effects, patient–clinician communication, and therapeutic outcomes. Journal of the American Medical Association 307 (6), 567–568.

Concato, J., Shah, N., Horwitz, R.I., 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine 243 (25), 1887–1892.

Council for International Organizations of Medical Sciences (CIOMS), Ethical Guidelines for Biomedical Research Involving Human Subjects, 2002. http://www.cioms.ch/publications/guidelines/guidelines_nov_2002_blurb.htm.

Council of Europe, Additional Protocol to the Convention on Human Rights and Biomedicine, Concerning Biomedical Research, 2005. http://conventions.coe.int/Treaty/en/Treaties/Html/195.htm.

Department of Health and Human Services, 2008. Food and Drug Administration. 21 Code of Federal Regulations Part 312. Human Subject Protection; Foreign Clinical Studies Not Conducted under an Investigational New Drug Application. http://www.fda.gov/ohrms/dockets/98fr/E8-9200.htm.

Department of Health and Human Services, 2013. Food and Drug Administration 21 Code of Federal Regulations. Chapter 1. Subchapter D. Part 314. Subpart D. Sec 314.126 Adequate and well-controlled studies. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.126.

Ellenberg, S.S., Temple, R., 2000. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. Annals of Internal Medicine 133 (6), 464–470.

Emanuel, E.J., 2012. Global justice and the "standard of care" debates. In: Millum, J., Emanuel, E.J. (Eds.), Global Justice and Bioethics. Oxford University Press, Oxford, pp. 181–212.

Emanuel, E.J., Miller, F.G., 2001. The ethics of placebo-controlled trials – a middle ground. New England Journal of Medicine 345 (12), 915–919.

Emanuel, E.J., Wendler, D., Grady, C., 2000. What makes clinical research ethical? JAMA 283 (20), 2701–2711.

Freedman, B., 1987. Equipoise and the ethics of clinical research. New England Journal of Medicine 317 (3), 141–145.

Freedman, B., Weijer, C., Glass, K.C., 1996a. Placebo orthodoxy in clinical research. I: Empirical and methodological myths. Journal of Law, Medicine & Ethics 24 (3), 243–251.

Freedman, B., Glass, K.C., Weijer, C., 1996b. Placebo orthodoxy in clinical research. II: Ethical, legal, and regulatory myths. Journal of Law, Medicine & Ethics 24 (3), 252–259.

Fried, C., 1974. Medical Experimentation: Personal Integrity and Social Policy. North-Holland Publishing Company, Amsterdam.

Friedman, L.M., Furberg, C.D., DeMets, D.L., 2010. Fundamentals of Clinical Trials, fourth ed. New York.

van der Graaf, R., van Delden, J.J.M., 2009. What is the best standard for the standard of care? American Journal of Bioethics 9 (3), 35–43.

van der Graaf, R., van Delden, J.J.M., 2012. Equipoise should be amended, not abandoned. Clinical Trials 9 (5), 628–629.

Hawkins, J.S., 2008. Exploitation and placebo controls. In: Hawkins, J.S., Emanuel, E.J. (Eds.), Exploitation and Developing Countries. The Ethics of Clinical Research. Princeton University Press, Princeton, pp. 246–285.

Hawkins, J.S., Emanuel, E.J. (Eds.), 2008. Exploitation and Developing Countries. The Ethics of Clinical Research. Princeton University Press, Princeton.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials, 2000. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf.

Kam-Hansen, S., Jakubowski, M., Kelley, J.M., Kirsch, I., Hoaglin, D.C., Kaptchuk, T.J., Burstein, R., 2014. Altered placebo and drug labeling changes the outcome of episodic migraine attacks. Science Translational Medicine 6 (218), 218ra5.

Kaptchuk, T.J., 1998. Powerful placebo: the dark side of the randomized controlled trial. Lancet 351, 1722–1725.

Kukla, R., 2007. Resituating the principle of equipoise: justice and access to care in non-ideal conditions. Kennedy Institute of Ethics Journal 17 (3), 171–202. http://www.ncbi.nlm.nih.gov/pubmed/?term=kukla+equipoise.

Lie, R.K., Emanuel, E., Grady, C., Wendler, D., 2004. The standard of care debate: the Declaration of Helsinki versus the international consensus opinion. Journal of Medical Ethics 30 (2), 190–193.

London, A.J., Kadane, J.B., October 2002. Placebos that harm: sham surgery controls in clinical trials. Statistical Methods in Medical Research 11 (5), 413–427.

London, A.J., 2001. Equipoise and international human-subjects research. Bioethics 15 (4), 312–332.

London, A.J., 2000. The ambiguity and the exigency: clarifying 'standard of care' arguments in international research. Journal of Medicine and Philosophy 25 (4), 379–397.

Lurie, P., Wolfe, S.M., 1997. Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. New England Journal of Medicine 337 (12), 853–856.

Macklin, R., 2004. Double Standards in Medical Research in Developing Countries. Cambridge University Press, Cambridge.

Marquis, D., 1983. Leaving therapy to chance. Hastings Center Report 13 (4), 40–47.

Miller, F.G., Joffe, S., 2011. Equipoise and the dilemma of randomized clinical trials. New England Journal of Medicine 364 (5), 476–480.

Miller, F.G., 2008. The ethics of placebo-controlled trials. In: Emanuel (Ed.), The Oxford Textbook of Clinical Research. OUP, Oxford, pp. 261–272.

Miller, F.G., Brody, H., 2007. Clinical equipoise and the incoherence of research ethics. Journal of Medicine and Philosophy 32 (2), 151–165.

Miller, F.G., Brody, H., 2003. Therapeutic misconception in the ethics of clinical trials. A critique of clinical equipoise. Hastings Center Report 33 (3), 19–28.

Miller, F.G., 2003. Sham surgery: an ethical analysis. American Journal of Bioethics 3 (4), 41–48.

Miller, F.G., Brody, H., 2002 Spring. What makes placebo-controlled trials unethical? American Journal of Bioethics 2 (2), 3–9.

Millum, J., Grady, C., 2013. The ethics of placebo-controlled trials: methodological justifications. Contemporary Clinical Trials 36 (2), 510–514.

Millum, J., Emanuel, E.J., 2012. Global Justice and Bioethics. Oxford University Press, Oxford.

Nuffield Council on Bioethics, 2002. The Ethics of Research Related to Health Care in Developing Countries. Nuffield Council on Bioethics, London.

Petryna, A., 2009. When Experiments Travel. Princeton University Press, Princeton.
Pogge, T., 2002. World Poverty and Human Rights. Polity Press, Cambridge.
Resnik, D.B., 1998. The ethics of HIV research in developing nations. Bioethics 12 (4), 286–306.
Rid, A., Wendler, D., 2011. A framework for risk–benefit evaluations in biomedical research. Kennedy Institute of Ethics Journal 21 (2), 141–179.
Rothman, K.J., Michels, K.B., 1994. The continuing unethical use of placebo controls. New England Journal of Medicine 331, 394–398.
Sacks, H., Chalmers, T.C., Smith Jr., H., 1982. Randomized versus historical controls for clinical trials. American Journal of Medicine 72, 233–240.
Sackett, D.I., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S., 1996. Evidence based medicine: what it is and what it isn't. British Medical Journal 312, 71–72.
South African Department of Health, 2004. Ethics in Health Research: Principles, Structures, and Processes. Research Ethics Guidelines. www.nhrec.org.za/wp-content/uploads/2011/ethics.pdf.
Stone, M., Laughren, T., Jones, M.L., Levenson, M., Holland, P.C., Hughes, A., et al., 2009. Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration. British Medical Journal 339, b2880.
Temple, R., Ellenberg, S.S., 2000. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. Annals of Internal Medicine 133 (6), 455–463.
Uganda National Council for Science and Technology (UNCST), 2007. National Guidelines for Research Involving Humans as Research Participants. http://www.uncst.go.ug/.
Varmus, H., Satcher, D., 1997. Ethical complexities of conducting research in developing countries. New England Journal of Medicine 337 (14), 1003–1005.
Weijer, C., Miller, P.B., 2004. When are research risks reasonable in relation to anticipated benefits? Nature Medicine 10 (6), 570–573.
World Medical Association (WMA), 2013. Declaration of Helsinki. Fortaleza, Brazil. http://www.wma.net/en/30publications/10policies/b3/.

Plagues and Diseases in History

José L Peset, CCHS – CSIC, Madrid, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In spite of the development of the medical science, during the twentieth century, individuals have observed the spread of new or reemerging diseases, from plague, cholera, and flu; measles, cancer, and malaria; to acquired immune deficiency syndrome, West Nile fever, resistant tuberculosis, virus of Ebola, Creutzfeldt-Jakob disease, and others. As individual illness is rooted in society and the environment, human life is tied up with the history of main endemic and epidemic diseases. Human health is very sensitive and adaptable to changes, so the history of disease and hygiene is the core of the new ecological history.

Disease is the sting of death and death is the necessary victory of the grave. If according to Xavier Bichat, life is resistance to death, it may be said that life is resistance to disease. Disease is involved in the mechanism of adaptation – that is, of attack, equilibrium, and defense – of the living being to its habitat. Human development and evolution, in its social, cultural, and ecological setting, might give rise to alterations in the human organism, both in body and mind, against which it must struggle. The evolution of the species, society, and culture takes place within a given environment, and this environment may alter the pathological spectrum. The human being seeks Hippocratic and Aristotelian equilibrium but, in doing so, it is frequently subjected to changes and movements that endanger this quest. As the Hippocratic texts teach, human nature is part of sacred Nature, and human life is inserted in Nature's life. Similar concepts may be found in the East in ancient Chinese and Indian cultures and in the West perhaps in the pre-Columbian cultures.

This Hippocratic thought is now renewed by ecological history, and ecological movements, highly influenced by the history of disease and the history of hygiene. As individual illness is rooted in society and the environment, human life is tied up with the history of the main endemic and epidemic diseases. It is also tied up with the history of environment: according to the new ecological history, human society develops in a cultural, biological, and physical *milieu*, acting and reacting on each other. Lands where we live are increasingly modified by human action. We are, and we will be, living the changes produced by human action (deforestation, industry, mining, new energies such as radioactivity from warfare and power stations, etc.) on the environment (biological and physical). If we hurt the Earth, Nature may hurt us. Human health is very sensitive and adaptable to changes, so the history of disease and hygiene is the core of the new ecological history.

These changes, quite frequently, depend on the human being itself, on its customs and knowledge, its social groupings, or its settlements, which lead it to avoid malaria by settling in mosquito-free zones or to control smallpox by vaccination. On other occasions, such changes do not depend on the human being but arise in the environment or within the diseases themselves, which may appear or disappear, or simply change their characteristics, of their own accord, as was the case with the 'sweating sickness' (sudor anglicus). On yet other occasions, these changes are merely due to cultural appreciations, as it has happened with certain sexual practices such as masturbation or homosexuality, which used to be framed as stigmatized diseases. On the other hand, disease is not always considered only harmful: thus it may be considered a distinction of the Gods, as was the case of epilepsy in ancient times, or among some historical aboriginal tribes. Nevertheless, the Hippocratic text *On the Sacred Disease* established the natural condition of this illness, similar to other medical affections. The same disease was associated with the devil by the Christians, and historically, it had always been a supposed distinction of great personages from Caesar to Napoleon. Disease can also be considered a way toward perfection or transformations, leading to the creation of art or to the salvation of the soul, as melancholy or sorrow were for centuries.

Disease may also have political repercussions: it is claimed that George the Third's madness gave rise to problems for the British Crown, while Roosevelt's fragile health was considered to have placed him in a weak position in the Yalta negotiations. Socially, disease has sometimes unjustly been considered the stigma of groups (as acquired immune deficiency syndrome (AIDS) for homosexuals, and poor or marginal groups, or alcoholism for black or aboriginal peoples). Medicine and politics have endeavored to preempt social changes by means of the contested eugenic theories and practices of social engineering, inspired by Francis Galton at the end of nineteenth century, which began with recommendations or laws to avoid marriages entailing risks (real or imaginary), and went on to such processes as sterilizations. Heredity is today considered an important predisposition for disease, but mostly not in a deterministic way. Hereditary conditions are thus intertwined with cultural, social, religious, and moral ones. Besides, in contrast with the stark determination of former times, there is today an attempt to draw new optimism from the possibilities of genetic modifications. Yet, in this latter respect, there is fear of a future where such a formidable force could rest in the hands of the rich and powerful.

An important renewal in the study of infectious diseases was driven by bacteriological discoveries, and the theories about infection and immunity. In this respect, Hans Zinsser in *Rats, Lice, and History* combined them with human history, proposing biographical interpretations of epidemiological

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62050-0

history. The study of the transmission of pathogens through animals and human beings (as vectors and hosts), living in a physical, biological, social, and cultural environment, was crucial in bringing about a new history of disease and also much later in fueling the most recent ecological history. In the meantime, the essential contributions of historical demography and social history licensed the consideration of the history of diseases as 'biographies,' shared by human cultures, living beings, and natural environment. But with the development of accurate technologies of diagnosis, the real identification of old plagues is more and more demanding.

Ancient and Medieval Times

Paleopathology enables us to discover the high incidence of diseases in the early inhabitants of the Earth, especially where such diseases left their mark on bones. It is thus possible to detect remnants of tumors, infections, necrosis, osteoporosis, and malformations, as also traumatisms, dental alterations, rickets, rheumatism, and other diseases of the bones. Today, laboratory analysis allows the possibility of detecting remnants of other diseases, germs, or lesions, in animal or human remains, including of course the famous mummies. From the first settlements in the Fertile Crescent, changes have taken place in the relationship of man and animals with their environment and it is possible to establish an extensive catalog of diseases that have largely subsisted down to this very day. Particularly significant are the forms of settlement, the crops and hunting, and the contact of man with cattle, pets, and parasites, as well as the relations between peoples, through mixing and trade, wars, and migrations.

Malaria originated in tropical Africa, accompanying the dawn of humanity, from plasmodia infecting animals and hominids. Caused by different plasmodia (*Plasmodium vivax* and *Plasmodium falciparum* among others, with different geographical distribution), its relation with human populations is shaped by migrations and deforestation, hunting and farming, climate and soils, crops, animals and foods, and a complicated immunity resulting from ancient genetic mutations and new infections. Thus, with the early human emigrations, it expanded through Eurasia, and in the third millennia BC, malaria had already set in the early civilizations in marshy areas where the water, the climate, and the crops provided the conditions for the *Anopheles* mosquitoes to infect in crowded populations. Evidence for the existence of intermittent fevers is already to be found in ancient cultures, finding suitable conditions along the Yellow, Indus, Ganges, Euphrates, or Nile Rivers.

Different interpretations considered the disease to originate from climatic or environmental factors, from particles or small animals, because of spirits or gods, or from unbalanced or disharmonic alterations. The connection between fevers, splenomegaly, and marshy areas appears in the Hippocratic texts, such as *On Airs, Waters and Places*. Deforestation, agriculture, temperature, and a greater population density, as well as military and commercial movements, favor the spread of the disease. In Italy, in Republican times, the disease becomes acute; and it was recommended that the marshy lands be either abandoned or sold and that people should dwell on high ground – the latter being traditional advice and here defense reasons also counted. Also, the cleaning and draining of swamps was considered, this being the origin of systems of engineering sanitation. The spread of malaria got worse and worse in Rome owing to negligence and flooding, climate and agricultural changes, wars, travels, and migrations, and it reached its zenith at the time of the fall of the Roman Empire, a cataclysm to which it may have contributed. But when the old Empire declined, another frightening ghost traveled through the Mediterranean Sea.

Plagues were considered in ancient cultures as diseases originating in God's punishment, with a violent pattern, and quick and terrible diffusion, affecting and killing a large number of persons. Destructive epidemic catastrophes were narrated frequently in Mesopotamian, Biblical, Egyptian, Indian, or Chinese sources, and in 430–426 BC, Thucydides describes the plague of Athens in the war against Sparta, but it is doubtful that this case refers to the bubonic plague. This epidemic death is considered the beginning of the decline of Athenian hegemony, the crisis of the democracy, and culture of the Pericles era. Fear, war, siege of the city, and its maritime port were accompanied by the death of the great ruler Pericles. Later on, the first reliable description of this disease is that given by Rufus of Ephesus in the AD first to the second century, in the epoch of Trajan; and in the sixth century, the Plague of Justinian initiates the first great epidemic cycles of bubonic plague. Soldiers and merchants, animals and merchandises, and slaves and prisoners all contribute to linking up the extremes of the known world. A consequence of this contact was the Antonine Plague in the second century. This plague –and others – is attributed to smallpox, a disease that had already existed as far back as 3000 years ago, as certain mummies show.

In the Middle Ages, leprosy spread widely, allegedly as a result of the increasing East–West relations trough trade, travels, and wars. This biblical and present day disease is accompanied by social repulse and malignant connotations since dirtiness and overcrowding provided the conditions for the spread of the otherwise not extremely contagious *Mycobacterium leprae*. In this respect, the way to Santiago de Compostela where pilgrims slept in groups and under unhealthy and dirty conditions provided a fertile ground for this disease. Another markedly religious character is to be found in the 'Sacred Fire,' 'St Anthony's Fire,' or ergotism; a disease that developed in people eating rye infected by ergot, it was soon prevented, but reappeared during the course of severe famines.

These diseases are linked to many others that owed their spread to squalor and poverty, for example, parasites and infections, scabies, mycoses and lice, anthrax and ophthalmia. Natural catastrophes, wars, poverty, and famines were escorted by mental diseases, tuberculosis and pneumonia, traumatisms and poisonings, diarrheal diseases and fevers, as well as smallpox and measles. In the Mediterranean basin, malaria continued to be endemic with the population tending to resort to residing on the hills and mountains, far from marshy lands. The extension of malaria involved south and southeast of Asia, Central China and Japan, and also North Europe. The plague set in the growing cities with the burden of rats and fleas, making this disease an explicative example par excellence of epidemic history. A second epidemic cycle began with the Black Death in 1348. Commerce and the cities collapsed, giving rise

to the origin of the modern European kingdoms. At this time, Venice established control over persons and ships, in order to avoid the spread of plague, arriving from distant lands. Dark Ages are ending, and New Worlds are opening in culture and geography.

Modern Times

European expansion meant a sudden traffic of diseases mostly toward the New World. The discovery of America, with its scattered population devoted to agriculture, where there were no dirty cities, and where the inhabitants were devoid of immunity to the diseases of old Europe, produced an enormous demographic crisis. Certain diseases, such as smallpox (and measles) ravaged the indigenous population; this very contagious and harsh illness was introduced in the Caribbean and Mexico, and destroyed American indigenous populations. Syphilis came supposedly from America (although other human treponematoses were present in other continents), brought back by the Spaniards and it caused havoc since Renaissance in Europe and soon all around the world. Troop movements spread an unknown disease called 'tabardillo,' or 'typhus exanthematicus,' bringing about serious suffering for armies and poor people. Sea voyages led to the development of scurvy, due to the lack of fresh food.

Other diseases also voyaged overseas, such as yellow fever; although it originated from Africa, due to the commerce of slaves and merchandise, special temperature conditions and mosquitoes as vectors were required for the contagion. In the seventeenth century, it took root in Brazil and also spread to the Caribbean and throughout America, reaching New York and Boston and likewise Colombia, Ecuador, and Peru. It affected warm America and Europe, through the Iberian Peninsula. Immunity and cutoffs in commerce stalled the disease, but it remains endemic in tropical areas of America and Africa. Also, malaria arrived to America with the European conquest in sixteenth century, beginning in the Caribbean and Central America and soon spreading to South and North America, becoming endemic in hot, wet, and low lands. Malaria was combated since the seventeenth century by the use of cinchona bark (*Peruvian bark*), found in Peru, as an indigenous medical practice. Later on, the quinine – the alkaloid obtained from the bark in French laboratories – gave rise to rich industry and trade. Trees were cultivated in British and Dutch colonies, mainly in Java. After being taken in the Second World War by the Japanese army, synthetic products such as atebrine and chloroquine were obtained. Getting worse with agriculture, mining and livestock changes, settlements, and slavery trade, malaria became – together with yellow fever – a scourge along the warm and hot reaches. Throughout the American conflicts between European empires, and later during the American Independence Wars, both diseases played an important role, infecting and killing nonimmune soldiers.

Plague slowed demographic growth in Europe since the wars between modern and powerful nations, the crowded and dirty cities, and the developing trades impeded protection against the disease. The great writer Daniel Defoe remembered the terrible contagion in London in 1665 in *Journal of the Plague Year*. After the great European plague of Marseille in 1720, the relative peace of the eighteenth century allowed Western Europe to set up sea and land defenses, with Austria becoming a solid bulwark in the face of the Ottoman Empire, thus sealing off all possible spread of the plague. Toward the middle of the nineteenth century, the third wave of the disease broke out and, with the exception of Europe, it spread to all countries including Asia, Africa, and paradises like America and Australia, leaving remnants in many places. The eradication of plague in Europe was a result of the advent of public hygiene, immunity to disease, and the disappearance of rats and of the old and dirty wooden buildings. The great London Fire in 1666 'purified' the city, since the hygienic rebuilding was carried out in stone, eliminating animals and dirt.

The disappearance of plague from Europe was followed by the outbreak of cholera, which had been endemic in India for centuries. This disease was described by Western travelers in the sixteenth century and its spread to Europe and America was a consequence of pilgrimages, trade, and a lack of cleanliness in water. A great pandemic broke out in 1817 from India and spread through Eastern Africa and Southern Asia, arriving to China and the Philippines, and in a second wave, it spread through Persia reaching Russia and Poland in 1830. Then this second wave swept across Europe and reached America by 1832. Throughout 100 years, terrible waves spread from the East. However, the timely and necessary cleansing of urban water supplies helped to progressively stall its advance. Nevertheless, as late as 1892, the free Hanseatic City of Hamburg suffered an outbreak of cholera epidemic, while the neighboring Altona, governed by the Prussian Reich, was able to avoid the disease, thanks to the successful filtering of its water. Local and national governments advanced in hygiene and they attempted to bring about healthier cities by means of appropriate public health measures. The nineteenth century sees the beginning of the demographic revolution in Europe, followed by developing countries, in America for example, with a sharp decline in mortality, especially maternal–infant mortality. Old inoculation and the new vaccination proposed in 1798 by Edward Jenner fought smallpox successfully. Studies about human immunity since the last decades of nineteenth century began a new medical approach to the prevention and understanding of illness, being successful in the contention of many diseases, which have a long record of burdens and deaths. Mother care also advanced both throughout pregnancy and at childbirth and during lactation.

The rise of great and industrial cities with low-grade outskirts and the emigration of peasants to the city favored the spread of maladies associated with poverty, hard work, squalor, lack of appropriate food, and exposure of people devoid of defenses. Some of the diseases were strictly occupational as in the case of miners and textile workers who were exposed to injury, and to industrial poisons that affected the proletariat. Sir Percivall Pott described in the eighteenth century the cancer produced in professional cleaners (chimney sweep) by chemical toxics contained in soot. Many other diseases, such as cancers, will be attributed to chemical and physical aggressions, including radioactivity. Other diseases were closely related to the harsh and dirty conditions of life such as spread of typhoid fever and pulmonary diseases. Tuberculosis spread over the turn between the eighteenth and nineteenth centuries, and while it was represented as the disease in fashion, affecting

notable and distinguished people, it struck the proletariat much more severely. The same can be claimed regarding the enormous spread of syphilis, and drug abuse, beginning with alcohol and continuing with cocaine and morphine.

Contemporary Times

The discovery of the microbiological origin of infectious diseases, and of effective therapies against them, and the development of public health changed the pathological landscape in developed countries. The long way between Ignaz Semmelweis and Alexander Fleming arrived to the contention of infections in health care. The twentieth century marked the descent of the high mortality rates in countries that reached high standards of sanitary development, investing in health care and public health. Such advances have been due to public health services, hospitals, antibiotics, surgery, and vaccination, and they have been reinforced by the developments in immunology and microsurgery, pharmacology, and biotechnology and with the promising future of genetic engineering. The international health solidarity promoted by the World Health Organization (WHO), and other governmental and nongovernmental organizations (NGOs), including church missionaries, were accompanied by better governance of nations, and internationalization of information and resources. The twentieth century established a serious change in geopathology of diseases. Smallpox is the first disease to be considered totally eradicated with only some samples of the virus being kept at a few laboratories for study purposes.

A Cuban doctor (Carlos Finlay) found the method by which yellow fever is transmitted through mosquitoes (*Aedes aegypti*). After the Cuban War and during the opening of the Panama Channel, yellow fever and malaria were studied and faced by US army and American sanitarians (Walter Reed, William Gorgas), and later by the Rockefeller Foundation. Impeding mosquito reproduction and avoiding bites and spreading chemical products were useful. Fortunately, an effective vaccine against yellow fever was later discovered. During nineteenth century, malaria expanded all around the world, arriving to the central extensions of America and Eurasia. Emigrations and settlements, wars, famines, climate change, and revolutions in travels, such as railroads and steamships, contributed. During the Napoleonic Wars and the American Civil War, malaria was seriously extended, as it was also during wars and revolutions in twentieth century.

In the interwar years, the Rockefeller Foundation and the League of Nations sponsored international campaigns against the disease. Many governmental campaigns also fought against malaria; several national programs were effective, from Taiwan and China to the United States, Brazil, and Argentina, passing across the Mediterranean basin, from Italy to Egypt. Opinions and campaigns oscillated between quinine treatment, fighting against mosquitoes (by dichlorodiphenyltrichloroethane (DDT)), and improvements in life, education, and land sanitation, clearing up the marshy areas. Malaria has been eradicated from Europe thanks to drainage, improvements in crop cultivation, mechanical barriers, quinine and modern drugs such as atebrine and chloroquine, and insecticides. Colonial settlements entered tropical areas backed by quinine, nets in beds, hygiene, and sanitation, but transmissible diseases were continuously a serious burden for colonial armies, in America, Africa, or Asia. Some chemical products such as DDT were a successful support for soldiers, travelers, merchants, or settlers. Nevertheless, in 1955, the World Health Assembly, meeting in Mexico, warned against the resistance of the mosquitoes to insecticides. Mosquitoes are still today carrying both diseases, and also dengue. Chagas disease produced by *Trypanosoma cruzi* is endemic in America, and sleeping sickness caused by *Trypanosoma brucei (rhodesiense* and *gambiense)* occurs in Africa.

Today, an effort is being made to involve governments in the fight against malaria, a disease that represents a danger to nearly half of the world's population. In Africa, the situation is very serious due to the changes in agriculture and irrigation as well as in work and migrations, the political and economic problems, the severe droughts, famines and wars, the bad sanitary conditions following in the wake of housing expansion and deforestation, and harsh social and economic exploitation. Together with the WHO and the United Nations International Children's Emergency Fund (UNICEF), regional institutions such as the Pan American Health Organization and the US government are financing programs of eradication. The WHO, through the 51st World Health Assembly of 1998, set up the program 'Roll Back Malaria' and, with aid from UNICEF and the World Bank, is endeavoring to bring about economic and sanitary improvements in developing countries. In twenty-first century, The Global Fund to Fight AIDS, Tuberculosis and Malaria is attaining important amount of resources and success. Nets treated with insecticides are very useful for protection against mosquitoes. Today, hope is also placed on vaccination and in the sterilization, or genetic modification, of the mosquitoes. Yet, malaria continues to be endemic in warm zones of America, Asia, and with severe cruelty, in Africa.

The rapid increase of world population in huge metropolitan areas was accompanied by new settlements and emigrations, wars, conflicts and revolutions, and marginalization of aboriginal cultures and disinherited peoples. Throughout the past two centuries, economical, social, and political expansion of the West led to studies on tropical medicine, creating hospitals and laboratories and institutions on public health and sanitary departments. Cholera continues to be feared in Asia, between India and the Far East, and also in warm zones in America and Africa. Wars and catastrophes, travels and migrations, as well as famines and unhealthy conditions allowed the expansion of the cholera germ, the *Vibrio cholerae* isolated by Robert Koch in 1883, as was the case of the recent tragedy of Haiti. At the end of nineteenth century, the third outbreak of bubonic plague allowed its bacteriological and epidemiological description. The germ of bubonic plague *Yersinia pestis*, discovered by Yersin and Kitasato, is still to be found throughout the world. The germ is carried by rats, *Rattus rattus*: rat fleas are the arthropod vectors transmitting 'epizootic' plague to humans from rodent hosts, and related species, but transmission among humans is also possible. Almost forgotten in the developed world, with several and very effective antibiotic therapies, there is, nevertheless, always the possibility that the devil may once more send his rats to the old, rich cities.

But other viruses have taken its place in the twentieth century. Thus the very old and common influenza caused several cruel and heavy outbreaks; among many others,

influenza pandemics began in 1918 and reappeared in 1957, 1968, and 2003: these were known, respectively, as the 'Spanish flu,' which perhaps originated in the United States, the 'Asian flu,' the 'Hong Kong flu,' and the 'Avian flu,' mostly coming from Asia. Vaccines are very useful in its prevention, as they are also in the control of many infectious childhood diseases, such as measles, rubella, mumps, chicken pox, whooping cough, diphtheria, and so on. Old diseases such as poliomyelitis exploded in the United States and Europe, affecting nonimmune children, and this disease has only been controlled in developed countries by means of vaccination. During the past century, individuals have observed astonished the spread of new or reemerging diseases, from plague, cholera, influenza, measles, and malaria, to AIDS, West Nile fever, Avian flu, severe acute respiratory syndrome, resistant tuberculosis, hemorrhagic fever (Ebola virus disease), transmissible spongiform encephalopathy (Creutzfeldt-Jakob disease), and others. Unfortunately today, many frequent or rare diseases do not have adequate treatment, and many germs are developing resistance to antibiotics, a serious threat for its affectivity.

AIDS has meant a development in the study of viral infections, and this disease has associations with the exploitation of poor people, unsafe sex, and drug trading, but it also affected an elite sector, which has sprung rapid research in the field. The world commotion surrounding this disease has served both to reveal human altruism on the one hand, and, on the other, contempt toward those infected by the disease, since AIDS became more and more the lot of the poor, mostly in large areas of Africa, or India, and of downcast or marginal groups such as the chronic patients, drug addicts, prostitutes, and homosexuals. Nevertheless, heterosexual transmission through unsafe sex and transmission from mother to child are today serious dangers. Safe sex and antiretroviral treatment are the best contention, while a vaccine will be obtained perhaps in the coming future.

Mental affections were considered till modern times, as devilish, criminal, or vicious behaviors, more an ethical or social than a medical problem. In the eighteenth century, the natural explanation of mental disease was established, according to Alexander Crichton, or Philippe Pinel, and in the nineteenth and twentieth centuries, its psychogenic process, following Sigmund Freud. From Pinel to Freud, the possibility of treatment and remedy of mental disease was established, leading the soul of the patient with convictions and health measures to the cure. The discovery in recent times of some effective pharmacological drugs acting on human mind, and conduct, has allowed better treatment. Reclusion was considered a doubtful possibility, restricted to some severe problems. The old lunatic asylums, founded since Middle Ages, were being abolished or completely renewed. Nevertheless, mental illnesses wreak serious havoc at present times all around the world, affecting all ages, genders, and conditions, without any distinction. Sometimes, the misunderstanding regarding mental diseases still produces cruel treatments, harsh restrictions, or punishment.

During the past decades, economic and social improvements led to quick globalization and urbanization, with longer human life and changes in disease patterns. Political, sanitary, and economic development managed to stamp out infectious morbidity, and forms of suffering or living illnesses are changing. UNICEF and WHO, in collaboration with governments, foundations, and NGOs, promote successful campaigns of vaccination, especially focusing on children. The burden of infectious, parasitic, and transmissible diseases changed to chronic diseases and sufferings related with aging and lifestyles, likewise, nutritional disorders, cancer, or heart and brain vascular diseases. On the other hand, the increase in life expectation favors the development of Alzheimer and other chronic and degenerative diseases, neurological and muscular diseases, mental affections such as schizophrenia, mania and depression, diabetes, vascular diseases, and of course, the terrible presence of cancer. Accidents and traumatisms, due to traffic and sports, factories and radiation, or even home accidents, are something that the twenty-first century is inheriting. Blindness, deafness, dumbness, and other frequent physical and mental disabilities, with multiple origins in traumatisms or accidents, genetic, metabolic and degenerative diseases, infectious diseases such as poliomyelitis, or cerebral and vascular affections, make daily life difficult and, at the same time, make economical, institutional, and social support necessary. Incapacitating and chronic diseases lead to severe dependency and so the need for protection of disabled individuals is leading to the founding of patients' associations looking for help and justice, creating new rights and demands. Ghettos and migrations, hard work, poverty and unemployment, and some of the old drugs such as alcohol, heroine, and tobacco or new synthetic ones are causing havoc even to the rich world.

But in developing countries, the old morbidity due to transmissible diseases continues to exist, maintaining a very low life expectation. Certainly, the increase of migrations and of urban population are requiring more water, food, and energy, producing climate warming, deforestation and agriculture changes, and dangerous issues such as waste, toxics, and pollutants. Potable water, health services (medical care or medical drugs), sanitation and hygiene are urgently requested. Environmental degradation and the confrontations between national, social, and ethnic groupings are a serious danger for healthy life. China, Japan, and South Korea and other expanding countries have lived through these challenges in different ways from diverse British colonies such as Australia or India. Different traditions and cultures are extremely important to understand the relations between peoples and diseases. Death is most rampant amid the least protected and poorest people, especially women, children, and the elderly, with harsh sufferings such as wars, famine, and exploitation. Under such circumstances, diseases caused by deprivation and infection continue to be the most rampant, causing high death rates among the population. Old and new diseases, such as malaria and AIDS, are ruining great stretches of Africa. Leprosy and cholera, tuberculosis and tetanus, and many child diseases, such as mumps, measles, tetanus, or meningitis are in poor countries and populations the salt of the Earth. Private and public funds, and international solidarity, are always necessary for relieving these harsh sufferings. The United Nations Millennium Development Goals are also facing them.

According to the WHO report *Global Health Risks* (2009), the better or worse conditions for mortality and for the burden of disease are seriously conditioned by several circumstances, related to lifestyles: blood pressure, blood glucose, physical activity, alcohol and tobacco, weight, safe sex, safe water,

sanitation, and hygiene. Obviously, these circumstances and their consequences are very different, depending on the social level of individuals and the public health governance of nations and peoples. If developing countries are freeing themselves from transmissible diseases, now they are fighting against noncommunicable diseases, related to social level, health organization, and hygienic customs. Some American, Asian, and African countries are still supporting a terrible burden of disease, which is also shared by low-income population in developed countries. The Rio Political Declaration in the World Conference on Social Determinants of Health (WHO-Rio de Janeiro, Brazil, 21 October 2011) is a new call looking for equity, justice, and universality of health. It has recommended adopting better governance for health and development; promoting information, justice, and participation in policy making and during the implementation process; including civil society like indigenous people; and reducing health inequities in the health sector. This Declaration promotes research on the relationships between social determinants and differences (economic, ethnic, and gender inequalities) and health equity. We are all convinced about the relation between poverty, social discrimination, low education and low sanitation, and diseases and death.

See also: Environmental History; Freud, Sigmund (1856–1939); Galton, Sir Francis (1822–1911); Historical Demography; Medicine, History of Western; Science, History of; Welfare State, History of.

Bibliography

Aberth, John, 2011. Plagues in World History. Rowman & Littlefield Publishers, Inc., Lanham, Maryland.

Arrizabalaga, Jon, Henderson, John, French, Roger, 1997. The Great Pox: The French Disease in Renaissance Europe. Yale University Press, New Haven, London.

Biraben, Jean-Nöel, 1975–1976. Les hommes et la peste en France et dans les pays européens et méditerranéens, 2 vols. Mouton & Co, École des Hautes Études en Sciences Sociales, Paris, La Haye.

Corbellini, Gilberto, Merzagora, Lorenza, 1998. La malaria tra passato e presente. Università di Roma La Sapienza, Università di Cassino, Roma.

Crosby Jr., Alfred W., 1972. The Columbian Exchange: Biological and Cultural Consequences of 1492. Greenwood Press, Westport, Connecticut.

Desaive, Jean-Paul, Goubert, Jean-Pierre, Le Roy Ladurie, Emmanuel, Meyer, Jean, Muller, Otto, Peter, Jean-Pierre, 1972. Médecins, climat et épidémies à la fin du XVIIIe siècle. Mouton & Co, École Pratique des Hautes Études, Paris, La Haye.

Kiple, Kenneth F. (Ed.), 1993. The Cambridge World History of Human Disease. Cambridge University Press, Cambridge, New York, Melbourne.

Lewis, Milton J., MacPherson, Kerrie L. (Eds.), 2008. Public Health in Asia and the Pacific. Historical and Comparative Perspectives. Routledge, London, New York.

Little, Lester K. (Ed.), 2007. Plague and the End of Antiquity: The Pandemic of 541–750. Cambridge University Press, The American Academy in Rome, Cambridge, New York, etc.

McNeill, John R., 2010. Mosquito Empires: Ecology and War in the Greater Caribbean, 1620–1914. Cambridge University Press, Cambridge, New York, etc.

McNeill, William H., 1976. Plagues and Peoples. Anchor Press.

Nash, Linda, 2006. Inescapable Ecologies: A History of Environment, Disease, and Knowledge. University of California Press, Berkeley, Los Angeles, London.

Peset, Mariano, Peset, José Luis, 1972. Muerte en España. Seminarios y Ediciones, S. A., Madrid.

Webb Jr., James L.A., 2009. Humanity's Burden: A Global History of Malaria. Cambridge University Press, Cambridge, New York, etc.

Zinsser, Hans, 1935. Rats, Lice and History. Printed and pub. for The Atlantic Monthly Press by Little, Brown, and Company, Boston.

Planning and Development

Vanessa Watson, University of Cape Town, Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article examines the two disciplinary fields of development studies and urban planning to consider what they have in common and why there has been so little connection between the two. One reason is that the focus of development studies has been on social and economic change in global South regions of the world, while planning has focused on global North regions and has been less interested in how these ideas 'travel' to the global South. The article uses public participation to illustrate how this divide between the two fields has come about.

Introduction

The article will bring together two areas of theory building, publication and practice, which, over the last several decades, have had little to do with each other: planning and development studies. It is somewhat surprising that there has been so little connection between the two fields. In broad terms, both are interested in societal change and improvement, although they have somewhat different areas of focus. The following definition of development by Beall and Fox (2009: 5) captures this sentiment:

> Thus development is fundamentally about change, and in particular progress and differentiation; development studies is an academic field of enquiry that seeks to define, critique and understand progress and differentiation; and those engaged in the practice of development seek to affect change through public action.

Beall and Fox (2009) go on to offer their own definition of progress and development, which draws on Amartya Sen's idea of freedom as both the means and end of development, and progress as the expansion of freedom to human populations across the globe. To this they add the notion of development as sustainable, implying that expanded freedom must happen in a way that does not threaten the freedom of present and future generations.

Turning to planning, a recent debate among planning scholars in various parts of the world produced a definition of urban planning for inclusion in the UN Habitat (2009: 19) Global Report on Human Settlements that makes the following points:

> Urban planning is … currently viewed as a self-conscious collective (societal) effort to imagine or re-imagine a town, city, urban region or wider territory and to translate the result into priorities for area investment, conservation measures, new and upgraded areas of settlement, strategic infrastructure investments and principles of land-use regulation …. At the core of urban planning is a concern with space … Planning also highlights a developmental movement from the past to the future … to decide between appropriate actions now in terms of their potential impact in shaping future socio-spatial relations. This future imagination is not merely a matter of short-term political expediency, but is expected to be able to project a trans-generational temporal scale, especially in relation to infrastructure investment, environmental management and quality of life.

As well as both being interested in social change and improvement, and environmental sustainability, the two fields have other aspects in common. Both are interdisciplinary in their own right: they draw on, and synthesize ideas from, a wide range of other disciplines such as geography, anthropology, economics, sociology, political science, and others in order to understand the context or problem at hand. They are also both concerned with the problem of linking knowledge and action, that is, how to translate the understanding of a context into policies, programs, and projects in order to bring about change and improvement. Both fields are also concerned with the functioning of cities. Planning, like development studies, occurs at the regional, national, and even the supranational scale, but most of the focus of planning is at the urban scale (which would include local neighborhoods within a city as well) and hence the focus in this article is on urban planning. In the case of development studies, the interest in urban development has been more ambiguous: at different periods in time cities have been sometimes seen as the engines of development and sometimes as a drain on national development resources. This will be discussed further below.

Perhaps the most significant difference, however, has been the geographical areas of focus of the two disciplines. The discipline of development studies has focused primarily on those parts of the world considered in need of development: these are the regions variously termed 'developing' or the 'global South' countries. Planning, on the other hand, emerged largely to address urban issues in 'developed' or 'global North' countries and from there the ideas were transferred and adapted to other parts of the world. It is perhaps this very different geographical focus that has resulted in the two disciplines somehow 'missing' each other, with scholars publishing in different journals and attending different conferences and hence not engaging with each other's debates and issues.

This article will discuss the formative influences in the planning and development studies fields and how both areas have engaged with urban issues. This will clarify how it has come about that the two areas with similar concerns have had little connection with each other. The section thereafter will focus on one area of research and action that has been central to both disciplines – that of participation or collaboration – in order to highlight how the two fields have approached the same issues differently.

Intellectual Influences

This section will periodize and locate the overarching formative influences on first, development studies, and then planning.

Development Studies

Beall and Fox (2009) explain how development thinking moved from a 'modernization' approach in the early 1950s to 'market-driven' approaches from the 1980s onward, along with a focus on governance and institutional issues in more recent times (Table 1).

Beall and Fox (2009: 8–19) trace thinking about development from its early roots at the time of the industrial revolution in Europe when debates centered on the nature of socioeconomic evolution and change. However, it was not until after the Second World War that the idea of development emerged as an international political agenda and a discipline in its own right. In this first era of thinking (1950s and 1960s), development was defined as 'modernization.' It was assumed that undeveloped or 'backward' economies could 'catch up' with advanced economies such as that of the United States given the right economic policies: Walter Rostow's thesis of '*Stages of Economic Growth*,' in which undeveloped countries could progress from a traditional, low-productivity economy to one of the capitalist high mass consumption, was highly influential. By the 1970s, these optimistic and confident development strategies were subject to growing critique. The 1960s economic boom had benefited many countries considered undeveloped or developing, yet in most there was little sign that these benefits had 'trickled down' to the poor. The emergence of a powerful dependency theory critique (associated with the writings of Andre Gunder Frank) suggested that poorer countries were undeveloped because of the uneven nature of the international capitalist system, which actively underdeveloped the 'peripheral' (developing) regions of the world in order to develop the 'metropole' (developed regions). A further criticism of the modernist theory of development was that the focus on macrolevel economic indicators meant that a concern with the microlevel conditions of the poor had been lost. Contributors to the development field called for a more nuanced understanding of the poor and their livelihoods, a recognition of gender differences, and approaches to development that incorporated the voices of the poor, that is, that development should be more participatory. Finally, it was at this time that a new environmental lobby began to emerge, and pointed to the negative impact of a modernist growth path on the globe's resources and environment.

While modernism and its critiques were directed primarily at the economies of the 'developing' world, the new dominant position on development of the 1980s affected all parts of the capitalist world, although in different ways. The global economic crisis of the 1970s ushered in an approach to economic policy now known as neoliberalism, in which free market liberalism together with greatly reduced state involvement in national economies was seen as an ideal. Neoliberalism remains a dominant economic paradigm even to the present day, although the economic crisis of 2008 placed assumptions of inherent market efficiency in question and has reopened debates about necessary government regulation of economies.

In countries of the global North, neoliberalism heralded the start of a period of public service privatization, monetarism, and challenges to public welfare systems. Planning, in particular, was seen as unnecessary state intervention that held back economic progress and introduced market inefficiencies. In countries of the global South, free market policies were strongly advocated as the new solution to development. Through international agencies such as the World Bank and International Monetary Fund, poorer countries were offered grants and loans to address pressing debt issues with conditionalities attached that required the introduction of neoliberal policies. These conditionalities have been termed structural adjustment policies and involved privatization of state enterprises, fiscal policy discipline and reduced public sector employment, reduced tariffs on imported goods, facilitation of foreign investment, tax reform, promoting private property rights, and deregulation of domestic markets. These policies proved disastrous for most of those countries that adopted them, and the 1980s has since been labeled the 'lost decade' of development, with rising levels of poverty and unemployment and declining health and education conditions.

Many of these policies are no longer viewed as being capable of promoting development and the nature of government and institutions was recognized as a central precondition for development. In both global North and South regions, the focus turned to the functioning of government and particularly its relationship with civil society and business. The term 'urban governance' was introduced to describe the kind of relationship between these 'stakeholders' that was needed to govern increasingly complex, globalized, and multiscalar urban and regional territories. Slater (2004: 98) has argued that the 'good governance' agenda is merely a shifting conceptual emphasis within neoliberalism, rather than a new approach to development. This agenda was shifted again in the

Table 1 Key trends in mainstream development theory, policy, and practice

1950s–1960s *Modernization/developmentalism*	*1970s* *Dependency: empowerment critique*	*1980s–1990s* *Neoliberalism*	*2000s* *Institutions and governance/security and development*
Industrial expansion through investment and interventionist economic policies; poverty reduction through economic growth and residual approaches to welfare.	Growth with redistribution; increased emphasis on poverty reduction through supporting production and basic needs; participation and empowerment.	Structural adjustment: privatization, liberalization, and deregulation of markets; reduction of subsidiaries and introduction of cost recovery.	Focus on improving accountability, transparency, rule of law, security of property rights, bureaucratic efficiency, and state effectiveness; linking of security and development agendas.

Reproduced from Beall, J., Fox, S., 2009. Cities and Development. Routledge, London, Table 1.1.

mid-1990s, he argues, with the introduction of the concept of 'social capital,' referring to the 'social glue' that creates 'institutional thickness,' i.e., networked and linked formal and informal institutions that can be operationalized to support economic activity.

Other elements of the more recent approaches to development serve to reinforce the aims of the good governance agenda. Decentralization has been promoted in order to improve governance efficiency and participation (Slater, 2004), nongovernmental organizations have increasingly been drawn into partnerships around participation and service delivery, and the concerns of earlier critics of development regarding gender, participation, and environment have now become mainstream in development approaches (Beall and Fox, 2009). Crime and violence in cities have emerged as more recent developmental concerns (World Development Report, 2011) and are seen as central obstacles to the achievement of the Millennium Development Goals. Here again, local institutions are seen as playing a major role in addressing the issue.

There is an ongoing critique of current theories of development, sometimes termed 'postdevelopment' thinking. Many of the writers in this field have drawn on anthropological perspectives and research methods to show how development programs have little benefit for the poor or directly worsen their circumstances. Li (2007), in particular, theorizes the 'inevitable' gap that occurs between development intentions and outcomes, often brought about by the requirement to frame problems in terms amenable to technical solutions. There is also a postcolonial critique from those who argue that there is an ongoing relationship of domination between metropole and a previously colonial periphery, in which intellectual domination is a key element. Scholars writing from this position argue that there are elements of a modernist development perspective, which still underlie development thinking, along with ideas that Western liberal democracy offers a universal political ideal. Authors such as Connell (2007), writing from sociology, argue that intellectual domination from the global North in this and other fields needs to be countered by building new theory from and about the global South. This is a challenge that is beginning to take hold in a number of disciplines including planning.

Cities and Development

This section considers how thinking about the role of cities in development has shifted over time, as this is important in understanding what connections there might be to the field of urban planning. Beall and Fox (2009: 19) explain that at some times, cities have been seen as drivers of modernization and at other times as obstacles to development. Early modernization theorist, Arthur Lewis, associated cities with modernity, industrialization, and economic development, and rural areas with 'backward' traditionalism and agriculture. His theory that cheap labor would flow from rural to urban areas to stimulate industrial economic growth was also present in Rostow's Stages of Economic Growth thesis. Regional planner John Friedmann linked cities to economic and political development and suggested that the encouragement of urban growth in 'backward' regional peripheries could stimulate development.

At the same time, however, rapid urbanization in 'developing' countries was giving rise to a growing informal economy and shack settlements, both regarded as the antithesis of modernization and development. Suggestions that 'over-urbanization' could represent a development problem took hold, and by the 1970s, the role of cities in development was perceived by many as negative. Rapidly growing cities were described as 'parasitic' entities that absorbed public and private resources at the expense of the rural areas and the rural poor, who were condemned to ongoing poverty. The 'urban bias thesis,' as it was termed, had a significant influence on the nature of development thought and practice, which shifted to a focus on rural development and ways to curb urbanization.

The influence of the antiurban bias thesis continues today in many regions of the global South, although this seems not to be the case in China where state-led urban development is seen to support the export orientation of the economy. Elsewhere, debates continue as to whether development stimulus should focus on urban or rural/agricultural areas. Nonetheless, there is frequently a neglect of cities and urban issues in development thinking and many international donor agencies still favor rural development projects over urban ones. In some regions, it is a position favored by local politicians whose voters are rurally based, and it allows urban governments to blame the growth of informal shack dwellers and traders on undesirable urbanization. For example, in Zimbabwe in 2005, Operation Murambatsvina (also termed Restore Order, Cleanup, and Drive out the Rubbish) targeted vendors' structures, informal businesses, and homes labeled as illegal by the government, and frequently belonging to members of the political opposition. Conservative estimates were that 700 000 people were evicted from their homes and 2.3 million people were affected in other ways. Decades of rural development (one aim of which has been to slow the movement of people out of the rural areas), has nonetheless been paralleled by rapid rates of urbanization and growing cities in many global South countries. Latin America was 78.3% urbanized in 2007 and Africa and Asia were far less urbanized at 38.7 and 40.8%, respectively, but both currently have high rates of urbanization. It is anticipated that by 2025 all the 'less developed regions' together will be 53.2% urbanized (UN Habitat, 2009), and this proportion will continue to grow. Development problems in the form of poverty, unemployment, slum settlement, lack of services, and poor health are concentrating in these growing cities, yet the focus of the development field remains on the rural areas.

Planning

Modern urban planning (Watson, 2009) emerged in the industrializing world in the latter part of the nineteenth century as a very direct response to concerns of rapid urbanization, unhealthy and polluted living conditions for the poor, vanishing open green space, and threatened political upheaval as a result. Broadly, the ambition of planning was to produce urban populations that would lead ordered and disciplined lives, in healthy environments, and develop and improve both physically and spiritually. Ways of

achieving these aims were strongly influenced by the time and place within which they were formulated. Early British town planning responded to radical and utopian socialist ideas of the time and a nostalgic longing for the village life of medieval England. One of the most influential planning forms of the time, Ebenezer Howard's Garden City, represented an attempt to recreate this village life through bringing 'green' back into towns made up of winding roads and separate cottage residences, through strict separation of land uses, and through controlling the size and growth of the town. The objectives here were twofold: social – the re-creation of a traditional way of life that was essentially anti-urban but was seen as preferable to the 'chaos' of the industrial city, and aesthetic – bringing the beauty of the countryside into the towns in the interests of both physical and spiritual health (Taylor, 1998). In France, the ideas of architect Le Corbusier in the 1920s and 1930s established the ideal of the 'modernist' city, which came to be highly influential internationally and still shapes planning in many parts of the world. Le Corbusier held that society should be highly regulated and controlled and this was to be achieved through an ideal city form that was neat, ordered, and efficient. Slums, narrow streets, and mixed-use areas should be demolished and replaced with efficient transportation corridors, residences in the form of tower blocks with open space 'flowing' between them, and land uses separated into monofunctional zones.

The primary legal tool through which these 'master' plans were implemented was the development control system or zoning scheme, which originated in Germany and spread across the United States and Europe in the early part of the twentieth century. In the United Kingdom, the important 1932 Town and Country Planning Act carried forward ideas of master planning and development control, and provided a model for much colonial planning. Land use zoning usually carried with it a particular view of urban form (in keeping with the visions of the early planners) and was enthusiastically adopted by urban middle and commercial classes who were able to use it as a way of maintaining property prices and preventing the invasion of 'less desirable' lower income residents, ethnic minorities, and traders. Early planning visions and the legal tool of land use zoning were central instruments of urban modernization from the 1950s onward, where older working class and industrial areas in global North cities were demolished to make way for urban renewal projects; freeways, which cut through urban fabric to connect center and suburbs; and new housing estates on the urban edge for the displaced poor.

These urban planning ideas came to symbolize the 'modern city': they were enthusiastically adopted in many global North cities and were exported to global South countries where they aligned comfortably with ideas of economic modernization and 'catching up with the West' as development ideals. In some parts of the global South, the planning of urban settlements was bound up not only with the 'modernizing and civilizing' mission of colonial authorities, but also with the control of urbanization processes and of the urbanizing population. On the African continent, this diffusion occurred mainly through British, German, French, and Portuguese influence, using their home-grown instruments of master planning, zoning, building regulations, and the urban models of the time – garden cities, neighborhood units, and Radburn layouts, and later, urban modernism. Many African countries still have planning legislation based on British or European planning laws from the 1930s or 1940s, revised only marginally. The most obvious problem with these forms of urban modernism is that they fail to accommodate the way of life of the majority of inhabitants in rapidly growing and largely poor and informal cities, and thus directly contribute to social and spatial marginalization.

The field of development studies shifted in the post-1960s to consider other and different approaches to development. Urban planning in cities of the global North also adopted a range of new ideas and approaches in the post-1960s, in response to neoliberalism, changing ideas on the role of government and environmental concerns. Innovation took place in the following areas: planning processes and decision making (shifts toward more participatory, democratic, and integrated processes, involving wider groupings within and beyond the state), forms of spatial planning (toward more flexible and market-friendly strategic planning at a range of scales), linking planning and environment (new concerns of environmental sustainability, climate change, and resource depletion), and some new directions in land use management. However, in many parts of the global South, planning approaches changed very little (UN Habitat, 2009). There is now a significant volume of literature that argues that urban planning in many parts of the global South fails to address the primary urban issues of the twenty-first century and that in some regions it is directly implicated in worsening poverty and the environment. The strong antiurban approach that has permeated thinking in the development studies field has influenced the planning field as well, and has (in many parts of the world) meant a relative neglect of urban planning and development issues, and even the use of planning controls by some governments to justify the discouragement of urban in-migration.

Shifts in the global economy and new development ideas (promoted particularly by international aid and development agencies) have had some impact on Southern planning, with significant regional variation. Structural adjustment policies of the 1980s, and the related cuts in the public sector, no doubt made the operation of urban planning systems slower and less efficient, and rates of urban land release that do not keep pace with demand continue to present problems in rapidly growing cities. The encouragement of decentralization that accompanied structural adjustment has not improved planning systems either, although some countries have managed better than others. Frequently, the lack of professional capacity at local government level as well as a lack of resources has meant that planning decisions *de facto* remained centralized in national government.

International development agencies have made various attempts to shift approaches to planning and administration in global South cities. Regarded as one of the largest global urban programs, the United Nations (UN) Urban Management Programme was started in 1986 by the Urban Development Unit of the World Bank in partnership with UN Centre for Human Settlements and funded by UN Development Programme. The program involved some 120 cities in 57 countries, with the overall mission of promoting socially and environmentally sustainable human settlements and adequate

shelter for all, and reducing urban poverty and social exclusion. The success rate of the program was generally mixed, for example, in Tanzania it influenced the 'forward planning' aspect of the planning system, particularly through the concept of the city development strategy, but the critical regulatory aspect of the planning system continued as before. By 2006, impatience with the conservative, exclusionary, and top–down nature of planning emerged clearly at the UN Habitat World Urban Forum in Vancouver, where the UN Habitat Executive Director accused planning of increasing social exclusion in cities through antipoor measures. She suggested that urban planning should be fundamentally reviewed to see if it was able to play a role in addressing issues in rapidly growing and poor cities and called on planning practitioners to develop a different approach that is pro-poor and inclusive, and that places the creation of livelihoods at the center of planning efforts (Watson, 2009).

Hence, while planning and development studies, as two different disciplines and professional fields of practice, have a great deal in common in terms of the overall focus of interest and objectives, they have tended to operate in relative isolation from each other. The following section expands on an example of this disconnect by examining the issue of public participation, which both disciplines have engaged with.

Participatory and Collaborative Development and Planning

In the field of development studies, public and community participation in development projects in global South regions has been a focus of theoretical and practical attention since the 1970s (Table 1). Large-scale government and agency-driven modernization projects had frequently shown little benefit (or 'trickle down') for the poor, and this revived an earlier interest in local and community management and decision making. A new participatory development movement took off in the early 1980s, strongly influenced by the work of Robert Chambers and his rapid rural appraisal approach, later termed participatory rural appraisal. The focus on rural communities in this work was in line with development thinking of the time. The method is 'bottom–up' and involves local communities in collecting and sharing information about their context and development issues. The assumption has been that participatory development projects will be better designed, have better targeted benefits, and be more cost-effective and less corrupt.

The idea of empowering poor people to take control of their own development process was taken on by agencies such as the World Bank and frequently incorporated as a condition of development funding. However, over the years, there has been extensive critique of participation within the development studies field. These critiques have focused on both the practice of participation where, it is argued, the intention has often been project efficiency and cost cutting rather than empowerment, and on the concepts of participation, where it has been argued that the issue of power relations has been misunderstood or ignored, and that concepts such as 'community' and 'social capital' have been used in a highly generalized way, ignoring the ways in which context makes a difference to the meaning of these terms. Neoliberalism has introduced a strong push toward the marketization of public services and private sector involvement and profit has often taken precedence over community empowerment.

By the early 2000s, critical development scholars were suggesting that public participation had become 'tyrannical' in its effects and was simply another form of political control (in effect an 'antipolitics machine'). The 'postdevelopment' critique, which has gained strength in the 2000s in the development studies field, has been centrally influenced by Foucauldian concepts of power and by agency-driven approaches within the field of anthropology. There appears to be little faith any more in localized, state-managed, and project-linked participatory processes to deliver empowerment. The interest has instead shifted to networks of social movements and advocacy campaigns (such as the 'right to the city' movement and the World Social Forum), in which local community-driven development practices are up-scaled into wider global actions. The work of Slum Dwellers International, which links affiliated, community-based, shack-dweller organizations in Asia, Africa, and Latin America, is an example of this. Some authors take a specifically postcolonial perspective, which challenges Western-centric forms of knowledge production, with its claims of universality, and counters this through a focus on context-specific and indigenous knowledge and recognition of social diversity and difference. Some social movements are prepared to engage directly with institutions of the state in order to secure gains (for example, the participatory budgeting process in Brazil, or Slum Dwellers International's approach of 'coproduction'); others seek greater autonomy and distance themselves from state structures. Engagements with state structures are sometimes seen as a way of transferring or reclaiming certain state functions (such as budgeting or urban services) from the state to civil society and then reconfiguring them to fit local needs.

The theory and practice of participation has therefore shifted significantly in the development studies field. The focus has been on participation linked to international agency or state-initiated development projects in global South regions, many of which have been in rural areas. A strong critique of these practices has developed over the years, although they continue in various forms. However, there is now more interest in the development studies literature in bottom–up and endogenous forms of community engagement with development processes, some of which have taken on a global reach and significance.

In the field of urban planning, the issue of participation (termed participatory, collaborative, or communicative planning) has also been a central concern, emerging in the 1980s and remaining today as a dominant issue in the planning theory literature. However, there have been very few links or connections between the two disciplines on the topic of participation over the last couple of decades: few ideas or concepts have been shared and documented practical examples have rarely crossed disciplinary boundaries. Scholars in the two fields publish in different journals and attend different conferences.

In planning, the focus of research and publication on participation has been very much on how decisions are made and agreements are reached in government-centered planning processes in countries of the global

North. Usually, this involves a situation where professional planners employed by the state are called upon to develop a long-term urban plan or a shorter term change of land uses, and are required by law to consult 'the public' and take note of their responses to planning proposals. Another possibility is that a property developer needs to secure state approval for a land investment and any public objections to this need to be taken into account through a consultative process. Much of the interest in the planning theory field has been in the nature of these processes, how democratic and fair they might be, and how they are able to actually change plans and outcomes if 'the public' wishes to do this. These situations are therefore rather different from development projects carried out in parts of the global South, although the principle of participatory decision-making processes is common to both fields.

Theorists in the urban planning field have drawn inspiration from social philosopher Jurgen Habermas to pose communication as the most important element of planning practice. They also echo Habermas' faith in civil society as a source of democracy, and as a vehicle for placing pressure on the state to act more responsively. As a result, planning theorists have argued that interaction (with stakeholders or interest groups), communicating ideas, forming arguments, debating differences in understandings, and finally reaching consensus on a course of action should form the primary activity of planners. Significantly, however, these ideas are based on a set of assumptions about the nature of society and government within which such processes can happen: stable and organized civil society where communities are prepared to rationally debate the long-term future of their environments; the possibility that consensus will be reached given sufficient debate in which dominant interests can be prevented from unreasonably exerting their power; and stable and capable governments and professionals able to manage consensus-seeking processes and interpret the results into plans and projects that can then be implemented.

In many parts of the global North, it is possible for plan decision-making processes to be managed in this way (although there is a growing literature that points to the difficulties of these processes). However, as with many other planning ideas, these ideas about communicative planning are assumed to have universal relevance and the ability to be applied in any other part of the world. But problems arise when planners in very different global South contexts attempt to apply these theories. In many parts of the global South, governments are weak, lack professional capacity, and have highly centralized decision-making processes. Planning systems and laws are often only marginally reformed from the colonial era, and their top–down and control-oriented tendencies are highly incompatible with the idea of participation and collaboration. The predominantly poor urban populations, many living in informal settlements, are usually very keen to have a say in state policies and plans that affect them, but as the ordering and controlling objectives of planning are frequently in conflict with informality, achieving consensus is far more difficult. Poorer communities themselves are usually divided in terms of their interests (e.g., tenants vs landowners), may not have a long-term commitment to particular places, and may be unwilling to peacefully accept decisions not in their favor. Where participatory planning processes are instigated, they are often an attempt at co-option rather than empowerment, echoing the conclusions of critics of participation in the development studies field.

In fact, there is a great deal that planners in the global South can learn from the field of development studies and particularly from the critiques of participatory processes in this discipline. Development studies has engaged with the issue of state-society relations in development for far longer than has been the case in the planning field, and extensive experience has been built up in global South contexts and documented. Increasingly, global South planners are being urged to adopt new approaches to urban planning (UN Habitat, 2009), which are less top–down and controlling, and more facilitative, flexible, and developmental. Hence, the practical work of planning and development is becoming more closely aligned, and it makes great sense that the two disciplines should share experiences and ideas – in the area of participation and in other respects as well.

Conclusion

This article has argued that the two disciplines of planning and development studies, while having much in common in both practical and theoretical respects, have had little to do with each other over the past decades. In part, this is because development studies have focused on problems in the global South (i.e., on those areas considered to be in need of 'development'), and planning has developed its practice and theorizing primarily in global North contexts. In planning, these ideas have then been exported to global South contexts where they may have succeeded in part in controlling and ordering urban development for urban elites, but their impact on the majority of the urban poor has been highly negative, resulting in sociospatial exclusion and marginalization in urban areas (UN Habitat, 2009; Watson, 2009).

The article has explored in more detail how the two disciplines have approached the issue of participation in development projects and urban planning. Both disciplines have been centrally concerned with this issue, yet both have developed theory and practice in different ways, using different conceptual informants and case studies. Development studies have a longer history of exploration of the field of participation but planning theory and practice has made little use of this work to inform its own theorizing. Certainly, those planning theorists interested in planning processes in global South cities would greatly benefit by drawing on the literature and debates from the development studies field. Moreover, as poverty and inequality increasingly affect cities in both the global North and South, the approaches of Southern planners to dealing with issues such as informality may well be of use to global North planners. Further, with the growing interest from planning theorists in countering the tendency to draw planning ideas only from the global North, and to also consider ideas that are informed by the contextual realities of global South cities, there is greater possibility that the two disciplines may begin to share ideas and experiences.

See also: Dependency Theory; Development Theory in Geography; Modernity.

Bibliography

Beall, J., Fox, S., 2009. Cities and Development. Routledge, London.
Connell, R., 2007. Southern Theory: The Global Dynamics of Knowledge in Social Science. Allen and Unwin, Australia.
Li, T., 2007. The Will to Improve. Governmentality, Development and the Practice of Politics. Duke University Press, Durham.
Slater, D., 2004. Geopolitics and the Post-colonial: Rethinking North-South Relations. Blackwell, Oxford.
Taylor, N., 1998. Urban Planning Theory since 1945. Sage, London.
UN Habitat, 2009. Global Report on Human Settlements: Planning Sustainable Cities. Nairobi.
Watson, V., 2009. 'The planned city sweeps the poor away…': urban planning and 21st century urbanization. Progress in Planning 3 (72), 151–193.
World Development Report, 2011. Conflict, Security and Development. World Bank, Washington.

Planning During Post-Socialism

Sonia A Hirt, Virginia Tech, Blacksburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article analyzes the state of public sector planning in countries that subscribed to a Marxist, state-socialist ideology during some part of the twentieth century, and especially those countries that comprise the former Eastern Bloc (today often referred to as the 'transitional countries'). Central economic planning was a defining feature of state socialism. With its collapse, this type of planning was abandoned. During the early years of postsocialism, all types of public sector planning, including urban spatial planning, were severely weakened as well. A tentative revival of urban planning can be only recently observed.

Thirty years ago, in 1985, nearly half the world comprised countries that regarded themselves, and were regarded by others, as socialist (Alternative terms are 'state socialist' or 'communist'. Without delving into the differences between these terms, for simplicity purposes from now on I will use the term 'socialist'.). In these countries, located on five continents (predominately Europe, Asia, and Africa), a single-party-dominated political apparatus subscribed to a Marxist–Leninist philosophy. Most urban land, large real estate and means of production belonged to the state and its subsidiaries in the form of various public agencies, from municipalities to workers' collectives. The stated goal of socialist regimes was to build a prosperous, egalitarian, and classless society. A planned expansion of the economy and a planned distribution of resources among the population were the major means through which this goal was to be achieved.

Yet, today, just a handful of countries carry on the socialist banner. In itself, this fact should amply illustrate that socialism's ostensibly noble goals fell short of realization. Even among those few countries where socialism remains the formal doctrine (e.g., China), capitalist elements of market-led production and resource distribution are now embedded into the socialist system.

This essay analyzes the evolution of public sector planning (from now on, 'planning' for short) in formerly socialist countries over the last quarter of a century: from the time when the 'velvet revolutions' spread down throughout Eastern Europe until today (i.e., 1989–2014). The discussion is based primarily on the experiences of the former Eastern Bloc, which comprised the Soviet Union and its East European satellites. Today, after the breakup of the Soviet Union and the two other Eastern Bloc federations, Czechoslovakia and Yugoslavia, the region includes 28 nations situated over a sixth of the world's landmass, with a total population of about 400 million people. These nations are often referred to as 'transitional' (Hirt and Stanilov, 2009).

Planning during Socialism

Marxist–Leninists regarded capitalism as inefficient and exploitative and, therefore, fundamentally flawed socioeconomic order that contained the seeds of its self-destruction. Capitalism's alternative, state socialism (in theory, later to be followed by communism) was to ultimately triumph in the world for a number of reasons. One of them was the introduction of systematic state planning of the national economy, which would efficiently and justly guide both the production and distribution of resources, thus ostensibly avoiding capitalist-style inequality and periodic economic crises. Prior to the advent of socialism, none of the countries in the region had a record of such planning. In the Soviet Union, the first Five-Year Plan covered the years from 1928 to 1933. Similar plans were adopted in the East European countries, once these countries converted to state socialism after World War II.

The nationwide plans were the poster child of Soviet-style governance. They were prepared by armies of technocrats and stamped by the states' top legislative bodies. They comprised the tip of a strictly hierarchical system of hundreds of lower-level economic plans, which translated the national goals into priorities and strategies for all other administrative divisions (e.g., republics, provinces, cities, towns) and all economic units (e.g., factories, agricultural collectives). This extraordinary scope of the planning activities was only possible because the governing communist parties had nearly full political control over their populations, and because most urban land, large real estate, and means of production were in state ownership. The planning system was heavily centralized and highly bureaucratic. Local, municipal planning was subordinate to national economic planning. The main function of local planners was to act as technical translators of the higher-level economic goals into the physical layout of cities. For example, if state economic goals were to increase a certain type of industrial production, these goals would have typically been translated into specific mandates for the building of new factories in various cities and regions, which would in turn require that new workers be brought from the countryside. Local planners would receive instructions on expected industrial and population growth. Their job would be to plan and design the new factories, infrastructure, housing, and services. Economic and physical plannings were not fully integrated in the sense that there was little negotiation between the economic planners (who operated at the national or republican level) and the physical planners (who operated at the local or municipal level). The ones in the former group held much greater powers, although they too were subordinate to the political apparatchiks.

Land-use projections were typically conducted at the local level for larger cities and at the republican level for small towns. Local forecasts had to comply with the economic objectives and projections and the spatial standards adopted at the state level; e.g., standards for green space and dwelling space per person (on how the Soviet-era planning model worked, see for example Andrusz, 1984).

The Transition Period

The communist regimes collapsed in a domino-wave fashion throughout the region between 1989 and 1992. The immediate effect was a major political and economic crisis and, in some countries, bloody ethnic wars and national breakdown. From 1989 to 1993, the East European countries lost about a quarter of their aggregate GDP; in the Soviet successor states, aggregated GDP contracted by nearly half. Most countries of the region saw a rise in unemployment rates exceeding a quarter of the population, an explosion of inflation rates into the double and triple digits, a collapse of government-provided safety nets, and a precipitous decline of living standards for the majority of the populations. Social stratification increased sharply as measured by the increases in the Gini coefficients of income distribution throughout the region. On several accounts, the Eastern Bloc meltdown was the deepest and widest economic crisis to occur in peaceful times in modern history. Its causes are well documented and include the institutional vacuum leftover after the collapse of the one-party regimes, the obsolete state of the socialist-era industries, and the erosion of human capital (Dobrinsky et al., 2006). Political reforms and economic recovery followed at varying speeds, with some countries, especially in Central Europe and the Baltic region, leading the pack and joining the European Union (The following former Eastern Bloc countries have joined the European Union: Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, and Slovenia.), and others still mired in poverty and unrest (Havrylyshyn, 2007). In most countries, there was a general consensus that the end-of-state socialism meant the beginning of reforms leading to market economies and democratic governance; hence, the term 'transition' – that is, a pathway from socialism to capitalism – which is commonly applied to the postsocialist period. But interpreting what market economy or what democracy means – and how to implement them – has varied significantly from country to country. In each, it has been shaped by history and traditions (e.g., see Stark, 1991 on 'path dependence'). Since strategies and end results have been so different, with some countries embracing a neoliberal version of capitalism (e.g., those in the Baltic region), others retaining elements of social-democratic welfare states (e.g., Slovenia) and third continuously exercising strong state control over resources (e.g., Russia, Kazakhstan), some authors have problematized whether the term 'transition' is accurate or applicable to all states region wide (on 'theorizing transition', see Pickles and Smith, 1998). Perhaps not all countries can be described as 'transitioning' toward capitalism; perhaps at least for some 'transformation' is a better term (on the diversity of postsocialisms, see for example, Bohle and Greskovits, 2012).

Privatization and Institutional Transformations

Regardless of the multiplicity and varying speeds of the post-socialist projects, some level of privatization of public resources was the transformative process common to all countries in the region. Privatization – the transfer of assets and functions from the public to the private sector – could well be described as the 'leitmotif' of the transition (Bodnar, 2001; Hirt, 2012). In the immediate aftermath of socialism, new constitutions guaranteed the sanctity of private property and mandated the privatization of most urban land, real estate (including almost all dwellings) and means of production, as well as many social services. Some assets were sold; others restituted to those private parties and their successors who could prove ownership predating socialist-era nationalization. In most countries, privatization and restitution were carried out during the early and mid-1990s. However, the process of restituting urban land was highly problematic. This is because often multiple private parties submitted claims to ownership and because much of the urban land was built upon during the years of socialism and thus could not be restituted back to its earlier owners without first demolishing the buildings located on it (which of course was impossible). In addition, the restitution of real estate and means of production was often characterized by corrupt dealings (e.g., large industrial enterprises were sold for pennies to former party apparatchiks). This process greatly contributed to the stratification of postsocialist societies.

Accompanied by the reintroduction of market tenets of production and consumption, privatization, and restitution fundamentally transformed economy and society, including the mechanisms of urban development and the planning process. Whereas public authorities had a near monopoly on urban development during the socialist years with the state and its subsidiaries acting as both urban planners and developers, the new paradigm of urban development involved intense competition for land and other resources between plural actors: property owners, private developers, builders, interest groups, political alliances, and nonprofit organizations, along with governments at all levels and the broader citizenry (e.g., Tsenkova and Nedovic-Budic, 2006; Stanilov, 2007). A wholly new actor in the development process were the large, multinational, typically Western (but in some countries also Chinese, Indian, Israeli, Middle Eastern, etc.) firms that entered different countries in the region at different times and in many cases secured preferential treatment from national and municipal governments, ostensibly in the name of faster economic growth. With both private and public resources severely shrinking, especially during the early years of the postsocialist transition, the competition over urban land and resources was particularly ruthless. The weakened and often passive public sector and the newly impoverished masses usually ended up on the losing side.

Another important factor transforming the urban planning and development process was institutional decentralization. National-government involvement in planning matters was sharply reduced. The practice of central economic planning was eliminated. Virtually all spatial planning powers were transferred to local, municipal institutions, whose sheer number in most countries increased dramatically. For example, in Hungary, the number of political governments doubled during

the 1990s; in Bulgaria and Slovenia it nearly tripled (Tosics, 2005). This institutional decentralization was carried out under new laws for municipal self-governance and was seen as an integral part of the overall democratization process. However, as a consequence of the economic downturn, local governments had scarce resources to engage in planning activity. Furthermore, many had no experience with proactive planning, since during socialism they had only followed national directives. As a result, local governments ended up in the peculiar position of having to plan and implement their plans but with little financial and institutional capacity to do so (Hamilton et al., 2005; Hirt and Stanilov, 2009).

Planning's Legitimacy Crisis

The ideological climate within which planning operates also changed radically throughout the region. In the aftermath of state socialism, most new governments made a sharp turn to the political right. Neoliberal doctrines espousing the superiority of free-market capitalism put into question the utility of public sector planning. Many politicians came to view planning as an unwanted vestige of the old socialist system – a phenomenon referred to as a 'legitimacy crisis' of planning (Nedovic-Budic, 2001, Hirt, 2005). This legitimacy crisis affected not just central economic planning – the type of planning that had malfunctioned during socialism and the first to be let go after its collapse. Moral devaluation affected all sorts of government policy, planning, and regulation activity. Spatial and housing policies – a traditional public-policy area in most capitalist democracies – were no longer considered an appropriate activity of higher-tier levels of government. The assumption was that the market will guide the production and distribution of resources in more efficient ways that government planners ever could. If government intervention in urban development were needed, it should be conducted by municipal governments, which are much closer to the real-life issues in their communities: hence, the transfer of planning activities to the local level, noted in the previous section. Whereas there were valid reasons to assume that local governments would be closer to the problems 'on the ground', in practice they lacked both financial and institutional resources to deal with the serious urban problems leftover from the socialist period: idle industry, decaying housing and infrastructure, environmental degradation, etc. New problems, such as growing poverty and unemployment that were induced by geopolitical and national-level events, were even further beyond what municipal governments could plan for or resolve.

Planning was seen as an outdated activity not just by politicians but by the broader citizenry as well. To a great extent, the problem arose from the legacy of the past. During the socialist period, planners were trained in the technocratic professions (architecture, civil engineering, economics, etc.). They applied their specialized knowledge to implement political priorities set by the communist leadership in the 5-year national economic plans. The highly centralized and bureaucratic system left no room for citizen participation in planning, as in other types of government decision-making. Thus, most citizens perceived planning as a faraway activity in which they had no input – a perception that greatly contributed to the profession's legitimacy crisis in the 1990s (Hirt, 2005). One of the immediate results of this crisis was that, in many towns and cities across the region, people took urban development into their own hands. To begin with, the intensity of building activities, public and private, generally decreased, as it does during any economic meltdown; new housing input dwindled especially dramatically. For example, in the early 1990s, the number of new dwelling units per year in Czechoslovakia dropped to a fifth of what it had been during the 1980s; in Bulgaria, it went down to a tenth (Stanilov, 2007: 173). But where housing and other construction did occur, an unusually high percentage of it was produced not just without long-term planning vision behind it but without government involvement of any sort, even without a building permit. Especially in the nations where ethnic conflict raged on for years, such as the former Yugoslavia and some of the post-Soviet republics, urban informality became the new norm (Tsenkova, 2012). The problem was aggravated by the fact that updated urban planning laws and regulations did not even exist. Through the 1990s, the last valid master plans of many cities dated back to the socialist 1960s and 70s. They were irrelevant and thus largely ignored. Urban development occurred on a piecemeal basis, with private parties often building whatever they wish where they wish – a process sometimes dubbed as 'investor urbanism' or even 'anarchitecture' (Hirt, 2012).

Whereas public sector withdrawal from urban involvement and ad hoc development became widespread, there are exceptions. In the capital cities of newly independent countries, such as those in the Baltic region and the former Yugoslavia, significant public-sponsored building activities were initiated with the goal of changing the architectural heritage to reflect new national identities (Czaplicka et al., 2009). Similar outcomes can be seen in some of large and economically powerful Soviet successor states, where national governments have invested enormous resources in endowing their capitals with a 'global-city' look (e.g., Astana, Kazakhstan) or where powerful city administrations have partnered with large private development corporations to the same end (e.g., Moscow and Saint Petersburg, Russia).

Pressing Challenges and New Planning Instruments

Many of the severe new challenges of the transition period, such as declining living standards and sharp social stratification, were a result of the broad economic decline and institutional instability of the transition period. However, other challenges arose specifically from the weakened status of planning. Perhaps ironically, during the years of sharp economic decline (the 1990s), the negative consequences of weak planning in many countries were partially offset by the fact that there was limited new building activity, as mentioned above. Thus, while the existing urban housing and infrastructure deteriorated, the sparse private resources invested in new construction guarded against the spread of poorly planned new building forms. This all changed about the year 2000, however, when economic growth picked up, leading to intense new construction in cities. The new construction did not always benefit the broader citizenry.

One obvious example of harmful new development is the loss of public green space. Socialist-era cities were generally more compact than their capitalist counterparts (Hirt, 2012). This is because the state controlled the periphery of cities, channeled the majority of the population toward the high-density mass-housing areas, and prevented sprawling suburban developments of the type common in capitalist contexts, especially in the United States. This changed dramatically after the end of socialism and specifically after the economic stabilization of the late 1990s. The outskirts of large cities were increasingly overtaken by upper- and upper-middle-class residential developments alongside big-box commercial establishments (malls, warehouses, business parks, entertainment complexes, etc.). Urban sprawl became one of the paradigmatic features of postsocialist urban development. The European Commission (2006) described cities in many East European countries (e.g., Bulgaria, Croatia, Estonia, Hungary, Latvia, Poland, Slovakia) as sprawling at rates far exceeding those in Western Europe. The Russian capital Moscow, for instance, lost 750 ha of forests in its once-lush greenbelt in just 2 years, from 1999 to 2001. During the same short-time period, forests in the metropolitan area decreased by 15% and grass areas declined by 55%, while impervious surface grew by 26% (Boentje and Blinnikov, 2007). The process of suburbanization, which devoured the once-compact edge of cities, was strengthened by the government fragmentation and decentralization that was instituted during the early postsocialist years. Although a welcome step toward a more democratic, 'closer to the people' government, the decentralization process created new institutional actors, small suburban-type governments, which in an era of scarcity competed vigorously to attract new taxes and economic development. The mechanisms they used to compete with their neighbors included making virgin land available for development, softening of building and zoning regulations, and adopting various incentives to attract large-scale developers. The outcomes of such policies are shown by Stanilov and Sykora (2012), for instance, the municipal competition contributed directly to the growth of sprawling developments in the vicinity of the Czech capital of Prague.

Loss of green space affected not only urban outskirts but also inner cities. Socialist-era regimes had demonstrated extraordinary generosity when it came to urban parks and neighborhood green spaces. This was because they had nearly full control over urban land, and because large public spaces well expressed the collectivist ideology of the socialist time period. But intense construction during the years of the post-socialist economy recovery devoured much of what may have been the most positive spatial legacy of the socialist years – the large public green spaces. In the Bulgarian capital of Sofia, for example, about 30% of public green spaces were 'lost in transition' and another 30% are awaiting the resolution of restitution claims (Hirt, 2012).

Another problem that may be attributed to weak planning during the postsocialist transition has been the loss of historic heritage. Although socialist-era regimes had maintained the landmark architecture in city centers, they had quietly left many historic neighborhoods to decay. Still, buildings in these neighborhoods were typically kept; i.e., they were not demolished and replaced with modern structures en masse as was the common case in some West European countries during the mid-1900s. Having survived decades of neglect, about the year 2000 historic areas in prime locations in East European cities were subjected to heavy market pressures. Developers were rarely interested in restoring the old structures. And whereas, in some cities and countries, the historic heritage laws were applied relatively stringently and demolitions were prevented, in others, especially those where growth pressures were immense, the laws were applied selectively thus allowing the destruction of invaluable historic buildings. Russia's St Petersburg, which is one of the largest UNESCO Heritage Sites in the world, is a well-known case (Trumbull, 2012). According to the local Architecture Preservation Society in the capital Moscow, some 400 buildings that were federally or locally designated as historic landmarks have been demolished since 1989 in order to make room for glitzy new developments (other experts put the number at 700; Hirt and Stanilov, 2009). And this count does not include the hundreds of other historic buildings that were not remarkable enough to make it into the official lists.

A fourth problem that planning failed to address is the change in mobility patterns. Socialist-era cities had fairly sophisticated mass-transit systems, which served the residents of their compact neighborhoods well. Following from the sprawl of the postsocialist metropolis and the limited investment in public infrastructure made in most countries, the urban citizenry became increasingly dependent on individual automobiles as the chief means of travel. In most East European countries, car ownership more than doubled during the postsocialist period; it is now rapidly increasing in the Caucasus region and Central Asia. For example, in Latvia, car ownership increased from 83 to 297 cars per 1000 people in about 20 years. Planners in this country's major cities, especially the capital of Riga, regard traffic congestion and air pollution as some of the worst current urban problems, yet little has been done to mitigate them. Auto-induced urban pollution has become an especially acute problem in the poorer countries, where much of the automobile stock comprises old second- and thirdhand cars imported from Western Europe, which do not meet contemporary emission standards (Hirt and Stanilov, 2009).

A fifth planning-related challenge stemmed from the sharpening stratification of postsocialist societies. Surely, urban planners and policy makers could have done little to prevent the yawning gap that developed between the handful of newly rich and the majority of the national populations during the first years of the postsocialist transition. However, in failing to produce an appropriate combination of policies, regulations and incentives to mitigate some of the negative outcomes of the postsocialist 'investor urbanism', planners contributed to converting social stratification into a spatial phenomenon. The socialist state had succeeded, albeit through authoritarian means, to reduce social segregation in cities. The mass-housing estates, specifically, were homes to wide and diverse segments of the population, from the intelligentsia to the working classes, and only the top regime apparatchiks lived in elite enclaves either in the centers of capital cities or in occasional securitized compounds at the urban edge. Assuming we consider spatial integration a positive phenomenon, we could say its benefits were all but lost in

postsocialist times. Middle and upper-middle-class residents of postsocialist cities have been increasingly exiting the mass-housing districts to settle either in the gentrifying neighborhoods of city centers or in the growing leafy suburbs. In some countries, the mass-housing areas are now occupied predominantly by elderly pensioners or ethnic minorities. In addition, in fast-growing cities, which attract residents from small towns and the countryside in search of jobs and services, the new peripheries very much resemble the type of slum areas found in the outskirts of the large cities of the developing world (e.g., Tsenkova, 2012). In the meantime, explicitly privatized urban forms, especially highly securitized gated communities, have gained popularity among new urban elites (e.g., Hirt, 2012).

The five planning-related challenges discussed above certainly do not comprise an exhaustive list. Many other serious urban spatial issues remain; for example, the need to adapt the extensive brownfields occupied by the derelict socialist-era industries to new use; the need to reconstruct the vast socialist-era housing stock, some of which is in perilously poor shape; and the need to deal with the poor quality of structures constructed during the lawless 1990s. These issues have been successfully addressed only in the most prosperous parts of the former Eastern Bloc: in the former East Germany and, to an extent, in the Central European states. However, there are signs that urban spatial planning is recovering from its crisis region wide and could, therefore, become more instrumental in solving some of the serious challenges. To begin with, the effects of chaotic urban development (congestion, pollution, etc.) of the 1990s have become so obvious that today there is a much broader political acceptance of planning as a legitimate municipal government function than there was a decade or two ago. Urban planning activities are definitely on the rise across the region. New master plans have by now been adopted for about every large city in the region; in many cases, the new master plans have in fact already undergone multiple revisions and amendments (Hirt and Stanilov, 2009). Alongside, the neighborhood-scale regulatory plans have been continuously updated and their implementation mechanisms have been strengthened, all of which sharply reduce the possibility of going back to the 1990s-style 'free-for-all' growth.

In addition, other types of planning are taking root as well. One new instrument is strategic spatial planning. Like master plans, strategic plans have a citywide focus. However, they are typically based on a shorter time span. Also, instead of attempting to address the full range of urban problems as master plans do, strategic plans aim to resolve only a specific set of pressing challenges. A 'typical' focus includes: improving city competitiveness in the era of globalization, fostering urban economic growth, promoting municipal financial reform, improving quality of life via modernizing housing and infrastructure, and enhancing urban sustainability (including through adopting local versions of Agenda 21). Many major cities in the region have by now passed strategic plans in addition to their traditional master plans (for example, Bratislava, Budapest, Prague, Riga, Sofia, Tallinn, Tirana, Zagreb). Because of their tangible focus, the strategic plans have attracted much wider interest from the citizenry than have traditional master plans. The second new type of planning is regional development planning. Regional plans are prepared by the regional layers of government created in the new member states of the European Union. They are essentially economic development plans, which the European Union requires in order to direct its massive structural funds. These funds aim to reduce regional disparities in income, wealth, and opportunities across the European Union and have been employed primarily in the new member states. The regional development plans do not necessarily have a spatial component. Still, in thinking through region-wide economic and other issues comprehensively, they have come to fill the planning vacuum that was created after the collapse of Soviet-style central economic planning. Finally, some planning activities have returned to the national level of government. Specifically, many countries have adopted national environmental strategies as well as action plans focused on specific environmental issues (e.g., air quality, water quality, biodiversity, waste disposal). This is especially true for the new European Union member states, which have been required to harmonize their national environmental legislation with European Union standards. Overall, whereas the era of virtually no planning in the transitional countries may be over, only time will tell whether the new planning activities will succeed in substantially improving the quality of life of the 400 million people who live in this part of the world.

See also: Cities Under Postsocialism; Critical Planning Theory; Planning Theory: The Good City and Its Governance; Regional Planning; Urban Planning in China; Urban Policy in Europe.

Bibliography

Andrusz, G., 1984. Planning and Urban Development in the USSR. State University of New York Press, Albany.

Bodnár, J., 2001. Fin de Millénaire Budapest: Metamorphoses of Urban Life. University of Minnesota Press, Minneapolis.

Boentje, J., Blinnikov, M., 2007. Post-soviet forest fragmentation and loss in the Green Belt around Moscow, Russia (1991–2001): a remote sensing perspective. Landscape and Urban Planning 82 (4), 208–221.

Bohle, D., Greskovits, B., 2012. Capitalist Diversity on Europe's Periphery. Cornell University Press, Ithaca.

Czaplicka, J., Gelazis, N., Ruble, B., 2009. Cities after the Fall of Communism: Reshaping Cultural Landscapes and European Identity. Johns Hopkins University Press, Baltimore.

Dobrinsky, R., Hesse, D., Traeger, R., 2006. Understanding the Long-term Growth Performance of the East European and CIS Economies. UNECE (United Nations Economic Commission for Europe), Geneva.

European Commission, 2006. Urban Sprawl in Europe: The Ignored Challenge. Directorate general Joint Research Center, Copenhagen.

Hamilton, F., Andrews, K., Pichler-Milanovic, N., 2005. Transformation of Cities in Central and Eastern Europe: Towards Globalization. United Nations University Press, New York.

Havrylyshyn, O., 2007. Fifteen Years of Transformation in the Post-Communist World. CATO Institute, Washington, DC.

Hirt, S., 2005. Planning the post-communist city: experiences from Sofia. International Planning Studies 10 (3–4), 219–240.

Hirt, S., 2012. Iron Curtains: Gates, Suburbs and Privatization of Space in the Post-Socialist City. Wiley-Blackwell, Oxford.

Hirt, S., Stanilov, K., 2009. Twenty Years of Transition: The Evolution of Urban Planning in Eastern Europe and the Former Soviet Union, 1989–2009. United Nations Human Settlements Programme, Nairobi.

Nedović-Budić, Z., 2001. Adjustment of planning practice to the new Eastern and Central European context. Journal of the American Planning Association 67 (1), 38–52.

Pickles, J., Smith, A. (Eds.), 1998. Theorizing Transition: The Political Economy of Post-Communist Transformations. Routledge, London.

Stanilov, K. (Ed.), 2007. The Post-Socialist City: Urban Form and Space Transformations in Central and Eastern Europe after Socialism. Springer, Dordrecht.

Stanilov, K., Sykora, L., 2012. Planning, markets, and patterns of residential growth in metropolitan Prague. Journal of Architectural and Planning Research 29 (4), 278–291.

Stark, D., December 1991. Path dependence and privatization strategies in East Central Europe. East European Politics and Societies 6, 17–54.

Tosics, I., 2005. City development in Central and Eastern Europe since 1990: the impact of internal forces. In: Hamilton, F.E.I., Andrews, K.D., Pichler-Milanović, N. (Eds.), Transformation of Cities in Central and Eastern Europe: Towards Globalization. United Nations University Press, New York, pp. 44–78.

Trumbull, N., 2012. Urban pulse—claiming "The Right to the City": architectural preservation in St. Petersburg as cultural and political catalyst. Urban Geography 33 (7), 1000–1007.

Tsenkova, S., 2012. Urban planning and informal cities in Southeast Europe. Journal of Architectural and Planning Research 29 (4), 292–305.

Tsenkova, S., Nedovic-Budic, Z. (Eds.), 2006. The Urban Mosaic of Post-Socialist Europe: Space, Institutions and Policy. Springer, Heidelberg.

Planning Ethics

Carlos N Silva, University of Lisbon, Lisbon, Portugal

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Planners and other professionals and officials in city government need to gain the confidence of the community they serve, and for that reason they must follow a standard of conduct that is far higher than what is strictly determined in the law. In order to achieve this, professional organizations in the multidisciplinary field of urban planning have adopted, in some cases as early as the beginning of the twentieth century, codes of professional conduct or codes of ethics. This article examines and discusses the different ethical perspectives that somehow frame the professional conduct of planners, explores a sample of these codes of ethics, its contents, and practical application, and identifies new challenges likely to result from the widespread use of information and communication technologies in the planning process.

Introduction

The main purpose of urban and regional planning is to pursue the interests of the community as a whole. However, with the increasingly diverse social and multicultural structure of our contemporary urban societies, that purpose is no longer as straightforward as it may have appeared in the past, since citizens with such highly different social and cultural backgrounds have potentially different and conflicting goals. In part for this reason, urban and regional planning is no longer seen as a politically neutral activity, contrary to the prevailing perspective during most of the twentieth century, namely within the framework of the rational planning paradigm. Planners and other professionals and officials in city government need to gain the confidence of the community they serve and for that reason he or she must commit to a standard of conduct that is far higher than what is strictly required by the planning law.

If in previous planning paradigms, from the City Beautiful and Garden City movement, at the end of nineteenth and early decades of the twentieth century, to the Congrès International d'Architecture Moderne discourse on Urbanism, or in the rational planning paradigm, urban planning was regarded as value free, this view was challenged since the late 1960s by postmodern approaches. This change was in part a recognition of the complex ethical dilemmas urban planners face in all stages of the planning process (Davey, 1996; Fox, 2000; Wachs, 1990; Younés and Paquot, 2000), related to the notion of public interest, obligations toward the community, colleagues, and employers to issues of intellectual property, privacy, confidentiality, as well as duties toward the environment, other forms of life, and future generations, and more recently also challenges associated with the extensive use of information and communication technologies, in particular the Internet, in the different stages of the planning process (Silva, 2010, 2012). Collaborative Planning (Healey, 1997) and New Urbanism (Silva, 2002) with their emphasis on citizen engagement in the planning process and in decision making highlight the importance of planners' values and the need to consider citizens' views and values as well.

In actual fact, planners face complex challenges in incorporating ethics into daily decisions over land uses, environmental conditions, social issues, and economic options, most of them multiscalar or multilevel, linking complex networks of actors, from the public and private sectors, in particular considering that an unethical conduct does not necessarily mean that it goes against the letter of the law, as Beatley (1994) and Perego (2010), among others, suggest. Despite that, and as the evidence shows, the discussion of these wider ethical issues within planning professional organizations is relatively recent (Hendler, 1990, 1991, 1999; Howe, 1990; Kaufman, 1990, 2001; Khakee and Dahlgren, 1990), including those basic professional issues normally dealt by the codes of ethics or codes of ethical conduct in the different professional organizations that make up the multidisciplinary field of urban planning (Fisher, 2000; Kaufman, 1990; Mahoney, 1992), as is recent the discussion (Howe and Kaufman, 1979, 1981) about the shift in the way ethical values were considered by planning professionals.

Nonetheless, some of these organizations, in North America, Europe, and in other parts of the world, adopted as early as the beginning of the twentieth century codes of professional conduct, which have been revised several times since then (Silva, 2005): in 1914, the American Society of Civil Engineers; in 1924, the International City/County Management Association (ICMA, 2004); in 1930, the Royal Town Planning Institute (RTPI); in 1962, the American Society of Planning Officials, the predecessor of the American Planning Association; and in 1971, the American Institute of Planners, the antecedent of the American Institute of Certified Planners (AICP). The European Council of Town Planners adopted, in 1985, guidelines for professional conduct (ECTP, 1985). The Canadian Institute of Planners approved a code of conduct in 1994. In Australia and New Zealand, in Brazil, South Africa, and India, there is also evidence of a similar process, to mention just a few more examples from other continents.

Ethical Issues in Urban Planning

Urban planners face three kinds of ethical issues (Silva, 2005). The first, related to individual rights, includes, for example, the

dilemmas associated with the choice of lifestyles and the preservation of the characteristics of a locality, paternalism, and the need to reduce damages. The second, related to social justice, includes obligations toward different social groups within the community, requiring planning decisions and actions to reduce economic and social inequalities, if necessary, through positive discrimination in favor of the most disadvantaged groups; in other words, a plan should reduce inequalities within the community, in part through positive discrimination in favor of the most vulnerable social groups, according to income, age, gender, ethnicity, and ability; in addition, planners have obligations toward future generations, since planning decisions, regarding the built environment and the natural patrimony, have far-reaching consequences. The third is related to the moral obligations planners face in relation to the environment and sustainable development.

Planning also has ethical duties toward other jurisdictions, since some of the issues it addresses are multiscalar by nature, requiring coordination with other subnational authorities, municipalities, and regional entities and transborder cooperation. This can be addressed by the creation of metropolitan or regional self-governments and institutionalized forms of transborder cooperation as is the case in some parts of the European Union. Transparency in the decision-making process is another ethical duty planners have, which requires nondiscrimination of social groups or individuals and the proactive inclusion of all those usually excluded from the public decision making, providing a fair participatory framework.

In sum, faced with the urban transition, in which more than half the world population now live in urban areas (UN-Habitat, 2010), with the challenges of climate change (UN-Habitat, 2011), and 20 years after the 1992 Rio Conference, or 25 years after the 1987 Brundtland Report, planners have now an accrued moral obligation to plan sustainable and resilient cities and regions (UN-Habitat, 2009; UN, 2012; UN-ISDR, 2005).

Ethical Perspectives in Urban Planning

Planners, as happens with other practitioners, are guided in their professional activity by an ethical framework that is the result of multiple influences, reflecting the socialization process they went through during their lifetime. These influences include, for example, values and experiences within the family context, influences from the school, friends, religion (i.e., religions differ, among other aspects, according to the degree of their anthropocentrism), political influences, and professional experiences as well.

Ethical perspectives in planning can be grouped, as proposed by Beatley (1994), according to two main dimensions. The first opposes the consequentialist perspective, according to which it is the outcome of an action (i.e., the outcome of a plan, project, or planning decision) that determines if such an action or decision is correct or not (Beatley, 1994; Darwall, 2003a, 2003b) to the deontological perspective that sees an action or behavior as having a value in itself, worth fulfilling independently of its consequences. In other words, it is possible to have planning practices with an ends-oriented ethics and other practices more means oriented.

The second dimension reflects the importance given to the *Homo sapiens* in the definition of the moral community, and opposes the anthropocentric to the nonanthropocentric perspective. While for the anthropocentric perspective, the *H. sapiens* is the center or the sole member of the moral community, for the nonanthropocentric also other forms of life and the environment as a whole are included in the moral community, with a value intrinsic in themselves independently of the relevance they might have for human beings. Therefore, planners can have rather different professional practices as a result of their ethical perspective. This has been empirically substantiated in the pioneering works of Kaufman, Khakee, and Dahlgree, referred by Silva (2005), as well as in more recent studies.

The deontological and the nonanthropocentric perspectives are relatively recent within the planning profession, emerging partially in association with the overall critique to modernism and rational planning, on one side, and the rise of the sustainable development agenda since the late 1980s, as argued before (Silva, 2005), and more recently also associated with the discourse on resilient urban development. From this perspective – deontological and nonanthropocentric – urban planning seeks to maximize the public interest without putting into question other moral duties in relation to other forms of life, as well as toward the creation of a more just city, in the sense Fainstein (2010) and Soja (2010), among others, put it, or more generally, a more just society.

A new ethical perspective seems necessary, if urban planning is to meet its obligations toward the preservation of the environment, not only for the needs of the *H. sapiens*, but also for other forms of life and future generations. Urban planners have an ethical duty to minimize the ecological footprint of cities, which requires a more compact urban form, the redevelopment of brown-field areas, higher densities, and a social and functional mix as well. In that sense, the United Nations sustainable development discourse, recently reinforced in the Rio+20 Conference (UN, 2012), and the New Urbanism plea for a more compact city and smart growth instead of continued urban sprawl, more pedestrian-friendly areas, biodiversity, and natural habitats protection, are just examples of some of the new ideas that contribute toward a more ethically responsible urban planning.

Planning Ethics in the Codes of Professional Organizations

In part as a consequence of these changes, it is now common for professional organizations of planners to have a statement of ethical principles or an enforceable code of professional conduct with the general purpose to set standards, values, and behaviors in the different stages of the planning process. It not only sets an ideal for the profession and its core values, but it can also act as a self-regulation reference and serve as a guide of ethical conduct to its members. It informs the community, employers, and clients, about the values and obligations a planner is committed to, and reinforces the position or status of the planning profession within the society

(Hendler, 1990, 1991; Kaufman, 1990, 2001). Planners are expected to adhere to these rules or guidelines and to practice the profession with the highest professional standard in order to gain the respect and trust of the community where they work and which they are supposed to serve.

The exact content of these codes of ethics or ethical guidelines vary greatly among professional organizations but it is common for these codes and guidelines to require members, for example, to promote equality of opportunity and to prohibit discrimination on the grounds of gender, age, race, religion, disability, sexual orientation, or any other social characteristic.

Even though general ethics courses and ethics codes from other professional organizations can also be useful for planners, there are certain ethical issues that are specific to professional planners and, for that reason, the demand for training on the ethical issues planners are likely to face during the planning process increased and it is now common for planning conferences to include sessions on ethics, in which planning professionals and ethics experts share personal experiences and discuss ethical scenarios.

Certified planners are expected to follow the professional code of conduct and may be held accountable to it. Members not in compliance with the Code of Ethics and Professional Conduct may be subject to sanctions, including the loss of membership in the most severe cases. For other professionals working in planning, be they public decision makers or advisors, it may be enough to have just a statement of ethical principles in planning that will serve as a guide. For example, the Association of American Planners' (APA) ethical guidelines are supposed to be adopted and followed by anyone working in planning, from planners, in both private and public sector, to appointed and elected officials, and citizen as well. On the contrary, only certified planners are held accountable to the AICP Code, although other planning professionals should also take it into consideration.

The APA adopted a 'Statement of Ethical Principles,' in 1987, replacing the 1962 code of the American Society of Planning Officials, followed, in 1992, by a new and revised 'Statement of Ethical Principles in Planning' (Barrett, 2001; Kaufman, 1990; Lucy, 1997). The 1992 statement, still applied two decades afterward, was adopted as "a guide to ethical conduct for all who participate in the process of planning as advisors, advocates, and decision makers. It presents a set of principles to be held in common by certified planners, other practicing planners, appointed and elected officials, and others who participate in the process of planning" (APA, 1992). This guide is organized into three main sections: the planning process must continuously pursue and faithfully serve the public interest; planning process participants continuously strive to achieve high standards of integrity and proficiency so that public respect for the planning process will be maintained; and APA members who are practicing planners continuously pursue improvement in their planning competence as well as in the development of peers and aspiring planners. They recognize that enhancement of planning as a profession leads to greater public respect for the planning process and thus serves the public interest.

Its twin organization, the AICP, adopted a code in 1978 (the AICP Code of Ethics and Professional Conduct), replacing the 1971 code of the preceding organization, the American Institute of Planners. In 2000, the AICP started the revision of the 1978 Code, a process that lead, after numerous meetings, to the new 'Code of Ethics and Professional Conduct,' which went into effect in 2005, being revised in 2009 (AICP, 2009). It is not an entirely new code but introduced a number of important changes, both in substance and in the procedures. For example, the previous advisory rulings that have been issued over the years, since 1978, and which became an addendum to the previous code, were incorporated into the main text of the 2005 Code. It is admitted that similar rulings may be adopted in the future as deemed necessary during the application of the present code.

The AICP Code of Ethics and Professional Conduct is divided into three main sections. The first section – Principles to Which We Aspire – contains the principles that constitute the ideals to which every member of the AICP is committed, which are grouped into three sections: responsibility to the public; responsibility to clients and employers; and responsibility to the profession and colleagues. The second section contains 26 rules of conduct that AICP members should adhere to, receiving sanctions if they fail to do so, the maximum of which is the loss of certification. The next two sections deal with different procedures, including those applied when planners are convicted of serious crimes.

As an example of the diversity of structures that can be found in the codes of ethics, the comparison between the AICP code, in the USA, with that of RTPI in the United Kingdom, provides a good illustration. The RTPI code of professional conduct, last amended in 2011, and effective since January 2012, includes 17 clauses, with requirements that are expected to guide the professional activities of every RTPI member and three annexes (Supplementary regulations; Relevant Bylaws; Disciplinary action) (RTPI, 2012a; RTPI, 2012b). According to this code, RTPI members "(a) shall act with competence, honesty and integrity; (b) shall fearlessly and impartially exercise their independent professional judgment to the best of their skill and understanding; (c) shall discharge their duty to their employers, clients, colleagues and others with due care and diligence in accordance with the provisions of the Code; (d) shall not discriminate on the grounds of race, sex, sexual orientation, creed, religion, disability or age and shall seek to eliminate such discrimination by others and to promote equality of opportunity; (e) shall not bring the profession or the Royal Town Planning Institute into disrepute." In addition, the RTPI adopted a set of procedures for investigating complaints against members and for taking disciplinary action. Among other aspects it set rules regarding the collection of preliminary material; formal investigation; findings that no breach of the Code of Professional Conduct has occurred; findings that a breach of the Code of Professional Conduct has occurred; suspension from membership; provision for appeal; and publicity of any findings of a breach of the Code.

In the Planning Institute of Australia' Code of Ethics and Professional Conduct, last revised in 2002 (PIA, 2002), the principle that 'a planner's responsibility to the community must take precedence over sectional interests' is implicit. It has two main sections, the first of which deals with the rules and

norms (professional standards; conflicts of interest; confidentiality and disclosure; true professional opinion; professional competence; marketing of services; fair competition), and the second with the procedures and penalties.

Besides planners, other professions directly engaged in urban and regional planning also adopted codes of ethics and professional conduct, such as geographers (AAG, 2009), architects (AIA, 2007), landscape architects (ASLA, 2006, 2009), civil engineers (ASCE, 2006), and sociologists (ASA, 1999). These codes, as those of professional organizations of planners, mentioned before, are mainly focused on issues of professional conduct. All these professional organizations seem to agree that a code of ethics is far from being enough. The profession may have a good reputation, but if there are occasional ethical lapses, the good reputation can soon disappear. It is therefore imperative to build a professional ethical culture, which although not as visible as other actions, seems to pay dividends in the long term, both for the profession and for the community every planner is expected to serve. For that, the ethical foundations of the planning professional organization need constant attention and training during the entire professional life of its members. It is important to share stories about specific ethical issues planners' face, to learn from each other, building gradually an ethical culture within the profession.

For this reason, planning organizations tend to include ethical issues in their professional development programs. In case of AICP–APA, a toolkit was prepared to assist their local chapters in the organization of training sessions on the standards of planning ethics, a tool focused on themes that are eligible for AICP certification maintenance credits. These training sessions can have different formats according to the needs of the audience: lectures; panel discussions; facilitated sessions; and plenary sessions (AICP, 2010). Another resource made available for planners by their professional organization is a confidential telephone counseling service, as the one offered by the Planning Institute of Australia, when new ethical issues or dilemmas do not fit or seem not to fit well into the written guidelines and rules. The publication of short texts or research articles on numerous topics relevant to ethical practice in the workplace, as well as questions on ethics that have been raised by members of the professional association, and the staff answers based on the Code of Ethics, has also been used to increase ethical awareness of planners. In sum, ethical education should not only be made an integral part of advanced undergraduate and graduate degrees, but also of development training programs promoted by planning professional organizations.

The widespread use of information and communication technologies in urban planning is gradually changing the way planners work and how they interact with citizens and other stakeholders (Silva, 2007, 2010). These changes raise new ethical issues, in particular in the context of the new possibilities offered by the Web 2.0 (Buchanan, 2012), in the use of online research methods (Suri and Patel, 2012), or the same kind of ethical issues with new configurations, for example, in the field of privacy and confidentiality. Even if we accept that it is only necessary to apply the basic ethical principles in the new digital setting in which e-Planning takes place, the codes of ethics of planning associations should include an explicit reference to it.

Conclusion

It is now common to admit that urban planning inevitably faces ethical dilemmas in all stages of the planning process, from the initial decision up to the implementation and revision of a given plan, to mention just few different contexts in which ethical issues, interests, and values emerge and need to be addressed by urban planners. Simultaneous with this increasingly greater awareness of ethical issues, new ethical concerns related to human rights, social cohesion, a deeper green consciousness, as well as new challenges associated with the widespread use of information and communication technologies in the planning process, emerged in the field of urban and regional planning in the last decades. All this and the increased emphasis on sustainable development and urban resilience seem to be moving urban planning toward a more deontological and nonanthropocentric ethical perspective, although in practice planning decisions seem to continue being predominantly based on utilitarian and anthropocentric principles.

Bibliography

AAG, 2009. Statement on Professional Ethics (adopted 1998, updated 2005, and revised November 1, 2009). Association of American Geographers, Washington.
AIA, 2007. Code of Ethics and Professional Conduct. The American Institute of Architects, New York.
AICP, 2009. Code of Ethics and Professional Conduct (adopted 2005 and revised 2009). American Institute of Certified Planners, Washington, DC.
AICP, 2010. Ethics in Planning: A Toolkit for Conducting Ethics Sessions. American Institute of Certified Planners (The American Planning Association' Professional Institute), Washington, DC.
APA, 1992. Statement of Ethical Principles in Planning. American Planning Association, Washington, DC.
ASA, 1999. Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics. American Sociological Association, Washington, DC.
ASCE, 2006. Code of Ethics (adopted in 1914 and last amended in 2006). American Society of Civil Engineers, Reston, VA.
ASLA, 2006. Code of Environmental Ethics (adopted in 2006 and last revised in 2006). American Society of Landscape Architecture, Washington, DC.
ASLA, 2009. Code of Professional Ethics (adopted 1995 and last revised in 2009). American Society of Landscape Architecture, Washington, DC.
Barrett, C., 2001. Everyday Ethics for Practicing Planners. Institute of Certified Planners, Washington, DC.
Beatley, T., 1994. Ethical Land Use. Principles of Policy and Planning. The John Hopkins University Press, Baltimore.
Buchanan, E., 2012. E-research ethics and e-planning: emerging considerations for transformative research. International Journal of E-Planning Research 1 (1), 5–15.
Darwall, S. (Ed.), 2003a. Consequentialism. Blackwell, Oxford.
Darwall, S. (Ed.), 2003b. Deontology. Blackwell, Oxford.
Davey, P., 1996. Work ethics (ergonomics in architecture). The Architectural Review 190 (download from Academic OneFile, 2012).
ECTP, 1985. International Agreement and Declaration by the National Institutes and Associations of Professional Town Planners Within the European Economic Community. Appendix C – Professional Conduct Requirements. European Council of Town Planners.
Fainstein, S., 2010. The Just City. Cornell University Press, Ithaca, NY.
Fisher, S., 2000. How to think about the ethics of architecture. In: Fox, W. (Ed.), Ethics and the Built Environment. Routledge, London.
Fox, W. (Ed.), 2000. Ethics and the Built Environment. Routledge, London.
Healey, P., 1997. Collaborative Planning. Palgrave, Basingstoke.
Hendler, S., 1990. Moral theories in professional practice: do they make a difference? Environments 20 (3), 20–30.
Hendler, S., 1991. Do professional codes legitimate planners' values? In: Thomas, H., Healey, P. (Eds.), Dillemmas of Planning Practice. Ethics, Legitimacy and the Validation of Knowledge. Avebury, Aldershot.
Hendler, S., 1999. Planning Ethics. Center for Urban Policy Research, New Brunswick.

Howe, E., 1990. Normative ethics in planning. Journal of Planning Literature 5 (2), 123–150.
Howe, E., Kaufman, J., 1979. The ethics of contemporary American planners. Journal of the American Planning Association 45 (3), 243–255.
Howe, E., Kaufman, J., 1981. The values of contemporary American planners. Journal of the American Planning Association 47 (3), 266–278.
ICMA, 2004. The ICMA Code of Ethics with Guidelines. International City/County Management Association, Washington.
Kaufman, J., 1990. American codes of planning ethics. Content, development and after-effects. Plan Canada 30 (5), 19–34.
Kaufman, J., 2001. Is the Increased Output of Planning Ethics Knowledge in North America Affecting Those Who Practice Planning? An Exploratory Investigation. Paper presented at World Planning Schools Congress, Xangai, China.
Khakee, A., Dahlgren, L., 1990. Ethics and values of Swedish planners: a replication and comparison with an American study. Scandinavian Housing and Planning Research 7, 65–81.
Lucy, W., 1997. APA's ethical principles include simplistic planning theories. In: Campbell, S., Fainstein, S. (Eds.), Readings in Planning Theory. Blackwell, Oxford.
Mahoney, J., 1992. Teaching Business Ethics in the UK, Europe and the USA. A Comparative Study. The Athlone Press, London.
Perego, M., 2010. Ethics versus the law… which is better? Just because something is legal doesn't make it ethical. Public Management 92 (4) (download from Academic OneFile, 2012).
PIA, 2002. Code of Ethics and Professional Conduct. Planning Institute of Australia, Kingston.
RTPI, 2012a. Code of Professional Conduct (as last amended by the Board of Trustees on 28 September 2011. Effective from 1 January 2012). The Royal Town Planning Institute, London.
RTPI, 2012b. Complaints Procedures: Procedures for Investigating Complaints Against Members and for Taking Disciplinary Action. The Royal Town Planning Institute, London.
Silva, C.N., 2002. A 'Carta do Novo Urbanismo' e a 'Nova Carta de Atenas': a utopia urbana do Século XXI? Cadernos Municipais – Revista de Acção Regional e Local XVI (79), 35–47.
Silva, C.N., 2005. Urban planning and ethics. In: Rabin, J. (Ed.), Encyclopedia of Public Administration and Public Policy. Marcel Dekker–Taylor & Francis, New York, pp. 311–316.
Silva, C.N., 2007. e-Planning. In: Anttiroiko, A., Malkia, M. (Eds.), Encyclopedia of Digital Government. Idea Group Publishing, Hershey, PA, pp. 703–707.
Silva, C.N., 2010. The e-planning paradigm. Theory, methods and tools. An overview. In: Silva, C.N. (Ed.), Handbook of Research on E-Planning: ICTs for Urban Development and Monitoring. IGI Global, Hershey, PA, pp. 1–14.
Silva, C.N. (Ed.), 2012. Online Research Methods in Urban and Planning Studies: Design and Outcomes. IGI Global, Hershey, PA.
Soja, E.W., 2010. Seeking Spatial Justice. University of Minnesota Press, Minneapolis.
Suri, H., Patel, F., 2012. Ethical considerations in online research methods. In: Silva, C.N. (Ed.), Online Research Methods in Urban and Planning Studies: Design and Outcomes. IGI Global, Hershey, PA, pp. 394–408.
UN, 12 June 2012. Rio+20, United Nations Conference on Sustainable Development. Outcome of the Conference: The Future We Want. United Nations.
UN-Habitat, 2009. Global Report on Human Settlements 2009: Planning Sustainable Cities. United Nations Human Settlements Programme, Nairobi.
UN-Habitat, 2010. State of the World's Cities, 2010/2011. Bridging the Urban Divide. Earthscan, London.
UN-Habitat, 2011. Global Report on Human Settlements 2011 – Cities and Climate Change: Policy Directions. United Nations Human Settlements Programme, Nairobi.
UN-ISDR, 2005. Hyogo Framework for Action 2005–2015: Building the Resilience of Nations and Communities to Disasters. United Nations Inter-Agency Secretariat of the International Strategy for Disaster Reduction, Geneva.
Wachs, M., 1990. Ethics and advocacy in forecasting for public policy. Business and Professional Ethics Journal 9 (1–2), 141–157.
Younés, C., Paquot, T., 2000. Éthique, architecture, urbain. La Découverte, Paris.

Planning Issues and Sustainable Development

Peter Newman, Murdoch University, Perth, WA, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11479–11482, © 2001, Elsevier Ltd.

Abstract

There is a growing use of sustainable development by planners. The concept is being adopted without much controversy to create more holistic planning frameworks and community-based planning processes. However, there is no clear agreement on what sustainability means in terms of city size, urban form, and transportation. The debates are summarized and some resolution suggested.

'Sustainable development' or 'sustainability' for short simply means that in a global context any economic or social development should improve and not harm the environment. This concept has developed from a global political process over the last three decades of the twentieth century into one that now touches every part of society. Sustainability has cut across all disciplines and professions, and has developed many complexities in relation to how people plan for their futures (Pezzoli, 1997). Its application to planning is full of interest and controversy.

Town and regional planning has its origins firmly in a desire to improve the environment and create a better future. The visions of Ebenezer Howard, Patrick Geddes, and Lewis Mumford, for example, were all rooted in the need to make a greener city. Today there are many planners who use sustainability to rationalize why they are planning the way they do. This article will try to clarify the ways that sustainability is shaping how planning is done, as well as the differences in the ways that planners relate to sustainable development.

The issues relating to sustainability that appear to be uniting planners are:

1. How necessary it is to use holistic planning frameworks; and
2. How planning processes must start with community/civil society perspectives.

The issues which appear to divide planners are:

1. How planners view the size of cities; and
2. How planners view density and car dependence.

Some resolution will be attempted on these issues.

The Concept of Sustainability

Sustainable development was defined by the Brundtland Commission in 1987 after they were given the task of resolving the apparent global conflict between development and the environment (WCED, 1987). They concluded that it is possible to develop the world in a way that improves the environment if:

1. High income nations see their responsibility to reduce their resource consumption with every new step in the development process;
2. Low income nations direct development to those who need it most as only then will population growth stabilize and misuse of natural resources begin to ease;
3. All nations need to see development as meeting community goals, thus development must be inherently a 'bottom up' process; and,
4. All professions and disciplines need to adjust to a more holistic, integrated approach if sustainability is to be converted into real change.

The thrust of this article is that, within the sustainable development framework of thinking, there are clear differences in what planners are suggesting sustainability means for cities. However, the overall situation cannot be seen as just a series of raging controversies, and indeed there are substantial areas of agreement.

The Agreements: Holistic Frameworks and New Planning Processes in the Age of Sustainability

Dyck (1998) suggests that sustainability is a concept which challenges prevailing planning theory because planning has emerged primarily from a modernist social context. In this perspective, modernism has associated growth or development with increased resource use, trickle-down economics, and 'expert-oriented' planning. Very few planners are, however, espousing a modernist view in the new millennium, and even the World Bank seems to agree with the basic conceptions of how sustainable development should now be viewed (Leitmann, 1999). In particular, the areas of emerging agreement are the need for new, more holistic planning frameworks and community-sensitive planning processes.

Holistic Frameworks

The principles of sustainability outlined above can be applied to cities, though the guidance on how this can be done was not very clear in Agenda 21 and other UN documents. It is probably true to say that the major environmental battles of the past were fought outside cities, but that awareness of the need to come back to cities is now recognized universally by environmentalists, government, and industry (United Nations Center for Human Settlements, 1996).

Anders (1991), in a global review of the sustainable cities movement, pointed out that "The sustainable cities movement seems united in its perception that the state of the environment

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74025-6

demands action and that cities are an appropriate forum in which to act" (p. 17).

One of the strongest themes running through the literature on urban sustainability is that, to solve problems, planners need to view the city as an *ecosystem*. As Tjallingii (1991: p. 7) puts it: "The city is (now) conceived as a dynamic and complex ecosystem. This is not a metaphor, but a concept of a real city. The social, economic and cultural systems cannot escape the rules of abiotic and biotic nature. Guidelines for action will have to be geared to these rules."

The city is an open ecosystem, having inputs of energy and materials. The main environmental problems (and economic costs) are related to sustaining the growth of these inputs, and managing the increased outputs. By looking at the city as a whole and by analyzing the pathways along which energy and materials (and pollution) move, it is possible to begin to conceive of management systems and technologies which allow for the reintegration of natural processes, increasing the efficiency of resource use, the recycling of wastes as valuable materials, and the conservation of (and even production of) energy.

There may be ongoing academic debate about what constitutes sustainability and how it applies to cities (Slocombe, 1993; Mitlin and Satterthwaite, 1994), but what is clear is that many strategies and programs around the world have begun to apply such notions both for new development and for redevelopment of existing areas. Holistic frameworks such as the Extended Metabolism Model (Newman and Kenworthy, 1999), the Ecological Footprint (Wackernagel and Rees, 1996), and Local Environmental Action Planning (Leitmann, 1999) are all emerging as new planning frameworks.

Planning Processes

One of the strengths of the sustainability concept is that it is not a set of regulations or even highly definitive principles. Its principles need to be examined by every local community to see how they can be applied to create better long-term futures for their children. Thus planners at the local level have often found something of the reality of sustainability through the use of Local Agenda 21 Plans which have emerged as a tool for planning since the Rio conference (ICLEI, 1996, 1997; Selman, 1996). Over 2000 local governments have produced such plans. Others have adopted approaches based on sustainability indicators such as Sustainable Seattle (City of Seattle, 1993) or the Local Environmental Action Plan (Leitmann, 1999), and many are now using ICLEI's Cities for Climate Protection system as a guiding approach.

All these new processes stress that planning has moved beyond traditional approaches such as Impact Assessment or Strategic Planning based on trend projections and functional separation. They have in common a desire to plan from a community base, integrated with the goals of the market and other levels of government, including the new global agendas generated from Rio, Kyoto, and beyond.

The stories of hope emerging at a local level are rich and diverse when examined at the grass roots level (e.g., Pathways to Sustainability Conference, 1997; ICLEI, 1997). The reason for this is probably that at the local level it is possible for government more easily to make the huge steps in integrating the environmental, economic, and social, in order to make policy developments that are sustainable. They are also closer to concerned people, and more distant from the single-issue powerful lobbies such as the fossil fuel and road lobbies, who are so obvious in shaping national priorities (Korten, 1990, 1995).

The Disagreements: Application of Sustainability to City Size, Urban Form, and Transportation

The disagreements that are apparent among planners, who all profess to want more sustainable cities, are reducible to issues concerning urban form and transportation. They relate to the size of cities and to their density/car dependence.

City Size and Sustainability

Many planners subscribe to the view that, on environmental grounds (and other criteria), cities would be better if they were smaller (Troy, 1996; Morris, 1982). This is an old planning debate, for example, Fischer (1976) shows how optimum size cannot be found on any social or economic variable, but the debate has new currency in light of the current move to seek 'smart growth' and growth management. The desire to keep settlements small is seen by many to be a major part of what sustainability means.

The evidence from studies of settlements shows that as cities get larger they become more efficient in energy use per person and other critical sustainability variables (Newman and Kenworthy, 1999; Naess, 1993; Smith et al., 1998). This appears to be related to the economies of scale and density associated with bigger cities, which applies equally to technologies seen to be appropriate for sustainable development such as transit, recycling, or the new information industries. It does not mean, however, that larger cities do not reach capacity limits in terms of their airsheds and water systems. But it does mean that simple ideas of decanting people from big cities to smaller ones would not help to create a more sustainable global future.

The evidence suggests that smaller cities find the process of achieving efficiency in economic and resource terms harder than larger cities. The strong emotional appeal of 'smallness' to our age is not, however, without its basis. E.F. Schumacher's 'Small is Beautiful' approach, with its attack on modern gigantism, is still relevant as it suggests that all technology needs to be at an appropriate scale for the community that it is meant to be serving. Community-scale technology is emerging as communities begin to assert their roles – whether the technology is in villages in the Third World or modern large cities (Korten, 1990). The thrust of the New Urbanism and the 'urban village' movement is that communities need to be physically designed for and given infrastructure at the scale of the community, in cities both large and small (Calthorpe, 1993).

Density and Car Dependence

In cities across the world the issue of urban sprawl is firmly on the agenda. Associated with it are the issues of density and levels of car dependence, and hence traffic. In the USA in 1998,

240 local governments had antisprawl initiatives on their ballots, and reductions in traffic had become a top priority across the country. However, debate among planners on these issues shows little sign of resolution.

The divide continues to widen between those advocating more compact cities and transit-oriented reurbanization (e.g., Cervero, 1999; Whitelegg, 1993; Roseland, 1998; Newman and Kenworthy, 1999; Calthorpe, 1993; Elkin et al., 1991), and those advocating reduced densities and only technological changes to cars (e.g., Gordon and Richardson, 1989; Troy, 1996; Wachs and Crawford, 1991; Johnson, 1993). Another category of planners sit on the density/car dependence fence (e.g., Haughton and Hunter, 1994), while others see the problems but tend to avoid the planning implications in low density, car-dependent planning (e.g., Leitmann, 1999; World Bank, 1996; Stren et al., 1992).

The data on density and car dependence is very clear. All cities appear to lie on the negative exponential curve of density and gasoline per capita (Newman and Kenworthy, 1999). The same relationship is seen across cities, so that as density increases, per capita energy goes down, well beyond any income factors, for example, Manhattan uses 500% less energy per capita than the outer suburbs of New York. The mechanism is understood easily in terms of the relative ability to use different modes as distances between land uses spread out. It suggests that density is a powerful determinant of gasoline use, and hence if cities are to develop sustainably they should be facilitating reurbanization rather than new fringe development. Other planning studies have also shown over many decades the link between low urban density and a range of environmental problems (RERC, 1974; Berry et al., 1974).

Despite this kind of evidence there is a long tradition of planners seeing density as a problem (Newman and Kenworthy, 1989). Those planners advocating the benefits of sprawl who are associated with land or automobile lobbies can be discounted in debates about urban sustainability (STPP, 1994). However, there are many environmental planners who also adopt the benefits of low-density planning (e.g., Gordon, 1990; Todd and Tukel, 1981; Coates, 1981). This suggests that there may be a deeper issue in this debate; an issue that also lies behind the question of city size and sustainability.

Radberg (1995), when reviewing the form of cities required to achieve sustainability, described the rationale for a more compact city which can conserve energy. He then set out the form of a city for achieving green objectives that are 'related to the ecological perspective' for 'recycling and cultivation'; this city is 'greener, more ruralized, more spread out.' He therefore raised the questions: Is there a conflict between the low-energy city and the green city? Is there a fundamental conflict between the need to have more urban land for the green city in order to do local ecological processing, and a low-energy city with its need to minimize travel distances and hence have a frugal use of land?

The resolution of this issue as suggested by Radberg (1995) was to enable some parts of cities to be developed more intensely for reductions in energy, and other parts to be more reduced in their density so that 'greener' activities could occur there. The evidence, however, is that those planners who support the reduction of density are not able to accept that any density increases are worthwhile (e.g., Troy, 1996). On the other hand, studies of urban ecology innovations by Scheurer (1999), confirmed by Beatley (2000), found that there are no increases in ecological activity in low-density areas. Indeed, the best examples of urban ecology (recycling, solar design, storm water retention, permaculture …) occurred in denser developments, mostly in inner city areas where communities were able to create innovative experiments, or in intentional rural or urban-fringe communities. In the car-dependent suburbs urban ecology innovations are rare.

Newman and Kenworthy (1999) tried to resolve the issue of density and the green city. They suggest that green innovations can indeed occur anywhere in a city, although experiments may best be applied in rural areas, especially in places where rural population decline is occurring. However, the broader global sustainability agenda cannot be subsumed by this new rationale for density-lowering; cities need to reurbanize while conducting simultaneous improvements in urban ecology. They suggest that the difference in attitude between the two groups is as much due to confusion over where different approaches to sustainability should be employed. In simple terms, the city should be allowed to become more urban, and the countryside more rural, if the multiple goals of planning and sustainability are to be met.

A resolution of this issue may not occur while people still use the word 'sustainable' to describe extensions of low-density, car-based development (even if designed with New Urbanist principles). The strong emphasis on reductions in energy that is now associated with sustainability will probably mean that car-based development increasingly will have to be described using other words that are less demanding of their global contribution.

See also: Environmental Economics; Sustainable Development: An Economic Perspective; Sustainable Transportation; Transportation Geography; Transportation Planning; Transportation, Supply, and Congestion; Urban Geography.

Bibliography

Anders, R., 1991. The Sustainable Cities Movement. Working Paper No. 2. Institute for Resource and Security Studies, Cambridge, MA.

Beatley, T., 2000. Planning for Sustainable Cities: The European Experience. Island Press, Washington, DC.

Berry, B., et al., 1974. Land Use, Urban Form and Environmental Quality. Research Paper No. 155. University of Chicago Press, Chicago.

Calthorpe, P., 1993. The Next American Metropolis: Ecology and Urban Form. Princeton Architectural Press, Princeton, NJ.

Cervero, R., 1999. Transit Planning. Island Press, Washington, DC.

City of Seattle, 1993. Sustainable Seattle: Indicators of Sustainable Community. City of Seattle, WA.

Coates, G., 1981. Resettling America: Energy Ecology and Community. Brick House, Andover, UK.

Dyck, R.G., 1998. Integrating planning and sustainability theory for local benefit. Local Environment 3 (1), 27–41.

Elkin, T., McLaren, D., Hillman, M., 1991. Reviving the City: Towards Sustainable Urban Development. Friends of the Earth, London.

Fischer, C.S., 1976. The Urban Experience. Harcourt Brace, New York.

Gordon, D. (Ed.), 1990. Green Cities. Black Rose, Montreal, PQ.

Gordon, P., Richardson, H.W., 1989. Gasoline consumption and cities: a reply. Journal of American Planning Association 55, 342–346.

Haughton, G., Hunter, C., 1994. Sustainable Cities. Taylor and Francis, Bristol, PA.
ICLEI (International Council on Local Environmental Initiatives), 1996. The Local Agenda 21 Planning Guide: An Introduction to Sustainable Development Planning. ICLEI, Toronto, ON.
ICLEI (International Council on Local Environmental Initiatives), 1997. Local Agenda 21 Survey. ICLEI, Toronto, ON.
Johnson, E., 1993. Avoiding the Collision of Cities and Cars. American Academy of Arts and Sciences with the Aspen Institute, Chicago.
Korten, D.C., 1990. Getting to the 21st Century: Voluntary Action and the Global Agenda. Kumarian, Bloomfield, CT.
Korten, D.C., 1995. When Corporations Rule the World. Kumarian, Bloomfield, CT.
Leitmann, J., 1999. Sustaining Cities: Environmental Planning and Management in Urban Design. McGraw-Hill, New York.
Mitlin, D., Satterthwaite, D., 1994. Cities and Sustainable Development. International Institute for Environment and Development, London.
Morris, D., 1982. Self Reliant Cities: Energy and the Transformation of Urban America. Sierra Club Books, San Francisco.
Naess, P., 1993. Energy use for Transport in 22 Nordic Towns. NIBR Report No. 2. Norwegian Institute for Urban and Regional Research, Oslo, Norway.
Newman, P.W.G., Kenworthy, J.R., 1989. Cities and Automobile Dependence: An International Sourcebook. Gower, Aldershot, UK.
Newman, P.W.G., Kenworthy, 1999. Sustainability and Cities: Overcoming Automobile Dependence. Island Press, Washington, DC.
Pathways to Sustainability Conference, 1997. In: Proceedings of UN Conference. City of Newcastle, New South Wales, Australia.
Pezzoli, K., 1997. Sustainable development: a transdisciplinary overview of the literature. Journal of Environmental Planning and Management 40 (5), 549–574.
Radberg, J., 1995. Termites heap or rural villages? the problems of urban density and sustainability. Proceedings of Conference, Copenhagen, April. In: The European City—sustaining Urban Quality. Ministry of Environment and Energy, Copenhagen, Denmark.
Real Estate Research Corporation, 1974. The Costs of Sprawl. Council of Environmental Quality, Washington, DC.
Roseland, M., 1998. Towards Sustainable Communities. New Society Press, Vancouver, BC.
Scheurer, J., 1999. An Evaluation of Danish Ecological Housing and Planning. Danish Building Research Institute, Horsholm, Denmark/ISTP. Murdoch University, Perth, WA.
Selman, P., 1996. Local Sustainability: Managing and Planning Ecologically Sound Places. Paul Chapman, London.
Slocombe, D.S., 1993. Environmental planning, ecosystem science and ecosystem approaches for integrating environment and development. Environmental Management 17 (3), 289–303.
Smith, M.A.F., Whitelegg, J., Williams, N., 1998. Greening the Built Environment. Earthscan Publications, London.
Stren, R., White, R., Whitney, J. (Eds.), 1992. Sustainable Cities. Westview Press, Boulder, CO.
STPP, 1994. A Citizens Guide to Transportation Planning and Livable Communities. Surface Transportation Policy Project, New York.
Tjallingii, S., 1991. The Responsible City. Paper presented at the International Federation for Housing and Planning. International Congress, Berlin.
Todd, J., Tukel, G., 1981. Rehabilitating Cities and Towns: Designing for Sustainability. Planet Drum, San Francisco.
Troy, P., 1996. The Perils of Urban Consolidation. The Federation Press, Sydney, NSW.
United Nations Center for Human Settlements, 1996. UN Global Review of Human Settlements. Habitat, Nairobi, Kenya.
Wachs, M., Crawford, M., 1991. The Car and the City: The Automobile, the Built Environment and Daily Urban Life. Michigan University Press, Ann Arbor, MI.
Wackernagel, M., Rees, W., 1996. Our Ecological Footprint: Reducing Human Impact on the Earth. New Society Publishers, Philadelphia, PA.
Walter, R., Arkin, L., Crenshaw, R. (Eds.), 1992. Sustainable Cities: Concepts and Strategies for Eco-city Development. Eco-Home Media, Los Angeles.
World Bank, 1996. Sustainable Transport—Priorities for Policy Reform, Development in Practice. World Bank, Washington, DC.
World Commission on Environment and Development, 1987. Our Common Future. Oxford University Press, Oxford.
Whitelegg, J., 1993. Transport for a Sustainable Future: The Case for Europe. Belhaven Press, London.

Planning Theory: The Good City and Its Governance

Patsy Healey, University of Newcastle, Newcastle, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This contribution reviews the discussion in the field of planning theory about the relation between cities, their governance, and planning interventions. It emphasizes the interaction between concepts about the nature, purpose, and method of planning; the context in which they have developed; and the practices at which they have been directed. It reviews changing ideas about the good city and about what constitutes good government, commenting on how concepts of place quality and governance capacity have coevolved. It then reviews three key themes in the planning field that have contributed to the development of these concepts: the relation between ideas and action, the interaction between the substance and process of planning ideas and interventions, and the transformative emphasis in planning theory. The piece concludes with a comment on the contribution of planning theory to transforming governance practices and enhancing place qualities.

During the twentieth century, the concept and practice of what variously is called town design, spatial planning, territorial management, and town and country planning has developed into a complex area of public policy. The theory informing this policy field has been the product of continual interaction among ideas about how to do planning, the experience of planning in practice, the intellectual context available to those thinking about planning, and the agenda of problems and opportunities that planning has been harnessed to address. Inherently, the planning enterprise is caught up in institutional contexts that shape its meaning, its material expression, and its consequences. This contribution illustrates the interaction between ideas about the nature, purpose, and method of planning; the context of their production; and the practices at which they have been directed.

Cities, Governance, and Planning

Few of the ideas about planning that evolved in the twentieth century in Europe and North America initially came out of academia. There was no significant academic infrastructure for the field until the last quarter of the twentieth century. The origins of twentieth-century planning ideas lay in the experience of rapidly urbanizing societies, and the perceptions of injustice, disorganization, and threats to human health and happiness, which urbanization seemed to bring. These perceptions promoted political movements, pressure group politics and expert evangelism, and attempted to transform city government to cope with the complexity of issues that arise in large cities. The focus of attention ranged from the neighborhood, the city, the region, the process of governance, and the overall organization of society. Two core ideas informed these otherwise-diffuse endeavors: first, that through appropriate governance action, it is possible to 'make a difference,' especially in making cities and societies more sustainable and just; and, second, that the spatiality of social processes and the qualities of place are important dimensions of well-being.

These two ideas have been developed in different ways, which are illustrated in the following sections, with the aim of highlighting both the situatedness of planning ideas and how their realization and impacts are highly dependent on the specifics of institutional milieux.

The Good City

In the early twentieth century, planning generally was understood as focused on cities, on how to manage the congested and sprawling consequences of the urbanization arising from industrialization and new transport technologies (Mumford, 1961). Throughout the century, there has been concern about the qualities of cities as living places, as aesthetic products, as business locales, and as biospheric relations. Theories of urban form and the arrangement of regional space dominated the discourses of city planning until at least the 1960s (Boyer, 1983; Hall, 1988). Protagonists, a mixture of campaigners and specialists in urban development (architects, engineers, and surveyors), battled over the relative merits of different forms. By the 1950s, most Western states had some form of urban planning and land use regulation embedded in their central and local government policy frameworks and competencies.

This success in institutionalizing the idea of city planning soon was overtaken by vigorous critiques of the predominant focus on urban form. Sociologists such as Herbert Gans in the United States argued that cities should be understood through the way people lived in them rather than through their physical form. Jane Jacobs, a fine observer of the way people used urban neighborhoods, argued against planners' ideas of demolishing old neighborhoods to replace them with modern buildings and the separation of living and working spaces. Regional economists and geographers challenged assumptions about the dynamics of urban growth embodied in the urban plans produced by architects, engineers, and surveyors. The emphasis of the earlier generation of town planners on achieving urban quality of life for the mass of the people was replaced by a concern with distributive justice and the 'socially just city.' This idea emphasized that the poor should not draw the short straw in the distribution of access to urban living space. By the 1970s, Marxist analysts had joined the debate. Manuel Castells showed, in a study of a French city, that the aim of urban planning to improve the quality of life was a mask behind

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74027-X

which developers and business interests could pursue strategies for realizing land value gains (Castells, 1977). This critical story had resonance in other parts of Europe where urbanization was proceeding apace, producing substantial development gains to landowners and developers. It was echoed in examples of urban renewal in the United States (Fainstein and Fainstein, 1986).

As a consequence of this critique, coupled with a slowdown in the pace of urban growth, in Europe at least, the idea of large-scale comprehensive planning fell into disrepute in the West. Urban management practices focused on the fine grain, the neighborhood, the site, and the urban renewal area. Attention in the 1970s and 1980s turned to projects and mechanisms. The focus on the physical form of the city was displaced by ideas about appropriate planning processes. Architects, engineers, and surveyors were replaced as a source of expertise for planning by regional economists, policy analysts, and those specially trained as 'planners.' A critique developed that suggested that the whole idea of attempting to 'plan' such a complex phenomenon as a 'city' was oppressive to the multiple realities and possibilities of urban life. Instead of designing the future city as a whole, attention turned to transformative innovations.

By the end of the century, a new wave of interest in the quality of cities, rural areas, and urban regions was building up in policy arenas and among commentators on planning (see Sandercock, 1998). The pressure for this came in part from an increasing awareness of new economic and social realities. The recognition of the openness of urban economies to the dynamics of global economic forces raised widespread concern about how to 'position' cities and urban regions in a landscape of competing regions. Traditional models of self-contained, hierarchically ordered cities no longer seemed to capture the reality of the complex urban agglomerations appearing around the world (Graham and Marvin, 2001). Most influential of all were the ideas associated with the rise of a new environmental consciousness in the last quarter of the twentieth century. This brought with it conceptions of the limits that the natural environment placed on human actions and of the need to consider transgenerational impacts not just short-term ones.

This led to a wave of interest in the 'sustainable city' (UN-Habitat, 2009). For some, this meant minimizing adverse impacts on natural environments, linked to ideas about 'ecological footprints' of urban development, and conceptions of the 'carrying capacity' and 'footprints' of environmental systems. For others, it emphasized keeping cities compact and reducing the land take for urban development. Others enlarged the concept to integrate the concern with social justice to that of 'environmental sustainability.' A version of sustainability that combined the principles of minimizing ecological impacts with ways of continuing to promote economic development had become a strong theme in public policy in Europe by the twenty-first century. But these two dynamics do not coexist comfortably. Tensions exist with notions of social justice, the multicultural city, and fostering innovative plurality. The reawakened emphasis on urban futures in planning theory thus opened up a complex set of issues that have since been explored vigorously by planning academics. Key themes in this exploration have centered on what constitutes the 'just city' (Fainstein, 2010) and on the contribution of informal development processes to improving place qualities (Roy, 2005). By this time, experiences from Asia and Latin America were beginning to have a significant impact on planning ideas.

These ideas and debates about the quality of the 'good city' and the 'good environment' were shaped by the changing dynamics of urbanization processes. In turn, concepts from planning theory have influenced the practices of urban management. Planners' ideas have molded the restructuring of existing urban areas, the design of new towns and settlements, and residential neighborhood design. They have affected the form and location of transport projects. They have shaped the evolution of urban street scenes through the impact of land use regulations. These regulations in turn have influenced the way land and property values are distributed and property development processes have evolved. Planners' ideas achieved this impact by framing the organizing principles and discourses used by governments in developing policy frameworks and organizing competences for planning 'systems.' The organizing principles and discourses of one era become embedded in practices and can come to outlive their usefulness. By the turn of the millennium, in many parts of the world, 'planning' as practiced was identified as 'part of the problem' of governance processes that do not match the aspirations and challenges of contemporary cities and regions (UN-Habitat, 2009).

Good Government

For the protagonists of early twentieth-century city planning, the everyday business of politics and government was 'part of the problem.' They sought to show a way forward by striding out ahead with their plans and designs, apart from this messy world. By midcentury, an alternative idea had taken root. Planning was a way of 'doing government' itself. In Europe, socialist movements advanced the role of planning as an alternative to allocating resources through market processes. In the United States, planning was promoted as a way to improve the quality of local government. Clientelist practices were well-developed in many cities, with adverse consequences for both the efficiency and effectiveness of public policy. Bureaucratic management was no better, as 'playing by the rules' led to government actions that did not adapt to the dynamics of changing situations or local specificities. The way forward, strongly advocated in the University of Chicago School of Planning in the 1950s, was to adopt principles of scientific management. This involved articulating the goals of an organization and translating them, through systematic steps, into policy principles or objectives that individual officials could use as guides as they developed responses appropriate to specific situations. This powerful idea, with its focus on 'means-ends relationships,' was strongly linked to the promotion of democratic government. Officials could be held to account in relation to their performance in achieving an organization's goals, whereas the process of strategy-formation made these goals and their justification transparent to citizens (Faludi, 1973; Friedmann, 1987).

Initially, it was assumed that the work of 'policy planning' would take place in a world of technical experts, reporting to politicians (Dahl and Lindblom, 1953). As Western democracies became increasingly educated and

pluralistic, distributing power and knowledge ever more widely among business groups, pressure groups, and citizen mobilization movements, this position has been challenged. Concepts of strategy, of evaluation of policy ideas and performance in relation to objectives, and of the systematic link between policy and implementation had become deeply embedded in expectations of government behavior in the West by the end of the twentieth century.

At the core of the original idea of a 'rational planning process,' as it was often called in the 1960s and 1970s, was the view that governments, and their teams of experts, needed to be able to explain why a particular strategy was appropriate in a particular context, how it would be implemented and whether it had been implemented. This idea became linked to the development of ways to make strategic and policy plans, create techniques of evaluating alternative options, appraise the impacts of projects, develop indicators to evaluate performance, and attempt to model situations, for example, land use and transport interactions, to assess the likely impacts of different courses of action. All these practices are still in common use in twenty-first-century urban and regional government, providing a methodological toolkit for planning office staff and consultants.

These early ideas attracted criticism almost as soon as they were propounded. The overall approach was challenged as deeply instrumentalist – defining the purpose of government in terms of achieving specific objectives effectively. The methodologies drew on welfare economics, with its utilitarian assumptions. The approach assumed that policy ideas were translated into action in a linear way, from top to bottom of an organization or government system. It was seen to be overgeneralized, failing to consider the differences that might arise in different policy fields and in different institutional contexts. It handed power to the experts, as the early city planning ideas had done. Politicians, lobbyists, and citizens were pushed to the sidelines. In the 1950s, a famous study by Meyerson and Banfield (1955) in Chicago showed that local politicians were not prepared to accept the restraints on their actions that the rational policy-planning model implied. In the 1960s and 1970s, scholars such as Charles Lindblom, Amitai Etzioni, John Dyckman, and Richard Bolan began to show how contexts might limit how far a policy-driven planning practice might go. An alternative was to accept the increasingly pluralistic form of local politics, and to develop a local politics around competing plans, produced by different interest groups. This approach, promoted by Paul Davidoff, became known as advocacy planning (Davidoff, 1965). The experience of policy practice began to look less and less like a strategy-driven process. Instead, analysts noted that the content of policy seemed to be negotiated in the flow of action, rather than being set in advance.

By the late 1970s, planning theorists turned their attention to issues of policy 'failure' and implementation 'gaps,' following landmark studies by Aaron Wildavsky and others. Wildavsky's work looked more closely at the institutional context in which policy frameworks were articulated and the ways planners, understood as experts and government officials, went about their tasks. It identified the problems that arose in cases in which policy ideas were not sufficiently 'owned' or understood by politicians, or by other parts of public administration that controlled significant resources or regulatory powers. There were close parallels with work undertaken in business studies. This analysis highlighted the difficulty of promoting strategy-driven governance processes where other ways of 'doing governance' controlled the culture of local government. It also showed that field-level actors, officials, and experts working up the details of policies, or acting at the interface between officials and the clients of public administration, developed their own ways to understand policies and how they might be applied.

This led to a critique of the rational planning process as a top-down viewpoint, challenged by a bottom-up viewpoint of the actors on the ground (Barrett and Fudge, 1981). The linear model of the policy process increasingly was challenged by conceptions of policy development as an interactive process, continuously evolving over time and drawing in different actors at different stages. In this formulation, planning was no longer either outside government or imposed on governments in an attempt to transform their processes. It was caught up inside politics and governance processes. In the 1980s, a significant strand of work associated with John Forester, Charles Hoch, and Howell Baum, began to address questions of the way planners committed to progressive values operated in these ethically challenging situations. This development, which became known by the end of the twentieth century as 'communicative planning theory' (Innes, 1995), was underpinned by an epistemological shift, drawing on nonpositivist philosophy, particularly pragmatism and the social theory of Anthony Giddens and Jürgen Habermas.

This philosophical shift, which by the 2000s had become widespread in the social sciences, focused attention on the construction of meaning in social contexts. Policy development was presented less as a technical exercise in calculation and design, and more as a social process to articulate ways of thinking and ways of acting that could command support among a wide range of players in urban and regional contexts in which power was likely to be diffused in complex ways rather than concentrated in the hands of a single elite. Key scholars in the planning field drew on the interpretive tradition as it was evolving in policy analysis (Wagenaar, 2011). They focused on the potential of collaborative and participative forms of policy making that were arising in many contexts to address complex and contested environmental and social dilemmas (Hajer, 1995; Hajer and Wagenaar, 2003; Healey, 2006).

The interest in collaborative and participative governance practices was partly a response to the search for deepening democratic practices, and there are close links between this work and the wider development of the idea of deliberative democracy. The interest was promoted by a critique of the paternalist and bureaucratic practices into which government in Western countries seemed to have sunk. Some argued for an overall reduction in the burden of government. This neoliberal trajectory was challenged, however, by the advocates of more participative democracy, who sought to reduce the perceived gap between the worlds of politics and government and the worlds of citizens.

By the late 1990s, ideas about consensus-building practices and collaborative strategy formation, collected and consolidated by planning theorists, were being rapidly circulated

internationally to help fulfill in practical ways the ambitions of politicians and officials trying to move in new directions. In this context, the value of policy effectiveness embodied in the rational planning process was supplemented by a concern with building democratic culture (respectful and citizen-centered styles of governance). During the first decade of the twenty-first century, there was an explosion of research about, and critique of, collaborative policy-making practices. Meanwhile, other contributions within the broad areas of postpositivist philosophy and the analysis of complex systems were emphasizing the unpredictability of the way futures emerge, and therefore the need for flexible, innovative governance responses (Hillier, 2007). These two strands coalesced in increasing interest in work on network governance and the governance of complex adaptive systems (Sorensen and Torfing, 2007; Innes and Booher, 2010).

The Coevolution of Place Quality and Governance Capacity

A key issue in these vigorous debates about place governance was the nature of the institutional context in which such new practices could flourish and how far their promotion could induce changes in governance processes and cultures. These developments lead to an interest in the different experiences of strategic planning initiatives and neighborhood transformations, notably in comparative work in Europe (Moulaert, 2000; Salet et al., 2003; Healey, 2007). This work centered on the transformative dynamics of governance capacity building, and the interaction between the content of policy initiatives and the governance processes through which such initiatives are realized. There remains a tendency, however, for the discussion of what makes a 'good city' to proceed separately from that of 'good governance,' Some agencies such as the World Bank and the United Nations have articulated universal principles of good governance.

The insight from work in the planning field, however, suggests that such universal statements neglect the situated contingencies of particular localities. Both what are valued as 'good' place qualities and 'better' ways of doing governance work to promote their realization are rooted in specific histories and geographies. These institutional dynamics need to be understood if transformative initiatives are to have a chance of achieving at least some of their ambitions. It is also clear, especially from examples in former colonized territories, that formal 'planning systems' inherited from the past may be part of the problem undermining place qualities in particular places, impeding effective governance, and sustaining unjust and environmentally destructive practices (UN-Habitat, 2009).

Planning Theories in Context

The enterprise of planning is profoundly committed to changing the ways things are, to transforming realities, whether those of cities or their governance. It is inherently normative in its value orientation, infused with the belief that deliberately designed actions today can have beneficial impacts tomorrow. Those who reject the idea that this is desirable or possible – anarchists, extreme free marketeers, and cultural postmodernists – will reject the enterprise of planning. Yet the idea that, in democratic states, governments should have policies and should be judged on how effectively they follow them through, is one of the great inventions of the twentieth century. It is being carried forward vigorously into the next century. What is changing, however, are the conceptions of where policies should come from, how they should be focused, and who gets involved in their articulation and implementation. These days, it is not enough for policies to be designed by experts. The quality of government increasingly is judged by the way its policy formation attends to the multiple voices of complex societies; by its skill in recognizing the complexity of economic, social, and environmental dynamics; and by its appreciation of the specific realities of the way policy programs are realized.

The contribution of planning theory to this enterprise has been to focus on critical themes surrounding the challenge of transformative collective action aimed at enhancing place qualities and improving governance cultures and practices. Next, I discuss three such themes.

Ideas and Actions

The first challenge focuses on the relation of theory to practice. The history of planning thought shows the many ways in which planning ideas filter into the construction of concepts, goals, and norms that guide specific actions: (1) through advocacy and campaigning that affects popular perception and politicians' agendas, (2) through the design of planning systems and formal norms, (3) through the ways these are interpreted in administrative and legal arenas, (4) through the formulation of explicit strategic frameworks and the ways these get translated into routinely used concepts and norms, and (5) through the training of 'planners' and the professional organizations that bring 'planners' together. Through continual interaction with these worlds, planning ideas themselves are changed. The contribution of planning thought in the twentieth century has been its preparedness, however ineffectively it often is developed, to link together the philosophy of normative action, the social science of urban and regional dynamics, and the fine grain of the practices of urban and regional governance. The challenge has been to draw out the significant strategic processes of governance that affects the qualities of places and territories, as lived, experienced, and given identity in multiple social processes (Friedmann, 1987).

Substance and Process

The geographer Ed Soja (1989) has remarked that twentieth-century thought has had a peculiar difficulty in perceiving the impact of 'being in a place' and of the spatiality of social relations. The notion of the autonomous individual in a 'free society,' or the social group as a 'class' in relation to other social classes, has been much more dominant. Contemporary culture, however, notices all kinds of cleavages and fractures, and stresses the complex relations individuals may have with others. Grasping the significance of place and territory in these complex representations and relations, and developing governance processes in which to give them voice, is a major challenge for urban and regional politics and policy (Massey, 2005). In planning theory, there has long been a debate

about how to relate the 'substance' of policy to the 'process' of policy development, with theories being criticized for giving too much weight to one or the other.

A second challenge for planning theory is to contribute to the coevolution of the 'substance' and 'process' of place governance. This has major implications for institutional design – who contributes what, when, and where. It also has implications for the articulation of values. There were major debates in the mid-twentieth century about whether planning was 'value-free' or 'value-full.' If 'value-full' were the values about process (fair, just, rational) or about substance (achieving specific goals, e.g., aesthetically pleasing environments, quality of life for ordinary people, redressing the unequal distribution of resources and power)? By the turn of the century, process values had been extended to include qualities of knowledgeability, inclusion, empowerment, and respect to all citizens, while substance values were infused with ideas about environmental sustainability and social justice. If the substance and process of governance for complex territorial problems are seen to coevolve, then it must be expected that values, too, coevolve. Several planning theorists have emphasized the importance of processes of social learning in action. Donald Schon introduced the important idea of the skilled practitioner reflecting in action (Schon, 1983), and John Forester has produced a corpus of work about the way practitioners go about 'doing' planning work (Forester, 1999, 2009).

System Maintenance or Transformation

The third challenge in planning theory lies in the approach to social change. Behind any idea of the 'good city' or 'good government' is some model of the nature and dynamics of social organization. For some advocates of planning ideas, it is the model itself that needs change. For others, the model needs correction to make it work better. As John Friedmann argued in the 1960s, some planning ideas sought to maintain basic systems. Others aimed to transform them. At the time, he was contrasting the management needed to trouble-shoot in Western democracies with the transformations needed in illiberal Latin American dictatorships, newly experimenting with democracy. By the 1970s, he had become much more radical, demanding 'social transformation' in overurbanized capitalist societies, dominated by global corporations. His work has resonance in many places experiencing the massive urbanization processes under way by the 2000s (Friedmann, 1973, 2011).

Throughout the twentieth century, people have been attracted to the planning field through transformative idealism. In the second part of the century in Europe, the target of transformation was a form of capitalism that dominated life chances, structured the way urban and regional space was used and damaged environments across the globe. For many, it was a shock to discover that the 'planning systems' and 'planning practices' that actually evolved seemed deeply committed to maintaining the very practices that they had sought to challenge. Ideas about alternative ways of organizing society, however, come up against the reality of the dynamics of social change. In some periods of tumultuous social change, any kind of planning is an impossibility. In some periods of entrenched stability, similarly, there are few cracks of opportunity to mobilize to do things differently. Most of the time, most situations are somewhere in between. In the contemporary world of complex relationships and diffused power, change happens both through big events (e.g., the 1970s Organization of the Petroleum Exporting Countries oil crisis, the collapse of communism, the 1987 property crash, the rise to power of neoliberal governments, and the financial crash and subsequent recession from 2008) and through lots of small shifts and mobilization episodes. Planning processes can help these shifts along. Instead of making plans that are assumed to be predictions of what will happen, plan making becomes more of an activity of social mobilization, encouraging new ways of understanding situations and new practices. Strategy development in a world understood as emerging in an evolutionary way takes on more of the character of collective invention and risk-taking in trying to shape futures rather than the mid-twentieth-century assumption that government could command resources and control futures (Healey, 2007; de Roo and Silva, 2010; Hillier, 2011).

The Contribution of Planning Theory

These debates in the planning field provide rich resources for imagining, developing, and critiquing governance practices as they affect urban and regional dynamics in the twenty-first century. Formalized in the planning theory literature, they range from philosophical debate to analysis of systems and practices and discussion of the situation of 'doing planning work.' In addition to its normative commitments, the planning field needs to find ways to combine imaginative foresight to help people think about how to work toward alternative futures with an astute and multifaceted engagement with evolving practices.

There are many windows through which this engagement has developed. One is that of the 'planner,' working within specific contexts and trying to make things happen differently. This has been the key contribution of John Forester, whose work speaks sympathetically to planners in complex situations around the world. Another is a focus on governance practices. In recent years, this work has been linked closely with the disciplines of urban politics and urban and regional geography, exploring questions about the relation between politicians, political regimes, experts, and citizens, and the way planning strategies and tools have been used in particular institutional milieux. A third window has focused on understanding the design of governance systems. This has examined competencies, law, resource allocation systems, and appropriate expertise, the top-down levers of policy change, and how these interact with bottom-up interpretive practices. Much of this work has been informed by the search for ways of changing governance practices. A fourth window has concentrated on ideas about the good city and the good society, especially in relation to ideas of environmental sustainability and multicultural societies, and on how these ideas have worked out in practice. The critical challenge for the development of planning as an enterprise is to recognize that these different windows do not represent separate compartments of the planning field. They merely cast a more intense lens over a particular set of

relations. All the others are still there, and need to be recognized as present, even if not the focus of a particular study.

The role of planning theory, as a distinctive task in this complex and dynamic institutional terrain, is to provide a critical imagination and structuring ideas for the dynamic invention processes that continually are unfolding in political communities in cities and regions. It should offer evaluative vocabularies for the recording and critiquing of these processes, both in relation to their content and their procedures. It should interpret emergent patterns and assess their consequences. It should draw in ideas from other fields and translate them in ways that are understandable to the various 'thought-worlds' caught up in city governance. It should consolidate the distinctive understanding evolved by engagement with this activity and make it available to those working in other fields. This implies that the task is always an in-between one. If the practitioners of planning theory withdraw too far from the worlds of those engaged in doing the planning, then they lose their capacity to record, shape, connect, and translate. If they stay too close to the doers, however, they are unable to see beyond what the doers themselves see, and therefore cannot add value to the work of doing. The challenge for planning theory in the twenty-first century is thus to enrich the evolution of public policy focused on place quality and to do this in a way that fosters the flow of ideas among the many arenas engaged in shaping urban and regional governance.

See also: Deliberation and Democracy; Just City; Pragmatism: Philosophical Aspects; Urban Informality: The Production and Regulation of Space.

Bibliography

Barrett, Susan, Fudge, Colin, 1981. Policy and Action. Methuen, London.
Boyer, Christine, 1983. Dreaming the Rational City. MIT Press, Boston, MA.
Castells, Manuel, 1977. The Urban Question. Edward Arnold, London.
Dahl, Robert A., Lindblom, Charles E., 1953. Politics, Economic and Welfare, fourth ed. Transaction Publishers, New Brunswick.
Davidoff, Paul, Nov 1965. Advocacy and pluralism in planning. Journal of the American Institute of Planners 31, 331–338.
de Roo, Gert, Silva, Elisabete A. (Eds.), 2010. A Planner's Encounter with Complexity. Ashgate, Farnham, Surrey, VT.
Fainstein, Susan, Fainstein, Norman (Eds.), 1986. Restructuring the City: The Political Economy of Urban Redevelopment. Longman, New York.
Fainstein, Susan, 2010. The Just City. Cornell University Press, New York.
Faludi, Andreas, 1973. A Reader in Planning Theory. Pergamon, Oxford.
Forester, John, 1999. The Deliberative Practitioner: Encouraging Participatory Planning Processes. MIT Press, London.
Forester, John, 2009. Dealing with Differences: Dramas of Mediating Public Disputes. Oxford University Press, Oxford.
Friedmann, John, 1973. Re-tracking America: A Theory of Transactive Planning. Anchor Press, New York.
Friedmann, John, 1987. Planning in the Public Domain: From Knowledge to Action. Princeton University Press, New Jersey.
Friedmann, John, 2011. Insurgencies: Essays in Planning Theory. Routledge, London.
Graham, Stephen, Marvin, Simon, 2001. Splintering Urbanism. Routledge, London.
Hajer, M., 1995. The Politics of Environmental Discourse. Oxford University Press, Oxford.
Hajer, Maarten, Wagenaar, Hendrik (Eds.), 2003. Deliberative Policy Analysis: Understanding Governance in the Network Society. Cambridge University Press, Cambridge, MA.
Hall, Peter, 1988. Cities of Tomorrow. Blackwell, Oxford.
Healey, Patsy, 2006. Collaborative Planning: Shaping Places in Fragmented Societies, second ed. Macmillan, London.
Healey, Patsy, 2007. Urban Complexity and Spatial Strategies: Towards a Relational Planning for Our Times. Routledge, London.
Hillier, Jean, 2007. Stretching Beyond the Horizon: A multiplanar Theory of Spatial Planning and Governance. Ashgate, Aldershot.
Hillier, Jean, 2011. Strategic navigation across multiple planes: towards Deleuzian-inspired methodology for strategic spatial planning. Town Planning Review 82 (5), 503–528.
Innes, Judith, 1995. Planning theory's emerging paradigm. Journal of Planning Education and Research 14 (3), 183–189.
Innes, Judith E., Booher, David E., 2010. Planning with Complexity: An Introduction to Collaborative Rationality for Public Policy. Routledge, London.
Massey, Doreen, 2005. For Space. Sage, London.
Meyerson, M., Banfield, E., 1955. Politics, Planning and the Public Interest. Free Press, New York.
Moulaert, Frank, 2000. Globalisation and Integrated Area Development in European Cities. Oxford University Press, Oxford.
Mumford, Lewis, 1961. The City in History. Penguin, Harmondsworth.
Roy, Ananya, 2005. Urban informality: towards an epistemology of planning. Journal of the American Planning Association 71 (2), 147–158.
Metropolitan governance and spatial planning. In: Salet, Willem, Thornley, Andy, Kreukels, Tom (Eds.), Comparative Studies of European City-Regions. E & FN Spon, London.
Sandercock, Leonie, 1998. Towards Cosmopolis. John Wiley, London.
Schon, Donald, 1983. The Reflective Practitioner. Basic Books, New York.
Soja, Edward W, 1989. Postmodern Geographies: The Reassertion of Space in Critical Social Theory. Verso, London.
Sorensen, Eva, Torfing, Jacob (Eds.), 2007. Theories of Democratic Network Governance. Palgrave Macmillan, London.
UN-Habitat, 2009. Planning Sustainable Cities: Global Report on Human Settlements 2009. Earthscan, London.
Wagenaar, Hendrik, 2011. Meaning in Action: Interpretation and Dialogue in Policy Analysis. M.E.Sharpe, New York.

Plato (428 BCE – 347 BCE)

Thomas Arnold and Jens Halfwassen, University of Heidelberg, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Plato is one of the most influential figures in the development of European thought. This article provides a short biography, deals with the complexities of interpreting Plato, and covers the main areas of Platonic thought (ontology, epistemology, ethics, and politics) and the major Platonic doctrines (Dialectic, Theory of Forms, and Unwritten Doctrine).

Introduction

Plato was born in 428 or 427 BC in Athens amid the turmoil of the Peloponnesian War and at the end of the Golden Age of Athens. He witnessed the decline of Athenian power and the violent change of political systems from democracy to oligarchy and back. However, as the offspring of one of the noble families of Athens he was also privy to the remarkable cultural achievements of Greek culture. Before meeting Socrates he was already acquainted with Cratylus, a disciple of Heraclitus and is likely to have had an excellent education in music, gymnastics, and grammar. After the execution of Socrates in 399 BC Plato undertook several voyages; although the details are hazy, he probably traveled to Megara, Cyrene, Egypt, and Magna Graecia, meeting with philosophers and politicians. Most notably, Plato became involved with the Sicilian court. Even though his first trip to Syracuse in 388 BC turned into a disaster (with some sources claiming that Plato was sold as a slave and only freed by chance), he returned twice, in 366 and 361 BC, still hoping to establish a form of philosophical government, but finally failing to achieve this goal. He died in 348 or 347 BC after a lifetime of teaching and research.

After his first trip to Sicily, Plato founded what we now call the Academy in the grove of a local hero, Academos, close to Athens. And even though Plato's Academy was a well-respected place of teaching, learning, and living, it was not just a school in the narrow sense, but also an 'academic' research institute in mathematics, astronomy, philosophy, law, grammar, and other subjects; Aristotle being its most famous student. While the site was destroyed by Sulla in 84 BC (when he needed the wood of the olive trees to build siege weapons to take Athens) the idea of the Platonic School lived on through later antiquity and the academy was reestablished and staffed by Neoplatonic philosophers until Justinian I had it closed in AD 529 as a threat to Christianization.

In developing his philosophy Plato could draw on many different strands of thinking, most of which are represented directly in his dialogues through proponents of the particular school of thought. The Heracliteans offered the idea that everything is in flow, Parmenides and Zeno held the idea that Being is one and eternal, the Pythagoreans thought that everything is ordered through numerical structures, and Anaxagoras saw the mind or intellect as the foremost principle of Being. Plato's most profound and famous source of inspiration is, at the same time, the one hardest to track, because Socrates, although present in nearly all Platonic dialogues and protagonist in most, never produced any written testimony of his philosophy. And even if Socrates appears in the works of others (Aristophanes and Xenophon) as well, his original teachings are still a matter of debate. If we judge the *Apology* to be the Platonic work closest to representing the original Socrates, his main concern was indeed the search for wisdom (*sophia*) (see Section The Meaning of Philosophy).

Plato's intellectual horizon is however not restricted to philosophy in our narrow sense. He alludes to mysteries and cults, quotes Homer, other poets, and playwrights. His interest in mathematics also shows up throughout the dialogues, where he often uses mathematical examples to illustrate philosophical points. The close connection between mathematics and philosophy is also echoed by the characters within the dialogues, as one of the most gifted interlocutors is the young Theaetetus, a friend of Plato and an eminent mathematician.

Another major influence on Plato's work is sophism. The term 'sophist' applied to a very diverse set of people, ranging from sages and scholars to lawyers and what we would call management consultants. They would offer practical education for certain fees, specializing in rhetoric and everything necessary for a successful career in politics. Plato's philosophy can be understood to a large extent as a reaction to sophism, especially its relativism (naturalism being the other opponent). Many sophists regarded truth and values as something merely relative, something to be negotiated by skilled orators depending on the occasion. Against this Plato sought to establish absolute truths and values, unveiled and defended by rigorous philosophical enquiry. Accordingly, many of Plato's dialogues center on discussions between Socrates and the sophists – some friendly, some rather heated.

Alfred North Whitehead's famous dictum sums up Plato's influence beautifully: "The safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato. I do not mean the systematic scheme of thought which scholars have doubtfully extracted from his writings. I allude to the wealth of general ideas scattered through them. His personal endowments, his wide opportunities for experience at a great period of civilization, his inheritance of an intellectual tradition not yet stiffened by excessive systematization, have made his writings an inexhaustible mine of suggestion" (Whitehead, 1929: p. 39). Plato's importance stems not just from his theories, but from the fact that he set the field for Western philosophy. He influenced Neoplatonism, Christian theology, the philosophy of the Renaissance, and German Idealism (Fichte, Schelling, Hegel) and continues to influence even those who disagree with or ignore him.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61261-8

Hermeneutical Complications

Even though Plato is one of the few writers of antiquity whose complete works we know, there is no straightforward way of telling 'what Plato thought on X' due to two facts. Firstly, Plato never wrote treatises but dialogues (and a few letters) and secondly, he offers a powerful critique of the written word as a medium of philosophy.

The Dialogues

Plato's works might have had the purpose of encouraging readers to engage with philosophy and visit the Academy, but they are not 'academic' in our narrow sense in that they are not treatises laying out the last word of the author. They are masterfully crafted, witty dialogues in which we encounter very different people talking and discussing many issues with one another, often straying from the original topic just to end up with a philosophical conundrum of one sort or the other. This form of presenting philosophy has several advantages, not the least, readability. It also draws attention toward the processual unfolding of thought and matters of philosophical performance, rhetoric, and style; terminology can be kept fluid and the reader is forced to form his own opinion on the matters discussed, thus making Plato's work an ideal entry point to philosophical discourse.

Which topics are discussed and what mistakes are being made in any given dialogue depends heavily on the cast and situation of the dialogue in question. If you are talking with an arrogant, obtuse anti-intellectual, the discussion probably would not go as well as with an interested, smart, and open-minded friend; and even debates between mutually acknowledged experts would not be free of errors. Importantly, the level of sophistication of the arguments rests on the intellectual capabilities of the interlocutors. Thus, Socrates might confront an opponent with a line of reasoning we easily unmask as faulty but which trips up his interlocutor nonetheless, while the mindful reader can pick up subtle clues leading to the solution of a problem. Moreover, Socrates might discuss the same issue with someone else in another dialogue and come up with very different theses. Sometimes, Socrates explicitly chooses to withhold certain ideas due to lack of understanding on the side of his partner. However, the fact that Plato depicts shifty rhetoric, false reasoning, and omission of information does not allow us to conclude his own reasoning to be faulty or incomplete.

Also noteworthy is his steady usage of myths in the dialogues, which pose difficult puzzles of their own, due to the fact that they often stand in for their exact opposite, that is, logical explication.

Critique of Writing

Interpretation is further complicated by Plato's critique of writing. In the *Phaedrus*, Socrates explains why the written text in general is an inadequate medium for philosophy as it cannot explain and elucidate itself but always says the same; it is unable to decide who to impart information upon and it cannot defend itself. Text is dead, and can only ever function as an aid to memory (Pheadrus 275). This does not just sound unfamiliar to someone used to academic communication in the form of books and journals, but it also poses a conundrum for the interpreter. Does Plato's critique encompass all writing, even his own? If so, can we take it seriously insofar as it is presented in a dialogue obviously written with a reader, not a listener in mind? The passage itself makes no exceptions to any form of text so we are forced to assume that Plato considered his own written works as secondary. Until the nineteenth century it was indeed commonly accepted that Plato's main metaphysical doctrine was unwritten or only hinted at in the dialogues, but nonetheless taught, discussed, and passed down in the Academy. Many sources, including Aristotle, support this view. The unwillingness to acknowledge an unwritten doctrine behind and beyond the dialogues stems from a strand of romanticist philology centered around certain hermeneutical ideas on the infinity of interpretation and the urge to free Plato's thought from a (misconceived) notion of esotericism.

These hermeneutical complications lead to an interestingly wide range of contemporary readings of Plato, focusing on very different aspects of his work and based on very different premises. Yet, even though we should be cautious of attributing theorems to Plato, we can nevertheless identify certain recurring themes and ideas that seem to be basic parts of his thought, regardless of our specific interpretative background.

The Meaning of Philosophy

Wisdom (*sophia*) is opposed to folly (*amathia*), ignorance of one's own epistemic limits. The love or search for wisdom is thus first of all a critical undertaking during which the philosopher questions his own beliefs and practices as well as those of his fellow human beings. Socrates describes himself as the barb or fly (*myops*) that keeps the noble but lazy horse Athens awake with these kinds of investigations (Apology 30). He strolls around Athens examining the knowledge claims of politicians, artists, and artisans, asking for justification and explanation (*logon didonai*), exposing their ignorance by leading them into *aporia*, "paralysing them like an electric ray" (Meno 80). How the interlocutor reacts to such a shocking *reductio ad absurdum* – with angry rejection or interest – shows whether he is philosopher-material or not. An important step in the search for wisdom is to admit one's ignorance and indeed this is the only epistemic advantage Socrates primarily makes a claim on: "He does not assume to know something where he really doesn't" (Apology 21) and "trusts the expert rather than the public opinion" (Crito 43).

Yet philosophy is not merely a negative or skeptical undertaking, quite the contrary. To avert epistemic overreach one needs to reflect on the nature and criteria for knowledge (*episteme*) as against mere traditional belief (*doxa*) and the way (*methodos*) of transforming the latter into the former (Meno 75); one aspect of philosophy is thus epistemology, the quest for the knowledge of (the nature of) knowledge (Charmides 169). The movement from belief to doubt to reflection to understanding and so forth is illustrated by the dramatic form of the dialogues.

This concern with epistemological and methodological matters (Politicus 286) sets Plato's philosophy apart from

what came before, and it also sets philosophy as a whole apart from other epistemic enterprises, such as mathematics. Plato never disputes the effectiveness or methodological autonomy of mathematics, but he draws attention to the fact that mathematics qua mathematics cannot found itself. For while the mathematician proceeds deductively from axioms, he is not able to answer certain questions regarding his presuppositions (Politeia 510), like "What is the ontological status of mathematical objects?" or "How can concrete signs represent abstract entities?" or – more generally – "What is knowledge?" (Theaetetus 145). Only philosophical reflection (as against mathematical deduction or empirical induction) catches up with its own preconditions and proceeds to the unconditioned, that is, the Absolute (see Section The Unwritten Doctrine). Thus, Plato introduces the idea of philosophy as a presuppositionless inquiry.

Finally, doing philosophy means learning to die. Death is the final separation of body and soul, an event during which the true human sheds the mortal hull in which it is entombed (Pheado 82) or by which it is encrusted like the statue of a sea-god is encrusted in clams (Politeia 611) – a separation the philosopher already seeks in life as he strives to use his mind without being affected by bodily urges or dominated by the lower, feral, and illogical parts of his soul. For Plato, philosophy ultimately is care for the soul through which we achieve an existence ordered and led by reason; philosophy brings us closer to the 'likeness to god,' that is, absolute rationality (Theaetetus 176). Therefore philosophy for Plato is never just a theoretical affair, but a thoroughly practical, life-encompassing matter.

Dialectic and the Forms

Philosophy proceeds first and foremost by talking (*dialegesthai*), thus the method of philosophy is aptly named 'the art of talking,' that is, 'dialectic.' A philosophical dialogue is not a rhetorical match – its aim is understanding, not winning (Gorgias 454); the first dialectical virtue is therefore the ability to properly raise and answer questions (Cratylus 390) in a way the interlocutor understands (Meno 75). Philosophical discourse tries to answer essential questions, so the second dialectical virtue is the ability to find the essential features of the issue in question (Politeia 534) by distinguishing (*diairesis*) genus and species (Phaedrus 265; Sophist 253). The third dialectical virtue concerns the attitude toward (pre)suppositions. The philosopher tries to explicate his presuppositions and consciously posits a thesis to see how far it carries and how firm it stands; if the thesis comes under attack, he will revert to a higher, more abstract thesis founding the initial thesis (Pheado 101). The goal is to reach an ultimate position, from which no further enquiry is possible.

Dialectic aims at uncovering what is. But what is cannot be grasped with the senses (Phaedo 78), as everything sensible is necessarily in a (Heraclitean) state of flux. The true objects of understanding and knowing are not individual things but their Forms (*morphe*) or Ideas (*idea*/*eidos*). When we see a table we see a steadily changing thing, yet we understand the table as table, that is, we do not just stare at a heap of matter but we grasp the general meaning 'table,' the table-ness of this particular table, so to speak. We understand the thing in front of us as partaking in the Idea of the table. This idea or form is also what the artisan has in mind while fabricating a thing; for when this breaks he will remake it not looking at the broken thing but at the idea of it (Cratylus 389). Similarly, with mathematics, the mathematician does not care about the chalk figure on the blackboard but about the abstract entity of a triangle. The same holds for philosophy. The philosopher is not interested in concrete instances of, say, virtue but he wants to identify its essence. Often Socrates' interlocutor runs into difficulties because he does not grasp the difference between Idea and example, naming examples, but failing to offer an essential definition. The theory of Forms thus functions first and foremost as a corrective, leading the partners of the dialogue away from the all-too-concrete toward more general notions as the true prize of dialectical enquiry.

Ontology

The theory of Forms is often taken as a variety of realism, a misinterpretation going back to Aristotle. According to this view Plato envisaged two realms, one 'down here,' full of sad and meager copies of the Ideas in the other realm 'up there'; in the heaven of Forms we find ideal entities, perfect horses on perfect lawns, perfect beds in perfect bedrooms, etc. This, of course, neither makes any sense nor is it endorsed by Plato. In some sense, the Form of X is indeed X; the Idea of the table is certainly table-y in the highest sense, the epitome of table-ness. On the other hand, it certainly is not a real table, as a table is made of some material and can be destroyed, occupies a position in space time, owns a definite number of legs, etc. The idea of the table is not made of any material, it cannot be destroyed and cannot possibly be located somewhere with a definite number of legs as that would exclude it functioning as the Form for all tables everywhere with every number of legs; Forms are 'one above many' (*hen epi pollon*). The term 'realism' is a complete misnomer as the Forms are not things (*res*) as Plato himself makes clear in the first half of the *Parmenides* – even if his talk of 'images,' 'copies,' and 'likenesses' invites these misunderstandings. At the same time, Plato's alternative is not some kind of nominalism – Platonic Ideas are not mere Lockean ideas either. They exist independently of us and our idea of table-ness is directed toward the Idea of table-ness and our concept of table is informed by the Form of the table; the Ideas are not identical with our thoughts (Parmenides 132).

Plato does however employ a descriptive dichotomy between being, and becoming, epitomized in the analogy of the line (Politeia 511). Imagine a line, split in two according to the golden ratio. The upper half represents being and the lower, becoming. The upper half is correlated to thought and the lower to perception. While actual tables can be perceived by the senses and are subject to change (becoming), the table Form is immutable (being) and can only be grasped in thought. The two halves are then again halved according to the golden ratio, whereby the two realms of being/thought and becoming/perceiving are further differentiated. The realm of being now includes Forms in the strict sense (at the highest section of the line) and mathematical objects (one rank below, but still located in the region of being); to the Ideas

corresponds *noesis* (i.e. immediate, non-discursive understanding), mathematical objects are grasped in *dianoia* (i.e., propositional understanding). The realm of becoming now consists of things and their copies (shadows, reflections, images, etc.). One part of Plato's critique of art relies on the low ontological standing of images (Politeia 595). Variations of this hierarchical ontology can be found throughout the dialogues.

At the very top of this hierarchy we find a set of highest Ideas (*megista gene*), forming a complex nexus of mutual participation (Sophist 254, Theaetetus 185). This set includes, for example, the Forms of Being, Identity, and Difference. All of these Forms partake in each other as well as themselves: All three Forms exist, that is, they partake in Being; all three Forms are also different from each other, yet identical with themselves, so they all partake in the Ideas of Identity and Difference. Everything partaking in one of them necessarily partakes in all others. And so the most abstract and fundamental Ideas intertwine, guaranteeing determinateness to each other and everything else.

The *Timaeus* finally offers a long monologue on cosmology, starting with the construction and ordering of the world by the godly *demiurge*, proceeding to the nature of materiality (*chora*) and the structure of matter, where Plato introduces the so-called Platonic Solids as the mathematical basis for the classical four elements.

Epistemology

Seeing that philosophy as dialectical enterprise relies on speech as the means for arriving at knowledge (*episteme*), Plato outlines at least a rudiment of logic as part of his wider epistemology (Sophist 263). A 'logos' is always about something; it is a concatenation of nouns and verbs disclosing something; it is true if it discloses what is the case and false if it does not. Plato also formulates an ontological version of the law of non-contradiction (Politeia 436; Theaetetus 165; Sophist 259).

Knowledge must always be logically present-able and defend-able (Politeia 534; Politicus 287); what you know you can put into words (Laches 190). A good criterion of whether someone knows anything is thus his ability to teach (Alcibiades 118). Another criterion is generality, as knowledge about anything implies a general understanding of some kind: If Io, the rhapsode, really knew that Homer is a good poet, he would be able to generally judge the quality of any poet in question (Io 531). The fact that Io lacks this ability disqualifies him from having knowledge. Generally, most people lack true knowledge, which is why one should never epistemically rely on the many, but only on designated experts (Alcibiades I 111). The most important difference between knowledge and mere belief or opinion, however, lies in the 'bonds' placed upon knowledge. While both knowledge and correct belief (*orthe doxa*) share the property of being true (Gorgias 454), knowledge is 'bound' by the chain (*desmos*) of reason (*logos*) (Meno 97). Despite this, Plato seems to oppose the definition of knowledge as justified ('bound') true belief, insofar as Theaetetus proposes exactly this definition in the eponymous dialogue but finds himself in an unresolved *aporia* at the end.

Knowledge is – like speech – always about something (Charmides 168). Plato now assumes a strict correspondence between the order of knowledge and the order of being. According to this onto–epistemological correlation only Being can possibly be an object of knowledge. Thus, technically, only mathematics and philosophy could yield true knowledge at all as they alone concern themselves with pure Forms and mathematical objects. Yet, only philosophy fulfills the strictest, final requirement of knowledge: presuppositionlessness. For Plato, everything below absolute knowledge is mere belief or a kind of technique (Politeia 511, 533). True knowledge is only obtained by *noesis*, acts of immediate, non-propositional, and non-discursive understanding.

Acquiring knowledge is described as a form of remembrance (*anamnesis*); as the myth goes, our eternal souls knew everything before being entombed in the body and all earthly knowledge is just reacquired (Meno 86; Phaedo 66; Phaedrus 247). It is still a matter of dispute as to how this myth is to be interpreted, whether Plato takes it to be literally true or whether he presents a theory of a priori knowledge in mythical form (or both).

Ethics and Politics

Taking his cue from the real Socrates, Plato's early dialogues largely deal with ethical issues. Socrates is pictured in conversation with friends and acquaintances, discussing matters of virtue, the good life, and education. He talks about courage with the old general Laches, prudence with the young and power-hungry Alcibiades, piety with the priest Euthyphron (who is on his way to sue his own father), friendship with the young Lysis and his friends, etc. Important questions include the nature of individual virtues, the problem of whether there are several distinct virtues or only one, and above all, the question of whether virtue can be taught, which he discusses in depth with the sophist Protagoras and the diplomat Meno in the dialogues of the same name. Experience seems to weigh in favor of the unteachability of virtue, as virtuous leaders and politicians are obviously unable to impart their virtue to their offspring (Protagoras 319), yet Socrates stays cautiously optimistic throughout the early dialogues and indeed develops a whole theory of education over large parts of the *Politeia*.

According to Plato, the human soul is made up of three parts (Politeia 440): (1) the desires, called 'the inner beast,' (2) the faculty of courage and decision, and (3) intellect or 'the inner man' (Politeia 588). A variation of this can be found in the *Phaedrus* where Socrates compares the soul to a chariot, lead by a charioteer, drawn by two horses, one good and obedient, and the other bad and wayward. To better understand the relationship between these parts, Socrates compares them to the castes of merchants and artisans, and warriors and leaders in a utopian city state (Politeia 368). Each part of the soul – and accordingly, each caste – has its own virtue (Politeia 442). The desires should be tempered and moderated (*sophosyne*), the faculty of courage should be brave (*andreia*), and the intellect should be wise (*sophia*); accordingly the merchants should not be greedy, the warriors should be brave, and the leaders should be wise. The overall property of a soul/state organized in this way is that of justice (*dikaiosyne*) insofar as every part does what it should

(Politeia 433). A virtuous (i.e., moderate, brave, wise, and just) soul/state is a unified, ordered soul/state: virtue is order (Gorgias 504). The task of politics is to mirror the inner order of the soul in the outer order of the city-state. A true politician makes his people better (Gorgias 517), and he threads and weaves together all lives and public activities into one ordered whole (Politicus 279). The best politician is the one not wanting the job because he does not seek power and finds his rewards elsewhere (Politeia 520). This leads to the theorem of the philosopher king: philosophers should become leaders or all leaders should start to philosophize, else the political future looks grim (Politeia 473). In the utopian *Politeia*, the form of the state is one of a benevolent dictatorship, where the post of king is held by highly educated men and women(!), selected purely on grounds of capability and merit.

Plato distinguishes three basic forms of governance: monarchy, oligarchy, and democracy (Politicus 302). All three forms come in a good and a bad version. Monarchy with a benevolent king is the best political order, tyranny the worst; a proper aristocracy is the second best, oligarchy the second worst; and democracy in both forms thus forms the middle ground. Book 8 of the *Politeia* offers a theory of (d)evolution of forms of governance corresponding to the (d)evolution of the individual soul. Regardless of the specific form of governance, the quality of political life is ultimately determined by the leaders having true knowledge instead of mere opinion (Politicus 293).

The ordering of the soul is the highest ethical goal; doing a crime is therefore much worse than becoming a victim, as criminal behavior upsets the inner order, while falling victim to a crime only concerns property or health, not the soul (Gorgias 469, 474). This is especially important seeing that Plato assumes the soul to be (partly) immortal (Phaedo 100) and its state to be the main factor determining heavenly punishment as well as future life (Politeia 612). Art can also perturb the order of the soul; next to the low ontological status this dangerous potency is the second reason for Plato's criticism of art.

Without virtues there is no good life, no happiness, nor are there any goods in the strict sense of the word as all things only gain their value through usage; property does not make anyone happy unless wisely used (Euthydemos 280), so wisdom alone is the source of values. This emphasis on intellect is essential to Platonic ethics. Virtue is not just order, but a form of understanding: it is impossible to act virtuously by chance, the ordered soul is a consciously ordered soul.

The Unwritten Doctrine

Although Plato himself deemed his theory of principles too important to be written down, many report its content, including such prominent figures as Aristotle, Sextus Empiricus, and the Neoplatonists. These written accounts of the unwritten doctrine allows us an 'esoteric' reading of the dialogues, giving meaning to many of the hints, winks, and deliberate gaps within Plato's writings, grounding the dialogical musings in a more systematic framework. And while the theory of principles is esoteric in that it was not shared with the public, it is quite easily put into words and hardly ever forgotten (Epistle 7, 344). In its core, it just states that the Good is the One, with the Many (also called the Indefinite Dyad) as its polar opposite.

The theory of principles discloses the ultimate ground and capstone of Plato's philosophy. Accordingly only a theoretical definition will grant an understanding of the unwritten doctrine; that is, only a complete grasp of Plato's theoretical edifice allows to fully conceive of the function of the principles. Yet, it is possible to offer a few clues as to the nature of the One and the Dyad. Plato seems to endorse a form of polar monism. The One is the highest principle, the pole toward which everything is orientated; it grants unity and therefore order – in the realm of Being as well as Becoming. The degree of unity determines the ontological rank of any entity, that is, unity orders the ontological hierarchy. The polar opposite of the One is the Many or the Indefinite Dyad. It is the principle of materiality, acting as the manifold upon which the One bestows order. Their interaction enables Being and Becoming. These very abstract structures recur on every level of Plato's philosophy: One virtue (justice) encompasses many virtues (moderation, courage, and wisdom); one Form determines many objects; one king rules many subjects, just as one part of the soul rules the others; and one man (Socrates) stands against the many (*hoi polloi*).

The principles are not entities, but above and below Being or Becoming, respectively. Within the dialogues, the principles take on different guises, most notably that of the Good, which grants Being and conceivability to everything and is 'beyond being' (Politeia 509). The One itself becomes thematic in the second half of the *Parmenides*, although it is not openly classified as a principle. Here the old Parmenides explicates dialectically why the One is no entity: taken in its absoluteness the One is literally nothing; it has no properties, no features, no Being, not even a name (Parmenides 141) – for if it were to have all these it would not remain purely and absolutely one. The polar opposite of the One likewise has no Being. As Indefinite Dyad the Many is an absolute multiplicity, yet instead of transcending being, it lacks being because it lacks unity. It is too diverse to be anything but nothing. In the guise of materiality (*chora*) it can only be grasped by "a kind of bastard-thought" (Timaeus 52).

See also: Dictatorship in History; Ethics and Values; Idealization, Abstraction, and Ideal Types; Justice: Philosophical Aspects; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Wisdom: Philosophical Aspects.

Bibliography

Primary Literature

Burnet, J. (Ed.), 1993. Plato, Opera. Oxford University Press, Oxford, New York (Greek Text).

Cooper, J.M. (Ed.), 1997. Plato: Complete Works. Hackett, Indianapolis (Translation).

Secondary Literature

Cherniss, H., 1945. The Riddle of the Early Academy. University of California Press, Berkley.

Erler, M., 2007. Platon. Schwabe, Basel.
Fine, G. (Ed.), 2008. The Oxford Handbook of Plato. Oxford University Press, Oxford.
Guthrie, W.K.C., 1986. A History of Greek Philosophy, vols. IV–V. Cambridge University Press, Cambridge.
Halfwassen, Jens, 2006. Der Aufstieg zum Einen. Walter de Gruyter, München, Leipzig.
Krämer, H.J., 1990. Plato and the Foundations of Metaphysics. State University of New York Press, New York.
Kraut, R. (Ed.), 1992. The Cambridge Companion to Plato. Cambridge University Press, Cambridge.
Kraut, R., Plato, Fall 2013. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/archives/fall2013/entries/plato/.
Nikulin, D., 2012. The Other Plato. State University of New York Press, New York.
Reale, G., 1997. Toward a New Interpretation of Plato. Catholic University of America Press, Washington.
Szlezak, T.A., 1999. Reading Plato. Routledge, London, New York.
Vlastos, Gregory, 1995. Studies in Greek philosophy. In: Graham, Daniel W. (Ed.), Socrates, Plato, and Their Tradition, vol. 2. Princeton University Press, Princeton.

Play, Anthropology of

Don Handelman, The Hebrew University, Jerusalem, Israel

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11503–11507, © 2001, Elsevier Ltd.

Abstract

Play is a malleable, changeable phenomenon, in terms of cognition, practice, and communication. Play places people into realities that are multiplicities of possibility. This article describes issues of the framing of play as separate from mundane reality, and contrasts play to the playful attitude that erupts suddenly into the mundane. Play is understood as a way of organizing activity, rather than any particular content of activity. Play is implicated in the making of change, and through this to the presence of violence. The aspect of violence is suppressed when play is rationally and neatly framed off from mundane reality, as make-believe and fictive. Studies of cosmology indicate that when play is implicated in the creation of cosmos, its effects are felt throughout the functioning of the cosmos that comes into existence. Otherwise, cosmos is perceived as peripheral and subversive. Finally, game is considered as a variant of play, a metaphor that has enjoyed especial popularity during the twentieth century.

Malleable, flexible, fluid, uncertain, changeable – all are adjectives that are used to describe play, whether as a medium of cognition, practice, or communication. Play refers to a complex of universal phenomena among human beings that also are found among numerous animal species (Fagen 1981). All peoples, though they may not distinguish in semantic terms between play and other behavior, likely engage in playing as they shift into a conditional or subjunctive sense of being, one that opens toward a multiplicity of possibility. The mathematician and the theoretical physicist play during the process of shaping their ideas no less than children imagine, practice, and repeat their inventions in their play. Useful overviews of play phenomena are given in Spariosu (1989), ,Schwartz-man (1978) and Sutton-Smith (1997).

Despite its pervasive presence, play is perhaps the least concrete of phenomena and therefore one of the most difficult to define. Any useful definition of play highlights its multiplicity of potential experiences. The anthropologist, Bateson (1972) perceives play as paradoxical, in that it is and is not what it seems to be. The classicist Spariosu (1989) terms play, amphibolous, or going in two directions at once. In a succinct manner, the anthropologist, Droogers (1996) defines play as, 'the capacity to deal simultaneously and subjunctively with two or more ways of classifying reality.'

Play and Reality

There is an intimate relationship between play and reality because play is first and foremost experiential, whether personal or social, and reality does not exist without its experiencing. In the creation and practice of realities of play, the imagination is crucial. Without the cognitive, the capacity to imagine realities, there is no play, just as, without the practice of experience there is no reality.

Play raises to the fore issues of alternative realities, the role of these and other multiplicities in social and personal life, and their consequences for human cognition and practice. The cognitive perceptions of play enable persons to enter into alternative realities, and perhaps into the simultaneity of multiplicity. Cognition is related intrinsically to how realities of play are imagined and framed differently from other realities. The practice of play entails the sensuous embodiment of these alternative realities; and, so, the subjective experiencing of play from within these imagined contexts that offer a sense of release, freedom, exploration, and perhaps exhiliration. The communicative component suggests that these specialized contexts of play are able to comment on the social orders from which they arise, and to send these messages back as the participants in play return to the reality of nonplay. Therefore contexts of play are metacommunicative in their relationships to social order: play communicates to its participants not only *how* to play but how such playing relates to social order. That play is framed as imaginative and malleable, and that through these qualities it generates and conveys metamessages informing players as to how to relate to aspects of social order, makes play perhaps the most fruitful of contexts of socialization, both for children (Briggs 1998) and for adults.

Framing Play

Given that play constitutes domains that cognize, practice, and communicate the shaping of imagination, are these domains to be conceptualized as incisively distinct from domains of serious reality, of not-play? This is the implication of Bateson's highly influential theory of the framing of play which separates play lineally from the serious, or not-play. According to Bateson, the cognitive metamessage, 'This is play,' brings into existence the play frame which cleaves between the seriousness of everyday reality and the imaginings which are practiced within the frame. One may adduce from this not only that play comes into existence as a corollary of its framing but also that play, and the practice of its imaginative multiplicities of possibility, is located within the restrictions of its framing. Viewed from outside the frame, play is make-believe activity that in fact does not exist. Experienced from within the frame, play may well be utterly engrossing in the embracing reality of its presence. These views accord well with Johann Huizinga's (1970) classic study of how civilization emerged through the

plastic premises of play; although in the process, play itself became a temporary sphere of make-believe activity, secluded, limited, and bounded in time and space. Once play is framed through lineal cognition, it becomes a world unto itself, one of temporary limited perfection with its own internal teleologies (Gadamer 1988). Within itself the practice of play may well proceed in deadly earnest.

Play and the Playful

Sutton-Smith (1997) offers a useful modification to the binarism of a reality either of play or of not-play, of make-believe or of serious reality. Sutton-Smith distinguishes between play and the playful. To build on his argument that the playful plays with the very framing of play itself, one can add that the playful is a mood, an attitude, that may permeate both serious reality and play, or that suddenly may surge into presence within the mundane. In either instance, the playful is full of the impulse to perceive and feel in ways other than those offered by the immediacy of a given reality. The abrupt presence of playfulness disrupts the most routine of expectations. Playfulness injects speedy uncertainty into the most expected of social practices.

The playful attitude within mundane reality may be integral to the existence and organization of the latter; while play itself may be more 'distant,' as it were, from mundane reality. Playfulness and play complement one another. The eruption of playful feeling and perception are a significant key to establishing the play frame in relationship to mundane reality. But the play frame itself directs the kinds of ludic or make-believe activities that take shape within it, relatively separated from mundane reality.

Play and Foundation-for-Form

In order for a play frame to come into existence there seems to be a need for a form of reference, a foundation-for-form or organic model that can be altered in systematic ways and turned into someone or something else. Nonetheless, whatever is changed retains crucial similarities to the form of its foundation and so remains intimately related to it. For example, The medieval European Feast of Fools which mocked the traditional Christian Mass needed the latter so that this could be inverted and altered. The play-mass, performed in church by priests, would have had no significance for participants were it not derived from and contrasted with its everyday analogue, the sacred Mass.

To take a more mundane example, when one little American girl, aged seven, turns to her sister, aged five, and says: 'Let's play sisters,' she is establishing a play frame within which these girls can play the foundational form of their sisterhood, imaginatively and perhaps with novelty. Play changes the known signs of form into something else by altering the reified boundaries that define and characterize the phenomenon. These girls will still be sisters, but they will be able to explore, practice, and experience their sisterhood with a difference, with license, in ways perhaps closed or unimagined within their usual relationship. They can play with the strangeness of being sisters. In turn, they may carry the messages of their experience back into their ongoing relationship as sisters, changing this in minute but additive ways. Hypothetically, they playfully may enter into play-sisterhood whenever they desire.

Attributes of Play

Play is a way of organizing activity, rather than any particular set or content of activities (Miller 1973). Any activity may be altered playfully, and played at. But in order to do this the everyday, instrumental perception of how means lead to and produce ends must be loosened, stretched, or perhaps uncoupled entirely. Goals exist in play, but the means to attain them are given degrees of autonomy. This is why play behavior can be endlessly repetitive or follow a crooked line around artificial obstacles. The patterning for play may be suggested by how the foundations-for-form that are used in playing function in the everyday. Say that the two little sisters and their friends place chairs in a row, the back of each facing the seat of the next, and that these children imagine these chairs to be a train. The children practice playing train, until they decide that the train is a plane, and the row sprouts wings, carrying its passengers, but then is turned into a bird flying off with children on its back, becoming a row of chairs in school and kindergarten that is also a plane. In each instance, one foundation for form (train, plane, bird) is imagined to take the place of another. Each such organic model is voluntarily turned into a foundation for imaginative practice that plays in turn with the multiplicity of these organic models, one after the other or as hybrids of one another (a row of chairs that with near simultaneity is train, plane, bird).

The perception and practice of play brings into being something that had not existed before in quite the same way. This is done by changing the shape and positioning of boundaries that categorize phenomena, thereby altering their meaning. Each shift by sisters into playing sisters seems like a repetition, yet is instead a different beginning since it imagines anew and plays with the very possibilities of sisterhood. Even a slight change in the boundaries that give shape and significance to phenomena involves some degree of destruction and recreation. In its cognition and feeling, play is associated closely with creativity, creation, and destruction.

Play is frequently implicated in the making of change. Change is the making of difference, even if this difference seems to mimic or eventually to reproduce an earlier state. Play introduces malleability into social and cultural life, taking apart the given and perhaps offering alternative visions and practices. Play often is involved in learning something new, through open-ended repetition, variation, exaggeration, and experimentation. Play is not unusual in the liminal phases of initiation rituals (Turner 1969), healing rituals (Kapferer 1983), and mortuary rituals of tribal peoples, when initiates, the ill, or the dead undergo profound transformations of being. Playfulness, in a more postmodern idiom is central to the practice of certain neo-shamanic rituals whose goal is the transformation of consciousness (Lindquist 1997). These and many other changing conditions involve shifts of cognition, practice, and communication. More significantly, change (creative or reproductive) usually involves fragmentation, destruction, or transformation of some sort, whether temporary or permanent, and these processes are often violent. The violence of change

and change through play are often closely related. Some form of destruction and violence (to cognition, feeling, body and being) is involved in the shattering suddenness of the playful and in play.

The Truth-Values of Play

The relationship between play and violence relates to how play has been conceived, and to the truth-values accorded to play. In his survey of philosophical ideas of play, Spariosu (1989) finds two trends, one that stresses the practice of play as prerational (with its roots in ancient Greek thought and poetry), and the other that emphasizes play as rational activity. In pre-rational thinking, play is full of violent agon and power, strongly and pleasurably emotional, arbitrary in its appearance, experienced perhaps as ecstatic, and opening toward unlimited freedom. In rational thinking, the naked, enjoyable forces of chaos and disorder, of irrational power and violence, are suppressed in favor of play as rationally framed activity that is set apart from serious reality. Here play is irreal in its irrationality, and therefore controlled, nonviolent, and limited in its freedom. To expand on Spariosu, one may add that the eruption of pre-rational play (akin to playfulness) into everyday life is full of its own unsuppressed truth-value. The truth-value of pre-rational play is integral to living within realities that overlap and that are continuous with one another – realities that relate to one another as multiplicities of possibility. By contrast, the make-believe status of rational play, framed and contained in special times and places (as Huizinga, Bateson, and others argue) denies any direct or immediate truth-value to play, enabling its practice to be treated as fictive, as 'just play,' 'mere play,' 'done only in play,' and so forth.

Play and Cosmology

The organization of cosmology demonstrates from another perspective the importance of the truth-values attributed to play. For example, in Indian cosmogen-esis, play (or more accurately, playfulness) is implicated in the very creation of cosmos. Thus, varieties of Hindu cosmology and religion embed premises of play – malleability, multiplicity, ongoing change – at a very high level of cosmic organization. In consequence, these self-same premises of play resonate throughout the levels of cosmos. The entirety of cosmos is influenced by premises of play. In this kind of cosmology, play is a top-down idea, responsible in no small degree for the multiplicities of reality that constitute Hindu cosmos and religions (Handelman 1992, Handelman and Shulman 1997).

Here play has high degrees of truth-value and is rarely referred to as make-believe. By contrast, mainstream monotheistic cosmologies, if they pay any attention at all to ideas of play, embed these at a low level of cosmic organization, distant from cosmogenesis and the ordering of religion and society. In monotheistic cosmologies, play is more of a bottom-up idea, lineally framed, rationally contained, and separated from serious reality. Bottom-up play enters cosmology and religion from their conceptual and emotional undergrowth, and therefore is perceived as threatening. Here the messages of play are often positioned in opposition to or as a negation of the order of things; and play is perceived as fantasy, illusion, and ephemeral; yet also as subversive and as resisting the order of things. But here the practice of play has low truth-value. Thus, the truth-value of top-down play is openly full of power and violence, while these qualities are veiled, hidden, and denatured in the low truth-values accorded to bottom-up play.

Game and Play

The idea and practice of game is closely related to that of play and may be understood as a variation of the latter, in the following senses. Game, like play, is a way of organizing activity. Game, too, loosens the relationship between means and ends. But the existence of game depends on relatively enduring rules, rather than on foundations-for-form that are easily changeable. Therefore games constitute distinct, autonomous realities, virtualities, given their internal, rule-based organization which to a high degree determines how they are practiced. Playfulness has less of a role in the practice of games. Games are less malleable and harder to alter than play.

Yet it is precisely because of their ruledness that games can be organized to make directed change through their practice. Levi-Strauss (1966) comments that game is a systematic way of creating inequality. Game begins with equality between its sides and produces differences which are perceived as significant. The logics of game, in fact, are often close to those of ritual which is intended to generate change through its operations. Traditional Mesoamerican ball-games had game-like aspects, yet they were ritual-like in their relationship between means and ends, perhaps ends of cosmic renewal that their practice was intended to accomplish. More generally, the presence of play and game within rites seem to contribute to the attainment of ritual goals. (This is powerfully exemplified in the roles of clowns in Native-American Pueblo rituals.)

Perhaps in more systemic ways than play, game may awaken interpretive, reflexive feelings about aspects of social and moral order. The highly agonistic Native-American Kwakiutl potlatch, a ritual full of conflict that both demonstrated and reordered social status, produced the 'play potlatch' (Codere 1956) in which participants laughed at their own strivings for power and hierarchy. The rigid gender divisions in traditional Iranian society produced games practiced by women in which these distinctions were called into question, mocked, and played with (Safa-Isfahani 1980).

The twentieth century has seen tremendous growth in the use of game as a metaphor for living and social order (Minnema 1998) – Wittgenstein's language games, Von Neumann's game theory, war games, educational games, the game of intelligence testing, games people play, the play of quantum mechanics, chaos theory, virtual realities, and so forth. In postmodernist terms, these visions are of life organized as interpenetrating games, none of which give any total or complete vision of reality, and no certainty about the stability of forms of living. The game metaphor reflects more fluid and mobile social orders that are perceived, correctly or incorrectly, as the creations of individual agency. Television as spectacle has contributed greatly to this growth by telecasting existing competitive game forms and aiding in the invention

of new ones. Sports are at the front of these developments. These multitudes of contest (in which violence is close to the social surface if not on it) continuously rank and re-rank participants over and again in hierarchies that are more kaleidoscopic, fragile, and ephemeral than they are enduring. Play as the multiplicities of possibility, and the playful as the eruption of these possibilities in the everyday, are competing with more unitary monotheistic ethoses, at least in the visions of the middle-class Western world.

Current thinking on play is reconceptualizing the problem of framing – the problem of relationships among realities, social and cognitive – to open alternatives to the ongoing rupture between play as rational and as irrational. Research on Internet chat groups plays into questions of framing (Danet et al. 1998). Issues of power in and through the practice of play are also related to questions of framing. Thus, ideas of 'dark play' (Schechner 1988), when players do not know that they are playing, and 'evil play,' when the manipulator of the frame positions herself outside it and plays others, yet pretends to be within the frame and controlled by it, relate to issues of play and power. They echo Bateson's still enduring thought that a more complex question to ask of a play phenomenon is whether it is indeed this, or whether it is pretending to be so. Imagining the pretense of play is even more powerful than play imagined as such.

See also: Game Theory; Sport, Anthropology of; Sport, Sociology of.

Bibliography

Bateson, G., 1972. Steps to an Ecology of Mind. Chandler, San Francisco.
Briggs, J.L., 1998. Inuit Morality Play: The Emotional Education of a Three-Year-Old. Yale University Press, New Haven CT.
Codere, H., 1956. The amiable side of Kwakiutl life: The potlatch and the play potlatch. American Anthropologist 58, 334–351.
Danet, B., Ruedenberg, L., Rosenbaum-Tamari, Y., 1998. Hmmm… where's that smoke coming from? Writing, play and performance on internet relay chat. In: Sudweeks, F., McLaughlin, M., Rafaeli, S. (Eds.), Network and Netplay: Virtual Groups on the Internet. AAAI Press, Menlo Park, CA and MIT Press, Cambridge, MA.
Droogers, A., 1996. Methodological ludism: beyond religionism and reductionism. In: van Harskamp, A. (Ed.), Conflicts in Social Science. Routledge, London.
Fagen, R., 1981. Animal Play Behavior. Oxford University Press, New York.
Gadmer, H-G., 1998. Truth and Method. Crossroad, New York.
Handelman, D., 1992. Passages to play: paradox and process. Play & Culture 5, 1–19.
Handelman, D., Shulman, D., 1997. God Inside Out: Siva's Game of Dice. Oxford University Press, New York.
Huizinga, J., 1970. Homo Ludens: A Study of the Play Element in Culture. Maurice Temple Smith, London.
Kapferer, B., 1983. A Celebration of Demons: Exorcism and the Aesthetics of Healing in Sri Lanka. Indiana University Press, Bloomington, IN.
Lévi-Strauss, C., 1966. The Savage Mind. Weidenfeld & Nicholson, London.
Lindquist, G., 1997. Shamanic Performances on the Urban Scene: Neo-Shamanism in Contemporary Sweden. (Stockholm Studies in Social Anthropology 39). Almqvist & Wiksell International, Stockholm.
Miller, S., 1973. Ends, means, and galumphing: some leitmotifs of play. American Anthropologist 75, 87–98.
Minnema, L., 1998. Play and (post)modern culture: an essay on changes in the scientific interest in the phenomenon of play. Cultural Dynamics 10, 21–47.
Safa-Isfahani, K., 1980. Female-centred world views in Iranian culture: symbolic representations of sexuality in dramatic games. Signs 6, 33–53.
Schechner, R., 1988. Playing. Play & Culture 1, 3–19.
Schwartzman, H.B., 1978. Transformations: The Anthropology of Children's Play. Plenum, New York.
Spariosu, M.I., 1989. Dionysus Reborn: Play and the Aesthetic Dimension in Modern Philosophical and Scientific Discourse. Cornell University Press, Ithaca, NY.
Sutton-Smith, B., 1997. The Ambiguity of Play. Harvard University Press, Cambridge, MA.
Turner, V.W., 1969. The Ritual Process. Aldine, Chicago.

Play During Infancy and Early Childhood: Cultural Similarities and Variations

Hui-Chin Hsu, University of Georgia, Athens, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Play is a common, yet complex, activity in infancy and early childhood across cultures. Cross-culturally, early play shows both similarities and differences. This article documents the coexistence of cultural-general patterns alongside cultural variations in the development (i.e., emergence and changes), varieties (e.g., object play and pretend play), structures (e.g., frequency and play partners), and functions (e.g., training and learning) of early play. Factors such as individual characteristics of the child and the caregiver, physical ecology, interpersonal context, cultural beliefs, and family immigration status that contribute to cultural variations are also reviewed.

Play is a common activity in infancy and early childhood across cultures. By nature, early play is complex, integrative, and characterized by developmental changes. Play during the relatively immature period of human development is believed to afford a low-cost and low-risk way for young children to learn about their environment for developing innovative and adaptive strategies (see Pellegrini, 2011). It is expected that play has both immediate and deferred benefits to children in multiple domains, including biological (e.g., signifying health and stimulating brain development), physical (e.g., exercising sensorimotor skills), cognitive (e.g., learning and creating knowledge), communication (e.g., making use of verbal and nonverbal skills), social (e.g., connecting and relating to others), emotional (e.g., expressing and regulating feelings), and motivational (e.g., initiating exploration) development. Early play also serves as an important socialization function to help young children learn cultural values to prepare for their adult roles.

Early play has been assumed to be a universal activity characterized by similar varieties and structures, following a similar developmental course, and promoting similar developmental skills across cultures. Converging evidence from anthropological and psychological research, however, suggests otherwise. A survey study (Singer et al., 2009) with mothers of children ages 1–12 from 16 countries in Africa, Asia, Europe, North America, and South America reported both similarities and differences in children's play. The cultural-general patterns shared between cultural communities included the following: playing outside was children's most enjoyable activity; children who resided in rural areas engaged in physical play more than those who resided in urban areas; boys engaged in more physical play than girls; and girls played more pretend than boys. Patterns indicating cultural diversity were also found; for example, children in the cultural communities with less advanced economic developments had higher rates of outdoor play than those in the cultural communities with more advanced economic developments; about half of the children played with toys; approximately a quarter of children engaged in pretend play; only children in Argentina, Ireland, Portugal, the United States, and the United Kingdom frequently participated in pretend play; and 15% of the mothers believed that pretend play was necessary for healthy development.

Taking a sociocultural approach, recent theorization challenges the universalists' view that early play is independent of culture. The socioculturalists' view suggests that play is not always valued across cultures, and that cultural variations are inherent in early play development. Early play is thought to reflect young children's experiences of growing up in a cultural setting. The regularities within settings, customs, and cultural belief systems organize children's developmental experiences and provide the information for children to construct their play. Through active practices in play, young children learn about their cultures (i.e., enculturation). Shaped by the specific cultural values and beliefs, early play is culturally determined and deeply rooted in daily routines and cultural practices (see Göncü and Gaskins, 2007).

Some developmentalists further argue that early play is not simply an outcome of a unidirectional influence that is passively shaped by the larger cultural beliefs systems. Rather, it is directly and indirectly influenced by a confluence of factors in a child's ecological niche. An ecological system view suggests that early play is a joint product of a multitude of distal and proximal influences, including biological and cognitive characteristics of the child, individual and behavioral characteristics of the parent, sociodemographics of the family, interpersonal experience, and the cultural context (see Bornstein et al., 2012; Bornstein's review in Göncü and Gaskins, 2007; Kenney, 2012; Haight et al., 1999). As a result of simultaneous interplay among these internal and external contributors, variations between individuals and across cultures are expected.

In this article, cultural-general (i.e., similarities) and cultural-specific (i.e., variations) patterns in the development, varieties, structures, and functions of play in infancy (the first year of life), toddlerhood (ages 1–2), and preschool age (ages 3–5) will be reviewed. Drawing on psychological and anthropological evidence, this article will discuss the factors contributing to cultural variations in early play by children and their play partners (e.g., mothers, fathers, siblings, and peers), including physical ecology and economic developments that determine the prevalence and varieties of play, interpersonal context that shapes cultural experience to be represented in play structures and themes, and cultural beliefs that influence the opportunities for and functions of play.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23171-1

Play during Infancy

Caregivers across cultures engage in playful interactions and games with their infants during the first year of life. Both cultural-general and cultural-specific patterns are observed in person- and object-oriented play between caregivers and infants. Cultural-general patterns in the development of game playing, person-oriented play, and object-oriented play will be discussed first, which are followed by the reviews on cultural variations, including parental investment in play, stylistic preferences in play, play structures, and play partners.

Cultural-General Patterns

Similarities in Game Playing

When playing games with their infants, caregivers demonstrate cultural-general behavioral patterns. The majority of mother–infant games (e.g., peekaboo) are invariant and repetitive in structure, involving rhythmical motor actions coordinated with vocal and facial expressions to highlight the sequenced movements. Game vocalizations by mothers across cultures share the same distinctive features of elongation and fluctuation accompanied by exaggerated positive facial expressions and gestures. Perhaps such universal design of providing redundant information is an adaptation to the infant's developing attentional and cognitive abilities (see Fernald's review in MacDonald, 1993). Despite repetitions and regularities, caregivers also improvise with varying speed, melody, words, and gestures to make games highly variable and unpredictable. Further, caregivers in different cultures are sensitive to their infants' developing ability. In response to infants' developing motor skills, both European-American and Puerto Rican mothers not only increase game playing, but also offer age-appropriate physical support during play (Miller and Harwood, 2002). With age, the role of caregivers in game playing gradually changes from acting as a coach providing active scaffolding to infant, to acting as a playmate engaging in reciprocal play with infant, and to acting as a spectator allowing for greater infant participation.

Similarities in Person- and Object-Oriented Play

Mothers in different cultures demonstrate comparable contingent behavioral response during playful social interactions. During person-oriented play (e.g., kissing and face-to-face talk), despite unique cultural variations in specific behavioral channels (e.g., Kärtner et al., 2008; Keller et al., 2010), American, Chinese, Indian, Japanese, German, and Cameroon Nso mothers all contingently respond to their 3- to 5-month olds (see Bornstein's review in Göncü and Gaskins, 2007; Kärtner et al., 2008). By 3–4 months, infants across cultures begin to show an interest in objects at about the same time. To elicit properties and extract information about the objects, infants' play with objects similarly progresses from simple and spontaneous sensorimotor exploration (e.g., mouthing) to functional manipulation (e.g., rolling a ball). Mothers and their infants in different cultures are also similarly coordinated and attuned to one another during object-oriented play. Five-month-old American, French, and Japanese infants, for example, share the same preferences of looking more at the environment than at their mother and producing more vocalization than fusses/cries. Their mothers similarly demonstrate cultural-general patterns of responding to their 5-month olds' exploration of the environment with encouragement, to infants' nondistress vocalizations with imitation, and to infants' distress vocalizations with nurturance. Mothers are equally sensitive to infants' nondistress and distress vocalizations and make more frequent vocal response to infants' vocalizing than their looking (see Bornstein's review in Göncü and Gaskins, 2007).

Cultural Variations

Differences in Parental Investment in Play

Caregivers are motivated to invest in their infants' protection and provision in order to maximize their survival. Play requires time and energy and involves a commitment of resources. Play with infants is thus expected to be more common among cultures with high-investment reproductive strategies (see MacDonald, 1993). For example, games and person-oriented play between caregivers and infants are more common in foraging societies than other subsistence cultures because high-quality care is needed to ensure infant survival (Lancy, 2007). With greater exposures to hazards and predators, Aka hunter-gatherers in central Africa were more concerned about the survivals of their infants than their neighboring Ngandu farmers, and thus engaged in more promixal interaction (e.g., physical contact) and less distal communication (e.g., looking) with their infants (Hewlett et al., 1998; Lancy, 2007). By 9 months of age, compared to Aka infants, Ngandu infants engaged in significantly more object play and solitary play without caregivers' involvement. Cultural variations in caregiver–infant play can also be attributed to caregivers' adaptation to their ecological environments. Mothers of Kpelle in West Africa, Gusii in East Africa, Micronesian island of Fais, and American Navajo Indians rarely engage in playful face-to-face interaction with their infants (see Field et al., 1981). Their low-investment parental style may be due to the high infant mortality rate, as parental indifference may be a coping strategy to reduce the sense of loss (Lancy, 2007; see Rooparine's review in Pellegrini, 2011).

Differences in Appropriateness of Play

Play between infants and their caregivers can be regarded as an expression of broader cultural expectations and rules. Cultures differ considerably in their views about the value of play with infants. Parents in cultures where play is believed to be valuable for promoting cognitive or social skills (e.g., the United States and the United Kingdom) often engage in play with their infants. In contrast, parents in cultures (e.g., India, Guatemala, Nigeria, and Zambia) where play is for enjoyment, not promotion of development, rarely engage in play with their infants (see Roopnarine et al., 1994). In India, mothers do not see themselves as the playmate of their infant; yet relatives may assume the role of playmate for infants. In Italy, mothers regard play as inevitable, requiring no adult intervention (see New's review in Roopnarine et al., 1994). Appropriate play behaviors with infants vary widely according to cultural scripts. Kaluli mothers in Papua New Guinea describe their infants as helpless and they rarely utter to their infants, make eye contact, or elaborate utterances initiated by the infant.

Because verbalization is viewed as the primary indicator of competence, nonverbal vocalization is not interpreted as meaningful in Kaluli. Kaluli infants are not treated as a play or communicative partner, but rather as an addressee. Reflecting the socialization emphasis on social ties and cooperation, Kaluli mothers often face their infants outward to be seen by the larger social group and model how to appropriately use utterances for the infant to a third person (see Lancy's review in Göncü and Gaskins, 2007). Caregivers also follow culturally appropriate rules of expressing affect during play. Gusii mothers in southwestern Kenya avoid bringing out a high intensity of affect in infants during play, which reflects their cultural beliefs that unrestrained expression of negative affect produces apprehension and that masking feelings can maintain social order and minimize group diverseness (see Field et al., 1981). Because it is culturally inappropriate to display affection openly, East Indian mothers display low emotional expressivity during play with their infants (see Roopnarine's review in Pellegrini, 2011). On the other hand, the high values of interpersonal dependency in Indian culture may be expressed in the greater overlaps and shorter pauses between mothers' talks with their infants' vocalization (Gratier, 2003). Infants in Marquesas Islands in French Polynesia rarely engage in face-to-face play with their mothers; they most frequently play with the child caregivers (e.g., older siblings). The playful interactions of infants with child caregivers are far more complex, reciprocal, and physically close than those with adult caregivers (see Field et al., 1981).

Differences in Social-Oriented Play

Differing in history, beliefs, customs, and practices, cultures can be broadly classified as with an independent or an interdependent orientation. According to Markus and Kitayama's (1991) conceptualization, cultures differ in their interpretations of self, of others, and of the balance between the two. Interdependent-oriented cultures (e.g., Argentina, Japan, rural India, Kaluli in Papua New Guinea, Nso in Cameroon, and Puerto Rico) are characterized by placing high values on relatedness among individuals and sensitivity to others. In contrast, independent cultures (e.g., France, Germany, and the United States) stress individuals' independence from others, attention to the self, and autonomy in personal choice and action. Differences in the broader cultural models inform socialization goals, guide child-rearing practices, and shape the development of parenting behaviors in specific ways (e.g., Kärtner et al., 2008; Keller et al., 2009, 2010). Whereas caregivers in the interdependent cultures commit to socialization goals such as interpersonal harmony, respect of others, obedience, and obligations, caregivers in the independent cultures place a socialization emphasis on individual achievement, self-affirmation and self-reliance, and assertiveness. Parent–infant social-oriented play reflects the broader socialization goals and is organized in culturally preferred ways. To foster infants' sense of individuality and mastery, mothers in independent cultures often engage in a distal style of playful interactions with their infants by making direct eye contact, orienting infants toward objects and events in the environment, and responding to infant vocal and facial signals with verbalization (Keller et al., 2009; see Bornstein's review in Göncü and Gaskins, 2007). In contrast, mothers in interdependent cultures engage in more person- than object-oriented play with their infants (see Bornstein's review in Göncü and Gaskins, 2007). Consistent with their cultural commitments, mother–infant person-oriented play in these cultures is characterized by a proximal style focusing on body contact and body stimulation (Keller et al., 2009). Compared to American mothers, Japanese mothers have a slower tempo of interactive play, are likely to touch their infants, use upper-body looming to punctuate facial expressions and vocalization, and attract their infants' attention to themselves (see Fogel et al.'s review in MacDonald, 1993). Puerto Rican mothers encourage social dependency, respect, and proper demeanor. Their play with infants is featured by leading turn-taking games requiring close physical proximity, issuing directive commands, focusing and signaling the infant attention on the relationship, and using physical restraint to keep infants oriented toward them (Miller and Harwood, 2002).

Differences in Object-Oriented Play

Although infants across cultures show a universal timing of developmental shift in their interest from person-oriented play to object-oriented play, objects are incorporated into play differently according to specific cultural rule. For East Indian parents, exposure to objects or the physical world is considered as inappropriate for young infants; thus, object stimulation during play is extremely rare (see Roopnarine et al., 1994). !Kung caregivers in northwestern Botswana do not give objects to infants or entertain them with objects. !Kung infants' interest in objects is largely ignored by their caregivers. Echoing the cultural emphasis on interpersonal relatedness, they consistently note infants' vocalizations and smiles occurring during object play (Bakeman et al., 1990). Middle-class caregivers in independent cultures engage in more object- than person-oriented play with infants, follow infant's lead, maintain physical distance, praise infants' play efforts, and issue directives in the form of comments and suggestions. During play, middle-class American mothers emphasize cognitively stimulating didactic actions, prefer to verbally direct infant attention to objects, and use toys as the topic of communication. In contrast, Japanese mothers use toys to mediate person-oriented play with infants (see Bornstein's review in Göncü and Gaskins, 2007; Bornstein et al., 2012). Moreover, although Japanese-American and South-American immigrant mothers living in the United States engaged in more object-oriented play than person-oriented play with infants, both groups of mothers reported that they engaged in more person- than object-oriented play with infants, which is consistent with the larger cultural values of mutual dependence. The paradox in the play behavior and perception in Japanese-American and South-American mothers may indicate that behaviors acculturate (i.e., adapt to a new culture) more readily than beliefs among middle-class immigrant mothers (Cote and Bornstein, 2009).

Differences in Physical Play

Fathers in independent cultures (e.g., Canada, Israel, Sweden, and the United States) engage in more vigorous physical play with their infants than mothers, which has led some researchers to conclude that rough and stimulating physical

play contributes to the development of father–infant attachment. However, physical play between fathers and infants appears to be a cultural-specific phenomenon. Both mothers and fathers in interdependent cultures (e.g., India, Malaysia, and Taiwan) rarely engage in rough and physical play with their infants. Parents in many subsistence cultures (e.g., Akas among the pygmies of Central Africa and Baka in Cameroon) practicing group care also do not engage in rough play with their infants (see Roopnarine's review in Pellegrini, 2011). In Italy, familial and nonfamilial women, but not parents, engage in rough play with infants (see New's review in Roopnarine et al., 1994).

Play during Toddlerhood

During the second and third year of life, toddlers' play becomes more complicated and incorporates imagination and pretense. Across cultures, toddlers engage in a variety of play not only with their parents but also with siblings and peers. Both cultural-general and cultural-specific patterns coexist during this developmental period. Cultural similarities are observed in the developmental emergence of symbolic (i.e., pretense) play and peer play, gender differences in caregiver–toddler symbolic play, and effects of adult play scaffolding. Cultures also differ significantly in adults' support to toddlers' play, caregiver–toddler play preferences, thematic topics of caregiver–toddler play, strategies used to scaffold object play, and toddlers' play partners.

Cultural-General Patterns

Similarities in the Emergence of Symbolic Play and Peer Play

Beginning in the second year of life, children' play becomes distanced from the properties of objects and takes on a nonliteral quality. Anthropological and psychological research indicates that the normative onset of symbolic play is similar across cultures at about 12–15 months. Toddlers begin to engage in symbolic play by reenacting activities by self and others (e.g., drinking from an empty cup). Without adult acknowledgment or support, for example, Mayan toddlers engage in brief episodes of symbolic play using household objects (see Göncü and Gaskins, 2007). Ready imitation of one another between peers during play also emerges at around the same time. Toddlers in Seltaman of Papua New Guinea and the United States are similar in generating a great variety of extended bouts of nonverbal coordinated action by playing reciprocal imitation games (Eckerman and Whitehead, 1999).

Similarities in Gender Differences and Mutual Specificity in Caregiver–Toddler Play

During play, caregivers in both independent (e.g., France and the United States) and interdependent (e.g., Argentina, Italy, and Japan) cultures are similar in their promptness, contingency, and appropriate reactions to toddlers (see Bornstein's review in Göncü and Gaskins, 2007; Suizzo, 2004). Regardless of culture, mothers' praise, affection, and elicitation are unrelated to toddlers' symbolic play. Within the dyadic parent–child play context, mothers in both independent and interdependent cultures engage in more object-oriented play with their boys and more symbolic play with their girls. Similarly, boy toddlers across cultures engage in more object-oriented play, whereas girl toddlers engage in more symbolic play with their mothers. Moreover, a cultural-general pattern of mutual specificity is observed, wherein mothers' frequent exploratory and symbolic play is specifically related to more exploratory and symbolic play in their toddlers, respectively (see Bornstein's review in Göncü and Gaskins, 2007).

Similarities in the Effects of Adult Play Scaffolding

Toddlers across cultures exhibit lower quality of play when playing alone than when playing collaboratively with a mature partner (e.g., mother or older sibling). Scaffolding by a mature play partner enhances the duration, diversity, and sophistication of toddlers' play (see Howes' and Lillard's reviews in Pellegrini, 2011). Regardless of culture, caregivers' redirecting of toddlers' attention away from their current interest similarly decreases play complexity, which may be due to the imposed interruption to the flow of activities. In contrast, caregivers' following of child attention increases play complexity, which may be attributable to the increased opportunity for adult modeling and child imitation of more sophisticated play (Sung and Hsu, 2009). Finally, parents who value play for its educational and cognitive benefits and are knowledgeable about play are likely to encourage toddlers' play, engage in play with their toddlers, and adjust the play level to that of their toddlers (see Bornstein's review in Göncü and Gaskins, 2007; Farver's review in MacDonald, 1993; Haight et al., 1999).

Cultural Variations

Differences in Adult Support

Influenced by cultural values and belief systems, caregivers in different cultures diverge on their views about the centrality of play in toddlers' learning and socialization. The views range from regarding play as an activity with minimal significance that needs to be curtailed to accommodate other responsibilities, such as work (e.g., Gusii and Ngeca in Kenya, Indonesia, and Mexico); to an activity with less significance that is only tolerated without active encouragement (e.g., Guatemala and Italy); to a highly valued activity with adult support and cultivation (e.g., the United States) (see Göncü and Gaskins, 2007). Middle-class American parents perceive play as important for toddlers' development and provide support to their play (see Göncü, 1999; Göncü and Gaskins, 2007). In contrast, mothers of Gusii and Ngeca in Kenya rarely engage in play with their toddlers; instead, they engage in more restricted training interaction to teach their toddlers appropriate social behaviors and work habits (see Field et al., 1981; Göncü and Gaskins, 2007). Because of their role as disciplinarians, mothers may be the least likely playmates for their toddlers in these cultures. Although Italian mothers acknowledge that it is natural for children to play, they do not consider playing with a child is part of maternal role, and only tolerate their children's playful activities (see New's review in Roopnarine et al., 1994). Despite the fact that Mexican and Indonesian toddlers engage in increasingly complex social pretend play, their mothers attribute little importance to play. When asked to play with their toddlers, Mexican mothers engage in

explicitly structured teaching, rather than a child-centered activity (see Farver's review in MacDonald, 1993). Mayan mothers in Guatemala do not recognize play as central to development because they believe that development is generated internally and not influenced by external forces. They discourage children's play, but sometimes tolerate it so that they can be relieved from child care activities and prevent child interference with their work. Less adult support for play is typically observed in cultures where adults have heavy workloads and/or where adult–child interactions generally are more hierarchically organized. American Indian mothers, for example, also facilitate play more often than participate in play (see Field et al., 1981; MacDonald, 1993; Pellegrini, 2011).

Differences in Caregiver–toddler Play Preferences

Even among cultures that value play, significant cultural variations exist. For example, parents in Argentina, France, Japan, and the United States share their views on the importance of play. Yet, because of differences in the history, beliefs, and child-rearing goals, cultural-specific patterns in mother–toddler play have also been documented. Similar to the patterns found in infancy, mothers in independent cultures (e.g., France and the United States) engage in more object-oriented play with their toddlers (e.g., pushing a truck), whereas mothers in interdependent cultures (e.g., Argentina and Japan) engage in more person-oriented and symbolic play, particularly other-directed symbolic play (e.g., feeding a doll) (see Bornstein's review in Göncü and Gaskins, 2007). Compared to Argentine and French mothers, American mothers describe objects and solicit their toddler to engage in symbolic play more often. Presumably, American mothers' preference in encouraging object-oriented play may reflect their emphasis on cultural values related to individual achievement, mastery, and personal choice. In contrast, French mothers tend to remain in a passive and spectator role and prefer an indirect approach of presenting children with options, rather than active stimulation. Argentine and Japanese mothers often solicit and demonstrate symbolic play and praise their toddlers, which are believed to reflect the cultural emphasis on control and interdependency (see Bornstein's review in Göncü and Gaskins, 2007). Mother–toddler play in immigrant families who are adapting to a new culture also demonstrates variations. Mother–toddler play in immigrant Japanese-American and South-American families is more similar to that in American families than to that in their cultures of origin. Immigrant Japanese-American and South-American mothers engage in more exploratory object play and less symbolic play with their toddlers than the dyads in Japan and Argentina (Cote and Bornstein, 2009). Findings with immigrant families support the idea that early play reflects cultural influences.

Differences in Parental Scaffolding of Object Play

Despite the universal benefit of attention-following scaffolding strategy to toddlers' play, mothers across cultures show a distinctive preference for using certain attention regulation strategies and referential speech patterns in guiding their toddlers' play. Mothers in independent cultures (e.g., the United Kingdom, the United States, and Finland) are likely to follow their toddlers' attention and make elaborative comments on their toddlers' actions or the objects involved. On the other hand, Chinese and Korean mothers who are committed to interdependent cultural values are likely to direct their toddlers' attention by introducing a new object or activity and make simple labeling or questions about the object or activity. Mothers' cultural-specific scaffolding strategies appear to echo their divergent cultural concerns and beliefs about early play as a joint activity for promoting individuality and autonomy versus as an educational context for teaching and learning, and for encouraging child compliance, respectively (Sung and Hsu, 2009).

Differences in Play Themes

The thematic topic of toddlers' play is cultural-specific and closely tied to their daily activities. In cultures where toddlers are integrated in adult activities, their play mirrors familiar everyday experiences. Indonesian toddlers, for example, initiate pretend play with their mothers and their pretend play often parallels adult housework activities with family role themes (e.g., Farver and Wimbarti, 1995). The themes of reciprocal imitation games between toddler peers also vary in line with the characteristic play settings. The themes of American toddlers' peer reciprocal games centered largely around play materials (e.g., toys), whereas the themes of peer play for the Seltaman toddlers of Papua New Guinea often resemble adult activities (e.g., beating rhythmically, as in drumming) (Eckerman and Whitehead, 1999).

Differences in Play Partners

In cultures where play is valued, toddlers' primary partners are usually their mothers and other adults. American toddlers play more frequently with their mothers than with their older siblings, and their play with mothers is more complex than with older siblings. In contrast, toddlers in Indonesia, Mexico, and among the Ngeca of Kenya play more with their older siblings than with their mothers. Sibling–toddler play often occurs in the context of sibling caretaking. Older siblings in these cultures provide cognitively challenging and emotionally complex stimulation and make comments and suggestions about play. Toddlers' play behaviors with siblings in these cultures are similar to American toddlers' play behaviors with their mothers. In contrast, mothers in these cultures use more restricted teaching behaviors that involve directing and correcting their toddlers' play (see Farver and Wimbarti, 1995; MacDonald, 1993; Göncü, 1999; Göncü and Gaskins, 2007).

Play during Preschool Age

Children's play reaches its peak at ages 3–5. At this developmental period, play is completely self-initiated without adult support. Even as household work becomes Mayan children's primary daily activity, play remains a dominant place in their activities (20–25% of time) (see Göncü, 1999; Göncü and Gaskins, 2007). Both cultural similarities and variations continuously coexist in preschoolers' play. Whereas gender differences, object use, nonsocial peer play, and the role of child characteristics in play behaviors are similarly observed across cultures, unique differences in ecological conditions,

interpersonal contexts, cultural activities, and socialization goals and practices contribute to cultural variations.

Cultural-General Patterns

Similarities in Object Use, Pretend Play, and Sex Segregation

Regardless of culture, objects are essential to play among preschool-age children. Middle-class children in cultural communities that are more advanced in economic developments typically incorporate manufactured toys (e.g., dolls and cars) into pretend play. In cultural communities where toys are scarce, children constructing props from the materials at hand (e.g., sticks, stones, and kitchen utensils) (Edwards, 2000). Despite marked variations in the frequency and themes of pretend play (see below), children across cultural communities, including those in subsistence societies (e.g., the !Kung San of Kalahari, Marquesan of Polynesia, and Hadza of Tanzania), all engage in pretend play (Edwards, 2000). To create shared understanding and prolong play episodes, children often enact themes that portray familiar adult daily activities (e.g., cooking). The patterns of gender differences in early play vary according to children's age and social contexts (e.g., dyadic parent–child play or peer group play). By age five during solitary and peer play, gender specificity among girls and boys becomes more evident universally. Whereas boys engage more in physical and motor activities, girls engage in more object play and pretend play (Edwards, 2000; Singer et al., 2009). As preschoolers increase their participation in same-sex playgroups, sex segregation becomes more common.

Similarities in Nonsocial Peer Play as Maladaptive

Following a predictable timetable, play progresses from solitary play in the presence of peers toward social play with peers between ages three and five. Play in this developmental period is fundamentally a social activity. Whereas social play with peers is adaptive for children across cultures, nonsocial play is evidenced to be maladaptive. Preschoolers who frequently engage in solitary-passive play (e.g., quiet object explorations and construction activities) and solitary-active play (e.g., repeated sensorimotor actions and solitary pretend play) have similar negative consequences in both independent (e.g., the United States) and interdependent cultures (e.g., China and Malaysia) (Choo et al., 2012; Nelson et al., 2012). Although independent and interdependent cultures diverge on their respective emphases on solidarity and social restraint, nonsocial play in preschoolers is associated with indicators of socioemotional maladjustments such as noncompliance and low prosocial behavior across cultures.

Similarities in the Role of Individual Characteristics

Play behaviors in preschoolers with different cultural backgrounds are similarly associated with the same individual characteristics of the child (e.g., social and cognitive skills). For example, behavioral dispositions including socially adaptive behavioral style, frequent positive social interaction with peers, and high creativity similarly predicted social pretend play behaviors in both Korean- and European-American preschoolers from middle-class families. Such a cultural-general pattern suggests that individual qualities may be just as important for preschoolers' expression of pretend play as culture-specific skills and behaviors (see Howes' and Roopnarine's reviews in Pellegrini, 2011).

Cultural Variations

Differences in Ecological Conditions

Children's play reflects the differences in their ecological environments (Singer et al., 2009). In cultural communities with more advanced economic developments, play is a preferred path to growth and development, which is often viewed as equally important to other goals such as academic success. In contrast, play in subsistence communities is typically utilized as a context to train specific skills for work. Children, particularly girls, as young as age three in many subsistence cultures begin to make a transition from play to work and become an asset to the household economy. In these cultures, as girls take on more work-related responsibilities such as child care and housework and spend less time in play, boys spend relatively more of their time playing (Edwards, 2000; see Göncü and Gaskins, 2007). The availability or constraint of physical space also modifies how children play. Children in subsistence cultures play in open outdoor space. Mayan children, for example, mostly play with objects and engage in gross motor play, but rarely play pretend. When they do, the pretend play tends to be spontaneous and often occurs in large mixed-age playgroups, organized by older children aged 6–11 (see Göncü and Gaskins, 2007). Taiwanese children in an urban city with limited space inside of their homes possess relatively few toys. They are likely to use social routines, not props or toys, to construct joint play with their mothers (Haight et al., 1999).

Differences in Interpersonal Contexts

Children's play depends in part on the availability of various partners (e.g., parents, siblings, peers, and extended family members). With a large number of children available for play, Mayan, Indonesian, Kenyan, and Puerto Rican children often play with siblings and engage in large-group peer play (Edwards, 2000; Trawick-Smith, 2010). A comparative study (Haight et al., 1999) showed that Irish-American preschoolers played pretend more frequently with other children due to their close social networks, whereas Taiwanese preschoolers pretended more with their caregivers due to their limited access to peers. In addition, as discussed above, adults' attitudes toward the significance of early play influence their beliefs about the appropriateness of participation in children's play. Consistent with their beliefs about the benefits of play for children's development and the greater informality in the relationships between children and adults, European-American parents and teachers emphasize play over academic tasks and frequently participate in children's play (see Roopnarine's review Pellegrini, 2011). On the other hand, Asian (e.g., East Indian, Chinese, Korean, Pakistani, and Thai) and Asian-American parents and teachers rarely play with children, viewing play as a culturally inappropriate adult behavior and an activity with little educational values (see Roopnarine's review in

Pellegrini, 2011). Caregivers in many subsistence cultures believe that children learn through work and not play, so they neither encourage children to play nor see themselves as adequate playmates for children (see Pellegrini, 2011; Singer et al., 2009). Caregivers in cultures where play is valued capitalize on children's interest in play and use it as an opportunity for socializing their preschoolers' culturally appropriate behaviors. For example, influenced by Confucian principles, mother–child play in Taiwanese families is characterized as didactic, with more mother-initiated interaction and a thematic highlight on proper conduct. Following a child-centered approach to socialization, European-American mothers allow their children to initiate and lead pretend play; their play is featured by mutuality and a thematic focus on child fantasies (Haight et al., 1999). Acculturation and assimilation of a new culture can modify immigrant parents' beliefs about play and play participation. Middle-class Korean-American mothers with greater cultural integration are more accepting and encouraging of preschoolers' play and report more parent–child play than those with less cultural integration. Korean-American children whose mothers are culturally assimilated and acculturated also participate in more pretend play (see Pellegrini, 2011).

Differences in Cultural Goals and Practices

Play is a form of personal expression by which children express their feelings and real-life experiences. Play also represents children's close observation of the common practices of their culture and attempts to make sense of their cultural experiences. During play, children often enact themes that are important to a particular culture, through which they practice cultural routines and customs. The themes of Mayan children's pretend play, such as playing house or selling produce, mirror adult life (see Gaskins' review in Göncü, 1999). Reflecting the collective orientation and oral storytelling traditions, music play is common among Caribbean preschool-aged children. Puerto Rican 4- and 5-year-olds often share songs, collaborate on the creation of dances, and pretend to be musical performers or dancers (Trawick-Smith, 2010). Thematic contents in peer play appear to resonate with differences in cultural orientations. Middle-class Korean-American children use greater formality and an honorific speech format to enact pretend roles with high status (see Pellegrini, 2011). Consistent with the cultural expectations for self-reliance, independent thinking, and creativity, European-American children engage in pretend play with fantasy themes more than Korean-American children (see Farver et al., 2000). As a reflection of Korean culture's emphasis on group harmony and cooperation, Korean-American children's play focuses on realistic everyday activity and familiar family role themes that can be easily shared. Communicative strategies used by children to structure and maintain pretend play are also consistent with the larger cultural rules. During dyadic peer play, European-American children are more self-assertive and disagreeable, as they describe their own actions, reject their partners' suggestions, and use directive commands. In contrast, Korean-American children are more agreeable and cooperative, as they describe their partners' actions and use polite requests and agreement statements (see Farver et al., 2000; Roopnarine's review in Pellegrini, 2011).

Conclusion

Children bring culture into play. Early play functions to express children's personal experiences of growing up in a cultural setting and to socialize them into accepting and practicing cultural customs for appropriate behaviors. Play may further shape children's cultural experience and bind them to their culture. As a dynamic co-construction between children and their culture, there are many cultural models of play that represent children's experiences in their ecological, interpersonal, and developmental contexts.

Early play by children and their play partners has both, invariant cultural-general and variable cultural-specific dimensions that result from the dynamic interplay between biology and cultural experience. Early play not only has neurobiological underpinnings within the child, but also is influenced externally by caregivers' culturally appropriate behaviors that may inhibit or promote early play. As a culturally mediated activity, play in infancy and early childhood varies in development, varieties, structures, and functions. Cultural variations in early play are not signs of deficiency or neglect (Lancy, 2007; Roopnarine et al., 1994). Rather, cultural differences are reflections of variations in economic and physical ecology (e.g., rural fishing village, urban suburb, or neighborhood), physical context (e.g., availability of space), arrangement of interpersonal context (e.g., availability of play partners), cultural scripts for everyday activities and social conventions (e.g., autonomy or relatedness emphasis), and adult beliefs and expectations for play (e.g., enhances necessary skills or hinders productive activity) (Haight et al., 1999).

Furthermore, a great range of differences within a culture (i.e., within-cultural variations) that change with the characteristics of the child (e.g., age and sex), the parent (e.g., education and acculturation), the family (e.g., socioeconomic status), and the neighborhood are also expected (see Cabrera et al., 2006; Farver et al., 2000; Kenney, 2012; MacDonald, 1993; Roopnarine et al., 1994; Pellegrini, 2011). A culture-sensitive theory of play should seek to understand how both between- and within-cultural variations contribute to the development of early play over time. Accounting for between- and within-cultural variations will require attention to a multitude of biological, developmental, social, and cultural factors. A longitudinal, multilevel, and multicultural approach to the dynamic, complex process of interplays among characteristics of the child, the play partner (e.g., parents and siblings), and the family will be necessary.

See also: Early Emotional Development and Cultural Variability; Human Development, Theories of; Play, Anthropology of; Pretend Play and Cognitive Development; Socialization in Infancy and Childhood.

Bibliography

Bakeman, R., Adamson, L.B., Konner, M., Barr, R.G., 1990. !Kung Infancy: the social context of object exploration. Child Development 61, 794–809.

Bornstein, M.H., Cote, L.R., Haynes, O.M., Suwalsky, J.T.D., 2012. Modalities of infant–mother interaction in Japanese, Japanese American immigrant, and European American dyads. Child Development 83 (6), 2073–2088.

Cabrera, N.J., Shannon, J.D., West, J., Brooks-Gunn, J., 2006. Parental interactions with Latino infants: variation by country of origin and English proficiency. Child Development 77 (5), 1190–1207.

Choo, M.S., Xu, Y., Haron, P.F., 2012. Subtypes of nonsocial play and psychosocial adjustment in Malaysian preschool children. Social Development 21 (2), 294–312.

Cote, L.R., Bornstein, M.H., 2009. Child and mother play in three U.S. cultural groups: comparisons and associations. Journal of Family Psychology 23 (3), 355–363.

Eckerman, C.O., Whitehead, H., 1999. How toddler peers generate coordinated action: a cross-cultural exploration. Early Education & Development 10 (3), 241–266.

Edwards, C.P., 2000. Children's play in cross-cultural perspective: a new look at the six cultures study. Cross-cultural Research 34 (4), 318–338.

Farver, J.A.M., Wimbarti, S., 1995. Indonesian children's play with their mothers and older siblings. Child Development 66, 1493–1503.

Farver, J.A.M., Kim, Y.K., Lee-Shin, Y., 2000. Within cultural differences: Examining individual differences in Korean American and European American preschoolers' social pretend play. Journal of Cross-cultural Psychology 31 (5), 583–602.

Field, T.M., Sostek, A.M., Vietze, P., Leiderman, P.H. (Eds.), 1981. Culture and Early Interactions. Lawrence Erlbaum, Hillsdale, NJ.

Göncü, A. (Ed.), 1999. Children's Engagement in the World: Sociocultural Perspectives. Cambridge University Press, New York.

Göncü, A., Gaskins, S. (Eds.), 2007. Play and Development: Evolutionary, Sociocultural, and Functional Perspectives. Lawrence Erlbaum, Mahwah, NJ.

Gratier, M., 2003. Expressive timing and interactional synchrony between mothers and infants: cultural similarities, cultural differences, and the immigration experience. Cognitive Development 18, 533–554.

Haight, W.L., Wang, X.-L., Fung, H.H.-T., Williams, K., Mintz, J., 1999. Universal, developmental, and variable aspects of young children's play: a cross-cultural comparison of pretending at home. Child Development 70 (6), 1477–1488.

Hewlett, B.S., Lamb, M.E., Shannon, D., Leyendecker, B., Scholmerich, A., 1998. Culture and early infancy among central African foragers and farmers. Developmental Psychology 34 (4), 653–661.

Kärtner, J., Keller, H., Lamm, B., Abels, M., Yovsi, R.D., Chaudhary, N., Su, Y., 2008. Similarities and differences in contingency experiences of 3-month-olds across sociocultural contexts. Infant Behavior & Development 31, 488–500.

Keller, H., Borke, J., Staufenbiel, T., Yovsi, R.D., Abels, M., Papaligoura, Z., Jensen, H., Lohausd, A., Chaudhary, N., Lo, W., Su, Y., 2009. Distal and proximal parenting as alternative parenting strategies during infants' early months of life: a cross-cultural study. International Journal of Behavioral Development 33 (5), 412–420.

Keller, H., Borke, J., Lamm, B., Lohaus, A., Yovsi, R.D., 2010. Developing patterns of parenting in two cultural communities. International Journal of Behavioral Development 35 (3), 233–245.

Kenney, M.K., 2012. Child, family, and neighborhood: associations with parent and peer interactive play during early childhood. Maternal and Child Health Journal 16, S88–S101.

Lancy, D.F., 2007. Accounting for variability in mother–child play. American Anthropologist 109 (2), 273–284.

MacDonald, K. (Ed.), 1993. Parents and Children Playing: Descriptions and Implications. Sate University of New York Press, Albany, NY.

Markus, H.R., Kitayama, S., 1991. Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review 98 (2), 224–253.

Miller, M.A., Harwood, R.L., 2002. The cultural organization of parenting: change and stability of behavior patterns during feeding and social play across the first year of life. Parenting: Science and Practice 2 (3), 241–272.

Nelson, L.J., Hart, C.H., Yang, C., Wu, P., Jin, S., 2012. An examination of the behavioral correlates of subtypes of nonsocial play among Chinese preschoolers. Merrill-palmer Quarterly 58 (1), 77–109.

Pellegrini, A.D. (Ed.), 2011. The Oxford Handbook of the Development of Play. Oxford University Press, New York.

Roopnarine, J.L., Johnson, J.E., Hooper, F.J. (Eds.), 1994. Children's Play in Diverse Cultures. Sate University of New York Press, Albany, NY.

Singer, D.G., Singer, J.L., D'Agostino, H., Delong, R., 2009. Children's pastimes and play in sixteen nations: Is free play declining? American Journal of Play 1, 283–312.

Suizzo, M., 2004. French and American mothers' childrearing beliefs: stimulating, responding, and long-term goals. Journal of Cross-cultural Psychology 35, 606–626.

Sung, J., Hsu, H., 2009. Korean mothers' attention regulation and referential speech: associations with language and play in one-year-olds. International Journal of Behavioral Development 33, 430–439.

Trawick-Smith, J., 2010. Drawing back the lens on play: a frame analysis of young children's play in Puerto Rico. Early Education and Development 21 (4), 536–567.

Plea Bargaining: A Comparative Analysis

Jacqueline Hodgson, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M. Heumann, volume 17, pp. 11507–11513, © 2001, Elsevier Ltd.

Abstract

Plea bargaining is a well-established phenomenon in the United States, such that it is regarded even by the Supreme Court as part of ordinary criminal justice practice, to which Constitutional rights should attach. Various forms of plea and sentence negotiation procedures also exist across Europe and the desire to minimize the number of fully contested trials appears to be a universal criminal justice objective. This article explores the types of bargaining in place in a range of jurisdictions. It focuses in particular on the implications of party-controlled negotiations within more inquisitorial procedures, which are generally characterized by a proactive and truth-seeking judicial function.

Introduction

Plea and sentence bargaining is characterized by several general features that we might recognize as common across a range of jurisdictions and procedural traditions. It requires an admission of guilt from the accused; the accused is offered some reward, incentive, or advantage, either in exchange for, or as a result of the plea; and there is some benefit to the criminal justice system – typically the avoidance of a more lengthy and expensive contested trial. The point in the criminal process at which these negotiations are initiated, the personnel involved, and the relative incentives and benefits available will of course differ. In some instances, explicit bargains are struck between prosecutor and defense lawyer; in others, there is direct judicial involvement; and in others, the system benefit may operate as an implicit reward, without any explicit bargaining taking place.

Growing criminalization coupled with the need to reduce the cost and delay brought about by increasing caseloads makes such bargaining an attractive option. The phenomenon of 'negotiated justice' is well established and has saturated the US criminal system, but we now see various forms of bargaining across the globe. In Europe, for example, in 1987, Recommendation R (87) 18 of the Council of Europe Committee of Ministers encouraged countries to make use of discretionary prosecution and the adoption of summary, simplified and out-of-court procedures for mass and minor offenses, in order to combat delay and to simplify the workings of criminal justice. Formal procedures have been introduced in many jurisdictions (e.g., the *conformidad* in Spain; the *confissão* in Portugal; the *composition pénale* and the *comparution sur reconnaissance préalable de culpabilité* in France), and informal procedures (such as the German *absprachen*) have been placed on a statutory footing. Bargaining exists even in international criminal justice, though the implications are rather different from those in domestic procedure (Langer, 2004, 2006). In both the domestic and the international sphere, the court's verdict represents an official truth or history of the case, a formal finding as to guilt. In the international criminal justice arena, negotiating what will be an important historical record of events has wider implications for reconciliation and justice for victims.

These procedures are not about acknowledging the remorse of the accused, or, arguably, even her guilt, so much as maximizing the efficiency of case management procedures. They are less widespread in continental European jurisdictions (they account for around one third of all cases dealt with by the courts in Italy, for example) but in England and Wales and the United States, guilty pleas have become the norm, even the expectation. Some 70% of cases before the courts are dealt with by way of guilty pleas in the Crown Court in England and Wales, and 68% in the magistrates' court (75% excluding cases that are discontinued at court). In the United States, around 95% of cases in both state and federal district courts are disposed of by way of guilty pleas (**Figure 1**), causing one commentator to observe: "This system loves punishment … This system hates trials" (Miller, 2004: p. 1212). Criminal justice systems are simply not equipped for a mass of contested trials. The full hearing of a case with live witnesses, in which the prosecution case is tested and rules of evidence applied, exists more as a promise or guarantee. It is the possibility of the trial that shapes the nature of the parties' pretrial behavior to some degree.

Understanding Plea Bargaining Comparatively

In Anglo-American systems of criminal justice, the parties are responsible for the investigation, collection, and presentation of evidence; the trial judge or the jury will then assess this evidence, whether written reports or live testimony, in order to determine the verdict. As untested evidence is brought directly by the parties, often complex rules of evidence determine admissibility. With the parties in control of the case in this way, prosecutors are at liberty to negotiate the evidence and the charges. For at least four decades, American scholars have marveled at the relative absence of plea bargaining in most continental European countries whose procedures are rooted in the inquisitorial tradition.

Europe is no longer a 'land without plea bargaining' (as Langbein referred to Germany in 1979), but while some form of plea or sentence bargaining probably now exists in most European countries, the nature and extent of the practice is still very limited compared with the routinized nature of negotiated justice in the United States and increasingly in England and Wales. (For an account of trial avoidance in various jurisdictions,

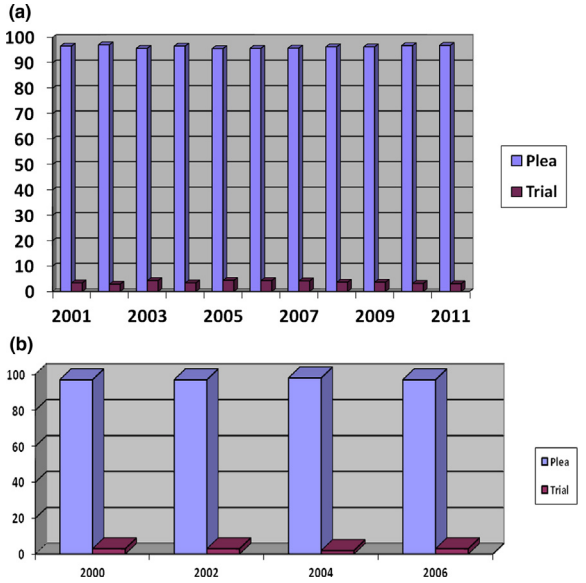

Figure 1 (a) Number of pleas vs trials in Federal Courts (federal data); (b) Number of pleas vs trials in 75 Major County Courts (state data). Source: Online Sourcebook of Criminal Justice Statistics, located at http://www.albany.edu/sourcebook/.

see Thaman, 2010.) In particular, the scope of plea bargaining is very different: plea bargaining is used across the full range of offenses in Britain and the United States, but in most inquisitorially rooted procedures, it is limited to less serious cases. Germany is perhaps the exception to this, where sentence bargaining practices have developed informally (they were not regulated by statute until 2009) and so they were free from the constraints of procedural principles and are applied to both serious and minor offenses (Weigend, 2008; Rauxloh, 2011).

In inquisitorially rooted systems such as France, Spain, or the Netherlands, the judge has a strong fact-finding and truth-seeking role, which is more investigative in nature than the passive and purely adjudicatory function of the judge within the adversarial tradition (e.g., in the United States). Inquisitorial procedures will typically include a pretrial judicial (or at least prosecutorial) function, in which a judicial officer is responsible for conducting or overseeing the investigation. This affects the relationship between trial and pretrial behavior as the prosecution case is considered the product of a judicial (or judicially supervised) investigation, rather than a partisan account to be tested by the defense. The trial of course yields the public pronouncement of the verdict, but rather than being the first hearing and testing of the evidence, it is the confirmation of an earlier set of data-collecting moments. As a result, the trial in more inquisitorial procedures is typically shorter, with fewer witnesses, and so is less resource intensive – it is a trial on the dossier (the case file of evidence), predictably often confirming the assertions of the 'judicial' prosecutor.

It might be presumed, therefore, that in inquisitorial-type systems there is less pressure toward a guilty plea given the more perfunctory trial process than exists in adversarial systems, which have a greater incentive to avoid the lengthy and complex process of the contested trial. Yet, even without a formal system of pleas, it is clear that in France, a confession is the primary goal of the police investigation despite it being carried out under the apparent supervision of the public prosecutor (Hodgson, 2001). Interestingly, some older studies in the United States have also doubted the extent to which caseload pressure is the real reason for trial avoidance (Mather, 1979; Schulhofer, 1984). But while the certainty of a confession is perhaps universally desired by those responsible for the investigation and prosecution of crime, greater opportunities to contest the evidence bring yet more pressure to close down the uncertainty of conflict. It would seem that the introduction of improved procedural defense rights immediately triggers the establishment of procedures through which these rights can be waived. In the 1990s, for example, when Italy was moving to a more accusatorial procedure in which written evidence was replaced by oral evidence at trial, realizing the huge time and resource implications, the reform retained a simplified 'trial by dossier' procedure (see generally, Pizzi and Montagna, 2004; Illuminati, 2005). It recognized that the adversarial ideal of the fully contested trial following a plea of not guilty would not be sustainable in all or even most cases.

Objections to Plea and Sentence Bargaining

At one level, encouraging the guilty accused to admit her guilt sooner rather than later seems quite logical. Yet, even in theoretical terms, the plea and sentence bargaining process might subvert criminal justice norms in several ways.

Supplanting the Role of the Court

Firstly, the court is no longer the arbiter of fact and judges no longer enjoy primary responsibility for case disposition. Instead, the court agrees to an outcome determined by the parties. This objection might be raised more generally against any system of guilty pleas, but it is especially problematic when incentives and rewards are offered. The shift in power from the judge to the parties poses particular problems in jurisdictions rooted in inquisitorial procedure, where the court, rather than the parties, has the primary responsibility for fact-finding.

For example, until relatively recently, France made no formal distinction between cases where the accused admitted the offense and those where she did not: cases were simply dealt with more or less rapidly depending on whether there was an admission (Hodgson, 2012). In 2004, a form of guilty plea procedure was introduced, underpinned by a process of sentence bargaining – the *comparution sur reconnaissance préalable de culpabilité* (CRPC). Where the accused has admitted the charges, the prosecutor can propose a reduced sentence of up to 1 year or half of the maximum sentence for the charge. If the defense agrees, the accused will have her case dealt with by the rapid guilty plea process. The court may only agree or reject the sentence – she may not adjust it in any way. It has been argued that this undermines the proper function of the judge and so the separation of powers. However, the Constitutional Council has ruled the procedure to be constitutional: the judge must do more than rubber stamp the agreement and may reject the plea and sentence if not supported by the facts, or after taking into account the

personality of the accused, the situation of the victim, or the public interest. (The same is true in Italy – the judge may only accept or reject the sentence proposed, but she is required to satisfy herself of the voluntariness of the admission and the evidence against the accused, thus retaining something of the traditional truth-finding judicial role.) In practice, the procedure is designed for a speedy resolution and it seems unlikely that the judge would go beyond the agreement presented to the court. The procedure has been extended to virtually all middle-ranking offenses (*délits*), and while it is currently applied immediately after the police detention and questioning of the suspect, it may now also take place at the close of the judicial investigation (*instruction*). Its use has increased each year and now accounts for 15% of all cases dealt with by the *tribunal correctionnel*.

The Invisible Exercise of Discretion

Second, in the process of plea and sentence bargaining, in most instances (Germany is an interesting exception here, where the defense lawyer negotiates directly with the judge), prosecutorial discretion is supreme and is largely invisible to scrutiny. These are deals behind closed doors, with no record of terms or reasons. In the United States, as power has shifted to the prosecutor, she has been given greater power to charge offenses and so greater bargaining power over charge, plea, and effectively sentence. Miller and Eisenstein (2005) found that increased cooperation between state and federal prosecutors, especially post 9/11, has moved sentencing power away from the judge to the prosecutor in federal cases. Miller (2004) argues that increased federal sentencing mandatory provisions, together with an increase in the range of crimes that are the focus of federal attention, have increased the power of the prosecutor. As a result, there are almost no trials taking place in federal court (the percentage of convictions resulting from guilty pleas having risen from 87 to 97% between 1991 and 2003), with the result that the federal prosecutor has become the federal sentencer. Mandatory sentences have also increased prosecutorial power in the same way, as the choice of charge will effectively determine sentence. In England and Wales, a recent development seeks to regulate the prosecutor's exercise of discretion in plea bargaining in serious fraud cases, ensuring greater fairness, avoiding improper pressure and keeping records of discussions. However, the disclosure provisions remain largely within the prosecutor's discretion, potentially undermining fairness (see Section Inequalities of Bargaining Power: Legal Assistance and Disclosure). Nonetheless, Watson (2010) argues that the guidelines may have a positive impact on the plea bargaining culture, creating a more formal and transparent system.

In jurisdictions where the prosecutor is a judicial officer, there is greater trust in the apparently nonpartisan nature of the prosecutor and the accused may appear to be more protected. For example, in France, the *procureur* is a *magistrat* along with the *juge d'instruction* and the trial judge. She is understood to act in the public interest, in contrast to the defense lawyer, who represents the accused in a more partisan way. However, research shows that in practice, the prosecution function is less judicial and public interest-centered and more partisan and prosecution-oriented than might be assumed (Hodgson, 2001, 2005). As a prosecuting party with hierarchical accountability to the executive, the European Court of Human Rights has also questioned the independence and so the judicial status of the French prosecutor in particular, though this applies potentially to many other jurisdictions whose prosecutors share her role and status (*Medvedyev v. France* 2010, 3394/03; *Moulin v. France* 2010, 37 104/06).

The public interest orientation of the continental public prosecutor should not be overplayed, but she does nevertheless play a different role from her British and American counterparts. While plea bargains in England and Wales and the United States might be struck between the prosecution and defense lawyers before being put to the accused by her own defense counsel for a plea of guilty to be entered, the French prosecutor's responsibility for the investigation makes her a more proactive prosecutor, engaging directly with the accused. She will first propose a sentence and then during a short hearing with the accused and her lawyer, the accused will be asked to confirm that the offense is admitted. If it is not, the prosecutor will simply remit it for ordinary trial. If it is, the prosecutor will listen to the defense mitigation, before proposing a revised, lesser sentence.

The power of the prosecutor as sentence negotiator is strongest where the prosecution function includes making a sentence recommendation to the court (such as in France) and so can control the entire case outcome. The prosecutor's discretion, and so her power as a negotiator, is weaker where there exists a judicially structured sentence discount, often framed either through guidelines or a fixed mandatory discount (though ironically this may increase her power in charge selection). In Spain, for example, defendants receive a mandatory one third sentence discount: there is an incentive, but the incentive itself is not negotiable. In these cases, the charges remain fixed and it is only the sentence that alters. More than 60% of cases in Spain are now dealt with through this procedure. In England and Wales, the prosecutor makes no sentence recommendation, but the judge may give an indication of sentence following a guilty plea (*Goodyear* [2005] EWCA Crim 888) and legislation provides for a sentence discount of up to one third. This is not mandatory in all cases as with Spain, but rather, the discount is greater the earlier the accused makes an admission. The court is required to state the sentence reduction applied and there are sentencing guidelines to structure this and provide a degree of consistency. In Italy too, where the parties agree the sentence (there is no charge bargaining), a discount of up to one third is available.

In sentence bargaining, such as that in France, there must be a clear admission to the offense being prosecuted. In England and Wales, an admission is not a precondition to the bargaining process, as it is the charge itself that will be negotiated. What is needed is a willingness to admit to something. This is built into the process; when advising the police on what to charge, Crown Prosecutors will note on the file from the outset the charges to which they are prepared to accept a plea of guilty. This, of course, runs the risk of even unconscious overcharging in order to build in space for later negotiations.

Inequalities of Bargaining Power: Legal Assistance and Disclosure

Third, equality of bargaining power is essential if the plea or sentence bargain is to be regarded as fair in any sense. There is

an assumption that bargaining takes place in the shadow of the trial and the promise it represents, without acknowledging the inequalities in bargaining power and the risk of a higher penalty if the right to trial is in fact exercised. There are two particular difficulties in ensuring that the accused is an equal negotiator in this process: she requires effective representation and full disclosure of the prosecution evidence.

The pressures of bureaucratic case management and the financial disincentives to take cases to trial make effective legal representation problematic even in adversarial-type procedures, and researchers have been critical of defense counsel's role in pressing accused persons to accept a plea deal (McConville et al., 1994). This is significant, as the courts assume defense counsel is a kind of guarantor of the voluntariness of the defendant's plea. For example, France ensured the right of the accused person to consult with a lawyer during the CRPC procedure.

Yet, despite their reliance on lawyers in this way, historically the judiciary has ignored the realities of plea bargaining and engaged in what Bibas (2011: p. 1119) describes as "anachronistic formalism focused on jury trials." The courts in England and Wales are also extremely reluctant to look behind a guilty plea and in particular to criticize legal advice on plea. In the United States, Bibas (2011) argues that *Padilla v. Kentucky* 130 S. Ct. 1473 (2010), in which the Supreme Court held that defective advice as to the collateral consequences of a guilty plea (here, deportation) could violate the defendant's Sixth Amendment guarantee of effective legal assistance, signals a shift toward acknowledging guilty pleas as the norm. Furthermore, it recognizes that 'plea bargaining needs tailored regulation in its own right, not simply a series of waivers of trial rights' (Bibas, 2011: p. 1138). The subsequent cases of *Lafler v. Cooper* 132 S. Ct. 1376 (2012) and *Missouri v. Frye* 132 S. Ct. 1399 (2012) have set this out more explicitly: the accused was prejudiced by counsel's failures during the plea bargaining process and the court ordered new trials. Justice Kennedy, writing for the majority, stated that plea bargaining is the norm of the criminal justice system, therefore 6th Amendment guarantees must apply there too, as well as during trials. Alschuler (2013) is skeptical of the significance of these two decisions, arguing that they will neither remove the temptation for lawyers to disregard their client's best interests, nor ensure effective legal assistance. Plea bargaining remains private and cases such as these two rely on defense counsel being forthcoming about their own unprofessional behavior.

The problem of legal assistance is more acute in many inquisitorially rooted procedures, where there is a less established defense role given the historic reliance on, and trust of, judicial investigations. In France, the defense had almost no role in the pretrial investigation until recently, and even now it is severely limited: the suspect may have a 30 min consultation with her lawyer and the lawyer may be present during the police interrogation but may not intervene or ask any questions until the questioning has finished. The lawyer is told only the basic reasons for arrest and detention. Placing the lawyer in a direct dialogue with the prosecutor during the CRPC procedure was therefore a new development and nearly a decade on, the procedure seems to work effectively. The prosecutor proposes the sentence, the defense lawyer points out weaknesses in the case and mitigates for the accused, usually resulting in the prosecutor further reducing the sentence offer.

If there is to be equality of bargaining and if guilty pleas are to be fair and not simply a way of enabling poor investigations to go unchecked, the accused needs to know the nature of the evidence against her. In all jurisdictions, disclosure at or before the first hearing will typically be a summary of the evidence together with the written statements of the key witnesses, though the stage and the extent of disclosure differ. For example, in France and the Netherlands, lawyers are relative newcomers to the police station environment and are told very little about the accusation during the detention and interrogation of the suspect. Lawyers in England and Wales have been present at this stage in the process for nearly 30 years. Relationships of trust have been built and the defense role has become more accepted. Lawyers are therefore told much more of the evidence against the suspect, allowing the defense to begin to structure the defense case at this early stage, including the offenses to which the accused should and should not plead guilty. In sentence bargaining such as the plea system in France, the essence of the bargain is that the accused accepts the facts as presented: this is a necessary precondition to the abbreviated procedure. In plea bargaining cases, however, such as in the United States and England and Wales, the prosecutor may also engage in fact bargaining, agreeing to adjust the agreed facts, by, for example, putting in evidence to the court that the victim was slapped but not punched. This bargaining requires a degree of trust between defense and prosecution, as there is no opportunity to test out the evidence gathered by the police. It is not clear that this trust is well-founded. The prosecutor has no involvement in the investigation and will often have no more idea of the nature or reliability of the evidence than the defense. This is true even in systems where the prosecutor is responsible for the investigation (such as France and the Netherlands) as supervision is, at best, by telephone and file review (Hodgson, 2001, 2005).

Disposing of Cases Earlier in the Criminal Process

Plea and sentence bargaining sit within a wider trend toward the earlier and more rapid disposition of cases. From abbreviated trials to penalties administered by prosecutors and even police officers, legal systems have developed an array of measures that limit the number of cases coming before the court. (See, for example, the French *composition pénale*, in which the prosecutor proposes measures similar to court penalties; the Italian *patteggiamento*, in which the accused agrees a penalty reduction with the prosecutor, reviewed by the judge and *giudizio abbreviato* or abbreviated trial where the penalty is reduced in exchange for an abbreviated trial on the dossier; and the German *strafprozessordnung*, which enables prosecution to be avoided in exchange for the defendant paying money to the victim, a charity, or the state.) As guilty pleas and alternatives to prosecution become the norm, the focus shifts not just to obtaining an admission, but rather to obtaining an admission as early as possible in the process. In England and Wales, out-of-court disposals peaked in 2007 when they outnumbered those cases disposed of at court, though they

have slowly declined since then (see Hodgson and Roberts, 2010). The inequalities of defense representation and evidence disclosure are greater the earlier in the process that the deal is struck.

Prosecution discretion is also stronger when bargaining takes place as an alternative to prosecution. The Netherlands, for example, has a target of disposing of 80% of cases at the police investigation stage, avoiding prosecution. This is tolerated in inquisitorially rooted procedures, because of the trust in the apparent public-interest orientation of the public prosecutor. However, even where the prosecutor is responsible for the investigation in countries such as France and the Netherlands, they will know little of the detail and strength of the evidence: it is unlikely that the police interrogation of the suspect will have been heard, or closed circuit television (CCTV) evidence viewed. Disposing of cases ever earlier in the process means that both defense and prosecution case knowledge is incomplete and the inequalities of bargaining power are therefore weaker.

Where suspects receive deal offers while in police custody or at the first court appearance, in many jurisdictions they will most likely be represented by a duty lawyer who has little or no knowledge of the accused or her case, placing them in a poor position to negotiate. Recent research on the organization and delivery of custodial legal advice shows that arrangements vary across jurisdictions (Blackstock et al., 2014). In England and Wales, custodial legal advice has been in place on a statutory footing since 1984 and is supported by a publically funded legal aid scheme. Duty lawyers are available but suspects may also select their own lawyer, and in most instances, the same lawyer (or at least the same firm) will go on to represent them at court.

This is very different from France and the Netherlands, where most suspects will be seen by one duty lawyer at the police station and a different attorney in court. In this way, legal assistance during police detention and interrogation is a discrete transaction, rather than part of a continuous engagement with the defense case. This has important consequences for the defense's ability to assess the strength of the prosecution case. For example, the prosecution will typically disclose a summary of the suspect's interrogation by the police. If the lawyer in court was not present at this interrogation, she cannot independently judge the accuracy of this summary – which is of concern, because the summary is prepared by police staff and tends to reflect the police interpretation of events, minimizing mitigation or exculpatory comments. In France, the defense may only view the video recording of the interrogation (used in juvenile cases) if the summary is disputed. But as the lawyer at court is unlikely to be the same attorney present during the interrogation, she will have no personal knowledge of the case beyond what the accused is able to tell her.

The quality of legal representation is also variable. In countries where custodial legal advice has recently been introduced or strengthened (largely as a result of the European Court of Human Rights case *Salduz v. Turkey*, 2008), duty rotas are staffed by lawyers with little or no criminal law experience in order to spread the workload – and in some instances to enable lawyers to build up a criminal practice. This replicates in part what happened in England and Wales in the early years following the introduction of the statutory right to custodial legal assistance: lawyers could not accommodate the additional workload and relied on inexperienced and unqualified staff (McConville and Hodgson, 1993). Suspects are the losers in this – the courts deal with them as though they have received effective legal assistance, when in fact they have not. As a result, accused persons are prevented from challenging evidence in court, as the presence of a defense lawyer is deemed to be adequate protection and a guarantee of procedural fairness.

Is Plea Bargaining Typical of Adversarial Procedure?

Plea bargaining is more predominant in adversarial procedures and scholars in both the United States and Britain have argued that it subverts the adversarial process: it removes cases from trial and so prevents the defendant from testing out the prosecution case. If the prosecution is not required to discharge the burden of proof in a public courtroom, this leaves space for weak cases to go unchecked and the adversarial ideal is undercut. Research suggests that this is precisely what does happen. Prosecutors file weak cases, in the knowledge that the vast majority will result in a guilty plea. Incentives and inducements can only accentuate this.

On the other hand, where plea and sentence bargaining has been introduced elsewhere, it is regarded as a more adversarial practice that undercuts and even subverts inquisitorial ideals and procedures. The essence of plea and sentence bargaining is that it is party centered – the plea and sometimes sentence is determined by the prosecution and defense, not by the court. As inquisitorially rooted procedures traditionally have a more active judicial role at trial – questioning witnesses and interrogating the evidence – passing the finding of guilt and the sentence to be imposed into the hands of the parties can be seen as deviating from the inquisitorial tradition and the truth-finding role of the judge. This was the problem raised in the introduction of the French guilty plea procedure. In Italy too, initial attempts at the introduction of an adversarial procedure in 1988 were resisted strongly by the judiciary and required constitutional change. Legal cultures are slower to adapt, and the idea of negotiated justice with contractual-type relationships does not sit well with the traditional public-interest-oriented Italian judicial role (Maffei, 2004).

Whatever the procedure within which bargaining takes place, it requires the accused to waive her trial rights and so the opportunity to test out the accusation against her. What does she receive in return? Is it a bargain, a lesser punishment than she deserves, as Justice Scalia claimed in *Lafler*? This certainly seems unlikely in the United States, where the "expected post-trial sentence is … like the sticker price for cars: only an ignorant, ill-advised consumer would view the full price as the norm and anything less a bargain" (Justice Kennedy in *Lafler*, quoting Bibas, 2011: p. 1138). In inquisitorial-type systems, bargaining is less widespread and not yet institutionalized to the extent of systematic overcharging. The more active judicial and prosecutorial functions also appear to guard against the worst excesses of purely negotiated justice, for the time being at least.

See also: Crime Policy in the United States; Policing; Prosecution; Punishment: Social and Legal Aspects; Trials.

Bibliography

Alschuler, A.W., 2013. Lafler and Frye: two small band-aids for a festering wound. Duquesne University Law Review 51, 673–707.

Bibas, S., 2011. Regulating the plea bargaining market: from Caveat Emptor to consumer protection. California Law Review 99 (4), 1117–1161.

Blackstock, J., Cape, E., Hodgson, J., Ogorodova, A., Spronken, T., 2014. Inside Police Custody. Intersentia, Antwerp.

Hodgson, J., 2001. The police, the prosecutor and the juge d'instruction: judicial supervision in France, theory and practice. British Journal of Criminology 41 (2), 342–361.

Hodgson, J., 2005. French Criminal Justice: A Comparative Account of the Investigation and Prosecution of Crime in France. Hart Publishing, Oxford.

Hodgson, J., 2012. Guilty pleas and the changing role of the prosecutor in French criminal justice. In: Luna, E., Wade, M. (Eds.), The Prosecutor in Transnational Perspective. Oxford University Press, Oxford, pp. 116–134.

Hodgson, J., Roberts, A., 2010. Criminal process and prosecution. In: Cane, P., Kritzer, B. (Eds.), Oxford Handbook of Empirical Legal Studies. Oxford University Press, Oxford.

Illuminati, G., 2005. The frustrated turn to adversarial procedure in Italy. Washington University Global Studies Law Review 4 (3), 567–581.

Langer, M., 2004. From legal transplants to legal translations: the globalization of plea bargaining and the Americanization thesis in criminal procedure. Harvard International Law Journal 45 (1), 1–64.

Langer, M., 2006. Rethinking plea bargaining: the practice and reform of prosecutorial adjudication in American criminal procedure. American Journal of Criminal Law 34, 223.

Maffei, S., 2004. Negotiations 'on evidence' and 'on sentence' – adversarial experiments in Italian criminal procedure. Journal of International Criminal Justice 2 (4), 1050–1069.

Mather, L.M., 1979. Plea-Bargaining or Trial: The Process of Criminal-Case Disposition.

McConville, M., Hodgson, J., 1993. Custodial Legal Advice and the Right to Silence. HMSO, London.

McConville, M., Hodgson, J., Bridges, L., Pavlovic, A., 1994. Standing Accused: The Organisation and Practices of Criminal Defence Lawyers in Britain. Clarendon Press, Oxford.

Miller, L.L., Eisenstein, J., 2005. The Federal/state criminal prosecution nexus: a case study in cooperation and discretion. Law & Social Inquiry 30, 239–268.

Miller, M.L., 2004. Domination & dissatisfaction: prosecutors as sentencers. Stanford Law Review 56, 1211.

Pizzi, W.T., Montagna, M., 2004. The battle to establish an adversarial trial system in Italy. Michigan Journal of International Law 25, 429–466.

Rauxloh, R., 2011. Formalization of plea bargaining in Germany: will the new legislation be able to square the circle? Fordham International Law Journal 34 (2), 296–331.

Schulhofer, S.J., 1984. Is plea bargaining inevitable? Harvard Law Review 97, 1037.

Thaman, S. (Ed.), 2010. World Plea Bargaining: Consensual Procedures and the Avoidance of the Full Criminal Trial. North Carolina Academic Press, Durham.

Watson, D., 2010. The attorney general's guidelines on plea bargaining in serious fraud: obtaining guilty pleas fairly? Journal of Criminal Law 74 (1), 77–90.

Weigend, T., 2008. The decay of the inquisitorial ideal: plea bargaining invades German criminal procedure. In: Jackson, J., Langer, M., Tillers, P. (Eds.), Crime, Procedure and Evidence in a Comparative and International Context. Hart Publishing, Oxford, pp. 39–64.

Pluralism

Richard J Ellis, Willamette University, Salem, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'pluralism' was first introduced into political science by the Englishman Harold Laski who used it to designate an early twentieth-century school of political and legal thought that sought to vindicate the rights and autonomy of associations by challenging the modern doctrine of unlimited, unitary state sovereignty. In the second half of the twentieth century the term became associated with a diverse group of American political scientists who traced their intellectual roots to Arthur Bentley and to the empirical study of pressure groups in the 1920s and 1930s. The emphasis in this postwar pluralism was less on the normative desirability of groups than on the empirical power of groups as an explanatory factor in political and social life. A term of many meanings, empirical as well as normative, pluralism has been a target of many criticisms, some of which are warranted, many of which are not. Ironically, while criticisms of pluralism as an empirical theory have intensified over the last decades of the twentieth century, the appeal of pluralism as a normative vision has undergone a marked revival in recent years.

The concept of 'pluralism' is usually associated with post-World War II American behavioral political science, but pluralism's roots extend back to the early twentieth century, and are as much European as American. The Englishman Harold Laski introduced the term 'pluralism' into political science in a number of essays published mostly while teaching at Harvard University between 1916 and 1920. For the term, Laski was indebted to the pragmatist philosophy of William James, whose book, *A Pluralistic Universe* (1909), Laski hailed as of "vital significance for political theory." In philosophy, pluralism, as distinguished from 'monism,' denoted a system of thought which recognizes more than one ultimate principle. Although Laski borrowed the word 'pluralism' from American pragmatism, the ideas he advanced under the rubric of pluralism were taken largely from a heterogeneous group of European, and in particular English, legal historians and political theorists whose work Laski encountered while a student at Oxford immediately prior to World War I. Among the most important of these were the German medievalist Otto Von Gierke; England's preeminent legal historian, Frederick Maitland; J. Neville Figgis; and Laski's Oxford tutor, Ernest Barker, who in 1914 coined the terms 'polyarchism' and 'the new federalism' to describe this new set of ideas.

The English pluralists (as they would come to be known by the time Laski returned to England in 1920) sought to vindicate the rights and autonomy of associations by challenging the modern doctrine of an unlimited and unitary state sovereignty. Pluralists drew an invidious contrast between the 'federalistic structure of medieval society,' in which a dense network of autonomous associations flourished, and the modern notion of state sovereignty that reduced "to impotence and then to nullity all that intervenes between Man and State" (Nicholls, 1975, p. 162). But pluralism was much more than a lament about the world that had been lost; it was also an empirical critique of the concept of sovereignty. Sovereignty was a myth, the English pluralists maintained. The world, Figgis explained, is "not, as a fact, composed of a few vast unities known as States, set over against crowds of isolated individuals." Rather each society "is a society of societies, each and all with rights, liberty and life of their own" (Nicholls, 1975, p. 140). The idea of unlimited state sovereignty, Laski agreed, was "out of accord with the facts of life" (1921, p. 236); the "history of societies" showed that organized groups, even unpopular ones like militant female suffragists, striking miners, or conscientious objectors, imposed limits on the allegedly sovereign state, forcing public officials to adopt policies to which they were opposed (1917, p. 12).

Although pluralists frequently appealed to "the real facts of human society" (Nicholls, 1975, p. 141), their challenge to what they termed 'the monistic theory of the state' was more normative than empirical, though typically the 'is' and the 'ought' were woven seamlessly together. In arguing that "the state is only one among many forms of human association" (Laski, 1919, p. 65), for instance, Laski meant not only to highlight the empirical fact of the social and political importance of associations and groups, but also, and more importantly, to advance a radically anti-statist normative vision. For the young Laski, the state was owed no more loyalty than other groups; each group, including the state, deserved only the loyalty it earned "by virtue of its achievement" (1921, p. 170). The 'ought,' in Laski's mind, derived from the 'is'; it was the "facts themselves" that "dissolve … the inherent claim of the State to obedience" (1917, p. 23).

Not all pluralists shared Laski's anarchic and syndicalist sympathies. Others, such as Figgis and Barker, allowed that the state had an important role to play in regulating conflict between associations and thus in creating the political conditions necessary for associations to pursue their own ends. Although Figgis and Laski disagreed on what should replace the monistic state, they both looked to pluralism less to explain the world than to change it. Their main objective was to help the groups they cared most deeply about, which, in the case of Figgis, an Anglican priest, were religious associations, and in the case of Laski as well as the guild socialist theoretician, G.D.H. Cole, were labor and the trade unions. And herein lies an explanation for English pluralism's precipitous decline in the 1920s. For pluralists failed to persuade others, or even eventually themselves, that pluralism was the most effective means to aid the groups they wanted to assist. Laski himself gradually abandoned or watered down his pluralist ideas

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93092-7

because, along with the rest of the left, he came to believe, particularly after the Great Crash of 1929, that labor had more to gain from an active, even paternal, state than from a minimal, pluralist state.

Ironically, at the same time that English pluralism as a normative vision and intellectual force was in decline, the empirical study of organized groups was gaining ground in American political science, particularly after the publication of Peter Odegard's *Pressure Politics: The Story of the Anti-Saloon League* (1928) and Pendleton Herring's *Group Representation Before Congress* (1929). These sorts of empirical studies of the role of organized groups in the American political process never interested Laski. Indeed, they only confirmed his belief that the trouble with American political science was that "everywhere the emphasis is on description and not principle" (Runciman, 1997, p. 192), a diagnosis diametrically opposed to that held by most American political scientists interested in the role of groups during this era. They believed, with Arthur Bentley, that there was too much moralizing about the evils of special interest groups and too little objective empirical description and analysis of group behavior in the political process. Over the next quarter century American political scientists worked diligently to remedy these empirical shortcomings, such that by 1950, a year before the publication of David Truman's landmark book, *The Governmental Process*, Merle Fainsod could conclude that "there are few fields which American political scientists have cultivated as intensively and profitably as the analysis of the dynamic interplay of interest groups as they give shape to public policy at almost all governmental levels" (Crick, 1959, p. 118).

In 1950 the discipline was relatively rich in empirical studies of interest group influence but the term 'pluralism' was rarely if ever used, though political scientists were beginning to speak of the "group basis of politics" (Latham, 1952) or the "group theory of politics." Truman as well as Bertram Gross had helped to revive interest in Bentley who, in his 1908 book, *The Process of Government*, had argued that groups, not laws, institutions, or constitutions, were the essential "raw materials" of government. Truman followed Bentley in trying to understand the government process by treating groups, specifically interest groups, as "a primary unit of analysis" (Truman, 1971/1951, p. xix). Over the next decade, this increasingly influential mode of analysis would become known by advocates and even more frequently by detractors as 'analytical pluralism' or simply 'pluralism.'

Critics were quick to charge pluralists with believing that groups were all that mattered, often pointing to Bentley's audacious remark that "when the groups are adequately stated, everything is stated" (1967/1908, p. 209). But most of those tagged as pluralists, including most especially Truman, never maintained anything so implausible. Pluralists looked to groups as "*a* major explanatory variable" (Truman, 1971/1951, p. xxii; emphasis added), but explicitly rejected the notion that government officials merely passively registered or mirrored the sum total of organized group preferences. The relative importance of groups and government officials in the shaping of policy, Truman emphasized, was an empirical question, not one to be settled by "preassigning weights and roles" (p. xxv). Analytic pluralists, to be sure, generally avoided using the term 'the state,' but this was not because their explanations were 'society-centered' or because they believed government officials lacked autonomy or importance vis-à-vis societal groups. Rather it was because they doubted, for epistemological as well as empirical reasons, that the myriad institutions and officials of government, particularly in the United States, were coherent enough to warrant a locution that treated the government as a unitary actor with a single will (Ellis, 1992). The same doubts fed pluralists' skepticism of concepts like 'the public interest' or 'the will of the people,' since society no less than the state was a complex mix of rival interests and objectives. In this respect, the modern American pluralists were kindred spirits of the early twentieth-century English pluralists.

Another strain in criticism of American pluralism objected not to its focus on societal groups but to its portrayal of the distribution of power within the United States as dispersed. Pluralists, it was claimed, slighted the severe inequalities in resources and power in American society, and believed mistakenly that the system was self-correcting: that the mobilization of one group would necessarily lead to the mobilization of an equal and opposite countervailing pressure. Pluralists were also said to believe that all important or legitimate interests and opinions were represented within the political system, or even that all groups had substantially equal access to the policy-making process. As the ideological dust kicked up in the 1960s slowly settled, scholars (e.g., Smith, 1990; Hirst, 1987) had begun to recognize that few, if any, of those commonly described as pluralists harbored such naïve views of the political process. The important question, in any event, is not whether it is possible to find someone myopic enough to believe that all groups have equal or nearly equal power, but whether pluralists are logically committed to these propositions. How widely dispersed power is within any given policy domain or in any given place or time is an empirical question that cannot be settled by methodological or theoretical fiat. Nor can it be answered satisfactorily without a comparative perspective: the same distribution of power may be reasonably characterized as either concentrated or dispersed depending on one's comparative referent or on one's normative benchmark. There is no inherent reason that an empirical commitment to the study of interest groups *à la* Truman need predispose the analyst to find any given distribution of political power. Nor need a normative commitment to pluralism, understood as wide dispersion of political influence and power, necessarily predispose one to find pluralism as a sociopolitical fact in any given setting.

The term 'pluralist' is a particularly confusing label because it is often applied to those who conclude that power in a given locale is dispersed rather than concentrated. Pluralism can be a useful term to describe either an empirical reality – one in which there is, for example, widespread dispersal of power, and bargaining rather than hierarchical decision-making – or a normative commitment to such dispersed power, but it does not help to call those scholars who find dispersed power 'pluralists.' When Robert Dahl (1961) finds a pattern of dispersed influence in contemporary New Haven politics, it no more makes him a pluralist than does his discovery of concentrated oligarchic power in early New Haven make him an elitist.

Dahl remains the archetypal pluralist in the minds of many, even though he is not a group theorist on the model of Bentley and Truman. Whether or not Dahl counts as a pluralist, he is

certainly among the keenest students of the dilemmas of pluralism. As a graduate student at Yale in the 1930s, Dahl took a year-long seminar on recent political thought from Francis Coker, who had written the entry on pluralism for the *Encyclopedia of the Social Sciences* (1934). There Dahl encountered Laski and the English pluralists as well as Coker's ambivalent critique, in which he praised pluralists for drawing attention to the importance of autonomous groups intermediate between the individual and the state, while also criticizing pluralists for slighting the need for centralized political control. State control was necessary, Coker argued, not only to referee intergroup conflict but also to protect the "unorganized interests of the general body of citizens" against the organized. Moreover, without active state intervention pluralist group autonomy would result in more powerful groups abrogating the rights of weaker groups. Of particular concern to Coker was the unequal relationship between employers and employees. Dahl amplified his teachers' arguments in *Politics, Economics, and Welfare* (1953). It was one thing, Dahl and his coauthor Charles Lindblom argued, to recognize (as Laski did) that government officials typically could not "enforce decisions by unilateral, hierarchical means" but must bargain with powerful and autonomous groups to negotiate "temporary armistices." It was "quite another to turn this social fact into a prescription of the desirable, and to argue in effect that politicians should not even attempt to exercise 'the last say,' but should turn that power over to national organizations bargaining among themselves" (p. 498). The fundamental dilemma of a pluralist democracy, in Dahl's (1982) view, was that autonomous associations, groups, or organizations are highly desirable and yet also capable of doing great harm if not controlled by a central authority.

Why is pluralism desirable? Advocates of pluralism see autonomous organizations fulfilling a variety of beneficial purposes. Some emphasize pluralism's role in facilitating individual development and self-expression (Rosenblum, 1998), while others emphasize the importance of groups in integrating the individual into society (Nisbet, 1975). Associations are also seen as seedbeds of civic virtue, an enabling environment in which citizens learn the habits of self-government (Putnam, 2000). A different emphasis is evident among those who stress pluralism's defensive political function, as an organizational counterweight to the coercive powers of the central government (Shklar, 1989, p. 28). Pluralism's vital contributions to democratic governance have long been Dahl's primary focus, a concern that became richly salient again as scholars, politicians, and citizens attempted to understand and direct the process of democratization in the former communist countries in Eastern Europe and the Soviet Union. In the absence of vibrant organizational pluralism, the experiences of these nations suggest, it is difficult for formal democratic institutions such as elections, parliaments, and presidents to curb hierarchy and domination (Howard, 2003).

Pluralism may be essential to the functioning of large-scale democracies but that does not mean, as Dahl emphasizes, that pluralism is without disadvantages. To begin with, group leaders may dominate or oppress group members, particularly where the costs of exit from a group are high. Less ominously, the group's elites may fail to represent the opinions of its membership – the elites may be more conservative or more radical, more compromising or more confrontational, than the membership. The English pluralists at times obscured these problems by treating organizations, whether churches or trade unions, as if they were undifferentiated organic wholes with 'real' personalities (even while they denied such a real personality to the state), but American analytic pluralism was rarely, if ever, guilty of reifying groups. Far from treating large groups as if they were homogenous, as is sometimes alleged (Kariel, 1968, p. 168), Bentley and other American group theorists were keenly aware, in Odegard's words, that all groups "are themselves almost infinitely divisible into subgroups whose interests are not always congruent or compatible with one another" (Bentley, 1967/1908, p. xxii).

A second problem with organizational pluralism is that all groups are not created equal: organized interests, all other things being equal, have an advantage over unorganized interests, and some interests are easier to organize than others (Olson, 1965). Autonomy for the lambs may well mean a quick dinner for the wolves. It is the weaker groups, as Schattschneider (1960) argued, who generally have the greatest interest in "enlarging the scope of conflict" by appealing to public authority and thereby modifying the imbalance in private power relations. Recognition that group autonomy may serve the interests of the most powerful groups is a potent argument against the radically antistatist vision of the English pluralists, but not nearly so troubling to the pluralist vision of postwar American political scientists, most of whom were New Deal Democrats. The problems of group autonomy are real, but they need to be counterbalanced with an appreciation for the equally real limits of central control. The articulation of group preferences and mobilization of group interests being absent, a central political authority would define who was a group and whose interests were legitimate. Without group pressures, the state would find it difficult if not impossible to gauge the intensity with which preferences are held. Group autonomy without central control is bad for the lambs, but central control without group autonomy is no better.

Other critics worry that, even if all groups were equal, pluralism may, in Dahl's words (1982, p. 43), "deform civic consciousness" by fostering particularistic or short-run demands at the expense of a broader common good or long-run social needs. Whether civic consciousness is higher or the common good is more closely attended to in those countries or policy domains where interest groups are less prevalent or less assertive is an empirical question, the answer to which is far from self-evident. A related concern is that private associations may be given essentially public functions, thereby removing or reducing democratic control over governmental decisions. One manifestation of this problem is European-style corporatism, in which bargains are worked out between peak associations representing labor, business, and farmers. In the United States, the problem of alienating democratic control may take the form of powerful groups capitalizing on the fragmentation of the system by 'capturing' governmental officials in a specialized policy domain (McConnell, 1966; Lowi, 1969). Dispersed power is not inconsistent with groups monopolizing power in specialized segments of the political system. Again the extent to which this describes accurately the policy-making in any given system or policy domain is an empirical question.

Among the most common criticisms of pluralism is that it was little more than a self-satisfied apologia for American or,

alternatively, Western, democracy. Whether or not those who were associated with pluralist ideas were satisfied with the state of democracy in America or Western Europe, pluralism as a normative vision is not inherently conservative. Hirst (1989), for instance, uses pluralist ideas to launch a blistering critique of the British political system, particularly Parliament's claims to unlimited sovereignty. Measuring Western democracies against an exacting yardstick of pluralist principles, Hirst finds severely wanting not only the West's narrowly plebiscitarian vision of democracy, in which "the people periodically choose who are to be their elected masters" (1997, p. 1), but also the over-centralization of the provision and delivery of public services. Taking pluralism seriously, Hirst maintains, means radically reshaping governance, not only in the nation state but also in international institutions such as the European Community. Dahl himself draws upon pluralist ideas to argue for a dramatic devolution of power to cooperatively owned economic enterprises. And in much of the non-Western world, even the tamest of pluralist ideas can pose a subversive challenge to entrenched state power and economic privilege.

Writing in the 1968 *International Encyclopedia of the Social Sciences*, Henry Kariel wrote an obituary for political pluralism, which "as an ideology … only lingers quietly as a submerged, inarticulate ingredient of Western liberalism" (p. 168). Almost half a century later, it is clear that reports of pluralism's demise have been greatly exaggerated. Particularly as a normative vision, as an antidote to plebiscitary democracy and to centralized state control, political pluralism has shown renewed appeal and relevance across the globe in the twenty-first century.

See also: Attitudes, Political and Civic Culture; Democratic Theory; Interest Groups, History of; Interest Groups; Party Systems; Political Representation; Politics and Participation; Voluntary Associations, Sociology of; Voluntary Organizations.

Bibliography

Baumgartner, F.R., Leech, B.L., 1998. Basic Interests: The Importance of Groups in Politics and in Political Science. Princeton University Press, Princeton, NJ.

Bentley, A.F., 1967[1908]. The Process of Government. Belknap Press, Cambridge, MA.

Coker, F., 1934. Pluralism. In: Seligman, E. (Ed.), Encyclopedia of the Social Sciences. Macmillan, New York.

Connolly, W.E. (Ed.), 1969. The Bias of Pluralism. Atherton Press, New York.

Crick, B., 1959. The American Science of Politics. University of California Press, Berkely, CA.

Dahl, R., Lindblom, C., 1953. Politics, Economics and Welfare. Harper, New York.

Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven, CT.

Dahl, R.A., 1982. Dilemmas of Pluralist Democracy: Autonomy vs. Control. Yale University Press, New Haven, CT.

Eisenberg, A.I., 1995. Reconstructing Political Pluralism. State University of New York Press, Albany.

Ellis, R.J., 1992. Pluralist political science and 'the state': distinguishing between autonomy and coherence. Polity 24, 569–589.

Gunnell, J.G., 1995. The declination of the 'state' and the origins of American pluralism. In: Farr, J., Dryzek, J.S., Leonard, S.T. (Eds.), Political Science in History: Research Programs and Political Traditions. Cambridge University Press, Cambridge, UK.

Herring, P., 1929. Group Representation before Congress. Johns Hopkins Press, Baltimore, MD.

Hirst, P.Q., 1987. Retrieving pluralism. In: Outhwaite, M., Mulkay, M. (Eds.), Social Theory and Social Criticism. Basil Blackwell, Oxford, UK.

Hirst, P.Q. (Ed.), 1989. The Pluralist Theory of the State: Selected Writings of G.D.H. Cole, J.N. Figgis and H.J. Laski. Routledge, London.

Hirst, P.Q., 1997. From Statism to Pluralism: Democracy, Civil Society and Global Politics. UCL Press, London.

Howard, M.M., 2003. The Weakness of Civil Society in Post-Communist Europe. Cambridge University Press, Cambridge.

James, W., 1909. A Pluralistic Universe. Longmans, Green, New York.

Kariel, H.S., 1961. The Decline of American Pluralism. Stanford University Press, Stanford, CA.

Kariel, H.S., 1968. Pluralism. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York.

Laski, H.J., 1917. Studies in the Problem of Sovereignty. Yale University Press, New Haven, CT.

Laski, H.J., 1919. Authority Is the Modern State. Yale University Press, New Haven, CT.

Laski, H.J., 1921. The Foundations of Sovereignty and Other Essays. Harcourt, Brace and Company, New York.

Lowery, D., Gray, V., 2004. A neopluralist perspective on organized interests. Political Research Quarterly 57, 163–175.

Lowi, T.J., 1969. The End of Liberalism. W.W. Norton, New York.

McConnell, G., 1966. Private Power and American Democracy. Knopf, New York.

Merelman, R.M., 2003. Pluralism at Yale: The Culture of Political Science in America. University of Wisconsin Press, Madison.

Nicholls, D., 1975. The Pluralist State. St. Martin's Press, New York.

Nisbet, R.A., 1975. Twilight of Authority. Oxford University Press, New York.

Odegard, P., 1928. Pressure Politics: The Story of the Anti-Saloon League. Columbia University Press, New York.

Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Polsby, N.W., 1980. Community Power and Political Theory, second ed. Yale University Press, New Haven, CT.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.

Rosenblum, N.L., 1998. Membership and Morals: The Personal Uses of Pluralism in America. Princeton University Press, Princeton, NJ.

Runciman, D., 1997. Pluralism and the Personality of the State. Cambridge University Press, Cambridge, UK.

Runciman, D., Ryan, M. (Eds.), 2003. Maitland: State, Trust and Corporation. Cambridge University Press, Cambridge, UK.

Schattschneider, E.E., 1960. The Semisovereign People. Holt, Rinehart and Winston, New York.

Schlozman, K.L., Verba, S., Brady, H.A., 2012. The Unheavenly Chorus: Unequal Political Voice and the Broken Promise of American Democracy. Princeton University Press, Princeton, NJ.

Shklar, J.N., 1989. The liberalism of fear. In: Rosenblum, N. (Ed.), Liberalism and the Moral Life. Harvard University Press, Cambridge, MA.

Smith, M.J., 1990. Pluralism, reformed pluralism and neopluralism: the role of pressure groups in policy-making. Political Studies 38, 302–322.

Stears, M., 2002. Progressives, Pluralists, and the Problems of the State. Oxford University Press, Oxford.

Truman, D., 1971[1951]. The Governmental Process: Political Interests and Public Opinion, second ed. Knopf, New York.

Walker Jr., J.L., 1991. Mobilizing Interest Groups in America: Patrons, Professions, and Social Movements. University of Michigan Press, Ann Arbor, MI.

Warren, M.E., 2001. Democracy and Association. Princeton University Press, Princeton, NJ.

Pluralism and Tolerance

Otfried Höffe, Universität Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is concerned with one of the most important characteristics of modern society – pluralism, and with the concept, which gives an answer to the various challenges pluralism poses – tolerance. After tracing the idea of pluralism, I focus on the various implications the fact of pluralism has for contemporary liberal societies. I then examine the concept of tolerance both historically and systematically, distinguishing between active and passive tolerance as well as between personal, political, and social tolerance.

One of the most important characteristics of modern society is its pluralism. Because of the challenges and problems pluralism raises, tolerance is needed as a counterpart. Though the idea of tolerance has a long and complex history, today it can primarily be understood as an answer to the challenges of pluralism. As pluralism has different aspects, the same is true for tolerance. Whereas pluralism emphasizes the diversity of equally entitled, free groups, tolerance stresses the 'live and let live' of others' liberty.

Though pluralism, and tolerance as its counterpart, are important concepts to explain modern society, they are one-sided concepts, which have no absolute value, but only a relative one, since beyond the plurality of values and life-forms, something has to be shared by all human members of society in order to prevent its dissolution. Pluralism and, correspondingly, tolerance cannot justify themselves – in both cases, the more fundamental ideas of freedom and justice are necessary for their justification.

The Idea of Pluralism

Pluralism can be understood in a metaphysical sense as designating those philosophies that allow for a diversity of entities in their explanations of the world, as did the Enlightenment-philosopher Christian Wolff (1720). Kant contrasts egoism with pluralism, which he understands to be "the way of thinking in which one is not concerned with oneself as the whole world, but rather regards and conducts oneself as a mere citizen of the world" (1798/2006: p. 18).

The pragmatist philosopher William James maintained the existence of an infinite number of limited facts, connections, and systems that are in constant flux; which always lead to new relationships and structures; and for the cognition of which, fixed categories and criteria do not exist, only perspective orientation points (James, 1909). Pluralism can also refer to the methodological neutrality of the social sciences of an empiristic orientation, to the multiplicity of cultural forms, from which "all sociology has its starting point" (Scheler, 1980: p. 25). Today, the term is primarily understood in a social and political sense in which empirical and normative elements converge.

Empirical and Normative Elements

Empirically speaking, pluralism denotes the diversity not only of confessions and religions (religious pluralism), but also of social groups together with their lifestyles (social pluralism), of basic values (pluralism of value), and of the powers, which determine politics (political pluralism). For this many-dimensional diversity, our lifeworld has primarily benefited from the religious wars of the Reformation, the European Enlightenment, and the Bourgeois and Industrial Revolutions. Through those factors, a process of differentiation was set into motion, which dissolved both the relatively homogenous and stable relationships of the so-called old-European societies and the connection between church and state. Moreover, this process brought forth open societies, the numerous and constantly expanding groups of which had their own interests, convictions, and manners of conduct. More recently, pluralism has gained a new momentum through the process of globalization (Höffe, 1999/2007), in particular also through migration movements, many refugees, and those looking for asylum.

Second, pluralism purports – and here one finds its normative content – that diversity and difference is acknowledged and approved of; all functionally different groups are equally entitled to develop themselves in a free and unconstrained manner.

The notion of pluralism is of particular relevance in modern democracy theory (Fraenkel, 1979; Oberreuter, 1980). It is considered here as a critical notion, namely as that central mediative category between the individual and the state, with which the liberal atomization of individuals just as the omnipotence and an all-encompassing capacity of the state are rejected. In a democracy, the empirical thesis of pluralism claims, many-faceted powers are at work, which, on the basis of a fundamental consensus over institutionalized arbitration with and against one another, struggle for influence and importance. As far as the normative content of pluralism is concerned, various social forces, notwithstanding their differences in content and function, have the same right to engage themselves in the political process by the means of, e.g., voting, of contributions to public deliberation, or by addressing themselves to members of parliament. These forces include trade groups, unions, and employers' associations, scientific organizations, churches, and political parties, and even civil action groups. Politics is not derived from the conception of a common good which is clearly recognizable, prior to all political confrontations and which does justice to all parties concerned. Political decisions are a result of the constant dispute between competing interest groups which have to mutually monitor one another and, in constantly new coalitions, 'scuffle' together into a compromise which might be accepted by a majority.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63061-1

The Critique of Pluralism

The criticism of the idea of pluralism is leveled partly at its empirical thesis and partly at its normative content; the normative criticism pertains to either pluralism of value or political pluralism.

Pluralism of value becomes visible in the form of the ideologically neutral state. It is reproached with relativism and even nihilism because it does not even acknowledge certain very general values as binding and is held responsible, therefore, for the crises of meaning and orientation, which threaten contemporary industrialized societies. Pushed to the fore, the ideologically neutral state will abolish itself, because it cannot legitimize the framework in the limits of which various powers are capable of freely expressing themselves.

The most significant criticism of *political pluralism* stems from Carl Schmitt (1927) and his followers, who fear the centrifugal, 'anarchistic' tendencies of pluralism. Through the openness and insecurity of pluralistic democracy, the state's power of decision making is left crippled; moreover, the compromises resulting from such a situation correspond to neither moral nor economical ideals of sound politics.

Supporters of the critical theory hold pluralism for desirable, but contest the existence of the empirical presuppositions for its realization. They point to the socially underprivileged, who are either effectively shut out of participating in the bargaining process of weighing competing interests, or are, at least, clearly at a disadvantage when up against more powerful organizations. The normative idea of an equal right for all groups is, therefore, rather a postulate than social reality. Other critics maintain a 'standardization of subjects', who are no longer capable of an autonomous definition of their interests and, hence, of a genuine pluralistic differentiation. Where a free concert of interest groups once existed, one now finds bureaucratic apparatuses in their place, the internal dynamics of which dominates the political scene.

Pluralism and Liberal Democracy

Political pluralism presupposes two conditions as given. On the one hand, all interests and the groups that represent them should have equal rights in public deliberation and in the process of political will-formation; on the other hand, the common good should realize itself through the struggle of interests between groups of equal rights. Even in developed democracies, both conditions are fulfilled at best in approximation. For in pluralistic society, citizens with their interests and convictions have no direct access to politics; their access stands in need of mediation through clubs, organizations, and political parties. In this manner, interests and convictions do indeed gain political power, however at different levels of intensity. The mediation of interests and convictions very often leads to perceptible distortions of their influence.

Due to varying historical or social conditions, many interests and convictions are represented more powerfully than others. The explanation of this fact might be seen in the tighter organization of certain associations, to the better support by more potent moneylenders, or to the fact that such groups have attractive media at their disposal, or better relations to the mass media. An equality of rights in modern times is not realized through the abolishment of legal privileges, characterizing hierarchical feudal or class societies. Pluralism does not simply describe our social and political reality. It is to a significant degree really a postulate, according to which societies need to be changed and modified.

A modification within pluralism, however, is possible only to a limited extent: first of all, because powerful groups do not freely relinquish their predominance; second, as can be observed sociologically (Olson, 1965), many interests are much easier to organize than others, which may lead to the superiority of certain economic groups (employers as much as employees). Furthermore, mergings and the oligarchical tendencies within organizations endanger civic freedom, the essence of pluralism. Dependence on the government gives way to dependence on both the bureaucracy of associations and its functionaries. Finally, even a fair arrangement between a plurality of interest groups does not simply lead to the realization of the common good; for, as a rule, long-term interests get the worst of short-term interests, just as general interests get the worst of particular interests. It is for this reason that the protection of the environment is so easy to push aside.

In taking an overview of these arguments, one must neither adhere to a conspiracy theory, nor to an extreme criticism of capitalism, in order to preserve a residue of skepticism with regard to the affirmation of an empirically existing pluralism. Hence, the plurality of intermediary groups is without a doubt a pillar of freedom; a pluralistic society is invariably better than a nonpluralistic one; and pluralism itself an important structural element of the liberal–constitutional democracy. Nevertheless, pluralism alone cannot warrant its own significance; it is only a helpful not in any way sufficient organizational form of freedom and equality. One should not easily identify pluralism and liberal democracy because the notion of pluralism brings to light only diversity and competition, and not the complementary elements of commonality and cooperation.

A liberal democracy consists not only of the coexistence and confrontation of a diversity, but also of the regulation of coexistence and confrontation. Regulation brings about community: the mutuality of problems to be solved and, in particular, that fundamental (constitutional) consensus, which defines the normative content and the rules of pluralism. In this manner, pluralism shows itself to be a one-sided, nondialectical notion that stresses plurality instead of unity, competition instead of cooperation.

A Relative Pluralism

Because empirical and normative elements converge in the notion of pluralism, the normative valuation has already begun in the empirical thesis. It is continued with a *pragmatic* consideration: the decentralizing of political power, the institutional and social separation of the process of will formation, just as the diversity of convictions and manners of conduct increase the chance of an efficient coexistence, in which superfluous frustrations and conflicts decrease.

The comprehensive moral evaluation begins with an ethical–anthropological insight. Because of its lack of instinct and a wanting specialization of organs, mankind is not fixed to definite manners of conduct. Due to its capacity for language

and reason, mankind is much more reflective in relation to its doings and omissions; this not only enables it to act freely and responsibly, but also constitutes a particular challenge. Therefore, mankind, be it as individual, a small or a large group, has the task and the right to develop its own convictions and practices, and to determine and pursue a particular conception of a good life. Different convictions and practices are to be expected, on account of the historical factors under which this process of personal as well as social self-determination takes place. The form as well as the results of self-determination are limited by different perspectives, and a single individual cannot lay claim to the unconditional 'truth' for himself (Rawls, 1993: p. 36). In the pluralism of interests, of values, and of life plans, the wealth of human self-realization is expressed just as well as the limitation of every concrete design.

On the other hand, different groups are already related to each other through the social division of labor; but the division of labor represents only one of the many aspects that make mediation necessary. Within one's own group, the recognition of interests and convictions is commonplace. However, with regard to other groups, these interests and convictions stand in both competition and conflict. Because the solution of conflicts is not secured by instincts, there exists the danger of collective self-destruction of mankind. In order to check this danger and, at the same time, to avoid a crude sort of social Darwinism that makes the weak even weaker and the vital and cunning even stronger, general obligations are necessary, which guarantee a humane and just management of conflicts. For that reason, not an absolute and merely a relative pluralism that develops in the context of fundamental commonalities is really desirable.

Pluralism of this sort can encounter the suspicion of being relativistic and nihilistic. Even if modern societies are skeptical of a timelessly valid code of values, even if they do not confer infallibility upon a particular personal, political, religious, or intellectual instance, they nonetheless need at least a few generally valid obligations removed from the opposition of particular designs of faith and existence. Therefore, human rights are accrued a particular meaning: as the personal right to liberty, the political right to participation, and as social and cultural rights, they are the condition and expression of the mutual recognition of fundamentally equal persons. Therefore, human rights refer to those claims which every individual may legitimately raise: they belong to a sphere beyond a legitimate pluralism and which every legal order, every state, especially pluralistic democracy owes to each individual. Through human rights, even those minorities that come off badly in the pluralistic struggle of competition are protected (Rawls, 1971; Höffe, 1987/1995).

The Idea of Tolerance

Originally, tolerance does not refer to tolerating somebody, but something. As a passive form of bravery, tolerance originally means the ability to suffer in the face of negative experiences: a patient enduring of bad things such as pain, torture, or strokes of fate. It is only later, under the conditions of pluralism that the meaning of the concept changes and now applies to tolerating the stranger, especially the stranger adhering to other religions and confessions. In this way, an attitude of how to live one's only life is transformed into a moral attitude how to treat others. The original concept of tolerance as one of individual ethics is transformed into a virtue of social ethics, a virtue of individuals (personal tolerance), communities (political tolerance), or cultures and societies (social tolerance).

Ever since the concept of tolerance changed into meaning the endurance of otherness, it is inevitably linked with pluralism as well as relativism. If all human beings belonged to the same religion, no religious tolerance would be needed, and if everybody lived the same way, no social tolerance would be required. It would also not be required in a society in which everything is totally arbitrary: if every conviction and lifestyle counts the same and if they all are equally valid, tolerance loses its area of application. Tolerating the others not only presupposes otherness; one's own otherness needs to be important to oneself, and the other's otherness thus objectionable. A person is not tolerant if he or she simply takes recourse to an attitude of detached indifference. Rather, the tolerant person actively tolerates persons with other convictions and another lifestyle. Tolerant is he or she who, contrary to mere indifference, actually experiences impositions, but, contrary to intolerance, endures them.

Its Origin in Religious Tolerance

Like the general principle of liberty, the particular principle of tolerance found its general recognition only after long intellectual and sociopolitical debate. In this context, religious tolerance plays a particular role (Püttner, 1977).

Many people consider tolerance to be an invention of modernity. A central principle, however, was developed already one and a half millennium before modernity. In the ancient Orient, religion, society, and the state were closely intertwined. In contrast to that, the demand from the New Testament "Render to Caesar the things that are Caesar's, and to God the things that are God's" (*Matt.* 22: 21) points toward a certain separation. The *Edict of Milan*, inspired by this demand, already in 313 AD lays down a corresponding right: Every person may believe what he or she wants. Because of that, earthly punishments for religious offenses are to be prohibited.

Since polytheistic religions make a lesser claim to truth than monotheistic ones, they do not face many difficulties as far as the toleration of diverse religious beliefs is concerned. Pre-Christian Rome allows all conquered peoples to practice and even propagate their own cults. Yet the Christians were excluded from the rule of tolerance, but not for religious reasons; from their refusal to observe the transmitted gods, one fears a backlash within the entire commonwealth.

As long as the Christians are persecuted, they appeal to the Roman Emperors, referring to the rights which all mankind are due: that freedom of religion is granted internally and externally. But after Christianity is promoted to the status of a state religion, it soon forgets this demand and its underlying principles contained in the New Testament: even the principle of reciprocity (*Matt.* 7: 12; also *Tob.* 4: 16); the sermon on the mountain; the parable of the weed under the wheat (*Matt.* 13: 24–30, 36–43); Jesus' conduct in general of welcoming, but not of forcing his imitation; moreover his magnanimous attitude toward sinners; finally, the apostle Paul's conception of freedom of religion as free from enforcement and his

admonition to mutual tolerance (*Col.* 3: 12f.; *Cor.* 1: 4,12; *Cor.* 2: 11.1; and *Gal.* 5: 1.13).

Whereas the *Edict of Milan* leaves religious belief to its practitioners and, moreover, differentiates between state and religion, the swift rise of schisms and heresies fills the bishops with such alarm, even in the patristic times, that they begin to bind, as in the ancient Orient, religion closely with politics. Especially the conversion to a foreign cult is punished to the same extent as capital crimes and the disruption of civil order.

Granted, medieval Christendom should not to be considered intolerant through and through (Patschovsky, 1998). Because it regards the transmittance of faith as a matter of free conscience, the conversion of pagans by force represents a clear exception. Jews are generally and, often, truly tolerated; and for a long time the Christians live in mutual tolerance with the Muslims. But, the heretic in Christendom, accused of both arrogance and impertinence, was considered, like the Islamic apostate, as it were, an outlaw in the eyes of the law. In the eleventh and twelfth centuries, capital punishment extends to the heretic and is maintained until well after the Reformation. The foundation of medieval intolerance, the all too intimate fusion of *Sacerdotium* and *Imperium*, of spiritual and secular matters are all well preserved in Protestant countries, more likely being advantageous to the state than to the church.

In the 'confessional age' (the age of the Reformation), the idea of (religious) tolerance has to be reestablished; and it is effected only after hefty intellectual debate and painful political experience. In order to encourage the victory of tolerance in the form of the religious neutrality of the state, the mutual autonomy of the state and the church, a few fundamental insights first have to be (re)attained. At first, the pragmatic consideration of the 'greater benefit', the fact that in most cases mildness is better suited than force in order to gain back 'lost souls'. Then one also has to reach the insight that the New Testament is based on tolerance and love, moreover, that not only the conversion but also the retention of faith is both an act of freedom and of mercy, and for that reason, cannot be enforced; furthermore, that the church is a free communion, whose task is to be fulfilled in the spiritual, not in the secular realm, and that for this task every possible constraint is useless, as it would at best lead to a feigned and not to a real recognition of faith.

The way for this new form of tolerance is prepared by Christian humanists like Dante, Marsilius of Padua, Nicolaus von Kues, and Sebastian Castellio, who reject all acts of violence against religious dissidents. However, they are still fixed to ideas of religious unity within Christendom, and, therefore, they search for a peaceful reconciliation by way of religious forum and dogmatic compromise between rivaling confessions. On the contrary, mystic spiritualists as well as some religious cults engage themselves directly for the religious liberty of each individual; they do this by appealing to the Holy Ghost, who forbids every form of slavery in the realm of the religious, including that of the institutional church.

Whereas no lasting success is allotted to either of these thoughts (peaceful reconciliation or religious individualism), the second half of the sixteenth century sees the problem shift from the religious to the 'secularized' level. Theological arguments are succeeded by considerations of politics and natural right, occasionally even of economics; unsuccessful reconciliatory talks give way to efforts of toleration, in which the denominational heterodox gains ground.

Starting in France, where the preservation of the nation is considered more important than the reestablishment of religious unity, the idea of civil tolerance, which unambiguously differentiates between the aims of the state and the aims of religion, is established against the rule of 'one confession, one law, one king'. With the idea of the confessional, later the religious neutrality of the state, the bloody experiences of confessional civil wars that, at the European level, ended in a stalemate, are finally digested. Consequentially, the relative autonomy of the state and church/religion and, with that, of confessional pluralism, is acknowledged as a necessary condition for political, even physical survival. Nevertheless, the numerous edicts of tolerance, which follow on the Peace of Augsburg (1555) and the Edict of Nantes (1598), are issued only reluctantly. In auspicious circumstances, as in France (1685), the edicts are repealed or, as in Great Britain, where the Catholics remain second-class citizens until the nineteenth century, they are not valid for everyone from the outset.

Because intolerance endangers the free development of business and trade, Europe and North America also have economic reasons to thank for the following two phenomena: that religious liberty does indeed assert itself in the end, and that the remnants of the medieval connection between the state and the church are vanquished. The European Enlightenment, however, plays a particular role in the ultimate victory of the principle of religious tolerance. Inspired by the idea of universal human reason and liberty, such figures as Grotius and Spinoza, Milton, Locke and Voltaire, Lessing (*Nathan the Sage*), later also Schiller (*Don Carlos*) turn to the remnants of dogmatic and political intolerance within the high church and to their corresponding, amalgamated states; moreover, they lay claim to religious liberty as a fundamental personal right to liberty. But, even those propagating the Enlightenment as a rule exclude atheists and nonbelievers from the idea of tolerance, granted, as with Locke, not out of religious but rather out of political reasons; for, whoever does not believe in God cannot take a legally admissible oath.

That religious freedom is established in the end is just in part due to a philosophical insight. The primary reasons for this establishment are to be found in the bloody religious and civil wars, as well as in economic and scientific developments – because intolerance threatens the free flourishing of trade and industry, as well as that of science and art. Tolerant states such as Brandenburg-Prussia and the Netherlands consequently flourish economically as well as culturally. Thus, already a well-considered self-interest calls for tolerance.

The justification from human dignity, however, is normatively more fundamental. This justification declares every human to be a free and equal person, supplied with the right to develop and pursue own convictions and beliefs. Recognition of corresponding rights for every other human being is, naturally, part of this principle. Tolerance ends where freedom and dignity of others are infringed.

The justification from freedom and dignity of every human being thus entails both: a measure for tolerance as well as a criterion for its limits. By making the demand that nobody should be allowed to a liberty, which cannot be allowed to every other person, as well, tolerance refuses to submit itself to

injustice. In cases where freedom and human dignity are threatened, tolerance transforms itself into critique, accusation, and protest. Paradoxically speaking, tolerance meets intolerance with intolerance.

Passive and Active Tolerance

The idea of tolerance kindled itself on the freedom of religion in Western Christianity. In our advanced, secularized world this idea no longer represents a special problem with regard to the social interaction of diverse confessions and religions. Tolerance is often enough primarily called for in relation to beliefs, which are socially, politically, and culturally differentiated. Thereby, a weaker, more passive form of tolerance, the mere toleration of otherness, is to be differentiated from a stronger, more active, and creative form, the openness to differences, yes, even to their candid recognition.

As pluralistic democracy cannot arise nor continue to subsist without the mutual toleration of otherness, tolerance is, at least in the weaker form, a fundamental principle of modern states (for developments see Lecler, 1955; Lutz, 1977; Becker, 1996). However, tolerance is not only necessary under the condition and in favor of a political pluralism. If not always immediately and distinctly perceptible, there are nonetheless differences among mankind: in need, talent, and interest; in taste and in sociohistorical origin; moreover, in the perception and interpretation of the personal or social situation in which one lives. Nor is one free of faults, prejudices, or mistakes. Therefore, at least passive tolerance is held to be a universally valid posture with regard to the civilized intercourse of mankind; it is through just such relations that the coexistence in marriage and family, in industry, in society, in the government, even between states is made bearable as well as efficient. So, even an enlightened self-interest speaks in favor of tolerance.

However, the core dimension of tolerance is, with that aspect, not yet achieved. Its stronger form, the candid recognition of the other and otherness, is grounded in the dignity and the liberty of every human being. Therefore, tolerance consists not only of that forbearance of the peculiarities and weaknesses of one's fellow man, which can be solicited from them and freely granted to them. As the coexistence of individually responsible persons with equal rights is not possible without reciprocal recognition, tolerance is a fundamental condition for the justice of mankind's coexistence in freedom; and, as a condition of justice, it is a claim that can be refused to no one.

For active tolerance, not merely the practical (moral–political) principle of liberty but also a theoretical interest in truth be adduced. Human perception being limited, the open debate of different opinions offers a better chance for establishing truth than the 'intolerant', namely, dogmatic and of a conviction considered infallible (Mill, 1989).

Tolerance, in any case, appears in two steps. The elementary step, which is in fact only a prestep, is passive tolerance. Here, tolerance is merely enduring the others, and often is combined with contempt. The second step, however, is where true, active, and authentic tolerance resides: Active tolerance is not content with merely accepting others unwillingly. This kind of tolerance rather overcomes mere acceptance and voluntarily and freely affirms the other's values and lifestyles, as well as their freedom and liberties. Thus, active tolerance is no fig leaf behind which moral indifference lurks: It is founded in the recognition of one's own value, one's self-esteem and self-respect.

Personal, Political, and Social Tolerance

Religious liberty is only one example that brings to the fore the specific intellectual and political difficulties that have to be overcome in order for the principle of tolerance to become historically significant. This principle, which goes beyond religious liberty, takes primarily three different forms: tolerance as (1) a personal attitude, (2) as a political principle of the state, and (3) as a principle of a society or culture. All three forms mutually supplementing and yielding one another.

Whoever affirms the self-determination (the right to life, the will to development, and the liberty) of the dissentient is realizing the individual attitude of tolerance: personal tolerance. Only when one has one's own convictions, can one speak of recognizing an 'other'; and only there, where one does not indulge in indifference, or in the type of cynical nihilism that plainly lets everything pass as valid, including even gross injustice, can one speak of this particular attitude. Tolerance does not mean the fig leaf, behind which moral indifference and intellectual weaknesses conceal themselves. It is a critical autonomy which does not even allow certain conflicts to arise in coexistence, especially not in situations of competition; and which does not go through the inevitable debates and quarrels with the inexorable severity of a religious war. Whoever is tolerant no longer pursues a life, which is founded upon self-assertion through violent conversion, or through the surmounting of adversaries. He or she endeavors to establish a communality on the basis of civil equality and communication to which the following qualities belong: the ability to listen, the capacity to take others seriously, and finally the willingness to let oneself be taught by new situations and new information. Tolerance is then accomplished by the willingness and ability to sympathize with the views and lifestyles of others.

Like pluralism, tolerance too is grounded in the following factors: the understanding that no one is entirely free from faults; the knowledge of the wealth as well as the perspectival limitations of every concrete instance of self-realization; and the insight into the limits of human knowledge. Tolerance, however, originates primarily in the recognition of the other as a free and equal person, who, on the basis of the inviolable dignity of mankind, has the right to form his or her own opinions, and to live according to them alone, or with like-minded individuals – provided that the exercise of this right does not encroach upon the same right of all others. The virtue of tolerance is one of the most important political virtues of a democratic citizen (Rawls, 1993: p. 194).

As tolerance does not represent an end in itself, but rather a condition and an expression of the liberty and dignity of mankind, it terminates at the point where liberty and dignity are violated. Tolerance does not consist, as Marcuse (1965) fears, of the stabilization of existing social states, in the preservation of the status quo, and even less of the (masochistic)

acceptance of injustice. Wherever the liberty and dignity of mankind needs to be protected, tolerance crosses over into accusation, criticism, and protest. The obligation to impose on no one a limitation of his or her liberty, which is not imposed on all others in the same manner, can be considered a criterion for the limits of tolerance. The highest principle of justice consists of this criterion of an equal liberty, which is primarily spelled out in terms of human rights.

Because the justification of tolerance is found in liberty and justice, it cannot be left to the discretion of individuals and political powers. Its recognition is not only a voluntarily assumed attitude, but also a duty of justice, the implementation of which the legal order, the state, is responsible. It is therefore a political task to institutionally guarantee basic elements of tolerance, such as human rights. The first task of tolerance as a principle of the legal order and the state is therefore found in the positivation of human rights and their recognition as basic rights (personal right to liberty, political right to participation, and social and cultural rights), or, as the case may be, as fundamental aims of the state (liberty, democracy, and welfare state).

As long as human rights are only recognized by coercion, tolerance remains nothing more than an acceptance accompanied by contemption toward those who think and live differently. In order to guarantee the stronger form of tolerance – coexistence based on mutual respect – tolerance as a duty of justice has to be acknowledged on its own terms; moreover, it must become through continuous practice a component of one's personal character, a genuine attitude. Tolerance as a political principle is only accomplished when this duty of justice is freely accepted as a moral virtue of the citizen.

Communities are tolerant in a weaker and passive sense if they simply endure minorities. They are, however, tolerant in an active sense if they affirm and grant public legal security to these minorities – as it is custom within liberal democracies. For a society to be tolerant, more needs to be the case. A social tolerance that reaches beyond the political level refrains from any pressure of conformity; it even allows for eccentric lifestyles provided they are nonviolent and peaceful.

At the international level, tolerance demands the mutual respect of different cultures and traditions, instead of valuing European culture or American lifestyle as the ideal of the development of all civilization and to devalue all other forms as primitive and barbarian. This type of respect was articulated first in Herder's (1774/2004) and later in Hegel's (1821/1991) theory of the 'spirits of nations' (*Volksgeister*), according to which different social and cultural patterns in which different peoples think and live assert themselves in the course of history. As these forms are coherent, they justify themselves to this extent.

One facilitates the recognition of other cultures' intrinsic value whenever one looks not only to that which divides but also to that which binds, that is, universal human attitudes such as readiness to aid, civil courage, and honesty. These commonalities are not only acknowledged as morally binding in the Christian West, but also in the most diverse of societies and epochs. The tolerance that began with religious liberty is accomplished only at the point, where the colorful abundance of personal and sociohistorical self-realization is freely acknowledged as both a task and a chance for creative humanity.

A Classification

Making use of the concepts of tolerance discussed so far, we can establish a classification of cultures and religions. Needless to say that within this classification, cultures are not measured by their cultural quality and religions are not measured by their religious quality. Instead, this classification uses the willingness to intercultural coexistence as a measure, which arises from an attitude of tolerance.

(1) At one extreme, there are those cultures that are as good as *helpless* with regard to modern tendencies and have no means to hold on to their distinctiveness – e.g., inhabitants of the rain forests and other indigenous people. If these cultures are not left to their own, they will be destroyed. (2) Other cultures like the Mennonites and the Amish people are *defensive*. Content with themselves, they are able to isolate themselves from others, provided that they are left alone. (3) *Reciprocal* cultures exist within a reciprocal give-and-take. (4) *Peacefully expansive* cultures give far more to others that what they adopt from them. They do not use any force or coercion, but are content with the attitude of a peaceful suggestion. (5) A *resentment–aggression* does not force itself upon others, but it isolates itself with violent means. Frequently, such a culture reacts with hatred toward the other; inner violence transforms itself into outer violence. (6) At the other extreme, finally, there *are aggressive–expansive* cultures, which force their elements upon others – not only religion, but also their language, their system of law and their custom, as well as their corporate culture.

See also: Democracy: Normative Theory; Multiculturalism; Relativism: Philosophical Aspects; Value Pluralism.

Bibliography

Becker, W., 1996. Anerkennung und Toleranz. Über die politischen Tugenden der Demokratie. Palm & Enke, Erlangen.

Forst, R., 2003/2012. Toleration in Conflict. Past and Present (C. Cronin, Trans.). Cambridge University Press, Cambridge. Orig: Toleranz im Konflikt. Suhrkamp, Frankfurt/M.

Fraenkel, E., 1979. Deutschland und die westlichen Demokratien, seventh ed. Kohlhammer, Stuttgart.

Hegel, G.W.F., 1821/1991. Elements of the Philosophy of Right (H.B. Nisbet, Trans.). vol. 7. Cambridge, New York. Orig: Grundlinien der Philosophie des Rechts, Werke in 20 Bänden. Suhrkamp, Frankfurt/M.

Herder, J.G., 1774/2004. Another Philosophy of History and Selected Political Writings (I.D. Evrigenis and D. Pellerin, Trans.). vol. 5. Indianapolis/IN. Orig: Auch eine Philosophie der Geschichte zur Bildung der Menschheit, Sämtliche Werke, B. Suphan (Ed.). Weidmann, Berlin.

Heyd, D. (Ed.), 1996. Toleration, an Elusive Virtue. Princeton University Press, Princeton/NJ.

Hill, T.E., 2000. Respect, Pluralism and Justice. Kantian Perspectives. Oxford University Press, Oxford.

Höffe, O., 1987/1995. Political Justice. Foundations for a Critical Philosophy of Law and the State (J.C. Cohen, Trans.). Polity Press, Cambridge. Orig: Politische Gerechtigkeit. Grundlegung einer kritischen Philosophie von Recht und Staat. Suhrkamp, Frankfurt/M.

Höffe, O., 1999/2007. Democracy in an Age of Globalization (D. Haubrich and M. Ludwig, Trans.). Springer, Dordrecht. Orig: Demokratie im Zeitalter der Globalisierung. Beck, Munich.

Horton, J., Mendus, S. (Eds.), 1991. John Locke: A Letter Concerning Toleration (1689). Routledge, London.

James, W., 1909. A Pluralistic Universe. Longmans, New York.

Kant, I., 1798/2006. Anthropology from a Pragmatical Point of View (R.B. Louden, Trans.). vol. 7. Cambridge University Press, Cambridge, pp. 117–334.

Orig: Anthropologie in Pragmatischer Hinsicht, Gesammelte Schriften. Königlich Preußische Akademie der Wissenschaften (Ed.). Berlin.
Kelso, W.A., 1978. American Democratic Theory. Pluralism and Its Critics. Greenwood Press, Westport, CT.
Lecler, J., 1955. Histoire de la tolérance au siècle de la réforme, 2 vols. Aubier, Paris.
Lutz, H. (Ed.), 1977. Zur Geschichte der Toleranz und Religionsfreiheit. Wissenschaftliche Buchgesellschaft, Darmstadt.
Marcuse, H., 1965. Repressive tolerance. In: Wolff, R.P., Marcuse, H., Moore, B. (Eds.), A Critique of Pure Tolerance. Beacon Press, Boston, MA.
Mendus, S. (Ed.), 1988. Justifying Toleration. Conceptual and Historical Perspectives. Cambridge University Press, Cambridge, UK.
Mill, J.S., 1989. On Liberty (1859). Cambridge University Press, Cambridge, UK.
Newey, G., 1999. Virtue, Reason and Toleration. The Place of Toleration in Ethical and Political Philosophy. Edinburgh University Press, Edinburgh.
Oberreuter, H. (Ed.), 1980. Pluralismus. Grundlegung und Diskussion. Leske und Budrich, Opladen.
Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Patschovsky, A. (Ed.), 1998. Toleranz im Mittelalter. J. Thorbecke, Sigmaringen.
Püttner, G., 1977. Toleranz als Verfassungsprinzip. Prolegomena zu einer rechtlichen Theorie des pluralistischen Staates. Duncker & Humblot, Berlin.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Raz, J., 1986. The Morality of Freedom. Oxford University Press, Oxford.
Scanlon, T.M., 2003. The Difficulty of Tolerance. Essays in Political Philosophy. Cambridge University Press, Cambridge/New York.
Scheler, M., 1980. Die Wissensformen und die Gesellschaft (1926), third ed. Francke, Bern.
Schmitt, C., 1927/2006. The Concept of the Political (G.D. Schwab, Trans.), second expanded edition. University of Chicago Press, Chicago. Orig: Der Begriff des Politischen. Duncker & Humblot, Berlin (1963).
Schnapper, D., 1998. La rélation à l'autre. Au coeur de la pensée sociologique. Gallimard, Paris.
Schultz, U. (Ed.), 1974. Toleranz. Die Krise der demokratischen Tugend. Rowohlt, Reinbek.
Seurin, J.L. (Ed.), 1981. La démocratie pluraliste. Économica, Paris.
Thierrey, P., 1997. La tolérance. Société démocratique, opinions, vertus. Presses Universitaires de France, Paris.
Walzer, M., 1997. On Toleration. Yale University Press, New Haven/CT.
Wolff, C., 1720. In: École, J. (Ed.), Vernünftige Gedanken von Gott, der Welt und der Seele des Menschen, auch allen Dingen überhaupt. Gesammelte Werke. Hildesheim.
Wong, D.B., 2006. Natural Moralities. A Defence of Pluralism. Oxford University Press, Oxford.

Polanyi, Karl (1886–1964)

György Lengyel, Corvinus University of Budapest, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Karl Paul Polanyi was born in Vienna in 1886, received his doctoral degree in Budapest in 1912, and ended his career at Columbia University lecturing in economic history and coordinating research into the origin of economic institutions. The major question of Polanyi's substantive view is how to identify the place of economies in societies. Embeddedness, according to Polanyi, is related to the fact that an economic actor is a social being, with a mixture of motives, serving to attain social recognition and social goods. Economy can therefore be interpreted as processes of production, transport, transactions, and disposition. Institutions lend unity and stability to these processes, which can be studied via the forms of integration: reciprocity, redistribution, and exchange. Markets are integrated into a dominant system only in exceptional historical cases. Substantivism became a chapter in the history of sociology and anthropology; Polanyi's embeddedness concept is a point of reference in economic sociology. Polanyi also influenced research on state and economy as well as socialist and postsocialist studies.

Biographical Notes

Karl Paul Polanyi was born in Vienna on 25 October 1886 to Michael Pollacsek and Cecile Wohl. They had married in 1881 and had six children: Laura, Adolf, Karl, Michael, Zsófia, and Paul. Karl became a recognized scholar in sociology and anthropology while Michael in physical chemistry and social sciences. Polanyi's father worked for the Swiss Railways and Vienna's Stadtbahn, before setting up on his own as a railway builder in Hungary. In the 1890s, the family moved to Budapest where Karl Polanyi began his secondary studies in the model school of the teachers' training institute in 1898. At the turn of the century, Polanyi's father's business went bankrupt and after indemnifying the creditors, the family was impoverished and the father was forced to take a job abroad again. In spite of that, there was lively intellectual activity in the parental home in Budapest. Polanyi's mother, Cecile Wohl, organized a literary salon attended, among others, by Oszkár Jászi, György Lukács, and Karl Mannheim. After finishing secondary school in June 1904, Polanyi enrolled in the Faculty of Law in Budapest University. He lost his maternal grandfather and father in 1905, both of whom had especially deeply inspired him. After attending courses in Vienna and Budapest, he took part in organizing demonstrations for the progressive thinker Gyula Pikler, a professor of the Philosophy of Law in 1907. Polanyi obtained his Doctoral degree on 26 June 1909 from Bódog Somló at the Kolozsvár (Cluj) University. In 1908 he became president of the recently formed Galilei Circle, a student body set up to disseminate the scientific world view in readings and debates, without – as Polanyi later self-critically recalled – being politically involved. He qualified as a lawyer in 1912 and began to co-edit the review *Szabadgondolat* (*Free Thought*) in 1913. His writings argued for free education and were devoid of the influence of the church. In 1914, Polanyi was involved in establishing the Radical Party, of which he became the secretary.

However, the next year he was sent to the front and discharged 2 years later as disabled. During the liberal Károlyi government of 1918, he was a member of the legal committee of the People's Council in Budapest and of the Students' Council. However, on the eve of the Hungarian Soviet Republic in 1919 he delivered a speech against the dictatorship, and then migrated to Vienna. He soon underwent a serious operation followed by a lengthy recovery. He was working on a voluminous unpublished work, *Behemoth*, concerned with the destructive capacity of science. In this he says that the future is constantly being remade, therefore it cannot give validity to our actions in the present.

In addition to Ludwig von Mises, Jakob Marschak, Felix Weil, and others, he contributed to the theoretical debate about the possibility of the planned economy in which he elaborated a proposal to solve the problem of 'Sozialistische Rechnungslegung' on the basis of the principles of guild socialism. He met a well-known figure of the revolutionary students' movement, Ilona Duczynska, in 1920. They married and had a child, Karoline Helene (Kari) in 1923. In Vienna, Polanyi first worked for Hungarian papers (Bécsi Magyar Ujság and Bécsi Kurir), and then from 1924 to 1933 he contributed to the Österreichische Volkswirt. Having lost his job because of the crisis and the rising danger of National Socialism, he emigrated to England where he was one of the masterminds behind the Christian Left movement, and concentrated on criticism of the essence of fascism.

This time he joined the Workers' Educational Association and the Extra-Mural Delegacies of Oxford and London, lecturing mainly in general economic history and social theory. His lectures focused on the shift from laissez-faire to regulated capitalism, when state and industry began to interpenetrate, and interference of government with the captains of industry became regular.

In 1940 he became a British citizen, but from that year, he was resident lecturer at Bennington College in the USA and toured America with lectures. He also worked on *The Great Transformation* first published in 1944 (Polanyi, 1957 [1944]). After returning to the UK during the war this book earned him the post of visiting professor at Columbia University, New York, in 1947. He lectured in economic history and coordinated research into the origin of economic institutions. These efforts were summed up in the posthumously published *The Livelihood of Man* (Polanyi, 1977). He continued to do research with a team after his retirement in 1953, publishing their findings in *Trade and Market in*

the Early Empires (Polanyi et al., 1957). The posthumously published *Dahomey and the Slave Trade* (Polanyi, 1966) was another result of the cooperation.

The family settled in Canada as US authorities had denied an entrance visa to Polanyi's wife. In 1963 they visited Hungary, where Polanyi gave a lecture about American economic sociology. He died of a stroke during the night of 23 April 1964 in Toronto, Canada.

The Substantive View

The major question of Polanyi's substantive view is how to identify the place of economies in societies (Polanyi, 1947, 1977). He argues that economic activity is embedded in the social and cultural setting and that the predominance of price-making markets does not apply to most primitive, archaic, or even contemporary economies. He therefore taught that the changing place of economies could be best described by combinations of comparative concepts that shed light on the alternative logic of economic reproduction. Polanyi's substantive theory was discernible in its essentials in *The Great Transformation* (Polanyi, 1957[1944]) and in the lectures he wrote in preparing it. The book, a critique of the market system, is a derivative of historical knowledge and skills accumulated in political journalism. This is borne out by his later remark that he had two aims with the work: to transfer the more combatant Austrian workers' culture into the English labor movement; and give an up-to-date criticism of capitalism, providing intellectual support to Roosevelt's New Deal. In Polanyi's view, the imminent problem of the market system was not the existence of markets but that land and labor were turned into commodities by this system. The analysis of the so-called Speenhamland law, temporarily hindering the creation of a competitive labor market as well as of the ensuing welfare acts, revealed that welfare legislation failed to efficiently remedy the mass impoverishment paradoxically entailed by growth and could only delay the emergence of a labor market from masses of wage workers.

The book was written during World War II, parallel with Hayek's *Road to Serfdom* (Hayek, 1976[1944]) and after Burnham's *The Managerial Revolution* (Burnham, 1960[1941]). While Hayek argued that deserting the road of self-regulation leads to fatal social deformation, Burnham envisaged the inevitable predominance of managerial-elitist hierarchies with the growing autonomy of control functions. Polanyi focused on the double movement of laissez-faire economy and self-protective society. He derived the respective distortions from the market system and from the degeneration of society's defensive mechanisms. At this point he basically agreed with Owen, that the institutional system was unfavorable to social happiness, because it produced unsettled, crude, migratory types of laborers and capitalists without self-esteem. From time to time he turned back to this normative problem of good society in his lectures and writings (Dalton, 1968; Polanyi, 2003[1959]).

What constitutes the continuity between these and his earlier ideas was the criticism of the market system, the rejection of faith in advancement, the doubt about interpreting class interests as material interests, and the stress on nonanonymous historical processes resulting in the market system.

These ideas can be detected in the study 'Our obsolete market mentality' (Polanyi, 1947) and in the volume *Trade and Market in the Early Empires* (Polanyi et al., 1957). However, the emphasis in the former is on the more accurate clarification of the concept of embeddedness, and in the latter, on the empirical analysis of primitive and archaic economic formations.

Embeddedness, according to the substantivist view, is related to the fact that an economic actor is a social being, with a mixture of motives, serving to attain social recognition and social goods. His economic efforts are therefore usually subordinated to these goals. Hunger and gain are not universal motives – Polanyi and his colleagues argued – thus the scarcity postulate, the cornerstone of formal economics, is inadequate. As indicated in the title of Polanyi's key study, economy can be interpreted as an 'instituted process' in a substantive sense. The term process refers to 'location and appropriate movements,' that is production and transport on the one hand and transaction and disposition on the other.

Institutions, as he put it, lend unity and stability to these processes, which can be approached via the forms of integration. Among these reciprocity presupposes the symmetry of groups, while redistribution is dependent upon collecting and distributing goods or disposal over them, hence it implies centricity. Exchange, in turn, depends on the existence of price-making markets. Not every exchange has integrative significance. Polanyi argued that neither simple operational exchange nor decisional exchange at a set rate integrate the economy, whereas bargained rate exchange does. The forms of integration are not arranged in phases, in stages of development as do the types of List or Bücher (Polanyi, 1968); apart from the predominant forms, others are also present, constituting varying combinations. Trade, money, and markets are not necessarily functional upon exchange in the substantive approach, unless they are fitted in a market system. The ports of trade, gift trade, and administered trade usually fall outside the market system.

Similarly, only one of the money uses coincides with exchange. The rest of the uses analyzed by Polanyi – payment use, standard, ideal unit, operational device – do not necessarily tally with the market, being the typical features of the redistributive integrative form of the economy. The market, Polanyi sums up, is not so old a phenomenon, as are trade and money, and though it may crop up in history, it only integrates into a market system in exceptional cases.

While the former book contained an implicit debate with Hayek and Burnham, this one was challenged by *Economy and Society*, the book of Parsons and Smelser, which sought to integrate economic and social theory and find an alternative explanation about the place of economy in society (Parsons and Smelser, 1957[1956]).

Polanyi's Influence

Polanyi's views were deeply influenced by the writings of Bücher, Weber, Malinowski, Thurnwald, and others (Polanyi, 1977) and Polanyi filtered these influences conveying them to new waves of social research. His ideas were discussed in several volumes, selections, and bibliographies (Stanfield, 1986; Polanyi-Levitt,

1990; Mendell and Salée, 1991; McRobbie and Polanyi-Levitt, 2000; Bugra and Agartan, 2007; Harvey et al., 2007; Hann and Hart, 2009; Dale, 2010). Substantivism and Polanyi appear to be integrated in the history of sociology and anthropology (Hindess, 1977; Smelser and Swedberg, 2005; Beckert and Zafirovski, 2006; Trigilia, 2002[1998]; Radaev, 1997; Gudeman, 2001; Carrier, 2005).

The *Trade and Market in the Early Empires* volume, as well as the posthumous works edited mainly by his disciples (Polanyi, 1966, 1977; Dalton, 1968) and the disputes coming to a head in economic anthropology suggest that a school was evolving around Polanyi. In economic history, Douglass C. North (1977) assessed Polanyi's influence in basically appreciative terms, though criticizing his market interpretation. Polanyi also influenced modern researches centering on the interrelation between state and economy (Block and Evans, 2005; Dobbin, 2004; Fligstein, 2001) as well as studies of socialist economy and its disintegration (Szelényi and Kostello, 1998; Nee, 1989).

According to an evaluation, in spite of his marginal academic position, Polanyi was one of the most influential thinkers of the twentieth century. As far as the sociological analysis of the economy is concerned, this statement might be proved. Polanyi's concept of embeddedness became a point of reference for new economic sociology, although not without controversies (Granovetter, 1992[1985]; Portes, 2010). Embeddedness in the Polanyian sense means subordination of the market to social institutions and interwoven economic, social, and cultural micromotives of individuals. Some representatives of new economic sociology put more emphasis on mesolevel interpretation of embeddedness, that is on relational and structural network position of economic actors (Granovetter, 1990; Uzzi, 2005[1996]). Polanyi referred mostly to historical societies in the analysis of social and cultural embeddedness of economic action. New economic sociologists study how contemporary economic institutions are culturally and socially shaped, or how formal economic motives infiltrate noneconomic life domains and vice versa (DiMaggio, 2002; Zelizer, 2011).

See also: Capitalism; Economic History; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Hierarchies and Markets; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Market Structure and Performance; Market and Nonmarket Allocation; Regulation, Economic Theory of.

Bibliography

Beckert, J., Zafirovski, M. (Eds.), 2006. International Encyclopedia of Economic Sociology. Routledge, London, UK and New York, USA.

Block, F., Evans, P., 2005. The state and the economy. In: Smelser, N.J., Swedberg, R. (Eds.), The Handbook of Economic Sociology, second ed. Princeton University Press, Princeton, USA and Oxford, UK, pp. 505–526.

Bugra, A., Agartan, K. (Eds.), 2007. Reading Karl Polanyi for the Twenty-First Century. Market Economy as a Political Project. Palgrave Macmillan, New York, USA.

Burnham, J., 1960[1941]. The Managerial Revolution. Indiana University Press, Bloomington, USA.

Carrier, J.C. (Ed.), 2005. A Handbook of Economic Anthropology. Edward Elgar Publishing Ltd, Cheltenham, UK.

Dale, G., 2010. Karl Polanyi. The Limits of the Market. Polity Press, Cambridge, UK.

Dalton, G. (Ed.), 1968. Primitive, Archaic, and Modern Economies. Essays of Karl Polanyi. Anchor Books, Garden City, New York, USA.

DiMaggio, P., 2002. Endogenizing "animal spirits": toward a sociology of collective response to uncertainty and risk. In: Guillén, M.F., Collins, R., England, P., Meyer, M. (Eds.), The New Economic Sociology. Developments in an Emerging Field. Russel Sage Foundation, New York, USA, pp. 79–100.

Dobbin, F., 2004. The sociological view of the economy. In: Dobbin, F. (Ed.), The New Economic Sociology. A Reader. Princeton University Press, Princeton, USA and Oxford, UK, pp. 1–46.

Fligstein, N., 2001. The Architecture of Markets. An Economic Sociology of Twenty-First-Century Capitalist Societies. Princeton University Press, Princeton, USA and Oxford, UK.

Granovetter, M., 1990. Old and new economic sociology: a history and an agenda. In: Friendland, R., Robertson, A.F. (Eds.), Beyond the Marketplace: Rethinking Economy and Society. Aldine de Gruyter, New York, USA, pp. 89–112.

Granovetter, M., 1992[1985]. Economic action and social structure: the problem of embeddedness. In: Granovetter, M., Swedberg, R. (Eds.), The Sociology of Economic Life. Westview Press, Boulder, USA, pp. 53–81.

Gudeman, S., 2001. The Anthropology of Economy: Community, Market, and Culture. Blackwell Publishing Ltd, Oxford, UK.

Hann, C., Hart, K. (Eds.), 2009. Market and Society. The Great Transformation Today. Cambridge University Press, Cambridge, UK.

Harvey, M., Ramlogan, R., Randles, S. (Eds.), 2007. Karl Polanyi. New Perspectives on the Place of the Economy in Society. Manchester University Press, Manchester, UK.

Hayek, F., 1976[1944]. The Road to Serfdom. The University of Chicago Press, Chicago, USA.

Hindess, B. (Ed.), 1977. Sociological Theories of the Economy. Holmes and Meier, New York, USA.

McRobbie, K., Polanyi-Levitt, K. (Eds.), 2000. Karl Polanyi in Vienna. The Contemporary Significance of the Great Transformation. Black Rose Books, Montreal, USA.

Mendell, M., Salée, D. (Eds.), 1991. The Legacy of Karl Polanyi. Market, State and Society at the End of the Twentieth Century. Macmillan, Basingstoke, UK.

Nee, V., 1989. A theory of market transition: from redistribution to markets in state socialism. American Sociological Review 54, 663–681.

North, D.C., 1977. Markets and other allocation systems in history: the challenge of Karl Polanyi. Journal of European Economic History 6 (3), 703–716.

Parsons, T., Smelser, N., 1957[1956]. Economy and Society. A Study in the Integration of Economic and Social Theory. Routledge and Kegan Paul, London.

Polanyi, K., 1957[1944]. The Great Transformation: The Political and Economic Origins of Our Time. Beacon Press, Beacon Hill, Boston, USA.

Polanyi, K., February 1947. Our obsolete market mentality. Commentary 3, 109–117.

Polanyi, K., Arensberg, C.M., Pearson, H.W. (Eds.), 1957. Trade and Market in the Early Empires. Economies in History and Theory. The Free Press and the Falcon's Wing Press, Glencoe, USA.

Polanyi, K. (in collaboration with Rotstein, A.), 1966. Dahomey and the Slave Trade. University of Washington Press, Seattle, WA.

Polanyi, K., 1968. Bücher, Karl. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 2. The Macmillan Company and the Free Press, New York, USA, pp. 163–165.

Polanyi, K., 1977. In: Pearson, H.W. (Ed.), The Livelihood of Man. Academic Press, New York, USA.

Polanyi, K., 2003[1959]. Arisztotelész és Galbraith a jólétről (Aristotle and Galbraith on affluence). Szociológiai Szemle 13, 137–142.

Polanyi-Levitt, K. (Ed.), 1990. The Life and Work of Karl Polanyi. A Celebration. Black Rose Books, Montréal, CA.

Portes, A., 2010. Economic Sociology. A Systematic Inquiry. Princeton University Press, Princeton, USA and Oxford, UK.

Radaev, V., 1997. Ekonomicheskaya Sotsiologia. Aspect Press, Moscow, RU.

Smelser, N.J., Swedberg, R. (Eds.), 2005. The Handbook of Economic Sociology, second ed. Princeton University Press, Princeton, USA and Oxford, UK.

Stanfield, J.R., 1986. The Economic Thought of Karl Polanyi. Lives and Livelihood. Macmillan, Basingstoke, UK.

Szelényi, I., Kostello, E., 1998. Outline of an institutional theory of inequality: the case of socialist and postcommunist Eastern Europe. In: Brinton, M.C., Nee, V. (Eds.), The New Institutionalism in Sociology. Russel Sage Foundation, New York, USA, pp. 305–326.

Trigilia, C., 2002[1998]. Economic Sociology. State, Market and Society in Modern Capitalism. Blackwell Publishing, Oxford, UK.

Uzzi, B., 2005[1996]. The sources and consequences of embeddedness for the economic performance of organizations: the network effect. In: Abolafia, M.Y. (Ed.), Markets. Edward Elgar, Cheltenham, UK, pp. 175–199.

Zelizer, V.A., 2011. Economic Lives. How Culture Shapes the Economy. Princeton University Press, Princeton, USA and Oxford, UK.

Police: A Sociology of Knowledge Approach

Peter K Manning, College of Criminal Justice, Northeastern University, Boston, MA, USA
Jeffrey Martin, University of Hong Kong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociology of the police concerns the study of the institution of policing, its structure, function, and evolution. This focus requires a sociology of knowledge, a framework that highlights how social factors shape current knowledge of police. The concept of police is a contested one and often assumed without definition. Its functions are associated with governance in general and the executive function in particular. The public policing form assembled by Robert Peel, a visible, reactive force, is an assumed paradigm for much police research, but it is one of several types of policing, and contrasts with high policing concerned with national security. Internationally, broad types or 'families' of policing have emerged as a result of colonialism. Four specific types of police are common in Western democracies: private policing, pubic policing, hybrid policing, and outsourcing or paying organizations to carry out functions previously restricted to police agencies. Metaphors for studying policing, such as dramaturgy, the policing web, and nodal policing, highlight selected features of the policing function. Many external factors shape policing and the future role and function of the police, given new forms of crime, is unpredictable.

The sociology of the police concerns the study of the institution of policing, its structure, function, and development. This perspective implies a sociology of knowledge, that is, how social interests and ideologies shape the sociological study of police. The independent analytic and theoretically grounded study of the police is in competition with a sociology for the police, a response to the practical-occupationally dictated concerns of police command. These two quite different foci lead to different research topics, findings, and definitions of the central topics of interest. The *concept* of police itself is contested, or more specifically, an ambiguous one, as it functions as an institution of governance, an extension of the executive branch. It is generally seen as managing formal social control within the context of the nation-state, and its evolution is related, therefore to social differentiation of functions, legitimate authority, and the emergence of a distinctive mandate.

Although many forms of formal policing exist and are embedded in culturally and legally defined sociolegal systems (Manning, 2010: p. 25), most of the present systematic research concerns policing in common law and continental systems, and the work of police engaged in everyday functions rather than in specific state security functions. It is well to remember that policing had a general function before the late eighteenth century as that which maintained the health or integrity of the body politic, and that the evolution of the concept must be traced through the innovations of the gendarmerie, of policing by Louis XIV and Napoleon. In this context, 'policing' was a combination of national economic interests, national and local security, and represented in time by distinctive organizations (Chapman, 1970; Emsley, 1994).

What is generally considered now the sociology of the police in the Anglo-American world is rather parochial and is traced to the founding of the Dublin City Police in 1786, and to later innovations associated with Robert Peel as Secretary for Ireland and later as Home Secretary. The London-based police of Peel, a response to complex societal changes, was an unarmed, visible, reactive officer standing symbolically between the people and the government. This imaginative and successful move set a paradigm for what is thought of as 'police.' It is only one sort of formalized policing, detached to some degree from national security and order maintenance. This 'Peel paradigm' absent in the armed semisecret riot police was adopted peacefully in counties throughout England by 1855, peacefully in the United States, Canada, and the Antipodes, and imposed by conquest in British colonies. Perhaps the most significant consequence of the diffusion of Peel's concept in time was a political commitment to varieties of localized, somewhat culturally sensitive, control over policing. This was not the case of course in Colonial nations and in non-English-speaking nations. Perhaps inevitably, then, the scope of 'police' and a sociology of the police shrank metaphorically and became synonymous with the study of public, legitimate, state-sponsored agencies that operated within a specified territory and sociolegal political system. Police, found in anomalous circumstances, exceptions to this tacit sketch, such as the gendarmerie, colonial police, police in conquered territories, secret and national-security police, police in authoritarian systems, and police in 'settler states' (Weitzer, 1990, 1995), have not been as extensively studied. The combination of scholarly interest, governmental funding, and the growth of cities in democratic societies led to an interest in and research about public police shortly after World War II. A series of important early studies in the 1960s and 1970s set the tone and focus of the research.

Scholarship

The development of scholarship on the public police in English has been well summarized (Reiner, 2010). But this work, broadly based and concerning police functions, by and large has assumed a well-ordered democratic state with sufficient resources, legitimacy, and a compliant citizenry. These assumptions have obscured 'deeper' questions of policing such as the loyalty of police organizations and their members to the state, their roles in revolution and a *coup de grace*, and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32107-9

their ideological commitments. They are armed and dangerous; where do their loyalties lie in a crisis? These questions, however, play an insignificant role in the current literature. Studies in the last 50 years or so, published in English with some exceptions, describe (1) socialization to the role; (2) functions of patrol officers and investigators, and lower participants rather than middle management or top command; (3) strategies of policing such as community policing or tactical variations such as 'crack downs' on high crime areas; (4) the organizational and occupational culture of the police; (5) variation in attitudes within the occupation based on race or gender; (6) accountability, corruption, and violence; and (7) the impacts of technology on police practice. There are very few studies of police unions, decision- and policy-making at the top, the internal politics of succession to command positions, or indeed the internal politics of the organization: who gets what, when, and where? There is very little understanding of the current dynamics, emotions, and politics within large urban departments. Whether there is indeed a systematic theory of the police is debatable.

Definition of Police as an Organization

As noted above, the growth of the sociology of police has taken place without much debate over the central yet contested definition of police and policing. If the police as an organization is to be broadly understood, then it must include private as well as public police and include the secret surveillance, tracking, and intelligence gathering that is an essential part of national security. Policing has a mandate, a tacit contract between the public and the police that sets out the boundaries of the actions, and denies these functions to others. Ironically, police are permitted to act outside the law in the pursuit of their obligations (Manning, 1977). Consider this definition of police in modern nation-states with common law or civil law traditions:

> The police, constituted of many diverse agencies, are authoritatively coordinated, legitimate organizations. They stand ready to apply force up to and including fatal force in legally defined territories. They seek to sustain politically defined order and ordering via tracking, surveillance, and arrest. As such, they require compliance to command from lower participants and citizens. Given their wide mandate, they have the ability to proceed by exception.

As discussed below, this definition may require reformulation as Asian forms of policing become better known and researched and the scope of police studies expands. The above definition is an organizationally based definition, which contains a number of ambiguous features, namely, the ambiguity of 'order' and 'ordering,' which cannot be defined in the absence of a sociolegal context, and the capacity for violence. It is thus a formal and normative definition that begs the question of police function. Bittner (1990: p. 131) eschews the normative view, and argues for a means-focus: "The role of police is best seen as a mechanism for the distribution of non-negotiable force employed in accordance with the dictates of an intuitive grasp of situational exigencies." These exigencies, in turn, are best determined by the here and now assessment of the officer. The above definition does include a matter that is implicit in the Bittner definition: the compliance of a pacified citizenry. This trust and related compliance is the source of the police mandate in modern states (Manning, 1977). The framing of the function of modern policing requires a distinction between law and other social norms and values and a *code* or interpretative context into which conflicts are entered. As Bittner (1970: p. 40) points out, the nature of intervention by modern police is not determined by the natural notion of solution that arises from citizens, but lies in the practices of the police in their management of the conflict. This interpretative code is not law, but rather the framing of the case so as to be entered into the legal system. The extension of this formulation is that Bittner focuses on the means, and the idea that, although force is the cornerstone of his idea, police violence is left undefined by the courts and (p. 34): "… police activities cannot be brought fully under the rule of law." Their independence of the courts, spelled out in detail in Bittner (pp. 22–35), facilitates and validates their intuitive actions. They are in effect permitted to act and account for their actions and decisions later. This formulation leads to the rather ironic conclusion that policing's rationality, since it has no known ends, is the rationality of the moment, and this deciding rarely lands in court or comes under significant judicial scrutiny. (Alternatively, one could argue following Jobard (2013) that they remain an instance of organized irrationality in the modern, rapidly rationalizing world. Furthermore, one might argue that their situated actions, because they do not and cannot draw on any fixed rules, combine ignorance and irrationality and are a form of performative magic. The idea that police draw on magic, or are magicians, is not flattery, but is derived from Durkheim's (1961: pp. 58–60) distinction between practitioners of sacred arts, who believe in what they are doing and know why are they acting as they are, and magicians, practitioners do not believe in the tacit premises that govern their actions. They aim to deceive.) This is one facet of the phrase 'proceed by exception' because no precise a priori limit can be set upon their actions. As the above definition states, for the democratic police to manage a pacified audience, they must be able to draw on their dramaturgical powers, the ability to draw on symbols and imagery, and carry out persuasive performances. They are naturally dramatic, and designed historically to produce trust.

Core versus Fringe Imagery

Let us consider another aspect of the sociology of police: the core versus the fringe imagery of police. Imagery is the ways in which the police present themselves – dramatize their actions and intentions – and how they are seen. The core imagery in Western industrialized societies is the visible, uniformed, reactive, service-oriented patrol officer, while the fringe functions, less visible and rarely considered or written about, are carried out by specialized paramilitary units, riot-control, surveillance and tracking functions, often now associated with antiterrorist activities, and units devoted to 'high

policing' (Brodeur, 2010). Nonuniformed officers are also engaged in such activities from time to time. This blindness to covert and less-than-visible activities has meant that a disproportionate amount of research in the sociology of police has been devoted to the activities and occupational culture of the uniformed officer, and relatively little to investigators, top command, and specialized units (Marks, 2005).

Historical Modes of Formal Policing: Centers and Peripheries

Most people in the world live under political arrangements that developed outside the historical core of the capitalist world system. Scholars in a number of disciplines have contributed accounts of histories and practices of policing in the 'alternative modernities' endemic to postcolonial or developmental environments. This diverse literature has recently begun to coalesce around the relativism of cultural anthropology into a sort of critical theoretical counterpoint in dialogue with the mainstream sociology of police (Garriott, 2013; Jauregui, 2013). Colonized societies saw anthropologists and police arrive together, as partners in the colonial system of control. Decolonization then forced contemporary anthropology to reevaluate the ways its inherited archives of 'cultural' information were implicitly structured by the practical concerns of the colonial officers who created them (Pels, 1997). Critical studies of the classification of populations in British South Asia (Cohn, 1987), Dutch Indonesia (Stoler, 2009), and elsewhere have illuminated some of the many ways cultural knowledge has historically functioned as a police technique. The contemporary anthropological study of policing takes this lesson to heart, examining policing as a field of knowledge/power in which a variety of institutions operate as means to diverse ends, structured by a plurality of hoped-for ideals of modern peace and/or prosperity (Comaroff and Comaroff, 2006).

Globally comparative discussions of policing arrangements are generally organized under historical and regional themes. Historically, there seems to exist a roughly tripartite typology of institutional 'families,' depending on whether a given policing system is derived from Anglo-American, Continental, or Soviet roots. Studies of policing in the former British Empire (Anderson and Killingray, 1992; Sinclair, 2006), e.g., show how the common law ideal can be differentially taken up as an element in a local 'policing vernacular' (Hornberger, 2011). Latin American policing, by contrast, is largely derived from the Civil Law tradition (Uildriks, 2009). Ethnographies of contemporary Latin American policing have focused on the 'disjunctive' qualities that arise where centralized policing systems serve to maintain radical economic inequalities (Caldeira, 2001). East Asian policing, finally, grows out of the Japan's nineteenth-century appropriation of French and German legal and administrative institutions (Martin, 2014). Subsequent to its 1949 revolution, China developed a totalitarian system that projected police power deep into the minutiae of everyday life, enforcing socialist order onto things ranging from the way words could be written to the number of children that could be born. China's subsequent success translating this authoritarian system into the developmental mechanism it has used to become a capitalist world power makes contemporary Chinese policing something of a *sui generis* formation, which is only beginning to be critically studied (Cao et al., 2014).

Types of Policing

Policing, as noted in the above formal definition, includes private as well as diverse forms of public policing, and as such, contrasts with *informal policing* or perhaps 'self-appointed or nominated policing,' which includes ad hoc policing, posses, militias, vigilante groups, and other informal, part-time policing individually or collectively that is carried out without uniform, pay, and specific conditions of employment. These informal police officers cannot be fired and thus their accountability is dubious. There are also 'minders' of public order who in the course of their duties sustain orderly activities, e.g., taxi drivers; traffic and parking attendants; bus, train, and subway drivers; emergency medical services and emergency personal; and members of the fire service. Thus, a sociology of policing differs from a sociology of the (formal) police. The range of functions carried out by the public police is diverse and complex. More is known about the workings of formal public police than about other forms.

Within the industrialized world, examination of formal policing reveals many types of policing with varying *mandates*, or legitimate validation of their claims to control the proper attitude toward the work, their rhetorical and resource-allocation strategies, and tactics and accompanying themes or *dramatic emphases* in their rhetoric, uniforms, and actions. In many ways, modern public police emphasize a theme of responsiveness, service, and crime control. They are in competition for legitimacy with other types of and of course informal social control. Here are some types of formal policing, their interests, and themes:

- *Private policing* – paid for actions in regard to maintaining and or expanding private interests. Usually connected to policing given territories. Policing with private interests as the mandate only marginally, if at all, serves the public welfare, even when policing quasi-public property. The functions carried out by private police are those connected with not only sustaining order in quasi-public spaces, but also in all those areas in which control of a citizenry is dubious. The theme here is prevention.
- *Policing carrying out traditionally public police functions* – using public funds to pay for agents to carry out functions connected the public good, e.g., traffic tickets, order maintenance; regulation that does not entail arrest in excess of 'everyday' citizen powers. This is also called 'outsourcing' on the grounds that it reduces costs and functions by contracting. Such policing functions have a mandate in the sense that they are licensed, legitimate, and have territorial limitations. They draw also on the mandate of public police by simulating their uniforms, interactional style, the decoration of their vehicles, and symbols. In the United States, they carry visible weapons. They may or may not include some arrest powers in private spaces for trespassing. The theme is protecting boundaries.

- *Hybrid policing* – varieties of policing, i.e., noticing, responding to, and perhaps sanctioning behavior with a quasi-public mandate. This sort of policing is now included in the 'police family.' The confusion in regard to legitimacy comes in that these functions can be carried out by police employed privately, police employed and paid by the public, or some amalgam of these. This can also include third-party policing and varieties of the police web. The theme here is service.
- *Public police* – the police organization includes warranted officers, those with arrest powers, and those who are not warranted or sworn. These 'civilians' can make up as much as 50% of the organization's employees. In addition, the organization can include reserve constables, cadets, and part-time officers. The organizational culture includes more than the segmented occupational culture of the police. The theme here is responsiveness and service.

In addition, the status of federal agents, the military and military police, and military investigative services in the United States, Australia, and Canada, is ambiguous. They carry out many functions that are unrelated to national security, overlap or compete with local jurisdictions, and operate primarily within federal laws and courts. There are also highly specialized police forces in the Anglo-American world with quite precise and limited mandates, which mean that they are not in competition with the above-listed types, e.g., British Railway Police; Amtrak (US railroad) police, and the police of an American rural development project, the Tennessee Valley Authority.

Policing and Metaphors

To study modern policing requires more than the study of police organization or even policing; it requires metaphoric work. The sociology of police rests on metaphor: a way of seeing something in terms of something else. If we conceive of the police as 'law enforcement,' we reduce them to a mere part of their whole; if we call them a 'business,' or an agency providing human services, we are treating the whole and denying many of the parts. The expansion of vision that has taken place in the last 20 years in the sociology of police, a result of transnational communication, migration, and travel, as well as transnational policing, now requires an understanding of how types of policing rather than merely public police are to be seen and seen to be in competition or conflict with each other. Brodeur (2010) has suggested the metaphor of a *policing web* or assemblage that connects loosely those agencies that police in some fashion. These organizations are a collection that is to some degree in conflict, in competitive, and/or conflict mode over time and across incidents. Think of this assemblage as a web: the strands are unequal in strength; the number of their connections varies, and the content, the communications exchanged, vary in density and frequency, e.g., information about crime, disorder, villains, and disasters. This shifts attention away from myopic focus on public police, draws attention to multiple and competing policing functions, and suggests the overlap of surveillance and control in which citizens are embedded. The web concept is, however, focused on agencies and organizations, not on interdigitated functions, and Brodeur's position is that the public police are and must remain the center of authority primacy of resources and collective obligations. A parallel and equally fascinating *trompe d'œil* is Shearing's notion of *nodal policing* (2005), which begins with charting functional capacities, technologies, mentalities, or socially grounded perceptions, and institutional arrangements. These clusters of functions are arrayed in some sense around social security or a collective sense of well-being. They do not stand as organizations but as nodes of function. These are seen as contingent arrangements for addressing emerging problems, rather than conventionally mandated responses to anomalies. In some sense, the idea is an analogue to the open market or invisible hand notion of organizing responses. There is an element of spontaneity about such responses. The public police are partners in this enterprise of security maintenance, and seen as cooperative partners in problem solving. These metaphoric moves open up new vistas and new avenues in the sociology of police.

The Role of External Factors

In the course of the development of the interdisciplinary field of police studies, the focus has been on the police organization as a dramaturgical actor within a field, the activities of officers, both investigators and uniformed officers, or the occupational culture and subcultures. In the last 20 years, the focus has been on tactics, crime control efforts, and efforts to measure and prevent crime. Surprisingly, the role of the organizational, political, and economic environment has been described in the abstract by economists and political scientists, but few empirical studies demonstrate how such external factors actually pattern policing practices such as arrest and enforcement targets. Developments in the last 30 years, inexpensive communication, travel, migration and immigration, the rise of terrorism and antiterrorism, global commerce, and international, transnational, and metanational policing (Bowling, 2010), have also widened the vision of scholars. The future problems of technocrime, international organized crime, and terrorism influence the police differently in developed nations. How and to what extent these influences will shape the police mandate, strategies, and tactics is yet to be determined.

Methods and Problematics

The study of police is problematic because access is difficult, the police are secretive and often suspicious of research and researchers, the work is diverse and spread out ecologically, and police data require context to interpret. They require translation and back translation (they are gathered in a context that is not the same as a research context). The early research was ethnographic and fieldwork based, and later studies expanded to include other police-gathered-data (officially recorded crime), interviews, questionnaires, and police records. Attention shifted in the 1980s to community policing and attitude data gathered from the public. Recent attention has been on crime reduction by manipulation of police presence and formal designs to measure the impact of tactics on the reduction of

crime and disorder. An abiding concern of researchers has been normative, concerned with corruption, excessive use of force, accountability, race profiling, gender and minority mistreatment both within the organization and outside it, and the quality of police–citizen interactions. Of late, reducing officially recorded crime has been a research focus. The impacts of new electronic technologies has not been closely examined, even as police develop Web sites and interactive blogs accessible to the public; employ social media, such as Facebook and Twitter; and maintain Web sites. While use of cell phones in the past was considered a breach of security as well as a violation of the chain of command, they are now considered part of the everyday equipment of all officers (it is likely now that most officers above the rank of sergeant carry at least two smart phones, one their own and one a departmentally issued phone).

The Future of the Sociology of Police

The future of the sociology of the police cannot be predicted. The sociological frames of reference within which police and policing are cast; the practitioners' interests, which bear an abiding influence on the research; and the inevitable fads and fashions in social science all will shape intellectual problems and research methods. Global issues such as the ongoing turmoil in the Middle East, global warming and related changes in the climate, bodies of water, and natural disasters, as they are seen to be in the ambit of the police, will shape the coming century of research. A corollary of the rise of democratic policing and governance is the resistance of the powerless in the form of terrorism and the response of antiterrorism tactics, paramilitary units, and increased surveillance. The issues that may arise are questions of democratic policing as it is associated with emerging nation-states, changes in the political economy of developing nations, and the growing importance of human rights.

See also: Colonization and Colonialism, History of; Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Ethnography in Applied Social Research; International and Comparative Criminology; Knowledge, Sociology of; Myths and Symbols: Organizational; Organizations and Culture; Policing; Public Bureaucracies; Public Sector Organizations; South Asian Legal Traditions.

Bibliography

Anderson, David, Killingray, David, 1992. Policing and Decolonization: Politics, Nationalism and the Police, 1917–1965. Manchester University Press, Manchester.
Bittner, Egon, 1990. Aspects of Police Work. Northeastern University Press, Boston.
Bowling, Benjamin, 2010. Policing the Caribbean. Oxford University Press, Oxford.
Brodeur, J.-P., 2010. The Policing Web. Oxford University Press, New York.
Caldeira, Teresa, 2001. City of Walls: Crime, Citizenship and Segregation in Sau Paulo. University of California Press, Berkeley.
Cao, Liqun, Sun, Ivan, Hebenton, William, 2014. The Routledge Handbook of Chinese Criminology. Routledge, London.
Chapman, Brian, 1970. The Police State. Pall Mall, London.
Cohn, Bernard, 1987. The census, social structure and objectification in South Asia. In: An Anthropologist among the Historians and Other Essays. Oxford University Press, Delhi, pp. 224–254.
Comaroff, Jean, Comaroff, John L. (Eds.), 2006. Law and Disorder in the Postcolony. University of Chicago Press, Chicago.
Durkheim, E., 1961. The Elementary Forms of the Religious Life. Collier Books, New York.
Emsley, Clive, 1994. The Gendarmes and the State in Nineteenth-Century Europe. Oxford University Press, Oxford.
Garriott, William (Ed.), 2013. Policing and Contemporary Governance: The Anthropology of Police in Practice. Palgrave Macmillan, New York.
Hornberger, Julia, 2011. Policing and Human Rights: The Meaning of Violence and Justice in the Everyday Policing of Johannesburg. Routledge, New York.
Jauregui, Beatrice, 2013. Cultures of legitimacy and postcolonial policing. Law & Social Inquiry 38 (3), 547–552.
Jobard, F., 2013. The concept of the police springer encyclopedia of criminology and criminal justice. In: Weisburd, D., Bruinsema, G., Manning, Peter K. (Eds.), Police Work. MIT Press, Cambridge.
Jobard, F., 2012. Democratic Policing in a Changing World. Paradigm Publishers, Boulder, CO.
Manning, Peter K., 1977. Police Work: The Social Organization of Policing. MIT Press, Cambridge.
Manning, Peter K., 2010. Democratic Policing in a Changing World. Paradigm Publishers, Boulder, CO.
Marks, Monique, 2005. Transforming the Robocops: Changing Police in South Africa. University of KwaZulu-Natal Press, Durban.
Martin, Jeffrey, 2014. The confucian ethic and the spirit of East Asian police. Crime, Law & Social Change 61 (4), 461–490.
Pels, Peter, 1997. The anthropology of colonialism. Annual Review of Anthropology 26, 163–183.
Reiner, Robert, 2010. The Politics of the Police, fourth ed. Oxford University Press, Oxford.
Shearing, Clifford, 2005. Nodal security. Police Quarterly 8, 57–63.
Sinclair, Georgina, 2006. The End of the Line: Colonial Policing and the Colonial Endgame, 1945–1980. Manchester University Press, Manchester and New York.
Stoler, A.L., 2009. Against the Archival Grain: Epistemic Anxieties and Colonial Common Sense. Princeton University Press, Princeton.
Uildriks, Niels, 2009. Policing Insecurity: Police Reform, Security and Human Rights in Latin America. Rowman & Littlefield, Lanham.
Weitzer, R., 1995. Policing under Fire: Ethnic Conflict and Police Community Relations in Northern Ireland. SUNY Press, Albany. http://books.google.com/books?hl=en&lr=&id=67hoMfy4kTsC&oi=fnd&pg=PR7&dq=Weitzer,+Ronald+&ots=ywjuAkFQ8m&sig=FVUFX4Ia5GLA9blnynJHwzpwHrY.
Weitzer, R., 1990. Transforming Settler States. University of California Press, Berkeley.

Police, Sociology of

Jonathon A Cooper, Indiana University of Pennsylvania, Indiana, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.-J. Albrecht, D. Nogala, volume 17, pp. 11532–11535, © 2001, Elsevier Ltd.

Abstract

Police as a formal institution of social control, organized within the framework of the nation state, emerged during the course of the eighteenth and nineteenth centuries. Although some issues of police and policing had been studied by the classic authors of sociology, it became a subject of genuine sociological study only after World War II. Early research was focused on empirical studies of police organizations and the actual work of police officers, coupled with theorizing. As a new and special terrain of research, sociology of police was at its height in the late 1960s and the 1970s among Anglo-American scholars. From then on a growing body of literature dealt with judicial, organizational, occupational, and political aspects of police work, however rarely under an explicit label of 'sociology of police.' Today, police is a topic for different academic disciplines, among them sociology.

Introduction

The *sociology of police* as a specialized area of study can be defined as the empirical and theoretical analysis of those institutions charged with the investigation and prevention of crime and the maintenance of public safety and order. Such organizations are usually, but not exclusively established by state governments. Such a subdiscipline is necessarily interdisciplinary, and embraces the systematic enquiry about the role of police in modern societies, the general characteristics of policing bodies and their relation to other institutions, the institutional development and internal structure of police forces, the collective and individual behavior of their members, and the interactions or relations of those agencies with individuals, different social groups, institutions, and the general public.

Police as a Matter of Sociological Concern

Both police and sociology are offsprings of the Enlightenment and the Industrial Revolution. Although the contours of both as we understand them today became obvious only in the nineteenth century – the former as a new formal institutionalization of social control, the latter as a new academic discipline – there is a line of discourse which can be traced back to ancient times, when Roman emperors struggled to get or keep their rule by using armed forces and when philosophers reasoned about the art of best governance. Despite the fact that the contemporary notions of 'police' and 'sociology' were not yet understood as such then, the reasoning about the necessity of establishing and maintaining a public force to control crime and to ensure public order – and the critical discussion of its actual appearance – is as old as the discourse on the *res publica*.

Later on, the issue of police was embedded into the broader philosophical topic of the state and government by early classic sociologists and social philosophers. For example, police were discussed mainly as an instrument for the sovereign's power to govern. Max Weber's often quoted notion that the state is that human community that successfully claims a monopoly of legitimate use of force within a certain territory still prevails within academic reflections on the police, either explicitly or implicitly (Bittner, 1991).

With the growing urbanization of the countries in industrial and technological transition, there was also an increasing body of literature on police matters from a pragmatic perspective in the first half of the twentieth century. Yet it took some time for the police to become an explicit and original matter in sociological thought and empirical inquiry, primarily in English-speaking countries. Such a change was largely due to a new era in the mid-twentieth century of the sort of civil transition that heralded creation of the police in the nineteenth century, such as race riots.

Determining the Subject

At first glance and seen from the surface, it seems rather trivial to determine who and what constitutes 'the police.' One would expect that the police as a public institution is represented by sworn officers, i.e., representatives of the state and its government, who are trained professionals uniquely invested with certain powers like the authority to search or arrest a person. Usually this 'apparent' kind of police will appear in public as officers on the beat, a crew in a patrol car, behind a desk in a police station, or as plainclothes detectives doing investigations in the field. This is the common and popular image (in the Western world) of what police actually are, transmitted by the media and supported by occasional ordinary encounters. From time to time this picture is widened by the appearance of riot police in full gear, underlining the state's authorization and capability of using legitimately physical force against disobedient citizens or rioting crowds.

When one looks again at who and what constitutes police, the semblance becomes more complicated. On the level of involved actors, it becomes apparent that a good portion of the workforce employed by the police consists of civilian staff, like secretaries or the clerks in the forensic laboratories; some sworn officers, such as staff who are responsible for conducting statistical analysis of criminal incidents or running a computer

program for matching data, are rarely out in the field, while certain field officers, like undercover agents or specialists in charge of surveillance of telecommunication, often do not act openly as police. Furthermore, there are other state agents who, having similar powers of investigation or intervention, perform certain functions of policing, but are not seen or labeled as 'police.' One might think here of custom officers, public health inspectors, water rights investigators, or prison guards.

As professionals trained to use state-sanctioned – and sometimes lethal – force, the police are unique. But on the level of routine work processes, this distinguishing feature is a rather rare event compared to the overall picture of police duties. Instead, research has shown that a considerable part of the modern police workload is not at all focused on crime control and investigation of criminal cases, but consists of responding to general emergency calls, mediating conflicts, regulating motor vehicle traffic, and communication with other institutions or agencies like social services, insurance companies, etc. Even the popular notion that an officer's shift consists of 7 h of absolute boredom and an hour of sheer terror exaggerates the reality.

Although it might be clear from a commonsense point of view what 'police' actually are, the subject in question gets more diverse the more we see it from a perspective of a peculiar organized social activity rather than as a matter of institutionalism. Thus, a true 'sociology of policing' would cover a wider area and embrace more organizations than initial 'police studies.' To wit, others appear in a policelike guise, but they are clearly not sworn officers (at least lacking their full powers): city wardens, commercial guards and patrols, hired investigators, vigilante organizations like the 'Guardian Angels,' bodyguards, stewards in a football stadium, or bouncers. This category might also include the pattern of civilianization among police agencies, where nonsworn officers are sent on routine reporting calls, such as cold burglaries.

It should not be overlooked that police as a subject of sociological interest is linked in many ways with other social systems, all of them carrying their own, often overlapping, bodies of literature. As an instrument of executive governance 'the police' can be seen as a segment of the *sociology of the state*. On the other hand, its quality of being an important part of the criminal justice system and its more or less explicit legal bindings makes it a component of the *sociology of law*, as Donald Black (1980) has argued. Finally, on the level of empirical studies, there are clear ties to the field of *organizational sociology*, as John Crank (2003) has suggested.

Changes in Focus and Emphasis

Although some academic work on the police can be traced back to the 1950s and earlier, the first surge of sociological studies of police work occurred during the aftermath of the increasing social and political changes that affected most countries of the Western world from the early 1960s on. In a time when social and political conflicts were brought onto the streets, and rising concerns about crime and disorder became a public issue, 'the police became increasingly visible, controversial and politicized in response to these tensions and pressures' (Reiner, 2000, p. 210).

According to Cain (1979), five types of studies, following a rough chronological sequence in the shift of areas of interest, can be identified for this period: (1) a law-related concern with civil rights when confronted with police powers and discretion; (2) examinations of the police organization by participant observations (Skolnick, 1966; Manning, 1977 are well-known examples); (3) studies of the roles that the police take in the social construction of deviants and the fabricating of public images of deviance; (4) analysis of the interactions and relationships of the police with the social environments they are working in; and (5) studies questioning the political contexts, presumptions, and consequences of police activities. It can be said that all these topics overlap in a way and have persisted during the proceeding decades.

There are two important points to note. First, that most of the aforementioned research was overwhelmingly conducted in the USA and the UK. Second, the main characteristics of this work, coming out of universities, private organizations such as the Police Foundation, and government agencies such as the Home Office, were motivated by a concern with issues of the use of police discretion and generally tended to be critical of the police. For example, the famous Kansas City Preventive Patrol experiment aimed to understand how police discretionary behavior (in terms of random patrol) impacted crime and fear of crime, and found, ultimately, that it did not necessarily have such an impact. This evaluation interest was grounded in the awareness that police did not always use their discretion as expected, or in moral ways, brought about by the aftermath of the 1960 race riots.

The policing research that was conducted by private think tanks such as the Police Foundation and the Police Executive Research Forum, and government agencies like the Bureau of Justice and the Home Office, tended (and tend) to focus on questions of good practice and evaluation of reform measures, and are generated under policy and managerialist perspectives (Reiner, 2000). And so we not only have the Kansas City Preventive Patrol Experiment, we have a general research agenda of *what works*, or, as the buzzword *par excellence* of the twenty-first century would put it, *evidence-based practices*. This followed a general trend in governance, but ultimately complicated the *role* of the police. As it became more apparent, from research, that police by themselves could do little to impact crime, their relationship to government, society, and community became increasingly complex.

Current Theory and Research

There has been precious little theorizing in policing over the last 30 years. The bulk of policing scholarship continues to draw on the 'classics' of policing theory, such as the works of Black (1980), Van Maanen (1974), Manning (1978), Wilson (1968), Jackson (1989), and Skolnick (1966). Still, there is a small group of theorists who have expanded policing theory, including Kane (2002), Klinger (1997), Bernard and Engel (2001), and Crank and Langworthy (1992). For the most part, these theories are quasifunctionalist with a Marxist bent, and rely heavily on institutional and psychological theories.

A number of trends in current policing research can be discerned, although much of it is carry-over from previous

decades. For example, we are still interested in aspects like recruitment and education, gender and minority issues, the role of unions and professionalism and specialization, and questions surrounding the use of force, particularly those conducted with electricity devices (e.g., the electroshock weapons for neuromuscular incapacitation manufactured by TASER). The question at stake here is not just what police are actually doing, but is it effective and why or why not. An enormous amount of scholarly energy and government monies have been expended in the last two decades trying to even understand what is meant by the term 'effective' when it comes to the police. This debate continues. Nevertheless, a 'what works' and 'evidence-based' research agenda largely dominates policing scholarship today, with or without theory backing it (Cooper and Worrall, 2012).

A good portion of police research literature is still concerned with issues of internal, as well as external, control mechanisms of individual and institutional police behavior. In this respect, the institutional autonomy of the police in the face of the government and the judiciary is one side of the question, while the control of the behavior of individual officers is another (Marx, 1988). The work of Samuel Walker (see especially 2005) and the movement in the USA and abroad to implement citizen oversight of police continue to be relevant.

Comparative Police Research

It is still the case that the bulk of police studies which are relevant for the sociology of police in one way or the other is produced in English-speaking countries, with the USA in a clear lead. This might be seen as the reason why, on an international level, the discourse on police and policing issues mainly reflects these countries. For example, debates on 'community policing' or 'zero-tolerance strategies' originate from the development of police policies in the USA. But as matters of cross-border and international police cooperation have become more important since the early 1990s, there is a growing interest in comparative perspectives (Brodeur, 1995; Findlay and Zvekic, 1993). This has especially been the case as other non-Anglo-American and even non-European countries have begun to take policing scholarship seriously – such as Belgium and the Netherlands, and South Korea. Indeed, it has only been very recently that the European Society of Criminology has formally adopted a police working group, the creation of which was pushed to a large degree by Scottish and Belgium scholars.

The contemporary sociology of police and policing has to take two aspects into account. First, there are quite a lot of significant differences in how police institutions are organized and how policing methods are actually achieved among different countries. This is mainly due to diverse national historical backgrounds. The diversity lies particularly in the arrangements of the police system that are influenced by legal, political, and occupational cultures (Busch et al., 1985; Monjardet, 1996), as well as the distinction between civil and common law systems, to say nothing of the added complication of federalism versus centralization of government. The other aspect is that when it comes to the level of policing (i.e., how police work is actually achieved and what kind of resources and strategies are used), some striking similarities and convergences can be noted. There seems to be an increasing transfer of issues in the international police discourse.

Increasingly, matters of police management and evaluation are discussed, especially those that are connected with institutional police reforms in the UK. Modernization and technology seems to have been an enduring issue in police reform since the 1970s throughout the globe.

Methodological Issues

There have always been some difficulties for police organizations granting outsiders access to internal functional mechanisms because it carries the risk of 'occupational secrets' becoming subjects of public discussion. This has become somewhat easier as police have come to recognize the importance and usefulness of evaluation research. Generally, from a methodological point of view, three points deserve attention. First, police studies that are relevant for sociological interests can be roughly divided into basic and applied science. In basic science, the police scholar is more concerned with theoretical issues and less concerned with practical issues; to oversimplify the matter, the opposite might be said to describe applied science. Of course, both applied and basic police science make use of theory and discuss practical implications. Nevertheless, contemporary policing research is often presented along these theoretical or atheoretical lines (cf Cooper and Worrall, 2012).

Another issue considers who initiates and/or carries out the research in question. It might make a difference for methodological approaches, as well as for the broadness of the theoretical perspective, if studies are done by researchers coming from police or other government institutions or if they are carried out by university institutes. When police behavior is observed from the outside – that is, by university researchers – especially in respect to its effects on citizens, it is a matter of external description and analysis. When researchers work from within an agency – or in collaboration with an agency – the internal perspective of the field actors can also be reflected in the research on police. Both have value and pitfalls that must be taken into account when evaluating any policing study.

Furthermore, there is a clear difference between research *for* and research *about* the police. Usually the former tends to be preoccupied with practical issues or management problems that need a quicker turn-around than rigorous social science can provide. Although there quite often exist interesting sociological aspects within these perspectives, a genuine sociology of police and policing is coupled with research *about* the police (i.e., its meaning in relation to the law, the state, or as a social response to deviance in general). To the extent to which research is *for* the police it may depart from the usual methodological considerations of the sociological imagination.

Finally, the term 'police' might be a category that is too broad when it comes to empirical analysis and theoretical conclusions. Not only do national police systems embrace a variety of more or less specialized forces, with specific and sometimes unique ways of division of labor among units, but police organizations also refer to very different geographical ranges of jurisdiction. This situation becomes more complicated when nationally focused studies are compared on an

international level: for a full comprehension of the differences and convergences, national particulars have to be taken into account.

Future Directions of Theory and Research

As any perusal of advertisements for criminology and criminal justice faculty positions will reveal, policing scholars – and their scholarship – are in high demand. It is perhaps ironic that many of the research threads that existed in the 1990s and early 2000s, or even earlier, are still being pursued. For example, scholars continue to explore the privatization and militarization of the police, their impact on not only crime but on disorder, as well as police officer characteristics and their effect on discretion, including officer satisfaction, motivation, race, gender, etc. Such research questions are hardly new in policing scholarship; what is new is a very strict positivist approach to these subjects. Prior research, following Weber's lead, combined empirical methodologies with an explication of data often more at home in the humanities than the scientific method. For the last decade, and the foreseeable future, policing research seems intently focused on more exact specifications that can be 'objectively' measured.

This *uber*-focus on measurement and outcome has – somewhat ironically given the seeming adoption of strict positivism – led a number of police scholars to either ignore theory altogether or to plug it in post hoc. Some policing research has, however, followed the rules of basic science more closely, either testing or articulating theory. For example, Wilson's (1968) book *Varieties of Police Behavior* continues to guide policing research to a great extent, as does Klinger's (1998) territorial explanation for police behavior.

More prominently, policing theory borrows from mainstream criminology. This pattern is less a matter of the evolution of police science, and more a matter of academic training. Most of the theories in the 1960s and 1970s were based on institutional and systems theory, which reflected the sociological and political science training of their authors. Today, more and more, police scholars come from criminology and criminal justice programs, and this theoretical background dominates their writing. Social learning theory, low self-control, and social disorganization (inter alia) have all been implicated in the last 15 years to explain the behavior of police officers and of the police.

Because policing research is, to a great extent, driven by American research, and because that research is driven, to a great extent, by monies from the government in the form of grants, it is no surprise that the future of policing research is mirroring the future of policing itself. Namely, as police departments began to use problem-oriented and situational crime prevention approaches more regularly, work with community groups, and engage in evaluation research, they concomitantly work more closely with researchers – both in-house and externally. The real questions facing police scholarship for at least the next 20 years should ask how does the scientific method apply to policing? What is the role of technology in this process? And finally, what are the implications of these changes, not only for civil order, but for civil rights as well?

See also: Control, Social; Crime Policy in the United States; Crime, Sociology of; Criminal Justice, Sociology of; Law, Sociology of; Law: Economics of Its Public Enforcement; Policing.

Bibliography

Bernard, T., Engel, R., 2001. Conceptualizing criminal justice theory. Justice Quarterly 18, 1–30.
Bittner, E., 1991[1980]. The functions of police in modern society. In: Klockars, C.B., Mastrofski, S.D. (Eds.), Thinking about Police. McGraw-Hill, New York, pp. 35–51.
Black, D., 1980. The Manners and Customs of the Police. Academic Press, London.
Brodeur, J.P. (Ed.), 1995. Comparisons in Policing: An International Perspective. Avebury, Aldershot, UK.
Busch, H., Funk, A., Kauss, U., Narr, W.D., Werkentin, F., 1985. Die Polizei in der Bundesrepublik. Campus, Frankfurt am Main, Germany.
Cain, M.E., 1979. Trends in the sociology of police work. International Journal of the Sociology of Law 7, 143–167.
Cooper, J., Worrall, J., 2012. Theorizing criminal justice evaluation and research. Criminal Justice Review 37, 384–397.
Crank, J.P., 2003. Institutional theory of police: a review of the state of the art. Policing. An International Journal of Police Strategies & Management 26, 186–207.
Crank, J.P., Langworthy, R., 1992. An institutional perspective of policing. Journal of Criminal Law & Criminology 83, 338–363.
Ericson, R.V., Haggerty, K.D., 1997. Policing the Risk Society. Clarendon Press, Oxford.
Findlay, M., Zvekic, U. (Eds.), 1993. Alternative Policing Styles: Cross-Cultural Perspectives. Law and Taxation Publishers, Deventer, The Netherlands.
Jackson, P., 1989. Minority Group Threat, Crime, and Policing: Social Context and Social Control. Praeger, New York.
Kane, R.J., 2002. The social ecology of police misconduct. Criminology 40, 867–896.
Klinger, D.A., 1997. Negotiating order in patrol work: an ecological theory of police response to deviance. Criminology 35, 277–306.
Manning, P.K., 1977. Police Work: The Social Organization of Policity. MIT Press, Cambridge, MA.
Manning, P.K., 1978. The police: mandate, strategies, and appearances. In: Manning, P.K., Van Maanen, J. (Eds.), Policing: A View from the Streets. Goodyear Publishing Company, Santa Monica, CA, pp. 7–31.
Marx, G.T., 1988. Undercover Police Surveillance in America. University of California Press, Berkeley, CA.
Monjardet, D., 1996. Ce que fait la police (Sociologie de la force publique). La Découverte, Paris.
Reiner, R., 2000. Police research. In: King, R.D., Wincup, E. (Eds.), Doing Research on Crime and Justice. Oxford University Press, Oxford, UK, pp. 205–235.
Shearing, C.D., Stenning, P.C. (Eds.), 1987. Private Policing. Sage, Newbury Park, CA.
Skolnick, J.H., 1966. Justice without Trial. Wiley, New York.
Van Maanen, J., 1974. Working the street: a developmental view of police behavior. In: Jacob, H. (Ed.), The Potential for Reform of Criminal Justice. Sage, Beverly Hills, CA, pp. 83–130.
Walker, S., 2005. The New World of Police Accountability. Sage, Thousand Oaks, CA.
Wilson, J.Q., 1968. Varieties of Police Behavior: The Management of Law and Order in Eight Communities. Harvard University Press, Cambridge, MA.

Police Use of Force

Geoffrey P Alpert, University of South Carolina, Columbia, SC, USA; and Griffith University, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The police have authority to ensure public safety by enforcing laws. Although the core of police work is cooperation, officers are sometimes required to use physical force. The nature and types of force used by the police create a perception of their legitimacy among the citizens they serve. This article discusses the use of force by police and provides a context to understand its use and abuse.

Introduction

In our social world, we rely on the police to protect us and ensure our safety. We provide them with the authority to use force and thereby differentiate them from other civilian employees (see Bittner, 1970). Although the authority to use force is the necessary and defining characteristic of the police, it brings with it exceptional scrutiny and oversight. This article discusses the use of force by the police and helps us understand its contours. We start with a brief understanding of when police use force, turn to the current legal standards that authorize it, provide an explanation of the context of police use of force and what our research findings teach us, and conclude with a conceptual framework to understand why police–citizen encounters result in the use of force.

Although we know the police have the authority to enforce the law and ensure public safety, we do not always know how much force or what type of force is reasonable and when force used by the police is unreasonable and excessive. We must recall Egon Bittner's (1970) clarification more than 40 years ago that while the use of force is at the core of police work, the true skill of an officer is to accomplish the police mission of saving lives and protecting society while avoiding the use of force or using only a minimal level of force as situationally justified. Bittner's explanation holds true today, although officers face more serious threats from citizens than they did in the 1960s and 1970s (see Alpert and Dunham, 2004). Similarly, the actions of the police that might have been reported on the evening news are often seen on YouTube shortly after they occur. Our digital world has increased the speed and flow of information, but the use of excessive force has the same damning effect on society and loss of trust as it did decades ago. Over the years, psychological and technological innovations have helped the police negotiate with suspects and deal with them in ways that reduce the likelihood of using force to control them. Skills including verbal judo, violence reduction, and other useful interpersonal communication techniques have all helped reduce police use of force and the likelihood of suspect resistance and injuries (Police Executive Research Forum, 2012; Prenzler et al., 2013). However, if all else fails, the use of force is the officer's last resort to control a suspect's resistance to a lawful order. Beyond the psychological advancements, less lethal technologies, including oleoresin capsicum (OC spray) and conducted electronic weapons (CEWs), have become very popular in the twenty-first century and an important part of police work.

The use of force by the police plays a critical role in how citizens view them, earning or eroding public trust, and developing or limiting cooperation with citizens to provide information on crimes and criminals. It is human nature to respond negatively to misconduct, or the rare and shocking abuses of power rather than positively to the routine, appropriate, and numerous thoughtful acts that are performed most frequently by the police. Citizens want to have positive attitudes toward the police but it is perception rather than reality that often paints the picture that impacts the public's views and responses to the police. Managing that perception can be impacted greatly by understanding why, how, when, and in what situations police use force, as well as the level and type of force that is applied. The true picture of police use of force is not as interesting as the one painted by the abuses that create attention and publicity. Police use of force is a rare event, estimated to occur in approximately 1–2% of all encounters between the police and the public, and even the majority of those uses of force are minor and do not result in significant injuries (Walker and Katz, 2013). However, the percentage of encounters in which force is used will fluctuate depending on how force is measured and how encounters with citizens are recorded. In today's world, many police agencies define force only as actions beyond verbal orders, handcuffing, pat-downs, and 'come-along' holds. They also do not record all police–citizen encounters but limit reporting to the most serious ones, which are differentially defined. Although the number and percentage of incidents in which force is used may vary by accounting procedures, the impact of each remains a critical moment for police–public relations. Most use of force applications are in response to citizens' aggressive or physical refusals to obey orders, but determining when each use of force crosses the line from reasonable to excessive, from appropriate conduct to misconduct, is a difficult determination.

Determining Excessive Force

As important as, and more problematic than, determining the nature and extent of police use of force, is the determination of the force as reasonable or excessive. The standard for this distinction is set by the courts, and interpreted by agency policies, procedures, training, and supervision. Police

departments rely on the standards established by dicta from the US Supreme Court in its opinion, *Graham v. Connor*, 490 US 386 (1989). In the *Graham* case, the Court provides the general framework for assessing whether a particular use of force is legal under the Fourth Amendment. This, as most general standards found in Fourth Amendment precedent, operates through a balancing test of the nature and quality of the intrusion on the individual's rights against the countervailing governmental interests. The facts of the case are important to help understand the dynamics of the police–citizen encounter. Mr. Graham was suffering from an insulin reaction and went to a store to purchase orange juice. Because of the long lines and Graham's need to get a 'fix,' he quickly exited and headed to a friend's house. A police officer who was observing Mr. Graham's behavior of quickly entering and exiting the store became suspicious and conducted an investigatory stop. The interaction between an officer who formed suspicion based on his observations and a citizen who was suffering from an insulin reaction resulted in misunderstanding, misread cues, physical force, and resistance. The officer's actions gave rise to a use of excessive force claim against the police. In its decision, the US Supreme Court provided guidance to determine the appropriate level of force used by the police. Specifically, the court used a three-question test to measure the reasonableness of a particular use of force:

1. What was the severity of the crime that the officer believed the suspect to have committed or be committing?
2. Did the suspect present an immediate threat to the safety of officers or the public?
3. Was the suspect actively resisting arrest or attempting to escape?

The court provided more direction and declared (at 396): "the calculus of reasonableness must embody allowance for the fact that police officers are often forced to make split-second judgments – in circumstances that are tense, uncertain, and rapidly evolving – about the amount of force that is necessary in a particular situation." The court also stated that the use of force should be measured by what the officer knew at the time, not by the '20/20 vision of hindsight.' In the nearly two decades since the *Graham v. Connor decision*, courts have refined the three-prong test and adopted several additional factors in the formula to determine the objective reasonableness of police use of force. For example, the degree of the threat posed by the suspect based on size, strength, and numbers has become part of the equation. When officers are outnumbered or confronted by particularly powerful suspects, higher levels of force may be justified than when officers are confronted by a lesser threat (see *Sharrar v. Felsing*, 128 F.3d 810 (3rd Cir. 1997)). Courts may also consider the immediate availability of less lethal weapons (*Tom v. Voida*, 963 F.2d 952 (7th Cir. 1992)). The subject's history of mental illness or level of impairment due to alcohol or drugs may also be considered when analyzing the threat posed by the suspect and the response by the police (see *Krueger v. Fuhr*, 991 F.2d 435 (8th Cir.), *cert. denied*, 510 US 946 (1993); *Hunt v. County of Whitman*, 2006 WL 2096068 (E.D. Wash. 2006)). Although the factors may change or be interpreted differently, the language remains the same: an officer may use only that force that is both *reasonable* and *necessary* to make an arrest or apprehend a suspect. Anything more is *excessive force* (*Payne v. Pauley*, 337 F.3d 767 (7th Cir. 2003)). Understanding the legal strictures sets the stage to understand what we know about police use of force.

The Context of Police Use of Force: What We Know

Research has been conducted on police use of force for more than 40 years and the results show important information about the nature and extent of force used by the police as they attempt to control suspects or respond to resistance from them (Alpert and Dunham, 2004; Alpert et al., 2004). The quality and quantity of information on police use of force incidents recognized by researchers and policymakers have increased dramatically in recent years. Findings from this body of research illuminate methodological shortcomings including work conducted from the perspective that use of force events are static, and that the data have focused on the officer's point of view (Alpert and Dunham, 2004). Additionally, the conceptual frameworks used to understand police use of force have not been linked together properly (see National Research Council, 2003).

Despite acknowledged shortcomings, research has provided important information to guide our understanding of police behavior. Most of the social science research on police use of force has been conducted since the 1970s (Garner et al., 2002) and in the largest police departments (Alpert and Dunham, 2004). As Herman Goldstein (1990, 9) informs us, "The police came under enormous pressure in the late 1960s and early 1970s as they were confronted with concern about a rapidly rising crime rate, civil rights demonstrations, racial conflicts, riots, and political protests of the war in Vietnam." This interest in studying the police and their control tactics was basically generated by their linkages to social and political events. These questions and concerns about police behavior led to an increase in federal funding for police research and a corresponding improvement in ways to study them. In other words, the reform and research movements were linked inexorably. Prior to the increase in systematic research, most information on police violence was anecdotal, but helped the reform movement gain traction. According to Garner et al. (2002: p.707):

> The narrative accounts by independent researchers tended to emphasize the researchers' personal interpretations of police work and to highlight alleged and sometimes confirmed incidences of unusual, dramatic, illegal, or inappropriate behavior by officers or civilians. These descriptions and insights provide a valuable basis for generating hypotheses about the nature of force and the situations in which force is used. However, because they have been limited to personal observations of a few officers in a few precincts in a few jurisdictions and have relied on qualitative assessments of single observers, these personal narratives provide an incomplete and potentially inaccurate picture of normal contact between the police and the public.

The recent era of research on police use of force has included traditional social science methods, including analysis of existing police records; surveys of officers, suspects, and members of the public; interviews with officers and suspects; and observations of the police in action. Overall, our knowledge is limited and results are mixed but our concerns are fervent. In some studies,

we learn that police use force more often and at higher levels against members of minority groups than white citizens, while in others there are no differences. The use of force against people with mental deficiencies or under the influence of alcohol or drugs remains an important topic that is difficult to study and understand, and research findings concerning how officers respond to this population are mixed as well. We investigate the use of force in specific areas, especially high-crime and lower socioeconomic areas, because it occurs more frequently in those areas than low-crime, middle-class areas. Generally, we investigate the use of force by officers in the larger departments and try to determine if some of them use force or use higher levels of force more often than other officers, against identifiable groups, or in specific places. We look at the age, gender, ethnicity, and experience of officers; when and where they are working; and the social, political, and economic conditions of their assignments. We look at training and supervision of officers, as well as the policies under which they operate. We consider the age, ethnicity, social class, and criminality of citizens who resist the police and on whom force is used. Although our research goals are appropriate and conventional, our results are mixed and difficult to interpret (National Research Council, 2003). Our concentration of research on people and places is premised on successful social science research strategies, but our results do not provide policymakers with definitive information or 'best practices' to train or guide officers.

Although the empirical findings from our collective research may lack specific policy mandates for managers, there have been advancements in the ways we look at police use of force and how we try to understand it. First, we have learned to consider use of force events as live, high-stress, adrenalin-driven interactive events that must be understood during the beginning, levels of interaction, and end (Alpert and Dunham, 2004). Second, we must consider not only an officer's version of the event, but also any physical evidence such as a video or audio recording of the event, any observation by a witness, and the story told by the suspect. Here we learn that how the story is told becomes critical in its interpretation and understanding (Rojek et al., 2012).

A significant advancement in use of force research is the improvement of research on the interactive model of police–citizen exchanges. We recognize that the goals of a police officer are to define and control a situation, and we agree that a citizen will likely cooperate with the officer unless the citizen believes the officer is going to take action that is contrary to his or her interest. When that occurs, citizens may resist verbally, with and without success. However, if neither actor acquiesces, both officers and citizens can act in ways that frustrate the other, giving rise to a subset of interactions that can escalate the situation until one of the actors resorts to force or physical resistance. Although we have not learned exactly why some people act differently than others, we do know that the goals and objectives of the encounter are different for the police and the citizen. Because the behaviors in a police–citizen encounter are not static but change as the event unfolds, it is necessary to understand more thoroughly how police and citizens respond to cues, other forms of nonverbal behavior, and the actions and reactions of the other actors, including witnesses.

Limited research has been conducted to capture the significance of the interaction sequences of police–citizen encounters. Alpert and Dunham (1997) developed the Force Factor, a measure of the level of police use of force relative to the level of suspect resistance. This interactive approach has been extended not to look at just the highest level of force and resistance but also to investigate how each step of the interaction influences the action (and reaction) of the other actor. The stages or frames of an encounter must be understood so officers can learn how best to react to a suspect and attempt to de-escalate the situation (Police Executive Research Forum, 2012; Prenzler et al., 2013). The examination of the frames and processes involved in police–citizen interactions will develop a much clearer picture of these very closely watched police–citizen events.

A typical difference between officer and suspect versions of an event can be understood by the following scenario (Alpert and Dunham, 2004). A suspect who claimed he was taken to the ground by the police who grabbed his hair and smashed his head onto the cement, causing serious injuries, files a complaint. The arresting officer answers by saying while he was chasing on foot a fleeing suspect, he tackled him and as they were falling to the ground, the suspect hit his head on a table causing the injuries. As there were no witnesses and the medical doctor reported that the injuries were consistent with either version, it is not clear what really happened. As in most of these 'he said, she said' scenarios, the officer's version becomes the official one (Noble and Alpert, 2009).

In most cases that have no strong evidence, such as a video recording or unbiased witnesses (Rojek et al., 2012), suspects report that the level of force the officer used against them was more than reported by the officers and was more than necessary. They also report quite often that they were not resisting at the level reported by the officer. Their version relates to issues of procedural justice. Juxtaposed to their version of the events, officers report, for the most part, that they used minimal force to control the suspects and do not mention using force after a suspect is under control. It is not likely that an officer will admit to using excessive force against a suspect. The officers report that their behavior followed policies, procedures, and the law. Not surprisingly, the officers often report a suspect using great and sometimes superhuman strength to resist and that the suspect would not stop fighting, which justifies their use of force and perhaps increased level of force as necessary and reasonable. We have learned that what we discover about these interactions must include information beyond a single version, whether the officer, suspect, or eyewitness. Recently, understanding police use of force has shifted to involve the availability and impact of technologically advanced weapons.

Technological Advancements

The development of OC spray and CEWs has changed the landscape of police use of force. Before these new tools became available to the police, officers had only their interpersonal skills, hands, fists, batons (and other weapons of convenience), and firearms to take suspects into custody. As technologies advanced, officers became better prepared to distract or use different types of pain compliance devices and types of force to

control suspects. Along with the innovative tools, came new problems (Smith and Alpert, 2000).

The uses of OC spray and CEWs by the police have similar histories. As OC spray became a popular tool for police in the 1990s, its use generated concerns about its abuse and lethality. Several advocacy groups, including the American Civil Liberties Union and Amnesty International, claimed its use was 'causing' severe injuries and in-custody deaths (American Civil Liberties Union of Southern California, 1995; Amnesty International, 1997). These claims encouraged research that investigated the safety and effectiveness of OC spray and resulting injuries (see Kaminski et al., 1998). Overall, the use of OC spray was determined to be quite effective (see Smith and Alpert, 2000) and its use was correlated with a reduction of injuries to both officers and suspects when compared to traditional 'hands-on' police use of physical force.

During the past few years, there has been a similar pattern of concern raised about the use of CEWs. Although there has been an explosion in the use of CEWs, research findings are just beginning to inform policymakers and trainers about best practices and potential problems (Alpert and Dunham, 2010; Police Executive Research Forum, 2011). In the majority of CEW applications, the application is successful as officers apprehend suspects without significant injuries to either them or the suspect. This leads to important claims that the availability and use of the CEWs assists in the safe apprehension of suspects. Although there is little doubt that the CEW is a popular and useful tool for the police, many questions remain. First, where should it be placed on a use of force continuum, or under what conditions is it an appropriate tactical choice? Second, how many applications are reasonable before turning to another tool or tactic? Third, does the application of electric shock impair the cognitive abilities of a suspect after the electricity is turned off? Fourth, does multiple application or prolonged use of the CEW cause injury or death in certain situations? Fifth, do officers resort to using a CEW too quickly and too often in an encounter as a substitute for lower levels of force? And sixth, does the ease of use or level of pain created by the weapon create an overreliance on it in lieu of hands-on policing, or the 'lazy cop syndrome?' Although many questions about the use of a CEW remain, CEW use continues to help the police manage encounters with suspects.

Research has shown that there are many ways to manage a police–citizen encounter. In a perfect world, there would be no need for the use of force, as citizens would comply with all requests from the police, and those requests would not put citizens at risk. As the world is not perfect and police–citizen encounters can result in force, we need to understand the reasons that there is conflict between the actors. The final section of this article proposes a framework to understand these interactions and why some result in the use of force.

Authority Maintenance Rituals: A Conceptual Framework to Understand Police–Citizen Encounters

The behavior of a police officer is greatly influenced by the nature of his or her interactions with citizens. Similarly, these citizens are likely to react to the officer's actions, comments, cues, and demeanor. Therefore, considering the interactive nature of police–citizen encounters is critical if the behavior of both officers and suspects is to be understood. Although research findings help us appreciate discretion and specific actions, studies have been generally limited to attitudes, personality traits, and demographic and situational characteristics of people and places. Unfortunately, theoretical and conceptual linkages among these behaviors are noticeably absent. We developed the 'authority maintenance theory' to explain the interactive processes involved in police–citizen encounters, with a focus on those that result in the use of force (for a discussion of the developed theory, see Alpert and Dunham, 2004). The term 'authority maintenance' captures the crucial role that authority has in police–citizen interactions, and stresses the officer's ritual to maintain an authoritative edge during any confrontation.

The authority maintenance theory is premised on theoretical strands from Blumer's (1937) symbolic interaction and Homans' (1958) exchange theory. Since these early concepts were developed, social scientists have integrated the concept of power (Emerson, 1972) and reciprocal exchanges (Molm et al., 1999) into these relationships. Sykes and Clark (1975) developed the deference exchange theory to study all police–citizen interactions, which is a necessary element of the authority maintenance theory. Police–citizen encounters turn on the officer's exercise of authority (coercion) and the citizen's submission to that authority (deference or resistance). Deference is only one response to authority. Other responses exist and range from submission to violent resistance. Deference exchange theory focuses only on deference, but the authority maintenance theory includes all behaviors across the deference–resistance continuum. Our theory explaining police–citizen interactions and their outcomes of force includes seven propositions (these propositions are taken from Alpert and Dunham, 2004, where they are explained fully) that account for the unique nature of police–citizen encounters, including the concepts of the asymmetry of authority and the breakdown of reciprocity.

Propositions

1. Police–citizen encounters must be understood as an interaction process, rather than as discrete events.
2. Police–citizen encounters are unique types of social interactions because the major criterion regulating the interaction is the balance of authority or power among actors.
3. Police–citizen encounters are more asymmetrical with respect to authority than most other types of interactions.
4. Police–citizen encounters are a unique type of social interaction because expectations and behaviors often violate the principle of reciprocity.
5. Officers respond to goal blockage with varying degrees of regulation depending upon the type of blockage.
6. Citizens respond to goal blockage with varying degrees of resistance.
7. The resistance/force sequence escalates until one party changes their expected goals voluntarily or involuntarily.

An Integrated Theory

The theory of authority maintenance is based on the notion that police–citizen encounters are interactive and asymmetrical with respect to authority. In other words, behavioral cues and responses to these cues are guided by the respective actor's power and status as well as his or her decision-making abilities and experiences (Alpert and Rojek, 2011). At the heart of the authority maintenance theory is the third proposition on the asymmetrical nature of police–citizen contacts. This supports the notion that an officer will use greater force when his or her authority is significantly challenged by the actions of a suspect. Future research should incorporate perceptual measures of authority from police officers that include the nonverbal gestures, dress, and demeanor of citizens. Debriefing officers after use of force incidents to describe their perceptions would provide important information on how to best operationalize the concept of authority and manage decision-making skills.

The authority maintenance theory goes beyond use of force, as suggested by the principle of reciprocity. Research is needed to identify the factors that trigger the breakdown of reciprocity in police–citizen encounters and result in negative outcomes, including the use of force. The social world of policing is composed of communication through ongoing processes or messages, many of which are threatening to other actors. During encounters, police and citizens must interpret and respond to these messages and incorporate them into their decisions. Peter Manning's (1977) use of 'drama,' or the presentation of behavior that sends a message to another actors, sets the stage for police training and role-playing.

Conclusion

Our effort to understand why, when, and under what conditions police use force is underscored by the critical role it plays in society. We have learned from previous research that there are more questions than answers, but we are learning how to investigate properly many of the issues raised over the years by researchers and policymakers. Police–citizen encounters that result in the use of force can be understood by a conceptual and theoretical framework that we call authority maintenance. This framework provides a way to comprehend why some people make certain decisions and respond in certain ways. Although this theoretical framework provides a way to look at police–citizen encounters, it does not provide all the answers. We are still searching for the reasons why some officers and suspects make decisions that challenge, coerce, or threaten other actors, and why the outcomes of these interactive encounters cannot be predicted successfully.

See also: Crime, Sociology of; Criminal Justice, Sociology of; Criminology: Development and Current Status of the Discipline; Fear of Crime; Guilt; Police, Sociology of; Policing; Wrongful Convictions.

Bibliography

Alpert, G., Dunham, R., 1997. The Force Factor: Measuring Police Use of Force Relative to Suspect Resistance. Police Executive Research Forum, Washington, DC.

Alpert, G., Dunham, R., 2004. Understanding Police Use of Force: Officers, Suspects, and Reciprocity. Cambridge University Press, Cambridge, NY.

Alpert, G., Dunham, R., MacDonald, J., 2004. Interactive police–citizen encounters that result in force. Police Quarterly 7, 475–488.

Alpert, G., Dunham, R., 2010. Policy and Training Recommendations Related to Police Use of CEDs: Overview of Findings From a Comprehensive National Study. Police Quarterly 13, 235–259.

Alpert, G., Rojek, J., 2011. Frontline Police Officer Assessments of Risks and Decision Making During Encounters with Offenders. Centre of Excellence in Policing and Security, Brisbane, Australia.

American Civil Liberties Union, 1995. Pepper Spray Update: More Fatalities, More Questions. American Civil Liberties Union, Los Angeles.

Amnesty International, 1997. USA: Police Use of Pepper Spray – Tantamount to Torture. Amnesty International, New York.

Bittner, E., 1970. The Functions of the Police in Modern Society: A Review of Background Factors, Current Practices, and Possible Role Models. National Institute of Mental Health, Center for Studies of Crime and Delinquency, Chevy Chase, MD.

Blumer, H., 1937. Social Psychology. In: Schmidt, E. (Ed.), Man and Society. Prentice-Hall, New York, pp. 144–198.

Emerson, R., 1972. In: Berger, J., Zelditch, M., Anderson, B. (Eds.), Exchange Theory, Part 1: A Psychological Basis for Social Exchange. Houghton-Mifflin, Boston, MA, pp. 38–57.

Garner, J., Maxwell, C., Heraux, C., 2002. Characteristics associated with the prevalence and severity of force used by the police. Justice Quarterly 19, 705–746.

Goldstein, H., 1990. Problem-oriented Policing. McGraw-Hill, New York.

Homans, G., 1958. Social behavior as exchange. American Journal of Sociology 62, 597–606.

Kaminski, R.J., Edwards, S.M., Johnson, J.W., 1998. The deterrent effects of oleoresin capsicum on assaults against police: testing the velcro-effect hypothesis. Police Quarterly 1, 1–20.

Manning, P.K., 1977. Police Work: The Social Organization of Policing. MIT Press, Cambridge, MA.

Molm, L., Peterson, G., Takahashi, N., 1999. Power in negotiated and reciprocal exchange. American Sociological Review 64, 876–890.

National Research Council, 2003. Fairness and Effectiveness in Policing: The Evidence. Committee to Review Research on Police Policy and Practices. Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

Noble, J., Alpert, G., 2009. Managing Accountability Systems for Police Conduct. Waveland Press, Long Grove, IL.

Police Executive Research Forum, 2011. 2011 Electronic Control Weapons Guidelines. Police Executive Research Forum, Washington, DC.

Police Executive Research Forum, 2012. An Integrated Approach to De-escalating and Minimizing Use of Force. Police Executive Research Forum, Washington, DC.

Prenzler, T., Porter, L., Alpert, G., 2013. Reducing police use of force: case studies and prospects. Aggression and Violent Behavior 18, 343–356.

Rojek, J., Alpert, G., Smith, H., 2012. Examining officer and citizen accounts of police use of force incidents. Crime and Delinquency 58, 301–327.

Smith, M., Alpert, G., 2000. Pepper spray: a safe and reasonable response to suspect verbal resistance. Policing: An International Journal of Police Strategy and Management 23, 233–245.

Sykes, G., Clark, J., 1975. A theory of deference exchange in police civilian encounters. American Journal of Sociology 81, 584–600.

Walker, S., Katz, C., 2013. The Police in America. McGraw-Hill, New York.

Policing

Mathieu Deflem, University of South Carolina, Columbia, SC, USA
Samantha Hauptman, University of South Carolina Upstate, Spartanburg, SC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.H. Skolnick, volume 17, pp. 11535–11541, © 2001, Elsevier Ltd.

Abstract

Policing organizations and activities are an important component in the wider constellation of crime control and criminal justice in society. This article reviews fundamental aspects of policing in terms of the following categories: the history of police systems; the functions and organization of policing; the relationship between policing and the community, including its cultural dimensions; the styles and strategies of policing and the role of technology; and the impact of globalization on the form and content of policing.

Policing refers to the whole of structures and processes that are associated with the control of crime and the maintenance of order in society. The history of the police function is not stable over time, and policing models and activities are not identical across societal contexts. The use of the term policing is itself not universally valid as other languages have no equivalent one-word expression for the activities associated with the police function on the basis of the verb associated therewith, and instead exclusively rely upon the noun police or a similar term. Additionally, it should be noted that the concept of police in the social sciences is conceived of as an institution (police in the singular), whereas in everyday language the police is more typically understood as referring to the whole of the body of men and women who occupy a particular position (police in the plural). Despite conceptual and associated institutional variability, trends can be observed toward the delineation of policing in association with crime and order objectives. The so-called private policing refers to security functions that are fulfilled, in more or less legally restricted ways, by privately owned companies and will not be further considered in this article.

Sociologically conceptualized within the broader field of the study of social control, defined as the societal mechanisms involved with the definition of and response to deviance and/or crime, the social reality of policing has been explored in relation to a variety of manifestations and patterns of change and continuity. The majority of insights into the social form and function of policing has come from sociology, with related issues having been explored in history, political science, anthropology, and other social and behavioral disciplines. The specialized treatment of police in the emerging fields of criminology and criminal justice has brought about additional insights concerning the administration of police as an applied field of study.

History

Historically, most societies have had some form of police institutions and policing activities for a long time, developing in relation to changing conceptions of government and other forms of authority (Bayley, 1985; Fogelson, 1977; Manning, 1977; Palmiotto and Unnithan, 2011; Skolnick, 2001; Walker, 1977). For much of early civilization, policing practices were informally organized and largely considered a private matter that consisted of little more than groups of citizens banding together to perform local watch duties as a means to protect their community from outsiders. As societies evolved and developed from simple mechanical and typically small communities, in which members are collectively responsible for the adherence to shared values, into complex organic and much larger states, in which specialization and interdependence become prevalent, a range of more formal and institutionalized forms of social control developed.

In the West, the last few centuries of the Middle Ages witnessed a period in which several European kings began to appoint public officials to formally assume an administrative role in the enforcement of laws. Yet policing duties were still mostly assigned to ordinary citizens, who remained largely responsible for their own protection and well-being. It was not until communities began to become more organized in the form of cities and national governments, along with complex social changes in the social fabric, that the organization of formal policing practices and institutions emerged.

Turning to modern times that began with the rise of the national state, the police function not only became more fully institutionalized, it also and importantly changed in meaning. Max Weber (1922) famously defined the state in instrumental terms with respect to the monopoly that is claimed in a political community over the means of legitimate physical coercion, including military and police as the two major, respectively externally and internally directed, components. However, the conceptual connection between police and politics, not to mention the deceptively obvious common etymological origin of both words, should not lead to overlook the fact that police and politics are variably (dis)connected (Deflem, 2002).

In the European context of national states of a strongly autocratic nature during the eighteenth and nineteenth centuries, police were closely aligned with the attempts of autocratic regimes to maintain power. Thus, policing primarily implied activities associated with the control and suppression of political opponents, such as the growing number of democrats, socialists, and other proponents of ideological political systems associated with ideals of democracy and political justice. As national states underwent processes of gradual democratization, the police function expanded and changed qualitatively.

Within the Western world and increasingly, with the march of globalization, also in other nations affected by developments or pressures originating there from, the police function has rationalized over time, concomitantly with the democratization of national states, in the form of the modern bureaucracy (Deflem, 2002). Manifesting a broader trend of societal rationalization toward the adoption of efficiency standards at the exclusion of other principles of conduct, bureaucratized policing practices are qualitative, different from premodern forms with respect to both means and objectives. Specifically, the means of policing became geared toward the reliance upon purposive-rational standards, such as scientific methods of criminal investigation and other aspects of police technique, while the objectives of policing became gradually detached from the political goals associated with the might of national states in favor of the development of distinctly criminal enforcement objectives.

Thus, modern policing is not only increasingly based on efficiency principles, it also becomes gradually detached from politics. Although the separation of police from politics is never absolute, the modern police function has rationalized in this particular manner in those societies that are democratically organized and that have achieved a modicum of peace. Conversely, under conditions of an absence of or decline in the democratic qualities of government in states or under conditions of warfare and similar occurrences of extreme societal upheaval, police institutions face pressures to once again align more closely with politics.

Functions and Organization

The functions of policing are not exclusively related to the understanding of police as law enforcement (Bittner, 1990; Manning, 1977; Palmiotto and Unnithan, 2011). Although many contemporary societies have organized police institutions that are formally referred to as law enforcement agencies, policing cannot rightly be conceived in terms of the enforcement of law alone. In the centuries leading to the formation of European national states, the policing function was very broadly conceived on the basis of a system of a concept of government that included many different objectives related to well-being, welfare, and the health and happiness of people (Deflem, 2002). Policing thus included a wide variety of social roles intended to regulate human interaction and maintain the social (rather than merely the legal) order in matters as diverse as crime and murder, health, fire, education, religion, and trade.

Over time, policing gradually became restricted to criminal enforcement duties, while other social security functions were delegated to specialized institutions no longer conceived of as police, such as fire brigades and food inspection agencies. The contemporary, more delineated conception of police in terms of crime control relates to the fact that police are authorized to enforce formal sanctions against acts of deviance that have been explicitly codified in systems of criminal law. Controlling crime, only formal police institutions are allowed, under specified circumstances, to rely on force, including deadly force. Broader policing functions, however, remain even in modern times, such as in the case of the regulation of traffic and the delivery of other social services.

A high degree of professionalization is an additional characteristic of modern police (Walker, 1977). Uniformed police officers have become the typical representatives of policing in agencies that are organized as paramilitary, hierarchically ordered organizations. The degree of the military nature and style of formal police institutions to this day vary across societies. In continental Europe and nations with policing systems imported there from, police can be very militaristic and highly centralized as a close ally of the military. Prototypical of these organizational forms are the French force of the *gendarmerie* and similar such military-style policing organizations in other continental nations. In recent times, there has been a trend in these societies toward a demilitarization of policing in favor of an elaboration of civilian police models, especially at the local level, which coexist with earlier developed national police systems.

Other nations, notably the United Kingdom and the United States, have a police system that originates from distinctly civilian forces that were primarily organized on a local level (Dempsey and Forst, 2014; Mawby, 1999; Palmiotto and Unnithan, 2011; Skolnick, 2001). In the United States, the sheriff is prototypical of this form. In the United Kingdom, professional police organizations began to develop from the mid-eighteenth century onward alongside of important societal changes, especially industrialization, increasing crime, and urban social disorder. In 1829, British Parliament passed the London Metropolitan Police Act, which created a civilian police force for the city of London. Led by Home Secretary, Sir Robert Peel, the new professional force was based on the notion that a cooperative relationship should exist between the police and the community they are meant to serve. Although initially seen as an external and hostile force, the new police organization gradually found acceptance among the citizenry and the system was adopted throughout the country.

In the United States, policing in colonial America embraced a night-watch system that was borrowed from the British and first established in Boston in 1636. Although the system initially spread to other large cities along the east coast, eighteenth-century developments of industrialization, immigration, and an upsurge of incidents of disorder, especially of crimes related to a general increase in mobility, brought about an elaboration of police institutions at different levels of government (local, state, and federal) without a national police that has jurisdiction over the whole of the country. Instead, local policing in the United States is executed at the municipal and county levels, administering social control over a limited jurisdiction in view of a variety of tasks, including service, order maintenance, crime control, and peacekeeping functions. In all the US states except Hawaii, state police agencies have jurisdiction over the whole of their respective state, with special concern for highway patrol, criminal investigation, and local law enforcement assistance. At the federal level, an increasing number of US law enforcement agencies has come to exist, each with narrowly defined and highly specialized enforcement tasks.

Community and Culture

Intimately related to its social control function, the police enjoys a special relation to the community of citizens it is

meant to protect and the cultural dimensions of both the police force and society at large. To some extent, police institutions will develop a structural position and cultural disposition that is uniquely tied to its function and organization, but at the same time policing will reflect broader societal characteristics of the structural and cultural context (Bayley, 1985; Scott, 2009; Skolnick, 2001).

With respect to the relation of policing to the wider community, police officers can be perceived in variable ways as either more distant from or more closely aligned with the citizenry. This relative degree of distance and proximity is influenced by structural characteristics, such as the military nature and degree of centralization of the organization of formal police institutions, by culturally conditioned attitudes on the proper and legitimate role of police in society, as well as by situational circumstances related to personal and group experiences with crime and crime control.

Because much police work, especially in the realm of crime control, by necessity must rely on information and other forms of cooperation from the public, it is critical for policing to develop and maintain fruitful relations with members of the community. For this reason, there has been a trend in the past few decades toward the development of so-called community policing strategies that are explicitly aimed at fostering close relations between police and members of the community (Dempsey and Forst, 2014; Greene and Mastrofski, 1991). Unlike traditional policing models in which the police officer is conceived of as a professional whose expertise can create a detachment from the public, the objective of community policing is to build an equal partnership between the police and the community on the basis of mutual trust and support. Beyond a mere improvement of public relations, community policing requires a problem-oriented approach that is based on the specific concerns that exist in a community on diverse matters ranging from crime control to safety issues and various aspects of well-being. The community is thereby ideally involved in both proactive and reactive ways. Modern proactive policing strategies, especially in the United States and the United Kingdom, have witnessed an expansion in recent years in the form of intelligence-led policing, whereby decisions for police to intervene rely upon the assessment of crime risks (Ratcliffe, 2008). Related prevention-oriented models can involve broken-windows policing techniques that adopt a zero-tolerance approach toward minor infractions in order to prevent major crimes from occurring (Kelling and Coles, 1996).

Related to the community structure, policing poses interesting issues in terms of the cultural characteristics of the society in which police institutions are established as well as with respect to the development of a police culture with its own distinctive characteristics (O'Neill et al., 2007). Cultural dimensions of the community influence the effectiveness of policing inasmuch as a degree of legitimacy is by definition required of all forms of authority. The police profession has in this respect typically enjoyed a less favorable standing than the military, especially in symbolic respects, as police officers are in many societies across space and time typically held in less esteem than military soldiers. Additionally, culturally differentiated subsections of society may hold sharply different perspectives of policing. Most notable in this respect are the often-strained relations between police and members of minority communities, such as recently arrived immigrants and a variety of diverse ethnic, religious, and otherwise culturally defined groups.

Among the most striking examples of cultural issues of policing, mention can be made of the organization of policing in the colonial context, the policing of racial minorities, such as African-Americans in the United States, and the policing of minority groups of Arab descent and/or Islamic faith across many parts of the world since the terrorist events of 11 September 2001. As a result, efforts are often made to increase the diversity within police agencies in order to gain a better understanding of minority populations and foster an open dialogue with a diverse variety of community members (Dempsey and Forst, 2014). Such initiatives also seek to address the discriminatory nature of police work that has been found to exist with respect to cultural and other relevant characteristics (Black, 1980).

The cultural aspects of police organizations themselves have been found to be sufficiently specific and distinct from the wider societal context that it is useful to speak of a police culture or a culture of policing (O'Neill et al., 2007; Palmiotto and Unnithan, 2011). The typical characteristics of the culture of policing relate intimately to relevant functions of social control. Police officers may develop an 'us-vs-them' mentality in their dealings with the public, due in large part to the fact that police are typically only called upon when a negatively defined situation occurs, thereby exposing them to the effects of criminal conduct, victimization, violence, and other problematic occurrences. As police agents are also uniquely authorized to use force and, at once, not necessarily treated with respect and understanding, citizens do not always display positive sentiments toward them. In response, a strong sense of social solidarity and cohesiveness promotes an independent police culture segregated from the rest of society. Among the possible negative implications, corruption and other forms of police misconduct may exist without effective ways to redress the problems that are posed (Corsianos, 2012; Sherman, 1978). The infamous code of silence that characterizes many police agencies may additionally prevent accountability.

Strategies and Technology

Besides cooperation with the citizenry, policing activities rely upon a wide range of strategies and technological means that involve different aspects of police work (Dempsey and Forst, 2014; Palmiotto and Unnithan, 2011). In matters of method, various strategies and styles of policing can be differentiated, such as proactive and reactive policing, and overt and undercover police work. In pursuit of investigative objectives, police agencies rely on a number of methods to supervise, control, and gain access to the community. Traditionally, the agents of formal police agencies relied upon minimal tools, such as a nightstick or a baton, and could not maintain communications with one another outside the police station. Advances in technology, in particular the development of police radio systems, would later allow agents to build a communications network. Depending on the style of policing, officers were armed with a variety of weapons. In terms of patrol in the

community, the development of methods such as foot patrol and vehicular pursuit are important elements that shape the variable ways in which policing can be effectively conducted.

Police agencies were historically and are until this day typically organized in distinct locales such as police stations that are responsible for specified jurisdictional areas typically referred to as police precincts. Police can then be contacted and called upon by citizens by such means as police call boxes or emergency phone numbers. Citizen access to police is highly variable across societies and over time and has been found to be a critical component of accountability in policing, especially in democratic societies. In response, policing methods have been developed that rely on rapid-response teams, assisted by two-way radio systems and the extensive use of police cars. Technological advances in telephone technology, particularly the widespread availability of cellular phones, have in recent years brought about a considerable increase in the number of calls to police.

Reactive strategies of policing are complemented by proactive methods meant to prevent or deter crime. The elaboration of proactive methods has gradually been accelerated through the development of methods of citizen support, such as neighborhood watch, seeking to prevent crime from occurring. Additionally, the reliance on technically advanced systems of data gathering and investigative aids has contributed to policing strategies of surveillance that are widely dispersed throughout the society. The installation of video cameras in streets and buildings is the prototypical manifestation of proactive surveillance in many contemporary societies.

Besides security tools and technologies of transportation, policing has greatly been influenced by the development of systems of criminal identification (Deflem, 2002). Originally, modern police organizations relied upon photographs and descriptions of the facial and other physical features of criminal suspects. From the late nineteenth century onward, the use of fingerprints, which were held to be unique and permanent, became standard practice among police. Fingerprint identification data were numerically expressed so that they could be shared among police by means of telegraph. Computerization has eventually brought about digital forms in which data can be stored and shared by means of electronic fingerprinting techniques. In recent decades, DNA identification has been added as the technically most advanced method to identify suspects.

The rise of digital-information computers has brought about important changes in policing (Manning, 2008; Dempsey and Forst, 2014). Many modern police agencies utilize computers at various levels of operations, giving police organizations the ability to tailor technology to supervise the community by means of a wide array of surveillance tools. Additional advanced technologies in contemporary policing include computerized reports, both at police stations and in police vehicles, and crime mapping technologies such as geographic information systems of a variety of variables associated with criminal incidents and crime patterns.

The use of video cameras and other data-collection techniques adds to the intrusive qualities of policing that have historically been most typically associated with covert methods relying on undercover agents (Marx, 1988). Undercover tactics are applied in a variety of settings, typically involving crimes of a hidden nature, such as illegal drug use, or crime taking place in specific communities such as the world of organized crime. The multiple goals of undercover policing include intelligence gathering, crime prevention, and the infiltration of suspect groups. Beyond its investigative objectives, undercover police work involves deception, thus making it, by its very nature, ironic from the viewpoint of the common perception of uniformed agents of social control. Undercover and proactive means of policing have been found to correlate inversely with traditional covert means of policing, the number of restrictions that are placed on the legitimate use of particular reactive methods, such as deadly force, and the extent to which police can legitimately resort to search and seizure procedures without due process and judicial oversight. Among its problematic implications, undercover policing may ironically contribute to an increase in crime by creating opportunities for criminal conduct to occur.

Globalization

Variable in function and organization over the course of history, policing also differs across national states and other delineated jurisdictions in relation to important societal characteristics such as political, economic, and cultural factors. At the same time, general trends and patterns can be observed as well (Mawby, 1999). Globally, the centralized structure of policing is the most common organizational style, existing in the democratic countries of continental Europe but also in former totalitarian regimes such as the Soviet Union. Strikingly, national states with an autocratic past, such as France and Germany, tend to continue to rely on policing institutions in the modern era that can be very similar in form and organization to their historical predecessors.

Developing nations, including countries in Asia and Africa, as well as traditionally Islamic countries employ centralized police systems whereby a central police headquarters, typically located in the nation's capital, has command over all other, subordinate police units (Sadykiewicz, 2005). The centralization of police power can in variable ways coexist with a likewise centralized nature of the accompanying political and legal systems. A decentralized policing structure, conversely, divides enforcement duties among police units at local and higher levels of jurisdictional authority without the existence of a national police. The US system of law enforcement typifies the decentralized policing model, whereby hosts of local, state, and federal agencies operate independently and often have specialized agencies that serve a variety of functions. Considered globally, police decentralization is uncommon except in most industrialized English-speaking countries. Police decentralization is in such national cultures also often associated with the safeguarding of democracy as a political ideal and the preservation of individual rights of liberty, such as the freedom of movement and the free expression of ideas.

In democratic societies, police agencies are important and arguably the most visible institutions in charge of guarding social order, aiding in the socialization of the citizenry, and serving as both moral and legal social control agents that operate under prevailing jurisdictional laws (Bayley, 2005; Sklansky, 2008). Harmonizing with the understanding of

democracy as involving an intimate relation between government and public, the democratic role of police assumes a balance between the authority to enforce laws, on the one hand, and the accountability to the public, on the other. At the same time, police must be responsive to the needs of society while working toward a cooperative relationship with the community. Democratic police organization is therefore often associated with community policing and similar styles of police involving support of the public.

The democratization processes that can be observed in recent decades in several parts of the world have brought out the relative independence of police systems. The transition to democratic systems of policing has indeed proven to be challenging for countries with vastly different political systems, such as the former Soviet nations and formerly repressive police states like South Africa and Iraq. The implementation of democratic principles, where police agencies were once highly politicized to exist as a repressive tool of the state, impedes the support of the public. In cases where there is a lack of financial or technical resources, the likelihood of success to establish accountable systems of police is compromised as well.

The growing interdependence among national states and otherwise confined jurisdictions has historically led to the development of a variety of international policing practices (Deflem, 2002). With the growth of national states in Europe, international policing was initially developed for political purposes to maintain autocratic rule, but gradually began to focus on distinctly criminal enforcement tasks. By the early twentieth century, police agencies of various nations began to develop permanent cooperation structures and the bilateral and multilateral levels, including the formation of Interpol as the International Criminal Police Commission in 1923. In the drastically changed international constellation of the post-World War II era, international policing expanded greatly on a variety of concerns, such as the international rendition of fugitives, the global drug trade, border control, and the importation and exportation of police techniques across national boundaries.

International policing activities take on multiple forms, ranging from very informal to highly structured methods (Casey, 2010; Deflem, 2002; Deflem and McDonough, 2010). Unilateral international and otherwise *trans*-jurisdictional strategies tend to be preferred as police agencies can thereby maintain control over their operations and the proper ways of execution. Cooperative efforts among police at various jurisdictional levels are typically initiated on a need basis, in order to successfully complete a specific investigation, and tend to be limited in terms of the number of participating agencies from other jurisdictions. Organizations of multilateral cooperation, such as Interpol and the European Police Office, can rely on a large number of partners from many different nations, but they are strictly collaborative in nature and do not involve the creation of a supranational police force.

Over the past decade, no matter has affected policing more than the focus on terrorism that has been propelled, more than by any other incident, by the events of 11 September 2001 (Deflem, 2010). In the wake of those attacks, police agencies in the United States as well as many other nations in the world have drastically enlarged and reoriented enforcement objectives toward the suppression of terrorist groups and individuals. In the United States, for example, the Department of Homeland Security has been established to coordinate a multitude of security agencies involved with protecting the borders and security of the country, most notably the newly created US Immigration and Customs Enforcement, which has begun to rival the Federal Bureau of Investigation as the premier federal American law enforcement agency. The policing of terrorism as a criminal law enforcement matter complements other counterterrorism strategies, especially those that are of a military nature and oriented toward the protection of national security interests under the heading of the so-called War on Terror.

In some nations, counterterrorism policing is tightly controlled by centrally organized governments in the furtherance of political power. In many democratic nations, however, police agencies are engaged in counterterrorism efforts on the basis of a depoliticized understanding of terrorism as a form of criminal conduct. Unlike military and related intelligence organizations, police agencies can rely upon extensive networks of international cooperation that are formed irrespective of the differences that may exist among nations in political and other respects.

With an eye toward the future of policing across the world, it can be postulated that policing will unquestionably remain an important component of crime control and order maintenance. Even as globalization processes continue to expand in political, economic, and other respects, policing systems can be expected to continue to exhibit nationally and regionally variable characteristics. Moreover, as the world as a whole and its various national states and other constituent parts can be assumed to further increase in scale and complexity, policing in the contemporary era of globalization will continue to expand and entail multiple forms and strategies in an increasingly intricate manner.

See also: Crime Policy in the United States; Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices; Crime, Sociology of; Criminal Justice, Sociology of; Law and Society; Police Use of Force; Police: A Sociology of Knowledge Approach; Professions and Professionalization, History of; Terrorism.

Bibliography

Bayley, D.H., 1985. Patterns of Policing: A Comparative International Analysis. Rutgers University Press, New Brunswick, NJ.

Bayley, D.H., 2005. Changing the Guard: Developing Democratic Police Abroad. Oxford University Press, New York.

Bittner, E., 1990. Aspects of Police Work. Northeastern University Press, Boston, MA.

Black, D.J., 1980. The Manners and Customs of the Police. Academic Press, New York.

Casey, J., 2010. Policing the World: The Practice of International and Transnational Policing. Carolina Academic Press, Durham, NC.

Corsianos, M., 2012. The Complexities of Police Corruption. Rowman & Littlefield Publishers, Lanham, MD.

Deflem, M., 2002. Policing World Society: Historical Foundations of International Police Cooperation. Oxford University Press, Oxford, UK.

Deflem, M., 2010. The Policing of Terrorism: Organizational and Global Perspectives. Routledge, New York.

Deflem, M., McDonough, S., 2010. International law enforcement organizations. In: Kethineni, S. (Ed.), Comparative and International Policing, Justice, and Transnational Crime. Carolina Academic Press, Durham, NC, pp. 127–148.

Dempsey, J.S., Forst, L.S., 2014. An Introduction to Policing, seventh ed. Delmar, Clifton Park, NY.
Fogelson, R., 1977. Big-City Police. Harvard University Press, Cambridge, MA.
Greene, J.R., Mastrofski, S.D., 1991. Community Policing: Rhetoric or Reality. Praeger, New York.
Kelling, G.L., Coles, C.M., 1996. Fixing Broken Windows: Restoring Order and Reducing Crime in Our Communities. Martin Kessler Books, New York.
Manning, P.K., 1977. Police Work: The Social Organization of Policing. The MIT Press, Cambridge, MA.
Manning, P.K., 2008. The Technology of Policing: Crime Mapping, Information Technology, and the Rationality of Crime Control. New York University Press, New York.
Marx, G.T., 1988. Undercover: Police Surveillance in America. University of California Press, Berkeley, CA.
Mawby, R.I. (Ed.), 1999. Policing across the World: Issues for the Twenty-first Century. UCL Press, London.
O'Neill, M., Marks, M., Singh, A.M. (Eds.), 2007. Police Occupational Culture: New Debates and Directions. Elsevier, Amsterdam.
Palmiotto, M.J., Unnithan, N.P., 2011. Policing and Society: A Global Approach. Delmar Cengage Learning, Clifton Park, NY.
Ratcliffe, J.H., 2008. Intelligence-Led Policing. Willan Publishing, Cullompton.
Sadykiewicz, M., 2005. Police structure: centralized/decentralized. In: Haberfeld, M.R., Sullivan, L.E. (Ed. in Chief) (Eds.), Encyclopedia of Law Enforcement. Sage Publications, Inc., Thousand Oaks, CA, pp. 1250–1253.
Scott, J., 2009. Police–community relations. In: Miller, J. (Ed.), 21st Century Criminology: A Reference Handbook. SAGE Publications, Inc., Thousand Oaks, CA, pp. 823–833.
Sherman, L.W., 1978. Scandal and Reform: Controlling Police Corruption. University of California Press, Berkeley, CA.
Sklansky, D.A., 2008. Democracy and the Police. Stanford University Press, Stanford, CA.
Skolnick, J.H., 2001. Policing. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences. Elsevier, Oxford, UK.
Walker, S., 1977. A Critical History of Police Reform: The Emergence of Professionalism. Lexington Books, Lexington, MT.
Weber, M., 1922/1978. Economy and Society: An Outline of Interpretive Sociology. University of California Press, Berkeley, CA.

Relevant Websites

www.europol.europa.eu – Europol.
www.fbi.gov – Federal Bureau of Investigation (FBI).
www.interpol.int – Interpol.
content.met.police.uk – Metropolitan Police Service.
www.ice.gov – U.S. Immigration and Customs Enforcement (ICE).

Policing in Divided Societies

Aogán Mulcahy, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In divided societies, the police tend to be closely aligned with the dominant social group and their actions focus on maintaining state security rather than providing routine policing services. As a consequence, policing often becomes a key axis of conflict and efforts to reform the police are central to the process of conflict resolution. The Patten Report reforms, for example, were a crucial part of the peace process in Northern Ireland and have been widely viewed as a success. However in other divided societies where police reform may comprise part of a peacekeeping intervention or international aid program, the impact of 'security sector reform' measures tends to be less clear-cut, partly because of the scale of the challenges involved, and also because of the difference in the priorities of the 'donor' and the host societies.

Introduction

The term 'divided societies' is used to describe societies characterized by a major schism between different social groups, often arising from their respective and fundamentally antagonistic orientation toward the prevailing political arrangements (Guelke, 2012). Whether arising from ethnic, national, or religious differences, the divisions that mark such societies often take the form of dominant/subordinate relations, with a corresponding and often stark differential in access to resources and the established routes to political power. While Northern Ireland, South Africa, and Israel/Palestine are typically highlighted as iconic examples of 'divided societies', the applicability of this term extends to a range of other societies and contexts, whether in relation to the former Yugoslavia, 'failed states' such as Somalia, or other societies marked by civil war or violent and enduring upheaval.

At the outset it is important to distinguish between 'divided societies' and 'societies with social divisions'. The latter characterization reflects the majority of societies worldwide, where factors including class, race, and gender may each exert a major influence on the ability of individuals to participate fully and equally in society and on the overall quality of the life they enjoy. In various ways, they exert a strong influence on the relationship between different social groups and the police. Although these divisions are tangible and far-reaching, the difficulties surrounding policing are of a different order when a society is structured around the key axis of division.

In divided societies, the dynamics of conflict often revolve specifically around policing and related issues of security and justice for several reasons. First, as an agency of the state, the police (or the public police at least) are symbolically associated with the dominant political order, as well as being a vital expression of its security insofar as their activities are crucial in protecting the state against its enemies, whether within or without. Second, and related to the above, while the police may continue to provide a civil policing service (dealing with 'ordinary' crime), primacy is given to securing the state. Relations with subordinate groups, therefore, are shaped by the view that they are a threat to be managed rather than part of the citizenry to be protected. Third, the nature of such societies ensures that political opposition generates an aggressive policing response; and police policies and behavior in dealing with social unrest or protest may exacerbate the existing divisions, as policies of crowd control or even the sustained surveillance of a suspect community generate further hostility, adding fresh grievances to extant opposition. These factors ensure that "conflict over policing is endemic in these societies" (Weitzer, 1995: p. 5), both as an underlying dimension and a dynamic aspect of the contours of conflict.

The remainder of this discussion is structured in two sections. First, I consider the defining features of policing in divided societies, including the role of nonstate actors in providing security and policing in authoritarian regimes. Second, I highlight the role that police reform plays in processes of conflict resolution, and in distinguishing between internally driven reform processes (using events in Northern Ireland as a case study) and those emanating from external intervention; and exploring the role that dealing with the past plays in the dynamics of this process.

Features of Policing in Divided Societies

Various authors have argued that policing in divided societies is characterized by a set of features that distinguish it from policing elsewhere. Brewer (1991) and Weitzer (1995) have proposed different typologies and these influential models are outlined below (see also Enloe, 1980) (**Table 1**).

Some differences are evident between these two models. Brewer (1991: pp. 181–184), for instance, claims that policing in divided societies is characterized by "chronic and endemic manpower shortage" due to the far-reaching demands placed on the police; "the diffusion of policing functions throughout the dominant group" through the involvement of volunteer or auxiliary groups in the policing role; and "close operational links between the police and the military". Weitzer (1995: p. 299), however, notes that these latter characteristics are "neither common in nor unique to divided societies", but certainly they are evident in some cases. Overall though, Brewer and Weitzer agree on core dimensions: the alignment of the police with the dominant social group against subordinate groups, the deployment of police resources accordingly, and the protracted, negative consequences arising from this.

While divided societies experience inequality and conflict across a range of domains, the issue of policing is uniquely

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96051-3

Table 1 Features of policing in divided societies

Brewer (1991)	*Weitzer (1995)*
• Selective enforcement of the law in favor of the dominant group. • Discriminatory practices which limit the exercise of the rights of the minority. • Political partisanship in upholding and enforcing the distribution of political power by allowing unequal rights to political protest • A lack of autonomy from the political system. • An absence of effective mechanisms of police accountability. • Relatively unrestrained use of force. • A dual role which arises from responsibility for ordinary crime and internal security, although as threats to internal order have grown, security tends to dominate police activities, strategic planning, and leadership. • The polarization of attitudes toward the police and their conduct. • A social composition biased toward the dominant group because of an inability or unwillingness to recruit from among the minority communities. • Chronic and endemic manpower shortage. • The diffusion of policing functions throughout the dominant group, as volunteer groups and other compatible agencies are drawn into the policing role. • Close operational links between the police and the military.	• Systematic bias in law enforcement. • Politicized policing, including strong police identification with the regime and vigorous police action against the regime's political opponents. • Dominant-group monopoly of the top positions in the police force. • Dual responsibility of the police for internal security and ordinary law enforcement. • Legal or extralegal powers giving the police great latitude, including powers in relation to the use of force. • An absence of effective mechanisms of accountability with respect to police abuses of power. • Polarized communal relations with the police, with the dominant group as a champion of the police and the subordinate group largely estranged from the police.

controversial. Again, though it is important to recognize the complexity of the issues involved here, the public can be a highly sophisticated audience, distinguishing between the more value-neutral issues of service delivery and the wider issue of legitimacy. For example, members of the public may express higher levels of satisfaction with the routine overall performance of the police than they do in relation to the police's involvement in public order or state security issues. Satisfaction with routine policing does not translate into acceptance of the police or the legitimacy of their role. Concerns over police legitimacy may be muted or latent in the absence of widespread disorder or an immediate political crisis. But if the political environment becomes more fractious and the issue of policing becomes the focus of attention – for example, through police responses to public protest – then the legitimacy deficit becomes all too apparent and the crisis of the state becomes a crisis of policing. Thus, reflecting on policing and the civil rights movement in the USA, Wright noted the spiraling effects of widespread concerns over policing, which the establishment largely ignored until this could no longer serve as a plausible response:

> once a group experiences exclusion and has no effective way of communicating what is happening to those who are not affected, the police become a lightning conductor. The worst things done by police, which are neither punished nor for which any kind of restitution is made, come to be seen as representative of a society which tolerates or encourages them. The best elements of the police are compromised by their complicity with the worst. And the most provocative amongst the people they are policing are seen to be acting under provocation themselves – a provocation which is conceived as an intent. If all other avenues of trans-ethnic communication and cooperation are absent, the substance of interethnic relations tends to revolve around the antagonistic axis of law and order.
>
> (1988: p. 180–181)

Similarly, in the context of Northern Ireland, Whyte (1990: p. 88) noted that Catholics and Protestants disagreed more over the "unhealed sore" of policing and security policy than they did on constitutional issues. The biggest disparity he found between Catholic and Protestant views in any public opinion survey, was not on political preferences but rather on the use of plastic bullets, a particularly controversial aspect of security policy, following the death of several members of the public after being struck by plastic bullets.

In highlighting these issues, it is important to bear a number of qualifications in mind. First, even in the most difficult environment, some level of 'normal' or 'routine' policing usually persists. In fact, 'routine policing' might even typify the nature of policing in communities which are at some remove from frontline conflict zones (Brewer and Magee, 1991). Second, there is a danger that the focus on intergroup relations overlooks internal divisions within either the dominant or the subordinate group. For example, while the dominant social group is likely to be largely supportive of the police, this is not always the case. Some communities within the dominant group may harbor serious concerns about police effectiveness and efficiency. For example, Hasisi and Weitzer (2007) demonstrate that whereas the Israeli-Palestinian conflict shaped public attitudes toward the police in significant ways, with large differences between the views of Jews and Arabs, substantial differences in opinion were also evident within the minority Arab community. In other contexts, pro-state groups may harbor deep concerns over police misconduct but may feel inhibited from expressing them, lest this be used to criticize and possibly undermine 'their' police (Weitzer, 1995).

Policing and Authoritarian Regimes

While the term 'divided societies' may imply a level of overt conflict, in reality many divided societies may be better

described by stability, albeit a stability arising from explicit domination and the threat of violence. In totalitarian or authoritarian states, for instance, the incumbent regime typically relied on an extensive state security apparatus to maintain control (supported by ideology and patronage, among other things). In the former Soviet Union, policing had an explicit political dimension, and police activities were directed against those who were perceived to pose a threat to the state (Shelley, 1996). Even the policing of low-level crimes of 'social disorder' was driven by an explicit concern that this posed a threat to the well-being of the wider society and to Soviet ideals (Shearer, 2009). The scale of security policing under these authoritarian regimes was enormous. Not only did it involve a large body of personnel, it also extended to a vast project of surveillance. The files of the STASI, East Germany's state security police, stretched for well over 100 km of shelf-space, and 174 000 people were identified as informants (2.5% of the population aged 18–60 years). Estimates suggest that the true number of informants could be significantly higher (perhaps by a factor of three of more), and that does not include those who were only occasional informants (Funder, 2004; Koehler, 2000. See also Browder's, 2004 discussion of policing in Nazi Germany).

The transition from such authoritarian regimes to a more democratic form of government inevitably brought issues of police reform to the fore. In addition to concerns over security policing, civil policing was also highly problematic and was characterized by a combination of routine misconduct, lack of accountability, bribery and corruption, and mistrust. In different former soviet societies, the legacy of this situation persists. Efforts to reform the police have proven limited in impact and levels of trust in the police remain relatively low, which itself may mirror low levels of trust in state institutions generally in post-Communist societies (Beck and Chistyakova, 2002; Beck and Robertson, 2009; Light, 2014; Semukhinaa and Reynolds, 2011; Shelley, 1999).

'Alternative Justice' in Divided Societies

The dynamics of policing and security in divided societies leave a considerable lacuna in terms of security provision. The focus on state security absorbs considerable policing resources; and hostility between the police and marginal groups leads to difficulties in conducting routine policing, including a paucity of information provided to the police (on whom they are dependent to investigate crimes successfully) as well as the physical dangers posed to the police. In those circumstances, public concerns over 'ordinary crime' can be considerable; and even though the demand for policing may be acute, it is not delivered at an appropriate level or to an adequate standard. The situation is further complicated by the use of criminals as informants by the police or security forces, as well as by concerns that communities opposed to the state are in effect punished for their stance by active neglect on the part of the police. Thus, providing a policing solution of sorts can be an important facet of the activities of nonstate actors.

The manner and extent to which this 'policing vacuum' is addressed varies from society to society, and while it may be relatively haphazard in some societies, in others it may evolve into a (relatively) structured system of 'alternative justice' (see, generally, Abel, 1982). In Northern Ireland, for instance, Republican and Loyalist groups each provided a system of 'alternative justice' during the conflict and beyond, incorporating a loose tariff of punishments – varying in prevalence over time and according to the wider political context – including punishment shootings (often 'kneecapping' of the victim), punishment beatings, banishment, shaming, and so on. The dynamics of this alternative justice system are highly complex. While it accrued a degree of political capital for those involved in administering this system, it also diverted resources away from what many paramilitaries viewed as their *raison d'etre* – violent opposition to or support for the prevailing state arrangements – and led to outraged criticism of the groups involved for the sheer brutality inherent in such punishments. The systems operated by Republican and Loyalist groups tended to reflect their different organizational structures as well as their respective relationships to the state. Republican groups tended to be more hierarchical in command structure. Loyalist groups, by contrast, tended to be less centralized and more loosely organized in structure, allowing for greater variation in the use of paramilitary punishments from group to group, while also allowing greater scope for these punishments to be used to settle personal scores rather than address crime specifically. While one of the untold stories of the Northern Ireland conflict lay in the difficulties and complexities of relations between Loyalist communities and the police, these difficulties were not of the same order as those found in Republican areas. Quite simply, there was not a sufficient gap in the policing market for Loyalist groups to operate a system of alternative justice that could generate the same levels of legitimacy that the Republican groups obtained.

In other settings, the inadequacy of the police response may lead various groups to resort to private security services to provide the level of policing/security they want, whether to protect their own personal security or specific commercial/corporate interests. In divided societies, however, the resort to private security is significant given the often highly militarized level of policing that such private contractors provide, and the lack of oversight characterizing their deployment (see, for example, Newburn, 2007; O'Reilly, 2010).

Political Transition and Police Reform

While policing is a key dynamic of conflict in divided societies, it is important to recognize that societies are not condemned to eternal conflict. Scope may exist to resolve the underlying issues and provide appropriate remedies through the development of political arrangements that can command widespread political support. Indeed, policing and 'security sector reform' generally loom large in any peace process and form a key dimension of negotiations and any ensuing political agreement. Police reform initiatives tend to occur through internal processes of political accommodation or through the intervention of outside parties and the deployment of external agencies.

Police Reform and Internal Agreement

One of the most prominent examples internationally of police reform is the recent case of Northern Ireland. The 1998 Belfast

Agreement outlined the basis for a political agreement to resolve the Northern Ireland conflict. Under this agreement the Independent Commission on Policing (ICP) – chaired by Chris Patten – was established, and its 1999 report laid the basis for a major transformation of the policing landscape in Northern Ireland. The ICP's recommendations were not oriented to the specific political circumstances prevailing in Northern Ireland at any one particular time, but instead focused on addressing "the challenge confronting policing in any modern society" (p. 6). Their recommendations were designed to be consistent with five key principles: "Does this proposal promote effective and efficient policing? Will it deliver fair and impartial policing, free from partisan control? Does it provide for accountability, both to the law and to the community? Will it make the police more representative of the society they serve? Does it protect and vindicate the human rights and human dignity of all?' (p. 5).

Overall, the report's most prominent recommendations coalesced around a number of key themes (see Table 2 below): to enhance accountability, oversight, and governance; to work in partnership with the community and to devolve decision-making to the lowest level possible; to embed human rights within the structures of policing; to make the police more representative of the population of Northern Ireland; and to make its culture and ethos more inclusive.

The ICP's most significant recommendations addressed issues of governance, accountability, and oversight through the establishment of new structures which would embed accountability in the institutional landscape of policing in Northern Ireland. In terms of governance, the ICP proposed that a 'Policing Board' replace the existing and largely ineffectual Police Authority, with the function of holding the Chief Constable publicly to account, and agreeing and monitoring the implementation of a policing plan; the Board would also be given powers to initiate inquiries and to require the Chief Constable to report on any issue.

At local authority level, a system of 'District Policing Partnership Boards (DPPBs)' was to be established, comprised of elected and independent members, and 'advisory, explanatory and consultative' in function (p. 35). Importantly, the ICP called for these to be given tax-raising powers of up to 3% (similar to the powers already available to District Councils for economic development) to purchase additional policing services if required, whether "from the police or other statutory agencies, or from the private sector" (p. 35).

The ICP endorsed an earlier recommendation that an independent police complaints system be established in the form of a 'Policing Ombudsman'. This would also have the power to investigate policies or practices where there was a public interest in doing so. To ensure that the reform process itself was fully implemented, the ICP called for the appointment of an 'Oversight Commissioner' with the explicit function of monitoring and reporting on the progress made in implementing the ICP recommendations. Underpinning this emphasis on oversight and accountability was an explicit commitment to embedding *human rights* within all aspects of policing, including training and appraisal. The report recommended a new oath and code of ethics and the appointment of a human rights lawyer to advise the police.

The ICP noted the complex nature of the many problems the police faced, but in seeking to address these effectively it called for a new model of 'policing with the community', a fundamental partnership between the police and the public. As it noted:

> it is not so much that the police need support and consent, but rather that policing is a matter for the whole community, not something that the community leaves to the police to do. Policing should be a collective partnership: a partnership for community safety.
> (p. 8)

It stated that 'policing with the community should be the core function of the police service and the core function of every police station' (p. 43), and that police decision-making should be devolved to the lowest level possible.

The ICP noted argued that through sustained efforts the police should be made more representative of the wider population, particularly Catholics who comprised 8% of the force in 1999 (at a time when Catholics nearly comprised half of the overall population). It proposed a new recruitment strategy to increase Catholic representation within the PSNI to approximately 30% (which the ICP appeared to view as a 'critical mass' figure) within a 10 year period. This entailed a system of '50:50 recruitment' whereby applicants who met the entry requirements would be placed into two separate pools of Catholics and Non-Catholics. Applicants would be ranked within each pool and an equal number of recruits would then be selected from each. Simultaneously, the report called for a significant

Table 2 Patten report – selected recommendations

Accountability, governance, and oversight	Policing Board to replace the existing Police Authority.
	District Policing Partnership Boards to be established at local authority level
	Police Ombudsman to be established as an independent agency to investigate police complaints
	Oversight Commissioner to monitor implementation of ICP recommendations
Human rights	New Oath
	Human Rights to be embedded within all training
Policing with the community	Policing with the community to be established as core function of the police
	Neighborhood policing teams to be established
	Decision-making to be devolved as much as possible
Size and composition	Reduction from approximately 13 000 full-time officers to 7500
	Introduction of 50:50 (Catholic:Non-Catholic) recruitment scheme
Culture and ethos	Royal Ulster Constabulary to be renamed
	Symbols to be inclusive

reduction in the overall force size, from 13 000 officers to 7500, which would be achieved specifically through a generous severance scheme. The report also recommended that the force be renamed – in 2001 it became the Police Service of Northern Ireland (PSNI) – and that 'force symbolism' should be neutral in terms of political affiliation.

The Chairperson of the Northern Ireland Policing Board lauded the reform program as a "blueprint for democratic policing anywhere in the world" and described the PSNI as "the most scrutinized and accountable police service probably anywhere in the world today" (see Bayley, 2008; Ellison, 2007; Mulcahy, 2013), while Chris Patten (2010: p. 25) described the PSNI as "the best working example of how to police a divided community". Such acclaim runs the risk of overlooking, among other things, the political antagonism which characterized the implementation of the Patten reforms, as well the ongoing difficulties in translating the report's focus on policing with the community into practice (see generally Ellison and O'Rawe, 2010; Mulcahy, 2013). Nevertheless, the police reform process in Northern Ireland quickly assumed the mantle of a success story, particularly for its establishment of a robust model of police governance and oversight.

If Northern Ireland, albeit with qualifications, stands as a success story, the development of policing in post-conflict settings is by no means a smooth process. In the context of South Africa, for instance, Brogden and Shearing (1993) and Marks (2005) highlighted the considerable difficulties involved in moving from a heavily militarized mode of policing to a civic oriented one. Conducted under a significant amount of international scrutiny, the reform program involved major institutional changes as well as extensive training and a turnover in personnel. Nevertheless, Ivkovic and Sauerman (2013) found that even two decades after the formal ending of the apartheid regime, a notable minority of officers did not recognize even the most severe forms of police corruption as violations of official rules. Baker (2002) links the state's failure to provide adequate public policing to the continuation and proliferation of nonstate forms of policing that sought to provide the security that the state did not, particularly as levels of violent crime remained extremely high and amidst the dawning realization that the end of apartheid did not guarantee any significant improvement in the living standards of the population. Private security in this context, however: "permits the continuation of authoritarian values and practices; divides communities on the basis of their ability to secure alternative policing for a failing state provision; and nurtures the view that the rule of law is more of an obstacle to maintaining social order than an effective guarantee of it" (Baker, 2002: p. 51). Such claims might not hold in every circumstance, but they nevertheless highlight the dangers of transitional processes and the centrality of security provision therein.

Security Sector Reform and External Intervention

In many conflict zones or transitional societies the issue of police reform is part of a wider process that often reflects a variety of international interests and interventions, due to the sheer scale of conflict and division that exists and the corresponding scale of the challenges that this poses. In 'developing democracies', for instance, Hinton and Newburn (2009: p. 6) argue that the public sphere tends to be characterized by: "weak democratic institutions"; "corruption and weak rule of law"; "significant levels of poverty and inequality"; "high crime and political instability"; and "poorly institutionalized channels of police accountability and responsiveness". In such circumstances, where routine law and order may have broken down, the challenges for policing are considerable. As Grabosky notes: "In many weak and failing states, and in other post-conflict situations, there may be no functioning police agency. Alternatively, the police in the host country may be corrupt, incompetent, abusive, or some combination of the three. Thus there is often a need for external assistance in building or in rebuilding an indigenous police service" (2009: p. 101). In violent contexts, police from a variety of countries may form part of international peacekeeping missions and police reform typically forms part of a wider program of 'security sector reform' (Greener, 2011).

The growth of interest in 'security sector reform' and the development of democratic policing worldwide arises from a number of factors. First, the success of development and aid programs, and even of political agreements, has increasingly come to be seen as dependent on ensuring basic levels of security (Bayley, 2005). Second, it reflects a heightened concern over insurgency and radicalism in the aftermath of the attacks of 11 September 2001 (9/11) on the United States. Third, the sphere of policy development is increasingly international in scope and characterized by greater attentiveness to 'policy transfer' and 'lesson drawing' between jurisdictions, which itself is underpinned by the rapid growth of the private security sector and the development of an international policy elite offering 'solutions' to police/security problems (O'Reilly, 2010).

International interventions have taken place in diverse contexts, and prominent examples include initiatives by the European Union and other bodies in the former Yugoslavia (Aitchison and Blaustein, 2013; Celador, 2008), as well as Australian-backed interventions in a variety of Pacific conflicts (Dinnen and Allen, 2013; Goldsmith and Dinnen, 2007). The provision of appropriate assistance, however, is complicated by various factors. First, in societies often characterized as 'weak or failing states', 'donor' societies may contribute to humanitarian and reconstruction programs partly to prevent the development of humanitarian disasters. However, this typically coincides with issues of donor self-interest, such as preventing the development of a political vacuum in which radical groups could flourish or the destabilizing effect that a conflict could have across an entire region e.g., through refugee flows as the spillover of violence (see, for example, Newman, 2010).

Second, the impact of external interventions is uncertain and contingent. This reflects "inflated expectations for police reform [which] are frequently trumpeted by domestic politicians and international policy elites, regardless of how staggering the rates of crime, how politicized and heavy-handed the police force, and how corrupt and lawless the government that is meant to enact and monitor the proposed reforms" (Hinton and Newburn, 2009: p. 1–2). It also reflects the general absence of long-term evaluations of such processes, and in effect, a default understanding of success in terms of the establishment of new institutions or training programs, rather than the impact they have on the wider social landscape of security and justice (Ellison and Pino, 2012).

Third, the eagerness with which western powers lecture transitional societies about 'international best practice' policing arrangements does not seem to be accompanied by a careful analysis of the specific needs of each society but rather by an assumption that Western models of police reform – relying heavily on vague models of 'community policing' – are the obvious 'solution' (Ellison and Pino, 2012). As Hills (2009) pointedly noted, external agencies tend to prioritize 'security', whereas local residents prioritize 'order' at an everyday level. Indeed, the involvement of international peacekeepers and initiatives to enhance the professionalism of the indigenous police can be counterproductive insofar as it can further undermine the reputation of the local police and lead to greater levels of dependency on international peacekeepers and other external parties (Dinnen and Allen, 2013).

Dealing with the Past

The legacy of the past is a formidable challenge for any society emerging from a period of conflict. Not only is the sheer scale of the conflict's human cost a major hurdle in its own right, but this is further complicated the covert nature of the protagonists' actions, the use of 'proxy' actors to pursue goals (Jamieson and McEvoy, 2005), and widespread strategies of 'othering' and 'denial' (Cohen, 2001). Yet the demand for justice and disclosure ensures that strategies to deal with the past often form a key dimension of the processes of conflict resolution. The South African Truth and Reconciliation Commission is probably the most renowned example of this, and while it was characterized as a means of laying bare the past and thereby providing a basis for political reconciliation, in its detail it often comprised a harrowing catalog of policing practices and the workings of the state security apparatus. The TRC was based on the simple premise that individuals providing full disclosure of their actions would be granted immunity from prosecution. This immunity was, in effect, presented as the price of political progress (see generally, Hayner, 2010).

In Northern Ireland, by contrast, the legacy of the past presided in its absence. Although this issue dominated popular debate and shaped in far-reaching ways the relations between the various parties that comprised the power-sharing executive, the British Government remained largely hostile to any formal process of truth recovery. While allegations surrounding different policing controversies formed one aspect of this debate (Lawther, 2010), one way in which some of the wider issues surrounding the conflict have been addressed is through the PSNI's establishment of a Historical Enquiries Team to investigate the approximately 1800 unsolved killings from the conflict, creating specific units to review police investigations and any available evidence. While there was a possibility of individuals being charged as a result of this review, it most cases it simply provided more information which could be relayed to the victims' families and thereby provide some form of closure. Yet as Lundy (2011: p. 102) notes, this format is more appropriate to truth recovery in respect of individual cases, than to investigating the 'macro-truths' relating to wider practices and policies of the various protagonists around which so much of the debate on the past has focused. The focus on policing the past has, however, proven costly, both in terms of absorbing scarce financial resources as well as in diverting political focus away from the need to agree on shared priorities for the future (Oversight Commissioner, 2007). These tensions highlight the manner in which current institutions can become enmeshed in the controversies surrounding legacy issues and this may remain the case until decisive steps are taken to embark on a society-wide process of truth recovery that provides an adequate means of resolving them.

Conclusion

In divided societies, policing is a uniquely controversial issue that often mediates the relationship between entire communities and the state. Typically, the police are closely aligned with dominant social groups and police resources are deployed accordingly against subordinate groups. The model of policing this generates is one characterized by close links between the police and the political sphere, policing strategies that are heavily reliant on controversial public order tactics and lacking in oversight, and deeply polarized experiences of and attitudes toward the police.

These difficulties have generated considerable interest in how policing in divided societies may be reformed. While recent developments in Northern Ireland – and particularly the template provided by the 1999 Patten Report – are often cited as an example of successful reform, especially in relation to oversight and accountability, in many other instances of police reform the outcomes are much more qualified. The impact of external interventions and the linkage of police reform to development aid programs is unclear and it has been heavily criticized for failing to pay due regard to the specific needs of individual societies. Even in wider processes of political transition, the significance of policing is evident through the shadow it casts on processes of 'truth recovery' and other measures to deal with the past.

See also: Ethnic Conflicts; Peace Processes; Political Repression.

Bibliography

Abel, Richard, 1982. The Politics of Informal Justice, 2 vols. Academic Press, New York.

Aitchison, Andy, Blaustein, Jarrett, 2013. Policing for democracy or democratically responsive policing? Examining the limits of externally driven police reform. European Journal of Criminology 10 (4), 496–511.

Baker, Bruce, 2002. Living with non-state policing in South Africa: the issues and dilemmas. The Journal of Modern African Studies 40 (1), 29–53.

Bayley, David, 2005. Changing the Guard: Developing Democratic Policing Abroad. Oxford University Press, New York.

Bayley, David, 2008. Post-conflict police reform: is northern Ireland a model? Policing 2 (2), 233–240.

Beck, Adrian, Chistyakova, Yulia, 2002. Crime and policing in post-Soviet societies: bridging the police/public divide. Policing and Society 12 (2), 123–137.

Beck, Adrian, Robertson, Annette, 2009. The challenges to developing democratic policing in post-Soviet societies: the Russian experience. Police Practice and Research 10 (4), 285–293.

Brewer, John, 1991. Policing in divided societies: theorising a type of policing. Policing and Society 1 (3), 179–191.

Brewer, John, Magee, Kathleen, 1991. Inside the RUC: Routine Policing in a Divided Society. Clarendon, Oxford.

Brogden, Mike, Shearing, Clifford, 1993. Policing for a New South Africa. Routledge, London.
Browder, George, 2004. Foundations of the Nazi Police State: The Formation of Sipo and SD. University Press of Kentucky, Lexington, KY.
Celador, G.C., 2008. Becoming 'European' through police reform: a successful strategy in Bosnia and Herzegovina? Crime, Law and Social Change 51 (2), 231–242.
Cohen, Stanley, 2001. States of Denial: Knowing about Atrocity and Suffering. Polity, Cambridge.
Dinnen, Sinclair, Allen, Matthew, 2013. Paradoxes of postcolonial police-building: Solomon Islands. Policing and Society 23 (2), 222–242.
Ellison, Graham, 2007. A blueprint for democratic policing anywhere in the world? Police reform, political transition, and conflict resolution in northern Ireland. Police Quarterly 10 (3), 243–269.
Ellison, Graham, O'Rawe, Mary, 2010. Security governance in transition: the compartmentalizing, crowding out and corralling of policing and security in Northern Ireland. Theoretical Criminology 14 (1), 31–57.
Ellison, Graham, Pino, Nathan, 2012. Globalization, Police Reform and Development: Doing it the Western Way? Palgrave Macmillan, Basingstoke, UK.
Enloe, Cynthia, 1980. Ethnic Soldiers: State Security in Divided Societies. Penguin, London.
Funder, Anna, 2004. Stasiland: Stories from behind the Berlin Wall. Granta, London.
Goldsmith, Andrew, Dinnen, Sinclair, 2007. Transnational police building: critical lessons from Timor-Leste and Solomon Islands. Third World Quarterly 28 (6), 1091–1109.
Grabosky, Peter, 2009. Police as international peacekeepers. Policing and Society 19 (2), 101–105.
Greener, Bethan, 2011. The rise of policing in peace operations. International Peacekeeping 18 (2), 183–195.
Guelke, Adrian, 2012. Politics in Deeply Divided Societies. Polity, Cambridge.
Harris, Vandra, Goldsmith, Andrew, 2012. Police in the development space: Australia's international police capacity builders. Third World Quarterly 33 (6), 1019–1036.
Hasisi, Badi, Weitzer, Ronald, 2007. Police relations with arabs and jews in Israel. British Journal of Criminology 47 (5), 728–745.
Hayner, Priscilla, 2010. Unspeakable Truths: Transitional Justice and the Challenge of Truth Commissions, second ed. Routledge, Abingdon.
Hills, Alice, 2009. Policing Post-Conflict Cities. Zed Books, London.
Hinton, Mercedes, Newburn, Tim (Eds.), 2009. Policing Developing Democracies. Routledge, Abingdon.
Ivkovic, Sanja, Sauerman, Adri, 2013. Threading the thin blue line: transition towards democratic policing and the integrity of the south african police service. Policing and Society. http://dx.doi.org/10.1080/10439463.2013.788651 (accessed 07.05.13.).
Jamieson, Ruth, McEvoy, Kieran, 2005. State crime by proxy and juridical othering. British Journal of Criminology 45 (4), 504–527.
Koehler, John, 2000. Stasi: The Untold Story of the East German Secret Police. Westview, Boulder, CO.
Lawther, Chris, 2010. 'Securing' the past: policing and the contest over truth in northern Ireland. British Journal of Criminology 50 (3), 455–473.
Light, Matthew, 2014. Police reforms in the Republic of Georgia: the convergence of domestic and foreign policy in an anti-corruption drive. Policing and Society 24 (3), 318–345.
Lundy, Patricia, 2011. Paradoxes and challenges of transitional justice at the 'local' level: historical enquiries in northern Ireland. Contemporary Social Science 6 (1), 89–106.
Marks, Monique, 2005. Transforming Robocops: Changing Police in South Africa. University Of KwaZulu-Natal Press, Pietermaritzburg.
Mulcahy, Aogán, 2006. Policing Northern Ireland: Conflict, Legitimacy and Reform. Willan/Routledge, Devon/London.
Mulcahy, Aogán, 2013. Great expectations and complex realities: the impact and implications of the police reform process in northern Ireland. In: Brown, Jennifer (Ed.), The Future of Policing. Routledge, Abingdon, UK, pp. 476–492.
Newburn, Tim, 2007. Governing security: the rise of the privatized military. In: Downes, David, et al. (Eds.), Crime, Social Control and Human Rights. Willan Publishing, Cullompton, Devon, UK, pp. 195–210.
Newman, Edward, 2010. A human security peace-building agenda. Third World Quarterly 32 (10), 1737–1756.
O'Reilly, Conor, 2010. The transnational security consultancy industry: a case of state-corporate symbiosis. Theoretical Criminology 14 (2), 183–210.
Oversight Commissioner, 2007. Overseeing the Proposed Revisions to the Policing Services of Northern Ireland – Report 19. Office of the Oversight Commissioner, Belfast.
Patten, Chris, 2010. Personal reflections on chairing the commission. In: Doyle, David (Ed.), Policing the Narrow Ground. Royal Irish Academy Press, Dublin, pp. 13–26.
Semukhinaa, Olga, Reynolds, Michael, K., 2011. Russian citizens' perceptions of corruption and trust of the police. Policing and Society 24 (2), 158–188.
Shearer, David, 2009. Policing Stalin's Socialism: Repression and Social Order in the Soviet Union 1924-1953. Yale University Press, New Haven, CT.
Shelley, Louise, 1996. Policing Soviet Society: The Evolution of State Control. Routledge, Abingdon.
Shelley, Louise, 1999. Post-Socialist policing: limitations on institutional change. In: Mawby, R.I. (Ed.), Policing across the World: Issues for the Twenty-first Century. UCL Press, London, pp. 75–87.
Weitzer, Ronald, 1995. Policing under Fire: Ethnic Conflict and Police-community Relations in Northern Ireland. State University of New York Press, Albany, New York.
Wright, Frank, 1988. Northern Ireland: A Comparative Analysis. Gill and Macmillan, Dublin.
Whyte, John, 1990. Interpreting northern Ireland. Clarendon, Oxford.

Policy Analysis

Aidan R Vining, CNABS, Simon Fraser University, Vancouver, Canada
David L Weimer, University of Wisconsin-Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public policy analysis seeks to promote social values by informing participants in the policy-making process. It typically involves framing problems as inefficiencies resulting from market and government failures or undesirable distributions of opportunities or outcomes. In its most general mode, multi-goal analysis, analysts identify and justify relevant substantive and instrumental goals, specify policy alternatives, and systematically compare alternatives in terms of goals to support a recommendation. Depending on the relevant goals and the capacity for analysts to quantify and monetize progress toward goals, policy analysis may be performed in the mode of cost–benefit or cost–effectiveness analysis.

Introduction: What Is Policy Analysis?

Government officials, whether pharaohs, kings, presidents, prime ministers, legislators, or bureaucrats, face decisions that create public policies affecting the security of their states as well as the economic, social, and political lives of those who reside in them. Many of these officials, especially in routine circumstances, act as their own policy analysts, identifying alternative courses of action and predicting the consequences of these alternatives. However, even in ancient times, they sought advice from analysts. Joseph in Genesis, probably the first noted policy analyst in the Western tradition, defined the problem suggested by Pharaoh's dreams as the risk of famine following years of plenty. He recommended the policy of stockpiling grain, and accurately predicted that this policy would benefit both the Pharaoh and his people. History records individuals in India and China playing similar roles in ancient times (Goldhamer, 1978; Radin, 2013, p. 13). Today, public policy analysts, whether identified by formal title or their actual practice, can be found in almost all government agencies, think tanks, interest groups, and consultancies.

Public policy analysts are those who do *public* policy analysis. Weimer and Vining (2011, p. 24) define public policy analysis as "client oriented advice relevant to public decisions and informed by social values." In this definition, the focus on public decisions, that is, those made through the exercise of the formal authority of the state, distinguishes public policy analysis from the analyses conducted to support private sector activity, such as marketing decisions by firms. The requirement that the analysis be informed by social values distinguishes those analysts who seek to contribute to better public policy from society's perspective from those who aim to advance a particular private interest (even though these advocates typically couch their analysis as being in the public interest). So, for example, a private sector analyst who considered only maximization of profit or a political analyst who considered only their client's electoral advantage would not be conducting public policy analysis under this definition. Finally, this definition emphasizes the requirement that there be a client. This emphasis distinguishes public policy analysis as a *professional* activity – public officials who do their own policy analysis can be thought of as having themselves as a client. This aspect of the definition may be the most controversial component because it clearly distinguishes public policy analysis from the broader field of policy-relevant research.

Many academics, especially those in the social sciences, do policy-relevant research. Yet they rarely have a specific client and their work typically seeks to add to general knowledge rather than inform a specific public policy decision. Absent a specific client, we label these activities public policy research rather than public policy analysis. As many readers of this article will be familiar, either as scholars or students, with published policy research, it is useful to consider a few of the ways in which policy research tends to differ from policy analysis (Weimer, 2009). Perhaps most importantly, policy research tends to emphasize particular aspects of policy problems and specific policy impacts rather than taking a comprehensive approach, largely because the audience for the work is mostly other researchers who value innovative techniques and models, rather than policy makers who seek advice about courses of policy action. Consequently, policy research tends to employ the methods of a single discipline (most frequently economics, political science, or sociology) rather than draw on multiple disciplines as is often needed to make a holistic assessment of impacts. It also reflects a timeframe dictated by the demands of the methodology employed rather than the demands of a specific decision. Nonetheless, although policy research and policy analysis differ in important ways, policy research is often an extremely valuable resource for policy analysis. Indeed, the weight of policy research related to a policy-relevant issue almost always plays an important evidentiary role in policy analysis.

The first row of **Table 1** summarizes the contrast between public policy analysis and public policy research. The second row shows a similar distinction between stakeholder analysis, which is conducted from the perspective of a specific client, and policy process research, which speaks primarily to scholars. As political feasibility and successful implementation are almost always relevant goals in public policy analysis, good public policy analysis usually goes beyond the narrow definition based on substantive goals (northwest cell) to include process analysis as well (southwest cell).

Having placed policy analysis in a broader policy context, in the rest of this article we seek to sketch the various modes of

Table 1 Policy analysis in perspective

		Client versus societal focus	
		Client focus	*Academic/societal focus*
Substantive versus process focus	**Substantive policy focus**	**Policy analysis (narrowly defined)** Problem-solving focus Economics pre-eminent Comprehensive: problem analysis (market & government failure, synthesis, solution analysis (alternatives, goals, assessment)) Goals clear, or at least emergently clear; efficiency (cost–benefit) or efficiency, equity, government revenue-expenditure (multi-goal)	**Policy research/policy sciences** Social science research on policy problems Policy problem discovery/exploration Solution (policy) discovery/exploration Broad range of social sciences, but economics, political science dominate Partial or fragmentary (in terms of policy problem) Goals contestable
	Policy process focus	**Political/organizational analysis (or stakeholder analysis)** Political, organizational, and inter-organizational analysis (including networks) Relevant for both adoption and implementation Strategic client focus Often informal and unwritten Primarily descriptive and predictive, rather than normative: goal is adoption and implementation	**Policy process research** All social science research, but dominated by political science Distributional and redistributional focus Theory somewhat contestable: interest group theory, advocacy coalition, path dependency, etc., but converging on contingent and comprehensive theory

Source: Adapted from Vining, Aidan R., Weimer, David L. Policy Analysis, Foundations of Public Administration Series (2010). http://www.aspanet.org/scriptcontent/index_par_foundationsseries.cfm.

public policy analysis, which we henceforth refer to as just policy analysis. However, as the appropriateness of the various modes depends on the nature of the goals relevant to policy choice, we first consider framing public policy problems in terms of the inefficiency resulting from market and government failures or as undesirable distributions of outcomes or opportunities.

Framing Public Policy Problems

Analysts often begin their work by being confronted with what clients or they themselves view as undesirable social, economic, or political conditions. The first task of an analyst is to understand these conditions: Has the condition gotten better, worse, or remained the same over time? Is the condition better, worse, or the same in this jurisdiction than in another comparable one? Why do or should we care about the condition? The second task of the analyst involves determining if the condition is amenable to public policy – if one can imagine changes in policy that would make the conditions less undesirable, then it becomes a public policy problem (Dery, 1984).

It is often helpful to frame problems as the inefficient use of resources because of market and government failures (Weimer and Vining, 2011). Inefficiency can arise in two ways. First, technical inefficiency occurs when some output could be produced at a lower opportunity cost. That is, some alternative mix of resources that could produce the same level of output has a lower value in its next best alternative use so that less would have to be given up to obtain the output. Second, allocative inefficiency occurs when some alternative set of outputs would be valued more by society. That is, there is an alternative allocation for which the difference between what people would be willing to pay in aggregate for it and its opportunity cost is greater than for the current allocation. So, for example, a municipal refuse system would be technically inefficient if it were possible to provide the same levels of refuse collection at a lower cost, such as by using trucks with robotic arms that reduced injuries to workers. It would be allocatively inefficient if residents valued more frequent pickup even taking account of the higher taxes they would pay to fund the additional service.

In an ideal competitive economy, individuals pursuing their own interests would efficiently use resources to achieve both technical and allocative efficiency. In a real economy, however, some circumstances, called market failures, prevent private activity from resulting in an efficient allocation of resources. The commonly recognized market failures include externalities (marginal private cost differs from marginal social cost so that individuals produce or consume too much or too little), natural monopoly (average cost declines over the relevant rage of demand so that only one firm can profitably survive in the market), public goods (nonrivalry of consumption, non-excludability of use, or both result in too little voluntary supply or too much consumption), and information asymmetries (one party to a transaction has information that, if revealed, would lead the other party to make a different decision so that either too little or too much is consumed). Additionally, various other limitations to market economies, such as imperfect insurance and capital markets, or incoherent individual choice, such as that resulting from imperfect cognition or addiction, can also result in private actions failing to achieve efficient allocations of resources. Correcting market failures to increase efficiency is one of the rationales for public intervention in markets.

Just as private actions can fail to produce an efficient allocation of resources, collective action can fail to correct market failures appropriately or actually create inefficiencies (Glazer and Rothenberg, 2001; Weimer and Vining, 2011; Wolf, 1979). Most fundamentally, no social choice rule can guarantee translation of individual preferences into a fair and coherent social

ordering (Arrow, 1951). Representative governments often give greater voice to concentrated and organized interests, sometimes creating inefficient policies that favor these interests. Problems of bureaucratic supply make it difficult for governments to use resources efficiently in the things that it directly produces. Decentralization of authority, often desirable to account for differences in geographic preferences, can nonetheless create incentives for subnational governments to adopt inefficient policies in their pursuit of fiscal advantage.

If efficiency were the only goal of public policy, then whether or not a policy problem involved a market or government failure would guide the appropriate policy response (Weimer and Vining, 2006, p. 28–31). If there is neither market nor government failure, then the market is working and should be left alone. If there is market failure and no government failure, then policy is appropriately corrective and should be left alone. However, if there is no market failure, but government failure has manifested, then a return to the market, such as through deregulation, is warranted. If a combination of market and government failure occurs, then current policy is inappropriate and should be redesigned.

Although efficiency is usually one appropriate goal of public policy, it is rarely the only one. Most often, some form of equity is relevant, whether creating greater opportunity for those with limited resources to participate in markets and civic life, providing floors on consumption to preserve human dignity, or creating a more equal distribution of consumption. Distributional goals seek greater equity in some relevant dimension.

More generally, policy analysis typically addresses multiple goals. These include substantive goals (the northwest cell of Table 1), those things analysts argue are desirable for society and often instrumental goals (the southwest cell), such as political feasibility, which are necessary for policies to realize substantive goals.

Policy Analysis Modes

The appropriate mode of policy analysis depends on answers to two questions (Vining and Boardman, 2006). First, how many distinct goals are relevant to policy choice? The framing of policy problems as market and government failures directly raises efficiency as a commonly relevant goal. Sometimes, it is the only goal. More often it is one of several substantive or instrumental goals. Whether efficiency is the only goal, one of two substantive goals, or one among many goals helps determine the appropriate mode of analysis. Second, to what extent can progress toward goals be quantified or monetized? The answer depends on both conceptual and practical considerations. Quantifying and monetizing all the relevant efficiency impacts of a policy may be conceptually possible, but not practical in terms of the data and time available to analysts. Or, although efficiency impacts can be quantified, analysts may not have sufficient information or willingness to monetize them. If a goal, such as efficiency, involves multiple impacts, then an inability or unwillingness to quantify the impacts in a common metric requires analysts to replace the single goal with multiple impact categories, effectively increasing the number of criteria across which alternatives must be compared.

Multi-Goal Analysis

The most general mode of policy analysis, multi-goal analysis, allows for multiple goals to be used to assess current policy and alternatives to it (Bardach, 2012; MacRae and Whittington, 1997; Weimer and Vining, 2011). It is always appropriate when there are more than two goals. If there are two goals, then multi-goal analysis is the most appropriate method if either one cannot be quantified. Multi-goal may even be appropriate if there is one goal, but it has multiple dimensions, one or more of which cannot be quantified. Herfindahl and Kneese (1974, p. 223) make the case that it is the only appropriate approach in some circumstances: "a final approach is that of viewing the various possible objectives of public policy as being substantially incommensurable on any simple scale and therefore necessitating the generation of various kinds of information, not summable into a single number, as the basis for political decision." (Also see Dodgson et al., 2001; Hrudey et al., 2001.)

Multi-goal analysis requires the analyst to complete a number of tasks. First, the analyst must identify and justify the selection of relevant goals. Second, identify impact categories for each of the goals. Third, specify alternatives to current policy. Fourth, systematically compare each alternative in terms of each of the impact categories. Fifth, recommend one of the policy alternatives, possibly current policy, based on the comparison.

Identify and Justify Goals

Goals encompass what analysts believe society should want from good public policy. As the assessment of alternative policies will depend on the selected goals, it is important that the goals be comprehensive and appropriate. As the enumeration of goals is an intermediate product of the analysis, the analyst should explain the rationale for the inclusion of a goal or goals and for the exclusion of other plausible goals.

A strong case can be made that efficiency should always be considered as one of the relevant goals in public policy analysis (Vining and Weimer, 2005). Thoughtful people would almost certainly agree that scarce resources should be used as efficiently as possible. Moving beyond efficiency, however, requires substantive knowledge of the specific policy area and the likely consequences of policy alternatives. Although little can be said in the abstract about what other goals to include in an analysis, there is one clear rule: Every valued impact of the policy alternatives must be taken into account under one of the goals. If an impact falls outside of the initially identified goals, then it is necessary to add an additional goal to take account of it. One implication of this rule is that changes in the set of alternatives may require a chance in the identified goals. Of course, if all the policy alternatives have the same impact relevant to a goal, then the goal is redundant with respect to comparison and should be eliminated.

In many policy contexts, equity is also an appropriate policy goal (Farrow, 2011; Myers, 2002; Rosen, 1975; Stone, 2012; Vining and Weimer, 2009). Indeed, equity has been codified in US federal regulatory practice as a goal. The United Kingdom *Green Book* (H.M. Treasury, 2003) also explicitly requires that equity be considered (and quantified) 'as much as possible' in assessing policy alternatives. The *Green*

Book suggests that equity not only be assessed for different income groups, but also for different age, gender, health status, and ethnic groups as well as by location. It is important to note that while efficiency and equity often involve a trade-off (increments of efficiency come at the expense of increments of equity), this is not always the case, especially in social policy arenas (for a further discussion of this, see Vining and Weimer, 2009, p. 2–4).

Apart from the goals of efficiency and equity, there is a legitimate instrumental reason to include the (net) impact on government revenue of policy alternatives as a policy goal, or sometimes as a constraint, frequently labeled 'revenue–expenditure impact.' The two major US (federal) policy analysis agencies, the Government Accountability Office and the Congressional Budget Office, posit three goals in most of their analyses: efficiency, equity, and impact on government revenues. An important rationale for the inclusion of revenue–expenditure impact for many governmental entities performing policy analysis is that they have effectively fixed budgets. Although other levels of governments may have the power to increase revenues through additional taxing, many subnational governments are constitutionally not permitted to run deficits. Thus, the impact of policies on government expenditures is often of crucial practical concern.

Ethical concerns often suggest other goals (Adler and Posner, 1999). For example, some argue that public policy should promote virtues, such as altruism or humane treatment of animals, not because of any social value they produce but rather because they are intrinsically desirable. Others argue for goals based on the value of preserving the natural environment, which might go beyond facilitating human use of the environment or even ecological goals such as maximizing the number of surviving species.

Finally, and most controversially, analysts have to consider whether to include political feasibility as a goal (or, again, at least as a constraint). The argument that political feasibility is an appropriate goal recognizes the reality that all major public policy decisions require the approval of political actors (Webber, 1986). Most policy analysts must take this reality into account in choosing among policy alternatives. However, analysts may sometimes ignore immediate political feasibility to initiate dialogue about an alternative that may become politically feasible in the future.

Specification of Impact Categories

Creating a comprehensive scorecard for comparing alternatives often requires analysts to specify impact categories for assessing progress toward multidimensional goals. For example, efficiency in the context of a public health problem may involve mortality, morbidity, and use of medical resources. If all three could be monetized, then their sum would incorporate all three impacts in a single efficiency measure. However, if mortality and morbidity cannot be monetized with available resources, then the goal of efficiency would require three impact categories for assessing alternatives.

Multiple impact categories can be required for other goals as well. For example, consider the goal of equal opportunity for political participation. Policies can have differential impacts on low-income groups and ethnic minorities. Even an instrumental goal, like political feasibility, may require multiple impact categories. For example, the analyst may believe that both public acceptability and impact on organized interest groups are relevant to the political feasibility of policy alternatives, but not have enough knowledge to assess their relative importance in making an overall prediction.

Specification of Policy Alternatives

The analyst begins with at least one policy alternative: current policy. Nonetheless, specifying current policy requires an explicit effort to identify the laws, norms, practices, procedures, and constitutional provisions that are relevant to the policy problem. The analyst must then develop and fully specify alternatives to current policy. An alternative is fully specified when its elements are defined in sufficient detail to allow predictions of its impacts. For example, predicting the impact of a higher alcohol tax on consumption requires a specification of at least the new level of the tax and the alcohol products to which it would apply.

How do analysts identify policy alternatives? Often organizations and other advocates offer alternatives that advance their interests. Analysts may find some of these alternatives worthy of development and may choose to include others simply because they will receive attention in the political process. More generally, the analyst can borrow, tinker, forward map, or backward map (Weimer, 1993).

One strategy, especially for analysts working for subnational governments, is to borrow either adopted or proposed policies from other jurisdictions (Schneider and Ingram, 1988). In the United States, organizations such as the National Conference of State Legislatures and the National Governors Association facilitate comparison of policies across states. Once a policy alternative is identified, whether borrowed or created anew, it provide a starting point for creating new alternatives through tinkering with elements of the alternative (May, 1981). Tinkering involves identifying the basic elements of the policy, identifying possible options for replacing each element, and then constructing new alternatives by assembling viable combinations of the options. For example, one version of a recycling program might involve no requirement for the separation of the recyclables by householders, weekly pickup, and processing by a city facility. Another alternative could instead specify separation by householders of paper from other recyclables, pickup every two weeks, and processing by a contractor. Just these two variations for each of the three elements would require the analyst to analyze eight possible alternatives.

Framing policy problems as market or government failures provides alternatives through "forward mapping" (Weimer and Vining, 2011): Each of these failures has associated with it common policy responses that can be the starting point for the development of policy alternatives. For example, a framing of a policy problem as resulting from a negative externality immediately suggests either a response based on internalizing the externality either through a tax on the externality producer or through a regulatory regime that reduces the harm produced by the externality. Of course, considerable specification will be needed to convert the generic policy response into a policy alternative for which the analyst can predict impacts.

Rather than through forward mapping, analysts may also develop alternatives through "backward mapping" (Elmore, 1979–80). Backward mapping starts with behavior that policy

seeks to change, then works back through a causal change to an initial policy decision. For example, consider the problem of rising municipal refuse costs. An analyst may seek to reduce or dampen this rising cost by encouraging residents to separate yard waste from their other household waste, usually by placing it in specially marked bags. A requirement for separation and a fine for not doing so, coupled with weekly pickup during the growing season, is hypothesized to change waste disposal behavior. Working backward toward a policy requires that bags must somehow be made available to residents, say through supermarkets. Some feasible method, such as allowing supermarkets to print advertising on the bags, must be found to provide the appropriate incentives for this altered behavior. The city council must specify the bag characteristics, authorize advertising on it, and provide resources to the sanitation department for pickup and rural disposal of the yard waste, and impose the bagging requirement. Reversing the behavioral causal chain results in a policy alternative to consider, beginning with the necessary actions by the city council.

Comprehensive Comparison

Comparing policy alternatives requires that the impacts of each policy be assessed in terms of each of the goals (or its impact categories). Analysts often find that the construction of a goals/alternatives matrix facilitates this comparison and its communication to clients. Typically, the rows of a matrix are labeled with goals (and impact categories) and the columns by policy alternatives. The cells then contain the predicted impacts of the policies. Analytic comprehensiveness requires that predictions be placed in all cells. The goals/alternatives matrix facilitates the identification of the various trade-offs among the identified alternatives in terms of achievement of the goals. Although it is possible to develop point systems or other criteria for providing summary assessments of alternatives (Davey et al., 1994), these approaches tend to be arbitrary and often obscure trade-offs. Therefore, they should be used with caution.

Making a Policy Recommendation

The final step in a comprehensive multi-goal analysis requires the analyst to make an explicit argument in favor of the alternative being recommended. This argument addresses the various trade-offs identified in the comparison. Clear and transparent arguments increase the chances that clients will benefit from the analysis even if they reject the recommended alternative. Notably, some government agencies (such as the CBO) do not include this final step.

Efficiency Analysis

When efficiency is the only goal, or one of two goals, analysts may be able to employ one of two particularly valuable analytical modes: cost–benefit analysis and cost–effectiveness analysis. Cost–benefit analysis is feasible and desirable when efficiency is the only goal and all impacts of alternatives relevant to efficiency can be monetized. Cost–effectiveness analysis is feasible and desirable when either (1) all but one efficiency impact can be monetized and the remaining efficiency impacts can be quantified or (2) all efficiency impacts can be monetized and the only other (non-efficiency) goal can be quantified.

Cost–Benefit Analysis

Cost–benefit analysis seeks to assess efficiency by aggregating all valued impacts of policy alternatives relative to current policy (Boardman et al., 2011; Farrow and Zerbe, 2013; Vining and Weimer, 2009). The guiding principle for monetizing impacts is the willingness of those affected to pay to obtain or avoid the impacts. Analysts have developed a variety of methods for predicting willingness to pay, ranging from drawing on preferences revealed by observable behavior in markets at one extreme to preferences about public goods elicited from surveys of relevant populations at the other. These methods allow analysts to estimate the net benefits of proposed policies relative to current policy. Among all the policies with positive net benefits, the set with the largest net benefits maximizes efficiency.

The history of the use of cost–benefit analysis in the United States begins with federal legislation supporting harbor, canal, and other infrastructure development at the turn of the twentieth century that required the Army Corps of Engineers to assess the costs and benefits of proposed projects (Hammond, 1966). Subsequently, efforts by environmental economists to apply it to natural resource projects led to the development of 'stated preference' and other methods that have since been employed in other policy areas. However, the greatest impetus for the actual use of cost–benefit analysis was President Reagan's Executive Order 12291 calling for Regulatory Impact Analyses, which require agencies to conduct cost–benefit analyses of proposed regulations that would have an effect on the economy of more than $100 million annually. Subsequent executive orders by Presidents Clinton and Obama extended the proposed rules subject to such analysis and required agencies to conduct retrospective analyses of already adopted rules. Although some states have adopted requirements for the application of cost–benefit analysis to specific policy proposals, such as proposed regulations or insurance mandates, there has been less routine application of cost–benefit analysis at the subnational government level. A major exception is the cost–benefit model developed by the Washington State Institute for Public Policy, which has been adopted in a number of other states with foundation support (Pew Charitable Trusts and MacArthur Foundation, 2014).

Cost–benefit analysis is seeing increased use in other countries as well. For example, the United Kingdom has recently increased emphasis on monetization for project appraisal and now requires "Wherever feasible, attributing monetary values to all impacts of any proposed policy, project and programme…" (H.M. Treasury, 2003). As economists and other researchers find ways to estimate "shadow prices," or imputed per unit social values, for various policy impacts, the feasibility of monetization will increase, as will the range of policies for which cost–benefit analysis is possible (Vining and Weimer, 2010).

Cost–benefit analysis serves as an appropriate decision rule only if efficiency is the only goal – a condition that is very rare in addressing major policy problems. However, it can also be viewed more generally as a protocol for assessing efficiency. Thus, cost–benefit analysis can be "embedded" within a multi-goal analysis in which efficiency is one of the goals (Vining and Boardman, 2006). Indeed, this is the role that

cost–benefit analysis actually plays in Regulatory Impact Analyses and in almost all other applications.

Cost–Effectiveness Analysis

When efficiency is the only goal, the ideal application of cost–benefit analysis requires that all impacts affecting efficiency be monetized. However, if one of the impacts can be quantified, but analysts are unable or unwilling to monetize it, then cost–effectiveness analysis may be helpful in identifying relatively efficient policies. For example, although there is a large and reasonably consistent literature on people's monetized willingness to pay for reductions in mortality risk (Viscusi and Aldy, 2003), many health policy analysts decline to apply this willingness to pay to monetize mortality reductions.

Cost–effectiveness analysis involves estimating the net incremental cost and net incremental effectiveness in moving from a reference policy to an alternative. For example, a program to increase vaccinations through a new public service announcement would estimate incremental cost as the cost of the new program and incremental benefit as the predicted number of additional vaccinations. A cost–effectiveness ratio for this program would be the ratio of the incremental cost to the incremental benefit.

It is important to note that cost–effectiveness ratios do not necessarily identify the most efficient program. Only if either all alternatives have the same incremental cost or all alternatives have the same incremental effectiveness does the cost-effectiveness ratio provide an unambiguous basis for ranking programs in terms of efficiency. The latter often holds in defense analyses in which military planners decide on some level of effectiveness, such as the "number of tons of supplies that can be delivered to a conflict theater in a certain time," and alternatives for achieving it are compared in terms of cost. In the more common situation in which both incremental costs and incremental effectiveness vary across alternatives, the relative efficiency depends on the shadow price applied to effectiveness – a value the analyst is apparently unwilling to specify. Further, comparison of cost–effectiveness ratios can be misleading if analysts fail to remove dominated alternatives, those alternatives with either the same effectiveness but higher cost or the same cost but lower effectiveness than other alternatives (Gerber, 2000). Consequently, considerable care must be taken by analysts in using cost-effectiveness to assess efficiency.

Cost–effectiveness may also be applied when all efficiency impacts can be monetized and there is one additional goal that can be quantified. For example, policies that would reduce the number of families living in poverty, a distributional goal, might involve different net social costs. These policies might be assessed in terms of net social costs per family removed from poverty. However, all the caveats that apply to the use of cost–effectiveness ratios in the case of a single goal of efficiency apply in this extended application as well.

Incomplete Analyses

Several commonly employed modes of analysis, although typically incomplete in their own right, are often of some value in developing an effective multi-goal analysis. Probably the most important of these in practice are fiscal analysis, stakeholder analysis, and program evaluation.

Fiscal Analysis

Many legislatures require analysts in supporting organizations to assess the fiscal impacts of proposed policies (essentially revenue–expenditure analysis). Simple analyses consider fiscal impacts only in the current fiscal year; more sophisticated analyses predict fiscal impacts over a longer time horizon. Sometimes fiscal analyses, such as those done by analysts for specific agencies, are even less connected to valued outcomes because they limit attention to budgetary implications for only that agency. In any event, unlike cost–benefit analysis, fiscal analysis counts only government revenues as benefits and only government revenues as costs. Consequently, net fiscal impact is not a measure of efficiency. Nonetheless, net fiscal impact is often a relevant instrumental goal in multi-goal analysis so that a well-done fiscal analysis may be a valuable component of a more comprehensive analysis.

Stakeholder Analysis

Various stakeholders, individuals, groups, and organizations, affect policy choice. Stakeholder analysis assesses the prospects for policy change by considering "the position, interest, influence, interrelations, networks, and other characteristics of stakeholders" in the past, present, and future (Brugha and Varvasovsky, 2000, p. 239). Its primary focus is predictive rather than normative. Nonetheless, stakeholder analysis may be useful in helping analysts predict political feasibility, a common instrumental goal in multi-goal analyses. Stakeholder analysis is also often useful in assessing the likelihood of successful implementation and adjusting policy designs to increase it. Implementation can be thought of as an assembly process where various stakeholders control the essential elements (Bardach, 1977). Understanding the interests and resources of these stakeholders helps analysts anticipate the extent to which they will cooperate in providing essential elements as well as strategies for increasing cooperation or getting along without it when it cannot be secured.

Program (Policy) Evaluation

Sometimes policy analysts, but more often policy researchers, assess the impacts that have been produced by some policy, commonly a program that employs some resources to achieve some policy goal. Although the assessments often cannot go beyond description, sometimes the evaluations of policies to help individuals are planned so that potential participants can be randomized into treatment (in program) and a control (not in program) or groups facilitating an assessment of causality – the likelihood that the program produced the observed outcomes. Absent such randomization, researchers must be clever in identifying valid comparison groups and employing statistical methods to assess causality (Langbein, 2012).

Program evaluation is retrospective. It becomes policy analysis when it is used prospectively to inform decisions about continuation, expansion, reduction, or termination of the evaluated program. It may also inform decisions about replicating or terminating similar programs, especially if it is one of many evaluations whose findings can be combined in a meta-analysis

(Lipsey and Wilson, 2001). Unfortunately, program evaluation done by policy researchers often fails to be comprehensive in its assessment of impacts, often focusing narrowly on the primary program goal and typically ignoring costs or even a systematic catalog of the required resources needed to implement the program. Thus, although it can be extremely valuable in helping analysts predict some of the program impacts, it often requires analysts to gather additional information to support policy decisions.

The Use of Policy Analysis

The primary focus of this article has been on the normative value of policy analysis. Thus far, it has not addressed the question: How widely is policy analysis actually used in policy making? One thing is clear, its direct impact is not clear! As Nilsson et al. (2008, p. 335) emphasize, "three decades of policy-analytic research reminds us that the inter-relationship between assessment tools, the evidence they reveal, and their use by policy makers is unlikely to be straightforward." (See also Weiss, 1979; Shulock, 1999; Radaelli, 2004; Howlett and Lindquist, 2007.) Additionally, the use and the directness of use of formal policy analysis vary widely across countries and jurisdictions. However, it is certainly the case that most developed countries have expanded their production of analysis over the last two decades.

In the United States, Whiteman (1995, p. 181) concludes, based on a study of communication surrounding four policy issues in the areas of health and transportation before the US Congress, that: "The results ... indicate that policy analysis clearly does flow through congressional communication networks. In three of the four issues examined, analytic information played a significant role in congressional deliberations." Almost certainly, these analyses have influenced members of Congress and their voting.

The most important use of policy analysis by the US federal government, however, is in the federal regulatory arena because of a series of Presidential executive orders requiring it (see Vining and Weimer, 2014). These executive orders routinely require the use of cost–benefit analysis (or something close to it) to assess new rules proposed by regulatory agencies as well as to review state and local government proposals for federally supported infrastructure projects. Although this oversight of the rulemaking process may appear somewhat narrow, it is important to recognize that much of the content of federal policy is set through rulemaking and that some of the proposed regulations would provide benefits and impose costs that amount to billions of dollars.

Legislatively, an important milestone was the passage of the Treasury and General Government Appropriations Act (P.L. 106-554). This Act requires the Office of Management and Budget (OMB) to report annually to Congress on the benefits and costs of federal regulations. The 2012 report documents that, during the first 10 years of reporting, 103 rules were both economically significant and featured cost–benefit analyses that monetized substantial portions of the reported benefits and costs (OMB, 2012).

At the state level in the United States, the use of policy analysis is highly variable. The impact of policy analysis – specifically cost–benefit analysis and cost–effectiveness analysis – appears to be significant in some states, such as Washington (Vining and Weimer, 2014).

Nilsson et al. (2008, p. 352) provide evidence that policy analysis does get used in the European Union countries (and the Union itself) but that, not surprisingly, "tools are likely to be selected primarily on the basis of organizational routines and standard practices and on the expectation that they will produce evidence that speaks directly to, and supports, the core beliefs of governing coalitions…to ensure that the evidence gathered buttress core interests."

Conclusion

Policy analysis in representative government can play a valuable role in specifying and predicting the consequences of alternative policies. In doing so, it can also address those interests in society that lack adequate articulation or representation in the arenas of collective choice. Efficiency is such an interest. However, because of the nature of collective choice, governments, especially representative governments, face tremendous pressures to adopt inefficient policies. Especially when viewed over a range of policies and over time, greater allocative efficiency almost always diffuses to benefit almost all members of society. At the same time, however, as stakeholder analysis nearly always highlights, almost all proposed policies have distributional consequences and the political process in representative government tends to favor policies that provide benefits for organized interests and impose costs on unorganized interests. Policy analysis, as we have defined it, can help counteract this pressure. But policy tools, as with other tools, are only as useful as the intentions of those wielding them. Some progress has been made, but much more can be done (Anthoff and Hahn, 2010).

See also: Applied Social Research, History of; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Evidence-Based Practice; Implementation Science; Needs Assessment; Prevention Research.

Bibliography

Anthoff, D., Hahn, R., 2010. Government failure and market failure: on the inefficiency of environmental and energy policy. Oxford Review of Economic Policy 26, 197–224.

Arrow, K., 1951. Social Choice and Individual Values. Wiley, New York.

Adler, M., Posner, E., 1999. Rethinking cost–benefit analysis. Yale Law Journal 109 (2), 165–248.

Bardach, E., 1977. The Implementation Game: What Happens after a Bill Becomes a Law. MIT Press, Cambridge.

Bardach, E., 2012. A Practical Guide for Policy Analysis, fourth ed. Sage, Los Angeles.

Boardman, A., Vining, A., Waters, W., 1993. Costs and benefits through bureaucratic lenses: example of a highway project. Journal of Policy Analysis and Management 12 (3), 532–555.

Boardman, A., Greenberg, D., Vining, A., Weimer, D., 2011. Cost–Benefit Analysis: Concepts and Practice, fourth ed. Prentice Hall, Upper Saddle River, N.J.

Brugha, R., Varvasovsky, Z., 2000. Stakeholder analysis: a review. Health Policy and Planning 15 (3), 239–246.

Davey, A., Olson, D., Wallenius, J., 1994. The process of multiattribute decision making: a case study of selecting applicants for a PhD program. European Journal of Operational Research 72 (3), 469–485.

Dery, D., 1984. Problem Definition in Policy Analysis. University of Kansas Press, Lawrence.

Dodgson, J., Spackman, M., Pearman, A., Phillips, L., 2001. Multi-Criteria Analysis: A Manual. Department of the Environment, Transport and the Regions (DTLR) and Office of the Deputy Prime Minister (ODPM), London.

Elmore, R.F., 1979–80. Backward mapping: implementation research and policy design. Political Science Quarterly 94 (4), 601–616.

Farrow, S., 2011. Incorporating equity in regulatory and benefit-cost analysis using risk-based preferences. Risk Analysis 31, 902–907.

Farrow, S., Zerbe Jr., Richard O. (Eds.), 2013. Principles and Standards for Benefit-Cost Analysis. Edward Elgar, Northhampton, MA.

Gerber, A.M., 2000. Advances in cost–effectiveness analysis. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics. Elsevier, New York, pp. 181–221.

Glazer, A., Rothenberg, L.S., 2001. Why Government Succeeds and Why It Fails. Harvard University Press, Cambridge.

Goldhamer, H., 1978. The Adviser. Elsevier, New York.

Hammond, R.J., 1966. Convention and limitation in benefit-cost analysis. Natural Resources Journal 6, 195–222.

Herfindahl, O., Kneese, A., 1974. Economic Theory of Natural Resources. Charles Merrill, Columbus, OH.

Howlett, M., Lindquist, E., 2007. Beyond formal policy analysis: governance context and analytical styles in Canada. In: Dobuzinskis, L. (Ed.), Policy Analysis in Canada: The State of the Art. University of Toronto Press, Toronto, pp. 86–115.

Hrudey, S., Adamowicz, V., Krupnick, A., McConnell, J., Renzi, P., Dales, R., Lippmann, M., 2001. Report of an Expert Panel to Review the Socio-economic Models and Related Components Supporting the Development of Canada-Wide Standards for Particulate Matter and Ozone to the Royal Society of Canada. Royal Society of Canada, Ottawa.

Langbein, L., 2012. Public Program Evaluation. A Statistical Guide, second ed. M.W. Sharpe, New York.

Lipsey, M.W., Wilson, D.B., 2001. Practical Meta-Analysis. Sage, Thousand Oaks.

MacRae, D., Whittington, D., 1997. Expert Advice for Policy Choice. Georgetown University Press, Washington, D.C.

May, P.J., 1981. Hints for crafting alternative policies. Policy Analysis 7 (2), 227–244.

Myers, S., 2002. Presidential address—analysis of race as policy analysis. Journal of Policy Analysis and Management 21 (2), 169–190.

Nilsson, M., Jordan, A., Turnpenny, J., Hertin, J., Nykvist, B., Russel, D., 2008. The use and non-use of policy appraisal tools in public policy making: an analysis of three European countries and the European Union. Policy Sciences 41 (4), 335–355.

Office of Management and Budget (Office of Information and Regulatory Affairs), 2012. Report to Congress on the Benefits and Costs of Federal Regulations and Unfunded Mandates on State, Local, and Tribal Entities. http://www.whitehouse.gov/sites/default/files/omb/inforeg/2012_cb/2012_cost_benefit_report.pdf.

Pew Charitable Trusts and MacArthur Foundation, 2014. Results First in Your State. http://www.pewstates.org/uploadedFiles/PCS_Assets/2013/Results%20First%20In%20Your%20State%20brief.pdf.

Radaelli, C., 2004. The diffusion of regulatory impact analysis: best practice or lesson drawing? European Journal of Political Research 43 (5), 723–747.

Radin, B.A., 2013. Beyond Machiavelli: Policy Analysis Reaches Midlife, second ed. Georgetown University Press, Washington, D.C.

Rosen, H., 1975. The role of cost–benefit analysis in policy making. In: Peskin, H.M., Seskin, E.P. (Eds.), Cost–Benefit Analysis and Water Pollution Policy. Urban Institute, Washington, D.C, pp. 367–368.

Schneider, A., Ingram, H., 1988. Systematically pinching ideas: a comparative approach to policy design. Journal of Public Policy 8 (1), 61–80.

Shulock, N., 1999. The paradox of policy analysis: if it is not used, why do we produce so much of it? Journal of Policy Analysis and Management 18 (2), 226–244.

Stone, D., 2012. Policy Paradox: The Art of Political Decision Making, third ed. W.W. Norton and Company, New York.

Svenson, O., 1979. Process description of decision making. Organizational Behavior and Human Performance 23 (1), 89–92.

Treasury, H.M., 2003. Green Book: Appraisal and Evaluation in Central Government. TSO, London.

Vining, A., Boardman, A., 2006. Metachoice in policy analysis. Journal of Comparative Policy Analysis 8 (1), 77–87.

Vining, A., Weimer, D., 2005. Efficiency and cost–benefit analysis. In: Peters, G., Pierre, J. (Eds.), Handbook of Public Policy. Sage, London.

Vining, A., Weimer, D., 2009. Assessing the costs and benefits of social policies. In: Weimer, D., Vining, A. (Eds.), Investing in the Disadvantaged: Assessing the Benefits and Costs of Social Programs. Georgetown University Press, Washington, D.C, pp. 1–16.

Vining, A.R., Weimer, D.L., 2010. An assessment of important issues concerning the application of benefit-cost analysis to social policy. Journal of Benefit–Cost Analysis 1 (1), 1–38.

Vining, A.R., Weimer, D.L., 2014. Cost–Benefit Analyses of Economic Policies (2000s–Present). In: Wright, R., Zeiler, T. (Eds.), Guide to U.S. Economic Policy. CQ Press, Thousand Oaks, CA, pp. 391–401.

Viscusi, W.K., Aldy, J.E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. Journal of Risk and Uncertainty 27 (1), 5–76.

Webber, D., 1986. Analyzing political feasibility: political scientists' unique contribution to policy analysis. Policy Studies Journal 14 (4), 545–564.

Weimer, D.L., 2009. Making education research more policy analytic. In: Sykes, G., Schneider, B., Plank, D.N. (Eds.), Handbook of Education Policy Research. Routledge, New York, pp. 93–100.

Weimer, D.L., 1993. The current state of design craft: borrowing, tinkering, and problem solving. Public Administration Review 53 (2), 110–120.

Weimer, D.L., Vining, A.R., 2011. Policy Analysis: Concepts and Practice, fifth ed. Longman, New York.

Weimer, D.L., Vining, A.R., 2006. Policy analysis in representative democracy. In: Gerber, A.S., Patashnik, E.M. (Eds.), Promoting the General Welfare: New Perspectives on Government Performance. Brookings Institution Press, Washington, D.C, pp. 19–39.

Weiss, C., 1979. The many meanings of research utilization. Public Administration Review 39 (5), 426–431.

Whiteman, D., 1995. Communication in Congress: Members, Staff, and the Search for Information. University of Kansas Press, Lawrence.

Wolf, C., 1979. A theory of nonmarket failures. Journal of Law and Economics 22 (1), 107–139.

Relevant Websites

http://www.appam.org/ – Association for Public Policy Analysis and Management.
http://www.cbo.gov/ – Congressional Budget Office.
https://research.tufts-nemc.org/cear4/ – Cost–Effectiveness Analysis Registry.
http://www.gao.gov/ – Government Accountability Office.
https://www.gov.uk/government/organisations/hm-treasury – H.M. Treasury.
http://www.regulations.gov/#!home – Regulations.gov.
http://benefitcostanalysis.org/ – Society for Benefit-Cost Analysis.
http://www.wsipp.wa.gov/ – Washington State Institute for Public Policy.

Policy Analysis: Evidence Based Policy-Making

Brian W Head, The University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Systematic use of rigorous evidence in the policy process is widely expected to produce more accurate policy advice. However, these scientific aspirations will always be limited by democratic political debates, stakeholder lobbying, and popular opinion. Policy analysts in government agencies attempt to utilize research and evaluation evidence relevant to the feasibility and effectiveness of current and proposed service programs and regulatory regimes. But their capacity to provide policy advice enriched by 'evidence-based' analysis is dependent on the availability of reliable socioeconomic data, use of advanced analytical techniques, investment in program evaluations, and support for transparent policy processes by administrative and political elites.

Introduction

There is a widely supported principle that public policy making should be based on the best available knowledge concerning the issues under consideration. In recent decades, the concept of 'evidence-based policy' has been more precisely elaborated and advocated. The essential meaning of the term is that objective knowledge from scientific research including rigorous evaluation studies should occupy a central place in policy decision making. Seeking rigorous and reliable knowledge and promoting its utilization within the policy process are core features of an evidence-based policy approach. Considerable efforts have been made in many countries to invest in the systematic data, the analytical skills, and the decision processes that would provide better foundations for a more evidence-informed approach. Evidence-based policy is an attractive concept for professionals concerned with building robust information bases and improving the techniques for analysis and evaluation. Nevertheless, alternative decision-making principles and practices – based on appeals to policy precedents, leaders' authority, or political self-interest – have been much more prevalent in most countries at most periods in time.

Historical Development and Overview

Over the last two centuries, there has emerged a growing conviction that reliable empirical knowledge, based on systematic analysis of evidence, could become a powerful instrument in providing more accurate and effective advice to governmental leaders. The scientific-technical thrust of this doctrine was initially directed toward the more efficient governance of nation-states in their quest to create industrial and commercial wealth. Technological sciences generated improvements in transportation, military capability, energy and water systems, and infrastructure for the growth of cities. In addition, epidemiological studies facilitated breakthroughs in public health, such as sanitation systems and control of communicable diseases. Since the 1920s, and especially since the 1950s, the empirical social sciences – such as sociology, economics, political science, social psychology – have opened up new dimensions and new tensions in the relationships between science and policy making. Historically, the enlightenment ethos has provided the normative underpinning of both the technological and social sciences, namely, the belief that scientific understanding of social and natural forces can generate economic wealth and social progress.

In advanced Western nations, the social sciences developed rapidly in the 1940s and 1950s, exemplified by Keynesian economics and welfare-oriented social and educational planning. In the United States, the 'policy sciences' championed by Lasswell and colleagues were clearly linked to the value bases of social progress and democracy (Lerner and Lasswell, 1951). Social scientists became deeply involved in the ambitious public programs addressing social welfare, educational reform, urban renewal, and race relations in North America, Europe, and Australasia over several decades from the 1960s and beyond. However, the results were often disappointing, and some of the blame fell on the inadequacies of social research, as well as the poor coordination and implementation capacities of government agencies.

How could the social sciences become more precise, reliable, and thus useful? By the 1970s, a growing group of leading social scientists were advocating the importance of rigorous behavioral and experimental methods, insisting that the social sciences should raise their 'scientific' standards as a basis for providing more accurate and trustworthy advice. Analysis of quantitative data derived from social experiments was strongly advocated as a means of providing more precise and reliable evidence for decision makers (e.g., Campbell, 1969). This focus on advocating the potential benefits of greater rigor and the use of advanced analytical methods, generated support in some government agencies, research centers, and research funding bodies, especially in the United States (Radin, 2000). A significant professional grouping of policy analysts and program evaluators developed, in step with the growing influence of the 'evidence-based' policy movement.

The UK central government, under Prime Minister Blair from 1997, attempted to develop a more coherent approach to policy development, championing evidence-based policy as a major element in developing the fresh thinking and increased policy capability that would be required by a reformist government (UK Cabinet Office, 1999; UK Treasury, 2007).

This approach became institutionalized through new public service units and cross-departmental teams working on complex issues. It was also reflected in initiatives for skills development and training for public servants and evaluators. Importantly, the UK Economic and Social Research Council (ESRC) also committed substantial funds to broad long-term research programs on evidence-based policy and practice, thereby linking the spheres of policy making, evaluation research, and improved professional practice in health, education, and social care. Issues targeted by the ESRC included the barriers and facilitators to research utilization, and programs to build techniques and capacities for evidence-based policy and professional practice. The diverse findings of these research programs have played a crucial role in legitimating rigorous approaches to applied research (Nutley et al., 2007). Moreover, international agencies, such as the Organization for Economic Cooperation and Development (OECD, 2010), have also been strong advocates for taking evidence-based approaches to policy development, innovation, and regulatory reform; while the European Commission and several agencies responsible for international development aid have also taken a strong interest in the effectiveness of funded programs.

The demand for rigorous social and economic research stems largely from government agencies gathering information on program outcomes, and from legislative committees that oversight the effectiveness of agency performance. In the United States, several state legislatures and Congressional committees have developed requirements for evidence-based justification for many social welfare programs. The federal Office of Management and Budget has repeatedly warned that evidence of cost-effectiveness is crucial for federal funding of programs, and this is all the more important in times of fiscal austerity. Government agencies and legislative bodies draw on diverse external bodies for expert information and advice, but there is also substantial capacity inside the public sector for gathering and processing social and economic information for the policy process. In federal nations (such as Germany, Canada, and the United States), policy analysis and evaluation capacity is required at both national and subnational levels, leading to some challenging problems of coordination and even competition in some public policy areas.

Evidence-based policy has become a catchcry in recent decades in those countries (especially in the OECD group) that take seriously the quality of policy analysis and the evaluation of program effectiveness. Where the political culture allows substantial elements of transparency and rationality in the policy process, decision makers are more likely to support increased utilization of policy-relevant knowledge. Supportive research cultures encourage and foster a commitment to rigorous methodologies for generating a range of policy-relevant evidence and analysis. Evidence-based policy has been a growth area in policy research because of these demand and supply factors: the needs of government decision makers for certain types of information about problems, programs and the effectiveness of options, and the increased supply of tools and techniques for analysis and evaluation of policy that emerged over recent years. The policy areas most committed to the principles of evidence-based policy are medical and health programs, technology and innovation policy, economic growth and stabilization, environment and natural resource management, education and training, social welfare, and criminal justice. But in all these areas, there are strong stakeholder views and vested interests that constrain and reshape the ways in which problems are framed and solutions proposed.

The reputation of clinical medicine in pioneering 'evidence-based' approaches remains very powerful. Thus, much of the literature on 'evidence-based' social evaluation begins by asking whether rigorous health-based methods are fully applicable to other areas of policy, such as social welfare, schooling, juvenile justice, and community health. While much of this literature concludes that the uptake of 'evidence-based' approaches is warranted, it is also noted that there are significant variations across policy domains in relation to how policies should be assessed and modified in the light of better information (Nutley et al., 2007; Bochel and Duncan, 2007).

Alongside the focus on using evidence for improving public policy decision making, the concept of 'evidence-based practice' has also been widely adopted, especially in health and social care. For example, major efforts have been made to 'translate' the findings and lessons of health research into clinical guidelines to be used by health care professionals (see, for example, the UK Web site http://www.nice.org.uk). Closely connected to this 'translational' research, the associated fields of 'implementation' research are concerned with how scientific findings can be fruitfully translated into a wide range of complex organizational contexts and professional settings, and how program evaluations can be used to improve policy and practice (see, for example, the journal *Implementation Science* and the Web site of the Coalition for Evidence-Based Policy http://coalition4evidence.org). However, the challenge of extracting 'lessons' from research findings and adopting them successfully in professional practice entails complex issues of interpretation, education, relationships, and collaboration across professional boundaries.

Forms of Evidence for Policy

The quest for evidence-based policy involves carefully using the findings of rigorous research to improve the quality of public policy making and program review. The availability of solid evidence may also better inform policy debates in the wider community. Good-quality research is intended to improve the accuracy of advice concerning the efficiency and effectiveness of policy settings and future alternative options. To underpin the scientific character of data analysis and its translation into policy advice, a substantial guidance literature considers methodological questions concerning data validity and analytical reliability. Specialists in the design of data collection and statistical analysis are crucial for ensuring the quality of information used in applied research on specific problems. Clearly, there are many kinds of relevant data, and many sources of evidence. For researchers and evaluators, a range of methods is available for analyzing information and for interpreting research and evaluation findings (Davies, 2012).

Perspectives on Evidence Reliability

In areas of knowledge uncertainty and policy debate, the quality and reliability of evidence should be important. The

most fiercely debated notion in 'evidence-based policy' is the claim that there is a 'hierarchy' in the reliability of research-based evidence. According to this viewpoint, the highest ranking is accorded to those findings derived from rigorous experimental methodologies, similar to those championed in medical research. This viewpoint is broadly consistent with the quantitative thrust in several social science disciplines, such as experimental criminology, or the cost-benefit analysis utilized in program evaluation. A more specific version of this argument is the claim that the findings from randomized controlled trials (RCTs) – first pioneered in health research to appraise clinical interventions – are inherently superior, with the associated claim that RCTs should be more widely used in the applied social sciences. (For advocates of RCTs, see the Web site of the Coalition for Evidence-based Policy http://coalition4evidence.org and the Web site of the Campbell Collaboration http://www.campbellcollaboration.org.)

In order to establish reliable knowledge in a policy or program area, including patterns of social causality and the effects of interventions, it is necessary to undertake 'systematic reviews' of available research, paying special attention to the findings drawn from those studies that exhibit rigorous methods (Petticrew and Roberts, 2005). According to the evidence 'hierarchy,' the very reliable comparative findings of RCTs can be contrasted with less reliable findings emerging from other research designs such as single case studies, or from interviews with experienced managers and practitioners. However, some disciplines in the social sciences (e.g., social anthropology, history, law, and political science) are critical of the high status accorded to experimental trials such as RCTs, instead assigning high importance to understanding the motives, values, and perceptions of stakeholders, service providers, and decision makers. Thus, some disciplines tend to focus on qualitative evidence about the 'lived experience' of participants – understanding in depth their attitudes, intentions, and contexts for action.

The counterarguments against the primacy attributed to RCTs are both practical and methodological. First, it is noted that the RCT measurement approach requires isolating a small number of variables within a stable social context: this is difficult to achieve in many sensitive areas of policy affecting vulnerable social groups. Moreover, it may be ethically challenging to provide benefits to some members of a vulnerable group and not to others. Second, the scientific validity of quasi-experimental results may not apply, once the pilot studies are expanded and transposed into mainstream programs that entail complex interactions with other programs and with sources of external change. Third, the advocates of RCTs necessarily downplay the field experience of professionals who have contextual knowledge of what works under what conditions (Schorr and Auspos, 2003). Studies of program implementation, including the study of how clients actually experience these programs, generally show that policy managers, funders, scientists, and service users may have different perspectives on what kinds of evidence about program outcomes are most trustworthy (Pawson, 2006). It is important to address the different levels of trust in the reliability of research findings. A key challenge is therefore how to combine the practical knowledge of service delivery professionals (and their clients) and the scientific findings emerging from quantitative studies of program outcomes.

Government agencies have generally been wary of prescribing a single preferred methodology, but have often had a preference for quantitative studies of cost-effectiveness. Providing economic cost-benefit ratios is highly attractive for decision makers, who are interested in the best return on investment. For example, James Heckman demonstrated in a series of studies that return on investment was higher for early childhood education and services than for schooling as a whole (see Heckman and Masterov, 2004); this research has influenced decision makers internationally to invest more heavily in 'early years' services especially for disadvantaged families. In the United States, which has the largest concentrations of professionals working in program assessment and analysis, the mandating of particular methodologies in program appraisal has proceeded further than in most other nations, for example, by legislatures requiring rigorous evaluation as a condition of continued program funding. Understanding complex issues and associated interventions will require methods that go beyond computer-assisted data modeling. Longitudinal panel studies have become an increasingly rich source of several types of evidence. Program evaluation professionals increasingly use a variety of methods as appropriate, building bridges across the divide between qualitative and quantitative information. Mixed methods are increasingly championed by analysts and evaluators who are attempting to explain complex problems and assess complex interventions (see Woolcock, 2009).

Types and Sources of Evidence in Policy Making

Policy decisions in the real world are not deduced from empirical-analytical models, but from politics and practical judgment. The cognitive world of decision makers comprises an interplay between facts, norms, interests, and various preferred courses of action. In the real world of policy making, what counts as 'evidence' is diverse and contestable. The policy-making process in democratic countries uses the rhetoric of rational problem solving and managerial effectiveness, but the policy process itself is fuzzy, political, and conflictual (Head, 2010). A skeptical-realist view was suggested many years ago by Charles Lindblom:

> Instead of reaching 'solutions' that can be judged by standards of rationality, policy making reaches settlements, reconciliations, adjustments and agreements that one can evaluate only inconclusively by such standards as fairness, acceptability, openness to reconsideration and responsiveness to a variety of interests. And analysis in large part is transformed from an evaluative technique into a method of exerting influence, control and power which we call partisan analysis.
>
> Lindblom, 1980: 122

Lindblom perhaps went too far in discounting the 'evidence-based' vision of social sciences that would seek to inform (if not shape) the exercise of political power. But Lindblom was correct in noting that the policy process comprises many activities in which scientific rigor rubs up against power, interests, and values. In many countries that enjoy high levels of research investment, scientific knowledge does provide an important contribution to policy making.

Such research and analysis is undertaken in many institutional contexts, whether inside government agencies, or elsewhere, such as academia, think tanks, consultancy firms, and business or community organizations. In particular, the role of rigorous evaluation has gradually become more substantial in the policy and review process of many countries, in conjunction with the increasing capability of evaluation practitioners to assess the effectiveness of complex interventions.

Thus, rigorous scientific analysis constitutes one important voice in the larger world of policy and program debate. Several other types of knowledge and expertise have legitimate voices in a democratic society. These other (nonscience) knowledge cultures include political, professional, and stakeholder knowledge:

- The political knowledge, strategies, and agenda setting of political leaders and their organizations set the 'big picture' of priorities and approaches. Political ideas mobilize values and marshal information to support political objectives and to build coalitions of support. Partisan arguments and the 'spin' of political communication may become more important than objective analyses of policy effectiveness.
- The professional knowledge of service delivery practitioners, program managers, and coordinators is vital for advising on feasibility and effectiveness. Their experience in service delivery roles, and their field experience in implementing and monitoring client services, gives them unique insights into social care, education, health care, environmental monitoring, etc. They wrestle with everyday problems of effective implementation, and develop practical understandings of what works in local contexts. Their understanding of how programs may be implemented in different settings is valuable.
- Third, the experiential knowledge of service users and stakeholders is vital for 'client-focused' service delivery. Ordinary citizens may have different perspectives from those of service providers and program designers; their views are increasingly recognized as important for program evaluation in order to ensure that services are appropriately responsive to clients' needs and choices, and that service standards are set realistically. Businesses that are subject to regulatory standards provide constant feedback about the scope and burden of regulations.

Publicly funded research has become, directly and indirectly, the most important source of social science research available for governments. The preferences of government agencies and funding bodies for certain types of research can significantly influence how research is conducted. Research organizations have improved their capacities to provide research findings on topics of interest to government agencies. Policy-relevant research capacities have been consolidated into research centers in universities, consultancy firms, private sector think tanks, and not-for-profit social welfare bodies. The topics of applied research are increasingly shaped by funders' priorities.

Factors Enhancing the Use of Evidence

Evidence-based policy requires that research and evaluation findings are effectively communicated and available to be utilized by policy and program managers. However, even where the quality of research is high, the overall level of impact or utilization by decision makers is often seen as problematic or disappointing. Thus, rigorous analysis can make an important contribution, but it needs to be articulated within other aspects of the policy process such as consultation, negotiation with stakeholders, and assessing options. The political context provides both opportunities and constraints. Only a handful of policy issues are open to rethinking by decision makers at a given point in time. Some areas of policy are tightly defined by government priorities, electoral promises, and ideological preferences. Hence, the scope for new policy thinking may be limited even when major new evidence about 'what works' becomes available. Moreover, on matters of deep controversy, research findings are very likely to be appropriated and mobilized as ammunition in partisan debate.

The general conditions for building a more evidence-informed public policy-making culture include five interrelated components. First, official statistics on socioeconomic trends and public expenditure need to be systematically collected, as a baseline investment in understanding and accountability. Second, skilled personnel are required both within government and in other organizations in order to analyze and interpret the flood of information. Third, the policy system needs strong capacity to provide performance information and to undertake rigorous analysis of policy options. Fourth, the policy system needs to develop robust processes for the evaluation and review of government programs and their impact on citizens. Finally, the policy system will benefit from an open political culture and liberal flows of knowledge and opinion.

To understand the nature and extent of research uptake in the policy process, it is convenient to follow the typology suggested by Weiss (1979). She distinguished between direct (instrumental) uses or impacts of research, which may involve the direct adoption/utilization of findings (especially when research has been tailored to address a specific issue identified by decision makers), and indirect forms of influence via research contributions that enhance understandings of social processes, e.g., through new frameworks for reconceptualizing problems or solutions. Weiss also drew attention to the political or symbolic uses of research findings, when they are taken up as weapons to be deployed in favor of preferred options and against the ideas of opponents. It might be suggested that some program areas achieve a more settled character over a period of time, where a dominant paradigm is consolidated for guiding how issues are understood and addressed. Here, evidence-based arguments to refine programs incrementally within a settled policy paradigm might be quite likely to gain traction.

Roles and Relationships

Recent research has considered the strategies available for improving the interaction between the research, policy, and professional practice sectors. Researchers who seek engagement with the policy process are generally seeking influence through a variety of roles as experts, consultants, and advisors. A report commissioned by the British Academy collected evidence on perceptions about the research/policy interface. Based on interviews with researchers, policy makers, and others, several

types of potential contributions by researchers to social policy were noted:

- acting as special government advisers;
- contributing to major national enquiries or commissions;
- raising public awareness of key problems and issues;
- providing answers to specific questions through modeling or evaluations;
- providing objective analysis of what works and what does not;
- monitoring and analyzing social trends;
- providing independent scrutiny of government initiatives and developments;
- challenging current paradigms, helping to identify new approaches, concepts, and principles (see British Academy, 2008: 20).

In practice, the relationships between research producers and policy advisors are complex and variable. It is necessary to understand the factors that facilitate or hinder flows of knowledge between these groups. Given the cultural and organizational differences between governmental and academic institutions, and the different incentives shaping performance, it is difficult to create effective and enduring forms of mutual interaction and influence between the policy sector and research sector (Nutley et al., 2007; Bogenschneider and Corbett, 2010). Some academics are wary about the possible dangers of co-option and dependency that arise from close relationships, such as writing contract-research reports. Other academics seek closer engagement, but are disappointed to find that policy makers are not open minded about the value of research critical of current orthodoxies. The British Academy report listed several reasons given by policy makers as to why they avoided or ignored certain types of external research. These included perceptions that (1) their own agencies did not highly value research and did not communicate research findings within the agency; (2) research undertaken by internal staff or by contracted consultants was more likely to be seen as relevant; (3) external academic research was not seen as timely, or not seen as relevant to users' current needs; and (4) research was less likely to be used when findings were controversial or upset the status quo (British Academy, 2008: 27).

Methods to enhance interaction need to take account of the known systemic obstacles to research-informed policy approaches. Thus, research findings that reinforce policy preferences of political leaders, and that consider carefully the political and financial feasibility of policy options, are more likely to be positively received. By contrast, rigorous research that queries the favored assumptions of political leaders will be unwelcome in some quarters and may undermine trust in the neutrality of policy research. In considering methods to promote the use of rigorous evidence, the early waves of 'science communication' emphasized the need to increase dissemination as a basis for the 'transfer' or 'transmission' of research knowledge into the policy sector. This approach is now regarded as too linear and simplistic. The challenge is to consider methods for mutual understanding and influence, such as jointly managed 'learning' processes, or creating new channels and forums for mutual understanding and influence.

A number of 'bridging' strategies have been developed in recent years to ensure that both the researchers and policy makers learn more about each other's business as a basis for better utilization of policy-relevant research. It is important here to recognize the substantial role of 'in-house' research and evaluation activities inside public agencies, as well as the findings of 'external' researchers in academic and independent bodies. The channels through which rigorous evidence might influence policy making are somewhat fragile, and therefore require specific care and attention. Supply-side provision of good research about 'what works' is not enough. Potential users of research findings will pay closer attention only if they are familiar with these potential inputs, understand the advantages and limits of the information, and are in a position to make use of the findings either directly or indirectly (Nutley et al., 2007; Ward et al., 2009). Thus, the social and institutional foundations for evidence-based policy do not emerge naturally from information systems adjacent to the policy process. Capabilities and understandings need to be built within and across organizations. If the suppliers of rigorous research and the policy community are mutually ignorant or indifferent, there is little prospect of evidence-based policy becoming more robust. This situation could be exacerbated by gaps in the technical knowledge of policy makers, but is more likely to arise from deficiencies in relationships, communication, and overlapping interests. Case-study research is beginning to demonstrate in more detail the contextual factors that influence policy choices and program effectiveness. These case-studies also draw attention to factors that may facilitate or hinder research utilization by policy makers and program professionals, and ultimately throw more light on the conditions for institutional learning (Nutley et al., 2010).

Challenges and Criticisms

The rationalist perspective aims to promote evidence-based policy making as a scientific contribution to achieving cumulative knowledge for iterative policy learning within a democratic policy process. For some countries with weak information systems, or with patronage-based political systems, this approach is seen as aspirational or unrealistic; the challenge for them is to make best use of diverse local knowledge while drawing on external expertise and continuing to build a more systematic information base for policy and performance management. Among the OECD countries, which generally have a robust information infrastructure, there are other concerns and challenges underlying attempts to build better evidence bases and strengthen the links into policy and program development.

For example, concerns have been expressed by critics who claim that evidence-based policy is inherently elitist or 'technocratic,' owing to its preference for scientific expertise and quantitative precision over other forms of knowledge. The relationships between expert technical knowledge and other forms of relevant knowledge are only weakly acknowledged by advocates of evidence-based policy making. This is unfortunate, as some of the earlier champions of progress in the social sciences (e.g., Lerner and Lasswell, 1951) recognized that democratic policy making requires community education and

strong processes for stakeholder involvement. A related concern has been the view that evidence-based approaches have become harnessed to the efficiency agendas of government rather than to the innovation and social benefit agendas. These different perspectives about rationalist policy making are grounded in different epistemologies and different views about the centrality of dialogue and participatory deliberation, and are therefore unlikely to be resolved.

A further area of concern has centered on the importance of long-term governmental commitment to improved processes for program implementation and evaluation. Ideally, the investment in evidence-based approaches would focus on important issues where research and evaluation findings could lead to better understanding of patterns, trends, and causes. In a complex policy intervention, clear program outcomes may not emerge for several years. But political leaders are often impatient for quick results, insisting that measurable outcomes be available in a short time frame. This pressure can lead to a focus on visible projects rather than on building foundations for sustainable benefits. Moreover, governments have a propensity to change a program before outcomes have been assessed, so that evaluations would be measuring moving targets.

Fortunately there are recent examples of solid commitment to program stability and to long-term evaluation. In the United Kingdom, for example, a complex program targeting early childhood development in disadvantaged communities ('Sure Start') underwent several changes to improve coordination, but the major ongoing evaluation project managed to provide an impressive series of assessments over several years (see Web site: http://www.education.gov.uk/childrenandyoungpeople/earlylearningandchildcare/evidence). Another area of major investment in policy-relevant research, in the United States and elsewhere, is the ongoing evaluation of 'welfare-to-work' programs, whereby parents in disadvantaged households are strongly encouraged to develop workforce experience and skills (e.g., Haskins, 2006).

A third challenge is how evidence-based approaches can deal with large and complex problems, rather than the simpler issues for which the variables can be specified and controlled. Experimental evaluations like RCTs work best in relatively simple environments. In complex programs with multiple objectives, such as antipoverty interventions where the clients are subjected to many sources of influence beyond the scope of the program, the challenges of accurate understanding are compounded. Moreover, many factors influence how these policy problems are framed, debated, researched, and managed. This diversity poses diverse challenges for evidence, analysis, and recommendation. The political requirement for solutions will sometimes encourage rapid generic responses, rather than the long pathway of detailed field research with careful evaluation of experience across a range of localities. Complex social problems (such as gun control, refugee resettlement, and juvenile justice) are often characterized by underlying clashes of values, which need to be handled politically rather than through larger amounts of scientific information. Conflict mediation, stakeholder coordination and provision of incentives for cooperation might need to be mobilized to improve the prospects of success in these complex areas of policy conflict.

Conclusions

The evidence-based policy movement in recent decades has been successful in expanding the scope of evidence and analysis available for decision makers, and demonstrating the advantages of a more rigorous approach to understanding trends, causality, and the effects of particular interventions. Tools for analyzing policy options and evaluating program outcomes are much more sophisticated, and more widely deployed, than some decades ago. The democratic dissemination of information and commentary, the wide accessibility of evaluation studies, and a proliferation of research-literate organizations, have all led to a more open policy-making system in many countries. In principle, this has entailed evidence-informed policy contestation and higher levels of public accountability for decision makers. In other countries that lack a democratic and pluralist political culture, the accountability benefits of evidence-based policy processes are sidelined, with the ruling elites selecting which technocratic initiatives will be developed.

In general, the large volume of policy-relevant research and evaluation has had a positive impact. Improvements in the scope and depth of research, and in the links with the policy process, are widely seen as worthwhile. However, establishing close linkages between scientific research and policy making will be a long struggle at many levels. The dream of achieving large and rapid improvements in policy and practice through closer integration with rigorous research has moderated somewhat. Advocates of closer links between rigorous evidence and policy making have pragmatically recognized the inevitability of working with a range of methodologies and forms of knowledge. Rigorous research findings are not available in many significant and emerging problem areas, so that policy decision making under conditions of uncertainty will remain the norm. The inherent institutional and cultural barriers between research, policy, and practice are considerable and long standing. Policy managers and political leaders are motivated and influenced by many factors besides research evidence. For these and other reasons, inserting research evidence into the center of policy decision making remains difficult.

Several features of democratic policy systems pose ongoing systemic obstacles to evidence-based policy making. First, the politicized context of government commitments and decision making creates dispositions in favor of certain policy ideas, regardless of objective evidence. This will remain a feature of political decision making. Second, evidence is always contestable and is often used in partisan ways. Hence, disputes about the interpretation and selection of evidence will continue. Third, public officials do not necessarily have up-to-date awareness of salient research, and much of their research-related effort may be directed at short-term problem fixing rather than at longer term analysis. Finally, the independent research community is often disconnected from the needs and expectations of public officials, with the consequence that research findings are not well targeted and not well communicated.

Thus, in the turbulent context of policy decision making, policy systems may well continue to become more evidence informed, but leaders remain unable to manage politically the large-scale problems of a globalized world, such as the policy

crises around economic stabilization, environmental stewardship, poverty, migration, and civil war. Rigorous scientific methods could serve to increase the credibility of social science approaches to major problems, but the policy system cannot become fully evidence-driven. Unrealistic expectations that better science will inevitably drive better policies have recently been scaled back, as reflected in the phrase 'evidence-informed' policy.

See also: Agenda Settting, Public Policy in; Civil Service; Governance; Policy Cycle; Policy Knowledge: Universities; Policy Networks; Quality of Government; Think Tanks.

Bibliography

Bochel, H., Duncan, S. (Eds.), 2007. Making Policy in Theory and Practice. Policy Press, Bristol.

Bogenschneider, K., Corbett, T.J., 2010. Evidence-Based Policy Making: Insights from Policy-Minded Researchers and Research-Minded Policy Makers. Routledge, London.

British Academy, 2008. Punching Our Weight: The Humanities and Social Sciences in Public Policy Making. British Academy, London.

Campbell, D.T., 1969. Reforms as experiments. American Psychologist 24 (4), 409–429.

Davies, P., 2012. The state of evidence-based policy evaluation and its role in policy formation. National Institute Economic Review 219, R41–R52.

Haskins, R., 2006. Work over Welfare. Brookings Institution, Washington.

Head, B.W., 2010. Reconsidering evidence-based policy: key issues and challenges. Policy and Society 29 (2), 77–94.

Heckman, J.J., Masterov, D.V., 2004. The Productivity Argument for Investing in Young Children. Committee for Economic Development, Washington.

Lerner, D., Lasswell, H.D. (Eds.), 1951. The Policy Sciences. Stanford University Press, Stanford.

Lindblom, C.E., 1980. The Policy-Making Process, second ed. Prentice-Hall, Englewood Cliffs, NJ.

Majone, G., 1989. Evidence, Argument, and Persuasion in the Policy Process. Yale University Press, New Haven.

Nutley, S., Walter, I., Davies, H., 2007. Using Evidence: How Research Can Inform Public Services. Policy Press, Bristol.

Nutley, S., Morton, S., Jung, T., Boaz, A., 2010. Evidence and policy in six European countries: diverse approaches and common challenges. Evidence & Policy 6 (2), 131–144.

Organization for Economic Cooperation and Development, 2010. Finland: Working Together to Sustain Success: Public Governance Review. OECD, Paris.

Pawson, R., 2006. Evidence-Based Policy: A Realist Perspective. Sage, London.

Petticrew, M., Roberts, H., 2005. Systematic Reviews in the Social Sciences. Blackwells, Oxford.

Radin, B.A., 2000. Beyond Machiavelli: Policy Analysis Comes of Age. Georgetown University Press, Washington.

Schorr, L.B., Auspos, P., 2003. Usable information about what works: building a broader and deeper knowledge base. Journal of Policy Analysis and Management 22 (4), 669–676.

UK Treasury, 2007. Analysis for Policy: Evidence-Based Policy in Practice. Government Social Research Unit, Treasury, London.

UK Cabinet Office, 1999. Professional Policy Making for the Twenty First Century. Cabinet Office, London.

Ward, V., House, A., Hamer, S., 2009. Knowledge brokering: the missing link in the evidence to action chain? Evidence & Policy 5 (3), 267–279.

Weiss, C.H., 1979. The many meanings of research utilization. Public Administration Review 39 (5), 426–431.

Woolcock, M., 2009. Toward a plurality of methods in project evaluation. Journal of Development Effectiveness 1 (1), 1–14.

Relevant Websites

http://www.campbellcollaboration.org – Campbell Collaboration.
http://coalition4evidence.org – Coalition for Evidence-Based Policy.
http://www.nice.org.uk – National Institute for Health and Clinical Excellence.
http://www.education.gov.uk/childrenandyoungpeople/earlylearningandchildcare/evidence – UK Children's Services.

Policy Cycle

Michael Howlett and Sarah Giest, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Analyzing policy development in terms of policy cycles and policy styles is useful for several reasons. Not only does it help to make sense out of the different models put forward earlier in the history of policy studies, such as the rational, incremental, and garbage can models of public policy decision making, it also helps to advance studies of these models by specifying the conditions under which different styles of policy making occur. This article sets out the logic of the policy cycle model and justifies its five-stage vision of that process.

The Policy Cycle Model of the Policy Process

The idea that the policy process can be thought of as a series of steps in a cyclical model of decision making was first broached systematically in the work of Harold Lasswell, a seminal figure in the development of the policy sciences at the University of Chicago and Yale University in the early 1950s (Lasswell, 1956, 1971). At present, a five-stage model of the policy process is the most commonly used, although many other versions exist with more fewer substages. In this model, 'agenda setting' refers to the first stage in the process when a problem is initially sensed by policy actors and a variety of solutions put forward (*see* Agenda Setting, Public Policy in). 'Policy formulation' refers to the development of specific policy options within government when the range of possible choices is narrowed by excluding infeasible ones and efforts are made by various actors to have their favored solution ranked highly among the remaining few. 'Decision making' refers to the third stage in which formal actors in government adopt a particular course of action. In the fourth stage of 'policy implementation,' governments put their decisions into effect using some combination of the tools of public administration in order to alter the distribution of goods and services in society in a way that is broadly compatible with the sentiments and values of affected parties. Finally, 'policy evaluation' refers to the fifth stage in the processes in which the results of policies are monitored by both state and societal actors, often leading to the reconceptualization of policy problems and solutions in the light of experiences encountered with the policy in question and the start of a new iteration of the cycle (Howlett et al., 2009).

This idea of policy making existing as a set of interrelated stages provides a general 'framework' for understanding the policy development process and points to several of the crucial temporal activities and relationships that should be examined in furthering the study of the issue. Although a useful conceptual or analytical tool, the policy cycle model has some limitations as an empirical description of policy-making reality due to its simplification of highly contingent and complex policy processes. The reality of policy making is not as systematic and linear as the model might suggest. Also, the stages are often compressed, skipped, or change their order entirely. For example, policy formulation can sometimes precede agenda setting, as 'solutions seek problems' to which they can be applied (Kingdon, 1984; Howlett et al., 2009). It also does not answer several key questions such as the actual substance of policy, the number and type of relevant actors involved in the process, the exact manner and sequence in which actual policy development processes occur, and whether there exist basic patterns of development in different issue areas, sectors, or jurisdictions (Sabatier, 1991).

Overall, when using the policy cycle idea, diligent attention has to be paid to the fact that the model is a guideline for the analysis of a far more complex reality.

Styles of Policy Behavior in the Policy Cycle

Empirical studies of policy-making processes undertaken to answer these questions and generate more precise models of the policy-making process have developed more specific knowledge about the interactions of policy actors at each stage of the cycle, summarized below. These draw attention to the limited number of common modes, or *styles*, which characterize policy processes throughout the world. A policy style is also often part of a larger 'policy regime' that emerges over time and includes not only how policy deliberations take place, but also the kinds of actors and ideas present. As such, it is useful in identifying long-term patterns in content and process of public policy making (Howlett et al., 2009). Common policy styles at each stage of the policy cycle characterize the principal components of policy dynamics. A policy style describes how governments go forward during, for example, the agenda-setting stage and which effects result from that stage for the ongoing policy-making process. They are tools for understanding the problem-solving directions of actors in the cycle and for revealing the institutional routines and procedures underlying policy processes.

Agenda Setting

In the scholarly literature on agenda setting, a useful distinction is drawn between the *systemic* or unofficial public agenda and the *institutional* or formal, official, agenda. The systemic agenda "consists of all issues that are commonly perceived by members of the political community as meriting public attention and as involving matters within the legitimate jurisdiction of existing governmental authority" (Cobb and Elder,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75031-8

1972, p. 85). This is essentially a society's agenda for the discussion of public problems, such as crime or health care, water quality or wilderness preservation. The formal or institutional agenda, on the other hand, consists of only a limited number of issues or problems to which attention is devoted by policy elites (Baumgartner and Jones, 1991; Kingdon, 1984).

Research conducted by John Kingdon into the dynamics of agenda setting in the US Congress in the early 1980s focused on the timing of agenda entrance. In his work, he focused on the role played by policy entrepreneurs both inside and outside governments in taking advantage of agenda-setting opportunities – policy windows – to move items onto formal government agendas (Kingdon, 1984). Kingdon argued that the characteristics of issues (the problem stream) combined with the characteristics of political institutions and circumstances (the politics stream) and the development of policy solutions (the policy stream) lead to the opening and closing of opportunities for agenda entrance. Such opportunities can be seized upon or not, as the case may be, by policy entrepreneurs who are able to recognize and act upon them. The policy windows resulting from this include routine, spillover, discretionary, and random types (Kingdon, 1984). Empirical evidence suggests that the frequency of occurrence of the window types varies by level of institutionalization – with the most institutionalized types occurring much more frequently than the least institutionalized. This lends the model a degree of predictability in terms of the occurrence and type of policy window.

Almost 40 years ago, the American political scientists Cobb, Ross, and Ross developed a model of typical agenda-setting styles. In their analysis, they argued that three basic patterns of agenda setting could be discerned, distinguished by the origins of a policy issue as well as the resources utilized to facilitate its inclusion on the policy agenda. In the outside initiation pattern, "issues arise in nongovernmental groups and are then expanded sufficiently to reach, first, the public [systemic] agenda and, finally, the formal [institutional] agenda" (Cobb et al., 1976). The *mobilization* case is quite different and describes 'decision-makers trying to expand an issue from a formal to a public agenda.' In this model, issues are simply placed on the formal agenda by the government with no necessary preliminary expansion from a publicly recognized grievance. In the third type of agenda setting, inside initiation, influential groups with special access to decision makers initiate a policy and do not necessarily want it to be expanded and contested in public.

Policy Formulation

Studies of policy formulation have emphasized the importance of actors interacting to develop and refine policy options for government (Freeman, 1955; Linder and Peters, 1990). But unlike agenda setting, where many members of the public may be actively involved, in policy formulation the relevant policy actors are typically restricted to those who not only have an opinion on a subject, but also have some minimal level of knowledge that allows them to comment, at least hypothetically, on the nature and feasibility of options put forward to resolve a policy problem.

The power of ideas and their stability in policy subsystems, in particular, has been a subject of much attention in studies of policy formulation in recent years. Carstensen (2011), for example, has suggested that ideas change incrementally, when new elements of meaning are added to them, resulting in a characteristic process of policy formulation in which similar ideas inform similar policy options developed over long periods of time. Focusing on the information or knowledge available to decision makers has led policy research in this vein to focus on the role of epistemic or policy communities in policy formulation (see Haas, 2012). These are loose groupings of experts who serve as knowledge 'providers' for the decision-making process – opening up opportunities for them to influence the content and timing of policies (Dunlop, 2009, p. 290; Haas, 1992).

Scholars over the years have developed a variety of taxonomies to help identify who the key actors in these policy subsystems are, what brings them together, how they interact, and what effect their interaction has on policy development. Most of these distinguish between a larger set of actors with some knowledge of the policy issue in question, and a smaller set in which actors not only have requisite knowledge, but also have established patterns of more or less routine interactions with each other (Knoke, 1993).

Membership in knowledge-based policy *communities* extends to actors such as state policy makers (administrative, political, and judicial), members of nongovernmental organizations concerned with the subject, members of the media who report on the subject, academics who follow or research the area, and members of the general public who, for whatever reason, have taken an interest in the subject (see Thatcher, 2012, Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks, Sabatier, 1987). In many issue areas, the policy community also involves members of other organizations such as businesses, labor unions, or various formalized interest groups and even international actors such as multinational corporations (Haas, 1992). A subset of these actors who interact within more formalized institutions and procedures of government are defined as members of *policy networks* (see Coleman, 2012, Policy Networks). These policy networks include representatives from the community, but are 'inner circles' of actors, who effectively hold the power to veto or disparage many policy options as untenable or infeasible.

In this view, the likely results of policy formulation are contingent upon the nature and configuration of the policy community and network in the specific sector concerned. A key variable that many observers have argued affects the structure and behavior of policy networks is their number of members, which affects aspects of networks such as their level of integration and the types of interactions members undertake with each other (van Waarden, 1992). What is important for policy communities, on the other hand, is not the number of participants in the community so much as the number of relatively distinct 'idea sets' that exist within it. This affects the nature of conflict and consensus that exist in the community and, as a result, impacts the behavior of community actors.

In open subsystems where networks have many members and communities share many idea sets, it can be expected

that a propensity for new, radical alternatives to the status quo may be generated in the policy formulation process. In closed subsystems, where networks have few members and communities are dominated by a single idea set, on the other hand, a status quo orientation is likely to emerge in the policy options developed and put before decision makers.

Although the networking framework has been criticized as being primarily a descriptive metaphor, due to the fact that "the independent variables, (of policy-making) are not network characteristics per se but rather characteristics of components within the networks" (Dowding, 1995, p. 137), it provides a general mechanism to organize the complex reality of multiple actors, institutions, and ideas found in policy formulation and generates a relatively parsimonious model of formulation styles (Howlett et al., 2009).

Decision Making

Many early studies of policy making in companies, governments, and organizations, conducted largely by the students of public and business administrations, argued decision makers attempt to follow a systematic method for arriving at logical, efficient decisions. They argued that policy makers achieved superior results when they first established a goal, explored alternative strategies for achieving it, attempted to predict its consequences and the likelihood of each occurring, and then chose the option that maximized potential benefits at least cost or risk (Cahill and Overman, 1990).

This model was 'rational' in the sense that it prescribed a standard set of procedures for policy making, which were expected to lead in all circumstances to the choice of the most efficient means of achieving policy goals. Such pure 'rational' models of decision making often thought of policy makers as neutral 'technicians' or 'managers,' who after identifying a problem would find the most effective or efficient way of solving it. Many of the latest efforts to enhance the efficiency and effectiveness of public policy decision making, such as the 'evidence-based policy movement' (*see* Policy Analysis: Evidence-Based Policy Making), focus on the application of a systemic evaluative rationality to policy problems in classic rational style.

Empirical research into decision-making processes, however, has discovered that political processes of bargaining and negotiation often outweigh 'rational' deliberations and calculations of costs and benefits in public policy decision making. Policy makers were often found to be neither necessarily neutral nor competent and other models of the public policy decision-making processes have argued this is not an accidental situation but rather an inherent and unavoidable characteristic of the policy-making exercise. Therefore, critics of the rational model, such as Herbert Simon, noted that there are definite cognitive limits to the decision makers' ability – 'bounded rationality' – to consider all possible options, which forces them to selectively consider alternatives. 'Satisficing' behavior, in which decision makers did not consider all options in an effort to maximize policy impact but rather simply agreed on a path that they found 'satisfactory,' was argued to better reflect the reality of decision making in a public policy context (Simon, 1955; Smith and May, 1980).

The well-known incremental model of policy making developed by Yale University political scientist Charles Lindblom incorporated these insights into the best known alternative to the classic rational model (Dahl and Lindblom, 1953; Lindblom, 1959). Lindblom summarized the elements of his model as consisting of 'mutually supporting set of simplifying and focusing stratagems.' These resulted, as he put it in his oft-cited 1959 article on "The Science of 'Muddling Through,' in decision makers working through a process of 'continually building out from the current situation, step-by-step and by small degrees' or 'increments'" in which policies were invariably developed through a process of 'successive limited comparisons' with earlier ones with which decision makers were already familiar (Lindblom, 1959).

Taking these insights into account, Braybrooke and Lindblom (1963) argued that different styles of decision making could be discerned depending upon the amount of knowledge at the disposal of decision makers, and the amount of change the decision involved from earlier decisions. Other authors, like Graham Allison, also developed similar models of distinct decision-making styles (Allison, 1971), but did not specify in any detail the variables that led to the adoption of a particular style.

Attempting to improve upon these models, John Forester (1984) suggested that decision making was affected by the number of agents involved in a decision, their organizational setting, how well a problem is defined, the information available on the problem, its causes and consequences, and the amount of time available to decision makers to consider possible contingencies and their present and anticipated consequences. In this model, decision makers situated in complex subsystems are expected to undertake adjustment strategies while those dealing with simple configurations of actors and ideas will be more prone to undertake search-type 'rational' strategies. The nature of the decision criteria, on the other hand, varies with the severity of the informational, time and other resource constraints under which decision makers operate. Hence, decision makers faced with high constraints will tend to favor satisficing over optimization, itself an outcome more likely to occur in situations of low constraint.

Policy Implementation Styles

Generally speaking, comparative studies have shown that governments tend to develop specific implementation styles as they go about the course of their business (Knill, 1998; Hawkins and Thomas, 1989; Kagan, 1991). These styles combine various kinds of policy instruments – such as public enterprises, regulation, or public consultation – into a more or less coherent whole, which is consistently applied in specific sectors of government activity – such as health policy or industrial policy. These policy tools provide the substance or content needed to implement what was planned in the formulation stage and decided upon afterward (Salamon, 1989).

These tools fall into two types. *Substantive instruments* such as regulation or public enterprises are those directly providing goods and services to members of the public or governments. They include a variety of tools or instruments relying on different types of governing resources – such as money, information, authority, or organizational resources (Hood,

1986) – for their effectiveness. *Procedural instruments*, on the other hand, such as public hearings or commissions of inquiry, are different from substantive ones in that their impact on policy outcomes is less direct. Rather than affect the delivery of goods and services, their principal intent is to modify or alter the nature of policy processes themselves.

Studies such as the work of Gunningham et al.'s (1998) 'smart regulation' led to the development of efforts to identify complementarities and conflicts within instrument mixes or tool 'portfolios' involved in more complex and sophisticated policy designs. Each tool mix decision combines advantages and disadvantages of each tool in its relationship to other tools as well as the effect it has on costs and benefits for government.

Why a particular combination of procedural and substantive instruments is utilized in particular policy issue areas is a key question affecting implementation and implementation studies. The preferences of state decision makers and the nature of the constraints within which they operate are among the factors affecting policy implementation and instrument choices that lead to the development of specific preferences or styles (Freeman, 1985). A state must have a high level of administrative capacity, for example, in order to utilize authority, treasure, and organization-based instruments in situations in which they wish to affect significant numbers of policy targets. When it has few of these resources, it will tend to utilize instruments like incentives or propaganda or rely on existing voluntary, community, or family-based instruments (Howlett et al., 2009). Similarly, a key feature of procedural instrument choice is a government's capacity to manipulate policy subsystems (Mayntz, 1975).

Policy Evaluation Styles

The last stage of the policy cycle is policy evaluation. For many early observers, policy evaluation was expected to consist of assessing if a public policy was achieving its stated objectives and, if not, what could be done to eliminate impediments to their attainment. Thus David Nachmias (1979) defined policy evaluation as "the objective, systematic, empirical examination of the effects ongoing policies and public programs have on their targets in terms of the goals they are meant to achieve." However, while analysts often resorted to concepts such as 'success' or 'failure' to conclude their evaluation, Ingram and Mann (1980) caution that such concepts are slippery and often highly subjective. That is, public policy goals are often not stated clearly enough to find out if and to what extent they are being achieved, nor are they shared by all key policy actors.

What is significant in the evaluative process is thus not so much ultimate success and failure, but that policy actors and the organizations and institutions they represent can *learn* from the formal and informal evaluation of policies in which they are engaged. This can lead them to modify their positions in the direction of greater substantive or procedural policy change, or it can lead them to resist any alteration to the status quo.

Policy evaluations do not necessarily result in major policy change. That is, while the concept of evaluation suggests an implicit 'feedback loop' is an inherent part of the policy cycle, in many cases this loop may not be operationalized or may result in only incremental changes to the status quo (Pierson, 1993). Path dependence, in which policies are set on 'trajectory' following some 'critical juncture' can hinder policy change and learning (Pierson, 2000). Organizational–institutional designs which limit 'absorptive capacity' are also often seen as barriers to learning from policy evaluations.

A significant variable in this regard is the capacity of an organization to absorb new information (Cohen and Levinthal, 1990). Only when state absorptive capacity is high would one expect any kind of learning to occur. If a relatively closed network dominates a policy subsystem, then learning is likely to be restricted to some form of 'lesson drawing,' in which policy makers draw lessons from past uses of specific policy instruments (Rose, 1991). Open subsystems allow for 'social learning' when administrative capacity is high and more informal evaluation takes place among all those affected by a policy (Howlett et al., 2009; Bennett and Howlett, 1992).

Conclusion: Policy Development as Policy Style

As Lasswell noted in the 1950s, envisioning policy development as a staged, sequential, and iterative process is a useful analytical and methodological device. Methodologically, such an approach reduces the complexity of public policy making by breaking down that complexity into a small number of stages and substages, each of which can be investigated alone or in terms of its relationship to any or all the other stages of the cycle. The policy cycle idea also helps to answer many key questions about public policy making regarding the effectiveness of different tools and the identification of bottlenecks in policy processes. The stages model allows for the identification of specific actors and actions in different phases of tackling a problem, which makes it easier to measure the impact of key variables on policy outcomes and to identify existing policy process styles.

Bibliography

Allison, G., 1971. Essence of Decision: Explaining the Cuban Missile Crisis. Little Brown, Boston.

Baumgartner, Frank R., Jones, Bryan D., 1991. Agenda dynamics and policy subsystems. Journal of Politics 53 (4), 1044–1074.

Braybrooke, David, Lindblom, Charles E., 1963. A Strategy of Decision: Policy Evaluation as a Social Process. Free Press of Glencoe, New York.

Bennett, Colin J., Howlett, Michael, 1992. The lessons of learning: reconciling theories of policy learning and policy change. Policy Sciences 25 (3), 275–294.

Cahill, Anthony G., Overman, E. Sam, 1990. The evolution of rationality in policy analysis. In: Nagel, S.S. (Ed.), Policy Theory and Policy Evaluation: Concepts, Knowledge, Causes, and Norms. Greenwood Press, New York.

Carstensen, Martin B., 2011. Ideas are not as stable as political scientists want them to be: a theory of incremental ideational change. Political Studies 59 (3), 596–615.

Cobb, Roger W., Elder, Charles D., 1972. Participation in American Politics: The Dynamics of Agenda-Building. Allyn and Bacon, Boston.

Cobb, Roger W., Ross, J.K., Ross, M.H., 1976. Agenda building as a comparative political process. American Political Science Review 70 (1), 126–138.

Cohen, Wesley M., Levinthal, Daniel A., 1990. Absorptive capacity: a new perspective on learning and innovation. Administrative Science Quarterly 35, 128–152.

Dahl, Robert A., Lindblom, Charles E., 1953. Politics, Economics and Welfare: Planning and Politico-Economic Systems Resolved into Basic Social Processes. Harper and Row, New York.

Dowding, K., 1995. Model or metaphor? A critical review of the policy network approach. Political Studies 43, 136–158.

Dunlop, Claire A., 2009. Policy transfer as learning: capturing variation in what decision-makers learn from epistemic communities. Policy Studies 30 (3), 289–311.

Forester, John, 1984. Bounded rationality and the politics of muddling through. Public Administration Review 44 (1), 23–31.
Freeman, J. Leiper, 1955/1965. The Political Process: Executive – Bureaucratic – Legislative Relations, rev. ed. Random House, New York.
Freeman, Gary P., 1985. National styles and policy sectors: explaining structured variation. Journal of Public Policy 5 (4), 467–496.
Gunningham, N., Grabosky, P., Sinclair, D., 1998. Smart Regulation: Designing Environmental Policy. Clarendon Press, Oxford.
Haas, Peter M., 1992. Introduction: epistemic communities and international policy coordination. International Organization 46 (1), 1–36.
Hawkins, Keith, Thomas, John M. (Eds.), 1989. Making Regulatory Policy. University of Pittsburgh Press, Pittsburgh.
Hood, Christopher, 1986. The Tools of Government. Chatham House, Chatham, NJ.
Howlett, Michael, Ramesh, M., Perl, Anthony, 2009. Studying Public Policy: Policy Cycles and Policy Subsystems. Oxford University Press, Toronto.
Ingram, Helen M., Mann, Dean E., 1980. Why Policies Succeed or Fail. Sage, Beverly Hills, California.
Kagan, Robert A., 1991. Adversarial legalism and American government. Journal of Policy Analysis and Management 10 (3), 369–406.
Kingdon, John, 1984. Agendas, Alternatives and Public Policies. Little, Brown, Boston.
Knill, Christoph, 1998. European policies: the impact of national administrative traditions. Journal of Public Policy 18 (1), 1–28.
Knoke, David, 1993. Networks as political glue: explaining public policy-making. In: Wilson, W.J. (Ed.), Sociology and Public Agenda. Sage, London, pp. 164–184.
Lasswell, H.D., 1956. The Decision Process: Seven Categories of Functional Analysis. University of Maryland, College Park.
Lasswell, H.D., 1971. A Pre-view of Policy Sciences. American Elsevier, New York.
Lindblom, C.E., 1959. The science of muddling through. Public Administration Review 19 (2), 79–88.
Linder, Stephen H., Peters, B. Guy, 1990. Research perspectives on the design of public policy: implementation, formulation, and design. In: Palumbo, Dennis J., Calista, D.J. (Eds.), Implementation and the Policy Process: Opening up the Black Box. Greenwood Press, New York.
Mayntz, Renate, 1975. Legitimacy and the directive capacity of the political system. In: Lindberg, Leon N., Alford, Robert, Crouch, Colin, Offe, Claus (Eds.), Stress and Contradiction in Modern Capitalism. Lexington Books, Lexington, MA, pp. 261–274.
Nachmias, David, 1979. Public Policy Evaluation: Approaches and Methods. St. Martin's Press, New York.
Pierson, Paul, 1993. When effect becomes cause: policy feedback and political change. World Politics 45, 595–628.
Pierson, Paul, 2000. Increasing returns, path dependence, and the study of politics. American Political Science Review 94 (2), 251–267.
Rose, Richard, 1991. What is lesson-drawing? Journal of Public Policy 11 (1), 3–30.
Sabatier, Paul A., 1987. Knowledge, policy-oriented learning, and policy change. Knowledge: Creation, Diffusion, Utilization 8 (4), 649–692.
Sabatier, P.A., 1991. Toward better theories of the policy process. PS: Political Science and Politics 24 (2), 144–156.
Salamon, Lester M. (Ed.), 1989. Beyond Privatization: The Tools of Government Action. Urban Institute, Washington.
Simon, Herbert A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69 (1), 99–118.
Smith, Gilbert, May, David, 1980. The artificial debate between rationalist and incrementalist models of decision-making. Policy and Politics 8 (2), 147–161.
van Waarden, F., 1992. Dimensions and types of policy networks. European Journal of Political Research 21 (1/2), 29–52.

Policy Knowledge: Foundations

Helmut Anheier, Hertie School of Governance GmbH, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Prewitt, volume 17, pp. 11586–11591, © 2001, Elsevier Ltd.

Abstract

This article assesses the nature of foundations and their role in policy and for civil society more generally. It covers the historical roots of foundations and also reflects on new developments in the field. Additionally it contemplates on the foundation's role in international development cooperation as well as within the American university system. Lastly the article discusses debates and criticisms generated around the topic and draws a comparison between the history of foundations in the United States and in Europe.

Foundations have been among the oldest social institutions, prevailing for thousands of years. Their existence in various forms and across many cultures throughout history has remarkably expanded in the immediate past few decades of flourishing international and national institutional philanthropy (Johnson, 2010). By the turn of the century, there were more foundations with more assets in more nations than ever before. Although endowment values dropped in 2001 and 2008 in most developed market economies, they still remain at historically high levels.

In enabling donors to reserve and invest charitable funds, setting terms for their contribution, providing funds to one or many grantees over time, and relocating funds from one charitable activity to another, foundations play an important role. In their civic capacity (comparable to associations and nonprofit organizations), they represent sources of wealth, influence, and initiative, independent of government and business. Foundations can help individuals discharge religious obligations and moral commitments; provide a secure base for religious, cultural, and scientific interests of minorities; and enable ambitious people to exchange economic wealth for social recognition and prestige (Anheier and Hammack, 2010).

Roots of Foundations

There is a lot of discussion about the origins of foundations, which depend on the definition of the term and its geographical focus. Smith and Borgman (2001) identify two prominent roots in the development of foundations in Europe; first religion, namely the Roman Catholic and the Protestant churches, and secondly the 'bourgeois' revolution of the nineteenth century, when the foundation's format was frequently used by the emerging urban middle class and the industrial elite. Early in history, philanthropy was regarded as a charitable attitude which was mainly exercised by religious groups or by royal families, as these institutions were influential in shaping society and making policies. In the introduction to *American Foundations: Roles and Contributions*, Anheier and Hammack (2010) point out how medieval and European philanthropies supported the saying of prayers, the education of religious leaders, the symbolic feeding of the poor, and the religious care of the sick and dying. Later in the eighteenth century, enlightened persons like Webster, Jefferson, and Smith thought of individual charitable giving and assistance as being in the general interest of civil society. In 'The Theory of Moral Sentiments,' Smith referred to benevolence as an universal part of what makes us human (Friedman, 2002: p. 7). Although charity still exists as a fundamental attitude, since the late twentieth century charitable activities have taken a more institutional form for long-term problem solving relating to the new social order. Furthermore, in the course of the twentieth century, scholars from all over the world have come to study nonstate actors and the nonprofit sector. Starting from the beginning of the twenty-first century, a very large body of scholars has focused on the nature and processes of philanthropic activity in the American and European nonprofit sector. Although there are still some limitations with regard to scientific findings, foundations and their role in civil society have become topics for research within the social sciences.

Roles of Foundations and Their Impact on Policy Making

Starting from the mid nineteenth century, foundations have adopted broad and ambitious goals. Health and education (e.g., funding for research, libraries, and educational programs) have been particularly well funded by foundations. Dennis Beatrice (1993) calls foundation's role "one of product development for government on politically charged issues" and Anheier and Hammack (2010) summarize the role of foundations in *American Foundations: Roles and Contributions* as follows:

> Philanthropy describes foundations efforts to create something new in one of the three ways: innovation in social perceptions, values, relationships and ways of doing things, which has long been a role ascribed to foundations, although any effort to innovate can fail and can yield negative as well as positive results, original achievements in the arts, in science and in the study of society and social and policy change whereby foundations foster recognition of new needs, seek to bring a wide range of perspectives to the take, encourage discussion of structural change in the interests of such general social goals as efficiency, equity, peace, social and moral virtues of all kinds, law and order or economic growth.
>
> p. 10

From acting as charitable facilitators and social contributors for a better society, foundations have become major players in policy making. Although they do not have the power to get strong public support, foundations, on the other hand, influence public policy by sharing knowledge. Through their activities, they help to bring issues to public attention and support the assessment of proposed solutions to problems. In many countries, foundations act to take care of problems years before governments notice them (Weissert and Knott, 1995). Although the role of foundations in policy making has been largely ignored within public policy literature, by funding researchers to perform high-quality research in policy areas, they contribute to a great extent to public policy discourses, without taking a specific stance.

Debates around Foundations

Despite the fact that foundations have been successful over many years, they also tend to be subjected to severe criticism. There are various interlinked levels of criticism of the foundation's role in and for society, its internal procedures, as well as around the 'philanthropic' concept as such.

It is argued that foundations stand for the inappropriate use of private funds to influence the public policy sphere and therefore interfere with the democratic process. Furthermore foundations are said to determine and pursue their own priorities with a mix of private and public (in terms of forgone tax through exemptions) funds, without control by the elected government. As institutions without shareholders, voters, or customers, foundations can perhaps be described as the most 'unaccountable' organizations in democratic societies (Anheier and Leat, 2013). Additionally they have been accused of acting in a neocolonialist manner – i.e., wealthier countries aim to dominate poorer states. Contributors to *Charity, Philanthropy and Civility in American History* (Friedman and McGarvie, 2002) point out how during Second World War and the Cold War, philanthropic movements became powerful instruments for governments to advance their visions and targets.

Regarding internal processes, foundations are criticized for the lack of accountability and for the inefficiency of their programs. They are accused of failing to recognize their own limitations and they are blamed for favoritism and paternalism (Anheier and Hammack, 2010).

However it is not only foundations that have been criticized; the intentions of major philanthropists have also been questioned over time. Slavoj Žižek (2006) argues:

> This is what makes a figure like [philanthropist] Soros ethically so problematic. His daily routine is a lie embodied: Half of his working time is devoted to financial speculations and the other half to humanitarian activities… that ultimately fight the effects of his own speculations. Likewise the two faces of Bill Gates: a cruel businessman, destroying or buying out competitors, aiming at virtual monopoly, employing all the dirty tricks to achieve his goals … and the greatest philanthropist in the history of mankind.
> pp. 42–43

Argandoña (1990) further contends in his paper about the ethical arguments of donorship and charity, that donorship always represents a kind of greed, pride, and superiority; this implies that the act of donating increases the ego of the giver and places him on a superior level in relation to the receiver (p. 6).

Nonetheless, one cannot deny that foundations have greatly contributed to social changes, particularly in the fields of education and health. Nowadays, foundations are everywhere in demand, for example, an artist looking for a patron, a local soccer team in financial difficulties, a community school planning to build new facilities or to buy computers, unemployed workers applying for financial help, all look for sponsorships from foundations. Eikenberry and Nickel (2006) argues that philanthropy is dependent upon the existence of marginal groups in need of assistance and a more powerful group in a position to offer this assistance.

While foundations have been publicly criticized, empirical studies have demonstrated that policymakers have pretty much a positive opinion about the work of these institutions. Using an empirical method, Weissert and Knott (1995) conducted a pilot study with a representative sample of policymakers in the United States. Respondents were asked to compare foundations with think tanks and interest groups in respect of their desire to serve the public interest, their compliance with nonpartisanship, and their advocacy for unpopular ideas. In contrast to the recent perception among politicians and academics, which perceived foundations as partisan, nearly three-fourths of the interviewed believed that foundations were nonpartisan compared with the other groups, and additionally, 80% said that foundations were more likely to spend money on novel ideas. Respondents were also asked to evaluate funding approaches and 78% of them accounted funding for protests as extremely useful or very useful for policy making; another 58% indicated the development of informative reports or commissions as especially useful instruments (Weissert and Knott, 1995).

In his book *American Philanthropy*, Robert H. Bremner (1988) argues in favor of philanthropists and claims that whatever the motives that animate the individual philanthropist, the purpose of philanthropy itself is to promote the welfare, happiness, and culture of mankind. Bankers Trust (2000) conducted a study of 112 wealthy households to examine the notion of 'wealth with responsibility.' The idea was to get a measure on how much the wealthy were giving back to society in terms of charitable giving; 90% of respondents stated that they have established an asset plan, which leaves 16% of their assets to charities, 47% to their heirs, and 37% to government in the form of taxation (Boettke and Coyne, 2002). However, the measurement of effectiveness of this charitable giving remains an ongoing debate.

Emergence of a New 'Philanthropy'

All the debates on the purposes, the ethics, and the effectiveness of foundations have also brought up new ideas on how to improve accountability, efficiency, and the managing strategies of these institutions. Therefore the past decade has witnessed the emergence of a new 'philanthropy' characterized by large-scale investment and more result-based mechanisms.

Critiques on traditional philanthropy range from accusing them of faults ranging from being too powerful and immune to public opinion, to being too accessible to organizations seeking to realize social change. Starting from the beginning of the twenty-first century, foundations and philanthropists believed that they could greatly improve the efficiency of their programs by using the same strategies and principles that contributed to their financial success. Since then, foundations have come to change the ways in which they operate and new 'venture philanthropic foundations' have emerged, mainly characterized by long-term and large-scale engagement and strategic similarities to venture capitalists. The term 'venture philanthropy' (VP) was probably first coined in 1969 by the American philanthropist, John D. Rockefeller III, in a hearing before the US Congressional Committee leading to the Tax Reform Act (John, 2006). Scholars are intrigued by this new trend in philanthropic activity and have been researching the major differences between traditional foundations and the new ones.

In her research on VP and charter school reform, Janelle Scott (2009) outlines some basic differences between the traditional form of foundations and VP foundations. She points out that contrary to traditional foundations (which carry out a specific ideology or social policies and share the funding costs between the same grantees), the new VP foundations provide unrestricted general operating support on a multi-year basis at much higher rates than traditional philanthropies. The two groups of foundations also differ in their philosophical approaches to funding. While traditional foundations typically identify broad program areas and invite applications for funding, VP foundations often choose to fund promising leaders and organizations instead. For instance, many of the new VP foundations do not accept grant proposals, for example the Robertson Foundation. It is a family foundation founded in 1996 which identifies the programs that are relevant to its tenets. Therefore its website describes the relating funding philosophy as follows: "The Robertson foundation takes a targeted, businesslike, results-oriented approach that is modeled more closely on private equity investing than on traditional grant making." In her publication, Scott also stated that the most relevant innovation is the utilization of terminologies: for VP foundations: grants have become investments, programs are ventures, and measures of impact generally involve the ability to scale up an initiative (p. 115).

While some scholars praise the new VP foundations in achieving greater results, others criticize and doubt that there is real innovation in their operations. Covington (1997) argues that although a new cadre of foundations has emerged pursuing more aggressive funding programs than traditional foundations, these funding initiatives have been largely already adopted by politically conservative foundations, which have been applying such strategies for some time. Cohen (2007) also doubts that the 'new' philanthropy is truly new. His examination of the new VP foundations revealed striking similarities between the strategies used by these newer philanthropies and those of politically and socially conservative foundations. Therefore, VP foundations seem to have much in common with the traditional foundations in terms of strategy and philosophy.

Though there is not a large amount of published literature on the new VP foundations, there is ongoing criticism about their actions. According to research by Ramdas (2011) VP practices seem to be motivated by technological solutions – the same 'fix-the-problem' mentality that allowed business people to succeed as hedge-fund managers, capital-market investors, or software-developers. This approach is designed to yield measurable and fairly quick solutions but may not always be effective in the world of philanthropy. Therefore, she points out that despite many good intentions, this version of philanthropy seems poorly suited to resolve the world's most pressing and deeply rooted problems.

Foundations and the American University System

As the United States has by far the world's largest foundation sector and offers something close to a 'gold standard' for other countries, it is important to highlight the role of foundations in the United States and especially its place in the American university system (Anheier, 2009; Anheier and Leat, 2013).

In the beginning of the twentieth century the universities in the United States were geographically widespread and mostly private institutions. They were mainly financed by the tuition fees of students and the quality of education was often of a low standard. Local foundations supported colleges with religious affiliation, mostly to strengthen Protestant education – and this gave the colleges a certain independence. However with the emergence of new foundations, among them Carnegie, Rockefeller, and Rosenwald, a new orientation of foundation work started. They strongly supported a system of academic research and professional teaching on a world class level with strategic interventions. Criticism evolved concerning the influence of these 'nouveau riche' philathropists, arguing that they try to control public welfare rather than serve society with an altruistic orientation. As a response to this criticism, foundations started focusing more on national needs.

Furthermore these new foundations started to create innovative systems, not only to support research projects, but also to establish a support system for academics in general as well as the development of selected individual disciplines. Since at the beginning of the twentieth century there were only a few world-class American scholars, the Rockefeller Foundation started to establish centers of excellence to make the United States competitive in international competition. The approach of these foundations consisted in supporting ultramodern laboratories, putting together clusters of excellent graduates, and bringing in highly qualified researchers from Europe to the United States. They also started to build up interuniversity institutions. For instance the Carnegie Institute and the Rockefeller Foundation built the most modern telescopes in years 1908 and 1917 in Texas. With investments of this kind, foundations generated essential support for interdisciplinary institutions. They also provided assets for the American Council of Learned Societies and the Social Science Research Council, through which they helped to build up the National Center for Economic Research. Furthermore they improved the overall economic situation of professors and academic employees.

The priority of most of the foundations was given to natural science rather than to humanities or cultural studies. But

already in these days, Rockefeller, Carnegie, Russell Sage, and others began to think about investment in the emerging field of social sciences. There were several approaches in this regard; for example, the Carnegie Corporation invested strongly in the study 'An American Dilemma,' one of the most important studies about the situation of minorities within the United States. However the problem with social science was its inherent evoking of political controversies. They dealt with topics, which lay at the core of religious and political conflicts. This was the reason why foundations acted in a very cautious manner and placed strong emphasis on accurate elaborated empirical studies. Social studies then had their breakthrough during the Second World War. With the financial support of the state they started engaging in strategic research on the morale of the troops and similar topics. This and other future-leading project funding could be based on (also established with the help of the foundations) the already created fundaments within the social sciences.

As the US system is a mixture of private and public institutions, the state is only expected to be partly held responsible for funding the universities. Since the 80s the Federal Government began shifting 'from grants to loans' and kept the funding of university-based research below the inflation rate. One can also observe a shift from state-funded to state-supported universities. Universities became more dependent on tuition fees, as well as on money from companies and foundations. The financial structure of private and public institutions was assimilated and the competition between the institutions increased. After Second World War, although the Federal Government still played a very important role in terms of funding universities, the strong relationship between foundations and universities still persisted, but evolved in new ways.

Nowadays foundations are hardly responsible for the core budgets of universities, but they invest in specific programs and projects as well as in the establishment and development of institutions. Next to the pattern of competition as the basic policy of foundations in respect of higher education, there are also longlasting cooperations and partnerships. Big parts of the Danforth Foundation, for example, are funding the Washington University of St. Louis, and recent foundations like the Hewlett Packard foundation are funding, to a large extent, the University of Stanford and the University of California. Within these cooperations there are recognizable disciplinary patterns, for example, the Mellon Foundation focuses on humanities, the Welch foundation on chemistry and the Wenner-Gren Foundation on anthropology. Although the donations of foundations for the university system represent a quarter of their assets, for the universities it is only a small part of their total income (below 5%). They receive the biggest parts through tuition fees and public funding, as well as through their property and state donations. These gains considerably increased in comparison to the funding from foundations for universities. Foundations still support innovation in the academic system and provide substantial resources, but they are not any longer able to influence universities to the extent they did at the beginning of the twentieth century. Nowadays they place more emphasis on certain restricted areas with special purposes and increasingly cooperate with other supporters and financiers (Anheier, 2009).

Foundations and International Development

Another topic that has attracted considerable attention from scholars is the involvement of foundations in international development efforts. The first interventions by foundations in the area of international development date back to the 1920s, with groundbreaking action in the public health sector undertaken by the Rockefeller Foundation (Sulla, 2007). Later on, with globalization and the increase in international social contributions for developing countries, foundations have evolved as new actors in the area of funding for international development.

In many developing countries or failed states, the incapacity of the government to meet the needs of the population has led to a growing preference for nongovernmental action as a means to solve societal problems. Foundations have therefore expanded their coverage from a regional level to an international level. As wealthy philanthropists invest billions of dollars on development efforts in Sub-Saharan, Asian, and Latin-American countries, their role in the international development agenda has evoked significant public attention. This kind of development assistance also stands for a world characterized by increasing interconnections beyond national boundaries and problems with global implications. Foundations have been playing an especially important role in fighting worldwide diseases, such as AIDS, Tuberculosis and Malaria. International foundations have also been engaged with the promotion of democratization and support for civil society in areas of limited statehood.

The overall international grants given by foundations from Western countries has doubled over the last decade, and the main recipients of the surplus are developing countries. According to The Foundation Center and Hudson Institute, foundations provided between US$7.0 billion and US $9.5 billion for activity in the sphere of international development. The Bill and Melinda Gates Foundation and Georges Soros' Open Society are concrete examples, both having invested billions of US dollars in Sub-Saharan, Asian, and Latin American countries. There is much speculation on the potential future role of private foundations in the area of development assistance, especially when comparing it with the official development assistance (OAD). Although international development represents only a small share of the overall spending of foundations and although their contribution in contrast to official aid is still small, one cannot deny their potential future impact on international development efforts.

According to Marten and Witte (2008), the Council on Foundations of the United States estimated that there were 68 000 grant-making foundations which gave $3.8 billion for international development in 2005, with the Bill and Melinda Gates Foundation at the top of the list and the Gordon and Betty Moore Foundation in second rank (its assets grew to $5 billion in 2003 and they funded the Andes-Amazon initiative to conserve biodiversity). However, although these numbers might seem to be impressive, only 10% of the grants given by US foundations goes to international development and it is mostly channeled through international institutions like the World Health Organization or through NGOs in Europe, rather than distributed directly to the countries (Sulla, 2007). Research by Sulla (Sulla, 2007) also demonstrates that

the main developing countries directly receiving international assistance from US foundations are the top emerging markets, including Russia, Brazil, India, China, Mexico, and South Africa. The least developed countries seem to be benefiting only slightly from the foundations' direct international flows; and this is mainly due to the difficulty of accessing information on the implementation of the assistance.

In the summer of 2007, the European Foundation Center (EFC), an umbrella organization with more than 200 members, conducted a survey (funded by the German Marshall Fund of the United States) to assess European foundations working in the area of international development. The results highlight the fact that European foundations involved in international development spent around $607 million in 2005; these investments were targeted at 126 countries with only 29 of them belonging to the least developed nations (Marten and Witte, 2008). Nonetheless, in contrast to the USA, Sulla (2007) found that the total international 'grant-giving' of European foundations has been slowly increasing over the last few years, although there is very little information on the sectors of activity to which European foundations have been making international grants.

Philanthropic foundations are said to have many advantages when it comes to international development. It is argued that because foundations can act outside official government channels, they can be more effective than official cooperation, take on more risk, and invest in more innovative projects. They can also reduce transaction costs and respond more quickly to emergencies. They are less subject to misappropriation because they deliver directly to civil society organizations. Additionally, they attract other foundations and potential donors for social development and institutional improvement in areas of limited statehood (Benjamin and Quigley, 2010).

Nonetheless, interventions of foundations in international development have raised criticism. Critics argue that private philanthropy is often supply driven and may not fully align with national development goals and strategies. Benjamin and Quigley (2010) also argue that the downside of foundations as first responders is that they often do not stick around long enough to ensure the long-term sustainability of the institutions that they help to build. Additionally, because their intervention is often off-budget, they might use systems that are parallel to those of governments and thereby complicate governmental financial management; they may also generate high overhead costs for administering small amounts of funds (UN Economic and Social Council Report, 2012).

Other critics fault international foundations with regard to their mode of operation in countries where they invest. Many large foundations working in international development do not have offices in the field and they manage their operations from headquarters in the United States or in Europe; in some other cases, foundations try to stay small although they are faced with the reality of huge budget disbursements that simply require more staff.

Some of the time, the private foundation's work is either led by so-called political philanthropists famous for their business successes or it is supported by celebrities interested in international development. In his book *Fighting Global Poverty: Who'll Be Relevant in 2020?*, Matthew Bishop (2007) argues that this power of personalities coupled with media visibility helps to mobilize and generate public support for development and also fosters interest in developmental issues in the broader public. However, the whole media attention can also exaggerate the work of these philanthropists.

Despite all this criticism, experts in international development tend to support the foundations' work in developing countries. Dr. Watson from the Institute for International Development Studies argues: "Philanthropy can contribute a huge amount to development cooperation. Both private and corporate foundations directed $52.5 billion toward international development in 2009. They can also contribute greatly in many other ways, for example in their convening power, in running pilot programs, in their ability to take risks, and in community engagement, research and advocacy. However, despite the amount of funding, there is slow progress with respect to child nutrition, education, sanitation, urban issues and female employment, and progress is extremely uneven between regions of the world" (http://www.ids.ac.uk/news/philanthropic-foundations-should-be-included-in-discussions-on-development-cooperation (accessed 04.03.14.)).

European and American Foundations: A Comparison

Comparative research on foundations across different countries has only begun in recent years. While research on American foundations started earlier, the study of European foundations is still in its infancy. For instance, research on German philanthropy has only begun in the last 10 years. The impetus for this research stems not from historical interest but from a change in German political culture (Adam, 2001).

While philanthropy has been historically performed by religious communities and royal families in early Europe, researchers have demonstrated that they did not develop a feeling of responsibility and philanthropic impulse. Historians rather have investigated that the state took responsibility for financing social and cultural public institutions. For this reason, philanthropy has been widely seen as an American phenomenon and therefore as a distinct aspect of modern life within the United States.

However in *Philanthropy, Patronage, and Urban Politics; Transatlantic Transfers between Europe and North-America in the Nineteenth and Early Twentieth Centuries*, Thomas Adam (2001) rejected this argument. He states that the travels of American and Canadian philanthropists to Great Britain and Germany in order to find models for museums, art galleries, and social housing projects demonstrate that philanthropy was not an American invention but a European one. Furthermore, by comparing social housing projects in Boston, Leipzig, London, and New York, Adam demonstrated the interconnections among philanthropists from these different countries and showed how American philanthropists implemented European models in US cities.

On the opposite side, Karsten Friedman in the same volume contradicted this view with his investigation of art museums in Germany and the United States after the turn of the century. Focusing on Berlin, he argued that there were fundamental differences between philanthropic support of museums in the United States and in Germany. Whereas American museums were privately owned, German museums were state institutions

that received additional financial support from 'museum associations.'

Toepler (2007) points out in *The Politics of Foundations: A Comparative Analysis* that there are elements that American and European foundations share in common, as well as fundamental differences. Both European and American foundations reject taking over the function of a substituting body, i.e., they are not willing to fulfil functions formerly supplied by the state – for example, redistributing economic resources from higher- to lower-income groups. They both place importance on innovation, which means they engage in promoting innovative enhancements in different fields. Additionally they share a halfhearted acceptance of their pluralistic function as promoters of experiments and diversity in general. However the contextualization of their roles differs. In the United States, the foundation's focus is mainly on social innovation with an undefined relationship to the public sector. The complementary role of serving under-supplied groups is not a big issue in their work. They operate more or less in an institutional vacuum and try to keep their distance from the government and its activities. In contrast European foundations compensate for the institutional shortcomings of the public sector and feel comfortable in their role as complementary actor in the work of the state.

See also: Charity and Philanthropy, Economics of; Charity and Philanthropy: Overview; Voluntary Organizations.

Bibliography

Anheier, H., Leat, D., 2013. Philanthropic foundations: what rationales? Social Research: An International Quarterly 80 (2), 449–472.

Anheier, H., Hammack, D. (Eds.), 2010. American Foundations: Roles and Contributions. Brookings Institution Press, Washington, DC.

Anheier, H., 2009. Stiftungen und Hochschulwesen. In: Schlüter, A., Strohschneider, P. (Eds.), Bildung? 26 Thesen zur Bildung als Herausforderung im 21. Jahrhundert. Berlin Verlag, Berlin, pp. 238–248.

Adam, T., 2001. Philanthropy, Patronage, and Urban Politics; Transatlantic Transfers between Europe and North-America in the Nineteenth and Early Twentieth. Indiana University Press, Bloomington, 2004.

Argandoña, A., July 1990. Sponsorship and Charity: The Ethical Argument. IESE Business School, University of Navarra. Working Paper WP-195.

Beatrice, D.F., Summer 1993. The role of philanthropy in health care reform. Health Affairs, 185–192.

Benjamin, L.M., Quigley, K.F., 2010. For the world's sake: U.S foundations and international grant making 1990–2002. In: Anheier, H., Hammack, D.C. (Eds.), American Foundations: Roles and Contributions. Brookings Institution Press, Washington, DC, pp. 205–261.

Bishop, M., 2007. Fighting Global Poverty: Who'll Be Relevant in 2020. Brookings Blum. Roundtable 2007 in Brookings Institution, Washington, DC.

Boettke, P.J., Coyne, C.J., 2002. The Political Economy of the Philanthropic Enterprise (Unpublished working paper).

Bremner, R., 1988. American Philanthropy, second ed. University of Chicago Press, Chicago, 312 pp.

Cohen, R., 2007. Strategic Grantmaking: Foundations and the School Privatization Movement. National Committee for Responsive Philanthropy, Washington, DC. Subject File Number: 2112007, 2007 (34 pp).

Covington, S., 1997. Moving a Public Agenda: The Strategic Philanthropy of Conservative Foundations. National Committee for Responsive Philanthropy, Washington, DC.

DECPG, International Finance Team, 2006. Philanthropic Foundations: Actual versus Potential Role in International Development Assistance.

Edwards, E., November 2011. The role and limitations of philanthropy. In: The Bellagio Initiative: The Future of Philanthropy and Development in the Pursuit of Human Wellbeing.

Eikenberry, A.M., Nickel, P.M., 2006. Towards a Critical Social Theory of Philanthropy in an Era of Governance (Unpublished manuscript).

Friedman, L.J., 2002. Philanthropy in America: historicism and its discontents. In: Friedman, L.J., McGarvie, M.D. (Eds.), Charity, Philanthropy and Civility in American History. Cambridge University Press, Cambridge, pp. 1–21.

Friedman, L.J., McGarvie, M.D., 2002. In: Charity, Philanthropy and Civility in American History. Cambridge University Press, Cambridge.

John, R., 2006. Venture Philanthropy: The Evolution of High Engagement Philanthropy in Europe. Skol Centre for Social Entrepreneurship, Oxford SAID Business School. Working paper, June 2006.

Johnson, Paula D., November 16, 2010. Global institutional philanthropy: a preliminary status report. Philanthropic Initiative, Inc.; Worldwide Initiatives for Grantmaker Support (WINGS).

Marten, R., Witte, J.M., 2008. Transforming Development: The Role of Philanthropic Foundations in International Development Corporation. Global Public Policy Institute. Research Paper Series No. 10.

Ramdas, K.N., 2011. Philanthrocapitalism: Reflections on Politics and Policy Making. Springer Science.

Renz, L.S., Lawrence, S., Smith, J., 2004. Foundation Funding for the Humanities: An Overview of Current and Historical Trends. The Foundation Center, New York.

Scott, J., 2009. The politics of venture philanthropy in charter school policy and advocacy. Educational Policy 23 (106). Available on: http://epx.sagepub.com/content/23/1/106 (accessed 04.03.14.).

Smith, J.A., Borgman, K., 2001. Foundations in Europe: the historical context. In: Schluter, A., Then, V., Walkenhorst, P. (Eds.), Foundations in Europe, Society, Management and Law. Directory of Social Change, London.

Sulla, O., February 2007. Philanthropic Foundations and Their Role in International Development Assistance. International Finance Briefing Note, New Series-Number 3.

Toepler, S., 2007. Foundation roles and visions in the USA: comparative note. In: Anheier, H.K., Daly, S. (Eds.), The Politics of Foundations: A Comparative Analysis. Routledge Taylor & Francis Group, London and New York, pp. 324–339.

United Nation Social and Economic Council, 2012. UN Economic and Social Council Report.

Weissert, C.S., Knott, J.H., 1995. Foundations' impact on policy making: results from a pilot study. Health Affairs 14 (4), 275–286.

Žižek, Slavoj, April 2006. Nobody has to be vile. London Review of Books 28 (7).

Policy Knowledge: New Liberalism

Mary O Furner, University of California Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

New Liberalism, a contrast term, refers to a movement in social knowledge, public philosophy, state policy, and social practice underway internationally between the 1880s and the 1940s, intended to address the 'social question' raised by instabilities and inequities of modern Western capitalism, and to provide new thinking about relationships between state, economy, and society. New liberals addressed a range of contested issues, among them business cycles and unemployment, rights and conditions of labor, wages and distribution of income and wealth, the limits of property rights, industrial organization and corporate reorganization, family welfare and gender roles, public administration, and issues of immigration and race. Deep divisions over the substance and implications of new empirical and theoretical knowledge on these subjects led to a branching of new liberalism into two distinct new liberal movements elaborated over several decades – a statist-democratic strand of reform thought born in the 1880s, and an opposing, corporate-voluntarist alternative that came together around the turn of the century. The account here identifies five phases in the bifurcated development of new liberalism: (1) market failure and the drive for new social knowledge, (2) conflicting approaches to framing the social question, (3) the 'turn' to social insurance, (4) liberal statism and New Deal planning, and (5) the containment of statist liberalism.

New Liberalism, a contrast term, refers to a movement in social knowledge, public philosophy, state policy, and social practice underway internationally between the 1880s and the 1940s, intended to address the 'social question' raised by instabilities and inequities of modern Western capitalism, and to provide new thinking about relationships between state, economy, society, and the individual. New liberals addressed a range of contested issues, among them business cycles and unemployment, rights and conditions of labor, wages and distribution of income and wealth, industrial organization and corporate reorganization, family welfare and gender roles, immigration and race. Deep divisions over both the substance and implications of new knowledge on these subjects led to a branching of new liberalism into two quite distinct 'new liberal' movements elaborated over several decades. The first to appear was a *statist-democratic* strand of reform thought, born in the 1880s, that proposed more active government in support of economic and social justice. An opposing, *corporate-voluntarist* analysis came together around the turn of the century, aimed at substituting voluntary collective action for much of the state action advocated by those seeking stronger government regulation of work, labor, and business. The term New Liberalism was invented initially by framers in England and the United States to distinguish its early, statist-leaning iteration from, and to set it against, 'old,' classical liberalism, as the latter had developed in the West between the seventeenth and mid-nineteenth centuries around the philosophies of John Locke, Adam Smith, and David Ricardo, and in response to arguments for laissez faire as a default public philosophy. In use internationally between the 1880s and the 1940s, the term referred to a broad-gauged movement intended to address the pernicious instabilities and distributive inequities that modern capitalism had engendered, and to provide new critical thinking about the optimal role of government. From the early twentieth century the terms 'progressive' and 'progressivism' supplanted for some the earlier naming, but both statists and voluntary associationalists could and did claim progressive credentials – a claim that some U.S. historians tend to support. Although the term 'new liberalism' was coined for, and continued to refer to, the move to enhance state power, historians of political ideology and political economy now recognize the binary character of conceptual innovation in this period.

Never a total repudiation of the older classical tradition, new liberalism demanded an exploration of how and where to amend existing commitments to the efficacy of free markets, individual autonomy, secure private property rights, and limited government in light of what had come to be seen as negative consequences of capitalist industrialization. Aiming to frame new methods for shaping and choosing among new civic and social moralities under construction in the industrializing nations of the West, new liberals were ultimately unable to sustain their critical assessment of the classical tradition without changing quite fundamentally their prescriptions for the relationship between society and state.

The account of new liberalism provided here refers mainly to the American case. It identifies five phases in the development of the bifurcated U.S. new liberalism that heightened the visibility of critical moments in its construction. These phases are as follows: (1) a drive for new knowledge prompted by a perception of market failure, (2) formation of conflicting approaches to framing a 'social question,' (3) the 'turn' to social insurance, (4) the rejection of national planning by pro-corporate new liberals, and (5) the containment of liberal statism from the late New Deal through World War II.

Market Failure and the Drive for New Social Knowledge

Chronic business depressions of the late nineteenth century, destabilizing waves of business bankruptcies, mass unemployment, and chronic, desperate poverty alongside ostentatious wealth produced an enormous disappointment of traditional expectations regarding the efficacy of the market.

New liberals recognized recurrent market failure and attendant social conflicts as evidence of a wider collapse, more disruptive to intellectual, social, and political order and more menacing to communities, of a social contract thought ideally to be based on unfettered economic competition between white males that had been exalted in policy discourse during the early nineteenth century. Critics of 'old' liberalism sought a higher rationality that would place knowledge in the service of a more democratically constituted society, which would maintain itself in part by allotting intellectuals and experts a special role in shaping a more vigorous, capable, institutionally developed, morally attuned, and purposeful national government.

Across the range of nations involved, the conceptual revolution that new liberals sought to promote was grounded institutionally in an international social science association movement that fostered transatlantic borrowing, and also in settlement houses, social surveys, and critical journalism, all of them cranked up to record, analyze, and disseminate empirical data, propound theory, encourage the development of political leadership responsive to the new social knowledge and improve the execution of policy. Within government, in investigative agencies such as improved censuses, labor, women's, and children's bureaus, trade boards, regulatory agencies, and bureaus of corporations, high-level civil servants such as Llewellyn Smith and Lord Beveridge in England, and Carroll Wright, Florence Kelley, Julia Lathrop, and Bureau of Corporations chief under Theodore Roosevelt, James R. Garfield, in the United States provided conduits for policy ideas from these burgeoning 'parastates' into government. Crucial early players were founders of the emerging academic social science disciplines, whose missions and boundaries were historically determined, differing from country to country. Academic professionalization and disciplinary specialization proceeded rapidly in the American research universities, fueled by the relative shortage of other types of creditable specialized intelligence.

From the beginning, progressive academics were active in defining what became known in the policy discourse of 1880s as 'the social question.' Strategic rhetorical steps were taken early that shaped the ties between American governance and intellectual life in the period. In the United States, Woodrow Wilson's early and generative 'Science of Administration' (1887) addressed republican fears of corruption by placing a putative firewall between politics and administration. Reflecting distaste for the intense, tribal partisanship of Gilded Age politics, leading academics and bureaucrats (unlike many female social scientists and social reformers, and under some duress) adopted nonpartisanship as a standard description of their professional ethic. This pristine canon elevated everyday standards of disciplinary and agency practice, but pressures for 'objectivity' often hampered the effectiveness of social scientists as advocates for change, ultimately routing important aspects of their activism through ostensibly apolitical venues such as foundations and think-tanks, which often ceded partial control of research agendas to corporate elites. To legitimate their involvement in policy research, academics often sought and accepted roles as 'commissioned' experts, in which they traded a degree of alignment for immunity from outside pressure. The practice of exporting problem solving to private, corporation-sponsored agencies became standard in the Hoover era, a trend reversed in the later New Deal.

Conflicting Approaches to Framing the Social Question

In the struggle to control the direction of change, new liberals of both persuasions deployed rhetorical strategies intended to link their conflicting programs – statist and voluntarist – to carefully selected visions of both the past and the desired future in economic and social policy. Each asserted a connection to the residue of an older individualist doctrine that enshrined private property and competition as the essential guarantees of individual virtue and general welfare, while also explaining how newly experienced social realities had challenged established categories (such as 'the' market) and demanded new forms of collective action. In addition, statist liberals expanded and socialized the republican conception of free labor as the basis of political and economic independence and civic virtue. The elaboration of these two conflicting reform discourses opened up greater possibilities for genuine policy innovation than existed either earlier or since, as in that discrete period, contrary to earlier or later eras, there was nothing close to a hegemonic policy paradigm.

The statist strand proposed translating an aroused social conscience into effective social action through a principled restriction of the meaning of 'private' in many of its applications and through a corresponding expansion of what were to be taken as properly 'public,' morally and socially necessary functions of government. Lamenting the reversal of state-level social regulations in recent court decisions based on newly potent laissez-faire jurisprudence, liberal statists mobilized public agencies in behalf of a reinvigorated *national* interest that they often saw as essential to replace initiatives of individual states and localities traditionally clothed with police powers.

The competing voluntarist vision rejected with equal conviction what, like their predecessor in the old liberal tradition, its founders referred to as 'state interference.' New liberal antistatists preferred to build new ordering strategies upon what they took to be the naturally evolving nongovernmental institutions of a modern society, premier among them the modern business corporation, and upon what they saw as organic, self-activating bodies such as trade associations, professions, charitable organizations, trade unions of strictly limited functions, and peak organizations of private entities, rather than around invigorated legislative democracy, political mobilization, and an empowered, expanding, bureaucratic, regulatory state. They were particularly interested in what could be achieved by way of ordering an excessively competitive and crisis-prone economy through innovative uses of the business contract, working within a powerfully enriched climate of information about the market and – by the turn of the century – through limited banking and tariff reform. Aimed as well at legitimating mergers and restraining coercive job actions by independent labor unions, this effort culminated during the 1920s in the liberal-corporatist design for peak-level business–government cooperation in achieving economic and social stabilization through energized philanthropy and local relief in hard times. Working through proxies, corporate liberals

following Herbert Hoover's lead pressured the worst-acting industrialists to curb voluntarily abuses such as the 12-h day in steel production while offering employee benefits through welfare capitalism.

Liberal statists placed greater confidence in deliberative democracy and less in capitalist rationality or the putative identity of the corporate and the public interest. At the start, drawing energy from heartland republicanism, evangelical Christianity, and German historicism, political economist Henry Carter Adams marked out a path that two subsequent generations of statist-leaning progressives would continue to define as a democratic, moderately collectivist, and communitarian public philosophy. The founding text for the liberal statist view was Adams' classic 1886 article, 'Relation of the State to Industrial Action,' which laid out the costs to society of relying upon the market alone. Adams, along with Richard Ely, Florence Kelley, Edmund James, Simon Patten, John R. Commons, Jane Addams, Mary Van Kleeck, and Charles S. Johnson among numerous others, established a 'context of refutation' for die-hard believers in laissez faire.

For the liberal statists, H.C. Adams pointed to three deficiencies of a regime of market competition that new liberalism would aim to correct: (1) its failure to allow society to enjoy the actual benefits of monopolies in increasing-returns industries without being forced to pay monopoly rents, (2) the inevitable decline of business ethics to the lowest level, absent legal establishment of labor standards, and (3) a general hollowing out of the potential of representative government in a deliberate, misguided weakening of certain key state capacities.

A rising group of new-liberal antistatists led by economists Arthur Hadley and Jeremiah Jenks began forming their own message in the 1890s. Close students of the sources of market failure, these two found great efficiencies in large, capital-intensive industries in transportation and manufacturing; yet they considered even indirect state control of production, wage, and profit levels a threat to the price and investment systems that progressive capitalism could not long endure. With later figures such as Adolph Berle and Hoover aide Edward Eyre Hunt, such thinkers hoped to make the new corporate elites responsible for trans-class and intercorporate cooperation, for providing chances for personal growth and upward mobility along career ladders provided by corporate business, and even for community economic and moral development – to the extent that experience within a corporate culture would inculcate values of efficiency, cooperation, interdependence, and service.

Leading statist liberals can be seen from one perspective as merely seeking to extend the long recognized police powers of the states, used for decades to regulate threats to the public health and safety. Prior to the 1880s, these powers had not been used in newly pressing areas, regulation of labor standards and monopoly. When state legislatures began to move on these issues, state and federal courts initially blocked their actions. As early as 1885 New York State's Court of Appeals, in *In re Jacobs*, struck down a law banning manufacture of cigars in tenement houses, ruling that it violated individual freedom of contract. In the emerging judicial philosophy known as legal formalism, such protective legislation was seen to violate individual due process rights guaranteed by the 14th amendment to the U.S. constitution. In *Ritchie v. People*, 1895, the Illinois Supreme Court applied the same ban against an hours law framed to protect working women. Regarding regulation of monopoly, the decision in *Munn v. Illinois* authorized regulation of rates charged by properties with monopoly positions, such as grain elevators and railroads, whose owners had placed them in the service of the public. But when these constraints on property rights were not extended to regulation of prices and profits in manufacturing monopolies, states turned to an outright ban against monopoly, elevated to the national level in the Sherman Antitrust Act, 1890. This act yielded an unexpected result (also promoted by other factors) when cartel-like combinations involving price fixing agreements between separate businesses were declared illegal, whereas outright mergers that fully absorbed a bunch of smaller manufacturing firms were judged legal. Following the Great Merger Movement, 1898–1902, which gobbled up around 1800 separate businesses, ownership in most core industries was highly concentrated, presenting new policy challenges to both statist liberals and corporate liberals.

As concentration of ownership proceeded in alarming proportions even as economic theorists posited overproduction and competitive price and wage cutting as causes of depressions, labor, populist, and consumer movements challenged new liberals to come up with new ways of addressing the issue. Led politically by Theodore Roosevelt, some liberal statists turned to national level antitrust prosecutions while groping for acceptably bureaucratic means for national regulation. To escape state-level populist pressures, corporate liberals called for transferring regulation of corporations upward, through a system of federal licensing of corporations that established their legality.

From the turn of the century, a dualistic discourse proceeded in which the two opposing strands of new liberalism agreed on departing from laissez faire. Classic texts such as Herbert Croly's *Promise of American Life (1909)* and the Progressive Party Platform of 1912 represented the maturing statist liberal inclination to move aggressively from regulating businesses into national management of natural resources, in which capable, informed governance could instruct a collective interest in conservation, and also into elevating labor standards, extending consumer protection, and establishing a more effective public presence in decisions regarding production, investment, and pricing. Whereas Woodrow Wilson differed on the degree of competition the public interest demanded, he was equally drawn to regulation through state building in areas of clear national-level authority. John Dewey's seminal *The Public and Its Problems* (1927), another powerful exemplar in the new liberal 'role of the state' genre, located the public interest at the place in social relations where effects of contractual relations spilled over onto third parties, who bore the often devastating ancillary consequences of others' private bargains. In economic theory, a corollary to this view was detected in a recognition of the problem of 'social costs' passed to third parties – workers' families and communities – when wages were too low.

Corporate liberals, meanwhile, also imagined institutional mechanisms for achieving economic and social stabilization without extensive government control. Their key innovations involved extensive support for production of policy knowledge, such as work by Wesley Mitchell's National Bureau of Economic Research on business cycles. Herbert Hoover, as

Commerce Secretary and President, backed off on antitrust prosecutions while promoting voluntary trade associations for promoting business cooperation. Rather than looking to assist workers through redistribution of income, the voluntarist, associative strategy here was to raise living standards by growing the economic pie.

Against this voluntarist current, statist progressives continued to look to independent unions protected by the state and to democratic politics as the means for achieving an 'economic republic' and an 'industrial democracy.' Their vision included political recognition of three inalienable rights of worker citizens – a 'living' wage sufficient to support a family in reasonable comfort, the right to organize collectively, and a labor voice in management – as well for some as a guarantee of employment. This earlier thrust is not always visible in histories of the U.S. welfare state.

The 'Turn' to Social Insurance

Underway in Europe much earlier, the new-liberal 'turn' to social insurance came in earnest in the U.S. case only in the second and third decades of the twentieth century. Rising in social work, social feminist, settlement, and bureaucratic circles, it signaled the onset of a second phase in new liberal thinking, running parallel, but in important respects contrary, to the older movement for achieving industrial democracy by empowering workers and regulating labor standards, thereby redistributing income. Several factors can be seen to have promoted adoption of a social insurance perspective, among them the arrival among educated elites of an environmentalist consensus on the causes of poverty and accumulating realism about distinctive constraints on labor's capacity to organize an independent political program within the U.S. version of non-ideological, coalition-based, patronage-dispensing political parties. As well, though increasing social tensions compelled American cities to provide some public relief during cyclical downturns from the 1870s forward, the expectation among many in the Protestant middle class that private charity should aid only the 'truly needy' remained strong.

In the rise of a compensatory approach through social provision, Atlantic-crossing academics with expert knowledge played a major role. Bismarckian introduction of several worker-appeasing social insurance programs through Great Britain's adoption in 1911 of broadly accessible unemployment insurance were faithfully cataloged in works such as Isaac Rubinow's *Social Insurance* (1913). As class conflict accelerated, inquiries staffed by social scientists, among them the famous Pittsburgh Survey (1907–08) and the United States Commission on Industrial Relations (USCIR, 1913–15), documented corporate abuses, the powerlessness of working class communities to combat them, and the toll exacted upon workers' families. To some thinkers these findings suggested a renewed drive for aggressive labor standards reform and protection for unions. Yet, contrary to the strong government role envisaged by the USCIR and in wartime programs for improving labor standards, in the post-World War I period, wartime wage codes and protections for collective bargaining were quickly dismantled, as government at all levels tolerated – even encouraged – an open shop movement based on individualist premises.

With Herbert Hoover in a prominent organizing role, the Wilson and Harding administrations arranged a series of high-level conferences for planning postwar stabilization concentrating initially on efforts to comprehend and moderate the business cycle. Allied with the President's Conference on Unemployment, Wesley Mitchell's National Bureau of Economic Research offered vital statistical work that suggested new policy options, among them enrichment of the informational climate for business decisions, accumulation of reserve funds for countercyclical public works at the local level, an actuarial approach to relieving unemployment through funds amassed by employers, unions, or government, and acceptance of state-level workers' compensation schemes, partly to avoid stricter employer liability enforcement. In what was clearly one of the most significant conceptual shifts in modern governance, these more conservative approaches transformed unemployment from a problem of the political order rooted in inequality into an insurable risk to be managed technocratically in the firm or industry.

But the impulse for what from the perspective of liberal-statist, laborist, and social-feminist proposals for extensive redistribution of income and political voice became a critical *retrenchment* in social policy thinking arose not only from these liberal corporatist initiatives, but from at least three other, at first glance unlikely sources as well: a maternalist turn in some of the women's movements of the long progressive era; negative visions of the immigrant poor propagated by the Chicago school of sociology most particularly; and the aggressive expansion of carefully targeted forms of social provision by the private sector, imagined as a strategy for making the 'modern manors' of welfare capitalism reliable instruments for securing the loyalty and uninterrupted productive service of industrial employees.

As politically conservative opposition to protective legislation extending the police powers accumulated and the suffrage crusade shifted to political mobilization, important segments of mainstream, middle-class feminism took on a more elitist, exclusionist cast, and a 'female dominion' of women's protective institutions increasingly followed a separate course in identifying reform goals. Leading female reformers built a case for special state-based income assistance to virtuous mothers and children and special provisions for safeguarding maternal and infant health. Thus maternalists helped to shape an American pattern that women's historians have identified as a 'two-channel welfare state.' By the 1930s one channel provided contribution-based coverage for male workers under comparatively generous income-replacing programs such as unemployment insurance and Social Security, whereas the other offered only meager, means-tested, minimally tax-funded 'welfare' for stay-at-home mothers and children, but little assistance of any kind for most female workers and people of color.

Against this trend toward the retreat to social insurance, two bastions of investigative activism in the national government continued to lobby for equal treatment for those workers. The wartime Women in Industry Service led by leftist Mary Van Kleeck, which became the U.S. Women's Bureau, sought gender equity on the job in war and peace. The short-lived Division of

Negro Economics (DNE) headed by National Urban League leader George Haynes aided the equitable integration of Southern Blacks into northern war work. Striving to 'integrate' the labor question, another group of mostly male social scientists and social investigators, Commons student Paul S. Taylor and New Deal artist Dorothea Lange among them, exposed equally pernicious myths about the needs and abilities of Mexican, Mexican American, and 'poor white' workers. Congress defunded DNE soon after the war and migrant workers, Hispanic as well as African American, remained (remain) vulnerable to endemic discrimination and periodic spells of outright persecution. In all these developments, social and economic investigation served a dual purpose: grounding public discussion in reliable information rather than what publicist Walter Lippmann termed 'pictures inside men's heads' and providing a basis for strategic and relatively safe advocacy.

Against the collectivist shadings of both statist and corporatist ideas, neoclassical economists cultivated statistical and mathematical models of analysis between the 1890s and the mid-twentieth century that extracted economic behavior from history and culture, (re)naturalized the workings of the market, and turned the discipline into a potent form of social engineering. Quantitative, behavioristic, and functionalist models vied with historicism and institutionalism in other social sciences as well, perhaps most visibly in two massive collaborative efforts in Hoover-era policy knowledge, *Recent Economic Changes* (1929) and *Recent Social Trends* (1933). The dominant sociological discourse of the 1920s, the 'Chicago School' established by Robert Park, W.I. Thomas, and their students at the University of Chicago, offered an alternative to the liberal-statist explanation for poverty and urban ills that took the form of a cultural theory of family and community 'disorganization.' Culminating with Gunnar Myrdal's *American Dilemma* (1944), this new cultural analysis of poverty separated the poor from the capitalist strategies that shaped their lives and rendered those historically constituted structures less visible. Noting the persistence of such tendencies helps to account for the distinctive pattern of U. S. state formation, and particularly welfare-state development, that students of social policy, often taking the more comprehensive European welfare states as the 'normal' path of history, have seen as tardy, insufficient, and incomplete.

When wartime controls on the economy were quickly abandoned, corporatism flourished in the United States during the Hoover era (as it did in different forms during the interwar period in Europe as well). The sociology of ethnic difference merged with developments in the new sciences of management to promote the refinement of corporation-based programs for linking carefully constructed forms of social provision to employment with specific employers, which served as well as a guarantee against worker turnover. As arts of advertising based on a new consumer psychology legitimized the desire for attractive and plentiful goods, theories of an administered market allowed liberal corporatists to proffer a new economy of mass consumption as a palatable alternative to nineteenth-century producerist dreams of worker control and to aspirations for distributive equity envisaged in the ideal of industrial democracy.

Resisting this trend only formally as they accommodated skilled worker unionism to the realities of class power, the American Federation of Labor assented to the views of industrial relations experts who were constructing a new profession in the 1920s in recognizing productivity increases, and not older republican-based claims for more equitable distribution to producers, as the sole and indispensable basis for increases in wages. Where company welfare programs existed in the 1920s, worker loyalties based on them ran quite deep, particularly in ethnic communities often played off against each other on the job. Thus the sudden collapse of welfare capitalist programs, the weak Hooverian response to the Great Depression crisis, the welcome proliferation of New Deal programs targeting unemployment, and the aggressive organizing efforts finally unleashed by national labor-rights legislation helped to ratchet up workers' expectations for the role of government. Pulling in long-time statist-leaning new liberals and their ideas, and despite early New Deal liberal corporatist tendencies, Franklin Roosevelt's turn to a class-edged defense of entitlement to jobs and to expanded social insurance programs reinvigorated impulses, largely quiescent since the immediate postwar World War I era, toward the development within government of planning capacities capable of heading off such emergencies in the future, and of democratizing both industrial governance and the political state.

The Limits of Liberal Statism

The broader vision of a deliberative process that could detect and shape the impact of American capitalism on society was the unifying insight at work in the planning agencies of the New Deal, most evident in the activities and discourses of the Public Works Administration, the National Resources Planning Board (NRPB), the Committee on Economic Security (CES) that devised alternative blueprints for Social Security, and the Temporary National Economic Committee (TNEC) that investigated how far monopoly had presaged or extended the Great Depression.

The so-called 'planning impulse' had a long history. In the 1910s as a conservationists' discourse of efficient natural resource management and in the 1920s often as a welfare corporatists' discourse it had drawn comparatively little fire. Stressing continuities in the development of a national planning capacity that bridged private and public, even the early New Deal National Recovery and Agricultural Adjustment Acts could seem like little more than continuations of corporatist strategies for controlling the volume of production through a beneficial collusion among competing producers to sustain prices, now under the cover of state compulsion.

Yet a parallel, consciously statist movement also appeared in the early 1930s, fueled by FDR's determination to create both government and private sector jobs to reduce catastrophic and politically unacceptable unemployment not amenable to associative solutions. Voices farther to the left demanded full utilization of idle resources and full employment of labor, strengthening the hands of New Dealers, brought to Washington in the hundreds, who were determined not to waste a crisis. Recruitment to government planning agencies of eminent social scientists, including a large number of historical

and institutional economists, energized and updated the liberal statist strand of new liberal thinking. By the mid-1930s policy innovations shifted the planning discourse toward using resources more efficiently over the longer term, expanding public infrastructure investment, rekindling demand, and supplementary job creation – policy goals institutionalists insisted would promote economic stability and distributive equity. The National Labor Relations Board in 1935 framed its mission in these terms, contending that ensuring workers' rights to organize independently equipped trade unions to play a role in redistributing income, thus supporting the workers' ability to consume while democratizing industry.

Economists' knowledge and planners' leanings toward infusing market structures with civic purposes also blended in agencies such as the venerable Bureau of Agricultural Economics and the community-building oriented though ultimately disappointing Tennessee Valley Authority. The Farm Security Administration addressed the issue of chronic farm surpluses, but also the economic and social marginality of people at the bottom of rural hierarchies who were unable, without the kinds of government assistance long ago advocated by 1890s Populists, either to synchronize production with world-market demand or to retain enough income from their crops to live decently. For the manufacturing sector, a major study of the concentration of economic power steered by Thurman Arnold and Leon Henderson for the TNEC referred the failure of prices to decline along with wages and the slow pace of investment in new industries back to the monopoly power that allowed large corporations to administer markets. An analysis of the 'stickiness of prices' by Gardiner Means of the NRPB on the basis of extensive study of investment and pricing patterns pointed in the same direction.

The meaning of planning was shifted most decisively by theoretical, institutional, and practical innovations generated by the NRPB, as Roosevelt enlarged that New Deal agency's role to guiding government efforts to achieve recovery and later to maintaining prosperity at war's end. A body of economic knowledge cobbled together in the early 1930s by economists such as Edward Chamberlain, Joan Robinson, Alvin Hansen, John Maynard Keynes, and Michael Kalecki explained how equilibrium well below the level of full employment of labor and capital had been reached and sustained over nearly a decade. Building upon this analysis, economists at NRPB and the Federal Reserve zeroed in on underconsumption as a factor in the slow, uneven pace of recovery, discovered the 'multiplier effect' of government spending, and insisted that gains could be achieved by more aggressive deficit financing of public expenditures in the form of worthwhile civil works.

By the definition of planning that flowed from this new thinking, the state became minimally responsible for continuously monitoring and occasionally augmenting or shifting investment flows in socially constructive directions not attractive to the private capital markets and for supporting incomes to support consumption. Inside the CES, one faction of planners insisted on continuing provision for government-subsidized full employment, but the victorious vision tilted more toward social insurance through an expanding welfare state. Among advanced elaborations of New Deal liberal statism, a report titled *Security, Work, and Relief Policies* (1942) conveyed NRPB's synthesis of three related but discrete government roles: financing and extending social insurance to care for those unable to participate in the work economy, guaranteeing work for the segment of the population securely attached to the labor market, and underwriting a still highly contested version of industrial democracy through defending the self-organization of labor. Through its promotion of this triad, the NRPB enunciated a moderately statist vision of a new social contract – one that, notably, did not include a civil rights dimension, reflecting not only FDR's and the New Deal's unwillingness to risk losing Southern Democratic votes but, as well, a more general silence on the race question among new liberals going back to the 1880s.

Though more credit has been given New Deal planners in light of post-1970s attempts to roll back New Deal reform, they managed only a partial and temporary fulfillment of the moralizing role that early liberal statists had charted for knowledge-based civic discourse. The Fair Labor Standards Act of 1938, which established a 40-h work week with time-and-a-half for overtime work, set a national minimum wage, and banned most child labor, was perhaps the apogee of labor standards reform; yet it exempted many members of the working class – most women, farm workers, domestic and managerial employees – from its protections. The Fair Employment Practices Commission also largely failed to address racial and gender discrimination.

Even so, opposed by antistatists, the NRPB's interest in postwar planning was one of the factors that led to the uncer-emonious termination of the agency by Congress in 1943. The silencing of NRPB was a milestone in the history of new liberalism and of the social sciences as well. The commitment of New Deal planners to full employment proved untenable in the face of rising opposition to the New Deal. Thus the Council of Economic Advisers created by the watered-down Employment Act of 1946 lacked the authority to trigger government spending at a level needed to achieve full employment, though the agency continued at least through the Truman era to press for a key government role in the postwar expansion of the economy, to jawbone against monopoly as an obstacle to growth, and to press for fuller socialization of the insecurities of modern living.

The Containment of Liberal Statism

Whereas statist elements of New Deal reform met with powerful opposition by the late 1930s, the rudimentary modernization theory embodied in Hoover's associative state was further elaborated during the Cold War, as U.S. manufacturing and credit extension underpinned the non-Soviet world economy. The need for uninterrupted production to reap the profits from rising demand formed the basis for a kind of truce between capital and strong industrial unions based on supervised collective bargaining, negotiated wage adjustments to meet inflation, a union crackdown on wildcat job actions, and elite theories that cast this system as the start of a 'post-capitalist' era of widespread abundance. Although President Dwight Eisenhower cautioned against the dangers of a 'military–industrial complex,' U.S. foreign policy dictated massive defense production alongside consumer goods manufacture, leading to low unemployment rates and good

wages. During the immediate post-World War II era, a mildly redistributive welfare state expansion including the GI Bill, social security extension, and minimum wage protection expressed an ongoing progressive commitment to some measure of social and economic democracy – though for whites only. Failed proposals for universal health care and full employment traced its limits.

The architects of U.S. corporate liberalism never intended it to become a European-style corporatism that involved mandatory codetermination of industrial and labor policy; in the diluted native form, it provided an alternative vision to liberal statism, supplying a degree of constraint on both income-supporting policies and union participation in economic planning that was present in anything like comparable strength nowhere else. More than other distinctive U.S. features, the fuller elaboration of the associative vision even during the deceptively radical 1960s goes far to explain both the special character of modern United States liberalism and the persistence of powerful linkages between ruling business elites and pro-corporate antistatists in political roles.

In light of recent scholarship, we now know much more about the mobilization against the New Deal, including how vehemently anti-New Dealers worked to defeat government planning for economic performance. Opponents of planning clustered around nodes represented by the resurgent free market liberalism of Friedrich von Hayek, who forecast a long slide down the slippery slope to communism for welfare states, by the romanticized heroics of libertarian novelist Ayn Rand's fictional entrepreneurs, and by concerns among business conservatives and Southern industrialists that federal government 'interference' might invade local prerogatives regarding labor organization and race. Aiding their project, a 'fiscal revolution' in economic policy thinking, underway by 1937 and dominant by the 1960s, offered modified, 'commercial Keynesian' policy knowledge suggesting that the measured use of fiscal and monetary levers would enable macroeconomic stabilization without requiring intrusion into capitalist investment practices or extensive interference with labor markets. Those who took this view in the early postwar decades typically saw some benefit in the collective bargaining structures mandated by the National Labor Relations Board (NLRB), whereas they preferred that all decisions regarding production be left to private ordering in the context of a mildly redistributive compensatory state.

Even more important foes of liberal statism were organizations where conservative theorists and high-level business executives engaged in counter-planning against all that the New Deal had inaugurated for the role of government. These efforts began quite early as business conservatives organized in the Liberty League, the National Association of Manufacturers, and the Chamber of Commerce to promote a pro-business agenda against labor unions and Social Security, and to finance economists such as Hayek, who organized the laissez-faire oriented Mount Pelerin Society in 1947. Against his own preference, this organization abandoned reservations about the moral exactions of modern capitalism that remained urgent for conservatives such as Hayek and early Chicago-school economist Frank H. Knight. The economists' laissez-faire movement, linked with anti-communism through the John Birch Society and informed by center-right economic theory through the Brookings Institution and the American Enterprise Institute, found a champion in staunchly anti-New Deal and pro-states rights department store magnate Barry Goldwater in his Phoenix-centered mobilization of business conservatives. Although his 1960 election campaign demonstrated the perils of attacking Social Security, it also set the stage for pro-market conservatives to turn the Republican Party into a vehicle for anti-statist ideology.

To sum up, whereas the first generation of American liberal-statist social scientists had reconstituted government as a knowledgeable, capable, morally responsible regulator and redistributor, and the second generation had modeled the state as social insurer and mediator of class antagonisms, the third, late New Deal generation of statist liberals conceived a state role that captured the planning role that originated in the conservation movement, the social provision role embodied in the Beveridge Plan, and the codetermination methods that took hold in postwar European social capitalism. It was strongly nationalist, and thus reminiscent of older Hamiltonian and Whig models of a development state responsible for energizing and supporting through provision of infrastructure voluntary community efforts leading to balanced economic, moral, and civic development of the commonwealth. Policy discourses of the 1950s, developed around Cold War internal security issues, an increasingly consensual commercial Keynesianism, modernization theories exported from the United States as technologies for liberal internationalism, dreams of 'post-modernity' in social relations, and domestic discourses of racial and gender identity, challenged the normative postures that liberal statists and left progressives had formulated between the 1880s and the Fair Deal and inaugurated a new ideological era. The legacy remains, embodied in resilient structures of disciplinary knowledge formation, though once durable discursive norms that demanded grounding political argument in social empiricism have been dramatically compromised in the post-1970s era of 'take no prisoners' partisanship that moved the U.S. political spectrum dramatically to the right. Still, absent the polarizing tendencies of a world divided between capitalism and state communism, and even given now the resurgence over recent decades of libertarianism and free market ideology, it becomes ever more obvious that the critical decisions in democratic governance in the West, and beyond, will henceforth be – as they were in the new liberal era, 1880s to 1940s – hard choices between variants of modern liberalism.

See also: Capitalism; Class: History of the Concept; Corporate Culture; Economics and Ethics; Economics, History of; Ideologies, Institutions, and the New Institutionalism; Labor Market Policy; Liberalism, Anthropology of; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Political Economy, History of; Regulatory Agencies; Social Justice and Social Inequalities; Social Question: Impact on Social Thought; State, History of; Think Tanks; Working Conditions.

Bibliography

Alchon, G., 1985. The Invisible Hand of Planning: Capitalism, Social Science, and the State in the 1920s. Princeton University Press, Princeton, NJ.

Balogh, B., 2009. A Government Out of Sight: The Mystery of National Authority in Nineteenth-Century America. Cambridge University Press, New York.

Barber, W., 1996. Design within Disorder: Franklin D. Roosevelt, the Economists, and the Shaping of American Economic Policy, 1933–1945. Cambridge University Press, New York.

Berkowitz, E., McQuaid, K., 1992. Creating the Welfare State: The Political Economy of Twentieth-Century Reform, rev. ed. University Press of Kansas, Lawrence, KN.

Brinkley, A., 1995. The End of Reform: New Deal in Recession and War. A. A. Knopf, New York.

Brick, H., 2006. Transcending Capitalism: Visions of a New Society in Modern American Thought. Cornell University Press, Ithaca.

Burgin, A., 2013. The Great Persuasion: Reinventing Free Markets since the Depression. Harvard University Press, 2012, Cambridge.

Burns, J., 2009. Goddess of the Market: Ayn Rand and the American Right. Oxford University Press, New York.

Clawson, M., 1981. New Deal Planning: The National Resources Planning Board. Johns Hopkins University Press, Baltimore, MD.

Collini, S., 1979. Liberalism and Sociology: L.T. Hobhouse and Political Argument in England, 1880–1914. Cambridge University Press, New York.

Eisenach, E., 1994. The Lost Promise of Progressivism. University Press of Kansas, Lawrence, KN.

Farr, J., 1989. Understanding conceptual change politically. In: Ball, T., Farr, J., Hanson, R. (Eds.), Political Innovation and Conceptual Change. Cambridge University Press, New York, pp. 24–49.

Fraser, S., Gerstle, G., 1989. The Rise and Fall of the New Deal Order, 1930–1980. Princeton University Press, Princeton, NJ.

Freeden, M., 1978. The New Liberalism: An Ideology of Social Reform. Clarendon Press, Oxford, UK.

Furner, M.O., 1975. Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905. The University Press of Kentucky, Lexington, KY.

Furner, M.O., Supple, B. (Eds.), 1990. The State and Economic Knowledge: The American and British Experiences. Cambridge University Press, Cambridge, UK.

Furner, M.O., 2009. From 'state interference' to the 'return of the market': the rhetoric of economic regulation from the old gilded age to the new. In: Balleisen, Edward, Moss, David (Eds.), Government and Markets: Toward a New Theory of Regulation. Cambridge University Press, New York.

Gerstle, G., 2001. American Crucible: Race and Nation in the Twentieth Century. Princeton University Press, Princeton, NJ.

Hawley, E., 1966. The New Deal and the Problem of Monopoly. Princeton University Press, Princeton, NJ.

Heclo, H., 1974. Modern Social Politics in Britain and Sweden. Yale University Press, New Haven, CT.

Hendrickson, M., 2013. American Labor and Economic Citizenship: New Capitalism from World War I to the Great Depression. Cambridge University Press, New York.

Hovenkamp, H., 1991. Enterprise and American Law, 1836–1937. Harvard University Press, Cambridge.

Horwitz, M., 1977. The Transformation of American Law, 1780–1860. Harvard University Press, Cambridge.

Horwitz, M., 1992. The Transformation of American Law, 1870–1960. Oxford University Press, New York.

Jackson, W., 1990. Gunnar Myrdal and America's Conscience: Social Engineering and Racial Liberalism, 1938–1987. University of North Carolina Press, Chapel Hill, NC.

Jacoby, S., 1997. Modern Manors: Welfare Capitalism since the New Deal. Princeton University Press, Princeton, NJ.

Joas, H., 1993. Pragmatism and Social Theory. University of Chicago Press, Chicago, IL.

Katznelson, I., 2013. Fear Itself: The New Deal and the Origins of Our Time. W.W. Norton & Company, New York.

Kessler-Harris, A., 2001. In: Pursuit of Equity: Women, Men, and the Quest for Economic Justice in 20th-Century America. Oxford University Press, New York.

Klein, J., 2003. For All These Rights: Business, Labor, and the Shaping of America's Public–Private Welfare State. Princeton University Press, Princeton.

Kloppenberg, J., 1986. Uncertain Victory: Social Democracy and Progressivism in European and American Thought, 1870–1920. Oxford University Press, New York.

Koven, S., Michel, S., 1990. Womanly duties: maternalist politics and the origins of welfare states in France, Germany, Great Britain, and the United States, 1880–1920. American Historical Review 95, 1076–1108.

Lacey, M.J., Furner, M.O., 1993. The State and Social Investigation in Britain and the United States. Cambridge University Press, Cambridge, UK.

Lacey, M. (Ed.), 1989. The Truman Presidency. Cambridge University Press, Cambridge, UK.

Leach, W., 1993. Land of Desire: Merchants, Power, and the Rise of a New American Culture. Pantheon Books, New York.

Novak, W., 1996. The People's Welfare: Law and Regulation in Nineteenth Century America. University of North Carolina Press, Chapel Hill.

O'Connor, A., 2001. Poverty Knowledge: Social Science, Social Policy, and the Poor in Twentieth Century US History. Princeton University Press, Princeton, NJ.

Patterson, J.T., 2000. America's Struggle against Poverty in the Twentieth Century. Harvard University Press, Cambridge, MA.

Persons, S., 1987. Ethnic Studies at Chicago, 1900–45. University of Illinois Press, Urbana, IL.

Phillips-Fein, K., 2009. Invisible Hands: The Making of the Conservative Movement from the New Deal to Reagan. W. W. Norton, New York.

Reagan, P., 1999. Designing a New America: The Origins of New Deal Planning, 1890–1943. University of Massachusetts Press, Amherst, MA.

Rodgers, D., 1998. Atlantic Crossings: Social Politics in a Progressive Age. Belknap Press of Harvard University Press, Cambridge, MA.

Ross, D., 1991. Origins of American Social Science. Cambridge University Press, Cambridge, UK.

Rossinow, D., 2008. Visions of Progress: The Left-Liberal Tradition in America. University of Pennsylvania Press, Philadelphia.

Rueschmeyer, D., Skocpol, T. (Eds.), 1996. States, Social Knowledge, and the Origins of Modern Social Policies. Princeton University Press, Princeton, NJ.

Silverberg, H. (Ed.), 1998. Gender and American Social Science: The Formative Years. Princeton University Press, Princeton, NJ.

Sklansky, J., 2002. The Soul's Economy: Market Society and Selfhood in American Thought, 1820–1920. University of North Carolina Press, Chapel Hill.

Sklar, K.K., 1993. The historical foundations of women's power in the creation of the American welfare state, 1830–1930. In: Koven, S., Michel, S. (Eds.), Mothers of a New World: Maternalist Politics and the Origins of Welfare States. Routledge, New York, pp. 43–93.

Sklar, M., 1988. The Corporate Reconstruction of American Capitalism, 1890–1916: The Market, Law, and Politics. Cambridge University Press, New York.

Skocpol, T., 1992. Protecting Soldiers and Mothers: The Political and Social Origins of Social Policy in the United States. Belknap Press of Harvard University Press, Cambridge, MA.

Smith, J.S., 2006. Building New Deal Liberalism: The Political Economy of Public Works, 1933–1956. Cambridge University Press, New York.

Stein, H., 1969. The Fiscal Revolution in America. University of Chicago Press, Chicago, IL.

Tomlins, C., 1985. The State and the Unions: Labor Relations, Law, and the Organized Labor Movement in America, 1880–1960. Cambridge University Press, Cambridge, UK.

Policy Knowledge: Universities

Michael Hoelscher, Universität Heidelberg, Heidelberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Court, volume 17, pp. 11605–11608, © 2001, Elsevier Ltd.

Abstract

In the knowledge society, political decisions increasingly have to be based on scientific results to be seen as legitimate. Governments increasingly hold universities accountable for providing such applicable knowledge. The article analyses (1) the current context, (2) the setup of this policy–research nexus, as well as (3) its epistemic impact on legitimate knowledge within science, and (4) the organizational consequences for universities and the social and behavioral sciences. One conclusion is that an increased openness toward policy-relevant issues has to be combined with research's independence from external influences to secure one of its most important contributions, namely a legitimate critical reflection on policy and society.

Introduction

One of the few unambiguously perceived trends of modern societies is the growing importance of knowledge, often described as the emergence of a knowledge society or knowledge economy. This process of scientification is defined by Weingart as "a process whereby the use of and claim to systematic and certified knowledge produced in the spirit of 'truth-seeking' science becomes the chief legitimating source for activity in virtually all other functional subsystems" (Weingart, 1997: 610). One of these subsystems is the political realm. Political decisions are, and have to be to gain legitimacy, increasingly based on scientific knowledge. However, this also has important repercussions for academia as the most important producer of this kind of knowledge (e.g., Boaz et al., 2008). While universities always generated useful and applicable knowledge as well as graduates well prepared for jobs in government, the administration, and the labor market more generally, current developments put much more pressure on higher education institutions to produce 'value for money.' Changes in the mode of knowledge distribution in the past decades, characterized by a smooth shift from supply- to demand-led knowledge production, resulted in a substitution, or at least complementation, of universities' traditionally claimed autonomy by a growing accountability toward the needs of government, society, and economy.

The following paragraphs will analyze the research–policy linkages from two perspectives and on two levels. First, looking outward, it will ask what kind of knowledge is needed in the policy realm and how it is transferred from academia into the policy realm. Looking inward, it will describe some implications of an increasingly demand-driven knowledge regime for the social and behavioral sciences.

At least as important as the growing dependence of political decisions on scientific knowledge is, second, a resulting growing demand for scientific literacy of citizens. Therefore, the topic of 'policy knowledge: universities' not only is understood as knowledge that is produced within universities and then used by governments to formulate policies but also incorporates the university-taught knowledge needed by responsible citizens to critically evaluate these policies. Taking this viewpoint results in two different levels of analysis within universities and their generation of policy-relevant knowledge: research and teaching (see **Figure 1**).

The structure of the article is as follows: In a first step, the relevant political and social context for today's universities is outlined. The next two sections discuss the implications of these trends for appropriate forms of knowledge and for the organizational setup of universities. The Section The Transfer of Knowledge from Universities into the Policy Realm discusses how knowledge is transferred between universities and policy. Then the Section Implications for the Social and Behavioral Sciences outlines some potential implications for the social and behavioral sciences. Finally the article ends with some conclusions and an outlook.

Before describing the current context, though, one caveat has to be mentioned. The situation is very heterogeneous across the world. However, this article is concentrating on the (though still very heterogeneous) situation in the developed countries, mainly from the OECD.

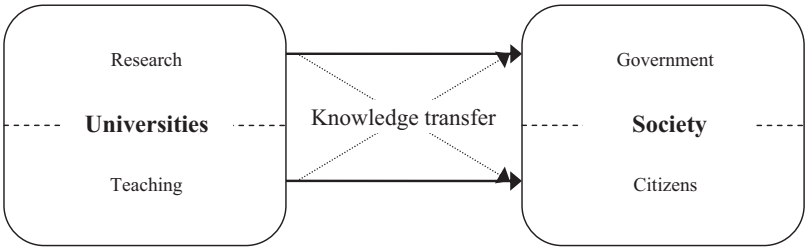

Figure 1 A model of knowledge transfer.

The Political and Social Context for Universities in the Twenty-First Century

Although universities as one of the oldest still existing institutions worldwide have always followed their own logic to some extent, they also have been influenced by their social, political, and economic contexts through the centuries. This paragraph outlines some transnational trends that are of special importance to the relationship between policy and universities, while acknowledging that national and regional developments are also relevant. Consequences of these trends will be discussed in the following sections of this article.

One of the most important developments is the worldwide expansion of participation in higher education during the past decades (Schofer and Meyer, 2005), turning universities from elite- to mass institutions (Trow, 1972) in many countries. This results in a need for additional funding and more generally in an increased importance of university's role in society.

Another important trend is the emergence of what is often called the 'knowledge society' or 'knowledge economy' (Delanty, 2001). Interestingly, this trend was also made popular by Bell's "The coming of post-industrial society" (1973) at the same time as Trow published his observations on expansion in higher education. Within the economy, knowledge has become ever more important, and at least in the developed countries it is sometimes seen as the dominant means of production, as increasingly industrial production sites are moved to countries with cheaper labor costs. Within modern societies, the process of constant 'rationalization,' as already described by Max Weber, calls for ever more information as the basis for decision-making on different levels (national, organizational, individual). Especially the political realm calls for information to legitimate and justify its decisions, so that it is able to deliver what is often called evidence-based policy-making. Universities contribute to this in two ways: by producing 'human capital' through educating students and by producing new knowledge through their research function (e.g., Jasanoff, 1990).

Globalization is a third trend influencing universities (Hoelscher, 2012; Scott, 1998). Growing flows of money, goods, people, and ideas across national borders open up opportunities, but also put pressure on higher education institutions. Eased exchange of ideas and researchers allows for transnational learning processes and creative innovations. However, as a corollary of globalization, one also finds increased competition and market orientation in the field of research and education (Slaughter and Rhoades, 2004).

These three trends are probably the most important reasons for an increased interest of policy in universities as well as academia's "expansion of interest in policy and application," mentioned already in the introduction to the first edition of this Encyclopedia: National governments as well as international organizations today see universities, main providers of human capital and of research-based innovation, as crucial for establishing or keeping a good position in the international economic competition. As a result, governments increasingly call on universities to provide the means to this end, with implications for the organizational setup of universities as well as for the appropriate forms of knowledge (e.g., Pawson, 2006: 2).

Implications for Appropriate Forms of Knowledge

Academics in many universities increasingly feel a pressure to produce directly applicable knowledge and educate graduates with regard to the labor market, and many criticize this as political and economic appropriation and exploitation of academia. Instead, they endorse the disinterested pursuit of truth and knowledge and claim that the academic freedom, whose basis was already laid out to some extent in the *Authentica Habita* as early as 1155 and which, for example, in Germany until today is guaranteed in the Basic Constitutional Law, is a necessary prerequisite of scientific progress as well as the basis for the critical engagement with existing social norms and values. Additionally, independence and autonomy of teaching in higher education is seen as important contribution to democracy and consequently as civil right (Dahrendorf, 1965).

While these critics are probably right in many points, one has to keep in mind that universities always educated their graduates to be able to fill their positions in society (Rüegg, 1992–2011). Already in the Middle Ages, the universities trained their graduates in the traditional fields of philosophy, theology, medicine, and law for certain professions. While the traditional task of the university was to preserve the culture and knowledge by passing it on to future generations, the modern university with a strong focus on research and innovation, which was until then often concentrated in the academies, began to evolve only in the eighteenth century and became extremely successful in many countries during the nineteenth century. This went along with a differentiation of taught disciplines. In this context the foundation of empirical social sciences was mainly based in the need of the modern state to acquire in-depth knowledge of its citizens. More recently, still, many institutions took social problems as their starting point for their successful and inspiring research, for example, the Chicago School of urban sociology.

Two founders of the modern idea of the university, Humboldt and Newman, both had occupational positions, especially positions that one could label as policy related, in mind when formulating their (different) approaches: Humboldt wanted the university to train independent scholars by promoting the unity of research and teaching, as he was convinced that such scholars were best prepared to later become civil servants for the Prussian State. For Newman, universities' task was to educate (then only) Gentlemen to become future leaders within the Commonwealth. The interesting twist, however, is, that both saw more or less independence of universities from the state/church, the teaching of nonvocational subjects and the fostering of a liberal intellectual culture as the best way to reach these aims. At least since the end of the Second World War, and additionally fostered through the student revolts during the 1960s, the idea of the university was crucially based on this dual task of independent research on relevant societal problems and the education of knowledgeable and critical-reflexive citizens. But how do the developments described above impact on the research and teaching function of higher education?

With regard to teaching, the expansion of higher education participation has raised questions of skills, competencies, and

knowledge needed for being successful in the labor market. While citizenship education is still an explicit aim at least in some countries, there is a tendency in many countries to make curricula much more directly applicable and in line with the demands of employers. The European Bologna Reform of Higher Education is an often criticized example for this shift. A possible reaction to this pressure is a diversification of higher education systems, with a top layer of universities still following the traditional ideal and a majority of higher education institutions educating more or less applied subjects preparing their graduates directly for the labor market. However, at least with regard to policy knowledge, it is still debated whether such knowledge can be directly taught in an applied way, or if it does not need to be carefully developed through fostering reflexive and contextual thinking, as was the aim in the traditional university. This refers to the broader discussion of a distinction between experts, solving specific problems through the narrow application of scientific knowledge, and intellectuals, developing larger meta-narratives and even utopian thinking and interpreting society and the world through these lenses (already Chomsky, 1967; Merton, 1945).

With regard to research, especially the rise of the knowledge society has led to the formulation of new requirements for larger parts of scientific knowledge production, which go beyond the divide between applied and basic research. One prominent concept in this context, covering the most important aspects of this shift, is called Mode 2 knowledge (Gibbons et al., 1994). In a critical review, Weingart points out the following aspects as being central for proponents of the approach: Knowledge is seen as becoming increasingly problem based and has to be produced within concrete contexts of application. As these complex problems and contexts cannot be grasped by one scientific discipline alone, research has to be transdisciplinary. Additionally, the research output and results not only are evaluated by scientific peers within academia any more but also have to convince those actors that are involved in the specific problem or situation. Thereby, "knowledge-production becomes socially accountable and reflexive" and "comes under new pressures of legitimation" (Weingart, 1997: 593).

While the concept of Mode 2 knowledge production has been criticized for being exaggerated, it highlights important current trends with interesting epistemic implications. Mode 2 is as much about practical knowledge (techné/τέχνη in Aristotle's terminology) as about theoretical knowledge (episteme/ἐπιστῆμη), and even includes aspects of beliefs and opinions (doxa/δόξα) as well as, at least in the case of social sciences, wisdom or intelligence (phronesis/φρόνησις) (e.g., Flyvbjerg, 2001). Scientific results increasingly do not only have to pass the critical assessment of the academic community but also shall prove their applicability and usefulness in domains outside science and for nonscientific audiences. This might, however, result in tensions, when for example public policy and business pose differing demands on university research. One possible reason for this might be that exploitation interests of economic actors collide with society's interest in public availability of research results. A prominent example from the natural sciences was the race to decipher the human genome.

Implications for the Organizational Setup of Universities

The stress on knowledge and innovation by governments and in society at large and the subsequent changes in the forms of knowledge and its production also have an impact on the setup of universities.

As described, the increased need for scientifically based knowledge in different parts of society, as well as increased funding for the higher education sector, results in an enhanced demand for accountability of universities by different stakeholders (Benneworth and Jongbloed, 2010). With regard to policy knowledge and the social and behavioral sciences, these stakeholders mainly stem from the political realm or civil society, but also actors from the economy are involved to a growing extent, for example, in the context of technology assessments and forecasting.

On the macro level, increased accountability in many countries is just the other side of a growing (formal) autonomy of higher education institutions from the state as a result of a strange mixture of neoliberal ideology and New-Public-Management-thinking on the side of governments, and efforts to gain independence on the side of the universities (Lazzeretti and Tavoletti, 2006). The idea behind this is that universities need additional freedom to be able to partake successfully in the global competition for the best students, the best researchers, and the best innovations. This increased competition between higher education institutions on the national and international level has led to what is often called the 'Entrepreneurial university' (Clark, 1998).

One example is the distribution of additional money for research, which is increasingly project driven, allowing the government and other funders to define research topics and goals (Harman, 2010). While government funding for research in higher education and other public research institutions remained on average pretty stable at around 0.6% of GDP between 1981 and 2000 for all OECD countries, especially larger countries experienced a substantial decline (OECD, 2003: 78). Additionally, one can observe an internationally visible shift from institutional (block grants) toward project funding (especially during the 1970s and 1980s), and within project funding from academic-oriented instruments toward thematic-oriented instruments (Lepori et al., 2007). This probably leads to a greater applicability of research results. Within the context of the currently very tight budgets, there is however the danger of a certain blindness, so that research only looks at problems relevant for these funders, and that at the same time equally relevant problems of less financially potent parts of society are neglected. Therefore, it is still an open question whether the two sides of the coin, increased autonomy and accountability, actually increase the independence of academia, or if this results in new, maybe even more binding dependencies.

Increased accountability has been implemented inter alia through different means of steering universities, for example, through evaluations or the introduction of quality assurance and management programs (Danish Evaluation Institute, 2003). With regard to state funding, one can observe a shift from input to output measures. For example, university budgets are often no longer assessed on the basis of numbers of

students and scientists, but on the basis of graduates and the number of publications.

Another important development on the macro level is the growing importance of additional knowledge providers besides the traditional universities, such as think tanks, NGOs, government agencies, or consulting firms. Their absolute number is negligible compared with universities, but they are gaining influence and visibility, especially in the field of policy knowledge. While they are often still dependent on the research input from universities, their advantages are a better transfer of knowledge into the policy realm and often their clear ideological position. As an effect, they might buffer universities from too strong claims for production of applicable policy knowledge, but they might also detract important research resources from them.

On the organizational meso level a 'third mission' for universities besides teaching and research has been formulated as a consequence of the described shifts. This third mission sees active engagement of universities in transferring knowledge into society as an important task (Zomer and Benneworth, 2011). Universities are creating specific facilities to promote patenting or research collaboration with external partners, mainly relevant for the natural and engineering sciences and medicine (where they are often called 'technology transfer offices'), but they are also trying to establish new forms of collaboration for the social and behavioral sciences, such as engagement in public science.

Another important development at the university level is the establishment of inter- or even transdisciplinary centers besides the traditional departments based on scientific disciplines. With this reaction to the shift toward what was called Mode 2 knowledge production above, ever more universities establish problem-oriented centers, which are often built around policy-relevant topics, such as the 'Heidelberg Center for the Environment,' and frequently include external funding, for example, like the 'INET@Oxford' with money from a New York-based nonprofit foundation. Another important characteristic is the common collaboration between researchers from the social and behavioral sciences and researchers from the natural and life sciences.

On the micro level of the single researcher, the developments lead to important tensions that have to be solved creatively. As academia is still mainly organized along the traditional disciplinary lines, it is rational for each researcher trying to follow an academic career to become a specialist in her field and to orient his research toward inner-scientific research problems, as it is still difficult to publish applied research results with the best journals. On the other hand, the additional efforts needed to establish cooperation with partners from outside the university and to get engaged with the 'real' world, sometimes on the expense of scientific rigor, could result in increased funding. While there are always people on both extremes of these dimensions, handling this balancing act might secure certain academic standards for applied research in most cases. However, increased competition and a growing dependence on external research funding present the danger that applied research results are more and more aligned with the expectations of the external funders. This, in turn, would decrease the utility of this kind of knowledge to policy, as it would be unreliable and therefore potentially misleading.

In general, one can observe growing diversification within universities as well as within whole higher education systems. The shift toward a knowledge economy and toward growing competition between and within institutions has led to very different reactions on the various levels of academia. While the above-described trends toward more accountability, Mode 2 knowledge and interdisciplinary cooperation are an important facet of current developments, at the same time a growing specialization within disciplinary fields and a concentration on inner scientific quality measures can be observed. Interestingly, it seems that those institutions with the highest academic prestige also possess the highest reputation in the policy realm. Whether this is because of their superior performance or due to the seemingly larger legitimacy they give to policies supported by them, as many sociological neoinstitutionalists would argue, cannot be decided here.

The Transfer of Knowledge from Universities into the Policy Realm

A growing interest for scientific knowledge in the policy realm and a growing awareness of such a demand within the universities is, however, not enough to guarantee successful knowledge transfer.

A general problem for the research–policy nexus is that scientific knowledge is always to a certain extent preliminary. Following Karl Popper, science is a constant process that is not able to prove or verify a natural law, let alone other theories and hypotheses, finally. Researchers can only make their theories stronger by trying to test or falsify them and being unsuccessful in this endeavor. Additionally, Thomas Kuhn has shown how different paradigms within science replace each other. The current need to base political decisions on the newest research available to secure the legitimacy of these decisions has the paradoxical effect that the research findings are often still controversial: "The ever more intense recourse to knowledge on which to base decisions also reveals ever more clearly the state of ignorance" (Weingart, 1997: 606). As a result, policy can neither wait nor expect final words of wisdom from academia. This is not to say that scientific knowledge is always relative and cannot claim high validity. However, especially with regard to very new and very complex topics, there is always a certain uncertainty, which leaves room for discussion and even abuse (e.g., Collingridge and Reeve, 1986), a situation sometimes labeled as *postnormal science* (Funtowicz and Ravetz, 1993). This makes the reliance of policy-making on scientific knowledge, while necessary, for politicians and their advisors quite risky and difficult.

An interesting example of such an abuse is the renouncement of anthropogenic climate change. While a seldom found consensus within scientific research on the topic states that climate change can be traced back to human influences (see the IPCC as a representative of this consensus), financially potent interest groups successfully block political decisions by claiming the incertitude of these findings under a scientific disguise (e.g., McCright and Dunlap, 2010).

With regard to social and behavioral sciences, the reflexive reactivity of its research objects is another problem for a smooth knowledge transfer from these disciplines into the

policy realm. While the weather does not change due to weather forecasts, a prognosis that many people will die because of obesity might lead to a change in peoples' eating behavior, in the provision of food by the economy, and in the political framework, thereby refuting the original prediction.

Besides such general problems, different contextual factors can hamper or foster policy–research linkages in a more direct way. These can be categorized into three streams of explanation: supply-side accounts, demand-led explanations, and sociopolitical thinking on knowledge utilization (Maxwell and Stone, 2005: 2 ff.). Important factors on the supply side might be, for example, the lack of a proper research infrastructure in the first place, a missing interest of researchers in transferable and policy-relevant knowledge, or an absent awareness of research output's potential usefulness. Corresponding factors on the demand side might be ignorance of politicians with regard to the existence of such knowledge, or an incapability or even the missing preparedness to use such knowledge to inform policy. The category of sociopolitical thinking comprises different aspects, for example, questions of who the 'user' is, on how valid and reliable the knowledge actually is, and on what can be known anyway. Illuminative in this last respect are discussions around what has been labeled as 'Science wars' (Flyvbjerg, 2001: 1 ff.).

Broadly one can differentiate between at least three channels through which knowledge can flow from universities into the policy realm: the first is depersonalized information in the form of research reports, and so on, the second are organizations, and the third are individuals. The depersonalized transfer of information seems to be pretty straightforward at first sight, but has some prerequisites. First, with reference to the just mentioned factors on the supply side, the relevant information has to be supplied, and supplied in a way that is understandable for the policy side. In recent years, a growing share of research projects has become aware of this and produces policy-related research briefings on top of their normal research reports.

On the policy side some receptor units able to absorb the scientific information and make it available for politicians are needed. Government agencies, ministries, and parties are establishing smaller research units to fulfill this function. Additional intermediary organizations would be the already mentioned think tanks.

With regard to people, one can further differentiate between graduates starting a job for example in a governmental organization, established scientists moving from university into the political sphere, and politicians or civil servants joining academia for a certain period of time. The state is still one of the most important employers for higher education graduates in many countries overall, and holding a university degree increasingly becomes a prerequisite for joining the political elite, as is the case with economic elites. To foster knowledge transfer through this channel, it is important to equip these graduates with research-based, and therefore most up-to-date knowledge in their fields of study, to teach them how to apply this knowledge in praxis, and to engage them in lifelong learning. In some countries, certain academic institutions are especially successful in placing their graduates in high political positions (or rather people with such ambitions choose these institutions). For example, 6 out of the last 10 presidents of the United States hold a degree from either Yale or Harvard, and 8 out of the 11 British Prime Ministers since 1945 studied at Oxford University. It is obvious from these figures that certain academic institutions are better positioned for knowledge transfer than others.

The exchange of personnel is another mechanism for knowledge transfer well known from innovation studies. However, the amount of exchange between the political and the academic realm differs widely, depending on the openness of the two systems and the attractiveness of such a change for the individuals. In many cases, the exchange is not organized through a radical shift from one realm to the other, but through cooperation or cooptation, for example, in advisory boards. However, again large national differences exist.

There are certain advantages of the knowledge transfer through people over the mere exchange of information. The exchange of people results in an overlap of the two realms, opening up the opportunity for longer term cooperation as well as an easier flow of knowledge in both directions, both helping to "create[…] common understandings and identities" (Maxwell and Stone, 2005: 5). Additionally, there is a tacit dimension to knowledge, which cannot easily be transferred into formal information (Polanyi, 1967).

Implications for the Social and Behavioral Sciences

The current trends offer opportunities as well as threats for the social and behavioral sciences. The older democratic mass university was the *natural realm* of the social and behavioral sciences (Delanty, 2001). However, most of these disciplines lost importance in the entrepreneurial university since the 1980s, as the natural and engineering sciences are easier applicable and easier to commodify (Zomer and Benneworth, 2011). Keywords would be patenting, the emergence of science parks, or the cooperation with economically potent partners from the industry more generally. The social benefits of social and behavioral sciences' research are often much more diffuse and less visible, especially when the valorization of knowledge is narrowly understood in monetary terms. In many universities, social and behavioral sciences are therefore under pressure to prove their usefulness vis-à-vis the natural and engineering sciences. However, to restrict the social and behavioral sciences to a mere mechanistic problem-solving science, trying to imitate the harder sciences, is probably no option. Much of social and behavioral sciences' research refers to questions of values, power, and culture. To increase its impact, it has to develop a reflexive and dialogical approach, nevertheless based on sound and reliable empirical research, thereby contributing to "society's capacity for value-rational deliberation and action" (Flyvbjerg, 2001: 167).

The current situation offers many opportunities for this road. The knowledge produced by social and behavioral sciences is needed in the reflexive modernity more than ever: For example, many of today's societal problems, such as the aging of populations, migration, growing inequality, or environmental pollution, cannot be solved by merely technical solutions. Governments and international organizations have reacted to this for example by funding or even

initiating new large-scale data projects in the past few years (e.g., the European Social Survey or the PISA studies) or by supporting the development of new and more comprehensive indicators, for example, for well-being, complementing traditional measures (e.g., the Gross Domestic Product). Additionally, possibilities for the application of scientific innovations (e.g., genetic engineering; nanotechnology, the Internet) have to be analyzed with regard to their wider societal contexts and ramifications. The growing opportunities for interdisciplinary work within universities allows to account for social and behavioral sciences' theoretical and methodological tools to approach these and other pressing problems. However, this demands an openness or even the positive will to collaborate with other disciplines as well as with people from outside academia.

A severe problem, though, is the ongoing shift toward project-driven funding, as much of it is exclusively related to issues stemming from outside academia and affirmative in perspective. There is the threat that the social and behavioral sciences increasingly only contribute to effective technocratic interventions to problems formulated by others, instead of questioning the aims of these others. However, social and behavioral sciences need capacities to criticize the state, the economy, and society effectively and search for alternatives, as this is part and parcel of its task of applied research and positive impact.

Some Conclusions and Outlook

In the emerging knowledge society, political decisions increasingly have to be based on scientific results to be seen as legitimate. This has led to a shift from a supply to a demand-driven knowledge production. Governments and society at large increasingly hold universities accountable for providing such applicable knowledge. With regard to organizational consequences, these trends are often discussed under the label of universities' third mission, referring to the universities' preparedness of transferring their knowledge into different realms of society. The concept of Mode 2 knowledge production, instead, emphasizes more the epistemic changes within what is seen as useful scientific knowledge. Research, it is claimed, has to become more interdisciplinary, problem-centered, contextualized, and assessed by practitioners from the field of application.

This calls for improved ways of knowledge transfer between the realms of academia and policy. Researchers and universities as a whole are increasingly aware that they have to prepare their findings in ways that are accessible not only to peer academics, but to the broader society, the media, politicians, and the economy. Political organizations, on the other hand, have developed different means, such as own smaller research units, that are able to absorb these informations. Besides this exchange of information, the exchange of personnel is a much more direct way of knowledge transfer from one realm to the other. In most countries, one can find a growing number of additional organizations, for example, think tanks, moderating and influencing this transfer.

The processes of a 'politicization of science' (Weingart, 1997: 606) and of a 'scientification of policy' are probably self-reinforcing, especially under the condition of ever more citizens holding a higher education degree, making them ever more literate with regard to scientific findings.

This situation opens up opportunities as well as risks for the social and behavioral sciences and the university as a whole. With regard to impact, an opportunity is that it allows science to become more influential than ever. However, a crucial question here is whether political decisions are really based on the state of the art of scientific knowledge, or if the latter is just selectively used to legitimize the former. Science and its research findings are always preliminary. This insecurity makes it difficult for researchers to develop clear guidelines, and for politicians to use research findings in a straightforward way. Additionally, political decisions gain legitimacy from different sources, for example, from voters' demands, and often have to be a compromise with regard to different claims (and research findings). Actually, providing solutions to diverging claims is one of the key tasks of policy. All these put the direct policy impact of research into perspective. Nevertheless, in sum the increase of educated citizens, the rationalization of the world and growing research capacities will probably increase the reliance of policy on research and science in the decades to come remarkably.

A risk for university and science is the loss of its traditional autonomy due to political demands for knowledge. Important proponents of the idea of a university, such as Wilhelm von Humboldt in Germany or John Henry Newman in the United Kingdom, have unambiguously argued that universities can best fulfill their functions when they are independent from state, church, and other influences. The current increase in project funding in most established higher education systems can be seen as a sign of decreasing autonomy. In many developing countries, however, governments deliberately found new universities to provide expertise as well as experts for the administration.

Another big risk is a narrow commodification of research. If research will be confined to those aspects that pay off in monetary terms or that help to manage society in a mere technical way to follow the ideas of the government only, the social and behavioral sciences will loose one of their most important functions, namely the critical reflection on society. It is a future task for all researchers to solve the ostensible paradox of combining independence from governmental and economic organizations with an increased engagement with/openness toward policy-relevant issues or the policy relevance of their research.

As social scientists, we should however be aware that 'scientific expertise' as well as the increasing claims to it are no 'overwhelming force' (Stehr, 1992: 136), but are socially constructed and therefore can also be influenced by society.

See also: Applied Social Research, History of; Attitudes, Political and Public Opinion; Data Bases and Statistical Systems: Applied Social Research; Evidence-Based Practice; Forecasting; Higher Education Market; Higher Education Research; Implementation Science; Intellectuals: History of the Concept; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Knowledge Society, History of; Knowledge, Sociology of; Legitimacy, Sociology of; Modernity; Modernization,

Sociological Theories of; Objectivity of Research: Ethical Aspects; Objectivity: Philosophical Aspects; Participatory Action Research in Environmental and Ecological Studies; Participatory Action Research in Social Research; Policy Analysis: Evidence Based Policy-Making; Policy Analysis; Policy Knowledge: Foundations; Policy Knowledge: New Liberalism; Policy Process: Nonprofit Institutions in the United States; Popper, Karl Raimund (1902–94); Practical Reasoning: Philosophical Aspects; Quality of Government; Science Communication; Science and Media; Science and Politics: Value Neutrality; Science and The State; Science, Public Engagement with; Scientific Controversies; Scientific Knowledge, Sociology of; Social Science and Universities; Social Scientists as Experts and Public Intellectuals; Societal Impact Assessment; Technology Transfer; Think Tanks; Universities and Science and Technology: Europe; Universities and Science and Technology: United States; Weber, Max (1864–1920).

Bibliography

Bell, Daniel, 1973. The Coming of Post-industrial Society: A Venture in Social Forecasting. Basic Books, New York.

Benneworth, Paul, Jongbloed, Ben, 2010. Who matters to universities? A stakeholder perspective on humanities, arts and social sciences valorisation. Higher Education 59 (5), 567–588.

Boaz, Annette, Grayson, Lesley, Levitt, Ruth, Solesbury, William, 2008. Does evidence-based policy work? Learning from the UK experience. Evidence & Policy 4 (2), 233–253.

Chomsky, Noam, 23 February 1967. The responsibility of intellectuals. The New York Review of Books 8 (3). Retrieved 14.02.13 from: http://www.nybooks.com/articles/archives/1967/feb/23/a-special-supplement-the-responsibility-of-intelle/.

Clark, Burton R., 1998. Creating Entrepreneurial Universities. Organisational Pathways of Transformation. Pergamon Press, Oxford/New York.

Collingridge, David, Reeve, Colin, 1986. Science Speaks to Power: The Role of Experts in Policy Making. Pinter, London.

Dahrendorf, Ralf, 1965. Bildung ist Bürgerrecht: Plädoyer für eine aktive Bildungspolitik. Nannen, Hamburg.

Danish Evaluation Institute, 2003. Quality Procedures in European Higher Education. An ENQA Survey. European Network for Quality Assurance in Higher Education, Helsinki.

Delanty, Gerard, 2001. Challenging Knowledge: The University in the Knowledge Society. Open University Press, London.

Flyvbjerg, Bent, 2001. Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again. Cambridge University Press, Cambridge, et al.

Funtowicz, Silvio O., Ravetz, Jerome R., 1993. The emergence of post-normal science. In: Schomberg, René von (Ed.), Science, Politics, and Morality: Scientific Uncertainty and Decision Making. Kluwer, Dordrecht, pp. 85–123.

Gibbons, Michael, Limoges, Camille, Nowotny, Helga, Schwartzman, Simon, Scott, Peter, Trow, Martin, 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, London.

Harman, G., 2010. Funding of university research. In: Peterson, P., Baker, E., McGaw, B. (Eds.), International Encyclopedia of Education, third ed. Elsevier, Oxford, pp. 279–285.

Hoelscher, Michael, 2012. Universities and higher learning. In: Anheier, Helmut K., Juergensmeyer, Mark (Eds.), Encyclopedia of Global Studies. SAGE, London, et al., pp. 1713–1718.

Jasanoff, Sheila, 1990. The Fifth Branch. Science Advisers as Policymakers. Harvard University Press, Cambridge.

Lazzeretti, Luciana, Tavoletti, Ernesto, 2006. Governance shifts in higher education: a cross-national comparison. European Educational Research Journal 5 (1), 18–37.

Lepori, Benedetto, van den Besselaar, Peter, Dinges, Michael, Potì, Bianca, Reale, Emanuela, Slipersæter, Stig, et al., 2007. Comparing the evolution of national research policies: what patterns of change? Science and Public Policy 34 (6), 372–388.

Maxwell, Simon, Stone, Diane, 2005. Global knowledge networks and international development. Bridges across boundaries. In: Stone, Diane, Maxwell, Simon (Eds.), Global Knowledge Networks and International Development. Bridges across Boundaries. Routledge, London, pp. 1–17.

McCright, Aaron M., Dunlap, Riley E., 2010. Anti-reflexivity: the American conservative movement's success in undermining climate science and policy. Theory, Culture & Society 27 (2–3), 100–133.

Merton, Robert K., 1945. Role of the intellectual in public bureaucracy. Social Forces 23, 405–415.

OECD, 2003. Governance of Public Research. Toward Better Practices. OECD, Paris.

Pawson, Ray, 2006. Evidence-Based Policy. A Realist Perspective. Sage, London, et al.

Polanyi, Michael, 1967. The Tacit Dimension. Routledge & Kegan Paul, London.

Rüegg, Walter, 1992–2011. A History of the University in Europe, 4 vols. Cambridge University Press, Cambridge, et al.

Schofer, Evan, Meyer, John W., 2005. The worldwide expansion of higher education in the twentieth century. American Sociological Review 70 (6), 898–920.

Scott, Peter (Ed.), 1998. The Globalization of Higher Education. SRHE and Open University Press, Buckingham.

Slaughter, Sheila, Rhoades, Gary, 2004. Academic Capitalism and the New Economy: Markets, State, and Higher Education. Johns Hopkins University Press, Baltimore, et al.

Stehr, Nico, 1992. Experts, counsellors and advisers. In: Stehr, N., Ericson, R.V. (Eds.), The Culture and Power of Knowledge. Inquiries into Contemporary Societies. de Gruyter, Berlin and New York, pp. 107–155.

Trow, Martin, 1972. The expansion and transformation of higher education. International Review of Education 18 (1), 61–84.

Weingart, Peter, 1997. From 'finalization' to 'mode 2': old wine in new bottles? Social Science Information 36 (4), 591–613.

Zomer, Arend, Benneworth, Paul, 2011. The rise of the university's third mission. In: Jürgen, Enders, et al. (Eds.), Reform of Higher Education in Europe. Sense Publishers, Rotterdam, pp. 81–101.

Policy Networks

William D Coleman, University of Waterloo, Waterloo, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of a policy network first emerged in the 1980s and early 1990s as a means to study changing forms of public policy making, particularly the shift from vertical to more horizontal decision-making in states. In many instances, this horizontal approach to policy making was, in turn, a response to the growing internationalization of public policy. In recent years, more and more policy making is taking place 'beyond the state,' with the growing importance of global governance. Consequently, global policy networks have become increasingly important political participants.

Introduction

The concept of a policy network first emerged in the 1980s and early 1990s as a means to study the changing forms of public policy making, particularly the shift from vertical to more horizontal decision making in states. As horizontal decision making became more institutionalized in the governments of many countries, important changes in governance were taking place beyond nation-states at regional and global scales. Intensifying globalization processes were leading to the institutionalization of more decision-making at these supranational scales. The concept of 'global governance' emerged in the late 1980s, pointing to these new global policy-making fora. The policy-making processes for areas of global governance involved newer, more complex forms of policy networks and policy making. The analysis of various forms of policy networks begins with an analysis of policy networks involving nation-states and is followed by a discussion of the more complex networks of networks that have emerged as part of global governance.

A policy network refers to a set of informal and formal interactions between a variety of usually collective public (state and international governmental organizations) and private actors, who have different but interdependent interests. Operating in a more or less institutionalized setting, these actors are engaged in horizontal, relatively nonhierarchical discussions and negotiations to define policy alternatives, formulate policies, or implement them. Policy networks vary along two core dimensions: the structural pattern according to which public power is shared among the members and the degree and patterns of integration among the members. These two dimensions developed along distinct conceptual routes, only merging into a unified concept for the study of nation-state policy making in the mid-1990s. These two dimensions are analyzed empirically using a variety of methodologies including quantitative network analysis, elite interviews, and close examination of documents and archival materials. The principal challenges for the future of the analysis of policy networks include the resolution of their theoretical status and adjustment to global governance.

The Changing Context of Nation-State Policy Making

The state-focused definition of policy networks introduced above, came as a response to changes in the form of political decision-making in many democratic states. In particular, the standard notion of a clear separation between state and society, with the state being seen as the highest center for social and political control gradually came into question. As Mayntz (1993) observed, social modernization brought more importance to formal organizations in all areas of society, and with their strengthened position came an increasing fragmentation of power. Policy networks represented one intellectual conceptualization for responding to the escalating functional differentiation of social subsystems. With the state no longer being the sole entity capable of organizing society, there emerged a dispersion of expertise and competence among private and public actors; a multiplication of channels for mediation and agreement; and the inclusion of different levels of decision-making from the local to the supranational (Jouve, 1995). Consequently, nation-state public policies became much less likely to be produced by a unitary state with full control over all the resources necessary for implementing decisions in the public interest. Rather, nation-state public policy came to result from strategic interaction among several or many policy actors, each with a particular understanding of the problem and its own individual or institutional self-interest (Scharpf, 1997; **Table 1**).

Table 1 Forms of public policy beyond the nation-state

	International public policy	*Transnational public policy*	*Global public policy*
Key actors	States	States, nongovernmental organizations, corporations, social movements, individuals	States, nongovernmental organizations, corporations, social movements, individuals
Scale of activity	Intraregional Interregional Global	Intraregional Interregional	Global

The Historical Development of the Concept

The sharing of political power between public and private actors became the first avenue of enquiry that led to the development of the policy network concept. In the mid-1970s, Schmitter and Lehmbruch (1977) suggested that the concept of corporatism be reconsidered as a conceptual tool for understanding the particular relationships between public and private actors. They pointed to these relationships emerging in the management of macroeconomic policy in several European democracies. The revitalization of the corporatism concept coincided with other new concepts suggested for capturing the changing patterns of power sharing between private and public actors: 'corporate pluralism' in Scandinavia, 'issue networks' in the United States (Heclo, 1978), 'policy communities' in the UK (Richardson and Jordan, 1979), and 'policy networks' (Katzenstein, 1978).

The early focus in this research was on broad, macro-characteristics of nation-state political systems. As these concepts became more familiar, however, a number of scholars noted their potential relevance for understanding private–public relations in more specialized sectoral policy arenas. Drawing on a growing number of empirical studies, these scholars began to develop typologies of different structural patterns of public–private interaction at the sectoral or meso-level (Atkinson and Coleman, 1989; van Waarden, 1992), prompting a certain spate of definitional disputes. By the mid-1990s, however, sectoral and state corporatism, pressure pluralism, clientelism, issue networks, and state-directed networks generally were accepted as core structural types at the nation-state level.

In addition to patterns of power sharing between public and private actors, a second line of inquiry developed, focusing on the degree of integration among the various actors involved in these horizontal policy-making settings. Rhodes (1990) and Rhodes and Marsh (1995) targeted the degree of integration as their key variable and outlined a typology of the different types of 'policy networks' that ranged from fairly tightly integrated 'policy communities' to very loose and open 'issue networks.' The horizontal arrangements in policy making also triggered the interest of political sociologists interested in patterns of communication between large organizations. Laumann and Knoke (1987) sought to identify network structures by examining how information was transmitted between policy actors, the nature of the resource transactions between actors, and the relative openness of boundaries in various policy domains. These studies employed quantitative social network analysis developed in sociology. By the mid-1990s, this quantitative approach had incorporated insights from other policy network writings (Pappi and Henning, 1998).

When speaking of the degree of integration of the policy network, it is usual to employ a concept that refers to the aggregate of actors involved. The term 'policy community' is utilized most commonly for these purposes, but terms like 'policy domain' or 'policy subsystem' are also employed. Research has established a number of factors that are related to integration:

1. Openness of the policy community. The more open the deliberations in the policy community tend to be, the less integrated they are likely to become.
2. Size of the policy community. The larger the number of members of the policy community, the greater the obstacles to high integration.
3. Frequency of interactions among community members and the degree to which these interactions become 'institutionalized' (as opposed to remaining ad hoc and fluid). Institutionalization normally enhances the degree of integration.
4. Membership stability. The more changeable the membership of a policy community, the less integrated it will be.
5. Density and structure of information exchange. The denser the ties for information exchange and the more direct the relationships, the more integrated the policy community.

Another, particularly important factor affecting the level of integration of a policy community is the degree to which its members share basic beliefs, values, and norms. Some actors within a given policy community may come to share a set of deep normative core beliefs and from these develop basic strategies and policy positions (Sabatier and Jenkins-Smith, 1993). These 'advocacy coalitions' create 'spaces of meaning' within the policy community that provide a framework for political action. The degree of integration of the policy community will be affected by how many distinct 'spaces of meaning' it holds. Sabatier and Jenkins-Smith suggest that policy communities will normally contain two to four advocacy coalitions. They add that the number will probably be higher when the policy community is newly formed and may dwindle to one dominant coalition as policy making becomes routine. The concept of an epistemic community developed in international relations has certain properties in common with Sabatier and Jenkins-Smith's 'advocacy coalition'. All or part of an epistemic community may act as an advocacy coalition. When this development occurs, these epistemic communities often provide important links between policy communities, whether in other nation–states or at the transnational level.

Different types of policy networks may be distinguished depending on whether state actors share political power with private actors and on the relative resource balance between private and public players. State actors may share political power by inviting private actors to participate directly in the formulation or implementation of policy. Here private actors become 'policy participants.' Corporatist and clientelist networks have this feature; or state actors may monopolize policy formulation and policy implementation, while listening to, or receiving submissions from private actors. In such a relationship, private actors perform as 'policy advocates.' Pressure pluralism is the most common network with this feature, but it also occurs in issue networks and state-directed networks. The decision by state actors to share political power affects the overall balance of resources between public and private actors in the policy network. Resources may be relatively balanced, they may favor the state actors, or they may advantage civil society actors. Both pressure pluralist and sectoral corporatist networks feature a balanced relationship. Private actors have the stronger position in clientelist and issue networks while public actors dominate in state-directed and state-corporatist networks.

The Theoretical Status of the Concept at the Nation-State Level

Although definitional disputes surrounding the concept have gradually been resolved, the overall theoretical status of policy networks remains subject to debate. Two fairly fundamental criticisms have been directed at the concept. First, Dowding (1995) has charged that the concept is nothing more than a metaphor and will thus fail to produce theories of the policy process. Secondly, the concept is said to have limited relevance because of an apparent lack of suitability for explaining policy innovations and policy change.

Policy Networks as Metaphors Only

Dowding has argued that the attempts to provide a middle range theory of policy making utilizing the policy network concept are bound to fail. He suggests that the driving forces of the explanation – the independent variables – are not characteristics of the network as such, but of the components within the networks. Variables such as the relative openness of a policy network are cast merely as reworked descriptions of policy phenomena. Related criticisms argue that policy network studies have failed to explain both the mutual dependence of public and private actors and the conditions under which particular policy networks arise (Thatcher, 1998). Critics also lament the lack of specificity of the hypotheses that do exist, a failing bound to render any empirical testing rather unconvincing.

Although useful for spurring discussion and reflection, there are several reasons why these criticisms may be premature. First, the criticisms undervalue the descriptive utility of the policy network concept. One of the most difficult empirical problems faced in the analysis of public policies is the specific determination of the actors of consequence in any policy domain. Although one might be able to begin to specify these actors by utilizing a good understanding of the formal institutions and the persons occupying key positions, such a specification is almost inevitably bound to be incomplete. Policy making is replete with informal relationships and the operation of subtle, somewhat opaque norms that empower certain actors in unexpected ways. Accordingly, the methodologies developed for delimiting a policy network have a broader use in policy analysis. By identifying the central actors in policy making, the analysis of the processes becomes more straightforward for determining the policy ideas that are translated into particular policy proposals through interchanges between identified sets of actors. This identification process becomes particularly helpful in the policy areas featuring polycentric governance discussed below.

Secondly, numerous empirical studies subsequent to the publication of Dowding's critique have provided convincing evidence of relationships between variations in policy networks and policy outcomes (**Table 2**). For example, Josselin's (1997) study of the negotiation of core directives for a single market in financial services in the EU, evaluates the extent to which different types of networks facilitate or hinder the promotion of the interests of member states in EU-level negotiation settings. She shows that a relatively closed corporatist network at the national level is a decided asset in promoting a unified position at the EU level. Conversely, however, a more fluid issue network makes such unity more difficult, while providing additional degrees of freedom in the fluid decision-making arenas that exist at the supranational level. Similarly, Daugbjerg (1998) relates the relatively closed character of corporatist networks to patterns of political change. The more closed and cohesive the network, the more incremental the pattern of change and the more stable the utilization of policy instruments. Coleman's (1996) study of the impact of globalization on the regulation of financial services at the domestic level traces relationships between particular types of policy networks and the varying adaptation of countries to these globalizing pressures.

Table 2 Variation in governance arrangements beyond the state

Level of institutionalization of international, transnational, global governing arrangements	*Levels of public sector activism in governing arrangements*	
	High	*Low*
High	Polycentric governance	Self-regulatory and private regimes
Low	Intergovernmental negotiations	Loose couplings

Finally, Dowding's critique is addressed by some scholars who are using more formal models. Pappi and Henning (1998) elaborated a complex model that suggests direct relationships between types of networks (pluralism, corporatism, clientelism, state-directed) and policy outcomes. Scharpf's (1997) 'actor-centered institutionalism' introduces network-based negotiation settings as one of several core types of forums for policy making. Drawing on game theory, he sets out a series of propositions that provide a basis for modeling negotiations between actors in such networks and deriving probable differences in policy outputs based on general welfare and distributive justice criteria. Both of these more formal theoretical approaches hold considerable promise for the development of a middle-level theory where policy networks will be an important independent variable.

Policy Networks and Policy Change

Policy networks have been conceptualized primarily using structural variables. Such structural characteristics are helpful for building descriptions of the policy process at a particular point in time and, more recently, for positing testable explanations of variations in policy outcomes. Dowding (1995) questions, however, whether these same variables are sufficient for explaining how and why policies change and why networks themselves change. Two avenues of investigation have been pursued for addressing these criticisms. The first focuses on the relationship between policy ideas and policy networks and the second draws on an analysis of boundary rules.

Sabatier and Jenkins-Smith (1993) argue that policy networks include one or more advocacy coalitions that are composed of actors who share deep-core beliefs, near-core policy beliefs, and secondary beliefs about the use of policy instruments. Changes in the relative dominance of advocacy

coalitions in a policy network may prompt policy change. Or members of coalitions may negotiate adjustment of secondary beliefs that, in turn, open the door to policy change. Paralleling Sabatier and Jenkins-Smith's work, Jobert (1995) argues that a 'political forum' and a 'scientific or expert' forum exist alongside a policy network. When politicians begin to experience problems that cannot be addressed well within a policy network, they begin to reach to these other fora for advice. To the extent that this advice is acted upon, the policy network may be dissolved, with new arrangements taking its place.

What both these theoretical arguments stress is the importance of looking at the interaction between policy ideas and the entities who share these ideas within a given policy network. Coleman et al. (1996) follow up on these insights by arguing that the degree of sharing of policy ideas in a given policy network will determine the approach to policy negotiations among the actors in the network. They hypothesize that the more that ideas are shared, the more policy discussions will take on aspects of what Scharpf (1997) has called 'positive coordination' or a 'problem-solving' decision making style. They suggest that positive coordination is more likely in closed networks, like corporatism, than in more open networks such as pressure pluralism. By coupling policy network structures to policy ideas, they then trace different patterns for how policy change occurs. Scharpf's work stresses the importance of looking at actor preferences and cognitive orientations in studying policy processes. The richness of his theoretical work offers further opportunities for improving theories of policy change.

The second avenue for improving explanations of policy change focuses on different types of boundaries within policy communities. Rhodes and Marsh (1995) note that a policy community can have several policy networks. These, in turn, may operate at two levels: a central level whose members dispose of considerable resources; and a peripheral level composed of those with few or no resources. Changes in the actors at the center of a policy community and thus in the dominant policy network, or movements from the peripheral networks to the dominant one, will be crucial intervening variables in explanations of policy change. Similarly, new actors may build peripheral networks over time and then contest the exclusivity of the existing dominant policy network. To the extent that new actors enter an existing policy network, they may bring about change in its structure. For example, entry of new actors into a previously closed corporatist network may foster sufficient dissent that the state refuses to involve private actors any longer in policy making. Once such actors are excluded, the network takes on pressure-pluralist characteristics, altering, in turn, the policy process and perhaps policy outcomes (Coleman et al., 1996).

Global Governance and Policy Networks

The growing policy making taking place beyond the state raises new challenges for policy network analysis because it has been used primarily to study nation-states and domestic policy making alone. In these changed circumstances, states remain important institutions in policy making but they are joined by other institutions creating situations of what Scholte (2005) has termed 'polycentric governance' as opposed to 'state governance.' Although scholars developed the concept of 'policy networks' to help in the study of policy making by nation-states, the concept retains its value in the analysis of global governance. In fact, the structural form of the network is dominant when it comes to policy making on a global scale.

Defining Global Governance

For the purposes of this discussion, I follow Scholte and define 'global governance' as: "a complex of rules and regulatory institutions that apply to transplanetary jurisdictions and constituencies" (2011: 8). In using the word transplanetary, Scholte is speaking of linkages between persons or organizations "situated at any inhabitable points on the earth" (2005: 59). The linkages are not confined to a particular country or to a region. They have the potential to be from anywhere to anywhere in the world. The definition also leaves open the possibility that global governance involves not only nation-states, but also other types of actors including business firms, civil society organizations, local governments, and regional groups and agencies (2005: 9).

Connected to the concept of global governance is the development of what might be termed a 'global public space' where policy making takes place. John Ruggie (2004: 219) refers to an emerging 'global public domain,' which he defines as: "an institutionalized arena of discourse, contestation, and action organized around the production of global public goods. It is constituted by interactions among nonstate actors as well as states. It permits the direct expression and pursuit of a variety of human interests, not merely those mediated (filtered, interpreted, promoted) by states. It 'exists' in transnational non-territorial spatial formations and is anchored in norms and expectations as well as institutional networks and circuits within, across, and beyond states." Similarly, Castells (2008) refers to a 'global public sphere' that exists with a political/institutional space that is not subject to any particular sovereign power. Rather it is shaped by relationships between states and global nonstate actors. He adds that "a variety of social interests express themselves in this international arena: multinational business, world religions, cultural creators, public intellectuals, and self-defined global cosmopolitans (2008: 80)." For both Ruggie and Castells, global governance takes place in these spaces when various constellations of the actors involved propose, debate, and agree upon policies and rules that are then implemented by both public and private actors including states.

Global Networks and Global Policy

The functioning of governance in these global spaces involves complex and variable networks. Scholars such as Ulrich Beck (2005), Manuel Castells (2008, 2009), and Phil Cerny (2010) argue that states, transnational corporations, civil society actors, and social movements transform themselves into collective actors in these global public spheres through the formation of networks. States form 'public' networks with one another for various functions, as Slaughter (2004) has observed. These might include sharing information,

coordinating policy, cooperating on enforcement, and exchanging particular regulatory forms (2004: 40). Castells (2008, 2009) argues that the "emerging network state" shares sovereignty and responsibility with other states and differing levels of government. There is also "flexibility of governance procedures and greater diversity of times and spaces in the relationship between governments and citizens compared to the preceding nation-state" (2009: 40). Other scholars such as Scholte see states as becoming less central to policy making. He speaks of the movement from "statism" to "polycentrism." Scholte argues that "governance has become distinctly multi-layered and trans-scalar. Regulation occurs at – and through interconnection among – municipal, provincial national macro-regional and global sites. No single level reigns over the others. Governance tends to be diffuse, emerging from multiple scales at once" (2005: 186).

As these examples indicate, globalization has brought on a need for global public policy, one of the functions of global governance. Global public policy refers to policies that are directed to the whole of the global polity, however disorganized and difficult the definition of its borders might be (Coleman, 2012). As Cerny (2010: 98) observes, "the constitution of the public itself is being transformed in the context of political (as well as economic and social) globalizing trends." For these reasons, 'global' public policy potentially draws input, advice and participants in its 'formulation' from anywhere in the world; the formulation of policies is potentially transplanetary. Similarly, once it is agreed upon, the given policy has the potential to be implemented in any place on the planet or in all places in the planet. This understanding does not mean that those involved in policy formulation come from everywhere in the world; rather it speaks to the potential that they could do so. Similarly, the definition leaves open the question of whether a given global policy is implemented in all places in the world; rather, it implies that no place in the world is necessarily excluded from receiving the benefits or complying with the requirements, of the policy.

The increased complexity of policy making on a global scale, in turn, leads to different structures for global policy networks. Generally speaking, these networks take the form of 'networks of networks.' Castells identifies two mechanisms that come into play in these circumstances. First, there is the "ability to constitute network(s)" and the capacity to "program/reprogram the networks" in pursuit of policy goals of the given networks (2009: 45). Secondly, there is the "ability to connect and ensure the cooperation of different networks by sharing common goals and combining resources" (2009: 45). Therefore, in this global policy context, the hypothesis is that the most effective global policy networks will be those which "possess the most transnationally interconnected resources, power and influence" (Cerny, 2010: 106). Cerny adds that successful builders of these networks of networks are those who are "able to build cross-border networks, coalitions and power bases among a range of potential allies and adversaries" (2010: 106), and are "able to coordinate and organize their strategic action on a range of international, transnational and translocal scales" (2010: 106).

The complexity, instability, and size of global policy networks pose significant challenges for their study. The structure and range of activities of these networks can vary widely from one policy area, like intellectual property rights or agricultural production, to others focused on global health or environmental degradation. Scholars approaching their studies from an international relations perspective have sought to conceptualize these networks by building on the notion of an international regime. Although an individual international regime (composed of states) might be relatively tightly organized, they are often not able to respond fully to global policy issues. In response to this limitation, Keohane and Victor note the emergence of 'regime complexes' as responses to global governance challenges. "Regime complexes are marked by connections between the specific and relatively narrow regimes but the absence of an overall architecture or hierarchy that structures the whole set . . ." (2011: 8). In such a scenario, the regime complexes participate to varying degrees in policy processes involving different levels of cooperation. In essence, they operate as networks of more narrowly defined regimes that have a particular common interest in a set of policy issues. What is distinctive about these particular networks is that the states remain the dominant, if not the only, players in the pursuit of global policy goals.

Other scholars argue that maintaining the centrality of states in studying global networking, underplays other structural forms where networks of states share authority with networks built with different types of institutions. Biermann et al. (2009) suggest that the kind of state centrism in the regime-complexes hypothesis can be avoided by working with the concept of "global governance architectures." They define such architectures as: "the overarching system of public and private institutions that are valid or active in a given issue area of world politics." This system comprises organizations, regimes, and other forms of principles, norms, regulations, and decision-making procedures. Architecture can thus be described as the "meta-level of governance" (2009: 15). Biermann et al. develop the architecture further by noting that these can vary in their degree of fragmentation. At one end, the 'governance architecture' can be 'synergestic,' that is, built around a 'core institution'; in contrast, the architecture defined as "conflictive" is comprised of "different, largely unrelated institutions" (2009: 19). In this latter ideal type, the degree of coordination among participating global policy networks will be low. Between these two ideal types, Biermann et al. define a third architecture titled "cooperative" where participating institutions and networks are loosely integrated (2009: 19–20).

Future Challenges

The concept of a policy network was designed in the first instance for the analysis of policy making at the national and more local levels. However, as globalization has increased the internationalization of many policy areas, the likelihood rises that several policy networks will be interacting with one another across different levels of decision-making. Future research, therefore, will need to focus on improving understanding of the relationships between policy networks at varying scales beyond the nation-state. Some of these networks will have the additional property of being transnational or global in character; that is, they will include

actors from nation-states and other actors operating at national, regional, and global scales.

The concept of policy community mediators offers one avenue for research on the relationships between policy networks (Coleman and Perl, 1999). In many of the more globally extensive policy domains, boundary rules will create overlapping memberships between policy networks, with mediators emerging from among those actors with joint memberships. Mediation can take place at two levels. First, a mediator may operate as a pragmatic policy broker who looks for opportunities to lessen conflict within and between policy networks in order to move the policy process forward. Secondly, and perhaps even more importantly, mediation takes place on an intellectual plane, involving the 'translation' of policy ideas and policy paradigms between policy networks. Mediators construct a new view of the world that has a place for the policy concerns of multiple networks and provides a language and set of norms that enable them to speak to one another. Research on linkages between policy networks and on the role of policy mediators will be crucial to further theorizing on policy networks in this globalizing era.

See also: Network Analysis; Social Networks.

Bibliography

Atkinson, M.M., Coleman, W.D., 1989. Strong states and weak states: sectoral policy networks in advanced capitalist economies. British Journal of Political Science 19, 47–67.

Beck, Ulrich, 2005. Power in the Global Age: A New Global Political Economy. Polity Press, Cambridge, UK.

Biermann, F., Pattberg, P., van Asselt, H., Zelli, F., 2009. The fragmentation of global governance architectures: a framework for analysis. Global Environmental Politics 9, 14–40.

Castells, Manuel, 2008. The new public sphere: global civil society, communication networks, and global governance. The ANNALS of the American Academy of Political and Social Science 616, 78–93.

Castells, Manuel, 2009. Communication Power. Oxford University Press, Oxford.

Cerny, Philip, 2010. Rethinking World Politics: A Theory of Transnational Neopluralism. Oxford University Press, Oxford.

Coleman, W.D., 2012. Governance and global policy. In: Levi-Faur, David (Ed.), Oxford Handbook of Governance. Oxford University Press, Oxford, pp. 678–685.

Coleman, W.D., 1996. Financial Services, Globalization and Domestic Policy Change. St. Martin's Press, New York.

Coleman, W.D., Perl, A., 1999. Internationalized policy environments and policy network analysis. Political Studies–London 47, 691–709.

Coleman, W.D., Skogstad, G., Atkinson, M., 1996. Paradigm shifts and policy networks: cumulative change in agriculture. Journal of Public Policy 16, 273–301.

Daugbjerg, C., 1998. Linking policy networks and environmental policies: nitrate policy making in Denmark and Sweden 1970–1995. Public Administration 76, 275–294.

Dowding, K., 1995. Model or metaphor? A critical review of the policy network approach. Political Studies–London 43, 136–158.

Heclo, H., 1978. Issue networks and the executive establishment. In: Beer, S.H., King, A. (Eds.), The New American Political System. American Enterprise Institute for Public Policy Research, Washington, DC.

Jobert, B., 1995. Rhétorique politique, controverses scientifiques et construction des normes institutionnelles: esquisse d'un parcours de recherche. In: Faure, A., Pollet, G., Warin, P. (Eds.), La Construction du Sens dans Les Politiques Publiques: Débats Autour de la Notion de Référentiel [The Construction of Meaning in Public Policies: Debates Surrounding the Concept of the Référentiel]. L'Harmattan, Paris.

Josselin, D., 1997. Money Politics in the New Europe: Britain, France and the Single Financial Market. Macmillan, Basingstoke, UK.

Jouve, B., 1995. Réseaux et communautés de politique publique en action. In: Le Galès, P., Thatcher, M. (Eds.), Les Réseaux de Politique Publique. Débat autour des Policy Networks [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris.

Katzenstein, P., 1978. Conclusion: domestic structures and strategies of foreign economic policy. In: Katzenstein, P.J. (Ed.), Between Power and Plenty: Foreign Economic Policies of Advanced Industrial States. University of Wisconsin Press, Madison, WI, pp. 295–336.

Keohane, Robert O., Victor, David G., 2011. The regime complex for climate change. Perspectives on Politics 9, 7–23.

Laumann, E.O., Knoke, D., 1987. The Organizational State. University of Wisconsin Press, Madison, WI.

Mayntz, R., 1993. Policy-netzwerke und die Logik von Verhandlungssystemen [Policy networks and the logic of negotiation systems]. In: Héritier, A. (Ed.), Policy-Analyse: Politische Vierteljahresschrift Sonderheft, vol. 24. Westdeutscher Verlag, Opladen, Germany.

Pappi, F.U., Henning, C.H., 1998. Policy networks: more than a metaphor? Journal of Theoretical Politics 10, 553–575.

Rhodes, R., 1990. Policy networks: a British perspective. Journal of Theoretical Politics 2, 293–317.

Rhodes, R., Marsh, D., 1995. Les réseaux d'action publique en Grande Bretagne. In: Le Galès, P., Thatcher, M. (Eds.), Les Réseaux de Politique Publique. Débat autour des Policy Networks [Public Policy Networks: Debates on Policy Networks]. L'Harmattan, Paris.

Richardson, J.J., Jordan, A.G., 1979. Governing under Pressure: The Policy Process in a Post-Parliamentary Democracy. Robertson, Oxford, UK.

Ruggie, John, 2004. Reconstituting the global public domain. European Journal of International Relations 10, 499–531.

Sabatier, P.A., Jenkins-Smith, H.C., 1993. Policy change over a decade or more. In: Sabatier, P.A., Jenkins-Smith, H.C. (Eds.), Policy Change and Learning: An Advocacy Coalition Approach. Westview Press, Boulder, CO, pp. 13–40.

Scharpf, F.W., 1997. Games Real Actors Play: Actor-Centered Institutionalism in Policy Research. Westview Press, Boulder, CO.

Schmitter, P.C., Lehmbruch, G., 1977. Trends toward Corporatist Intermediation. Sage Publications, London.

Scholte, J.A., 2011. Global governance, accountability and civil society. In: Scholte, J.A. (Ed.), Building Global Democracy: Civil Society and Accountable Global Governance. Cambridge University Press, Cambridge, UK.

Scholte, J.A., 2005. Globalization: A Critical Introduction. Palgrave Macmillan, London.

Slaughter, Anne-Marie, 2004. A New World Order. Princeton University Press, Princeton, NJ.

Thatcher, M., 1998. The development of policy network analyses: from modest origins to overarching frameworks. Journal of Theoretical Politics 10, 389–416.

van Waarden, F., 1992. Dimensions and types of policy networks. European Journal of Political Research 21, 29–52.

Policy Process: Nonprofit Institutions in the United States

Donald T Critchlow, Arizona State University, Phoenix, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the development of expertise and policy knowledge from the late nineteenth century through today within the context of a strong antistatist tradition. The growth on nonprofit institutions, especially think tanks, emerged in the early twentieth century to bring nonpartisan expertise to public policy and government administration. In the post-World War II period, policy expertise became increasingly politicized but continues to play an essential role in public policy within the American system.

Modern governments requires specialized knowledge and administrative expertise. In the USA, government bureaucracies rely heavily on private research organizations and nonprofit foundations, operating outside of government, to provide this expertise. The expanded role of the federal government in the post-World War II era strengthened the network of relations among public agencies and private groups including nonpartisan research institutes. The peculiar relationship reflects a strong antistatist tradition in America, evident in the role of private reform organizations in the early and late nineteenth century. In the late twentieth century, nonpartisan research organizations such as the Brookings Institution, the American Enterprise Institute, the Heritage Foundation, and the Institute for Policy Studies gained prominence for providing specialized knowledge to government policymakers, even while these organizations reflected ideological positions in their research orientations.

The complexity of advanced industrial systems requires governments to acquire an enormous range of specialized knowledge and administrative expertise with which to order their affairs. Uniformity in demand, however, has not produced uniformity in the character and supply of knowledge for government purposes. There are significant variations, related to diverse cultural traditions in the means of acquiring knowledge and applying methods of expertise. For example, central government bureaucracies in Western Europe and Japan directly provide many more sources and opportunities for the development of policy expertise than they do in the USA, which in modern times has relied heavily on private research organizations operating outside government bureaucracies for functional expertise. As the federal role in economic and social policy expanded in the post-World War II era, increased federal funding did not produce a commensurate increase in the size of government. Instead, it strengthened the complicated network of relationships among the federal government, state and local governments, and private groups, including nonpartisan research institutes (Alchon, 1985; Critchlow, 1985; Ricci, 1993; Smith, 1991).

In this policy system, the federal government operates indirectly, through intermediary organizations in state and local government, and in civil society (Heclo, 1980). This reliance on private, nonpartisan research organizations imparts a unique quality to the acquisition and use of knowledge by the administrative state in modern America.

The vast size, complexity, and expense of this system of acquiring knowledge by government in the USA suggests the pluralistic nature of the American polity, deeply rooted in its republican distrust of centralized government. This distrust of centralized government found expression in the nation's constitution, drafted in Philadelphia in 1787. Its famous beginning 'We the people' asserted that the source of its authority rested in the people rather than state or centralized government. This distrust of \centralized government found expression in the antistatist Jeffersonian–Jacksonian tradition that emerged early in the nation's history.

Social Investigation in the Nineteenth Century

In the environment of relatively weak centralized government prior to the US Civil War, the craft of social investigation developed within a middle-class subculture of reformers, many of them women, who documented the plight of slaves, the poor, prisoners, and the insane. The affinity of antebellum social reform and feminism became evident in reports issued by American abolitionists who documented the conditions of slave life, as well in reports on the care of prisoners, the indigent, and the insane. In the 1840s, social reformer Dorothea Dix issued a series of groundbreaking reports to state legislatures calling for the humane care of the mentally ill. This tradition of empirical research into social problems conducted by private individuals associated with reform movements was extended late into the nineteenth century (Marshall, 1937; Nelson, 1992; Leach, 1989; Daniels, 1968; Dupree, 1957).

By the outbreak of the Civil War, state legislatures in the most populous states regularly received detailed narrative and statistical reports on dependency, crime, public health, labor conditions, and economic development. These reports relied more on democratic and humanitarian sentiments associated with moral reform than they did on claims of science and disinterested expertise. Female reformers, as suggested by the role of Dorthea Dix, played an especially important role in this early endeavor to bring social knowledge to state legislatures and the public (Furner and Lacey, 1993; Ross, 1991).

While the capacity for social knowledge grew on the state level, the role of the federal government's activities in the pursuit of knowledge remained limited generally to surveying, land exploration, and modest technological development

through private contracts. This relative institutional impoverishment reflected a unique antistatist tradition in America, reinforced by sectionalist politics of the period. The major institution in Washington for the gathering and accumulation of scientific and social knowledge remained the Smithsonian Institution, which opened in 1846 at the instigation of the individual for whom it was named. In the decades following the Civil War, the Smithsonian Institution became a center for people such as Francis Walker, John Eaton, John Wesley Powell, and Carroll D. Wright, who took leading roles in establishing a rudimentary infrastructure of social investigation in the nation's capital. These individuals served as conduits for a transition from moral reform to scientific reform.

The Twentieth Century

By the beginning of the twentieth century a large resident class of scientific and professional civil servants had emerged in Washington. In 1904, the US Bureau of Census reported that 3919 civil servants fell into the classification of 'professional, technical, and scientific' workers. Within this group, nearly 1100 people were described as 'special agents, experts, and commissioners,' while another 300 or so were designated simply as 'scientific experts and investigators' (Lacey, 1993: pp. 127–170, Ross, 1991; Blumer et al., 1991; Keller, 1977; Skowronek, 1982; Wilson, 1909).

Many of the early organizers in the Smithsonian Institution played a critical role in the establishment of the American Social Science Association (ASSA) in 1865. The ASSA sought to serve as a clearinghouse of social knowledge in order to promote reform in education, public health jurisprudence, finance, and social economy. Directly influenced by the achievements of the Sanitary Commission during the Civil War and modeled on the British National Association for the Promotion of Social Science, the ASSA attracted commercial elite groups interested in promoting sound fiscal and monetary policy; lawyers, doctors, ministers, and teachers who sought professional affirmation; and middle-class reformers who combined humanitarian moral reform with the new scientific spirit. In this way, the ASSA, and other professional organizations that arose during this period, paved the way for a new liberalism, grounded in social empiricism (Haskell, 1977; Furner, 1975, 1993).

These reformers called for new measures of efficiency and economy in voluntary charity organizations and in public administration. Behind this concern with efficiency and economy lay a deep conviction that the polity was threatened by corruption of partisan political machines serving special economic interests. These political machines, run by local and state political bosses, consistently manipulated the mass of voters to maintain political power in flagrant disregard of any notion of public morality or efficiency and economy in government. Thus progressive reformers called for new social knowledge to help restore political order and representative government to American society. Beneath the catchphrase 'efficiency and economy' – the rallying cry of reformers – lay strong antimajoritarian values fused with antiparty perspectives (Critchlow, 1985; Alchon, 1985; Karl, 1974).

This movement for efficiency found early expression in late nineteenth-century charity endeavors. Amid concern that relief was being wasted on the undeserving poor, new professionals were brought into charity organization to centralize recipient roles, provide vouchers for relief funds to be distributed to recipients, institute periodic checks on recipients, and implement budget systems within agencies to keep a close watch on funds. Essential to this reform was the use of data gathering through statistical measures and new accounting procedures. Behind this movement lay a sentiment expressed by industrialist Andrew Carnegie, who declared, "One of the most serious obstacles to the improvement of our race is indiscriminate charity." Carnegie's infatuation with the new social sciences would lead him to create a number of new nonprofit philanthropic institutions concerned with the gathering of public knowledge (Lagemann, 1989).

At the vanguard of this movement for efficiency and economy in government were university trained and highly specialized social scientists. In the latter half of the nineteenth century, the basis of social science research was transformed from deductive reasoning derived from natural law into inductive procedure founded on historical analysis and empirical research. This new methodology allowed social scientists to mask any preoccupation with political matters in the terminology of scholarly objectivity and empirical research. The transformation of political impulses into the quest for public knowledge and administrative solutions reflected the aspirations of a new professional class anxious to gain influence in a highly industrialized society through expertise. They sought reform by placing 'the best people' – objective-minded experts – in the government (Ross, 1991; Haskell, 1977).

The call for efficiency and economy in government, and the need to develop new public agencies for the gathering of public knowledge essential to restoration of representative government, began on the municipal level first in New York with the creation of a Bureau of Municipal Research in 1905, organized by social scientists and civic-minded philanthropists and businessmen. Municipal research bureaus would be established in several other US cities. Central to these municipal research bureaus was the creation of executive budget systems that allowed a strong executive office, public officials, and the general electorate to track how public funds were spent. Essential to the functioning of an executive budget was the creation of an agency staffed by nonpartisan experts who would prepare and analyze budgetary finances and expenditures (Critchlow, 1985; Karl, 1974; Smith, 1991).

This demand for an executive budget system reached the national level under the administration of William H. Taft in 1910. When Taft assumed office in 1909, the federal budget stood at over a billion dollars and the federal debt had reached disturbing proportions, largely as an aftermath of the 'Panic of 1907.' In these circumstances, Taft was able to persuade Congress to appropriate $100 000 for a commission on efficiency and economy in government. To staff the commission, Taft called on the best social scientists in the country. After 2 years of painstaking work, the Taft Commission came out with its report, *The Need for a National Budget*. Yet, because the proposal was made in an election year, a recalcitrant Congress refused to enact an executive budget system. The failure of the Taft Commission did not dismay reformers, however. In 1915 they created the Institute for Government Research – the predecessor to the Brookings Institution. Only in 1921 during

the Harding administration would budget reform be enacted in the Budget and Accounting Act of 1921.

Taft's successor, Woodrow Wilson reflected in many respects the deep ambivalence that middle-class Americans felt about the state during the early twentieth century. An array of new government agencies was now involved in conservation, consumer protection, corporate regulation, trade practices, and monetary and financial policies. Yet Wilson, in his pioneering essay on public administration warned that the model of Prussian administration should be avoided in America. He declared that European bureaucratic practices would have to be adapted to fit the highly decentralized form of government in the USA. If the European model is to be employed in the USA, he told his fellow reformers, "We must Americanize it … We must learn our constitutions by heart and we must get the bureaucratic fever out of its veins" (quoted in Cuff, 1993: p. 390).

Until America's entry into the war in April 1917, neither the federal government's defense department nor civilian departments had considered mobilization tasks. As a consequence, private planning proceeded public planning. Financial and manufacturing firms such as J.P. Morgan and Du Pont Corporation, by receiving Allied war orders, were among the first to assess the financial, technical, and labor requirements of munitions production (Brownlee, 1993).

War, however, forced the enormous administrative expansion of the federal government in the short period between April 1917 and November 1919. New government agencies appeared in haphazard fashion, drawing corporate and academic personnel into government management. This involvement led many to the belief that the capacity for public knowledge and for effective social action rested in the mobilization of private, public-spirited professional and business groups willing to work in cooperation with a sympathetic state. The war demonstrated for these individuals how to forge a uniquely American combination of state–society relations.

Herbert Hoover sought to put this insight into practice as Secretary of Commerce in the Harding and Coolidge administrations, and later as president, by using the techniques of private–public cooperation he had learned as a wartime administrator. Hoover sought to develop state–voluntary cooperation in the collection of specialized data required for rational decision making in both the private and public sectors. For Hoover and others concerned with public knowledge, the war revealed just how essential government knowledge was for economic and social planning. Through the development of an administrative state he sought to develop new mechanisms for collecting public knowledge without the coercive element of state power (Hawley, 1974).

In keeping with his views about the credibility of privately financed researchers, Hoover arranged for the Rockefeller Foundation, founded in 1913, to fund a comprehensive survey of social changes in American life between 1900 and 1930. The 3-year study resulted in the massive 1600-page *Recent Social Trends in the United States: Report of the President's Research Committee of Social Trends* (1933). This study presented the full panoply of available knowledge for public purposes (Karl, 1974; Hawley, 1974; Alchon, 1985).

Although state capacity for gathering public knowledge expanded dramatically during the Great Depression era, World War II, and the postwar period, the federal government continued to rely indirectly on intermediary organizations in state and local government, and in civil society, for the accumulation of knowledge. The commensurate increase in the size of the federal government led to the creation of fluid networks of policy specialists, who operated in and between government and the private sector, identifying policy issues, generating 'solutions' to policy problems, and overseeing and evaluating the implementation and administration of policies within their specific areas of competence (Heclo, 1980).

Role of the Think Tanks

Essential to these policy networks are private policy research institutes or 'think tanks,' a peculiarly American social invention. By the 1990s there were an estimated 1200 nongovernment think tanks of various descriptions. World War II and the postwar continuation of a large defense program, as well as expansion of federal social programs, was a watershed in the evolution of the American think tanks.

At the end of World War II, Americans made a political decision to have government train, subsidize, and regularly consult a sizable population of civilian experts on such issues as defense, welfare, health, and economic stability. Of the major think tanks, only the Brookings Institution (1916) and the Carnegie Endowment for International Peace (1910) were founded before World War II. The American Enterprise Institute was founded during the war in 1943 (Critchlow, 1985; Smith, 1991; Lagemann, 1989; Dickson, 1971).

Although think tanks are ostensibly nonpartisan, in many instances they function as extensions of state power, their influence rising and falling with changes in governments and shifts in the ideological climate of the country. In other cases, think tanks function more independently, questioning and monitoring state strategies and structures. For example, the Rand Corporation of the 1940s and 1950s was designed to strengthen the programs of the US Air Force, whereas the Institute for Policy Studies has been a major center of criticism of US foreign policy, domestic social programs, and policies in the areas of civil rights and poverty.

Often the ideological orientation of these think tanks reflected changes in the historical context and political climate, as indicated by the dominant trends in high and popular culture, changes in voting behavior, and shifts in political leadership (Ricci, 1993).

The course of the Brookings Institution from the New Deal through the end of the twentieth century reflected these ideological currents. While Brookings studies in the early 1930s criticized New Deal programs, the Brookings Institution under its president Harold Moulton tried to avoid becoming directly involved in partisan politics. By 1937, however, the institution had emerged as a vociferous critic of New Deal fiscal policy.

In response to the new postwar environment and the reluctance of foundations to fund an institution they perceived as ineffective or out of touch, Robert Calkins, former head of the General Education Fund at the Rockefeller Foundation, agreed to head the institution. Calkins reorganized Brookings and recruited social scientists with liberal credentials and government experience. This new group had close ties with

government, and unlike the devotees of the earlier nonpartisan ideal, aligned themselves closely with presidential administrations. In 1965, when Calkins retired, Brookings was representative of mainstream Keynesian economic thinking, and its growing influence was reflected in renewed foundation support, especially from the Ford Foundation. Under Calkins' successor, Kermit Gordon, Brookings' reputation as a liberal democratic think tank was well entrenched. As a result, the institution became a major research center and a holding area for policy experts who were temporarily out of office. In these years, the Brookings Institution became a major center for policy innovation in welfare, health care, education, housing, and taxation policy (Critchlow, 1985; Smith, 1991).

Divisions over foreign and domestic policy opened cracks in the liberal consensus in the early 1970s, however, leading many observers to see the institution as isolated from large segments of business. In 1976, the board of trustees appointed Bruce MacLaury to head the institution. A former regional Federal Reserve banker and treasury official, MacLaury successfully courted business support, increased corporate representation on the board of trustees, and moved the institution toward a more moderate ideological stance.

By the 1970s, the Brookings Institution confronted competition from other major policy research institutions, especially the American Enterprise Institute and the Heritage Foundation – both viewed as conservative research institutions close to the Republican Party.

The American Enterprise Institute (AEI), which was founded in 1943 as the American Enterprise Association (AEA), illustrates the experience of a conservatively oriented research institution that expressed deep ambivalence about the post-World War II policy consensus (Critchlow, 1985; Smith, 1991). The key figure behind the establishment of the AEA was Lewis Brown, chairman of Johns-Manville Corporation. Brown realized that the New Deal had changed the political environment for business and, although he expressed anxiety about the level of government intervention, he accepted much of the New Deal program. From the start, the AEA reflected a conservative bias.

In 1954, A.D. Marshall, head of General Electric, assumed the institution's presidency and immediately hired William Baroody S. and W. Glenn Campbell, both staff economists at the US Chamber of Commerce, to head the research program. Under their guidance, AEA was gradually built into a modern research institute under its new name, the AEI. Principal support came from private foundations and major corporate sponsors. The institution's reputation was enhanced when the Nixon administration called upon a number of AEI associates to fill government positions. The AEI also emerged as a successful proponent of economic deregulation.

In 1977 Baroody retired and his son, William Baroody Jr., took over the presidency of the institution. To improve its standing in the academic community, the AEI assembled an impressive staff including Melvin Laird, William Simon, Robert Bork, Michael Novak, and Herbert Stein. The tenure of William Baroody Jr., however, ended abruptly in the summer of 1987, when an increasingly restive board of trustees forced his resignation because of cost overruns and declining revenues. Moreover, despite the large number of AEI staffers who held appointments in the Reagan administration (30 left AEI to join the administration in its first 2 years alone), members of the New Right, suspicious of what they considered the centrist orientation of the institution, voiced bitter criticism of the AEI.

Baroody's successor, Christopher DeMuth, bolstered the conservative orientation of the institute by bringing on board several former Reagan administration officials with strong rightist reputations.

The founding of the Heritage Foundation in 1973 revealed a new ideological climate in the analysis of public knowledge (Edwards, 1997; Peschek, 1987). Founded by Edwin Feulner and Paul Weyrich to provide rapid and succinct legislative analysis on issues pending before Congress, the Heritage Foundation sought to promote conservative values and demonstrate the need for a free market and a strong defense. The Heritage Foundation's articulation of conservative values in social policy, education, and government activities placed it at the forefront of New Right activity. Although sympathetic to the Reagan wing of the Republican Party, Heritage's commitment to conservative economic theories and social values allowed the institution to be critical of both republicans and democrats. The Heritage Foundation remained relatively small in its early years, but the election of Ronald Reagan to the presidency in 1980 enhanced the institution's prestige. By the mid-1980s the Heritage Foundation had established a solid place in the Washington world of think tanks as a well-organized, efficient, and well-financed research organization that called for the turning over of many government tasks to private enterprise, a strong defense, and a cautious approach to Russia and China.

During these years, myriad other think tanks emerged in Washington representing a range of ideological positions and specialized policy interests, including the left-oriented Institute for Policy Studies (1963) and the libertarian-oriented CATO Institute (1977). Think tanks concerned with national security included the Center for Strategic and International Studies (1962) and the Center for National Security Studies (1962) affiliated with the American Civil Liberties Union (Friedman, 1983).

These private institutions served to supplement social research conducted by federal agencies and institutions. By the 1990s the number of personal and committee staff of Congress exceeded 11 000, with a further 20 000 people responding to the information needs of the legislature from the Congressional Budget Office, the Office of Technological Assessment, and the General Accounting Office (Lacey, 1993). The executive branch commands far greater numbers. By the 1990s the scale of spending by the federal government for the acquisition and application of public knowledge was massive. Funding is currently routed through an estimated 180 bureaus and agencies. This research includes a mixture of 'in-house' research conducted by career civil servants, and a larger body of research conducted by outside parties via grants and contracts on a 'for-profit' basis and some of it by university based scholars on a 'not-for-profit' basis. An estimated one-third of this research activity is supported by the federal government to meet its needs, while the remaining two-thirds is authorized for state and local government, school systems, policy forces, welfare agencies, and industrial and agricultural groups. In addition, a vast array of trade and professional organizations and 'public interest' groups conduct research in the areas of environmental policy, consumer affairs, civil rights, and foreign policy issues (Critchlow, 1985). While the government relied on think tanks

for policy innovation, program review, and policy assessment, the federal government relied on nonprofit service providers to implement social programs. In this way, the expansion of the welfare state in the 1960s and 1970s created a seeming paradox that remains apparent in the contemporary regime: the rapid growth of the nonprofit, voluntaristic sector occurred at the same time that federal social policy became increasingly centralized and active. The parallel development of the centralized welfare state and the voluntary sector was no coincidence. The growth of federal social programs mobilized the nonprofit sector, while in turn the nonprofit sector, through policy innovation and program implementation, helped legitimize the modern welfare state.

Thus, as the traditional federal system that had existed since the founding of the nation eroded in the 1960s, when the federal government extended its powers at the expense of state power, the nonprofit sector was called upon to help design, implement, and administer social programs that had once been within the jurisdiction of the public sector (Hall, 1992; Kameran and Kahn, 1989; Salamon, 1992). As a consequence, the nonprofit sector expanded in the 1960s at the very time that the federal government enlarged its powers over the states. The result produced an unusual phenomenon – government centralization produced greater nonprofit sector voluntarism (Critchlow, 1999).

The 'Mixed' Welfare State

Between 1965 and 1980 the American social welfare state underwent a significant transformation, with government spending on social welfare increasing by 263% in inflation-adjusted dollars (Gilbert, 1983; Critchlow, 1999). The expansion of the welfare state promoted the emergence of a government–nonprofit partnership. Increasingly in the 1960s, the federal government generated the funds but then turned over the delivery of services to other public state and local organizations, and private nonprofit organizations. The private, nonprofit sector became a major beneficiary of this system of services. By the 1980s, the federal government alone provided $40 billion in support to the private nonprofit sector, while state and local governments provided additional amounts from their own resources (Salamon, 1992).

Federal demand enabled the nonprofit sector to grow by leaps and bounds. In 1940 there were only 12 500 charitable, tax exempt organizations. By 1990, there were over 700 000, with most of this growth taking place after the 1960s (Salamon, 1992; Hall, 1992; Critchlow, 1999). While the federal government centralized social policy through congressional mandates, administrative and legislative regulations, and new funding arrangements, the delivery and implementation of these programs were placed in the hands of state and local governments and nonprofit organizations.

The growth of philanthropic foundations in the twentieth century, especially in more recent years has contributed significantly to role of the increase in policy knowledge. Increasingly philanthropic foundations such as the Rockefeller Foundation, the Ford Foundation, and the Gates Foundation operate in an increasingly complex environment reflecting the diversification of the federal government's role in public policy. For example, the Low Income Housing Tax Credit Act of 1986 has encouraged public funds to match grants from the Ford Foundation through the Local Initiatives Support Corporation and Enterprise Community Partners. In supporting these activities, Ford Foundation urban policy experts have undertaken studies and work on local community low-income housing problems. Similarly, the nonprofit Trust for Public Land has worked with private individuals and for-profit corporations to preserve land. The Davide and Lucile Packard Foundation spent over a $100 million since 1998 for the Conservating California Landscape Initiative to present 343 000 acres. The rapid growth of community foundations have been used to support expert studies and policy initiatives in the areas of conservation, tax policy, economic development, education at all levels, and cultural and arts projects. The well-funded Gates Foundation, for example, is extensively involved in social and health policy studies and projects in the United States and the world. While nonprofit foundations after 9/11 experienced strict regulation to ensure that funds were not being directed to groups supporting terrorism, nonprofits have accelerated their contributions to the increase in policy knowledge.

The emergence of the 'second welfare state' in the last three decades of the twentieth century – contrary to the concerns of those who feared that government would replace the nonprofit sector altogether – created a vibrant nonprofit sector. Government, in turn, came to rely on this sector to design and implement programs, undertake demonstration projects, contract with other private agencies, and provide outside consultation and independent review agencies and activities.

As this second welfare state emerged, the nonprofit sector was called upon to fulfill a number of distinct functions. Nonprofit organizations served as *funding intermediaries* distributing funds to other parts of the nonprofit sector, *service-providing organizations* involved in health care, education, counseling, adoption assistance, and *political action agencies* supporting particular pieces of legislation or candidates for political office (Salamon, 1992). By providing public services, these nonprofit organizations developed a symbiotic relationship with the federal government. By 1990, government grants, contracts, and reimbursements accounted for 31% of nonprofit service organizations' income. Just 18% of the total income that nonprofit organizations receive comes from private giving (Salamon, 1992; Hall, 1992).

The proliferation of government and nonprofit institutions involved in public policy marshals expert against expert, testing the credibility of each and enriching the information environment in which policy discourse occurs and decisions are taken. Indeed, on highly technical issues, opposing experts often directly contradict one another. Yet the growth of public and private agencies involved in collecting public knowledge allows for a more democratized policy process – much closer, in a sense, to the Jeffersonian–Jackson vision of an open, competitive democracy.

Still, two qualifications are in order: neither Jefferson nor Jackson envisioned a world of competing experts representing various interests and ideological positions; and secondly, this contemporary world of competing institutional research is not equally open to all comers. Access to financing is one limitation, and access to those in power is another.

The proliferation of expertise and the expansion of government knowledge that has arisen in response to the demand for

more policy inquiry have not necessarily improved the information base on which political decisions are made. The emergence of more complicated patterns of political mobilization by a larger and more diverse array of interests and groups has necessarily politicized the nature of policy knowledge and thereby redefined the nonpartisan ideal espoused at the beginning of the twentieth century. Competition for attention to research places heavy demands on the time and attention of legislators and policy-makers who, in a value-laden universe, are left, ultimately, to decide what information counts for them on the basis of their own values. A widespread acceptance of relativism, embodied institutionally in public and private research organizations, has irrevocably altered the process of social investigation and the uses of social knowledge.

See also: Intellectuals, Sociology of; Nonprofit Sector, Geography of; Policy Knowledge: Universities; Think Tanks.

Bibliography

Alchon, G., 1985. The Invisible Hand of Planning: Capitalism, Social Science, and the State in the 1920s. Princeton University Press, Princeton, NJ.

Anheiirer, H.K., David, C.H. (Eds.), 2010. American Foundations: Roles and Contributions. Brookings Institution Press, Washington, DC.

Bensel, R.F., 1990. Yankee Leviathan: The Origins of Central State Authority in America, 1859–1877. Cambridge University Press, Cambridge, UK.

Blumer, M., Bales, K., Kish Sklar, K. (Eds.), 1991. The Social Survey in Historical Perspective. Cambridge University Press, Cambridge, pp. 1889–1940.

Brewer, J., 1990. The Sinews of Power: War, Money and the English State, 1688–1783. KNOPF, Cambridge, MA.

Brock, W.R., 1984. Investigation and Responsibility: Public Responsibility in the United States, 1865–1900. Cambridge University Press, Cambridge, MA.

Brownlee, W.E., 1993. Social investigations and political learning in the financing of world war II. In: Lacey, M.J., Furner, M.O. (Eds.), The State and Social Investigation in Britain and the United States. Cambridge University Press, Cambridge, pp. 323–361.

Cassedy, J.H., 1969. Demography in Early America: Demography in Early America: Beginnings of the Statistical Mind. Harvard University Press, Cambridge, MA.

Cohen, P.C., 1982. A Calculating People: The Spread of Numeracy in Early America. Chicago University Press, Chicago.

Critchlow, D.T., 1985. The Brookings Institution, 1916–1952: Expertise and the Public Interest in a Democratic Society. Northern Illinois University Press, DeKalb, IL.

Critchlow, D.T., 1999. Intended Consequences: Birth Control, Abortion and the Federal Government. Oxford University Press, New York.

Cuff, R.D., 1993. War mobilization, institutional learning, and state building in the United States, 1917–1941. In: Lacey, M.J., Furner, M.O. (Eds.), The State and Social Investigation in Britain and the United States. Cambridge University Press, Cambridge, pp. 388–427.

Daniels, G.H., 1968. American Science in the Age of Jackson. Columbia University Press, New York.

Dickson, P., 1971. Think Tanks. New York.

Dupree, Hunter, A., 1957. Science in the Federal Government. Harvard University Press, Cambridge, MA.

Edwards, L., 1997. The Power of Idea: The Heritage Foundation at 25 Years. Jameson Books, Ottawa, IL.

Friedman, J. (Ed.), 1983. The First Harvest: An Institute for Policy Studies Reader. Grove Press, New York, pp. 1963–1983.

Furner, M.O., 1975. Advocacy and Objectivity: A Crisis in the Professionalization of American Social Science, 1865–1905. University of Kentucky Press, Lexington, KY.

Furner, M.O., 1993. The republican tradition and the new liberalism: social investigation, state building, and social learning in the gilded age. In: Lacey, M.J., Furner, M.O. (Eds.), The State and Social Investigation in Britain and the United State. Cambridge University Press, Cambridge, pp. 171–242.

Furner, M.O., Lacey, M.J., 1993. Social investigation, social knowledge and the state: an introduction. In: Lacey, M.J., Furner, M.O. (Eds.), The State and Social Investigation in Britain and the United States. Cambridge University Press, Cambridge.

Furner, M.O., Supple, B. (Eds.), 1990. The State and Economic Knowledge: The American and British Experiences. Cambridge University Press, Cambridge.

Gilbert, N., 1983. Capitalism and the Welfare State: Dilemmas of Social Benevolence. Yale University Press, New Haven, CT.

Hall, P.D., 1992. Inventing the Nonprofit Sector and Other Essays on Philanthropy, Voluntarism, and Nonprofit Organizations. Johns Hopkins University Press, Baltimore.

Haskell, T., 1977. The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth Century Crisis of Authority. Illinois University Press, Urbana.

Hawley, E., June 1974. Herbert Hoover, the commerce secretariat, and the vision of an associative state, 1921–1929. Journal of American History 61.

Heclo, H., 1980. Issue networks and the executive establishment. In: King, A. (Ed.), The New American Political System. American Enterprise Institute Press, Washington, DC.

Kameran, S.B., Kahn, A.J., 1989. Privatization and the Welfare State. Princeton University Press, Princeton, NJ.

Karl, B., 1974. Charles Merriam and the Study of Politics. University of Chicago Press, Chicago.

Karl, B., 1983. The Uneasy State: The United States from 1915 to 1945. University of Chicago Press, Chicago.

Keller, M., 1977. Affairs of State: Public Life in Late Nineteenth Century America. Harvard University Press, Cambridge, MA.

Lacey, M.J., 1993. The world of the bureaus: government and the positivist project in the late nineteenth century. In: Lacey, M.J., Furner, M.O. (Eds.), The State and Social Investigation in Britain and the United States. Cambridge University Press, Cambridge.

Lagemann, E.C., 1989. The Politics of Knowledge: The Carnegie Corporation, Philanthropy, and Public Policy. University of Chicago Press.

Leach, W., 1989. True Love and Perfect Union: The Feminist Reform of Sex and Society. Basic Books, New York.

Marshall, H.E., 1937. Dorothea Dix: Forgotten Samaritan. University of North Carolina Press, Chapel Hill, NC.

National Academy of Sciences, 1978. The Federal Investment in Knowledge of Social Problems. Washington, DC.

Nelson, W., 1992. The Roots of American Bureaucracy, 1830–1900. Cambridge.

Peschek, J.G., 1987. Policy-planning Organizations: Elite Agendas and America's Rightward Turn. Temple University Press, Philadelphia.

Ricci, D., 1993. The Transformation of American Politics: The New Washington and the Rise of Think Tanks. New Haven, CT.

Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, Cambridge.

Salamon, L.M., 1992. America's Nonprofit Sector: A Primer. Foundation Center, New York.

Skowronek, S., 1982. Building a New American State: The Expansion of National Administrative Capacities, 1877–1920. Cambridge University Press, Cambridge.

Smith, J.A., 1991. The Idea Brokers: Think Tanks and the New Policy Elite. The Free Press, New York.

Wilson, W., 1909. The State: Elements of Historical and Practical Politics, rev. ed. D C Health and Company, Boston.

Politeness and Language

Penelope Brown, Max Planck Institute of Psycholinguistics, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article assesses the advantages and limitations of three different approaches to the analysis of politeness in language: politeness as social rules, politeness as adherence to an expanded set of Gricean Maxims, and politeness as strategic attention to 'face.' It argues that only the last can account for the observable commonalities in polite expressions across diverse languages and cultures, and positions the analysis of politeness as strategic attention to face in the modern context of attention to the evolutionary origins and nature of human cooperation.

What Is Politeness?

If, as many have claimed, language is the trait that most radically distinguishes *Homo sapiens* from other species, politeness is the feature of language use that most clearly reveals the nature of human sociality as expressed in speech. Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors' social status and their social relationship. Politeness – in this broad sense of speech oriented to an interactor's public persona or 'face' – is ubiquitous in language use. Since, on the whole, taking account of people's feelings means saying and doing things in a less straightforward or more elaborate manner than when one is not taking such feelings into consideration, ways of being polite provide probably the most pervasive source of indirectness, reasons for not saying exactly what one means, in how people frame their communicative intentions in formulating their utterances.

There are two quite different kinds of feelings to be attended to, and therefore there are two distinct kinds of politeness. One kind arises whenever what is about to be said may be unwelcome, prompting expressions of respect, restraint, avoidance ('negative politeness'). Another arises from the fact that long-term relationships with people can be important in taking their feelings into account, prompting expressions of social closeness, caring, and approval ('positive politeness'). There are many folk notions for these kinds of attention to feelings – including courtesy, tact, deference, demeanor, sensibility, poise, discernment, rapport, mannerliness, urbanity, as well as for the contrasting behavior – rudeness, gaucheness, social gaffes, and for their consequences – embarrassment, humiliation. Such terms label culture-specific notions invested with social importance, and attest both to the pervasiveness of notions of politeness and to their cultural framing.

How Can We Account for Politeness?

Since politeness is crucial to the construction and maintenance of social relationships, politeness in communication goes to the very heart of social life and interaction; indeed it is probably a precondition for human cooperation in general. Politeness phenomena have therefore commanded interest from theorists in a wide range of social sciences. Three main classes of theoretical approach to the analysis of politeness in language can be distinguished.

1. *Politeness as social rules*. To the layman, politeness is a concept designating 'proper' social conduct, rules for speech and behavior stemming generally from high-status individuals or groups. In literate societies such rules are often formulated in etiquette books. These 'emic' (culture-specific) notions range from polite formulae like *please* and *thank you*, the forms of greetings and farewells, etc., to more elaborate routines for table manners, deportment in public, or the protocol for formal events. Politeness is conventionally attached to certain linguistic forms and formulaic expressions, which may be very different in different languages and cultures. This is how the 'person on the street' tends to think about politeness, as inhering in particular forms of words.

Some analytical approaches to politeness are formulated in terms of the same sorts of culture-specific rules for doing what is socially acceptable, for example, the work by Ide and others on Japanese politeness as social indexing or 'discernment' (see Watts et al., 1992). In these approaches, politeness is a matter of social norms, and inheres in particular linguistic forms when used appropriately as markers of pre-given social categories. This approach is most appropriate for fixed aspects of language use – the more or less obligatory social marking of relatively unalterable social categories.

2. *Politeness as adherence to Politeness Maxims*. An alternative approach takes the position that the linguistic forms of politeness are not a matter of arbitrary convention but are motivated by general principles. During the 1970s this perspective was formulated in linguistic pragmatics in Gricean terms, with politeness seen as a set of social conventions coordinate with Grice's Cooperative Principle for maximally efficient information transmission ('Make your contribution such as required by the purposes of the conversation at the moment'), with its four 'Maxims' of Quality, Quantity, Relevance, and Manner (*see* Pragmatics, Linguistic). Lakoff (1973) suggested that three 'rules of rapport' underlie the choice of linguistic expression, rules which can account for how speakers deviate from directly expressing meanings. Choice among these three pragmatic rules ('Don't impose,' 'Give options,' 'Be friendly') gives rise to distinct communicative styles. Leech's more detailed

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53072-4

proposal (1983) is in the same vein. Complementary to Grice's Cooperative Principle, Leech postulates a Politeness Principle – 'Minimize the expression of impolite beliefs,' with the six Maxims of Tact, Generosity, Approbation, Modesty, Agreement, and Sympathy. As with Grice's Maxims, deviations from what is expected give rise to inferences. Cross-cultural differences derive from the different importance attached to particular maxims.

The conversational maxim approach shares with the social norm approach the emphasis on codified social rules for minimizing friction between interactors, and the view that deviations from expected levels or forms of politeness carry a message.

3. *Politeness as face management*. A more sociological perspective places 'face work' at the core of politeness. Goffman (1967) considered politeness as an aspect of interpersonal ritual, central to public order. He defined face as an individual's publicly manifest self-esteem, and proposed that social members have two kinds of face requirements: positive face, or the want for approval from others, and negative face, or the want not to offend others. Attention to these face requirements is a matter of orientation to Goffman's "diplomatic fiction of the virtual offense, or worst possible reading" (Goffman, 1971: 138ff), the working assumption that face is always potentially at risk, so that any interactional act with a social–relational dimension is inherently face-threatening and needs to be modified by appropriate forms of politeness.

Universal Principles of Politeness

Brown and Levinson (1987, henceforth B&L) introduced a new perspective by drawing attention to the detailed parallels in the construction of polite utterances across widely differing languages and cultures, and arguing that universal principles underlie the construction of polite utterances. The parallels they noted are of two sorts: how polite utterances are linguistically constructed, and how the polite expression of utterances is modified in relation to social characteristics of the interlocutors and the situation. At least three social factors are involved in deciding how to be polite: (1) one tends to be more polite to social superiors; (2) one tends to be more polite to people one doesn't know. In the first case, politeness tends to go one way upwards (the superior is less polite to an inferior); in the second, politeness tends to be symmetrically exchanged. In addition, (3) in any culture there are norms and values affecting the degree of imposition or unwelcomeness of an utterance, and one tends to be more polite for more serious impositions. In language there are also detailed parallels, with the linguistic structures for realizing particular kinds of politeness showing remarkable similarities across unrelated languages. The politeness of solidarity is characterized, for example, by use of intensifiers, in-group identity markers and address forms, exaggerated intonation patterns, and forms for emphasizing agreement and avoiding disagreement. Avoidance-based politeness is characterized by forms expressing self-effacement, formality, restraint, deference, with the use of honorifics, hedges, indirect speech acts, and impersonalizing mechanisms like pluralization of pronouns, nominalization, and passive. All of these involve elaborating the expression of communicative intent in specific ways that deviate from direct Gricean communication.

Why then are these kinds of detailed parallels across languages and cultures to be found in the minutiae of linguistic expression in socially analogous contexts? Explanations in terms of arbitrary social norms or rules can account for politeness in a particular social group, but not for the cross-cultural patterns, which seem to require a strategic account in terms of what people generally are trying to do when they are being polite. B&L proposed that such deviations from Gricean efficient communication are motivated by the universal human assumption that interactors have 'face,' in the form of desires for the interactional acknowledgment of their social persona. Following Goffman, they posited two kinds of face: positive face desires for affirmation and social closeness, and negative face desires for separation, personal territory, and freedom from unexpectable impositions. These give rise to two styles of politeness – one addressed to ratifying the personality of the interlocutor (positive or solidarity politeness), and the other designed to mitigate or minimize impositions (negative or respect politeness).

B&L (1987) proposed an abstract model of politeness wherein human actors are endowed with two essential attributes: *face* and *rationality*. Face consists of two specific kinds of wants: positive face (i.e., the desire to be approved of, admired, liked, validated), and negative face (the desire to be unimposed upon, unimpeded in one's actions). The second ingredient in the model – rationality – provides for the ability to reason from communicative goals to linguistic means that would achieve these goals. From these two assumptions – face and rationality – and the assumption that speakers mutually know that all speakers have these two attributes, B&L developed a model of how speakers construct – and interpret – polite utterances in different contexts on the basis of assessments of three social factors: the relative power (P) of speaker and addressee, their social distance (D), and the intrinsic ranking (R) of the face-threateningness of an imposition. P, D, and R are seen as abstract social dimensions indexing kinds of social relationship (P and D) and cultural values and definitions of impositions or threats to face (R).

B&L (1987) argued that the actions people do when talking with one another – for example, requesting, informing, offering, complaining – have implications for each others' face, and they identified a wide array of speech act types that have the potential to threaten face: 'face-threatening acts' or FTAs. They distinguished five general types of strategies of politeness, ranging from avoiding an FTA altogether, to carrying it out but 'off record' (indirectly). On-record realization of an FTA can be done without any redressive action at all ('baldly'). It may be carried out with positive redress, which is essentially approach based, addressing the hearer's positive face wants by emphasizing closeness and solidarity. Politeness may also be carried out with negative redress, which is essentially avoidance based, addressing negative face wants for distance, deference, and freedom from impositions. Speakers are assumed to choose the linguistic framing of their utterance from this set of strategic possibilities according to the weightiness of the FTA, which is assessed with reference to the three contextually dependent social factors P, D, and R. For low levels of FTA threat, bald-on-record or positive politeness is most

appropriate and cost effective; for higher levels, negative politeness is required; for the highest threats, indirectness is the safe option.

The claim, then, is that, however culturally variable the kinds of social relationship and kinds of face threat might be, underlying them are human attributes (face, means-ends reasoning abilities) and pan-cultural social dimensions (relative power, social distance, ranking of face-threateningness) which universally go into the reckoning and the interpretation of strategic language choice, and hence one can derive the cross-cultural similarities in choice of linguistic realizations of politeness strategies that empirically seem to be in evidence. B&L (1987) claimed further that this model of politeness universals could be applied in particular cultural settings as an ethnographic tool for analyzing the quality of social relationships. Stable social relationships are characterized in part by stable patterns of language use, which may distinguish particular societies or social groups. Analyzing these stable patterns (for example, the distribution of T/V pronoun usage across individuals and social groups) can reveal important aspects of social relationships and their changes over time.

Critiques of the B&L Politeness Model

The goal of B&L to formulate a cross-culturally applicable 'etic' set of concepts in terms of which politeness can be analyzed in 'emic' terms for any particular society was ambitious. Etic analytical concepts are drawn from a universal set, defined from outside a particular cultural perspective and used to compare behavioral or linguistic systems across different cultural groups, in contrast to emic ones which are meaningful within a specific culture and are used to describe a system in its own terms. Critics of the B&L model reveal several major points of contention about whether it is legitimate at all to generalize across cultural systems and, if so, how a theory of politeness should be formulated.

The Universality of Face

Many critics have challenged B&L's formulation (via Goffman and Radcliffe-Brown) of positive and negative face wants as a valid way of conceptualizing the universal underpinnings of politeness. Negative face, in particular, considered as wants for freedom from imposition, appears entirely too embedded in Western individualism to sit well with conceptions of face in some other (e.g., east Asian) cultures. In part, this is due to a misconstrual: in the B&L model, face wants are abstract, they do not necessarily correspond clearly to conscious emic notions. What B&L (1987) claimed is that *underlying* diverse folk notions is a core of two interactionally relevant wants (for ratification, and freedom from imposition) which seem to be cross-culturally applicable, as desires concerning one's public self-image in the context of the moment which are *assumptions oriented to in interaction*. Other theorists (e.g., O'Driscoll, 1996; Arundale, 1999) have argued for notions of positive/negative face that are even more abstract, in terms of merging/individuation or closeness/separation, as the universal heart of politeness.

Challenges to the universality of the model also extend to the proposed hierarchy of increasing politeness (from bald-on-record to positive to negative to indirectness). Assessments of the P, D, and R factors are situationally and culturally very variable; it is possible to accumulate different strategies in one utterance and to balance elements of negative politeness with positive politeness in one act, and indirectness is not always seen as the most polite option. These observations have led some researchers to argue against the possibility of identifying any kind of universal basis for polite behavior; politeness is simply incommensurate across societies. Those who take this extreme relativistic line can have no explanation for the observable cross-cultural parallels in patterns of language use, for how people manage (sometimes) to understand others from culturally different backgrounds, or for cross-linguistic parallels in the diachronic sources of honorifics from politeness strategies.

Politeness as Communicated or Taken for Granted

In contrast with rule-based approaches, B&L (1987) insist that politeness inheres not in words or in sentences per se, but in utterances uttered in a context, by virtue of the successful communication of a polite attitude or intention. Polite utterances are not necessarily communicating 'real' feelings about anothers' social persona, but expressing contextually expected concern for face. This concern is an 'implicature,' an inference of polite intentions, not a feature inextricably attached to particular linguistic forms. Politeness is ascribed to a speech act, or to an interactional move (if you prefer), not to a strategy or its linguistic realization per se.

In other approaches (for example, Fraser's (1990) 'conversational contract,' Watts et al.'s (1992) 'politic behavior'), politeness is taken to be the expected *background* to interaction; it is normally not communicated but consists in following expectations as to appropriate behavior.

Broad versus Narrow Scopes for Politeness Theory

A narrower view takes politeness to be strategic orientation to potential face threats; there are some situations (e.g., task-oriented ones) where politeness may be subsumed to other goals, and there are many reasons for being indirect in speech other than politeness (e.g., humor, irony, rhetorical force). Many motivations other than politeness guide human behavior. A more inclusive view sees politeness as orientation to the social-relationship dimension of every interaction, with attention to face taken to be an omnipresent necessity. The whole continuum from extreme politeness through a quite neutral level of politeness (maintaining the status quo, 'discernment') to rudeness (outright intentional threat) then needs to be brought into the theory.

Politeness from the Point of View of the Individual, the Dyad, or the Social Group

The B&L model takes the interacting dyad as its unit; it is about how interlocutors make inferences of politeness from one another's deviations from Gricean efficient communication, and how stable patterns of strategies characterize dyadic interaction, providing an index to the quality of the social relationship. Many politeness strategies are quintessential examples of 'intersubjective perspective-taking' – putting yourself in the others' position. A major goal of B&L (1987) was to insist on the centrality of social interaction as a significant level of social life, intermediate between the individual and society, where social/cultural facts (status, role, values, norms, rights, and obligations) are integrated with individual ones (goals, plans, and strategies).

Yet the Gricean foundation of the theory and the speech-act-based formulation of the B&L strategies have made many see the model as purely psychological (how a speaker calculates how to frame an utterance). Arundale (1999), for example, argues instead for a theory of how face is *jointly constituted* in ongoing interaction. In fact, we need both perspectives: face is indisputably interactionally created, negotiated, and manipulated. Nevertheless, it can be considered from the point of view of the individual speaker or hearer (as B&L do in their production/comprehension model), or of the society or social group (as 'face constituting theory' and most sociolinguists do).

The Indeterminacy Problem

B&L (1987) proposed their model as a concrete discovery method for analyzing social relationships in any society by studying members' speaking practices. However, the usefulness of this model as an ethnographic tool is undermined by a number of complexities that arise when examining naturally occurring language usage data. The problems of mixing strategies in a single utterance and ordering of strategies across utterances are compounded by the indeterminacy of context-dependent P, D, and R assessments and by difficulties of establishing whether an utterance is intended to be taken as ironic, joking, or teasing. All of these make it hard to code levels of politeness in any concrete situation and to assess how different expressions interact to produce a composite message in an utterance. These are problems with any theory in terms of actors' intentions when applied to empirical data; as both interactors and conversation analysts know, it is not always possible to be certain what interlocutors' intentions are at a particular point in natural interaction. The B&L model of politeness as originally formulated clearly needs elaboration and revision.

New Directions in Research on Politeness

The intellectual climate of research has changed radically in the four decades since the B&L model was fostered at the University of California, Berkeley, in the early 1970s. Since then an enormous amount of research has addressed politeness from a variety of perspectives, in work coming from all the social sciences that have an interest in social interaction. Empirical studies of politeness have greatly increased the amount of information we have about social interactional styles in different contexts and different societies. Yet the research emphasis in work on politeness has been largely on cross-cultural differences in politeness, with little attention addressed to the cross-linguistic and cross-cultural parallels which motivated the theory. Indeed, very few studies provide the crucial kinds of evidence necessary to test the universality of any theory of politeness, namely, for a range of particular societies, an 'ethnography of speaking' providing evidence of naturally occurring language usage across a wide range of different contexts, showing how politeness is interactionally modulated in relation to local social factors in that society.

Researchers, impressed by different cultural views of face, propriety, conviviality, and the individual, have generally preferred to study culture-specific patterns of language use. Yet the significance of politeness lies far beyond the culture-specific rules of appropriate behavior and speech that seem so salient to members of the culture. This wider significance lies in the fact that by regular patterns of language choice humans interactively construct their social relationships. Research in this area therefore needs to be based in a theory of social interaction that takes into account both our common human nature and ability to communicate cross-culturally, and the cultural differences which can sometimes lead us to misunderstand one another. Despite its shortcomings, the B&L model retains its hold on research on politeness largely because it provides a coherent set of concepts for analytically dissecting polite speech in different societies and contexts. It is also the only theory that begins to provide an account for the detailed cross-cultural parallels in polite speech.

In the new millenium, research agendas and priorities have changed. The study of social interaction has turned to focus on the interactional foundations of language, its cognitive underpinnings, and its instantiation in the brain. Research continues on the sociocultural contexts of language use, with a major emphasis on face as co-constructed in interaction and on the sequential development of this interactional process, with impoliteness as well as politeness as the object of enquiry. A corresponding emphasis on the sequentially developing contexts for face and related politeness issues (preference structures, entitlement, 'ownership' of knowledge, 'fairness') appears in the field of Conversation Analysis, for example, in Pomerantz and Heritage (2013), Lerner (1996), Heinemann (2006), Curl and Drew (2008), Heritage and Raymond (2005), and Heritage (2013). There is also an increasing trend toward the comparative study of naturally occurring interaction from a conversation analytic perspective (Sidnell, 2009), as well as new theoretical approaches to the study of cultural models (e.g., Strauss, 2004) and of human social relations – for example, Agha's (2007) view of social relations as created through language use shaped by reflexive models of social life. All of these developments can be expected to feed into our understanding of politeness.

Another new direction is the emerging emphasis on the cooperative basis of human sociality and on the interactional underpinnings to language, its evolution, and its ethological base; Goody (1995), Enfield and Levinson (2006), Tomasello (2008, 2009), and Stivers et al. (2009) can be mentioned here. There are recent investigations into the cognitive and brain underpinnings to interaction (e.g., De Ruiter et al., 2010; Noordzij et al., 2009, 2010), work which forms a new context within which the search for universal bases for politeness can be pursued, exploring the cognitive prerequisites for politeness, such as recursive theory of mind, without which the elaborate demonstration of mutual regard would be impossible. This context suggests that the inhibition of aggression, the soothing character of politeness work, and the elaborate interpersonal ritual that politeness represents, may have played a crucial role in the evolution of the specifically human forms of social life and culture, and, indeed, may have been a fundamental factor driving the evolution of human cognition.

Conclusions

Politeness has attracted an enormous amount of research attention since the 1970s, and continues to be a major focus for research in disciplines concerned with social interaction.

The study of politeness phenomena can provide insight into widely differing issues; foci of interest correspondingly differ widely. They include, for example, analyses of the sequential development of politeness in natural interaction, the study of cross-cultural pragmatics and misunderstanding, the ethnography of speaking of face and politeness in different cultures and contexts, politeness strategies as providing the stylistic coherence of particular types of interaction (e.g., gender differences in speech style), politeness as a functional motivation for linguistic structure (e.g., honorifics), the social psychology of face management and interpersonal perception, applications of politeness theory to the analysis of formal ritual and to a view of culture as 'rhetoric,' or forms of effective expression.

Research on politeness has, however, been much weaker on the theoretical front. A major limitation is the kinds of data used in analyses. A large proportion of studies take as their data people's conscious evaluations of politeness expressed in sentences, judgments which tend to be both prescriptive and stereotypical. Far fewer studies use as data recordings of situated conversational exchanges to explore how politeness is achieved sequentially in naturally occurring discourse, and only a handful provide the crucial kind of evidence necessary to test the universality of any theory of politeness: for a particular society, an 'ethnography of speaking' providing evidence across different contexts to show how politeness is modulated in relation to social factors (P, D, R, and others) in that society.

The emphasis in research has been largely on cross-cultural *differences*, with insufficient attention addressed to the cross-linguistic/cross-cultural parallels which tend to be taken for granted when they are not disputed. But politeness has a significance far beyond the P's and Q's of appropriate behavior and speech. The wider significance of politeness is in the interactional, communicative, day-to-day basis of social life and the conduct of social relationships. Research on politeness needs to be anchored in a theory of social interaction that takes account *both* of our common human nature and our ability to communicate cross-culturally and of cultural differences which make us sometimes misunderstand one another.

See also: Communicative Competence: Linguistic Aspects; Cooperation: Sociological Aspects; Discourse, Anthropology of; Intercultural Communication: An Anthropological Perspective; Intergroup Relations; Linguistic Anthropology; Pragmatics, Linguistic; Prosocial Behavior and Empathy; Social Psychology; Status and Role, Social Psychology of.

Bibliography

Agha, A., 2007. Language and Social Relations. Cambridge University Press, Cambridge, UK.

Arundale, R., 1999. An alternative model and ideology of communication for an alternative to politeness theory. Pragmatics 9, 119–153.

Brown, P., Levinson, S.C., 1987[1978]. Politeness: Some Universals in Language Use. Cambridge University Press, Cambridge, UK.

Brown, R., Gilman, A., 1968[1960]. Pronouns of power and solidarity. In: Fishman, J. (Ed.), Readings in the Sociology of Language. Mouton, The Hague, Netherlands, pp. 252–276.

Coulmas, F. (Ed.), 1991. New Perspectives on Linguistic Etiquette. International Journal of the Sociology of Language Mouton de Gruyter, New York; Berlin. P40 .I58 no. 91–94.

Culpeper, J., 2011. Impoliteness: Using Language to Cause Offence. Cambridge University Press, Cambridge, UK.

Curl, T., Drew, P., 2008. Contingency and action: a comparison of two forms of requesting. Research on Language and Social Interaction 41 (2), 129–153.

De Ruiter, J.P., Noordzij, M.L., Newman-Norlund, S., Hagoort, P., Levinson, S.C., Toni, I., 2010. Exploring the cognitive infrastructure of communication. Interaction Studies 11, 51–77.

Dufon, M.A., Kasper, G., Takahashi, S., Yoshinaga, N., 1994. Bibliography on linguistic politeness. Journal of Pragmatics 21, 527–528.

Enfield, N.J., Levinson, S.C. (Eds.), 2006. Roots of Human Sociality. Cambridge University Press, Cambridge, UK.

Fraser, B., 1990. Perspectives on politeness. Journal of Pragmatics 14, 219–236.

Goffman, E., 1967. The nature of deference and demeanor. In: Goffman, E. (Ed.), Relations in Public: Microstudies of the Public Order. Harper and Row, New York.

Goody, E., 1995. Social Intelligence and Interaction. Cambridge University Press, Cambridge, UK.

Goffman, E., 1971. Relations in Public: Microstudies of the Public Order. Basic Books, New York.

Grice, H.P., 1975. The logic of conversation. In: Cole, P., et al. (Eds.), Syntax and Semantics 3: Speech Acts. Elsevier, pp. 41–58.

Heinemann, T., 2006. "Will you or can't you?" Displaying entitlement in interrogative requests. Journal of Pragmatics 38, 1081–1104.

Heritage, J., Raymond, G., 2005. The terms of agreement: indexing epistemic authority and subordination in assessment sequences. Social Psychology Quarterly 68, 15–38.

Heritage, J., 2013. Epistemics in conversation. In: Stivers, T., Sidnell, J. (Eds.), Handbook of Conversation Analysis. Wiley/Blackwell, London, pp. 370–394.

Ide, S., 1989. Formal forms and discernment: two neglected aspects of linguistic politeness. Multilingua 8 (2/3), 223–248.

Kasper, G., 1990. Linguistic politeness: current research issues. Journal of Pragmatics 14, 193–218.

Lakoff, R., 1973. The logic of politeness or minding your p's and q's. In: Papers from the Ninth Regional Meeting of the Chicago Linguistic Society Chicago, pp. 292–305.

Leech, G., 1983. Principles of Pragmatics. Longman, London.

Lerner, Gene, 1996. Finding "face" in the preference structures of talk-in-interaction. Social Psychology Quarterly 59, 303–321.

Linguistic Politeness Research Group (Ed.), 2011. Discursive Approaches to Politeness. Mouton de Gruyter, Berlin.

Matsumoto, Y., 1988. Reexamination of the universality of face: politeness phenomena in Japanese. Journal of Pragmatics 12, 403–426.

Matsumoto, Y., 1989. Politeness and conversational universals – observations from Japanese. Multilingua 8 (2/3), 207–221.

Noordzij, M., Newman-Norlund, S.E., De Ruiter, J.P., Hagoort, P., Levinson, S.C., Toni, I., 2009. Brain mechanisms underlying human communication. Frontiers in Human Neuroscience 3, 14.1–14.13.

Noordzij, M.L., Newman-Norlund, S.E., De Ruiter, J.P., Hagoort, P., Levinson, S.C., Toni, I., 2010. Neural correlates of intentional communication. Frontiers in Neuroscience 4, E188.

O'Driscoll, J., 1996. About face: a defence and elaboration of universal dualism. Journal of Pragmatics 25, 1–32.

Okamoto, S., 1999. Situated politeness: coordinating honorific and non-honorific expressions in Japanese conversations. Pragmatics 9 (1), 51–74.

Pomerantz, A., Heritage, J., 2013. Preference. In: Stivers, T., Sidnell, J. (Eds.), Handbook of Conversation Analysis. Wiley-Blackwell Press, Hoboken, NJ, pp. 210–228.

Sidnell, J., 2009. Conversation Analysis: Comparative Perspectives. Cambridge University Press, Cambridge, UK.

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.-E., Levinson, S.C., 2009. Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences of the United States of America 106 (26), 10587–10592.

Strauss, C., 2004. Cultural standing in expression of opinion. Language in Society 33 (2), 161–194.

Strecker, I., 1998. The Social Practice of Symbolization. Athlone, London.

Strecker, I., 2010. Ethnographic Chiasmus: Essays on Culture, Conflict and Rhetoric. Lit and Michigan State University Press, Michigan; Berlin.

Tomasello, M., 2009. Why We Cooperate. MIT Press, Cambridge, MA.

Tomasello, M., 2008. Origins of Human Communication. MIT Press, Cambridge, MA.

Tracy, K., 1990. The many faces of facework. In: Giles, H., Robinson, W.P. (Eds.), Handbook of Language and Social Psychology. Wiley, Chichester, UK, pp. 209–226.

Watts, R.J., Ide, S., Ehlich, K. (Eds.), 1992. Politeness in Language. Mouton, Berlin.

Political Advertising

Stephen Ansolabehere, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11624–11628, © 2001, Elsevier Ltd.

Abstract

Campaign messages broadcast through radio, television, or the Internet, or circulated via handbills, newspaper advertisements, or word of mouth provide people with much of their political information. Research on the effects of such messages has had both a positive and normative component. The main strain of positive research concerns how information translates into electoral opinions and behavior. The main stream of normative research concerns whether people are adequately informed, or even misinformed, and how we can improve the information content of political advertising. This article briefly considers recent empirical research on both matters. The principle finding of empirical research in this area is that advertising mainly reinforces viewer's predispositions. As a result, advertising imparts considerable information to viewers, but appears to have modest effects on election outcomes.

Political advertising has been the subject of intense scholarly investigation for nearly a century. Two agendas underlie this strain of social science inquiry.

The first agenda seeks to measure how political advertising, and campaigns generally, produce votes. Candidates' campaign advertisements are a major source of information about electoral choices that voters face. The intellectual origins date to Gosnell's studies of voting in Chicago in the 1920s (Gosnell, 1927, 1935). How much and how do advertisements influence election outcomes? The answers to these questions have very practical uses for political practitioners, who seek information that can guide campaign strategies, and for social scientists, who seek to explain how elections are won. Many social scientists question whether political advertising has any influence on elections, because elections are readily forecasted using indicators of peace and prosperity. Key writes in *The Responsible Electorate* that "The hullabaloo of a presidential campaign so commands our attention that we ascribe to campaigns great power to sway the multitude. Campaigning does change votes and it does bestir people to vote. Yet other influences doubtless outweigh the campaign in the determination of the vote" (1966, p. 9). On election forecasts, see Rosenstone (1983).

The second agenda concerns the extent of voter autonomy. Are votes hollow echoes of campaign messages, or does political discourse enable voters to reach reasonable decisions that reflect their own preferences and values? Electoral reform motivates much research on advertising and voter autonomy. If political advertising easily misleads voters, then restrictions on campaign advertising or some sort of 'political consumer protection' may be necessary in order to make democracy work. To the extent that there is a scholarly consensus on this matter it is that concerns about voter autonomy are overdrawn. Behavioral research in a wide variety of settings indicates that advertising has modest but important effects on election outcomes: advertising can make a difference in a close election, but it does not appear to persuade large numbers of voters to turn against a candidate or party that is well-known, well-liked, and represents the voters' interests.

Advertising and Vote Production

Political advertising covers a wide range of media that candidates use to mobilize supporters to turn out or to persuade people to vote for them. Commentary on political advertising often equates it with television advertising, as, for instance, when European commentators descry the 'Americanization' of their politics. But even in the US there is, in fact, much less television in politics than is often supposed. Campaigns for the US House of Representatives, for instance, use as much direct mail as television advertising. And in local campaigns candidates use little television but a vast range of other media, including mail, billboards, lawn signs, telephones, and door-to-door canvassing (Key, 1966).

Campaigns' usage of different media appears to follow the same principles that apply to firms. Candidates choose what media to use based on the productivity of the media and their relative prices. Variation in the mix of media used at different levels of office suggests that there are economies of scale to different sorts of advertising. Personal contact is reputed to be the most effective mode of political communication and it is a staple of local elections everywhere. But it is impossible to spend 15 min with each of the millions voters in a national election.

Mass media are most cost-effective at reaching broad constituencies. Evidence of the rational use of media comes from US House races. The average household in the typical House district receives about two pieces of mail and sees at least two television advertisements. However, candidates choose a mix of advertising media based on the prices and productivity of those means of reaching potential voters (or consumers). Campaigns for the US House in districts with expensive broadcasting media (such as Los Angeles and New York) spend $2.50 on mail for every $1 broadcasting, or three pieces of mail and 15 s of advertising per person. House campaigns in districts with inexpensive broadcasting (such as Dayton, Ohio and El Paso, Texas) spend $10 on broadcasting for every $1 on mail, or one piece of mail and 8 min of advertising per person. (For details see Ansolabehere et al., 1999a.) To my knowledge, similar sorts of calculations have not been conducted for other countries, though it is evident

from patterns of media usage that parties and candidates in Europe, Australia, and elsewhere are similarly responsive to media prices. (See, for example, Blumler, 1983 and Australian Election Commission, 1997.)

Social science research has focused on how voters respond to the overall volume and different sorts of campaign advertising that they receive. The estimated effects depend in part on the methods used.

Three different approaches have been taken to the study of advertising. First, laboratory and field experiments have been used to gain precise measures of the effects of advertising exposure in specific campaign settings. Such studies control what messages individuals receive and then measure differences in opinions or behaviors in response to different sorts of messages (e.g., Gosnell, 1927, 1935; Ansolabehere and Iyengar, 1996; Gerber and Green, 1999; Norris 1999).

Second, studies of aggregate data correlate the observed vote to levels of campaign spending. Some of these studies design their inquiries around 'quasi-experiments' which exploit factors that change levels of spending (such as personal wealth) in order to measure the effect of spending on the vote. The seminal work in this vein, for both American and European researchers, is Jacobson (1978). Subsequent research has improved on this model; see, for example, Gerber (1998). Races involving three or more candidates pose statistical problems in estimation of vote production functions, for which widely accepted statistical models have not yet been developed. Third, surveys associate reported vote or political attitudes with reported or recalled exposure to messages. Almost all research outside of the US relies on survey methods. For a discussion of the problems with survey-based measures of campaign effects see Ansolabehere et al. (1999b). Analogous with the aggregate studies, instrumental variables can solve such problems.

Experiments, which offer high internal validity, show a remarkably consistent set of results about the capacity of advertising to produce votes. Researchers observe exactly what messages citizens see and can design the study so that other factors do not confound the effect of the experimental manipulation. Perhaps the earliest experimental study of political advertising and campaigning took place in the 1920s in the US. To measure the mobilizing effects of direct campaigning, Gosnell (1927) sent a cadre of graduate students out to canvass a set of blocs of the city of Chicago in an effort to manipulate the levels of voter contact. They found higher in turnout ranging from 4 to 7% in the areas that the students canvassed over the areas that they did not canvass. Fast forward to the 1990s. Gerber and Green (1999) replicated Gosnell's study in several towns in southern Connecticut. They found increases in participation similar to those reported in the Chicago studies 70 years earlier. They also found increases in vote shares of candidates in the neighborhood of 4 percentage points for two pieces of mail. In a series of laboratory experiments conducted in Los Angeles from 1990 to 1993, Ansolabehere and Iyengar (1996) found strikingly similar results for television advertising. Exposure to a 30-s television advertisement raised the percentage of viewers who intended to vote for a candidate by 4 to 10 points. These and many other experiments show substantial effects of political advertising on candidates vote shares. Those who see a broadcast message, receive a mailing, or talk to a campaign worker are between 4 and 10% more likely to support the candidate being promoted.

Aggregate studies of campaign spending yield similar, albeit lower estimates of the effectiveness of campaign advertising. Using fairly complicated, and controversial, statistical models, researchers have attempted to measure the marginal effect on the vote of spending an additional campaign dollar. Such techniques yield estimates ranging from miniscule effects of a dollar spend on the vote to estimates of approximately $10 per vote. The key papers in this subfield are Jacobson (1978), Green and Krasno (1988), and Gerber (1998).

Such estimates are in the right ballpark, but the wide range of estimates in the literature on campaign spending suggests that additional work finding valid statistical models is in order before we can put much confidence in the estimates emerging from this literature.

The problem is that it is difficult to sort out cause and effect in aggregate electoral data. Consider, for example, two similar legislators: one holds a highly competitive seat and the other holds a very safe seat. The legislator in the competitive seat must spend large amounts of money to win re-election; the legislator in the safe seat is unlikely to be challenged and if they are they do not need to spend much to retain the seat. In this example the correlation between spending and vote shares is negative! To fix this problem researchers have taken a 'quasi-experimental' approach, using factors such as candidate wealth, which affect directly how much candidates spend and in turn affect their vote shares. Improving such statistical models remains an important research challenge.

Survey research has presented the most modest estimates of the effectiveness of advertising. A long line of studies report minimal effects of campaigns in general and of advertising in particular on the vote (in the US, see Campbell et al., 1960; in the UK, see Norris, 1999). Serious problems of question validity lead to serious doubts about these research findings. In the laboratory setting, Ansolabehere and Iyengar (1996) document that survey-based measures of advertising exposure miss one exposure for every one that they detect. This degree of misreporting is sufficient to explain why surveys often produce insignificant findings or even estimates in the 'wrong' direction. Quasi-experimental techniques analogous to those used in aggregate studies can correct measurement problems (Bartels, 1993).

Beyond differences in media and methodologies, a further theme in the study of political advertising is the mechanism through which messages become votes. There are three. Advertising may solidify an uncertain vote (reinforcement), it may stimulate a supporter to vote (mobilization), and it may lead a voter to change their mind (conversion). This typology is due to Berelson et al. (1954). Although political observers often talk of 'swing' voters and movement in the polls, there appears to be relatively little evidence in experiments or surveys of conversion. Few Republicans decide to vote Democrat and few Democrats decide to vote Republican in response to political advertisements. Perhaps the strongest evidence against conversion, though, comes from nonpartisans. Nonpartisans, especially low information independents, are the least responsive to political advertising.

Reinforcement appears to be the dominant mechanism in political communication. For instance, in the experimental

research reinforcement produces the clearest effect on the vote. Nonpartisans and viewers of the opposite party as the sponsor of the ad showed no statistically significant response to a candidate's ads. Low information partisans of the same party as the sponsor showed by far the strongest favorable response to political advertising. That effect shrank, though it remained statistically and substantively large, with the partisan's level of information and interest in politics (Ansolabehere and Iyengar, 1996). This pattern suggests that political advertising fits neatly into the more general model of information acquisition developed by Zaller, in which both predisposition to believe a piece of information and prior levels of knowledge determine responsiveness to new information (Zaller, 1992).

Mobilization has proved the most controversial of the three mechanisms. Experimental research on political advertising has called into question the longstanding belief among political scientists that campaigns are inherently stimulating affairs. Berleson et al. (1954) describe the process in detail. This is essentially the model behind retrospective voting developed by Key (1966) and widely used to understand elections in the US, Europe, and elsewhere. A large number of experimental studies have demonstrated that not all viewers are stimulated to vote by political advertisements. For example, nonpartisan and low information viewers exposed to negative advertising appears express significantly lower intentions to vote than others. As a result, negative campaigning can produce lower turnout among the key group of 'swing' voters and an electorate that reflects two ideologically opposed groups.

The demobilizing effects of negative advertisements on nonpartisans underscores the difference between reinforcement and mobilization. Negative advertising might win the highest vote share, mainly by reinforcing partisans, but it can lead to lower overall participation.

One important caveat about the state of social science research on political advertising is in order. The bulk of the research on the effectiveness of political advertising has been conducted in the United States. Many of the findings about mobilization and reinforcement likely depend on the electoral system, which influences the number of parties or candidates. The dynamics of campaigns and the consequences of advertising appear to be quite different in races involving more than two candidates. Johnston et al. (1992) provide excellent examples of political communications having complicated effects on competition in Canada's multiparty system. Careful experimental research on multicandidate and multiparty races, though, is in its nascency.

Voter Autonomy

The central problem for democratic theory in modern societies is that people apparently know little about the choices they face and they have little incentive to seek new information. Candidates and parties, then, have very strong incentives to provide information to voters at little or no search costs: candidates try to reach us, rather than us, them. As a result we are highly reliant on what politicians choose to say in their advertisements.

This is the problem of voter autonomy. By one telling voters have abdicated to political elites the important jobs of choosing what issues will be discussed, what elections and therefore government will be about, and even what we will think as people walk into the voting booth. From this view, people are highly susceptible to false or deceptive advertising and will vote for the candidate whom we see the most rather than the one we think is best representative. Regulation of the volume and content of political advertising may, therefore, be necessary.

The counter argument is that voters, even though they do not know much by way of details, have a clear sense of what they themselves value or prefer, and, in order to win elections, politicians must adjust their messages to what the voters want. In crafting their advertisements, a candidate selects issues on which he or she can make electoral gains, because the voters care about it, because the candidate has a strong record, or because the candidate has have taken the more popular position. This leads to majoritarian biases in what is discussed, but that is the essence of elections.

Voter autonomy presents a much more difficult set of research questions. In order to say that people are or are not misled by political communications, we must establish what it is that they prefer or value a priori. The effects of advertisements must then be judged against that baseline. Two questions are central.

First, do advertisements lead people to choose a candidate who is less good for their interests or values? Research showing strong reinforcement and few conversion effects suggests that the answer to this question is probably no. In both survey and experimental data it appears that people who lean toward a party become more solidly in favor of that party as the campaigns progress (Ansolabehere and Iyengar, 1996, Chap. 4; Gelman and King, 1990). An important variation on this arises from tactical voting in Britain, Canada, and other countries. As campaigns progress, voters supporting a candidate or party in third place or lower might vote for the second place candidate if they prefer that candidate over the front-runner. This seems to have happened in the 2000 Mexican election as support for Cardenas fell over the course of the campaign and support for Fox, who was in second place most of the race, rose.

Nonpartisans tend to move very little, and tend not to vote. The immediate political implication is that campaign discourse is polarizing: it ultimately pits Democrats against Republicans, Conservatives against Labourites. The low amount of conversion suggests, however, that people fundamentally are not misled.

Second, can advertising set the public agenda, or are people mainly responsive to what they themselves care about? While agenda setting effects have been widely documented in other areas of study, they have received surprisingly little careful study in the field of political advertising. Early evidence of agenda setting is provided by Bereleson et al. (1954, Chap. 12). Two careful studies stand out. Jamieson's (1992) *Dirty Politics* documents how the Republican and Democratic strategists use their advertising and the news media to magnify their own agendas, especially Republican's usage of crime and race. Johnston et al. (1992) demonstrate that the debate over free trade completely transformed the Canadian voters' priorities, without changing their preferences on free trade itself.

Reinforcement and agenda setting can hold conflicting implications for voter autonomy. Partisan reinforcement

suggests that there is little evidence that people vote against their expressed preferences; agenda setting suggests that people are vulnerable to what the campaigns choose to discuss. There may, however, be no contradiction here. It may indeed be the case in each of the examples discussed that voters were simply very receptive to new agendas, and that elections gave voters an opportunity to change public policies on trade, affirmative action, and other issues from those set decades earlier.

New Directions

Political advertising poses both methodological and normative challenges for social scientists, and our understanding on both levels has advanced considerably over the last two decades. Advertising has important effects on election outcomes, but it does not appear to lead people to vote for candidates with whom they fundamentally disagree. It, therefore, does not on the whole undercut the normal workings of majority rule in democratic elections.

In terms of methods, a high priority for further inquiry is bridging the gaps between experimental, aggregate, and survey methods, and developing more powerful quasi-experimental approaches to correct measurement and simultaneity problems. To this end, research on vote production, which has relied mainly on quasi-experiments in aggregate data, should make greater use of experiments to map out features of the vote production function more fully, especially the effects of repeated exposures to messages and interactions between the messages and the viewers preferences.

In addition, there is considerable need for research on political advertising outside the US. The American focus of campaign research stems in part from existing regulations on elections and media. The US imposes few limits on how candidates and parties raise and spend campaign funds, on their access to media, and on what can be said on the various media. As a result, the US has proved a rich field to study advertising effects. Research on campaigning in European countries has placed a much greater emphasis on electoral laws, political parties, and tactical voting. There is, however, growing interest in campaign effects in European political science.

Beyond methodological differences, there is the matter of why. Reinforcement and mobilization are a sufficiently well-established phenomenon that it is worth opening up the black box. Why are these the routes through which political advertising works, and why not conversion? The highly inductive approach of empirical researchers is likely to pay the biggest dividends, as political theory has yet to produce a powerful model of why people vote.

See also: Mass Communication: Normative Frameworks; Mass Media, Representations In; Political Communication; Political Money and Party Finance; Political Science and Polling; Voting, Sociology of.

Bibliography

Ansolabehere, S., Gerber, A., Snyder Jr., J.M., 1999a. How Campaigns Respond to Media Prices. MIT, Cambridge, MA.
Ansolabehere, S., Iyengar, S., 1996. Going Negative: How Political Advertising Shrinks and Polarizes the Electorate. Free Press, New York.
Ansolabehere, S., Iyengar, S., Simon, A., 1999. Replicating experiments using surveys and aggregate data: the case of negative advertising. American Political Science Review 93, 901–910.
Australian Election Commission, 1997. Funding and Disclosure Report, 1996. AEC, Sydney, Australia.
Bartels, L., 1993. Message received. American Political Science Review 87, 267–285.
Berelson, B., Lazarsfeld, P., McPhee, W., 1954. Voting. University of Chicago Press, Chicago.
Blumler, J., 1983. Communicating to Voters: Television in the First European Parliament Elections. Sage, London.
Campbell, A., Converse, P., Miller, W., Stokes, D., 1960. The American Voter. University of Chicago Press, Chicago.
Fritz, S., Morris, D., 1992. The Handbook of Campaign Spending, 1990. CQ Press, Washington, DC.
Gelman, A., King, G., 1990. Why are presidential election campaign polls so variable when votes are so predictable? British Journal of Political Science 23, 409–452.
Gosnell, H.F., 1927. Getting-Out-the-Vote: An Experiment in the Stimulation of Voting. University of Chicago Press, Chicago.
Gosnell, H.F., 1935. Machine Politics: Chicago Model. University of Chicago Press, Chicago.
Green, D., Krasno, J., 1988. Salvation for the spendthrift incumbent. American Journal of Political Science 32, 884–907.
Gerber, A., 1998. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. American Political Science Review 92, 401–411.
Gerber, A., Green, D., 1999. Does canvassing increase voter turnout? A field experiment. Proceedings of the National Academy of Sciences 96, 10939–10942.
Jacobson, G., 1978. The effects of campaign spending in congressional elections. American Political Science Review 72, 469–491.
Jamieson, K.H., 1992. Dirty Politics. Oxford University Press, New York.
Johnston, R., Blais, A., Brady, H.E., Crete, J., 1992. Letting the People Decide: Dynamics of a Canadian Election. Stanford University Press, Palo Alto, CA.
Key, V.O., 1966. The Responsible Electorate. Vintage, New York.
Morris, D., Gamache, M., 1994. The Handbook of Campaign Spending, 1992. CQ Press, Washington, DC.
Norris, P., 1999. Political communications in British election campaigns: reconsidering media effects. British Elections and Parties Review 5, 190–211.
Rosenstone, S., 1983. Forecasting Presidential Elections. Yale University Press, New Haven, CT.
Zaller, J., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.

Political Anthropology

Christian Krohn-Hansen, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Spencer, volume 17, pp. 11628–11631, © 2001, Elsevier Ltd.

Abstract

Anthropologists have approached the comparative study of politics in different ways, depending on their general theoretical perspectives. This article provides an overview of the history of political anthropology from around 1940 until now. The author first discusses different anthropological attempts to define politics, or the political, and looks at the classic, British structural-functionalist anthropology of politics in the 1940s and 1950s. The author thereafter examines a series of significant changes that the anthropology of politics has registered since the early 1970s and rounds off by discussing features of the last two or three decades' anthropology of the modern state, and by reflecting briefly on political anthropology's relationship to, and interest in, studies of forms of political violence. The article maintains that since the 1980s political anthropology has become less strictly disciplinary, and has gradually opened itself more to academic conversations with interests and views derived from political philosophy and political science.

Anthropology and the Definition of Politics

Political anthropology is a subfield of sociocultural anthropology, but like anthropology as a whole, it remains immune to precise delimitation. The core of political anthropology is the comparative, fieldwork-based examination of politics in a broad range of historical, social, and cultural settings. Today, it is common to see political anthropologists combine ethnographic work with history. Some analyze the symbolic forms and practices of a specific state bureaucracy, others a form of political activism, and yet others the perpetration of terror or torture, or the political effects associated with the everyday and ritual construction of a particular collective memory. The field of political anthropology has been, and continues to be, extensive, diverse, and shifting.

Anthropologists have defined politics in different ways, depending on their general theoretical–analytical interests and research questions. A typical anthropological point of departure today states that politics, or the political, must not be treated as a separate domain or field, but must be investigated as articulation between power relations, cultural processes, and historical trajectories. As one contemporary anthropologist, Steven Gregory, put it in his *Black Corona* (1998, p. 13), "For my purposes, politics refers to a diverse range of social practices through which people negotiate power relations. The practice of politics involves both the production and exercise of social relationships and the cultural construction of social meanings that support or undermine those relationships." The work that founded and gave shape to the classic, British structural-functionalist anthropology of politics in the 1940s and 1950s, *African Political Systems* edited by Meyer Fortes and Edward Evans-Pritchard, viewed a society's political organization as those structures or relations that maintain "social order within a territorial framework, by the organized exercise of coercive authority through the use, or the possibility of use, of physical force" (1940, p. xiv).

In the 1950s and 1960s, anthropologists gradually moved away from the more rigid structural-functionalist views, and many came under the influence of versions of action theory and interactionism. The reformulated political anthropology that emerged during these two decades saw politics in formal terms: political anthropology examines comparatively 'political behavior' – typically defined as (1) the search for power or influence and (2) a human constant, a universal aspect of societies.

In a remarkable study, *Society against the State* (1977 [1974]), the French anthropologist Pierre Clastres criticized all traditional political anthropology for having universalized a particular (Western) notion of political power, or politics, a notion that identifies political power with coercion or violence: "the kinship is closer than seems apparent between Nietzsche, Max Weber (state power as the monopoly of the legitimate use of violence), and contemporary" anthropology, he wrote (p. 4). What, Clastres asked, do we do with Amerindian societies in which "if there is something completely alien to an Indian, it is the idea of giving an order or having to obey" – in other words, societies where political life functions as a process 'beyond hierarchical subordination' and totally 'beyond coercion and violence' (p. 5). Are these societies not only without 'a state' but also without 'politics'? Or is there something wrong with a particular way of thinking about the nature of political power? Clastres replied that a genuinely cross-cultural political anthropology needs a wide and open definition of politics: "all societies, whether archaic or not, are political, even if the political is expressed in many voices, even if their meaning is not immediately decipherable; and even if one has to solve the riddle of a 'powerless' power" (p. 14).

One year before Clastres published his book, Talad Asad launched a pathbreaking critique of structural-functionalist anthropology, and of the mainstream anthropological project more generally, in *Anthropology and the Colonial Encounter* (1973). Asad's study and similar efforts helped partly to launch and partly to strengthen anthropologists' interest in colonial and postcolonial studies, and in history. Today, most in the discipline consider the issue of how to define politics in terms of a certain history: the notion of 'the political' appeared in Western Europe after the Reformation,

alongside the modern concepts of the state and democracy. The geographical and historical origin of the discourse of politics did not, however, preclude its being exported to the rest of the world. The anthropology of politics must therefore study varying, historically and culturally constituted definitions of the political – in brief, whatever the actors (leaders and masses, states and nongovernmental organizations) across the globe consider politics to be.

The use of the term 'political anthropology' was widespread in anthropology in the 1960s and early 1970s. Subsequently, it came to be used far less, indeed to be less fashionable, even among those in the discipline who evidently studied political phenomena. In 2002, an authoritative, instructive reader was published under the title *The Anthropology of Politics* (Vincent, 2002).

The Making of a Political Anthropology, *c.*1940–70

Before the publication of *African Political Systems*, structural-functionalist anthropologists had not formulated an explicit approach to the comparative study of politics. They had worked on 'religion' and 'kinship,' but not on 'politics.' That volume gave decisive direction to anthropological examinations of politics for at least two decades. Its authors' approach had five important features. First, to them the political question was primarily a matter of understanding how a certain 'order' was created and recreated in society, how particular institutions and mechanisms contributed to the maintenance of a sufficient measure of peace and cohesion in the social system. Second, although the structural-functionalists' general thinking about, and definition of, a political organization seemed premised upon universalization and naturalization of a specific, a European model, or a set of ideas about the nation-state, they did not study the state. Instead they followed Radcliffe-Brown, who in his Preface to *African Political Systems* suggested that instead of wasting time on the study of what he classified as a purely ideological construct, namely, 'the State,' it would be better simply to study forms of politics, forms of power and authority, and forms of government. The state, he wrote, is most often "… represented as being an entity over and above the human individuals who make a society, … and sometimes spoken of as having a will … or as issuing commands. The State in this sense does not exist in the phenomenal world; it is a fiction of the philosophers. What does exist is an organization, i.e., a collection of individual human beings connected by a complex system of relations …. There is no such thing as the power of the state" (1955 [1940], p. xxiii).

Third, these researchers were concerned to develop a taxonomic classification of types of political systems. They argued, for example, that Africa included three main types of society: those societies that had centralized authority, administrative machinery, clearly defined judicial institutions, and economic stratification; those societies that lacked centralized authority, administrative machinery, and judicial institutions – in brief, 'stateless' societies; and finally, those very small societies in which even the largest political unit embraced a group of people all united by kinship bonds, such as a roving extended family. Fourth, the structural-functionalists sought mainly to explain political systems in terms of synchronic examinations of mutual interdependence between institutions; generally they were little concerned with studying the past. And fifth, they demonstrated an impressive will, and ability, to show with rich ethnographic material why any productive anthropology of power and politics needs to remain intimately connected both with the anthropology of kinship and with the comparative study of religion. They generated a series of works documenting that the exercise of politics can be embedded, indeed apparently concealed, in the ordinary organization of kinship and marital bonds, and in forms of ritual life.

The best known example is Evans-Pritchard's (1940, 1955 [1940]) ethnographically based study of the 'stateless' political system of the Nuer, a pastoral people of southern Sudan. His analysis of the Nuer political order embodied the general views outlined above. The Nuer, Evans-Pritchard argued, were without chiefs or leaders with coercive or absolute authority over others, and he described their political systems as 'ordered anarchy' based on the principle of 'segmentary opposition.' Society's members were organized into clans and lineages based on male lines of descent from founding figures. Local groups constituted 'segments' of larger, more inclusive kinship units defined in terms of descent. This hierarchy of nested lineage segments of differing scale, Evans-Pritchard claimed, was derived from crucial moral-political values maintained by the Nuer. Obligations to back and help others in fighting were articulated in terms of kinship. Groups that were opposed at one level of segmentation, that of minor lineages, for instance, had a duty to unite in a conflict between the higher segmentary unit to which they all belonged and another unit of the same hierarchical level, such as a major lineage. These dynamics of 'fission and fusion' also ensured that Nuer society was able to act as a collective unity in conflicts with other 'tribes.' Evans-Pritchard's ahistorical, structuralist narrative of the Nuer quickly became enormously influential and was generalized into models of 'acephalous (primitive) political systems' and 'segmentary lineage theory.'

The reformulated political anthropology that gradually emerged in the 1950s and 1960s remained essentially based on synchronic investigations, and little preoccupied with the state. In addition, it continued to produce detailed ethnographies of connections between the practice of politics on the one hand and, on the other, the organization and uses of kinship and rituals. But in some ways, this so-called processual approach to the anthropological study of politics deviated from the earlier structural-functionalist school.

A major contribution on the road to the changed approach was Edmund Leach's *Political Systems of Highland Burma* (1954), which argued in favor of, and demonstrated, a more dynamic, more process-oriented type of examination. This study, which sought a way to treat and explain forms of social and political change, rested upon a simple, basic assumption: that a conscious or unconscious wish to gain power is an extremely common, probably universal human motive – and that persons faced with a choice of action will typically choose what helps to obtain power and recognition. Two important voices when the processual approach was given shape in the 1960s were the Norwegian Fredrik Barth (who wrote his doctoral work with Leach as supervisor) and the Englishman Fredrick G. Bailey. Both, like Leach, conducted research outside Africa, in Asia, and in the Middle East. Although it would be wrong to

consider the process approach as a coherent paradigm or theory, the transformed focus may be summed up under three headings.

First, the general interest in producing and discussing classification of types of political systems disappeared; already in Barth's first major work, *Political Leadership among Swat Pathans* (1959), this interest was entirely absent. Second, the new political anthropology shifted the focus of analysis away from those structures that secure order to the many activities and forms of action everywhere that express competition for influence, control, and leadership – in brief, forms of power struggle. Process analysis was based on the idea that political behavior and political processes ought to be studied in terms of networks of individuals and groups that employ power to achieve particular results. The objective became finding a way to model features of such competition and struggle for power. And third, new terms were put to use to study politics cross-culturally: terms like 'process,' 'strategy,' 'tactic,' 'maximizing,' 'political faction,' and 'political arena.' Most of these concepts have since been used by anthropologists, albeit to varying extents.

Culture, Power, and History

During the late 1960s and 1970s, the discipline registered rapidly growing interest in forms of Marxist-inspired perspectives, world-systems theory, and history. Anthropology developed a stronger awareness of the need to understand the historical process of Western global domination. Since the late 1960s and early 1970s, the discipline has also seen the emergence of symbolic anthropology, or the construction of a strikingly strong concern with the study of meaning. From the second half of the 1970s, the ideas of Foucault, Bourdieu, and Gramsci exercised considerable influence on anthropologists studying the relationship between power and meaning, including political forms and practices. As already noted, the category or label 'political anthropology' soon became old-fashioned. Yet anthropologists still continued to examine political phenomena. In the period from the early 1970s to the late 1990s, anthropologists studying political practices and processes were preoccupied with at least three wide, overlapping problem areas: 'power and culture,' 'domination and resistance,' and 'power and history.' Let us consider each of these research interests in turn.

Power and Culture

Increasingly, anthropologists realized that the production of meanings is what relationships of power, myths of authority, and more formalized forms of politics and state building everywhere in the contemporary world is all about. In 1974, Abner Cohen published an important study that argued that Political Man is also Symbolist Man, that man is two dimensional. His *Two-Dimensional Man* examined the possibilities for systematic anthropological study of what Cohen described as the dialectical interplay between power relationships and uses of symbols in modern, complex, industrial societies. Yet far more influential was another book – Clifford Geertz's *The Interpretation of Cultures* (1973). Geertz's ideas about the analysis of cultural life helped to frame and shape much later work, not only in anthropology, but also in history and other academic disciplines. In 1980, Geertz published *Negara: The Theatre State in Nineteenth-Century Bali*. Here he maintained that much political analysis had been reductionist: it had ignored and silenced the significance of rituals, symbols, and meanings in the construction and reconstruction of forms of politics. The rituals of the Balinese theatre state "were what there was" (1980, p. 136); they were not means of representing power or masking its true nature – they actually constituted the state.

The 1980s and 1990s saw the appearance of a strongly intensified, interdisciplinary concern with the study of ethnicity and nationalism, especially the nation form. Anthropologists contributed by investigating nations and nationalisms as symbolic forms, as particular cultural worlds and meaning formations. An instructive example is Bruce Kapferer's *Legends of People, Myths of State* (1988), an unconventional comparison of two forms of nationalist thinking – one expressed through Australia's 'aggressively' democratic egalitarianism, the other through the sharply hierarchical thinking of Buddhists in Sri Lanka. Kapferer examined contemporary political worldviews "on their own terms and in the contexts of their ideas" (1988, p. xii). In so doing, he brought out powerfully both the enormous cultural complexity of forms of politics in the contemporary world and the way in which assumptions and perspectives derived from the heart of symbolic anthropology could be used to illuminate such forms.

Another significant anthropological contribution to the scholarship on identity politics was the anthology *Naturalizing Power: Essays in Feminist Cultural Analysis*, edited by Sylvia Yanagisako and Carol Delaney (1995). Like Kapferer, the authors of this anthology drew on fairly mainstream anthropological work on uses of imaginaries and rituals in order to shed light on forms of power differentials and authority in the contemporary world. But *Naturalizing Power* had a broader scope; it treated a whole specter of different types of self-understanding – family identity, racial identity, national identity, and religious identity. These authors argue that, although differences of identity and power often seem to be given in nature or ordained by God, they are in fact not logical extensions of a natural order, but rather an integral part of cosmologies, including myths of origins, descent, and kinship. A line goes from the classic structural-functionalists' preoccupation with the need to examine specific thinking about descent and kinship, to be able to comprehend forms of political agency, to *Naturalizing Power*'s and today's anthropologists' clearly overlapping and related interests.

Domination and Resistance

A central focus in anthropology has been how subordinate communities including minorities, for example, indigenous peoples, have been exploited, silenced, marginalized, or destroyed by ruling strata or interests – wealthy landowners, big companies, the state, and so on. This is a wholly reasonable and natural perspective in many contexts. However, in the 1980s and early 1990s, anthropologists increasingly explored

the ways in which subordinated and apparently powerless individuals and groups actively offer resistance, nonviolently, using whatever means and strategies they have access to.

A study that inspired many was James C. Scott's *Weapons of the Weak* (1985), based on field research among rich and poor in a rice-farming Malaysian village. In this book, he maintains that the response to routine domination or daily exploitation is always the subordinates' general and unremitting efforts to fight back. His eyes are fixed on a variety of prosaic weapons used by weak groups, like foot dragging, dissembling, false compliance, pilfering, feigned ignorance, setting of fires, sabotage, and desertion. Such everyday forms of resistance have often proved the most effective in the long run. This shifts the focus of analysis from open, organized (formal) activities (for example, collective stonewalling or rebellion) to what can be called the essentially unwritten and muted history of political protest. *Weapons of the Weak* goes so far as to invert the ideas of Gramsci. The latter argues that the 'radicalism' of subordinates is more strongly present in their actions than in their ideas, whereas Scott contends that the weak are most severely restricted in the sphere of behavior, while the oppressed are least shackled in the sphere of notions and interpretations, where they can safely be ventured. He underscores the necessity of never associating conformity automatically with intellectual or moral compliance, much less with legitimation.

The wave of resistance studies in the 1980s and 1990s contributed to better understanding of the dynamics of resistance, and the possible space for it, in power relations, but these studies were also met with criticism. They have been criticized for idealizing or romanticizing resistance among the poor and oppressed, and for being ethnographically 'thin': thin on the complexity of cultural meaning formation by dominated groups, and on the internal politics of those groups.

Power and History

With a growing anthropological interest in history, there also appeared a sharper recognition of the inevitable complexity, and the contested nature, of the field of power and knowledge that one calls the past. Or to put it another way, researchers increasingly took for granted that, if one wishes to understand political processes, it is necessary to examine history in at least two different manners, or to study it in two different senses of the word: first, to seek to reconstruct history according to academic criteria for history production and history writing (using archives, libraries, and so on), and second, to study the political agents' own histories, their own narratives of the past, present, and future. As many have shown, collective and individual memories are often at the core of contemporary politics. Memories give form to the practice of politics in a deep sense. Ideas about the past shape ethnic, national, and religious identities, authorize claims, and provide political leaders and masses with a horizon and a direction. Yet all narratives of the past are anchored in time. Memories reflect, and express, particular power relations and particular cultural forms and practices. The construction of stories about the past is a profoundly historical process.

An early example of a many-faceted anthropological treatment of the relationship between power and the making of history is Richard Price's classic *First-Time: The Historical Vision of an Afro-American People* (1983). The book contains a thought-provoking collection of oral testimonies about the lives of eighteenth-century Saramakas, 'Maroons' who escaped from slavery and settled in the rainforest of Suriname. Yet at the same time, the text provides a double account of Saramaka history. Saramaka communities offered Price their most secret and powerful memories. Hence, the book allows us to witness slaves escaping, surviving, and producing a new understanding of themselves in brutal surroundings. To this 'native' history, Price juxtaposes systematically a Western historical account and perspective. "Saramaka collective identity is," Price emphasizes, "predicated on a single opposition: freedom versus slavery. The central role of First-Time [their narrative of the past] in Saramaka life is ideological; preservation of its knowledge is their way of saying 'Never again'" (1983, p. 11). In a recent book, Rainforest Warriors (2011), Price tells movingly of how Saramakas continue to preserve and employ their First-Time knowledge, while they have been able, incredibly enough, to harness international human rights law to gain control politically and legally of their own piece of the Amazonian forest and secure their survival.

Tracking the Political, Reconstituting the Anthropology of the State

Today, many share Price's basic ideas about a necessary principle to be accepted for fruitful political anthropology: that political activities and struggles are not confined to a narrow set of (political) arenas, but take place in all situations where people negotiate power. Anthropologists thus seek to investigate what constitutes the political about the processes they study, but insist that politics, or the political, should not be treated as a separate category, but should be examined as connections between power relations, cultural forms and practices, and historicity. A striking change in political anthropology since the late 1980s has been the gradual emergence of a thriving new anthropology of the modern state. Most of this reconstituted and renewed anthropology of the state has been shaped by, and remains rooted in, the same ideas that I have just sketched: that is, ideas emphasizing the need to move from a focus on studies of 'politics' to a focus on explorations of culture, power, and history. In some ways there is a clear overlap between Radcliffe-Brown's conspicuous suspicion of political philosophy in the Preface to *African Political Systems* and aspects of the last two or three decades' anthropology of the state. As noted earlier, Radcliffe-Brown dismissed the idea of the state as a myth of philosophers. In contrast to this, authors of the new literature on the state take for their basis that anthropologists must study states empirically and ethnographically and take the comparative study of the state seriously. But they agree that typically philosophers, social scientists, politicians, and citizens have provided, and continue to provide, the state with an unreal concreteness. The recent anthropological literature on the state revolves around two core questions: what is the 'state,' and how can one fruitfully study it? Below the author outlines three features of this literature, before briefly mentioning a few examples of it.

The first fundamental idea behind the new anthropology of the state emphasizes that political theory, as Foucault (1980, pp. 97, 121) expressed it, needs to cut off the king's head – that is, that anthropologists should seek to unmask 'the state' and write against fetishizations of it. In an important essay, Philip Abrams stressed long ago that too much academic work has reified the state (1988, p. 76). As an alternative, he suggests that one examine the "state-system: a palpable nexus of practice and institutional structure centered in government and more or less extensive, unified and dominant in any given society" (1988, p. 82), as well as the 'state-idea,' an ideological construct "attributing unity, morality and independence to the disunited, amoral and dependent workings of the practice of government" (1988, p. 81). On the basis of ideas of this type, anthropologists have sought to answer such questions as, in what ways, and through what practices and processes, does the state appear as an objective entity over society? Through what imageries are states conceptualized, understood, and represented?

The second important idea upon which the new literature rests has to do with the understanding of the complexity of the state. According to these works, the state may most fruitfully be understood as a set of translocal institutions that are made manifest and visible in localized (or specific) practices. As Penelope Harvey (2005, p. 138) has formulated it, considering the state as 'more than' the local (or worse, as 'over' or 'above' the local, or as 'encompassing' the local) seems not very helpful. Why has this position been crucial, not to say liberating? It means accepting that researchers may produce valuable knowledge of a part of the complexity inherent in the constitution and reconstitution of a particular state-system through investigating a group of localized processes and practices, or through making an ethnography based on (localized) participant observation. Another way of putting this is to argue, as Veena Das and Deborah Poole (2004) have done, that what are often considered as the state's territorial and/or social 'margins' are not peripheral to its workings but, on the contrary, constitute manifest and prominent key features of its routine functioning.

The third major idea behind the new studies stresses the historical and cultural specificity of constructions of the state and the need for a thoroughly critical perspective on the growing, global domination of the Western conceptual apparatus. The building of modern states has typically been understood and studied within a framework that distinguishes between 'state' and 'civil society.' This conceptual distinction was created on the basis of the historical experience of Western Europe. The basic problem, however, is not that the distinction between state and civil society expresses Eurocentrism. The problem is that the separation between society, or civil society, and the state, or between the social sphere and the political sphere, does not exist in reality – not even in the West. To reiterate, the renewed anthropological research on the state claims that a modern state must be understood as produced by a broad and shifting field of power relations, everyday practices and formations of meaning. Instead of operating with a sharply confining and static starting point, the framework of a distinction between the state and civil society, one should begin by recognizing that one needs a more open conception of the state. This invites the researcher to examine in detail how a particular state is built and given shape in everyday encounters at the local level – in those contexts where the state bodies' representatives and individuals and groups interact. It is true that this makes the presence of the state in social life a good deal more fluid than is often thought. But it is precisely this fact – that the state's presence in social life really is incoherent and fluid – that makes critical, grounded anthropological work so necessary if one wishes to comprehend state-building processes.

One who inspired others at an early date to look again at the anthropology of the modern state was Michael Taussig. Through a series of thought-provoking – but also in some ways obscure – writings, first *The Nervous System* (1991) and then *The Magic of the State* (1997), he urged readers to reflect on the apparently magical powers of the state and the objectification and the fetishization of the state. Others who published important anthropological works on the state in the second half of the 1990s were Ana Alonso (1995) and Fernando Coronil (1997), writing, respectively, on the Mexican and the Venezuelan state. Since the late 1990s, two influential theorists in the discipline in general and in political anthropology in particular have been James Ferguson and Akhil Gupta. Ferguson and Gupta's coauthored, programmatic 2002 *American Ethnologist* article *Spatializing States: Toward an Ethnography of Neoliberal, Governmentality* has been enormously important. Or to put it 'more generally, these two authors' perspectives and works have, to a substantial degree, helped to define and give form to recent years' anthropology of the state, including the three key sets of central ideas listed above about the state and the study of the state. Gupta is in addition coeditor, with Aradhana Sharma, of the reader *The Anthropology of the State* (2006), a collection of historical and contemporary writings. In 2012, he published an innovative, ethnographically driven study of bureaucratic practices in India. In *Red Tape*, Gupta sees the relation between the state in India and the poor as one of structural violence that every year kills 2–3 million people. But, he argues, the country's poor are not helpless; they actively participate politically. Nor is the Indian state indifferent to the difficulties of the poor; it finances many poverty amelioration programs. Yet the Indian state care is arbitrary and irrational in its effects, he shows, and that arbitrariness is systematically produced by the very activities and projects that are meant to alleviate social suffering.

In brief, the state long ago returned as a basic concern in political anthropology, and since the late 1990s a similar shift has, in effect, occurred with a series of closely related and partly overlapping concepts or foci – such as 'sovereignty' (Hansen and Stepputat, 2006), 'citizenship' (Lazar, 2008), 'democracy' (Paley, 2008), 'authoritarianism' (Krohn-Hansen, 2009), and 'security' (Goldstein, 2010). That is, political anthropology has engaged itself increasingly in the broader academic and political conversation about globally extended uses of categories and perspectives derived from one particular, historically constituted political theory. Today anthropologists study important elements of liberal political philosophy – sovereignty, citizenship, democracy, security, and so on – but instead of approaching that tradition as if it belonged to nature, or reifying it, they deal with it as one more political cosmology, more fruitfully seen as an object of ethnographic-historical

examination than as a platform and framework for generating objective understanding.

Political Anthropology and Violence

Violence has an ambiguous status in Western thought; it often functions as a sign of (the Other's) savagery or primitiveness, but it is simultaneously a legitimate political and state means and a resource. Small wonder, perhaps, that the whole issue of how to account for and write about, or represent, violence is also a difficult challenge for political anthropologists. Classic British political anthropology, with its structural-functionalist ideas, tended to overlook violence or relegate it to the background, or else to rationalize it as functional. Such approaches have evidently helped to maintain the image of anthropologists as romanticizing and utopian. But since the early or mid-1980s, anthropology has seen a renewed and growing interest in violence, including various forms of political violence such as terror, religiously or ethnically inspired violence, and torture (Taussig, 1984; Feldman, 1991; Bloch, 1992; Daniel, 1996; Das, 2006; Kelly, 2011). The last two to three decades' anthropological writings on political violence do not represent a uniform theoretical tradition, though several shared intellectual roots are evident. But they have in common a couple of crucial features. First, these works examine forms of political violence not as incomprehensible 'exceptions' but as 'rules' – that is, not as phenomena outside a daily and ritually constructed symbolic, social and political reality of a particular kind, but, on the contrary, as practices and meanings which belong to (in the sense of being both products and producers of) a context, a society. Second, these studies analyze political violence as historically and culturally constituted processes, and as features of structures of power.

Much of the recent anthropology of the state takes it for granted that construction and reconstruction of the structures of power and knowledge that one calls modern states are inseparable from the deployment of state violence. Hence, one now has an important anthropological literature on phenomena like state terror, state massacres, and civil war (Sluka, 2000; Coronil and Skurski, 2005; Kwon, 2006; Theidon, 2013). This literature documents convincingly that state violence is used and contested in the idiom of a social formation's specific history, that such violence is shaped and energized by each society's myths of authority and identity, by each society's historically constituted cultural formations. The anthropology of modern states and violence demonstrates in addition that, as Walter Benjamin put it, testimonies of civilization are simultaneously testimonies of barbarism (1969, p. 256), and that the validity of Foucault's conceptual distinction between premodern sovereign power and modern disciplinary power is highly questionable – that his underscoring of the rationality of modernity and modern forms of power is too tidy. Foucault's dichotomy obscures too much of what should attract one's attention. The distinction is in practice blind to the violence through which states across the contemporary world reproduce the conditions of their existence, and it prevents us from fully recognizing that the roots of modern forms of power and politics, in both the global center and global periphery, lie in imperialism and colonialism.

Conclusion

In some ways, political anthropology began and ended with a set of processes occurring in a limited number of countries between 1940s and the early 1970s. But as I have shown, this is not the whole story of the anthropological study of political life. The generation of political anthropology in the 1940s and 1950s helped to define and shape views and questions which, albeit in changed forms, still remain part of today's anthropologists' investigations of political phenomena; also today's political anthropologists take it for granted that political processes and political authority can seldom be understood satisfactorily in isolation from explorations of kinship and ritual life – and that the cross-cultural study of forms of power and rule has to remain solidly rooted in examinations of everyday life and in distrust of narratives that reify or naturalize the state.

Yet at the same time, political anthropology has changed substantially. Since the early 1970s, anthropological studies of politics have generally mirrored growing and more complex interest in both meaning and history, including Western imperial histories and world history. Since the 1980s, the anthropology of politics has become less insular or strictly disciplinary, and has gradually opened itself somewhat more to forms of dialogue with interests and perspectives derived from political philosophy, political history, and political science. In so doing, it has recovered and reinforced its critical preoccupation with, and scrutiny of, the dominant categories of contemporary politics – concepts like the state, the nation, sovereignty, citizenship, and democracy.

See also: Advocacy in Anthropology; Anarchism; Colonialism, Anthropology of; Communalism; Conflict: Anthropological Aspects; Factionalism; Governmentality; Militaries and Militarization, Anthropology of; Political Economy in Anthropology; State: Anthropological Aspects; Violence in Anthropology; War: Anthropological Aspects, Historical Development of.

Bibliography

Abrams, P., 1988. Notes on the difficulty of studying the state. Journal of Historical Sociology 1, 58–89.
Alonso, A.M., 1995. Thread of Blood. University of Arizona Press, Tucson, AZ.
Asad, T., 1973. Anthropology and the Colonial Encounter. Ithaca Press, London.
Barth, F., 1959. Political Leadership among Swat Pathans. The Athlone Press, London and Atlantic Highlands, NJ.
Benjamin, W., 1969. Illuminations. Schocken Books, New York.
Bloch, M., 1992. Prey into Hunter. Cambridge University Press, Cambridge.
Clastres, P., 1977 [1974]. Society against the State. Urizen Books, New York.
Cohen, A., 1974. Two-Dimensional Man: An Essay on Power and Symbolism in Complex Society. Routledge and Kegan Paul, London.
Coronil, F., 1997. The Magical State. Nature, Money, and Modernity in Venezuela. University of Chicago Press, Chicago, IL, and London.
Coronil, F., Skurski, J. (Eds.), 2005. States of Violence. The University of Michigan Press, Ann Arbor, MI.
Daniel, E.V., 1996. Charred Lullabies. Chapters in an Anthropography of Violence. Princeton University Press, Princeton, NJ.
Das, V., 2006. Life and Words: Violence and the Descent into the Ordinary. University of California Press, Berkeley, CA.
Das, V., Poole, D. (Eds.), 2004. Anthropology in the Margins of the State. School of American Research Press, Santa Fe, NM.
Evans-Pritchard, E.E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Clarendon Press, Oxford.

Evans-Pritchard, E.E., 1955 [1940]. The Nuer of the southern Sudan. In: Fortes, M., Evans-Pritchard, E.E. (Eds.), African Political Systems. Oxford University Press, London, pp. 272–296.
Feldman, A., 1991. Formations of Violence: The Narrative of the Body and Political Terror in Northern Ireland. University of Chicago Press, Chicago, IL.
Ferguson, J., Gupta, A., 2002. Spatializing states: toward an ethnography of neo-liberal governmentality. American Ethnologist 29, 981–1002.
Fortes, M., Evans-Pritchard, E.E. (Eds.), 1955 [1940]. African Political Systems. Oxford University Press, London.
Foucault, M., 1980. Two lectures. In: Gordon, C. (Ed.), Power/Knowledge. Pantheon Books, New York, pp. 78–108.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Geertz, C., 1980. Negara: The Theatre State in Nineteenth-Century Bali. Princeton University Press, Princeton, NJ.
Goldstein, D.M., 2010. Toward a critical anthropology of security. Current Anthropology 51, 487–517.
Gregory, S., 1998. Black Corona: Race and the Politics of Place in an Urban Community. Princeton University Press, Princeton, NJ.
Gupta, A., 2012. Red Tape: Bureaucracy, Structural Violence, and Poverty in India. Duke University Press, Durham, NC, and London.
Hansen, T.B., Stepputat, F., 2006. Sovereignty revisited. Annual Review of Anthropology 35, 295–315.
Harvey, P., 2005. The materiality of state-effects: an ethnography of a road in the Peruvian Andes. In: Krohn-Hansen, C., Nustad, K. (Eds.), State Formation: Anthropological Perspectives. Pluto Press, London and Ann Arbor, MI, pp. 216–247.
Kapferer, B., 1988. Legends of People, Myths of State: Violence, Intolerance, and Political Culture in Sri Lanka and Australia. Smithsonian Institution Press, Washington, DC.
Kelly, T., 2011. This Side of Silence: Human Rights, Torture, and the Recognition of Cruelty. University of Pennsylvania Press, Philadelphia, PA.
Krohn-Hansen, C., 2009. Political Authoritarianism in the Dominican Republic. Palgrave Macmillan, New York.
Kwon, H., 2006. After the Massacre: Commemoration and Consolation in Ha My and My Lai. University of California Press, Berkeley, CA.
Lazar, S., 2008. El Alto, Rebel City: Self and Citizenship in Andean Bolivia. Duke University Press, Durham, NC, and London.
Leach, E., 1981 [1954]. Political Systems of Highland Burma. The Athlone Press, London.
Paley, J. (Ed.), 2008. Democracy. Anthropological Approaches. School for Advanced Research Press, Santa Fe, NM.
Price, R., 1983. First-time: The Historical Vision of an Afro-American People. Johns Hopkins University Press, Baltimore, MD, and London.
Price, R., 2011. Rainforest Warriors: Human Rights on Trial. University of Pennsylvania Press, Philadelphia, PA, and Oxford.
Radcliffe-Brown, A.R., 1955 [1940]. Preface. In: Fortes, M., Evans-Pritchard, E.E. (Eds.), African Political Systems. Oxford University Press, London.
Scott, J.C., 1985. Weapons of the Weak. Everyday Forms of Peasant Resistance. Yale University Press, New Haven, CT.
Sharma, A., Gupta, A. (Eds.), 2006. The Anthropology of the State: A Reader. Blackwell Publishing, Oxford.
Sluka, J. (Ed.), 2000. Death Squad: The Anthropology of State Terror. University of Pennsylvania Press, Philadelphia, PA.
Taussig, M., 1984. Culture of terror – space of death. Roger Casement's Putumayo Report and the explanation of torture. Comparative Studies in Society and History 26, 467–497.
Taussig, M., 1991. The Nervous System. Routledge, New York.
Taussig, M., 1997. The Magic of the State. Routledge, New York.
Theidon, K., 2013. Intimate Enemies: Violence and Reconciliation in Peru. University of Pennsylvania Press, Philadelphia, PA.
Vincent, J. (Ed.), 2002. The Anthropology of Politics: A Reader in Ethnography, Theory, and Critique. Blackwell Publishers, Oxford.
Yanagisako, S., Delaney, C. (Eds.), 1995. Naturalizing Power. Essays in Feminist Cultural Analysis. Routledge, New York and London.

Political Communication

Pippa Norris, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition article by P. Norris, volume 17, pp. 11631–11640, © 2001, Elsevier Ltd., with revisions made by the Editor

Abstract

Political communication is an interactive process concerning the transmission of information among politicians, the news media, and the public. The process operates downward from governing institutions toward citizens, horizontally in linkages among political actors, and also upward from public opinion toward authorities.

Political communication has always been central to the electoral and policymaking process but since the early 1990s certain important developments have fundamentally altered this process, particularly postwar trends in the mass media moving from the traditional world of newspapers, radio, and television broadcasting toward the Internet. This article outlines alternative interpretations of the nature of these trends and reflects on their consequences for the process of socioeconomic and political development around the globe.

The literature in political communication can be subdivided into three major categories, using a simple systems model of the process illustrated in Figure 1 to distinguish between production, contents, and effects.

Production Processes

Work on the production process focuses on how messages are generated by political actors like parties and interest groups, and then transmitted via both direct channels like political advertisements and indirect channels including newspapers, radio, and television. Many studies have focused on the increased professionalization of political marketing campaigns in the postwar era, including the rise of the class of political consultants, pollsters, advertising executives, and their coterie, and the consequence of this process for strategic communications by political parties and interest groups. A large literature, particularly within Europe, has also studied the changing structure of the news industry, notably the economic basis of the newspaper industry and the legal structure regulating broadcasting and the press. Comparative studies have also commonly analyzed the news culture, especially the values that journalists, broadcasters, and editors employ as 'gatekeepers' in deciding 'what's news,' as well as the organizational structure of newsrooms. Recent work within this area has generated a growing body of research on the rise of new communication and information technologies, and use of the Internet by parties, new social movements, and the news media.

One of the major challenges in this subfield is to widen the scope of comparative research on a systematic basis so that we move from studies of the structure of the political communication process within particular nations to conceptual typologies, broader theories, and empirical generalizations that can be tested across different types of societies. Much of this work has traditionally focused upon postindustrial nations, particularly the USA and Western Europe, although in the 1990s increased attention has been paid to new democracies in Central and Eastern Europe, Latin America, and Asia, as well as, to a lesser extent, to the role of the news media in authoritarian regimes like Burma, China, and Cuba where the role of the free press and opposition movements remain severely restricted.

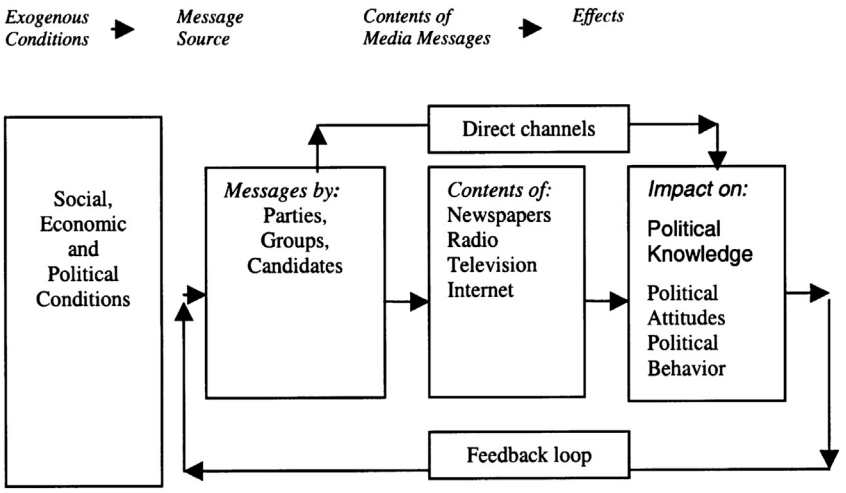

Figure 1 Systems model of political communication process.

Contents

Another related mainstream research tradition has examined the content of the messages that are produced by this process, such as the amount and tone of political reporting presented in television news, the partisan balance in the press, the coverage of election campaigns and particular events, the agenda-setting reporting of policy issues, and the representation of social minorities in the news media. The most common forms of comparison are between different media, for example, differences in the messages conveyed during an election campaign in paid political advertisements, party press releases, newspaper columns, and TV news stories.

Other forms of comparison examine trends over time, such as the coverage of political scandals or social minorities in recent decades. More rarely, collaborative teams have attempted comparison across nations, for example, concerning a specific period or event in selected major newspapers. Work in this tradition has drawn largely upon systematic forms of content analysis of a representative random sample of stories among different media, although alternative qualitative techniques for deconstructing textual and visual messages are also common. Moving beyond systematic description, which is valuable but limited, the main challenges in this area are to relate the content of the messages to either the production process (to examine their possible causes) or to their potential impact (to understand their effects).

Effects

Lastly perhaps the largest body of research, certainly in the USA, has focused at individual level on the study of the potential effects of exposure to different types of political communication messages. The most common method has been to draw upon cross-sectional or panel representative surveys or, more rarely, experimental methods. The key issues have focused upon analyzing the potential impact of exposure to different type of mediated messages (such as watching an advertisement or news story) upon either political knowledge and opinions (such as awareness about an issue, civic knowledge, or recognition of political leaders), political attitudes and values (such as support for a particular party or issue), and political behavior (such as voting turnout). Most work has focused upon the impact of the messages on the mass public or particular subgroups, like women or undecided voters, although some studies have also analyzed the effects upon middle-level elites involved in the policymaking process.

The primary challenges are threefold: (1) to expand generalizations beyond the USA, to see how far they continue to hold up within very different contexts; (2) to move beyond cross-sectional surveys, which cannot determine issues of causality, toward more dynamic designs such as panel surveys and pre–post experimental designs; and (3) to link studies of the individual-level analysis of effects to both what we know about the structure of the news industry and the contents of the messages, generating multilevel and multinational research designs. This work has made considerable progress as the study of political communications has moved increasingly into the mainstream within political science since the 1980s, but nevertheless the subfield remains predominantly US and European, and the process of internationalization in what is increasingly a global society is only starting to produce more systematic cross-national and multilevel research.

Political communication has, therefore, always been central to the electoral and policymaking process but since the 1990s certain important structural developments have fundamentally altered this process, particularly postwar trends in the mass media moving from the traditional world of newspapers, radio, and television broadcasting toward the Internet. The rest of this article outlines alternative interpretations of the nature of these trends and reflects on their consequences for the process of socioeconomic and political development around the globe.

Postwar Trends in Newspapers

Concern about traditional standards of journalism has been fuelled by major changes in the newspaper industry during the postwar era. In the USA the daily press has experienced dwindling readership and sales, especially among the younger generation, a loss of advertising market share to the electronic media, and growing concentration of ownership in larger multiple newspaper chains or a few multimedia conglomerates (McManus, 1994; Baker and Dessart, 1998). All these developments have had a major impact upon the profitability and economic viability of the print sector in the USA, particularly for smaller outlets (De Bens and Ostbye, 1998).

Yet although the demise of newspapers has been widely predicted for decades, we should not underestimate their continued popularity and technological adaptation to new forms of production and distribution. If we compare postwar trends in circulation, controlling for population growth, the evidence shows that sales of the daily press in most postindustrial societies has not been affected by the growing availability of electronic media. The long-term trend in newspaper sales in OECD countries, per 1000 population, has proved fairly stable. Average circulation in all OECD states was 271 newspapers per 1,000 population in 1950, rising modestly in 1980, before subsiding slightly to 263 per 1000 in 1996 (UNESCO). In postindustrial societies, despite the massive surge in the availability of television since the 1950s, about one quarter of the population continues to buy a daily newspaper, and readership figures are even higher. The electronic media have therefore increased the choice and diversity of news outlets and formats, but at the same time they have not killed sales of the printed press. Given growing educational levels and affluence characteristic of postindustrial societies, consumption of news has not proved a zero sum game.

Moreover there are considerable variations among societies in use of the traditional mass media. Figure 2 illustrates the penetration of newspapers and TV sets per 1,000 population worldwide. The pattern confirms how far newspaper and TV penetration levels are associated with basic patterns of socioeconomic development; use of the mass media is related to a more affluent and educated population with greater leisure time. Postindustrial societies tend to be the heaviest consumers of the mass media; nevertheless the pattern also shows considerable variations among these nations. Some outliers like the USA and Canada prove far more television-oriented

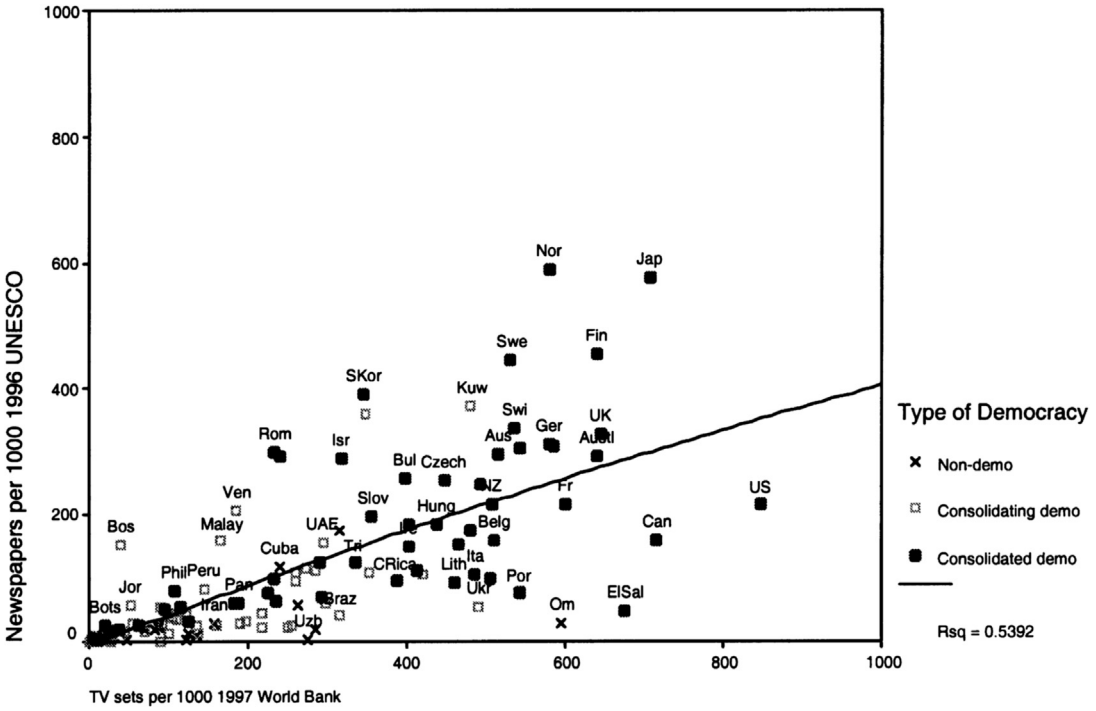

Figure 2 Worldwide distribution of newspapers and TV sets, mid-1990s. (Sources: Newspapers: UNESCO Statistical Yearbook 1997; TV sets: World Bank, 2000. World Development Indicators 2000. www.worldbank.org/data. Type of democracy: Based on the Freedom House classification of political rights and civil liberties.)

than average (along with Southern Mediterranean Europe and much of Latin America), while others nations like Norway and Sweden (along with much of Northern, Central, and Eastern Europe) remain more newspaper-oriented.

It remains unclear whether systematic trends in the newspaper industry have changed traditional journalism, producing an increased focus on crime, sex, and entertainment, as is assumed by some critics. What seems equally plausible across OECD countries is an expansion of both lowbrow and highbrow news media in recent decades, representing a diversification of the market. A recent review of the comparative literature by Kees Brants concluded that the few available content analysis studies provide an ambiguous and sometimes contradictory picture of the growth of 'infotainment' news in different countries, rather than showing a uniform pattern: 'Where for the European countries as a whole we might see a slight tendency towards the popularization of news, there is little evidence that politicians and politics are more dramatically personalized and sensationalized than before' (Brants, 1998).

Brants found that the available content analysis shows a mixed picture of the growth of 'infotainment' news in different European countries, rather than a uniform trend. Frank Esser concluded that there were marked contrasts between Germany, the UK, and the USA in the popularity of tabloid news, and that the nature and degree of competition in a particular media market is the decisive factor explaining the degree of tabloidization (Esser, 1999). Moreover systematic research on long-term trends in British newspapers from 1952–97 found that the amount of political coverage in the tabloid sector had not declined over time, as many critics assume. Instead, the tabloid press in the UK has expanded its coverage of entertainment but also maintained its political news over this period (McLachlan and Golding, 1999; McLachlan, 1999).

Therefore across all postindustrial societies newspaper circulation levels have remained largely stable during the postwar era, yet at the same time the range of papers published in these states has contracted. The number of daily newspapers published in OECD nations fell on average by 15% during the postwar era, from 160 per country in 1960 to 130 in 1996, producing greater concentration of readership in fewer outlets. Many countries have introduced measures to maintain press diversity, on the assumption that we need diverse outlets for an effective civic forum. Antitrust regulations have attempted to insure competition in the ownership of the press, for example, limiting the proportion of cross-media ownership by a single company, administered by fair trade bodies like the British Monopolies and Merger Commission or the German Federal Cartel Office. Other societies like Sweden and Norway have used press subsidies as a policy instrument to protect the financial viability of the more vulnerable sectors of the press (McQuail and Siune, 1998). Countries with provincial and localized newspaper markets like the USA and Germany proved particularly prone to media mergers and acquisitions, reducing pluralism and competition in many cities. Papers in smaller countries like Austria and Belgium also often experienced takeovers or closure because of a limited domestic market and imports from neighboring states with a shared common language.

Concentration of ownership in the hands of a few multinational corporations with multimedia empires has become increasingly common, notably Rupert Murdoch's News International, and the vast holdings of Bertelsmann in Germany, or

Fininvest in Italy (Tunstall and Palmer, 1991; Smith, 1991; Sanchez-Tabernero, 1993). Hence Rupert Murdoch, who started with two small Australian newspapers, built an empire in News Corp. that includes 20th Century Fox films, the Fox TV network, a number of US stations, 50% ownership of Sky TV, a majority interest in the STAR Asian satellite, ownership of *The Sun* and *The Times* in the UK, additional television stations in Latin America, and the book publisher HarperCollins, as well as investments in Internet companies. In the USA, Time Warner's purchase of Turner Broadcasting Systems Inc (including CNN) in 1996 created the largest media firm in the world with strong print, cable, and programming divisions. Walt Disney Company's acquisition of Capital Cities/ABC Inc for $19 billion in 1995 created the second largest media conglomerate with movie, television, cable, and Internet interests, although the purchase proved costly since ABC's profitability moved sharply into the red four years after acquisition. Conrad Black's acquisition of Southam in Canada in 1996 gave his company, Hollinger Inc, control of two-thirds of the newspapers in that country.

What will be the consequences of these developments? Many commentators like Ben Bagdikian (1997, see also Bogart, 1995) fear that media mergers have concentrated excessive control in the hands of a few multinational corporations, who remain unaccountable to the public, and that only greater economic competition can change this situation. Recognition of this problem has arisen from the understanding that economic controls can constrain the media just as significantly as political controls. There is nothing new about this concern, which was often expressed during the interwar era of the press barons when proprietors like Beaverbrook and Rothermere actively intervened to further their political ambitions. Yet others like Robert Picard (1988, 1998) remain more sanguine about recent developments, arguing that we need to distinguish between concentration defined by considering the number of media outlets held by dominant firms and concentration defined by dominance in a clear geographical market. It is the latter – which can harm consumers by producing fewer choices, poorer services, and higher prices – which is important for the availability of alternative sources of political information in a democracy.

Monopolies in the local market for ideas can be harmful for pluralism. Nevertheless it must be recognized that we need to look beyond any single media sector to establish the harmful political effects of concentration, since consumers use and have access to multiple sources of news and information, from newspapers to radio, television, and the Internet. Moreover the trends toward greater concentration are not universal, as some OECD countries have seen a significant expansion in the circulation and range of daily newspapers being published in the postwar era, particularly states like Mexico and Greece where educational and literacy rates have been rising sharply, as well as more modest growth evident in newer democracies like Hungary, the Czech Republic, and South Korea.

Trends in Broadcasting

Just as there are serious concerns about the future of newspapers, so many believe that in recent decades traditional standards of television news and public affairs have come under threat from technological and economic developments. The critical factors transforming broadcasting include the proliferation of channels on terrestrial, cable, satellite, digital, and now broadband services, fragmenting the mass audience; the crisis of identity and funding facing public service television, which once enjoyed a monopoly throughout most of Europe, following the rise of myriad commercial competitors; and lastly the more recent technological convergence with the digitization and compression of images, sounds, and data which has produced a new multimedia environment breaking down the traditional boundaries between telecommunications, the audiovisual industries, and computers.

These trends have affected all OECD countries to different degrees although their impact and timing has been strongly mediated by the existing communications landscape. In the Thatcherite 1980s, deregulation and privatization had the most profound influence on public service broadcasters throughout Western Europe (Achille and Ibanez Bueno, 1994). Following the fall of the Berlin Wall, the transition to democracy in the early 1990s produced an even more radical jolt to public television in Central and Eastern Europe (O'Neill, 1998). Meanwhile in the USA the long-standing rule of the three major networks experienced an equivalent coup d'etat, cut down in the 1980s by myriad competitors on cable and satellite (Kimball, 1994). Despite the long-standing contrasts between the commercially dominant major networks in the USA and public television in Europe, in recent years both have faced a strikingly similar multiplication of outlets and fragmentation of television audiences, raising new concerns about the standards of programming.

Despite the deluge of commercial alternatives, the public channels remain popular; on average across all OECD states public channels maintain a 42% share of the television audience. This varies substantially, however, by country. Public television RAI1 and TVE1 remain market leaders in Italy and Spain, while NRK in Norway and SVT in Sweden had most of the best-rated shows in their countries. In contrast public TV has a far smaller share in some other societies such as PBS (3%) in the United States and NHK (18%) in Japan (Mediametrie, 1999). OECD states can be classified into three major types: those that remain predominately *public* systems (based on an audience share of public channels of 60% and above), *mixed* systems (with a public share of 40–59%) and predominately *private* systems (with a public share of less than 40%). Where there is comparable data, only three OECD nations can be categorized as predominately public (Austria, Denmark and Hungary), 11 represent mixed systems, while 10 can be classified as predominately private systems. Therefore although there has been growing transatlantic convergence over the years, nevertheless broadcasting systems continue to bear the distinct imprint of their origins with radio in the 1920s, and public television continues to remain popular in many countries, rather than being swamped by commercial services.

The Rise of the Internet

Since the early 1990s the most important change to the political communication process has occurred through the rise of

the Internet, particularly in postindustrial societies that are at the forefront of the information society such as the USA, Australia, and Sweden. Networked computing and computer-mediated email have existed for the scientific elite since the early 1960s but the number of users was too small to monitor through mass surveys. The key historic development transforming the Internet into the world's favorite virtual reference library, post office, and shopping mall was a series of rapid innovations: the birth of the World Wide Web (in 1990) and the launch of popular browsers to access materials including those by Mosaic (1993), Netscape Navigator (1994), and Microsoft Internet Explorer (1995) (see Berners-Lee, 2000). Subsequent technological applications, like the easy transfer of .mp3 music files and video formats, and WAP-enabled digital telephony, while representing important innovations, cannot yet claim to have had an impact equal to the basic invention of point-and-click browsers.

As yet no official government statistics on the online population are collected by international agencies like UNESCO and the International Telecommunications Union (ITU), although many indirect measures of technological diffusion are available, including investment in scientific Research and Development, the spread of computer hardware, and the rate of telephone density. The most comprehensive worldwide guide estimating the size of the online population is provided by NUA. This organization regularly monitors and collates survey results conducted by different market research companies in each nation. The surveys ask a representative sample of the public in each country about use of the Internet from home, work, or elsewhere during the previous 3 months. NUA's database 'How Many Online' (http://www.nua.ie) currently collects data from 179 countries, covering 5.7 billion people.

The NUA evidence highlights the dramatic rise in popularity of the Internet in recent years: between 1995 and 2000 the total number of Internet users surged from about 26 to 377 million worldwide, an explosive jump within the space of a few years. The Internet became a truly global phenomenon as more and more users came online from around the world and the proportion of Americans in the online community dropped from 70 to 40% in 1995–2000. Despite this remarkable expansion, about one in 20 of the world's population was online in 2000, with highly uneven diffusion globally.

According to NUA estimates, in Spring 2000 Scandinavia and North America led the world in rates of Internet penetration, with one-third or more of the population online, followed by Western Europe, with about one in 10 online. Central and Eastern Europe, Asia, the Middle East, and South America fall below the world average, all with less than one in 20 online, while minimal diffusion is evident in Sub-Saharan Africa, with only 36 users per 1000 people. In terms of levels of human development, there are stark contrasts between rich and poor nations. Most of the world's online community (87%) lives in highly developed nations, as measured by the UNDP index (see UNDP, 1999). In comparison, the 35 societies classified by the UNDP with low levels of human development, like Nigeria, Bangladesh, and Uganda, contained only 5% of the online population, although home to half a billion people.

A finer-grained comparison of countries ranked by the online population reveals a pattern of widespread adoption in four clusters of societies:

1. Throughout the smaller Nordic social democratic welfare states, especially Sweden, Norway, Iceland, and Finland;
2. In larger Anglo-American and English-speaking nations including the USA, Canada, Australia, and the UK;
3. In the Asian 'tiger' economies of Singapore, South Korea, and Taiwan, as well as Japan; and lastly,
4. In a few smaller European nations with above-average Internet use such as the Netherlands, Belgium, Switzerland, Slovenia, and Estonia.

At the bottom of the national rankings, with less that 0.5% of the population online, few Internet users are found throughout most of the poorer countries of sub-Saharan Africa (with the exception of South Africa), as well as in many states in central Europe, the Middle East, Asia, and Latin America.

Yet at the same time if technological diffusion can be achieved in poorer societies, and it is a big 'if,' then many observers hope that the Internet will provide multiple opportunities for socioeconomic and democratic development. Digital networks have the potential to broaden and enhance access to information and communications for remote rural areas and poorer neighborhoods, to strengthen the process of democratization under transitional regimes, and to ameliorate the endemic problems of poverty in the developing world. With connectivity as the umbilical cord, enthusiasts hope that the Internet will eventually serve multiple functions as the world's favorite public library, school classroom, and medical database; post office and telephone; marketplace and shopping mall; channel for entertainment, culture, and music; daily news resource for headlines, stocks, and weather; and heterogeneous global public sphere. In the heady words of the G-8 Okinawa Charter:

> Our vision of an information society is one that better enables people to fulfill their potential and realize their aspirations. To this end we must ensure that IT serves the mutually supportive goals of creating sustainable economic growth, enhancing the public welfare, and fostering social cohesion, and work to fully realize its potential to strengthen democracy, increase transparency and accountability in governance, promote human rights, enhance cultural diversity, and to foster international peace and stability.
>
> *G-8 Okinawa Charter on Global Information Society 23 July 2000.*
> http://www.g8kyushu-okinawa.go.jp/w/documents/it1.html

The Internet may allow societies to leapfrog stages of technological and industrial development. On the production side, if Bangalore companies can write software code for IBM or Microsoft, and if Costa Rica can manufacture chips for Intel, then potentially entrepreneurs can offer similar services from Malaysia, Brazil, and South Africa. The Internet encourages market globalization: small craft industries and the tourism industry in Bali or the Maldives can deal directly with customers and holidaymakers in New York and London, irrespective of distance, the costs of advertising, and the intermediate distribution chains of travel agents and retail businesses (International Telecommunications Union, 1999: p. 7; Dugger, 2000).

The Internet also offers promise for the delivery of basic social services like education and health information across the globe, a function that may be particularly important for middle-level professionals serving their broader community (Hayward, 1995; Wresch, 1996). Potentially, local teachers or community officials connected to the digital world in Lagos, Beijing, or Calcutta can access the same electronic journals, books, and databases as students at the Sorbonne, Oxford, or Harvard. Distance learning can widen access to training and education, via open universities in India, Africa, and Thailand, and language websites for schools (Arunachalam, 1999). Networks of hospitals and healthcare professionals in the Ukraine, Mozambique, and Stockholm can pool expertise and knowledge about the latest research on AIDS. Peasant farmers using village community centers can learn about storm warnings and market prices for their crops, along with employment opportunities in local towns. Where peripheral regions lack access to the traditional media, the convergence of communication technologies mean that potentially the Internet can deliver virtual local newspapers, streaming radio and television video, as well as other services.

Numerous examples can be cited to show the potential of digital technologies for fostering new opportunities for development in societies around the world (World Economic Forum, 2000). Many South East Asian nations seek to emulate the Japanese model of development in the postwar era of reconstruction, and the knowledge-based economy in Singapore, South Korea, and Taiwan. In Malaysia, for example, the Multimedia Super Corridor has been developed to bring investment from telecommunications, multimedia, and electronics companies, and the production of silicon wafers and software. The corridor has attracted major players such as Microsoft, Sun Systems, and NTT (Japanese telecom). Under the 'Vision 2020' plan, Malaysia now boasts cellular telephone penetration rates of one in every 10 people, more and more wired schools, and 21 Internet hosts per 1,000 people. Revenue generated by the production of information and communication technology goods, like office equipment, telecommunications, and consumer audiovisuals, shows that the USA leads the world but many Asian countries are close rivals, including Japan (second), Korea (third), Singapore (fourth), Taiwan (seventh) and Malaysia (eighth) (OECD, 2000: p. 24, Table 2).

Southern India is most often cited as an important area of software development, producing an estimated $3.8 billion in revenues, with this figure doubling in the last years of the 1990s. Over one-half of India's software services are exported to the USA (OECD, 2000). The Bangalore area has attracted inward investment from many major corporations, not least from the diaspora of the Asian dot.com entrepreneurs thriving in California's Silicon Valley and Cambridge's Technology Park (for a discussion see Yourdon, 1996). In rural Bangladesh many isolated communities lack landline telephones. An innovative program by Grameen Telecom supplies cellular mobile phones to village women, who rent calls in their community to repay the loan and sustain thriving micro enterprises. (By Spring 2000, over 1000 phones had been provided, serving 65000 people, and the eventual target is 40000 phones. See Richardson, 2000, see also http://www.grameenphone.com.). With this service, local communities benefit by direct links to job, weather, and health information, as well as more efficient markets for their produce. Village telecom centers are being developed with email and fax services, along with computer literacy projects in selected schools.

In Central and Eastern Europe, Slovenia, Estonia, and Slovakia have made great strides in moving their populations online, moving well ahead of Portugal, Greece, and Austria in levels of connectivity. Hungary's ambitious Schoolnet program has allowed students in two-thirds of all secondary schools to browse the Web from their classrooms, with extensive teaching resources, interactive discussion forums, events, and competitions (see http://www.SuliNet.hu). In the Baltic, the Estonian government has provided public access points for the Internet throughout the country, using schools, post offices, community centers, libraries, police stations, and health clinics. The program has been highly successful; today more than one in 10 Estonians is online, with personal computer ownership well above average for Central and Eastern Europe (UNDP, 1999: p. 64).

Progress has been slower in Africa, but nevertheless plans have been announced by Global Crossing, Lucent Technologies, and Africa One for an ambitious $1.9 billion project to link up the whole continent by 2002 through a high-speed underwater fiber optic cable, with interior countries connected through terrestrial cables, microwave, or satellite facilities, overcoming many of the current problems of the inadequate telephony infrastructure (Wired News, 2000. For further details see www.AfricaOne.com). Given a high-speed backbone, and market liberalization of telecommunication services, African nations may also be able to 'leapfrog' stages of industrialization through new technology by investing in fully digitized telecommunications networks rather than outdated analog-based systems. Cellular telephony is rapidly expanding as an alternative to conventional network services; the number of subscribers in the OECD region reached almost one-quarter of the population in 1998 (OECD, 2000: p. 81). This growth has had even greater impact in the developing world. In postindustrial economies there were 20 times as many mobile phones in 1998 as there were in 1990, and in developing economies there were 160 times as many, an astonishing rise (World Bank 2000, p. 299). Over a third of all telephone subscribers in Cote d'Ivoire, Cambodia, and Paraguay, for instance, are now connected via mobiles, a far higher proportion than in the USA (International Telecommunications Union, 1999).

The Consequences for Democracy and Democratization

What will be the political consequences of changes in the traditional news media and the subsequent rise of the Internet? In many ways it remains far too early to say (for a detailed discussion, see Norris, 2001). The Internet has generated deeply contested alternative visions about the future. Luddites fear for the worse, but technophiles hope for the better.

The most positive perspective is held by cyber-optimists who emphasize the Panglossian possibilities of the Internet for the involvement of ordinary citizens in direct democracy. In its more utopian manifestations, this view has been dubbed 'technoromanticism' (Coyne, 1999), expressed in earlier eras in

response to Samuel Morse's electric telegraph, Alexander Graham Bell's telephone, and Guglielmo Marconi's wireless radio (see, e.g., Rhodes, 1999; Pursell, 1995; Lubar, 1993). In this account, digital technologies hold promise as a mechanism facilitating alternative channels of civic engagement such as political chatrooms, electronic voting in general elections and for referenda issues, and the mobilization of virtual communities, revitalizing levels of mass participation in public affairs (see Barber, 1998). The use of the Internet by groups and social movements is often believed to exemplify digital politics.

The more utopian visions of the Internet suggest a future society in which virtually unlimited qualities of information become available, civic society flourishes, government decision making becomes more open and transparent, and nation-state borders are eroded as people build virtual communities for work, learning, and leisure spanning traditional boundaries of time and place. Although still in its adolescence, the core transformative capacities of the Internet include its potential for radically shrinking communications and information costs, maximizing speed, broadening reach, and eradicating distance. Compared with radio, television, and newspapers, controlled by editors and broadcasters, the World Wide Web facilitates a virtually unlimited choice of information and communication one-to-one (e.g., via email), one-to-many (e.g., via a personal home page or electronic conference), many-to-one (e.g., via an electronic poll) and, perhaps most importantly, many-to-many (e.g., an online chatroom), with a minimal role for gatekeepers or government censors (see, e.g., Shapiro, 1999; Weare, 2000). Internet messages have the capacity to flow further, faster, and with fewer intermediaries.

As a result many hope that recent developments, especially the spread of new information and communication technologies, will serve to undermine authoritarian regimes, creating a 'dictators' dilemma' in countries like Burma, China, and Cuba (Drake et al.). Leaders in these nations want to facilitate economic development through the Internet and yet at the same time they seek to restrict political access. It is also believed that the Internet will have major consequences for electoral democracies, countries like Russia, Taiwan, and Mexico that are seeking to consolidate democratic political institutions.

Yet as the Internet evolves, a darker vision has been articulated among cyber-pessimists, who regard digital technology as a Pandora's box unleashing new inequalities of power and wealth, reinforcing deeper divisions between the information rich and poor, the tuned-in and the tuned-out, the activists and the disengaged. This account stresses that the global and social divides already discussed mean that Internet politics will disproportionately benefit the elite (see, e.g., Golding, 1996, 1998, 2000). In this perspective, despite the potential for technological innovations, traditional interests and established authorities have the capacity to reassert their control in the virtual political sphere, just as traditional multinational corporations have the ability to reestablish their predominance in the world of e-commerce (see, e.g., McChesney, 1999: pp. 182–5). In this view, so far the potential of the Internet has failed to have a dramatic impact on the practical reality of 'politics as usual,' for good or ill, even in countries at the forefront of digital technologies (Margolis and Resnick, 2000). There are fears that continued inequalities in the spread of new technologies will exacerbate the traditional North–South divide evident in access to the traditional news media like newspapers, radio, and television. As such, while some developing countries may manage to leapfrog earlier technologies in the race towards the information society, other poorer societies may drop even further behind.

The debate between the cyber-optimists and pessimists continues and as the Internet continues to evolve over the first decade of the twenty-first century, then the impact of new technologies will become evident, whether for good (as some hope) or ill (as others fear). What is certainly clear is that political communications via the old world of newspapers and television is in the process of fundamental change and this process holds both threats and promises for the future of socioeconomic and political development.

Editor's Note

A comprehensive and current update of the literature on political communications is Esser, F., Hanitzsch, T. (Eds.), 2012. The Handbook of Comparative Communication Research. Sage, Thousand Oaks, CA. See especially the article by Norris, "Comparing Cross-border Information Flows and their Effects" (pp. 353–369). Another important resource is Norris, P. (Ed.), 2010. Public Sentinel: News Media and Governance Reform. The world Bank, Washinton, DC. Other recent and worthwhile books on the topic include Richard M. Perloff, 2013. Perloff, R.M., 2013. Political Communication: Politics, Press and Public in America. Routledge, New York and Davis, A., 2010. Political Communication and Social Theory. Routledge, London. The leading journal in this area of scholarship is the Taylor and Francis journal Political Communication, the official journal of the political communications section of the American Political Science Association. One highly relevant piece from a recent issue of that journal is Sobieraj, S., Berry, J.M., 2011. From civility to outrage: Political discourse in blogs, talk radio and cable news. Political Communication 28 (1), 19–41.

See also: Advertising: Effects; Advertising: General; Broadcasting: Regulation; Campaigning: Political; Journalism and Journalists; Media Effects; News: General; Political Advertising; Political Money and Party Finance.

Bibliography

Achille, Y., Ibanez Bueno, J., 1994. Les télévisions publiques en quête d'avenir. Presses Universitaires de Grenoble, Grenoble, France.

Arunachalam, S., 1999. Information and knowledge in the age of electronic communication: A developing country perspective. Journal of Information Science 25 (6), 465–476.

Bagdikian, B., 1997. The Media Monopoly, fifth ed. Beacon Press, Boston.

Baker, W.F., Dessart, G., 1998. Down the Tube. Basic Books, New York.

Barber, B.R., 1998. Three scenarios for the future of technology and strong democracy. Political Science Quarterly 113 (4), 573–589.

Berners-Lee, T., 2000. Weaving the Web. Harper Business, New York.

Bogart, L., 1995. Commercial Culture. Oxford University Press, New York.

Brants, K., 1998. Who's afraid of infotainment? European Journal of Communication 13 (3), 315–335.

Coyne, R., 1999. Technoromanticism: Digital Narrative, Holism, and the Romance of the Real. MIT Press, Cambridge, MA.

De Bens, E., Ostbye, H., 1998. The European newspaper market. In: McQuail, D., Siune, K. (Eds.), Media Policy: Convergence, Concentration and Commerce. Sage, London, pp. 7–22.
Drake, W., Kalathil, S., Boas, T.C. Dictatorships in the Digital Age: Some Considerations on the Internet in China and Cuba. iMP: The Magazine on Information Impacts.
Dugger, C.W., 28 May 2000. Connecting rural India to the World. New York Times. http://www.nytimes.com/library/tech/yr/mo/biztech/articles/28india.html.
Esser, F., 1999. Tabloidization of news: A comparative analysis of Anglo-American and German press journalism. European Journal of Communication 14 (3), 291–324.
Golding, P., 1996. World wide wedge: Division and contradiction in the global information infrastructure. Monthly Review 48 (3), 70–85.
Golding, P., 1998. Global village or cultural pillage? The unequal inheritance of the communication revolution. In: McChesney, R.W., Meiksins Wood, E., Foster, J.B. (Eds.), Capitalism and the Information Age: The Political Economy of the Global Communication Revolution. Monthly Review Press, New York.
Golding, P., 2000. Information and communications technologies and the sociology of the future. Sociology 34 (1), 165–184.
Hayward, T., 1995. Info-rich, Info-poor: Access and Exchange in the Global Information Society. Saur, London.
International Telecommunications Union (ITU), 1999. Challenges to the Network: Internet for Development. ITU, Geneva, Switzerland.
Kimball, P., 1994. Downsizing The News: Network Cutbacks in the Nation's Capital. Woodrow Wilson Center Press, Washington, DC.
Lubar, S., 1993. Infoculture: The Smithsonian Book of Information Age Inventions. Houghton Mifflin, Boston.
Margolis, M., Resnick, D., 2000. Politics as Usual: The Cyberspace 'Revolution'. Sage, Thousand Oaks, CA.
McChesney, R.W., 1999. Rich Media, Poor Democracy. University of Illinois Press, Urbana, IL.
McLachlan, S., Golding, P., 1999. Tabloidization in the British Press: A Quantitative Investigation into Changes within British Newspapers from 1952–1997. Communication Research Centre Working Paper #9. Loughborough University, Loughborough, UK.
McLachlan, S., 1999. Who's afraid of the news bunny? The changing face of the television evening news broadcast. Information and Democracy Research Project: Working Paper No. 3. Loughborough, UK: Loughborough University.
McManus, J.H., 1994. Market Driven Journalism. Sage, London.
McQuail, D., Siune, K., 1998. Media Policy: Convergence, Concentration and Commerce. Sage, London.
Mediametrie, 1999. One Television Year in the World. 1998. Mediametrie, Paris.
Norris, P., 2001. Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide. Cambridge University Press, Cambridge, UK.
O'Neill, P., 1998. Post-communism and the Media in Eastern Europe. Frank Cass, London.
Organization for Economic Cooperation and Development (OECD), 2000. OECD Information and Technology Outlook. OECD, Paris.
Picard, R.G., 1988. Press Concentration and Monopoly: New Perspectives on Newspaper Ownership and Operation. Ablex, Norwood, NJ.
Picard, R.G., 1998. Media concentration, economics, and regulation. In: Graber, D., McQuail, D., Norris, P. (Eds.), The Politics of News: The News of Politics. CQ Press, Washington, DC.
Pursell, C.W., 1995. The Machine in America: A Social History of Technology. The Johns Hopkins University Press, Baltimore, MD.
Rhodes, R. (Ed.), 1999. Visions of Technology. Simon & Schuster, New York.
Richardson, D., 2000. Grameen's Telecom's Village Phone Programme in Rural Bangladesh. Telecommons Development Group, ON. http://www.telecommons.com.
Sanchez-Tabernero, A., 1993. Media Concentration in Europe: Commercial Enterprises and the Public Interest. John Libbey, London.
Shapiro, A.L., 1999. The Control Revolution: How the Internet is Putting Individuals in Charge and Changing the World we Know. Public Affairs, New York.
Smith, A., 1991. The Age of Behemoths: The Globalization of Mass Media Firms. Priority Press, New York.
Tunstall, J., Palmer, M., 1991. Media Moguls. Routledge, London.
United Nations Development Programme (UNDP), 1999. Human Development Report 1999. UNDP/Oxford, New York.
UNESCO, 1964. UNESCO Statistical Yearbook. UNESCO, Paris.
Weare, C., August 2000. Technology and politics: Linking the Internet to changes in democratic governance. Paper presented at the International Political Science Association World Congress, Quebec.
Wired News, 5 June 2000. Africa One project targets 2002.
World Bank, 2000. World Development Indicators 2000. www.worldbank.org/data.
World Economic Forum, 2000. From the Global Digital Divide to the Global Digital Opportunity. Proposals submitted to the G-8 Kyushu-Okinawa Summit 2000. www.ceip.org.
Wresch, W., 1996. Disconnected: Haves and Have-nots in the Information Age. Rutgers University Press, New Brunswick, NJ.
Yourdon, E., 1996. The Rise and Resurrection of the American Programmer. Prentice-Hall, New York.

Political Ecology: Anthropological Perspectives

Bengt G Karlsson, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political ecology is a transdisciplinary research field addressing nature – society interrelations, often with a focus on contentions and struggles over land and natural resources. Power asymmetries and social inequalities are critical points of departure, and many scholars in the field pursue a kind of emancipatory engagement with subalterns or marginalized people whose livelihoods depend on the local resource base. Capital accumulation and political economy more generally provide the overall framework for understanding such instances of dispossession and displacement of local communities by global forces of state and market. Political ecology has from its very inception remained a rather loosely defined research field. During the last two decades, the field has expanded rapidly. From an earlier largely rural focus, recent work increasingly engages environmental politics in urban settings and addresses contemporary questions such as climate modeling, genetically modified organisms, food industries, pollution, city planning and infrastructure development. The way ahead is for political ecology is to enter into new conversations with related strands of scholarship, like, for example, with science and technology studies, with studies that concern human – animal relations, and recent work in anthropology on ontology and radical alterity. This article aims to give a short introduction to central aspects of contemporary political ecology and its emergence.

Definition

Political ecology is a transdisciplinary research field addressing nature–society interrelations, often with a focus on contentions and struggles over land and natural resources. A commonly asked question is whose use of, claims to, or perceptions of the environment prevail. Power asymmetries and social inequalities are hence critical points of departure and many scholars in the field pursue a kind of emancipatory engagement with subalterns or marginalized people whose livelihoods depend on the local resource base. Mining, logging, nature conservation, plantations, dams, and various forms of industrial undertakings tend to encroach on and appropriate the land and resources of such communities, and these are also themes that political ecologists have been engaged with from the beginning of the 1970s onward. Capital accumulation and political economy more generally provide the overall framework for understanding such instances of dispossession and displacement of local communities by global forces of state and market.

Political ecology has nevertheless from its very inception remained a rather loosely defined research field, lacking a coherent theory. It can perhaps best be described as a perspective, approach, or research framework. During the last two decades, the field has expanded rapidly. More recent research often has an even more eclectic theoretical approach inspired by, for example, poststructuralist theory (Escobar, 1996; Fletcher, 2010), feminist and gender studies (Rocheleau et al., 1996), and science studies (Forsyth, 2003; Goldman et al., 2011). Focus has increasingly shifted to urban settings (cf. Swyngedouw and Heynen, 2003; Baviskar, 2011), addressing things as different as biomedical interventions, climate modeling, genetic engineering, fair trade commodity chains, waste recycling, machine technology, energy system, transport, and sustainable city designs. The global dimension of politicized environments remains a defining feature, and so also the concern with power and the normative persuasion for a sustainable and equitable future. Political ecology is especially influential within the academic fields of anthropology and geography and has a lot in common with other social science and humanities fields preoccupied with questions about how humans shape and are shaped by the environment (for example, environmental anthropology, human ecology, and environmental history). Political ecology is methodologically rooted in in-depth analysis of specific cases, often combining ethnographic methods with archival research, spatial techniques, and ecological analysis. The findings of the study in a particular place are commonly related to aggregated data on a regional and transnational level.

This article aims to give a short introduction to central aspects of contemporary political ecology and its emergence.

Ecosystems and World-Systems

Welcome to the Anthropocene! Scientists tell us that we have entered a new geological era where human beings have collectively become a force of global environmental change. The planetary boundaries are being pushed to the limits and as a result the most likely scenario is rising temperatures and runaway climatic changes with far-reaching consequences for people all over the world. The overall message is the urgency of the matter: the time for complacency is over. Business as usual is, in other words, not an option any longer. As one might expect, not all agree with such alarmist prognosis (cf. Hulme, 2009). Neither is this necessarily news, since a similar message has been professed by the environmental movement since the late 1960s (Milton, 1993). The difference now, however, is that the science community at large, as well as political leaders and heads of major corporations, seem to agree on the seriousness of the matter: radical measures to mitigate a burgeoning global ecological crisis are required. Yet when the world comes together to act, as in the UN conference COP 15 in Copenhagen 2009, nothing much is achieved. Individual countries push for their specific interests at the expense of common environmental action. For example, China, India, and other emerging economies in the Global South seek exceptions and insist on agreements based

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12215-9

on per capita emissions rather than gross emissions such as the US delegation pushed for in Copenhagen. The later UN climate conferences in Cancun and Doha have seemed equally impotent in crafting a powerful global framework to replace the expiring Kyoto Protocol.

In their recent book *Global Political Ecology*, geographers Richard Peet, Paul Robbins, and Michael Watts note that critical social science has been largely absent from the debates about global climate change, and hence call for political ecology to take up this challenge, calling as it were for research on global climate change "as science, policy and politics" (2011: p. 3). Such calls for the study of global nature might appear far from earlier decades of political ecology research dealing with issues closer to the ground, i.e., with peasants, herders, shifting cultivators, and other subsistence people that struggle to eke out a living in difficult social and environmental circumstances. But, as we will see, there are important continuities.

Anthropologist Eric Wolf was the first to use the term 'political ecology' with its contemporary meaning. In his article 'Ownership and Political Ecology,' a short preface to a special issue of *Anthropological Quarterly* dealing with ownership and inheritance of land in the Alps, Wolf suggested that:

> … we shall need to combine our inquiries into multiple local ecological contexts with a greater knowledge of social and political history, the study of inter-group relations in wider structural fields. (1972: pp. 204–205)

The term political ecology only figures in the title of the article, but as he elaborates, different social arrangements like ownership regimes do not follow from or are adaptions to the particular ecosystem, but need to be understood in relation to wider historical and political processes. Ecological factors matter, but to grasp aspects like access, control, or ownership of land in emerging capitalist relations, the influence of nonlocal elites is critical (ibid.: pp. 201–205). This might appear self-evident today, but anthropology was at this point in time under the influence of various forms of environmental determinism.

Wolf was a student of Julian Steward whose cultural ecology had enormous impact on American anthropology after World War II. According to Steward, the 'culture core' of a society is an adaptation to a specific ecological setting, mediated through the technology or form of subsistence of that particular society (Steward, 1955). He thought that the key question to ask in cultural ecology is how people make a living, and that once one understands the subsistence system, aspects of culture and social organization follow. Another of Steward's influential students was Roy Rappaport. In 1968, Rappaport published the much discussed book *Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People*, where he applies a form of systems theory to capture the interaction of culture, in the form of ritual, and the environment. The local ecosystem is foregrounded; that which set the stage for the unfolding ritual cycle which among other things involved killing a large number of pigs. The pigs required increasing feed and labor input, and this in its turn degraded the local resource base and put enormous pressure on women who were in charge of the pigs. There is also a nutrition cycle at play where the subsequent distribution and consumption of pork served to mitigate deficiencies and with high-protein food in their stomachs people could take up fighting with neighbors, also using the pork as a bargaining stick to secure alliances and support in the warfare. Altogether this is a fascinating account of where society and environment are intertwined and function as a self-regulating system. Rappaport points to a regional system of which the Tsembaga is part, but this is not something that is given any direct importance in the study. In short, one could say that both history and the influence of the surrounding world are ignored. To recall, this was what Wolf insisted while invoking the notion of political ecology. Most commentators on Rappaport's study today would hold that his application of cybernetic systems thinking works to ignore global relations, that it reduces culture to nature, and subsequently turns human agency into systemic responses (cf. Biersack, 2006: p. 7). Even if Rappaport completely ignores Steward and in passing states that he is concerned with "*regulation*, or processes by which systems maintain their structure, rather than *adaptation*" (1968: p. 241, italics mine), *Pigs for the Ancestors* is commonly taken as the cultural ecology text par excellence. This aside, of salience here is the link between earlier forms of ecological anthropology and the political ecology thinking that started to form in the 1970s. Geographer Roderick Neumann goes so far as to claim that "political ecology is constructed primarily through the merger of cultural ecology with political economy" (2005: p. 9).

Let us briefly return to Wolf. Although he did not explicitly pursue political ecology, by insisting on the need for history in anthropology and with his concern with commodity chains spanning the world, he paved the way for later developments in historical political ecology research. Wolf's seminal study *Europe and the People Without History* (*1982*) accounts for the European expansion into the Global South and hence how the Dutch, Spanish, Portuguese, and British conquered people and territories in the pursuit of valuable raw materials. At least for 500 years, Wolf argues, we have had a world-system that affects people and their livelihoods even in the most out-of-the-way places. It is connections that are important; neither societies nor cultures are isolates. Wolf thus opens the critique against the anthropological notion of culture as fixed, internally coherent and homogeneous. This billiard ball idea of culture, Wolf argues, "must give way to a sense of the fluidity and permeability of cultural sets" (1982: p. 387). With such a critique, the functionalist idea of societies as self-regulating systems received a final blow (see Friedman (1974) for an early critique of functionalism in ecological anthropology). The systems thinking, as we will see, was not dismissed but rather relocated on another scale, i.e., moving from a local society adapted to a particular ecosystem to the world with global flows of resources and a joint division of labor.

Immanuel Wallerstein's world-systems approach is a key reference here. The modern world can be divided into countries located in the center or the core, and those in the periphery and the semiperiphery. The system works to redistribute surplus value from raw material exporting countries in the periphery to the industrial countries in the core, according to Wallerstein (1974). What we have then are systemic features that reproduce inequalities and exploitation of people in the Global South. That the former colonies gained independence in the twentieth century did not change this. Environmental

anthropologist Alf Hornborg is a contemporary proponent of historical political ecology. In a series of recent publications along with his colleagues, he elaborates the systemic properties of global capitalism and its environmental legacy (Hornborg et al., 2007, 2012; Hornborg and Crumley, 2006). A key trope in this work is that of a 'metabolic rift,' a separation of social life from nature under capitalism. Hornborg writes:

> Marx was very much concerned with the asymmetric exchange of nutrients and other material resources between town and countryside in 19th-century Europe, which among other things resulted in the impoverishment of rural soils and the accumulation of garbage and sewage in urban areas. The deterioration of European soils prompted capitalist entrepreneurs to exploit phosphates in Oceania and deposits of guano along the west coast of South America, and to develop artificial fertilizers requiring significant input of energy. Marx's crucial observation can today be expanded to the metabolism of the entire world-system (…), some parts of which are ecologically impoverished while other parts are smothered with garbage, air and water pollution and other forms of material overload (…). (2009: p. 247)

If neoclassic economic theory ignores the social and environmental aspects of market transactions, scholars like Hornborg seek to render these visible by looking at the physical properties of traded goods. Concepts like 'embodied land,' 'ecological footprints,' or 'ecologically unequal exchange' are applied for this purpose (ibid.: p. 246). Embodied land, for example, would then be a calculation of the total amount of land required to produce certain goods. Critical political ecology analysis commonly goes against the grain of global ecological governance models that are based on the idea of putting a price on nature and hence encouraging environmental stewardship through market mechanisms. The most well-known example of this is the trade of emission reductions that was introduced with the Kyoto protocol in 1997. From the political ecology horizon, market-based measures enhance the commodification of nature and alienate human beings from the biophysical realities that survival on planet earth ultimately depend on. As put by two of Hornborg's colleagues, "the social metabolism of capitalism is increasingly placed in opposition to the requirements that sustain/support natural cycles and ecosystems" (Clark and York, 2011: p. 26). What this suggests is that to tackle present environmental crises, the root causes have to be addressed. For many within political ecology, one thus has to engage the systemic features of contemporary capitalism.

What has been suggested in this section is that political ecology analysis is historical and multiscalar, looking at resource struggles over centuries spanning the local as well as the global through processes of capital accumulation.

Different Actors, Different Perspectives

Political ecology studies typically begin with a particular problem. In a pioneering study, Piers Blaikie and Harold Brookfield (1987) address the issue of land degradation. They start by noting that the etymology of the word degradation implies 'reduction to a lower rank.' But already here one has to introduce a perspective, a particular use or an interest. "One person's degradation is another's accumulation," as they famously put it (Blaikie and Brookfield, 1987: p. 14). A forest, for example, can be denuded and hence reduced to a lower rank (with detrimental changes in soil, hydrology, and microclimate), yet from the perspective of people involved in agriculture and pastoralism this can well be perceived as improvement. Different systems of production thus avail different 'capabilities' of the land. Measuring land degradation hence becomes a matter of reduced capability vis-à-vis the particular demands that are made upon the land. Such demands, as mentioned, can differ over time (ibid.: pp. 6–13). Based on such reasoning, Blaikie and Brookfield formulate the important proposition that there are "competing social definitions of land degradation" and, following from this, that there are plural perceptions, problem definitions, and rationalities to consider (ibid.: p. 16).

Blaikie and Brookfield propose that the starting point of the analysis of environmental problems should be with the land managers – those that stand in a direct relationship to the land – and that one subsequently should work one's way upward; looking at relations with other land users as well as the other social groups in the region and, on the next level, look at the state and, finally, the world economy and the various institutions operating on this level (ibid.: p. 27). They further introduced the notion of marginalization a process whereby socioeconomic marginalization pushes people into ecologically marginal areas which eventually leads into a spiraling degradation of the landscape and increasing impoverishment – "a cycle of social and environmental degradation," as Paul Robbins aptly sums up their idea (2004: p. 77). With the intensified integration in a capitalist mode of production, resource-rich margins in the global south tend to resort to environmentally destructive resource extraction, which might enrich elite groups but most often undermines the livelihoods of the majority of the local people (cf. Karlsson, 2011). The main focus, in other words, is on the connections between social and ecological change.

The research design suggested by Blaikie and Brookfield still structures much of the work in political ecology. Critical here is to map the differently placed actors involved in resource conflicts. Geographers Raymond L. Bryant and Sinéad Bailey (1997) suggest, for example, the following typology of actors that usually are present in conflicts relating to the environment: the state, multilateral institutions, business or corporate interests, environmental NGOs, and grassroots actors. It is not only the different interests of these actors that are important, but also how their respective claims are being articulated and the very basis of those claims. Here one needs to consider the different world views or perceptions of the environment that might be at play in resource conflicts. The risk here is to impose coherence on actors that are themselves internally differentiated and riven by opposing interests. Earlier political ecology has also been criticized for treating 'the state' or 'the community' as single interests. Arun Agrawal (2005: p. 211) has pointed to another related problem in political ecology analysis and that is the tendency to take actors and interests as 'already given,' hence failing to 'examine how they are made' or emerge in situations of conflict.

A recent example of a political ecology study that is aware of such problems is Paige West's study *Conservation is Our Government Now: The Politics of Ecology in Papua New Guinea* (2006). The book deals with conflicts regarding nature conservation and community development. The main actors involved are the Gimi people, who populate the forests that have become a biodiversity hotspot, and foreign conservationists seeking to preserve precious bird and wildlife habitats and then promote community development. Even if West, like Rappaport, as previously discussed, carries out a village ethnography, the focus here is on outside connections that create the place rather than the internal dynamics of the village and the local ecosystem that preoccupied Rappaport and earlier generations of ecological anthropologists. West introduces a set of outsiders who over centuries have found their way into this out-of-the-way place and who in different ways have taken part in its creation; tradesmen, colonial patrols, missionaries, scholars, tourists, Peace Corps volunteers, aid professionals, pilots, and state officials. The life of the Gimi revolves today around the small airstrip constructed at the fringe of the village, trafficked by occasional planes and helicopters that bring people and goods in and out. West also has a photo of the airstrip on the cover of her book, showing a place surrounded by forest-clad hills. Getting to the village on land requires a good one and half day's walk. The pigs are all gone, slaughtered for good as the Seventh Day Adventists have established themselves in the village, bringing along their particular dietary rules banning pork among other things. West could not even find a single pig in the settlements that constitute the village, which as she puts it, "might seem unbelievable to those familiar with Papua New Guinea subsistence practices" (2006: p. 74). Since the 1970s, people have been engaged in coffee production, which still remains the key source of income (see also West, 2012). The development practices initiated by transnational conservationist NGOs and foreign donors are based on neoliberal ideas, with the market as driver of desirable change. West focuses on two income-generating schemes in the village, both of which have turned the environment into a commodity. One is the Harpy Eagle Project which provide men with income by being guides and helping with logistic support to visiting researchers and tourists interested in the rare and exotic eagle. The other relates to handicraft production, the sale of the traditional string bag that women make out of fiber from tree bark. As West argues, the commodification of the string bags and the harpy eagle alters social relations in the village, not least gender relations, alienates people from the products of their labor and further changes their relationship and identification with the environment (2006: p. 211).

West's study illustrates well the themes and focus of recent rural political ecology writings. Besides ethnography, she builds her story on historical sources and other written documents. The process of commodification of nature and the societal changes this brings about are critical. The multiscalar dimension is foregrounded and the Crater Mountain Wildlife Management Area is represented as a place constituted through interactions of peoples, ideas, and things that span more or less the entire world. Nature or the environment might not be the subject of in-depth ecological analysis, but figures more as an arena for politics and human appropriation. Whether this is to be considered a strength or a weakness is highly contested. Critics argue that political ecology has become too preoccupied with politics at the expense of ecology and ecological processes independent of human action.

Against Political Ecology

In the article 'Against Political Ecology' (1999), ecological anthropologist Andrew P. Vayda along with geographer Bradley B. Walters launched a fierce critique against what they described as the ideologically driven and tendentious research carried out under the rubric of political ecology. In their views, political ecology is a misnomer. If the field once emerged as a reaction against 'ecology without politics' it has now become 'politics without ecology,' and thus better belongs to fields like political science, natural resource politics, or political anthropology (ibid.: p. 168). Above all, they question what they see as a priori judgments and programmatic statements that "political influences" and "especially political influence from outside, from the so-called wider political-economic system" always should be given priority. If environmental change is what one grapples with, Vayda and Walters argue, one has to begin with a careful analysis of the 'actual environmental changes' taking place, and in so doing also be open to the biophysical factors causing change (ibid.: p. 173). Geographer Peter A. Walker follows up on this point asking "Where is the ecology in political ecology?" (2005: p. 73). He nuances Vayda and Walters' critique saying that there are indeed political ecologists dealing with biophysical ecology, but that this is a somewhat threatened concern in more recent scholarship in the field.

Environmental anthropologist Eugene N. Anderson (2011) argues that the political ecology mode of analysis reduces human behavior to the pursuit of power. Particularly troubling is the tendency to invest agency in abstract entities like 'colonialism, neoliberalism, globalization,' apparently assuming that such entities make things happen on their own (ibid.: p. 59). For Anderson, the 'overattention to power' prevents anthropologists to raise other and more important questions, in his case, why certain people have proven success in surviving over millennia 'without wiping out their subsistence base' (ibid.: p. 60).

All scholarly perspectives draw attention to certain things at the expense of others. Theory guides the scholar in the selection of aspects of human or natural life to concentrate on. This is the case not only with political ecology, but scholars within this field have perhaps been more explicit about their theoretical point of departure. When Vayda and Walters claim that political ecology does not address ecology, they evoke a particular understanding of what constitutes nature or the biophysical realm. Even if most political ecologists profess a realist ontology assuming that there exists a world independent of human consciousness (Biersack, 2006: pp. 27–28, Neumann, 2005: pp. 9–10), they would commonly also align with the epistemological position that the researcher has no unmediated access to this reality. Arturo Escobar points to this (1999: p. 3) by suggesting a political ecology that strives for a 'balanced position' that acknowledges the constructedness of nature and yet purporting a realist understanding of nature as having independent existence. How to achieve this is a central problem in Escobar's excellent monograph *Territories*

of Difference: Place, Movement, Life, Redes (2008). Other scholars would probably hold that there are no alternative epistemologies of nature to consider, or for that matter any need for theoretical elaborations regarding the ontological status of the objects of their empirical investigations, that is, 'actual environmental changes.'

Ontological and epistemological debates are a mined terrain that seems to lead us astray from the concrete problems that political ecology research commonly seeks to address. Yet, these are unavoidable issues to address. Leading environmental anthropologist Tim Ingold has for some time pursued a way of thinking that questions both positivist/realist and constructivist approaches, arguing that both place the human observer outside the environment. Ingold (2000) proposes instead a view where human beings and other organisms are understood as already part of or constantly unfolding in the field of relations that constitute their environment, i.e., 'organism-in-the-environment.' Ingold is particularly interested in people's practical engagement with the world, that is, their mode of 'dwelling.' In his most recent book, he rephrases his approach to insist more on 'movement' and 'path' instead of on place, which the notion of dwelling suggests. As he puts it, "wayfaring is the fundamental mode by which living beings inhabit the earth" (ibid.: p. 12). Put differently, what is suggested here is the impossibility to think of the environment as external to the social; instead one must engage with the co-constitution of nature and society. Along similar lines Bruno Latour challenges political ecology for not questioning the conceptual pair of 'politics' and 'nature' (2004). To push the argument, Latour calls for a political ecology that lets go of nature. In doing so, he reminds us that for many of the people anthropologists study there is no equivalent of the Western category 'nature' (ibid.: p. 43). Like Anderson, Latour is also more generally wary of the preoccupation in the social sciences with power. As he puts it, "be sober with power" (2005: p. 260). In anthropology, leading scholars like Marshall Sahlins have similarly decried "the obsession with power" (for the larger debate see Gledhill, 2009). While there is reason to be wary of the mere flagging of power as the key explanatory device, much of recent political ecology analysis does offer sophisticated elaborations on how power operates in particular contexts.

Looking Ahead

In what direction should political ecology go? Political ecology is a multistranded research field and to what extent it will branch out or consolidate its core in the future is hard to tell. At this juncture, there seem to be a set of tensions that energizes political ecology. One such tension is between systemic thinking and poststructuralist or postmodern approaches that insist on difference and fragmentation. Another related one is between a firmer engagement with the biophysical or actual ecological processes on the one hand and the pursuit of narration and the discursive construction of nature on the other. Such tensions have of course a methodological equivalent, for example, a tension between a turn to more solid ecological research or toward discursive analysis. But rather than pitching the future in terms of either pathway, it is perhaps more fruitful to think in terms of political ecology entering into new conversations. An example of this is the recent call for political ecology to engage with science and technology studies (STS) (cf. Goldman et al., 2011). Earlier political ecology analysis tended to simplify science as a homogenous field of knowledge production, mainly attending to "the social interests of science (focusing on who benefits)" (Forsyth, 2011: p. 43). Bringing STS debates into political ecology can provide more sophisticated understandings of the differences and contradictions within science itself, as well as the multiple ways of knowing nature that are in circulation and that inform local resource use as well as state environmental management policy.

Another such conversation would reclaim political ecology as a central part of environmental anthropology. Pushing for this, one could follow the trail of Emilio Moran who suggests that an engaged environmental anthropology that wishes to contribute to debate about climate change has to turn to spatial approaches and apply the new tools of geographic information systems and satellite remote sensing (2011). These techniques allow for a wider reach in, for example, studies of vegetation change or land use patterns. Moran also envisions multidisciplinary research teams, rather than the solitary scholar that dominates in anthropology. As part of environmental anthropology, one could also envision a turn inward, that is, to look at the very structures of life itself and engage recent developments within biology. A critical question here is if a new figure of the human is emerging as a consequence of genetic modifications and creations of transgenetic organisms and landscapes. The work of Donna Haraway is critical here. Take the case of her call to rethink human–animal relations; a 'becoming with' instead of the human as measurement and ultimate goal of all living beings (Haraway, 2008). This can also be pushed to incorporate anthropological research on different ontologies, studies of people who endow nature with agency and consciousness. Anthropologist Julie Cruikshank gives a fascinating account of how her indigenous First Nations interlocutors portray glaciers as sentient beings that listen, smell, and are responsive to human action (2005).

But one can as well see political ecology taking a firmer grip on the political dimension of resource struggles. With the new economic giants like India and China joining the global scramble for land, minerals, and other scarce natural resources, the process of accumulation by dispossession is escalating. In the case of China, the state figures as a key agent of capitalist extraction. This happens alongside the development of new forms of neoliberal governance and control of nature. The World Bank, for example, has reinvented itself as a champion of green governance. In the case of Laos, it has managed to introduce a hybrid construct of large-scale hydropower exploitation along with the creation of a network of national parks and forest reserves. The socialist Lao government, with support from a number of transnational wildlife conservation organizations, has pursued this massive project over the heads of local communities which stand to lose their lands and livelihoods (Goldman, 2005). Global interventions in the name of saving the planet are bound to produce evermore such dispossessions. A critical role of political ecology is to insist that a sustainable future must be built on global social justice.

See also: Agroecology and Agricultural Change; Anthropocene Epoch; Ecology and Health; Land Change: The Merger of Land Cover and Land use Dynamics; Land-Use Indicators; Social Ecology; World Systems Theory.

Bibliography

Agrawal, A., 2005. Environmentality: Technologies of Government and the Making of Environmental Subjects. Duke University Press, Durham.

Anderson, E.N., 2011. Drawing from traditional and 'indigenous' socioecological theories. In: Kopnina, H., Shoreman-Ouimet, E. (Eds.), Environmental Anthropology Today. Routledge, London and New York.

Baviskar, A., 2011. Cows, cars and cycle-rickshaws: bourgeois environmentalism and the battle for Delhi's streets. In: Baviskar, A., Ray, R. (Eds.), Elite and Everyman: The Cultural Politics of the Indian Middle Classes. Routledge, New Delhi.

Biersack, A., 2006. Reimagining political ecology: culture/power/history/nature. In: Biersack, A., Greenberg, J.B. (Eds.), Reimagining Political Ecology. Duke University Press, Durham and London.

Blaikie, P.M., Brookfield, H.C., 1987. Land Degradation and Society. Methuen, London and New York.

Bryant, R.L., Bailey, S., 1997. Third World Political Ecology. Routledge, London and New York.

Clark, B., York, R., 2011. Techno-fix: ecological rifts and capital shifts. In: Hornborg, A., Clark, B., Hermele, K. (Eds.), Ecology and Power: Struggles Over Land and Materila Resources in the Past, Present, and Future. Routledge, London & New York.

Cruikshank, J., 2005. Do Glaciers Listen? Local Knowledge, Colonial Encounters, and Social Imagination. University of Washington Press, Seattle.

Escobar, A., 1996. Constructing nature: elements for a poststructuralist political ecology. In: Peet, R., Watts, M. (Eds.), Liberation Ecologies. Routledge, London.

Escobar, A., 1999. After nature: steps to an anti-essentialist political ecology. Current Anthropology 40 (1), 1–30.

Escobar, A., 2008. Territories of Difference: Place, Movement, Life, Redes. Duke University Press, Durham and London.

Fletcher, R., 2010. Neoliberal environmentality: towards a poststructuralist political ecology of the conservation debate. Conservation and Society 8 (3), 171–181.

Forsyth, T., 2003. Critical Political Ecology: The Politics of Environmental Science. Routledge, London and New York.

Forsyth, T., 2011. Politicizing environmental explanations: what can political ecology learn from sociology and philosophy of science. In: Goldman, M., Nadasdy, P., Turner, M.D. (Eds.), Knowing Nature: Conversations at the Intersection of Political Ecology and Science Studies. The University of Chicago Press, Chicago and London.

Friedman, J., 1974. Marxism, structuralism and vulgar materalism. Man 9 (3), 444–469.

Gledhill, J., 2009. Power in political anthropology. Journal of Power 2 (1), 9–34.

Goldman, M., 2005. Imperial Nature: The World Bank and Struggles for Social Justice in the Age of Globalization. Yale University Press, New Haven.

Goldman, M., Nadasdy, P., Turner, M.D. (Eds.), 2011. Knowing Nature: Conversations at the Intersection of Political Ecology and Science Studies. The University of Chicago Press, Chicago and London.

Haraway, D.J., 2008. When Species Meet. University of Minnesota Press, Minneapolis and London.

Hornborg, A., 2009. Zero-sum world: challenges in conceptualizing environmental load displacement and ecologically unequal exchange in the world-system. International Journal of Comparative Sociology 50 (3–4), 237–262.

Hornborg, A., Crumley, C. (Eds.), 2006. The World System and the Earth System: Global Socio-Environmental Change and Sustainability since the Neolithic. Left Coast Press, Walnut Creek.

Hornborg, A., Clark, B., Hermele, K. (Eds.), 2012. Ecology and Power: Struggles over Land and Material Resources in the Past, Present, and Future. Routledge, London and New York.

Hornborg, A., McNeill, J.R., Martinez-Aleir, J. (Eds.), 2007. Rethinking Environmental History: World-System History and Global Environmental Change. AltaMira Press, Lanham.

Hulme, M., 2009. Why We Disagree about Climate Change: Understanding Controversy, Inaction and Opportunity. Cambridge University Press, Cambridge.

Ingold, T., 2000. The Perception of the Environment: Essays in Livelihoods, Dwelling and Skill. Routledge, London and New York.

Ingold, T., 2011. Being Alive: Essays on Movement, Knowledge and Description. Routledge, London and New York.

Karlsson, B.G., 2011. Unruly Hills: A Political Ecology of India's Northeast. Berghahn Books, New York and Oxford.

Latour, B., 2004. Politics of Nature: How to Bring the Sciences into Democracy. Harvard University Press, Cambridge, MA and London.

Milton, K. (Ed.), 1993. Environmentalism: The View from Anthropology. Routledge, London and New York.

Neumann, R.P., 2005. Making Political Ecology. Hodder Arnold, London.

Rappaport, R., 1968. Pigs for the Ancestors: Rituals in the Ecology of a New Guinea People. Yale University Press, New Haven.

Robbins, P., 2004. Political Ecology: A Critical Introduction. Blackwell Publishing, Oxford.

Rocheleau, D., Thomas-Slater, B., Wangari, E. (Eds.), 1996. Feminist Political Ecology. Routledge, London and New York.

Steward, J., 1955. Theory of Cultural Change: The Methodology of Multilinear Evolution. The University of Illinois Press, Urbana.

Swyngedouw, E., Heynen, N.C., 2003. Urban political ecology, justice and the politics of scale. Antipode 35 (5), 898–918.

Vayda, A.P., Walters, B.B., 1999. Against political ecology. Human Ecology 27 (1), 167–179.

Wallerstein, I., 1974. The Modern World System, Capitalist Agriculture, and the Origins of the European World Economy in the Sixteenth Century. Academic Press, New York.

West, P., 2006. Conservation Is Our Government Now: The Politics of Ecology in Papua New Guinea. Duke University Press, Durham and London.

West, P., 2012. From Modern Production to Imagined Primitive: The Social World of Coffee from Papua New Guinea. Duke University Press, Durham and London.

Wolf, E., 1972. Ownership and political ecology. Anthropological Quarterly 45, 201–205.

Wolf, E., 1982. Europe and the People without History. University of California Press, Berkeley.

Political Economy, History of

Margaret Schabas, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political economy emerged as a discourse in early modern Europe and reached full maturity in the first half of the nineteenth century. It is the progenitor to modern economics, but there are several distinctive characteristics that set it apart. Political economy was a literary rather than a mathematical science; it openly acknowledged its place within moral and political philosophy, and it cultivated a historical stance. It also did not draw a firm demarcation between what came to be known, *c.*1940, as microeconomics and macroeconomics. Political economy is still an acknowledged subfield within economics and political science, albeit with a different set of problems and methods.

Premodern Economic Thought

Aristotle (384–322 BC) analyzed a number of key economic phenomena in the context of his moral and political philosophy. He discerned several essential properties of money, for example, that a monetary price is a numerical ratio of two bartered goods. He ascribed three key functions to money: a measure of value, a store of wealth, and a medium of exchange. Although trade and commerce were part of Athenian life, he discredited the pursuit of profits or the practice of usury, since these rendered money into an end rather than a means to a more virtuous life.

The late medieval commentaries by Christian scholars sustained Aristotle's critical stance toward commerce, in step with Charlemagne's prohibition of all forms of usury in AD 789. To sell a good for more than the customary or just price would, in most cases, require deception and was thus a sin. Likewise, to charge interest on a loan violated the Golden Rule. If you were fortunate enough to have idle funds, then you ought to assist others. To accept more than the principal when the loan was repaid, Aquinas (1225–74) argued, was unjust because money, like wine, is a consumptible good. Money was not yet seen as an instrument for investment, since most loans were for singular events such as a dowry. Nevertheless, the discourse and the law gradually admitted exceptions to these rules. If goods were transported to a place where they were scarce, a higher price might be charged. If in the period of the loan, the lender incurred a loss that would otherwise have been averted, this might justify the payment of interest on the loan. The increasing diffusion of money, and frequent and disruptive debasements by the Crown, prompted the fourteenth-century philosophers Jean Buridan and Nicole Oresme to seek solutions to monetary instability. They downplayed Aristotle's conventional stance on money in favor of the intrinsic discipline offered by the limited supply of gold and silver. Oresme also discerned what came to be known, by the sixteenth century, as Gresham's law, by which the clipped coins drove the full-bodied coins out of circulation. The pamphlet literature of the sixteenth century tracked the transition to the legalization of usury, even though these practices were widespread since at least the thirteenth century, in the form of partnerships, bills of exchange, or annuities.

Early Modern Political Economy

In the early modern period, with the rise of overseas trade and with banking in full sway, attention shifted to stabilizing the supply of money and prices. Nicolaus Copernicus (1473–1543), Navarrus (1491–1586), Jean Bodin (1530–96), Thomas Mun (1571–1641), William Petty (1623–87), and John Locke (1632–1704) are some of the more prominent contributors, but there are hundreds more, and virtually every major philosopher of the period weighed in on economic debates. Copernicus, Navarrus, and Bodin articulated ideas that grew into the quantity theory of money, by which the influx of new money results in a rise in overall prices. Mun, a leading mercantilist and advocate of net exports as the key to England's prosperity, recognized the concept of price elasticity and, more profoundly, market forces that override legal restrictions. His short treatise, *England's Treasure by Forraign Trade* (1664), was in wide circulation till the middle of the next century. It advanced the notion of a virtuous merchant and attended to the economic rivalries between England, the Netherlands, and Turkey to control the trade of cloth and spices. Contrary to the popular view that mercantilists sought to stockpile gold, Mun downplayed the importance of money. Rather, a favorable balance of trade would provide the revenues needed to improve the manufacturing and agrarian sectors, such as the recent rise of silk production in London or the increase of arable land in East Anglia. His views on economic rivalry were boosted by mercantilist John Cary, whose *Essay on the State of England* (1695) also held sway for much of the next century.

Petty was one of the first to devise quantitative measurements of economic phenomena, including population, money supply, and per capita output. Gregory King (1648–1712), William Fleetwood (1656–1723), and Charles Davenant (1656–1714) also contributed to the discipline of Political Arithmetic, compiling data on prices and money supplies. King and Davenant discerned the law of demand, as well as the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03033-6

price elasticity of demand, with a chart of corn prices as a decreasing function of the annual yield. John Graunt (1620–74) and Christiaan Huygens (1629–95) devised population estimates and mortality tables. This in turn inspired leading mathematicians such as Abraham De Moivre (1667–1754) and Jacob Bernoulli (1654–1705) to join concepts of probability to demography, and to spell out the core elements, by 1720, of the binomial theorem or law of large numbers, an invaluable tool for statistical thinking and econometrics that took hold in the nineteenth century.

Petty and Locke, respectively, refined the quantity theory of money by devising the two additional variables of the velocity of money and the level of transactions. Given the dire shortage of specie, they each considered alternative solutions to avert price fluctuations, either increasing the velocity or altering the rhythm of rent payments. Locke was a primary influence on Isaac Newton, then Warden of the Mint, during the massive recoinage of 1696, whereby the clipped coins were summoned and replaced by full-bodied coins at par. Locke also developed a monetary theory of the interest rate, as the price for loanable funds. He urged the dismantlement of the current legal ceiling and the promotion of competitive banking such that the market rate of interest would converge on the natural rate.

Locke broached a labor theory of value that equated the labor input to the relative price. He also wedded the virtue of industry to the right to private property or estate, and motivated the individual rights of association and mobility that in turn rendered the tacit contracts of the market as the locus of liberty. Others at the time also helped forge the idea of what came to be known as the liberal doctrine of political economy, notably Josiah Child (1630–99) and Pierre le Pesant de Boisguilbert (1646–1714), who coined the term laissez faire. Their advocacy of a freer movement of goods and prices was in direct opposition to long-standing regulations and crown privileges. As finance minister to King Louis XIV, Jean-Baptiste Colbert (1619–83) had promoted the liberalization of the internal trade of France while at the same time instilling measures that would protect it against foreign imports. Colbertism was an early example of political economy in the sense in which the power of the nation-state was closely conjoined with policies for economic development in both the agrarian and artisanal sectors.

Eighteenth-Century Political Economy

In the eighteenth century, economic discourse grew dramatically. In France, the center of Enlightenment thought, there were over 3000 publications on economic topics, in the form of essays, pamphlets, dictionaries, encyclopedia entries, and full-fledged treatises. In France and Italy of the 1750s and 1760s, specialized journals, such as the *Journal Oeconomique* or *Il Caffé*, provided forums for economic debates. The crowning achievement, however, came from Scotland. Over 1000 pages in length, Adam Smith's *An Inquiry into the Wealth of Nations* (1776) became the dominant text by the 1790s. But Smith's work did not spring from a vacuum. Some key pre-Smithian texts are John Law's *Money and Trade* (1705), Jean-Francois Melon's *Essai politique sur le commerce* (1734), and Richard Cantillon's *Essay on the Nature of Commerce in General* (1755). Most influential was David Hume's *Political Discourses* (1752), a set of 12 essays on political economy that underwent 11 British editions and was translated into four other European languages, including 12 translations into French alone, all before Hume died in 1776. Ferdinando Galiani's texts, *On Money* (1751) and *Dialogues on the Grain Trade* (1769), were also widely read and cited.

There was no consensus on the name for the discourse on economic phenomena until c.1800, when political economy became the operative term. Antoine de Montchrétien issued a lengthy book entitled Traicté de l'oeconomie politique (1615), but until James Steuart's Inquiry into the Principles of Political Economy (1767), most works had titles with more singular topics such as money, taxes, commerce, or trade. Most deemed their inquiry a science, the science of commerce for example, and positioned it squarely within the broader rubric of the science of man or moral sciences. In Book IV of the Wealth of Nations, Smith introduced the science of political economy, although he defined it as the science of the legislator.

In the Scandinavian and German-speaking regions of Europe, cameralism, the science of state management or applied political economy, flourished throughout the early modern period as an integral part of higher education. Some of the more prominent contributors are Johann Joachim Becher (1635–82), Ernst Ludwig Carl (1682–1743), Carl Linnaeus (1707–78), Johann Heinrich Gottlob von Justi (1717–71), and Joseph von Sonnenfels (1732–1817). Cesare Beccaria (1735–94) was appointed the first chair of Cameral Science at the University of Milan, in 1768. Cameralism preached a strong version of autarky or economic self-sufficiency, including sumptuary laws, and promoted the study of natural history as the means to harness nature's bounty through mining and forestry.

Cameralism is usually sharply distinguished from the mainstream of eighteenth-century political economy, which was increasingly cosmopolitan and forward-looking; global trade and economic expansion implied greater prosperity in the future and, more significantly, a diminution of war and conflict. Predilections for new goods, conveniences, or luxuries, were seen in an increasingly favorable light, as the means to insure trade and thus goodwill between nations. In addition to Hume and Smith, many of the leading Enlightenment philosophers, for example, George Berkeley (1685–1753), Montesquieu (1689–1755), Voltaire (1694–1778), and Jean-Jacques Rousseau (1712–78), also wrote about economic topics, money, wealth, and trade. Two leading savants, Jean d'Alembert and Denis Diderot, secured a place for the science of economics, in their highly influential *Encyclopédie* (17 vols. 1751–65). In addition to those cited, other important contributors to eighteenth-century political economy include Bernard Mandeville (1670–1733), François Quesnay (1694–1774), Antonio Genovesi (1712–69), Condillac (1714–80), and Anne Robert Jacques Turgot (1727–81).

At Versailles, Quesnay founded one of the first schools of political economy, known as the Physiocrats or *les économistes*. Members include Victor Mirabeau (1715–89), Pierre Mercier de la Rivière (1719–1801), and Pierre-Samuel du Pont de Nemours (1739–1817). Physiocracy means the rule of nature and, as a doctrine, it advocated a strong agrarian sector with unrestricted domestic trade. All artisanal activities were deemed sterile, merely transmuting materials, leather into shoes, for

example, but not creating genuine wealth; this could only arise from the gifts of nature, the sun, the rain, and the soil that allows one seed in the spring to become two in the fall. Quesnay devised one of the first models in economic theory, his celebrated *tableau économique*, which depicted an annual diffusion of money and goods between three economic classes, the farmers, the artisans, and the landowners. The *tableau* captured the notion of capital as an advance, producing a net surplus over the course of a year. It also broached an implicit version of the multiplier by which each transaction prompted additional partial expenditures such that the initial capital doubled its output.

Hume made two enduring contributions to monetary theory: the specie-flow mechanism by which prices and money equilibrate over the entire globe; and the causal path by which an unanticipated growth in money due to a trade surplus could stimulate economic growth. Hume also advanced a nonmonetary interpretation of the interest rate and broached the law of the falling rate of profit. Although Hume's account leaves many loose threads, it is all the richer for his insights into human nature and political order, particularly as imbedded in his historical account of commercialization.

Adam Smith (1723–90) canvassed the entire field of economics, including the theories of value and distribution, trade and development, public finance, even economic history and the history of economics. Book I, on value and distribution, establishes labor as the original price of all commodities and capital goods. Labor is estimated objectively in terms of hours of exertion and subjectively in terms of the loss of leisure. Once property rights and money come into existence, capital and labor acquire returns in the form of wages and profits, respectively. The spectrum of wages can be attributed to five circumstances, and the spectrum of profits to two, the disagreeableness and insecurity of the particular business, but in both cases there exists an underlying natural rate of return. Prices are reducible to three components, wages, profits, and rent. The market price, the one observed on a quotidian basis, fluctuates around the natural or long-term price. Smith also lays out his celebrated account of the division of labor, in the sense of division within a single trade, the production of metal pins, for example, into specific tasks to glean economies of scale. He also emphasized the importance of specialization by trade to take advantage of natural talents and thus secured one of the more enduring functional relations in economics, namely, that the division of labor by trade is a function of the size of the market.

Books II and III focus on capital accumulation and economic growth, first within a nation and then with respect to global trade. Smith observes that, for most of the past, human institutions have distorted the natural progress of opulence. Instead of the agrarian surplus of a town sparking a manufacturing surplus and then overseas trade, growth was most manifest in the trading cities of Venice, Hamburg, or Amsterdam that then prompted manufacturing and finally more agriculture to feed the growth in local population. Books IV and V expand out to comparative political economy, domestic and foreign trade policies, the provision of public works, and fiscal policy. Building on Hume's account of trade, Smith delivers on the question as to why some countries are poor and others rich, and offers a dynamical account that peers far into the future.

Although mentioned only once in the *Wealth of Nations*, the invisible hand has come to be synonymous with Smith. It draws on but softens Mandeville's insight that private vices have public virtues; rather, self-interested actions not only yield unintended and beneficial consequences, but also may yield better outcomes than actions that profess the public good. The invisible hand is also mentioned once in Smith's only other book, the *Theory of Moral Sentiments* (1759), and is used in a very different light, to motivate the sense in which there is distributive justice even in a world of rich and poor. For Smith, the real purpose of life is to achieve peace and tranquility and, for the most part, this is an inverse function of wealth and rank. The poor sleep more soundly than the rich and, as long as the basic necessities of life are met, have an equal if not superior claim on a happy life. Wealth, for Smith, is a futile and hollow end, born out of a restless desire for trinkets and gadgets. But because we are deceived into thinking that wealth brings happiness, we work hard and thus keep the world in motion. Wealth also garners the approval of others, and this is part and parcel of our more fundamental disposition to cultivate fellow feeling. Arguably no one has surpassed Smith in weaving such a penetrating account of human nature into such an expansive theory of material well-being and the rise of modern commerce.

Nineteenth-Century Political Economy

Smith's ideas were extended and refined by, most notably, Thomas Robert Malthus (1766–1834), Jean-Baptiste Say (1767–1832), David Ricardo (1772–1823), and John Stuart Mill (1806–73), a tradition now known as the Classical school of political economy. The Classical economists emphasized a cost-of-production (labor) theory of value and the competing claims of the three economic classes (laborers, landowners, and capitalists) over the annual product, aptly depicted as the fall harvest of corn (wheat). The capitalists made a justifiable contribution by bringing foresight and prudence into a world that was otherwise myopic. The so-called iron law of wages proclaimed the futility of raising wages much above subsistence, since this would increase population and within 10 years increase the supply of labor thus bringing wages back to their original level. The only ways to avoid this were either to engage in the moral restraints that Malthus advocated to keep population down, or to accumulate capital, since that would increase the demand for labor.

For Ricardo, the price of the wage good, corn, was determined at the margin of cultivation, where the last units of labor and capital were deployed. Rent, which Ricardo defined as a return to the powers of the soil, was a pure residual and did not, as Smith had posited, enter into the formation of prices. Also contrary to Smith, Ricardo argued that an increase in wages was not inflationary. Because wages and profits are inversely related, an increase in wages would prompt the movement of capital from the labor-intensive (agriculture) to the capital-intensive (manufacturing) sectors, unleashing an adjustment in relative prices but not in the overall price level. Population growth implied the cultivation of inferior grades of land and diminishing returns implied the rise in the price of corn. This meant a long-term tendency for the profit rate to

fall, even if the capital stock continued to accumulate. Ricardo's analysis was exacting and cemented the method of abstract deductive reasoning. Ricardo also worked out the principle of comparative advantage in trade, predicated on the international immobility of capital. Even if a country such as Portugal can produce both cloth and wine more cheaply than its trading partner, England, it is more efficient if Portugal produces just the one good, wine, in which it has the greatest comparative advantage (lower labor costs of production) and trades that for English cloth, the good for which England has the least comparative disadvantage. Mill extended the analysis by introducing nation-specific patterns of demand and the interference effects of foreign exchange rates.

For the Classical economists, pricing came about after distribution, because the returns to the three factors of production, land, labor, and capital, were part of the natural order. Malthusian laws of population growth and diminishing returns were seen to be inexorable, the result of forces deeper than any human action or process of deliberation. Even money, via the specie-flow mechanism, was seen to have a motion that transcended human agency. This in turn bolstered the doctrine of laissez faire, most championed by Say, who argued against the possibility of persistent gluts in production, and thus for the efficiency and justice of market mechanisms. Say's *Traite d'economie politique* (1803) was one of the most widely read texts of the first half of the nineteenth century and became the target of the socialist movement that emerged at this time, led by Henri de Saint-Simon (1760–1825), Charles Fourier (1772–1837), and Pierre-Joseph Proudhon (1809–65).

Classical economists worried about the onset of the stationary state, whereby the profit rate fell to the point that capital accumulation would halt, and overpopulation result in a world of 'standing room only.' Scarcity, doom, and gloom were prevalent motifs of the dozens of texts by the Classical economists, notwithstanding the dramatic growth of the Western European economy throughout the nineteenth century. They clearly underestimated the potential for birth control, for technological inventions and innovations, both in the form of labor-saving and capital-saving procedures, and for the dramatic returns in the agrarian sector as new techniques of plant breeding, crop rotation, fertilization, and transportation took hold. The reasons for their pessimism are still not fully understood.

Early Neoclassical Economics

Despite numerous appreciations by the Classical economists of the importance of utility in the determination of prices, it was only in the 1870s that mainstream economic theorists discarded the labor theory of value and embraced the idea of marginal utility as the key determinant of a price. The leading instigators of this so-called marginal revolution were, in Britain, William Stanley Jevons (1835–82), Francis Ysidro Edgeworth (1845–1926), and Alfred Marshall (1842–1924) and, on the Continent, Léon Walras (1834–1910), Vilfredo Pareto (1848–1923), and Carl Menger (1840–1921). In the United States, John Bates Clark (1847–1938) and Irving Fisher (1867–1947) were at the vanguard. Inspiration came from Jeremy Bentham (1748–1832), who believed that utility could be quantified and all actions reduced to a calculus of pleasure and pain. Insofar as individuals strive to maximize utility and minimize labor, wealth is construed as a mental not material entity, and thus unbounded. Herein lay the germ of the more optimistic tenor that characterizes the Neoclassical period.

The most salient feature of economic theory since the 1870s is its systematic use of mathematics, particularly the calculus. Much of the original momentum to transform literary political economy into a mathematical discipline was drawn from the aspiration to emulate physics. John Stuart Mill was the first to argue that political economy must adopt the axiomatic-deductive method of Newton. What motivated the early Neoclassical economists to develop a mathematical theory was the realization that the phenomena themselves were quantitative, and that the mind deliberated in terms of infinitesimal units and thus the calculus. They drew inspiration from new developments in logic and psychology, as well as the techniques of constrained maximization as worked out in Lagrangian and Hamiltonian mechanics. Walrasian economics became the dominant theory by the 1920s, and by the 1950s, the *Foundations of Economic Analysis* (1947) by Paul Samuelson (1915–2009) insured that every future economic theorist must first acquire training in mathematics.

Political economy gained considerable standing over the course of the nineteenth century, having traditionally been part of the study of jurisprudence and moral philosophy. In 1805, Malthus was appointed to the first post as Professor of Political Economy, at the East India College, and in the 1820s, chairs were established at the more renowned universities at Cambridge, Oxford, and London. In France, Charles Gide (1847–1932) led the establishment of chairs in political economy at every French university by 1878, and founded the *Revue d'économie politique* (1887). Later efforts at discipline formation were made with the discarding of the adjective 'political' from the name of the discipline, as evident by the publication of Marshall's *Principles of Economics* (1890), along with the formation of the British Economic Society and the *Economic Journal* in 1891. Marshall's chosen successor at Cambridge, Arthur C. Pigou (1877–1959), founded welfare economics in the 1920s, and insisted that economics was now no longer tied to moral or political philosophy.

American economists garnered professional momentum with the founding of the American Economic Association in 1885, and the creation of three journals, the *Quarterly Journal of Economics* (1887), the *Journal of Political Economy* (1892), and the *American Economic Review* (1891). Thorstein Veblen (1857–1929), John R. Commons (1862–1945), and Wesley Clair Mitchell (1874–1948), among others, founded a school known as Institutionalism that left a profound imprint on subsequent American economics. Mitchell and his successor, Arthur F. Burns (1904–87), developed the National Bureau of Economic Research (founded in 1920), into the spawning ground for many of the leading American economists of the mid-twentieth century, including Burns's student Milton Friedman (1912–2006) who became the dominant voice of neoliberalism.

Marshall's most famous student, John Maynard Keynes (1883–1946), decried the problems of persistent unemployment and inequalities of wealth and income. The solution lay

with increased government intervention in the form of public works, severe inheritance taxes, and reformed fiscal policy. Keynes envisioned a world of developed economies run by decentralized institutions, whereby individuals would no longer be confined by economic necessity nor driven by greed, where they could enjoy leisure, artistic improvement, and enter the 'ambit of altruism.' In addition to his social philosophy, Keynes's *General Theory of Employment, Interest and Money* (1936) remains one of the most influential books in the history of political economy for its theoretical integration of a number of key tools and concepts, for example, the marginal propensity to consume and to save, the multiplier, the saving-investment function, and the demand theory of money. Many of these tools, including involuntary unemployment and business cycles, were worked out with his fellow economists at Cambridge, Ralph Hawtrey (1879–1975), Dennis Robertson (1890–1963), Richard Kahn (1905–89), and Nicholas Kaldor (1908–86), or can be found in the earlier work of Knut Wicksell (1851–1926). Piero Sraffa (1898–1983) and Joan Robinson (1903–83), both colleagues of Keynes, cultivated neo-Ricardian theories that drew attention to capital rigidities and imperfect competition. *Value and Capital* (1939), by Oxford economist John Hicks (1904–89), made canonical both Walrasian general equilibrium theory and the graphical tools for Keynesian economics that mapped functions for investment and money demand in terms of the interest rate and aggregate nominal income (the IS-LM diagram).

Although the Neoclassical theory became the dominant position by the 1930s, there were many alternative streams of thought, some of which still survive to this day. In Germany, for example, such thinkers such as Wilhelm Roscher (1817–94), Gustav von Schmoller (1838–1917), Werner Sombart (1863–1941), and Max Weber (1864–1920) were adamantly historical and sociological in their approach to the study of economic phenomena. Others followed Karl Marx (1818–83), such as Karl Kautsky (1854–1938) and Rosa Luxemburg (1871–1919), emphasizing class conflict, and the rights of labor to its full product. Marx's major treatise in political economy, *Das Kapital* (vol. I, 1867), was left unfinished; not until Friedrich Engels completed Volume III in 1894 was there sufficient material to spawn a distinct school. And although Marx was much indebted to Ricardo, Marxist economics is best viewed as a twentieth-century pursuit. It has become increasingly analytical and mathematical, but still adheres to some of the distinctive principles of exploitation, alienatation, and increasing immiseration. Eastern Europe produced several economists who made important analytic contributions to mainstream economics, notably Eugen Slutsky (1880–1948), Nikolai Kondratiev (1892–1938), Oskar Lange (1904–65), and Wassily Leontief (1906–99), who emigrated to the States. In England, new schools of thought were founded that were either historical or socialist in orientation. Under the inspiration of Thorold Rogers (1823–90) and Arnold Toynbee (1852–83), economic history secured a firm toehold in the university curriculum, and matured in the hands of William Ashley (1860–1927), G.D.H. Cole (1889–1959), and Phyllis Deane (1918–2012). The Fabians, notably Beatrice Webb (1858–1943) and Sidney Webb (1859–1947), founded the London School of Economics in 1895 and were instrumental in steering Britain toward the welfare state.

In Vienna, Menger attracted a number of students, including Friedrich von Wieser (1851–1926) and Eugen von Böhm-Bawerk (1851–1914), giving rise to a full-fledged Austrian School. The second generation of Austrians, Ludwig von Mises (1881–1973), Joseph A. Schumpeter (1883–1950), and Friedrich von Hayek (1899–1992), moved to Britain or the United States, and in the process created an interesting blend of Austrian rationalism with American libertarianism. Schumpeter's *Capitalism, Socialism and Democracy* (1942) and Hayek's *Road to Serfdom* (1944) are two of the most popular nontechnical works of twentieth-century political economy. Like Keynes, they grappled with the fate of capitalism as an economic system and proposed political measures to sustain the status quo.

If economics is what economists do, then the 1930s and 1940s constitute the most pivotal decades by which political economy retreated from center stage and transformed into an interdisciplinary subfield within the academy, focusing on global and comparative national development. It has been adopted by faculty across the social sciences, and given new meaning by, for example, the sociologist Immanuel Wallerstein, the political scientist Elinor Ostrum, or the economist Doran Acemoğlu. There is no simple explanation for this shift, but one critical move can be attributed to the fundamental divide undertaken by Keynes and the early founders of econometrics, Ragnar Frisch (1895–1973), Jan Tinbergen (1903–94), and Trygve Haavelmo (1911–99). They segregated macroeconomics (the study of national output and growth) from microeconomics (the study of prices and markets) and political economy took on a more specialized and subservient place in the configuration of economic theory.

See also: Capitalism; Economics, History of; Macroeconomics, History of From 1933 to Present; Microeconomics, History of.

Bibliography

Arnon, A., 2011. Monetary Theory and Policy from Hume and Smith to Wicksell: Money, Credit, and the Economy. Cambridge University Press, Cambridge.

Backhouse, R., 2002. Ordinary Business of Life: A History of Economics from the Ancient World to the Twenty-first Century. Princeton University Press, Princeton.

Faccarello, G. (Ed.), 1998. Studies in the History of French Political Economy: From Bodin to Walras. Routledge, London.

Hollander, S., 1987. Classical Economics. Blackwell, Oxford.

Hont, I., 2005. Jealousy of Trade. Belknap Press, Cambridge, MA.

Hutchison, T., 1988. Before Adam Smith: The Emergence of Political Economy, 1662–1776. Basil Blackwell, Oxford.

Kaye, J., 1998. Economy and Nature in the Fourteenth Century. Cambridge University Press, Cambridge.

Larrère, C., 1992. L'invention de l'économie au XVIIIe siècle. Presses Universitaires de France, Paris.

Medema, S., 2009. The Hesitant Hand: Taming Self-interest in the History of Economic Ideas. Princeton University Press, Princeton.

Medema, S., Boettke, P. (Eds.), 2005. The Role of Government in the History of Economic Thought. Duke University Press, Durham and London.

Mirowski, P., 1989. More Heat than Light: Economics as Social Physics, Physics as Nature's Economics. Cambridge University Press, Cambridge.

Morgan, M.S., Rutherford, M. (Eds.), 1998. From Interwar Pluralism to Postwar Neoclassicism. Duke University Press, Durham and London.

Murphy, A.E., 2009. The Genesis of Macroeconomics: New Ideas from Sir William Petty to Henry Thornton. Oxford University Press, Oxford.
Redman, D.A., 1997. The Rise of Political Economy as a Science. MIT Press, Cambridge, MA.
Reinert, S.A., 2011. Translating Empire: Emulation and the Origins of Political Economy. Harvard University Press, Cambridge, MA.
Roncaglia, A., 2005. The Wealth of Ideas: A History of Economic Thought. Cambridge University Press, Cambridge.
Schabas, M., 2005. The Natural Origins of Economics. University of Chicago Press, Chicago.
Screpanti, E., Zamgani, S., 2005. An Outline of the History of Economic Thought. Oxford University Press, Oxford.
Vaughn, K., 1994. Austrian Economics in America: The Migration. Cambridge University Press, Cambridge.
Wakefield, A., 2009. The Disordered Police State: German Cameralism as Science and Practice. University of Chicago Press, Chicago.
Wennerlind, C., 2011. Casualties of Credit. Harvard University Press, Cambridge, MA.
Winch, D., 1996. Riches and Poverty: An Intellectual History of Political Economy in Britain, 1750–1834. Cambridge University Press, Cambridge.

Political Economy in Anthropology

Joel S Kahn, The University of Melbourne, Parkville, VIC, Australia; and La Trobe University, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term *political economy* has been used by anthropologists since the 1970s, particularly in North America, to signal a shift in the focus of anthropological analysis away from isolable local communities and toward large-scale regional political/economic systems, along with the adoption of broadly Marxist, or at least materialist, theoretical frameworks. In this article, the field is also taken to include the field of Marxist anthropology based mainly on the work of French structural Marxists, which emerged in the same period. Although both Political Economy and Marxist anthropology have fallen out of favor in recent years, many of the concerns of the 1970s have returned in the more recent critical engagement with globalization and neoliberalism and in new work on the anthropology of money and finance and the ethnography of capitalism.

Introduction

In the history of economics, political economy is a widely accepted designation for a general style of analysis of national wealth and international trade that emerged in the early nineteenth century in the writings of classical economists like Adam Smith and David Ricardo. The term was borrowed by anthropologists in the 1970s, and mainly in North America, to signal a shift in the focus of anthropological analysis away from isolable local communities and toward large-scale regional political/economic systems, along with the adoption of broadly Marxist, or at least materialist, theoretical frameworks. Particularly in North American anthropology, political economy took 'the form of studying the effects of capitalist penetration upon these communities' (Ortner, 1984: 141).

Doubtless, American efforts in this direction were distinguished by their debt to the work of dependency theorist Andre Gunder Frank and that of the sociologist Immanuel Wallerstein on the capitalist world system, as well to the influences on them of earlier materialist trends in American anthropology. However, it would be misleading to see this shift as being limited to North American anthropology in the 1970s. At least some anthropologists were already analyzing the effects of such broader links well before 1970, while contemporaries outside North America were often well in advance of their American colleagues in the development of Marxist approaches to anthropological materials.

Notably in France and Britain, and often under a rather different set of theoretical influences, most notably from contemporary Marxist theorizing, anthropologists also began to examine the relationships between capitalism and what were often described as precapitalist modes of production or social formations in Asia, Africa, and Latin America.

Any assessment of the general move on the part of anthropologists of the 1970s to engage with global economic factors and their influence on the local groups they were more used to studying – and to do so from a broadly Marxist perspective – must then examine developments outside American anthropology as well, even if they may not have attracted the label of political economy.

Political Economy, Anthropology, and Global Capitalism

The development of political economy in North American anthropology was linked to a more general intellectual and cultural challenge in America's imperial expansion, itself stimulated by opposition to the war in Southeast Asia. Consequently, political economy in the American context was on the one hand theoretically somewhat eclectic and, on the other, concerned primarily with issues of global capitalism. Given the tendency of anthropologists to engage with regions, societies, and cultures outside or on the margins of North Atlantic capitalism, much of the energy of political economists in America was devoted to analyzing the impact of capitalism on individuals and communities on what were sometimes called, following Wallerstein, the 'peripheries' of the world capitalist system. This global emphasis in their approach to more traditional anthropological concerns, moreover, made American political economists less likely to classify these marginal individuals and groups as precapitalist. They were consequently also less interested than their French counterparts in any attempt to develop a theory of modes of production that was general and comparative, seeing it to be of little relevance to the concerns of social and cultural anthropologists doing field studies in the present.

As already noted, materialist analyses were well established in American anthropology before the emergence of political economy, although it could hardly be argued that these analyses were based on any genuine engagement with, or even sympathy for, the work of Marx himself. This materialist current is commonly associated with the name of Leslie White, who is assumed to have been familiar with work in the Marxist tradition, if not with Marx's own writings, although he was prevented from acknowledging the link explicitly because of a generalized American hostility to Marx. Marvin Harris was an influential advocate of materialist analyses from the 1960s, although unlike White he was openly hostile to Marxist theorizing, preferring a form of techno-ecological determinism that was rather close to that of White.

Closer to the interests of the new generation of political economists was the work of Julian Steward, who had

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12132-4

formulated a more open multilinear theory of evolution. He also headed the Puerto Rico project, which was unusual for the ways it tackled issues to do with social change and colonial history as central concerns of anthropology.

But more than anyone else, a group of anthropologists studying in New York in the 1940s – including subsequently well-known figures such as Eric Wolf, Sydney Mintz, Elman Service, and Stanley Diamond – became the key figures in anthropology's direct engagement with colonialism and global capitalism from within a materialist perspective, an engagement that subsequently attracted the label *political economy*. Wolf and Mintz, in particular, became increasingly concerned with the implications of European global expansion for the anthropological analysis of peasant communities in the Caribbean and South America. Both wrote anthropological monographs that were unusual for their discussion not just of local systems of culture, kinship, and social organization but also of the formative links established in the long history of European colonization between local communities and the centers of the world economy in Europe and North America. Later, both also produced influential general works concerning the development of a world capitalist system, paying particular attention to the interconnections established in this process between Europe and North America on the one hand, and on the other hand, the only apparently isolated peoples on its periphery (Wolf, 1982; Mintz, 1985).

Others, both from the New York group and outside it, contributed, if less directly, to the development of political economy. Unlike many materialists in American anthropology, Stanley Diamond made important contributions in Marxist theory, founding the journal *Dialectical Anthropology* in the 1970s and working on a critical/phenomenological Marxism that contrasted rather sharply with contemporary Marxist theorizing in France.

Mention must also be made of Eleanor Leacock, who made her own contributions to the development of a Marxist anthropology through the pursuit of the themes of kinship, family, and gender raised in Engels' *Origins of the Family*, as well as to political economy in some important work on the impact of the fur trade on indigenous North Americans.

Two further influences served to broaden the popularity of such analyses in American anthropology by the early 1970s. First, as noted, there was the publication of the very influential early works of the dependency theorist Andre Gunder Frank and, shortly thereafter, the works of sociologist Immanuel Wallerstein (see Frank, 1969; Wallerstein, 1974). What these works brought, in particular, was a more sophisticated analysis of the imperatives of the global economy and the compatibility, indeed even the functionality, within it of apparently noncapitalist economic and political forms. Anthropologists like Jane and Peter Schneider, Katherine Verdery, Gavin Smith, Carol Smith, William Roseberry, and others associated with the new political economy school did not necessarily accept wholeheartedly the capital-centrism of these more global approaches, but they did enter into a sustained and fruitful debate with them over the global versus the local determinants of socioeconomic systems in the so-called Third World as well as in the European semiperiphery.

The second formative influence on materialist anthropology in America in this period was, of course, the growing movement against America's military intervention in Indochina. This contributed greatly to the radicalized political and intellectual atmosphere on North American campuses, where a new generation of anthropologists was being trained. But it also has specific implications for anthropology, raising as it did the possibility that the discipline itself could be implicated in the project of empire. In 1968, in an article that generated a heated debate, one anthropologist went so far as to describe anthropology explicitly as a 'child of imperialism' (Gough, 1968). Similar concerns led to significant conflicts within the professional anthropological body itself (the American Anthropological Association), which debated the involvement of anthropologists in covert operations at its annual meetings.

The French Marxist Heritage

If in North America the main parameters of political economy were laid down by the engagement of anthropologists with a predilection for materialist explanations, with issues arising out of the development of a world capitalist economy, in Europe the situation was somewhat different. This is not to say there was no interest in the relationship between global capitalism and local socioeconomic organization. On the contrary, this central concern in American anthropology found important parallels, especially in France and Britain. There, however, the debates were informed by a far more rigorous attempt to develop a Marxist theory appropriate to the analysis of anthropological materials, a level of theoretical concern largely absent in contemporary American anthropology. And the center of this reworking of Marxism was to be found in France.

Developments in French anthropology were subject to three distinctive influences. The first was the opening up of the Marxist left, spurred in part by the developing Marxist critique of Stalinism which, among other things, produced a space for Marxist intellectuals, both inside and outside the Communist Party, to criticize the kinds of mechanical and evolutionist interpretations of Marx that until then were the only versions permitted by Marxist parties. Second, and related to this, the Sino-Soviet conflict inspired at least some Marxist intellectuals and political activists to search for alternative models of socialism and of the ways in which it could be achieved. Third, French anthropology was dominated by the imposing intellectual figure of Claude Lévi-Strauss, so it is not surprising that when French anthropologists turned to the work of Karl Marx for inspiration, they developed a structural version of Marxism that was highly critical of the existential and humanist Marxism whose best-known exponent was Jean-Paul Sartre. It is not surprising that the two main schools in French Marxist anthropology – the first associated closely with the theories of the communist philosopher Louis Althusser, the second with the structural Marxism of Maurice Godelier (a student of Lévi-Strauss) – although critical of mainstream 'mechanical Marxism,' should both also have been equally if not more critical of the main alternative. These ideas may not have been particularly influential in the wider student movement, but Althusser's anthropological admirers, most notably Emmanuel Terray and Pierre Philippe Rey, along with Godelier and his students, all sought to reformulate a Marxism that was

'scientific,' based on Marx's 'mature' work, and avoided what they took to be the pitfalls of an approach based on assumptions about a universal human subjectivity thwarted by capitalism.

As this suggests, structural Marxists rethinking the Marxist heritage had to go back to the work of Marx on capitalism, and what they found in so doing were the beginnings of a general theory of historical materialism within which anthropology could play a very significant role. That role was to be the formulation of a general theory of modes of production and the laws of transition from one to another. As opposed to the thinking of some of their critics, this was not seen as a mere exercise in antiquarianism because such a theory would provide, it was argued, the theoretical basis for a political practice aimed at stimulating not just a transition to communism but the shaping of the nature of the communist future as well.

It has been suggested that there was no single Marxist approach in French anthropology but instead, a collection of competing and often antagonistic camps. Yet in spite of differing allegiances – to Althusser, to Godelier's version of structuralism, and so on – at least when viewed from the outside, French Marxist anthropology did seem to constitute a fairly unified discursive field for debate.

French Marxist anthropologists made important contributions to a number of areas of concern to both Marxism and anthropology (for an overview, see Kahn and Llobera, 1981). A good example is provided by the debates over the nature of lineage-based societies in Africa that were sparked by the early publications of Claude Meillassoux (see, for example, Meillassoux, 1964, 1967). Scholars, including most centrally Meillassoux himself, Emmanuel Terray, and Pierre Philippe Rey (both former students of Louis Althusser), as well as Maurice Godelier, Arienne Deluz, Pierre Bonte, and others used Meillassoux's material on the Gouro of the Ivory Coast, as well as ethnographic data on other uncentralized African societies, to debate most of the basic concepts of historical materialism – class, mode of production, productive forces and relations of production, exploitation, articulation, social formation – as well as the central theoretical tenet of historical materialism – economic determinism (or what one participant called Marx's hypothesis of economic determinism in the last instance). This important debate, more than any other, serves to illustrate much of the potential theoretical benefits, as well as some of the main pitfalls, of the project of a Marxist anthropology as it was conceived by these scholars, namely a project to recover, by means of a careful reading of Marx's own work, his 'mature' theoretical project and to examine its more general validity in situations largely unknown to its originator. Moreover, the debate added considerably to the understanding of more traditional anthropological problems related to the nature of authority and social cohesion in kinship-based societies and societies without centralized political systems.

Noteworthy in this context were the critiques of a number of feminist anthropologists, including Felicity Edholm, Olivia Harris, Kate Young, and Maila Stivens, who built on the Marxist concept of social reproduction to generate important insights into gender inequality in precapitalist as well as capitalist societies.

Marxism, Articulation, and Political Economy

From the perspective of the study of the linkages between global capitalism and local social and economic systems, the most important effect of these debates was in contributing to the formation of a theory of what some called the 'articulation of modes of production.' Drawing on the theoretical work of Althusser himself, Etienne Balibar, Maurice Godelier, anthropologists, and others with an interest in the concerns of political economy produced analyses of the relationships between the local and the global, which were understood as an articulation of capitalism with different precapitalist modes of production. It is not often appreciated that exploring this relationship was always a central concern of Meillassoux in a series of publications, including his early monograph on the Gouro. A sophisticated Althusserian analysis of articulation is to be found in the work of the brilliant Marxist anthropologist Pierre Philippe Rey (see especially Rey, 1973). And other important contributions to this project were made in Britain by a variety scholars influenced by French structural Marxism, including Ernesto Laclau, Barry Hindess, Paul Hirst, Talal Asad, and others, and by a group associated with the journal *Critique of Anthropology*, including Felicity Edholm, John Gledhill, Olivia Harris, Joel S. Kahn, Josep Llobera, Stephen Nugent, and Maila Stivens. Mention should be made also of the important work of Jonathan Friedman on Marxist theory, anthropology, and globalization.

Beyond Political Economy

The central concerns of political economy have certainly not disappeared in anthropology. But it would be fair to say that in general there have been significant shifts at the levels of theory and of analytic style as well as in the selection of problems and issues for research, such that very few anthropologists now describe themselves as political economists. The reasons for this declining popularity of political economy are varied. They certainly include an acknowledgment that earlier attempts to theorize the relationship between the global and the local in either a dependency/world systems mode or by means of a concept of the articulation of modes of production suffered from problematic theoretical tendencies. These included an overly functionalist understanding of the needs of capitalism, a unidimensional productionism (or economism), an inattention to the spheres of meaning and agency, and so on. Political economists have, like many others, been criticized for their eurocentrism, for their inattention to gender as a primary category of sociocultural differentiation, and for failing to appreciate that race, culture, and gender are not mere epiphenomena of social systems. Like those of many others, the claims of truth by political economists have been subjected to wide-ranging critique by scholars influenced by poststructuralism, postmodernism, feminism, and postcolonialism.

Social and cultural anthropology, in particular, was also affected by the critiques of those associated with the so-called writing culture school, which undermined claims about the reality of those other lives constructed in anthropological texts and led to a demand for greater degrees of reflexivity. Political economy which, despite the name, doubtless focused on

economic issues more than any others, was naturally affected by these shifts. Finally, European versions of political economy, precisely because of their greater concern with issues of (Marxist) theory, have been damaged by the general intellectual and political rejection of Marxism, particularly since the late 1980s.

Although political economy in the various forms it took in Europe and North America in the 1970s and early 1980s will probably not regain the central role it once had in the discipline, for better or worse, it would be wrong to pronounce it dead and buried. The tendency to see in the rise of Foucauldian and poststructuralist paradigms a complete rejection of, particularly, the basic conceptions of structural Marxism is probably misleading. It is possible to argue that these merely took the structural Marxist project one step forward, from a critique of ideology to a critique of discourse *tout court*.

Certainly, some of the central concerns of political economy have re-emerged in more recent attempts by anthropologists to theorize the phenomenon of globalization. An anthropology seeking its roots in a variety of new social movements is very much on the agenda of anthropologists styling themselves as radical and/or committed. Some of the concerns of an earlier generation of political economists are evident in more recent feminist anthropology as well as in various attempts to construct an anthropology of human rights. Recent attempts to forge a link between critical social theory and anthropology through various reworkings of the concept of modernity may be seen as an attempt to correct the perceived failings of structural Marxism. There are even calls for another return to Marx.

The Critique of Neoliberalism

Perhaps the field in which the concerns of the political economists have most clearly reasserted themselves in the new century has been that of the largely critical analyses of so-called neoliberalism. To quote one contributor to the field: "Across the globe, policies called neoliberal have had devastating impacts on the people with whom anthropologists work. Referring to neoliberalism has become a shorthand way of signaling all that is wrong with the ethnographic present, much as previous generations of anthropologists would refer to globalization [and, one should add, their political economist predecessors would refer to capitalism]" (Elyachar, 2012: 76) (Just how central the critique of neoliberalism has become to the discipline is evident in the fact that two of the most important anthropology journals in North America and Europe, respectively, devoted special issues to the topic in 2012 (Cultural Anthropology 27, (1) and Social Anthropology 20, 1)). Particularly influential have been the writings of anthropologists James Ferguson and Aihwa Ong, along with those of geographer-turned-anthropologist David Harvey (see for example Harvey, 2005; Ferguson, 2006; Ong, 2006).

Originally coined in the interwar years by the Viennese economist Frederich Hayek to signal what he thought were the failings of classical liberalism, as well as his implacable opposition to state involvement in economic planning, the term *neoliberalism*, as used in turn-of-the-century anthropology, means somewhat different things to different people. For some it refers simply to the triumph of unfettered market forces, along with the corresponding decline in the powers and economic role of the state, that appear to date from the Reagan-Thatcher years. Others, however, have suggested that the neoliberal era has not been characterized by the sort of withdrawal of the state favored by both free-market radicals and libertarians, but instead by what Loïc Wacquant, based on research on Chicago's South Side and the housing projects of outer Paris, has called a new 'articulation of state, market and citizenship that harnesses the first to impose the stamp of the second onto the third' (2012: 66). Wacquant's formulation, in which structural links are posited between commodification, disciplinary social policy, penal policies that are 'expansive and pornographic,' and the 'trope of individual responsibility,' has the advantage of theoretical rigor and therefore the advantage in explicating what is actually 'neo' about neoliberalism. However, it can be criticized on very much the same grounds, as were earlier attempts at theorizing the capitalist mode of production as a structure in dominance of three semiautonomous levels, or instances, namely, that when juxtaposed with the global diversity of actually existing neoliberal regimes, the necessity of the connections among these elements is called into question (see, for example, Hilgers, 2012). Like the abstract concept of the capitalist mode of production, in other words, the more rigorous formulations of a concept of neoliberalism become extremely difficult to operationalize in precisely those contexts in which anthropologists work.

Toward an Anthropology of Capitalism

If the critical anthropological analyses of neoliberalism build on political economy's concerns with the impact of (global) capitalism on the individuals and communities they study, the development of what could be loosely termed an anthropology of capitalism based on the mainly ethnographic study of capitalists themselves has served to flesh out the somewhat abstract or reified understandings of capitalism that prevailed in political economy.

It is impossible to do justice in a short survey to the full range of contributions to this emerging field, but it certainly includes ethnographic studies of corporate life, such as those by Patricia Sloane-White in Malaysia (Sloane, 1999). Another area in which a good deal of innovative work has been done is the ethnographic study of banking and finance. The ethnographic study of financial institutions was pioneered by scholars like Mitchel Abolafia (Abolafia, 1996), while more recent contributions include the work of Caitlin Zaloom on futures markets, Karen Ho on investment bankers, and Alexandra Ourousoff on Wall Street (Zaloom, 2006; Ho, 2009; Ourousoff, 2010). The scholar who has probably done the most to establish the study of finance in the discipline has been Bill Maurer (see Maurer, 2005, 2012).

Historical and Comparative Political Economy

The broadly comparative and historical projects of at least some of 1970s political economy have not disappeared, although they have involved a turn toward theorists other than Marx, or at least a blending of Marxist and non-Marxists

theoretical approaches. With his research on Islamic finance, Maurer himself, of course, retains something of the political economists' comparative project. But perhaps closer to the project of the historical materialists is the broadly comparative and historical work of anthropologists like Keith Hart on money and markets, C.M. Hann on property relations, and David Graeber on value and debt. Although to different degrees remaining within the Marxist tradition, such scholars have also reached out to non-Marxist theorists and traditions: Marcel Mauss and Karl Polanyi in Hart's case, classical anthropological work on property as a bundle of rights in Hann's case, and the anarchist tradition in Graeber's case (see Hann, 1998; Hart, 2001; Graeber, 2011, among others).

Many of the concerns of those anthropologists who were content to be labeled political economists, or even Marxist anthropologists, in the 1970s are of continuing importance to the discipline, even if very few would use either of these terms to describe themselves today. It is therefore, perhaps, premature to pronounce political economy in anthropology dead and buried.

See also: Center–Periphery Relationships; Cultural Critique, Anthropological; Development: Social-Anthropological Aspects; Economic Anthropology; Finance, Anthropology of; Globalization, Anthropology of; Markets: Anthropological Aspects; Modes of Production; Money: Anthropological Aspects; Property: Anthropological Aspects.

Bibliography

Abolafia, M., 1996. Making Markets: Opportunism and Restraint on Wall Street. Harvard University Press, Cambridge, MA.

Elyachar, J., 2012. Before (and after) neoliberalism: tacit knowledge, secrets of the trade, and the public sector in Egypt. Cultural Anthropology 27, 76–96.

Ferguson, J., 2006. Global Shadows: Africa in the Neoliberal World Order. Duke University Press, Durham, NC.

Frank, A.G., 1969. Capitalism and Underdevelopment in Latin America: Historical Studies of Chile and Brazil. Monthly Review Press, New York.

Gough, K., 1968. Anthropology: child of imperialism. Monthly Review 19 (11), 12–27.

Graeber, D., 2011. Debt: The First 5,000 Years. Melville, New York.

Hann, C.M. (Ed.), 1998. Property Relations: Renewing the Anthropological Tradition. Cambridge University Press, Cambridge, UK.

Hart, K., 2001. Money in an Unequal World. Texere, New York.

Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press, New York.

Hilgers, M., 2012. The historicity of the neoliberal state. Social Anthropology 20, 80–94.

Ho, K., 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham, NC.

Kahn, J.S., Llobera, J.R. (Eds.), 1981. The Anthropology of Precapitalist Societies. Macmillan, Houndmills, Basingstoke.

Maurer, B., 2005. Mutual Life Limited: Islamic Banking, Alternative Currencies, Lateral Reason. Princeton University Press, Princeton, NJ.

Maurer, B., 2012. Finance 2.0. In: Carrier, James G. (Ed.), A Handbook of Economic Anthropology, 2nd ed., pp. 183–201.

Meillassoux, C., 1964. Anthropologie Économiques des Gouro de Côte d'Ivoire. Paris-Le Haye: Mouton.

Meillassoux, C., 1967. *Recherche d'un niveau de détermination dans la société cynégétique*. L'Homme et la Société 6.

Mintz, S.W., 1985. Sweetness and Power: The Place of Sugar in Modern History. Viking, New York.

Ong, A., 2006. Neoliberalism as Exception: Mutations in Citizenship and Sovereignty. Duke University Press, Durham, NC.

Ortner, S., 1984. Theory in anthropology since the sixties. Comparative Studies in Society and History 26, 126–166.

Ourousoff, A., 2010. Wall Street at War: The Secret Struggle for the Global Economy. Polity Press, Cambridge, UK.

Rey, P.P., 1973. Les Alliances de Classes. Maspero, Paris.

Sloane, P., 1999. Islam, Modernity and Entrepreneurship Among the Malays. St Martin's Press, Oxford and New York.

Wacquant, L., 2012. Three steps to a historical anthropology of actually existing neoliberalism. Social Anthropology 20, 66–79.

Wallerstein, I.M., 1974. The Modern World-system, vol. 1. Academic Press, New York.

Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley, CA.

Zaloom, C., 2006. Out of the Pits: Traders and Technology from Chicago to London. Chicago University Press, Chicago, IL.

Political Economy of Renewable Energy

Volkmar Lauber, University of Salzburg, Salzburg, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Renewable power from wind, solar PV, or biomass reached market penetration with large full-scale deployment within about four decades of the first oil crisis. This required overcoming ideological hurdles and electricity incumbents' vested interests. Chief ideological hurdle was neoclassical economics with its restricted role for public policy in technology innovation, thus policy initiatives came about in only a few countries with special political constellations. EU institutions reflected those conflicts; their renewed intensification in the 2010s led to a setback for renewables in Europe. But by then the breakthrough on costs put renewables in a good position to thrive on all continents.

Introduction

Overcoming fossil fuel dominance in electricity generation by a shift to renewable energy sources is a long-term social and political process as well as a technological and economic one which began about four decades ago. Actors from very different milieus and worldviews need to be won over to this idea before it can become effective and prevail over defenders of the status quo, above all interests linked to fossil fuels. The present article analyses how renewable electricity generation made its breakthrough over the past two decades due to special policies mostly in Europe and the United States, and how Europe may now fall behind again on this development.

An Idea Emerges and Gains a Foothold

The idea to shift from fossil to renewable energy usually appears at times of crises resulting from increasing fossil fuel scarcity, higher prices or difficulties with imports, or from the hazards resulting from fossil (or nuclear) energy use. Renewable technologies can avoid these problems while promising greater self-reliance and an energy system more responsive to citizen concerns. In recent years, the strongly reduced cost of renewable power became an additional argument.

Many of these things came together in the 1970s. For the United States, geologist M. King Hubbert had predicted in the 1950s that oil production would peak in the 1970s; the Club of Rome in its first report on the Limits to Growth in 1973 spread a message of a broader resource scarcity coming up. When the first oil crisis broke out in 1973, OPEC's oil price increase and embargo for the Unites States gave Americans a first impression of energy scarcity (more for heat and transport though than for electricity). Around that time the cost of nuclear power plants in the US was also becoming prohibitive, and nuclear power suffered another severe setback with the Three Mile Island accident of 1979. All this contributed to reconsider energy technologies and helped the 'discovery' of renewable energy. The energy crisis put pressure on governments to set initiatives on this issue. Civil society also reacted, often in ways quite different from governments.

Nuclear power had become a collective myth of progress in the 1950s but made slow headway until the energy crises of the 1970s when many governments formulated plans for its rapid expansion. In several countries, this led to a strong antinuclear movement (Germany, France, Denmark, Austria, to name but a few). In Germany and Denmark, support for 'solar' energy in architecture, home heating, and electricity generation became part of an energy counterculture which was quickly taken up by the media and public opinion. More in the background, reliance on coal was also stepped up to replace oil. Electricity from wind, biomass, and solar power was still an abstract idea rather than an embodied technology, but it started to draw attention. This was hardly a case for business entrepreneurs as potential profits were uncertain and far away; thus progress seemed to demand government support, and this is where most calls were addressed. Many organizations such as universities, research organizations, and high-tech firms were eager to move into this field with the help of public funding.

Surprisingly, private initiatives came into being nonetheless, notably with 'tinkerers.' In Denmark, they were the pioneers on wind energy, often coming from the cooperative and folk high school movements; in Germany they pioneered work on biogas plants and to a lesser extent also on wind energy. Unlike business actors, they shared inventions freely without regard to patents, even across borders. They mostly came from the antinuclear counterculture and social movements.

Government Ambivalence: Subsidies for Nuclear and Coal, R&D Funding for Renewable Energy

The U.S. government was the first to react in big style to the oil crisis of 1973/74 with the creation of a well-endowed new government department (which became the Department of Energy in 1977) and comprehensive R&D programs. Though the main emphasis was on coal and nuclear power, generous R&D funding for wind and solar power was expected to bring down their cost quickly. Several prototypes of wind turbines were developed and demonstration programs started. The cost of solar cells was also brought down substantially. In 1978, a federal law (the Public Utilities Regulatory Policies Act, PURPA) was passed which required utilities to purchase renewable electricity from qualifying facilities at the avoided cost of own power. Half a decade later – probably unexpected

for the law's authors – this act unleashed remarkable business entrepreneurship in the field of wind energy during the California wind power boom in the early 1980s which started under governor Jerry Brown, supported by favorable tax credit legislation, and led to the first mass market for (mostly Danish) wind turbines. But the big spending programs for R&D and RD&D programs were wound down again when Ronald Reagan replaced Jimmy Carter in the White House and turned the focus back to oil and gas; Republican governor Deukmejian in California brought a similar change (van Est, 1999).

In Europe, no government undertook such a drastic step as PURPA in the 1970s or even the 1980s, but things changed there as well. Most governments stepped up R&D on renewable energy in the second half of the 1970s, to reach a peak in 1980 followed by decline thereafter; as in the US this support remained tiny when compared to that given to nuclear power. Nevertheless, several governments funded the development of wind turbine prototypes (Belgium, Denmark, Germany, the Netherlands, Sweden, the UK), usually large machines full of technical problems. The idea prevailed that utilities would be the natural users of such turbines, even though utilities were quite ambivalent about new and decentralized renewables (in Germany a 3 MW turbine was even built with public money to prove that wind power was no viable alternative to nuclear). Only Denmark and Germany also supported small turbine development; these were the first viable machines. In the UK, wave power was also subsidized, particularly in a big project that literally sunk in a storm. Solar power research took place in fewer countries; Germany was leading as German firms had been producing PV for U.S. satellites for two decades. In the 1970s, some high-tech firms thought that a breakthrough to cheap solar power could be achieved quickly but lost interest when things turned out otherwise.

By the 1980s, government R&D had produced a variety of prototypes still far removed from market readiness and with only few niche markets (there was one for small wind power among a community of believers in Denmark) and tiny specialty markets in which prices were not that important (e.g., off-grid applications for PV such as buoys, telecommunication towers, or pipelines). But on the whole, those prototypes could not generate electricity at prices competitive with fossil generation. Only Danish wind turbines – smaller and sturdier than most others, built on earlier experience and with state quality control – for some years found a sizable market in California until the boom ended in 1984 with the termination of PURPA standard offer contracts. Voluntary green electricity markets could not create sufficient demand. Change became more urgent with the Chernobyl accident in 1986 and the emergence of climate change as a major political challenge; both called for a reconsideration of fossil and nuclear (fission) electricity.

Yet for adherents of neoclassical economic theory – the dominant economic theory of public management, reinforced by the new neoliberal ideal of lean government – the state had no legitimation to go beyond providing R&D and perhaps also RD&D support; for them, things were now up to the market. In practice, this attitude would have left the prototypes in a lurch, with market actors – and particularly utilities – shying away from costly adventures. In general the utilities – but also most governments and the IEA (International Energy Agency) – viewed renewables as simply too small to be taken seriously before the middle of the twenty-first century, their potential too marginal under the meteorological conditions of Europe and their price too high. This is precisely what happened with most prototypes. It took some 'political accidents' to allow the necessary next step in renewables development in at least a few countries: systematic market creation for renewable power. This inspired considerable political controversy.

The Issue of Market Creation: by Whom, for Whom, and by What Instrument?

Mobilizing public resources to help firms to be viable is more easily associated with declining sectors than with learning investments for young industries, at least in the 'advanced' industrial countries (even though it was practiced as development economics in the 1970s in the 'newly industrializing economies' of Southeast Asia). In European states, such decisions were usually made by the Finance or the Economic Affairs ministry – at that time mostly strongholds of neoclassical economics for whom supporting a technology beyond R&D and a prototype was not considered useful since at this point it was up to the market to take over. Exceptions were made in cases of high politics for reasons of national priority such as nuclear power programs or even coal subsidies (to secure an important domestic energy source). But fledgling renewable power technologies hardly belonged in this category.

In some states, parliaments, more responsive to social movements than bureaucracies that often identified with electricity incumbents, took on the cause of renewables against this orthodoxy. In some cases, breakthroughs came about as a result of a political accident that let a bill slip through parliament before the electricity incumbents had mobilized their supporters.

One such case was the first instance of market creation for renewables which took place in the United States with PURPA 1978 (see above) and which was not expected to play a big role for renewables but helped – together with tax credits – to set off an unprecedented boom for wind power in California. But the tax credits were problematic: investors received the credits once a turbine was installed, even if it failed shortly afterward. Thus the standard offer contracts for investors under PURPA were terminated in 1984 and the wind boom stopped. This was hardly an inspiring success story for imitation by others.

In Denmark, the government from 1979 combined an investment subsidy with the threat of regulation by law to induce utilities to conclude voluntary agreements with wind turbine owners – usually cooperatives – by which the latter would receive 85% of the consumer price for electricity and part of the connection costs. These agreements had to be renewed every few years. In the early 1990s, the utilities wanted lower payments per kilowatt hour and delayed a new agreement, whereupon the minister of energy in 1992 declared that the matter would henceforth be regulated by a law which maintained the 85% figure and which was passed by the alternative 'green' majority in parliament that existed alongside the government coalition. Within a few years this led to the biggest wind boom in Denmark involving several thousand cooperatives and some 100 000 owners; in addition, the Danish government asked utilities to develop wind farms on

their own. It also supported wind energy deployment via a state-financed quality program for turbines, rules for determining turbine locations etc. Parliament rejected nuclear power in 1985 and in 1990 made clear that no big coal plants would be permitted in the future, a step taken by few countries so far. In 2001, a new government stopped most support measures, but in the meantime Danish wind turbine makers were global leaders and leading exporters (van Est, 1999; Nielsen and Heymann, 2012; Hvelplund, 2005).

In Germany, market creation for renewables was first introduced by parliament against government and electricity incumbents. After Chernobyl, public opinion became solidly and permanently antinuclear. The Greens then committed to shut down nuclear plants should they come to power; the Social Democrats (also in opposition) to phasing them out gradually. Rising concerns over climate change led to a comprehensive consensus in parliament that market creation for renewables in the form of feed-in tariffs should be instituted. The Ministry of Economic Affairs – traditionally the patron of the electricity incumbents – opposed what it called 'perpetual subsidies.' Against strong resistance from their leadership, Conservative MPs in the late 1980s submitted members' bills requiring utilities to purchase and pay feed-in tariffs for all renewable electricity tendered to them by mostly small generators, at a rate of 90% of the consumer price (clearly inspired by Danish experience but not voluntary). The government reluctantly gave its consent since the cause was widely popular, elections were upcoming, and the incumbent utilities absorbed by taking over their East German counterparts during German reunification. Thus the Feed-in Law of 1990 was passed unanimously, although the utilities (and Conservative and Liberal leaders) viewed this as an accident in need of repair (Jacobsson and Lauber, 2006; Laird and Stefes, 2009). But for almost two decades these attempts at 'repair' remained without success, be it in the courts, parliament, or even EU institutions.

When in 1998 a Social Democratic-Green coalition replaced the Conservative-Liberal government, it started *Energiewende* or energy transition (i.e., to renewables and efficiency) by adopting the Renewable Energy Act of 2000 plus the phaseout of nuclear power against the preference of an Economic Affairs minister who would have preferred a quota/certificate system (see below) or even voluntary green tariffs. The philosophy of this act was to accelerate renewable power deployment by providing great security for investors in such technologies via an unlimited purchasing obligation of utilities for all renewable electricity, priority dispatch, and technologically differentiated, 'cost-covering' feed-in tariffs fixed for 20 years after installation and providing a modest return to installations properly built and managed. The extra cost of the tariffs would be borne by all electricity consumers via a surcharge, though exceptions were soon made for energy-intensive industries facing international competition. Cost and windfalls would be limited by taking account of size, quality of the site etc. Rapid and stable deployment should encourage mass production, technological learning, and price reductions even for technologies that were still expensive at the time the law was enacted, and also create new employment and exports. This investment in learning would be amortized by clean and cheap domestic electricity within a few decades and free Germany from import dependence (Jacobsson and Lauber, 2006; Mitchell et al., 2006; Hirschl, 2008).

This was a policy for the long haul and not dependent on the good will of utilities. Liberals and part of the Conservatives objected to the long suspension of 'market rules' (no competition between generators but 'only' between producers of turbines, solar cells etc.). Market creation was highly successful in terms of deployment. Onshore wind grew (cumulative capacity) from 55 MW in 1990 to 6.1000 MW in 2000 and 34.000 in 2013; PV from 2 MW (1990) and 114 MW (2000) to about 35.500 MW in 2013. In terms of electricity generation, the share of new renewables without hydro grew from near zero in 1990 to about 22% in 2013. Half of the investment came from about 1.5 million private persons. In 2010, the four big incumbents owned only about 6% of new renewables' capacity (trend:research, 2011). By that time, the renewable power sector had also created about 350 000 jobs.

Spain was the third European country to rapidly deploy wind and solar power, mostly by means of a market premium (fixed and technology specific, added to the electricity price). Mostly incumbent utilities profited from this highly lucrative system (given Spain's high level of electricity prices); rising demand meant that conventional generation was not threatened. For small generators a feed-in tariff offered lower but safer returns (Rio González, 2008).

An entirely different approach to market creation prevailed in the UK. Liberalization and privatization of electricity (Thatcher's reform projects) were meant to intensify competition and reduce prices. They also exposed that all nuclear power plants but one were making losses, not profits. Distributors were then required to carry a specific quota of (noncompetitive) nuclear power. This nuclear assistance program was baptized 'nonfossil fuel obligation (NFFO).' A creative bureaucrat decided that renewables were also 'nonfossil' and could thus bid for the quota. Bidding rounds were organized irregularly for renewable generators – originally financed by a fossil fuel levy – to make this a highly competitive process though almost meaningless in terms of renewable power installed. Intense competition for the windiest locations led to neglect of landscape considerations or local preferences and unleashed an aggressive antiwind power movement ('Country Guardians,' many of them friends of nuclear) even before a significant numbers of turbines were actually installed. It also brought down bid prices so far that many winning projects were never built. In 2002, the Blair government replaced NFFO by the Renewables Obligation (RO), based on similar economic philosophy to support only the currently cheapest sources to drive down their cost even further. The RO set an annual quota of renewable electricity (increasing over time) for all distributors, which they could meet from own generation or by buying RO certificates from others, or by paying a buyout price. Renewable generators receive the market price for electricity and the certificate price (plus some other elements). The system resembles emissions trading: volatile certificate prices, low security for investors, and thus need for high rates of return. This keeps out small generators, reserving a very lucrative business to utilities ambivalent about renewable power and helping them to keep transition to renewables under control by regularly under fulfilling the quota (Mitchell et al., 2006; Woodman and Mitchell, 2011).

A similar system was introduced in most U.S. states since the late 1990s: the renewable portfolio standard (quota). This quota system also favors the cheapest technologies (unless there is a set-aside e.g., for solar PV) and like the RO acts as a cap on deployment; if the quota is over fulfilled, certificate prices collapse (Lewis and Wiser, 2007). The resulting low security for investors and slow growth long prevented the rise of a renewables equipment industry in both the UK and the US. In the US, this was reinforced by the on–off character of federal tax credits for renewable power which led to huge swings in annual installations and long discouraged especially wind turbine producers from settling there.

When the European Commission tried to accelerate the internal market integration of the national renewable power sectors by legislation (renewable energy directives 2001/77/EC and 2009/28/EC), it argued that a quota system would be the most cost-efficient and also the most effective solution. Chief supporters of this move were Eurelectric (the association of European electricity incumbents) and the U.K. government. But in 2008, Commission staff submitted studies on the actual performance of the two support systems. They showed that quota systems generally paid the highest revenues and profits to renewable power generators and were the lowest on effective deployment. Feed-in tariffs combined high deployment with lower revenues and modest profits. On both occasions, Parliament and Council prevented the banning of feed-in tariffs pushed by the Commission (Lauber and Schenner, 2011).

In Japan, MITI launched an investment subsidy program for PV rooftops which made Japan the biggest PV market in the early 2000s but subsidies were reduced rather quickly, presumably due to the neoclassical aversion against supporting long-term learning investments and a clear preference for nuclear power as the mainstay of electricity supply (Chowdhury et al., 2014).

Theses examples show that market creation policies varied strongly with economic ideologies and the institutions that formulated those policies. Parliaments were decidedly more favorable to effective deployment often inspired by technological innovation system approaches (Jacobsson et al., 2009) while ministries and technocrats tended toward neoclassical restraint. Neoclassical quota systems such as RO and renewable portfolio standard (RPS) favored the cheapest renewable power technologies (often waste combustion and landfill gas), capped deployment at low levels, tended to produce large windfalls, profits for incumbents and contributed little to initially more expensive technologies such as solar PV, offshore wind etc. German-style feed-in tariffs supported wind, PV, and other sources simultaneously and without cap, and limited windfalls. Had the neoclassical approach prevailed – as favored in the late 1990s/early 2000s by the European Commission (and also IEA and the World Bank) – renewable energy development would not be nearly as advanced as it is now. PV would still be prohibitively expensive and wind energy less advanced. Under quota systems, growth of capital goods industries and technological learning concerning renewable power would have been much slower; up until about 2010 they were heavily concentrated in Denmark, Germany, and Spain. Electricity incumbents lukewarm about such technologies would still tell the story of high-priced renewables (in fact, 'noncompetitive' German feed-in tariffs for wind are lower than 'competitive' British RO payments despite the UK's better wind resources) and probably continue with fossil and nuclear for several more decades.

Global Surge of Photovoltaics Followed by Contraction in Europe

Between 2000 and 2010, wind power (worldwide) grew at an average annual rate of 27% (doubling every 4 years; see **Figure 1**); solar PV at a rate of 42% (doubling every 2 years; see **Figure 2**). Worldwide investment flows into the renewable electricity sector set new records every year. During this decade Europe was clearly leading the world in RES-E (renewable energy sources – electricity) development and deployment, with the US and China closing up. Some thought that the EU Emissions Trading System (EU ETS) would also promote renewable energy development, with rising carbon prices expected to reduce the role of coal. But until auctioning of certificates started in 2013, its perverse effects prevailed and ETS windfall profits for coal generators actually encouraged coal plants (Pahle et al., 2011).

In the mid-to-late 2000s, many European countries introduced a feed-in tariff for solar PV since quota systems – unless reformed by technological banding (successfully pushed in Britain by parliament) – failed to support this technology. PV's rapid deployment had unexpected effects. Economies of scale and technological learning made module prices fall faster than tariffs and stimulated demand even more; this resulted in speculative demand bubbles (and high legacy costs of tariffs). It also led to partial outsourcing of cell and module production to Southeast Asia. China's local and regional government levels supported the rise of a PV export industry with cheap land, low-cost long-term credit, and other privileges. Now subsidizing exports is outlawed by the World Trade Organization (WTO), but few European governments even tried invoking WTO rules to stem the onslaught of Chinese modules until the damage was largely done (i.e., by 2013). Many European governments compounded that damage by radical support cuts, sometimes even retroactive and particularly damaging to investor confidence, drastically reducing installations and driving many European cell and module producers into insolvency. China by contrast multiplied its domestic PV demand several times over to save its industry. The combined result of rapid technological learning plus trade war was that PV became reasonably cheap very quickly (cost reduction by about 80% in less than a decade) but was still perceived as expensive by many governments (due to its legacy costs, possibly also to intellectual inertia, and political opposition; see below).

The surge of PV had another effect explaining some of the slowness of many European governments to help this sector. It dealt a hard blow to incumbents relying on fossil and nuclear generation, biting into their profits and accelerating their decline. These incumbents now claimed the surge of renewables – particularly of PV – not only caused grid instability, the risk of blackouts, and energy poverty; worse, it threatened Europe's industry, employment, and global competitive position. Governments should therefore slow down this process in the public interest. This drama gathered intensity over the past few years.

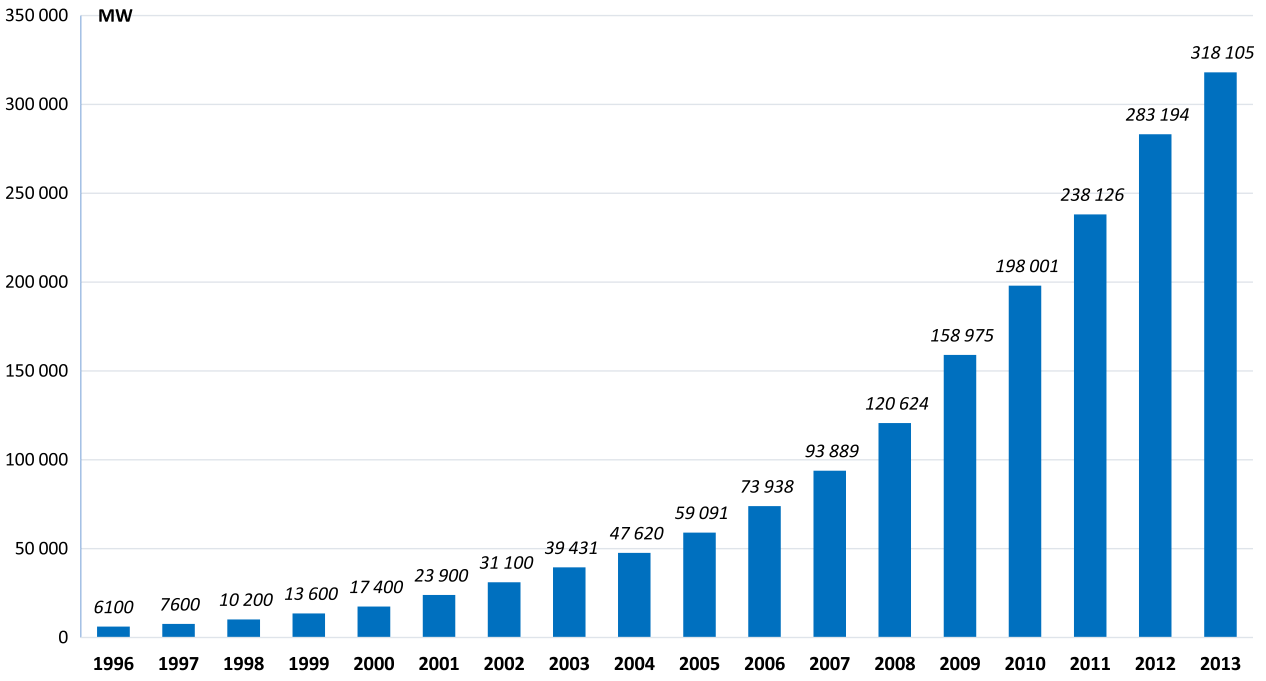

Figure 1 Evolution of global cumulative installed wind capacity, 1996–2013. Adapted from: Global Wind Energy Council (2014) Global Statistics. http://www.gwec.net/global-figures/graphs

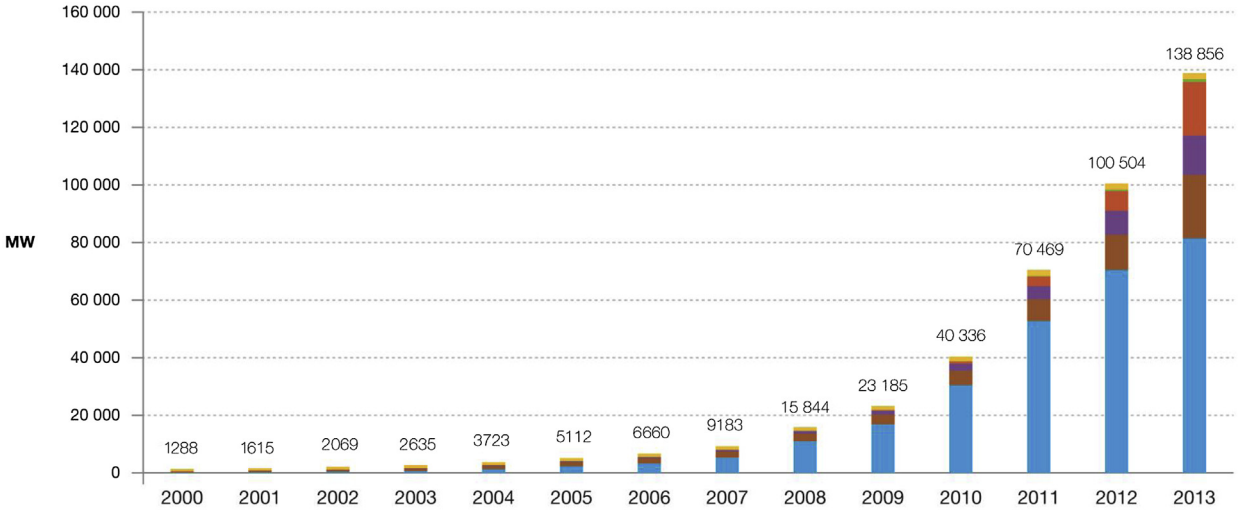

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
RoW	751	807	887	964	993	1003	1108	1150	1226	1306	1590	2098	2098	2098
MEA	n/a	n/a	n/a	n/a	1	1	1	2	3	25	80	205	570	953
China	19	24	42	52	62	70	80	100	140	300	800	3300	6800	18 600
Americas	21	24	54	102	163	246	355	522	828	1328	2410	4590	8365	13 727
APAC	368	496	686	916	1198	1502	1827	2098	2628	3373	4951	7513	12 159	21 992
Europe	129	265	399	601	1306	2291	3289	5312	11 020	16 854	30 505	52 764	70 513	81 488
Total	**1288**	**1615**	**2069**	**2635**	**3723**	**5112**	**6660**	**9183**	**15 844**	**23 185**	**40 336**	**70 469**	**100 504**	**138 856**

RoW: Rest of the World. MEA: Middle East and Africa. APAC: Asia Pacific.
Methodology used for RoW data collection has changed in 2012.

Figure 2 Evolution of global cumulative installed PV capacity, 2000–13. European Photovoltaics Industry Association, 2014 Global Market Outlook for Photovoltaics, 2014–18, p. 17. www.epia.org/news/publications

The Impact of RES-E Deployment on Fossil and Nuclear Electricity Incumbents

Since the first half of the twentieth century, electricity incumbents – some of the most capital-intensive firms globally – developed a business model characterized by increasingly large-scale thermal technologies, highly centralized with a limited number of plants connected to urban and industrial consumption centers. This system grew along with electricity consumption. Until liberalization in the late twentieth century, it was largely protected against competition, by territorial monopolies whose regulation meant constant interaction with governments. Their corporate culture defined the right size of generation units in hundreds of MW to more than 1 GW; small units were a nostalgic utopia. With liberalization, substantial overcapacities emerged. The fact that new renewable energy technologies took off in parallel presented an additional challenge for the sector; it was natural that electricity incumbents would try to stop them or to subject them to their own system requirements (e.g., by a quota system).

Beginning in the mid-2000s the surge of renewable generation made itself felt in the countries leading on deployment. First, incumbents lost market shares, which diminished business volume but not infrastructure costs. Second, their most expensive generation plants were displaced by the merit-order system typical of liberalized markets under which generators are dispatched according to their marginal cost; this eliminated oil and then gas plants and reduced wholesale prices (and thus revenues) for all generation technologies. Third, they lost an increasing share of their most profitable generation – during the daily peak – to solar PV. These peaks now tended to become mid-day troughs in countries with large deployment of PV such as Germany and Italy (plus some neighbor countries sharing the same electricity exchange). Worse, PV deployment came with lightning speed when compared to the long gestation periods (10 years and more) for coal or nuclear plants – and some of these incumbents (e.g., in Germany) had built new coal plants with windfall profits from the EU ETS (Pahle, 2010), increasing surplus capacity. For many generators this meant heavy losses, poor credit ratings, and thus high costs of capital – and most of all, the perspective of further contraction and eventual demise. This fall from the throne came about within just a few years (Economist, 2013).

Political Efforts to Slow Down the Shift to Renewable Energy

This led to far-reaching political reactions. Politicians following neoclassical logic refurbished their old claim that RES-E support which ignored market competition was bound to lead to disaster and needed to be replaced with market-compatible instruments or scrapped entirely. They attacked feed-in tariffs, reduced them beyond plausible need or canceled them altogether. In Germany, the government in 2012 started to question the basic instrument of energy transition or Energiewende, i.e., the Renewable Energy Act, pointing to the apparently excessive cost of renewable electricity as reflected in the growing surcharge paid by consumers and burdening German industry. It called for capping RES-E growth and succeeded in changing the RES-E discourse even though its interpretation was singularly one-sided. It neglected that merit-order effects caused by renewable electricity had actually decreased wholesale electricity prices (which were benefiting half of industry but not passed on to household consumers and small industry by generators, distributors, or suppliers), and that half of industry was actually benefiting – not suffering – from the combination of lower wholesale prices with exemptions from the surcharge and grid fees.

In September 2013, nine of the biggest European utilities urged EU institutions to make sure that electricity prices reflect only market costs, not 'political measures.' This is directed particularly at market-creation policies which finance learning investments. They demanded that support for 'mature' RES-E technologies must be eliminated and restricted to R&D for infant technologies. By contrast conventional plants on standby for balancing fluctuating renewables should be supported by capacity payments. The emphasis in climate policy should be placed on emissions trading, a system that so far failed to produce the expected impact. By February 2014, the EU Commission supported the utilities' push in its proposals for nonbinding, low renewable energy targets for 2030, and for guidelines on state aid to renewable energy eliminating feed-in tariffs. On the national level, a similar policy approach has prevailed in some countries that formerly were leading RES-E development. Thus the German government is intent on putting the brakes on 'excessive' RES-E growth ever since 2010, despite Fukushima and subsequent talk of Energiewende. In Spain, successive governments dramatically reduced or eliminated premium payments for RES-E. In Italy, the government canceled a feed-in tariff that had led to the highest penetration of PV in Europe (7%), considerably higher than in Germany (5%). The U.K. government proposed to offer an incredibly lucrative feed-in tariff ('contract for difference') to a new nuclear plant but reduced plans for the use of abundant and cheaper wind power.

Renewable Electricity Development: Slower in Europe, on the Rise Globally

So in 2014 the world region which contributed the most to renewables development for about two decades – the European Union – is headed for a slowdown, with many of its leaders arguing that it cannot carry this burden alone. These arguments are put into practice just as RES-E development is bearing fruit in the form of lower prices, something that would call for an adjustment of policy but not for a change of goals. However, it is precisely this fruition which upset fossil electricity incumbents and motivated politicians to come to the rescue of this declining industry.

The US is in middle position. PV is finally taking off in the country that first massively invested in its development. However, many utilities fear that its most lucrative customers may become self-suppliers and look to displace or reduce net metering, or burden it with new fees; this conflict could yet hurt deployment. At the same time on–off tax credits and fracking constrain wind power development. Some utilities however are eager to capture – and if possible restrict – the solar PV rooftop business for themselves.

But this does not mean that RES-E deployment will slow down globally. In the early 2010s, leadership in production, diffusion, and increasingly technological development of renewable electricity shifted to China, which in 2013 was about at par with the EU in terms of RES-E installations but had much higher growth rates. Japan was suddenly propelled to a leading role in RES-E deployment due to a feed-in tariff voted after Fukushima. And while it was common to argue that only wealthy countries could afford wind turbines or PV, this has changed massively in recent years as the price for these devices has come down. The financial crisis dampened deployment particularly in Organisation for Economic Cooperation and Development (OECD) countries, but not so in others. Renewable electricity deployment is taking big steps in developing countries, with economics coming to its support. The virtual absence of a comprehensive grid in many of these countries encourages RES-E. Meteorological conditions and thus wind and solar resources are usually much more favorable than in Europe. The extraordinary capital cost of new coal and nuclear plants are a handicap, even for big utilities in wealthy countries. Furthermore, in 2013 the U.S. government, and later the World Bank, the European Investment Bank and the European Bank for Reconstruction of Development plus the U.K. government all pledged to no longer finance coal plants in developing countries except in very special situations. The import bill for fossil fuels is a heavy burden for many such states, and their prices are likely to go up over the long term. The price reductions for RES-E are creating new opportunities there. All across Asia, Latin America, and Africa numerous projects have sprung up in the current decade. Europe may take a back seat at its own risk; but it will not stop the rapid advance of RES-E technologies in the rest of the world.

See also: Bounded Rationality; Energy and Society; Energy, Anthropology of; Environmental Movements; Green Parties; Oil Age: Energy in Transition; Policy Analysis; Policy Networks; Public Policy.

Bibliography

Becker, P., 2011. Aufstieg und Krise der deutschen Stromkonzerne. Ponte Press, Bochum.
Chowdhury, S., Sumita, U., Islam, A., Bedja, I., 2014. Importance of policy for energy system transformation: diffusion of PV technology in Japan and Germany. Energy Policy 68, 285–293. http://dx.doi.org/10.1016/j.enpol.2014.01.023.
Economist, 12 October 2013. European Utilities. How to Lose Half a Trillion Euros. Europe's Electricity Providers Face an Existential Threat.
Elliott, D., 2007. Sustainable Energy: Opportunities and Limitations. Palgrave.
van Est, R., 1999. Winds of Change. International Books, Utrecht.
Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, 2013. Renewable Energy Sources in Figures. Berlin. http://www.erneuerbare-energien.de/fileadmin/Daten_EE/Dokumente__PDFs_/ee_in_zahlen_en_bf.pdf (last accessed 05.02.14.).
Hirschl, B., 2008. Erneuerbare Energien-Politik. VS Research, Wiesbaden.
Hvelplund, F., 2005. Renewable energy: political prices or political quantities. In: Lauber, V. (Ed.), Switching to Renewable Power. Earthscan, London, pp. 228–245.
Jacobsson, S., Lauber, V., 2006. The politics and policy of energy system transformation. Energy Policy 34 (3), 256–276.
Jacobsson, S., Bergek, A., Finon, D., Lauber, V., Mitchell, C., Toke, D., Verbruggen, A., 2009. EU renewable energy support policy: faith or facts? Energy Policy 37 (6), 2143–2146.
Kemfert, C., 2013. Kampf um Strom. Murmann, Hamburg.
Kitzing, L., Mitchell, C., Morthorst, P.E., 2012. Renewable energy policies in Europe: converging or diverging? Energy Policy 51, 192–201.
Laird, F., Stefes, C., 2009. The diverging paths of German and U.S. policies for renewable energy: sources of difference. Energy Policy 37 (7), 2619–2629.
Lauber, V., 2011. Wind power policies in Germany and the UK: different choices leading to divergent Outcomes. In: Szarka, J., et al. (Eds.), Learning from Wind Power. PalgraveMacmillan, Basingstoke, pp. 38–60.
Lauber, V., Schenner, E., 2011. The struggle over support schemes for renewable electricity in the European Union: a discursive-institutionalist analysis. Environmental Politics 20 (4), 508–527.
Lewis, J.I., Wiser, R.H., 2007. Fostering a renewable energy technology industry: an international comparison of wind industry policy support mechanisms. Energy Policy 35 (3), 1844–1857.
Lovins, A.B., October 1976. The Road Not Taken, Foreign Affairs, pp. 65–96.
Mendonca, M., 2007. Feed-in Tariffs. Earthscan, London.
Mitchell, C., Bauknecht, D., Connor, P., 2006. Effectiveness through risk reduction: a comparison of the Renewables Obligation in England and Wales and the feed-in system in Germany. Energy Policy 34 (3), 297–305.
Nielsen, K.H., Heymann, M., 2012. Winds of change: communication and wind power technology development in Denmark and Germany from 1973 to ca. 1985. Engineering Studies 4 (1), 11–31.
Pahle, M., 2010. Germany's dash for coal – explaining drivers and factors. Energy Policy 38 (7), 3431–3442.
Pahle, M., Fan, L., Schill, W.-P., 2011. How emission certificate allowances distort fossil investments: the German Example. Energy Policy 39, 1975–1987.
Rio González, P. del, 2008. Ten years of renewable electricity policies in Spain: an analysis of successive feed-in tariff reforms. Energy Policy 36 (8), 2917–2929.
Rogol, M., Rogol, S.H., 2012. Explosive Growth. Live Oak Books, Austin.
Smith, A., Raven, R., 2012. Reconsidering niches in transitions to sustainability. Research Policy 41 (6), 1025–1036.
Stenzel, T., Frenzel, A., 2008. Regulating technological change – the strategic reactions of utility companies towards subsidy policies in the German, Spanish and UK electricity markets. Energy Policy 36 (7), 2645–2657.
trend:research, 2011. Marktakteure Erneuerbare Energien-Anlagen in der Stromerzeugung (Internet, accessed 15.12.12.).
Woodman, B., Mitchell, C., 2011. Learning from experience? The development of the Renewables Obligation in England and Wales, 2002–2010. Energy Policy 39, 3914–3921.

Relevant Websites

http://www.erec.org/ – Association of renewable energy industry.
http://about.bnef.com/ – Bloomberg New Energy Finance.
http://www.journals.elsevier.com/energy-policy/ – Energy Policy – Elsevier Journal.
http://www.eurosolar.de/en/ – European Renewable Energy Council, association of European renewable energy industry.
http://www.google.at/#q=EurObserv'ER&safe=off – EurOberv'ER renewable energy barometer.
http://www.eufores.org/ – European Parliamentarians for a Sustainable Energy Future.
http://www.epia.org/home/ – European Photovoltaics Industry Association.
http://www.ewea.org/ – European Wind Energy Association.
http://www.iea.org/ – International Energy Agency.
http://www.irena.org/home/ – International Renewable Energy Agency.
http://www.countryguardian.net/ – Leading group of windpower opponents, UK.
http://www.photon-international.com – Leading trade magazine on solar energy.
http://www.windpowermonthly.com/ – Leading trade magazine on wind energy.
http://www.res-legal.eu/ – Regulations on renewable energy in Europe.
http://www.renewablesinternational.net/ – Renewable energy newsletter.
http://www.ren21.net/REN21Activities/GlobalStatusReport.aspx – Renewable Energy Global Status Report.
http://www.wind-works.org/cms/ – Paul Gipe renewable energy newsletter.

Political Geography

Alexander B Murphy, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political geography is a subdiscipline of geography concerned with the spatial arrangement of political practices and understandings on Earth's surface, and the role of politics in shaping geographical patterns and processes. Political geography investigates why formalized political spaces develop, and how the configuration of those spaces reflects and shapes cultural, social, economic, and environmental circumstances. It also explores the geographical character of political and social movements, as well as the impacts of politics on the geographical organization of human activity and human–environment interactions. The academic study of geopolitics represents an additional important area of political-geographic inquiry.

Development of Modern Political Geography, 1890s–1970

The roots of political geography can be traced as far back as Aristotle's effort to describe the ideal state in terms of its territory, population, and internal geographical structure. The German geographer, Friedrich Ratzel (1844–1904), is often credited with being the progenitor of modern political geography. Ratzel argued that an understanding of the political world needed to be grounded in the human and physical character of the concrete territorial spaces occupied by states. He is most known for a Darwin-inspired 'organic theory,' which treated states as competitive territorial entities vying with one another for control over parts of Earth's surface. Ratzel's seminal book, *Politische Geographie* (1897), was the first major work to use the term political geography.

Ratzel's organic state theory attracted attention in both Europe and the United States. The Swedish political scientist, Rudolf Kjellén (1865–1922), drew on Ratzel's ideas to develop an elaborated vision of the state as an organism, in the process giving birth to the term 'geopolitics' (*see* Geopolitics). Contemporaneously, the American naval historian Alfred Thayer Mahan (1840–1914) published an influential set of works arguing that control over sea lanes and adjacent land areas was critical for the maintenance and expansion of political power. His work was soon followed by an alternative, more widely circulated geopolitical vision developed by the English geographer Sir Halford Mackinder (1861–1947). Mackinder emphasized the importance of the interior of the Eurasian landmass (the heartland) for the maintenance and spread of political power.

Given the relevance of these geopolitical contributions, and the interest they attracted, it is not surprising that the political geography of the early twentieth century was strongly identified with geostrategic thinking (Claval, 1994). Ratzel's and Kjellén's work became the foundation for a school of geopolitics (Geopolitik) in Germany led by Karl Haushofer (1869–1946), which eventually was appropriated by the Nazi war machine to justify its expansionist policies. In other parts of Europe and the United States, debates over the relative merits of Mahan's and Mackinder's conceptualizations attracted considerable attention in both academic and policy circles, and led to yet other geopolitical visions – most notably those of the Dutch-born, American-educated international relations scholar, Nicholas John Spykman (1893–1914), who focused attention on the strategic importance of the so-called rimland (coastal fringe) of Eurasia.

The first half of the twentieth century also saw political geographers being called upon to provide background on boundary disputes, internal divisions within empires, and the impacts of new transportation technologies on political areas. The director of the American Geographical Society, Isaiah Bowman (1878–1950), was a key advisor to President Wilson during and after World War I. Bowman was the chief territorial specialist of the American delegation to the Paris Peace Conference, where a new European political map was forged. In many cases, however, political compromises meant that Bowman's proposals for the redrawing of boundaries were not followed.

During the interwar period, the study of political geography expanded modestly in colleges and universities, although relatively few geographers defined themselves primarily as political geographers. Studies emphasized the historical territorial development and contemporary geographical characteristics of individual states. Bowman produced the first systematic political geography text of the post-World War I political order: *The New World: Problems in Political Geography* (1921). It concentrated on the territorial problems of the new countries of the world – particularly those in Europe. Richard Hartshorne (1899–1992) viewed this work as an extension of Ratzel's concern with the environmental character of state areas; Hartshorne's own influential work focused on the integrative (centrifugal) and disintegrative (centripetal) potentials of differing state territorial characteristics and founding territorial ideologies (summarized in Hartshorne, 1954).

The formal study of political geography declined during the middle decades of the twentieth century. Geopolitics in particular fell into wide disfavor during the middle of the twentieth century because of its association with Nazi efforts to justify claims to territory. In light of the close association between the study of geopolitics and of political geography, for decades after the war the Soviets opposed any use of the term political geography, and in Germany political geography ceased to be taught. Political geography maintained a foothold in the English-speaking academy, but its importance waned for a time. Nonetheless some efforts were made to examine new

geopolitical circumstances (Cohen, 1963), and to describe how geography affected individual states through the study of the spatial morphology, internal organization, locational characteristics, and boundaries of states (e.g., Prescott, 1965).

Political geography became somewhat disconnected from conceptual and empirical developments in other parts of geography when the discipline took a positivist turn during the late 1950s and 1960s. Partial exceptions could be found in studies of electoral patterns and political redistricting, which lent themselves to quantitative analysis. By the end of the 1960s, however, mainstream political geography was not a prominent subfield of geography.

Political Geography's Late-Twentieth-Century Revival

Political geography did not remain in a marginal position within the discipline of geography for long. An important edited compilation from the end of the 1960s (Kasperson and Minghi, 1969) provided a useful framework for teaching and research. Shortly thereafter, Jean Gottmann's (1973) study, *The Significance of Territory*, opened up new ways of thinking about the nature and meaning of the world political map. Moreover, 1970s-era scholars trained in urban and economic geography began to turn their attention to the role politics plays in the distribution of services, people, and wealth. The latter development spawned political-geographic work focused at local and regional scales. Notably, Kevin Cox's (1973) seminal book on the politics of urban services gave rise to an influential set of studies in 'welfare geography' that drew on Marxism and public choice theory.

Contemporaneously, expanded work on the spatial organization of elections, voting patterns, and the geography of representation made electoral geography an increasingly widely recognized subspecialty within political geography in the 1970s (Taylor and Johnston, 1979). Electoral geography widened the scope of political geography and attracted interest among a new generation of geographers. At the same time, the different scales and issues on which votes are taken encouraged a broadening of political geography's earlier preoccupation with the scale of the state.

By the early 1980s, the new work being undertaken on urban and electoral themes was attracting growing attention and helping to connect political geography to issues and concerns being taken up by other subdisciplines of geography and by the social sciences more generally. Moreover, traditional political-geographic themes were taking on new life, with expanding interest in shifting cold-war geopolitical relations, the growth of substate nationalist movements, and the emergence of new regimes of control in the world's oceans. The founding of a new journal, *Political Geography* (originally entitled *Political Geography Quarterly*), in 1982; the establishment of major research and study groups in the International Geographical Union and various national geographic organizations; and the founding of a boundary research institute at the University of Durham signaled the growing importance of the subdiscipline.

At the heart of political geography's revivification was a growth in interest in the role of politics in the larger discipline of geography and the effort by political geographers to develop a more sophisticated theoretical foundation for the subdiscipline. Impetus for the theoretical turn came from the combination of insights being offered by studies of urban social conflict from a political-economy perspective and an interest in making connections with work going on outside the subdiscipline. Explicit engagement with theory in political geography took four overlapping, sometimes conflicting forms: a concern with developing a theory of the state, a concern with understanding how systemic influences shape political-geographic processes, a concern with theorizing the human–territory relationship, and a concern with the role place plays in the development of political identities and practices.

The work of Gordon Clark and Michael Dear (1984) is particularly associated with the effort to theorize the state in political geography in the 1980s. Inspired by, but extending beyond, Marxian conceptions of the role states play in the maintenance and expansion of capitalism, Clark and Dear treated the state as a product of institutionalized political-economic patterns and associated spatial arrangements that influence how states develop. Their work helped to foster wide-ranging inquiries into the spatial dimensions of state policy initiatives, as well as studies of the 'local state' and its relationship to the national state. The latter was also heavily influenced by the writings of David Harvey (e.g., Harvey, 1985), who focused attention on how the capitalist accumulation process was linked to the political production of spatial inequalities within cities and substate regions.

The strand of work in political geography concerned with systemic influences focused particularly on the political-economic underpinnings of the state system and their role in shaping territorial practices and behaviors at the state and substate scales. British geographer Peter Taylor helped to advance this line of work – particularly through his influential text, *Political Geography: World-Economy, Nation-State and Locality* (Taylor, 1985 and later editions coauthored with Colin Flint). Drawing on the World Systems ideas of Immanuel Wallerstein, Taylor situated his analysis of the changing political geography of the modern world in the context of the territorial dynamics of the emerging capitalist world economy. He argued that the study of political-geographic processes at national and subnational scales could not be divorced from the positions states and local political units occupy in the global political-economic order.

Robert Sack's work on *Human Territoriality* (Sack, 1986) helped to catalyze a body of work that treated territory not simply as a framework for the projection of political power, but as a product of human ideas and behaviors. Sack argued that a critical component of human social behavior was the effort to influence people and ideas through the control of specific geographical areas, and that a concern with 'human territoriality' could therefore provide important insights into social and political processes. Sack's work directly challenged the dominant tendency to treat political territories as a priori spatial givens – in the process opening an important theoretical window for political-geographic work investigating how and why particular territorial arrangements come into being, and how those arrangements shape economic, social, cultural, and environmental processes (e.g., Murphy, 1996).

Work on the role place plays in the development of political identities and practices was spurred in particular by John

Agnew's (1987) influential volume *Place and Politics*. Agnew argued that political studies are generally based on the assumption that, in the modern world, local and regional identities are being replaced by national communities segmented by socioeconomic differences – not by geographic sectionalism. Agnew showed that this 'nationalization thesis' fails to account for the enduring impacts of local context on social and political behavior. Drawing on Italy as a case study, he demonstrated that political perspectives and identities were influenced not just by generalized economic circumstances but by local political culture. His work in this vein helped to foster a turn toward more localized studies in political geography.

The new theoretical perspectives coming into political geography during the 1980s influenced the types of issues being studied and how they were studied. The relationship between political geography and uneven development became a particularly significant focus of attention (see Smith, 1984). Political geographers also turned their attention to processes of nation-building and substate nationalism, seeing these as predicated in significant part on the territorial dynamics of the state system. Studies in electoral geography also began to look at the relationship between spatial structures of power and electoral outcomes.

Work on the politics of place and on territoriality developed in conjunction with the late-twentieth-century turn toward post-structuralist approaches in human geography – in the process helping to open up new vistas for political geography. The centrality of politics and political relations to post-structuralist thought greatly abetted this process (Painter, 1995). The academic study of geopolitics was particularly influenced by the post-structuralist turn, giving rise to a body of work known as critical geopolitics. Influenced by Foucauldian ideas, the critical geopolitics literature looks at the discourse of geopolitical practice to gain insight into the ways in which spatial understandings have been developed, deployed, and manipulated in the pursuit of political and military objectives (see Ó Tuathail, 1996). Foucauldian perspectives also promoted new relational approaches to territory, which cast territory not simply as the historical outcome of struggles for control over Earth's surface, but as the outcome of a set of relationships linking individuals and groups with the material and discursive environments in which they are situated (see, e.g., Raffestin, 1980; Storey, 2001).

Contemporary Approaches and Themes

Much of human geography as currently practiced is about politics – making it difficult to establish exactly what distinguishes political geography from other branches of human geography. Yet as Agnew (1997: 3) has argued, "there is nevertheless still the need for an area of study of the more formalized and institutionalized arenas of power such as the spatiality of states, interstate geopolitics, the geographies of political and social movements, and territorial conflicts involving national and ethnic groups." These are the matters that continue to be most closely associated with the subdiscipline of political geography, and they are not particularly confining. Indeed, the contemporary political geography literature addresses themes that range from the ways in which gendered identities undergird particular political-geographic practices to the implications of new telecommunications technologies for the territorial integrity of states.

Much of the recent work in political geography is rooted in approaches that are traceable to three intellectual currents that have shaped the subdiscipline over time: (1) political geography's tradition of describing and analyzing the spatial configuration, geographic character, and boundaries of politically organized areas; (2) the political-economy perspectives on political-territorial developments that began to influence the subdiscipline in the 1970s and 1980s; and (3) the more recent cultural turn that reflects growing interest in the insights that post-structuralist theories of society can provide on the spatial dimensions of political processes. Some studies can still be neatly classified in one or another of these analytical traditions, but there is also considerable blurring among them. Nevertheless, tensions persist among proponents of these different perspectives, the most acute of which is between those who posit a world of 'real' political spaces/arrangements that we can comprehend through some combination of empirical and critical analysis, and those who view reality as unknowable outside of the individuals and societies that construct it.

The substantive topics on which contemporary political geographers are working are a product of the more general effort to make sense of the rapidly changing political, social, and economic environment of recent decades. The following subsections highlight a number of key areas of concern. There is considerable overlap among and between these areas of research, but each has a core literature attached to it.

Political-Geographic Assumptions of Social and Political Theory

There is a pervasive tendency in the social sciences to frame analyses of social and political developments in terms of the political territories in which they are situated. Yet the territories themselves are usually treated as little more than locational referents for whatever is being studied. A body of work in political geography seeks to expose the shortcomings of this way of thinking. Much of that work focuses particularly on the importance of understanding what is hidden and what is revealed when the modern territorial state is treated as an unproblematic spatial container for social and political processes. John Agnew in particular has argued that this approach is the product of a cognitive 'territorial trap' that leads scholars and other commentators to ignore critical historical and geographical processes (see Agnew, 1994). In support of this line of reasoning, a number of political-geographic studies have shown how the political organization of space is itself implicated in social and political processes – even shaping the very concepts of identity on which political communities are based. The nature and meaning of territory has also come under increased scrutiny, most notably in the form of a major treatise tracing the historical development of the concept in the Western tradition (Elden, 2013).

Geopolitical Conceptualizations and Practice

Geopolitics is clearly back on the political geography agenda, as evidenced by the interest in critical geopolitics, by the publication of a spate of recent books, and by the founding of

a journal titled *Geopolitics* in 1996 (its name for the first 2 years was *Geopolitics and International Boundaries*). The recent literature on the subject reveals some continuing interest in the development of formal geopolitical models, but recent studies in this vein have concentrated on the changing geopolitical environment for interstate relations and on the worldviews and priorities of political actors. Considerable attention has been paid to the emergence of the modern world economy and its impacts on the conceptualization and pursuit of geopolitics. Feminist perspectives on geopolitics have also shaped the subdiscipline in recent years (Hyndman, 2003) – encouraging consideration of the ways in which gendered identities and power relations affect geopolitical conceptions and practices. Political geographers with an interest in geopolitics have also examined topics as diverse as the locational components of conflict, the geopolitical dimensions of environmental change, and the ways in which popular culture reflects and shapes geopolitical practice. The end of the Cold War and the initiation of a US-led 'War on Terror' represent a particularly important empirical focus of investigation, with studies examining the ways in which geopolitical ideas and practices are influenced by the inertia of colonial relationships and by the understandings of territory and terror that developed in conjunction with the modern state system (Gregory, 2004; Flint, 2005; Elden, 2009).

Deterritorialization, Reterritorialization, and the Changing Nature of Sovereignty

There is a lively debate in political geography and beyond about the degree to which the traditional role of the state is being undermined by the increased mobility of capital and labor, the growing power of multinational corporations, the advent of new transportation and communications technologies that decrease or eliminate the friction of distance, and the rise of extrastate trading blocs and political organizations (see generally Agnew, 2009; Murphy, 2013). This debate finds concrete expression in studies focused on the ways in which these developments are fostering either deterritorialization and a concomitant undermining of traditional sovereignty arrangements, or reterritorialization and a reassertion of traditional state sovereignty. In an effort to unravel the complexities of the shifting political-territorial order, political geographers are raising questions about the changing nature and function of boundaries; the distribution of powers among and between substate, state, and superstate political-territorial entities; and the means by which governments at different scales are seeking to integrate and defend their territories. Such efforts have led political geographers to focus growing attention on political spaces above and below the scale of the state, as well as to study new political spaces of cooperation along international boundaries and geographically distant political communities that are forging links outside the framework of states.

The Rescaling of Political Practice

Closely related to the prior theme is a concern with the ways in which the scale of political practice is changing. The Keynesian welfare state of the early to mid-twentieth century reinforced the importance of the state scale in the development and implementation of economic and social policy. Over the past few decades, however, changes in the dominant modes of economic production and capital accumulation have worked together with intensified globalization to produce a global shift toward neoliberal political-economic arrangements. Many traditional state practices have been pushed downward to the substate level, whereas others have been taken up by supranational entities and institutions. The extent and political-geographic implications of this rescaling of political power have emerged as a significant theme in political geography – bringing political geographers into conversation with social and economic theorists interested in the changing nature of governance. The geographical dynamics of multilevel governance have attracted particular attention, as has urban government, which has often had to bear the brunt of the abdication of traditional state responsibilities in the social and economic arenas (e.g., Brenner, 2004). At even smaller spatial scales, there is a growing body of political-geographic work focused on political practice in the home, schools, and other microscale settings. Informed by feminist scholarship, studies in this vein have examined, for example, on the impacts of neoliberal political-economic arrangements on 'private' spaces (see, e.g., England, 2003).

Boundaries, Identity, and Citizenship

Political geographers have long been interested in the ways in which the political organization of space reflects and shapes patterns of identity – emphasizing, for example, the often discordant relationship between the map of nations, in the original sense of the term (a group of people sharing a common sense of cultural and political identity), and the map of states. The rise of so many visible instances of ethnonational conflict over the past few decades has moved the relationship between political territoriality and identity to the center of the political geography agenda. Political-geographic research has focused on the competing territorial ideologies and strategies of majority and minority groups, the ways in which the political organization of territory affects intergroup interaction and intragroup identity, and the structures and arrangements that lead to the reproduction of state nationalism. The infusion of post-structuralist ideas into political geography has also promoted interest in the ways in which governmentality (in the Foucauldian sense) and the practices of everyday life influence the nature and meaning of citizenship in an increasingly mobile world (see Painter, 2006; Staeheli et al., 2012), as well as the meanings that are ascribed to particular territories and nationalized landscapes. Indeed, there is growing recognition that how political spaces are understood or imagined can, in many instances, be more socially and politically influential than how they are actually configured (see, e.g., Sidaway, 2002).

The Political Geography of the Environment

One of the clearest obstacles to the effort to confront environmental problems is the fragmentation of the planet into

political spaces, few of which are meaningful ecological units. The issue is not just that ecological and political boundaries do not coincide; the very roles that territorial states play in the international state system and the world economy frequently work against collective action in the interests of the environment (Johnston, 1996). Building on a tradition of examining the relationship between spaces of governance and environmental geographies, political geographers are increasingly turning their attention to environmental questions. Recent work has focused on new transnational regulation regimes and the special issues presented by environments that do not lie within the territories of individual states: the world's oceans and Antarctica. Political geographers have also focused considerable attention on questions of environmental security and the geopolitical implications of environmental change (e.g., Dalby, 2009). At smaller scales, political geographers have examined the circumstances giving rise to the emergence of environmental movements in particular places, the problems and prospects of collective resource management in different political contexts, and the use of environmental strategies in the pursuit of political-territorial objectives. Some of this work is closely aligned with the emerging field of political ecology, with its focus on the role of political-economic arrangements in the production of local environmental outcomes.

Political Movements

Studies in electoral geography have long focused on the spatial dynamics of a particular type of political movements that seek to influence formal elections. Political geographers continue to study electoral patterns – influenced increasingly by structuralist and post-structuralist theorizations of space, place, and power (Warf and Lieb, 2011). With new political movements increasingly challenging long-standing political arrangements, a variety of other types of political movements are commanding attention as well. Work in political geography has focused on the political-territorial contexts that give rise to such movements, the spatial character of movement support, and the impacts of movement objectives on the political organization of space. A significant body of recent literature focuses on political movements that have coalesced around such social categories as race, gender, sexual identity, and ethnicity. Political geographers have also explored the geographical foundations and consequences of antistate movements and other forms of resistance that seek to counter dominant the political-economic structures and arrangements (see, e.g., Pickerill and Chatterton, 2006) – a scholarly endeavor that traces its lineage to Peter Kropotkin's famous 1885 essay, 'What Geography Should Be.' The concept of place – as politically and socially constructed – is central to much of the work on political movements because places provide the foundations for the development of such movements and because identification with place often shapes social and political identities. A considerable body of scholarship focuses on the ways in which dominant interests create spaces or places of marginalization that reflect and reinforce existing power relations and social prejudices.

The Way Forward

Political geography has emerged as one of the most vibrant and active areas of inquiry within the discipline of geography, and it is likely to remain so since the political-territorial upheavals of recent years show no signs of abating. Interest in political geography will likely be fueled by growing challenges to the role of the state – challenges that will further undermine the long-standing tendency throughout the social sciences to treat states simply as taken-for-granted spaces in which social processes play out. Rapid changes in the spatial organization of politics will also likely foster growing efforts to comprehend the nature and meaning of the changing political-territorial order. Studies in political geography will likely continue to be interdisciplinary in character, while underscoring the idea that political practices and outcomes cannot be divorced from the changing geographical spaces and contexts within which they are situated.

See also: Electoral Geography; Geographic Thought, History of; Geopolitics; Military Geography; Postcolonial Geography; Social Movements, Geography of; State, Geography of; Territoriality: Geographical; War and Terrorism, Geography of.

Bibliography

Agnew, J.A., 1987. Place and Politics: The Geographical Mediation of State and Society. Allen and Unwin, Boston.
Agnew, J.A., 1994. The territorial trap: the geographical assumptions of international relations theory. Review of International Political Economy 1, 53–80.
Agnew, J.A. (Ed.), 1997. Political Geography: A Reader. Arnold, London.
Agnew, J.A., 2009. Globalization and Sovereignty. Rowman & Littlefield, Lanham, MD.
Brenner, N., 2004. New State Spaces: Urban Governance and the Rescaling of Statehood. Oxford University Press, Oxford.
Clark, G.L., Dear, M., 1984. State Apparatus: Structures and Language of Legitimacy. Allen & Unwin, Boston.
Claval, P., 1994. Géopolitique et Géostratégie: La Pensée Politique, l'Espace et le Territoire au XXe Siècle. Editions Nathan, Paris.
Cohen, S.B., 1963. Geography and Politics in a World Divided. Random House, New York.
Cox, K.R., 1973. Conflict, Power and Politics in the City: A Geographic View. McGraw Hill, New York.
Dalby, S., 2009. Security and Environmental Change. Polity Press, New York.
Elden, S., 2009. Terror and Territory: The Spatial Extent of Sovereignty. University of Minnesota Press, Minneapolis, MN.
Elden, S., 2013. The Birth of Territory. University of Chicago Press, Chicago, IL.
England, K., 2003. Towards a feminist political geography. Political Geography 22, 611–616.
Flint, C. (Ed.), 2005. The Geography of War and Peace. Oxford University Press, Oxford.
Gottmann, J., 1973. The Significance of Territory. University of Virginia Press, Charlottesville, VA.
Gregory, D., 2004. The Colonial Present: Afghanistan-Palestine-Iraq. Blackwell Publishing, Malden, MA, and Oxford.
Hartshorne, R., 1954. Political geography. In: James, P.E., Jones, C.F. (Eds.), American Geography: Inventory and Prospect. Syracuse University Press for the Association of American Geographers, Syracuse, NY, pp. 167–225.
Harvey, D., 1985. The geopolitics of capitalism. In: Gregory, D., Urry, J. (Eds.), Social Relations and Spatial Structures. St. Martin's Press, New York, pp. 128–163.
Hyndman, J., 2003. Mind the gap: bridging feminist and political geography through geopolitics. Political Geography 23, 307–322.
Johnston, R.J., 1996. Nature, State, and Economy: A Political Economy of the Environment, second ed. John Wiley & Sons, New York.
Kasperson, R.E., Minghi, J.V. (Eds.), 1969. The Structure of Political Geography. Aldine Publishing Company, Chicago.

Murphy, A.B., 1996. The sovereign state system as political-territorial ideal: historical and contemporary considerations. In: Biersteker, T.J., Weber, C. (Eds.), State Sovereignty as Social Construct. Cambridge University Press, Cambridge, England, pp. 81–120.
Murphy, A.B., 2013. Territory's continuing allure. Annals of the Association of American Geographers 103, 1212–1226.
Ó Tuathail, G., 1996. Critical Geopolitics: The Politics of Writing Global Space. University of Minnesota Press, Minneapolis.
Painter, J., 1995. Politics, Geography, and Political Geography. Arnold, London.
Painter, J., 2006. Prosaic geographies of stateness. Political Geography 25, 752–774.
Pickerill, J., Chatterton, P., 2006. Notes towards autonomous geographies: creation, resistance and self-management as survival tactics. Progress in Human Geography 30, 730–746.
Prescott, J.R.V., 1965. The Geography of Frontiers and Boundaries. Aldine, Chicago.
Raffestin, C., 1980. Pour une géographie du pouvoir. Librairies Techniques, Paris.
Sack, R.D., 1986. Human Territoriality: Its Theory and History. Cambridge University Press, Cambridge.
Sidaway, J.D., 2002. Imagined Regional Communities: Integration and Sovereignty in the Global South. Routledge, London and New York.
Smith, N., 1984. Uneven Development. Basil Blackwell, Oxford.
Staeheli, L.A., Ehrkamp, P., Leitner, H., Nagel, C.R., 2012. Dreaming the ordinary: daily life and the complex geographies of citizenship. Progress in Human Geography 36, 628–644.
Storey, D., 2001. Territory: The Claiming of Space. Prentice Hall, Harlow.
Taylor, P.J., 1985. Political Geography: World-economy, Nation-state and Locality. Longman, London and New York.
Taylor, P.J., Johnston, R.J. (Eds.), 1979. Geography of Elections. Penguin, London.
Warf, B., Lieb, J. (Eds.), 2011. Revitalizing Electoral Geography. Ashgate, Farnham, England, and Burlington, VT.

Political History: History of Politics

James J Sheehan, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 17, pp. 11667–11673, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The first and still one of the greatest political histories ever written is Thucydides's *Peloponnesian War*, in which we find many of the themes that continue to fascinate students of politics. I shall begin by examining the development of international political history from Thucydides to the present. Another important subject of political history, also with classical roots, is the development of law and institutions. Here I will be especially concerned with the significance of legal history for the development of modern historiography. In the nineteenth century, political history became closely associated with the nation-state, whose story was at once the record and instrument of nationbuilding. But while political history is often allied with established authority, it can also become a weapon of political dissent by providing an alternative narrative to aspiring political parties, social movements, and subject nationalities. For most of the history of historiography, political history dominated the discipline. Beginning in the 1960s, however, first in the United States and then in Western Europe, traditional political history came under attack, first by social, then by cultural historians. After considering the character of the ensuing debate, I will conclude with some tentative thoughts about the role of political history in the future.

Political history is the study of how legitimate power has been defined, distributed, and justified in the past (*see* Power). Two imperatives, both with ancient roots, link historiography and politics. The first is the need to affirm the identity of the political community by uncovering its roots in the past. This search for historical identity engaged both the Hebrews and Greeks at the very beginning of European culture; it moved the Romans to search for the sources of their state and prompted succeeding generations to study the history of their cultures, religions, and nations. "No question," wrote Jean Bodin in the sixteenth century, "has exercised the writers of histories more than the origin of peoples" (Kelley, 1998). The second link between history and politics is the desire to find lessons in the past that might help solve the political problems of the present. "I shall be content," Thucydides remarked in the first chapter of *The Peloponnesian War*, "if (this history) be judged useful by those inquirers who desire an exact knowledge of the past as an aid to the understanding of the future" (Thucydides, 1996).

From the Ancient World to the End of the Middle Ages

Political history in the West begins with two masterpieces: Herodotus's history of the Persian Wars, written between 445 and 425 BC and Thucydides's history of the long conflict between Athens and Sparta, which he began in 424 BC and left unfinished when he died a quarter of a century later. Products of the extraordinary burst of cultural energy that transformed fifth- and fourth-century BC Greece, Herodotus and Thucydides combined a deep sense of the importance of their own age – Herodotus wrote about the recent past, Thucydides about events in which he himself had taken part – with an awareness of the historian's responsibility to establish the truth about the past. Both men reflected on the problems of historical research, the reliability of eye witnesses, and the necessity of critically analyzing source material. More than any particular way of writing history, this was their most significant legacy for future historians (*see* Historical Thought and Historiography: Classical Period (Especially Greece and Rome)).

The study of history was more important in Rome than it had been in Greece, perhaps because Romans were always conscious of their complex historical relationship to the Greek culture that they had both vanquished and absorbed. Roman historians examined the problem of politics from a variety of perspectives; they wrote about leadership and institutions, military strategy and court intrigue, republican virtues, and imperial ambitions. But behind these disparate approaches was a common theme: the origins, triumph, and – eventually – the decline of the Roman state, which, as republic and empire, dominated the public life of the Mediterranean world for 700 years and then continued to haunt Europeans' historical imagination until well into the nineteenth century.

The emergence of Christianity and its adoption as the Roman Empire's official religion transformed all of Western culture, including both politics and history. Christianity's most immediate contribution to political historiography was to pose the problem of church and state, which would decisively shape the theory and practice of European politics from the end of the Roman Empire to the beginning of the twentieth century. Unlike the pagans whom they displaced or the Hebrew traditions upon which they built, Christians created religious institutions that were connected to, but not coincident with, the sources of political authority. "Spiritual and secular power," Ranke (1962) wrote, "could approach one another and be intimately related, but they could be fused only under exceptional circumstances and for a brief period of time. Their relationships and conflicts provide one of history's most important themes." From Eusebius's history of the church, which reflected the triumph of Constantine's conversion at the beginning of the fourth century, to Otto of Freising's twelfth-century chronicle, *The Two Cities*, which described the conflict between empire and papacy, to Ranke's own history of the popes, which bore the marks of nineteenth-century's struggles

between church and state, historians searched the past for clues to the proper relationship between God and Caesar, pope and emperor, altar and throne, religious autonomy and secular authority (*see* Historiography and Historical Thought: Christian Tradition).

History and the State

Beginning in the 1490s with the French invasion of the Italian Peninsula, a series of international, constitutional, and religious crises disrupted European politics and helped to create a new subject for political history. This subject was, in Quentin Skinner's (1978) concise summary, "the concept of the state – its nature, its powers, its right to common obedience." The concept of the state emerged from the dissolution of two closely connected assumptions that had shaped European political thought since Augustine: first, that the secular realm of politics had to be seen as part of a sacred order; and second, that Europe's components – its free cities, ecclesiastical states, and various monarchies – were part of a single polity, governed by the universal laws of the church and sustained by the institutional legacies of imperial Rome. A good deal of medieval political theory and practice had been devoted to reconciling these assumptions with the stubborn realities of political life; in the first decades of the sixteenth century, Europeans began groping their way toward new ways of thinking about politics – and, therefore, toward a new kind of political history.

One of the first signs of this can be found in the work of Niccoló Machiavelli. As a Florentine who had been forced into exile by the collapse of his city's constitutional order, Machiavelli felt the political crisis of the early sixteenth century at first hand. In *The Prince and The History of Florence*, Machiavelli sketched a political world that Thucydides would have recognized immediately, a world driven by struggles for power, beset by violence, subject to forces that are often beyond human capacities to control. A similar picture emerges in the work of Machiavelli's fellow citizen and contemporary, Francesco Guicciardini, whose histories provide an even darker picture of instability and violence. But however limited people's ability to control events, history, for both Machiavelli and Guicciardini, was a totally human enterprise, untouched by the hand of God, unfettered by a transcendent order.

Machiavelli wrote about 'the state,' but without giving it a clear institutional definition. This would be the great achievement of sixteenth- and seventeenth-century legal humanists, whose study of the law provided a new foundation for both politics and history. Of course law had always been an important part of European culture, but scholars like Guillaume Budé, Jean Bodin, François Baudouin, and François Hotman combined philological methods and historical analysis to create a new science of jurisprudence that enabled them to establish the origins, nature, and limits of the state's authority. Bodin, for example, provided the classic definition of sovereignty as the supreme source of law within a commonwealth, exercised by, but clearly independent from, the ruler's person. The work of these scholars stands at the source of one of political history's richest traditions: the study of legal institutions and constitutional development, which includes among its practitioners such giants as Montesquieu, Sauvigny, and Maine (Kelley, 1970).

Law was also important to those who sought a vision of the international order to replace medieval universalism. Hugo Grotius, for example, who wrote during the Thirty Years War in the first half of the seventeenth century, argued that the international order was based on natural laws, which were divinely ordained but could be discovered by examining the past; like the French legal humanists, Grotius regarded the study of law and history as part of a single enterprise. By the end of the eighteenth century, under the influence of a new set of crises generated by the French Revolution (*see* French Revolution, The) and Napoleonic conquests, a number of scholars searched for an international system that was no longer based on God-given laws but was still subject to underlying principles that the historian could discover. Heeren (1834), for instance, wrote that the 'essential property' of the European state system was its "internal freedom, that is, the stability and mutual independence of its members. To set forth how this was formed, endangered, and preserved…is the great object of the historian."

The historical image of a world of independent states struggling for existence within an international system was given its definitive formulation in the middle decades of the nineteenth century by Leopold von Ranke, perhaps the greatest political historian of the modern era (*see* Historiography and Historical Thought: Modern History in the West since the Eighteenth Century). Von Ranke is best known for his archival work and his insistence on the critical use of primary sources. But inseparable from Ranke the great empiricist was Ranke the moral philosopher of statehood, who regarded states as the 'ideas of God' that move through the historical order 'like celestial bodies.' His work, which combined scholarly rigor with stylistic grace, was read widely and admired throughout Europe; his methods of teaching and research were adopted everywhere (see Krieger, 1977).

History and the Nation

Of all kinds of history, the history of the nation has the deepest roots and the most widely extended branches (*see* Nation: Sociological Aspects of). Its first and most persistently influential expression was in the Hebrew Bible, which recounts the archetypical story of a nation's origins, defines its special historical mission, traces its survival in the face of terrible trials, and promises its eventual triumph. Their story, the Hebrews' historians insisted, is a story like no other – among the many ways in which national histories resemble one another, none is more striking than their insistence on being unique. As for so many other nations without states, this national history sustained the Jewish people's sense of collective identity through centuries of dispersal and oppression. And once the Jewish nation finally had its state, the story of the national past became – and this is also characteristic of the genre – the site of political debates about the nation's present identity and future direction (*see* Historical Thought and Historiography: Jewish Tradition).

Although national history has a genealogy that extends to the dawn of Western culture, its most intellectually vigorous and politically potent versions were written in the nineteenth century when, for the first time, political history became virtually synonymous with the history of nations and nation-states (*see* State, History of). The nineteenth century was, of course, the great age of historical writing, a time when historical consciousness touched virtually every aspect of European culture, from art and literature to philosophy and the social thought and even the natural sciences. For most nineteenth-century intellectuals, to think about the world was to think about historical origins, evolution, and identity. This was a time when statesmen – men like Guizot and Macaulay – wrote history and when historians – Bancroft, Droysen, Treitschke – were active in political affairs. And this was also the era when the institutional foundation for modern historical scholarship – document collections, libraries and archives, journals and reference works – was created, much of it by, and often for, national governments (*see* Historicism).

While every nation has a story to tell about itself, every national story is different, shaped by the political imperatives, cultural climate, and institutional arrangements of each national community. Nineteenth-century French historians, for instance, whose state had an unusually long and apparently uninterrupted existence, were able to imagine a single national history. Their problem was how to find a place in this history for the trauma of revolutionary upheavals and prolonged warfare that engulfed France after 1789. For almost two centuries, French historiography was shaped by political struggles over the revolution's legacy, just as French politics was influenced by historical struggles about the revolution's character and significance. Among the works engaged by this project are such classics of political historiography as Tocqueville's *Ancien Regime and the French Revolution* and Michelet's *History of the French Revolution* (*see* Tocqueville, Alexis de (1805–59)). It was not until the 1970s that François Furet could declare that the French Revolution was finally over, by which he meant that it had lost its central place in the nation's political discourse and public memory.

Each national experience poses a distinctive ensemble of questions for its historians to confront. In Britain, these are most often questions about institutional continuity rather than revolutionary ruptures. In the US, the central historical issue is the question of identity itself, which was addressed in works like Frederick Jackson Turner's classic study of the frontier in American history (*see* Frontiers in History). The problem of identity also dominated German historiography in the nineteenth century: before the creation of a German nation-state, this meant a competition between different versions of the national past, after national unification, it involved demonstrating that the nation created by Bismarckian 'blood and iron' was the only authentic destination for the German people's long historical journey toward political unity. In the twentieth century, German history was shaped by efforts to locate the catastrophe of National Socialism in the national past: was National Socialism, as some claimed, the destination of Germany's 'special path' (*Sonderweg*) to modernization, or was it a particularly toxic example of a broader crisis within modern culture and politics (*see* National Socialism and Fascism)?

An important element in the public life of every nineteenth-century state, history was even more significant for those national communities that did not have a state of their own. For the Poles, the Irish, the Czechs, and, in the course of the twentieth century, for the peoples of the world colonized by Europeans, a shared history was the surest source of present cohesion and the brightest hope for future nationhood. After all, what else did these nations possess besides their belief in a common ancestry? Often enough, this belief had to be constructed from the scattered remnants of an imagined past – literary works discovered and transcribed by patriotic scholars, ancient sites unearthed by archaeologists, memories of lost greatness brought to life in poems and plays and popular histories. Often enough the character of the national past was bitterly contested – by Czechs and Germans in Bohemia, Protestants and Catholics in Ireland, conservatives and liberals in partitioned Poland, Muslims and Christians in the Balkans. In these parts of the world, the historical significance of John Hus, whether the town was called Derry or Londonderry, the meaning of the Polish uprising of 1863, or how to commemorate the battle of Kosovo could swiftly become matters of life or death. Every national history reflects ideological divisions and political conflict, but in nations without states, the seam between politics and history was nearly invisible, so that it was usually impossible to determine where the one began and other ended.

Subversive Histories

Although historians have often been the servants of those in power, there is also a long, rich tradition of subversive historiography, written by scholars who turned to the past in order to question the established order and suggest new paths for the future. In the nineteenth century, at the same time that historians were drawn into the service of the nation-state, subversive historiography was given its most influential formulation by Karl Marx. For him, the political movements that had interested Thucydides, Machiavelli, and Ranke are surface phenomena, reflections of the deeper social conflicts from which history takes its shape and direction. Politics is merely an instrument of the ruling class, a way of justifying and defending its hold on the means of production. Traditional political historians, therefore, view the past through a distorting lens that turns the world upside down, obscuring the essential sources of change with their concern for artificial conflicts, empty slogans, and trivial events. The Marxist historian's purpose is to substitute for these ideological distortions the radical clarity of scientific materialism, which will reveal to the proletariat the roots of its oppression and thus help prepare them for their revolutionary role in history cosmic drama. From the last decades of the nineteenth century until the end of the twentieth, Marxism – in all its many forms – remained the most influential form of subversive historiography (*see* Marx, Karl (1818–83); Field Theory).

Well before the final collapse of international communism in 1989, Marxist historiography had begun to lose its intellectual attraction and political influence. In part this was because of the diminution of ideological energies both in the communist regimes of Eastern Europe and in labor movements

throughout the world. Moreover, by the 1970s, the political passions that had drawn intellectuals to Marxism during the struggle against fascism in the 1930s and against colonialism in the 1960s had cooled. The Marxist view of history was also undermined by a growing skepticism about all grand narratives, a turning away from the aspiration to grasp the meaning of human history as a whole. And finally, Marxism declined because it could not explain very well the subjects to which more and more historians now turned, especially problems of ethnicity, race, and gender, which did not fit into the categories defined by the means of production. Not surprisingly, the form of Marxism that remained most influential in Britain, the US, and most of Western Europe was associated with historians like E.P. Thompson, who tried – not always with success – to fuse class analysis with a concern for culture and politics (*see* Social History).

Since the decline of Marxism, the subversive impulse in historiography has been fed from many other sources of which the most important, at least in the last two decades of the twentieth century, was certainly the work of the French philosopher and historian, Michel Foucault (*see* Boserup, Ester (1910–99)). Like Marx, Foucault's enemies were oppression and inequality, his goals, freedom and community. But unlike Marx, he did not propose an alternative view of history or a critical science of society. Foucault's project was willfully fragmented; he called his approach to the past a genealogy rather than a narrative history. Deeply concerned about what he regarded as the repressive power of organized knowledge, he was nevertheless unwilling to abandon efforts to understand and subvert the existing order of things. Above all, this meant to understand and subvert power, which was exercised not only by the agents of the state but rather was diffused "in the multiple forms of subjugation that have a place and function within the social organism." Emancipation, therefore, required a struggle against all of what Foucault regarded as the "polymorphous techniques of subjugation" (Dirks et al., 1994).

Marx (and also, if much less influentially, Freud) had sought to subvert traditional political history by showing that politics was really a reflection of deeper social (or psychological) conflicts (*see* Psychohistory). Foucault subverted the conventional image of politics from the opposite direction, that is, not by showing that politics should be reduced to something else, but rather by insisting that everything – including what was once thought of as the most intimate spheres of private life – is politically charged. Both perspectives are useful since both can illuminate new problems and neglected connections. Whether they can – and should – lead to the abandonment of the subjects that have concerned political historians since classical times is much more open to question.

The Varieties of Political History in the Twentieth Century

Scholars' interest in the past has always been influenced by contemporary crises. Thucydides was moved to write his history by the Athenians' disastrous war with Sparta, Josephus recorded the destruction of the Jewish people by the Romans, and Augustine composed his universal history under the shadow of the Vandals' sacking of Rome. In the modern era, the French Revolution inspired Tocqueville, the Bolshevik Revolution, Trotsky, the rise of Italian Fascism, Salvemini, Nazi totalitarianism, Arendt, the World Wars, Churchill – the list might easily be extended to include many other occasions when, to borrow Hegel's phrase, men's hopes were destroyed on history's slaughter bench.

The War that shattered the world order in 1914 dominated the study of international political history for a quarter of a century (*see* First World War, The). Basing their work on the documents published by all of the belligerents, historians examined both the immediate causes of the war and its deeper roots in the great power rivalries of the nineteenth century. Some of this work was obviously partisan, written to absolve the historian's own nation of responsibility; but much of it sought to strike a balance between commitment and detachment. A notable feature of this historiography was the participation of American scholars. Responding to the hopes and disappointments that attended the failure of Woodrow Wilson's vision of a conciliatory peace, these scholars – Sidney Fay, Robert Binkley, William L. Langer, Raymond J. Sontag, and many others – played an active role in controversy over the causes and consequences of the war. By the 1930s, there was a rough scholarly consensus that no one power was responsible for the war, which was thought to have come from a series of miscalculations by several different governments. In the 1960s, this consensus was shattered by the work of Fritz Fischer and his students who attempted to blame the war on Germany's reckless bid for world power. Fischer's view was bitterly attacked and then widely accepted, but no new consensus has emerged to explain what George Kennan, himself a distinguished student of international history, once called "the great seminal catastrophe of the century" (Stevenson, 1997).

In contrast to World War I, the causes of World War II did not attract much scholarly attention; virtually everyone agreed that this was Hitler's war (*see* Second World War, The). But the war's character and consequences generated an enormous historical literature, including some great works by participants such as Winston Churchill and Charles DeGaulle. Historical controversies surrounded issues such as the air war (especially the decision to use atomic weapons against Japan), wartime diplomacy, and the role of resistance and collaboration in the territories occupied by Germany and Japan. Among the historical questions about the war in Europe none are more painfully complex than those surrounding the mass murder of European Jews, conventionally known as the Holocaust. Was it, as some scholars believe, tied to a distinctive sort of German anti-Semitism? Or did it arise within the climate of violence created by the war itself? Is it best seen as a problem of German history? Or of modernity as a whole? Could it have been prevented or at least diminished by the action of Germany's enemies? Central to the historical self-understanding of both perpetrators and victims, the Holocaust has become the century's exemplary disaster (Marrus, 1987) (*see* Holocaust, The).

The origins of the Cold War, the global geopolitical rivalry between the Soviet Union and the US that dominated international politics from 1945 to 1989, were a matter of intense scholarly controversy throughout its long history. Was the Cold War, as some claimed, a defensive response by the Western powers to communist aggression and expansionism?

Or, as the so-called 'revisionists' argued, did the US over estimate, perhaps even manufacture, the communist threat in order to advance its own aims throughout the world? Until the collapse of the Soviet Union, both sides in this controversy depended largely upon sources from the Western powers, but after 1989, the empirical basis of Cold War historiography was transformed by the opening of archives in both the former Soviet Union and its Eastern European allies. As a result, historians have begun to create a new picture of international history in the second half of the twentieth century (Gaddis, 1997).

Among the great themes of world history is the expansion of European power which began in the late fifteenth century and ended during the second half of the twentieth century, when the colonial states either abandoned or were expelled from their overseas possessions (*see* Imperialism, History of). European expansion was the subject of many important historical studies, including the analyses by J.A. Hobson and V.I. Lenin that inspired several generations of anti-imperialist writers. Decolonization also had a lively historiography, stimulated in part by the traumatic wars fought by France in Algeria and by the US in Indochina. In the 1990s, historians became interested in what became known as postcolonialism, that is, the persistent and reciprocal relationship between colonized and colonizer both during and after the colonial experience itself. While many postcolonial studies take up issues that seem far distant from conventional political history – for instance, questions of gender, sexuality, and cultural identity – they are all connected to the ways in which political power is defined and distributed.

In the last decades of the twentieth century, political historians have become increasing aware of the importance of regional and loyal allegiances and institutions. A number of recent works on the revolutions of 1848, such as Jonathan Sperber's account of the Rhineland, have shown how tensions between central and local authorities affected both the origins and outcome of the revolution. Historians have also become sensitive to the historicity of the national state: Eugen Weber, for instance, has described the long and uneven process through which the French state imposed its values in the countryside, while Linda Colley, among many others, has uncovered the way in which Britain became a focus of people's interests and loyalties. The upheavals that destroyed Yugoslavia in the 1990s produced a cluster of new works on the historical sources of ethnic identity and changing nature of nationalism in the twentieth century (*see* Nationalism: General; Nationalism, Historical Aspects of: The West). In these studies, traditional questions about state power and politics are enriched by a new concern for ethnicity, culture, and identity.

Conclusion: Political History in the Twenty-First Century

Although it no longer has the unquestioned historiographical primacy it once enjoyed, the history of politics remains a vigorous and remarkably diverse enterprise. For the general public, political history is a source of apparently endless fascination. Each year a staggering number of books, especially biographies and battle narratives, are published about subjects such as the American Civil War and World War II. Among professional historians there are signs that political history may have recovered some of the ground lost to social and cultural history in the 1970s and 1980s. An increasing number of scholars are now seeking to apply social scientific methods and concepts to the study of both domestic and international politics (*see* History and the Social Sciences). At the same time, insights from cultural anthropology (*see* Anthropology and History) and literary criticism have raised interesting new questions about the nature of political power and identity.

The history of politics is important because individuals' lives are shaped and limited by political institutions, while the very existence of their communities depends on how political power is exercised and restrained. The illusion that politics does not matter is a luxury that only those fortunate enough to live in a rich and orderly state can afford. Moreover, while the importance of national history may have declined in Western Europe, for much of humanity the question of national identity remains vitally relevant; in many places people are still willing to die – and to kill – in behalf of their own version of the national story. And the emergence of a new set of problems and opportunities from the wreckage of the Cold War system (*see* Cold War, The) will certainly raise a new set of questions about the past and future of the international politics.

Does knowledge of the past help us, as Thucydides and Machiavelli and so many others believed, to master our own political problems? To be sure, few now share the confidence expressed by Lord Acton in his inaugural lecture of 1895 that "the science of politics is the one science that is deposited by the streams of history, like the grains of gold in the sand of a river; and the knowledge of the past, the record of truths revealed by experience, is eminently practical, as an instrument of action and a power that goes to making the future" (Acton, 1957, pp. 25, 26). And yet the impulse to extract lessons from the past seems irresistible; historical references and analogies are everywhere a prominent part of political discourse. If nothing else, the study of past politics should help to assay the ore left by the stream of history and to recognize the fool's gold so often on display in the rhetoric of public life.

See also: Citizen Participation; Democracy, History of; Democracy: Normative Theory; Democratic Transitions; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; Parliaments, History of; Political Anthropology; Political Economy, History of; Politics and Participation; Republicanism: Philosophical Aspects.

Bibliography

Acton, J., 1957. Essays on Freedom and Power. Meridian Books, New York.

Breisach, E., 1983. Historiography: Ancient, Medieval, and Modern. University of Chicago Press, Chicago.

Burrow, J.W., 2008. A History of Histories: Epics, Chronicles, Romances and Inquiries from Herodotus and Thucydides to the Twentieth Century. A.A. Knopf, New York.

Dirks, N.B., Eley, G., Ortner, S.B. (Eds.), 1994. Culture/Power/History. Princeton University Press, Princeton, NJ.

Finkelman, P., Wallenstein, P., 2001. The Encyclopedia of American Political History. CQ Press, Washington, DC.
Gaddis, J.L., 1997. We Now Know: Rethinking Cold War History. Clarendon Press, Oxford, UK.
Gilbert, F., 1965. Machiavelli and Guicciardini: Politics and History in Sixteenth Century Florence. Princeton University Press, Princeton, NJ.
Gooch, G.P., 1959. History and Historians in the Nineteenth Century. Beacon, Boston.
Heeren, A., 1834. A Manual of the History of the Political System of Europe and Its Colonies. Clarendon Press, Oxford, UK.
Iggers, G.G., Parker, H.T., 1979. International Handbook of Historical Studies. Greenwood Press, Westport, CT.
Kelley, D.R., 1970. Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance. Columbia University Press, New York.
Kelley, D., 1998. Faces of History: Historical Inquiry from Herodotus to Herder. Yale University Press, New Haven, CT.
Krieger, L., 1977. Ranke: The Meaning of History. University of Chicago Press, Chicago.
Kamrava, M., 2005. The Modern Middle East: A Political History since the First World War. University of California Press, Berkeley.
Kazin, M., Edwards, R., Rothman, A., 2010. The Princeton Encyclopedia of American Political History. Princeton University Press, Princeton.
Laistner, M.L.W., 1963. The Greater Roman Historians. University of California Press, Berkeley, CA.
Marrus, M., 1987. The Holocaust in History. University Press of New England, Hanover, NH.
Momigliano, A., 1990. The Classical Foundations of Modern Historiography. University of California Press, Berkeley, CA.
Mackay, C.S., 2004. Ancient Rome: A Military and Political History. Cambridge University Press, Cambridge.
Price, J.J., 2001. Thucydides and Internal War. Cambridge University Press, Cambridge.
Phillips, K., 2002. Wealth and Democracy: A Political History of the American Rich. Broadway Books, New York.
Plant, R., 2001. Politics, Theology, and History. Cambridge University Press, Cambridge, UK.
Ranke, L. von, 1962. Die Römischen Päpste in den letzten Vier Jahrhumderten. Phaidon Verlag, Vienna, Austria.
Stevenson, D., 1997. The Outbreak of the First World War: 1914 in Perspective. St. Martin's Press, London.
Thucydides, 1996. The Landmark Thucydides. Free Press, New York.
Zelizer, J.E., 2012. Governing America: The Revival of Political History. Princeton University Press, Princeton, NJ.
Zagorin, P., 2005. Thucydides. Princeton University Press, Princeton, NJ.

Political Lawyering

Terence C Halliday, American Bar Foundation, Chicago, IL, USA
Lucien Karpik, École des Mines and École des Hautes Études en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political lawyering refers to a mode of lawyers' collective politics in struggles over the rise and fall of political liberalism. Extensive historical and comparative research demonstrates that lawyers, in certain circumstances, will fight for a moderate state, civil society, and basic legal freedoms. Those struggles always occur through the politics of a legal complex of all legally trained occupations continuing to act in a legal capacity. Configurations of the legal complex frequently ally with civil society groups, including the media, justice-related associations, and religious organizations. Consequently, transitions from authoritarian and illiberal regimes to liberal political societies, including postcolonial societies, frequently are spearheaded by political lawyers in coalitions with elements of the legal complex and civil society. This approach to studying the politics of lawyers is contrasted with the study of cause lawyering and of lawyers' role in transforming state power.

Scholarship on the comparative and historical politics of lawyers took a significant turn in the late 1990s. In previous decades, two sharply contrasting approaches to lawyers' politics had been observed.

At the height of functionalist theory, which dominated the theory of professions during the years 1950–70, politics was essentially absent from studies on the link between knowledge and social needs. Society delegated to a community of professionals a set of duties, undergirded by 'institutionalized altruism,' and rewarded professionals highly for their fiduciary contributions to society. Neither politics nor collective action, nor historical or comparative analysis, were needed to explain the integrative functions of lawyers.

A sharp critique of functionalism then emerged from a sociology that was deeply skeptical about lawyers' professional ideologies and practices of 'service' and civic-mindedness. Magali Sarfatti Larson's (Larson, 1977) highly generative book, *The Rise of Professionalism*, presented a counternarrative to functionalist accounts that managed to neglect class, conflict, and occupational self-interest. Lawyers, and other professionals, had engaged historically in collective action for a double purpose: collective upward mobility in the class system and creation of professional services as a commodity. From this nuanced and sophisticated analysis was derived a theory of market control whereby the politics of lawyers became a mechanism of rent seeking. Lawyers mobilized, argued Abel and Lewis (Abel and Lewis, 1988a,b) and many of their collaborators, to monopolize markets and maximize profits. Collective action was largely taken for granted and mobilization for reasons other than economic interests entirely neglected.

Even the paradigmatically ecological theory of professions presented by Abbott (Abbott, 1988), which offered acute insights into ways that professions mobilize knowledge in their scramble for control over work, did not treat either the political processes by which professions secure and maintain jurisdictional control, or the activities of professions beyond control of work domains.

In the 1990s, several strands of scholarship on the politics of lawyers emerged. New line of historical and comparative scholarship emerged to critique the limitations of a market-oriented lawyers' politics and to explore theoretically and empirically the engagement of lawyers in the politics of political liberalism, state building, and cause advocacy. This article encapsulates the broad contours of lawyers' collective politics in struggles over the rise and fall of political liberalism, a mode of action labeled 'political lawyering.' The article concludes with brief contrasts between political lawyering and alternative formulations of lawyers' political activity that also emerged in the 1990s, which emphasizes lawyers as utility maximizers or as cause lawyers.

Political Lawyering

Extensive historical and comparative research cumulatively has demonstrated that fights for political liberalism have been integral to the politics of lawyers and to the politics of states in a wide variety of times and places (Halliday and Karpik, 1997; Halliday et al., 2007a, 2013). These politics, while diverse and variegated, nevertheless frequently register a coherence that can be bundled within the concept of political liberalism.

Political Liberalism

For lawyers, political liberalism has three principal components:

1. A moderate state, which fragments political power into constitutionally structured contestation among elements of the state, including legislatures, executive agencies, and courts. For lawyers, the most critical element of the moderate state is a degree of autonomy of the judiciary from other branches of government.
2. Civil society, which is constituted as social organization that exists substantially independent of the state. Civil society expresses itself in two forms. It is a web of associations that fills a middle ground between the state and tribe or family. It does not owe its existence to the state and can stand against it. Associational life is expressed empirically in voluntary groups, religious organizations, unions,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86092-4

interest groups, women's associations, among others. Civil society constitutes a public sphere where opinions are freely expressed, ideas circulate, and politics becomes informed by discursive encounters (Habermas, 1989). For lawyers, their associations both constitute civil society and look to civil society for political allies and supportive publics in struggles for political liberalism.

3. Basic legal freedoms, which comprise three parts. First, there are core legal rights to protect citizens from arbitrary state power (e.g., rights to due process in law, habeas corpus, right to a lawyer, freedom from arbitrary arrest, torture, death). Second, there are basic political rights of freedom of speech, association, and movement, all of which are foundational for the existence of civil society. Third, there are property rights, which offer protections for citizens against arbitrary appropriation of land or other property by the state or others.

Three important qualifications must immediately be noted. First, there is nothing inevitable or invariant about lawyers' mobilization for political liberalism. Because lawyers frequently do *not* mobilize on behalf of these three features of political liberalism, this is a contingent theory that requires attention to the conditions under which mobilization occurs. Second, when lawyers do mobilize collectively most often it is around this tight set of goals. Like other citizens, lawyers divide on many substantive political issues, but even deeply divided legal professions can find common ground on the core attributes of political liberalism. That implies the use of ideal types. Third, lawyers historically have not mobilized as a whole for economic and social freedoms. The political liberalism of lawyers is limited.

In sum, wide-ranging interdisciplinary research across the world and on the last 300 years to the present shows that a political commonality can frequently be found in lawyers' readiness to mobilize collectively to build political liberalism or to fight against retreats from a moderate state, the dismantling of civil society, the control of the public sphere, or erosions of basic legal freedoms.

Five Propositions

In European and North American history, the political activism of the legal profession as a periodic but by no means consistent mode of action could be expressed in five propositions:

1. The autonomy of the judiciary and the autonomy of the bar are the principal conditions for the fight of lawyers on behalf of a moderate state. Unless the courts have some measure of independence from executive power, or judges are prepared to fight for autonomy from political control, then effective mobilization will be unlikely. And unless the bar itself has a degree of independence from executive or other state control, then its collective action will be impaired or impossible. It is not surprising, therefore, that struggles for judicial autonomy by the French *parlements* in eighteenth-century France (Karpik, 2000), or periodic defenses of the English bar to maintain their independence as 'little republics' in the English body politics, or contemporary struggles by a vanguard of lawyers in China, have resonances in common.
2. Historically, lawyers' collective defense of civil rights did not extend to economic and social rights. Most often lawyers' collective action has been possible only for core civil rights, not for second or third generation rights, such as universal education, social welfare, universal voting rights, and the like. Rarely did lawyers as a whole mobilize again racism, or xenophobia, or gender discrimination.
3. Lawyers most often act reactively, not proactively. That is, they were provoked into action by arbitrary or illegal actions by governments. Sometimes mobilization coalesces around *causes célèbres*, mostly notably, widely publicized court trials. They frequently have sought solidarity and support from publics and civil society.
4. Courts can become the central locus of battles on behalf of liberal political society. Courts have three benefits for lawyers' politics: they are often protected spaces, allowing speech and opinions that would not be permitted elsewhere; they are lawyers' battleground of choice, since they operate by a *legal* rather than a political logic, and thus give lawyers a technical advantage; and they are the arena into which society's most vexing disputes about state power eventually can appear (Karpik, 2007).
5. Lawyers have generic capacities to act as spokespersons for publics. Insofar as lawyers are occupationally charged with representation of others, they frequently parlay that individual lawyer–client role into a collective lawyer–society role where they may claim to speak to power on behalf of publics and civil society in favor of political liberalism (Karpik, 2007).

The Legal Complex and Political Liberalism

As research findings accumulated from Latin America, Africa, and Asia, it became clear that a politics of lawyers alone was inadequate as an explanatory theory of mobilization on behalf of political liberalism. Almost always lawyers' mobilization was intricately intertwined not only with the actions of judges but with other legally trained occupations. These findings led to the naming of a new collective actor – the legal complex (Karpik and Halliday, 2011).

The Legal Complex Is Defined by Five General Characteristics

1. It is composed, in principle, of *all legally trained occupations* that are continuing to act in a legal capacity. It includes practicing private lawyers, who usually form the core of the legal complex; legal academics; judges; prosecutors; law officials inside government; and lawyers in the military. In some countries, it includes grassroot lawyers who offer legal advice or lawyers whose licenses have been withdrawn for political reasons.
2. It is defined behaviorally in terms of *current actions*, not past training. It excludes persons exclusively trained as lawyers, but who later become government officials, business persons, nonprofit leaders, politicians, and are not acting in legal roles in those positions.
3. It is a concept intended to capture *collective action*. That is, the legal complex putatively acts across the conventional divides that segregate academic research on courts and judges from those on private lawyers from those on prosecutors.

4. It expresses itself in varieties of *structures*. Research on the legal complex requires attention to the structures that enable or preclude mobilization across conventional cleavages among categories of legally trained professions. For instance, in relations between lawyers and judges within the legal complex, analysis will attend to relations that are cooperative, or oppositional, or detached, or crosscutting, or multiplex. Each kind of relationship and each structural property may have explanatory significance for capacities to mobilize. In other words, the diversity of the interdependency in relations among groups is expressed by different structures and by different types and degree of mobilization.
5. It is highly *dynamic*. Not only may the legal complex mobilize in very different ways on different issues, but the legal complex may be differentially constituted and structured in different places and times across different aspects of political liberalism.

Dynamics of the Legal Complex toward Political Liberalism

Research on several continents in the twentieth century shows that mobilization by the legal complex has occurred in at least four patterns (Halliday et al., 2007b). These show the flexibility of the concept for social science and historical analysis and the contingency of actions by the legal complex, which require further specification in the development of theory on political lawyering and political liberalism.

1. There are many instances where the legal complex in several of its components – lawyers and judges, sometimes with academics and prosecutors, or with civil servants or military lawyers – has acted on behalf of basic legal freedoms. In several countries, this has led to relatively robust liberal political societies. Mobilization by many lawyers, then judges, prosecutors, and legal academics was integral to dramatic regime transformation in South Korea and Taiwan as military dictatorships yielded slowly to constitutional political orders. In other countries, such as Egypt in the 1990s, or Venezuela in the 1980s and 1990s, mobilization by a vanguard of lawyers before sympathetic courts led to short-lived legal and political openings, which eventually resulted in political or military backlash (Moustafa, 2007). Even in colonial Hong Kong, some government lawyers found common cause with barristers to defend basic rights.
2. There are occasions when lawyers mobilize without the support of judges and therefore may be compelled to rely more on legal academics or find support outside the legal complex in civil society. In Japan from 1886 to the1920s and after the 1980s, lawyers pressed for politically liberal reforms, even for a more active and autonomous judiciary, but they did so in the face of judicial passivity. In China since the early 2000s, a vanguard of lawyers has sought to mobilize the bar as a whole, with some support from the legal academy, to protect basic legal freedoms, but it has confronted an 'iron triangle' of judges, police, and prosecutors who have resisted. After 9/11 in the United States, many private lawyers, some bar associations, and even military lawyers sought greater protection for detainees, but judges were slow to permit those protections.
3. In some situations, a legal profession and judiciary, which previously has fought for political liberalism instead fails to mobilize. Two cases illustrate where selective mobilization can occur, both focused on security. In one set of cases, mobilization fails in response to *external* threats to security. In Israel, an otherwise vocal legal complex remained silent about certain illiberal policies of Israel's state and courts. In the U.S., the fear of terrorism, which followed 9/11, temporarily silenced otherwise activist lawyer groups. In another set of cases, rampant criminality *internal* to a country, such as Argentina and Brazil in the 1990s, led government prosecutors and judges to withhold action against police brutality and killings of suspects (Brinks, 2008).
4. There are many instances where all or most of the legal complex were hostile to defense of basic legal freedoms. Research shows that lawyers and judges in 1930s Weimar Germany, lawyers in 1920s Fascist Italy, lawyers and judges in militarizing Japan in the 1930s, lawyers confronted with Pinochet's State of Siege in 1970s Chile, among others, could not or did not stand collectively or in force against dictatorial regimes.

Political Liberalism in the Postcolony

While much research showed evidences of political lawyering in widely diverse polities across time and place, it did not confront one of the most fundamental transformations in twentieth-century politics – the fall of the European colonial empires. Since British imperialism at once held in tension the contradictory actions of illiberal conquest with tempered liberal ideals of rule of law, it became an empirical and theoretical challenge to discover whether political lawyering extended to resistance to illiberal colonial politics and to construction of postcolonial liberal polities.

Postcolonial histories and world histories have yet to study extensively the politics of lawyers and the legal complex in the postcolony. Halliday et al. (2013) build upon interdisciplinary scholarship on South Asia, South East Asia, and Africa to propose that postindependence countries exemplified three political trajectories. The role of the legal complex varied systematically across these trajectories.

A *liberal-legal* political order emerges and is sustained in countries such as India and Namibia. Despite their differences and deviations, these countries historically have shared extension in practice of basic legal rights, including core political rights, for their citizens; place legal restraints on state power; limit recourse to use of emergency powers; and permit growth of civil society and an open public sphere.

A *despotic* political order, observed in states as diverse as Sudan, Singapore, and Sri Lanka, severely limits exercise of core political rights, minimizes restraints on executive powers by judiciaries, relies extensively on emergency powers for the security apparatus, and offers little protection of basic legal freedoms to those deemed political opponents of an authoritarian or totalitarian state.

A *volatile* political order, evident in Malaysia, Pakistan, Zambia, and Malawi, lurches back and forth historically in its movements sometimes toward political liberalism and

sometimes toward despotism but never institutionalizes either of these. These almost cyclical histories are distinctive in themselves because they persist over decades.

Research and theory is underdeveloped on the extent of legal complex involvement in these respective trajectories, but some preliminary conclusions have been proposed.

It appears that the probability of a postcolonial polity that is politically liberal increases (1) if the legal complex already had been fighting for independence during the colonial period; (2) if divisions of race, language, and religion are transcended in microcosm within the legal complex by common commitment to liberal legalism; (3) if the legal complex has found allies in civil society; and (4) if a national legal complex can draw upon an international legal complex for resources and support.

The probability of a postcolonial polity becoming a despotic order appears higher when (1) a colonial bar or bench has been passive in its relations with the colonial regime, (2) when politically active lawyers are targeted for repression by a state and bar associations are controlled or silenced by the regime, (3) when courts are packed with politically subservient judges, (4) when the legal complex is so divided by race or class or religion that it cannot mobilize as a collective legal actor, and (5) when states repress associations and voices in civil society that are potential allies for a politically liberal-legal complex.

A volatile political order appears to be associated with an unstable legal complex where: (1) the bar retains a memory or has a vision of a liberal-legal political order, but is too weak or divided to fight for it; (2) there is a judiciary that follows political currents rather than has some autonomy from politics; and (3) there exists a civil society that also expands and contracts as political freedoms open or close.

International Contexts for Political Lawyering

During the heyday of the British Empire, the seeds of political liberalism were scattered across the world alongside a coercive colonial apparatus that those seeds would eventually dislodge. Following World War II, the breakup of the European empires, and a resurgence of globalization and law, political lawyering in most countries could not be entirely localized, even by authoritarian regimes. Although research remains limited, fragments of evidence show that international influences often matter and matter a great deal. They may flow across borders in the form of ideas, as in the influences of Continental European thinkers on judges in Franco's Spain. They may be facilitated by close ties built on security relationships, such as the impact of the U.S. rights ideas on Korean and Taiwanese political lawyers in the liberalization that gathered momentum in the 1980s. They include the influences of international bar associations and legal NGOs, which may monitor countries, issue reports, and attempt to shame repressive governments. They stem from global norms promulgated by the U.N. or human rights or rule of law organizations. They can be tangible, in flows of money and advice, as was the case in Egypt in the 1990s and early 2000s. They are magnified by the international media, as contemporary worldwide coverage of China's human rights abuses frequently feature political lawyers.

These studies point to attributes of states and civil society before and after independence that interact with the properties of the legal complex to more or less facilitative of political lawyering. The variety and complexity of these remain understudied, especially as they vary across empires.

Political Lawyering and Civil Society

Repeatedly, in metropolitan, colonial, postcolonial, and other societies, the interplay between political lawyering and civil society constitutes a key part of the history and theory of political liberalism. In general it can be concluded that the availability of civil society associations as political partners for the legal complex depend upon: (1) whether associational life is viable; (2) what ideological orientations are adopted by associations, most importantly, whether they engage 'political' issues; (3) which leaders and resources are available to social groups outside the state; and (4) whether voluntary associations are prepared to join alliances or even be led by the legal complex.

Four types of civil society groups recur in lawyers' politics:

1. Justice-related associations may be most proximate to the legal complex and therefore are frequent political allies or even prompts for the legal complex. These include human rights groups, monitors of police behavior or judicial decisions, drafters of manifestoes for legal/political change, critics of the justice system, and defenders of prisoners. Frequently they are led or heavily populated by lawyers.
2. The media offer much to political lawyers and the legal complex but the media also have an ambivalent position in struggles for political liberalism. On the one hand, there are many instances (e.g., Spain, Egypt, Brazil, India, Malaysia) where the old media have reported mobilization by lawyers and judges, published op-eds by legal academics, publicized abuses of rights, shown corruption in the courts, among others. On the other hand, the media frequently are targets of state control (e.g., Singapore, Venezuela) or are restrained by market pressures precisely because they threaten the authoritarian rule directly or in alliance with the legal complex (Rajah, 2012). A significant new development for many legal professions (e.g., China) has been the explosion of new media, which allow instant reports, graphic images, and exchanges between political lawyers and publics, in some cases producing instant public action.
3. Religious groups also have an ambivalent relationship with political liberalism. While earlier histories (Bell, 1994) found common cause between a 'Protestant-like' strand of eighteenth-century Catholicism and struggles by lawyers and judges for liberal politics, there is much evidence of Christian groups, which have resisted political liberalism. Numerous twentieth-century examples can be found of religious groups mobilizing with the legal complex, including fragments of the Roman Catholic Church in Franco's Spain, Pinochet's Chile, Brazil's Dirty War, Lee Kuan Yew's Singapore, and Zambia's resistance to emergence of a Big Man regime. Protestants in Kenya, Zambia, and South Korea and prominent leaders within

the legal complex itself, such as China's *weiquan* lawyers, similarly have mobilized. Too little is yet known about other world religions, such as Buddhism, Islam, and Taoism. Not enough attention has been paid to whatever affinities there might explain the resilience of liberal politics in India and its predominant Hindu religion. Yet there are many instances where religion has been complicit, or indifferent, or silent, or divided in the face of political repression, including repression against lawyers. A contingent theory of religion and political lawyering remains to be developed, especially given religious resurgence across the world in recent years.

4. Political parties, too, can support or subvert political liberalism. One-party states by definition remain largely resistant to political liberalism since the essence of the moderate state is the checking of state powers controlled by a single party. But multiparty politics can also subvert the legal complex if and when the politics within the legal complex are not a politics of legalism but an internalization of statewide party contests. In this latter case the legal complex is riven with party conflict and the prospect of transcending party politics to find a common ground of legalism (e.g., defense of basic legal freedoms) dims or fades away completely. This has been an enduring problem in Latin American tensions between political liberalism and democracy (Hilbink, 2007).
5. In some African countries, the national grassroots influence of women's groups has been an aid to political lawyering for liberal politics. Trade unions, if independent and if not deliberately crushed or controlled, also offer an organizational ally for activist bar associations.

Nevertheless, as Karpik demonstrated for eighteenth-century France, political mobilization can be spearheaded by lawyers acting as spokesmen for publics (Karpik, 2000). Notable instances in Pakistan and Malaysia since the 1990s demonstrate that this can occur collectively, when the bar as a whole takes to the streets and has its voice magnified by old and new media. Yet individual notables in the legal complex – a chief judge, a leading barrister, a law dean – can also purport to speak on behalf of publics as if they had been appointed spokespersons. In these senses, the public square can be infused with debate led by notables in the legal complex or even by lawyers in the new media whom authoritarian governments would prefer to render invisible.

Alternative Approaches to Lawyers and Politics

Political lawyering as a distinctive approach to understanding lawyers' politics, which emphasizes their collective action through the legal complex for the goal of political liberalism is compared with two other approaches to the analysis of lawyers' politics.

Cause Lawyering

A substantial body of scholarship accumulated in the 1990s and 2000s on cause lawyering. Scheingold and Sarat, the leaders of this scholarly movement, stated, "cause lawyering is about using legal skills to pursue ends and ideals that transcend client service" (Scheingold and Sarat, 2004: 3). Unlike the classic image of the lawyer as a 'hired gun,' exercising zealous advocacy and technical expertise irrespective of the merits of a client's case or its moral implications, the common element of cause lawyering is that the lawyer mobilizes the law on behalf of clients whose moral and political ideals they share. By mobilizing the arsenal of legal weapons on behalf of their own ideals, cause lawyers meld their personal beliefs with their legal work on behalf of clients who share their values. Nevertheless, Scheingold and Sarat argued, a single, cross-cultural and historically valid definition that would encompass all cause lawyering was not possible.

The range of causes supported by cause lawyers can be enormously variegated, ranging from pay equity, animal rights, capital punishment, Bedouin land rights, struggles against military repression and capital punishment, fighting for women's rights, peasant movements, and human rights (see case studies in Sarat and Scheingold, 1998, 2001, 2006). In their definitive summing up of 15 years of cause lawyering writing and research, Scheingold and Sarat show how the concept became increasingly expansive over time. Although initially focused on causes associated with the liberal left in the U.S. and elsewhere, the definition came to embrace lawyering on behalf of causes associated with the political right as those activities expanded, including antiabortion, libertarian resistance to state regulation of the environment, and property rights.

Cause lawyering is a form of doing politics (Scheingold and Sarat, 2004). Politics comes in two guises in this literature.

On the one hand there are *political ends*. Thus "cause lawyering is first and foremost a creature of and contributor to liberal democracy" (Scheingold and Sarat, 2004: 122). Yet, cause lawyering can also embrace ideals that may be antithetical to liberal democracy, such as three movements they label as evangelical advocacy, social democratic, and emancipatory democratic politics (Scheingold and Sarat, 2004: ch. 5). In principle, it appears, lawyers' causes could extend to defense of illiberal politics and authoritarian states. That is, the definition of 'cause' is now emptied of any particular political content.

On the other hand there are *political means*. Cause lawyers may mobilize through the law itself, utilizing its weapons and waging their campaigns in courts and other legal arenas. Cause lawyers may also mobilize outside the law, using conventional political strategies, such as aligning with social movements, adopting tactics of insurgent politics, and the like.

The evolving and expanding concept of cause lawyering converges at some points with the core of political lawyering, insofar as Scheingold and Sarat argued that all cause lawyers, at least in the U.S., concurred on a commitment to first generation or basic legal rights, such *as habeas corpus* (2004: 132). Both political and cause lawyers are activists. Beyond those points, however, Karpik (2007) shows there are fundamental differences in these forms of political activism.

First, political lawyers are focused on a tight core of goals potentially common to all in the legal complex, whereas cause lawyers traverse a limitless range of moral and political ideals that divide the legal complex. Political lawyering reaches historically to the collective identity of the bar and the fight for

and defense of basic legal freedoms that are integral to that identity, Cause lawyers, say Scheingold and Sarat (2004: 6), "put themselves ethically at odds with the vast majority of lawyers."

Second, political lawyering has a long history, emerging over several hundred years, whereas most cause lawyering treated in recent scholarship has flourished in recent decades, initially in developed countries, later in developing countries.

Third, since political lawyering centers on an age-old struggle between states and civil society over core freedoms, political lawyers benefit from long historical precedent that 'to defend individual rights has become self-evident.' Cause lawyers cannot take for granted that an issue will be considered a 'cause' and therefore must justify their mobilization without precedent to draw upon.

Fourth, their long history has endowed political lawyers with the resources of "a voluminous political philosophy, a repertoire of forms of action and a privileged position within the liberal state." Cause lawyers by contrast and by definition more often have limited symbolic, technical and financial resources, especially for new causes, where justifications and resources must be built from the ground up.

In summary, says Karpik (2007: 492):

> Political lawyers and cause lawyers do not share the same issues, the same history, the same references and they share only partly the repertoire of forms of action. They do not occupy the same position in the bar; they do not receive the same solidarity. Their relationships with history, with political philosophy, with the state, with civil society are different.

Lawyers in the Field of State Power

Another sustained line of writings on the politics of lawyers has concentrated on their involvement in the transformations of state power, mostly in the development contexts of Latin America and Asia. Led by Yves Dezalay and Bryant Garth (Dezalay and Garth, 2002, 2010), this approach has a continuity with political and cause lawyering of a focus on lawyers as political actors, and an apparent continuity with aspects of political lawyering insofar as lawyers strive for rule of law or restructuring the state by promoting more powerful and independent judiciaries.

Nevertheless, sharp discontinuities distinguish it from other approaches.

First, this approach is state-centered. It concerns itself with the ways power is structured by the state and exercised by the state, whether in colonial or postcolonial situations. Thus, Dezalay and Garth show how the British colonial power in India invested in law and courts whereas in Indonesia there was limited investment in law by the Dutch, who were mostly content to rely on local law for governance. This approach does not concern itself with the constitution of civil society nor the representation of marginal individuals most vulnerable to assaults on their basic legal freedoms or personal vulnerabilities.

Second, this is a politics of elite lawyers in contrast both to cause lawyers, who are more often marginal, and to political lawyers, who range from professional elites to ordinary practitioners. It does not refer to the bar as a whole, nor acknowledge any core identity common to the bar, nor attend significantly to collective action by the bar.

Third, these lawyers engaged in the field of state power mobilize not on behalf of legal, moral, or political ideals about a good society or good polity, but to maximize their own 'professional rewards' (Dezalay and Garth, 2010: 10–11). They do not act for their clients or the bar writ large but for their personal career ambitions. A kind of legal opportunism is juxtaposed with occasional 'legal idealism' (Dezalay and Garth, 2010: 60) but for the most part these lawyers are characterized as capital-accumulating utility optimizers. Underlying this orientation to lawyers' politics is an implicit motivational theory, derived from Bourdieu, on lawyers' accumulation of various sorts of capital – social capital, 'legal capital,' 'family and state capital' (Dezalay and Garth, 2010: 3). These lawyers are economistic rational actors variously 'investing' in 'capital' and 'profiting' from returns on investment.

Fourth, if earlier investments yield later rewards, then attention to lawyers' careers becomes a paramount analytic strategy. These elite lawyers are constantly mobile as their portfolio of capital expands and diversifies. At one moment they may invest in trade and commerce, in another in human rights, in another in governance. For example, U.S. lawyers who affected the field of state power in the Philippines simultaneously invested in commerce but redeemed it by also investing in foreign policy, the state and international relations (Dezalay and Garth, 2010: 52). Indeed, it is a strength of this approach that it does not draw a sharp dichotomy between those lawyers who practice in the market and those who enter politics. One may enable the other and indeed many legal elites move back and forth between them. Paradoxically this point of view repeats aspects of the theory of market control that has been subjected to sustained critique.

In contrast to the tone of the political lawyering and cause lawyering approaches, which accept that lawyers may be driven by ideals beyond self-gratification, the state power approach offers a skeptical theory of lawyers' politics. Speaking on behalf of publics or civil society, for instance, is less a means of fighting for basic legal freedoms than a source of prestige or a means of accumulating political capital (Dezalay and Garth, 2010: 26ff).

In sum, the lawyers-and-state orientation involves legal elites exercising power. These elites mobilize individual resources accumulated as an aggregation of 'capitals' during their careers. While they are situated in social structures high in the prestige hierarchies of professional work, they tend to represent or depend upon the collective mobilization of the bar as a whole. Unlike the political lawyering and cause lawyering perspectives, they evince no coherent or consistent set of values that would alter the structure of state power in one way or another. Nevertheless, their individual impacts may be substantial in the field of state power.

Conclusion

Lawyers' politics have emerged as a significant site of theory and research since the 1990s. Those politics are salient to a wide range of scholarly and pragmatic issues, ranging from comparative politics, postcolonialism, economic

development and citizenship to constitutionalism, human rights, rule of law, and state structures.

Political lawyering expresses a form of lawyers' politics, embedded in a wider legal complex, that have deep historical provenances, strong philosophical foundations, and variegated comparative expressions. Insofar as political lawyering is integral to the viability of liberal polities, arguably a necessary condition of democracy, foundational for constitutionalism and rule law, and demonstrably an incipient danger to authoritarian rulers, military dictatorships, one-party politics, and Big Man regimes, then it must be the case that the extension of empirical research in varieties of settings and the refinement of contingent theories of mobilization for basic legal freedoms is an imperative for scholarly inquiry.

See also: Civil Liberties and Human Rights; Civil Rights; Constitutionalism, Comparative; Courts and Adjudication; Globalization and Religion; Globalization: Legal Aspects; Judges; Law and Democracy; Law and Development; Law, Mobilization of; Religion and Economic Life; Religion, Sociology of; Religion: Mobilization and Power; Rights: Legal Aspects; Rule of Law (and Rechtsstaat).

Bibliography

Abbott, Andrew, 1988. The System of Professions: An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.

Abel, Richard L., Philip, S.C. Lewis (Eds.), 1988a. Lawyers in Society: The Civil Law World. University of California Press, Berkeley.

Abel, Richard L., Philip, S.C. Lewis (Eds.), 1988b. Lawyers in Society: The Common Law World. University of California Press, Berkeley.

Bell, David A., 1994. Lawyers and Citizens: The Making of a Political Elite in Old Regime France. Oxford University Press, Oxford.

Brinks, Daniel M., 2008. The Judicial Response to Police Killings in Latin America. Cambridge, New York.

Dezalay, Yves, Garth, Bryant G., 2002. The Internationalization of Palace Wars: Lawyers, Economists, and the Contest to Transform Latin American States. University of Chicago Press, Chicago.

Dezalay, Yves, Garth, Bryant G., 2010. Asian Legal Revivals: Lawyers in the Shadow of Empire. University of Chicago Press, Chicago.

Habermas, Jürgen, 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. MIT Press, Cambridge, MA.

Halliday, Terence C., Karpik, Lucien (Eds.), 1997. Lawyers and the Rise of Western Political Liberalism: Europe and North American from the Eighteenth to Twentieth Centuries. Clarendon Press, Oxford.

Halliday, Terence C., Karpik, Lucien, Feeley, Malcolm M. (Eds.), 2007a. Fighting for Political Freedom: Comparative Studies of the Legal Complex for Political Change. Hart Publishing, Oxford.

Halliday, Terence C., Karpik, Lucien, Feeley, Malcolm M., 2007b. The legal complex and struggles for political liberalism. In: Halliday, Terence C., Karpik, Lucien, Feeley, Malcolm M. (Eds.), Fighting for Political Freedom: Comparative Studies of the Legal Complex for Political Change. Hart Publishing, Oxford, pp. 1–42.

Halliday, Terence C., Karpik, Lucien, Feeley, Malcolm M. (Eds.), 2013. Fates of Political Liberalism in the British Post-colony: The Politics of the Legal Complex. Cambridge University Press, New York.

Hilbink, Lisa, 2007. Judges beyond Politics in Democracy and Dictatorship: Lessons from Chile. Cambridge University Press, New York.

Karpik, Lucien, 2000. French Lawyers: A Study in Collective Action, 1274–1994. Oxford University Press, Oxford.

Karpik, Lucien, 2007. Postscript–political lawyers. In: Halliday, Terence C., Karpik, Lucien, Feeley, Malcolm M. (Eds.), Fighting for Political Freedom: Comparative Studies of the Legal Complex for Political Change. Hart Publishing, Oxford, pp. 463–494.

Karpik, Lucien, Halliday, Terence C., 2011. The legal complex. Annual Review of Law and Social Science 7, 217–236.

Larson, Magali S., 1977. The Rise of Professionalism. University of California Press, Berkeley.

Moustafa, Tamir, 2007. The Struggle for Constitutional Power: Law, Politics, and Economic Development in Egypt. Cambridge University Press, New York.

Rajah, Jothie, 2012. Authoritarian Rule of Law: Legislation, Discourse and Legitimacy in Singapore. Cambridge University Press, New York.

Sarat, Austin, Scheingold, Stuart (Eds.), 1998. Cause Lawyering: Political Commitments and Professional Responsibilities. Oxford University Press, Oxford.

Sarat, Austin, Scheingold, Stuart (Eds.), 2001. Cause Lawyering and the State in a Global Era. Oxford University Press, Oxford.

Sarat, Austin, Scheingold, Stuart A. (Eds.), 2006. Cause Lawyers and Social Movements. Stanford University Press, Stanford, CA.

Scheingold, Stuart A., Sarat, Austin, 2004. Something to Believe in: Politics, Professionalism, and Cause Lawyering. Stanford University Press, Stanford, CA.

Political Money and Party Finance

Thomas E Mann, The Brookings Institution, Washington, DC, USA
Trevor Potter, The Campaign Legal Center, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Healthy democratic politics depends on a free flow of information among citizens, groups, candidates for public office, and political parties. The right to speak diminishes in significance without the ability to be heard, and that speech is often conditional on the availability of financial resources to pay for public communication. Yet the manner in which political money is raised and spent can undermine the legitimacy of democratic politics. Efforts to prevent concentrations of wealth from undermining political equality or introducing corruption may conflict with bedrock freedoms of speech and association. Every democracy struggles to reconcile the need for political money with the problems it begets. None have found entirely satisfactory and long-lasting solutions. The political finance tools available to policy makers – public subsidies for candidates and political parties, limits and controls on political contributions and expenditures, disclosure requirements, and regulation of campaign activity – often leave reformers unsatisfied and produce unanticipated consequences. Not all political finance measures are doomed to fail, nor are they miracle cures for the ills arising from money in politics. Democracies will continue to struggle not to solve these problems once and for all, but to contain and manage them as best as they can.

Healthy democratic politics depends on a free flow of information among citizens, groups, candidates for public office, and political parties. The right to speak diminishes in significance without the ability to be heard. In the legendary words of California Assembly Speaker Jesse Unruh, "Money is the mother's milk of politics." Yet the manner in which political money is raised and spent can undermine the legitimacy of democratic politics and effective governance. Ensuring that parties and candidates have adequate resources to compete effectively in elections is problematic, given the inevitable tension between the reality of economic inequality and the ideal of political equality. Efforts to prevent concentrations of wealth from undermining political equality or introducing corruption may conflict with bedrock freedoms of speech and association. Every democracy struggles to reconcile the need for political money with the problems it begets, ranging from classic *quid-pro-quo* corruption or extortion to the perception or reality of undue access to or influence over policy makers and government decisions. None have found entirely satisfactory and long-lasting solutions. The political finance tools available to policy makers – public subsidies for candidates and political parties, limits and controls on political contributions and expenditures, disclosure requirements, and regulation of campaign activity – often leave reformers with unfulfilled objectives and unanticipated consequences.

Institutional Context

The prominence and nature of money in politics vary across institutional settings. Party-centered electoral systems differ substantially from candidate-centered systems. In the former, most often associated with the parliamentary systems of Western Europe, political money means party finance. Most, if not all, political funds are raised and spent by political parties, and the primary concerns have to do with the adequacy of funds available for routine party activities, the work of party caucuses in parliament, and election campaigns. Direct and indirect public subsidies are often provided to the parties, including cash and free broadcasting time on state-owned networks. There is substantial variation in the extent of regulation of party finance, with disclosure of large private contributions and prohibitions on paid political advertising the most typical, and limits on the size of contributions the least typical. Party-centered systems have not been immune from escalating campaign costs and scandals, as the clean hands investigation in Italy, the 'illegal earnings scandal' involving Helmut Kohl's Christian Democrats in Germany in the late 1990s, and the more recent charges involving former French President Nicolas Sarkozy illustrate. But the problems associated with political money in party-centered systems are relatively contained and finance practices are generally viewed as in need of monitoring and repair, not radical surgery.

In candidate-centered systems, by contrast, the focus is less on political party activity and more on campaign fund-raising and spending by individual candidates. In these systems, the term of choice is campaign finance rather than party finance. The United States is, of course, the most prominent example of this class of democracies and the most atypical one in its various components. While parties once dominated the recruitment of candidates and the waging of campaigns in the United States, a host of developments, including the creation of a professional civil service, the spread of the direct primary, the rise of television and other paid media, the reform of party rules, and the provision of substantial staff resources to public officials, diminished their role. Party organizations in the United States became largely appendages of incumbent officeholders – more instruments for raising money within and beyond the constraints of the law, rather than independent political actors. The recent increase in partisan polarization in the United States has given parties and their networks of activists and organizations a more important role in campaign finance but candidates remain the dominant players. Political money in the United States and other candidate-centered systems is widely viewed as highly problematic, and much

time and energy are devoted to reporting on it and debating how best to ameliorate its harmful effects.

The demand for and supply of political money are, however, shaped by more than the extent of party or candidate-centered elections. In the United States, the sheer number of contests across the federal system, the separation of executive and legislative elections, plurality rules, constitutional protections of free speech, the high cost of reaching the large American audience, the professionalism and length of campaigns, and the permeability of the policy process all contribute to the large sums raised and spent in election campaigns. The perceived stakes of elections, the incentives and opportunities for raising and spending private resources in campaigns, and the laws and norms governing its role all condition how political money flows in democracies.

Problems Associated with Political Money

Rising campaign costs, scandals, and equity concerns in the early and mid-twentieth century precipitated waves of regulation of political finance and subsidization of party and campaign costs in democracies around the world (Paltiel, 1981). Similar concerns kept issues of the regulation and subsidy of political finance on the public agenda in more recent years. Problems arise as parties and candidates struggle to raise adequate resources to compete effectively in times, when those costs have risen sharply and groups and individuals increasingly seek to influence election outcomes and policy decisions. At one extreme lies bribery and extortion, the illegal linkage of private resources and public authority for personal gain. Short of such explicit *quid-pro-quo* exchanges lies problems, such as the dependence of parties and candidates on large financial contributors, conflicts of interests arising when politicians receive contributions from those who seek favorable action by a legislative or administrative body, and the impact on democratic legitimacy of the actuality and appearance of special access and influence by financial contributors. And then there are the effects of the money chase itself – the unending scramble for political money – on how political parties identify and recruit potential candidates, how party and public officials allocate their time, who they see and listen to, and how both affect the deliberative processes of government.

Another set of problems are associated directly with the electoral process. Rising campaign costs and inadequate political funding depress electoral competition, discourage able people from seeking public office, give a distinct advantage to incumbent officeholders and wealthy self-financed candidates, and weaken democratic accountability by muffling the voices of individuals, groups, candidates, and parties.

A final set of problems arise from the very act of trying to cope with the others. Reforms designed to alter the flow of money in elections often produce problematic unanticipated consequences. Political money is fungible, and legal constraints can divert it to less-accountable passageways. Ambitious politicians and actors naturally exploit the weaknesses and push the boundaries of the law, threatening the effectiveness and credibility, if not, the complete collapse of the regulatory regime.

Political Finance Tools

Policy makers have utilized a conventional toolbox in responding to the problems associated with political money, but the mix of tools and particular applications have varied widely among democracies.

Public Subsidies

In response to concerns about rising campaign costs and excessive dependence on financial contributors, public support of parties and campaigns has become standard among democracies (Katz, 1996). According to the International Institute for Democracy and Electoral Assistance (IDEA), 154 of 214 countries surveyed currently provide some form of public funding to political parties, including 119 countries that provide free or subsidized media. Subsidies take the form of direct cash assistance, the provision of goods and services (e.g., free broadcast time, mailings, voter brochures, and meeting space), and indirect support such as tax credits or deductions for contributions or reduced rates for political advertising. The subsidies are given to parties or candidates, and directed toward party maintenance or campaign expenditures. Direct cash subsidies to parties are typically allocated in proportion to electoral strength or representation in an elected body, while services and indirect subsidies are usually provided on a more equitable basis. In the United States, public matching funds are provided to candidates for their party's presidential nomination, who raise the requisite number of private contributions and agree to limit their spending; full public grants are provided to the major party candidates for the general election as long as they agree to raise and spend no private funds. But unlike most other democracies, the United States provides no free broadcasting time for parties or candidates and no public subsidies for candidates for Congress, although state and local candidates receive such subsidies in some states and localities in the United States.

Public subsidies are not without challenges and risks. In the United States, the voluntary presidential public financing system that began in 1976 in response to the Watergate scandal has been rendered obsolete by the rapid growth in campaign spending. For example, in 2012 the presidential nominee of each major party could have become eligible to receive a public grant of $91 million to fund their general election campaign. By way of comparison, President Barack Obama and Republican nominee Mitt Romney, who chose not to participate in the voluntary public funding system, spent more than $710 million on general election advertisements in battleground states alone. During the 2008 presidential election, Democratic Barack Obama became the first major party candidate to reject public financing for the general election. In 2012, neither Barack Obama nor Mitt Romney accepted public financing for either the primary or general elections. Public financing systems can quickly become outdated if not adjusted to account for changing demands for campaign funds.

Even when adequately funded, public subsidies can reinforce the dominance of established parties, increase the potential for control by the state, and weaken the connection of parties with their grassroots. Public financing, conditioned on very low qualifying thresholds, can also encourage fringe and

nonserious candidates and parties. To the extent that public money is used to subsidize speech or candidates that citizens find disagreeable or offensive, the subsidized speech can itself erode public support for the public financing system. Full public financing systems also generate intense pressure for private resources to find alternate routes to influence elections and policy. To the extent such avenues are available, the ability of the public financing system to alleviate the problems associated with political money is undermined.

Contribution Limits

While public subsidies are the norm in political financing, limits on contributions are less common. Many democracies place no restrictions on the source or amount of contributions to parties or candidates. Among those that do, prohibitions on particular sources (e.g., civil servants, foreign donors, corporations, and unions) are most common. According to IDEA, 38 countries currently prohibit corporate donations to candidates or political parties, 41 countries prohibit union donations to political parties, 40 countries prohibit union donations to candidates, 113 countries prohibit foreign interests from donating to political parties, and 85 countries prohibit foreign interests from donating to candidates.

In the United States, corporations and unions have been precluded for many decades from making political contributions to federal candidates or political parties from their treasuries (with a variety of regimes at the state level), but they may spend their funds to form and administer voluntary political action committees that may raise funds from certain corporate or union personnel and use those funds to make contributions to parties and candidates. But, in recent years, even the long-standing prohibitions on contributions from corporate and union treasuries have been rendered largely ineffectual due to deregulatory decisions of the Supreme Court of the United States allowing unlimited corporate direct spending on election messages, which has led to the emergence of the so-called 'outside groups' that can receive unlimited, and in many cases undisclosed, corporate and union contributions and spend those funds on election advertising so long as they do not directly coordinate their efforts with political parties or candidates.

Some countries have placed statutory limits on the size of contributions from permissible sources, including Brazil, France, Israel, Japan, the Russia Federation, and the United States. According to IDEA, 68 countries currently limit the amount that a donor can contribute to a political party, and 54 countries currently limit the amount that a donor can contribute to a candidate. The contribution limits in the United States are notably low, and even though many of the limits are indexed for inflation, the rapid rise in campaign costs has far outpaced these inflation-related adjustments. This contributes to the extensive fund-raising demands on politicians and creates increased pressure for private funding to flow through other channels, such as 'outside groups,' which are characteristic of problems associated more generally with contribution limits.

Expenditure Limits

A number of democracies place limits on spending by parties and candidates. According to IDEA, 75 of 172 countries surveyed limit the amount candidates can spend, and 51 countries limit the amount that political parties can spend. However, these limits often apply only to a brief period of time (e.g., the formal campaign in parliamentary systems), to candidates but not parties (or vice-versa), or they exempt entire categories of political activity that influence election outcomes. For example, in Japan, candidates for the Diet face limits on campaign expenditures but these are largely irrelevant because they apply only to the 12-day official campaign period. For well over a century, Britain limited the expenses individual candidates for Parliament can incur in their constituencies (and prohibited outside groups from spending on their behalf), but did not regulate political party expenditures at the national level, where most of the action is. However, a law adopted in 2000 limits national campaign expenditures by political parties; it also limits the amount of spending by outside organizations and individuals for political activity during a general election.

Mandatory spending limits adopted in the United States in 1974 were declared unconstitutional by the Supreme Court of the United States in 1976, while voluntary caps tied to the receipt of public funds in presidential elections remain permissible. But, even if presidential candidates were to opt in to the voluntary public financing system and thereby become subject to the spending caps, presidential candidates, political parties, and interest groups would likely work around these limits by conducting political advertisements through outside groups that the courts have ruled may spend unlimited amounts on election advertisements so long as they do not directly coordinate their advertisements with candidates or political parties.

Public Disclosure

Transparency in the flow of political money is embraced by many democracies as an efficient and effective deterrent to corruption or influence peddling, and as a means of providing citizens information they can use to evaluate candidates and to hold elected officials accountable. In 2004, the United Nations Convention against Corruption urged nations to take steps to "enhance transparency in the funding of candidates for elected public office and, where applicable, the funding of political parties." At its core, an effective public disclosure system requires political parties and candidates to disclose assets, income (including contributions), and expenditures in a timely fashion to an independent agency.

According to the IDEA, 125 of 172 surveyed countries currently require political parties to regularly report on their finances, and 103 of 171 surveyed countries currently require candidates to report on their campaign finances. While disclosure of political money is gaining in acceptance in the democratic world – Britain, for example, broke with its long-standing practice in 2000 by requiring that large donations to and expenditures by political parties be publicly disclosed – some Scandinavian countries such as Sweden and Norway continue to consider disclosure requirements applied to election campaigns as an unacceptable violation of privacy.

Even among the countries that have adopted disclosure requirements, however, the comprehensiveness and effectiveness of disclosure regimes vary widely. At one extreme are

systems requiring only the filing of a single report on party expenditures well after the election. At the other are systems, like the United States, with requirements for frequent reporting of political contributions and expenditures (as small as those totaling $200) by candidates, political parties, and politically active groups, and making this information publicly available on a timely basis. But, even in the United States, where there has existed an historic political consensus in favor of disclosure requirements as the cornerstone of campaign finance regulation, a significant amount of electioneering occurs outside the boundaries of the disclosure requirements. For example, during the 2012 elections it is estimated that almost $600 million spent to influence federal elections in the United States was not fully disclosed. Achieving transparency can be complicated by a lack of transparency in a country's broader economy. For example, despite India's relatively comprehensive disclosure requirements applicable to candidates and political parties, a significant amount of election-related spending occurs in India's so-called 'shadow economy' through off-the-books cash transactions. Election authorities have attempted to cut off this 'black money' market by enacting expenditure monitoring measures and even seizing stores of rupees from political parties and candidates, but significant challenges remain. This demonstrates one of the key challenges of public disclosure requirements as regulatory tool: transparency is only as valuable as it is comprehensive. As long as those that desire to have their contributions or expenditures go undisclosed are permitted to achieve that goal – be it due to insufficiently comprehensive disclosure requirements or inadequate enforcement mechanisms – the value of the disclosure system as a whole is compromised.

Regulation of Campaign Activity

One way of reducing the demand for political money is to prohibit or limit certain costly campaign activities. A surprisingly large number of countries ban paid political advertising on television, although, most couple the ban with the provision of blocks of free TV time to the parties. Such prohibitions are unthinkable in the United States, where the free speech guarantees of the First Amendment to the United States Constitution, as historically interpreted by the United States Supreme Court, preclude any significant restraint on political communication. Although the European Court of Human Rights recently upheld the United Kingdom's far-reaching ban on political advertising against challenge brought by an animal welfare group, many other countries have moved to assert that a legal protection of freedom of expression limits the extent to which campaign activity can be subjected to regulation. For example, the Australian High Court declared a 1991 law that prohibited the broadcasting of political advertising during election periods unconstitutional, on grounds that it infringed the implied right to freedom of political communication.

Enforcement and Effectiveness

The success of any measures designed to redirect the flow of or shine light on political money depends upon the ingenuity, comprehensiveness, and timeliness of the particular mix of regulatory and subsidy measures and on the mechanisms for enforcement. Short of blatant violations of criminal statutes (e.g., bribery or extortion), departures by parties and candidates from the letter and spirit of campaign finance laws often go unpunished. Even when violations are pursued, the punishment is often not levied until long after the election during which the violation occurred takes place. Politicians are reluctant to invest the requisite authority in public agencies to enforce political finance laws, a phenomenon tellingly illustrated by the structure and operation of the U.S. Federal Election Commission (FEC). The FEC, whose six members effectively include three from each of the two major political parties, operates under very tight statutory and budgetary reins from Congress and often deadlocks on regulatory and enforcement decisions, as a majority vote of the members is required for the Commission to take action, leading to split party-line votes. Putting in place a reasonable and workable set of regulations, appropriate criminal and civil sanctions, and administrative and judicial mechanisms has proven a daunting task in many democracies.

For all of these reasons, it is not surprising that political finance reforms often fall short of reducing corruption, limiting the influence of monied interests, slowing the increase in campaign expenditures, and increasing electoral competition. Other institutions and policies are, perhaps, more effective levers for combating corruption and enhancing democratic accountability. These include broad economic policies that limit the discretionary authority of public officials; institutions and practices nurturing the rule of law; an independent judiciary; adequate pay and strong ethical requirements and accompanying enforcement for politicians and civil servants; transparent budget and financial control systems; and party systems supporting stable patterns of competition, accountability, and alternation of power (Kaufmann, 1997; Pope, 1996).

Not all political finance measures are doomed to fail, nor are they miracle cures for the ills arising from money in politics. But, since neither the abolition of all private funds from politics nor the complete deregulation of political finance are feasible or sustainable, democracies will continue to struggle not to solve these problems once and for all, but to contain and manage them as best as they can.

Acknowledgments

Bryson B. Morgan, Investigative Counsel, United States House of Representatives Office of Congressional Ethics. The views expressed herein do not represent the views of the Office of Congressional Ethics.

See also: Campaigning: Political; Congress: United States; Mass Media and Politics; Political Parties, History of.

Bibliography

Alexander, H.E., Shiratori, R. (Eds.), 1994. Comparative Political Finance among the Democracies. Westview Press, Boulder.

The Campaign Legal Center. Generally, available at: http://www.campaignlegalcenter.org/.

Corrado, A., Mann, T.E., Ortiz, D.R., Potter, T., 2001. The New Campaign Finance Sourcebook. Brookings Institution Press, Washington, DC.

Gunlicks, A.B. (Ed.), 1993. Campaign and Party Finance in North America and Western Europe. Westview Press, Boulder.

Katz, R.S., 1996. Party organizations and finance. In: Leduc, L., Niemi, R.G., Norris, P. (Eds.), Comparing Democracies: Elections and Voting in Global Perspective. SAGE Publications, Thousand Oaks, pp. 107–132.

Kaufmann, D., 1997. Corruption: the facts. Foreign Policy, 114–131.

Paltiel, K.Z., 1981. Campaign finance: contrasting practices and reforms. In: Butler, D., Penniman, H.R., Ranney, A. (Eds.), Democracy at the Polls. American Enterprise Institute, Washington, DC, pp. 138–173.

Pope, J. (Ed.), 1996. The Transparency International Source Book. Transparency International, Berlin.

Political Finance Database, International Institute for Democracy and Electoral Assistance (IDEA). Available at: http://www.idea.int/political-finance/?ul=es (updated March, 2013).

Report for Congress, 1997. Campaign Financing of National Elections in Selected Countries. Law Library, Library of Congress, Washington, DC.

Scarrow, S., 2007. Political finance in comparative perspective. Annual Review of Political Science 10, 193–210.

Sorauf, F.J., 1994. Inside Campaign Finance: Myths and Realities. Yale University Press, New Haven.

Political Parties, History of

Estevao C de Rezende Martins, University of Brasilia, Brasilia, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political parties are considered as formal civil organizations that are active in public life, in which they represent specific class, professional, or ideological interests, within the framework of the modern constitutional State, and whose goal is to run for election and, upon winning, to take power as the government. There are three nuances of the term 'party' from the historical point of view: rival groups, orders, states, or dynasties; theological or ideological critics; and modern formal social organization. Political parties – in the contemporary sense of the term – became common in the nineteenth century, alongside the development of representative institutions and the suffrage, becoming gradually universal. Parties began to elicit attention from philosophers and the interest of modern historians from the time of the English revolutions of the seventeenth century. A party emerges when a political current within society finds expression through a specific (national or regional) organization. The typical action of parties in modern societies emerged out of the political consciousness that defined and established the notion of a universal citizenship.

The definition of a political party requires a historical perspective. The idea of a party is multifaceted and developed over time, in the prevailing political, social, economic, and (especially) legal realities in which it is formed. In current language – and in the specialized terminology of the social sciences – the term 'party' has been in use for longer than the term 'social class.' Etymologically, the term derives from the verb *party*, which, in archaic French, meant 'partake,' or 'share.' The present article regards political parties as formal civil organizations that are active in public life, in which they represent specific class, professional, or ideological interests, within the framework of the modern constitutional State, and whose goal is to run for election and, upon winning, to take power as the government.

Approaches and Definitions

In the past, the expression 'political party' was also used to describe other political phenomena. Throughout the history of societies and of their political bodies, since antiquity, the expression 'be party to' affairs of public life has been considered a common mode of human behavior. The translators of Aristotle use the term 'party' to designate opposing social groups in Athens. Specialists on ancient Rome have used the concepts of 'Plebeian Party' and 'Patrician Party,' and the famous *De viris illustribus Romae* relates to the party disputes that dilacerated the Roman Republic. The supporters of the Crown and the defenders of the rights of the nobility were also parties. Critics of the theological hegemony of the church and of the clergy were parties within the context of the political order of Christian Europe. Noble families in opposition to the Crown were parties in the marking; religious and political reformers were parties. These are broad meanings of the term that are frequently to be found in the literature. Also under the broad heading of 'parties' can be categorized all manner of social organizations that seek political ends, even if they never stand for (or win) elections, as is the case for the miniparties or revolutionary organizations that have no constituency among the electorate and that do not accept the validity of the electoral system, or even the dominant groups in totalitarian states. Thus, the expression 'political parties' is also often applied to interest groups; propagators of ideologies, doctrines, and political positions; political model makers; and mediators between citizens and government, including a vast multiplicity of bodies that participate in public life.

There are three nuances of the term 'party' from the historical point of view: rival groups, orders, states, or dynasties; theological or ideological critics; and modern formal social organization. The third interpretation has, in part, inherited the historical characteristics of earlier forms, making it possible to make such distinctions as conservative and liberal parties; religious and lay groups; corporative interest and those of the masses; revolutionaries and reformers; left and right; center and radical; and progressive and reactionary. Such distinctions are not exclusive, are directly proportional to the degree of institutional solidity and of integration that the parties enjoy in political society, and always relate to the ideas that separate – or unite – individuals or groups. Such multiplicity is evident in the classification of parties, according to the ideological density of their manifestos, the form of internal structure, competitive or noncompetitive organization, and of party systems (single-party, two-party, or multiparty; unipolar, bipolar, or multipolar) concerning the political and legal frame they are put in. Unlike political science, historical research does not treat these issues in isolation, but rather, seeks to reconstruct how all such aspects relate to each other.

Political parties – in the contemporary sense of the term – became common in the nineteenth century, alongside the development of representative institutions and the suffrage, becoming gradually universal, especially in Western Europe and the United States. In the aftermath of the Congress of Vienna (1814–15), political parties had not yet been formally recognized as civil associations. Any association, assembly, or grouping united by specific goal or political interest was regarded as a party. Conversely, the pursuit of private interests and the application of abusive pressures were regarded as a constraint upon political life, as the goal of winning elections

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62094-9

and of influencing government and public policy became the established criteria for identifying a group of citizens as a political party. In the United States, by the 1830s, parties as such were fully active. In the Western European states, it was in the final decades of the nineteenth century and the first decade of the twentieth century that this development took place. Modern political parties are full players in the public life of democratic societies. In effect, only constitutional states that formed and consolidated over the past 200 years have the institutional framework necessary to ensure regular political participation on the part of their citizens. Such a framework determines the legal principles for party actions, articulates position of the parties in relation to the system of government and the electoral system, and places them within the framework of civil law. Their social origin and ideological orientation notwithstanding, it was the placing of parties within a legal framework that created the party system – understood both as a juridical and political frame of the participation of citizens in political life.

Parties began to elicit attention from philosophers and the interest of modern historians from the time of the English revolutions of the seventeenth century, on through the revolutionary turmoil of 1848, and up until the death of Marx in 1883. They were observed by Bolingbroke and Madison, Hume, Burke, Benjamin Constant and Blüntschli, and the German social democratic thinkers. At the turn of twentieth century, party political phenomena began to be the object of serious scientific methodological analysis. The earliest works in this field were published after the consolidation of parliamentary forms of democratic government, which had been in place since the mid-nineteenth century. The pioneering works of Ostrogorski (1903) and Michels (1911) examined the functioning of parties in the late nineteenth and early twentieth centuries in the United Kingdom, France, and the United States. These authors sought to put forward a general theory, beginning with highly particular experiences and specific historical developments, such as the establishment of federative republicanism of the United States or the burgeoning power of Parliament in the Government of the United Kingdom. There are many descriptive and historical studies on political parties in specific countries and also comparative studies that seek to define possible models. Research trends generally give greater emphasis to the western model of representative democracy, in which formalized parliamentary, electoral, and political party systems play an essential role in public life and the practices of government (Epstein, 1993; Budge and Keman, 1990; Fenske, 1994).

With Max Weber, who couches his remarks on parties from a sociological standpoint, modern political parties came to be understood as a product of democracy and of universal suffrage. Jean Charlot (1971), according to LaPalombara and Weiner (1966), provides a systematically historical view of parties. In order for a group to be regarded as a political party, it should fulfill four requisites. It must (1) possess a visible and stable organization, for the articulation of its local, regional, and national components; (2) ensure continuity of organization, independently of its current managers; (3) have the explicit goal of achieving and holding decision-making power, either alone or as part of a coalition with others, at all levels; and (4) have organizational efficiency, with a view to capturing the vote and harnessing popular support. It is, however, a work by Duverger (1951) that systematizes and compiles knowledge on political parties. Two of his research elements remain valid: one puts forward a theory of the origin and of the proliferation of political parties, from an institutional perspective, relating to the party phenomenon and the dynamics of institutions, whereas the other resides in a typology of parties based on the nature of their organization. After World War II, the discredit and mistrust with which people regarded the traditional political parties of Europe – owing to their failure to avoid the emergence of fascism and totalitarian dictatorships in Italy, Germany, Spain, and Russia – hampered the reintroduction of democracy, and parties were left fragmented and disunited. The understanding came up that healthy democratic life requires political parties within a system of government that establishes clear rule for the formation of governments to exercise the will of the majority, solely on the basis of the judgment expressed through the popular vote. The political instability of the Fourth Republic in France, and of the government of Italy, for example, provided Duverger with a rich material on which to base the most ambitious article written on the theme of political parties in the mid-twentieth century. Duverger regards the parties of the masses as the inevitable product of universal suffrage and regards many of the residual parties as being archaic, especially in postwar France. His view is supported by the movements for ideological concentration in European public life at the Cold War and has as its inspiration the two-party systems – which he regarded as being successful – in the United States, United Kingdom, and West Germany.

Although there is still a lack of broad-based comparative studies, the contributions of Robert Dahl (1966), Lipset and Rokkan (1967), and Epstein (1993) take steps in this direction. The ideological nature of parties is given an innovative treatment by Stein Rokkan, comparable with the contribution of Duverger on party organization. The analysis of party systems was advanced by Sartori (1976) and Blondel (1978), whereas electoral sociology – beginning with the pioneering work of André Siegfried on electoral statistics and geography (1913, 1927), by Richard Rose (1984), David Butler (1981), and Alain Lancelot (1988). More recently, Daalder and Mair (1983), Pierre Avril (1990), Klaus von Beyme (1985), Daniel-Louis Seiler (1986), Mair (1990), Mehra et al. (2003), and Tremblay (2008) have sought to formulate studies from an institutional and systemic viewpoint.

Stages of Development (Phases) Change Over Time

The first expressions of party life began with the defense of specific interests, although in different contexts and times, such as those of the aristocracy against the power of kings and governments, or of the working class against the political, economic, and social dominance of ruling groups or elites. Such defense was not passive. It was carried out by means of consistent and prolonged actions within the representative institutions, such as the estates, consultative assemblies, and parliaments. It also comprised the revolutionary movements in which the respect for rights – both effectively held and merely claimed – was demanded of the political authorities, the governing establishment or the despot.

A party emerges when a political current within society finds expression through a specific (national or regional) organization. The first steps toward the organization of a political group can be observed through the 'parties' into which England polarized in the mid-seventeenth century. Although these parties were still characterized by the discipline and discourse of religion, the political effect of the change from one system of government to another imposed by those in opposition to the power of the throne ushered in a new form of the exercise of power within society. It goes without saying that the great revolutions of the seventeenth and eighteenth centuries – in England, France, and the United States – also gave rise to factions that the majority of historians would call 'parties.'

The typical action of parties in modern societies emerged out of the political consciousness that defined and established the notion of a universal citizenship, evolves in the new space for political protagonists (both for and against the State), the emergence of organized political parties during the nineteenth century and investiture of political parties as being the only protagonists capable of adopting practices, expressing ideas, and taking power and exercising public office within the State. The parties of the bourgeois in the early nineteenth century protagonized opposition between the aristocracy and the monarchy; a similar process can be observed among the emergent revolutionary or proletarian parties of the end of the same century. In revolutionary France, there were embryonic experiences of political party organization: the Girondins, Montagnards, and the Jacobins. The greater part of these 'parties' consisted of small groups, without formal organizational structures. It can thus be stated that no formal parties existed prior to 1820. Three phases in the development of party systems can be identified in the nineteenth century: the earliest phase between 1820 and 1830, an intermediate phase in the 1860s, and a final phase around 1880.

In the 1820s and the 1830s, the British Tories and Whigs emerged and consolidated long-held positions. At the same time, in many countries of Western Europe, and notably in France, there was a movement toward a more constitutional state, expanding suffrage and consolidating the system through application of the principle of majority rule in the formation of governments. It was in this phase that the earliest political parties emerged in Latin America, as in the case of the *Partido Liberal* in Colombia (1815) and the *Colorado* and *Nacional* parties in Uruguay (1836), which arose as an immediate consequence of the wars of independence against Spain and of the ideological influence of the United States. In the intermediate phase, there was a clear expansion of political parties: the Conservative (1840) and the Liberal (1860) parties in Brazil; the Liberal Party in Belgium (1846), the Christian – Democrat and Radical – Democrat parties in Switzerland (1848), the Democratic (1848) and Republican (1856) parties in the United States, the Liberal Party in Italy (1848), the *Partido Radical* in Chile (1863), and the Conservative (1854) and Liberal (1867) parties in Canada.

In Latin America, where most countries had thrown off colonial rule in the first decades of the nineteenth century, only gradually did political parties begin to gain importance. As of 1850, the national aspirations of countries of South America came under the spreading influence of North American two-party system. The first expression of political life was the affirmation of a national identity. The main banner of the 'political parties' up until the mid-1800s was nationalism, defended as the means whereby these new states could affirm their independence in face of their former colonial masters. The 1880s witnessed the emergence of 'classical' political parties, generally in twos, and of either conservative or liberal persuasion – but always nationalist – in the former Spanish colonies. It should be noted that the emergence of political parties did not necessarily mean that they assumed a decisive role in political life. Under the Brazilian Empire (1822–89), a two-party system existed as a means for the rural aristocracy to maintain its hold on power and, under the Republic proclaimed in 1889, this oligarchic system remained essentially unchanged at least until 2003 as the first worker became an elected president.

In the mid-nineteenth century, growing social tensions and bitter social criticism led people with a political outlook to express their social and economic (and occasionally revolutionary) demands through formal political action, by means of the creation of political parties. Thus began the 'socialist' or 'social democrat' phase, typified by formation of the Danish Social Democratic Party (1871), the Spanish Socialist Party (1879), the German Social Democratic Party (1863, 1869, 1875), the Swiss Socialist Party (1880), the Belgian Workers Party (1885), the Indian National Congress (1885), the Austrian Social Democratic Party (1889), the Swedish Social Democratic Party (1889), the Italian Socialist Party (1892), and many others.

World War I forced a pause for reflection and redefinition of the party systems in Europe. The victory (1917) of the Bolshevik Party (1903), which upon seizing power set up a revolutionary and internationalist government in Russia, and, simultaneously, the conservative bias of the Treaty of Versailles (1919) mark the polarization of two separate and irreconcilable ideological currents. On one side, moderate, conservative democratic, and liberal parties found themselves up against hard-line right- or left-wing revolutionary opponents of a nationalist or internationalist persuasion, with scant respect for suffrage, democracy, or tolerance. The twentieth century opened with the founding of a series of communist parties (Russia, 1903; Austria, 1918; Greece, 1918; United States, 1919; United Kingdom, 1920; Spain, 1920; Italy, 1921). These new protagonists on the political scene achieved varying degrees of importance in different places and their role was to become more prominent during the cold War. Communist ideology and the proposals of revolutionary Marxist analyses were to assume the weight of a political mortgage at two moments: in Europe in the 1920s and the 1930s, and throughout the world from 1945 to 1990. Social demands and economic crises deeply disturbed the political environment. The emergence of totalitarian and fascist dictatorships was to undermine political party structures throughout Europe until 1945. Nevertheless, these dictatorships organized political life by means of an imposed adherence to a single party. Only formally such a party was maintained as an element of legitimacy in the exercise of state power, despite the fact that authoritarian regimes could dispense this.

The success of the Western model of parliamentary democracy based upon a three-legged political system (representative and participative democracy, the electoral system, and

the political party system) is thus a recent phenomenon. During the Cold War, the ideals of representative democracy as being the end product of the modern liberal constitutionalism, with roots in the 1790s, appeared to be heritage limited principally to Western Europe and the United States. A few other countries, notably former members of the British Empire – such as Canada, Australia, New Zealand, and India – have practiced representative democracy since achieving their independence. In countries such as Brazil and India, where democratic pluralism exists under specific historical and cultural conditions, many political activities take place outside the formal political party structure.

Between the world wars, growing social awareness of rights and of the need to manage an increasingly literate populace led to structural modifications among the political parties. Economic crises in Central Europe were accompanied by increasing mistrust of parliamentary institutions and the emergence of manipulative right-wing political parties. Fascism and communism radically changed the way in which political parties were regarded. Into the parties of the masses flowed all the social dissatisfaction and frustration, making them easy prey for manipulation by dictators of every stripe. In Europe, as also in other regions, the trend toward mass parties, which in principal appeared to be a sing of the triumph of democracy, in the end generally resulted in quite the opposite effect as authoritarian rulers assumed power. In many regions, however, expectations with respect to the freedom to participate in party activities were enhanced, especially after the experience of World War II.

The triumph of the Western model of liberal democracy with parliamentary representation was one of the consequences of the new international order that emerged after 1945, except in regions under the yoke of communism (Eastern Europe, China, and Cuba). Communist regimes, however, do not dispense with mass parties or even with formal parliaments, although the role of these institutions is merely figurative. After World War II, a new phase of political life began, with the restoration of the values of liberal democracy, universal suffrage, civil liberties, and parliamentary representation. The Western tradition of recording political events – at least since the Code Napoleon (1804) and the British Reform Act (1832) – has provided a framework for the organization of political life in such a manner that the formal and legal structuring of parties has advanced in parallel to the campaign for constitutional government. This has caused parties to organize and represent themselves as both defenders of an overall social vision and as pursuers of political goals. Thus, parties assume the posture of political businesses whereby they compete within a legal framework to advance the political convictions, opinions, and beliefs – and to defend the preferences and class interests – of their constituents. Diverse and intersecting interests find expression among groups, clubs, associations, and, above all, political parties. The social economic and ideological cutlines present in the social structures were key factors throughout the nineteenth century. Conservatives and Liberals, labor parties and revolutionaries, opposed each other in Europe and also in Latin America. In the United Kingdom, the Conservatives and the Liberals succeeded the Tories and the Whigs and were in turn superseded by the Conservative and Labour parties of the twentieth century. In the United States, in the period following the Civil War, the Democrats and the Republicans definitively consolidated their hegemony over the political scene (Katz and Mair, 1992).

The history of political parties clearly demonstrates that the political opportunities in the social competition for power in the State have an effect upon the parties that, for their part, have an effect upon the structure of political opportunity, i.e., parties and institutions mutually affect each other. In democracies that work regularly, disciplined voters arise, as in the European Union, for example. These voters do respond to party activities, although they are increasingly critical of the means whereby parties pursue their ends. 'Belonging' to a party tends to be more a sociological identification with a political tendency, rather than involving formal party affiliation. This suggests a salutary clarification of the boundaries between parties and their environment. The process whereby a party is identified with its leader is an interesting historical phenomenon in the political organization of societies. For example, the Braybrooke Manuscript (1788) refers to the parties 'of the Crown,' 'faithful to Mr. Pitt,' and 'of Mr. Fox.' Another interesting and more contemporary example is of the emergence of a 'Gaullist' party in France, after 1946. The tradition of having a strong personal leader continues very much alive in Latin America, even in the late twentieth century. This tradition has influenced the social practices of parties in various ways. In Mexico, the Institutional Revolutionary Party – PRI (1911) – dominated the Mexican political scene single-handedly until 2000.

With the advance of democracy and of democratic negotiating practices, conducted with a view to making government of contemporary states feasible, all parties (including the new parties that emerged out of the former communist parties of Eastern Europe, and also certain parties in Western Europe) have had to become connected. The difficulty of winning elections based solely upon their different platforms and making their will prevail has, since the 1970s, obliged parties to emphasize their capacity to draw together to form and support coalitions. Finally, all modern parties are obliged to adopt businesslike organizational structures or run the risk of failing to fulfill the demands of modern life, in terms of communication, social penetration, and funding.

Evidently, the aim of assuming the mantle of government remains the principal motivation of political parties. The historical trajectory of activities undertaken by political parties is closely associated with an awareness of the structure of political opportunities within a given political system. Revolutionary organizations that reject existing institutions, social clubs, debating societies, pressure, or advocacy groups with independent objectives can be regarded as entities that discreetly play significant political roles without being formally political parties. Formal political parties must, as such, compete to control a political process whose agenda they do not establish by themselves, but which cannot function without their participation. This experience is deeply ingrained in the political parties of the late twentieth century.

Unanswered Questions and Issues for Debate

Some important questions remain unanswered. For example, the emergence of political participation – as an alternative to

traditional in liberal democracies – is often treated both as an essential element for modernization and as its critical consequence. To assume the mantle of government and to perpetuate the values of parliamentary democracy, strong parliamentary institutions are necessary if modern political parties are to perform their roles. This is true even in countries such as the United States – and many others in Africa and Latin America – in which the presidency is a strong and 'personalized' institution. This applies even in cases where presidents are elected by direct vote and are endowed with special powers, as occurs in France and Portugal.

Although much is known with respect to the details of the emergence of specific party systems, the capacity to generalize as to the reasons why different types of parties emerge in the ways and forms in which they have done so remains incipient. There are few works that examine or seek to explain the vast variety of different contexts, and comparative studies among regions with different levels of development are rare. In the 1980s, the political development of multinational regional institutions – which initially had a solely economic focus – such as the European Economic Community, Association of Southeast Asian Nations (ASEAN), and (in the 1990s) MercadoComum do Sul (MERCOSUL), has unveiled a whole new field for research into the political life of such institutions, their relations with member states, and the entire theme of supranational integration. From a historical research perspective, only now are the first comparative studies of the political parties being carried out. This applies not only to research on existing parties in specific countries or groups of countries (as is the case in the European Union) but also with respect to an examination of the historical roots of contemporary political parties.

Not only parties but also party systems can be analyzed from the institutional standpoint. Both the parties and the party systems of the 1970s were little different from those of the 1920s. With the exception of the ecological movement, the majority of the party groupings, i.e., similar parties, can be presented according to their similarities. Practically all the present parties – of the left, of the right, and of the center; liberal; and revolutionary parties – consolidated themselves in the 1920s, and some much earlier. Political parties are present at the heart of all activities of the state in liberal democracies. However, even under communist regimes the role of the political party is performed in a well-established manner. In the 1990s, however, it could not safely be predicted that political parties would face a future with development and prestige, similar to that they enjoyed in the recent past. At the present time, a new type of political culture is emerging. Political awareness and politics – as subject for discussion – open up a new avenue of research into general participation in political life on the threshold of the twenty-first century. Culture may help in understanding as to why certain social cleavages are successfully transposed by a stable party system, whereas others are not. Political culture and the manner in which it is created are responsible for some significant differences in the traditional organization of states, or of bodies associated with them. Political culture includes not only ideals, such as values, beliefs, and attitudes, but also political practices and societal behavior.

There are certain themes in the study of political parties that still require further research from a historical point of view. At the present time, issues on gender representation, minorities' rights through political representation, religious militancy in politics, and reevaluation of subnational or national claims within historically centralized states won on importance and political relevance. Therefore, the question of the balance in the use of political power and the politics of containment became recurrent themes in political science and in history as well. One of the most important contentions of political activity is the institutional framework of the state. Contemporary political and historical research has made clear the importance of institutional factors in political life. Aside from this new institutional focus, the study of political parties has demonstrated a renewed appreciation of history, of events, of the rule, and of political regimes as the active forces of political life. It is perhaps commonplace to state that 'history matters.' The 'frozen cleavages' of Lipset and Rokkan, the development models of communism, fascism and the parliamentary democracy of Moore (1968a), and the theories of 'critical realignment' of Burnham (1970) are good examples of historical reviews and models. The legacy of history is thus one of the limitations to which institutionalism is subject.

Socioeconomic constraints are generally accepted without going into the mythical and overworked concepts that have been served up since the time of Marx. Rapid changes in international finances and economic scenarios have produced intriguing works. The exercise of social power is also constrained by cognitive means, in the use of pure reason (and more specifically, of the practical one). The knowledge society and information society will have to take into consideration nonrational features of political life, and the social and ideological factors that influence the belief systems of the masses (Meny, 1989). The final two decades of the twentieth century witnessed a swift retreat in the positivist distinction between fact and value. Specialists recognized that political agents were also ethical players (Rawls, 1993), who have their own values and act in accordance with them, and that occasionally they may be persuaded (by political philosophers, perhaps) to adopt other, better values. Any attempt to divorce facts and values from the mental processes and from the underlying political dynamics would be doomed to failure. Similarly, political science increasingly seeks to develop studies that take into account the complexity of the structures, processes, and results that it investigates. In order to do this, one must have a theoretical framework within which the history of the processes and the prevailing political philosophy must occupy a space that formally would have drawn opposition from previous generations.

See also: Conservatism: Historical Aspects; Democracy, History of; Interest Groups, History of; Liberalism: Historical Aspects; Parliaments, History of; Party Systems; Political History: History of Politics; Socialism: Historical Aspects.

Bibliography

Avril, P., 1990. Essais sur les Partis Politiques. Payot, Paris.

Blondel, J., 1978. Political Parties: A Genuine Case for Discontent? Wildwood House, London.
Budge, I., Keman, K., 1990. Parties and Democracy. Coalition, Formation and Government Functioning in Twenty States. Oxford University Press, Oxford, UK.
Butler, D., 1981. Democracy at the Polls. AEI, Washington, DC.
Burnham, W.D., 1970. Critical Elections and the Mainsprings of American Politics. W. W. Norton and Co, New York.
Charlot, J., 1971. Les Partis Politiques. Payot, Paris.
Daalder, H., Mair, P., 1983. Western European Party Systems: Continuity & Change. Sage Publications, London and Beverly Hills.
Dhal, R., 1966. Political Oppositions in Western Democracies. Yale University Press, New Haven, NJ.
Duverger, M., 1951. Les Partis Politiques. Colin, Paris.
Epstein, L.D., 1993. Political Parties in Western Democracies. Transaction Publishers, New Brunswick, NJ.
Fenske, H., 1994. Deutsche Parteiengeschichte. Schöningh, Munich, Germany.
Galli, G., 1990. Storia dei Partiti Politici Europei. Rizzoli, Milan, Italy.
Goodin, R.E., Klingemann, H.-D. (Eds.), 1966. A New Handbook of Political Science. Oxford University Press, Oxford, UK.
Graham, B.D., 1993. Representation and Party Polities. A Comparative Perspective. Blackwell, Cambridge, UK.
Huntington, S.P., 1968. Political Order in Changing Societies. Yale University Press, New Haven, NJ.
Kaase, M., Newton, K. (Eds.), 1995. Beliefs in Government. Oxford University Press, Oxford, UK.
Katz, R.S., Mair, P. (Eds.), 1992. Party Organizations. A Data Handbook on Party Organizations in Western Democracies. 1960–90. Sage Publications, London.
Kitschelt, H., Kirk, A.H., Luna, J.P., Rosas, G., 2010. Latin American Party Systems. Cambridge University Press, Cambridge, UK.
Lancelot, A., 1988. Les élections sous la Vème République. PUF, Paris.
LaPalombara, J., Weiner, M. (Eds.), 1966. Political Parties and Political Development. Princeton University Press, Princeton, NJ.
Lawson, K. (Ed.), 1994. How Political Parties Work: Perspectives from Within. Praeger Publishing, Westport, CT.
Lipset, S.M., Rokkan, S., 1967. Cleavage structures, party systems and voter alignments: an introduction. In: Lipset, S.M., Rokkan, S. (Eds.), Party Systems and Voter Alignments: Cross-national Perspectives. Free Press, New York.
Lloyd, C. (Ed.), 1983. Social Theory and Political Practice. Wolfson College Lectures. Clarendon Press, Oxford, UK.
Lloyd, C., 1993. The Structure of History (Studies in Social Discontinuity). Blackwell, London.
Lovenduski, J., Pippa, N. (Eds.), 1993. Gender and Party Polities. Sage Publications, London.
Mainwaring, S., 1999. Rethinking Party Systems in the Third Wave of Democratization: The Case of Brazil. Stanford University Press, Stanford, CA.
Mair, P. (Ed.), 1990. The West European Party System. Oxford University Press, Oxford, USA.
Maisel, L., Cooper, J. (Eds.), 1978. Political Parties: Development and Decay. Sage Publications, London.
Maisel, L., Sacks, P.M. (Eds.), 1975. The Future of Political Parties. Sage, Beverly Hills, CA.
Mayhew, D.R., 1986. Placing Parties in American Politics: Organization, Electoral Settings, and Government Activity in the 20thCentury. Princeton University Press, Princeton, NJ.
Mehra, A.K., Khanna, D.D., Kueck, G.W. (Eds.), 2003. Political Parties and Party Systems. Sage Publications, London.
Meny, Y. (Ed.), 1989. Idéologies, Partis Politiques, et Groupes Sociaux. Press de la FNSP, Paris.
Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.
Michels, R., 1911. Soziologie des Parteiwesens. W. Klinkhardt, Leipzig, Germany.
Moore Jr., B., 1968a. Injustice. The Social Bases of Obedience and Revolt. Princeton University Press, Princeton, NJ.
Moore Jr., B., 1968b. Moral Aspects of Economic Growth and Other Essays. Cornell University Press, Ithaca, NY.
Norton, P. (Ed.), 1996. The Conservative Party. Oxford University Press, Oxford, UK.
Offerlé, M., 1997. Les Partis Politiques. PUF, Paris.
Ostrogorski, M., 1903. La démocratie et l'Organisation des Partis Politiques. Calmann-Lévy, Paris.
Pippa, N. (Ed.), 1999. Critical Citizens: Global Support for Democratic Government. Oxford University Press, Oxford, UK.
Pombeni, P., 1994. Partiti e Sistemi Politici Nella Storia Contemporanea. Il Mulino, Bologna, Italy.
Rawls, J., 1993. Political Liberalism. Columbia University Press, TN.
Rohe, K. (Ed.), 1990. Elections, Parties and Political Traditions. Social Foundations of German Politics and Party Systems, 1867–1987. Berg, New York.
Rose, R., 1984. Understanding Big Government: The Programme Approach. Sage Publications, London and Beverly Hills.
Sartori, G., 1976. Parties and Party Systems. Cambridge University Press, Cambridge.
Seiler, D.L., 1986. De la Comparaison des Partis Politiques. Economica, Paris.
Skocpol, T. (Ed.), 1999. Democracy, Revolution, and History. Cornell University Press, Ithaca, NY.
Tremblay, M. (Ed.), 2008. Women and Legislative Representation: Electoral Systems, Political Parties, and Sex Quotas. Palgrave-MacMillan, New York.
Tufte, E.R., 1980. Political Control of the Economy. Princeton University Press, Princeton, NJ.
von Beyme, K., 1985. Political Parties in Western Democracies. Palgrave-MacMillan, London.
Ware, A., 1996. Political Parties and Party Systems. Oxford University Press, Oxford, UK.
Weber, M., 1919. Politik als Beruf. J.B. Mohr, Tübingen, Germany.
Wong, J., Friedman, E., 2008. Political Transitions in Dominant Party Systems: Learning to Lose. Routledge, London.

Political Parties: Communist

DS Bell, University of Leeds, Leeds, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Anderson, volume 4, pp. 2331–2334, © 2001, Elsevier Ltd.

Abstract

Communism is a loosely used term applied to a variety of political parties, but in this article the focus is the Communist Party as established and run by Lenin and its organizing principle of democratic centralism and relation to the Soviet Union. These parties were parts of the world Communist system and had the vocation to run the societies that they took over, whether by revolution as in Russia itself or put in power by the Soviet Army as in the Eastern European regimes.

In the discussion of the Communist Parties, several issues need to be disentangled. In the first place the Communist Parties are Marxist Parties, but the Marxist vision of industrial society was appropriated by most of the left-wing Socialist movements in the late nineteenth century and Marxism (with a few exceptions) became the common property of the left in different versions. Thus, at that time, many parties of the left would have claimed to be 'Communist Parties' and 'revolutionary' without having links to the Russian Party or Moscow's institutions. Secondly, Communist Parties were organized on the basis of 'democratic centralism' and that distinguishes them from other Socialist parties of the left and from their rivals. Thirdly, Communist Parties were established to support the Russian Bolsheviks after Lenin's 1971 coup d'état (in orthodox Communist parlance the 'October Revolution'), and subordinate Communist Parties formed parts of a worldwide system until the collapse of the Soviet Union itself. Their *raison d'être* was the defense of the USSR, the land of 'really existing Socialism,' the workers' fatherland where the Socialist society was being built. Thus their lodestar and guiding principle was the defense of the Soviet Union as defined by the Russian Communist Party's institutions in Moscow.

Communist Parties claim to be the authentic interpreters of Marxism but, of course, their views should not be confused with Marx's philosophy, which is a social theory about the coming revolution in industrial societies. In it Marx argued that the working class would be subject to such oppression that they would ultimately be forced down below the threshold of poverty and then would rise up and overthrow capitalism. Thus freed from capitalism, the workers would install a new (Socialist) society that would in turn lead to Communism. Socialist society, in a phrase used by Communist Parties but (having poisonous connotations) dropped latterly, was the 'dictatorship of the proletariat.' But the proletariat, or industrial workers, having no ulterior or selfish interest, would free humanity from oppression. Marxism is thus a theory of revolution that predicts how the revolution will happen and which class in society will carry it out. It does not say anything about how to organize revolution, the future society, or how Socialism is to be run. In fact, as Popper pointed out, the very flimsy depiction provided by Marx of the Socialist society is one that looked somewhat like mid-twentieth century Europe and not a Communist utopia. However, crucially, Marx does not say how the revolution is to be accomplished and Marx does not provide a handbook for government. Marx ignored the crucial political problems of how to take power and how government should be organized. These, Marx assumed, would be spontaneously solved or painlessly resolved when the class of the future, the proletariat, took power. Initially Communist Parties had a naive view of the state and organization, but Lenin developed the theory and practice of the Revolutionary Party: the guide and tutor of the working class leading it to its vocation to liberate humanity, and the ruler of the Socialist state. Alone, without the Communist Party as its vanguard, the working class would not rise to its mission and would content itself with minor reforms (Lenin *What is to be done?*).

However, by the turn of the nineteenth century the big European Socialist parties had realized that the prediction made by Marx was wide of the mark by a substantial margin; European society was not getting poorer and the propensity to revolution was absent from industrial settings (there were dramatic demands about wages and conditions). During the first decades of the twentieth century, for many on the left, the theory of 'revolution' as set down by Marx was abandoned or watered down by many mainstream left wing parties.

The Leninist Party

Lenin, an isolated figure before the First World War but the sovereign theorist of Communist Parties, maintained the 'revolutionary faith' and the Party was fashioned to implement that vision. Lenin's Bolsheviks, a faction inside the Russian Social Democrats, were organized around obedience and discipline. This organizing principle is 'democratic centralism' and changes in doctrine and alliance are secondary to this practice. Adoption of this 'democratic centralism' was the seventh of Lenin's 'Twenty-One Conditions' that local Communist Parties had to accept to be admitted to the Third International. In June 1921 the Komintern declared that the Communist Parties were the directing army of the proletariat through capitalism and socialism into Communist society. As it is usually formulated (and it is not the subject of extended exegesis like other Communist precepts) democratic centralism may appear to be straight forward: it is described as something like open discussion and then unity around the policy, once it is decided. In practice it is highly centralized and dictatorial, providing leaders with a paramilitary structure. Any criticism of the leadership and its prevailing policy

can be denounced as 'factionalism' and as aid to the class enemy (Carew-Hunt, 1950; Kolakowski, 1978).

In addition, at the Communist Party's core were professional bureaucracies staffed by a narrow group, in reality appointed by the leadership from the top. Communist Party practice amounts to a militaristic discipline and the execution of orders. In 'democratic centralism' lower bodies are subordinate and submit to the higher organs, and contacts across the base of the party or members grouping together are outlawed as 'factional activity.' Internal criticism is severely constrained and the internal elections amount to nomination by the leadership. Promotion depends on the leadership. It is a hierarchy that mimics 'bourgeois society' but the elected Communist representatives (parliamentarians, for example) will be low in the structure and under the purview of superiors in the Party.

Trotsky, before joining Lenin's Bolsheviks, was one of the first to observe the consequences of 'democratic centralism': "Lenin's methods lead to this – the party organisation at first substitutes itself for the party as a whole. The central committee substitutes itself for the party organisation, and finally a single man substitutes himself for the central committee" (Waller, 1981: p. 117). Lenin, who was surely aware of this dynamic, fashioned the Party for himself although it was taken over by Stalin, Lenin's 'magnificent Georgian.' It was 'democratic centralism' and not revolution that was the main contention on the left after Lenin seized power. Communist Parties, established on Leninist lines, were also the embryo of future society and, in this sense, were 'totalitarian' parties destined (where they came to power) to control all aspects of society. In a way the Communist Parties, without the familiar Browning motion of political parties, are instruments for revolution and control, and have eliminated the politics of negotiation, bargaining, and debate as it is usually conceived in open societies (Polan, 1984). In this way, trade unions, for example, become the instrument of the Communist Party (the 'transmission belt' to the masses); and fronts and pressure groups in the Communist orbit functioned in the same way, promoting Communist Party positions, though tailored to the particular audience (peasants, industrial workers, women, peace movements, and so on).

Communist Parties have, in ideal type and as a general model, a pyramid structure with the apex – leadership – issuing orders. At the base of the Communist Party is the cell (sometimes called the 'primary party organization') formed with three or four Party members. These small groups are for electoral purposes and are usually organized in electoral districts; but this was not the priority (the basic unit being electoral in other parties) and cells were usually set up in factories, farms, recreational organizations, and workplaces. Cells have a Secretary and receive instructions from the hierarchy but cells are isolated from each other and cannot campaign or develop policies contrary to those of the leadership. This is a robust system that enables both control by the leaders and clandestine activity by members who are *in partibus infidelium* – in capitalist society – or who are controlling society in Communist Party states. Further up the pyramid there will be a Federal Committee, but at the top are the Central Committee, a Political Bureau, and above that is a small Secretariat. Central Committees, sometimes misleadingly called the 'parliament of the Communist Party,' are 'elected' by party congresses and in turn designate the small Political Bureau and the smaller Secretariat. These elections are in reality nominations with one candidate provided by the Secretariat and are invariably unanimous.

Party debates are severely circumscribed and opposition, if expressed, is isolated and crushed. Communist Congresses are characterized by unanimous or near-unanimous votes (usually taken in public with a show of hands) and so are the votes of higher bodies. Because the Secretary General has the power of nominations, the central figure in the Communist Party (outside times of 'collective leadership' when no dominant figure has emerged) is the Secretary General. Moreover, the Secretary General emerges from the central core apparatus of the Communist Party bureaucracy. This apparatus regularly produces the 'prodigious mediocrities' that run Communist Parties and Communist states.

This pyramid of power is run by the central apparatus that Lenin established: the professional revolutionary. Communist Parties are essentially the central core around which are wrapped the activists and front or parallel organizations. Importantly, discipline also enabled the orchestrating of campaigns and the simulacrum of grass roots upsurges in opinion. This professional core is the lever with which Lenin intended to move the world. Communist Party functionaries, 'apparatchiks,' are nominated, promoted, and made by the leaders and, like the Duke of Newcastle's Bishops, they worship their maker. Full time organizers and workers on the Communist Party payroll play a very important part in Communist politics and are a powerful resource in an area of social life that is rarely organized and depends on amateur or volunteer support. Thus, in a fully developed Communist Party the main posts will be held by people who have ascended the apparatus and who are capable of taking orders and applying them with vigor and efficiency (however, counter-intuitive they may be). This structure may not be evident to all of those working in the Communist Party and Party employees are described as 'workers' in industry. It is, of course, a larger establishment than other parties and is expensive to run. Subsidies from the Soviet Union and control of trade with Soviet satellites were important in the past but the local Communist Parties developed resources of their own (local government and unions, for example).

Third International Parties

It was the First War that gave the far left its crucial boost and the Bolsheviks, led by Lenin, were able to take advantage of war weariness and increasing privations. But the Russian Revolution also raised a major disagreement that crystallized on the socialist left after 1917. When the Lenin's (Communist) Third International was set up in 1919, Western Socialist Parties determined that the route to future progress would be electoral and endorsed the rights in the constitutions of modern states. By the mid-twentieth century, all but a few of the Socialist parties in Europe had abandoned Marxism even in watered-down versions and it was left to Communist parties, although other minor splinter parties also laid claim to Communism. (McInnes, 1975)

After the Russian Revolution the Western Socialist parties supported the new government. Tsarist Russia, corrupt and

repressive, had few friends on the left wing of politics and the installation of a Socialist government under Prime Minister Kerensky was hailed as great step forward. This support generally remained after Lenin deposed the Social Democrats and installed the Communists (Bolshevik Party) in power. There was a degree of confusion on the left because Lenin appropriated the Revolution and support for the new Russia was widespread. When Communist Parties were founded in the early 1920s, many former pacifists and idealists joined on the assumption that they were joining a new formation dedicated to defending the Revolution, but essentially a more determined variant of the Socialists. There followed a period of 'bolshevization' of the Communist Parties, which saw the eviction of the initial leaderships and the installation of Moscow loyalists.

This is where the distinctive features of the Communist Parties become crucial. Parties were centralized and disciplined where possible (local difficulties, like civil wars, meant different rates of progress). This process of so-called 'bolshevization' took place under the auspices of the Third International executive in Moscow, the Komintern founded in 1919. In this, the local elites who had rallied to the Third International were replaced with Moscow's reliable cadres. Those who joined in the initial wave were often at odds with the Komintern outlook: usually because they were admirers of the Russian Revolution who had joined out of antiwar sentiment, because the Socialist Parties did not – as they saw it – defend workers sufficiently vigorously in the postwar recession, or in the upwelling of idealism that the Communist Parties tapped into. In this phase of development the Communist Parties, where they were free to do so, set up their own unions and competed directly with the 'reformist' Socialist trade unions and Socialist parties. Trade unions were alerted to this threat and in many industrial countries perceived Communist Parties as hostile to unions and to union influence.

Communist Parties had a vocabulary that portrayed them as the artisans of working class unity and as the most advanced part of the proletariat. They would not allow that they were the factor of division of the proletariat but that other malefactors or class enemies were, through 'false consciousness' or 'economism' (devotion to short-term wage increases and the like) disrupting the unity of the class with the revolutionary vocation. But the Communist lexicon enabled a considerable tactical flexibility. Communist Parties could emphasize or downplay parts of the Marxist message according to circumstances. Churchill said that the Communist Parties zig-zagged but you never know whether they were on a zig or a zag.

It was the Soviet leadership that decided the exact tactics that parties would take. Communists differentiated between 'unity from below' and 'unity at the summit.' 'Unity from below' was an attack on 'bourgeois society' and its outriders the Socialist Parties and unions; and 'unity at the summit' involved collaboration between parties and non-Communist institutions (which could be conservative or Christian) in pursuit of some aim. Communist Parties at this time took a revolutionary line in which the Socialist Second International Parties were the principal impediment to the working class realizing its revolutionary potential; and attacks on 'Social Fascists' continued even after Hitler came to power in Germany.

From 1917 and until the late 1920s Communist Parties were set up, usually by enthusiasts for what they saw as the Revolution's ideals. These parties sent delegates to the Third International (Komintern) in Moscow and formed the nucleus of the world Communist movement. Komintern's establishment of its authority in the mid-1920s created the hierarchical and bureaucratic Communist Parties under the Komintern Executive Committee with, usually, local leaders placed in position by Moscow. This 'bolshevization' should not, as is often the case, be confused with stalinization. A power struggle in Moscow after Lenin's death saw the emergence of Stalin as the Soviet Party's leader, and many on the Komintern Executive had been opponents of Stalin's rise or not enthusiastic about him. Stalin's preponderance was signaled by a change of line to 'Socialism in one country' which recognized the failure of revolution in Europe, and the focus on the defense of the building of Socialism in the Soviet Union. Leaders like Zinoviev were replaced and reliable placemen installed. This process of eliminating the unreliable that Lenin urged and which he referred to as 'purges' (sometimes loosely referred to as 'stalinization') took place both in Moscow and at the national level, although many leaders changed horses convincingly enough to be retained. They had usually worked to convince the Moscow leadership of their reliability and willingness to execute orders, however suicidal from a local point of view they might appear to be. Palmiro Togliatti, the Italian official who apparently had Stalin's ear, was superior to other Europeans in the Komintern hierarchy and may have had greater leeway than others. The French Communist Party leader Maurice Thorez, although leading the biggest Communist Party, was subordinate. Georgi Demitrov, the Bulgarian Communist, took over as the head of the Komintern Executive and in the 1930s there were purges of foreign Communists who were, in some cases, handed over to enemies.

In 1928 Stalin's increasing domination was behind a further change in policy to 'class against class' sectarianism: the aggressive attack on 'bourgeois' institutions and the repudiation of the big working class parties ('Social Fascists' in Communist terms) and institutions on the grounds that the capitalist crisis was entering its final phase of convulsion. This policy, disastrous for local parties which lost support, meant vigorous attacks on the Socialist Parties and led to the rise of Nazi Party, which Communists persisted in seeing as an epiphenomenon (Mortimer et al., 1979: Chapter 2). However, the rise of Hitler could not be dismissed in this cavalier fashion and a U-turn in policy to inaugurate cooperation against the Hitler Right was introduced by Moscow in 1934. This was the 'Popular Front' strategy intended to bring together all anti-Nazi forces: these forces were mainly the Socialist Parties but the strategy was a Soviet foreign policy *demarche* and not a domestic, local, one (Joll, 1959: p. 56). Hence the range of the appeal was not limited to the left and the Communists, and in pursuance of this objective, the Popular Front strategy stripped out the radical elements of their appeal and rhetoric and emphasized the patriotic and national elements of politics. Thus the reforms with which the Popular Fronts are sometimes credited, were Socialist and not Communist in origin.

Although Komintern was dissolved during the Second World War in May 1943 in deference to US and UK Allied sensibilities, the change was more apparent than real and the central control of the Communist movement was maintained through the International Department of the Central

Committee of the Soviet Communist Party. However, Stalin's control, although indirect, remained the guiding force for the Communist Parties. After the Liberation, Stalin's immediate priority appears to have been the establishment of Communist Party control in the countries overrun by the Red Army in Eastern Europe, and the Western Parties played a low-key and accommodating role in reconstruction. This changed when the United States returned to the European arena in support of the non-Communist countries and the line again changed to one of aggressive attacks on 'bourgeois' institutions. This was established in September 1947 after the Marshal Aid programme was implemented by the United States, when the Kominform was set up (Marcou, 1987).

Kominform was to inaugurate a new offensive line in Western Europe that saw the Communists orchestrate strikes and riots in many countries in opposition to the Marshal Plan. This new line probably took the Western Communist Parties by surprise and the delegates from France and Italy at Komintern's Szlarska Poręba first conference in Poland found themselves in the dock, subject to volleys of abuse for lack of revolutionary spirit. Despite the creation of this new focus for Communist Parties, the Komintern was less important to parties than direct relations with the CPSU International Department. In 1956 Khrushchev dissolved the Kominform and the CPSU's International Department took over the office of coordination. This process was assisted by the editing committee of the journal *World Marxist Review* published in Prague after 1958. The *World Marxist Review* was a largely Russian enterprise but was destined for Communist Parties across the world, and it set out the Communist position coordinating Parties on issues of the day, until it was wound up in 1990.

Postwar Fortunes

At the Liberation, the second wave of Communist electoral growth began and the emergence of the Communist Parties in France and Italy as major forces was consolidated. In Southern Europe, Portugal, Spain, and Greece, Communism emerged from the dictatorships and became electorally significant. The mechanism here was probably similar to the Second World War Occupation in France and Italy: the Communists, secretive and well-suited to clandestine work, were the principal opposition to the occupations and to dictators and benefited from being the main opposition to the regime in these countries. If people were disposed to struggle against the dictatorships they would very likely find themselves in the Communist ambit – if not the Party itself – and participating in an opposition portrayed as patriotic first and Socialist/Communist second

Khrushchev, however, changed the Communist Parties' line and moved to peaceful coexistence and an understanding based on the nuclear stalemate in Europe that froze the two blocs until the late 1980s. Communist Parties, henceforth, would seek the 'parliamentary road to power.' In the history of Communist Parties and in Western Europe, the key date is 1956 when the Communists changed tactics from revolution in Europe to the 'peaceful road to power' through parliamentary elections. This was Khrushchev's repositioning strategy and reflected a *realpolitik* recognition of the balance of power in Europe, and by contrast, the revolutionary possibilities in the decolonizing European Empires in Africa and Asia. Khrushchev's strategy was, it was stressed, not to be confused with the 'parliamentary cretinism' of the Socialist Parties. Socialists put all the emphasis on reform through 'bourgeois' institutions, in the belief that a more moderate form of capitalism could be developed that provided welfare and security without destroying the market mechanism. Communist Parties did not accept this 'sell-out.' Communist Parties were expected to use the opportunities of 'bourgeois' institutions without abandoning their ultimate revolutionary vocation.

Hence Communist Parties in Western Europe began to seek partners within the party systems of Europe for coalition alliances and this meant – inevitably – targeting the Socialist Parties. However, Socialist Parties had broken links with the Communist Parties in the Cold War and were suspicious of Moscow's intentions. With the memory of the take-over of Eastern Europe and with Communist coups in mind (and those with close union links were aware of the suppression of unions in the Soviet bloc), there was no rush to seal deals with the Communist Parties. Communist Parties seeking alliances were largely unsuccessful and they met notable rebuffs in Italy, Spain, and Portugal; and only in France was there a modicum of success with the formation of the Socialist-Communist Common Program alliance in 1972. In 1981 President Mitterrand offered the Communist Party four minor ministerial portfolios in a Socialist-dominated government (that had no need of Communist support). When they quit 4 years later the experiment was terminated.

If Khrushchev's strategy had meager returns, it did help to loosen the ties of the Communist Parties with Moscow. There was a tension in the Communist strategy of alliances within a parliamentary system: local Communist Parties had to show that they were not controlled by Moscow and this led to the 'Eurocommunist' interlude in the 1970s. Communist Parties in France, Italy, and Spain announced their differences with Moscow (and even picked quarrels) in order to demonstrate their patriotic autonomy. This was, however, more a presentational device than a real independence. However irritating it was to the CPSU and to Moscow loyalists, the divisions were superficial and cosmetic. Thus the Italian slogan 'NATO out of Italy, and Italy out of NATO' may have looked even-handed, but the Warsaw Pact forces which suppressed uprisings against the Communist Parties in East Germany, Hungary, and Czechoslovakia were on the continent and under Soviet control. Likewise the protests against the Cruise and Pershing missile bases set up in the 1980s were in historical fact a response to Soviet missile deployment and not a one-sided NATO initiative. To some extent this tension between the USSR and local Communists subsided in the 1980s, when the French Party supported the Russian invasion of Afghanistan, the Spanish Party imploded, and the Italian Communist Party lost support.

Communist Party Control

In government the Communist Party's top-down command structure and its permeation of society constituted a control mechanism for the whole of society. In a Communist state, the so-called 'People's Democracies,' government departments, and civil society were under the oversight of the Party and

under its control, with the result that a highly centralized and responsive state–society relation was constructed, but one at the command of the leadership. Communist Party control of civil society more or less eliminated intermediate institutions and those that remained were subordinate to the Party state. This was not a once-for-all process and the incorporation of institutions into the Communist system took place with varying degrees of efficiency and some, such as the Polish Church, never fell under Party control. However, the Party supervised and ran the ostensibly independent trade unions, writers' organizations, theater and film makers, and so on. In the production of the economic plan this central control was crucial: the Party took over and ran the factories and workshops and independent businesses with only peasant agriculture (partially in some countries) escaping this extensive control.

In its developed stages in the Soviet Union, this party–state interaction was a coherent and almost monolithic structure with internal dissent subdued and usually eliminated. As the system matured, the establishment of 'baronies' in regions or in some sections of society and industries began to emerge and the control by the center weakened, although the central command structure and the Party organization was still the central feature. Communist Parties found it increasingly difficult to maintain central control and to devolve power to enable innovation and initiative, with the result that the system did not bend under pressure, but broke. Whether the remaining Communist Party systems can introduce the flexibility required by marketisation of the economy and retain an absolute political dominance remains a moot point.

Electoral Support

Though they have not neglected the electoral arena, traditionally, Communist Parties have not regarded the electoral arena or parliaments as the key to political power. In the major societies the Communist Parties have usually been marginal, where there is a system of competitive elections; and before the Second World War they had a party system impact only in France, Germany, and Finland. In the 1930s there was not a great deal of electoral success and in many countries they were outlawed. After the Second World War they had more success, but in open societies few Communist Parties were major components of the left and in only Finland, France, and Italy (and Cyprus later in the 1970s) were they genuine mass parties. In the late 1940s the Communist Parties had some success claiming to be the most advanced part of the Socialist movement and promising peaceful change; moreover they had the prestige of the USSR, a victorious ally, and the Soviet system (about which at that time little was known by most people) behind them. In many countries they made a respectable showing at the polls perhaps winning 5% or more after the Liberation: over 28% in France, 23.5% in Finland, 13% in Belgium, and 12.5% in Denmark. (Setting aside such minuscule oddities as San Marino, where the Communist Party polled 42% in 1949.) Elsewhere they were sect-like but adept at infiltrating unions and organizations or running fronts, if not at winning votes. (McInnes, 1975) In some countries, Germany, the USA, Canada, Australia, New Zealand, and the UK notably, they made no electoral impact whatsoever. This slight ascension of the Communist Party vote was reversed with the onset of the Cold War and in most countries they were, by the 1950s, reduced to groupuscular status. Communist Parties were successful when they were presented as social democratic parties, promoting reform and workers' rights, but when they were hostile to Western institutions and vaunted their attachment to Moscow their vote fell or was reduced to virtually nothing.

Not the least of the paradoxes of the Communist Parties (the parties of the working class) is that the successes, such as they were, came in the 'wrong' places: in the backward and peasant countries of Russia, China, Vietnam, and Cuba. In the mid-twentieth century (during and after decolonization) the Communist Parties made their advances in the backward and peasant societies of the Third World – Africa in particular – but not in the industrial world. Marx's theory was about the development of mature and decaying capitalism and, if the theory had been correct, the site of revolutions would have been the advanced industrial countries, such as Germany, the USA and the UK where, as Marx saw it, the newly emerging proletariat in the factories and workshops would sweep away the old systems. Even in Western Europe the Communists found themselves positions in the mainly poor agricultural areas, in traditional industries of mining, steel, and coal and to some extent in the public sector. Electorally the same paradox appears in that the Communist Parties were strongest in the poor agricultural regions like the borders of the Central Massif in France, North Finland, the Alentejo in Portugal, and the declining heavy 'rust belt' industries of coal, steel, and textiles. Communist Parties, despite the emphasis on science and progress, were part of the nineteenth century sectors and found it impossibly difficult to break into the new science-based and information-based businesses.

Post-Wall Communist Parties

In Europe after 1992, it needs to be said that the collapse of the Soviet Union caused the parties identifying with Moscow to transform themselves into quite different party organizations or to claim to be new in some respect. Hence it is difficult, in the absence of a Communist International, to say precisely which are Communist and which have moved away from the initial model . Very few, even of the post-Communist Parties, retained 'democratic centralism' as an operating system, although there are a number of Trotskyite parties that claim to be the 'real' inheritors of the Communist message and which are Leninist. In the twenty-first century, those proclaiming themselves as Communist parties include the former Soviet loyalists who emphasized the original vocation in Spain, Portugal, Greece, France, and the Third International dissidents – Trotskyites – who have not abandoned their ancient rancors. Other Communist Parties became part of rainbow alliances of ecologists, decentralizers, and new leftists in minor parties. (Courtois, 2011: Part 4, esp. Chapter 23)

Governing Communist Parties, in the Eastern bloc but also in the Soviet Republics and the East, also repudiated the Soviet past and transformed themselves into something else: this was usually as a Socialist Second International Party, and the Communist Parties in their initial incarnation disappeared.

There are exceptions: the best known are North Korea, Cuba, and China where the ruling Communist Parties have not experimented with open elections and where the solidity of the one-party state has remained intact. In North Korea the military state remains intact but under the façade of the Communist Party and its long-running dynasty. Cuba has opened sectors to the outside but remains a one-party state organized around the remaining revolutionary elite (Rigoulot, 2007). But China, the most innovative of the post-Soviet states, has, by unleashing market forces in the economy but remaining in control of the political institutions, found new ways to shore up the totalitarian system. One of the unknowns about the remaining state Communist parties is whether they can manage a transition to a market economy while maintaining the pervasive political domination of the Communist Party (Kissinger, 2012). This proved impossible in the Soviet Union and its satellites but a one-party state running a market economy ('market-Leninism') cannot be excluded.

See also: Dictatorship; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Party Systems; Social Democracy; Socialism; Totalitarianism.

Bibliography

Aron, R., 1957. The Opium of the Intellectuals. Secker and Warburg, London.
Backs, U., Moreau, P. (Eds.), 2008. Communist and Post-Communist Parties in Europe. Vandenhoeck and Ruprecht, Göttingen.
Bull, M.J., Heywood, P. (Eds.), 1994. West European Communist Parties after the Revolutions of 1989. Palgrave, London.
Carew-Hunt, R.N., 1950. The Theory and Practice of Communism. Penguin, London.
Colas, D., 1982. Presses universitaires de France. Le Léninisme, Paris.
Courtois, S. (Ed.), 2011. Sortir du communisme, Changer d'époque. Presses universitaires de France, Paris.
Furet, F., 1999. The Passing of an Illusion. Chicago University Press, Chicago.
Gombin, R., 1975. The Origins of Modern Leftism. Penguin, London.
Gotovitch, J., Delwit, P., De Wael, J.-M., 1992. L'Europe des communistes. Editions Complexe, Brussels.
Hobday, C., 1986. Communist and Marxist Parties of the World. Longman, London.
Holmes, L., 1986. Politics in the Communist World Oxford. Oxford University Press.
Joll, J., 1959. The Making of the Popular Front. St. Antony's Papers, Oxford.
Kissinger, H., 2012. On China. Penguin, London.
Kolakowski, L., 1978. Main Currents of Marxism. Volume 3 Chs 17 and 18. Oxford University Press, Oxford.
Lazar, M., 1988. Communism in Western Europe in the 80's. The Journal of Communist Studies 4 (3), 243–257.
Lenin, V.I., 2014. Lenin: Collected Works. http://www.marxists.org/archive/lenin/works/1906/may/20c.htm.
McInnes, N., 1975. The Communist Parties of Western Europe. Oxford University Press, London.
Marantzidis, N., Lamprini, R., 2011. L'évolution électorale et politique des partis communistes en Europe 1076–2009. In: Marcou, L. (Ed.), 1990. Le Mouvement Communiste International. Presses universitaires de France, Paris.
Marcou, L., 1987/1947. la guerre froide: L'engrenage (La Mémoire du siècle. Editions Complexe, Brussels.
Mortimer, E., Filo della Torre, P., Storey, J. (Eds.), 1979. Eurocommunism: Myth or Reality. Penguin, London.
Nove, A., 1983. The Economics of Feasible Socialism. George Allen and Unwin, London.
Polan, A.J., 1984. Lenin and the End of Politics. Methuen, London.
Popper, K., 1966. The Open Society and Its Enemies, vol. 2. Routledge, London.
Primus, R., November 2, 1996. A brooding omnipresence: totalitarianism in postwar constitutional thought. Yale Law Journal 106, 423–457.
Rigoulot, P., 2007. Corée du Nord, Etat voyou. Buchet Chastel, Paris.
Schapiro, L., 1972. Totalitarianism. Pall Mall, London.
Schapiro, L., 1985. The Communist Party of the Soviet Union. Methuen, London.
Thomas, H., 2011. Cuba: A History London. Penguin.
Tiersky, R., 1986. Ordinary Stalinism. Allen and Unwin, London.
Walker, M., 1972. The Language of Communism. Bodley Head, London.
Waller, M., 1981. Democratic Centralism. St Martin's Press, New York.
Waller, M., Courtois, S., Lazar, M. (Eds.), 1991. Comrades and Brothers: Communism and Trade Unions in Europe. Cass, London.
Webb, C., 1979. Eurocommunism and Foreign Policy. Policy Studies Institute, London.

Political Parties: Extreme Right

Christopher T Husbands, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes how 'extreme right' became part of the vocabulary of political science. It discusses the introduction of the cognate term 'radical right.' It offers its own comprehensive definition of extreme-right parties intended to cover a range of examples in different locations since the early nineteenth century. It discusses whether some parties labeled by many authors as 'extreme right' or 'radical right' ought rather or also to have 'populist' as a component of their labeling. It periodizes twentieth and twenty-first century extreme-right parties and discusses theorizing and research on the extreme right. It focuses first on debates about German Nazism, discussing contemporary and more recent theories about it. Various theories about postwar extreme-right parties are then described. Findings about the social base of extreme-right parties are summarized. Alternative forms of extreme-right movement are described. The tendency of extreme-right parties to splinter is discussed. Finally, some possible trajectories about the future of extreme-right parties are introduced.

Political parties that current conventional labeling by many political scientists and other informed analysts describes as 'extreme right' have emerged in many countries and on many occasions since at least the early nineteenth century, notwithstanding a debate about appropriate descriptors. The label, as employed in this article, is necessarily an encompassing one because it is applied generically to numerous parties that nonetheless differ from each other in various ways and on to which, in some cases, not every recognized specialist is willing to confer such a label.

The Term 'Extreme Right'

The word 'extreme' has been used in a political sense, usually with some of the same meaning and connotations as today, since at least the mid-nineteenth century. However, the term 'extreme right' and its cognates are surprisingly recent arrivals in political science's standard vocabulary, having been regularly employed in the discipline only since the 1950s. Lipset (1959a/1963), for example, used the term very occasionally in his *Political Man* and published an article in 1959 with 'right-wing extremism' in its title given in quotation marks (Lipset, 1959b), it became conventional usage among European (except German) commentators during the 1960s, and it achieved final recognition in American political science with Lipset and Raab's (1970/1971) *The Politics of Unreason*. One reason for the slowness with which the term came to be adopted in the United States was the popularity of the alternative, and still widely used, 'radical right' (e.g., Kitschelt and McGann, 1995; Norris, 2005) in order to describe political phenomena that other present-day analysts would call 'extreme right.' Lipset had also been responsible for the introduction of the term 'radical right' into American political science, and so for its popularity during much of the following decade. It was perhaps taken from the usage in German of *rechtsradikal*. This neologism, although also postwar, can be traced back to its introduction into German in the early 1950s, probably influenced by the fact that Germany's most significant extreme-right party was then the later-proscribed *Sozialistische Reichspartei* (SRP). This designation, though sometimes also *extrem rechtsradikal*, was widely used by German authors in the 1960s and 1970s, and has only relatively recently been superseded among a majority of German political scientists by *rechtsextrem/rechtsextremistisch* (e.g., Greß et al., 1990; Jaschke, 1994); the America-based Minkenberg (1998) was an apparent exception, although he has also shown his readiness to use the word 'extremism' as a synonym for these phenomena (e.g., Minkenberg, 2009). The distinction was also a bureaucratically important one in Germany, since a purported, if casuistical, distinction between *rechtsradikal* and *rechtsextrem* was the basis on which the German *Bundesamt für Verfassungsschutz* (Federal Office for the Protection of the Constitution, BfV) justified varying monitoring and control strategies toward different right-wing formations. By the later 1990s, extreme-right activities in Germany made this distinction increasingly dubious. The translation into English of *rechtsradikal* as 'radical right' long remained peculiar to American political science; the translation given for it in the standard German–English dictionaries published in the 1970s, both British and American ones, was still 'right-extremist,' 'extreme right-wing,' or simply 'rightist'; 'radical' was introduced later into the translated meaning.

A Definition of Extreme-Right Parties

An inclusive definition of extreme-right parties has necessarily to be extensive and complex; in order to include all types of party that have attracted the designation, though not necessarily a consensus attribution by all analysts, the definition must be appropriate for such parties and movements as diverse in time and location as the American Anti-Masonic Movement of the 1820s; the Know-Nothing Movement of 1850s America; the French Boulangist Movement of the 1880s; interwar fascism in Italy and other European countries; Nazism in Germany; McCarthyism in 1950s US; Poujadism in 1950s France; George Wallace's 1968 American Independent Party (AIP); recent European examples such as the French *Front National* (FN), the German *Nationaldemokratische Partei Deutschlands* (NPD), the *Deutsche Volksunion* (DVU, which merged with the NPD on 1 January 2011), and *Die Republikaner* (REPs); the British National Party (BNP); Vladimir

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93087-3

Zhirinovsky's *Liberal'no-Demokraticheskaya Partiya Rossii* (Liberal Democratic Party of Russia); the current Hungarian *Jobbik Magyarországért Mozgalom* (Movement for a Better Hungary, henceforth *Jobbik*); some of the Indian Hindu nationalist parties such as *Shiv Sena* (Shiva's Army); certain types of new Christian right fundamentalism, especially as seen in the United States within the Republican Party; and purportedly more mainstream political parties of the present such as Austria's *Freiheitliche Partei Österreichs* (FPÖ) and Switzerland's *Schweizerische Volkspartei* (SVP). The contemporary American militias are usually regarded as extreme right, but they are idiosyncratic examples not regarded as political parties, their most obvious European analogue perhaps being the militant neo-Nazi 'defense groups' of Germany and some Scandinavian countries. Although far from all analysts agree on the 'extreme right' designation for every one of these, each has nonetheless attracted such an attribution from at least a critically minimum number of historians, political scientists, or informed journalist observers. Commentators from some countries have been more ready than those elsewhere unwaveringly and without collateral justification to label contemporary examples as 'extreme right.' An inclusive definition of the extreme right has to be based on ideological features such as those below and, although each would not necessarily apply with equal emphasis and significance, or at all, to all extreme-right parties, all political scientists using this concept would understand that many or most would, at leadership or mass-support level, characterize a political party so designated:

- selective inclusion, especially within the nation and often based on preferences or assumptions about ethnic or religious similarity (which may take the form of ultranationalism, ethnocentrism, or hostility to a particular religion such as Islam or Roman Catholicism);
- selective exclusion, based on xenophobia or directed against indigenous or immigrant ethnic minorities, but also against social minorities such as homosexuals, travelers (Roma), asylum seekers, or in the past even the Masons, and to whom are denied equality and equality rights;
- racism, based on supposedly biological perspectives about racial inferiority or on the fabrication of ethnic and cultural boundaries;
- anti-Semitism, which may be both specific and directed at contemporary targets, but which may in some cases involve denying, or mitigating the scale of, the Holocaust and glorifying the deeds of the Third Reich or fascist Italy;
- a preference for authoritarian initiatives by a strong state;
- the cult of a leader figure;
- a preference for a hierarchical social order;
- antisystem, antipluralist, and antiparliamentary political perspectives;
- overt hostility to political opponents, often associated with the use, encouragement, or tolerance of violence;
- low tolerance of social change; and
- nostalgia for the past.

However, it is important to note the nature of some of the variations. For example, the Hungarian party *Jobbik* (now the country's most important extreme-right formation after uniting in 2005 with the earlier *Magyar Igazság és Élet Pártja*, Hungarian Justice and Life Party), like other east-European extreme-right phenomena, is less concerned with opposing multiculturalism and immigration and more focused on scapegoating and harassing the country's internal minorities, in particular the country's Roma minority, which comprises about 8% of its population. *Jobbik* also has the very specific irredentist aim to nullify the 1920 Treaty of Trianon, under which the country's size was greatly reduced after World War I, so placing significant Hungarian-speaking minorities in present-day Slovakia and Romania. The extreme-right parties in the latter countries, such as *Slovenská národná strana* (Slovak National Party, SNS) and the *Partidul România Mare* (Greater Romania Party), are equally keen to resist the loss of territory so implied and to oppose any granting of special rights to their country's Hungarian-speaking minority as being the thin end of the wedge toward territorial autonomy. Other countries' extreme-right parties also have concerns specific to their national situations. Israel's extreme-right parties, of which there are several, have supported Jewish settlement in the occupied territories. *Eretz Yisrael Shelanu* (Our Land of Israel) for example, was formed in 2008 to seek to prevent creation of a state of Palestine and to oppose the dismantling of Jewish settlements in the West Bank. One strand of the Israeli extreme right sought the establishment of a state of Greater Israel, normally defined as the current state of Israel and the Palestinian territories. The extreme-right *Srpska radikalna stranka* (Serbian Radical Party) aspires to a Greater Serbia, a territorial concept that has varied in extent but would certainly now include areas from contemporary Croatia, Bosnia and Herzegovina, Montenegro, and Macedonia.

Attitudes to the economy, however, offer little key to a definitional criterion, for there is considerable divergence among these parties. Some, such as those like the French FN or the British BNP that seek to have or to maintain a working-class appeal, have an economic program with strongly social-democratic elements, whereas parties like the Flemish *Vlaams Belang* (Flemish Interest, VB, till 2004 the *Vlaams Blok*) or the Swiss SVP are economically liberal and in favor of free-market initiatives.

Because definitional arguments have been such a significant component of the academic literature on the extreme right, it is relevant to elaborate this debate, which has been continuing among political scientists since at least the 1980s. Although the positions of contributing authors are individually nuanced, the crux of the argument is whether it is theoretically appropriate to distinguish between right-wing extremism (or right-wing radicalism) and right-wing populism, and indeed whether some parties are too readily labeled extreme-right when they are in essence populist and might better be described by a label including that word. The terms 'neo-populist,' 'right populist,' and 'national-populist' also occur in the literature. Betz (1994), using the label 'radical right-wing populist,' was a major contribution using the distinction, and Mudde (2007) is one of the most forthright and extensively sourced recent arguments for it. Mudde opts for 'populist radical right' as his descriptor, claiming that such parties are "(nominally) democratic, even if they oppose some fundamental values of *liberal* democracy, whereas the extreme right is *in essence* antidemocratic, opposing the fundamental principle of sovereignty of the people" (Mudde, 2007: 31 (italics in original)). However, other writers – some identified with a more overtly left-wing stance – regard such a distinction as overdone and dismiss or at least

diminish it. Minkenberg (1998: 33), for example, though using his preferred term of right-wing radicalism, merely includes the idea of populism in his definition of such radicalism, which he says "is directed against liberal democracy and its fundamental values of liberty and equality, as well as against individualism and universalism." Others dismissive of the extremism–populism distinction note that all these parties are to a greater or lesser degree tainted by exclusionism and racism. Confining this encyclopedia entry only to parties fitting a reductive definition of the extreme right would necessarily cover very few significant current examples. Moreover, some parties about whom such a definitional discussion might be held have a dual strategy – participating at least for the expedient present in formal electoral politics, while at the same time having an antidemocratic and authoritarian basis to their ideology. Such an accusation might be made against, for example, the Italian *Alleanza Nazionale* (AN), a successor and palimpsest of the earlier and incontrovertibly neofascist *Movimento Sociale Italiano* (MSI), though despite its origin the AN is classified by some political scientists as not extreme right but as having refocused to become nationalist-conservative.

If, however, there is a distinction to be made among the parties that are all, at least to those on the left, to be seen as extreme right, it is partly one of organization and style and less of ideology. The Dutch *Partij voor de Vrijheid* (Party for Freedom, PvdV) of Geert Wilders is very much the creation of its founder, as was a similar Dutch party of the early 2000s, the *Lijst Pim Fortuyn*, that briefly burned brightly but after the murder of Fortuyn was quickly eclipsed by rapid descent into schism after its electoral success in the Dutch general election of May 2002. Wilders' party, like Fortuyn's before, opposes multiculturalism and is particularly hostile to Islam, wanting to retain what it claims are traditional Dutch values against an Islamic threat. It favors gay marriage and women's rights and is, in general, socially liberal. It would be unfair to equate these parties, which many Dutch political scientists now writing about them label populist and not extreme right, with, for example, the aggressive extreme right seen in Hungary's *Jobbik* or Greece's *Chrysi Avgi* (Golden Dawn, GD), with their sometimes denied but often overt neo-Nazism, their militant wings, and their unashamedly vigilante activities against specific outgroups in their countries. Indeed, in October 2013 the leader of GD and several of its MPs were arrested and accused of leading a party that was in reality, a violent criminal organization. Nonetheless, both GD and *Jobbik* secured three seats each in the European Parliament elections of May 2014.

Extreme-Right Phenomena in the Twentieth and Twenty-First Centuries

It is useful to give a very general, if crude, periodization for extreme-right phenomena of the twentieth and early twenty-first centuries, although in each of these periods there was nonetheless a considerable variety of such parties:

- 1920–45. The era of 'classic' anti-Semitic Nazism and fascism.
- 1945–70. The era of the 'old wave' extreme right, often comprising anti-Semitic neo-Nazi/neofascist parties with a yearning for the 'classic' era (e.g., Germany's SRP or Italy's MSI), but also nostalgic anticommunist movements (e.g., McCarthyism).
- 1970–2000. The era of the 'new wave' extreme right, appealing particularly on a basis of xenophobia and racism against ethnic-minority immigrant groups and asylum seekers.
- 2001–present. The era of increased globalization that has seen the emergence of new dynamics of the extreme right such as Islamophobia and the global financial collapse since 2008, including the effects of the Eurozone crisis since 2009.

This periodization is not perfect and there are anomalies. McCarthyism, for example, was a movement within the Republican Party rather than a party in its own right and it lacked the vigorous anti-Semitism of 'old wave' extreme-right parties. The American AIP anticipated the issues of the 'new wave' extreme right. However, this periodization is a useful heuristic device for discussing theoretical approaches to the analysis of these parties.

Theoretical Approaches to Explaining the Extreme Right

Fashions in political science toward the analysis of the support of extreme-right parties have not been constant. While it is important not to exaggerate these changes, it is true that research in the 1950s and into the 1970s, conducted on the vote for Nazism (the *Nationalsozialistische Deutsche Arbeiterpartei*, NSDAP) and on 'old wave' postwar parties, was especially concerned with answering questions about the partisan origins and social bases of these examples of extreme-right support. Lack of any individual-level data on voters for the NSDAP led to the use of different combinations of aggregate data in a substantial body of research about Nazism's partisan origins and social base. The consensus that emerged from the comprehensive research in this tradition has accepted the case for considerable social variety in the NSDAP's later support. Postwar parties for whose support there existed survey data were frequently analyzed in the first instance in terms of the social categories to which they particularly appealed. Ideological and attitudinal dispositions toward the extreme right were not ignored, of course (especially not by research drawing on perspectives in *The Authoritarian Personality* (Adorno et al., 1950)), but they were not always treated as the principal explicandum; indeed, many authors (e.g., some of those writing about McCarthyism in *The Radical Right* (Bell, 1963/1964)) simply imputed motive. Researchers working on parties that have emerged in the 'new wave' era have been more inclined to recognize that most extreme-right parties draw from a variety of social sources and have given greater attention to the attitudinal, psychological, and contextual bases of such support. The disproportionate support for many 'new wave' extreme-right parties among certain economic and demographic groups has been recognized, but it has not been a central focus of theorizing about them.

In giving an overview of the theoretical perspectives toward extreme-right parties, it is essential to distinguish between the pre-World War II and postwar cases. Many of the interbellum parties played major political roles, especially so, of course, in

the cases of Italy and Germany. Even in countries where such parties did not achieve power, they often had substantial impact upon national politics during the 1930s (as in France or The Netherlands). On the other hand, in the postwar period extreme-right parties, though not without some important effects upon the content of political agendas, have not usually achieved actual governmental office – significant exceptions being Italy briefly in 1959 and intermittently from the mid-1990s, Austria in 2000, Slovakia in the 1990s and after 2006, as well as the problematic case of contemporary Switzerland in the light of the metamorphosis of parts of the SVP toward the extreme right. Spain, a fascist dictatorship till 1975, is a special case.

Most theoretical analyses of the period of the interbellum extreme right focused on German Nazism and Italian fascism. Some influential perspectives came from self-avowed Marxists and, though incorporating views about the social base of these movements, attributed to them a particular role in the balance of class forces. Perhaps, the more subtle of these Marxist perspectives were those by Gramsci, who used a concept of Caesarism, and by Trotsky and Thalheimer, who drew Bonapartist analogies and regarded fascism as a resort by capitalists to perpetuate their interests in a situation of class stalemate between the proletariat and capital. Postwar Marxists have sometimes resurrected such perspectives in looking at Italian fascism and Nazism and have even sought to apply them to postwar movements, but, to most analysts, these approaches are no longer relevant. Instead, the great majority of political scientists writing about the 'old wave,' 'new wave' and present-day extreme right have sought, albeit in varying ways, to focus on the social situations that may produce extreme-right sympathy and behavior.

A widespread feature of numerous theories of both the 'old wave' and 'new wave' extreme right is to see them as supported by those who in some way have suffered, or see themselves to have suffered, from social change. This theorizing may take the form of predicating status loss (e.g., Hofstadter and Lipset on McCarthyism in Bell, 1963/1964), or disorientation as a result of social strain (e.g., Parsons also in Bell, 1963/1964), or a reaction against the growth of postmaterialist values (e.g., Ignazi, 2003), or material loss due to processes of modernization or globalization (e.g., Heitmeyer et al., 1992; Scheuch and Klingemann, 1967). Theories with a social-psychological basis, as by those who have applied perspectives derived from *The Authoritarian Personality* and the Frankfurt School to 'old wave' and 'new wave' parties, also assume a resistance to, and intolerance of, certain types of social change. 'New wave' extreme-right parties since the 1970s, because in many cases they have made racism and xenophobia the principal basis of their appeal, have particularly attracted theorizing seeking to explain the presence of such attitudes among individual voters and in certain groups of voters.

Writers on this subject in the more recent period, while not neglecting these theories of the circumstances of individual susceptibility to the appeal of the extreme right, have rather shifted the explanatory focus more to what the literature calls 'opportunity structures' – the contextual situations, whether political or social, that foment or assist the rise and persistence of extreme-right phenomena. That has certainly been the emphasis in the literature on the most controversial and publicized examples of the recent extreme right, whether the most militant cases of the genre like *Jobbik* in Hungary and GD in Greece or the 'softer,' arguably populist, ones such as the PvdV in The Netherlands. Extreme-right parties emerge and – a separate issue – succeed in sustaining themselves only in certain particular conditions. The analytic consensus is that there has to be an electorate attracted by the appeal made by the party, usually on some one particular issue on which feelings are strong (such as hostility to multiculturalism); that there is widespread disillusion with established parties, the cause of which may be a range of issues or may be their failure successfully to address in the eyes of those attracted to the extreme-right party the concerns raised by the issue; that electors so affected feel powerless and neglected as a result; and, critically important, that there is a party leadership seen to have credibility and competence to promote the cause concerned. All these dynamics may also contain the key to whether the party can sustain itself or whether its momentum will falter. If their chosen party disappoints because little of the discontent that sustained its rise is seen to have been assuaged, or if one or more of the established parties successfully convinces one-time supporters of the extreme-right party that the established party has adopted policies to address their concerns, the success of the extreme-right party can be short-lived. If the extreme-right party's leadership loses the aura of competence or is revealed as corrupt, or if the extreme-right party descends into factionalism, its support is equally likely to decline. There are numerous historical examples of these processes. The success of Britain's National Front of the 1970s was thwarted by the take-over by the Conservative Party of a sanitized dilution of its appeal and then later by its internal schism. The German REPs in the 1990s lost support as the party descended into schism. The Dutch Centrumdemocraten, also in the 1990s, undoubtedly suffered from leadership deficiencies. On the other hand, a leadership has sometimes been sufficiently convincing to sustain the party, despite attempts by other parties to coopt the issues that concern it. Again, there are numerous historical examples: the FN in France and, more recently, the PvdV in The Netherlands.

Social Bases of Extreme-Right Parties' Support

Even within the undoubtedly generally eclectic support base of the prewar extreme right there were concentrations in specific circumstances, and the same is true of extreme-right phenomena of the more recent period, though political scientists have nonetheless offered generalizations about the social base of such parties. Lubbers et al. (2002) gave a comprehensive pan-west-European analysis of extreme-right support – albeit including some parties that would be regarded by Mudde and others as merely radical-right populist rather than extreme right – and from this certain general conclusions about such support in the mid-1990s can be drawn: individual such supporters tended to be male, younger, less educated and, in terms of employment status, manual, self-employed, or unemployed – findings that were consistent with research in the 1980s about such examples as the FN and the German DVU. Norris (2005)' analysis – including non-European parties

and parties not extreme right on the exclusionary definition – gave party-specific breakdowns and showed that an analysis such as the one of Lubbers and his colleagues, which did not differentiate between individual parties could lose some details of difference. The male preponderance, however, is one finding that is almost universal and extends to activists as well as voters. However, there are examples, such as the Austrian FPÖ and the New Zealand First party, whose support in the 1990s was especially from older rather than younger voters. Not all extreme-right parties are supported disproportionately by the less educated, the Israeli extreme right being one of the exceptions, as more recently has been electoral support for *Jobbik*. The class base of the extreme right is by no means uniform – in some countries the disproportionate contribution of the self-employed is significant but this is not universal, and nor is disproportionate support from manual workers or the unemployed. Hungary and Israel stand out as examples of the extreme right's sometimes appealing particularly to more affluent strata. One problem with these generalizing approaches using many examples is that they are largely fixed to a single point in time and are less suited to detect evolutions in such support. For example, the center of gravity of support for both the FN and the FPÖ evolved during the 1980s and 1990s from nonmanual to manual voters, as these countries' social-democratic parties' traditional hold on the support of manual workers slipped; the FPÖ also became disproportionately attractive to much older voters as its support increased during the 1990s.

Any niceties of variation in social base assume a lesser explanatory importance from the repeated findings that, irrespective of individual social characteristics, it is attitudinal dispositions that are by far the better predictors of extreme-right support: such attitudes are typically racist or antiimmigrant/antiasylum-seeker opinions or antisystem attitudes toward the mainstream parties. However, even in the first case there are exceptions – attitudes to immigration do not particularly distinguish FPÖ supporters from voters for other Austrian parties and, among the some of the east-European examples, attitudes to immigration are less significant than those toward indigenous ethnic or linguistic minorities.

Alternative Forms of Extreme-Right Movement

Although this article is about extreme-right political parties, it would leave a wrong emphasis not briefly to note that there are other types of extreme-right activity. Many countries have seen forms of the extreme right that have not participated significantly in formal electoral politics. In many countries there have been demonstrations and marches designed by the most committed activists of the extreme right to demonstrate their 'street presence.' For example, until 2005 the small German town of Wunsiedel in northeast Bavaria suffered an annual Rudolf Hess Memorial March by neo-Nazis, this being the town where Hess' remains were then buried. Other extreme-right movements have their own 'pilgrimage' locations. The annual IJzerbedevaart (Pilgrimage of the Yser) in the Flemish city of Diksmuide, intended to celebrate peace and to focus on Flemish autonomy, attracted the unwelcome attention of European neo-Nazis for a number of years until they were banned; the attraction to them was that the site had been the location of collaborationist ceremonies during World War II. The English Defence League (EDL), claiming to be merely anti-Islamist, may be cited as a further example of an alternative form of extreme-right movement. Except for a desultory link in 2011 to a very small British extreme-right party, the now-defunct British Freedom Party, it has eschewed formal political involvement in favor of media appearances by its leaders and an on-the-street militancy using marches in areas with high levels of Muslim population.

Since at least the 1970s, there have been yet more forms of extreme-right activity. The 'skinhead' phenomenon has been associated with the extreme right in a number of countries, including France and Germany. Related, and also significant for the culture of parts of the extreme right, has been a White Power music scene as an activity that from early on was able to transcend national frontiers. Another arena has been a certain type of support of particular football clubs in several countries, including Great Britain, Germany, Italy, and The Netherlands; in Italy the clubs AS Roma and SS Lazio are especially notorious. The most extreme manifestations of this behavior – anti-Semitic and racist chanting and giving the Nazi salute – have been largely stamped by the stewarding authorities in northern Europe, but they remain an issue among certain clubs in southern and eastern Europe, where they sometimes also emerge in support for national teams.

A yet further forum for extreme-right activity and mobilization has been the Internet, which has also facilitated the easy transcendence of national borders by the extreme right and, for example, brought a range of extreme-right American material, protected by the First Amendment in the United States, within easy reach of European extreme-right supporters. Indeed, the popularity of the Internet for the dissemination of extreme-right material and its consumption by an unspecified number of users raise the question whether the customary method used by social scientists to measure the strength of extreme-right support – numbers of votes in elections and numbers of identified activists in known extreme-right organizations – is now outmoded. The number of lone but Internet-connected extreme-right supporters – individually exemplified most dramatically by the case of the Norwegian terrorist and murderer Anders Behring Breivik – is undoubtedly far greater in some countries than the numbers of voters and activists.

The Tendency of Extreme-Right Parties to Schism

Extreme-right parties frequently exhibit distinctive organizational characteristics in comparison with mainstream political parties. The former have a particular tendency to split, and contain strong schismogenetic features. This is sometimes because of personal animosities between leaders, who are often psychologically disposed to be authoritarian and are intolerant both of different perspectives and also of rivals. Because of the variety of ideological appeals offered by extreme-right parties, their leaderships and supporters often have a mixture of social origins and this too can lead to divisive differences over minutiae of ideology (a characteristic shared with extreme-left movements). Extreme-right parties, particularly those in the

postwar 'new wave' period, have exhibited a remarkable tendency to fracture. In France, Germany, Great Britain, The Netherlands, and Switzerland, to name only some examples, the indigenous extreme right has had a history of division and internal dissension. Indeed, many of the apparent electoral successes in the 1980s and 1990s by extreme-right parties foundered on the organizational collapse of the elected legislative grouping, something that was a particular feature of the German extreme-right parties, the REPs and the DVU. The British BNP – apparently on the crest of a breakthrough at the beginning of the 2000s in some locations in northern England where it appealed to local anti-Muslim sentiments – became mired in internal dissensions and suffered a loss of credibility.

The Future of Extreme-Right Parties

Even 10 years ago it was possible, viewed from a democratic pluralist perspective, to be relatively sanguine about the future of the extreme right. In few countries was it a well-supported phenomenon. Even in countries where its support was stronger, such as France or Flanders, the political system had excluded it from governmental participation, at least at the national level, although in the 1990s the FN, for example, had a few successes in the south of France in obtaining municipal control and extreme-right parties in some other countries had similarly enjoyed some local government successes. However, the FN generally glorified in its pariah status; a faction wanting to seek accommodations with mainstream right-wing parties had caused a damaging split in 1998, but it was the breakaway *Mouvement National Républicain* of Bruno Mégret wanting such a rapprochement that failed electorally against the FN when their respective popularities were put to an electoral test. True, the FPÖ in Austria had been part of a national coalition government for a short period, the Slovak SNS had been small part of a coalition government in the 1990s, and the Swiss SVP, part of the multiparty coalition that had governed Switzerland since 1959, was turning in some cantons increasingly into an extreme-right party and increased its support and national influence in doing so. However, the accession to power of the FPÖ had led to revulsion and boycott by some other countries in the European Community (as then called) and had significantly damaged the country's international reputation. Even the SNS had, and still has, trouble reaching the country's five-percent electoral threshold. Parties in other countries, such as Australia's One Nation party, had burned briefly but then guttered and died – 'like bees: once they have stung, they die,' in Hofstadter's memorable and much-quoted aphorism about the American People's Party.

A decade and more later, the state of the extreme right is rather different. The dynamics that became important in the period after 2000 show little sign of diminishing in intensity. Rising and persistent unemployment in many countries in Europe and the Americas, combined with the likelihood of long-term economic recession or depression, offers the breeding ground for extreme-right mobilization. The decline, continuing a process that in many countries can be traced back to the 1970s, in committed mainstream political partisanship, especially of working-class attachments to social-democratic parties; the general rise of electoral volatility; the evidence in many countries of increasing disenchantment with mainstream politicians and political parties: all these are factors that provide a pool of less attached voters who may be attracted to the extreme right. A pool of such voters has always existed in most countries, as has an underlying, if otherwise dormant, propensity to some form of extreme-right mobilization. The issue for the likelihood of future success for extreme-right politics is whether that pool is becoming critically larger.

The extreme right has had various levels of success in securing parliamentary representation and has been assisted in this in some countries by a wholly or partly proportional franchise system. In an assembly such as the European Parliament its presence has had little impact. However, in The Netherlands the PvdV and in Denmark the *Dansk Folkeparti* (Danish People's Party) achieved significant policy influence by their support for minority center-right coalition governments. In Hungary, where the mainstream right party *Fidesz* (*Fidesz – Magyar Polgári Szövetség*, Hungarian Civic Union) governs with a smaller right-wing party and has a comfortable parliamentary majority, it has nonetheless been influenced on several policy positions by the presence of *Jobbik* to its right on the political spectrum. Even majoritarian electoral systems have not wholly frustrated the extreme right; in France, where in the 2002 Presidential election the success of Jean-Marie Le Pen's first-round 16.9% of votes was somewhat soured for him by more than 82% of voters opposing him in the second round, the FN succeeded in securing seats in regional elections, though its success in more local elections was otherwise patchy. Increasing support for extreme-right parties is not inexorable; electoral support for all the longer-lasting examples has ebbed and flowed according to the vicissitudes of leadership, policy and credibility, though it is parties such as the FN and VB with more capable leaderships and better organizations that have been deft enough to survive periodic downturns of fortune.

Perhaps the greatest longer term dynamic of the extreme right in some Western countries is Islamophobia, which has fueled support for new extreme-right parties such as those in The Netherlands, has provided an issue on to which older extreme-right parties such as the French FN have latched, and has provided the pretext for street-level militancy in a number of countries (the EDL in Great Britain, for example). As long as there is a view, even a minority one among a small number of indigenous Muslims, that the presence of some Western countries in the Islamic world is illegitimate, Islamist reactions to this will be a potent source of Islamophobia in those countries and then of extreme-right benefits from this.

The issues are rather different, however, for the extreme right in eastern and southern Europe; such parties have so far been qualitatively different from those elsewhere. The trajectory of these extreme-right parties is unclear, partly because their dynamics are contested and not all countries in this region have equally significant such parties. Some parties have ultranationalist and often irredentist concerns, having emerged from long-held resentments that arose between 1919 and 1949, and are therefore seen by some as a type of revival of the interbellum extreme right. Others are shaped by the structures and attitudes induced by Leninist state socialism between 1949 and 1989. Others were simply products of the chaos after the post-1989 dissolution of the Soviet Bloc and may increasingly come to resemble west-European extreme-right parties. Quite

how each will evolve, and so what danger to democracy it will pose, is a quandary.

See also: Attitudes, Political: Authoritarianism and Tolerance; Immigration: Political Aspects; Islamic Fundamentalism; Populism.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.

Bell, D. (Ed.), 1963/1964. The Radical Right: The New American Right Expanded and Updated. Anchor, Garden City, NY.

Betz, H.-G., 1994. Radical Right-Wing Populism in Western Europe. Macmillan, Basingstoke, UK.

Betz, H.-G., Immerfall, S. (Eds.), 1998. The New Politics of the Right: Neo-populist Parties and Movements in Established Democracies. St. Martin's Press, New York.

Cheles, L., Ferguson, R., Vaughan, M. (Eds.), 1995. The Far Right in Western and Eastern Europe, second ed. Longman, London and New York.

Eatwell, R., 1995/2003. Fascism: A History. Pimlico, London.

Greß, F., Jaschke, H.-G., Schönekäs, K., 1990. Neue Rechte und Rechtsextremismus in Europa: Bundesrepublik, Frankreich, Großbritannien. Westdeutscher Verlag, Opladen.

Griffin, R. (Ed.), 1995. Fascism. Oxford University Press, Oxford, UK.

Hainsworth, P. (Ed.), 2000. The Politics of the Extreme Right: From the Margins to the Mainstream. Pinter, London.

Heitmeyer, W., Bushse, H., Liebe-Freund, J., Möller, K., Müller, J., Ritz, H., Siller, G., Vossen, J., 1992. Die Bielefelder Rechtsextremismus-Studie: Erste Langzeituntersuchung zur politischen Sozialisation männlicher Jugendlicher. Juventa, Weinheim, Germany.

Ignazi, P., 2003. Extreme Right Parties in Western Europe. Oxford University Press, Oxford, UK.

Jaschke, H.-G., 1994. Rechtsextremismus und Fremdenfeindlichkeit: Begriffe, Positionen, Praxisfelder. Westdeutscher Verlag, Opladen.

Kitschelt, H., McGann, A.J., 1995. The Radical Right in Western Europe: A Comparative Analysis. University of Michigan Press, Ann Arbor, MI.

Lipset, S.M., 1959a/1963. Political Man: The Social Bases of Politics. Anchor, Garden City, NY.

Lipset, S.M., 1959b. Social stratification and 'right-wing extremism'. British Journal of Sociology 10 (4), 346–382.

Lipset, S.M., Raab, E., 1970/1971. The Politics of Unreason: Right-Wing Extremism in America, 1790–1970. Heinemann, London.

Lubbers, M., Gijsbert, M., Scheepers, P., 2002. Extreme right-wing voting in Western Europe. European Journal of Political Research 41 (3), 345–378.

Mammone, A., Godin, E., Jenkins, B. (Eds.), 2013. Varieties of Right-Wing Extremism in Europe. Routledge, London & New York.

Melzer, R., Serafin, S. (Eds.), 2013. Right-Wing Extremism in Europe: Country Analyses, Counter-strategies and Labor-Market Oriented Exit Strategies. Friedrich-Ebert-Stiftung, Berlin.

Minkenberg, M., 1998. Die neue radikale Rechte im Vergleich: USA, Frankreich, Deutschland. Westdeutscher Verlag, Opladen & Wiesbaden.

Minkenberg, M., 2009. The radical right in Europe: challenges for comparative research. In: Stiftung, Bertelsmann (Ed.), Strategies for Combating Right-Wing Extremism in Europe. Verlag Bertelsmann Stiftung, Gütersloh, pp. 13–28.

Minkenberg, M. (Ed.), 2010. Historical Legacies and the Radical Right in Post-Cold War Central and Eastern Europe. Ibidem-Verlag, Stuttgart.

Mudde, C., 2007. Populist Radical Right Parties in Europe. Cambridge University Press, Cambridge, UK.

Norris, P., 2005. Radical Right: Voters and Parties in the Electoral Market. Cambridge University Press, Cambridge, UK.

Ramet, S.P. (Ed.), 1999. The Radical Right in Central and Eastern Europe since 1989. The Pennsylvania State University Press, University Park, PA.

Stiftung, Bertelsmann, 2009. In: Strategies for Combating Right-Wing Extremism in Europe. Verlag Bertelsmann Stiftung, Gütersloh.

Schain, P., Zolberg, A., Hossay, P. (Eds.), 2002. Shadows over Europe: The Development and Impact of the Extreme Right in Western Europe. Palgrave Macmillan, New York.

Scheuch, E.K., Klingemann, H.D., 1967. Theorie des Rechtsradikalismus in westlichen Industriegesellschaften. In: Ortlieb, H.-D., Molitor, B. (Eds.), Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik. J. C. B. Mohr, Tübingen, pp. 11–29.

Political Parties: Manifestoes

Ian Budge, University of Essex, Colchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Manifestoes are defined and described physically. Their unique standing as the only authorized expression of collective party preferences gives them a central place in guiding policy voting and effecting the party mandate. Their strategic role in election campaigns causes them to be written in a particular way: stressing party concern with their own issues and downplaying rivals' issues, and avoiding specific policy stands. Nevertheless, manifestoes indicate what priorities parties will pursue if they get into government, and in this sense form a unique link between electors' preferences and government actions.

Introduction

Manifestoes or platforms are the programs usually issued by parties in campaigning for elections. They give an analysis of current problems and future developments endorsed – uniquely – by the party as a whole through its formal procedures. As such they are crucial to the operation of the democratic mandate: the idea that parties are chosen by electors for government on the basis of the competing policies they present. Only in the manifestoes can such competing policies be found, since the manifesto is generally the only policy statement endorsed officially by the party as a whole – in the election or indeed at any other time. The unique standing of the manifesto as a statement of official party policy renders it to be useful practically, as a guide to the media and (indirectly) electors, about what the party stands for; and also crucial analytically, as the best indicator of what the party's collective preferences are. It is, however, a coded statement designed to win over votes. Hence its interpretation has to be based on assumptions about how parties compete for votes, which are described below.

The Nature of the Manifesto

In the UK, United States, and some continental European countries, manifestoes are published by the party as booklets at the official start of the election campaign. There have been clear physical changes in their appearance and length during the postwar period. They have become markedly longer and noticeably glossier. Starting out as badly printed, unillustrated pamphlets immediately after World War II, they have in some later cases approached the size of books. A more recent tendency is to have a plethora of glossy photographs, colored charts, and figures to reinforce and break up the text, which itself is headed by and interspersed with sound bites.

Elsewhere, practices in regard to the publication of election programs differ. In France, programs have sometimes been published commercially as books, which have attained the best-seller list! In Scandinavia and Canada, they have at times been split as separate pamphlets addressed to women, youth, workers, and other demographic groups. In Japan, they are widely diffused as summaries of interviews with the party secretaries, published in the leading mass circulation newspapers. In Australia and New Zealand, they do not even get into print, but are substituted by an hour-long television address from the party leader – the content of which is organized along conventional lines, and so covers much the same range of topics broken into the same kinds of sections as in the United States and UK.

These variations may reflect a growing realization that most electors do not read party programs, even if they are pushed through their door. Instead, they get an impression of them from media discussion. Media presenters, on the other hand, read programs very carefully, using them to start debates and confront party spokespeople if they try to pull back from concerns expressed in the document. In this sense, the manifesto constitutes a policy equilibrium point for the party which is set at the outset of the campaign (Budge, 1994).

Reading the document casually, however, one might be hard pressed to discover much policy content at all. Until recently, specific promises have been limited and confined mostly to pledges on peripheral problems where the party knew it could deliver with some certainty if it became (part of) the government (Rallings, 1987). Most of the text discusses general trends and developments under such headings as Youth, Unemployment, the Economy, and so forth. Depending on whether the party is in government or opposition, it describes the situation in optimistic or pessimistic terms, sometimes going into the historical achievements of the party or stressing the importance of the area.

Rarely, however, is the text more specific than this. The manifestoes of competing parties do nonetheless succeed in presenting quite contrasting diagnoses of the current situation – one painting a glowing picture of free enterprise surging ahead amid general prosperity, another pointing to growing social inequalities and the pressing need for government intervention. Without further information, one could not even be sure if they appeared at the same election or at different ones!

Electoral Competition and the Manifesto

The reason for structuring election programs in this way is to be found in the view party strategists take of competition for votes. In line with Riker's (1993: p. 105) 'dominance principle,' electors are seen as overwhelmingly endorsing one 'obvious'

line of policy in each area: cutting taxes, strengthening law enforcement, extending social provision, and so forth. Particular parties, because of their record and ideology, are associated with each favored line of policy. If you want to cut taxes you will not vote Socialist but you will if you want to extend welfare.

For the parties, competition thus consists in getting 'their' issues – the ones on which they are associated with the popular line of action – on to the agenda. To do this, they need to emphasize the importance of such issues compared with those which favor rivals. Their manifesto advances this process by talking a lot about 'their' party issues, ignoring areas 'belonging' to rival parties. As the task is to convince media and electors of these issues' importance, the manifestoes can ignore fine details of policy in favor of general discussion – a tactic which, conveniently, does not tie parties down to specific actions when in government.

The view of party competition, which emerges from analyses of manifestoes (Robertson, 1976: Budge and Farlie, 1983: Petrocik, 1996) is thus one in which relative emphases on different issues substitute for direct confrontation over differing policies on the same issue. This is not to say, however, that parties are not putting forward different programs of government. Parties that talk a lot about taxes mean to cut them and parties discussing welfare mean to extend it. Manifestoes are about priorities between tax, welfare, and other areas, even if the other side to the debate – cutting services or increasing taxes – goes unstated.

Using Manifestoes in Political Research

The way manifestoes are written makes them relatively easy to analyze quantitatively. By counting sentences or words into different policy areas and percentaging them out of the total to standardize for length, one can contrast left-leaning and right-leaning parties and even characterize their attitudes within different policy areas (Klingemann et al., 2006). Word counts of electronic texts can be performed easily by computer programs, opening up new possibilities of research once the techniques are better validated. Sentence counts have been carried out manually for all significant parties in 51 countries over the postwar period, from the time of the first democratic election (Budge et al., 2001; Volkens et al., 2013).

Given the manifestoes' standing as the sole authorized statement of party collective policy preferences, such analyses can give an exact picture of parties' ideological progression and shifting policy stances over the postwar period. This is illustrated in **Figure 1**, which traces the US Democrats' and Republicans' left–right movement over the postwar period. (Left–right is measured by adding all percentage mentions of 'rightist' issues and subtracting from this the total percentage mention of 'left-wing' issues (Budge et al., 2001: pp. 21–24).)

Figure 1 shows first of all that Democrats and Republicans take up the standard left and right positions assumed by Socialist and Conservative parties in other countries. The validity of interpreting US party competition in left–right terms is supported:

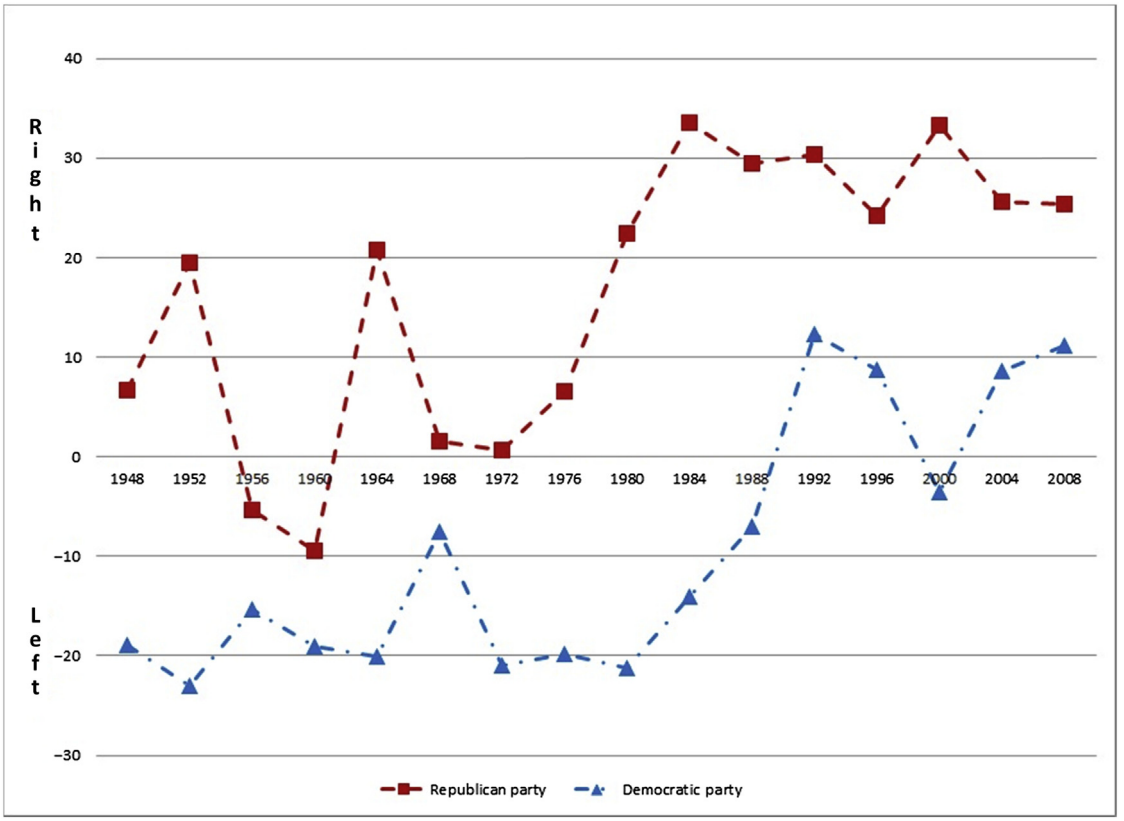

Figure 1 Right–left movement of American parties 1945–2008 as shown in their platforms.

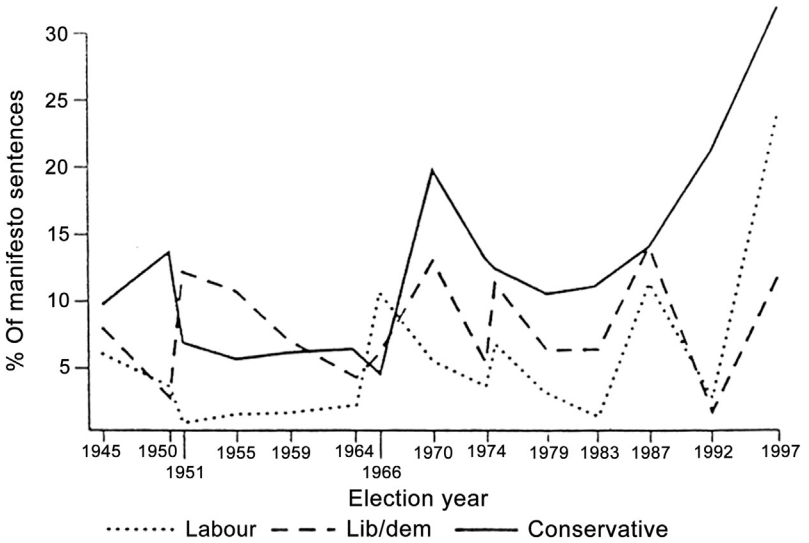

Figure 2 British parties' support for social conservatism, estimated from manifestoes 1945–97.

1. by the way in which contrasts between the parties mirror the historical record – for example, far apart at the 'Goldwater' election of 1964, closer in 1992 with Clinton's famous move to the right;
2. by the high correlation between presidential platform positions and the relative 'liberalism' of subsequent presidential policy (McDonald et al., 1999). (This finding incidentally illustrates the usefulness of comparing manifesto commitments with actual policy as a method of checking out mandate theory (McDonald and Budge, 2005: 107).)

Figure 1 also serves as a check on Downsian theories of party competition, which predict policy convergence between parties as both seek to position themselves near the median elector (Downs, 1957: pp. 110–117). Clearly this expectation is not fulfilled as the US parties always maintain their ideological distance from each other, even where they converge to some extent, as in 1992. Their patterns of movement are explained better by the 'saliency' theory, where parties are stuck with their traditional issue positions and can act strategically only by varying the emphasis they put on them (Budge et al., 2010).

Such changes can be monitored for the separate issue areas by seeing what percentage of sentences are devoted to them in the manifesto at different elections. Figure 2 traces British parties' concerns with traditional morality and law and order ('social conservatism') from 1945 to 1997. The Conservative Party generally appropriates this area, where it presents itself as the defender of 'family values.'

Linking Manifestoes to Election Change

As well as supplementing historical analyses of party change, such 'maps' of party movement can also be related to voting behavior. Instead of estimating party positions indirectly by asking electors about them, we can actually study the effect of a party's strategic adjustments on electors' perceptions and subsequent voting decisions. This extends the possibilities of linking such phenomena as 'revisionism' to electoral reactions, to see if it actually has the effects claimed for it (Evans and Norris, 1999: pp. 87–101). Methodologically, one can marry textual analysis to survey research to deepen the insight provided by both.

Linking Manifestoes to Government Action

Manifestoes mediate between electors' preferences – expressed by voting for the closest party – and government actions as mirrored in expenditure or legislation. As well as investigating what impact they have on electors, the priorities in the manifesto can be related to government policy outputs. Research has shown them to be quite closely linked (Budge et al., 2012: pp. 178–193) thus upholding 'mandate' ideas that the party or parties in government have been popularly authorized to carry through their program and indeed have a responsibility to carry it through.

This moral commitment is reinforced by two other characteristics of manifestoes, which give them an impact in government.

1. Manifestoes reaffirm party ideology, applying it to specific policy areas. Thus most party representatives, who belong to a party in the first place because they subscribe to its principles, actually want to carry out manifesto commitments regardless of electoral considerations.
2. The manifesto is often the only worked out program the party has for government. Even where coalition partners make a policy agreement about what to do, the separate party manifestoes feed into it. For practical reasons of coordination and planning in government therefore, the manifesto is an indispensable document.

Overall Role and Significance of Manifestoes

Cynicism about the extent to which parties keep their promises, and hence about the significance of the manifesto which contains them is easy. The election program has often been seen as a 'mere piece of paper' forgotten immediately the party gets into government. The research cited above, however,

shows that the manifesto deserves to be taken more seriously, both by voters and political analysts. Parties are responsible, both in maintaining consistency with previous policy and in carrying it out in government. This would hardly be so, if they just shifted policies around in response to electoral expediency. What Figures 1 and 2 show is considerable stability in the positions taken by the only policy document which parties issue collectively. This renders it an invaluable source for studying party intentions and preferences, and a good guide to what they will do in government.

See also: Conservatism: Theory and Contemporary Political Ideology; Party Identification and Nonpartisanship.

Bibliography

Budge, I., Klingemann, H.-D., Volkens, A., Bara, J., Tanenbaum, E., 2001. Mapping Policy Preferences: Estimates for Parties, Voters and Governments 1945–2008. Oxford University Press, Oxford.

Budge, I., 1994. A new spatial theory of party competition. British Journal of Political Science 24, 443–467.

Budge, I., Ezrow, L., McDonald, M.D.M., 2010. Ideology party factionalism and policy change. British Journal of Political Science 40, 781–804.

Budge, I., Farlie, D.J., 1983. Explaining and Predicting Elections. Routledge, London.

Budge, I., Keman, H., McDonald, J.D., Pennings, P., 2012. Organizing Democratic Choice: Party Representation over Time. Oxford University Press.

Budge, I., Robertson, D., Hearl, D. (Eds.), 1987. Ideology, Strategy and Party Change. Cambridge University Press, Cambridge, UK.

Downs, A., 1957. An Economic Theory of Democracy. Harper, New York.

Evans, G., Norris, P., 1999. Critical Elections. Sage, London.

Klingemann, H.-D., Volkens, A., Bara, J., Budge, I., McDonald, M.D., 2006. Mapping Policy Preferences II. Oxford University Press, Oxford.

McDonald, M.D., Budge, I., Hofferbert, R.I., 1999. Party mandate theory and time series analysis. Electoral Studies 18, 587–596.

McDonald, M.D., Budge, I., 2005. Elections Parties Democracies: Confirming the Median Mandate. Oxford University Press, Oxford.

Petrocik, J.R., 1996. Issue ownership in presidential elections. American Journal of Political Science 40 (3), 825–850.

Rallings, C., 1987. In: Budge, Robertson, Hearl (Eds.), The Influence of Election Programmes, pp. 1–14.

Riker, W.H., 1993. Agenda Formation. University of Michigan Press, Ann Arbor, MI.

Robertson, D., 1976. A Theory of Party Competition. Wiley, London.

Volkens, A., Bara, J., Budge, I., McDonald, M.D., Klingemann, H.-D. (Eds.), 2013. Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts. Oxford University Press, Oxford.

Political Protest and Civil Disobedience

Dennis Chong, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political protest refers to a multitude of methods used by individuals and groups within a political system to express their dissatisfaction with the status quo. Although protest can be registered through voting, letter writing, campaign contributions, participation in interest groups and political parties, and other traditional means, the term more commonly refers to those kinds of mass political activism that occur outside of conventional parliamentary channels, such as demonstrations and marches, labor stoppages, boycotts, and other methods that combine communication and persuasion, direct action, and noncooperation with political, economic, and social institutions.

Civil disobedience is a particular form of political protest that is marked by the deliberate violation of the law with the express goal of correcting a perceived injustice. Action that would now be regarded as civil disobedience has been a regular feature of social protest throughout history even though the term itself, attributed to an 1849 essay by Henry David Thoreau, is a relatively recent creation.

Political protest refers to a multitude of methods used by individuals and groups within a political system to express their dissatisfaction with the status quo. Although protest can be registered through voting, letter writing, campaign contributions, participation in interest groups and political parties, and other traditional means, the term more commonly refers to those kinds of mass political activism that occur outside of conventional parliamentary channels, such as demonstrations and marches, sit-ins, blockades, labor stoppages, boycotts, and other methods that combine communication and persuasion, direct action, and noncooperation with political, economic, and social institutions.

Much social science research on political protest has concentrated on explaining organized, sustained forms of mass collective action such as social movements. A social movement consists of an array of organizations and individuals who share a common identity and agenda of social and political goals, and who are regarded as a collective entity by other political actors. Social movements typically pursue their goals by mixing protest actions with conventional political strategies, such as lobbying, litigation, public education, and electoral mobilization. Discontented groups often resort to direct action when they are denied equal access to conventional means of participation or when such avenues are regarded to be ineffectual or inadequate for registering the intensity of their grievances.

Protest as an Assurance Game

Mass protest movements often can be represented by a type of strategic interaction known in game theory as an 'assurance game,' in which individual decisions to contribute are contingent on the mutual cooperation of a sufficient number of people. Supporters of a movement therefore try to coordinate their decisions with those of other activists, participating when it appears that the movement has enough popular support to be politically effective, but not otherwise. Such decisions, however, typically have to be made with considerable uncertainty about both the intentions of other individuals and the prospects of the movement as it develops.

Consider how these sources of uncertainty alter the calculations of activists. Each individual has to determine how the costs and benefits of participation vary according to the aggregate level of participation. Costs include the opportunity costs and risks of political activism, while benefits include both collective benefits (e.g., policy changes) that go to all supporters of the movement whether or not they participate, and selective benefits (e.g., material rewards, social esteem) that go only to those who participate. Let us assume that the benefits of participation increase, while the costs decrease, with the size, N, of the movement. Therefore, at low levels of N, the movement generates few, if any, benefits and is extremely costly for individual contributors, making abstention the preferred alternative. If the movement is able to attract more widespread support, several favorable developments occur: its prospects for success improve, selective incentives become more salient, and the costs of participation decline with the growing safety of numbers. In accordance with the assurance game, I assume that there is some threshold level of participation, $N = n'$, above which cooperation provides net benefits.

Estimates of the threshold value for participation will vary among group members depending on their assessment of the prospects and benefits of collective action, and estimates of the probability that there will be a sufficient number of cooperators will also be unsettled. Obviously, much depends on the responsiveness of the government or other petitioned party to the demands of the movement. Someone who thinks that collective action has little chance of succeeding will set very high levels for n'; in the most pessimistic case, n' will exceed the size of the group seeking change – in other words, even total cooperation will be assumed to be futile. To further complicate matters, not only will an individual be uncertain in his or her own mind about these parameters, but he or she will also know that others are uncertain about them too, and that everyone is aware of each other's uncertainty. Given this individual calculus, a number of inferences can be drawn about the

resources, strategies, goals, and political conditions that will be required to coordinate and organize mass social protest.

Leadership

Prior to organizing, people suffer diffuse grievances and contemplate various means to address them. Leadership is therefore necessary to develop a common understanding of the problems people face and to provide convincing arguments in favor of collective action. All analysts of social movements point to the importance of having an ideology that interprets people's circumstances and offers prescriptions for action. The task of leadership is to use such ideological appeals to create a group identity among those who are discontented and to train the energies of group members on a set of targets and solutions to their problems. Leaders therefore help to define the assurance game for others by persuading them of the promise entailed in collective action. In addition to defining the game for others, leaders help mightily to solve it by constituting, in effect, the critical mass that instigates the growth of collective action.

Community

Knowledge of the calculations and intentions of fellow group members is, of course, alleviated when collective action is built out of existing organizations and community relationships. The sociological literature is rich with examples of social movements capitalizing on existing social networks of individuals with common interests and goals.

Groups in general are more readily mobilized than individuals. Individuals in groups – bonded by a distinct ethnicity, race, or social status; ties of friendship and kinship; and common membership in religious, fraternal, social, and political clubs, associations, and organizations – have greater solidarity and are in closer communication than groups that are not constrained by similar functional and social linkages. Solidarity and affection, in turn, go hand in hand with the establishment of personal obligations and exchange relationships. Members of these preexisting networks have reputations to protect, respect, and affection for each other, and reliable information about the preferences and probable behavior of others.

Mass Communications

Modern communications technology can compensate for weak grassroots organizations and expand the popular base of a social movement by facilitating the rapid transmission of information. While small groups can be coordinated through meetings and face-to-face interaction, the coordination of large masses of people in disparate locations depends on central sources of information transmission. Through the electronic media, people can learn about the plans of the movement to initiate new campaigns, the size and impact of demonstrations in distant cities, how authorities are responding to collective action, whether third parties are sympathetic or hostile to the struggle, and other relevant information that updates their evaluation of the movement's prospects. Various means of communications not only inform interested parties, but also convey information that the recipient will interpret as common knowledge based on his or her assumption that access to this information is widely shared.

A key factor behind the massive size of the 1989 upheaval in Eastern Europe is that people were extremely well informed and aware of political developments in their own country and abroad. With the aid of modern telecommunications, the opposition movements were able to coordinate strikes, demonstrations, and other activities. For example, one of the most effective ways that Solidarity leaders in Poland transmited information was to provide foreign radio networks with details about their strike activities, which would then be broadcast in news reports throughout the country. When the Czechoslovak opposition gained access to the radio and television waves, it was able to provide updates on the progress of the movement to parts of the country beyond Prague. Moreover, through print, radio, and television, people throughout Eastern Europe were able to stay apprised of developments in the West and in the Soviet Union. It was easy to learn from the media about Gorbachev's support for the upheaval in Eastern Europe, his plans for reform in the Soviet Union, and the consequences of those reforms, namely, competitive elections and the development of an independent national legislature.

Tacit Coordination Using Cultural Resources

Overt leadership solves the coordination problem; but leadership can often be identified and eliminated by the authorities trying to prevent collective action. To reduce the need for explicit communication, activists can capitalize on famous locations, significant anniversary dates, focal events such as deaths and funerals, and other cultural signposts when staging mass demonstrations and protests. A striking feature of the revolutions in Eastern Europe was the manner in which the opposition movements took advantage of salient historical precedents to coordinate their activities. Much of the protest activity in Hungary, Poland, and Czechoslovakia in 1989 occurred at the same sites where earlier protests had been held. There were obvious parallels between the processions organized by the opposition in Budapest in 1989 and similar processions staged during the 1848 and 1956 Hungarian Revolutions. In Czechoslovakia, the demonstrators conducted their rallies at the symbolic Wenceslas Square, the historic setting for most of the nation's greatest political events. Similarly, Leipzig's famous tradition for opposing authority, dating from the Protestant Revolution and extending through the Peasants Rebellion, the 1848 revolution, the workers' uprising in 1923, and the sympathy protests for the Prague Spring, may help to explain why the mass demonstrations that brought down the regime in East Germany were initiated in that city before diffusing to the rest of the country. All of these examples illustrate how history and tradition can supplement overt leadership and formal organizations in coordinating mass political activism.

Demand and Supply

Participation in collective action may begin with the belief among individuals that enough pressure can be applied to win

concessions from the opposition. Sooner or later, however, collective action will have to produce real dividends to sustain the motivation of activists. Therefore, in addition to trumpeting the potency of collective action, the leaders of a movement will try to wrest concessions from their adversary as quickly as possible. Doing so establishes a level of proficiency and success that reinforces their followers and makes the movement attractive to others. Notable successes spur a bandwagon of new participants to join based on their revised estimates of the prospects of the movement. In addition, concessions establish the responsiveness of the authorities to political pressure.

The premium that is placed on continual success gives rise to an evolution of tactical strategies over the course of the movement based on what succeeds. Tactics that work in forcing the opposition to supply concessions will become popular, while those that prove ineffective will be abandoned. Each innovative tactic will lose its appeal within a movement once its effectiveness is reduced by the tactical countermaneuvers of the opposition.

Leaders of collective action must therefore devise a continuing series of novel tactics in order to retain their followers and maintain the leverage of the movement. During the American civil rights movement, for example, several innovative tactics, such as the bus boycott, sit-in, and freedom ride, grabbed the spotlight at different times. After a new tactic was introduced, the following months of activity in the civil rights movement tended to be dominated by copycat campaigns in other locales around the country. The popularity of various tactics used over the course of the civil rights movement rose and fell according to their success in gaining media attention, provoking southern authorities, and extracting meaningful concessions.

Political Opportunities

The incentive structure of the assurance game also shows how fluctuations in demands for social change will be linked to the disposition of the agencies capable of supplying such change. A sympathetic government or a government that is expected to be responsive to popular demands will obviously encourage collective action. It seems reasonable, for example, to suppose that extremely repressive countermeasures in the early stages of the demonstrations in Eastern Europe in the late 1980s could have halted the popular movements there. Once compliance with government ideology dissolved at both the private and public levels, the only restraints on mass dissent were the twin threats of internal repression and Soviet intervention. The two methods of repression were not independent, however. State resistance and countermeasures were evident in Poland, East Germany, and Czechoslovakia in the early stages of these movements; but once it became apparent that the Soviet Union would not interfere – except to send encouraging signals about its nonintentions – the option of internal repression also became limited. By the summer of 1989, fears that either form of repression would occur had diminished greatly even in the more hard-line states of Czechoslovakia and East Germany. What is remarkable, nevertheless, is the speed with which the publics in Eastern Europe sized up current events and seized the moment when it became apparent that the Communist system was ready to collapse.

The Power of Social Media

The advent of social media has changed the technology of political protest, accelerating its dynamic in ways that are predictable from social movement theories. Facebook, Twitter, and YouTube have been used extensively to inform and coordinate participants in the 2010–11 revolutions in Tunisia and Egypt, and in subsequent popular protests against socioeconomic inequality in Western nations.

Social media reduce the costs and practical barriers to coordination by facilitating communication that can help to solve collective action problems. The value of social media has been to allow people to discover the extent of discontent in society and the willingness of people to engage in political protest. Online resources can be used to strengthen the development of networks and social movement communities that are built on face-to-face interactions and ongoing relationships.

Social media take on greater importance in authoritarian societies that do not enjoy freedom of the press, speech, or assembly, and where the government closely monitors organized groups and frequently jails its leaders. The availability of new means of communication that can bypass these barriers is a significant development that has weakened the ability of governments to control and contain the diffusion of information. Virtual forums of communication can be safer sites of discussion than in-person gatherings. Citizens can learn of government abuses, inadequate social conditions, and the disaffection felt by other citizens. Social media can therefore both reveal and coordinate widespread discontent, which are essential stages in mobilizing political protest.

The technology also affects the strategies and methods of social protest and government responses to dissent. The ability of citizens to document government reactions to protest can also place constraints on government misconduct, thereby lowering the cost of protest. Mobile phones with picture and video capabilities allow 'citizen journalists' to record and publicize protests and to bypass state-censored coverage. Through content created on social media, political activists are able to bring outside pressure to bear on the governments facing popular unrest. Such international support for demonstrations and protests against unpopular regimes can also strengthen resolve among the demonstrators.

Social Media in the Arab Spring

Political protests in Tunisia in 2010–11 succeeded over a repressive government that maintained tight controls over the media. Online social networks allowed elite activists to circumvent government censorship and build a digital network that became the foundation for the 2010 uprising. Following a dramatic triggering event (the self immolation of a fruit vendor), social media allowed transmission of information to international media, especially Al Jazeera, despite efforts of the government to limit reporting on internal dissent. Al Jazeera had limited capacity to cover events within the country, especially in more remote areas; web activists stepped into the breech and served as journalists by providing video to Al Jazeera showing atrocities committed by government police forces that was then broadcast to Tunisian citizens.

Images and videos circulated on the Internet reduced the legitimacy of the Ben Ali regime in Tunisia. Reports of the abuses of the regime in response to the protests served as a catalyst that broadened the social bases of support for the uprising. Social media were critical in publicizing the size of the protest demonstrations and in creating expectations about the magnitude of future protest gatherings, thus promoting safety in numbers and reducing the fear of protest. Individuals learned about future protest events online and could also anticipate that these events would be well attended based on online responses to announcements of future protests. The information that opponents of the government obtained from the Internet increased their confidence that the Ben Ali regime would be brought down by the protests.

The collapse of the Ben Ali government in the face of mass protests gave inspiration to Egyptians to organize similar protests in early 2011 against their government. The Egyptian Revolution was over quickly – 18 days from start to the removal of President Hosni Mubarak – but the social and organizational foundation of the movement was laid in preceding years. The April 6 youth movement was formed online several years earlier during an effort to organize support for a nationwide strike. The creation of a Facebook page around the April 6 strike led to over 70 000 individuals joining the group. The movement took advantage of a political opportunity created in 2005–06 when the George W. Bush administration pressured Mubarak to conduct multiparty elections. This opening allowed individuals to organize a loose coalition of activists around the goal of preventing Mubarak from handing power to his son. At the same time, the Internet was becoming more readily available in Egypt and became an important tool of political communication that substituted for face-to-face meetings.

Dissenters harnessed the new technologies associated with social media to publicize police brutality and social and economic problems. An emerging group consciousness grew around these grievances, especially among university graduates who suffered from high unemployment rates. Popular protests arose over poor living conditions, inflation, low wages, and state repression. The event that instigated the 2011 protests leading ultimately to the removal of Mubarak was the killing by Egyptian police of a blogger, Khaled Saeed, who had posted incriminating video of police misconduct. A Facebook page publicizing the killing and showing pictures of Saeed's body became a focal point for the expression of grievances against the Mubarak regime.

Protest organizers used Facebook and Twitter to organize the initial protest on 25 January 2011. Facebook was a platform for sharing ideas and plans, causing the government to try to block Internet access in the country. The size of the crowds in Tahrir Square grew rapidly over the course of the protests, but it was critical that the first day of protest drew an overwhelming crowd because this was a powerful signal of the amount of discontent among the public that encouraged others to join. Online communications were coupled with conventional means of spreading information through leaflets and word of mouth to reach individuals without the Internet access. Social media informed people and strengthened ties and commitments among them that were critical to sustaining protests in the face of opposition.

Social Movements Now and Then

Social media have facilitated the rapid diffusion of symbols and methods across nations, as well as solidarity for common causes. The Occupy movements in Western cities in 2011 were in many ways inspired by the Egyptian Revolution. The original announcement of the idea to 'occupy' Wall Street was carried in an Adbusters poster with a Twitter hashtag that went viral over the Internet.

Some of the appeal of recent social protests against economic inequality in Western nations has been their spontaneity and lack of hierarchy, discipline, and leadership. The protests have marshaled and aggregated discontent about the economy, government, social services, equality, and political representation. Not everyone who participates will be or needs to be equally committed to the goals of the movement. When large numbers of protesters are joined in opposition to a regime, the presence of weakly committed individuals may, nonetheless, make a crucial difference in communicating the size and breadth of social support for the movement. For example, Internet protests in China that are held together by loose online ties have often been sufficient to instigate a government response to mass reactions to unjust outcomes or decisions.

However, the elements of successful social movements are not fundamentally changed by new technologies, even if those technologies provide additional resources that can be employed in contemporary protests. Organizations may have become less central to the mobilization of people with the advent of social media, but weakly structured movements that do not have clear leadership also exercise less control over the participants and the demands and agenda of the movement.

Movements still require leadership, strategy, and interpersonal and organizational networks to establish goals and to ensure discipline and continued participation. Inexperienced activists without prior participation in protests may be less committed and less likely to sustain their participation in the face of government opposition or repression. The spontaneity of the Occupy movement throughout Western nations in 2011–12 and its themes of economic equality and corporate responsibility enjoyed considerable short-term success, but there was no agenda for achieving policy changes that may have sustained the movement.

Civil Disobedience

Unlike protest marches, picketing, and other kinds of legal protest, civil disobedience is a particular form of political protest that is marked by the deliberate violation of the law for social purposes. Civil disobedience is not merely conscious law breaking, but rather it is law breaking with the express goal of correcting a perceived injustice. Action that would now be regarded as civil disobedience has been a regular feature of social protest throughout history even though the term itself, attributed to an 1849 essay by Henry David Thoreau, is a relatively recent creation.

The act of civil disobedience is fundamentally an act of political protest because the noncompliance is intended to convey a message of discontent to the government, media, and

public. The acts of defiance are typically overt, public acts aimed at marshaling collective opposition to unjust or unconstitutional laws and policies. Civil disobedience is not restricted to protests over laws and public policies, but is also employed against the practices of private institutions such as corporations and universities, as occurred, for example, in the divestment campaign on American university campuses in the early 1980s, which pressured universities (both private and public) to cut their financial ties to companies that conducted business in South Africa in an effort to bring an end to apartheid.

Forms of Civil Disobedience

Within this broad conceptualization, civil disobedience can emanate from different motivations and assume a number of forms. Civil disobedience can be active or passive, direct or indirect. An individual can either actively commit a prohibited act or passively refuse to conform to a prescribed action. For example, an antiwar protester may destroy his draft card or, alternatively, he may refuse to register for the draft. Both actions qualify as civil disobedience because the protester intentionally violates the law as an expression of his opposition to the war, knowing the legal consequences of his actions. Thoreau himself engaged in a form of passive civil disobedience when he chose to register his opposition to the institution of slavery by refusing to pay the Massachusetts state poll tax.

Civil disobedients often break laws that are unrelated to the law or policy they are protesting. For example, peace activists who disrupt operations at defense laboratories are not protesting laws against trespassing, but are using civil disobedience to dramatize their opposition to nuclear weapons. Supreme Court Justice Abe Fortas objected to the indirect use of civil disobedience on the grounds that it would give license to people to break any law they wished in the service of a goal that they perceived to be nobler. But few scholars have subscribed to this restriction because, on many issues, there is no feasible direct method to engage in civil disobedience. There may, for example, be no direct method by which activists can use civil disobedience to protest the government's foreign policy or its failure to pass laws addressing issues of inequality and poverty.

A narrower definition of civil disobedience would exclude illegal actions taken by individuals who do not accept the constitutionality of the law they are violating. By this definition, an activist who breaks the law by setting fire to the American flag is *not* engaging in civil disobedience if he believes that his actions are protected by the First Amendment, but a flag burner who accepts the constitutionality of a law against flag burning *is* engaging in civil disobedience. For the purpose of defining civil disobedience, however, it may be unnecessary to make such a distinction based on the activist's frame of mind, because the crucial assumption underlying civil disobedience is that the law being broken is generally viewed to be valid by those in authority, so the actions of the civil disobedient are generally viewed to be illegal. Therefore, both flag burners are committing civil disobedience in the sense that they are breaking what is generally understood to be the law in an effort to change it, and the main difference between them is whether they are appealing to the courts or to the executive and legislature for remedial action.

Although Thoreau assumed that 'civil' protest would be conducted in an orderly and peaceful fashion, scholars disagree over whether civil disobedience is necessarily nonviolent by definition. Some commentators insist that civil disobedience must be nonviolent, or else it cannot be distinguished from rebellion, rioting, and other types of violent direct action. The two individuals who are most responsible for the association of civil disobedience with nonviolence are Mohandas Gandhi (see e.g., Gandhi, 1961) and Martin Luther King, Jr (see e.g., King, 1964). Indeed, contemporary understandings of civil disobedience have been substantially shaped by the mass campaigns for equal rights organized by Gandhi in South Africa and India, and by the King in the United States.

Both Gandhi and King possessed the extraordinary combination of political 'savvy' and unwavering moral and spiritual commitment to the philosophy of nonviolence. In his teachings, Gandhi insisted that civil disobedients must scrupulously obey laws that are unrelated to the target of the disobedience campaign and submit themselves to the punishment of the state for breaking a law they believed to be unjust. Civil disobedients do not flee or resist arrest, for in accepting penalties for their actions, they bring attention to the injustice they wish to correct and arouse the sense of injustice in others. Gandhi's concept of *satyagraha* or 'truth force' refers to the moral conversion of one's adversary. In their quiet suffering, civil disobedients demonstrate their honor and integrity and their capacity to differentiate between just and unjust laws.

King regarded Gandhi as his mentor and shared his belief that the purpose of civil disobedience is to persuade the state and its citizens that an existing law violates a higher moral principle. Violence, therefore, must be avoided because it potentially undermines the communicative intent of civil disobedience. Adherence to nonviolence, on the other hand, indicates that one bears no malice or hatred for one's opponent. Like Gandhi, King believed that the ultimate goal of a campaign of nonviolent protest is to achieve a new moral understanding with one's opponents. Nonviolent resistance works, not by defeating and humiliating the opponent, but by winning the opponent's empathy and friendship.

Justification of Civil Disobedience

In addition to disagreements over what kinds of actions qualify as civil disobedience, there is debate over the circumstances in which civil disobedience can be justified. Civil disobedience raises issues about the nature of authority, obligation, and consent in society. The practice originates from the idea that a person must be true to his conscience and that he is not obligated to submit to the laws or commands of the state when they violate his conception of justice. Civil disobedience therefore assumes both a right to individual freedom and the existence of a natural law that takes precedence over civil law.

Philosophical discussions of the justification for civil disobedience generally fall between legal and moral individualism. In justifying civil disobedience, it is necessary to accept that there are standards against which governments can be evaluated, and citizens have the right to exercise such judgments. Few would argue that the law must always be obeyed; and similarly, few would contend that obeying the

law is entirely a matter of individual judgment about what is morally right and wrong. One extreme supports despotic governments, while the other suggests anarchy. Whether civil disobedience undermines respect for the law is a matter for empirical inquiry. The penalties for breaking the law are usually sufficiently great that it is unlikely that people will engage in civil disobedience unless they are intensely motivated by a sense of injustice.

But do the moral criteria for civil disobedience reside only in conscience? Or are there intersubjective criteria that permit rational assessment of the morality of a law or policy? Without such objective criteria, we are in danger of every individual invoking arbitrary subjective criteria for disobeying different laws. We invite moral anarchy unless there are some agreed-upon criteria for evaluating whether a law or policy is unjust or immoral. At a minimum, the civil disobedient must be able to demonstrate that the law or policy violates principles of welfare, equality, or justice. The law or policy must be shown either to do more harm than good, to distribute benefits unequally, or to operate without the consent of certain groups. Legal methods of petition to change the law must be tried first, and the activists must grant the same rights of civil disobedience to others who believe that another law is unjust. Some would argue that the civil disobedient must also have a reasonable chance of overturning the law. According to these normative criteria, only those who believe in reciprocity and adhere to a sophisticated set of general rules of political conduct can rightfully be said to be engaged in justifiable acts of civil disobedience.

Effectiveness of Civil Disobedience

What are the mechanisms that make civil disobedience effective in practice? Gandhi recognized that civil disobedience could not be effectively employed for a global, distant goal such as national independence, but had to be used for specific, local issues on which the opposition had the power to compromise and make concessions. True to form, civil disobedience campaigns led by Gandhi were most successfully applied in local disputes concerning such issues as land revenue assessments and mill workers' wages, in which small numbers of close knit, disciplined protesters were united by their interest in the issue and capable of conforming to a norm of nonviolence.

In theory, if civil disobedience is to work through persuasion, then the protesters must invoke principles or values that are shared by a large proportion of the population. In practice, however, the preferences that are converted are more likely to be those of the spectators to the conflict than those of the adversaries of the protesters. A small number of disobedients can seldom, by their actions alone, render a law inoperative, but they may be able to disrupt the social order sufficiently to draw public attention to their grievances. Although Gandhi believed that the suffering of protesters in jail would apply pressure on the government, he also recognized that the state's use of violence would serve the interests of the protesters by influencing third parties who were witnesses to the conflict. British repression of Indian dissenters confirmed to onlookers that the British ruled by force and not through the consent of the Indian population. Similarly, in the course of the civil rights movement, it was the contrast between the nonviolent protesters and the violent response of the southern police that helped to galvanize public opinion and mobilize federal government intervention in support of racial equality.

In response to a campaign of civil disobedience, the defenders of the status quo must decide whether to offer concessions to the protesters or to escalate the conflict by using force or summoning the power of the state. It is the threat of continued disruption of the normal operations of business and government that 'persuades' the targets of the protest to negotiate with the activists and to yield to their demands. During the civil rights movement, the early success of direct action campaigns in some southern cities aimed at forcing businesses to provide equal treatment to blacks made it easier for similar campaigns to succeed elsewhere, because businesses foresaw the normal course of these protests and sought to cut their losses sooner rather than later. There is little evidence that the attitudes of white southerners were converted by direct action; indeed, the attitudes of many were hardened by the confrontational tactics of the movement.

Civil disobedience thus rarely works through persuasion and attitude change alone, but also depends significantly on pressure and coercion. Actions that are aimed at transforming preferences may leave attitudes unchanged, but still be effective at promoting social change because the opponents are pragmatic enough to calculate that they are better of making concessions than dealing with the disorder created by continual protests. Civil disobedience in this respect resembles a pragmatic tactic within a repertoire of potential strategies that are available to a protest group rather than an option that is resorted to only when the conventional channels of participation have been exhausted.

See also: Citizen Participation; Civil Rights Movements; Civil Society, Concept and History of; Collective Behavior, Sociology of; Peace and Nonviolence: Anthropological Aspects.

Bibliography

Bay, C., 1968. Civil disobedience. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. MacMillan, New York.
Chong, D., 1991. Collective Action and the Civil Rights Movement. University of Chicago Press, Chicago.
Gandhi, M.K., 1961. In: Kumarappa, B. (Ed.), Non-violent Resistance. Schocken Books, New York.
Gewirth, A., 1982. Human Rights: Essays on Justification and Applications. University of Chicago Press, Chicago.
Greenawalt, K., 1987. Conflicts of Law and Morality. Oxford University Press, New York.
King, M.L., 1964. Why We Can't Wait, first ed. Harper & Row, New York.
McAdam, D., 1982. Political Process and the Development of Black Insurgency 1930–1970. University of Chicago Press, Chicago.
McAdam, D., Tarrow, S., Tilly, C., 2001. Dynamics of Contention. Cambridge University Press, New York.
Powers, R.S., Vogele, W.B., 1997. Protest, Power, and Change: An Encyclopedia of Nonviolent Action from ACT-UP to Women's Suffrage. Garland Publishing, New York.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Sharp, G., 1973. The Politics of Nonviolent Action. Porter Sargent, Boston.
Tufekci, Z., Wilson, C., 2012. Social media and the decision to participate in political protest: observations from Tahrir Square. Journal of Communication 62, 363–379.
Zinn, H., 1968. Disobedience and Democracy: Nine Fallacies on Law and Order. Random House, New York.

Political Psychology

Christian Staerklé, University of Lausanne, Lausanne, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political psychology is an interdisciplinary field of inquiry concerned with the analysis of political thought and behavior of individuals within politically organized communities. Survey, experimental, and qualitative research examines political behavior at the individual and the collective level, in the general public and among political elites, and as formal and informal processes of political participation. Key topics in political psychology include political beliefs and values, cognitive processes in political decision making, political communication, media effects, political rhetoric, international conflict, racism and prejudice, ethnic identities, and collective action.

Political Psychology

Political psychology is an interdisciplinary scientific field of inquiry concerned with the study of political processes from a psychological perspective. At the most general level, political psychology is concerned with political thought and behavior of individuals within politically organized communities. Research in political psychology examines political behavior at the individual (e.g., decision making) and at the collective level (e.g., collective action), it concerns processes occurring in the general public (e.g., public opinion) and among political elites (e.g., psychology of leadership), and it relates to formal (e.g., voting) and informal (e.g., community involvement) processes of political participation.

The boundaries between political psychology and the various fields of inquiry it draws upon are blurred. The discipline uses theory and methodology derived from cognitive, personality, social, developmental, and organizational psychology, from research traditions within political science interested in political behavior and decision making, from sociology, education, public opinion, and communication research. It is therefore difficult to demarcate without ambiguity the boundaries of research in political psychology from studies in the respective neighboring disciplines.

Its interdisciplinary nature has been an essential feature of political psychology since its beginning. During the short history of the discipline, the two disciplines have often existed side by side, for example, when either psychology (personality analyses, political cognition) or political science (political behavior) defined the main paradigms under investigation. Yet, the question of how to articulate and integrate the different methodologies, worldviews, and epistemologies associated with psychology and with political science remains a recurrent concern. Some approaches may be accused of 'too much psychology' (neglecting political, organizational, and systemic explanations) while others may be seen as 'too much political science' (using simplified models of psychological capacities and motivations and ignoring how citizens actively make sense of political situations). Nevertheless, much research has successfully integrated the two disciplines by creating original and innovative research paradigms in political psychology.

In the next section the historical and institutional context of political psychology is briefly outlined, followed by summaries of the most common fields of investigation of political psychology.

The Context of Political Psychology

A Short History of Political Psychology

Political psychology has a long past, but a short history. While the origins of the interest in the relationship between psychological processes and political organization can be traced back to ancient Greece, the philosophy of enlightenment and nineteenth century social and political science (with widespread interest in crowd psychology, for example), political psychology as a field and later as a discipline emerges in its own right after WWII. McGuire (1993) distinguishes three phases in the short history of political psychology: in the 1940s and 1950s, researchers were concerned with behavioral pathology and the impact of personality on political processes, inspired by psychoanalytic and behaviorist theories. During the 1960s and 1970s, scholarly interest turned toward political attitudes and voting behavior, largely based on newly available survey data and statistical techniques that allowed, for example, investigating the impact of political campaigns on attitudes and voting behavior. During the 1980s and early 1990s, the field was largely focused on political cognition and individual decision making, thereby reflecting the cognitive trend observable in other behavioral sciences as well. The citizen was represented as a limited information processing machine that was nevertheless able to take good enough decisions based on various cognitive heuristics. While the first handbook of political psychology was published in the early 1970s (Knutson, 1973), it is during this period that a number of textbooks explicitly devoted to the discipline of political psychology appear, all clearly marked by the cognitive approach to political psychology (e.g., Iyengar and McGuire, 1993; Lau and Sears, 1986).

Since the mid-1990s, the field has grown significantly and has become ever more diversified. Today, no single approach can be considered as dominant or organizing the field as a whole. 'Classical' research topics on personality profiles of leaders, psychodynamic explanations of political behavior, cognitive approaches to political decision making, or analyses of prejudice and racism in political processes continue to

attract the attention of researchers. Political psychology has been thriving during the past two decades and an evident sign of this trend is the increasing number of new books published under the header of political psychology, including edited books (Borgida et al., 2009; Monroe, 2002), a textbook (Cottam et al., 2010), a reader (Jost and Sidanius, 2004), and a handbook (Sears et al., 2003; Huddy et al., 2013; for the second ed.). Moreover, reflecting the trend toward integration of biological and social processes, political behavior is increasingly explained with neurological and genetic models (Marcus, 2012).

The Societal Turn

In 1993, McGuire was correct in predicting that a fourth wave of political psychology would concern research on intergroup relations, social change, and ideological beliefs. These new topics seem to reflect a deep change that has occurred in political psychology since the mid-1990s. Much like social psychology, political psychology has moved toward closer scrutiny of societal issues related to power relations between groups, political legitimacy, and civic participation. One may speculate that this changing focus is the result of changing historical conditions of the recent past, including the breakdown of cold war certainties in 1989, growing economic inequalities in contemporary free market societies, an increasing number of grievances and claims by minority groups, tense political polarization in many Western countries, and the various terrorist attacks that have marked recent history.

This theoretical expansion of political psychology is associated with a significant geographical development of the discipline. Since its beginning as a modern scientific discipline, political psychology has largely been defined through North American research. Accordingly, until the mid-1990s, non-American political psychology was virtually inexistent. This situation of American predominance has now radically changed: the single most important development in the recent history of the discipline is the growing interest for and the massive expansion of research in political psychology in non-American countries, in particular Europe. This is evidenced with increasing attendance from European researchers at political psychology conferences, with the number of papers published by European researchers, and with the founding of the new *Journal of Social and Political Psychology* in 2013 that explicitly seeks to attract a broad international audience. More importantly, this move toward Europe has also sparked new theoretical developments, for example, critical political psychology that focuses on collective action, globalization, and the migration experience (Kinnvall et al., 2014; Tileagă, 2013) and a social representations approach to political psychology that proposes dynamic, contextual, and normatively determined analyses of political behavior (e.g., Howarth et al., 2014).

Methods in Political Psychology

The field relies on a variety of methodological approaches. Survey research is privileged by political scientists to study research questions requiring correlational and longitudinal analyses, for example, dynamics of public opinion and voting behavior, political values and ideological beliefs, and cross-cultural differences in political attitudes. Experimental approaches are more often used by psychologists to test causal models, for example, concerning cognitive processes in rational choice decision making. The recent shift toward research concerned with societal issues has brought qualitative approaches to the forefront of political psychology, including discursive and rhetorical approaches that enrich the traditionally positivist worldview of political psychology with a more constructivist epistemological orientation.

Institutionalization

While research in political psychology has existed since the 1940s, proper institutionalization of the discipline began in 1978 when the International Society of Political Psychology was founded by Jeanne Knutson from the University of California, Los Angeles (UCLA). In 2013, the association had about 800 members many of which attend the Annual Meetings that have been held every year since then, with rotating locations between North America, Europe, and other locations around the globe. A summer academy in political psychology for young scholars is linked to the Annual Meetings. The society publishes Political Psychology, the respected quarterly flagship journal of the discipline, and, starting in 2014, the annual series Advances in Political Psychology.

Moreover, a 3-week summer school in political psychology is organized by Stanford University, and major associations such as the American Political Science Association and the European Consortium for Political Research have active political psychology sections.

Graduate and undergraduate courses under related titles are taught in leading universities around the world, and many American universities offer degrees in political psychology, mostly in political science departments. While Queens University, Belfast, is the first European university to offer a complete political psychology program, many other universities have graduate programs that run under different labels but nevertheless offer curricula closely related to political psychology (e.g., various social psychology programs).

Overview

In this text, some of the most emblematic research in political psychology will be presented, but many important topics are necessarily left out (e.g., biological, genetic and evolutionary approaches, deliberation and network research, emotional and motivational approaches, psychobiography, conflict resolution). Research in political psychology can be organized into three major categories of research, and the presented approaches are deemed to represent some of the key characteristics of each of these categories.

Individual-level processes are concerned with personality profiles of citizens and political leaders, with political beliefs and values, with the role of socialization and individual development in the formation of political beliefs, and with the examination of cognitive processes in political decision-making. A second strand of research is concerned with *political communication*, including media communication, political

rhetoric, and constructivist approaches. A third field of research examines *group-level processes* in political thought and behavior, including international relations and conflict, the role of racism in politics, group identities and cultural diversity, and collective action and social change.

Individual-Level Processes

Personality Profiles of Citizens and Political Leaders

Personality is understood as being composed of traits, needs, motives, and self-conceptions that predispose individuals to relatively stable and consistent patterns of political thought and behavior. Hitherto referred to as 'temperament' or 'character,' personality profiles have been amply used to explain the behavior of both political leaders and followers. Psychodynamic approaches were dominant in the early days of research on personality, referring to unconscious drives and mechanisms. The *Authoritarian personality* syndrome (Adorno et al., 1950) was developed in the aftermath of WWII to explain mass submission to political authorities, aggressive behavior against minority groups, and uncritical endorsement of totalitarian ideologies. This pioneering study was further developed with research on dogmatic personalities that leads individuals to be more receptive to rigid, closed-minded belief systems (Rokeach, 1960). But it was not until the 1990s that the 'Big five' became an almost consensually accepted general system to describe personality traits through five dimensions (extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience). Much research since then has examined the relationship between personality traits and political orientation, showing, for example, that openness to experience is generally associated with left-wing support, whereas conscientiousness is more likely to lead to support for right-wing parties. Other studies have shown persistent links between extraversion and openness to experience and various forms of political participation, including voting, attending meetings, campaigning, and signing petitions. Another personality dimension concerns general needs and desires, differentiated by McClelland (1985) into the three basic motives of achievement, affiliation, and power. Research in the motivated social cognition tradition, in turn, distinguishes between epistemic, existential, and relational needs, and showed that political conservatism is associated with the epistemic need for cognitive closure and the existential need for safety (Jost et al., 2003).

Studies on *personality profiles of political leaders*, in turn, suggest that history is at least partly shaped by leaders' idiosyncratic ways of perceiving and interpreting social reality (e.g., distortions, confirmatory bias), private memories (e.g., war experiences in childhood), personal goals (e.g., power orientation), cognitive styles (e.g., rigidity), and emotional reactions to stress (e.g., low tolerance for stress). Autocrats' grandiose self-conceptions, fears, and obsessions also play a role in their strategies to maintain power. The major difference with large-scale personality studies is that leader personalities can generally only be assessed with indirect methods, including psychobiography (accounting for leader personality through an analysis of childhood experiences), content analyses of formal and informal speeches, and personality ratings by individuals close enough to leaders to judge them.

Political Beliefs and Values

Political beliefs differentiate individuals and groups according to their ideological orientations and political preferences. Debates revolve around the question of the structure of beliefs systems. In many studies, a single dimension (conservatism vs liberalism or left- vs right-wing orientation) is used to differentiate individuals, but it becomes increasingly clear that one-dimensional models are no longer adequate to account for the heterogeneity and variation of political beliefs. There is today a large consensus that at least two independent dimensions are required, one organized around social and cultural issues (traditional morality vs cultural freedom), and the other one around economic positioning (egalitarianism vs meritocracy). An important set of studies has, for example, analyzed the complex and context-dependent relationships between *Social Dominance Orientation* (assessing attitudes toward group-based inequality) and *Right-Wing-Authoritarianism* (the ideological equivalent to the authoritarian personality described above).

Values play also a central role as individual-level organizers of political judgments and behavior. Schwartz's (1992) theory of the human values system paved the way for the systematic analysis and comparison of the impact of values on political behavior. Typically, left-wing political orientations are associated with values of universalism and benevolence, whereas right-wing positions are related to the values of self-enhancement, power, achievement, and conformity.

The origins of political beliefs and values can partially be traced back to *socialization* during childhood and early adult years. There is increasing evidence that early learning of politically relevant information, including political beliefs, intergroup attitudes (e.g., racism), and social identities (e.g., national identity) remain stable across time, giving rise to considerable political continuity over an individual's life course (Sears and Brown, 2013). There is also support for the hypothesis of preadult family transmission, with studies showing substantial parental transmission of political party identification to adolescent children. Nevertheless change in sociopolitical behaviors, attitudes, and opinions still occurs in adult years, as a function of new contingencies and changes in the wider sociocultural environment.

Cognitive Approaches to Political Decision Making

Political decision making is one of the most important research domains in political psychology, and rational choice theory is the most commonly used theoretical framework to explain decision-making processes. Basic assumptions of (normative) rational choice theory are that individuals have a coherent set of preferences, gather the necessary information to reach an informed decision, evaluate alternative actions, and choose actions that are optimally related to their beliefs and values. Such decisions are expected to further individuals' self-interest and are therefore deemed rational. However, the rational choice approach is confronted to a paradox, since research has amply shown that in practice political decision-making virtually never follows these principles. People lack consistency in

their opinions, use information incorrectly, are overconfident in their own choices, fail to adapt existing evaluations in light of new information, draw unwarranted conclusions from insufficient data, and express prejudiced opinions. Moreover, political decision making, in particular voting, is only weakly related to actual self-interest.

Following the implausibility of a 'full' rational choice model, researchers have developed models of 'bounded' rationality that are based on similar assumptions as rational choice theories, but that recognize individual and contextual variation in decision-making processes and outcomes (Kahneman, 2011). To compensate limited information, cognitive biases and lack of motivation, bounded rationality models such as behavioral decision-making theory describe a number of cognitive strategies that help individuals to make 'good enough' decisions: simplifying the decision task, selectively filtering new information, interpreting information as a function of preexisting ideological predispositions, and making sense of political issues through evaluations of social groups involved in the decision (e.g., beneficiary groups of welfare programs). Individuals also refer to opinions of others to make up their mind, for example, by relying on expert judgments, by socially validating their opinion through comparisons of opinions with those of relevant others (informational influence), or by aligning their opinion with the perceived majority opinion in their community to avoid marginalization (normative influence). Finally, they may also defend the interest of their group (collective self-interest) rather than only their individual self-interest.

Nevertheless the amount of *factual knowledge* citizens possess does make a difference for the decision-making process (Delli Carpini and Keeter, 1996). High levels of knowledge are, for example, likely to increase the alignment of one's decision with one's beliefs and preferences. Yet, while many studies highlight low levels of political knowledge and lack of political sophistication of citizens, the actual implications of this general lack of knowledge remain debated.

Political Communication

Media Communication

Communication between citizens and between citizens and political leaders is another key topic for political psychology. Research on mass media has analyzed how media shape political outcomes by directing the public's attention to major issues (agenda setting) and by describing those issues as a function of the communicator's interests (framing effects). *Agenda setting* shows how successful the media are in determining what citizens 'think about,' based, for example, on research showing nearly perfect correspondence between the issue priorities of the public and news organizations. Research on *framing* in turn has shown that individuals are highly sensitive to the way policies are described and survey questions worded, and that even minor changes in the labeling of political alternatives can produce strong effects on public opinion. For some, such findings challenge the validity of democratic opinions, since opinions appear to be highly volatile as a function of irrelevant pieces of information. Antidotes to framing effects are deliberation, expertise, and exposure to competing frames (Druckman, 2004). Nevertheless, framing effects can also be viewed as part and parcel of everyday political reasoning, since people necessarily rely on peripheral cues to form an opinion. Framing has also a strong strategic dimension, since messages are deliberately framed by political elites (parties, interest groups) to move opinions in the direction of their interests. Extremist groups, for example, may be portrayed as exercising their right to freedom of expression, or they may be described as a risk for public safety. Similarly, poverty may be explained with individual behaviors or with political choices, thereby altering attributions of responsibility and ultimately affecting political decisions.

Constructivist, Critical, and Rhetorical Approaches

The growing number of constructivist and critical approaches to political psychology can be seen as one of the most important theoretical developments in the recent history of the discipline. These approaches offer a conceptual alternative to the causal models characterizing a majority of studies in political psychology (Tileagă, 2013). In a constructivist perspective, political thought and behavior is contextualized within the social and cultural environment.

This constructivist perspective can be exemplified with the social representations approach that has recently entered political psychology. This approach explores the processes by which citizens come to a shared understanding of political issues and how this shared understanding relates to their possibilities for political action (Elcheroth et al., 2011). Communicating knowledge creates communities of belief that transform individual opinions into social representations, that is, into knowledge shared and debated by members of the group. Such shared knowledge (e.g., awareness of excessive social inequalities) then becomes the basis for political action, both individual (e.g., voting) and collective action (e.g., mass demonstrations). In line with constructivist principles, social representations do not just reflect social reality but they constitute that social reality, and sometimes change the nature of that reality. Hence, political thought and behavior is as to large extent shaped by what people think others are thinking (meta-representations).

Given its focus on the social context that produces representations, a social representations approach highlights the role of power relations in the communication and dissemination of representations between individuals and between groups. Political participation, for example, can be conceptualized as the power to construct and convey particular representations over others (Howarth et al., 2014). In this view, the 'framing' of concepts is a sign of the inescapably contested nature of political concepts such as freedom, equality, democracy, or justice. Such abstracts notions are open to debate and dispute and much of political action and discourse can be viewed as competitive attempts to impose specific interpretations of political concepts, events, and situations consistent with the actor's political goals. Opposing sides may, for example, refer to a same concept that is assumed to appeal to the people: for the Democratic Party in the United States, the Affordable Care Act (Obamacare) brings 'justice' to the American people, whereas for the Republican Party only the repeal of the same law would do so. Quite clearly, the interpretations of 'justice'

vary wildly in these two rhetorical perspectives. A social representations approach thus examines the construction of legitimate and dominant political knowledge, and the strategies used to marginalize and silence alternative ways of knowing and being. While this type of power is typically afforded to high-status groups who have the power to construct what is 'real' or 'true,' social representations nonetheless allow for resistance and agency through the active reappropriation of existing knowledge.

Another variant of a constructivist perspective are political rhetoric approaches that take up a critical stance toward conventional models of decision making and psychological functioning, arguing that dispute and conflict are essential features of political thought and behavior (Condor et al., 2013). According to this line of work, citizens are embedded in a sociocultural environment characterized by ideological dilemmas and opposing worldviews that shape citizens' attitudes toward political issues. Rather than being expressions of an irrational common sense, such contradictions are the preconditions for political thought and for rhetorical deliberation within and between members of a society, and consequently for a lively democratic political culture.

Group-Level Processes

International Conflict

Research on international relations and international conflict is a traditionally important aspect of political psychology. Rooted in a social psychological perspective, this strand of research offers a complementary view on structural, strategic, and realist analyses of international conflict. It assumes that subjective perceptions of the conflict determine actions and responses by the conflicting parties, that these perceptions develop in the context of dynamic and changing intergroup relations, and that ongoing interaction between conflicting parties determines the development and outcome of the conflict (Fisher et al., 2013). International conflict is thus a process driven by collective needs and fears rather than by objective national interests, in particular concerning identity, recognition, and autonomy of ethnonational groups.

This focus on subjective perceptions of conflict is evident in research on stereotypical images of countries. These images are held by both the general population and political elites and define the perceived threats and opportunities countries are confronted with. They are stable, difficult to change, and easily become taken-for-granted assumptions that produce routinized responses of what is seen as in a country's interest or against it. A limited number of such images seem to organize international relations, including enemy, ally, degenerate, colony, and barbarian images (Herrmann, 2013). An enemy image implies, for example, a tendency to dehumanize its population and to interpret any activity by its government as evidence of resolve and aggressive intent. Hence, international relations are severely prone to misperceptions with potentially catastrophic consequences; for example, when a military conflict escalates due to distorted or selective perceptions of the conflict situation. Other factors that have been shown to contribute to conflict escalation are the normative reinforcement of bellicose attitudes in times of conflict and psychological commitments that prevent individuals and groups deeply involved in the conflict to actively engage in deescalation and conflict resolution. This is the case when worldviews are so strongly built around the conflict that they would be threatened by an end to the conflict.

Racism and Politics

Negative attitudes toward groups – prejudice – is another central group-level topic addressed by political psychologists. This research has its roots in U.S. American race relations that have shaped American political debate since the nineteenth century. The current state of research on the role of prejudice in politics can be traced back to the 1960s, a period during which a number of crucial changes in American race relations have occurred, most notably the Civil Rights Act of 1964 that outlawed any institutionalized discrimination based on racial criteria. Up to this period, overt and blatant racism against Blacks was largely seen as legitimate, but with the deep changes of the racial relations following the civil rights era, beliefs about biological inferiority of Blacks became more and more unacceptable. Instead, new, more subtle forms of racism emerged, called symbolic, modern, or ambivalent racism that were characterized by the rejection of claims about biological inferiority and instead emphasized value differences between racial groups and value violations by Blacks (e.g., in claims that racial inequality was due to lacking work ethic by Blacks).

It is in this context that researchers have studied the role of prejudice and racism in American politics (see Kinder, 2013). Research has shown that racial animosity played and still plays a central role in many political processes, including party identification, voting, and policy opinions. Identification with the Republican and the Democratic Party closely follows racial divides, following the major political realignment in 1964 that shifted Blacks' identification from the Republican to the Democratic Party in the wake of the parties' starkly opposing positions on the Civil Rights Act.

Racial prejudice may act as a shortcut to political decisions regarding support or opposition to social policies. Many studies have demonstrated that public opinion is shaped by the attitudes citizens hold toward the groups deemed to be beneficiaries or victims of a given policy, in particular concerning measures destined to correct social inequalities (e.g., welfare, employment, and affirmative action policies). Prejudice has been shown to explain widespread opposition to welfare (Gilens, 1999), and the fierce opposition to the Affordable Care Act (Obamacare) in 2013 can also be seen as being based on racial resentment. It must be noted, though, that the interpretation of the observed relationships between prejudice and political outcomes is itself hotly debated. For some, the effects are a mere expression of principled conservatism without any connection to prejudice, whereas for others cumulated survey evidence is proof enough that prejudice plays the assumed causal role in these political processes.

Group Identities and Cultural Diversity

The recent focus on issues of social change and political legitimacy has led to an intensified use of theories of intergroup relations and political identities in political psychology. These

approaches conceptualize groups as the link between societal conditions and psychological processes to address issues such as multiculturalism, social inequality, intergroup dominance, and discrimination from the perspective of political psychology. In line with the principle that political thought and behavior is context dependent, many approaches further assume that psychological processes are contingent upon the social position of individuals. Research therefore challenges the idea of universal psychological processes and differentiates analyses of psychological processes between dominant versus subordinate groups, or between ethnic majority versus minority groups.

A key concept in this intergroup approach to political psychology is political identity. Based on different intergroup approaches developed in social psychology (e.g., Social Identity Theory, Realistic Conflict Theory), political identities refer to identification with and meaning attributed to membership in politically relevant groups, including political parties and national, ethnic, linguistic, or gender groups (Huddy, 2013). Research has, for example, shown that individuals who strongly identify with their group are more likely to act on behalf of a group-related cause, to view the political environment in antagonistic terms, and to act defensively in face of group criticism.

Cultural diversity of contemporary societies provides the background to much of group identity and intergroup relations research in political psychology (Green and Staerklé, 2013). The most widely studied instances of (majority) group identity concern different forms of national attachment and identification. Two broad forms can be distinguished: patriotism denotes positive feelings of pride toward the nation and its symbols. It may be accompanied by critical attitudes that invite group members to speak up and criticize the nation when necessary. While patriotism is fundamentally non-comparative, nationalism, in turn, together with its derived forms of chauvinism and blind patriotism, is a comparative form of attachment that asserts national superiority in comparison to other countries, and may in its extreme form question the mere presence of immigrant and other minority groups on national soil.

A related strand of research is concerned with the ways majority members of receiving societies react to the presence of migrant groups, often based on analyses of perceived symbolic and material threat evoked by immigrants and of various forms of intergroup contact. New developments include the widespread use of multilevel techniques to examine the role of political and ideological contexts (assessed with macro-level measures at the national or regional level, e.g., strength of political parties) in explanations of hostile attitudes and behaviors toward immigrants.

Even though group identities are important for both majority and minority groups, their role in political processes has often been studied with minority groups, in particular ethnic or migrant groups. Strong minority group identities increase, for example, the likelihood that group members define their (material) interests according to the collective interests of the group rather than as a function of their individual self-interest. Highly identified minority group members develop a sense of common fate with other group members that leads them to express claims and grievances in the name of the group and to defend their own and fellow group members' interests.

Other research has examined subjective migrant experiences within receiving societies from a discursive perspective, focusing on the active construction of intercultural identities and everyday practices of intercultural interactions. Studies have, for example, shown that while ethnic identities involve beliefs in commonality, shared kinship, and ancestry, they also capture processes of ongoing negotiation of ethnic boundaries, both within the migrant group (e.g., regarding the maintenance of cultural traditions in the receiving society) and between the migrant group and the receiving society (e.g., the nature of assimilation).

Social Change and Collective Action

Research on civic participation and collective action in political and social psychology has increased significantly as part of the new interest of researchers in issues of social change and contestation. This increase goes hand in hand with the rise of social movements all around the world during the past two decades. Collective action research is concerned with movement politics as opposed to party politics, that is, with active participation in mass demonstrations, protest events, occupations of public sites, boycotts, blockades, riots, and the like (Klandermans and van Stekelenburg, 2013).

Early approaches to collective action in the 1950s and 1960s explained protest by high levels of discontent experienced by individuals, followed by resource mobilization and political process theories that claimed that aggrieved people engage in collective action when the necessary organizational resources are available. Within these structural approaches, participation in collective movements was explained with rational choice models according to which individuals evaluate the costs and benefits of participation.

In the early 1990s, new social constructionist approaches to protest began to emerge in response to the dominant structural approach. These new approaches attempted to move away from static, decontextualized explanations of protest to more dynamic and contextualized models. These new models highlighted processes of social interaction and negotiation among participants, opponents, and bystanders, and analyzed the role of collective emotions and subjective definitions of the situation to explain collective action. In line with earlier theories, grievances, opportunities, and resources were still required, but grievances were understood as more nuanced responses to illegitimate inequality (assessed with judgments of distributive justice and group relative deprivation), suddenly imposed grievances (e.g., new roads or buildings) and violated principles (e.g., betrayed trust).

Moreover, since grievances are ubiquitous in contemporary society, the more important question now was to examine the circumstances under which grievances do actually give rise to actual protest movements. To answer this question, the various contemporary theoretical frameworks on collective action converge on the key role played by two factors: first, a politicized collective identity is a crucial component of engagement in collective political action. Many studies have shown that the more individuals identify with a group involved in a protest activity, the more likely they are to participate in that activity.

Movements offer indeed the opportunity to act on behalf of one's group and to celebrate the collective identity with symbols, rituals, marches, and songs. The second factor is efficacy as the belief that the movement will be able to redress grievances effectively. In addition, ideological motivations and intergroup emotions (in particular anger) have also been identified as key antecedents of collective actions.

Once these elements are in place, organizational processes involved in the mobilization of protesters explain collective action. Mobilization processes turn sympathizers into active participants and transform grievances into political claims. Through them, movements attempt to disseminate their analysis of the situation, create moral outrage among participants, and point out targets. Discursive approaches highlight identity functions of mobilization discourse: political communicators may attempt to regroup a heterogeneous audience into a single rhetorical entity ('we are the 99%') or to present a political program as an instantiation of the dominant norms and values of a given identity category ('we all welcome foreigners in our community'). Finally, researchers now recognize the importance of virtual networks and social media such as Twitter and Facebook in shaping 'connective action,' but the actual workings of these new forms of mobilization leading to coordinated collective action are currently far from clear.

See also: Emotions and Intergroup Relations; Immigration: Social Psychological Aspects; Intergroup Relations; Racism: Social Psychological Perspectives; Representations, Social Psychology of; Social Constructionism; Social Dominance Orientation; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper Row, New York.

Borgida, E., Federico, C.M., Sullivan, J.L. (Eds.), 2009. The Political Psychology of Democratic Citizenship. Oxford University Press, Oxford, UK.

Condor, S., Tileagă, C., Billig, M., 2013. Political rhetoric. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 262–297.

Cottam, M.L., Dietz-Uhler, B., Mastors, E., Preston, T., 2010. Introduction to Political Psychology, second ed. Psychology Press, New York.

Delli Carpini, X.M., Keeter, S., 1996. What Americans Know about Politics and Why It Matters. Yale University Press, New Haven, CT.

Druckman, J., 2004. Political preference formation: competition, deliberation, and the (ir)relevance of framing effects. American Political Science Review 98, 671–686.

Elcheroth, G., Doise, W., Reicher, S., 2011. On the knowledge of politics and the politics of knowledge: how a social representations approach helps us rethink the subject of political psychology. Political Psychology 32, 729–758.

Fisher, R.J., Kelman, H.C., Nan, S.A., 2013. Conflict analysis and resolution. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 489–521.

Gilens, M., 1999. Why Americans Hate Welfare. Race, Media and the Politics of Antipoverty Policy. University of Chicago Press, Chicago.

Green, E.G.T., Staerklé, C., 2013. Migration and multiculturalism. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 852–889.

Herrmann, R.K., 2013. Perceptions and image theory in international relations. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 334–363.

Howarth, C., Andreouli, E., Kessi, A., 2014. Social representations and the politics of participation. In: Kinnvall, K., Capelos, T., Dekker, H., Nesbitt-Larking, P. (Eds.), Palgrave Handbook of Global Political Psychology. Palgrave, London.

Huddy, L., 2013. From group identity to political cohesion and commitment. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 737–773.

Huddy, L., Sears, D.O., Levy, J.S. (Eds.), 2013. Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK.

Iyengar, S., McGuire, W. (Eds.), 1993. Explorations in Political Psychology. Duke University Press, Durham, NC.

Jost, J.T., Glaser, J., Kruglanski, A.W., Sulloway, F., 2003. Political conservatism as motivated social cognition. Psychological Bulletin 129, 339–375.

Jost, J.T., Sidanius, J. (Eds.), 2004. Political Psychology: Key Readings. Psychology Press, London.

Kahneman, D., 2011. Thinking, Fast and Slow. Penguin Books, London.

Kinder, D.R., 2013. Prejudice and politics. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 812–851.

Kinnvall, K., Capelos, T., Dekker, H., 2014. Nesbitt-Larking, P. (Ed.), Palgrave Handbook of Global Political Psychology. Palgrave, London.

Klandermans, B., van Stekelenburg, J., 2013. Social movements and the dynamics of collective action. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 774–811.

Knutson, J.N., 1973. Handbook of Political Psychology. Josey Bass, San Francisco.

Lau, R.R., Sears, D. (Eds.), 1986. Political Cognition: The 19th Annual Carnegie Symposium on Cognition. Lawrence Erlbaum, Hillsdale, NJ.

Marcus, G.E., 2012. Political Psychology: Neuroscience, Genetics, and Politics. Oxford University Press, Oxford, UK.

McClelland, D., 1985. Human Motivation. Cambridge University Press, New York.

McGuire, W.J., 1993. The poly-psy relationship: three phases of a long affair. In: Iyengar, S., McGuire, W.J. (Eds.), Explorations in Political Psychology. Duke University Press, Durham, NC, pp. 9–35.

Monroe, K.R., 2002. Political Psychology. Lawrence Erlbaum, Mahwah, NJ.

Rokeach, M., 1960. The Open and Closed Mind. Basic Books, New York.

Schwartz, S., 1992. Universals in the content and structure of values: theoretical advances and empirical tests across 20 countries. Advances in Experimental Social Psychology 25, 1–65.

Sears, D., Brown, C., 2013. Childhood and adult political development. In: Huddy, L., Sears, D.O., Levy, J.S. (Eds.), Oxford Handbook of Political Psychology, second ed. Oxford University Press, Oxford, UK, pp. 59–95.

Sears, D.O., Huddy, L., Jervis, R. (Eds.), 2003. Oxford Handbook of Political Psychology. Oxford University Press, New York.

Tileagă, C., 2013. Political Psychology: Critical Perspectives. Cambridge University Press, Cambridge, UK.

Political Religions: History of the Concept

Sabine A Haring, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of Western modernity must not only be regarded as a linear process of secularization. It is rather characterized by the ambiguities of the modern age, by both the 'disenchantment' and the 'reenchantment.' Worldly replacements such as science and political movements partly took over the function and the meaning of traditional religions. Political 'religions' not only adopted, transformed and amalgamated the contents of Christian texts and imaginations and used Christian rites and symbols, but also fulfilled special functions for their 'believers.' This article discusses the history of the concept 'political religion' from the period when Voegelin's book 'The Political Religions' was published in 1938 to recent works concerning the topic.

Introduction

The development of modern western societies is the result of a number of transformation processes in the fields of politics, economy, law, and culture which can be described in more detail by terms such as democratization, industrialization, urbanization, secularization, rationalization, and individualization. In the social-scientific discourse since the 1950s these processes have mostly been subsumed under the term 'modernization' (Haring and Scherke, 2000). Analogously to the predominant insight in the current debate that there is no empirical evidence for any universal modernization process concerning the different societal subsystems and therefore we have to speak of several kinds of modernity or different ways of modernization, also a concept of secularization which equates 'secularization' with increasing rationalization seems to have become obsolete. More recent research shows on the one hand that the idea of secularization being an integral element of modernization *cannot* be transferred to non-Western societies but that rather the West is the exception from the global rule. On the other hand this indicates that also the history of western modernity cannot be described as a simple, linear-progressive process of secularization but shows a variety of sometimes even contradicting tendencies of secularization and resacralization (Hildebrandt et al., 2001). Modernity cannot be described only as a period of pluralism, subjectivism and the destruction of all intellectual and moral foundations; the radical questioning of all traditional values produced also a search for unity behind particular phenomena, a search for new foundations, systems and structures. In this context, as Haring demonstrates (2008), 'reenchantment' appeared in the specific guise of traditional religions as well as worldly substitutes for religion took over the *topics* and *functions* of traditional religions. Although Norbert Elias observed in certain fields the intrusion and finally predominance of scientific explanations, in "social matters, people are exposed even today to pressures and anxieties they cannot comprehend. Since people in distress are unable to live without some explanation, the gaps in understanding are filled out by fantasy." (Elias, 1978, p. 27) (As one example of providing explanations for social distress and excitement, Elias gives the 'National Socialist myth.') Precisely in times of growing insecurity and lack of orientation – for example after the traumatic experiences during and after the First World War – many people tend toward gathering even more closely around the ancient gods, toward erecting new idols, or toward leaving themselves to the uncoordinated, chaotic course of events (see Gadshijew, 1996, p. 76).

In the interwar period the contemporaries tried to interpret and understand the development of these new totalitarian regimes. Thus, the ways in which the new revolutionary regimes referred to themselves – Communism, Fascism, National Socialism – were joined by references by others; at first travelers, journalists, and writers told about those political movements which had newly come to power. Soon one tried to find appropriate names for these new phenomena; they became the topic of an international philosophical-political debate. Finally these attempts at new concepts and terms for the new political movements resulted in larger patterns of interpretation: the concepts of totalitarianism and of political religions are considered the best-known ones (see Maier, 2003, p. 9). One tried to grasp common grounds of National Socialism and Stalinism by help of the totalitarianism concept. Authors such as Eric Voegelin, on the other hand, by employing the concept of political or secular religions, focused on the religious dimension of the totalitarian movements. As Hans Maier (p. 19) points out, the 'classical' vocabulary was most present (and was systematically renewed also at a later stage) in the Anglo-Saxon countries, whereas it looks most marginalized in Germany and Italy. In these countries the concept of dictatorship has been reactivated during the 1920s. In remaining Europe, after 1923 and after 1938 the concepts of totalitarianism and political religions have been spreading. In this context, 'mediators' of East European origin, such as Waldemar Gurian (and later Zbigniew Brzezinski), exert crucial influence. Then in the 1930 and 1940s the totalitarianism thesis conquers also the Anglo-Saxon countries. However, there it is always challenged to a certain degree. Even an important supporter of the totalitarianism concept such as George Orwell, whose attention is most of all attracted by what happens in Germany and Russia, uses the adjective 'totalitarian' as a synonym of 'tyrannical' and 'despotic.' Against the background of the new political movements and regimes the concept of 'political religion' tries to emphasize the specific changing relationships between religion and politics and especially between religion and totalitarianism. (On the different analyses of the complex

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03110-X

relationships between religion and politics see Lovin, 1992; Gebhardt, 1995.) Emilio Gentile shows that "practically every scholar of totalitarianism during the interwar period […] agree that the sacralisation of politics (variously defined as lay religion, secular religion, earthly religion, political religion, political mysticism, and political idolatry) was one of the most distinctive elements, if not the most dangerous, of the totalitarian phenomenon" (Gentile, 2000, p. 18).

Eric Voegelin

In most cases, the various contributions to the topic of 'political religions' start out from political scientist and philosopher Eric Voegelin, whose work has attracted more attention in the German-speaking countries only in the past 10 years. Voegelin – just as French sociologist Raymond Aron, with whom he corresponded for years (Weiss, 2003, pp. 41–42) – interpreted the ideological movements of the twentieth century as *one* kind of 'religion' in so far as they enriched their essential symbols with sacral substance, to *affectively* bind the masses and make them expect salvation in a way which could be politically exploited.

Voegelin's book *The Political Religions* – published by Bermann-Fischer in Vienna in the month of Austria's annexation to the Reich and just a few months before his flight to the USA – marks a break in his works. (In his early period of work, he had first of all been dealing with issues of political science, constitutional theory, legal theory and empirical cultural sciences. However, with the publication of *The Political Religions* these issues increasingly retreated to the back, and from then on Voegelin dealt predominantly with his political theology and philosophy of history. See Henkel, 1998, p. 10–11, 25 as well as 169; Porter, 2002, most of all 151–160.) As Voegelin remarked in his *Autobiographical Reflections*, the choice of this title was based on certain contemporary analyses – as examples he gave Louis Rougier's study 'Les mystiques politiques' (contemporaines et leurs incidences internationals from the year 1935; the author) – which emphasized the religious dimension of ideological movements (Voegelin, 1994, p. 70). Now, in *The Political Religion* Voegelin assumed that main experiences his/her existence as "being natural (kreatürlich) and, therefore, questionable" (Voegelin, 2000, p. 30). The wealth of the different ways and degrees in which humans perceive this and finally indeed express it "unfolds on one single dimension of religious experience, namely, in connection with the agitations of naturalness" (p. 31). In this context, whole worlds of symbols, linguistic signs and concepts organize themselves around the respective holy center. Finally they consolidate to systems that can be grasped with the help of the spirit of religious excitement and can be fanatically defended as the 'correct' order of being.

By referring explicitly to Scheler, particularly to the latter's work *Die Stellung des Menschen im Kosmos*, Voegelin started out from certain basic anthropological assumptions: both the realm of politics and the realm of religion are said to have common roots in human nature which is a unity and must be accepted as a unity. (Voegelin dissolved the antagonistic relation between religion and politics, which had developed in the course of the Modern Age.) In *Race and State* (1933), he had discussed the problem of the connection between bodily physical and spiritual-political existence (Voegelin, 1994, pp. 34–37), in *The Political Religion*, he turned to the connection between religious and political existence by trying to work out the religious nature of political mass movements – having in mind first of all National Socialism and Italian Fascism. (H.C. Kraus considers precisely the new way of defining the relation of politics and religion that what makes Voegelin's approach fruitful; and that precisely at a time "when awareness of this had almost been lost." See Kraus, 1997, p. 80.) Voegelin considered precisely the present as he experienced it to be characterized by a multitude of 'religious' orders being at work. Man, he stated, is torn apart, the old certainities have dissolved: "And wide is the space for places of the soul from which ecstasies arise, the experiences in which humans transcend their existence: from the *unio mystica* in spirit to the exaltation in the celebration of community, to the dedication to the fraternity of companions, to the loving extension of oneself into the landscape, the plants and the animals, up to carnal convulsion in the sexual act and in bloodlust." (Voegelin, 2000, p. 33)

His writings were influenced, among others, by his teachers Othmar Spann and Hans Kelsen (Voegelin, 1994, pp. 19–25, 38–41; Neisser, 2003, pp. 19–25), by Max and Alfred Weber, but also by the Stefan circle around Stefan George as well as by Karl Kraus (Porter, 2002, pp. 154–155), by Max Scheler, John Dewey and William James (Voegelin, 1994, pp. 46–48, 92–93), the philosophers and theologians of Neo Thomism – G. Sertillanges, J. Maritain and E. Gilson – as well as by Henri Bergson and the members of the so called 'Geistkreis' (Voegelin, 1994, pp. 23–25; Herz, 1996, pp. 195–204) named several elements of the symbolic forms that ca be found in all "very advanced civilizations": "the hierarchy, in which the sacral substance branches out from a transcendent God to the community of creatures; the ecclesia as the sacral communal substance; the apocalypse as the revelation of the empire; the holy kings as God's mediators and personality carries of the community" (Voegelin, 2000, pp. 70–71).

Hierarchy and Ecclesia

The clearly determined sacral hierarchy reached from the divine top via the hierarchy of rulers and offices as far as to the last obedient subject. It was symbolically expressed, for example, by the image of the sacral outpouring: Pharaoh Akhenaton's sun cult, which is said to have much influenced Plotin, the symbol of outpouring in Maimonides, the rays of light in Dante's 'Monarchia, ' and Louis XIV's sun image illustrate the continuity of symbolic manifestations over long time periods. Finally Bodin, the great theoretician of the French monarchy, is said to have rationalized the symbol of the sacral hierarchy as far as to the finest details of state and legal hierarchy. Another kind of symbol which has been transformed from being sacral to being secular is the 'ecclesia.' The process of closing off a hierarchically organized community, Voegelin (pp. 44–50) stated, had demanded that unity experienced itself as a center of existence resting within itself. However, finally the concept of Christianity – as it was developed in the epistles of Paul and the closely related Epistle to the Hebrews – understood the 'ecclesia,' the community, to be the mystical Body of the Christ

(p. 45). The offices of the 'ecclesia' were organized like the 'limbs' of the body, to then, finally, form the unity of the earthly body, with Christ being its *pneuma*. The model of the early community, as it had been drafted by Paul, could be extended by political functions of the ruler, which way the ruler's function was given its status within the *corpus mysticum*. Throughout all transformations of the Christian 'concept of the ecclesia,' Voegelin pointed out (pp. 45–47), a basic frame was kept, which became obvious both in terms of Christian Churches and the state communities. But not only there: The concept of freedom and equality served as the state religion during the French Revolution, the Pauline concept of extending the *corpus mysticum* toward mutually completing functions of the community was introduced into the theories on the division of labor of British and French economists and sociologists, and the demand for spiritual conformity as the basic form of the mystic corpus which was united by the *pneuma* was still alive in, for example, the political religion of National Socialism.

Another process which was significant for the development of inner-worldly communities, Voegelin said (pp. 47–50), was the split of the 'ecclesia' into the 'spiritual' and 'temporal' lines: as examples of this Voegelin gave St. Augustine's distinction of the heavenly and the earthly civitas as well as St. Thomas Aquinas's definition of the relation of the functions of Church and Prince. Finally, the contradictions between 'spiritual' and 'temporal' had become meaningless, after the occidental 'ecclesia' had dissolved into state subcommunities. The 'ecclesia' became one particular community out of others; state and Church formed a close unity which made any distinction obsolete, although not all parts of Europe experienced this development at the same time. "In the course of the political-religious history of Europe," as Voegelin emphasized, "the particular ecclesia developed further into a closed, inner-worldly ecclesia and also gave rise to the corresponding empires of Satan." (p. 58) After all, the symbolism of the completely secluded inner-worldly 'ecclesia' needed only one short step beyond the symbol of the Leviathan, although now the connection to God had been cut and replaced – in its function as a legitimation of the community as a person – by the community itself. (For Voegelin the great theologian of the particular 'ecclesia' was Thomas Hobbes, the creator of Leviathan. See Voegelin, 2000, pp. 53–58.)

For the explanations on hierarchy and 'ecclesia,' history was observed so to speak from the outside, as a sequence of changes of a world of religious symbols. By a next step, in his explanations on the apocalypse, Voegelin discussed the "interpretation of the development from inside, from the standpoint of the people and powers involved" (p. 50).

Apocalypse

The basic structure of a religious interpretation of history had already been found in the Pauline classification of world history into three areas (p. 50): the *lex naturalis*, the *lex mosaica* of the Old Testament, and third the Christian empire; however it was further developed by being enriched with topics which were of significance for each respective period. In this context the topic was understood to be an expression of God's will in history, the Holy Numbers functioning as a rational structure: the 'three' of the Trinity, the 'seven' for the gifts of the Holy Spirit, the hexameron, the 6 days of creation, the Old Testament's 42 generations from Abraham to Christ, and half of that number as represented by the 21 generations before Abraham. According to Voegelin (pp. 51–52), it was Joachim of Fiore (around 1132-1202) who, by his Trinitarian eschatology, defined the symbols which, up to our days, determine the self-interpretation of the modern political community: firstly, understanding history as a sequence of three ages, the last of which will be the Third Empire – an idea which is found both in the humanist and encyclopedic structuring of history into antiquity, the Middle Ages and the Modern Age and in the models of Turgot and Comte, of Hegel and Marx as well as in the National Socialist symbol of the Third Reich; secondly, the symbol of the leader, which is found, for example, with the sect movements during the Renaissance and the Reformation as well as in Dante or Condorcet, in Comte and Marx; thirdly, the symbol of the prophet of the new age which implicitly assumes that history as a whole is open to human recognition – may it be by way of revelation or speculative gnosis; and fourthly, the symbol of the brotherhood of autonomous individuals, meaning the abandoning of institutions, still had the abandoning of the Church in mind, which as a mediator for grace and administer of the sacraments will finally become obsolete because in the third age the humans will have direct access to salvation. This idea of a community of perfected humanity without any institutional authority was soon found – in various ways – with the sects of the Middle Ages and the Renaissance as well as with the Puritan Churches of the Saints (see Voegelin, 1999, pp. 113–116). Finally, the immanentization of the meaning of history, which was started already by Joachim, was fully developed by the process of secularization with thinkers such as Comte and Marx, where finally the meaning of history became a completely inner-worldly phenomenon without any transcendental breaks. In the nineteenth century, Voegelin says, 'metaphysics' became a bad term and religion was discredited as the 'opium of the people' or unmasked as an 'illusion.' However 'counter-formulas' to the 'spiritual religions': the 'scientific Weltanschauungen,' 'scientific Socialism,' the 'scientific race theory' (p. 60) developed. But as the humans are not able to make the problems of their existence disappear, even if God has become invisible behind the modern concept of the world, the topics of the world have become new gods. According to Voegelin (pp. 60–62) both, Christian and inner-worldly religions, have their own apocalypse, although the latter takes the symbols it creates from the stock of the sciences. Here, the Final Empire is no longer a heavenly community of the spirit but an earthly state of perfected mankind. Kant, Fichte, Comte and Marx developed models which were appropriate to this postulate and also created the appropriate symbolism of the devil.

Belief (Glaube)

Finally, at the end of *The Political Religions*, Voegelin returned to his initial consideration: from the experience of being a creature there develops the longing to go beyond the conditions of human existence and to be dissolved and redeemed by something which is beyond man, greater than man and more perfected than man. Analyzing Gerhard Schumann's *Lieder vom Reich*, one of the "strongest expressions of political-religious

agitation" (p. 67), he sketched those movements of the soul "that form the basis for the symbols and historical reality of the inner-wordly community" (p. 67). As forms of religious excitement, they are rooted "in the experience of creatureliness, but the *realissimum* in which they find redemption are not spiritual but enchored in desire, and they end in the bloodlust of the deed" (p. 67) (By Schumann, Voegelin selected a National Socialist *par excellence* who is considered an exemplary representative of the young generation of National Socialists, in particular of a group of authors born between 1905 and 1914, among whom there counted a. o. Hans Baumann, Herbert Böhme, Kurt Eggers, Herybert Menzel, Eberhard Wolfgang Möller, Hans Jürgen Nierentz and Baldur von Schirach. See Vondung, 1997, p. 22.).

Voegelin interpreted National Socialism as a religion less because it had adopted symbols of the Christian language and cultic practices of the Christian Church but rather due to the fact that National Socialism was based on religious experiences which, he said, had led to the manifestation of a new *realissimum*. That subcontent of the world which the National Socialists had made their *realissimum* was the ethnic community (*Volksgemeinschaft*) as a unity of common blood, which Voegelin called a particular 'ecclesia' in contrast to the universal Christian 'ecclesia.' Already in the late 1930s his analyses made aware of the *ambivalences* of the Modern Age. Voegelin did not revise the thesis according to which the rise of the ideological mass movements was based on essential religious needs of broad parts of the population, which were met by these ideologies, even if later, in his book *Science, Politics and Gnosticism* he called the term 'political religions' a 'makeshift solution' (Voegelin, 1994, p. 70) which now had to be replaced by 'Gnosticism.' (During his *Walgreen Lectures* at the University of Chicago, as Henkel points out (1998, pp. 49, 75), Voegelin had for the first time developed the Gnosticism thesis by its more extended theoretical and historical embedding. He used the category of Gnosticism, by which originally one meant religious sects on the Mediterranean in late antiquity, a. o. to characterize ideologies and – following Max Weber – based his interpretation of history on it, in the sense of a 'pneuma-pathological idealtype' (G. Weiss) which found expression by many realizations.) After all, he interpreted the political mass movements such as National Socialism and Italian Fascism as expressions and symptoms of a severe crisis of western civilization, as heresies which do not do justice to the essence of man.

From Rudolf Otto to Waldemar Gurian

That modern political movements can be described and categorized with the help of religious categories can only be explained against the background of religious-philosophical and phenomenological research which started already before the First World War. Here the concept of religion – if we just consider Otto (*Das Heilige – Über das Irrationale in der Idee des Göttlichen und sein Verhältnis zum Rationalen*, published at Breslau in 1917), van der Leeuw or, later, Eliade – also included features of the 'numinous,' of the fascinating-horrifying and of the provoking (see Gentile, 2000, pp. 29–30): "The horrifying, the tremendum and fascinosum are rediscovered as aspects of religious experience." (Maier, 1996, p. 244) But Voegelin was the first theoretician, who developed systematically the pattern of 'political religions,' to describe the common grounds of modern mass movements, even if earlier ways of using the term 'political religion' are found already in Tommaso Campanella, Daniel Clasen and Christoph Martin Wieland (see Seitschek, 2003, pp. 109–120). Shortly after the publication of Voegelin's study, also other authors pointed out to the religious elements of National Socialism (see Vondung, 1997, pp. 33–34). In 1939 Kenneth Burke stated the frequent appearance of religious images and symbols in Hitler's language and called Hitler's thought pattern a perverted or caricatured kind of religious thought. Hermann Rauschning identified the roots of the National Socialist striving for world power in the Catholicity of the new belief in the Führer as embodying God. Walter Künneth interpreted National Socialism as an imitation of the Catholic Church, and Albert Camus in *L'Homme révolté* described the features of the divinization of the irrational and of National Socialist mysticism. Finally, Friedrich Heer analyzed the political 'religiosity' of Adolf Hitler as someone coming from the Austrian-Catholic milieu. In her essays, Lucie Varga had characterized National Socialism as a political religion already in 1937. Until her early death in 1941 this medievalist had dealt with the heretic movements of the Cathar and Albigenses. She wrote a number of essays, reports and a serial novel where she tried to describe and understand the atmosphere and 'religious' ideology of National Socialism. Other than Voegelin, for whom – embedded in a universal-historical theory of consciousness – National Socialism was *one* example of a gnostic mass movement or political religion, Varga was interested in concrete individual analyses which partly – like her studies on the current sociopolitical change in rural areas (Montafon Valley and Enneberg Valley) – were based on her own field research. The topic of political religion, as Schöttler points out, is basically found in all of Lucie Varga's texts, independently of her subject and her kind of discourse. More or less explicitly it is always about phenomena of conversion or identification, of orthodoxy or heresy, and not at last about those 'scapegoats' into whom in times of crisis all unhappiness can be so wonderfully projected. Outsiders, former leaders and fallen angels, heretics, witches or devils, and finally: "the Jews" (Schöttler, 1997, p. 192). When interpreting National Socialism, we may identify interesting appeals and parallels to Varga's observations, both in her husband, Franz Borkenau, and also in her friend and teacher, Lucien Febvre. Borkenau emphasized the religious dimension of National Socialism particularly in his books *The New German Empire* (1939) and 'The Totalitarian Enemy' (1940), even if – different from Varga – he never spoke of concrete religious rituals or experiences of conversion but instead of 'Nazi psychology' in the sense of a paranoid way of thinking which turned against everything and everybody (see Burleigh, 2000, pp. 3–4). Febvre asked – beyond 'intellectuality' – about the moods which came to bear by the National Socialist turnaround (see Schöttler, 1997, pp. 192–197). In 1938, also Luigi Sturzo in *Politica e Morale* had pointed out to the religious components of totalitarian movements. According to Sturzo, the entire common life is permanently mobilized by way of parades, feasts, and sports events, to create a common consciousness (*l'unanime consenso*) and to stimulate a collective atmosphere

of enthusiasm and excitement. The entire cult of the state, he says, is oriented at a leader – Hitler, Mussolini, Lenin or Stalin – who finally becomes a 'semigod.' However, he says, the thus addressed permanent mobilization of the people can only be achieved by way of a 'sfruttamento del sentimento religioso' (abusive exploitation of the people's religious feelings). Sturzo had used this formulation already in 1925. For Sturzo, a Catholic priest and the founder of the 'Partito Populare Italiana,' the separation of state and religion was the key element of the totalitarian problem. The history of modern politics, he says, is a history of the increasing idolization of the state. God is replaced by the state. It concentrates the religious feelings of its citizens, to then finally, with the help of the means of modern propaganda, exploits them for its own purposes: here Sturzo's observations are similar to Eric Voegelin's (see Schäfer, 1996, pp. 43–44). In the context of totalitarian regimes, also Waldemar Gurian spoke of 'political religion' and later of 'totalitarian religion' which, he says, is created by controlling the public opinion: "Its content is not adoration of a ruler or a dynasty but the mass which is embodied in the leader and the ruling elite. This new religion does not only demand outward participation but inner approval." (Hürten, 1996, p. 64) The post-World War I totalitarian movements – as Gurian writes in his article 'Totalitarian religions,' published in *Review of Politics* in 1952 – are fundamentally 'religious' movements striving for reshaping the nature of man and society. In this context they claim the only true and necessary knowledge of life and its meaning. According to their claim to absolute validity, everything not being congruent with their ideology must be 'cleared,' if not in reality, then at least in the people's consciousness (see Hürten, 1996, p. 68).

From Raymond Aron to Emilio Gentile

The fact that National Socialism and Communism show religious aspects or at least aspects which are similar to religion and that they employ certain ways of presentation which are related to those of the Christian Churches attracted the attention of so called 'theoreticians of totalitarianism' such as Raymond Aron or Hannah Arendt. Whereas for Voegelin, notwithstanding a strong emphasis on positive religion, after all the development of modern totalitarianism was a consequence of a vacuum due to secularization, that is the result of a decay process, for Aron who, as a university student at the University of Cologne, witnessed Hitler's rise in Germany in the early 1930s, it was due to incomplete Enlightenment. On Aron's stay in Germany as well as on his analysis of National Socialism see Aron, 1993.) For Aron, modern totalitarianism symbolized a residual of the past (see Maier, 2003, p. 14). Already in 1936 Aron hinted, as Seitschek shows (2003, p. 161), that in the case of German National Socialism he had identified religious features of totalitarian regimes. Three years later, in 1939, in a review of Halévy's *L'ère des tyrannies*, he speaks explicitly of 'political religions.' In 1941, once again he mentions the term 'political religions,' to characterize the phenomena of totalitarian mass movements. Later, from 1944 on, however, Aron speaks of 'secular religions, ' he keeps this term and reaches back to it long after the Second World War, for example in 1979, in various essays. Aron understood political or secular religions to be imitations of traditional religion and at the same time to be sharpening with increase of ideology, as Seitschek demonstrates (2003). He identified its following characteristics: firstly, making their own doctrine a dogma and reformulating the topics of their own systems into fixed tenets which finally claimed absolute truth; secondly, the concentration of power in the hands of one party or movement which in most cases culminated in an almighty, prophetic leader; thirdly, the construction of inner and outer enemies who embodied all that opposed their own, 'true' doctrines and were thus to be 'destroyed'; and fourthly, the intention to create not only a new world order but also a new kind of man. Secular religions attempt to redeem mankind but only after an apocalypse at the end of a distant future and by reshaping and restructuring society. Thus, they restructure the entire course of history in their favor by interpreting it as the embodiment of values. Modern mass propaganda help secular religions to succeed by reaching broad parts of the population, supplying them with the topics of their ideologies and thus indoctrinating them. Already in 1932, Aron saw that the masses had been stirred up by a collective belief of a religious kind – i.e., the belief in Adolf Hitler who had prophesized a chiliastic future for Germany. Later in his book *The Opium of the Intellectuals* (in French published in 1955) he dealt with the religious dimensions of Marxism, in particular with its attractiveness for intellectuals (Aron, 1957, p. 9). (Here Aron reaches back to Marx's terminology, who had called religion the 'opium of the people.' See Aron, 1957, p. 348.) In Marxism, he says, passions which otherwise would have been lived out in the religious sphere are subjected to political goals. An age of salvation is prophesized: There will be classless society, first formed by a privileged avant-garde of intellectuals. The Party as a Church includes the avant-garde and generates those dogmas that become valid also for the masses. Apart from classless society, the final goal is the creation of New Man. The doctrine of the party demands absolute obedience, the 'not-chosen-ones' being enemies who must be fought with the help of violence.

Whereas up to the 1970s Bolshevism, Fascism and National Socialism had predominantly been dealt with as political phenomena and one attempted to structure them by help of totalitarianism or Fascism theories, since then – after the Holocaust had become a focus of contemporary research – historical considerations have increasingly taken their 'religious' dimension into consideration, eventhough applying the concept of religion to political phenomena was often rejected (see Maier, 1997, p. 299). By reaching back to authors such as Eric Voegelin and Raymond Aron, one increasingly paid attention to the 'religious' elements of the National Socialist movement, worked out analogies and differences to the Christian religion and looked at the transformations of Christian topics and symbols into secular ones. Exemplarily, here we may refer to the works by Michael Burleigh, Hans Maier, Klaus Vondung, Claus-Ekkehard Bärsch, and Michael Ley (see Haring 2008: 297). In the 1990s, after the fall of the Iron Curtain, we observe not only a revival of the totalitarianism theory but also – connected, among others, to a real 'Voegelin Renaissance' – a strong research interest in the 'religious' aspects of the totalitarian movements of the twentieth century, in the context of which it is probably true that political religions as a concept or typology or type do not rule out the concept of totalitarianism

and are also not identical with it but grasp only part of the phenomenon (see Gentile, 2000, pp. 19–21). As Haring points out (2008), political 'religions' not only adopted, transformed and amalgamated the contents of Christian texts and imaginations (in the area of dogma, eschatology, apocalypse and messianism) and used Christian rites and symbols, but also fulfilled special functions for their 'believers.' One of the most important functions both of traditional religions and political religion is to evoke emotions and to give opportunities to act on emotions (see Haring, 2008a). Political religions succeed – like Eric Voegelin said – in activating people emotionally. For many authors such as Emilio Gentile their first manifestations of the sacralization of politics in the modern sense of the term are found during the American and French Revolution (see Gentile, 2000, pp. 29, 34–35), when a "political movement confers a sacred status on an earthly entity […] and renders it an absolute principle collective existence, considers it the main source of values for individual and mass behaviour and exalts it as the supreme ethical precept of public life." (pp. 18–19). (See Gentile, 2000, p. 21: "The term, the 'socialization of politics' means the formation of a *religious dimension in politics that is distinct from, and autonomous of, traditional religious institutions*." But Gentile emphasizes that "the relationship between the socialization of politics and traditional religion is, in reality, a very complex one.") And the "myth of the nation and revolutionary faith became the driving force behind the sacralisation of politics over the 200 years that followed" (p. 35). Distinguishing from the concepts of "political theology" or "civil religion" the concept of 'political religions' has steadily got more significance for the analyses of kinds of ideologies and the everyday political practice of Communism and National Socialism precisely in the past few years. To "be able to understand the mentality of that worldwide civil war: the readiness to kill and die in the service of people, class or race, 'political religion' has meanwhile become key category" (Berghoff, quoted in Albrecht, 1999, p. 10).

Acknowledgment

This article was translated from German into English for the most part by Mirko Wittwar, to whom I would like to say sincerely 'thank you.'

See also: Aron, Raymond (1905–83); Communism, History of; First World War, The; National Socialism and Fascism; Secularization: History of the Concept; Totalitarianism.

Bibliography

Albrecht, C., 1999. Ist die Säkularisierungstheorie eurozentrisch? Zur globalen Bedeutung politischer Religion. DVPB aktuell 3, 9–12.

Aron, R., 1957. Opium für Intellektuelle oder die Sucht nach Weltanschauung. Kiepenheuer & Witsch, Köln-Berlin.

Aron, R., 1993. In: Stark, J. (Ed.), Über Deutschland und den Nationalsozialismus. Frühe politische Schriften 1930–1939. Leske & Budrich, Opladen.

Bärsch, C.-E., 1998. Die politische Religion des Nationalsozialismus. Die religiöse Dimension der NS-Ideologie von Dietrich Eckart, Joseph Goebbels, Alfred Rosenberg und Adolf Hitler. W. Fink, München.

Brocker, M., et al., 2003. Einleitung: Religion – Staat – Politik: Zur Rolle der Religion in der nationalen und internationalen Politik. In: Brocker, M., et al. (Eds.), Religion – Staat – Politik. Zur Rolle der Religion in nationalen und internationalen Politik. Westdeutscher Verlag, Wiesbaden, pp. 9–30.

Burleigh, M., 2000. National socialism as a political religion. Totalitarian Movements and Political Religions 1 (2), 1–26.

Elias, N., 1978. What Is Sociology? (Translated by S. Mennell, G. Morrissey, with a Foreword by R. Bendix) Hutchinson & Co, London.

Gadshijew, K., 1996. Betrachtungen über den russischen Totalitarismus. In: Maier, H. (Ed.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs, vol. 1. Schöningh, München-Wien-Zürich, pp. 75–80.

Gebhardt, J., 1995. Politik und Religion. In: Nohlen, D., Schultze, R.-O. (Eds.), Lexikon der Politik, vol. 1. Beck, München, pp. 435–442.

Gentile, E., 2000. The sacralisation of politics. Definitions, interpretations and reflections on the question of secular religion and totalitarianism. Totalitarian Movements and Political Religions 1 (1), 18–55.

Haring, S.A., Scherke, K., 2000. Einleitung. In: Haring, S.A., Scherke, K. (Eds.), Analyse und Kritik der Modernisierung um 1900 und um 2000. Passagen, Wien, pp. 11–32.

Haring, S.A., 2008. Verheißung und Erlösung. Religion und ihre weltlichen Ersatzbildungen in Politik und Wissenschaft. Passagen, Wien.

Haring, S.A., 2008a. Mysterium tremendum und mysterium fascinans. Die emotionale Dimension politischer 'Religionen'. Österreichische Zeitschrift für Soziologie (ÖZS) 33 (2), 43–62.

Henkel, M., 1998. Eric Voegelin zur Einführung. Junius, Hamburg.

Henningsen, M., 2000. Editor's Introduction. In: Henningsen, M. (Ed.), The Collected Works of Eric Voegelin, Modernity without Restraint, vol. 5. University of Missouri Press, Columbia-London, pp. 1–17.

Herz, D., 1996. Der Begriff der "politischen Religionen" im Denken Eric Voegelins. In: Maier, H. (Ed.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs. Schöningh, München-Wien-Zürich, pp. 191–209.

Hildebrandt, M., et al., 2001. Einleitung: Säkularisierung und Resakralisierung in westlichen Gesellschaften. Ideengeschichtliche und theoretische Perspektiven. In: Hildebrandt, M., et al. (Eds.), Einleitung: Säkularisierung und Resakralisierung in westlichen Gesellschaften. Ideengeschichtliche und theoretische Perspektiven. Westdeutscher Verlag, Wiesbaden, pp. 9–28.

Hürten, H., 1996. Waldemar Gurian und die Entfaltung des Totalitarismusbegriffs. In: Maier, H. (Ed.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs, vol. 1. Schöningh, München-Wien-Zürich, pp. 59–73.

Kraus, H.-C., 1997. Eric Voegelin redividus? Politische Wissenschaft als Politische Religion. In: Ley, M., Schoeps, J.H. (Eds.), Der Nationalsozialismus als politische Religion. Philo, Frankfurt am Main, pp. 74–88.

Lovin, R., 1992. Religion and politics. In: Hawkesworth, M., Kogan, M. (Eds.), Encyclopedia of Government and Politics, vol. 1. Routledge, London-New York, pp. 521–533.

Maier, H., 1996. Konzepte des Diktaturvergleichs: "Totalitarismus" und "politische Religionen". In: Maier, H. (Ed.), "Totalitarismus" und "Politische Religionen". Konzepte des Diktaturvergleichs, vol. 1. Schöningh, München-Wien-Zürich, pp. 233–250.

Maier, H., 1997. 'Politische Religionen' – Möglichkeiten und Grenzen eines Begriffs. In: Maier, H., Schäfer, M. (Eds.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs, vol. 2. Schöningh, Paderborn-München-Wien-Zürich.

Maier, H., 2003. Einführung: Zur Deutung totalitärer Herrschaft 1919–1989. In: Maier, H. (Ed.), Totalitarismus und Politische Religionen. Konzepte des Diktaturvergleichs, Deutungsgeschichte und Theorie, vol. 3. Schöningh, Paderborn, pp. 9–28.

Neisser, H., 2003. Eric Voegelin und die Zwischenkriegszeit. In: Ley, M., et al. (Eds.), Politische Religion? Politik, Religion und Anthropologie im Werk von Eric Voegelin. W. Fink, München, pp. 19–25.

Porter, C.F., 2002. Eric voegelin on nazi political extremism. Journal of the History of Ideas 63 (1), 151–171.

Schäfer, M., 1996. Luigi Sturzo als Totalitarismustheoretiker. In: Maier, H. (Ed.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs, vol. 1. Schöningh, München-Wien-Zürich, pp. 37–47.

Schöttler, P., 1997. Das Konzept der politischen Religionen bei Lucie Varga und Franz Borkenau. In: Ley, M., Schoeps, J.H. (Eds.), Der Nationalsozialismus als politische Religion. Philo, Frankfurt am Main, pp. 186–205.

Seitschek, H.O., 2003. Frühe Verwendungen des Begriffs "Politische Religion": Campanella, Clasen und Wieland. In: Maier, H. (Ed.), Totalitarismus und Politische Religionen. Konzepte des Diktaturvergleichs, Deutungsgeschichte und Theorie, vol. 3. Schöningh, Paderborn, pp. 109–120.

Voegelin, E., 1994. Autobiographische Reflexionen. In: Opitz, P.J. (Ed.). W. Fink, München.

Voegelin, E., 1999. Der Gottesmord. Zur Genese und Gestalt der modernen politischen Gnosis. In: Opitz, P.J., Hollweg, T. (Eds.). W. Fink, München.

Voegelin, E., 1995. Die Größe Max Webers. In: Opitz, P.J. (Ed.). W. Fink, München.

Voegelin, E., 2000. The political religions (1938). In: Henningsen, M. (Ed.), The Collected Works of Eric Voegelin, Modernity without Restraint, vol. 5. University of Missouri Press, Columbia – London, pp. 19–73.

Vondung, K., 1997. "Gläubigkeit" im Nationalsozialismus. In: Maier, H., Schäfer, M. (Eds.), "Totalitarismus" und "Politische Religionen." Konzepte des Diktaturvergleichs, vol. 2. Schöningh, Paderborn-München-Wien-Zürich, pp. 15–28.

Weiss, G., 2003. Die europäische Rezeption Eric Voegelins. In: Ley, M., et al. (Eds.), Politische Religion? Politik, Religion und Anthropologie im Werk von Eric Voegelin. W. Fink, München, pp. 26–46.

Political Representation

Dennis F Thompson, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The process of political representation involves relationships not only between representatives and their constituents, but also among the constituents and among the representatives. Each set of relationships raises fundamental issues of democracy and has stimulated extensive research in both normative political theory and empirical political science. One of the most notable developments is the significant expansion in the number and variety of individuals and groups who claim representation.

Political representation is a process in which one individual or group (the representative) acts on behalf of other individuals or groups (the represented) in making or influencing authoritative decisions, policies, or laws for a polity.

The representative is typically a legislator, but not exclusively. A growing body of research recognizes that a representative may hold any of a number of other offices (such as executive, administrator, ambassador, judge, lobbyist, party leader) and may even be a collectivity (a governmental agency or political organization, international as well as domestic) (Urbinati and Warren, 2008). Individual citizens themselves may act as representatives when brought together in citizens' assemblies, a relatively novel institution that has been established in Canada, the Netherlands and other countries in recent years (Warren and Pearse, 2008). There are also 'self-authorized representatives,' individuals, and organizations that do not hold official positions but have some claim to speak and act for others (Urbinati and Warren, 2008).

The represented may be almost any individual or group, but they are usually understood to be those individuals or groups, typically called constituents, who are entitled to select the representatives. As the variety and range of the represented has expanded vastly in modern times, new issues for the theory of representation have arisen (Urbinati and Warren, 2008; Rehfeld, 2005; Shapiro et al., 2010). When the constituents from several nations are represented in the same government, as in the European Union, the question of who is to be represented becomes especially problematic (Thompson, 1999; Kroger and Friedrich, 2012).

The process of representation involves relationships (1) between the representatives and the represented, (2) among the represented, and (3) among the representatives. With respect to (1), the issue was once framed as a dichotomy between delegate and trustee conceptions of representation – whether representatives follow their own judgment or their constituents' preferences. But it is now recognized that the dichotomy is too simple and the literature distinguishes many different modes of this relationship (Pitkin, 1967; Thompson, 1988; Rehfeld, 2009).

As for (2), the key question concerns the criteria by which constituencies are defined (Rehfeld, 2005). Although many different criteria (including gender, ethnicity, religion occupation, and class) have been proposed and applied in various systems, the most common in modern electoral systems are party and geography. One of the most fruitful but controversial areas of inquiry raises issues of representation of identity groups of various kinds (Williams, 1998).

With respect to (3), scholars now recognize that the interactions among representatives are no less important than the linkages between representatives and constituents, which tend to receive the larger share of scholarly attention (Thompson, 1988). The increased polarization, especially in the US Congress, and the decreased capacity for compromise have prompted several new lines of research (see Section The Performance of the Representative Body below).

Origins and Development

Political representation arose not as a democratic but as a monarchical and aristocratic practice in medieval government in Europe. Both the theory and practice evolved over a long period from about the fourteenth to seventeenth century as kings and nobles convened assemblies to collect revenue and conduct wars. The bodies, usually composed of members from various estates or classes, gradually assumed independent power and came to be thought of as representative institutions. In the debates during the Civil War period in England, defenders of Parliament connected the idea of representation to the idea of consent of the governed. In 1651, the members of Parliament were for the first time referred to as 'representatives' (Pitkin, 1967; also see Larsen, 1966; Manin, 1997).

By the eighteenth century, the link between representation and popular rule (though not democracy) had become sufficiently close that Baron de Montesquieu could praise the English constitution for establishing a republic in a large state by giving the people a voice through their representatives in the legislature. Defending the new American constitution, James Madison argued that representation not only made republics possible in large territories, but also mitigated their dangers, especially the influence of factions. It was John Stuart Mill who in the late nineteenth century provided the framework for the theoretical synthesis that combined the traditional virtues of representation and the emerging ideals of democracy (Thompson, 1976).

Representative government was not yet representative democracy, because much of the adult population neither served as a representative nor counted as the represented. The practice – notably the expansion of franchise that is essential to

making representation democratic – lagged behind the theory. Universal suffrage came slowly and fitfully in the United States. Barriers to participation and biases in influence persist today. Even the so-called advanced democracies cannot yet claim to be inclusively representative. Furthermore, as the scope of government has expanded and the interactions among societies increased, many people are affected by the actions of governments over which they have no formal control or practical influence. Despite its elitist origins and democratic deficiencies, representation is now defended as a valuable component of democracy (Urbinati, 2008).

The Role of the Representative

The relationship between the representative and the represented is usually analyzed in terms of the role of the representative. It is sometimes called dyadic representation to differentiate it from the more complex relationship that exists between all representatives in the legislature and the electorate or public as a whole (Ansolabehere and Jones, 2013). (Compare 'dynamic representation': Stimson et al., 1995).

Three aspects of the role of the representative should be distinguished. The first concerns the *mode* of representation – how constituents are represented. To what extent does a representative express the views of constituents, attempt to change their views, and by what methods and over what time periods? The second aspect is the *locus* of representation – which citizens are regarded as constituents by the representative? (This aspect involves the identification of the represented.) Among the possibilities are the nation as a whole, a political party, the state or district, and various groups and organizations. The district is more complex than often assumed. For example, many legislators are bound to local elites and national groups whose interests may diverge from those of other individuals and groups in the district; some legislators see themselves as representatives of underrepresented groups, such as blacks and women. The third aspect refers to variation in the *content* of representation – the types of issues with which a representative deals. A legislator may act, for example, as a delegate of the district majority on welfare issues, as a delegate of the national majority on civil rights, but as a trustee of both on foreign policy.

Much of the empirical literature on the linkages between the policy preferences of constituents and legislators has been influenced by the seminal model presented by Miller and Stokes (1963), which highlights four major variables: legislators' personal attitudes, roll call votes, constituency opinion, and legislators' perceptions of constituency attitudes. The vast literature on this linkage shows that the conditions under which each has an effect are quite variable. It is now recognized that roll call votes fail to capture much important legislative activity and do not always reflect the actual views of legislators.

Given this complexity, it is not surprising that most normative theories today decline to privilege any one of the aspects of the roles: what each representative should do depends on the nature of the particular issues, what other representatives are doing, and the biases in the political system. Some of the most interesting normative work is revitalizing the ethics of representation (see Dovi, 2007; Beerbohm, 2012). Similarly, when political scientists examine the simpler conceptions of role, they find little general correspondence between the role and actual behavior. The most robust accounts of legislators' conceptions posit decision trees with multiple branches and a complex set of factors (Kingdon, 1981; Thomassen, 1994). In some legislatures, such as Congress, representatives specialize in particular issues and in effect delegate their decision-making to other representatives (Hall, 1998).

On any account of the role of the representative, there is the further question of how representatives can be held accountable. For Congress, Mayhew's model, which posited that representatives are motivated solely by the desire for reelection, dominated the literature for many years (Mayhew, 2004). The 'electoral connection' has been shown to explain much of the behavior of members of the House, and in modified form can be seen in earlier periods in American history as well (Carson and Jenkins, 2011). But the resurgence of partisanship and party control in recent years stands in tension with Mayhew's insistence that individual members, not parties, are the most significant unit of analysis (Aldrich and Rohde, 2001).

Beyond the United States, the methods of accountability vary depending on the political system and on assumptions about what political accountability means (Przeworski et al., 1999). There has been an 'explosion' of comparative studies that examine representational linkages in many different countries (Powell, 2004).

Since at least the eighteenth century students of legislatures have criticized the influence of money and wealth on the behavior of representatives. Concern about this influence has increased in recent years, especially in the United States. As reformers have called attention to the problem, researchers have devoted more attention to its causes and potential remedies (Lessig, 2011; Francia et al., 2003). Political scientists have not generally found strong correlations between contributions and roll call votes on broad policy areas even in the US system. But they have found that money can buy access – marginal time, energy, and legislative support (Hall and Wayman, 1990). The increasing availability of data in other countries has led to a growth in comparative research on the role of money in politics (Scarrow, 2007).

The Identification of the Represented

All legislative representation is group representation. Even in societies governed by individualist norms such as 'one person-one vote,' the definition of the constituency implicitly identifies a group or groups that are to be represented. Electoral rules and other features of the political system shape the relative weights of various groups in the representational process.

One major line of research takes parties as the main unit of representation and measures representativeness by the ratio of legislative seats to votes (Powell, 2004). Proportional systems have been found to translate seats into votes most accurately (and presumably more fairly). In the most common forms, the share of votes that each party receives determines how many of the candidates from its list win seats in the legislature. The system usually produces multiple parties, and thus gives voters a wide range of ways to express the interests they wish to have

represented. However, it also gives voters less choice about which particular candidate should serve as their representative, and in this respect differs significantly from the purer form of proportional representation, the system of the single-transferable vote, in which every successful candidate has a unanimous constituency. Because the more common forms of proportional representation rarely produce majority governments, much of the research considers how the strategies of coalition building and party discipline affect representation (Martin and Vanberg, 2011).

Geographical representation relies more on single-member plurality systems and tends to produce only two or three parties. The parties represent broader interests than do the parties in modified proportional systems, but if party discipline is weak, as in the United States, the geographical basis encourages the representation of particular regional and other interests.

No single type of representational system has been shown to be superior in all or most circumstances, and the choice generally involves a trade-off: the wider the range of interests represented in the legislature, the less the likelihood that any particular interest will be fully represented in legislation actually adopted. The most useful scholarship on these and related questions has been comparative (Norris, 2004; Farrell, 2011).

Another important line of research begins with citizens' preferences rather than with their votes, and compares their policy views to those of their representatives (again typically by party). This research begins with the Miller–Stokes paradigm but moves from the dyadic relationships (individual representatives/groups of constituents) to analyze the electorate as a whole in relation to the legislature or the legislative party (Powell, 2004). One of the most striking findings is that citizens are generally more divided than their representatives. That means that there are some voters in each party who favor positions that are supported exclusively by representatives of other parties. In this respect, a party cannot be considered as the primary form of representation, at least for these citizens (Holmberg, 2000).

In systems such as the United States, where many of the districts that define constituencies do not follow any natural geographical boundaries, the questions of how the districts are drawn and whether they produce fair representation have stimulated important normative and empirical research. Political scientists have examined the partisan and incumbency biases in various redistricting schemes (Bullock, 2010). Some of the most valuable theoretical work in this area combines empirical and normative approaches. One study has developed more analytically robust concepts of 'fair representation' in electoral systems (Beitz, 1989), and another has revived the theory of descriptive representation for 'marginalized groups' (Williams, 1998). Scholars have investigated to what extent descriptive representation supports substantive representation by studying the presence of women in parliaments (Wängnerud, 2009).

Several important studies have shown that representation (not only in the legislature) is increasingly biased in favor of wealthy constituents and corporate interests. The phenomenon is most salient in the United States. (Bartels, 2008; Hacker and Pierson, 2011; Gilens, 2012). However, a growing body of research is finding that other systems are showing similar patterns (Beramendi and Anderson, 2011; Lefkofridi et al., 2012). The causes of the inequality, and the appraisal of its effects, remain controversial. Scholars investigating what is called dynamic representation casts doubt on whether the representation is unequal in any way that should be objectionable. They find that the views of groups move in parallel, shifting left or right more or less together, and that, because representatives follow the shift, groups are treated equally. Stimson (2011) argues that this apparent conflict can be resolved if we recognize that within the broad left and right shifts, the modest biases toward the wealthy accumulate over time and produce gross inequalities in the effects of policies.

The principle that 'everyone who is affected by the decisions of a government should have the right to participate in be represented in that government,' though not without problems, is widely accepted in democratic theory and practice (Goodin, 2007). Yet it has potentially radical implications for representation in an increasingly interdependent world. As the effects of government policy spill over national boundaries, the principle and related criteria would seem to require granting some form of representation to individuals who have not traditionally been included as part of the represented. This line of thinking has led, for example, to research on new forms of accountability in environmental policy (Mason, 2008).

As national boundaries become more porous, the representation of noncitizens and expatriates has also become a more pressing issue. The debate about this issue, conducted under the rubric of transnational citizenship, focuses on enfranchisement of the residents who are not citizens and citizens who are not residents (Lopez-Guerra, 2005; Owen, 2011). More generally, scholars urge that more attention be given to what has been called 'moral constituents' – not only individuals such as foreigners who are affected by decisions in which they have no formal voice, but also future citizens who cannot be formally represented because they do not yet exist (Gutmann and Thompson, 1996; ch. 4; Thompson, 2010).

The Performance of the Representative Body

Much of the research on the relationships within legislatures and other representative bodies deals with the tension between individual and collective behavior in the legislature. The general problem concerns what is sometimes called legislative individualism, which includes the tendency of legislators to act in ways that promote their own goals without regard to the effect on collective product of the legislature (Corey, 2008). This tendency has been shown to produce a surplus of 'particularized benefits' (dams in certain regions, tariffs for local industries, and military bases in the district) and the neglect of more general public benefits (such as education, health, and the environment (Evans, 2013)). The explanations for these outcomes vary, but they are generally structural, locating the problem in the incentives of the system, rather than in the character of legislators (e.g., Crisp et al., 2004). They assume that, given the structure of the political system, individually rational behavior by legislators produces collectively suboptimal outcomes in the legislature.

If all legislators are pure individualists seeking only particularized benefits for their own constituents, how (under

conditions of majority rule and weak parties) can the legislature adopt any legislation at all, even the pork barrel kind? This puzzle has been studied by formal theorists, who propose various solutions for achieving policy equilibria in legislative institutions (Shepsle and Weingast, 1995). One solution relies on a committee system to set an agenda and limit the scope of conflict.

Shared interests and cross-cutting cleavages are still prevalent in any actual system. These are more likely to exist or to arise in legislatures that encourage or at least permit representatives to change their positions through interaction with their colleagues. This is claimed to be one of the advantages of deliberative theories of legislative representation, which allow for changes in the preferences of both representatives and the represented over time (Gutmann and Thompson, 1996). Political scientists have begun to conduct empirical studies to learn how deliberative democracy actually functions in legislatures (Steiner et al., 2004).

In the United States, the increasing polarization of Congress has drawn the attention of scholars because it is thought to be one of the causes of the deadlock and dysfunction, many ascribe to the institution in recent years. Political scientists disagree about the source of the polarization itself. Is it only elites who have become more polarized, or also the electorate? and if both, to what extent are representatives the cause? (Fiorina and Abrams, 2008; Abramowitz, 2010). The challenge of governing in a polarized legislature has revived interest in the classic problem of political compromise (Gutmann and Thompson, 2012).

See also: Campaigning: Political; Citizenship, Anthropology of; Deliberation and Democracy; Democracy: Normative Theory; Democratic Theory; Global Governance; Interest Groups; Intergenerational Justice; Law and Democracy; Multiculturalism; Politics and Participation; Proportional Representation.

Bibliography

Abramowitz, A.I., 2010. The Disappearing Center: Engaged Citizens, Polarization, and American Democracy. Yale University Press, New Haven, CT.

Aldrich, J.H., Rohde, D., 2001. The logic of conditional party government: revisiting the electoral connection. In: Dodd, L., Oppenheimer, B. (Eds.), Congress Reconsidered, seventh ed. CQ Press, Washington DC, pp. 269–292.

Ansolabehere, S., Jones, P.E., 2013. Dyadic representation. In: Schickler, E. (Ed.), Oxford Handbook on Congress. Oxford University Press, New York.

Bartels, L., 2008. Unequal Democracy: The Political Economy of the New Gilded Age. Princeton University Press, Princeton, NJ.

Beerbohm, E., 2012. In: In Our Name: The Ethics of Democracy. Princeton University Press, Princeton, NJ.

Beitz, C.R., 1989. Political Equality. Princeton University Press, Princeton, NJ.

Beramendi, P., Anderson, C.J. (Eds.), 2011. Democracy, Inequality, and Representation: A Comparative Perspective. Russell Sage Foundation, New York.

Bullock III, C.S., 2010. Redistricting: The Most Political Activity in America. Rowman & Littlefield, Lanham, MD.

Corey, J.M., 2008. Legislative Voting and Accountability. Cambridge University Press, Cambridge, UK.

Carson, J.L., Jenkins, J.A., 2011. Examining the electoral connection across time. Annual Review of Political Science 14 (1), 25–46.

Crisp, B.F., Escobar-Lemmon, M.C., Jones, B.S., Jones, M.P., Taylor-Robinson, M.M., 2004. Vote-seeking incentives and legislative representation in six presidential democracies. Journal of Politics 66 (3), 823–846.

Dovi, S., 2007. The Good Representative. Wiley-Blackwell.

Evans, D., 2013. Pork barrel politics. In: Schickler, E., Lee, F.E. (Eds.), The Oxford Handbook of the American Congress. Oxford University Press, New York.

Farrell, D.M., 2011. Electoral Systems: A Comparative Introduction, second ed. Palgrave-Macmillan, New York.

Fiorina, M.P., Abrams, S.J., 2008. Political polarization in the American public. Annual Review of Political Science 11 (1), 563–588.

Francia, P.L., Green, J.C., Herrnson, P.S., Powell, Lynda W., Wilcox, C., 2003. The Financiers of Congressional Elections: Investors, Ideologues, and Intimates. Columbia University Press, New York.

Gilens, M., 2012. Affluence and Influence: Economic Inequality and Political Power in America. Princeton University Press, Princeton, NJ.

Gutmann, A., Thompson, D., 1996. Democracy and Disagreement. Harvard University Press, Cambridge, MA.

Gutmann, A., Thompson, D., 2012. The Spirit of Compromise: Why Governing Demands It and Campaigning Undermines It. Princeton University Press, Princeton, NJ.

Goodin, R., 2007. Enfranchising all-affected interests and its alternatives. Philosophy & Public Affairs 39 (1), 40–67.

Hacker, J.S., Pierson, P., 2011. Winner-Take-All Politics: How Washington Made the Rich Richer-and Turned Its Back on the Middle Class. Simon & Shuster, New York.

Hall, R.L., 1998. Participation in Congress. Yale University Press, New Have, CT.

Hall, R.L., Wayman, F.W., 1990. Buying time: moneyed interests and the mobilization of bias in congressional committees. American Political Science Review 84 (3), 799–820.

Holmberg, S., 2000. Collective policy congruence compared. In: Miller, W., et al. (Eds.), Policy Representation in Western Democracies. Oxford University Press, New York, pp. 87–111.

Kingdon, J.W., 1981. Congressmen's Voting Decisions, second ed. Harper and Row, New York.

Kroger, S., Friedrich, D. (Eds.), 2012. The Challenge of Democratic Representation in the European Union. Palgrave Macmillan, New York.

Larsen, J.A.O., 1966. Representative Government in Greek and Roman History. University of California Press, Berkeley, CA.

Lefkofridi, Z., Giger, N., Kissau, K., 2012. Inequality and representation in Europe. Representation 48 (1), 1–11.

Lessig, L., 2011. Republic, Lost: How Money Corrupts Congress–and a Plan to Stop It. Hatchette, New York.

Lopez-Guerra, C., 2005. Should expatriates vote? Journal of Political Philosophy 13 (2), 216–234.

Manin, B., 1997. The Principles of Representative Government. Cambridge University Press, Cambridge, UK.

Mansbridge, J., 2003. Rethinking representation. American Political Science Review 97 (4), 515–528.

Martin, L.W., Vanberg, G., 2011. Parliaments and Coalitions: The Role of Legislative Institutions in Multiparty Governance. Oxford University Press, New York.

Mason, M., 2008. The governance of transnational environmental harm: addressing new modes of accountability/responsibility. Global Environmental Politics 8 (3), 8–24.

Mayhew, D., 2004. Congress: The Electoral Connection, second ed. Yale University Press, New Haven, CT.

Miller, W.E., Stokes, D., 1963. Constituency influence in Congress. American Political Science Review 57 (1), 45–56.

Norris, P., 2004. Electoral Engineering: Voting Rules and Political Behavior. Cambridge University Press, Cambridge, UK.

Owen, D., 2011. Transnational citizenship and the democratic state: modes of membership and voting rights. Critical Review of International Social and Political Philosophy 14 (5), 641–663.

Pitkin, H.F., 1967. The Concept of Representation. University of California Press, Berkeley, CA.

Powell Jr., G.B., 2004. Political representation in comparative politics. Annual Review of Political Science 7 (1), 273–296.

Przeworski, A., Stokes, S.C., Manin, B. (Eds.), 1999. Democracy, Accountability and Representation. Cambridge University Press, New York.

Rehfeld, A., 2005. The Concept of Constituency: Political Representation, Democratic Legitimacy and Institutional Design. Cambridge University Press, Cambridge.

Rehfeld, A., 2009. Representation rethought: on trustees, delegates, and gyroscopes in the study of political representation and democracy. American Political Science Review 103 (2), 214–230.

Scarrow, S.E., 2007. Political finance in comparative perspective. Annual Review of Political Science 10 (1), 193–210.

Shapiro, I., Stokes, S., Wood, E.J., Kirshner, A.S. (Eds.), 2010. Political Representation. Cambridge University Press, Cambridge.

Shepsle, K.A., Weingast, B.R., 1995. Positive Theories of Congressional Institutions. University of Michigan Press, Ann Arbor.

Stimson, J.A., Mackuen, M.B., Erikson, R.S., 1995. Dynamic representation. American Political Science Review 89, 543–565.

Stimson, J.A., 2011. The issues in representation. In: Enns, P.K., Wlezien, C. (Eds.), Who Gets Represented? Russell Sage, New York, pp. 347–360.

Steiner, J., Bächtiger, A., Spörndli, M., Steenbergen, M., 2004. Deliberative Politics in Action: Crossnational Study of Parliamentary Debates. Cambridge University Press, Cambridge, UK.

Thomassen, J., 1994. Empirical research into political representation. In: Jennings, M.K., Mann, T.E. (Eds.), Elections at Home and Abroad. University of Michigan Press, Ann Arbor, MI, pp. 237–264.

Thompson, D.F., 1976. John Stuart Mill and Representative Government. Princeton University Press, Princeton, NJ.

Thompson, D.F., 1988. Representatives in the welfare state. In: Gutmann, A. (Ed.), Democracy and the Welfare State. Princeton University Press, Princeton, NJ, pp. 131–155.

Thompson, D.F., 1999. Democratic theory and global society. Journal of Political Philosophy 7 (2), 1–15.

Thompson, D.F., 2010. Representing future generations: political presentism and democratic trusteeship. Critical Review of International Social and Political Philosophy 13 (1), 17–37.

Urbinati, N., Warren, M.E., 2008. The concept of representation in contemporary democratic theory. Annual Review of Political Science 11 (1), 387–412.

Urbinati, N., 2008. Representative Democracy: Principles & Genealogy. University of Chicago Press, Chicago, IL.

Wängnerud, L., 2009. Women in parliaments: descriptive and substantive representation. Annual Review of Political Science 12, 51–69.

Williams, M.S., 1998. Voice Trust and Memory: Marginalized Groups and the Failings of Liberal Representation. Princeton University Press, Princeton, NJ.

Warren, M., Pearse, H., 2008. Designing Deliberative Democracy: The British Columbia Citizens' Assembly. Cambridge University Press, Cambridge, UK.

Political Repression

Linda C Keith, University of Texas at Dallas, Richardson, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political repression continues to be a harsh reality in many states today. A large body of empirical literature seeks to understand why states imprison, torture, force disappearance of, or kill their own citizens. The empirical literature has demonstrated that several factors are associated with decrease in political repression; the most significant being domestic threat, especially civil war. Other factors are associated with increased repression; the most consistent being the level of democratization achieved within the state. The research suggests that commitment to rights through treaties and constitutional provisions, especially in tandem with an independent judiciary, may be one potential path to lessen repression, as is the role of international NGOs in the state.

Political repression by one's own state continues to affect almost half of the global set of countries today. For example, citizens in almost 60 countries live in a state with extensive political imprisonment, unlimited detention, and extrajudicial executions or other political murders and brutality. In more than 30 other countries, political murders and disappearances are a common part of life and may affect even those who are not politically engaged. Over the last two decades social scientists have produced a growing body of studies that systematically and rigorously examine theoretically important factors that may explain why states politically repress their own citizens. The empirical study of political repression focuses on two categories of state abuse of human rights and typically involves the creation and dissemination of large cross-national datasets that encompass the global set of countries and increasingly longer periods of time. Political repression is broadly defined as "involving the actual or threatened use of physical sanctions against an individual or organization, within the territorial jurisdiction of the state, for the purpose of imposing a cost on the target as well as deterring specific activities and/or beliefs perceived to be challenging to government personnel, practices or institutions" (Davenport, 2007b: 2). Most of the literature focuses on one of two forms of political repression separately – either addressing the more severe form of repressions, namely violations of personal or physical integrity (imprisonment, torture, killing, and disappearances) (e.g., Poe and Tate, 1994; Poe et al., 1999) or addressing the broader category of civil liberties restrictions that are typically of less severe form (e.g., Davenport, 1995). While the theoretical perspectives on the two forms of repression are largely indistinguishable, the most comprehensive examination of both categories of repression shows that while they are similarly influenced by the same agentic and structural factors, significant differences emerge with regard to the influence of domestic and external threats, an independent judiciary and constitutional provisions (Keith, 2012). Henderson (2004) makes an important contribution in expanding the traditional conceptualization and measurement of political repression to forms that are specifically targeted at females, which includes such acts as harassment, degrading treatment, denial of sanitary devices for women's hygiene, virginity examinations, hostage taking to pressure male family members, imprisonment with men, absence of female guards, threats of rape, actual rape, forced prostitution, and pregnancy.

Measuring Political Repression

Social scientists seeking to empirically measure state repression face significant measurement issues, but over time political scientists have developed several measures that capture the level, frequency, and type of political repression. Most empirical studies to date have primarily addressed the most severe form of repression – violations of personal integrity (e.g., Poe and Tate, 1994; Poe et al., 1999; Cingranelli and Richards, 1999; Keith, 2002a) – and have tended to employ the dominant indicators in the field, for example, the Political Terror Scale (PTS) which is maintained by Gibney et al. (2014) (www.politicalterrorscale.org) and is a standards-based measure of state use of political imprisonment, torture, killing, and disappearances that is operationalized as a five-point ordinal scale that ranges from one, where repression is rare, to five where repression is extensive and threatens the entire population. **Figure 1** shows the percentage of countries at each level of political repression over time. We see a significant proportion of states achieving a score of one, the lowest level of repression or no repression, which ranges from 25 to almost 35%, with a significant drop following 2001. The proportion of countries with the highest level of repression (five) represents the smallest segment of countries, ranging from 2 to 12.5%. Level four of repression represents the second smallest segment here. We see the mid-levels of repression (two and three) affect approximately the same percentage of states over time as the lowest level of repression, with some variation over time. With the addition of a new measure, the CIRI physical integrity measure, which is maintained by Cingranelli et al. (2013) (www.humanrightsdata.org), empirical studies can now examine the same four behavioral components individually or as an additive index. Some limitations and benefits adhere to each of these measures (see http://www.politicalterrorscale.org/datafiles/hrq%20pts.pdf); most scholars deal with the trade-offs by conducting parallel analyses employing both measures. Conrad and Moore's (2014) new Ill-Treatment and Torture (ITT) measures, while focusing on a single dimension of repression of personal integrity rights (torture), offers

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93178-7

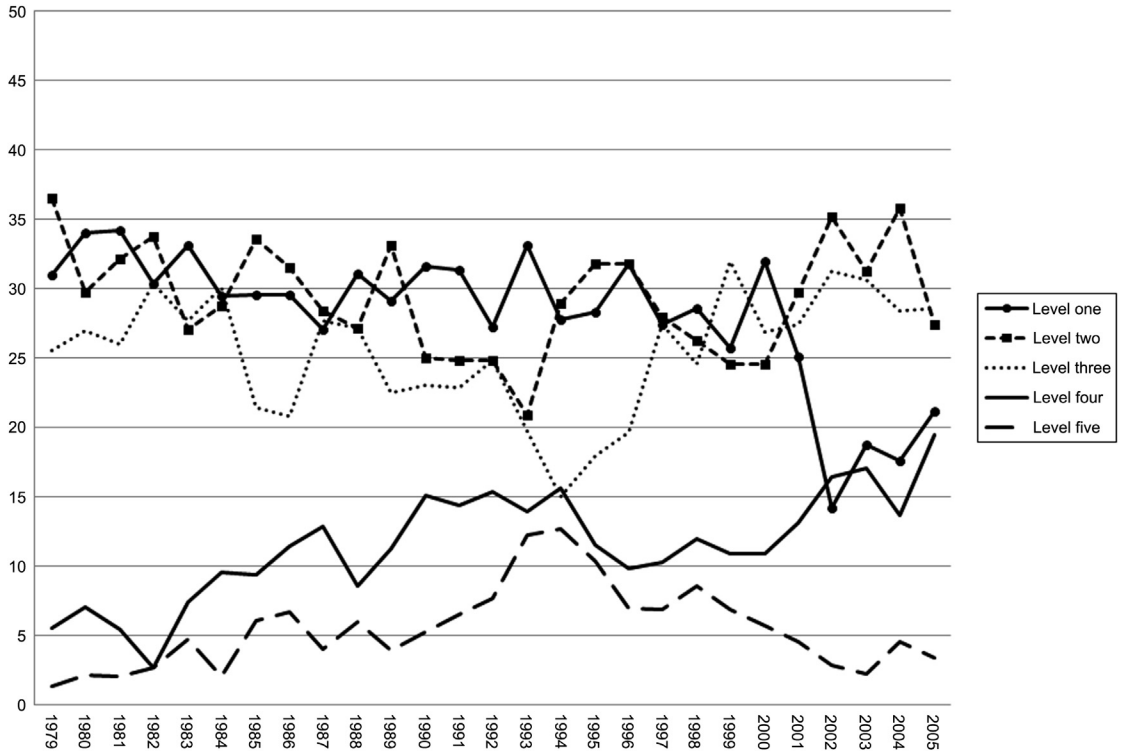

Figure 1 Political repression of Personal Integrity Rights, 1979–2005.

significant advantages over both the CIRI and PTS measures, in which it includes indicators related to the agents of torture, the victim type, and the level of torture (http://faculty.ucmerced.edu/cconrad2/Academic/ITT_Data_Collection.html).

Measuring the level of state restrictions on civil liberties has proven to be more difficult. The two earliest popular measurements of political repression, Taylor and Jodice's (1983) state coercive behavior and Freedom House's civil liberties index (http://www.freedomhouse.org/report/freedom-world/freedom-world-2014#.Uyh57vldX-o), represent significant data collection efforts, but each of them has significant limitations. Taylor and Jodice's measurements are events counts that are limited in country and time coverage. The construction of the Freedom House's civil liberties measure is unclear and the two of its four core components are problematic in that they do not specifically examine civil liberties protection per se, but rather measure circumstances and institutions that may facilitate or hinder the protection of civil liberties or deal with the abuse of these rights which would appropriately be classified as oppression of social and economic privileges falling outside the definition of repression above. Scholars now have access to new standards-based measures within the CIRI data project mentioned above that allow them to examine separately some core individual liberties such as freedom of religion, speech and press, and assembly and association. While the range of civil liberties they cover is limited and the pairing of rights is not theoretically intuitive, the measures do not suffer from many of the shortcomings of the Freedom House measure. **Figure 2** presents the mean repression scores for CIRI's freedom of speech and press, freedom of assembly and association, and freedom of religion measures. As with most studies, I have inverted the CIRI measures here to reflect repression of these freedoms, thus each of the three measures is now a three-point scale that ranges from zero to two with zero representing no repression and two the highest level of repression. We see several interesting trends. First, the most obvious pattern is that repression of religious freedom is consistently higher than repression of the other two sets of freedoms, and while there is some variation over time, there is virtually no improvement in the entire time period. Indeed in 2005 we see a higher mean of repression than the beginning date of 1981. Second, we see a clear improvement in the level of repression for the other two categories of freedom, beginning with the end of the Cold War and leveling off in the mid-1990s with an increase in repression of speech/press beginning in after 2001, which could be related to the shift in emphasis to national security priorities following September 11. We do not see the same pattern, however, for freedom of assembly and association, which during most years is repressed less than speech and press freedoms.

Dominant Theoretical Approaches to Understanding Political Repression

Most of the current empirical work on political repression has approached the state's behavior from the perspective of international relations theory, primarily utilizing a 'soft' rational choice perspective (e.g., Poe et al., 1999; and Davenport, 2007a) that assumes that political leaders are rational actors and that they choose from a menu of repressive tools because they perceive these tools to be the most effective means to

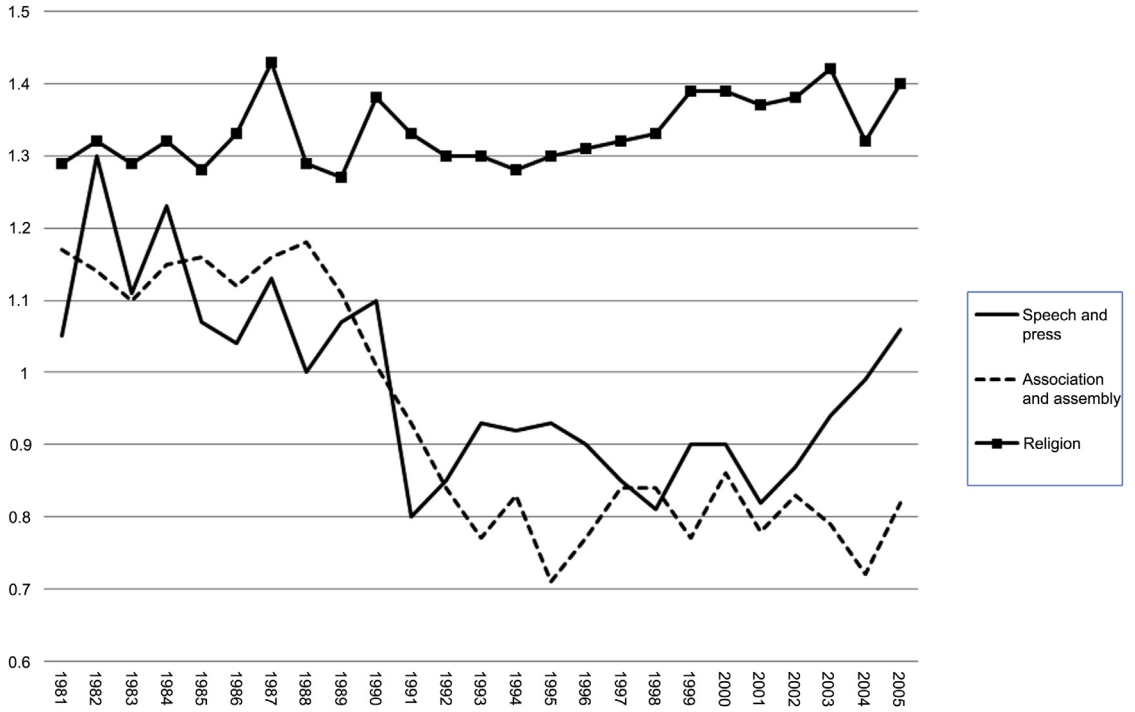

Figure 2 Mean repression of freedom of speech and press, freedom of association and assembly, and freedom of religion.

achieve their chief end, which is to stay in power. The most pervasive factor that increases leaders' willingness to employ the tools of repression is a threat to their rule, whether real or perceived, and that typically the more serious the perceived threat, the more willing state leaders are to employ repression (Keith and Poe, 2004; Keith, 2012). Or more broadly, as Davenport argues, state actors utilize actual or threatened physical sanctions in order to impose "a cost on the target as well as deterring specific activities and/or beliefs perceived to be challenging to government personnel, practices or institutions"; however, state actors carefully weigh the costs and benefits of engaging in repressive action, and also consider a menu of alternative mechanisms of control, as well assessing the odds of achieving their goals with these tools (2007a: 2–4).

The repression literature has drawn upon the theoretical insights of the domestic institutions/democratic peace perspectives – another stream within the rational actor approach. However, these perspectives dismiss the assumption of a unitary state actor and instead recognize the role of numerous domestic actors and institutions within the state (particularly in democratic regimes) which may affect the regime's menu of acceptable policy options and/or the regimes' calculation of costs and benefits related to commitment to international and national human rights norms and with subsequent (non)compliance to formal commitment to the norms. Poe et al. (1999) argue that state leaders in democratic regimes have less opportunity and less willingness to employ the tools of repression, even when facing domestic or international conflict or threats (293, and citations within). Opportunity is constrained because the structure and limited nature of democratic governments, which may include separation of powers and/or checks and balances, make it difficult for state actors to coordinate extensive use of repression. The necessity to utilize repression as a tool to solve conflict can be mitigated in democracies because they provide a variety of alternative mechanisms through which conflict can be channeled for possible resolution (Poe et al., 1999; Davenport, 2007b), in turn, weakening the state's justification for coercive state action. In addition, the need to employ the tools of repression is also dampened by "the socialization processes that guide citizens of democratic polities toward the belief that nonviolent means of resolving conflicts are preferred over violence" (Poe et al., 1999: 293). The democratic norms of compromise, toleration, and facilitation may influence the decisions of the rulers by "increasing the cost of human rights violations as well as decreasing the value to quiescence" (Davenport, 1999: 96). The cost of repression is increased in democracy as the electoral process "provides citizens (at least those with political resources) the tools to oust potentially abusive leaders from office before they are able to become a serious threat" (Poe and Tate, 1994: 855). From the domestic legal institutions perspective it is not just democratic electoral processes, but also the legal institutions associated with democratic systems that can provide the public and other political actors the tools and venues through which they can hold the regime accountable, should it fail to keep its formal commitments, both domestic (Keith, 2002: 2012) and international (e.g., Powell and Staton, 2009). A truly independent judiciary, in particular, "should be able to withstand incursions upon rights because (1) the courts' power and fiscal well-being are protected, (2) the courts have some ability to review the actions of other agencies of government, and (3) the judges' jobs are constitutionally protected" (Keith et al., 2009: 649). Powell and Staton (2009) posit that in states where the courts are influential (effective),

the use of repression may lead to rights claims in which the regime will incur a loss of resources as punishment for violating its obligations (154).

Another set of theoretical approaches generally focuses on transnational or international interaction and socialization which is believed to drive the creation of and commitment to international human rights norms, rather than rationalist calculations. These perspectives emphasize the transformative power of international normative discourse on human rights and the role of activism by transnational actors (international organizations and nongovernmental actors) who through repeated interactions with state actors induce states to accept new norms and also support local efforts to press for human rights commitment. Thus, their emphasis is on the processes through which state actors adopt formal constraints on the state's opportunity to repress and through which state actors eventually internalize international human rights norms or values which over time removes many of the tools of repression from the state's menu of appropriate policy options. The transnational advocacy networks posits that international human rights norms are diffused through networks of transnational and domestic actors who "bring pressure 'from above' and 'from below' to accomplish human rights change" (Risse and Sikkink, 1999: 18). A second perspective, the world society approach (e.g., Meyer et al., 1997) perceives states to be embedded in an integrated cultural system that "promulgates cognitive frames and normative prescriptions that constitute the legitimate identities, structures, and purposes of modern nation-states" (Cole, 2005: 477). Thus, with the proliferation of human rights treaties codifying human rights norms, states' legitimacy, or 'good nation' identity is increasingly linked to the formal acknowledgment of these norms and newly emerging or independent states will be likely to adopt standardized forms of government (Go, 2003). However, as Cole notes, many states join the traditionally weak human rights regime, "not out of deep commitment, but because it signals their probity to the international community" and thus "a decoupling is endemic to the human rights regime" (p. 477).

Empirical Findings

Standard Model

While much of the early repression work was quite thin in terms of applying broad theoretical approaches, as Davenport and Armstrong (2004) note, over time a 'standard' model has emerged in the repression literature (Poe and Tate, 1994; Poe et al., 1999 and the numerous citations within) that includes nine variables: level of democratization, which is discussed separately in the following sections; two forms of threat (civil war and international war), which are also discussed in the following sections; two additional regime types (military and leftist regimes); and socio-economic conditions (economic development, population size, and colonial legacy). All of but one of the variables within the model represent the domestic environment of the state that either constrains or creates the regime's opportunities to repress or that affects the regime's decision-making calculation of the advantage or disadvantage to exercising repressive tools as a means to achieve its policy goals. Over time the body of empirical work has expanded beyond domestic influences to examine a variety of international or transnational influences, including bilateral aid, multilateral lending programs, international treaty regimes, trade relations, and foreign investments. As the models of repression have expanded, the latter five variables of the standard model above have performed inconsistently. And while they continue to be used as control variables, other variables have become more theoretically and statistically important.

Democracy

Democracy is the most widely tested and consistently performing variable in the standard model. While the literature has tended to assume a linear association between the level of democracy and repression, recent theoretical and empirical work suggests that this oft unstated assumption does not accurately reflect the true relationship between democracy and repression. In more recent work, Davenport and Armstrong (2004) as well as Bueno de Mesquita et al. (2005) argue and demonstrate empirically that the relationship between democracy and repression has been misspecified and that a threshold effect is a more accurate characterization. Other scholars have perceived democracy as a multidimensional concept and have found that some dimensions of democracy are more likely to lessen the use of political repression – most particularly, the competitiveness of political participation (Keith, 2002a; Bueno de Mesquita et al., 2005).

Domestic and External Threats

Theoretically, the most significant factor is the presence of domestic opposition or a perceived challenge to the regime's hold on power, especially if there is a threat (actual or perceived) that the regime's challengers may resort to violent tactics. The regime may choose to employ coercive force to prevent, contain, or punish such threats. The most potentially disruptive form of domestic threat, civil war, has received the most empirical attention, and its effect has been demonstrated to be the strongest of many factors, in terms of impact, statistical robustness, and consistency across a variety of measures of repression and other human rights behavior. To illustrate the magnitude of its impact, Poe et al. (1999) demonstrate that an ongoing civil war would increase the level of repression over time by around one and half points on a five point scale of repression, with other factors in the model held equal (p. 308). The literature has also sought to understand the impact of less severe forms of threat, as well as other dimensions of threat, such as the variety of strategies and frequency of conflict. For example, Davenport (1995) found that deviance from past norms of conflict, the frequency of conflict, and the variety of strategies engaged, each increase the likelihood that a state would resort to civil liberties restrictions; however, once these factors are controlled the presence of domestic violence becomes statistically insignificant. Poe, Tate, Keith, and Lanier (2000) found that the repressive response to domestic threats was dependent not only on the level of the threat but also upon the prior level of repression. Regan and Henderson (2002) also found that the level of threat was positively associated with repression, and that the impact of threat was greater than that

of regime type. External threats are also theoretically important, and the empirical literature has consistently confirmed the effect of international war participation on domestic political repression, and thus over time, international war has become a key component of the standard model of human rights behavior (Poe and Tate, 1994; Poe et al., 1999; Cingranelli and Richards, 1999; Davenport and Armstrong, 2004); however, the impact of international war is substantially lower than that of civil wars. For example, Poe et al. 1999 found the impact of participation in a civil war to be at least three times the impact of being a participant in an international war. This is somewhat intuitive even though external threats such as interstate war are potentially more serious than civil wars, often civil wars are a greater threat to regime due to the proximity and ability of the opposition to impose damage, especially compared to some interstate wars, which may be fought far away from at least some of the belligerents' states, such as in the Persian Gulf War or the War in Iraq.

Domestic Legal Commitments and Institutions

A growing body of research has examined the question of whether constitutional provisions for rights and an independent judiciary constrain the likelihood of states repressing their own citizens. Most of these studies have examined the role of specific rights, typically imbedded bills of rights (e.g., Davenport, 1996; Keith, 2002a; Keith et al., 2009; Keith, 2012) and to a lesser extent have examined institutional constraints such as provisions for an independent judiciary (Blasi and Cingranelli, 1996; Keith, 2002b; Keith, 2012), states of emergency clauses (Keith and Poe, 2004; Keith, 2012), or provisions for federalism (Blasi and Cingranelli, 1996). Most of the earliest analyses found evidence that not only were the constitutional rights provisions not associated with improved rights behavior, the associations were in the opposite direction, suggesting a harmful effect. As studies have improved in their temporal and spatial dimensions and have become more sophisticated methodologically, the observed effect of constitutional rights provisions has become somewhat more consistent and suggests a beneficial effect overall. The effect of constitutional provisions for judicial independence appears to be more appropriately conceptualized as influencing *de facto* judicial independence, which in turn consistently decreasing the probability of repression of all forms (Keith, 2012). One set of constitutional provisions remains problematic – studies have consistently argued and demonstrated that certain provisions for states of emergencies (such as derogation clauses), which have been promoted by international organizations such as the International Lawyers Association and International Commission of Jurists, have unintended consequences and are largely associated with increased levels of repression.

International Treaties, Norms, and INGOs

Measuring norms is probably the most difficult task that empiricists face in testing normative approaches, and thus most empirical studies of compliance have examined surrogate indicators. Following Goodliffe and Hawkins' (2006) work on treaty commitment, Keith et al. (2009) use regional and global commitment rates to the International Covenant on Civil and Political Rights (ICCPR) as measures of norms, but find no effect in a variety of models of repression. Powell and Staton (2009) test regional and global norms, as indicated by past rates of torture, in their Convention Against Torture (CAT) models; however, they find too little evidence that norms influence state torture practices. Most empirical studies of state compliance with international human rights treaty obligations focus on the state's provision or protection of the guaranteed rights embedded in the document. The bulk of the empirical compliance literature, while grounded in rich theoretical debate concerning the influence of human rights treaties, has largely been limited to tests of whether being a state party produces a positive or negative effect, controlling for a variety of factors. It does not enable us to determine which of the mechanisms that are hypothesized to be at work are actually operative. While the literature examines a broad range of human rights behavior, much of the research has focused on rights abuse that would appropriately be labeled as political repression. Extant empirical evidence of influence of participation in the international human rights regime has been rather mixed. For example, Keith (1999) finds no effect from the ICCPR on either form of political repression, unless controlling for state derogations. Hathaway (2002) finds that most treaties within the human rights regime do not significantly affect human rights behavior, and that participation in some of the treaties, such as the CAT produce negative effects, a result which is confirmed by Hafner-Burton and Tsutsui (2005) in regard to a wide range of treaties. On the other hand the most exhaustive compliance studies (Landman, 2005) found consistent evidence of an association between rights behavior and state commitment to a variety of treaties within international human rights regime. Landman finds that commitment to the ICCPR and the CAT, even while controlling for the level of reservations decreased state repression of personal integrity rights, torture, and civil and political rights. On the other hand, Hill's (2010) sophisticated analysis of the CAT and the ICCPR, finds both to be associated with increased repression. Keith's (2012) comprehensive analysis of both forms of political repression (and a variety of measures), found that the ICCPR has both a beneficial unconditional effect and an interactive effect with international NGOs (INGOs); however, the Optional Protocol to the treaty, only has an effect in conjunction with the presence of INGOs. The CAT has neither conditional nor unconditional effect on any form of repression, regardless of measures. Thus, our understanding of international treaties' ability to reduce repression remains sketchy.

The presence of transnational networks or civil society pressure (typically measured as the number of INGOs in which citizens have membership) is much easier to measure than the presence of norms; however, the direct link between treaty ratification and civil society pressure is not typically tested in human rights models. Both Hafner-Burton and Tsutsui (2005) and Powell and Staton (2009) find that as the number of INGOs to which citizens belong increases, the level of protection of human rights increases. Neumayer (2005) provides the most rigorous analysis, examining specifically the effect of treaty commitment on human rights behavior, conditioned upon the strength of civil society organizations. However, his results are rather mixed: he does find that the stronger the state's participation in INGOs the greater the beneficial effect of

ratification of the CAT on human rights behavior; however, the results do not hold for ICCPR ratification. He also finds mixed results in regard to regional treaties. For example, while the benefit of ratification of the Inter-American torture convention increases as INGO participation increases, the effect does not hold for the European torture convention.

Overall, the observable effect of norm diffusion is rather weak and inconsistent at best. The evidence of a strong positive influence by NGOs on states human rights practices is more convincing; however, we must keep in mind that the civil society linkage through treaty ratification is not directly demonstrated except in Neumayer's work. The evidence of the beneficial effect of treaty commitment continues to grow with more sophisticated and rigorous analyses, with the exception of the CAT, which has no observable benefit.

Bilateral, Multilateral, and Global Economic Influences

As the field has grown over time, increasing attention has focused on bilateral, multilateral, and global economic influences. While much attention has been given to impact of repression on bilateral aid allocation, relatively little attention has been given to the question of whether U.S. aid influences human rights practices of recipient states (e.g., Shoutlz, 1981; Regan, 1995; Apodaca, 2001) beyond the two studies supported by USAID itself (Finkel et al., 2006, 2008) and by Knack (2004), by a senior research economist for the World Bank who examined the impact of western foreign assistance on democratization in recipient states. While there are strong theoretical links between aid and improved human rights practices, both direct and indirect, there is also some justification to expect that aid would not influence human rights or that it might even have a harmful effect. Overall, the empirical evidence suggests that bilateral aid has limited (Regan, 1995; Apodaca, 2001) or no effect (Knack, 2004), and sometimes a harmful effect (Shoultz, 1981). Two comprehensive analyses of the USAID democracy and governance aid commissioned by the USAID itself have consistently demonstrated that while spending in the other subsectors of aid (elections, civil society, and governance) all produce the desired effects, spending on the human rights component of the rule of law program has a strong negative effect on states' human rights protection, regardless of numerous controls, including controls for endogeneity, and regardless which measure of human rights is employed (Finkel et al., 2006, 2008). As the investigators note in the final report of their second study, their "effort to untangle the web of relationships that may underlie this distressing and presumptively anomalous relationships and to model them statistically has been largely unsuccessful in its basic purpose" (Finkel et al., 2008: 57). While scholars seem somewhat reluctant to expect these null or negative findings, we are left with the dilemma that aid may actually promote political repression, at least under some circumstances. Recent empirical studies of the influence of multilateral aid programs have been just as pessimistic. A growing number of empirical studies have demonstrated the IMF structural adjustment programs lead to increased repression. The most comprehensive and rigorous study to date (Abouharb and Cingranelli, 2008) demonstrates that the structural adjustment programs of both the IMF and World Bank increase civil conflict which in turn increases the state's level of physical integrity. The evidence in regard to global economic influences, such as trade openness and foreign economic penetration is mixed but somewhat more optimistic than that of the aid relationship.

Empirical studies that have tested liberal economic theory on the link between trade and human rights protection, although small in number, have consistently confirmed these expectations (e.g., Apodaca, 2001; and Harrelson-Stephens and Calloway, 2003). The evidence examining the effect of foreign investment has been much more inconsistent, often dependent upon the category of repression with FDI reducing only the likelihood of political repression of civil liberties not the personal integrity (Apodaca, 2001; Richards et al., 2001). Keith's (2012) more comprehensive study of a three-decade period consistently found that all indicators of global embeddedness (trade, FDI, and WTO membership) were associated with increased odds of both forms of repression. Overall, the effect of bilateral and multilateral aid, as well as those of global economic influences, proves consistently to be in a direction that is harmful to human rights, and to increase the likelihood of repression.

The Future of Empirical Study of Political Repression

Several limitations across the literature suggest the direction for future research. While the theoretical approaches that have informed the empirical examination of state repression has expanded to draw a wider range of disciplines, an integrative grand theory is yet to evolve. And while the reification of the state as a unitary actor has faded somewhat with the integration of the domestic institutions approaches, the empirical literature continues largely to ignore the individual, actual agents of the state (or private actors sanctioned by the state) who directly engage in the human rights violations. This lack is likely linked, in part, to severe data limitations that restrict the growth of the field. For this body of research to progress, social scientists must create data that move the temporal dimension of the data beyond the arbitrary calendar year. These finer delineations are necessary to capture the response of states to particular events, especially in terms of perceived threats to the regime. In addition, scholars need to follow the lead of Conrad and Moore and create variables that identify the state or state-sanctioned actors (and their relevant characteristics) who engage in or facilitate repression, and similarly, variables that identify the targets of repression. Finally, scholars should move toward capturing a more nuanced understanding of the modalities of repression, especially in regard to the variation in legality. This improvement would allow scholars to examine the substitutability of tools of repression.

Bibliography

Abouharb, R., Cingranelli, D., 2008. Human Rights and Structural Adjustment. Cambridge University Press, New York.

Apodaca, C., 2001. Global economic patterns and personal integrity rights after the cold war. International Studies Quarterly 45, 587–602.

Blasi, G., Cingranelli, D., 1996. Do Constitutions and Institutions Help Protect Human Rights? In: Cingranelli, D. (Ed.), Human Rights and Developing Countries. JAI, Greenwich.

Bueno de Mesquita, B., Downs, G., Smith, A., Cherif, F., 2005. Thinking inside the box: a closer look at democracy and human rights. International Studies Quarterly 49, 439–457.

Cingranelli, D., Richards, D., 1999. Measuring the pattern, level, and sequence of government respect for human rights. International Studies Quarterly 43, 407–417.

Cingranelli, D., Richards, D., Clay, C., 2013. The CIRI Human Rights Dataset. Version 2013.12.05. http://www.humanrightsdata.org (accessed 14.04.01).

Cole, W., 2005. Sovereignty Relinquished? Explaining Commitment to the International Human Rights Covenants, 1966–1999. American Sociological Review 70, 472–495.

Conrad, C., Moore, W., 2014. The Ill-Treatment and Torture Data Collection Project. http://faculty.ucmerced.edu/cconrad2/Academic/ITT_Data_Collection.html (accessed 14.04.01).

Davenport, C., 1995. Multi-Dimensional threat perception and state repression: an inquiry into why states apply negative sanctions. American Journal of Political Science 39, 683–713.

Davenport, C., 1996. Constitutional promises' and repressive reality: a cross-national time-series investigation of why political and civil liberties are suppressed. Journal of Politics 58, 627–654.

Davenport, C., 1999. Human rights and the democratic proposition. Journal of Conflict Resolution 43, 92–116.

Davenport, C., 2007a. State Repression and the Domestic Democratic Peace. Cambridge University Press, New York.

Davenport, C., 2007b. State repression and political order. Annual Review of Political Science 10, 1–23.

Davenport, C., Armstrong, D., 2004. Democracy and the violation of human rights: a statistical analysis from 1976 to 1996. American Journal of Political Science 48, 538–554.

Finkel, S., Pérez-Liñán, A., Seligson, M., Azpuru, D., 2006. Effects of U.S. Foreign Assistance on Democracy Building: Results of a Cross-National Quantitative Study. http://pdf.usaid.gov/pdf_docs/Pnade694.pdf (accessed 14.04.01).

Finkel, S., Pérez-Liñán, A., Seligson, M., Tate, N., 2008. Deepening Our Understanding of the Effects of US Foreign Assistance on Democracy Building Final Report. http://pdf.usaid.gov/pdf_docs/PNADL961.pdf (accessed 14.04.01).

Gibney, M., Cornett, L., Wood, R., Haschke, P., 2014. Political Terror Scale 1976–2012. http://www.politicalterrorscale.org/ (accessed 14.04.01).

Go, J., 2003. A globalizing constitutionalism? views from the postcolony, 1945–2000. International Sociology 18, 71–95.

Goodliffe, J., Hawkins, D., 2006. Explaining commitment: states and the convention against torture. The Journal of Politics 68, 358–371.

Hafner-Burton, E., Tsutsui, K., 2005. Human rights in a globalizing world: the paradox of empty promises. American Journal of Sociology 110, 1373–1411.

Harrelson-Stephens, J., Callaway, R., 2003. Does trade openness promote security rights in developing countries? examining the liberal perspective. International Organizations 29, 143–158.

Hathaway, O., 2002. Do human rights treaties make a difference? Yale Law Journal 111, 1935–2042.

Henderson, C., 2004. The political repression of women. Human Rights Quarterly 26, 1028–1049.

Hill, D., 2010. Estimating the effects of human rights treaties on state behavior. The Journal of Politics 72, 1161–1174.

Keith, L., 2002a. Constitutional provisions for individual human rights (1976–1996): are they more than mere "Window Dressing?". Political Research Quarterly 55, 111–143.

Keith, L., 2002b. International principles for formal judicial independence: trends in national constitutions and their impact (1976 to 1996). Judicature 85, 194–200.

Keith, L., 2012. Political Repression: Courts and Law. University of Pennsylvania Press, Penn.

Keith, L., Poe, S., 2004. Are constitutional state of emergency clauses effective? an empirical exploration. Human Rights Quarterly 26, 1071–1097.

Keith, L., Tate, N., Poe, S., 2009. Is the law a mere parchment barrier to human rights abuse? Journal of Politics 71, 644–660.

Knack, S., 2004. Does foreign aid promote democracy? International Studies Quarterly 48, 251–266.

Landman, T., 2005. Protecting Human Rights: A Comparative Study. Georgetown University Press, Washington, DC.

Meyer, J., Boli, J., Thomas, G., Ramirez, F., 1997. World Society and the Nation-State. American Journal of Sociology 103, 144–181.

Poe, S., Tate, N., 1994. Repression of human rights to personal integrity in the 1980s: a global analysis. American Political Science Review 88, 853–872.

Poe, S., Tate, N., Keith, L., 1999. Repression of human rights to personal integrity revisited: a global cross-national study covering the years 1976–1993. International Studies Quarterly 43, 291–313.

Poe, S., Tate, N., Keith, L., Lanier, D., 2000. The continuity of suffering: domestic threat and human rights abuse across time. In: Davenport, C. (Ed.), Paths to State Repression: Human Rights and Contentious Politics in Comparative Perspective. Roman Littlefield, Boulder.

Powell, E., Staton, J., 2009. Domestic judicial institutions and human rights treaty violation. International Studies Quarterly 53, 149–174.

Regan, P., 1995. US economic aid and political repression: an empirical evaluation of US foreign policy. Political Research Quarterly 48, 613–628.

Regan, P., Henderson, E., 2002. Democracy, threats, and political repression in developing countries: are democracies internally less violent? Third World Quarterly 23, 119–136.

Richards, D., Gelleny, R., Sacko, D., 2001. Money with a mean streak? foreign economic penetration and government respect for human rights in developing countries. International Studies Quarterly 45, 219–239.

Risse, T., Sikkink, K., 1999. The socialization of international human rights norms into domestic practices: introduction. In: Risse, T., et al. (Eds.), The Power of Human Rights: International Norms and Domestic Change. Cambridge University Press, New York, pp. 1–38.

Schoultz, L., 1981. Human Rights and United States Policy toward Latin America. Princeton University Press, Princeton, NJ.

Taylor, C., Jodice, D., 1983. World Handbook of Political and Social Indicators. Yale University Press, New Haven.

Political Science and Empirical Implications of Theoretical Models

Jim Granato, University of Houston, Houston, TX, USA
Melody Lo, College of Business, University of Texas at San Antonio, San Antonio, TX, USA
Man Chiu Sunny Wong, University of San Francisco, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes a framework for the Empirical Implications of Theoretical Models (EITM) initiative. The objective of EITM is to encourage political scientists to test and estimate empirical models that are directly connected to an analytical model. The practice of merging formal and empirical analysis can reduce nonfalsifiable research and acquire social scientific accumulation. The EITM framework can be considered a three-step procedure. The first step is for researchers to unify theoretical mechanisms and applied statistical concepts. Step two is to develop measurable devices ('analogs') for these mechanisms and concepts. The final step is to unify the analogs. Using a set of plausible facts or axioms, scholars can build a rigorous mathematical model to potentially identify causal relations that explain empirical regularities.

Introduction

Background

In July 2001, the Political Science Program at the National Science Foundation (NSF) convened a workshop seeking ways to improve technical–analytical proficiency in Political Science. (The 2001 NSF EITM Workshop was recorded and transcribed. The written transcript is available on the NSF Political Science Program Web Site: www.nsf.gov/sbe/ses/polisci/reports/eitm709.pdf and www.nsf.gov/sbe/ses/polisci/reports/eitm710.pdf. A report of the EITM initiative (EITM Report 2002) is also available at: www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf.) The workshop, termed the Empirical Implications of Theoretical Models (EITM) Workshop, suggested constructive approaches so that the NSF Political Science Program could develop linkages between formal and empirical modeling. (A quantitative model can be characterized as a construct that is represented by a set of logically – and in this case – quantitatively connected variables. Quantitative models include formal analysis (modeling) and applied statistical analysis. Formal analysis (i.e., social choice, linear difference equations, and differential equations) refers to deductive modeling that includes a theorem and proof presentation or computational modeling requiring simulation. Applied statistical analysis involves data analysis (from experimental or secondary sources) using statistical tools.)

The workshop participants acknowledged that a schism had developed between those who engage in formal modeling, focusing on quantifying abstract concepts mathematically, and those employing empirical modeling emphasizing applied statistics. (The participants and their commentaries are listed in Appendix B of the 2002 EITM Report.) As a consequence, a great deal of research in political science is competent in one technical area, but lacking in another. To put it another way, the impaired competency is reflected by either a formal approach with substandard (or no) empirical tests or an empirical approach without formal clarity.

This 'siloed' research – and attendant training regimen – contributes to a failure in identifying the proximate causes explicated in a theory and, in turn, increases the difficulty of achieving a meaningful increase in scientific knowledge. A reason is that researchers limit themselves to the strengths and weaknesses of a single methodological approach. For formal modelers, this manifests itself in not respecting the facts; for researchers who rely exclusively on applied statistics, we find data mining, garbage-can regressions, and statistical patches (i.e., omega matrices) (Granato and Scioli, 2004).

EITM is meant to correct these current methodological practices. When using EITM, scholars are making the ultimate focus of a model, and test is to support a cumulative scientific process, a process geared toward finding a causal mechanism. The ability of a researcher to parse out specific causal linkages among the many factors encountered is fundamental to the scientific enterprise. But, it also should be noted that no one engaged in the scientific enterprise would call finding causal mechanisms – easy. What makes methodological unification so useful in this process is that specifying a model linking both formal and empirical approaches alerts researchers to outcomes when specific conditions are in place. In other words, EITM promotes a dialog between models and tests to not only aid in predictive accuracy but also in understanding the workings of a system (particularly through formalization), which can sometimes occur long before tests and data are available (Powell, 1999; Coase, 1994).

We hasten to add that methodological unification – linking formal and empirical analysis – in the social and behavioral sciences is not new and can be traced primarily to the accomplishments of the Cowles Commission in the 1930s. Created in the 1930s, the Cowles Commission was designed "to foster the development of logical, mathematical, statistical methods of analysis for application in economics and related social sciences" (see http://cowles.econ.yale.edu/about-cf/about.htm). The basis for this linkage is the notion that random samples are governed by a latent and probabilistic law of motion (Haavelmo, 1944; Morgan, 1990). Further, this view argues that formal models, related to an applied statistical model, could be interpreted as creating a sample drawn from the underlying law of motion. A well-grounded test of a theory could be accomplished then by relating a formal model to an applied statistical model and testing the applied statistical model.

Current Methodological Practices: The Shortcomings

The Cowles contributions to the social sciences were far reaching. It was the Cowles Commission that explored the differences between structural- and reduced-form parameters (Morgan, 1990; Heckman, 2000: p. 46). Conditions for identifiability were introduced to aid in this differentiation. The Cowles methodology created new research aimed at determining valid inference through the emphasis of the properties identification and invariance. For *identification*, rules (i.e., rank and order conditions) were devised so that an equation of a model could reveal one and only one set of parameters consistent with both the model and the observations (Koopmans, 1949). A second issue involved the *invariance* of a (structural) relation. If an underlying mechanism is constant in the past and future, then the path of the relevant variable(s) will be predictable from the past, apart from random disturbances (Marschak, 1947, 1953). This work is now part of standard texts in econometrics. We adopt Heckman's (2000: p. 59) terminology below:

> Structural causal effects are defined as the direct effects of the variables in the behavioral equations … When these equations are linear, the coefficients on the causal variables are called structural parameters (emphasis added), and they fully characterize the structural effects.

Heckman also notes there is some disagreement about what constitutes a structural parameter. The disagreement centers on whether one uses a linear model, a nonlinear model or, more, recently a fully parameterized model. In the latter case, structural parameters, can also be called 'deep' to distinguish between "the derivatives of a behavioral relationship used to define causal effects and the parameters that generate the behavioral relationship" (p. 60).

Current methodological practice marginalizes the Cowles Commissions contributions, but also served for the motivation for the EITM initiative and framework. Here we identify two adverse consequences: (1) when used in isolation, formal models miss key empirical findings and can only obscure a more accurate understanding of the inner workings of a system let alone impair predictive accuracy; and (2) empirical modelers' excessive reliance on statistically *significant* results means nothing when the researcher makes little attempt to identify the precise origin of the parameters in question. Absent this identification effort, it is not evident where the model fits facts.

Formal Models and Facts

Formal models can fail to incorporate empirical findings that would assist in providing a more accurate depiction of the relations that are specified. This results in modeling efforts that yield inaccurate predictions or do not fit findings. In fact, data may contradict not just a model's results but also its foundational assumptions. The problem is not just the unreal assumptions, for one way to build helpful models is to begin with stylized assumptions, test the model's predictions, and then modify the assumptions consistent with a progressively more accurate model of reality. Yet, these follow-up steps are too often not taken or left unfinished – with the result being a model that does little to enhance understanding and advance the discipline.

Applied Statistical Procedures that Fail to Identify Mechanisms

Three harmful but common applied statistical practices exist: data mining, garbage cans, and the use of statistical weighting and patching ('omega matrices'). These applied statistical practices lack overall robustness as they obscure fundamental specification error. To see how this occurs, we summarize each practice (Granato and Scioli, 2004):

1. *Data mining*: This practice involves putting data into a statistical package with minimal theory. Regressions (likelihoods) are then estimated until either statistically significant coefficients or coefficients the researcher likes are found. This *stepwise* search is not random and has little relation to identifying causal mechanisms (Lovell, 1983; Denton, 1985).
2. *Garbage cans*: This practice, related to data mining, is where a researcher includes, in a haphazard fashion, a plethora of independent variables into a statistical package and gets 'significant' statistical results somewhere. Researchers who use garbage-can models usually pay little attention to potential confounding factors that could corrupt statistical inferences. Efforts to identify an underlying causal mechanism are also few and far between (Achen, 2005).
3. *Omega matrices*: Data mining and garbage-can approaches virtually are guaranteed to break down statistically. The question is what to do when these failures occur. There are elaborate ways of using (error) weighting techniques *to correct* model misspecifications or to use other statistical patches that influence $s.e.(b)$. In almost any intermediate econometrics textbook one finds a section that has the Greek symbol: Omega (Ω) (Johnston and DiNardo, 1997: p. 189). This symbol is representative of the procedure whereby a researcher weights the data that are arrayed (in matrix form) so that the statistical errors, ultimately the standard error noted above, are altered and the t-statistic is manipulated.

In theory, there is nothing wrong with knowing the Omega matrix for a particular statistical model. The standard error(s) produced by an Omega matrix should only serve as a check on whether inferences have been confounded to such an extent that a Type I or Type II error has been committed. Far too often, however, researchers treat the Omega weights (or alternative statistical patches) as the result of a true model. This attitude hampers scientific progress because it uses a model's mistakes to obscure flaws. An example of this practice is the 'correction' for serial correlation (Mizon, 1995).

We can also evaluate these current practices in relation to how they fail to contribute to a modeling dialog between theory and test (see **Figure 1**). What we see is that the processes of prediction and validation are never directly applied. Instead, the empirical test(s) remains in a loop or dialog with itself. An iterative process of data mining, overparameterization, and the use of statistical patches (Omega matrices) replaces prediction, validation (falsification), and an understanding of the process.

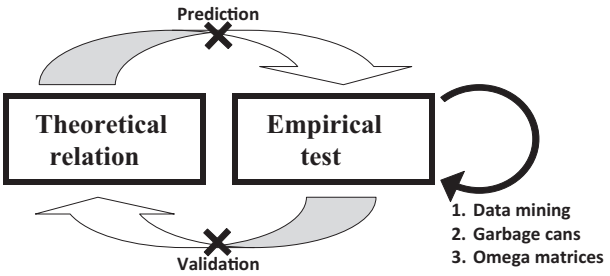

Figure 1 Noncumulative applied statistical practices.

It is important to note when we use the word validation, this is not simply a process of a single theory being tested by the accuracy of its predictions but also in comparison to competing theories on the same phenomena. As Kuhn (1979) states:

> Anomalous observations … cannot tempt [a scientist] to abandon his theory until another one is suggested to replace it…In scientific practice the real confirmation questions always involve the comparison of two theories with each other and with the world, not the comparison of a single theory with the world (p. 211).

The EITM Framework

The EITM framework builds on the Cowles Commission approach. (see Granato (2005) and Granato et al. (2010a,b, 2011) for a description and examples of the EITM framework. An extensive and earlier treatment on the linkage between formal and empirical analysis can be found in Morton (1999).) This framework takes advantage of the mutually reinforcing properties of formal and empirical analysis addressing the challenge(s) above. Why develop a framework? As Ostrom argues the development of a framework is important since "without a framework to organize relevant variables identified in theories and empirical research, isolated knowledge acquired from studies … by … social scientists is not likely to cumulate (Ostrom, 2009: p. 420)."

Mechanism Operationalization

The EITM framework supports on mechanism operationalization. It involves creating measurable devices (what we term analogs) not only on the formal and behavioral side, but also on the empirical side. Behavioral concepts include (but are not limited to) expectations, learning, social interaction, decision-making, strategic interaction, and more. Empirical concepts include (but are not limited to) persistence, measurement error, simultaneity, prediction, nominal choice, and more.

Traditional operationalization involves finding measures for behavioral concepts. These measures are usually a variable of some sort. Mechanism operationalization encompasses this tradition, but also extends it to include the use of operators, frameworks (that can include variables), and tools to represent both behavioral and empirical concepts. EITM places emphasis on finding ways to model human behavior and action and, thereby, aids in creating realistic representations that improve upon simple socioeconomic categorization. Numerous social science disciplines focus a good deal of research effort on the interactions between agent behavior and public policies. The EITM framework make use of structural equation modeling associated with the Cowles Commission, but it also includes alternative probability approaches, computational methods, and experimental methods (Poteete et al., 2010).

From Concepts to Analogs

The EITM framework also uses what we know about theoretical and applied statistical concepts, provides a rigorous basis for these concepts through the use of their respective analogs, and then merges these theoretical analogs with the applied statistical analogs. A *concept* can be thought of as an abstract or general idea inferred or derived from specific instances. An *analog* can be thought of as a device in which a concept is represented by continuously variable – and measurable – quantities. What would develop is a road map for others to modify, correct, or follow. More importantly, one would be providing a transparent linkage between a theory and test. This is not to say the model is correct. Instead, what you would be doing is meeting a minimal requirement that the theory and test are related and, therefore, serve as a basis for evaluating the inner workings of the system, predictive accuracy, or both.

The Three-Step Framework

The EITM framework contains three steps:

Step 1: Identify Focal Concepts

The goal of the EITM framework is a closer integration of theoretical and empirical models. We start, however, not with the development of mathematical structures but with the identification of *concepts*. It is of course standard to suggest that research starts with concepts. We have in mind, however, not the substantive concepts central to political science (such as legitimacy, democratization, attitude formation), but instead to the general attributes of the thing being researched. Consider, for example, the concept of democracy and democratization. Its substantive importance is undisputed, but precisely where one goes from there is unclear. There are simply too many possible ways to proceed.

The goal of this first step in EITM is to transform the focus from the substantive topic to the underlying behavioral process. A student of democracy might focus, for example, on *choice*: how do demographic and attitudinal variables drive individual selection over political parties. Another student might focus on *uncertainty and learning*: given the lack of a 'track record' among political parties in newly democratizing states, how do individuals come to form expectations regarding those parties, and how do those expectations shift in response to political and economic changes? A third student might concentrate on the idea of *bargaining*: how do the various party leaders face the trade-offs between maximizing their potential influence in the political system and maintaining the promise to democratize? The idea is not to ignore the substantive aspects, but to look at substance from a different perspective, one that not only helps clarify the focus of the research, but

also suggests common *behavioral* concerns that make it easier to communicate across subfields and find common elements and approaches. This leads to the second step.

Step 2: Develop Formal and Applied Statistical Analogs

Once the focal concept has been identified, work moves to the second step, the development of formal and applied statistical analogs. Consider, for example, the concept of *persistence*. Theoretically, this conjures up a host of potential analogs: coordination equilibria, Markov chain steady states, locally and globally stable solutions to systems of differential equations, or still lifes in cellular automata models. Similarly, persistence can be tied to a number of statistical models, including autoregressive estimation and asymptotic theory.

Step 3: Unify and Evaluate the Analogs

By starting with the concept and then moving to the theoretical and applied statistical analogs, we guarantee that there must be *something* in common between the theory and the empirical analysis. The idea, then, is to locate the parameters of interest in each that reflect the underlying concept, and then use those to build clearer and stronger links between the mechanisms of the theoretical model and the specification of the statistical methods. The specified linkage not only draws theory and empirics closer, but also provides a way for research to build by showing potential sources of inaccuracies and model failure.

EITM in Practice: Economic Voting

We now demonstrate the EITM framework using published research on economic voting. These studies have refined earlier work and present models of voter sophistication and new applied statistical tests. In the former instance, voters possess the capability to deal with uncertainty in assigning blame or credit to incumbents for good or bad economic conditions. For the latter, applied statistical tests include some of the more advanced tools in time series analysis.

The example we use examines how the behavioral concepts of expectations and uncertainty, when linked to the applied statistical concept of measurement error, create distinct hypotheses in the economic voting literature. The analogs for these concepts are as follows: the formal analog(s) are: conditional expectations modeling, recursive projections, and the law of iterated projections (or expectations). The applied statistical analog is error-in-variables regression.

Step 1: Relating Behavioral and Applied Statistical Concepts: Expectations, Uncertainty, and Measurement Error

Starting with Kramer (1983) and continuing with work that includes (but is not limited to) Alesina and Rosenthal (1995), Suzuki and Chappell (1996), and Lin (1999), there is a substantial quantitative literature on how the economy affects voting behavior. For our purposes we demonstrate how the combined work of Alesina and Rosenthal and Suzuki and Chappell have in their own way contributed constituent parts to the study of economic voting.

The uncertainty that voters face can ideally be tested using the EITM framework. The EITM relation can be found in the following way: when expectations and uncertainty analog(s) are linked to a measurement error analog (error-in-variables regression), it can be shown that parameters in the regression reflect the increasing variance (i.e., uncertainty) of the variables of interest. In other words, as expectations and uncertainty analogs change, the testable economic effect on voting changes as well.

Step 2: Devise Formal and Applied Statistical Analogs

We start with a formal model that uses the behavioral concepts: expectations and uncertainty. Alesina and Rosenthal (1995) provide such a formal model (pp. 191–195). Their model of economic growth is based on an expectations-augmented aggregate supply curve:

$$\widehat{y}_t = \widehat{y}^n + \gamma(\pi_t - \pi_t^e) + \varepsilon_t \qquad [1]$$

where $\widehat{y}_t$ represents the rate of economic growth (GDP growth) in period t, $\widehat{y}^n$ is the natural economic growth rate, π_t is the inflation rate at time t, and π_t^e is the expected inflation rate at time t formed at time $t-1$.

With voter inflation expectations established, we now turn to the concept of uncertainty. Let us assume that voters want to determine whether to attribute credit or blame for economic growth (y_t) outcomes to the incumbent administration. Yet, voters are faced with the uncertainty of determining what part of the economic outcomes is due to incumbent 'competence' (i.e., policy acumen) or simply good luck.

Equation [2], the shock in eqn [1], can be modified to present the analog for uncertainty:

$$\varepsilon_t = \eta_t + \xi_t \qquad [2]$$

The variable ε_t represents the shock that is comprised of the two unobservable characteristics noted above – competence or good luck. (This following modification is commonly referred to as a 'signal extraction' or measurement error problem.) The first, represented by η_t, reflects 'competence' that can be attributed to the incumbent administration. The second, symbolized as ξ_t, are shocks to growth that are beyond administration control (and competence). Both η_t and ξ_t have zero mean with variance(s) σ_η^2 and σ_ξ^2, respectively.

Note also that competence can persist and *support* reelection. This feature can be characterized as an MA(1) process:

$$\eta_t = \mu_t + \rho\mu_{t-1}, \quad \rho \in (0, 1] \qquad [3]$$

where μ_t is $iid(0, \sigma_\mu^2)$. The parameter ρ represents the strength of the persistence. The lag or lags allow for retrospective voter judgments.

If we reference eqn [1] again, let us assume that voters' judgments include a general sense of the average rate of growth ($\widehat{y}^n$) and the ability to observe actual growth ($\widehat{y}_t$). Voters can evaluate their difference ($\widehat{y}_t - \widehat{y}^n$). Eqn [1] also implies that when voters predict inflation with no systematic error (i.e., $\pi_t^e = \pi_t$), the result is noninflationary growth that does not adversely affect real wages.

Next, we tie economic growth performance to voter uncertainty. We formalize how economic growth rate deviations from the average can be attributed to administration competence or fortuitous events:

$$\widehat{y}_t - \widehat{y}^n = \varepsilon_t = \eta_t + \xi_t \qquad [4]$$

Equation [4] shows that when the actual economic growth rate is greater than its average or 'natural rate' (i.e., $\widehat{\gamma}_t > \widehat{\gamma}^n$), then $\varepsilon_t = \eta_t + \xi_t > 0$. Again, the voters are faced with uncertainty in distinguishing the incumbent's competence (η_t) from the stochastic economic shock (ξ_t). However, because competence can persist, voters use this property to make forecasts and give greater or less weight to competence over time.

To demonstrate this behavioral effect, substitute eqn [3] in [4]:

$$\mu_t + \xi_t = \widehat{\gamma}_t - \widehat{\gamma}^n - \rho\mu_{t-1} \qquad [5]$$

We now determine the optimal estimate of competence, (η_{t+1}), when the voters only see $\widehat{\gamma}_t$. We demonstrate this result by making a one-period forecast of eqn [3] and then solve for its expected value (conditional expectation) at time t:

$$E_t(\eta_{t+1}) = E_t(\mu_{t+1}) + \rho E(\mu_t|\widehat{\gamma}_t) = \rho E(\mu_t|\widehat{\gamma}_t) \qquad [6]$$

where $E_t(\mu_{t+1}) = 0$. Using this analog for expectations in eqn [6], we see that competence, η_{t+1}, can be forecasted by predicting μ_{t+1} and μ_t. Since there is no information available for forecasting μ_{t+1}, voters can only forecast μ_t based on observable $\widehat{\gamma}_t$ (at time t) from eqn [5].

Step 3: Link the Formal and Applied Statistical Analogs

Now, using the method of recursive projection and eqn [5] the behavioral analog for expectations is linked to the empirical analog for measurement error (an error-in-variables 'equation'):

$$E_t(\eta_{t+1}) = \rho E(\mu_t|\widehat{\gamma}_t) = \rho\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\xi^2}(\widehat{\gamma}_t - \widehat{\gamma}^n - \rho\mu_{t-1}) \qquad [7]$$

where $0 < \rho\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma_\xi^2} < 1$. Equation [7] shows that voters can forecast competence using the value of $\widehat{\gamma}_t - \widehat{\gamma}^n$, but also the 'weighted' lag of μ_t (i.e., $\rho\mu_{t-1}$).

In eqn [7], the expected value of competence is *positively* correlated with economic growth rate deviations. Voter assessment is filtered by the coefficient, $\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma_\xi^2}$, which represents a proportion of competence that voters are able to interpret and observe. The behavioral implications are straightforward. If voters interpret that the variability of economic shocks come solely from the incumbent's competence (i.e., $\sigma_\xi^2 \to 0$), then $\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma_\xi^2} \to 1$. On the other hand, the increase in the variability of uncontrolled shocks, σ_ξ^2, confounds the observability of incumbent competence since the signal coefficient $\frac{\sigma_\mu^2}{\sigma_\mu^2+\sigma_\xi^2}$ decreases. Voters assign less weight to economic performance in assessing the incumbent's competence.

To sum up, Alesina and Rosenthal provide an EITM connection between eqns [1], [3] and their empirical tests. *They link the behavioral concepts – expectations and uncertainty – with their respective analogs (conditional expectations and measurement error) and devise a signal extraction problem*. While the empirical model resembles an error-in-variables specification, testable by dynamic methods such as rolling regression (Lin, 1999), instead they estimate the variance-covariance structure of the residuals.

Summary and Discussion

We have described the weaknesses of current research practices and how they do not contribute to a valid modeling dialog. The failure to tie theories to tests and provide meaningful feedback means scientific accumulation suffers. The EITM framework is a response to this methodological status quo in formal and empirical analysis.

It should also be noted that implementation of the EITM framework *requires* some changes in graduate training. If methodological instruction continues on its current path, then successive generations will only repeat the shortcomings noted.

As a final thought, and if we were to sum up the EITM framework, we would argue it is a guide to unification. But, there is no set formula for what tools to use. What dictates the tools to use is how you characterize your idea. At that point, you will need to look in a tool box and use the specific formal and empirical techniques available and appropriate. Competing characterizations are appropriate and should be evaluated on their ability to represent accurately the workings of the system as well as their predictive power.

Appendix: Notes on Analogs

Expectations and Uncertainty: The Tools and Components that Build the Behavioral Analog

The tools that allow us to link an error-in-variables regression with eqn [7] are more complicated than in the first example. The discussion below is based in part on Sargent (1987: pp. 223–238). The behavioral (formal) analogs for expectations and uncertainty center on *rational (conditional) expectations* modeling. The rational expectations hypothesis can be characterized as individuals forming their subjective expectations of a variable (y^e) consistent with an objective mathematical expectation of (y) that is conditional on all available information (Ω). (Note: this symbol should not be confused with the omega matrix discussion in Section Applied Statistical Procedures that Fail to Identify Mechanisms.)

Formally, rational expectations can be defined as a mathematical expectation (projection) of the probability distribution of y given a set of information, $\Omega = \{x_0, x_1, \ldots, x_n\}$:

$$y^e = P(y|\Omega) \qquad [1a]$$

Assuming that the mathematical expectations are formed in a linear expression:

$$P(y|\Omega) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n = \sum_{i=0}^{n} a_ix_i \qquad [1b]$$

where $x_0 = 1$. Given the orthogonality principle where $E[(y - P(y|\Omega))x_i] = 0$ for all i, we have the following normal equations

$$\begin{bmatrix} Ey \\ Eyx_1 \\ Eyx_2 \\ \vdots \\ Eyx_n \end{bmatrix} = \begin{bmatrix} 1 & Ex_1 & Ex_2 & \cdots & Ex_n \\ Ex_1 & Ex_1^2 & Ex_1x_2 & \cdots & \\ Ex_2 & Ex_1x_2 & \ddots & & \\ \vdots & \vdots & & \ddots & \\ Ex_n & & & & Ex_n^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \qquad [1c]$$

Based on the orthogonality principle, a′ can be uniquely determined (see Sargent, 1987: pp. 223–226). From [1c], we can calculate the vector of a's:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \left[Ex_ix_j\right]^{-1}\left[Eyx_k\right] \qquad [3d]$$

Now apply the formulation above to the following linear expression and apply the least squares rule for $n = 1$:

$$P(y|1, x_1) = a_0 + a_1x_1$$

For eqn [1c], we have:

$$\begin{bmatrix} Ey \\ Eyx_1 \end{bmatrix} = \begin{bmatrix} 1 & Ex_1 \\ Ex_1 & Ex_1^2 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$$

where in normal equation form, the result is:

$$a_0 = Ey - a_1Ex_1$$

and then applying the least squares rule:

$$a_1 = \frac{E(y - E(y))(x_1 - Ex_1)}{E(x_1 - Ex_1)^2} = \frac{\sigma_{x_1y}}{\sigma_{x_1}^2}$$

where σ_{x_1y} is the covariance between x_1 and y, and $\sigma_{x_1}^2$ is the variance of x_1.

Relating the Behavioral Analogs to the Empirical Analog for Measurement Error: Error-in-Variables Regression

The linkage between conditional expectations with the law of iterated expectations has a natural relation with error-in-variables regression (measurement error). There are many examples of this 'EITM-like' linkage and they generally fall under the umbrella of *signal extraction* problems. This economic voting example is a signal extraction problem where there is uncertainty over attributing growth performance to incumbent competence or good luck.

Signal extraction problems can be expressed in the following way. Suppose an agent wants to estimate a random variable 's' but only 'sees' the variable x:

$$x = s + n$$

where $Esn = 0$, Es^2, $En^2 < \infty$, $Es = En = 0$.

Therefore using a projection operator and applying the least squares rule we have:

$$P(s|1, x) = a_0 + a_1x$$

and

$$a_1 = \frac{E(xs)}{Ex^2} = \frac{E[(s+n)s]}{E(s+n)^2} = \frac{Es^2}{Es^2 + En^2}$$

but

$$a_0 = 0$$

The result is that there can be a behavioral basis for an error-in-variables regression (see Johnston and DiNardo (1997: pp. 153–155)). The identity below is in the same form as eqn [7]:

$$P(s|1, x) = \frac{Es^2}{Es^2 + En^2}x = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_n^2}x$$

Acknowledgment

We wish to thank Frank Scioli for his comments and assistance.

See also: Bounded Rationality; Economic Evaluations and Electoral Consequences; Politics, Formal Modeling of; Voting, Explanations of: Social Class.

Bibliography

Achen, C.H., 2005. Let's put garbage-can regressions and garbage-can probits where they belong. Conflict Management and Peace Science 22, 327–339.
Alesina, A., Rosenthal, H., 1995. Partisan Politics, Divided Government, and the Economy. Cambridge University Press, New York.
Coase, R.H., 1994. How should economists choose? In: Coase, R.H. (Ed.), Essays on Economics and Economists. University of Chicago Press, Chicago, pp. 15–34.
Cooper, G., 1948. The role of econometric models in economic research. Journal of Farm Economics 30, 101–116.
Denton, F.T., 1985. Data mining as an industry. The Review of Economics and Statistics 67, 124–127.
EITM Report, 9–10 July 2002. Political Science Program, National Science Foundation, Directorate of Social, Behavioral, and Economic Sciences. Arlington, Virginia. www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf.
Granato, J., 2005. Scientific progress in quantitative political economy. The Political Economist 12 (4), 1–3, 9, 11–13, 20.
Granato, J., Lo, M., Wong, M.C.S., 2010a. The empirical implications of theoretical models (EITM): a framework for methodological unification. Política y gobierno XVII (1), 25–57.
Granato, J., Lo, M., Wong, M.C.S., 2010b. A framework for unifying formal and empirical analysis. American Journal of Political Science 54 (3), 783–797.
Granato, J., Lo, M., Wong, M.C.S., 2011. Modeling and testing the diffusion of expectations: an EITM approach. Electoral Studies 30, 389–398.
Granato, J., Scioli, F., 2004. Puzzles, proverbs, and omega matrices: the scientific and social significance of empirical implications of theoretical models (EITM). Perspectives on Politics 2, 313–323.
Haavelmo, T., 1944. The probability approach in econometrics. Econometrica 12, S1–S115.
Heckman, J., 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. Quarterly Journal of Economics 115, 45–97.
Johnston, J.J., DiNardo, J., 1997. Econometric Methods. McGraw-Hill, New York.
Koopmans, T., 1949. Identification problems in economic model construction. Econometrica 17, 125–144.
Kramer, G., 1983. The ecological fallacy revisited: aggregate versus individual-level findings on economics and elections and sociotropic voting. American Political Science Review 77, 92–111.
Kuhn, T.S., 1979. The Essential Tension: Selected Studies in Scientific Tradition and Change. University of Chicago Press, Chicago.
Lin, T.M., 1999. The historical significance of economic voting, 1872–1996. Social Science History 23, 561–591.
Lovell, M.C., 1983. Data mining. The Review of Economics and Statistics 65, 1–12.
Marschak, J., 1947. Economic structure, path, policy, and prediction. American Economic Review 31, 81–84.
Marschak, J., 1953. Economic measurements for policy and prediction. In: Hood, W.C., Koopmans, T.G. (Eds.), Studies in Econometric Method. Wiley, New York, pp. 1–26.
Mizon, G., 1995. A simple message to autocorrelation correctors: don't. Journal of Econometrics 69, 267–288.
Morgan, M., 1990. The History of Econometric Ideas. Cambridge University Press, Cambridge.
Morton, R.B., 1999. Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science. Cambridge University Press, New York.
Ostrom, Elinor, 2009. A general framework for analyzing sustainability of social-ecological systems. Science 325, 419–422.
Poteete, A.R., Janssen, M.A., Ostrom, E., 2010. Working Together: Collective Action, the Commons, and Multiple Methods in Practice. Princeton University Press, Princeton, NJ.
Powell, R., 1999. In the Shadow of Power. Princeton University Press, Princeton, NJ.
Sargent, T., 1987. Macroeconomic Theory. Academic Press, Orlando, FL.
Suzuki, M., Chappell Jr., H.W., 1996. The rationality of economic voting revisited. Journal of Politics 58, 224–236.

Political Science and Polling

Robert Y Shapiro, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the history and development of polling that has been relevant to political science in the United States. It covers the early 'straw polls' to the advent of 'scientific polling' based on statistical sampling, to telephone surveys, exit polls, and later technological developments and ensuing debates. It describes the general use of survey research to study public opinion and political behavior; the sources of errors in surveys and other methodological issues; and the topics in political science regarding public opinion and democracy that polling has helped examine.

Introduction

Opinion *polls* or *surveys* are widely used in political science and other fields to study the behavior of individuals and publics in the *aggregate* – including voters and smaller population subgroups. Academicians and practitioners engaged in opinion polling have preferred to call themselves 'survey researchers,' in contrast to 'pollsters' who carry out nonacademic polls to predict election results and report in the news media the mass public's or others' opinions and attitudes; or who work as consultants doing proprietary research (including market and public relations analysis) for political parties, political candidates (including leaders in office), or other clients (Converse, 1987; Moore, 1992; Jacobs and Shapiro, 1995; Eisinger, 2003). Polls and surveys are synonymous, but their quality – validity and reliability – can vary widely. Researchers select and ask questions concerning attitudes, behavior, personal characteristics, and other information that survey respondents can provide. These responses are counted, analyzed statistically, and interpreted depending on the types of questions asked and research topics examined. This article reviews the history and development of polling that has been relevant to political science in the United States from the early 'straw polls' to the advent of 'scientific polling' using statistical sampling, to telephone surveys, exit polls, and later technological developments and ensuing debates (on polling's broader relevance in comparative perspective, see e.g., Dalton and Klingemann, 2007; Frankovic et al., 2009). It focuses on the general use of survey research to study public attitudes, opinions, and political behavior; the sources of errors in surveys and other methodological issues; and the topics in political science regarding public opinion and democracy that polling has helped examine.

Straw Polls

Polling had expanded greatly worldwide by the late-twentieth century. Its origin occurred in the United States beginning with election 'straw polls' in the nineteenth century. These were motivated by people's curiosity about public opinion and voting in upcoming elections, and there was a market in the mass media for this kind of information. Opinion surveys have been a major part of the rise of quantification in society consistent with Max Weber's writing about the role of calculation and rationality in social life (Herbst, 1993). The earliest straw poll counterparts of modern opinion surveys occurred in the 1824 presidential election and were reported in the partisan press; the use of different types of these in politics and journalism increased until the emergence of more 'scientific' polling in the 1930s (Smith, 1990). These straw polls used 'convenience samples' in public gatherings and setting and took many different forms, including oral counts, hand raising, paper ballots, and preferences recorded in 'poll books.' The participants were *self-selected* in ways that were likely to produce *selection bias*, in which the results of these polls would be influenced by the characteristics of those who were able to participate and who might be different from the electorate or population to which the poll results would be generalized. Reporters and partisans, nonetheless, treated these polls as indications of candidates' election chances, and the track record of large-scale straw polls was good: the 1896 *Chicago Record*'s postcard ballots returned by over 200 000 registered voters from Chicago and eight surrounding states found that Republican presidential candidate William McKinley far ahead of Democrat William Jennings Bryan; and in Chicago it got within 0.04% of the actual vote. More (in)famously, *The Literary Digest*'s ballot cards sent to millions of its subscribers and individuals from telephone, car registration, and voter lists provided accurate estimates of presidential election winners from 1916 to 1932 (see Frankovic et al., 2009).

'Scientific' Polling

The problems associated with the representativeness and precision of such polls became clear publicly when the accuracy of the *Literary Digest* poll was called into question by George Gallup and other new 'public pollsters,' statisticians, and social scientists. Gallup became the most widely known, if not the most important figure, in what became modern-day survey research (along with Elmo Roper and Archibald Crossley; with Paul F. Lazarsfeld, Hadley Cantril, and others emerging in the academic world). He was a vocal proponent of polling to facilitate democratic responsiveness (Gallup and Rae, 1940), while ultimately using his public polls to promote his renowned proprietary survey research company. Although Gallup's first national election poll in 1936, done through

geographic and quota sampling and in-person interviews, actually underestimated the margin of Democrat Franklin Roosevelt's election victory over Republican Alfred Landon, what is remembered is that it visibly triumphed over the *Literary Digest* in predicting the winner.

The Gallup poll represented the electorate far better than did the *Digest*, confirming the superiority of more 'scientific-' minded survey methods. In principle, these methods allowed each potentially eligible adult voter (or eligible respondent, given the research objectives of a particular poll) much closer to an equal probability of selection, compared to the *Digest* that introduced an upper status selection bias from the initial lists it used. In addition, there was a further bias in the same direction as a consequence of those who chose to return the ballots (Squire, 1988). The New Deal partisan realignment that occurred by 1936 was a class-based transformation of the electorate, so that the underrepresentation of economically downscale and increasingly democratic voters was much more consequential for the *Digest* – and dramatically so – than in previous elections.

These early polls involved sending interviewers to many locations around the country and then allowed interviewers to select their respondents to fill 'quotas' based on the population characteristics to be represented in the poll. This also led to an upper status bias, though less severe than in the *Digest* poll. This explains why Gallup got the winner correct, but was numerically off in predicting Roosevelt's victory *margins* in the 1940 and 1944 elections. These and other problems in the early polls, which included ceasing their interviews weeks before an election, were debated widely among pollsters and social scientists after the polling debacle in the 1848 presidential election, when the polls predicted incorrectly that Republican Thomas Dewey would defeat Democratic President Harry Truman (Mosteller et al., 1949). Gallup and other pollsters subsequently improved their sampling methods. Also beginning to conduct large-scale opinion and voting studies at this time, survey professionals and academic researchers engaged in a concerted effort to improve and defend opinion polling and to measure and examine closely public opinion and mass political behavior (see, e.g., Sheatsley and Mitofsky, 1992; Mitofsky, 1998; on election polling and its future, see Hillygus, 2011).

Emerging Uses of Public Opinion Research

Public opinion analysis advanced professionally in the United States with the establishment of the journal, *Public Opinion Quarterly*, in 1937 and the American Association for Public Opinion Research in 1947, followed later by other associations of American and international polling organizations and researchers (see Frankovic et al., 2009). While Gallup emphasized that polling could and should make the impact of public opinion felt more broadly in government and society (Gallup and Rae, 1940; cf Herbst, 1993), the journal was founded on the premise that mass opinion had already shown its importance and what was needed was research concerned with influences on public opinion – whether through constructive education and leadership, or less benign manipulation and propaganda (see the emphasis on leadership in Key, 1961) – and methodologies for studying these opinions, including especially polling.

University-based survey research had its beginnings in 1939–40 with sociologist Paul Lazarsfeld's Office of Radio Research, later renamed the Bureau of Applied Social Research at Columbia University. This history is described by Rossi (1959), Natchez (1985), and especially Converse (1987). Lazarsfeld and his colleagues in *The People's Choice* (Lazarsfeld et al., 1944) did the first sophisticated survey study of presidential elections in 1940 in Erie County, Ohio. This was followed by *Voting* (Berelson et al., 1954), a study of the 1948 election in Elmira, New York. These early studies were highly sensitive to methodological issues in survey research and contributed to the development and use of panel surveys (the interviewing of the same respondents over two or more points in time) and complex research designs. Two of the largest and most widely known research centers that developed the capacity to do long-term national surveys are the National Opinion Research Center (NORC), which was established in 1941 at the University of Denver and moved to the University of Chicago in 1947; and the University of Michigan's Survey Research Center (SRC, at the Institute for Social Research (ISR)), established in 1946. With the publication of *The Voter Decides* (Campbell et al., 1954), the funding and locus for national election surveys shifted to Michigan's SRC, which led to the publication of *The American Voter* (Campbell et al., 1960) and the continuing series of American National Election Studies and studies based upon them (Nie et al., 1976/1979; Miller and Shanks, 1996; Lewis-Beck et al., 2008; Bafumi and Shapiro, 2009; Abramson et al., 2012).

The years from Roosevelt's election through World War II were important for professionalized polling. Roosevelt and the Democratic Party had the first recognizable political pollster/consultant Emil Hurja in their 1932, 1934, and 1936 election victories (Holli, 2002). Government commissioned survey research played a significant role in national leaders' and administrators' dealing with issues related to the World War II mobilization efforts and the economy at home; this included monitoring reactions to prices, rationing, and war bond drives, and examining labor and other war-relevant policy areas (see Sparrow, 2011). A major large-scale survey study of the military was conducted by Samuel Stouffer et al. in *The American Soldier* (1965[1949]), which examined army life, morale, and ultimately issues related to demobilization at end of the war. Government surveyors became leaders in academic survey research after the war, such as at Michigan's ISR/SRC.

During the war Roosevelt had the first 'presidential pollster,' Hadley Cantril, whose Office of Public Opinion Research provided information about public opinion related to the conducting of the war, as Roosevelt and his advisors felt the constraints of public opinion and attempted to anticipate, lead, and – arguably – manipulate it. This began with Roosevelt easing the United States into aiding the British against Nazi Germany prior to the United States entering the war after the Japanese attack on Pearl Harbor; then his administration's efforts to gain support for giving priority to the war in Europe versus the war against Japan; through the final decision to so require unconditional surrender from both the Germans and the Japanese (see Casey, 2001). Into the postwar years and the Cold War, the State Department commissioned NORC to

conduct what were 'secret' foreign policy surveys. Upon learning of these surveys, Congress cut off funding for them, arguing that members of Congress were the representatives of public opinion and that separate government polling had no place in representing the people (Eisinger, 2003). This did not, however, preclude the State Department's United States Information Agency from conducting surveys abroad to inform American foreign policy-making.

Subsequent Expansion

Professional polling and academic survey research expanded after 1950 (see Frankovic et al., 2009). The polling landscape included additional public pollsters of note who provided their survey results to the mass media and other subscribers along with some free public releases of poll results (e.g., Louis Harris, Daniel Yankelovich, and later the Pew Research Center, among others). Polling experienced its greatest expansion beginning in the 1970s and especially the 1980s, when it was found that sufficiently representative and accurate polls could be done less expensively by telephone than in person (based on experiments comparing polls conducted by phone with those done face-to-face), eventually leading most pollsters to do their surveys by phone, as the penetration of phones nationwide reached the 95% level. *Random Digit Dialing* (RDD and variants of it) provided more direct probability samples of households with phones, and the advent of *Computer-Assisted Telephone Interviewing* (CATI) expedited telephone surveying and data processing. The national news media, which had relied on polling by public pollsters, began sponsoring and conducting their own polls, establishing their own, often joint telephone polling operations (e.g., CBS News/*New York Times*, and ABC News/*Washington Post* polls) or making arrangements to contract-out polls in which they controlled the content and reporting of results (see Gollin, 1980; Mann and Orren, 1992). Moreover, in this way the media had *independent* sources of data from the polling results that self-serving political parties, candidates, and interest groups might send their way or publicize. The media polls greatly expanded the polling data that political scientists and other researchers could draw on.

Nonpolitical government survey research continued for public administration and policy-related purposes, including U.S.-Census-related surveys and survey data collection bearing on public health and the nation's general welfare (Converse, 1987; see Frankovic et al., 2009). The political side of polling has been dominated by market researchers and others who expanded their enterprises into consulting for political parties and candidates (e.g., see Moore, 1992). Presidential polling has been paid for by the political parties or by campaigns, not through government funding. After Cantril's polling for the Roosevelt administration, the history of presidential pollsters picks up with Louis Harris working for candidate and later President, John F. Kennedy, as polling and political consultants became a staple of political campaigns and as leaders brought them along informally and later more formally as they entered office. Kennedy was followed by Presidents Lyndon Johnson and Richard Nixon, with pollsters Oliver Quayle, David Derge, the Opinion Research Corporation, and Robert Teeter (Market Opinion Research); President Jimmie Carter and pollster Patrick Caddell; President Ronald Reagan and pollster Richard Wirthlin; President George H.W. Bush and Robert Teeter; President Bill Clinton and Stanley Greenberg, followed later by Dick Morris, Mark Penn, and Douglas Schoen; and Presidents George W. Bush and Barack Obama with various pollsters/consultants. Most notably, again, this polling was not used to respond to mass opinion as Gallup had posited, but rather to help presidential administrations and other leaders determine how to move the public, especially voters, in the direction of their preferred policies or ideological goals (see Jacobs and Shapiro, 1995, 2000).

Academic survey research flourished as well. In addition to the American National Election Studies that started in 1948, the NORC General Social Surveys conducted since 1972 have been widely used by social scientists including political scientists. A number of state universities established survey research centers, which could be used to poll the nation, states, and localities. Later, private colleges and universities began doing polls for purposes of promoting their schools' visibility for fundraising and other purposes, some of which expanded to include media partnerships in doing especially state- and local-level polling. Perhaps most important, archives were established to assemble and make available polling results and data sets for academic and other research, including especially the surveys done by the public polling and media organizations. The leading archives in the United States include the Roper Center for Public Opinion Research (University of Connecticut), the Inter-university Consortium for Political and Social Research (University of Michigan), and the Odum Institute (University of North Carolina). These archives are crucial to historical research on public opinion in the United States, including studies of long-term opinion trends – and their causes and consequences – which utilize the available polling data since the 1930s (Page and Shapiro, 1992; Mayer, 1993; on the study of public opinion and 'American Political Development,' see Shapiro, 2014).

Methodological and Technical Issues

While polling conducted by these different organizations depends on an understanding and use of scientific sampling to draw inferences about a target population (e.g., national or local, the entire adult population or particular subpopulations, 'likely voters,' or elites), polls can be conducted in different ways depending on the research goal, the kind of survey questionnaire to be administered, and the available resources (i.e., money and time). In-person national polling – typically done by first randomly sampling geographic locations, then households in these locations, and then random sampling within households – is the most expensive and time-consuming survey mode because of travel and other costs. Face-to-face interviews can involve longer surveys, the use of visual aids, and 'secret ballots' (or turning over the interviewer's portable computer to respondents in Computer-Assisted Personal Interviews) for responses to sensitive questions.

In-person election 'exit polls' (usually with paper questionnaires that respondents can fill out) that are used as a factor in projecting election results and that, more importantly,

provide the basis for more in-depth statistical analysis of voting behavior, are short and efficient surveys, based on random samples of voters exiting election sites. Exit polling has been controversial in the United States due to the 'early' reporting of results that might affect the behavior of those who vote later in the day. Moreover, the data have to be statistically weighted for the actual vote results due to selection biases that have come to overrepresent Democratic Party voters. These and other complications contributed to controversies regarding the 2000 and 2004 presidential elections (see Frankovic et al., 2009).

Besides the advantages already noted, telephone polling is easier to monitor and validate than in-person interviews; however, it excludes households without telephones, and it has higher rates of nonresponse due to households that are not reached (no answers, busy signals, answering machines, and refusals to participate); it is also easier in person to make contact with respondents and establish rapport. Mail surveys are the least expensive but produce low response rates unless extensive follow-ups are made (Dillman et al., 2009). Low response rates and the increasing potential for selection or nonresponse bias have emerged as the most pressing methodological problems currently facing survey researchers, initially as the result of the proliferation of telephone surveys, telemarketing, and solicitations that interrupt and impose upon people's time. This problem has been compounded dramatically by the proliferation of the use of cell phones, which are more complicated to call in the United States where federal law prohibits the use of the automated dialers typically used in CATI systems (e.g., see Groves and Couper, 1998; Groves et al., 2009; Frankovic et al., 2009; Brick, 2011; Couper, 2011; Baker et al., 2013). Sampling cell phones, however, has the advantage of directly sampling individuals, but how readily individuals can – and will – participate in cell phones surveys adds another layer of cost and complexity. At this writing, more than 50% of the adult population is 'mostly' or only reachable by cell phone (Blumberg and Luke, 2013), so that telephone surveys must include cell phones.

These developments and added costs have led to efforts to employ new technology, which has contributed to the expansion in new directions of polling in the United States. This has included interactive voice response polls ('Robo-Polls') that are fully automated surveys with a recorded interviewer asking questions and respondents entering answers by touching numeric phone keys. These polls save interviewer costs and standardize the verbalizing of the questions (avoiding 'interviewer effects'), but they provide no control over respondent selection, do not include cell phones unless they include a manual element, and have lower response rates.

Further, to deal with costs as well as with the difficulty in reaching individuals by phone, and to expedite the polling process, online/Internet surveying has had substantial appeal. It provides greater ease with which respondents can complete self-administered surveys at their convenience, without the disruption of a phone call (and without interviewer effects). Researchers can administer complex questionnaires, include visual aids and graphic or multimedia presentations, and thereby, if desired, conduct greatly enhanced experiments with random assignment of 'treatments' than might be conducted through CATI or other means. While the penetration of online access has expanded enormously, Internet surveys present serious methodological challenges related to the ability to sample fully from the population they want to cover and thereby offer valid statistical conclusions. Since there is no available listing of all electronic e-mail addresses, or a method akin to RDD, it is impossible to directly draw a probability sample. One approach in the United States has been to use RDD, address-based sampling, or in-person survey probability sampling methods to draw random samples of adults, and then giving online access to respondents without such access in establishing 'Internet panels' of individuals who are available to participate subsequently in several surveys; there are then cost savings in not having to draw and survey multiple new samples from the population at large. A less expensive and increasingly used, but somewhat more controversial, method involves recruiting thousands of volunteers to join online panels recruited through invitations on Internet sites, and to adjust for any selection bias through sophisticated methods of 'propensity score' weighting or 'matching' as fully as possible to the demographic composition of the population. It is also possible to do surveys using mixed or multiple modes, but these have not been used extensively in public opinion and election surveys (see Frankovic et al., 2009; Dillman et al., 2009; Baker et al., 2013).

Other Sources of Error

Beyond the above issues of population coverage and nonresponse bias that can lead to the need, at minimum, to apply weights to the data, there are other sources of error in polling. Statistical 'sampling error' is the most widely cited and it refers to error in survey results that is a consequence of having data for a *sample* of a certain size and not for the entire population being sampled. There are, however, other important and less directly quantifiable sources of error. These include the potentially substantial effects of the *mode* of surveying used (in person, phone, mail/paper, online); question wording and question order or context effects (including the order in which response categories are ordered); coding and data processing errors; problems in imputing missing data; and the effects that interviewers – including their ascriptive characteristics such as race, gender, and age – can have on responses to questions. Particularly important, there may be *conceptualization* and *specification* problems that bear on the ultimate validity of opinion measures. For example, in some cases, researchers assume that respondents have genuine attitudes and opinions, or particular issues are highly salient, when they may not be, so that responses to questions may be transitory survey responses or 'nonattitudes.' Care needs to be exercised in drawing inferences about actual behavior from survey measures of opinion, and self-reports of future – or even past – behavior. *Reported* voting in a postelection survey may not match actual behavior confirmed by information on 'validated' votes. Last, errors can occur in how survey results are *reported* or *interpreted*. While issues of reporting and interpretation can be contentious, efforts to address this have involved making information on survey data more

widely available and transparent, especially through free online sources including '*aggregators*' of survey data (e.g., *pollster.com*, *realclearpolitics.com*, *fivethirtyeight.com*, *pollingreport.com*) and the 'transparency initiative' of the American Association of Public Opinion Research (e.g., see Schuman and Presser, 1981; Cantril, 1991; Mann and Orren, 1992; Traugott and Lavrakas, 2007; Frankovic et al., 2009; Dillman et al., 2009; Asher, 2012; Schaeffer and Dykema, 2011; Newport et al., 2013).

Contributions to Political Science: Public Opinion and Democracy

With technological and methodological issues under control, polling gave political scientists in the United States the opportunity to study empirically the nature and role of public opinion in American democracy, as well as to engage in basic social scientific research on the political psychology and behavior of individuals and of the public as a collective.

One initial debate as professional and academic polling began and that still continues concerns the conceptualization of 'public opinion' as an aggregation of individuals' equal opinions that can simply be counted up in polls. That is, polling led to *authoritative* statements about public opinion based on *data*. While from a populist standpoint, counting everyone equally in this way could be justified normatively, for critics of polling the crystallizing and channeling of public opinion into politics involved broader processes that included social forces, organized groups, political parties, and the role and context of institutions, which polls could not capture (see Blumer, 1948; Rogers, 1949). Some also argued that the existence of polling gives the public a false sense that its voice is sufficiently represented in the political process (Herbst, 1993).

Second, the findings of the early Columbia and Michigan academic studies of public opinion and voting fueled criticisms regarding the normative role of public opinion in a democracy. The early survey results raised questions about whether citizens exhibited the desired 'democratic competence' to warrant political leaders and government to be responsive to them. The public did not appear to fit the civics book expectations of knowledgeable and attentive citizens who held well reasoned and stable opinions, and who voted based on issues not on less substantive concern. Rather, individuals fell short in these respects but were influenced by affective partisan loyalties and other nonsubstantive concerns in their personal socioeconomic environments. This is still widely debated and a different picture of public opinion emerges in looking at the opinions and behavior of the public in the *aggregate* over time, not as separate individuals, and in changing social, economic, political, and informational contexts or environments. Patterns of statistical relationships found in public opinion data, and change and stability in public preferences over time, occur for identifiable reasons and are 'rational' in that sense. This does not mean that public is necessarily wise: the extent to which it depends on the information made available to it by the nation's leaders and others who offer guidance through the mass media – including what the media provide through independent analysis and commentary (see Key, 1961; Sniderman et al., 1991; Page and Shapiro, 1992; Glynn et al., 2004; Gilens, 2012; cf Converse, 1964; Zaller, 1992). The increasing partisan polarization and conflict in the United States that has occurred at the elite level since the 1970s, if not earlier, penetrated to the level of mass opinion, in a way that has produced partisan attachments among the public that are much more issue and ideologically based than these loyalties were found to be in the early voting studies, indicating that those findings that reflected badly on the American public were time bound. But it is uncertain what long-term consequences this change in the nature of public opinion will have for American politics (see Bafumi and Shapiro, 2009; Shapiro and Bloch-Elkon, 2008; Shapiro and Jacobs, 2011).

Polling has enabled political scientists to examine a normative and empirical question central to theories of democracy: what is the role of democratically elected governments and leaders with respect to public opinion? Should they respond to or lead the public? Have government leaders and policies been responsive to public opinion? The normative debate is an ongoing one (e.g., see Newport et al., 2013). Empirically, as noted earlier, polling has not been used explicitly by leaders and policy-makers to respond to public opinion. Nonetheless, there is substantial evidence that public opinion matters in ways consistent with the expected role of electoral incentives and accountability. Shapiro (2011; Newport et al., 2013) provides a review of the now-enormous body of empirical research, which ranges from studies of constituency representation, state-level politics, and policy-making; studies of consistency or 'congruence' between national policies and majority public opinion, to time series studies of trends in public preferences, ideology, and government policies. The overall finding from this work is a highly robust but far from complete connection between public opinion and policy-making. The relationship is stronger for highly salient issues and, under other conditions, consistent with electorally accountable leaders responding to the opinions of voters (see Erikson et al., 2002; Shapiro, 2011).

These robust correlations over extended time periods between public opinion and policies do not preclude the possibility that government may be more responsive to some segments of the public than others; that is, to those who are more affluent and thus revealing political inequality due to economic inequality. Survey research has long revealed in the political science research on political participation that there is a socioeconomic bias in participation (for a review, see Schlozman et al., 2012), and recent research has looked more closely at the policy consequences of this. Some research has found that the policy preferences of the affluent were better represented by U.S. senators than those of lower income groups, and that beyond legislative voting, national policies were more likely to be enacted when supported by larger percentages of the affluent as opposed to support from those with middle and lower incomes (Bartels, 2008; Gilens, 2012).

In addition to these debates about such important questions about democracy in the United States and elsewhere, survey research has also contributed to advances in basic social science research in the areas of political psychology and political communication, which have been concerned with the broad range of possible influences on public opinion and voting. Much important research has come from the use of

experimental methods in laboratory settings and in both telephone and Internet survey experiments with national cross-sectional samples of adults (see Mutz, 2011; Vavreck and Iyengar, 2011). The innovative 'deliberative polls' developed and conducted by Fishkin (2009) have included large-scale experiments to evaluate the effects of both information and group discussions/deliberations on the opinions of representative national survey samples of the American public that have been brought together in one location to learn about and debate important policy issues.

The range of media (including *agenda-setting*, *framing*, and *priming* effects) and other influences on public opinion that have been studied is enormous as well as are the types of opinions examined (see the chapters in Shapiro and Jacobs, 2011). These include both *cognitive* and *affective* psychological influences, including the role of personality (e.g., Marcus et al., 2000; Mondak, 2010; Lodge and Taber, 2013); the social and biological bases of public opinion, including gender; the role of values, ideology, and partisanship; racial and ethnic differences in public opinion and political behavior; the role of religion, class, and knowledge and sophistication; and age and socialization processes. These influences play out across voter choices and public opinion toward different domestic and foreign policy issues – from economic welfare and 'big government issues,' including regulation; to the history of civil rights, civil liberties, and criminal justice in the United States; to other present day issues including abortion, gay rights, immigration, environmental protection, energy policy, and gun control; and foreign policy toward all regions of the world, war, national security, and terrorism (e.g., see Page and Shapiro, 1992; Shapiro and Jacobs, 2011; Erikson and Tedin, 2011).

See also: Attitudes, Political and Public Opinion; Deliberation and Democracy; Democratic Theory; Majoritarianism and Majority Rule; Mass Media and Politics; Media Effects; Party Identification and Nonpartisanship; Political Communication; Political Representation; Politics and Participation; Voting, Explanations of: Social Class; Voting: Turnout.

Bibliography

Abramson, P.R., Aldrich, J.H., Rohde, D.W., 2012. Change and Continuity in the 2008 and 2010 Elections. CQ Press, Washington, DC.

Asher, H., 2012. Polling and the Public: What Every Citizen Should Know, eighth ed. Congressional Quarterly, Washington, DC.

Bafumi, J., Shapiro, R.Y., 2009. A new partisan voter. The Journal of Politics 71, 1–24.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., Tourangeau, R., 2013. Report of the AAPOR Task Force on Non-Probability Sampling. American Association for Public Opinion Research.

Bartels, L.M., 2008. Unequal Democracy: The Political Economy of the New Gilded Age. Russell Sage Foundation, New York. Princeton University Press, Princeton, NJ.

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting: A Study of Opinion Formation in a Presidential Campaign. University of Chicago Press, Chicago, IL.

Blumberg, S.J., Luke, J.V., 2013. Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July-December 2012. National Center for Health Statistics.

Blumer, H., 1948. Public opinion and public opinion polling. American Sociological Review 13, 542–549.

Brick, M., 2011. The future of survey sampling. Public Opinion Quarterly 75 (5), 872–888.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E., 1960. The American Voter. University of Chicago Press, Chicago, IL.

Campbell, A., Gurin, G., Miller, W.E., 1954. The Voter Decides. Row, Peterson, Evanston, IL.

Cantril, A.H., 1991. The Opinion Connection: Polling, Politics, and the Press. Congressional Quarterly, Washington, DC.

Casey, S., 2001. Cautious Crusade: Franklin D. Roosevelt, American Public Opinion, and the War against Nazi Germany. Oxford University Press, New York.

Converse, J.M., 1987. Survey Research in the United States: Roots and Emergence, 1890–1960. University of California Press, Berkeley, CA.

Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D. (Ed.), Ideology and Discontent. The Free Press, New York, pp. 206–261.

Couper, M.P., 2011. The future of modes of data collection. Public Opinion Quarterly 75 (5), 889–908.

Dalton, R.J., Klingemann, Hans-Dieter (Eds.), 2007. The Oxford Handbook of Political Behavior. Oxford University Press, Oxford.

Dillman, D.A., Smyth, J.D., Christian, L.M., 2009. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, second ed. Wiley, New York.

Eisinger, R.M., 2003. The Evolution of Presidential Polling. Cambridge University Press, New York.

Erikson, R.S., Tedin, K.L., 2011. American Public Opinion, eighth ed. Longman, New York.

Erikson, R.S., MacKuen, M.B., Stimson, J.A., 2002. The Macro Polity. Cambridge University Press, New York.

Fishkin, J.S., 2009. When the People Speak: Deliberative Democracy and Public Consultation. Oxford University Press, Oxford.

Frankovic, K.A., Panagopoulos, C., Shapiro, R.Y., 2009. Opinion and election polls. In: Pfeffermann, D., Rao, C.R. (Eds.), Handbook of Statistics 29A: Sample Surveys: Design, Methods, and Applications. North Holland, The Netherlands, pp. 567–596.

Gallup, G., Rae, S.F., 1940. The Pulse of Democracy. Simon & Schuster, New York.

Gilens, M., 2012. Affluence & Influence: Economic Inequality and Political Power in America. Russell Sage Foundation, New York. Princeton University Press, Princeton, NJ.

Glynn, C.J., Herbst, S., O'Keefe, G., Shapiro, R.Y., Lindeman, M., 2004. Public Opinion, second ed. Westview, Boulder, CO.

Gollin, A., 1980. Exploiting the liaison between polling and the press. Public Opinion Quarterly 44, 445–461.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowsky, J.M., Singer, E., Tourangeau, R., 2009. Survey Methodology, second ed. Wiley, New York.

Groves, R.M., Couper, M.P., 1998. Nonresponse in Household Interview Surveys. Wiley, New York.

Herbst, S., 1993. Numbered Voices: How Opinion Polling Has Shaped American Politics. University of Chicago Press, Chicago, IL.

Hillygus, D.S., 2011. The evolution of election polling in the United States. Public Opinion Quarterly 75 (5), 962–981.

Holli, M.G., 2002. The Wizard of Washington: Emil Hurja, Franklin Roosevelt, and the Birth of Public Opinion Polling. Palgrave, New York.

Jacobs, L.R., Shapiro, R.Y., 1995. The rise of presidential polling: the nixon white house in historical perspective. Public Opinion Quarterly 59, 163–195.

Jacobs, L.R., Shapiro, R.Y., 2000. Politicians Don't Pander: Political Manipulation and the Loss of Democratic Responsiveness. University of Chicago Press, Chicago.

Key Jr., V.O., 1961. Public Opinion and American Democracy. Alfred A. Knopf, New York.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., 1944. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. Columbia University Press, New York.

Lewis-Beck, M.S., Jacoby, W.G., Norpoth, H., Weisberg, H.F., 2008. The American Voter Revisited. University of Michigan Press, Ann Arbor, MI.

Lodge, M., Taber, C.S., 2013. The Rationalizing Voter. Cambridge University Press, New York.

Mann, T., Orren, G. (Eds.), 1992. Media Polls in American Politics. Brookings Institution, Washington, DC.

Marcus, G.E., Neuman, W.R., MacKuen, M., 2000. Affective Intelligence, and Political Judgment. University of Chicago Press, Chicago.

Mayer, W.G., 1993. The Changing American Mind: How and Why American Public Opinion Changed between 1960 and 1988. University of Michigan Press, Ann Arbor, MI.

Miller, W.E., Shanks, J.M., 1996. The New American Voter. Harvard University Press, Cambridge, MA.

Mitofsky, W.J., 1998. Was 1996 a worse year for polls than 1948? Public Opinion Quarterly 62, 230–249.

Mondak, J.J., 2010. Personality and the Foundations of Political Behavior. Cambridge University Press, New York.
Moore, D.W., 1992. The Superpollsters: How They Measure and Manipulate Public Opinion in America. Four Walls Eight Windows, New York.
Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S., Truman, D.B., 1949. The Pre-election Polls of 1948. Social Science Research Council, New York.
Mutz, D.C., 2011. Population-Based Survey Experiments. Princeton University Press, Princeton, NJ.
Natchez, P.B., 1985. Images of Voting: Visions of Democracy. Basic Books, New York.
Newport, F., Shapiro, R.Y., Ayres, W., Belden, N., Fishkin, J., Fung, A., Herbst, S., Lake, C., Page, B., Page, S., Pinkerton, J.P., Selzer, J.A., Warren, M., 2013. Polling and Democracy: Report of the AAPOR Task Force on Public Opinion and Leadership. American Association for Public Opinion Research.
Nie, N.H., Verba, S., Petrocik, J.R., 1976/1979. The Changing American Voter, enlarged ed. Harvard University Press, Cambridge, MA.
Page, B.I., Shapiro, R.Y., 1992. The Rational Public: Fifty Years of Trends in Americans' Policy Preferences. University of Chicago Press, Chicago, IL.
Rogers, L., 1949. The Pollsters: Public Opinion, Politics, and Democratic Leadership. Alfred A. Knopf, New York.
Rossi, P.H., 1959. Four landmarks in voting research. In: Burdick, E., Brodbeck, A.J. (Eds.), American Voting Behavior. The Free Press, Glencoe, IL.
Schaeffer, N.C., Dykema, J., 2011. Questions for surveys: current trends and future directions. Public Opinion Quarterly 75 (5), 909–961.
Schlozman, K.L., Verba, S., Brady, H.E., 2012. The Unheavenly Chorus: Unequal Political Voice and the Broken Promise of American Democracy. Princeton University Press, Princeton, NJ.
Schuman, H., Presser, S., 1981. Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context. Academic Press, New York.
Shapiro, R.Y., 2011. Public opinion and American democracy. Public Opinion Quarterly 75, 982–1017.
Shapiro, R.Y., 2014. Public opinion. In: Valelly, R., Mettler, S., Lieberman, R. (Eds.), The Oxford Handbook of American Political Development. Oxford Universiity Press, Oxford.
Shapiro, R.Y., Bloch-Elkon, Y., 2008. Do the facts speak for themselves? partisan disagreement as a challenge to democratic competence. Critical Review 20 (1–2), 115–139.
Shapiro, R.Y., Jacobs, L.R., 2011. The democratic paradox: the waning of popular sovereignty and the pathologies of American politics. In: Shapiro, R.Y., Jacobs, L.R. (Eds.), The Oxford Handbook of American Public Opinion and the Media. Oxford University Press, Oxford, pp. 713–731.
Sheatsley, P.B., Mitofsky, W.J. (Eds.), 1992. A Meeting Place: The History of the American Association for Public Opinion Research. American Association for Public Opinion Research, Ann Arbor, MI.
Smith, T.W., 1990. The first straw: a study of the origins of election polls. Public Opinion Quarterly 54, 21–36.
Sniderman, P.M., Brody, R.A., Tetlock, P.E., 1991. Reasoning and Choice: Explorations in Political Psychology. Cambridge University Press, Cambridge.
Sparrow, J.T., 2011. Warfare State: World War II Americans and the Age of Big Government. Oxford University Press, New York.
Squire, P., 1988. Why the 1936 literary digest poll failed. Public Opinion Quarterly 52, 125–133.
Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams Jr., R.M., 1965 [1949]. The American Soldier: Adjustments during Army Life, vol. 1. Wiley, New York.
Traugott, M.W., Lavrakas, P.J., 2007. The Voter's Guide to Election Polls, fourth ed. Rowman and Littlefield, Lanham, MD.
Vavreck, L., Iyengar, S., 2011. The future of political communication research: online panels and experimentation. In: Shapiro, R.Y., Jacobs, L.R. (Eds.), The Oxford Handbook of American Public Opinion and the Media. Oxford University Press, Oxford, pp. 156–168.
Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.

Political Science, History of

Erkki Berndtson, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political science as a discipline began to develop at American universities in the late nineteenth century. A new departmental structure of American universities, a birth of the disciplinary association and the founding of specialized publications created an academic community with transferable merits. After the war, the American model started to influence the study of politics worldwide, beginning in Western Europe. Today, the discipline is rapidly advancing especially in Asia and in Latin America. The past 60 years have witnessed political science becoming a pluralist discipline with different research traditions.

Study of Politics vs Political Science as a Discipline

Politics has been discussed, evaluated and studied since the ancient times. It has also been taught at universities for centuries. However, the academic discipline of political science dates only from the late nineteenth century. This contradiction is due to the nature of academic disciplines. Although universities have existed at least since the founding of the University of Bologna in 1088, the medieval universities did not have much in common with modern universities starting to develop in the nineteenth century. New European higher education ideologies (Humboldt in Germany, Newman in Britain, the Napoleonic reforms in France) changed universities thoroughly. At the same time, American universities adopted and transformed European ideologies into a new mold. An important American innovation was the departmental structure of universities, which was created between 1890 and 1910. At that time, it was an internationally unique system. Departments stood between individual professors and the university as a corporate body. They allowed the better recognition of disciplines and formed a group of scholars with transferable merits (Abbott, 2001: pp. 122–123).

In spite of the long heritage of political thinking in Europe, from the Sophists to Hegel and Marx, the modern political science was born in the United States. The main reason for this is the organizational and institutional change of American universities. On the other hand, university departments are not enough for a subject of teaching to be a discipline. A discipline needs a community of scholars interacting and communicating with each other across institutions. It also needs publishing outlets, journals, and disciplinary series. Also in this respect, American political scientists created conditions for the separate political science discipline in the late nineteenth/early twentieth century.

Scientific disciplines also need a favorable social environment to develop. For political science, free interaction and communication between scholars are basic conditions for research. That is why political science, more than many other disciplines, has been tied to the development of democracy (Easton et al., 1995). Without the freedom of association and speech, the study of politics has been one sided, narrow, and ideological. For this reason, authoritarian tendencies have often limited the study of politics in many countries. The United States as 'the first modern democracy' had an advantage over many countries also in that respect. As Alexis de Tocqueville (1805–59) wrote in his *Democracy in America* (1835–40), 'a new political science is needed for a world itself quite new' (Tocqueville, 1994: p. 12).

The Early Years of the Discipline of Political Science

The birth of political science as a discipline can be illustrated by comparing American and European trajectories in the study of politics. It is sometimes claimed that the first political science professorship was founded at the University of Uppsala in Sweden in 1622 (a professor of Discourse and Politics). However, a similar professorship had been founded at the Dutch University of Leiden already in 1613. Politics had also been taught at many German universities since the fourteenth century. It became an important subject of teaching at the Faculties of *Staatswissenschaft* (the sciences of the state) in the nineteenth century. Mainly for political reasons, the subject disappeared after the 1848/49 political upheavals. In the late nineteenth century, however, the teaching of politics was introduced as *Allgemeine Staatslehre* (the general theory of the state) at German law faculties. In that sense, the study of politics has been well rooted in European universities for centuries. However, before the World War II, it was conducted undifferentiated from neighboring fields, which later became disciplines, such as history, philosophy, and constitutional law (see the chapters in Klingemann, 2007).

At the beginning of the disciplinary development, American political science did not differ much from the study of politics in Europe. Teaching and research were based on European intellectual traditions. Francis Lieber (1798–1872), who has been called the first American political scientist (Farr, 1990), was a German, who had emigrated to the United States in 1827 after having studied in German universities (as well as being wounded at the Battle of Waterloo and participating in the Greek War of Independence in 1821). Starting his life in the United States as a swimming instructor, he became a Professor of History and Political Economy at the South Carolina College in 1835. In 1857, he moved to the Columbia College in New York, becoming a Professor of History and Political Science, the first time political science was recognized as part of an academic title. Through his career, Lieber had close relations with European scholars (being, for instance,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03042-7

one of Tocqueville's sources of information on American democracy).

Furthermore, around the turn of the twentieth century, hundreds of American PhD students went to study in Europe, mainly in Germany, and introduced German scholarship and teaching methods in American academia after their return (Haddow, 1939: p. 172). One of these 'exchange scholars' was John W. Burgess (1844–1931), who studied in Germany and France during 1869–73 and initiated a plan for the School of Political Science at Columbia College (later University). One of Burgess's main inspirations for the School was the *Libre Ecole des Sciences Politiques*, which had been founded in Paris in 1872 to train civil servants, diplomats, and journalists as well as to educate laymen in politics, economics, sociology, journalism, and law. The Columbia School was opened in 1880 and became an early center of the discipline of political science. Burgess himself became a Professor of Political Science and Comparative Constitutional Law. Because of the founding of the School, Burgess has often been taken as one of the founding fathers of political science (Somit and Tanenhaus, 1982).

The model of the *Ecole Libre* was imitated not only at Columbia, but in many European countries as well. Some of the new institutions which were founded included, e.g., the *Facoltà di scienze politiche* in Florence (1874), schools for political and social sciences at *the Catholic University of Louvain* and at the *Free University of Brussels* (1893), the *London School of Economics and Political Science* (1895), the *Ecole des Sciences Sociales et Politiques* at Lausanne (1902), the *Deutsche Hochschule für Politik* in Berlin (1920), and a *Higher School of Political Sciences* in Athens (1927) (see the chapters in Klingemann, 2007).

The early Columbia School reminded its European counterparts. The journal, *Political Science Quarterly*, which the Columbia School had begun to publish in 1886, manifested this well. It was a journal for political *sciences*, the articles were written as well by economists, historians, and public law scholars as political scientists. The focus of political science was on the study of the state. In the first issue of the journal, the managing editor Munro Smith wrote, "[t]he principal legal question of the day, in our country is: to what organ or organs of the state shall the development of law be trusted – to the judicial and legislative, or to the legislative alone? … Whether the increasing of the state be deplored or applauded, the fact remains that it is rapidly becoming, if it is not already, the central factor of social evolution" (Smith, 1886: p. 8). In the 1890s, the School developed into a Faculty with three internal groupings, 'economics and social science', 'history and political science', and 'public law and comparative jurisprudence'. The Department of Political Science evolved only later from the last of the three groupings (Somit and Tanenhaus, 1982).

If Columbia had close relations with German scholarship and public law, another important political science institution, the Johns Hopkins University, was influenced by historical scholarship and the Anglo-Saxon tradition. Johns Hopkins began its first semester in 1876. In 1882, a Department of Historical and Political Science was founded with Herbert Baxter Adams (1850–1901) in charge of the Department. In 1883, Adams started the book series, *The Johns Hopkins Studies in Historical and Political Science*, which was one of the first publishing outlets in the field of political science. Although also Adams had studied in Germany he was influenced more by the British historical scholarship.

Columbia and Johns Hopkins were the two main universities usually recognized as the 'founders' of political science as a discipline. However, there were similar developments in Cornell, the University of Michigan and Yale. Besides, the first separate political science department was created at the University of Chicago when it opened its doors in 1892. At the same time, Chicago established the separate departments of political economy, history, sociology, and philosophy.

Although the American and European study of politics did not differ much from each other intellectually in the late nineteenth/early twentieth century, organizationally American political science began to develop rapidly. By 1914, there were 38 separate political science departments at American universities. Political science was also offered in 216 departments with other disciplines (Anderson in Haddow, 1939: p. 263). The number of independent political science departments increased and with them the number of political scientists. The *American Political Science Association* (APSA) was founded in 1903. In 1904, the membership of APSA was 214, by 1920 it had increased to some 1300. Separate political science departments, a disciplinary association and specialized publications in the field (the first real political science journal, the *American Political Science Review*, was founded in 1906) made the study of politics a discipline in the United States, while in Europe it remained part of other academic subjects without any organized interaction between the students of politics.

Political Science after the World War I

Between the World Wars I and II, the American and European study of politics began to differ from each other intellectually. Political participation, political parties and groups, became a new focus of research in the United States. The study of the functioning of mass democracy required new research methods. Historical and judicial analyses were replaced by the use of current documents and first attempts to use interviews as data and statistics as a research method appeared. Psychology was seen as an important discipline to explain citizens' political activity. Other social sciences, especially sociology, were seen as allies instead of the old partners, history, and constitutional law. The department of political science at the University of Chicago under the leadership of Charles E. Merriam became an innovative center of the discipline focusing on *New Aspects of Politics* (Merriam, 1925).

Old concepts were replaced by new ones. Power gave a more dynamic view on politics than sovereignty. One of the theoretical innovations was the theory of pluralism. Although the concept had European roots, as it had been introduced earlier by such scholars as Leon Duguit in France and Harold Laski in England, in the United States it became a theory that emphasized group activity as the basis of politics. Pluralism focused on the limitations of state sovereignty and argued that the state was not the only unit in society, which commanded the allegiance of its subjects. More importantly, pluralism was used as a theoretical framework in empirical studies on political parties and pressure groups.

Empirical research also influenced a new theory of democracy, which later has been called 'the relative theory of democracy' (Purcell, 1973). It emerged as a response to national and international challenges. In the 1920s, American political scientists discussed the unforeseen tendencies of democracy, such as low political participation (nonvoting) and the role of special interests in politics. These tendencies did not fit well with the ideal of democracy. In the 1930s, American political scientists still faced another problem. The development of authoritarian political systems in Europe and in Asia posed a threat to American democracy struggling with economic recession. A new theory of democracy tried to solve these internal and external challenges. It was argued that nonvoting was not a problem, as it guaranteed the stability of the political system. Instead, the most important condition on democracy was a democratic culture, a common way of acting. Furthermore, a new theory praised the value of compromises in politics. Democracy was an agreement to disagree. There was no place for absolutes in politics.

The relative theory of democracy emphasized empirical research. However, it soon had to face a theoretical challenge, which was posed by European refugees, especially from Germany. Many of them represented philosophical traditions, which challenged the principles of the new American empirical research. These refugee scholars influenced theoretical discussion in American social science (e.g., Leo Strauss, Hannah Arendt, Max Horkheimer), but at the same time, they were creating a rift between the empirically minded and more theoretically oriented political scientists, leading into a division between empirical political science and normative political theory (Gunnell, 1993).

Political Science after the World War II

After the World War II, the model of empirical American political science was introduced in other parts of the world. For a number of reasons, the American influence was the biggest in Western Europe. The time was ripe for the empirical study of politics in many European countries, as many thought that a better understanding about politics could have prevented the catastrophe of the War. At the same time, UNESCO wanted to create international social science associations. Political science was one of the first of these (sociologists establishing their own international association at the same time).

The *International Political Science Association* (IPSA) was founded in 1949 under UNESCO's auspices although political science as a discipline hardly existed outside North America (Coakley and Trent, 2000). IPSA was designed as an association with collective members (national political science associations). However, in addition to the American association, only the Canadian (established in 1929), the Finnish (1935), the Indian (1938), and the Japanese (1948) associations were in existence. The Chinese association (1932) had become inactive during the war. At the beginning, IPSA had to rely on individual political scientists to be able to begin its activities. In the meeting, which drafted the Constitution in 1949, it was decided that IPSA would become into legal existence, when four national associations had joined it. This happened later in 1949 after the French association had been set up and became a founding member of the IPSA together with American, Canadian, and Indian associations. Other national associations followed. The first of these were established in the United Kingdom (1950), the Netherlands (1950), Belgium (1951), and Germany (1951).

At the beginning, most European scholars were skeptical about the American model, which emphasized the scientific study of politics as an independent discipline. In the late 1940s, the general opinion in Europe was that "[p]olitical science, as a distinct branch from speculation concerning political phenomena or the history of these phenomena, is of fairly recent development, more recent, certainly, than other social sciences such as law, political economy and sociology" (Salvadori in UNESCO, 1950: p. 1). Besides, there were differences between the British, French and German traditions in the study of politics. These differences were displayed also in other countries, as the British study of politics was influential in most Commonwealth countries (including India), the French legal tradition was followed in the Mediterranean countries, and even in many Latin American countries, while the German tradition of *Allgemeine Staatslehre* had followers in Austria, the Netherlands, Scandinavian countries, as well as in Japan (Coakley and Trent, 2000: p. 3).

The British scholars argued that the history of political ideas must have a prominent place in political theory and one cannot understand government without knowing its historical background and growth (Robson in UNESCO, 1950: p. 306). The best example of the British skepticism toward American political science was Bernard Crick's book *The American Science of Politics* (1959), which argued that the idea that politics can be understood by the method of the natural sciences is embedded in American political culture (Crick, 1959: p. v). On the other hand, French scholars thought that the term 'political science' is familiar mainly to scholars in the Anglo-Saxon countries. In France, one was used to understand the study of politics as part of the "political sciences", as "practically all of the social sciences could be qualified, at least in certain aspects, as political sciences" (Kopelmanas in UNESCO, 1950: pp. 647–648). Germans were more divided in their attitude toward the scientific study of politics as an independent discipline. The lost war had discredited the German scholarship and the study of politics was highly fragmented after the war. At the same time, some returning émigré scholars introduced American political science in Germany, which made German political science responsive to the idea of the scientific study of politics.

In spite of the early skepticism about the idea of political science as an independent discipline, the separate departments of political science were founded in many West European universities between 1945 and 1960. The pace of development varied, however, as the discipline took off first in northern Europe, where the old German *Allgemeine Staatslehre* tradition was replaced with the American model of political science. The British and German political scientists started to adopt American theories and methods only later. Close relations with American universities and the founding of new universities in both countries helped to move the discipline closer to American political science in the 1960s.

For cultural and organizational reasons the disciplinary development in France and in southern Europe was much

slower. The French political science has been a divided discipline with the majority of political science chairs at law faculties and the minority at eight *Institut d'Etudes Politiques*, where political science has often been understood as political sociology. Although the situation has changed, the idea of an independent political science discipline raises still opposition among some French quarters. Organizational difficulties have also hindered the development of Italian political science. In Italy, the first political science chair was established at the University of Florence as late as 1966. In Italy, political science has been part of the faculties of political sciences, where paradoxically the discipline has been a minor subject. The same organizational difficulties have been noticeable also in Portugal and Spain. Besides, because of the authoritarian political systems until the 1970s, the discipline has been able to develop in these two countries only since the 1980s (see the chapters in Klingemann, 2007).

The organizational development of the discipline has been still slower in the other parts of the world. The idea of political science as an independent discipline has been accepted in many countries only since the 1960s on (Jinadu, 1991 on Africa; Atal in Easton et al., 1995 on Asia).

American Behavioralism

The themes that had emerged during the interwar period in American political science developed further in the 1950s and 1960s. The discipline adopted a more systematic perspective on the study of politics. One of the early calls for a more scientific approach was David Easton's book *The Political System* (1953) which criticized American political science about hyperfactualism (empiricism without theory). The new political science was soon labeled as behavioralism. Dwight Waldo has expressed the change in the following way:

> By almost any measure the most important aspect of post–World War II political science has been the rise of 'behavioralism': the controversies it engendered, its success to coming to dominate much of organized political science, the changes it brought in the matters to which political scientists attend, and the manner in which they are addressed.
>
> (Waldo, 1975: p. 58)

The heyday of behavioralism was from the mid-1950s to the mid-1960s (Somit and Tanenhaus, 1982: p. 185). It was the main intellectual approach that 'Americanized' the study of politics in Europe and in the other parts of the world. Writing about European political science in the early 1990s, Ken Newton and Josep Vallès noted that "it is, nevertheless, true that the 'behavioral revolution' transformed large parts of political science in Western Europe, and its effects were so profound that some on the European side of the Atlantic claim that their traditions of political inquiry were abandoned or forgotten to such an extent that the American approach colonised the West European profession" (Newton and Vallès, 1991: pp. 234–235). Although Newton and Vallès see the colonization thesis as an exaggeration, they do not deny the American influence. This influence was also due to an increasing interest in comparative politics. American political scientists traveled around the world gathering data and trying to develop a cross-cultural theory of politics (see the chapters in Daalder, 1997).

In spite of its importance, behavioralism has not been easy to define. Most histories of political science agree that its origins are "somewhat obscure" (Waldo, 1975: p. 58) and "attempts at coming to any complete definition … are probably futile given the diversity of those who followed its banner" (Seidelman and Harpham, 1985: p. 151). However, one of the most influential definitions has been David Easton's list of the characteristics of behavioralism: (1) search for regularities, which can be expressed in generalizations or theories with explanatory and predictive value, (2) the verification of propositions by testable data, (3) the development of research techniques which can be used for the rigorous observation and analysis of behavior, (4) the precision of data and findings with quantification and measurement when possible and relevant, (5) keeping ethical evaluation and empirical explanations analytically distinct, (6) the systematization of research with theory and empirical research as intertwined parts of a coherent body of knowledge, (7) the application of knowledge as part of the scientific enterprise, but pure science preceding the application of knowledge, and (8) the integration of political science with other social sciences to analyze the whole human situation (Easton, 1965: p. 7).

One of the main dividing lines within behavioralism has been between those who conduct empirical studies based on surveys (mainly voting studies) and those who have wanted to develop theoretical (conceptual) frameworks for political analysis, such as systems analysis (David Easton), structural functionalism (Gabriel A. Almond), or cybernetics (Karl W. Deutsch). On the other hand, the most important empirical theory that emerged as part of behavioralism was the pluralist theory of democracy. One of its main early works was Robert A. Dahl's *Who Governs?* (1961). Dahl criticized the elitist conception of power (the existence of one power elite in society) and described politics as a struggle for power between many unequal elites. Different elites were making decisions in different policy areas.

Postbehavioral Revolution

In the end of the 1960s, the opponents of behavioralism launched a counterattack, which changed the contours of the discipline. Criticism arose from different sources, conservative, radical, and liberal. Conservative political philosophers were often unsatisfied with the results and ideology of empirical political science. The most outspoken critics were Leo Strauss and his followers, who accused behavioralism for relativism and for pretending to represent value free research while promoting liberal democracy (Storing, 1962). Another source was a heterogenous group of left wing academics, who founded the *Caucus for a New Political Science* in 1967 in response to the social upheavals of the times. Many of these scholars participated in the antiwar, civil rights, and feminist movements. The postbehavioral revolution included even liberals, often policy analysts hoping to make political science a more socially relevant discipline.

In 1969, it was David Easton again, who summarized the views of this heterogenous movement as postbehavioral revolution and identified seven of its tenets: (1) substance must precede technique, as it is better to be relevant and sometimes vague rather than nonrelevantly precise, (2) behavioral science conceals an ideology of empirical conservatism, (3) behavioral research has lost touch with reality, (4) research about and constructive development of values are part of the study of politics and one must be aware of one's own value premises, (5) political scientists bear the responsibility of all intellectuals to protect the humane values of civilization, (6) knowledge means the responsibility to act and to engage in reshaping society, and (7) the politicization of political science is inescapable as well as desirable (Easton, 1969: p. 1052).

The postbehavioral revolution ended the hegemony of behavioralism. However, behavioralism has not disappeared, but exists along new competing research traditions. Political philosophy, policy analysis, and various critical approaches, including historical institutionalism, have become strong research areas. In addition, since the publication of Anthony Down's *The Economic Theory of Democracy* (1957), rational choice theory has become a widely held paradigm in American political science. The theory has many variations, however, from the fiscal conservative public choice theories to William H. Riker's studies on voting and group behavior and to a Nobel laureate Elinor Ostrom's *Governing the Commons* (1990). These developments have led to the fragmentation of political science into competing theoretical schools.

Fragmentation of Political Science

As the history of political science shows, the development of the discipline has proceeded through continual attempts to legitimize and delegitimize various research traditions. In that sense, it is no wonder that the so-called *Perestroika* movement emerged in 2000 to oppose the strong position of rational choice theory in American political science (Monroe, 2005). The movement criticized the elitist nature of the *American Political Science Association* with its white and male leadership. It also criticized the irrelevancy of the *American Political Science Review*, which was seen to contain mainly articles based on statistics and game theory, while ignoring articles based on other research methods. The revolt has led into a more pluralist American discipline, as the APSA has responded to criticism by renewing its decision-making procedures, encouraging qualitative research and offering new publishing outlets to its members. The change has vitalized the discipline, but at the same time it has become still more fragmented into different theoretical schools.

The American postbehavioral revolution coincided with the strengthening of the discipline in Western Europe. The *European Consortium for Political Research* was established in 1970 (Newton and Boncourt, 2010). This has led to an increasing cooperation between European political scientists and for further growth of the discipline. The breakdown of the European communist regimes in 1989/91 has also changed the landscape of political science (Eisfeld and Pal, 2010). Still in 1982, the editor of the *International Handbook of Political Science*, William G. Andrews (1982: p. 3), estimated that the American political scientists represented 75–80% of the world's supply. Recently, however, Hans-Dieter Klingemann (2008: p. 376) has estimated that around 2005 there were some 10 000 political scientists in Europe, half of them being in the Central and Eastern Europe. That roughly equals the number of American political scientists. Besides, the growth of the political science profession has been impressive in Asia during the last 30 years. There are some 5000 political scientists in Japan and some 2500 in South Korea alone (Inoguchi, 2012: p. 17). In addition, India has a long tradition of political studies. The discipline has developed rapidly also in Latin America during the last two decades. Although there are no exact figures available, it is safe to say that the increase of the number of political scientists around the world during the past 30 years has made political science a global discipline and diminished the previous central role of American political science in the profession.

Especially many European scholars have influenced theoretical discussion in the social sciences all over the world. German critical theory (Jürgen Habermas), French philosophy and sociology (Michel Foucault), as well as British intellectual history (Quentin Skinner) have opened new perspectives also for the study of politics. There are no paradigmatic theories in contemporary political science, but many competing theories and intellectual cultures. The globalization of the discipline has increased the diversity of theoretical perspectives. Different countries have their own intellectual traditions and scholars have become aware that they have to take into account local as well as global conditions when explaining politics. A good example is a study of democracy. There is no single theory of democracy today. Instead, the theories of deliberative, participatory, delegative, cosmopolitan, and global democracy, among others, are offered as solutions to the problems of contemporary politics.

The fragmentation of the discipline is displayed by the growth of different subdisciplinary specialties. Many scholars identify themselves more as members of a specific research field than as political scientists in general. International relations and public administration have founded their own associations and journals, and especially the study of public administration has in many countries cut itself off from the rest of political science. The identity politics has also become important, especially in regard to gender and race. Feminist political theory has become one of the fastest growing fields within political science.

Globalized Political Science

This historical account has focused mainly on the United States and Europe. This can be justified by the fact that the birth of political science as a discipline has been closely linked to the American and European scholarship. One cannot deny the organizational and theoretical influence of American political science in other countries (Inoguchi, 2012: p. 11). Since the 1970s, also European political science has grown in importance. This is not to deny the role of political science in other parts of the world. However, political science is still an unevenly developed discipline worldwide. One of the signs of this is the existence of political science associations in the world

and their membership in the *International Political Science Association*. Currently (2013) there are 51 collective IPSA members, 31 of them are from Europe, 2 from North America, 3 from Africa, 8 from Asia, 6 from Latin America, and 1 from Oceania. However, the change in that respect is evident in the future, as recent developments in Asia and Latin America show. Political science has become an international discipline, but it has not reached its limits yet.

See also: Cultural Studies of Science; Dahl, Robert A (1915–2014); Democracy: Normative Theory; Democratic Theory; Discipline Formation in the Social Sciences; Educational Institutions, History of; Educational Systems: North America; Intellectual History; Knowledge, Sociology of; Pluralism; Political Sociology; Political Theory, History of; Quantification in the History of the Social Sciences; Science and Politics: Value Neutrality; Science, History of; Science, Sociology of; Scientific Disciplines, History of; Scientific Knowledge, Sociology of; Social Science and Universities; Tocqueville, Alexis de (1805–59).

Bibliography

Abbott, A., 2001. Chaos of Disciplines. The University of Chicago Press, Chicago.

Andrews, W.G., 1982. Introduction: freaks, rainbows, and pots of gold. In: Andrews, W.G. (Ed.), International Handbook of Political Science. Greenwood Press, Westport, CT, pp. 1–6.

Coakley, J., Trent, J., 2000. History of the International Political Science Association 1949–1999. IPSA, Dublin.

Crick, B., 1959. The American Science of Politics. Its Origins and Conditions. Routledge and Kegan Paul, London.

Daalder, H. (Ed.), 1997. Comparative European Politics. The Story of a Profession. Pinter, London.

Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven.

Downs, A., 1957. An Economic Theory of Democracy. Harper & Row, New York.

Easton, D., 1953. The Political System: An Inquiry into the State of Political Science. Knopf, New York.

Easton, D., 1965. A Framework for Political Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Easton, D., 1969. The new revolution in political science. The American Political Science Review 55 (4), 550–559.

Easton, D., Gunnell, J.G., Stein, M.B. (Eds.), 1995. Regime and Discipline. Democracy and the Development of Political Science. The University of Michigan Press, Ann Arbor, MI.

Eisfeld, R., Pal, L.A. (Eds.), 2010. Political Science in Central-East Europe. Diversity and Convergence. Barbara Budrich Publishers, Opladen & Farmington Hills.

Farr, J., 1990. Francis Lieber and the interpretation of American political science. The Journal of Politics 52 (4), 1027–1049.

Gunnell, J.G., 1993. The Descent of Political Theory. The Genealogy of an American Vocation. The University of Chicago Press, Chicago and London.

Haddow, A., 1939. Political Science in American Colleges and Universities 1636–1900 (Edition with an Introduction and Concluding Chapter by Anderson, W.). D. Appleton-Century Company, New York and London.

Inoguchi, T., 2012. Political science in three democracies, disaffected (Japan), third-wave (Korea) and possibly fledgling (China). In: Trent, J., Stein, M. (Eds.), The World of Political Science: A Critical Overview of the Development of Political Studies Around the Globe: 1990–2012. Barbara Budrich Publishers, Opladen, Berlin, and Toronto, pp. 11–39.

Jinadu, L.A., 1991. Political science in 'Anglophone' Africa. Its context and developmental logic in historical perspective. In: Easton, D., Gunnell, J.G., Graziano, L. (Eds.), The Development of Political Science. A Comparative Survey. Routledge, London and New York, pp. 252–274.

Klingemann, H.-D. (Ed.), 2007. The State of Political Science in Western Europe. Barbara Budrich Publishers, Opladen and Farmington Hills.

Klingemann, H.-D., 2008. Capacities: political science in Europe. West European Politics 31 (1–2), 370–396.

Merriam, C.A., 1925. New Aspects of Politics. The University of Chicago Press, Chicago.

Monroe, K.R. (Ed.), 2005. Perestroika! the Raucous Rebellion in Political Science. Yale University Press, New Haven and London.

Newton, K., Boncourt, T., 2010. The ECPR's First Forty Years 1970–2010. ECPR, Colchester.

Newton, K., Vallès, J.M., 1991. Introduction: political science in Western Europe, 1960–1990. European Journal of Political Research 20 (3–4), 227–238.

Ostrom, E., 1990. Governing the Commons. The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.

Purcell Jr., E.A., 1973. The Crisis of Democratic Theory. Scientific Naturalism & the Problem of Value. The University of Kentucky Press, Lexington.

Seidelman, R., Harpham, E.J., 1985. Disenchanted Realists. Political Science and the American Crisis, 1884–1984. State University of New York Press, Albany, NY.

Smith, M., 1886. Introduction: the domain of political science. Political Science Quarterly 1 (1), 1–8.

Somit, A., Tanenhaus, J., 1982. The Development of American Political Science: From Burgess to Behavioralism. Irvington Publishers, Inc., New York.

Storing, H.J. (Ed.), 1962. Essays on the Scientific Study of Politics. Holt, Rinehart and Winston, New York.

Tocqueville, A. de, 1994 (orig. 1835–1840). Democracy in America. G. Lawrence, Trans. and Mayer, J.P. (Ed.). Fontana Press, London.

UNESCO, 1950. Contemporary Political Science. A Survey of Methods, Research and Teaching. UNESCO, Paris.

Waldo, D., 1975. Political science: tradition, discipline, profession, science, Enterprise. In: Greenstein, F.I., Polsby, N.W. (Eds.), Handbook of Political Science, vol. 1. Addison-Wesley, Reading, MA, pp. 1–130.

Relevant Websites

http://www.apsanet.org/ – American Political Science Association.
http://www.auspsa.org.au/ – Australian Political Studies Association.
http://www.ecpr.eu/ – European Consortium for Political Research.
http://www.ecpsa.org/ – European Confederation of Political Science Associations.
http://www.ipsa.org/ – International Political Science Association.
http://www.jpsa-web.org/ – Japanese Political Science Association.
http://alacip.org/ – Latin American Political Science Association.
http://www.saaps.org.za/ – South African Association of Political Studies.

Political Sociology

Tyson Patros and Judith Stepan-Norris, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Allardt, volume 17, pp. 11701–11706, © 2001, Elsevier Ltd.

Abstract

Political sociology is a major subfield on the border of sociology and political science, combining explanatory factors and research interests of both disciplines. It developed from the work of the early founders of social science to include areas of inquiry that tackle the salient social and political processes of the twentieth century and the contemporary world. We discuss important theoretical and empirical advancements and the subfield's continuous preoccupation with the interrelationships of 'politics' and 'society', political and social changes, and with issues of power, domination, and exploitation.

Political sociology bridges the fields of sociology and political science by addressing issues of power and authority with a focus on state/civil society relations. Political sociology differs from political science in that it includes and often focuses on the civil society side of the equation rather than placing an emphasis on the state and/or political elites. Core areas of research include state formation and change, forms of political rule, major social policies, political institutions and challenges to them (including reform-oriented and revolutionary social movements), political parties and the social bases of political attitudes and behaviors, class/power relations, and the political consequences of globalization. The field includes distinct major approaches, yet theoretical combination and synthesis is common. Many early and contemporary studies utilize comparative historical analysis, especially with regard to critical junctures and historical processes and developments, whereas current work has become methodologically more eclectic. Contemporary political issues and events, regimes in power, and cases relating to the United States and Europe tend to garner the most scholarly attention, though there is a steadily growing body of theory and empirical work beyond the core capitalist democracies.

We consider four somewhat distinct periods in the development of political sociology, while prioritizing the most recent period: (1) We trace political sociology's inspiration in the mid-nineteenth century to the earlier philosophers who considered the state and social life, but note that it was not until the founding fathers of nineteenth-century social science began thinking of society and the state as distinct entities and analyzing the relationships between them, that political sociology was born; (2) The post–World War II period, which shifted the subfield's focus to the prerequisites of democracy and voting patterns; (3) The late twentieth-century period, which saw political sociologists move toward a focus on state building and political change, using Marxist theories of capitalism and class as well as other theories; and (4) The contemporary period, which is characterized by a proliferation of topics relevant to politics such as globalization, race, gender, and culture. (Sections Historical Developments through Mid- and Late Twentieth-Century Period are largely abridged, paraphrased, and enhanced versions of the previous survey by E. Allardt in this publication. See Allardt (2001) for full references to these sections. Additionally, for broad works on political sociology, see Alford and Friedland, 1985; Janoski et al., 2005; Amenta et al., 2012.)

Historical Developments

The foundations of social sciences emerged in a context of epochal social and political transformations, including the development and unfolding of capitalist social systems and the modern state. These developments encouraged conceptualizations of the state and society as separate entities, making it possible to investigate their interrelations. Relying on earlier works on social life and politics by philosophers, historians, and legal theorists, sociology's three founding fathers made important contributions to political sociology's beginnings. Karl Marx challenged the idea that power concentrates solely in political offices and officeholders. Instead he focused on politics as emerging primarily from class conflict, just as the bourgeoisie and industrial working classes were becoming the principal social groups. He also emphasized the role of ideology in sustaining the powers that be. Max Weber responded to and critiqued Marx and added important analyses of authority, status categories, and social institutions (especially bureaucracy). Sociology's third founding father, Emile Durkheim, did not have an explicitly political focus, yet his emphasis on social order, integration, and solidarity had important implications for analyses of society and politics. Others more directly associated with the emergence of political sociology include: Talcott Parsons, who was inspired by Durkheim, Vilfredo Pareto, and Gaetano Mosca, who made independent contributions by emphasizing the role of elites for social and political change (along with others who continued this concern, including Robert Michels, C. Wright Mills, G. William Domhoff, Alexis de Tocqueville, and Seymour Martin Lipset).

Marx and Class Domination, Capitalism, and Revolution

Marx was among the first to articulate an empirical political sociology, marking out historical materialism in his systematic general insistence on exploring the relationships between the mode of economic production and social property relations with state/political forms, social and political struggles, and consciousness, with a particular focus on capitalism and classes

in capitalist societies. His was a firmly historical sociology, which emphasized human agency and historical contingency as being conditioned and constrained by structural features and processes, and consequently critically analytical of the character and dynamics of capitalism. Broadly, Marx argued that the division of societies into social classes is based on objective relations of domination and that the dominant class' exploitation of dominated classes is central to many historically specific forms of economic production. In such situations, the state or central political power and authority generally maintains and reproduces the basic social systems, which are riven by the potentially explosive or revolutionary social antagonisms rooted in these relations of domination and exploitation. Despite varying and often contradictory interpretations of his extensive body of work, Marx's framework is often seen as the primary source of political economy and class approaches.

Like Marx himself, several later Marxists had wide-ranging influence both in and outside of academia. Vladimir Lenin developed theories on revolutionary politics and change, imperialism, the state, and liberal democracy. Leon Trotsky's work on the Russian Revolution is a milestone in scholarship on revolution and historical sociology, prioritizing mass political participation and the study of the political processes of the masses themselves in relation to political leaders and parties. He emphasized the historical unfolding of revolutionary struggles and the necessity of analyzing the structural conditioning, if not constitution, of such intense sociopolitical struggles in forms of economic development (hence the theory of combined and uneven development). Antonio Gramsci offered influential theories on hegemony, political conflict, and the distinction and relation between political society and civil society in capitalist democracies. Central to his work was the notion that bourgeois cultural values and institutions, (both those centered in the state and those that were not), help create situations of consented coercion, with continuous state formation acting to balance the interests of the 'fundamental group and those of the subordinate groups'. These analyses demonstrate that political sociology includes politically-oriented analyses, and also that important historical actors are included as political sociologists.

Weber and Political Authority, Bureaucracy, and the Modern State

Max Weber wrote many of his works in response to Marx, with considerable overlap and agreement as well as divergence. Also largely historical, Weber traced the "elective affinity" of capitalist development and Protestantism in Europe, highlighting their functional compatibility and the eventual coconstitutional development of economic forms with culture and ideology. Additionally, Weber's work examined ideal typical forms of legitimate authority and domination, bases of individual and group social action, and perhaps most important, he offered a general theory of rationalization (i.e., instrumental rationality) as embodied in bureaucratic organizational forms and the wide-ranging bureaucratization of modern life – the latter process tied into the dynamic development of capitalism and the modern state. In contrast to Marx, who saw the modern state more or less in relation to powerful class interests, class struggles, and capital accumulation processes, Weber emphasized the modern state as an autonomous source of interests and power. Cultural scholars at times refer to Weber's wide-ranging body of work and theory, yet his work is more often than not associated with and taken as a basis for statist approaches.

Many theorists have been influenced by Weber, including those who furthered the study of elites and bureaucracy. Vilfredo Pareto and Gaetano Mosca were pioneers in the sociological study of elites. Pareto identified the governing elite, who ruled with a psychological preference for either force or fraud. He argued that the elite class is maintained by the circulation of especially apt nonelites into the ruling group. Mosca takes a more sociological approach to the study of elites by emphasizing structural and organizational factors in the maintenance of elite classes, such as their superior organization and their control of resources.

Robert Michels' seminal work criticized party bureaucracies in modern states and addressed central themes such as the role of elites, leftist political movements, bureaucratization, and the gap between democratic theory and practice. Michels developed a thesis called "the iron law of oligarchy", which argues that organization necessarily leads to oligarchy. In Michel's words: "it is organization which gives birth to the dominion of the elected over the electors, of the mandataries over the mandators, of the delegates over the delegators. Who says 'organization', says 'oligarchy'" (1911/1968).

The Durkheimian Tradition and Pluralism

While Marx and Weber were clearly social and political theorists, Emile Durkheim is less often noted as a scholar of politics. However, his concern with social order, integration, and solidarity had clear implications for analyses of politics and society (Allardt, 2001). Durkheim attempted to overhaul the traditional concept of an increasing division of labor and social cohesion, in his insistence that modern societies can achieve a degree of cohesion surpassing their predecessors. He is often associated with those who emphasize culture, values and norms, and functionalism, with a linkage to later pluralist theorists that would come to dominate American social science.

The English translations of Alexis de Tocqueville's *Democracy in America* and *The Old Regime and the Revolution* are standards in political sociology. In the former, he specified the conditions for pluralism and institutional democracy, arguing that in industrially developed societies, social cleavages, and group/strata competition are necessary for consensus. Plural democracy is created and sustained by voluntary associations and active, relatively autonomous local communities. Thus, de Tocqueville emphasized the role of social conflicts and cleavages in developing consensus in democracies (Allardt, 2001).

The Institutionalization of Political Sociology after World War II

In the postwar period, the prerequisites of democracy and the role of parliamentary elections as the major mechanism for establishing and securing democratic rule became important foci. After the war there were great hopes attached to the

possibility of building a better world with the aid of social science and research. As political science became more preoccupied with constitutional problems and modes of state management and sociology focused on social structure and social group behavior, political sociology moved between the disciplines (Allardt, 2001).

Political sociologists undertook large-scale studies of voting behavior and developed fruitful theories and hypotheses, such as Lazarsfeld's argument that cross-pressures lead to political passivity (Lazarsfeld et al., 1944; Allardt, 2001). Seymour Martin Lipset had already published important works, and his *Political Man: The Social Bases of Politics* (1960) became a leading text of the period and helped give birth to institutional political sociology. Lipset's and Rokkan's *Party Systems and Voter Alignments: Cross-National Perspectives* (1967) would later apply Parson's A-G-I-L scheme (an analytic model designed to explain stability by considering Adaption, Goal Attainment, Integration, and Latency) in an attempt to combine structural and institutional approaches to understand the social bases of politics and the effects of politics on social structures (Allardt, 2001).

It was in this context that the pluralist conceptualizations of social and political action arose in earnest. In this vein, Smelser emphasized the state, collective behavior, and cultural-value change (1967); Easton focused on the political system (1965); and Gurr on political violence and relative deprivation (1970). Broadly, these studies tended to equate the political system with the state and considered individuals with distinct preferences and values to be the constitutive units of both organizations and societies (Alford and Friedland, 1985: p. 35).

While the study of electoral behavior and voting patterns became a specialized area of inquiry, political sociologists' basic interest was in studying the conditions for democracy, regime type and breakdowns, state and nation building, modernization, and other processes of social and political change (Allardt, 2001).

Political sociologists, for example, have explored many aspects of democracy: the conditions for democracy in labor unions (Lipset et al., 1956), in the local community (Robert Dahl, 1961) and the conditions of conflict regulation in industrial society (Dahrendorf, 1959). Others focused on the breakdown of democracy and the rise and appeal of Fascism and Communism, such as Raymond Aron in *The Opium of Intellectuals* (1955) and William Kornhauser in *The Politics of Mass Society* (1959). C. Wright Mills *The Power Elite* (1990) examined the relationships and shared interests among US corporate, political, and military leaders; leading to the concentration of economic power and cultural influence in the hands of the relatively few and interchangeability of positions within these three institutions.

The Mid- and Late Twentieth-Century Period

With the maturity and incorporation of electoral studies and studies of some other original political sociology topics in the 1960s and 1970s along with new critiques from the left, these topics disappeared from political sociology's central agenda. Increasingly, the focus became large-scale patterns of societal change. S.N. Eisenstadt (1963/1993) examined the formation and fall of empires, Rokkan studied the historical formation of European centers and peripheries (Rokkan and Urwin, 1983) and Juan Linz analyzed the breakdown of democratic regimes (Linz and Stepan, 1978) (Allardt, 2001).

Yet the period was perhaps most substantially marked by the outpouring of Marxist or Marxist-oriented scholarship on class politics and major sociopolitical transformations as well as the structural contours and dynamics of global capitalism (Allardt, 2001). Unfortunately a full review of this literature is not possible, here, and thus several important examples must suffice. Barrington Moore wrote an often-quoted major work *The Social Origins of Dictatorship and Democracy* (1966), using in-depth historical case studies to show how the varying class coalitions that emerged during agricultural commercialization impacted the forms of 'political modernization' in the modern world. E.P. Thompson (1963) explored workers' life experiences and cultural outlooks in his historical analysis of the formation and consciousness of the British working-class. Jeffery Paige (1975) explored how specific agrarian organization and class relations encouraged certain forms of political action and revolution. Maurice Zeitlin (1984) and Zeitlin and Ratcliff (1988) analyzed the development of the Chilean state in light of struggles between segments of the dominant class and the interactions between international and domestic forces. G. William Domhoff (1967, 1990) combined class and institutional analysis to build his class dominance theory and explore how the power elite in the US (owners and top managers of major corporations) actually dominate policy-making. Immanuel Wallerstein (1974/2011) focused on international divisions of labor in his study of the origins and structural contours of the world capitalist system. On the borders of political sociology were the critical sociologists in the Frankfurt School of dialectical sociology and the Dependency theorists who pointed out how underdeveloped countries were enmeshed in a world dominated by international capitalism.

Interest in master patterns of change and in state and nation building proliferated in the 1980s and 1990s. Theda Skocpol (1979) helped reintroduce the state as the originator of social outcomes and Charles Tilly (1992) stressed a complicated web of warfare, fiscal policy, bureaucratization, and state-making as forming national developments. Goran Therborn's (1995) studies of European states emphasized that national development is multidimensional and may be reversible.

Developments from this period onward made the borders of political sociology more permeable. For instance studies considering the roles of ethnicity, ethnic groups, and ethnic identities in the formation of nations straddle the line between political sociology and race/ethnic studies. Studies in the political sociology tradition often emphasize the macro-level role that ethnicity and ethnic identities may play in the formation of nations and political development (Rokkan and Urwin, 1983).

The Contemporary Period

Over the last few decades, scholars have expanded political sociology's theoretical and empirical scope. Established approaches have persisted and developed, and have been challenged or supplemented by theories of rational choice (Tsebelis,

1990; Chong, 2000), culture (B. Anderson, 1983/2006; Bourdieu, 1984/2013; Inglehart, 1997), race (Gilroy, 1991; Winant, 1994), gender (Orloff, 1993; Paxton et al., 2007), and several institutionalisms (Amenta and Ramsey, 2010; Thelen, 1999). Oftentimes, these approaches influenced political sociology after migrating from political science, economics, or the humanities. While 'new' theoretical perspectives and explanatory models may initially seem to displace established approaches, their overly bold early claims may recede as scholars implicitly or explicitly adopt theoretical synthesis or accommodation, or bolster their approaches by accounting for or appropriating salient factors or processes emphasized by others. So, while Marxist scholarship which flourished in the 1970s lost much of its early fervor by the mid 1980s, political sociology has remained theoretically eclectic.

Scholars are also more likely to explore historical or contemporary empirical phenomena of diverse geographical settings, a trend bolstered by the growing numbers and diversity of academics and research institutions worldwide, e.g., political sociologists regularly examine social and political processes in cases outside of Europe and the United States. Further, not only do emergent and changing social and political circumstances tend to impact what is being explored (by providing cases) but also salient political trends and/or the configuration of political power holders may impact the explanations given.

Moreover, recent scholarship has benefited from methodological advances. While quantitative methods may predominate in the social sciences, the flourishing of work in comparative and historical methodologies (Abrams, 1982; Ragin, 1987; Mahoney and Rueschemeyer, 2003) has found fertile ground in political sociological research, both through within-case analyses and cross-case comparisons.

In this context, the following section not so much adjudicates theories but instead introduces empirical studies covering a range of research questions asked by scholars from different theoretical traditions. It is not an exhaustive review of recent literature; rather, it provides a brief look at some vital political sociology research areas, such as states and society, revolution, globalization, culture, immigration, macro-social theory, and also notes other substantive themes.

Scholarship on state formation continued in the decades after World War II, exploring processes of the formation of early modern European states, the global diffusion (uneven and varied) of their general political models, variations in regime type, welfare state formation, developmental states, and so on. While states differ along multiple axes, including over time and place, it is commonly held that national states originally emerged in early modern Europe (along with capitalism) and have thus received due diligence to the degree that it was from this temporal and geographical source that they exerted extensive gravitational pull on world history, so that state formation processes in later centuries – often accompanying Western imperialism – were reproduced across the globe as approximations of a single model. More generally, scholars explore modern states' formation with varying emphases on processes of capitalist development, war, international relations, resource extraction, class or elite coalitions, and culture.

Political sociologists emphasize the ways in which social factors impact state formation processes, considering capitalism and class as well as culture and institutions. Seminal Marxist explanations focus on rural classes and capitalist agricultural development to explain the origins of liberal democracies and Communist or Fascist dictatorships in major states (Moore, 1966); or investigate the centralized states of European Absolutism, its regional variation, and its role in the transition from feudalism to capitalism (Anderson, 1974/2013). Tilly (1992) adopts a bellecist approach to explore the formation of national states in an environment of incessant military-strategic competition, and system-wide convergence on that model in Europe. The interplay of war making, state making, and capitalist consolidation, Tilly argues, proceeds in a dialectic of the increasing concentration and accumulation of capital and coercion, pressed by war making and preparation, fiscal extraction, administrative expansion, and the making of social alliances. Ertman (1997) prioritizes local political institutional path dependency and the timing of the onset of sustained geopolitical-military competition, to account for the diverse range of regime types and state infrastructures across Europe in the eighteenth century, whereas Gorski (2003) shifts to culture to unravel the emergence of strong centralized states in Germany and the Netherlands, arguing that Calvinism and its emphasis on public order and discipline heightened state capacities and functions, regarding education, crime and punishment, and military effectiveness.

While extending insights from European cases to other regions, scholars often avoid transplanting some universal European trajectory and note the impact of colonialism, Western geopolitical dominance, and how being 'late' developers may impact state formation in the Third World. Heydemann (2000) notes that interstate war between Middle Eastern and North African states was rare in the twentieth century, but that extensive war preparation and militarized government produced deep political consequences, with variation in state formation linked to discrete modes of resource extraction. Centeno (2003) likewise describes interstate warfare as rare in post-independence Latin America, and argues that this absence corresponds to the weakness of Latin American states and the national elites' unwillingness to ally with them.

Scholars also study the development of welfare states and public social provision, the emergence and character of such systems, and/or their perceived ongoing decline in neoliberal capitalism (see Marshell 1950; Korpi, 1983; Huber and Stephens, 2001; Pierson, 1994). In general, political sociology emphasizes social inequalities along various dimensions and their impact on welfare states. Esping-Andersen (1990) explores the origins and role of welfare states in European capitalist countries, emphasizing the impact of working class power and varying class coalitions on the character of welfare states and regimes of public social provision. Fox (2012) explores the American social welfare system and how different racial and immigrant groups receive differential access to social welfare programs, whereas Misra (1998, 2003) analyzes family allowance policies, and the role of women's movements and the perceived value of women's paid and unpaid labor for variations in public social provision. Venturing to the global South, Sandbrook et al. (2007) look at social democratic movements and democratic developmental states, and how some peripheral states have balanced the achievement of economic growth through

globalized markets with relatively progressive social and political policies.

In a world of stark international inequalities, studies of development attempt to unravel the social and political processes behind prosperous economies with a focus on class relations, the state, and economic institutions. These works often center on the concerted efforts of post–World War II 'developmental' states, traversing state building, economic development, and postcolonialism. A major thrust in this literature explores conditions in which states and state leaders gain the wherewithal to ignite and sustain industrialization. One broad statist explanation focuses on state capacity and autonomy to guide and discipline capitalists (and labor), thus bolstering state planning agencies to drive developmental policies (Amsden, 1989; Evans, 1995; Wade, 1990). Some challenge, or refine, this state autonomy/capacity thesis. Chibber (2003) brings back in structural political economy to consider class alliances and interests, state capacity and autonomy, and international economic opportunity to explain situations where local capitalists gain from, and thus support, the strengthening autonomous state economic planning agencies (as in South Korea) and, conversely, (in cases where capitalists would suffer from, and thus stifle these institutions), robust development agencies and planning (as in India). His study illuminates crucial class and state processes impacting relative success versus relative failure, and deepens the understanding of capitalist classes in developmental states. Finally, Mahoney (2010) studies the impact of colonial arrangements on postcolonial development, and how institutional fit (or non-fit) between colonizer and colonized greatly influenced longer-term economic development.

Recent scholarship on revolution has shifted focus from the 'great' social revolutions (Moore, 1966; Skocpol, 1979) to political revolutions and revolutionary movements, and their varying forms and outcomes, in the post–World War II world. Political sociologists emphasize social compositional factors, regime types, and socioeconomic structure. Parsa (1989) offers a sophisticated articulation of the multi-class coalition thesis, arguing that the 'success' of the Iranian revolution was rooted in multiple social classes mobilizing in a common oppositional front outside the formal power structure of the Shah's autocratic regime. Accepting the import of class struggle, Goodwin (2001) refines a statist approach in his study of revolutions in Latin America, SE Asia, and East Europe, arguing that particular neopatrimonial state-regime types may not only incubate and encourage revolutionary movements but also impact the success of such movements in engendering political revolution. In a cultural turn, Slater (2009) holds that the political posture and emotive appeals of communal elites – those with religious and cultural authority – determines the emergence and outcome of revolutionary mobilization. Finally, Achcar (2013) extends and modifies Marx's theory of social revolution to analyze recent uprisings in the Arab world, arguing that it is the specific modality of an economic mode of production, especially in the political and legal structures that block social and economic development, rather than the generic economic mode itself, that may provide crucial causes for popular uprisings and thus become the object of political revolution.

Globalization naturally attracts considerable scholarly attention, providing numerous areas of inquiry. Political sociologists tend to focus on power relations in global capitalism and the consequences for development, the political institutions that undergird attendant policies, as well as salient spatial and organizational transformations. Dependency scholars Cardoso and Faletto (1979) critique Modernization theory's universal linear stage model of development, and prioritize asymmetrical and partially constitutive international power relations and their affects on national political and economic development. The world-systems tradition has long analyzed national and regional development in the light of the character and dynamic of global capitalism (Wallerstein, 1974/2011) and imperial hegemonies (Arrighi, 1994); yet many drop a level of analysis to examine crucial shorter-term political processes linked to a globalizing world. Babb (2009) explores institutional linkages between the American state (especially the legislative branch) and multilateral development banks, and Washington's political sway over them. Others explore China's rise and how it has helped reshape the structure and dynamics of global capitalism, as well as its geopolitical impact on China's close neighbors (Hung, 2009), global cities (Sassen, 2001), international organizations (Boli and Thomas, 2003), and so on.

Relatedly, neoliberal globalization, many argue, fundamentally includes financialization and attendant transformations that diminish the quality of public social provision, and political sociologists have analyzed multiple facets of these phenomena. Tabb (2012) investigates the restructuring of capitalism from the 1970s onward, specifically with regard to banking practices, financial motives, and regulatory policies. Krippner (2011) studies the historical development of the US financial market and the creation of market policies conducive to financialization, arguing that these were not policymakers' deliberate goals but the inadvertent results of attempts to solve pressing economic problems. Prechel and Morris (2010) explore the causes of financial malfeasance from 1995 to 2004 and how certain social structures create dependencies, incentives, and opportunities to engage in such behavior, i.e., neoliberal policies permit such activities, and largely result from well-financed and systematic corporate political strategy.

Immigration and citizenship in the globalizing world continues to capture scholars' attention, with the migration of people across countries clearly linked to potentially salient social, political, and economic transformations. Portes and Rumbaut (2006) provide a prominent overview of immigrant social and political dynamics in the US, honing in on micro- and macro-level processes impacting immigrants' individual lives as well as the wider social and political formations to which they belong. Bloemraad (2006) queries how countries that allow the entrance of immigrants and refugees can foster civic cohesion and political community, and tempers the widespread emphasis on features of immigrant communities by considering the impact of state policy for outcomes of citizenship acquisition and political participation. Furthermore, the global salience of 'market fundamentalism', argues Somers (2008), subjugates notions and practices of citizenship to market logic and yields deleterious results for people's lives.

Scholars continue to investigate social bases of politics and political behavior, which has been a hallmark of political sociology since at least the 1950s. Domhoff (1990) and Domhoff and Webber (2011) refine the class dominance theory to look at social networks of power and the role of capitalist class segments on the origins of major social policies; while Manza and Brooks (1999) consider more broadly the role social group cleavages have on electoral politics. Clawson et al. (1998) investigate the role of financial contributions in policy formation, and the relationships that may develop between politicians and private contributors. Scholars in this general line of inquiry also explore issues surrounding political parties (Panebianco, 1988; Shefter, 1994), public opinion (Dalton, 2008), and interest groups and other political organizations (Wilson, 1995).

The concept of culture is not only disputed in terms of its specific meaning(s), but it is also broad, multifaceted, and at times muddled, with cultural studies of politics employing the concept in wide-ranging and varying ways. Using a dynamic theoretical framework, Campbell (1998) explores how ideas – paradigms, public sentiments, programs, and frames – may influence major policy-making innovations. For Ermakoff (2008), subjective orientations and patterns of interactions impact major political change and the surrender of democratic to nondemocratic authority, while Ikegami (2005) details the interconnections of arts and esthetics with state–society relations in a case of state formation where esthetic socialization compensated for state policies that fostered extreme social fragmentation. Others explore folk music's unifying and galvanizing power, and the relation between cultural forms and social or labor movement activities (Roscigno and Danaher, 2004). In a different vein, important studies of the media examine the mass media in multiple ways, seeing a powerful and effective ideological institution that props the political status quo (Chomsky and Herman, 1988/2008) – society's "master arena" where individual or collective political actors engage in contests over social meanings (Gamson and Andre, 2004), the interactions of media discourse and public opinion formation (Gamson and Modigliani, 1989), and the news media as a political institution (Schudson, 2002). Also, in an ambitious macrophenomelogical approach and in macro-social theory, the world societal perspective prioritizes global culture in the production and spread of ideas, policies, and institutions (Meyer et al., 1997).

Other than world-systems scholars (Wallerstein and Arrighi), Michael Mann (1986, 1993, 2012, 2013) is one of the few contemporary political sociologists who attempts to rigorously identify the principal social-structural trends across history and to explain the development and expansion of fundamental power structures. Mann proposes the IEMP model, which holds that human societies form around four power relations – ideological, economic, military, and political – that are intertwined although none is purely reducible to another. For Mann, contemporary globalization is the "plural extension" of these four relations, with the primary modern power organizations (capitalism, nation-states, and empires) congealing around them. This globalization thus consists of three main institutional processes, the globalization of capitalism, the nation-state, and empires (eventually just the American empire). These processes, however, crystalize in various, often-competing forms and have thus been geographically and institutionally polymorphous.

Concluding Remarks

Certainly there are remarkable areas of social and political inquiry not included in the above survey, and even those discussed are only briefly introduced. There are veritable literatures on the political processes of empire and imperialism (Harvey, 2003; Wood, 2005), nationalism (Breuilly, 1994; Hobsbawm, 1990), labor and labor movements (Stepan-Norris and Zeitlin, 2003; Fantasia, 1989), ethnicity and ethnic conflict (Varshney, 2003; Gutiérrez, 1995), and so on. Nonetheless, there are several clear features or trends that manifest in this survey. Political sociology literature continues to emphasize civil society and potentially vital socioeconomic factors, in contrast to much of political science. With the aid of methodological advances, political sociologists continue to develop various theoretical frameworks and to expand the substantive breadth and depth of this border field. Scholars enter empirical cases from differing levels of analysis to address an inherently wide set of questions. With these diverse research approaches, political sociologists continue to cultivate useful knowledge through rigorous empirical research on social and political change as well as the basic problems afflicting societies, and thus remain relevant to policy-oriented debates. Echoing the previous entry from the first edition, political sociology's main general problem areas continue to revolve around the origins, character, and practices of social and political power and conflict; dominant and emerging cleavages; patterns of social, political, economic, and cultural change; the formation of states and nations as well as the breakdown of social and political orders; the conditions of different regime-types (e.g., democracy or authoritarianism); capitalist development; and multifaceted international relationships. For better or worse, these are vital areas of inquiry that will continue to provide attendant puzzles and problems to be scrutinized by political sociologists.

See also: Attitudes, Political: Authoritarianism and Tolerance; Bureaucracy, Sociology of; Capitalism: Global; Capitalism; Class: Social; Dependency Theory; Historical Sociology; Imperialism; Marxism in Contemporary Sociology; Movements, Social; Power in Society; Power; Revolutions, Sociology of; State, Sociology of The; Working Classes, History of; World Systems Theory.

Bibliography

Abrams, Philip, 1982. Historical Sociology. Cornell University Press, Ithaca.

Achcar, Gilbert, 2013. The People Want: A Radical Exploration of the Arab Uprising. University of California Press, Los Angeles and Berkeley.

Alford, Robert R., Friedland, Roger, 1985. Powers of Theory: Capitalism, the State, and Democracy. Cambridge University Press, Cambridge.

Allardt, Erik, 2001. Political Sociology. In: Neil Smelser, J., Paul Baltes, B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences, vol. 11. Elsevier, New York and Amsterdam, pp. 11701–11706.

Amenta, Edwin, Ramsey, Kelly, 2010. "Institutional Theory" Handbook of Politics. Springer, New York, pp. 15–39.
Amenta, Edwin, Nash, Kate, Scott, Alan (Eds.), 2012. The Wiley-Blackwell Companion to Political Sociology. Wiley-Blackwell, Oxford.
Amsden, Alice, 1989. Asia's Next Giant: South Korea and Late Industrialization. Oxford University Press, Oxford.
Anderson, Benedict, 1983/2006. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, New York.
Anderson, Perry, 1974/2013. Lineages of the Absolutist State. Verso, New York.
Aron, Raymond, 2009. The Opium of the Intellectuals. Transaction Books, New York.
Arrighi, Giovanni, 1994. The Long Twentieth Century: Money, Power, and the Origins of Our Times. Verso, New York.
Babb, Sarah, 2009. Behind the Development Banks: Washington Politics, World Poverty, and the Wealth of Nations. University of Chicago Press, Chicago.
Bloemraad, Irene, 2006. Becoming a Citizen: Incorporating Immigrants and Refugees in the United States and Canada. University of California Press, Los Angeles and Berkeley.
Boli, John, Thomas, George M., 2003. Constructing World Culture: International Nongovernmental Organizations since 1875. Stanford University Press, Palo Alto.
Bourdieu, Pierre, 1984/2013. Distinction: A Social Critique of the Judgement of Taste. Routledge, New York.
Breuilly, John, 1994. Nationalism and the State. University of Chicago Press, Chicago.
Campbell, John L., 1998. Institutional analysis and the role of ideas in political economy. Theory and Society 27 (3), 377–409.
Cardoso, Fernando, Faletto, Enzo, 1979. Dependency and Development in Latin America. University of California Press, Los Angeles and Berkeley.
Centeno, Miguel A., 2003. Blood and Debt: War and the Nation-State in Latin America. Pennsylvania State University Press, University Park.
Chibber, Vivek, 2003. Locked in Place: State-Building and Capitalist Industrialization in India, 1940–1970. Princeton University Press, New Haven.
Chomsky, Noam, Herman, Edward S., 1988/2008. Manufacturing Consent: The Political Economy of the Mass Media. Random House, USA.
Chong, Dennis, 2000. Rational Lives: Norms and Values in Politics and Society. University of Chicago Press, Chicago.
Clawson, Dan, Neustadtl, Alan, Weller, Mark, 1998. Dollars and Votes: How Business Campaign Contributions Subvert Democracy. Temple University Press, Philadelphia.
Dahl, Robert A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven.
Dahrendorf, Ralf, 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Palo Alto.
Dalton, Russell J., 2008. Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies. Sage, New York.
Domhoff, G., Webber, Michael J., 2011. Class and Power in the New Deal: Corporate Moderates, Southern Democrats, and the Liberal-Labor Coalition. Stanford University Press, Palo Alto.
Domhoff, William G., 1967. Who Rules America? Prentice Hall, Englewood Cliffs, NJ.
Domhoff, William G., 1990. The Power Elite and the State: How Policy is Made in America. Aldine de Gruyter, Hawthorne, NY.
Easton, David, 1965. A Systems Analysis of Political Life. Wiley, New York.
Eisenstadt, Shmuel N., 1963/1993. The Political Systems of Empires. Transaction, New York.
Ermakoff, Ivan, 2008. Ruling Oneself Out: A Theory of Collective Abdications. Duke University Press, North Carolina.
Ertman, Thomas, 1997. Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe. Cambridge University Press, Cambridge.
Esping-Andersen, Gøsta, 1990. The Three Worlds of Welfare Capitalism. Polity Press, Cambridge.
Evans, Peter, 1995. Embedded Autonomy: States and Industrial Transformation. Princeton University Press, Princeton, New Jersey.
Fantasia, Rick, 1989. Cultures of Solidarity: Consciousness, Action, and Contemporary American Workers. University of California Press, Los Angeles and Berkeley.
Fox, Cybelle, 2012. Three Worlds of Relief: Race, Immigration, and the American Welfare State from the Progressive Era to the New Deal. Princeton University Press, New Jersey.
Gamson, William A., Andre, Modigliani, 2004. Bystanders, public opinion, and the media. In: Snow, David A., Soule, Sarah A., Hanspeter, Kriesi (Eds.), The Blackwell Companion to Social Movements. Blackwell Publishers, Malden, pp. 242–261.
Gamson, William A., Modigliani, Andre, 1989. Media discourse and public opinion on nuclear power: a constructionist approach. American Journal of Sociology 95 (1), 1–37.
Gilroy, Paul, 1991. 'There Ain't No Black in the Union Jack': The Cultural Politics of Race and Nation. University of Chicago Press, Chicago.
Goodwin, Jeff, 2001. No Other Way Out: States and Revolutionary Movements, 1945–1991. Cambridge University Press, Cambridge.
Gorski, Philip S., 2003. The Disciplinary Revolution: Calvinism and the Rise of the State in Early Modern Europe. University of Chicago Press, Chicago.
Gurr, Ted, 1970. Why Men Rebel. Paradigm, New York.
Gutiérrez, David G., 1995. Walls and Mirrors: Mexican Americans, Mexican Immigrants, and the Politics of Ethnicity. University of California Press, Los Angeles and Berkeley.
Harvey, David, 2003. The New Imperialism. Oxford University Press, Oxford.
Heydemann, Steven (Ed.), 2000. War, Institutions, and Social Change in the Middle East. University of California Press, Los Angeles and Berkeley.
Hobsbawm, Eric, 1990. Nations and Nationalism since 1780: Programme, Myth, Reality. Cambridge University Press, Cambridge.
Huber, Evelyne, Stephens, John D., 2001. Development and Crisis of the Welfare State: Parties and Policies in Global Markets. University of Chicago Press, Chicago.
Hung, Ho-fung, 2009. China and the Transformation of Global Capitalism. Johns Hopkins University, Baltimore.
Ikegami, Eiko, 2005. Bonds of Civility: Aesthetic Networks and the Political Origins of Japanese Culture. Cambridge University Press, Cambridge.
Inglehart, Ronald, 1997. Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies. Princeton University Press, New Jersey.
Janoski, Thomas, Alford, Robert R., Hicks, Alexander M., Schwartz, Mildred A. (Eds.), 2005. The Handbook of Political Sociology. Cambridge University Press, Cambridge.
Kornhauser, William, 1959/2013. Politics of Mass Society. Routledge, New York.
Korpi, Walter, 1983. The Democratic Class Struggle. Routledge & Kegan Paul, London.
Krippner, Greta, 2011. Capitalizing on Crisis: The Political Origins of the Rise of Finance. Harvard University Press, Cambridge.
Lazarsfeld, Paul F., Berelson, Bernard, Gaudet, Hazel, 1944. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. Columbia University Press, New York.
Linz, Juan J., Stepan, Alfred C. (Eds.), 1978. The Breakdown of Democratic Regimes, Latin America. Johns Hopkins University Press, DC.
Lipset, Seymour M., 1960. Political Man: The Social Bases of Politics. The Johns Hopkins University Press, Washington, DC.
Lipset, Seymour M., Rokkan, Stein (Eds.), 1967. Party Systems and Voter Alignments: Cross-National Perspectives. Free Press, New York.
Lipset, Seymour M., Trow, Martin A., Coleman, James S., 1956. Union Democracy: The Internal Politics of the International Typographical Union. Free Press, Glencoe.
Mahoney, James, 2010. Colonialism and Postcolonial Development. Cambridge University Press, (Cambridge), New York.
Mahoney, James, Rueschemeyer, Dietrich (Eds.), 2003. Comparative Historical Analysis in the Social Sciences. Cambridge University Press, Cambridge.
Mann, Michael, 1986, 1993, 2012, 2013. The Sources of Social Power, vols. 1–4. Cambridge University Press, Cambridge.
Manza, Jeff, Brooks, Clem, 1999. Social Cleavages and Political Change: Voter Alignments and US Party Coalitions. Oxford University Press, Oxford.
Marshall, Thomas H., 1950. Citizenship and Social Class: And Other Essays. Cambridge University Press, Cambridge.
Meyer, John W., Boli, John, Thomas, George M., Ramirez, Francisco O., 1997. World society and the nation-state. American Journal of Sociology 103 (1), 144–181.
Michels, Robert, 1911/1968. Political Parties. Taylor & Francis, New York.
Mills, Wright C., 1956/1999. The Power Elite. Oxford University Press, Oxford.
Misra, Joya, 1998. Mothers or workers? The value of women's labor: women and the emergence of family allowance policy. Gender & Society 12 (4), 376–399.
Misra, Joya, 2003. Women as agents in welfare state development: a cross-national analysis of family allowance adoption. Socio-economic Review 1 (2), 185–214.
Moore, Barrington, 1966. Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World. Beacon Press, New York.
Orloff, Ann S., 1993. Gender and the social rights of citizenship: the comparative analysis of gender relations and welfare states. American Sociological Review 58 (3), 303–328.
Paige, Jeffrey M., 1975. Agrarian Revolution. Free Press, New York.

Panebianco, Angelo, 1988. Political Parties: Organization and Power. Cambridge University Press, Cambridge.
Parsa, Misagh, 1989. Social Origins of the Iranian Revolution. Rutgers University Press, New Jersey.
Paxton, Pamela, Kunovich, Sheri, Hughes, Melanie M., 2007. Gender in politics. Annual Review of Sociology 33, 263–284.
Pierson, Paul, 1994. Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment. Cambridge University Press, Cambridge.
Portes, Alejandro, Rumbaut, Ruben G., 2006. Immigrant America: A Portrait. University of California Press, Los Angeles and Berkeley.
Prechel, Harland, Morris, Theresa, 2010. The effects of organizational and political embeddedness on financial malfeasance in the largest US corporations. American Sociological Review 75 (3), 331–354.
Ragin, Charles, 1987. The Comparative Method: Moving beyond Qualitative and Quantitative Strategies. University of California Press, Los Angeles and Berkeley.
Rokkan, Stein, Urwin, Derek W., 1983. Economy, Territory, Identity: Politics of West European Peripheries. Sage, London.
Roscigno, Vincent J., Danaher, William F., 2004. The Voice of Southern Labor: Radio Music and Textile Strikes, 1929–1934. University of Minnesota Press, Saint Paul.
Sandbrook, Richard, Edelman, Marc, Heller, Patrick, Teichman, Judith, 2007. Social Democracy in the Global Periphery: Origins, Challenges. Cambridge University Press, Cambridge.
Sassen, Saskia, 2001. The Global City: New York, London, Tokyo. Princeton University Press, New Jersey.
Schudson, Michael, 2002. The news media as political institutions. Annual Review of Political Science 5 (1), 249–269.
Shefter, Martin, 1994. Political Parties and the State: The American Historical Experience. Princeton University Press, New Jersey.
Skocpol, Theda, 1979. States and Social Revolutions. Cambridge University Press, Cambridge.
Slater, Dan, 2009. Revolutions, crackdowns, and quiescence: Communal elites and democratic mobilization in southeast asia. American Journal of Sociology 115 (1), 203–254.
Smelser, Neil J., 1967. Theory of Collective Behavior. Taylor & Francis, New York.
Somers, Margaret R., 2008. Genealogies of Citizenship: Markets, Statelessness, and the Right to Have Rights. Cambridge University Press, Cambridge.
Stepan-Norris, Judith, Zeitlin, Maurice, 2003. Left Out: Reds and America's Industrial Unions. Cambridge University Press, Cambridge.
Tabb, William, 2012. The Restructuring of Capitalism in Our Time. Columbia University Press, New York.
Thelen, Kathleen, 1999. Historical institutionalism in comparative politics. Annual Review of Political Science 2 (1), 369–404.
Therborn, Goran, 1995. European Modernity and beyond. The Trajectory of European Societies 1945–2000. Sage, London.
Thompson, Edward P., 1963. The Making of the English Working Class. IICA, London.
Tilly, Charles, 1992. Coercion, Capital, and European States, AD 990–1992. Blackwell, Oxford.
Tsebelis, George, 1990. Nested Games: Rational Choice in Comparative Politics, vol. 18. University of California Press, Los Angeles and Berkeley.
Varshney, Ashutosh, 2003. Ethnic Conflict and Civic Life: Hindus and Muslims in India. Yale University Press, New Haven.
Wade, Robert, 1990. Governing the Market: Economic Theory and the Role of Government in East Asian Industrialization. Princeton University Press, New Jersey.
Wallerstein, Immanuel, 1974/2011. The Modern World-System: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century I. University of California Press, Los Angeles and Berkeley.
Wilson, James Q., 1995. Political Organizations. Princeton University Press, New Jersey.
Winant, Howard, 1994. Racial Conditions: Politics, Theory, Comparisons. University of Minnesota Press, Minneapolis.
Wood, Ellen M., 2005. Empire of Capital. Verso, New York.
Zeitlin, Maurice, 1984. The Civil Wars in Chile, or, the Bourgeois Revolutions that Never Were. Princeton University Press, Princeton, New Jersey.
Zeitlin, Maurice, Ratcliff, Richard E., 1988. Landlords & Capitalists: The Dominant Class of Chile. Princeton University Press, New Jersey.

Political Theory, History of

Bjorn Wittrock, Swedish Collegium for Advanced Study in the Social Sciences, Uppsala, Sweden

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11706–11712, © 2001, Elsevier Ltd.

Abstract

Historically, the focus of political thought has been on a triad of related problems, namely those of constituting a community bound by a normative order, of establishing the legitmacy and terms of enforcement of such an order, and of articulating criteria for allocations of resources and entitlements. In early modern Europe political thought was largely concerned with the nature and legitimacy of new forms of normatively binding orders and their terms of enforcement. The concern with the constitution of political community became the central one in political thought from Hobbes and Pufendorf over Locke all the way to Rousseau. The intellectual transformation of the late eighteenth and early nineteenth centuries entailed a shift in the thinking about human beings, about society, and about the very nature of political order. The end of World War I and the interwar years marked the triumph and crisis of liberal political thought and the emergence of a range of forms of political thought that transcended liberal traditions. The history of political thought towards the end of the twentieth century has been characterized by efforts to reconceptualize the nature human agency, whether in a rationalistic-deliberative form or in a linguistic-interpretative form. The first stream of research has led to formulations of the nature of political community in contractual terms. It has also sought to establish normative criteria for the proper allocation of resources. The second stream has resulted in programs for the writing of conceptual and contextual intellectual histories. It has also involved efforts to position political thought in a wider global context.

Reflection about political phenomena in a wide sense of the term is as old as human community itself. It focuses on a triad of related problems that might be broadly delimited as follows. First, there is the problem of constituting a community bound together by a normative order, secondly that of establishing the legitimacy and the terms of enforcement of such an order, and, thirdly, that of finding criteria for a proper and just allocation of resources as well as of entitlements. Jointly these three sets of problems entail that political theory and the history of political theory, as distinct from, say, a theory of scientific validation or a theory of argumentation, will be characterized by an inherent duality. Thus it will not only, as these other types of theory, be an account of rules of contestation and signification. It will also have to involve an account of which such rules are and which are not upheld by a normativity that is, in the last instance, enforceable by violent means. Put in everyday language, political theory will in its core be a theory both about meaning and communication and a theory about power, enforcement, and, ultimately, violence. It is easy to see that different political theorists have come to emphasize different strands in this duality, but the duality itself is irreducible and constitutive of political thought.

Political Thought in Historical Context

Thought about such matters is found in clearly articulated form from the middle of the first millennium BCE onwards in all the major civilizations of Afro-Eurasia. Thus it is manifested in written texts in the sixth century by Confucius and two centuries later by Mencius, in China and the writings during the same period in India, as well as in the prophetical age of Jewish religion and during the era of classical philosophy in Greece. The emergence of imperial-like political orders across Afro-Eurasia in the last centuries of the first millennium BCE also gave rise to a range of texts about the nature of political order. This is equally valid for texts about the Roman and Byzantine empires in the Mediterranean region to texts highlighting features of rulership in Parthian and later also in Sassanian Persia, in Maurya India, and in the Han Empire of China.

The recording of political thought in the form of the writing of histories of such thought is however a much more recent phenomenon. At its focus it is the relationship of community-wide rules to forms of enforcement and an effort to arrive at an understanding of the historical context in which particular types of such relationships occur. The emergence of clearly articulated forms of historical thought is largely coterminous with the emergence of history and social science as forms of scholarly activity performed within reformed or newly created universities and higher education institutions from the early nineteenth century onwards. In order to understand the key features and characteristics of the writing of histories of political thought it is first of all necessary to bring out key conditions of political thought in the region in which such an early disciplinary condition first occurred, namely in Europe.

Political Thought and Early Modernities

In a broad sense political thought in early modern Europe was a way to understand the nature of ruling in human communities and to grasp the conditions of normative argumentation and enforcement in such communities. The need for new understandings of these types of phenomena was itself directly related to the gradual decline and eventual demise of the two most prominent institutional forms of medieval Europe. Those were, first, the cultural world of Western Christendom,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03043-9

epitomized first by the all-embracing Christian Church and second by its secular counterpart, namely the pervasive, but elusive, political order of the Holy Roman Empire. By the twelfth and thirteenth centuries, however, city republican rule, particularly in the northern half of Italy, had emerged as a new form of political order. Originally it was often modeled on trading communities and was created out of a sense of the necessity of mutual support and protection in areas exposed to incursions and with little or no protection provided by the nominal Imperial rulers.

Similarly, incipient national monarchies were starting to assert their authority vis-à-vis both pope and emperor. In this secular process the French king played a pioneering role. In particular, he had earlier than most other kings and rulers demonstrated his ability to strengthen his own position and financial basis, e.g., by confiscating lands and goods of the order of the Templars. This example was to be followed by kings across Europe in their confiscating the lands and riches of monasteries and religious orders in centuries to come.

In this sense, the Reformation and the Counterreformation meant that in Protestant and Catholic countries alike, national monarchies, normally of a composite nature (cf. Elliott, 1992), were created or reformed. This occurred on the basis of the previous dominant ideals of church, empire, and city republics, but in a form that synthesized parts that had previously been separated. Furthermore, the new national and composite monarchies almost invariably involved efforts on the part of rulers to appropriate lands and resources that had previously been controlled by the church or religious orders. Thereby they were also able to secure a fiscal basis for the new forms of political ruling, if often at a horrendous price of protracted religious civil wars and cleansings.

Under these conditions the pre-eminent political questions became that of grasping the nature and legitimacy of a normatively binding communal order and to outline the range of legitimate means for upholding and enforcing that order. Many of the early modern classical political thinkers, from Machiavelli to Jean Bodin, provide answers to these very questions. They have, however, significantly less to say about the two other sets of questions discerned above and that were to become prominent in later political thought. These were the questions of the very constitution of a community of human beings and of the principles underlying decisions about the allocations of resources or the granting or withholding of entitlements.

All this was to change with the final breakdown of the old political and cultural order in Europe in the middle of the seventeenth century. In China, even in the embattled transition from the Ming to the Ching dynasty at roughly the same time, the general linguistic and political order was preserved. In Japan a long period of internal warfare and turmoil came to an end with the establishment of one of the most effective forms of absolutist rule, namely the Tokugawa shogunate. Across the Islamic world, the three parallel empires of the Ottomans, the Safavi, and the Mughals, constituted stable political entities. In Europe, however, a new system of states took form and was promulgated in the Westphalian peace treaty of 1648, based on an explicit recognition of religious diversity and monarchical pre-eminence. It was also based on the tacit recognition that an all-encompassing imperial polity was no longer even possible as a viable political ideal and that religious diversity was not to be overcome.

Political Thought as the Constitution of a Political Community

In this unfamiliar context, there was an urgency about finding ways of granting legitimacy and meaning to the new order that emerged in the wake of the Westphalian peace treaty and the provisional end of a century and a half of bloody civil wars in Europe. Thereby a sea change occurred in political thought, namely a deep-seated shift from a concern with the enforcement of normative political order to a concern with the nature and very emergence of political community in the first place.

The concern with the constitution of political community becomes the central one in political thought from Hobbes and Pufendorf to Locke all the way to Rousseau. Within this wide range of answers there is, however, a recurring and predominant concept, namely an inquiry into the conditions for the establishment of a social contract. The existence of such a social contract was normally seen also to entail an answer to the question of which types of enforcement of the normative order constitutive of the political community are legitimate. More specifically only those types of enforcement are sanctioned that, implicitly or explicitly, follow from the terms of the social contract itself.

Political theory from Hobbes onwards is centrally preoccupied with the nature of and the consequences of a transition from an imaginary state of nature to the contractually constituted political community. Ultimately all such forms of theorizing seem to depend on a limited set of assumptions about human beings and the nature of their joining together in societies. One such set of assumptions refer to the assignment of some kind of human rights, i.e., rights that accrue to human beings irrespective of their origins and other accidental traits and characteristics. Furthermore, there is a set of assumptions that refer to the thought-experiment of society being created through a process, normally described in terms of a contract, whereby human beings agree about their association in political form.

Questions about enforcement, separate from an answer to the question of the constitution of a political order, based upon legitimately agreed upon contractual arrangements, were of subordinate and derivative importance relative to this overriding concern. The same is true of questions pertaining to the allocation of resources and the assignment of entitlements to the members and subjects of a political community. Thus such matters have to be dealt with and negotiated in accordance with the terms of those agreed upon in the original contractual arrangements.

Thus European political theory in this classical form emerged in a particular historical context both in European and world history. It also involved forms of theorizing that portrayed an imaginary contractual past as the point of constitution of polities. However, it was not in itself characterized by an effort to reflect on the history of political theory as a key component of developing such a theory. This form of reflexive consciousness was to become a key characteristic of yet another major transformation in intellectual and political history.

The Formation of Modernity and the Transition from Political Philosophy to Social Science and History

The late eighteenth and early nineteenth centuries are often described in terms of the coterminous occurrence of major transformations of both technological and economic practices and of political order. Often these transformations are captured with terms such as the industrial revolution and the political transformations inherent in the French and American revolutions. However, in this period there also occurs a major restructuring of ideas about nature and society and, more broadly, in the forms of knowledge and beliefs about the world. This intellectual transformation had deep repercussions for the emergence of new forms of political thought and reflection about the historical context of such forms of thought and for the decline of other forms of political thought.

In this period broad genres of thought about the world gradually gave way to more differentiated forms of knowledge. Slowly natural philosophy and political philosophy came to be replaced by a variety of different natural and social sciences. At the core of this transformation were epistemic shifts where ideas about biological and historical processes, about language and interpretation, about human beings and their capacities and bonds were created or reformulated. Some scholars have even spoken of this period as a second scientific revolution, where modern disciplines emerge in rudimentary form and where there is a shift in the institutional locations of scholarly and scientific activities toward new research laboratories and universities. In this process, society itself may even, to quote one scholar (Manent, 1994, pp. 80–83), be said to have been discovered in its modern form, as distinct from the political order of the prerevolutionary period, whether it had the form of an absolutist monarchy or a constitutional one. Contrary to older forms of contractual reasoning, which postulated social contract as the hypothetical medium that had created political order, the new forms of social thought were distinctly historical and could also draw on anthropological experiences of human beings and societies in different parts of the world (Fox et al., 1995).

Broadly speaking, four distinct forms of thinking about human beings, societies, and political order emerged. These new forms of reasoning and the transition from generalized moral and political philosophy to social sciences had deep repercussions for political thought.

First, there was, with roots in the political economy and moral philosophy of the Scottish Enlightenment, a view of human beings as rational and deliberative individuals and of society as a compositional collective. This also made possible a new way of thinking about political community and its legitimacy. More precisely, it provided the possibility for thinking about political order as formed through the consent of individual human beings. However it was not, as in older contractual thinking, a consent that would have ushered in a distant and hypothetical social contract but rather a consent that could be assessed and rationally evaluated in terms of its reasonableness in its present and historical context.

Second, there emerged, often enough out of the experiences of the relative inability of the French Revolution to create stable institutions, a conception of human agency as formed and constrained by structural and systemic properties and of society as an organic totality. Such a conception would in a conservative version lead to theories of the necessity to restore an organic political order. In a radical version it could, however, also usher in demands for the overthrow of unnatural and illegitimate order. In a more broadly defined liberal or republican tradition, it might and would be compatible with ideas of the type that thinkers, often inspired by Durkheim, would propound toward the end of the nineteenth century in their combat against the ideological enemies of the republic. Thus it could support the idea that modern society was not an atomistic and distorted form of human community but could relate to an organic solidarity compatible with the life of free citizens in a free republic.

Third, and often in relation to historical and linguistic thought among representative of German philosophical idealism, a conceptualization of human beings took form in terms of the linguistic-interpretative nature of human agency and of society as an historically emergent totality. In this sense, political community did not have to derive its legitimacy from dynastic tradition but from the historically grown experience of a community. Such communities in turn were constituted by human beings sharing a language and a history and where the community of shared life forms was seen as deeper, more genuine, and more real than the artificial boundaries of a minor princely state. It is easy to see that demands for national self-determination and for both popular rule and a nation-state easily flow from a conception along these lines.

Finally, in parallel to the dramatic development of forms of statistical analysis, there emerged a statistical-inductive way of representing the interactions of human beings and of society as a kind of systematized aggregate. Such views did not immediately entail a specific form of legitimate political community, but it would mean that any such community would have at its disposal infinitely more detailed knowledge of the polity and its members than even the most absolutist monarchy (cf. Brian, 1994).

Thus the great intellectual transition entailed a deep shift in the history of political thought which affected both conceptions of the nature of a political community and the nature of legitimate and illegitimate enforcement. By virtue of this it also had profound implications for conceptions of the proper distribution of entitlements. Thus it was simply no longer possible, as observed most clearly perhaps by Tocqueville (1964, [1850]), to deny demands for popular participation with reference to tradition or dynastic legitimacy. The thought of a right to the political entitlements belonging to all members of a society, that had been manifested in the proclamations of the revolutionary upheavals, could never again, despite all efforts of the Vienna Congress and the Holy Alliance, be unthought. They were to pervade the political struggles first of European and American societies throughout the nineteenth and the early twentieth centuries and eventually in the world at large.

Triumph and Crisis of Liberal Political Thought

The end of World War I marked a sea change in world history. It spelt the end of European global pre-eminence. It also meant that the transition in Western Europe, normally from forms of

constitutional monarchical rule, to parliamentary democratic rule was dramatically increased. In the Eastern half of Europe all of the four imperial political entities, the Russian, the Habsburg, the Wilhelminian, and the Ottoman, crumbled under the impact of external military defeat and internal revolutionary upheavals. However, the very triumph of liberal principles of representative democracy was coterminous with the emergence of new forms of political thought that transcended the traditions of the prewar world.

Leninist political thought, which before and during the war had been a radical version of classical Marxism, gradually came to take on the features of an official ideology of the new Soviet state. Precisely in its historical imaginary, Leninism differed from both revisionist and orthodox thought in a Marxist tradition, as epitomized by Bernstein and Kautsky respectively, of the Social Democratic mainstream of the Second International in the late nineteenth and early twentieth centuries. Thus Leninism cast its own claims to legitimacy and its right to use forms of violent enforcement, previously unthought of in late nineteenth century socialist thought, in terms of an historical account and of a projected visionary transition to a stateless society of communism.

Experiences of the mass violence of the war itself, as well as of the violent national and social struggles of the immediate postwar period, formed the environment out of which fascism emerged as a violent rejection of both Marxism and of the whole mainstream tradition of liberal political thought. Even if fascism had roots in prewar forms of radical conservative thought, it marked a form of transcendence of such traditions and indeed of the legacy of modernity itself. Thus even if fascism in power, and its radicalized form of National Socialism, embraced technological modernization, it represented a rejection of basic premises of political thought of modernity in the form it had taken since the late eighteenth century. In particular this is true of the basic presupposition of political thought of modernity of a principled equal right to participation in governance. It is also true of fascism's rejection of another basic presupposition of modernity, namely the legitimacy of a public sphere with free access for all and a principled freedom of every member of a polity to speak not only about but also to the rulers.

In a sense fascist political thought was a form of political thought that invoked history, or rather historical mythology. However, it was not in itself characterized by historical reflection. In Germany, the challenge of fascism to the political order of the Weimar republic was manifested in a range of debates, one of the most important of which concerned the nature of legal order and of the state itself. It was in this debate that the contesting positions of the radical and pro-Nazi decisionism of Carl Schmitt was confronted by the legal positivism of Hans Kelsen and the complex and sophisticated thought of Hermann Heller.

The outlines of a radical rethinking of the history of Western philosophical and political thought can be discerned in the elective affinity between fascism and some forms of philosophical thought, in particular that of Heidegger of the late 1920s and early 1930s. However, such a program was never elaborated or extended, nor was the decisive response and resistance to it on the part of a thinker such as Ernst Cassirer. Instead the military defeat of fascism in World War II was seen to entail the definitive demise of features of political thought characteristic of fascism.

In the French context, the response to the challenges of fascism was to be found in several quarters. These would range, in political terms, from Marxism to radical Catholic political thought, and in scholarly terms, from sociology in the tradition of Durkheim to the profound renewal of historical studies of the Annales School. In the period after World War II, the French intellectual scene may have been the most important one in Europe in terms of a renewal of political thought. Thus it would comprise the nexus of contributions made by scholars elaborating a Marxian philosophical existentialist position, most importantly of course Sartre, but also thinkers such as Claude Lefort, who, starting out from such a position came to question the core of Leninist political thought. It would also comprise some of the most articulated liberal political thought as epitomized by Raymond Aron.

In the interwar years, the new challenges to liberal political thought had already given rise to a profound renewal of political thought in historical context. These were the years when new standard histories of political thought were being written. The most prominent example in the Anglo-Saxon world might have been a book that was to become a classic for generations, namely George Sabine's (1937) majestic volume *A History of Political Theory*. Refugees from continental Europe made key contributions to this debate. A classical example is the analysis of the meaning of democracy in historical context that was articulated by Joseph Schumpeter (1943) in *Capitalism, Socialism, and Democracy*. Most importantly, perhaps, Karl Popper transcended the bounds of philosophy of science and wrote one of the most quoted works in political thought during the twentieth century in his two-volume opus on *The Open Society and Its Enemies* (Popper, 1966, [1945]). This was, of course, a work that was pre-eminently characterized by the links between historical analysis and political philosophy of immediate and contemporary relevance.

In parallel, leading representatives of Scandinavian political and legal science made some pathbreaking contributions both to an understanding of fascism and of the nature of democracy. In this respect the names of Herbert Tingsten and Alf Ross stand out, particularly with Tingsten's pioneering studies of fascism and his and Ross's contributions to a political theory of democracy.

One prominent characteristic of antiliberal political thought of the first half of the twentieth century was its emphasis on the right of the polity to resort to enforcement and violence against the enemies of the polity without being bound by any other constraints than those of pure convenience. Thus the constitutive boundary between members and enemies of the political community, between those chosen to fulfill the destiny of that community and those deemed to stand in the way and worthy of being removed from community, and ultimately from life, became the most crucial one. The political community itself was seen as constituted in terms of a historical account, whether of the struggle of a class for its liberation against oppression or of a people for its legitimate sphere of life and domination. Resistance to such types of political thought of a historicist nature formed the core of the argument of Popper's political philosophy. It also formed the focus for

a range of contributions criticizing totalitarian political thought. Among the main contributors to this critique were thinkers as different as Jacob L. Talmon, Hannah Arendt, and Claude Lefort.

Renewal and Critique in Political Thought

The history of political thought in the first part of the latter half of the twentieth century is characterized by three parallel developments. First, political thought in this period cannot but reflect the deep tensions in a world with two superpowers in a state of cold war. At the same time large parts of the Third World were emerging out of the experiences of a colonial past and seeking a place in this new world of fundamental tensions. One manifestation of this historical context, was the absolute absence of publications in the tradition of liberal political thought in the countries under Soviet dominance, but also the relative absence of Marxist political thought at many universities in the English-speaking world. Another indirect consequence was a persistent interest in exploring the empirical basis of liberal democracy. In this sense, the contributions by scholars such as Robert Dahl and some of his colleagues at Yale university, not least Charles Lindblom, stand out as exemplary. Second, a classical mainstream tradition of historically orientated political philosophy made significant advances. Isaiah Berlin and Judith Shklar are pre-eminent representatives of this type of scholarship, and the works of Charles Taylor and Norberto Bobbio should also be included here. Third, however, philosophical and theoretical contributions were made that were to reshape the history of late twentieth century political thought. I shall highlight some of these contributions and their consequences in terms of two of the major conceptualizations of human agency and of society that were discerned in the section above about the rise of the social sciences at the turn of the eighteenth century.

In a rationalistic-deliberative tradition, a landmark was established by the publication of Kenneth Arrow's (1951) *Social Choice and Individual Values*. A few years later, economic theorizing would enter political science at large with Anthony Downs's (1957) book *An Economic Theory of Democracy*, to be followed in the early 1960s by a range of books applying game theory and social choice theory to political phenomena. (Incidentally, as testified by Gabriel Almond (1990, p.117) the institutional role of the Center for Advanced Study in the Behavioral Sciences at Stanford was a crucial one for some of the interactions involving Arrow, Dahl, and others. These interactions were of immediate importance to the very publication of Downs's book but also more generally to the breakthrough of rationalistic forms of analysis).

A few decades later, what had once appeared as an interesting but rather marginal concern of some specialized political theorists had evolved into a broad tradition of rational choice analysis with pervasive influence on both the theory and practice of political thought, particularly in North America. Society as a compositional collective was described not just as an analytically convenient metaphor but as a representation of societal order and with consequential claims concerning the nature of legitimate political action. One thrust of theorizing in this tradition was to demonstrate the limitations, or even the impossibility, of delimiting an unambiguous common good, whether arrived at with the help of a social welfare function or of a majority decision. It is in this context that John Rawls's (1971) *A Theory of Justice* appeared as a pathbreaking book. Drawing on an individualistic and contractual form of reasoning, succinctly criticised by several scholars, most notably perhaps Michael Sandel, Rawls was able to reintroduce discussions of justice and just distributions and allocations into the core of political philosophy. Even after decades of discussions and critique, both Arrow's and Rawls's contribution retain their stature. However, both of them signal a rupture with historical and contextual reasoning rather than the opposite.

In the broad tradition of linguistic-interpretative conceptions of human action and of society as a form of emergent totality in historical context, important new developments also occurred. Thus contributions by continental European phenomenological and hermeneutic philosophers in the 1950s and 1960s came to merge with influences from Anglo-Saxon linguistic philosophy in the tradition of Wittgenstein and Austin. Three forms of such thought, in particular, emerged and have exerted a deep influence on the ways of studying the history of political thought.

In 1959, Reinhart Koselleck (1988, [1959]), drawing on a tradition of hermeneutic and historical scholarship, published a book on critique and crisis. This book had the deep transformation of intellectual and political order at the turn of the eighteenth century as its focus but also contained the core of a research program on the history of key concepts of society and political order in a long-term perspective. This program was realized through the publication during the following decades of a set of nine massive volumes tracing the history of a limited number of such concepts, primarily in continental Europe.

Roughly at the same time, Jürgen Habermas elaborated another research program, drawing on the joint traditions of critical Marxist thought, hermeneutics, and speech act theory. In 1962 Habermas published a volume that had an analogous empirical focus as that of Koselleck (Habermas, 1989). Thus it was a reinterpretation of the emergence of a new type of public sphere in the late eighteenth century. Habermas's subsequent long-term program has consistently linked these different traditions and may in one sense be seen as a form of historical positioning but also a radical generalization of the philosophical insights of linguistic philosophy and speech act theory.

This later tradition also inspired Quentin Skinner's (1978) two volume major opus *The Foundations of Modern Political Thought* (a work, incidentally, written largely at the Institute for Advanced Study at Princeton). Skinner came to be the major representative of a new school, often identified with his – and John Pocock's old – university, namely Cambridge. It involved the setting of new and significantly more rigorous standards for the writing of the history of political thought, particularly in its demand that the linguistic and intellectual context of the actual writing of major works be carefully brought to the fore of analysis. After decades of discussions, there can be little doubt that intellectual history has made advances in this period, inspired by the works of Skinner and several of his colleagues, that still by no means have been fully explored. At the same

time, it is also clear that these works still leave room for a more extensive inquiry into the relationship between long-term institutional and intellectual change.

Finally, links between political thought and political agency are prominent in studies of mechanisms of inclusion and exclusion and on the granting and withholding of entitlements for human beings. Gender studies has clearly entailed a renaissance of the writing of the history of political thought with pioneering contributions by Carole Pateman and Seyla Benhabib, just to mention two of the most distinguished scholars on the late twentieth century.

Historically orientated scholarship has in recent years seemed to open up the possibility for situating political thought in a much more encompassing context than has hitherto been customary. Not least, scholars writing in countries outside of the North Atlantic and West European context seem now to be about to help bring about a more truly global effort to rethink the history of political thought. That would also, almost inevitably, contribute to a transcendence of some of the most severe limitations of contemporary political thought, in particular the following ones:

> The relative inability of political thought in the rationalistic tradition to transcend a theory of agency that is manifestly impoverished vis-à-vis available historical research but also vis-à-vis demands that social and political thought itself cannot but make concerning a theory of agency.
>
> (cf. Joas, 1996, 2000)

> The relative inability of the so-called communitarian critics of this tradition to supply a sustained historical account of their preferred community and to escape a tendency of entering into circular justifications of those very societies which they themselves happened to be located or identified with.
>
> The relative inability of research in the traditions of conceptual history and linguistic contextualism to contribute to an understanding of the institutional formation of political order in a long-term and also extra-European perspective.

In order to be able to advance beyond these limitations, it seems that political thought must become significantly more prepared to engage in studies of long-term processes of intellectual and institutional transformations. This will in all likelihood entail efforts to invigorate historical social science of a type that was characteristic of the classics of social science, not least Weber (cf. Schluchter, 1996). It would then also stimulate culturally and contextually sensitive social science of a type that has been advocated most ambitiously and consistently by scholars across a range of the social and historical sciences and with S. N. Eisenstadt (1995) as one of its most prominent representatives. The history of political thought in the last three decades of the twentieth century involved impressive advances. At the turn of the millennium, however, the need for historically and comparatively orientated research programs able to transcend the limitations of contemporary political thought also stand out as possibly more urgent than at any time since the emergence of social science itself in the end of the eighteenth century.

See also: Democratic Theory; Enlightenment; Field Theory; Hermeneutics, History of; Integration, Social; Intellectual History; National Socialism and Fascism; Scientific Disciplines, History of; Solidarity: History of the Concept.

Bibliography

Almond, G.A., 1990. A Discipline Divided: Schools and Sects in Political Science. Sage, Newbury Park, CA.
Arrow, K., 1951. Social Choice and Individual Values. Wiley, New York.
Benhabib, S., 1992. Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics. Routledge, New York.
Brian, E., 1994. La Mesure de l'Etat: Administrateurs et Géometres au XVIII Siècle. Albin Michel, Paris.
Downs, A., 1957. An Economic Theory of Democracy. Harper, New York.
Eisenstadt, S.N., 1995. Power, Trust and Meaning. University of Chicago Press, Chicago.
Elliott, J.H., 1992. A Europe of composite monarchies. Past and Present 137, 48–71.
Fox, C., Porter, R., Wokler, R. (Eds.), 1995. Inventing Human Science: Eighteenth Century Domains. University of California Press, Berkeley, CA.
Habermas, J., 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. MIT Press, Cambridge, MA.
Harré, R., 2000. One Thousand Years of Philosophy: From Ramanuja to Wittgenstein. Blackwell, Malden, MA.
Heilbron, J., Magnusson, L., Wittrock, B. (Eds.), 1998. The Rise of the Social Sciences and the Formation of Modernity: Conceptual Change in Context, 1750–1850. Kluwer, Dordrecht, The Netherlands.
Heilbron, J., 1995. The Rise of Social Theory. Polity Press, Cambridge, UK.
Joas, H., 1996. The Creativity of Action Polity. University of Chicago Press, Chicago, IL.
Joas, H., 2000. The Genesis of Values Polity. University of Chicago Press, Chicago, IL.
Koselleck, R., 1988 [1957]. Critique and Crisis: Enlightenment and the Pathogenesis of Modern Society. MIT Press, Cambridge, MA.
Lefort, C., 1988. Democracy and Political Theory. University of Minnesota Press, Minneapolis, MN.
Manent, P., 1994. An Intellectual History of Liberalism. Princeton University Press, Princeton, NJ.
Pagden, A. (Ed.), 1987. The Languages of Political Theory in Early Modern Europe. Cambridge University Press, Cambridge, UK.
Pocock, J.G.A., 1975. The Machiavellian Moment. Princeton University Press, Princeton, NJ.
Popper, K., 1966 [1945]. The Open Society and its Enemies, 2 vols. Routledge, London.
Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.
Sabine, G., 1937. A History of Political Theory. Holt, New York.
Schluchter, W., 1996. Paradoxes of Modernity: Culture and Conduct in the Theory of Max Weber. Stanford University Press, Stanford, CA.
Schumpeter, J.A., 1943. Capitalism, Socialism and Democracy. Allen & Unwin, London.
Shklar, J.N., 1998. Political Thought and Political Thinkers. University of Chicago Press, Chicago.
Skinner, Q., 1978. The Foundations of Modern Political Thought, 2 vols. Cambridge University Press, Cambridge, UK.
Skinner, Q., 1996. Reason and Rhetoric in the Philosophy of Hobbes. Cambridge University Press, Cambridge, UK.
Talmon, J.L., 1960. The Origins of Totalitarian Democracy. Praeger, New York.
Tocqueville, A.de, 1964 [1850]. Oeuvres Complètes, Vol. XII. Gallimard, Souvenirs. Paris.

Political Thought, History of

Rachel Hammersley, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of past ideas and writings concerning politics has been undertaken not just by historians, but by scholars from a variety of different disciplinary perspectives. This article focuses primarily on historical approaches to the study of past thinking about politics. In particular, three approaches that have been especially influential in the late twentieth and early twenty-first centuries are explored in detail: the Straussian approach initiated by Leo Strauss; the Cambridge School interpretation that was developed by J.G.A. Pocock and Quentin Skinner; and Begriffsgeschichte or Conceptual History that was pioneered by Otto Brunner, Werner Conze, and Reinhart Koselleck.

Definition

The history of political thought is a field of study which explores past ideas and writings concerning politics and can cover a whole range of issues including democracy, freedom, constitutionalism, and political obligation. As the label suggests, it is technically a subdiscipline of history and consequently, many practitioners emphasize the commonality in terms of subject matter, approach, and methodological techniques with other branches of history. However, the history of political thought also shares much in common with other disciplines and subdisciplines including philosophy (particularly ethics and political philosophy); politics, political science, and political theory; intellectual history, cultural history, and even literary history. Moreover, some of those who have made important contributions to the history of political thought are more comfortable describing themselves as political scientists or political philosophers than historians of political thought.

While scholars from different disciplines can and do work together and build on each other's findings, they have sometimes approached the study of past political thought with rather different purposes in mind. For example, some of those concerned with the study of politics or government have presented their own ideas against the background of accounts by previous writers, stretching back to Plato and beyond. Yet, in these cases, the study of these texts and ideas is viewed from the perspective of present concerns. The same has also been true of philosophers interested in political and ethical issues. Such scholars tend to focus on a canon of key texts or figures who are effectively seen as exploring the same issues and as engaging in dialogue with one another. In reaction to this, others in these fields have challenged the relevance of past discussions of political issues in favor of an empirical social science approach that focuses on how politics operates today or a philosophical analysis of political concepts exploring the grammar underpinning their current use. Set against these various trends (as well as against those historians who downplay the relevance of political thought, or even ideas more generally, to historical enquiry) is the orientation of those historians who have been interested in the history of political thought.

Even among those who consciously study the history of political thought there is no single approach to the subject. Rather a number of different methodologies have been developed. Consequently, the remainder of this article will focus on three approaches that have proved particularly dominant over the past 50 years and remain so today: the Straussian approach that takes its lead from the work of Leo Strauss; that of the Cambridge School, the leading figures of which are J.G.A. Pocock and Quentin Skinner; and Begriffsgeschichte or Conceptual History that was pioneered by the German scholars Otto Brunner, Werner Conze, and Reinhart Koselleck.

Approaches

Leo Strauss and the Straussians

Though difficult to comprehend in the current climate, in the early to mid-twentieth century the history of political thought was widely perceived to be an obsolete field, having been supplanted by a scientific approach to politics more typical of the social sciences. The German-born American political philosopher Leo Strauss (1899–1973) analyzed this situation; offered a robust critique of the 'new political science'; and so helped to bring about the revival of the discipline (Smith, 2009: 2).

The major starting point of Strauss's thought was the idea that the modern world had reached crisis point: "The crisis of modernity reveals itself in the fact, or consists in the fact, that modern western man no longer knows what he wants – that he no longer believes that he can know what is good and bad, what is right and wrong" (Strauss, 1975: 81). And, of course, this has important implications for questions such as what sort of government is desirable or when opposition to rulers is legitimate. In an attempt to understand this crisis Strauss drew a sharp distinction between the ancients and the moderns, inverting the outcome of the early eighteenth-century 'querelle des anciens et des modernes' by asserting the superiority of the former over the latter. On Strauss's analysis it was Niccolò Machiavelli and Thomas Hobbes who signaled the beginning of the modern era and initiated the first of three waves of modern thought (Strauss, 1975: 81).

Machiavelli was among the first to challenge the belief in a set of fundamental and eternal values governing human morality and politics, as well as the notion that there is a single ultimate form of human perfection. He thus shifted the focus of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62163-3

political thought from how humans ought to live (the ideal) to how they actually live (the pragmatic or practical). In doing so Machiavelli was accused by Strauss of opening the door to the relativism and nihilism that characterized the twentieth century, and which arguably is still with us today. At the same time Machiavelli introduced the idea that rather than having simply to submit to chance, as ancient thinkers had believed, human beings actually had the capability to shape and even control fortune, if they were sufficiently cunning. This idea was then developed in the midst of the scientific revolution of the sixteenth and seventeenth centuries, becoming manifest in the Enlightenment belief that through science and technology human power could be used to conquer and control nature and to reform society. The application of scientific methods not just to the natural world, but also to human nature and society posed a particular problem for Strauss since science deals in facts not value judgments, whereas the human and social sciences, such as political philosophy, depend upon deep-seated value judgments that must simply be accepted rather than being rationally validated. Ultimately, Strauss was concerned at the idea, prevalent in the social sciences at the time and epitomized by Max Weber, that a scientific approach could supplant the ancient role of philosophy. Thus, the ancients were superior, in his eyes, because their philosophers (though not necessarily the people more generally) appealed to eternal values and had a clear sense of what was just, virtuous, and good. This was true not only of the classical philosophy of Plato and Aristotle, but also of the account offered in the Bible: in this "crucial respect there is agreement between classical philosophy and the Bible, between Athens and Jerusalem, despite the profound difference and even antagonism between Athens and Jerusalem" (Strauss, 1975: 86).

In his analysis of western political thought Strauss sought to counter this loss of knowledge by both uncovering the innovations of the moderns (including positivism and historicism) and reemphasizing and reinvigorating the eternal concepts and values of the ancients. This also had an impact on the way in which he approached the history of political thought and his methodology. In studying the ancients Strauss was primarily concerned with their understanding and appreciation of fundamental values. When he came to the moderns he focused on their undermining of these values and the consequences that ensued. His works thus read as though the canon of great thinkers were essentially engaged in discussion with each other and the focus is almost exclusively on their development and transformation of the field of political philosophy. His methodology therefore focused almost entirely on the texts. He paid very little attention to the historical and intellectual contexts in which they were written, firmly rejecting the idea that an appreciation of these is necessary in order to fully understand the ideas themselves.

In challenging the path taken by modern thinkers, Strauss drew inspiration from the approach to philosophy that had been taken by Arabic thinkers during the medieval period, in particular Abu Nasr Alfarabi (*c*.870–950). These thinkers offered, so Strauss contended, a deeper interpretation of the ideas of ancient philosophers such as Plato and Aristotle than was common in parallel western accounts, with the Arabic writers retaining a belief in certain eternal ethical and political concepts. Strauss also drew from these thinkers his controversial concept of esotericism. It was based on the idea that most political thinkers in the past had to conceal their deepest ideas to avoid persecution. This called for a reinterpretation of the canon and a new emphasis on 'careful reading' to uncover the hidden meanings of the great thinkers of the past.

Though Strauss himself died in 1973, his approach has survived and continues to be influential, particularly in the United States. Not all Straussians agree with each other and there are particular distinctions between the East and West Coast groups. One topic that has received emphasis from Straussians in recent years (partly in response to its emphasis among historians of political thought who adopt the rival Cambridge School approach) is republicanism. Paul Rahe offered a fundamentally Straussian account of the history of republican thought in his enormous book *Republics Ancient and Modern*, which appeared in 1992. As its title suggests, it was an attempt to underline the fundamental caesura introduced by Machiavelli. Crucial here was the shift away from seeing civic virtue as inherently good and valuable toward a much more instrumental understanding of it as a means of achieving freedom and individuality. In his more recent book, *Against Throne and Altar*, which focuses more closely on the period of the English Revolution, Rahe developed his account of the emergence of modern republicanism more fully while also paying closer attention to the legacy and influence of Arabic authors including Alfarabi and his followers (Rahe, 2008). This Straussian interpretation of republicanism has been further explored by other scholars including Thomas Pangle and Vicki Sullivan.

The Cambridge School

The series of approaches known collectively as the 'Cambridge School' were devised, at least in part, as a response by historians to Strauss's political philosophy and other philosophical approaches. The leading figures of this school, J.G.A. Pocock and Quentin Skinner, were motivated by similar concerns and had shared aspirations, but each developed his own particular approach. This school derives its name from the fact that it was at Cambridge that Pocock and Skinner first developed their methodological ideas and that institution also housed other pioneers and practitioners of the approach, most notably John Dunn, Peter Laslett, and Gareth Stedman Jones. Before exploring the particular approaches of Pocock and Skinner in detail it is worth focusing on the concerns that prompted their methodological work and the common features that their investigations share.

Essentially, the approach of the Cambridge School was designed to cut a middle way between history and philosophy. Pocock, in particular, was concerned that the study of past political thought had tended to be dominated by philosophers who lacked a proper historical training. Consequently such philosophers were less concerned by what past authors intended by their claims and more interested in what "he in his present can make it mean: what he can do with it for purposes of his own, which may or may not – and therefore do not have to – coincide with those of the author" (Pocock, 1971: 7).

A further consequence of this philosophical focus was the tendency to understand the history of political thought in terms of a canon of greater thinkers and texts. As Pocock

explained: "Alone among the major branches of historical study in the middle twentieth century, the history of political thought was treated as the study of a traditional canon" (Pocock, 1971: 5). Skinner too was concerned by this tendency, directing his criticism more specifically at the Straussians and their apparent assumption that the great political thinkers of the past had effectively been engaged in a dialogue with each other, rather than being largely inspired by and engaged with the troubles, debates, and thinkers of their own time. Thus, members of the Cambridge School were keen to re-establish the historical dimension to the study of political thought and to construct it as a firmly grounded historical discipline.

At the same time advocates of the Cambridge School also expressed their discontent with a marked tendency among historians to downplay or even dismiss ideas completely. This was not only a feature of high political and Namierite interpretations that saw ideas as little more than a means of rationalizing what were in reality selfish political interests, but also of the Marxist approach. Here political thought tended to be viewed as mere 'idealism,' with ideas simply reflecting more fundamental social and economic factors. Indeed even where Marxist-inspired scholars did treat ideas more seriously, as with C.B. Macpherson in his influential, but controversial work, *The Political Theory of Possessive Individualism*, they tended still to be portrayed as superficial representations of deeper social phenomena (Pocock, 1971: 10 and 37).

Ironically, given the concerns about philosophical approaches to the study of the history of political thought, the methodologies that Pocock and Skinner developed to chart their middle course were themselves grounded in philosophy, more specifically the linguistic philosophy of the 1950s. "What seemed to many, about 1956, the subversion of political philosophy by linguistic analysis helped to liberate the history of political thought by converting it from a history of systematization ('philosophy' in an old sense) into one of linguistic use and sophistication ('philosophy' in a new)" (Pocock, 1971: 11–12). Thus members of the Cambridge School drew on several philosophical ideas and theories. These included Ludwig Wittgenstein's explorations of the diverse language games in terms of which human beings make sense of the world and J.L. Austin's notion of speech acts. Also of influence was Thomas Kuhn's treatment of the history of science in terms of paradigms and paradigmatic shifts.

Pocock's approach sought to move the focus away from particular thinkers or texts by adopting political languages as the key unit of analysis and examining the ways in which those languages were developed, adapted, and even subverted over time. What he meant by political language, he explained, was the idioms, rhetorics, specialized vocabularies, and grammars – the modes of discourse or ways of talking about politics that developed at particular points in the past (Pocock, 1971: 21). It was the task of the historian of political thought, he insisted, to learn these languages by reading many texts in order to be in a position to detect not only uses of them, but also developments, refutations, and subversions. In his own work, Pocock has focused on what he saw as a number of particularly significant early-modern political languages including the common law, civic humanism (or republicanism), and political economy, but his approach has also inspired the in-depth study of other languages, including natural law.

Though undoubtedly influenced by Pocock's emphasis on languages, Skinner has adopted a slightly different approach by emphasizing the need for historians of political thought to address the authorial intention or the 'illocutionary force' of particular speech acts. In other words, Skinner insisted that an historian of political thought should seek to uncover exactly what a particular thinker or writer intended or meant by a particular statement or work; how it fitted within, altered, or subverted the conventional understandings of the time; and what (s)he was seeking to achieve by it.

While both of these approaches require the close reading of particular texts in order to uncover their precise meaning, they also depend on a deep understanding of the immediate context in which the author was writing. Moreover, this includes not only the historical or political context, the particular events or issues of the time that stimulated the author into producing the ideas or works that (s)he did, but also the intellectual context: the other works on which the author drew or against which (s)he was arguing and the debates in which (s)he was engaged. This emphasis is perhaps best illustrated in *The Foundations of Modern Political Thought*, which Skinner published in 1978. The aim of these two volumes (covering the Renaissance and the Reformation, respectively) was to provide the background or context in which some of the great works in the history of political thought had been written. Unlike normal textbooks in the history of political thought these did not simply provide a summary of the key works themselves, rather they were designed to be read in conjunction with the texts and would help the reader to situate the works in their appropriate contexts. More recently, this approach has been further supported by Cambridge University Press through two series devoted to the history of political thought. The first, *Cambridge Texts in the History of Political Thought* (colloquially known as the blue texts series on account of the distinctive covers), offers cheap and accessible editions not only of the great canonical works, but also of other works of interest and distinction from the ancient world right up to the present. Each volume also includes a scholarly introduction designed to outline the context in which the work was produced. Secondly, *Ideas in Context*, a collection of scholarly monographs and edited collections relating to various aspects of the field.

As always with advocates of new and influential methodologies, the Cambridge School has not escaped criticism. Some have questioned the universalist claims that have been made for an approach that has centered largely on political texts produced between around 1450 and 1800. Others, more specifically, have derided the elitism inherent in the approach and have criticized the alleged absence of any real exploration of social background or institutional structure or indeed of the reception of texts and ideas. Still others have questioned whether Skinner, in particular, really applied his own theories in practice. Despite placing emphasis on intellectual context and lesser known ideologies and thinkers, his own work has continued to center on key figures such as Machiavelli and Hobbes. At the same time, though preaching a properly historical approach, some have argued that his work is teleological in its focus on the emergence of modern concepts such as the state. Despite these criticisms, over the last 40 years, the Cambridge School approach has become the dominant method for pursuing the history of political thought in Britain

and has also proved influential abroad, although on the continent it has had to contend with a distinctive and distinguished alternative in the form of Conceptual History.

Conceptual History

In Continental Europe, and particularly in a German and Dutch context, a different approach to intellectual history and the history of political thought has been pursued by advocates of Begriffsgeschichte or Conceptual History. The origins of this approach are similar in a number of respects to those of the Cambridge School. Like Pocock and Skinner, advocates of Begriffsgeschichte were trained historically and consequently their approach is firmly grounded in historical methodology. Once again, it is in part a response to the Straussians. It also shares with the Cambridge School (and especially Pocock) its emphasis on language (particularly political language) and the ways in which meanings and usages change over time. However, in contrast to Pocock and Skinner, advocates of Begriffsgeschichte focus their attention on key concepts and treat these – rather than texts, thinkers, or languages – as central to their analysis.

The founding work of this movement was the Geschichtliche Grundbegriffe (GG), a huge historical dictionary or encyclopedia of concepts, which was first announced in 1967 and the first volume of which appeared 5 years later. There were three key figures involved in this project. First, Otto Brunner, whose background was in medieval philology, but who had anticipated many of the ideas of Begriffsgeschichte in his influential book Land und Herrschaft (*Land and Lordship*) of 1939. Secondly, Werner Conze, a social historian and a pioneer of labor history in Germany (Van Horn Melton, 1996: 33). Thirdly, Reinhart Koselleck, who is perhaps best thought of as an intellectual historian and is the only one to have seen the project through from beginning to end. It was Koselleck who apparently came up with the idea of producing a dictionary of concepts while working for Conze (Tribe, 1989: 180). But it was also Koselleck who completed the project after the death of Brunner in 1982 and Conze in 1986. The initial plans for the project were relatively limited. A single volume focused on German concepts was envisaged. However, by the time the final volume appeared in 1997, the work consisted of eight volumes, comprising 120 entries and more than 7000 pages.

The breadth of interests among the editors is reflected in the GG and in Begriffsgeschichte more generally. The focus is not simply on political thought, narrowly defined, there is also a strong social history dimension and a real attempt to integrate the two. From the outset, Koselleck was keen to take full account of sociopolitical experience rather than simply focusing on ideas and to emphasize the interrelationship between them (Tribe, 1989: 181; Koselleck, 1985: 73–91). As he explained: "Without common concepts there is no society, and above all, no political field of action. Conversely, our concepts are founded in politicosocial systems that are far more complex than would be indicated by treating them simply as linguistic communities organized around specific key concepts" (Koselleck, 1985: 74). Moreover the emphasis is on exploring the debates surrounding the meaning of particular concepts and uncovering the relationship between conceptual change and more concrete social structural transformations (Richter, 1990: 41).

Another key difference between the Cambridge School and German advocates of Begriffsgeschichte is that members of the former have tended to focus their attention on the period between the fifteenth and eighteenth centuries, whereas the latter were focused more narrowly on the period between the mid-eighteenth and mid-nineteenth centuries. Brunner had already presented the nineteenth century as constituting a fundamental break with the past in his earlier work and he saw conceptual history as a means of bridging that gulf between past and present (Van Horn Melton, 1996: 22). However, it was Koselleck who dubbed this period the Sattelzeit or saddle period, and defined it as an era of rapid political and social change during which a number of distinctively modern political and social concepts were said to have emerged. As Koselleck himself explained, the aim was: "to investigate the dissolution of the old world and the emergence of the modern world through the history of its conceptual comprehension" (Tribe, 1989: 183). In the GG a distinction is drawn between three types of concepts (Koselleck, 1985: 82). First, those long in use, in particular terms that had been applied in antiquity and are still current in the modern world. Secondly, those concepts whose meaning was completely transformed during the shift to modernity (for example, state/estate). Thirdly, neologisms which were invented generally during the Sattelzeit (such as communism or fascism). The project also highlighted four processes that commonly affected basic political and social concepts in the German-speaking world between 1750 and 1850. First, Verzeitlichung, which refers to the tendency: "to insert modern political and social concepts into one or another philosophy or horizon of history set out teleologically in terms of periods, phases, or stages of development" (Richter, 1996: 12). Secondly, Demokratisierung or the democratization of political and social vocabularies that had previously been confined to an elite. Thirdly, Ideologiesierbarkeit or the incorporation of concepts into ideologies. Finally Politisierung or the politicization of concepts and their use for political purposes.

While initially developed in relation to the German-speaking world, Begriffsgeschichte has subsequently been applied to other European languages. Rolf Reichardt and E. Schmidt adopted a similar approach to French history in their Handbuch polititisch-sozialer Grundbegriffe in Frankreich 1680–1820 (*A Handbook of Political and Social Concepts in France*, 1680–1820), which appeared in 1985. Reichardt has also continued to pursue this approach in his more recent work, which has often set the exploration and interpretation of visual images alongside more traditional written texts.

In the 1990s, a group of Dutch scholars were also attracted to this approach and several of them established a research project in Dutch conceptual history. The initial project focused on three key concepts that were seen to be important in a Dutch context: Liberty, Fatherland, and Citizenship. Three interdisciplinary research groups were set up to explore each of the concepts in a Dutch setting. Since then the project has expanded in several ways. In the first place, the group has begun to work on additional concepts, including honor, gender, state/estate, and modernity. The ambitions are perhaps less grand (in this sense at least) than the GG, with talk of up to 15 concepts in total. However, attempts have also been made to develop a more comparative approach. A series of workshops has already been held in which the use of the three

initial concepts in a Dutch context was contrasted and compared with the ways in which those same concepts developed in British, French, and German settings. A comparative approach was also offered in a volume that was produced by the group in 1998 and was entitled *History of Concepts: Comparative Perspectives* (Hampsher-Monk and Tilmans, 1998). Interestingly, the Dutch scholars set out with the specific intention of exploring the distinctiveness of the Dutch experience. This led to a greater focus on the seventeenth century (as a particular high point in Dutch history) and the expectation that the Sattelzeit would be of less significance in a Dutch context. However, the preliminary findings suggest that this was not the case and that in Dutch, just as in German, it was the period between 1750 and 1850 that witnessed the emergence of the modern conceptual world.

Similar projects have also been proposed focusing on other linguistic areas. In his contribution to the American conference to celebrate the completion of the GG, Melvin Richter spoke of the possibility of a Hungarian project and referred to an additional initiative relating to keywords of the Chinese Revolution from 1911 to the present (Richter, 1996: 19). More recently British and French scholars have become interested in exploring keywords or concepts and several projects based around this idea are currently taking shape.

Although conceptual history appears to be on the rise, there are still problems and drawbacks with the approach pioneered by the GG. In particular the alphabetical organization of that work presents obstacles to its usage, hindering – for example – attempts to use it to construct vocabularies or languages that were dominant at particular times. Interestingly the editors had originally hoped to implement a more complex organization that would have rendered these kinds of connections easier to make, but the complexities involved in doing so pushed them toward the lexicon format. One of the central tenets of Richter's attempts to integrate the Cambridge School and Begriffsgeschichte approach has been to stress what the latter could learn from the focus on languages or ideologies that is characteristic of the work of the Cambridge scholars. Criticism has also concerned the particular periodization adopted by the GG, and especially the emphasis on the period 1750–1850. Yet, given the tentative findings of the Dutch project it may be that this focus is more justified than some have thought.

Conclusion

As this overview has demonstrated, there are different ways of approaching and studying the history of political thought. These approaches each have their own strengths and weaknesses, and their practitioners might indeed learn from each other as Richter suggests. While there is no single dominant approach, what can be said with confidence is that the fortunes of the field have been transformed over the last century. Having been deemed obsolete in the early twentieth century it would seem that the history of political thought is alive and well at the beginning of the twenty-first.

See also: Cultural History; Freedom, History of; Historical Thought and Historiography: Current Trends; Historicism; History and Literature; Intellectual History; Linguistic Turn and Discourse Analysis in History; Political History: History of Politics.

Bibliography

Hampsher-Monk, Iain, Tilmans, Karen, 1998. History of Concepts: Comparative Perspectives. Amsterdam University Press, Amsterdam.

Koselleck, Reinhart, 1985. Future Past: The Semantics of Historical Time. MIT Press, Boston, MA.

Lehmann, H., Richter, Melvin (Eds.), 1996. The Meaning of Historical Terms and Concepts: New Studies on Begriffsgeschichte. Occasional Paper. German Historical Institute, Washington. http://www.ghi-dc.org/publications/ghipubs/op/op15.pdf.

Macpherson, C.B., 1962. The Political Theory of Possessive Individualism. Clarendon Press, Oxford.

Pangle, Thomas L., 2004. Straussian approaches to the study of politics. In: Gaus, Gerald F., Kukathas, Chandran (Eds.), Handbook of Political Theory. Sage, London.

Pocock, John Greville Agard, 1971. Politics, Language and Time: Essays on Political Thought and History. Methuen & Co. Ltd, London.

Pocock, John Greville Agard, 1987. The concept of a language and the metier d'historien: some considerations on practice. In: Pagden, Anthony (Ed.), The Languages of Political Theory in Early-Modern Europe. Cambridge University Press, Cambridge.

Rahe, Paul A., 1992. Republics Ancient and Modern. University of North Carolina Press, Chapel Hill, NC.

Rahe, Paul A., 2008. Against Throne and Altar: Machiavelli and Political Theory under the English Republic. Cambridge University Press, Cambridge.

Richter, Melvin, 1990. Reconstructing the history of political languages: Pocock, Skinner and the Geschichtlich Grundbegriffe. History and Theory 29, 38–70.

Richter, Melvin, 1996. Appreciating a contemporary classic: the Geschichtliche Grundbegriffe and future scholarship. In: The Meaning of Historical Terms and Concepts: New Studies on Begriffsgeschichte. Occasional Paper. German Historical Institute, Washington, DC. http://www.ghi-dc.org/publications/ghipubs/op/op15.pdf.

Robertson, Neil G., 1998. The closing of the early modern mind: Leo Strauss and early modern political thought. Animus 3, 1–16.

Skinner, Quentin, 1978. The foundations of modern political thought. In: Volume 1: The Renaissance. Volume 2 The Age of Reformation. Cambridge University Press, Cambridge.

Skinner, Quentin, 2002. Visions of politics. In: Regarding Method, vol. 1. Cambridge University Press, Cambridge.

Smith, Steven B., 2009. Introduction: Leo Strauss today. In: Smith, Steven B. (Ed.), The Cambridge Companion to Leo Strauss. Cambridge University Press, Cambridge.

Strauss, Leo, 1975. The three waves of modernity. In: Gilden, Hilail (Ed.), Political Philosophy: Six Essays by Leo Strauss. Bobbs-Merrill, Indianapolis, pp. 81–98.

Tilmans, Karen. Applying Begriffsgeschichte to Dutch History. Some Remarks on the Practice and the Future of a Project. At https://docs.google.com/viewer?a=v&q=cache:PsSoDGHVIOUJ:dare.uva.nl/document/43514+Dutch+conceptual+history&hl=en&gl=uk&pid=bl&srcid=ADGEESgcJ58OtVZsuRdl15ZExydXzqmnjXNwR-RyctuOEUjxlsqDP_GXs4W_me_sn300nWFTeTTPJGLKFRywe_k1vw1tvh5tc-Gu9Os56y2CPNjED5wWcwu4MFMfa0JWsASkDlHsfVfX&sig=AHIEtbSfpNKwJOdMitw9njUumyo82HrTQw

Tribe, Keith, 1989. The Geschichtliche Grundbegriffe project: from history of ideas to conceptual history. A review article Comparative Studies in Society and History 31, 180–184.

Tully, James, 1988. Meaning and Context: Quentin Skinner and His Critics. Princeton University Press, Princeton, NJ.

Van Horn Melton, James, 1996. Otto Brunner and the ideological origins of Begriffsgeschichte. In: The Meaning of Historical Terms and Concepts: New Studies on Begriffsgeschichte, Occasional Paper. German Historical Institute, Washington, DC. http://www.ghi-dc.org/publications/ghipubs/op/op15.pdf.

Whatmore, Richard, 2006. Intellectual history and the history of political thought. In: Whatmore, Richard, Young, Brian (Eds.), Palgrave Advances in Intellectual History. Palgrave Macmillan, Basingstoke, pp. 109–129.

Political Tolerance

Sandra T Marquart-Pyatt, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political tolerance is essential to vibrant, healthy democracies. It involves taking key democratic principles like support for civil liberties and political rights and applying them in practice to disliked groups. Scholars have tracked public opinion on tolerance for more than 60 years using group-specific and content-controlled measures. Recent works have expanded the focus from the United States to other established democracies and to nascent democratic regimes. These studies reveal intolerance to be widespread around the world and strive to understand why this is the case.

Introduction

Political tolerance, defined as the willingness to grant political rights to disliked groups, is central to studies of democratic values given its links with civil liberties and political freedoms. For more than six decades, scholars have tracked public opinion on political tolerance given its role in supporting democratic values and political rights like freedoms of speech and association and the right to vote, demonstrate publicly, and hold an elected political office. Taken as a whole, these studies have yielded many insights. Of these, perhaps most significant is that establishing and maintaining tolerant publics, even in long-standing democratic nations, remains a challenge. Nontrivial gaps between tolerance in principle and in practice remain in both established and nascent democracies (Marquart-Pyatt and Paxton, 2007). While in some contexts intolerance toward specific groups has fluctuated and may have even attenuated, it is also possible in other contexts that levels of tolerance have remained the same or perhaps even increased (Gibson, 2008, 2013; Shafer and Shaw, 2009; Sullivan and Hendriks, 2009). In addition, in the United States a shift in attitudes toward granting the government authority to suspend civil liberties such as in the aftermath of 9/11 has had a profound influence on this work (Davis, 2007; Sullivan and Hendriks, 2009). Although a growing number of studies attest to the importance of investigating tolerance worldwide, research reveals it to be contextually grounded as well as potentially developing at uneven rates globally, leading to complexities with regard to how it is realized locally, nationally, and globally.

Since the seminal work on political tolerance in the United States by Samuel Stouffer (1955) describing how the American public viewed the civil liberties of communists during the McCarthy era, research has investigated different ways of measuring tolerance, broadened in geographic scope to national and cross-national work, and considered a number of theoretical frameworks. Studies have employed different measures of political tolerance, including fixed-group and content-controlled measures, for a number of different social, political, and ethnic groups. Much of this research has been country-specific, focusing on tolerance in the United States (Gibson, 2008, 2013; Sullivan and Hendriks, 2009; Sullivan and Transue, 1999). Research on political tolerance in the American public has examined tolerance over time toward left-wing political groups like communists and socialists, atheists, religious groups like Muslims, and right-wing racists (Sullivan and Hendriks, 2009). As described in this article, research has also explored tolerance in other established democracies like Australia, Britain, and Israel, as well as in newer democracies like Poland, the former Soviet Union, and South Africa. Few of these studies attempt to compare across new and established democracies, however, and even fewer attempt comparisons across a substantial number of countries. A handful of recent works examine political tolerance across more than two dozen countries. A variety of explanations have been used to characterize individual-level variation in tolerance including democratic norms (Sullivan et al., 1982, 1985), democratic learning and protest (Guerin et al., 2004; Marquart-Pyatt and Paxton, 2007; Peffley and Rohrschneider, 2003), and threat perception (Sullivan et al., 1982, 1985). An emerging line of inquiry seeks to include macro-level frameworks like democratic learning (Peffley and Rohrschneider, 2003).

Political tolerance entails the extension of political rights to groups with opposing or disliked views and is a pivotal democratic value essential for democratic functioning and survival. Political tolerance involves taking key democratic principles like support for civil liberties and minority rights and applying them in practice to disliked groups. Healthy democracies require citizens to put up with or tolerate the views and political participation of other groups, even those groups whose views they might oppose or whose participation they may perceive as threatening (Almond and Verba, 1963; Sullivan and Transue, 1999). Political intolerance targets the political freedoms of particular groups and ideologies through restrictions on civil liberties like freedom of expression and association. Tolerance is especially important in nascent, unconsolidated democracies where intolerance may cause instability in providing transitions from one regime to another as groups resist ceding power to rivals or try to institutionalize restrictions on civil liberties for particular groups (Guerin et al., 2004). Tolerance is also salient in established democracies given its crucial role in solidifying a broad array of civil liberties and political rights for a wide range of groups.

In this article, I provide an overview of published research on the topic of political tolerance that covers three areas: (1) measures of political tolerance in the literature; (2) national, international, and comparative research grouped into single country case studies, comparative small sample (N) and large N

works; and (3) major explanatory frameworks and variables investigated. Special emphasis is placed on work published in the last two decades to facilitate discussion of the road ahead that includes a brief overview of existing datasets available for secondary analysis of political tolerance across nations and contexts and conceptual issues that remain, laying a foundation on which future theoretical and empirical work can build.

Measuring Political Tolerance

The extant literature employs a number of different measures of political tolerance. Although conceptually related, political tolerance is neither synonymous with civil liberties nor is it necessarily linked with social or ethnic tolerance (see Gibson, 2008). At least three approaches to measuring tolerance can be gleaned from this scholarship. Two are vital to distinguish given their long-standing influence on tolerance studies: group-specific or fixed-group and content-controlled measures. A third is also important given how it engages policies related specifically to restrictions on civil liberties with corresponding implications for democratic theory. In the following, each is described in turn, although the bulk of attention is devoted to the first two measures.

One measurement strategy in tolerance research is a group-specific or fixed-group approach. Group-specific tolerance research asks questions about particular groups such as communists, socialists, and atheists (Stouffer, 1955; Nunn et al., 1978). These items are routinely gathered in the General Social Survey (GSS) (Smith et al., 2013). The GSS is a nationally representative sample of the US public, conducted since 1972, containing information on demographic, attitudinal, and behavioral questions on a range of social issues. Regarded as an excellent source of societal trend data, the GSS has included questions about a few groups in many years and some in specific years including, for instance, communists, socialists, militarists, atheists, feminists, gays and lesbians, and immigrants.

Studies using fixed-group measures of tolerance over time reveal that some groups included in this set of survey items have become more socially accepted while others have not necessarily been recipients of more welcoming attitudes. A recent trend article using the GSS illustrates broad patterns among survey respondents of mildly increasing tolerance toward people who differ from them, especially for gay and lesbian individuals (Shafer and Shaw, 2009). Rather than being illustrative of a broad shift, however, Shafer and Shaw (2009) suggest this may indicate that other groups in this time frame from the early 1990s to 2007 may have moved into the category of least-liked. This confirms the work of Sullivan et al. (1979, 1982). A few cross-national studies on tolerance have used group-specific measures. For example, Karpov (1999a, 1999b) examines tolerance toward communists, militarists, and atheists, while Duch and Gibson (1992)'s work investigated fascists. These studies reveal similar patterns of levels of tolerance in the United States and Poland in the 1990s (Karpov, 1999a, 1999b) and across 12 Western European countries (Duch and Gibson, 1992).

A second measurement strategy in tolerance research is the least-liked or content-controlled approach. Content-controlled measures of tolerance seek to disentangle ideological orientation and tolerance (Lawrence, 1976; Sullivan et al., 1981, 1982, 1985). With this measure, respondents first pick the group that they like least from a list of possible groups. Then, respondents are asked whether they would extend civil liberties (such as the right to make a speech in the community) to their *least-liked* group. As an example, in the 1995–97 *World Values Survey* (*WVS*) (Inglehart et al., 2000) individuals were asked to name their most disliked group from the following list of choices. (The list of least-liked groups varied in a few countries, reflecting national differences in the relevance of different groups. For example, in the United States respondents were asked about immigrants without visas, right-wing extremists, Islamic fundamentalists, welfare recipients, criminals, homosexuals, and left-wing extremists. The WVS has gathered data since 1990, and as of 2014 is currently collecting its sixth wave of data in 50 countries. Regarding population coverage, the WVS gathers data from nationally representative samples from 9 out of 10 people worldwide, with some differing national samples included in different cross-sectional studies. It is viewed as the largest source of information on human beliefs and values across a wide range of social, political, and economic topics, particularly regarding how it can be used to study changes in these views. The content-controlled questions on political tolerance were included only in the 1995–97 wave.):

1. Jews
2. Capitalists
3. Communists
4. Immigrants
5. Homosexuals
6. Criminals
7. Neo-Nazis

After the respondent identified his or her least-liked group from the list of approximately seven choices, they were asked three follow-up civil liberties questions – whether members of their least-liked group should be allowed to:

1. hold public office
2. teach in schools
3. hold public demonstrations

Content-controlled measures like the one described above seek to disentangle a respondent's political ideology with their level of tolerance. That is, if a respondent leans to the left, then he or she should appear more tolerant than a right-leaning respondent if asked about the civil liberties of a leftist group like socialists. However, if the civil liberties of fascists are in question, then the right-leaning respondent may appear more tolerant. In either case, it is the political ideology of the respondent that is driving their answers, rather than their actual level of tolerance. To avoid this problem of confounding a respondent's political ideology with their level of tolerance, content-controlled measures of tolerance were developed (Lawrence, 1976; Sullivan et al., 1981). By allowing respondents to choose a least-liked group, content-controlled measures attempt to isolate political tolerance from political ideology. For some studies of tolerance, especially comparative analyses, using a content-controlled

measure is particularly important. If only a single group was considered, individual responses to that group could vary due to national differences in political history or values, rather than individual differences in tolerance. This would apply, for instance, in the case of fascists or neo-Nazis, which have a different political history in Germany than in the United States.

Research employing content-controlled measures of tolerance reveals a powerful role of context in varied settings like the established democracies of the United States, Israel, and New Zealand (Sullivan et al., 1982, 1985). Briefly, even in investigations that include controls for the content of political ideology, tolerance and intolerance are revealed to be linked to the liberty or right in question (Sullivan et al., 1982), are situationally and group-dependent (Sullivan et al., 1985), and least-liked groups are more diffuse in some countries than others where tolerance appeared to be more focused (Sullivan et al., 1985). In essence, contexts vary substantially. In addition, respondents selected least-liked groups that tended to be on the opposite end of the political spectrum from their views (Sullivan et al., 1985). A trend study using content-controlled measures has not been published to this author's knowledge given limitations of data availability described later as well as assertions that temporal variability in least-liked groups may limit the feasibility of such a work. Rather than attempting a temporal comparison, recent work has begun to use a content-controlled measure in more than a dozen countries gathered in the aforementioned 1995–97 wave of the WVS. These studies reveal striking similarities in levels of intolerance across 12, 13, and 17 countries (Guerin et al., 2004; Marquart-Pyatt and Paxton, 2007; Peffley and Rohrschneider, 2003).

A third approach to measuring tolerance gauges respondents' views on the suspension of civil liberties by the US government, minimizing sentiment toward specific groups (Gibson, 2013). Questions partially shift the focus from groups and instead query respondents about curbing civil liberties of all citizens including the right to demonstrate in military-style uniforms (Gibson and Bingham, 1985) and privacy of e-mail communication (Davis, 2007). Although less prevalent, these studies reveal some support for restrictions on civil liberties. One applies, for instance, to support for laws regarding the rights of neo-Nazis marching through Skokie, Illinois (Gibson and Bingham, 1985). Others reveal that respondents were willing to consider suspending some civil liberties in order to have greater personal safety and security when faced with the threat of terrorism in the aftermath of the terrorist attack of 9/11 in the United States (Davis and Silver, 2004). Combined, results are to some degree also contextually specific in that they are tied to particular policies.

Research on Political Tolerance: Single Country, Comparative Small and Large N Studies

Social scientists have tracked political tolerance for more than 60 years, attesting to its importance for scholarship on democracy and democratic values. Pointedly, although growing numbers of scholars and studies attest to the importance of tolerance, research reveals that not only is it a complex construct that is difficult to measure, as described in the previous section, but that it can be realized differently depending on a variety of factors like national and/or cultural contexts in addition to characteristics of individuals.

Many studies of tolerance are country-specific. For quite some time, these country-specific studies tended to focus on established democratic nations like the United States (Gibson, 2008; Mondak and Sanders, 2003; Schwadel and Garneau, 2014; Stouffer, 1955; Sullivan et al., 1979, 1981, 1982), Great Britain (Barnum and Sullivan, 1989), Canada (Sniderman et al., 1991, 1996), West Germany (Weil, 1989, 1991), Italy (Sniderman et al., 2000), and Israel (Shamir, 1991; Shamir and Sagiv-Schifter, 2006; Shamir and Sullivan, 1983). Studies of political tolerance have broadened in the last three decades to encompass newly democratizing countries including Russia (Bahry et al., 1997; Gibson, 1992, 1996, 1998; Duch and Gibson, 1992; Gibson and Duch, 1993), Poland (Golebiowska, 2009, 2014; Karpov, 1999a, 1999b), Estonia (Kaplan, 1995), and South Africa (Gibson and Gouws, 2000, 2003).

Notwithstanding recent growth, comparative studies of tolerance are less frequent in the literature, and typically focus on countries with similar political systems. Studies have explored tolerance in established democracies, comparing, for instance, the United States, New Zealand, and Israel (Sullivan et al., 1982), the United States and Israel (Shamir and Sullivan, 1983), politicians in New Zealand, Israel, Britain, and the United States (Sullivan et al., 1993), 12 and 17 Western European countries (Duch and Gibson, 1992; Weldon, 2006), and 13 former communist countries that shared a political legacy until the early 1990s (Guerin et al., 2004). Studies that explore tolerance in established and newer democracies at the same time, however, are less prevalent (cf Karpov, 1999b; Rohrschneider, 1999). A recently expanding set of works using the 1995–97 WVS attempt comparisons across a substantial number of countries (Guerin et al., 2004; Marquart-Pyatt and Paxton, 2007; Peffley and Rohrschneider, 2003). These vary, however, in the number of countries explored and range from 12 to 17 nation-states. Large N, cross-national scholarship on political tolerance that includes more than a handful of countries has recently expanded.

The majority of work on political tolerance is cross-sectional, which only provides a snapshot of tolerance that may be either group-specific or more diffuse. A few works provide a comparative or longitudinal view for the American public (Gibson, 2008, 2013; Schwadel and Garneau, 2014; Shafer and Shaw, 2009). Regarding building a comparative understanding of political tolerance across nation-states, these recent large N, comparative works are limited by large-scale data-gathering initiatives collecting indicators of tolerance across countries. Of vital importance is considering how tolerance is measured, whether it is fixed-group or content-controlled, and the extent to which individual and national or contextual factors are integrated into these works. Further, as described in the next section, while some similarities can be identified in terms of individual-level explanations of tolerance, these studies differ broadly in conceptual focus and empirical models. This variability has consequences in terms of knowledge accumulation, making

comparisons across studies difficult and to some degree hampering our understanding of political tolerance comparatively and cross-nationally. For instance, it remains possible that the levels and sources of political tolerance differ across countries and regions. These differences may be driven by contextual factors like democratic political histories and experience with the translation of democratic principles into democratic practice through internalizing democratic norms.

Major Explanatory Frameworks and Predictors Investigated

Studies of tolerance and the conceptual frameworks guiding them can be divided into those exploring individual-level theories and those examining contextual arguments related to the expression of tolerant judgments. These categories can be further subdivided into groupings of important predictors at the individual level like psychological dispositions, political orientation, and demographic characteristics and those at the contextual level including democratic learning and territorial threat.

A core set of individual-level variables typically included in studies of political tolerance include threat perception, democratic norms and activism, and demographic characteristics (Sullivan and Hendriks, 2009; Sullivan and Transue, 1999). Models of individual political tolerance often include measures of threat perception and/or fear (Duch and Gibson, 1992; Marcus et al., 1995; Sullivan et al., 1982). Disliked groups can be perceived by others as threatening to their core political values, to their social position, or to the constitutional order. In theory, threatened individuals should be more willing to put aside abstract principles of tolerance to reduce their immediate feelings of threat. Some research on political tolerance has focused on predispositions to feeling threatened (see Sullivan and Transue, 1999: 632), where some individuals are seen as psychologically more prone to feelings of threat. Other research asks about feelings of threat from the disliked group directly (Duch and Gibson, 1992; Sullivan et al., 1985).

Previous research has also stressed that political attitudes should affect political tolerance and that they may do so in more than one way (Bahry et al., 1997; Duch and Gibson, 1992; Sullivan et al., 1982, 1985). A general interest in politics might lead to an increased desire to engage complex political topics like tolerance. Also, internalization of abstract beliefs about democracy should enhance an individual's willingness to apply those norms to particular situations (Gibson, 1992; Lawrence, 1976; Marcus et al., 1995; Sullivan et al., 1982). Political ideology or left/right placement on the political spectrum may also affect tolerance. The findings on political ideology are mixed, however, linked with whether a group-specific or content-controlled measure was used (Duch and Gibson, 1992; Sniderman et al., 1989; Stouffer, 1955; Sullivan et al., 1982). In addition, individuals who directly use civil liberties may be more tolerant. That is, individuals who display 'democratic activism' by making use of their freedom to associate, right to free speech, or freedom to protest, practice their civil liberties and should therefore move beyond tolerance in principle to true increased levels of tolerance (Peffley and Rohrschneider, 2003; Guerin et al., 2004). Finally, demographic characteristics like age, gender, and religion have also been linked with tolerance (Stouffer, 1955; Sullivan et al., 1982, 1985).

These individual-level explanations of tolerance share an emphasis on learning tolerance. Individuals can acquire political tolerance much as they acquire other democratic norms – through contact with democratic institutions. That is, exposure to institutions in a democratic society increases exposure to the dominant democratic norms, which encourages political tolerance. As an example, the effect of education on political tolerance is an example of a learning effect. Individual education increases familiarity with diverse ideas and with diverse people. At the same time, educational institutions in a democratic society socialize individuals to democratic norms (Weil, 1991). Stouffer's (1955) seminal work on tolerance identified education as a crucial predictor of tolerance, and the effect of education has been demonstrated in many subsequent studies. Generally, increased levels of education should lead to increased levels of tolerance.

The recent availability of opinion data for more than a handful of countries has led to renewed focus on the democratic learning perspective. Moving beyond individual-level explanations, studies emphasize democratic institutions and how similarities and/or differences in the levels and sources of political tolerance may be best understood. For instance, citizens living in established democratic societies may have more opportunities to have contact with democratic institutions or engage in the 'rough-and-tumble' of democratic politics (Peffley and Rohrschneider, 2003). The democratic learning hypothesis would suggest that lower levels of tolerance in Eastern Europe result from these countries' lack of experience with democracy, such that citizens may not yet have learned how to translate general democratic values into applied tolerance. The longevity of democracy in a nation constitutes an important contextual feature. For instance, as noted earlier, three recent studies use the WVS. These studies revealed lower levels of political tolerance to be widespread in postcommunist countries in Eastern Europe (Guerin et al., 2004; Marquart-Pyatt and Paxton, 2007), and that factors typically associated with tolerance in the United States replicated weakly in other countries, notably so in Eastern Europe (Guerin et al., 2004; Marquart-Pyatt and Paxton, 2007; Peffley and Rohrschneider, 2003).

Recent work has identified important themes comprising emerging areas of interest worth noting including how widespread intolerance is within established and nascent democracies, how tolerance might shift over time and/or across groups, and the continuing relevance of the measurement of tolerance for comparative and cross-national research. Although empirical investigations are constrained by the availability and quality of existing data, prospects for future research on political tolerance are promising. A growing number of countries and questions tapping into democratic values like tolerance have periodically been added to existing data-gathering initiatives like the Eurobarometer Surveys, The European Values Survey, GSS, the International Social Survey

Program, and the WVS, making new comparisons possible. In addition, as knowledge is gained about the possibility of contextual and individual explanatory factors being salient, analytical techniques being applied to this stream of research are becoming increasingly sophisticated. For instance, multilevel or hierarchical linear modeling offers numerous advantages to cross-national, comparative research on a range of important topics, including tolerance.

Concluding Thoughts for Future Research

Recent scholarship lays a promising yet challenging foundation on which subsequent studies might build. These thematic areas include variability among established and newer democracies, multiple ways of measuring political tolerance and intolerance, and the question of the extent to which least-liked groups may have changed over time within a single country. As was described in the previous section, future work can potentially take on at least two possible forms: (1) comparing within established democratic countries over time using fixed-group and content-controlled group measures and (2) comparing across democratic regimes and time to advance our knowledge about the level and sources of political tolerance in comparative and cross-national context. Given existing data infrastructures, future research is likely to continue along at least the following three lines. First, single-country or case studies that include a significant temporal element are increasingly probable (Schwadel and Garneau, 2014). Second, the number of studies comparing across countries is also likely to grow. A third, possible emerging area is scholarly work that embarks on comparisons across datasets (i.e., of more than one cross-national secondary dataset), offering a promising addition to the existing literature on tolerance in the American public (Gibson, 2013). Although challenges remain for political tolerance worldwide, the implications of intolerance and its potential for restricting civil liberties are important for established and unconsolidated democracies.

See also: Ethnocentrism and Prejudice: History of the Concepts; Prejudice and Discrimination; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Race and Racism in the Twenty-First Century; Racism, History of; Racism, Sociology of; Racism: Social Psychological Perspectives; Religiosity and Prejudice against Minorities; Values, Psychology of; Values, Sociology of.

Bibliography

Almond, G.A., Verba, S., 1963. The Civic Culture: Political Attitudes and Democracy in Five Nations. Princeton University Press, Princeton, NJ.

Bahry, Donna, Boaz, Cynthia, Gordon, Stacey Burnet, 1997. Tolerance, transition, and support for civil liberties in Russia. Comparative Political Studies 30, 484–510.

Barnum, David G., Sullivan, John L., 1989. Attitudinal tolerance and political Freedom in Britain. British Journal of Political Science 19, 136–146.

Davis, Darren W., 2007. Negative Liberty: Public Opinion and the Terrorist Attacks on America. Russell Sage Foundation, New York.

Davis, Darren W., Silver, Brian D., 2004. Civil liberties vs security: public opinion in the context of the terrorist attacks on America. American Journal of Political Science. 48, 28–46.

Duch, Raymond M., Gibson, James L., 1992. 'Putting up with' fascists in Western Europe: a comparative, cross-level analysis of political tolerance. The Western Political Quarterly 45, 237–273.

Gibson, J.L., 2008. Intolerance and political repression in the United States: a half century after McCarthyism. American Journal of Political Science 52, 96–108.

Gibson, James L., 1992. The political consequences of intolerance: cultural conformity and political freedom. The American Political Science Review 86, 338–356.

Gibson, James L., 1998. A sober second thought: an experiment in persuading Russians to tolerate. American Journal of Political Science 42 (3), 819–850.

Gibson, James L., 2008. Intolerance and political repression in the United States: a half century after McCarthyism. American Journal of Political Science 52, 96–108.

Gibson, James L., 2013. Measuring political tolerance and general support for pro-civil liberties policies: notes, evidence, and cautions. Public Opinion Quarterly 77, 45–63.

Gibson, James L., Bingham, Richard D., 1985. Civil Liberties and Nazis: The Skokie Free Speech Controversy. Praeger, New York.

Gibson, James L., Duch, Raymond M., 1993. Political tolerance in the USSR: the distribution and etiology of mass opinion. Comparative Political Studies 26 (3), 286–329.

Gibson, James L., Gouws, Amanda, 2000. Social identities and political intolerance: linkages within the South African mass public. American Journal of Political Science 44, 278–292.

Gibson, James L., Gouws, Amanda, 2003. Overcoming Intolerance in South Africa: Experiments in Democratic Persuasion. Cambridge University Press, New York.

Golebiowska, Ewa, 2009. Ethnic and religious tolerance in Poland. East European Politics and Societies 23, 371–391.

Golebiowska, Ewa, 2014. The Many Faces of Tolerance: Attitudes toward Diversity in Poland. Routledge, New York.

Guerin, Daniel, Petry, Francois, Crete, Jean, 2004. Tolerance, protest, and democratic transition: survey evidence from 13 post-communist countries. European Journal of Political Research 43, 371–395.

Hutchison, Marc L., 2014. Tolerating threat? The independent effects of civil conflict on domestic political tolerance. Journal of Conflict Resolution 58 (5), 796–824.

Hutchison, Marc L., Gibler, Douglas M., 2007. Political tolerance and territorial threat: a cross-national study. Journal of Politics 69 (1), 128–142.

Inglehart, R., et al., 2000. World Values Surveys and European Values Surveys, 1981–1984, 1990–1993, and 1995–1997 [Computer file]. ICPSR version. Institute for Social Research [producer], 1999. Ann Arbor, MI, Ann Arbor, MI. Inter-university Consortium for Political and Social Research [distributor].

Kaplan, Cynthia S., 1995. Political culture in Estonia: the impact of two traditions on political development. In: Political Culture and Civil Society in Russia and the New States of Eurasia. M.E. Sharpe, Armonk, NY, pp. 227–269.

Karpov, Vyacheslav, 1999a. Political tolerance in Poland and the United States. Social Forces 77 (4), 1525–1549.

Karpov, Vyacheslav, 1999b. Religiosity and political tolerance in Poland. Sociology of Religion 60 (4), 387–402.

Lawrence, David G., 1976. Procedural norms and tolerance: a reassessment. American Political Science Review 70 (1), 80–100.

Marcus, Georges E., Sullivan, John L., Theiss-Morse, Elizabeth, Wood, Sandra L., 1995. With Malice toward Some: How People Make Civil Liberties Judgments. Cambridge University Press, Cambridge.

Marquart-Pyatt, Sandra T., Paxton, Pamela, 2007. In principle and in practice: learning political tolerance in Eastern and Western Europe. Political Behavior 29 (1), 89–113.

McClosky, Herbert, 1964. Consensus and ideology in American politics. American Political Science Review 58, 361–382.

McClosky, Herbert, Brill, Alida, 1983. Dimensions of Tolerance: What Americans Believe about Civil Liberties. Russell Sage Foundation, New York.

Mondak, Jeffrey, Sanders, Mitchell, 2003. Tolerance and intolerance, 1976–1998. American Journal of Political Science 47 (3), 492–502.

Nunn, C.Z., Crockett Jr., H.J., Williams Jr., J.A., 1978. Tolerance for Non-Conformity. Jossey-Bass Publishers, Washington.

Peffley, Mark, Rohrschneider, Robert, 2003. Democratization and political tolerance in seventeen countries: a multi-level model of democratic learning. Political Research Quarterly 56, 243–257.

Rohrschneider, Robert, 1999. Learning Democracy. Oxford University Press, New York.

Schwadel, Philip, Garneau, Christopher R.H., 2014. An age-period-cohort analysis of political tolerance in the United States. The Sociological Quarterly 55, 421–452.

Shafer, Chelsea E., Shaw, Greg M., 2009. The polls-trends: tolerance in the United States. Public Opinion Quarterly 73 (2), 404–431.

Shamir, Michal, 1991. Political intolerance among masses and elites in Israel: a reevaluation of the elitist theory of democracy. The Journal of Politics 53 (4), 1018–1043.

Shamir, Michal, Sullivan, John, 1983. The political context of tolerance: the United States and Israel. The American Political Science Review 77, 911–928.

Shamir, Michael, Sagiv-Schifter, Tammy, 2006. Conflict, identity, and tolerance: Israel in the Al-Aqsa Intifada. Political Psychology 27 (4), 569–595.

Smith, Tom W., Marsden, Peter V., Hout, Michael, 2013. General Social Surveys, 1972–2012: Cumulative Codebook. NORC, Chicago.

Sniderman, Paul A., Tetlock, Philip E., Glaser, James M., Green, Donald Philip, Hout, Michael, 1989. Principled tolerance and the American mass public. British Journal of Political Science 19, 25–45.

Sniderman, Paul M., Fletcher, Joseph F., Russell, Peter H., Tetlock, Philip E., Gaines, Brian J., 1991. The fallacy of democratic elitism: elite competition and commitment to civil liberties. British Journal of Political Science 21 (3), 349–370.

Sniderman, Paul M., Fletcher, Joseph F., Russell, Peter H., Tetlock, Philip E., 1996. The Clash of Rights: Liberty, Equality, and Legitimacy in Pluralist Democracy. Yale University Press, New Haven, CT.

Sniderman, Paul, Peri, Pierangelo, Rui, J., de Figueiredo Jr., P., Piazza, Thomas, 2000. The Outsider: Prejudice and Politics in Italy. Princeton University Press, Princeton, NJ.

Stouffer, Samuel A., 1955. Communism, Conformity and Civil Liberties: A Cross-Section of the Nation Speaks Its Mind. Doubleday, Garden City, NY.

Sullivan, John L., Hendriks, Henriet, 2009. Public support for civil liberties pre- and post-9/11. Annual Review in Law and Social Science 5, 375–391.

Sullivan, John L., Transue, J.E., 1999. The psychological underpinnings of democracy: a selective review of research on political tolerance, interpersonal trust, and social capital. Annual Review of Psychology 50, 625–650.

Sullivan, John L., Piereson, James, Marcus, George E., 1979. An alternative conceptualization of political tolerance: illusory increases 1950s–1970s. American Political Science Review 73 (3), 781–794.

Sullivan, John L., Marcus, George E., Feldman, Stanley, Piereson, James E., 1981. The sources of political tolerance: a multivariate analysis. American Political Science Review 75 (1), 92–106.

Sullivan, John L., Piereson, James, Marcus, George E., 1982. Political Tolerance and American Democracy. University of Chicago Press, Chicago.

Sullivan, John L., Shamir, Michael, Walsh, Patrick, Roberts, Nigel S., 1985. Political Tolerance in Context: Support for Unpopular Minorities in Israel, New Zealand, and the United States. Westview Press, Boulder.

Sullivan, John, Patrick, Walsh, Michal, Shamir, David, G., Barnum, James Gibson, 1993. Why politicians are more tolerant: selective recruitment and socialization among political elites in Britain, Israel, New Zealand, and the United States. British Journal of Political Science 23, 51–76.

Weil, Frederick D., 1989. The sources and structure of legitimation in Western democracies: a consolidated model tested with time-series data in six countries since World War II. American Sociological Review 54, 682–706.

Weil, Frederick D., 1991. Structural determinants of political tolerance: regime change and the party system in West Germany since WWII. In: Research in Political Sociology, vol. 5. JAI Press, Greenwich, CT, pp. 299–332.

Weldon, Steven A., 2006. The institutional context of tolerance for ethnic minorities: a comparative, multilevel analysis of Western Europe. American Journal of Political Science 50 (2), 331–349.

Political Trials

Leora Bilsky, Tel Aviv University, Tel Aviv, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the theory and practice of political trials in liberal democracies as distinct from 'show trials.' It explores the dilemmas raised by such trials when they occur in transitions to democracy, within established democracies, and in international and transnational trials of atrocity, and explains scholars' call to recognize the inevitably political character of legal proceedings.

The term 'political trial' is popularly associated with 'show trials' like the Moscow Purge trials conducted by Stalin during the 1930s. Such trials represent an extreme example of political repression and are viewed as sham legal proceedings designed by authorities to dramatize specific political campaigns and/or to eliminate prominent individuals. In this context, the political trial is taken to be a 'show' in which politicians and not jurists pull the strings and determine the results of the trial in advance. This pejorative use of the term also helps designate regimes that resort to such techniques as the antithesis of liberal democracies. Yet, as Otto Kirchheimer observed, the most effective usage of political trials occurs precisely within liberal democracies, in societies committed to the rule of law. What can explain this gap between popular perceptions and the actual practice of political trials? Can we distinguish a political trial from a 'show trial'? Should we continue to view political trials as opposed to the very essence of liberal democracies, or should we accept them as part and parcel of this system, deserving a sustained theoretical elaboration?

In his seminal book *Political Justice*, Kirchheimer defines political trials as instances in which "court action is called upon to exert influence on the distribution of political power" (Kirchheimer, 1961: p. 49). This broad definition includes the use of the court by the ruling elites or by out-groups and dissenters to achieve political goals by judicial means. Many writers regard the central factor in such trials as the political motive behind them. The connecting thread between the trial of Socrates for corrupting the youth of Athens (399 BC), the trial of Jesus for blasphemy and sedition (AD 30), the trial of Joan of Arc (1431) for heresy and witchcraft, the trial of Thomas More (1534) for remaining silent when asked about Henry VIII's supremacy in religion, the trial of Galileo (1633) for heretically suggesting that the earth moved around the sun, and many other trials is that, in each case, men in power believed that the defendant was a threat to them (Becker, 1971; Belknap, 1981). For this reason the prototype of a political trial has been the criminal trial of a political adversary for political reasons. This broad definition encompasses trials in which the defendant has directly attacked the established order by acts such as treason or sedition, but also trials involving common crimes in which only the personality or the motive of the offender indicates their political nature. Other writers try to refine this definition by expanding its scope to legal proceedings other than criminal trials such as impeachment, congressional hearings, and truth commissions. It has also been suggested that the definition be extended to include trials involving larger group conflicts such as ethnic, race, or labor struggles. Some move the emphasis from the political motive of the authorities to the political identity of the defendant. Others conclude that the important factor that distinguishes such trials is not the motive but the competition of stories over the identity of the polity. Notwithstanding their differences, all writers on the subject agree that political trials are inevitable not only in authoritarian regimes but also in constitutional democracies, although this is generally denied, in particular by legal practitioners. Critical of such denials, these scholars set out to replace them with an honest recognition of the existence and function of political trials in both types of regimes.

Why should the term 'political trial' cause such anxiety and widespread denial? Because law tends to stabilize the status quo, it is not surprising that there are always people who feel justified in ignoring the law in order to bring about drastic and necessary changes. This is especially true for powerless groups that are prevented from participating effectively in important institutional decisions affecting their lives (Hakman, 1972; Fletcher, 1995). When the authorities put on trial those who challenge the basic value distribution in society, especially those who are willing to disregard existing laws to this end, the result is a political trial. Despite this regular occurrence of political trials within democratic societies, the term is widely resisted because of its connotation with 'show trials' in which either the charge is a 'frame-up' to conceal the political motivation of the prosecution, or the hearing itself fails to meet minimum standards of judicial impartiality and independence. For this reason, writers attempt to distinguish among different types of political trials. One classification distinguishes among politically motivated trials, politically determined trials, and trials with substantial political consequences (Friedman, 1970). Another differentiates between political 'trials' (abridgment of due process guarantees), the 'political' trial (camouflaged as a criminal trial), and the political trial (politically motivated) (Becker, 1971). These efforts at classification have proved frustrating because of the immense historical variations and the widespread existence of political trials in very different regimes. Writers have identified a variety of characteristics of political trials but none has proved to be a sufficient and necessary condition. This literature has largely been descriptive but has failed to assess the compatibility of political trials with liberal values and thus to overcome the negative connotations the term arouses. In order to address

this question, it is necessary to expose the hidden engine of political trials – their legitimizing function – and its relation to democracy.

The legal system offers a tempting opportunity for those in power to damage enemies, tarnish their image, and isolate them from potential allies by casting them as criminals. This temptation becomes especially strong in democracies that constrain the ruling elite from resorting to more direct devices of repression. What distinguishes political trials from other political devices is their legitimizing function, that is, their ability to turn a political adversary into a criminal defendant and thereby reinforce the legitimacy of those in power. However, this function can be served only if the public perceives the trial as providing the defendant with a 'fighting chance.' In constitutional democracies, this 'fighting chance' is generally assured by the relative autonomy of the judges and the constraining effects of the rule of law (Kirchheimer, 1961). Moreover, these guarantees introduce an element of uncertainty into the trial, and encourage political defendants to use it to expose the illegitimacy of the policy or value system advanced by the authorities. Indeed, from Socrates's trial (399 BC) to the Chicago Conspiracy trial (1969), there have always been defendants who have exploited the limited forum that the political trial provided them to articulate their oppositional views. Some have adopted a defiant stance, proclaiming their right to commit the crime (Solomon Tellirian – as a protest against the Armenian massacre by the Turks in 1921; Dr Jack Kevorkian – advocating euthanasia in 1999), or refusing to respect traditional courtroom decorum altogether (Bobby Seale – Black Panthers in 1969; Klaus Barbie – Nazism and Vichy France in 1987); others have played by the rules in order to win an acquittal that would confirm their claim that they were being prosecuted for their political beliefs (Dr Spock – Vietnam War in 1968). Insight into this legitimizing function and the element of uncertainty that it introduces into political trials in democratic regimes exposes the tension between politics and justice inherent in them. This has made political trials a fascinating area for research since the 1960s.

Political Trials in the Transition to Democracy

Political trials are most salient in times of transition between regimes, especially when a new democratic regime confronts crimes of the old regime. Indeed, the first attempts to seriously consider the compatibility of political trials with liberal democratic values appear in the literature on the transition to democracy. Unlike a military revolution that sustains its authority by brute force, democratic regimes are committed to the rule of law and are inclined to address the evils of the previous regime with the help of legal devices. However, the new regime's commitment to the rule of law also makes it aware of the dangers of using ex post facto laws and indulging in 'victor's justice.' At such times, the various expectations from the law – to punish the guilty, ascertain the truth about the old regime, and enhance reconciliation in society – seem to overwhelm the legal system and to push it in opposite directions. Moreover, the forward-looking direction of the architects of democracy who are concerned with the efficacy of the transition often conflicts with the backward-looking direction of legal proceedings and their narrow focus on individual guilt (Osiel, 1997; Nino, 1991; Teitel, 1997). For these reasons, trials of transition bring to the foreground the clash between politics and justice. Two main approaches to the problem have evolved since World War II: exemplary criminal trials and truth commissions, both of which can be considered political trials of sorts and have received extensive theoretical elaboration.

At the end of World War II, the Allies established an international military tribunal in Nuremberg (1945) to judge the leaders of the Nazi regime. This was the first time in history that the leaders of a defeated country had faced criminal prosecution for war crimes, crimes against the peace, and crimes against humanity by an international tribunal. The charge applied in the trial – conspiring to wage an aggressive war – allowed the prosecution to tell a general historical narrative. Could the obvious political end of using the court to teach a history lesson (through ex post facto laws, and victors' judges) be reconciled with the demands of liberalism? Confronted with this dilemma, Judith Shklar made the first serious effort to reconcile the legacy of liberalism with political trials occurring in the transition to democracy (Shklar, 1964). A legalistic understanding of liberalism, she argued, could justify the Nuremberg trials only at the high cost of denying their political nature altogether. In her view, procedural safeguards to defendants could only guarantee that the trial would not deteriorate into a show trial, but this in itself could not justify deviating from basic liberal demands such as an independent judiciary, established court, nonretroactive laws, and so forth. Paradoxically, Shklar came to the conclusion that the political dictate of founding a democracy justified the divergence from strict legalism in the Nuremberg trials. In other words, it was only an honest recognition of the political aspects of the trials, their educational and symbolic contribution to the building of democracy in postwar Germany, that could justify them. Later writers accepted Shklar's consequentialist approach to transitional justice but limited its applicability to trials that guarantee the due-process rights of the defendants (Osiel, 1997).

A very different attempt to address the crimes of the Nazi regime was made in the Eichmann trial (1961). This time it was not the international community but an Israeli court that undertook to judge one of the central figures in the massacre of the Jews. The political nature of the trial was evident, among other things, from its reliance on the particularistic law of 'crimes against the Jewish people.' Likewise the heavy reliance on victims' testimonies about their suffering (and through them the telling of the Jewish Holocaust) conflicted with the liberal demand that the trial concentrate on proving the deeds of the accused. Hannah Arendt and other writers criticized the Israeli prosecution for politicizing the trial in this way (Arendt, 1963; Segev, 1993; Lahav, 1992). Others see the decision as stemming from the general dilemma of political trials of how to strike a balance between the need to establish the truth about an oppressive regime, the need to give voice to the victims, and the need to ascertain the guilt of the defendants (Bilsky, 2004).

This dilemma is compounded when the former authoritarian regime retains enclaves of power and opposes any attempt to bring to justice the officials of the former regime. In South America, Argentina is one of the few nations to have

prosecuted military officers in public trials for their role in repression carried out under the former regime. The political nature of these trials was evident in the decision to prosecute only the masterminds and the worst offenders, and in the greater priority given to the value of learning about the evils of the authoritarian past than to prosecuting every individual who was involved. Again, the justification for such selective prosecution is given in consequentialist terms – the extent to which the trial serves to strengthen the emerging democracy (Nino, 1991). A central dilemma that emerges in transitional trials of this kind is the breakdown of common discourse, which undermines the legitimacy of the trial for some segments of society. Thus, the Argentinian court's decision to limit the evidence to proving the deeds of the accused, without allowing them to be situated within the wider historical narrative about an alleged war of self-defense against attacks on the regime by guerrilla forces, was deemed political and undermined the legitimacy of the trial in the eyes of the military (Osiel, 1997).

Alongside efforts to bring to trial leaders of former authoritarian regimes for gross violations of human rights developed the institution of truth commissions established by national or international organizations. Truth commissions reduce the tensions that arise in transitional trials by separating the various functions of learning the truth about the past, of ascertaining individual guilt, and of giving voice to victims. The Truth and Reconciliation Commission in South Africa, for example, was divided into a Committee on Human Rights Violations (victims' testimonies), a Committee on Amnesty (offenders' testimonies), and a Committee on Reparation and Rehabilitation. In each committee, different procedures were used (1995 Act). It further made the granting of individual amnesty conditional upon providing a full account of the offenses committed (Minow, 1998). Truth commissions have proved to be far more effective than court proceedings in furnishing a dramatic medium for theatricalizing the new official history, because the narratives of victims are rarely interrupted by lawyers and there are incentives for offenders to relate their part in the repression. Indeed, in several countries the commission's proceedings have been broadcast daily, and in others the final report has become a bestseller (Brazil, Argentina). Truth commissions present us with a kind of political trial in which the balance between politics and justice has been struck in ways more compatible with liberal concerns. However, the adoption of the discourse of healing by truth commissions has been sharply criticized for drawing attention away from the political context in which abuse is committed, as well as signaling that the commission of mass crimes does not carry serious consequences (Leebaw, 2011).

Political Trials in Established Democracies

The political nature of transitional justice stems from the fact that there are no overarching legal norms that are accepted as legitimate by the two successive regimes. Moreover, in these cases the trials have overwhelmingly been used to fulfill social and political functions other than ascertaining individual guilt. Liberal scholars, as we have seen, have been willing to recognize the political nature of such trials and to justify it, but have not extended their theories to trials in established democracies. The general view is that in normal times law should endeavor to keep politics out of the courtroom so that every political trial is perceived as a corruption of the rule of law (which may be termed the 'pathology thesis'). This distinction, however, fails to account for the striking similarities between transitional situations and periods in which there has been a significant increase in political trials in democratic societies.

In the United States, political trials have occurred whenever the status quo has been challenged, generally during periods of social and political ferment unleashed by such forces as war, economic conflict, or racial discord (Belknap, 1981). For example, the decade that began with the escalation of the Vietnam War in 1965 brought an epidemic of political trials culminating in the Chicago Conspiracy trial (1969–70). This case represents a microcosm of American political justice of the era, as it was directed against representatives of major antiwar groups, youth counterculture, and the Black Panther Party, who all believed that the war was illegal and sought to mobilize the public against it. Although the charges were clearly political (conspiracy to cross state lines with the intent to incite riot), the trial also demonstrated the ability of the defense to politicize the trial and turn it into a forum for social protest by flouting the norms of the courtroom and ridiculing the judge (Ely, 1981). In such trials as the Chicago Conspiracy trial, the same breakdown of common discourse was manifested as in political trials during transitions to democracy.

The social turmoil and political trials of the late 1960s influenced the development of radical theories of law (Gordon, 1982), as scholars in critical legal studies, critical race theory, and feminist legal studies began to question the plausibility of the liberal ideal of separating law from politics. For these scholars the very attempt to delimit the boundaries of the category 'political trials' obscures the way in which politics enters every trial and every field of law. They argue that the liberal reliance on a distinct category of political trials often serves to legitimize the status quo. Critical theory understands politics not in the narrow sense of the term (motive of authorities, identity of defendant, and so forth) but as the hegemonic ideology that shapes the interpretation of law while presenting it as neutral. The main contribution of critical writings to understanding political trials has been their rejection of the 'pathology thesis' about the relation between law and politics and their systematic efforts to uncover the ideological structures that shape different areas of law (Kelman, 1987; Kairys, 1982). This broad definition of political trials, however, had its drawbacks because it no longer called for a systematic investigation of the unique features of classic political trials, and shifted the attention of scholars from the court drama and social reception of the trial to the appellate court interpretation of the law.

Other scholars, still committed to the liberal framework, have identified a new form of political trial that emerged during the 1980s and 1990s. While in the old political trial the ruling authorities selected certain individuals to stand for an opposition that the state wanted to eliminate, in the new political trial a section of the public turns the trial into a political trial by identifying with the victims (who are not a formal party to the trial) or with the defendant. In these cases, the state does not intend a political trial and has very little control over its

politicization (Fletcher, 1995). The people mobilized around these trials are usually out-groups protesting their marginalization by the legal and political institutions. Examples are famous rape trials and self-defense trials of battered women who killed their abusive spouses in which women's groups identified with the victim and politicized the trial. The Rodney King trial and O.J. Simpson trial likewise became political trials in which African-American groups were mobilized to protest against white justice in America. The study of these trials draws particular attention to the role of the media in these cases.

The focus on group conflict and on the narrative and rhetorical aspects of political trials in ongoing democracies provides scholars with the key to understanding their dynamics from a pluralistic perspective. The famous trials of Socrates, Jesus, Dreyfus, and others are examples of heroes unjustly prosecuted, but they can also be viewed as major junctions in the life of the republic where society's conflicting values are played out not only, and not even mainly, through the learned interpretation of the law but through the human drama in and around the courtroom. In these trials, the social conflict is transformed into competing narratives that capture the public's attention and offer an opportunity for collective self-reflection. Through an examination of competing values and loyalties they bring together for public consideration society's basic contradictions (Christenson, 1999). From this perspective we can see how political trials, while threatening the rule of law, at the same time may contribute to the development of a more critical and democratic society.

Political Trials in the Transnational Legal Order

As we saw, although the Nuremberg trials were condemned as victors' justice, the Eichmann trial was criticized as victims' justice. Ironically, in the 1990s Nuremberg was invoked as the most important precedent for the establishment of ad hoc international criminal tribunals and a few years later the permanent International Criminal Court (ICC). The Eichmann trial became a precedent for the exercise of universal jurisdiction by domestic criminal courts. In both cases, the justification for turning to courts outside the national jurisdiction with closest links to the crimes was to end impunity, that is, to counter domestic politics' ability to obstruct the wheels of justice. This turn was accompanied by a growing reliance on a new legalism connected to the human rights discourse. At the center of these trials stood the new crimes of genocide and crimes against humanity, viewed as supranational crimes exploding spatiotemporal legal coordinates by overcoming prescriptive periods and territorial limits to criminal jurisdiction (Douglas, 2007).

Underlying these developments in international criminal law is an unarticulated assumption about political trials: if the danger of impunity came from too much proximity to domestic politics, a cure was to be found in distancing the tribunal from the influence of the political authorities. In other words, the turn to third party criminal courts and international tribunals was meant to ensure the triumph of the rule of law over political considerations. However, far from ending the debate about political trials, new criticism was raised by famous defendants, such as former Chilean dictator General Augusto Pinochet, former Yugoslav president Slobodan Milosevicz, and former Israeli prime minister Ariel Sharon, who claimed to be the victims of a new form of show trial. International criminal law thus poses the following puzzle: are international tribunals and third party criminal courts the source of danger of the politicization of international law, or are they the best guarantee against 'political justice,' piercing the shield of sovereign immunity?

I explore this question by examining the new political trial in its three contemporary manifestations: third party domestic courts exercising universal criminal jurisdiction, international criminal tribunals, and transnational tort actions brought by victims in third party civil courts.

Universal Jurisdiction by Domestic Criminal Courts

If the problem of political trials was attributed to too much proximity to the political authorities, can too much distance produce new dangers of politicization? During the 1990s, frustrated groups of victims discovered the tool of universal jurisdiction, first recognized by the Eichmann court, which had remained dormant throughout the Cold War, as a way to seek justice in foreign courts. Prosecutors further relied on legal doctrines such as crimes against humanity and genocide, which have been interpreted by some national courts as not being subject to statutes of limitation. However, this development created new dangers of politicization. The removal of territorial and temporal limitations on prosecution makes the legal process vulnerable to political opportunism in the timing of the prosecution. For example, Ariel Sharon faced the possibility of trial in Belgian courts in connection with the 1982 Sabra and Shatila massacre in Lebanon, following complaints brought only after he was elected prime minister of Israel, 20 years after the massacre. Ordinary prescriptive periods force the prosecution to bring charges as soon as possible and to concentrate on the strength of the legal case, thus precluding political considerations of this kind. Yet understanding the significance of crimes of such magnitude may require the passage of long periods of time and the changing of generations, lending support to the relaxation of procedural constraints on prosecution.

Another sense of politicization that undermines more profoundly the integrity of the criminal process stems from the disjunction that universal jurisdiction creates between the judge and the relevant political community. This disjunction operates at two levels. First, critics worry that allowing foreign judges to adjudicate cases related to societies to whom they are largely unaccountable threatens the democratic foundations of criminal law (Orentlicher, 2004). Second, since the courts exercising universal jurisdiction operate at a remote distance from the communities that are deeply affected by their judgment, the solidarity between judge and community is undermined, thus weakening in the eyes of the community the legitimacy of the judges who do not take the risk of living under the results of their judgments. This problem is exacerbated in cases in which the judge fails to appreciate the intricacies of the cultural, political, and historical context of the case. We are used to thinking about the rule of law in terms of maintaining a distance between the court and the parties to the dispute, but at the same time the rule of law requires that a meaningful link be established between the

court and the community injured by the deed. It seems that underlying this tension is a basic assumption that the criminal law belongs to a political community and is an exercise in self-governance.

In the eyes of supporters of universal jurisdiction, the dangers of unaccountable sovereign power seem to outweigh these new dangers of politicization. That position, however, takes for granted that the two alternatives are exhaustive. A middle ground can be found in the requirement of a meaningful link to the court, recently added by legislators and courts, in response to the political uses of universal jurisdiction made during the 1990s (Bilsky, 2010). Similarly, the Princeton Principles on Universal Jurisdiction, a set of nonbinding guidelines drafted by jurists in 2001, recommend that in case of competing national jurisdictions, weight be given to the place of the commission of the crime and the nationalities of the alleged perpetrator and victims, among other factors (Principle 8 of the Princeton Principles on Universal Jurisdiction).

Criminal Trials in International Criminal Courts

International criminal tribunals appear to enjoy more legitimacy than domestic criminal courts exercising universal jurisdiction, as they are the product of consent (where the country is a party to the treaty establishing the ICC, or the case is self-referred) or the will of the international community (where the case was referred by the U.N. Security Council). Yet the distance between judge and community raises some of the same problems as universal criminal jurisdiction, and therefore also calls for ways to reconnect the relevant community to the trial. We can view the principle of complementarity embedded in the Rome Statute establishing the ICC as aspiring to find a middle ground between universality and territoriality. By providing that a case is inadmissible before the ICC if it is being investigated or prosecuted by a state having jurisdiction over it, this gives priority to national courts in trying the case (Holmes, 1999). When there is such an investigation or prosecution, the ICC can take up the case only when the state "is unwilling or unable genuinely to carry out the investigation or prosecution" (art. 17 of the Rome Statute). This mechanism provides strong incentives to states to prosecute crimes they would have preferred to ignore, nevertheless leaving the door open for the ICC to intervene in exceptional cases. The principle of complementarity therefore appears to address the problem of sovereign impunity while limiting the dangers of politicization that arise from the distance between the trial and the concerned community. However, a recent study of compliance and resistance to the Inter-American Court among domestic judges suggests that a referral system such as the one created by the European Court of Justice might lead to greater compliance (Huneeus, 2011). Another middle ground is achieved in 'mixed' or 'hybrid' tribunals, staffed by both national and foreign judges and prosecutors, and likewise blending international and national criminal procedural law (Shraga, 2009).

Is the ICC immune, then, to political uses? The U.N. Security Council's significant role in the establishment of international criminal tribunals (the International Tribunal for the Former Yugoslavia, and the International Criminal Tribunal for Rwanda) and its power to make referrals to the ICC certainly provide openings for political considerations in the choice of situations brought to justice. Critics point to the fact that, to date, the ICC has only investigated crimes committed by Africans. But it is not only powerful states that can draw political advantage from these trials. Domestic authorities too can use ICC proceedings for political purposes, even though they do not control the decision to prosecute (this decision is taken by the ICC's prosecution office). If Kirchheimer's classic definition of the political trial is the use of legal means for political ends, turning political adversaries into criminal defendants and thereby reinforcing the legitimacy of those in power, recent research on the local impact of ICC prosecutions related to Sudan and Uganda reveals the same dynamic (Nouwen and Werner, 2012). According to that study, the Ugandan government referred rebel movements to the ICC as part of a military strategy, in the hope that the ICC could capture the rebel leaders. Moreover, the government successfully used the proceedings as part of an international reputation campaign in which rebels came to be branded as internationally wanted 'criminals' while the Ugandan government could portray itself as a friend of humanity. Although referral carried the risk identified by Kirchheimer that Ugandan government officials themselves be prosecuted, the government's reputation campaign succeeded in silencing criticism of its actions in the Congolese territory, which the ICC prosecutor has not investigated (Nouwen and Werner, 2012).

Thus the question is not whether international criminal trials are political, but what are the characteristics of the new political trial that emerges from the ICC. Nouwen and Werner have pointed to the fact that because the court is international and not controlled by domestic authorities, it provides new possibilities for political opponents to initiate the trial through referrals. Yet in their view this heightened risk to the government is balanced by the fact that international officials, not victims, control the proceedings, together with the Court's dependence to a large degree on the cooperation of the relevant national government for numerous procedures, including the enforcement of arrest warrants and the issuance of visas to ICC staff. The dependence on state cooperation may also lead ICC staff to play into the friend/enemy of mankind dichotomy used by national governments. This institutional interest in being in good relations with the authorities is exacerbated by the ICC's concern with establishing its own legitimacy as an international court. The fact that contrary to the tribunals established at Nuremberg, the ICC can adjudicate atrocities even as an armed conflict is ongoing further raises the political stakes of the ICC's interventions. These political aspects specific to international tribunals are obscured by the universalist vocabulary of crimes against humanity, which itself might intensify political struggles. In other words, although the dynamics of international political trials are analogous to those in the domestic sphere, the stakes are much higher as the ruling of an international court brands the defendant as an enemy of humanity, leaving little space for political contestation or deliberation.

The distance of the ICC from the relevant political community enables it, on the other hand, to attend to victim groups that are often excluded or persecuted by the political authorities. Thus, the Rome Statute has enshrined victims' rights to both protection and participation in the ICC's

proceedings, through a range of procedural rules and mechanisms, and the creation of a special Victim and Witness Unit charged with coordinating protection measures. However, victim participation does not amount to control of the case, which remains in the hands of the prosecution. Moreover, because the ICC's legitimacy depends in part on the claim that it is acting on behalf of victims, proceedings tend to emphasize victims' innocence, obviating any discussion of 'imperfect' victims who collaborated with the regime. This leads to criticism that the ICC remains legalistic and fails to address the gray zone between perpetrators and victims that characterizes collective crimes (Leebaw, 2011).

Transnational Tort Human Rights Litigation in Domestic Courts

We have seen that much of the denial of the political nature of trials and the reemergence of legalism is connected to the form of the criminal trial. Can civil proceedings allow for a better legal treatment of mass violence? Can they end impunity by creating a direct channel for victims to bring charges against state officials without the filter of a public prosecutor? Since the 1980s, victims of grave human rights abuses have invoked the Alien Tort Statute of 1789, which grants US federal courts "jurisdiction of any civil action by an alien for a tort only, committed in violation of the law of nations or a treaty of the United States" (28 U.S.C. §1350). The US courts have generally interpreted this statute as enabling them to exercise a form of universal civil jurisdiction, by opening the door to cases unrelated to the United States. These claims filed by individuals as well as classes of victims have targeted both low-level state officials and prominent politicians (such as former Filipino dictator Ferdinand Marcos), sometimes in parallel to international criminal prosecutions. Thus, former Bosnian Serbian leader Radovan Karadzic was sued under the Alien Tort Statute and later indicted by the International Criminal Tribunal for the Former Yugoslavia, in both cases for war crimes committed in Bosnia. This type of litigation has also redirected attention to a new sort of defendant: the multinational corporation, accused of supporting or enabling atrocities committed by state authorities. Alien Tort Statute litigation therefore holds the promise of greatly reducing impunity, by not only distancing the legal process from the relevant political community as in universal criminal jurisdiction, but also by avoiding reliance on a public authority to bring a claim and targeting not only individuals but also the corporations that enable mass violence.

However, privatizing prosecution by granting full control of the litigation to victims creates new types of politicization. Organized groups of victims can dominate the proceedings at the expense of other victims. For example, in restitution cases brought against Swiss and German corporations for participation in the Nazi crimes, Jewish organizations tended to predominate over other victim groups (Adler and Zumbansen, 2002). A different problem arises from the lack of screening by a public prosecutor, which shifts to the court all political discretion in allowing the case to proceed. Although this has led many to condemn Alien Tort Statute litigation as court interference in foreign policy, others have countered that various legal doctrines reduce the possibility of a clash with US foreign policy, and that in any event the strict conditions imposed by the US Supreme Court means that in practice very few cases succeed. Nevertheless, the very filing of a lawsuit by victims might paralyze the political process in cases of transition to democracy. Furthermore, damage awards to victims might conflict with the distributive plans of transitional governments. One such example is the class action against Marcos, where a US jury granted a class of 10 000 victims of torture damages in excess of 2 billion dollars. Although legally a success, the victims have collected very little of the damage award, in part because of conflicting claims to the Marcos's fortune by the democratic government of the Philippines. Although some writers argue that the real contribution of the civil lawsuit is symbolic rather than material, even symbolic victories have political implications, potentially empowering victims but also enabling some victims to the exclusion of others to affect the domestic political agenda.

Unique to civil litigation is the possibility of settlement. The settlement of Holocaust-related claims by groups of victims against European corporations was criticized for undermining the normative functions of the court as well as the didactic functions of clarifying history and shaping collective memory (Marrus, 2009). Moreover, observers unaccustomed to American class action practice perceived the negotiation process and the monetary settlements that followed as victories of political pressure over the rule of law (Marrus, 2009). Others counter that it is precisely the American class action that gave victims the necessary leverage against giant corporations that had enjoyed *de facto* immunity for decades due to their political lobbying in Europe (Bilsky, 2012a). Moreover, settlement provided the flexibility required for dealing with complex political issues and the admission of moral responsibility in situations of collaboration with authoritarian regimes. It also released the judge from the controversial role of creating official history by incentivizing private corporations to investigate their past and publish new histories (Bilsky, 2012b). Yet this new power of corporations to create soft law and commission history raises new concerns about the political use of legal proceedings by corporate actors.

The possibility of suing individuals and corporations under the Alien Tort Statute appears to have been seriously restricted by the 2013 decision of the Supreme Court in *Kiobel v. Royal Dutch Petroleum*, in which it was held that the statute should not be interpreted to apply to conduct outside the United States. Nevertheless, it is now apparent that civil litigation is crucial for addressing the involvement of larger sections of society, including business entities, in collective crimes, and the creation of a globalized class action might be considered in parallel to the ICC.

Conclusion: Is the Political Trial Still a Relevant Category?

If we accept – following Shklar, Douglas, Bilsky, Nouwen and Werner, and others – that trials, whether domestic, international, transnational, criminal, or civil, inevitably have political consequences, the question of whether a trial is political obscures more than it reveals. Instead, we should ask what sort of politics each mechanism promotes, and what kind of procedures and legal doctrines are available to the court to

balance the politics while fighting niches of immunity. The question, in short, is not, "Is law political?" but "What sort of politics can law maintain and reflect?" (Shklar, 1964: p. 144). It seems that trials addressing collective violence inevitably have substantial collective significance, and attention should therefore be turned to finding ways to harness the law so as to foster participatory democracy. Enhancing the participatory character of legal proceedings is particularly pressing at a time when the nation-state is weakened and new transnational political actors carve for themselves areas of power without corresponding responsibilities.

See also: Globalization: Legal Aspects; Justice and Law; Law, Mobilization of; Political Lawyering; War Crimes Tribunals, Crimes against Humanity and Transitional Justice.

Bibliography

Adler, Libby, Zumbansen, Peer, 2002. The forgetfulness of noblesse: a critique of the German foundation law compensating slave and forced laborers of the Third Reich. Harvard Journal on Legislation 39 (1), 1–61.

Arendt, Hannah, 1963, 1964. Eichmann in Jerusalem, A Report on the Banality of Evil. Penguin Books, New York.

Barrett, Wilfred Phillips, 1931. The Trial of Jeanne d'Arc. George Routledge & Sons, London.

Becker, Theodore Lewis (Ed.), 1971. Political Trials. Bobbs-Merrill, New York.

Belknap, Michal R., 1981. American Political Trials. Greenwood Press, Westport, CT.

Bilsky, Leora, 2001. Between justice and politics: the competition of storytellers in the Eichmann trial. In: Aschheim, Steven (Ed.), Hannah Arendt in Jerusalem. University of California Press, Berkeley.

Bilsky, Leora, 2004. Transformative Justice: Israeli Identity on Trial. University of Michigan Press, Ann Arbor.

Bilsky, Leora, 2010. The Eichmann trial and the legacy of jurisdiction. In: Benhabib, Seyla (Ed.), Politics in Dark Times: Encounters with Hannah Arendt. Cambridge University Press, New York, pp. 198–218.

Bilsky, Leora, 2012a. Transnational Holocaust litigation. European Journal of International Law 23 (2), 349–375.

Bilsky, Leora, 2012b. The judge and the historian: transnational Holocaust litigation as a new model. History and Memory 24 (2), 117–156.

Boyle, James (Ed.), 1994. Critical Legal Studies. New York University Press, New York.

Christenson, Ron, 1999. Political Trials, Gordian Knots in the Law. Transaction Publishers, New Brunswick, Canada.

Conot, Robert, 1983. Justice at Nuremberg. Harper & Row, New York.

Cover, Robert M., 1983. The Supreme court, 1982 term—foreword: nomos and narrative. Harvard Law Review 97, 4–70.

De Brito, Alexandra Barahona, 1997. Human Rights and Democratization in Latin America: Uruguay and Chile. Oxford University Press, Oxford, UK.

Dorsen, Norman, Friedman, Leon (Eds.), 1974. Disorder in the Court, Report of the Association of the Bar of City of New York, Special Committee on Courtroom Conduct. Pantheon Books, New York.

Douglas, Lawrence, unpublished. Crimes of Atrocity, the Problem of Punishment and the Situ of Law, unpublished manuscript.

Douglas, Lawrence, 2007. Shattering Nuremberg, toward a jurisprudence of atrocity. Harvard International Review.

Du Plessis, Lourens, 1994. Amnesty and transition in South Africa. In: Boraine, Alex, Levy, Janet, Scheffer, Ronel (Eds.), Dealing with the Past: Truth and Reconciliation in South Africa. IDASA, Capetown, South Africa.

Duff, Antony, 2001. Punishment, Communication, and Community. Oxford University Press, Oxford.

Ely, J.W., 1981. The Chicago conspiracy case. In: Belknap, Michal R. (Ed.), American Political Trials. Greenwood Press, Westort, CT.

Fletcher, G.F., 1995. With Justice for Some, Victim's Rights in Criminal Trials. Addison-Wesley, Reading, UK.

Friedman, Leon, 1970. Political power and legal legitimacy: a short history of political trials. Antioch Review 30, 157–171.

Gagne, Patricia, 1998. Battered Women's Justice. Twayne, New York.

Gordon, Robert W., 1982. Some critical theories of law and their critics. In: Kairys (Ed.), The Politics of Law. Basics Books, New York.

Hakman, Nathan, 1972. Political trials in the legal order: a political scientist's perspective. Journal of Public Law 21, 73–126.

Hariman, Robert (Ed.), 1990. Popular Trials, Rhetoric, Mass Media and the Law. The University of Alabama Press, Tuscaloosa, AL.

Harris, Whitney R., 1954. Tyranny on Trial, the Trial of the Major German War Criminals at the End of World War 2 at Nuremberg, Germany, 1945–1946. Southern Methodist University Press, Dallas, TX.

Hesse, Carla, Post, Robert (Eds.), 1999. Human Rights in Political Transition. Zone Books, New York.

Holmes, J.T., 1999. The principle of complementarity. In: Lee, Roy S. (Ed.), The International Criminal Court – The Making of the Rome Statute: Issues, Negotiations, Results. Springer, The Hague, pp. 41–78.

Huneeus, Alexandra, 2011. Courts resisting courts: lessons from the Inter-American Court's struggle to enforce human rights. Cornell International Law Journal 44, 493.

Kairys, David (Ed.), 1982. The Politics of Law, A Progressive Critique. Basic Books, New York.

Kelman, Mark, 1987. A Guide to Critical Legal Studies. Harvard University Press, Cambridge.

Kirchheimer, Otto, 1955. Politics and justice. Social Research 22, 377–398.

Kirchheimer, Otto, 1961. Political Justice, the Use of Legal Procedure for Political Ends. Princeton University Press, Princeton, NJ.

Lahav, Pnina, 1992. The Eichmann trial, the Jewish question and the American-Jewish intelligensia. Boston University Law Review 72, 555–575.

Leebaw, B Anne., 2011. Judging State-Sponsored Violence, Imagining Political Change. Cambridge University Press, Cambridge.

MacKinnon, Catharine A., 1989. Toward a Feminist Theory of the State. Harvard University Press, Cambridge, MA.

Marrus, Michael Robert, 2009. Some Measure of Justice: The Holocaust Era Restitution Campaign of the 1990s. University of Wisconsin Press, Madison, WI.

Minow, Martha, 1998. Between Vengeance and Forgiveness, Facing History after Genocide and Mass Violence. Beacon Press, Boston, MA.

Nino, Carlos S., 1991. The duty to punish past abuses of human rights put into context: the case of Argentina. Yale Law Journal 100, 2619–2640.

Nino, Carlos S., 1996. Radical Evil on Trial. Yale University Press, New Haven, CT.

Nouwen, Sarah M.H., Werner, W.G., 2012. Doing justice to the political: the international criminal court in Uganda and Sudan. European Journal of International Law 21 (4), 941–965.

Orentlicher, Diane F., 1991. Settling accounts: the duty to prosecute human rights violations of a prior regime. Yale Law Journal 100, 2537–2615.

Orentlicher, Diane F., 2004. Whose justice? Reconciling universal jurisdiction with democratic principles. The Georgetown Law Journal 92, 1057.

Osiel, Mark, 1997. Mass Atrocity, Collective Memory and the Law. Transaction Publishers, New Brunswick, Canada.

Princeton Principles on Universal Jurisdiction, 2001. Available at: http://lapa.princeton.edu/hosteddocs/unive_jur.pdf.

Safferling, Christoph J.M., 2011. The role of the victim in the criminal process: a paradigm shift in national German and international law. International Criminal Law Review 11 (2), 183–215.

Segev, Tom, 1993. The Seventh Million: The Israelis and the Holocaust. Hill and Wang, New York.

Shklar, Judith N., 1964. Legalism, Law, Morals and Political Trials. Harvard University Press, Cambridge, MA.

Shraga, D., 2009. Mixed or internationalized courts. In: Cassese, Antonio (Ed.), The Oxford Companion to International Criminal Justice. Oxford University Press, Oxford, pp. 424–426.

Stone, Isidor F., 1989. The Trial of Socrates. Doubleday, New York.

Teitel, Ruti, 1997. Transitional jurisprudence: the role of law in political transformation. Yale Law Journal 106, 2009–2080.

Zalaquett, Jose, 1988. Confronting human rights violations by former governments: principles applicable and political constraints. In: Henkin, Alice (Ed.), State Crimes: Punishment or Pardon. Aspen Institute, Queenstown, MD, pp. 23–71.

Politics and Participation

Carole Jean Uhlaner, University of California Irvine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 16, pp. 11078–11082, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Political participation refers to voluntary activities undertaken by the mass public to influence public policy, either directly or by affecting the selection of persons who make policies. Examples of these activities include voting in elections, helping a political campaign, donating money to a candidate or cause, contacting officials, petitioning, protesting, and working with other people on issues. Particular activities cluster into modes of participation. Substantial cross-national data have shown that wealthier and better-educated people participate at higher rates than the less advantaged, although this relationship is weaker in countries where strong parties or other political organizations provide alternative resources. Research associated with the Civic Voluntarism Model (CVM) has documented that people transfer to political activity skills which they acquire in organizations. The CVM and other recent studies have also demonstrated that people participate more in response to recruitment or mobilization. (Both terms refer to efforts by one person to increase the activity of another.) The Political Action project has shown that people have expanded their 'political action repertory' to encompass both protest and conventional participation. The patterns of participation in authoritarian regimes and in economically less-developed countries raise other issues but substantially follow these broad outlines.

Political participation refers to the activities of the mass public in politics, including, for example, voting in elections, helping a political campaign, giving money to a candidate or cause, writing or calling officials, petitioning, boycotting, demonstrating, and working with other people on issues. Political participation figures in philosophical discussions of democracy and representative government since ancient times because it provides the means for citizens to communicate their views. Normative interest in the topic remains vigorous (*see* Democratic Theory). This article focuses upon the empirical study of political participation. Scholars look for systematic patterns that explain why some individuals are inactive and why others choose certain activities. The patterns reflect the political context and both reflect and affect the structure of political power.

Participation provides the link from the mass public to the political elite who are presumed – and sometimes shown – to respond (*see* Political Representation).

Differing Definitions of Participation

Scholars differ in their definition of political participation. The currently dominant view limits participation to actions that might affect others, as in the following definitions. "By political participation we mean activity by private citizens designed to influence government decision-making" (Huntington and Nelson, 1976, p. 3). "By *political* participation, we refer simply to activity that has the intent or effect of influencing government action – either directly by affecting the making or implementation of public policy or indirectly by influencing the selection of people who make those policies" (Verba et al., 1995, p. 38). The restriction to private citizens is meant to exclude from the concept activity undertaken in their official capacity by those for whom politics and governing are a vocation. A few scholars (cf Milbrath, 1965) include political involvement and activities to support the regime as participation. For these scholars, reading about politics is also political participation, while under the dominant definition it is not (since it does not have direct effects on others).

Another definitional debate in the field concerns the extent to which actions that have the effect of influencing policy without that being the actor's intent also count as political participation. Adopting the more common view, Verba et al. (1995, pp. 38–39) restrict their study to voluntary activity, which they define as follows: "By voluntary activity we mean participation that is not obligatory – no one is forced to volunteer – and that receives no pay or only token financial compensation." In contrast, Huntington and Nelson (1976, p. 7) explicitly include not only *autonomous* participation but also *mobilized* participation, defined as "activity that is designed by someone other than the actor to influence governmental decision-making." Under that definition, they would include as a participant a worker who attends a rally because his boss threatens to demote him otherwise, even if the worker has no intent of affecting government. In practice, it can be difficult empirically to distinguish mobilized from autonomous activity. For answering questions about the impact of participation upon political outcomes, the two are equivalent. The distinction is useful, however, when one focuses on the motives for participation or the impact of participation upon the actor.

The Study of Participation

Participation as Electoral and Hierarchical: Early Studies

The modern empirical study of political participation began developing in tandem with the modern study of electoral behavior using sample survey data. Election outcomes depend not only upon people's preferences but also upon which citizens choose to exercise their preference at the polls. Partially reflecting that link, many of these studies truncated political participation to electoral participation, that is, voter turnout

and occasionally some campaign activities (*see* Voting: Turnout). Much empirical work was done in the United States, and much of that was connected to the Michigan Survey Research Center, especially under the leadership of Warren Miller and Philip Converse. These investigators also collaborated with colleagues outside the United States, especially in Europe, producing further studies addressing participation, especially voting. Under the leadership of Stein Rokkan, Scandinavian scholars produced a substantial body of work on participation as part of their more extensive studies of political behavior (see for example Rokkan, 1970). These studies painted a clear socio-demographic and psychological picture of voters (and occasionally of participants in campaigns). In contrast to nonparticipants, participants in general had more money and education, were more interested in politics, more partisan, more efficacious, more likely to be men, more involved in organizations, and more rooted in their community. Some, including Rokkan, also paid close attention to the relationship between political behavior, including participation, and the citizen's position in the social cleavage structure.

Milbrath summarized the findings to date in 1965. He included political support as participation as well as electoral activities beyond voting. He promoted the influential organizing idea that participation lay on a hierarchical scale – there was a ladder of activities, with the easiest at the bottom and the most difficult at the top. 'Persons who engage in the topmost behaviors are very likely to perform those lower in rank also' (Milbrath, 1965, pp. 17–18). Given the emphasis on electoral activities and given the frame created by the ladder concept, levels of participation in the United States were perceived as low.

Large-Scale Cross-National Studies

Interest in democratic stability prompted cross-national projects that took political participation as an important variable. For Harry Eckstein, participation provided the link upwards from subordinates to superordinates (Eckstein and Gurr, 1975, pp. 60–67). For Almond and Verba in *The Civic Culture* (1963), the participatory behavior of citizens in a country was one important component in characterizing their political culture, but they were just as interested in attitudes and beliefs about hypothetical behavior as in activity. Nonetheless, the data, collected in 1959 and 1960 in five countries (the United States, Mexico, Italy, West Germany, and Great Britain) provided the first major database used for cross-national comparisons of political participation, especially beyond voting turnout. Some of these analyses contributed the important conclusions that persons involved in organizations participate more in politics and that societies with denser organizational structures exhibit higher levels of participation.

Within the decade (1966–71) Verba was in the field again, with an international team of collaborators, studying political activity. Surveys were completed in the United States, Japan, Austria, the Netherlands, India, Nigeria, and the former Yugoslavia. Some of the first products of this research (e.g., Verba and Nie, 1972) overturned the conception of participation as a hierarchy of electoral activities. Instead, participation was expanded to include nonelectoral activities and reconceptualized as modes of activity in any one or more of which individuals might specialize. Although often these are positively correlated, meaning that a participant in one is more likely to take part in others than a randomly selected person, many people nonetheless partake in one or some to the exclusion of other forms of participation. In particular, substantial numbers of people who avoided electoral politics did engage in nonpartisan participation, such as involvement with community groups.

As reported in Verba and Nie (1972) and Verba et al. (1978), as well as other publications from the project, wealth and education were reaffirmed as important predictors of participation. But the main theoretical point is that the strength of that relationship varies across societies as a function of the strength of other, 'group,' resources. Where these are strong, such as in countries with strong political cleavages tied to ascriptive characteristics, or with sociological or occupational segments with strong political leadership, the relationship of socioeconomic status (SES) to participation was relatively weak. Certain well-off segments of the population were 'inhibited' from participation while other, possibly disadvantaged, segments were 'mobilized,' thereby together attenuating the relationship between SES and participation. While others had noted similar phenomena, the finding now had wide cross-national grounding.

The Verba and coworkers' studies deliberately ignored political protest, that is, participation in activities outside accepted conventional norms, including some that are illegal. This omission seemed unfortunate by the 1970s in light of the frequency of real-world protest. Out of the study of revolution and political violence (*see* Revolution) came renewed interest in relative deprivation. But many protest activities are not violent (*see* Political Protest and Civil Disobedience).

A group of European and American scholars launched the Political Action project, a cross-national study focused on protest. They conducted surveys in the United States, West Germany, the Netherlands, Great Britain, and Austria in 1973–74, in Italy, Switzerland, and Finland 2 years later, and a second wave in the first three countries in 1979–80. The results appeared in *Political Action* (Barnes et al., 1979), *Continuities in Political Action* (Jennings et al., 1989), and other publications. One objective of this research was to understand the relationship between protest and conventional participation. The Political Action project clearly showed that enough persons who participated in conventional politics were adding protest to their 'action repertory' to conclude that conventional and protest participation were complements rather than substitutes. The Political Action studies differed from their predecessors in showing that variations in values predict both who will be active and the form of their activity. Specifically, persons with postmaterialist values are more active and more likely to protest than are materialists. Interpretation of the results is hampered slightly by the inclusion of respondents' hypothetical willingness to protest in some of the measures of protest activity.

Other scholars investigated the effect of political institutions on participation. In particular, they showed that cross-national variations in voter turnout rates reflect differences in electoral systems, party competitiveness, registration systems, government structure, and requirements to vote.

Attention to participation in organizations shades into studies of social movements. Where a political scientist sees a participant in a voluntary organization addressing a specific problem (and thus somehow engaged in the communal mode of participation), a sociologist sees someone active in a social movement. The latter have produced another extensive, related literature.

Approaches and/or theories from these cross-national projects have been influential in many country-specific studies of participation. These studies have contributed much that is new, but their conclusions have also tended to support the theories developed in the large projects.

Mobilization as a Predictor of Participation, Not a Type, and the CVM

The next major theoretical development in the study of participation involved incorporation of the recognition that people are more likely to participate when they are asked to do so (i.e., when they are 'recruited' or 'mobilized') and, moreover, that mobilization follows systematic patterns. Rosenstone and Hansen (1993) developed this idea in their study of participation in the United States from 1956 to 1990, claiming that a decline in mobilization over that period accounts in large part for the observed decline in participation.

Verba et al. (1995) use the related concept of 'recruitment' cross-sectionally in their study of political participation in the United States based on a 1989–90 survey. It is the third component of the CVM; the first two are 'resources' and 'engagement.' They argue that higher levels of any of these factors lead to greater participation – people participate in politics because they can, because they want to, and because someone asks.

'Resources' are most readily measured by income and education, but these serve as proxies for more fundamental measures. The most interesting components of resources are the 'civic skills,' operationalized by Verba et al. (1995) as writing a letter, making a speech, and attending or planning a meeting where decisions are made in one or more nonpolitical settings, such as on the job, at church, or in a voluntary association. Civic skills turn out to explain the long-observed association between organization membership and participation. 'Engagement' includes interest, efficacy, political information, and partisan attachment, and also specific issue or ideological concerns. Note the connection between this part of the model and the finding that post-materialists are more participatory. 'Recruitment' includes both requests from individuals and those from persons in authority (e.g., on the job, at church, or in an organization). This latter route provides one path by which group resources operate.

People with greater wealth and education in general have more resources, are more engaged, and are more likely to be recruited, so overall the relationship between socioeconomic status and participation is positive. The CVM model provides a fuller understanding however of why the relationship holds. Moreover, were it applied cross-nationally, it would point to components that need measuring. For example, churches play a role in the United States that is not likely to be replicated elsewhere.

Cost and benefit analyses inspired by rational actor models influence much contemporary work on participation, including that of Rosenstone and Hansen (1993) and Verba et al. (1995).

Many of the factors that lead to high levels of political participation are associated with high levels of social capital (*see* Social Capital) notably including substantial organizational involvement. Since social capital is also measured with political participation, the relationship among the concepts and the phenomena need more attention.

Conceptual Issues in Research on Political Participation

Participation Outside Advanced Democratic Countries

Most of the research discussed above has focused on economically advanced democratic countries, with the occasional addition of other research sites. The restriction to democratic countries reflects the particular role of political participation in a democracy. Participation provides the means by which people control their government, most decisively so in elections which determine who rules. The unique power of the vote to change rulers provides the force for most of the other modes of participation. Campaigning, expressions of opinion, community organizing, and other actions command attention in part because ultimately the opposition can win an election.

In an authoritarian regime, the electoral connection no longer exists. Nonetheless, there is still political participation. Although the people no longer select the rulers, the rulers still need to keep the population reasonably satisfied to forestall revolt, and, to do so, they need information on the people's preferences, and they need to respond to discontent. Several very interesting studies have examined participation in authoritarian regimes, such as Friedgut's study of the USSR (1979) and the work by Shi (1997) on China.

In economically less-developed countries, one frequently finds that government institutions are also less well developed, that the associational networks are weak, or that loyalties to local groups outweigh those to governmental institutions. In these circumstances, much political participation may consist of informal involvement in the local arena or in traditional institutions (see, for example, Dietz, 1998).

Issues of cross-national and cross-cultural comparability loom large in all studies of participation, but they are especially salient for research in nondemocratic or less advanced countries. One must rethink what counts as 'political' in these contexts in light of the institutional situation. For example, some of these studies led to the suggestion that participation includes activity directed toward altering the implementation of policy, rather than just its formulation, or that bribery is an important mode. Certain actions that are nominally the same can serve very different functions in differing contexts. For example, 'voting' may just show regime support in one country but select a government in another. The independent variables of importance also vary. For example, because of the particular circumstances in China, 'education' does not usefully measure resources there (Shi, 1997, pp. 145–148). In many countries, other institutions are likely to play the role of churches in the United States as a source of civic skills.

Matters become more complex when an authoritarian regime democratizes. Many current studies examine participation in Eastern Europe and the republics of the former Soviet Union. With the abrupt changes in regime, the process of political socialization (*see* Politics and Socialization) will probably also change, providing good conditions in which to observe the effect of political attitudes on participation.

Leadership

If people participate more when they have resources, are engaged, and are recruited, then elite action can significantly alter the level of participation by affecting engagement and recruitment. Political leaders provide resources to make participation easier (such as a babysitter, a ride to the polls, or information on what to support). Leaders contact people, directly or through proxies, asking them for action. Leaders can increase engagement by framing the public debate in a way that increases the stakes of political competition by communicating this message to their listeners.

The contemporary mass media plays a role very similar that of to leaders. They frame issues and strongly affect (at times, negatively) engagement and recruitment. However, their motivations differ from those of leaders as they seek attention rather than support, and their actions are not targeted.

The political world has many persons vying for power, each of whom needs large numbers of supporters to achieve power. Only some are successful in recruiting others and, as noted above, theory has not yet explained the circumstances of success. But political participation will increase when many are active and successful, and it will be higher than otherwise expected for persons who are targets of their efforts.

Editors' Note

The years following the publication of Uhlaner's essay above have witnessed renewed emphases on many topics in participation research. Four topics are:

1. What participation is – In relative terms, less emphasis has been placed on defining participation in general, and more emphasis has been given to newer activities within each major type (electoral, communal, protest), as well as to newer types of participation such as boycotting/buycotting as a newer protest activity, and political discussion and social media use as a newer participation type (see, e.g., Dalton, 2014, Chapter 3).
2. Has participation changed – For example, voting turnout has decreased but remains the most frequently performed electoral activity; signing petitions has increased and is the most frequently engaged in protest activity, with its frequency now similar to that of voting turnout in some countries; and communal activities occurring more or similarly often in the United States compared with other countries (see Dalton, 2014, Chapter 4; see also Attitudes, Political and Civic Culture; Political Protest and Civil Disobedience; and Voting: Turnout in this encyclopedia).
3. Who participates – Current research continues to focus on whether and if so, then why differences in electoral, communal, and protest activities have contracted or expanded between people who are less educated, ethnic-minority members, of lower socioeconomic status, and/or younger, and their counterparts (see Dalton, 2014, Chapter 4; Leighley and Nagler, 2013). With respect to younger people, there is now more focus on the effects of civic education, and of Internet and other social media uses, on their engagement in electoral, communal, or protest activities (see, e.g., Dalton, 2009; Milner, 2010; see also Generations: Political and Politics and Socialization in this encyclopedia).
4. How participation can/should be studied – This refers to the design of studies capable of providing the data necessary to test which of several alternative models provide(s), or whether a coherent, composite model produces, the best explanation(s) of electoral, communal, and/or protest participation. Notable models are cognitive engagement (information interest), civic voluntarism (resources mobilization), equity fairness (relative deprivation), general incentives (costs–benefits–norms), and social capital (networks trust) (for research reviews and test results, see, e.g., Clarke et al., 2004, Chapters 7 and 8; Whiteley et al., 2013, pp. 105–108).

See also: Alliances: Political; Campaigning: Political; Citizen Participation; Democratic Theory; Efficacy, Political; Party Identification and Nonpartisanship; Political Protest and Civil Disobedience; Political Representation; Politics and Socialization; Power; Revolution; Social Capital; Voluntary Organizations; Voting, Explanations of: Social Class; Voting, Sociology of; Voting: Turnout.

Bibliography

Almond, G., Verba, S., 1963. The Civic Culture. Princeton University Press, Princeton, NJ.

Barnes, S.H., Kaase, M., Allerbeck, K.R., Farah, B.G., Heunks, F., Inglehart, R., Jennings, M.K., Klingemann, H.D., Marsh, A., Rosenmayr, L., 1979. Political Action: Mass Participation in Five Western Democracies. Sage, Beverly Hills, CA.

Conway, M. Margaret, 2000. Political Participation in the United States, third ed. CQ Press, Washington, DC.

Clarke, Harold D., Sanders, David, Stewart, Marianne C., Whiteley, Paul F., 2004. Political Choice in Britain. Oxford University Press, Oxford.

Dietz, H., 1998. Urban Poverty, Political Participation, and the State. University of Pittsburgh Press, Pittsburgh, PA.

Dalton, Russell J., 2014. Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies, sixth ed. Sage Publications/CQ Press, Los Angeles.

Dalton, Russell J., 2009. The Good Citizen: How the Young Are Reshaping American Politics, revised ed. Congressional Quarterly Press, Washington.

Eckstein, H., Gurr, T.R., 1975. Patterns of Authority: A Structural Basis for Political Inquiry. Wiley, New York.

Friedgut, T.H., 1979. Political Participation in the USSR. Princeton University Press, Princeton, NJ.

Huntington, S.P., Nelson, J.M., 1976. No Easy Choice. Harvard University Press, Cambridge, MA.

Jennings, M.K., van Deth, J.W., Barnes, S.H., Fuchs, D., Heunks, F.J., Inglehart, R., Kaase, M., Klingemann, H.-D., Thomassen, J.J.A., 1989. Continuities in Political Action. de Gruyter, Berlin.

Leighley, Jan E., Nagler, Jonathan, 2013. Who Votes Now? Demographics, Issues, Inequality and Turnout in the United States. Princeton University Press, Princeton.

Milbrath, L.W., 1965. Political Participation. Rand McNally, Chicago.
Milner, Henry, 2010. The Internet Generation: Engaged Citizens or Political Dropouts. Tufts University Press/University Press of New England, Lebanon, NH.
Rokkan, S., 1970. Citizens Elections Parties. Oslo University Press.
Rosenstone, S.J., Hansen, J.M., 1993. Mobilization, Participation, and Democracy in America. Macmillan, New York.
Shi, T., 1997. Political Participation in Beijing. Harvard University Press, Cambridge, MA.
Verba, S., Nie, N.H., 1972. Participation in America. Harper & Row, New York.
Verba, S., Nie, N.H., Kim, J.-O., 1978. Participation and Political Equality. Cambridge University Press, Cambridge, UK.
Verba, S., Schlozman, K.L., Brady, H.E., 1995. Voice and Equality. Harvard University Press, Cambridge, MA.
Whiteley, Paul, Clarke, Harold D., Sanders, David, Stewart, Marianne C., 2013. Affluence, Austerity and Electoral Change in Britain. Cambridge University Press, Cambridge.

Politics and Socialization

M Kent Jennings, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14522–14525, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

In its broadest sense political socialization refers to the content, processes, and dynamics of political learning across time and space. Related terms include civic education and citizenship training. The practice of political socialization is as old as political life. However, the term itself, and systematic research, did not emerge until the post-World War II period. An initial burst of scholarship in the 1960s and 1970s was followed by a prolonged decline. Interest revived during the 1990s, fueled in part by democratization movements around the world and, ironically, by concerns in established democracies about the declines in civic virtue among upcoming cohorts. The consequences of political socialization have been gauged in terms of its impact on the political character of individuals, the perpetuation of political cultures, and the functioning of political systems. Although most attention has been devoted to socialization in the pre-adult years, research in the area often merges into issues involving life cycle, generational, and historical effects. Survey research has been the most commonly employed methodology in the field. Longitudinal and cross-national research designs have been especially valuable in advancing the state of knowledge.

In its broadest sense political socialization refers to the content, processes, and dynamics of political learning across time and space. Related, more restricted terms include civic education and citizenship training. Because of the importance commonly attached to the consequences of early socialization, scholarship in this area has focused on the pre-adult years though not to the exclusion of later periods as well. Political socialization figures heavily in discussions of political culture, the relation of the individual to the political system, and political system stability and change. This article covers the major foci of the field, methodological approaches, and the field's evolution and development since the 1950s.

Major Topics and Questions

Some scholars, political scientists in particular, are concerned about the implications of political socialization for how political institutions function. Other scholars focus more on what socialization means for individuals and how they cope with their political environment.

Scholars have tended to borrow from other disciplines rather than developing original perspectives. Applications of social learning and cognitive growth theories predominate. Learning about politics occurs as a result of intentional and direct efforts as well as unintentional and indirect efforts.

The range extends from broad topics such as national identity and consensus political norms to narrower ones such as partisanship and issue preferences. The broader topics are especially important in terms of how political culture is preserved, while the narrower ones help explain individual level behavior and attitudes.

The 'agents of socialization' is one of the most explored topics, with the role of the family, schools, and political events receiving heaviest scrutiny. Complex research designs are often employed here so that the individuals being socialized can be matched up with the agents providing the socialization cues.

There are intracountry variations. Political socialization may proceed in different ways and generate different outcomes according to individual and contextual characteristics. Most attention has been paid to the role of social class, gender, race, region, religion, and birth cohort.

On the other hand cross-country variations exist. Commonalties based on individual developmental processes and differences based on institutional arrangements have been observed. Some of these conclusions are based on systematic, multicountry investigations whereas others are derived from a comparison of independent, one-country studies.

Although much emphasis has been placed on transitions from early childhood up through young adulthood, the study of political socialization increasingly is being viewed from a lifelong perspective in order to comprehend the residues of earlier socialization as well as the impact of adult life stage transitions. For many nonconsensus political beliefs and practices the formative years appear to be young adulthood, after which beliefs and practices tend to solidify. For beliefs and values of a more integral and widely shared nature, the formative years occur earlier.

Of interest to a wide range of scholars, generational and historical effects may reflect either gradual changes in the social composition of succeeding cohorts or sharper discontinuities due to secular events. Rising education levels and affluence exemplify generational effects due to social composition. Wars, depressions, revolutions, and student protest movements are examples of more abrupt, event-driven effects.

Methodology

The majority of the literature rests on a foundation of empirical methodology. Attention has focused more on individuals than on institutions and systems as units of analysis.

There are various modes of data gathering. Surveys of probability and nonprobability samples compose the major

mode. Three variants have been used. One consists of self-administered questionnaires of elementary and secondary school students. The use of such instruments for young children has been questioned. A second variant consists of personal interviews, either face to face or by telephone. Some of these surveys employ a structured format while others are more semi-structured. Investigators using either mode typically employ statistical analysis. Other modes of data collection include intensive interviews, sometimes combined with participant observation and projective techniques, with a small number of subjects (Coles, 1987). Analysis in this instance is usually qualitative in nature. Although not typically classified under the heading of political socialization, analytic biographies of political figures – employing a variety of data sources—also constitute a mode of data gathering. Studies based on experiments are relatively rare.

Most investigations are cross-sectional, that is, the subjects were studied only once. However, a few over-time studies of the same individuals exist, thereby enabling more insight into developmental patterns. Some studies stretch from pre to full adulthood and involve either special populations or more general ones. Other longitudinal projects, especially those assessing the impact of educational institutions and political events, cover a shorter period of time.

Replicated surveys of the same populations provide historical trend data. Longitudinal studies of pre-adult populations are now available in a number of countries via national and international archives. Similarly, replicated surveys of adult birth cohorts are widely available. These surveys facilitate analyses devoted to assessing the probable impact of preadult experiences as cohorts age.

Initial Development and Ascendance of the Field

The phenomenon of political socialization has a long history. The preparation of individuals for their roles in the political world is as old as political life itself. Both political authorities and ordinary citizens have been subject to the practices and outcomes of socialization regardless of political regime type. By contrast, the systematic study of political socialization has a relatively short history. Around the turn of the twentieth century educators in the United States conducted studies of school children that included choices of exemplars, national figures most admired by the children (Greenstein, 1965). These 'surveys' and occasional other investigations in a variety of settings marked the first empirical studies of what would later be called political socialization.

It was not until the post- World War II era that more sustained, disciplined research emerged. The term itself surfaced in the mid-1950s. By the early 1960s the label had achieved wide academic currency among political scientists in the United States and quickly spread to other western countries. Early work in the Unites States, the initial center of the empirical investigations, focused on pre-adolescents and emphasized the positive and relatively benign processes and outcomes of political socialization. Although more realistic perceptions and understandings set in with advancing age and cognitive development, the overall effect was viewed as salutary both for the successful functioning of the political system and the citizenry (Easton and Dennis, 1969).

Following in the wake of these pioneering studies came a flood of projects in the United States and elsewhere. These inquiries tended to elaborate upon, modify, and in some instances to reject the conclusions of the earlier work. Variations according to social class, race, and location were identified. Political institutions conditioned the content, processes, and outcomes in fundamental ways. Considerable attention was devoted to the agents of socialization. Although the family appeared to be the key agent, it became clear that many other forces—including individuals themselves—also contribute to the formation and maintenance of political orientations. In general the more concrete, affect-laden, and reinforced the attribute in question, the more likely it was to be reproduced in the person being socialized and to persist over time.

Waning Interest

Scholarly inquiry in this area began to decline after the mid-1970s, especially in the United States. What had been a bull market turned bearish. Two factors appeared to be at work beyond the fading effect often associated with scholarly trends. One source lay in the dwindling novelty of studying school children. Political scientists in particular were accustomed to dealing with the 'real world' of politics, one occupied and ruled by adults. A second likely source rested in the appearance of youth movements throughout much of the western world in the late 1960s and early 1970s. This development not only shifted attention away from pre-adults but also appeared to undermine the proposition that the trusting orientations observed in many of the earlier studies would persist over time. Indeed, the question of persistence posed a serious threat to the field.

Ironically, even as fresh research efforts tailed off, the concept and the term itself became embedded in any number of subfields, including public opinion, electoral behavior, the politics of race and gender, political culture, and regime stability. Political socialization came to be assumed, rather than observed. Still, research continued on in a number of countries (Percheron, 1993; Ichilov, 1990) and longitudinal studies began to appear (Alwin et al., 1991; Jennings and Niemi, 1981).

Resurgence and Future Prospects

After a prolonged lull, scholarly inquiry began to increase in the 1990s and has continued onward. Three developments helped lead to a resurgence of interest. One development was the crumbling of the Soviet Union and the appearance of transitional regimes and new democracies around the globe. These changes generated numerous surveys of adults that emphasize generational differences and studies of how and with what effect pre-adults are being socialized under the new regimes (e.g., Flanagan and Sherrod, 1998). New, large-scale immigration patterns also presented opportunities for studies of adult resocialization. Thus, dramatic events in the political world have provided a natural laboratory for examining the processes

and outcomes of political socialization. A second development springs from the decline in social capital and civic virtue among upcoming cohorts said to characterize most Western countries. In an effort to understand what is happening and to propose possible solutions, a variety of institutions and scholars have turned to the question of education and training of the young. National and cross-national assessments dealing with the character and impact of curriculum and school characteristics are being employed (Smith, 1999). Much of this work is being carried out under the banner of civic education inquiries (Torney-Purta et al., 2001).

A third development consists of renewed interest in the question of persistence and the dynamics of learning. Results from the long-term panel studies form part of this renewal as do the shorter-term studies of children and adolescents. The accumulation of replicated surveys as represented by such projects as national election studies, the General Social Surveys, and the Eurobarometer Surveys are also relevant because they permit the tracing of birth cohorts over extended periods of time. In doing so, they enable informed interpretations drawing upon the conditions under which pre-adult socialization occurred and how these cohorts respond to later developments. These new developments augur well for continued scholarly inquiry. At the same time, the prospects remain bleak with respect to the study of young children, a major focus of the early research efforts.

Editor's Note

M. Kent Jennings is one of the 'founding fathers' of political socialization research and theory and among the first to recognize that if we seek to understand the political views and behaviors of adults, we need to start by looking at how political thinking is formed in childhood. Jennings' earliest publication on the topic dates to 1968 ("The Transmission of Political Values from Parent to Child" (with Richard G. Niemi), American Political Science Review, 62 (March, 1968), 169–184.) His article on the topic in the first edition of the Encyclopedia is reproduced here as a legacy entry.

See also: Attitudes, Political and Civic Culture; Citizen Participation; Citizenship: Political; Educational Sociology; Generations: Political; Opinion Formation, Theory of; Political Communication; Political Protest and Civil Disobedience; Politics and Participation; Socialization and Education: Theoretical Perspectives.

Bibliography

Alwin, D.F., Cohen, R.L., Newcomb, T.M., 1991. Political Attitudes Over the Life Span: The Bennington Women after Fifty Years. University of Wisconsin Press, Madison, WI.

Coles, R., 1987. The Political Lives of Children. Houghton Mifflin, Boston.

Easton, D., Dennis, J., 1969. Children in the Political System. McGraw Hill, New York.

Flanagan, C.A., Sherrod, L.R. (Eds.), 1998. Political development: youth growing up in a global community. Journal of Social Issues, 54, pp. 447–627.

Greenstein, F.I., 1965. Children and Politics. Yale University Press, New Haven, CT.

Ichilov, O. (Ed.), 1990. Political Socialization, Citizenship Education, and Democracy. Columbia University, New York.

Jennings, M.K., Niemi, R.G., 1981. Generations and Politics: A Panel Study of Young Adults and their Parents. Princeton University Press, Princeton, NJ.

Percheron, A., 1993. La Socialization Politique (Political Socialization). Armand Colin, Paris.

Sears, D.O., Valentino, N.A., 1997. Politics matters: political events as catalysts for preadult socialization. American Political Science Review 91, 45–65.

Smith, E.S., 1999. The effects of investments in the social capital of youth on political and civic behavior in young adulthood: a longitudinal analysis. Political Psychology 20, 553–580.

Torney-Purta, J., Lehmann, R., Oswald, H., Schulz, W., 2001. Citizenship and Education in Twenty-Eight Countries: The Civic Knowledge and Engagement of Adolescents. International Association for the Evaluation of Educational Achievement. Amsterdam.

Politics, Formal Modeling of

Jeffrey S Banks, California Institute of Technology, Pasadena, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11717–11719, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article describes the noncooperative game-theoretic approach to studying political phenomena, using examples to illustrate the various concepts and moving parts. It examines the ability of this approach to generate conclusions about the influence of different political institutions on policy outcomes. This approach is put in the context of earlier results from the cooperative game-theoretic approach, for example the median voter theorem and the chaos theorems.

Positive theories of politics provide explanations for why political phenomena occur in the manner that they do. Examples of such phenomena include which parties or candidates are elected at certain times, what policies legislative bodies adopt, and when and how wars are fought between countries. Formal models complement these explanations by using the language and analysis of mathematics to derive the phenomena as necessary consequences of certain sets of underlying factors. As with mathematical models in other disciplines, the main advantage of such an approach is twofold: to give an unambiguous representation of the relevant moving parts of a theory (the clarity of the argument) and to enforce a logical coherence on a theory (the consistency of the argument). The main drawback is found in the required simplification of the phenomena down to a manageable number of moving parts, and the subsequent possibility of simplifying away the very essence of the problem.

The Basic Structure

Most formal models of politics fall in the category of rational choice where the presumption is that observed outcomes are the result of decisions made by a certain relevant set of individuals. These could be voters and candidates in the first example above, elected representatives and appointed ministers in the second, or heads of state in the third. The general behavioral postulate is that individuals attempt to influence the outcome in a manner reflecting their preferences (tastes, values, etc.) over the set of possible outcomes. For instance, in selecting campaign platforms, a candidate who only cares about winning the current election might look to adopt policies appealing to the greatest number of voters; conversely, a candidate who cares intrinsically about the policies to be implemented might shy away from politically popular but personally unpalatable alternatives. Whatever their specific form, individuals' preferences are taken to be primitives of the model, and so are held fixed throughout the analytic determination of an outcome. However, different specifications of preferences can give rise to different outcomes. Indeed, to the extent the predicted outcome varies with changes in the underlying individual preferences, one has a potential explanation for any variation in observed outcomes as being due to variation in these preferences. The extent to which these preferences can themselves be modeled as functions of observable parameters, such as wealth or legislative seniority, permits one to generate a theoretical relationship between these parameters and observable political outcomes as well.

The influence of preferences on outcomes occurs through the choices the individuals make. The most common assumption is that each individual's choice is optimal with respect to their preferences, for example, in the above scenario the candidate selects a vote-maximizing policy. This optimality assumption may be far from a realistic representation of individual decision-making, implying as it does that individuals can effortlessly solve what are at times relatively difficult mathematical problems. However, assuming otherwise typically requires the presence of auxiliary factors (for instance, decision costs) to rationalize *non*-optimizing behavior, factors which, while reasonable, tend to create additional complications in the model. Also the optimality assumption allows one to import existing results from optimization theory to aid in the analysis.

As seen in the examples above, most formal models of politics involve more than one individual, and so this individual level of optimization often takes place within the confines of non-cooperative game theory. Such models explicitly describe the interaction among the individuals that ultimately produces a collective decision. This description begins with a 'game form' consisting of (1) the set of strategies available to each individual, and (2) the specification of which outcome occurs when a given profile of strategies is chosen (with each profile consisting of one strategy per individual). Continuing with the electoral setting, one example of a game form is where the set of strategies for a candidate is equal to the set of available policies, a voter's strategy describes which candidate to vote for based on the candidates' chosen policies, and an outcome, consisting of who wins the election and with what policy, is determined by the candidate receiving the most votes. The remaining piece of the puzzle then is the derivation of the individuals' strategies, with the fundamental concept being that of the 'Nash equilibrium (Nash, 1950a, b; 1951).' Note that, even assuming all individuals have well-defined preferences over the set of outcomes, the notion of optimality in this multiperson environment is necessarily conditional, as an individual's best choice of strategy may depend nontrivially on the strategy choices by the other individuals. For instance, in the electoral game above, the location of

a candidate's optimal policy may depend on the voters' behavior as well as the policies endorsed by rivals. A Nash equilibrium is a profile of strategies wherein each individual's strategy is optimal, given the strategies of the others, that is, conditional on the choices by the other individuals, no one individual has an incentive to alter her choice. A Nash equilibrium outcome is thus a collective decision consistent with individually optimal behavior and constitutes the basic object of analysis in any noncooperative model. (In certain game forms, some Nash equilibria are inherently more plausible than others and so refinements of Nash equilibria, such as subgame perfection, are used.)

As with the choice of preferences, the choice of game form is in the hands of the modeler. This choice may be influenced by observable features of the political phenomenon in question, as in the above electoral example. In such situations, there will invariably be a trade-off between making the game form descriptively accurate and keeping the game form analytically tractable. Thus while the electoral rule may easily be incorporated, simplifying the campaign stage of the process to a single choice of policy, thereby avoiding the complexities found in actual elections, may be necessary to guarantee the existence of and to solve for equilibrium outcomes. At other times, there may not exist any such observable features to leverage, an example of which would be when two heads of state are bargaining over land and other resources in the shadow of war. Here one would be hard-pressed to assert that the unstructured negotiations taking place can be definitively captured in any one particular game form. Hence descriptive accuracy is not an issue, and the modeler is relatively unconstrained in the search for a 'good' game form, that is, a game form capable of generating new and valuable insights into the political phenomenon in question.

Yet to the extent that the game form incorporates observable features of the political process and that the equilibrium outcomes vary with these features, one has a second source of variation in outcomes above and beyond changes in preferences, namely, changes in the political process. For instance, the equilibrium outcomes under the above plurality electoral rule may differ from those associated with a run-off rule (where, if no candidate receives a majority of the votes, the two candidates receiving the highest number of votes are paired in a second election, in which majority rule determines the winner) even if the candidates' and voters' preferences are held fixed. This allows one to make statements about the degree to which differences in observed collective choices are due to differences in preferences, or to differences in political institutions. More ambitiously, this type of comparative institutional analysis can be used to predict the effect changes in existing political institutions will have on collective outcomes.

An Alternative Approach

While noncooperative game theory has tended to dominate the modeling landscape in recent years, much of the early formal work employed the techniques of cooperative game theory. Rather than explicitly modeling outcomes as the aggregation of individual decisions, under this approach one looks for outcomes that are stable in some sense, with this 'sense' implicitly characterizing the (unmodeled) underlying strategic interaction. Given the normative appeal and frequent use of majoritarian-based political procedures, the focus was primarily on identifying outcomes which are majority-preferred to all others in pair-wise comparisons. Such an outcome would then be immune to changes proposed by any majority coalition, whereas all other outcomes would not.

It turns out that a majority-stable outcome necessarily exists in certain environments, the most compelling being when the collective choice concerns the selection of a single number. For example, suppose the issue before a legislature is the size of the nation's military budget, the set of possible outcomes can be taken to be the set of nonnegative real numbers. Suppose further that the legislators have 'single-peaked' preferences: each has some best number, and the value to them of the outcome falls as one moves away from this best number. Then a stable outcome will exist and, if the number of individuals is odd, will be unique and equal to the median of these best numbers. This 'median voter theorem' has been at the core of numerous and varied empirical studies of collective choice, from local school budgets in the United States to European Union decision-making. Its comparative advantage as a formal model is obviously its parsimony, requiring as it does only a description of how individual preferences over outcomes are related to any underlying parameters of interest (e.g., household income in the local school budget case). Further, by imposing additional assumptions on preferences, one finds that the median voter's preferences coincide with those of the majority over the entire set of outcomes, implying that the median voter is in a literal sense representative of the collective as a whole. The benefits from this result accrue principally to the noncooperative modeler: suddenly a one-dimensional model of the interaction between the US President and the 435 members of the US House of Representatives over the size of the military budget, with majority rule dictating the decision-making among the latter, can be simplified to a two-person interaction between the President and the median member of the House.

The main shortcoming of the median voter theorem of course is its restriction on the set of outcomes to a single dimension. In certain situations this is benign, as in the above example when the policy choice in question consists of selecting the military budget. However, a policy choice consisting of the distribution of this budget across more than two groups is inherently a multidimensional problem and the conclusions about majority rule for multidimensional problems, relative to their single-dimensional cousins, could hardly be more dramatic. In particular, for most descriptions of preferences, no majority-stable outcome exists (where 'most' has a mathematically precise definition in this setting). As an example of such non-existence, suppose in the above scenario that each legislator cares only about how much their favorite group receives, with one-third of the legislators associated with each of three groups. Then for any distribution of the budget, the legislators for two groups (which together constitute a majority) prefer to take away any amount going to the third group and re-distribute it amongst themselves, thereby guaranteeing that no distribution is majority-stable. Furthermore, when no stable outcome exists, not only is there no representative voter but in fact one can construct majority

preference cycles throughout the policy space (i.e., policy *a* is majority-preferred to policy *b*, which is majority preferred to policy *c*, which … which is majority-preferred to policy *z*, which is majority-preferred to policy *a*). Thus when majority rule breaks down, it breaks down completely. Subsequent research has shown that these negative conclusions continue to hold, to a greater or lesser degree, for most non-majoritarian rules as well.

As with the median voter theorem, it is difficult to overstate the impact these 'chaos theorems' have had on the formal modeling of politics. Clearly, in a multi-dimensional world, the modeler has lost the analytic shortcut of representing a collective's decision-making as being governed by a single individual (or, at least, the logical rationale for doing so). More importantly, the theorems gave rise to a new-found appreciation for political institutions, such as committee systems, restrictive voting rules, and the like as key elements in the determination of collective outcomes. Indeed, such institutions are often depicted as necessary ingredients in any multidimensional model of politics, in that, but for such institutions, collective decision-making would fall prey to the inevitable onslaught of cycling and chaos. This interpretation, while quite popular, appears inconsistent with the central tenet of formal modeling, that of equilibrium analysis, as the chaos theorems are properly seen as nonequilibrium results. Put differently, these theorems do not predict that (as commonly phrased) 'anything can happen,' for as nonequilibrium results they do not predict anything at all. Rather, they demonstrate the impossibility of any general theory of collective behavior based solely on the (majority) aggregation of preferences, and hence the necessity of being more explicit about the decision-making process so as to provide the analytical traction required to formulate a well-posed model. From this perspective, the lesson to be learned from the chaos theorems is one for the modeler of collective decision-making, rather than as one about collective decision-making itself.

Editor's Note

Especially influential has been since then the work of E.Ostrom(1990; 1998; 2005) ; Ostrom/Gardner/Valker, 1994; Dietz/Ostrom/Stern, 2003, She analyzed the functioning of institutions from a rational choice and game theoretical point of view. Starting with the tragedy of the commons (Hardin, 1968) which has been a major concept in environmental research and policy discussions. One assumes that a common pool exists, which can be utilized by a lot of persons. This is true for example for the pacific ocean or the atlantic ocean from which many nations harvest fish. Overfishing has been taken place and may lead finally to the destruction of the resources. The central dilemma is as follows. If one nation stops to circumvent the destruction of the resources andtheothers however still continue, destruction goes on. A lot of research has been performed in the last thirty decades to clarify the conditions under which destruction takes place or constructive solutions are found. This stream of research has now combined mathematical models, laboratory experiments, field studies and case studies. Such a mixed method approach has become more frequently also in other applications.

Additionally game theory became a much more visible place in a lot of doctoral programs especially in the United States but also in the Netherlands, Great Britain, Germany and Japan.

Furthermore in the large summer schools from ICPR, ECPR and in the Essex Summer School game theory courses are regularly offered. Several general introductions to game theory and its application to political science have been now published (Austen-Smith/Banks, 1998; McCarthy/Meirowitz, 2007; Gates/Humes, 1997).

See also: Action, Theories of Social; Bounded Rationality; Collective Action; Game Theory: Noncooperative Games; Model Testing and Selection, Theory of.

Bibliography

Austen-Smith, D., Banks, J.S., 1998. Social choice theory, gametheory and positive political theory. Annual Review of Political Science 259–287.
Austen-Smith, D., Banks, J.S., 1999. Positive Political Theory I: Collective Preference. University of Michigan Press, Ann Arbor, MI.
Dietz, T., Ostrom, E., Stern, P.C., 2003. The struggle to govern the commons. Science 302, 1907–1912.
Gates, S., Humes, B.D., 1997. Games, Information and Politics. University of Michigan Press.
Hardin, G., 1968. The Tragedy of the Common. Science 162, 1243–1248.
McCarthy, N., Meirowitz, A., 2007. Political Game Theory An Introduction. Cambridge University Press, Cambridge.
Morrow, J.D., 1994. Game Theory for Political Scientists. Princeton University Press, Princeton, NJ.
Nash, J.F., 1950a. Equilibrium Points in n-Person Games. Proceedings of the Natioal Academy of the United States of America 36, 48–49.
Nash, J.F., 1950b. The bargaining problem. Econometrica 18, 155–162.
Nash, J.F., 1951. Non-cooperative games. The Annals of Mathematics 54, 286–295.
Ordeshook, P.C., 1986. Game Theory and Political Theory: An Introduction. Cambridge University Press, Cambridge, UK.
Ostrom, E., 1990. Governing the Commons. The Evolution of Institutions for Collective Actions. Cambridge University Press, Cambridge.
Ostrom, E., Gardner, R., Valker, J., 1994. Rules, Games and Common Pool Ressources. Michigan University Press, Ann Arbor.
Ostrom, E., 1998. A behavioral approach to the rational choice theory of collective action. American Political Science Review 92, 1–22.
Ostrom, E., 2005. Understanding Institutional Diversity. Princeton University Press, Princeton.
Shepsle, K.A., Bonchek, M.S., 1997. Analyzing Politics: Rationality, Behavior, and Institutions. Norton, New York.

Pooled Cross-Sectional and Time Series Analyses, in Political Science

Ellen M Key, Appalachian State University, Boone, NC, USA
Matthew J Lebo, Stony Brook University, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Time can either be treated as a nuisance or as a substantively interesting variable across a variety of political science subfields and a wide array of research questions. The popularity of studying political phenomena over time has increased over the past 40 years. Encompassing any study where data are collected at multiple points in time, time series analysis includes studies of duration, event counts, (pseudo-)panels, repeated cross-sections, pooled cross-sections, and studies of volatility. Beginning with a discussion of the development and integration of 'long' time series methodology in the political science literature, we also briefly discuss other types of time series. We conclude with some suggestions for future research.

Introduction

While most empirical analyses in political science study cross-sectional units, the use of political data measured over time has become increasingly popular over the past 40 years. The broad category of political time series would include any study where data are collected at more than one point in time. It contains many types of analysis including longer time series, event counts, studies of volatility, duration analysis, forecasts, pooled cross-sectional time series (PCSTS), panel models, pseudo-panels, and repeated cross-sectional analyses. This article begins with a discussion of 'long' time series methods and the integration and development of these approaches into the empirical literature in political science. Following that is a brief discussion of other types of time series.

There are many ways that time can be a factor in collecting and analyzing political data. The most prevalent example is the use of a variable collected at multiple points in time. Such a variable, denoted Y_t, could be the percent of samples giving a particular answer on identical survey measures at different points in time like monthly presidential approval (e.g., Mueller, 1970), the actions of one country toward another within multiple discrete time periods like the data collected by the Kansas Events Data System (e.g., Schrodt and Gerner, 1994), a measure of government actions or performance that is collected at regular intervals such as the percent of times a president wins his battles with Congress (e.g., Lebo and O'Geen, 2011), or one of the countless other examples.

Studying a political variable over time offers many advantages. To begin, the study of political dynamics is itself interesting – studying important political phenomena over time is certainly as promising as examining variation across units fixed in time. The movement of variables such as macropartisanship (MacKuen et al., 1989) and the gender gap (Box-Steffensmeier et al., 2004), for example, tell us a great deal about the long-term movement of the electorate and make important complements to cross-sectional studies. (The choice of the level of analysis is critical when beginning any project and the appropriate level of analysis depends on the research question. If it is an atomistic question about the behavior of individuals, then analysis should be conducted at the individual level. If, however, the question is about how the citizenry, electorate, or economy moves as a whole, then macroanalysis is the most appropriate.)

The value of aggregate-level studies has been demonstrated particularly well in the economic voting literature. As Erikson, MacKuen, and Stimson note "elections are won and lost in the aggregate. And the aggregate moves at the margin; a few people doing this or that systematically produces big, sometimes shocking, aggregate effects" (2002: xix). Although Converse (1964), Campbell et al. (1960), and others find the average citizen to show little ideological constraint, the degree of constraint varies throughout the population. The well informed are more likely to have stable opinions over time and the public as a whole is responsive to social, political, and economic change because opinion averages over the well and ill informed (Kinder, 2006; Page and Shapiro, 1992). In fact, it is these more informed citizens who contribute disproportionately to movement in the aggregate. Ill-informed citizens cancel each other out (Erikson et al., 2000). Although at the individual level individuals do not appear to have mastered complex political and economic ideas, in the aggregate they do appear to have done so. (Moreover, individual-level data is plagued with problems such as response bias and measurement error, problems which are diminished when responses are aggregated (Kramer, 1983).)

Thus, time series analysis offers a useful lens into political phenomena and the method has become prevalent as a research tool. **Figure 1** presents a count of the number of articles mentioning 'time series' in the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *Political Analysis* from 1970 to 2012. From a handful of articles in the early 1970s, there has been a steady increase to around 70 articles in recent years.

The increasing popularity of time series is in large part due to increased data availability. Economists have had over half a century of rich temporal data to draw upon, but analogous data were not available for political scientists until the past few decades. Systematic data collection programs have led to the increased availability of long t time series focusing on political phenomena. Hand in hand with the growing body of data are the development of research methods that allow the study of complex dynamic relationships.

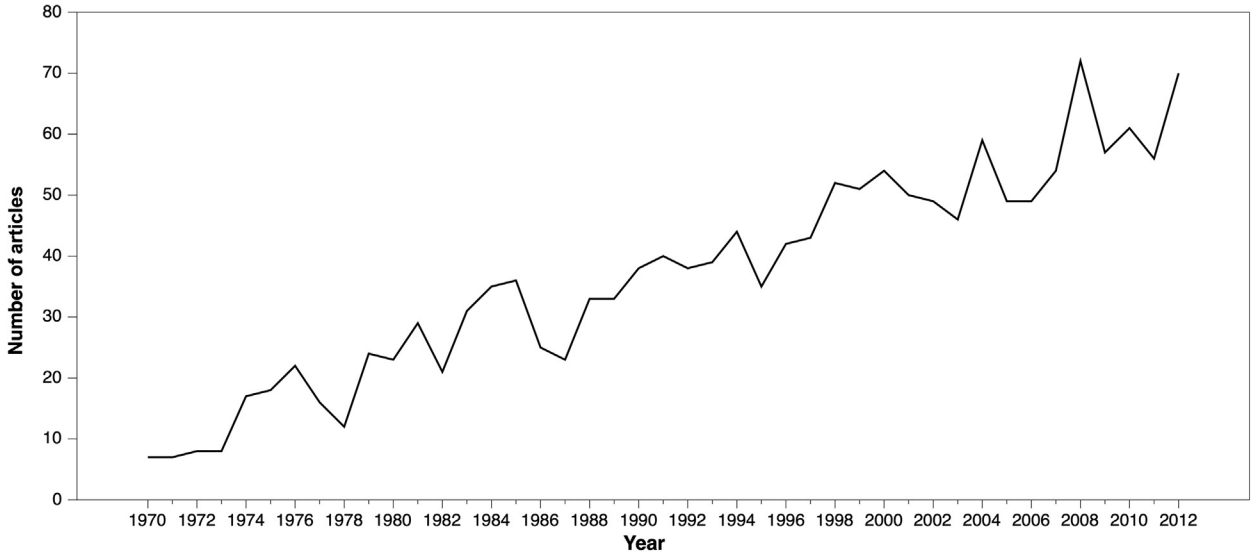

Figure 1 Time series articles per year, 1970–2012.

Advantage and Pitfalls of Time Series Analysis

Studying the evolution of data over time allows insights that prove more difficult or impossible for cross-sectional data. For one, studying just the univariate characteristics of a variable can be extremely useful. Knowing the way in which a time series's value depends on its past values means that we can make forecasts of the series's values ignorant of other predictors (Hibbs, 1977). Studying incidents of terrorism (e.g., Enders and Sandler, 2005) or the swings of the electoral pendulum (e.g., Lebo and Norpoth, 2011; Norpoth, 1996), for example, can lead to useful forecasts even without incorporating independent variables. The forecasting of elections may add independent variables to a model but will usually rely as well on identifying the pattern in a small number of data points and extrapolating to the next one. Such exercises have been an entertaining field roughly since the 1992 American presidential election (Campbell, 1992).

The ordering of observations also makes it possible to stretch beyond correlation and catch glimpses of causation. For example, if by incorporating past values of X_t we can improve our the predictions of Y_t beyond what the history of Y_t would predict, we can say that X_t 'Granger causes' Y_t (Granger, 1969; Freeman, 1983). Thus, researchers are able to study questions such as whether the economy affects economic judgments or vice versa (De Boef and Kellstedt, 2004). Multiple equation setups such as vector autoregression (VAR, Freeman et al., 1989; Sims, 1980), near-VARS (Enders, 2004), and seemingly unrelated regressions (e.g., Lebo et al., 2007) allow the study of complex dynamic relationships that might include reciprocal causality and feedback. In this way, time series analysis is protected from some critiques of observational studies developed in recent years as political scientists are paying more attention to the logic of causal inference.

But the ordering of observations also creates a range of complications. Much of the statistical theory is concerned with random samples of independent observations, but in time series, successive observations are usually not independent. In a cross-sectional survey, for example, the answers respondent #100 gives are unlikely to have any special relationship with the responses of person #101. The ordering of the observations is simply random. In time series, on the other hand, the value of a variable at time-point 100 is likely to be highly correlated with its value at time-point 101. This is a form of autocorrelation – cases nearer to each other in time will be more highly correlated with each other than observations more distant. This means that the statistical notion of identically and independently distributed observations is unlikely to hold with time series data.

This temporal dependence also means that popular techniques like ordinary least squares (OLS) encounter serious problems with time series data. This has led to an entirely separate branch of statistical methods (e.g., Box and Pierce, 1970; Box and Jenkins, 1976). Time series analysis generally takes careful steps to separate the two components of a time series: the deterministic and the stochastic. The deterministic portion is the part that is predictable based on past values and the stochastic component is what remains and might be partially explained by independent variables. Techniques like the Box and Jenkins (1976) transfer function approach build multivariate models only after first modeling how each variable depends upon its own past history (McCleary and Hay, 1980).

Political Science and Time Series Analysis

The development of time series analysis within political science has been very much aligned with popularity functions beginning in the early 1970s with the work of Mueller (1970, 1973) who studied the decay of job approval across the president's term. Mueller modeled a linear decline in popularity due to the fragmentation of a 'coalition of minorities' created during the presidential term. Subsequent research by

others focused on the appropriate way to model approval as a time series. Stimson (1976) argued the effect was quadratic, rather than linear, with approval declining after the start of the term but then experiencing an uptick near the end of the term. He attributed the initial decline to the disillusionment of uninformed citizens after the buildup of expectations surrounding the election. Similarly, the subsequent increase in approval at the end of the term is a result of increasing expectations leading into the next election. Kernell (1978) argued approval does not decline solely due to time, but rather the ebb and flow of approval can be attributed to political and economic events properly specified. Likewise, Monroe (1978) criticized Stimson for misspecifying the effects of the economy by using a time counter rather than objective economic indicators and actual expenditures.

Indeed, multivariate time series methods have developed in large part in tandem with models of economic voting. The compilation of monthly series of subjective economic evaluations – Gallup data in Britain and the Index of Consumer Sentiment in the United States – led to long debates regarding identifying the most important economic factors to voter decisions. On one hand, Key (1966) claims that voters punish or reward incumbents for the state of their pocketbooks, a view that supports personal retrospective evaluations. In contrast, Sanders (1993) used personal expectations as a key to forecasting the surprise Conservative victory in the 1992 British General Election. Looking at the effects of macroopinion on the national economy, MacKuen et al. (1992) see the electorate as forward-looking 'bankers' rather than retrospectively focused 'peasants' (Norpoth, 1996). Importantly, Clarke and Stewart (1994) pushed the discipline forward by sorting through the methodological issues in the data at issue. The concept of unit roots was long studied in economics but had not made its way to political science. When Clarke and Stewart (1994) allowed for and tested this possibility they found elements of truth in both theoretical approaches – the electorate is both prospective and retrospective. Further, Clarke and Stewart (1994) introduce the concepts of cointegration and error correction (Engle and Granger, 1987) to political scientists.

Series are cointegrated if there exists a mean-stationary linear combination of the variables. This expansion of the methodological toolkit allows the study of long-run equilibrium relationship between variables that would be obscured by traditional differencing techniques (see e.g., Clarke and Stewart, 1995; Clarke et al., 1998; Clarke and Lebo, 2003). As shocks drive cointegrated series apart, an error correction mechanism can be used to capture how quickly the series return to equilibrium. Several political and economic variables are cointegrated. In their study of the South African economy, Dunne and Vougas (1999) found economic growth and military spending are cointegrated, showing the negative effect protracted military conflicts have on economic health. Noneconomic variables may also be cointegrated. For example, the norm of consensus on the Supreme Court eroded around the time of the New Deal. Calderia and Zorn (1998) show the rate of concurrences and dissents, which are cointegrated, are affected by the decline of this norm. Another set of noneconomic cointegrated series is prime minister popularity and governing party support in Great Britain (Clarke et al., 2000). In all, a variety of political variables are cointegrated and error correction mechanisms allow for the study of these complex relationships.

As in many areas of political methodology, the 1990s saw a boom of new methods and approaches to time series analysis. Aside from the expansion of tools, important work began studying the ways in which political data differed from the majority of economic time series. So many of the methods used by political science had simply been imported from work in economics without much thought regarding how the basic data generating processes between the fields might differ.

For example, Box-Steffensmeier and Smith (1996) showed the need for fractional integration methods, a small corner of econometric research, to be widely used with political data. Although series can be analyzed in level form – without transforming its values – if it is not mean-reverting the failure to account for this may result in spurious regressions where it is falsely concluded that a relationship exists between two factors.

One way to correct for this is to first difference the data, however this, too may be insufficient. While differencing the data provide significant improvement over the level-form analysis of nonstationary series, it may overcorrect by overdifferencing the data if the series is not truly a unit root, building in an moving average (MA) parameter that is not present in the data generating process. The best way to avoid overdifferencing while still making a series stationary is to fractionally difference the data. Fractional integration is theoretically motivated when heterogeneity at the individual level exists due to individual-level variation in the persistence of new information (Granger and Joyeux, 1980). Some individuals have long or perfect memory (or a strong autoregressive process) and others less so (Box-Steffensmeier and Smith, 1996). This heterogeneity at the individual level produces aggregate series that are neither stationary nor unit roots, but instead have a mix of properties. Several variables of interest to political scientists are fractionally integrated, including partisanship (Box-Steffensmeier and Smith, 1998) and presidential approval (Lebo et al., 2000). Autoregressive fractionally integrated moving average techniques allow for a series to have long, but not perfect memory. Shocks to series can persist but eventually be discounted over time.

Other Approaches to Time Series Research – Volatility, Time-Varying Parameters, and Dynamic Effects

There are multiple ways to trace the dynamics of a series or a system of series. Because the relationships studied by political scientists are extremely complex, there are many reciprocal relationships and feedback loops. One way to study the dynamic interrelationship between variables is to use an impulse response function (IRF). IRFs trace the cumulative effect of a shock to one variable as it moves its way through a VAR system (Sims, 1980). For example, Box-Steffensmeier and Smith (1996) use an IRF to determine the persistence of shocks to macropartisanship. These changes to the series decay slowly and persist the effect can be felt for years. Similar

patterns are evidenced in the Republican and Democratic series, indicating that macropartisanship does change on the scale of years rather than decades or months. IRFs have also been used to study political bureaucratic adoption (Wood and Waterman, 1993), presidential rhetoric and economic performance (Wood et al., 2005), and fatalities on Israeli–Palestinian relations (Jaeger and Paserman, 2005).

Another branch of analysis goes beyond studying how time series rise and fall and looks at changes in the volatility. Although variables are often assumed to have constant error variance across time, in reality many go through different periods of stability and volatility. Volatility clustering, first studied in the financial markets, refers to that phenomena that "[w]hen volatility is high, it is likely to remain high, and when it is low it is likely to remain low" (Engle, 2003). Autoregressive conditional heteroskedasticity (ARCH) and the generalized form (GARCH) allow for modeling the variance as well as the mean of a series. This has been used primarily to study volatility in popularity functions, e.g., Gronke and Brehm (2002), using a modified ARCH approach, found volatility in presidential approval to have increased over time. The effect of events on approval volatility, however, is conditional on the partisan identification of the respondent. Positive presidential events create cognitive dissonance in out-party identifiers, leading to increased volatility (Kriner and Schwartz, 2009). Mounting wartime casualties may even increase volatility among the president's most strident supporters, as was the case with Franklin Roosevelt and federal assistance recipients during World War II (Kriner, 2006). Just as one would expect volatility to be low among those most predisposed to support the president, one would also expect macropartisanship to be relatively stable over time. However, electoral cycles and waning party identification create periods of high volatility in aggregate partisan identification (Maestas, 1997). If periods of high volatility in either presidential popularity or macropartisanship were to coincide with an election, the outcome would be much more difficult to forecast.

Other lines of research are opened when one considers that the relationship between variables may change over time in interesting and predictable ways due to circumstances such as elections, recessions, wartime, or other factors. Three common ways to test for dynamic relationships between time series are moving windows, Kalman Filter, and dynamic conditional correlations (DCC, Lebo and Box-Steffensmeier, 2008). To use a moving window estimator design, the researcher first selects the size of the window (s), and runs a model using just the first s observations. The process is then repeated in an iterative fashion as the window moves across the data, using observations 2 to $s+1$, 3 to $s+2$, etc. After model estimation is complete, the correlation (or regression coefficients) provides insight into the changing nature of the relationship. Moving windows are useful when the time period being studied is shorter or if the researcher has strong expectations of when the relationship should change over time (e.g., presidential administrations). Moving windows have been used to study the relationship between international trade and the opportunistic timing of elections (Kayser, 2006), social spending and partisanship in OECD countries (Kwon and Potusson, 2010), and economic health's effect on policy liberalism (Ferguson et al., 2013).

The Kalman Filter is a recursive system that starts with some data, points 1 to s, and then adds more data iteratively (e.g., Bond et al., 2003). Unlike the moving window, past observations are not dropped out of the analysis as new observations are added. Eventually, the coefficients converge upon the estimate one would obtain if the relationship was time invariant. Thus, a Kalman Filter allows us to see the evolution of a coefficient over time. Stimson et al.'s (1995) model of policy responsiveness, dubbed 'dynamic representation,' was estimated using a Kalman Filter. As public opinion becomes more liberal, governmental institutions respond with more liberal policy outputs. Similarly, public opinion about the president's job performance, a factor so heavily colored by partisanship that it should have little explanatory power, affects the president's legislative success (Bond et al., 2003).

A relatively new technique for political researchers is the dynamic conditional correlation model developed by Engle (2002) and introduced to political science by Lebo and Box-Steffensmeier (2008). Building on multivariate-GARCH models, DCC models allow researchers to study the direction and strength of the relationship while taking into account the volatility in the times series. One can find periods of no correlation, as well as positive and negative correlation, between two or more series. Consumer sentiment and presidential approval, for example, are two series that have nonconstant correlation (Lebo and Box-Steffensmeier, 2008). The DCC method's ability to tell what a correlation is *now* shows that, over 30 years of data, the height of correlation between presidential approval and economic sentiment was during the 1992 presidential election – it was, in fact, 'the economy, stupid.' Wars and terrorist events have sometimes even produced negative correlations between economic and political assessments.

Event History Models

Many of the data that are of interest to political scientists denote change from one 'regime' to the next. Beyond just the transition from one state to another, the factors precipitating the change may also be of interest. This leads to research questions that are focused on the timing and change as well as the ultimate outcome. Event history models, also known as duration or survival models, have grown in popularity in the social science literature since the 1990s, culminating in the publication of Box-Steffensmeier and Jones' (2004) key book on the subject (see also Box-Steffensmeier and Jones, 1997).

Initially, the temporal structure of event data was treated as a nuisance, rather than being of substantive interest. As a result the temporal ordering was ignored and post hoc corrections were applied to correct for autocorrelation in the data. In the early 1990s, Bartels and Brady encouraged political scientists to incorporate more techniques specifically designed to analyze event history data (1993). Heeding their advice, researchers began to import methods from biostatistics to measure the 'risk' that an event would occur and the likelihood of an observation's 'survival' at the end of the study.

There are several problems inherent in analyzing event data. The data are always nonnegative, leading traditional estimation methods to return nonsensical predictions. Furthermore, survival at time t is conditional on the observation having not experienced the event by time $t-1$, so this dependence must be accounted for. Finally, there is the problem of censoring. At the end of the period of study, some observations at risk will not have experienced the event. This leads to biased estimates if not corrected.

This durational framework has been applied to a variety of political science subfields. King et al. (1990) study cabinet government dissolution with a unified model that allows them to account for both the particular attributes of the cabinet as well as the stochastic event process. Although there was disagreement in the literature about the proper specification of the hazard rate, this model was later applied to the duration of governmental leaders. Alt and King (1994) found that the risk of experiencing the event, in this case the removal of the leader from power, increased over time just as it did for cabinets. While these studies included characteristics of the political system, economic factors were omitted. The addition of economic indicators such as the inflation and unemployment rates improved the predictive power of the models and showed that poor economic health contributes to the collapse of governments (Warwick, 1992).

The duration of armed conflict or peace is common event history topics in international relations. Many conflict scholars are interested in the role third-party intervention plays in the duration of civil and interstate wars. The effect of the international community on duration is conditional on both the actor and the actor's relationship with the warring parties. For instance, Aydin and Regan (2012) found that engagement by third-party state increases the duration of civil wars if the states support opposite sides of the conflict. Duration is decreased, however, if the states unilaterally support the same side. Similarly, the outcome of a conflict, e.g., negotiated settlement, military victory, and the presence of international peacekeeping forces, have also been shown to affect the likelihood of a conflict reignigting (Mason et al., 2011).

In the U.S. context, duration models have been used to study policies, institutions, and behavior. The widely used Berry and Berry model of policy diffusion was developed in the event history context (Berry and Berry, 1990, 1992, 1994). The model allows for both internal and regional characteristics to influence the adoption of new policies and has since been applied a variety of policy areas, such as the adoption of climate protection and antismoking legislation (Krause, 2010; Pacheco, 2012, respectively). Committee tenure (Katz and Sala, 1996) and tenure in office more generally (Jones, 1994) are influenced by electoral systems and the calendar. The political environment also affects the duration of presidential appointee confirmation, with the hazard function decreasing as ideological distance between the president and Congress increases (Shipan and Shannon, 2003; Segal et al., 2010). An event history framework has also been applied to the timing of voting decisions. Box-Steffensmeier and Sokhey (2010) show that political engagement, ideology, and strength of partisanship all increase the hazard rate thereby reducing the time it takes for a voter to go from 'undecided' to 'decided.'

Time Series and Cross-Sectional Analyses

In many instances, data can be collected where cases are spread both within and across time points. Cross-section data have multiple units observed at a given time point but PCSTS involves multiple cross-sections observed at different points in time. Since OLS assumes errors are independently and identically distributed, the PCSTS framework requires its own set of methods. PCSTS methods attempt to address the complex multilevel structure of multiple units observed over time. A unit x_{it} is correlated with x_{jt} by virtue of being observed at the same point. Depending on the type of cross-sectional data, x_{it} may also be correlated with x_{it-1} if the same unit is observed at multiple time points. There are several types of PCSTS data, and each structure poses different challenges to data analysts.

For starters, a true panel design involves an identical set of cases repeated over multiple time points. Panels are frequently comprised of survey respondents who are interviewed a small number of times such as a special subset of election study respondents. Panel election studies began with the Columbia school (e.g., Lazarsfeld, 1948), although repeated interviews during the course of a single campaign limited the representativeness of the sample. In addition to problems with generalizability, election panels can also suffer from attrition. A variety of factors, from frequency of contact, method of contact, and even mortality can all affect the recontact rates.

Alternatively, the same set of countries observed over the same set of years also represents a type of panel design. As with most time series methods, models for analyzing panel data were developed to handle econometric data with a large enough N to capitalize on the estimator's asymptotic properties. Methods such as dynamic panel analysis, Markov-switching models, differencing, and lagged dependent variables are available.

Deviations from a true panel fall into a range of types we can classify as pseudo-panels in which there are multiple cross-sections of data but not all cases are found in each wave. One variety of pseudo-panel structures is the unbalanced panel where some cases are observed more than once, but each observation does not appear in every time point, e.g., Carson et al. (2010) study Members of Congress running for reelection from 1956 to 2004. The range of methods available to account for time dependencies quickly dwindles here. Since some cases appear in the data but once, differencing data of using lagged variables will drop out those case and potentially cause problems of selection bias. Fixed effects by time point or panel-corrected standard errors are common, but imperfect, solutions (Beck and Katz, 1995).

Two last types of pseudo-panels are the rolling and repeated cross-sectional (RCS) designs. In both of these cases each observation appears a single time in a data set that consists of multiple waves. In a rolling design, typically an election study, the survey designers choose their set of respondents at the beginning and then roll out the survey to subsets of respondents at prespecified dates. The American National Election Studies conducted the first rolling cross-section in 1984 and it was later adopted by the Annenberg Election Study and the British Election Study.

Repeated cross-sectional designs include the wider set of designs in which observations are collected *post hoc* and assembled into a larger data set (Lebo and Weber, 2014). An RCS data set may include hundreds of months of individual-level data from consecutive public opinion surveys, or it could be Supreme Court cases nested with terms of the Court, or some other case for which the time the data are collected requires attention in the modeling framework. Lebo and Weber (2014) demonstrate how RCS data can be studied at two levels of analysis – aggregate and individual level – in a multilevel model that includes time-varying parameters. In any study where cases are nested within distinct time periods, researchers should pay close attention to the error structure to be sure model estimates are trustworthy.

Conclusion

Overall, the wide variety of data structures requires researchers to think carefully about how time plays a role in the phenomenon they study. Time can either be treated as a nuisance or as a substantively interesting variable across a variety of political science subfields and a wide array of research questions. Those treating time as substantively interesting have continued to advance time series analysis since its introduction to the field in the early 1970s. This is in part due to the increased availability of a range of data spanning sufficiently long time periods. In addition to more readily available data, traditional data sources are being measured with increasing frequency. For example, presidential approval was once only available quarterly yet is now available daily from Gallup. Similarly, the Annenberg Election Survey (The survey will resume in 2016.) tracks opinion about campaigns, candidates, and issues on a daily basis.

To date, most methods used to analyze temporal effects have been imported by political scientists from other social science disciplines. Although these methods are theoretically applicable to political phenomena, they may require tweaking to adapt them to political science data structures. Researchers should pay special attention to the properties of their data before wholesale adoption of imported techniques. Monte Carlo studies are an underutilized way to determine how an estimator responds to changes in level of integration, seasonality, and series length – all of which vary between data sources. Having a better grasp on how the methods respond to political science data allows a more informed decision about the appropriate estimator.

It should come as no surprise that unpacking the relationship between various political phenomena over time is challenging as many factors are closely intertwined. Time series analysis provides a unique opportunity to approach questions of causality using observational data. While Granger causality tests can identify if changes in one variable of interest temporally precede changes in another, it is a common misperception that these tests alone are enough to rule out endogeneity (Charemza and Deadman, 1997). Indeed, a battery of tests exist to determine if a one variable is weakly or strongly exogenous to another. If one series is weakly exogenous, there is no reciprocal causality and modeling is simpler. However, if there is unmodeled reciprocal causality, estimates become biased and inefficient. Thus researchers should explore exogeneity tests and utilize multiple equation (near-)VARs to disentangle reciprocally causal relationships where they exist.

Overall there is much to be learned from incorporating time into studies of politics. The many statistical problems inherent in data that evolve over time can wreak havoc on our understanding of political relationships. But, even setting those aside, modeling temporal effects provides interesting insights that cannot be gained from the single snapshot in time that cross-sectional analyses provide.

See also: Attitudes, Political and Public Opinion; Congress: United States; Economic Evaluations and Electoral Consequences; Event History Analysis; Party Identification and Nonpartisanship; Political Science and Polling; Survival Analysis: Introduction; Time Series: Advanced Methods; Time Series: Cointegration; Time Series: General; Time Series: State Space Methods; Time Series: Unit-Roots and Nonstandard Limiting Distributions; Voting, Explanations of: Social Class.

Bibliography

Alt, James, King, Gary, 1994. Transfers of governmental power: the meaning of time dependence. Comparative Political Studies 27 (2), 190–210.

Aydin, Aysegul, Regan, Patrick M., 2012. Networks of third-party interveners and civil war duration. European Journal of International Relations 18, 573–597.

Bartels, Larry M., Brady, Henry E., 1993. Political Science: The State of the Discipline II. American Political Science Association chapter. The State of Quantitative Political Methodology.

Beck, Nathaniel, Katz, Jonathan N., 1995. What to do (and not to do) with time-series cross-section data. American Political Science Review 89 (3), 634–647.

Berry, Frances Stokes, Berry, William D., 1990. State lottery adoptions as policy innovations: an event history analysis. The American Political Science Review 84 (2), 395–415.

Berry, Frances Stokes, Berry, William D., 1992. Tax innovation in the states: capitalizing on political opportunity. American Journal of Political Science 36, 715–742.

Berry, Frances Stokes, Berry, William D., 1994. The politics of tax increases in the states. American Journal of Political Science 38, 855–859.

Bond, Jon R., Fleisher, Richard, Wood, B. Dan, 2003. The marginal and time-varying effect of public approval on presidential success in Congress. Journal of Politics 65 (1), 92–110.

Box, George E.P., Jenkins, Gwilym M., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box, George E.P., Pierce, David A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association 65 (332), 1509–1526.

Box-Steffensmeier, Janet M., Sokhey, Anand E., 2010. Event history models. In: Leicht, K.T., Jenkins, J.C. (Eds.), Handbook of Politics: State and Society in Global Perspective. Spring Science.

Box-Steffensmeier, Janet M., Jones, Bradford S., 1997. Time is of the essence: event history models in political science. American Journal of Political Science 41 (4), 1414–1461.

Box-Steffensmeier, Janet M., Jones, Bradford S., 2004. Event History Modeling: A Guide for Social Scientists. Cambridge University Press.

Box-Steffensmeier, Janet M., Smith, Renée M., 1996. The dynamics of aggregate partisanship. American Political Science Review 90, 567–580.

Box-Steffensmeier, Janet M., Smith, Renee M., 1998. Investigating political dynamics using fractional integration methods. American Journal of Political Science 661–689.

Box-Steffensmeier, Janet M., DeBoef, Suzanna, Lin, Tse-Min, 2004. The dynamics of the partisan gender gap. American Political Science Review 98, 515–528.

Calderia, Gregory A., Zorn, Christopher, 1998. Of time and consensual norms in the supreme court. American Journal of Political Science 42 (3), 874–902.

Campbell, Angus, Converse, Philip E., Miller, Warren E., Stokes, Donald E., 1960. The American Voter. Wiley, New York.

Campbell, James E., 1992. Forecasting the presidential vote in the states. American Journal of Political Science 36, 386–407.

Carson, Jamie L., Koger, Greg, Lebo, Matt J., Young, Everett, 2010. The electoral costs of party loyalty in Congress. American Journal of Political Science 54 (3), 598–616.
Charemza, Wojciech, Deadman, Dereck, 1997. New Directions in Economic Practice: General to Specific Modeling, Cointegration, and Vector Autoregression. Edward Elgar.
Clarke, Harold D., Stewart, Marianne C., 1994. Prospections, retrospections and rationality: the 'bankers' model of presidential approval reconsidered. American Journal of Political Science 38, 1104–1123.
Clarke, Harold D., Stewart, Marianne C., 1995. Economic evaluations, prime ministerial approval and governing party support: rival models reconsidered. British Journal of Political Science 25, 145–170.
Clarke, Harold D., Lebo, Matthew, 2003. Fractional (co)integration and governing party support in britain. British Journal of Political Science 33 (02), 283–301.
Clarke, Harold D., Ho, Karl, Stewart, Marianne C., 2000. Major's lesser (not minor) effects: prime ministerial approval and governing party support in Britain since 1979. Electoral Studies 19 (2–3), 255–273.
Clarke, Harold D., Stewart, Marianne C., Whiteley, Paul F., 1998. New models for new labour: the political economy of labour party support, January 1992–April 1997. American Political Science Review 559–575.
Converse, Philip E., 1964. The nature of belief systems in mass publics. In: Apter, David E. (Ed.), Ideology and Discontent, 206–261. University of Michigan Press, Ann Arbor.
De Boef, Suzanna, Kellstedt, Paul M., 2004. The political (and economic) origins of consumer confidence. American Journal of Political Science 48 (4), 633–649.
Dunne, Paul, Vougas, Dimitrios, 1999. Military spending and economic growth in South Africa: a causal analysis. The Journal of Conflict Resolution 43 (4), 521–537.
Enders, Walter, 2004. Applied Econometric Time Series, second ed. Wiley, New York.
Enders, Walter, Sandler, Todd, 2005. After 9/11 is it all different now? Journal of Conflict Resolution 49 (2), 259–277.
Engle, Robert F., 2002. Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20 (3), 339–350.
Engle, Robert F., 2003. Risk and Volatility: Econometric Models and Financial Practice. Nobel Foundation.
Engle, Robert F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica 55, 251–276.
Erikson, Robert S., MacKuen, Michael B., Stimson, James A., 2000. Electoral Studies 19, 295–312.
Erikson, Robert S., MacKuen, Michael B., Stimson, James A., 2002. The Macro Polity. Cambridge University Press, New York.
Ferguson, Grant, Kellstedt, Paul M., Linn, Suzanna, 2013. How does the economy shape policy preferences? Electoral Studies 32 (3), 544–550.
Freeman, John R., 1983. Granger causality and the time series analysis of political relationships. American Journal of Political Science 27, 327–358.
Freeman, John R., Williams, John T., Lin, Tse Min, 1989. Vector autoregression and the study of politics. American Journal of Political Science 33, 842–877.
Granger, Clive W.J., Joyeux, Roselyne, 1980. An introduction to long-memory time series models and fractional differencing. Journal of Time Series Analysis 1 (1), 15–29.
Granger, Clive W.J., 1969. Investigating causal relations by econometric models and cross spectral methods. Econometrica 37, 424–438.
Gronke, Paul, Brehm, John, 2002. History, heterogeneity, and presidential approval: a modified ARCH approach. Electoral Studies 21 (3), 425–452.
Hibbs Jr., Douglas A., 1977. On analyzing the effects of policy interventions: Box-Jenkins and Box-Tiao vs structural equation models. Sociological Methodology 8, 137–179.
Jaeger, David A., Paserman, M. Daniele, 2005. The cycle of violence? An empirical analysis of fatalities in the Palestinian-Israeli conflict. The American Economic Review 98 (4), 1591–1604.
Jones, Bryan D., 1994. Reconceiving Decision-making in Democratic Politics: Attention, Choice, and Public Policy. University of Chicago Press, Chicago.
Katz, Jonathan N., Sala, Brian R., 1996. Careerism, committee assignments, and the electoral connection. American Political Science Review 90, 21–33.
Kayser, Mark A., 2006. Trade and the timing of elections. British Journal of Political Science 36 (3), 437–457.
Kernell, Samuel, 1978. Explaining presidential popularity. American Political Science Review 72, 506–522.
Key Jr., V.O., 1966. The Responsible Electorate. Harvard University Press, Cambridge.
Kinder, Donald R., 2006. Belief systems today. Critical Review 18 (1–3), 197–216.
King, Gary, Alt, James E., Burns, Nancy E., Laver, Michael, August 1990. A unified model of cabinet dissolution in parliamentary democracies. American Journal of Political Science 34, 846–871.
Kramer, Gerald H., 1983. The ecological fallacy revisited: aggregate versus individual level findings on economics and elections and sociotropic voting. American Political Science Review 77, 92–111.
Krause, Rachel M., 2010. Policy innovation, intergovernmental relations, and the adoption of climate protection initiatives by U.S. Cities. Journal of Urban Affairs 33 (1), 45–60.
Kriner, Douglas, 2006. Examining variance in presidential approval: the case of FDR in World War II. Public Opinion Quarterly 70 (1), 23–47.
Kriner, Douglas, Schwartz, Liam, 2009. Partisan dynamics and the volatility of presidential approval. British Journal of Political Science 39 (3), 609–631.
Kwon, Hyeok Young, Potusson, Jonas, 2010. Globalization, labour power, and partisan politics revisited. Socio-Economic Review 8 (2), 251–281.
Lazarsfeld, Paul F., 1948. People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign, second ed. Columbia University Press, New York.
Lebo, Matthew, Norpoth, Helmut, 2011. Yes, prime minister: the key to forecasting British elections. Electoral Studies 30 (2), 258–263.
Lebo, Matthew J., McGlynn, Adam J., Koger, Greg, 2007. Strategic party government: party influence in Congress, 1789–2000. American Journal of Political Science 51 (3), 464–481.
Lebo, Matthew J., O'Geen, Andrew J., 2011. The president's role in the partisan Congressional arena. Journal of Politics 73 (3), 718–734.
Lebo, Matthew J., Weber, Christopher, 2014. A solution to the repeated cross sectional design. American Journal of Political Science 0 (0), 1–17.
Lebo, Matthew J., Box-Steffensmeier, Janet M., 2008. Dynamic conditional correlations in political science. American Journal of Political Science 52 (3), 688–704.
Lebo, Matthew J., Walker, Robert W., Clarke, Harold D., 2000. You must remember this: dealing with long memory in political analyses. Electoral Studies 19 (1), 31–48.
MacKuen, Michael B., Erikson, Robert S., Stimson, James A., 1989. Macropartisanship. American Political Science Review 83, 1125–1142.
MacKuen, Michael B., Erikson, Robert S., Stimson, James A., 1992. Peasants or bankers: the American electorate and the U.S. economy. American Political Science Review 86, 597–611.
Maestas, Cherie D., 1997. Macropartisan volatility: a new perspective using GARCH modeling. In: American Political Science Association Meetings.
Mason, T. David, Gurses, Mehmet, Brandt, Patrick T., Quinn, Jason Michael, 2011. When civil wars recur: conditions for durable peace after civil wars. International Studies Perspectives 12, 171–189.
McCleary, Richard, Hay, Richard A., 1980. Applied Time Series Analysis. Sage, Beverly Hills.
Monroe, Kristen R., 1978. Economic influences on presidential popularity. Public Opinion Quarterly 42 (3), 360–369.
Mueller, John, 1970. Presidential popularity from Truman to Johnson. American Political Science Review 65, 18–34.
Mueller, John, 1973. War, Presidents and Public Opinion. Wiley, New York.
Norpoth, Helmut, 1996. Presidents and the prospective voter. Journal of Politics 58 (3), 776–792.
Pacheco, Julianna, 2012. The social contagion model: exploring the role of public opinion on the diffusion of antismoking legislation across the american states. Journal of Politics 74 (1), 187–202.
Page, Benjamin I., Shapiro, Robert Y., 1992. The Rational Public: Fifty Years of Trends in Americans' Policy Preferences. University of Chicago Press, Chicago.
Sanders, David, 1993. Forecasting the 1992 British general election outcome: the performance of an "economic" model. British Elections & Parties Yearbook 3 (1), 100–115.
Schrodt, Philip A., Gerner, Deborah J., 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982–92. American Journal of Political Science 825–854.
Segal, Jeffrey A., Cameron, Charles M., Key, Ellen M., 2010. The politics of scandal. In: The Annual Meeting of the American Political Science Association.
Shipan, Charles R., Shannon, Megan L., 2003. Delaying justice(s): a duration analysis of supreme court confirmations. American Journal of Political Science 47 (4), 654–668.
Sims, Christopher R., 1980. Macroeconomics and reality. Econometrica 48, 1–48.
Stimson, James A., 1976. Public support for American presidents a cyclical model. Public Opinion Quarterly 40 (1), 1–21.
Stimson, James A., MacKuen, Michael B., Erikson, Robert S., 1995. Dynamic representation. American Political Science Review 89 (3), 543–565.
Warwick, Paul, 1992. Economic trends and government survival in West European parliamentary democracies. American Political Science Review 86 (December), 875–886.
Wood, B. Dan, Waterman, Richard W., 1993. The dynamics of political-bureaucratic adaptation. American Journal of Political Science 37 (3), 497–528.
Wood, B. Dan, Owens, Chris T., Durham, Brandy M., 2005. Presidential rhetoric and the economy. Journal of Politics 67 (3), 627–645.

Popper, Karl Raimund (1902–94)

Peter Schroeder-Heister, Universität Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Karl Popper established falsificationism as a basic orientation in modern philosophy of science. Falsificationism, as opposed to verificationism, claims that the main research activity is to invalidate a theory by observation or experiment; confirmation or corroboration of a theory relies on the failure of attempted falsifications. Following Hume, Popper denied the validity of induction from limited observations to universal regularities. For Popper, the dynamics of theories consists of replacing falsified hypotheses with better ones, with the aim of approaching objective truth. Popper was an objectivist and realist, strongly opposed to subjectivistic approaches; he considered probability as a strictly objective measure rather than a degree of (rational) belief. Popper later adopted a form of evolutionary epistemology, which considers all life to be problem solving by trial and error. Most fundamentally, he extended falsificationism to the methodology of critical rationalism, encompassing all areas of reasoning, including metaphysics. Popper defended indeterminism as the basis of free human action. He proposed a triad of causally interacting worlds of body, mind, and contents of thought. In his social and political philosophy, he was opposed to historicism as a deterministic view of history, and proclaimed the idea of an open society as a social framework for responsible action.

Popper's Life

Karl Raimund Popper was born in Vienna, Austria, on 28 July 1902. His father, Simon (Siegmund Carl) Popper (1856–1932), who had come from Bohemia to Vienna, was a well-established lawyer with his offices and family residence in the center of Vienna. Karl Popper's mother, Jenny Popper (neé Schiff) (1864–1938), came from a musical family and was a gifted pianist herself. Both Simon and Jenny Popper were Jews by origin, but converted later in their lives to Protestantism. Besides their son Karl, they had two older daughters, Dora (1893–1932) and Anna Lydia (1898–1975).

Karl Popper left high school (Franz Josef Gymnasium) at the age of 16 without a degree, moved out of his parents' home and attended lectures at the University of Vienna on various subjects, in particular mathematics, philosophy, physics, and psychology. Besides occasional jobs to earn his living (e.g., coaching American students), Popper was engaged in social activities. For some time, he worked as an assistant of Alfred Adler on neglected children. Hints in Popper's autobiography indicate that this connection (through which he got acquainted with Adler's 'individual psychology') contributed to Popper's later critique of psychoanalysis as a 'pseudoscience.' In his youth, Popper had a strong favor for democratic socialism. For a few months, he even considered himself a communist, from which he refrained, in 1919, after witnessing a shooting incident in which socialist workers were killed by police forces. This experience of a useless sacrifice of lives demonstrated to him the futility of ideologies. A second, more intellectual experience in 1919 was his taking interest in Eddington's test of Einstein's eclipse predictions carried out at an expedition of the Royal Astronomical Society, whose results gained widespread publicity at the time. During a total solar eclipse, which was observed from West Africa and Brazil, a shift of apparent position of stars in the presence of the sun's gravitational field could be established, as predicted by the general theory of relativity. This coined Popper's later insistence that scientific theories develop by exposing them to possibly falsifying experience. Around that time, Popper visited a lecture by Einstein in Vienna.

The young Popper also tried manual work. From 1922 to 1924, he was an apprentice to a cabinet maker. At the same time, he passed the high school exams ('Matura') as an external participant and formally enrolled as a student at the University of Vienna. He also considered becoming a musician. After being disappointed with the sort of music cultivated in the Schönberg circle, he enrolled as a student of Church Music at the Vienna Academy of Music ('Konservatorium') and stayed there for 1 year. He was admitted upon submission of a fugue for organ in F sharp minor. Besides his scientific studies at the University of Vienna, in 1924 he qualified as a primary school teacher. Since no appropriate teacher's post was available at the time, Popper worked as a social worker with neglected children.

In 1925, Popper became a student at the Pedagogic Institute of the City of Vienna, where he stayed until 1927. This institute, in which Popper was admitted to due to his status of a social worker, had been newly founded. It was linked to the University, though formally independent. In today's terms, its purpose was to scientifically accompany and evaluate the school reform movement. This school reform movement had been politically supported by the Austrian social democrats since 1919 and was in 1925 still very strong in Vienna. It was characterized by ideas of abolishing school learning as an accumulation of facts in favor of developing a more autonomous and participating attitude. The influence of these ideas on Popper should not be underestimated. His later criticism of what he calls the 'bucket theory of knowledge,' i.e., the inductive generation of knowledge by collecting facts, is related to pedagogic proposals on how learning proceeds or should proceed. Actually, Popper's first published papers were in this field and appeared in journals belonging to the school reform movement. The very first one *Über die Stellung des Lehrers zu Schule und Schüler* (*On the teacher's attitude toward school and students*, 1925) is very significant in this respect.

At the pedagogic institute, Popper met Josefine Henninger (1906–85), who was training as a teacher and who was to become his wife in 1930. (They had no children.) He also met the psychologist Karl Bühler and the philosopher Heinrich Gomperz, both of them professors at the University. They were of crucial importance to Popper's intellectual development. It was through discussions with Gomperz and reading his works that Popper shifted away from any psychologistic approach to the philosophy of science. From Bühler, he adopted the distinction of expressive, signaling, and descriptive functions of language, which still figured prominently in Popper's late metaphysical writings.

In 1928, Popper completed his PhD in Psychology with Karl Bühler and Moritz Schlick as examiners. The subject of the thesis was *Zur Methodenlehre der Denkpsychologie* (*On the Methodology of the Psychology of Thinking*, 1928), a work that combined Popper's interest in psychology with the methodological bias, which he had finally shifted to during his time at the Pedagogic Institute. In 1929, Popper qualified as a teacher for secondary schools ('Hauptschulen'), and in 1930, he became a teacher in Vienna.

Since 1922, Popper had been in contact with members of the Vienna circle, a group of philosophers around Moritz Schlick who propagated what they called 'logical positivism,' a version of empiricism, which made strong use of modern mathematical logic as a tool of philosophical analysis. Mathematical logic had already been developed by Gottlob Frege (1848–1925) and Bertrand Russell (1872–1970) decades ago, but had not been appropriately appreciated within the established academic philosophy. Though never a member of the circle, Popper attended its sessions occasionally and was in contact with several of its members, in particular Rudolf Carnap, Victor Kraft, and Herbert Feigl. Although he shared certain basic attitudes of the Vienna circle, in particular the plea for a scientifically oriented philosophy, he was highly critical of other tenets, such as the theory of meaning and the form of empiricism propagated. Motivated by Feigl, he put this criticism into a large manuscript in two volumes *Die beiden Grundprobleme der Erkenntnistheorie* (*The Two Basic Problems of Epistemology*, published in 1979, the second volume only as a fragment). A shorter version of the second volume of this manuscript was published in 1934 as *Logik der Forschung* (*Logic of Scientific Discovery*), which was to become his main work.

The book was well received. Reviews appeared in many journals. Even Einstein wrote Popper a letter after the book was sent to him through intermediates. In 1935 and 1936, Popper received invitations to England, where he lectured at Bedford College and Imperial College, London, at the London School of Economics and at Cambridge. He made a lot of contacts. Of most importance for his later career were those with Ernst Gombrich and Fritz Hayek. However, attempts to find a position in England did not materialize, and in 1937 he emigrated with his wife to New Zealand, accepting an offer of a lectureship at Canterbury University College.

During the time in exile, Popper intensified his interest in social and political philosophy and in the philosophy of history, besides continuing his research in logic and philosophy of science. From that time on, he published mainly in English. He wrote a series of articles *The Poverty of Historicism* (Popper, 1957a originally 1944–45); and, more importantly, his two-volume classic *The Open Society and Its Enemies* (Popper, 1945). For both works, he had to struggle to find a publisher. Due to the *Open Society*, he acquired his initial postwar reputation and, in particular, his post, from which he carried out all subsequent work. Through the intervention of Hayek who, together with Gombrich, had made the publication of the *Open Society* with the publisher Routledge possible, he was offered a readership in Logic and Scientific Method at the London School of Economics in 1945, which Popper took up early in 1946, shortly after publication of the *Open Society*. In 1949, he received a professorship of the same title at the London School of Economics, which he held until his retirement in 1969.

During the first 5 years of his stay in Britain, he was sort of a 'star.' The *Open Society* made him a quite popular theoretician of political liberalism. It has often been said that this book filled some intellectual vacuum in the postwar period. Popper was invited to many places and, at the same time, had a substantial audience at the London School of Economics. In the 1950s, he shifted more to working from home with less institutional involvement. At the same time, his research interests moved away from political and social philosophy and back toward the subject of the *Logik der Forschung*, i.e., the methodology of the natural sciences. However, the first English edition of this book appeared only in 1959, with a lot of material added compared with the German original. Many other books and plenty of papers followed. Up to his death, Popper produced an enormous scientific output and had a vast correspondence. From the mid-1960s on, Popper received a great number of prizes and honorary degrees in many countries, including Knighthood and Membership of the Royal Society in Britain. He died on 17 September 1994 in London. His ashes were buried in Vienna, where his wife had died in 1985.

Popper's Work

Popper's work covers nearly all areas of theoretical philosophy. His overall profile is that of a great philosophical classic. However, most important are his contributions to the methodology of science, where he established falsificationism as one of the basic orientations in the modern philosophy of science.

The basic claims of falsificationism are put forward in the *Logik der Forschung* (1934, official year given in the book: 1935). Most of Popper's later methodological contributions can be viewed as elaborations or amendments of what was already stated in his first book. Even Popper himself treated the *Logik der Forschung* like a classical text, with all changes in the first English translation (*The Logic of Scientific Discovery* 1959) as well as in the later 10 German and 10 English editions, as a system of footnotes and appendices to the 1934 edition.

Falsificationism can be described as a 'why not' methodology in contradistinction to a 'why' methodology with respect to scientific theories. According to Popper, a scientific theory is posed without any prior justification. It gains its confirmation, or 'corroboration,' which is Popper's preferred term in later writings, not by experimental evidence demonstrating *why* it should hold, but by the fact that all attempts at showing *why* it does *not* hold have failed. The main activity of an experimental

researcher is trying to refute or '*falsify*' a scientific theory, rather than to justify or '*verify*' it. Theoreticians who design theories should do this in such a way that they can be refuted most easily. In no case must a theory be immunized against possible falsification, e.g., by the introduction of ad hoc hypotheses. We rely on a theory only because we have not managed to falsify it (so far), not because we have any positive or direct reasons in its favor.

His own theorizing, and especially the idea of falsificationism, is considered by Popper as proposing 'methodological rules.' These methodological rules have to prove fruitful in the understanding of scientific research and, particularly, of scientific progress. The main argument in favor of his falsificationist proposal is that the converse proposal, which is called 'verificationist,' is bound to fail. Verificationism, which underlies the 'inductivist' methodology, claims that a scientific theory is supported by a collection of verifying instances. However, as has already been established by David Hume (1711–76), the inductive reasoning from a finite set of observations to a general statement is not logically valid if the domain of reasoning is infinite or at least indefinite. Furthermore, any attempt at creating a special nondeductive form of reasoning (an 'inductive logic') can only be justified by means of induction itself, thus ending up in a vicious circle. To quote Popper's famous example, no observation of white swans can verify the general statement that all swans are white.

Popper's falsificationist methodology rests on the assumption that unlike verification, falsification of a general statement by experimental or observational means is without any logical problem. The statement 'all swans are white' can be falsified by the observation of a single black swan. The logic involved here is just ordinary deductive logic, due to which the falsity of an instance derived from a general hypothesis falsifies the hypothesis itself. Seemingly singular statements such as 'The planet Pluto has a moon,' which at first glance seem to be verifiable but not falsifiable, are viewed by Popper not in isolation, but as embedded in a general theory and therefore are no counterexample to what he calls the 'asymmetry between verification and falsification.'

More complicated is the example of statistical theories, where definite falsification is not possible. Here, Popper applied as an additional methodological rule that what is extremely unlikely can be neglected as something, which cannot be reproduced deliberately. This allows Popper to cope with the standard way of rejecting statistical theories by means of a critical test (usually defined by stipulating a level of significance).

In traditional methodology, particularly that of the Vienna circle, inductivism was tied to empiricism and guaranteed the connection between theoretical statements and their empirical basis. In falsificationism, the empirical significance of theoretical statements is kept by relating them to basic statements, which possibly falsify them. However, these basic statements are no longer purely observational in some epistemological sense, but are themselves theory-laden from the very beginning. Even elementary perceptual evidence has to rely on theoretical assumptions. Thus, Popper rejects any semantic division between theoretical and observational statements and, especially, the claim that a particular sort of certainty comes with the latter. According to Popper, the validity of basic statements can only be fixed in the course of the testing of a theory, and, in the end, this fixing is a matter of convention. This variant of conventionalism, which enters Popper's approach in addition to the conventionalism inherent in the choice of methodological rules, has repeatedly been the subject of criticism.

For Popper, the distinction between a theory that can be falsified by means of basic statements and one that is not falsifiable serves as a demarcation criterion for empirical science versus metaphysics. However, metaphysics, though not falsifiable, is not meaningless. Metaphysical statements may serve useful purposes, even within science. For example, the principle of causality ('any event can be causally explained'), which cannot be falsified and is therefore metaphysical, can be viewed as the useful methodological rule never to abandon the search for laws that explain observed facts. This view of metaphysics is less rigid than that of the Vienna circle, who considered metaphysics to be meaningless throughout. It allows Popper to sharply criticize certain nonfalsifiable theories as 'pseudosciences' (such as psychoanalysis), but nevertheless to make sense of certain traditional philosophical claims or to appreciate the fruitful role of metaphysical research programs in the history of physics ('A metaphysical epilogue,' in *Quantum Theory and the Schism in Physics* 1956, published 1982b).

In his later works, Popper extended the methodological approaches of the *Logik der Forschung* toward a general account of the dynamics of scientific theories ('Truth, rationality, and the growth of scientific knowledge,' in *Conjectures and Refutations* 1963). According to this view, theoretical progress is made by successive critique and revision of existing theories, which is governed by the idea of objective truth. Scientific progress works by replacing a falsified theory with a new one, which, in the ideal case, is more encompassing than the previous one in that, while coping with the new results, it also covers the unfalsified instances of the old theory. This appeared to be the case with classical and relativistic gravitational theories in the presence of the results of the 1919 solar eclipse observations, which were so impressive to the young Popper. His later claim is that the new and more encompassing theory is nearer to the truth than the older and more special one, even if it should, at some point, become falsified itself. Popper has tried to characterize the distance of a theory from truth, its 'truthlikeness' or 'verisimilitude,' in numeric terms. Considerable work has been carried out in recent decades to develop improved measures.

Within the revision of scientific theories, Popper strongly distinguishes between hypotheses being tested and background knowledge, which, though testable in principle, is not being questioned at the same time. This distinction makes the falsification of parts of a theory possible, and not only the falsification of a theory as a whole. Therefore, Popper pleads to formulate theories in a modular way such that its parts can be tested independently of each other. He is opposed to 'holistic' views of theories, according to which, in a single test, the whole theory (Pierre Duhem) or, even more radically, all our theoretical knowledge including logic and mathematics (W.V.O. Quine) is at stake. Popper defended his model of scientific progress as a change of scientific hypotheses by means of falsification against historic approaches, which denied the rational character of scientific change. So he was strongly opposed to Thomas S. Kuhn's account of scientific revolutions

as irrational changes without the concept of 'progress' being applicable. Similarly, he argued with Imre Lakatos (1922–74) and Paul K. Feyerabend (1924–94), who started as Popperians, but later on made changes in falsificationist methodology by weakening and relativizing Popper's strict idea of progress by critical testing.

In *Evolution and the Tree of Knowledge* (1961, published 1972 in *Objective Knowledge*), Popper embedded the idea of theory change by falsification and revision within a general theory of the evolution of knowledge. Here 'knowledge' is viewed as a form of adaptation of an organism to the environment. Popper's schema of progress from problems via falsification to new problems is interpreted as the evolutionary transition from organisms via natural selection to those that are better adapted. By viewing all life as problem solving (*All Life is Problem Solving* 1994c), Popper claimed to combine theories of biological evolution with theories of scientific progress, relating his approach to evolutionary epistemology in the sense of D. Campbell. However, it must be emphasized that, unlike most evolutionary epistemologists, Popper was not a naturalist. He insisted that the cultural scientific evolution is rational and based on conscious and deliberate criticism, not on natural selection governed by chance.

In his later work, Popper also replaced the narrow concept of 'falsification' with a broader concept of 'criticism' or 'critical examination.' This broader concept not only applies to scientific theories that are formulated in a logical language, but also to philosophical claims that are metaphysical. It involves tests of consistency and the investigation of how far the theory in question is able to solve particular problems. This orientation, with critical examination as its cornerstone, was called 'critical rationalism' by Popper and has since become the standard denomination of the Popper school. The main examples of Popper's dealings with metaphysical problems are his treatment of determinism and his philosophy of mind. Quite independent of problems of quantum mechanics, which often serve as arguments against determinism, in *The Open Universe: An Argument for Indeterminism* (1956, published 1982a), he developed an indeterministic world view. In 'Of clouds and clocks' (1966, reprinted in *Objective Knowledge* 1972), he extended this view with arguments for the openness of the future for free action. In the mind–body debate, Popper defended a mind–body dualism ('Language and the Body–Mind Problem,' 1953, reprinted in *Conjectures and Refutations* 1963), which he later extended to a model of three worlds of (1) physical entities, (2) mental states, and (3) objective contents of thought, which causally interact. This approach gained much attention through his discussions with J.R. Eccles which, together with their separate contributions, are presented in their joint book *The Self and Its Brain* (1977).

Epistemologically, Popper was an outspoken objectivist and realist. On the Popperian view, our theories are not just methodological instruments or means of acquiring (some degree of) certainty about something, but attempts at explaining the real world. As reality is described by means of true sentences, this is another way of claiming that the goal of science is truth. Part of Popper's objectivism and realism is his rejection of any subjective interpretation of probability. In particular, he was strongly opposed to Bayesianism and its view that scientific progress is just changing the probabilistic degree of belief in certain propositions, which means that scientific dynamics is probability dynamics (*Probability Magic or Knowledge Out of Ignorance* 1957b). Both the probability of hypotheses, for which Popper coined the term 'logical probability,' and statistical probability, which may be part of a hypothesis, are objective probabilities reflecting properties of what they are attributed to. Logical probabilities are inversely proportional to the content of a theory, the less probable theory having a higher content. Statistical probabilities, which in the *Logik der Forschung*, were interpreted as relative frequencies, were later viewed as objective 'propensities' of experimental designs to develop certain outcomes (*Realism and the Aim of Science*, 1956, published 1983). This propensity interpretation of probability claims to be a solution of the problem of assigning probabilities to singular events, rather than only to events as members of a given sequence or set of related events.

Popper's attitude against probabilities as degrees of belief does not contradict the fact that he attempted to measure the degree of corroboration of a theory in probabilistic terms. However, first of all, Popper's degree of corroboration, though relying on probabilities, is not a probability itself, i.e., it does not obey the laws of the calculus of probability. Second, the degree of corroboration is intended as a strictly objective measure, which reflects the way the hypothesis in question was subjected to falsification, and is definitely not meant as a degree of rational belief in a hypothesis. According to Popper, epistemology is the methodological treatment of objective scientific contents and not a theory of the knowing subject ('Epistemology without a knowing subject,' 1967, reprinted in *Objective Knowledge* 1972).

Popper's social and political philosophy, which was developed during his emigration in New Zealand and published in *The Poverty of Historicism* (1944–45) and *The Open Society and Its Enemies* (1945) can, to some extent, be viewed as an application of falsificationism to the area of social interaction. By 'historicism,' Popper denoted the claim that the methodology in the historical and social sciences differs fundamentally from those in the natural sciences but, nevertheless, allows one to formulate objective laws of historical development, which cannot be changed by subjective action. The standard example is the Marxist philosophy of history (historical materialism). In contradistinction to that, Popper put forward the idea of 'piecemeal social engineering,' i.e., of a development of social institutions in small steps, which can be revised and whose potential damage can be controlled. According to Popper, piecemeal social engineering can be combined with a falsificationist form of social research, which is based on trial and error as it generates models for social development. For areas of historical and sociological research that are interested in the explanation of single events rather than the testing of general laws, Popper proposed the procedure of a 'logic of the situation,' which describes the acting of individuals under the hypothesis of rational purpose-oriented behavior within a certain social situation. In this context, he defends the autonomy of sociology against psychologism, as this situation is described in objective terms by reference to social institutions, etc. Since 1961, Popper defended this account of social theorizing against representatives of the neo-Marxist Frankfurt school in Germany, in particular, Theodor W. Adorno and Jürgen Habermas, in a dispute, which became known as the

'positivist dispute in German sociology,' and which made Popper well known to a wider audience in the German-speaking countries.

The *Open Society* extends the critique of historicism to a critique of the philosophical theories of state and society by Plato, Hegel, and Marx, which Popper considers to be philosophical precursors of political totalitarianism. According to Popper, they advocate the idea of a closed society, which is a collectivistic and organismic whole, without the distinction between natural and cultural laws. In an open society, which Popper strongly favors, individuals are personally responsible for their decisions and can therefore be critical to social regulations. Popper argues that the traditional question 'who should rule' ought to be replaced by the question of how to minimize the risk of bad rulers and how to get rid of them once they are in power. For Popper, the long-standing philosophical discussion about the best rulers belongs to justificationist thinking that has to be abandoned in favor of critical thinking, just as in general methodology. The utilitarian principle of maximization of happiness should give way to the idea of minimizing suffering.

Although there appears to be a lot of moralizing both in the *Open Society* and in many short papers and interviews on questions of morals and politics given toward the end of his life, this should not be understood as fragments of an ethical theory. For Popper, there is no such thing as ethics, because ethical sentences are not propositions in the logical sense that one could argue for or that one could reject. However, there are moral actions. These are actions that are carried out by someone who is aware of their consequences. Since the consequences of actions can be critically examined in the sense of critical rationalism, moral decisions are not irrational.

Relevance and Impact of Popper's Philosophy

Popper's methodological program was revolutionary as it reverted the established conceptual order between certain fundamental notions. He put 'negative' concepts such as 'falsification' and 'rejection' first, and made 'positive' concepts such as 'acceptance' or 'corroboration' dependent on them. While, according to the traditional philosophical view of science established ever since Bacon, acceptance of a theory is based on the direct generalization from observation, according to Popper, it can only rely on failed rejection. It is therefore correct to speak of Popper's 'negative methodology.' As Popper tried to show in his later writings, this is not just a program for the natural sciences, but applies to any area of scientific reasoning and even to metaphysics, where negative concepts such as that of 'criticism' come first. The credit we give to a conjecture is based on the fact that refuting it by criticism has been unsuccessful. This program is even the basis of Popper's political philosophy. The negative question of how to restrain political power and to remove bad rulers from power is of primary concern compared with the positive one of whom to give power. Popper's negative methodology is a unified view of rationality under the header 'critical rationalism.'

This comprehensive perspective of theoretical reasoning and social and political action makes Popper's philosophy comparable with that of great classics like Aristotle or Kant. It was also one of the sources of the great success, which Popper enjoyed still during his lifetime. As a general and unifying view, 'critical rationalism' is, of course, liable to oversimplification and, among certain adherents of Popper there has, in fact, been a considerable amount of that. On the other hand, it was only critical rationalism as a general orientation that could publicly compete with the eloquence of hermeneutic or neo-Marxist approaches. Philosophical theories like logical empirism, though with respect to intellectual clarity and responsibility by no means behind Popper's, would have been too academic to develop any major public significance.

In fact, it is not very difficult to give an outline of the basic principles of critical rationalism to somebody who is not a trained philosopher. Critical rationalism is an approach that offers something at any depth of understanding. Popper's works both cover extremely intricate discussions, e.g., of the methodological role of probability, whose significance can only be judged by highly specialist researchers, and general questions of reasoning, society, and politics at a level accessible to anybody who is really interested. With the exception of the chapters on probability and on quantum mechanics and of the appendices, the *Logik der Forschung* can easily be studied without deep expertise in academic philosophy. Even from a first reading, the average student would gain a preliminary understanding of falsificationism. Matters of presentation are very important in philosophy, and Popper was a master in that. His language is extremely clear and nonpretentious, which distinguishes him from many other philosophers, particularly in the German tradition.

Popper made a strong impression on politicians of various directions, although he was himself never involved in party politics. Both conservative and social democratic or labor politicians have referred to Popper to justify their respective programs and proposals. The term 'open society,' which Popper coined, has become widely accepted as a designation of basic features of a democratic state.

Popper's lasting impact on the shape of philosophy of science was not only due to his methodology in the narrower sense, but also due to his drawing attention to the *dynamics* of science. It is significantly due to Popper's influence that the philosophical and logical reconstruction of scientific reasoning is not just a modeling of static theories in logical terms but, to a great extent, an account of theory dynamics, i.e., of what it logically means that a theory is being given up or revised. The combination of philosophy of science with the history of science, which is reflected in many divisions of university departments, denominations of chairs, etc., emerged on this background, as philosophical theories of theory dynamics have to be tested against the historical development of theories, and, conversely, the proper understanding of historical developments requires appropriate methodological tools. Although today's theorizing differs in many respects from what Popper argued for and often contradicts his claims (i.e., Bayesian approaches in the field of belief revision), the research program of putting dynamics of theories first in the methodology of science is unthinkable without Popper's revival of the induction problem and his negative methodology as a proposed solution.

The influence of Popperian theories on the 'working scientist' should not be underestimated. If scientists have any interest in the methodology of science at all, they would know of Popper's approach. This is particularly true of the social and

behavioral sciences. There, falsificationism figures prominently in textbook chapters dealing with methodological questions. One may speculate why falsificationism is so appealing to social scientists. A possible answer might be that statistical methodology, which lies at the heart of the social sciences, has a bias toward falsification. According to the most widespread statistical methodology, which is a combination of accounts by R.A. Fisher and by J. Neyman and E. Pearson, the rejection or falsification of a statistical null hypothesis at a certain level of significance is a direct and uncontroversial result, whereas the acceptance of a positively formulated statistical hypothesis is only indirect and preliminary (as it depends on features such as sample size) and, in any case, is technically more involved.

Certain results of modern computer science have challenged Popper's view on induction. His methodology comes along with the claim that we always start with theories or hypotheses which, in the course of time, are then revised to better hypotheses by trial and error. How a new hypothesis is formed was not the subject of rational investigation for Popper. In artificial intelligence and, in particular, in learning theory, algorithms have been developed to generate theories that fit given data. This does not contradict falsificationism, as these heuristically developed hypotheses would then have to be tested, just as falsificationism claims. However, the availability of such procedures destroys the myth that a theory is the extrarational product of a genius or some sort of accidental mutation, and lays the foundation for algorithmic inductive heuristics as a rational enterprise.

There are a lot of questions in the technical development of the Popperian approach, e.g., his treatment of probability and of logic, whose significance for future developments has still to be proven. Something similar holds for Popper's metaphysical theories. However, these are less important details compared with his basic methodological claims, which will remain a standing challenge.

See also: Causation: Physical, Mental, and Social; Explanation: Conceptions in the Social Sciences; Historicism; Liberalism: Historical Aspects; Liberalism: Impact on Social Science; Liberalism: Political Doctrine and Impact on Social Science; Positivism, History of; Positivism, Sociological; Sociology, Epistemology of; von Hayek, Friedrich A. (1899–1992).

Bibliography

Boylan, T.A., O'Gorman, P.F. (Eds.), 2008. Popper and Economic Methodology: Contemporary Challenges. Routledge, London.

Corvi, R., 1997. An Introduction to the Thought of Karl Popper. Routledge, London.

Gattei, S., 2009. Karl Popper's Philosophy of Science: Rationality without Foundations. Routledge, London.

Geier, M., 1994. Karl Popper. Rowohlt, Hamburg.

Hacohen, M.H., 2000. Karl Popper – the Formative Years, 1902–1945: Politics and Philosophy in Interwar Vienna. Cambridge University Press, Cambridge, UK.

Jarvie, E., Pralong, S. (Eds.), 1999. Popper's Open Society after Fifty Years. Routledge, London.

Jarvie, I., Milford, K., Miller, D. (Eds.), 2006. Karl Popper: A Centenary Assessment, 3 vols. Ashgate, Aldershot, UK.

Keuth, H., 2005. The Philosophy of Karl Popper. Cambridge University Press, Cambridge, UK (German ed.: Die Philosophie Karl Poppers, Mohr Siebeck, Tübingen 2000, second ed. 2011).

Magee, B., 1985. Philosophy and the Real World: An Introduction to Karl Popper. Open Court, La Salle.

O'Hear, A., 1995. Karl Popper: Philosophy and Problems. Cambridge University Press, Cambridge, UK.

O'Hear, A. (Ed.), 2004. Karl Popper. Critical Assessments of Leading Philosophers, 4 vols. Routledge, London.

Popper, K.R., 2001–2015 [Collected Works, German]. Gesammelte Werke, 15 vols. Mohr Siebeck, Tübingen.

Popper, K.R., 1925. Über die Stellung des Lehrers zu Schule und Schüler: Gesellschaftliche oder individualistische Erziehung? Schulreform 4, 204–208 (New ed. in Gesammelte Werke, vol. 1).

Popper, K.R., 1928. Zur Methodenlehre der Denkpsychologie. originally unpublished (New ed. in Gesammelte Werke, vol. 1).

Popper, K.R., 1934. Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft. Julius Springer, Vienna (Tenth ed. Mohr Siebeck, Tübingen 1994 [last ed. prepared by Popper], eleventh ed. by H. Keuth as Gesammelte Werke, vol. 3); English: The Logic of Scientific Discovery, Hutchinson, London 1959 (tenth ed. 1980).

Popper, K.R., 1945. The Open Society and Its Enemies, Vol. I: The Spell of Plato, Vol. II: The High Tide of Prophecy: Hegel Marx, and the Aftermath. George Routledge, London (Second ed. Routledge & Kegan Paul, London 1952, many reprints) (German ed. as Gesammelte Werke, vols. 5–6).

Popper, K.R., 1957a. The Poverty of Historicism. Routledge & Kegan Paul, London (Originally as a series of papers in the Journal Economica 11 (1944) and 12 (1945)) (German ed. as Gesammelte Werke, vol. 4).

Popper, K.R., 1957b. Probability magic, or knowledge out of ignorance. Dialectica 11, 354–372.

Popper, K.R., 1963. Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge & Kegan Paul, London (German ed. as Gesammelte Werke, vol. 10).

Popper, K.R., 1972. Objective Knowledge: An Evolutionary Approach. Clarendon Press, Oxford (second ed. 1979) (German ed. as Gesammelte Werke, vol. 11).

Popper, K.R., 1976. Unended Quest: An Intellectual Autobiography. Fontana, London (=Intellectual Autobiography In: Schilpp 1974) (German ed. as Gesammelte Werke, vol. 15).

Popper, K.R., 1979. Die beiden Grundprobleme der Erkenntnistheorie (ed. Hansen, T.E.). Mohr Siebeck, Tübingen (Third ed. as Gesammelte Werke, vol. 2).

Popper, K.R., 1982a. The open universe: an argument for indeterminism. From the Postscript to the Logic of Scientific Discovery (ed. Bartley III, W.W.). Hutchinson, London (German ed. as Gesammelte Werke, vol. 8).

Popper, K.R., 1982b. Quantum theory and the schism in physics. From the Postscript to the Logic of Scientific Discovery (ed. Bartley III, W.W.). Hutchinson, London (German ed. as Gesammelte Werke, vol. 9).

Popper, K.R., 1983. Realism and the aim of science. From the Postscript to the Logic of Scientific Discovery (ed. Bartley III, W.W.). Hutchinson, London (German ed. as Gesammelte Werke, vol. 7).

Popper, K.R., 1984. Auf der Suche nach einer besseren Welt: Vorträge und Aufsätze aus dreißig Jahren. München, Piper (English: In Search of a Better World: Lectures and Essays from Thirty Years, Routledge, London 1992) (New German ed. as Gesammelte Werke, vol. 14).

Popper, K.R., 1985. Popper Selections (ed. Miller, D.). University Press, Princeton.

Popper, K.R., 1990. A World of Propensities. Thoemmes, Bristol.

Popper, K.R., 1994a. The Myth of the Framework: In Defence of Science and Rationality (ed. Notturno, M.A.). Routledge, London (German ed. as Gesammelte Werke, vol. 13).

Popper, K.R., 1994b. Knowledge and the Body-Mind Problem: In Defence of Interaction (ed. Notturno, M.A.). Routledge, London (German ed. as Gesammelte Werke, vol. 12).

Popper, K.R., 1994c. Alles Leben ist Problemlösen: Über Erkenntnis, Geschichte und Politik. München, Piper (English: All Life is Problem Solving, Routledge, London 1999).

Popper, K.R., 1996. The Lesson of This Century: Karl Popper Interviewed by Giancarlo Bosetti. Routledge, London.

Popper, K.R., 1998. The World of Parmenides. Essays on the Pre-socratic Enlightenment (ed. Peterson, A.F.). Routledge, London.

Popper, K.R., 2012. After the Open Society: Selected Social and Political Writings (ed. Turner, P.N., Shearmur, J.). Routledge, London.

Popper, K.R., Eccles, J.C., 1977. The Self and Its Brain: An Argument for Interactionism. Springer International, Berlin etc.

Rowbottom, D.P., 2011. Popper's Critical Rationalism: A Philosophical Investigation. Routledge, London.

Schilpp, P.A. (Ed.), 1974. The Philosophy of Karl Popper, 2 vols. Open Court, La Salle.

Shearmur, J., 1996. The Political Thought of Karl Popper. Routledge, London.

Simkin, C.G.F., 1993. Popper's Views on Natural and Social Science. Brill, Leiden, Köln etc.

Popular Culture

Purnima Mankekar, University of California Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article I present some conceptual frameworks for examining popular culture. I discuss how popular culture includes media, sports events, advertisements, street fairs, and tourism. I then describe how popular culture provides us with a lens to trace larger historical processes and public debates on issues as wide-ranging as family and kinship, nation, intimacy and sexuality, and the transnational. I posit that, in the contemporary world, it is essential to examine popular culture in its transnational dimensions. Popular culture is also a critical node in the generation and circulation of not just ideologies but also regimes of affect. In every sense, popular culture is socially and politically consequential. I trace different scholarly perspectives on popular culture, such as feminist critique, critical theory, cultural studies, and historical frameworks. I conclude by considering how digital media have altered how we think of popular culture, such that our previously held assumptions about sociality and politics, distance and proximity, and virtuality and reality are now altered.

For most of us, popular culture is a source of entertainment. It encapsulates media (such as film, television, print media, and digital ('new') media), and includes sports events, advertisements, street fairs, and tourism. We engage with popular culture because we get pleasure from it; at the same time, we are also informed by it. We usually associate information with news rather than with entertainment. But other forms of popular culture that we often associate with for 'mere entertainment' – for instance, tourist practices, popular music, films, sports, and television programs – also inform us in profound ways. For example, a romance novel might instill in us expectations about intimacy, gender and sexuality. Popular culture plays an important role in shaping our personal and collective identities. It also provides us with an analytic lens to understand sociohistorical processes such as class, gender and sexuality, nationalism, and transnationalism.

Changes in popular cultural texts often articulate in wider social and cultural changes. Consider the contrast between the harmonious, heteronormative nuclear family portrayed in the *Cosby Show* in the 1980s, and the family we see in the U.S. sitcom *The Modern Family* which depicts the lives of members of three interrelated families. *The Modern Family* is multiracial and includes a gay couple and an adopted Vietnamese child. The past few years have also witnessed an efflorescence of digital media that have been immensely powerful in altering the cultural landscape of many societies: a range of relationships are now forged on Facebook, flash mobs converge through Twitter, and the internet has become a vital source of connectivity in many parts of the world. In recent years, digital media have also played an important role in politics, whether through online political fund raising or by galvanizing rebellion in societies where other media may be censored.

What Is Popular Culture?

In the modern world, popular culture is available to what we may call 'the public' or 'masses' of people. For this to happen, there have to be established means of production, distribution, and reception. For example, for a soap opera to exist there have to be a crew of script writers, actors, a director, a producer – in other words, a crew of people who work behind the scenes and in front of the camera. There also has to be a means to finance this production: in capitalist societies, soap operas are financed by sponsors who pay for the production and telecast of a soap opera in exchange for being able to advertise on it. This is how the term 'soap opera' originated: soap operas were sponsored originally by manufacturers of detergents. Even today, the manufacturers of detergents and other household goods like washing machines, kitchen appliances, and cleaning agents are key advertisers on soap operas. There also needs to be an infrastructure for its distribution. This is where television networks come in. Very often, television networks are part of larger media conglomerates or groups of companies. Television production is, therefore, part of a larger system of industrial production.

Popular culture presents us with a vantage point on the world: it is a domain that shapes our sense of self or our identity. Not only is it generally available to a large number of people, it also responds to and creates points of identification among large groups within a community or nation. Popular culture plays an important role in constituting and representing the public to itself. It is public in two senses: not only does it circulate widely and is often associated with public exhibition and display, it also configures 'the public,' 'the nation,' 'society,' 'culture,' 'family,' and other collectivities. In the contemporary world, transnational media play a primary role in the circulation of popular cultural texts. From a sociological viewpoint, popular culture is probably the most important influence on the lives of millions. However, it is important to recognize that viewers are not just unthinking targets of popular culture. Rather, in order to understand the cultural significance of popular culture, we need to examine how its consumers respond to, reinterpret, and employ its messages.

Popular culture is a site of struggle between dominant discourses and forces of resistance. It contains 'points of resistance' as well as ' moments of supersession'; it forms a 'battlefield where no once-for-all victories are obtained but where there are always strategic positions to be won and lost'

(Hall, 1981, p. 233). Mass media may open up a 'zone of debate' about issues current in a given community (Gupta, 1995). For instance, in the 1990s, television shows like the U.S. sitcom *Ellen* enabled the rearticulation of debates around homosexuality. Popular culture, then, does not simply 'reflect' the lives and worlds of audiences; nor does it 'impact' these worlds or lives in a simple way. Instead, popular cultural texts are social formations in and of themselves.

In addition, we need to attend to three other dimensions of popular culture in the contemporary world. One, we need to examine popular culture in terms of its transnational dimensions. Transnational media have played a critical role in connecting viewers/readers/consumers in different parts of the world. Satellite television allows us to keep up with the news, weather updates, and entertainment programs in different parts of the world: thus, for instance, viewers living in London are able to watch television soap operas at the same time as their relatives in the United States. The internet has changed previous conceptions of 'breaking news' by enabling us to engage with events even as they unfold in different parts of the world.

Second, the proliferation of media has altered how we engage the world of information and entertainment: we do not just listen to music, we watch music videos; we stream movies on our computers; we listen to ringtones on our cellphones; and so on. As a result, rather than focussing on a single medium, scholars of media prefer to trace the different media that form a media-saturated *landscape* or media *ecology*.

Finally, we need to conceive of popular culture as a node for the generation and circulation of not just ideologies but, equally importantly, of affects. Media scholars now recognize that in addition to being ideologically interpellated by media, media also work on us through the affective regimes they generate and circulate. These regimes of affect blur the line between feeling and thought, the corporeal and the cognitive, and the individual and the collective: the regimes of affect circulated through media are transitive, yet have a profound effect on how we position ourselves in the world. For instance, media played a critical role in the generation and circulation of the affects of fear and suspicion in the aftermath of terrorist attacks in the United States on 11 September 2001. These affects of fear and suspicion were transnational in their provenance and enabled some people in the Western world to feel a strong sense of identification with the victims of these terrorist attacks: for instance, the French newspaper *Le Monde* ran a front page headline stating: "We are all Americans." Equally important, the affects of fear and suspicion that circulated after 11 September 2001 had the effect of demonizing racial and religious Others as potential terrorists: in many Western nations, this resulted in hate-based violence against those who 'appeared to be' terrorists and to their imprisonment and deportation.

Popular culture, then, is socially and politically consequential. Our subjectivities are constituted in the process of making sense of the world around us and in our interaction with others and, in many cases, making sense of the world entails engaging with it not just cognitively but, also, through affect. In capitalist and postcapitalist societies, these processes are refracted by media. Film, television, popular music, and mass market fiction constitute and represent the public realm in cognitive, moral, and affective terms. But these representations do not impact us as if we are blank screens – instead, 'they occupy and rework the interior contradictions of feeling and perception … they do find or clear a space of recognition in those who respond to them' (Hall, 1981, p. 232). Thus, although representations are neither all-powerful nor all-inclusive, they have material effects on how people see themselves and lead their lives.

Scholarly Perspectives on Popular Culture

Definitions of popular culture have shifted according to the disciplinary emphases of its students. Historians of early modern Europe (e.g., Peter Burke (1978) and Robert Darnton (1984)) stress how members of different classes created and participated in popular culture and, simultaneously, were shaped and constrained by it. These historians analyze the factors influencing the popular cultures of early modern Europe, chiefly, the rise of cities and the advent of printing. They supplement thematic analyses of books with information about audiences and, thus, aim at providing the context for the meaning and uses of books.

One of the foundational works on the relationship between capitalism and mass media is Max Horkheimer and Theodor Adorno's *The Culture Industry: Enlightenment as Mass Deception* (1944). Horkheimer and Adorno focused on how culture and the individual are 'produced' by the culture industry. They claimed that the culture industry's power lies in its ability to disseminate ideas and beliefs on a mass scale. They were concerned about how the management and control of leisure time and the socialization of the ego by the culture industry leads to a 'mass' culture that undermines the private realm.

The relationship between modern forms of popular culture and the growth of capitalism is also examined by some cultural studies scholars. Some of the most influential work in this area was produced (and has been influenced by) the Centre for Contemporary Cultural Studies at the University of Birmingham, UK. These scholars are concerned with how media shape dominant ideologies and discourses. Their central project is to examine the relationship between the production and consumption of popular culture and social axes of inequality, such as class, race, and gender.

Although cultural studies' earliest borrowings were from literary criticism, chiefly the work of Raymond Williams and Richard Hoggart, recently its primary focus has been on the practices of everyday life. The greatest debts of this genre of cultural studies are to Marxist theories of culture. Its basic premise is that cultural production often takes the form of commodities. Other important propositions are as follows: (1) popular culture is inextricably tied up with social relations; (2) popular culture can both enable social groups to define and realize their needs and reproduce asymmetries; and (3) popular culture is a terrain where social differences and struggles are articulated. These concerns reflect the influence of Gramsci on the ideological and popular struggles involved in the construction of common sense. In addition, some cultural studies scholars influenced by Lacanian analyses of film and television reconceptualize ideology in terms of how texts

construct subject positions. Others have rethought the role of ideology in popular culture in terms of Foucault's analyses of discursive practices. Breaking with Horkheimer and Adorno's assumption of the audience as a passive and undifferentiated mass of receivers, these cultural studies scholars conceive of audiences or consumers who actively 'decode' and interpret popular culture texts.

In addition to studying everyday practices, some cultural studies scholars also use semiotic methods of textual analysis. They treat cultural products as 'texts' to be 'read' according to interpretive strategies, including the literary analysis of narrative, semiology, and deconstruction. These analytical strategies reveal that literary conventions and forms have greater social and cultural significance than we might first suspect. For example, feminists have pointed to the place of fictional representations of romantic love in cultural constructions of femininity. Far from being 'mere' literary constructions, these representations play a crucial role in the constitution of dominant practices of gender and sexuality.

Anthropologists bring important theoretical and methodological insights to the study of popular culture because of their commitment to ethnographically analyzing the relationship between popular culture and other social institutions. Initially, mass media were neglected by anthropologists (exceptions include Bateson, 1943; Powdermaker, 1950) because they were deemed 'too redolent of western modernity and cultural imperialism for a field identified with tradition, the non-western and the vitality of the local'(Ginsburg, 1999, p. 297). Some anthropologists have insisted on foregrounding the *social practice* of media (Abu-Lughod, Ginsburg, and Larkin, eds.) and examine the cultural implications of popular culture through an ethnographic attention to the everyday practices and lives of producers and/or consumers of media.

Anthropological perspectives underscore that popular culture cannot be generalized across different historical and cultural contexts simply on the basis of the media apparatuses. Producers and consumers of media are, fundamentally, social and historical subjects and media are, primarily, cultural institutions that need to be studied in their concrete particularity. Hence, we cannot conceptualize relationships between popular culture and the social actors who engage its myriad texts in the abstract: neither can be studied outside their location in specific sociohistorical contexts. Several anthropologists have engaged in analyses of popular culture. These scholars, together with those who study the reception of media, carefully situate popular culture in the discursive and historical contexts in which it is produced and consumed. These studies undermine unilinear models of 'communication flows,' and foreground the importance of examining the specific conditions in which popular culture circulates.

Other scholars have examined how media have shaped our sense of the public. Drawing on and critiquing notions of the public sphere first articulated by Jurgen Habermas, these scholars have examined subaltern counterpublics and multiple publics, or have proposed notions of public culture as an alternative to those of the public sphere. Others have pointed to how gender and conceptions of intimacy play a critical role in how we think of the public. Many scholars foreground the role of transnational media in the formation of multiple publics. Far from proposing a dichotomous relationship between the public and the private, all these scholars assert that media blur the line between the two.

Popular Culture as a Lens for Studying Other Sociocultural Processes

Popular culture is a terrain where identities are articulated, negotiated, and consolidated. For example, class position might provide individuals with 'cultural constraints' and a common 'outer frame of meaning' (Darnton, 1984, pp. 6–7). A mutually influencing relationship exists between elite culture and popular culture. Popular culture also provides a site for the negotiation and articulation of class identities and other hegemonic discourses. Thus, in nineteenth-century India, Bengali elites took upon themselves the task of 'purifying' popular culture produced by working-class women: while their objective was ostensibly to 'emancipate' women of their class, their efforts resulted in the consolidation of their own middle- and upper-class identities through the marginalization of working-class popular culture (Bannerjee, 1989). At the same time that these interventions influenced (and were influenced by) hegemonic discourses of class and gender, they also became foundational to discourses of culture that were fundamentally nationalist.

The role of popular culture in mediating discourses of gender and sexuality has long been acknowledged. Two pioneering works in this regard are Laura Mulvey's *Visual Pleasure and Narrative Cinema* (1975) and Janice Radway's *Reading the Romance* (1984). Laura Mulvey examined how the cinematic forms and narratives of classic Hollywood film 'suture' viewers into 'spectatorial positions' that reinscribe dominant ideas about gender and sexuality. Mulvey's work provided feminist critics of popular culture with a paradigm for examining the role of mass media in the constitution of discourses of gender and sexuality. Janice Radway's work laid the foundation for studying the role of a specific genre of texts, romance fiction, in the constitution of discourses of gender and sexuality. Radway situated women's reading practices in their social relationships and everyday lives, and studied their interpretations of different romances in order to analyze what the books mean to their readers.

Since the publication of these works by Mulvey and Radway, the relationship of popular culture to gender and sexuality has been analyzed by scholars too numerous to cite. Some of the most noteworthy studies in this regard, however, come from scholars, chiefly anthropologists, who have examined media in non-Western contexts. These include the analysis of the relationship between popular culture, gender and sexuality in Japan (Allison, 1996), and television and notions of modern womanhood in Egypt and India (Abu-Lughod, 1993; Mankekar, 1999 respectively). The significance of these studies lies in their decentering of Eurocentric perspectives dominant in cultural studies, media studies, and feminist studies.

Some scholars have argued that the conceptualization of popular culture was itself engendered by the emergence of nationalism (Burke, 1978). Building on Benedict Anderson's insights about the role of print capitalism in the construction of nations as imagined communities (1983), several scholars have examined the role of popular culture in the formation of

national identity. For instance, diverse forms of popular culture, including tourism, exhibitions, and material culture and consumption, have enabled the constitution of the national imaginary in China (Anagnost, 1997). In some parts of the developing world, television has been consciously deployed to construct a 'national community': in Mexico and Venezuela, indigenous soap operas called telenovelas have combined entertainment with 'social messages' about development. In India, state-owned television was harnessed to the explicitly nationalist goals of development and national integration (Mankekar, 1999) and, furthermore, participated in the consolidation of an exclusionary and strident Hindu nationalism (Rajagopal, 2000).

Yet, at the same time that popular culture enables the constitution of local and national communities, some of its forms are inherently transnational. Thus, the telenovelas described above circulated not only within but across national borders in Latin America and were adopted as models for the television programs in India. In addition, the transnational circulation of popular culture has resulted in the creation of new forms of diasporic identities. The transnational circulation of newspapers, film, television programs relayed via satellite, advertisements, and mass market fiction have created, what anthropologist Arjun Appadurai has termed, 'mediascapes' (1996). Mediascapes 'refer both to the distribution of the electronic capabilities to produce and disseminate information … now available to a growing number of private and public interests throughout the world, and to the images of the world created by these media' (1996, p. 35). The boundaries of cultures and nations have thus proved to be permeable to globalized and transnational forms of popular culture. However, as scholars have argued for several years, globalization has not resulted in a homogenization of culture. At different historical moments, globalization has resulted in a recasting of 'tradition' and processes of creolization (Hannerz, 1996, pp. 67–77). Transnational popular culture thus forces us to revisit previously held notions of culture as bounded and systemic (Appadurai, 1996; Gupta and Ferguson, 1997).

In recent years, digital media have been critical to the further expansion of our sense of transnational popular culture. Digital media or 'new media' may alter previous conceptions of spatiality and temporality by muddling our sense of distance and proximity, the virtual and the actual. Scholars have now gone beyond *either* valorizing new media as liberatory and empowering *or* condemning it as divisive and oppressive. We have learned how the internet provides sites for arranging marriages; it is also a locus for the formation of hate-based collective identities. Online communities may offer a sense of solidarity to victims of child abuse; at the same time, the internet has also provided opportunities for child pornography and hate-based bullying leading, in some cases, to tragic consequences. Digital technologies may shift power dynamics to provide marginalized groups with access to representation. But, here again, we are also reminded that 'new media,' as much as 'old media' like television, radio, and film, are imbricated in existing power relations (Wilson and Peterson, 2002).

It is important to note that popular cultures generated by new media coexist with other media like film, television, and radio. Nevertheless, recent developments in new media studies also underscore that we may need to rethink previous assumptions about the relationship between humans and machines, the social and the technological, and organic and inorganic forms of life. Thus, for instance, digitization has compelled us to rethink migration and labor (as in the 'virtual migration' of call center workers whose labor travels even as their bodies remain in place (Aneesh, 2006)); electronic surveillance technologies and cyberwarfare have transformed how we think of national security; digital archives have changed how we think of privacy; and webcams and Ipods now seem to have extended our senses of sight and hearing (Balsalmo, A., 2008; Berry et al., 2003; Constable, 2013; Escobar, 1994; Ito et al., 2010; Kolko et al., 2000; Kroker and Kroker, 2008; Morse, 1998; Nakamura and Chow-White, 2011; Stratton, 2008; Tsatsou, 2009; Turkle, 2005, 2012). Furthermore, it is important to remember that even as the digital moves at the speed of light in the form of code, it is refracted by race, class, sex, and gender; conversely, domains of popular culture cohering around digital media compel us to revisit our assumptions about race, class, sex, and gender. Ultimately, digital popular cultures, like popular cultures of print or electronic media, are constituted through sociality even as they transform our sense of the social.

See also: Creolization: Sociocultural Aspects; Everyday Life, Anthropology of; Globalization, Anthropology of; Music: Anthropological Aspects; Nationalism: General.

Bibliography

Aneesh, A., 2006. Virtual Migration: The Programming of Globalization. Duke University Press, Durham.

Abu-Lughod, L., 1993. Finding a place for Islam: Egyptian television serials and the national interest. Public Culture 5, 493–514.

Allison, A., 1996. Permitted and Prohibited Desires. Westview Press, Boulder, CO.

Anagnost, A., 1997. National Past-times: Narrative, Representation, and Power in Modern China. Duke University Press, Durham, NC.

Anderson, B., 1983. Imagined Communities. Verso, London.

Appadurai, A., 1996. Modernity at Large. University of Minnesota Press, Minneapolis.

Balsalmo, A., 2008. The virtual body in cyberspace. In: Kennedy, Barbara, Bell, David (Eds.), The Cybercultures Reader. Routledge, New York.

Bannerjee, S., 1989. Marginalization of women's popular culture in nineteenth century Bengal. In: Kumkum, S., Sudesh, V. (Eds.), Recasting Women: Essays in Colonial History. Kali, New Delhi, India, pp. 127–180.

Bateson, G., 1943. Cultural and thematic analysis of fictional films. Transactions of the New York Academy of Sciences 5, 72–78.

Introduction, 2003. In: Berry, C., Martin, F., Yue, A. (Eds.), Beep – Click – Link: Mobile Cultures. Duke University Press, Durham.

Burke, P., 1978. Popular Culture in Early Modern Europe. Harper and Row, New York.

Constable, N., 2013. Correspondence marriages, imagined virtual communities, and countererotics on the internet. In: Media, Erotics, and Transnational Asia. Duke University Press, Durham.

Darnton, R., 1984. The Great Cat Massacre and Other Episodes in French Cultural History. Vintage, New York.

Escobar, A., 1994. Welcome to Cyberia: notes on the anthropology of cyberculture. Current Anthropology 35, 211–232.

Ginsburg, F., 1999. Shooting back: from ethnographic film to indigenous production/ethnography of media. In: Miller, T., Stam, R. (Eds.), A Companion to Film Theory. Blackwell, Malden, MA, pp. 295–322.

Gupta, A., 1995. Blurred boundaries. American Ethnologist 22, 375–402.

Gupta, A., Ferguson, J. (Eds.), 1997. Culture, Power, Place. Duke University Press, Durham, NC.

Hall, S., 1981. Notes on deconstructing 'the popular'. In: Samuel, R. (Ed.), People's History and Socialist Theory. Routledge, London, pp. 227–240.
Hannerz, U., 1996. Transnational Connections. Routledge, London.
Horkheimer, M., Theodor, A., 1944. Dialectic of Enlightenment. Continuum Press, New York.
Ito, Mizuko, et al. (Eds.), 2010. Hanging Out, Messing Around, and Geeking Out. MIT Press, Boston.
Kolko, B.E., Nakamura, L., Rodman, G.B. (Eds.), 2000. Race in Cyberspace. Routledge, New York.
Kroker, Arthur, Kroker, Marilouise (Eds.), 2008. Critical Digital Studies: A Reader. University of Toronto Press, Toronto.
Mankekar, P., 1999. Screening Culture, Viewing Politics. Duke University Press, Durham, NC.
Morse, M., 1998. Virtualities: Television, Media Art, and Cyberculture. Indian University Press, Bloomington.
Mulvey, L., 1975. Visual pleasure and narrative cinema. Screen 16 (3), 6–18.
Nakamura, L., Chow-White, P. (Eds.), 2011. Race and New Media. Routledge, New York.
Powdermaker, H., 1950. Hollywood: The Dream Factory. Houghton Mifflin, Boston.
Radway, J., 1984. Reading the Romance. University of North Carolina Press, Chapel Hill, NC.
Rajagopal, A., 2000. Retailing Hinduism. Cambridge University Press, Cambridge, UK.
Stratton, J., 2008. Cyberspace and the globalization of culture. In: Kennedy, Barbara, Bell, David (Eds.), The Cybercultures Reader. Routledge, New York.
Tsatsou, P., 2009. Reconceptualising 'time' and 'space' in the era of electronic media and communications. Journal of Media and Communication 1, 11–32.
Turkle, S., 2005. The Second Self: Computers and the Human Spirit. MIT Press, Boston.
Turkle, S., 2012. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York.
Wilson, S.M., Peterson, L.C., 2002. The anthropology of online communities. Annual Review of Anthropology 31, 449–467.

Popular Sexual Cultures in Western Europe (1930s to the Present)

Franz X Eder, Universität Wien, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins with general considerations about popular sexual beliefs and cultures in Europe since the 1930s. Then it shows how state interventionism, conservative family values, and commercial eroticization shaped the sexual culture during the late 1930s until the early 1960s. The 'sexual revolution' from the mid-1960s to 1980s is discussed in different aspects between liberalization of sexual mores and new forms of sexualization and regulation. Since the 1980s the western European culture is highly sexualized. Sex thereby has become a consumer product between negotiation, risk, and virtualization. After the sexual liberalization of the 1960s and 1970s and the gender revolution of the 1980s, the 1990s saw a radical change in relationship culture – matrimony decreased and long-term cohabiting partnership rose. The preferred sexual relationship is now one in which both partners communicate on an equal basis.

Popular sexual beliefs and cultures in Europe were subject to swift and, at times, radical changes during the second half of the twentieth century. New sexual images, norms, and identities affected peoples' erotic fantasies, sexual knowledge, and attitudes. During this period, changes in popular culture effectively became people's day-to-day 'lifeworld' experiences (Schütz and Luckmann, 2003). At the same time, the potential for communication increased, supported by the expansion of educational institutions and growing mass media that opened up yet further new sexual ideas, fields of knowledge, and images to an ever-increasing audience; the educational institutions reached out to a broader range of social classes; and the extensive distribution of mass media such as books, newspapers, and magazines, and the development of new media such as movies, radio, TV, and the Internet reached even more. Therefore, the second half of the twentieth century saw the term 'popular sexual culture' evolve to refer to a culture 'which is widely disseminated and consumed by large numbers of people' (Hinds, 2006: p. 163).

Media-based communication resulted in traditional procurers of values and knowledge – such as parents, teachers, and priests – loosing their pedagogic importance and being replaced by new knowledge informers – such as sex researchers, guidebook authors, politicians, writers, or movie directors. However, conventional understandings that had up to then been conveyed by 'closed' lifeworlds were not substituted altogether. Social and cultural traditions continued, passed down via gender-specific education, socialization within the peer group, value conveyance within the professional realm, and group-specific ideals relating to love, partnership, and family. More often than not, they were concurring with the generally advanced knowledge about sex and the eroticizing images displayed by the media. This 'productive' tension within the sexual culture of the twentieth century called forth relentless and emotional public debates, contributing in turn to changes in behavior and thought within just a few decades.

State Interventionism, Conservative Family Values, and Commercial Eroticization (the Late 1930s until the Early 1960s)

Propelled by the Great Depression of the late 1920s and the early 1930s, Western countries embraced conservatism while developing sexual morals and cultures. Franco's Spain, Salazar's Portugal, and Stalin's Soviet Union underwent a Puritan change. Fascist governments of Italy (as of 1922), Germany (as of 1932–33), and Austria (as of 1934), and subsequently of all countries occupied by the Germans, made sexuality an essential part of their ideological quest (Quine, 1996). Consequently, one's personal sex life was no longer a private matter but was turned into a topic of debate around nationalism and race. Supported by pronatalistic measures, the supposedly pure and 'high-grade' (Aryan) segment of the population was encouraged to expand. Measures included symbolic awards to women who 'joyfully' had given birth many times, tax benefits for large families, the displacement of women from the labor market, the reestablishment of men as heads of the family, and harsh penalty on contraception and abortion. In the long run, these measures failed to generate any substantial change in sexual conduct. Other countries, too, began to upgrade their public health departments and medical counseling centers, and as a consequence their medical competence.

The fascists' cult of the body made nudity socially acceptable, transporting it into racist, nationalist, and aesthetic concepts. Bodies 'steeled' by gymnastics and sports symbolized strength, discipline, and the eagerness to procreate, and at the same time allow for the publication of nude pictures without it being stigmatized as pornography. As a consequence, an inadvertent mass reproduction of nude photos took place decades prior to the 'sexual revolution'. The sexualization of images was, however, two-sided: the beauty of the pure body was counteracted by 'degenerate art' depicting undisciplined sexual craving; and the menacing art of lecherous Jewish, homosexual, and Black artists. The same applied to erotic glamour girls displayed in French magazines and in photos and postcards sold under the counter. In general, fascist regimes tended to assume a prosexual attitude as far as premarital and extramarital relations were concerned. Apparently, the image of the presumably sex-averse Nazi regime was one of the myths materializing in the late 1960s and 1970s in the wake of the so-called struggle to come to terms with National Socialism (Herzog, 2005; Eder, 2007).

The era's prostitution did, in fact, evoke ambivalence: On one hand, it was generally accepted that a man should occasionally frequent professional prostitutes, provided he did not contract any venereal diseases (VDs). On the other hand, this ubiquitous probability of infection made prostitutes a threat to

the whole population, not just families, and the military. In addition, homosexuals were seen as a menace to moral purity through their employment of seduction techniques and the ensuing 'contagion'. The fascist regimes' homophobia resulted partly from the problematic demarcation of homosocial and homoerotic male relationships within their own – usually unisexual – organizations. Homosexuality was suspected to be a contagious disease spreading rapidly within male and youth organizations. Consequently, threats of penalties, the banning of relevant magazines, and the policing of gay and lesbian subcultures, increased. Homophobia reached a state where neighbors, colleagues, and even family members volunteered to denounce suspected homosexuals. At the same time, during the 1930s, the Nazi regime utilized the population's antihomosexual resentments when attacking government critics; for instance, when employing political purges upon church representatives and aberrant party members after incriminating them to have committed homosexual offenses.

During World War II, restrictive attitudes toward sexuality were on the rise as distress and dearth left their mark. Contraceptives were difficult to obtain, abortion rates rocketed, and VDs spread. Married couples became alienated after long separations. Military officials therefore recommended sex during home leaves seeing it as essential to the troops' morals. Military brothels – often abusing women, mainly forced laborers, and concentration camp captives – served the same purpose. Women in general faced increasing sexual violence: countless forcible rapes were committed by members of the German Wehrmacht in their occupied territories (Beck, 2004) and by members of the allies, particularly the Russian soldiers, in Germany and Austria in 1945. These traumatizing experiences had dramatic effects on those involved but were hardly acknowledged after the war. Many of the women were scarred for life from the enforced physical and mental assaults, and their sex lives were permanently affected by the ordeal. Marriages frequently failed as women could not explain their feelings and experiences to their husbands or because some husbands considered the sexual assault to be conjugal infidelity (Kral, 2004).

Numerous voluntary sexual relations between soldiers of the NS Wehrmacht and local women took place in the occupied territories. German-occupied areas in western and northern Europe, for instance, remained largely undisturbed by acts of war until the allies' invasion and retention of 'civil life' meant that sexual contacts of occupiers with local women took place, as reported from France, Belgium, the Netherlands, and Norway. After the war, many of those women were seen as collaborators and labeled as 'Germans' whores'; along with their children, they were stigmatized and frequently attacked. Attitudes toward allied occupiers were similarly ambivalent. GIs, the soldiers of the US army, in particular represented a new type of man, who seduced women with offers of food and extra rations. They were popular with the local young women who, after years of wartime deprivation, wanted to have a good time; others acted out of sheer monetary need. At the same time, GIs gained a bad reputation in England. They were attacked for being 'overpaid, oversexed, over here' (Hall, 2000: p. 145), suspected of being the driving force behind the VD epidemic, and were blamed for the alleged moral decay of young women and their sexual permissiveness.

A 'sexual crisis' developed. Owing to the long separations, spouses drifted apart. Women had grown more independent, and men had returned changed or 'broken' by war trauma. Divorce rates increased as rapidly as did illegitimacy, abortions, and VDs. Couples, who were reunited, often lacked the contraceptives necessary for having sex without fear. From 1947, Beate Uhse, the founder of the future worldwide sex company, realized the situation and started distributing information on contraception in a brochure marketing the Knaus–Ogino temperature method; this enabled a woman (or women) to calculate her fertile and infertile phases.

Young people developed a craving for adventurous sexual experiences, while facing an imbalance in the ratio of women to men owing to the war's death toll in men. With the loosening of censorship, educational advice literature came onto the market supplemented with erotic magazines styled in the US manner; the latter depicted pinup girls that had been part of every Western allied soldier's field kit during the war (Hanson, 2004). Meanwhile, VDs lost much of their threat once penicillin was introduced after the war.

The evident liberalization of sexual attitudes during and after the war was confirmed by surveys conducted in several countries at the end of the 1940s. They revealed that contraception was by now widely used and premarital sex was indeed accepted, if only with prospective spouses. Opinions on abortion, prostitution, and masturbation varied, but people mainly remained antagonistic toward homosexuality. A healthy sexual relationship was now regarded as being essential to a successful marriage. Yet, in spite of these liberal tendencies, for the most, sexual activities remained more acceptable for couples who were already married.

Postwar 1940s and the early 1950s saw a clampdown in reaction to the perceived sexual chaos: The neoconservative (nuclear) family ideal was at the heart of this moral reconstruction. Sex manuals published clear definitions of a woman's role: Their job included being a willing and enthusiastic sexual partner. Sexual freedom was curtailed and sex became a quintessentially private and nonpolitical act. The sexual fidelity of a woman was a highly treasured value.

Premarital restraint and sexual integrity headed the Christian prescription for a future love match and a happy matrimonial life. Contrary to men, women were seen to be more emotional, eager to procreate, and sexually passive by nature. Among the approved sexual ideals for adolescents were virginity, restraint from masturbation, and a stable heterosexual orientation. In European countries, sexual education was regarded as the responsibility of the parents, but brochures and films gained territory as commercial interests increased (Martin, 2003: pp. 135–136). German and Austrian porn laws came out around 1950 aimed at preventing the distribution of 'pulpy and pornographic literature'. Marketing and distribution of contraceptives were subjected to strict regulations, resulting in soaring sales of 'discreet' ordering services.

Several 'sex problems' of the war and postwar period, such as sexually transmitted diseases and prostitution, became less significant in the wake of medical advance and the stabilization of the family system. And even though rates were somewhat higher than in the 1930s, the topic of illegitimacy and abortion also faded into the background. As economic prospects began to brighten, many illegitimate children, often conceived by long-time couples who had postponed marriage in the hope of a more stable economic future, were legitimized by fast

marriages. Homosexuality, on the other hand, largely remained an indecent topic, yet the predicaments of – mainly male – homosexual life managed to reach the public via press releases on sensational arrests and legal proceedings.

Homosexuals were seen as the antithesis of the heterosexual married couple; Finland's media became particularly concerned "when the issues of family, motherhood, and women's employment were also very much in the forefront" (Löfström, 1997: p. 33). The statements of medical doctors, psychologists, lawyers, and politicians reinforced prejudiced stereotypes, even in countries where they did not strictly prohibit homosexual acts. In France, homosexuals were blamed for refusing to procreate and were equated with Nazis and Arabs (Copley, 1989: p. 218). More open discussions were taking place in Britain with the establishment of the British Wolfenden Committee (1957), which suggested decriminalization of consensual homosexual acts among adults (over 21 years of age) and same-sex prostitution.

The 1950s and the early 1960s can be perceived as a period of conservative sex morals, yet they were also marked by a media-hyped boom of eroticism. A commercialization of sex took place in popular culture, particularly in the music and movie industries, as investors realized the potential value of sex. 'Elvis the Pelvis' and other rock 'n' roll musicians began to attract the attention of young people. 'Wild' dancing styles opened new opportunities to young men and women to express themselves more openly than to their parents from the frugal and sacrificing war generation. 'Sexy' movie idols, such as James Dean in 'Rebel Without a Cause' (1955), played angry and rebellious, and yet highly attractive young heroes. Bikinis and 'atomic boobs' adorned the female body on magazine covers. Beauty contests became the 1950s' platform for marketing and adapting erotic body ideals to national and regional styles. Smut and men's magazines such as 'Razzle' and 'Playboy' (1953) published articles along with pictures of pinups (including the famous 'Vargas Girls') in glamour shots. American-style movie, music, and youth magazines conveyed 'gossip' and fashion to youngsters directed at the new 'teen' market. Everyone who had the opportunity went to see one of the new scandalous erotic movies showing naked actresses. Françoise Arnoul shed her clothes in 'L'Epave' (1949) and 'Le fruit défendu' (1952), and Hildegard Knef stepped naked (if only for a couple of seconds) in front of the camera as an artist's model in 'The Sinner' (1950).

Parent's adverse reaction to the media's sex symbols made them all the more appealing to their teenage children and ensured their attraction to the forbidden erotic culture. When Ingmar Bergman's *Tystnaden* (The Silence, 1963) depicted scenes of masturbation and sex scenes devoid of love or affection, the media cried out in protest, yet few people resisted going and marveling at the sensation: Some 10 million spectators made the movie a great commercial success.

A Revolution, or: Catching Up on Liberalization (Mid-1960s to 1980s)

Sexual lives in (western) Europe had, in fact, been 'revolutionized' decades prior to the so-called sexual revolution of the late 1960s and the 1970s; public moral standards and attitudes merely trailed behind established practice. What we refer to as 'the sexual revolution' is therefore not so much a swift and radical change of sexual conduct as a culmination of a lengthy process. From the mid-1960s onward, this process was widely discussed in Western societies and within a few years had led to a liberalization of attitudes and sex-related norms. In a strictly historical context, therefore, one should perhaps refer to a sexual 'liberalization' than 'revolution'. Without doubt, the discourse on liberalization via media, politics, youth, and student movements contributed to convincing more and more people to adopt the new liberal approach, eventually making them standard. Yet the sexual liberalization's entire impact was not entirely revealed until the 1970s, when politicians succumbed and modified the provisions of the law that had been in force for decades or even centuries. The new sexual culture showed itself in continuously decreasing marriage and birth rates. In devout Catholic countries such as Italy, Spain, or Portugal, the 'sexual revolution' came a few years later. However, Spain, immediately following the end of Franco's dictatorship in 1975, was flooded by a radical sex and porn wave.

During the 1960s, rapid economic development saw the change from productive to consumption society that made 'sex' a positive asset of culture and society. For the first time people were able to forget the years of deprivation and poverty and turn their attention to consumption and experience. Even in countries most affected by the war, the economy turned from demand-oriented to supply-oriented – more goods and services were available than demanded. Marketing and advertisement soon utilized sex as a communication tool to support products' and media's competitive edge (Reichert, 2003: pp. 133–166). Although traditional sex norms of the Churches were still in place, certain laws were relaxed: Business investors quickly took advantage of the opportunities that had opened up to make money from trading in erotic and sex-related products. Meanwhile, while sex has still being widely considered a 'dirty secret' by their parents' generation, young people happily absorbed any information on the subject. Yet sexualization was not a phenomenon restricted to youth and student culture. It was widespread, having initially emerged among the working class and lower middle class before reaching better-off juveniles and students.

Furthermore, liberal attitudes were supported by political and democratic developments such as feminist and antiracism movements. Increasing individualization also had its impact, allowing people (young people in particular) to develop their own and very private living. This private space became more widely accepted as a realm where one had a right to pursue one's sexual pleasures (provided they were within the law). In some countries, several years passed before governments changed their laws and fell in line with popular practice. In Holland, it was the Catholic and Protestant clerics that advocated a positive approach to sex life and sexual questions regarding education or marriage – including contraception and divorce.

One of the prevalent factors propelling the change in attitude toward sexuality was, indubitably, the popularization of sexual knowledge provided in guidebooks and educational movies. They shifted attitudes from focusing on marital intercourse to an imperative to the pursuit of one's own individual

experimental experiences. Professional sex 'science' and psychiatry remained, for the time being, dedicated to sexual pathology, reaching but a small circle of interested intellectuals.

Schools also eventually opened up the topic for debate: Sex education classes were discussed for years before finally being put in place in the 1970s. This move also revealed the need to first educate the schoolteachers who were to teach the subject, not only the social and cultural aspects of sex life but also the biological ones.

From the middle of the 1960s onward, magazines depicting half-naked female bodies eagerly reported on various ways and practices of 'physical lovemaking'. At the same time, the first 'modern' educational movies were shown. From 1970, the commercial exploitation of the sexualized – mostly female – body served the basis for movies such as *Schulmädchen-Report* (*Schoolgirl Report*), viewed by some 25 million spectators all over Europe.

During the late 1960s and early 1970s the pornography laws were abolished. Denmark was the first to legalize the distribution of 'classical' erotic literature in 1979. Two years later, all the prohibitions of erotic texts, images, and objects were dropped. In October 1969, Copenhagen hosted the world's first 'erotic fair', attracting national and international – mostly male – visitors of all social classes. Sweden overturned its pornographic laws in 1971, and advanced quickly to be a dominant supplier of related products. By the middle of the 1970s, pornography was available to all adults in most other European countries.

Representatives of the new feminist movement criticized the sexual objectification of the female body, and objected to the pandering to men's power-obsessed domination fantasies. Nonetheless, the market for erotic and soft-core magazines such as the 'weekend magazines' flourished, and hard-core images and movies began to appear, concentrating on showing the sex act at the expense of a proper plotline. Television, now the leading medium during the 1960s, eventually fell in line.

'The pill' made a huge impact on popular culture. In 1962, the first hormonal contraceptive 'Anovlar' was sold in Europe. By the middle of the 1960s, advertising and press reports had generated sufficient publicity for the oral contraceptive to be prescribed and purchased on a larger scale. A survey from 1989 revealed that around 75% of all English women born between 1945 and 1959 were using the pill, lifting the threat of unwanted pregnancy from their sex lives (Cook, 2004: p. 281). For the first time, young women could take control of their sex lives without requiring their partner's support; the pill was also reliable, relieving much of the anxiety of unwanted pregnancy. Meanwhile, feminist movements voiced their concerns over the lack of surveys on long-term use and possible adverse reactions, and suggested that women had once again ended up with yet another burden of unshared responsibility.

The youth movement was another force propelling sexuality into public debate. Striving to confront their authoritarian parental generation and the capitalistic consumerism they represented, young people banded together to establish alternative lifestyles involving flower power, beat music, and drugs. They introduced the concept of 'free love', the message 'Let's Spend the Night Together' proclaimed by the Rolling Stones acting as the leitmotif for popular youth culture and their leading medium, rock music. The spirit of freedom in the Summer of Love (1967) was seized by the youth movement as the beginnings of a new time of sexual liberation with an expectation that a revolution of general social standards would follow. The most prominent example was Berlin's 'Kommune 1' who conveyed their criticism on society via sex scandals. 'Make Love Not War' was their motto, aimed especially against the Vietnam War. Wilhelm Reich and Herbert Marcuse supplied them with the theoretical ammunition to develop a sex-based explanation for the bourgeois repression of sexuality. According to these authors, liberalization of sex was mandatory in order to create a more just and nonrepressive society. The limitations of the youth and student movements were exposed by the women's liberation movement of the 1970s, who equated social dominance of men with phallic dominance in bed and put a noisy end to the 'Myth of Vaginal Orgasm'.

As a result of public debate, laws against abortion were abolished. In Germany, for instance, hundreds of women in a spectacular campaign of 'Stern' magazine (1971) admitted that they had had abortions, pointing out the uselessness of the law. In the wake of the protest movements and their exposure in the media, the laws were changed in various countries and abortion was permitted within the first 3 months of pregnancy or under special circumstances agreed by the doctors (GDR 1972, Austria 1973–75, Sweden 1974). On a practical basis, however, it took years to supply the required number of doctors and medical outpatient clinics sufficiently equipped to undertake the demands for abortion.

The 1970s saw the liberalization of regulations regarding family and divorce, both for contracting and dissolving marriages of two (heterosexual) partners based on their mutual consent alone, civil contracts were allowed, devoid of any religious ceremony. Gays and lesbians vehemently demanded legal and social emancipation with the 'Stonewall Riots' attracting considerable media attention: On 27 June 1969, for the first time homosexuals openly rebelled against police harassment and discrimination after a raid at the 'Stonewall Inn' located at New York's Christopher Street.

In western European countries, gays and lesbians made their voices heard in demonstrations and demanded change(s). Laws against homosexuality were annulled, but different regulations concerning the age of consent, matrimony, hereditary, and adoption remained unchanged for decades to come. The cultural identity of gays and lesbians, however, did change during the 1970s. Traditional role concepts of effeminate men and masculine women faded to be replaced by so-called normal homosexuals who increasingly adopted their own genders' identities. Most homosexual couples now socially positioned themselves akin to heterosexuals, and formerly marked age differences between same-sex couples diminished.

Comparative studies on German students confirmed the shifting of attitude toward homosexuality. Between 1966 and 1981, they increasingly tolerated homosexuals while their own heterosexual orientation became more fragile. At the same time, heterosexualization became further entrenched in the 1970s and 1980s, seen in decreased homosexual contact during adolescence (Schmidt et al., 1998). This was a result of coeducation and better sex education. All those having sexual

contact with another person of the same gender were now forced to contemplate the possibility that they were homosexual and would be subject to stigmatization as homosexual.

In general, as far as sexual behavior and attitudes of 'conservative' groups were concerned (of which students were a part), they both changed rapidly. As has been the case with young working class members during the 1950s and 1960s, premarital sexual intercourse now grew to be standard among other classes, and the number of sex partners and frequency of intercourse increased. The employment of a variety of sex techniques became a sexual imperative just as did the acceptance of masturbation prior to and during partnership or matrimony. This progress was most obvious among women. Premarital virginity ceased to be regarded as a virtue; young people mostly rejected the sexual double standard as a relic from their parents' generation. Supported by the feminist movement, sex guidance literature, and the media, more and more women actively demanded that their specific sexual desires should be satisfied.

Sex between Negotiation, Risk, and Virtualization (the 1980s to Present Day)

As a result of liberalization in the late 1960s and the 1970s, sex was soon fully integrated into the commercial supply of the Western market. Products and services were sold in the media as sexual promises. This oversexualization rapidly reduced the effect of sexual stimuli forcing the advertising industry to push back the barrier of what could or should be explicitly depicted (Reichert, 2003: p. 364). Visual sexual stimulation was easily obtained in the rapidly expanding market of erotic and pornographic magazines and movies, offered for sale at every kiosk or video store's adult section. This gave birth in the 1980s to a feminist movement named PorNo, fighting against pornographization of the female body and the accompanying violent sex fantasies. At the peak of the sex war in the middle of the 1980s, the battles involved "the regulation of pornography, the scope of legal protection for gay people, the funding of allegedly 'obscene' art, the content of safe-sex education, the scope of reproductive freedom for women, the extent of sexual abuse of children in day care centres, the sexual content of public school centers, and more" (Duggan and Hunter, 1995: p. 1). Alternatively, 'porn for women' voted for a different kind of eroticism and pornography that would cater to the needs of women.

A satisfactory sex life had now come to be a requirement for (post)modern men and women. Terms such as 'performance', 'innovation', and 'technique' entered the sexual arena, aimed at improving one's own and one's partner's pleasure. Sex manuals suggested remedies for impotence and suggested on how to work on lack of sexual interest. Therapists and sexual medical doctors offered their services, and in the middle of the 1990s, Viagra came to the aid of those suffering from impotence. Those who did not manage to act out their fantasies and desires with a lover or sex partner would soon find substitutes in new technology such as telephone or the Internet sex. Increasingly, masturbation came to be regarded as a separate autoerotic sexual realm offering uncomplicated sexual satisfaction or simply relaxation.

In the face of affordable; easily obtainable; and, most importantly, safe contraceptives, sex was no longer tied to procreation. Meanwhile, new reproduction technologies (such as *in vitro* fertilization) detached procreation from sexual desire. Sexually transmitted diseases lost some of their threat owing to effective remedies. During the 1980s, however, AIDS became the new fear but was then thought to affect only a marginalized minority. Moralizers even claimed the 'gays' pest' to be a divine retribution for the promiscuity of homosexuals. Nevertheless, the AIDS panic only held up the increasingly liberal approach to homosexual men and women for a short while. Intense coverage of homosexuality in the media even opened up new insights into the social and relational cultures of gays and lesbians – with it came an increase in tolerance and acceptance. During the second half of the 1980s, it became evident that unprotected heterosexual intercourse had led to a rise in the spread of VD, putting an end to sex without worries and unprotected sexual liberty.

During the 1990s and the early twenty-first century, horrifying cases of pedophilia, child pornography, and sexual murder of children were exposed. Although the number of such cases had not increased – most assaults on children still occurred within the circle of family and acquaintances – the number of occurrences that were uncovered did increase, mainly owing to a higher sensitivity to these offenses. The highly emotional media coverage did, however, open the debate on age of consent and possible sexual dependencies between adults and children/adolescents. For some groups of migrants, sexual liberalization and sexualization went beyond their realms of tolerance that were defined by their traditional cultures' approaches to gender and sexuality matters. Violent assaults on the members of the gay community were reported from some of the new EU member states, where after 1989 latent resentments erupted as new homophobia.

In the 1990s also for adolescents and young adults, solid relationships and sexual fidelity grew more important, again as a reaction to a culture that had abandoned traditional sex morals and replaced them with negotiation or consent morals (Schmidt, 1996). The desire to experiment in promiscuous relationships was increasingly replaced by a need for closeness and safety in a relationship. Now also the male youths increasingly connected sex to love and emotional ties and young women and girls demanded sexual pleasure more actively (Sachdev, 1993: p. 105). Using contraceptives was by now commonplace, yet first sexual contacts often still took place without protection, mainly due to embarrassment. In comparison with the sexual 'revolutionists' of the 1960s and 1970s, both men and women had come to perceive sex as a rather nonmythical and undramatic part of life, with no requirement for liberations or attachments to a political Utopia. This does not necessarily imply a sexual counterrevolution. The rigid attitude that had prevailed prior to the sexual liberalization never reestablished itself (McLaren, 1999: pp. 193–205). But in comparison with the liberal 1970s, some people displayed less tolerance to promiscuousness, to premarital and extramarital sex and homosexuality. "This ending of liberal advance might reflect a homeostatic cycle of reform or a reaction to problems created by liberalism in

general or sexual permissiveness in particular (e.g., increases in sexually transmitted diseases and non-marital births, possibly a desire for commitment rather than casual, recreational sex, etc.)" (Smith, 1994: p. 91). In Europe, the fueling of 'sex panics' repeatedly proved useful in introducing neoconservative and restrictive moral standards to everyday political discourse (Irvine, 2002).

The mass production of sexual cultural images and texts has had the effect of blurring gender and sex identities. Beyond polarized heterosexuality and homosexuality, there exists a wide range of sexualities today: hetero-, homo-, and multisexuals, androgyny, queer, and transgender sexualities, and so forth. Homosexual desire is widely accepted within the 'normal' range of sexualities – which is one of the reasons why sexuality has ceased to be a suitable basis for the constitution of a specifically gay or lesbian identity. According to today's confessional TV shows, everybody should or could be a little 'perverted'. The fetishizing of images of sex has not stopped at normalizing perversion. Currently, the Internet offers a conglomeration of diverse sexual practices, some of which, only a few decades ago, would have been subjected to penal law or, at the very least, investigation by psychiatrists. Furthermore, intimate bodily contact had been replaced by cybersex, the autoerotic pleasures of steel prostheses nipple claps, and machine-generated stimulation such as the Sybian (a masturbatory device for women), but such cold technology failed to take hold.

After the sexual liberalization of the 1960s and 1970s and the gender revolution of the 1980s, the 1990s saw a radical change in relationship culture – matrimony decreased and long-term cohabiting partnership rose. The preferred sexual relationship is now one in which both partners communicate on an equal basis (Giddens, 1992). Whether the relationship endures or not is influenced by mutual compatibility; which includes as a central factor; excitement, and satisfaction in one's sex life. As was the case in the 'romantic love' partnership of the eighteenth and nineteenth centuries, sex is now being integrated into the overall vision of a loving relationship. Instead of gaining only satisfaction in sex, a rounded loving relationship is being sought – yet sex is only one component of consumption, leisure, and individualism, an everyday experience along with many others. But the egalitarian, emotion-oriented, and soothing sex ideal misjudge vital constituents of sexual desire, such as difference in power relations and sexual tensions, and this may be one of the reasons why the sexual desire decreases instead of increases (Sinfield, 2004). Also, in most Western nations, criticism of sexual liberalization is increasing, with increasing authoritarianism eradicating freedoms gained since the 1960s and 1970s. Recent surveys have revealed that the number of people who are dissatisfied with their sex lives is rising – possibly a result of the growing gap between the media produced ideal, in all its glossy variations; and the unvaried everyday bedroom practices between spouses or partners. Compared to the other increasing differentiations in the world of consumption, Western capitalism has not diversified the average citizen's sex life, but restricted it (Hekma, 2005: p. 220): In the past two decades, normalized heterosexual monogamy has gained momentum, leaving promiscuous, homosexual, or prostitutive sex practices far behind. Prior to the 'sexual revolution', Europeans practiced sex in every way they pleased regardless of the law or moral reprobation. Now, it seems, the highly sexualized Western contemporary culture ignores the practices and forces us to make use of whatever sexual images and fantasies are available to us.

A longer version of this article was published as: Franz X. Eder, Sex, Popular Beliefs and Culture. In: Gert Hekma (Ed.), A Cultural History of Sexuality in the Modern Age. Oxford/New York 2011, pp. 149–175 and 243–246.

See also: Gay Men's Emancipation Movement in Germany: 1897–1933; Gay and Lesbian Movements; Heterosexism and Homophobia; Pornography; Prostitution and Sex Work; Queer Geography; Rape and Sexual Coercion; Sexual Behavior and Social Networks; Sexual Orientation and the Law; Sexual Orientation: Historical and Social Construction; Sexual Risk Behaviors.

Bibliography

Beck, Birgit, 2004. Wehrmacht und sexuelle Gewalt: Sexualverbrechen vor deutschen Militärgerichten 1939–1945. Schöningh, Paderborn.

Cook, Hera, 2004. The Long Sexual Revolution: Women, Sex, and Contraception in England 1800–1975. Oxford Univ. Press, Oxford.

Copley, Antony, 1989. Sexual Moralities in France, 1780–1980: New Ideas on the Family, Divorce and Homosexuality. Routledge, London/New York.

Duggan, Lisa, Hunter, Nan D. (Eds.), 1995. Sex Wars. Sexual dissent and Political Culture. New York.

Eder, Franz X., 1999. Sexual cultures in Germany and Austria, 1700–2000. In: Eder, Franz X., Hall, Lesley A., Hekma, Gert (Eds.), Sexual Cultures in Europe: National Histories. Manchester Univ. Press/St. Martins Press, Manchester/New York, pp. 138–172.

Eder, Franz X., 2002. Kultur der Begierde: Eine Geschichte der Sexualität. C.H. Beck, München.

Eder, Franz X., 2007. The Nationalists' 'healthy sexuality' was followed by America's influence: sexuality and media from National Socialism to the sexual revolution. In: Günter, Bischof, Anton, Pelinka, Josef, Kostlbauer (Eds.), Sexuality in Austria, Contemporary Austrian Studies, vol. 15. Transaction Publishers, New Brunswick/London, pp. 102–130.

Franz X. Eder. Bibliography of the History of Western Sexuality. http://www.univie.ac.at/Wirtschaftsgeschichte/Sexbibl with over 25,000 entries.

Giddens, Anthony, 1992. The Transformation of Intimacy. Sexuality, Love, and Eroticism in Modern Societies. Stanford Univ. Press, Stanford.

Hall, Lesley A., 2000. Sex, Gender and Social Change in Britain since 1880. Macmillan Press, Houndmills/London.

Hanson, Dian (Ed.), 2004. The History of Men's Magazines: vol. 1: From 1900 to Post-WW II. Taschen, Cologne, Germany.

Hekma, Gert, 2005. How libertine is the Netherlands? Exploring contemporary Dutch sexual cultures. In: Bernstein, Elizabeth, Schaffner, Laurie (Eds.), Regulating Sex. The Politics of Intimacy and Identity. Routledge, New York, pp. 209–224.

Herzog, Dagmar, 2005. Sex after Fascism: Memory and Morality in Twentieth-Century Germany. Princeton Univ. Press, Princeton/Oxford.

Hinds, Harold E., 2006. A holistic approach to the study of popular culture. Context, text, audience, and recoding. In: Hinds, Harold E., Motz, Marilyn F., Nelson, Angela M.S. (Eds.), Popular Culture Theory and Methodology. A Basic Introduction. Univ. of Wisconsin Press, Madison/London, pp. 163–180.

Irvine, Janice M., 2002. Talk about Sex. The Battles over Sex Education in the United States. Univ. of California Press, Berkeley.

Kral, Silke, 2004. Brennpunkt Familie. 1945 bis 1965. Sexualität, Abtreibungen und Vergewaltigungen zwischen Intimität und Öffentlichkeit. Jonas Verlag, Marburg.

Löfström, Jan, 1997. Sexuality and the performance of manliness: sketching the historical trajectory of male fear. Ethnologia Scandinavica 27, 21–40.

Martin, Michelle H., 2003. No one will ever know your secret! Commercial puberty pamphlets for girls from the 1940s to the 1990s. In: Nelson, Claudia, Martin, Michelle H. (Eds.), Sexual Pedagogies: Sex Education in Britain, Australia, and America, 1879–2000. Palgrave Macmillan, New York, pp. 135–154.

McLaren, Angus, 1999. Twentieth-Century Sexuality: A History. Blackwell Publishers, Oxford.

Quine, Maria S., 1996. Population Politics in Twentieth-Century Europe: Fascist Dictatorships and Liberal Democracies. Routledge, London.

Reichert, Tom, 2003. The Erotic History of Advertising. Prometheus Books, Amherst.

Sachdev, Paul, 1993. Sex, Abortion, and Unmarried Women. Westport/London, Greenwood.

Schmidt, Gunter, et al., 1998. Veränderungen des Sexualverhaltens von Studentinnen und Studenten 1966–1981–1996. In: Schmidt, Gunter, Strauß, Bernhard (Eds.), Sexualität und Spätmoderne. Über den kulturellen Wandel der Sexualität. Enke, Hamburg, pp. 118–136.

Schmidt, Gunter, 1996. Das Verschwinden der Sexualmoral: Über sexuelle Verhältnisse. Klein, Hamburg.

Schütz, Alfred, Luckmann, Thomas, 2003. Strukturen der Lebenswelt. UVK Verlagsgesellschaft, Konstanz.

Sinfield, Alan, 2004. On Sexuality and Power. Columbia Univ. Press, New York.

Smith, Tom W., 1994. Attitudes toward sexual permissiveness: trend, correlates, and behavioral connections. In: Rossi, Alice S. (Ed.), Sexuality across the Life Course. Univ. of Chicago Press, Chicago.

Population Aging: Economic and Social Consequences

Jacques Légaré, Université de Montréal, Montréal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Demographic aging is a modern phenomenon: never before has there been a society that was so old, although within societies there were some old individuals. As of now, all societies will become old because they control both their mortality and fertility; they will be characterized by demographic transition and as a result, demographic aging. This change inevitably brings demographic and socioeconomic consequences. To be properly managed, they must be empirically determined and rationally studied, because the aging of populations is one of the great achievements in the history of civilization. Increasing numbers of people are reaching retirement and old age, and their relative weight in the population will be greater than ever before. The costs of health care and pensions will increase and their method of financing is a concern to many. In countries that do not have social security systems, innovative systems will have to be established, and in those that do, in-depth reforms will be required to insure that beyond traditional solidarity, we secure for the coming generations a degree of intergenerational equity.

The notion of aging is quite natural for human beings: from the age of reason, we understand that with the passage of time, every living thing ages and finally dies. Young children experience this with their goldfish or dog, adults with their parents, and old people with their friends. By analogy, we apply the notion of aging even to nonliving things: for example, as a car deteriorates, we say that it is getting old and eventually we relegate it to the scrap lot. Although a good wine may improve with age, in the end it becomes undrinkable.

From the individual, the notion of aging is extended to the group formed by the individuals: thus, we speak of an aging fleet of cars or an aging population, meaning that the average length of life of the entire group of these individual units changes with the passage of time (*see* Ecology of Aging). However, unlike the case for the individuals making up the group, the group itself is able not to age, or even to get younger. This is the case for human populations. For them, the aging process is strictly structural, not biological. The decay generally associated with the aging of individuals should not be assumed to apply to aging populations.

Population Aging: A Modern Demographic Phenomenon

A population changes through the interaction of the fertility, mortality, and migration of its members. In recent centuries, we have witnessed remarkable changes in human fertility and mortality, which, uncontrolled since time immemorial, had always been high. In the past, there was a 'natural' fertility and mortality, and population numbers throughout the world were relatively low. Eventually, people found ways to reduce first mortality and then fertility, after a certain time lag. In theoretical terms, this historical process is known as the demographic transition (Chesnais, 1992). During this process, the population grows at a higher or lower rate, depending on the extent of this time lag. At the end of the process, according to the theory, both fertility and mortality are low and the population stops growing. But an inescapable law of demography says that a population either grows or it ages (Coale, 1987). As all populations around the world have reduced or stopped their growth (United Nations, 2011), they are all bound to age. In some countries the fertility level is well below the replacement level, and thus their populations will decrease unless they experience in-migration, a possibility not foreseen by classical transition theory.

There are several ways of measuring the aging of a population. Instinctively, one could say that a population is aging if the average – or median – age of its members as a group increases over time. Alternatively, one could look for an increase in the proportion of old people in the population. But at what age does a person become old? We do not want to enter into a debate over the exact threshold of old age or retirement, but for purposes of illustration we will use the percentage aged 60 and over in the population (*see* Age Stratification; Age Structure). To avoid delving too far into the past, we will restrict our discussion of demographic aging to the period from 1950 to 2100.

During the twentieth century, the growth in the number of people aged 60 and over in different parts of the world was relatively small, and there were roughly as many people old (according to this definition) in more developed regions as in less developed regions. However, from now until 2100, this number will grow spectacularly, mainly because of the increase in less developed regions (**Table 1**). But if we look at the proportion of people aged 60 and over in the population as a measure of demographic aging between 1950 and 2100 (**Table 1**), we see that the situation differs greatly from one continent to another. Africa will remain young while the more developed regions as a whole will have a bit less than one-third of their populations in the 'old' category. Asia will age at an increasing rate: in 2100, it will have near 1500 million people aged 60 and over. This is half of the world's elderly population.

In the industrialized regions, the process of population aging is relatively slow, but in developing regions it will be more intense and rapid. This can be clearly shown by a comparison of Canada and the BRICS countries, a group of emerging economies formed by Brazil, Russia, India, China,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.34041-7

Table 1 Population aged 60 and over, 1950–2100

Regions	*1950*	*1975*	*2000*	*2025*	*2050*	*2075*	*2100*
(a) Population aged 60 and over, in thousands							
World total	204 304	348 177	609 286	1 198 458	2 031 337	2 538 904	2 831 027
More developed regions[a]	95 245	161 851	231 926	350 298	418 326	412 719	432 831
Less developed regions[a]	109 059	186 325	377 360	848 160	1 613 011	2 126 186	2 398 196
Least developed countries[a]	10 267	17 385	33 549	72 353	181 568	361 437	561 814
Less developed regions, excluding least developed countries	98 792	168 940	343 811	775 806	1 431 442	1 764 749	1 836 381
Less developed regions, excluding China	67 114	119 410	243 995	557 613	1 161 352	1 716 404	2 068 354
Africa	12 226	20 873	41 963	90 828	215 249	434 686	716 398
Latin America and the Caribbean	9392	21 057	43 558	101 371	187 869	237 917	236 282
Northern America	21 288	35 310	50 985	95 879	120 797	142 732	165 564
Asia	93 643	157 176	320 725	698 549	1 252 588	1 486 421	1 473 324
Europe	66 338	111 390	147 880	203 309	241 828	220 389	219 435
Oceania	1417	2371	4175	8522	13 006	16 759	20 023
(b) Population aged 60 and over, percentage of total population							
World total	8.1	8.5	10.0	15.0	21.8	25.6	28.0
More developed regions[a]	11.7	15.5	19.5	27.2	31.9	31.5	32.4
Less developed regions[a]	6.3	6.1	7.6	12.6	20.2	24.7	27.3
Least developed countries[a]	5.2	5.0	5.1	6.3	10.5	15.9	20.9
Less developed regions, excluding least developed countries	6.5	6.3	8.0	13.9	22.8	27.9	30.1
Less developed regions, excluding China	5.8	5.7	6.7	10.5	17.4	22.9	26.4
Africa	5.3	5.0	5.2	6.4	9.8	14.7	20.0
Latin America and the Caribbean	5.6	6.5	8.4	14.9	25.0	32.2	34.4
Northern America	12.4	14.6	16.3	24.7	27.0	28.9	31.5
Asia	6.7	6.6	8.6	14.8	24.4	29.9	32.1
Europe	12.1	16.5	20.3	27.3	33.6	32.3	32.5
Oceania	11.2	11.0	13.4	19.1	23.5	27.0	30.4

[a]More developed regions comprise Europe, Northern America, Australia/New Zealand, and Japan. Less developed regions comprise all regions of Africa, Asia (excluding Japan), Latin America, and the Caribbean plus Melanesia, Micronesia, and Polynesia. Least developed countries are 48 countries, 33 in Africa, 9 in Asia, 5 in Oceania, plus one in Latin America and the Caribbean.
Source: United Nations Department of Economic and Social Affairs, 2011. World Population Prospects: The 2010 Revision. United Nations, New York.

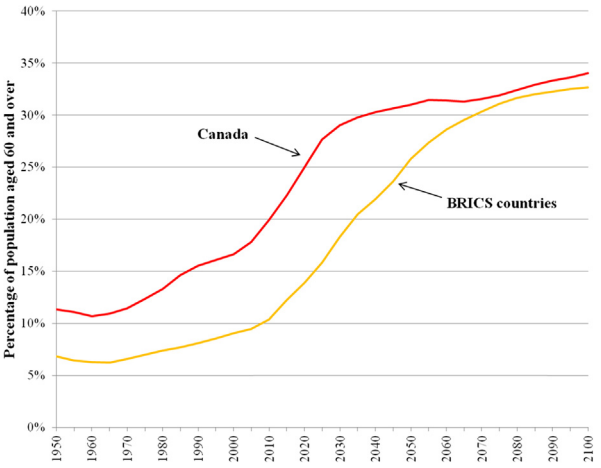

Figure 1 Population aging in Canada and BRICS countries, 1950–2100. Source: United Nations Department of Economic and Social Affairs, 2011. World Population Prospects: The 2010 Revision. United Nations, New York; Laïdi, Z., 2012. BRICS: sovereignty power and weakness. International Politics 49, 614–632.

and South Africa (Laïdi, 2012) (Figure 1). In 2100, BRICS countries will be world leaders in demographic aging with almost 930 million people aged 60 and over. This is more than 30% of their total population.

In the face of globalization, the changes occurring over this century will give rise to a number of socioeconomic issues that would not have made themselves felt if population aging affected only the currently industrialized countries. There will be both demographic and economic and social consequences for our societies.

Demographic Consequences

As is easily demonstrated, in the early stages of the demographic transition aging is much more the result of a decline in fertility than in mortality. When mortality declines from a high level, it is primarily early deaths that are eliminated, in particular deaths of children and young adults. The result is a younger (not older) population, as was seen in Africa between 1950 and 1975 (Table 1).

But when life expectancy is already very high, as is the case now in industrialized countries, mortality becomes an

important factor in demographic aging. If mortality is low and decreasing, we get a double effect. First, increasing numbers of people reach the old age threshold – because of lowered mortality before that age. Second, people live for increasing lengths of time beyond this threshold – owing to a continued decline of mortality at old age (Shumanthy et al., 2012). The main consequence of this development is that old people no longer form a homogeneous group: we now have the young old, who generally are in quite good health, and the oldest old, who generally are more dependent for income security and health. Thus, we have moved from having three main lifecycle stages to four. There is a new third age that extends from retirement to the oldest-old stage. This is an entity that never existed before in human history (Laslett, 1989). Another important feature in mortality changes in many societies has been an increased mortality gap between men and women. Even if attenuated in the present century (Avdeev et al., 2011), very frequently old people in these societies are and will continue to be primarily old women.

Like any other population, the elderly population is subject to a process of incessant inward and outward movement of cohorts. In most industrialized countries the incoming generations will be clearly differentiated with respect to their numbers and their characteristics. The generations now entering retirement are the baby boomers born after World War II. They will be followed by the baby busters, who are far less numerous. These generations, particularly the women, will have characteristics very different from those of today's old people (Légaré and Ménard, 2007), especially in terms of education.

The changes in the age structure of our societies will cause major realignments that affect not only the elderly but the entire population. In terms of health and financial security, will old people be more independent or more dependent than before? The increasing presence of very old people tends to suggest dependence, while the characteristics of the future younger elderly appear to suggest greater autonomy. Will the younger adults, who are far less numerous than the baby boomers, be willing to pay for a golden twilight for tomorrow's seniors? We now turn to these questions in the following sections.

Economic Consequences

The Aging Process and the Working-Age Population

At the beginning of the population aging process, a standard age pyramid shows the population cohorts by age increasingly reduced by mortality. As the process gradually proceeds, positive impacts on the working-age population appear. The decline in mortality eliminates early deaths and increases population numbers. Then, as the effects of fertility decline come to the fore, cohort size is increasingly reduced. When mortality improvements lead to significant reductions in old-age deaths, the members of the retired cohorts become as numerous as, and sometimes more numerous than, those of working-age cohorts. The pyramid changes into a top and finally into a cylinder. This means that at the end of the aging process, all generations are of approximately the same size except for the very old. This process is reflected in the variations of what are commonly called the dependency ratios, one of which is the ratio of people aged 65 and over to those aged 19–64. Unfortunately, the use of the adjective 'dependency' may unduly give an impression of economic dependence. These indicators are merely indices of the age structure and it would be more appropriate to speak of demographic dependency ratios (*see* Age Structure).

In studies of aging we frequently find the members of the working-age population equated to actual workers. While in the past this was true for men when their labor force participation rate was close to 100%, it is now less so, in particular for women. To assess the economic impacts of aging and our ways of dealing with them, we needed to take into account not only population numbers, but perhaps more importantly, labor force participation rates and worker productivity (OECD, 1998). If the shift from agricultural to industrial societies had great impacts on the world of work, the shift to aging postindustrial societies will have an equally great impact. This is particularly true for the change in the rate of labor force participation by women and by older workers.

The radical decline in fertility and otherwise changing family arrangements has freed women from a myriad of family duties and made them available for increased participation in the paid labor market. The women's liberation movement also contributed to changes in attitudes and women are becoming increasingly independent, not only at working age, but also, and perhaps especially, during retirement.

One of the corollaries – and achievements – of the development of industrial societies has been the right to retirement and associated pensions. In the second part of the twentieth century, we have witnessed a decline in the labor force participation rate of older workers, along with a systematic decline in the age of leaving the labor market. This ever-earlier retirement from the labor market has now been reversed. As life expectancy increases, especially at old age, and as the time spent taking education is extended, the life segment in retirement is growing longer and the segment in which one works is getting shorter. When a large and growing share of the costs of social security, including pensions and health insurance, is borne by those who work, some adjustments must be considered. It is not possible to maintain contribution and benefit levels and, simultaneously, to bring down, or even maintain, the age of access to benefits.

There is a need for greater flexibility, both in wages and responsibilities for the population approaching retirement. For older workers, one should consider a gradual progression from wages to pension income, perhaps following a normal rather than S-shaped curve. This may call for a change in attitudes, not only among the workers themselves but also among employers and trade unions.

Solidarity among generations presupposes a minimum of equity between groups. When social security programs in the industrialized countries were reorganized after World War II, both the population and the economy were growing: the populations were young. Pay-as-you-go financing did not create any visible problems, and it eliminated some perverse effects of funded programs, many of which had collapsed in Europe during the Depression and the war. In aging societies – particularly those which have experienced a large baby boom followed by a baby bust – a hybrid system combining pay-as-

you-go and funded elements is likely to be more equitable across generations.

At present, the labor market is undergoing major changes. Workers will no longer hold full-time jobs for life the way they did previously in industrial societies. New workers will enter the labor market later in life. They will hold part-time jobs with varying employers, more often and possibly for a significant part of their working life, or be self-employed. From time to time they will have periods of unemployment or of retraining. Moreover, such patterns, which formerly applied mainly to women, will now apply to men as well. Consequently, retirement systems based on previous labor market participation and involving individual employers are subject to in-depth revision. In addition, one should take the value of unpaid work into account. In particular, work performed to raise children or to care for aging parents needs to be recognized.

An Aging Society: A Burden?

Because population aging is equated to individual aging all too often, one tends to highlight problems for aging societies, in particular problems related to the costs of health and pension programs for the elderly. On this subject some issues need to be clarified and some myths should be dispelled.

First, the situation of old people in most countries has improved greatly largely as a result of post–World War II social programs. There have even been hard economic times when their situation has deteriorated less than that of younger people.

Second, health and pension costs will certainly continue to increase in aging societies, but at what rate will this happen? With regard to health Zweifel et al. (2004) have shown that increases in medical costs are more closely linked to the number of old-age deaths than to the size of the elderly population, and the number of old-age deaths is increasing much less rapidly than the size of the elderly population (Légaré et al., 2006). The same is not true for pensions, whose costs are linked directly to the number of old people. However, one should not forget that retired people also pay taxes, both on their income and on consumption. As people age, their material needs (other than those related to health) decrease, however, and this might be reflected in their pension requirements.

Third, although incomes are on average lower at old age than at other ages, the elderly often hold a large share of the collective wealth through their assets, in particular by owning of their dwellings (*see* Intergenerational Solidarity).

In brief, far from being an impediment to economic development, elderly people in aging societies can sometimes serve as an important actor if not the engine of such development (Dirn and Mendras, 1984).

Social Consequences

Increasingly, the group of old people is not homogeneous with regard to health status and wealth, and this has numerous consequences for their life-style. First, let us recall that to a large extent the world of the elderly is a world of old widows. This feature has become increasingly prominent over the twentieth century, as excess male mortality has grown while the difference in age at marriage has remained stable. In addition, the twentieth century has been characterized by the quest for greater independence at all ages. Old people in industrialized countries prefer to live alone if they do not live with their spouse, particularly when they are in good health. But this choice does not necessarily imply isolation. Alternatively, living together with people other than one's late spouse may become a strategy for dealing with financial or health problems, perhaps as a stage that precedes institutionalization.

Multigenerational households are much more common in less industrialized countries, mainly for cultural and economic reasons. We can only wonder what will happen in the future in these societies. In a context of rapid urbanization and scarcity of dwellings, there will be an acceleration in aging following the sharp decline in fertility. With its one-child policy, China is a particularly poignant case. Demographic constraints may become apparent as daughters and daughters-in-law become increasingly less numerous and less often available to take care of old parents.

In societies with low mortality (or equivalently, long life expectancy), early deaths are reduced to a minimum and are usually related to lifestyle rather than caused by exogenous diseases. This produces a great deal of interest in the population's health status, and public health policy objectives may concentrate on insuring quality of life for the aging population more than on preventing early adult deaths. The focus moves from mortality to morbidity, from life expectancy to healthy-life expectancy (Robine and Cambois, 2013). By means such as surveys on seniors' functional capacities, it is possible to identify impediments to independent existence and to develop policies that will minimize the adverse effects of biological and social aging.

Such policies are centered on support networks (*see* Age Policy). Assistance provided to old people may come from professionals hired by the state, the community, or private enterprise. Alternatively, they may be offered by the family, friends, or neighbors. The expanding need for services provided through such formal and informal networks is an increasing source of concern. For the informal networks, the many changes that have occurred in family structures leave future prospects unclear; for the formal networks, it is mainly the cost that worries public policy makers. In this picture one should not forget that seniors are not only receivers of assistance, but very often also providers.

Conclusion

Some believe that the twenty-first century will see the emergence of intergenerational conflicts similar to the gender conflicts of the twentieth century. Major transformations of the population age structure have altered the ratio between the retired and the working population and will continue to do so. In addition to solidarity among generations, any reforms to social security programs taking account of the new demographic reality need to aim to promote intergenerational equity. For this purpose, we have seen the development of new measurement techniques, known as demographic accounting (Auerbach et al., 1999). Even if not yet often applied, by making use of individual-level data and by making adequate

long-term projections we will be able to benefit from these new methodologies.

In conclusion, we observe that for centuries, life in all its forms was regulated by nature. Gradually, human societies became able to control their destiny by largely controlling their mortality and fertility. This resulted in great changes for the societies, and in particular it accelerated their aging. The fact that most now live to old age and that our populations progressively age can only be regarded as great cultural progress (Légaré, 2005). However, like all progress, individual and societal aging must be managed skillfully. Moreover, we are in the midst of a process of shifting control from society to the citizen (Légaré and Marcil-Gratton, 1990). Decisions concerning demographic behavior, like whether or not to get married, whether to have children, how many and when, whether to leave an existing marriage and contract another one, and so on, are now made on the individual level more than ever before. In a world where centenarians will become rather common, people will doubtless also want to have a say on the timing of their departure from life (Légaré and Marcil-Gratton, 1990).

See also: Adulthood: Dependency and Autonomy; Age Policy; Age Stratification; Age Structure; Age, Sociology of; Aging and Health in Old Age; Aging, Theories of; Cognitive Aging; Motivation: Life Course and Sociological Perspectives; Old Age and Centenarians: The Human 'Warranty Period'; Retirement and Encore Adulthood: The New Later Life Course.

Bibliography

Auerbach, A.J., Kotlikoff, L.J., Leibfritz, W. (Eds.), 1999. Generational Accounting around the World. University of Chicago Press, Chicago.

Avdeev, A., Eremenko, T., Festy, P., Gaymu, J., Le Bouteillec, N., Springer, S., 2011. Populations and demographic trends of European countries, 1980–2010. Population-E 66 (1), 9–130.

Chesnais, J.C., 1992. The Demographic Transition: Stages, Patterns, and Economic Implications. Oxford University Press, New York.

Cliquet, R., Nizamuddin, M. (Eds.), 1999. Population Ageing. Challenges for Policies and Programmes in Developed and Developing Countries. United Nations Population Fund, New York.

Coale, A., 1987. How a population ages or grows younger. In: Menard, S.W., Moen, E.W. (Eds.), Perspectives on Population: An Introduction to Concepts and Issues. Oxford University Press, New York.

Dirn, L., Mendras, H., September 1984. Le troisième âge animera la société française. Futuribles 80, 3–28.

Laïdi, Z., 2012. BRICS: sovereignty power and weakness. International Politics 49, 614–632.

Laslett, P., 1989. A Fresh Map of Life. The Emergence of the Third Age. Weidenfeld and Nicolson, London.

Légaré, J., Marcil-Gratton, N., 1990. Individual programming of life events: a challenge for demographers in the 21st century. Annals of the New York Academy of Sciences 610, 98–105.

Légaré, J., 2005. Economic, social and cultural consequences of the aging of the population. In: Casseli, G., Vallin, J., Wunsch, G. (Eds.), Demography: Analysis and Synthesis. A Treatise in Population Studies, vol. 3. Academic Press, New York, pp. 327–336 (Chapter 82).

Légaré, J., Bourbeau, R., Desjardins, B., Deblois, C., 2006. Variation in cohort size and lower mortality in the elderly: implications for pay-as-you-go healthcare systems. In: Yi, Z., et al. (Eds.), Longer Life and Healthy Aging. Springer, Heidelberg, pp. 305–318.

Légaré, J., Ménard, P.O., 2007. Generations succeed but do not resemble each other. The X, Y and baby-boomers generations in Quebec. Horizons 9 (3), 9–14.

Martin, L.G., Preston, S.H. (Eds.), 1994. Demography of Ageing. National Academy Press, Washington, DC.

Mathers, C., Boerma, T., Ma Fat, D., 2008. The Global Burden of Disease: 2004 Update. World Health Organization, Geneva.

Organisation for Economic Co-operation and Development (OECD), 1988. Ageing Populations. The Social Policy Implications. Organisation for Economic Co-operation and Development (OECD), Paris.

Organisation for Economic Co-operation and Development (OECD), 1998. Maintaining Prosperity in an Ageing Society. Organisation for Economic Co-operation and Development (OECD), Paris.

Organisation for Economic Co-operation and Development (OECD), 2000. Reforms for an Ageing Society. Organisation for Economic Co-operation and Development (OECD), Paris.

Robine, J.M., Cambois, E., 2013. Healthy life expectancy in Europe. Population & Societies 499.

Shumanthy, R., Bourbeau, R., Légaré, J., September 2012. Mortality in Canada for cohorts 1921–1981. In: European Association for Population Studies (EAPS). Working Group on Health, Tallinn, Estonia, pp. 5–7.

Suzman, R., Willis, D.P., Manton, K.G., 1992. The Oldest Old. Oxford University Press, New York.

United Nations Department of Economic and Social Affairs, 1956. The Ageing of Populations and Its Economic and Social Implications. United Nations, New York.

United Nations Department of Economic and Social Affairs, 1988. Economic and Social Implications of Population Ageing. United Nations, New York.

United Nations Department of Economic and Social Affairs, 2011. World Population Prospects: The 2010 Revision. United Nations, New York.

World Bank, 1994. Averting the Old Age Crisis: Policies to Protect the Old and Promote Growth. Oxford University Press, New York.

Zweifel, P., Felder, S., Werblow, 2004. Population ageing and health care expenditure: new evidence on the "Red Herring". The Geneva Papers on Risk and Insurance 29 (4), 652–666.

Population Change in the Former Communist States of Europe and Asia

Timothy Heleniak, George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 29 former Communist states of Europe and Asia underwent significant demographic change along with the political, economic, and social transitions during the 1990s and 2000s. This included steep declines in fertility, increases in mortality, and large-scale migration including large flows of refugees and internally displaced persons. These demographic trends lead to declines and rapid aging of the population of most countries in the region. In contrast to global population trends, most of the countries in the Eurasia region are projected to decline significantly in the future.

Introduction

The 29 formerly Communist, formerly centrally planned states in Europe and Asia (Figure 1) have undergone considerable economic, social, political changes over the past two decades which have had significant demographic impacts on these countries. This group of countries were once referred to collectively as the 'transition states,' because of the transitions they underwent starting in the late 1980s with the symbolic collapse of the Berlin Wall (World Bank, 1996). This consisted of three different transitions: an economic transition away from centrally planned economies toward market economies; a political transition of increased pluralism and control by lower levels of government marked most notably by the breakup of the Soviet Union, Czechoslovakia, and Yugoslavia, increasing the number of

Figure 1 The Eurasia region, 2010.

states in the region from 8 to 29; and a social transition toward more liberal societies.

While they share in common a legacy of at one time being part of the Communist block, they differ considerably in the degree and length of time that Communism influenced those societies. More importantly, they differed considerably in their levels of development while under Communism and have diverged even more so in the post-Communist period. They have also embarked on divergent political paths with some countries joining the EU and others becoming dictatorships. There is also considerable demographic diversity among these countries in terms of population size and growth and their stage of demographic transition. However, they are treated together because of their shared historical legacy. As is shown in this article, this shared legacy has resulted in the countries in the region combining demographic characteristics of both developed and less developed countries.

The article begins by examining the overall population change in the former Communist states during the two decades from the beginning of the economic, political, and social transition around 1990 to 2011. It then discuses the components of population change – fertility, mortality, and migration. A final section looks at the implications of the recent demographic trends and projections of the population of the countries in the region. For lack of a better term, Eurasia is used to describe the region, understood here as the current 29 former Communist states in Europe and Asia (Kosovo's independence is not universally recognized and given its recent independence, data are not always available). The year 1989 is chosen as the first year in many of the tables because it is the year of the last population census conducted in the Soviet Union. There are data issues in many countries in the region owing to the breakup of the three federal states and the necessity for all to make the transition from monitoring social and economic trends in centrally planned economies to market economies.

Population Change in the Transition States of Eurasia

Over the past two decades, there have been large differences among the Eurasian countries in their rates of population change and in the relative contribution of natural increase and net migration (Table 1). Over the past two decades, the global

Table 1 Population change in the Eurasian States, 1989–2012 (beginning of year; thousands)

	Total population		*Absolute change*			*Percent change*		
	1989	*2012*	*Total*	*Natural increase*	*Migration*	*Total*	*Natural increase*	*Migration*
Russia	147 022	143 056	−3966	−12 324	8358	−2.7	−8.4	5.7
Ukraine	51 707	45 453	−6254	−5598	−656	−12.1	−10.8	−1.3
Belarus	10 152	9465	−687	−615	−72	−6.8	−6.1	−0.7
Moldova	4338	3 560	−778	116	−894	−17.9	2.7	−20.6
Latvia	2667	2042	−625	−229	−396	−23.4	−8.6	−14.8
Lithuania	3675	3008	−667	−75	−592	−18.2	−2.0	−16.1
Estonia	1566	1294	−271	−74	−197	−17.3	−4.8	−12.6
Armenia	3449	3274	−174	511	−685	−5.1	14.8	−19.9
Azerbaijan	7021	9235	2214	2256	−42	31.5	32.1	−0.6
Georgia	5401	4498	−903	354	−1257	−16.7	6.6	−23.3
Kazakhstan	16 465	16 675	211	3330	−3119	1.3	20.2	−18.9
Kyrgyzstan	4254	5552	1298	1906	−608	30.5	44.8	−14.3
Tajikistan	5109	7807	2699	3651	−952	52.8	71.5	−18.6
Turkmenistan	3518	5402	1884	1692	192	53.5	48.1	5.4
Uzbekistan	19 882	29 555	9673	10 898	−1224	48.7	54.8	−6.2
Poland	37 885	38 538	654	1058	−404	1.7	2.8	−1.1
Czech Republic	10 360	10 505	145	−138	283	1.4	−1.3	2.7
Slovakia	5264	5404	140	199	−59	2.7	3.8	−1.1
Hungary	10 589	9958	−631	−793	162	−6.0	−7.5	1.5
Albania	3182	2800	−382	821	−1203	−12.0	25.8	−37.8
Bulgaria	8987	7327	−1659	−798	−862	−18.5	−8.9	−9.6
Romania	23 112	21 356	−1756	−505	−1251	−7.6	−2.2	−5.4
Slovenia	1996	2055	59	18	41	3.0	0.9	2.1
Croatia	4762	4398	−364	−122	−241	−7.6	−2.6	−5.1
FYR Macedonia	1881	2060	179	255	−76	9.5	13.6	−4.1
Bosnia-Herzegovina	4435	3843	−592	310	−903	−13.4	7.0	−20.4
Serbia	7516	7241	−275	41	−316	−3.7	0.5	−4.2
Montenegro	652	621	−31	50	−81	−4.8	7.7	−12.5
Serbia and Montenegro	10 445	–	–	–	–	–	–	–

Sources and notes: UNICEF Regional Office for CEECIS. 2013, TransMONEE 2013 DATABASE, released in June 2013 (www.unicef.org/ceecis). Data for Turkmenistan are for 1989–2007. Data for Serbia and Montenegro are for 2000–2012.

population increased by 30%, with less developed countries increasing by 36% and more developed countries increasing by just 8%. The Eurasia region as a whole decreased by 2%, thus bucking a global pattern of population increase.

The populations of the Central Asian countries of Turkmenistan, Tajikistan, Uzbekistan, Kyrgyzstan, and Azerbaijan increased the most among Eurasian countries, with rates near or above those for global population growth and the growth rate for less developed countries. Nine Eurasian countries – Albania, Ukraine, Bosnia-Herzegovina, Georgia, Estonia, Moldova, Lithuania, Bulgaria, and Latvia – had population declines of 10% or more over this period. Those countries in the Eurasia region which grew the fastest did so because of high rates of natural increase due to their younger age structures. Most also had high net emigration. Those Eurasian countries which had large population declines usually did so because of a combination of both natural decrease and net emigration. One exception is Albania which continued to have high natural increase but declined in population size because of large-scale emigration of more than one-third of the population. The significance of migration as a component of population change is evidenced by 13 countries having had emigration of more than 10% of their populations (Bulgaria, Montenegro, Estonia, Kyrgyzstan, Latvia, Lithuania, Tajikistan, and Kazakhstan) and another five when had emigration of more than 20% (Armenia, Bosnia-Herzegovina, Moldova, Georgia, and Albania).

There is a high correlation between level of economic development and being further along on the demographic transition and the Eurasian countries are at quite different stages of the transition. In 1989, Tajikistan, Turkmenistan, Uzbekistan, Kyrgyzstan, Azerbaijan, Albania, Armenia, Kazakhstan, and Macedonia could be classified as being in the second or third stage of the transition with high birth rates and low death rates leading to rapid population growth. Large portions of the region's population were in countries in the fourth stage where births only barely exceeded deaths or were roughly the same and population growth was slow or declining.

By 2011, with the steep transition-era population declines there had been some shifts among many Eurasian countries in their place along the demographic transition. In the same group of Central Asian countries births exceeded deaths and their populations continued to grow, albeit at much lower rates. In most countries in Eastern Europe, deaths and births are about equal so the populations in those countries have stabilized or are declining slowly (often because of out-migration). In Russia, Ukraine, Belarus, Hungary, Bulgaria, Latvia, and Serbia, deaths exceed births by considerable amounts leading to population decline. Demographers have recently begun to discuss a fifth stage of the demographic transition or a second demographic transition. This is characterized by very low fertility rates, high rates of non-marriage and non-childbearing, and high rates of non-marital childbearing. Many countries in Western Europe and a number of the Eurasian countries are in this situation.

Fertility

While the high mortality rates in many countries in the region have received the most attention of scholars examining demographic trends, it is actually the low fertility rates and steep declines during the 1990s which have had the largest numeric impact on population change. Prior to the economic transition, the social welfare system in the Eurasia region was decidedly pro-natalist with long and generous maternity benefits and significant state support for early child care and education. Contraceptive prevalence was low by international standards and abortion had become the primary means of fertility regulation in nearly all countries in the region. Marriage and childbearing were nearly universal. The pattern over large portions of the region was early marriage and childbearing, followed by subsequent abortions of subsequent pregnancies (David, 1999).

In 1989, much of the population resided in countries where the fertility rate was at about replacement level as many women had settled into a two-child pattern as the norm. Fertility rates in the Caucasus (Armenia, Azerbaijan, and Georgia) were slightly above this between two and three children per woman. Fertility rates in Central Asia and Albania were the highest between three and five children per woman. With the economic transition and steep economic decline in the 1990s and changes in the social support system, fertility rates declined in all countries in the region and in most by large amounts. The fertility rate in those Eurasian countries with rates of about replacement fell to rates of between 1.1 and 1.3 children per woman in the late 1990s, levels demographers call 'lowest-low' fertility (Kohler et al., 2001), and early 2000s before recovering somewhat in the late 2000s (**Table 2**). A number of Eurasian countries have fertility rates which are among the lowest in the world. Fertility in the high-fertility Eurasian countries of Central Asia and Albania have also declined by about 1.5 children on average. In 2009, only Azerbaijan and the five Central Asian countries had fertility rates above replacement level. Thus, in Eurasian countries encompassing about 85% of the region's total population, the fertility rate is well below replacement level and is not expected to increase back up to that level anytime in the foreseeable future. This trend, coupled with already older populations and a declining number of women in the childbearing ages, will inevitably lead to population decline across much of the region.

Many observers attribute the transition-induced economic declines across the Eurasian region as the reason for the steep declines in fertility. While that fact is certainly a contributing factor, there are other aspects of social change that have contributed to the decline as well (Brainerd, 2007). The changing role of women and circumstances of child rearing were also contributing factors. The proximate determinants framework can be used to analyze factors influencing changes in fertility across the region over the past two decades (Bongaarts, 1978).

The number of marriages across the region declined by 28%, from 3.6 million in 1989 to a low of 2.3 million in 2000, with the decline being universal across all countries (Heleniak, 2010). This had a large effect because most births still occur within marriage. Modern forms of contraception were not a priority in the region during the Communist period and since 1990, there has been a universal increase in contraceptive prevalence rates across all countries for which data are available. The Soviet Union

Table 2 Total fertility rate in the Eurasian States, 1989–2011

	1989	*2011*	*Lowest TFR*	*Year of lowest TFR*
Russia	2.01	1.58	1.17	1999
Ukraine	1.90	1.46	1.10	2000
Belarus	2.03	1.52	1.23	2003
Moldova	2.46	1.23	1.21	2002
Latvia	2.04	1.34	1.11	1998
Lithuania	1.98	1.55	1.23	2002
Estonia	2.22	1.52	1.28	1998
Armenia	2.61	1.50	1.19	1999
Azerbaijan	2.79	2.38	1.83	2001
Georgia	2.13	1.69	1.39	2005
Kazakhstan	2.84	3.00	1.80	1999
Kyrgyzstan	3.80	3.10	2.40	2000
Tajikistan	5.08	3.80	3.27	2005
Turkmenistan	4.30	2.60	2.60	2002
Uzbekistan	4.18	2.53	2.36	2003
Poland	2.05	1.30	1.22	2003
Czech Republic	1.87	1.43	1.13	1999
Slovakia	2.08	1.45	1.20	2001
Hungary	1.78	1.24	1.24	2011
Albania	2.96	1.71	1.60	2007
Bulgaria	1.90	1.51	1.09	1997
Romania	2.20	1.33	1.23	2001
Slovenia	1.52	1.56	1.20	2003
Croatia	1.63	1.41	1.33	2003
Macedonia	2.09	1.46	1.46	2005
Bosnia	1.70	1.21	1.18	2006
Serbia	–	1.41	1.38	2007
Montenegro	–	1.65	1.60	2005
Serbia and Montenegro	2.06		–	–

Sources and notes: UNICEF Regional Office for CEECIS, 2013. TransMONEE 2013 DATABASE, released in June 2013 (http://www.unicef.org/ceecis). Data for Turkmenistan listed under 2011 are for 2006, for Uzbekistan are for 2009, and for Romania are for 2010. These are the latest data available.

was the first country in the world to legalize abortion and as a result, an abortion culture developed where it was an acceptable means of fertility regulation with little moral stigma attached and little public opposition. The number of recorded abortions fell from 7.2 million in 1989 to 2.3 million in 2007, a two-thirds decline. The decline was universal across all countries in the region. In spite of this decline, abortion rates in some Eurasian countries remain among the highest in the world. These factors have caused the number of births in the region to decline from 6.9 million in 1989 to a low of 4.3 million in 2001 before increasing somewhat to 5.5 million in 2009.

There has been a shift in the patterns of childbearing in the region, with later marriage, later childbearing, and higher shares of birth outside of marriage. There has been a shift in the relative contribution of women at different ages to the overall fertility with women aged 25–29 now having more births than those aged 20–24 who were the cohort that gave birth to the most children during the Communist period. This is the result of less social support for childbearing and rearing and increased returns to education causing many women and couples to delay or defer children in favor of increased education and starting careers. Thus, the pattern of fertility across much of the region is beginning to resemble that of Western Europe.

Mortality

At the beginning of the transition period in the early 1990s, the populations living in a large number of Eurasian countries had very poor health outcomes that were inconsistent with their levels of income and development (Feshbach and Friendly, 1992). In Russia, aside from a brief period during the antialcohol campaign in the mid-1980s, the highest life expectancy achieved for Russian men was in 1964 – nearly a half century ago! Russian women only achieved their 1964 peak in 2007. Thus, Russia and a number of other countries in the Eurasian region have completely missed the large declines in mortality and morbidity that has taken place across both developed and developing countries (Lutz et al., 1994). Life style factors – poor diets, high levels of tobacco and alcohol consumption, lack of exercise, and stress – partly account for the gap in Eurasia–Western mortality and morbidity patterns (World Bank, 2005). This pattern of excess mortality applies mainly to the Slavic states of the former Soviet Union (Russia, Ukraine, and Belarus), the three Baltic states, and Kazakhstan, all which all have large Slavic populations. However, the pattern of high mortality affects a large portion of the Eurasian population.

Thus, much of the Eurasian population had poor health outcomes and high mortality even before the disruptive period of extreme social and economic change and initial period of independence of the 1990s when mortality increased and life expectancy declined precipitously across much of the region. The declines in life expectancy are unprecedented for countries not impacted by war or natural disasters. The increases in mortality and declines in life expectancy were universal across the region, albeit to vary levels and with different timing. Life expectancy for Russian men fell by 7.1 years from 1989 to 1994 and by 3.3 years for women in Russia over the same period (**Table 3**). Life expectancy for Russian males has since increased but is still not at the level it was at the beginning of the transition period (Heleniak, 2012). The same is true for the other Slavic states of Ukraine and Belarus. Males in the other Slavic countries of Ukraine and Belarus and the Baltic states had similar but smaller declines. In general, the Slavic states and those countries with large Slavic populations such the Baltics and Kazakhstan had the steepest declines in life expectancy over the period. The countries of Eastern Europe had earlier and smaller declines in life expectancy and faster recoveries, in part attributable to active labor market measures put in place to cushion the restructuring taking place in the labor market (Cornia and Paniccia, 2000). For most of the countries of Eastern Europe, the declines in life expectancy in the early transition period were a temporary blip and life expectancy has since continued an upward trend.

The unique mortality crisis across a large portion of the Eurasian region spawned a large literature examining its causes (Becker and Bloom, 1998; Bobadilla et al., 1997; Eberstadt,

Table 3 Life expectancy in the Eurasian States, 1989–2011

	Males				*Females*			
	1989	*2011*	*Lowest life expectancy*	*Year of lowest life*	*1989*	*2011*	*Lowest life expectancy*	*Year of lowest life*
Russia	64.4	64.0	64.0	1994	74.4	75.6	74.4	1994
Ukraine	66.4	66.0	66.0	1995	75.0	75.9	75.0	1995
Belarus	67.1	64.7	64.7	1999	76.3	76.7	76.3	1999
Moldova	65.5	66.8	65.5	1995	72.1	74.9	72.1	1995
Latvia	65.3	68.8	65.3	1994	75.2	78.7	75.2	1994
Lithuania	66.9	68.5	66.9	1994	76.3	79.1	76.3	1994
Estonia	65.7	71.2	65.7	1994	74.7	81.1	74.7	1994
Armenia	69.0	70.7	69.0	1993	74.7	77.5	74.7	1993
Azerbaijan	66.6	71.2	66.6	1993	74.2	76.5	74.2	1995
Georgia	67.3	70.2	67.3	1993	75.0	78.6	75.0	1993
Kazakhstan	63.9	64.0	63.9	1996	72.9	74.0	72.9	1996
Kyrgyzstan	64.2	65.7	64.2	1995	71.8	73.7	71.8	1995
Tajikistan	66.8	70.9	66.8	1993	71.7	74.1	71.7	1993
Turkmenistan	61.8	66.2	61.8	1994	68.2	72.9	68.2	1995
Uzbekistan	65.8	70.2	65.8	1989	71.6	74.9	71.6	1994
Poland	66.8	72.4	66.8	1991	75.5	80.9	75.5	1991
Czech Republic	68.1	74.7	68.1	1990	75.4	80.7	75.4	1989
Slovakia	66.8	72.2	66.8	1990	75.2	79.4	75.2	1991
Hungary	65.4	70.9	65.4	1993	73.8	78.2	73.8	1990
Albania	69.6	75.2	69.6	1993	75.5	78.3	75.5	1993
Bulgaria	68.6	70.4	68.6	1996	75.1	77.4	75.1	1997
Romania	66.5	70.0	66.5	1997	72.4	78.0	72.4	1989
Slovenia	68.8	76.6	68.8	1989	76.7	82.9	76.7	1989
Croatia	68.6	73.9	68.6	1991	76.0	80.0	76.0	1991
FYR Macedonia	70.1	72.7	70.1	1994	74.4	77.0	74.4	1994
Bosnia	–	72.4	72.4	–	–	77.7	77.7	–
Serbia	–	72.0	72.0	–	–	77.0	77.0	–
Montenegro	–	71.7	71.7	–	–	76.6	76.6	–
Serbia and Montenegro	68.7	–	–	–	73.8	–	–	–

Sources and notes: UNICEF Regional Office for CEECIS, 2013. TransMONEE 2013 DATABASE, released in June 2013 (http://www.unicef.org/ceecis). Data for Turkmenistan and Uzbekistan under 2011 are for 2006, for Macedonia are for 2010, and for Montenegro are for 2009. These are the latest data available. Data listed under 1989 for Croatia and Macedonia are for 1991. These are the earliest data available.

2010; United Nations in Russia, 2008; United Nations Development Programme, 2009). These studies have noted and come to several conclusions. The mortality crisis across much of the region was not due to a collapse of the health care system (there were too many doctors to begin with), worsening environmental conditions (which actually improved as a result of declines in industrial output), or worsening diets. Much of the increase was due to poor coping mechanisms on the part of many in the region as a result of large output declines, periods of high inflation, open unemployment becoming a reality for the first time, large-scale privatization in many states causing rapid labor turnover, the breakup of the three federal states in the region, two with some violence, the large movement of peoples around and within these states, the deterioration of the social safety net – education, health, the social protection system of guaranteed wages and pensions, and the re-writing of the social contract. During periods when mortality increased in the early 1990s, it was those in the middle ages who had the largest increases and the causes of deaths which increased the most were cardiovascular causes and external causes (homicide, suicide, accidents, and drownings) (World Bank, 2009). In Russia, Ukraine, and Belarus, the Baltic states, and to a lesser extent some countries in Eastern Europe such as Poland, Hungary, the Czech Republic, and Slovakia, adult mortality levels are rather inconsistent with infant and child mortality levels, which are relatively low, and inconsistent with their levels of development. In Central Asia and the most of the former Yugoslav states, both child and adult mortality are in line with their income levels.

Another trend that marks a large portion of the population of the region is the large difference between male and female mortality. Globally, women outlive men by 4 years. In more developed countries, the gap is about 7 years and in the less developed about 4 years. In much of the Eurasian region, women outlive men by 8 years and in Russia, Ukraine, and Belarus and three Baltic states, women outlive men by more than 10 years. These are currently the largest female–male life expectancy advantages in the world.

International Migration

International migration has become a major demographic, social, and economic influence on both the people and

countries in the region over the past two decades (Mansoor and Quillin, 2007). Under Communism, migration was tightly controlled by the state and emigration was prohibited except in rare circumstances. There was an attempt to equalize the standard of living across regions and between urban and rural areas, which kept the income disparities between regions rather narrow, thus dampening a major driver of migration.

In the post-Communist period, a number of factors have contributed to greatly increasing the amount and impact of migration. The first has been the liberalization of most of the countries in the region which has included allowing freedom of movement and the right to emigrate. The second has been the increased income disparities among countries in the region and between and countries outside the region. A third factor has been the breakup of three federal states increasing the number of countries in the region from 8 in the late 1980s to 29 currently. Only Poland, Hungary, Romania, Bulgaria, and Albania have remained in their same borders. The breakup of these states added some 45 million persons to the global stock of migrants as defined by the United Nations as somebody residing outside their country of birth.

Each of the Eurasian countries is a nation-state based on one ethnic group. At the time of the breakup of the Soviet Union, 43.4 million persons (17%) of the 15 ethnic groups upon which the successor states are based were residing outside their ethnic homelands (Heleniak, 2013). One of the largest ethnic diasporas to emerge from the breakup of the Soviet Union was the 25.2 million Russians who lived in the non-Russian states of the former Soviet Union. The diaspora population created by the breakup of Yugoslavia in the early 1990s counted approximately 3.4 million ethnic diaspora members, which partially factored into the ethnic conflicts in the region. Thus, a portion of the migration during the 1990s was diaspora migration, people moving to either their place of birth or ethnic homeland.

Migration in the region during the 1990s was marked by large refugee and internally displaced populations, rapid depopulation of regions and countries whose economies were no longer viable under new conditions, migration to outside the region for the first time, large flows of diaspora populations migrating to their ethnic homelands, and new forms of migration in the region such as labor migration. After 2000, the splintering of new states slowed, ethnic conflicts were either resolved or became stalemates, and economic recovery began in all countries in the region. As a result, refugee and IDP (internally displaced persons) flows declined, the return of diaspora populations slowed, labor migration became a fact of life for a number of countries in the region, and many of the countries in the region joined international organizations which measure and regulate international migration flows. A number of Eurasian countries have become members of the European Union and have had large numbers of their citizens migrate to the older EU countries for work.

Income gaps between the Eurasian countries and those of high-income countries in Europe and North American remain wide in spite of recent faster economic growth among the Eurasian countries. During the 2000s, the economy of Russia grew much faster than most of the other CIS countries which largely explains its increased attraction as a migration destination for countries in the region. Differences in levels of economic development within the CIS are significant and widening giving an impetus for migration as predicted by neoclassical theories of migration. Income per capita in Moldova, Uzbekistan, Tajikistan, and the Kyrgyz Republic is less than one-fifth that in Russia. In the first decade of the 2000s, several of these countries have the highest shares of remittances as a share of GDP in the world.

The significance of migration to the countries in the region is reflected in the size of the diaspora populations (based on place of birth) residing outside these countries and the size of the migrant populations within each. Some of these large migrant populations can be attributed to the unique historical circumstances of the emergence of 24 new countries but also the large migration flows which took place in the post-Communist period. Globally, 3.1% of the population reside outside their country of birth. Overall, about 10% of the population originating in a country of Europe and Central Asia reside outside that country (**Table 4**). Albania and Bosnia and Herzegovina have 40% or more of their populations outside their countries, while Armenia, Georgia, Kazakhstan, Macedonia, and Moldova have roughly one-quarter. A number of other countries of Europe and Central Asia have smaller but still significant shares of their populations residing outside their home countries. The Europe and Central Asian country with the smallest percent of its population residing outside the country is the Czech Republic with 4%.

The Eurasian countries are also major receivers of migrants or had large migrant populations when they became independent. In 2010, there were 29 million foreign-born migrants within the Eurasian countries, a decline of 3.3 million since 1990, due to increased restrictions on migration and the increase in various forms of temporary or labor migrant movements which do not include a permanent change in residence.

All of the countries of the former Soviet Union except for Russia had declines in their stocks of migrants. The largest percentage point declines were in Lithuania, Armenia, Latvia, Kyrgyzstan, and Estonia. For the three Baltic states and Kyrgyzstan, the reason was the large decline in the ethnic Russian and Russian-speaking populations. From Armenia, this was due to out-migration of native born ethnic Armenians. Russia's migrant stock increased from 11.5 million in 1990 (7.8% of the population) to 12.3 million in 2010 (8.7%).

Because of the unique historical circumstances of the disillusion of three large federal states, nearly all of the Eurasian countries have larger migrant stocks as a percent of their populations than the global average. There are several groupings of countries in the Eurasian region based on the sizes of migrant stocks. In Kazakhstan, Latvia, Estonia, Ukraine, Moldova, and Belarus, 10% or more of their populations are foreign born because of the large flows of Russians and other Slavic groups to these states during the Soviet period. A similar situation took place in the former Yugoslav states of Croatia, Slovenia, Montenegro, Macedonia, and Serbia. Though 'foreign-born,' many of these

Table 4 Native-born populations residing outside the Eurasian states and migrant stocks in the Eurasian countries, 2010

Country	*Native-born populations residing outside each Eurasian state*	*Percent of population residing outside country*	*Country*	*Foreign-born population*	*As percent of population*
Albania	1 438 451	45.4	Kazakhstan	3 079 491	19.5
Bosnia and Herzegovina	1 460 639	38.8	Croatia	699 947	15.9
Armenia	870 458	28.2	Latvia	335 022	15.0
Georgia	1 058 300	25.1	Estonia	182 464	13.6
Kazakhstan	3 719 766	23.6	Ukraine	5 257 527	11.6
Macedonia	447 138	21.9	Moldova	408 319	11.4
Moldova	770 528	21.5	Belarus	1 090 378	11.4
Belarus	1 765 877	18.4	Armenia	324 184	10.5
Croatia	753 529	17.1	Russia	12 270 388	8.7
Azerbaijan	1 433 513	16.0	Slovenia	163 894	8.1
Bulgaria	1 201 191	16.0	Montenegro	42 509	6.8
Ukraine	6 525 145	14.4	Macedonia	129 701	6.3
Lithuania	429 016	13.2	Serbia	525 388	5.3
Romania	2 769 053	13.1	Czech Republic	453 041	4.4
Estonia	169 213	12.6	Uzbekistan	1 175 935	4.2
Latvia	275 177	12.3	Tajikistan	284 291	4.0
Kyrgyz Republic	621 076	11.2	Kyrgyzstan	222 731	4.0
Tajikistan	791 618	11.2	Turkmenistan	207 700	4.0
Slovakia	519 716	9.6	Georgia	167 269	4.0
Poland	3 155 509	8.3	Lithuania	128 855	4.0
Russia	11 034 681	7.9	Hungary	368 076	3.7
Uzbekistan	1 954 460	7.0	Azerbaijan	263 940	3.0
Slovenia	131 895	6.5	Albania	89 106	2.8
Turkmenistan	260 953	5.0	Slovakia	130 682	2.4
Hungary	462 418	4.6	Poland	827 453	2.2
Czech Republic	369 737	3.6	Bulgaria	107 245	1.4
Serbia	196 013	2.0	Bosnia and Herzegovina	27 780	0.7
Montenegro	36	0.0	Romania	132 757	0.6
Eurasia region, total	44 585 106	10.2	Eurasia region, total	29 096 073	7.2
			World	213 943 812	3.1

Sources and notes: World Bank, 2013. Global Bilateral Migration Database (http://data.worldbank.org/data-catalog/global-bilateral-migration-database). Data for Serbia and Montenegro are presumably an underestimate because many of the migrants are still counted as being part of a combined Serbia and Montenegro.

people in both the former Soviet Union and former Yugoslavia are long-term residents who are well integrated into their states of current residence. As time goes on, with aging, these migrant stocks in these countries will continue to decrease.

Population Projections and Implications of Trends

The global population is projected to increase by nearly 30% by 2040, to 8.9 billion from 6.9 billion in 2010, with nearly the entire increase taking place in less developed countries. Over the next three decades only the five Central Asian countries of Tajikistan, Kazakhstan, Kyrgyzstan, Turkmenistan, and Uzbekistan, plus Azerbaijan are projected to have population growth (Table 5). All of the others are projected to have population declines. Belarus, Bosnia and Herzegovina, Latvia, Lithuania, Moldova, and Ukraine are projected to have population declines of more than 10% and Bulgaria and Georgia are projected to have declines of more than 20%.

For demographers, the trends in the region of extreme declines in fertility to nearly unprecedented levels, steep increases in mortality, and large-scale emigration from many countries during a period of significant social and economic upheaval provide interesting case studies for re-evaluating some demographic theories. Aside from periods of war or famine, countries of this size have never had such large population declines. Because of the demographic trends of the past two decades, the countries of Eurasia will confront a number of demographic challenges over the coming decades. These include shrinking populations and labor forces because of below replacement-level fertility and older age structures, high mortality in a large portion of the region, aging populations and high dependency ratios, shrinking youth populations and less new entrants to the labor force, and relatively immobile populations (Chawla et al., 2007). Sixteen Eurasia countries are among 20 in the world where the number of deaths exceed the number of births. A large portion of the population in the Eurasia region live in countries where male mortality is far below what is expected given their income and levels of development.

Table 5 Population in the Eurasian countries, 1980–2040 (millions)

	1980	*2010*	*2040*	*Change total*	*2010–40 (%)*
Albania	2.7	3.2	3.2	0.0	−0.8
Armenia	3.1	3.1	3.0	−0.1	−1.8
Azerbaijan	6.2	9.2	11.3	2.1	22.7
Belarus	9.7	9.6	8.4	−1.2	−12.1
Bosnia and Herzegovina	3.9	3.8	3.2	−0.5	−13.9
Bulgaria	8.9	7.5	5.9	−1.6	−20.8
Croatia	4.4	4.4	4.0	−0.4	−8.6
Czech Republic	10.3	10.5	10.7	0.2	1.7
Estonia	1.5	1.3	1.3	−0.1	−6.2
Georgia	5.1	4.4	3.5	−0.9	−20.6
Hungary	10.7	10.0	9.4	−0.6	−5.7
Kazakhstan	14.9	16.0	20.0	4.0	25.1
Kyrgyzstan	3.6	5.3	7.3	1.9	36.2
Latvia	2.5	2.3	2.0	−0.3	−12.2
Lithuania	3.4	3.3	2.9	−0.4	−11.8
Montenegro	0.6	0.6	0.6	0.0	−1.6
Poland	35.6	38.3	36.4	−1.9	−5.0
Moldova	4.0	3.6	2.9	−0.7	−18.7
Romania	22.2	21.5	19.5	−2.0	−9.4
Russian Federation	138.7	143.0	131.3	−11.7	−8.2
Serbia	8.9	9.9	9.2	−0.7	−6.9
Slovakia	5.0	5.5	5.4	−0.1	−1.0
Slovenia	1.8	2.0	2.0	0.0	0.0
Tajikistan	4.0	6.9	9.9	3.0	44.3
Macedonia	1.8	2.1	2.0	−0.1	−4.1
Turkmenistan	2.9	5.0	6.5	1.4	28.4
Ukraine	50.0	45.4	38.1	−7.3	−16.2
Uzbekistan	16.0	27.4	34.8	7.4	26.9
Eurasia, total	382.1	405.0	394.7	−10.3	−2.5
World	4453	6896	8874	1978	28.7
More developed regions	1081	1236	1307	71	5.7
Less developed regions	3372	5660	7567	1907	33.7

Source: United Nations, 2011. Population Division of the Department of Economic and Social Affairs of the, World Population Prospects: The 2010 Revision (http://esa.un.urg/unpd/wpp/index.htm).

Several Eurasia countries have among the oldest median-age populations in the world.

See also: Adult Mortality in Eastern Europe and the Former Soviet Republics; Fertility Trends in the Formerly Socialist Countries of Europe; International Migration and Remittances; Migrants and Migration in Europe; Second Demographic Transition.

Bibliography

Becker, C.M., Bloom, D. (Eds.), 1998. World Development, Special Issue, vol. 26 (11). The Demographic Crisis in the Former, Soviet Union, pp. 1913–2103.

Bobadilla, J.L., Costello, C.A., Mitchell, F. (Eds.), 1997. Premature Death in the New Independent States. National Academy Press, Washington, DC.

Bongaarts, J., 1978. A framework for analyzing the proximate determinants of fertility. Population and Development Review 4 (1), 105–132.

Brainerd, E., 2007. The Baby Decision Amid Turmoil: Understanding the Fertility Decline in Russia of the 1990s. National Council for Eurasian and East European Research Working Paper. http://www.nceeer.org/ (accessed 24.07.13.).

Chawla, M., Betcherman, G., Banerji, A., 2007. From Red to Gray: The "Third Transition" of Aging Populations in Eastern Europe and the Former Soviet Union. World Bank, Washington, DC.

Cornia, A.G., Paniccia, R. (Eds.), 2000. The Mortality Crisis in Transitional Economies. Oxford University Press, New York.

David, H.P., 1999. From Abortion to Contraception: A Resource to Public Policies and Reproductive Behavior in Central and Eastern Europe from 1917 to the Present. Greenwood Press, Westport, CT and London.

Eberstadt, N., 2010. Russia's Peacetime Demographic Crisis: Dimensions, Causes, Implications. The National Bureau of Asian Research, Seattle, Washington.

Feshbach, M., Friendly Jr., A., 1992. Ecocide in the USSR: Health and Nature Under Siege. Basic Books, New York.

Heleniak, T., 2010. Causes and demographic consequences of fertility decline in the former Soviet Union and Central and Eastern Europe. Marriage and Family Review 46 (1), 79–106.

Heleniak, T., 2012. Population trends in Russia. In: Wegren, S. (Ed.), Return to Putin's Russia: Past Imperfect, Future Uncertain, fifth ed. Rowman and Littlefield Publishers Inc., London, Boulder, New York, Toronto, Plymouth, UK.

Heleniak, T., 2013. Diasporas and Development in Post-Communist Eurasia, Migration Information Source. http://www.migrationinformation.org/ (accessed 24.07.13.).

Kohler, H., Billari, F.C., Ortega, J.A., 2001. Towards a Theory of Lowest-Low Fertility, Max Plank Institute for Demographic Research, MPIDR Working Paper 2001–2032.

Lutz, W., Scherbov, S., Volkov, A. (Eds.), 1994. Demographic Trends and Patterns in the Soviet Union before 1991. International Institute for Applied Systems Analysis, Laxenburg, Austria.

Mansoor, A., Quillin, B., 2007. Migration and Remittances: Eastern Europe and the Former Soviet Union. World Bank, Washington, DC.

UNICEF, 2013. Regional Office for CEECIS. TransMONEE 2013 DATABASE, released in June 2013. www.unicef.org/ceecis.

United Nations in Russia, 2008. Demographic Policy in Russia: From Reflection to Action. United Nations, Moscow.

United Nations Development Programme, 2009. Russia Facing Demographic Challenges: National Human Development Report, Russian Federation 2008. UNDP, Moscow.

United Nations, 2011. Population Division of the Department of Economic and Social Affairs of the, World Population Prospects. The 2010 Revision. http://esa.un.org/unpd/wpp/index.htm.

World Bank, 1996. World Development Report 1996: From Plan to Market. World Bank, Washington, DC.

World Bank, 2005. Dying Too Young – Addressing Premature Mortality and Ill Health Due to Non-Communicable Diseases and Injuries in the Russian Federation. World Bank, Washington, DC.

World Bank, 2009. Confronting "Death on Wheels": Making Roads Safe in Europe and Central Asia. World Bank, Washington, DC.

World Bank, 2013. Global Bilateral Migration Database. http://data.worldbank.org/data-catalog/global-bilateral-migration-database.

Population Composition by Race and Ethnicity: North America

Jenifer Bratter and Ellen Whitehead, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Morning, volume 17, pp. 11744–11749, © 2001, Elsevier Ltd.

Abstract

This article examines the history and current state of race/ethnic composition within North America, encompassing Canada, the United States, and Mexico. It summarizes general trends in demographic change as well as the complexity of demographic composition by race and major ethnic group. We begin with a brief overview of the historical moments shaping ethnic and racial composition and then provide the most recent estimates of ethnic and racial composition for each region. At the close we provide a comparative analysis, focusing on what the composition bodes for the future.

The racial and ethnic composition of the continent of North America, encompassing Mexico, the United States, and Canada, has undergone a variety of transitions, reflecting in many cases both demographic processes as well as increased complexity of race/ethnic categories. While race and ethnicity are defined differently, their measurement is a common feature of many population registers across the globe (Morning, 2008). The compositions of all nations have become more differentiated, particularly in the United States and increasingly Canada, as 'non-European' subgroups occupy an expanding share of the nation as a whole, translating into a more diverse landscape. On one hand, this reflects the expanding immigrant presence as well as the population momentum driven largely by the children of immigrants and as the growth of interracial unions and multiracial identities (Passel, 2011; Humes et al., 2011; Lee and Boyd, 2007). On the other hand, shifts in composition have also occurred alongside ever-evolving discussion of how best to capture racial and ethnic identity through census classification (Perez and Hirschman, 2009; Goldscheider, 2002), and this increased complexity raises new questions about how race/ethnicity will be captured in the future.

By contrast, Mexican race/ethnic composition has been remarkably stable as the key distinction is linguistic, between Spanish and indigenous speaking peoples, with few official mentions of the small African origin population (de Vidas and Hoffman, 2012). This is not to say that race does not factor into variation within Mexico, as other physical distinctions are a salient part of social variation among Mexican people (i.e., skin color, indigenous appearance). However, capturing race and ethnic variation is less central to the demographic record.

Historical Perspective of North America's Ethnic Composition

We wish to refer readers to the previous article (Tienda and Morning, 2001) for an excellent overview of the historical origins of the peopling of North America. We begin with a brief overview of historical shifts in race/ethnic composition and then explore historical construction of categories that provides the basis of these distinctions. The major categories of race in most records mark the following broad ancestral origins: North American indigenous ancestry, or origins of those immigrating from European, Africa, Asian, or Latin America. The outlying composition of racial and ethnic groups has fluctuated tremendously, fueled by waves of immigration (both forced and voluntary) and differential natural increase. Consistent across all nations within North America is the depopulation of the indigenous populations. The size of the precontact American Indian population has been estimated to range between 1 and 18 million in the United States and Canada (Thornton, 1989) and around 4.5 to 25 million in Mexico (Alba, 1977). Due to deaths caused by infectious illnesses carried by the European settlers, warfare between the Europeans, and genocide, this number is estimated to have dwindled to less than 250 000 as of the 1800s. The total depopulation in America between 1500 and 1890 is estimated to have been around 1.75 to 4.75 million people (Feagin, 2000).

The diminishing indigenous presence corresponded with waves of immigration from Europe, Asia, and Africa. Europeans maintained a large majority in the United States and Canada (between 80 and 90% of the population) (McDaniel, 1995). Some of the earliest nonindigenous presence within the North American territories arrived from Spain and occupied what is now the United States and Mexico, predating that of the English colonists. This presence existed over two centuries, longer than that length of the independent nation of the United States (Rumbaut, 2005: 24–28). In contrast, the early presence of foreigners in Canada primarily consisted of French and English settlers up until the early twentieth century and became more diverse in terms of European origins. Over time, the United States also experienced an influx of English, German, French, Dutch, Irish, and Scotch-Irish. The African presence within North America evolved differently across the various nations with Africans entering as slaves within the United States (known then as the colonies) and Mexico (Feagin, 2000), but entering Canada largely as a free population escaping slavery in the United States and the Caribbean. Between 1600s and 1820s, an estimated 8 million Africans were transported via the Middle Passage, and between 15 and 30% of those kidnapped died aboard ships (Desmond and Emirbayer, 2009). Although the presence of slaves was especially prevalent in the United States, many were transported to Mexico and eventually settled in indigenous villages (Carroll, 2009).

The size and share of the Asian population reflected the impact of restrictive immigration laws and anti-immigrant sentiment that reduced the growth of this group during late 1800s on into the latter part of the twentieth century. The first Asian immigrants to the United States were the Chinese, who arrived between 1840 and 1860 during the California gold rush

and later worked on railways, with the Japanese being the most dominant immigrant group between 1910 and 1950 (Xie and Goyette, 2004). Although small, Chinese immigration to Mexico occurred at the beginning of the twentieth century (Augustine-Adams, 2009). Meanwhile, the major waves of Asian immigrants in Canada began after 1960 when restrictions were largely lifted, inspiring a similarly dramatic increase in the number of non-European immigrants of various backgrounds.

Notably, this demographic growth occurred alongside changes in demographic measurement in ethnic and racial groups. Intermixture between groups, as well as variable notions of how and on what basis to define racial groups, created ambiguity in how these categories could be applied. The category 'White,' for example, did not initially include members of Italian, Spanish, or Irish descent as it did for persons of French or English descent (Perez and Hirschman, 2009). While membership within the African-American or Black category has been defined under the 'one-drop rule' (i.e., returning anyone of any African admixture to this category), categories of partial African descent (e.g., mulatto, quadroon, octoroon) were classified during the 1800s and early 1900s. The number of Asian ethnic groups on the census expanded from one census to another, reflecting increasing concerns over immigration from this region. Persons of Mexican origin, who may be considered Latino today, were classified as 'White' as a condition of the annexation of Mexico into Texas. Only in 1930 would there be a 'Mexican' on the census, which was removed in the following year. Changes and measurement of race/ethnic composition have historically evolved.

Contemporary Trends of Ethnic and Racial Population Composition in North America

As we move through the second decade of the new millennium, we can see that patterns of ethnic and racial change have taken a variety of forms, mostly that maintain the trajectory established in the beginning of the decade. On one hand, we observe a diminishing share of populations that occupy the race/ethnic majority while 'non-White' immigrant groups expand. On the other hand, classification schemes have also become even more complex, with options to express multiple origins and more detail and specificity provided to particular ethnic and racial categories. Taken together, these patterns point to two salient issues in social demographic study of race/ethnic composition: (1) the ever-expanding diversity and (2) the complexity of race/ethnic measurement.

Canada

Canada's multicultural landscape, like the United States, has been characterized by immigration that has spurned growth among non-European ethnics, altering the balance between these groups and the European majority. Following the removal of ethnic restrictions that limited immigration from non-European countries, immigration from these countries increased, particularly from Asian. Characterizing this changing composition in the demographic record has tended to emphasize ethnic distinctions as opposed to 'race.' Table 1 shows the distribution of the population across ethnic origin designations, according to the quinquennial Census, from 1996 to 2011. Although the government conducted an 'Ethnic Diversity Survey' in 2002, conclusions are difficult to draw from a one-time study, and data from the past 10 years would be missing, as well. What is immediately apparent is the shift in usage and level of detail of particular categories of ethnic groups between 1996 and 2001 (with empty cells noted with N/A). This is an oft noted feature of Canadian census composition (see Goldscheider, 2002 for discussion), limiting the appraisal of changing composition over time.

Some shifts occur on what aspects of ethnicity are recorded, with the use and then removal of parental ethnicity (specifically from male ancestors) or the inclusion of racial designations via the 'visible minority' classifications (see Table 2). Others reflect changes in usage, naming, and degree of detail. For example, as of 1996, residents can declare their origins as 'Canadian' or the

Table 1 Canada: population by ethnic origin, 1996–2011

	1996		*2011*	
Ethnic origins	*N*	*%*	*N*	*%*
Single origins	18 303 625	64.16	19 036 295	57.95
North American Aboriginal	477 635[a]	1.67	628 460	1.91
Canadian	5 326 995	18.67	6 030 470	18.36
European	9 694 260[b]	33.98	7 229 120	22.00
British Isles	3 267 525	11.45	2 521 360	7.67
French	2 683 840	9.41	1 170 620	3.56
Western	N/A	N/A	991 555	3.02
Northern	N/A	N/A	143 860	0.44
Eastern	N/A	N/A	901 865	2.75
Southern	N/A	N/A	1 357 735	4.13
Caribbean	N/A	N/A	360 380	1.10
Latin, Central, and South American	118 635	0.42	285 070	0.87
African	137 315	0.48	465 235	1.42
Central and West	N/A	N/A	105 785	0.32
North	N/A	N/A	175 520	0.53
Southern and East	N/A	N/A	120 725	0.37
Asian	1 968 465[c]	6.90	4 022 085	12.24
West Central Asian and Middle Eastern origins	N/A	N/A	534 185	1.63
South Asian	590 150	2.07	1 306 380	3.98
East and Southeast Asian	1 271 450	4.46	2 169 270	6.60
West Asian	106 865	0.37	N/A	N/A
Arab	188 430	0.66	N/A	N/A
Oceania	N/A	N/A	15 475	0.05
Pacific Islands	5765	0.02	9570	0.03
Multiple origins	10 224 500	35.84	13 816 025	42.05
Total	28 528 125	100.00	32 852 320	100.00

[a]The 1996 number for 'North American Aboriginal' represent the category for 'Aboriginal' origins.
[b]The 1996 numbers for 'European' origin represents the sum of British Isles, French, and European origins, replicating the way the category is constructed in the 2011 census.
[c]The 1996 numbers for 'Asian' represents the sum of South Asian, East and Southeast Asian, and West Asian origins, replicating the presence of this broader category in the 2011 census.
Source: Statistics Canada. 1996 Census: Total Population by Ethnic Origin. www.statcan.ca; 2011 National Household Survey Profile: Ethnic Origin Population. www12.statcan.gc.ca.

Table 2 Canada: visible minority population

Visible minority status	1996 N	1996 %	2011 N	2011 %
Total visible minority population	3 197 480	11.21	6 264 750	19.07
South Asian	670 590	2.35	1 567 400	4.77
Chinese	860 150	3.02	1 324 750	4.03
Black	573 860	2.01	945 665	2.88
Filipino	234 195	0.82	619 310	1.89
Latin American	176 970	0.62	381 280	1.16
Arab	N/A	N/A	380 620	1.16
Arab/West Asian	244 665	0.86	N/A	N/A
Southeast Asian	172 765	0.61	312 075	0.95
West Asian	N/A	N/A	206 840	0.63
Korean	64 835	0.23	161 130	0.49
Japanese	68 135	0.24	87 270	0.27
Visible minority, not included elsewhere	69 745	0.24	106 475	0.32
Multiple visible minorities	61 575	0.22	171 935	0.52
Not a visible minority	25 330 645	88.79	26 587 575	80.93
Total	28 528 125	100.00	32 852 320	100.00

Sources: Statistics Canada. 1996 Census: Visible Minority Population. http://www12.statcan.gc.ca; 2011 National Household Survey Profile: Visible Minority Population. http://www12.statcan.gc.ca.

formerly distinct category of *Arab* has been absorbed into a broader category of *West Central Asian and Middle Eastern Origins* in the most recent census. Some groups shift their place in broader categorization, such as *British Isles* and *French* origins being included among the broader category of 'European' in the most recent census. Even ethnic origin groups that have always been present on the census rolls, specifically African and Aboriginal, have changed their naming over time or have been tracked with greater detail within the category. This ever-changing nature of the Canadian census reveals that political nature of census classification, impacted by needs of policy makers and the public.

Over the past 15 years, the population of Canada has become more ethnically and racially diverse, marked by an increasing shift in demographic balance favoring non-European origin groups and visible minorities. The population has grown overall, from 28 million in 1996 to 32 million in 2011, an increase of about 15%. The majority of Canada's population is single origin, though increasingly Canadians are reporting multiple ethnic origins; the single-origin share has diminished from roughly two-thirds (64%) to a little more than half (57%) in 2011. This change is driven by the decreasing share of the population that identifies as European, from 34% in 1996 to 22% in 2011, and Canadian, from 23% in 2001 to 18% in 2011. Absolute numbers of these groups have also dropped with the British Isles population, declining from 3.3 to 2.5 million, and the drop in the French origin population from 2.7 to 1.2 million. Meanwhile, all other non-European origin groups have expanded their shares, with the percent of Latin, Central, and South American; African, and South Asian origin more than doubling in this time span. The Asian population, the fastest growing and largest non-European ethnic group, has doubled in numbers from around 2 million in 1996 to over 4 million in 2011 and its share from 6% in 1996 to 12% in 2011. Overall, the expanding groups are those hailing from most recent waves of immigration that is non-European, drawing greater attention to distinguishing racial and as well an ethnic variation.

We now turn to variation in visible minority status, shown in Table 2. These figures show the impact of continual immigration from non-European nations, as well as the population momentum created by their presence. Visible minorities refer to "persons, other than Aboriginals, who are non-Caucasian in race or non-white in skin color" (Hamplova and Le Bourdais, 2010: 1538). A minority but growing share and size of the population identifies with one or more of the visible minority categories, growing from 11%, or 3.1 million persons, to 19%, or 6.2 million persons, during this period. Each visible minority population, such as the numerous Asian categories, Blacks, Filipinos, and Latin Americans, have made small but steady increases during this time period, both proportionately and in absolute numbers. Unlike the United States, the largest racial minority groups are ethnically Asian. South Asians and Chinese are the largest groups, encompassing roughly 6–8% of the population (derived from summing these two groups). Blacks are the next largest group, growing from 2% to nearly 3% in 2011, followed by Filipinos, Latin Americans, and Arabs, whose share is a little more than 1% of the population as of 2011. Among the remaining groups, South Asians, Koreans, and those declaring multiple categories have more than doubled their numbers.

United States

Similar to Canada, the United States is increasingly diverse. However, unlike Canada, this process has been faster, with population projections predicting ethnic/racial plurality with no clear majority in the next 30 years. The change is already apparent in births, as one-fourth of nation's children as of 2010 can be considered 'immigrant youth;' that is, they are either foreign-born or are children of first-generation parents (Passel, 2011). Children are the most diverse age group, as the United States witnessed its first birth cohort with no racial majority as non-Hispanic Whites encompassed (slightly) less than half of babies born. This increasingly diverse composition has a spatial component as well, with four states the District of Columbia, and one-tenth of all counties have 'majority–minority' populations (Humes et al., 2011).

In Table 3, we show the composition of the US population by race and Hispanic ethnicity from 1990 to 2010. The categories set forth by the Office of Management and Budget, the federal agency charged with setting the standards for race/ethnic classification, place Hispanics, regardless of race, in a separate category and provide race categories for non-Hispanics. We provide more detail here, showing the race of Hispanics and non-Hispanics. In addition, as respondents are now (since 2000) allowed to declare multiple race responses, the major categories specify those persons who marked this category 'alone.' We also place all persons selecting more than one race in a separate category as well, among Hispanics and non-Hispanics. The US population continues to grow, adding 60 million people between 1990 and 2010. The majority of population 'non-Latino,' though Latinos are among the fastest growing group, increasing their share from 9 to 16% from 1990 to 2010. Latinos are now the second largest ethnic/racial group,

Table 3 Race/ethnic composition of the United States by Hispanic ethnicity, 1990, 2000, and 2010 estimates

	1990		*2000*		*2010*	
	Estimate	*Percentage*	*Estimate*	*Percentage*	*Estimate*	*Percentage*
Non-Hispanic/Latino(all races)	226 355 814	**91.0**	246 116 088	**87.5**	258 267 944	**83.7**
White alone	188 128 296	75.6	194 552 774	69.1	196 817 552	63.8
Black or African American alone	29 216 293	11.8	33 947 837	12.1	37 685 848	12.2
American Indian and Alaska Native alone	1 793 773[a]	0.7	2 068 883	0.7	2 247 098	0.7
Asian alone	6 968 359[b]	2.8	10 123 169	3.6	14 465 124	4.7
Native Hawaiian and other Pacific Islander alone	n/a	n/a	353 509	0.1	481 576	0.2
Some other race alone	249 093	0.1	467 770	0.2	604 265	0.2
Two or more races	n/a	n/a	4 602 146	1.6	5 966 481	1.9
Hispanic/Latino(all races)	22 354 059	**9.0**	35 305 818	**12.5**	50 477 594	**16.3**
White alone	11 557 774	4.7	16 907 852	6.0	26 735 713	8.7
Black or African American alone	769 767	0.3	710 353	0.3	1 243 471	0.4
American Indian and Alaska Native alone	165 461[a]	0.1	407 073	0.1	685 150	0.2
Asian alone	305 303[b]	0.1	119 829	0.0	209 128	0.1
Native Hawaiian and other Pacific Islander alone		n/a	45 326	0.0	58 437	0.0
Some other race alone	9 555 754	3.8	14 891 303	5.3	18 503 103	6.0
Two or more races		n/a	2 224 082	0.8	3 042 592	1.0
Total population	248 709 873		281 421 906		308 745 538	

Figures in bold represent larger categories that sum to 100 percent.
n/a: Category not available at the corresponding Census year.
[a]The category 'Asian alone' was the category 'Asian or Pacific Islander' in the 1990 Census.
[b]The category 'American Indian and Alaska Native alone' was the category of 'American Indian, Eskimo, or Aleut' in the 1990 Census.

replacing African-Americans, who had been the largest 'non-White' group since the first Census that was taken in 1790. This group represents a stable 12% of the population. Asians are also quickly expanding their numbers, from less than 3% to nearly 5% in 2010. In addition to immigration and population momentum, the composition of the United States has also been affected by increasing interracial unionship, evidenced by the growing number of persons claiming two or more races are also growing, from 1.6 to 1.8%.

We now turn to more detailed examination of ethnicity in **Table 4**, providing ethnic information for Hispanic and Asian populations. The majority of the Hispanic population (see panel A), nearly three-fourth, is ethnically Mexican, followed by Puerto Ricans and Cubans, together who make up less than 20%. A sizable share of this group also hails from Central and South America. This ethnic mix include a wide range of perspectives on issues of racial and ethnic identity and as such, some argue that ethnicity for Hispanics provides more specific information on origins than race, given that nearly half identify racially as 'Some other race' (Perez and Hirschman, 2009). Turning to Asians (in panel B), this racial category masks a wide range of ethnic diversity, as no one group represents more than 21%. Chinese are the largest ethnic group, followed by Asian Indians, Filipinos, and Vietnamese. This also reflects the impact of recent immigration, as Japanese and Koreans were more prevalent shares of previous flows of Asian immigrants (Xie and Goyette, 2004). The variation in immigration histories of these 'new' groups provides important complexity to the pathways of Asian incorporation.

Mexico

Unlike Canada or the United States, the key ethnic or racial distinctions in the official record of Mexico is linguistic, specifically whether a person speaks Spanish or an indigenous language. As a nation which is known for emigrants and low immigration, this division in language emphasizes the groups that already reside in Mexico, unlike the United States and Canada, for whom tracking immigrants has been salient in their record keeping. By contrast, the Mexican census gives little focus to newcomers to the country that may vary linguistically. This system of classification based on spoken languages allows an assumed distinction between those groups which have existed in Mexico for centuries or more: the indigenous Mexican population and those of whole or partial European origin. Exclusion of a *racial category* from the Mexican census is based upon the supposition that almost every national of this country is 'mestizo' (Spanish for mixed race), as well as the government's perception that Mexican society had become increasingly panracial and integrated (Augustine-Adams, 2009). As shown in **Table 5**, the number of Spanish speakers and indigenous language speakers have both increased over the past decade and a half, but the population who speaks only Spanish has increased at a slightly faster rate (a linear increase of 25.4 compared to 22.1). Those who solely speak Spanish still constitute the vast majority of the population, comprising 93% of Mexicans compared to 7% who are indigenous speakers.

Looking at their proportion to Spanish speakers in the population, most indigenous language speakers were concentrated in the southern states of Mexico in 1995, and this trend continues up through 2010. Over one million indigenous speakers resided in both Oaxaca and Chiapas in 2010, and Veracruz, Querétaro, and Yucatán all had over half a million indigenous speakers this same year. All but one state (Yucatán) increased in absolute numbers of their population of indigenous speakers, yet half decreased in terms of the proportion of the total state population these speakers constituted. These numbers indicate a gradual trend toward a nation with a smaller indigenous population – or at least, a nation with a small number of

Table 4 Ethnic composition of Asians and Latino/Hispanics, 2010

Ethnic group	*Population*	*Percent*
Panel A. Hispanic/Latino		
Mexican	31 798 258	62.99
Puerto Rican	4 623 716	9.16
Cuban	1 785 547	3.54
Dominican	1 414 703	2.80
Central American	3 998 280	7.92
South American	2 769 434	5.49
Other Hispanic or Latino	4 087 656	8.10
Total Hispanic or Latino:	50 477 594	100.00
Ethnic Category	*Population*	*Percent*
Panel B. Asian Alone (selected one ethnic category)		
Asian Indian	2 843 391	19.86
Bangladeshi	128 792	0.90
Bhutanese	15 290	0.11
Burmese	91 085	0.64
Cambodian	231 616	1.62
Chinese (except Taiwanese)	3 137 061	21.92
Filipino	2 555 923	17.86
Hmong	247 595	1.73
Indonesian	63 383	0.44
Japanese	763 325	5.33
Korean	1 423 784	9.95
Laotian	191 200	1.34
Malaysian	16 138	0.11
Nepalese	51 907	0.36
Pakistani	363 699	2.54
Sri Lankan	38 596	0.27
Taiwanese	196 691	1.37
Thai	166 620	1.16
Vietnamese	1 548 449	10.82
Other Asian	239 558	1.67
Total	14 314 103	100.00

Source: 2010 Decennial Census (Summary File 1).

Table 5 Mexico: indigenous language speakers, 1995–2010 population of age 5 years and above

	Language spoken			
Year	*Spanish*	*Indigenous*	*Total*	*Percent indigenous*
1995	74 735 782	5 483 555	80 219 337	6.84
2000	78 749 907	6 044 547	84 794 454	7.13
2005	84 255 223	6 011 202	90 266 425	6.66
2010	93 715 582	6 695 228	100 410 810	6.67

Source: Instituto Nacional de Estadística, Geografía e Informática (INEGI). Banco de Información; INEGI 1995–2010. http://www3.inegi.org.mx/sistemas/biinegi/?e=31&m=0&ind=1005000040.

people who speak these languages. These numbers do not reflect the small but significant African origin or Black population in Mexico (Augustine-Adams, 2009).

Comparative North American Perspective: Looking toward the Future

The future of race/ethnic composition in the North American region is likely to be one of increased complexity. According to US Census projections, the United States will become the first nation in the region to have no ethnic/racial majority. Meanwhile, Mexico, as in other Latin American countries, is beginning to undergo a shift in classification to allow for more detail to recognize Indigenous and Black communities (de Vidas and Hoffman, 2012). In Canada, populations of multiple ethnic origins expand tremendously, but visible minority remains small. This raises the question whether classifications approaches will remain stable or, might there be another change on the horizon as the population becomes more ethnically fluid.

In addition, interracial family formation has become more common yielding both more mixed-race families and more multiracial offspring. Both nations have official mechanisms in place to capture multiple origins to track the growth of multiple-ancestry population and to enable incorporating such groups into official statistics. Importantly, expansion among the 'multiple origin' or the two or more race populations may undo notions of rigid race/ethnic categorization around single-origin groups. However, single-origin groupings have tremendous political significance, and therefore it is also highly possible that classification will accommodate for increased blurring without fundamentally shifting how populations are organized.

See also: Age Structure; Assimilation of Immigrants; Assortative Mating in the Marriage Market; Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Ethnic Intermarriage; Ethnic, Racial, and Nationalist Movements; Global Migration; Immigration Policy; Immigration, Applied Research on; Immigration, Citizenship, and Integration: Social Work Connections; Immigration, Emigration, and Citizenship; Immigration: Health and Mortality; Immigration: Political Aspects; Migration and Development; Migration and Fertility; Migration, Economics of; Migration, Ethnicity, Aging and Social Work Practice; Minorities; Population Dynamics: Momentum of Population Growth; Population Forecasts; Racial Relations; Residential Segregation in the United States; Slavery as Social Institution; Social Classification.

Bibliography

Alba, Francisco, 1977. La Población de México: Evolución y Dilemas. El Colegio de México, México, D.F.

Augustine-Adams, Kif, 2009. Making Mexico: legal nationality, Chinese race, and the 1930 population census. Law and History Review 27 (1), 113–144.

Carroll, Pat, 2009. Black Aliens and black Natives in new Spain's indigenous communities. In: Vinson III, Ben, Restall, Matthew (Eds.), Black Mexico. University of New Mexico Press, Albuquerque, New Mexico, pp. 72–95.

Desmond, Matthew, Emirbayer, Mustafa, 2009. Racial Domination, Racial Progress: The Sociology of Race in America. McGraw-Hill, New York.

Feagin, Joe, 2000. Racist America: Roots, Current Realities, and Future Reparations. Routledge, New York.

Goldscheider, Calvin, 2002. Ethnic categorizations in censuses: comparative observations from Israel, Canada, and the United States. In: Kertzer, David I., Arel, Dominique (Eds.), Census and Identity: The Politics of Race, Ethnicity, and Language in National Censuses. Cambridge University Press, New York, pp. 71–91.

Hamplova, Dana, Le Bourdais, Celine, 2010. Visible minorities and 'White'–'non-White' conjugal unions in Canadian large cities. Ethnic and Racial Studies 33 (9), 1537–1560.

Humes, Karne R., Jones, Nicholas A., Ramierz, Roberto, 2011. "Overview of Race and Hispanic Origin" 2010 Census Briefs. U.S. Department of Commerce, Economics and Statistics Administration. http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf.

Instituto Nacional de Estadística, Geografía e Informática (INEGI). Banco de Información INEGI 1995–2010. http://www3.inegi.org.mx/sistemas/biinegi/?e=31&m=0&ind=1005000040.

Lee, Sharon, Boyd, Monica, 2007. Marrying out: comparing the marital and social integration of Asians in the US and Canada. Social Science Research 37, 311–329.

McDaniel, Antonio, 1995. The dynamic racial composition of the United States. Daedalus 124 (1), 179–198.

Morning, Ann, 2008. Ethnic classification in global perspective: a cross-national survey of the 2000 census round. Population Research and Policy Review 27 (2), 239–272.

Passel, Jeffrey, 2011. Demography of immigrant youth: past, present, and future. Future of Children 21 (1), 19–41.

Perez, Anthony Daniel, Hirschman, Charles, 2009. The changing racial and ethnic composition of the US population: emerging American identities. Population and Development Review 35 (1), 1–51.

Rumbaut, Ruben, 2005. Making of a people. In: Tienda, M., Mitchell, F. (Eds.), Hispanics and the Future of America. National Academy Press, Washington, DC, pp. 16–65.

Statistics Canada, 2001. 2011 Census: Ethnic Origin. http://www12.statcan.gc.ca.

Thornton, Russell, 1989. Cherokee tragedy: the Ridge family and the decimation of a people, second ed. The American Indian Quarterly 13 (2), 202–204.

Tienda, Marta, Morning, Ann, 2001. Population composition by ethnicity and race in North America. In: Smelser, Neil J., Baltes, Paul B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Pergamon, Oxford, pp. 11745–11749.

U.S. Census Bureau. 2001. 2000, 2010 United States Census, Summary File 1.

de Vidas, Anath Ariel, Hoffman, Odie, 2012. Beyond reified categories: multidimensional identifications among black and Indian groups in Columbia and Mexico. Ethnic and Racial Studies 35, 1596–1614.

Xie, Yu, Goyette, Kimberly A., 2004. A Demographic Portrait of Asian Americans. Russell Sage Foundation, New York. Population Reference Bureau, Washington, DC.

Population Cycles, Formal Theory of

Shripad Tuljapurkar, Mountain View Research, Los Altos, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11755–11758, © 2001, Elsevier Ltd.

Abstract

A population cycle is a roughly periodic variation over time in the size or composition of a population. Here, composition refers to the component structure of the population, for example into distinct age classes. The formal theory of population cycles consists of mathematical tools that yield qualitative and quantitative information about the characteristics of cycles, such as their amplitude or period. There are three broad types of cycle: *transient cycles* are a response to a perturbation of finite duration such as a fertility transition; *externally driven cycles* are triggered and sustained by a persistent external perturbation such as an economic cycle; *internally driven cycles* are maintained by interactions within the population such as a feedback between cohort size and cohort fertility. This article discusses the theory for these three types of cycles, dealing mainly with age-structured populations, although other aspects of population structure are mentioned.

A population cycle is a roughly periodic variation over time in the size or composition of a population. Here, composition refers to the component structure of the population, for example into distinct age classes. The formal theory of population cycles consists of mathematical tools that yield qualitative and quantitative information about the characteristics of cycles, such as their amplitude or period. There are three broad types of cycle: *transient* cycles are a response to a perturbation of finite duration such as a fertility transition, *externally driven* cycles are triggered and sustained by a persistent external perturbation such as an economic cycle; *internally driven* cycles are maintained by interactions within the population such as a feedback between cohort size and cohort fertility. This article discusses the theory for these three types of cycles, dealing mainly with age-structured populations, although other aspects of population structure are considered briefly.

Mechanisms Driving Population Cycles

To explain the different mechanisms that drive cycles, and the kinds of formal theory that are required, it is useful to begin with Lotka's classical demographic equation which relates births $B(\mathrm{t})$ at the present time $t > 0$ to past births, by summing the reproductive contributions $\phi(a, t)$ made by individuals of all ages a, as

$$B(t) = \int_0^t \mathrm{d}a\phi(a,t)B(t-a) + H(t). \qquad [1]$$

Here $H(t)$ specifies births to individuals who were present at $t = 0$, and immigration is ignored. The theory of this classical demographic model is discussed in Population Dynamics: Theory of Stable Populations; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Theory of Nonstable Populations.

Transient cycles typically occur in a population whose vital rates, i.e., mortality and fertility, change over a time interval, say from $t = 0$ to $t = t_1$, and are constant thereafter. A practical example of such a perturbation is a decline from high fertility to replacement fertility that starts at $t = 0$ and is completed at $t = \mathrm{t}_1$. After time t_1, the population is described by classical demographic theory and undergoes damped oscillations towards the classical stationary state. But the dynamics during the transition are also oscillatory and require a different treatment.

Externally driven cycles result from a persistent perturbation of vital rates. One example is a periodic variation over time in fertility that is generated by an exogenous force, such as an economic cycle in wages. Another example is an irregular, stochastic variation in fertility or mortality that produces an ongoing stochastic response that has cyclical components. Formally, such perturbations are described by an appropriate time-dependence in the function $\phi(a, t)$.

Internally driven cycles result from changes in vital rates that are driven by changes in population size and composition. In such cases, the net reproduction $\phi(a, t)$ has a functional dependence on the past trajectory of births, or on past vital rates. One example is a relationship between the fertility of a cohort in its childbearing years with the size of the cohort: for a substantive discussion of such relationships, *see* Fertility Transition: Economic Explanations. Another example is provided by inverse projection in historical demography (Lee 1977), in which observed counts of births and deaths over time are used in conjunction with the Lotka equation to estimate the time-variation in the level of mortality rates. Here, the constraint of matching observed counts results in a feedback between the level of mortality and population counts (Wachter 1986).

The formal theory of cycles driven by each of these mechanisms is discussed below, followed by a brief consideration of future directions.

Transient Cycles

A perturbation such as a fertility transition may take decades to complete, and a formal theory must provide formulas for the birth sequence $B(t)$ both during and after the perturbation. Li and Tuljapurkar (1999b) provide an explicit if cumbersome

description of the decaying population cycles that result from a perturbation. This theory shows that the population undergoes damped cyclical changes during and after the transition. The characteristics of these cycles have not been explored except in particular cases; but this analysis makes possible the computation of population momentum for all gradual transitions (Li and Tuljapurkar 1999a). It is worth noting that the analytical results for this case rest on properties of the roots of Lotka's equation. The relationship between these roots and the age-pattern of reproduction has been studied by approximate (Keyfitz 1985) and exact (Wachter 1984) methods that are also useful in the analysis of externally and internally driven cycles.

Externally Driven Cycles

An illuminating treatment of population cycles driven by external periodic variation in reproduction was given by Coale (1972), who showed that the population responds with periodic variation in the numbers and relative sizes of its age components. The challenge for the theory lies in the nonlinearity of eqn. (1) when $\phi(a,t)$ contains a periodic component. To see this, suppose that $\phi(a,t) = \phi(a) + \psi(a,t)$ with ψ being periodic in time t, and suppose that the resulting birth sequence in eqn. (1) has the form $B(t) = B + b(t)$, with b also a periodic function. Substitution of these forms into eqn. (1) shows that the unknown b(t) depends on a nonlinear product $\psi(a, t-a)\ b(t-a)$. General results for nonlinear renewal eqns. of this type are not known. There are analytical approximations which employ an expansion in the amplitude of the periodic perturbation ψ; these describe the amplitude response of the population and the phase difference between population cycles and the external driving cycle (Tuljapurkar 1985). An interesting class of periodically perturbed Lotka models has been described and analyzed exactly by Kim and Schoen (1997), Schoen and Kim (1994, 1997). The latter work demonstrates that it is sometimes possible to solve exactly the homogeneous form of equation (1), i.e., with $H = 0$, for an explicitly time-dependent $\phi(a, t)$. Such analytical examples can be valuable as a source of insight into the dynamics, and as a check on approximate theories intended for less tractable problems.

The analysis of periodic perturbations suggests that a persistent stochastic perturbation of vital rates will also generate a superposition of cycles of population change. The dynamics of a population subject to stochastic perturbation are also nonlinear, so again the theory has relied upon approximations. Lee (1975) pioneered the application of time-series methods to the linearized analysis of stochastic perturbations. To go beyond linearization, e.g., if the perturbations are large or have strong positive serial correlations, a higher order approximation may be used as discussed by Tuljapurkar (1990), Tuljapurkar and Lee (1997).

Internally Driven Cycles

The formal analysis of internally driven cycles is quite distinct from the analyses used in classical demography, and draws on developments in nonlinear dynamics (for a review of modern nonlinear techniques for dynamics, see Guckenheimer and Holmes (1983)). To start with, it is important to recognize that there are different forms of internal feedback. One class of feedback is exemplified by cohort size effects on fertility: large birth cohorts go on to have lower fertility than their parents and the opposite for small birth cohorts (Lee 1974, 1976, Wachter and Lee 1989). A second class of feedback is generated by social norms or expectations: for example, individual fertility is influenced by the average fertility in the population (Kohler 2000). A third class of feedback arises through an interaction between a population and external variables – for example, economic conditions, environmental conditions, or resources such as land holdings – and these external variables in turn influence fertility or mortality (Bonneuil 1990, 1992, 1997, Chu 1998, 1999, Feichtinger and Sorger 1990, Feichtinger and Dockner 1990, Lee 1987, Milik and Prskawetz 1996).

The formal theory of such feedbacks shares several elements that may be illustrated by considering a nonlinear version of Eqn. (1). Suppose that the net reproduction $\phi(a,t)$ in that equation changes with the population structure: for example, the fertility of individuals aged a at time t depends on the size of that cohort relative to the size of their parents' cohort. Such feedbacks result in a functional relationship of the form

$$\phi(a,t) = \phi(a)h(g,a,\{B(t-s), 0 < s < S\}). \qquad [2]$$

Here the function h on the right describes the effect of the other birth cohorts, born as long as $S > 0$ years ago, on the fertility of the cohort aged a. The parameter g controls the strength of the feedback, for example, by controlling the gradient of h. Lotka's equation is now a nonlinear renewal equation, which can be written in abstract form as

$$B = \mathscr{L}(\mu, B). \qquad [3]$$

The nonlinear operator $\mathscr{L}$ maps past values of births into their current value. This verbal description can be formalized suitably; for an example of the mathematical details see, e.g., Cushing (1978).

In most applications the nonlinear eqn. (3) has a constant solution, $E(\mu)$. The dynamic stability of this solution is examined by studying small deviations $x(t) = \{B(t)-E(n)\}$: stability means that $x(t)$ should decrease over time. An expansion of the nonlinearity in eqn. (3) yields a series of terms of the form

$$x = \mathscr{H}(\mu,x) + \mathscr{G}(\mu,x,x) + \mathscr{R}. \qquad [4]$$

The operator $\mathscr{H}$ is a linear operator, $\mathscr{G}$ is quadratically nonlinear, and $\mathscr{R}$ involves higher-order nonlinear terms. Local stability is determined by the eigenvalues of the linear operator $\mathscr{H}$, and these in turn depend on the elasticity of the linear response (Wachter 1991a). In typical examples, the equilibrium is locally stable when $g < g^*$, say, but becomes unstable when $g \geq g^*$. Local instability means that small deviations from the equilibrium births will change with time as $\exp\{(p + iq)t\}$ where $p > 0$ and q are given by eigenvalues of the linear operator $\mathscr{H}$. The local change in $x(t)$ follows a cycle of period $(2\pi/q)$. The central question now is: when such local instabilities grow and experience the nonlinear forces in equation (4), will they turn into sustained, stable cycles? If they

do, increases in the parameter g past the threshold value g^* will generate a sustained population cycle, called a *limit cycle*. Since $E(\mu)$ is still a constant solution, eqn. (3) (or its solution) is said to have undergone a bifurcation at $g = g^*$.

For any value of g above but not far from the threshold g^*, the nonlinear limit cycle may be represented by a superposition of periodic solutions to the linear part of eqn. (3). Numerical illustrations of such limit cycles are given by Frauenthal and Swick (1983). A systematic procedure for computing the amplitude, period, and stability properties of this limit cycle beyond the bifurcation point is described by Tuljapurkar (1987), and substantive applications that illustrate these computations are given by Wachter (1991b) and Chu (1998). These procedures may be extended to many other models with feedback, and provide an important part of the theory of population cycles. Other feedback mechanisms may result in a different approach to bifurcation (e.g., bifurcation in fertility level, Kohler 2000).

New Directions

The formal theory of cycles has moved considerably beyond classical demographic theory. The nonlinear bifurcation analysis discussed above can be considerably extended. Wachter (1994) has explored global properties of feedback models, addressing the behavior of cycles far above a bifurcation threshold. Bonneuil (1992), Prskawetz and Feichtinger (1995), and Tuljapurkar et al. (1994) show that increasing feedback can lead to more complex types of population cycles including aperiodicity and chaos. Purely internal feedbacks may not be strong enough to generate demographic cycles in human populations, but feedbacks mediated by social or external interactions may well be strong enough. The challenges that face the theory now lie in two areas: confrontation with data to test and modify theories, and the development of new models that incorporate a wider range of demographic dynamics and interaction. As yet, nonlinear cycles in human populations have not been examined in the realistic and rich way that biologists have explored cycles in the structure of laboratory populations (de Roos 1997, Cushing 1997).

Bibliography

Bonneuil, N., 1990. Turbulent dynamics in a 17th century population. Mathematical Population Studies 2, 289–311.

Bonneuil, N., 1992. Attractors and confiners in demography. Annals of Operations Research 37, 17–32.

Bonneuil, N., 1997. Transformation of the French Demographic Landscape, 1806–1906. Oxford University Press, Oxford.

Chu, C.Y.C., 1998. Population Dynamics: A New Economic Approach. Oxford University Press, Oxford.

Chu, C.Y.C., 1999. Predator-prey models with endogenous decisions. Mathematical Population Studies 8, 55–71.

Coale, A.J., 1972. The Growth and Structure of Human Populations: a mathematical investigation. Princeton University Press, Princeton, NJ.

Cushing, J.M., 1978. Nontrivial-periodic solutions of some Volterra integral equations. Lecture Notes in Mathematics 737, 50–66.

Cushing, J.M., 1997. Nonlinear matrix equations and population dynamics. In: Tuljapurkar, S., Caswell, H. (Eds.), Structured Population Models in Marine, Terrestrial and Freshwater systems. Chapman and Hall, New York.

de Roos, A.M., 1997. A gentle introduction to physiologically structured models. In: Tuljapurkar, S., Caswell, H. (Eds.), Structured Population Models in Marine, Terrestrial and Freshwater systems. Chapman and Hall, New York.

Feichtinger, G., Dockner, E.J., 1990. Capital accumulation, endogenous population growth, and Easterlin cycles. Journal of Population Economics 3, 73–87.

Feichtinger, G., Sorger, G., 1990. Capital accumulation, aspiration adjustment, and population growth: limit cycles in an Easter-lin-type model. Mathematical Population Studies 2, 93–103.

Frauenthal, J.C., Swick, K.E., 1983. Limit cycle oscillations of the *human population*. Demography 20, 285–298.

Guckenheimer, J., Holmes, P., 1983. Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. Springer, New York.

Keyfitz, N., 1985. Applied Mathematical Demography. Springer, New York.

Kim, Y.J., Schoen, R., 1996. Populations with sinusoidal birth trajectories. Theoretical Population Biology 50, 105–123.

Kohler, H.P., 2000. Social interactions and fluctuations in birth rates. Population Studies 54, 223–227.

Lee, R.D., 1974. Formal dynamics of controlled populations and the echo, the boom and bust. Demography 11, 563–585.

Lee, R.D., 1975. Natural fertility, population cycles, and the spectral analysis of births and marriages. Journal of the American Statistical Association 70, 295–304.

Lee, R.D., 1976. Demographic forecasting and Easterlin hypothesis. Population and Development Review 2, 459–468.

Lee, R.D., 1977. Methods and models for analyzing historical series of births, deaths and marriages. In: Lee, R.D. (Ed.), Population Patterns in the Past. Academic Press, New York.

Lee, R.D., 1987. Population cycles. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave; A Dictionary of Economic Theory and Doctrine. MacMillan, UK.

Li, N., Tuljapurkar, S., 1999a. Population momentum for gradual demographic transitions. Population Studies 53, 255–262.

Li, N., Tuljapurkar, S., 1999b. The solution of time-dependent population models. Mathematical Population Studies—J. Demog. 7, 311–329.

Milik, A., Prskawetz, A., 1996. Slow–fast dynamics in a model of population and resource growth. Mathematical Population Studies 6, 155–169.

Prskawetz, A., Feichtinger, G., 1995. Endogenous population growth may imply chaos. Journal of Population Economics 8, 59–80.

Schoen, R., Kim, Y.J., 1994. Cyclically stable populations. Mathematical Population Studies 4, 283–295.

Schoen, R., Kim, Y.J., 1997. Exploring cyclic net reproduction. Mathematical Population Studies 6, 277–290.

Tuljapurkar, S., 1985. Population dynamics in variable environments. 6 Cyclical environments. Theoretical Population Biology 28, 1–17.

Tuljapurkar, S., 1987. Cycles in nonlinear age-structured models. I. Renewal equations. Theoretical Population Biology 32, 26–41.

Tuljapurkar, S., 1990. Population Dynamics in Variable Environments. Springer, New York.

Tuljapurkar, S., Boe, C., Wachter, K., 1994. Nonlinear feedback dynamics in fisheries: analysis of the DeRiso–Schnute model. Canadian Journal of Fisheries and Aquatic Science 51, 1462–1473.

Tuljapurkar, S., Lee, R., 1997. Demographic uncertainty and the stable equivalent population. Mathematics and Computer Modelling 26, 39–56.

Wachter, K.W., 1984. Lotka's roots under rescalings. Proceedings of the Natural Academy of Sciences of the USA 81, 3600–3604.

Wachter, K.W., 1986. Ergodicity and inverse projection. Population Studies 40, 275–287.

Wachter, K.W., 1991a. Pre-procreative ages in population stability and cyclicity. Mathematical Population Studies 3, 79–104.

Wachter, K.W., 1991b. Elusive cycles: are there dynamically possible Lee-Easterlin models for U.S. births? Population Studies 45, 109–135.

Wachter, K.W., 1994. The Cohort Feedback model with symmetric net maternity. Mathematical Population Studies 5, 25–44.

Wachter, K.W., Lee, R.D., 1989. U.S. births and limit cycle models. Demography 26, 99–115.

Population Dynamics: Classical Applications of Stable Population Theory

Thomas W Pullum, University of Texas, Austin, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11758–11762, © 2001, Elsevier Ltd.

Abstract

If age-specific fertility and mortality rates are constant for a long period of time, a population will approach a fixed age distribution, and fixed overall rates, which can be calculated from the specific rates. An assumption of stability may thus enable the indirect estimation of missing data when only a few population characteristics are known. Even when the assumption is not satisfied, the stable population model provides a useful approximation to the relationships among survivorship, reproduction, and age structure. Several applications are provided, ranging from the simple exponential model of population growth to models that estimate numbers of kin.

Introduction

The theory of stable populations is presented in both continuous and discrete form in the article Population Dynamics: Theory of Stable Populations. Virtually any real population, whether contemporary or historical, deviates from stability, but the stable model produces a standard for comparison based on simplifying assumptions that allow us to handle data limitations. Some applications of the theory appear in the introductory article just mentioned and in other articles, including Demographic Techniques: Indirect Estimation; Population Dynamics: Momentum of Population Growth. The present article will review several applications within a structure that specifies the amount of information that is required. The applications are 'classical' in the sense that they are relatively well known.

It is helpful first to distinguish between two classes of models that are widely used: survival processes and renewal processes. The prototypical example of a survival process is the single-sex, single-decrement life table, described in the article Life Table. A survival process describes a cohort, real or synthetic, from age 0 until some terminal age ω. The main concepts are (a) the survival function, $l(x)$, namely the proportion of the cohort that survives to each age x, and (b) the mean waiting time or remaining life expectancy, e_x^o, at any specific age x.

A renewal process, by contrast, is a survival process with reproduction. With this component, the process generates successive cohorts that are (in general) larger than necessary simply to replace the persons who have died. The cohort interpretation is supplemented with the asymptotic implications of fixed and stable overall rates (of growth, mortality, and fertility), and a fixed and stable age distribution, none of which depends on initial conditions.

If a renewal process is in equilibrium – that is, if reproduction is exactly at replacement value, then the stable population is stationary. It has an intrinsic growth rate of zero, and one does not need explicit age-specific birth rates to analyze it, for its overall birth rate is identical to the overall death rate. There are several classical applications and generalizations of this scenario, all of which can be regarded as applications of stable population theory because a survival process is a special stable renewal process.

A stable population is typically motivated as the asymptotic consequence of projecting the population indefinitely with fixed age-specific rates. Its overall rates and age distribution converge to functions that do not depend on the initial conditions. However, unless the intrinsic growth rate is zero, there is a fundamental implausibility in this asymptotic interpretation. If the intrinsic growth rate is positive, then the overall population size will increase exponentially, an outcome that cannot be sustained indefinitely in a real population. If the intrinsic growth rate is negative, then the overall population size will *decrease* exponentially until eventual extinction. In probabilistic terms, the fixed-rates assumption implies almost certain ultimate extinction or explosion (*see* Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems). It is probably more useful to regard the stable population parameters simply as a synthesis of the underlying age-specific probabilities or rates.

Applications Not Requiring Schedules of Rates

Applications of stable population assumptions can be organized according to the amount of information they utilize. Probably the simplest possible applications occur when the only parameter utilized is the intrinsic growth rate, r. If we assume stability between two times t_1 and t_2, and that the population size at those two times is n_1 and n_2, then

$$n_2/n_1 = e^{r(t_2-t_1)}$$

Solving for r, we get

$$r = (\ln n_2 - \ln n_1)/(t_2 - t_1)$$

We can then estimate the population size at an intermediate time, $t_1 \le t \le t_2$, to be $n_t = n_1 e^{r(t-t_1)}$. This formula, or variations upon it, is often used when the population is not stable, in the absence of better information. Another simple and well-known application is to determine the doubling time associated with a given value of r. Setting $n_2 = 2n_1$ and letting T be the time required for the population to double, we obtain $2n_1 = n_1 e^{rT}$, leading to $T = (\ln 2)/r = 0.6931/r$. For example, if the annual growth rate is 0.01, or 1 percent, a stable

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31134-5

population will double in 69.31 years, or approximately seventy years. These and related applications are discussed by Hinde (1998: pp. 154–8), for example.

Suppose that in addition to the intrinsic growth rate, r, we also know the intrinsic birth rate, b (and hence the intrinsic death rate $d = b - r$). Keyfitz (1977: pp. 12–14) uses such information to estimate the number of persons who have ever lived on Earth, which is the same as the number of persons ever born. His strategy is to piece together long intervals of time within which stability (at different levels) is assumed. If the population is stable between t_1 and t_2, then the number of person-years lived in that time interval will be

$$PY = \int_{t_1}^{t_2} n_1 e^{r(t-t_1)} dt = (n_2 - n_1)/r$$

PY can be converted to the corresponding number of births by multiplying by b:

$$B = bPY = b(n_2 - n_1)/r$$

(If $r = 0$, then $n_1 = n_2 = n$, $PY = n(t_2 - t_1)$, and $B = bn(t_2 - t_1)$.)

Applications that Assume Schedules of Rates

Another class of applications arises if we know r, b, and d, and do not know the survival and fertility schedules but are willing to make some assumptions about them. Coale and Demeny (1983) blended empirical life tables into four sets of so-called regional model life tables (north, south, east, and west) for males and females, keyed to the life expectancy e_0^o at birth and e_{10}^o at age 10 (*see* Demographic Models). They also developed plausible fertility schedules, varying according to the mean age of childbearing but adding up to one child, so that the entire schedule could be multiplied by a hypothetical total fertility rate (TFR). With this information, much of it hypothetical, they were able to generate sets of stable age distributions. It is then possible to estimate a wide range of population characteristics, using only a few pieces of summary information (along with a specification of region and sex). The summary information typically includes certain combinations of the following: a measure of growth, such as r; a measure of fertility, such as b, the gross reproduction rate (GRR), or the TFR; a measure of mortality, such as d or the average age at death; or a measure of the age distribution, such as the proportion of persons below age 20. Stable age distributions generated in this way can also be used to approximate real age distributions in settings of deficient data. Some methods for the computation of population parameters using the stable population model are described in Demographic Techniques: Indirect Estimation.

Applications that Require Schedules of Rates

A number of analyses of population dynamics have resulted from examining the sensitivity of population characteristics to disturbances of a stable population. These applications require knowledge of the age-specific schedules for mortality (say m_x for each age x), of fertility (say f_x for each x), and for any other transitions in the model. For example, Preston (1974) took partial derivatives of the basic stable population formulas to estimate the effect of various age-specific mortality changes upon the intrinsic growth rate, birth rate, and age composition. He found that the effects could be represented as simple functions of the age pattern of the cumulative changes in the mortality rates. Probably the best known of these sensitivity analyses has been on the topic of population momentum, which in its simplest form measures the eventual effect on population growth of an immediate shift to replacement fertility (*see* Population Dynamics: Momentum of Population Growth).

The classical stable population model has no in- or out-migration, but several authors have used the model in innovative ways to evaluate the impact of emigration (Keyfitz, 1971) or immigration (Lesthaeghe, et al., 1988). Mitra (1990) considered how a population could maintain stationarity through a combination of below-replacement fertility and a steady stream of immigration. With assumptions about the ages and fertility of the immigrants, it is possible to determine the necessary size of the stream of immigration. This model is relevant, at least as a first approximation, to the current mix of below-replacement fertility and immigration in Europe and North America.

The classical formulation of stable population theory only allows individuals to be born, procreate, and die, but later developments permit individuals to move between several statuses or states during their lifetime, as in multistate models (*see* Multistate Transition Models in Demography). In one obvious application the states represent regions of a territory (Feeney, 1970). In another, Keyfitz developed a linkage between aging and career mobility in stable populations (1973; the title of this paper is misleading because he included stable as well as stationary populations). He supposed that advancement in status across a set of discrete categories is essentially based on the steepness of the age distribution. In a setting of high growth, age (as a proxy for experience) tends to produce earlier advancement because there are relatively few people at higher levels. In a setting of low growth, seniority is less valued, because there are relatively many people at higher levels, and promotion will take longer. The approach is again a type of sensitivity analysis, in which the age at reaching a sequence of discrete gateways is expressed as a function of stable population parameters (in particular, the intrinsic growth rate).

Applications that Link Individuals to One Another: Kinship

Assume now that we have access to complete schedules of survival and fertility. In continuous form, say that $l(x)$ is the probability that a woman will survive to age x and $m(x)\,dx$ is the probability that a woman will have a daughter between ages x and $x + dx$ (with $m(x) > 0$ for x in the reproductive age range $\alpha < x < \beta$ within the full range of ages $0 < x < \omega$).

Single-sex kinship provides a structure by which individuals are linked through lines of descent or ancestry. Two specific models for kinship structures will be described here. In the context of kinship, reference individuals (ego) are linked to their direct descendants, such as daughters; to progenitors, such as mothers; or to their lateral kin, such as sisters. Two-sex approximations are possible. The main outputs are the

probability that a specific progenitor is still alive when ego is age a, and the expected number of descendants and lateral kin (in a specific category) who are still alive, or were ever born, when ego is age a. These are obtained from various convolutions of the survival and fertility functions.

The first kinship model to be described was developed by Goodman et al. (1974, 1975). The following exposition will be limited to surviving kin when ego is age a. It can be modified to refer to kin ever born simply by deletion of the outermost survival function (that is, setting it equal to 1).

Each kinship role is expressed with a different formula, but the strategy for generating these formulas is the same for all roles. The steps to link ego with a consanguineal relative are to identify the relevant sequence of births, and the ages of the respective mothers at those dates; to incorporate the fertility rates for those ages, the probability of surviving to those ages, and, for the final link, the probability of surviving to the time when ego is age a; and integrating over all allowable combinations of ages of the linked kin. Note that the birth of ego herself, and her survival to age a, are assumed.

For example, to link ego with her daughter, the relevant birth is that of the daughter, which occurred when ego was some previous age x. If the daughter is still alive when ego is age a, then the daughter's age is $a - x$. Thus ego experienced the fertility rate for age x, and the daughter survived to age $a - x$. Multiplying in reverse order and integrating over allowable ages x, the expected number of surviving daughters for a woman age a is

$$D(a) = \int_{\alpha}^{a} l(a-x)m(x)\mathrm{d}x$$

(If $a < \alpha$ then $D(a) = 0$ and if $a > \beta$ then $D(a) = D(\beta)$.)

The stable population probability that ego's mother is alive when ego is age a was originally calculated by Lotka (1931), and is useful for estimating adult mortality when some empirical information about the incidence of orphanhood is available. If the mother was age x when ego was born, then the probability that she is still alive a years later is the probability of surviving to age $a + x$, $l(a + x)/l(x)$. Obviously this depends on the value of x, as well as a, and these survival ratios must be weighted by the probability that the mother was age x to age $x + \mathrm{d}x$ when ego was born. In a stable population these weights will be given by $W(x)\mathrm{d}x = \mathrm{e}^{-rx}l(x)m(x)\mathrm{d}x$ (from the characteristic equation, this is the density function for mothers' ages and integrates to 1). Therefore,

$$M_1(a) = \int_{\alpha}^{\beta} [l(a+x)/l(x)]W(x)\mathrm{d}x = \int_{\alpha}^{\beta} e^{-rx}l(a+x)m(x)\mathrm{d}x$$

is the probability that ego's mother is alive when ego is age a.

The expected number of surviving sisters when ego is age a involves a similar logic but will not be given here. Ego is linked to her sisters through their mother. It is necessary to distinguish between older and younger sisters because the survivorship of the mother is problematic after, although not before, the birth of ego. The formulas can be extended to aunts, nieces, first cousins (mother's sister's daughters), and indeed any category of kin.

A noteworthy feature of this application of the basic model of a stable population is that it requires only the basic survival function and maternity function, and a structure to link individuals. As an alternative to linkage by kinship, it should be possible to substitute an organizational structure or some form of vertical and horizontal connections, but we are not aware of any analogous applications.

The formulas are deterministic, rather than stochastic, and will not produce distributions. They assume that the birth of a child is independent of the number of children a woman has already had, that survivorship is independent of biological relationship, and that the population is homogeneous. Although these limitations are severe, it would be difficult to find data corresponding to more realistic assumptions. Calculations using these multiple integrals require some form of quadrature.

In a second application of kinship (Pullum, 1982), counts are not conditional upon the survival of ego or any other kin, but are defined as 'eventual expected' numbers of kin. For example, the net reproduction rate

$$N = \int_{\alpha}^{\beta} l(x)m(x)\mathrm{d}x$$

is the average number of female births that will be produced, a generation later, from a single female birth. This is equivalent to the eventual expected number of daughters that a woman (ego) will have, defined at the point of ego's birth and the daughters' births.

The eventual expected number of sisters, defined at the point of ego's births and the daughters' births, can be referred to with the symbol S. N is not equivalent to any of the functions defined by Goodman et al. (1974) but S does appear within their scheme as the (expected) number of sisters (older and younger added together) ever born by the time ego has reached age $\beta - \alpha$, the maximum age at which a sister can be born. This formula superfluously conditions on the survival of ego, but allows for all sisters to be born.

It can be shown that N and S are sufficient to obtain the eventual expected numbers of any other types of kin. For example, an average of N^2 granddaughters (daughters of daughters) will be born, SN first cousins (daughters of sisters of the mother), SN nieces (daughters of sisters), and so on. The sex ratio at birth will also be constant in a stable population, and by including it one can easily extend to a two-sex kinship structure. For example, define s to be the ratio of males to females at birth, so that for every female birth there will be $1 + s$ births of males *and* females. This implies that for every female child there will be $1 + s$ children, so N is replaced by $(1 + s)N$; for every sister there will be $1 + s$ siblings, so S is replaced by $(1 + s)S$. (Note that instead of one mother there will be exactly two parents.) The number of grandchildren (sons and daughters of sons and daughters) will be $(1 + s)^2N^2$; the number of first cousins (sons and daughters of brothers and sisters of the mother or father) will be $2(1 + s)^2SN$; the number of nieces/nephews (sons and daughters of brothers and sisters) will be $(1 + s)^2SN$; and so on.

Like the applications given in Section Applications that Assume Schedules of Rates above, this second model of kinship only requires summary measures of the stable population (in this case S and N), and does not require any schedules of rates.

Conclusion

These applications have been presented largely at a conceptual level, without the specific data or numerical conclusions that might help to demonstrate their value (the bibliography and other entries in this encyclopedia describe several further examples). In each case, however, the original goal was to utilize stable population theory to draw approximate conclusions about real populations in situations where data are scarce. The model approximation has certain elegant formal properties and in effect permits the simulation of data that are generally not available for real populations. Indeed, significant improvement in the availability of demographic data is a likely reason why there have been relatively few new applications of the theory in recent decades.

See also: Demographic Techniques: Indirect Estimation; Life Table; Multistate Transition Models in Demography; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Stable Populations.

Bibliography

Bourgeois-Pichat, J., 1971. Stable, semi-stable populations and growth potential. Population Studies 25, 235–254.

Cerone, P., 1987. population theory with immigration. Demography 24, 431–438.

Coale, A.J., 1981. Robust estimation of fertility by the use of model stable populations. Asian and Pacific Census Forum 8, 5–10.

Coale, A.J., Demeny, P., 1983. Regional Model Life Tables and Stable Populations, second ed. Academic Press, New York.

Espenshade, T.J., Bouvier, L.F., Arthur, B., 1982. Immigration and the stable population model. Demography 19, 125–133.

Feeney, G.M., 1970. Stable age by region distributions. Demography 7, 341–348.

Goodman, L.A., Keyfitz, N., Pullum, T.W., 1974. Family formation and the frequency of various kinship relationships. Theoretical Population Biology 5, 1 -27. Addendum 1975 Theoretical Population Biology 8, 376–381.

Hinde, A., 1998. Demographic Methods. Arnold, London.

Keyfitz, N., 1971. Migration as a means of population control. Population Studies 25, 63–72.

Keyfitz, N., 1973. Individual mobility in a stationary population. Population Studies 27, 335–352.

Keyfitz, N., 1977. Applied Mathematical Demography. Wiley, New York.

Keyfitz, N., 1985. The demography of unfunded pensions. European Journal of Population 1, 5–30.

Keyfitz, N., Nagnur, D., Sharma, D., 1967. On the interpretation of age distributions. Journal of the American Statistical Association 62, 862–874.

Lapkoff, S.F., 1991. A research note on Keyfitz's 'The demographics of unfunded pensions. European Journal of Population 7, 159–169.

Lesthaeghe, R., Page, H., Surkyn, J., 1988. Are immigrants substitutes for births? Unpublished paper, European University, Florence. Published in German in 1991 as Sind Einwanderer ein Ersatz für Geburten? Zeitschrift für Bevölkerungswissenschaft 3, 281–314.

Lotka, A.J., 1931. Orphanhood in relation to demographic factors: A study in population analysis. Metron 9, 37–109.

Mitra, S., 1983. Generalization of the immigration and the stable population model. Demography 20, 111–115.

Mitra, S., 1990. Immigration, below-replacement fertility, and long-term national population trends. Demography 27, 121–130.

Preston, S.H., 1974. Effect of mortality change on stable population parameters. Demography 11, 119–130.

Pullum, T.W., 1982. The eventual frequencies of kin in a stable population. Demography 19, 549–565.

Schoen, R., 1986. The two-sex multiethnic stable population model. Theoretical Population Biology 29, 343–364.

Population Dynamics: Mathematic Models of Population, Development, and Natural Resources

Alexia Fürnkranz-Prskawetz, Vienna University of Technology, Vienna, Austria; Vienna Institute of Demography, Austrian Academy of Sciences, Vienna, Austria; and Wittgenstein Centre of Demography and Global Human Capital (WIC), Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Whereas interactions of population and natural resource growth have generally been modeled using large-scale numerical simulation studies, we review small-scale models that permit partially analytical results and can capture some of the salient features observed in history. This article refers only to models that concentrate on renewable resources since these are the most interesting and least understood as regards changes in population structures and dynamics. It starts by reviewing models of population and resource dynamics that refer to subsistence economies. Within this class of models it is possible to investigate how the equilibrium between the resource and population stock changes, dependent on the functional relations that govern the dynamics of population and resources. But as societies become less bound to the land, more urbanized, and more technological, resources might still be at risk, albeit no longer from positive population growth but rather from the environmental costs of consumption and production. To capture the impact of population growth on the environment, demographic impact models have been developed aimed at decomposing the effect on the environment from population on the one hand and that from consumption and technology on the other. The article concludes our review with a look at models of environmental influences on population dynamics.

Although the importance of the links between population and environment is widely appreciated, there is little agreement in the literature about the nature of these links.

The dynamic link between population and natural resources – influences are never entirely unidirectional or constant over time – constitutes the main difficulty in assessing formal models and unified theories. To complicate matters further, this interrelationship depends on the level and type of economic activities, technological development, institutional settings, and cultural systems. Even the variables themselves are not uniquely defined, but depend rather on the level and purpose of the analysis.

This article does not survey numerical simulation models of demographic/economic/environmental interactions (see Sanderson, 1994 for a review), of which the World 3 Model (Meadows et al., 1972) is the most popular. Instead, it looks at some simple mathematical models that capture stylized facts of the link between population and natural resource dynamics. Such 'mathematical cartoons' help us to explain and understand how various population and environmental characteristics might affect each other, and they can teach us how to respond most effectively to various demographic and environmental developments.

Although it was exhaustible resources that first received theoretical attention, this article concentrates on renewable resources since these are the most interesting and least understood as regards changes in population structures and dynamics. In contrast to nonrenewable resources, the use of which is controlled largely by market prices, renewable ones are very often open-access resources. The overuse of resources if property rights are absent or poorly defined is called the 'first tragedy of the commons' (Hardin, 1968) and can be corrected by restricting access to the resource. But as emphasized by Lee (1990), if the population is not fixed – as is presupposed in the models presented in this article – free access through reproduction constitutes the 'second tragedy of the commons'. More recently the externality of childbearing in relation to pollution externality has been studied in Harford (1997, 1998) and Schou (2002).

Malthusian Population Dynamics and Natural Resources

The study of interactions between population and resources has a long history. According to Malthus, population growth reduces material welfare due to diminishing returns to labor on a fixed supply of land. On the other hand the higher the level of material welfare the higher the population growth rate will be. The Malthusian model predicts that "population will equilibrate with resources at some level mediated by technology and a conventional standard of living" (Lee, 1986). Improvements in technology will be offset in the long run by increases in the size of the population, but the standard of living will not be related to the level of technology. As such, the Malthusian model provides a description of a rather primitive society with incomes not too far above the subsistence level and where local renewable resources are an important part of the economic production process.

Renewable resources (agricultural land, forests, lakes, etc.) are not in fixed supply as Malthus assumed. Renewable resources regenerate, but if the rate of utilization (harvest) exceeds the rate of regeneration, a renewable resource will be depleted or, in the extreme case, irretrievably exhausted. By adding the dynamics of renewable resource growth to the dynamics of population growth, the Malthusian model is capable of explaining patterns of population growth and resource degradation that do not necessarily end up in a single

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31005-4

equilibrium (Malthusian trap) (see Brander and Taylor, 1998; Prskawetz et al., 1994).

The structure of these models can best be described in terms of prey–predator dynamics, with the resources, R, being the prey and the population, P, acting as the predator. The dynamics of renewable resources (eqn [1]) is commonly described by the standard model of mathematical bioeconomics, where the net growth of the renewable resource, $\mathrm{d}R(t)/\mathrm{d}t$, is affected by two counteracting factors: indigenous biological growth, $g[R(t)]$, and the harvest, $H[R(t), P(t)]$, which depends on the stock of resources available to be harvested and the number of people who are harvesting. Indigenous resource growth is modeled by the logistic growth function $g[R(t)] = aR(t)[K - R(t)]$, where the coefficient K determines the saturation level (carrying capacity) of the resource stock (i.e., K is the stationary solution of R if the resource is not degraded) and parameter a determines the speed at which the resource regenerates. The functional form of the harvest function is determined by the prevailing economic structure, and it establishes the link between the stock of resources and population. The population growth rate, $(\mathrm{d}P(t)/\mathrm{d}t)/P(t)$ (eqn [2]), is modeled as an increasing function of material welfare $\gamma(t) = \gamma[P(t), H(t)]$, which is determined by the level of the harvest and will be reduced by population growth. Whenever material welfare falls below the subsistence level, the population will decline. These Malthusian population dynamics imply that population growth may well adapt to resource constraints in contrast to models with exogenous positive population growth, where the economy collapses if resources do not regenerate quickly enough.

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = g[R(t)] - H[R(t), P(t)] \qquad [1]$$

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = n[\gamma(t)]P(t) \qquad [2]$$

Within this class of models it is possible to investigate how the equilibrium between resources and population stock changes, dependent on the functional relations that govern the dynamics of population and resources. For instance, the efficiency of harvest technology, the degree of substitution between labor and resources, the indigenous rate of resource growth, and the carrying capacity of the resource stock will have an effect on material welfare and hence on population growth. In turn, fertility and mortality will affect the resource dynamics via the input of changing labor stocks in the harvest. An empirical calibration of the model for the case of the small Pacific Easter Island characterized by pronounced population fluctuations was carried out by Brander and Taylor (1998) in the framework of a general-equilibrium model.

During the last decade, the model by Brandner and Taylor has been extended in several directions. Dalton and Coats (2000) studied the introduction of different structures of property rights (common access vs personal ownership of resources) as represented by alternative specifications of the link between the dynamics of natural resources and population growth. Their results indicate that the steady-state solutions for population and resources are unaffected by the institutional setup, whereas the transition toward the long-run equilibrium will depend on the prevailing institutional setup exhibiting less (more) fluctuations in the dynamics of population and resources under private (common) property resources. These results will change if institutions also affect other parameters and functional forms of the model such as technology, preferences, fertility, and mortality (e.g., D'Alessandro, 2007). Institutional settings are also discussed in Anderies (2000); both for the Brandner and Taylor model of Easter Island as well as the model of the Tsembaga society of New Guinea (Anderies, 1998). The Tsembaga people is provided as an example, where institutional settings, such as a ritual cycle in which pigs are killed and warfare is initiated, serve to reestablish a sustainable equilibrium of population and resource stocks. Modifying the original Brandner and Taylor model by introducing a subsistence level of consumption, Anderies (2000) shows that changes in population and resources are more rapid as compared to the original Brandner and Taylor model. Anderies then argues that these sudden changes – that fit the historical evidence quite well – did not allow for institutional adaption. Explicit conservation policies (a consumption tax and quotas on total resource harvest, total harvest effort, and per capita harvest effort) are introduced in Pezzey and Anderies (2003). These policies may counteract the destabilizing effect of a larger subsistence level in consumption although their implementation and control may indeed be rather difficult. Besides institutional reforms, several authors have also studied the role of technological progress to prevent the collapse of population and resources in Easter Island. Reuveny and Decker (2000) assume that the growth rate of resources and their carrying capacity, as well as the productivity of harvesting positively depends on technological development. However, as technological progress may allow for faster population growth, it may thereby even foster the collapse of the resource stock and population.

An alternative explanation of the dynamics of population and resources in primitive societies such as Easter Island was proposed by Maxwell and Reuveny (2000). The authors refer to recent research on resource scarcity as it may initiate conflict in developing countries and extend this causal link by studying the repercussion of conflict situations on resource and population dynamics. A variant of the model by Maxwell and Reuveny can be found in Prskawetz et al. (2003). Instead of assuming two separate models that present either a conflict or no-conflict situation, they postulate that changes in the labor force participation rate, the death rate, and the growth rate of resources are 'smooth' functions of the prevailing level of per capita resources. Most recently de la Croix and Dottori (2008) explain the history of Easter Island by introducing a model of noncooperative bargaining between two competing groups where population growth acts as a means to gain power. Fertility behavior between both groups is therefore guided by strategic complementarities – a new motive to explain population dynamics. The authors discuss the conditions under which a population race may set in, which ultimately results in resource degradation and the collapse of the society. A recent review of the various extensions of the Brandner and Taylor model is presented in Nagase and Uehara (2011).

Contrary to the Malthusian predictions, increasing population densities might be beneficial, as argued by Boserup (1981), and could well increase the human carrying capacity of the earth. Higher population densities will initiate technological innovations in agriculture, thereby increasing the yields so that the natural environment can support a larger population without reducing the level of welfare. Similar positive feedback mechanisms are captured in a simple mathematical cartoon of the interdependence between the growth in population $P(t)$ and carrying capacity $K(t)$ (expressed in numbers of individuals) by Cohen (1995) as

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = rP(t)[K(t) - P(t)] \qquad [3]$$

$$\frac{\mathrm{d}K(t)}{\mathrm{d}t} = c\left[\frac{\mathrm{d}P(t)}{\mathrm{d}t}\right] \qquad [4]$$

with $r > 0$ and c either negative, zero, or positive. The parameter c, which captures the effect of an increment of population on the carrying capacity, determines the long-term population dynamics, and it can represent technological innovation. When $c = 1$ population size grows exponentially (Euler, eighteenth century); when $c < 1$ population grows logistically (Verhulst, nineteenth century); and when $c > 1$ population grows faster than exponentially (Forester and coworkers, twentieth century). The positive correlation between population growth and technological progress is taken up by Dalton et al. (2005) who extend the model by Reuveny and Decker (2000) allowing for endogenous technological progress. Whether long-term economic growth with positive levels of populations and resources is viable depends on the division of technological progress between being 'resource-depleting' (by increasing the efficiency of harvest) or 'resource-conserving' (by increasing the growth rate of resources) (Dalton et al., 2005: p. 36).

Malthusian results (in the sense that population will grow to the point where material welfare matches subsistence demand) can be further undermined by allowing population growth to be a choice variable. Zero population growth may well be the optimal choice for individuals in the economy. Since environmental changes are often slow over the course of an individual life span, and since environmental damage may outlive its perpetrators, overlapping generation models (Eckstein et al., 1988) provide an appropriate demographic structure. Intra- as well as intergenerational conflicts can be modeled in such a framework, taking account of the effects of increases in population and resource exploitation on the future population's quality.

Economic Development and the Environment

The models presented in Section Malthusian Population Dynamics and Natural Resources represent traditional societies in which populations derive their living from primary occupations (agriculture, hunting, fishing, etc.) that depend on the availability of resources. But as societies become less bound to the land, more urbanized and more technological, they not only use the environment as a source of natural resources but also as a dump for waste products arising from human activity. Furthermore, an economy's production possibilities are no longer determined by the maximum sustainable yield of renewable resources. Improvements in technology can increase the sustainable yields or reduce the resource stock required for production, and economic growth will allow for the use of artificial capital in place of natural resources. In open economies with trade, technological change, and economic growth, there is no simple and direct relationship between population growth on the one hand and the environment on the other hand. The paper by Lehmijoki and Palokangas (2010) nicely demonstrates – based on a stylized dynamic optimization model – how gains from trade in developing countries may first increase and then decrease environmental pollution and population growth, thereby replicating the time path of the environmental Kuznets curve.

Resources might still be at risk, albeit no longer by positive population growth alone. The risk can stem from the environmental costs of consumption and production. It is therefore of great importance to understand how these environmental impacts depend on different population structures.

An interesting contribution that considers the role of consumption for environmental quality but also allows for investment to increase environmental quality is presented in John and Pecchenino (1994). Based on an overlapping generations model, where the utility of households depends on consumption and environmental quality, they study the link between economic growth and the environment. The short-lived agents may be better-off if a long-lived social planner takes into account the externalities of consumption and maintenance of the environmental quality. The model is extended to allow for population growth in John et al. (1995). Lower population growth together with lower levels of consumption are shown to imply higher levels of environmental quality and economic output in the steady state. Similar to John and Pecchenino (1994) the authors show that a social planner who has a lifetime similar to the environment may internalize the externalities that are not taken into account by short-lived individual households.

The role of preferences and technology in the relation between population size and environmental quality is studied in Cronshaw and Requate (1997) within a static framework, where agents' utility also depends on the aggregate emissions of the economy. The very long-term relation between energy use and economic development, allowing for endogenous population development, is studied in Fröling (2011) within the framework of the unified growth theory.

Besides the mere number or growth of the population, its age structure will be an important variable for long-term economic growth and the environment. Dalton et al. (2008) introduce population age structure into an energy-economic growth model. Calibrating their model to the US shows that population aging reduces long-term emissions and the effect may be similar to the role of technological change.

The Impact of Population on the Environment

In a seminal paper, Ehrlich and Holdren (1971) express the impact on the environment, I (e.g., the value of some

pollutant), as the product of population, P, and per capita impact, $F(P)$:

$$I = PF(P) \quad [5]$$

Depending on whether diminishing returns or economies of scale are dominant, per capita impact can be increasing or decreasing in the total population. The most familiar form of such a demographic impact model is the so-called I-PAT identity, which divides per capita impact $F(P)$ into affluence A, and technology T. The accounting model I-PAT stands for the environmental Impact decomposed into Population size P, Affluence A, and Technology T. Affluence can be measured as Gross National Product (GNP) per capita, and technology represents the mean impact of an extra dollar of production (impact/GNP). The latter variable encompasses the cleanliness of technology, the effects of institutions, and other factors that can alter the impact efficiency of production.

Demographic impact models (Wexler, 1996) constitute multiplicative identities aimed at decomposing the effect on the environment of population on the one hand and that of consumption and technology on the other. The number of variables on the right-hand side of eqn [5] can be any number greater than two. Although these models are commonly used, they face several criticisms. They suppress any feedback relation between the variables on the right-hand side; their implications are highly sensitive to the choice of decomposition variables; and the method of decomposition and the role of institutions and culture are ignored. Recent research on the I-PAT identity has dealt with some of these problems. Expressing the identity in terms of variances rather than means (Preston, 1996) allows one to incorporate the links between the decomposition variables. Considering households instead of individuals as the unit of consumption (MacKellar et al., 1995) highlights the effect of economies of scale as household size increases. Typical results of such analyses are that the population has a much stronger impact on the environment in developing countries than in developed countries. In the case of environmental hazards produced by industrial processes, population growth plays a minor role both in developing and in developed countries. A stochastic version of the I-PAT identity (STIRPAT) and a discussion among the alternative variants of measuring the impact on the environment are presented in York et al. (2003). STIRPAT constitutes a stochastic model and stands for Stochastic Impact by Regression on Population, Affluence, and Technology.

Milik et al. (1996) embed the I-PAT identity into a dynamic model where population P, affluence A, and technology T influence each other and vary over time. Pollution I reduces the regenerative capacity of natural resources, which in turn influences the dynamics of population and affluence. The model highlights the fact that understanding and monitoring the impact of pollution on the environment is important even when outward signs of damage (as a reduction in the resource stock) are not yet visible. Ecosystems often absorb stresses over long periods of time and then eventually reach a disruption level at which the cumulative consequences of stress appear in critical proportions. A policy of waiting until the first signs of environmental deterioration are observed before actions are taken to reduce the impact on the environment may well prove to be disastrous. When action is initiated, the impact on the environment might already be irreversible. Mathematically, these systems can best be described in terms of 'slow–fast' dynamics, where during certain periods some variables vary much more rapidly than others. While population and economic variables can be regarded as slow-moving variables, the environment changes most rapidly once the system has absorbed too much stress. For instance, the emission of carbon dioxide occurred almost unobserved for several decades before any attention was paid to the impact of greenhouse gases on global warming.

It becomes difficult to quantify the impact of population on the environment if the demand for local resources is not related to local population growth and if institutional environments are not stable. In fact, "commercialisation by converting a limited and inelastic subsistence demand to a limitless and elastic export demand (from the standpoint of the region, at least) can lead to much more rapid rates of exploitation than would be implied by population growth alone" (Repetto and Holmes, 1983).

Modeling Environmental Influences on Population Dynamics

Existing models of environmental influences on population exhibit two shortcomings. First, they generally do not distinguish between the separate effects the environment may have on fertility and mortality. Second, the environmental impact on population growth is commonly modeled to work only through the economic and social variables of the model. There are only few examples in the literature that model the direct effect of environmental stress on fertility and mortality.

The first example (Nerlove, 1991) is based on the vicious circle of poverty, population growth, and environmental degradation as evidenced in small rural communities of sub-Saharan Africa. In these economies simple tasks like fetching water and collecting fodder and wood are mostly carried out by children. This means that children are not only valuable to parents for future income but also as a source of current income. Higher population growth leads to more resource depletion, which reduces the marginal productivity of the resource. To offset this effect families have more children, thus depleting the resource even more. In a short parable, Nerlove demonstrates that increasing the birthrate when the environment deteriorates will be the optimal choice for a household that maximizes per capita harvest when 'the perceived marginal product of an additional child increases as environment deteriorates'. In addition, in a system in which land rights are acquired through cultivation, a large number of children can imply increased claim to land ownership. This positive feedback between fertility and environmental degradation is in stark contrast to Malthusian dynamics, where population equilibrates with resources. Since the marginal private and social costs of reproduction are not the same, Nerlove then shows how social interventions in the form of taxes or subsidies can be used to induce a specific birthrate corresponding to a socially desirable stationary state.

The underlying mechanism of the above-mentioned vicious circle is neither population growth nor environmental degradation but poverty, which prevents the substitution of

alternative fuel sources, and the low status of women and girls, which devalues the large amount of time and effort that they must devote to daily gathering activities. To escape this vicious circle one needs to alleviate poverty and educate women.

As suggested by Chu (1998), the increasing specialization and level of economic activity typical of advanced societies may also imply a substantial dependency of population dynamics on the state of the environment. Increasing the level of human activities destroys the earth's biodiversity, so that the ability of the environment to absorb negative shocks will decline, while increased specialization makes people less adaptive to environmental shocks. Chu suggests a simple mathematical model of the survival probability $f(\cdot,\cdot)$ to capture both effects:

$$\begin{aligned} f(m_t, x_t) &< f(m_t', x_t) \quad \text{if } x_t > 0 \\ f(m_t, x_t) &\geq f(m_t', x_t) \quad \text{if } x_t \leq 0 \end{aligned} \qquad [6]$$

where m_t denotes the logarithm of the number of intermediate goods (a measure of the division of labor) with $m_t < m_t'$ and x_t is a realization of the environment value X_t with the critical state $x_t = 0$. The mean and the variance of the random variable X_t will be increasing functions of m_t. By reducing the ability to adapt, increased specialization m_t' will decrease the survival probability if the environment is already in a disastrous state $x_t < 0$, while the opposite relation will hold if the environmental state is favorable $x_t > 0$. These considerations are integrated into a modern Malthusian theory in Chu and Tai (2001).

In a recent contribution Mariani et al. (2010) consider the role of environmental quality for life expectancy. Their model is based on previous work by John and Pecchenino (1994) but accounts for the empirically observed positive correlation between life expectancy and environmental quality. More specifically they assume an overlapping generation framework where individuals live for three periods. While survival into the first and second period is quite certain, the survival into the third period is uncertain and depends on environmental quality. During their second period of life, individuals choose consumption and environmental maintenance (given an exogenous wage rate) such as to maximize their utility – which positively depends on second-period consumption and third-period environmental quality – and taking into account the effect of consumption and environmental maintenance on the dynamics of environmental quality. The model is capable of capturing the stylized facts of the positive correlation between survival and environmental quality. In addition, the authors show that multiple equilibria are possible that represent environmental poverty traps with low life expectancy and low environmental quality as opposed to high-level equilibria with high life expectancy and high environmental quality. The results are robust toward the extension of endogenous human capital formation that spurs growth and also life expectancy.

Another recent contribution that endogenizes the role of pollution on mortality is Lehmijoki and Rovenskaya (2010). Based on a long-term consumer optimization model and the assumption that pollution as a by-product of production negatively affects survival, the authors study the role of economic growth for reduced pollution and lower mortality.

Concluding Assessment of the Literature

Although there is a growing awareness of the population/environment linkage, as evidenced by the increasing number of empirical (case) studies, research has only slowly progressed in the Malthusian spirit of modeling the complex interaction in simple mathematical cartoons. Often dynamic interactions between population and environment are neglected – or increasingly complex simulation models are set up where anything can happen.

Further research on this topic is needed. The concern should be shifted away from physical limits to growth toward the ability of social units to respond to environment/population linkages. Although they are more difficult to formalize, such models should also include organizational and institutional issues.

See also: Boserup, Ester (1910–99); Carrying Capacity of the Environment; Demographic Impact of Disasters; Environment and Development; Environment and Health; Human–Environment Interactions: Case Studies; Ipat (Impact, Population, Affluence, and Technology); Kuznets Curves; Limits to Growth; Local Economic Development; Malthus, Thomas Robert (1766–1834); Sustainable Agriculture; Sustainable Development: An Economic Perspective.

Bibliography

Anderies, J., 1998. Culture and human agro-ecosystem dynamics: the Tsembaga of New Guinea. Journal of Theoretical Biology 192 (4), 515–530.
Anderies, J., 2000. On modeling human behavior and institutions in simple ecological economic systems. Ecological Economics 35, 393–412.
Boserup, E., 1981. Population and Technological Change: A Study of Long-term Trends. University of Chicago Press, Chicago, IL.
Brander, J.A., Taylor, M.S., 1998. The simple economics of Easter Island: a Ricardo–Malthus model of renewable resource use. The American Economic Review 88 (1), 119–138.
Chu, C.Y.C., 1998. Population Dynamics: A New Economic Approach. Oxford University Press, New York.
Chu, C.Y.X., Tai, C., 2001. Ecosystem resilience, specialized adaptation and population decline: a modern Malthusian theory. Journal of Population Economics 14, 7–19.
Cohen, J.E., 1995. Population growth and earth's human carrying capacity. Science 269, 341–346.
Cronshaw, M.B., Requate, T., 1997. Population size and environmental quality. Journal of Population Economics 10 (3), 299–316.
de la Croix, D., Dottori, D., 2008. Easter Island's collapse: a tale of a population race. Journal of Economic Growth 13, 27–55.
D'Alessandro, S., 2007. Non-linear dynamics of population and natural resources: the emergence of different patterns of development. Ecological Economics 62, 473–481.
Dalton, M., O'Neill, B.C., Prskawetz, A., Jiang, L., Pitkin, J., 2008. Population ageing and future carbon emissions in the United States. Energy Economics 30, 642–675.
Dalton, T.R., Coats, R.M., 2000. Could institutional reform have saved Easter Island? Journal of Evolutionary Economics 10, 489–505.
Dalton, T.R., Coasts, R.M., Asrabadi, B.R., 2005. Renewable resources, property-rights regimes and endogenous growth. Ecological Economics 52, 31–41.
Eckstein, Z., Stern, S., Wolpin, K.I., 1988. Fertility choice, land, and the Malthusian hypothesis. International Economic Review 29 (2), 353–361.
Ehrlich, P.R., Holdren, J.P., 1971. Impact of population growth. Science 171, 1212–1217.
Fröling, M., 2011. Energy use, population and growth, 1800–1970. Journal of Population Economics 24, 1133–1163.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Harford, J.D., 1997. Stock pollution, child-bearing externalities, and the social discount rate. Journal of Environmental Economics and Management 33, 94–105.
Harford, J.D., 1998. The ultimate externality. American Economic Review 88 (1), 260–265.
John, A., Pecchenino, R., 1994. An overlapping generations model of growth and the environment. The Economic Journal 104 (427), 1393–1410.
John, A., Pecchenino, R., Schimmelpfenning, D., Schreft, S., 1995. Short-lived agents and the long-lived environment. Journal of Public Economics 58, 127–141.
Lee, R.D., 1986. Malthus and Boserup: a dynamic synthesis. In: Coleman, D., Schofield, R. (Eds.), The State of Population Theory: Forward from Malthus. Basil Blackwell, Oxford, pp. 96–130.
Lee, R.D., 1990. Comment: the second tragedy of the commons. Population and Development Review 16 (Suppl), 315–322.
Lehmijoki, U., Palokangas, T., 2010. Trade, population growth, and the environment in developing countries. Journal of Population Economics 23 (4), 1351–1370.
Lehmijoki, U., Rovenskaya, E., 2010. Environmental mortality and long-run growth. In: Crespo Cuaresma, J., Palokangas, T., Tarasyev, A. (Eds.), Dynamic Systems, Economic Growth, and the Environment, vol. 12. Springer, Berlin, Heidelberg, pp. 239–258.
MacKellar, F.L., Lutz, W., Prinz, C., Goujon, A., 1995. Population, households, and CO_2 emissions. Population and Development Review 21 (4), 849–865.
Mariani, F., Pérez-Barahona, A., Raffin, N., 2010. Life expectancy and the environment. Journal of Economic Dynamics and Control 34, 798–815.
Maxwell, J.W., Reuveny, R., 2000. An inquiry into renewable resource scarcity and conflict in developing countries. Journal of Peace Research 37 (3), 301–322.
Meadows, D.H., Meadows, D.L., Randers, J., Behrens, W.W., 1972. The Limits to Growth. Universe Books, New York.
Milik, A., Prskawetz, A., Feichtinger, G., Sanderson, W.C., 1996. Slow–fast dynamics in Wonderland. Environmental Modeling and Assessment 1, 3–17.
Nagase, Y., Uehara, T., 2011. Evolution of population-resource dynamics models. Ecological Economics 72, 9–17.
Nerlove, M., 1991. Population and the environment: a parable of firewood and other tales. American Journal of Agricultural Economics 73, 1334–1347.
Pezzey, J., Anderies, J.M., 2003. The effect of subsistence on collapse and institutional adaptation in population-resource societies. Journal of Development Economics 72, 299–320.
Preston, S.H., 1996. The effect of population growth on environmental quality. Population Research and Policy Review 15, 95–108.
Prskawetz, A., Feichtinger, G., Wirl, F., 1994. Endogenous population growth and the exploitation of renewable resources. Mathematical Population Studies 5 (1), 87–106.
Prskawetz, A., Gragnani, A., Feichtinger, G., 2003. Reconsidering the dynamic interaction of renewable resources and population growth: a focus on long-run sustainability. Environmental Modeling and Assessment 8, 35–45.
Repetto, R., Holmes, T., 1983. The role of population in resource depletion in developing countries. Population and Development Review 9 (4), 609–632.
Reuveny, R., Decker, C.S., 2000. Easter Island: historical anecdote or warning for the future? Ecological Economics 35, 271–287.
Sanderson, W.C., 1994. Simulation models of demographic, economic, and environmental interactions. In: Lutz, W. (Ed.), Population–Development–Environment: Understanding Their Interactions in Mauritius. Springer, Berlin, Heidelberg, pp. 33–71.
Schou, P., 2002. Pollution externalities in a model of endogenous fertility and growth. International Tax and Public Finance 9, 709–725.
Wexler, L., 1996. Decomposing Models of Demographic Impact on the Environment. Working Paper. IIASA, Laxenburg, Austria. WP-96–85.
York, R., Rosa, E.A., Dietz, T., 2003. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts. Ecological Economics 46 (3), 351–365.

Population Dynamics: Momentum of Population Growth

Thomas J Espenshade and Jonathan BC Tannen, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Y.J. Kim, volume 17, pp. 11766–11769, © 2001, Elsevier Ltd.

Abstract

This article summarizes the recent literature on population momentum, usually defined as the subsequent population growth or decline after replacement conditions have been achieved. We review total momentum, its constituent parts (nonstable and stable momentum), trends over time, and how momentum is affected by immigration and emigration. Numerical examples are provided. The unifying concept is that momentum is created by a deviation between a population's observed age distribution and its underlying stationary (or sometimes stable) age distribution, especially at younger ages.

Introduction

It is natural to assume that, in a closed population, the rate of population growth is determined by the level of fertility and mortality. After all, the observed growth rate at time t, $r(t)$, is simply the difference between the crude birth rate, $b(t)$, and the crude death rate, $d(t)$, or

$$r(t) = b(t) - d(t)$$

It is true that in a long run stable equilibrium population the intrinsic growth rate, r, depends only on the fertility schedule, $m(a)$ – the annual rate of bearing daughters for women at exact age a; and on the survival function, $p(a)$ – the probability of surviving from birth to exact age a, given the underlying force of mortality schedule, $\mu(a)$ – the annual rate of dying for women at exact age a. In the short run, however, the shape of the current age distribution also matters.

If the proportion of women in the female population at age a and time t is given by $c(a, t)$, then

$$b(t) = \int_0^\omega c(a,t)m(a)\mathrm{d}a$$

where ω is the oldest age attained, and

$$d(t) = \int_0^\omega c(a,t)\mu(a)\mathrm{d}a$$

so that

$$\begin{aligned} r(t) &= \int_0^\omega c(a,t)m(a)\mathrm{d}a - \int_0^\omega c(a,t)\mu(a)\mathrm{d}a \\ &= \int_0^\omega c(a,t)[m(a) - \mu(a)]\mathrm{d}a \qquad [1] \end{aligned}$$

The term in square brackets in eqn [1] is negative below the age when childbearing begins (α) and above the age when it ends (β), and is usually positive in between. This means that all other things being equal, the so-called middle-heavy age distributions that have relatively large numbers of women at or near the peak ages of childbearing tend to maximize population growth rates. Even when a schedule of fertility at replacement, $m_0(a)$, is substituted for $m(a)$ in eqn [1], the rate of population growth at time t will not necessarily be zero, unless $c(a, t)$ is the age distribution of the stationary population.

Here is where the idea of population momentum enters. Equation [1], with $m_0(a)$ substituted for $m(a)$, shows the contribution of age composition to population growth in a given year t when fertility is at replacement. Population momentum measures the cumulative contribution of age composition to population growth (or decline) in all future years after fertility has been set at replacement.

Total Population Momentum

An Empirical Approach

A population typically does not stop growing or declining the instant its fertility reaches replacement level. Instead, in a closed population, growth or decline slows gradually until a zero-growth stationary population is reached, in much the same way that a car gradually comes to a complete stop after the driver's foot is lifted from the accelerator (Schoen and Kim, 1991). The amount of momentum contained in a population's age distribution is usually measured in a relative sense by the ratio of the size of the long-run stationary population to that of the observed population when replacement fertility is first attained. Momentum coefficients in AD 1935 ranged from 1.08 in Austria, Belgium, and France to 1.60 in Puerto Rico and Honduras (Vincent, 1945). In AD 2005, momentum coefficients among United Nations countries fell between a low of 0.81 for Bulgaria and a high of 1.76 in Oman (Espenshade et al., 2011). Coefficients of momentum can be greater than 1.0 (indicating positive momentum) or less than 1.0 (when momentum is negative). Momentum can play a large or a small role in population dynamics (Bongaarts, 1994, 2007).

It may seem paradoxical that population growth (or decline) does not necessarily cease the moment replacement fertility is reached. Consider two examples. Suppose, first, the case of a rapidly growing population in which the total fertility rate (TFR) has been above five for the past several generations. In this situation, young daughters are numerous relative to the number of mothers, and mothers are in plentiful supply relative to the number of grandmothers. Now suppose that fertility is suddenly reduced to replacement and the net reproduction rate (NRR) is 1.0. Going forward, girls in the population who have

not yet entered childbearing will have only enough daughters to replace themselves, so that eventually there will be roughly equal numbers of young daughters and mothers (allowing for mortality). But the population of grandmothers swells in the short run as their relatively more numerous daughters age into their later years (Preston, 1986). The total population continues to grow for a while after fertility has been lowered, and the growth is largely concentrated at the older ages.

Alternatively, consider a population that has experienced a sharp fertility decline to below replacement levels, or one where, hypothetically, fertility has been below replacement for a very long time. Children in this population are likely to be relatively few compared to the number of mothers. If fertility is raised instantaneously to replacement, then going forward these children will replace themselves as they move through their childbearing years. But when they age into their later years, they will succeed larger cohorts of their mothers. The ensuing population decline will be largely concentrated at the older ages.

These two situations are illustrated in Table 1, which contains data for the female populations of Nigeria and Bulgaria. Nigeria's female population in the mid 2010s totaled 78.2 million. Its TFR for the previous 5 years was 5.61; and this high level of fertility had been sustained for many years, with only small declines from above 6.0 in the past decade (United Nations, 2011). The population is approximately stable, as indicated by the similarity of the observed annual rate of population growth (2.56%) compared to the intrinsic growth rate (2.49%). Nigeria's coefficient of population momentum for AD 2010 was calculated by projecting the AD 2010 population forward 300 years by which time a stationary population had been achieved, using observed rates for AD 2005–10 and then setting fertility to replacement by normalizing the maternity schedule by the NRR. A stationary population with 114.0 million women implies a momentum coefficient of 114.0/78.2 = 1.457. The impact of age composition alone is sufficient to propel population growth, resulting in an eventual stationary population nearly 46% larger than the observed population in 2010. It is clear from Table 1 that this growth is not shared equally by all age groups, but is concentrated above about age 30.

Bulgaria, also shown in Table 1, represents a country with a dramatic fertility collapse. As recently as AD 1985–90, Bulgaria's TFR was 1.92, just below replacement, and it had been comfortably above replacement for most prior years dating back to AD 1950. But the TFR fell to 1.51 in AD 1990–95 and further to between 1.2 and 1.3 for the period AD 1995–2005. By AD 2005–10 it had rebounded slightly to 1.46 (United Nations, 2011). This fertility decline is reflected in Bulgaria's female age distribution for AD 2010. Just 12.9% of its population is under age 15 years, compared to 42.4% in Nigeria. The observed annual growth rate was −0.59% and the intrinsic rate was −1.36%. The momentum coefficient for Bulgaria shows that, even if fertility were raised immediately to replacement and remained fixed, the female population would nevertheless decline from 3.9 million in AD 2010 to 3.0 million – a decline of more than 22%. The overall momentum coefficient is 0.777, but the values for individual age groups indicate that, as anticipated, the shrinkage is confined to ages 20 years and above.

An Analytic Approach

The previous discussion has suggested that what determines total population momentum is the shape of a population's observed

Table 1 Population momentum values for Nigerian and Bulgarian females, 2010[a]

	Nigeria			*Bulgaria*		
Age	*Observed population*	*Stationary population*	*Stationary ÷ observed*	*Observed population*	*Stationary population*	*Stationary ÷ observed*
0–4	12 974.3	9479.8	0.731	181.2	194.4	1.073
5–9	10 913.5	9187.8	0.842	160.3	194.2	1.211
10–14	9276.4	8983.1	0.968	157.6	194.0	1.231
15–19	8031.5	8819.3	1.098	194.4	193.7	0.996
20–24	7142.1	8633.2	1.209	245.6	193.3	0.787
25–29	6219.1	8365.7	1.345	256.3	192.8	0.752
30–34	5158.9	8002.5	1.551	279.3	192.2	0.688
35–39	4016.9	7562.8	1.883	273.1	191.4	0.701
40–44	3152.4	7103.7	2.253	258.1	190.0	0.736
45–49	2688.9	6682.5	2.485	258.8	187.6	0.725
50–54	2303.9	6295.9	2.733	269.4	184.0	0.683
55–59	1925.6	5881.9	3.055	276.4	178.7	0.647
60–64	1543.5	5358.5	3.472	280.3	171.4	0.611
65–69	1205.2	4658.2	3.865	222.6	160.0	0.719
70–74	832.4	3751.8	4.507	197.8	140.6	0.711
75–79	495.1	2679.3	5.412	178.5	111.8	0.626
80–84	237.8	1591.4	6.692	115.5	75.1	0.650
85+	104.5	970.4	9.283	66.3	62.0	0.935
Total	78 222.0	114 008.0	1.457	3 871.5	3 007.1	0.777

[a]Female population in thousands.
United Nations, 2011. World Population Prospects: The 2010 Revision. Department of Economics and Social Affairs, Population Division. Retrieved from: http://esa.un.org/wpp/index.htm (on 22.05.13.) and authors' calculations.

age distribution in relation to the age distribution of the long-run stationary population. Young populations, especially those with high proportions below age 15 years, tend to have a substantial amount of built-in positive momentum, whereas older populations with relatively few children typically have latent population decline even after fertility is raised to replacement.

These notions are formalized in the following analytic expression for total population momentum (Espenshade et al., 2011).

$$\text{Total Momentum} = \int_0^\beta \frac{c(x)}{c_0(x)} \int_x^\beta p(a)m_0(a)\mathrm{d}a\mathrm{d}x/A_0 \qquad [2]$$

where $c(x)$ is the observed proportionate age distribution, A_0 is the mean age of childbearing in the long-run stationary population, and $c_0(x)$ and $m_0(a)$ are the stationary population counterparts to $c(x)$ and $m(a)$. The subscript 0 is used here to indicate replacement values in the stationary population where $r = 0$ by definition. An identical formula for total momentum appears in Preston and Guillot (1997: pp. 20–21), and it was anticipated in somewhat different form by Keyfitz (1985: pp. 155–157).

Equation [2] shows that relative deviations between a population's observed age distribution, $c(x)$, and its stationary population age distribution, $c_0(x)$, below the oldest age of childbearing, β, largely determine population momentum. Young populations (in relation to their stationary forms) will contain large amounts of positive momentum; whereas populations with relatively few young people in relation to the stationary population will have negative momentum. In addition, the second integral in eqn [2] is a weighting function. These weights sum to unity and are greatest before the onset of childbearing. Therefore, the deviations between $c(x)$ and $c_0(x)$ that matter most are those that occur at the youngest ages. Finally, it is apparent from eqn [2] that if the observed population is already stationary so that $c(x)/c_0(x) = 1.0$ everywhere, then the coefficient of momentum is identically 1.0 because all momentum has been wrung out of the age distribution (Espenshade et al., 2011; Preston and Guillot, 1997).

Disaggregating Population Momentum

Nonstable Momentum

Total population momentum, or the cumulative contribution of age composition to the change in future population size, has two principal parts. The first part involves the transition to a stable equilibrium population. Imagine an initial arbitrary female population with size P and proportionate age distribution $c(x)$. We know from eqn [1] that the observed rate of population growth is influenced by the shape of $c(x)$ along with patterns of fertility and mortality. Let the contribution of fertility and mortality to $r(t)$ be captured by the intrinsic growth rate, r. Then the residual $r(t) - r$ measures the contribution of age composition at time t. If fertility and mortality are held constant, $r(t)$ will converge to r. But in the meantime, the population grows or declines due to age composition alone at rate $r(t) - r$. By the time $r(t) = r$ so that $r(t) - r = 0$, the population will have reached size Q due solely to the effects of age composition.

Q will be a stable population. In fact, it is called the stable equivalent population, because if populations P and Q are projected forward using fertility and mortality observed at time $t = 0$, they will eventually become indistinguishable in terms of size, age composition, and rate of growth (or decline). Another way to see how Q is determined is to imagine projecting population P forward using constant fertility and mortality until it becomes a stable population, and then reverse projecting the resulting stable population back to time $t = 0$ using the intrinsic growth rate, r.

The first part of total population momentum, then, is the ratio Q/P. It is called 'nonstable' momentum. It is the transient or ephemeral part of total momentum, because it becomes smaller and smaller (and eventually disappears) as the initial population P gets closer to a stable population. If P were stable to begin with, then $P = Q$, and there would be no nonstable momentum.

Formally, nonstable momentum can be written as

$$\begin{aligned}\text{Nonstable Momentum} &= \frac{Q}{P} \\ &= \int_0^\beta \frac{c(x)}{c_r(x)} \int_x^\beta e^{-ra} p(a)m(a)\mathrm{d}a\mathrm{d}x/A_r\end{aligned} \qquad [3]$$

where $c_r(x)$ is the stable age distribution corresponding to intrinsic growth rate r, and A_r is the mean age of childbearing in the stable population (Espenshade et al., 2011). Nonstable momentum is a weighted average of relative deviations between the observed age distribution and the implied stable age distribution. Because the weights (in the second integral) are nonincreasing, the deviations that matter most are those in the early years of life before childbearing begins. The importance of nonstable momentum has been recognized by Bourgeois-Pichat (1971), Espenshade and Campbell (1977), Feeney (2003), and Schoen and Kim (1991).

Stable Momentum

Instead of starting with an arbitrary female population at time $t = 0$, suppose we begin with a stable population (for example, population Q). Now find momentum in the usual way by setting fertility at replacement, holding fertility and mortality constant, and projecting the population until it becomes stationary with eventual constant size S_2. Then stable momentum is the ratio S_2/Q, or,

$$\begin{aligned}\text{Stable Momentum} &= \frac{S_2}{Q} \\ &= \int_0^\beta \frac{c_r(x)}{c_0(x)} \int_x^\beta p(a)m_0(a)\mathrm{d}a\mathrm{d}x/A_0\end{aligned} \qquad [4]$$

It is called 'stable' momentum because, just like other stable population concepts, it depends only on fertility and mortality schedules and not on the current age distribution (Espenshade et al., 2011). The value for stable momentum will depend on relative deviations between the stable and stationary age distributions, with the youngest ages carrying the most weight.

Keyfitz (1971) popularized the concept of population momentum in a stable population. He showed that, starting with a stable population, total momentum is the ratio

$$\frac{S_2}{Q} = \frac{b\mathring{e}_0}{rA_0R_0}(R_0 - 1) \qquad [5]$$

where b is the stable birth rate, R_0 is the initial NRR, and $\mathring{e}_0$ is life expectancy at birth. It is straightforward to see that eqns [4] and [5] are identical. Alternative formulations for momentum in initially stable populations have been proposed by Goldstein and Stecklov (2002), Kim and Schoen (1993, 1997), and Schoen and Jonsson (2003). Some of this work assumes that fertility moves gradually (and not instantaneously) to replacement.

Fitting the Pieces Together

Return to the initial, arbitrary population P and recall how total momentum is calculated. Put fertility at replacement, hold fertility and mortality constant, and project P until it becomes stationary with size S_1. Total momentum is then S_1/P. The identity

$$S_1/P \equiv Q/P \times S_2/Q \times S_1/S_2$$

tells us that total momentum can be factored into the product

$$\text{Total Momentum} \equiv \text{Nonstable Momentum} \times \text{Stable Momentum} \times S_1/S_2 \qquad [6]$$

where S_1/S_2 is an offset factor. But if $S_1 = S_2$ we would have achieved an exact factorization of total momentum into the product of nonstable and stable momentum.

On the basis of an analysis of 176 UN countries, Espenshade et al. (2011) concluded that S_1 and S_2 are almost identical. In roughly 40% of the cases, the deviation was less than ±0.1%. In two-thirds of all cases, the deviations were contained within ±0.2%. And in three-fourths of all cases, the relative difference between S_1 and S_2 lay within ±0.3%. One implication is that the product of nonstable and stable momentum is an extremely good approximation to total momentum. More important, the identity in eqn [6] provides a unifying framework within which the types of momentum that various authors have considered can be situated.

There are several special cases in which S_1 is identically equal to S_2 – if the initial population is already stable or if current fertility is already at replacement. Finally, there is the degenerate case where fertility is at replacement and the population is stationary. In this instance, there is no momentum of any kind: neither total, nor nonstable, nor stable (Espenshade et al., 2011).

There are distinct regional patterns for total momentum and its constituent parts. Total momentum in AD 2005 for more-developed countries (MDCs) was typically near 1.0 and frequently below it, whereas total momentum for the least-developed countries (LDCs) clustered around 1.5. Stable momentum was usually below 1.0 and often near 0.5 for MDCs, but close to 1.5 for LDCs. Nonstable momentum, on the hand, was near 1.5 for MDCs but closer to 1.0 for LDCs. Other developing countries had momentum values intermediate between those for MDCs and LDCs. This analysis shows that as the level of development increases, overall momentum values fall, nonstable momentum becomes stronger, and stable momentum grows weaker (Espenshade et al., 2011).

Population Momentum Over Time

Population momentum is usually considered as a static concept, calculated for a given point in time. Blue and Espenshade (2011) describe the arc of momentum over time in 16 populations, selected to represent different stages of the demographic transition. These stages are divided into five categories: (1) pretransition, when birth and dates rates are uniformly high, (2) early transition, with falling mortality and still relatively high fertility, (3) mid-transition, characterized by a robust rate of natural increase and the maximum difference between birth and death rates, (4) late transition, when falling fertility approaches low mortality, and (5) the second demographic transition, with low birth and death rates, and subreplacement fertility.

It is difficult to identify today any national population that falls into Stage 1 of the demographic transition, because no country remains untouched by improvements to mortality. There are, however, numerous examples of countries in each of the four later stages. Blue and Espenshade (2011) rely for their examples on Ethiopia, Haiti, Kenya, and Nigeria for Stage 2, on Egypt, India, and South Africa for Stage 3, on Argentina, Brazil, Indonesia, and Turkey for Stage 4, and on Belgium, Finland, France, Japan, and Sweden for Stage 5.

Measures of total population momentum, calculated at 5-year intervals starting as early as reliable data permit and continuing with projected data to AD 2050, support the following conclusions. Especially for the examples from developing countries, total momentum tends to rise at the beginning of the period of observation as mortality decline (with nearly constant fertility) creates a younger observed population and an older stationary population, thereby widening relative deviations between $c(x)$ and $c_0(x)$. Later, as fertility falls, total momentum also falls. Toward the end of the transition, if fertility appears to stabilize at below-replacement levels, then total momentum values stabilize below 1.0.

In particular, results for all 11 developing countries that comprise the examples for transition Stages 2, 3, and 4 are quite similar. In each case, total momentum tends to peak in an interval between 1.4 and 1.6. The main difference among the stages is in the timing of transition. The peak occurs about AD 1980 for countries in Stage 4, about AD 1995 for Stage 3 countries, and is projected to arrive about AD 2015 for countries now in the earliest stages of the demographic transition. For the five developed countries in Stage 5, the patterns are less uniform. The transition occurred more gradually, and therefore the rise and fall in total momentum also occurred more slowly. In addition, the peak momentum value for Stage 5 countries tends to be no higher than about 1.2. Japan appears to be a conspicuous outlier. Total population momentum was about 1.4 for Japan in AD 1950. But because of Japan's sudden fertility collapse, total momentum fell to roughly 0.8 by AD 2000 and is projected to decline further to close to 0.6 by AD 2045.

Blue and Espenshade (2011) also traced the evolution of stable and nonstable momentum in these 16 countries. If we imagine values for stable momentum graphed along a horizontal axis and those for nonstable momentum along a vertical axis, then demographic transition theory predicts that coordinates for stable and nonstable momentum will emanate from the origin (a value of 1.0 for both stable and nonstable momentum), form a counter-clockwise swirl pattern, and eventually return back to the origin (if fertility at the end of the demographic transition returns to replacement level) or to a point on the horizontal axis to the left of the origin if end-of-transition fertility stabilizes below replacement. Developing countries may exhibit larger swirls than those in developed countries because the former typically pass through the demographic transition more quickly.

The evidence we have supports the counter-clockwise swirl pattern for developing countries (especially those not hit hard by HIV), because these countries have largely avoided major fertility reversals. For example, we pick up the data for India in AD 1950 when stable momentum is 1.24 and nonstable momentum is 0.96. The curve then heads off to the Northeast until AD 1975, when total momentum reaches 1.5, stable momentum equals 1.43, and nonstable momentum is 1.05. From that point, the curve changes direction heading in a Northwesterly direction until AD 2030 when stable momentum will have fallen back to 0.82 and nonstable momentum is at its maximum (for India) of 1.44. Beyond that point, the momentum swirl heads southward. There is little further change in stable momentum, though nonstable momentum declines to 1.21 by AD 2050. Developed countries tend to follow this pattern, too, but only up to about AD 1950. The ensuing fertility swings from the baby boom inject some jagged fluctuations into these patterns along a vector running from the Northwest to the Southeast.

Population Momentum Incorporating Migration

This section draws on recent work by Espenshade and Tannen (2013). Discussions of population momentum typically assume a closed population. When these models are extended to include immigration and emigration, however, new possibilities emerge. In this expanded context, a given population can be associated with a range of momentum values depending on how the task of reaching replacement conditions is shared between fertility change and immigration change.

In the expanded framework, the survival function is no longer the survival function, $p(a)$, described above. Instead, one can exit a population by dying or by emigrating, and one can enter by immigrating above age 0. To accommodate these possibilities define a migration-inclusive survival function $\pi(a)$ as

$$\pi(a) = e^{-\int_0^a (\mu(x)+\nu(x)-\iota(x))\mathrm{d}x} \qquad [7]$$

with mortality rates $\mu(x)$, in-migration rates $\iota(x)$, and out-migration rates $\nu(x)$. Notice that, unlike $p(a)$, $\pi(a)$ can increase over some age intervals if the immigration rate exceeds the sum of mortality and out-migration rates, and it can even exceed 1.0. But eventually $\pi(a)$ must return to zero. If we define maternity rates $m(a)$ as before, then a population of size P with proportionate age structure $c(x)$ has an intrinsic growth r^* that satisfies

$$\int_0^\beta e^{-r^* a}\pi(a)m(a)\mathrm{d}a = 1 \qquad [8]$$

In closed-population scenarios, replacement fertility is generally achieved by dividing the maternity schedule by the NRR, effecting the same proportionate fertility change (higher or lower) at all ages. Here, if replacement conditions are attained entirely through adjustments to fertility, suppose fertility is modified according to

$$m_0(a) = e^{-r^* a}m(a)$$

Alternatively, suppose fertility is left unchanged and the immigration rate schedule is moved up or down by a constant amount r^* at all ages to achieve replacement. In particular, let

$$\begin{aligned}\pi_0(a) &= e^{-r^* a}\pi(a)\\ &= e^{-\int_0^a (\mu(x)+\nu(x)-[\iota(x)-r^*])\mathrm{d}x}\end{aligned}$$

These two extreme cases can be combined into a set of fertility and migration-inclusive survival functions that together achieve replacement. There are, in fact, many such combinations that jointly satisfy replacement conditions. Define new replacement-level fertility and migration-inclusive survival functions as

$$m_0(a|\gamma) = e^{-(1-\gamma)r^* a}m(a) \qquad [9]$$

$$\pi_0(a|\gamma) = e^{-\gamma r^* a}\pi(a) \qquad [10]$$

where γ is a parameter that ranges between 0 and 1 and represents the proportion of the overall rate change absorbed by immigration. If $\gamma = 0$, then all of the work of reaching replacement is being done by fertility. If $\gamma = 1$, then all of the adjustment to replacement falls on immigration. When these two new schedules are used in tandem, they imply a migration-inclusive intrinsic growth rate r^* equal to zero.

In the context of open populations that incorporate migration, two aspects of the long-run stationary population need to be examined: the shape of its age distribution and its size. Both will depend on the particular value assigned to γ. The age distribution is

$$\begin{aligned}c_0(a|\gamma) &= b_0(\gamma)\pi_0(a|\gamma)\\ &= b_0(\gamma)e^{-\gamma r^* a}\pi(a)\end{aligned} \qquad [11]$$

where $b_0(\gamma)$ is the birthrate in the migration-inclusive stationary population and equal to $1/\int_0^\omega \pi_0(a|\gamma)\mathrm{d}a$.

How the age distribution responds to changes in γ will depend on the value of r^*. The age distribution becomes younger as γ varies from 0 to 1 whenever $r^* > 0$. For example, Iceland in AD 2008 had a value for $r^* = 0.015$. As γ increases from 0 to 1, the share of Iceland's migration-inclusive stationary population that is under age 15 years rises from 11.7 to 20.2%, while the share over age 75 years falls from 16.2 to 8.7%. On the other hand, if $r^* < 0$, then the stationary population becomes older with increases in γ. Germany had a negative r^*-value in AD 2009 of -0.011. As γ ranges from 0 to 1, the share of Germany's stationary

population over age 75 years rises from 13.5 to 20.9%, while the percentage under age 15 years falls from 16.9 to 10.6. In the special case where $r^* = 0$, the shape of the stationary age distribution is invariant with respect to changes in γ.

Extending work by Preston and Guillot (1997) and Espenshade et al. (2011), we may define migration-inclusive momentum for a population with size P and age structure $c(x)$ as

$$M(\gamma) = \frac{S(\gamma)}{P} = \int_0^\beta \frac{c(x)}{c_0(x|\gamma)} \int_x^\beta e^{-r^* a} \pi(a) m(a) \mathrm{d}a \mathrm{d}x / A_0 \quad [12]$$

where $S(\gamma)$ is the size of an eventual stationary population that is achieved after converting fertility and immigration instantaneously to γ-dependent replacement, and then holding fertility, mortality, and both in-migration and out-migration rates constant until an equilibrium stationary population is reached. A_0 is the mean age of childbearing in the long-run stationary population. Equation [12] indicates that γ influences total momentum only through the shape of the stationary population age distribution. The form of eqn [12] is much like eqns [2]–[4]. Total momentum in the migration-inclusive case is once again a weighted average of age-specific relative deviations between the observed age distribution and the long-run migration-inclusive stationary population age distribution, with weights that are largest below the beginning of childbearing and decline toward zero with advancing age.

Equation [12] also shows that, unlike the closed-population example, a given open population will typically have many different coefficients of momentum associated with it, depending upon the particular value assigned to γ. Moreover, when $r^* > 0$, $c_0(x|\gamma)$ becomes younger as γ increases, and, consequently, total momentum falls. To take the case again of Iceland in AD 2008 with an r^*-value of 0.015, total momentum declines from 1.66 to 1.08 as γ increases from 0 to 1. On the other hand, for negative values of r^*, total momentum rises as γ increases. In the case of Germany in AD 2009, for instance, migration-inclusive momentum goes up from 0.87 to 1.25 when γ ranges from 0 to 1. Finally, if $r^* = 0$, then the stationary population age distribution is invariant with respect to changes in γ, and total momentum will not change.

Espenshade and Tannen (2013) analyzed data for all European Union countries having complete fertility, mortality, immigration, and emigration data. Two important policy lessons can be drawn from their analysis. First, depending on how replacement conditions are achieved and, therefore, on the value of γ that is chosen, in roughly two out of every five of their 58 country-year observations, the long-run stationary population size can be larger or smaller than the current population size. There are 34 instances in which population momentum was consistently positive, that is, in which momentum coefficients were always greater than 1.0. In some cases, the variation in momentum was relatively small, as in the case of Denmark in AD 2010 (1.055, 1.059) or the Netherlands in AD 2009 (1.087, 1.105). But in others, as γ ranges between 0 and 1, there can be wide swings in momentum, as in Iceland AD 2008 noted above (1.082, 1.663) or in Luxembourg AD 2008 (1.096, 1.816). However, there are 24 country-year observations – 41% of the total – for which momentum can be positive or negative depending on the choice of γ. In Lithuania in AD 2010, for instance, migration-inclusive momentum is 0.57 when $\gamma = 0$, but 1.39 when $\gamma = 1$. With such wide variation, policymakers have considerable leeway when developing judgments about future population size and age composition.

Second, migration-inclusive population momentum is often grossly underestimated by total momentum calculated for a closed population, at least for these European Union countries. In almost 60% (34 out of 58 cases) of the country-years for which the Espenshade and Tannen (2013) study provides data, the value for closed-population momentum falls short of the open-population momentum range, and it does so by a substantial margin two-thirds of the time. Moreover, in slightly more than half of these cases (18 out of 34), both high and low open-population estimates suggest that momentum is positive, whereas the closed-population estimate points in a negative direction. Therefore, in 31% of all cases, both the magnitude and the direction of the closed-population momentum estimate give misleading information.

Future Research

Several new research directions are suggested by the current state of our knowledge. First, typically in closed populations, momentum estimates are produced by reducing fertility immediately to replacement by dividing the maternity schedule by the NRR. Mitra (1976) proposed moving to replacement by an abrupt, but exponential, fertility decline. Other authors have examined the effects of a gradual shift to replacement fertility (Frauenthal, 1975; Goldstein, 2002; Goldstein and Stecklov, 2002; Li and Tuljapurkar, 1999, 2000; O'Neill et al., 1999; Schoen, 2005; Schoen and Jonsson, 2003; Schoen and Kim, 1998). It may be useful to investigate gradual transitions to replacement in fertility–migration combinations.

Second, Rogers (1975) outlined a framework for thinking about population momentum in a multiregional setting. Rogers (1995), Rogers and Willekens (1978), and Schoen (2002, 2006) applied this framework to investigate momentum in a two-region, urban–rural context. Typically in this work, the fertility schedule bears the full responsibility for getting to replacement. It is worth extending the analysis to let fertility and migration rates share this responsibility in some proportion. In addition, Raymer et al. (2012) propose three alternatives for incorporating flows of international migrants into multiregional models. One alternative that they do not consider, but that may be worth examining, is simply to expand an N-dimensional multiregional system to $N + 1$ dimensions to incorporate the 'rest of the world'.

Third, analyses of stable and nonstable momentum have so far been limited to single-region, closed populations. It would be productive to study the behavior of these two constituent parts of total momentum in a multiregional setting.

Acknowledgments

Partial support for this research was provided by a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Grant #5R24HD047879) and from the National

Institutes of Health (Grant #5T32HD-007163). We are grateful to Laura Blue for helpful comments and suggestions.

See also: Fertility Change: Quantum and Tempo; Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems; Population Dynamics: Theory of Nonstable Populations; Population Dynamics: Theory of Stable Populations; Second Demographic Transition.

Bibliography

Blue, L., Espenshade, T.J., 2011. Population momentum across the demographic transition. Population and Development Review 37 (4), 721–747.

Bongaarts, J., 1994. Population policy options in the developing world. Science 263, 771–776.

Bongaarts, J., 2007. Population Growth and Policy Options in the Developing World. The Population Council, New York (Paper presented at the Beijing Forum, Beijing, November 1–4.).

Bourgeois-Pichat, J., 1971. Stable, semi-stable populations and growth potential. Population Studies 25 (2), 235–254.

Espenshade, T.J., Campbell, G., 1977. The stable equivalent population, age composition, and Fisher's reproductive value function. Demography 14 (1), 77–86.

Espenshade, T.J., Olgiati, A.S., Levin, S.A., 2011. On nonstable and stable population momentum. Demography 48 (4), 1581–1599.

Espenshade, T.J., Tannen, J.B.C., 2013. Population Momentum with Migration. Unpublished manuscript. Office of Population Research, Princeton University, Princeton, NJ.

Feeney, G., 2003. Momentum of population growth. In: Demeny, P., McNicoll, G. (Eds.), Encyclopedia of Population, vol. 2. Macmillan Reference USA, New York, pp. 646–649.

Frauenthal, J., 1975. Birth trajectory under changing fertility conditions. Demography 12 (3), 447–454.

Goldstein, J.R., 2002. Population momentum for gradual demographic transitions: an alternative approach. Demography 39 (1), 65–73.

Goldstein, J.R., Stecklov, G., 2002. Long-range population projections made simple. Population and Development Review 28 (1), 121–141.

Keyfitz, N., 1971. On the momentum of population growth. Demography 8 (1), 71–80.

Keyfitz, N., 1985. Applied Mathematical Demography, second ed. Springer-Verlag, New York.

Kim, Y.J., Schoen, R., 1993. Crossovers that link populations with the same vital rates. Mathematical Population Studies 4, 1–19.

Kim, Y.J., Schoen, R., 1997. Population momentum expresses population aging. Demography 34 (3), 421–427.

Li, N., Tuljapurkar, S., 1999. Population momentum for gradual demographic transitions. Population Studies 53 (2), 255–262.

Li, N., Tuljapurkar, S., 2000. The solution of time-dependent population models. Mathematical Population Studies 7 (4), 311–329.

Mitra, S., 1976. Influence of instantaneous fertility decline to replacement level on population growth: an alternative model. Demography 13 (4), 513–519.

O'Neill, B.C., Scherbov, S., Lutz, W., 1999. The long-term effect of the timing of fertility decline on population size. Population and Development Review 25 (4), 749–756.

Preston, S.H., 1986. The relation between actual and intrinsic growth rates. Population Studies 40 (3), 343–351.

Preston, S.H., Guillot, M., 1997. Population dynamics in an age of declining fertility. Genus 53 (3–4), 15–31.

Raymer, J., Abel, G., Rogers, A., 2012. Does specification matter? Experiments with simple multiregional probabilistic population projections. Environment and Planning A 44, 2664–2686.

Rogers, A., 1975. Introduction to Multiregional Mathematical Demography. John Wiley & Sons, New York.

Rogers, A., 1995. Multiregional Demography: Principles, Methods and Extensions. John Wiley & Sons, New York.

Rogers, A., Willekens, F., 1978. The spatial reproductive value and the spatial momentum of zero population growth. Environment and Planning A 10 (5), 503–518.

Schoen, R., 2002. On the impact of spatial momentum. Demographic Research 6 (3), 49–66.

Schoen, R., 2005. Intrinsically dynamic population models. Demographic Research 12 (3), 51–76.

Schoen, R., 2006. Dynamic Population Models. Springer, Dordrecht, The Netherlands.

Schoen, R., Jonsson, S.H., 2003. Modeling momentum in gradual demographic transitions. Demography 40 (4), 621–635.

Schoen, R., Kim, Y.J., 1991. Movement toward stability as a fundamental principle of population dynamics. Demography 28 (3), 455–466.

Schoen, R., Kim, Y.J., 1998. Momentum under a gradual approach to zero growth. Population Studies 52 (3), 295–299.

United Nations, 2011. World Population Prospects: The 2010 Revision. Department of Economics and Social Affairs, Population Division. Retrieved from: http://esa.un.org/wpp/index.htm (accessed on 22.05.13.).

Vincent, P., 1945. Potentiel d'accroissement d'une population [Growth potential of a population]. Journal de la Société de Statistique de Paris, 1re Série, Janvier-Février 86 (1–2), 16–39.

Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems

Peter Jagers, Chalmers University of Technology, Gothenburg, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relation between individual reproduction and the probability that populations die out is given. Populations that decrease on average will, of course, always die out, but populations whose expected sizes grow can also have a high probability of extinction. Malthus's law of exponential growth of populations, i.e., not dying out, holds in general, not only for populations of independently reproducing individuals, but also under some types of interaction. The stable age distribution and general stable composition, appearing as a consequence of exponential growth, are described. Finally, populations whose size and composition may influence individual reproduction are described.

Extinction, Stable Size, or Explosion?

Historically, probabilistic population theories originate from the *extinction problem*: suppose you know the probabilities of begetting 0, 1, 2, 3, … children. Now determine the probability that a population of a known number of individuals will eventually die out. Already Malthus had noted that a rapid increase of the whole population could occur together with frequent extinction of the separate families constituting the population. Rephrased in the terminology of evolutionary biology, frequent extinction of species does not contradict a rapid growth of the biomass of the world.

The mathematical explanation of this apparent paradox is due to Bienaymé (1845) and independently, 30 years later and only in parts correctly, to Galton and Watson, who were later to lend their names to the simplest, generation-counting version of the so-called branching processes that provide the mathematical structure underlying most stochastic population dynamics. Indeed, if p_k denotes the probability of giving birth to k children, $k = 0, 1, 2, 3, \ldots$, then the extinction probability of a population stemming from one ancestor will be the smallest nonnegative solution q of the equation,

$$\sum_k p_k q^k = q.$$

If we disregard the trivial case $p_1 = 1$, q will equal 1 if and only if the mean number of children,

$$m = \sum_k p_k k \leq 1.$$

If $m > 1$, then the population growth rate per generation will be m, and even though q must be less than 1 it can still be pretty large: it is easy to formulate demographically or biologically plausible cases, where, say, m exceeds 2 and q is still around 0.8.

The preceding argument presumes that individuals reproduce independently of one another. In deterministic population theories dependence structure is not made precise, and stable population sizes can arise, cf Keyfitz 1968. The situation in stochastic models with interaction or dependence upon the population as a whole is therefore of great interest.

Population Growth

To describe population growth in real time rather than generation-wise, a more sophisticated description of individual life and reproduction is needed. Probabilistic population dynamics is *individual-based* in the sense that it starts from such a description. Deterministic theories, on the other hand, often start directly from an assumed behavior of the population as a whole. Demographic theory is often *pseudo-probabilistic*, in that, it uses probabilistic notions but makes conclusions only about the expected behavior of the population as a whole. (This is the reason why there exist deterministic populations with stationary sizes: if the underlying stochastic population is critical in the sense that on average each individual is replaced by one child, then the expected total population size will be stationary, whereas the population itself will die out sooner or later.)

The most general probabilistic formulation of individual life goes as follows: at birth an individual inherits a *type* from her mother. The type determines her survival distribution as well as the distribution of age at childbearing. It also determines the types of the children she will give birth to in her turn. (Think of the type as a genotype, or a genotype combined with the social and biological conditions of life.)

In the *supercritical* case, where there is a positive chance of nonextinction, the probability laws of survival and childbearing determine a positive *Malthusian* parameter that gives the rate of exponential growth. In the theory of branching processes this rate is usually denoted by α, so that the population will grow like $e^{\alpha t}$, as time t passes. In deterministic theory the conventional notation is often r instead.

Reproduction Influenced by the Population as a Whole, Extinction or Quasi-Stationarity

Now assume that individual reproduction can be influenced by the present state of the population as a whole, its size or its

composition or both, or only by special individuals like neighbors or close relatives.

If this is the case, the situation can be drastically different and it is easy to exhibit examples of populations with high growth rate ($m > 1$), which still are bound to extinction (Haccou et al., 2005). It can, however, be proved very generally that a stochastic population, which is not dying out, must grow beyond all limits (Jagers, 1991; see also Haccou et al., 2005). This is a consequence of stochasticity and the asymmetry between extinction and big sizes: however big the population, there remains an albeit small but still positive risk of extinction, whereas an extinct population cannot be resurrected.

Even though exponential growth appears to be a consequence of the independence assumed (tacitly or explicitly) between the reproduction of different individuals, it is interesting to note that such growth has been proved also to occur in many cases where individual reproduction depends upon the population as a whole, e.g., its size. For such results in the case of simple Galton–Watson branching processes, see Klebaner (1984). In cases where the reproduction approaches criticality as the population grows, nonMalthusian growth can occur nevertheless, e.g., in the form of linear growth (Klebaner, 1994).

But a finite world cannot sustain infinite growth. One can therefore argue that probabilistic population dynamics show that all real populations will die out sooner or later. But before that happens there may be a long period of growth, or even – as recent research has shown – long periods of quasi-stationarity.

Indeed, consider a population where individual reproduction at any time is influenced by total population size, and assume that there is a *carrying capacity* K, a numerical value above which the population is subcritical and cannot sustain itself, whereas it is supercritical below K. Then, if starting small, it may die out directly, or else in time span of order log K, for large K, it will reach a band around the carrying capacity, where it will then stay during a period whose length is exponential in K, until it finally dies out (Jagers and Klebaner, 2011).

Stable Population Theory

In contrast to population size, the composition of growing populations, where individuals reproduce independently, tends to stabilize. The description of this composition, or rather the various stable compositions pertaining to different reproduction and survival laws, is the object of stable population theory.

The most well-known aspect of stable population composition is the stable age distribution, first described by Euler in 1760 and since then rediscovered repeatedly: if the Malthusian parameter is $\alpha > 0$ and all individuals have the same survival function ℓ, then the age of an individual sampled at random in an old population will have the density

$$\frac{e^{-\alpha a}\ell(a)}{\int_0^\infty e^{-au}\ell(a)\mathrm{d}u}$$

at age a, exactly as in deterministic theory. This is, of course, a law-of-large-numbers effect; we are considering ever-increasing populations.

If, as in the general setup mentioned above, there are several types of individuals, each with its own life span distribution, then the situation will be more complex. First of all, the type distribution will converge to a stable type law, determined by an eigen-measure property. Then the survival function l, above will have to be replaced by an average of all the different survival functions weighted by the stable type distribution.

But the stable composition describes many other aspects of a balanced growing population besides ages. For example, the probability of being firstborn can be calculated as well as other aspects of family structure. In modern evolutionary biology the genealogy of individuals and, say, the mutational history play an important role. In an old population this is described by a stable, typical ancestry, which can be used to determine time back to some event like divergence between species. In a strict sense, such results have yet been proven only for populations of independently reproducing individuals. Little is known about stabilization during the quasi-stationary stage in population size dependence.

See also: Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Population Dynamics: Momentum of Population Growth; Population Dynamics: Theory of Nonstable Populations.

Bibliography

Athreya, K.B., Ney, P., 1972. Branching Processes. Springer-Verlag, Berlin.
Haccou, P., Jagers, P., Vatutin, V., 2005. Branching Processes: Variation, Growth, and Extinction of Populations. Cambridge University Press, Cambridge, MA.
Harris, T., 1963. The Theory of Branching Processes. Springer-Verlag, Berlin, Heidelberg.
Jagers, P., 1975. Branching Processes with Biological Applications. Wiley, Chichester.
Jagers, P., 1991. The growth and stabilization of populations. Statistical Science 6, 269–283.
Jagers, P., Klebaner, F.C., 2011. Population size dependent, age structured branching processes linger around their carrying capacity. Journal of Applied Probability 48A, 249–260.
Keyfitz, N., 1968. Introduction to the Mathematics of Population. Addison-Wesley, Reading, MA.
Klebaner, F.C., 1984. Geometric rate of growth in population size dependent branching processes. Journal of Applied Probability 21, 40–49.
Klebaner, F.C., 1994. Asymptotic behaviour of Markov population processes with asymptotically linear rate of change. Journal of Applied Probability 31, 614–625.

Population Dynamics: Theory of Nonstable Populations

Cyrus CY Chu, Academia Sinica, Taipei, Taiwan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article introduces the demographic theory of nonstable populations. The analysis of population dynamics was originally a branch of mathematical biology but is also important for social demography because it provides elegant models for how populations ebb and flow.

Background

To introduce the theory of nonstable population dynamics, we start by presenting a Malthusian model studied by Lee (1974). Consider an overlapping-generation framework in which each individual lives one or two periods. The first period is childhood and the second is adulthood, and all surviving adults will be in the labor force. The Malthusian model can be characterized by the following two equations: $W_t = f(N_t)$, $b_t = g(W_t)$, where W_t is the wage rate (at time t), N is the size of the adult group, b is the crude birth rate, the $f(.)$ function characterizes the wage/employment relationship in the labor market, and the $g(.)$ function specifies the influence of the labor market reward on fertility.

Suppose the number of children born in period t is B_t, and the child survival rate is l, then $N_t = l \cdot B_{t-1}$. The above two equations can be combined in the following reduced form: $B_t = h(B_{t-1})$, where $h(B_t) = g(f(l \cdot B_{t-1})) \cdot l \cdot B_{t-1}$. The above expression is a recursive equation of B_t, from which the dynamic pattern can be analyzed, and the possible existence of cyclicity can be studied. Here, the Malthusian model is interpreted as a model of density dependency, where the density is characterized by the birth size of the previous period. It turns out that the Malthusian model described above may not converge to a 'dismal' steady state, as Malthus suggested. Mathematically, the B_t series may converge, diverge, or be cyclic, and the volatility of the birth series B_t crucially hinges on the feedback elasticity of $h(.)$ with respect to B_{t-1}. When this elasticity is sufficiently large, the B_t series can easily have cycles. This is the classical case of nonstable population. Modern versions of nonstable population will be discussed below, after we introduce the general population structure.

The General Population Structure

Let $N_t \equiv (N_{t,\ 1}, \ldots, N_{t,\ n})$ be an n-type population vector at period t, where $N_{t,\ i}$, $i = 1, \ldots, n$ indicates the size of type-i population at period t. The type here may refer to any criterion that is used to classify the population, such as age, sex, and occupation. A general formulation of population dynamics can be written as follows:

$$N_{t+1} = Q(N_t)N_t \qquad [1]$$

where $Q(N_t)$ is the transition matrix. The simplest case is when $Q(N_t) = Q$, a time-independent matrix. In this situation, the above equation reduces to a time-invariant Markov chain: $N_{t+1} = QN_t$. It is a well-known implication of the Frobenius–Perron theorem that if Q is positively regular, meaning that there is flexible cross-period mobility across types, then the population vector N_t will converge to a steady state, in which all elements of the N_t vector grow at the same exponential rate, thereby the population composition is also time invariant. It has been shown by Chu (1998) that, for most population models with time-independent Q, this positive regularity condition is likely to be satisfied, therefore any nonstable behavior of the population would appear only when $Q(N_t)$ is not a constant matrix, or equivalently when the relationship between N_{t+1} and N_t in eqn [1] has some nonlinearity. Below we will introduce several special cases of the generic setting in eqn [1], and discuss their empirical relevance. The reader can see that each model corresponds to a special interpretation of the generic setting in eqn [1].

Age-Structured Cycles

For demographers, the most familiar version of the dynamic models in eqn [1] is the Lotka–Leslie age-specific models, where the type criterion is age. Let $N_{t,\ i}$ be the number of population aged i at period t, B_{t-i} the birth size at period $t - i$, and l_i probability that a person can survive to age i, then the following relationship must hold: $N_{t,\ i} = B_{t-i} \cdot l_i$. The transformed dynamics in terms of B_t is the well-known Lotka's equation:

$$B(t) = \sum_i B_{t-i} \cdot l_i \cdot m_{t,i} = \sum_i B_{t-i} \cdot \phi_{t,i} \qquad [2]$$

where $m_{t,\ i}$ is the average number of births per surviving member aged i at period t, and $\phi_{t,\ i} \equiv l_i \cdot m_{t,\ i}$ is the net maternity function. Evidently, the Malthusian model presented in the previous section is a special case of eqn [2] with two age types. As we mentioned, when the fertility function $m_{t,\ i}$ is independent of t, then the process of B_t follows a time-invariant Markov chain, and the usual convergence result applies. But if $m(t, i)$ is time dependent, then cycles or more volatile behavior may appear. One typical case of such time dependency is when the age-specific fertility rates are functions of previous birth sizes $m_{t,\ i} = m_i \cdot (B_t)$, where $B_t \equiv (B_{t-1}, B_{t-2}, \ldots)$.

Suppose the birth series corresponding to eqn [2] has a steady state. Taking a Taylor expansion around this steady state, Lee (1974) showed that the volatility of the birth series characterized by eqn [2] hinges upon the elasticities of the net maternity function with respect to B_t. When previous birth

sizes are larger, usually the density pressure is larger. Intuitively, larger feedback elasticities imply that present reproduction would be more sensitive to previous birth shocks; hence the birth series would be more volatile. Easterlin (1961) argued that the baby-boom/baby-bust cycles observed in the United States are typical examples of a feedback cycle.

How large the feedback elasticities are is an empirical question. In order to estimate such elasticities, two models have been proposed: The first is to assume that $\phi_{t,\ i}$ is only a function of the cohort size aged i: $\phi_{t,\ i} = \phi_i \cdot (B_{t-i})$ and the second is to assume that $\phi_{t,\ i}$ is a function of the weighted average of birth sizes of different ages $\phi_{t,\ i} = \phi_i \cdot (\sum_j w_j \cdot B_{t-j})$. The former is referred to as the cohort model, and the latter as the period model. However, empirical evidence has so far failed to back up either of these two models in producing persistent birth cycles that fit all characteristics (such as amplitude and period). Moreover, even if there does exist a cyclic solution to eqn [2], the information concerning feedback elasticity is not sufficient for us to tell whether the cyclical solution in question is stable. Related analysis is rather technical, and the reader is referred to Tuljapurkar (1987) for details.

Predator–Prey Cycles

The predator–prey model describes a two-type population structure, where the types are prey and predator, respectively, referred to as type-1 and type-2. The simplest formulation of the predator–prey model is characterized by the following difference equations:

$$\begin{aligned} \Delta N_{t+1,1} &\equiv N_{t+1,1} - N_{t,1} \\ &= aN_{t,1} - bN_{t,1}N_{t,2} \\ \Delta N_{t+1,2} &\equiv N_{t+1,2} - N_{t,2} \\ &= -cN_{t,2} + dN_{t,1}N_{t,2}, \quad a,b,c,d > 0 \end{aligned} \qquad [3]$$

Equation [3] is obviously a special case of eqn [1]. The prey is considered the only food resource available to the predator. Thus, if $N_{t,\ 1} = 0$, the predator population decreases exponentially at the rate C. With the nonexistence of predators, the prey population grows exponentially at the rate a.

In the steady state of eqn [3], $\Delta N_{t,\ 1} = \Delta N_{t,\ 2} = 0$, there is only one nontrivial solution: $(N_1^*, N_2^*) = (c/d, a/b)$. Let the Jacobian matrix of eqn [3] around the nontrivial solution be denoted J. It turns out that the J matrix can be written as

$$J = \begin{pmatrix} a - bN_2^* & -bN_1^* \\ dN_2^* & -c + dN_1^* \end{pmatrix} = \begin{pmatrix} 0 & -bc/d \\ ad/b & 0 \end{pmatrix}$$

Hence, we have $|J| = ac > 0$ and the trace of J is zero.

Thus, we know that the eigenvalues of J are purely imaginary, meaning that besides the equilibrium point $(c/d, a/b)$ itself, the solution trajectories of eqn [3] are cyclic and nonstable. But it was also pointed out that when the predator–prey model embodies the concerns of diminishing or increasing returns, then the solution becomes locally stable or divergent.

Occupation–Switch Cycles

Based on stylized facts found in Chinese history, Chu and Lee (1994) proposed an occupation-specific population model, in which the population is separated into three types: peasants, soldiers, and bandits. Peasants grow crops and pay taxes, soldiers (rulers) collect taxes and hunt for bandits, and bandits rob food and clash with peasants and soldiers. It is assumed that soldiers are drafted by the government, and all other civilians can choose to become peasants or bandits. Their occupation choice is made depending on which occupation is expected to generate a higher utility. When the bandit/soldier ratio is relatively high, it is likely that the incumbent power regime will be overthrown by the bandit group, and the dynasty switches. It turns out that this pattern of population dynamics characterizes the dynamic evolution of dynasties in Chinese history, and the cycles were called dynastic cycles by historians.

There is a unique feature associated with the above-mentioned occupation-switching cycles. For most animals, it is believed that when density pressure occurs, the environment becomes less favorable, and the reproduction rate of animals is reduced. One might expect that rational human decisions may be able to weaken the outside density pressure through institutional and rational regulations. But the scenario of dynastic cycles seems to be a counterexample.

Human beings' occupational choice between peasants and bandits has a 'demonstration effect,' which tends to magnify the originally weak density pressure and 'destablize' the dynamics of the population compositional structure.

Laborer–Capitalist Cycles

The cyclical pattern of population composition mentioned above is not the only case we find in human societies. Some researchers have considered a neoclassical economic growth model, and focused on the two economic classes in the economy: the laborers and the capitalists. Let u_t be the income share attributed to labor, and v_t be the employment rate of labor. Under some reasonable assumptions, Goodwin (1967) has shown that the dynamics of u_t and v_t mimic that of eqn [3]. Thus, the conclusion that the capitalists' economy appears to be 'permanently oscillating' was reached. But just as the original predator–prey model is sensitive to variations in institutional specifications, Goodwin's model after some minor modification also generates qualitatively different results. Definite evidence of the nonstability of human population is yet to be found.

Human Development Cycles

The final population dynamics we discuss is related to the economic development process of human beings. The Neolithic Revolution about 10 000 years ago triggered the human society from hunter-gatherer to agricultural. The Industrial Revolution 250 years ago marked the beginning of the modern phase. In terms of population dynamics, we observe that in the agricultural epoch the population was dominated by the Malthusian rule with a stable population size and a constant per-capita income, whereas after we entered the industrial revolution we first experienced a demographic transition and then a near-replacement level population growth.

The demographic transition changes parents' choices from quantity to quality of children, and the increased level of child education leads to the growth of per capita income. In short, we have a subsistence level family income and stagnant population size 300 years ago, and a growing population size (estimated 9.3 billion by 2050) with drastically different income levels across countries. What is behind this big regime change of dynamics?

It was the contribution of Galor (2011) to have explained the above-mentioned regime change in a unified model. Among others, there are two factors that dominated the change of population dynamics. The first is the binding subsistence level of consumption in the Malthusian period, when parents cannot spare their resources to increase children's quality. The second is the improved life expectancy and the change of market/household comparative advantage, which facilitate women to adopt a quantity–quality trade-off. Countries in the modern world are therefore divided into pre-transition, transition-leaders and transition-followers. The first category has a fast population growth but yet a low income level, and the latter two categories have a stable population size and a growing income level.

See also: Demographic Models; Population Dynamics: Classical Applications of Stable Population Theory; Population Dynamics: Theory of Stable Populations.

Bibliography

Chu, C.Y.C., 1998. Population Dynamics: A New Economic Approach. Oxford University Press, New York.

Chu, C.Y.C., Lee, R.D., 1994. Famine, revolt and the dynastic cycle: population dynamics in historic China. Journal of Population Economics 7, 351–378.

Easterlin, R.A., 1961. The American baby boom in historical perspective. American Economic Review 51, 860–911.

Galor, O., 2011. Unified Growth Theory. Princeton University Press, Princeton, NJ.

Goodwin, R.M., 1967. A growth cycle. In: Feinstein, C.H. (Ed.), Socialism, Capitalism and Economic Growth. Cambridge University Press, Cambridge, UK.

Lee, R.D., 1974. The formal dynamics of controlled population and the echo, the boom and the bust. Demography 11, 563–585.

Tuljapurkar, S., 1987. Cycles in nonlinear age-structured models I: renewal dynamics. Theoretical Population Biology 32, 26–41.

Population Dynamics: Theory of Stable Populations

John H Pollard, Macquarie University, Sydney, NSW, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11773–11777, © 2001, Elsevier Ltd.

Abstract

A population with an age distribution which remains constant over time is referred to as stable. A stable population of constant size is said to be stationary. These concepts are of considerable antiquity. Subject to age-specific mortality and fertility rates which remain unchanged over time a population will eventually develop a stable age distribution which depends on those mortality and fertility rates but is independent of the initial age distribution of the population. This result, due to Sharpe and Lotka (1911), caused renewed interest in stable populations among demographers, and the theory underlying it is outlined. Although demographers usually think of a stable population as one with an age distribution which is unchanging, such a definition is restrictive, because populations can also exhibit stability in respect of other characteristics as well as age and in more general situations. The discrete approach of Bernardelli, Leslie, and Lewis, which readily permits generalizations to a range of population stability situations is therefore described, as well as extensions covering migration, multiregional populations, populations studied by parity of woman, and membership of particular organizations. Applications are discussed only very briefly with cross-references to the relevant article. Limitations of the underlying models are also mentioned, again with a cross-reference.

A population with an age distribution which remains constant over time is referred to as 'stable.' A stable population of constant size is said to be 'stationary.' These concepts are of considerable antiquity. Euler, for example, knew in 1760 the stable age distribution a population needed to have if it were to have a given growth rate. Compilers of some of the early life tables such as Graunt in 1662 appear to have assumed populations which were stationary, and nineteenth-century actuaries certainly made use of the stationary population concept (King, 1902).

Subject to age-specific mortality and fertility rates which remain unchanged over time a population will eventually develop a stable age distribution which depends on those mortality and fertility rates but is independent of the initial age distribution of the population. This result, due to Sharpe and Lotka (1911) caused renewed interest in stable populations among demographers. Although demographers usually think of a stable population as one with an age distribution which is unchanging, such a definition is restrictive, because populations can also exhibit stability in respect of other characteristics as well as age and in more general situations.

The Population Model of Sharpe and Lotka

Sharpe and Lotka considered the male population alone and assumed that the growth of the female population would be such as to justify assumptions of constant age-specific fertility and mortality for the males. Subsequent writers have usually applied the model to the female component of the population because of its shorter reproductive age span and the fact that the number of females in the reproductive age range plays a more significant role in determining the number of births than the number of males (Pollard, 1973).

Applied to the female population, the model assumes $F(x, t)\,\mathrm{d}x$ females in the age range $(x, x + \mathrm{d}x)$ at time t and $B(t)\mathrm{d}t$ female births between times t and $t + \mathrm{d}t$. Then, if the average number of daughters born per female aged exactly x in time element $\mathrm{d}t$ is $\lambda(x)\mathrm{d}t$ and the probability that a female survives from age u to age $u + v$ is ${}_vp_u$, the following relationship is immediately apparent:

$$B(t) = \int_0^\infty F(x,t)\,\lambda(x)\mathrm{d}x \qquad [1]$$

Noting that those alive and aged x at time t, enumerated in $F(x, t)\,\mathrm{d}x$, must either have been born at time $t - x$ $(x < t)$ or else have been aged $x - t$ at time 0 $(x > t)$, eqn [1] may be rewritten as

$$B(t) = \int_0^t B(t-x)\,{}_xp_0\,\lambda(x)\mathrm{d}x + G(t) \qquad [2]$$

with

$$G(t) = \int_0^\infty F(x,0)\,{}_tp_x\,\lambda(x+t)\mathrm{d}x \qquad [3]$$

The solution to the integral equation (2) reveals that for large values of t, under quite general assumptions

$$B(t) \cong C\,e^{rt} \qquad [4]$$

where the constant r (the *intrinsic growth rate* of the population) is the unique real solution to the integral equation

$$\int_0^\infty e^{-rx}\;\;{}_xp_o\,\lambda(x)\mathrm{d}x = 1 \qquad [5]$$

and the constant C depends on the initial age distribution of the population as well as the age-specific mortality and fertility rates (Keyfitz, 1968; Pollard, 1973).

Since the females aged between x and $\mathbf{x} + \mathrm{d}x$ at time t are the survivors of those born x years earlier between times $t - x$ and $t - x + \mathrm{d}x$, it follows from eqn [4] that

$$F(x,t) \cong C\,e^{r(t-x)} \qquad {}_xp_o = \left(C\,e^{rt}\right)\left(e^{-rx}\;{}_xp_0\right) \qquad [6]$$

The first factor in the right-hand side of eqn [6] indicates that the population grows exponentially asymptotically at rate r, while the

second reveals that the age distribution is asymptotically independent of t. In other words, a *stable age distribution* is reached.

The Discrete Time Model of Bernardelli, Leslie, and Lewis

Whilst the continuous time model of Sharpe and Lotka is elegant and predicts a stable population asymptotically under conditions of constant age specific mortality and fertility, the discrete time formulation of Bernardelli (1941), Lewis (1942) and Leslie (1945), which makes similar assumptions, is more readily adapted to analyze stability in more complicated situations.

The number of females aged x last birthday at time t is represented by $n_{x,t}$. There is no migration, and no one can live beyond age k last birthday. The average number of daughters born to a female while she is aged x last birthday, those daughters surviving to be enumerated as being aged zero at the next point of time, is F_x, and the proportion of females aged x last birthday surviving to be aged $x+1$ last birthday a year later is P_x. Under these assumptions, the following matrix recurrence equation must apply to the numbers of females in successive years:

$$\begin{pmatrix} n_{0,\,t+1} \\ n_{1,\,t+1} \\ n_{2,\,t+1} \\ n_{3,\,t+1} \\ \cdots \\ n_{k,\,t+1} \end{pmatrix} = \begin{pmatrix} F_0F_1 & F_2\cdots & & F_{k-1}F_k \\ P_0 & & & \\ & P_1 & & \\ & & P_2 & \\ & & \ddots & \\ & & & P_{k-1} \end{pmatrix} \begin{pmatrix} n_{0,\,t} \\ n_{1,\,t} \\ n_{2,\,t} \\ n_{3,\,t} \\ \cdots \\ n_{k,\,t} \end{pmatrix} \qquad [7]$$

This equation may be written more concisely as

$$\mathbf{n}_{t+1} = \mathbf{A}\,\mathbf{n}_t \qquad [8]$$

so that

$$\mathbf{n}_t = \mathbf{A}^t\,\mathbf{n}_0 \qquad [9]$$

Under fairly general conditions, matrix $\mathbf{A}$ is positive regular, in which case it has a dominant eigenvalue α which is positive, of multiplicity one, and greater in absolute size than any other eigenvalue. It follows from eqn [9] therefore that for large t,

$$\mathbf{n} \cong C\,\alpha^t\,\mathbf{z} \qquad [10]$$

where C is a constant which depends on the initial age distribution of the population, and $\mathbf{z}$ is the column eigenvector corresponding to the eigenvalue α. If Π_s is written instead of the product $P_0P_1...P_S$, the elements of $\mathbf{z}$ are $\Pi_0\alpha^{-0}, \Pi_1\alpha^{-1}, ..., \Pi_k\alpha^{-k}$. It follows, therefore that the number of females aged x last birthday at time t is

$$n_{x,t} \cong \left(C\ \alpha^t\right)\ \left(\alpha^{-x}\ \Pi_x\right) \qquad [11]$$

This is the discrete analogue of the Sharpe and Lotka formula, eqn [6]. The second factor defines the asymptotic stable age distribution; the first shows the exponential growth of the population (Keyfitz, 1968; Pollard, 1973).

Stability with Migration

The most appropriate model for a population subject to significant migration must depend on the pattern that migration takes. In the case of a country accepting a constant number of immigrants each year with a constant age pattern, for example, the discrete time recurrence of eqn [8] needs to be modified as follows:

$$\mathbf{n}_{t+1}\ \ \mathbf{A}\,\mathbf{n}_t + \mathbf{b} \qquad [12]$$

where $\mathbf{b}$ is a constant vector of immigrants. Under this model, immigrants are assumed to adopt immediately the mortality and fertility patterns of their new homeland. Repeated application of eqn [12] for $t = 0, 1, 2, ...$ reveals that, provided the dominant eigenvalue of $\mathbf{A}$ is not equal to one,

$$\mathbf{n}_t \ = \ \mathbf{A}^t\,\mathbf{n}_0 + \left(\mathbf{I} - \mathbf{A}^t\right)\left(\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{b} \qquad [13]$$

If the rate of natural increase of the population is less than zero, α will be less than one, and for very large t

$$\mathbf{n}_t \ \cong\ \left(\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{b} \qquad [14]$$

The population will become stationary with its age structure determined by the ages of the immigrants (vector $\mathbf{b}$), as well as the mortality and fertility of the host country (matrix $\mathbf{A}$). Where the fertility of the host country is greater than replacement level, α will be greater than one, and the following asymptotic formula follows from eqn [13] for very large t:

$$\mathbf{n}_t \ \cong\ C'\ \alpha^t\,\mathbf{z} \qquad [15]$$

In this case, the population grows asymptotically with the same exponential growth rate as it did in the discrete non-immigration case of Section 2. The asymptotic stable age distribution (vector $\mathbf{z}$) is also the same. The constant C', however, is different, depending on the initial age structure, the vector of immigrants $\mathbf{b}$, and the mortality and fertility of the population.

In the pathological case where the levels of mortality and fertility in the population correspond exactly to replacement, α is equal to one, and a different algebraic approach is necessary to the solution of eqn [12], which yields, for very large t, the following asymptotic result:

$$\mathbf{n}_t \ \cong\ C''\ t\ \mathbf{z} \qquad [16]$$

The population grows linearly asymptotically and the asymptotic age distribution is the same as in the previous case. The constant C'' depends on the immigration vector, as well as the fertility and mortality rates of the population.

For a population experiencing net emigration with a proportion β_x of females aged x departing each year, the matrix model of Section 2 can be utilized with each of the P_x in matrix $\mathbf{A}$ replaced by P_x–β_x. The mathematical theory can then be applied in the same manner as before, and an asymptotic stable population emerges. The dominant eigenvalue and eigenvector will of course be different from those in Section 2. Depending on whether the dominant eigenvalue of the modified matrix $\mathbf{A}$ is less than, equal to, or greater than one, the population will ultimately decline, become stationary or grow in size.

The above models are only two of a very wide range of possible migration models appropriate in different circumstances. They demonstrate how the discrete time model of Section 2 may be readily modified to take account of migration and the fact that population stability will usually emerge in the presence of migration when the migration parameters, mortality, and fertility remain unchanged over time.

Generalizations

The discrete time models of the earlier sections are readily adapted to study populations partitioned according to other characteristics as well as age. Multiregional demography is the best known example (Rogers, 1966, 1975). Females of a population are characterized by age and geographical region in an extended column vector $\mathbf{n}_t$ of dimension mk, where k is the maximum age last birthday attainable and m is the number of regions in the country. As well as surviving and reproducing in the manner of Section 2 with rates that may differ from region to region, females can move from one region to another with transition rates which depend on the regions involved and may also depend on age. The matrix recurrence equation is again eqn [8], but with $\mathbf{n}_t$ of dimension mk and $\mathbf{A}$ of dimension $mk \times mk$. Conditions for a unique dominant eigenvector are readily determined, and asymptotic stability for the population by age and geographic region is predicted.

Another partition characteristic sometimes proposed is the *parity* of woman. Females at time t are enumerated in a vector $\mathbf{n}_t$ with elements $\{n_{x,y,t}\}$ which record the numbers aged x last birthday who already have y children. The only possible transitions for a female aged x and parity y remaining within the system are to age $x + 1$ and parity y or to age $x+ 1$ and parity $y + 1$ (ignoring the small possibility of multiple births). Where there is a change in parity, a new female will be introduced into the population aged 0 last birthday and parity 0 if the mother's change in parity corresponds to a female birth. Recurrence matrix $\mathbf{A}$ contains these transition rates. The fundamental recurrence equation is again eqn [8]. For all realistic populations, the enlarged matrix $\mathbf{A}$ will have a unique dominant eigenvector, and the population enumerated by age and parity will approach stability (Keyfitz, 1968).

The discrete-time models of Sections 2 and 3 have also been applied to workforce planning (Bartholomew, 1973; Pollard, 1967), membership of learned societies (where Fellows are elected on merit and there is a danger of the society becoming a rapidly aging institution), and membership of other groups including pension schemes (Sherris and Pollard, 1980). A learned society, for example, might have a policy of electing five new Fellows each year. The membership of the society at time t can be summarized by a vector $\mathbf{n}_t$ with elements $\{n_{x,t}\}$ representing the numbers at the various ages at that time. The number of new members (five) and their age distribution can be summarized in a vector $\mathbf{b}$. The transition rates in the recurrence matrix are simply the proportions $\{P_x\}$ surviving from age x last birthday to age $x + 1$ last birthday a year later, for all relevant ages. Assuming that the ages are listed in their natural increasing order, the transition matrix $\mathbf{A}$ is particularly simple in this case: all its elements are zero, except those immediately below the main diagonal which comprise the survival proportions $\{P_x\}$. With $\mathbf{n}_t$, $\mathbf{b}$, and $\mathbf{A}$ defined in this manner, the recurrence equation for the society is eqn [12]. All the eigenvalues of $\mathbf{A}$ are zero, and for t greater than the number of ages involved, $\mathbf{A}^t = \mathbf{O}$. The society will approach a stationary state given by eqn [14].

All such linear models with realistic assumptions concerning transition between states, entry and reproduction can be shown to predict stable populations asymptotically.

Applications

No observed population experiences constant levels of mortality and fertility. Nevertheless the concept of a stable population finds frequent application. Keyfitz's elegant formula for the effect of population momentum (*see* Population Dynamics: Momentum of Population Growth) assumes that the population is stable immediately prior to fertility falling to replacement level. Yet the formula works remarkably well for observed populations which are not stable. Stable population theory may be helpful in deducing results in respect of populations with unreliable and/or limited demographic data. Equation (4), for example, can be used to infer an approximate value for the crude birth rate of a population where contraception is not widely practiced, using only a reliable census of the population and a more or less appropriate life table (Bourgeois-Pichat, 1957). Many of the applications of model life tables rely on stable population theory (*see* Population Dynamics: Classical Applications of Stable Population Theory).

Limitations

The discrete time model of Bernardelli, Leslie, and Lewis, like the continuous (female) formulation of Sharpe and Lotka, assumes that the development of the male component of the population is such as to justify the assumption of unchanging female age-specific mortality and fertility. Linear models which include the male component of the population alongside the females have also been proposed, and mathematical analysis analogous to that presented in Section 2 predicts a stable age–gender composition for the whole population. Nonlinear models have also been proposed to accommodate the 'two-sex problem' (*see* Population Dynamics: Two-Sex Demographic Models). Many of these nonlinear models would also seem to predict populations which are asymptotically stable under conditions of unchanging rates of mortality, fertility and nuptiality.

For many developing countries, for which stable population theory had been an extremely helpful tool, mortality began to decline in the second half of the twentieth century, but fertility remained essentially unchanged from previous levels. To accommodate the improvements in mortality, the concept of a quasi-stable population was introduced: one with improving mortality but constant fertility (Coale, 1963).

Bibliography

Bartholomew, D.J., 1973. Stochastic Models for Social Processes, second ed. Wiley, New York.

Bernardelli, H., 1941. Population waves. Journal of Burma Research Society 31, 1–18.

Bourgeois-Pichat, J., 1957. Utilisation de la notion de population stable pour mesurer la mortalité et la fécondité des populations des pays sous-développés. Bulletin de l'Institut International de Statistique (actes de la 30e Session).

Coale, A.J., 1963. Estimation of various demographic measures through the quasi-stable age distribution. Emerging Techniques in Population Research: Proceedings of the 1962 Annual Conference of the Milbank Memorial Fund.

Keyfitz, N., 1968. Introduction to the Mathematics of Population. Addison-Wesley, Reading, MA.
King, G., 1902. Life Contingencies, second ed. Charles & Edward Layton, London.
Leslie, P.H., 1945. On the use of matrices in certain population mathematics. Biometrika 33, 183–212.
Lewis, E.G., 1942. On the generation and growth of a population. Sankhya 6, 93–96.
Pollard, J.H., 1967. Hierarchical population models with Poisson recruitment. Journal of Applied Probability 4, 209–213.
Pollard, J.H., 1973. Mathematical Models for the Growth of Human Populations. Cambridge University Press, Cambridge, UK.
Rogers, A., 1966. The multiregional matrix growth operator and the stable interregional age structure. Demography 3, 537–544.
Rogers, A., 1975. Introduction to Multiregional Mathematical Demography. Wiley, New York.
Sharpe, F.R., Lotka, A.J., 1911. A problem in age-distribution. Philosophical Magazine, Ser 6 (21), 435–438.
Sherris, M., Pollard, J.H., 1980. Application of matrix methods to pension funds. Scandinavian Actuarial Journal 77–95.

Population Dynamics: Two-Sex Demographic Models

Robert Schoen, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by R.A. Pollak, volume 17, pp. 11773–11777, © 2001, Elsevier Ltd.

Abstract

Demographic models of marriage and fertility that incorporate the behavior of both males and females encounter the so-called 'two-sex problem.' That problem arises because observed one-sex male and female rates of marriage and birth are influenced by the total age–sex composition. The leading two-sex solution that has been advanced is the harmonic mean approach, which assumes that the sum of male and female marriage or birth rates is independent of such compositional effects. Although other serious alternatives have been proposed, the harmonic mean approach is based on the most plausible demographic foundation.

The marriage squeeze is the best known manifestation of two-sex population dynamics. Using the harmonic mean approach, analyses indicate that over the past 50 years marriage squeezes have had significant impacts on the timing of marriage. Nonetheless, marriage squeezes have had less influence on the probability of ever marrying, and are unlikely to have played a major role in large-scale social change.

Human populations have two sexes, males and females, but demographic models generally depict only one sex. The major reason is the so-called 'two-sex problem,' the fact that conventional marriage and birth rates are shaped by the composition of the entire population.

Here we examine the nature of the two-sex problem, alternative approaches that have been proposed to deal with it, and touch on some demographic models that incorporate the experience of both sexes. Then we consider the 'marriage squeeze' and the 'birth squeeze,' real world manifestations of two-sex demography, and examine their impact on population dynamics.

The Two-Sex Problem

To begin, let us consider a hypothetical population of 100 males and 100 females, which experiences 10 marriages during the course of a year. Conventional one-sex incidence/exposure rates of marriage are given by the number of marriages divided by the population at risk. Here, the male marriage rate, W_m, is $10/100 = 0.1$. That is identical to the female marriage rate, W_f. If the underlying propensity to marry is unchanged and the number of males and females continues to be equal, rates of 0.1 will continue to yield the annual number of marriages in the population.

Now assume that the propensity to marry and the size of the female population remain unchanged, but the size of the male population increases to 200. Applying the male rate to the male population gives 20 marriages, while applying the female rate to the female population yields 10 marriages. Given the traditional definition of marriage as involving one male and one female, it is clear that both the male and female rates cannot persist when relative population size changes. In a real population, we would expect the number of observed marriages to increase, but not to double, because the number of females did not double. Hence the male marriage rate would fall below 0.1 while the female marriage rate would rise above 0.1. The two-sex problem is to determine exactly how male and female marriage rates would change. While here we only consider traditional marriage, a 'two-sex' problem would still exist if partnerships were formed independent of gender, as long as propensities to partner varied by age or some other demographic characteristic.

The Harmonic Mean Solution

Schoen (1981) argued that if there were no change in the propensity to marry, the rise in the marriage rate for one sex would exactly offset the fall in the marriage rate for the other sex, leaving the sum of the male and female rates constant. That sum of male and female rates, termed the 'magnitude of marriage attraction,' would reflect the underlying propensity to marry between males and females, and would be independent of population composition. Magnitudes of marriage attraction, like rates, could be made specific to age and other characteristics of the men and women marrying. If two populations have the same magnitudes of marriage attraction, then, in every case, the sum of the characteristic-specific male and female marriage rates would be equal, even though the corresponding male and female rates might differ.

The above approach is a harmonic mean solution because the magnitude of marriage attraction, as the sum of male and female rates, equals the number of marriages divided by half the harmonic mean of the number of males and females. The magnitude of marriage attraction is zero if either the number of males or females is zero, and if one sex is large relative to the other, is proportional to the number in the less numerous sex. If the number in both sexes increases by factor f, then the number of marriages increases by factor f. The denominator of the magnitude of marriage attraction can also be interpreted as the total number of male–female pairs divided by the sum of the male and female populations. Hence the 'two-sex population' denominator is the average number of possible pairs per person.

An important property of a two-sex solution is sensitivity to the age–sex composition of the entire 'marriage market,' i.e., all persons eligible to marry. For example, the chance that a 25 year old man will marry a 24 year old woman should be influenced by competition from the number of men aged 24 and women aged 23, and indeed by the number of eligible men and women at all ages. It is not obvious that the harmonic mean approach, which determines the magnitude of marriage attraction from just the number of men aged 25 and women aged 24, does so. Yet, the harmonic mean approach is sensitive to the marriage market because of the intimate relationship between population stocks (numbers of persons) and flows (here, marriages).

Consider two points. First, the number of persons in the (midyear) population eligible to marry and the number of observed marriages reflect outcomes from the competitive environment of the marriage market. Second, recall that rates are based on the experience of a given year. Now assume that we are at the beginning of a year of observation and know the array of magnitudes of marriage attraction and the initial number of unmarried persons by age and sex. To calculate the number of marriages by age and sex, we need the midyear, not beginning of the year, populations. Those midyear populations are affected by the marriages that occur during the first half of the year, thus the size and composition of the midyear populations bears the imprint of the competitive context of the marriage market.

Because of its desirable properties, the harmonic mean approach is generally accepted as 'a' solution to the two-sex problem. However, there is no consensus as to 'the' solution, and a number of other serious approaches have been advanced.

Alternative Two-Sex Solutions

Schoen (1988: Chapter 6) reviewed a number of solutions to the two-sex problem that have been advanced. Here, we focus on 4 approaches: (1) other simple means or combinations of rates; (2) iterative proportional fitting; (3) panmictic circles; and (4) the birth matrix-mating rule approach. Since the essence of the two-sex problem is how population composition impacts male and female rates, methods based only on relative numbers of males and females are not explored.

Other Simple Means or Combinations of Rates

The simple arithmetic mean, i.e., half the sum of the male and female rates, is an obvious possibility. However, it has received little attention because it yields marriages even when there are no persons of one sex.

The geometric mean, i.e., the square root of the product of male and female rates, is a simple average that avoids the problem of marriages when there are no persons of one sex. Goodman (1967) discussed a geometric mean variant where the two sexes are given unequal weight, that is where the male rate would be raised to the power α ($0 \leq \alpha \leq 1$) and the female rate would be raised to the power $(1 - \alpha)$. Parameter α would indicate the 'dominance' of one sex.

Rather than averaging male and female rates, one could simply take the minimum of the two rates. That 'minimum' approach avoids the problem of marriages in the absence of one sex, and has been incorporated into more sophisticated models.

Iterative Proportional Fitting

McFarland (1975) applied a variation of the statistical technique of iterative proportional fitting (IPF), also known as the Deming-Stephan procedure or biproportional adjustment. Essentially, IPF begins by creating a marriage array, i.e., a cross-tabulation of marriages by age of bride and age of groom. The rows of the array contain elements showing the numbers of marriages for a given age of the bride and each age of the groom, plus an additional element equal to the total unmarried female population that age minus the total number of women marrying at that age. The columns contain elements showing similar numbers of marriages for a given age of the groom, plus an additional element equal to the total unmarried male population at that age minus the number of men marrying at that age. The row and column sums of the array, i.e., the marginals, are thus the total marriage eligible populations at each age. To apply IPF, change the marginals to the values for the population whose marriages are to be found. Then repetitively, or iteratively, multiply each row of the array by a factor that will yield that row's marginal. Next multiply each column of the array by a factor that will yield that column's marginal. Continue until no further row or column adjustments are needed. The process will always converge to a unique solution, and the resulting array will give the desired distribution of marriages by age and sex. If only one marriage age is involved, IPF is identical to the geometric mean approach. IPF is quite flexible, preserves the cross-product ratios of the original array, and is a maximum entropy (or minimum information) solution.

Marriage within Panmictic Circles

Henry (1972) viewed the marriage process as one involving men and women meeting and marrying within overlapping 'circles,' or large social groupings, within which marriage is 'panmictic' in the sense that age does not influence partner choice. Henry set forth a process for dividing male and female population distributions into a series of such circles, arguing that a maximum of six circles would be sufficient to characterize national level data. Within circles, male and female age-circle-specific marriage rates can be calculated, and those rates applied to other populations that have been decomposed into panmictic circles. The male rate yields one number of marriages and the female rate yields another, with the minimum number used. The procedure is not readily visualized, but seems to respond to competition in the marriage market.

The Birth Matrix-Mating Rule Model

Pollak (1990), building on the work of Feeney (1972), advanced the birth matrix-mating rule (BMMR) model consisting of three fundamental components. The first component is a birth matrix, indicating the number of children born to each possible union type (e.g., males age y and females age x). The second is a mating rule, which gives the number of unions

of each kind from the number of males and females in each category. The third has the applicable male and female age-specific death rates.

Under a number of possible mating rules, the BMMR model provides the long term male and female age distributions and the age pattern of marriage. The mating rule used must satisfy two fairly weak conditions: it must not yield negative numbers and it cannot yield more male (female) unions in an age category than there are males (females) in that category. The model yields a unique result if an increase in the number of persons in any age category does not decrease the number of births.

The BMMR model offers a framework for two-sex model solutions, rather than a specific solution. The harmonic mean, geometric mean, and IPF approaches, among others, can provide satisfactory rules whose demographic implications can be found using the BMMR approach.

Evaluating Alternative Two-Sex Solutions

Demography is empirically oriented, so it is logical to try and evaluate alternative approaches by applying them to actual data. The problem is that the standard for comparison – the correct result – is unknown. The most promising approach is to compare the same population over time, but that still requires the rather strong assumption that all of the propensities have remained constant. Schoen (1981) made such comparisons for Sweden, using single year of age male and female population and marriage data for the years 1961, 1962, 1963, and 1964. The harmonic mean, iterative adjustment, and panmictic circles approaches were each applied to the data, using 1961 data to predict 1962 marriages, 1962 data to predict 1963 marriages, and 1963 data to predict 1964 marriages. None of the three approaches performed well. All of them tended to be off in the same direction, but in an inconsistent and erratic fashion. A recourse to data thus seems unlikely to provide a useful evaluation of alternative two-sex solutions.

Another approach has been to posit a set axioms or conditions that a satisfactory method must satisfy, e.g., those set forth by Parlett (1972) or those required by the BMMR model. That approach has not worked well either, as no set of axioms has been generally accepted, and the leading two-sex solutions satisfy virtually all of the required properties that have been proposed.

There appears to be no alternative to evaluating two-sex solutions on theoretical or conceptual criteria. To be considered an acceptable solution, a two-sex approach should not only yield reasonable results empirically, but be based on a plausible, demographically meaningful rationale. In that respect, the harmonic mean emerges as the most satisfactory approach. No clear explanation has been offered for why the cross-product ratios of a marriage matrix should be preserved, why marriage should be independent of age within a few large 'circles,' or why the geometric mean captures the interaction between the sexes. In contrast, the harmonic mean method is based on magnitudes of attraction that reflect the marginal change in the number of marriages (or births) given a marginal change in the number of couples, and that imply a two-sex population at risk equal to the average number of pairs per person.

Two-Sex Demographic Models

Traditional one-sex demographic models can be extended to incorporate two-sex marriage or two-sex fertility. Along the three dimensions of single state/multistate, stationary/stable, and one-sex/two-sex, Schoen (1988: p. 120) identified six distinct two-sex nuptiality models and two two-sex fertility models.

The conventional, one-sex life table follows a cohort, or closed group of persons, over time, often recognizing more than one cause of decrement (loss) from the cohort. In particular, a one-sex nuptiality–mortality life table follows single person from birth to decrement from either first marriage or death, based on rates from a source population. The corresponding two-sex nuptiality–mortality life table follows a male cohort and a female cohort from birth to first marriage or death. The source population provides the one-sex male and female rates as well as the two-sex magnitudes of marriage attraction.

The classical one-sex stable population depicts the implications of a set of male or, more usually, female age-specific birth and death rates, showing the long term exponential population growth and constant age composition that those rates imply. Similarly, a two-sex (fertility) stable population shows the implications of the source population's male and female death rates and magnitudes of fertility attraction, i.e., sums of male and female age-specific birth rates.

A number of two-sex life table and stable population models have been constructed and analyzed, though such analyses are still infrequent. Two-sex modeling demands extensive input data, and is hampered by a lack of consensus on the appropriate two-sex methodology.

Manifestations of Two-Sex Behavior

Two-sex population dynamics, or the interaction between the sexes with regard to marriage and birth, is important to study because it influences marriage and fertility behavior. An imbalance between the number of men and women at the prime ages of marriage (or reproduction) alters the marriage (or fertility) rates of both sexes, and the effect can be substantial. A gender imbalance at the marriage ages produces what has been termed a 'marriage squeeze.' The analogous phenomenon, a 'birth squeeze,' can arise with respect to fertility. We examine each in turn, using measures based on the harmonic mean approach.

The Marriage Squeeze

Marriage squeezes typically arise when the number of births is rapidly rising (or falling). In most populations, men marry women several years younger than themselves. Thus when the number of births is increasing, men are looking for partners born in larger cohorts while women are looking for partners born in smaller cohorts. The result is a marriage squeeze disadvantaging females. If the number of births is falling, then the same process yields a marriage squeeze disadvantaging males.

Some observers have theorized that the marriage squeeze played a key role in social change during the twentieth century. Heer and Grossbard-Shechtman (1981) argued that a marriage squeeze against females reduced the proportion of US women

who married during the 1960s, changing the 'terms of trade' between men and women and increasing women's labor force participation. Davis and van den Oever (1982) interpreted the feminist movement as a response to lower fertility, less marriage, and the marriage squeeze. Guttentag and Secord (1983) argued that whenever there are *Too Many Women* (the title of their book), as in the United States during the 1960s and 1970s, marriage and family life are devalued and sexual libertarianism prevailed. Such claims by serious researchers deserve careful consideration and analysis.

Because two-sex dynamics change one-sex rates, we want a marriage squeeze measure that reflects how compositional changes have influenced male and female behavior. One such measure is marriage squeeze index S (cf Schoen, 1988). Index S is the male life table number of male marriages, minus the female life table number of female marriages, divided by the total number of marriages in the two-sex nuptiality life table. If $S > 0$, there is a marriage squeeze against females, if $S < 0$, there is a marriage squeeze against males, and if $S = 0$, there is no marriage squeeze at all. The larger the magnitude of S, the greater the squeeze. For example, if there were 98 marriages in a male life table, 92 marriages in a female life table, and 96 marriages in a two-sex table, then $S = 6/96 = 0.0625$.

Schoen (1983) looked at index S in 25 countries during the period 1966–75, using data from the 1976 United Nations Demographic Yearbook. The largest marriage squeezes were found in Jordan 1970 ($S = 0.14$) and Costa Rica 1973 ($S = 0.12$). More developed countries had S values that were either small or negative, the most negative being Sweden 1975 ($S = -0.04$). Schoen and Baj (1985) examined trends in index S in five Western countries (Belgium, England and Wales, Sweden, Switzerland, and the United States) during the period 1910–75. In the post–World War II period, S values were small, though Switzerland's was about 0.03 and Sweden's became negative. However, in the 1910s and 1920s, England and Wales, Sweden, and Switzerland had S values of 0.125 or above.

Schoen (1983) also looked at the marriage squeeze and its implications for the United States during the period 1950–90, assuming that 1970 magnitudes of marriage attraction applied over that period. Index S was always quite small, varying between 0.0013 and 0.0073, showing a slight marriage squeeze against females. Closer examination, however, revealed significant changes in the distribution of male and female marriages. The male mean age at marriage varied by 0.52 years, the female mean age at marriage by 0.62 years, and the mean difference between partner marriage ages fluctuated between 1.84 and 2.99 years. The variance of the distribution of both male and female ages at marriage also changed substantially.

In sum, studies suggest that marriage squeezes of some magnitude do characterize the contemporary world. While marriage squeeze effects on the proportion ever marrying appear modest, at least in Western countries, they have exerted substantial influence on the timing and distribution of marriages, including the age difference between spouses. Still, the evidence falls short of supporting the marriage squeeze as a cause of large-scale social change. In the 1910s and 1920s, sizable marriage squeezes in the West passed with little comment, and most social change since the 1960s, when the marriage squeeze has been small, seems attributable to more fundamental social and economic transformations.

The Birth Squeeze

The same demographic dynamic that produces marriage squeezes produces birth squeezes, even when marriage is not a necessary societal prelude to fertility.

The standard measure of fertility is the total fertility rate (TFR_f), the sum of age specific female birth rates over the female reproductive age span. The TFR_f can be interpreted as the average number of children a woman would bear in her lifetime in the absence of mortality. Fertility analysis has traditionally emphasized female birth rates because of the narrower female reproductive age span and the fact that data on fathers are often incomplete. Though less frequently calculated, TFR_m, the male total fertility rate, is the sum of male fertility rates over the male life course.

The approach expressed here, i.e., that two-sex dynamics affect one-sex rates, but that the sum of male and female rates is composition independent, applies to fertility as well as marriage. Following Schoen (1985), define the two-sex total fertility rate, TFR2, as one-half the sum of TFR_m and TFR_f. Paralleling index S, birth squeeze index U can be defined as $(TFR_m - TFR_f)/TFR2$. When $U > 0$, there is a birth squeeze against females, when $U < 0$, there is a birth squeeze against males, and when $U = 0$, there is no birth squeeze.

Schoen (1985) examined index U for 22 countries during the 1963–74 period, using data from the United Nations Demographic Yearbooks. Developed countries in the sample typically had small or negative values of U. For the United States 1970, U was 0.12. Four countries, Phillipines 1973, Jordan 1974, Tunisia 1971, and Puerto Rico 1970, had U values over 0.25. The Phillipines 1973, where U was 0.30, was the highest, having a TFR_m of 4.6 and a TFR_f of 3.4.

Since population growth is associated with birth squeezes against females, it is not surprising that growing populations have sizable values of U. Less obviously, when $TFR2 > TFR_f$, an argument can be made that the conventional TFR is understating the population's actual propensity for fertility.

Conclusion

Conventional one-sex demographic rates are unable to capture the dynamics underlying two-sex marriage and fertility behavior, or to model such demographic phenomena as the marriage squeeze. The harmonic mean approach, the leading two-sex solution to date, affords an opportunity for such analyses. The results have provided strong evidence that marriage squeezes have had considerable influence on the timing and pattern of marriage, but little support for a belief that marriage squeezes have caused major societal change.

See also: Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demographic Techniques: Data Adjustment and Correction; Population Dynamics: Theory of Nonstable Populations.

Bibliography

Das Gupta, Prithwis, 1973. Growth of the U.S. population, 1940–1971, in the light of an interactive two-sex model. Demography 10, 543–565.

Davis, Kingsley, van den Oever, P., 1982. Demographic foundations of new sex roles. Population and Development Review 8, 376–395.

Feeney, Griffith, M., 1972. Marriage Rates and Population Growth: The Two-Sex Problem in Demography (Ph.D. thesis). University of California, Berkeley, CA.

Goodman, Leo A., 1967. On the age-sex composition of the population that would result from given fertility and mortality conditions. Demography 4, 423–441.

Guttentag, Marcia, Secord, Paul F., 1983. Too Many Women? The Sex Ratio Question. Sage, Beverly Hills.

Heer, David M., Grossbard-Shechtman, Amyra, 1981. The impact of the female marriage squeeze and the contraceptive revolution on sex roles and the women's liberation movement in the United States, 1960 to 1975. Journal of Marriage and the Family 43, 49–65.

Henry, Louis, 1972. Nuptiality. Theoretical Population Biology 3, 135–152.

Iannelli, Mimmo, Martcheva, Maia, Milner, Fabio A., 2005. Gender Structured Population Modeling: Mathematical Methods, Numerics, and Simulations. Society for Industrial and Applied Mathematics, Philadelphia.

McFarland, David D., 1975. Models of marriage formation and fertility. Social Forces 54, 66–83.

Parlett, Beresford, 1972. Can there be a marriage function? In: Greville, T.N.E. (Ed.), Population Dynamics. Academic, New York.

Pollak, Robert A., 1990. Two-sex demographic models. Journal of Political Economy 98, 399–420.

Schoen, Robert, 1981. The harmonic mean as the basis of a realistic two-sex marriage model. Demography 18, 201–216.

Schoen, Robert, 1983. Relationships in a simple harmonic mean two-sex fertility model. Journal of Mathematical Biology 18, 201–211.

Schoen, Robert, 1985. Population growth and the birth squeeze. Social Science Research 14, 251–265.

Schoen, Robert, 1988. Modeling Multigroup Populations. Plenum, New York.

Schoen, Robert, Baj, John, 1985. The impact of the marriage squeeze in five western countries. Sociology and Social Research 70, 8–19.

Population Forecasts

Juha M Alho, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Population forecasts are used to summarize existing knowledge of population change to help decision makers. Since the 1920s, the previously favored curve fitting methods were replaced by the cohort component method that adds age and sex details. This requires separate forecasting of age-specific fertility, mortality, and migration. A defect in the classical cohort component methods is that they fail to convey to the forecast users an impression of the statistical uncertainty of the forecast in an understandable way. In particular, the high–low variants that have often been used suffer from logical defects. Probabilistic approaches have been developed for the task. Due to the availability of fast computing, simulation can be used to generate a population of sample paths for the population by age and sex, for the future years of interest. The sample paths provide a numerical approximation to the predictive distribution of the future population. Such results can be used, for example, in macroeconomic analyses of aging, sustainability of public finances, or planning of health care or education.

Population forecasts are used to summarize existing knowledge of population change to help decision makers. Most methods of forecasting that are currently used by official statistical agencies were developed during 1920s to 1940s. Previously, the calculations concerning future population usually involved the fitting of mathematical curves (exponential or logistic, for example) to the trajectory of population size. The so-called cohort component method added age and sex detail. It can give important insight into how current vital rates (fertility, mortality, and migration) and current age distribution combine into the observed growth rate. Nevertheless, to improve overall accuracy, good forecasts of the age-specific vital rates are needed. Most population forecasts are purely demographic, that is, they do not take explicitly into account the social and economic factors that influence the vital rates. This is because such factors are often harder to predict than the demographic phenomena themselves. Similarly, the possibility that a forecast might become a self-fulfilling or self-negating prophecy is usually ignored. A defect in the classical cohort component methods is that they fail to convey to the forecast users an impression of the statistical uncertainty of the forecast in an understandable way. Probabilistic approaches have been developed for the task. Due to the availability of fast computing, simulation can be used to generate a population of sample paths for the population by age and sex, for the future years of interest. The sample paths provide a numerical approximation to the predictive distribution of the future population.

Population forecasts summarize existing knowledge about the change in the size and structure of populations. They support decision making by providing broad perspectives for future development, for problems related to aging, pension, health systems, etc. Although simple deterministic forecasts can provide us with important insights, policy design and risk management in the context of such macroeconomic issues require that uncertainty be realistically quantified. Although the fundamental reasons for the unpredictability of demographic processes are only poorly understood, the development of new statistical models and the availability of fast computing allow a realistic description of forecast uncertainty in terms that are understandable to nonspecialist users.

Development of the Cohort-Component Method

Curve Fitting Approach

Early calculations concerning the renewal of populations were based on simple mathematical models. Suppose $V(t)$ is the total population at time t. If the rate of growth is constant, or $V'(t)/V(t) = r$, then the population grows exponentially, $V(t) = V(0)e^{rt}$, from the starting, or *jump-off* population, $V(0)$. If the rate of growth decreases with increasing population according to $V'(t)/V(t) = r(M - V(t))/M$, where M is a limiting population size, then the growth is *logistic*: $V(t) = Me^{rt}/(1 + e^{rt})$. A discrete time version of the exponential model was used, for example, by T. Malthus in his *Essay on the Principle of Population* in 1798. The logistic model was introduced by F. Verhulst in 1838.

Such models were widely criticized for the following reasons: (1) they do not give age detail; (2) they do not allow a change from increase to decrease; (3) they are too mechanistic for human populations; (4) even though the exponential and logistic models are empirically indistinguishable when $V(t)$ is small relative to M, they give completely different long-term forecasts.

Incorporation of Age Structure

Methods for describing age-structured populations were developed around the turn of the nineteenth century in many countries; first in England by E. Cannan in 1895. By the end of the 1920s, forecasts by age and sex had also been made for the Soviet Union, The Netherlands, Sweden, Italy, Germany, France, and the United States (de Gans, 1999).

Suppose $V(x, t)$ is the size of the female population at age x at the beginning of year t. Then, for ages $x > 0$, we have the bookkeeping equation $V(x + 1, t + 1) = V(x, t)\, p(x, t) + N(x + 1, t + 1)$, where $p(x, t)$ is the probability of surviving from age x to age $x + 1$ during year t, and $N(x + 1, t + 1)$ is the net number of migrants during year t who are at age $x + 1$ at the beginning of year $t + 1$. For age $x = 0$ we have, $V(0, t + 1) = \sum_x f(x, t)V(x, t) + N(0, t + 1)$, where the summation is over the childbearing ages (often $x = 15, \ldots, 49$), and $f(x, t)$ is the expected number of children at age 0 at the

beginning of year $t + 1$ that were borne by women at age x during year t. Corresponding formulas can be written for the male population, and they can be extended to comprise several regions. This system of calculation is called the *cohort-component* (c-c) method.

Advantages of the Cohort-Component Method

The c-c method can be viewed as a response to the criticisms of the Section Curve Fitting Approach in the following sense: (1) Since the c-c method treats the population by age groups, it provides direct estimates of the future population by age and sex. That this is useful for the planning of housing, schools, hospitals, etc., was understood from the beginning. (2) The c-c method makes no assumption about the growth rate. It shows that, other things being equal, the growth rate increases with the proportion of women of childbearing ages. In 1925, L. Dublin and A. Lotka introduced the concept of intrinsic rate of growth. For a closed population with unchanging fertility and mortality, it gives the asymptotic rate of growth. As such it provides important insight into the renewal of populations that cannot be detected from current growth rates. Many of the early cohort-component forecasts assumed unchanging fertility and mortality, and no migration, and showed that the size of the childbearing population, and with it the total population, would cease to grow, even though the populations had been growing fast during the early 1900s. This contradicted the forecasts by exponential and logistic models. (3) The reason why the c-c forecasts are less mechanistic than those based on exponential or logistic curves is that they utilize the available information or assumptions about the change of the *vital rates* (of fertility, mortality, and migration). The forecasters of the 1920s knew that fertility and mortality had been declining for decades. How far down would they go and how fast? Examples from other countries or from privileged social groups were used as leading indicators for future mortality. The spreading of the urban lifestyle was seen to lead to a decrease in fertility. Opinion surveys were developed around 1940 to collect data on desired family size (Whelpton et al., 1947). (4) The early forecasters knew that the exponential, logistic, and other curves often gave similar fits to past data, but led to very different forecasts. Since statistical *model choice* depended on judgment but had a drastic effect on the results, why could one not use judgment in other ways as well? The solution by Whelpton to this problem was to set up targets for future vital rates, and to connect the most recently observed rates to the targets by suitable mathematical curves. Dutch forecasters had also used similar methods. This practice has been widely followed, and most national population forecasts use judgmental methods for specifying future vital rates, often via explicit targets (Andrews and Beekman, 1987).

Although the c-c method was an improvement over the curve fitting methods, it was soon understood that it was really a bookkeeping framework, and without an accurate specification of the future change in the vital rates improvements in forecast, accuracy would not materialize.

As an example, we may consider a forecast of Sweden that was crafted by S. Wicksell in 1926. The forecast had 1920 as the jump-off year with a population of 5.90 million. It had five fertility variants. For 1940 the most likely forecast was 6.53 million inhabitants. The high forecast was 6.87 million and the low forecast was 6.28 million. The actual population size was 6.37 million, hence the error was 2.5%. For 1970 the most likely value was given as 6.67 million and the high–low interval was [6.26, 8.44] million. The true value was 8.08 million so the error was −17.5%, but the high–low interval still gave a realistic indication of the uncertainty of the forecast. This is in contrast to many official forecasts, whose high–low intervals have been so narrow that the actual population count has escaped from the interval in a decade, if not sooner.

Foundational Aspects in the Forecasting of the Vital Rates

Most national forecasts of the vital rates are purely demographic, that is, they do not explicitly incorporate social or economic factors. Only rarely do they take into account the fact that the forecast itself might influence the future development. Partly because of these factors, and partly because of the poor accuracy of past forecasts, there has been a lengthy debate about whether the results of the calculations would be better interpreted as mere 'projections' rather than 'forecasts.'

The Difficulty with Explanatory Models

Refinement in demographic analysis often refers to the partitioning of the population at risk into more meaningful subgroups. For example, instead of relating births and deaths to the total population (which gives the crude birth and death rates), one may relate births to age of mother and consider deaths by age and sex (which gives the age-specific fertility and mortality rates) (*see* Demographic Measurement: General Issues and Measures of Fertility; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Demographic Techniques: Lexis Diagram). The use of crude rates in exponential or logistic models can imply growth forever, whereas the use of the age-specific rates in a c-c model may imply a future decrease of the same population.

The analysis can be further refined. Age-specific rates may be analyzed for birth cohorts and by parity; marital and nonmarital fertility may be considered separately; or one may consider fertility by race, for example. Empirical evidence of whether this improves accuracy is mixed.

However, the more fundamental social or economic factors that determine the fertility, mortality, and migration of population subgroups are rarely used explicitly in forecasting. This is because future changes in such factors can be harder to forecast than the vital processes themselves. Keyfitz (1982) reviews many of the difficulties.

One attempt at solving the problem has been *conditioning*. By making detailed assumptions about the future course of relevant social and economic factors, forecasters have sought to formulate conditional forecasts, or *scenarios*, in which the evolution of the vital processes would follow assumed relationships with their determinants. A drawback of such an approach is that the manner in which the conditioning is done can have a major effect on the results. A *sensitivity analysis* that systematically varies some of the assumptions may then be helpful.

Self-Fulfilling or Self-Negating Prophecies

In some countries such as Germany and France the early population forecasts were motivated by the fear of population decline and loss of political power. In the Netherlands overpopulation was the concern. In both cases the forecasts were seen as a means of influencing social policies (de Gans, 1999). Inasmuch as the forecasts were influential, they were self-negating.

In other cases such as regional forecasting, a forecast of population growth may cause the local society to develop the infrastructure of an area to allow for future growth. The favorable circumstances created may then turn the forecast into a self-fulfilling prophecy.

Although examples of both types of 'prophecies' exist, it has been difficult to demonstrate the causality in a historical context. Deliberate government attempts at lowering (or increasing) fertility, or lowering mortality, have seldom had long-lasting effects that could be attributed to the policies as opposed to other concurrent factors.

Forecast Errors and the Debate on Terminology

From the very start, forecasters have been aware that the future vital rates cannot be known with certainty. Evaluations of past forecasts have vividly demonstrated the unpredictability of demographic change. For example, the post-World War II baby boom was missed by all forecasters, in Europe and the United States. After the baby boom fertility, forecasts have tended to be too high. Notably, the decline of fertility by one-third in Central and Southern Europe in the 1980s, came as a surprise.

Mortality has declined faster than anticipated. In his 1926 forecast Wicksell expected that Swedish life expectancy would increase from 58 years to an eventual maximum of 69.5 by 2025, but in 2010 life expectancy was already 83.5 for women and 79.5 for men. Comparable errors occurred in many other countries. As another example, the purported maximal life spans for humans have often been exceeded by the time they have been published (Oeppen and Vaupel, 2002).

To safeguard their reputation some forecasters have occasionally tried to communicate the uncertainty by using the terms 'projection' or 'illustrative projection' for the results of their calculations. This is a reasonable practice, for example, if the calculation consists of a c-c forecast with constant vital rates, when no one expects the rates to remain constant. However, as pointed out by H. Dorn in 1950, the producers of official forecasts typically publish calculations that are based on what they consider as the most likely future vital rates, and users also interpret the results in that way. Indeed, we may define a calculation based on what is regarded as the most likely assumptions as a *forecast*.

Statistical Uncertainty in Forecasting

Although probabilistic terminology is used in many branches of demography (e.g., probability of death in life tables) (*see* Life Table), and experience has shown that forecasts of the future population have been highly uncertain, the forecasting models considered so far have all been deterministic. The recognition of the uncertainties in probabilistic terms opens up new possibilities to enhance the usability of the forecasts.

Uncertainty in the Events and in the Rates

Consider a childbearing population of size V. Suppose the number of births during a year is B. We can think of B as a sum of V random experiments with an average probability f of a birth. The expected number of births is $E = fV$. Based on a Poisson approximation one can estimate that in a sequence of B values the relative variability (coefficient of variation) is $E^{-1/2}$. In a small regional population we may have $E = 100$, for example. Then, the births vary by about 10% each year due to random variation alone. However, if $E = 10\,000$, they only vary by about 1%, and if $E = 1\,000\,000$ they vary by 0.1% each year. It was pointed out by J. Pollard in 1973 that much larger relative variations are observed in the numbers of births in different countries, so this type of Poisson variation does not account for the variability of the observed birth processes.

The situation changes if we allow f to change from year to year with a coefficient of variation c (or by $100c$ percent). Then the relative variation is at least c, irrespective of the expectation E. The conclusion is that for forecasting large populations, the uncertainty lies in the randomness of the rates rather than in the Poisson type variation in the events themselves.

Logical Flaws in the High and Low Variants

Over the years it became customary to produce high and low variants, in addition to the most likely medium variant, to give the forecast users an indication of the plausible range of the future developments. Users have typically ignored the high and low variants. One reason has been the added complexity of having more than one number in a planning context. More fundamentally, an additional reason may have been that the high and low variants have not had a logically coherent interpretation. There are two reasons for this. First, suppose $[B_L, B_H]$ and $[D_L, D_H]$ are high–low intervals for births B and deaths D, respectively. In general, there is no guarantee that the two intervals cover an equal range, because no probability content for the intervals has traditionally been given. Any calculation based on intervals with incomparable ranges is difficult to interpret. Second, suppose that the intervals are given a probabilistic interpretation, so that the probability is, for example, 80% that each one of them contains the true value. This was first suggested by L. Törnqvist in 1949. Suppose there is no migration. Then the increase of the population is $I = B - D$. Unfortunately, one can show that the usual high–low interval $[B_L - D_H, B_H - D_L]$ for I has a higher probability than 80% to cover the increase. Only if the number of births and deaths have correlation -1 does the high–low interval have probability content 80%. But the assumption of perfect negative correlation is known empirically to be false.

In a traditional c-c forecast a very large number of vital rates are given high–low variants, and these are combined in complex ways to produce high–low intervals for population sizes at different ages. No uniform interpretation can be given to the latter intervals. For example, a United Nations forecast of the total world population published in 2010 gives a high–low interval of [7.9, 8.8] billions for the year 2030 and [8.1, 10.6]

for 2050. Since it is not known what probability contents these intervals are intended to have, it is not clear how they should be interpreted.

Expressing Uncertainty via Predictive Distributions

Modern probability theory provides a language that can be used to describe the uncertainty of the future vital rates and vital events in a consistent manner. This leads to a *predictive distribution* of the future population given what we have observed in the past.

The availability of fast computing allows the use of stochastic simulation to describe numerically the predictive distribution. In practice, this is done so that a set of age-specific vital rates is sampled from their respective predictive distributions. A c-c calculation is then made based on the sampled values to produce a possible *sample path* of the population in each age and sex group. This is repeated thousands of times and the results are stored. For any age–sex group, or for any larger population aggregate, one can represent the uncertainty of the forecast in terms of graphics (histograms, box plots, etc.), or in terms of statistical summaries (means, medians, standard deviations, etc.). Such techniques have been available for single region populations since the 1990s, and examples of their application for multiregional (or multistate) populations, household forecasts, and other functional forecasts have appeared in the 2000s.

The key element in such calculations is the formulation of the predictive distributions of the vital rates based on statistical analysis of the past data (Alho and Spencer, 2005). This can be based on formal time series methods of a frequentist or Bayesian type (Booth and Hyndman, 2008; Girosi and King, 2008), the analysis of the accuracy of the past forecasts (Keilman, 1990), and judgment (Lutz et al., 1998). Each method entails its own strengths and weaknesses.

Official stochastic population forecasts have been published in the Netherlands from 1999. The United Nations has produced probabilistic forecasts for all countries in the world since the 2010 World Population Prospects (Raftery et al., 2012).

Uses of Probabilistic Forecasts

Predictive distributions provide the forecast users with a richer view of the possible future developments than the traditional population forecasts. For users needing a point forecast only, one would typically offer the median of the predictive distribution. At the other end, a complete predictive distribution is available to a sophisticated user who may want to use it in macroeconomic calculations or *risk analyses*. Stochastic forecasts have been used in Europe and the United States in the analysis of the financial stability of pension funds and other social security systems. In practice, such applications require that suitable macroeconomic tools (e.g., generational accounting or computable overlapping generations models) are used in the context of a simulated predictive distribution of a population (Alho et al., 2008).

See also: Aging and Health in Old Age; Demographic Measurement: General Issues and Measures of Fertility; Demographic Techniques: Small-area Estimates and Projections; Family Theory: Economics of Intergenerational Relations; Fertility Theory: Theory of Intergenerational Wealth Flows; Forecasting; International Migration; Period and Cohort Analysis in Demography; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Ahlburg, D.A., Land, K.C., 1992. Population forecasting. International Journal of Forecasting 8 (Special issue).

Alho, J. (Ed.), 1997. Special section on statistical analysis of errors in population forecasting and its implications on policy. Journal of Official Statistics, vol. 13, pp. 201–304.

Alho, J.M., Hougaard Jensen, S.E., Lassila, J. (Eds.), 2008. Uncertain Demographics and Fiscal Sustainability. Cambridge University Press, Cambridge.

Alho, J.M., Spencer, B.D., 2005. Statistical Demography and Forecasting. Springer, New York.

Andrews, G.H., Beekman, J., 1987. Actuarial Projections for the Old-Age, Survivors, and Disability Insurance Program of Social Security in the United States of America. Actuarial Education and Research Fund, Itasca, IL.

Booth, H., 2006. Demographic forecasting: 1980 to 2005 in review. International Journal of Forecasting 22, 547–581.

de Gans, H.A., 1999. Population Forecasting 1895–1945: The Transition to Modernity. Kluwer, Dordrecht, The Netherlands.

Girosi, F., King, G., 2008. Demographic Forecasting. Princeton University Press, Princeton.

Hyndman, R.J., Booth, H., 2008. Stochastic population forecasts using functional data models for mortality, fertility and migration. International Journal of Forecasting 24, 323–342.

van Imhoff, E., Keilman, N.W., 1991. Lipro 2.0: An Application of a Dynamic Demographic Projection Model to Household Structure in the Netherlands. Swets & Zeitlinger, Amsterdam.

Keilman, N.W., 1990. Uncertainty in National Population Forecasting: Issues, Backgrounds, Analyses, Recommendations. Swets & Zeitlinger, Amsterdam.

Keyfitz, N., 1982. Can knowledge improve forecasts? Population and Development Review 8, 729–751.

Land, K.C., 1986. Methods for national population forecasts: a review. Journal of the American Statistical Association 81, 888–901.

Lee, R.D., Tuljapurkar, W., 1994. Stochastic population forecasts for the U.S.: beyond high, medium and low. Journal of the American Statistical Association 89, 1175–1189.

Lutz, W., Vaupel, J.W., Ahlburg, D.A., 1998. Frontiers of population forecasting. Population and Development Review 24 (Suppl.).

National Research Council, 2000. Beyond Six Billion. Forecasting the Worlds Population. National Academy Press, Washington, DC.

Oeppen, J., Vaupel, J., 2002. Broken limits to life expectancy. Science 296 (5570), 1029–1031.

Raftery, A.E., Li, N., Sevcikova, H., Gerland, P., Heilig, G.K., 2012. Bayesian probabilistic population projections for all countries. PNAS 109, 13915–13921.

Smith, D.P., 1992. Formal Demography. Plenum, New York.

United Nations, 2010. World Population Prospects: The 2010 Revision. http://esa.un.org/wpp/Excel-Data/population.htm.

Whelpton, P.K., Eldridge, H.T., Siegel, J.S., 1947. Forecasts of the Population of the United States 1945–1975. U.S. Government Printing Office, Washington, DC.

Population Geography

Adrian J Bailey, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Population geography grew as a subdiscipline of human geography to describe, analyze, and reflect upon the geographical organization of human populations. Contributing research on migration, population/environment/development links and transition theory, the growth of multidisciplinary approaches reflect broader interests in social difference, lifecourse, and population issues under a neoliberal and restructuring capitalist global system. Emerging discussions of transdisciplinarity may connect scholars inspired to research the interdependencies of societies' demographic, spatial, and political organization, and reinvigorate a heritage of service and interest in social justice.

Introduction

Many commentators regard population geography as the subdiscipline of human geography that describes, analyzes, and reflects upon the geographical organization of human populations (Trewartha, 1953; Clarke, 1965). Since Brunhes' preliminary discussion in 1910, the scholarship of population geography has varied along three axes: the degree to which studying the areal differentiation and geographic diversity of population could shed light on the unitary and holistic nature of place; the degree to which population was the apex of geography's concerns; the manner in which the field would coproduce knowledge with other academic specialisms (Bailey, 2005). In practice, it has been this last dimension that has stimulated the most debate. Hence, in reviewing the field's major contributions to social and behavioral knowledge, I emphasize how the field's traditional embrace of cross-disciplinary exchange between geography and demography has been supplemented by multi- and transdisciplinary approaches.

Spatial Demography

Clarke (1965: p. 2) argued that a clear disciplinary division of labor encouraged cross-disciplinary exchange: "[with its] emphasis particularly on space … population geography distinguishes itself from demography, which is the science of population viewed as a single topic." Thus oriented, spatial demography explored geographic variations in population, migration, population systems, and the role of modernization in population change.

Population Distribution

Geographic tools including visualization, areal classification, and spatial decomposition illuminate the diversity of population and the complexities of its uneven growth. Demography suggested that factors of fertility, mortality, and migration (from the balancing equation) and age–sex composition (from stable population theory) contributed to population diversity (Woods, 1982). Normative classifications used to further explore the components of diversity, including race, nationality, religion, education, occupation, family structure, urban–rural residence, development level, and customs (Trewartha, 1953: p. 88), gradually took center stage as objects of analysis (cross-reference to geography of youth, geography of race, children's geographies).

Visualizing population diversity continues to stimulate hypothesis testing. For example, neighborhood maps of the changing distribution of minority groups accompanied research on segregation and processes of population turnover. John Snow's map of cholera deaths surrounding water pumps in Victorian London's Soho district highlighted a cluster of cases and promoted further study on how one hand pump influenced the nature of disease transmission. As fresh public and commercial challenges arise, and as Censuses and surveys contribute new data, the production of such population maps proliferates. Recent applications draw attention to uneven life expectancies and the geographies of missing girl babies, female infanticide, aging, HIV-AIDS deaths and HIV-orphans, and childhood social exclusion. Geodemographic applications consider, for example, public and commercial facility location, emergency service management, and lifestyle marketing.

Demography and Economics of Human Migration

Among the three demographic components of change, migration receives the most attention in population geography. Migration flows are typically classified on the basis of motive (for example, voluntary vs forced, economic vs environmental), distance moved (for example, residential mobility vs international migration), and direction moved (for example, new vs repeat or circular). Statistical analyses use demographic rates (for example, migration rate) and spatially referenced indicators (for example, in- and out-migration, net migration, migration efficiency, and migraproduction rate, Clark, 1986).

Residential Mobility

Residential mobility is an influential component of local population growth and decline, with implications for service provision (for example, schools, clinics) and neighborhood change (for example, gentrification). The generally observed exponential distribution of the distance of such moves is cited as support for Ravenstein's observation that most moves are

close moves, and continues to stimulate social physics-inspired gravity modeling of migration flows. This stimulated the use of gravity models based on the social physics concept that interaction is a function of the size of places migrants move between and their relative proximity. Moreover, the commonly observed suburban-wards directionality to much (but not all) residential mobility was consistent with social ecological precepts used by Chicago-school sociologists. Less deterministic explanations emphasized the joint role of housing and family factors which, as Rossi and, later, Lee noted, unfolded in systematic ways through the life cycle. The role and price of information about opportunity and constraints emerged as a powerful locus of research, which accelerated the spatial application of microeconomic and push–pull frameworks. Cross-fertilized by Morrill's ideas about migration information fields and Wolpert's writing on place utility, Brown and Moore acknowledge the dynamic trade-offs that individuals make around such factors as accessibility to employment, housing, schools, and amenity. Such insights extended life cycle approaches by pointing to the strategies individuals used to manage risk and disappointment and learn-by-doing.

Regional Migration

Uneven and fluid patterns of regional growth, rural depopulation, urbanization, suburbanization, counterurbanization, nonmetropolitan growth, and the rise of ghost cities all have significant consequences for economic and political structures. In analyzing the distribution and dynamics of gross and net migration flows between regions, gravity models have been upscaled to consider regional structures (competing destinations), and nationally comparative work has pioneered the application of entropy theory to migration. Research has also developed accounting frameworks and multiregional models referencing cohort-based approaches themselves built on the concept of migration channels, which John Fraser Hart earlier used to explain patterns of African-American urbanization in the United States. The use of increasingly sophisticated estimates of disaggregated life tables, transition probabilities, and estimates of fertility and mortality for key subgroups provides critical information to local and regional governments on, for example, the impact of aging scenarios upon regional economies and housing demand.

Growing interest in different approaches to labor market policy stimulate research on the labor market behavior of particular groups of migrants, including unemployed and family migrants. Sjaastad's influential human capital account of migration proved flexible enough to explain why young people and some unemployed were more likely to migrate (longer time horizon over which to gain benefits), to inspire work on sex differences in returns to migration (Mincer's family migration framework hypothesized the wife's sacrifice), and to explain why many third world residents moved to areas of high unemployment (the Harris–Todaro model of rural to urban migration emphasized expected returns rather than actual wage differences).

International and Circular Migration

Long distance human migration is closely associated with differences in political cultures (for example, US, Canada, and Australia narrate themselves as 'classic' countries of immigration, and fiercely debate modes of migrant belonging including assimilation and multiculturalism) and development outcomes. The field's accounts of such migration are based on movements across officially recognized borders of nation-states, and have been organized into quite disparate camps. Refugees (more generally, forced migrants) were distinguished from economic migrants, who were themselves subdivided by level of skill (skilled and unskilled) and legal status (for example, guest workers, the so-called illegal/undocumented/irregular migrants). Explanations upscale accounts of, for example, skilled migrants being channeled between global cities, and refugees being propelled like billiard balls following the social physics logic of kinetic energy. International circulation assumes a commitment to return migration and is linked to the need for seasonal labor in livelihood systems that often predate colonially imposed borders. In these accounts, migrations are seen as outcomes of economic and political arrangements and as having limited impact upon these systems. Recent attention focuses on the expanding population of international student migration and international return migration, particularly among the second generation.

Population Systems

Widely applied in population geography, general systems theory calls attention to the interrelations of demographic flows and events within a given geographic area. Akinlawon Mabongunje explains why he felt a systems application to rural to urban migration in Nigeria could extend orthodox explanations: "one of the major attractions of this approach is that it enables a consideration of rural–urban migration as no longer a linear … cause-effect movement, but as a circular, interdependent, progressively complex, and self-modifying system" (1970: p. 16). His rural–urban migration system considered urban workers, urban consumers, communication systems, government development policy, agricultural practices, education policy, and social development. A similar approach has enabled research to analyze the links between internal (regional) migration and immigration, and how this affects the regional location and nature of new immigrant destinations, the changing nature of segregation, and the direction of causality between migration and labor market change. A cross-national inventory of local and regional accounts of new immigration communities has arisen, alongside more system-oriented analyses of the changing role of regions, gateways, and spatial structure in a national economy.

Systems approaches also stimulate research on links between population and environment. At a national and regional scale, this includes the neo-Malthusian 'Limits to Growth' model, and the 'IPAT' framework, which models impact as a function of population, affluence, and technology in a region. At a regional and local scale, John Hunter's field-based analysis of the Nangodi region of Ghana revealed that, between 1948 and 1960, population declined by 15%, despite the availability of medicines and falls in mortality. The immediate cause was sex selective outmigration to southern cities, triggered by overpopulation. This arose as river blindness spread into Nangodi and reduced the cultivable land area, putting intense pressure on the regional livelihood system

(reviewed in Bailey, 2005: p. 100). This approach resonates more generally with Ackerman's notion of resource regions, and emphasizes how regional systems contain interdependencies between population, resources, and technology that can be offset through the movement of demographic (migration) and nondemographic (trade) factors of production.

Modernization and Population Transitions

Population geography adds a spatial dimension to accounts of how modernization leads to systematic population transitions. The classic demographic transition model describes a hypothetical sequence of changes in a country's population from a high fertility–high mortality–low growth rate regime to a low fertility–low mortality–low growth rate stage. In postulating two intermediate phases, where mortality falls more quickly than fertility, and overall growth increases, studies have focused on the diffusion of modern sanitation and antibiotics, and processes of secularization (for example, the adoption of a norm of a smaller family in urban society). By regionalizing areas (countries) with similar mortality and fertility dynamics, scholars establish expected speeds of transition and draw attention to deviations from such norms. For example, noting transition disparities in societies with below replacement fertility, rapid immigration, demographic aging, where fertility has declined with no increase in economic development, and where mortality has increased to levels seen a 100 years ago, research now attends to the mediating role of population composition, including family structure, ethnicity, and age profile. Discussions of the so-called second demographic transition are located within a broader agenda considering demographic convergence. Critics point to the value-laden assumption that Western concepts of secularization and ideal family size apply universally, without reference to local culture.

Modernization affects uneven spatial patterns of mortality and morbidity through developments in nutrition, manufactured products, medical advances, public health provisions, and patterns of disease distribution. Research considering the case of HIV-AIDS has greatly complicated the glib assertion that the diffusion of medical technology plays the driving role in the epidemiological transition. As he vividly illustrates for Haiti, Paul Farmer (1999) argues diffusion is a spatial process that is socially mediated: "AIDS in Haiti is all about proximity not distance. AIDS in Haiti is a tale of ties to the United States…about steep grades of inequality, both local and transnational. AIDS in Haiti has … to do with the pursuit of trade and tourism in a dirt-poor country" (p. 126).

Zelinsky's mobility transition similarly investigates the systematic rhythms of migration that accompany the onset of modernization. He tracks shifts from a localized population with little internal movement, through a phase of falling mortality and rising growth rate, to a more urban system with much rural to urban migration (and frontier migration in developing economies, Bailey, 2005: pp. 92–94). As one idea of modernization diffuses – the idea that spatial and social mobility can be substituted – he argues urban to urban migration also diminishes, to be replaced by circulation. His framework stimulates research on, for example, how and when rural fertility regimes change in response to outmigration, and how communication technology absorbs geographic mobility.

Population Studies

The rise of multidisciplinary accounts has accompanied the wider use of the term population studies by many practitioners, and a set of critiques about the empiricist and positivist bases of spatial demography (White and Jackson, 1995). One concern is the paradox that, despite being about population, people (their agency, their suffering, etc.) are often absent from many accounts. Moreover, the use of normative classifications and a focus on events may obscure the complexity of social processes, and lead to unhelpful essentialization of population groups. Similarly, the field's treatments of space can be superficial and uninspired. Taken together, and particularly influenced by feminism and neo-Marxism, an emerging multidisciplinarity in population geography shared at least three features: a desire to give greater attention to debates within human geography, particularly about the role of space and place in social life; an appetite for pluralism and the use of concepts, theories, and methods from areas such as sociology, anthropology, political science, gender studies, postcolonial studies, critical race theory, and development studies; a tolerance of thematic specialization. While by no means exhaustive, four areas where such multidisciplinarity grew include population geographies of the lifecourse, environment and development, difference, and sociospatial relations.

Lifecourse

Partly inspired by behavioral and feminist critiques of the deterministic role of age in life cycle models, population geographies of the lifecourse emphasize the patterns of order and order of patterns in the practices of everyday life. Like Hagerstrandian and time geographic approaches more widely, the lifecourse provides a framework for understanding interdependencies between: population events (including the act of moving, partnering, becoming unemployed, graduating); careers (including residential locations, migration decisions, employment, family life); linkages between lives and between institutions; the learned and social nature of time. The fecundity of such a lens is suggested by, for example, growing interest in the process of studentification as lifecourse planning. This debate straddles the intersection of formerly separate categories of analysis (life cycle nest-leaving, rental housing choices, international migration), while powerfully shaping the way housing markets function, the acquisition of cultural capital by transnational families, changing subject positions within extended households, immigration policy, and community relations. Another emerging debate on time squeeze references diverse factors like the rise of dual career couples, traffic congestion, the sandwich generation, and social media and studies how the changing expectations of, and realities of coordination and synchronization reproduce gender relations, age relations, and familialism.

A connected thread of lifecourse research explores the structural nature of population inequalities by examining how events and experiences at one time impact subsequent outcomes. Comparisons between population groups (cohort-based by socialization, race and ethnicity-based, class-based etc.) show how context affects the ways in which early life medical conditions, education investments, and migration

behaviors affect later life opportunity and outcomes. The approach draws attention to how healthy places reduce morbidity, defamilialization and how population experiences vary across welfare regimes, how visa regulations affects family reunification and ongoing subject positions within families, and how population events become embedded and transmitted as cultural survival strategies in ecologically dynamic regions such as the Sahel.

Environment and Development

Multidisciplinary analyses have extended traditional notions that population events are crisis responses to environmental changes and that population phenomenon can independently trigger ecological transformation. Whether narrated as a population–environment–development nexus, a population–environment nexus, a population–development nexus, or a migration–development nexus, the field uses ideas from political ecology about the dialectics of nature society relations, resilience and adaptability, and the nature of ecological regulation by the state and other bodies (Findlay, 2005). Research analyzes the vulnerabilities that populations experience in and through their ecological settings and practices, and the structures of adaptability and resiliency that emerge.

Explorations of the cultural and interdependent nature of the population geographies of environment and development also accompany the rise of a more progressive cultural ecology in the field. This studies how ecological change can influence gender, class, and ethnic/racialized positions, with implications for population structures. For example, Gray and Mueller (2012) use rural-household survey data to show how place-based gender norms mediate the migration response to drought in Ethiopia. They report that male outmigration increases for livelihood reasons but the same local shortages of capital restrict the potential for dowry payments and limit female marriage migration. Indeed, as situations of ecological collapse may require young males to outmigrate (and remit) or be deployed to more risky means of livelihood, either response changes their gendered position of influence, such that ecological, class, and gender systems are recalibrated. Careful ethnographic and empirical analysis reveals similar cultural transformations in the context of forced displacement from dam and infrastructural projects.

In exploring the cross-cultural ways in which adaptability and resilience is understood and transmitted through norms and practices, population geography draws attention to how local discourses may counter hegemonic, often neoliberal understandings of how certain population groups may need 'protection' from hostile environments and, more controversially, from their own 'local' practices. The field similarly contributes a more spatial and politically aware discussion of population as 'the elephant in the climate change room.' Research attends to how spatial context affects the interdependencies between climate change inputs, including droughts, floods, extreme events, shifts in the ecumene, and population, migration and, in turn, geopolitical consequences (for example, famine-induced conflict). More broadly, the need to carefully model complex relations between environmental, economic, and social systems has stimulated the development of Geographical Information Science and spatially disaggregated, longitudinal, and historical climate models. Through offering sophisticated analyses of ecological, population, and cultural interdependence, the field can counter overly economic and neo-Malthusian accounts of climate change and may provide ammunition for those who wish to diffuse hysterical media hype around, for example, floods of environmental refugees and climate-induced wars.

Difference

Inspired by feminist and critical interpretive approaches, the field contributes to understandings of how population categories, including events, groupings, and distributions, are social constructions rather than normative structures. This project is intimately tied to humanities research on the nature of social difference, including gender, race, ethnicity, age, sexuality, and disability. One thread looks at how families, states, and other social institutions originate and mediate discursive productions, which affect identity, with research-exploring gendered migration (Silvey, 2004) and the place-based othering of queer communities, asylum seekers, and Romani travelers (Tyner, 2009). For example, Anderson (1987) shows how an idea of Chinatown emerges as a state constructed population category, which circulates further, explicitly racialized discourses.

Ongoing research on the ascribed identities of diasporans accents their subject positions as cosmopolitan elites and patriotic remitters, and probes the racialized manipulation of discourses of belonging and home that states regard as influential in exploiting their economic contributions. The intensification of transnational, cross-border activity has significant consequences for gender, ethnic, racialized, faith-based, family-based, and political identities, which in turn transform social structures (including split families, astronaut families), cultural structures (including ethnoscapes, faith-based community networks), economic structures (notably the impact of remittances upon livelihoods, communities, and the growth of the migration–development nexus), and political structures (including institutions of governance, multiculturalism, citizenship, and the trajectory of the European suprastate). Regarding citizenship, the intersection of gender, ethnicity, and class (employment type) brings systematic disadvantage for women migrants in accessing family reunification provisions. Indeed, such research on intersectionality has stimulated further interest in *how* identities come together in particular spaces and places at particular times (Silvey, 2004).

Most widely, population geographies of difference accent the role of space in negotiating identity, intersectionality, and for promoting identity politics. For example, discussions of segregation and NIMBYism have laid bare the social norms held by the public and state concerning how certain population groups should organize themselves spatially and temporally. In the UK, the multidisciplinary debate about the extent to which minority groups and natives lived 'parallel lives' and society was 'sleepwalking into segregation' turns on spatial expectations about the speed of assimilation and the persistence of the so-called ethnic ghettoes (cf underclass debate in the United States). The relevance of context is also shown by work on the uneven embodiment of transmigrant life, and how transnational networks are not blank deterritorialized canvasses, but

actively reterritorializing networks of networks (Dunn, 2010). Such emplaced performances have been studied through accounts of visual and material cultures in the home, and comparative analyses of language use in home and classroom settings.

Sociospatial Relations

Research on the sociospatial relations of population geographies locates the recursive connections between identity/belonging and space/place/network within broader concepts of capitalist restructuring. Population organization and strategy has been tied to capitalist rescaling in work on everyday transnationalism. Examining Chinese globalization Pieke et al. (2004: p. 162) show how Fujianese migrants create a transnational social space from below that is rooted to both China and Europe: "through practice and discourse [migrants] are tied to a transnational social space that spans the globe…[and] configures the perspective through which migrants look at the geographic locations that migration brings them to." By changing geographic perspectives migrants rescale the networks connecting China and Europe, and have the potential to shift neoliberal political structures (cf the more deterministic treatment of flexible citizenship in other accounts of Asian mobilities under exceptional neoliberalism, Bailey, 2005). The broader point, expressed cogently by Ley in his analysis of the performed sociospatial relations of trans-Pacific migrants (2010), is that migrant rescaling goes beyond the crude binary categorization of globalization, and free mobility versus localism and constraint. Such transnationalism is not just 'from below' or 'from above' but always situated, placed, and transformative.

Similar themes run through research on the population geographies of vulnerability. Populations exposed to structural risk, be it economic (contemporary global recession), cultural (hate crimes and xenophobia), or environmental (for example, the Indian Ocean tsunami) reproduce sociospatial relations, which transform political, social, and religious institutions; the nature of borders and bordering processes; and shape capitalist restructuring (Findlay, 2005). Another debate examines if and how migration and state constructions of overseas workers respond to, or drive the structure of labor markets and systems of global cities (Tyner, 2009). In contrast to orthodox research that argued skilled migrants moved according to human capital and contractual dictates, and in the direction of global capital, analyses of the sociospatial relations of contemporary international mobility shows distinctive links between skilled and unskilled migration flows, the rise of global care chains, the appearance of a migrant division of labor, distinctive second generation return mobility, and transformations in the nature of work that undermine the long-held distinction between temporary/informal and permanent/formal labor.

Transdisciplinary Population Geographies

Transdisciplinarity aspires to transformational conceptual engagement between and across disciplines and knowledge partners, and often uses case-driven methods to reorient the politics and ethics of knowledge. While method was ontologically unproblematic for spatial demography, pluralism and specialization associated with multidisciplinarity made it harder for practitioners to articulate a cohesive position on population matters. With the cultural turn in human geography igniting a debate about relevancy, signal events like 9/11 derailed an abstract and dry globalism that proclaimed the local and global interdependence of lives and rerouted it as a globalism bristling with menacing and threatening population issues, including disaffected immigrant youth, ecological refugees, global pandemics, NEETs, and a pensions crisis. Within the field, Philo (2005) urges scholars to not just pay lip service to ideas such as vulnerability, but to inspect what lies beyond the label, i.e., chains of influence and power. Tyner (2009) imagines new solidarity networks around topics, themes, forms of measurement, across disciplines, and between universities and their local communities. In practice, and using preliminary examples concerning violence, and bodies, transdisciplinary population geographies have acknowledged the global, cross-cultural production of knowledge, the possibility of an emerging global consciousness, and the need for an ethical intervention – a view from somewhere.

Violence

Drawing on different elements of the diverse corpa of Michel Foucault and, to a lesser extent Henri Lefebvre, a number of population geographers have studied the systematic, historicized, and placed relationships between population, inequality, and power. A growing number of studies examine how members of population groups, including migrants, experience vulnerability arising from security concerns. Post-9/11, post-transnational fluidity, and under global recession, such others are not just different, but threatening. Mountz (2010) explores why one group of migrants – those smuggled – come to be seen as 'exceptional,' and what this means for their lives. She uses a bespoke institutional ethnography to 'study up' how the daily actions and narratives and silences of individuals associated with state regulation of immigration and asylum in Canada have shifted the population geographies of protection for smuggled asylum seekers. Her framework draws widely on literature in geography, anthropology, sociology, and political science to interrogate 'the long tunnel thesis' whereby 'arriving migrants' (i.e., those interdicted in Canadian waters) would be treated as if they were walking through the long underground corridors of an airport, always approaching but never quite landing (Mountz, 2010: p. xiv). Tyner's treatment of violence seeks to: "emphasise the production of knowledge claims that legitimize and justify the perpetuation of violence through the marginalisation of bodies and populations" (Tyner, 2009: p. 45). Tracing the manifestation of violence through war, he considers the cases of rape of Tutsi women, the killing fields of Cambodia, and the body count in Viet Nam. He extends the Foucauldian concept of biopower by combining anatomopolitics (separation and discipline through classification) and biopolitics (including the exercise of state power through manipulations of what counts as a population and what counts as a population event, including death).

Bodies Making Population

By holding that emplaced and practiced performances rework social and cultural norms and relations between constituting institutions researchers reposition populations and their acts (things like childbirth, moving, staying put, dating, being represented) as questions of bodies in/through space. A focus on embodiment points out that bodies do not just appear/happen, but come with messy materialities, such that space and place is a lived and vital part of any population (act/being). In the recent Body/State, Cameron et al. (2013: p. 2) further interrogate the roles of states, noting: "this book has come into being because … there has been a proliferation of ideas, insights and knowledge surrounding the biopolitics of bodies across an ever-widening range of social, political, and academic concerns." Their empirical matter discusses female circumcision, hunger strikes, organ transplants, body monetization, male sex work, the disabled body, picketing and unruliness, home-schooling, the biopolitical regulation of liminal transsexual bodies, pitied bodies, unmapped bodies, and virtual bodies. For Longhurst et al. (2009), participating with migrants in New Zealand in cooking a symbolic dish illuminated daily experience as a set of embodied and scaled relations of body–mind, biology–social, edible–nonedible, here–there, and human–nonhuman. Such research has the potential to expand the dimensions on which difference – but also connection – is understood, with Rosi Braidotti offering a tantalizing glimpse of how embodied and emplaced subjects are critical to counter contemporary, false proclamations of social progress. Noting that "advanced capitalism has consumed and subsumed the old gender system," and drawing on Deleuze's concept of empirical transcendental, her nomadic theory calls for research on: "power relations that construct technologies as liberating while using them for the most confining, profit-making, war-seeking, and sexually-conservative ends" and focuses on bodies as nomadic subjects that reject liberal individualisation and connect with respect for complexity and interconnection – "a collectively-oriented, externally-bound, multiple subject" (2011: p. 79).

Conclusion

Ebbs and flows in the popularity of the term 'population geography' mask a persisting commitment to, and ever wider production of spatial demographic knowledge about populations in their spatial settings. Yet, as academic divisions of labor evolve, and public and political consciousness about population softens and hardens, multidisciplinary and transdisciplinary scholarship further expands the approaches, topics, and ethics of the field. It perhaps seems fair to suggest that population geography is becoming a community of practice that can connect scholars inspired to research the interdependencies of societies' demographic, spatial, and political organization, significantly drawing on a heritage of service and interest in social justice. Drawing together diverse knowledge from diverse places, all the time making a difference, may require new approaches to activities many may consider beyond the scope of a subdiscipline, including having impact in a neoliberal landscape, and inspiring students with discussions of peace and activism.

See also: Children's Geographies; Global Migration; Migration, Economics of; Political Geography; Race and Racism, Geography of; Scale in Geography; State, Geography of; War and Terrorism, Geography of; Youth, Geography of.

Bibliography

Anderson, K.J., 1987. The idea of chinatown: the power of place and institutional practice in the making of a racial category. Annals of the Association of American Geographers 77 (4), 580–598.
Bailey, A.J., 2005. Making Population Geography. Hodder, London.
Braidotti, R., 2011. Nomadic Theory. Columbia University Press, New York.
Cameron, A., Dickinson, J., Smith, N. (Eds.), 2013. Body/State. Ashgate, Farnham.
Clark, W.A.V., 1986. Human Migration. Sage, Beverly Hills.
Clarke, J.I., 1965. Population Geography. Pergamon, Oxford.
Dunn, K., 2010. Embodied transnationalism: bodies in transnational spaces. Population space and place. Special Issue 16 (1), 1–83.
Farmer, P., 1999. Infections and Inequalities: The Modern Plagues. University of California Press, Berkeley.
Findlay, A.M., 2005. Vulnerable spatialities. Population space and place. Special Issue 11, 429–439.
Gray, C., Mueller, V., 2012. Drought and population mobility in rural Ethiopia. World Development 40 (1), 134–145.
Ley, D., 2010. Millionaire Migrants: Trans-Pacific Life Lines. Wiley Blackwell, Oxford.
Longhurst, R., Johnston, L., Ho, E., 2009. A visceral approach; cooking 'at home' with migrant women in Hamilton, New Zealand. Transactions of the Institute of British Geographers 34, 333–345.
Mabogunje, A.L., 1970. Systems approach to a theory of rural-urban migration. Geographical Analysis 2, 241–260.
Mountz, A., 2010. Seeking Asylum: Human Smuggling and Bureaucracy at the Border. University of Minnesota Press, Minneapolis.
Philo, C., 2005. The geographies that wound. Population Space and Place 11, 441–454.
Pieke, F.N., Nyíri, P., Thunø, M., Ceccagno, A., 2004. Transnational Chinese: Fujianese Migrants in Europe. Stanford University Press, Stanford.
Silvey, R., 2004. Power, difference and mobility: feminist advances in migration studies. Progress in Human Geography 28, 1–17.
Trewartha, G., 1953. The case for population geography. Annals of the Association of American Geographers 43, 71–97.
Tyner, J.A., 2009. War, Violence, and Population: Making the Body Count. The Guilford Press, London.
White, P., Jackson, P., 1995. (Re)theorising population geography. International Journal of Population Geography 1, 111–123.
Woods, R.I., 1982. Theoretical Population Geography. Longman, London.

Population Size Estimation and Capture–Recapture Methods

Ruth King and Antony Overstall, School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by R. Cormack, volume 17, pp. 11809–11813, © 2001, Elsevier Ltd.

Abstract

The question 'How many?' arises in many scientific areas. To address this question, Capture–recapture studies are often used to partially observe the population of interest. In these studies, several sources are used to observe individuals in the population, which are subsequently collated together and the data summarized by the number of individuals observed by each distinct combination of sources. Statistical models are fitted to the observed data, permitting the estimation of the number of individuals not observed by any of the sources, and hence (when combined with the number of observed individuals) the total population size.

Introduction

The question 'How many?' arises in many different scientific areas. For example, "How many leopards are there in a given jungle/area?," "How many injecting/problem drug users are there in Glasgow and/or Scotland?," or even "How many typographical errors are there in a given text?." Unfortunately, identifying and enumerating all the given members of the population of interest is typically infeasible or often simply impossible for numerous reasons.

An early attempt to address this question of 'How many?' was made by Laplace in 1786 who wished to estimate the population of France. The underlying idea that was used is as follows. Consider a subgroup of the population of interest, where the size of this subgroup is known to be n_1. Randomly sample n_2 individuals from the whole population of interest and record the number of individuals that are also in the subgroup, denoted m_2. (Laplace used the register of births in France for the subgroup of known size, and parish registers for the subsequent random sample of the whole population.) Under some assumptions (see below), the proportion of individuals from the sampled population that are in the known subgroup should be equivalent to the proportion of individuals from the whole population that are in the given subgroup, so that

$$\frac{n_1}{N} = \frac{m_2}{n_2}$$

Rearranging the equation gives an estimate of the total population size of $\widehat{N} = n_1 n_2 / m_2$. For many applications, the subgroup itself is obtained by taking a random sample of individuals from the total population at a given capture event/occasion, which are subsequently marked to make them identifiable. This corresponds to an initial capture event, the subsequent resampling of the population corresponds to the recapture event. This description leads to the commonly used name capture–recapture (or mark–recapture).

The above estimate of the total population size is typically referred to as the Lincoln–Peterson estimate. A number of assumptions are made, including (1) the population is closed between capture occasions (i.e., there are no births/deaths/migration within this period); (2) all individuals in the population have the same capture probability at each capture event; (3) at the recapture occasion all previously observed individuals can be identified (i.e., there is no mark loss). The Lincoln–Peterson estimator is a consistent estimator of the total population size, however, it is biased for finite sample sizes. This has led to the Chapman estimator (Chapman, 1951):

$$N = \frac{(n_1 + 1)(n_2 + 1)}{(m_2 + 1)} - 1$$

which has smaller bias. See, for example, McCrea and Morgan (2014; Chapter 3 and references therein) for further discussion of the Lincoln–Peterson estimator and underlying assumptions. Note that this two sample capture method is also referred to as the dual system estimation problem.

Multiple Recaptures

The previous two sample capture–recapture experiment can be extended by taking multiple recapture events. The study design is as follows. At the initial capture occasion all observed individuals are uniquely marked and released back into the population. At each subsequent capture occasion all observed individuals are recorded, previously unmarked individuals are uniquely marked, and all individuals released back into the population. The data obtained correspond to the set of capture histories for each individual observed within the population, detailing at which capture occasions each of the individuals are observed. For example, suppose that there are 10 capture occasions and a given individual is initially observed at time 2 and subsequently observed at capture occasions 3, 4, and 9. Their corresponding capture history would be represented as:

0 1 1 1 0 0 0 0 1 0

The 0/1 indicates that the individual is unobserved/observed at the given capture occasion. This experimental design is typically undertaken with ecological studies, where the capture occasions correspond to capture events, where observers go into the field and record all individuals seen within a given period of time. However, for other populations of interest, notably within epidemiological applications, including human populations, the capture occasions typically

refer to different 'sources' or 'lists' that can observe individuals. The corresponding capture history details the set of sources that have observed each individual in the study period. Thus for such studies, the capture histories do not provide any temporal information. For example, for the given capture history above, this would correspond to an individual being observed by sources 2, 3, 4, and 9, but does not indicate the order in which they were observed by the different sources. This naturally leads to different forms of models for ecological and epidemiological studies (see below). For simplicity we will generally use the term 'source' to refer to all of the different forms of capture occasions, however, we do revert back to the use of capture occasions/events in the context of ecological studies.

Capture–recapture data are often summarized (particularly when the number of sources is small) via the form of an incomplete 2^k contingency table, where k denotes the number of sources. The factors of the contingency table correspond to the sources, each of two levels corresponding to being observed or not observed. The cell entries correspond to the number of individuals observed by a given combination of sources. The contingency table is incomplete as the number of individuals in the population not observed by any source is unknown. **Table 1** provides an example of an incomplete contingency table using three different sources (Hook et al., 1980).

Statistical models can be fitted to the observed data of the incomplete contingency table, which then permit estimation (or prediction) of the unobserved cell entry. This estimate of the unobserved cell is combined with the total observed number of individuals in order to obtain an estimate of the total population size. These statistical models are dependent on the form of the sources used (i.e., epidemiological capture sources or ecological temporal capture occasions) and the underlying modeling assumptions. The models are fitted to the data, and inference obtained, either via the calculation of maximum likelihood estimates (MLEs) or via the Bayesian posterior distribution of the model parameters with associated posterior point estimates presented, such as the posterior mean or median (see below for further discussion of the Bayesian approach). The estimate of the total population size is usually presented with an associated uncertainty interval, namely, a confidence interval for the classical approach or a credible interval for the Bayesian approach. We discuss the associated models in more detail next.

Table 1 An example of a 2^3 contingency table for babies born with spina bifida in New York State between 1969 and 1974 presented by Hook et al. (1980). Source 1 corresponds to birth certificates, source 2 to death certificates, and source 3 to medical records. For example, 12 individuals are observed by all three sources and 142 individuals are observed by sources 1 and 2 but not source 3

		Source 3	
Source 1	*Source 2*	*1*	*0*
1	1	12	142
1	0	112	247
0	1	4	49
0	0	60	?

Models

A range of models have been proposed to fit to capture–recapture data to incorporate different underlying assumptions, and in particular different forms of heterogeneity. We present these separately for ecological (temporal) capture–recapture studies and social/epidemiological (nontemporal) capture–recapture studies. Chao et al. (2001) provide a summary and comparison of many of these models.

Social/Epidemiological Capture–Recapture Studies

For social (or epidemiological) capture–recapture studies the capture sources often relate to different 'registers' or 'lists' that each attempt to enumerate the number of individuals in the population of interest. These often relate to hidden (or difficult to observe) populations of societal interest, such as problematic drug users, individuals with a given disease, the number of homeless individuals in an area, etc. Registers of such individuals are often compiled by a number of different agencies that can be cross-classified to form the incomplete contingency table using identifiers such as initials, date of birth, postcode, etc. to uniquely identify each observed individual. For such data the sources (or registers) are in general not independent of each other. For example, if an individual is observed by one register then they may be more or less likely to be observed by another register, leading to the idea of a two-way interaction. For any two given registers, there is said to be a positive (or negative) dependence if an individual is more (or less) likely to be observed by one register if they are observed by the other register (and vice versa). Assuming that the registers are independent will lead to an overestimation in the total population size if two sources are negatively dependent and an underestimation if two sources are positively dependent. Higher-order interactions (i.e., three-way, four-way, etc.) are also possible, that are able to represent more complex relationships between the registers. Such interactions are able to represent relationships corresponding to the propensity of being observed by one register being dependent on the combination of other registers that an individual is observed by.

Log-linear models specify the contingency table cell entries to be independent observations from Poisson distributions, where the corresponding log of the expectations are a linear function of main effect terms, corresponding to the capture propensity of each source, and interaction terms (Fienberg, 1972). In general, for such incomplete contingency tables, when there are k sources, interactions up to order $(k-1)$ may be estimated (when there are only two sources, so that $k=2$, it is not possible to consider an interaction between the sources, and they are assumed to be independent). For example, when there are three sources, denoted A, B, and C, there are three possible two-way interactions that can be estimated, corresponding to each pairwise interaction between the three sources, AB, AC, and BC. In addition, there are eight possible log-linear models, corresponding to the inclusion/exclusion of each of these interactions (see below with regard to model selection). Note that, typically, only hierarchical log-linear models are considered – these are defined such that if a given interaction is present in the model, all of the corresponding associated lower-order interactions (and main effect

terms) are also present. Graphical models and decomposable graphical models are often also considered due to their associated properties, and are special cases of hierarchical log-linear models. For a given log-linear model, the associated likelihood function can be constructed and an estimate of the total population size obtained (either using maximum likelihood or a Bayesian approach). Bishop et al. (1975) provide detailed descriptions of log-linear models for capture–recapture data.

Categorical covariate information is often collected on the observed individuals within the study. For example, gender, geographical region, and/or age-group (as noted above these are often used to uniquely identify individuals permitting cross-classification across different registers). In the presence of such covariate data, when it is believed that the different subclasses may represent different capture propensities for (and/or interactions between) the different registers, it is commonplace to stratify the data, corresponding to the set of categorical covariates. This leads to multiple incomplete contingency tables, corresponding to each combination of categorical covariates. Traditionally, each individual's incomplete contingency table is analyzed independently of each other, providing an estimate of the total population size for each combination of categorical covariates. However, stratifying in this manner can lead to 'small' contingency tables (in terms of the cell entries, including zero cell entries) and unstable estimates of population size. An alternative approach is to construct a single incomplete contingency table, where each covariate is regarded as a factor (note that the resulting contingency table has multiple unobserved cells). Log-linear models are defined in the standard way with additional main effect terms for the new categorical covariates and associated interactions (of only covariates, or covariates and registers). For example, see King et al. (2014) for such data relating to injecting drug users in England, where additional two-level covariates of gender and age-group are considered.

Ecological Capture–Recapture Studies

For ecological capture–recapture studies, there is a temporal component to the survey design, with captures of individuals associated with a given (ordered) set of capture events/occasions. Otis et al. (1978) described three possible forms of heterogeneity that may affect the associated capture probabilities: temporal (t), behavioral (b), and individual (h) heterogeneity. Temporal effects relate to variability in capture probabilities across occasions (for example, due to environmental conditions, changing survey effort, etc.); behavioral effects corresponding to a change in an individual's capture probability following their initial capture (an increase in capture probability is referred to as a 'trap happy' response; a decrease as 'trap shy'); and individual heterogeneity corresponding to variability in capture probabilities across individuals.

A common modeling approach is to specify the capture probabilities to be of logistic linear form. For example, let p_{it} denote the initial capture probability for individual i at time t; and c_{it} the recapture probability (i.e., their capture probability following their initial capture). The logistic linear model allowing for temporal, behavioral, and individual heterogeneity specifies the initial capture probabilities to be of the form

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha_t + \gamma_i$$

with associated recapture probabilities, $\log\left(\frac{c_{it}}{1-c_{it}}\right) = \log\left(\frac{p_{it}}{1-p_{it}}\right) + \beta$. The term α_t denotes the temporal effects; β the behavioral response ($\beta > 0$ corresponds to trap happy; $\beta < 0$ to trap shy); and γ_i the individual effects. Submodels, excluding a given heterogeneity is obtained by setting constraints on the parameters and/or terms equal to zero. For example, if there is no temporal heterogeneity, $\alpha_t = \alpha$ for all capture times t; if there are no behavioral effects, $\beta = 0$; in the absence of individual heterogeneity, $\gamma_i = 0$ for all individuals i. The temporal effects are essentially equivalent to the main effect terms for log-linear capture–recapture models; and the behavioral effects can be viewed as a constrained set of log-linear interaction terms. The Rasch model is a special case when there are only temporal and individual heterogeneity effects present (i.e., $\beta = 0$), which can also be viewed as a log-linear quasi-symmetry model (Fienberg et al., 1999).

For models with only temporal and/or behavioral heterogeneity, there is an explicit likelihood function. In the presence of the individual heterogeneity component, additional complexities arise. Specifying distinct (arbitrary) individual parameters (i.e., γ_i values) does not permit the total population size to be estimated (how do we know what the individual parameter values are for the unobserved individuals?). To incorporate individual heterogeneity, finite or infinite mixture models are typically assumed for the individual component terms. Finite mixture models specify the population to be composed of a set of homogeneous subpopulations, such that within each subpopulation individuals have common capture probabilities (see for example, Pledger and Philpot, 2008). For infinite mixture models (also called random effects models), the individual capture probability component is assumed to be a random sample from some distribution. The individual effect terms (i.e., the γ's in the above) are integrated out numerically or analytically, dependent on the model specification, within the corresponding likelihood expression (see, for example, Coull and Agresti, 1999; Dorazio and Royle, 2003). Additional individual-level covariate information can be incorporated within the logistic regression, permitting the capture probabilities to be dependent on additional factors (Huggins, 1989; Alho, 1990).

Bayesian Paradigm

The above statistical models can be fitted to observed data and the MLEs of the model parameters calculated, providing an estimate of the total population size. However, recently, there has been increasing interest in the development of Bayesian model-fitting approaches. The underlying philosophy of the Bayesian approach is that parameters are random variables (as opposed to fixed but unknown quantities in the classical framework). The mechanics of Bayesian inference are as follows. We begin with a prior distribution: this represents our beliefs and uncertainty about the model parameters prior to

collecting the data. Once we have collected the data, we evaluate the posterior distribution by combining the information from the data (via the likelihood function) with the prior distribution using Bayes theorem:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The posterior distribution represents our updated beliefs and uncertainty about the model parameters, in light of the observed data. Bayesian approaches have been developed for both social/epidemiological and ecological capture–recapture studies (see, for example, Madigan and York, 1997; King and Brooks, 2001, 2008).

The philosophy of the Bayesian approach means that probability statements can be used to quantify uncertainty in the model parameters. For example, a 95% credible interval for the total population size will contain the total population size with probability 0.95. These types of statements are typically easier for scientists and epidemiologists to interpret than the corresponding statements under the classical approach where model parameters are fixed but unknown (i.e., relating to confidence intervals). In addition to the interpretability, a further advantage of the Bayesian approach is its ability to incorporate (independent) external information through the prior distribution, fully incorporating all available information within a formal framework. For example, King et al. (2014) are interested in the number of injecting drug users in England. They use independent data corresponding to the number of drug-related deaths attributed to injecting drug users and associated prior expert beliefs on the injecting drug-related death rate to form a prior distribution on the total population size of injecting drug users. In the absence of any prior information on the model parameters, uninformative (or vague) priors are specified on the model parameters.

Model Selection

The estimation of the total population size is generally highly dependent on the model fitted to the data. For example, for log-linear models, this corresponds to the set of interactions that are present in the model; for logistic regression models, the heterogeneities that are assumed to be present in the capture probabilities. Thus, within an analysis of the capture–recapture data, the first step is usually to determine, or select, the model. One possible approach is to fit the saturated (most complex) model to the data. However, this often leads to low precision and unstable estimates of the total population size. Thus, in general, the idea of parsimony is advocated. This approach essentially involves finding the 'simplest' model that is deemed to adequately fit the observed data. A range of parsimonious model selection criteria exist. Within classical statistics, likelihood ratio tests can be used to compare nested models (i.e., where one model is a special case of the other model). More generally (including for non-nested models), information criteria can be used. Information criteria are defined such that there is a trade-off between the fit of the model to the observed data (usually defined via the deviance) and the corresponding complexity of the model. The most commonly applied information criterion for capture–recapture data is Akaike's information criterion (AIC), and associated variants, including corrected AIC (AICc) and quasi AIC (QAIC) for small sample sizes and overdispersed data, respectively (Burnham and Anderson, 2002). Within a Bayesian framework, models may be quantitatively compared via posterior model probabilities or Bayes factors.

Ideally, for a given model selection criterion, such as the AIC statistic, each possible model should be fitted to the data in turn. The (final) model chosen is the model that is deemed optimal for the given model selection criterion, and the corresponding estimate(s) for the model reported. However, fitting all possible models to the data is often infeasible, due to the total number of models. In these circumstances, a model search algorithm is often implemented using a 'bottom–up' approach (starting in the smallest model and proposing to add terms one at a time) or a 'top–down' approach (starting in the saturated model and proposing to remove terms one at a time), or some combination of these approaches. However, this does not guarantee that the optimal model will be chosen. Automated (and stochastic) computational algorithms have also been developed to search over the set of possible models in order to identify the optimal model.

Providing an estimate of the total population size conditional on only a single model, deemed optimal via some model selection criteria, assumes that the model identified is correct. This, in turn, may lead to an undesirable overconfidence in the estimate obtained of the total population size. An alternative approach is to provide a model-averaged estimate for the population size. Such estimates involve using a weighted average of the estimated total population size for each possible model, thus accounting for model uncertainty in the estimation of the total population size. For example, within classical analyses, AIC weights are often advocated; within Bayesian analyses the weights correspond to the posterior probability of each model. Model-averaging approaches have been applied to both social and ecological capture–recapture studies (see, for example, Madigan and York, 1997; King and Brooks, 2001, 2008).

Statistical Software

In order to implement the statistical techniques and models to analyze data, a range of (more-or-less) user-friendly computer packages have been developed (Bunge, 2013). These may be in the form of stand-alone packages, or packages developed for standard software, such as the freely available statistical software R. In particular, these include the R package conting (Overstall and King, 2014) for Bayesian log-linear analyses and the stand-alone package MARK (White and Burnham, 1999) for ecological capture–recapture models. Such packages permit the fitting of many of the standard models presented above for capture–recapture data.

Relaxing the Assumptions

Any statistical model that is fitted to data makes certain underlying assumptions. These assumptions may or may not be able to be assessed as to their validity (for example, the validity of the extrapolation to unobserved individuals, given the observed individuals is untestable). In order to assess, or remove, a given assumption, additional statistical models have been developed, extending the traditional capture–recapture models.

For capture–recapture data, the construction of the capture histories of the observed individuals, and hence incomplete contingency table data itself assumes that all individuals are correctly matched across different sources. However, what is the confidence attributed to the matching of individuals across the different sources: are any mismatches likely to occur? If so, are these more/less likely to be for given sources, or are they simply random occurrences across each source? Mismatches can lead to false positives (two distinct individuals identified as the same individual by two different sources) and false negatives (the same individual identified as two different individuals by two different sources). For social applications, an individual may have his/her name spelt differently by two different registers, leading to them failing to be matched as the same individual when the sources are collated (a false negative) or two distinct individuals with similar names being incorrectly identified as the same individual (a false positive). See Fienberg and Manrique-Vallier (2009) for a discussion of these issues. Within ecological capture–recapture data studies, traditionally physical marks have been applied to individuals so that they are uniquely identifiable. However, photo identification and DNA profiling are becoming increasingly popular due to their noninvasive nature but may lead to false positives/negatives (see, for example, Wright et al., 2009 with regard to DNA genotype matching).

The sources used to obtain the capture histories may also raise issues. For example, individuals observed by the given registers are assumed to be identifiable as to their membership of the given population of interest. However, this may not always be the case. For example, Overstall et al. (2014) use a hepatitis C virus database as a capture–recapture source for identifying injecting drug users. This database identifies injecting drug users as those individuals who specify the potential source of contracting hepatitis C as injecting drugs. However, the population of interest is that of current injectors only, so that this database also records past injectors who are no longer injecting, and hence not in the population of interest. Standard capture–recapture models can be extended to directly model such additional data complexities (Overstall et al., 2014).

Many other issues may arise with regard to the associated modeling assumptions, and should be assessed for their validity for any given problem. For example, in the estimation of total population size, the above models all assume that the population is closed (i.e., there are no births/deaths/migrations). For studies conducted over an extended period of time, this assumption may not be reasonable. This has lead to the development of stopover models within ecological applications, where individuals may enter and leave the population within the study period, so that the total population size is not constant throughout the study. Whenever the underlying assumptions are violated, the population estimates may become positively or negatively biased.

Summary

Estimating population sizes will continue to be an important aspect of many areas of the sciences and humanities. Capture–recapture methods are often used to obtain data that can be used to estimate the population size of interest. Understanding the limitations of the statistical models and associated assumptions that are made within the corresponding analyses will aid in the interpretation of the estimates obtained (and their use within, for example, social policy, etc.). Within any analysis, the population to be estimated needs to be clearly defined and appropriate sources used for the given population of interest. For example, suppose that the population of interest is the number of injecting drug users in a given area. Considering only sources that are dependent on such individuals also being prescribed or taking methadone leads to the population size being estimated relating to injecting drugs and methadone users (and not simply injecting drug users only). Understanding the data collection survey and collation of data between the different sources will provide further insight into the requirements of the subsequent statistical models used to analyze the data. Implementing careful study designs and standardized policies may help to minimize the necessary assumptions (and potential capture heterogeneities) and lead to improved estimation of the population size of interest.

See also: Bayesian Statistics; Estimation: Point and Interval; Log-Linear and Marginal Models; Model Selection and Model Averaging.

Bibliography

Alho, J.M., 1990. Logistic regression in capture-recapture models. Biometrics 46, 623–625.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis. MIT Press, Cambridge, MA.

Borchers, D.L., Buckland, S.T., Zucchini, W., 2002. Estimating Animal Abundance: Closed Populations. Springer, New York.

Bunge, J.A., 2013. A survey of software for fitting capture-recapture models. WIREs Computational Statistics 5, 114–120.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Chao, A., Tsay, P.K., Lin, S.-H., Shau, W.-Y., Chao, D.-Y., 2001. The applications of capture-recapture models to epidemiological data. Statistics in Medicine 20, 3123–3157.

Chapman, D.G., 1951. Some properties of the hypergeometric distribution with applications to zoological censuses. University of California Publications in Statistics 1, 131–160.

Coull, B., Agresti, A., 1999. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. Biometrics 55, 294–301.

Dorazio, R.M., Royle, J.A., 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics 59, 315–364.

Fienberg, S.E., 1972. The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika 59, 591–603.

Fienberg, S.E., Johnson, M.S., Junker, B.W., 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. Journal of the Royal Statistical Society: Series A 162, 383–405.

Fienberg, S.E., Manrique-Vallier, D., 2009. Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. Advances in Statistical Analysis 93, 49–60.

Hook, E.B., Albright, S.G., Cross, P.K., 1980. Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in live births and the completeness of vital record reports in New York state. American Journal of Epidemiology 112, 750–758.

Huggins, R.M., 1989. On the statistical analysis of capture experiments. Biometrika 76, 133–140.

King, R., Bird, S.M., Overstall, A., Hay, G., Hutchinson, S., 2014. Estimating prevalence of injecting drug users and associated death rates in England using regional data and incorporating prior information. Journal of the Royal Statistical Society: Series A 77, 209–236.

King, R., Brooks, S.P., 2001. On the Bayesian analysis of population size. Biometrika 88, 317–336.
King, R., Brooks, S.P., 2008. Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. Biometrics 64, 816–824.
Madigan, D., York, J.C., 1997. Bayesian methods for estimation of the size of a closed population. Biometrika 84, 19–31.
McCrea, R.S., Morgan, B.J.T., 2014. Analysis of Capture-Recapture Data. CRC Press, Boca Raton.
Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R., 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62, 1–135.
Overstall, A.M., King, R., 2014. Conting: an R package for Bayesian analysis of complete and incomplete contingency tables. Journal of Statistical Software 58 (7).
Overstall, A.M., King, R., Bird, S.M., Hutchinson, S.J., Hay, G., 2014. Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Statistics in Medicine 33, 1564–1579.
Pledger, S., Philpot, P., 2008. Using mixtures to model heterogeneity in ecological capture-recapture studies. Biometrical Journal 50, 1022–1034.
White, G.C., Burnham, K.P., 1999. Program MARK: survival estimation from populations of marked animals. Bird Study 46 (Suppl.), 120–138.
Wright, J.A., Barker, R.J., Schofield, M.R., Frantz, A.C., Byrom, A.E., Gleeson, D.M., 2009. Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65, 833–840.

Relevant Websites

http://cran.r-project.org/web/packages/conting/ – R Package conting.
http://www.r-project.org/ – Statistical Package R.
http://www.phidot.org/software/mark/ – Software MARK.

Populism

Koen Abts, University of Leuven, Leuven, Belgium
Stijn van Kessel, Loughborough University, Leicestershire, UK; and Heinrich-Heine-Universität, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.B. Collier, volume 17, pp. 11813–11816, © 2001, Elsevier Ltd.

Abstract

Populism is a concept applied to a wide range of political movements and actors across the globe. There is, at the same time, considerable confusion about the attributes and manifestation of populism, as well as its impact on democracy. This contribution identifies the defining elements of the populist ideology and discusses the varieties in which populism manifests itself, for instance as a component of certain party families. We finally discuss various normative interpretations of populism, and the tension between populism and liberal democratic values.

The Russian Narodniki movement of the 1860–70s is mostly seen as the first appearance of modern populism. This short-lived grassroots movement of intellectuals never achieved mass adherents but represented a radical articulation of the idea that wisdom resides with the common people and that socialism should be based on the traditional peasantry as its heartland (Taggart, 2000). American populism, beginning in the 1880s and continuing particularly in the South throughout the twentieth century, also referred to themes of moral decay, conspiracy, and betrayal of the ordinary people by the corrupt elite. The history of American populism could be portrayed as reactionary and a nativist movement of great 'democratic promise' (Goodwyn, 1976). In addition, Latin American populism of the mid-twentieth century has received much attention. The emphasis was on nation-building under a centralized and charismatic leadership (see Weyland, 2001; Roberts, 2006). From the 1980s, there is a resurgence throughout the world of movements and leaders that have been dubbed populist. These include, for instance Collor, Menem, Chavez, Obrador in Latin America; Le Pen, Haider, Wilders, Dewinter in Western Europe; Buchanan, Perot, Manning in the United States and Canada; Hanson in Australia; and Meciar, Milosevic, Lukashenko, Orbán in post-communist states. The populist label has thus been attached to a variety of political movements. It has been used as a term of approbation only by some movements themselves and mostly as a pejorative term by critics.

Populism: Mobilization, Leadership, Style, or Ideology?

The core idea of populism rests with the claim to represent or act in the name of the people, understood as the 'common people' and the 'silent majority.' Nevertheless, the main problem of defining populism is that conceptualizations encompass very different traits, while analysts apply the term to diverse phenomena. One approach suggests that populism is essentially a strategy of *political mobilization* using a typical style of political rhetoric (Kazin, 1995). It considers populism to be a tool for a leader to seek and exercise power. Populist parties and leaders appeal to the power of the common people to challenge the legitimacy of the current political establishment. A second approach focuses on populism as a *type of organization*. Populism presents itself not as an ordinary party characterized by different factions and an appeal to a specific section of society, but as a unified bloc or movement of the people. The movement is characterized by a politics of personality centered on a strong, personalized, even charismatic leader who is said to embody the will of the common people and who is able to speak on their behalf. The leader–mass linkage is direct, rather than mediated by organizations. As a result, populist framing is linked to a particular *style of communication* (Tarchi, 2002). Populism utilizes an antitheoretical rhetoric and anti-intellectual oratory to exploit feelings of resentment politically (Betz, 1994; Abts, 2012). Populists offer simplistic solutions to complex political problems in a very direct and demagogic language, appealing to the common sense of the people and denouncing the intellectualism of the established elites. Although charismatic leadership and simplistic language might be associated with populism, these action frames could, arguably, only be understood as expressions of an underlying populist ideology. However, populism needs to be understood as a *thin-centered ideology* because it does not provide a comprehensive vision of society (Mudde, 2004; Abts and Rummens, 2007; Stanley, 2008). It only gives a precise meaning and priority to certain key concepts of political discourse, thereby generating a particular ideological picture of parts of the political domain.

The Elements of the Populist Ideology

Populism is intrinsically "a thin-centered ideology which advocates the sovereign rule of the people as a homogeneous body" (Abts and Rummens, 2007: p. 409). Three elements of populist ideology are recurrently highlighted in the literature. First, it is argued that populism revolves around a central *antagonistic relationship between 'the people' and 'the elite.'* Populism is an "appeal to 'the people' against both the established structure of power and the dominant ideas and values of society" (Canovan, 1999: p. 3). The establishment is attacked for its alleged privileges, its corruption, and especially, for its lack of accountability

to the people. Elites are accused of representing only their own interests and of being alienated from the real interests, values, and opinions of the common man.

Secondly, populism tries to give power back to the people and restore *popular sovereignty*. Populists believe that politics should be based on the immediate expression of the general will of the people. They speak and behave as if "democracy meant the power of the people and only the power of the people" (Mény and Surel, 2002: p. 9). Populist ideology favors more direct forms of democracy, such as majority rule or referenda, which should replace current representative and intermediary institutional arrangements. Thereby, the will of the people is considered to be transparent and immediately accessible to those willing to listen to the *vox populi*. Thus, populism is wary of compromise and accommodation, and emphasizes the need for a politics of will and decision (Urbinati, 1998: pp. 116–118). As a consequence, there is no need for elaborate discussion or party politics, and populist movements and parties are often headed by charismatic leaders who can speak and act directly on behalf of the people. Populist democracy can be understood as an attempt to achieve an immediate identity of governed and governing. Populists claim to present and proclaim, not to represent, the essentialist will of the people. In this way, populism opts for the idea of popular sovereignty, if necessary, at the expense of constitutional guarantees.

Thirdly, the transparency of the will of the people is possible because populism conceptualizes the *people as a homogeneous unity* (Abts and Rummens, 2007). In populist ideology, 'the people' functions as a central signifier, which receives a fundamentally monolithic interpretation. The people are united and indivisible, fully formed, self-aware, and identifiable by the majority of numbers. Accordingly, the people are not seen as a heterogeneous collection of social groups and individual subjects with diverse values, needs, and opinions. On the basis of a supposed shared identity, the people are considered to form a collective body, which is capable of having a common will and a single interest and which is able to express this will and to take decisions (Canovan, 2002: p. 34). Populist ideology, however, only implies *that* the people constitute a homogeneous body, it does not say *what* this substantive identity should be. All actual populist movements need to supplement their thin-centered populist ideology with additional values and beliefs that give content to this substantive unity. Populism, then, cohabits with other, more comprehensive ideologies depending on the context and the values of the heartland to which it appeals. Paradigmatic possibilities here would be a leftist version of populism that identifies the people in socioeconomic terms as the working-class exploited by a bourgeois elite or a right-wing populist movement that refers to ethnonational characteristics to identify the people with the (ethnic) nation (see Mudde and Rovira Kaltwasser, 2013). The presumed unity of the people also implies that populism cultivates antagonistic relationships toward those who do not fit in and therefore threaten the homogeneity. Depending on the specific nature of the populist image of the people, this might include, for example, cultural and economic elites, foreigners, minorities, welfare recipients, or others. It can then be argued that populism is an inherently antipluralistic ideology.

Based on these three core elements, scholars present different, yet overlapping, descriptions. Mudde (2004: p. 543) defines populism as "an ideology that considers society ultimately separated into two homogeneous and antagonistic groups, 'the pure people' versus 'the corrupt elite,' and argues that politics should be an expression of the *volonté générale* of the people." Albertazzi and McDonnell (2008: p. 3), on the other hand, conceptualize populism as "an ideology which pits a virtuous and homogeneous people against a set of elites and dangerous 'others' who are together depicted as depriving (or attempting to deprive) the sovereign people of their rights, values, prosperity, identity and voice."

Confronted with a great number and diversity of both movements and (ideological) traits it seems that populism occurs in various forms (e.g., style, strategy, and core ideological attribute), and is not confined to one specific type of political actor. The diversity of populist movements is accommodated by different typologies of populism – types differentiated in terms of traits in the core definition, which they lack. In their groundbreaking edited volume on populism, Ionescu and Gellner (1969) already observed that populism was used to refer to actors from a wide-ranging set of political ideologies. Later, Canovan (1981: p. 133) argued that it is not possible to unite different populist movements "into a single political phenomenon with a single ideology, program or socioeconomic base." Notwithstanding the diversity among populist parties, several patterns can be distinguished. Populist leaders in Latin America, for example, tend to be characterized by a socioeconomically left position, whereas populism in Europe is often associated with the radical right (Mudde and Rovira Kaltwasser, 2013). In the United States, in turn, populism is often seen as an intrinsic feature of politics, in view of the more general wariness of big government and the Washington elite (see Ware, 2002). More recently, the term has often been associated with the culturally conservative and anti-statist Tea Party. In organizational terms, populism in Europe is predominantly seen as a feature of political parties rather than individuals – even though many of those parties are characterized by centralized leadership and a strong figurehead. Irrespective of the general tendency of populism to cohabit with the xenophobic right, populism in Europe has also been associated with other 'party families.'

Populism in Party Families

Mudde (2007) has treated populism as a core element of the populist radical right (PRR), but also identified two other types of populist parties: neoliberal populists and social populists (see also March, 2011). Parties from the latter two families are mainly concerned with socioeconomic themes – promoting free-market economics and welfare protectionism, respectively – while PRR parties are, besides their populism, primarily characterized by their nativism and authoritarianism. Thus, even though it would not be accurate to speak of a populist 'party family,' because populist parties can come in various ideological shapes and guises, several party families can be distinguished for which populism is a defining characteristic.

Again, while there have been prominent examples of neoliberal populist parties (e.g., those of Silvio Berlusconi in Italy and Pim Fortuyn in the Netherlands) and social populists

(such as Die Linke in Germany and Sinn Fein in the Republic of Ireland), the majority of the cases of populism in Europe appear to fall within the PRR category. In Western Europe, these parties are for instance characterized by their hostility against immigration and emphasis on the need for foreigners to adapt to the nation's customs. Several such parties have been around for quite a long time: the Front National (FN) in France was founded in 1972, the Vlaams Belang (VB, previously Flemish Block) in 1978. Other parties with a longer history have transformed themselves into successful PRR parties more recently – the Austrian Freedom Party (FPÖ) and the Swiss People's Party being prime examples. In post-communist countries in Central and Eastern Europe, where immigration levels have been low or even negative, populist radical right parties have typically identified ethnic minorities – and in some countries the Roma population in particular – as the nonnative elements. Examples include parties such as Jobbik in Hungary, the Greater Romania Party (PRM), and the Slovak National Party (SNS).

It must be noted that populism is not necessarily a phenomenon confined to certain parties only. In Europe, too, populism can take the form of a more fleeting rhetorical device used to blame electoral competitors for having lost touch with the 'ordinary people.' In post-communist countries, where political trust tends to be low and public sector corruption is often a salient electoral theme, many (newly established) political parties have voiced populist antiestablishment rhetoric (Van Kessel, 2014). The term populism has also regularly been used to describe large mainstream parties such as the Hungarian FIDESZ and the Polish Law and Justice (PiS).

The difficulty in pinpointing exactly which actors are populist or not has added to the concept's unsystematic use and the more general confusion surrounding the term. Another problem of 'populism' is that the term is used rather randomly as a term of abuse and that populism is, without much substantiation, often automatically treated as a phenomenon that threatens the quality of democracy.

Populism and Democracy

Populism is regularly applied as a synonym for demagoguery, simplistic solutions, or opportunism (e.g., Mudde, 2004; Taggart, 2000). Especially in the vernacular sphere, the term is often used pejoratively to refer to vote-winning policy proposals, attempts to pander to public opinion, or anti-immigration attitudes (Bale et al., 2011). The academic debate about the concept may be more sophisticated, but also in scholarly literature, populism is frequently seen as a negative phenomenon. Betz (1994: p. 4), for instance, sees populism as a means of political opportunism that is unscrupulous and exploitative of the anxieties of the electorate. Others view populism's support for unmediated popular sovereignty as a threat to liberal democratic 'checks and balances' and the protection of minorities. Abts and Rummens (2007: p. 4), for instance, argued that populism is inherently incompatible with democracy; populism's conception of the people as a homogeneous body is fictional and "generates a logic which disregards the idea of otherness at the heart of democracy and aims at the suppression of diversity within society."

A more optimistic view is that populism is an indicator of the health of representative democracy. Although not denying the risks of populist politics, some scholars argue that populism emerges when the political elite loses track of the popular will, or when the 'constitutional' or 'liberal,' as opposed to the 'democratic' or 'popular' pillar of democracy, is seen to become too dominant. Taggart (2002: p. 63), for instance, argued that populism acts as a 'bellwether' for the health of representative politics. Mény and Surel (2002: p. 17) have described populism as "a warning signal about the defects, limits and weaknesses of representative systems," and argue that "in spite of its often unpleasant tones, it may constitute an effective reminder that democracy is not a given, but is instead a constant enterprise of adjustment to the changing needs and values of society." Canovan (1999: p. 11) speaks about the tension between the 'pragmatic' and 'redemptive' faces of democracy and argued that "When too great a gap opens up between haloed democracy and the grubby business of politics, populists tend to move on to the vacant territory, promising in place of the dirty world of party manoeuvring the shining ideal of democracy renewed. Even from the point of view of pragmatic politics, the vital practices of contestation and accountability grow weak without the energy provided by democracy's inspirational, mobilizing, redemptive side."

Still, populism is seldom seen as an unequivocally good thing. Not many scholars appear to subscribe to Laclau's (2005: p. 48) argument that populism and democracy are essentially interchangeable terms and that "the end of populism coincides with the end of politics." Panizza (2005: p. 30) instead described populism as a 'mirror in which democracy can contemplate itself,' but argued that 'populism is neither the highest form of democracy nor its enemy.' In the same volume, Arditi (2005) argued that populism can appear in three possible modes: populism as a mode of representation, as a symptom, or as an underside. The first mode is compatible with liberal-democratic politics, the second presents a disturbance of democracy, whereas the latter entails an actual interruption of democracy. Pasquino (2008: p. 28), in turn, argued that the appearance of populism is often a sign of a poorly functioning democratic regime, but that populism, for instance due to its unrealizable promises, has a negative impact on the democratic framework itself.

An edited volume by Mudde and Rovira Kaltwasser (2012) attempted to test empirically the consequences of populism in several countries in Europe and the Americas. Based on the various case studies, the editors concluded that populism can be a corrective as well as a threat to democracy. Populist actors can, for instance, place issues on the agenda which have been ignored by the political establishment and give voice to excluded sections of society. Yet especially in unconsolidated democracies, populism can also undermine liberal democratic institutions in view of its monist conception of society and disdain for 'checks and balances.'

As a consequence, an obvious question is how established political actors could react to emerging populism. A quick glance reveals that Western Europe is divided about how to deal with PRR parties. In Italy, Austria, Denmark, and the Netherlands a rather open and accommodating approach prevails whereby populist challengers are treated as ordinary political opponents and forms of cooperation remain possible.

In other countries, such as France and Belgium, institutional and political actors pursue a much more repressive and antagonistic approach whereby the populist parties are delegitimized and treated as genuine enemies of the democratic system (Eatwell and Mudde, 2004). Consequently, the public debate is largely determined by different stances toward the well-known dilemma of 'tolerance for the intolerant.' On the one hand, the *procedural view* conceives of democracy as a set of procedures organizing the struggle for power. In the electoral marketplace, all opinions and parties are tolerated as long as they follow the procedural rules. As a result, freedom of expression and association is almost unlimited. The adherents of the procedural view propose to treat populist parties as legitimate opponents in the political struggle, while it would be inappropriate to challenge political opponents with repressive rules of exclusion. On the other hand, the *substantive view* advocates a more repressive stance toward populism, in case it leads to extremism. Democratic procedures are not an end in themselves but only the means for realizing and protecting a set of substantive democratic values and rights. Consequently, all political parties are required to underwrite the fundamental values of freedom, equality, respect, and tolerance. As far as populists threaten these values, the tolerance for the intolerant may be limited. In this sense, repressive measures like a cordon sanitaire, which aim to obstruct extremist parties, are not allowed but actually required (Capoccia, 2007: pp. 55–67). Instead, Rummens and Abts (2010) propose a concentric containment policy for dealing with populist parties in trying to overcome the traditional opposition between procedural and substantive views of democracy. On the procedural side, this policy emphasizes the importance of tracking all the relevant concerns of citizens in the public sphere, whereas, on the substantive side, it stresses the need for an adequate filtering which guarantees the compatibility of actual policies with the core values of liberty and equality. Their proposed twofold requirement of tracking and filtering translates into a guideline of decreasing tolerance toward populists as they approach the centers of formal decision-making power. The resulting containment policy listens to populist voters, but puts simultaneously unremitting civilizing pressure on populist parties themselves.

See also: Agendas: Political; Political Parties: Extreme Right.

Bibliography

Abts, K., 2012. Social Resentment and Ethno-Populism: Late Modernity, Ressentiment, Ethnocentrism, Political Cynicism and Radical Right Voting. Ph.D. manuscript. KU Leuven, Leuven.

Abts, K., Rummens, S., 2007. Populism versus democracy. Political Studies 55 (6), 405–424.

Albertazzi, D., McDonnell, D., 2008. Introduction: the spectre and the spectre. In: Albertazzi, D., McDonnell, D. (Eds.), Twenty-First Century Populism. The Spectre of Western European Democracy. Palgrave, Basingstoke, pp. 1–11.

Arditi, B., 2005. Populism as an internal periphery of democratic politics. In: Panizza, F. (Ed.), Populism and the Mirror of Democracy. Verso, London, pp. 72–98.

Bale, T., van Kessel, S., Taggart, P., 2011. Thrown around with abandon? Popular understandings of populism as conveyed by the print media: a UK case study. Acta Politica 46 (2), 111–131.

Betz, H.-G., 1994. Radical Right-Wing Populism in Western Europe. St. Martin's Press, New York.

Canovan, M., 1981. Populism. Harcourt Brace Jovanovich, New York and London.

Canovan, M., 1999. Trust the people! Populism and the two faces of democracy. Political Studies 47 (1), 2–16.

Canovan, M., 2002. Taking politics to the people: populism as the ideology of democracy. In: Mény, Y., Surel, Y. (Eds.), Democracies and the Populist Challenge. Palgrave, Basingstoke, pp. 25–44.

Capoccia, G., 2007. Defending Democracy: Reactions to Extremism in Interwar Europe. John Hopkins University Press, Baltimore.

Eatwell, R., Mudde, C. (Eds.), 2004. Democracy and the New Extreme Right Challenge. Routledge, London.

Goodwyn, L., 1976. Democratic Promise: The Populist Movement in America. Oxford University Press, New York.

Kazin, M., 1995. The Populist Persuasion: An American History. Basic Books, New York.

Ionescu, G., Gellner, E. (Eds.), 1969. Populism, Its Meanings and National Characteristics. Weidenfeld and Nicolson, London.

Laclau, E., 2005. Populism: what's in a name? In: Panizza, F. (Ed.), Populism and the Mirror of Democracy. Verso, London, pp. 32–49.

March, L., 2011. Radical Left Parties in Europe. Routledge, Oxon.

Mény, Y., Surel, Y., 2002. The constitutive ambiguity of populism. In: Mény, Y., Surel, Y. (Eds.), Democracies and the Populist Challenge. Palgrave, Basingstoke, pp. 1–21.

Mudde, C., 2004. The populist Zeitgeist. Government and Opposition 39 (4), 542–563.

Mudde, C., 2007. Populist Radical Right Parties in Europe. Cambridge University Press, Cambridge.

Mudde, C., Rovira Kaltwasser, C. (Eds.), 2012. Populism in Europe and the Americas. Threat or Corrective for Democracy?. Cambridge University Press, Cambridge.

Mudde, C., Rovira Kaltwasser, C., 2013. Exclusionary vs. inclusionary populism: comparing contemporary Europe and Latin America. Government and Opposition 48 (2), 147–174.

Panizza, F., 2005. Introduction: populism and the mirror of democracy. In: Panizza, F. (Ed.), Populism and the Mirror of Democracy. Verso, London, pp. 1–31.

Pasquino, G., 2008. Populism and democracy. In: Albertazzi, D., McDonnell, D. (Eds.), Twenty-First Century Populism. The Spectre of Western European Democracy. Palgrave MacMillan, Basingstoke, pp. 15–29.

Roberts, K., 2006. Populism, political conflict, and Grass-Root organisation in Latin America. Comparative Politics 38 (2), 127–148.

Rummens, S., Abts, K., 2010. Defending democracy: the concentric containment of political extremism. Political Studies 58 (4), 649–665.

Stanley, B., 2008. The thin ideology of populism. Journal of Political Ideologies 13 (1), 95–110.

Taggart, P., 2000. Populism. Open University Press, Buckingham and Philadelphia.

Tarchi, M., 2002. Populism Italian Style. In: Mény, Y., Surel, Y. (Eds.), Democracies and the Populist Challenge. Palgrave, Basingstoke, pp. 120–138.

Taggart, P., 2002. Populism and the pathology of representative politics. In: Mény, Y., Surel, Y. (Eds.), Democracies and the Populist Challenge. Palgrave, Basingstoke, pp. 62–80.

Urbinati, N., 1998. Democracy and populism. Constellations 5 (1), 110–124.

Van Kessel, S., 2014. The populist cat-dog. Applying the concept of populism to contemporary European party systems. Journal of Political Ideologies 19 (1), 99–118.

Ware, A., 2002. The United States: populism as politics strategy. In: Mény, Y., Surel, Y. (Eds.), Democracies and the Populist Challenge. Palgrave, Basingstoke, pp. 101–119.

Weyland, K., 2001. Clarifying a contested concept – populism in the study of Latin American politics. Comparative Politics 34 (1), 1–22.

Pornography

Gert M Hald, University of Copenhagen, Copenhagen, Denmark
Neil M Malamuth, University of California Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by N.M. Malamuth, volume 17, pp. 11816–11821, © 2001, Elsevier Ltd.

Abstract

The following is a review of pornography consumption patterns and effects of pornography on attitudes, cognitions, and behaviors. The review shows that pornography is widely circulated and used; predominately nonviolent pornography is used; main media outlets for consumption are the Internet, videos/DVDs, television, and magazines; and most often pornography is watched at home, either alone or with a regular sexual partner. Further, although causality cannot be conclusively established, an increasing amount of diverse empirical evidence suggests a variety of associations between pornography and certain cognitions, attitudes, and behaviors for a subgroup of users. This implies that possible links between pornography and certain cognitions, attitudes, and behaviors are moderated by individual and contextual differences.

Introduction

The topic of pornography has been discussed from many vantage points and is often hotly debated by scholars and researchers. Some have focused on the functions and effects of pornographic mass media as they pertain to understanding the impact of pornography on individual consumers, for example, in the realms of fantasy, sex education, esthetic/artistic expression, the sex life, and reproduction. Yet others have approached the topic at a broader level, exploring such topics as the relevance of pornography to male–female relationships, family and group relations, social institutions, the role of government regulation, and social change.

Traditionally, opponents of pornography have claimed that pornography may have damaging effects on beliefs, morals, values, attitudes, and behaviors, and they hold pornography responsible for a variety of adverse effects, including wrecking marriages, negatively changing men's perceptions of women and women's perceptions of themselves, sexual addiction, unhealthy attitudes, and problematic (sexual) behaviors. On the contrary, proponents of pornography have claimed that little or no such effects of pornography consumption are evident. Rather, pornography may benefit the individual by enhancing the sex life, contributing to knowledge about sex, providing a recreational sexual outlet, buffering against sexual assaults, or helping to assess or cure common sexological dysfunctions.

It is beyond the scope of this article to consider all of these diverse approaches to the topic of pornography. Instead, the focus here may be described as a 'social science-centered' approach revolving around two central questions: First, what are the findings on pornography consumption patterns? Second, what are the findings of effects of pornography on attitudes, cognitions, sexual behaviors, and sexually aggressive behaviors?

Given the controversial nature of pornography, it is important to stress that the following review should not be seen as an attack on or defense of pornography but as an attempt to synthesize current empirical literature on pornography.

Defining Pornography

Originally, the term *pornography* referred to the writings about or depictions of harlots or prostitutes. The term derives from the Greek word *pornographos*, which again derives from the Greek words *pórne*, meaning 'whore,' and *graphein*, meaning 'to write.' Thus, the term originally referred to 'writings on whores.'

In terms of legal definitions, the terms *pornography* and *obscenity* need to be distinguished. Obscenity refers to pornographic content that has been judged by the legal system to be illegal. In the United States, most states use the 1973 Supreme Court decision in *Miller vs. California* as the legal definition of *obscene*. To be judged obscene under this case the material must satisfy three conditions: (1) The average person, applying contemporary community standards, would find that the work, taken as a whole, appeals to the prurient interest; (2) the work depicts sexual conduct in a patently offensive way; and (3) the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.

According to Diamond (2009) the most common definition of pornography used by researchers today is: "media basically construed as intended to entertain or arouse erotic desire" (p. 304). Nonviolent pornography may be defined as sexually explicit materials without any overt coercive content, but which may sometimes imply acts of submission and/or coercion by the positioning of the models, use of props, or display of unequal power relationships. Violent pornography may be defined as sexually explicit materials in which nonconsensual, coercive, and/or violent sexual relations are explicitly portrayed (see also Hald et al., 2010). However, as elaborated upon in the following section, some ambiguity exists in how to define 'violent pornography.'

Pornography Consumption Patterns

Prevalence of Pornography Consumption

Some of the most consistent findings in the area of pornography research are the relatively high consumption rates especially among men and the large gender differences in

consumption patterns. Following are summaries of such and related findings.

Consumption rates

With few exceptions, international studies have reported consumption rates in the range of 50–99% amongst men and 30–86% amongst women. However, these rates have been found to be heavily influenced by sample characteristics (e.g., age), time period assessed (e.g., pornography consumption in the past six months, year, ever), sexual experience, definition of pornography, and study methodology and design.

Gender differences

Compared to women, men have been found to use pornography more frequently, spend more time on pornography consumption, be exposed to pornography at a younger age, and use pornography more often during masturbation. Further, compared to women, men prefer pornography with many different actors as compared to pornography with the same actors performing different acts, are more attracted to both a wider range of hardcore pornography and hardcore pornography devoid of relationship context and emotional attachments, and are generally, more sexually aroused by pornography (Hald and Malamuth, 2008).

Frequency of exposure

According to Edelman (2009), based on a 2008 monitoring of Internet usage patterns, 36% of users visited at least one adult entertainment website per month. Among these 36%, the average number of visits to adult websites was 7.7 times a month with an average visit lasting 11.6 minutes. It has been estimated that around 25% of all Internet search engine requests are for pornography, and some of the most popular search engine request words are sexually related.

Age of first exposure

Age at first pornographic exposure varies from culture to culture, but most findings suggest that first nonaccidental exposure on average happens between the ages of 11 and 15 with men being exposed at a significantly younger age than women.

Pornography Consumption Outlets

Financially, the porn industry has been 'big business' with the largest companies being listed on stock exchanges and many other large companies (e.g., hotel chains, cable networks, etc.) gaining much revenue from such media content. In a 2006 estimate, the AVN Media Network, an adult entertainment trade organization, estimated the total retail sales of pornography to approximately US$ 13 billion. Up until the 1990s the dominant media outlets for pornography consumption were magazines, videos/DVDs, and television. However, this changed in the wake of the Internet and smartphone revolution where high-speed connections increasingly allowed for easy, cheap, and convenient access to all types of pornography (Edelman, 2009). Today, the predominant media outlets of pornography consumption are the Internet, videos/DVDs, television, and magazines.

Types of and Reasons for Pornography Consumption

Increased recent concerns regarding the consumption of pornography have focused on the easy and inexpensive access to all types of pornography. This especially concerns the use of violent pornography or more atypical kinds. Unfortunately, systematic longitudinal studies of dominant types/themes of consumption of pornography over the past several decades are not available.

Content of pornography

Most of the pornography consumed today is primarily of a nonviolent nature according to the definitions described herein. The dominant themes of consumption include: vaginal-, anal-, oral-, and group sex. In total, approximately 10–15 % of users prefer same-gender, violent/coercive, fetish, Bondage & Domination (B&D), or Sadomasochism & masochism (S&M) pornography (Hald et al., in press).

However, potential ambiguity in how to define 'violent pornography' is highlighted by a recent content analysis of the leading 'adult' films (Bridges et al., 2010). This study content analyzed a randomly selected representative sample of 50 adult film videos from the 2004–05 list of best-selling adult film videos. The authors noted the problem of applying the term *violence* and chose instead to focus on the frequency of 'aggression' in adult films, including a wide range of physical acts to define physical aggression:

> (a) pushing or shoving; (b) biting; (c) pinching; (d) pulling hair; (e) spanking; (f) open-hand slapping; (g) gagging (defined as when an object or body part, e.g., penis, hand, or sex toy, is inserted into a character's mouth, visibly obstructing breathing); (h) choking (when one character visibly places his or her hands around another character's throat with applied pressure); (i) threatening with weapon; (j) kicking; (k) closed fist punching; (l) bondage or confining; (m) using weapons; and (n) torturing, mutilating, or attempting murder (p. 1075).

Also coded separately in this content analytic study were verbally aggressive acts that included (1) name calling or insulting and (2) threatening physical harm. In terms of agreement among coders, trained coders reached very high levels of agreement regarding the various types of aggression portrayed. The findings indicated that using this highly inclusive definition of physical aggression, the clear majority of adult videos included some forms of aggression. For instance, over 75% of the scenes included portrayals of spanking and over 50% of gagging. Choking occurred in about 28% of the scenes.

However, the portrayal of what may be described as violence (according to widely accepted social scientific definitions of this concept) in this sexual content was actually very rare: none of the scenes portrayed the use of threat with a weapon, hitting with a closed fist, torture, or mutilation. Using coercive language occurred in only 1.6% of the scenes. Thus, at least in 'mainstream' commercially released pornography, it would appear that violent pornography is rare. Nevertheless, earlier trends, documented some time ago, appear to have shown considerable increases in the accessibility of violent pornography in recent years, fueled by the Internet revolution (Gaddam and Ogas, 2011).

Reasons for Consumption

A variety of reasons for consuming pornography are given by consumers, with the most prevalent being: involuntary/chance exposure (e.g., spam, pop-ups), for entertainment/fun, curiosity, sexual learning, mood management, out of habit, to build and maintain relationships, for fun, to enhance sexual fantasies/the sex life, and as a prelude to or intensifier in autoerotic activities, or sexual intercourse (Emmers-Sommer and Burns, 2005).

Situational and Interpersonal Characteristics of Pornography Consumption

Few quantitative scientific studies have examined preferences in, and situational and interpersonal characteristics of pornography consumption. For both men and women, these studies have found that the main place of use is 'home,' and that the interpersonal context of use is either 'alone' or with a 'regular sexual partner' (e.g., Hald, 2006). Although teenagers in particular also report watching pornography with friends, this is not common fare. Using both American and Scandinavian samples research finds that female users are significantly more likely than male users to report watching pornography with a sexual or dating partner, whereas male users are significantly more likely than female users to report watching pornography alone (Hald et al., in press).

Summary: Patterns of Pornography Consumption

Pornography is widely circulated and used. In general, many gender differences in pornography consumption patterns have been consistently demonstrated. Although also used by women, pornography consumption amongst men can be conceptualized as a cornerstone of their autoerotic sexuality.

Most of the pornography consumed today is primarily of a nonviolent nature. The predominant themes of preference include vaginal-, oral-, anal-, and group sex, with only a smaller group of users preferring themes such as violent/coercive, fetish, B&D, or S&M pornography. Presently, the dominant media outlets are the Internet, videos/DVD, television, and magazines. Pornography is most often watched at home, either alone or with a regular sexual partner. Main reasons for watching pornography include involuntary/chance exposure, entertainment/fun, curiosity, sexual learning, to enhance sexual fantasies/the sex life, and as a prelude to or intensifier in autoerotic activities or sexual intercourse.

Effects of Pornography

Effects of Pornography on Attitudes

In the controversy about effects of exposure to pornography, attitudes have long held a central and decisive role. The widely accepted conceptualization of 'attitudes' usually incorporates three components, including affective responses, cognitive evaluations, and behavioral predispositions toward an entity (see also Hald et al., 2010). Predominately, attitudes supporting violence (ASV) against women (including rape supportive attitudes and beliefs), sexist attitudes, sexually permissive attitudes, and attitudes toward pornography have been targeted in pornography research.

Attitudes Supporting Violence against Women

ASV have been conceptualized as (1) affective responses to acts such as rape, other types of sexual aggression, and partner violence, (2) evaluative cognitions, and (3) behavioral predispositions or attractions toward such aggressive acts (Malamuth et al., 2012). Consequently, ASV also incorporates rape supportive attitudes and rape myth acceptance (see also Allen et al., 1995b).

Generally, data from a wide variety of experimental and correlational studies have converged to show that for some individuals pornography consumption may increase ASV (Allen et al., 1995a; Hald et al., in press). In a recent meta-analysis on the relationship between pornography consumption and ASV in nonexperimental studies, Hald et al. (2010) found a significant association between pornography consumption and ASV, with a significantly stronger effect for violent than nonviolent pornography. Furthermore, Hald et al. (2010) found heterogeneity among the included studies to a degree indicative of moderating variables, although the literature encompassed in this meta-analysis did not enable the authors to identify the basis for such moderation. However, another study by Malamuth et al. (2012) using a large representative sample of American men made this possible. It was found that the significant association between pornography consumption and ASV was moderated by individual differences in risk of sexual aggression, such that consumers at the highest risk level of sexual aggression who reported the most frequent use of pornography also held the highest level of ASV. These results from nonexperimental studies converge with results from experimental studies. For example, in a meta-analysis on pornography and ASV in experimental studies, Allen et al. (1995a) reported a significant effect of pornography consumption on ASV with a stronger effect for violent than for nonviolent pornography.

Sexist Attitudes

Sexism has been defined as a form of negative attitude toward women (or men) including hostility, negative prejudices, and stereotypes (Glick and Fiske, 1996). Sexist attitudes have been linked to a number of adverse effects. These include discrimination against women, gender rigidity, hostility, and anger toward women, greater self-reported likelihood of using violence in relationships with women, acceptance of rape myths, and the commission of real-life sexual aggression, including rape (Hald et al., in press). Studies on the association between pornography and sexist attitudes present an ambiguous and inconsistent picture. Thus, a number of previous survey and experimental studies fail to demonstrate significant associations between pornography and sexism or find only weak positive associations or even negative associations between pornography and sexism. This may be due to the lack of distinction between differing kinds of sexist attitudes (see also Glick and Fiske, 1996) or to the overall scientific quality of some of these studies, which has been seriously questioned (Malamuth et al., 2000), or because

many of these studies fail to consider individual differences as potentially important effect moderators. In view of the findings from other areas of the pornography effects research that specifically highlight the importance of moderators, this seems a critical shortcoming.

Other research studies, among them prospective and longitudinal studies, report that pornography consumption of various kinds increases 'notions of women as sex objects' and is associated with significantly stronger gender stereotypic attitudes and cognitions in sexually coercive males as compared to both normal controls and young males with a conduct disorder. Further, studies report that men who habitually consume high amounts of Internet pornography are more likely to view women in stereotypical gender roles, perceive a world inhabited by masculine men, and report the highest level of self-perceived masculine personality characteristics. Further, for both males and females, early exposure to sexually explicit media has been found to predict less progressive gender role attitudes, and, for males only, greater sexual harassment perpetration 2 years later (Brown and L'Engle, 2009).

Sexually Permissive Attitudes

Exposure to pornography has been shown to be significantly associated with sexually permissive attitudes, with stronger associations for Internet pornography than for more traditional types of pornography consumption (Kingston et al., 2009). For example, using both American, Dutch, Scandinavian, and Taiwanese samples of youths and young adults, it has been shown that more frequent exposure to sexually explicit Internet material is associated with more positive attitudes toward uncommitted sexual exploration, for example, sexual relations with casual partners/friends or with sexual partners in one-night stands, sexually permissive attitudes, positive attitudes toward sex, sexually permissive attitudes, and an expansive sexuality including more positive attitudes toward oral and anal sex as appealing. Further, in a large-scale longitudinal study of American youths, Brown and L'Engle (2009) found both large significant positive associations between pornography consumption and permissive personal sexual norms at baseline, and that, among men, increased exposure to pornography at baseline predicted more permissive personal sexual norms 2 years later, even after controlling for baseline sexual norms, demographic variables, and sensation seeking.

Attitudes toward Pornography

One of the best documented effects of pornography exposure is that repeated exposure results in more favorable attitudes toward pornography (ATP). American, Danish, Finnish, Italian, Norwegian, and Swedish studies all indicate gender differences in ATP (see also Hald et al., in press). In these studies, it was found that women more than men favor restrictions on pornography and are more likely to describe pornography as dull, not exciting, or repulsive. Further, compared to women, men express more positive ATP and are more likely to view pornography as a means of sexual enhancement.

Summary: Attitudes

Using a variety of subject samples, study designs, and cross-cultural research, significant positive associations between pornography consumption and both ASV, sexually permissive attitudes, and attitudes toward pornography have consistently been demonstrated. Research including individual differences has revealed the critical importance of individual differences in the pornography-attitudinal relationship. For the relationship between pornography consumption and sexist attitudes, findings are more conflicting and inconclusive.

Effects of Pornography on Cognitions

Abnormalities in the perceptual-, memory-, or information-processing systems may lead to maladaptive, offense supportive cognitions and problematic interpretations of stimuli and social encounters (see also Ward and Beech, 2006). Further, these systems may work as preattentive filters biasing the intake and processing of information and prime offense supportive cognitions (Ward and Beech, 2006). In turn, offense supportive cognitions may evoke emotional, attitudinal, and behavioral responses, increasing the probability of sexually aggressive behaviors. Pornography exposure may affect these cognition-related systems.

Cognitively, pornography may prime, maintain, and reinforce offense supportive cognitions, for example, that women are constantly receptive to men's needs, deceptive, or treacherous (Ward and Beech, 2006). When primed repeatedly, such cognitions may result in the chronic accessibility of attitudes that reduce the responsibility of men who commit acts of aggression against women and generally reinforce the acceptance of dominating, controlling, and perhaps even violent acts of aggression against women.

Individual differences moderation found in previous research (e.g., Malamuth et al., 2012) suggests that such priming, maintenance, and reinforcement by pornography may be particularly probable for men classified at high risk of sexual aggression (see also Malamuth et al., 2000; Ward and Beech, 2006) as these men often already hold hostile/power schemas associated with women and sexuality and view women dichotomously as either 'whores' or 'madonnas.'

Thus, the availability and accessibility of offense supportive schemas similar to that of (some) pornography may already be in place for, as well as perceptually and emotionally attended to, by this group of men. In keeping with these findings, exposure to pornography may not activate similar schemas in men classified at low risk of sexual aggression because of the low occurrence or accessibility of such schemas among this group of men.

These findings may also help explain the interaction of pornography and individual differences in the development of stronger offense supportive schemas supporting violence against women, as men at greater risk for sexual aggression are more likely to choose to expose themselves to violent or similar forms of pornography (Malamuth et al., 2012).

Summary: Cognitions

Associations between pornography consumption and certain (offense supportive) cognitions have been hypothesized, as

have some of the possible mechanisms by which such associations may occur. However, studies testing these hypotheses and probable mediating mechanisms are still sparse. Further, the direction of causality between pornography consumption and cognitions is yet to be established.

Effects of Pornography on Behavior

Perhaps the most critical questions in pornography effects research have been whether, if, and how pornography consumption influences behaviors. Traditionally, two groups of behaviors have mainly been studied, namely sexual behaviors and sexually aggressive behaviors.

Sexual Behaviors

Using a variety of cross-cultural samples, methodology, and designs research has generally and with great consistency demonstrated associations between pornography consumption and sexual behaviors. For example, significant positive relationships have been found between pornography consumption and frequency of masturbation, varied sexual experiences, precoital behaviors, ever to have sold sex, lower age at first intercourse, number of sexual partners during the past year, number of overall and casual sexual intercourse partners, having multiple lifetime partners, having engaged in oral, anal, or group sex, not to have used contraception during the last intercourse, and to having had a sexually transmitted infection (Hald et al., 2013).

Even though pornography consumption has been consistently found to be associated with a diversity of sexual behaviors, this is not evidence of causal links between pornography consumption and sexual behaviors. As proposed by Hald et al. (2013) and others, associations between pornography consumption and sexual behaviors may be spurious, confounded, and/or (better) accounted for by third factors, such as individual differences in personality or sexual background variables. Indeed, Hald et al. (2013), using a large Dutch youth sample, found the added significant prediction values of pornography consumption on sexual behaviors (and vice versa) to be limited when controlling for the combined influence of a comprehensive array of factors previously identified as important influences of the pornography consumption-sexual behavior relationships, for example, (sexual) sensation seeking.

Sexually Aggressive Behavior

Sexually aggressive behavioral outcomes include verbal manipulation, taking sexual advantage of an intoxicated person, or physically forcing someone into sexual activities against their will (see also Hald et al., 2013). Longitudinal studies, reviews, and meta-analyses of existing literature have all shown positive associations between pornography consumption and sexually aggressive behaviors (e.g., Allen et al., 1995b; Kingston et al., 2009; Malamuth et al., 2000). However, these associations have also been shown to be moderated by various individual differences (see also Kingston et al., 2009). Such research suggests that once an individual possesses certain other key risk factors, then and only then does pornography consumption add significantly to the confluence or interactive combination of these key risk factors in the prediction of sexual aggression. This line of thinking is evident in leading integrative or multimodal models of sexual aggression, such as the Confluence Model of Sexual Aggression (e.g., Malamuth et al., 2000) and the Three Path Model (e.g., Knight and Sims-Knight, 2011).

In known groups of sexual offenders, pornography use has been assessed in terms of frequency and type of pornography used. However, much of this research has been equivocal and remains inconclusive (see also Kingston et al., 2008). In contrast, and in keeping with the models described before, a relatively recent study examining the unique contribution of pornography consumption to the longitudinal prediction of criminal recidivism in a large sample of child molesters found that, after controlling for general and specific risk factors for sexual aggression, pornography still added significantly to the prediction of recidivism (Kingston et al., 2008). Further, this study also indicated that frequency of pornography consumption was primarily a risk factor for higher-risk offenders.

Summary: Behavior

Across studies, cultures, and population samples, associations between pornography consumption and a diversity of sexually related behaviors have been consistently demonstrated. In the area of sexually aggressive behaviors, the research has indicated that it is the confluence or interactive combination of general antisocial factors, contextual factors, and factors specific to sexual aggression, including pornography use, that have the strongest predictive power of sexually aggressive behaviors. Consequently, potential effects of pornography on sexually aggressive outcomes are best conceptualized and understood within comprehensive integrative and empirically validated theoretical frameworks.

Further, research has demonstrated that increased risk of sexually aggressive outcomes on the basis of pornography consumption may only occur for a small subgroup of users, namely those already at relatively high risk of committing sexual aggression and amongst the most frequent pornography consumers. Although findings of pornography-behavioral links are fairly consistent throughout the literature this does not imply causality, nor that this relationship may not be (better) accounted for by third factors.

Beneficial Effects of Pornography

Predominately, pornography effect research has focused on adverse effects of pornography (consumption). However, especially in recent years, increasing focus has been on possible positive or beneficial effects of pornography. Thus, empirically, pornography consumption has been shown to benefit the individual by enhancing the sex life, contributing to knowledge about sex, providing a recreational sexual outlet, expanding sexual horizons through normalization of sexual behaviors, facilitating a willingness to explore new sexual behaviors and sexual relationships through empowerment, helping to assess or cure common sexological dysfunctions, assisting in medical settings in which semen is collected for reproductive purposes,

and improving the quality of life in general (Hald and Malamuth, 2008).

Many of the findings cited herein rely on self-reported and self-perceived effects of pornography. While proponents of pornography would argue that the individuals themselves are in the best position to judge such effects, and that results from studies such as the above should therefore be taken at face value, critics of pornography have described the gradual greater recognition of pornography as beneficial, and de facto, positive as one of the most insidious effects of consumption (see also Hald and Malamuth, 2008).

Conclusions

Pornography research has often been influenced by ideological/political perspectives with vested interests in particular conclusions. This may have led to framing of research questions and design of some studies in ways that encourage simple conclusions, while not readily accommodating more nuanced conclusions. Nevertheless, research in this area has progressed considerably. In future work, it is essential not to use an 'either–or' perspective in which research is cast simply into questions such as whether pornography exposure is generally harmful or not. As suggested in the research summarized here, differing conclusions may result depending on a variety of factors such as the cultural milieu, the individual's background and risk profile, the particular content of the stimuli, the type of responses focused on, the content of exposure, the consumer's environmental circumstances, and the way 'harm' is defined. Some may wish for a 'one-handed' set of conclusions, but the research more accurately justifies a 'multihanded' perspective that reveals the richness and complexity of the issues related to the study of pornography.

See also: Attitudes and Behavior; Gender and Feminist Studies in Geography; Media Effects on Children; Media Effects; Personality and Crime; Rape and Sexual Coercion; Violence and Media.

Bibliography

Allen, M., Dalessio, D., Brezgel, K., 1995a. A meta-analysis summarizing the effects of pornography. 2. Aggression after exposure. Human Communication Research 22 (2), 258–283. Available from: ISI:A1995TD50800004.

Allen, M., Emmers, T., Gebhardt, L., Giery, M.A., 1995b. Exposure to pornography and acceptance of rape myths. Journal of Communication 45 (1), 5–26. Available from: ISI:A1995QL84000002.

Bridges, A.J., Wosnitzer, R., Scharrer, E., Sun, C., Liberman, R., 2010. Aggression and sexual behavior in best-selling pornography videos: a content analysis update. Violence Against Women 16 (10), 1065–1085. Available from: ISI:000283510600002.

Brown, J.D., L'Engle, K.L., 2009. X-rated sexual attitudes and behaviors associated with US early adolescents' exposure to sexually explicit media. Communication Research 36 (1), 129–151. Available from: ISI:000262083300006.

Diamond, M., 2009. Pornography, public acceptance and sex related crime: a review. International Journal of Law and Psychiatry 32 (5), 304–314. Available from: PM:19665229.

Edelman, B., 2009. Markets red light states: who buys online adult entertainment? Journal of Economic Perspectives 23 (1), 209–220. Available from: ISI:000263886700010.

Emmers-Sommer, T.M., Burns, R.J., 2005. The relationship between exposure to Internet pornography and sexual attitudes toward women. Journal of Online Behavior 1 (4).

Gaddam, S. Ogas, O. 2011. A Billion Wicked Thoughts: What the World's Largest Experiment Reveals about Human Desire. New York.

Glick, P., Fiske, S.T., 1996. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology 70 (3), 491–512. Available from: ISI:A1996TZ88300007.

Hald, G.M., 2006. Gender differences in pornography consumption among young heterosexual Danish adults. Archives of Sexual Behavior 35 (5), 577–585. Available from: PM:17039402.

Hald, G.M., Kuyper, L., Adam, P.C.G., De Wit, J.B.F., 2013. Does Viewing explain Doing? Assessing the Association between Sexually Explicit Materials Use and Sexual Behaviors in a Large Sample of Dutch Adolescents and Young Adults. Journal of Sexual Medicine. (Advance Online Publication).

Hald, G.M., Malamuth, N.M., 2008. Self-perceived effects of pornography consumption. Archives of Sexual Behavior 37 (4), 614–625. Available from: PM:17851749.

Hald, G.M., Malamuth, N.M., Yuen, C., 2010. Pornography and attitudes supporting violence against women: revisiting the relationship in nonexperimental studies. Aggressive Behavior 36 (1), 14–20. Available from: PM:19862768.

Hald, G.M., Seaman, C., Linz, D., (in press). Sexuality and pornography. In: Tolman, D., Diamond, L., (Editors-in-Chief), Bauermeister, J., George, W., Pfaus, J., Ward, M., (Associate Editors), APA Handbook of Sexuality and Psychology: Vol. 2. Contextual Approaches. Washington, DC: American Psychological Association.

Kingston, D.A., Fedoroff, P., Firestone, P., Curry, S., Bradford, J.M., 2008. Pornography use and sexual aggression: the impact of frequency and type of pornography use on recidivism among sexual offenders. Aggressive Behavior 34 (4), 341–351. Available from: ISI:000256645800001.

Kingston, D.A., Malamuth, N.M., Fedoroff, P., Marshall, W.L., 2009. The importance of individual differences in pornography use: theoretical perspectives and implications for treating sexual offenders. Journal of Sex Research 46 (2–3), 216–232. Available from: ISI:000264481400008.

Knight, R.A., Sims-Knight, J.E., 2011. Risk factors for sexual violence. In: White, J.W., Koss, M.P., Kazden, A.E. (Eds.), Violence against Women and Children, Mapping the Terrain, vol. 1. American Psychological Association, Washington, DC, pp. 125–172.

Malamuth, N.M., Addison, T., Koss, M., 2000. Pornography and sexual aggression: are there reliable effects and can we understand them? Annual Review of Sex Research 11, 26–91. Available from: PM:11351835.

Malamuth, N.M., Hald, G.M., Koss, M.P, 2012. Pornography, individual differences in risk and men's acceptance of violence against women in a representative sample. Sex Roles 66, 427–439.

Ward, T., Beech, A., 2006. An integrated theory of sexual offending. Aggression and Violent Behavior 11 (1), 44–63. Available from: ISI:000234581900004.

Positivism, History of

Steve Fuller, University of Warwick, Coventry, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11821–11827, © 2001, Elsevier Ltd.

Abstract

Positivism is the name of a social and intellectual movement that tried to learn from the mistakes of the Enlightenment project that eventuated, first, in the Reign of Terror following the French Revolution of 1789, and second, in the irrationalism of the Weimar Republic following Germany's defeat in World War I. While it has been customary to distinguish between the quasipolitical movement called 'positivism' originated by Auguste Comte in the 1830s and the more strictly philosophical movement called 'logical positivism' associated with the Vienna Circle of the 1930s, both shared a common sensibility, namely, that the unchecked exercise of reason can have disastrous practical consequences. Thus, both held that reason needs 'foundations' to structure its subsequent development so as not to fall prey to a self-destructive scepticism. In this respect, positivism incorporates a heretofore absent empiricist dimension to the risk-averse orientation to the world historically associated with Platonism.

Positivism is the name of a social and intellectual movement that tried to learn from the mistakes of the Enlightenment project that eventuated, first, in the Reign of Terror following the French Revolution of 1789, and second, in the irrationalism of the Weimar Republic following Germany's defeat in World War I. While it has been customary to distinguish between the quasipolitical movement called 'positivism' originated by Auguste Comte in the 1830s and the more strictly philosophical movement called 'logical positivism' associated with the Vienna Circle of the 1930s, both shared a common sensibility, namely, that the unchecked exercise of reason can have disastrous practical consequences. Thus, both held that reason needs 'foundations' to structure its subsequent development so as not to fall prey to a self-destructive scepticism. In this respect, positivism incorporates a heretofore absent empiricist dimension to the risk-averse orientation to the world historically associated with Platonism.

Positivism's Platonic Political Roots

In terms of Western intellectual history, positivism inflects Plato's original philosophical motivation through a secularized version of the Christian salvation story, in which Newton functions as the Christ figure. This captures both the spirit of Auguste Comte's original project and its residual effects in twentieth century logical positivism, which dropped the overt historicism of Comte's project, while retaining the fixation on Newton as the model for what it means to express oneself scientifically and a vague belief that greater scientific knowledge will deliver salvation. Indeed, positivism's core conceptual problem has been to define a scientific vanguard capable of both offering guidance to the unenlightened and itself changing in light of further evidence and reflection. Its Platonic roots are worth recalling.

Plato was affected deeply by the volatile political environment created by the public dialectical encounters of Socrates and the other Sophists who mentored the leaders of his Athenian youth. Thus, after the fall of Athens, Plato was keen to restrict the dialectic to a cloistered setting, only to be unleashed to the larger society once initiates were sufficiently secure in their own knowledge and judgement that they would not be unduly swayed by popular demands. Where Plato had hoped to produce implacable philosopher-kings who would rule as absolute monarchs, positivists typically have envisaged a more differentiated but no less authoritative (authoritarian?) rule by experts, each an oligarch over their domain of knowledge. In this respect, positivism is bureaucracy's philosophical fellow traveler.

Not surprisingly, positivism's relationship to democracy has been chequered. Like Plato, positivists have feared protracted public disagreement most of all and hence have tended to demonize it as 'irrational' and 'noncognitive.' Their image of 'plural' authority presumes nonoverlapping competences, such that legislative questions are reduced to judicial ones concerning the expertise to which one should defer. This is in sharp contrast to most republican conceptions of democracy, which base representation on interest groups, each invested with a competence appropriate to the promotion of its interests but typically about matters in which other interest groups might also have competence. Since matters of public policy typically affect several such groups at once, collective decisions do not turn on identifying the right group whose judgement should hold sway. Rather, it is expected that public debate will issue in a solution that transcends all the groups' starting points but nevertheless manages to serve their respective interests.

Thus, there is a fundamental ambiguity in positivism's appeal to organized reason, or 'science,' in the public sphere. Sometimes this ambiguity is finessed by saying that positivists regard science as the main source of political unity. At the very least, this implies that it is in the interest of all members of society to pursue their ends by scientific means, as that may enable them to economize on effort and hence allow more time for the fruits of their labour to be enjoyed. Ernst Mach (discussed in Section. Antithesis: Machian Positivism) comes closest to defending this position in its pure form. It fits comfortably with the libertarian idea that democratic regimes should enable maximum self-empowerment.

However, many positivists have drawn a further conclusion that can thwart this libertarian impulse. From Comte onward, it has been common to argue that science can unify the polity by resolving, containing, or circumventing social conflict. Here a well-established procedure or a decisive set of facts is supposed to replace more 'primitive' and volatile forms of conflict resolution such as warfare and sometimes even open debate – all of which supposedly compromise the integrity of opposing viewpoints in the spirit of expedience. Accordingly, a scientific politics should not merely satisfy the parties concerned: it should arrive at the 'correct' solution.

To be sure, even this mentality admits of a democratic interpretation, as positivist social researchers have been in the forefront of presenting 'data' from parties whose voices are unlikely to be heard in an open assembly. Typically, this has occurred in surveys designed to represent the full empirical range of a target population. Nevertheless, the question remains of exactly who reaps the political benefits of these newly articulated voices: the people under investigation, the investigators themselves, or the investigators' clients? Moreover, once a target population has been registered empirically, do its members remain 'objects of inquiry' or are they promoted to full-fledged inquirers capable of challenging the original investigators' findings and methods? Probably the most sophisticated treatment of these questions in the context of positivistically inspired US social policy research is to be found in Campbell (1988).

These delicate questions arise because ultimately positivism turns Plato on his side by converting a static hierarchy into a temporal order. Where Plato imagined that authority flowed downward from the philosopher-king in a caste-based social structure, positivists have envisaged that all of humanity may pass (at a variable rate) through a sequence of stages that retrace the socioepistemic journey from captivity to autonomy. In the positivist utopia, it is possible for everyone to be an expert over their own domain. Moreover, there is a recipe for the conversion of Platonism to positivism. It proceeds by isolating a domain of inquiry from the contingencies surrounding its manifestations so that its essential nature may be fathomed. Whereas Plato reserved such inquiry to philosopher-kings, positivists have more often turned to state-licensed professional bodies. And instead of Plato's intellectual intuition (*nous*), positivists attempt to gain epistemic access by comparative historical and experimental methods.

This recipe can be illustrated in the work of Otto Neurath, an organizer of the Vienna Circle. He wanted to isolate the essence of the 'war economy' so that its efficient central planning mechanism can be transferred to environments where it would have more socially salutary consequences. Here Neurath anticipated what Alvin Gouldner would call the Janus-faced character of the 'welfare-warfare state,' whereby the same organizational structure (in this case, a concentration of resources in the nation–state) can have radically different consequences, depending on the supporting political environment. Neverthelesss, as Neurath's many critics pointed out, positivism seems to have inherited Platonism's political naivete, which confuses the fact that, say, the 'war economy' can be identified analytically as a feature of many societies and the analyst's ability to transfer it to new social environments – bar ring the imposition of sufficient force to hold all other environmental factors constant. If anything deserves the name of the 'positivist fallacy' it is this too easy assimilation of the forum to the laboratory. The history of positivism can be neatly captured as a Hegelian dialectic, the three moments of which are epitomized by the work of Auguste Comte (thesis), Ernst Mach (antithesis), and the Vienna Circle (synthesis). However, these moments have historically overlapped, occasionally coming together in figures such as Neurath. The career trajectories of positivism's standard bearers help explain the direction taken by their thought.

Comte was an early graduate of the *Ecole Polytechnique* who believed that its Napoleonic mission of rendering research a vehicle for societal transformation had been betrayed, once he himself failed to achieve a permanent academic post. Mach was a politically active physicist on the losing side of so many of the leading scientific debates of his day that his famous chair in Vienna, from which the logical positivists sprang, was awarded on the strength of his critical-historical studies, not his experimental work. Finally, the intellectual leader of the Vienna Circle, Rudolf Carnap, had to abandon physics for philosophy because his doctoral dissertation topic was seen as too 'metatheoretical' for a properly empirical discipline. For Carnap and others who came of age in World War I, physics had devolved into another specialized field of study, rather than – as it had still been for Einstein – natural philosophy pursued by more exact means.

Thesis: Comtean Positivism

Auguste Comte coined the word 'positivism' to refer to the positive side of the Enlightenment's essentially destructive approach to traditional social and intellectual orders. Comte agreed with the Enlightenment wits that scientific rationality is the model of social rationality, the exemplar of which is Newtonian mechanics. However, according to Comte, the French Revolution had refuted the wits' belief that people's 'natural reason' would come to light once they were freed from their captivity to religion and other traditional forms of authority. Comte thus designed positivism to replace both the church and the sovereign with scientific experts who together would function as a secular priesthood for modern society. 'Sociology,' another of Comte's coinages, referred at once to this political vision and the scientific project that would render it a legitimate form of authority.

Positivism in this basic sense has remained the philosophy underwriting science as a specialized profession of 'scientists' with an identifiable order of merit and career structure. Ironically, none of the movement's first phase protagonists – including Comte himself, John Stuart Mill, and Herbert Spencer – ever called a university home. Nevertheless, they all earned their living from producing knowledge (some would call it 'ideology'), unlike those for whom the pursuit of knowledge was a refined way of consuming wealth. In this important respect, positivism was aligned with the ascendant bourgeoisie and hostile to such aristocratic forms of learning as the gentleman-scientist and the academic humanist. Comte arrived at this position from his mentorship under Count Henri de Saint-Simon, a spiritual godfather of today's management gurus, who at the

dawn of the industrial age anticipated that organized knowledge would be the principal factor of economic production and political control in what Daniel Bell would call, 150 years later, 'post-industrial society.'

Given its explicitly anticlerical posture, positivism was more often found in medical schools and polytechnic institutes than traditional universities. (French universities did not become fully secular institutions until the Third Republic in 1870.) The medical interest is noteworthy because, in the nineteenth century, positivism was used to legitimize everything from phrenological and psychosomatic research to the experimental approach to medicine championed by Claude Bernard. What they all shared was the recognition that physical constitution plays a significant role in defining psychic, or spiritual, life. This basic assumption was sufficient to challenge religion's ministerial monopoly.

More generally, positivism's anticlericalism has been among its most powerful export values, especially in Latin America (where Comte's motto 'Order and Progress' remains emblazoned on the Brazilian flag). Without a thoroughgoing Protestant Reformation, Latin Americans found Comte's secularized version of Catholicism much less utopian than did their European counterparts. To be sure, background cultural allegiances divided positivist loyalties, with Francophile Latin Americans (e.g., Brazil) preferring Comte and Anglophile ones (e.g., Argentina) Spencer. In either case, though, positivism was equally likely to justify authoritarian and democratic regimes. Perhaps the most lasting intellectual residue of positivism's Comtean origins has been the default association that social scientists make between reliable knowledge and statistically significant empirical correlations. Hempel (1942), the canonical positivist account of historical explanation, placed the Vienna Circle's distinctive stamp on this point. Nowadays, its import is easily lost in methodological discussions that trivialize correlations as deficient accounts of causation, proof that positivism owes too much to Francis Bacon. However, following Karl Pearson, perhaps the first self-declared positivist to hold a professorship in Britain, the appeal to correlations served at least four distinct functions.

First, correlations enabled an economy of thought by summarizing observations already made. This would decrease the likelihood of 're-inventing the wheel,' thereby expediting overall human progress. Second, correlations forced investigators to look beyond the singular striking success to more generalizable, albeit mundane, tendencies. Here the positivism helped moderate the hype surrounding new inventions and discoveries. Third, the focus on correlations implied that scientific work could be mechanized by treating individual experiences as instances of more general types. This helped dispel the idea that scientists must have a special 'genius' for finding hidden causes beneath the welter of data. Fourth and probably most significant in Pearson's own turn-of-the-century context, the appeal to correlations echoed the 'uniformitarian' approach to natural history championed by Darwin, whose theory of evolution presupposed that all biological change has occurred by principles that emain in effect today. Thus, there was no need for either a divine creator behind the scenes whose actions transcend observed correlations or a mystification of statistically extraordinary events as 'miracles.'

Antithesis: Machian Positivism

Ernst Mach symbolized the second moment of positivism's dialectic: he gave the movement a dose of its own medicine, namely, a scientific sense of science's own limitations. Mach's counsel emerged in the midst of the ironic, perhaps perversely self-defeating, character of *fin de siecle* Viennese culture, whereby Karl Kraus's quip that psychoanalysis is the disease for which it is cure was eventually incorporated into positivistic lore as Ludwig Wittgenstein's dark saying that philosophy is the ladder that must be climbed in order to be discarded (Janik and Toulmin, 1973).

In the long term, this sense of scientific self-restraint has made its strongest impression in the dual doctrines of 'value neutrality' and 'academic freedom,' which are most closely associated in the social sciences with Max Weber. Although these doctrines normally are treated as separate, it is difficult to motivate either without examining their historical interconnection. The missing link is the idea that, absent further empirical demonstration, the value of theoretical discourse – the most obvious epistemic marker distinguishing experts from the lay public – is purely heuristic. Theory enables those who already know something to know much more; however, it obscures the vision of those who have yet to know anything.

Thus, the various Mach-inspired positivist projects of 'reducing' theoretical discourse to its empirical bases have been motivated by a keen sense of the contingency of when and where new scientific insights arise. The distinction that Hans Reichenbach and Karl Popper drew between the contexts of discovery and justification canonized this point for contemporary philosophy of science. The unique means by which a discovery is made is neither necessary nor sufficient for demonstrating its validity. Elementary logic textbooks demonize the failure to abide by this point as the 'genetic fallacy.' For Mach, the contingent nature of discovery places a special burden on scientists to render their insights as widely as possible, without denying themselves the right to follow it up in their own way. The Vienna Circle positivists may be seen as having compromised this legacy by limiting the context of justification to the received canons of inductive and deductive reasoning. But perhaps they had good reason.

Mach may have been too keen to let the lay public draw whatever conclusions they wished from their studies of science. The second moment of positivism enabled the conversion of scientific knowledge into a pure instrument detached from any prior theoretical or normative commitment. In practice, this 'neutral' stance made science available to the highest bidder, who in turn could hold a monopoly over its use (or nonuse). Consider the fate of the law of diminishing marginal utility (LDMU), which states that if there is enough of a good to satisfy a basic level of need, then each additional increment of the good satisfies the need less than the previous increment. LDMU first appeared in John Stuart Mill's normative political economy, the source of modern welfare economics. The German translation of Mill's *System of Logic* in 1849 had introduced the first phase of positivism to Mach and such contemporaries as Wilhelm Wundt and Wilhelm Dilthey, initiating the long-standing debates over the reducibility of the methods of the *Geisteswissenschaften* (Dilthey's rendering of Mill's 'moral sciences') to those of the *Naturwissenschaften*.

While Mill's later work, *On Liberty*, would inspire Mach's own views linking academic freedom to the right of dissent, Mach refused to follow Mill in drawing normative conclusions from putatively scientific facts, even though Mill's aim of removing the accumulated advantage of the rich was close to Mach's own heart.

Mill treated LDMU as a scientific basis for redistributionist social policies. He inferred that each additional increment of income earned by the rich would improve their lives less than the same increment transferred to the poor. Mill presumed a common standard of interpersonal utility and the idea that different interpersonal endowments could be explained as reversible historical accidents. Moreover, he modeled LDMU on a certain interpretation of Newton's laws, which holds them to be true in an ideal physical medium but not in actual physical reality. Thus, Mill made the ideality of the scientific laboratory stand for the normative basis of social reform. What is now called the 'marginalist revolution' in economics reinterpreted the Newtonian precedent for LDMU in the value-neutral terms that captured the social situation of Mach and other second generation positivists who called a university home.

In the 1870s, Mill's academic nemesis, William Stanley Jevons, had begun interpreting LDMU as the emergent product of interactions between agents whose respective utilities are in principle incalculable, that is, an empirically stable resolution of unknowable constitutive forces. Today we would say that LDMU is a macrolevel market effect, in which the parameters of the relevant market are left unspecified. Thus, cases that fail to conform to LDMU do not force a difficult normative judgement about whether the law or the case is at fault; rather, one casts around for an appropriately defined market to which LDMU might apply, since one may not have considered all the effects over a large enough scale, over a long enough period.

Second generation positivists came to regard LDMU as holding no clear implications for social policy. This opened the door to the curious twentieth century 'noninterventionist' policies that Milton Friedman and others have used to oppose scientific justifications of the welfare state. Here positivism's antitheoretical and antinormative stance is explicitly identified with the inductively self-organizing order of the so-called invisible hand. The shift from Jevons to Friedman is indicative of what logicians call the 'modal fallacy': The denial that X implies Y is interpreted to mean that X implies not Y, where X=LDMU; Y=State-managed redistribution of income. Thus, an economic principle that was seen originally as not necessarily licensing a particular policy intervention came to be seen as an outcome that obviates the need for any policy intervention whatsoever.

Nevertheless, the transition between the first two moments of positivism was neither smooth nor clearly separated from the third. Free market liberals continued to invoke Comte's original sense of positivism to stigmatize large-scale, usually Marx-inspired, social planning of the economy (e.g., Hayek, 1952). A target of these studies was Vienna Circle organizer, Otto Neurath, who supposed that a specialist understanding of the persistent underlying structures of political economy provides a better handle on social policy than the spontaneously aggregated experiences of individual producers and consumers. While Neurath's Marxist orientation to social planning was close to Comte's privileging of expert over lay judgement, he also accepted the Machian view that useful knowledge should be spread as widely as possible. In that sense, Neurath's interest in designing an ideographic language ('Isotype') to enable the populace to understand capitalism's inequities was born of the same impulse as his debates with Rudolf Carnap over the construction of the 'protocol statements' that constitute the fundamental language of science. Here the second moment of positivism naturally shades into the third.

Synthesis: The Vienna Circle

The Vienna Circle was started by Moritz Schlick, another physics-trained philosopher who succeeded Mach to the chair of the history of inductive sciences at the University of Vienna. Much has been made of the Vienna Circle's preoccupations with the conceptual and empirical foundations of science, the demarcation of science from nonscience, and the unification of the special sciences under a single metaphysical and/or methodological rubric. Less has been said to explain the Circle's partial return to positivism's communitarian and authoritarian origins, especially after Mach had made positivism virtually synonymous with the liberalization of epistemic authority. Nevertheless, there were few true liberals in the Vienna Circle: Schlick and Wittgenstein were conservatives, Carnap and Neurath socialists. Popper, the Circle associate closest to Mach's own politics, was not easily accepted into the fold.

This antiliberal turn can be explained by changes in science's social relations in the German-speaking world over the first quarter of the twentieth century. Seen against this background alone, the Vienna Circle appears utopian and even prone to nostalgia; but with Nazism's ascendancy in the 1930s, the 'logical positivist' movement quickly acquired a defiant and progressive cast. At first, the Circle recoiled from Machian positivism because of the ease with which Mach's neutral instrumentalism allowed scientists, especially chemists, to harness their knowledge for aggressive military–industrial ends. Consequently, Germany's humiliating defeat in World War I paved the way for the antiscientific backlash of Weimar culture that formed the backdrop to the Circle's meetings. In recognizably Platonic fashion, the Circle forged a philosophy that protected science from this backlash, while divesting it of any potentially corrupting links to technology. The key was the idea that a genuine science must satisfy certain logical and empirical criteria relating to the public testability of knowledge claims. The model for this requirement was ongoing debates over the foundations of mathematics, which Carnap and Reichenbach had already transferred to debates over the presuppositions of relativity and quantum theory.

However, one decidedly anti-Platonic insight that the Vienna Circle retained from Mach, which increased in importance over the years, was that esoteric forms of knowledge must be 'reducible' to their core 'cognitive' components so as to facilitate both their evaluation and utilization. Specifically, the Vienna Circle aimed to demystify 'metaphysical' jargon that encoded prejudice without extending insight. (Carnap famously cited Heidegger in this respect.) At the peak of its popularity in the 1940s, logical positivism was linked with such American best sellers of the 'general semantics movement'

as Hayakawa (1978). Here the foe had shifted from Heidegger to Hitler, with the positivists arguing against the irrationalism of mass propaganda. Perhaps the most famous single work of the positivist movement, Ayer (1936), written by the British observer at the Vienna Circle, can be seen as a transitional work between these two phases.

Given this history, logical positivism appears in a rather ironic light in the highly publicized *Positivismusstreit* that transpired in 1961–8 between Karl Popper and Frankfurt School doyen, Theodor Adorno. It turns out that both Popper and Adorno demonized 'positivism' in social science for the exact same ills as the logical positivists had demonized Heidegger and Hitler a quarter-century earlier, namely, a 'jargon of authenticity' and its contribution to an 'authoritarian personality' (two relevant Adornian phrases). Moreover, the two antagonists presupposed the logical positivists' ideal of unified inquiry and the inadequacy of defining society as simply the totality of 'social facts.' Where Popper and Adorno clearly differed from the positivists was over the latter's promotion of a sharp fact-value distinction, whereby the presumed unity of factual knowledge offsets the irreducible plurality of value orientations. Here positivism courted a 'relativism' that both antagonists strongly opposed. Nevertheless, the ascendancy of relativism, in the guise of 'postmodernism,' would render Popper and Adorno's wranglings over 'critical reason' beside the point by the 1980s.

In detecting positivism's latent relativism, Popper and Adorno not only anticipated the future; they also recalled the 'value neutrality' that had originally turned Machian positivism into a vehicle of power politics. In the social sciences, Max Weber's distinction between charisma and bureaucracy as modes of legitimation enabled both politicians and scientists to turn a blind eye to each other's activities, as the two groups were supposedly driven by radically different normative orientations: the politician acts, while the scientist provides reasons. Similar criticisms have been made of other movements in the social sciences affiliated with positivism, most notably Hans Kelsen's legal positivism and B. F. Skinner's radical behaviorism. Yet, even Neurath was happy to call his project 'neutral Marxism.'

These cases are united by a refusal to blur the boundary between science and politics, which in practice has meant a rejection of social democratic, welfare-state policies in favor of more extreme political stances: either complete withdrawal or outright authoritarianism. Once again, positivism would seem to have fallen victim to the modal fallacy: because politics and science can contaminate each other, therefore they inevitably will, and hence each must be left to its own devices.

Final Assessment: Does Positivism Have a Future?

After the leading members of the Vienna Circle migrated to the US in the 1930s, logical positivism seeded that country's analytic philosophy establishment for the second half of the twentieth century. However, this is the only context in which positivism possibly dominated an established discipline. For the most part, positivism has been embraced by disciplines that have yet to achieve academic respectability, even in the natural sciences, where Mach found his strongest support among chemists, biomedical scientists, and psychologists – not physicists. (It is often overlooked that positivism's reliance on Newtonian mechanics as the model for all science generally was not appreciated by a physics community jealous to guard its guild privileges.) Unsurprisingly, positivism's most ardent supporters have been social scientists, not for the Comtean reason that sociology is the pinnacle of all science but for the more mundane reason that positivism seemed to offer a strategy for rendering one's activities 'scientific.'

These matters came to a head with the publication of Kuhn (1970), which Carnap enthusiastically endorsed as the final instalment of the logical positivists' *International Encyclopedia of Unified Science*. It provided an account of scientific change based largely on the history of physics that was quickly embraced by social scientists, as Kuhn stressed science's self-organization over its larger societal impact. However, unlike previous positivist accounts, Kuhn's explicitly fixated on particular scientific disciplines (or 'paradigms') without presuming that science as a whole is heading toward a unified understanding of reality. By implication, then, all scientific standards are discipline-relative. Kuhn's approach suited what is now called the 'postmodern condition,' which the *Positivismusstreit* had desperately tried to stave off.

In retrospect, the popularity of Kuhn (1970) is better understood as signifying positivism's decadent phase than, as it was originally seen, a fundamental challenge to positivism. The decadent state of contemporary positivism can be gleaned from the ongoing debates over the scientific status of 'sociobiology.' Each side of the debate captures part of the positivist legacy: proponents like E.O. Wilson hark back to positivism's Comtean vision of unified knowledge in aid of societal transformation, while opponents like Richard Lewontin draw on positivism's more recent restriction of the bounds of science to discipline-specific criteria.

If positivism has a future, it lies in rekindling a sense of 'Science' that transcends the boundaries of particular scientific disciplines. This is how Comte originally thought about the discipline he called 'sociology.' He claimed that it was the last to develop, not simply because of the complexity of its human subject matter, but more importantly, sociology had to reconstitute the (natural) sciences that historically preceded it. Too often the history of positivism's quest for unified science has been interpreted as exclusively a matter of applying the methods of physics to less developed sciences. The reciprocal movement is really more important, namely, the application of sociological findings to the future direction of science as a whole. Such interdisciplinary projects as 'social epistemology' and the 'science of science' have tried to fulfill this side of the Comtean promise, which may be redeemed once the present intellectual climate changes.

See also: Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Empiricism, History of; Enlightenment; Kuhn, Thomas S. (1922–96); Logical Positivism and Logical Empiricism; Mill, John Stuart (1806–73); Popper, Karl Raimund (1902–94); Positivism, Sociological; Wittgenstein, Ludwig (1889–1951).

Bibliography

Adorno, T. (Ed.), 1976. The Positivist Dispute in German Sociology. Heinemann, London (trans. Adey G, Frisby D).

Ayer, A.J., 1936. Language, Truth and Logic. Gollancz, London.

Campbell, D.T., 1988. Methodology and Epistemology for Social Science: Selected Papers. University of Chicago Press, Chicago.

Cartwright, N., Cat, J., Fleck, L., Uebel, T., 1996. Otto Neurath: Philosophy between Science and Politics. Cambridge University Press, Cambridge, UK.

Fuller, S., 1993. Philosophy of Science and Its Discontents, second ed. Guilford Press, New York (orig. 1989).

Fuller, S., 2000. Thomas Kuhn: A Philosophical History for our Times. University of Chicago Press, Chicago.

Hayakawa, S.I., 1978. Language in Thought and Action, fourth ed. Harcourt Brace and Jovanovich, New York (orig. 1939).

Hayek, F., 1952. The Counter-Revolution of Science: Studies on the Abuse of Reason. University of Chicago Press, Chicago.

Hempel, C.G., 1942. The function of general laws in history. Journal of Philosophy 39, 35–48.

Janik, A., Toulmin, S., 1973. Wittgenstein's Vienna. Simon and Schuster, New York.

Kolakowski, L., 1972. Positivist Philosophy From Hume to the Vienna Circle. Penguin, Harmondsworth, UK (trans Guterman N).

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago (orig. 1962).

Proctor, R., 1991. Value-Free Science? Purity and Power in Modern Knowledge. Harvard University Press, Cambridge, MA.

Turner, S.P., 1986. The Search for a Methodology of Social Science. Kluwer, Dordrecht, Netherlands.

Zea, L., 1963. The Latin American Mind. University of Oklahoma Press, Norman, OK (trans. Abbott J A, Dunham L).

Positivism, Sociological

RW Outhwaite, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.H. Turner, volume 17, pp. 11827–11831, © 2001, Elsevier Ltd.

Abstract

Positivism and sociology have a common origin, and positivism remains a significant approach in sociology and the other social sciences. In positivist sociology, the scientific study of the social world is identified with empirical research, statistical methods, and often the pursuit of general laws of social life which can be tested against experience.

The Rise of Positivism

Auguste Comte (1798–1857) argued in his *Course of Positive Philosophy* (1830–1842) that the various sciences developed through three stages – theological, metaphysical, and, eventually, positive. A positive science does not rely on spirits or abstract forces but on precise relationships of the kind expressed in Newton's law of gravity. The sciences move gradually into the positive stage, with sociology representing the culmination of this process, in the work of Comte himself. Comte, a strong influence on John Stuart Mill (1806–73) popularized the name 'sociology' to refer to what had earlier been called social physics, and a positivist approach was taken forward in sociology by Herbert Spencer (1820–1903) and later Emile Durkheim (1858–1917). Spencer made much of the parallels, as he saw them, between biological and social evolution and between societies and (other) organisms and the functional relations between their component parts. Durkheim's evolutionism was more muted, though his study of religion looked to Australian totemism for the *Elementary Forms of Religious Life* (1912) and his earlier *Rules of Sociological Method* (1895) argued for the study of 'social facts' as 'things.' Insisting on the distinctiveness of sociology and in particular its difference from psychology, Durkheim argued, notably in his classic study of *Suicide* (1897), that one should look for the social causes of regular differences in suicide rates in religious attachments or marital status, rather than invoking psychological explanations or speculating about the motives of suicidal individuals. Durkheim's distinction between 'normal' and 'pathological' states of society echoes Comte's principle that science can lead to prediction and thereby to the control of natural or social processes.

Positivism also had a presence in the study of history, for example in the work of Henry Thomas Buckle (1821–62) and others across Europe, and it was in Germany that this provoked the strongest antipositivist reaction, with the philosopher of history J.G. Droysen (1808–84) and his contemporary Wilhelm Dilthey (1833–1911), insisting on the distinctiveness of what came to be called the 'human sciences' or Geisteswissenschaften. In these sciences, as Dilthey put it, the mental activity of humans and of some other animals, and its products, can be *understood*. Dilthey developed what we would now call a research program for history and the other human sciences based on the distinctiveness of human psychic expressions and the understanding of those expressions. In a move which was to become a definitional feature of later interpretive social science, Dilthey emphasized the continuity between everyday understanding and more formal processes of interpretation.

In sociology, Georg Simmel (1858–1918) and Max Weber (1864–1920) also argued strongly for the importance of understanding (Verstehen) in history and the other social sciences, and Weber presented in his posthumously published *Economy and Society* (1920) the principle that sociology, as he conceived it, should attempt the interpretive understanding of social processes in order to explain them. His earlier classic study of *The Protestant Ethic and the 'Spirit' of Capitalism* (1904–1905) had applied this two-track approach by first showing (at least to his own satisfaction) that ascetic protestants such as Calvinists were *in fact* more economically innovative than Catholics in early modern Europe and then offering an explanation which '*made sense*' of their conduct.

By 1920, then, Weber had developed a strong alternative to positivistic sociology and offered a way of combining understanding and causal explanation which continues to attract many sociologists. In the years following his death, however, sociology became more polarized between more radical versions of both positivist and interpretivist approaches.

The Vienna Circle and the Transformation of Positivism

Positivist philosophy of science had taken a strongly empiricist direction in the work of Ernst Mach (1838–1916) who inspired the Vienna Circle, which met regularly in 1910–12 and again from 1926 into the 1930s. These physicists, mathematicians, and philosophers founded the Ernst Mach Society in 1928 and published, in 1929, a manifesto under the editorship of the sociologist Otto Neurath (1882–1945) (Neurath et al., 1929).

Comte's positivistic conception of science had stressed theory and opposed empiricism, arguing that "the next great hindrance to the use of observation is the empiricism which is introduced by those who, in the name of impartiality, would interdict the use of any theory whatsoever" and that "No real observation of any kind of phenomena is possible, except in as far as it is first directed, and finally, interpreted, by some theory." Viennese positivism or logical empiricism drew more on the formalization of logic and mathematics in the early years of the twentieth century and its extension to the philosophical analysis of language by Ludwig Wittgenstein (1889–1951),

whose *Tractatus Logico-Philosophicus* (1922) was a strong influence on members of the Circle. As they saw it, language maps onto the world through precise observation statements, to which the language of science has to be ultimately reducible. Whereas Comte had stressed the distinct domains of the various sciences, on this model the social sciences were reducible to psychology and ultimately to physics, which was no longer just the lead science in the march toward a positive state, as it had been for Comte, but the fundamental basis of 'unified science.'

This approach was disseminated in the 1930s as the members of the Vienna Circle fled from Nazism, with several figures settling in the United States, where a similar approach to science had developed independently (Platt, 1996). (The American physicist Percy Williams Bridgman's operationalism, had, for example, been a major influence on the Vienna Circle itself.) Two philosophers who popularized logical empiricism were Alfred Ayer (1910–89) in the United Kingdom and Karl Popper (1902–94), who emigrated from Austria to New Zealand and then also settled in London. Popper introduced the most important modifications to the Vienna Circle model and also addressed issues of social and political philosophy. One of the central principles of logical empiricism had been the 'verification principle,' the Bridgmanian idea that the meaning of an empirical statement depends on, or in some versions is identical with, its means of verification. (One consequence, enthusiastically promulgated by Ayer (1936), was that value judgments were meaningless, being neither empirical nor logical or mathematical.) Popper (1934) demonstrated the impossibility of conclusive verification in science and stood the problem on its head, arguing that testability was better understood in terms of falsification. The statement that all swans are white can never be conclusively *verified*, unless we are sure that we have observed all swans that ever existed and that they are now extinct, whereas the statement can be *falsified* by the observation of just one black swan. Scientific theories expose themselves to testing, as Einstein had done by making a prediction about the position of Mercury which was proved accurate in 1910. In Popper's analysis, this did not prove the theory but merely gave it greater 'verisimilitude' since it had passed a crucial test. Like a car which has passed an annual safety inspection, it could be conditionally relied on until the next time. Pseudotheories by contrast, such as Marxism and psychoanalysis, could amass lots of supporting evidence but did not expose themselves to falsification. Social science, Popper argued, could make conditional predictions, but should avoid the temptation of 'prophecy.'

Positivism and Sociology

This modified version of Viennese positivism has been massively influential in the philosophy of science, including social science. Popper had undercut the search for certainty which had been the Achilles heel of logical empiricism and offered an alternative model of scientificity. In the social sciences, new quantitative techniques were developed for theory testing, and punched cards gave way to computers. In the 'covering-law' model of causal explanation, empirical findings were attached to general laws in the manner alleged to operate in the natural sciences. (In the classic example, the general law that water freezes at 0° explains the state of my car radiator after a cold night.) This covering-law approach to explanation coexisted with the rise of functionalism; Talcott Parsons suggested, quite implausibly, that his system theory might one day be open to empirical testing. Max Weber's plea for the avoidance of value judgments *in social science* (which in Weber's practice coincided with an almost existentialist position on our choice of ultimate values and a variety of intensely political engagements) became the dominant position justifying an *apolitical* stance for social science – which in practice tended to mean an endorsement of the ideological status quo.

However, the history of positivism, like that of Stalinist economic planning, is largely the history of attempts to modify or reform it, and it unraveled from a number of directions. In philosophy, thinkers influenced by North American pragmatism such as W.V.O. Quine (1908–2000) emphasized the importance of theoretical frameworks, questioning the idea of piecemeal testing or instance confirmation. More seriously for Popper, the historian of science Thomas Kuhn (1962) showed that scientists tend to work most of the time within unquestioned frameworks or 'paradigms,' switching from one to another only in periods of scientific revolution, in a process close to religious conversion. Paul Feyerabend (1924–94) drew the extreme conclusion that a good empiricist should be an epistemological anarchist and accept that 'anything goes': all theories should be taken seriously. Wittgenstein (1953) had dropped his earlier theory of language in favor of a much more holistic and sociological approach to 'language games' located in 'forms of life.' This linked up with antipositivist interpretive approaches in sociology and social anthropology, as demonstrated by the British philosopher Peter Winch (1958). Meanwhile the Canadian philosopher Charles Taylor (1964), influenced both by analytic philosophy and by continental European hermeneutics, argued that the behavior of humans and other higher animals should be understood in terms of intentions and purposes rather than just causal impulses.

In a further twist, Karl-Otto Apel (1967) demonstrated in more detail the affinities between analytic philosophy of language and the Geisteswissenschaften tradition, and Jürgen Habermas, who had joined in the critique of positivism in West German sociology in the early 1960s (Adorno et al., 1976), showed how phenomenology, analytic philosophy and Hans-Georg Gadamer's version of philosophical hermeneutics could combine against social scientific positivism and yield a 'critical social science' (Habermas, 1968a,b). This type of science combined causal explanation and understanding in an approach, exemplified by psychoanalysis and the marxist critique of ideology, which identified and removed causal obstacles to understanding. Positivism, for Habermas (1968b: p. vii) and his associate Albrecht Wellmer (1967), meant an absence of reflection and the reduction of epistemology to methodology, whether in natural or social science. In the latter case, however, the imitation of a natural scientific model was particularly dangerous because it reinforced tendencies to technocracy which undermined the possibilities of political deliberation. Habermas later reformulated his model of emancipatory science, but it remains in a modified form in his idea of reconstructive sciences, which explicate capacities which we exercise in daily life. His initial example was linguistics, which identifies the rules which we

unconsciously follow as native speakers of a language, but his own theory of communicative action (Habermas, 1981) also embodies this approach.

Whereas critical theory tends toward a dualism of natural and social science, but in an increasingly muted form, the assumption that opposition to positivism also entailed dualism or antinaturalism was, however, also put in question in the late 1960s and early 1970s by the realist metatheory of science developed by Rom Harré and Roy Bhaskar. Both Harré and Bhaskar, like Habermas, were substantially motivated by the desire to undermine positivistic theories and approaches in the social sciences. They were interested in giving a more adequate account of science as a whole, in world composed of relatively enduring structures and mechanisms. Some of these could be isolated in scientific experimentation, given the contingent existence of *Homo sapiens* and *Homo scientificus*. An important aspect of the realist program developed by Harré, Bhaskar, and others was a conception of explanation as involving not an essentially semantic reduction of causal statements to general laws (which in sociology tended to be either vacuous or riddled with exceptions) but a reference to the causal powers of entities, structures, and mechanisms. The real explanation of the state of my radiator is not the mere generalization that water freezes at 0° but the causal properties of water molecules: their capacity to freeze or evaporate. Causal tendencies may be outweighed by countervailing tendencies, and two causal tendencies may neutralize one another, as do the centrifugal force of the Earth's rotation and its gravitational attraction, with the convenient consequence that human beings and other animals are safely anchored close to the Earth's surface but can jump over small obstacles.

This and other features of realism meant that the whole issue of naturalism could be rethought. Human beings could be seen as having causal powers and liabilities, just like other entities; it no longer mattered if their relations rarely sustained any universal generalizations of an interesting kind, but only sets of tendencies regular enough to be worth exploring. It also seemed natural to include among the causes of human action the agents' reasons for acting – reasons which must be understood as far as possible. Reasons, such as a smoker's reasons for giving up the habit, might form part of a causal explanation of the cure. The fact that many of the entities accorded causal force in social scientific explanations were necessarily unobservable was not, as it was for empiricism, a problem of principle. And the understanding of meaning could, as Bhaskar (1979: pp. 58–59) put it, be seen as in some ways equivalent to measurement in the natural sciences. Realists diverge on the issue of naturalism, with Bhaskar arguing for it and Harré defending a more interactionist and social constructionist position, focusing on interpersonal action and questioning the existence of social structures.

Realism found a more comfortable home in sociology and the emergent discipline of international relations than in academic philosophy, and in Britain and Scandinavia more than in other parts of Europe or in North America. Together, however, all these antipositivist currents meant that it was even harder than before to find explicit defenders of positivism. Jonathan Turner (1981, 1985, 1992, 2013) in the United States is a significant example, and Peter Blau also identified explicitly with positivism. For the most part, however, positivism tended to be equated, by both supporters and critics, with empirical research and in particular with quantitative methodologies. Turner is an exception, since his positivistic defense of scientific theory in sociology was also highly critical of the vogue for quantification in the United States. As he noted in an earlier version of this article, positivist books on 'theory building' and 'theory construction' in the late 1960s, beginning with Hans Zetterberg's (1965) *On Theory and Verification in Sociology* and continuing through the 1960s and 1970s (e.g., Blalock, 1969; Dubin, 1969; Gibbs, 1972; Hage, 1972; Reynolds, 1971), did not retain their initial appeal at a time when functionalism was also going into decline.

As the theoretical wave of the later twentieth century passed over sociology and began to dissipate, with more interdisciplinary conceptions of 'social theory' tending to break away from mainstream sociology, the theoretical debates between positivism and hermeneutic and other antipositivist approaches tended to be trivialized into a choice between quantitative and qualitative methods or between 'empiricism,' identified with any empirical research, and 'theory,' meaning any sociological work not involving empirical research. Some prominent figures labeled as theorists did empirical research, for example, Pierre Bourdieu in France; others, like Zygmunt Bauman or Anthony Giddens in the United Kingdom, did not, but this was not generally felt to be a significant difference.

The Prospects for Positivism

Positivism in the very general sense of an aspiration to scientificity and to the construction and empirical testing of formal theories remains a significant presence in contemporary sociology, and to a much greater extent in economics, political science, and some parts of international relations. As Peter Halfpenny (2001: p. 382) notes,

> …large numbers of investigators continue to produce research that conforms to the positivist image of science, explaining social activities in loosely deductive-nomological terms, and these explanations are used to guide and evaluate a wide variety of social programmes.

It may be that the important dividing line is now between formal conceptions of theory and more informal ones such as the 'theory' of structuration or risk society or theoretical approaches such as actor-network theory. This dividing line (Turner, 2013; Outhwaite, 2014) runs through several theory families, such as rational choice or rational action theory, which encompasses a highly formalized, even formalistic mainstream, following Gary Becker (1976), and more flexible and sophisticated approaches such as those of Martin Hollis (1938–98), for whom "rational action is its own explanation" (Hollis, 1977: p. 135) or Jon Elster (1989, 2007). Hollis and Nell (1975) deconstructed the powerful alliance between positivism and neoclassical economics, while Hollis and Smith (1990) provided what

remains one of the best statements of what is at stake in the oppositions between 'explaining' and 'understanding' in the social sciences and between 'agency' and 'structure.' As Smith wrote in the exchange with which they end the book:

> I, too, reject the Positivist notion that there is a world waiting to be mapped. There may be regularities in human affairs but I do not accept the idea that we can construct a neutral theory, valid across time and space, that allows us to predict in the same way as occurs in the natural sciences. (Hollis and Smith, 1990: p. 203)

Evolutionary sociology, mentioned above in relation to Herbert Spencer, has recently experienced a certain revival, though in forms which are mostly dismissive of Spencer's earlier efforts (O'Malley, 2007). Marxism is another example, with formal and mathematical reformulations of central Marxist propositions (including French structuralist Marxism, analytical Marxism, also known as 'no-bullshit Marxism,' and 'rational choice Marxism'), coexisting with more humanistic variants which shade off into critical theory, in both its Frankfurt School and broader literary and cultural variants. As Peter Halfpenny (1982: p. 120) concluded,

> Positivism may be dead in that there is no longer an identifiable community of philosophers who give its simpler characterisations unqualified support, but it lives on philosophically, developed until it transmutes into conventionalism and realism.

Since then, the term 'postpositivism' has come into use. Like most 'post' formulations, it tends to be used in a wide variety of senses, ranging in this case from modifications of an essentially still positivist program to strong versions of social constructionism (Alvesson and Sköldberg, 2010). 'Positivism' has tended to become a term of abuse, while the positivistic impulse toward formalization and (presumed or hoped-for) certainty continues to pervade our increasingly rationalized and performance-oriented social sciences.

See also: Comte, Auguste (1798–1857); Durkheim, Emile (1858–1917); Hermeneutics, History of; Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Positivism, History of; Sociological Theory; Sociology, Epistemology of; Spencer, Herbert (1820–1903); Wittgenstein, Ludwig (1889–1951).

Bibliography

Adorno, T.W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., Popper, K.R., [1972]1976. The Positivist Dispute in German Sociology. Harper & Row, New York.
Alvesson, M., Sköldberg, K., 2010. Reflexive Methodology. New Vistas for Quantitative Research. Sage, London.
Apel, K.-O., 1967. Analytic Philosophy of Language and the Geisteswissenschaften (H. Holstelilie, Trans.). D. Reidel, Dordrecht and New York.
Ayer, A.J., [1936]2001. Language, Truth and Logic. Penguin, 2001, London.
Becker, G., 1976. The Economic Approach to Human Behavior. Chicago University Press, Chicago.
Bhaskar, R., 1979. The Possibility of Naturalism. Harvester Press, Brighton.
Blalock Jr., H.M., 1969. Theory Construction: From Verbal to Mathematical Formulations. Prentice Hall, Englewood Cliffs, NJ.
Comte, A., 1830–1842. Cours de philosophie positive. Bachelier, Paris, Various partial translations into English.
Dubin, R., 1969. Theory Building. Free Press, New York.
Durkheim, E., [1895]1982. The Rules of Sociological Method: And Selected Texts on Sociology and its Method. In: Steven Lukes (Ed. and Introduction), (W.D. Halls, Trans.). Macmillan, London.
Durkheim, E., [1897]2002. Suicide: A Study in Sociology (John A. Spaulding, George Simpson, Trans.). Routledge, London.
Durkheim, E., [1912]1995. The Elementary Forms of Religious Life (Karen E. Fields, Trans.). Free Press, New York; London.
Elster, J., 1989. Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.
Elster, J., 2007. Explaining Social Behaviour: More Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.
Gibbs, J., 1972. Sociological Theory Construction. Dryden Press, Hinsdale, IL.
Giddings, F., 1901. Inductive Sociology. Macmillan, New York.
Habermas, J., [1968a]1988. On the Logic of the Social Sciences (Shierry W. Nicholsen, Jerry A. Stark, Trans.). MIT Press, Cambridge.
Habermas, J., [1968b]1987. Knowledge and Human Interests (Jeremy J. Shapiro, Trans.). Polity, Cambridge.
Habermas, J., [1981]1987. The Theory of Communicative Action, 2 vols. (Thomas McCarthy, Trans.). Polity, Cambridge.
Hage, J., 1972. Techniques and problems of theory construction in sociology. Wiley,, New York.
Halfpenny, P., 1982. Positivism and Sociology: Explaining Social Life. Allen & Unwin, London.
Halfpenny, P., 2001. Positivism in the twentieth Century. In: Ritzer, G., Smart, B. (Eds.), Handbook of Social Theory. Sage, London, pp. 371–385.
Halfpenny, P., McMylor, P. (Eds.), 1994. Positivist Sociology and its Critics, 3 vols. Edward Elgar, Aldershot.
Hollis, M., 1977. Models of Man. Oxford University Press, Oxford.
Hollis, M., Nell, E.J., 1975. Rational Economic Man. Cambridge University Press, Cambridge.
Hollis, M., Smith, S., 1990. Explaining and Understanding International Relations. Clarendon Press, Oxford.
Johnson, W.M., 1983. The Austrian Mind: An Intellectual and Social History, 1848–1938. University of California Press, Berkeley, CA.
Keat, R., 1981. The Politics of Social Theory: Habermas, Freud and the Critique of Positivism. Blackwell/University of Chicago Press, Oxford.
Kuhn, T., [1962]1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.
Mach, E., 1893. The Science of Mechanics (T.J. McCormack, Trans.). Open Court, La Salle, IL.
Montesquieu, C., 1900. The Spirit of Laws, 2 vols. London. Colonial [1748].
Neurath, O., et al., 1929. Wissenschaftliche Weltauffassung - der Wiener Kreis. In: Neurath, Marie, Cohen, Robert (Eds.), (Marie Neurath, Trans.), Empiricism and Sociology. Veröffentlichungen des Vereins Ernst Mach, Vienna (D. Reidel Publishing Company, Dordrecht-Holland; Boston-USA).
O'Malley, M., 2007. Evolutionary approaches in the social sciences. In: Outhwaite, W., Turner, S.P. (Eds.), The Sage Handbook of Social Science Methodology. Sage, London, pp. 333–357.
Outhwaite, W., 2014. Sociological theory. In: Holborn, M. (Ed.), Contemporary Sociology. Polity, Cambridge.
Platt, J., 1996. A History of Sociological Research Methods in America 1920–1960. Cambridge University Press, Cambridge.
Popper, K., 1959. The Logic of Scientific Discovery. Hutchinson, London.
Popper, K., 1969. Conjectures and Refutations. Routledge & Kegan Paul, London.
Reynolds, P.D., 1971. A Primer in Theory Construction. Bobbs-Merrill, Indianapolis, IN.
Spencer, H., 1873. The Study of Sociology. Routledge & Kegan Paul, Boston, MA.
Taylor, C., 1964. The Explanation of Behaviour. Routledge & Kegan Paul, London (Humanities Press, New York).
Turner, J.H., 1985. In defense of positivism. Sociological Theory 3, 24–30.
Turner, J.H., 1992. The promise of positivism. In: Seidman, S., Wagner, D.G. (Eds.), Postmodernism and Social Theory. Blackwell, Oxford, UK, pp. 75–94.
Turner, J.H., 2013. Contemporary Sociological Theory. Sage, Thousand Oaks, CA.

Turner, S.P., Turner, J.H., 1990. The Impossible Science: An Institutional Analysis of American Sociology. Sage, Newbury Park, CA.
Weber, M., [1904–1905]2009. The Protestant Ethic and the Spirit of Capitalism (Stephen Kalberg, Trans. and Introduced). Oxford University Press, New York; Oxford.
Weber, M., [1920]1968. Economy and Society: An Outline of Interpretive Sociology. In: Guenther Roth, Claus Wittich (Eds.), (Ephraim Fischoff et al., Trans.). Bedminster Press, New York.
Wellmer, A., 1967. Methodologie als Erkenntnistheorie: Zur Wissenschaftslehre Karl R. Poppers. Suhrkamp, Frankfurt am Main.
Winch, Peter, 1958. The Idea of a Social Science and its Relation to Philosophy. Routledge and Kegan Paul, London.
Wittgenstein, L., [1922]1971. Tractatus Logico-philosophicus (D.F. Pears, B.F. McGuinness, Trans.). Routledge & Kegan Paul, London.
Wittgenstein, L., 1953. Philosophical Investigations (G.E.M. Anscombe, Trans.). Basil Blackwell, Oxford.
Zetterberg, H.L., 1965. On Theory and Verification in Sociology, third ed. Bedminster Press, New York.

Possession, Linguistic

Jüergen Broschart, The GEO Magazine, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In linguistics, possession is conceived as a relationship of appurtenance between a possessor and a possessed (*possessum*, *possessee*) for which there are specific paradigms of expression in the languages of the world. On the conceptual side, this relationship of appurtenance ranges from material possession via kinship and body-part relationships to the domain of abstract relations (someone's ideas, someone's activities, etc.). On the formal side, one distinguishes mainly between attributive possession (e.g., 'my house,' 'the man's house') and predicative possession ('the house is mine,' 'the house belongs to me,' 'I have a house'). However, it is not always easy to distinguish the conceptual domain of possession from other cognitive domains, and also on the formal side there is a continuum of expression rather than some strictly defined set of constructions. Especially in the area of inalienable possession (body parts, kinship relations) one often finds peculiar constructions where the possessor of body parts is 'raised' to become an argument of the verb as in 'he hit me on the head' versus 'he hit on my head.' Generally speaking, the proper understanding of linguistic possession includes matters of the relative discourse prominence of possessor and possessum in the event described, depending on criteria like animacy, egocentricity, etc.

In linguistics, possession is conceived as a relationship of appurtenance between a possessor (P'or) and a possessed (P'ed, possessum, possessee) for which there are specific paradigms of expression in languages of the world. On the conceptual side, this relationship of appurtenance ranges from material possession via kinship and body-part relationships to the domain of abstract relations (someone's ideas, someone's activities, etc.). On the formal side, one distinguishes mainly between attributive possession (e.g., 'my house,' 'the man's house') and predicative possession ('the house is mine,' 'the house belongs to me,' 'I have a house').

However, it is not always easy to distinguish the conceptual domain of possession from other cognitive domains, and also on the formal side there is a continuum of expression rather than some strictly defined set of constructions. For instance, in many languages (e.g., in English or Icelandic) constructions like 'I have ploughed an acre' (expressing an event in time) historically go back to possessive expressions whose literal meaning was 'I have an acre as something ploughed,' and there are also numerous ways of crossing the boundary between attributive and predicative sentential constructions. For instance, many languages alternate between attributive constructions like German *der Sohn des Mannes* (ART:NOM son:NOM ART:GEN man:GEN, 'the man's son'; where ART = article, GEN = genitive, NOM = nominative) and *dem Mann sein Sohn* (ART:DAT man:DAT his:NOM son:NOM, lit. '(to) the man his son'), where the dative constituent (DAT) functioned originally as an indirect object in predications like *sie entführten dem Mann seinen Sohn* (they:NOM kidnapped ART:DAT man:DAT his:ACC son:ACC, lit. 'they kidnapped (to) the man his son'; where ACC = accusative). These latter constructions (which are typically associated with some psychological involvement on the part of the P'or) are known as *dativus sympaticus*-constructions or are referred to in the context of possessor raising or external possession.

Since psychological involvement also correlates with the relative 'closeness' and 'inalienability' of the possessive relation, it is customary to distinguish between constructions for inalienable and alienable possession. Alienable possession normally refers to material belongings as opposed to kinship and body-part relationships, and there are many languages that have standardized formal ways of distinguishing these domains. One has to bear in mind, however, that the main purpose of these constructions is not the expression of possession as such (see below). Rather, these constructions typically reflect considerations of discourse pragmatics with respect to two nominal referents. In other words, while the concept of possession extends beyond linguistics, the formal category of possession carries properties that are specific to language and linguistics, respectively. This suggests that to a certain extent linguistic possession must be differentiated from nonlinguistic concepts of possession.

Approaches to Possession

The Literature on Possession

The linguistic literature on possession concentrates on rather different aspects. Seiler's monograph (Seiler, 1983) is still the most comprehensive attempt at dealing with the cross-linguistic spectrum of constructions, which according to Seiler can be placed on a formal and semantic continuum in the domain of possession (see Section Details on Typological Research on Possession below). Heine (1997) and, more recently, Stassen (2009) focus mainly on predicative possession and on typical ways in which such constructions develop in the process of grammaticalization. Attributive possession is usually extensively described in traditional grammars under the headings of possessive pronouns and genitival constructions. McGregor and Chappell (1996) focus on the typology of inalienable possession, and Payne and Barshi (1999) focus on the typology of external possession. Other publications include discussions on the grammatical status of specific items from the domain of possession. For instance, it is sometimes

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52040-6

unclear whether possessive pronouns behave like genitival arguments or like adjectives. The discussions in this area are rather technical and must be seen in the context of DP-analysis (for reference see Abney, 1987; Löbel, 1995). As far as language typology is concerned, in the last decade, substantial work has been added on the various ways in which possession is expressed in many different languages of the world, for instance by Motoki (2010), Grkovic-Major (2010), Gunkel and Zifonun (2011), Koptjevskaja-Tamm (2002, 2003), McAnallen (2011), and Stassen (2009). Also the *World Atlas of Language Structures* by Haspelmath et al. (2005) presents an abundance of typological and geolinguistic facts. In what follows, this article will focus on some important issues from the typology of possessive constructions.

Details on Typological Research on Possession

The most difficult problem for the typological description of possession is the great variety of expression and the fact that there is no simple correspondence between a conceptual relationship of possession and traditional linguistic categories like possessive pronouns or possessive verbs (in fact, possessive verbs are cross-linguistically rather rare). For instance, a Latin construction like *mihi est liber* (me:DAT is book:NOM, lit. 'to me is (a) book') contains no 'classical' possessive category, but conceptually it figures as a possessive construction. Conversely, a construction like 'my train' in 'I've got to catch my train' does not really refer to the train which I own, though it does involve a 'classical' possessive pronoun. Finally, there are instances of so-called 'possessor deletion' as in the German *er hob die Hand* (lit. 'he raised the hand') versus the English 'he raised his hand,' which unlike the German example contains an explicit possessive pronoun.

As a solution to these problems, Seiler (1983) proposes that the expression of possession constitutes a cognitive operational program ranging from minimal explicitness of the relationship (indicativity) to maximal explicitness (predicativity). Indicativity literally means to 'point' at what one may take for granted, while predicativity establishes a relationship by overt means. The 'indicative' and less explicit techniques are mainly characteristic of the attributive pole (with the same constructions ranging from the clearly possessive 'my house' to fairly unspecific relations like 'my train'), while explicitly possessive verbs belong to the most predicative type of construction (with constructions like 'I own a house').

In this context, it is no coincidence that inalienable, inherent possessive relationships usually need less formal 'effort' to be encoded than alienable relationships – as the relationship is usually self-understood, it literally goes without saying. For example, in the Amerindian language Cahuilla the expression 'his eye' would just involve the direct juxtaposition of an uninflected personal pronoun and the lexical stem for 'eye,' while the expression 'his dog' needs the addition of a 'possessive classifier' meaning 'pet': compare *he-puc* (3.SG-eye, 'his eye'; where SG = singular) versus *he-ʔas ʔawal* (3.SG-pet dog, lit. 'his pet dog'). Here, the (inalienably conceived) body-part term 'eye' is inherently relational, such that the possessive relation between 'he' and 'eye' is more or less 'self-understood' and need not be made explicit; for 'dog' the situation is different, so that the relationship between 'he' and 'dog' must be established by some relational item like 'pet.' By way of contrast, one would need far more effort in Cahuilla to use a word like 'eye' in a 'derelationalized' sense. Thus, while 'the dog' is simply *ʔawal*, 'the eye' (in some abstract sense) would be *puc-il* (eye-DER; where DER = derationalizer).

Seiler lists a number of techniques commonly employed for the expression of possessive relationships, which can all be placed on a continuum between maximal implicitness and maximal explicitness of the relationship. These techniques include plain juxtaposition of P'or and P'ed ('N–N-constructions,' see above), connectives (CONN), possessive classifiers, (possessive) cases, existential copula constructions, and possessive verbs. The Persian *iza:fa*-construction would be an example for connectives (cf *ha:ne-je pedar* (house-CONN father, 'the father's house')), and for possessive classifiers see Cahuilla. Latin *domus patris* (house:NOM father:GEN) is an instance of a case construction. A case construction is a more 'explicit' technique than the connective technique in the sense that so-called 'cases' may allow far more oppositions than all-purpose connectives, etc.; for instance, apart from using a possessive genitive, many languages also employ locatives in attributive possession (cf Modern Icelandic *höfuðið á mér* (lit. 'the head on me,' 'my head'), as opposed to *augun í mér* (lit. 'the eyes in me,' 'my eyes'), which is a fairly explicit way of specifying how the body-part relates to the P'or). Case constructions also include expressions like the German *der Mann mit dem Hut* ('the man with the hat'). Here, however, the relationship of possession is expressed from a different perspective than *der Hut von dem Mann* (lit. 'the hat from the man,' 'the man's hat'); in the former construction, the nucleus is the possessor, while in the latter the nucleus is the possessed. (For the change of perspective also see Section The Linguistic Side of Possession below.)

With the help of copulas, the case constructions can often be turned into literally predicative expressions of possession (cf the German *der Hut ist von dem Mann* (lit. 'the hat is from the man,' 'the hat belongs to the man') and Icelandic *ég er með bíl* (lit. 'I am with (a) car,' 'I have a car with me')). Finally, languages may also employ actual verbs of possession as in 'I have or own a book' and 'the book belongs to me,' etc. This technique allows the greatest number of oppositions (the number of lexical verbs is virtually limitless), and since this technique is the most predicative one in the sense of Seiler, it is hardly 'natural' in the context of inalienable and 'self-understood' possession. For example, clauses like 'I have a father' or 'I have a head' are considered deviant, unless modifications are added (cf 'I have a nice father,' 'I have a swollen head'). Nevertheless, though in principle inalienable possession tends toward the 'indicative' pole, and alienable possession tends toward the 'predicative' pole, there are always possibilities of making the implicit explicit and the explicit rather implicit. For instance, Indo-European possessive pronouns probably go back to locative constructions of much the same type as the ones which are now commonly used for attributive body-part possession in Icelandic (see above), and the Icelandic attributive constructions in question like *höfuðið á mér* (lit. 'the head on me') originate in turn from predicative *dativus sympaticus*-constructions like 'he hit (on) me on the head.' Therefore, there appears to be some shift from the more explicit and predicative to the less explicit in the process of grammaticalization. At the same time, as a reaction to

a decrease of explicitness, new explicit constructions may be coined. In this context, Seiler claims that indicativity and predicativity represent complementary principles that are in competition with each other in the operation of giving expression to conceptual possession.

Heine (1997), too, is interested in grammaticalization channels, but he concentrates on the lexical sources of (mostly nonattributive) possession and on the ultimate destiny of possessive items in other domains of grammar. As for nonattributive possession ('predicative possession' in a more restricted sense than in Seiler's framework), Heine (1997: p. 47) presents the following typology based on various conceptual 'schemata': 'X takes Y' ('action schema'), 'Y is located at X' ('location'), 'X is with Y' ('companion'), 'X's Y exists' ('genitive'), 'Y exists for/to X' ('goal'), 'Y exists from X' ('source'), 'as for X, Y exists' ('topic'), 'Y is X's (property)' ('equation'). As an example for the action schema Heine mentions Portuguese, while Turkish has a location schema. Somali has a companion schema, while the genitive schema is found again in Turkish. The goal schema is manifested in Tamil and Hungarian, the source schema manifests itself in Ik (Nilo-Saharan), Cahuilla has a topic schema, and Russian has an equation schema. In many of these languages different schemata coexist. Again, there is no clear separation between possession and very many other kinds of conceptual domains.

For the different schematic types mentioned it is possible to identify very common paths of development and transitions to other categories. Especially postural verbs (e.g., 'stand,' etc.) often develop into locative copulas, and the resulting location schema ('X is located at Y') may develop into a companion schema ('X is with Y'); both the location schema and the companion schema then tend to be reemployed as expressions of possession (for companion cases ('be with') cf Icelandic *ég er með bíl* (see above), and for location cf Finnish *Liisalla on mies* (lit. 'at Liisa is (a) husband,' 'Liisa has a husband')). The same construction types also often grammaticalize to expressions of aspect (Swahili has a 'be with'-progressive, and Icelandic has a 'be at'-progressive). Aspectual categories also commonly develop from possessively interpreted action schemata like 'to take X' (>'to have X'>'to have gone'), and aspectual categories may ultimately be interpreted as tense categories ('past tense').

Stassen (2005), too, classifies possessive categories into different types. His five categories for predicative possession are the following: locational possessive ('P'ed is at P'or'), genitive possessive ('P'ed is of P'or'), topic possessive ('as for P'or, there is P'ed'), conjunctional possessive ('P'or is with P'ed'), and the 'have'-possessive ('P'or has P'ed' and 'P'ed belongs to P'or,' respectively). As the accompanying map (which is accessible online) illustrates, the allegedly canonical 'have'-construction only accounts for about a fifth of the predicative structures actually employed in the languages of the world. There are also regional clusters: conjunctional possessives are particularly common in the southern part of Africa, while topic possessives are frequent around Indonesia (where 'have'-verbs are practically absent).

In his latest book on predicative possession, Stassen (2009) claims to have found strong structural ties between the kind of possessive structure used in this context and semantically fairly different expressions such as *John sang and Mary danced*. For instance, the predicative encoding of possession (for instance, 'be with') is paralleled by the encoding of temporal sequencing 'and then'; cf Stassen (2009: p. 90f). Granted that this is actually true, it still leaves the question why this should be the case. Why should possessive constructions be closely tied to constructions of a very different sort? Again, answers must often not be looked for in the realm of possession as such, but rather in the relevant discourse strategies (see ch. 1.3 below).

The shortcomings of traditional descriptions of possessive constructions are most noticeable in the area of inalienable possession. The *dativus sympaticus*-constructions are extremely common, and the cases used in possessive constructions for the P'or or noun phrase include not only genitives, datives, locative constructions, and the plain juxtaposition mentioned above, but occasionally even accusatives (e.g., Faroese kinship relations). As a consequence, it is necessary to give special attention to alternatives to traditional grammatical encodings of possession. The book edited by McGregor and Chappell (1996) presents a rich cross-linguistic survey of inalienable possession; common strategies associated with inalienables apart from 'he hit me on the head'-constructions are 'incorporating' strategies like 'he is bigheaded' (vs? 'he is bigcarred'). Quite typically, the P'or may not be mentioned at all on the inalienably possessed noun phrase (cf the phenomenon of 'possessor deletion' (see above)). But conversely, inherent relationality may sometimes have to be obligatorily marked (lit. 'his hands-his'). As a variant of possessor raising giving prominence to psychologically involved possessors, also different honorifics may be used in the context of body-part relations (cf Japanese, see Tsunoda, 1996).

Payne and Barshi's definition (1999: p. 3) of 'external possession' covers more than the constructions used in possessor raising: here external possession denotes any construction in which the P'or figures as a core argument of the verb and in a constituent separate from that which contains the P'ed. This definition includes applicatives (e.g., Mixean, see Zavala, 1999), 'double subject constructions,' and expressions where the P'or is the subject of a verb with an incorporated P'ed (Mohawk, see Baker, 1999). What is common to all these constructions is some difference in 'empathy,' 'discourse prominence,' and 'fore-grounding' in the relationship between P'or and P'ed as compared to the 'standard' possessive constructions. In fact, all kinds of possessive constructions can be ordered in terms of the comparative discourse prominence for P'or and P'ed. With this in mind, it is necessary to be aware of some specifically linguistic functions of possessive categories. Since surprisingly little work has been done on this aspect of linguistic possession, I will present some of the evidence for possession and discourse pragmatics in relative detail here.

The Linguistic Side of Possession

Different degrees of discourse prominence can be tested, among other things, with criteria like accessibility for relativization, etc. Languages differ considerably in the extent they 'normally' give prominence to possessors and possessed items, but there is usually variation between various options in the same language, too. Compare English, where the ordinary possessive

pronoun 'her' cannot be taken up in a relative clause, but the phrase 'of hers' can, as in 'her dog *who (=she)' versus 'this dog of hers, who (=she).' Another instance of this kind is Icelandic *bíllinn sinn, *sem* ('his car, who') as opposed to *bíllinn hans, sem* ('the car of his, who'): here, the 'eius'-possessive pronoun *hans* is a genitive form and can be taken up by a relative clause, such that the P'or gains relative discourse prominence, while the 'suus'-possessive pronoun cannot be taken up by a relative pronoun and has very little discourse prominence. The only purpose of *sinn* is to properly identify the P'ed – because *bíllinn* 'the car' would not be explicit enough. Conversely, constructions like Icelandic *Pabbi or faðurinn* ('daddy' or 'the father') may go without specification. They are inherently relational and are naturally associated with the speaker.

Yet more generally speaking, what is called 'possession' in linguistics is a special way of relating two different, but more or less salient referents in the discourse. This distinguishes possession clearly from 'identity' (same referent) and 'location' (space expressions are hardly salient or subject-worthy in the discourse). From these facts alone follow quite a lot of structures observed in the languages of the world.

First of all, it is natural in possessive expressions to find structures with two 'nominal' and thus typically referential items. This is true of genitival constructions like *the man's house*, but also of predicative constructions like Latin *mihi est liber*, lit. 'to me is (a) book,' 'I have a book,' as well as German *das Buch ist mir*, lit. 'the book is to me,' 'the book belongs to me.' It is also quite natural that the use of verbs of possession is not as frequent as the use of verbs in action scenarios – as the relationship is fairly stable in time. And speaking of discourse pragmatics, it is conceivable that mainly alienably possessed items are 'isolatable' enough to figure as subjects (cf *the book belongs to me* vs **the head belongs to me*).

If, however, for some reason, the expected 'norms' are violated, and an inalienable P'ed becomes a subject, this requires extra formal 'effort' – as, for instance, in the German phrase *mein Bauch gehört mir* 'my belly belongs to me,' which was used as a slogan of the feminist movement in the debate on abortion. Here, the possessive pronoun *mein* 'my' makes it clear that *Bauch* 'belly' is not a canonical, absolute subject. Other languages, for instance in Austronesia, employ specific possessive classifiers to indicate that the expression is not a canonical subject.

On the other hand, a possessive relationship is conceptually more controlled and closer than a locative relation, and therefore there is a tendency for original locatives to become grammaticalized as either genitives or datives in the domain of possession. And just as there is a change of perspective with respect to the use of 'P'or has P'ed'- versus 'P'ed belongs to P'or'- constructions, similar things can be said with respect to the use of prepositions. Compare, for instance, *das Buch von dem Mann* versus *der Mann mit dem Buch* (lit. 'the book from/of the man' vs 'the man with the book'). Interestingly, while we are talking about the same possessive relationship, the relation is viewed from different angles: in the ordinary, genitive-like construction with the preposition *von* 'from/of,' the P'ed is conceived as something literally 'distant from' or 'off' the main point of reference (genitives typically derive from 'off'-expressions). Conversely, the preposition *mit*, which goes back to the locative noun *Mitte*, 'middle,' marks the P'or as the actual 'middle' or 'center' of the relationship. In other words, the shift of perspective from P'or-NP to P'ed-NP is accompanied by a shift of attention from the (natural) center to the periphery or unnatural center. This interplay of center and periphery refers not to the possessive relation as such, but rather to the psychological way of observing the relation in question. In other words, psychological 'gestalt'-principles are an important structural element of language and possessive constructions, respectively (compare Broschart, 2007).

With this in mind, it comes as no surprise that attributive constructions like *das Buch von dem Mann* are far more common than *der Mann mit dem Buch* – simply because an NP like *der Mann* is usually readily identifiable as an 'isolate' figure in the discourse, while belongings are not. For this reason, *the book* usually needs an attribute, while *the man* does not.

But the story does not end here. As was shown above, there are a great many options of giving different degrees of discourse prominence to P'or and P'ed. The lowest discourse prominence for a speech act participant like 'I' or 'you' is a possessive pronoun or adjective, respectively (like Icelandic *minn* 'my' (see above)), while case forms of these pronouns (like genitives, *dativus sympaticus*, or even possessive accusatives) gain more and more discourse prominence. This is particularly true in events where the P'or is the actual target of the action (as in *he hit me on the head*).

As for common P'or-nouns, they are high in discourse prominence when they occur inside a referential (case or prepositional) phrase; prominence is lowest, however, in compounds such as German *Stuhlbein*, lit. 'chairleg,' 'the foot of a chair.' Yet even here, naturally salient, animate referents often appear with case-like items: *von Kind-es-beinen an*, lit. 'from a child's legs onward,' 'right from the age of a child.'

Thus, if there is anything that can order the incredible variety of possessive constructions cross-linguistically, it is the interplay of norm and exception with respect to subject-worthiness or relative discourse prominence, respectively. Any deviation from the norms (e.g., using a P'ed as a subject) requires additional effort (possessive pronouns or classifiers) such that the norms must not be forgotten (in Broschart (2007), I compare this to the 'cybernetic' mechanism of a thermostat in a heating system which controls the amount of energy needed, yet always relates back to normal). And it is also clear that if a relationship is self-understood, it need not be made explicit (*daddy* is sufficiently relational without the addition of *my*). Normally, there is some continuum of increasing possessor prominence from possessive pronouns via personal pronouns in the genitive to actual arguments of the verb (in the dative or accusative, etc., as in 'possessor raising' constructions). And while the P'ed is totally inaccessible for relativization in 'the blue-eyed man *which (=the eyes) …,' the P'ed can be relativized in 'the man with the blue eyes which (=the eyes) …,' etc.

As was noted earlier, in this context, possessive constructions do not simply express some extralinguistic concept, i.e., the constructions in question must also be judged with respect to their function in the discourse. This is particularly conspicuous in connection with genitival constructions. It is true that 'normally' the genitive phrase 'expresses' the possessor (cf 'the

man's car'), but sometimes the situation is reversed. Compare German *das Auto des Mannes ist alt* (lit. 'the car (P'ed:NOM) of the man (P'or:GEN) is old') versus *der Besitzer des Autos ist alt* (lit. 'the owner (P'or:NOM) of the car (P'ed:GEN) is old'). Clearly, it would be wrong to say that the genitival phrase always expresses the P'or – after all, the nucleus NP *der Besitzer* explicitly states that what follows is a possessed item. So the function of the genitive phrase cannot sufficiently be described by saying that it expresses a P'or. At best, one might say that the genitive phrase 'normally' expresses a P'or, and that in the deviant cases one has to create derived nouns like *Besitzer* to balance the effect of the deviation. However, what is generally true of the function of the genitive phrase is that it is used for the identification of the referent in the discourse.

'Normally,' it will be human P'ors who are 'better known' for identification than inanimate referents or dependent referents like body parts, etc., but in examples like German *der Besitzer des Autos mit dem Kennzeichen XY wird gebeten, sich mit der Polizei in Verbindung zu setzen* ('the owner of the car with the registration number XY is requested to contact the police'), it is the inanimate referent that happens to be better identifiable. In this case, the P'or will have to be expressed by a special relational noun (*Besitzer* 'owner'), but the genitival phrase is still basically used for identification in the discourse, even though it now encodes the P'ed.

Since discourse relevance is a direct correlate of high animacy and conceptual independence, this explains why the linguistic treatment of possessive relations will tend to be associated with distinctions like [+/-animate] and [+/-alienable]. As animacy also correlates with control just like any relationship of possession, it is not surprising that there are conceptual shifts from possession to distinctions of agency (cf 'I have beaten' (agentive) vs 'I am beaten' (nonagentive)) and vice versa. Agent (AG) and patient (PAT) are also differentiated in nominalized possessor-of-the-act constructions (cf 'Peter's (AG) beating (action) of the dog (PAT)' or Tongan *ko 'eku* (alienable) *taa'I 'o* (inalienable) *e kulí* (lit. 'it is my (active) beating of the (passive) dog,' 'I am currently beating the dog'); cf Broschart (1994)). Apart from that, agent and patient relations have also frequently been observed to coincide with differences of imperfective versus perfective aspect. And note that even the shift from an indefinite P'ed in a 'have'-construction to a definite P'ed typically leads to new aspectual constructions. Compare, for instance, German *ich habe ein Haus* 'I have/own a house' as opposed to **ich habe das Haus* (lit. 'I have the house'): the latter is considered strange and would be replaced by the more natural *das Haus gehört mir*, 'the house belongs to me.' But there are related 'have'-constructions with a definite P'or-object that tend to become reanalyzed as temporal or aspectual, as in *ich habe das Haus gekauft* ('I have bought the house' or 'I have the house as sth. bought'). Hence, from the point of view mentioned above, the typical correspondences between possession, voice, and aspect are not surprising. Like the voice distinction, possession implies some shift of perspective from an 'ordinary' discourse referent (which is animate and independent) to some referent which is usually inanimate and dependent. And the 'have'-constructions with definite objects are borderline cases of possession that tend to be reanalyzed as aspectual or temporal constructions.

It is these shifts of perspective and the many alternative forms of encoding that make possession a highly interesting domain of study for linguistics and cognitive science. But again, it must be emphasized that it is not possession as such that makes it a highly salient concept in language. For instance, cultures that originally had very little sense of personal belongings may still have hundreds of possessive pronouns – as is the case in Polynesian Tongan. And while languages differentiate very frequently between alienable and inalienable possession, I know of none that grammatically differentiates between cheap and expensive belongings.

Rather, it just so happens that many characteristics of possessive semantics correlate with important ingredients of the speech act, such as control or agency, familiarity in the discourse, etc. It is these characteristics that get grammaticalized, not some extralinguistic notion of belonging. However, knowing what belongs to whom sets the stage for many different kinds of interaction – including trade and warfare. For minimizing potentially bad effects of interaction, communication is essential – like employing a common tool which we all share – 'our' language.

See also: Discourse and Indentity; Inferences in Discourse, Psychology of; Linguistic Turn and Discourse Analysis in History; Linguistic Typology; Pragmatics, Linguistic.

Bibliography

Abney, S., 1987. The English noun phrase in its sentential aspect. MIT Press, Cambridge, Massachussets.

Baker, M., 1999. External possession in Mohawk: body parts, incorporation, and argument structure. In: Payne, D., Barshi, I. (Eds.), External Possession. Benjamins, Amsterdam, pp. 293–324.

Broschart, J., 1994. Präpositionen im Tonganischen. Zur Varianz und Invarianz des Adpositionsbegriffs. Brockmeyer, Bochum.

Broschart, J., 2007. Über die Sprachwissenschaft hinaus. Sprache und Linguistik aus transdisziplinärer Sicht. LIT-Verlag, Berlin.

Grkovic-Major, J., 2010. The development of predicative possession in Slavic languages. In: Motoki, N. (Ed.), The Grammar of Possessivity in South Slavic Languages. Synchronic and Diachronic Perspectives. Slavic Research Center, Hokkaido, Chap. 2, pp. 35–54. Available online at: http://src-h.slav.hokudai.ac.jp/coe21/publish/no24_ses/21_Grkovic.pdf (accessed 20.02.13.).

Gunkel, L., Zifonun, G. (Eds.), 2011. Deutsch im Sprachvergleich. Institut für Deutsche Sprache, Mannheim.

Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), 2005. The World Atlas of Language Structures (WALS). Oxford University Press, Oxford. Available online at: http://wals.info/ (accessed 20.02.13.).

Heine, B., 1997. Possession. Cognitive Sources, Forces, and Grammaticalization. Cambridge University Press, Cambridge, UK.

Koptjevskaja-Tamm, M., 2002. In: Adnominal Possession in the European Languages: Form and Function, vol. 2. STUF, 55, pp. 141–172.

Koptjevskaja-Tamm, M., 2003. Possessive NPs in the languages of Europe. In: Plank, F. (Ed.), Noun Phrase Structure in the Languages of Europe. Mouton de Gruyter, Berlin, pp. 621–722.

Löbel, E., 1995. Zur kategorialen Bestimmung der Possessiva in der NP/DP: Possessiva als Adjektive. In: Tappe, T., Löbel, E. (Eds.), Struktur der Nominalphrase. Bergische Universität & Gesamthochschule, Wuppertal, pp. 54–86.

McAnallen, J., 2011. The History of Predicative Possession in Slavic: Internal Development vs. Language Contact. ProQuest, Ann Arbor.

McGregor, W., Chappell, H. (Eds.), 1996. The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation. Mouton de Gruyter, Berlin.

Motoki, N. (Ed.), 2010. The Grammar of Possessivity in South Slavic Languages. Synchronic and Diachronic Perspectives. Slavic Research Center, Hokkaido. Available online at: http://src-h.slav.hokudai.ac.jp/coe21/publish/no24_ses/contents.html (accessed 20.02.13.).

Payne, D., Barshi, I. (Eds.), 1999. External Possession. Benjamins, Amsterdam.

Seiler, H., 1983. Possession as an operational dimension of language. Narr, Tübingen.

Stassen, L., 2005. Chapter 117: predicative possession. In: Haspelmath, et al. (Eds.), The World Atlas of Language Structures (WALS). Oxford University Press, Oxford. Available online at: http://wals.info/chapter/117 (accessed 20.02.13.).

Stassen, L., 2009. Predicative Possession. Oxford University Press, Oxford.

Tsunoda, T., 1996. The possession cline in Japanese and other languages. In: McGregor, W., Chappell, H. (Eds.), The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation. Mouton de Gruyter, Berlin, pp. 565–630.

Zavala, R., 1999. External possessor in Oluta Popoluca (Mixean): applicatives and incorporation of relational terms. In: Payne, D., Barshi, I. (Eds.), External Possession. Benjamins, Amsterdam, pp. 339–372.

Post Traumatic Amnesia

Thomas M McMillan, University of Glasgow, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For some seven decades, posttraumatic amnesia (PTA) has been viewed by most to be a key indicator of the severity of head injury and a predictor of long-term outcome. Discussed here is a historical introduction, definition of PTA, description of assessment tools, behavior and cognitive function and recovery during PTA, and of relationships between PTA and outcome and the evidence for behavioral interventions that attempt to reduce PTA duration.

Historical Introduction

Early work by Symonds (1928) associated the duration of loss of consciousness and stupor with severity of cerebral injury, and noted that it was common for there to be a period of amnesia after regaining consciousness:

> On regaining his senses he is found to have an amnesia for the period of clouded consciousness. Such patients are often said to have been unconscious during this whole period, but it is important to distinguish between complete unconsciousness or coma and the condition of stupor or clouded consciousness which is characteristic of major contusion.

The concept of posttraumatic amnesia (PTA) was, however, more formally developed by W. Ritchie Russell. While an Assistant Physician in the Royal Edinburgh Hospital, he presented a paper on 'Amnesia Following Head Injuries' to the International Neurological Congress in London in 1935:

> After head injuries accompanied by immediate loss of consciousness, certain disturbances of memory are well known to occur. On full recovery of consciousness, the patient is found to have little or no recollection of events during the period of confusion following the injury.

Russell later moved to Oxford where he developed his interest in head injury through working with casualties in World War II and developed the concept of PTA with which clinicians are so familiar today. In further defining PTA, he cites case examples where individuals could recollect events leading up to the injury early in their confusional state but later have no memory of these events. He later qualified this (Symonds and Russell, 1943) and noted that this endpoint can be followed by a further period of amnesia for hours or days. Russell makes clear that individuals may be able to answer simple questions and yet be disorientated for time and place (Russell, 1935). The emphasis from that time has been on a return of continuous memory for day-to-day events as the endpoint of PTA.

Russell and Nathan (1946) argue that the duration of PTA can be determined long after the injury occurred and that it provides an index of the severity of closed traumatic brain injury (TBI). In support of this, they report data on return to duty in 274 members of the armed services; extrapolating from their table (my analyses) the proportion who returned to duty from hospital (Chi squared = 46.03; $p < .0001$) or to full duties at follow-up (Chi squared = 52.32; $p < .0001$) was greater in those with PTA < 1 day than in those with PTA 1–7 days.

In these early works Russell describes 'islands' of memory during PTA. He describes 'delayed PTA' whereby there is a lucid interval followed by later confusion and amnesia, which he notes to occur more often in the case of gunshot wounds (in about 2.5% of $n = 100$) or relatively little disturbance of consciousness at the time of injury. He links 'islands' of memory to the occurrence of special events such as a visit from a relative, and indicates that they are more likely when confusion appears slight. He notes that behavior during PTA can 'closely approximate to normal.'

Russell also discusses recall of the period in PTA during barbiturate hypnosis, where curiously, one patient remembered aspects of his journey to hospital (some of which were independently verified) that he had not otherwise remembered before; several others remembered no more information under Pentothal. Russell concludes that under Pentothal or occasionally during a lucid period in PTA some can remember aspects around the time of the injury event that they have not remembered before. Also of interest are early attempts by Russell to link preinjury factors such as educational attainment, mental health problems, poor work record and family history (e.g., alcoholism) to outcome, implying a model in which severity of head injury is only one factor and indeed such a model would find current empirical support.

Definition of Posttraumatic Amnesia

PTA as defined by Russell is the duration of disturbance of consciousness following a head injury. This includes coma and a reduction in alertness that is reflected in amnesia for contemporaneous events. Hence, loss of consciousness, disorientation, confusion, and amnesia are included within a definition of PTA. The endpoint of PTA occurs when continuous memory for daily events returns. This avoids erroneously equating PTA with the first memory recalled after a head injury, which may be an 'island' of memory in a sea of amnesia. A formal definition of PTA that equates with the perspective of Russell (1971) is:

> the interval of time between injury and the return of continuous memory for everyday events.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51053-8

Some have argued that it is not appropriate to include coma in the duration of PTA because the relative proportion of time in coma can vary between individuals whose postcoma duration of PTA is the same (Levin et al., 1982). This issue may be of little interest where loss of consciousness is very brief, but where coma is prolonged and the duration of confusion and disorientation is relatively modest, the severity of head injury may be underestimated. A counterargument may be that the total duration from injury to return from orientation and amnesia gives a better estimate of severity of injury and of outcome and this view currently prevails. Hence, most assessment tools define PTA from the onset of loss of consciousness. The endpoint of PTA can be difficult to define where head injury (HI) is extremely severe and anterograde memory impairment continues to be severe. Cases of this kind tend to be relatively few in number and additional assessment of behavior and of daily function may be required.

It is of note that PTA is unlikely to be a valid measure of the severity of penetrating head injury. Here, there can be little or no disturbance of consciousness despite focal damage that results in significant limitations to daily function. In his series of World War II survivors of gunshot head injuries, Russell (1971) observes that 25% had no PTA and in 20% PTA was for less than an hour. Salazar et al. (1986) report that 51% of 342 military casualties with penetrating brain injuries had momentary or no loss of consciousness.

Assessment

Levin et al. (1979) were the first to publish a tool for the prospective assessment of PTA, the Galveston Orientation and Amnesia Test. The original version of the Galveston Orientation and Amnesia Test comprises 10 items; 8 focus on orientation and 2 on anterograde and retrograde memory. The test can be given repeatedly and 'PTA' is defined as a score below 75/100. The Galveston Orientation and Amnesia Test has strong associations with other measures of severity of injury (Glasgow Coma Scale) and disability outcome (Glasgow Outcome Scale). Although simple to administer, the Galveston Orientation and Amnesia Test has been criticized for not including a formal test of anterograde memory, as this can result in an underestimate of PTA (Tate et al., 2006). Artiola i Fortuny et al. (1980) added a test of recall and recognition in the Oxford Test and asked further about first memories after the injury. The end of PTA is indicated on the first day of a perfect score, if there is then a perfect score on two subsequent and consecutive days. The Westmead Post Traumatic Amnesia Scale (Shores et al., 1986) was based on the Oxford Test; it standardized procedures and provided a scoring form. Like the Oxford Test, a perfect score on three consecutive days is required to confirm that PTA has ended. A shortened version, the Revised-Westmead Post Traumatic Amnesia Scale that can be repeated at hourly intervals has been developed for use with less severe head injuries in emergency departments (Ponsford et al., 2004). The Julia Farr Centre PTA Scale (Forrester et al., 1994) also requires a perfect score on three consecutive days for PTA to be considered to be at an end; recent memory is only tested if orientation has returned, and this may not be valid in those with mild TBI (see Recovery and Outcome). One study compared PTA duration in relation to function and behavior after very severe HI on the Westmead, the Modified-Oxford PTA Scale and the Galveston Orientation and Amnesia Test (Tate et al., 2006). The Westmead and the Modified-Oxford PTA Scale differed in terms of the content of recall and recognition tests and the Galveston Orientation and Amnesia Test has no new memory component. Longer PTA durations were found for Westmead and the Modified-Oxford PTA Scale than for the Galveston Orientation and Amnesia Test with overall differences between Westmead/Modified-Oxford PTA Scale being nonsignificant. In a subgroup, PTA duration defined by the Westmead was longer than on the Modified-Oxford PTA Scale, reflecting greater difficulty in the new memory component on the Westmead (Tate et al., 2006). The Orientation-Log is a brief assessment designed for people in rehabilitation and offers the potential for cueing and a range of scores per item (Novack et al., 2000). There are pediatric versions of the Galveston Orientation and Amnesia Test and the Westmead (Ewing-Cobbs et al., 1990; Marossezky et al., 1993). The Westmead may not be valid for children aged below 7 years. The Orientation Group Monitoring System (Corrigan et al., 1988) uses serial ratings by trained ward staff that in addition to orientation and memory includes compliance with therapy and the ability to plan daily activities.

Clearly assessing PTA prospectively is impossible for clinicians who see people later after brain injury in the community. PTA is often not assessed prospectively in hospital, because people with TBI are often admitted to nonspecialist general medical or orthopedic wards. Hence there is a need to consider retrospective assessment of PTA. Indeed, Russell (1971) recommended that the endpoint of PTA is assessed by interview retrospectively. However, there have been few studies comparing retrospective and prospective PTA. One of these compared the assessment of PTA prospectively using the Galveston Orientation and Amnesia Test with a retrospective assessment 3–6 years later using a list of identifiable postinjury events. At reassessment, the assessor was blind to the original Galveston Orientation and Amnesia Test scores (see McMillan et al., 1996). The correlation between prospective and retrospective measures was very high (0.87) and both correlated significantly with several indicators of outcome in daily life. Others have used a protocol for assessment that can be used retrospectively (King et al., 1997). Although prospective measures might be more exact, the need for precision needs to reflect the clinical use of PTA. Clinically and sometimes in research, a range of PTA durations is used to categorize TBI severity, for example as mild, moderate, or severe TBI (see Recovery Severity and Outcome below), in which case there is no need for a precise estimate.

The duration of PTA can be longer in people with severe TBI, if intoxicated at the time of injury (Kaplan and Corrigan, 1992). It is not clear whether intoxication at the time of injury results in a more severe injury; for example, by potentiating effects of the TBI, causing a slower recovery, or a combination of these. The chronicity and severity of alcohol abuse are also likely to be important in relation to outcome. Other factors that may cause or prolong 'PTA-like' memory difficulty during hospital admission include surgery, opioid medication, and anxiety (Kemp et al., 2010).

Severity of TBI, Recovery, and Outcome

PTA is often used to categorize the severity of TBI. The mild, moderate, and severe categories defined by Russell and Smith (1961) were later expanded by Jennet and Teasdale (1981) into very mild (PTA < 5 min), mild (5–60 min), moderate (1–24 h), severe (1–7 days), very severe (1–4 weeks), and extremely severe (>4 weeks). Recovery from PTA often tends to occur in an order that reflects earlier return of longer term autobiographical material; that is with date of birth recovering first and recall of new information recovering last (Leach et al., 2006). Hence, reorientation to family and friends should occur before that for recognition or recall of the names of ward staff. Orientation most often returns in the order person, place, and time, consistent with that found following electroconvulsive therapy for depression (Daniel et al., 1987; Geffen et al., 1991) and irrespective of whether the TBI is mild, moderate, or severe (High et al., 1990). Some inconsistencies between studies in the order of recovery of components of orientation are likely to reflect differences in the items included in categories for 'place and time.' In cases with severe brain injury, recognition memory for new information can often occur prior to re-orientation for time (Tate et al., 2000). Where the brain injury is mild, and PTA is therefore short, reorientation may occur before recovery from amnesia.

Several studies report that PTA predicts late outcome in terms of disability, and that it is superior to other indices of severity of injury such as the Glasgow Coma Scale score or duration of loss of consciousness (Bishara et al., 1992; McMillan et al., 1996).

Cognitive Impairment and Behavior during PTA

One early study reports impairment of general intellect during PTA after severe TBI. Decrements in IQ were 25–40% on average compared to the same participants after recovery from PTA and compared to a separate group of severe TBI's who were not in PTA (Mandleberg, 1975). Impairment in general intellect during PTA was more pronounced for the Wechsler Adult Intelligence Scale Performance IQ (nonverbal) than for verbal IQ. People in PTA tend to show a wider range and more severe cognitive deficits than chronic amnesics or people with severe TBI who are no longer in PTA (Wilson et al., 1992). Most research on cognition and PTA has, however, focused on attention and memory. Severe impairment of declarative memory at recall and recognition and accelerated forgetting is found during PTA (Levin et al., 1988). For some time it was believed that permanent learning and memory largely did not occur during PTA. Later, it became evident that there is a degree of dissociation between motor and verbal learning. For example, Ewert et al. (1989) reported an ability to improve on tests requiring procedural learning (mirror reading, pursuit rotor, and maze tasks) during PTA at a time when deficits in verbal memory were evident. This may reflect neuropathology; the medial temporal areas that serve declarative memory are commonly compromised after severe TBI, whereas the basal ganglia that subserve procedural learning are rarely damaged nor is the occipital cortex that is involved in perceptual priming. A clinical example of this might be the patient with PTA on a ward who learns to find their way to the toilet, but cannot tell you where the toilet is. Deficits in procedural learning tend to be more pronounced in PTA than in chronic amnesics (Wilson et al., 1992). Chronic amnesics are also capable of implicit learning of verbal material (e.g., cannot recall a list of words but can complete the words if given the first syllable such as given 'MEA' for the target, 'meadow') and there is some evidence for implicit learning of verbal material during PTA from a small N study (Glisky and Delaney, 1996). Orientation tends to recover first, including attention required for simple tasks; then recognition, then cued recall and then free recall, with tests that require more effort taking longer to recover (Geffen et al., 1991; Stuss et al., 1999); these findings exemplify the need to incorporate a test of free recall in acute assessment of PTA duration.

Confusion, disorientation, and anterograde amnesia characterize PTA. There is often a lack of insight and sometimes to the extent that the patient may not believe they have had any serious injury. The behaviors that can be associated with this include agitation, irritability, aggression and sometimes violence (e.g., if goal-directed behavior is thwarted such as attempting to leave the ward or when staff is assisting with personal care), and wandering. When seen in the context of poor insight, impaired judgment and amnesia, the individual can simply believe that they are being unreasonably prevented from going about their daily business or even that they are being assaulted or have been imprisoned. In others, with less severe brain injuries behavior can seem normal during PTA. This is exemplified in sports players who sustain a head injury and continue to play, but later that day or the next they may have no memory of the game (Yarnell and Lynch, 1973). If not properly assessed, this can result in a premature return to play with enhanced risk of further injury or poorly planned hospital discharge.

Rehabilitation

Although there is evidence to suggest a better outcome following rehabilitation early after TBI, few studies have looked at treatments specifically for PTA.

Reality orientation is most commonly associated with older adults, where meta-analysis suggests that it can improve orientation in people with dementia, although studies are few in number and small in scale (Woods et al., 2012). Reality orientation was originally developed for war casualties and has been investigated as an intervention to reduce PTA. The Orientation Group Monitoring System incorporates reality orientation in a daily 30 min session in a group format with aims of assessing PTA duration, effects of other interventions, or complications in recovery and potentially as an intervention in itself (Mysiw et al., 1996). As yet, there is no evidence that the Orientation Group Monitoring System is superior to other methods of assessing PTA or that it has therapeutic benefit. The North Star Project in Montreal also considered integration of a reality orientation program in early rehabilitation during PTA. Although a trend toward shorter PTA is reported, this was not statistically different from controls (De Guise et al., 2005). Alderson and Novack (2002) found no difference in recovery of orientation in groups given reorientation sessions daily or

three times a week using the 'Orientation-Log.' Larson et al. (2011) used a three-dimensional virtual reality intervention in an attempt to improve attention early after severe TBI; 60% of their sample of 15 were in PTA. They report improvement in attention in PTA and non-PTA sub-groups; a significant limitation is the absence of a control given that natural recovery is likely to be rapid during this early phase.

There is no strong evidence to support reorientation as an effective technique for reducing PTA duration. More generally, there is as yet no evidence that rehabilitation for confusion and disorientation after TBI is effective and often multidisciplinary neurorehabilitation is recommended to begin after PTA has resolved; hence assessing the endpoint of PTA is clinically important for this reason.

See also: Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory; Brain Damage: Neuropsychological Rehabilitation; Confabulation; Implicit Memory; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory; Reconstructive Memory, Psychology of; Short-Term Memory: Psychological and Neural Aspects; Working Memory, Psychology of.

Bibliography

Alderson, A.L., Novack, T., 2002. Measuring acute recovery of orientation during acute rehabilitation for traumatic brain injury: value and expectations of recovery. The Journal of Head Trauma Rehabilitation 17, 210–219.

Artiola i Fortuny, L., Briggs, M., Newcombe, F., et al., 1980. Measuring the duration of post traumatic amnesia. Journal of Neurology Neurosurgery and Psychiatry 43, 377–379.

Bishara, S.N., Partridge, F.M., Godfrey, H.P.D., Knight, R.G., 1992. Post traumatic amnesia and the Glasgow Coma Scale related to outcome in survivors in a consecutive series of patients with severe closed head injury. Brain Injury 6, 373–380.

Corrigan, J., Arnett, J.A., Houck, L.J., et al., 1988. Reality orientation for brain injured patients, group treatments and monitoring of recovery. Archives Physical Medicine and Rehabilitation 66, 626–630.

Daniel, W.F., Crovitz, H.F., Weiner, R.D., 1987. Neuropsychological aspects of disorientation. Cortex 23, 169–187.

De Guise, E., Leblance, J., Feyz, M., Thomas, H., Gosselin, N., 2005. Effect of an integrated reality orientation programme in acute care on post traumatic amnesia in patients with traumatic brain injury. Brain Injury 19, 263–269.

Ewert, J., Levin, H.S., Watson, M.G., Kalisky, Z., 1989. Procedural memory during post traumatic amnesia in survivors of severe closed head injury. Archives of Neurology 46, 911–916.

Ewing-Cobbs, Levin H.S., Fletcher, J.M., Miner, M.E., Eisenberg, H.M., 1990. The childrens orientation and amnesia test. Relationship to severity of acute head injury and to recovery of memory. Neurosurgery 27, 683–691.

Forrester, G., Encel, J., Geffen, G., 1994. Measuring post-traumatic amnesia (PTA): an historical review. Brain Injury 8, 175–184.

Geffen, G., Encel, J.S., Forrester, G.M., 1991. Stages in recovery during post traumatic amnesia and subsequent everyday memory deficits. Neuroreport 2, 105–108.

Glisky, E.L., Delaney, S.M., 1996. Implicit memory and new semantic learning in posttraumatic amnesia. The Journal of Head Trauma Rehabilitation 11, 31–42.

High, W.M., Levin, H.S., Gary, H.E., 1990. Recovery of orientation following closed head injury. Journal of Clinical and Experimental Neuropsychology 12, 703–714.

Jennet, B., Teasdale, G., 1981. Management of head injuries. Brain Injury 6, 337–349. F.A. Davis, Philadelphia.

Kaplan, C.P., Corrigan, J.D., 1992. Effect of Blood Alcohol Level on Recovery from Severe Closed Head Injury. Brain Injury 6 (4), 337–349.

King, N.S., Crawford, S., Wenden, F.J., Moss, N.E.G., Wade, D.T., Caldwell, F.E., 1997. Measurement of post traumatic amnesia: how reliable is it? The Journal of Neurology, Neurosurgery, and Psychiatry 62, 38–42.

Kemp, S., Agostinis, A., House, A., Coughlan, A.K., 2010. Analgesia and other causes that mimic post-traumatic amnesia (PYA): a cohort study. Journal of Neuropsychology 4, 231–236.

Larson, E.B., Ramaiya, M., Zollman, F., Pacini, S., Hsu, N., Patton, J.L., Dvorkin, A.Y., 2011. Tolerance of a virtual reality intervention for attention remediation in persons with severe TBI. Brain Injury 25, 274–281.

Leach, K., Kinsella, G., Jackson, M., Matyas, T., 2006. Recovery of components of memory in post traumatic amnesia. Brain Injury 20, 1241–1249.

Levin, H.S., O'Donnell, V.M., Grossman, R.G., 1979. The Galveston orientation and amnesia test: a practical scale to assess cognition after head injury. Journal of Nervous and Mental Disease 167, 675–684.

Levin, H.S., Benton, A.L., Grossman, R.G., 1982. Neurobehavioural Consequences of Closed Head Injury. Oxford University Press, New York.

Levin, H.S., High, W.M., Eisenberg, H.M., 1988. Learning and memory during post-traumatic amnesia in head injured patients. The Journal of Neurology, Neurosurgery, and Psychiatry 51, 14–20.

Mandleberg, I.A., 1975. Cognitive recovery after severe head injury. The Journal of Neurology, Neurosurgery, and Psychiatry 38, 1127–1132.

Marossezky, N., Batchelor, J., Shores, E.A., Marossezky, J.E., Klein-Boonschate, M., Fahey, P.P., 1993. The performance of hospitalised head injury children on the WPTAS PTA scale. Clinical Neuropsychology 7, 83–95.

McMillan, T.M., Jongen, E.L.M.M., Greenwood, R.J., 1996. Assessment of post traumatic amnesia after closed head injury: retrospective or prospective? The Journal of Neurology, Neurosurgery, and Psychiatry 60, 422–427.

Mysiw, W.J., Bogner, J.A., Arnett, J.A., Clinchot, D.M., Corrigan, J.D., 1996. The orientation group monitoring system for measuring posttraumatic amnesia and assessing therapeutic interventions. The Journal of Head Trauma Rehabilitation 11, 1–8.

Novack, T.A., Dowler, R.A., Bush, B.A., et al., 2000. Validity of the orientation log relative to the GOAT. The Journal of Head Trauma Rehabilitation 15, 957–961.

Ponsford, J., Cameron, P.C., Willmot, C., et al., 2004. Use of the Westmead PTA scale to monitor recovery after mild head injury. Brain Injury 18, 603–614.

Russell, W.R., 1932. Cerebral involvement in head injury. Brain 69, 549–603.

Russell, W.R., 1935. Amnesia following head injuries. The Lancet, 762–763.

Russell, W.R., 1971. The Traumatic Amnesias. Oxford University Press, Oxford.

Russell, W.R., Nathan, P.W., 1946. Traumatic amnesia. Brain 69, 280–300.

Russell, W.R., Smith, A., 1961. Post traumatic amnesia in closed head injury. Archives of Neurology 5, 4–17.

Salazar, A.M., Grafman, J.H., Vance, S.C., Weingartner, H., Dillon, J., Ludlow, C., 1986. Consciousness and amnesia after penetrating head injury: neurology and anatomy. Neurology 36, 178–187.

Shores, E.A., Marosszeksy, J.E., Sandaman, J., et al., 1986. Preliminary validation of a clinical scale for measuring duration of post-traumatic amnesia. Medical Journal of Australia 144, 569–572.

Stuss, D.A., Binns, M.A., Carruth, F.G., Levine, B., et al., 1999. The acute period of recovery from traumatic brain injury: post traumatic amnesia or post traumatic confusional state. Journal of Neurosurgery 90, 635–643.

Symonds, C.P., 1928. Observations on the differential diagnosis and treatment of cerebral states consequent upon head injuries. British Medical Journal 2 (3540), 829–832.

Symonds, C.P., Russell, W.R., 1943. Accidental head injuries. The Lancet 2, 7–10.

Tate, R.L., Pfaff, A., Jurjevic, L., 2000. Resolution of disorientation and amnesia during post traumatic amnesia. Journal of Neurology Neurosurgery and Psychiatry 68, 178–185.

Tate, R.L., Pfaff, A., Baguley, I.J., Marosszeky, J.E., Gurka, J.A., Hodgkinson, A.E., King, C., Lane-Brown, A.T., Hanna, J., 2006. A multicentre randomised trial examining the effect of test procedures measuring emergence from post traumatic amnesia. The Journal of Neurology, Neurosurgery, and Psychiatry 77, 841–849.

Wilson, B.A., Baddeley, A., Shiel, A., Patton, G., 1992. How does post traumatic amnesia differ from the amnesic syndrome and from chronic memory impairment? Neuropsychological Rehabilitation 2, 231–243.

Woods, B., Aguirre, E., Spector, A.E., Orrell, M., 2012. Cognitive stimulation to improve cognitive functioning in people with dementia. Cochrane Collaboration. http://dx.doi.org/10.1002/14651858.CD005562.pub2.

Yarnell, P.R., Lynch, S., 1973. The 'ding': amnestic states in football trauma. Neurology 23, 196–197.

Postal Systems

Richard R John, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Postal systems are administratively coordinated communication networks that originated in antiquity, yet remain important today. During their 4000-year history, they have changed significantly, making it hard to generalize about their scale, scope, and accessibility. Until fairly recently, for example, postal systems often conveyed people in addition to information and goods. Most surveys of the world's postal systems are institutional genealogies. Postal systems beget one another in a grand procession of names, places, and dates. Such an approach has the merit of bringing together a wealth of anecdotal material, not easily found elsewhere. Yet, it exaggerates continuity, discourages systematic comparison, and obscures the role of the postal system as an agent of change. This article takes a comparative institutional approach. The world's postal systems are divided into three groups: imperial, corporate, and national. The section on National Postal Systems is subdivided into three parts: early national postal systems, postal reform, and postal systems in the twentieth and twenty-first century. The article concludes with a brief discussion of postal systems and social science, with some suggestions for future research.

Postal systems are administratively coordinated communications networks that originated in antiquity, yet remain important today. During their 4000-year history, they have changed significantly, making it hard to generalize about their scale, scope, and accessibility. Until fairly recently, for example, postal systems often conveyed people in addition to information and goods.

Overview

Most surveys of the world's postal systems are institutional genealogies. Postal systems beget one another in a grand procession of names, places, and dates. Such an approach has the merit of bringing together a wealth of anecdotal material, not easily found elsewhere (Scheele, 1970). Yet, it exaggerates continuity, discourages systematic comparison, and obscures the role of the postal system as an agent of change.

This article takes a comparative institutional approach. The world's postal systems are divided three groups: imperial, corporate, and national. The section on National Postal Systems is subdivided into three parts: origins, postal reform, and recent developments. The article concludes with a brief discussion of postal systems and social science, with suggestions for future research.

Imperial Postal Systems

Imperial postal systems enabled political leaders to conduct diplomacy, plan military campaigns, and oversee routine administration. The most highly developed were those of the Persians and the Romans. The Persian postal system under Emperor Darius (521–486 BCE) employed teams of mounted couriers to transmit clay tablets at 14-mile intervals along a 1600-mile route that stretched from the Tigris River to the Aegean Sea. The Greek historian, Herodotus, praised its speed and regularity: "Neither snow, nor rain, nor heat, nor gloom of night stays these couriers from the swift completion of their appointed rounds."

The Roman postal system, the *cursus publicus*, was established by Emperor Augustus (29 BCE–14 CE) and lasted 500 years. Relays of couriers with horse-drawn carts transmitted official dispatches along with government officials and their baggage. Routes extended throughout the empire and, during military campaigns, to generals in the field. Messages were written on papyrus, parchment, or wax-coated tablets – paper remaining unknown in Europe until after 1000 CE. The relays explain the etymology of 'postal'; it was derived from *positus*, the past participle of *ponere*, 'to place.' Couriers traveled 50 miles a day – a speed rarely exceeded until the nineteenth century.

Access to the *cursus publicus* was largely restricted to government officials. For this reason, some scholars term it a government communication network, and reserve the term postal system for networks open to the general population. Merchants relied on private messengers, seafarers, or friends. Maritime links were particularly important for Christian apostles such as Paul, who used them to send epistles throughout the Mediterranean world.

Corporate Postal Systems

Long after the fall of Rome, imperial postal systems continued to flourish in Asia, South America, and the Middle East. In Europe, however, no comparable institution existed for centuries. When postal systems revived in the 1200s, their organizers were no longer imperial regimes, but corporate bodies – including the University of Paris, the monks of Cluny, and the German butchers' guild.

Corporate postal systems rarely maintained relays of mounted couriers and often lacked the trappings of power. Yet, in several respects, they marked a distinct advance. Many crossed political boundaries, establishing links *between* states as well as *within* them. Few were restricted, like the Roman *cursus publicus*, to government officials. Open access was facilitated by the recent introduction of paper, a light-weight writing medium that was easy to transport. The policy also helped defray costs, since travel was expensive and many corporate postal systems were self-supporting. Henceforth, the primary

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95026-8

constraint for potential postal patrons was financial: postal rates were typically quite high, limiting access to the well-connected and the well-to-do.

Among the most important corporate postal systems were those established by the Italian city-states. By the end of the thirteenth century, couriers linked Venice with Constantinople, and Florence, Genoa, and Siena with northern France. The most enduring was the *Corrieri Bergamaschie* that Amadeo Tasso established in Bergamo, near Milan, in 1290. This institution much impressed the Habsburgs, who, in the 1400s, enlisted the Tasso and Delle Torre families to extend it throughout the Holy Roman Empire.

To oblige the Habsburgs, the families Germanicized their names to Thurn and Taxis. Shortly thereafter, in 1489, Johann von Taxis became the Habsburgs first postmaster. Within the next few years, he and his brother Franz established postal links between the Habsburgs' far-flung domains and Rome, France, and Spain, and opened the service to the general population (Dallmeier, 1977). Couriers carried a distinctive posthorn, an instrument that remains to this day a symbol of the German Deutsche Post. Though the rise of the nation-state greatly restricted its scope, the Thurn and Taxis postal system survived until 1867, when it was bought out by the Prussian government (Behringer, 1990). It has been justly credited with hastening a 'communications revolution' that established the technical preconditions for the emergence of the newspaper in the opening decades of the seventeenth century (Behringer, 2006).

National Postal Systems

Origins

Corporate postal systems stymied the rulers of the emergent nation-states in several ways. Most obviously, they created an unregulated communication channel that threatened political stability. In addition, they deprived rulers of a useful tool with which to monitor the flow of information and ideas. In response, nation-states throughout Europe – led by France and England – established postal systems under tight government control (Coase, 1955).

The French and English postal systems originated, like the imperial postal systems of antiquity, as government communication networks. In France, the founder was Louis XI (1461–83); in England, Henry VIII (1509–47). In both countries, the early 1600s marked a decisive turning point. Within a few decades, both governments granted access to the general population, instituted a monopoly, established a rate structure and delivery schedule, and organized a surveillance mechanism to monitor their subjects' correspondence: the *cabinet noir* in France; the secrets office in England (Robinson, 1948; Vaille, 1946–55). Shortly thereafter, the English government mandated a major rate increase, transforming its postal system into a revenue source, which it would remain until the twentieth century (Daunton, 1985).

Beginning in the 1600s, national postal systems did much to hasten the emergence of the newspaper – at the time, a new form of political communication. This relationship was particularly close in England, British North America, the Netherlands, and Prussia. By permitting the regular transmission of information on current events, postal systems became, in effect, an early broadcast media. Having gained control of the means of communication, governments used it to foster a distinctive worldview. In so doing, they helped forge the national identities that are such a fundamental feature of the modern world (Mueller, 1986).

Nowhere was the relationship of the postal system and the newspaper closer than in the United States. In the 1790s, following the adoption of the federal Constitution, the American government invested its postal system with an educational rationale far more expansive than the fiscal rationale that had previously prevailed. The new mandate, codified in the Post Office Act of 1792, hastened the circulation of newspapers by conveying them at extremely low rates, while enormously expanding the number of localities to which they could be sent (John, 1995). A landmark in postal history, this policy may well have marked for the first time a government that deliberately subsidized the transmission of a particular kind of information to facilitate the circulation of information on public affairs.

By 1828, the American postal system, with almost 8000 post offices, was larger and more geographically extended than any postal system in Europe. At this time, the United States boasted 74 post offices per 100 000 inhabitants; Britain, 17; and France, 4 (John, 1995). Prior to the coming of the railroad and telegraph, the American government had established the preconditions for the disembodied public sphere that French traveler Alexis de Tocqueville would explore so perceptively in *Democracy in America* (1835, 1840).

Postal Reform

Victorian commentators often hailed the postal system of their days as a fundamentally *new* institution that was little older than the railroad and telegraph. For the majority of the population, this was often true. Prior to the 1840s, postal rates remained so high that letter-writing was, in practice, confined to the few. Even in the United States, if a farmer or artisan received anything in the mail, it was more likely to be a newspaper than a letter (Fuller, 1974).

In both Britain and the United States, the high cost of letter postage hastened the establishment of clandestine letter-delivery services in direct violation of the postal monopoly. Following the advent of the railroad in the 1830s, these operations greatly expanded, hastening the major rate reduction known as postal reform.

Postal reform began in Britain, where it was championed by educator Rowland Hill. Hill recognized that high letter rates encouraged illicit mail carriage, and posited that a major reduction would hasten a massive expansion in letter-writing. To buttress his argument, he demonstrated in a famous pamphlet – *Post Office Reform: Its Importance and Practicality* (1837) – that rates bore no relation to expenses. All but a tiny fraction of the cost of delivering a letter, Hill calculated, derived from its *routing* through the system rather than from its *conveyance* from sender to recipient.

Hill prevailed and, in 1840, the British government simplified the rate schedule and reduced the basic letter rate to a penny. Hill styled his reform uniform postage; it was commonly called penny postage. Hill's reform greatly increased mail volume, just as he had predicted. In addition, it hastened the advent of several familiar postal practices – including

mandatory prepayment, mail boxes, and the postage stamp (Daunton, 1985).

Many governments followed Hill's lead. The United States lowered its letter rates in 1847 and 1851; France in 1848. Japan, eager to emulate the best practices in the West, established in 1870 a postal system modeled closely on Britain's (Westney, 1987). Increased mail volume, in conjunction with the advent of the ocean-going steamship, hastened the proliferation of transnational postal treaties. These diplomatic initiatives were regularized in 1874 with the establishment of the Universal Postal Union (UPU), today a branch of the United Nations. The organization of the UPU owed much to Heinrich von Stephan, a gifted postal administrator who had helped to establish a unified German postal system following the creation of the German Empire in 1871.

Closely related to postal reform was the substitution of the railroad for the stagecoach as the preferred mode of conveyance. By the 1830s, stagecoach service had attained a high level of development, especially in Britain and France. Interestingly, Rowland Hill had based his famous cost calculations on the stagecoach, rather than the railroad, which had yet to be widely adopted. Stagecoach speeds often averaged 10 miles an hour, a standard rarely matched since the demise of the *cursus publicus* of imperial Rome. The use of the railroad to convey mail posed various scheduling dilemmas and could increase postal costs; in the United States, for example, it quickly ran the postal system into debt. Yet it proved indispensable, greatly increasing carrying capacity and revolutionizing postal routing. In Britain, the United States, and most industrialized nations, it was long customary to sort the mail onboard moving trains, a practice known as railway mail.

Postal reform ushered in a golden age of letter-writing (Henkin, 2006). In many countries, the exchange of holiday greeting cards became a popular custom. Soldiers' letters emerged, especially in wartime, as a notable mode of literary expression. Though letter-writing would, in the twentieth century, face competition from telephony, the habit persists, and, since the popularization of e-mail, may well be enjoying a revival.

In the decades preceding the First World War, each of the world's leading postal systems – Britain, France, Germany, and the United States – undertook major new responsibilities. Each established a parcel post and instituted a rudimentary banking service known as postal savings. And each, with the exception of the United States, took over the telegraph and telephone, establishing a ministry of Posts, Telegraphs, and Telephones, known in France and elsewhere as the PTT (Bertho, 1981; Perry, 1992; John, 2010). This expansive era in postal history reached its apogee around 1900, when it was common for journalists to hail the postal system as the greatest business in the world (John, 2012). The advent of commercial aviation in many countries, including the United States, owed much to the generous mail contracts that postal administrators awarded airline firms, a relationship that underscored the longstanding role of the postal system in hastening transportation innovations.

Recent Developments

In both the developed and the developing world, postal systems today remain indispensable elements of the information infrastructure (Sinclair, 1984). In the past few decades, however, much has changed. Beginning around 1970, rapid changes in communications technology led many governments to reevaluate the union of the postal system, the telegraph, the telephone. Deregulation followed. In Britain, the breakup came in 1981; in France, in 1991.

The Internet poses a further challenge, yet, in a sense, it is nothing new. For a century and a half, postal systems have coexisted with earlier modes of electric communications. The United States today is the originator not only of the Internet, but also of 40% of the entire world's mail. A century ago, pundits routinely predicted that telegraphy would soon render mail delivery obsolete. Today telegraphy has been everywhere supplanted, while postal systems persist.

Additional challenges stem from the emergence of new competitors – DHL in Europe; Federal Express and UPS in the United States. Express firms rely on improvements in air travel to compete head-on with the national postal systems in key high-value markets. In response, many governments have reexamined – and, in several cases, narrowed – the scope of the postal monopoly (Campbell, 1994).

Postal administrators have met these challenges in various ways. Many have transformed national postal systems into public corporations, a transformation that is sometimes called privatization. If this trend continues, postal systems in the next century may share less with the national postal systems out of which they evolved than the corporate postal systems of medieval Europe. The French postal system – known today as La Poste – has been a publicly traded corporation for over 30 years (Le Roux, 2002). The British postal system, Royal Mail, has aggressively sought new markets (Campbell-Smith, 2012). So, too, has the Deutsche Post, which issued its first securities in 2000, and the U.S. Postal Service, which has not been a government department since 1970 and no longer receives a government subsidy.

Postal Systems and Social Science

Scholars agree that the printing press was an agent of change. Only recently, however, have postal systems been studied in a similar way. The last edition of the *International Encyclopedia of the Social Sciences* (1968) contained no entry on the postal system, and no references to it in the index.

The rise of the new institutionalism in history, sociology, and political science bodes well for postal studies. So too does the continuing vogue for French literary theory, with its focus on the complexity of the communication process (Siehart, 1993). Recent developments in information technology promise further insights, offering new perspectives on modes of communication that are seemingly being displaced. For postal scholars, Hegel's dictum holds true: the owl of Minerva flies at dusk.

Research opportunities abound. Comparative data remains scarce, especially for the period prior to 1840. Without it, it is hard to frame credible generalizations about what postal systems did and how they worked. Topics that invite exploration include the development and significance of scheduling; censorship and surveillance; postal privacy; communication networks and the rise of nationalism; challenges to the postal monopoly; changing rationales for postal systems; and the shifting meaning of the equal access ideal.

See also: Communication and Transportation, History of; Mass Communication: Normative Frameworks; Telegraph; Telephone.

Bibliography

Behringer, W., 1990. Thurn und Taxis: Die Geschichte ihrer Post und ihrer Unternehmen. Piper, Munich.

Behringer, W., 2006. Communications revolutions: a historiographical concept. German History 24.

Bertho, C., 1981. Telegraphes et Telephones de Valmy au Microprocesseur. Librarie Generale Francaise, Paris.

Campbell, R.M., 1994. The Politics of the Post: Canada's Postal System from Public Service to Privatization. Broadview Press, Peterborough, ON.

Campbell-Smith, D., 2012. Masters of the Post: The Authorized History of the Royal Mail. Penguin, London.

Coase, R.H., 1955. The postal monopoly in Great Britain: an historical survey. In: Eastham (Ed.), Economic Essays in Commemoration of the Dundee School of Economics, 1931–55. William Culross & Son, Coupar Angus, pp. 25–37.

Dallmeier, M., 1977. Quellen zur Geschichte des europäischen Postwesens, 1501–1806, 3 parts. Lassleben, Kallmünz.

Daunton, M., 1985. Royal Mail: The Post Office since 1840. Athlone, London.

Fuller, W.E., 1974. The American Mail: Enlarger of the Common Life. University of Chicago Press, Chicago.

Henkin, D.M., 2006. The Postal Age: The Emergence of Modern Communications in Nineteenth-Century America. University of Chicago Press, Chicago.

John, R.R., 1995. Spreading the News: The American Postal System from Franklin to Morse. Harvard University Press, Cambridge, MA.

John, R.R., 2010. Network Nation: Inventing American Telecommunications. Harvard University Press, Cambridge, MA.

John, R.R. (Ed.), 2012. The American Postal Network, 1792-1914. Pickering & Chatto, London.

Le Roux, M. (Ed.), 2002. Histoire de la Poste de l'Administration à l'Énterprise. Editions Rue d'Ulm, Paris.

Mueller, M., 1986. The Currency of the Word: War, Revolution, and the Temporal Coordination of Literate Media in England, 1608–1855 (MA thesis). University of Pennsylvania, Philadelphia, PA.

Perry, C.R., 1992. The Victorian Post Office: The Growth of a Bureaucracy. Royal Historical Society, London.

Robinson, H., 1948. The British Post Office: A History. Princeton University Press, Princeton.

Scheele, C.H., 1970. A Short History of the Mail Service. Smithsonian Institution Press, Washington.

Siehart, B., 1993. Relais: Geschicke der Literarur als Epoche der Post, 1751–1913. Verlag Brinkmann & Bose, Berlin [1999. Literature as an Epoch of the Postal System. Stanford University Press, Stanford].

Sinclair, J., 1984. Uniting a Nation: The Postal and Telecommunication Services of Papua New Guinea. Oxford University Press, Melbourne.

Vaille, E., 1946–55. Histoire Generale des Postes Francaises, 7 vols. Presses Universitaires de France, Paris.

Westney, D.E., 1987. Imitation and Innovation: The Transfer of Western Organizational Patterns to Meiji Japan. Harvard University Press, Cambridge, MA.

Postcolonial Geography

Jane M Jacobs, Yale-NUS College, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the emergence of postcolonial geography in the 1980s and the intellectual contributions it has made since, including reference to key thinkers. It discusses the relationship with postcolonial studies more generally and outlines the distinct lines of inquiry taken in postcolonial geography, namely: historical geography revisions of colonialism; contemporary studies of the geographies of postcolonial formations and expressions; and critiques of Western-centric geographical theory building and empirical scholarship. It also discusses the methodological legacies of postcolonial geography approaches.

The contemporary geographical scholarship that describes itself as 'postcolonial' has much in common with other parts of the social sciences and humanities that use this term to describe their approach and matters of concern. That said, there are certain historical connections between geography and the making of empires, which add an imperative to the way geographers engage with postcolonial thought and phenomena. Pragmatic geographies of surveying and mapping were the frontier technologies of imperial exploration and the operations that brought distant territories under the control of metropolitan centers as colonies. As historian Edward Said noted in *Culture and Imperialism* (1993: p. 271), imperialism is "an act of geographical violence through which virtually every space in the world is explored, charted, and finally brought under control." The imperial expansions of eighteenth and nineteenth century Europe appropriated the resources of other peoples and lands. That material interest was motorized by a complex value system, which made audacious and often-violent appropriations make perfect sense to the colonizers. Ideas about the racial hierarchy of humankind placed Caucasians over and above all others and enabled European colonizers to assume a racial superiority and civilizational worth that legitimated their imperialism. The modern discipline of geography lives uneasily with this past and in the 1990s that ambivalence offered a moral compass for the emergence of postcolonial geography.

Original Scope of Postcolonial Geography

Those geographies described as 'postcolonial' use the term in a range of ways, as is the case in other parts of the social sciences and humanities. First, it is used to describe the formal political circumstance of a nation or political entity becoming independent from colonial rule or an imperial force. Second, the term is used to describe a set of social, cultural, and political formations that come into being because of, and usually in opposition to, colonization. Such formations may temporally precede a formal state of postcolonial independence, but they may also emerge well after that independence is realized due to the lingering after effects of colonialism. Third, the term is often used to describe research that is informed by postcolonial theory and which applies its analytical tools to, say, race relations, processes of development, subaltern or nativist intellectual and political formations, or creative expressions. Postcolonial geographies may be either historical or contemporary in focus. For example, a persistent interest of more historically focused postcolonial geographers has been revisiting the site of colonialism and offering revisionist accounts. Another body of work examines contemporary conditions of decolonization, even amid the insistent presence of imperial logics. Yet more work looks at the emergence of hybridized systems that are both dependent upon lingering colonial logics but operate subversively. Many have questioned the 'post' of postcolonialism. Postcolonial formations remain complexly entangled in their colonial pasts, just as they are constantly negotiating various kinds of neocolonialism, be it in the form of geopolitical influence, cultural effect, or economic leverage. The persistence of various kinds of colonialism has led scholars to note that the term postcolonialism is prematurely celebratory.

While having much in common with other variants of postcolonial studies, the distinguishing feature of postcolonial geography is that the research emphasis is on the spatial logics and expressions of colonial and postcolonial phenomena (Sharp, 2009). Historically inflected scholarship within postcolonial geography offers a critical reappraisal of the spatialized forms of knowledge, political systems, and social identities authorized under colonialism. In 1993, Livingstone (1993) published an account of the geographical tradition, within which he paid particular attention to the relationship between European geography, race, and empire. Driver (1992, 2000) has similarly 'unmasked' geography's disciplinary relationship to European colonization. By demonstrating the way geographical knowledges framed other worlds within imperial projects, this article has empirically thickened the analytic framework set by Edward Said's landmark contribution *Orientalism* (1978). One particular geographical knowledge form has come under specific scrutiny: the map, and its associated practical sciences of survey and cartography. This imperial knowledge system made meaningful both peoples and lands that, from a Eurocentric worldview, were so radically different that ofttimes they could only be understood as 'savages' or 'wastelands.' Harley's (1988) reappraisal of cartographic practices and colonialism seeded an ongoing scholarship on cartography and power. More recent scholarship has

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72040-X

examined the relationship between new technologies of mapping such as geographic information system and new nonterritorial styles of imperialism. Others have turned their attention to how digital mapping technologies have been used in the counter cartographies of postcolonial nation building (Radcliffe, 2009) or struggles by indigenous groups to have their land rights and cultural sites recognized (Dunn, 2007).

The 1990s to the early 2000s offered over a decade of sustained research in the area of postcolonial geography, the full breadth of which is showcased in a number of edited collections (Godlewska and Smith, 1994; Bell et al., 1995; Blunt and McEwan, 2002). Some significant lines of inquiry are worth mentioning. Geographers became interested in the ways in which the largely racial logics of differentiation that underscored colonialism were intersected with other axes of difference, such as gender (Blunt and Rose, 1994; Phillips, 1997). Other geographers denaturalized the racial category of whiteness itself, looking at the ways in which it generated, and held in place, assumptions about authority and civility that shaped landscapes, past, and present (Bonnett, 2000; Shaw, 2007). Geographers also attended to the ways the built environment reflected colonial values, imperial nostalgias, and modernizing aspirations (Smith, 1992; Jacobs, 1996; Yeoh, 1997; Driver and Gilbert, 1999; Jazeel, 2013). Colonial cities were important sites in the transfer of modern, European values to new worlds and functioned as important centers of power for administering colonized people and resources. In these cities, town-planning ordinances about cleanliness and sanitation became the mechanism through which colonial assumptions about racial difference were translated into on-the-ground technologies of segregation and exclusion.

Postcolonial Geography Revisions

By 2003, the editors of a special issue of the *Singapore Journal of Tropical Geography* on postcolonial geographies could conclude that the subfield was 'rich and contentious.' One contention came from the emphasis on historical examples, as well as the reliance on interpretative approaches to representational and discursive data. This bias was unsurprising given that geographers took inspiration from a wider field of postcolonial studies, which was grounded in the fields of literature and history. In 2003, Jennifer Robinson called for geographers to expand the scope of critical scholarship bequeathed by postcolonial geography. She took inspiration from an argument about 'postcolonial modernity' proposed by the historian and social theorist Dipesh Chakrabarty in his *Provincializing Europe* (2001). Chakrabarty pointed out that much of what is said about the experience of postcolonial modernity is constructed in and through European social scientific concepts. As Chakrabarty puts it: "European thought is at once both indispensable and inadequate in helping us to think through the experiences of political modernity in non-Western nations." Given that many of the core analytical tools of the so-called universal social science were grounded in the specific experience of Europe, the next phase of the postcolonial social scientific project became one of 'provincializing Europe': exploring how its social scientific thought 'may be renewed from and for the margins' (Chakrabarty, 2000: p.16).

Robinson's plea set postcolonial geography on a new course, which might be better described as an 'after-the-West geography.' She called for a 'cosmopolitan' geographical scholarship that looked beyond the West, both in social theoretical terms as well as in terms of practical knowledges and actions of world building. As Robinson (2003: p. 273) puts it: "A geography whose intellectual vision is limited to the concerns and perspectives of the richest countries in the world has little hope of effectively participating in the debates that will matter in the twenty-first century." Robinson gave empirical flesh to this proposition by thinking about urbanization in non-Western contexts. She argued that cities in such contexts were often compared negatively to models of urbanization generated from the circumstances of Western cities. She called upon geographers to generate revised theories of urbanization that took as their starting point the processes exhibited in the many other 'ordinary cities' (Robinson, 2006). In a similar vein, other geographers began to apply this after-the-West version of postcolonial scholarship to existing theories of economic and political development. Using such a lens meant that non-Western experiences of development were not simply seen as slower or lesser or mimetic versions of the West, but offering alternative models of what modernity and development might be (McEwan, 2009; Pollard et al., 2011). These new directions in postcolonial geography had considerable intellectual overlap with what has come to be known as postdevelopment studies (Escobar, 1995; McFarlane, 2006; Sidaway, 2007).

Legacies and New Directions

Postcolonial geographical perspectives have resulted in important methodological legacies. First, there is a new self-reflexivity among geographers with respect to how they construct their geographies and the power differentials that exist when they work across cultural or development divides. Second, and relatedly, there is a new awareness of the need to reflect upon the origins of apparently neutral or universal explanatory frameworks drawn upon in geographical explanations. Third, there is a commitment to creating geographies that readjust colonial or neocolonial structures of authorization with respect to who speaks and with what authority. For example, geographers have used multiple authorship to this end (Huggins et al., 1995), as well as various models of participatory or action research (Pain and Kindon, 2007). Fourth, there has been renewed and revised application of comparative methods in the discipline (Robinson, 2011). This new comparativism is distinct from earlier methods, and instates a cosmopolitanism necessary to the construction of geographical knowledges in an era of transnational movements and the emergence of non-Western centers of power.

Has the time for postcolonial geographies passed? Many of the methodological and theoretical developments bequeathed by a postcolonial approach will leave a permanent legacy. Furthermore, there are many current events that ensure ongoing relevance for a postcolonial geography. The transnational logics of contemporary times leads all too easily to

conclusions that the world is succumbing to the homogenizing force of globalization. That suffocating fact can be challenged by adhering to a postcolonial perspective, which is attuned to alternative modernities and trajectories of development. Furthermore, global connectivity in communication and travel has led to new opportunities for transnational solidarities among groups seeking postcolonial recognition of their rights, generating a still emergent era of 'global indigeneity' (Sidaway et al., 2014). And there are many colonial presents. Some of them are starkly territorial, as Gregory (2004) shows with respect to the contemporary geopolitics of the Middle East, which still draws sustenance from persistent orientalisms. Others are variations of seemingly familiar developmental expansions, such as the effects of Rising Asia, including China's presence in Africa or the South China Sea (Mohan and Power, 2009; Agnew, 2012; Anand, 2012; Raghuram et al., 2014). Yet others operate in and through the logic of culture, such as the imperialism of the English language, which creates profound challenges for creating fully cosmopolitan geographies inclusive of other knowledges and voices (Kitchen, 2005). Finally, the planetary ill effects of humankind on the environment, what has been dubbed the age of the Anthropocene, offer new imperatives for postcolonial geographies (Jackson, 2014; Sidaway et al., 2014). This developing 'planetary' circumstance is calling various social institutions (be they nations, organizations, or ethnically defined groups) into new relationships of responsibility and accountability with respect to the environment, resources, and productivity.

See also: Colonialism, Anthropology of; Geography: Overview; Globalization and World Culture; Imperialism, History of; Racism, History of; Subaltern History.

Bibliography

Agnew, J., 2012. Looking back to looking forward: Chinese geopolitical narratives and China's past. Eurasian Geography and Economics 53, 301–314.
Anand, D., 2012. China and India: postcolonial informal empires in the emerging global order rethinking Marxism. A Journal of Economics, Culture and Society 24, 68–86.
Bell, M., Butlin, R., Heffernan, M. (Eds.), 1995. Geography and Imperialism, 1820–1940. Manchester University Press, Manchester.
Blunt, A., Rose, G. (Eds.), 1994. Writing Women and Space: Colonial and Postcolonial Geographies. Guilford Press, London and New York.
Blunt, A., McEwan, C., 2002. Postcolonial Geographies. Continuum, London and New York.
Bonnett, A., 2000. White Identities: Historical and International Perspectives. Prentice Hall, Harlow, UK.
Chakrabarty, D., 2001. Provincializing Europe. Princeton University Press, Princeton, NJ.
Driver, F. (Ed.), 1992. Geography's empire: histories of geographical knowledge. Environment and Planning D: Society and Space 10, 23–40.
Driver, F., 2000. Geography Militant: Cultures of Exploration and Empire. Blackwell Publishers, Oxford.
Driver, F., Gilbert, D., 1999. Imperial Cities: Landscape, Display and Identity. Manchester University Press, Manchester, UK and New York.
Dunn, C.E., 2007. Participatory GIS — a people's GIS? Progress in Human Geography 31 (5), 616–637.
Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton.
Godlewska, A., Smith, N. (Eds.), 1994. Geography and Empire. Blackwell, Oxford, UK.
Gregory, D., 2004. The Colonial Present. Blackwell Publishing, Oxford.
Harley, J.B., 1988. Maps, knowledge and power. In: Cosgrove, D., Daniels, S. (Eds.), The Iconography of Landscape. Cambridge University Press, Cambridge, UK, pp. 277–312.
Huggins, J., Huggins, R., Jacobs, J.M., 1995. Kooramindanjie: place and the postcolonial. History Workshop Journal 39, 165–181.
Jackson, M., 2014. Composing postcolonial geographies: postconstructivism, ecology and the limits of critique. Singapore Journal of Tropical Geography 35, 73–88.
Jacobs, J.M., 1996. Edge of Empire: Postcolonialism and the City. Routledge, London and New York.
Jazeel, T., 2013. Sacred Modernity: Nature, Environment and the Postcolonial Geographies of Sri Lankan Nationhood. Liverpool University Press, Liverpool.
Kitchen, R., 2005. Disrupting and destabilizing Anglo-American and English-language hegemony in geography. Social and Cultural Geography 6, 1–15.
Livingstone, D.N., 1993. The Geographical Tradition. Blackwell, Oxford, UK.
Marcus, G.E. (Ed.), 1992. Environment and Planning D; Society and Space, vol. 10.
McEwan, C., 2009. Postcolonialism and Development. Routledge, London and New York.
McFarlane, C., 2006. Transnational development networks: bringing development and postcolonial approaches into dialogue. The Geographical Journal 172, 35–49.
Mohan, G., Power, M., 2009. Africa, China and the 'new' economic geography of development. Singapore Journal of Tropical Geography 30, 24–28.
Pain, R., Kindon, S., 2007. Participatory geographies. Environment and planning A 39, 2807–2812.
Phillips, R., 1997. Mapping Men and Empire: A Geography of Adventure. Routledge, London and New York.
Pollard, J., McEwan, C., Hughes, A. (Eds.), 2011. Postcolonial Economies. Zed Books, London.
Radcliffe, S.A., 2009. National maps, digitalization and neoliberal cartographies: transforming nation-state practices and symbols in postcolonial Ecuador. Transactions of the Institute of British Geographers 34, 426–444.
Raghuram, P., Noxolo, P., Madge, C., 2014. Rising Asia and postcolonial geography. Singapore Journal of Tropical Geography 35, 119–135.
Robinson, J., 2003. Postcolonising geography: tactics and pitfalls. Singapore Journal of Tropical Geography 24, 273–289.
Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routledge, London and New York.
Robinson, J., 2011. Cities in a world of cities: the comparative gesture. International Journal of Urban and Regional Research 35, 1–23.
Sharp, J.P., 2009. Geographies of Postcolonialism: Spaces of Power and Representation. SAGE, London.
Shaw, W.S., 2007. Cities of Whiteness. Wiley-Blackwell, London.
Sidaway, J.D., 2007. Spaces of postdevelopment. Progress. Human Geography 31, 345–361.
Sidaway, J.D., Woon, C.Y., Jacobs, J.M., 2014. Planetary postcolonialism. Singapore Journal of Tropical Geography 35, 4–22.
Smith, D.M. (Ed.), 1992. The Apartheid City and beyond. Routledge, New York and London.
Yeoh, B.S.A., 1997. Contesting Space: Power Relations and the Urban Built Environment in Colonial Singapore. Oxford University Press, Oxford, UK.

Postcolonial Law

Eve Darian-Smith, University of California Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Postcolonialism and the problems associated with postcolonial law are today more widely recognized than ever before among scholars, political philosophers, and legal practitioners. However, the terms of the conversation have shifted since postcolonial theory first emerged as a new arena of critical scholarship. Today, postcolonial theory addresses the oppression of all communities historically treated as racially and ethnically inferior to Europeans, whether or not these communities self-identify as 'indigenous' or think of themselves as formerly colonized. This article explores the challenges presented by postcolonial law and related analyses of legal orientalism in the twenty-first century. Acknowledging the postcolonial dimensions of legal exchange between the global North and global South demands a rethinking of the prevailing Eurocentric, state-bound understanding of what constitutes law in the current global era.

Postcolonial Law in the Twenty-First Century

Postcolonialism and the problems associated with postcolonial law are today more widely recognized than ever before among scholars, political philosophers, and legal practitioners. However, the terms of the conversation have shifted to more aptly apply to current global geopolitical realities. Whereas the language of earlier postcolonial theorists focused on the dialectic between colonizing nations and the colonized, contemporary scholars talk in terms of the relations between what is commonly referred to as the global North and global South. This shift in terminology expands the lens of analysis from state-centered law in the context of specific national colonial enterprises to a more global post-westphalian worldview that takes into account transnational, regional, and state interrelations (Falk, 1998). It also opens up the conversation to include the oppression of all communities historically treated as racially and ethnically inferior to Europeans, whether or not these communities self-identify as 'indigenous' or think of themselves as formerly colonized.

Contemporary scholars of postcolonial law draw upon an intellectual legacy that traces back to postcolonial studies that emerged in the 1980s to talk about postcolonial law in a more encompassing way. These scholars are interested in the underlying orientalist assumptions in national and international law that affirm essentialized constructions of cultural difference (Ruskola, 2002; Falk, 2009; Pahuja, 2011), and inform many countries' flawed polices of multiculturalism (Bhandar, 2009; Marés, 2007). Importantly, what links the legacy of postcolonial studies to contemporary analyses of legal orientalism is a central focus on the endurance of historically structured racial and ethnic divides between Western and non-Western societies despite a growing appreciation of their respective interdependencies.

Postcolonial theory is largely associated with a rethinking of a dominant European historiography that places the West at the center of the world. Associated with South Asian scholarship, subaltern and literary studies as well as analyses of resistance, postcolonial theory emerged out of the global South in the 1980s and gained an increasing presence in Anglo-European universities (Prakash, 1990, 1992; O'Hanlon and Washbrook, 1992; Chakrabarty, 1992). With the proliferation of postcolonial research, there was, and remains, much debate over the meaning and scope of postcolonial terminology and its political agenda (see Williams and Chrisman, 1994; Schwarz and Ray, 2000; Loomba, 2005; Mongia, 2009). Despite these ongoing deliberations, it is helpful to turn to the most significant implications of a postcolonial perspective, which mark it as distinctly different from, yet complementary to, other critical investigations associated with critical race theory, critical legal theory, and poststructuralist and feminist theoretical perspectives. This process will help to flesh out the contours of postcolonial studies and, at the same time, underscore its intellectual lineage and relevance to contemporary legal analyses.

Below I discuss some of the insights postcolonial theory brings to bear on an understanding of legal orientalism in contemporary legal processes. This is most notably achieved through the theoretical insights of Edward Said, who was a leading figure in postcolonial criticism and whose book *Orientalism* (1978) established a long trajectory of critical thinking about the subjugation of non-Western peoples. Specifically, postcolonial theory reveals the violence and the technologies of power involved in understanding concepts such as modernity and capitalism as well as contemporary state and trans-state nationalisms (Baxi, 2000; Spivak, 1999; Darian-Smith and Fitzpatrick, 1999; Comaroff and Comaroff, 2006). Moreover, postcolonial theory provides the intellectual bridge linking historical colonial injustices to contemporary global racial injustices and asymmetries of power between a global North and global South. In short, postcolonial theory provides the intellectual platform from which to identify, analyze, and assess what is encompassed in the wider geopolitical framing of the term 'postcolonial law' in the twenty-first century.

Defining a 'Postcolonial' Perspective

Postcolonial theory acknowledges and recovers the ongoing significance of colonized peoples in shaping the epistemologies, philosophies, practices, and shifting identities of dominant and taken-for-granted Western subjects and subjectivities (Santos, 2007). Postcolonial scholars bring to the foreground the cultural and psychological relations between the former

colonized and colonizers, whom, they argue, cannot be understood except in conjunction with each other (Gandhi, 1998). Postcolonial theorists do not claim that colonialism was experienced in the same way under different regimes, just as they recognize that today colonialism operates in very different ways from its earlier configurations. Nonetheless, while paying attention to the details of specific contexts, postcolonial scholars agree that in order to understand all contemporary histories of peoples and places, irrespective of whether there historically existed in any given site an explicit colonial regime, it is important to remain aware of the enduring presence of discourses that posit 'civilized,' 'progressive,' and 'lawful' Europeans against 'barbaric,' 'static,' and 'lawless' native populations.

According to these scholars, colonial assumptions of Western superiority endure across time and undermine contemporary attempts to build more inclusive multicultural societies. This is because categories of racial difference were used to varying degrees by colonial governments to gain power and control over locally subjugated peoples (Stoler, 1989; Dirks, 1986). Today, despite claims of increasing acceptance of the concept of multiculturalism and cultural diversity, racial categories, and racialized differences continue to exist, though often in less overt manifestations (Goldberg, 2008; Lentin and Titley, 2011). Moreover, according to many postcolonial theorists, these boundaries of difference are insurmountable, because the psychological intersubjectivity between former colonizers and colonized people constantly invites re-representations of difference (see Fitzpatrick, 1999).

Drawing on a variety of theoretical perspectives – including an Hegelian master/slave dialectic, phenomenological essentialism, and psychoanalytical insights gleaned from Franz Fanon and Jacques Lacan – postcolonial theorists such as Homi Bhabha, Gayatri Spivak, and Benita Parry bring to the fore complex understandings of how oppressed peoples resist and seek empowerment. In recognizing processes of mutual desire and negation between the 'master' and 'slave,' postcolonial scholars have had to grapple with alternative historical narratives and identities other than those conventionally supplied by the West (Trivedi and Mukherjee, 1996). The irony is that the process of self-determination by peoples formerly colonized requires the adoption of European knowledge, including concepts such as 'progress,' 'development,' 'individualism,' and what it means to be human. It also means the use of Western forms of government, state-building, and perhaps most importantly of all, law. Thus, according to a postcolonial perspective, all assertions of freedom and self-awareness require elements of mimicry and voyeurism. As Douglas Robinson has noted:

> Postcolonial or subaltern scholars claim it is at once essential and impossible to forge a 'new' postcolonial identity: *essential*, because those colonial constructs were at once alien and negative, because they came from the outside and destroyed much of value in the indigenous cultures, and because an effective postcolonial politics requires the development of more positive indigenous visions: but also *impossible*, because colonial discourse continues to inform even these postcolonial attempts to break free of it, and tends to condition even the imagination of a 'new' (postcolonial) identity along 'old' (colonial) lines (Robinson, 1997: 19–20).

A prominent postcolonial theorist, Dipesh Chakrabarty, has responded to the ironies of alterity by calling for the 'provincializing,' or decentering, of Europe and European epistemological knowledge. Chakrabarty (and others) argue that non-Western knowledge has been historically ignored and precluded from historiographical accounts of humanist understanding and intellectual endeavor (Chakrabarty, 1992, 2000). In an effort to critique "the 'Europe' that modern imperialism and (third world) nationalism have, by their collaborative venture and violence, made universal," Chakrabarty urges scholars "to write into the history of modernity the ambivalences, contradictions, the use of force, and the tragedies and the ironies that attend it" (Chakrabarty, 1992: 20–21).

Postcolonial Law

Nowhere in the histories of modernity has the 'use of force, and the tragedies and the ironies that attend it' been as obvious as they are in the context of legal engagement. European law, in a variety of ways, was the formal mechanism and institutional frame through which colonial governments oppressed and controlled indigenous peoples. Of course, there are many varieties of imposition and reception of law, a few even categorized as voluntary. In some cases, such as in Africa (Mamdani, 1996), this involved co-opting native chiefs and traditional procedures of arbitration and dispute resolution. In other cases, such as Australia, the British declared the continent *terra nullius* or vacant, and on the basis that native populations were thought to be less than human, they were deemed to have no law at all (Behrendt, 2003). As a result, traditional local methods of peacekeeping were often overlooked or deliberately obliterated. Law – emblematic of European rationalism, individual property rights, and sovereign state authority – provided the justification for domination and exploitation based on racial, ethnic, or religious inferiority.

It is impossible to separate colonial from postcolonial law because each developed in collaboration and conjunction against and through each other (Merry, 2004; Darian-Smith, 1996). Hence it can be said that nineteenth century metropolises such as New York, Paris, and London were always connected with and influenced by their nation's colonial peripheries, be these geographically distanced outposts beyond state boundaries or socially and politically isolated enclaves (such as native reservations) within it. Similarly, colonial outposts were intimately connected back to their colonizing oppressors. In short, colonies emerged over time as intricately entangled hybrid societies incorporating European and non-European sensibilities. Hence the decolonized countries of the post-1945 era – such as India, Algeria, South Africa, and the Philippines – could not entirely excise their Euro-American colonial legacies. Moreover, these newly independent nations had little choice but to adopt European forms of state building and the institutions, bureaucracies, and constitutions of their former masters. Such formalities were necessary in order for new states to declare themselves 'liberal' and 'modern' and to participate in national and international political and economic organizations such as the United Nations. The rule of law exemplifies the postcolonial dilemma that required – and still requires – self-

determining nations to mimic and adopt Euro-American legal concepts and structures.

Globalization and Postcolonial Law

Prevailing discussions about law and globalization epitomize the ironies presented by postcolonial law. For instance, The United Nations, World Bank, International Monetary Fund, and other international legal arrangements all require that in order for a country to participate in the global political economy, it must demonstrate commitment to, and adherence with, the foundational values of Western law (Halliday and Osinky, 2006: 455–456). In a postcolonial world, just as in the colonial situation, there are always ongoing modifications and appropriations between all interacting communities. One only has to think of human rights to appreciate the extent to which its apparently universal application has to be constantly translated and modified to fit the social, political, and economic values of particular peoples living in localized places (Merry, 2006; Sarat, 2001). Similarly, Euro-American law is not entirely impervious to the influence of laws emanating from the global South, as illustrated by the adoption of the United Nations Declaration on the Rights of Indigenous Peoples in 2007. That being said, the global dominance of Euro-American law in the early decades of the twenty-first century is not easily destabilized. And as a result, Euro-American law structurally institutionalizes the enduring asymmetries of power between the global South and global North and pervasively fashions and legitimates certain legal practices, meanings, and imaginations that have their origins in the global North (Darian-Smith, 2000).

Within much scholarship there lingers a deeply embedded assumption of the global North's legal superiority vis-à-vis the rest of the world (Darian-Smith, 2013). This is evidenced by a majority of United States and European scholars continuing to treat the rule of law as a discrete entity and not as a dynamic product of historically contested and culturally informed colonial–postcolonial interactions. Thus well into the twenty-first century, in arguments both for and against the enduring significance of the nation-state in the increasing context of globalization, the dominance of Western legalism is taken as a given. As a result, analysts of law tend to look only at the privileged domains of legal interaction among lawyers, judges, businesspeople, and entrepreneurs, and to ignore the perspectives of ordinary people whose culturally informed normative understandings of law may be very different (Ewick and Silbey, 1998). In short, scholars should not ignore the 'contribution of the masses' (Rajagopal, 2003: 402). We need to be concerned with how globalization affects peoples in different ways, and studies should include both the small, cosmopolitan, plane-hopping legal elite *and* the millions of peoples from various classes, cultures, ethnicities, and religions whose understanding of law may appear to those embedded in a Western legal heritage as 'traditional' or 'backward.'

Legal Orientalism and Postcolonial Law

Postcolonial law requires that scholars and practitioners come to terms with the fact that there is no universal legal code and no such thing as pure legal objectivity, but rather a complex overlapping plurality of legal systems and legal meanings. If this view was to take hold, the naturalized centrality and superiority of a Euro-American legal perspective would be dislodged and, borrowing Chakrabarty's terminology discussed above, would become necessarily 'provincialized.' However, given the relations between law, capitalism, and a global political economy, it is perhaps not surprising that Western legal scholarship has largely ignored (some would argue deliberately) the challenging presence of postcolonial law. In an attempt to move beyond this deadlock, some scholars are coming at the problem of legal plurality by talking about the issue of legal orientalism. These scholars take a long historical view in arguing that legal orientalism has shaped the development of modern Euro-American law from the sixteenth century to the present (Ruskola, 2002; Anghie, 2006; Falk, 2009: 39–54). This argument forces us to think about how racial and cultural biases continue to inform globally dominant legal concepts and assumptions of Western legal superiority, and may in turn open up ways to challenge or resist these dominant legal understandings of the world (Santos and Rodriguez-Gavarito, 2005).

What is legal orientalism? As mentioned above, the concept of legal orientalism draws expressly upon the work of Edward Said, a leading figure in postcolonial theory. Said coined the word 'orientalism' to refer to the ways European societies throughout the nineteenth century constructed their identity and self-understanding through imagining their difference from the Arab and Muslim world (Said, 1978, 1993). Essential in this process was the West's stereotyping of the Orient, which included a range of Eastern cultures that included the Middle East as well as China, Japan, and South Asia. Orientalist discourses emanating from Europe were not exactly the same as those emanating from the United States because they were usually directed toward the Middle East and China, while in the United States orientalist rhetoric was usually directed to the Philippines and targets closer to home (Little, 2008; Brody, 2010; Francavigilia, 2011). These differences typically correlated with a country's imperial and colonial interests and often changed over time. However, what united these various forms of orientalist rhetoric and material practice was the assumed oppositional relations between an exoticized Orient and a civilized Occident.

Typically nineteenth century orientalist discourses about the East were negative and reinforced a presumed hierarchy of Western superiority and Eastern inferiority. However, this was not always the case, as seventeenth and eighteenth centuries Enlightenment philosophers and missionary Jesuits often praised Chinese people for their ingenuity and skill (see Gregory, 2003; Mungello, 2009). However, by the nineteenth century European attitudes about Asian peoples had crystallized into derogatory stereotypes (Ruskola, 2002). Europeans promoted themselves as modern, rational, moral, and lawful in contrast to a projection of Eastern societies as premodern, irrational, immoral, and lawless. Hence, at the same time that commentators such as Alexis de Toqueville were remarking upon the emphasis given to law in the United States in the 1830s, historians and social theorists were pointing to the lack of law in countries such as China, which was essentially viewed as a backward, 'stagnant' society in which lawlessness reigned

(Ruskola, 2002: 181–187, 213–215). But as insisted upon by Said, this did not mean that the "Orient was essentially an idea, or a creation with no corresponding reality" (Said, 1978: 5). Rather, "The Orient is an integral part of European *material* civilization and culture … with supporting institutions, vocabulary, scholarship, imagery, doctrines, even colonial bureaucracies and colonial styles" (Said, 1978: 2).

Legal orientalism served a variety of purposes. The most obvious of these was that it helped confirm on the world stage the marginality of the East and the centrality of the imperial West. European and American scholars argued that Eastern jurisprudential traditions were based on custom, ritual, and religion in contrast to the so-called rational and scientific legal systems of modern Western nations. Declaring non-Western legal systems inferior helped to justify European law and culture as a superior civilization, worthy of world leadership and dominance. Orientalist rhetoric also provided the rationale for Western nations to marginalize Asian peoples within their domestic jurisdictions. For instance, in the United States orientalist rhetoric provided the basis for the Chinese Exclusion Act (1882). This act suspended Chinese immigration into the country and prevented those Chinese people already living in the United States from ever being granted citizenship. Under the act, it was argued that Chinese people were nonlegal subjects because they were incapable of understanding American law and so deserved to be excluded from the new republic (Park, 2004; Ruskola, 2002: 215–217).

Postcolonial legal scholars argue that the oppositional rhetoric between Eastern and Western legal traditions was essential for the development of modern Euro-American law. In other words, European law emerged historically through a perceived difference from non-Western legal concepts. According to the sociolegal scholar Duncan Kennedy, international law must be understood in relation to "a distinction between the West and the rest of the world, and the role of that distinction in the generation of doctrines, institutions and state practices" (Kennedy, 1997: 748). This perceived difference helped shape the international legal system, which required the 'invention of legal primitivism' to legitimate the West's universal aspirations (Gathii, 1998; Bowden, 2005; Anghie, 2006; Wilf, 2009).

If one accepts this argument, then it follows that Western law has orientalist assumptions historically built into its language, structure, and procedures. This suggests that contemporary Euro-American law, and the international legal system on which it is built, remains to this day intrinsically and pervasively cultural and racially biased (see Pahuja, 2011; Westra, 2011). In short, legal orientalism endures in twenty-first century international law and global legal relations (see Otto, 1996; Falk, 2009; Haldar, 2007). Legal orientalism continues to fuel assumptions about the global North's legal superiority over the global South and has been deployed in a range of national and international legal forums such as asylum and refugee claims (Akram, 2000). Moreover, legal orientalism is evident in the ways the global North interprets law in the Middle East, particularly in the wake of the events of 9/11, and how Western nations view law in China, Africa, and Latin America. However, as the international legal scholar Teemu Ruskola remarks, the point of recognizing the presence of contemporary legal orientalism is not to overcome our cultural biases – an impossible task – but rather to ask why certain orientalist images of law developed, why they continue to resonate in the contemporary world, and what can be done to dilute these negative stereotypes that undermine international law and prevent sincere global dialogue and creative legal collaboration (Ruskola, 2002: 222).

One thing that is certain – whether one frames asymmetrical power relations in terms of postcolonial law or legal orientalism – is the need to move past a modernist hierarchy of legal authority based on simplistic binaries of rational versus nonrational and civilized versus uncivilized legal systems. De-orientalizing the twenty-first century's normative global order and stereotyped legal divides is seen, by some scholars and analysts at least, as ultimately necessary for the stability and peace of global, international, national, regional, and local relations (Santos, 2007; Onuma, 2010). As the Nigerian legal scholar Ikechi Mgbeoji has eloquently stated, "the North and South are mutually vulnerable, sharing a common destiny, which cannot be realized unless notions of a civilized self and barbaric other are abandoned" (Mgbeoji, 2008: 152).

Postcolonial insights offer some insights as to how to get past historically structured racialized divides between peoples and communities. The political theorist Duncan Ivison in his book *Postcolonial Liberalism* (2002) argues for the need to create a "genuine 'multilogue' not just between the state and indigenous peoples, but between them and other cultural and national groups as well" (Ivison, 2002: 163). In his arguing that indigenous peoples can make considerable contributions in the thinking of how to build more inclusive societies, Ivison notes that this will take time and a firm commitment to "the ideal of a political order in which different national groups, with different modes of belonging and different conceptions of the good and the right, nevertheless share a willingness to live under political arrangements that reflect this plurality" (Ivison, 2002: 166). Ivison's argument underscores the political challenge of a postcolonial perspective with respect to law. In order for there to be "a context-sensitive and embedded form of public dialogue and deliberation" (Ivison, 2002: 163), we must first rethink the prevailing Eurocentric, state-bound understanding of what constitutes law. In other words, embracing postcolonial law and coming to terms with its deeply problematic histories of colonial oppression is perhaps the first step in a process toward building a global legal future that is more inclusive, responsible, and equitable.

See also: Law and Development; Modernity: History of the Concept; Modernity; Nationalism, Historical Aspects of: Africa; Nationalism, Historical Aspects of: South Asia; Nationalism, Historical Aspects of: The West; Nationalism: General; Orientalism; Postcoloniality.

Bibliography

Akram, Susan M., 2000. Orientalism revisited in asylum and refugee claims. International Journal of Refugee Law 12 (1), 7–40.

Anghie, Anthony, 2006. The evolution of international law: colonial and postcolonial realities. Third World Quarterly 27 (5), 739–753.

Baxi, U., 2000. Postcolonial legality. In: Schwarz, H., Ray, S. (Eds.), A Companion to Postcolonial Studies. Blackwell, Oxford, pp. 540–555.

Behrendt, Larissa, 2003. Achieving Social Justice: Indigenous Rights and Australia's Future. The Federation Press, Sydney.
Bhandar, Brenna, 2009. The ties that bind: multiculturalism and secularism reconsidered. Journal of Law and Society 36 (3), 301–326.
Bowden, Brett, 2005. The colonial origins of international law: European expansion and the classical standard of civilization. Journal of the History of International Law 7, 1–23.
Brody, David, 2010. Visualizing American Empire: Orientalism and Imperialism in the Philippines. University of Chicago Press, Chicago, IL.
Chakrabarty, D., 1992. Postcoloniality and the artiface of history: who speaks for 'Indian' pasts? Representations 37, 1–26 (Winter).
Chakrabarty, D., 2000. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton, NJ.
Comaroff, Jean, Comaroff, John (Eds.), 2006. Law and Disorder in the Postcolony. University of Chicago Press, Chicago, IL.
Darian-Smith, E., 1996. Postcolonialism: a brief introduction. Social and Legal Studies 5 (3), 291–299.
Darian-Smith, E., 2013. Laws and Societies in Global Contexts: Contemporary Approaches. Cambridge University Press, Cambridge.
Darian-Smith, E., Fitzpatrick, P. (Eds.), 1999. Laws of the Postcolonial. University of Michigan Press, Ann Arbor, MI.
Darian-Smith, E., 2000. Structural inequalities in the global legal system. Law & Society Review 34 (3), 809–828.
Dirks, N.B., 1986. From little king to landlord: property law, and the gift under the Madras permanent settlement. Contemporary Studies in Society and History 28, 307–333.
Ewick, Patricia, Silbey, Susan, 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago, IL.
Falk, Richard, 1998. Law in an Emerging Global Village: A Post-westphalian Perspective. Transnational Publishers, Ardsley, New York.
Falk, Richard, 2009. Achieving Human Rights. Routledge, London and New York.
Fitzpatrick, P., 1999. Passions out of place: law, incommensurability, and resistance. In: Darian-Smith, E., Fitzpatrick, P. (Eds.), Laws of the Postcolonial. University of Michigan Press, Ann Arbor, MI, pp. 39–60.
Francavigilia, Richard, 2011. Go East, Young Man: Imagining the America West as the Orient. Utah State University Press, Logan, Utah.
Gandhi, L., 1998. Postcolonial Theory: A Critical Introduction. Colombia University Press, New York.
Gathii, James Thuo, 1998. International law and eurocentricity. European Journal of International Law 9, 184–211.
Goldberg, David Theo, 2008. The Threat of Race: Reflections on Racial Neoliberalism. Wiley-Blackwell, Oxford.
Gregory, John S., 2003. The West and China Since 1500. Palgrave Macmillan, Basingstoke, Hampshire, UK.
Haldar, Piyel, 2007. Law, Orientalism and Postcolonialism: The Jurisdiction of the Lotus Eaters. Routledge-Cavendish, Abingdon.
Halliday, Terence, Osinky, Pavel, 2006. Globalization of law. Annual Review of Sociology 32, 447–470.
Ivison, Duncan, 2002. Postcolonial Liberalism. Cambridge University Press, Cambridge.
Kennedy, Duncan, 1997. Review of the rights of conquest. American Journal of International Law 91, 745–748.
Lentin, Alana, Titley, Gavan, 2011. The Crises of Multiculturalism: Racism in a Neoliberal Age. Zed Books, London.
Little, Douglas, 2008. American Orientalism: The United States and the Middle East since 1945. University of North Carolina Press, Chapel Hill, NC.
Loomba, A., 2005. Colonialism/Postcolonialism, second ed. Routledge, London and New York.
Mamdani, M., 1996. Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism. Princeton University Press, Princeton, NJ.
Marés, Carlos Frederico, 2007. Multiculturalism and collective rights. In: Santos, Boaventura de Sousa (Ed.), Another Knowledge is Possible: Beyond Northern Epistemologies. Verso, London and New York, pp. 75–104.
Merry, Sally Engle, 2004. Colonial and postcolonial law. In: Sarat, Austin (Ed.), The Blackwell Companion to Law and Society. Blackwell Publishing, Oxford, UK, pp. 569–588.
Merry, Sally Engle, 2006. Human Rights and Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago, IL.
Mgbeoji, Ikechi, 2008. The civilized self and the barbaric other: imperial delusions of order and the challenges of human security. In: Falk, Richard, Rajagopal, Balakrishnan, Stevens, Jacqueline (Eds.), International Law and the Third World: Reshaping Justice. Routledge-Cavendish, London and New York, pp. 151–165.
Mongia, Padmini (Ed.), 2009. Contemporary Postcolonial Theory: A Reader. Bloomsbury, New York.
Mungello, D.E., 2009. The Great Encounter of China and the West, 1500–1800 (Critical Issues in World and International History), third ed. Rowman & Littlefield, New York.
O'Hanlon, R., Washbrook, D., January 1992. After orientalism: culture, criticism, and politics in the Third World. Comparative Studies in Society and History 34 (1), 141–167.
Onuma, Yasuaki, 2010. A Transcivilizational Perspective on International Law. (Hague Academy of International Law.) Martinus Nijhoff Publishers, Leiden, Boston, MA.
Otto, Dianne, 1996. Subalternity and international law: the problems of global community and the incommensurability of difference. In: Darian-Smith, Eve, Fitzpatrick, Peter (Eds.), Special Issue on Law and Postcolonialism. Social & Legal Studies 5 (3), 337–364.
Pahuja, Sundhya, 2011. Decolonizing International Law: Development, Economic Growth and the Politics of Universality. Cambridge University Press, Cambridge.
Park, John S.W., 2004. Elusive Citizenship: Immigration, Asian Americans, and the Paradox of Civil Rights. New York University Press, New York.
Prakash, G., 1990. Writing post-orientalist histories of the Third World: perspectives from Indian historiography. Comparative Studies in Society and History 32 (2), 383–408.
Prakash, G., January 1992. Can the 'subaltern' ride? A reply to O'Hanlon and Washbrook. Comparative Studies in Society and History 32 (1), 168–184.
Rajagopal, Balakrishnan, 2003. International Law from Below: Development, Social Movements and Third World Resistance. Cambridge University Press, Cambridge, UK.
Robinson, D., 1997. Translation and Empire: Postcolonial Theories Explained. St. Jerome Publishing, Manchester, UK.
Ruskola, Teemu, 2002. Legal orientalism. Michigan Law Review 101 (1), 179–234.
Said, Edward W., 1978. Orientalism. Vintage, New York.
Said, Edward W., 1993. Culture and Imperialism. Alfred A. Knopf, New York.
Santos, Boaventura de Sousa (Ed.), 2007. Another Knowledge is Possible: Beyond Northern Epistemologies. Verso, London and New York.
Santos, Boaventura de Sousa, Rodriguez-Gavarito, Cezar (Eds.), 2005. Law and Counter-hegemonic Globalization: Toward a Cosmopolitan Legality. Cambridge University Press, Cambridge, UK.
Sarat, Austin (Ed.), 2001. Special issue human rights: new possibilities/new problems. Studies in Law, Politics, and Society 56.
Schwarz, H., Ray, S. (Eds.), 2000. A Companion to Postcolonial Studies. Blackwell, Oxford, UK.
Spivak, G.C., 1999. A Critique of Postcolonial Reason. Harvard University Press, Cambridge, MA.
Stoler, A., 1989. Rethinking colonial categories: European communities and the boundaries of rule. Contemporary Studies in Society and History 31 (1), 134–161.
Trivedi, H., Mukherjee, M. (Eds.), 1996. Interrogating Post-colonialism: Theory, Text and Context. Indian Institute of Advanced Study, Rashrapati Nivas, Shimla, India.
Westra, Laura, 2011. Globalization, Violence and World Governance. Brill, Leiden and Boston.
Wilf, Steven, 2009. The invention of legal primitivism. Theoretical Inquiries in Law 10, 485–509.
Williams, P., Chrisman, L. (Eds.), 1994. Colonial Discourse and Post-colonial Theory: A Reader. Colombia University Press, New York.

Postcolonial Science Studies

Warwick Anderson, University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

During the past 20 years, postcolonial approaches have emerged to explain the distribution and local adaptation of science, technology, and medicine around the world. Such studies contest European and North American hegemony in science, situating and thereby dismantling global or universal claims. Critical postcolonial analysis should be distinguished from anodyne talk of 'global flows' and from actor-network theory, which avoids discussion of power differences and structural inequalities. This article explains the origins of postcolonial science studies, describes their various forms and functions, and suggests future elaboration of this sort of critique.

Postcolonial Science Studies

Since the late 1980s, some postcolonial attitudes, or at least postcolonial premises, have slowly worked their way into science and technology studies, though their presence often is hidden or even denied. Studies that contest European and North American hegemony in science, situating and thereby dismantling global or universal claims, represent what might be called the postcolonial turn. Recognition of creditable knowledge making beyond North Atlantic shores constitutes a postcolonial approach. So too does the emphasis on hybridity, heterogeneity, and indeterminacy in what once appeared to be sovereign, uncontaminated categories. A postcolonial orientation directs attention to the complexities of relations in any contact zone. It re-examines the terrain that empire has tilled across the world, showing that dominance is never absolute – that imperial or authoritative knowledge, despite colonial fantasy and amour propre, must always adapt to local conditions, mix with other traditions, and incorporate difference. In this sense, the argument that we have never truly been modern (Latour, 1993) is implicitly postcolonial. Thus an analysis that deconstructs imperial binaries such as nature–culture, modern–traditional, global–local builds on a postcolonial, or decolonizing, platform. Even if explicit recourse to postcolonial theory remains rare in science and technology studies, a postcolonial sensibility has infiltrated its critical scholarship.

'Postcolonial,' however, is a tricky word. It can signify a simple temporal distinction – what comes after the colonial – or a complicated analytic framework. Even more confusing, postcolonial analysis implies the opposite of postcolonial temporality, as it suggests the persistence of the complexly refigured colonial: that the colonial as a sort of contact zone, or place where differences meet, is always, and everywhere, with us. In this sense, the 'colonial' now needs to be deconstructed as a site of contestation, difference, and unstable hybridity – not the old colonial wish fulfillment of dominance and submission, of polar opposites. The inability to distinguish the temporal (and perhaps meretricious) postcolonial from the critical, wide-ranging analytic postcolonial leads to perplexity and bafflement even among experienced scholars (not to mention copy editors with an aversion to hyphens). Perhaps we should substitute 'postindependence,' or 'decolonized,' or 'national,' for the temporal postcolonial. But while this may allay confusion, it would be regrettable if critical postcolonial analysis were to sever some of its historical ties, attenuated as they are, to the earlier liberatory projects. Instead, we should remember that as shape-changing empire endures, the need to decolonize knowledge, or to deflate its colonial puffery, never ends.

Postcolonial studies of the spread and refiguring of European science and biomedicine offer an antidote to the pharmakon, the multivalent drug, of global flows and circulations, which has proven so popular in the twenty-first century. Sadly, consumption of hydraulic metaphors continues to exert a soporific effect on scholars of the globalization of science. Scientific knowledge and practice, if they no longer 'diffuse' out from Europe, now seem to flow and circulate unimpeded around the world, as in some alluring imperial fantasy. The history of science once was proudly universal, latterly national and social, sometimes comparative across nation states, but now it is customary to explain the globalization of science, technology, and medicine as a process of laminar flow. Postcolonial approaches, though, can reveal the terrain that channels such circulation, showing the historical and political forms and interactions, the systematic exclusions and inclusions that make these flows turbulent. Globalization narratives tend to produce ornamental differences, whereas postcolonial framings show how we perceive and deal with radical otherness. For many decades, critical scholars in science studies have excelled in situating scientific and biomedical knowledge and practice, in putting them in their place – but how do we connect these multiple sites? The imperial gaze sees smooth, inescapable global flows; postcolonial critics instead see messy, uneven politics and diverse, contending agents amid the historical debris.

Postcolonial Genealogies and Antecedents

Postcolonial studies of science boast a complex genealogy, including relations with South Asian subaltern studies, indigenous knowledge traditions, feminist standpoint theory, colonial histories of science and medicine, and critiques of development. The connections with formal postcolonial theory are surprisingly sparse and feeble. While the cultural and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85057-6

textual analysis of Edward W. Said, Homi K. Bhabha, Gayatri Chakravorty Spivak, and others often has stimulated inquiry and suggested key terms, its abstract style and narrow literary scope restrict its appeal. Esoteric postcolonial theorists have found it hard to light out into new territories. In contrast, postcolonial scholars in science and technology studies, drawing on different sources, have favored richly textured historical and anthropological narratives – indeed, they are notably inclined to combine historical and anthropological approaches. Thus conventional criticisms (Thomas, 1994; Cooper, 2005) of the obscuring abstractions of postcolonial theorizing prove inapplicable and redundant in the case of densely empirical postcolonial studies of science.

Starting in the 1980s, South Asian scholars and activists have engaged in vigorous debate about the status and significance of European science, technology, and medicine, viewing the relations of imported practices with local knowledge traditions through Gandhian and postcolonial framings. Critics of European scientific dominance such as Ashis Nandy (1988, 1995) and Shiv Visvanathan (1997) have valorized Indian knowledge traditions and traced the diverse terrain of scientific practice. For Gyan Prakash (1999: 7), colonial science was 'a sign of Indian modernity,' translated and appropriated by Hindu nationalists to advance their causes. These critics of colonial and neo-colonial science and medicine render their topic of inquiry far more heterogeneous, hybrid, and contested than it seems in simple diffusionist accounts. Regardless of claims for the transcendence or universalism of science, in practice it appears to be doing specific political and cultural work on the ground. Importantly, postcolonial studies of South Asian science and medicine have expanded the range of what constitutes science and who is recognized as a scientist. Like their subaltern studies associates, critics of colonizing science reorient Antonio Gramsci, perceiving European dominance without hegemony.

In troubled, self-critical settler societies like Australia, New Zealand (Aotearoa), and Canada – and less so in more complacent ones like the United States – postcolonial scholars have focused on the interactions of scientists and indigenous people. With the Yolngu of eastern Arnhem Land, Australia, Helen Verran (1998: 238) has described "the politics waged over ontic/epistemic commitments," whether to science or traditional knowledge. Verran hopes for a community that "shares imaginaries and articulates those imaginaries as part of recognizing the myriad hybrid assemblages with which we constitute our worlds" (1998: 238). Like other postcolonial critics, she appreciates the messiness, contingency, and ineluctable heterogeneity of knowledge practices (Watson-Verran and Turnbull, 1995; Verran, 2002). Similarly, David Turnbull has studied the "interactive, contingent assemblage of space and knowledge" in diverse settings, arguing that "all knowledge traditions, including Western technoscience, can be compared as forms of local knowledge so that their differential power effects can be compared but without privileging any of them epistemologically" (Turnbull, 2000: 6; see also Chambers and Gillespie, 2000). Exploration of the contact zones between knowledge traditions has stimulated other efforts to multiply the sites of knowledge making and to destabilize geopolitical dominance. Such efforts include Linda Tuhiwai Smith's validation of indigenous research practices in *Decolonising Methodologies* (1999) and Raewyn Connell's (2007) affirmation of "Southern theory" in the social sciences.

Since the 1990s, philosopher Sandra Harding (1998: 305) has tried to 'locate modern sciences on the more accurate historical and geographical maps produced by the postcolonial accounts.' Relying more on feminist standpoint theory than on any particular postcolonial doctrine, Harding imagines a multicultural science that might one day encompass the knowledge and experience of women and the peoples of the Third World. She seeks to integrate indigenous science into Western 'ethnoscience' and build support for autochthonous Third World projects. Harding draws on historical studies of imperial science and medicine, anthropological descriptions of ethnoscience, criticism of development programs, and the scholarship and activism of South Asian intellectuals, such as Nandy, Susantha Goonatileke, and Vandana Shiva. Teetering on the edge of essentialism, she wants to take up the standpoints of non-European cultures in order to make a better future. She laments the "persistence of legacies of male supremacy and Eurocentrism in the work of even progressive science and technology scholars" (Harding, 2008: 214). But Harding's framing of Western and other sciences is surprisingly bipolar, with each pole representing an opposing standpoint, eschewing heterogeneity and hybridity (Harding, 2011). Indeed, it is hard to envisage the twain ever meeting.

The critical anthropology of 'development' is another rich source of postcolonial analysis of science, technology, and medicine. To some extent, such analysis responds to Paul Rabinow's call to "anthropologize the West" – to show "how exotic [the West's] constitution of reality has been; emphasize those domains most taken for granted as universal (this includes epistemology and economics); make them seem as historically peculiar as possible; show how their claims to truth are linked to social practices and have hence become effective forces in the world" (Rabinow, 1986: 241). In *Encountering Development*, Arturo Escobar (1995: 19) thus sought to investigate modernity "as a culturally and historically specific phenomenon." Akhil Gupta (1998: 20), similarly, in his study of agricultural development in India, has used a postcolonial framework to describe "hybrid discourses and practices and to delineate the intertwining of 'local' practices with global and national projects of development." Like other postcolonial scholars, Gupta wants to destabilize the "binaries of colonial and nationalist thought in pointing to the imbrication of the Indigenous in modernist discourse" (1998: 20). Other anthropologists of development and technoscientific modernity, such as Stacy Leigh Pigg (1997) and Chris J. Shepherd (2004), have helped to foster the postcolonial analysis of science and medicine.

The emergence of postcolonial studies of science and medicine is a particular manifestation of more general efforts, gaining force in the 1990s, to decolonize knowledge, to make visible subjugated practices and conjugated subjects. Thus Walter D. Mignolo has pushed for the destabilizing of Occidentalism through "border thinking," which implies the recognition and recuperation of the radical differences that global translations attempt to obliterate. For Mignolo (2000: x), "border thinking is more than a hybrid enunciation. It is a fractured enunciation in dialogic situations with the territorial

and hegemonic cosmology." Border thinking functions as a "machine for intellectual decolonization" (2000: 45). Similarly, Fernando Coronil (1996) urged us to move beyond Occidentalism in recreating our conjectures and methods. In this sense, our theories about the world are catching up with, or being co-opted by, the dispersive logic and multiple registers of contemporary globalization.

Postcolonial Actors and Scientific Travels

During the past 40 years, we have sought to situate science, technology, and medicine ever more securely and intricately in time and place. No longer transcendent or universal, science has come to appear as a series of complicated local accomplishments, the result of knotty social negotiations on site. While scientists may think globally, they act locally. But how do we connect the multitude of local case studies? "We need to understand," writes Steven Shapin (1998: 6–7), "not only how knowledge is made in specific places but also how transactions occur between places." Having put the view from nowhere in its place, how do we get from there to somewhere else? "If facts depend so much on … local features, how do they work elsewhere?" asks Simon Schaffer (1992: 23).

Actor-network theory (ANT) has provided one of the more appealing explanations of how science travels and becomes stable in various places. Originally, ANT was supposed to demonstrate how a series of translations through a network might keep science and technology invariant in different settings. The extension of networks could fix facts, rendering them as 'immutable mobiles.' Bruno Latour described how an object becomes robust and stable through developing better articulation with human and nonhuman actors. Society, nature, and geography therefore are the outcomes, not the causes, of these mobilizations, translations, and enrolments. Polarities of center and periphery, global and local, make little sense in these distributed networks. As Latour (1993: 117) insisted, "even a longer network remains local at all points." But the local in early ANT often was a very abstract place, a point where lines intersected, a location depleted of historical and social specificity. Shapin (1998: 7) went so far as to criticize the "militaristic and imperialistic language that is so characteristic of Latour's work." It seems that ANT initially resembled colonial discourse, not postcolonial critique. But later versions of the theory emphasized the modification and adaptation of objects and practices as they traveled – they suggested a less orderly topology of science. In a study of the Zimbabwean bush pump, Marianne De Laet and Annemarie Mol (2000) show how the object shifted shape and meaning from one village to the next, refiguring relations as it did so, all the while remaining identifiably a bush pump. De Laet and Mol recognize the displacement, transformation, and contestation that occur when science and medicine travel, the proliferation of hybridity and refashioning of identities and places that happen with contact. Even so, they continue to express reluctance to become too engaged with Africa, to take the postcolonial turn.

Recently, Latour has positioned ANT alongside postcolonial studies as ways of doing the anthropology of modernity. ANT, he argues, is a means of localizing the 'global,' of recognizing the connected locales where global scale making happens. Thus "globalization circulates along miniscule rails resulting in some glorified form of provincialism" (Latour, 2005: 190). Just as postcolonial historian Dipesh Chakrabarty (2000) urged us to provincialize Europe, Latour seeks to provincialize "global" science. Additionally, Latour (2005: 204) wants to 'redistribute' what counts as local: "no place dominates enough to be global and no place is self-contained enough to be local." We should concentrate on how forms of knowledge are translated from site to site, how practices, people, and things move around, adapt, become standardized, are seen as comparable and commensurable. Latour's postcolonial predicates become evident when we substitute 'empire' for his deconstructed 'society,' or for the 'global.' Postcolonial criticism and ANT both challenge claims of sovereignty, both situate scale making, and both shift focus from structures of dominance to patterns of interaction and connection. But Latour's absolute aversion to structure renders him unable to consider power and inequality – he resists the thought that some people and institutions have more power to shape interactions and make connections than others. He provides us with an implicitly postcolonial analysis, shorn of critique. While Latour complains that critical studies of science too often put the cart – filled with structures, contexts, and histories – before the horse, in his vision the cart disappears, leaving nothing but the unburdened horse.

Postcolonial Technoscience

In the late 1990s, a number of historians of science and medicine, stimulated by new critical and heterogeneous imperial histories, came to embrace a postcolonial thought style, not only in their historical analysis, but also in explaining the dispositions of our contemporary science, technology, and medicine (Anderson, 1998, 2004). In effect, this meant thinking more anthropologically, working athwart conventional European trajectories. These postcolonial scholars regularly issued manifestos and other programmatic statements. They arranged workshops and special symposia within meetings of organizations like the Society for Social Studies of Science and the American Anthropological Association. When introducing a postcolonial thought style to science studies, they also were seeking to expand the scope of postcolonial studies, to add science and medicine to its usual literary repertoire.

In 2002, Gabrielle Hecht and I edited a special double issue of *Social Studies of Science* (32: 5–6), with the ambiguous title, "Postcolonial Technoscience." In the introduction, I wrote:

> A postcolonial perspective suggests fresh ways to study the changing political economies of capitalism and science, the mutual reorganization of the global and the local, the increasing transnational traffic of people, practices, technologies, and contemporary contests over 'intellectual property' … We hope that a closer engagement of science studies with postcolonial studies will allow us to question technoscience differently, find more heterogeneous sources, and reveal more fully the patterns of local transactions that give rise to global, or universalist, claims.
>
> (Anderson, 2002: 643)

The assembled articles, which addressed African nuclearity, Aboriginal fire regimes, Guiana space research, and American biosecurity fears, advanced our knowledge of "the localness of technoscientific networks, the situated production of 'globality', the transnational process of displacement and reconfiguration, the fragmentation and hybridity of technoscience" (Anderson, 2002: 651). Before long, the journal *Science and Culture* (14: 2, 2005) also offered a special issue on postcolonial studies of science (see the introduction by McNeill, 2005).

Some scholars suspected the politics of this enterprise, fearing that it might be harnessed to emerging nativist and nationalist knowledge claims. Itty Abraham, for example, supported postcolonial approaches that aim to "understand both the situatedness of local knowledge practices [and] their movement through space" (Abraham, 2006: 210). But he was concerned when the postcolonial was identified as a "site for understanding the clash of knowledges and the formation of alternative modernities." The rise of Hindu nationalism made him worried that "when the postcolonial is linked to a fixed site of irreducible knowledge claims, it articulates an ontology that ties knowledge to a location as a singular and essential quality of space" (Abraham, 2006: 210). Postcolonial scholars might end up as fellow travelers with the nationalists who assert their own special modernity. Abraham's warnings perhaps were most pertinent to those postcolonial theorists who tend to essentialize ethnosciences or discern place-specific 'standpoints.' Even so, scholars like Helen Verran, in her engagement with the Yolngu people in Australia, may argue that essentialized Indigenous knowledge performs a different politics in the settler society, where it operates as a decolonizing strategy, not a nationalist one. Although South Asia has been the modal site for postcolonial thinking, there may be times when we need to provincialize India in the postcolonial analysis of other sites.

When Vincanne Adams and I next surveyed postcolonial studies of science, we chose, somewhat quixotically, to make particular recourse to Said's 'traveling theory,' his advocacy of a critical spatial consciousness, recognizing the territorial entanglements of Western cultural forms with others. As the imperial project is taken apart, or heard as contrapuntal, "its incorporative, universalizing, and totalizing codes [are] rendered ineffective and inapplicable" (Said, 1983: 29). Adams and I believed:

> This requires a multiplication of the sites of technoscience, revealing and acknowledging hidden geographical notations and power relations, and further study of the mechanisms and forms of travel between sites. It means we have to be sensitive to dislocation, transformation, and resistance; to the proliferation of partially purified and hybrid forms and identities; to the contestation and negotiation of boundaries; and to recognizing that practices of science are always multi-sited.
>
> (Anderson and Adams, 2008: 183–184)

Although by this stage such statements may have become redundant, we went on to claim that postcolonial analysis offers a "flexible and contingent framework for understanding contact zones of all sorts, for tracking messy translations and transactions that take place between different cultures and social positions, including between different laboratories and disciplines even within Western Europe and North America" (Anderson and Adams, 2008: 184). That is, the postcolonial is not limited to former empires, but is available as a means of deconstructing any claims to sovereignty, including the scientific and medical ones asserted along North Atlantic shores. Peter Galison provides a vivid example of this methodological amplitude when he draws on postcolonial Pacific anthropology to describe the 'wordless pidgins' and 'wordless creoles' that allow communication between theorists, experimentalists, and instrumentalists in the 'trading zones' of Big Science in the United States – an analogy that recasts the epistemic privilege often assigned to theory in physics (Galison, 1997: 463–464).

Aside from various manifestos, and often indifferent to them, good work in postcolonial studies of science and medicine has proceeded apace. In view of the profusion of postcolonial inquiries, it is perhaps invidious, though necessary, to select a few recent examples. In *Being Nuclear*, Hecht (2012) displaces the standard geopolitics of things 'nuclear,' making Africa central to this science and industry, and thereby reshaping the 'global' nuclear order (see also Hecht, 2002, 2011). Claire Wendland (2010, 2012) examines medical training in Malawi, recasting Africa as a site of knowledge making and intellectual hybridity, not just of lack and inadequacy. In a provocative study, Johanna Crane (2013) shows how African suffering and structural inequalities permit the development of 'global health' research and other neo-colonial biomedical projects, which marginalize and discount local expertise. Gabriela Soto Laveaga (2009) puts Mexican campesiños, formerly invisible to historians, into the story of the development of the contraceptive pill, taking seriously their own scientific claims, and tracing how involvement in science changed their sense of themselves and their place in the world. In a series of articles, Amit Prasad (2005, 2006) maps the postcolonial trajectories of magnetic resonance imaging technology in India and the United States, revealing transnational shape shifting, the concatenation of epistemes and methods, in modern technoscience (see also Prasad, 2008). My own book on the transnational investigation of the brain disease kuru, *The Collectors of Lost Souls* (Anderson, 2008), also exemplifies the postcolonial study of technoscience. Applying insights from Melanesian anthropology to explain the complex exchange relations between scientists, local people, and other scientists, *Lost Souls* traces modern kula rings and other long-distance trade partnerships, casts scientists as sorcerers, and lets isolated 'primitives' speak as cosmopolitan figures. In it, I treat "ostensibly global science as though it were assembled out of a set of rather peculiar local achievements" (Anderson, 2008: 7). Such studies – a small sample of those available – are among the salients in the field, holding out against the sovereign 'global flows,' which prove so alluring yet so facile.

Conclusion

Postcolonial approaches in science studies allow for a recognition of the patchy decolonizing of scientific production and medical practice, a process occurring gradually for more than 100 years, and greatly amplified in the past 20 years. This style of analysis takes on the heterogeneity and hybridity

that finds expression in any contact zone, but it does so aware of the persistent or recursive colonial inequalities and exclusions that shape these situated interactions. Postcolonial scholars and activists seek to render more turbulent the supposedly laminar global flows, to deconstruct and impede grandiose claims to sovereignty and hegemony. Therefore they examine critically the multiple local sites where scale making and translation take place in science, and trace the connections and the ruptures between such places, across a rugged terrain.

And yet, it could be said that the imperatives and methods of science studies themselves have remained almost untouched by postcolonial critique. Escobar argues that the postcolonial 'crisis' in our representation of modernity calls for "new theories and research strategies; the crisis is a real conjunctural moment in the reconstruction of the connection between truth and reality, between words and things, one that demands new practices of seeing, knowing, and being" (Escobar, 1995: 223). Can science studies participate in postcolonial analysis without changing its methods? Is it business as usual over a bigger geographical range and with a larger cast of actors? In becoming postcolonial, how might science studies be transformed? What, for example, would making "Asia as method" (Anderson, 2012) entail for the sort of questions we ask of science, medicine, and technology – and for how we might answer them? Conventional historical and anthropological studies probably are inadequate for the task. We are waiting for the new generation of scholars and activists, who will show us what the postcolonial studies of science might become.

See also: Anthropology and History; Anthropology: Overview; Center–Periphery Relationships; Colonialism, Anthropology of; Colonization and Colonialism, History of; Cultural Geography; Cultural Rights and Culture Defense: Cultural Concerns; Cultural Studies of Science; Development Theory in Geography; Development: Social-Anthropological Aspects; Global South; Health: Anthropological Aspects; Indigeneity; Integration and Multiculturalism: Focus on Western Europe; Latour, Bruno (1947–); Medical Sociology; Postcoloniality; Science and Development; Science and Technology, Anthropology of; Standpoint Theory, in Science.

Bibliography

Abraham, I., 2006. The contradictory spaces of postcolonial techno-science. Economic and Political Weekly 41 (3), 210–217.

Anderson, W., 1998. Where is the postcolonial history of medicine? Bulletin of the History of Medicine 72, 522–530.

Anderson, W., 2002. Postcolonial technoscience. Social Studies of Science 32, 643–658.

Anderson, W., 2004. Postcolonial histories of medicine. In: Warner, J.H., Huisman, F. (Eds.), Medical History: The Stories and Their Meanings. Johns Hopkins University Press, Baltimore, MD, pp. 285–307.

Anderson, W., 2008. The Collectors of Lost Souls: Turning Kuru Scientists into Whitemen. Johns Hopkins University Press, Baltimore, MD.

Anderson, W., 2012. Asia as method in science and technology studies. East Asian Science, Technology and Society: An International Journal 6, 445–451.

Anderson, W., Adams, V., 2008. Pramoedya's chickens: postcolonial studies of technoscience. In: Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge, MA, pp. 181–204.

Chakrabarty, D., 2000. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton, NJ.

Chambers, D.W., Gillespie, R., 2000. Locality in the history of science: colonial science, technoscience, and indigenous knowledge. Osiris 15, 221–240.

Connell, R., 2007. Southern Theory: The Global Dynamics of Knowledge in Social Science. Polity Press, Cambridge.

Cooper, F., 2005. Colonialism in Question: Theory, Knowledge, History. University of California Press, Berkeley, CA.

Coronil, F., 1996. Beyond occidentalism: toward non-imperial geo-historical categories. Cultural Anthropology 11, 51–87.

Crane, J.T., 2013. Scrambling for Africa: AIDS, Expertise, and the Rise of American Global Health Science. Cornell University Press, Ithaca, NY.

De Laet, M., Mol, A., 2000. The Zimbabwe bush pump: mechanics of a fluid technology. Social Studies of Science 30, 225–263.

Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.

Galison, P., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago, IL.

Gupta, A., 1998. Postcolonial Developments: Agriculture in the Making of Modern India. Duke University Press, Durham, NC.

Harding, S., 1998. Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies. Indiana University Press, Bloomington, IN.

Harding, S., 2008. Sciences from below: Feminisms, Postcolonialities, and Modernities. Duke University Press, Durham, NC.

Harding, S., 2011. The Postcolonial Science and Technology Studies Reader. Duke University Press, Durham, NC.

Hecht, G., 2002. Rupture talk in the nuclear age: conjugating colonial power in Africa. Social Studies of Science 32, 691–727.

Hecht, G. (Ed.), 2011. Entangled Geographies: Empire and Technopolitics in the Global Cold War. MIT Press, Cambridge, MA.

Hecht, G., 2012. Being Nuclear: Africans and the Global Uranium Trade. MIT Press, Cambridge, MA.

Latour, B., 1993. We Have Never Been Modern (C. Porter, Trans.). Harvard University Press, Cambridge, MA.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, New York.

McNeill, M., 2005. Introduction: postcolonial technoscience. Science as Culture 14, 105–112.

Mignolo, W.D., 2000. Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking. Princeton University Press, Princeton, NJ.

Nandy, A. (Ed.), 1988. Science, Hegemony and Violence: A Requiem for Modernity. Oxford University Press, New Delhi.

Nandy, A., 1995. Alternative Sciences: Creativity and Authenticity in Two Indian Scientists. Oxford University Press, New Delhi.

Pigg, S.L., 1997. 'Found in most traditional societies': traditional medical practitioners between culture and development. In: Cooper, F., Packard, R. (Eds.), International Development and the Social Sciences. University of California Press, Berkeley, CA, pp. 259–290.

Prakash, G., 1999. Another Reason: Science and the Imagination of Colonial India. Princeton University Press, Princeton, NJ.

Prasad, A., 2005. Scientific culture in the 'other' theatre of 'modern science': an analysis of the culture of magnetic resonance imaging (MRI) in India. Social Studies of Science 30, 463–489.

Prasad, A., 2006. Beyond modern vs. alternative science debate: analysis of magnetic resonance imaging research. Economic and Political Weekly 41, 219–227.

Prasad, A., 2008. Science in motion: what postcolonial science studies can offer. Electronic Journal of Communication Information and Innovation in Health 2, 35–47.

Rabinow, P., 1986. Representations are social facts: modernity and post-modernity in anthropology. In: Clifford, J., Marcus, G. (Eds.), Writing Cultures: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA, pp. 234–261.

Said, E.W., 1983. Traveling theory. In: Said, E.W. (Ed.), The Word, the Text, and the Critic. Harvard University Press, Cambridge, MA, pp. 226–247.

Schaffer, S., 1992. Late Victorian metrology and its instrumentation: a manufactory of ohms. In: Bud, R., Cozzens, S.E. (Eds.), Invisible Connections: Instruments, Institutions and Science. SPIE Optical Engineering Press, Bellingham, WA, pp. 23–56.

Shapin, S., 1998. Placing the view from nowhere: historical and sociological problems in the location of science. Transactions of the Institute of British Geographers 23ns, 5–12.

Shepherd, C.J., 2004. Agricultural hybridity and the 'pathology' of traditional ways: the translation of desire and need in postcolonial development. Journal of Latin American Anthropology 9, 235–266.

Smith, L.T., 1999. Decolonizing Methodologies: Research and Indigenous Peoples. University of Otago Press, Dunedin, NZ.

Soto, L.G., 2009. Jungle Laboratories: Mexican Peasants, National Projects, and the Making of the Pill. Duke University Press, Durham, NC.

Thomas, N., 1994. Colonialism's Culture: Anthropology, Travel, and Government. Princeton University Press, Princeton, NJ.

Turnbull, D., 2000. Masons, Tricksters and Cartographers: Comparative Studies in the Sociology of Scientific and Indigenous Knowledge. Harwood Academic, Amsterdam.

Verran, H., 1998. Re-imagining land ownership in Australia. Postcolonial Studies 1, 237–254.

Verran, H., 2002. A postcolonial moment in science studies: alternative firing regimes of environmental scientists and aboriginal landowners. Social Studies of Science 32, 729–762.

Visvanathan, S., 1997. Carnival for Science: Essays on Science, Technology and Development. Oxford University Press, New York.

Watson-Verran, H., Turnbull, D., 1995. Science and other indigenous knowledge systems. In: Jasanoff, S., Markle, G.E., Petersen, J.C., Pinch, T. (Eds.), The Handbook of Science and Technology Studies. Sage, Thousand Oaks, CA, pp. 29–63.

Wendland, C., 2010. A Heart for the Work: Journeys through an African Medical School. University of Chicago Press, Chicago, IL.

Wendland, C., 2012. Moral maps and medical imaginaries: clinical tourism at Malawi's College of Medicine. American Anthropologist 114, 108–122.

Postcoloniality

John D Kelly, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11844–11849, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Postcoloniality is the unprecedented, complex predicament faced by individuals, societies, and governments after European colonial rule. This article addresses elements of postcolonial predicament including prospects for agency among dominated 'subaltern' groups, teh power of the West, liberation in theory and reality, and the relation of postcoloniality and globalization. Differences between diverse eighteenth-century and twentieth-century decolonizations lead to contrasting postcolonial ideologies: while theorists have recently sought ways to make 'postcolonialism' a viable political perspective for excolonies, in a different way, US ideology and power also manifest a postcolonial political will in globalization.

Postcoloniality is the unprecedented, complex predicament faced by individuals, societies, and their governments after the end of European colonial rule. An increasing number of social scientists, especially in the 1980s and 1990s, have found conceptions of postcoloniality useful for characterizing many contemporary societies. Theories of postcoloniality have emerged especially as more established alternatives – theories of modernization, of underdevelopment, of the Third World, and of revolution – have lost their power to explain the present and predict or guide the future. The premise of most postcoloniality theories is that 'The history of the West and the history of the non-West are by now irrevocably different and irrevocably shared' (Sangari, 1995, p. 147).

Writings on postcoloniality range widely in discipline and scholarly methods, and might be typifiable first of all by an attitude, even a mood: disillusionment coupled with commitment, disillusionment with the conspicuous failure of decolonization to produce a world of peace, prosperity, and real liberation, coupled with commitment to diagnose that failure and seek improvements or alternatives. In a British colonial prison during World War II, Jawaharlal Nehru wrote that 'Out of this crucible of war, wherein so much is melting, we hope that something finer will emerge for the West as well as the East, something that will retain all the great achievements of humanity and add to them what they lacked' (Nehru, 1982 [1946], p. 81). Thirty-seven years later, as he helped launch the 'Subaltern Studies' school of history, Ranajit Guha declared the 'historic failure of the nation to come to its own' in postcolonial India, its failure as either a bourgeois democracy on the European model or a socialist 'new democracy.' He declared the study of this failure 'the central problematic' for the historiography of India (Guha, 1982, p. 7). Similarly, Frantz Fanon wrote in the late 1950s, in the midst of the Algerian anticolonial war, that 'Decolonization is the veritable creation of new men,' that 'the European game has finally ended; we must find something different,' and that 'It is a question of the Third World starting a new history of Man' (Fanon, 1968, pp. 36, 312, 315). Thirty-five years later, in the Afterword added to his classic philological study *Orientalism*, Edward Said specifically abjured the idea of 'fixed identities battling across a permanent divide' of Europeans and others (Said, 1994, p. 335), in favor of more nuanced multicultural critique of power. While some theorists voice a continuing optimism that, especially in Southern Africa, Eastern Europe, and Southeast Asia, the end of the twentieth century has inaugurated a new age of revolutions (e.g., Comaroff, 1996), the dominant tendency in most studies of postcoloniality is critical dissection of utopian anticipations.

To date, discussion of postcoloniality has been more extensive in the Humanities than in the Social Sciences. Critical, often reflexive, theorizing about postcolonial identity and experience has been more extensive than social scientific research into political, economic, and cultural dynamics of postcoloniality. However, controversies in postcolonial theory have provoked fruitful research in anthropology, history, geography, sociology, political science, and other social sciences, characteristically entailing profound political as well as analytic questions. Good available readers on postcolonial studies include Williams and Chrisman (1994) and Ashcroft et al. (1995). This article centers on efforts to address three powerful antinomies in postcolonial theory, concerning (a) the agency of dominated groups, whether they determine their own fate, (b) the power of the West, whether Western paradigms are inevitable, and (c) the goal of liberation, and when and how postcoloniality might ever end. The article concludes with the tensions between postcoloniality theory and postmodernism, and with efforts to relate postcoloniality and globalization, arguing against portrayals of globalization as the end of postcoloniality, in favor of depiction of the two as intrinsically related.

Can the Subaltern Speak?

In 1988, literary critic Gayatri Spivak posed the above question as a challenge to the 'Subaltern Studies' school of historians (Spivak, 1988). The Subalternists, following Gramsci's theory of class politics in culture, have sought to uncover hidden realities of dominated people's experience and agency in South Asia. The 'Subaltern Studies' project, and Spivak's question, have led to debates and research globally on class, status, culture, and power especially in colonial and postcolonial societies.

The premises of the Subalternists have been, first, that prior histories of colonialism, decolonization, and the postcolonial nation-state have focused on elites, and second, that especially under colonial conditions, elite culture has not been hegemonic, i.e., has not established rituals and routines for social order unchallenged by 'subalterns' or lower class and status, dominated people. In order to explain why a unified postcolonial nation has failed to take shape, Subalternists seek to understand the social experience, collectivities and initiatives of 'subaltern' groups. The original South Asian 'Subaltern Studies' of peasant insurgencies and labor struggles have joined a much broader stream of scholarship on colonial and postcolonial politics, broader not only geographically but also topically, including especially in the 1990s many studies of sex and gender relations, institutions, and imaginaries. However, the issue of the failure of decolonization and the postcolonial nation-states remains central. Efforts have begun to identify positive, existing alternatives to 'the nation' as a form of social organization, not only in South Asia but also in East Asia (notably Duara, 1995), the Americas and elsewhere, and also to understand the social legacies of specifically colonial and postcolonial state structures (e.g., in Africa (Mamdani, 1996), in Indonesia (Pemberton, 1994), and in India (Pandey, 1990)).

In the context of research seeking to isolate colonial pathologies and celebrate successful resistances to them, Spivak's question attracted enormous attention because it challenged a second tier of scholarly premises. Phrased with a literary critic's focus on voice, 'Can the Subaltern speak?' quickly metamorphosed into a question about coherent, autonomous agency, and initiative. Is it possible for 'subaltern' people to shape their own society and culture? Precisely as a radical among philologists, Spivak had insisted that representation in a literary or semiotic sense must be reconsidered in connection with representation in politics, representation in the sense of any capacity for a person to be the agent of, to stand for, the will of other people. Do subalterns have any real power to represent themselves, or is that precisely the lack that constitutes them as subaltern?.

Social scientists have come to recognize empirical as well as theoretical issues here. In research on postcolonial societies, the question of subaltern political agency has intersected with studies of class and ethnic structure of anticolonial movements and postcolonial governments, and with studies of democracy in postcolonial conditions. While one tendency in contemporary scholarship is to discover unnoticed ways and means by which subaltern groups preserve themselves, maintain autonomous self-control, and even augment their life chances (e.g., Scott, 1985), another trend is to seek aspects of political structure, global and local, that might account for the reproduction of extremities of domination and violence in postcolonial societies (e.g., Tambiah, 1996), and a third trend is to seek to reintegrate the questions about colonial and postcolonial subalterns with studies of domination in metropolitan societies (e.g., Stoler and Cooper, 1997). Are postcolonial governments especially likely to be bound to the interests of urban, elite, and ethnically marked subpopulations? Are 'subaltern groups' in such societies less likely to exercise routine rights as citizens, and more likely to sustain whole institutional apparatuses of avoidance and resistance, 'informal' economies, even labor migrations kept deliberately invisible to the postcolonial state? No consensus has emerged. However, it does grow increasingly clear at least that, systematically, the nation-state does not 'come into its own' in postcolonial societies, and 'subaltern' classes in them are neither powerless nor as powerful as citizens in the West.

Provincializing Europe

Is the postcolonial world destined (or doomed) to follow the West's lead? To fit Western molds? Is the history of the West the central axis of world history? If it isn't, then how can sense be made of colonial history? Research assessing the successes and failures of postcolonial societies that sought alternatively to emulate and to refuse Western models (China, Singapore, Japan, Malaysia, Indonesia, India, and many sites in Africa), especially concerning successes of non-Western capitalisms, and the failures of self-proclaimed socialist revolutions, suggest that efforts to understand the West either as 'exceptional,' successful for particular cultural or material reasons, or as 'universal,' a distillation of necessities, both fail to grasp the more complex dynamics of the actual history of capitalism, colonialism, and the rise of the nation-state. Anticipations kindred to those of Fanon and Nehru, hoping more or less radically for a new order to supplant the civilization of Europe, were common among anticolonial intellectuals of the mid-twentieth century. These anticipations frequently entwined in a variety of ways with socialist and communist projects for social reconstruction and/or revolution, an end to a capitalist era. For this and other reasons, decolonization became deeply enmeshed in the narrow frame of the Cold War, and was dramatically reconceived within it. The question of the future relation of excolonies and excolonizers became, temporarily, reduced to matters of allegiance in a grand global dualism. Social scientists operated under the premise that all 'Third World' new nations were destined as they 'modernized' to join either the 'free' or 'Communist' worlds (see Pletsch, 1981). However, by the end of the century much that was taken for granted in this Cold War common sense was in doubt, was under scrutiny, or was repudiated outright, not only concerning the viability of a revolution against capitalism but also concerning modernization and the nation-state.

In the late-twentieth century, great popularity accrued among scholars for Benedict Anderson's vaguely Marxist theory that the nation as an 'imagined community' was central to the century's global culture (Anderson, 1983), but at the same time, disquiet about actual postcolonial nations and states led to new critical questions. Partha Chatterjee mounted the most influential challenge to the premise of the nation-state for excolonies. Chatterjee amplified and reconsidered the implications of parts of Anderson's argument, that 'the nation' was a form for politics linked intrinsically to Enlightenment social theories, a political format pioneered in the Americas and Europe, and then 'pirated' in already 'modular' format as colonies were refashioned and new nations built. Insofar as this is what happened, then, decolonization was not actually a break from Western hegemony. 'Europe and the Americas, the only true subjects of history, have thought out on our behalf not only the script of colonial enlightenment and exploitation,

but also that of our anticolonial resistance and postcolonial misery' (Chatterjee, 1993, p. 5). Chatterjee doubts that 'our imaginations must remain forever colonized' (Chatterjee, 1993, p. 5). However, the question of how to decolonize political imagination, and whether or where it has already been decolonized, has been a provocative one for postcolonial studies. While another influential critic has argued that postcolonial critique of Western universalisms has opened a 'third space' outside both local, cultural particularisms, and modernist universalism (Bhabha, 1994, pp. 36–9), others have defined postcolonialism instead as the dream that this could be done, 'the need, in nations or groups which have been victims of imperialism, to achieve an identity uncontaminated by universalist or Eurocentric concepts and images' (During, 1987, p. 3), or the quest for a new relationship with 'the intimate enemy' (Nandy, 1983). In this sense, as Lenin once hoped that a United Front politics in colonies could use bourgeois nationalism as a stepping stone to a proletarian revolution, a postcolonial perspective is proffered by some as the key to a new political future. With or without such a dream, the matter remains of how to understand the impact of 'the West' and its (possibly parochial) universals on one-time colonies in both the past and the present. Touchstones for controversy have included 'world system' models of a globally coercive Western-centered economy, and models of globally hegemonic Western culture. World system economic models have largely been encompassed by approaches emphasizing the irreducibility of both global and local economic forces, and the necessity to understand both. While support increases for models of economic 'globalization' that see global economic forces and relations increasing in their influence over time, the causes of the rise of the West as the pre-eminent capitalist center, and the place of colonialism in that history, are still much debated. Similar controversies have followed models of globally hegemonic Western culture, notably Edward Said's landmark study of the kind of discourse he named 'Orientalism,' discourse that constitutes the East and West comparatively, the East as the inferior antithetical other. Said's study has been criticized, ironically, for itself overriding the possibility of non-Western political discourses, for itself undermining the possibility of a world history centered anywhere but the West. Studies have also begun of processes of cultural globalization and of the crucial process of the localizing intermixture of global and local cultural elements that clearly accompanies the development of global cultural connections, the process Hannerz has named 'creolization' (Hannerz, 1992).

At their extremes, globalization models can be as simple in their symmetries as conceptions of postcoloniality can be stark in their asymmetries. Recharacterizations of global culture in terms of flows, links, and hybridities, and inquiry into 'modernity at large' (Appadurai, 1996), have added an important complexity to studies of postcoloniality, undermining the starkness and simplicity of the colonizer–colonized dualism, and challenging the premise of continuing cultural superiority of the West. However, the situation of postcoloniality requires attention to asymmetries. Reconsidering the past, present, and future place of the West, postcolonial theorists have grown as skeptical of triumphalist visions of reversal, of a non-Western overcoming of Western culture, as they are critical of intrinsically colonial visions of Western superiority over the rest. In *Provincializing Europe*, Dipesh Chakrabarty suggests a way out of the antinomies generated by efforts to reckon with Western domination (Chakrabarty, 2000). The paradox in all postcolonialisms that seek a triumphant postcolonial succession to Europe at history's center, Chakrabarty shows, is that even in such a success, the historical structure would still be 'First in Europe, and then elsewhere.' Europe indeed pioneered the form of historicist explanation that gives time a single processural center, giving capital and its culture of Enlightenment a unified place, time, and pattern of development in Europe, while occluding all inassimilable complexities of its earlier and later history elsewhere. However, Chakrabarty argues for abandoning such historicism, in favor of a method which never confines universal history, and provincializes Europe not from a new center, but instead by understanding Western cultural history, also, as one of creolizations (to use Hannerz's term here). Following Chakrabarty (or the similar, 'anti-Hegelian,' argument of Duara (1995)), one can provincialize Europe, confront and unmake colonial asymmetries, without a triumphalism that will inevitably, fatally, unmask as mimicry of European colonialism's own naive historicist egocentrism. Reflection on how, in recent but past centuries, Europe encountered and came in particular, powerful, but not all-powerful, ways to embody and develop universals of capital, democracy, science, etc., might point toward alternative developments in the past and the present, and possibilities not yet recognized for the future, rather than renewing an imperative to overcome empires already dead. In any case, postcolonial theorists have grown weary of battle with colonial ghosts.

When, and How, Can Postcoloniality End?

It will now be useful, and easy, to clarify the distinction between postcolonialism and postcoloniality. If postcolonialism is the aspiration for a profound political perspective to emerge from the experience of decolonization, postcoloniality is the unprecedented political predicament that decolonization has delivered many societies. The antinomies that fuel the struggle to constitute a postcolonialist political vision are, in less ideological form, equally inescapable for those without postcolonialist aspirations. Postcolonialist quests for political mission via subaltern standpoints struggle over what to expect from and for subaltern political agency. However, vast, dominated, and politically active populations no doubt exist in subaltern positions in postcolonial societies. Any coherent postcolonialism would seem to require a history both centered on and decentering of Europe, but, no doubt, Europe did once dominate, in various and varying ways, the societies it colonized. The question of how, and whether, that domination can be (or has been) ended is crucial to the third antinomy for postcolonialism, the third aspect of predicament in postcoloniality.

Anticolonial theorist and activist Franz Fanon insisted on the goal of liberation, not mere decolonization, for European colonies. Decolonization was something done for, and to, one-time colonies, where the goal should be actual liberation. In fact, as Fanon perceived, in the aftermath of World War II, the European empires disintegrated with extraordinary rapidity, and

(except for the French and Portuguese) with little violence: Fanon perceived why. In the atmosphere of the Cold War, Western powers and especially the newly pre-eminent United States greatly feared radical new states emerging from wars of liberation, and preferred to shape successor regimes and constitutions through a pre-emptive orderly process. As Fanon put it, the message flashing was 'for God's sake let's decolonize quick,' as the pieces of European empires were reconfigured into the nation-state form of 'self-determination' required by the new, US-fashioned United Nations (Fanon, 1968, pp. 70, 79).

By many diagnoses, decolonization and not genuine liberation was what the new 'Third World' received. But the problem of what would constitute genuine liberation, and which groups merit it, has become increasingly complex. As Firth (2000) has demonstrated, rhetorics of decolonization and liberation have expanded and proliferated after the breakup of most of the European empires has already been accomplished, expanding to articulate aspirations both of groups within nation–states (Biafrans in Nigeria, Kurds in Iraq) and new kinds of groups (women, indigenous peoples, even environments). If the premise of the United Nations has been that the nation-state as a political form constitutes the ultimate form of self-determination for all peoples, that simple, symmetric premise has been called into question. While most reckonings of postcoloniality focus upon the need to overcome asymmetries past, another approach entirely is to reckon it the onset, rather than the aftermath, of something, the predicament constituted for capital-poor excolonies by their re-establishment within the formal symmetries, and limited liabilities, of the United Nations model (Kelly and Kaplan, 2000). From this perspective, the goal of liberation itself has become a very specific trap, as formally independent and autonomous political standing comes systematically to entail perpetual clientage, expanding debts and ongoing substantive subservience to outside agencies and agendas. Like the 'free laborer' perceptively critiqued by Marx as free of all access to means of production, these liberated states are, by their very freedom and formal equality, radically limiting of the life chances of their citizens, especially their subalterns. In an interesting recent controversy, Tambiah (1996) rejects the arguments of Pandey (1990) explaining postcolonial communalist violence in South Asia as a legacy of colonial policies that emphasized and institutionalized communal political differences. Tambiah argues that endemic political violence is a phenomenon of democracy especially in postcolonial conditions, conditions that engender a spiral of destructive, effervescent leveling crowds. The crowds seek not liberation but leveling, destruction of privileges perceived as unmerited, especially across ethnic lines.

Boundary issues have arisen concerning which societies are postcolonial, especially concerning East Asian societies never formally colonized, and the Americas where the legacy of earlier, Iberian domination and earlier independence movements must be calibrated with the history of European colonies in Africa, Asia, and elsewhere. Sociologist Haejoang Cho has argued forcefully that Korea and other East Asian societies were colonized by the West, with becoming colonized defined as coming to require foreign means, methods, and tools simply to live and to think (Cho, 1992–94). For the Americas, Klor de Alva (1992) has argued that few or none of the societies outside the Caribbean are postcolonial in fact, that regardless of mestizaje ideologies, Latin American societies are descended not from colonized peoples but largely from the colonizers. As with most of Latin America, whether the United States is postcolonial depends upon the definition adopted. The question is most interesting when connected to different boundary issues concerning postcoloniality: its relations with postmodernism, and with globalization.

Is the post- in postcolonial the post- in postmodern? (So asked another literary critic, Appiah (1992).) Can either end? Dramatic, relativistic postmodernist proclamations that the Enlightenment universals are dead have been countered, Appiah argued, by the fact of misery in postcolonial societies, leading postcolonial political movements to ground their calls for new reforms in ethical universals. While postmodernism offered a first alternative vision to the stalled modernization logic of the Cold War, at present conceptions of both postcoloniality and globalization offer more dynamic alternative formulations. How, then, are postcoloniality and globalization connected? Some scholars envision globalization as the successor, the late capitalist liberator from postcoloniality, as the globe moves from Western modernity to modernity at large or to alternative modernities, negating all vestiges of asymmetric colonial relations. The counter argument (Kelly, 1998) is that postcoloniality and globalization are two sides of the same coin, the coin of American power, phenomena not necessarily of a 'late' capitalism but certainly of the new form capitalism has taken under US leadership in the twentieth century. If this is the case, if decolonization and globalization are the fruit of US-led quests for political self-determination and economic open doors, then Europe has already been provincialized by an explicitly anti-imperial agenda, and the predicaments of postcoloniality are the consequences, precisely, of a postcolonialism already instituted. David Scott has perceptively observed, in the first book with 'After Postcoloniality' in its title, that in the late twentieth century, in the wake of the global resurgence of liberal democracy and the failures of socialist states both Soviet and non-aligned, 'it is no longer so clear what "overcoming" Western power actually means' (Scott, 1999, p. 14). Perhaps, then, the time has come for dividing 'the West' into its colonial European and twentieth-century American versions. In that case, postcoloniality could end only when American power is as thoroughly confronted as European power has been, and the limitations intrinsic to the formal symmetries of the political present are as fully overcome as have been the formal asymmetries of the colonial past.

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Globalization and World Culture; Hegemony: Anthropological Aspects; Imperialism, History of; Postcolonial Law; Subaltern History; Third World.

Bibliography

Anderson, B., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Appiah, K.A., 1992. The postcolonial and the postmodern. In: Appiah, K.A. (Ed.), In My Father's House: Africa in the Philosophy of Culture. Oxford University Press, New York, pp. 137–157, chap. 7.
Asad, T., 1993. Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam. The Johns Hopkins University Press, Baltimore, MD.
Ashcroft, W., Griffiths, G., Tiffin, H. (Eds.), 1995. The Post-colonial Studies Reader. Routledge, London.
Bhabha, H., 1994. The Location of Culture. Routledge, New York.
Chakrabarty, D., 2000. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton, NJ.
Chakrabarty, D., 2009. Provincializing Europe: Postcolonial thought and historical difference. Princeton University Press, Princeton, NJ.
Chatterjee, P., 1993. The Nation and its Fragments: Colonial and Postcolonial Histories. Princeton University Press, Princeton, NJ.
Cho, H., 1992–94. T'alsikminji Chisikin ui Kullikki wa Samilkki [Reading Texts, Reading Life: The Point of View of a Post-Colonial Intellectual], 3 vols. Ttohaniuimunhwa Press, Seoul.
Comaroff, J., Comaroff, J.L., 2012. Theory from the South: Or, how Euro-America is evolving toward Africa. Anthropological Forum 22 (2), 113–131.
Comaroff, J., 1996. Ethnicity, nationalism, and the politics of difference in an age of revolution. In: Wilmsen, E.N., McAllister, P. (Eds.), The Politics of Difference: Ethnic Premises in a World of Power. University of Chicago Press, Chicago, pp. 162–183.
DelVecchio, M., Hyde, S.T., Pinto, S., Good, B.J. (Eds.), 2008. Postcolonial Disorders. University of California Press, Berkeley.
Duara, P., 1995. Rescuing History from the Nation: Questioning Narratives of Modern China. University of Chicago Press, Chicago.
During, S., 1987. Postmodernism or post-colonialism today. Textual Practice 1, 32–47.
Fanon, F., 1968. The Wretched of the Earth. Grove Press, New York.
Firth, S., 2000. Decolonization. In: Borofsky, R. (Ed.), Remembrance of Pacific Pasts: An Invitation to Remake History. University of Hawaii Press, Honolulu, HI.
Guha, R., 1982. On some aspects of the historiography of colonial India. In: Guha, R. (Ed.), Subaltern Studies I. Oxford University Press, Delhi, pp. 1–8.
Hannerz, U., 1992. Cultural Complexity: Studies in the Social Organization of Meaning. Columbia University Press, New York.
Hansen, T.B., Stepputat, F. (Eds.), 2009. Sovereign bodies: Citizens, migrants, and states in the postcolonial world. Princeton University Press, Princeton, NJ.
Kelly, J.D., Kaplan, M., 2000. Nation and Decolonization. University of Chicago Press, Chicago, IL.
Kelly, J.D., 1998. Time and the global: against the homogeneous, empty communities in contemporary social theory. Development and Change 29, 839–871.
Klor de Alva, J., 1992. Colonialism and postcoloniality as (Latin) American mirages. Colonial Latin American Review 1, 3–23.
Mamdani, M., 1996. Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism. Princeton University Press, Princeton, NJ.
Nandy, A., 1983. The Intimate Enemy: Loss and Recovery of Self Under Colonialism. Oxford University Press, New York.
Nehru, J., 1982[1946]. The Discovery of India. Oxford University Press, New Delhi.
Pandey, G., 1990. The Construction of Communalism in Colonial North India. Oxford University Press, New Delhi.
Pandian, A., 2008. Pastoral power in the postcolony: On the biopolitics of the criminal animal in South India. Cultural Anthropology 23 (1), 85–117.
Pemberton, J., 1994. On the Subject of 'Java. Cornell University Press, Ithaca, NY.
Pletsch, K.E., 1981. The three worlds, or the division of social scientific labor, circa 1950–1975. Comparative Studies in Society and History 23, 565–590.
Said, E., 1994. Orientalism. Vintage Books, New York.
Sangari, K., 1995. The politics of the possible cultural critique 7. In: Ashcroft, W., Griffiths, G., Tiffin, H. (Eds.), The Post-colonial Studies Reader. Routledge, London, pp. 143–147.
Scott, J.C., 1985. Weapons of the Weak: Everyday Forms of Peasant Resistance. Yale University Press, New Haven, CT.
Scott, D., 1999. Refashioning Futures: Criticism After Postcoloniality. Princeton University Press, Princeton, NJ.
Spivak, G.C., 1988. Can the subaltern speak? In: Nelson, C., Grossberg, L. (Eds.), Marxism and the Interpretation of Culture. Macmillan Education, Basingstoke, UK, pp. 271–313.
Stoler, A.L., Cooper, F., 1997. Between metropole and colony: Rethinking a research agenda. In: Cooper, F., Stoler, A.L. (Eds.), Tensions of Empire: Colonial Cultures in a Bourgeois World. University of California Press, Berkeley, CA, pp. 1–56.
Stoler, A.L. (Ed.), 2013. Imperial Debris: On Ruins and Ruination. Duke University Press, Durham.
Tambiah, S.J., 1996. Leveling Crowds: Ethnonationalist Conflicts and Collective Violence in South Asia. University of California Press, Berkeley, CA.
Williams, P., Chrisman, L. (Eds.), 1994. Colonial Discourse and Postcolonial Theory: A Reader. Columbia University Press, New York.

Postindustrial Societies

Brian A Hoey, Marshall University, Huntington, WV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term postindustrial society presupposes categorizing society based on an economic means of classification. Its use rests on assessing the relative status of manufacturing industry as an economic sector. Significant adjustment in sectoral location and nature of employment precipitated by late-twentieth-century deindustrialization in the developed world led many social theorists and critics to predict broad changes throughout domains of everyday life. Some began to speak not only of sectoral transformation but also of an emergent 'postindustrial society.' Following earlier agrarian and industrial 'revolutions,' postindustrialism suggested yet another revolution that would again transform how societies were organized.

General Introduction

To employ the term 'postindustrial' to society is to seek a manner of categorizing society largely dependent on an economic means of classification – or, at least, one using an economic point of reference. Specifically, of central concern is the relative status of manufacturing industry as a delineated sector in an overall economy. As the importance of this sector declines in a particular country as measured by its contribution to job creation and that society's gross domestic product – the market value of formally recognized goods and services produced in a given period of time – many social observers have chosen to describe such societies as postindustrial in order to emphasize displacement of the industrial sector.

In the late nineteenth century, Western countries shifted from economies largely dependent on agricultural production to one focused on industrial production of goods. This was a shift that led legions of workers to abandon farming and rural areas for higher wages in growing urban centers of manufacturing. For much of the twentieth century, countries that had led the Industrial Revolution – including England and the United States – experienced robust growth in industrial as well as service sectors. Beginning in the 1970s, however, it was evident that the industrial, or secondary, sector was joining the primary sector, defined by agriculture, in decline as practically all net employment growth could now be measured in an ascendant service sector – generally referred to as the tertiary.

This kind of broad economic change may be referred to as 'sectoral transformation' of the labor force. Importantly, while characterization of a society through reference to the relative importance of an economic sector might suggest a limiting perspective, exploring implications of a possible shift to a postindustrial economy has incorporated a broad range of subjects that will be at least partly addressed in this article. In the developed world, such a significant alteration in the sectoral location and nature of employment led many to predict far-reaching social changes throughout varied domains of everyday life. Some social theorists began to speak not only of sectoral transformation but also arrival of a 'postindustrial society.' That is, following both the agrarian and industrial revolutions, postindustrialism suggested yet another revolution that would again transform the way that a society itself was organized.

Historical Context for Development of Term

Although used as early as the second decade of the twentieth century, it was not until the years following World War II that the term postindustrial was again applied in discussions of social development. In his *The Age of Extremes: A History of the World, 1914–1991*, the historian and social theorist Eric Hobsbawm (1994) asserted that even while many societies had been profoundly transformed through this war, it was difficult to characterize a world of seeming constant change. In the period following the global watershed that this military conflagration represents, a variety of historians and social scientists sought to describe an emerging age and to confer an all-encompassing identity that sensibly pulled together disparate changes that were occurring since the war's end. Most attempts appended the prefix 'post' to one or another core concept used in characterizing the preceding, prewar era. Accordingly, scholars have offered such suggestive combinations as 'postcolonial,' 'post-Fordist,' 'postmodern,' and, of course, 'postindustrial.'

The latter two terms are clearly related in the attempts of those using them to make sense of significant changes taking place during the second half of the twentieth century. The term 'postmodern' has come to designate a period thought to have begun in the 1960s and generally thought to be defined by a decentered array of cultural paradigms coexisting without any one being dominant and that is economically determined by the particular qualities of a form of capitalism distinct from the prewar era. The economist Amitai Etzioni (1968) applied the term postmodern to mark the end of the modern period by virtue of sweeping changes after the war in the quality of communications, knowledge production, and energy. For the philosopher and literary theorist Jean-François Lyotard, knowledge in the postmodern is produced through questioning existing models, inventing novel ones, and rejecting the 'grand' or 'meta' narratives that he considered defining features of modernity. Such narratives would include expressions of modernist faith in progress – as achieved through science and technology – as the driving force behind history. Explicitly linking a postmodern period with emergent postindustrial society, Lyotard (1984: p. 3) writes that "Our working hypothesis is that the status of knowledge is altered as societies enter what is known as the

postindustrial age and culture enters what is known as the postmodern age." For Lyotard, social organization in the postindustrial society is built on rapidly expanding scientific knowledge, computer-processed information, and advanced technology while postmodern culture, at the same time, encourages deconstruction of master narratives and rejection of totalizing theory.

Not unlike social theorists a century earlier – such as Henri de Saint-Simon, Karl Marx, and Fredrich Engles – who tried not only to make sense of but also shape developments that they observed, there is a prescriptive element to the pronouncements of many late-twentieth-century social theorists. That is, one can see a degree of reflexivity that suggests an attempt to shape the direction of public policy, or at least the character of scholarly or public debate, through the basic quality of their analyses and accompanying assertions.

Orienting Theoretical Developments

In his *Old Worlds for New: A Study of the Post-Industrial State*, a collection of previously published essays, architect and social critic Arthur Penty (1917) called for abolition of industrial practices, development of what he referred to as the 'postindustrial state,' and a corresponding rise of decentralized, artisanal workshops. Ultimately, it was Penty's desire that – through these workshops – social spheres of work, leisure, and family could be brought together after having been functionally separated during the Industrial Revolution. For Penty, an avowed antimodernist, World War I marked the end of an era and time for fundamentally rethinking the nature of society. Further organization of industry through principles of 'scientific management,' those that characterized the prescriptions of the influential mechanical engineer Frederick Taylor, promised a decreased demand for labor that would exacerbate conditions that had just led to worldwide conflict.

Frederick Taylor's (1911) concepts and practices – such as those captured in his most cited text *The Principles of Scientific Management* – not only represented the beginning of a managerial era in industrial production but also hastened the craft era's end in the United States. Taylor's work highlights a burgeoning relationship between rationalization in general and labor control methods in particular. Although Taylor's principles were specifically intended to address perceived problems of labor while increasing production efficiencies in the workplace, effects of his scientific management spread widely from the factory floor to shape practices basic to the conduct of everyday life. This was a process facilitated, at least in part, by the influence of economist John Maynard Keynes whose theories provided a ready vehicle through which 'Fordist' ideals of consumption and the promotion of consumerism – a necessary complement to mass production – could be elevated, with the help of state agencies, to the status of primary engine for economic growth.

Henry Ford is widely credited with employing 'assembly lines' in his automobile factories as a means of streamlining routinized mass production of a standardized product. Embracing Taylor's principles, Ford maximized efficient use of inputs – both material and human capital – in order to maximize corporate profit. Born of his marked achievement putting scientific management to work on the assembly line, 'Fordism' has been used to describe not only what was increasingly practiced in the industrial sector during the twentieth century but also the ways of work generally and even the basic form of a society wrought by the imperatives of standardization and mechanization. Widespread application in the twentieth century of both Taylorist and Fordist principles of work rationalization led to what some have decried as a kind of degradation of work.

By the late 1960s, Fordism began to break down as America's industrial production began its steep decline. With arrival of the 1980s, the basic Fordist cycle of increasing wages coupled with rising production and consumer demand could no longer be sustained. Inflexible bureaucratic business organization and the work practices associated with them appeared too slow to respond to changing consumer demands. It was at this time that the term 'downsizing' entered the American vernacular as a means to describe the way in which a variety of companies sought to cut costs by employing fewer workers even while they sought to maintain or increase overall productivity.

An American sociologist drawing almost exclusively on evidence from the United States, Daniel Bell (1973) attempted to provide an explanation – with publication of his book *The Coming of Post-Industrial Society* – for emerging economic, social, and cultural conditions in developed countries during inchoate industrial restructuring. Increased efficiency in the industrial sector, first stimulated through scientific management and later achieved through technologically fueled automation, had led to sharp declines in manufacturing jobs. Bell asserted that changes taking place as a result of this shake-up constituted not merely a shift from an economy based on the production of goods to one based on the production of knowledge, but a fundamental change in the nature of knowledge itself. Specifically, this transformation would be accompanied by increasing importance of scientific rationality expressed in economic, political, and social spheres.

This change in the production and quality of knowledge would be aided by technological advances such as powerful computers that could automate data collection and manipulation and thus place increased value on analytical, higher order knowledge skills. Daniel Bell's (1960) work on postindustrial society built on his earlier book *The End of Ideology* in which he argued – with modernist optimism – that ideologically driven politics would give way to a more deliberate, conscious guidance of the economic sphere through expert application of scientific and theoretical knowledge. While *Coming of Post-Industrial Society* may be seen as a rethinking of his previous position that technocratic doctrine would bring a socially engineered peace facilitated by informed experts seeing to a rational distribution of wealth, Bell held that this emergent postindustrial society was one in which a free-market system alone cannot be left to set social priorities.

While Bell identified most of the transformations that he envisioned for a postindustrial society as broadly beneficial, in Daniel Bell's (1976) book *The Cultural Contradictions of Capitalism* he nevertheless identified a growing disconnect between social structure and culture wherein many people would find themselves with less clearly defined roles than in

earlier times. In the postwar period, service sector jobs expanded rapidly leading to changes in both the character and social significance of employment. Developed nations like the United States were beginning to contend with emergent issues born of what many began to refer to as 'mass' society – including mass consumption and, specifically, mass culture. Writing a book that would help define a genre of sociological and other inquiries, law professor and later sociologist David Riesman's (1958) book *The Lonely Crowd* claimed that in those countries that had entered a stage of mass culture, individuals were increasingly guided or 'directed' by norms and values conveyed to them by mass media – enabled by technological advances in communication – as well as by peer groups defined, at least in part, through shared engagement in particular leisure activities.

In this way, Riesman felt that a defining, even 'revolutionary,' characteristic of postindustrial society would be a shift from an 'inner-directed' social character derived from norms and values of the family and, to a lesser extent, community, to an 'other-directed' individual who internalizes the perceived norms and values of an idealized, general peer group not necessarily like people living nearby. In his writing, Riesman used the term postindustrial society to connote growing importance of leisure over work and specifically growth of leisure as a central value of lifestyle. Similarly, sociologist Ralf Dahrendorf (1959) described how – in what he termed emergent 'postcapitalist' societies – occupational roles would lose their ability to comprehensively mold the social identity of workers as other spheres of individual and collective behavior took on greater importance.

Continuing in this vein, sociologist Dean MacCannell (1976) asserts that elements of leisure, in the realm of consumptive behavior, may now dislodge work as an organizing force behind social arrangements of contemporary society. Specifically, MacCannell holds that the domain of work has exhausted what he described as a capacity for 'regeneration' such that new forms of social organization would no longer materialize in the realm of production. According to Dean MacCannell (1976: p. 6), 'Lifestyle,' a generic term for specific combinations of work and leisure, is replacing 'occupation' as the basis of social relationship formation, social status, and social action."

Such a vision proposes that advances in automation and productive capacity of science – realized through technology – will together supplant the work ethic with an emphasis on leisure as source of personal and collective meaning. This resonates with the notion that in a postindustrial socioeconomic order, the immaterial conception and marketing of goods, as well as the broad means through which they are channeled to consumers, diminishes the social importance attributed to production and the work associated with it. Daniel Bell (1973: p. 348) provides a simple causal factor for this transition from social value being placed in productive activities to it being placed in consumptive activities in his singular statement that "A postindustrial society is one in which the majority of those employed are not involved in the production of tangible goods."

As a management consultant, the self-styled 'social ecologist' Peter Drucker (1969) is typically credited with early recognition of key aspects of what many now refer to as the New Economy in his description of an ascendant class of 'knowledge workers' and, indeed, arrival of a 'knowledge society.' References to 'new economy' may serve as a sometimes alternative reference to the postindustrial and is used to describe an extensive revision of the established paradigm of a capitalist economy. Like others attempting to describe socioeconomic changes over the past several decades, Peter Drucker (1993: p. 8) described an emergent society based on this new economy as one in which "the basic economic resource – 'the means of production,' to use the economist's term – is no longer capital, nor natural resources, nor 'labor.' It is and will be knowledge." Similarly, the sociologist Alain Touraine (1988: p. 104; see also Touraine, 1971) states that "The passage to postindustrial society takes place when investment results in the production of symbolic goods that modify values, needs, representations, far more than in the production of material goods or even of 'services.' Industrial society had transformed the means of production: post-industrial society changes the ends of production, that is, culture."

While defining oneself by way of a job might have been something embraced by urbanist and writer William Whyte's (1956) memorable character, the 'Organization Man,' who as corporate employee not only worked for but also – effectively and affectively – belonged to 'the Company,' social scientists such as Richard Sennett (1998) assert that the economic conditions of a new culture of capitalism – one that reshapes the very nature of work – provide no guarantee of stability in one's job as a source for self-definition. While such ambiguity is not new, Richard Sennett (1998: p. 31) explains that when compared with any other time in history, uncertainty in the 'flexibility' that appears to characterize a postindustrial society is peculiar only by virtue of the fact that it "is woven now into the everyday practices of a vigorous capitalism. It is meant to be normal."

Departing from the ideal of a standardized, Fordist industrial world, today's worker must exhibit flexibility by being malleable and forever learning. The cultural historian, John Gillis (1996: p. 232) suggests that in the postmodern era, postindustrial societies encourage people of all ages "to think of themselves in a perpetual state of becoming … [who must] retrain, re-educate, and recycle." One of the defining features of a postindustrial society may be that workers are expected to see themselves as a set of disembedded skills rather than as a complete, integrated individual with a particular, discrete job.

At a time when economic restructuring had roiled the US economy, the term 'contingent work' was coined by Audrey Freedman (1985: p. 35), an economist at the US Bureau of Labor Statistics, to describe "conditional and transitory employment arrangements as initiated by a need for labor – usually because a company has an increased demand for a particular service or a product or technology, at a particular place, at a particular time." As companies moved to what came to be known as 'just-in-time' manufacturing during the 1980s, it was only a short step to begin following a model of just-in-time employment. Employers today increasingly want what are variously referred to as 'portfolio workers,' 'career gypsies,' 'project workers,' 'temps,' and 'task consultants' to do specific but transitional work.

The term 'free agent' has been used to characterize a growing number of workers who are in some manner working for themselves and living lives documented by social commentator Daniel Pink (2001) in his popular book *Free Agent Nation: The Future of Working for Yourself*.

While Pink and others suggested a postindustrial society composed of the *de facto* self-employed, some envisioned a world without workers. Declaring that the 'Information Age' had arrived through advances in information technology and communication, economist Jeremy Rifkin (1995) held that 'intelligent machines' would promote eventual elimination of human labor from production processes and usher in a 'near-workerless' world. In his book, *The End of Work*, Rifkin (1995: pp. 56, 293) asserts that while "the century-old utopian dream of a future techno-paradise is within sight," its arrival would not come without potential peril for he also warns "the end of work could spell a death sentence for civilization as we have come to know it." Having been transformed by changes in the world of work, Rifkin proposed that postindustrial societies must provide productive outlets for displaced human labor through work in a 'third sector' that would stand apart from either private or governmental sectors of the economy. This would be a realm of nonprofit, voluntary, community-based work basic to civil society. In Rifkin's view, with governments abandoning their long-standing position as backup employer and with a formal, private sector increasingly ineffective at creating enduring jobs for millions of people in search of employment all over the world, it is necessary to reimagine work itself through developing a robust third sector to absorb workers displaced by economic and governmental restructuring. Thus, postindustrial society would be one built out of the need to address the fundamental problem of widespread unemployment.

Some theorists point to a shift from what Lash and Urry (1987) refer to as 'organized capitalism,' characteristic to the mid-twentieth century, to a 'neoliberal' form of (late) capitalism. Important in this move is the end of Fordism as a framework for defining the relationship between capital and labor. Specifically, while under Fordism – born in the industrial sector of the economy – organized labor fared reasonably well in terms of pay and job security, in the neoliberal, post-Fordist socioeconomic order, workers are contingent, disposable. At the same time, Keynesian principles for the relationship between the economy and government are thought to have shifted from one of regulation and investment in social programs to deregulation and disinvestment – a trend that may be seen as either challenging or hastening fulfillment of the vision outlined by Rifkin.

Tensions and Competing Visions

Given that rapid urbanization has been tied in large part to the growth of industrial economies during the previous two centuries, processes of deindustrialization within developed countries – understood as variously linked to the emergence of postindustrial society – have contributed to declining economic opportunity and outmigration from urban areas in most regions formerly dependent on large industrial employers. Thus, in light of economic restructuring leading to the closure of these factories, planning and policy discussions in the United States regarding urban and regional development often associate postindustrial society with alternative models for growth and, specifically, attempts by those communities most affected by deindustrialization to attract and retain residents as well as economic capital. The prevailing approach to encouraging capital investment in many such places has been to cut taxes and provide cheap land and labor in order to attract big employers. It is this 'smokestack chasing' strategy – as it is sometimes called – that various communities now challenge in their attempt to usher in a postindustrial economy, at least at the local level, and assure some stake in what is taken to be emergence of a broader, postindustrial society.

Many towns and cities in the northern tier of the midwestern and northeastern United States that collectively comprise what many refer to as the Rust Belt – in order to conjure images of decaying industrial places from another economic era, have embraced the promises of New Urbanists and others. At the core of their planning prescriptions is a call to create 'healthy neighborhoods' defined by walkable scale, open spaces for public recreation, a range of housing options and businesses in 'mixed-use' design, and a 'sense of place' that evokes traditional, human-scale urbanism. Proponents assure economic benefit from such development through attracting an emergent social demographic referred to as 'cultural creatives' by sociologist Paul Ray and psychologist Sherry Anderson (2000) and by urban studies theorist Richard Florida (2002) as the 'creative class.' For former industrial places, the image of a postindustrial society is taken as an ideal to achieve through nurturing such 'creatives' for whom community development refers to enhancements in quality of life enabled by way of progressively vibrant and diverse economic and social environments.

For some, postindustrial society promises a world wherein traditional social and cultural boundaries might be transcended. Perhaps reproducing a technocratic 1950s confidence and sense that modernization would make possible what seemed – following the war – a necessary break with an outmoded and menacing past, late-twentieth-century futurists held that the new nature of capital would provide everyone equal opportunities to benefit from developments in information and technology and lead to an emphasis on educational achievement. Further, a fundamental reliance on human – rather than natural – resources and capital would place less strain on the environment as the very character of work shifted even farther away from a direct confrontation with nature.

Alvin Toffler's (1980) book *The Third Wave* might be considered an expression of this broadly idealized vision of postindustrial society. Reflecting an understanding that human history could be understood in terms of two previous 'waves' entailing the agricultural and industrial revolutions, Toffler projected an incipient wave that would radically transform the world of work and end the dehumanizing effects of industrial labor so maligned by Penty and others. Toffler forecast creation of a distinctly postindustrial lifestyle that would be rooted in an idealized, pastoral landscape where workers could relocate from overcrowded and often fractious metropolitan areas for quality of life considerations while remaining electronically connected to service- and

information-based jobs located elsewhere. These postindustrial workers would value careful use of resources and environmental protection. In his studies of migration trends within the United States, economist Jack Lessinger (1986) suggests that in postindustrial society one will see a category of person appear that he terms the 'caring conserver' who is born of dissatisfaction with status quo values that define fragmenting relationships of work, family, and community that have characterized industrial society, generally, and American suburbia, in particular. In his view, the suburb will fall to a vast, deconcentrated 'penturbia,' as he brands the nonmetropolitan areas of the country, through the force of changing consumer tastes.

For those with far less boosterish views of postindustrialism, the relatively utopian visions offered by Toffler and others grossly exaggerate the transformative, leveling powers of information. For these theorists, the intellectual demands necessary to be competitive – coupled with ever-decreasing demand for manual forms of labor – would put currently marginalized groups at even further social and economic disadvantage and serve to deepen existing achievement gaps. Thus, historically advantaged groups who would take the lead in developing new information-based, service industries could maintain their economic, social, and cultural privilege. Generally in the pessimistic camp, Alain Touraine held that the benefits of education and knowledge would continue to be unevenly distributed among members of society and, further, that one would likely see conflict grow between those whose knowledge embodied humanistic principles and those who purposefully developed what were now more socially valued scientific and technical forms of knowledge.

Not only would social class remain an influence in educational opportunity within developed and developing countries alike, but differentiation in opportunity and access would also be deepened as historically poor countries were left unable to catch up given the rapid pace of change. In this sense, postindustrialism would be based on emergence of a highly developed and fundamentally unequal division of labor internationally in which established industries could transfer their production (and other) processes. In the 1980s, the term 'outsourcing' (like 'downsizing') joined American speech and tellingly described the turn to contracting out certain elements of business. Later, the term 'offshoring' emphasized wholesale relocation of certain functions out of country. According to geographers Allen Scott and Edward Soja (1996), although major sites of industrial production may lie largely outside of those countries that could be labeled postindustrial, their societies remain fundamentally bound to industrial production as a basic engine of growth and focus for processes of capital accumulation. Such an assessment identifies service sector expansion in the developed world with redistribution of manufacturing jobs to relatively low-cost and regulatory-lax developing countries in Asia and Latin America. In this interpretation, one could expect the traditional compromise of industrial capitalism between social and environmental justice and economic productivity to continue.

Basically, the comparatively dystopian perspective on postindustrial society lends itself to the assertion that it is wrong to hold that information-based capitalism is in any way meaningfully different than its forebear in industrial capitalism. From this viewpoint, extant postindustrial economic organization can only reproduce a capitalism fraught with persistent problems tied to class struggle and inequality rather than usher in a fundamental transformation of society. Thus, while it became accepted practice in the 1990s to refer to Eastern Europe as 'postcommunist,' virtually no scholar has embraced the term 'postcapitalist,' applied by Ralf Dahrendorf (1959) decades earlier as a label to describe Western Europe. Peter Drucker (1993), however, is a notable exception in his insistence that a rise of the knowledge worker must challenge the understanding of capitalism given that these workers have ownership of the means of production in the form of their accumulated education, ideas, and skills. One can see that this, at least partly, is a reflection of Dahrendorf's position that he was witness to a shift in 'authority relations,' of power, from control of the means of production, or property generally, to managerial and administrative control of capital. At the same time, however, Drucker nevertheless describes a society that continues to be defined by a basic dichotomy – a class division – between an educated elite of 'knowledge workers' and those that toil without comparative education, skills, or autonomy as 'service workers' in conditions not fundamentally different from those of the scientifically managed assembly line.

In *The McDonaldization of Society*, sociologist George Ritzer (1993) speaks to the split between classes of workers described by Drucker by centering on the milieu of legions of service workers. Ritzer reconceptualized Max Weber's arguments about the bureaucratic organization of society through applying his model for the fast-food industry wherein four primary components are efficiency, predictability, calculability, and increased control through replacement of human labor and capital with technology. Ritzer holds that the profitable operationalization of these components leads to further rationalization of basic forms of human interaction and broader social structures in both the economic and social spheres. Ritzer's position is that principles employed in the fast-food restaurant – a major component of the service sector – would come to dominate many aspects of society far from their origins in those businesses. That the components of McDonaldization demand organizationally defined efficiencies, highly quantified assessments, maintenance of systemization, formalization, and routine, and a rigid concern for managing corporate uncertainty, challenges assertions that societies in the developed world can be properly described as 'post-Fordist.'

The historian Christopher Lasch (1972: p. 47) argued that "The post-industrial order, far from transcending the contradictions inherent in capitalism, embodies them in an acute form." In his review of Bell's *The Coming of Post-Industrial Society*, Christopher Lasch (1974) contends that Bell posited a simplistic sequence of social development in which a postindustrial society summarily replaces a fundamentally capitalist, industrial one. This would suggest that Bell assumed that society developed through distinct phases and that industrial capitalism was but one of them, in manner akin to Karl Marx. In fact, Bell was quite critical of Marxist insistence on a binary class system and suggested, for example, that companies were increasingly owned by a fluctuating assemblage of shareholders – to whom they

must ultimately answer – rather than a singular class of wealthy individuals. In his new foreword to a revised edition of his book, Daniel Bell (1999: p. xv) referred to this as an economic, if not social, system wherein one approached the anomaly of having 'capitalism without capitalists.'

Examining restructuring of the economic system during the 1980s in developed countries like the United States, Manuel Castells (1996) described a rise of what he described as 'informational capitalism,' the creation of 'network society,' and the 'feminization' of work. In his book *The Rise of Network Society* (1996), the first of a triology titled *The Information Age: Economy, Society and Culture*, Castells spoke of how networks were becoming a key feature of 'social morphology,' and would serve as basic units of postindustrial societies. In these emergent social arrangements, 'networking logic' not only modified the operation and outcomes of production – through technological change and a shift from vertical bureaucratic structures to decentralized, horizontal organizations – but also informed cultural changes associated with networking. Transcending space and time in a capacity for pervasive, perpetual communication, Castells holds that the contemporary global economy is founded on instantaneous flows – a so-called space of flows wherein economic and cultural capital are exchanged – that shape both productive processes and patterns of consumption. Many social observers point to ways in which 'social networking' made possible by forms of so-called new media – as might be seen in the use of Facebook or Twitter – can be taken as illustrations of how network society organizes and, in particular, how this organization can take place outside the control of state apparatus.

Within this emerging socioeconomic order, Castells (2000: p. 12) describes how the "flexible woman gradually replace(s) the 'organization man' as the harbinger of a new type of worker" as well as transforms existing social relationships through redefining such basic categories as gender, family, and sexuality. Despite this possibility for transformation, like Drucker, Castells maintained that a fundamental divide would remain among classes of workers. In the networked society, this divide exists between 'informational' or networked labor – workers who are 'self-programmable,' adaptable, highly educated, and skilled at managing information – and 'generic' laborers, such as those in the service sector, who are relatively deskilled, exchangeable, and disposable.

Although quite a number of theorists of postindustrial society have advanced the primacy of knowledge and information not only as basic to the conduct of contemporary business but also as fundamental principles for social organization, it may be argued that the basic values of capitalism – even if described, perhaps, as 'late-capitalism' – determine what kinds and by what means informational media, knowledge, and technology are produced, distributed, and consumed. The postmodern cultural logic of a late capitalism has been described by Frederic Jameson (1984) and others as fundamentally derived from an emphasis on fragmentation, commodification, acceleration, and simulation that threatens a controlling influence in all spheres of contemporary life. Speaking on this control, the social theorist Jean Baudrillard (1980: p. 137) referred to postindustrial society as one dominated by a 'neo-capitalist cybernetic order' of electronic media where "we live in a world of proliferating information and shrinking sense" that effectively damages the existing social order by destroying the 'real' and substituting 'simulacra,' complex simulations that provide spectacle instead of meaning. For his part, Jameson (1984: p. 78) insists that the consumer capitalism that characterizes a possible postindustrial society is "the purest form of capital yet to have emerged, a prodigious expansion of capital into hitherto uncommodified areas."

Over a century ago, Emile Durkheim envisioned that 'mechanical solidarity' among members of preindustrial societies was transformed into an 'organic solidarity' among members of industrial societies that was shaped by an elaborate division of labor accompanied by a system of rights and duties that bound people together in an enduring fashion. While acknowledging a global shift to economic postindustrialism, the economist Daniel Cohen (2009) wonders whether it is possible to herald a postindustrial society when the Durkheimian solidarity that characterized industrial society appears lost to systematic disassembling of the components of large industrial enterprises, increasing contractual or contingent work, and the physical and structural distancing of workers from each other. In a similar manner, Eric Hobsbawm (1994: p. 334) writes that "The cultural revolution of the later twentieth century can thus best be understood as the triumph of the individual over society, or rather, the breaking of the threads which in the past had woven human beings into social textures."

Concluding Thoughts

There appears a fundamental tension in thinking about the possibility for a postindustrial society. There are those who appear ready to embrace the term 'postindustrial' when applied to the economic sphere, but decline to assert arrival of post-industrial society. On the other hand, others are disinclined to describe any society as economically postindustrial given that they find industrialism remains essential to the global economic system of which all societies are a part.

The term 'postmodern' has generally been applied to discussions of present-day culture – as opposed to society – and typically refers to how the status of knowledge has been altered through the peculiar qualities of a form of capitalism that may be distinct from that of a widely industrial age. It is generally believed, by those applying the term, that this is about much more than a shift from an economic order based on production of goods to one based on production of knowledge. Rather, it entails fundamental changes in the nature of knowledge itself and attendant changes in the footings for social organization through evolving conditions of labor and, indeed, the very meaning and place of work in contemporary culture and society. While industrial society transformed the means of production through the large-scale assembly of standardized, material goods, what may be characterized as the 'post-industrial' appears to transform culture through the production of symbolic goods, ascendance of media, and emergence of mass society.

See also: Economic Anthropology; Globalization, Anthropology of; Industrial Societies; Modernity; Modes of Production.

Bibliography

Baudrillard, J., 1980. The implosion of meaning in the media and the implosion of the social in the masses. In: Woodward, K.M. (Ed.), The Myths of Information: Technology and Postindustrial Culture. Coda Press, Madison, WI.

Bell, D., 1960. The End of Ideology: On the Exhaustion of Political Ideas in the Fifties. Free Press, Glencoe, IL.

Bell, D., 1973. The Coming of Post-Industrial Society: A Venture in Social Forecasting. Basic Books, New York.

Bell, D., 1976. The Cultural Contradictions of Capitalism. Basic Books, New York.

Bell, D., 1999. The Coming of Post-Industrial Society: A Venture in Social Forecasting, Special anniversary ed. Basic Books, New York.

Castells, M., 1996. The Rise of the Network Society. Blackwell Publishers, Malden, MA.

Cohen, D., 2009. Three Lectures on Post-Industrial Society (McCuaig, Trans.). MIT Press, Cambridge, MA.

Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford, CA.

Drucker, P., 1969. The Age of Discontinuity: Guidelines to Our Changing Society. Harper & Row, New York.

Drucker, P., 1993. Post-Capitalist Society. HarperBusiness, New York, NY.

Etzioni, A., 1968. The Active Society: A Theory of Societal and Political Processes. Free Press, New York.

Florida, R., 2002. The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life. Basic Books, New York, NY.

Freedman, A., 1985. The New Look in Wage Policy and Employee Relations. Conference Board Report. The Conference Board, New York.

Gillis, J., 1996. A World of Their Own Making: Myth, Ritual, and the Quest for Family Values. Basic books, New York.

Hobsbawm, E., 1994. The Age of Extremes: A History of the World, 1914–1991, first American ed. Pantheon Books, New York.

Jameson, F., 1984. Postmodernism, or the cultural logic of late capitalism. New Left Review 146, 53–92.

Lasch, C., 1972. Toward a theory of post-industrial society. In: Hancock, M.D., Sjoberg, G. (Eds.), Politics in the Post-Welfare State: Responses to the New Individualism. Columbia University Press, New York, pp. 36–50.

Lasch, C., 1974. An Exchange on Post-Industrial Society. New York Review of Books, New York, pp. 49–52.

Lash, S., Urry, J., 1987. The End of Organized Capitalism. University of Wisconsin Press, Madison, WI.

Lessinger, J., 1986. Regions of Opportunity: A Bold New Strategy for Real-Estate Investment with Forecasts to the Year 2010. Times Books, New York.

Lyotard, J., 1984. The Postmodern Condition: A Report on Knowledge. University of Minnesota Press, Minneapolis, MN.

MacCannell, D., 1976. The Tourist: A New Theory of the Leisure Class. Schocken Books, New York.

Penty, A., 1917. Old Worlds for New: A Study of the Post-Industrial State. G. Allen & Unwin Ltd, London.

Pink, D., 2001. Free Agent Nation: How America's New Independent Workers are Transforming the Way We Live. Warner Books, New York.

Ray, P., Anderson, S., 2000. The Cultural Creatives: How 50 Million People are Changing the World. Three Rivers Press, New York.

Riesman, D., 1958. The Lonely Crowd: A Study of the Changing American Character. Yale University Press, New Haven, CT.

Rifkin, J., 1995. The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-market Era. G.P. Putnam's Sons, New York.

Ritzer, G., 1993. The McDonaldization of Society: An Investigation into the Changing Character of Contemporary Social Life. Pine Forge Press, Newbury Park, CA.

Scott, A., Soja, E., 1996. The City: Los Angeles and Urban Theory at the End of the Twentieth Century. University of California Press, Berkeley, CA.

Sennett, R., 1998. The Corrosion of Character: The Personal Consequences of Work in the New Capitalism. W.W. Norton, New York.

Taylor, F., 1911. The Principles of Scientific Management. Harper & Brothers, New York.

Toffler, A., 1980. The Third Wave. Bantam Books, New York.

Touraine, A., 1971. The Post-Industrial Society: Tomorrow's Social History, Classes, Conflicts and Culture in the Programmed Society, first American ed. Random House, New York.

Touraine, A., 1988. Return of the Actor: An Essay in Sociology. University of Minnesota Press, Minneapolis, MN.

Whyte, W., 1956. The Organization Man. Simon and Schuster, New York.

Relevant Website

http://portal.unesco.org/ – "Observatory on the Information Society" at UNESCO.

Postmodern Urbanism

David B Clarke, Swansea University, Swansea, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by M.J. Dear, volume 17, pp. 11856–11860, © 2001, Elsevier Ltd.

Abstract

Beyond signaling a departure from the tenets of modern urbanism, postmodern urbanism remains an essentially contested concept. Shooting to fame in the 1980s and 1990s, it was greeted by many as a new urban logic, forged as modernity's future-oriented temporality yielded to a postmodern concern with spatiality. Yet the postmodern transformation of the city was dismissed as capitalism's new 'cultural clothing' by urban political economists. Paradoxically, the term's premature disappearance may signal its triumph rather than its demise, the situation it was meant to grasp perhaps becoming generalized to the point where the need for specialist terminology is rendered redundant.

Defining the Indefinable

Postmodern urbanism, like the 'postmodern' itself, remains an essentially contested concept. One might superficially define it in terms of an urban form that defies the predictable ordering principles and expected land-use patterns of the modern city (such that Los Angeles is typically cited as the paradigmatic postmodern city, in contrast to that modern urban archetype, Chicago) (Dear, 2000; Soja, 1996). Or one might relate the concept to an architectural aesthetic that playfully subverts the modernist logic of 'form following function' (Ellin, 1996). Or, indeed, one might allude to the kind of 'disorientating' urban environment that results from either or both of the above, noting the iconic status of Los Angeles's Bonaventure Hotel in this regard (Jameson, 1984). Or, again, given that the world of art and letters invariably functions as the vanguard of theory, one might gesture, with Harvey (1989), toward the urban sensibilities captured in a book like Jonathan Raban's 'Soft City' or point to a film such as Ridley Scott's 'Blade Runner' (particularly when the latter is contrasted with the archetypical modern city of Fritz Lang's 'Metropolis'). Such superficial definitions are useful, but only up to a point.

To say that postmodern urbanism is essentially contested is to say that it is impossible to provide a single, universally accepted definition. The term is invoked in innumerable ways, the diversity of meaning arising from the fact that interpretations may vary along at least three different dimensions: first, in relation to the various senses attributable to the term 'postmodern' (principally, whether one accepts that the authority of so-called 'metanarratives' has been undermined (Lyotard, 1984) or not (Harvey, 1989), which effectively amounts to accepting or rejecting the postmodern thesis per se); second, in terms of the way in which modernity and postmodernity are configured in relation to space and time (some align modernity with temporality and postmodernity with spatiality (Soja, 1989), while others reject such formulations); and, third, with regard to how the urban itself is conceived (in broadly sociospatial terms, for instance (Dear, 2000), or primarily in architectural terms, relating to the built environment and urban landscape (Ellin, 1996; Ley and Mills, 1992)). The combinatorial possibilities arising from these differences would be enough to grant 'postmodern urbanism' significant semantic complexity. Add to this the incompatibility of certain combinations, and one can readily appreciate the definitional difficulties it presents.

Whenever definitions of postmodern urbanism are offered, they generally help to clarify the term's meaning, yet cannot but reveal the predilections of those proffering the definitions. Thus, for Dear (2000, 140), the problematic of a distinctively postmodern urbanism "is based on a simple premise: that just as the central tenets of modernist thought have been undermined, its core evacuated and replaced by a rush of competing epistemologies, so too have the traditional logics of earlier urbanisms evaporated, and in the absence of a single new imperative, multiple urban (ir)rationalities are competing to fill the void." As helpful as this may be, when Harvey (1987, 279) refers to postmodern urbanism in relation to the American city as "nothing more than the cultural clothing of late capitalism," it is implicitly understood that the alleged undermining of the 'central tenets of modernist thought' in favor a 'rush of competing epistemologies' is something calling out for explanation, not something to be regarded as an explanation in its own right.

"Discourses of the postmodern are having profound effects on a range of urban practices, from philosophy to planning," wrote Gibson and Watson (1995, 1), in a major collection on the topic released at the height of its influence: "It would seem that, whether we like it or not, the postmodern is here to stay." As it turned out, nothing could have been further from the truth. Despite the intense debate constituting an initial period of engagement with the topic, the term has, by now, been largely abandoned. One should, however, interpret its disappearance with caution. Though the terminology may have run its course and fallen out of favor, this may signal the concept's triumph as much as its demise, the situation that it was meant to grasp having become so generalized as to render the need for specialist terminology redundant.

The City in Ruins

Postmodern urbanism is predicated on the idea of the 'end of modernity.' A direct corollary is that the modern city, one of

modernity's crowning glories, lies in ruins. An immediate note of caution should be sounded here, however, for neither the notion of 'the end' nor the sense of 'ruination' is quite what it seems. Let us take each in turn. The hypothesis of the end of modernity signals something other than a conventionally understood terminal point, of the kind that might be located at the far end of a linear trajectory. Insofar as the notion of linear time, progressing toward some distant (utopian) location in the future, is itself a thoroughly modern one, then the end of modernity would also signal the end of linear time – and with it the end of the illusion of 'the end' itself. This leaves us trapped in the endless repetition of a 'perpetual present' that some call schizophrenic. The end of modernity, therefore, equates to the 'end of history,' inasmuch as history ineluctably belongs to the world view instated by modernity (and not the other way round). The idea of a linear chronology is itself perfectly modern. Holding to it in interpreting the end of modernity would be self-defeating, since it necessarily entails that modernity's uninterrupted progress is still in train, thereby undermining its own reasoning or begging the question (petitio principii).

Similarly disrupting conventional meanings, the idea that the fate of the city after modernity is one of ruination is far from a straightforward allusion to physical decay and decrepitude, even if the rhetorical force of such tropes sometimes finds favor. The dystopian Los Angeles of 2019 envisioned by Blade Runner, for example, was in perfect accord with the postmodern thesis in view of its uncompromising departure from modernist fantasies of a utopian urban future. Yet, in keeping with the contradictory qualities of the postmodern, a sense of ruination may just as easily be derived from precisely the opposite point of view: in terms of the loss of any solid, authentic reality in favor of the superficial glitz or insubstantial froth (Soja, 1996, 247) speaks of 'bubbling, postmodern complexity' associated with the postindustrial city, the consumer society, gentrification, urban boosterism, image-driven place promotion accompanied by a thin veneer of soulless regeneration, and so on. For Marxist urban scholars, especially, such processes stand accused of sanitizing the real, gritty urban environment, replacing it with a fantasy construct, detached from workaday lives of the silent majority, and expressing the privileged vision of a yuppified urban elite (Harvey, 1987). Marxists do not intend to romanticize the alienating dead weight of the modern city, of course, as much as expose the 'castles in the air' of a fake postmodern utopia. However much the 'end of modernity' signaled by the postmodern thesis problematizes the idea of the end, in Marxist circles this involves a surreptitious defense of an indefensible concept, illegitimately sidestepping the need to examine its veracity.

A key facet of the discourse of postmodern urbanism, which made it central to the postmodern debate as a whole, is that a major attempt to fend the notion off was provided by the foremost Marxist urban scholar, David Harvey (1989). Harvey (1987) initially examined postmodernism in relation to the American city, assembling a wealth of evidence to account for the latest round of 'postmodern' urban transformations as expressive of a new capitalist regime of 'flexible accumulation' – a regime emerging from the demise of the preceding Fordist regime of mass production, mass consumption, and Keynesian demand management. Accordingly, for Harvey (1989), there was something suspect in the postmodern thesis propounded by Jean-François Lyotard (1984), insofar as it sought to rule out of court the kind of explanation Harvey offered, suggesting that it had been 'delegitimized' by the very changes he sought to explain. The inherent circularity of the argument is giddying. To invoke another metaphor, it bears all the hallmarks of an intractable debate; a confrontation of the 'irresistible force meets immovable object' variety. Lyotard included Marxism among the modern 'metanarratives' whose legitimacy had been undermined by the changing nature of knowledge. For Harvey, such pronouncements are themselves explicable in terms of the Marxist 'metanarrative,' which unveils them as prime instances of ideological obfuscation: good, old-fashioned false consciousness. Let us take this more slowly, first putting forward Lyotard's case, then returning to Harvey.

Lyotard (1984, xxiv) defined the postmodern as 'incredulity toward metanarratives.' Whereas, once upon a time, stories were simply stories, at the Enlightenment, certain discourses were elevated to previously unknown heights, granting them unprecedented legitimacy on account of their privileged access to the truth. Knowledge, it was claimed, could lead humanity on the road to emancipation (knowledge serving as a means to an end); to a full and final understanding of reality itself (knowledge serving, here, as its own, 'speculative' end). For Lyotard, there is no question that knowledge could ever actually provide access to the real (as promising as the emancipatory rhetoric sounds). One can never dive down beneath language to see directly how well it maps on to reality: metanarratives merely sustain the belief that one can. Metanarratives were a way of modern knowledge legitimating itself, precluding other, potentially competing 'language games' from measuring up against its declarative pronouncements. The fact that this thesis is even possible might suggest a waning faith in the power of modern knowledge. Add to it the ever-growing toll of modernity's unintended consequences, the countless atrocities and injustices committed in its name, the damage done by those claiming to act on the authority of the truth, and the Enlightenment blueprint of progress-through-knowledge begins to lose its former allure. Lyotard (1984) merely ruminated on the fact that knowledge today occupies a wholly different position from the one it enjoyed throughout modernity, where the alignment of knowledge and power had reigned supreme. If modernity solemnized the marriage between power and knowledge – producing a reciprocally reinforcing 'power/knowledge:' knowledge-assisted power coupled with power-assisted knowledge – then postmodernity presides over the divorce. Lyotard (1984) saw nothing to lament in this state of affairs. Modern metanarratives had always gained legitimacy by delegitimizing other, competing narratives ('paralogies', if you will). For Lyotard, a strict adherence to realism and the regulative ideal of true knowledge necessarily did violence to the other. Even emancipatory projects, such as historical materialism, occluded the possibility of sensing the many other injustices that were simply not permitted to appear within a given frame of reference. (Interestingly, Harvey's (1989) defense of the Marxist frame of reference created a backlash from feminist scholars, who saw the potential of

a 'postmodernism of resistance' that might embrace more than class struggle alone.)

Historical materialism, in seeking to shape a better world on the basis of privileged access to the truth of history, meets all of Lyotard's criteria for a metanarrative. One can see why Marxists would demur. A fundamental principle of historical materialism is that social consciousness itself is a reflection of the material mode of production of society, and not the other way round: knowledge forms a part of the ideological superstructure, arising on the base provided by the mode of production. On this understanding, tracing a cultural logic back to its economic determination seems a robust tactic, not least where the capitalist mode of production is concerned. For Harvey (1989), this provides sufficient ammunition to demonstrate that his own favored metanarrative (so-called) remains fully operational: not least in that it is eminently capable of exposing the sort of ideological discourse that would aim to undermine its credibility. Harvey thus turns the tables on Lyotard, putting postmodernism in the dock, precisely where Lyotard had placed Marxism. There is, however, a significant blindness in reactions to the postmodern that saw it as a superficial, 'merely cultural' phenomenon – yet no more so than there is in those who misguidedly regarded the 'postmodern' as marking a complete break with the past, as much 'postcapitalist' as 'post-' anything else. In retrospect, the postmodern debate demonstrates that an attitude of 'business as usual' for Marxist urban theory represented an inadequate response (Dear, 2000), while those seeking to maintain a clean break with modernity – especially evident in attempts to periodize modernity and postmodernity as epochal concepts – were on a hiding to nothing.

Urban Revolutions

Postmodern urbanism juxtaposes a profound discontinuity – the end of modernity – with a marked continuity – the survival of capitalism. Much of the debate rehearsed above was premised on a false choice between the two. In this regard, it is unsurprising that the resources offered by another Marxist, Henri Lefebvre, should be called upon to conceptualize postmodern urbanism, most notably by Ed Soja (1989, 1996, 2000). While some have (not unreasonably) objected to the postmodern appropriation of Lefebvre (Kipfer et al., 2008), it is undeniable that Lefebvre's (2003) work prefigured the tenor, if not the tenets, of postmodern urbanism. Originally published in 1970, that work was forged in the aftermath of the events of May 1968, which exposed the French Communist Party's willingness to subjugate spontaneous revolutionary desire to a self-interested concern with establishing power on its own terms. Lefebvre's titular 'urban revolution' may be read, on the one hand, as the twentieth-century counterpart of the nineteenth-century industrial revolution and, on the other hand, as a recognition of an epistemological revolution accompanying the arrival of a fully urbanized society. "The words 'urban revolution' do not in themselves refer to actions that are violent. Nor do they exclude them" (Lefebvre, 2003, 5).

Urban space has, for Lefebvre, become a crucial locus of difference; of that which is not recoupable to any established order. In Lefebvre's (2003, 37) words, the urban revolution "is constituted by a renewed space-time, a topology that is distinct from agrarian (cyclic and juxtaposing local particularities) and industrial (tending toward homogeneity, toward a rational and planned unity of constraints) space-time. Urban space-time, as soon as we stop defining it in terms of industrial rationality – its project of homogenization – appears as a differential, each moment existing only within a whole, through the contrasts and oppositions that connect it to, and distinguish it from, other places and moments." For Lefebvre (2003, 38), isotopy (a term expressing linguistic redundancy, borrowed for the purposes of 'reading' space), refers to 'everything that makes a place the *same place*.' Heterotopy, in contrast, turns it into a different place, another place: securing its difference by ensuring that it differs from itself. Both these notions apply to real places, rather than to imaginary or symbolic places like utopia or its negative double, dystopia. They signal, however, the inherently contested nature of space, its dialectical charge.

In a subsequent work, originally published in 1974, Lefebvre (1991) elaborated and refined this idea, formulating two conceptual triads to grasp the different 'moments' of the production of space: 'spatial practice,' 'representations of space,' and 'representational spaces,' which may be assimilated to the phenomenological realms of the 'perceived' (*perçu*), the 'conceived' (*conçu*), and the 'lived' (*vécu*). The emphasis placed on production relates to Lefebvre's insistence on space as a concrete abstraction – an abstraction whose truth is forged in practice – and on its ineffably 'open' quality, which derived from its ceaseless production. "The original sense of 'production,' as Baudrillard (1987, 21) reminds us, "is not in fact that of material manufacture; rather, it means to render visible, to cause to appear and be made to appear: *producere*." If space is produced, it is never static, and it is produced differently at different times and by different societies. In an oft-cited passage, Lefebvre (1991) adumbrates the crisis of a specifically modern space – the sort of ordered space, predisposed toward surveillance, that finds its origins in the Quattrocento space of perspectival art. Here, the extent to which Lefebvre's concerns foreshadow postmodern urbanism again becomes apparent:

> The fact is that around 1910 a certain space was shattered. It was the space of common sense, of knowledge (*savoir*), of social practice, of political power, a space thitherto enshrined in everyday discourse, just as in abstract thought, as the environment of and channel for communications; the space, too, of classical perspective and geometry, developed from the Renaissance onwards on the basis of the Greek tradition (Euclid, logic) and bodied forth in Western art and philosophy, as in the form of the city and town. … Euclidean and perspectivist space have disappeared as systems of reference, along with the other former 'commonplaces' such as the town, history, paternity, the tonal system in music, traditional morality, and so forth. This was truly a crucial moment (Lefebvre, 1991, 25).

When Lefebvre's erstwhile assistant at Nanterre, Jean Baudrillard (1987, 126) – later dubbed the 'high-priest' of postmodernity – declared that "There is no more system of reference to tell us what happened to the geography of things," the resonance is clear. Their views of the world, however, are far from consonant. While Lefebvre is concerned

to trace the contours of a spatial dialectic, for Baudrillard (1990a, 7), "the world is not dialectical, it is sworn to extremes, not to equilibrium; sworn to radical antagonism, not to reconciliation or synthesis."

The incautious treatment of such resonances will, as we shall see, prove deleterious to framing postmodern urbanism as a meaningful discourse. Bearing this in mind, it is noteworthy that Lefebvre's (2003) articulation of 'differential' space in terms of the 'heterotopic' chimes with Foucault's (1986, 25) 'heterotopia,' the topic of a 1967 lecture. Heterotopias – Foucault's examples include cemeteries, theaters, brothels, colonies, ships, hotels, and gardens – are "capable of juxtaposing in a single real place several spaces, several sites that are in themselves incompatible" (Foucault, 1986, 25). Foucault's opening remarks, moreover, bear an uncanny resemblance to the thesis of The Urban Revolution:

> The great obsession of the nineteenth century was, as we know, history: with its themes of development and of suspension, of crisis and cycle, themes of the ever-accumulating past, with its great preponderance of dead men and the menacing glaciation of the world. … The present epoch will perhaps be above all the epoch of space. We are in the epoch of simultaneity: we are in the epoch of juxtaposition, the epoch of the near and far, of the side-by-side, of the dispersed (Foucault, 1986, 22).

The proximity of Foucault's and Lefebvre's formulations should not, however, blind us to their differences. Foucault's concern to relativize history is, somewhat impishly, turned by Lefebvre (2003, 191) into a symptom of the 'urban revolution' (just as Foucault might see Marx as symptomatic of the industrial revolution): "Using borrowed concepts and terms we can say that the urban … rises above the horizon, slowly occupies an epistemological field, and becomes the episteme of an epoch."

> During this new period, what once passed as absolute has become relativized: reason, history, the state, mankind. We express this by saying that those entities, those fetishes, have died. There is something true in this claim, but fetishes do not all die the same death. The death of 'man' only affects our philosophers. … Reflective thought allows itself to be captivated by these dramas and turns its gaze from the field before it, which remains blind. To explore this field, to see it, change is necessary, the abandonment of earlier viewpoints and perspectives. During this new period, *differences* are known and recognized, mastered, conceived and signified (Lefebvre, 2003, 36–37).

Foucault is implicitly castigated for scotomization, for a misdirected, 'reflective' thought, out of step with a society in the throes of urban revolution – streetwise and attuned to difference – that is already detecting unfamiliar shapes emerging in a visual field from which the philosophers avert their gaze. Could Lefebvre's work amount to a postmodern urbanism avant la lettre? This implication is explored in an ambitious exegesis of postmodern urbanism presented in a trilogy of studies by Soja (1989, 1996, 2000).

Soja's dazzling, eclectic performances are justly famous for, among other things, juxtaposing Foucault and Lefebvre. A good deal of fancy footwork, however, makes it extremely difficult to keep one's eye on the ball, and their success belies some pervasive problems. Soja privileges difference to the extent that anything goes, including substituting the dialectic with the neologism 'trialectic' – the result, unfortunately, being equally unsatisfactory to dialecticians and postmodernists. "Noting the recurrent use of conceptual triads in Lefebvre, Soja reifies them into 'trialectics' (sic) of ontology (spatiality, historicity, sociality) and epistemology (first, second, third space)," observe Kipfer et al. (2008, 20): "Lefebvre's triad of the production of space is transformed into an 'epistemological trialectics' of objectivist (first), subjectivist (second), and lived (third) space." For Soja (1996, 56–57), "*Everything* comes together in Thirdspace: subjectivity and objectivity, the abstract and the concrete, the real and the imagined, the knowable and the unimaginable…" and so on. Establishing itself on an etymological simplification invoking the binary, however, Soja's remedial 'trialectic' manages to miss the complex senses of 'across,' 'through,' 'throughout,' and 'complete' possessed by the prefix 'dia-' (one might note, analogously, that 'dialogue' need not be between two people). In so doing, he neglects the significance of space being a concrete abstraction: "for Lefebvre, the third term in a triad is not an expression of trialectics but a product of three-dimensional dialectical transformation" (Kipfer et al., 2008, 20). Indeed, 'de-' and 'dis-' are the favored postmodern prefixes, implying subtraction and disjunction rather than addition (+1) and conjunction, yet Soja's (1996) preferred tactic is to adjoin everything in a 'real-and-imagined' space that leaves intact what would otherwise spontaneously deconstruct ('real' and 'imaginary' are left unscathed by hyphenation). As belated as the lesson may be, Lefebvre does not provide a template for postmodern urbanism.

Lost in Hyperspace

Were one forced to pick a single term to sum up postmodern urbanism, 'hyperspace,' with its sci-fi overtones, would admirably fit the bill. It was pressed into service by Fredric Jameson (1984), for whom postmodernism represents the 'cultural logic' of late capitalism – not merely its 'cultural clothing' but a fully integral logic, as if base and superstructure were fused by an unforeseen short circuit. Architecture is identified as the privileged aesthetic language expressing 'the bewildering new world space of late or multinational capital,' with certain examples, of which downtown LA's Bonaventure Hotel looms large, engendering a vertiginous experience that announces this 'latest mutation in space – postmodern hyperspace' (Jameson, 1984, 54, 83). Jameson's essay has received enormous critical attention. Yet despite the animus it has often attracted, it is not unreasonable that Jameson should have seen the labyrinthine layout of the Bonaventure Hotel as an apt metaphor for the disorientation induced by the epochal convulsions of postmodernity. Reminiscent of Baudrillard's reading of the Beaubourg (or Pompidou Centre) in Paris, 'hyperspace' resonates with Baudrillard's (1994) 'hyperreality' (the 'hyperreal' being 'more real than real'). However, while Davis (1985, 106) suggests that, for Jameson, the Bonaventure represents an 'impossible representation' of hyperspace – the implication being that representation is defeated by this latest

spatial mutation; that the real now resists representation and the territory can no longer be adequately mapped – Jameson's claim that the postmodern involves the transformation of reality into images is scarcely Baudrillardian. It is not that Baudrillard holds to the opposite hypothesis: that images have been transformed into reality. In fact, both hypotheses might well apply, but – Baudrillard's point – they would retroactively scramble the very terms in which they are framed. Unlike Soja's (1996) syncretic 'real-and-imagined,' for Baudrillard (1993, 72), "The hyperreal effaces the contradiction of the real and the imaginary." The terms 'reality' and 'imaginary' still have currency, but their meanings have been ineluctably altered and their indistinction assured. If postmodernity produces 'unreal' urban environments, then, it must be recognized that this 'unreality no longer implies the imaginary' (Baudrillard, 1990b, 30). 'Simulacrum' and 'simulation' are the apposite terms for describing this altered state (Baudrillard, 1994).

If the heterotopic qualities of hyperspace resist representation, this outcome was written into modernity from the off. In a fragmentary text by Borges, Baudrillard (1994) discovers an apt allegory for the project of modernity in the insane enterprise of a 1:1 scale map, which once covered the entire territory of a distant empire. Discontented with the world as it found it, modernity determined to make the world over in its own image. Not content with merely mapping the world, modernity resolved to change it wherever it failed to measure up to expectations: to smooth out the untidy crumples in reality; to remove the unsightly blemishes, straighten out unnecessary deviations, relocate whatever was out of place, correct any offending irrationalities: in other words, to impose order on an unruly world. In this, modernity pulled off a remarkable feat. It presented its own imaginary as the 'reality principle,' casting the world itself as its inadequate representative. In its failure to live up to the principle, reality was found wanting. Reality, therefore, was never quite what it seemed. In a classically modern move, the 'real' was forged in opposition to the 'illusory,' the 'imaginary,' the merely 'apparent.' If flighty, fickle appearances are deceptive, then solid, steadfast reality is trustworthy. What its opposing terms lack, reality possesses. Modernity's ordering zeal set out to eradicate the enigmatic, ambivalent qualities of the premodern world. Accordingly, "The true revolution of the nineteenth century, of modernity, is the radical destruction of appearances, the disenchantment of the world and its abandonment to the violence of interpretation and history" (Baudrillard, 1994, 160): not the industrial revolution, nor any other revolution, but "the immense process of the destruction of appearances (and the seduction of appearances) in the name of meaning (representation, history, criticism, etc.)." This process is, however, hoist by its own petard. The closer one gets to a world in which appearances are done away with, that world succumbs to a fatal reversibility. As appearances are vanquished, the real loses all sense of reality: not as the result of a deficit or lack of reality but in consequence of a surplus or excess: hyperreality. As with compulsory reality, so too with compulsory meaning: an excess of meaning is no longer meaningful. When something means everything, it means nothing: "by giving you a *little too much* one takes away everything" (Baudrillard, 1990b, 30). Modernity, then, is diverted from its course – not in spite of, but precisely because of, its own best intentions. Its furtive destiny lies elsewhere. Little wonder, then, that Baudrillard (1994, 160–161) should refer to the "second revolution, that of the twentieth century, that of postmodernity, which is the immense process of the destruction of meaning, equal to the earlier destruction of appearances."

From the off, modernity was on course for utopia. Even as it charted its progress, however, it went astray, "crossing into a space whose curvature is no longer that of the real" (Baudrillard, 1993, 2). The upshot is that postmodern urbanism needs to be understood, not in terms of the production of space, but in terms of its seduction – from "*se-ducere*: to take aside, to divert from one's path" (Baudrillard, 1990b, 22). Modernity inaugurated a new conception of time. In the wake of the 'querelle' of Ancients and Moderns in the 1690s, Christian Europe's time reckoning was no longer "organized around a fixed point in the slowly receding past," but was set in motion, "attaching it firmly to Europe's own thrust toward colonizing the future in the same way as it had colonized the surrounding space" (Bauman, 1987, 110). "Modernity," Bauman (1997, 86) avers, "furnished time with certain traits which only space possesses 'naturally': modern time had direction, just like any itinerary in space." Modernity was a journey; destination – Utopia. Hence the fact that 'progress' should become modernity's watchword. A future-oriented temporality, geared toward what might yet be, necessarily casts the present as inadequate. Utopia was destined to lie forever in the future, beyond the infinitely receding horizon. The secret curvature in modernity's linear trajectory, however, changes everything. The 'end of modernity' does not signal that utopia failed to appear, that society grew impatient with the arduous journey that never reached its end – but, rather, that it failed to appear in the expected location. At a certain point in time, the realization took hold that utopia was already behind us. The end of modernity signals the abandonment of the pursuit of utopia because utopia has already been achieved, no matter that it failed to materialize in the manner anticipated. An achieved utopia retrospectively reveals that utopia was never what it purported to be. Ultimately, therefore, the postmodern means not 'reaching the end,' but 'having passed beyond the end.' And beyond the end, it becomes apparent that the end no longer means anything – and never meant anything in the first place.

One should not be surprised to discover then that, beyond the end, postmodern urbanism does not look entirely dissimilar to modern urbanism. Postmodern urbanism implies in a state of hysteresis: a state defined by the continuation of the effect ('reality effects') in the absence of the cause (the 'destruction of appearances' set in sway by modernity). Insofar as it retroactively cancels out its distinction from what went before, the postmodern cannot be taken in a periodizing sense. That distinction, like the torsion in a möbius strip, occupies an indeterminate location. Only a cut localizes it, at which point it disappears, "without the reversibility of surfaces being resolved" (Baudrillard, 1994, 18). Hence Lyotard's (1984, 81) quip that the "*Post modern* would have to be understood according to the paradox of the future (*post*) anterior (*modo*)."

Zygmunt Bauman – once dubbed the 'prophet of postmodernity' – concurs:

> The 'after' bit which the concept of 'postmodernity' entailed seemed to me suspicious from the start. … It seemed to me that the 'post-modern perspective' which allowed the scrutiny of modernity's failures and the debunking of many of its undertakings as blind alleys, far from being in opposition to modernity or growing on its grave, had from the start been its indispensable *alter ego*: that restless, perpetually dissentful voice that enabled modernity to succeed in its critical engagement with found reality and the many realities sedimented by that engagement. I liked Lyotard's quip: one cannot be truly modern without first being postmodern (Bauman and Tester, 2001, 74–75).

The same logic pervades postmodern urbanism, not least in the verisimilitude of its architecture. In a manner apposite to the end of history, Koolhaas's experimental notion of Junkspace "means no more than perpetual renovation … perpetual recycling," according to Jameson (2003, 74), who cities Koolhaas to this effect: "Restore, rearrange, reassemble, revamp, renovate, revise, recover, redesign, return … redo, respect, rent: verbs that start with *re-* produce Junkspace." One could just as readily cite Tschumi (1994, 225): "dis-integrated, dis-located, dis-juncted, deconstructed, dismantled, disassociated, discontinuous, deregulated … de-, dis-. … These are the prefixes of today." There is no longer any meaningful gap between the 'de-' and the 're-': no interval capable of conferring on either a stable identity. Like characters in a cartoon caper, relentlessly chasing each other round in circles, repeatedly swapping (and ultimately losing) their identities as 'pursuer' and 'pursued,' everything loses its meaning in the vertiginous cycle of repetition. Postmodern urbanism, properly understood, defines an 'ecstatic space' (from 'ex-stasis' – to put out of place): "Ecstasy is the quality proper to any body that spins until all sense is lost, and then shines forth in its pure and empty form" (Baudrillard, 1990a, 9). To put it another way, 'postmodern urbanism' is the technical term for the city in ruins: "Hornstrumpot! We shall not have succeeded in demolishing everything unless we demolish the ruins as well. But the only way I can see of doing that is to use them to put up a lot of fine, well-designed buildings" (Jarry, 2002, 106).

See also: Cities: Capital, Global, and World; Consumption, Geography of; Cultural Geography; Foucault, Michel (1926–84); Marxist Geography; Space and Social Theory in Geography; Spatial Thinking in the Social Sciences; The Right to the City; Time-Space in Geography; Urban Design and Public Space; Urban Entrepreneurialism; Urban Geography; Urban Neoliberalism; Urban Studies: Overview.

Bibliography

Baudrillard, J., 1987. Forget Foucault. Semiotext(e), New York.
Baudrillard, J., 1990a. Fatal Strategies. Pluto, London.
Baudrillard, J., 1990b. Seduction. Macmillan, London.
Baudrillard, J., 1993. Symbolic Exchange and Death. Sage, London.
Baudrillard, J., 1994. Simulacra and Simulation. University of Michigan Press, Ann Arbor, MI.
Bauman, Z., 1987. Legislators and Interpreters: On Modernity, Post-modernity and Intellectuals. Polity, Cambridge.
Bauman, Z., 1997. Postmodernity and Its Discontents. Polity, Cambridge.
Bauman, Z., Tester, K., 2001. Conversations with Zygmunt Bauman. Polity, Cambridge.
Davis, M., 1985. Urban renaissance and the spirit of postmodernism. New Left Review 151, 106–113.
Dear, M.J., 2000. The Postmodern Urban Condition. Blackwell, Oxford.
Ellin, N., 1996. Postmodern Urbanism. Blackwell, Oxford.
Foucault, M., 1986. Of other spaces. Diacritics 16, 22–27.
Gibson, K., Watson, S., 1995. Postmodern spaces, cities and politics. In: Watson, S., Gibson, K. (Eds.), Postmodern Cities and Spaces. Blackwell, Oxford, pp. 1–10.
Harvey, D., 1987. Flexible accumulation through urbanization: reflections on post-modernism in the American city. Antipode 19, 260–286.
Harvey, D., 1989. The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change. Blackwell, Oxford.
Jameson, F., 1984. Postmodernism, or the cultural logic of late capitalism. New Left Review 146, 53–92.
Jameson, F., 2003. Future city. New Left Review 21, 65–79.
Jarry, A., 2002. The Ubu Plays: Ubu Rex, Ubu Cuckolded, Ubu Enchained. Methuen, London.
Kipfer, S., Goonewardena, K., Schmid, C., Milgrom, R., 2008. On the production of Henri Lefebvre. In: Goonewardena, K., Kipfer, S., Milgrom, R., Schmid, C. (Eds.), Space, Difference, Everyday Life: Reading Henri Lefebvre. Routledge, London, pp. 1–23.
Lefebvre, H., 1991. The Production of Space. Blackwell, Oxford.
Lefebvre, H., 2003. The Urban Revolution. Minnesota, Minneapolis.
Ley, D., Mills, C., 1992. Can there be a postmodernism of resistance in the built environment? In: Knox, P. (Ed.), The Restless Urban Landscape. Prentice Hall, Englewood Cliffs, NJ, pp. 255–278.
Lyotard, J.-F., 1984. The Postmodern Condition: A Report on Knowledge. Manchester University Press, Manchester.
Soja, E.W., 1989. Postmodern Geographies: The Reassertion of Space in Critical Social Theory. Verso, London.
Soja, E.W., 1996. Thirdspace: Journeys to Los Angeles and Other Real-and-Imagined Places. Blackwell, Oxford.
Soja, E.W., 2000. Postmetropolis: Critical Studies of Cities and Regions. Blackwell, Oxford.
Tschumi, B., 1994. Architecture and Disjunction. MIT Press, London.

Postmodernism in Sociology

Alex Preda, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the intellectual roots of postmodernism in sociology and its evolution over the past three decades. It briefly examines the main areas of research developed as a response to the emergence of new media and communication technologies since the late 1990s. It defines postmodernism as an intellectual project developed since the 1970s mainly within philosophy and the humanities, which has been adopted, adapted, and enriched within sociology as a response to the theoretical and empirical challenges raised by the cultural features of contemporary developed societies, including here the advent of communication technologies such as the Internet. The article distinguishes between two aspects of postmodernism in sociology: the critique of theoretical assumptions considered as modernist, on the one hand, and the empirical investigation of cultural phenomena taken to be characteristic for postmodern societies, on the other hand. Empirical postmodern research has evolved from an initial dominating focus on consumption and consumerism to the investigation of 'digital cultures,' understood as communication formats such as electronic socialization platforms, instant messaging, or online gaming.

Postmodernism in sociology is an intellectual project with a controversial, yet established legacy, characteristic for the late twentieth century. This project comprises (1) an analysis of the social and cultural features of late capitalism (postmodernity), (2) a critique of sociological theory as a modernist project, (3) an extension of sociological inquiry into new domains, and (4) new forms of sociological expression. The term postmodernism made its way into sociology from literary theory and criticism in the 1970s; at the same time, its adoption continued and reformulated a series of important sociological topics. The key concepts of sociological postmodernism are subject, identity, text, and symbol. On these grounds, postmodernity is characterized as a form of social order in which the electronic media play a prominent role, symbolic codes are pervasive, and social identities are fragmented. These terms are also employed in the critique of sociological theory as a product of modernity. According to this critique, scientific theories work as symbol systems enabling certain courses of social action, and the knowing subject is the product of social circumstances. Sociological postmodernism claims a structural similarity between texts and social order: both are network-like systems of symbols. On the methodological side, sociological postmodernism emphasizes qualitative research methods, especially discourse and visual analysis. Writing-related issues also occupy an important place.

Prominent fields of postmodernist research have been the Internet, work and organizations, political action, science and technology, commodification, consumption, gender, media, and popular culture. Classical sociological topics like symbolic consumption have been further explored and enriched. During the 2000s, postmodernist research has oriented more and more toward investigating new media formats, generically known as digital cultures or as 'virtual worlds.'

Sociological postmodernism appears as a form of sociological reflection, a kind of sociological sensibility, and a specific take on a cluster of social phenomena. Its varieties share a series of theoretical and methodological premises, but differ in their conclusions and research programs. Core common elements are (1) avoiding recourse to a set of universally valid assumptions as theoretical and methodological foundations, together with (2) the key role ascribed to notions like subject, identity, text, and symbol in the analysis of society.

Origins of Postmodernism

The term 'postmodernism' has been used in Latin American literary criticism since the 1930s, and in Anglo-American debates since the 1940s, in order to designate new forms of expression in their relationship with the aesthetic of modernism. In history, the term was used by the British historian Arnold Toynbee in *A Study of History* in 1947 and designated the latest phase of Western civilization. In sociology, it was introduced by Amitai Etzioni's book, *The Active Society* in 1968. Starting with the early 1970s, the term appeared more and more frequently in sociological texts in France, North America, and the United Kingdom. At that time, it was already common stock in literary theory and criticism. An essay written by the French philosopher Jean-François Lyotard, *The Postmodern Condition: A Report on Knowledge* (1979) played a seminal role in making this term widely known and used in the social sciences. Therefore, postmodernism was partially a conceptual import; its diffusion in the 1970s was part of a wider innovative movement in the discipline, marked, among others, by the rise of social constructivism and feminism. At the same time, it continued and reformulated a series of topics already present in sociology.

The intellectual roots of sociological postmodernism can be identified in the works of some key nineteenth and early twentieth century philosophers, sociologists, and linguists. Karl Marx's critique of capitalism, Georg Simmel's analysis of modernization processes, Friedrich Nietzsche's critique of epistemology, and Ferdinand de Saussure's theory of language play here a prominent role. The French structuralist movement (Claude Lévi-Strauss's anthropology, Roland Barthes's semiotic theory, Jacques Lacan's psychoanalysis) and the post-structuralist one (Michel Foucault, Jacques Derrida), Ludwig Wittgenstein's later works, and Clifford Geertz's anthropology have also played a considerable role. Among other significant influences are the Frankfurt School (Walter Benjamin, Theodor

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32111-0

Adorno, Max Horkheimer, and Herbert Marcuse) and symbolic interactionism.

The Tenets of Postmodernism

There are two main tenets, mirroring the distinction between postmodernism and postmodernity: while the latter designates the social and cultural features of late capitalism, distinguishing them from earlier phases (modernity), the former designates a theoretical and methodological program, as distinct from earlier (modernist) ones.

The Analysis of Postmodernity

The analysis of postmodernity developed at the beginning of the 1970s, continuing a preoccupation already found in the works of Georg Simmel, Thorstein Veblen, and the Frankfurt School. Postmodernity is a late phase in the evolution of the capitalist system, seen either as a continuation and intensification of modernity, or as a substantially new stage. It is characterized by the transition from manufacture to a decentralized services economy; science and complex technological systems play a key role here. Decentralization, together with the pervasiveness of expert and technological systems in all domains of everyday life, leads to a fragmentation and diversification of social identities. The electronic media (television, computers, video) are prominent in the constitution and presentation of social reality; visual and linguistic symbols become pervasive, while popular cultural forms diversify and expand. Society is therefore fragmented and heterogeneous; social order is always local and contextual, and works as an assemblage of symbolic codes. Production and consumption, social stratification, as well as personal and collective identities become dependent on such codes. They are understood as a support, a medium for, and an outcome of social action. This notion becomes a general explanatory paradigm: every domain of societal organization can be seen as a symbolic system and analyzed accordingly.

The emergence, spread, and popularity of electronic media, particularly of socialization platforms and of multi-user games, are accompanied by forms of social interaction and activities that do not entirely overlap in their structure with the face-to-face interactions prevalent in everyday life (Boellstorff, 2008).

Consequently, sociology (an outcome of modernity itself) should take fragmentation, decentralization, relativization, and the mediated character of social reality into account; it should adapt its own tenets and claims as well as its own modes of analysis and representation to these changes. Starting from these premises, Zygmunt Bauman (1992) has argued that notions like sociality, habitat, self-constitution, and body are better suited for the analysis of contemporary social processes than the traditional concepts of society and order. The increasingly dominant focus on virtual worlds' or 'digital cultures' (Miller, 2011) has been accompanied by claims that fragmentation, disconnectedness of the parts, and to some extent disembodiment (or dematerialization, meaning the dominance of symbols over the material features of social life) characterize social reality.

The Critique of Modernist Sociology

During the 1980s, a second theme made its way from philosophy and literary theory into sociological debates. It was a critique of the status of sociological theory as a modernist product, and it drafted a new methodological program. This debate raised a series of questions about the validity claims of scientific theories, the nature of scientific truth, and the status of the knowing subject, arguing that (1) all knowledge is contextual and local; (2) the validity claims of any scientific theory are not to be found in some abstract, universal criteria, but are rather the results of either negotiated consensus or power struggles; and (3) as a consequence, the knowing subject does not dispose of universal criteria to ascertain the validity and truth of his or her knowledge. The subject should not be seen as autonomous with respect to an objective, given world, but rather as the product of social circumstances. Since knowledge has a contextual and local character, scientific theories are reinterpreted according to local conditions: they are open-ended and cannot be controlled by their authors. Scientific theories are not determined by external validity and truth criteria; they work as systems of symbols enabling certain courses of social action and promoting a certain worldview. On these grounds, there should be a structural similarity between texts and social order: both are locally determined, network-like, open-ended systems of symbols requiring certain kinds of action from the actors embedded in them.

Notions like text and performance should occupy a prominent place in the analysis of social phenomena. Sociology should focus on investigating and describing how symbol systems work in particular social contexts, instead of producing all-encompassing theories. A corollary of these arguments is that sociological theories should acknowledge the fragmentation and diversity inherent in the local character of knowledge; they should explore their consequences for the world we live in, and should experiment with new modes of representation – like poetry, or dialogical forms. These epistemological claims overlap with those of the first strand of postmodernist sociology. However, while the first strand derives the fragmentation, relativization, and diversification of knowledge from social changes, the second one takes them as stemming out of a number of epistemological, universally valid first principles.

At the end of the 1980s and in the early 1990s, such metatheoretical topics were intensely debated in sociology and, for a short time, even eclipsed the first strand of postmodernism. However, the acceleration of globalization processes and the advent of new communication technologies has brought social analysis to the forefront again.

The Methodology of Postmodernist Sociology

One important methodological aspect is the antipositivist and antiverificationist stance: there is no totalizing explanatory model and no universally valid verification procedure. Postmodernist sociology stresses the role of qualitative methods in the analysis of contemporary social phenomena, as a consequence of the prominence of visual and linguistic symbols. While some authors (Richardson, 1997) emphasize interpretation and subjectivity, others (Lemert, 1997) argue that methodology

must preserve objectivity and logical consistency. There has been an increased sociological interest in discourse analysis, understood as a cluster of techniques for grasping the emergence, functioning, and context-boundedness of symbolic codes. More recently, discourse analysis has been complemented by a heightened interest in the analysis of visual assemblages such as the imagery of video games or of Internet videos. The prominence and sometimes prevalence of images over text in electronic communication formats have been accompanied not only by the growth of visual studies (Elkins, 2003), but also by a marked interest for visual modes of sociological expression, such as film. The emphasis placed on modes of expression includes writing processes and images as sources of conceptual enrichment in sociology. Poetry, drama, or film is promoted as a legitimate complement to the established, academic expository modes. Hybrid forms of sociological expression have appeared too – for example, in works combining academically written texts, poetry, and artwork (Haraway, 1997).

Fields of Postmodernist Research

Postmodernist analyses have been prominent in domains like work and organizations, political action, science and technology, commodification and consumption, gender, media, and popular culture. George Ritzer (1993), for example, examined the new forms of work that emerge in the knowledge-based, global economy: they are characterized by fragmentation, local specialization, flexibility, and mobility, on the one hand, and by global dispersion and typification, on the other. Under these circumstances, the temporal and spatial conditions of work change substantially: while the lifespan of products shortens, they need to be replaced faster, and worktime expands. Tensions arise between requirements for flexibility and mobility, on the one hand, and typification, on the other. New directions in commodification and consumption complete this picture. Commodification is understood as the relentless expansion of market exchange, which reshapes social relationships according to its own logic; the pendant of this process is the increased social significance of the act of consumption. Jean Baudrillard (1972), for example, has argued that symbolic consumption becomes one of the main social mechanisms through which individual and collective identities, as well as social relationships are defined and reproduced.

In political sociology, the analysis has focused on political discourses and symbolisms, and on the spectacular form taken by political events. Another direction has been to examine the strains put on democracy by the media-induced forms of political action, and by the globalized economy. Authors like Chantal Mouffe (1992) have asked whether the new economic and political conditions do not require devising new forms of democratic participation and of political action.

With knowledge lying at the core of the postmodern society, science has been a domain of sustained investigation. The focus has been on the changes brought about by the new information and biomedical technologies in social life, as well as on the role and effects of complex technological systems. Bruno Latour (1991) and Donna Haraway (1997), among others, have argued that such systems are so pervasive and deep-reaching and that social life is dependent on them to such an extent that new forms of social symbiosis have emerged; these are described with the help of notions like the cyborg society, hybrids, or networks of human and nonhuman actors. These authors have called for a revision of the conceptual apparatus of sociology on the grounds of the relationships between human actors and technical systems.

A further domain of investigation, situated at the crossroads between feminist and postmodernist sociology, has been that of gender and sexuality. Postmodern diversification brings forth a fragmentation and multiplicity of gender and sexual identities. The distinctions between genders become blurred; some forms of sexual identity formerly considered marginal gain a new significance, and new sexual identities emerge (e.g., transsexuality). Owing to the local and contextual character of knowledge, the gender specificity of knowledge and experience has to be taken into account. Several authors (e.g., Smith, 1990) have explored the ways and the implications of gender-specific knowledge in the workplace, the field of politics, popular culture, and the public sphere.

One of the key domains in which the impact of postmodernist sociology has been felt is that of the electronic media and communication. The popularity of Internet socialization platforms, and also of multiuser games and instant messaging, has brought to the fore the issue of identity formation in various interaction formats (Castronova, 2005). Online identities can be tied to features different from those of everyday life identities; they can be less stable and more prone to variability. Online communication structures (such as the allocation of speaking rights) are different from their offline counterparts. This has raised the issue whether virtual worlds change the ethical underpinnings of social interactions. Finally, postmodernist sociology has also considerably influenced the subfields of cultural and urban studies, reinvigorating a research tradition that goes back at least to the Chicago School of sociology.

The State of Postmodernist Research

Postmodernist sociology has been around for approximately four decades; while controversial and contested for some of the more radical theoretical stances, many postmodernist ideas have made inroads into mainstream sociology. At the same time, it has continued and enriched a series of themes already present in classical sociology – like the sociology of consumption or the study of popular cultures. It has also pioneered some new domains of research, such as the study of Internet cultures, of the media, or (together with ethnomethodology) that of the impact of computers. Together with feminist sociology, it has had a considerable influence on the sociology of gender and sexuality. Its empirical, domain-oriented impact has been more significant than that of its pure theoretical tenets.

New Research Directions

The cultural forms induced by the electronic media (e.g., the culture of software programmers, of Internet users) have been a major focus of study since the 2000s. The changes in consumption patterns and market relationships brought about

by the Internet, the rapid expansion of online communications, or the Internet's role in the globalization processes have taken center stage in empirical investigations. As electronic media continue to expand and diversify in often surprising ways, they will also continue to provide a major challenge to postmodernist research.

See also: Consumption, Sociology of; Cultural Relativism, Anthropology of; Globalization, Anthropology of; Modernity; Postmodern Urbanism; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Relativism: Philosophical Aspects; Social Identity in Social Psychology; Sociology, Epistemology of.

Bibliography

Baudrillard, J., 1972. Pour Une Critique de l'Économie Politique du Signe. (For a Critique of the Political Economy of the Sign.) Gallimard, Paris.
Bauman, Z., 1992. Intimations of Postmodernity. Routledge, London.
Boellstorff, T., 2008. Coming of Age in Second Life: An Anthropologist Explores the Virtually Human. Princeton University Press, Princeton, NJ.
Castronova, E., 2005. Synthetic Worlds: The Business and Culture of Online Games. The University of Chicago Press, Chicago, IL.
Elkins, J., 2003. Visual Studies: A Skeptical Introduction. Routledge, New York.
Etzioni, A., 1968. The Active Society: A Theory of Societal and Political Processes. Collier-Macmillan, London.
Haraway, D., 1997. Modest Witness @ Second Millenium: FemaleMan® Meets OncoMouse™; Feminism and Technoscience. Routledge, New York.
Latour, B., 1991. Nous n'avons jamais été modernes. (We Have Never Been Modern.) La Découverte, Paris.
Lemert, C., 1997. Postmodernism Is Not What You Think. Blackwell, Cambridge, MA.
Lyotard, J.F., 1979. La Condition Postmoderne: Rapport sur le Savoir. (The Postmodern Condition: A Report on Knowledge.) Editions de Minuit, Paris.
Miller, V., 2011. Understanding Digital Culture. Sage, London.
Mouffe, C., 1992. The Return of the Political. Verso, London.
Richardson, L., 1997. Fields of Play. Rutgers University Press, New Brunswick, NJ.
Ritzer, G., 1993. The McDonaldization of Society: An Investigation into the Changing Character of Contemporary Social Life. Pine Forge Press, Newbury Park, CA.
Smith, D.E., 1990. Texts, Facts, and Femininity: Exploring the Relations of Ruling. Routledge, London.
Toynbee, A., 1947. A Study of History. Oxford University Press, London.

Postmodernism: Methodology

Pauline V Rosenau, University of Texas, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11868–11872, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Postmodernism is considered here to be a methodology. However, because postmodern social science is based on intuition, emotion, faith, subjective judgment, imagination, and introspection as methods, several methodologies exist. Accordingly, this article describes postmodern methodology in opposition to quantitative and qualitative methods and, in general, to any method seeking truth and generality in any way. It emphasizes interpretation as introspective, intersubjective, and antiobjectivist.

If methodology is defined as referring to the foundations of knowledge, as how we perceive and understand reality, as well as how we study it, then there is not only a single implicit postmodern methodology. There are several. But all challenge the methodological assumptions associated with rigorous, modern social science inquiry, be it qualitative or quantitative (Dickens and Fontana, 1994). Some postmodernists look to deconstruction as a method. Others employ introspective, intersubjective, and antiobjectivist interpretation. Still others look to wisdom, lived experience, storytelling, empathy, introspection, imagination, subjective judgment, and faith, as method. A few reject methodology altogether, constructing an intellectual framework within which there is no need to bother about it.

Indeed, there are serious philosophical differences between postmodernism and the enlightenment norms of traditional science. Some postmodernists are uninterested in generalization, definitive explanation, replication, validity, reliability, etc. Neither prediction nor theory building is a major concern. They look to deconstruction as method. Deconstruction tears a text apart, revealing its contradictions and assumptions. Other postmodernists criticize modern methodology and seek to revise it, to build upon it. They look for alternatives to what they call the old, past, closed conventional social science methods inherited from the natural sciences. But they do so without entirely rejecting reason, consistency, and coherence of argument. Although they abandon foundations, they retain what they call preferences.

Many postmodernists see the need for new standards for evaluating knowledge. Their criteria are likely to be subjective in nature, including, for example, flexibility, sensitivity, and interactivity. They look for beauty, strength, or force in a text. Some value elegance of expression and style, and seductiveness of content while others say that such criteria are community specific and so cannot be enumerated. Still other postmodernists willingly employ more direct valuations, such as negation of oppression, exploitation, and domination, or the affirmation of liberation, freedom, insubordination, and resistance as key standards (Rosenau, 1992).

A Postmodern Methodological Focus

Postmodern methodology focuses on the margins and postmodern social scientists highlight the unique and the unusual. They concentrate on the enigmatic and appreciate the unrepeatable. They seek to multiply paradox, to invent ever more elaborate repertoires of questions as much as to determine answers. They look to 'différance' (Derrida, 1972) rather than sameness, regularity, and synthesis. 'Différance' is a structuring principle that tells us that definition rests not on some inherent characteristics of a thing but rather in terms of its negative and positive references to other texts. It also suggests that in the end, definition is never final, always postponed, deferred forever (Derrida, 1981).

Postmodern methodology seems driven by substantive concerns quite different from those modern social scientists focus upon (Cullenberg et al., 2001). Postmodernists direct their attention to everyday life and they gravitate toward mini-narratives or stories. Postmodern social science is a 'systematic commentary on the knowledge of daily life' (Bauman, 1990). This means examining what is assumed to be quite ordinary and attributing a new importance to it. Postmodern social science makes room for the unexpected, the mystical, and the metaphysical. Concepts such as fate and spirituality are, once again, legitimate (Griffin, 1988).

Linking Substance and Method

If there is a close association of substance and method, then conventional social science methods are not helpful given the postmodern focus on the margins, the unexpected, the paradoxical, and the unrepeatable. Empirical methods require repeatability so that careful measurement can be carried out and statistical norms met. But the subject matter of interest to postmodernists does not recur necessarily and its very interest may be in the fact that it is transient, constantly in flux, and difficult to capture. Because of their substantive foci, these postmodernists require methods that are open and flexible, rather than highly structured, codified, and technically detailed. They must be adaptable to a great variety of situations because of the range of potential application. There is no easy methodological recipe for doing this type of research.

Going Beyond Qualitative Methods

Postmodernists would reject any argument that their methodology is simply qualitative social science in a different guise.

The gamut of modern qualitative approaches developed over the years (from observation to participant observation) retains a preoccupation with generalization and replication as much as an appreciation of complexity. Those employing conventional qualitative methods in social science have invested enormous energy in separating out what is procedurally acceptable from the unqualified. All this is abandoned in the postmodern methodology that embodies an attitude of humility and respect. It does not try to convince others that its own methodological perspective is right, superior, truthful, objective, the best assessment, or the source of the only correct answer. It does not settle on solutions or give privileged voice to any particular, singular point of view. It does not attribute authority to authorship. It queries and interrogates constantly. It follows its own advice: no new metanarrative. A metanarrative is a global worldview, a 'mastercode' that assumes the validity of its own truth claims.

An Upbeat Postmodern Methodology

What methodological alternatives does postmodernism have to offer vis-à-vis modern social science methodology? It emphasizes personal wisdom, lived experience, and knowledge as rich sources of material that inform writing and research. We are counseled to look to music, art, poetry, and literature instead of census data and public opinion polls. A postmodern social scientist would have more confidence in feelings, faith, empathy, emotion, insight, sensitivity, intuition, imagination, subjective judgment, and introspection as method. Postmodern social scientists are obviously and openly more comfortable with these than with an exclusive confidence in reason and logic. But by praxis, if not discourse, some postmodernists communicate that their alternative methodologies are of most value if disciplined by reason and logic. This amounts to coming to terms with the best of the Enlightenment heritage but without being intimidated by it. It suggests a refusal to accept the authority of modern science but also a desire to salvage from it whatever is of interest, such as the possibility of an interactive and reciprocal relationship between, say, reliable knowledge and meaningful participation in the public sphere, the creation of social life and political action, the enhancement of personal development, self-expression, self-determination, and collective praxis (Jennings and Graham, 1996). Doing social science might also have to be fun, enjoyable, entertaining rather than somber, listless, and mundane (Bakhtin, 1973; Todorov, 1984). It would not seek to demonstrate the obvious so much as to highlight the overlooked and the unrecognized.

Methodology as Art and Literature

Postmodern social science is often viewed as vision and art more than trained observation. It disperses any clear distinction between self and other; fact and values; author, text, and reader. It is inspired by the cultural and contextual preoccupations of anthropology where the distinction between literature and science was never clearly established. Anthropology had to be content with differing 'evocative devices of metaphor and imagery' because it was so difficult to arbitrate between different narratives from the past (Richardson, 1988). Postmodern social science methodology centers on 'listening to,' and 'talking with' the other (Tyler, 1986). The reader 'writes' the postmodern text and in so doing postmodern social science gives voice to those previously silent. It empowers through the methods of interpretation and criticism.

Postmodern methodology employs a literary capacity and it reinvigorates the mininarrative as a story that is worthy of notice even if it is not necessarily true in the modern sense of science. Storytelling is both a method and a means to emancipatory power (Newman and Holzman, 1999). The need to reconcile various stories, narratives, or research findings may or may not be urgent. The postmodernism tolerance for diversity and its indulgence of discrepancy reflect the fact that there is no rush to create, to construct, to accumulate, and to impose new knowledge in the modern sense.

Methodology as Interpretation

Postmodern methodology is interpretive but postmodern interpretation bears but slight resemblance to the more conventional social science forms of interpretation (Alvesson and Sköldberg, 2009). Modern social science understands interpretation to be the careful consideration of data with the goal of locating patterns. Meaning is not arbitrary and some interpretations are assumed to be better than others. But neither does postmodernism resemble hermeneutic interpretation that is more subjective and less ambitious than modern social science's interpretation. Hermeneutic interpretation, unlike postmodern interpretation, seeks to probe the 'silences,' to uncover a deeper meaning, masked, and hidden perhaps, but waiting to be discovered (Dreyfus and Rabinow, 1983; Madison, 1988). There is a current within postmodernism that says there simply is no 'meaning' to uncover. In many ways hermeneutic interpretation accepts truth, universality, and a shared discourse (a community of understanding).

Postmodern interpretation is *introspective*, *intersubjective*, and *antiobjectivist*, a form of individualized understanding. Postmodern theories of interpretation no longer 'ground,' they rather 'mediate' (Kugler, 1988). They range from 'interpretative reason' (Bauman, 1990) to prayer and chanting as 'interpretative communication.'

Methodology for a Complex Intertextual World

Postmodernists see the world as so complex that any direct causal statements relating to reality would necessarily be simplistic. They argue that it is impossible to established temporal priority in a simple fashion. They similarly are suspicious of modern versions of time that clearly delineate the past from the present, and include a strict version of linear time or geography. Their preference is for a present-oriented social science that erases linear time and ignores strictly bounded forms of geography. Their analysis emphasizes postmodern intertextuality, including, for example, chaos theory with its view of nonlinear social dynamics, and a definitely more complex version of reality (Young, 1992).

If the social universe for these postmodernists is intertextual, this has an impact on their methodology. Relationships, interconnections, and interactions are areas of substantial

interest. Conventional research tools are hardly appropriate. Quantitative methods isolate and simplify what is being studied, holding as many extraneous variables as possible constant so as to tease out causal, predictive associations. These postmodernists seek methodologies that integrate what is being studied with all that is related to it even if this complicates the research task, because in the end explanation in any absolute sense may be impossible.

Relativism and Questioning Postmodernists

Postmodernism is plagued, because of its European philosophical precursors, with the charge of relativism (Gellner, 1992) but many postmodernists defy this heritage and reject relativism. This is especially true for the postmodernists who come from a philosophically committed heritage: marxism, feminism, environmentalism, peace movements, ecology groups, and religious studies (Griffin, 1988, 1990; Kumar, 1995; Leonard, 1997). This leads them to argue that while deconstruction is an interesting first step to a postmodern methodology, it is limited because its goal is to undo all constructions. Deconstruction's intent is not to improve, revise, or offer a better version of the text, so they reject it. Postmodernism, in this view, must ultimately move to reconstruction and by extension, methodology remains essential.

Methodology Defined Narrowly

If methodology is defined narrowly as the rules and regulations of inquiry, then postmodernists must question the whole undertaking, arguing that there is no such thing as a postmodern methodology. 'Rules,' as such, are modern, not postmodern in character because they imply hierarchy, a series of good–bad evaluative criteria, an assessment on the basis of a set of agreed upon standards. Postmodernists reject rule-bound approaches as 'metanarratives.' Critics of postmodernism, however, respond that the absence of rules is itself a rule, the rule that 'anything goes' (Feyerabend, 1975; Heller, 1987) and that there is, thus, a postmodern methodology.

Rejecting Methodology Altogether

Postmodernists, inspired by Heidegger and Nietzsche, are antiscience and anti-Enlightenment, and this leads to an absolute relativism and to methodological nihilism. This postmodern orientation emphasizes the hopelessness and meaninglessness of life; theirs is the pessimistic, dark side of postmodernism (Rosenau, 1992).

Some such postmodernists practice epistemological impossibilism. They do not intend to provide 'clear reasoned defense.' They tell us their texts offer 'no consistent argument' and that 'not all points made are "necessary."' They boast that by design 'some essential matters are forgotten' (Corlett, 1989). To insist that these postmodernists develop a self-conscious methodology is to 'do violence' to their views, to impose one's own perspective, and thus diminish them by forcing them to acknowledge your views (Ashley and Walker, 1990a,b; Lyotard, 1988). In the end, from this postmodernist methodological perspective, we can never really know anything, not even our own feelings and emotions. These postmodernists speak of the death of the author (as agent) and of the absence of the subject (as a real person) with a specific identify. This means that they escape agency, accountability, and responsibility. They deny causality, question efforts to represent reality (through social science theory), and they contend that all interpretations are equally interesting (Miller, 1981).

All textual meaning, all interpretation, is undecidable within this postmodern perspective, because there is no final sense of a text and everything is a text in this postmodern world. No interpretation can be regarded as superior to any other (Latour, 1988). There is a multiplicity of readings and these postmodernists refuse to privilege one statement over another. "Every understanding is a misunderstanding" (Culler, 1982; Norris, 1988). Within this intellectual framework there is no need to bother about methodology at all.

Conclusion

Postmodern methodologies will be judged in the long term by their product. In the arts and literature, postmodern methods can be fascinating, challenging, and interesting. They stimulate creativity and innovation as exemplified by the writings of Umberto Eco, Italo Calvino, Milorad Pavic, and Salman Rushdie. They dissolve the boundary between fiction and social critique as the works of Jean Baudrillard and Michel Foucault so well illustrate. The fruitfulness of postmodern methodologies in the applied social sciences is yet to be demonstrated. Many feel them to be seriously flawed and epistemologically dubious. Postmodern methods contributed little toward the solution of the greatest social problems of the twentieth century. But then, that is not what they set out to do; it was never the intent.

See also: Symbolic Interaction: Methodology.

Bibliography

Alvesson, M., Sköldberg, K., 2009. Reflexive Methodology: New Vistas for Qualitative Research. SAGE Publications, Thousand Oaks, CA.
Ashley, R.K., Walker, R.B.J., 1990a. Reading dissidence/writing the discipline: crisis and the question of sovereignty in international studies. International Studies Quarterly 34, 367–416.
Ashley, R., Walker, R.B.J., 1990b. Speaking the language of exile: dissident thought in international studies. International Studies Quarterly 34, 259–268.
Bakhtin, M., 1973. Problems of Dostoevsky's Poetics. Ardis, Ann Arbor, MI.
Bauman, Z., 1990. Philosophical affinities of postmodern sociology. The Sociological Review 38, 411–444.
Corlett, W., 1989. Community without Unity: A Politics of Derridian Extravagance. Duke University Press, Durham, NC.
Cullenberg, S., Amariglio, J., Ruccio, D.F. (Eds.), 2001. Postmodernism, Economics, and Knowledge. Routledge, London.
Culler, J., 1982. On Deconstruction: Theory and Criticism after Structuralism. Cornell University Press, Ithaca, NY.
Derrida, J., 1972. Marges de la philosophie. Editions de Minut, Paris.
Derrida, J., 1981. Positions. University of Chicago Press, Chicago.
Dickens, D.R., Fontana, A. (Eds.), 1994. Postmodernism and Social Inquiry. Guilford Press, New York.
Dreyfus, H.L., Rabinow, P., 1983. Michel Foucault, beyond Structuralism and Hermeneutics, second ed. University of Chicago Press, Chicago.

Feyerabend, P., 1975. Against Method: Outline of an Anarchistic Theory of Knowledge. New Left Books, London.
Gellner, E., 1992. Postmodernism, Reason, and Religion. Routledge, London.
Griffin, D.R. (Ed.), 1988. The Reenchantment of Science: Postmodern Proposals. State University of New York Press, Albany, NY.
Griffin, D.R. (Ed.), 1990. Sacred Interconnections: Postmodern Spirituality, Political Economy, and Art. State University of New York Press, Albany, NY.
Heller, A., 1987. Mouvements culturels et changements des modeles de vie quotidienne depuis le deuxieme guerre. In: Corten, A., Tahon, M.B. (Eds.), La Radicalite du quotidien. VLB editeur, Montreal, PQ.
Jennings, L.E., Graham, A.P., 1996. Postmodern perspectives and action research: reflecting on the possibilities. Educational Action Research 2.
Kugler, P., 1988. From modernism to post-modernism: some implications for a depth psychology of dreams. Psychiatric Journal of the University of Ottawa 13, 60–65.
Kumar, K., 1995. From Post-Industrial to Post-Modern Society: New Theories of the Contemporary World. Blackwell, Oxford, UK.
Latour, B., 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA.
Leonard, P., 1997. Postmodern Welfare. Sage Publications, Thousand Oaks, CA.
Lyotard, J.F., 1988. L'inhumaine Causerie sur le Temps. Galilee, Paris.
Madison, G.B., 1988. The Hermeneutics of Postmodernity: Figures and Themes. Indiana University Press, Bloomington, IN.
Miller, J.H., 1981. The disarticulation of the self in Nietzsche. The Monist 64, 247–261.
Newman, F., Holzman, L., 1999. Beyond narrative to performed conversation. Journal of Constructivist Psychology 12, 23–40.
Norris, C., 1988. Deconstruction and the Interests of Theory. Pinter, London.
Richardson, L., 1988. The collective story: postmodernism and the writing of sociology. Sociological Focus 21, 199–208.
Rosenau, P.M., 1992. Post-Modernism and the Social Sciences. Princeton University Press, Princeton, NJ.
Todorov, T., 1984. Mikhail Bakhtin: The Dialogical Principle. University of Minneapolis Press, Minneapolis, PA.
Tyler, S., 1986. Post-modern ethnography: from document of the occult to occult document. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.
Young, T.R., 1992. Reinventing sociology. In: Mission and Methods in Postmodern Phenomenology. The Red Feather Institute, Weidman, Michigan.

Postmodernism: Philosophical Aspects

Paul Patton, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Postmodernism is an unstable concept because of the complex ways in which it overlaps with other concepts such as postmodernity, poststructuralism, and pragmatism. It may be identified by reference to an open-ended list of themes, including a preference for plurality, difference, and identity over unity, universality, and identity. Postmodernism also privileges a nonreferential conception of signs and images as simulacra rather than representations, although no less real in their effects. It also seeks to account for the possibility of criticism and to maintain commitments to freedom and progress, however redefined.

'Postmodernism' is a term that defies simple definition. Although earlier uses are recorded, the term has been widely used since the 1960s to refer to a range of antimodernist attitudes and strategies in art, architecture, and literature. References to postmodernism in philosophy and social theory first appeared in work by Jean-François Lyotard and Richard Rorty during the early 1980s. Philosophical postmodernism became associated with the idea that there was a 'crisis' in representation, with the abandonment of belief in a universal subject of reason and value, and with the rejection of unitary schemas of progress inherited from the Enlightenment. In place of these foundational motifs of modern thought, postmodernism emphasized difference, diversity, irreducible conflict, and the absence of progress in human affairs.

Bertens' *The Idea of the Postmodern: A history* begins with illustrations of the sometimes contradictory ways in which the term has been employed. For example, by 1970, modern painting had long abandoned the idea that representation or narrative was essential to its art. In explicit reaction against modernism understood as the self-reflexive exploration of the formal possibilities of painting, postmodern painting involved a reprise of representation and narrative. In a similar fashion, postmodern architecture rejected the formalism of the postwar International style and reembraced historical and vernacular styles. By contrast, at least in some of its manifestations, literary postmodernism involved a move in the opposite direction away from narrative and representation toward a form of radical self-reflexivity. Bertens concludes that, depending on the artistic discipline, "postmodernism is either a radicalization of the self-reflexive moment within modernism, a turning away from narrative and representation, or an explicit return to narrative and representation. And sometimes it is both" (Bertens, 1995: p. 5).

There are several reasons why postmodernism should be "a particularly unstable concept" (Bertens, 1995: p. 12). First, to the extent that postmodernism is considered to be a cultural movement alongside others such as romanticism, positivism, or structuralism, then like these it will encompass a variety of distinct aesthetic and intellectual strategies. Second, commentators and users of the term hesitate over whether postmodernism is primarily a temporal or a typological concept. The prefix and the manner in which the term has been employed in the arts suggest a temporal reference, even though estimates of the beginning of the postmodern era range from the 1860s to 1980s. By contrast, some theorists consider postmodernism in the arts and culture to be a radicalization of tendencies already present in the modern. For example, Lyotard argues that postmodernism in painting stands for an aesthetic of the sublime that embraces the ever-renewed attempt to 'present the unpresentable' (Lyotard, 1983). Following this line of thought, Umberto Eco suggests that postmodernism should not be chronologically defined but rather treated as "an ideal category – or better still, a *Kuntswollen*, a way of operating. We could say that every period has its own postmodernism, just as every period would have its own mannerism" (Eco, 1985: p. 66). Third, the concept of postmodernism stands in a variety of complex relations to other concepts such as modernity, postmodernity, modernism, poststructuralism, and deconstructionism. Since the meaning of these concepts is similarly uncertain, so our understanding of 'postmodernism' will be affected by the particular meaning attached to these related concepts. In view of the many different phenomena to which it has been applied, it is common for discussions of postmodernism, to begin, as this one has, with the claim that the term has no clear meaning (Eco, 1985: p. 65; Rorty, 1999: p. 262).

Modernity/Modernism

Just as modernism is often regarded as a response to the process of modernization that transformed the material conditions of European society during the nineteenth century, so postmodernism is often presented as a consequence of the restructuring of late capitalism at the end of the twentieth century. The historical period of European modernization, which extended from the middle of the eighteenth through to the middle of the twentieth century, saw the development of industrial infrastructure and means of production such as railways, automobiles, air transport, electricity, and telegraphic and telephonic communication. It also saw the growth of cities, the spread of public schools, hospitals, and prisons, and the development of bureaucratic and rational procedures for the government of social life. 'Modernism' commonly refers to an artistic sensibility that embraced innovation and change. In its early stages, modernism involved a positive sense of living at the dawn of a new era. To be modern implied a heroic

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63062-3

commitment to action and to revolutionizing oneself as well as the means of artistic production. In its later stages at the end of the nineteenth century and the beginning of the twentieth, modernism involved experimentation with the style and form of the various arts, hence the dynamic of successive avant-gardes in literature, music, and visual arts.

Theorists of postmodernity such as Fredric Jameson and David Harvey assert a parallel connection between the changes associated with the restructuring and globalization of late capitalism and the emergence of a postmodern culture (Jameson, 1991; Harvey, 1989). They argue that changes in technology, urban space, and forms of consumption have created the conditions of a new mode of experience of self and others. In particular, the 'compression' of social space and time brought about by new electronic and communications technologies is reflected in a new cultural sensibility. Just as the destruction of traditional ways of life and thought wrought by modernization was often linked to a tragic sense of loss, so theorists of postmodernity argue that a similar kind of homelessness is generated by the late twentieth century mutations in our experience of space and time. Among those who take a negative view of the cultural experience of postmodernity, Jameson argues that it does not sustain a sufficiently robust sense of self and history. He suggests that the rapid evolution of postmodern urban space has overtaken "the capacities of the individual human body to locate itself, to organize its immediate surroundings perceptually, and cognitively to map its position in a mappable external world" (Jameson, 1991: p. 83). On his account, postmodernity is experienced as incapacity in the face of complex and discontinuous social spaces and as disorientation in the absence of a rich sense of the past or future. The result is a widely documented fragmentary and schizophrenic experience of subjectivity, along with a tendency toward affectless repetition or 'blank parody' of past forms of artistic and cultural production.

For others, modernist and postmodernist alike, the appropriate affective response to change is less a matter of nostalgia for older ways of being at home and more a question of orientation toward the future. This was the attitude of Friedrich Nietzsche, who in the view of many, is the source of much postmodern philosophy. For Gianni Vattimo, for example, "It could legitimately be argued that philosophical postmodernity is born with Nietzsche's work" (Vattimo, 1991: p. 164). Writing at the end of the nineteenth century in response to what he took to be the defining event of modernity, namely the fact that belief in the Christian God had become unsustainable, Nietzsche employed the condition of 'homelessness' as a metaphor for the situation of those who no longer believe in a divine providence but nevertheless look forward to 'the strengthening and enhancement of the human type.' Among the Europeans of his time, he wrote, there is no shortage of those "who have a right to call themselves homeless in a distinctive and honorable sense … "We children of the future – how *could* we be at home in this today?" (Nietzsche, 2001: p. 241).

Unity and Plurality

A feature of many cultural responses to the transient, fleeting, and ephemeral character of modern life is their attachment to ideas of unity, whether in the object to be represented or the artifact to be produced. Thus, for example, modernists believed that a machine or a city could be a unified, smoothly functional object, and a novel or work of art could be a complete, self-contained, meaningful entity. Even where there is experimentation with diverse modes of representation, this is undertaken in the attempt to convey a single complex and multifaceted reality. And even where there is a dizzying succession of conceptions of the nature of a particular art such as painting, there is, nevertheless, a conviction at each stage that at last the true form of the art has been discovered.

Even though 'modernism' does not have the same currency as a classificatory term in philosophy and social theory as it does in the arts, there are epistemological parallels to this modernist attachment to unity. Modernist thought would be that which tolerated perspectivism only on the assumption that there is a single, intelligible reality to be presented. Even when there is an acceptance of surface incoherence, as in Marx's account of the contradictions of capitalism or Freud's account of conscious behavior, this is balanced by the discovery of an underlying structure that produces such contradictory effects. In this sense, Harvey argues that modernism "took on multiple perspectivism and relativism as its epistemology for revealing what it still took to be the true nature of a unified, though complex, underlying reality" (Harvey, 1989: p. 30). By contrast, postmodernism is associated with the acceptance of difference and plurality in both ways of being and ways of knowing. In opposition to the ideal of unified objects of knowledge, whether texts or artifacts, postmodernism refers instead to networks, open systems, or the dissemination of meaning. In opposition to the modernist ideal of unified science, postmodernism upholds the idea of irreducible difference and even incommensurability among the varieties of knowledge.

In their efforts to make positive sense of such difference, along with the experience of being estranged from the social world in which they must live, postmodernists often have recourse to the concept of irony. Umberto Eco argues that the postmodern attitude is encapsulated in the irony that acts and speaks in full awareness of the degree to which all our ways of acting and speaking involve repetition of the past: "the postmodern reply to the modern consists of recognizing that the past, since it cannot really be destroyed, because its destruction leads to silence, must be revisited, but with irony, not innocently" (Eco, 1985: p. 67). Rorty contrasts metaphysical thinkers committed to the unity of truth and a hierarchy of values with ironists who are conscious of the contingency and mutability of the 'final vocabularies' employed to make sense of their lives, yet nonetheless committed to the values that those vocabularies embody (Rorty, 1989: p. 73). He argues that, to the extent that postmodernism stands for anything distinctive in philosophy, it refers to a pluralism shared by a range of modern thinkers such as Nietzsche and William James. After having used the term 'postmodernism' in the 1980s to describe his own philosophy and that of others such as Heidegger and Derrida who were, in Lyotard's phrase, "distrustful of metanarratives" Rorty became skeptical of its utility. In the Introduction to *Essays on Heidegger and Others* (Rorty, 1991), he says that he had given up on "the attempt to find something common to Michael Grave's buildings, Pynchon's and Rushdie's novels, Ashberry's poems, various sorts of

popular music, and the writings of Heidegger and Derrida" (Rorty, 1999: p. 1). He gave up on the attempt to periodize every aspect of culture in favor of efforts to periodize particular disciplines and genres of cultural activity, suggesting that so-called postmodern philosophy is better described as post-Nietzschean and pragmatist philosophy. In a later essay, he argued that as various as the definitions of 'postmodern' are, "most of them have to do with a perceived loss of unity" (Rorty, 1999: p. 262).

Postmodernism and Poststructuralism

Postmodernism in philosophy and social theory is widely associated with the work of a number of French thinkers, notably Baudrillard, Deleuze, Derrida, Foucault, Irigaray, and Lyotard. Because these authors emerged from the shadow of the structuralism of Althusser, Lacan, Levi-Strauss, and others at the end of the 1960s and beginning of the 1970s, they are also commonly referred to as post-structuralists. By different intellectual paths, these thinkers rejected the idea that theories and concepts simply represent a preexisting reality. For some, such as Derrida or Baudrillard, this conclusion was reached via a radicalization of Saussurian theses about the arbitrariness of signs and the manner in which their sense is derived from differential relations to other signs. For others, such as Foucault, similar conclusions were reached via Kantian and Heideggerian views of the role played by language and thought in the construction of its own objects. A consequence of Foucault's historicization of the Kantian idea that objects of knowledge are determined in part by the underlying rules governing discourse is that 'truth' acquires a history and can no longer be measured solely against an independent reality.

While the work of Foucault and others does imply a form of historical relativism with regard to truth, it is overstating the case to suggest that the resultant 'crisis in representation' amounts to "a deeply felt loss of faith in our ability to represent the real, in the widest sense" (Bertens, 1995: p. 11). It is not our ability to represent reality but rather what Deleuze calls the representational "image of thought" that is impugned by poststructuralist epistemology (Deleuze, 2011: pp. 164–213). Foremost among the influential philosophical assumptions that make up the representational image of thought is the idea that thought has an 'affinity' with truth, that it seeks truth by nature and that it is fully equipped to attain a true representation of the world. For Deleuze as for Heidegger, it is not recognition or representation but creation that is the highest expression of thought: "thought only thinks when constrained or forced to do so, in the presence of what 'gives cause to think' or of what is to be thought – and what is to be thought is just as much the unthinkable as the unthought, in other words the perpetual fact that 'we do not yet think'" (Deleuze, 2011: p. 182).

The turn to language in Heidegger's later philosophy was paralleled by the linguistic turn in analytic philosophy, inspired by Wittgenstein's posthumous *Philosophical Investigations*. Both of these philosophical tendencies converge in Rorty's neo-pragmatism (Sheehan, 2004: p. 24). For Rorty, the core of postmodernism in philosophy is the abandonment of the correspondence theory of truth and the "theologicometaphysical belief that Reality and Truth are One – that there is One True Account of How Things Really Are" (Rorty, 1999: p. 262). A less negative characterization might say that postmodernism does not, so much, reject the idea that thought represents reality as embrace more complex views of what representation entails. In this manner, for example, Foucault's genealogies of modern punishment and sexuality drew attention to the ways in which what passes for truth in the social sciences is bound up with the exercise of power over individuals and populations. More generally, postmodernism has drawn attention to new aspects of social reality, such as the systems of thought and language, which condition public discourse and what we take to be the truth. It has also given rise to new kinds of analysis of cultural and social phenomena: genealogical, deconstructive, narratological, and so on. In these ways, postmodernism has given rise to new research methods and new objects of analysis in anthropology, sociology, and other social sciences, as well as in jurisprudence and the philosophy of law (Clifford and Marcus, 1986; Dickens and Fontana, 1994; Litowitz, 1997).

Universalism and Freedom

One of the most influential formulations of postmodernism in philosophy was provided in Lyotard's *The Postmodern Condition* (Lyotard, 1984), first published in 1979 (Sim, 2005: pp. 6–10). Here, Lyotard uses 'modern' to refer to any form of knowledge that seeks legitimation by appeal to some 'grand narrative, such as the dialectics of the spirit, the hermeneutics of meaning, the emancipation of the rational or working subject, or the creation of wealth.' He then defines the postmodern attitude as one of "incredulity toward metanarratives" (Lyotard, 1984: pp. xxiii–xxiv). Lyotard is not concerned with postmodernity understood as an historical epoch, since his inquiry is limited to "the condition of knowledge in the most highly developed societies" (Lyotard, 1984: p. 13). Rather, his aim is to show that the supposed transcendent status of scientific knowledge is an illusion sustained by a certain narrative language game. He argues that the privileged status of scientific knowledge is destined to disappear once it is accepted that this status depends upon a narrative that privileges science over other kinds of language game. In European philosophy, this narrative of legitimation has either followed the French Enlightenment link between the pursuit of knowledge and human emancipation, or the German Idealist construal of history as the process of the realization of reason through the sciences. Lyotard asserts not only that these 'metanarratives' have lost their appeal, but also that the very idea of transcendent legitimation has given way to forms of legitimation, which are local and immanent in particular social practices and forms of life. The postmodern condition therefore is the state in which we must accept this plurality of micronarratives and find other ways to address the resulting conflicts that do not involve the imposition of a single master narrative.

In *Modernity—An Incomplete Project*, Habermas (1983) provided an influential response to Lyotard and other post-structuralist thinkers. He characterized the philosophical project of modernity as the desire for unification of the three spheres of cultural value distinguished by Kant – the

theoretical, the practical, and the aesthetic – and for the application of the results of these distinct kinds of judgment to the improvement of everyday life. According to this view, the project of modernity consisted in the effort of Enlightenment thinkers "to develop objective science, universal morality and law, and autonomous art, according to their inner logic. At the same time, this project intended to release the cognitive potentials of each of these domains from their esoteric forms. The Enlightenment philosophers wanted to utilize this accumulation of specialized culture for the enrichment of everyday life – that is to say for the rational organization of everyday social life" (Habermas, 1983: p. 9). The major obstacle preventing the realization of this project, according to Habermas, lay in the manner in which these spheres have become separated from one another, both institutionally and in their internal logics and forms of argumentation, and in the manner in which they have become increasingly separated from a lifeworld dominated by instrumental rationality.

Habermas's solution to this problem appealed to the procedural rationality implicit in language use, or what he calls 'communicative reason,' as a means to restoring the possibility of intersubjective consensus. Both his diagnosis and his solution to the problem of modernity are founded upon the Enlightenment belief that universal standards of judgment are necessary, both for the development of reason and for moral progress. By contrast, postmodern thinkers reject the idea of progress in human affairs in favor of the idea that there are many ways in which the present differs from the past. In common with Lyotard, they reject the idea of universal standards of judgment, value or 'legitimation.' Although Habermas characterizes French thinkers such as Foucault, Derrida, and Lyotard as "young conservatives" (Habermas, 1983: p. 14), most of them do not wish to abandon the ideals of freedom and emancipation but only the approach that links these to universality, unity, and consensus and does so in the name of the Enlightenment.

In response to Habermas's criticisms, Foucault presented his practice of genealogy as serving the cause of freedom, while suggesting that it represented a critical ethos or attitude toward the present that had always belonged to the Enlightenment (Foucault, 1984). Genealogical criticism does not seek to identify the universal structures of human knowledge and moral action, but rather takes the form of an historical investigation into the means by which we have been constructed as certain kinds of subject. Against the presumption that human reason is an unchanging and undivided faculty, Foucault views reason as divided among the distinct rationalities embodied in particular systems of thought and practice such as 'medicine,' 'discipline,' or 'sexuality.' The aim of this type of historical investigation, Foucault suggests, is to determine "in what is given to us as universal, necessary, obligatory, what place is occupied by whatever is singular, contingent, and the product of arbitrary constraints" (Foucault, 1984: p. 45). Thus, in contrast to the idea that progress in human freedom depends upon the extension of the sphere in which universal rationality governs social life, Foucault presents progress in human freedom in negative terms as the escape from particular forms of domination. For many poststructuralist theorists, freedom is not linked to the achievement of universal standards of value or to the hope that dissensus will eventually give way to consensus. Rather, it is dissensus and the experience of irreducible difference that are the conditions of freedom and creativity or change in human affairs.

Difference and Multiplicity

In his later work, Lyotard introduced the concept of the differend in order to characterize the postmodern condition with respect to political and moral judgment. A differend is a conflict between parties who inhabit different regimes of utterance or discourse, the sentences of which are incommensurable with one another (Sim, 2005: p. 9). As a result, the conflict between the parties cannot be resolved peacefully for lack of a rule of judgment common to both (Lyotard, 1988: p. 9). In practice, one party tends to enforce its discursive regime on the other. Foucault also embraced the idea of radical difference between regimes of discourse. He described his critical attitude toward the present as one that endorsed differentiation from the past rather than movement toward some future identity (Foucault, 1984). However, the recent origins of this concern with difference must be sought in the philosophical work of Deleuze and Derrida published during the 1960s. Both sought to elaborate a nondialectical concept of difference, which would see it neither as the simple contrary of identity, nor as 'dialectically' identical with identity (Descombes, 1980: p. 136). Hegel was a focus of their criticism because he represented the culmination of a metaphysical tradition that treated identity as primary and difference as the derivative or secondary term. In *Difference and Repetition*, Deleuze (2011) explicitly rejects the link which Hegel forged between difference and contradiction, arguing that contradiction is not the condition or ground of difference. On the contrary, 'It is not difference which presupposes opposition but opposition which presupposes difference, and far from resolving difference by tracing it back to a foundation, opposition betrays and distorts it' (Deleuze, 2011: p. 62). Limitation or opposition is a distortion of difference, according to Deleuze, because difference in itself implies 'a swarm of differences, a pluralism of free, wild, or untamed differences' (Deleuze, 2011: p. 61). Similarly, in *Of Grammatology*, Derrida (1976) drew upon Heidegger's concept of the ontological difference between being and beings and Saussure's concept of the differentiation that gives rise to units of signification in a sign system to draw attention to the movement of deferral and differentiation (*différance*) that underlies all production of meaning (Derrida, 1976: pp. 56–65).

Both Deleuze and Derrida were concerned with the problem of how to conceive of the identity of objects such as meanings, texts, or signs that are subject to continuous variation. They sought to describe a form of unity that is not identical with itself. To this end, Derrida relied upon a series of concepts such as *différance*, the trace, and iterability, while Deleuze relied upon a concept of open or variable multiplicity derived from Bergson. In *A Thousand Plateaus* (Deleuze and Guattari, 1987), this concept of multiplicity provided the formal structure for the elaboration of a series of philosophical concepts of such things as desire, subjectivity, language use, knowledge, and forms of social organization. The concept of multiplicity, he and Guattari suggest, 'was created precisely in order to escape the abstract

opposition between the multiple and the one, to escape dialectics, to succeed in conceiving the multiple in the pure state, to cease treating it as a numerical fragment of a lost Unity or Totality or as the organic element of a Unity or Totality yet to come, and instead distinguish between different types of multiplicity' (Deleuze and Guattari, 1987: p. 32). Deleuze and Guattari defended this invention of novel concepts as a new practice of philosophy, the ultimate aim of which was to bring about new ways of speaking and acting, to enable us to perceive the world differently and thereby to bring about new worlds and new peoples (Patton, 2010).

A distinction is often drawn between 'postmodern' theorists, such as Deleuze–Guattari, Derrida, and Foucault, and theorists of postmodernity such as Jameson and Harvey. The theorists of postmodernity tend to adopt a broadly Marxist approach to society and culture, and to the cultural phenomena they see as representative of the social totality, in order to pass judgment. Their 'modernist' social theory presupposes that a total picture is possible and that cultural phenomena can be understood as effects of a certain economic 'logic.' They proceed on the assumption that critical distance is possible and that an appropriate interpretative framework will provide such distance. By contrast, the 'postmodern' theorists tend to abandon economic or other forms of historical determinism, along with the idea that a totalizing theory of society and history is possible or desirable, in favor of the idea of multiple, overlapping, and divergent histories written from the distinctive perspective of particular social groups, peoples, or interpretative frameworks (power, desire, the body, etc.). This distinction is not absolute, since there are those such as Baudrillard, Foucault, and Lyotard who might be included in both camps although at different stages of their careers. However, it is characteristic of postmodern theorists that their work involves a self-conscious attempt to theorize in a different manner, while also searching for critical strategies that do not rely upon a normative standpoint outside or apart from the object of criticism.

Simulacra and Simulation

Philosophies that seek to make difference, an object of affirmation, must do more than reverse the traditional hierarchy between identity and difference. The mere inversion of hierarchy does not change the fundamental relation between the elements involved, nor does it change the nature of those elements. For this reason, Derrida's practice of deconstructive philosophy always envisaged a further stage after the initial hierarchy has been overturned, which would involve "the irruptive emergence of a new 'concept,' a concept that can no longer be, and never could be, included in the previous regime" (Derrida, 1981: p. 42). In the course of his essays, he produced or drew attention to a series of such paradoxical concepts, including the trace, writing-in-general, metaphoricity, and Plato's *pharmakon*, all of which served to disrupt conceptual oppositions and hierarchies in philosophy. Similarly in *Plato and The Simulacrum* (Deleuze, 1990: pp. 253–266), Deleuze argued that the concept of simulacra served to undermine or 'overturn' the primacy of identity as this was embodied in Plato's conception of a model or form of which reality is only an imperfect copy.

However, it was the sociologist Jean Baudrillard who took up the concept of simulacra and applied it to aspects of everyday life. In his earlier work, Baudrillard had argued that industrial society at the end of twentieth century had entered a new phase in which consumption rather than production served to structure social relations. He argued that the Marxian view of commodities as useful items, produced in order to realize exchange value, must be replaced by a semiotically informed view of commodities as signs, produced in order to realize their 'sign value' as differential elements of a social code of signification. In a series of essays published during the early 1980s, Baudrillard built upon this analysis and other accounts of the power of mass media to argue that society had entered a new era in which simulation had replaced representation as the dominant principle of signification. The essays translated in *Simulations* (Baudrillard, 1983) combined a Nietzschean-inspired critique of metaphysics with a descriptive account of the present stage of social life as a regime of simulacra in which the distinction between reality and the forms of its representation has broken down. According to Baudrillard, there are many areas of social life, such as the media, where signs or other representations do not faithfully represent a prior reality but function in its stead. As a result, for practical purposes we live in a hyperreal social world of representations without referents, images without originals. The proliferation of media transmitted information and the speed with which it circulates creates a perpetual cultural present in which contemporary society has begun to lose its capacity to retain its own past.

Criticism without Foundations

While critics object to the nihilistic, impressionistic, and highly selective nature of Baudrillard's descriptions, his analyses of technological, media, and political events as simulations served to draw attention to this pathological feature of the experience of social reality at the end of the twentieth century. His essays have been criticized for their failure to provide any stable foundation for criticism (Norris, 1992). However, Baudrillard openly acknowledges their ironic and hyperbolic style, describing them as 'theory–fictions' intended to provoke and to challenge contemporary common sense rather than convince by evidence or argumentation. Such a provocative function is often invoked in defense of postmodern theory. At best, it is argued, postmodern approaches serve to problematize established concepts and to draw attention to ways in which traditional universalist and humanist assumptions function to exclude or marginalize differences.

Postmodern theory is also criticized for its limited and negative approach to normative issues. Critics assert that the antiuniversalism of Foucault, Lyotard, and others is incapable of providing grounds for moral or political judgment much less comprehensive programs for social reform. Defenders argue that their aim is both more limited and more modest, namely, to work on the limits of what it is possible to think or to say and thereby to contribute to the emergence of new norms, which may also become bases for judgment. To those who argue that the task of problematizing existing ways of thinking or challenging existing norms is ultimately of limited importance, postmodernists reply that not to do so is to limit the possibilities

for thinking and acting differently. For this reason, they argue, the ethos or *Kunstwollen* of postmodernism should be seen as a necessary element of every genuinely critical approach.

See also: Deconstruction; Foucault, Michel (1926–84); Modernization and Modernity in History; Nietzsche, Friedrich (1844–1900); Phenomenology: Philosophical Aspects; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Baudrillard, J., 1983. Simulations (P. Foss, P. Patton, P. Beitchmann, Trans.). Semiotext, New York.

Bertens, H., 1995. The Idea of the Postmodern: A History. Routledge, London.

Best, S., Kellner, D., 1991. Postmodern Theory: Critical Interrogations. Macmillan, Basingstoke, UK.

Clifford, J., Marcus, G.E. (Eds.), 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Deleuze, G., 1990. In: Lester, M., Stivale, C., Boundas, C. (Eds. and Trans.), The Logic of Sense. Columbia University Press, New York and Athlone Press, London.

Deleuze, G., 2011. Difference and Repetition (P. Patton, Trans.). Continuum International Publishing Group, London. Reprinted edition with corrections.

Deleuze, G., Guattari, F., 1987. A Thousand Plateaus: Capitalism and Schizophrenia (B. Massumi, Trans.). University of Minnesota Press, Minneapolis, MN and Athlone Press, London.

Derrida, J., 1976. Of Grammatology (G.C. Spivak, Trans.). Johns Hopkins University Press, Baltimore, MD.

Derrida, J., 1981. Positions (A. Bass, Trans.). University of Chicago Press, Chicago.

Derrida, J., 1982. Margins of Philosophy (A. Bass, Trans.). University of Chicago Press, Chicago, IL.

Descombes, V., 1980. Modern French Philosophy (L. Scott-Fox, J.M. Harding, Trans.). Cambridge University Press, Cambridge, UK.

Dickens, D.R., Fontana, A., 1994. Postmodernism and Social Inquiry. Guildford Press, New York.

Eco, U., 1985. Reflections on the Name of the Rose. Secker and Warburg, London.

Foucault, M., 1984. What is enlightenment? In: Rabinow, P. (Ed.), The Foucault Reader, first ed. Pantheon, New York, pp. 32–50.

Habermas, J., 1983. Modernity—an incomplete project? In: Foster, H. (Ed.), The Anti-esthetic: Essays in Postmodern Culture. Bay Press, Port Townsend, WA, pp. 3–15 [Reprinted 1981 version of Modernity versus postmodernity. New German Critique 22: 3–14].

Harvey, D., 1989. The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change. Blackwell, Oxford, UK.

Jameson, F., 1991. Postmodernism, or, the Cultural Logic of Late Capitalism. Duke University Press, Durham, NC.

Litowitz, D.E., 1997. Postmodern Philosophy and Law. University Press of Kansas, Lawrence, KS.

Lyotard, J.-F., 1983. Answering the question: what is post-modernism? In: Hassan, I., Hassan, S. (Eds.), Innovation Renovation: New Perspectives on the Humanities. University of Wisconsin Press, Madison, WI, pp. 329–341.

Lyotard, J.-F., 1984. The Postmodern Condition: A Report on Knowledge. University of Minnesota Press, Minneapolis, MN.

Lyotard, J.-F., 1988. The Differend: Phrases in Dispute. University of Minnesota Press, Minneapolis, MN.

Nietzsche, F., 2001. In: J. Nauckhoff, Trans., Williams, B. (Ed.), The Gay Science. Cambridge University Press, Cambridge, UK.

Norris, C., 1992. Uncritical Theory: Postmodernism, Intellectuals and the Gulf War. Lawrence and Wishart, London, UK.

Patton, P., 2010. Deleuzian Concepts: Philosophy, Colonization, Politics. Stanford University Press, Stanford, CA.

Rorty, R., 1989. Contingency, Irony, and Solidarity. Cambridge University Press, Cambridge, UK.

Rorty, R., 1991. Essays on Heidegger and Others: Philosophical Papers, vol. 2. Cambridge University Press, Cambridge, UK.

Rorty, R., 1999. Afterword: pragmatism, pluralism and post-modernism. In: Philosophy and Social Hope. Penguin, London, pp. 262–277.

Sheehan, P., 2004. Postmodernism and philosophy. In: Connor, S. (Ed.), The Cambridge Companion to Postmodernism. Cambridge University Press, Cambridge, UK, pp. 20–42.

Sim, S., 2005. Postmodernism and philosophy. In: Sim, S. (Ed.), The Routledge Companion to Postmodernism, second ed. Routledge, London and New York.

Vattimo, G., 1991. The End of Modernity: Nihilism and Hermeneutics in Postmodern Culture. Johns Hopkins Press, Baltimore, MD.

Post-Socialism: Political Economy

Aleksandra Sznajder Lee, University of Richmond, Richmond, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the political economy of the reforms undertaken in the contentious climate of fragile democratic transition following the collapse of Soviet-style communism in Central and Eastern Europe (CEE). It first surveys the enormity of the tasks facing the reformers and the historical and political domestic context that shaped the initial debates about reform, and, ultimately, forged the reform pathways. The article then turns to the strong external pressures influencing economic reform, most significantly to those of the European Union. Collectively, these external pressures facilitated structural changes that resulted in highly liberalized and globally integrated CEE economies.

Introduction

After the fall of Soviet-style communism in Central and Eastern Europe (CEE), the countries in the region faced the daunting challenge of rejoining the West. This article focuses on economic reform, which was subject to contentious political debates in the young, and often fragile, democracies. Although local political contexts shaped the initial reform pathways in CEE, these trajectories soon became subject to strong external pressures emanating from international financial institutions, world markets, and most importantly, the European Union (EU). The resulting structural changes turned these economies into some of the most liberalized and globally integrated economies in the world.

Economic Reform Agenda

As the communist system collapsed, countries in the region faced the unprecedented challenge of moving from centrally planned to market economies. Within a very short time span these countries both sought and were expected to reestablish market institutions and rejoin the global capitalist economic order. With the EU at their doorstep, the overarching goal of the 'return to Europe' quickly became the battle cry of reformers from the former Soviet satellite states. It was a goal to which even the less ardent supporters of reform paid lip service. As will become apparent, the task of economic (re)integration was complicated by the type of development these countries had undergone under the communist rule, as well as by the complex institutional evolution that had taken place in the West.

Although the reform priorities, sequencing, and design differed depending on the political context of the country at hand, the list of tasks facing reformers was commonly agreed upon (Mitra and Selowsky: p. 14). First came the need to liberalize prices and trade in order to improve allocative efficiency and create a competitive market environment. Preexisting inflationary pressures, as well as the aftermath of rapid liberalization created the need for anti-inflationary measures that would stabilize the macroeconomy (Roland, 2000: p. 11; Mitra and Selowsky, 2002: p. 14). These included restrictive monetary policy (high interest rates) and balanced budgets. The latter was especially problematic given the legacy of the 'soft budget constraint' (Kornai, 1980). Rather than forcing the enterprises to adjust their production practices, inefficient enterprises that habitually operated in the red were being bailed out by the responsible ministry. Eliminating or severely curbing subsidies and other forms of assistance to both enterprises and banks proved to be an extremely demanding task politically.

A related task was creating an environment conducive to the development of the private sector. This entailed the overwhelming challenge of privatizing existing state-owned enterprises (SOEs), especially the large ones, as well as the encouragement of the entry of domestic entrepreneurs into the market. Another (still ongoing) reform, that went hand-in-hand with privatization but proved to be extremely difficult, was providing an institutional structure conducive to good corporate governance and enterprise restructuring. This, in turn, called for a series of legal and judicial reforms. Yet another set of reforms that were critical to the development of the private sector pertained to banking and to the creation of securities markets, and nonbank, financial institutions. The reform to-do list also included the reform of the tax system and restructuring of public expenditures, as well as the institutional reform of the public sector more generally (Mitra and Selowsky, 2002: p. 14; European Bank for Reconstruction and Development, various years).

Communist Legacies and Challenges to Economic Reform

These daunting reforms were complicated by the type of development transition economies had undergone during the communist system. All of the countries discussed here had large industrial sectors, ranging from 30% GDP in Hungary to 59% in Bulgaria and Czechoslovakia (European Bank for Reconstruction and Development, 1999). More problematic than the sheer extent of these economies' reliance on industrial production was the type of industrial production in which they were engaged. The effort to build a communist economy entailed a massive modernization project where heavy industry, rather than industries producing consumer goods, played the primary role. Because success was measured by

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10126-6

output, and the enterprises did not have a hard bottom line, industries such as steel, heavy machinery, armaments, and chemicals suffered from overcapacity and inefficiency in the context of a market economy. The case in point was Czechoslovakia, where the communist authorities sought to close the developmental gap between the Czech lands (Bohemia and Moravia) – the most industrially developed in CEE – on the one hand, and the largely agricultural Slovakia, on the other. They succeeded, but industrial development took, in part, the form of armaments factories that serviced the needs of the Warsaw Pact. The end of the Cold War rendered these obsolete and greatly complicated the reform process in Slovakia (Wolchik, 1994: p. 64). The monoindustrial towns throughout the post-communist region, where the entire local economy revolved around a single giant enterprise faced particular economic and social challenges.

Despite their shared legacy of central planning, membership in the barter-based, Soviet-sponsored Council for Mutual Economic Assistance, and political dependence on the Soviet Union, the former communist countries did not all begin the reform process from the same starting point. Even though they shared the same basic challenges, their exact nature and severity varied by country. These differences were rooted in the historically contingent nature of communist rule in each country. There are nuanced typologies of communist regimes in CEE that take into consideration the relationship between the communist authorities and society (see Kitschelt et al., 1999). For the sake of simplicity, however, communist regimes can be broadly divided into more 'liberal' and 'orthodox' in their approach to political and economic developments.

The more liberal regimes, most notably Poland and Hungary, allowed for greater personal freedoms than their more orthodox counterparts, and they also engaged in forms of economic liberalization. Both countries allowed private enterprise to exist. Private sector's share in GDP equaled 30% in Poland and 25% in Hungary in 1990 (European Bank for Reconstruction and Development, 1999: pp. 228, 252). The relatively large private sector in Poland resulted from aborted efforts at collectivizing agriculture, which then remained largely in the hands of individual farmers. It also stemmed from limited forms of small private businesses. Hungary's private sector reflected the Hungarian regime's tolerance of the so-called second economy, or parallel markets, where individuals, often using machinery leased from state enterprises, would produce consumer goods undersupplied by the state (Stokes, 1993: p. 81).

Both countries' relative openness to the West also had its price as it led to accumulation of large amounts of foreign debt in the 1970s. The abundance of petrodollars and ease of credit created incentives to engage in a Western loan-financed, large-scale industrial investment project extravaganza. The projects had disappointing results, and the countries were left with the bill. By contrast, Czechoslovakia had very little foreign debt, as did East Germany, which nonetheless benefited from significant (unofficial) West German subsidies. Bulgarian foreign debt only ballooned in the late 1980s while Romania paid off the entirety of its foreign debt as part of its leader's brutal campaign for total independence from the West (Stokes, 1993: p. 143).

The communist-era economic reform efforts in both Poland and Hungary led to partial liberalization prior to system change, but these measures created the need and opportunity for further reform. Partial price liberalization fueled inflationary pressures. So did the self-management of enterprises where the managers faced worker pressures for increased wages. The same was the case in Yugoslavia, which was not a Soviet satellite state and followed its own version of socialist self-management. Interestingly, similar problems also surfaced in the aftermath of Gorbachev's *perestroika* reforms in the Soviet Union, which preceded its breakup.

If the more 'liberal' communist states had already been exposed to some of the elements of rudimentary market principles, they were also westward-oriented in terms of seeking assistance with further reforms. The case in point was the joining of the International Monetary Fund and the World Bank by Hungary in 1982 and by Poland in 1986, and the allowing of the creation of joint ventures with foreign capital. The pinnacle of westward orientation was a move that dramatically flew in the face of communist ideology: the decision of the Hungarian Socialist Workers' Party to attempt to sell more than 50 SOEs to foreign owners in January 1989 (Stark and Bruszt, 1998: p. 54).

Neoliberals and Their Critics

The West indeed took an active interest in the Soviet satellite states as communism collapsed, especially in those in which it had a financial stake, such as Poland and Hungary. Western advisers, most prominently Jeffrey Sachs, flocked to the region and supported the local reformers in their efforts. Although the totality of the reform agenda, outlined above, may not have been controversial, the priorities, sequencing, and the means through which the reforms were to be undertaken differed substantially among countries. They also generated significant debate. The reform choices were closely rooted in the domestic political conditions, as will be outlined in broad brushstrokes below. First, however, we turn to the biggest controversy that plagued practitioners and academics alike during the first decade of transition: the neoliberal approach to reform. Broadly speaking, two camps emerged: neoliberal proponents of the 'Washington Consensus' and their critics.

The neoliberal camp, which dominated the transition process and consisted of both foreign advisors as well as domestic reformers, advocated the primacy of the policy trio of the Washington Consensus: liberalization, stabilization, and privatization. The key was to get the price signals right by eliminating distortions, which called for a quick and comprehensive 'shock therapy' package consisting of liberalizing and anti-inflationary measures. Privatization was to accompany the shock therapy package as quickly as possible because it was seen as a tool to "create market structures to encourage private businesses" (Åslund, 2007: p. 146).

Rapid reforms were considered advantageous for several reasons. First, they would take advantage of the unique 'window of opportunity' offered by the short-lived popular enthusiasm for reform (Balcerowicz, 1995). Given that under shock therapy the material situation of the population deteriorates quickly before it improves, the reformers saw the need to carry out painful reforms before the losers of the reform process had the time to organize to reverse them (Przeworski, 1991).

Rapid reforms would offer two additional benefits that were especially salient in the case of privatization: they would make the reforms irreversible, and they would create a constituency of winners who would support the continuation of the reform efforts.

The dominant neoliberal paradigm was criticized on several counts. First, proponents of gradual reforms argued that these would result in less social dislocation as the poststabilization recession would not be as deep (Przeworski, 1991). Second, institutional economists, sociologists, and political scientists argued that the neoliberals presented a much oversimplified vision of market economy. The frequently cited quote by Jeffrey Sachs that "[m]arkets spring up as soon as central planning bureaucrats vacate the field" (1994: p. xiii) became symbolic of the neoliberals' catchy but simplistic invisible hand approach that ignored the centrality of institutions in market exchange. Some of the fiercest critics even leveled the charge of 'market Bolshevism' as they lambasted the neoliberals for assuming that they could simply use shock therapy to 'install' institutions while ignoring the specificity of local conditions and preexisting institutions (Stiglitz, 1999; Murrell, 1992). Rather, they argued that institutions take time to evolve and are socially embedded. Hence, they were skeptical that any reform blueprint could simply be imposed on what was the supposedly blank slate of transitioning countries. Rather, as a prominent sociologist put it, capitalism was being built not on but *with* the ruins of socialism (Stark, 1992). The local political and social conditions, and the institutions left over from the previous system would determine both which reforms would be adopted and how they would be implemented.

Economic Reform Pathways

Although the neoliberal blueprint became the dominant reform paradigm in policy circles, the extent and the mode of its implementation were rooted in the domestic political conditions. Space considerations preclude an in-depth discussion of individual countries, but the broad brushstroke presentation of reform pathways illustrates the diversity in the paths chosen. Poland, facing significant macroeconomic imbalance, became the path-breaker with its bold 1990 shock therapy recipe, named the Balcerowicz Plan after Poland's finance minister. Due to internal political disagreements over the choice of privatization method for large SOEs, the so-called 'large privatization' was left until later, much to the disappointment of the international financial institutions, and proceeded in a gradual fashion. At the same time, research has shown that, faced with hard budget constraints, SOEs adjusted much better to market conditions than expected (Pinto et al., 1993).

A year later, Czechoslovakia followed with its own shock therapy plan, accompanied by an innovative voucher privatization scheme aimed largely at small and medium enterprises, with sales of only minority stakes in large enterprises. By contrast, Hungary chose the gradualist path, but the rising macroeconomic instability eventually called for a shock therapy package, the so-called Bokros Package (also named after Hungary's finance minister), which was implemented in 1995. The Hungarians were also unique for their initial ability to attract relatively high levels of foreign capital as part of their privatization efforts.

The East German pathway was different altogether as East Germany became absorbed by the much bigger West German economy in the aftermath of the German reunification in 1990 and could take advantage of both West German market and state institutions, as well as of their significant transfer payments. These were very much in order as the East German economy collapsed in the aftermath of a currency union between East and West Germany, which although quite advantageous for the East Germans, had the side effect of making East German industry uncompetitive virtually overnight (Stokes, 1993: p. 185). The reforms in Romania and Bulgaria, on the other hand, proceeded in fits and starts. In both cases, initial stabilization and privatization efforts were significantly compromised by the heavy presence of the communist-era managerial interests that still had the ear – and the purse – of government officials (Ganev, 2007; Gallagher, 2005; Roland, 2000).

Privatization as an Example of Economic Reform Challenges

The preceding account illustrated the diversity of the extent and timing of both macroeconomic stabilization policies and privatization. The privatization experience offers insight into the unique challenges of reforming a postcommunist economy, but also into the importance of regulatory institutions. The challenges of privatization were well summed up by a quip Janusz Lewandowski, the Polish Minister of Privatization, made in 1990: "[p]rivatization is when someone who doesn't know who the real owner is and doesn't know what it's really worth sells something to someone who doesn't have any money" (Stokes, 1993: p. 196). Therein lay the two key problems encountered by reformers. First, given the legacy of central planning, it was difficult to appraise the assets that were being offered for sale prior to privatization. Second, buyers were not easy to come by.

Foreign investors had the money and the know-how to turn SOEs into profitable businesses, but they were initially wary of entering these markets for several reasons. First, they faced uncertainty about the unprecedented reform process, and their unease was not allayed by the limited domestic support for sales of the 'family jewels' to foreigners in the region. Second, foreign investors feared investing in enterprises with excess production capacity, obsolete equipment, overstaffed workforce, and assets completely unrelated to production in their care (company housing units, medical clinics, kindergartens, recreational facilities, etc.). In other words, they needed some form of preprivatization restructuring by state authorities before committing to some of the bigger enterprises. State-led restructuring, and reform (as opposed to excising) of state institutions more generally, were certainly not at the top of the initial reform agenda (King and Sznajder, 2006).

Domestically, individuals with significant financial assets were largely communist party cadres, especially in Poland and Hungary. There, the communist-era liberalization efforts led to significant business opportunities for well-connected, entrepreneurial, communist party members and state security apparatus officials. As a result, many reformers and the population at large, were skeptical of capital sales to individuals. In an

effort to popularize privatization, but also to resolve the problem of equity – after all, state property belonged to everyone under the communist regime and therefore, everyone should now benefit – voucher privatization became a solution to the privatization issue touted by the neoliberal reformers. There were several schemes in CEE. Unlike the Russian one where the workers and managers obtained free shares in their own enterprises, in the Czech (and initially Slovak) variant, all citizens were eligible for buying voucher booklets which they could then use in order to bid on enterprises they preferred. In the Polish case, the eventual, relatively small-scale voucher program provided for citizen stakes only in the holding companies that managed shares of the enterprises they were assigned.

Voucher privatization meant that even though the state would have to forego income from the sale of these enterprises, state assets would get into the hands of private owners, who would, in theory, do a better job of restructuring them than the state could. According to the Czech design, by having to bid on enterprises they were interested in, voucher holders would become personally invested in the success of the reform process while receiving a crash course in market economics. The expected fragmented ownership structure was seen as a temporary phenomenon that would get consolidated over time. The voucher experiment, however, had disappointing results, though its proponents steadfastly defend its achievements. The fiasco of the Russian program of distributing the vouchers to enterprise insiders is well known. However, even the touted Czech program, very skillfully marketed at home and abroad by the Czech Prime Minister Václav Klaus, was gravely plagued by inadequacies of corporate governance. In the Czech case, investment companies that sprung up overnight, bought up vouchers and through cross-ownership of banks and enterprises siphoned off, or 'tunneled out' assets. This, in turn, contributed to the Czech banking crisis of 1997 (Orenstein, 2001).

Voucher privatization remains controversial. Its detractors argue that it led to massive asset stripping and theft. Even in its mildest incarnation in Poland, where the abuses were curtailed by better regulation, the resulting holding companies did not engage in much, if any, investment and had a very passive ownership role (King and Sznajder, 2006). Supporters of voucher privatization, however, insist that poor corporate governance notwithstanding, its results were better than the alternatives (Åslund, 2007: p. 180).

Illustrating the highly political nature of privatization, each CEE state followed its own path to private ownership. While the Czech Republic raced ahead with voucher privatization and also with direct sales to domestic owners, the Hungarian government mostly relied on direct sales and investment tenders, often to foreign investors who felt secure about the stability of the reforms in the country. In Poland, political disagreements over privatization methods led to reliance on bottom-up privatization which consisted of employee and management buyouts, in addition to the eventual combination of limited voucher privatization and direct sales. In Slovakia and Romania, management and employee buyouts also constituted the primary form of privatization while in Bulgaria, privatization took the form of sales (often corrupt) to outside owners. Both in Slovakia and Bulgaria, voucher privatization represented secondary forms of privatization (Åslund, 2007: p. 156).

While the initial round of economic reforms focused on macroeconomic stabilization and privatization, state and regulatory institutions were not on the forefront of the reform agenda. Yet, the importance of state capacity for economic reform, including privatization, became increasingly recognized throughout the reform process (Åslund, 2007: p. 179). Profound transformation of state institutions, including those responsible for regulation, became part and parcel of the EU accession effort, which took off in 1998 following the opening of EU membership negotiations with Hungary, Poland, the Czech Republic, Estonia, and Slovenia. In 2000, these negotiations were extended to Romania, Bulgaria, Slovak Republic, Latvia, and Lithuania.

The Challenge of EU Accession

As stated at the outset of this piece, the idea of the 'return to Europe' was the guiding principle of political and economic reform in CEE. However, EU membership was not immediately on the agenda as EU countries opted for association agreements, also known as Europe Agreements, which established far-reaching trade liberalization with the CEE countries. These were signed between 1991 and 1993. It soon became apparent that these were not enough in the face of clear desire of CEE states for EU membership. Post-Cold War moral imperative on the part of Western European states notwithstanding, the EU realized that eastern enlargement would solidify reforms and therefore enhance the stability of the entire continent. It would also offer new economic opportunities for Western European players on emerging markets to their east.

The institutional gulf dividing the EU from the reforming countries made it apparent that any accession process would be long and arduous. After all, European integration had been ongoing since the 1950s and the postcommunist countries were facing a continually transforming moving target, which was already bound by the body of the so-called *acquis communautaire*. The latter represented a nearly 50-year worth of directives, regulations, and court decisions that were binding on all EU member states (Cameron, 2003). When the EU made the decision to open up membership to CEE states at the 1993 European Council meeting in Copenhagen, it produced the so-called 'Copenhagen criteria,' according to which the new states had to meet three conditions in order to become EU members. The first objective was the stability of institutions guaranteeing democracy, the rule of law, human rights, and respect for and protection of minorities. The second objective was economic and pertained to the existence of a functioning market economy that had the capacity to cope with competitive pressure and market forces within the Union. Third, the new members would have to accept the Community *acquis* and have the ability to take on the obligations of membership, including commitment to adopting the euro as their currency (European Commission, 'Conditions for membership,' http://ec.europa.eu/enlargement/policy/conditions-membership/index_en.htm).

To emphasize the scope of this challenge, the countries that 5 years earlier were still guided by the central planners and

(most of which) had virtually no private property of note, were told that in order to join the EU, they would have to operate on the same market and according to the same rules as the seasoned players coming from such industrial and economic powerhouses as Germany. More details of what it meant to be a functioning market economy and be able to operate on the EU market were spelled out in a 1995 White Paper produced by the European Commission (the equivalent of the executive branch in the EU) that divided the body of EU legislation into subject-related chapters. When the accession negotiations opened with the first tier of reformers in 1998 and with the second one in 2000, these 31 chapters became the basis for negotiations with the Commission.

At the core of EU membership negotiations was the so-called harmonization of national laws with EU law. For each of the thousands of measures comprising the *acquis communautaire*, accession states had to present evidence of how the EU rules would be implemented into their own law at the time of accession. The European Commission monitored the process and evaluated progress made in its annual 'report cards' called the regular reports. Thus, under the Commission's watchful eye, the accession countries were being re-created as states (Cameron, 2003: p. 25). At the same time, although many portray the Commission as imposing its rules on the new member states, some scholars argue that the Commission actually mobilized private and public actors and constructed a nonhierarchical mode of governing (Bruszt and Langbein, 2014). Thus, the new EU institutional requirements became embedded in the national and local conditions.

The Effects of EU Accession and Implications for Capitalism in Europe

Overall, the EU accession process had several profound effects on accession states' economic reform trajectory. On the one hand, among the leading reformers, it strengthened the domestic proreform coalitions but also pushed for reform in areas that were politically difficult (Drahokoupil, 2009; Sznajder, Lee forthcoming). At the same time, among transition laggards, such as Romania and Bulgaria, it tipped the political scale in favor of reformers (Vachudova, 2005; Bohle and Greskovits, 2012).

EU membership thus had a profound effect on the trajectory of market reforms, and it served as a powerful signaling device for foreign investors. After 1994, the initial Hungarian advantage in attracting foreign capital diminished, as the other transition leaders, such as Poland and the Czech Republic, demonstrated their commitment to reform. Moreover, they actively attempted to attract more foreign investment. In fact, the EU effect was so strong that economic studies "underestimated the volume of FDI inflows as the volume of foreign capital stock more than tripled between 1998 and 2004" (Medve-Bálint, 2014: p. 38).

The EU accession process, moreover, also strengthened domestic coalitions that sought to use 'second stage reforms' such as pension privatization, introduction of flat taxes, and cutting of corporate tax rates as a positive signal to attract foreign investors. It should be stressed that with the exception of pension privatization, these reforms went above and beyond what the EU, or even the international financial institutions, such as the International Monetary Fund, advocated (Appel and Orenstein, 2014). Thus, the neoliberal reform trajectory continued beyond EU accession.

The large quantities of foreign investment had profound developmental implications for these countries. They yielded economies that were more transnational than those of the original EU member states, and in which the 'old' EU member states dominate by controlling about two-thirds of the foreign-controlled production (Medve-Bálint, 2014: p. 36).

The banking sector is an extreme, though telling example of the extent of foreign capital in the region. The asset share of foreign-owned banks in national banking systems equals or exceeds 80% in (in rising order): Bulgaria, Romania, Czech Republic, Slovakia, Croatia, Lithuania, Bosnia and Herzegovina, Albania, and Estonia (in the latter, it approaches 100%). In Poland and Hungary, the shares hover about 70 and 65%, respectively. This is in stark contrast to the Western European EU member states. For example, in Spain, Netherlands, France, and Italy the same share comes out to less than 10% (European Bank for Reconstruction and Development, 2012: p. 48). In fact, the heavy overall presence of foreign capital in CEE even led to the economic system in the region being labeled as 'dependent market economies,' where foreign capital and its imported institutions represent the principal coordinating mechanism in the economy (Nölke and Vliegenthart, 2009).

Even though these countries' economies are transnational, some scholars highlight existing differences in their socioeconomic attributes. Dorothee Bohle and Béla Greskovits, for example, argue that putting the label of "dependent market economies" does not reflect the salience of diverse kinds of foreign investment, whether financial capital or investment in the country's industrial production base (Bohle and Greskovits, 2012). Second, various countries struck different types of balance between following the precepts of neoliberalism and attempting to preserve social cohesion and political legitimacy. As a result, countries can be divided into the neoliberal type in the Baltic states and – belatedly – Bulgaria and Romania, the embedded neoliberal type among the so-called Visegrád countries composed of the Czech Republic, Hungary, Poland, and Slovakia, and the neocorporatist type in Slovenia.

Thus, whereas the Baltic states combined radical market reforms with very limited compensation for the heavy social costs of transformation, the Visegrád states searched for social cohesion in the face of neoliberal reforms. Moreover, the former largely shut out organized social groups from democratic decision-making, but the latter were more inclusive – if less efficient – in their reform efforts. Finally, Slovenia stands apart as its marketization strategy was the least radical, most compensatory, and least exclusive, as it followed the principles of democratic corporatism that brought state, business, and labor together in search of compromise solutions (Bohle and Greskovits, 2012).

The Impact of the Financial Crisis and Future Challenges

The new, transnational capitalism established in CEE faced its biggest challenge yet during the onset of the global financial crisis. It hit the region early, prior to the emergence of Greece's financial troubles, and dealt a harsh blow to CEE economies.

However, the effect of the crisis was not uniform, and it was very much tied to the type of foreign investment – whether through the financial sector or through manufacturing and exports – that dominated in the countries at hand (Bohle and Greskovits, 2012; Myant and Drahokoupil, 2012). As scholars have shown, countries belonging to the neoliberal type were, among other external imbalances, particularly vulnerable to 'financialization' or the dependence on foreign finance to support bank loans, which left them helpless when foreign capital pulled out. On the other hand, the embedded neoliberal regimes (and, to lesser extent, Slovenia) were less exposed to capital flight, as foreign direct investment was more concentrated in the export-oriented manufacturing sector. These sectors were hit hard when foreign demand fell, but they bounced back rather quickly due, in no small part, to the fairly strong German economic growth in 2010 and 2011 (Bohle and Greskovits, 2012; Myant and Drahokoupil, 2012). Hungary stands out in this regard, as it was very negatively affected by the high share of foreign currency-denominated loans to private households.

The transition economies have weathered the economic and financial crisis quite well in that the economic institutions in place proved resilient in the face of tremendous pressure. At the same time, the crisis had several effects on the region. First, the enthusiasm for neoliberal reforms waned and some of them were reversed. Overall, the level of policy liberalization fell, compared to the one from before the crisis (Jacoby, 2014: p. 68). The extreme case is that of Hungary, where the populist government of Viktor Orbán's FIDESZ party has thrived on antiforeign capital rhetoric and took punitive actions. These included forcing foreign banks to covert foreign currency-denominated mortgage loans to forints at a disadvantageous fixed rate. Second, some neoliberal policy reversals took place, mainly in the area of pension privatization. In Hungary, pension privatization was reversed, while the Polish government markedly reduced the amount of money allocated to the private pension accounts, thereby increasing the share of state-held assets (Appel and Orenstein, 2014: p. 26).

Third, the crisis also led to questioning of the desirability of membership in the Eurozone, which was one of the commitments made during EU accession, albeit this response was not uniform. The small Baltic states that were particularly vulnerable to boom and bust cycles sought quick membership in the Eurozone, with Estonia joining in 2011 and Latvia in 2014. Lithuania is set to join in 2015. However, in Hungary, joining the Eurozone was placed on hold indefinitely, only in part because the government is not close to meeting the strict membership criteria. At the same time, the Czech Republic has encountered domestic skepticism of membership, as did Poland. In Poland, the initial postcrisis skepticism has most recently been replaced by clear signals from the current government, the President, and the head of the Polish National Bank that Eurozone membership would further embed Poland in the Western institutional framework. This is seen as an additional security measure in the face of Russia's increasingly hostile stance vis-à-vis the EU.

Reflecting on the progress made over the past quarter century, the countries in CEE have made incredible strides in terms of integrating with the Western economic and political order. The reforms have certainly been stimulated along the way by the process of European integration. Yet, although the economic divide between the 'West' and 'East' has narrowed, the gap resulting from the historical underdevelopment of the region has not yet been bridged. Further progress will require overcoming deeply ingrained obstacles such as corruption but also attracting investment that will spur innovation (Epstein, 2014: p. 31).

See also: Russian and Post-Soviet Studies: Culture; Russian and Post-Soviet Studies: Education; Russian and Post-Soviet Studies: Environment; Russian and Post-Soviet Studies: Politics; Russian and Post-Soviet Studies: Society.

Bibliography

Appel, H., Orenstein, M., 2014. The Triumph and Crisis of Neoliberalism in Eastern Europe: Competitive Signaling and Policy Interdependency. Paper presented at the Annual Meeting of the American Political Science Association, Washington, DC.

Åslund, A., 2007. How Capitalism Was Built: The Transformation of Central and Eastern Europe, Russia, and Central Asia. Cambridge University Press, Cambridge.

Balcerowicz, L., 1995. Socialism, Capitalism, Transformation. Central European University Press, Budapest.

Bohle, D., Greskovits, B., 2012. Capitalist Diversity on Europe's Periphery. Cornell University Press, Ithaca.

Bruszt, L., Langbein, J., 2014. Anticipatory Integration and Orchestration: The Evolving EU Governance of Economic and Regulatory Integration during the Eastern Enlargement. Paper presented at the Annual Meeting of the American Political Science Association, Washington, DC.

Cameron, D.R., 2003. The challenges of accession. East European Politics and Societies 17 (1), 24–41.

Drahokoupil, J., 2009. Globalization and the State in Central and Eastern Europe: The Politics of Foreign Direct Investment. Routledge Series on Russian and East European Studies, London.

Epstein, R., 2014. Overcoming 'economic backwardness' in the European Union. Journal of Common Market Studies 52 (1), 17–34.

European Bank for Reconstruction and Development, 1994–2013. Transition Report. European Bank for Reconstruction and Development, London.

European Bank for Reconstruction and Development, 1999. Transition Report 1999: Ten Years of Transition. European Bank for Reconstruction and Development, London.

European Bank for Reconstruction and Development, 2012. Transition Report 2012: Integration Across Borders. European Bank for Reconstruction and Development, London.

Gallagher, T., 2005. Modern Romania. New York University Press, New York.

Ganev, V., 2007. Preying on the State: The Transformation of Bulgaria after 1989. Cornell University Press, Ithaca.

Jacoby, W., 2014. The EU factor in fat times and in lead: did the EU amplify the boom and soften the bust? Journal of Common Market Studies 52 (1), 52–70.

King, L., Sznajder, A., 2006. The state-led transition to liberal capitalism: neoclassical, organizational, world systems, and social structural explanations of Poland's economic success. American Journal of Sociology 112 (3), 751–801.

Kitschelt, H., Mansfeldova, Z., Markowski, R., Tóka, G., 1999. Post-communist Party Systems: Competition, Representation and Inter-party Cooperation. Cambridge University Press, Cambridge.

Kornai, J., 1980. Economics of Shortage. North-Holland Publishing Company, Amsterdam.

Mitra, P., Selowsky, M., 2002. Transition, the First Ten Years: Analysis and Lessons for Eastern Europe and the Former Soviet Union. World Bank Publications, Washington, DC.

Myant, M., Drahokoupil, J., 2012. International integration, varieties of capitalism and resilience to crisis in transition economies. Europe-Asia Studies 64 (1), 1–33.

Medve-Bálint, G., 2014. The role of the EU in shaping FDI flows to east central Europe. Journal of Common Market Studies 52 (1), 35–51.

Murrell, P., 1992. Conservative political philosophy and the strategy of economic transition. Eastern European Politics and Societies 6 (1), 3–16.

Nölke, A., Vliegenthart, A., 2009. Enlarging the varieties of capitalism: the emergence of dependent market economies in East Central Europe. World Politics 61 (4), 670–702.
Orenstein, M., 2001. Out of the Red: Building Capitalism and Democracy in Post-communist Europe. University of Michigan Press, Ann Arbor.
Pinto, B., Belka, M., Krajewski, S., 1993. Transforming State Enterprises in Poland: Evidence on Adjustment by Manufacturing Firms. Brookings Papers on Economic Activity 24 (1), 213–270.
Przeworski,, A., 1991. Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America. Cambridge University Press, Cambridge.
Roland, G., 2000. Transition and Economics: Politics, Markets, and Firms. MIT Press, Cambridge.
Sachs, J., 1994. Poland's Jump to the Market Economy. MIT Press, Cambridge.
Stark, D., Bruszt, L., 1998. Postsocialist Pathways: Transforming Politics and Property in East Central Europe. Cambridge University Press, Cambridge.
Stark, D., 1992. Path dependence and privatization strategies in East Central Europe. East European Politics and Societies 6 (1), 17–53.
Stiglitz, J., April 28–30, 1999. Whither reform? Ten years of the transition. Keynote Address at the 1999 Annual Bank Conference for Development Economics, Washington DC, 1–34.
Stokes, G., 1993. The Walls Came Tumbling Down: The Collapse of Communism in Eastern Europe. Oxford University Press, New York; Oxford.
Sznajder Lee, A. Transnational Capitalism in East Central Europe's Heavy Industry: From Flagship Enterprises to Subsidiaries, University of Michigan Press, Ann Arbor, forthcoming.
Vachudova, M.A., 2005. Europe Undivided: Democracy, Leverage and Integration after Communism. Oxford University Press, Oxford.
Wolchik, S.L., 1994. The politics of ethnicity in post-communist Czechoslovakia. East European Politics and Societies 8 (1), 153–188.

Relevant Websites

http://www.ebrd.com/pages/homepage.shtml – European Bank for Reconstruction and Development.
http://www.ebrd.com/pages/research/publications/flagships/transition/archive.shtml – European Bank for Reconstruction and Development. Transition Report archive.
http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/ECAEXT/0,contentMDK:22587588~pagePK:146736~piPK:146830~theSitePK:258599,00.html – European Bank for Reconstruction and Development and the World Bank, The Business Environment and Enterprise Performance Survey (BEEPS).
http://ec.europa.eu/enlargement/policy/from-6-to-28-members/index_en.htm – European Commission, Directorate General for Enlargement.
http://ec.europa.eu/enlargement/index_en.htm – European Commission, Enlargement.
http://ec.europa.eu/regional_policy/what/future/index_en.cfm – European Commission, Regional Policy, Cohesion Policy 2014–2020.

Postsocialist Societies

Maya Grekova, Sofia University 'St. Kliment Ohridski', Sofia, Bulgaria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'postsocialist societies' is applied to societies that were formerly identified as socialist and that have since renounced the socialist course of development since the 1989–90 changes in government. This classification is practically void of meaning as it defines these societies vis-à-vis their past – as societies in transition, which regard themselves, and are regarded, as being in a process of democratization and marketization, and building a civil society. Differences rooted in the history of the transformation of the respective societies into socialism, as well as in their history as socialist societies, have appeared in the course of this transition. Notwithstanding all differences in the pace of transition and in the current or future hardships of postsocialist societies, several common problems attract the researcher's attention.

Definition

The term 'postsocialist societies' itself raises a serious problem: Its use implies that societies that were considered (called themselves or identified as) socialist, and that have renounced the 'socialist course of development' since the 1989–90 changes in government, have no distinctiveness of their own but still are defined by their past. These societies are classified in one and the same group, defined in one and the same way from the perspective of their past – their existence as 'socialist societies' – which they have renounced. On the other hand, this definition from the perspective of their past suggests that they are presumed to be societies in the course of self-identification – seen as being in transition from totalitarianism to democracy, from centrally planned to market economy – which have embarked on the arduous course toward building civil society.

This raises a serious question of the possible criteria by which a concrete national society would shed its 'postsocialist' label; that is, ultimately would have defined itself and be defined from the perspective of the present rather than of the past. At what stage does the present existence of the respective society provide sufficient grounds for regarding it as a society with sustained development based on values asserted in it (and shared by) its citizens?

Changes in the form of government are a necessary but not a sufficient prerequisite for radical restructuring of social life as well as for significant and sustainable changes in the concepts of the world, human relations, and individual actions. That is precisely why the term used is 'postsocialist societies' rather than 'postsocialist states.' The form of state government may or may not be socialist, whereas the declared renunciation of the 'socialist course of development' in itself does not free a society of the values, ideas, and expectations accumulated over the years, nor of the established relations and groups, and the very attitudes to and mechanisms of their establishment. At the same time, the course of self-identification of the 'postsocialist societies' is indefinite – their point of departure is known, but one cannot know for sure what they ultimately will become.

This is a generalization only, because the majority of 'postsocialist societies' (those in Europe) are expected to embrace precisely democratic values, a market economy, and civil society as the pillars of self-identification. Most of these societies have not merely declared this but are making real progress in this respect. They apparently are aware that they will no longer be labeled as 'postsocialist societies' once they complete the process of accession to the European Union (EU). In other words, for those national societies, the criteria of recognition of a completed process of self-identification are the same as the criteria of fully fledged EU membership, which means that they already have chosen a course of development and that their development is predetermined – all they have to do is attain the ultimate objective. Postsocialist societies outside the purview of the EU, of course, face different problems.

Differences Stemming from the Past

The term 'postsocialist' creates a false impression of uniformity of the societies in question (even though this definition is indefinite itself) and, respectively, of uniformity in the process of departure from their previous development as socialist. This idea is based largely on the previous term 'socialist societies' that has the ideological function of hiding or belittling the internal distinctions of the respective 'socialist societies,' as well as of rendering as entirely insignificant, from the perspective of their existence as 'socialist,' the differences in their previous development and the concrete process of their 'transformation' to socialism. They are 'socialist societies,' and all details of the history of their transformation to socialism and of the national specificity of their existence thus are deemed insignificant. Yet it is precisely those formerly 'insignificant' differences (declared and considered insignificant, but to a large extent also virtually insignificant from the perspective of 'the general line' of development, of the rallying of the respective societies into the 'socialist camp') that largely determine the form of renunciation of the 'socialist course of development' constituting significant accelerators or impediments (Fowkes, 1999; Nagle and Mahr, 1999; Offe, 1996; Rose et al., 1998). Generally speaking, these national variations of the road to state socialism are significant determinants of the national specificity of the existence of the concrete societies as 'postsocialist,' as well as of the pace and consistency of the transition period, of the ways of reaching national consensus on the parameters of the change, and of the actual concept

and definition of 'national consensus' (for a consistent pursuit of such an understanding, see Brown, 2009; Burawoy and Verdery, 1999; Priestland, 2009).

Forgetting the history of the formation of a particular nation-state and the way in which it 'embarked on the road to socialist development' – or failing to take into consideration the influence of national specificity on the existence of the respective society as socialist – usually results in overemphasizing the impact of international institutions (economic and financial, as well as political) whose actions are guided by economic and geopolitical interests, as well as of the role of the personality of the respective political leaders (in the concrete 'postsocialist society'). Thus, the changes and their pace in 'postsocialist societies' (especially in Europe) prove directly and crucially dependent on foreign interests or the individual abilities of politicians. Consequently, individual and group responsible action proves impossible or, rather, pointless. Even if it were possible, its ineffectiveness is justified readily by counteracting external forces or inactive internal factors.

Social scientists, among others, are accountable for the formation of explanatory models centered on these two variables. To explain the transition period and its problems, they deviate from their scientific position under pressure from their partisan allegiance; concentrate on the present, forgetting history; resort to biased selection of historical facts to legitimate their interpretation of events; build a macroexplanatory model of transition in general, ignoring national differences; or cite separate facts or data from opinion polls without the historical-theoretical framework requisite for their interpretation.

The production and public presence of such 'explanatory models' of the problems of the transition period generate further tensions in the postsocialist society. First, they are explanatory and deliberately created for the purpose of introducing ultimate clarity among ordinary citizens about current developments. Second, they are even more convincing on the strength of their authors' scientific prestige. Third, they are qualitatively heterogeneous – rather than just different – interpretations, and at the same time each one of them claims to be the only truthful interpretation. This lopsided way of 'understanding' developments, history, and the origin of problems in social life, as well as events and political actions, proves to be yet another factor (along with political allegiance and social position) that antagonizes individuals and groups.

Present Similarities or Key Problems of Transition

Two ranges of problems may be identified: Problems associated with identity and problems associated with social interaction. Although they are not confined to postsocialist societies only, here these problems have acquired special significance and a specific form. They may also be generalized as 'political,' 'economic,' or 'social' from a macrosociological perspective on social reality, on the processes occurring within the latter, and on their specific intertwining. The microsociological perspective proposed is based on the understanding that the restructuring of a society that has renounced socialist principles and norms (or, generally speaking, any society in the course of self-identification as something that is qualitatively different) proceeds from the presumption of the major significance of the changes in the ways of thinking and action and is impeded seriously by the continuing effect of previous models of thinking of the world and the others, patterns of behavior, and mutual relations of individuals and groups.

Problems of Identity

These problems concern apparently quite heterogeneous phenomena that, however, are experienced by the individual or group in a similar way as problems of self-identification: from 'nationalism,' which has become indicative of postsocialist societies, to 'unemployment' and 'poverty,' which rarely are treated as problems of identity (defined foremost as social and economic problems).

The preoccupation with national and ethnic identity, which apparently has divided (albeit in quite different ways) the postsocialist societies, stems from two specific features of the ideology upheld in socialist society: National and ethnic specificities within the boundaries of a separate state and between the states from the 'socialist camp' are insignificant – the significant self-identification is 'socialist citizens' (which does not mean that ethnocultural differences have been forgotten within the separate states or that they never were an undisclosed reason for actual social inequality). Accordingly, the significant difference is associated directly with significant self-identification – the localization of everything 'strange' there, that is, beyond the 'iron curtain.'

The disintegration of 'the camp' and the fall of 'the wall' dividing the two worlds, inevitably rejecting the validity of significant self-identification and difference, generated the need for 'national' self-identification. Whereas the question 'Who are we?' focuses the attention inward on the 'national community' and back in history, on the national community's existence and formation before its 'depersonalization,' the attempt at defining 'us' (who are united within a state) has exposed differences and has led to intolerance and discrimination against different ethnic and religious minorities (Opalski, 1998; Rupesinghe et al., 1992; Tismaneanu, 1998). This process, common to all postsocialist societies, has acquired qualitative specificity (largely determined according to the constitution of the respective 'socialist societies'): from violent disintegration of multinational states (the Soviet Union and Yugoslavia), to the peaceful division of Czechoslovakia, to the more or less conflict-ridden confrontations between 'national majority' and 'ethnic–national minorities' in Bulgaria, Romania, Hungary, and elsewhere. Yet the problem of emphasizing internal differences (by origin, language, religion) has remained common despite their qualitative specificity: Internal differences that are thought to be significantly distinctive of people to the point of their existence within more or less closed groups, and perception of respective individuals as significantly predetermined by their 'belonging' to those groups. Regardless of the self-perception of the respective individuals, 'the others' ascribe identity to them and treat them (engage in or avoid relations with them) from the perspective of this identity.

The problematic process of identification with European (Western) values unfolds against the background of these 'internal' problems of self-identification and ascription of identity. This process is problematic mainly because of the

complex and contradictory combinations of three commonplace ideas: 'We are European,' 'We have had a tragic destiny (not without the involvement of the West), now the West should help us become like them,' and 'Once we are admitted to the EU, we shall become Europeans.' Although the images and expectations underlying the apparently unambiguous images of Europe and the West and the notions that 'we are European' and 'we shall become Europeans' are quite different, their focus is arguably on 'becoming like them.' Conversely, the principal dimension of the process of 'becoming' is the expected well-being.

This is not accidental. And it is directly associated with another aspect of the identity crisis in postsocialist society – unemployment and poverty (Kornai, 1995, 1997; Michta and Prizel, 1992). These two related but different problems are no doubt indicative of the scope of the economic crisis, but their personal aspect concerns the deepest layers of individual identity, largely because the status of being unemployed is unfamiliar, virtually absent in the experience of the 'socialist individual,' whereas poverty is a process that individuals who have found themselves in a postsocialist society do not associate with democracy and a market economy.

Problems of Social Interaction

The change (or, rather, the orientation toward change) in the principles of organization of social life triggers tensions at all levels of social interaction. At the level of the individual, these tensions take the form of a large-scale crisis in human relations. At the level of institutionalized interactions, they are manifested in 'corruption' of significant proportions.

The 'crisis in human relations' could be reworded, too, as a moral and a political crisis as well as a problem of interethnic relations. For the everyday person, however (insofar as we can talk about an everyday person in a world in which what is taken for granted is always questioned), this problem appears as tensions in relations, as a failure to understand the 'other' (or vice versa), avoidance of certain others, isolation of and from strangers, and so forth. That problem covers the whole range from the closest social environment – family, relatives, friends – to neighbors and colleagues, to functional relationships with anonymous others.

The problem generally defined as a 'crisis in human relations' actually results from the intertwining and superimposing of diverse manifestations and echoes of political, economic, and cultural values and other changes on the macrosocial level. The realized opportunity of being different in our political preferences and allegiance and in our economic status, as well as of choosing between different cultural models and of adhering to different values, inevitably has destroyed established models of family, friendly, neighborly, business, and interethnic relations. Yet new models may appear only as a result of a long and complex process of construction, mutual adjustment, and assertion. They are not something that can be borrowed (imported, grafted) or something that becomes valid simply because it is (or appears to be) better.

Similar to nationalism, corruption has become emblematic of postsocialist societies. And just like nationalism, corruption is not an unfamiliar phenomenon in Western societies. Moreover, corruption is an affliction of democratic society insofar as the latter has sufficiently large opportunities for responsible decision making, instead of the strictly prescribed rules for any action that are typical of any totalitarian society and relieve its members of responsibility. Yet the opportunity for responsible decision making is also an opportunity for irresponsible action. In the postsocialist society, the complex and contradictory intertwining of still-effective old schemes of control and decision making, and the deficiency of new norms (both on the institutional level of laws, regulations, ordinances, and so on and on the level of individual and group consciousness, including role models, values, and notions) turn corruption into an epidemic. The affliction acquires huge proportions due to, inter alia, the internally contradictory perception of corruption: on the one hand, it is seen as an inevitable relation in a situation of transition, of reformulating and reregulating of institutions, social positions, rules, and so forth, which are invisibly present in the social sphere and are manifested in status, security, possessions, and wealth of those who have effected the replacement. On the other hand, corruption is embodied in the bribe-taker, insofar as bribe-takers are those who have rare resources, by rule of power, which many other persons would like to use. The bribe-giver, however, is simply one of many: Bribe-givers do not control the rules of replacement, the way of negotiations, the size of the bribe to the bribe-taker, they simply are the person who has happened to be in this position and, in this sense, anyone may end up in this position. Thus, the bribe-taker is regarded as an irresponsible power-holder and the bribe-giver as anyone of us forced by the circumstances and therefore not responsible for their involvement in relations of corruption.

From Postsocialist Society to EU Membership

The general orientation that all European postsocialist societies indisputably follow is democratization of social life, market economy, freedom of speech and association, multiparty political system, and human and minority rights. This orientation was anticipated to lead the respective states to full membership in the EU, which, once achieved, would secure an end to their postsocialist, transitory, and undetermined state of being. In different years, the European postsocialist states declared their wish to join the EU, engaged in accession negotiations, and acquire the status of candidates for membership. By June 2012, the majority of these countries already had been accepted as full member states of the EU. The foundation for the EU enlargement was laid by the Copenhagen criteria adopted by the EU in 1993, which should be met by the candidate countries to make them eligible for full membership in the EU (see Copenhagen criteria, at http://www.europarl.europa.eu/enlargement/ec/pdf/cop_en.pdf).

Former Postsocialist Societies – New Members of the EU

On 1 May 2004, the Czech Republic, Slovakia, Poland, Hungary, Estonia, Latvia, Lithuania, and Slovenia (together with Malta and Cyprus) joined the EU. It was assessed that they had implemented the Copenhagen criteria. Separated from this group were Bulgaria and Romania, which were evaluated to not

have covered all criteria and, respectively, their accession was postponed and took place on 1 January 2007.

The successful implementation of the criteria for integration in the EU does not mean that the differences among the postsocialist states, as well as between them and the old member states, have been erased. During the socialist regime, the differences among the socialist states were presented as insignificant from the viewpoint of 'building socialism' and 'the bright communist future.' Even in the socialist era, however, there were emblematic differences: The Hungarian revolution of 1956 and the Prague Spring of 1968 were examples of publicly articulated resistance against the communist regime, which were suppressed by the military force of the Warsaw Pact; Romania under Nicolae Ceaușescu (1918–89) refused to join the bloc's military intervention in Czechoslovakia in 1968 but endured his extremely repressive totalitarian dictatorship in the following decades; Hungary, Czechoslovakia, and the German Democratic Republic (GDR) achieved considerable economic development and constituted a 'showcase of socialism'; the Socialist Federal Republic of Yugoslavia dominated by Serbia under the leadership of Josip Broz Tito (1892–1980) maintained a relatively autonomous ideological standing and pursued an economic policy that deviated from the socialist principles; Albania was concealed from and for the world as a stronghold of Enver Hoxha (1908–85). In addition to these differences among the states in the socialist period, other specificities should be taken into account, including specificities that originated in the presocialist times and later were downplayed as insignificant for the socialist development of the respective societies. These specifics include the traditionally strong influence of Catholicism in Poland, the imperial past of Czechoslovakia and Hungary within the Austro-Hungarian Empire, and the legacy of Bulgaria as a former domain of the Ottoman Empire. These differences, among others, exert their impact on the national specificity of the process of 'transition' despite the shared socialist past of the countries, as well as on the processes of overcoming the socialist legacy and joining the EU. Precisely these differences among the European socialist states as well as between the variety of socialisms in European and non-European contexts increasingly attract the research interests of social scholars (Blokker and Dallago, 2009; Boje et al., 2007).

The Difference between Postsocialist and Former Socialist Societies and Countries

The integration of the former socialist states in the EU shifted the emphasis in their self-identification: They already have completed the process of transition and are no longer postsocialist societies but rather are democratic societies that have severed all links to their socialist past. Regardless of the political, economic, and social problems that to a varying degree characterize the contemporary situation of all European democratic societies, the former European socialist states are now part of a united Europe. Differences still exist among them; for example, Bulgaria and Romania are not yet members of the Schengen Area of free movement because they are not in compliance with certain political criteria for joining; Bulgaria, the Czech Republic, Latvia, Lithuania, Hungary, Poland, and Romania are not yet members of the Euro-zone because they not in compliance with certain economic criteria. These are differences, however, that newly accepted EU member states presumably are capable of overcoming.

The term 'postsocialist societies' still is used as the current designation of the societies or states that so far have not been accepted into the EU: Croatia became a member of EU on 1 July 2013; Serbia, Macedonia, and Montenegro are candidates for membership; Albania, Bosnia and Herzegovina, and Kosovo are potential candidates. With the exception of Albania, these are all states that declared their independence and subsequently were recognized by the EU after the military disintegration of the Socialist Federal Republic of Yugoslavia.

The designation 'postsocialist societies' labels an obscure transitory condition. 'Former socialist states,' for its part, designates a state regime that was left behind in history.

Are Non-European (Former) Socialist Countries Postsocialist Societies?

If we apply the previous definition of a postsocialist society, we should assume that China, Vietnam, and Cuba are not postsocialist societies. In their state governance, they continue to maintain the socialist ideology and praxis, whereas the repudiation thereof is a necessary, although insufficient, condition to determine a particular society as 'postsocialist.'

The worldwide changes, in the course of which the former socialist countries in Europe, the countries established after the disintegration of the Union of Soviet Socialist Republics, and Mongolia embarked on a transition toward democratic governance and market economy (by different means and with different pace and success), no doubt have an impact on the countries that adhere to state socialism. Among them, the development of the People's Republic of China is certainly the most interesting case from the viewpoint of the ongoing socioeconomic transformations. This is similar to the development course of the Socialist Republic of Vietnam, which, however, has a much weaker presence on the world economic and political stage. These transformations are induced largely by the existence of private business – albeit under the umbrella of the state authority, the Communist Party – which emerged in the end of the 1980s and by now has absorbed a great part of the productive capabilities of the country and become an important constituent of the global economy (Brown, 2009; Tsai, 2007). What is the appropriate way to define a society, in which socialist rule, capitalist economy, and agriculture based on small family property and traditional–patriarchal values and relations are all in place in a relatively unproblematic coexistence (Brown, 2009; Nee and Opper, 2012; Szelenyi, 1998)? We can conceptualize it as a 'transitional society' inasmuch as it remains unclear in which of the familiar types of society this 'transition' will end and insofar as it might be assumed that the present combination of socialist rule, capitalist economy, and traditional social relations holding together the majority of the population cannot be sustained in time.

The case of Cuba is different. In the Republic of Cuba (the 'Island of Freedom'), a socialist society continues to be

reproduced despite certain changes in the state governance that in no way contest its fundamental socialist principles and values, and a limited 'openness' to the world that tolerates access while resisting outside influence.

See also: Democratic Transitions; Eastern European Studies: Culture; Ethnic Conflicts; Minorities; Nationalism, Sociology of; Post-Socialism: Political Economy; Poverty, Sociology of; Prejudice in Society: Psychological Perspectives; Russian and Post-Soviet Studies: Culture; Russian and Post-Soviet Studies: Politics; Social Identity in Sociology; Socialist Societies: Anthropological Aspects.

Bibliography

Blokker, P., Dallago, B. (Eds.), 2009. Regional Diversity and Local Development in the New Member States. Palgrave/Macmillan, Houdmills, UK, New York.

Boje, T., Van Steenbergen, B., Walby, S., 2007. European Societies: Fusion or Fission? Routledge, London.

Brown, A., 2009. The Rise and Fall of Communism. Doubleday, Canada.

Burawoy, M., Verdery, K. (Eds.), 1999. Uncertain Transition: Ethnographies of Change in the Postsocialist World. Rowman & Littlefield, Lanham, MD.

Fowkes, B., 1999. The Post-Communist Era: Change and Continuity in Eastern Europe. St Martin's Press, New York.

Kornai, J., 1995. Highway and Byways: Studies on Reform and Post-Communist Transition. MIT Press, Cambridge, MA.

Kornai, J., 1997. Struggle and Hope: Essays on Stabilization and Reform in a Post-Socialist Economy. Edward Elgar, Cheltenham, UK.

Michta, A., Prizel, I. (Eds.), 1992. Post-Communist Eastern Europe: Crisis and Reform. St Martin's Press, New York.

Nagle, J.D., Mahr, A., 1999. Democracy and Democratization: Post-Communist Europe in Comparative Perspective. Sage, London.

Nee, V., Opper, S., 2012. Capitalism from Below: Markets and Institutional Change in China. Harvard University Press, Cambridge, MA.

Offe, C., 1996. Varieties of Transition: The East European and East German Experience. Polity Press, Cambridge, MA.

Opalski, M. (Ed.), 1998. Managing Diversity in Plural Societies: Minorities, Migration and Nation-Building in Post-Communist Europe. Nepean Forum Eastern Europe, Nepean, ON, Canada.

Priestland, D., 2009. The Red Flag: A History of Communism. Penguin Books Ltd, London.

Rose, R., Mishler, W., Haerpfer, C., 1998. Democracy and Its Alternatives: Understanding Post-Communist Societies. Johns Hopkins University Press, Baltimore, MD.

Rupesinghe, K., King, P., Vorkunova, O. (Eds.), 1992. Ethnicity and Conflict in a Post-Communist World: The Soviet Union, Eastern Europe, and China. St Martin's Press, New York.

Szelenyi, I. (Ed.), 1998. Privatizing the Land: Rural Political Economy in Post-Communist and Socialist Societies. Routledge, London.

Tismaneanu, V., 1998. Fantasies of Salvation: Democracy, Nationalism, and the Myth in Post-Communist Europe. Princeton University Press, Princeton, NJ.

Tsai, K.S., 2007. Capitalism without Democracy: The Private Sector in Contemporary China. Cornell University Press, Ithaca, NY.

Relevant Websites

http://www.europarl.europa.eu/enlargement/ec/pdf/cop_en.pdf – Copenhagen Criteria.
http://ec.europa.eu/enlargement/index_en.htm – EU Enlargement.
http://ec.europa.eu/economy_finance/eu/euro_area/index_en.htm – Euro Area.
http://www.historiasiglo20.org/europe/amsterdam.htm – History of the European Union.
http://europa.eu/legislation_summaries/justice_freedom_security/free_movement_of_persons_asylum_immigration/l33020_en.htm – Schengen Area.

Post-Traumatic Stress Disorder

J David Kinzie, Oregon Health and Science University, Portland, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Newman, volume 17, pp. 11881–11885, © 2001, Elsevier Ltd.

Abstract

Since the description of post-traumatic stress disorder (PTSD) in the second edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-II) in 1980, PTSD has become a well-established and well-researched clinical condition. DSM-5 further refined the criteria but it is not consistent with the proposed 11th version of International Classification of Diseases criteria for PTSD. PTSD is fairly common and its prevalence increases with increase in traumatic events and is associated with alcohol abuse, depression, hypertension, and diabetes. The general theory of symptoms of PTSD is that the uncontrollable state of fear persists even after the threat is gone. Recent biological research indicates that biological factors such as genetics increased activity in the amygdale, reduced cortisol, and hyperactivity of adrenergic response occur in chronic PTSD. Psychological treatment has usually been trauma-focused behavioral therapy, but it is not clear if it is necessary to focus specifically on the trauma rather than supportive therapy of current life stressors. Medicine can reduce major symptoms such as nightmares, irritability, sleep disturbances, and needs to be used with psychotherapy. Future research will provide better understanding and treatments of this often chronic disabling disorder.

Introduction

Since ancient times, psychological traumas have been described in various literature. The Sumerians in 2000 BC lamented the destruction of Nippur, leaving the population in total anguish and suffering. There was not a good clinical description of psychological effects of trauma until the railroad spine syndrome, which was described by Erikson in 1887 (Kinzie and Goetz, 1996). The symptoms from railroad train accidents included anxiety, memory and concentration problems, irritability, disturbed sleep, sudden alarmed distressing dreams, and multiple somatic complaints. Similar complaints also include chest pain, rapid pulse, and shortness of breath were noted during the American Civil War. Natural disasters, such as earthquakes and volcano eruptions, resulted in clinical descriptions of acute reactions and even psychosis with clouded consciousness, amnesia, disorientation, fearful affect, hallucinations, and perceptual disturbances.

An increase in post-traumatic symptoms were described during and after World War I under the name 'shell shock,' which later also became known as 'war neurosis.' This was a heterogeneous group of symptoms, which were originally explained by organic brain damage but was later known to have mostly psychological cause. Early theory of post-traumatic stress symptoms were dominated by Freudian psychology, which was a neurosis related to sexual conflict. However, this was not universally accepted because neurosis emphasized intrapsychic processes and downplayed the effects of the trauma itself. World War II saw many severe reactions to war, namely 'war fatigue,' 'war neurosis,' and 'soldier syndrome,' and these had marked similarities to those described previously to traumatic events. Concentration camps of the Holocaust provide further evidence of severe psychological impairment from very traumatic events including mass execution, brutality, starvation, and humiliation. Among the symptoms were restricted mental horizons, indifference, and irritability. The symptoms after the concentration camps were diagnosed as Gross Stress Reaction in the first edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-I) and usually involved a poor prognosis.

No similar reaction diagnosed in DSM-II, but in DSM-III edition printed 1980 after Vietnam War veterans returned there was a description of the major aspects of post-traumatic stress disorder (PTSD) under the category of anxiety disorders. The symptoms had broad categories of reexperiencing avoidance and autonomic arousal. The DSM-IV specified the trauma criteria in which a person experiences, witnesses, or confronted with events involving actual threatening of death or serious injury or the psychological integrity. In conjunction with the event, the person who suffered responses must include intense fear, helplessness, or horror with multiple symptoms of persistent reexperiencing including nightmares, distress, intrusive thoughts, and intense psychological distress when exposed to external cues. Persistent avoidance was another group of symptoms involving numbing of response or avoidance; effort to avoid thoughts and avoid activities reminding one of the traumas they experienced, an inability to recall important aspects of the trauma, decreased interest in participation of significant activities, feeling of detachment, or enstrangement from others. Hyperarousal symptoms involve difficulty falling asleep or staying asleep, difficulty concentration, irritability, hypervigilance, and exaggerated startle response. The duration of these symptoms lasted longer than 1 month and caused significant impairment in social, occupational, or other important aspects of functioning. PTSD was separated from acute stress disorder in which there are dissociative symptoms and the disturbances last for a minimum of 2 days and a maximum of 4 weeks.

DSM-5 of the American Psychiatric Association Diagnostic and Statistical Manual published in 2013 involved major changes in the diagnosis of PTSD. In DSM-5, PTSD was included in new category not under 'anxiety disorders.' It is in a separate chapter called 'Trauma and Stress-Related Disorders.' This chapter specifically is for diagnosis in which a traumatic event resulted in the onset of symptoms. In DSM-5, the removal of the subjective response, which was a fear-based response (fear, helplessness, or horror), because research

indicated that these fear reactions did not happen for many people who subsequently developed PTSD. Summarized here, there must be at least one exposure to an actual or threatened death, serious injury, or sexual violence. Criteria A: presence of at least one intrusive symptom such as recurrent memories of distressing dreams. Criteria B: at least one symptom of persistent avoidance. Criteria C: at least two negative alterations in mood or cognition such as the inability to remember important aspects of the traumatic events or impairment in relationships and is not associated with substance use. Criteria D: hyperarousal symptoms such as irritability hypervigilance with exaggerated startle response. Criteria E: symptoms muse last longer than 1 month and must cause significant impairment in social, occupational, and other important areas of functioning.

DSM-5 notes that cultural factors may influence the expression of symptoms and also resilience. Certain cultures have more exposure to cultural events such as genocide and impact of trauma may influence the entire social–cultural society such as inability to perform funeral rituals or residing in population of unpunished perpetrators.

Early research indicates that the DSM-5 changes probably do not significantly change the prevalent rate of PTSD. A controversial issue was to include a category called Complex PTSD in DSM-5. This would involve events that are severely traumatizing including prolonged torture or in the developmental stages, i.e., severe child abuse. Eventually, it was ruled that there was not enough evidence to support this category despite strong clinical experience that some forms of traumas are so severe that they involve special consideration, treatment, and prognosis (McDonnell et al., 2013).

The WHO international classification of diseases (ICD), eleventh version, has proposed two related diagnosis: PTSD and complex PTSD. PTSD symptoms include core elements reexperiencing of traumatic reminders and excessive hypervigilance or exaggerated startle response. The syndrome has fear at its heart and conceptualizes PTSD as a fear condition.

Complex PTSD has basic PTSD symptoms as well as three domains of self-organization, affective dysregulation, negative self-concepts, and interpersonal problems. A latent profile analysis (Cloitre et al., 2013) confirmed that chronic traumas was more predictive of complex PTSD and that complex PTSD had more impairment than PTSD.

The development of ICD-11 and DSM-5 were to harmonize the two classification symptoms of mental disorder. As shown above, the symptoms are markedly different have the potential to cause confusion to clinicians, patients, and researchers (Bisson, 2013).

Prevalence, Comorbidity, and Prognosis

Although many people (up to 90%) experience a traumatic event in their lifetime, most do not develop PTSD. By the National Comorbidity Survey (Kessler et al., 1995), the lifetime prevalence of PTSD is 7.8%. It is higher among women traumatized by rape and sexual molestation and is also higher among men with combat exposure. A recent study has a lifetime prevalence rate at 10.1% and a 12-month prevalence rate of 3.1% in 2012 (Kessler et al., 2012). Report by the US Military personnel are thought to be 14–15% for lifetime prevalence of PTSD in combat veterans. A study of torture survivors who are applying to the United States as refugees or asylum seekers, the rate of PTSD is 80% by treating clinicians' evaluation.

PTSD is highly comorbid with other disorders. In a national survey (Kessler et al., 1995), about 49% had comorbid major depression. Alcohol abuse or dependence was 52% among men and 28% among women. In clinical interviews, it was found that 63% of Vietnam veterans with PTSD had comorbid major depressive disorder (MDD) as compared to 45% of Afghani/Iraqi veterans (Pukay-Martin et al., 2012). Among torture survivors, the comorbid rate of MDD was 62% (unpublished data, National Consortium of Torture Treatment Programs USA). The latter populations had almost no alcohol abuse or dependence, probably related to the population predominantly Muslim and Buddhist.

For many, the course of PTSD is quite chronic. More than a third of people with index PTSD have failed to recover after many years and it has shown among follow-up studies of patients with PTSD that it has a fluctuating course with remissions and exacerbations under stress (Chopra et al., 2014). Fluctuating course, especially among the elderly, gives an overall lower quality of life. It is now known that veterans of recent wars in Afghani/Iraqi region have a high suicide rate. Studies have shown that veterans with PTSD and major depression have a high suicidal ideations and that depression may be the mediating factor of suicidal behaviors.

There is increasing evidence that chronic PTSD can significantly increase the association with type 2 diabetes (Lukaschek et al., 2013). In addition, there is evidence it may increase the blood pressure and hypertension as a consequence of the arousal symptoms of this disorder (Paulus et al., 2013). Furthermore, evidence has accumulated that childhood and probably adult trauma can increase the risk of cognitive impairment in old age, leaving a higher prevalence of dementia in this group (Burri et al., 2013).

In summary, it appears that PTSD is fairly common in a national sample population. The prevalence of PTSD increases when an increase of traumatic stress happens, such as it would be for combat veterans and particularly torture victims. In national studies, women have more prevalence of PTSD than men, and in both, it is highly comorbid with MDD. In American veterans, PTSD is correlated with alcohol abuse, but not with samples from torture victims. It is now known that suicide is a long-term concern of veterans with PTSD and depression, and now, diabetes, hypertension, and dementia may also be a long-term consequence of PTSD.

Etiology

PTSD is the only major psychiatric disorder for which there is a known cause; that is a traumatic event that is a threat to the personal integrity or life of an individual. There has been a movement to go away from an intrapsychic dynamic of neurotic mechanisms, as was emphasized in earlier psychological literature, to one where there is an emphasis on the real trauma itself. Using a Pavlovian model, it is thought that the response to traumatic events occurs not just at a psychological level, but also at a physiological level due to classical conditioning. PTSD, with its continuation of symptoms, is consistent

with the uncontrollable state of fear. In this model, PTSD represents a failure of extinction. Recent research (Bryant et al., 2013) document that PTSD has a complex and fluctuating course, 44% of these 24 months after the event had PTSD, but did not meet criteria at 3 months. Part of this was explained by the number of stress events experienced between 3 and 24 months. This finding is consistent with a sensitization model, which suggests that an excessive response to a subsequent less-stressful event occurs because neural pathways are sensitized after an initial traumatic event. Sensitization model helps explain delayed onset and reactivation symptoms. (For a review of memory mechanisms involved in PTSD, see Parsons and Ressler, 2013.)

Biological Factors in PTSD

One of the most robust biological findings in PTSD is elevated heart rate during both trauma cue studies and resting states. They support the idea that PTSD is associated with elevated psychophysiology (Pole, 2007). As mentioned before, hypertension is found in many PTSD patients (Paulus et al., 2013). Other biological factors have been intensely studied in PTSD and have been summarized in two expert's reviews (Pitman et al., 2012; Sherin and Nemeroff, 2011).

Genetics

Recent estimates are that more than 30% of the variance associated with PTSD is related to hereditary components. These genetic factors are common to other major illnesses: depression, anxiety, and panic disorder. The genetic influences on PTSD are probably polygenetic and at least 17 gene variances have been associated with PTSD. These have been involved in various central nervous systems. There is a lack of consistency of association of specific genetic variances of PTSD, which may be explained by the modification of the other genes or environmental factors not accounted across studies, for example, extreme traumatic stress. Further genome-wide association studies perhaps offer the best opportunities to identify novel 'true-untrue' variances in PTSD. One current concept from genetic studies is that the effects may not directly be the result from the trauma itself, but may predispose to trauma or symptoms. Predisposition may be involved in personality traits in risky behavior, increased violence, or the expression of symptoms once trauma has been occurred. Genetic factors may interact with environmental factors to cause PTSD. A study found that a stress-related gene (FKPB5) is a polymorphic variant, which interacts with child abuse severity as a predictor of adult PTSD (Binder et al., 2008). There is also evidence that the serotonin transporter gene, which regulates serotonin, and thus emotional response, is involved in a PTSD system. However, there are contradictory results, which specific variants of the serotonin transporter gene are involved in PTSD in moderately but not severely traumatized individuals.

Imaging Studies

In magnetic resonance imaging (MRI) studies of PTSD, the most consistent finding shows lower volume of the hippocampus and a lower volume of the central prefrontal cortex. There is a controversy whether the small hippocampus is a result of traumatic exposure or rather represents a risk factor for PTSD that is of genetic or environmental origin. Interesting suggestion is that pharmacological treatment of PTSD with the selective serotonin reuptake inhibitors (SSRI), paroxetine has reported to increase hippocampus volume. Thus, reduced hippocampus volume could be an acquired reversible abnormality even as it is a preexisting risk factors.

Functional Neuroimaging Studies

Functional imaging studies using positron emission tomography or functional MRI (fMRI) have shown many activity changes in the brain of people with PTSD. The amygdala is involved in recognizing stimuli signaling danger and the fear response and has exaggerated activities in individuals with PTSD. In contrast, activities of the ventromedial prefrontal cortex (VMPC) are reduced. The VMPC reduces the fear response and its activity and is negatively correlated with PTSD symptoms. The inactivity of VMPC and the continued exaggerated amygdala response may indicate why extinction does not occur in PTSD patients.

Much research has found that cortisol is either reduced or normal in individuals with PTSD and as not as high as expected when in a state of chronic stress. Cortisol is a hormone, released from the adrenal cortex in an acute response to stress in a threatening or uncontrollable nature. The concept of cortisol deficiency in the pathogenesis of PTSD is supported by a preliminary study of high-dose cortisol administration. The study showed that administration of cortisol within hours of a traumatic event reduced the development of PTSD.

Neuroendocrine Studies

There is much evidence of the sympathetic nervous system hyperactivity adrenergic response (adrenaline or norepinephrine) in PTSD. An increase of norepinephrine may influence a formation of durable traumatic memories. It also contributes of hyperarousal symptoms including nightmares, flashbacks, intrusive memories, and a physiological response to traumatic stimuli. The efficacy of prazosin and clonidine, central nervous system blocking agents, in treating hyperarousal symptoms are consistent with these findings (Boehnlein and Kinzie, 2007).

Neurobiological Research

This brief summary of emerging and increasingly sophisticated studies of the neurobiology involved in PTSD has shown robust findings with contradicting results. They do point out that there are several clear biological mechanism involved in the brain during severe trauma and are hopeful that further research will lead to more definitive treatments. Neurobiological research has helped to elucidate basic mechanism of PTSD and lead to rational understanding of the psychological therapies as well as pharmacological treatments.

Psychological Treatment

Psychological interventions are widely used in the treatment of PTSD; an analysis of types of intervention include trauma-focused cognitive behavior therapies, exposure therapies, stress management, and eye movement desensitization and reprocessing . An early report indicated that there is evidence for the effectiveness of all these treatment of people with PTSD (Bisson and Andrew, 2007). Nontrauma-focused therapies did not reduce PTSD symptoms as significantly as trauma-focused therapies. More recently, it has suggested that prolonged exposure therapy (PET) as a first line treatment for PTSD has been demonstrated to reduce comorbid issues such as anger, guilt, and depression, in diagnostically complex population. The main components are psychological education, prolonged imaginative exposure, and emotional reprocessing. These have been supported by the American Veteran's Health Administration for training of mental health professionals (Rauch et al., 2012). In a comparison of exposure versus cognitive therapy, it was found there were no significant differences in the relative effectiveness of the two in PTSD (Ougrin, 2011). It has been universally accepted concept that trauma-focused therapies are superior to nontrauma-focused therapies for PTSD. However, there is now evidence that this may not be true. The results may be influenced by a selection bias and not controlling for nonspecific or common factors. It may that suggests that nontrauma-focused intervention may be just as effective (Wempolda et al., 2010). A nontrauma-focused therapy called present-centered therapy was originally designed as a control treatment but is now listed as a research-supported intervention for people with PTSD. There is a point of research to see if trauma-focused interventions are necessary for treating PTSD (Imel et al., 2012).

Clearly, the amount of severe trauma presented by PTSD patients strains the therapist's ability as well as the treatment paradigm in delivering psychotherapy for people with PTSD. Complex and atypical patients such as survivors of child abuse or severe trauma like concentration camps require a more modified approach than the often protocol indicated from research design. There is a need for therapist flexibility and deviation from the usual treatment approaches. Clear cases are traumatized refugees who have undergone severe trauma as well as multiple losses, perhaps over a long period of time. An important ingredient in successful treatment may be the nonspecific factors such as the establishment of trust, a good relationship with the therapist, and the professional's compassion in working with these difficult patients (Mirdel et al., 2012; Kinzie, 2001). The nonspecific factors have relevance for the Western-traumatized patients where previously specific techniques have been emphasized.

Combined Medicine and Psychotherapy

A recent article using PET and paroxetine SSRI for PTSD-related symptoms to the World Trade Center attack found that the combination therapy resulted in significant improvement compared to placebo in both response rate and quality of life (Schneier et al., 2012). In combined supportive psychotherapy with medicine proved to be very helpful for severely tortured refugees after 1 year of treatment. This was a nontrauma-focused therapy with emphasis on the relationship and treatment consistent with the culture, 20 out of 22 had a good treatment response at 1 year (Kinzie et al., 2013).

Pharmacological Therapy for PTSD

There is a great deal of evidence now that many of the symptoms of PTSD respond to medication. Paroxetine is FDA approved for the treatment of PTSD, but there is little reason to suggest it is better than any other SSRI. Probably the most consistent approach psychiatrists have used is to use targeted agents for specific symptoms of PTSD and their comorbid disorders (Jeffreys et al., 2012). The following suggestions are from Jeffrey's article and from the personal experience of the author treating thousands of patients over 30 years. It is clear that medicine cannot work by themselves and need to be combined with ongoing psychotherapy, either trauma focused or relationship focused such as interpersonal or supportive. Also, the guidelines are only suggestions and that most medicines have side effects and must be individualized to the patient's needs and symptoms. Most pharmacological therapies include SSRI or a combination of serotonin–norepinephrine reuptake, such as venlafaxine. They often are ineffective for nightmares and reexperiencing phenomena. The most common complaints for severe PTSD are poor sleep, nightmares, and intrusive thoughts, to which tricyclic antidepressant such as imipramine or doxepin, which are more sedative, have greater effects on these complaints, and are very helpful. Nightmares and reexperiencing are often helped by prazosin and clonidine (Boehnlein and Kinzie, 2007). Aggressive-agitated behavior is often helped by a low dose of antipsychotic risperdal. The SSRI and tricyclic antidepressants can treat the comorbid MDD. Most medical treatment needs to be long term and can sometimes be overridden by ongoing or renewed trauma. The rapid relief of sleep disorders, nightmares, and aggressions with medication often gives the patient immediate relief and help them to maintain in psychotherapy treatment.

New and Emerging Research

Clonidine, an adrenergic blocking agent in the central nervous system can disrupt fear memory retention. Thus, when administered after retrieval of traumatic memories, it could prevent the reconsolidation and augment exposure therapy (Gamache et al., 2012). An unusual treatment is to use deep trans-cortical magnetic stimulation with brief exposure therapy in PTSD. A study has found that there was significant improvement in the intrusive component of PTSD after deep brain stimulation to the traumatic event script (Isserlei et al., 2013). Other agents involved in PTSD treatment trials include steroids, glutamate, and dopamine, used in combination with exposure therapy for PTSD (Dunlop et al., 2012).

A problem for all PTSD research and research for many psychiatric disorders is that the diagnosis is often made as of result of self-reported symptoms and sometimes clinician's observation. There is no real biomarker that would definitively diagnose and reliably provide objective evidence of PTSD. Such

a long-term study is underway at New York University and they lead to reliable markers for PTSD as well as mild traumatic brain injury (*New York Times,* 7 February 2013, page 18).

See also: Antidepressant Drugs; Asylum Seekers; Mass Trauma: Psychopathological Effects across the Life Span; Military and Disaster Psychiatry; Torture and its Consequences, Psychology of.

Bibliography

Burri, A., Maercker, A., Krammer, S., Simmen-Janevska, K., 26 February 2013. Childhood trauma and PTSD symptoms increase the risk of cognitive impairment in a sample of former indentured child laborers in old age. PLoS One [Epub ahead of publication].

Binder, E.B., Bradley, R.G., Liu, W., et al., 2008. Association of FKBP5 polymorphism and childhood abuse with risk of post traumatic stress disorder symptoms in adults. Journal of the American Medical Association 299, 1291–1305.

Bisson, J., 2013. What happened to harmonization of the PTSD diagnosis. The divergence of ICD 11 and DSM 5. Epidemial Psychiatry Science 22, 1–3.

Bisson, J., Andrew, M., July 18, 2007. Psychological treatment of post traumatic stress disorder (P75D). Chochrane Database Systematic Review.

Boehnlein, J.K., Kinzie, J.D., 2007. Pharmacologic reduction of CNS noradrenergic activity on PTSD: the cure for clonidine and prazosin. Journal of Psychiatric Practice 13, 72–78.

Bryant, A.B., O'Donnell, M.L., Creamer, M., McFarlane, A.C., Silove, D., 2013. A multisite analysis of the fluctuating course of posttraumatic stress disorder. JAMA Psychiatry 70, 839–896.

Chopra, M.P., Zhang, H., Koiser, A.P., Moye, J.A., Liorente, M.P., Oslin, D.W., 10 October 2014. Sptro A 3rd. P74D is a chronic, fluctuating disorder affecting the mental quality of life in older adults. American Journal of Geriato Psychiatry 22, 86–97 [Epub ahead of print].

Cloitre, M., Garvest, D.W., Brewin, C.R., et al., 2013. Evidence for proposed ICD 11and complex PTSD: a latent profile analysis. European Journal of Psycho-traumatology 10, 3402.

Dunlop, B.W., Manson, E., Gerardi, M., 2012. Pharmacological innovations for post traumatic stress disorder and medication- enhanced psychotherapy. Current Pharmaceutical Design 18, 5645–5658.

Gamache, K., Pitman, R.K., Nader, K., 2012. Pre-clinical evaluation of reconsolidation blockade by clonidine as a potential novel treatment for post traumatic stress disorder. Neuropsychopharmacology 37, 2789–2796.

Imel, Z.E., Loska, K., Jaropcak, M., Simpson, T.L., January 21, 2012. Meta-analysis in drop-out in treatments for post traumatic stress disorder. Journal of Consulting and Clinical Psychology.

Isserlei, M., Shalew, A.Y., Roth, Y., et al., 2013. Effectiveness of deep transcranial magnetic stimulation combined with a brief exposure procedure in post-traumatic stress disorder - a pilot study. Brain Stimulation 6, 377–383.

Jeffreys, M., Capehart, B., Freidman, M.J., 2012. Pharmacotherapy for posttraumatic stress disorders: review with clinical applications. Journal of Rehabilitation Research and Development 49, 703–715.

Kessler, R.C., Sonnega, A., Bromet, E., Hughes, M., Nelson, C., 1995. Post traumatic stress disorder in the national comorbid survey. Archives of General Psychiatry 52, 1048–1060.

Kessler, R.C., Petukhova, M., Sampson, N.A., et al., 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. International Journal of Methods in Psychiatric Research 21, 169–189.

Kinzie, J.D., Goetz, R.R., 1996. A century of controversy surrounding post traumatic stress- spectrum syndromes: the impact on DSM-III and DSM-IV. Journal Traumatic Stress 9, 159–179.

Kinzie, J.D., 2001. Psychotherapy for massive traumatized refugees: the therapist variable. American Journal of Psychotherapy 55, 475–490.

Kinzie, J.D., Kinzie, J.M., Ledighi, B., et al., 2013. Prospective one year treatment of torture refugees: a psychiatric approach. Torture 22, 1–10.

Lukaschek, K., Boument, J., Kruse, J., et al., January 29, 2013. Relationship between post traumatic stress disorder and type 2 diabetes in population based cross-sectional study with 2970 participants. The Journal of Psychosomatic Research 74, 340–345 [Epub ahead of print].

Mirdel, G.M., Rydrig, E., Essendrop Sondry, M., 2012. Traumatized refugees, other therapies, and their interpreters: three perspectives on psychological treatment. Psychology and Psychotherapy 85, 436–455.

McDonnell, M., Robjant, K., Katana, C., 2013. Complex post traumatic stress disorder and survivors of human right violations. Current Opinion in Psychiatry 26, 1–6.

Ougrin, D., 2011. Efficacy of exposure versus cognitive therapy in anxiety disorders systematic review and meta-analysis. BMC Psychiatry 11, 200.

Parsons, R.G., Ressler, K.J., 2013. Implications of memory modulation for post traumatic stress and fear disorders. Nature Neuroscience 16, 146–153.

Paulus, E.J., argo, T.R., Egge, J.A., 2013. The impact of post traumatic stress disorder on blood pressure and heart rate in a veteran population. Journal of Traumatic Stress 26, 169–192.

Pitman, R.K., Rasmusson, A.M., Koenen, K.C., et al., 2012. Biological studies of post traumatic stress disorder. Nature Reviews Neuroscience 13, 769–789.

Pole, N., 2007. The psychophysiology of post traumatic stress disorder: a meta-analysis. Psychological Bulletin 133, 725–746.

Pukay-Martin, N.D., Pontoski, K.E., Maxwell, M.A., et al., 2012. The influence of depressive symptoms on suicidal ideation among US Vietnam-era and Afghanistan/Iraqi era veterans with posttraumatic stress disorder. Journal of Traumatic Stress 25, 578–582.

Rauch, S.A., Effehari, A., Rizer, J.I., 2012. Review of exposure therapy a gold standard for PTSD treatment. Journal of Rehabilitation Research and Development 49, 679–687.

Sherin, J.E., Nemeroff, C.B., 2011. Post traumatic stress disorder: the neurological impact psychological trauma. Dialogue Clinical Neuroscience 13, 263–278.

Schneier, F.R., Neria, Y., Pavlicova, M., et al., 2012. Combined prolonged exposure therapy and paroxetine for PTSD related to World Trade Center attack, a randomized controlled trial. American Journal of Psychiatry 169, 80–88.

Wempolda, B., Imelb, Z.E., Lasksa, K., et al., 2010. Determining what works in the treatment of PTSD. Clinical Psychology Review 30, 923–933.

Post-Traumatic Stress in Social Work

Orit Nuttman-Shwartz, The Israeli National Council for Social Work, School of Social Work, Sapir College, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over 60% of all individuals around the world have been exposed to traumatic events. Responses to those events range from pathological symptoms and vulnerability to healthy responses, hardiness, and resilience. These responses derive from individual, family, community, and social resources, as well as from the ability of individuals to cope with the events, and to work through the consequences of their exposure. This article relates to the role of social workers before, during, and after traumatic events. A brief definition and description of the main responses to these situations is provided, followed by a discussion of social workers' approaches and strategies of intervention. Implications for educational and professional training, the cost of care, and for practice are also discussed.

Introduction

Over 60% of all individuals around the world have been exposed to traumatic events such as violence, abuse, terror, civil and political wars, racial strife, human trafficking, forced migration, human immunodeficiency virus/acquired immunodeficiency syndrome, and technological and natural disasters. Traumatic events are private, public, or international, and they significantly affect individuals, families, groups, and communities in a variety of contexts such as the war in Croatia (1991–95), the London bombings (2005), the September 11 terrorist attack (2001), and the Israeli-Palestinian conflict. In addition, many traumatic events have occurred in civilian environments. For example, in the United States there have been the shootings at Virginia Tech University (2007), at a movie theater in Dallas (2012), and at an elementary school in Connecticut (2012). Traumatic events also result from natural disasters such as Hurricane Katrina in New Orleans (2005), the Tsunami in Asia (2004), and the earthquake in Haiti (2010).

In recent years, there has been increasing recognition that individual differences in responses to traumatic events cover a broad spectrum, ranging from pathological to healthy responses and from short-term to long-term responses. They also include the ability to cope, recover, and work through these events, as well as to identify risk and protective factors. Finally, numerous transpersonal factors can contribute to or detract from healthy functioning in the aftermath of such events. These factors combine person-centered variables (e.g., disposition and personality) with sociocontextual variables (e.g., family interaction and community support systems) and risk and protective factors.

These understandings have led to a shift from a medical approach, which focuses on traumatic responses, to a psychosocial approach, which is based on the results of theoretical and empirical research evidence indicating that individual responses to traumatic events are affected by one's subjective perspective of the event. This approach is well known as stress theory (Lazarus and Folkman, 1984). It has also been argued that responses to traumatic events are the result of personal and social resources that enable people to cope with the aftermath of those events, as coined in the Conservation of Resources (COR) theory (Hobfoll et al., 1995).

In addition to the increasing rate of traumatic events, the above understandings have been heightened by the growing identification of social workers with the field of trauma and their need to face unprecedented challenges in responding effectively to their clients' trauma-related experiences. Therefore, this article will relate to the role of social workers before, during, and after traumatic events that result from civil and national traumas as well as from natural disasters. Following a brief definition and description of the main responses, social workers' approaches and methods of intervention in situations of trauma will be discussed. The article will conclude with a critical reflection on the uniqueness of being a social worker dealing with traumatic events in today's world, which is full of violence.

The Ramifications of Trauma

Two seemingly opposing notions relate to the negative ramifications of trauma on the one hand, and positive responses in the immediate and long-term period after traumatic events on the other. The current understanding is that the responses to a potentially traumatic event fall within a spectrum ranging from pathological to salutogenic (healthy) responses, and from vulnerability to resilience on the other.

Pathological Responses

The first notion, which is most prevalent, is based on the assumption that trauma has a pathogenic effect. Traumatic events jeopardize physical and psychological equilibrium, and give rise to a wide range of physical and mental health complications. There is extensive empirical research evidence in support of this view, which has documented increased rates of posttraumatic stress disorder (PTSD), depression, anxiety, somatization, and alcoholism following traumatic events (Dohrenwend et al., 2006). PTSD is one of the most substantial long-term psychopathological sequelae of such distress. Its symptoms can be grouped into three main clusters. The first consists of persistent reexperiencing of the traumatic event, such as recurrent dreams and flashbacks. The second is persistent avoidance of internal or external cues, such as avoiding

thoughts and activities associated with the trauma, diminished interest, detachment, restricted affect, and sense of foreshortened future. Finally, increased arousal is manifested in difficulty sleeping, irritability, difficulty concentrating, hypervigilance, and exaggerated startle responses (American Psychiatric Association, 1994). The core symptoms of PTSD (intrusive images and thoughts, avoidance, and arousal) severely impair day-to-day functioning, and have a significant negative impact on one's intrapersonal and social life. Although over 90% of the exposed population shows acute stress responses immediately after traumatic events, the majority successfully returns to normal routines, and cope remarkably well. Only 5–11% of these individuals have long-term distress effects, and have been diagnosed with PTSD (Bonanno and Mancini, 2012). Although most of the existing literature on trauma and posttraumatic stress focuses on the effects of trauma on the primary survivor, i.e., the person who directly experienced the traumatic event, direct exposure to trauma not only jeopardizes a person's own physical and mental health, but also undermines that person's familial, occupational, and social functioning (Solomon, 1993).

Salutogenic Responses

An alternative perspective proposes that trauma has a salutogenic effect, and that survivors may gain psychological benefits. According to this perspective, which is in line with positive psychology theory, individuals can develop a positive outlook and further experience positive psychological changes in the wake of traumatic events. The commonly held concept of posttraumatic growth (PTG, Tedeschi and Calhoun, 1996) proposes that individuals transform in new ways that go beyond their pretrauma level of psychological functioning. PTG is described as the subjective experience of positive psychological change reported by an individual as a result of coping with trauma. PTG relates to a variety of positive psychological changes, including increased appreciation of life, setting of new life priorities, a sense of increased personal strength, identification of new life possibilities, improved closeness in intimate relationships, appreciation of life, and/or positive spiritual change.

Without underrating the pathogenic impact of trauma, there is growing research evidence of PTG reported by survivors following various physical and psychological traumas. Such outcomes have been documented following a variety of traumatic events, including natural disasters, war, and terror (for a review, see Linley and Joseph, 2004).

Continuing with the Routine and Resiliency

Despite these emotional reactions, many people tend to maintain their usual routines. Several theorists have used the term resilience in reference to the ability to overcome, to recover, and even to cope well with extreme stressors (Bonanno et al., 2011).

The capacity for resilience has been defined as "the ability of human being in otherwise normal circumstances who are exposed to an isolated and potentially highly disruptive event, such as the death of a close relation or a violent or life-threatening situation, to maintain relatively stable, healthy levels of psychological and physical functioning" (Bonanno, 2004: p. 20). Ungar (2008), using a broader psychosocial perspective, defines resilience as "The capacity of individuals to navigate their way to resources that sustain well-being; the capacity of individuals' physical and social ecologies to provide those resources; and the capacity of individuals, their families and communities to negotiate culturally meaningful ways for resources to be shared" (p. 225).

Ungar's definition highlights the importance of considering the broad systems surrounding the individual, such as family, community, nation, and state, as factors that explain the ability to cope with exposure to traumatic events and the society's role in times of crisis (Breckenridge and James, 2010).

This current and broader definition of resilience is in line with stress theory, which focuses on how individuals perceive traumatic events (Lazarus and Folkman, 1984), and with COR theory (Hobfoll, 1989), which is based on the idea that both the availability of resources and the change in resources that often results from highly disruptive events play a crucial role in human adaptability to extreme stress. These resources are economic and material resources (e.g., income and income loss), energy resources (e.g., the availability of health insurance, of money and knowledge), interpersonal resources (e.g., positive and negative interpersonal support, availability of affinity groups), and work resources (e.g., gainful employment or loss of employment).

This concept is also consistent with the 'person-in-environment' approach, which is a prevalent approach in social work, and supports the idea that it is important to interweave social and environmental resources while working with service users (Hare, 2004). Based on this perspective, the social work profession seeks to augment the ability of individuals, families, and communities to solve their problems and realize their potential (Breckenridge and James, 2010).

Trauma and Social Work

In social work, trauma is often defined and subsumed within its impact on microlevel fields of practice such as child protection, family violence, military social work, and social and interpersonal violence. These definitions present social work education with the conceptual and philosophical challenge of aligning concepts of trauma with notions of strength, recovery, and resilience.

Historically, there is a plethora of literature substantiating the importance of responding to individuals who have experienced trauma. More recently, however, there is evidence of a growing recognition emanating from both the disciplinary and interdisciplinary literature regarding the effects of traumatic events on communities (including nations) as well as the importance of community responses to promote well-being and healing among traumatized individuals. Moreover, individuals and communities themselves are more aware of the potential effects of trauma on health and well-being, thus creating a consumer-led demand for appropriate services. In addition, in the mid-twentieth century, social work journals began publishing articles using the crisis intervention framework and

promulgating specialized training of social workers in agencies that work with clients in crises (NASW, 2008).

Simultaneously, social workers have been trained to work with clients experiencing both acute and chronic stress, to intervene before and during crises, and to implement prevention and postintervention services. These services include individual, family, and group interventions aimed at helping trauma victims achieve 'reasonable mastery' and caring relationships after a traumatic event. In addition, today social workers routinely work in the global arena with people who have experienced multiple traumas. As a result, it is essential for social workers and all human service providers to have an understanding of trauma and the range of its possible effects in order to provide effective, multifaceted intervention in health and welfare contexts.

Moreover, in line with its mission, social work aims to promote social justice and well-being for all people, particularly those who are vulnerable, oppressed, marginalized, and live in poverty. The majority of clients that social workers encounter have experienced numerous and diverse traumatic experiences. As a result, trauma intervention has been incorporated as part of the concepts and values of social work. Within this, the Person in Environment (PIE) approach has been used as a basis for training and for intervention with individuals, families, communities, and societies that have experienced trauma (Joseph and Murphy, 2013). Cultural, ethnic, racial, and other socioeconomic differences have been addressed in an attempt to create policies that strengthen people who need to cope with traumatic events, to reduce conflicts and inequity, and to promote reconciliation and social justice.

In this context, social workers and professionals from other disciplines have been increasingly engaging in forms of 'multifaceted intervention,' which both encompass and depart from traditional approaches that focus primarily on individuals. Notably, these more comprehensive responses promote the integration of community development approaches with psychosocial interventions.

Not surprisingly, the use of combined or multifaceted responses has found increasing support in a number of trauma-related fields, including work with refugees, interventions in disaster-affected communities, and interventions in schools affected by youth suicide and violence, as well as interventions with victims of sexual and physical violence. The rapid development of research-informed practice that has occurred in the field of trauma presents a new and immediate challenge for educators: to prepare students and professionals to provide informed multifaceted interventions in their future practice.

Diversity of Client Populations

Besides providing first aid to trauma victims, the first obligation of social workers is to populations served by various social service agencies. These populations include people with physical and mental disabilities, and children and adolescents at risk living in residential out-of-home settings, as well as young people with learning disabilities, older persons in institutional and community settings (state, municipal, and private), and prisoners. These populations are defined as disadvantaged, excluded, and even invisible, and have been referred to as the 'first recipients of assistance' (Lobley, 2007).

The basic assumption is that as results of social structural problems, these populations are at high risk for developing pathological responses following exposure to traumatic events. They have limited environmental and personal resources at their disposal at times of crisis, because they usually rely on assistance from social workers in routine situations. Accordingly, social workers have to help make decisions, such as those regarding the need to evacuate people from danger zones during natural (e.g., hurricane or flooding) or (hu)man-made disasters (e.g., war and terror attacks) and the need to give them access to psychosocial assistance as well as provide life-saving medications, food, water, shelter, and clothing. In addition, social workers provide psychological and emotional assistance to help them cope with danger, exposure to threat, and the consequences of such situations.

The 'second recipients of assistance' refers to populations that are known to social services but are not in active contact with a helping professional at the time of the traumatic event. In this context, social workers have to evaluate the extent to which these individuals and their families need assistance following the traumatic event. It is important to examine whether they require assistance due to the new situation that has arisen, what individual and family support systems are available to them, and what role the community they live in can play in helping them. In emergency situations, social workers have to actively initiate assistance, because they are supposed to have the information about individuals and communities that have been exposed to traumatic event.

The 'third recipients of assistance' refers to the community at large, the region, or the state/nation. In this context, social workers need to make constant, long-term help available in accordance with the nature and scope of the event. Assistance provided at this level includes intervention in acute stress situations as well as identification, mediation, and cooperating with key figures from various disciplines and helping the community return to a daily routine. In addition, before being exposed to the traumatic event, social workers need to initiate policies that facilitate their implementation in situations of crisis and trauma by influencing processes of decision making and policy making. In addition, they need to train the system of psychosocial services before the event so that the procedures for working with the population will be clear. These procedures should be based on pooling community, environmental, and national resources. Moreover, it is no less important to base the procedures on strengthening the community and promoting its resilience during and after the events. In so doing, social workers not only need to be concerned with bringing the community back to a normal routine in terms of physical, social, and emotional functioning and with self-care but also need to worry about themselves. The last circle relates to the trauma workers themselves including social workers as direct and indirect exposure to the traumatic events (see STR subsection below).

Types of Interventions

Social workers deal with trauma among individuals, families, groups, communities, and policy makers while focusing on

interventions aimed at prevention, therapy, and return to routine. For example, they analyze the differences between supportive counseling interventions and interventions that aim to process trauma and reduce symptoms of distress. In relation to the community, social workers highlight the importance of local and informal social and support networks, as well as the role of public acknowledgment of traumatization. Given the nature of their role, they focus mainly on minorities and subculture populations, and deal with the trauma at the levels of ambiguous loss (of country, culture, and place) and real loss (e.g., Atkinson, 2006).

Traditionally, social workers engage in individual direct practice, which is based on a variety of psychological approaches such as crisis intervention, first aid, and narrative therapy. Recently, there has been a shift to behavioral and cognitive therapy techniques such as critical incident stress debriefing, cognitive behavioral therapy, prolonged exposure, and eye movement desensitization and reprocessing (e.g., Lee and Drummond, 2008), and trauma and stress management, which have been found to be more effective than traditional psychodynamic ways of intervention (e.g., Foa et al., 2000). In addition, family child therapy has become more common as a result of the acknowledgment that there is a relationship between parental responses to traumatic events and the responses of children, babies, young people, and adolescents. In the same vein, child–parent therapy that focuses on a behavioral approach has become common (e.g., Fredman et al., 2011).

In light of the increasing prevalence of traumatic events, which have affected a wide range of populations, group intervention methods have become very popular. These interventions range from debriefing groups and short-term interventions to open groups, long-term interventions, and self-help groups. Through the groups, trauma victims can share their stories with their partners or with others who have experienced similar situations (see also Weinberg et al., 2005).

Recently, as part of the strengths-based approach, the multidisciplinary perspective of mindfulness and creative therapy has become more common as a way in which the community/social network facilitates healing and recovery among individuals. In addition, social workers can help individuals and communities engage in social action or legal reform efforts aimed at increasing awareness and exercising rights to security in their unstable situation before or after the traumatic events.

Finally, global issues are illuminated when vulnerable populations in different parts of the world experience natural or technological disasters. In those situations, social workers tend to engage in disaster and crisis management in contexts such as refugee camps, which require cultural sensitivity and relativity (Gray, 1996).

The Trauma Helpers – Secondary Traumatization, Vicarious Traumatization, and Compassion Fatigue

It has been recognized that a traumatic event may also have far-reaching consequences for significant others, especially for family members. A wide range of emotional symptoms has been identified among family members of Holocaust survivors, children of combat soldiers, and therapists working with victims of violence (Dekel and Monson, 2010). The ramifications of working with trauma survivors have been described through a variety of well-known concepts, including vicarious traumatization (Pearlman and MacIan, 1995), compassion fatigue (Figley, 1995), secondary traumatic stress (Stamm, 2002), and burnout (Maslach and Leiter, 1997). Specifically, Figley (1995) proposed the term 'compassion fatigue' to describe the long-term cumulative stress resulting from the 'cost of caring.' He articulated a comprehensive description of the cognitive, emotional, behavioral, spiritual, and somatic symptoms that can manifest themselves in compassion fatigue, as well as ways in which it can negatively influence one's personal relations and work performance. Secondary traumatic stress, by contrast, can occur suddenly and is directly related to the client's experience of trauma rather than to cumulative professional stress per se. Secondary trauma symptoms are akin to those of posttraumatic stress, and involve anxiety, depression, avoidance, and hyperarousal. These symptoms are in response to the client's trauma narrative, and the clinician's reaction may mirror that of the client. Highly empathic and neophyte clinicians are more prone to develop secondary traumatic stress than are seasoned mental health professionals. Vicarious trauma, like secondary traumatic stress, occurs in the context of work with trauma survivors. Unlike secondary traumatic stress, however, it is not considered a pathological process. Rather, emphasis is on the cognitive and emotional transformations that occur as a result of empathic engagement with trauma survivors. As with secondary traumatic stress, these changes are most pronounced among highly empathic and neophyte therapists as well as among clinicians with a previous trauma history. These changes take place in the totality of the therapist's life. They interfere with the therapist's feelings, cognitive schemata, memories, self-esteem, and/or sense of safety, and they include permanent alterations in one's self-concept and worldview. Vicarious trauma is a unique, albeit common consequence of trauma work, and it does not reflect psychopathology in either the therapist or the survivor client (Pearlman and Saakvitne, 1995).

Shared Traumatic Reality

In the wake of the recent increase in acts of terror and natural disasters, both the clinical and research literature have focused more attention on the problems that arise in a situation where both the social worker and the client are exposed to a similar threat, and on the implications of such exposure for the helping professions.

This situation, which has been referred to as *shared* (*traumatic*) *reality* (STR), is defined as the affective, behavioral, cognitive, spiritual, and multimodal responses that clinicians experience as a result of exposure to the same collective trauma as their clients. Like vicarious traumatization, these reactions have the potential to cause permanent alterations in the clinician's existing mental schema and worldview. However, in contrast to vicarious traumatization, the primary experience of shared trauma causes therapists to become potentially more susceptible to posttraumatic stress, as well as to blurring of professional and personal boundaries, and to increased

self-disclosure. Shared trauma symptoms are attributed to the dual nature of the exposure. Findings indicate that under these circumstances, therapists feel that their ability to help is impaired, and they experience heightened work-related stress as a result of greater demands on their professional time, the sense that they are professionally unprepared for the situation, and loss of boundaries between their personal and professional selves (Cronin et al., 2007; Tosone et al., 2012).

The Other Side of the Coin – 'Vicarious Resilience'

Parallel to the PTG process, Hernandez et al. (2007) have also discussed a relatively new concept referred to as 'vicarious resilience' (VR). This concept provides a basis for considering how the stories of trauma survivors may also sustain and empower workers, and is congruent with many of the strengths-based approaches to practice. VR offers a counterbalance to the negative effects of trauma work on therapists. This concept suggests that if the therapists are open to and aware of the possibility and utility of VR, their ability to reframe negative events and coping skills may be enhanced through work with trauma survivors. Research findings have also shown that therapists who succeed in finding new methods of intervention that help both their clients and themselves have reported an increased sense of competence (Bauwens and Tosone, 2010). A supportive environment and a part-time position can also enhance their ability to develop VR in their work with disadvantaged and traumatized populations and in STR situations (Stevenson et al., 2011).

Models of Training for Intervention in Situations of Stress and Trauma

Considering how the experiences of trauma pervade social work practice, there is a growing body of literature on models that have been employed in trauma education. The most common one is the classic model, a theoretical course on the nature and effects of trauma, which provides an essential foundation for helping professionals and students to understand the context in which they operate (Cunningham, 2003). In addition, training workshops that emphasize practical skills have also been developed (Fournier, 2002). Besides providing theoretical knowledge and skills in respect to stress situations, trauma, and PTSD, the workshops also elicit active participation to create a suitable learning environment and fashion the group as an educational-experiential space, which is very common in the social work education process. A third model often used in training social work supervisors places emphasis on creating a supportive group space that aims to help social workers manage stress successfully and reestablish and enhance their confidence in their professional and personal capabilities (Cwikel et al., 1993). This model is of particular value in a shared traumatic reality, where it is essential to combine client-focused training and work with the care providers themselves, who have been exposed to the same anxiety and insecurity.

As a rule, these models have all been employed with professionals. However, studies conducted among students of social work have revealed that the students are especially vulnerable due to lack of suitable training and insufficient knowledge (Cunningham, 2003). Moreover, the students have not yet acquired and internalized stable theoretical foundations for their work, nor have they developed a clear understanding of the meaning of secondary traumatization or adequate coping strategies. In addition, students have been found to display a tendency for isolation, scapegoating, and interpersonal distance (Cunningham, 2004). Thus, social workers as supervisors need to address the issues raised by students and prevent the development of isolation and feelings of being neglected. The social workers need to keep in mind their important role in supporting students, to help them implement emotional self-regulation, offer a containing and holding environment, and legitimate concerns about being less effective while they themselves are exposed to a threat.

Another relevant issue is coping with trauma in a traumatized society, particularly against the background of national conflict and threats to security. This situation also impacts on the dynamic processes that typify helping relationships, as well as on study and supervision groups. As a result, efforts have been made to develop a global curriculum for social work in the context of social and political traumatic events and conflict, and to incorporate this content in professional core knowledge (Duffy et al., 2013).

Summary and Conclusion

Although social work and other disciplines share universal aspects of understanding and responding to the needs of trauma survivors, the perspective of social work is unique, in that it reflects the values of the profession and emphasizes two aspects. First, social work focuses on understanding the development, functioning, challenges, and needs of people and social systems in situations of trauma from a relational viewpoint and within the context of their environments (e.g., immediate and extended family relationships and cultural, social, racial, ethnic, and religious/spiritual affiliation). Second, social work views people and systems through a holistic, multidimensional lens, which addresses individuals, families, groups, and communities from a perspective that recognizes strengths, resilience, and the ability to cope and thrive in addition to challenges, difficulties, and pathologies (Berger, 2012).

In many cases, social workers serve as 'first-aid helpers' who provide assistance immediately after traumatic events, in addition to providing long-term assistance that is similar to the help provided by other professionals. The multidisciplinary approach to trauma is expressed in training provided at social work schools as well as in the establishment of professional organizations and accreditation. The Council on Social Work Education, for example, has recently published a manual for social workers on work in situations of trauma. In addition, there are studies that have highlighted the complexity and difficulty of social work with clients who have experienced traumatic events, including research reports on the implications of working in a shared reality.

Nonetheless, the question arises: How do social workers respond to the unique challenges of their profession, which

combines interventions ranging from prevention and policy making to short- and long-term interventions that focus not only on the microlevel but also on the macrolevel, as well as on promoting growth, strengths, and resilience? Another question is whether the methods and levels of interventions implemented by social workers are consistent with evidence-based practice and cultural sensitivity, and whether they meet the demands of the complex reality preceding traumatic events, and especially following those events. Research findings have shown that more efforts need to be invested in training social workers and providing them with methods for working with clients who have lived through loss, extreme stress, and trauma when they have been exposed to the same traumatic events as those clients.

To conclude, in line with the residual concept of social work, the contribution of trauma intervention by social workers is as follows:

1. *Incorporating the PIE approach before, during, and after the event*: Emphasis should be placed on prevention, provision of responses, returning to a routine, and even creating and promoting sustainability.
2. *Seeing the person and environment as a whole*: Incorporating the positive and strengths-based approaches that have been found to be effective in promoting healing, recovery, health, and well-being as a result of the fact that the majority of the population is resilient and can cope successfully with the traumatic events. As such, the chances of recovery also depend on individuals and their social resources, identity, sense of belonging, and sense of support.
3. *Incorporating culturally sensitive concepts* in the intervention methods used by social workers while responding to traumatic events and providing suitable assistance to minorities, different cultural subgroups, and others.
4. *Implementing interventions at all levels* such as working with policy makers, communities, families, and individuals. This will also help social workers devote time to working with populations at risk and advocating for those who have been referred to as 'invisible populations.'

In the light of the above, future effort is needed to implement evidence-based practice, as well as to expand existing knowledge and update this with biomedical and alternative perspectives. Toward this end, there is a need to incorporate relevant cultural, biological, and historical dimensions of trauma. In addition, social workers in the field of trauma need to address global processes. It is important for social workers to be trained in dealing with international traumas and the social effects of those traumas (e.g., forced relocation and displacement). This work should be based on antioppressive and postcolonial approaches (Marlowe and Adamson, 2011).

See also: Psychosocial Work; Social Work Theories.

Bibliography

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., DSM-IV. APA, Washington, DC.

Atkinson, J., 2006. Trauma Trails—Recreating Song Lines. The Transgenerational Effects of Trauma in Indigenous People. Spinafex Press, North Melbourne.

Bauwens, J., Tosone, C., 2010. Professional posttraumatic growth after a shared traumatic experience: Manhattan clinicians' perspectives on post 9/11 practice. Journal of Loss and Trauma 15 (6), 498–517.

Berger, R., 2012. Trauma and social work practice. In: Figley, C. (Ed.), The Encyclopedia of Trauma. Sage, Thousand Oaks, pp. 700–703.

Bonanno, G.A., 2004. Loss, trauma, and human resilience: have we underestimated the human capacity to thrive after extremely aversive events? American Psychologist 59 (1), 20–28.

Bonanno, G.A., Mancini, A.D., 2012. Beyond resilience and PTSD: mapping the heterogeneity of responses to potential trauma. Psychological trauma: theory, practice. Research and Policy 4 (1), 74–83.

Bonanno, G.A., Westphal, M., Mancini, A.D., 2011. Resilience to loss and potential trauma. Annual Review of Clinical Psychology 7, 511–535.

Breckenridge, J., James, K., 2010. Educating social work students in multifaceted interventions for trauma. Social Work Education 29 (3), 259–275.

Cronin, M.S., Ryan, D.M., Brier, D., 2007. Support for staff working in disaster situations: a social work perspective. International Social Work 50 (3), 370–382. http://dx.doi.org/10.1177/0020872807076050.

Cunningham, M., 2003. Impact of trauma work on social work clinicians: empirical findings. Social Work 48 (4), 451–459.

Cunningham, M., 2004. Teaching social workers about trauma: reducing the risks of vicarious traumatisation in the classroom. Journal of Social Work Education 40 (2), 305–318.

Cwikel, J., Kacen, L., Slonim-Nevo, V., 1993. Stress management consultation to Israeli social workers during the gulf war. Health and Social Work 18 (3), 172–183.

Dekel, R., Monson, C.M., 2010. Military-related post-traumatic stress disorder and family relations: current knowledge and future directions. Aggression and Violent Behavior 15 (4), 303–309.

Dohrenwend, B.P., Turner, J.B., Turse, N.A., Adams, B.G., Koenan, K.C., Marshall, R., 2006. The psychological risk of Vietnam for US veterans: a revisit with new data and methods. Science 313 (5789), 979–982.

Duffy, J., Ramon, S., Guru, S., Cemlyn, S., Lindsay, J., Nuttman-Shwartz, O., 2013. Developing a social work curriculum on political conflict: findings from an IASSW funded project. European Journal of Social Work. http://dx.doi.org/10.1080/13691457.2012.757215.

Figley, C.R. (Ed.), 1995. Compassion Fatigue: Coping with Secondary Traumatic Stress Disorder in Those Who Treat the Traumatized. Brunner Mazel, New York.

Fredman, S.J., Monson, C.M., Adair, K.C., 2011. Implementing cognitive-behavioral conjoint therapy for PTSD with the newest generation of veterans and their partners. Cognitive and Behavioral Practice 18 (1), 120–130.

Foa, E.B., Keane, T.M., Friedman, M.J. (Eds.), 2000. Effective Treatment for PTSD. Guilford Press, New York.

Fournier, R., 2002. A trauma education workshop for posttraumatic stress. Health & Social Work 27 (2), 113–124.

Gray, M., 1996. Moral theory for social work. Social Work/Maatskaplike Werk 32, 289–295.

Hare, I., 2004. Defining social work for the 21st century the international federation of social workers' revised definition of social work. International Social Work 47 (3), 407–424. http://dx.doi.org/10.1177/0020872804043973.

Hernandez, P., Gangsei, D., Engstrom, D., 2007. Vicarious resilience: a new concept in work with those who survive trauma. Family Process 46 (2), 229–241.

Hobfoll, S.E., 1989. Conservation of resources: a new attempt at conceptualizing stress. American Psychologist 44 (3), 513–524. http://dx.doi.org/10.1037/0003-066X.44.3.513.

Hobfoll, S.E., Dunahoo, C.A., Monnier, J., 1995. Conservation of resources and traumatic stress. In: Freedy, J.R., Hobfoll, S.E. (Eds.), Traumatic Stress: From Theory to Practice. Plenum Press, New York, pp. 29–47.

Joseph, S., Murphy, D., 2013. Trauma: a unifying concept for social work. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs207.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Lee, C.W., Drummond, P.D., 2008. Effect of eye movement versus therapist instructions on the processing of distressing memories. Journal of Anxiety Disorders 22 (5), 801–808.

Linley, P.A., Joseph, S., 2004. Positive change following trauma and adversity: a review. Journal of Traumatic Stress 17 (1), 11–21.

Lobley, K. (Ed.), 2007. In Case of Emergency: How Australia Deals with Disaster and People Who Confront the Unexpected. ETN Communication, Artarmon.

Marlowe, J., Adamson, C., 2011. Teaching trauma: critically engaging a troublesome term, social work education. The International Journal 30 (6), 623–634.

Maslach, C., Leiter, M.P., 1997. The Truth about Burnout: How Organizations Cause Personal Stress and what to Do about it? Jossey-Bass, San Francisco.

National Association of Social Workers, 2008. Military Service-Related PTSD. Retrieved from: http://www.socialworkers.org/research/nasw Research/0907Military/default.asp.

Pearlman, L., Maclan, P., 1995. Vicarious traumatization: an empirical study of the effects of trauma work on trauma therapists. Professional Psychology: Research and Practice 26 (6), 558–565.

Pearlman, L., Saakvitne, K., 1995. Trauma and the Therapist: Countertransference and Vicarious Traumatization in Psychotherapy with Incest Survivors. Norton, New York.

Stamm, B.H., 2002. Measuring compassion satisfaction as well as fatigue: developmental history of the fatigue and satisfaction test. In: Figley, C.R. (Ed.), Treating Compassion Fatigue. Brunner Mazel, New York, pp. 107–119.

Solomon, Z., 1993. Combat Stress Reaction: The Enduring Toll of War. Plenum Press, New York.

Stevenson, A.D., Phillips, C.B., Anderson, K.J., 2011. Resilience among doctors who work in challenging areas. British Journal of General Practice 61 (588), 404–410.

Tedeschi, R.G., Calhoun, L.G., 1996. The posttraumatic growth inventory: measuring the positive legacy of trauma. Journal of Traumatic Stress 9 (3), 455–471.

Tosone, C., Nuttman-Shwartz, O., Stephens, T., 2012. Shared trauma: when the professional is personal. Clinical Journal of Social Work 40 (2), 231–239. http://dx.doi.org/10.1007/s10615-012-0395-0.

Ungar, M., 2008. Resilience across cultures. British Journal of Social Work 38 (2), 218–235.

Weinberg, H., Nuttman-Shwartz, O., Gilmore, M., 2005. Trauma groups: an overview. Group Analysis 38 (2), 187–205.

Potlatch in Anthropology

ME Harkin, University of Wyoming, Laramie, WY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Potlatch' is a key anthropological concept, which has spread to many other fields of study. It refers to a ceremonial exchange system that arose in the nineteenth century on the Northwest Coast of North America. Although it takes various forms, most attention has been paid to the agonistic displays of the Kwakiutl potlatch.

'Potlatch' is a term, derived from Chinook jargon, used to refer to a variety of exchange practices among aboriginal groups on the Northwest Coast of North America. These practices all involve ceremonial exchange as the central feature, although secondary features, as well as the function and meaning of the exchange, vary considerably from group to group. The concept of 'potlatch' is an anthropological reification of a range of cultural practices characteristic of this area, all of which bear a certain family resemblance, but which cannot be considered a unitary phenomenon. Most discussion of the potlatch focuses on the Kwakiutl (Kwakwaka'wakw) of Vancouver Island, who settled around Fort Rupert in the late nineteenth century and were studied by Franz Boas and his research partner George Hunt. Among the elements stressed by Boas and his followers were agonistic display, balanced reciprocity, loans and repayment with interest, and extreme psychological states. Other scholars, especially those working in the northern region, have stressed themes of cooperation, harmony, and the creation of public meaning. Materialist scholars have stressed the role of the potlatch in spurring economic overproduction or in redistributing resources across microenvironments. The potlatch was outlawed by the Canadian government from 1885 to 1951. However, it persists among many groups in Canada and the United States to this day.

In the broadest sense, 'potlatch' refers to a ceremonial exchange complex of aboriginal peoples of the Northwest Coast of North America. Originating in precontact practices, the potlatch evolved in the nineteenth century into a quasi-commercial system of acquiring prestige through large gifts – and occasionally destruction – of native and trade goods. Among the Kwakiutl and neighboring groups, the potlatch came to include a supplemental economy of loans and interest, allowing for presentations so large that they beggared the host. This extreme form, appearing between roughly 1870 and 1925, earned it a central place in the ethnographic imagination.

History and Origins

Precontact Trade and Ceremonial Exchange

'Potlatch' derives from the Chinook jargon word *patshatl*, to give. Chinook jargon was a trade language centered in the Columbia River basin, but extending north to the central Northwest Coast. It predates European contact and served as a *lingua franca* among groups engaged in aboriginal trade. Archeological evidence suggests the existence of intergroup ceremonial exchange alongside precontact trading relations. However, it is clear that such exchange was both quantitatively and qualitatively different from the historic potlatch. It involved modest amounts of animal pelts and a few types of prestige goods and was known under different labels in different areas. It reaffirmed but likely did not alter rank relations within and between groups. Thus, it is fair to say that the potlatch, as a translocal institution, was a product of the fur trade era.

The Fur Trade: Influx of Wealth

The early fur trade of the 1770s and 1780s involved Russians, Spanish, British, and Americans. Trade focused on the sea otter and was conducted primarily from ships. This had a limited effect on aboriginal groups. However, by the late 1780s, the sheer number of trading vessels, and the shift to land-based trading, had begun to bring changes. The transfer of the Spanish colony on Juan de Fuca Strait to the British in 1793 marked the beginning of a process that would see Hudson's Bay Company forts established all along the coast, and eventually a substantial colony at Victoria. To the north, the Russians established a fort at Sitka, in Tlingit territory, and to the south, Americans established posts on the Columbia River in the 1810s. The shift to land-based trade brought with it an influx of wealth and trade goods, as well as epidemic disease, and a broad range of new cultural influences.

Despite the negative effects of contact, the early nineteenth century was characterized by an efflorescence of art and expressive culture on the Northwest Coast. Nowhere is this seen more clearly than in the plastic arts. The introduction of metal woodworking tools and, somewhat later, commercial paint, allowed for an explosion of art in the form of masks, architectural elements, storage vessels, and free-standing monumental sculpture. Such objects were commissioned with new commercial wealth; these objects, in turn, were vehicles by which this wealth was displayed and validated. Wealth had always been closely associated with rank. The problem arose in that rank was ideologically a closed system: a limited number of ranked positions, associated with titles, privileges, and symbolic property ('crests') were believed to exist, and these positions were filled in accordance with regular principles of descent and succession. However, during the fur trade period, *nouveaux riches* rose to prominence. New crests were created for them by artists who were experimenting with both themes and techniques. Simultaneously, such persons also staked claim to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12135-X

unused traditional crests or those in dispute; given high mortality from a series of epidemics, application of the rules of normal succession (e.g., to nephew or son-in-law) was often impossible. The potlatch was the forum in which such claims were made public and were either validated or rejected by the larger community.

The Dominion Period: Response to Traumatic Change

After the 1860s, which saw both Canadian Dominion and the purchase of Alaska by the United States, aboriginal groups were increasingly encapsulated within Euro-American political, social, and economic institutions. Simultaneously, the full effects of disease mortality were being felt, producing upheavals in aboriginal society. The potlatch, an effective means of dealing with demographic and social change, was mobilized to address issues such as the amalgamation of groups, the lack of personnel to fill important rank positions, changing relations with Euro-Americans, and changes in the relative status of local groups. In the process, potlatches became larger and more expansive. A wide range of aboriginal practices, such as shamanic winter dances, became incorporated within the potlatch. Simultaneously, the potlatch replaced intergroup violence, which was suppressed, as a means for gaining prestige at neighbors' expense.

During this period, the potlatch was suppressed by Canadian authorities, who passed a 'Potlatch Law' in 1885. Although unenforceable for most of its existence, it did result in famous prosecutions and convictions of Kwakiutl potlatchers in 1923. In addition to criminal punishment, the state seized 'potlatch goods,' including ritual objects. These provided the basis of ethnological collections at museums in Ottawa and Toronto. In the 1970s and 1980s, these items were returned to Kwakiutl communities at Cape Mudge and Alert Bay, which have built museums to house them.

Geographical Range and Variation

The Northern Potlatch

The northern region, consisting of matrilineal societies, used the potlatch primarily in the context of funerals and succession. For the Tlingit and Haida, potlatches were between moieties, and stressed cooperation and solidarity between hosts and guests. Moieties, a division of society into two halves, provided an image of social and cosmological totality. In the case of funerals, the opposite moiety provided comfort and assured the reincarnation of the deceased. Nevertheless, a minor theme of competition was often introduced, with two separate guest groups attempting to outdo each other in the degree of 'comfort' provided to the host. To the south, Tsimshian (including Nishga and Gitksan) potlatches were similar, but, lacking moieties, were somewhat more agonistic. They were associated with marriages, as well as other life-crisis rituals; marriage was seen in oppositional terms. In addition, potlatches were used to make territorial claims.

The Southern Potlatch

South of the Tsimshian, social organization was cognatic, with a matrilineal orientation in the north becoming more patrilineal in the south. Among the Nootka (Nuu-chah-nulth), potlatches were essentially private entertainments sponsored by a chief for his people, and thus stressed social harmony. Elsewhere, however, potlatches were primarily agonistic. Marriage potlatches, including those held to repay bridewealth, were often couched in military terms. As group membership and rank were less fixed in cognatic societies, they were more the object of political manipulation, and thus of representation in the potlatch.

The Kwakiutl Potlatch

The extreme example of agonistic prestation was provided by the Kwakiutl, who live on northeastern Vancouver Island and the adjacent islands and mainland. During the 60 years after Dominion, the potlatch intensified to the point that Benedict (1934), in her colorful if flawed analysis of Boas' material (Boas, 1966), described the Kwakiutl as 'megalomaniacal' and 'paranoid.' Fierce competition among brothers, cousins, and affines for succession to titles, the existence of numerous latent titles, the rise of a *nouveaux riche* class called 'Eagles,' and a system of intergroup ranking, all contributed to the intensity of the potlatch, which participants viewed as warfare by other means. Potlatches opened with guests arriving in war regalia, and a mock battle on the beach, reminiscent of the murderous raids that were endemic to this area.

Kwakiutl potlatches were predicated on rivalries between chiefs and, by extension, their groups. In addition to the usual means of amassing wealth, would-be hosts engaged in a system of loans with 100% interest, which allowed lenders to build fortunes quickly and borrowers to obtain last-minute infusions. The currency was Hudson's Bay Company blankets, supplemented and eventually replaced by Canadian currency. Copper shields were the ultimate valuable; bid up in previous transactions, 'coppers' reached cash values of several thousand dollars. In addition, trade goods played an increasing role. By the start of the twentieth century, manufactured items such as shotguns, sewing machines, and pool tables were given away, along with quantities of flour, rice, cloth, and utensils. In some of the most dramatic potlatches, goods including coppers were destroyed.

Functional Accounts

The Potlatch as Total Social Phenomenon

The French sociologist Marcel Mauss (1990) noted that the potlatch is an integral element of the rank system of Northwest Coast societies. It is a mechanism by which claims to rank status are validated or challenged, and changes in social order are communicated to the larger community. The potlatch rests on opposition between host and guest. The host – a chief or high-ranking person and, by extension, his social group – was concerned with conducting certain types of business: naming heirs, contracting marriages, conducting funerals, and proclaiming new titles and privileges. This business was witnessed, in a juridical sense, by the guests, who were paid for their service. Potlatch gifts went beyond mere payment for services, however, since they were also taken as prima facie evidence for the validity of the claim.

Wealth was a necessary concomitant of high rank. High-rank positions implied a certain degree of extraordinary power, similar to the *mana* of Polynesian chiefs, which necessarily resulted in wealth. Inability to provide an abundant potlatch would seriously undermine the legitimacy of the claimant's assertions. By the same token, an extravagant potlatch could 'fasten on' a title, merging the holder's social identity with the permanent qualities of the title. Such acts were seen as a sort of reincarnation, by which the living title holder became incorporated in the title, which encompassed the holder's predecessors and mythical founder. For chiefly positions, the ascendance of a new holder was often marked by events symbolic of the (re)creation of the social world, including human sacrifice, the razing of old chiefly houses and poles, and their replacement by new ones.

In chiefly succession potlatches, a broad range of political, legal, and economic issues were implicated. In all groups, chiefly titles involved territorial claims of some sort. These titles were based on mythical narratives, set locally and taking place at the beginning of human history, thus staking aboriginal claim to land. Among most groups, such territorial claims were very specific and resembled Western notions of ownership. In all Northwest Coast cultures, chiefs were minimally seen as stewards of the land and its inhabitants were responsible for regulating relations between human and nonhuman populations (which were believed to include more than the flora and fauna recognized by biological science). This was accomplished by ritual means, including prayers, attraction magic, and first fruit rituals. The latter were used to regulate the seasons on fish and game. Although the informing ideology was moral, such regulation had the effect of conserving and managing resources in a sustainable manner.

The potlatch thus involved relations between the social group and the nonhuman outside world, the prehuman myth world, and the group's ancestors. In this way, the group was situated within a larger universe and was thus defined and delimited. Potlatches likewise served to delineate the internal organization of society and the functional relations obtaining between its components. In a general sense, one potlatches to one's affines, who constitute the friendly other. In northern matrilineal groups, this opposition was absolutely clear. Especially in the moiety system of the Tlingit and Haida, one's affines were a separate corporate group; the cooperation of the two sides presented an image of social totality cosmological in scope.

The Potlatch and Marriage Exchange

Potlatches were associated with marriage in most areas; even where they were not, the potlatch was transacted between affines. Politically, potlatches were vital in renewing alliances among groups. Among the northern groups, whose aristocrats practiced preferential cross-cousin marriage, long-term stable exchange relations involving two or more local groups were established. Such groups provided political, military, and economic support, as well as external validation of chiefly claims. Among cognatic groups, no such long-term structural relations existed but were rather produced and reproduced in a series of exchanges tied to specific marriages (payment and repayment of bridewealth, naming of children, and funeral of a spouse). The connection between marriage and potlatching was so strong that in groups such as the Kwakiutl, Bella Coola, and Heiltsuk, aristocrats would engage in serial monogamy and even fictive marriage in order to obtain prestige from potlatching. Potlatching varied along several dimensions (size, number, occasion, agonism, and others) depending on the precise configuration of kinship, descent, and marriage rules. Among the Kwakiutl and neighboring groups, marriage was viewed as a limited contract. The repayment of the bridewealth (often many times over) in a potlatch by the wife's natal family effectively ended that contract. For aristocratic women (nonnobles were said not to marry at all, since they did not potlatch), a typical career involved marrying multiple times.

Potlatching is a profoundly gendered institution. It is a prototypical masculine activity (along with fishing, hunting, and warfare). Men derive both tangible and intangible (but culturally highly important) property, as well as prestige and political power. Women gain as well, but can be viewed as another class of object exchanged among men. Indeed, Lévi-Strauss's notion of the circulation of women and property between groups has been shown by Rosman and Rubel to have considerable relevance to Northwest Coast ethnology (Rosman and Rubel, 1971). At the same time, women are consciously advancing their own careers. While not eligible for chiefly titles, they nevertheless were said to become more 'heavy' with successive potlatches. Heaviness is the archetypal quality of aristocrats; they are seen as more 'real' than others. Heaviness is symbolically associated with copper, the form and symbol of wealth in traditional Northwest Coast cultures. As women were married or had their bridewealth repaid, they were said to be 'heavy with wealth.' This was (and still is) symbolized by the sewing of wealth items, including Canadian currency, into the dress worn by a bride.

The Potlatch and Ecological Variation

Since the 1960s, materialist scholars have viewed the potlatch as a virtual homeostatic mechanism, distributing personnel and food efficiently through space and time. The earlier conception of the Northwest Coast as free from food pressures is justly rejected. The potlatch is seen as having four distinct but interrelated functions: (1) spurring overproduction, (2) encouraging storage of food, (3) allowing for the exchange of food for prestige goods, and (4) regulating group membership and territorial boundaries, matching resources with personnel. This approach assumes local variation in resources, a condition that does not exist everywhere but is more pronounced in the north. Item (4) was affected more often by warfare than by potlatching. However, these remain at least potential functions of the potlatch. More recently, scholars have examined the ways in which potlatching, especially in its more cooperative versions, shared resources, and hence an interest in preserving those resources, among different groups, thus enhancing the sustainability of the fisheries (Trosper, 2009).

Rank, Class, and Political Power

Northwest Coast society was stratified by both rank and class. Potlatching was the domain of chiefs and nobles; however, both commoners and slaves played important roles in amassing resources. By definition, only persons with titles were in

a position to potlatch. Potlatching thus reinforced vertical, as well as horizontal, social divisions. Simultaneously, potlatching was the mechanism by which social mobility was integrated into the social system.

The potlatch both manifested and represented political power. By drawing together a large number of guests, especially if they were from far away or their chiefs were especially high ranking, a host was demonstrating his prestige and the size of his political network. The material goods and food distributed were an index of his land holdings and their productivity, which ultimately depended upon the chief's legitimacy. Potlatches were fora for authoritative political discourse, defined technically as monologic, multichannel, redundant communication. A significant aspect of this communication was an illocutionary speech act. This was supplemented with nonverbal media, including plastic arts and dance. The political success of a potlatch was usually measured in terms of the degree to which these utterances remained unchallenged. They may be challenged directly, in rivalry gestures, or more subtly through various sorts of corrosive discourse, including humor. However, in some groups, such as the Kwakiutl, such elements were institutionalized within the structure of the potlatch. This had the effect of heightening tensions within the potlatch and giving it a more gamelike structure, in which a wide range of outcomes was possible.

These agonistic aspects of the potlatch especially in its Kwakiutl version have drawn the attention of anthropologists and scholars outside the region, making it one of the most famous examples from the anthropological literature. Generosity so great as to be unreciprocable, the symbolic violence of giving on this scale, the destruction of property in dramatic action, the humiliation of rivals: these features have made it a sort of logical limit case and hence useful in theoretical argument. For this reason, David Graeber has employed it to represent the pole of nonreciprocal exchange, which he sees as the origin of money and of modern political power (Graeber, 2001). In this case, logical clarity is purchased at the price of ethnographic veracity. Even among the Kwakiutl of the era of the classic potlatch, various modalities of exchange were employed to a variety of social, political, and religious ends.

Problems of Reference

Reification

Given the degree of variation over space and time, it is unclear that one can say that the potlatch, *as such*, exists. Rather, the understanding of it is conditioned by reifications imposed by government agents and missionaries, anthropologists, and aboriginal people themselves. The Potlatch Law of 1885 first tried to specify "the Indian festival known as the 'Potlatch.'' Problems of reference significantly impeded successful prosecution of the law. During this intervening time, the symbolic labor of many people served to bring descriptions and practice closer together.

Invention of Tradition

Aboriginal people themselves reified the potlatch by defining their ceremonial praxis as potlatch, a foreign word and concept mediated through Euro-American society. By the time Franz Boas recorded ethnographic data among the Kwakiutl in the 1890s (Boas, 1966), the concept of potlatch was present. This consciousness of 'the potlatch,' on the part of both Boas and the Kwakiutl, was shaped largely by the existence of the Potlatch Law. Clearly, aboriginal practice and self-awareness of practice were conditioned by exogenous expectations and categories. It is not surprising that during a period of rapid change such an invention of tradition would occur. It allowed aboriginals to retain symbolic control of their destiny, by redefining novel structures and practices as traditional. Because of this useful quality, potlatching remains a common practice today.

The Potlatch as Anthropological Totem

The potlatch remains one of the most durable anthropological concepts, despite the reanalyses and deconstructions to which it has been subjected. Beginning with Boas, anthropologists defined potlatch as both self and other, transparent and exotic. This contradictory and mediational quality has made potlatch an especially useful metaphorical vehicle. Potlatch has been viewed as, inter alia, metaphorical banking, warfare, and mental illness (Benedict, 1934; Boas, 1966; Codere, 1950). At the same time, various aspects of Western societies, including politics and consumer capitalism, have been compared to potlatch. Its hybrid nature and its origins in the desire on the part of Euro-Americans to describe indigenous peoples, and indigenous peoples to represent change and continuity, make it an attractive concept, but one that resists critical usage.

See also: Boas, Franz (1858–1942); Colonialism, Anthropology of; Elites, Anthropology of; Exchange in Anthropology; Exchange: Social; Mauss, Marcel (1872–1950); Political Anthropology; Totemism.

Bibliography

Adams, J.W., 1973. Gitksan Potlatch: Population Flux, Resource Ownership, and Reciprocity. Holt, Rinehart, and Winston, Toronto, Canada.

Bataille, G., 1985. In: Stoekl, A., (Ed. and Trans.), Visions of Excess: Selected Writings, 1927–1939. University of Minnesota Press, Minneapolis, MN.

Benedict, R., 1934. Patterns of Culture. Houghton Mifflin, Boston.

Boas, F., 1966. Kwakiutl Ethnography. University of Chicago Press, Chicago.

Bracken, C., 1997. The Potlatch Papers: A Colonial Case History. University of Chicago Press, Chicago.

Codere, H., 1950. Fighting with Property: A Study of Kwakiutl Potlatching and Warfare, 1792–1930. Augustin, New York.

Cole, D., Chaikin, I., 1990. An Iron Hand upon the People: The Law against the Potlatch on the Northwest Coast. University of Washington Press, Seattle, WA.

Drucker, P., Heizer, R.F., 1967. To Make My Name Good: A Reexamination of the Southern Kwakiutl Potlatch. University of California Press, Berkeley, CA.

Graeber, D., 2001. Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams. Palgrave Macmillan, New York.

Harkin, M.E., 1997. The Heiltsuks: Dialogues of Culture and History on the Northwest Coast. University of Nebraska Press, Lincoln, NE.

Kan, S., 1989. Symbolic Immortality: The Tlingit Potlatch of the Nineteenth Century. Smithsonian Institution, Washington, DC.

Lévi-Strauss, C., 1982. The Way of the Masks. University of Washington Press, Seattle, WA.
Mauss, M., 1990. The Gift: The Form and Reason for Exchange in Primitive Societies (W.D. Hall, Trans.). Routledge, London.
Mauzé, M., Harkin, M., Kan, S., 2004. Coming to Shore: Northwest Coast Ethnology, Traditions, and Visions. University of Nebraska Press, Lincoln, NE.
Rosman, A., Rubel, P.G., 1971. Feasting with Mine Enemy: Rank and Exchange among Northwest Coast Societies. Columbia University Press, New York.
Schulte-Tenckhoff, I., 1986. Potlatch: Conquète et Invention: Réflexion sur un Concept Anthropologique. Éditions d'en Bas, Lausanne.
Seguin, M., 1985. Interpretive Contexts for Traditional and Current Coast Tsimshian Feasts. National Museums of Canada, Ottawa, Canada.
Trosper, R., 2009. Resilience, Reciprocity, and Ecological Economics: Northwest Coast Sustainability. Routledge, New York.
Walens, S., 1981. Feasting with Cannibals: Metaphor and Morality in Nineteenth Century Kwakiutl Culture. Princeton University Press, Princeton, NJ.

Poverty, Culture of

Phillippe Bourgois, University of California, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11904–11907, © 2001, Elsevier Ltd.

Abstract

The 'culture of poverty' is a concept popularized by the anthropologist Oscar Lewis during the 1960s in his best-selling ethnographic realist books on family life among the urban poor. Drawing from Freudian culture and personality theory, which dominated US anthropology in the post-World War II period, Lewis listed over 50 traits that he claimed were shared by approximately 20% of the poor, including 'orality,' 'strong present-time orientation,' and a 'high tolerance for psychological pathology.' These traits were said to be transmitted cross-generationally within families and prevented individuals from taking advantage of economic opportunities. Despite Lewis's social democratic politics, his culture of poverty concept resonated with the moralistic condemnation of the unworthy poor deeply ingrained in US popular ideology. It spawned a polemical response from social scientists who criticized its blame-the-victim and psychological reductionist implications. Lewis himself was not concerned with theory, and did not particularly believe that his ethnographic work documented the culture of poverty. The vituperative debates during the 1970s through to the 1990s over whether or not a culture of poverty really exists were not fruitful theoretically. They tend to degenerate into political name calling or into an empirical denial of the lived experience of social suffering among the persistently poor. The uses and misuses of the culture of poverty illustrate how research on social inequality reflects societal biases. They also demonstrate the inadequacy of the culture concept when it is used to explain hierarchy in an essentializing, atheoretical vacuum that ignores history and structural power dynamics.

The culture of poverty concept was developed in the USA during the 1960s primarily through the best- selling ethnographic realist publications of the cultural anthropologist Oscar Lewis, who tape-recorded eloquent life histories of the urban poor. He reprinted numerous versions of his definition of the term 'culture of poverty' in short journal articles and also in the introductions to his books on family life among Mexicans, Puerto Ricans, and Cubans living in shanty towns and ghettos (Lewis, 1961, 1966a,b, 1967). Lewis's culture of poverty struck an academic identity politics nerve, and at the turn of the millennium the concept remained enmired in a bitter polemic over how to analyze and engage politically the persistence of poverty in the midst of postindustrial plenty.

Ideological Backdrop to the Culture of Poverty

In the USA, irrespective of the theoretical orientation of researchers, most discussions on poverty polarize around value judgments concerning individual self-worth or around racial/ethnic stereotypes. US attitudes towards poverty are rooted in the country's colonial Calvinist/Puritanical heritage and are exacerbated by the historical importance of racialized hierarchies that have legitimized genocide, slavery, colonization, and immigration control. This helps explain why the culture of poverty concept continues to generate so much emotional heat while shedding so little conceptual light. The uses and misuses of the concept offer a fascinating case study in the sociology of knowledge illustrating the political interfaces between theory, empiricism, art, and ethnocentric moralizing in the social sciences.

Poverty research throughout history has been more successful at reflecting the biases of an investigator's society than at analyzing the experience of poverty. The state of poverty research in any given country emerges almost as a litmus for gauging contemporary social attitudes toward inequality and marginaliz- ation. For example, while Lewis's books are read by a US public as an individualistic interpretation of the persistence of poverty that blames victims, in France his work is interpreted as a critique of society's failure to remedy the injuries of class-based inequality under free market capitalism.

Defining the Culture of Poverty

The socialist sociologist Michael Harrington was the first prominent academic to use the phrase 'culture of poverty' in a major publication. His book, *The Other America*, documented rural poverty in Appalachia and represented a moral call to action that anticipated the War on Poverty initiated by President Johnson in 1964 (Harrington, 1962). As a first-generation son of impoverished Jewish immigrants who was influenced by Marxism in his youth, Lewis shared Harrington's social democratic commitment to combating poverty (Rigdon, 1988; Harvey and Reed, 1996). Ironically, however, Lewis's popularization of the culture of poverty concept is said to have tolled an intellectual death knell to the optimistic idealism of the mid-1960s that advocated eradicating poverty through direct state intervention (Katz, 1989). This is because Lewis's definition of the culture of poverty stressed that a significant minority of the poor (approximately 20%) were trapped in self-perpetuating cycles of dysfunctional behaviors and attitudes: 'By the time slum children are age six or seven they have usually absorbed the basic values and attitudes of their subculture and are not psychologically geared to take full advantage of changing conditions or increased opportunities' (Lewis, 1965a: p. xlv).

This kind of psychological reductionist and individualistic interpretation of the persistence of poverty resonated with US popular blame-the-victim discourse. Ironically, in the same articles or book introductions in which he defined the culture of poverty, Lewis also included radical political statements

contradictory to the implication that poverty is caused by self-perpetuating deficient value systems. For example, in his *Scientific American* version of 'The Culture of Poverty' he quotes Frantz Fanon, praises Castro's Cuba, and criticizes 'free-enterprise, pre-welfare-state stage capitalism' for spawning the culture of poverty (Lewis, 1966b). At the same time he states that 'it is much more difficult to undo the culture of poverty than to cure poverty itself,' and advocates 'psychiatric treatment' for poverty in the USA (Lewis, 1966b: p. 25).

In other words, the culture of poverty concept was confused theoretically at its inception. Unfortunately, Lewis never managed to clarify what he intended to mean. His published correspondence reveals that he was profoundly disturbed by the blame-the-victim interpretation of the causes of poverty that he triggered in the USA: 'There is nothing in the concept that puts the onus of poverty on the character of the poor' (Lewis, 1967: p. 499).

The notion of a culture of poverty, consequently, should not be treated as a full-blown theory. As presented by Lewis, it was merely a bundle of some 70 traits which he did not link to a particular processual or dynamic logic. In fact, he never even listed all 70 of the traits that he claimed existed. The theoretical sloppiness of the culture of poverty concept may well be a product of the McCarthyist anticommunism in US academia that impinged on Lewis during his formative years in the 1950s. Four decades after their inception, his lists of culture of poverty character traits appear embarrassingly arbitrary, ethnocentric, and psychologically reductionist (Lewis, 1966a: p. xlviii):

> … a high incidence of maternal deprivation, of orality, of weak ego structure, confusion of sexual identification, a lack of impulse control, a strong present-time orientation with relatively little ability to defer gratification and to plan for the future, a sense of resignation and fatalism … male superiority, and a high tolerance for psychological pathology of all sorts.

The Policy Implications of the Culture of Poverty

In the late 1960s and 1970s Lewis's culture of poverty concept produced an outpouring of political and academic reactions, primarily in the USA but also in Mexico and Puerto Rico (Leacock, 1971; *Alteridades*, 1994). Despite conceiving of his work as a call for the expansion of public-sector intervention on behalf of the poor, his concept took the popular intellectual spotlight off the need for structural economic reform, and glossed over the social power dynamics revolving around class, ethnic, gender, and colonial inequalities. Policy makers, if they paid any attention to the culture of poverty concept, interpreted it as advocating the need to rehabilitate the deficient cultural value systems of poor children through the agency of psychiatric social workers. For example, in the applied policy realm Lewis consulted in the development of the Head Start Program in the USA (Rigdon, 1988), which has been criticized retrospectively as an attempt to 'take inner-city preschoolers who live in lead-painted, rat- infested tenements without steady heat or hot water, and metamorphose them into bright-eyed, upper-middle-class overachievers' (Bourgois, 1995: p. 325). Significantly, in the year 2001, Head Start was still identified by both liberals and conservatives as one of the only successful anti-poverty programs of the 1960s.

The Theoretical Implications of the Culture of Poverty

Unfortunately, most of the hostile academic responses to Lewis's culture of poverty concept have limited themselves to contradicting Lewis's empirical assertions, rather than to critiquing theoretically his psychological reductionism, his sloppy use of the culture concept, and his failure to link in a dynamic manner macrostructural political and economic forces – including gender power relations – to ideology, culture, and individual values (Valentine, 1968; Stack, 1974; for a political economy exception, see Katz, 1989; for a feminist literary criticism exception, see Franco, 1989). The bulk of the negative reaction hinges on a political concern for replacing the negative imagery of Lewis's painful but expressive ethnographic portraits of the everyday suffering of urbanized families, with positive images of the worthy poor, struggling for upward mobility against all odds. A late 1990s rehabilitating of the culture of poverty concept from a Marxist perspective dismissed the virulence of the US progressive reaction against the culture of poverty concept as a sectarian 'ultra Bolshevism' that swept the New Left when the general public was drifting ideologically to the Right following the War on Poverty. This precipitated a 'fruitless game of radical one-upmanship' among frustrated intellectuals, who were completely marginal to public political discourse, and who chose instead to devote their energies to proving their dedication to protecting the image of the poor (Harvey and Reed, 1996). More importantly, the urgent righteousness of the anti-culture of poverty social science literature is comparable to the polemics against Moynihan's 1967 patriarchal attribution of the 'tangle of pathology' in the black family as being the central cause for the persistence of poverty among urban African-Americans (Rainwater and Yancey, 1967).

The angry denial by academics of the existence of the types of violence and self-destructive behaviors described ethnographically by Lewis among the vulnerable families that he tape-recorded and described reveals how far removed intellectuals can be from the inner-city street. Although Lewis's writing deserves criticism for presenting his subjects in a decontextualized pornography of violence, sexuality, and emotional brutality, none of the behaviors or personalities described by Lewis should shock anyone who is familiar with everyday life in the US inner city or Latin-American shanty towns. On the contrary, Lewis's ethnographic realist descriptions, unfortunately, still ring true four decades after they were written. His disturbing material, however, demands theoretical explanation and political contextualization, and that is where both Lewis and his critics and admirers have largely failed. By confining the debate to a worthy versus unworthy poor dichotomy, the internecine squabbles between leftist, liberal, and conservative academics mimetically reproduce the right-wing hegemony in popular US culture that equates poverty with sinfulness.

Arguably, the polemics of righteousness that the culture of poverty prompted scared a generation of social scientists away from ethnographic analyses of inner-city poverty in the USA and, to a lesser extent, around the world (Wilson, 1987: pp. 13–16).

Indeed, accusations of supporting a 'culture of poverty interpretation' are still frequently invoked in polemical identity politics attacks between academics over any representation of poverty that is not flattering to poverty's victims (cf Lassalle and O'Dougherty, 1997). Hence the virulence of the 'underclass debate' in sociology spawned by William Julius Wilson's book *The Truly Disadvantaged* during the late 1980s through the 1990s (Katz, 1989; Wacquant, 1997).

From a theoretical perspective, the legacy of the culture ofpoverty debate has impoverished research in the social sciences on the phenomenon of social suffering, everyday violence, and the intimate experience of structural oppression in industrialized nations. Most importantly, by remaining mired in debates driven by identity politics, researchers have minimized the painful experience of day-to-day survival among the persistently poor. Epidemiological data on the associations between social class interpersonal violence, domestic violence, health outcomes, education outcomes, substance abuse, etc. are simply ignored by most poverty researchers in the USA.

The vacuum of critical intellectual engagement with the phenomenological experience of poverty has allowed right-wing academics subscribing to facile neoliberal blame-the-victim interpretations to capture popular imagination and policy debates – especially in the USA. Social or cultural reproduction theory, which emerged out of studies of poor youth at the intersection of the disciplines of education, sociology, and anthropology during the 1980s and early 1990s, offered a critical theoretical alternative. By focusing on the power dynamics ofthe interface between culture and social inequality, social/cultural reproduction theorists address the empirical reality of the existence of patterns of interpersonal self-destruction without obscuring structural political forces. Although vulnerable to critique for being overly functionalist, these theories allow for the reinscription of agency among the poor, as well as an autonomous role for culture in political economy (Bourdieu and Passeron, 1977; Willis, 1981).

Few serious social science researchers, if any, at the dawn of the twenty-first century would dare utilize the term 'culture of poverty' in their analysis, despite the fact that their empirical and theoretical work addresses cultural expressions of social suffering due to the entrenchment of urban poverty in industrialized and postindustrialized societies. The problematic analytical and political utility of the culture of poverty concept demonstrates how dangerously essentializing the phrase 'culture of …' can become with respect to any concept. Indeed, anthropologists cannot agree upon a useful definition for culture; nor do they understand how it operates without turning it into a black box of totalizing essences (Gupta and Ferguson, 1997).

The culture of poverty furore reminds us that academics fight so hard over so little especially when marginalized political perspectives are at stake. At the turn of the millennium, much of the world's population survives precariously in shanty towns, housing projects, tenements, and homeless encampments where mind-numbing, bone-crushing experiences of poverty engulf the socially vulnerable. Meanwhile, concerned academics continue to fiddle in their ivory towers, arguing over how to talk correctly about the structural violence of poverty.

See also: Class: Social; Income Distribution and Inequality: Measurement Issues; Income Inequality; Poverty Policy; Poverty, Sociology of; Poverty: Measurement and Analysis; Racism, History of; Racism, Sociology of; Underclass; Urban Poverty in Neighborhoods; Wealth Distribution.

Bibliography

Alteridades, 1994. Oscar Lewis Revisitado. Alteridades 4 Special Issue.
Bourdieu, P., Passeron, J.-C., 1977. Reproduction in Education, Society, and Culture. Sage Publications, London.
Bourgois, P.I., 1995. In Search of Respect: Selling Crack in El Barrio. Cambridge University Press, New York.
Franco, J., 1989. Plotting Women: Gender and Representation in Mexico. Columbia University Press, New York.
Gupta, A., Ferguson, J. (Eds.), 1997. Culture, Power, Place: Explorations in Critical Anthropology. Duke University Press, Durham, NC.
Harrington, M., 1962. The Other Poverty: Poverty in the United States. Macmillan, New York.
Harvey, D.L., Reed, M.H., 1996. The culture of poverty: an ideological analysis. Sociological Perspectives 39, 465–495.
Katz, M.B., 1989. The Undeserving Poor: From the War on Poverty to the War on Welfare. Pantheon Books, New York.
Lassalle, Y., O'Dougherty, M., 1997. Death without weeping: The violence of everyday life in Brazil. Radical History Review 69, 243–260.
Leacock, E., 1971. The Culture of Poverty: A Critique. Simon and Schuster, New York.
Lewis, O., 1961. The Children of Sanchez: Autobiography of a Mexican Family. Random House, New York.
Lewis, O., 1966a. La Vida: A Puerto Rican Family in the Culture of Poverty—San Juan and New York. Random House, New York.
Lewis, O., 1966b. The culture of poverty. Scientific American 215, 19–25.
Lewis, O., 1967. The children of Sanchez—Autobiography of a Mexican family. Current Anthropology 8, 480–500.
Rainwater, L., Yancey, W.L. (Eds.), 1967. The Moynihan Report and the Politics of Controversy. MIT Press, Cambridge, MA.
Rigdon, S.M., 1988. The Culture Facade: Art, Science, and Politics in the Work of Oscar Lewis. University of Illinois Press, Urbana, IL.
Stack, C.B., 1974. All Our Kin: Strategies for Survival in a Black Community. Harper and Row, New York.
Valentine, C.A., 1968. Culture and Poverty: Critique and Counterproposals. University of Chicago Press, Chicago.
Wacquant, J.D., 1997. Three pernicious premises in the study of the American ghetto. International Journal of Urban and Regional Research 21 (2), 341–353.
Willis, P., 1981. Learning to Labor. Columbia University Press, New York.
Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.

Poverty, Geography of

Anne Bonds, University of Wisconsin–Milwaukee, Milwaukee, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines themes in geographical research on poverty. Given the highly contested nature of the term, I begin with an overview of debates about definitions of poverty. This discussion contrasts absolute and relative poverty measures and critically assesses some of the assumptions embedded with dominant poverty definitions. Next, the article considers the place of geographic research on poverty in critical development studies, emphasizing how the idea of poverty has been fundamental to projects of development. The article then charts shifts in poverty dynamics associated with economic restructuring and welfare reform in the United States. Finally, I analyze the discursive construction of poverty and consider how ideas about poverty categorize and differentiate impoverished places and the poor.

Poverty remains one of the world's most pressing challenges. It is also the subject of intense debate and controversy that reveals a great deal about political economic dynamics, culture, and society. Extremes of wealth and poverty across the globe underlie conditions of health and illness, educational opportunity and illiteracy, land ownership and forced migrations, and employment protections or unregulated 'flexible' work. In spite of numerous agendas to eradicate poverty at various geographic scales over the past century, global poverty remains entrenched and billions of people continue to live in poverty. The World Bank reports that 1.3 billion people were living on less that $1.25 per day in 2008. In the United States, the 2010 Census reports poverty rates at 15.3%, which is an 18-year high (U.S. Census Bureau, 2011). Given these trends, it is perhaps not surprising that poverty remains a prominent focus within governmental policies and in the agendas of international institutions and nongovernmental organizations. The World Bank, for example, now operates under the slogan 'working for a world free of poverty' (www.worldbank.org) and the United Nation's Millennium Goals declare a commitment to ending extreme poverty by 2015 (http://www.un.org/millenniumgoals/poverty.shtml).

Geographic research has been key in demonstrating the highly uneven social and spatial distribution of poverty. At the global scale, the vast majority of the poor live in so-called 'developing nations'. A growing number of the world's poor live in cities. Rapid urbanization resulting from neoliberal policies that have reoriented previously more rural and agrarian economies has contributed to the rise of informal urban settlements and a deepening of urban poverty in the global South. The urbanization of poverty, however, exhibits regional variation. For example, in Central Asia and in Eastern Europe, poverty has deepened in rural areas, while in Latin America poverty in urban areas has intensified (Ravallian et al., 2007). Poverty is distinctly patterned in the United States and the United Kingdom. In the US, poverty rates are differentiated across cities (14.6%), suburbs (11.4%), and rural areas (16.5%) and there is a clear regional trend, with deeply entrenched poverty in areas of the South, Appalachia, the US–Mexico border, and on Native American reservations (Glasmeier, 2006; Kodras, 1997). In the UK, poverty is markedly higher in urban areas and has a clear North–South dimension, with more pronounced poverty in the North and lower rates of poverty in the South (Milbourne, 2010: pp. 161–162). Poverty is highly racialized, with communities of color and racial and ethnic minorities experiencing markedly higher rates of poverty across the globe. Significantly, women of all ages and from all racial and ethnic groups are far more likely to live in poverty than are their male counterparts, a phenomenon that is known as the 'feminization of poverty' (see Pearce, 1978).

Yet many critical human geographers contend that poverty is much more than a statistical measurement mapped onto particular places, arguing that it is better conceptualized as a lived experience shaped by both social and spatial relations. From this position, a reliance exclusively on poverty estimations in specific spatial units constructs poverty subjects and impoverished places in ways that can flatten and homogenize the diverse experiences of the poor and the multifaceted dimensions of poverty in a range of places. As Milbourne (2010) notes, a potential oversight emerging from a focus on statistical measurements is an emphasis on "spatial concentrations of poor people in particular (inner-city) places" and a minimization of "the presence of poverty in other (urban and rural) spaces" (2010: p. 161). Furthermore, these techniques have a tendency to emphasize spatial expressions and distributions of the poor rather than the actual *processes* producing poverty (Milbourne, 2010: p. 161; Lawson, 2012; Kodras, 1997).

Defining and Measuring Poverty

Poverty remains a widely contested and difficult concept to define. Conceptualized as a lack of access to essential basic needs and material goods, poverty is measure of economic health defined through various income thresholds designating the financial means necessary to support households of a particular size and composition (Glasmeier, 2006). Accordingly, the poverty rate is the percentage of the population with income levels lower than the poverty threshold. Poverty thresholds are absolute in that they are a standard applied across social, economic, and spatial contexts.

Increasingly, the efficacy of absolute poverty measures has been called into question. In particular, poverty researchers and policymakers have criticized the rigidity of absolute poverty measures in defining subsistence income levels. These challenges are founded on the basis that poverty varies across places and across time. For example, in the United States, while the absolute number of people in poverty in 1960 was higher (about 20%) than contemporary figures (15.2%), "a poor family today is much worse off compared to an average family than a poor family was in 1960" (Glasmeier, 2006: p. iv). That is to say, although the poverty rate itself has improved, the gap in incomes between an average poor family and an average middle-class family is much wider today than it was in 1960 (Glasmeier, 2006). Additionally, geographic differentiation means that what constitutes poverty and how poverty is experienced can differ dramatically across places. Factors such as local social and economic structures, cost of living differences, and access to public assistance and transportation systems all have the potential to shape the place-specific dynamics of poverty. Absolute poverty measures, however, do not take these geographic distinctions into account.

Another criticism of dominant measures of poverty centers on the way in which subsistence costs for the average family are conceptualized. In U.S. Census poverty measures, for instance, these costs are based upon an assumed 'basket of goods', originally conceptualized by the federal government in the 1960s (Glasmeier, 2006). This hypothetical 'basket of goods' has been contested for resting on outdated assumptions about family costs while minimizing or excluding expenses essential for contemporary household subsistence. For example, when this bundle of goods originated, it was presumed that families spent about one-third of their household income on food (Glasmeier, 2006) whereas today, food consumes much less than one-third of family income. Other expenses like housing, transportation, and healthcare now make up a significant portion of household expenditures but are minimized within Census definitions. Importantly, childcare, which comprises a significant and growing cost for households, is entirely excluded from census-defined 'basket of goods' (Glasmeier, 2006). Feminist critics argue that this conceptualization rests on a highly gendered household ideal comprised of a male breadwinner and a female family caretaker. This critique suggests that 'basket of goods' estimates do not reflect contemporary social and economic realities, namely the rise of – and growing necessity for – dual-income households, a shift that is further underscored by the decline of high-wage, manufacturing employment, and the rise of single-parent, female-headed households.

Taken together, these challenges lead some to suggest that poverty thresholds have the potential to significantly undercount the poor, which, in turn, has far-reaching implications. Federal poverty calculations shape policy responses and the flow of scare resources to low-income communities and programs for the poor. In the United States, these income thresholds unambiguously define who is and is not able to access public services, including critical housing and food supports. In addition, stark income-based definitions fail to account for growing numbers of the 'working poor', or employed individuals earning incomes slightly above the poverty line (see Shipler, 2004).

Some poverty experts advocate using relative rather than absolute poverty measures in order to account for some of these dynamics. Relative poverty defines poverty in relation to median income, rather than as a fixed level of money (Milbourne, 2010; Glasmeier, 2006). For example, in Great Britain, all households with income levels falling below 60% of median income are defined as poor. However, poverty measures based on income thresholds remain widely utilized as a means to count the poor. Calculating poverty through income guidelines allows for standard measures to be made across locations and time periods, facilitating historical comparison and the evaluation spatial patterns. Income-based, absolute poverty measures are therefore useful for broad-scale analysis, and their availability from governmental sources make them accessible for analysis. Nonetheless, interest is growing in alternative measures of poverty, including relative and 'living wage' measures, better suited for capturing geographic variation, new income demands, and the lived experience of poverty.

Geographic Conceptualizations of Poverty

While social inequality has long been of interest to geographers, poverty has been relatively marginalized in human geography compared to other social sciences such as sociology and anthropology. While not excluded from research, poverty has often been folded in with other concepts, such as inequality, deprivation, and welfare in scholarship on socioeconomic inequity and uneven development. Moreover, a bifurcation exists in geographic research on poverty between, on one hand, development scholarship theorizing the poverty in the global South as produced through uneven capital accumulation, unequal trade policies, and development interventions; and, on the other hand, poverty in the global North (especially the US and the UK) in relation to urban and rural restructuring and the reworking of national welfare systems. I reproduce this dichotomy here by separating the discussion of poverty research in the United States from the consideration of poverty and development. However, a burgeoning body of literature in geography theorizes poverty across this divide to connect processes producing poverty in the both the West and the so-called 'developing' world (e.g., Lawson, 2007, 2012; Roy, 2010).

Global Poverty and Development

Much of the research in geography on poverty has emphasized its systemic production through political economic relations. This scholarship connects poverty to the imperatives of capitalism and its attendant dynamics of uneven development and exploitation. These themes are particularly apparent in Marxian development geography, which has focused on poverty in the global South as emerging from histories of imperialism, colonization, and unequal access to power and resources with the global economic system. Geographic scholarship in critical development studies challenges conceptualizations that explain poverty in 'Third World' nations as an outcome of stalled economic progress, physical geography (i.e., the absence of resource endowments), and conflict (Lawson, 2007).

Instead, this research situates poverty as an outcome of historically unequal political economic relationships established through colonization and sustained through contemporary divisions of labor, development interventions, and international trade policy. These studies, therefore, emphasize poverty in relation to exploitation and structural transformations in economy.

Scholarship in this vein has been highly critical of international institutions like the World Bank and the International Monetary Fund (IMF) and their strategies to eradicate poverty. Rooted in Eurocentric conceptualizations of progress and the economic trajectories of the industrial West, these multilateral agencies, it is argued, emerged from a post-World War II, cold war context and have always represented the geopolitical and economic aims of the West. Against the background of decolonization, post-war rebuilding, and concerns about the spread of communism, poverty became a central justification for development interventions and the rise of newly established multilateral institutions focused on integrating recently independent nations into Western circuits of capital (Lawson, 2007; Roy, 2010; Escobar, 1995). As explained by Arturo Escobar (1995), "[i]t was commonly accepted in the early 1950s that if poor countries were not rescued from their poverty, they would succumb to communism" (p. 34). Across the decades of the 1950s and 1960s Keynesian modernization and industrial development policies were prioritized in the global South as a purported means to alleviate poverty, expand global markets, and support democracy. Poverty has thus long been entangled with the policies of influential multilateral institutions like the IMF and the World Bank, providing a key rationalization for the projects of development and the promotion of Western capitalist models of growth.

By the 1980s, neoliberal ideology was profoundly shaping the efforts of global institutions to address poverty. Broadly defined, neoliberal development approaches advocate market liberalization, deregulation, privatization, and fiscal austerity as a means to facilitate economic expansion. The 'Washington Consensus' of the 1990s (see Williamson, 1990) enshrined the rationalities of neoliberalism in development thought, advocating (among other things) the privatization of public services and state enterprises; the deregulation of industries and social welfare programs; the elimination of barriers to international trade; and the reduction of public welfare spending (Lawson, 2007: pp. 84–85; Peck, 2001). These logics were embedded in stabilization programs from the IMF and in structural adjustment programs (SAPs) promoted by the World Bank, a suite of development tactics often referred to as 'shock therapy' for the dramatic restructuring they engendered. Such efforts institutionalized free trade agendas and reduced public expenditures in heavily indebted global South nations and had the effect of dramatically reducing spending on poverty, education, and health care. The poor – comprised disproportionally of women and children – bore the brunt of structural adjustment's 'shocks', including the effects of dramatically reduced poverty supports, deteriorating infrastructure, wage declines, and destabilized domestic markets resulting from trade liberalization (see Benería and Feldman, 1992; Elson, 1991). Rather than alleviating poverty, such programs had the effect of entrenching poverty in the global South. While structural adjustment programs have since been abolished, international development institutions continue to operate from the neoliberal logic that poverty can be solved through market reforms and private enterprise (Roy, 2010; Lawson, 2007; Yapa, 1996).

More recently, development agencies have defined poverty through a social capabilities approach, which emphasizes the social dimensions of poverty including access to education and broader cultural resources facilitating the ability to live with dignity. Jeffery Sachs (2005), for example, sees persistent poverty as an outcome of market failure and believes that development must include infrastructure and health and that rich nations must promote the end of poverty through aid (p. 19). Sachs' ideas about poverty and development are encapsulated in the United Nation's Millennium Development Goals (MDGs).

Geographers and critical development scholars, however, are highly critical of Sachs' ventures. In particular, geographic research has challenged the ways in which his conceptualizations of poverty rest upon a geographic determinism that equates physical geography (rather than political economic processes) with poor places (e.g., Lawson, 2007: p. 13). This is particularly evident in Sachs' contention that 'adverse' geography sustains the 'poverty trap' (Sachs, 2005: p. 59). Scholarship critical of Sachs argues that his solutions to poverty remain entrenched within normative models of economic growth. Sachs' claims about a 'ladder of economic development', for example, are highly reminiscent of Rostow's model of stages of development, which has been much critiqued for its conceptualization of development as a linear trajectory of industrialization based upon the history of the West (Roy, 2010). According to Ananya Roy (2010), Sachs is representative of contemporary attitudes about global poverty that tend to emphasize charity and aid while eschewing questions about the global distribution of wealth, power, and privilege.

Ideas about poverty – and its origins, impacts, and solutions – have played a central role in contemporary development. Indeed, eradicating poverty has been a key focus of international development agendas, even as poverty alleviation strategies have often reinforced the political economic status quo and deepened the need for the most vulnerable.

Poverty, Welfare Reform, and Restructuring in the US

Geographers theorizing poverty in the United States and in the United Kingdom similarly share a concern with political economic transformation and the production of poverty. This research has especially emphasized the restructuring of Keynesian welfare states and the new forms of poverty emerging under the neoliberal transformation of labor markets and public policy (e.g., Kodras, 1997; Wolch and Dear, 1993; Peck, 2001; Wolch and Dihn, 2001; Lawson et al., 2010).

Over the past 30 years, the gap between the wealthy and the poor in the US has dramatically increased, even as neoliberal globalization has facilitated an unprecedented expansion of capital and markets (Piven, 2001). Since the 1970s, workers' median weekly earnings, adjusted for inflation, have *declined* and now remain far below their 1973 peak (Piven, 2001). This wage stagnation has disproportionately affected the poor, as

increasing numbers of full-time, low-wage workers remain unable to secure wages above the poverty level (Schram, 2000). A number of important policy and economic shifts underscore the dynamics of growing poverty and inequality amid increasing wealth. Markedly, following dramatic market downturns and economic crisis in the 1970s, states and firms adopted new strategies to reduce labor costs and to improve the 'flexibility' of global capital (Harvey, 1989). These strategies included a range of corporate policies that destabilized the Fordist contract between labor and capital, such as an increased reliance on part-time, subcontracted, and seasonal labor (Goode and Maskovsky, 2001). Perhaps the most pervasive cost-saving tactic was the relocation of production activities to low-wage locales, first in nonunionized regions of the South, and subsequently, across borders, in nations with low-wage rates and levels of unionization and fewer environmental and labor regulations.

The establishment of regional and multilateral trade agreements lowering trade barriers further enhanced this process and helped to establish a new international division of labor defined by the concentration of manufacturing and production activities in the global South and an expanded – and highly segmented – service sector in Western, industrialized nations. In the U.S., 'flexibility' in employment and production systems led to the rapid growth of a low-skill, poorly paid and benefited workforce, even as there was a parallel expansion of a highly paid financial services sector at the opposite end of the employment spectrum (Piven, 2001). This highly divided labor market was reinforced by state policies that embraced and institutionalized the neoliberal ethos of antiunionization, privatization, deregulation, and trade liberalization (Peck, 2001). These processes unfolded unevenly across space and produced new patterns of wealth and inequality. In industrialized nations, they effectively reduced the power of workers and created dramatic losses for the middle class and the poor.

A concomitant reworking of welfare programs and reductions in federal social welfare programs amplified the unequal impact of economic restructuring. In the United States, an assortment of programs for the poor are popularly referred to as 'welfare'. These programs include cash assistance to families, food vouchers, and Medicaid (health care coverage for the poor). However, the program most closely associated with welfare is the cash assistance program, originally called Aid to Families with Dependent Children (AFDC) and now referred to as Temporary Assistance to Needy Families (TANF). AFDC was created in 1935 following the Great Depression as part the Social Security Act. Though the Act was not without its flaws – most notably its exclusion of people of color, immigrants, and indigenous peoples from supports – it represents the largest and 'most effective antipoverty program' to be implemented in the U.S. (Lawson, 2012: p. 1).

The shifting aims of AFDC/TANF reveal a great deal about changing social and cultural norms in the United States. AFDC was initially designed to address childhood poverty among households without the presence of a male breadwinner (see Boushey, 1997). The federal program provided cash assistance to poor, predominantly white, single mothers who were deemed eligible by case workers and local administrators. Black women, who had long been active in the paid labor market, were largely excluded from the program as a consequence of racialized assumptions about employment outside of the home. Therefore, AFDC was geared toward white single mothers and, in fact, was conceived as a means to keep these women *out* of the labor market (Albelda and Whithorn, 2002).

Overtime, however, there was a significant shift in the emphasis in the program. By the 1970s and 1980s, AFDC became more focused on moving female recipients into the labor force (Mink, 1998). Supported by popular myths about 'welfare queens', increasingly the program gave priority to 'discouraging' behaviors seen to prevent women from paid work, such as childbearing outside of marriage. These efforts were promoted as a means of decreasing welfare abuse and 'dependence'; however, geographic and feminist scholarship has underscored the way in which such language elevated ideals about normative family and gender structures while simultaneously providing the political justification for widespread public welfare disinvestment (Wolch and Dihn, 2001; Mink, 1998; Albelda and Whithorn, 2002). In the 1980s, U.S. President Ronald Reagan – the figure perhaps best representing the conservative rally against liberal social welfare – initiated a 'war on welfare', which symbolized a significant shift away from the 'war on poverty' of the 1960s. Subsequently, federal spending on public assistance programs diminished throughout the 1980s.

Most decisively, in 1996, politicians from both political parities came together to create legislation that ended welfare, slashed funding for welfare programs, and devolved federal responsibility for welfare to the state and local levels. The 1996 Personal Responsibility and Work Opportunities Reconciliation Act (PRWORA) concluded 60 years of welfare as a nationally funded entitlement by eliminating AFDC and replacing it with Temporary Assistance to Needy Families (TANF). This gave states full authority to define who is poor, who is eligible for assistance, and the level of benefits available to those in need. The 1996 Act also mandated a series of new punitive guidelines designed to move recipients from cash assistance to work. Some of these reforms include a significant tightening of eligibility parameters, the imposition of new work requirements, the implementation of strict time limits for cash assistance (a 5-year lifetime limit), and the enactment of criteria designed to penalize particular behaviors, such as out of wedlock birth (Peck, 2001b; Mink, 1998; Wolch and Dihn, 2001). Following the implementation of welfare reform, the number of people receiving cash assistance decreased significantly (Schram, 2006).

Welfare reform is significant to geographic research on poverty for several reasons. First, it signals a clear shift away from Keynesian-style welfare, premised on a federally funded social safety net, to neoliberal workfare, centered on market liberalization, localization, and a retrenchment in social services (Peck, 2001b). This change represents a fundamental re-scaling of poverty policy, such that state and local governments were given increased discretion in the provision of social services, creating sharp spatial differences in definitions of need, eligibility, and entitlements across places. Thus, the rescaling of welfare exacerbates already uneven geographies of poverty.

Second, welfare reform represents a key change in the way that poverty is understood and managed. While poverty and

welfare policies have always been specifically racialized and gendered, social scientists and geographers have increasingly documented the way in which the neoliberal management of poverty criminalizes the poor (Peck, 2003; Bonds, 2009). This body of scholarship connects the retrenchment of social services, the unprecedented escalation of mass incarceration in the US, and the disciplinarity embedded within welfare regimes, to argue that there has been a transition away from the 'social welfare management' of poverty to its treatment through incarceration (Wacquant, 2001: p. 98). Accordingly, incarceration works as a mechanism to manage social divisions and rising insecurity resulting from economic deregulation and public welfare disinvestment.

Together, economic restructuring and welfare reform have reworked geographies of poverty in the United States. The dismantling of the liberal welfare state has created a new economic order that valorizes the market as the solution to poverty. Neoliberal restructuring has thus given rise to new spatialiaties as poverty governance is localized and shifting regulations reframe poverty.

Discursive Constructions of Poverty and Place

Geographic research emphasizes not just the political economic underpinnings of poverty, but also the ways in which it is lived, experienced, and discursively (re)produced. Constructions of poverty and cultural representations of the poor reflect social hierarchies and are infused with prevailing conventions about gender, race, ethnicity, and class. These discourses establish and reinforce distinctions between the deserving poor and the undeserving poor (see Katz, 1989). Discourses about poverty construct and position the poor and shape policies that have very real implications for everyday life.

A significant intervention of geographic research in this area has been its interrogation of the sociospatial underpinnings of discursive representations of poverty. In particular, this scholarship examines the ways in which dominant understandings spatialize poverty, making poverty a simplistic problem of geography rather than one stemming from complex structural and discursive processes. Put another way, place becomes the unproblematized explanatory basis for poverty. For example, Crump's (2002) analysis of discourses surrounding the spatial concentration of urban poverty reveals how the poor are pathologized by culture of poverty explanations that define poverty as reproduced in inner city 'breeding grounds' (p. 583). Such conceptualizations identify poverty as endemic to the inner city – and suggest that poverty can be solved simply by deconcentrating the poor – rather than viewing poverty and the inner city as mutually produced through the sociospatial processes of racism, deindustrialization, disinvestment, and out-migration (Crump, 2002). This uncomplicated understanding of poverty disconnects the urban ghetto – and the socioeconomic processes producing it – from other places (for example, the suburbs), thereby neutralizing the unequal spatial relations that sustain urban inequality. In a similar vein, Lawson (2007) examines the ways in which development narratives conceive of 'failed places' – those in need of outside interventions – and 'successful places' in ways that abstract place from the historical processes and interconnections that have produced poverty in some places and led to the accumulation of wealth in others (pp. 52–53).

Research on the discursive production of the poor and poor places thus challenges the ontological foundations of poverty (Lawson, 2012). Poverty statistics, calculations, and measurements not only enumerate the poor and 'failing places' in need of development (Lawson, 2007: p. 52), they are also generative. That is, these categories have the power to (re) create poverty subjects, poor places, and the stigma of being poor (Lawson, 2012). Consequently, processes of enumeration reinforce material and discursive processes of differentiation that shape actions and the way that individuals perceive their ability to act in the world.

The cultivation of poor subjects and spaces through discourses, measurements, and practices of calculation both reflects and rationalizes prevailing social conventions and ideologies. The material processes producing poverty, especially economic restructuring and the implementation of social policies that have diminished public welfare investments, are strengthened by discursive iterations that describe poverty as the outcome of individual flaws or place-specific deficiencies. Such explanations justify retrenchment, disciplinary practices, and particular development interventions and sustain inequitable economic relations.

Conclusion

In situating poverty and understanding it as both materially and discursively produced in place, critical research on poverty reveals that poverty is not merely a statistic or a '"moral" condition', but rather a 'geographically and historically contingent process' (Goode and Maskovsky, 2001: p. 17). By attending to both the material and discursive production of poverty, critical poverty research analyses how processes of neoliberal economic restructuring and capital accumulation are extended and discursively mapped on to particular bodies and places. Moreover, it is essential to think about poverty relationally; that is, to analyze poverty as created and sustained through sets of relationships between places, between social groups. Such a focus means that research on poverty should focus not just on the poor and on impoverished places, but on all social groups in order to understand how the production of the poverty is intimately connected to the production of the nonpoor (Lawson, 2012; Milbourne, 2004).

While mapping and enumerating poverty remains important for illustrating broad-scale dynamics and the spatial positioning of the poor, critics have warned of the problematic assumptions built in to techniques of measuring poverty, of the dangers of categorizing the poor, and of the limitations of emphasizing patterns over processes. Geographic scholarship on poverty underscores its multilayered dimensions and the interconnected nature of poverty dynamics. This emphasis challenges researchers to think across scales in order to conceptualize poverty as produced through the intertwining of global processes of capital accumulation, macrolevel policy shifts, and social and cultural norms as well as local particularities that give rise to distinct geographies of poverty and opportunity.

See also: Climate Change and Global Poverty; Development Geographies; Development Theory in Geography; Economic Restructuring: Geographic Aspects; Marxist Geography; Poverty Policy; Poverty in History; Poverty, Sociology of; Poverty: Measurement and Analysis; Urban Poverty in Neighborhoods; Welfare Reform; Welfare Retrenchment; Welfare and Work; Welfare.

Bibliography

Albelda, R., Whithorn, A., 2002. Lost Ground: Welfare Reform, Poverty, and Beyond. South End Press, Cambridge.

Benaria, L., Feldman, S., 1992. Unequal Burden: Economic Crises, Persistent Poverty, and Women's Work. Westview Press, Boulder.

Bonds, A., 2009. Discipline and devolution: constructions of poverty, race, and criminality in the politics of rural prison development. Antipode: A Radical Journal of Geography 41 (3), 414–438.

Boushey, H., 1997. Embracing discrimination? The interaction between low wage labor markets and policies in aid of the poor. In: Mutari, E., Boushey, H., Fraher, W. (Eds.), Gender and political economy: incorporating diversity into theory and policy. M.E. Sharpe, Armonk, N.Y.

Crump, J., 2002. Deconcentration by demolition: public housing, poverty, and urban policy. Environment and Planning D: Society and Space 20, 581–596.

Elson, D., 1991. Male Bias in the Development Process. Manchester University Press, Manchester.

Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.

Glasmeier, A.K., 2006. An Atlas of Poverty in American: One Nation, Pulling Apart. Routledge, New York.

Goode, J., Maskovsky, J., 2001. The New Poverty Studies: The Ethnography of Power, Politics, and Impoverished People in the United States. New York University Press, New York.

Harvey, D., 1989. The Condition of Post Modernity. Blackwell, Oxford.

Katz, M., 1989. The Undeserving Poor: From the War on Poverty to the War on Welfare. Pantheon Books, New York.

Kodras, J.E., 1997. The changing map of American poverty in an era of economic restructuring and political realignment. Economic Geography 73 (1), 67–93.

Lawson, V., 2007. Making Development Geography. Hodder Arnold, New York.

Lawson, V., Middle Class Poverty Politics Research Group, 2012. Decentering poverty studies: middle class alliances and the social construction of poverty. Singapore Journal of Tropical Geography 33, 1–19.

Lawson, V., Jarosz, L., Bonds, A., 2010. Articulations of poverty and place: dumping grounds and unseen grounds in the American Northwest. Annals of the Association of American Geographers 100 (3), 655–675.

Mink, G., 1998. Welfare's End. Cornell University Press, Ithica, New York.

Milbourne, P., 2004. The local geographies of poverty. Geoforum 35, 559–575.

Milbourne, P., 2010. The geographies of poverty and welfare. Geography Compass 4 (2), 158–171.

Pearce, D., 1978. The feminization of poverty: women, work, and welfare. Urban and Social Change Review 11 (1,2), 28–36.

Peck, J., 2003. Geography and public policy: mapping the penal state. Progress in Human Geography 27 (2), 222–232.

Peck, J., 2001. Neoliberalizing states: thin policies/hard outcomes. Progress in Human Geography 25 (3), 445–455.

Peck, J., 2001b. Workfare States. The Guilford Press, New York and London.

Piven, F.F., 2001. Welfare reform and the economic and cultural reconstruction of low-wage labor markets. In: Goode, J., Maskovsky, J. (Eds.), The New Poverty Studies: The Ethnography of Power, Politics, and Impoverished People in the United States. New York University Press, New York, pp. 135–151.

Ravallion, M., Chen, S., Sangraula, P., 2007. New evidence on the urbanization of global poverty. Population and Development Review 33 (4), 667–701.

Roy, A., 2010. Poverty Capital: Microfinance and the Making of Development. Routledge, New York.

Sachs, J., 06 March 2005. The End of Poverty. Time Magazine. http://www.time.com/time/magazine/article/0, 9171, 1034738,00.html (accessed 13.03.13.).

Sachs, J., 2005. The End of Poverty: Economic Possibilities of Our Time. The Penguin Group, New York.

Schram, S.F., 2006. Welfare Discipline: Discourse, Governance, and Globalization. Temple University Press, Philadelphia, PA.

Schram, S.F., 2000. After Welfare: The Culture of Postindustrial Society. New York University Press, New York.

Shipler, D., 2004. The Working Poor: Invisible in America. Vintage Books, New York.

U.S. Census Bureau, 2011. American Community Survey Briefs: Poverty in 2009 and 2010. www.census.gov/prod/…/acsbr10-01.pdf (accessed 25.04.14.).

Wacquant, L., 2001. Deadly symbiosis: when ghetto and prison meet and mesh. Punishment and Society 3 (1), 95–133.

Williamson, J., 1990. What Washington means by policy reform. In: Williamson, J. (Ed.), Latin American Adjustment: How Much Has Happened? Institute for International Economics, Washington, DC.

Wolch, J., Dihn, S., 2001. The new poor laws: welfare reform and the localization of help. Urban Geography 22 (5), 482–489.

Wolch, J.R., Dear, M.J., 1993. Malign Neglect: Homelessness in an American City. Jossey-Bass Publishers.

Yapa, L., 1996. What causes poverty? A postmodern view. Annals of the Association of American Geographers 86 (4), 707–728.

Relevant Websites

http://livingwage.mit.edu – Living Wage Calculator.
http://povertyinamerica.mit.edu/ – Poverty in America.
http://povertyinamerica.mit.edu/ – United Nations Millennium Development Goals.
http://www.unmillenniumproject.org/mv/ – United Nations Millennium Villages Project.
http://www.census.gov/hhes/www/poverty/ – United States Census Poverty Page.
http://www.worldbank.org/en/topic/poverty – World Bank Poverty Page.

Poverty in History

W Fischer, Freie Universität Berlin, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11907–11911, © 2001, Elsevier Ltd.

Abstract

Poverty can be found throughout human history once a definition of 'relative deprivation' within a given society is accepted. The nature and extent of poverty for much of human history is of course known. In the early modern period in Europe, estimates vary from a small proportion of the population to the vast majority. This article distinguishes between people who belonged to the 'structural poor' (physically or mentally challenged individuals, orphans, widows, and many of the elderly) and the great mass living on marginal lands or insecurely employed (small peasants and handicrafts men, day laborers, serfs, and slaves). It goes on to analyze the causes of poverty and various remedies that have been commonplace since the nineteenth century, and more especially with the growth of ideas about welfare states.

There is no universal concept of poverty, which can be applied to all cultures, societies, or times. Poverty can be defined reasonably only as 'relative deprivation,' within a given society. The size of poverty is unknown for great periods of history. In early modern Europe estimates run from 5 to 10% in wealthier cities to more than half of the population in England 1688 and even more in times of crisis. In all societies physically and mentally handicapped, orphans, widows, and many elderly people belonged to the 'structural poor.' The great mass were, however, those living on marginal lands, or on insecure jobs (small peasants and handicrafts men, day laborers, serfs, and slaves). Varying numbers lived from begging. The causes for poverty can be natural (poor soil, extreme climatic conditions or unreliable weather with periods of dearth and flooding, poor communication because of natural hindrances like mountains or deserts), demographic (large families, sharp population growth), sociopolitical (discrimination of certain groups, classes, or casts), or economic (poor productivity, structural changes, cyclical or occasional crisis). Policies to relieve poverty are known for many centuries. Churches, fraternities and communities, and finally governments tried to help and remove some of the causes for poverty, usually with little success. A decisive step was taken only with the introduction of obligatory social security schemes since the last decades of the nineteenth century.

The Concept of Poverty

There is no universal concept of poverty, which can be applied to all cultures, societies, or times. Nowadays, poverty is usually defined in relation to the medium-income level of a society, and the poverty line is drawn where a family has less than 50% of this medium income available. Such seemingly 'exact' definitions were not possible before general income statistics were generated by statistical offices in the twentieth century (with forerunners in some of the most developed countries in the nineteenth century, and even earlier in some cities).

Some researchers have sought to define an absolute level of poverty by drawing the borderline at the minimum input of calories, which a person needs to survive. But even this definition cannot be used as a universal measure since this minimum differs with age, sex, climate, the individual body, health, and the kind of work a person does. Moreover, poverty certainly begins before a person dies of hunger. Therefore, "poverty can be defined objectively and applied consistently only in terms of the concept of relative deprivation" (Townsend, 1979: p. 31). But what is relative deprivation?

The Size of Poverty

If we were to take present standards of living in advanced societies as a yardstick, most people, probably more than 90%, in all cultures and ages, would have to be defined as paupers. Only rulers, their immediate entourage, and some large landholders and merchants were free of poverty. Most people working in handicrafts or on the land whom we would term today as the middle classes lived at least at the borderline of poverty. When the harvest failed, epidemics arrived, or war devastated a region – and this occurred frequently even in Europe up to the seventeenth and eighteenth centuries – they could be driven into misery. In other regions of the world, where periods of drought or torrential rains recur regularly, it is difficult even today for the majority of people to work themselves out of poverty.

For Western Europe we have sporadic estimates of the size of poverty for single cities and smaller regions since the later Middle Ages, and by Gregory King for a whole country, England, in 1688. They vary considerably, partly because the criteria used often remained vague and depended on the prejudice and purpose of the author or institution taking the numbers. In some relatively well-to-do cities like Toledo, Lyons, Verona, Augsburg, Amsterdam, and Norwich in the sixteenth and seventeenth centuries, the poor numbered between 5 and 10% of the population in 'normal' times, rising to between 20 and 30% in times of wars, epidemics, or harvest failures. Similar figures have been found for some rural hinterlands of such cities, e.g., around Nuremberg or Zurich (Jütte, 1994: p. 53).

According to Gregory King, however, paupers constituted the majority of the population of England at the end of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62051-2

seventeenth century. While half a million families (comprising about 2.7 million persons) increased the wealth of the country, about 850 000 families (comprising more than 2.8 million persons) decreased it by depending on others for their livelihood (Mathias, 1983: p. 24). Other contemporaries give, however, different figures.

In eighteenth century, French estimates of the size of poverty vary according to the definition between one-third and half of the population. When, during the French Revolution, the definition of poverty was sharpened by excluding beggars, vagabonds, and others 'unwilling to work.' The Comité de Mendicité stated that 39% of the workingmen were not able to support a family of five persons (Hufton, 1974: p. 22).

One can distinguish at least three different measures used by officials in Western Europe in early modern times to count their paupers. The first was to count the recipients of welfare in one way or the other. This depended, of course, on the welfare policy applied at a particular location. Known figures vary from 2% for the city of Berlin in 1665 (but 7.2% in 1799) to 14.8% in Trier in 1623. But it is also known that in nearly all cities, the majority of people who applied for welfare were refuted. (In Amsterdam in 1799, 37 500 received help, while 81 080 were refuted.) The second measure were censuses of the poor. These were rare, and the few available figures vary between 9.5% in Valladolid, Spain, in 1561 and 20–25% in Huddersfield, England, in 1622. The third method is to count the 'fiscal poor': the people (mostly heads of households) exempted from taxes because of poverty. Here we find a much higher level. The lowest available figure is 4.7% for Cologne in 1582 while the highest was 77% for Verona in 1635. Most other cities record shares of the tax-exempt households between 35 and 50% (Jütte, 1994: pp. 46–56).

Taking into account that most of these figures stem from cities in the wealthier parts of Europe, it is easy to imagine the degree of poverty in other regions, particular the notoriously poorer mountainous regions, or the countryside in Eastern Europe, not to speak of the less developed continents.

Who Are the Paupers?

Certain groups of paupers can be found in all cultures and centuries, for instance, the bodily or mentally handicapped who are not able to work, and father- or motherless children (except those who live in extended families). Very often mothers with fatherless children are being reported among the poor. Finally, elderly people, particularly widows, frequently fell into poverty, at least before old age pensions were introduced in the late nineteenth and twentieth centuries. Even then they rarely received pensions high enough to support them. These people can be called 'structural poor.' They probably formed the 5–10% of paupers registered even in well-to-do cities and regions. In addition are those who have become temporarily unable to work because of sickness or injury, particularly during wars or natural disasters. One could term them 'cyclical poor.' Their numbers could swell easily to more than half of the population if events like epidemics, crop failure, or war devastated a region. At such times usually all those who lived on the margin of subsistence – until very recently the majority of the population – fell into deprivation.

All these groups have been seen since the Middle Ages by both worldly and religious authorities as 'deserving poor,' to help whom was a task for church, local or central governments, and private charity. They are to be distinguished from the 'undeserving poor': able-bodied idlers, beggars, vagrants, vagabonds, and prostitutes, who had to be educated or penalized, being expelled or – since the sixteenth century in Europe – put into workhouses or prisons (often one and the same).

For modern observers the biggest group is, however, the 'working poor': small peasants or handicrafts men, day laborers in cities as well as in the countryside, and the unemployed. Nor should slaves be forgotten. Though not all slaves in ancient Europe or seventeenth to nineteenth century America were poor – some had learned crafts or worked as privileged servants, even as artists or managers – the great majority certainly were, and slavery existed also in other cultures and in some European countries as late as the seventeenth and eighteenth centuries.

The Causes of Poverty

For analytical purposes one can distinguish between different causes of poverty but one has to be aware that in reality they often were intertwined. Jütte (1994) distinguishes between accidental, cyclical, and structural causes. Accidental causes are, for instance, sickness, which counted according to the English Poor Law Commission in 1909 for 'at least one-half of the total cost of pauperism' (p. 21). In early modern Europe plague, influenza, typhus, typhoid fever, or smallpox did not only kill large segments of the population, but also impoverished many of those who had survived. Warfare is another 'accidental' cause. Bad harvests, as well as periods of accelerated population growth and structural changes in the economy, Jütte counts as 'cyclical,' while he sees 'structural' reasons mainly in demographic patterns: persistent large families with many children and a nonsecured old age. One would have to add seasonal causes to this typology. In northern countries each winter produced unemployment and poverty.

Another analytic scheme could be to distinguish between natural, demographic, sociopolitical, and economic causes of poverty. In all preindustrial societies man was subject to natural conditions and disasters: drought or flooding, fire, unbearable heat in summer, or ice-cold storms in the winter time could stop his activities, and the natural fertility of his land limited his productivity. Richer or poorer soil, the availability of transport routes on rivers, lakes, and oceans, the availability of raw materials, like wood, copper, or iron ores determined the results of his labor. Demographic factors played an important role – as long as productivity could not be raised markedly, population growth could lead to increased poverty. Sociopolitical factors influenced the degree of, and the distribution of, wealth and poverty in all cultures, societies, and ages. Centralized or decentralized power, the strictness or looseness of social and political hierarchies, the degree and distribution of freedom, serfdom, or slavery among the population influenced also the living standards. Where geographic mobility was restricted, as in the case of slaves and serfs, people could rarely evade poverty by moving to better places. Where upward social mobility was restricted, even hard work could help little to improve the lot of

the lower strata of society. Here India with her caste system is probably the worst case, but Chinese and Eastern European peasants also fared worse than Western European ones, or Europeans in their new settlements overseas. Economic conditions are closely connected to the political and social order. Where market conditions could develop, a growing minority at least had the chance to escape poverty. But economic development did not occur harmoniously and evenly. The rise of one sector or region could cause impoverishment in another. When Dutch and English woolen products flourished in northwestern Europe in the sixteenth and seventeenth centuries, Italian cities suffered from the new competition. When Indian cotton products appeared on the European markets, the European woolen and linen producers had a difficult time, and when the British and Western European textile industries mechanized their production in the late eighteenth and nineteenth centuries Indian calico producers lost out.

Since the Industrial Revolution, numerous such displacements took place and led to regional and branch- or industry-wide pauperization. Most, but not all, were temporary and could be stopped by adjustments to new technologies or other activities. Social and geographical mobility helped to absorb such shocks. There was, however, no overall impoverishment connected with industrialization. On the contrary, rather, the growth of productivity connected with mechanization and rationalization improved even the lot of the poorer segments of the population in the long run. The social problems at the beginning were caused rather by the population growth, which accelerated noticeably from the later eighteenth century. Without industrialization and the growth in productivity in agriculture too, Malthus' law that population grows faster than productivity would not have become obsolete.

There has been a long discussion among economic historians of Britain about whether the standard of living of the working classes during the Industrial Revolution improved or deteriorated. There is agreement that it improved from the 1830s when productivity grew recognizably and regularly and the growth of industry created more better-paid jobs. The pessimistic school maintained, however, that between about 1780 and 1830 the living standard of the workers deteriorated. This discussion suffers from a lack of adequate statistics, particularly of consumer prices and regional differences. As always when structural changes occur, there were winners and losers. The winners were skilled laborers and those who could move to better-paid jobs in growing regions; the losers were the immobile, unskilled people in areas not enjoying growth in industry, trade, and services (Fischer and Bajor, 1967, and an ongoing discussion in the *Economic History Review*).

The discussion was first aired among British social reformers and Members of Parliament in the 1820s and 1830s, culminating in F. Engels' famous book on the situation of the working classes in England (1845) in which he painted an unrealistic picture of the idyllic life of the workers in the putting-out industry and contrasted it with that of the industrial workforce. It was also influenced by the last grave subsistence crisis, which occurred in parts of Western Europe during the second half of the 1840s. Ireland in particular, but also regions on the continent like Silesia, was hit by bad harvests including a nearly total loss of potatoes, the poor man's main food. Millions died or emigrated. The not-yet-mechanized linen industry could not compete with mechanized cotton industry. Pauperism seemed on the rise again, but this was a cyclical rather than a structural phenomenon. Since the middle of the nineteenth century, living standards of workers improved in most countries of Western Europe and North America without, however, erasing poverty. In many countries serious empirical work on the size and causes of poverty began only now, culminating (among others) in the 17 volumes Charles Booth published between 1889 and 1897 about *Life and Labour of the People of London*.

Welfare Policies

There is agreement among students of poverty that (modern) welfare policy began only in the nineteenth century, with the onset of social security schemes. But to care for the poor was seen as a moral obligation in all major religions and it was institutionalized at very early stages of history. Compassion was, however, only one reason for this: the other was fear, for paupers were regarded as dangerous. This twin motivation runs right through the centuries. The second motive, fear, seems to have dominated public policy in the Roman Empire, but early Christians insisted that *caritas* was essential. The Early Fathers of the Church, like St Augustine, preached that poverty was a virtue and some founders of Christian orders, as in the case of St Francis of Assisi, obliged their brethren to adhere to it. The Church and its monasteries were thus the main bodies responding to the fate of the poor during most of the Middle Ages, though not the only ones. Self-help played a role not only among families and neighborhoods, but also in the craftsmen's guilds, which cared for their sick people and sometimes for the widows and elderly masters. Brotherhoods existed not only for journeymen but sometimes also for unskilled laborers, and in the later Middle Ages in some cities such as Frankfort, Strasbourg, or Zurich even for 'the blind and the lame' (Fischer, 1982: p. 29). In other countries, such as Spain, confraternities that enrolled men and, in some cases, women from different occupations, offered mutual aid alongside spiritual welfare to their members. Christian belief convinced wealthier citizens to secure their salvation by giving alms, establishing foundations for the poor, or leaving parts of their fortunes to institutions such as hospitals.

Hospital pauperum already constituted small centers of welfare in the early Middle Ages. Often they cared for 12 persons, symbolizing the 12 apostles. In the twelfth and thirteenth centuries many such hospitals were founded, not only by bishops and monasteries but also by kings, princes, merchants, guilds, and cities. About 600 of such hospitals existed in England in the middle of the fourteenth century. Some of them specialized, for example, in the care of children or of old people, or of those with certain maladies like leprosy. Often they supplied only minimal help, not even food was necessarily provided, and their inhabitants received permission to beg. Medical care was an exception since very few physicians were available. Most of the hospitals in the smaller towns remained all-round institutions for helpless people, but a few housed several hundred people.

In the later Middle Ages cities took over some such hospitals or founded new ones. They installed poor-relief administrators,

who began investigations of individual cases of poverty and made surveys for their towns or districts. Poor relief was to be made more efficient and rational; pure almsgiving was no longer enough. The 'deserving poor' were to be separated from the 'undeserving' ones. Orderly procedures were being introduced in administering poor relief. This was particularly marked in Northern and Central Europe while in France, Spain, and Italy worldly authorities remained more reluctant to take poor relief out of the hands of the Church.

This rationalization and bureaucratization began well before the Reformation, and afterward Catholic and Protestant regimes did not necessarily follow different policies – both, for instance, tried to educate the poor, particularly the children, though with meager results. But there were differences as to the centralization or decentralization of such policies. Mainly Protestant critics centralized their institutions into a 'common chest' to exercise better control. In the German Empire the Imperial Diet of 1497 declared that each community should take measures to care for its own poor, and this principle was enacted in the Imperial Police Ordinance of 1530, which remained in force until the end of the Empire in 1806. Switzerland, the Netherlands, Britain, and the Scandinavian countries followed similar paths, while in the Mediterranean countries the many decentralized institutions remained more or less undisturbed.

While Britain is seen as the pioneer of a general poor tax in 1572, the municipality of Paris had introduced such a tax already in 1551 after the French king had given it the authority "to supervise and conduct all things required for the treatment and the relief of the poor" (Jütte, 1994: p. 119). But Britain pioneered early modern poor regulations in several respects. Already in the middle of the fourteenth century Parliament ordered that the working poor should be distinguished from beggars, and in 1388, tried to restrict the mobility of beggars. From the 1530s to around 1600 numerous statutes were enacted, of which those of 1572 and 1596 became known as the 'Old Poor Laws,' which remained in force until 1834. The judges of the peace had to select four men of substance to act as overseers of the poor. These men had to raise the poor tax, distribute help, and take care of the education of orphans. They were supposed to punish beggars and vagabonds, and could establish houses for correction. As in other countries, the result of their endeavors was very uneven. The poor taxes always fell short of the needs. It has been estimated that only 7% of the expenditure for the poor originated from taxes – by far the greater part came from private donations and charitable trusts, which were established by wealthy citizens. One rich Londoner alone set up institutions for the poor in 216 communities (Fischer, 1982: p. 43f). Such trusts also ran hospitals and schools.

In the minds of the early modern poor administrators and the wider public, education to work became more important than poor relief. Therefore, workhouses were established in the major cities and some smaller towns. The hope was that they would both accustom the poor to work and produce some profit. Usually they failed in both respects. Particularly in the Netherlands, the Scandinavian countries, and in Germany they became rather a combination of workhouse and house of correction, the famous example being the Amsterdam *tuchthuis* founded in 1595. In smaller towns such houses often tried to perform four functions: educating orphans, getting adults to work, penalizing beggars and criminals, and taking care of handicapped or sick people.

The English Poor Law of 1834 was supposed to remedy the ages-old deficiencies in welfare policies. After a Parliamentary Commission had found that the 'Old Poor Law,' by prescribing assistance to every needy person, had actually provided a disincentive to work, the new law stipulated that only such paupers who were willing to live (and work) in poorhouses should get relief. It turned out that there were not enough such institutions to house everybody who did not find work on his own or did not earn enough to survive, particularly with a larger family. The new law did not remove the ancient causes of poverty, neither could it take care of the temporary unemployment of industrial workers, but it revived the discussion about social problems in the country. A whole set of measures had to be taken to improve the lot of the poorer classes: while more employment was to be created by industry, transport business, and services. Parliament regulated child and female labor, and began to enforce safety measures at the workplace, while local communities improved the housing and sanitary conditions. In 1876, about 19% of the population were insured by voluntary Friendly Societies, relatively more than in Germany's obligatory health insurance in 1885, but the most needy could not afford the contributions. Private charity remained, therefore, important in Britain. In 1897, Britain introduced obligatory accident insurance, in 1908 old age pensions, and in 1911 a combined health and unemployment insurance.

Other European countries were faster in creating obligatory accident-, health-, and old-age insurance. But unemployment, one of the most important reasons for poverty, was being regarded as noninsurable even at the end of the nineteenth century. In Germany, unemployment insurance came into force only in 1927. Most of these insurances were quite inadequate at the beginning. They covered industrial workers only in some branches, not day-, farm-, and home-workers, thus excluding the poorest part of the labor force. Bismarck's Old Age Insurance provided means only at the age of 70 when most people had died already, and widows received only 20% of the earning of their husbands in accident insurance. But the principles of a modern welfare state were established, and the basic institutions for it were created before or shortly after World War I.

See also: Income Distribution and Inequality: Measurement Issues; Inequality: Comparative Aspects; Poverty Policy; Poverty, Culture of; Poverty, Sociology of; Poverty: Measurement and Analysis; Social Inequality in History (Stratification and Classes); Social Stratification; Wealth Distribution; Welfare Economics; Welfare Policy Comparisons; Welfare State, History of; Welfare.

Bibliography

Abel, W., 1974. Massenarmut und Hungerkrisen im vorindustriellen Europa. Versuch einer Synopsis [Mass Poverty and Subsistence Crisis in Pre-industrial Europe]. Paul Parcy, Hamburg and Berlin.

Aldcroft, D.H., Catterall, R. (Eds.), 1996. Rich Nations – Poor Nations. The Long Run Perspective. Edward Elgar, Cheltenham, UK.

Beaudoin, S.M., 2007. Poverty in World History. Routledge, London.

Clark, R.F., 2002. The War on Poverty: History, Selected Programs, and Ongoing Impact. University Press of America, Lanham, MD.

Drabble, J.H., 2000. An Economic History of Malaysia, *c.*1800–1990: The Transition to Modern Economic Growth. St. Martin's Press, New York.
Dunlevy, M., 2011. Pomp and Poverty: A History of Silk in Ireland. Yale University Press, New Haven.
Fischer, W., 1982. Armut in der Geschichte. Erscheinungsformen und Lsungsversuche der 'Sozialen Frage' in Europa seit dem Mittelalter [Poverty in History]. Vandenhoeck and Ruprecht, Göttingen, Germany.
Fischer, W. (Ed.), 1995. Lebensstandard und Wirtschaftssysteme [Living Standards and Economic Systems]. Fritz Knapp, Frankfurt, Germany.
Fischer, W., Bajor, G. (Eds.), 1967. Die Soziale Frage. Neuere Studien zur Lage der Fabrikarbeiter in den Frühphasen der Industrialisierung [The Social Question, Recent Studies about the Situation of Industrial Workers in the Early Phases of Industrialization]. Koehler, Stuttgart, Germany.
Fischer, W., van Houtte, J.A., Kellenbenz, H., Mieck, I., Vittinghoff, F. (Eds.), 1980–1993. Handbuch der Europäischen Wirtschafts- und Sozialgeschichte [Handbook of European Economic and Social History], 6 vols. Klett-Cotta, Stuttgart, Germany.
Floud, R., Johnson, P., 2003. The Cambridge Economic History of Modern Britain. Cambridge University Press, Cambridge.
Flynn, M., 1989. Sacred Charity: Confraternities and Social Welfare in Spain, 1400–1700. London.
Geremek, B., 1988. Geschichte der Armut [History of Poverty]. Artemis, Zürich, Switzerland.
Gutton, J.P., 1974. La Société et les pauvres en Europe, 16e–18e sicle [Society and the Poor in Europe, 16th–18th Centuries]. Presses universitaires, Paris.
Himmelfarb, G., 1984. The Idea of Poverty: England in the Industrial Age. London.
Hippel, W., 1995. Armut, Unterschichten, Randgruppen in der frühen Neuzeit [Poverty, Lower Strata, Marginal Groups in Early Modern Times]. Oldenbourg, Munich, Germany.
Hufton, O.W., 1974. The Poor of Eighteenth-Century France 1750–1789. Clarendon, Oxford, UK.
Illife, J., 1987. The African Poor. A History. Cambridge University Press, Cambridge, UK.
Jütte, R., 1994. Poverty and Deviance in Early Modern Europe. Cambridge University Press, UK.
Landes, D.S., 1998. The Wealth and Poverty of Nations. Why Some Are So Rich and Some So Poor. Norton, New York.
Lindenmeyr, A., 1996. Poverty is Not a Vice: Charity, Society and the State in Imperial Russia. Princeton University Press, Princeton, NJ.
Lines, T., 2008. Making Poverty: A History. Zed Books, London.
Lis, C., Soly, H., 1979. Poverty and Capitalism in Pre-industrial Europe. Harvester, Hassocks, UK.
Mathias, P., 1983. The First Industrial Nation. An Economic History of Britain 1700–1914, second ed. Methuen, London.
Mink, G., O'Connor, A., 2004. Poverty in the United States: An Encyclopedia of History, Politics, and Policy. ABC-CLIO, Santa Barbara, CA.
Mollat, M. (Ed.), 1974. Études sur l'histoire de la pauvreté (Moyen Age–XVI sicle) [Studies in The History of Poverty: Middle Ages to 16th Century], 2 vols. Sorbonne, Paris.
Orleck, A., Hazirjian, L.G., 2011. The War on Poverty: A New Grassroots History, 1964–1980. University of Georgia Press, Athens.
Pimpare, S., 2008. A People's History of Poverty in America. New Press, New York.
Pullan, B., 1994. Poverty and Charity, 1400–1700. Europe, Italy and Venice. Variorum, Aldershot, UK.
Riis, T. (Ed.), 1981–1990. Aspects of Poverty in Early Modern Europe, 3 vols. Klett-Cotta, Stuttgart, Germany.
Rimlinger, G., 1971. Welfare Policy and Industrialisation in Europe, America and Russia. Wiley, New York.
Salama, P., 1994. Pauvretés et inegalités dans le tiers monde [Poverty and Inequality in the Third World]. Paris.
Thorp, R., 1998. Progress, Poverty and Exclusion: An Economic History of Latin America in the 20th Century. Washington, DC.
Tomaskovic-Devey, D., 1988. Poverty and Social Welfare in the United States. University of Colorado Press, Boulder, CO.
Townsend, P., 1979. Poverty in the United Kingdom. A Survey of Household Resources and Standards of Living. Penguin, Harmondsworth, UK.
Williamson, J.G., 1991. Inequality, Poverty and History. Blackwell, Cambridge, UK.
Winch, D., 1996. Riches and Poverty. Cambridge University Press, Cambridge.
Woolf, S., 1986. The Poor in Western Europe in the Eighteenth and Nineteenth Centuries. Methuen, London.

Poverty Law: United States

Jeffrey Selbin, UC Berkeley School of Law, Berkeley, CA, USA
Scott L Cummings, UCLA School of Law, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.F. Handler, volume 17, pp. 11911–11917, © 2001, Elsevier Ltd.

Abstract

'Poverty law' in the United States refers to policy and lawyering strategies to contest inequality. The rise of the federal welfare state shaped the contours of poverty law in the first half of the twentieth century. This combined with the rights revolution at mid-century to mobilize legal services lawyers and courts in the War on Poverty, which was the zenith of the antipoverty movement. The welfare state's subsequent decline and federal court retrenchment channeled the antipoverty movement in new directions forged by decentralization, privatization, and globalization: moving it *downward* (from federal to local), *outward* (from state to market), and *beyond* (from domestic to global).

Introduction

The United States is a nation of extremes. By gross domestic product, it is the wealthiest country in the world. But it is also marked by persistently high rates of poverty and inequality. The U.S. lags behind peer countries in most indicators of health and social welfare with some of the highest rates of infant mortality, maternal mortality, and teen pregnancy. Life chances are shaped by economic status: poor people are at much greater risk of losing their jobs, housing, healthcare, children, and liberty as a consequence of their poverty. And poverty rates and related disparities continue to correlate closely with racial and gender inequality.

American law and lawyers have long sought to redress the damaging impacts of poverty and inequality. During the last century, government antipoverty programs have gone through periods of expansion and retrenchment, at times achieving meaningful success in reducing poverty. Activist and civic-minded lawyers have fought to establish and enforce rights to basic social and economic goods, often realizing important victories on behalf of poor people. Courts have mediated many of these distributive struggles, sometimes at the vanguard of change and sometimes forestalling or reversing it.

In the U.S., 'poverty law' refers to these interrelated policy and lawyering strategies, which are deployed and contested across a range of lawmaking spaces. The rise of the federal welfare state shaped the contours and content of poverty law in the first half of the twentieth century. This combined with the rights revolution at mid-century to mobilize legal services lawyers and courts in the War on Poverty, which was the zenith of the antipoverty movement. The welfare state's subsequent decline and federal court retrenchment channeled the antipoverty movement in new directions forged by decentralization, privatization, and globalization: moving it *downward* (from federal to local), *outward* (from state to market), and *beyond* (from domestic to global).

The Rise of the Federal Welfare State

The nature of social obligation – what is owed to and expected from fellow community members – is at the core of antipoverty policy. Social welfare programs have often made moral distinctions between the worthy and unworthy poor (Handler and Hasenfeld, 2007; Zatz, 2012). Such distinctions hinge on divergent understandings of the causes of poverty, and reflect prevailing social norms, such as the value of promoting certain types of families. Moral views, in turn, are influenced by economic concerns, which focus on the degree to which aid reduces economic incentives to work and creates dependency, and who bears the cost of any transfer scheme.

These considerations have informed the development of two basic types of aid in the United States: social insurance programs, which are contributory (and what some consider 'earned'); and public assistance or welfare programs, which are non-contributory (and what some consider 'unearned'). U.S. antipoverty policy can be traced along a rough historical arc of increasing federal support for such programs culminating in the War on Poverty. Although political pressure has been placed on both types of programs, the redistributive aspect of welfare has made it more vulnerable, causing the federal role to be both curtailed through burden-shifting to states and localities and rechanneled through market-based mechanisms.

The first national welfare program – and arguably the inception of the modern U.S. welfare state – came in the form of pensions paid out to northern veterans of the Civil War starting in the 1860s. For several decades, Civil War pensions represented roughly one-fourth of the federal budget, and by the turn of the century more than one quarter of all white men age 65 and older were receiving benefits (most southern states adopted their own pension systems for Confederate veterans by the 1890s) (Skocpol, 1995).

During the Progressive Era in the U.S. (1890s–1920s), states established widows pensions directed primarily at poor single mothers in 'suitable homes' (Trattner, 1998). As a result of the Great Depression, many of these state programs collapsed, ushering in a new level of federal investment in aid to the poor. President Franklin Roosevelt's New Deal, enacted during the depths of the Depression, included the Social Security Act of 1935, which: (1) created a state-based Unemployment Insurance program; (2) directed grants to the states for the blind, elderly poor and dependent children (the latter of which we now think of as 'welfare'); and (3) established

contributory insurance benefits for retirees in certain industries (what we now think of as 'Social Security').

World War II led to further expansion of the federal safety net through the Servicemen's Readjustment Act of 1944, popularly known as the G.I. Bill. The G.I. Bill provided extensive assistance to veterans and their families, including extended unemployment compensation, low-interest mortgages, generous business and farm loans, and free educational expenses for college or vocational training. While not expressly targeting the poor, the G.I. Bill kept or lifted many families out of poverty and is often credited with helping to create the post-war middle class (Skocpol, 1995).

But the benefits and the economic gains in the 1940s and 1950s were not shared by all. The Social Security Act of 1935 maintained deference to the states on almost all matters beyond contributory social insurance, leaving behind large numbers of poor people, including non-white mothers, elderly African-Americans, and younger workers of all races who were abled-bodied but could not find work. To appease southern Democrats in Congress, Social Security did not cover mostly black agricultural and domestic workers (Skocpol, 1995). New Deal work relief programs like the Works Progress Administration and the Civilian Conservation Corps employed millions of people, but were short-lived.

The next great burst of antipoverty legislation was a direct outgrowth of internal migration, post-war prosperity, and an emergent 'rights consciousness.' As the South continued to enforce racial segregation and a sharecropper economy, African Americans embarked on a 'Great Migration' to the North and West, drawn to industrial jobs in major urban areas. In two waves beginning roughly in 1910 and 1940, 6 million African Americans left the Jim Crow South to seek a better life. While African Americans made important gains on the employment front, continued discrimination in the private and public sectors helped to fuel both the civil rights and the welfare rights movements.

These movements stimulated – and in turn were supported by – greater public focus on those left behind by the New Deal, powerfully captured by Michael Harrington's 1962 book on the 'other America' (Harrington, 1962). Presidents John Kennedy and Lyndon Johnson responded to this persistent inequality through the New Frontier and Great Society programs. To increase independence and reduce welfare rolls, President Kennedy's 1962 Public Welfare Amendments to the Social Security Act expanded the federal government's contribution to the states for job training, job placement, and other support services to recipients.

In January 1964, President Johnson elevated the federal role further by declaring an 'unconditional war on poverty' and calling upon Congress to enact a multi-prong strategy to confront a 'domestic enemy which threatens the strength of our Nation and the welfare of our people.' As part of the War on Poverty, Congress passed the Food Stamp Act, which converted an agricultural surplus subsidy into a program to assist millions of low-income families. The 1965 amendments to the Social Security Act created Medicare, which provided the closest thing to national health insurance for millions of older Americans; and Medicaid, which provided grants to the states for a subset of the medically-needy poor (mostly those on welfare).

A centerpiece of the War on Poverty was the Economic Opportunity Act of 1964, which created an Office of Economic Opportunity (OEO) to administer a range of innovative and at times controversial social programs. The Community Action Program (CAP) was designed to empower the poor through their 'maximum feasible participation' in local antipoverty efforts. Although successful in building a number of powerful community action agencies in big cities, CAP was criticized for both doing too little and doing too much: some commentators lamented the lack of authentic community input, while big city mayors opposed federal government funding of aggressive local mobilization.

The Economic Opportunity Act included funding for community development corporations engaged in neighborhood-based revitalization efforts through affordable housing and job creation initiatives – thus stimulating the nascent community economic development movement. The Act also established other enduring antipoverty programs, such as college programs for urban youth (Upward Bound), a domestic peace corps (VISTA), and preschool funding for disadvantaged children (Head Start). Praised and decried as expansions of the welfare state, these legislative initiatives were designed to bring low-income urban and rural communities into the economic mainstream through increased opportunity.

Legal Services and the Rights Revolution

The institutional field of poverty lawyering in the United States evolved from the early charity days of legal aid societies, to its federally-funded zenith during the War on Poverty, to a more decentralized and market-dependent system today. The OEO Legal Services Program and the early years of the successor Legal Services Corporation (LSC) marked the high water point of federal support for the legal aid movement and transformed the delivery of legal assistance in low-income communities. Poverty lawyers in this period turned to the federal courts to instantiate and extend core principles of the modern welfare state, first with success, then with increasing failure, marking the end of the rights revolution – and forcing poverty lawyers to pursue alterative advocacy strategies.

Like the welfare state, legal aid traces its roots to the aftermath of the Civil War, when the federally sponsored Freedman's Bureau was created to provide free legal services to newly emancipated slaves in the Reconstruction South. The first legal aid program, New York's Legal Aid Society, was started in 1876 to 'render legal aid and assistance, gratuitously, to those of German birth, who may appear worthy thereof, but who from poverty are unable to procure it.' During the early twentieth century, charitable legal aid organizations spread in poor neighborhoods, where they relied on staff attorneys to provide free services to specific categories of poor clients, and were subsidized, usually very meagerly, by philanthropic contributions. The organized bar was not involved in establishing a legal aid society until 1909 (Cummings, 2004).

Stung by the publication of Reginald Heber Smith's *Justice and the Poor* (1919), which denounced the glaring inequality in legal services, the bar took on a greater role in funding legal aid, stimulating its expansion over the next 40 years. While bar involvement supported growth, it also placed operational

constraints on legal aid work. Controversial clients were generally avoided, cases that could generate fees were rejected, and client income eligibility was maintained at levels acceptable to private attorneys competing for lower-income clients. These limitations, combined with those imposed by charitable and local business funders, led legal aid lawyers to downplay reform-oriented advocacy and concentrate on resolving individual disputes.

Change came in the form of the OEO Legal Services Program, which institutionalized a more reform-oriented approach within a newly created federal agency committed to attacking poverty. The Economic Opportunity Act did not mention legal aid. However, the Ford Foundation funded many programs – including legal services – that informed antipoverty strategies. In *The War on Poverty: A Civilian Perspective* (Cahn and Cahn, 1964), Edgar and Jean Cahn argued that neighborhood legal services should be a part of the antipoverty effort. With others, they persuaded OEO's first director, Sargent Shriver, to fund legal services for the poor.

The location of legal services within the federal government allowed it to depart from the legal aid model. Propelled forward by the power of the federal purse, the new program was freed of constraints that forced bar-funded legal aid societies to adopt an individualized equal access approach. In the program's first 2 years, grants were made to 300 legal services organizations at a level that constituted an eightfold increase over the 1965 investment in legal aid. By 1981, federal funding for legal services reached its peak at $321 million (Johnson, 1974).

Not all legal services offices were activist. Because the government administered legal services funding on a block-grant basis, many legal aid societies that received grants continued to provide individual services and avoid reform activities (Abel, 1985). However, there was an undeniable shift in tone as well as strategy. As the legal services program's first director, Clinton Bamberger, implored bar presidents in 1966:

> We cannot be content with the creation of systems of rendering free legal assistance to all the people who need but cannot afford a lawyer's advice. This program must contribute to the success of the War on Poverty. Our responsibility is to marshal the forces of law and the strength of lawyers to combat the causes … of poverty … [,] remodel the system which generates the cycle of poverty[,] and design new social, legal and political tools and vehicles to move poor people from deprivation, depression and despair to opportunity, hope and ambition. (Johnson, 1974)

Spurred by this commitment – and buoyed by the new talent attracted by the program's prestige and resources – legal services lawyers did in fact employ more aggressive legal tactics: taking a higher percentage of their cases to court than their legal aid counterparts, making over 1000 appeals every year, and bringing 219 cases to the U.S. Supreme Court in the first 5 years of the program's existence (Lawrence, 1990).

Poverty lawyers, inspired by the impact litigation model pioneered by the NAACP Legal Defense and Education Fund in *Brown v. Board of Education* (1954), thus became active participants in shaping poverty law. Courts had, of course, played crucially important roles in the development of poverty law before the civil rights period. In *Helvering v. Davis* (1937) and *Steward Machine Company v. Davis* (1937), for example, the Supreme Court upheld key elements of the Social Security Act of 1935. Yet the emergent rights revolution of the 1960s – powered by the ideal of progressive social change through law centered on the Warren Court – created new energy around poverty law litigation.

Part of this energy was directed at government and business practices affecting the poor. Lawyers at California Rural Legal Assistance (CRLA), for example, successfully sued the state to restore millions of dollars in Medi-Cal benefits, expanded food stamps and school lunch programs, forced the adoption of a state minimum wage for farm workers, and curtailed the importation of Mexican braceros. Legal services lawyers prevailed in *Fuentes v. Shevin* (1972) when the U.S. Supreme Court struck down on due process grounds state statutes that allowed private parties summarily to repossess goods. Poverty lawyers also sought change through regulatory advocacy on welfare, consumer, housing, and other issues within the now-powerful administrative state.

The most high-profile litigation campaign of this era ambitiously aimed to constitutionalize welfare rights (Davis, 1993). Building upon the intellectual foundation of academic work like Charles Reich's *The New Property* (1964) and led by front-line groups and back-up centers, like Ed Sparer's Center on Social Welfare Policy and Law, lawyers won a string of victories that established key constitutional protections for welfare recipients. In *King v. Smith* (1968), a unanimous Court struck down Alabama's 'man in the house' provisions. In *Shapiro v. Thompson* (1969), the Court invalidated welfare residency requirements. And in *Goldberg v. Kelly* (1970), the Court established the right to an evidentiary hearing before reduction or termination of public benefits, which some lawyers hoped would lay the foundation for a constitutionally guaranteed minimum income.

Taken together, these cases substantially limited states' discretion to legislate morality (*King*), restrict eligibility (*Shapiro*), and terminate benefits (*Goldberg*) at precisely the moment when welfare rolls were growing and becoming disproportionately African-American. As a result of this success, poverty lawyers were criticized for expanding the welfare state and contributing to social decline. Like their civil rights counterparts, poverty lawyers were also criticized for assuming too much power, undercutting the movement's organizing base and moving too far ahead of public opinion in ways that invited backlash (Alfieri, 1991).

Down and Out: Poverty Law after the Welfare State

Just as the Economic Opportunity Act marked the legislative apex of antipoverty policy, so too did *Goldberg* constitute the height of the law reform project. Political resistance to and backlash against the legislative and legal victories of the antipoverty movement – and New Left movements more broadly – began to erode support for the robust exercise of federal power to redistribute resources, and created traction for the ideology of small government and limited judicial intervention on behalf of minority rights. These ideas were fueled by, and reinforced, an ascendant conservative movement that reshaped federal policy making and the judiciary.

Beginning with Richard Nixon's 1968 presidential election and the 1969 appointment of Justice Warren Burger, and culminating in the presidency of Ronald Reagan, all three branches of the federal government retreated from the antipoverty project. Federal courts in general and the Supreme Court in particular began withdrawing constitutional support for welfare and other antipoverty policies. The executive branch became less receptive to advocacy by poverty lawyers, who themselves came under attack. And Congress began unraveling the federal safety net.

These shifts did not spell the end of antipoverty law and advocacy, but instead initiated a political realignment that had parallel effects on both law and lawyers, moving policy making, implementation, and advocacy *downward* (to the state and local levels) and *outward* (from the government to the market).

Poverty Law

A crucial lesson of the antipoverty movement was that a progressive law reform agenda asserting the rights of the poor against government bureaucracies and private businesses was vulnerable to judicial and political backlash. Federal court welfare rights retrenchment began with the election of Richard Nixon and his appointment of four Supreme Court justices from 1969 to 1972. Deeming welfare an economic and social issue and not a fundamental right, the Court in *Dandridge v. Williams* (1970) upheld a Maryland law that capped public assistance grants regardless of family size. *Dandridge*, decided only 2 weeks after *Goldberg v. Kelly*, marked the end of the welfare rights movement's litigation campaign to establish a constitutionally protected right to a minimum income (White, 1999).

The Court issued other opinions outside of the welfare rights domain that nonetheless had a significant impact on inequality. Notably, in *San Antonio Independent School District v. Rodriguez* (1973), the Court upheld a Texas school financing regime that resulted in lower per-pupil funding in poorer school districts, holding that education was not a fundamental right and wealth was not a suspect classification for equal protection purposes. This narrow definition of rights and deference to the legislature in the economic and social arena came to define poverty law jurisprudence – and marked the federal courts as an inhospitable venue for poverty lawyers (Brodie et al., 2014).

Nixon also sought to remake the federal welfare state. His 1969 Family Assistance Plan (FAP) failed to garner liberal support – despite its relatively generous assistance by today's standards – primarily because it would have introduced work requirements for most welfare recipients. Although it would not come to full fruition until the 1990s, the effort to tie public assistance to work resulted in the 1975 passage of the Earned Income Tax Credit (EITC), which provided tax credits to working people with children. The administration also succeeded in passing the Community Development Block Grant program in 1974, which decentralized urban antipoverty money by distributing it in block grants to states and localities.

The failure of the FAP and the creation of the EITC actually combined to increase overall welfare spending during the Nixon era, which fed the growing political reaction that found its champion in Ronald Reagan. In the mid-1970s, then-presidential candidate Reagan introduced the 'welfare queen' into the public lexicon. Though his example was both fictitious and misleading – the caricature never represented an actual person, and whites were still the majority of welfare recipients – this barely coded racial critique resonated strongly with a white middle class fearful of losing further economic ground during the 'stagflation' of the 1970s.

Such attacks helped to fuel a grassroots anti-tax movement opposed to 'big government' generally and social welfare spending in particular. Reagan himself rode this anti-tax sentiment to the White House, where he succeeded in reducing welfare state spending. For example, his administration slashed public housing programs by 80%, and promoted market-oriented alternatives, including the Housing Choice Voucher Program (1983), which shifted federal funding from housing production to vouchers given to recipients to use in the private rental market, and the Low-Income Housing Tax Credit Program (1986), which subsidized private companies that invested in affordable housing development.

This populist sentiment was given intellectual heft in the 1980s by critics who challenged existing policies on moral and philosophical grounds. In *Losing Ground* (Murray, 1984), Charles Murray led the moral charge against so-called 'transfers' (not only from rich to poor, but also between low-income people). In *Beyond Entitlement* (Mead, 1986), Lawrence Mead argued that public assistance should come with reciprocal obligations (what he called the 'social obligations of citizenship'). Responses to Murray and Mead that took seriously the role of structural barriers in poor communities, like William Julius Wilson's *The Truly Disadvantaged* (1987), did not hold sway with the voting public and elected officials.

When President Bill Clinton signed into law the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA, also known as 'welfare reform'), it was the culmination of several decades of conservative critique of federal control and government spending. Welfare reform ended the 60-year federal entitlement to benefits for eligible families established by the Social Security Act. In its place, PRWORA created Temporary Assistance to Needy Families (TANF) through a block grant program to the states. TANF benefits were lifetime-limited to no more than 60 months and came with stringent work requirements. This federal devolution and emphasis on market-based incentives has become a signature feature of contemporary antipoverty policy.

As the 1990s welfare debate receded, successive administrations have tried to address spiraling healthcare costs coupled with the tens of millions of uninsured Americans. In 2010, President Barak Obama signed into law the Patient Protection and Affordable Care Act (PPACA), which aims to reduce the number of uninsured Americans and slow the growth of system-wide costs. The PPACA is the most important healthcare reform since the establishment of Medicare and Medicaid in 1965, and extends the logic of decentralization and privatization. In contrast to single-payer or government-run systems in most peer countries, the PPACA will rely on market mechanisms implemented primarily at the state level.

The PPACA also reaffirmed the importance of courts in the poverty law movement, with advocates now defending the constitutionality of a law under attack by opponents. In

a bitterly divided decision in *National Federation of Independent Businesses v. Sebelius* (2012), the U.S. Supreme Court affirmed the central provision of the PPACA. The Court held that Congress had the power under its taxing authority to require individuals to obtain health insurance or pay a fine, thus ensuring implementation of one of the most expansive antipoverty programs since the War on Poverty. At the same time, the Court limited the federal government's authority to expand Medicaid, perhaps auguring a new phase in the political contest over the scope and meaning of poverty law.

Healthcare reform reflects another way in which poverty itself – and hence the poverty law movement – has changed over time. By expanding Medicaid to more people and otherwise subsidizing the acquisition of private insurance by those unable to afford it, the PPACA spotlighted the unmet healthcare needs of the 'working poor.' Since the mid-twentieth century, the growth of those living below the poverty line despite holding jobs has been part of a larger story of economic restructuring related to declining unionization, deindustrialization, foreign outsourcing, free trade, and immigration.

Poverty law's traditional concern with safety net programs has thus been extended to policies and advocacy strategies that redress economic insecurity and living conditions produced by unstable and low paying work (Munger, 2002). The expansion of the EITC into one of the nation's largest antipoverty programs underscored both the decline of the traditional welfare state and the expansion of low-wage work. As the scope of welfare benefits shrunk and the number of working poor grew, the movement to advance antipoverty policy – through both policy reform and legal advocacy – has become deeply linked to a broader range of issues relating to economic inequality, including community development, labor and employment law enforcement, immigrant rights, prisoner reentry, consumer lending, and environmental justice.

Poverty Lawyers

Federal support for legal services has undergone a similar transformation. The success of legal services in pressing a redistributive agenda to expand entitlements and reform business practices became a liability as the political winds shifted to the right. The LSC Act of 1974 created the successor to OEO Legal Services, but placed restrictions on program activities and defunded national law reform centers. In exchange for continued government support, legal services lawyers were required to eschew reform in favor of 'access to justice' and individual client service.

The pressure on legal services increased with the presidential election of Ronald Reagan, who was a vocal critic of LSC from his days battling CRLA as the state's governor. As president, Reagan moved aggressively against legal services, proposing to eliminate LSC in his first budget (and every budget thereafter). Congress restored 75% of LSC's funding, but Reagan appointed a hostile board. He also succeeded in shifting a portion of LSC funds away from field offices to support 'private attorney involvement,' which was deemed a less political effort to promote market-based delivery through *pro bono* and Judicare systems (Cummings, 2004).

As with welfare reform, the attacks on legal services reached their peak in 1996, when Congress slashed LSC's budget by another third and imposed crippling restrictions. Congress banned LSC-funded offices from redistricting challenges, lobbying, class action lawsuits, representing most noncitizens, political advocacy, collecting attorney's fees, abortion litigation, prisoner representation, welfare reform activities, and defending public housing tenants evicted for drugs. Most drastically, this legislation prohibited LSC offices from using non-LSC funds to engage in any of the banned activities. Courts have generally upheld the restrictions, though the ban on welfare litigation was struck down, and in 2010 Congress lifted the ban on attorney's fees.

In the face of these challenges, legal services to the poor has not so much declined as changed. Like antipoverty policy more generally, it is now characterized by resource decentralization and greater emphasis on private sector delivery. While LSC funding has decreased dramatically in inflation-adjusted terms since 1981, system-wide funding for legal aid has actually grown because of increases from other sources, particularly state, local, and private funds. Support for legal services was approximately $1.5 billion in 2010 – or almost twice the real dollar value of the high water mark of LSC funding – with at least two-thirds derived from non-federal sources like private giving and state-managed Interest on Lawyers' Trust Accounts (IOLTA), which distributes to legal services programs bank interest accumulated in private attorneys' client trust accounts (Houseman, 2009).

In addition to this shift in funding, legal services has undergone changes in how it is distributed, with a greater role for market-based delivery. A key development is the increasing role of private sector lawyers providing *pro bono* services, which has been driven by several related trends (Cummings, 2004). First, LSC's private attorney involvement mandate – designed to weaken field offices – shifted resources to *pro bono* attorneys. As a result, the number of *pro bono* referral programs has grown from about 90 in 1980 to approximately 900 such programs today, accounting for between one-quarter and one-third of free civil legal services (Sandefur, 2007).

Second, as LSC funding decreased and restrictions were imposed, the bar promoted *pro bono* as a crucial bulwark against the decline of available federal legal aid. In an important early example, the 1983 Model Rules of Professional Conduct for the first time asserted that a 'substantial majority' of *pro bono* hours were to be targeted to 'persons of limited means' or organizations that advocate on their behalf. Other factors have supported the institutionalization of *pro bono*, including the growth in size and number of large law firms, the advent of law firm *pro bono* rankings by *The American Lawyer*, and evolving professional norms of voluntarism (Cummings, 2004; Granfield and Mather, 2009). As a result, *pro bono* service has become a major resource for poor and underserved clients, with large law firm *pro bono* programs leading the way (Boutcher, 2009).

The turn to nongovernmental support and private *pro bono* has involved trade-offs. Market-based funding and delivery mechanisms are not countercyclical. Thus, in a period of economic crisis like the 2008 Great Recession, legal services funding may decrease as a result of lower interest rates on IOLTA funds, shrunken foundation endowments, and reduced individual and corporate giving. Evidence from the Great Recession suggests that *pro bono* activity may also decrease in

times of economic crisis. When government is not able or willing to increase spending to counteract the decline of private supply, the reliance on market mechanisms may mean that legal services to the poor will drop at the very moment that it is needed most.

In addition, coordinating volunteer services across nonprofit and for-profit firms generates significant transaction costs as front-line legal aid providers have to conduct client intake and then spend time finding volunteers willing to take the cases. Despite substantial progress, such as the widespread hiring of *pro bono* coordinators, there are still organizational impediments to *pro bono* activity in large firms. In her survey of *pro bono* activity, Deborah Rhode reported that only "a quarter of surveyed lawyers' workplaces (25%) fully counted pro bono work toward billable hours" (Rhode, 2005).

Finally, the pressure to prioritize billable hours, combined with the lack of strong supervision by partners, may contribute to less than exemplary *pro bono* representation. In *Maples v. Thomas* (2012), the U.S. Supreme Court excused a death row inmate's habeas filing error when his Sullivan & Cromwell *pro bono* attorneys left the firm and failed to inform him – thus causing him to miss the filing deadline. Although such extreme cases are atypical, quality concerns are not. As Rhode reports: "About three fifths of [surveyed public interest] organizations experienced some quality concerns [with pro bono lawyers]; fourteen percent … reported extensive problems, 33% reported moderate problems, and 8% reported limited problems" (Rhode, 2008).

Private sector delivery also raises questions about which clients get which types of services. Within the *pro bono* system, law firms are reluctant to work on cases that threaten corporate client interests; thus, they take fewer cases in the areas of employment and labor law, environmental justice, and consumer law. Overall, the *pro bono* system is more oriented toward individual dispute resolution and less toward systemic reform. This preference reinforces rather than supplements the federal legal services system, where the vast majority of cases pertain to private disputes or individual claims for government benefits.

In the face of continued scarce resources (LSC, 2009), poverty lawyers are pushing in several directions. On one end, advocates are moving away from the goal of full-scale legal representation for the poor and toward limited scope representation (Trubek, 2001). Proponents of 'unbundling' legal services suggest that legal problems can be broken down into smaller parts (writing a demand letter, filing a grievance, attending a mediation session), and that clients and nonlawyers can be trained to undertake some of these tasks through telephone hotlines and pro se assistance centers without the need for lawyers. Critics argue that unbundling, while useful in very basic situations, risks leaving poor people with a second-class version of legal assistance that can harm their interests.

On the other end, the 'Civil Gideon' movement – drawing its name from the 1963 Supreme Court case mandating state provision of counsel for indigent defendants in criminal cases – has pressed for a judicial or legislative right to counsel in civil matters where the stakes are high (Abel, 2009; Pastore, 2009). In 2009, California passed a law providing nearly $10 million to support pilot projects to provide counsel in civil cases involving 'critical issues affecting basic human needs,' such as housing, child custody, and domestic violence. In 2010, the American Bar Association called for the provision of "counsel as a matter of right at public expense to low income persons in… adversarial proceedings where basic human needs are at stake, such as those involving shelter, sustenance, safety, health or child custody…"

The U.S. Supreme Court, however, is not prepared to extend *Gideon* to the civil context, even when liberty is at stake. In *Turner v. Rogers* (2011), a unanimous Court held that an indigent defendant facing jail time for failure to pay child support was not automatically entitled to state-provided counsel. Though five justices held that Turner did not receive due process in the instant case, the Court held that sufficient procedural safeguards could be met with less than full lawyer representation. The Court's refusal to establish a right to counsel in a case threatening a significant deprivation of liberty disappointed Civil Gideon proponents and reinforced the notion that federal courts could not be counted on to advance the interests of poor people. Partly for this reason, poverty lawyers have turned to state courts to litigate Civil Gideon and other important antipoverty issues.

Beyond Borders: Poverty Law in a Global Age

As federal retrenchment has channeled poverty law activity downward to the local, and outward to the market, it has also interacted with increasing globalization in ways that have linked U.S. policy and practice to change beyond American borders (Cummings, 2008). The impact of globalization can be understood from two perspectives. Internally, global economic integration has produced migratory patterns that have placed issues facing noncitizens at the forefront of the U.S. antipoverty agenda. Externally, U.S. poverty law activism has increasingly interacted with the development of poverty law around the world.

Within the United States, globalization has reshaped the composition of the antipoverty constituency and informed the tactical approach of poverty lawyers. The increase of unauthorized immigration in the 1980s and 1990s, powered in part by economic disruptions caused by free trade, confronted poverty lawyers with clients defined by the absence of legal status. As LSC prohibited the provision of legal services to most undocumented immigrants, immigrant rights legal groups emerged to redress the full range of issues caused by legal marginalization and its attendant economic insecurity. Particularly as federal antipoverty policy sought to exclude many immigrants – most notably in the denial of some benefits to legal immigrants in the 1996 welfare reform bill – immigrant rights lawyers focused on reconstituting an immigrant safety net at the state level and advocated for comprehensive immigration reform at the federal level.

As poverty law's client base has become more globalized, so too has its approach. U.S. poverty lawyers facing the limits of domestic advocacy through law have pursued strategic linkages to the international arena. A number of high profile antipoverty campaigns have drawn upon international human rights law and proceeded through international venues. For example, the Poor People's Economic Human Rights Campaign challenged the 1996 welfare reform bill before the Inter-American

Commission on Human Rights as a violation of the Universal Declaration of Human Rights. That it did not aim to win the legal case, but rather to mobilize international pressure on U.S. lawmakers and courts to change domestic policy, both reflected the strategic innovation of the poverty law movement and underscored how far conservative backlash had forced it to travel to advance its claims.

Outside the U.S., countries have made significant investments in poverty law (Cappelletti and Garth, 1978). While the U.S. trails Western European and Commonwealth nations in per capita governmental spending on legal aid (Johnson, 2000), countries in the Global South have incorporated lessons from the U.S. experience – sometimes following the U.S. lead and other times charting a distinctive path. India's legal aid system is based on a constitutional guarantee to equal justice and implemented by legal services authorities (Galanter and Krishnan, 2004). Brazil recognized the right to legal aid in its 1988 constitution and developed a public system designed to provide free assistance in civil and criminal cases (Vieira, 2008). In 2003, China codified a state-sponsored legal aid system that uses government-funded and private lawyers to represent poor people in a range of individual disputes (Liu, 2008). Many other countries have developed state-sponsored and nongovernmental civil legal services sectors.

In addition, U.S. and international funders have directed civil justice resources to support the rule of law and complement international development. Beginning in the 1990s, Northern donors turned to poverty law as a tool to promote governmental legitimacy and increase access to justice. Major foundations like the Open Society Institute and the Ford Foundation – and governmental and international agencies, such as the U.S. Agency for International Development and the United Nations – made significant investments in supporting access to justice (legal aid and *pro bono*) in diverse countries across the globe, including Bangladesh, Bulgaria, Cambodia, China, Hungary, Liberia, Mongolia, Poland, and Russia.

Clinical legal education has also been an important vector of global exchange, with major funders promoting the development of clinics with an antipoverty agenda (Bloch, 2010). The Global Alliance for Justice Education was founded in the late 1990s to facilitate the network of clinical and practice-oriented law school professors from around the world interested in promoting social justice pedagogy. Partly as a result of such exchanges, law schools in the Global South have themselves been important sites for the incubation and practice of poverty law.

U.S. poverty lawyers have played important roles as advisors in developing international programs, which have been greeted with mixed responses by local populations pleased to receive resources but resistant to the imposition by powerful Northern actors of external models untailored to local needs. Thus, as the poverty law movement becomes more globalized, it confronts anew its central dilemma: how to mobilize elite support and leverage public and private resources toward the end of not only transferring resources to poor people, but ultimately building their political power.

See also: Civil Rights; Equality and Inequality: Legal Aspects; Justice, Access To; Law as an Instrument of Social Change; Public Interest Law: The United States and Beyond.

Bibliography

Abel, Laura K., 2009. Keeping families together, saving money, and other motivations behind new civil right to counsel laws. Loyola of Los Angeles Law Review 42, 1087.

Abel, Richard L., 1985. Law without politics: legal aid under advanced capitalism. UCLA Law Review 32, 474.

Alfieri, Anthony V., 1991. Reconstructive poverty law practice: learning lessons of client narrative. Yale Law Journal 100, 2107.

Bloch, Frank (Ed.), 2010. The Global Clinical Movement: Educating Lawyers for Social Justice.

Boutcher, Steven A., 2009. The institutionalization of pro bono in large law firms: trends and variation across the AmLaw 200. In: Granfield, Robert, Mather, Lynn (Eds.), Private Lawyers and the Public Interest: The Evolving Role of Pro Bono in the Legal Profession, p. 135.

Brodie, Juliet, Pastore, Clare, Rosser, Ezra, Selbin, Jeffrey, 2014. Poverty Law, Policy & Practice.

Cahn, Edgar S., Cahn, Jean C., 1964. The war on poverty: a civilian perspective. Yale Law Journal 73, 1317.

Cappelletti, Mauro, Garth, Bryant, 1978. Access to justice: the worldwide movement to make rights effective: a general report, in access to justice. In: Cappelletti, Mauro, Garth, Bryant (Eds.), A World Survey, Book I 3, vol. 1, pp. 54–120.

Cummings, Scott L., 2004. The politics of pro bono. UCLA Law Review 52 (1).

Cummings, Scott L., 2008. The internationalization of public interest law. Duke Law Journal 57, 891.

Davis, Martha, 1993. Brutal Need: Lawyers and the Welfare Rights Movement, 1960–1973.

Galanter, Marc, Krishnan, Jayanth K., 2004. "Bread for the poor": access to justice and the rights of the needy in India. Hastings Women's Law Journal 55, 789.

Granfield, Robert, Mather, Lynn, 2009. Pro bono, the public good, and the legal profession. In: Granfield, Robert, Mather, Lynn (Eds.), Private Lawyers and the Public Interest: The Evolving Role of Pro Bono in the Legal Profession, p. 1.

Handler, Joel F., Hasenfeld, Yeheskel, 2007. Blame Welfare, Ignore Poverty and Inequality.

Harrington, Michael, 1962. The Other America: Poverty in the United States.

Houseman, Alan W., 2009. Civil Legal Aid in the United States: An Update for 2009. http://www.clasp.org/admin/site/publications/files/CIVIL-LEGAL-AID-IN-THE-UNITED-STATES-2.pdf (last visited March 28, 2013).

Johnson Jr., Earl, 2000. Equal access to justice: comparing access to justice in the United States and other industrial democracies. Fordham International Law Journal 24, S94.

Johnson Jr., Earl, 1974. Justice and Reform: The Formative Years of the OEO Legal Services Program.

Lawrence, Susan E., 1990. The Poor in Court: The Legal Services Program and Supreme Court Decision Making.

Documenting the Justice Gap in America: The Current Unmet Civil Legal Needs of Low-Income Americans, 2009. Legal Services Corporation.

Liu, Titi M., 2008. Transmission of public interest law: a chinese case study. UCLA Journal of International Law & Foreign Affairs 13, 263.

Mead, Lawrence, 1986. Beyond Entitlement, the Social Obligations of Citizenship.

Munger, Frank (Ed.), 2002. Laboring Below the Line: The New Ethnography of Poverty, Low-Wage Work, and Survival in the Global Economy.

Murray, Charles, 1984. Losing Ground: American Social Policy, 1950–1980.

Pastore, Clare, 2009. A civil right to counsel: closer to reality? Loyola of Los Angeles Law Review 42, 1065.

Reich, Charles, 1964. The new property. Yale Law Journal 73, 733.

Rhode, Deborah L., 2005. Pro Bono in Principle and Practice. Public Service and the Professions.

Rhode, Deborah L., 2008. Public interest law: the movement at midlife. Stanford Law Review 60, 2027.

Sandefur, Rebecca L., 2007. Lawyers' pro bono service and American-style civil legal assistance. Law & Society Review 41, 79.

Skocpol, Theda, 1995. Social Policy in the United States: Future Possibilities in Historical Perspective.

Smith, Reginald H., 1919. Justice and the Poor.

Trattner, William, 1998. From Poor Law to Welfare State: A History of Social Welfare in America.

Trubek, Louise G., 2001. Old wine in new bottles: public interest lawyering in an era of privatization. Fordham Urban Law Journal 28, 1739.

Vieira, Oscar V., 2008. Public interest law: a Brazilian perspective. UCLA Journal of International Law & Foreign Affairs 13, 219.

Wilson, William J., 1987. The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy.

White, Lucie, 1999. Specially tailored legal service for low-income persons: pragmatism or capitulation? Fordham Law Review 67, 2573.
Zatz, Noah, 2012. Poverty unmodified? Critical reflections on the deserving/undeserving distinction. UCLA Law Review 59, 550.

Relevant Websites

http://www.gaje.org/ – Global Alliance for Justice Education.
http://www.lsc.gov/ – Legal Services Corporation.
http://ncforaj.org/ – National Center for Access to Justice.
http://maximinlaw.wordpress.com/ – Poverty Law for Professors and Legal Academics.
http://www.probonoinst.org/ – Pro Bono Institute.
http://povertylaw.org/ – Sargent Shriver National Center on Poverty Law.

Poverty: Measurement and Analysis

Robert H Haveman, University of Wisconsin–Madison, Madison, WI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 17, pp. 11917–11924, © 2001, Elsevier Ltd.

Abstract

The issue of low income and poverty has concerned social scientists for centuries. A primary issue is to define poverty precisely, and then to document its extent, consistent with this definition. Several definitions of poverty are presented, along with their advantages and disadvantages. Having identified the poor population, questions regarding the nature, causes, and cures of poverty follow naturally. The process of measuring poverty and analyzing its causes and consequences has advanced social science research in several areas, including identifying the underlying causes of poverty, understanding social mobility, status attainment, and income dynamics, and measuring the behavioral effects of antipoverty policy interventions. It has also contributed to the development of social science methods, including randomized social experimentation, econometric techniques for dealing with selectivity, and microdata simulation modeling.

The issue of low income and poverty has concerned social scientists for centuries. In 1776 Adam Smith linked economic poverty directly with the lack of "not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without" (Smith, 1776: Book V, Chap. II, Pt. II, Article 4th). Interest in this problem was a natural accompaniment of efforts to understand why nations experience economic growth. If the economy of a nation grows, it is only natural to wonder how the fruits of this growth will be distributed among people and why some people are excluded from the benefits of economic prosperity. A primary issue is to define poverty precisely, and then to document its extent consistent with this definition, that is, to count the poor. Having identified the poor population, questions regarding the nature, causes, and cures of poverty follow naturally. The process of measuring poverty and analyzing its causes and consequences has advanced social science research methods and techniques.

The Measurement of Poverty

The most prominent early effort to measure poverty was undertaken by B.S. Rowntree in 1902, in his attempt to assess the extent of poverty in York, England (Rowntree, 1902). Since that time, there have been numerous attempts to measure poverty in a variety of countries or more narrowly defined areas.

Although reducing poverty is a nearly universal goal among both nations and scholars, there is no commonly accepted way of identifying who is poor. Some argue for a multidimensional poverty concept, such as being 'housing poor' or 'health poor' or lacking in social contacts, thought of as social deprivation. Economists tend to prefer a concept of hardship that reflects economic position or economic well-being, somehow measured; that concept has come to dominate the measurement of poverty.

The measurement of economic poverty seeks to identify those families whose economic position, defined in terms of *command over resources*, falls below some minimally acceptable level. In addition to requiring a definition of economic well-being, the measure must specify a *minimum level of well-being* (or 'needs') in terms that are commensurate with 'resources.' Sen (1983) considered the needs standard (or poverty line) to have 'some absolute justification of its own,' a level below which "one cannot participate adequately in communal activities, or be free of public shame from failure to satisfy conventions" (p. 167). An ideal poverty measure would allow for differentiation according to household size and composition and would enable people to compare the level of access with resources over time.

Having determined the measure of household resources, the needs standard – or poverty line – appropriate for various households allows one to compare the level of household resources with its needs, and thereby to indicate if the family is poor or not poor. When this process is followed for all households in a nation (or some other jurisdiction), one obtains the *poverty rate*, or the percentage of the people (or households) in the jurisdiction who are living in poverty.

Within this economic perspective, a number of economic well-being indicators have some claim to identifying the economic position of households. For example, the official US poverty measure relies on the *annual cash income* of a family, and compares this with some minimum income standard or 'poverty line.' Alternative – and equally legitimate – measures suggest that the level of *annual consumption* better reflects a family's level of living, or that some measure of a family's *capability to be self-reliant* or a family's *own assessment of economic well-being* are superior concepts for measuring the level and composition of poverty. No matter which perspective is adopted, a large number of judgments and choices are necessary for developing a concrete poverty measure. For example, if income is taken to be the best indicator of economic status, is annual, multiyear, or lifetime income the appropriate measure? Should we examine pretax, pretransfer income, or income after accounting for taxes and/or transfers? Should in-kind income be counted or not? Each decision implies both a different poverty rate and poverty population and a different set of antipoverty policies. Finally, irrespective of the decisions made,

the resulting poverty measure can be either absolute or relative. The indicator is absolute if the definition of needs is fixed, so that the poverty line does not vary with the standard of living. A relative poverty measure reflects changes in the overall standard of living of the society.

Absolute Income Poverty Measures

The most prominent absolute income poverty measure is that of the USA, which is the only industrialized country to have an official poverty measure. See Ruggles (1990) and Citro and Michael (1995) for further discussion of the US definition and other definitions. Since the late 1960s, the nation has reported on the level of national poverty and the composition of the poor population. This indicator seeks to identify those families that lack sufficient annual cash income to meet what is judged to be their annual needs. The needs indicator is a fixed income threshold, constructed originally as a multiple of the food needs of households of different sizes and compositions. This threshold changes only as the price level changes, but not as average incomes in society change.

This and other income poverty indicators have a particular philosophical basis. The implied social objective is that, together, the community's efforts and those of the individual should ensure that some minimal level of living is attained. Although people may experience hardship in many dimensions – education, housing, food, social contacts, security, environmental amenities – only a sufficiently low level of money income matters in determining who is poor. Such a measure also presumes that money income is a good proxy for welfare (or utility), and that a particular year's income is an acceptable indicator of longer-run income.

Aside from taking exception to the social objective that underlies the official measure, most other criticisms of it focus on the adequacy of the annual cash income measure of 'command over resources.' While the current cash income numerator of the poverty ratio may reflect the extent to which the family has cash income available to meet its immediate needs, it indicates little about the level of consumption spending potentially available to the family. For many families, annual income fluctuates substantially over time owing to unemployment, decisions to change jobs, or changes in health. Even as an indicator of a family's ability to meet its immediate needs, the current income measure is flawed – it reflects neither the recipient value of in-kind transfers (e.g., food or medical care subsidies), which support the economic well-being of low-income families nor the taxes for which the family is liable. Similarly, while current cash income – and hence the official poverty measure – reflects financial flows in the form of interest and dividends from the assets held by individuals, the assets themselves are not counted, nor is the value of leisure (or voluntary nonwork) time reflected in the measure. The arbitrary nature of the denominator of the poverty ratio – the minimum-income-needs indicator – has also been criticized. Given the weak conceptual basis and the crude empirical evidence on which the dollar cut-offs rest, including the equivalence scales designed to account for the income needs of different family sizes and structures, the official U.S. poverty lines are viewed generally as arbitrary.

These weaknesses have stimulated efforts in the US to improve the official poverty measure. In 1995, a panel appointed by the National Research Council issued its report suggesting revisions in the official measure. The poverty thresholds were modified so as to reflect a broader commodity bundle of needs (relative to the food need basis of the current thresholds) and a new set of equivalence scales. Moreover, the measure of resources was revised to include the value of in-kind transfers received by a household and to exclude estimates of the labor and income taxes for which the household is liable (see Citro and Michael, 1995). While alternative versions of this new measure are being published, the new measure has not yet been adopted formally.

Relative Income Poverty Measures

Use of a relative income poverty measure takes into consideration changes in the overall economy (in terms of wages and prices); it stresses the relationship of the resources of a living unit to those available to other living units in the society.

One prominent relative poverty definition, used extensively by the European Community, considers those with incomes less than one-half of the median income to be in poverty, hence reflecting the view that poverty is only meaningful when compared with overall income or spending levels. Often, a one-half of median income standard is used in the analysis of poverty in the European Community, but more recently, a 60% of the median standard has been adopted (see Bradshaw, 2001).

In addition to the arbitrariness of the 50% (or 60%) threshold, this measure has other weaknesses. Critics point out the weakness of a relative measure in assessing the efficacy of antipoverty efforts. Moreover, the relative income standard ensures that the poverty threshold will rise most rapidly in periods of economic growth, during which time those at the bottom of the distribution also experience real growth in both earnings and consumption. Hence, even though poor families may perceive themselves as better-off during a prosperous period, the poverty rate may not fall, thus overstating the poverty problem. As Ruggles (1990) has stated: "poverty cannot decline under a relative poverty measure without some change in the shape of the income distribution as a whole."

Consumption-Based Poverty Measures

A primary criticism of income poverty measures is that the annual income concept on which they rest is a poor indicator of the permanent income (or lifetime resources) of the family unit. Using such a measure, a wealthy family with a well-educated head and substantial assets, but a year of low income, would be classified as 'poor.'

One proposal designed to avoid this problem involves use of measured family consumption to determine poverty status, based on the argument that family consumption is a superior proxy for the family's permanent income, or family command over resources. Slesnick (1993) argues that income measures of poverty are 'severely biased' (p. 2) in that *temporarily* low-income families are regarded as *permanently* needy. The consumption-based poverty indicator proposed by Slesnick uses household real consumption expenditure per equivalent

adult (taken to be the quotient of real household consumption and a household-specific cost-of-living index and an equivalence scale) as the indicator of resources.

Slesnick's consumption poverty measure has been criticized on several grounds. One particularly salient criticism concerns the arbitrary and potentially misleading equivalence scales developed by Jorgenson and Slesnick (1987) that are employed (see Triest, 1998). While other consumption-based indicators (see Cutler and Katz, 1991) have shown a growing poverty rate over the last two decades, the Slesnick measure suggests that poverty in the USA has been decreasing secularly over that period.

In general, the largest impediment to utilizing a consumption-based index is the difficulty of obtaining complete and accurate family expenditure data, at least relative to measuring a family's income. Furthermore, consumption may not fully reflect a family's true well-being; it is possible that simple frugality may be mistaken for poverty.

Subjective Measures of Poverty

An alternative view of poverty is the 'Leyden School' approach, which relies on the subjective responses of individuals to questions that seek their impressions of their own economic position or well-being, relative to some norm (see Hagenaars, 1986; van Praag et al., 1982). As with income and consumption poverty measures, subjective poverty indicators are based on an 'access to resources' concept, but because subjective thresholds change over time, these measures are relative poverty indicators.

Subjective poverty measures are based on survey responses to questions regarding the minimum level of income or consumption that respondents consider to be 'just sufficient' to allow them to live a minimally adequate lifestyle. If families respond that their own level of living exceeds that minimum, it is possible by observing their actual income to infer both a poverty line and a poverty rate. Because subjective measures are based on individual opinions of what constitutes 'minimally adequate' or 'enough to get by,' they require that individual perceptions of these notions reflect the same level of real welfare for all respondents.

A number of subjective measures have been developed and tested, mainly in Europe. Despite rather minor differences in terminology and question phrasing among these measures, they have yielded highly diverse results. The effectiveness of subjective measures is also limited by the nature of the data collection method. Most estimates are based on small sample sizes, yielding large standard errors.

Inability to be Self-reliant as a Poverty Measure

One additional poverty indicator should be mentioned – one that views those who are without the capability to be self-reliant as the ones with the lowest economic position. This view rests on two considerations, one conceptual and the other practical. The *conceptual reason* is the more basic. This position has its foundations in the writings of Sen (1992, 1997), who argued, "that the basic failure that poverty implies is one of having minimally adequate capabilities," (1992: p. 111) and hence that "poverty is better seen in terms of capability failure than in terms of the failure to meet the 'basic needs' of specified commodities" (1992: p. 109).

There is also a *policy-related reason* for developing a capability measure of poverty. In recent years, there has been renewed civic debate regarding appropriate norms and standards for individual responsibility and behavior, and hence the appropriate role of the state. A prominent viewpoint in this debate has emphasized the merits of individual independence (relative to reliance on government programs), the negative effects of government programs on individual behavior, and the desirability of a smaller economic and social policy role for government. If society is to base policy on the central social goal of 'economic independence,' then it would seem important to identify the size, composition, and growth of the population of citizens who do not have the capability to be independent in a market economy.

An example of such a capability-based poverty measure is based on the concept of a family's net earnings capacity (the amount that the adults in a family could earn if they worked full time, full year, adjusted for the constraints on working at capacity due to health problems or disability and the expenses required if all of a family's working-age adults did work at capacity). If earnings capacity is less than a poverty threshold, the family is defined as 'unable to be self-reliant,' and in poverty (see Haveman and Bershadker, 2001). A related measure of family capability is Becker's (1965) concept of 'full income,' which includes both income realized through market work and the value of leisure time.

Measuring 'self-reliant poverty' requires several implicit conventions, norms, and assumptions. For example, such a measure is reliable only if the 'norm' of full-time, full-year work is an accepted, socially determined norm representing the full use of human capital. Of necessity, attribution of poverty status to any particular family requires prediction from statistical estimates rather than values measured in survey data (such as income), and hence may be viewed as less objective than, say, income-based measures.

The Analysis of Poverty

Since the mid-1960s and the US declaration of a War on Poverty, social science researchers have addressed a number of issues concerning the nature, causes, and cures of poverty. In the process, substantive additions to economic knowledge and to research methods have been developed.

Understanding the Determinants of Poverty: Demographics, Discrimination, Human Capital, Immigration, and Macroeconomic Performance

A major objective among social science researchers has been to understand the processes which combine to result in some people failing to attain minimal standards of economic well-being. Although the number of such processes and institutions is very large, and the interactions among them are complex, research has been concentrated in several areas.

The underlying *demographic structure* of the population – age, gender, family arrangements, fertility choices – plays a major role in determining whether families or individuals

are poor. For example, the share of the population that is elderly, and hence unable to be self-reliant through work and earnings, influences the prevalence of poverty in a society. Research has tracked the poverty status of the older population and evaluated the ability of older people to escape poverty given their own resources and social support. Similarly, changes in the prevalence of single-parent families caused by rising divorce and non-marital childbearing rates are also fundamental determinants of poverty. Families with children but only a single adult tend to be restricted in their ability to be self-reliant from own earnings. Social scientists have studied the trends and determinants of these choices in an effort to better understand the causes of poverty.

If the attitudes of the majority of the population toward minority groups are such as to lead to lower payments for equivalent work or to inferior schooling, housing, or mobility opportunities, poverty rates among minority groups would tend to be high. Given the prevalence of racial minorities among the poor population, the possible causal role of *labor market, education, and housing discrimination* has been an important topic of study, especially by economists. While the existence and extent of racial income gaps had been studied prior to the 1960s, later research delved deeper than previous work in measuring and explaining the sources of these income gaps. See Cain (1986) for a review of this line of analysis. Given that competition would be likely to erode the effects of discriminatory behavior by employers (Becker, 1957), research turned to explanations involving 'statistical discrimination' and 'market signaling' as the basis for continued discrimination as a cause of poverty.

The role of *education and training* – human capital – in enabling people to attain self-sufficiency also became a focus of poverty research and analysis. The question of the effect of increases in schooling on individual earnings occupied the efforts of numerous researchers, and education came to be analyzed as a process that created human capital. Estimates of the rates of return on schooling seemed to justify diverting resources from private sector investments in physical capital to public investments in education. Rosen (1987) surveys the literature in this area. While economists focused on the earnings returns from schooling, other social scientists sought to document the effect of schooling on student attainment (such as test scores), social mobility, and economic status. This research documented the complexity of the processes by which some children succeed while others fail, and led to a search for improved data and statistical methods capable of disentangling the effects of education on attainments. This 'human capital' approach to schooling carried over to numerous studies of the effectiveness of governmental programs designed to provide skills and training to adult workers. See Heckman and Lochner (2000) for a review and interpretation of the evidence on the economic effects of education and training.

Concern with the problem of poverty in the USA coincided with a rapid increase in *immigration*, often of individuals with low levels of schooling or language competence. The potential impact of immigration in eroding wage rates for unskilled workers in urban labor markets containing large numbers of immigrants became an important policy and research issue.

The effect of the overall performance of the economy on the poverty rate was also the subject of substantial research and analysis, beginning in the 1960s. During the 1970s, economic growth and low unemployment seemed to be powerful instruments in reducing unemployment. However, during the 1980s and early 1990s, economic growth seemed to have lost its antipoverty effectiveness. Evidence suggests that during the 1990s, the level of poverty was again responding to macroeconomic performance. See Haveman and Schwabish (2000) for a review of this research and for recent evidence.

Understanding Social Mobility, Status Attainment, and Income Dynamics

Research on the causes of poverty have tended to focus on static issues – factors that appear to be causally related to low economic position. Simultaneously, a complementary line of research addressed questions of the dynamic processes by which people either succeed or fail to attain economic position – processes that indicate the openness or fairness of society. While research on these social mobility processes had been undertaken prior to the 1960s, those studies were limited by the absence of both a conceptual framework and data on individuals over time – longitudinal data – necessary to trace these dynamic patterns. The early 1960s witnessed a number of important research contributions designed to increase understanding of the process of social mobility. One of the most central of these presented a causal model of the socioeconomic life cycle, distinguishing the interactions among three primary life phases: family, schooling, and work (see Duncan and Hodge, 1963). The socioeconomic status of the job was taken to be the ultimate indicator of success, with family back-ground and schooling (as an intervening variable) viewed as the determinants of that status. This framework stimulated several empirical studies of these relationships, yielding insights into changing intergenerational correlations involving occupations and income, changes in the degree of social mobility, changes in the role of education in fostering mobility, and changes in mobility patterns between whites and nonwhites. A variant of this approach sought to identify the factors that contribute to social or economic status, distinguishing those that relate to exogenous family and background factors from those amenable to policy influence. Two prominent examples of this approach are Jencks et al. (1972, 1979). Family characteristics and choices and intellectual inheritance were seen as playing important roles, but school quality and the racial mix of neighbors and peers were found to be less important. Haveman (1987) reviews and critiques this research.

While these analyses were revealing, they demonstrated the difficulty of making progress in this area without detailed data on individuals and families that tracked their attributes and performance over time. The Panel Study of Income Dynamics, the National Longitudinal Surveys, and the Survey of Income and Program Participation are among the most prominent and high quality of these data collection efforts. Extensive use of them has enabled major advances in understanding the nature and causes of poverty, including the lengths (or persistence) of poverty and welfare stays among the poor, trends in income

trajectories among families with various characteristics, the identification of the long-term or permanently poor population, and estimates of the relative roles of family background, schooling choices, neighborhood characteristics, and labor market discrimination in determining the probability of being poor. A complementary field of study has used these data to explore the determinants of these family characteristics (e.g., poverty-level incomes, single-parent status, low parental education, welfare recipiency), schooling quality, neighborhood characteristics, and own schooling and fertility choices on a variety of children's attainments (see Haveman and Wolfe, 1995; Duncan and Brooks-Gunn, 1997).

Antipoverty Effects of Policy Interventions

Poverty researchers have also sought to describe the various policy measures in place to assist poor families and to identify changes in or alternatives to these policies that might be effective in reducing poverty. Probably because the main anti-poverty efforts were in the area of income support measures, analysts concentrated their assessments on such policies. A voluminous literature describing the nature of income support policies and critiquing their incentive (efficiency) and equity characteristics has influenced policy changes over the period since the 1960s. These analyses revealed low 'take-up' rates among people eligible for support, horizontally uneven support patterns among equally poor people, the uncertain rationale for and implications of providing support via in-kind benefits versus tax subsidies versus cash transfers, the relative effectiveness of alternative programs in reducing poverty rates or the intensity of poverty, and the presence of incentives for undesirable behaviors (e.g., reductions in work, increases in family breakups, and nonmarital childbearing) implicit in the structure or administration of the programs.

These descriptive and policy analysis studies played an important role in guiding numerous antipoverty policy reform efforts that have occurred over the past several decades, including proposals for 'negative income tax' plans, work-related subsidies, the substitution of in-kind for cash benefits, and 'welfare-to-work' initiatives. They also provided a catalog of important questions for which answers would be relevant in designing effective policy: how large are work effort reductions in response to income transfers? Has public income support reduced family stability, increased unemployment, reduced savings, or increased nonmarital childbearing? (see Haveman, 1987).

Labor Supply, Family, and Fertility Effects of Income Support Policy

Concerns with the work incentive effects of income support policy have dominated debates over this antipoverty strategy. Numerous analysts responded to this concern with empirical studies documenting the extent of these behavioral responses to public income support programs. These studies, employing a variety of survey data and empirical methods, found that public benefits tend to reduce employment probabilities, lower work hours per year if working, and possibly reduce wage rates among both recipients and potential recipients of benefits.

Most of the early studies employed cross-sectional survey data in order to measure the responses of individuals to the increased income from transfer benefits and the reduced effective wage rate associated with working. Later studies relied on the results from carefully designed 'social experimental' studies of these potential responses (see below). The essential characteristic of these studies is the random assignment of families to alternative forms of transfer program designs (in terms of the generosity of income guarantees and the extent of 'benefit reduction' rates), and the observation of the behavior and performance of these families over subsequent years. Both forms of studies confirmed that, as expected, work effort falls as nonlabor income rises and as the effective wage rate falls. While the estimates varied widely, nearly all of them raised serious questions about the efficiency and effectiveness of this income support approach to helping poor people.

While these studies focused on the work effort responses of people to the availability of income support, other analyses (including randomized approaches) tested the family structure, childbearing, health care utilization, housing quality, and welfare utilization choices of low-income families in response to both welfare-type policies and reform efforts in the educational, housing, health insurance, and public employment areas. Danziger et al. (1981) and Moffitt (1992) have surveyed this research.

Poverty Research and Social Science Methods

In addition to adding to the body of knowledge regarding the nature, causes, and cures of poverty, the growth of poverty-related research since the 1960s has also contributed to the development of new research methods of general applicability in the social sciences (Haveman, 1987). Four contributions deserve mention.

The first contribution consists of a new field of evaluation analysis and research. As analysts, in response to the demands of policy makers, focused on the poverty problem, an emerging analytical area, evaluation research, was also being spawned. Interest in this applied research approach extended far beyond the poverty issue, but research on social and antipoverty policy issues provided this then-nascent subdiscipline with some of its basic methods and case studies. Without the emphasis of policy makers on the need to evaluate the effectiveness of the numerous new antipoverty initiatives, it is unlikely that this field of analysis, with its journals and professional organizations, would have developed as rapidly as it has, and in the same directions.

Second, as noted above, randomized social experimental research was first applied to proposed policy interventions in the poverty area – the negative income tax proposals, school and housing vouchers, employment and training measures, and so on. Without the emphasis on the effect of antipoverty policy interventions on family economic status during the 1960s and 1970s, it is unlikely that this important social science research method would have attained the status it now holds among social science researchers.

One of the problems that analysts face in assessing the effects of policy and other changes on the behavior of people is

the need to rely on samples of observations that may be composed of individuals who are selectively included in the sample, and hence unrepresentative of the larger population for which inferences are desired. This concern dominated the study of labor markets well before demands to assess the impacts of antipoverty measures, especially in the employment and training area, where factors that determined the presence of a person in a research sample were also determinants of the effect that was being studied (see Ashenfelter, 1978). Several approaches to this problem were developed by econometricians (see especially Heckman, 1979), and these statistical tools became a standard part of applied statistical analyses.

Finally, in part because of the development of large representative samples of national populations, analysts began developing computer models designed to simulate the effects on the income, well-being, and behavior of each observation in a data set of actual or proposed policy measures. The application of such microdata simulation models found its most welcomed audience among analysts concerned with the effects of changes in taxation and income transfer policy on low-income families. Without the impetus provided by the need to assess the impacts on poor people of proposed policies, this advance to predictive model building would not have occurred. This research methodology is assessed in Citro and Hanushek (1991).

See also: Income Distribution and Inequality: Measurement Issues; Mobility, Social; Poverty Policy; Poverty in History; Poverty, Culture of; Poverty, Sociology of; Skills Training: Public Policy Aspects; Social Inequality in History (Stratification and Classes); Wealth Distribution.

Bibliography

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. Review of Economics and Statistics 60, 47–57.

Becker, G.S., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.

Becker, G.S., 1965. A theory of the allocation of time. Economic Journal 75, 493–517.

Bradshaw, J., 2001. The measurement of absolute poverty. In: Schokkaert, E. (Ed.), Ethics, Poverty and Inequality and Reform in Social Security. Ashgate, Aldershot, UK.

Cain, G.C., 1986. The economic analysis of labor market discrimination: a survey. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics. North Holland, Amsterdam.

Citro, C., Hanushek, E., 1991. Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. National Academy Press, Washington, DC.

Citro, C., Michael, R., 1995. Measuring Poverty: A New Approach. National Academy Press, Washington, DC.

Cutler, D.M., Katz, L.F., 1991. Macroeconomic performance and the disadvantaged. Brookings Papers on Economic Activity 2, 1–61.

Danziger, S., Haveman, R., Plotnick, R., 1981. How income transfer programs affect work, savings, and the income distribution: a critical review. Journal of Economic Literature 19, 975–1028.

Duncan, G.J., Brooks-Gunn, J. (Eds.), 1997. Consequences of Growing Up Poor. Russell Sage Foundation, New York.

Duncan, O.D., Hodge, R.W., 1963. Education and occupational mobility. American Journal of Sociology 68, 629–644.

Hagenaars, A.J.M., 1986. The Perception of Poverty. North Holland, Amsterdam.

Haveman, R., 1987. Poverty Policy and Poverty Research. University of Wisconsin Press, Madison, WI.

Haveman, R., Bershadker, A., 2001. The 'inability to be self-reliant' as an indicator of poverty: trends for the US 1975–1997. Review of Income and Wealth 47 (3).

Haveman, R., Schwabish, J., 2000. Has macroeconomic performance regained its antipoverty bite? Contemporary Economic Policy 18 (4), 415–427.

Haveman, R., Wolfe, B., 1995. The determinants of children's attainments: a review of methods and findings. Journal of Economic Literature 33 (4), 1829–1878.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 147, 153–161.

Heckman, J., Lochner, L., 2000. Rethinking education and training policy: understanding the sources of skill formation in a modern economy. In: Danziger, S., Waldfogel, J. (Eds.), Securing the Future: Investing in Children from Birth to College. Russell Sage Foundation, New York.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., Michelson, S., 1972. Inequality: A Reassessment of the Effect of Family and Schooling in America. Basic Books, New York.

Jencks, C., Bartlett, S., Corcorau, M., Crouse, J., Eaglesfield, D., Jackson, G., McClelland, K., Mueser, P., Olneck, M., Schwartz, J., Ward, S., Williams, J., 1979. Who Gets Ahead? The Determinants of Economic Success in America. Basic Books, New York.

Jorgenson, D.W., Slesnick, D.T., 1987. Aggregate consumer behavior and household equivalence scales. Journal of Business and Economic Statistics 5, 219–232.

Moffitt, R., 1992. Incentive effects of the US welfare system: a review. Journal of Economic Literature 30 (1), 1–61.

van Praag, B.M.S., Hagenaars, A.J.M., van Weeren, H., 1982. Poverty in Europe. Review of Income and Wealth 28 (3), 345–359.

Rosen, S., 1987. Human capital. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 2. Palgrave Publishers, Hampshire, UK, pp. 681–690.

Rowntree, B.S., 1902. Poverty, a Study of Town Life. Macmillan, London.

Ruggles, P., 1990. Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy. Urban Institute Press, Washington, DC.

Sen, A., 1983. Poor, relatively speaking. Oxford Economic Papers 35, 153–169.

Sen, A., 1992. Inequality Reexamined. Russell Sage Foundation and Harvard University Press, Cambridge, MA.

Sen, A., 1997. On Economic Inequality. Clarendon Press, Oxford, UK.

Slesnick, D.T., 1993. Gaining ground: poverty in the postwar United States. Journal of Political Economy 101 (1), 1–38.

Smith, A., 1776. Wealth of Nations. Modern Library Edition, 1993. Random House, New York.

Triest, R.K., 1998. Has poverty gotten worse? Journal of Economic Perspectives 12 (1), 97–114.

Poverty Policy

Hilary Silver, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by Z. Ferge, volume 17, pp. 11924–11928, © 2001, Elsevier Ltd.

Abstract

Antipoverty policies include social assistance programs and 'safety nets' for low-income individuals and households. These policies place a floor under consumption levels or provide a minimum income by transferring cash or in-kind assistance or redistributing access to assets or services. Although some transfers are universal, antipoverty policies typically target social categories or apply means tests for benefits, thereby defining who is poor and deserving of aid. Recent reforms made entitlements to assistance more conditional on work requirements or other active obligations.

Definition and Overview

Antipoverty policies include social assistance programs and 'safety nets' for poor and vulnerable individuals and households. These policies place a floor under consumption levels or provide a minimum income by transferring cash or in-kind assistance or redistributing access to assets or services. Some antipoverty benefits are universal rights of citizenship, while others are conditional on fulfilling obligations. Typically, social assistance policies target categories of people who face risks associated with the inability to work – the disabled, the elderly, expectant and new parents, the unemployed. Such benefit programs may apply means tests of eligibility, developing official poverty thresholds that define who is poor and deserving of aid. Poverty measures are also used to monitor and evaluate the performance of antipoverty policies over time. Antipoverty policies are typically financed from general taxation rather than personal contributions to insurance against risks over the life course. However, unemployment, disability, and old-age benefits also prevent destitution among former workers.

Historically, religious charity to the needy preceded public antipoverty programs. Private nonprofit aid programs, sometimes encouraged by public subsidies or tax exemptions, supplement public assistance to varying extents. Public programs may provide resources through tax systems as well as monetary and in-kind transfer payments. The structure or mix of taxes may entail trade-offs with poverty reduction. Contrary to conventional wisdom in economics, progressive income taxes can reduce poverty without necessarily reducing economic growth. More expected, payroll taxation does apparently reduce job growth. Tax credits to top up low earnings may discourage added work hours.

Recently welfare states have reformed guaranteed entitlements, making benefits more conditional. Like the historical workhouses, programs increasingly enforce work requirements or other active obligations of the poor in return for assistance or 'workfare.' However, working is no guarantee of escaping poverty. In recent years, the number of working poor has increased. Rising employment rates have not reduced aggregate poverty in many countries of the world. Therefore, antipoverty transfer policies are supplemented by tax credit and other programs like mandatory minimum wages that 'make work pay.' Training services indirectly enable poor people to leave poverty on their own. These are sometimes called preventive, promotion, or enabling policies, activation, or social investment.

The notion of social investment currently widespread in the European Union (EU) points to the ways antipoverty policies contribute to the overall public good. Indeed, some antipoverty policies contribute to economic growth by enhancing worker productivity. Even when the business cycle turns down, benefits maintain aggregate demand or reduce unemployment. Antipoverty programs may also contribute to social peace and public order. These societal benefits help justify general taxation to pay for social assistance.

Antipoverty approaches and institutions vary widely across regions of the world. The Anglo-American approach emerged from early, local-level poor relief policies establishing the principle of 'less eligibility' and distinguishing between the deserving and undeserving, able-bodied poor. Increasingly this 'neoliberal' distinction is discerned in conditional cash allowances adopted in developing countries. In contrast, socialist and social democratic antipoverty policies traditionally minimized the distinction between workers and the dependent poor, emphasizing full employment and universal citizenship rights to a minimum income. The European Council's Recommendation 92 of 1992 recognizes "the basic right of a person to sufficient resources and social assistance to live in a manner compatible with human dignity." Yet, even in European countries with a right to a minimum income, there is a trend toward encouraging more active participation of social assistance recipients.

History and Development of Anti-Poverty Programs

Public policies to help the poor evolved from religious charity. For millennia, poverty was considered inevitable, but could be alleviated by benevolence. Even today, charities and religious institutions at all scales of operation provide assistance and services to poor people in much of the world. In Europe, government involvement in managing poverty began at the local level in response to pauperism and itinerant laborers.

The earliest publicly financed 'workhouses' were found in Amsterdam around 1600, but, in line with mercantilist thinking, developed extensively in Britain. The Elizabethan Poor Laws instructed parishes to provide 'poor relief' for old age, widowhood, disability, illness, or unemployment. Initially restricted to residents, the 'workhouse' or 'indoor relief' included all those dependent on aid, from the disabled to unwed mothers, who were obliged to contribute to their own support and refrain from immoral conduct.

Prior to the late eighteenth century, the dominant school of economic thought saw poverty as a social good, essential for economic development because it enforced the work ethic, kept wages low, and created a globally competitive economy (Ravallion, 2013). Such liberal welfare principles attribute poverty to individual causes or personal characteristics, such as insufficient investment in human capital, or to bad choices and moral failings, such as promiscuity or insufficient work effort. In 1834, Britain amended its Poor Law because parishes had provided a minimum level of subsistence that exceeded the level of wages. To remove this apparent work disincentive hindering the rise of a labor market in this first industrial nation, the Poor Law reform established the general principle of 'less eligibility' in which relief or public assistance in workhouses had to be less generous than what workers could earn in wages. Anglo-American welfare programs have retained this principle by keeping wage replacement levels in social assistance programs low, eligibility rules narrow, and work requirements strict. For example, at the turn of the twentieth century, Seebohm Rowntree inquired into the wages of the poor relative to the minimal subsistence costs for housing, clothing, fuel, and food, identifying a basic diet of 3478 calories per day. Low benefit levels and narrow eligibility for welfare thus 'regulate the poor' and discipline the wage labor force.

Except for relief under the Poor Laws in England and Wales, there was little increase in the low levels of private or public support for the poor for most of the nineteenth century. Gradually, the poor were sorted into categories. Those considered 'deserving' of public support were provided 'outdoor relief' and could live independently, while others deemed able-bodied were considered 'undeserving' of help and put to work in return for support. These categories persisted so that today, the sick and disabled, women and children are still disproportionately poor and dependent upon distinct programs of social assistance.

Beginning in the late nineteenth century, after numerous cycles of boom and bust, industrializing societies began to establish social insurance programs to cover temporary poverty among able-bodied workers who suffered from job losses or accidents that were no fault of their own. By contributing to insurance funds during good times, workers could protect themselves from the risk of destitution during recessions or incapacity. Thus, for much of the twentieth century, poverty programs *sensu stricto* shrunk to cover small categories of people – widows, the aged, and the infirm – who were not expected to work in the labor market. Often, local governments were charged with providing social assistance for these special needs populations, in contrast to nationalized social insurance funds that created large risk pools.

After the Oil Shocks of the 1970s, welfare states began to change. Slower growth, higher unemployment, and population aging raised fiscal concerns that existing social protection commitments were unsustainable. These 'austerity' conditions prodded many countries to institute 'neoliberal' reforms, cutting social expenditures relative to GDP, tightening eligibility, lowering benefit levels, shortening terms for assistance, and enforcing conditions in return for public support. This is not to say that different welfare regimes began converging, but they were changing in similar ways.

Distinctions among Poverty Programs

Antipoverty programs differ in many respects: eligibility rules; scale and method of governance; activation or work conditions; benefit levels; adjustment for household, life stage, and other categorical differences; administrative costs; and total budget. During the postwar era, mature Western welfare states usually distinguished between insurance against risks and assistance with current needs. This distinction is crosscut by one of universalism versus categorical programs. Universalism entitles all citizens to benefits by right, in contrast to categorical programs that restrict assistance to particular groups or conditions. Universalist programs have been effective at poverty reduction although their effects on inequality vary. Even if low-income households do not receive higher benefits than affluent ones, the benefits constitute a larger share of the poor's total income. Universal pensions may provide a minimum income to all elderly citizens regardless of prior earnings, but better paid workers often enjoy greater or supplemental old-age payments. In turn, taxes may claw back the higher income. Unlike universalist 'social democratic' welfare states, 'Bismarckian' systems, widespread in Continental Europe, are less progressive in the sense of reducing inequality. Even if they offer a floor of income, they tend to preserve inequalities derived from current or prior earnings. In contrast, 'social security' in the US pays retired low-wage workers more than they contributed over a lifetime, while capping maximum payouts to high earners. This progressive system is credited with reducing poverty among older Americans even though the payroll taxes financing it are also capped and regressive. Recent cross-national research suggests poverty rates decline with a larger share of transfers in household income, with homogeneity of transfers, and with targeting of transfers on low-income households.

Categorical programs tend to be more efficient than universalist ones at targeting those in greatest need. The categories vary, but usually include those whom society does not deem able to work. They may include people with disabilities, the aged or children, and sometimes, home-based caregivers, expectant mothers, or lone parents. Aid to single nonworking parents has been declining, while paid parental leave from work has increased. The efficiency of targeting those in greatest need may reduce the overall cost of poverty programs, but savings may be offset by increased administrative costs to police eligibility. Targeting also undermines broad political support.

Categorical programs may or may not 'means-tested.' Sometimes categorical programs like family allowances are universally provided to parents with equal benefits per additional child, regardless of means, but when poor households have

higher birth rates, they may disproportionately benefit. Universal child benefits may also be taxed based on income, clawing back public resources from wealthier families. Nevertheless, eligibility for antipoverty programs per se is usually restricted to those who can demonstrate a lack of other resources, including insurance coverage. Providing means-tested subsidies through the tax system eliminates the expensive and stigmatizing process of documenting assets and income and of meeting a case worker that traditional social assistance or 'welfare' programs often require.

Finally, transfer payments may be made in money or in-kind. Food, housing, and other voucher programs not only have higher administrative costs than cash assistance, but the visible redemption of vouchers also subjects recipients to stigma and shame. In-kind assistance also conveys a society's distrust of the poor, questioning their ability to make good choices. In contrast, minimum or basic income programs often rest upon a guaranteed human right to subsistence.

Entitlements versus Conditions

Social assistance may be provided as of right or on condition that the poor meet obligations. Most mature welfare states now have minimum income schemes that support those below a recognized threshold and who have no recourse to contributory insurance protection. Rather than being targeted at social categories, minimum incomes are quasi-universal, offering a guaranteed right to support. They vary widely in generosity. Countries with the most generous and effective minimum incomes see them as social safety nets that prevent the unemployed from becoming so demoralized and excluded that they are incapable of social participation or effectively seeking work.

A basic income guarantee (BIG) is an unconditional, universal, government-insured citizen entitlement to enough income to meet basic needs. In theory it is not means-tested, so it costs less administratively than even a negative income tax, although it may have potential work disincentives. The guarantee makes it possible for poor individuals to make high-return investments in job search, child care, etc. that poverty might otherwise make prohibitive. An example of a BIG is the South African Child Support Grant, an unconditional, means-tested cash transfer program introduced in 1998. One study found that receiving a Child Support Grant was associated with increased school attendance, decreased child hunger, and increased broad labor force participation, while it had no identifiable effect on narrow labor force participation or employment.

Related to BIGs are work guarantees, albeit in unskilled jobs at low minimum wages. In India, the Maharashtra Employment Guarantee Scheme, started in 1973, became the model for the national Mahatma Gandhi National Rural Employment Program, passed in 2005. The latter guarantees landless laborers and small farmers the right to a minimum of 100 days of paid work per household per year. Because of administrative limitations, however, it is forced to ration the work, providing 50 days of employment a year to around 50 million households. Almost half of the beneficiaries are women, reducing their high poverty gap. In 2001, India's Supreme Court also mandated the provision of cooked lunches to all primary school children. The mid-day meal program distributes surplus grain to about 55% of Indian children. According to the Center for Social Equity in New Delhi, the program has a demonstrable effect on school attendance, especially of girls, and encourages caste mixing.

In contrast to such guarantees, social assistance may be made contingent upon work effort or other conditions, such as motherhood or disability. Traditionally, minimum income and categorical means-tested programs expected able-bodied beneficiaries to seek work, but these obligations were not strictly enforced until recently. Beginning in the 1980s in the United States, mature welfare states have reformed their 'passive' social assistance benefits requiring little in return into programs 'activating' recipients to engage in work or other forms of social participation.

'Workfare' is a derogatory term for mandatory work or community service requirements in return for income support. Many consider these requirements to be a punitive throwback to the nineteenth-century 'workhouse' or a form of 'labor recommodification.' The term originated in the United States where, after many state-level experiments, the 1996 welfare reform (Personal Responsibility and Work Opportunity Reconciliation Act) imposed work requirements on single parents receiving Temporary Assistance to Needy Families (TANF, which replaced Aid to Families with Dependent Children, also known as 'welfare.'). 'Workfare' included carrots and sticks. Although progressive states added funding for work supports, such as child care subsidies, subsidized health insurance, nutrition assistance, transportation and clothing allowances, conservative states introduced more punitive measures, such as family caps, sanctions, and lifetime time limits on support shorter than the federal one of 5 years. Many applicants lost benefits if they did not accept community service requirements, even because they were attending school. Others were diverted to 'work first' before they were ever granted benefits.

Welfare reform had the effect of halving the number of beneficiaries on the rolls. Mandatory work (or work-related activities) reduced welfare use and increased earnings and employment. More generous child care subsidies also promoted maternal employment. But welfare reform also increased the number of 'working poor' households, i.e., those with insufficient earnings to move them above the poverty line, who more than ever relied on other support services and tax credits to survive. The expansion of the Earned Income Tax Credit as people were moving from welfare to work did help 'make work pay.' Negative income taxes like Earned Income Tax Credits are poverty policies based on neoliberal assumptions that welfare undermines work incentives and creates poverty traps, and employers need wage subsidies to price welfare recipients into work.

Most of the evidence suggesting positive impacts of welfare reform was collected during a period of economic expansion, prior to the Great Recession. Initially, the most economically disadvantaged families received the largest amount of support, but by the mid-2000s, families that near or just above the poverty line received substantially larger transfers than in the past. It is too soon to assess how well the program served as a safety net during the Great Recession. Welfare rolls did rise

from 2007 to 2009, but TANF still had work requirements and time limits.

Minimum Incomes and Activation

While liberal welfare states offered low welfare payments and short-term unemployment benefits, Continental Europe long considered social assistance policies to be marginal to the welfare state. Assistance addressed 'residual' or categorical needs not covered by social insurance, was financed from general taxation, and was provided locally with poor transparency. Some needy people were uncovered and fell through the cracks in the safety net. In fact, France had no universal minimum income program until 1988, and Portugal introduced one as late as 1997. Since the 1980s, as high unemployment persisted, more and more jobless Europeans remained on relatively generous unemployment insurance benefits. Gradually, benefit durations were shortened, replacement rates reduced with time, and more employment services provided. With rising costs, welfare states tried to shift long-term unemployed insurance recipients into less generous social assistance or merged the programs, as in Germany's Hartz reforms. Over time, minimum income or social assistance recipients increased, suggesting a spiral into 'social exclusion.'

Therefore, most European minimum income programs have adopted stricter tests on ability to work, enforced job search obligations, required jobseekers to accept employment offers even in different occupations and at lower wages than before, and instituted 'activation' policies. Activation programs differ from 'workfare' in that they activate not only social assistance recipients but also those receiving unemployment, disability, and other benefits, they may use taxes and benefits as incentives, and they may require activities besides paid market employment (Lodemel and Moreira, 2014). These activities may include participation in public or subsidized employment, or in other activities, such as training and job search. Activation schemes often target hard-to-employ groups: the long-term unemployed, ethnic minorities, lone mothers, and so on.

Activation programs are modeled on the active labor market policies (ALMP) of the Nordic welfare states. These policies discouraged long-term dependence on what were, comparatively speaking, generous minimum incomes. Besides the macroeconomic intention of income maintenance during downturns, the whole complex of Scandinavian welfare state institutions – health, education, training, employment services, labor exchanges, transportation, child care – worked together to keep people employed through relocation or retraining or to rapidly moving them back into jobs in demand before their skills atrophy. Yet Nordic ALMP rests upon a whole range of human services, usually public, that do not exist elsewhere. Also, unlike activation of minimum incomes, ALMP eligibility was not tied to income.

Nordic welfare states reconcile decommodification policies insuring basic needs with policies encouraging rapid reemployment. With active labor market expenditures and short-term, rather than long-term unemployment insurance, Scandinavians had shorter average unemployment spells but high wage replacement rates. Today Denmark, the Netherlands, and other welfare states pursue 'flexicurity' policies to allow firms to dismiss unneeded workers who are nonetheless protected from poverty during jobless transitions and are helped to find new jobs. Similarly, in Germany, flexible measures to prevent unemployment, such as 'mini-jobs' and shortened hours ('Kurzeit'), keep workers from being laid off entirely during downturns. In addition, wage subsidies may compensate employers for retaining their workers or hiring new ones.

Over time, European activation programs began to exhibit some common trends. Training and active labor market programs contracted. Work requirements were enforced with sanctions, and paid work was emphasized over subsidized, sheltered employment, or community service. Greater financial incentives were offered to make work pay. Benefit recipients received more tailored, individualized attention from case managers, but also had to sign personal action plans or integration contracts specifying a long-term course of action toward integration. Meanwhile, the administration of activation was outsourced to private employment services that made greater use of Internet-based job searches and on-line job banks. Job search supports, personalized employment services, and enforcement of job searches vary cross-nationally. Yet the impact of activation on poverty appears limited. One study of 26 European countries found that, controlling for social assistance benefit levels, there was no clear additional effect of public services or ALMP on material deprivation.

Ironically, even though many governments have reduced long-term or on-the-job training that could enhance labor productivity, the European Union has adopted the idea of 'social investment' and 'flex-security' policies. This 'inclusionary' approach aims to reduce long-term joblessness by encouraging mobility and reemploying people in high wage jobs. Yet pressuring the jobless to accept any job offer by threatening to reduce benefits for refusals leaves vulnerable many extremely poor people who are not ready for work – the hungry, homeless, ill, addicted, and materially deprived children. Tacitly acknowledging two tiers of unemployment, the EU instituted a Fund for European Aid to the Most Deprived. It aims to help 4 million people with nonfinancial material assistance (food, clothing, shoes, soap, etc.) and will be financed from the Cohesion Policy budget, rather the European Social Fund, the main instrument of the 'social investment' approach to fight poverty. Thus, an un/deserving distinction among the poor is arising in Europe too.

Growth, Employment, and Poverty

Alongside explicit antipoverty programs, most countries rely on macroeconomic growth and labor market policies to reduce poverty indirectly through insuring gainful employment for all. Over time, industrial societies expected antipoverty policies to contribute to economic growth, not impede it, as investments in human capital – health, education, and training – enhanced productivity. To enable workers to share in rising affluence, governments statutorily increased the returns to work through minimum wages or mandated employer benefits, or encourage organized labor and business to bargain collectively over these. In recent years, Brazil, China, and other emerging economies have raised the minimum wage as an

antipoverty measure. US evidence shows that increases in the minimum wage to 'livable' levels substantially raise earnings at the bottom of the distribution, reduce wage inequality, and do not impede job creation, although there are probably spillover effects on those earning slightly above the minimum wage (Belman and Wolfson, 2014).

Many consider economic growth to be both a necessary and sufficient condition for reducing extreme poverty. In line with longstanding active labor market programs, the EU has tied poverty reduction and social inclusion to employment and growth. Europe 2020 is the EU's post-Lisbon social investment strategy for smart, sustainable, and inclusive growth. This politically attractive approach to poverty reduction aims to lift at least 20 million people out of poverty and social exclusion and to increase employment of the population to 75%. To reach these targets, the strategy includes a flagship initiative, the Platform against Poverty and Social Exclusion. The Platform calls for coordination of labor market policies, minimum income support, health care, education, housing, and access to basic banking across EU countries; earmarked European Social Funds (through PROGRESS) for fighting poverty and social exclusion; evidence-based social policy evaluations; and partnerships with civil society and participation of people experiencing poverty.

The modernization of social protection in EU states is intended to promote social investment throughout the life course. This means guaranteeing youth a job or training, promoting prime age employment, integrating social assistance and unemployment insurance, and making pensions sustainable. Workforce development, dual apprenticeship, on-the-job training, and adult and continuing education programs aim to increase human capital and thereby, wages. Internships, transitional jobs, and other forms of temporary or trial employment are opportunities for jobless workers to convince employers to hire them permanently. To the extent that people can more easily find jobs that match their skills, then employment services, job search programs, and mobility assistance will also contribute to reducing poverty.

Yet, given reduced spending on training, some scholars are concerned that relatively expensive social investment programs may 'crowd out' traditional welfare programs. Indeed, the EU poverty rate has remained around 15% since the mid-1990s, despite rising employment rates. This was partly because jobs went to new labor market entrants, not the long-term unemployed. But it was also because employment in the new economy no longer insures an escape from poverty. Hours of work are as important as minimum wage rates and earned income tax credits in determining poverty.

Europe's social investment approach to poverty reduction apparently entails trade-offs, or a 'trilemma' making it hard simultaneously to ensure that: (1) the unemployed are not poor, (2) budgetary constraints are recognized, and (3) employment growth is sufficient to reduce benefit dependency. Until the Great Recession, the trilemma was avoided by making it more difficult to claim and stay on benefits. Even without reducing the level or replacement rates of benefits, social spending was kept under control, wage inequalities remained largely unchanged, and employment rose significantly. European governments balanced the books mainly by lowering social protection for those outside the labor market. Externalizing costs kept outsiders' poverty rates high (Cantillon, 2011). The economic crisis added to social exclusion, material deprivation, and poverty, requiring special measures for the 'most deprived.'

Conditional Cash Transfers

Conditional cash transfers (CCT) impose obligations on poor recipients besides work. Varying across countries in generosity, cash assistance is made conditional upon certain eligibility criteria and actions such as enrolling children in public schools, getting vaccinations or regular doctor's check-ups, and other investments in human capital. CCTs address the intergenerational transmission of poverty and foster social inclusion and development.

CCTs expanded rapidly after 2001. The World Bank says over 30 developing countries now have CCT programs and the number is growing. Some of the best known are Brazil's Bolsa Escola program started in the 1990s and its Bolsa Familia that provides monthly cash payments to poor households if their children are enrolled in school. Mexico's Oportunidades, originally called Progresa, is the principal antipoverty program of the Mexican government. It helps poor families in rural and urban communities improve the education, health, and nutrition of their children. Supplemented by Apoyo Alimentario, Oportunidades reaches about a fourth of those in need and has reduced hunger despite rising food prices. Chile's Solidario, Colombia's Familias en Acción, Peru's Juntos, and other CCTs throughout Latin America and the Caribbean all pair cash transfers with human capital investments. More recently, CCTs are being developed in the Middle East and Asia as well.

CCTs have resulted in sizable reductions in poverty among recipients – especially when the transfer has been sufficient, well targeted, and structured in a way that does not discourage recipients from taking other actions to escape poverty. The spread of cash transfers alongside other kinds of social safety net programs helped an estimated 50 million people in 2010 raise their incomes above $1.25 a day (Ravillion, 2013). Because CCTs provide a steady income, they have helped protect poor households from the worst effects of unemployment, catastrophic illness, and other sudden income shocks. Since virtually all CCTs make cash transfers to women, they may have increased female bargaining power too.

The transfers do have weaknesses. Although intended to be universal, people with disabilities or in remote areas with insufficient public services or transportation may be excluded from such programs. Poor education and health systems limit the benefits of CCT programs. There is also some evidence that, once recipients enter the labor force, CCTs do little to increase the quality of education, learning, nor salaries.

Poverty Policies in the Global South

Less developed countries have more poverty than the OECD countries, and fewer means to reduce it. Besides CCTs, the Global South frequently relies on in-kind safety nets, mostly subsidized food and fuel. In the 1990s, international agencies often helped developing countries subsidize basic commodities,

but international loans created large, unsustainable debts. Increasing prices of staple goods could set off social unrest as well. Therefore, international agencies and developing states now approach poverty reduction through the pursuit of economic growth and development. Whereas Western welfare states provide essential public services and redistribute income in response to democratic demands, Eastern states are sometimes called 'productivist' or 'developmental' because they integrate social programs with top-down macroeconomic growth policies.

To the extent that poverty reduction programs have been introduced in the Global South, they are mainly found in middle-income countries. Different regions have education, health, and social insurance programs with different levels of coverage and different public and private mixes. In most countries, both democratic and nondemocratic, public employees are usually offered good social protection. Newly democratic middle-income countries in East Asia and Eastern Europe are establishing or expanding unemployment insurance, although it is limited in Latin America. Some East Asian countries have instituted universalist, redistributive social policies with high coverage, while authoritarian Malaysia and Singapore took a more liberal approach to safety nets. Latin America had 'informal security regimes' that until recently relied upon traditional family and community relationships to meet basic needs. More recently, neoliberal reforms have expanded CCTs as well as the privatization of old-age insurance, to produce a 'liberal-informal' welfare regime. Yet some Latin American countries also offer widely available, but means-tested health and education programs. After the fall of communism, Eastern Europe used early retirement disability and family allowances to cope with unemployment, but the programs did not help people escape poverty. They are turning to unemployment insurance.

China and some other command economies do provide income support and subsidized commodities, but tie them to a fixed place of residence and workplace. Since the 1980s, China has also dedicated resources to the development of poor areas in expectation of geographic spillovers. This territorial approach has both impeded labor mobility and left illegal urban migrants exposed to poverty, even though poverty in China has fallen with economic growth.

Another, more indirect approach to poverty reduction in low-income countries is micro-credit or micro-finance programs. Making it easier for poor people to borrow makes it possible to invest and improve their incomes and provide insurance to smooth consumption during downturns. Reflecting the large size of informal economies where small enterprises and self-employment prevail, access to credit – often through mutual savings funds or rotating credit associations – can move people, women especially, over the poverty threshold. While often rooted in ethnic or local bonds of enforceable trust, these micro-credit programs are nevertheless vulnerable to external risks and rarely enable micro-enterprises to grow much in size. However, social entrepreneurship is an increasingly popular neoliberal approach to development among international organizations.

Many countries lack domestic safety net or antipoverty programs entirely. Without stabilizers of domestic demand, global recessions can propel large numbers of people into poverty and create political unrest. Lower-income countries lack the resources and administrative capacity to establish full-fledged social assistance programs. At best, a few like Kenya and Zambia have launched pilot programs. Some therefore refer to African welfare regimes as 'insecurity' or 'ill-fare' regimes. The World Bank's Social Funds aim to help the poorest countries institute targeted antipoverty programs and finance public works in poor communities. Provided corruption does not divert funds, such development projects can contribute to human development and provide a rapid local economic stimulus. Since the 1990 *World Development Report,* a 'world free of poverty' became the World Bank's overarching goal.

See also: Cash Transfers; Climate Change and Global Poverty; Conditional Cash Transfers: Influence on Marriage and Fertility; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Poverty: Measurement and Analysis; Urban Poverty in Neighborhoods; Welfare Economics; Welfare Reform; Welfare Retrenchment; Welfare and Education; Welfare and Work; Welfare.

Bibliography

Barrientos, A., Hulme, D. (Eds.), 2010. Social Protection for the Poor and Poorest: Concepts, Policies and Politics. Palgrave Macmillan.

Belman, D., Wolfson, P., 2014. What Does the Minimum Wage Do? Upohn, Kalamazoo, MI.

Brady, D., 2009. Rich Democracies, Poor People: How Politics Explain Poverty. Oxford University Press, Oxford.

Cantillon, B., 2011. The paradox of the social investment state: growth, employment and poverty in the Lisbon era. Journal of European Social Policy 21, 432–449.

Fiszbein, A., Schady, N., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank Publications, Washington.

Frazer, H., Marlier, E., Nicaise, I., 2010. A Social Inclusion Roadmap for Europe 2020. Garant, Antwerpen/Apeldoorn.

Gough, I., Wood, G., Bevan, P., Barrientos, A., Davis, P., Room, G., 2004. Insecurity and Welfare Regimes in Asia, Africa and Latin America: Social Policy in Development Contexts. Cambridge University Press, New York.

Haughton, J., Khandker, S., 2009. Handbook of Poverty and Inequality. The World Bank, Washington.

Kenworthy, L., 2011. Progress for the Poor. Oxford University Press, Oxford.

Lodemel, I., Moreira, A. (Eds.), 2014. Activation or Workfare? Governance and Neo-Liberal Convergence. Oxford University Press, Oxford.

Mares, I., Carnes, M., 2009. Social policy in developing countries. Annual Review of Political Science 12, 93–113.

Marx, I., Nelson, K. (Eds.), 2013. Minimum Income Protection in Flux. Palgrave Macmillan, New York.

Ravallion, M., 2013. The Idea of Antipoverty Policy. National Bureau of Economic Research, Boston. http://www.nber.org/papers/w19210.

Walker, R., 2014. The Shame of Poverty. Oxford University Press, Oxford.

Ziliak, J.P. (Ed.), 2009. Welfare Reform and Its Long-Term Consequences for America's Poor. Cambridge University Press, New York.

Poverty, Sociology of

Tim Smeeding, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T.M. Smeeding, volume 17, pp. 11928–11932, © 2001, Elsevier Ltd.

Abstract

Sociologists and other social scientists have used a variety of approaches in explaining and describing poverty, as well as methods to define and measure poverty. Herein, we examine several approaches and concepts of poverty measurement. Each approach has its strengths and weaknesses, but all end up with some definition or description of poverty and the attributes of the poor. Also debated are the sociological theories of why poverty exists, and its relation to family, culture, and context. Sociologists have provided important tools for testing causal hypotheses concerning the explanations, causes, and cures for poverty regardless of how it is measured. These include place, culture, stratification, space, race, gender, imprisonment, and other popular theories of poverty in sociology. Sociological studies of poverty have both a rich heritage and have become an important area of research in increasingly unequal rich and poor societies.

Disclaimer

The author thanks David Chancellor for expert research and manuscript preparation assistance. Any views expressed in this article are those of the author alone and not those of the Institute for Research on Poverty.

Sociologists have long written about the topic of poverty, as have economists and other social scientists. The fundamental concept of poverty concerns itself with having too few resources or capabilities to participate fully in a society. As Blank (2008) reminds us, "poverty is an inherently vague concept and developing a poverty measure involves a number of relatively arbitrary assumptions." Ultimately, all social scientists need to first establish the breadth and depth of this social phenomenon called 'poverty' before they can meaningfully analyze it and explore its ultimate causes and remedies. We begin here with sociological approaches to the origins and nature of poverty, both measurement and theory. Then we move on to concepts and techniques used to measure poverty by sociologists and other social scientists in both national and cross-national contexts. In all circumstances, we focus on sociologists' contributions to understanding the nature, causes, and consequences of poverty.

Sociologists, Poverty Theory, and Poverty Measurement – Some History

Sociologists have traditionally been more interested in the explanation of poverty than in its measurement, though almost all sociologists believe that poverty statistics are meaningful social indicators of basic needs (Piachaud, 1987; Townsend, 1979; Ringen, 1985; Brady, 2003a). In general, sociologists are critical of the economist's perspective of choice models whereby individuals control their own destiny and are therefore often the cause of their own poverty (Piachaud, 1987). Some such as Gans (1972) even feel that poverty has a functional role to play in capitalist society. In any case, social scientists need to first establish the breadth and depth of poverty in order to analyze it and explore its ultimate causes and remedies. The 'whats' of poverty measurement are fairly well accepted by all social scientists and are covered below. But the reasons why we have poverty, why it persists in some places and not in others, and its consequences are more controversial and here sociologists have played a prominent role.

For instance, consider 'social stratification,' a categorization of people into rankings of socioeconomic tiers based on factors like wealth, income, social status, occupation, and power. Different theories of sociology lead to different explanations for poverty. Structural-functionalists maintain that stratification and inequality are inevitable and beneficial to society at large, or at least to the well-off in society (Davis and Moore, 1945; Rank, 2011). The role of social stratification in poverty has been studied by Massey (2008) and popularized by others such as Beeghley (2007) and Eitzen and Eitzen Smith (2008). Conflict theorists argue that stratification is dysfunctional and harmful to society, and that it results in competition between the rich and the poor as individuals act for their own economic advantage (Kerbo, 2012). Some believe that poverty and inequality are created and maintained by people trying to regulate or control other individuals (Piven and Cloward, 1993).

Such differences of opinion also spread over into sociological interest centered around the ideas of the 'culture' of poverty and the effects of 'place' on poverty – which are often measured differently and are therefore hard to compare. The urban/central city/ghetto aspects of this issue are discussed in the work of Rainwater (1970), Wilson (1987), Anderson (1999), Massey and Denton (1993), and Massey (1996). Harrington's classic work, *The Other America*, covers rural deprivation as well (Harrington, 1962, 1981). Roles of culture, power, social structure, and other factors largely out of control of the individual are the main forces which sociologists use to explain poverty. The basic working hypothesis is that individuals are strongly influenced by the physical and cultural context in which they live. 'Neighborhood' exerts a strong influence on behavior and concentrated poverty in central city ghettos and therefore has a strong negative effect on future life chances and long-run deprivation.

Sociologists and demographers also have importantly pioneered the effects of family structure on poverty, especially its effects on children and social mobility. Most prominent here is

the work of McLanahan (1985, 2004, 2009), McLanahan and Sandefur (1994), and McLanahan and Percheski (2008). Eggebeen and Lichter (1991) and Brown (2004) have also made important contributions here. Some of this work spills over into the realm of parenting and social class per se (Kohn, 1963; Lareau, 2002), as well as the environment of child poverty (Evans, 2004). Closely related is social mobility out of poverty where works by Vaisey (2010), Vermeulen (2000), and Young (2006) have become standard references as far as mobility up from poverty is related to culture and race. In contrast, the structural limits to cultural perspectives are also emphasized by Brady (2003b) and Rank et al. (2003).

In one particularly important area of stratification, that is related to criminal justice, sociologists are the leaders in research on the effects of incarceration on poverty and disconnection from work, public income support, and effects on the family. Most recently, a spate of articles and books on poor younger men and issues related to poverty and incarceration, jobs, social safety nets, and minority men have appeared as we begin to finally address the serious negative effects of prison on neighborhoods, families, and life after prison. Perhaps the most appropriately titled contribution is "Collateral Costs" (Western and Pettit, 2010). But then see also Goffman (2014), Smeeding et al. (2011), Harding et al. (2014), Morenoff and Harding (forthcoming), Wyse et al. (2014), Harding (2010), Wakefield and Wildeman (2013), Western and Wildeman (2009), Wildeman (2009), Western (2006, 2010), Pettit (2012), Pager (2007), and Manza and Uggen (2006), all of whom are among the most prominent researchers in this arena.

Resources other than access to material goods might be important determinants of poverty. The literature on the 'time poor' has grown in recent years after a nascent period. The argument is that time is often the scarcest commodity in market work-oriented societies. Further, time spent on home production is often undervalued or not valued at all. The gender bias in time use is therefore at the heart of these types of analyses (Vickery, 1977). In particular, research has focused on the time crunch on single parents who must both earn a living in the labor market and care for their children (England and Folbre, 1999; Folbre, 1994; Shipler, 2005).

Concepts of Poverty and Techniques of Poverty Measurement

The rest of our discussion is framed by **Figure 1**, which reviews most of the possibilities of poverty concepts and measures. The conceptual underpinnings of poverty measures come from sociology (Grusky and Kanbur, 2006), economics (Lampman, 1964), social statistics (Orshansky, 1965), and social policy (Rowntree, 1901; Booth, 1903). All of the social sciences use or refer to objective and quantitative poverty measurement, usually but not always using a single dimension of 'resources,' income, and several notions of 'needs' standards: those that are relative, absolute, and, closely related, anchored poverty lines (e.g., see Iceland, 2005). We often choose income poverty because of its domination in modern (post-1960s) poverty studies and because of its linkages to the income inequality literature that follows. But wealth poverty is also commonly addressed by sociologists (Shapiro, 2004), as well as family budget models, related to making ends meet as pioneered by Stack (1978) and, more recently, by Edin and Lien (1997) and Halpern-Meekin et al. (2015).

Income or 'living-standards' poverty measurement began in the Anglo-Saxon countries and dates back at least to Rowntree (1901), who was the first to employ the concept of a poverty line in his empirical work on York, England. And thanks to his enterprise, and that of Booth (1903), who invented the idea of a poverty line for London, we have a meaningful social indicator of basic needs (see for instance, Townsend, 1979, 1993; Piachaud, 1987; Ringen, 1985; Ravallion, 2014). We also note that official poverty measurement began as an Anglo-American social indicator. But since then, 'official' measures of poverty (or measures of 'low income') now exist in over 100 countries and for Europe as a whole (Eurostat, 2005). The United States (DeNavas-Walt et al., 2012; Short, 2013) and the United Kingdom (Department for Work and Pensions, 2012) have

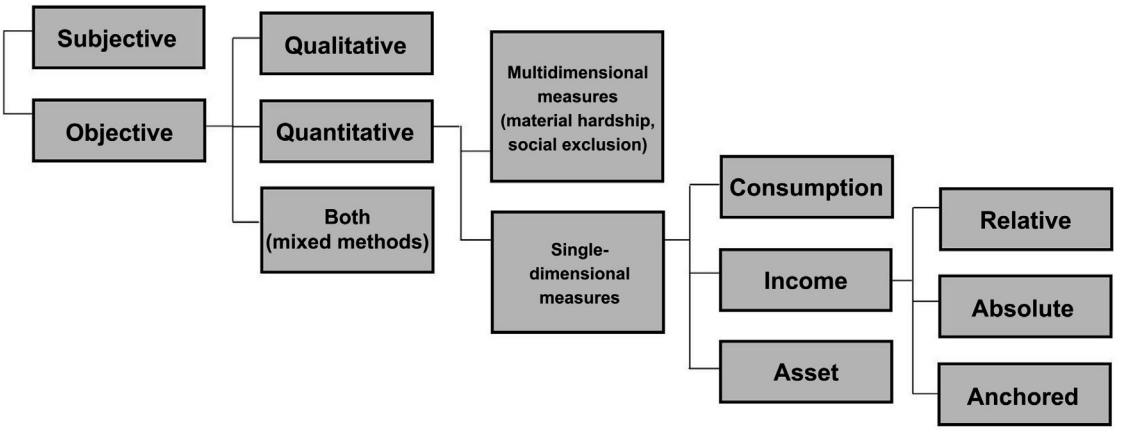

Figure 1 Concepts and measures of poverty. Source: Adapted from Dhongde, S., 20 March 2013. Measuring Multidimensional Poverty in the U.S. Presentation at Institute for Research on Poverty, University of Wisconsin–Madison, Madison, WI; Smeeding, T.M., 2014. Poverty measurement. In: Brady, D., Burton, L. (Eds.), The Oxford Handbook of Poverty and Society. Oxford University Press.

long standing 'official' poverty series. Statistics Canada publishes the number of households with incomes below a 'low-income cutoff' on an irregular basis, as does the Australian government with those below the 'Henderson line.' In Northern Europe and Scandinavia, the debate centers instead on the level of income at which minimum benefits for social programs should be set. In other words, their concept of insufficient 'low income' is directly fed into programmatic responses to social needs (Björklund and Freeman, 1997; Marx and Nelson, 2013; Ravallion, 2014). In addition to these 'objective' poverty measures, several have used 'subjective' measures of poverty and well-being, including income sufficiency (Ravallion, 2014; Van Praag, 1968; Groedhart et al., 1977; Hagenaars and van Praag, 1985; Clark and D'Ambrosio, 2014).

Figure 1 suggests that objective studies of poverty can be classified as quantitative, qualitative (ethnographic), or both. And in the latter two areas, sociologists have made the major contributions to the field. The classic ethnographic studies of poor neighborhoods (both urban and rural), jobs, street corners, and the culture of poverty were pioneered by Harrington (1981), Leibow (1967), Lewis (1969), Wilson (2010), Patterson (2001), and Bourgois (2001). More recently, new, younger sociologists have taken the field to new levels, including Harding et al. (2010), Goffman (2014), often looking specifically at race, ethnicity, and immigration issues, as in Linn and Harris (2008). And ethnographies of low-wage work such as those by Shipler (2005) and Ehrenreich (2001) show how fragile is the edge between poverty and utter destitution for the working poor.

Perhaps, the most powerful approaches to poverty measurement are situations where both qualitative and quantitative works are combined. This type of 'mixed methods' research is done by sociologists in both rich countries (Edin and Lien, 1997; Halpern-Meekin et al., 2015; Tach and Halpern-Meekin, 2014) and poor countries (Carvalho and White, 1997). This research brings both the richness of qualitative descriptions of life in poverty and a large enough sample of budgets or quantitative measures to add some structure and depth to the ethnographies and personal vignettes that support the numbers.

Most measures of quantitative poverty are single dimensional. But concern about the concept of multidimensional poverty is also flourishing in sociology and economics. Official measures of material deprivation, and material hardship exist mainly in Europe, where the European term for a similarly disaffected population is a measure of "social exclusion" (Room, 1999; Hills, 1999; Glennerster et al., 1999). Europe adopted the official Laeken set of social indicators to measure social exclusion in 1995, including the at-risk-of-poverty indicator, with an explicit objective of reducing poverty and social exclusion (Marlier et al., 2007). Indeed, indicators of material deprivation now form part of the Europe 2020 target of poverty reduction (Atkinson and Marlier, 2010: Chapter 6). The study of material hardship in the United States has attracted some attention by sociologists (e.g., see Heflin et al., 2011; Mayer and Jencks, 1989). And it is also related to the so-called 'underclass' which goes beyond poverty alone to include all persons with dysfunctional behaviors living in 'bad' neighborhoods (Mincy et al., 1990; Jencks and Peterson, 1990).

Because there is no single commonly accepted way to measure poverty among social scientists, there is a desire to go beyond the popularly used income poverty definition as a single dimension measure (**Figure 1**), and so there exists a wide variety of additional poverty measures which substitute for or complement the preponderance of income-based measures used by quantitative sociologists and economists (e.g., see Haveman, 2009; Ruggles, 1990; Boltvinik, 2000).

In principle, if poverty is a multidimensional concept, it should reflect several aspects of personal well-being, not a single one, should it be income, consumption, or wealth, as shown in **Figure 1**. Forms of deprivation other than these can certainly be relevant to poverty measurement and to anti-poverty policymaking. A number of authors have suggested that separate measures of needs ought to be developed for different goods and services such as health care or education (Aaron, 1985). Housing and health care are most often mentioned in this context, though the latter is particularly of interest in medically unequal nations such as the United States, while the former is of much greater interest in the United Kingdom (UK, 1993).

Approaches to Poverty Measurement

While poverty measurement is an exercise that is particularly popular in the English-speaking countries, most rich nations share the Anglo-Saxon concern over distributional outcomes and the well-being of the low-income population. There is no international consensus on guidelines for measuring poverty (Townsend, 1993), but international bodies such as the United Nations Children's Fund (UNICEF, 2000), the United Nations Development Programme (UNDP, 1999), the Organization for Economic Cooperation and Development (OECD, 2008, 2013), and the European Statistical Office (Eurostat, 1998, 2005), have published several cross-national studies of the incidence of poverty in rich countries. A large majority of these studies by sociologists as well as economists have been based on the Luxembourg Income Study (LIS) database which can be accessed at www.lisdatacenter.org. Some examples of these studies include Brady (2003a, 2003b, 2005), Brady et al. (2010), Förster (1993), Jäntti and Danziger (2000), Smeeding et al. (2000), Kenworthy (1999), Smeeding et al. (1990), and Smeeding (2006). More recently, the European Union (2005) and the OECD have regularized measurement of poverty by using different standards and data sources. Today one can also find harmonized measures from the World Bank that use both secondary (published) data- and micro data-based measures of consumption and income to determine those living below some particular amount of income per person per day, from $1.25 to $2.00 (Ravallion and Chen, 2011; Chen and Ravallion, 2012).

Most broadly, the quantitative single dimension measurement of poverty in rich, poor, and middle-income nations involves the comparison of some index of household well-being, like income, with household needs. When command over economic resources falls short of needs a household (or person or family) is classified as poor. Here well-being is economic, referring to the material resources available to a household. The concern with these resources among most

social scientists is generally not with material consumption alone, but also with the capabilities such resources give to household members so they can participate fully in society (Sen, 1983, 1992; Brandolini et al., 2010). These capabilities are inputs to social activities, and participation in social activities in return gives rise to a particular level of well-being as argued by sociologists like Rainwater (1990) and Coleman and Rainwater (1978). Methods for measuring a person's or household's capabilities differ according to the context in which one assesses them, either over time or across nations or among subpopulations within a nation, e.g., rural versus urban India.

As most sociologists agree, all advanced societies are highly stratified and, hence, some individuals have more resources than others. The opportunities for social participation are affected by the resources that a household disposes. These resources, the "family income package" (Rainwater et al., 1986; Rainwater and Smeeding, 2003), come from personal effort (earnings), family efforts, and others outside the household, principally governments but also other third parties like nonprofits. Money income is therefore a crucial resource. Of course, there are other important kinds of resources considered by sociologists, such as social capital, wealth, noncash benefits, primary education, and access to basic health care, all of which add to human capabilities (Coleman, 1988). These resources may be available more or less equally to all people in some societies, regardless of their money incomes. But there are also many forces in rich societies which reduce well-being by limiting capabilities for full participation in society, including inadequacies in neighborhoods where people live, racial and ethnic discrimination, neighborhood violence, low-quality public schools and other social services, lack of good jobs, and job instability. These factors increase economic insecurity, reduce human capabilities, and increase poverty, as dozens of analysts, including those mentioned above, have established.

Both consumption poverty and asset poverty have been proposed as an alternative to income poverty in rich nations (Meyer and Sullivan, 2012a, 2012b; Brandolini et al., 2010). And in a few nations, asset and income poverty can be combined into a joint measure (Gornick et al., 2009) as can consumption and income poverty (Meyer and Sullivan, 2012a, 2012b). But consumption and asset poverty measures are not yet ready for widespread use on a cross-national basis, despite their usefulness for some types of poverty measurement (e.g., many income-poor elderly consume more than their incomes due to dissaving and spending from assets).

Absolute, Relative, and Anchored Poverty in Rich and Selected Middle-Income Countries

Having settled on an index of resources, like income, one also needs a measure of need, or a poverty line to complete the objective, quantitative poverty measure in **Figure 1**. Economists and sociologists are often at odds on whether poverty is to be an absolute or relative measure. An absolute poverty standard is defined in terms of a level of purchasing power that is sufficient to buy a fixed bundle of basic necessities at a specific point in time. A relative standard, on the other hand, is defined relative to the typical income or consumption level in the wider society. The purchasing power of a relative poverty standard will change over time as society-wide income or consumption levels change, while an absolute poverty standard will change only with the prices of commodities it can buy. Most cross-national comparisons use the relative definition of poverty, especially since purchasing power parities to convert any absolute measure to country currency are subject to fluctuation and sometimes severe measurement error (Jäntti and Danziger, 2000), and most sociologists believe poverty ought to be measured relative to the time, place, and the standard of living where people live (Rainwater and Smeeding, 2003; Brady, 2003a).

And, in the broadest sense, all measures of poverty or economic needs are relative, because context is important to the definition of even 'absolute' needs. The World Bank uses poverty measures of $1.25 to $2 per person per day – or $1095 to $2190 per year for a family of three – for the developing nations of Africa, Central Asia, or Latin America (Chen and Ravallion, 2012). The measure makes no sense for a rich country today, but it might have served as poverty line for the United States set in the early 1900s. In contrast, the 2011 US 'absolute' poverty threshold was about $18 000 for a family of three – 8 to 17 times the World Bank's poverty line (depending on family size). And absolute and relative may also differ depending on the lower part of the income distribution in the country of interest. For instance, one-half of median income, the preferred relative poverty standard in the United States, is another 25–30% above the official US poverty line or 10–21 times the absolute poverty standard in poor countries. Moreover, as economic inequality has increased in most rich societies over the past 20 years, the study of relative deprivation and poverty has taken on a new life (Gottschalk and Smeeding, 2000; Gornick and Jäntti, 2013; OECD, 2011, 2013).

Cross-national comparisons of poverty in rich countries therefore rely heavily on openly and directly relative concepts of poverty, which are a reflection of the fact that a poverty standard or a minimum income standard ought to reflect the overall standard of living in society. One early source of this formalization (Abel-Smith and Townsend, 1965) came about in arguing that the officially defined minimum level of income in the United Kingdom, as represented by the National Assistance scale, should increase with the rising standard of living, and not just with consumer prices. It was Townsend's work in the early 1960s culminating in his famous 1979 book that really launched the relative poverty approach on a much wider scale.

As Townsend (1979: 31) wrote

> Individuals, families and groups in the population can be said to be in poverty when they lack the resources to obtain the type of diet, participate in the activities and have the living conditions and the amenities which are customary, or at least widely encouraged or approved in the societies to which they belong.

The measurement of relative poverty has more recently been generally operationalized with a definition of the poverty

line as a fraction of median income. Cross-national studies typically compare the percentage of persons living with income below some fraction of the family-size-adjusted national median income.

Measurement of relative poverty in the United States also began in the 1960s and was pioneered by Fuchs (1967), who followed the thinking of Townsend and Abel Smith and linked relative and absolute income poverty measurement. When Fuchs began his study, the absolute poverty measure in the United States begun by Lampman (1964) and then Orshansky (1965) was based on a poverty line of about $3000 for four persons. Fuchs pointed out that this was half of median income at that time and that one could think differently about relative poverty compared to absolute poverty (Gilbert, 2008: 136).

A relative poverty measure comparison is also consistent with a well-established theoretical perspective on poverty (Sen, 1983, 1992; and again, Townsend, 1979). However, the fraction of income at which the poverty line ought to be set is open to debate. Most cross-national studies (LIS, OECD) focus on half of the median income, following Fuchs and others. But many feel that a 50% of median standard is too low. It implies a poverty cutoff well below half the mean in unequal societies, and it also affects the country rankings. The European Statistical Office Working Group on Poverty Measurement has employed 60% of the national median income as the common poverty threshold for European Community poverty studies in the new millennium (Eurostat, 2005).

It is worthwhile to understand both absolute *and* relative poverty measures because they tell different things about absolute living standards as well as relative deprivation. Increasingly, the idea of 'anchored' poverty measures have become important and can be employed to indicate both relative and absolute poverty trends within a given nation. Anchored poverty measures begin with the same relative measure in 1 year (t) and then compare relative poverty in some future year (say year $t + 10$) to poverty measures against a poverty line that has been changed only for prices between year t and year $t + 10$. These measures are especially useful in periods of rapid expansion *or* contraction in an economy, where relative poverty may not change by a lot, but where absolute poverty does change due to economic growth or contraction (see Atkinson et al., 2002; Smeeding, 2006; Johnson and Smeeding, 2012; OECD, 2013). Any absolute poverty line is also therefore an anchored poverty line. The difference is that an anchored poverty line can be updated to any period which is policy relevant, given the analysis.

Summary

Sociologists and other social scientists have used a variety of approaches in explaining and describing poverty, as well as methods to define and measure poverty. Each approach has its strengths and weaknesses. More debated are the sociological theories of why poverty exists, and its relation to family, culture, and context.

Concepts and measures of poverty abound. Most quantitative poverty measures use income as a measure of resources, compared to some needs standard or poverty line. Both absolute and relative poverty line measurements tell us something about changes in living standards for low-income people. In order to measure both relative and absolute progress in countries, analysts increasingly use the idea of anchored poverty, whereby one fixes a poverty line within a country at the relative value in a given year and prices that forward to the most recent year. When one uses these concepts across a range of rich nations, one can trace out both changes in real or absolute living standards for low-income people, as well as relative progress or regress compared to the middle person in society.

Sociologists have also provided important tools for testing causal hypotheses concerning the explanations, causes, and cures for poverty regardless of how it is measured. These include place, culture, stratification, space, race, gender, imprisonment, and other popular theories of poverty in sociology. They are complemented by social indicator-based measures of inadequacy of time for market work, home work, and the nurturing and care of family members. Sociological studies of poverty have both a rich heritage and have become an important area of research in increasingly unequal rich and poor societies.

See also: Climate Change and Global Poverty; Culture of Poverty: Critique; Earnings Inequality; Income Distribution and Inequality: Measurement Issues; Income Inequality; Inequality Measurement; Inequality, Social; Poverty Law: United States; Poverty Policy; Poverty in History; Poverty, Culture of; Poverty, Geography of; Poverty: Measurement and Analysis; Social Inequality in History (Stratification and Classes); Spatial Thinking in the Social Sciences; Urban Poverty in Neighborhoods.

Bibliography

Aaron Henry, J., 1985. Comment: The evaluation of Census Bureau procedures for the measurement of noncash benefits and the incidence of poverty, Proceedings of the Bureau of the Census Conference on the Measurement of Noncash Benefits. US Department of Commerce, Washington, DC. 57–62.

Abel-Smith, B., Townsend, P., 1965. The Poor and the Poorest. Occasional Papers in Social Administration 17. Bell, London.

Anderson, E., 1999. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. W. W. Norton, New York.

Atkinson, A.B., Cantillon, B., Marlier, E., Nolan, B., 2002. Social Indicators: The EU and Social Inclusion. Oxford University Press, Oxford.

Atkinson, A.B., Marlier, E., 2010. Income and Living Conditions in Europe, Eurostat At: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-31-10-555/EN/KS-31-10-555-EN.PDF.

Beeghley, L., 2007. The Structure of Social Stratification in the United States, fifth ed. Pearson Publishing, London.

Björklund, A., Freeman, R.B., 1997. Generating equality and eliminating poverty – the Swedish way. In: Freeman, R.B., Topel, R., Swedenborg, B. (Eds.), The Welfare State in Transition: Reforming the Swedish Model. University of Chicago Press, Chicago, pp. 33–78.

Blank, R., 2008. How to improve poverty measurement in the United States. Journal of Public Analysis and Management 27 (2), 233–254.

Boltvinik, J., 2000. Poverty Measurement and Trends, United Nations Development Programme, SEPED Series on Poverty Reduction. United Nations, New York. www.undp.org/poverty/publications/pov_red.

Booth, C., 1903. Life and Labour of the People of London Second Series, Industry, vol. 5. Macmillan, London.

Bourgois, P., 2001. Culture of poverty. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Pergamon, Oxford, pp. 11904–11907.

Brady, D., Fullerton, A., Moren Cross, J., 2010. More than just nickels and dimes: a cross-national analysis of working poverty in 18 affluent democracies. Social Problems 57, 559–585.

Brady, D., 2003a. Rethinking the sociological measurement of poverty. Social Forces 81 (3), 715–752.

Brady, D., 2003b. The politics of poverty: left political institutions, the welfare state, and poverty. Social Forces 82, 557–588.

Brady, D., 2005. The welfare state and relative poverty in affluent western democracies, 1967–1997. Social Forces 83, 1329–1364.

Brandolini, A., Magri, S., Smeeding, T.M., 2010. Asset-based measurement of poverty. Journal of Policy Analysis and Management 29 (2), 267–284.

Brown, S., 2004. Family structure and child well-being: the significance of parental cohabitation. Journal of Marriage and Family 66, 351–367.

Carvalho, S., White, H., 1997. Combining the Quantitative and Qualitative Approaches to Poverty Measurement and Analysis: The Practice and the Potential. World Bank, Washington, DC.

Chen, S., Ravallion, M., 2012. Absolute Poverty Measures for the Developing World. Chapter 20 in Measuring the Real Size of the World Economy. World Bank, Washington, DC.

Clark, A., D'Ambrosio, C., 2014. Attitudes to income inequality: experimental and survey evidence (Chapter 14). In: Atkinson, A.B., Bourguignon, F. (Eds.), 2014. Handbook of Income Distribution, vol. 2A. Elsevier.

Coleman, J., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Coleman, R.P., Rainwater, L., 1978. Social Standing in America: New Dimensions of Class. Basic Books, New York.

Davis, K., Moore, W., 1945. Some principles of stratification. American Sociological Review 10, 242–249.

DeNavas-Walt, C., Proctor, B.D., Smith, J.C., 2012. Income, Poverty, and Health Insurance: 2011. Current Population Reports, P60–243. U.S. Census Bureau. At: http://www.census.gov/prod/2012pubs/p60-243.pdf.

Department for Work and Pensions (UK), 2012. Households below Average Income Available at: http://research.dwp.gov.uk/asd/index.php?page=hbai.

Dhongde, S., 20 March 2013. Measuring Multidimensional Poverty in the U.S. Presentation at Institute for Research on Poverty, University of Wisconsin–Madison, Madison, WI.

Edin, K., Lien, L., 1997. Making Ends Meet: How Single Mothers Survive Welfare and Low-Wage Work. Russell Sage Foundation, New York.

Edin, K., Tach, L., Halpern-Meekin, S., 2014. Tax Code Knowledge and Behavioral Responses Among EITC Recipients: Policy Insights from Qualitative Data. Journal of Policy Analysis and Management 33 (2), 413–439.

Eggebeen, D., Lichter, D., 1991. Race, family structure, and changing poverty among American children. American Sociological Review 56 (6), 801–817.

Ehrenreich, B., 2001. Nickel and Dimed: On (Not) Getting by in America. Metropolitan Books, New York.

Eitzen, D.S., Eitzen Smith, K., 2008. Experiencing Poverty: Voices from the Bottom. Allyn & Bacon, New York.

England, P., Folbre, N., January 1999. The cost of caring. Annals of the American Academy of Political and Social Science 561, 39–51.

European Union, 2005. Income Poverty and Social Exclusion in the EU25. Statistics in Focus. Official Publications of the European Communities. Eurostat, Luxembourg.

Eurostat, 1998. Recommendations of the Task Force on Statistics on Social Exclusion and Poverty. European Statistical Office, October, Luxembourg.

Eurostat, 2005. Income Poverty and Social Exclusion in the EU25. Statistics in Focus. Official Publications of the European Communities, Luxembourg.

Evans, G.W., 2004. The environment of childhood poverty. American Psychologist 59 (2), 77–92.

Folbre, N., 1994. Who Pays for the Kids? Gender and the Structures of Constraint. Routledge, London.

Förster, M., 1993. Comparing Poverty in 13 OECD Countries: Traditional and Synthetic Approaches. Luxembourg Income Study Working Paper Series.

Fuchs, V.R., 1967. Redefining poverty and redistributing income. The Public Interest 8, 88–95.

Gans, H.J., 1972. The positive functions of poverty. American Journal of Sociology 78 (2), 275–289.

Gilbert, G., 2008. Rich and Poor in America: A Reference Handbook. ABC-CLIO Press, New York.

Glennerster, H., Lupton, R., Noden, P., Power, A., 1999. Poverty, Social Exclusion, and Neighborhood: Studying the Area Bases of Social Exclusion. CASE Paper 22. Centre for the Analysis of Social Exclusion. London School of Economics, London, UK.

Goffman, A., 2014. On the Run: Fugitive Life in an American City. University of Chicago Press, Chicago.

Gornick, J.C., Jäntti, M., 2013. Income Inequality: Economic Disparities and the Middle Class in Affluent Countries. Stanford University Press, Stanford.

Gornick, J.C., Sierminska, E., Smeeding, T.M., 2009. The income and wealth packages of older women in cross-national perspective. The Journals of Gerontology: Psychological Sciences 64B (3), 402–414.

Gottschalk, P., Smeeding, T.M., 2000. Empirical evidence on income inequality in industrialized countries. In: Atkinson, A.B., Bourgignon, F. (Eds.), Handbook of Income Distribution. Elsevier-North Holland Publishers, New York, pp. 261–308.

Groedhart, T., Halberstadt, V., Kapteyn, A., Van Praag, B., 1977. The poverty line: concept and measurement. Journal of Human Resources 12, 503–520.

Grusky, D.B., Kanbur, R., 2006. Introduction: the conceptual foundations of poverty and inequality measurement. In: Grusky, D., Kanbur, R. (Eds.), Poverty and Inequality. Studies in Social Inequality. Stanford University Press, Stanford, pp. 1–29.

Harding, D.J., Morenoff, J.D., Herbert, C., 2013. Home is hard to find: neighborhoods, institutions, and the residential trajectories of returning prisoners. Annals of the American Academy of Political and Social Science 647, 214–236.

Harding, D.J., Wyse, J.J.B., Dobson, C., Morenoff, J.D., 2014. Making ends meet after prison. Journal of Policy Analysis and Management. http://dx.doi.org/10.1002/pam.21741.

Harding, D.J., Lamont, M., Small, M.L., 2010. Reconsidering culture and poverty. Annals of the American Academy of Political and Social Science 629, 6–27.

Harding, D.J., 2010. Living the Drama: Community, Conflict, and Culture Among Inner-city Boys. University of Chicago Press, Chicago.

Hagenaars, A.J.M., van Praag, B., 1985. A synthesis of poverty line definitions. Review of Income and Wealth 31 (2), 139–154.

Halpern-Meekin, S., Edin, K., Tach, L., Sykes McLaughlin, J., 2015. It's Not like I'm Poor: The Financial Lives of the Working Poor. University of California Press, Berkley.

Harrington, M., 1962. The Other America: Poverty in the United States. Touchstone Press (A Division of Simon & Schuster), New York, pp. 1–19, 158–174.

Harrington, M., 1981. The Other America: Poverty in the United States, second Exp. ed. Penguin, Harmondsworth, Middlesex.

Haveman, R., 2009. What does it mean to be poor in a rich society? In: Cancian, M., Danziger, S. (Eds.), Changing Poverty, Changing Policies. Russell Sage Foundation, New York, pp. 387–408.

Heflin, C., London, A.S., Scott, E.K., 2011. Mitigating material hardship: the strategies low-income families employ to reduce the consequences of poverty. Sociological Inquiry 81 (2), 223–246.

Hills, J., 1999. Social Exclusion, Income Dynamics, and Public Policy. Annual Sir Charles Carter Lecture, Northern Ireland Economic Council, Report No. 129. NIEC, Belfast.

Iceland, J., 2005. Measuring poverty: theoretical and empirical considerations. Measurement 3 (4), 199–235.

Jäntti, M., Danziger, S., 2000. Income poverty in advanced countries. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution. North-Holland, Amsterdam, pp. 309–378.

Jencks, C., Peterson, P.E., 1990. The urban underclass and the poverty paradox. In: Jencks, C., Peterson, P.E. (Eds.), The Urban Underclass. The Brookings Institution, Washington, DC, pp. 3–27.

Johnson, D.S., Smeeding, T.M., 2012. A Consumer's Guide to Interpreting Various U.S. Poverty Measures. Fast Focus 14-2012. Institute for Research on Poverty, Madison, WI. Available at: http://www.irp.wisc.edu/publications/fastfocus.htm.

Kenworthy, L., 1999. Do social welfare policies reduce poverty? A cross-national assessment. Social Forces 77, 1119–1139.

Kerbo, H.R., 2012. Social Stratification and Inequality. McGraw-Hill, New York.

Kohn, M., 1963. Social class and parent-child relationships: an interpretation. American Journal of Sociology 68 (4), 471–480.

Lampman, R., 1964. The problem of poverty in America. Economic Report of the President, 1964. U.S. Government Printing Office, Washington, DC, pp. 55–79.

Lareau, A., 2002. Invisible inequality: social class and childrearing in black families and white families. American Sociological Review 67, 747–776.

Leibow, E., 1967. Tally's Corner: A Study of Negro Streetcorner Men. Rowman and Littlefield, Lanham, Maryland.

Lewis, O., 1969. The culture of poverty. In: Moynihan, D.P. (Ed.), On Understanding Poverty: Perspectives from the Social Sciences. Basic Books, New York, pp. 187–220.

Linn, A., Harris, D., 2008. The Colors of Poverty: Why Racial and Ethnic Disparities Persist. Russell Sage, New York.

Morenoff, J.D., Harding, D.J. Prisoner reentry and communities. Annual Review of Sociology, forthcoming.

Manza, J., Uggen, C., 2006. Locked Out: Felon Disenfranchisement and American Democracy. Oxford University Press, New York.

Marlier, E., Atkinson, A.B., Cantillon, B., Nolan, B., 2007. The EU and social inclusion: facing the challenges. Journal of Common Market Studies 45 (2), 518–533.

Marx, I., Nelson, K., 2013. Minimum Income Protection in Flux. Palgrave Macmillan, Basingstoke.
Massey, D., 2008. Categorically Unequal: The American Stratification System. Russell Sage, New York.
Massey, D., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge.
Massey, D., 1996. The age of extremes: concentrated poverty and affluence in the 21st century. Demography 33 (4), 395–412.
Mayer, S.E., Jencks, C., 1989. Poverty and the distribution of material hardship. Journal of Human Resources 24 (1), 88–113.
Mayer, S.E., Jencks, C., 1993. Recent trends in economic inequality in the United States: income versus expenditures versus material well-being. In: Papadimitriou, D.B., Wolff, E.N. (Eds.), Poverty and Prosperity in the USA in the Twentieth Century. St. Martin's Press, New York.
McLanahan, S., Percheski, C., 2008. Family structure and the reproduction of inequalities. Annual Review of Sociology 34, 257–276.
McLanahan, S., Sandefur, G., 1994. Growing Up with a Single Parent: What Helps, What Hurts. Harvard University Press, Cambridge.
McLanahan, S., 1985. Family structure and the reproduction of poverty. American Journal of Sociology 90 (4), 873–901.
McLanahan, S., 2004. Diverging destinies: how children fare under the second demographic transition. Demography 41 (4), 607–627.
McLanahan, S., 2009. Fragile families and the reproduction of poverty. Annals of the American Academy of Political and Social Science 621, 111–131.
Meyer, B.D., Sullivan, J.X., 2012a. Identifying the disadvantaged: official poverty, consumption poverty, and the new supplemental poverty measure. Journal of Economic Perspectives 26 (3), 111–136.
Meyer, B.D., Sullivan, J.X., 2012b. Winning the war: poverty from the great society to the great recession. Brookings Papers on Economic Activity, Economic Studies Program. The Brookings Institution 45 (2), 133–200.
Mincy, R., Sawhill, I., Wolf, D., 1990. The underclass: definition and measurement. Science 248, 450–453.
OECD, 2008. Growing Unequal? Income Distribution and Poverty in OECD Countries. OECD Publishing, Paris.
OECD, 2011. Divided We Stand: Why Inequality Keeps Rising. OECD Publishing, Paris.
OECD, 2013. Crisis Squeezes Income Puts Pressure on Inequality and Poverty. OECD Publishing, Paris.
Orshansky, M., 1965. Counting the poor: another look at the poverty profile. Social Security Bulletin 28 (1), 3–29.
Pager, D., 2007. Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration. University of Chicago Press, Chicago.
Patterson, O., 2001. Taking culture seriously: a framework and an Afro-American illustration. In: Harrison, L.E., Huntington, S.P. (Eds.), Culture Matters: How Values Shape Human Progress. Basic Books, New York, pp. 202–218.
Pettit, B., Western, B., 2004. Mass imprisonment and the life course: race and class inequality in U.S. incarceration. American Sociological Review 69, 151–169.
Pettit, B., 2012. Invisible Men: Mass Incarceration and the Myth of Black Economic Progress. Russell Sage Foundation,, New York.
Piachaud, D., 1987. Problems in the definition and measurement of poverty. Journal of Social Policy 16 (2), 147–164.
Piven, F.F., Cloward, R.A., 1993. Regulating the Poor. Vintage Books, New York.
Van Praag, B., 1968. Individual Welfare Functions and Consumer Behavior. North-Holland, Amsterdam.
Rainwater, L., 1970. Behind Ghetto Walls. Pelican Books, New York.
Rainwater, L., Rein, M., Schwartz, J.E., 1986. Income Packaging in the Welfare State: A Comparative Study of Family Income. Oxford University Press.
Rainwater, L., 1990. Poverty and Equivalence as Social Constructions. Luxembourg Income Study Working Paper. 55, at http://www.//lisdatacenter.org/wps/liswps/55.pdf.
Rainwater, L., Smeeding, T.M., 2003. Poor Kids in a Rich Country. Russell Sage Foundation, New York.
Rank, M.R., 2011. Rethinking American poverty. Contexts 10, 16–21.
Rank, M.R., Yoon, H.S., Hirschl, T., 2003. American poverty as a structural failing: evidence and arguments. Journal of Sociology and Social Welfare 30 (4), 3–29.
Ravallion, M., Chen, S., 2011. Developing world is poorer than we thought, but no less successful in the fight against poverty. Quarterly Journal of Economics 125 (4), 1577–1625.
Ravallion, M., 2014. The idea of antipoverty policy. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution, vol. 2A. Elsevier, Amsterdam.
Ringen, S., 1985. Toward a third stage in the measurement of poverty. Acta Sociologica 28, 99–113.
Room, G.J., 1999. Social exclusion, solidarity and the challenge of globalization. Inter Social Welfare 8, 166–174.
Rowntree, B.S., 1901. Poverty: A Study of Town Life. Macmillan, London.
Ruggles, P., 1990. Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy. Urban Institute Press, Washington, DC.
Sen, A., 1983. Poor, relatively speaking. Oxford Economic Papers 35, 153–169.
Sen, A., 1992. Inequality Reexamined. Harvard University Press, Cambridge, MA.
Shapiro, T., 2004. The Hidden Cost of Being African American: How Wealth Perpetuates Inequality. Oxford University Press, New York.
Shipler, D.K., 2005. The Working Poor: Invisible in America. Vintage, NY.
Short, K., November 2013. The Research Supplemental Poverty Measure: 2012. Current Population Reports: P60–247. U.S. Census Bureau, Washington, DC. At: https://www.census.gov/prod/2013pubs/p60-247.pdf.
Smeeding, T.M., 2014. Poverty measurement. In: Brady, D., Burton, L. (Eds.), The Oxford Handbook of Poverty and Society. Oxford University Press.
Smeeding, T.M., 2006. Poor people in rich nations: the United States in comparative perspective. Journal of Economic Perspectives 20, 69–90.
Smeeding, T.M., O'Higgins, M., Rainwater, L., 1990. Poverty, Inequality and the Distribution of Income in an International Context: Initial Research from the Luxembourg Income Study (LIS) Project. Wheatsheaf Books, London. Urban Institute Press, Washington, DC.
Smeeding, T.M., Rainwater, L., Burtless, G., August 2000. United States Poverty in a Cross-National Context. Mimeo, Center for Policy Research, Syracuse University, Syracuse, NY.
Smeeding, T.M., Garfinkel, I., Mincy, R. (Eds.), 2011. Young Disadvantaged Men; Fathers, Families, Poverty, and Policy. Annals of the American Academy of Political and Social Science 635 (1).
Stack, C., 1978. All Our Kin. Basic Books, New York.
Townsend, P., 1979. Poverty in the United Kingdom. Penguin, Harmondsworth.
Townsend, P., 1993. The International Analysis of Poverty. Harvester Wheatsheaf, London, UK.
UNICEF Innocenti Research Centre, June 2000. A League Table of Child Poverty in Rich Nations. Innocenti Report Card 1. UNICEF, Florence. At: http://www.unicef-icdc.org.
United Kingdom Department of Social Security, 1993. Households below Average Income. Government Statistical Service, London.
United Nations Development Programme (UNDP), July 1999. Human Development Report. United Nations, New York.
Vaisey, S., 2010. What people want: rethinking poverty, culture, and educational attainment. Annals of the American Academy of Political and Social Science 629 (1), 75–101.
Vermeulen, H., 2000. Introduction: the role of culture in explanations of social mobility. In: Vermeulen, H., Perlmann, J. (Eds.), Immigrants, Schooling, and Social Mobility: Does Culture Make a Difference? St. Martin's Press, New York, pp. 1–22.
Vickery, C., 1977. The time poor: a new look at poverty. Journal of Human Resources 12 (1), 27–48.
Wakefield, S., Wildeman, C., 2013. Children of the Prison Boom: Mass Incarceration and the Future of American Inequality. Oxford University Press, New York.
Western, B., Wildeman, C., 2009. The black family and mass incarceration. Annals of the American Academy of Political and Social Sciences 621, 221–242.
Western, B., 2010. Incarceration and social inequality. Daedalus 139 (3), 8–19.
Western, B., 2006. Punishment and Inequality in America. Russell Sage Foundation, New York.
Western, B., Pettit, B., 2010. Collateral Costs: Incarceration's Effect on Economic Mobility. The Pew Charitable Trusts, Washington, DC.
Wildeman, C., 2009. Parental imprisonment, the prison boom, and the concentration of childhood disadvantage. Demography 46, 265–280.
Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass and Public Policy. University of Chicago Press, Chicago.
Wilson, W.J., 2010. Why both social structure and culture matter in a holistic analysis of inner-city poverty. Annals of the American Academy of Political and Social Science 629 (1), 200–219.
Wyse, J.J.B., Harding, D.J., Morenoff, J.D., 2014. Romantic relationships and criminal desistance: pathways and processes. Sociological Forum 29 (2), 365–385.
Young, A., 2006. The Minds of Marginalized Black Men: Making Sense of Mobility, Opportunity, and Future Life Chances. Princeton University Press, Princeton, NJ, pp. 3–36, 107–136, 199–206.

Power

Andrew S McFarland, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Political power-as-causation is discussed: A causes B to change course against B's initial resistance. Models of power in policy-making such as pluralism, multiple elitism, and neopluralism; domination through socialization and social structure; Foucault's theory of power; and power in international systems and regimes are considered. Power-as-causation is contrasted to power-as-resources.

Leading political theorists, such as Machiavelli and Hobbes, have devoted much attention to 'political power.' As Robert A. Dahl noted (1957), power has a certain universality, whether it is called "power, influence, control, pouvoir, puissance, Macht, Herrschaft, Gewalt, imperium, potestas, auctoritas, potential, etc." This article first focuses on the concept of power defined as a type of social causation.

Power as Causation

During 1953–75, several leading political scientists developed the view of power as social causation, converging with an earlier, similar definition by Max Weber (Gerthe and Mills, 1946, 180). For power, after rejecting alternative definitions, Herbert Simon (1953: 503) cited a definition by Lasswell and Kaplan (1950): "The exercise of influence (influence process) consists in affecting policies of others than the self." In the formulation best known to political scientists, Dahl stated (1957: 202–203): "My intuitive idea of power, then, is something like this: A has power over B to the extent he can get B to do something that A would not otherwise do." James G. March stated (1955: 437): "… the set of all influence relations is here defined to be that subset of all causal relations such that the behavior of an individual appears as the terminal point in the causal linkage. Alternatively we can say that two individuals are in an influence relation if their behaviors are linked causally."

This leads to distinctions among different types of causal relations as they might relate to power. Political power refers to causal relations among humans, not to inanimate forces changing human behavior, as a hurricane might, for instance. Second, a distinction causing much confusion, one needs to distinguish situations in which the changed behavior reflects the intentions of the powerful actor. One might change the behavior of another in a way not intended, as when an incompetent lobbyist changes a legislator's vote to one different from one favored by the lobbyist. Some political scientists, such as March, are more comfortable with power causing changes in the direction intended by the powerful actor, and referring to the more general category of causing any changes as *influence*. On the other hand, the seminal articles by Simon and Dahl do not restrict power to intended changes in others' behavior.

Other examples of such causal relations would include power-as-suggestion: when A proposes a behavior change, B adopts the change without resistance. Another category of causal relations occurs when B resists, and A exercises rewards and sanctions, perhaps even force, to get B to change his/her behavior. This type of power may be called coercion.

In regard to power as coercive causation, John Harsanyi (1962) emphasized costs of exercising power and also the costs of compliance with the exercise of power. For instance, a lobbyist might change a legislator's vote with little cost if it just means arranging a meeting with one of the legislator's staff. On the other hand, the cost is high if the lobbyist must mobilize hundreds of persons to petition the recalcitrant legislator. From the standpoint of the legislator, a vote may be cheap, offending almost no one, or it may endanger victory in the next election. Still another causal situation occurs when A's preferences cause changes in B's preferences, causing B to act differently. This is widely known as *anticipated reactions*, in which B, perhaps to please A and to avoid A's exercise of sanctions, changes behavior in line with an anticipation of what A wants him/her to do, even though A did not explicitly state such preferences (Nagel, 1975).

An important, implicit aspect of defining power as causation is its link to empiricism. A statement of power then implies observations of instances of A changing B's behavior. It rules out defining power in terms of resources (wealth or force, etc.) used to get B to comply. One first has to observe the causal relationship; then one analyzes how power was wielded. Such resources were called by Dahl a power base, and both he and Simon (1953) stated that the power base should not be confused with power itself. In addition, Dahl stated that power-as-causation, to be an empirically useful concept, had to be used with four concomitant dimensions of power: base, amount, scope, and domain. *Base* then refers to the resources or means that A uses to cause changes in others' behavior. *Amount* refers to the idea that some instances of power refer to greater changes in behavior than others. A lobbyist exercises more power when he/she changes the legislator's vote from no to yes, for instance, than if the change is only from abstain to yes. The *domain* of an actor's power consists of those persons subject to an actor's power. The *scope* of power consists of the matters subject to the actor's power.

Dahl's major contribution to the study of political power is his insistence that domain and scope always be clearly stated in discussions of power. Or to put it in another way, an actor's power in one situation should not be immediately generalized to postulating equivalent power over other persons in other

situations. The pharmaceutical lobbyist may have power over just one or over a score of legislators regarding the regulation of generic drugs. The same lobbyist might not have equivalent power over Medicare payments for prescriptions; this lobbyist may have no power at all over any legislators in regard to civil rights or to gun control.

Scope and domain of power were key issues in the still widely remembered *community power debate* of the 1960s. Following the example set by Dahl in *Who Governs?* (1961), political scientists studied political power in localities by conducting case studies of public policy-making and charting which persons changed the course of events. Such studies normally found that the domain and scope of power was fragmented; different persons had power in different areas of policy. Other social scientists – tending to be sociologists – followed a different conception of power, conceiving power to be possession of wealth or high social status. Such observers usually described inequalities of wealth and status and stated this finding as a finding of power in a community. Political scientists, on the other hand, for the most part did not consider wealth and status to be power itself, but instead to be resources that could be used to win political power, if such wealthy or high-status persons intended to do so. But such intentions are an empirical question, in this view, and must be observed in case studies.

This famous (or notorious) controversy extended for at least a decade, was expressed in scores of books and hundreds of journal articles, before the frustrated debate participants, unable to persuade one another, for the most part went on to study questions other than community power (Polsby, 1980).

Proponents of power-as-resource generally overlook a strong criticism of the power-as-causation model. Social causation is ordinarily very complex; case studies are not easy to do; power as causation thus is ordinarily complex, requiring a great of effort to study. Such writers as James March (1955), Herbert Simon (1957), John Harsanyi (1962), and William Riker (1964) left us with elaborate statistical models for studying power, models which require costly and difficult data collection. Jack Nagel (1975) wrote a concluding work in this line of analysis, in which power-as-causation was stated as path analysis, a form of statistics.

Power and Policy-Making

William Riker (1964) and James March (1966) were both noted for asking the question in respect to political analysis: "what is the power of power?" Political scientists answered this question, not by flocking to do more studies of the holders of wealth and status, but by stating models of power structure applied to political systems, both national and local. Robert Dahl (1961) had pressed forward the definition of power-as-causation in the course of his empirical refutation of the power elite model of Floyd Hunter (1953) as applied to the city of Atlanta and the later statement of the elite model by Wright Mills (1956). Such writers were among those who described power by describing characteristics and activities of an elite group, whose actual exercise of power usually went undemonstrated. Scholars such as Dahl became known as describing a *pluralist model of power*, finding power to be fragmented among various scopes and domains, as opposed to the concentration of power within a single elite.

Scholars such as Murray Edelman (1964), Grant McConnell (1966), and Theodore Lowi (1979), on the other hand, agreed that political power in America is fragmented, but argued that stable, unrepresentative elites often control policy within specific scopes of power – such as setting grazing fees on federal lands, developing nuclear power plants, setting airline and interstate trucking schedules and fees before they were deregulated. Such analyses, which flourished briefly, also were based on case studies of policy-making and constitute a school which might be termed *multiple-elitist*, finding not a single power elite but hundreds of elites, each specific to a particular scope of action and not subject to control by elected officials. But multiple elitism in turn faced a revisionist interpretation, *neopluralism*, as found in studies by James Q. Wilson (1980) and Jack Walker (1991), for instance, who found that while power within specific scopes of policy was normally unequal, it was nevertheless shared by so many organized interest groups, politicians, and government agencies that even within narrowly defined policy scopes, stable and controlling elites are seldom found (McFarland, 2004).

The power-as-causation model was developed during 1953–75, and the examples cited in this literature were almost always power among individual persons. But given the great difficulties in gathering such data, political scientists seldom conducted such studies, and instead, the analysis of political power flourished with the development of models of policy-making just cited. Such scholars saw no need to worry about possible criticisms of ignoring the power bases of wealth and status, and continued to conduct case studies of policy-making, following Dahl's example, even as many argued with Dahl's pluralist model. Such case studies were not conducted according to the canons of path analysis or other forms of statistics; instead such researchers used historical materials, interviews, and current documents to ascertain which groups, leaders, and government agencies caused changes in the course of public policy (Baumgartner et al., 2009). Relatively simple empirical methods, assuming power-as-causation, have been enough to generate interesting models of power structure, but perhaps more complex methods will be needed to resolve differences among various schools of political power analysis.

The study of political power has developed satisfactorily since Herbert Simon's seminal article in 1953, but a great theoretical difficulty remains – the problem of *hegemony*. Hegemonic writers stress the great importance of the power to determine which issues are on the political agenda of a political system (Lukes, 1974). Community power theorists Peter Bachrach and Morton Baratz (1962) argued that there are 'two faces of power,' the first face being power over overt activity on issues already considered in the system as studied by Dahl. Still to this face must be added a second face: power over the formulation of political issues yet to be considered by the political system. In arguing for the understanding of hegemony, the Marxist Antonio Gramsci argued (1971) that capitalism maintained its political power through the socialization of persons into supportive values and norms as much as through the machinations of bourgeois government. Social values shape the preferences of the actors in their search for political power, but how can the scholar make empirical

statements about one set of values displacing perhaps hundreds of conceivable alternative value systems? This seems to presuppose a comprehensive theory of social relations.

Hegemony can be distinguished from the power of *anticipated reactions*, power exercised when B anticipates A's intentions, and acts in accordance with them, perhaps fearing the costs of A's overt coercion. Republican members of Congress may have considered blocking appropriations for a government department, but anticipating President Obama's effective reactions, they think better of the idea and do little but discuss the matter. But since the matter received some discussion, the theorist of political power may make an empirical observation.

During the 1980s mainstream political scientists normally accepted the three-point progression in the understanding of political power: (1) power as causation in which person A gets person B to do something A intended, but B initially resisted, that is causing intended changes in another's behavior; (2) adding to the foregoing or the 'first face of power,' the 'second face of power,' control over the appearance of those issues over which persons contended to realize their intentions; and (3) control over the development of individual intentions, even before such intentions become expressed as political issues. Acceptance of the power-as-causation model accompanied an understanding that others, often critics of the social status quo, preferred to refer to power in terms of resources (especially wealth), resources to gain one's way in the exercise of political power.

After accepting this threefold understanding, many political scientists accepted Charles E. Lindblom's structural model of power, which accepted power-as-causation but proclaimed the necessity of understanding the social structures which shaped the issues getting to the political agenda and the political socialization processes that shaped the goals behind such issue definition. Accordingly, in regard to political power in the United States, Lindblom stated 'the privileged position of business' in having the power to shape intentions through advertising, and the power to affect the placement of issues on the political agenda, particularly through the capacity to move jobs from one place to another, or to refuse to invest unless receiving favorable treatment in systems of taxation and government regulation. Lindblom's structural regime theory became a widely accepted, although infrequently explicitly stated assumption, for those conducting empirical studies of political power exhibited in elections or lobbying (Lindblom, 1977).

Such concerns were brought together in the landmark study of lobbying and political power in issues considered by the American national congress. Accepting the power-as-causation model, five political scientists and their research assistants conducted 315 interviews and manuscript background research for 98 case studies to develop narratives of who caused what to happen on 98 separate issues. The study only weakly demonstrated the power of business, because it seemed that the status quo already exhibited business power, and most political power was exhibited in vetoing those seeking to change the status quo, including businesses seeking to expand their power over political issues. From another perspective, this *Lobbying and Policy Change* study demonstrates the great effort needed to gather sufficient data to demonstrate who has power according to the power-as-causation concept. Individual researchers need be aware of this, and lacking the resources to develop many case study narratives, then be prepared to make plausible, evidence-based statements about successful exercise of political power, which might leave some element of uncertainty due to the difficulty of estimating the g complex causation involved in political interactions (Baumgartner et al., 2009).

Foucault and Power

Beginning in the 1990s, the ideas of French social theorist Michel Foucault (1926–84) provoked a great deal of discussion among American social scientists as they had done among the French somewhat earlier. Foucault's views about power differed from those in the causation, limitation by general social structure, or resource perspectives mentioned above. Foucault viewed power as manifested within normative systems controlling human behavior within specific institutional structures developed by processes of modernization. These institutional structures included prisons, factories, schools, universities, hospitals, mental hospitals, military barracks, and charity and philanthropic institutions. Systems of knowledge pertaining to each institution define and comprise it and control human behavior within the institution. Such systems include philosophies of education, professional norms such as those of medicine and law, academic disciplinary norms, criminology, social work and sociological theory, military leadership, and especially psychoanalysis and other systems of psychological theory. Foucault's views even extended to the perpetuation of norms controlling behavior by the family and by sexual interaction (Foucault, 1977).

Opposed to defining power as one agency dominating another, Foucault defined power as control through a system (although he avoided the use of such general, abstract terms). Foucault devoted years to research concerning 'discipline and punish,' the overt use of social sanctions to prevent deviant behavior and to change the behavior of miscreants back to the group norm. Yet Foucault clearly believed that almost all persons most of the time accepted the social norms of the institutional system and did their best to behave accordingly. In particular, the institutional rank and file do not differ from institutional leaders and disciplinarians in their views of right and wrong behavior. Almost everyone accepts the legitimacy of system norms and in fact participants' behavior embodies the norms (Brass, 2000).

Institutional participants are constantly observed by others in the system and additionally participants (perhaps best termed 'subjects') constantly observe themselves to identify and to correct behaviors that deviate from the system norm. Institutions have a knowledge base such as professional or academic disciplinary norms to control behavior, and examinations are given to subjects to measure proficiency in such knowledge. The institution accumulates data about each of its subjects to observe and examine how well each subject is adhering to the institutional norm. The institution is thus a self-perpetuating system. Thus the modern political world is characterized by subjects (lacking individual autonomy), not citizens.

Foucault's theory of power thus can be linked to Putnam's theory of civic engagement, probably the theoretical

perspective getting the most attention from American political scientists in the period roughly 1998–2008. Foucault's institutional systems can be characterized as inducing a high degree of cooperation, to the degree that we can neglect individual agency. Robert Putnam and Theda Skocpol among others have been concerned about the decline and frequent disappearance of voluntary associations (somewhat neglected by Foucault) as well as face-to-face interaction among neighbors. They argue that face-to-face interaction within associations promotes interpersonal trust, subsequently enhancing the capacity to cooperate, thus promoting democratic citizenship. The Putnam–Skocpol theory assumes individual agency among collapsing institutions; Foucault's theory assumes a lack of agency among self-perpetuating institutions. But both theories posit a lack of citizenship capacity among citizens, either because they lack individuality or because they as individuals cannot cooperate with other individuals. In such conditions, society obtains power elites (Putnam, 2000; Skocpol, 2004).

International Relations

The term 'power' appears in international relations research probably more often than it appears in any other scholarly field. International relations scholars, foreign ministry leaders, and military officers have been centrally preoccupied for centuries with the state of 'the balance of power.' Currently both ordinary language usage and scholarly writings about international relations generally refer to power in the sense of necessary resources for national security, particularly military and economic resources. In current language such resources are termed 'hard power,' as opposed to 'soft power,' generally the nonmaterial resources that boost a state's prestige, helping it to compete in the international system: its moral stature, such as in human rights, its export of prestigious cultural activities.

However international relations scholars have developed more technical uses for the term power. Within this field, the most common technical definition of the term power has been Robert A Dahl's basic definition (see above) with the addition of the element of intentionality: the ability of A to get B to do what A wants even though B initially resists A's intention (Barnett and Duvall, 2005: 49). This means, however, that power is a type of social causation because A changes B's behavior, but in many international situations to be studied, causation and hence power is complex, hence difficult to ascertain.

Much of international relations theory is systems oriented, meaning that the system of interaction constrains the behavior of decision-makers, so that causal definitions of power do not add much to our understanding of state behavior. It is best here to refer to system determination rather than the exercise of power by an individual agency. Yet some clear cases for the causal definition of power can be found in the relatively simple interactions within a short time frame, such as militarized disputes among two nations, or in general what can be considered to be a single foreign policy crisis between two nations. In such situations it is relatively clear that one state caused a change in behavior (policy) by a second state.

International relations literature places emphasis on a state's acquisition of resources that can be converted into power within the system (i.e., achieving intentions against initial opposition). This sometimes implies manifesting resources (military, economic) that deter others from resisting or deter others from achieving their intentions, opposed by the first power. In the study of domestic politics, such deterrence is referred to as 'anticipated reactions,' meaning actor B anticipates that actor A will oppose some action by B, and that B will subsequently be the loser as having less power resources than A, so that B never attempts the action at all. An observer can say that A deterred the action by B through B's anticipated reactions. As a parallel international relations scholars frequently refer to 'deterrence' between or among states in regard to possible state actions. The well-known and common use of deterrence refers to Cold War interactions between the U.S. and the U.S.S.R., as each deterred the other from the use of nuclear weapons to gain foreign policy goals. This is based on the idea of power as causation, in that we might say that the U.S. deterred or caused changes in the behavior of the U.S.S.R., dreading the use of nuclear weapons, while the reverse was also the case. However, anticipated reactions and deterrence theory in international relations is usually dependent upon counterfactual hypotheses; the observation is that something did not occur that otherwise would have occurred due to strategic restraint on the part of a state or actor. Dependency on counterfactuals is certainly a difficulty in the use of power as causation to analyze deterrence situations, although study of documents and interviews of participants, usually at a much later time, can provide data to understand what did not happen (Jervis, 2013).

International relations scholars have dealt with the difficulties in using power-as-causation by using 'end-run' strategies to flank the problems posed by complex causation. One such strategy is to rely on the concept of international system as the major causal factor. Another such theoretical strategy is 'regime theory,' which focuses on constraints on the behavior of international actors by interlocking behavior norms and shared framing of the meaning of international behaviors. The major example of regime theory is the category 'liberal international economic regime,' indicating a widespread belief in and practice of free trade among nations, combined with aspects of market theory such as rules for repayment of international loans made by states or by international agencies such as the International Monetary Fund and the World Bank. In using the regime theory framework, one can avoid directly causal statements about power and instead focus on describing the interpretive frames (particular simplifications of complex situations) that constitute the system of regime norms and how such interpretations of norms and beliefs are employed by state and nonstate actors (Krasner, 1983).

Michael Barnett and Raymond Duvall (2005) write that in international relations theory, it is best to state that power is fourfold. A particular study might focus on one of the four ideas of power, although in addition, scholars can frequently find links between the types. The Barnett and Duvall system is another attempt to avoid the complexities of observing power-as-causation. Their basic definition of power goes beyond gaining one's way in a specific social interaction and adds a dimension of producing social effects, beyond changing the other's course of action, that enable further control in subsequent social situations: "Power is the production, in and

through social relations, of effects on actors that shape their capacity to control their fate" (Barnett and Duvall, 2005: 45). Their first definition is 'compulsory' power, one actor compels another to change course in line with the first actor's intentions in a specific or series of specific social interactions. This is much the same as Dahl's power-as-causation. Their second definition of power is 'structural power,' meaning that social structure directly affects the power of various groups, classes, and so on, in their interaction within the social system. An application of the theory of structural power to international relations is Cold War systems theory in which a balance of deterrence shaped the behavior of the two major powers. Barnett and Duvall's third type of power is 'institutional' power, in which the intentions of actors are mediated through institutions to bring changes in social relations that enhance actors' capacity to control their own fate. For instance the State Department might influence International Monetary Fund officials to grant a loan to a friendly government in order to stabilize its position. However, the concept of institutional power is subject to difficulties in causal analysis, as, for instance, if the friendly government stays in power, this might have been the result of several interacting factors. Barnett and Duvall's fourth type of power is 'productive' power, similar to the ideas of Lindblom and Gramsci noted above, and constituting a generalization from the idea of regime theory. Productive power indicates producing social effects that are not immediate by affecting the perceptions and the language of the participants. Productive power focuses on the nature of discourse and how it affects the definitions of social identity among system participants, and its most well-known theorist is Michel Foucault.

See also: Agendas: Political; Pluralism.

Bibliography

Bachrach, P., Baratz, M., 1962. Two faces of power. American Political Science Review 56, 947–952.

Barnett, M., Duvall, R., 2005. Power in international politics. International Organization 59, 39–75.

Baumgartner, F., Berry, J., Hojnacki, M., Kimball, D., Leech, B., 2009. Lobbying and Policy Change. University of Chicago Press, Chicago.

Brass, P.R., 2000. Foucault steals political science. Annual Review of Political Science 3, 305–330.

Dahl, R.A., 1957. The concept of power. Behavioral Science 2, 201–215.

Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven, CT.

Edelman, M., 1964. The Symbolic Uses of Politics. University of Illinois Press, Urbana, IL.

Foucault, M., 1977. Discipline and Punish. Pantheon, New York.

Gerthe, H.H., Mills, C.W., 1946. From Max Weber: Essays in Sociology. Oxford University Press, New York.

Gramsci, A., 1971. In: Hoare, Quintin, Nowell Smith, Geoffrey (Eds.), Prison Notebooks. International Publishers, New York.

Harsanyi, J.C., 1962. Measurement of social power, opportunity costs, and the theory of two-person bargaining games. Behavioral Science 7, 67–80.

Hunter, F., 1953. Community Power Structure. University of North Carolina Press, Chapel Hill, NC.

Jervis, R., 2013. Causation and responsibility in a complex world. In: Finnemore, M., Goldstein, J. (Eds.), Back to Basics: State Power in a Contemporary World. Oxford University Press, New York, pp. 313–338.

Krasner, S. (Ed.), 1983. International Regimes. Cornell University Press, Ithaca, NY.

Lasswell, H.D., Kaplan, A., 1950. Power and Society. Yale University Press, New Haven, CT.

Lindblom, C.E., 1977. Politics and Markets. Basic Books, New York.

Lowi, T.J., 1979. The End of Liberalism, second ed. W.W. Norton, New York.

Lukes, S., 1974. Power: A Radical View. Macmillan, London.

March, J.G., 1955. An introduction to the theory of measurement of influence. American Political Science Review 49, 431–451.

March, J.G., 1966. The power of power. In: Easton, D. (Ed.), Varieties of Political Theory. Prentice-Hall, Englewood Cliffs, NJ, pp. 39–70.

McConnell, G., 1966. Private Power and American Democracy. Knopf, New York.

McFarland, A.S., 2004. Neopluralism: The Evolution of Political Process Theory. University Press of Kansas, Lawrence, KS.

Mills, C.W., 1956. The Power Elite. Oxford University Press, New York.

Nagel, J., 1975. The Descriptive Analysis of Power. Yale University Press, New Haven, CT.

Polsby, N.W., 1980. Community Power and Political Theory, second ed. Yale University Press, New Haven, CT.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.

Riker, W.H., 1964. Some ambiguities in the notion of power. American Political Science Review 58, 341–349.

Simon, H.A., 1953. Notes on the observation and measurement of power. Journal of Politics 15, 500–516.

Simon, H.A., 1957. Models of Man. Wiley, New York.

Skocpol, T., 2004. Voice and inequality: the transformation of American civic democracy. Perspectives on Politics 2, 3–20.

Walker Jr., J.L., 1991. Mobilizing Interest Groups in America. University of Michigan Press, Ann Arbor, MI.

Wilson, J.Q., 1980. The Politics of Regulation. Basic Books, New York.

Power in Society

Stewart Clegg, University of Technology, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Power has been seen as an 'essentially contested' concept. Some of the major recent contours of this contestation will be addressed in this article. The argument is that power is a concept whose meaning is variable in different language games and that it is important to be clear which concepts, from which language games, one is using in inquiry.

Introduction

Power has been seen as a concept that is 'essentially contested' (Lukes, 1974/2005). By this, it is meant that its essence, its central meaning, is fundamentally contested. Lukes (1974/2005) suggests this is because the concept cannot be divorced from normative perspectives. While he sees three essentially distinct normative perspectives in what he terms the liberal, the reformist, and the radical perspectives, many of the major recent contours of contestation do not follow this set of distinctions and depart in a number of ways from Lukes' concerns. The main reason for this is the eclipse of Marxist notions, which sustained the idea of their being a 'radical' view of human choices. Each of the three dimensions took a more or less voluntarist perspective on the nature of subjectivity: for the liberal, it was a matter of choices being exercised on relatively independent issues; for the reformist, it was a matter of choices constrained by a subjective and probabilistic calculation; for the radical, it was a matter of choice being constituted by structures of cultural domination. Lukes locates the basic concept of power in the work of Weber, amongst other theorists.

Weber: Ideal Types of Rule and Power

At the core of modernism, according to Weber, lies the rationalization of various institutional arenas such as the market, technology, law, state, the growing pervasiveness of rational calculation in all spheres of life, and the disenchantment of more spiritual or superstitious forms of calculation. Disenchantment means exposure to a world in which meaningfulness is never 'given.' Instead, meanings clash, struggle against the resistance of others, and are never wholly secure. Rational action consists precisely in the capacity to respond to the uncertainties of a world without meaning. Analytically, he conceived of three main types of rule around which rational action might revolve: principles of charisma, tradition, or rational legality. Each of these three types is a source of a particular type of *Herrschaft* (authority, or in better translation, domination).

In Weber's view, rationality had only a limited number of sources. It might derive from those few individuals who choose to impose value and meaning upon the world by confronting the world as something other than resolutely, unyieldingly, absolutely, and unbreakably cast in a mold. There were precious few such people in Weber's vision, other than charismatic leaders. Charisma is often identified with historical figures regarded as ordained with exceptional grace and a sense of calling that they were able to transmit and project to others, such as Jesus. Charisma is not a purely historical category, however, but it is only the rare charismatic leader, whose personal grace, sense of calling, duty, and devotion to some ultimate ideal is unshakeable, who would be able to impose such meaning on life. Hence, charisma was one of the key bases of authority that Weber identified. Genuine charisma, typically, is thought of as something above the grubby theater of mundane political wheeling and dealing: think of celebrities such as Princess Diana, who was able to parlay her celebrity into a passionate social advocacy on behalf of the victims of landmines and in so doing, bring about substantial changes in government policies. Here, was a charismatic celebrity for the modern age – especially in her death. What gave her charisma and celebrity was a fortunate marriage that rapidly proved to be ill-starred in many ways, a constant broadcast media focus culminating in an unfortunate death, all captured in images that were broadcast widely. In the past, celebrity and charisma could become deliberately entangled because of broadcast control of the image: the cult of the personality that modern leaders like Mao, Stalin, and Hitler cultivated were deliberate attempts to use tight control of the state to fashion and stage-manage a celebrity that transcended the political sphere and which sought to invest them with charisma. In modern times, celebrity and charisma are closely related but by no means the same.

Tradition functions as another ideal type, best seen in claims to authority based on lineage or dynasty, in principles of 'blood and breeding' or exclusion from opportunities to claim authority because of the lack of such blood and breeding. The succession principles of monarchy, or inheritance rights attached to property, are prime examples of rule by tradition. Most pervasive, however, as a form of authority, was what Weber took as the quintessential modality of modern rule: rational-legal bureaucracy, where rule is conducted according to abstract principles. The dominance of what Weber referred to as the rational-legal does not overwhelm all other bases for rule and authority. However, its pervasiveness is often summed up today in a popular fear of 'economic rationalism,' referring to those situations where only a legally mandated economic calculus is applied to policy formulation (such as competition policy), without regard for other ways of accounting for the impact of such policies on the human spirit, communities, and the fabric of society.

Each principle of rule posits a particular type of authority, and it is a simple matter to construct a matrix of these relations in order to generate some idea of what power might be. Weber defined power in terms of social action: power is "the probability that one actor in a social relationship will be in a position to carry out his will despite resistance, regardless of the basis on which this probability rests" (Weber, 1978: 53). An evident example of resistance to power thus arises when there are cases of conflict between principles of rule propounded by the power yielder and those thought legitimate by the resistant subject. Here, power is seen in terms of actions situated in and resisted by opposed principles or ultimate values.

Power as a Generic Concept

While Weber's definition of power is quite generic, the concept need not be premised on the conflict of rival, ultimate values. Steven Lukes (1974/2005), for instance, in one of the most influential books on the subject, regarded Weber's definition as essentially similar to the generic definition of power contained in the work of Robert Dahl (1957). Here, power is defined as the situation that prevails when A gets B to do something that they would not otherwise do. While this sounds a methodologically sparse and modern behavioral definition it is one that reaches back into the earliest development of the modern era – to Thomas Hobbes's conception of power in a causal, mechanical world.

The Community Power Debate that formed around Dahl's (1961) work concerned the distribution of power in postwar USA. Dahl (1961) insisted that such questions could only be answered empirically, rather than through a broad ranging essaying of concepts that were then current, such as their being a 'military–industrial complex.' Additionally, he thought them better answered at the local, community level rather than at the national level. To determine the distribution of power required an operational definition for empirical investigation. For Dahl, power was indissoluble from its exercise: someone had to be seen doing something to another for us to say that power had been exerted. Unlike Weber, he was not concerned with ultimate values, but he did think it important that the exercise of power should be determined by competition over issues that people thought important. Basically, he sought to determine what the key issues were in a community and then establish the outcome of the politics surrounding these decisions, to see who gained or yielded power.

Power, Real Interests, and Hegemony

By the time that Lukes (1974; second expanded edition 2005) published his book titled *Power: A Radical View*, Dahl's approach had already been criticized for leaving out of consideration 'what does not happen' by Bachrach and Baratz (1970). Nonissues, nondecision making, the construction of agendas, and the mobilization of bias through everyday routines were introduced to show how power is accomplished through covert means, through inaction, rather than through overt exercise of causal power.

Lukes took the critique one step further. He argued that the critics had only sought to qualify Dahl's behaviorism rather than transcend it through radical critique. What was required, he argued, was a conception of how power worked not only against people's expressed and implicit preferences but also their 'real interests.' The notion of real interests came from Marxist discourse – originally designed to address situations of 'false consciousness' – where a class of people were unable to recognize that the system within which they were embedded was the source of their entrapment. Lukes followed the Italian Marxist theorist, Antonio Gramsci, in introducing a 'cultural spin' on the idea of false consciousness by using the notion of hegemony. Where people were deluded about what they wanted and what they might authentically be, because of their inability to think in concepts other than those of domination and the dominated, they suffered from hegemony. The basic idea was that they were subject to a 'ruling culture' disseminated through the media, the state, education, and other institutional arenas.

Theories of hegemony became variations upon the theme of working-class acquiescence secured through normative incorporation by dominant ideology (Abercrombie et al., 1980). The narrative was one of concealment: Marxist analysis, in theory, should have provided revelation about their conditions of existence that the repressed would act on – but this rarely seemed to happen. To cover for the inadequacy of a theory whose propositions were not confirmed by practice, the theory developed an account of why the conditions that it predicted did not occur. The explanation was couched in terms of the collective inability of the repressed to be able to see and analyze the conditions of their repression, because they lacked the critical concepts to be able to do so. The repressed, in the form of the working class, were seduced by the ideas, blandishments, and inducements offered by the dominant system of repression, capitalism. They were blinded to their fate as exploited and repressed subjects.

Not everyone was blinded, however: theorists armed with the appropriate radical theory of Marxism could see the truth of the mechanisms that, despite and even through resistance, secured and reproduced consent, even while the subjects of analysis had the nature of their real interests concealed from them. However, Western Marxism, the source of these insights, was impotent at best and at worst irrelevant, in lifting the veils of occlusion (Anderson, 1976). In late twentieth century in the West, socialism became increasingly utopian (Bauman, 1977) while, in the East, it became increasingly dystopian (Feher et al., 1983), as the collapse of 'dictatorship over needs' produced its implosion.

The Discursive Turn

The ideological arguments, with their fundamental denial of the liberal sovereign subject, receded in importance in the 1980s. One response to this situation within the Marxist camp came from Laclau and Mouffe (1985), who, instead of thinking of hegemony as a state of mind, regarded it as a set of practices. These practices were primarily discursive. Hegemony thus came to be seen not in terms of its substantive content but its form. Hegemony existed when the indefinite possibilities of

signifying elements and their relations were strategically foreclosed (Laclau and Mouffe, 1985). Again, the emphasis is on things that do not happen. Any stable set of articulations is thus hegemonic: only permanent Cultural Revolution would not be, although the implications are not spelled out.

Real interests are what radical politics and theories claim to uncover and reveal. One way of doing this, suggested Haugaard (1998), that is immune to the criticism raised previously, might be to explain to actors the way in which their structured practices feed into the reinforcement of a regime of domination contrary to interests already known to them. What may be at issue are structural and institutional relations of domination rather than real but unknown interests. Once actors reflect upon their practical consciousness, they then enjoy relative autonomy. Power thus consists not so much in the suppression of real interests but the denial of the relative autonomy required for reflection (Haugaard, 1998, p. 141).

Haugaard (1998) notes that conceptions of interests form discursively, even to the extent that one can persuade some people that they have real but unrealized interests blocked by their present conditions of existence. Such discursive realities help shape them to resist the sources of domination identified. Politics mobilize power and resistance through discursive expression of existing, past, or future conditions of existence. Thus, resistance is not endemic but depends on the practical consciousness of people and the categorical devices through which they lead their everyday lives in relation to the discursive consciousness proffered by various media of representation – parties, press, radio, television, the Internet, public intellectuals, and so on. Only in rare moments is such diversity represented or resisted in the name of some great unifying narrative or theme. Most probably, these will be those relatively rare moments where a privileged identity is posited as a historical subject: the subject people, the subject party, the subject race, the subject nation, the subject church, the subject gender, the subject sexuality, the subject ecology, and so on.

A plurality of life-worlds, with interpretative resources, and translation devices for switching between them, minimizes the chances of hegemony. People from sophisticated cultures routinely have a multiplicity of interpretative horizons available to them, as a result of exposure to different life-worlds through various media, educational experiences, travel or migration, or interpersonal relations. Of course, to the extent that these experiences are compartmentalized, one way or another, the ability to achieve critical distance from any posited identity is diminished.

The theorization of autonomy in terms of moves between interpretative horizons means that the ability of actors to transcend their social environments and structured context does not presuppose an ability to 'escape' the processes of socialization by adopting a neutral, culturally unbiased, position. Autonomy does not involve a flight into any transcendental realism of absolute truth, nor does it involve the transcending of false consciousness. Critical faculty does not imply the objectivity of truth, nor does it entail an undersocialized concept of agency whereby autonomy is gained by transcending the basis of social culture. It requires only relative reflexivity on the part of subjects. Such a conception of the relation of power and knowledge, with its emphasis on conditions of relative autonomy, turns out to be a way of reapproaching the classical questions about power and community that were at the center of Dahl's work. The difference is that, this time, the approach has been through Foucault rather than Dahl.

Foucault and Power

Foucault provides a detailed history of some power practices and techniques that have characterized modernity in a way that is nonreductive and complements Weber (Dandekker, 1991), although he understates the importance of Max Weber's (1978) argument concerning discipline. Nonetheless, there are common themes (see Szakolczai, 1998). Different emphases on power occur through the course of Foucault's work. First, there is his work on genealogy as the natural history of specific practices; second, the work on the 'care of the self,' where histories are used to critique the present, by uncovering the constituted nature of what have been taken as the here-and-now and third, the discussion of government or 'governmentality.'

In Foucault's (1979) most influential genealogical model of 'surveillance and discipline,' power is constituted in an intimate relationship with liberty. Because the exercise of power presumes a free subject, there will always be resistance. Where liberty is extremely limited, because power is stable and hierarchical, Foucault terms it domination. (The parallelism of this account with that of Weber's 1978 discussion of *Herrschaft*, is evident.) Two models of power are compared and contrasted: the repressive and the positive. For the latter, power shapes subject dispositions in a variety of structural ways, of which the Benthamite panoptical device is only one colorful metaphor. (Foucault's 1979 concern with the facilitative and productive aspects of power relates to the kind of arguments made most clearly by Parsons (1963), although it is not acknowledged as such.)

In general Foucault does not offer a critique of power. It would be impossible to do so because, to coin a phrase, power has no other: it is inescapable. Liberation from power is only a dream (or a nightmare). However, specific forms of power may and do change. The 'punitive' power of corporal excess may be contrasted with the 'positive' power of bureaucracy. However, there is no space outside of power, no knowledge that is pure truth. (In this way utopia and dystopia are avoided.) In reading Foucault, it is necessary to uncouple a sociological from a philosophical question concerning truth. The former deals with what passes for truth rather than asking what is the truth and with how truth-effects are produced, in the Foucauldian vernacular. It is the latter that concerns Foucault, rather than the former.

In Foucault's model of power relationships through government – the structuring of action through conduct – which he defines as the capacity of people to regulate their own behavior through various rationalities, his treatment of liberalism suggests that it is conditions of relative autonomy that define the nature of political community.

Community Politics after Foucault

Foucault suggests that democracy is neither an ideal with which to measure actual practice nor a method for producing

utopias. It is the messy and politically expeditious ways of making sense of politics its participants routinely use. It is this fundamental approach that Flyvbjerg (1998) takes from Foucault. He provides a deeply documented, rich narrative of one complex set of issues and processes associated with the many ways in which political communities were constructed and deconstructed in the laborious pursuit of local interests and issues surrounding a proposed traffic plan for the city of Aalborg in Denmark.

Aalborg was not unusual with respect to its power relations: These relations were sufficiently asymmetric as to verge on domination, with commercial interest overriding more civic interests. Those who were dominant in power relations were well aware of their dominance and found little need to explain themselves or their actions in rational terms compared to those relatively weaker. The authority of technical rationality was a weapon of the politically weak – those without other resources at their disposal. Hence, not surprisingly, technically rational plans were constantly undercut and attempts at securing democratic outcomes based on that technically rational basis proved weak. Neither a normative emphasis on democracy nor rationality proved sufficient to avoid domination by superior power relations. Such power defined reality and, not surprisingly, the rationality accepted depended on power rather than legality. One of the main strategies of the dominant group was to use techniques of rationalization as power: indeed, the greater the power the lesser the rationality. Thus, we might conclude that power is more ordinarily stable than conflictual and that its relations are constantly being produced and reproduced in process. Where power is stable, power relations can more characteristically be legitimated in terms of a prevailing mode of rationality.

The Language Games of Power

The idea that power is an essentially contested concept presumes that there is an essence that is contested, as suggested earlier. However, having reviewed these different ways of theorizing power, it seems increasingly difficult to hold that this is the case. Either there is a great deal of error concerning the true nature of power or there is, perhaps, no true nature. Consider this: Dahl defines power in terms of its exercise; Foucault sees it in terms of subject dispositions; Weber in terms of might and domination, while Lukes sees it in terms of hegemony. Can they all be correct? Is one right and are the others wrong? Or, as Haugaard and Ryan (2012: 9) suggest, if, rather than being an essentially contested concept, power is a key term in various analytical language games that analysts of society engage in, then it behooves them to make clear how they are using the concept analytically rather than merely making normative statements. Given this, it is better to eschew discussion of essences being eternally contested. Instead, we should see power as a term that features in many related but distinct language games that, in Wittgenstein's (1967) terms have only limited family resemblances. Hence, some language games will center on domination, and others on power as productive of subject dispositions. The different concepts of power are tools with which to do things. Just as one would not have a toolbox comprising only hammers so that every problem becomes a nail to be hammered, so one would not work with an omnibus concept of power but would draw on different language games in different contexts of and for explanation and understanding. It is this understanding that characterizes Clegg's (1989) account of the different circuits of power through which empirical action might move and still be considered as power. Lacking an essence, power might be conceptualized in a manner that suits the task at hand, drawing on different traditions of analysis embedded in distinct language games that become more or less tangled up with each other in practice (see Clegg, Courpasson and Phillips 2006; Clegg and Haugaard 2012; Haugaard and Clegg 2012 for examples.).

Conclusion

In conclusion, we may note that in the genealogy of power that has been constructed here, power is not defined as only one thing, or in one set of terms. The terms shift with debate but there is a family resemblance among them. The emphases have changed with the values underlying scholarship. Lukes' importation of the Gramscian notion of hegemony seemed to take us a long way from the political communities that Dahl studied with such methodological scruples. Laclau and Mouffe's introduction of Foucault to the debate, somewhat surprisingly, brings debate back full circle to issues of political community because of the centrality their conception of power placed on discourse and identity. The theme of the relative autonomy of political processes dominated the late work of Foucault and was taken up empirically by Flyvbjerg. Thus, nearly 50 years after the Community Power Debate first kicked off contemporary sociological interest in the conceptual relation of power and society, the concepts have come back to their empirical point of departure, but they are much richer and more sophisticated than when they first embarked on that journey, because the family of power concepts has grown more sophisticated and developed more branches over that time.

See also: Action, Theories of Social; Authority, Social Theories of; Charisma, Social Aspects of; Class: Social; Community Sociology; Foucault, Michel (1926–84); Political Sociology; Weber, Max (1864–1920).

Bibliography

Abercrombie, N., Hill, S., Turner, B.S., 1980. The Dominant Ideology Thesis. Allen and Unwin, London.
Anderson, P., 1976. Considerations on Western Marxism. Verso, London.
Bachrach, P., Baratz, M.S., 1970. Power and Poverty: Theory and Practice. Oxford University Press, Oxford, UK.
Bauman, Z., 1977. Socialism: The Active Utopia. Allen & Unwin, London.
Clegg, S.R., 1989. Frameworks of Power. Sage, London.
Clegg, S.R., Courpasson, D., Phillips, N., 2006. Power and Organization. Sage, London.
Clegg, S.R., Haugaard, M. (Eds.), 2012. Power and Organizations, vols. 4. Sage, London.
Dahl, R.A., 1957. The concept of power. Behavioural Science 2, 201–205.

Dahl, R.A., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven, CT.
Dandekker, C., 1991. Surveillance, Power and Modernity. Polity Press, Cambridge, UK.
Feher, F., Heller, A., Markus, G., 1983. Dictatorship Over Needs. Blackwell, Oxford, UK.
Flyvbjerg, B., 1998. Rationality & Power: Democracy in Practice. Chicago University Press, Chicago.
Foucault, M., 1979. Discipline and Punish. Penguin, Harmondsworth, UK.
Haugaard, M., 1998. The Constitution of Power. Manchester University Press, Manchester, UK.
Haugaard, M., Clegg, S.R. (Eds.), 2012. Power in Politics, vols. 4. Sage, London.
Haugaard, M., Ryan, K. (Eds.), 2012. Political Power: The Development of the Field. Barbara Budrich, Opladen.
Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy. Verso, London.
Lukes, S., 1974. Power: A Radical View. Macmillan, London.
Parsons, T., 1963. On the concept of political power. Proceedings of the American Philosophical Society 107, 232–262.
Szakolczai, A., 1998. Max Weber and Michel Foucault: Parallel Life–Works. Routledge, London.
Weber, M., 1978. Economy and Society. University of California Press, Berkeley, CA.

Power, Politics, and Influence in Organizations

John N Harris, Zachary A Russell, Liam P Maher, and Gerald R Ferris, Florida State University, Tallahassee, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of the related concepts of power, politics, and influence in organizations has been a topic of interest for well over a century, with research expanding considerably in the last several decades. Power, politics, and influence are fundamental, and effectively ubiquitous, components of organizational life, serving to help explain how individuals act on their environments and interact with other individuals in organizations. This article briefly covers the existing knowledge of power, politics, and influence in organizations through a discussion of seminal works in these areas. Emerging areas of study and directions for future research are also discussed.

Power, Politics, and Influence: Historical Perspective and Literature Review

The History of Power, Politics, and Influence

Power, politics, and influence all represent components of human agency, which acknowledge that rather than being characterized as simply passive reactors to their contexts, individuals are active creators of their environments. Theories of human agency suggest that people act on their environments in ways that create, transform, preserve, and even destroy it, and they also act on themselves to adapt to contextual conditions (Bandura, 2006). As such, this implies the proactive initiation of goal-directed behavior.

The study of influence is perhaps the oldest experimental paradigm in the behavioral sciences, dating back to Triplett's (1898) investigation demonstrating that cyclists performed better in the presence of other cyclists than when alone (i.e., thus initiating the early underpinnings of the subsequently developed theory of social facilitation). Of course, many would date the study of influence even further back, to the sixteenth-century works of Niccoló Machiavelli (e.g., *The Prince*), which, among other things, illustrated how to gain and maintain political power. Machiavelli's work was likely one of the first to break from an abstract view of ideal ethics and political behavior, and instead focus on the effective truth of how power is amassed and preserved.

Nearly four centuries later, Dale Carnegie's self-help book *How to Win Friends and Influence People* brought the idea of influence into popular press, and predated much of the academic work in organizational politics and influence by decades. The tenants of the book were rooted in amicable and sincere interpersonal influence tactics and were applied to both professional and personal venues. Specifically, the book discussed the fundamentals of handling people, techniques to make people like you, ways to win people over to your way of thinking, and how to manage change within others. With over 30 million copies sold worldwide, there is no doubting the reach of Carnegie's message, and this work is recognized across disciplines by a wide variety of practitioners and academics alike. Carnegie's book drew attention to the role of social influence tactics in people's everyday lives and helped spawn the academic investigation into power, politics, and influence in organizations. This article will continue with a review of the academic literature that exists and provide some directions for future investigation.

Literature and Reviews of Power, Politics, and Influence

In the same year as Carnegie's classic mentioned earlier, Lasswell (1936) defined politics as "who gets what, when, and how," and argued that it was the manner in which aspects of the environment were manipulated that predicted success. Furthermore, Lasswell was one of the first scholars to identify and discuss the key role of social astuteness and influence execution as a determinant of the success of influence attempts (Almond, 1987). Twenty-five years later, Burns (1961) conducted one of the first scholarly discussions of politics in organizations. He suggested that organizations represent arenas or systems where competitions are staged by members for valued resources.

Over time, scholarly interest in power, politics, and influence in organizations has continued to flourish. This has provided the field with a rich collection of ever-expanding research examining these three aspects of the organizational sciences. Just some of what has been learned about power, politics, and influence is reported here. We encourage readers to explore the wealth of other published work on these topics, beyond the classics and reviews mentioned.

Power in Organizations

French and Raven's (1959) work examined The Bases of Social Power by conceptually deriving logical bases for social influence and psychological change, which are dependent on the relationship between the person and social agent. From this relationship, they developed five different bases of power. Expert power states that the social agent has knowledge above and beyond that of the person. Referent power refers to a situation where a person or people share a high level of identification with a social agent. Legitimate power implies that the social agent has a higher hierarchical standing than the person. Coercive power is used when a social agent is able to exert some sort of harm on the person, which is counter to reward power, where the social agent is able to reward the person for a specific performance.

Emerson (1962) sought to corral the contemporary literature on power-related concepts and derive a simple but

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22027-8

generalizable theory of power through social relations. This theory gave no attention to trait or contextual factors and instead examined the power–dependence relationship between actors. Emerson reasoned that power and dependence between actors has a reciprocal nature, and the theory at its rudimentary level states that power is equal to the dependence of actor A on actor B. This dependence is positively related to motivational investment toward goals that are mediated by the other party. When there is an imbalance in this power and dependence, such that the dependence of actor A on actor B, moderated by the motivational investment of a goal, becomes greater than or less than the dependence of actor B on actor A, then one is said to have power over the other. By definition, power does not exist without the property of resistance, which occurs when an actor yields to something he or she would normally resist in an effort to work toward a treasured goal.

Emerson (1962) theorized that there are costs associated with one party meeting the demands of the other party, and that actors employ strategies to reduce these costs. Generally, these costs consist of value or identity shifts that make the process of meeting the other person's demands less painful for the dependent. Cost reduction strategies are, by definition, not necessarily associated with balancing operations, which specifically aim to change the balance of power rather than psychologically facilitating the demands of the more powerful actor. There are four balancing operations that bring a two-party interaction back toward power equilibrium. The first two are reducing motivational investment in goals moderated by the more powerful actor and finding alternative sources of gratification for those goals. The third balancing operation is for the more powerful actor to increase motivational investment in a goal moderated by the less powerful actor, which is contingent on the fourth operation that no alternative source of gratification is available for the more powerful agent to achieve this goal.

Utilizing the idea of power as the ability of A to modify B without modifying his or her own behavior, Schein (1977) conceptualized the bases of individual power, the intent of a powerholder, and the means used by a powerholder. Building upon French and Raven's (1959) sources of power, Schein noted that individuals use power in self-serving ways to "a) obtain more power, b) maintain current power, and c) eliminate other's power" (p. 66). All actions and movements within an organization are to achieve more power, and often are accomplished through behaviors that can be either congruent or incongruent with the organization's goals. Behaviors that are incongruent were deemed political by Schein, and involve the use of overt or covert actions that align with the individual's current power bases. Which power bases are utilized and the type of behaviors selected often will be determined by the intent of the individual.

Politics in Organizations

Between the mid-1970s and the mid-1980s, several scholars proposed political perspectives on organizations (i.e., Mintzberg, 1983; Pfeffer, 1981; Porter, 1976; Salancik and Pfeffer, 1974). These political perspectives seemed to share in common the view that to obtain power in organizations, it took the skillful use of political behavior and/or influence tactics/strategies, which, Pfeffer and Mintzberg argued, required political skill. The recent emphases by scholars in this area today have maintained the intertwined nature of these constructs, demonstrating the commonalities or how the different components work together synergistically. For example, Pfeffer (2010) has argued that power is obtained through the skillful selection and execution of mechanisms of influence (i.e., politics or political behavior and influence tactics), which require political skill to ensure effectiveness in execution and goal attainment.

Salancik and Pfeffer (1974) were among the first scholars to examine power above the individual level by examining subunits in university academic departments. They found that the departments able to acquire the most outside resources (i.e., grants and other funding sources), also received the greatest amount of scarce resources from the university. Indeed, the "power derived from acquiring resources is used to obtain more resources, which in turn can be employed to produce more power – the rich get richer" (p. 470).

Taking a more structural approach, Astley and Sachdeva (1984) synthesized what they believed to be a fairly fragmented theoretical literature on organizational power by developing a theoretical model that incorporated three structural sources of power. Hierarchical authority is grounded in the conception of power being inherent in an official role or position. Resource control suggests that power can come from the ability to distribute or withhold vital resources. Finally, network centrality yields power when a person holds a central network position within multiple interdependencies, thus making him or her a conduit to resource facilitation. After running pair-wise comparisons among the three sources of power, these authors concluded that there was a mediating effect of the third, suggesting a more interconnected theoretical relationship.

Influence in Organizations

This previous theoretical work laid the framework for the empirical study of intraorganizational influence tactics by Kipnis et al. (1980). Whereas most research during this time focused on executive influence, this piece focused on the political influence tactics employed at lower levels of the organization by middle managers. Using content analysis, they identified eight dimensions of influence: assertiveness, ingratiation, rationality, sanctions, exchange, upward appeals, blocking, and coalitions. This study not only provided practical insight into how organizational agents influence each other, but also tangentially supported the notion that influence in organizations is not just bound by leadership. Indeed, there are political underpinnings to multiple levels of the organization.

A core component of how individuals may exert influence, participate successfully in politics, and gain power is through the utilization of select influence tactics known as impression management techniques. Quite literally, impression management represents the behaviors in which actors engage to manage the identities they have selected or been assigned by others by way of influencing the impressions others have of them (Tedeschi and Melburg, 1984).

Although there are a variety of impression management tactics available, they can be described primarily in terms of two dimensions: defensive-assertive and tactical-strategic (Tedeschi and Melburg, 1984). Assertive impression management deals

mostly with establishing an identity, and thus, tends to be active in nature. On the other hand, defensive impression management is more reactive, occurring when actors face some sort of issue or threat to an identity. Tactical impression management is geared toward short-term objectives, while strategic impression management is long-term focused. All impression management techniques essentially can be classified along these two dimensions, so tactics can be described as defensive-tactical (e.g., apologies), defensive-strategic (e.g., learned helplessness), assertive-tactical (e.g., ingratiation), and assertive-strategic (e.g., status).

Liden and Mitchell (1988) examined ingratiation, a form of impression management common in organizations and popular in organizational research, a major component of many social influence attempts, and a component of social interaction processes. They defined ingratiation as behaviors employed by a person to make himself or herself more attractive to another. The linear conceptualization of the ingratiation process explores the motivation of ingratiation, choice of ingratiation strategy, ingratiation behavior, target reaction, and outcomes. This portion of the model is impacted by situational determinants.

Ingratiation behaviors are motivated by a need to be liked stable trait, or a situational opportunity to defend or promote one's self. A risk assessment would determine the strategy used to ingratiate the target, as there are potential costs associated with ingratiatory behavior. If the ingratiator chose to move forward, three distinct behaviors are possible. Target-directed behaviors represent interpersonal exchanges between the ingratiators and targets, and are meant to make targets feel good about themselves. Self-presentation deals specifically with impression management of one's self, and could include positive self-description or, in a more defensive situation, apologetic language. Third-party tactics involve influencing a knowing or unknowing assistant to convey positive messages about the ingratiator to the target. The targets' perception will depend upon whether the attempt is perceived as assertive or defensive, and whether it is sincere or manipulative. At the time of this study, little was known about the short- and long-term outcomes of ingratiation behaviors.

Leary and Kowalski (1990) described impression management (i.e., also referred to as self-presentation) as the process by which people control the impressions others form of them. They compiled the relevant impression management literature and developed a two-component model that conceptualizes impression management as two distinct processes: impression motivation and impression construction. Impression motivation is defined as the degree to which people are motivated to control how others see them, which is impacted by both situational and dispositional factors. In their model, a host of factors are examined using three distinct dimensions. Goal relevance of impression is the extent to which impressions have an impact on a goal or set of goals. The value of desired goals is rooted in the well-accepted notion that there is a positive relationship between the importance of a goal, and one's motivation to do so. Finally, if there is a high perceived discrepancy between desired and current images, people will tend to be motivated to act in a way that changes these perceptions.

The second component of the model addresses how impressions are constructed. Some criticisms of the impression management models stem from the notion that people are trying to deceive the target of their impression management. While this can be the case, it is also true that individuals simply want to portray themselves in a positive light, and thus their self-concept is an important factor in determining what type of impression to build. When individuals feel the need to fabricate or embellish aspects of themselves, they will draw upon desired and undesired identity images to construct an impression. Expectations about how an individual in a particular role should behave create a role constraint that guides the observable behaviors of a person and thus impacts impression formation. Much literature exists suggesting that impressions often are tailored to meet the target's values in attempts to gain favor, and even may do so by portraying values with which senders may not agree as long as they are compatible with the target's values. Finally, impression strategies are constructed based upon the senders' perception of how they are currently viewed by the target.

So, as noted, power, organizational politics, and social influence represent distinctive yet overlapping areas of theory and research in the organizational sciences that have been studied actively for decades. Furthermore, periodically over those years, many comprehensive reviews of these literature have been published, some selectively reviewing just the work regarding each separate construct, and other recent reviews that incorporate two or even all three of the constructs together. As a result, collectively and individually, we as a field know quite a bit about power (e.g., Emerson, 1962; French and Raven, 1959; Mintzberg, 1983; Salancik and Pfeffer, 1974), organizational politics (e.g., Ferris et al., 2002a,b; Ferris and Hochwarter, 2011; Silvester, 2008), and social influence (e.g., Kipnis et al., 1980; Leary and Kowalski, 1990; Liden and Mitchell, 1988; Tedeschi and Melburg, 1984).

Power, Politics, and Influence in Action

More recently, research in this area has begun examining how individuals can effectively employ their power, politics, and influence. Among the most prevalent explanations for this effectiveness is the construct known as political skill (Ferris et al., 2012). Originally described in the 1980s (e.g., Mintzberg, 1983; Pfeffer, 1981), political skill did not start to catch on in continued scientific work until the late 1990s and mid-2000s with the development of the Political Skill Inventory, and the theoretical development of the political skill construct (Ferris et al., 2007).

Political skill is defined as "the ability to effectively understand others at work, and to use such knowledge to influence others to act in ways that enhance one's personal and/or organizational objectives" (Ferris et al., 2007: p. 291). The construct consists of four dimensions: social astuteness, interpersonal influence, networking ability, and apparent sincerity. These four dimensions together form the overall construct, which is a particular form of what is known as a social effectiveness construct. These sorts of constructs are individual differences that each person possesses, to a greater or lesser degree, which help to facilitate easy and successful social interactions for those who score highly in them, through the dimensions listed earlier.

Social astuteness allows for keen understanding of the other people and social situations around an individual.

Interpersonal influence deals with an individual's ability to be flexible, to adapt to particular situations, and be able to still effectively influence others. Networking ability is quite literally how well an individual creates, maintains, and utilizes valuable contacts or networks of contacts. Finally, apparent sincerity might be the most important of the dimensions because, without this, success with any of the other dimensions is much less likely. Apparent sincerity deals with politically skilled individuals appearing genuine or honest, regardless of their underlying intentions. Influence attempts made when the target of the attempt does not believe the intentions of the influencer are sincere are much less likely to be successful than they would be if the influencer has a high political skill and target believes the influencer has no ulterior motives (Ferris et al., 2012).

Political action and the acknowledgment of politics in organizations are inherently perceptual. An action that one target of influence might interpret as politically motivated for a negative cause might be interpreted in a positive manner by another target, and might even go unnoticed and not deemed political by yet another target. The perception of organizational politics impacts organizations in many ways. Because of this, the antecedents, consequences, and moderating factors of the perceptions of organizational politics have received a greater amount of attention over the past few decades (e.g., Ferris et al., 2002a).

Researchers have found organizational influences (e.g., centralization of power, formalization, hierarchical level), work environment influences (e.g., advancement opportunities, feedback, interactions with coworkers), and personal influences (e.g., Machiavellianism, affect) to directly lead to individuals' perceptions of politics in organizations, which are then associated with a number of outcomes. Perceptions of politics have been found to be positively related to job anxiety, organizational cynicism, organizational withdrawal, and political behavior, among many others, and negatively related to job satisfaction, job performance, and organizational commitment (Ferris et al., 2002a).

The direct relationships between the antecedents of perceptions of organizational politics listed earlier and its consequences can be moderated by many individual variables, such as age, gender, self-efficacy, perceived control, and understanding of organizational politics. Thus, these moderator variables can result in considerable variation across individuals regarding how politics is perceived and the consequences such perceptions produce. What one might view as a threat and intimidating political landscape could be perceived as an environment full of exciting opportunities by another. The political skill of an individual is yet another moderating factor that can impact how politics is viewed and approached by individuals. Politically skilled employees are better able to navigate political settings, would have much different perceptions of an organization's political atmosphere, and thus would be able to work through the politics to ensure positive consequences (Ferris et al., 2002a).

Measurement of Politics in Organizations

The fact that political behavior and the political environment of organizations are perceptual and can change from person to person is one of the factors associated with the difficulty measuring power and politics in organizations. As mentioned earlier, perceptions of politics can vary widely from person to person depending upon certain demographic, personality, and skill-set factors. Thus, determining a 'true' level of politics within an organization is likely impossible, as the construct can be construed as largely perceptual in nature.

Other issues exist when attempting to measure political behavior. In the organization sciences in general, reviewers and methodologists often criticize the prominent use of self-report measures. Due to the tainted general attitude toward politics, self-report measures are even more of an issue due to social desirability bias. Many of the items in existing measures are negatively toned (Ferris and Hochwarter, 2011), and because of this negative tone, individuals completing a survey who prefer to view themselves in a positive light would be less likely to accurately admit they partake in certain behaviors deemed as political. Additionally, individuals in organizations who are politically skilled may not wish to accurately complete surveys that they feel could inhibit their functioning within the organization's political network.

A final issue with measurement revolves around the lack of agreement on what constitutes political behavior. Many authors have acknowledged the definitional issues that persist in the power and politics literature (e.g., Ferris et al., 2002a; Silvester, 2008; Ferris and Hochwarter, 2011). Some definitions focus primarily on the negative aspects of political behavior in organizations, whereas others acknowledge that good can result and take a more neutral stance. As the literature on power and politics expands, it is important for authors to focus on aspects of these constructs that are clear and observable.

It is impossible to measure a construct that does not have a clear definition, so by providing accurate and precise definitions scholars can encourage future research that builds on previous work in a constructive manner. The measurement of power and politics has many issues that require attention, from the development of valid measures to the development of better methodologies. In order to develop these measures and methodologies, clearer and more precise definitions of what is actually being studied are necessary.

The study of power, politics, and influence in organizations has received extensive attention in many types of organizations, both large and small, public and private. However, it is somewhat ironic that organizational scholars have paid very little attention to partisan politics: the study of professional politicians, the skills they possess, and the ways they operate within their professional roles (Silvester, 2008). Some researchers have suggested that the roles of professional politicians and regular workers are very different, but little evidence exists about such differences. Silvester noted that more consideration should be given to how the study of politicians at work and political activity in organizations may overlap. Political behavior certainly is prominent in politicians, and although the behavior is expected and sanctioned for politicians, they still need to effectively use their power and influence to satisfy their various constituents. Like traditional workers in organizations, they must monitor and navigate their environments and utilize political skill to succeed. Silvester suggested that further synthesis of the bodies of research constructed by sociologists and political theorists could greatly contribute to the study of politics in organizations and vice versa.

Research on Power, Politics, and Influence: Key Issues and Directions

Rather than simply continuing to reiterate what other fine reviews already have stated, we focus our paper on several important issues in the area of power, politics, and influence on which we need much more research. We present these issues in no particular order of importance.

Nonlinear Effects of Influence

Although a significant amount of research has been devoted to examining social influence and its effects as a linear process, a discussion of the potential nonlinear effects of influence in the organizational sciences is warranted. All too often, antecedent variables are presented, theorized, and tested as linear in their relation to organization outcomes (e.g., job performance, job satisfaction), but, in reality, a growing number of empirical studies have demonstrated that variables believed to have a positive, linear relationship with these outcomes actually are associated with negative outcomes and effects (Pierce and Aguinis, 2013). This nonlinearity is likely no different for much, if not all, of the research in power, politics, and influence.

Pierce and Aguinis (2013) describe this pervasive nonlinearity as the too-much-of-a-good-thing effect (TMGT effect). They propose that all linear, positive relationships eventually reach a context-specific downturn, and level-out or become negative at high levels or when taken too far, resulting in a curvilinear relationship (e.g., an inverted U-shape). This effect was proposed to account for the seemingly paradoxical relationships found in a variety of relationships.

In the power, politics, and influence literature, the linearity of antecedents' relationships with outcomes cannot be as quickly inferred as it might have once been. An example of this can be found in the work of Ames and Flynn (2007), who found a curvilinear relationship between assertiveness and leadership effectiveness. Leadership was less effective at both the high and low extremes of assertiveness, tipping down to increasingly less effectiveness at a context-specific inflection point. Evidence such as this and the meta-theoretical framework of the TMGT effect suggest that there is actually an optimal level for many constructs that is not always the highest, or most extreme, values. In many instances, as in the one illustrated by Ames and Flynn, more is not always better.

Power, Politics, and Influence as Positive (Not Simply All Negative)

Despite the stigma attached to the word politics, not all behaviors dealing with power, politics, and influence are inherently negative. Notwithstanding this widespread misunderstanding, there has been very little research done on the positive aspects of these constructs. Hochwarter (2012) is a notable exception, and his review leads our discussion of the positive, or dualistic, aspect of power, politics, and influence. Continued research into the positive aspects of these areas is presented as a clear direction for future research.

The Greek philosopher Aristotle once noted that "man is by nature a political animal"; an observation that advocates that political behavior as ingrained in who we are as a species. Indeed, evolutionary research suggests that the ability to acquire and manage resources, oftentimes with the influence and assistance of others, will separate those who thrive from those who perish. Whether this phenomenon is bread out of our DNA, or through social construction, it is clear that an organizational actor's survival is predicated on the ability to properly acquire and manage contextually relevant resources (Hochwarter, 2012).

In practice, organizations are a collective of actors working in a hyperchanging environment, who hold asymmetrical resources, which cause significant uncertainty and situational anxiety. Political behaviors provide the ability to adapt in these environments, and it has been shown that those organizational actors who can skillfully swap, acquire, or share resources or information among other organizational actors have been able to reduce ambiguity and thus reduce associated stress levels.

Political behavior also can be effectively employed to combat injustices within the organization. Organizational actors at all levels can utilize political behaviors to bring balance to an untold number of unjust situations such as discriminatory promotion practices, workplace harassment, and compensation. While it is acknowledged that self-serving political behavior has the power to create or maintain unjust situations, it is also the case that personal orientation and context can lead to equally powerful instances of other-serving political behaviors aimed at balancing injustices that negatively impact other actors who reside at different or identical levels of the organization.

Perhaps most famously, political skill has been positively linked with job performance and career success (Hochwarter et al., 2007; Ng et al., 2005). This relationship may once again conjure up images and anecdotes of organizational actors schmoozing and backstabbing their way to success at the expense of others, and although this can be the case, there are many demonstrated positive political behaviors that improve personal outcomes. The initial stages of any young career or contextual situation can be classified as weak and ambiguous, granting the newcomer very few cues about how to act appropriately and effectively. For example, proactive employees who effectively engage in information-seeking activities are able to reduce information incongruences, and build task-specific skills, as well as learn to properly self-regulate within the organizational setting.

The previous discussion suggests many instances and situations in which political skill, power, and influence are able to further the cause of the individual organizational actor, and the organization itself. There is immeasurable work to be done examining the positive aspects of these constructs. Hochwarter (2012) has suggested that a clear distinction be made between positive politics and other casual subordinate-benefitting activities, such as organizational citizenship behaviors. Additionally, the genetic, situational, and motivational antecedents of positive politics are yet to be explored. Fedor et al. (2008) proposed that research be done on the organizational actor's perception of political behavior and the resulting positive outcomes. Put simply, do the ends justify the means? Future research will hopefully answer these and other related questions and lead us to a more equitable and balanced understanding of organizational politics.

Conclusion

Power, politics, and influence represent key components of human agency, and they play important roles in interpersonal interaction in organizations and in daily life. Their importance has been recognized for hundreds of year, but it is just in the past 30 years that the academic investigation into these concepts' impact in organizations has received great attention. We have attempted to summarize and present the majority of the central findings involving power, politics, and influence in organizations, but, for sake of concision, much great research has been omitted. We have also presented some areas of politics in organization that we believe deserve greater attention and will be rewarding paths of investigation. In sum, although much has been done to investigate and understand the role of power, politics, and influence in organizations, there is still much to be done. We call for scholars to continue to build upon previous findings and provide greater insight into the role these constructs have in organizational life.

See also: Industrial–Organizational Psychology: Science and Practice; Management: General; Organizational Behavior, Psychology of; People in Organizations; Power; Work Motivation.

Bibliography

Almond, G., 1987. Harold Dwight Lasswell, 1902–1976. Biographical Memoirs 57, 149–274.

Ames, D., Flynn, F., 2007. What breaks a leader? The curvilinear relation between assertiveness and leadership. Journal of Personality and Social Psychology 92, 307–324.

Astley, W., Sachdeva, P., 1984. Structural sources of intraorganizational power: a theoretical synthesis. Academy of Management Review 9, 104–113.

Bandura, A., 2006. Toward a psychology of human agency. Perspectives on Psychological Science 1, 164–180.

Burns, T., 1961. Micropolitics: mechanisms of institutional change. Administrative Science Quarterly 6, 257–281.

Emerson, R., 1962. Power-dependence relations. American Sociological Review 27, 31–41.

Fedor, D., Maslyn, J., Farmer, S., Bettenhausen, K., 2008. The contribution of positive politics to the prediction of employee reactions. Journal of Applied Social Psychology 38, 76–96.

Ferris, G., Adams, G., Kolodinsky, R., Hochwarter, W., Ammeter, A., 2002a. Perceptions of organizational politics: theory and research directions. In: Yammarino, F., Dansereau, F. (Eds.), Research in Multi-Level Issues, The Many Faces of Multi-level Issues, vol. 1. JAI Press/Elsevier Science, Oxford, UK, pp. 179–254.

Ferris, G.R., Hochwarter, W.A., 2011. Organizational politics. In: Zedeck, S. (Ed.), APA Handbook of Industrial and Organizational Psychology, vol. 3. American Psychological Association, Washington, DC, pp. 435–459.

Ferris, G.R., Hochwarter, W.A., Douglas, C., Blass, R., Kolodinsky, R., Treadway, D., 2002b. Social influence processes in organizations and human resources systems. In: Ferris, G., Martocchio, J. (Eds.), Research in Personnel and Human Resources Management. JAI Press/Elsevier Science, Oxford, UK, pp. 65–127.

Ferris, G.R., Treadway, D.C., Brouer, R.L., Munyon, T.P., 2012. Political skill in the organizational sciences. In: Ferris, G.R., Treadway, D.C. (Eds.), Politics in Organizations: Theory and Research Implications. Routledge/Taylor and Francis, New York, pp. 487–528.

Ferris, G.R., Treadway, D.C., Perrewé, P.L., Brouer, R.L., Douglas, C., Lux, S., 2007. Political skill in organizations. Journal of Management 33, 290–320.

French, J.R.P., Raven, B., 1959. The bases of social power. In: Cartwright, D., Zander, A. (Eds.), Group Dynamics. Harper Row, New York, pp. 150–167.

Hochwarter, W.A., 2012. The positive side of organizational politics. In: Ferris, G.R., Treadway, D.C. (Eds.), Politics in Organizations: Theory and Research Implications. Routledge/Taylor and Francis, New York, pp. 27–65.

Hochwarter, W., Ferris, G., Zinko, R., Arnell, B., James, M., 2007. Reputation as a moderator of political behavior – work outcomes relationships: a two-study investigation with convergent results. Journal of Applied Psychology 92, 567–576.

Kipnis, D., Schmidt, S.M., Wilkinson, I., 1980. Intraorganizational influence tactics: explorations in getting one's way. Journal of Applied Psychology 65, 440–452.

Lasswell, H.D., 1936. Politics: Who Gets What, When, How. McGraw-Hill, New York.

Leary, M.R., Kowalski, R.M., 1990. Impression management: a literature review and two-component model. Psychological Bulletin 107, 34–47.

Liden, R.C., Mitchell, T.R., 1988. Ingratiatory behaviors in organizational settings. Academy of Management Review 13, 572–587.

Mintzberg, H., 1983. Power in and around Organizations. Prentice-Hall, Englewood Cliffs, NJ.

Ng, T.W.H., Eby, L.T., Sorensen, K.L., Feldman, D.C., 2005. Predictors of objective and subjective career success: a meta-anlayis. Personnel Psychology 58, 367–408.

Pfeffer, J., 1981. Power in Organizations. Pitman, Boston.

Pfeffer, J., 2010. Power: Why Some People Have It and Others Don't. HarperCollins, New York.

Pierce, J.R., Aguinis, H., 2013. The too-much-of-a-good-thing effect in management. Journal of Management 39, 313–338.

Porter, L.W., September 1976. Organizations as Political Animals. Presidential Address, Division of Industrial-Organizational Psychology, 84th Annual Convention of the American Psychological Association, Washington, DC.

Salancik, G.R., Pfeffer, J., 1974. The bases and use of power in organizational decision making: the case of a university. Administrative Science Quarterly 19, 453–473.

Schein, V., 1977. Individual power and political behavior in organizations: an inadequately explored reality. Academy of Management Review 2, 64–72.

Silvester, J., 2008. The good, the bad and the ugly: politics and politicians at work. International Review of Industrial and Organizational Psychology 23, 107–148.

Tedeschi, J.T., Melburg, V., 1984. Impression management in the organization. In: Bacharach, S.B., Lawler, E.J. (Eds.), Research in the Sociology of Organizations, vol. 3. JAI Press, Greenwich, CT, pp. 31–58.

Triplett, N., 1898. The dynamogenic factors in pacemaking and competition. American Journal of Psychology 9, 507–533.

Practical Reasoning: Philosophical Aspects

RJ Wallace, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Practical reasoning is reasoning that aims to resolve the question of what an agent is to do. This article considers the theoretical commitments of reasoning of this kind, as well as the motivational constraints to which it is subject. How is it possible for reasoning to engage the will? Can we make sense of practical reasoning within a broadly naturalistic conception of agents and the world? What are the distinctive forms that practical reasoning can take?

Practical reasoning is reasoning that aims to resolve the question of what one is to do. Such reasoning is practical in at least two senses. First, it is practical in its subject matter, insofar as it is concerned with human *action*. But it is also practical in its issue, insofar as reflection about action can itself directly move people to act. Reasoning of this kind raises two sets of philosophical problems. One concerns the question of how reflection can succeed in being practical in its issue. What assumptions must be made both about agents and about the processes of reasoning they engage in to make sense of the fact that such reflection can give rise to action? Can justice be done to this fact while preserving the idea that practical reasoning is genuinely a form of reasoning? A second set of issues concerns the content of the standards that are brought to bear in practical reasoning. Which norms for the assessment of action are binding on us as agents? Do these norms provide resources for critical reflection about our ends, or are they exclusively instrumental? Under what conditions do moral norms yield valid standards for reasoning about action? The first set of issues is addressed in the following three sections, while the last two sections cover the second set of issues.

Practical and Theoretical Reasoning

Practical reasoning takes place from a distinctive standpoint. When agents deliberate about action, they think about themselves and their situation in characteristic ways. What are some of the salient features of the practical point of view?

One way to interpret this point of view is to contrast it with the standpoint of theoretical reflection. Theoretical reflection is sometimes understood as reasoning about questions of explanation and prediction. Looking backward to events that have already taken place, it asks why they have occurred; looking forward, it attempts to determine what is going to happen in the future. In these ways, theoretical reflection is concerned with matters of fact and their explanation. Furthermore it treats these issues in impersonal terms that are accessible (in principle) to anyone. The natural and social sciences can be viewed as giving paradigmatic expression to theoretical reasoning, so construed.

Practical reasoning, by contrast, takes a distinctively normative question as its starting point. It typically asks, of a set of alternatives for action none of which has yet been performed, what one ought to do, or what it would be best to do. It is thus concerned not with matters of fact and their explanation, but with matters of *value*, of what is to be done. In practical reasoning, agents attempt to assess and weigh their *reasons for action*, the considerations that speak for and against alternative courses of action that are open to them. Moreover they do this from a distinctively first-personal point of view, one that is defined in terms of a practical predicament in which they find ourselves (individually or collectively – people sometimes reason jointly about what they should do together; *see* Joint Action).

There is, however, a different way of understanding theoretical reasoning, stressing the parallels rather than the differences between the two forms of reflection. According to this interpretation, theoretical reflection too is concerned with a normative rather than a factual question, namely with the question of what one ought to believe. It attempts to answer this normative question by assessing and weighing reasons for belief, the considerations that speak for and against the particular conclusions one might draw about the way the world is. Furthermore, it does this from a standpoint of first-personal reflection: the stance of theoretical reasoning in this sense is the committed stance of the believer, not the stance of detached contemplation of one's beliefs themselves (Moran, 2001). Seen in this way, the contrast between practical and theoretical reflection is essentially a contrast between two different systems of *norms*: those for the regulation of action on the one hand, and those for the regulation of belief on the other.

Theoretical reflection, interpreted along these lines, addresses the considerations that recommend accepting particular claims as to what is or is not the case. That is, it is reasoning with an eye to the truth of propositions, and the reasons for belief in which it deals are considerations that speak in favor of such propositions being true, or worthy of acceptance. Practical reasoning, by contrast, is concerned not with the truth of propositions but with the value of actions. The reasons in which it deals are considerations that speak in favor of particular actions being good, or worthy of performance in some way. This difference in subject matter corresponds to a further difference between the two kinds of reasoning, in respect of their consequences. Theoretical reflection about what one ought to believe produces changes in one's overall set of beliefs, whereas practical reason gives rise to action; it is, as noted earlier, practical not only in its subject matter, but also in its issue.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63063-5

Two observations are in order about this side of practical reasoning. First, the contrast just drawn might suggest that there is a categorial difference in the effects of theoretical and practical reasoning, insofar as the former produces changes in our mental states, whereas the latter gives rise to bodily movements. But this would be misleading. Practical reasoning gives rise not to bodily movements per se, but to intentional actions, which are intelligible as such only to the extent they reflect our mental states. It would thus be more accurate to characterize the consequences of both theoretical and practical reasoning as *attitudes*, noting that the former leads to modifications of our beliefs, whereas the latter leads to modifications of our *intentions* (Harman, 1986; Bratman, 1987).

Second, in neither case do the modifications in question occur infallibly. There is room for *irrationality* both in the theoretical and the practical sphere, which in its strongest form involves a failure to acquire the attitudes that one acknowledges to be called for by the reasons one has reflected on. Thus a person might end up watching another hour of television, while judging that it would be better on the whole to go back to work. Practical irrationality of this latter kind is known as *akrasia*, incontinence, or weakness of will, and its possibility makes clear that practical reasoning is not automatically practical in its issue. A better way to represent the consequences of practical reasoning would be to say that it generates appropriate intentions to the extent an agent is rational (Korsgaard, 1996; *see* Irrationality (Philosophical Aspects)).

Naturalism and Normativity

The connection of practical reasoning with intentional action raises large questions about its credentials as a genuine form of reasoning. As noted earlier, intentional action is not mere bodily movement, but reflects a mental state of the agent, viz. intention. To be in this kind of mental state is to have a goal or aim of some kind, which one seeks to realize through one's action. Intention seems in this respect to be strikingly unlike belief. The propositions that give the content of beliefs have a representative function; they aim to fit the way the world is, so that if one discovers that the world is not how one previously took it to be, one's belief will automatically be modified in the process. With intentions, however, matters are otherwise (Smith, 1987). The intention to go shopping on Tuesday, for instance, is not a state that would or should be abandoned upon discovering that one has not (yet) gone shopping on Tuesday; rather the person with such an intention will ordinarily try to bring the world into alignment with the intention, by going shopping when Tuesday comes around. Intentions are in this way more like blueprints than like sketches of an already completed structure.

Reflection on this contrast between belief and intention has led some philosophers to ask whether practical reasoning might not be something of a misnomer. The concern, in a nutshell, is how processes of thought that count as genuine reasoning could by themselves generate states with the peculiar, nonrepresentative function of intentions. Reasoning seems to consist of cognitive operations, whereas intentions are distinctively noncognitive states, insofar as they do not aim to reflect the way things happen to be in the world.

Expressivist accounts give voice to this skeptical attitude about practical reasoning. These accounts offer interpretations of the normative language that distinctively figures in practical reflection about action. As was seen in the first section, such reflection is an attempt to assess an agent's reasons for acting in one way or another; conclusions about such reasons are characteristically expressed in claims about what it would be *good* to do, or about the actions that one *ought* to perform. According to the expressivist, however, such claims do not represent genuine cognitive achievements, judgments that are literally capable of being true or false. Rather they give expression to desires, sentiments, and pro-attitudes, the kinds of noncognitive state that motivate people to action. Only if normative assertions are understood in this way – so the expressivist contends – can sense be made of the capacity of practical reflection to generate states with the peculiar structure and function of intentions.

Expressivism in this form can be seen as a naturalistic interpretation of practical reasoning, one that is appropriate to the enlightened commitments of the modern scientific world view. It is naturalistic metaphysically, insofar as no assumptions need be made about the objective existence in the world of such questionable entities as values, norms, or reasons for action. If normative claims do not represent genuine cognitive achievements, then their legitimacy does not depend on our postulating a realm of normative facts to which those claims might correspond. It is also naturalistic psychologically, insofar as it promises explanations of intentional human behavior that are basically continuous with explanations of the behavior of non-human animals. In both the human and the non-human cases, behavior is seen as the causal product of noncognitive attitudes, operating in conjunction with the creature's factual representation of how things are in its environment. The difference is just that humans have much more sophisticated linguistic methods for giving voice to their motivating noncognitive attitudes. Indeed, many contemporary expressivists would contend that these expressive resources are sufficiently powerful to explain the features of practical reflection that initially give it the appearance of a genuine form of reasoning (Blackburn, 1998; Gibbard, 1990).

Other philosophers remain unimpressed with these naturalistic explanations. One ground for dissatisfaction with them is the following. The expressivist strategy relies on an initial contrast between practical reflection and the genuine forms of cognitive activity characteristic of theoretical reasoning. There has to be some basic sense in which such discourse does not come up to the standards of authentic cognitive discourse in the literal sense; otherwise the contention that normative discourse is expressive rather than cognitive will lack any significant content. But the contrast between theoretical and practical reflection required for this purpose seems elusive. As was seen in the first section, theoretical reasoning is no less a normative enterprise than practical reasoning. It concerns itself with reasons for belief, the evidence and other considerations that speak for and against particular conclusions about the way things are in the world. In this respect, theoretical and practical reasoning would seem equally problematic from the naturalistic perspective – assuming, that is, that it leaves no place for such normative considerations as reasons. But if *naturalism* calls into question

the credentials of theoretical reason, it thereby undermines the contrast between genuine reasoning and noncognitive normative discourse on which expressivists themselves rely (*see* Emotion and Expression).

Reasons and Motivation

The capacity of practical reasoning to generate intentional action divides even those philosophers who agree in rejecting the expressivist strategy discussed earlier. Such philosophers are prepared to grant that there are reasons for action, and to accept the cognitive credentials of normative claims about such reasons. But they differ in their accounts of the truth conditions of these normative claims. Two approaches may be distinguished.

The first of these, frequently referred to as internalism, holds that reasons for action must be grounded in an agent's prior motivations (Williams, 1981). According to this position, a given agent s can have reason to do x only if x-ing would speak to or further some element in s's 'subjective motivational set.' There must be some rational route that connects s's x-ing with the subjective motivations to which s is actually already a subject; otherwise the claim that s has reason to x must be rejected, as false or incoherent. Behind this internalist position lies the idea that practical reasoning is practical in its issue. Internalists contend that sense can be made of the capacity of such reasoning to generate new intentions only if it is conditioned by motivational resources that are already to hand. Reasoning about action is a matter of working out the implications of the commitments contained in one's subjective motivational set. In this sense, *motivation* is prior to practical reason, and constrains it (though it might function as a background condition of practical reason, rather than a focus of attention for agents when they deliberate about what to do: Schroeder, 2007).

Externalists reject this assumption, contending that one can have reasons for action that are independent of one's prior motivations. Most of them agree that practical reasoning must be capable of generating new intentional actions, and hence of giving rise to motivation. They agree, in other words, that if agent s has reason to do x, it must be possible for s to acquire the motivation to x through reflection on the relevant reasons. But they deny that such reasoning must in any way be constrained by s's subjective motivations prior to the episode of reasoning. Practical reasoning is not exclusively a process of working out the implications of one's existing commitments; it is also an attempt to get clear about what it would objectively be good to do. Normative considerations of this kind are independent of one's prior motivations and can open up new motivational possibilities (Parfit, 1997).

Underlying this dispute are diverging approaches to the explanation of intentional action. Internalists are impressed by the differences between intentions and the cognitive states that figure in paradigmatic examples of theoretical reasoning. Pointing to these differences, they ask how practical reasoning can succeed in producing new intentions if it is not grounded in something of the same basic psychological type: a motivation or desire that is already part of the agent's subjective motivational equipment. Many externalists find this contrast between intentions and cognitive states overdrawn. They observe that we need to postulate basic dispositions of responsiveness to reasons to account for the capacity of theoretical reflection to generate new beliefs, and question why these same dispositions cannot explain the fact that practical reasoning is practical in its issue. Cognitive or not, intentions belong to the broad class of attitudes that are sensitive to normative judgments, and this may account for the capacity of practical reflection to give rise to new intentions (Scanlon, 1998). A third possibility is that intentions are products of the *will*, a capacity for self-determination distinct from the dispositions that render theoretical rationality possible. Depending on how it is developed, the theory of the will may offer a different way of accounting for the practical consequences of practical reflection, without assuming that reasons for action are grounded in an agent's subjective motivations (Wallace, 1999).

Instrumental and Maximizing Rationality

Among the substantive norms of practical reasoning, those of instrumental rationality appear least controversial. *Instrumental rationality*, in its most basic form, instructs agents to take those means that are necessary to achieve their ends. In the modern era, this form of rationality has widely been viewed as the single unproblematic norm of practical reasoning. The instrumental principle makes no assumptions about the prospects for rationally scrutinizing peoples' ends. Rational criticism of this kind would seem to presuppose that there are objective values, standards for assessment of ends that are independent from psychological facts about what people happen to be motivated to pursue. In keeping with the naturalistic attitude sketched in Section Naturalism and Normativity, however, it may be doubted whether such standards can be reconciled with the metaphysical commitments of contemporary scientific practice (*see* Value Pluralism). A world that is shorn of objective values or norms leaves no room for rational criticism of people's ends, but only for Weberian *Zweckrationalität*: the rational determination of means to the realization of ends that are taken as given, as a matter of psychological fact.

It is important to bear in mind, however, that instrumental rationality is itself the expression of an objective normative commitment. The instrumental principle tells us that we have *reason* to take the means that are necessary to achieve our ends; if the principle is valid, then people are open to rational criticism to the extent they fail to exhibit this kind of instrumental consistency, regardless of whether they want to comply with the principle or not. If naturalism really leaves no place for objective norms or values, it may be wondered how an exception can possibly be made for the instrumental norm. The consistent naturalist should perhaps reject even *Zweckrationalität* in favor of a skeptical attitude toward practical reasoning in all its forms (Hampton, 1998). Further questions can be raised about the plausibility of the instrumental norm as a self-standing principle of practical reason. The norm says that one should take the means that are necessary to achieve one's ends. But how can the fact that a given means is necessary, relative to some end, give a person reason to choose the means, if the end is not itself something it would be good to achieve in some

way? The instrumental principle seems to function as a binding standard for practical reasoning only if it is taken for granted that there are additional, independent standards for the assessment of our ends (Korsgaard, 1997).

Many proponents of the instrumental principle would agree that it does not generate reasons for action. The fact that a given means is necessary, relative to one's given ends, is not a reason to take the means. The instrumental principle functions, rather, as a structural requirement on one's attitudes (Broome, 1999). Thus, suppose one intends end E, and believes (truly) that E can be achieved only if one intends to do M. It appears that there are two ways in which one could revise one's attitudes in response to these considerations, compatibly with the instrumental principle: one could form the intention to M, or one could abandon one's original intention to E. The instrumental principle, considered in itself, is indifferent as between these two possibilities; it should be understood as a wide-scope requirement, governing combinations of attitudes, rather than a source of detachable normative conclusions about what one has reason to do. (Modus ponens represents a similar rational requirement in the domain of theoretical reason, governing combinations of beliefs.)

Rational requirements of this kind have recently become a subject of intensive philosophical debate. The idea that there are structural requirements on our attitudes appears to be common ground among philosophers who differ significantly in their views about the nature and scope of practical reason. It is accepted by most Humeans, for instance, who believe that there is no scope for the rational criticism of individual ends, and also by Kantians, who think that the requirements of reason ultimately constrain us to choose in accordance with the moral law. From the perspective of practical and theoretical deliberation, we commonly grant the force of these structural requirements, acknowledging a kind of rational pressure to bring our beliefs and intentions into compliance with the instrumental principle and other standards of consistency and coherence.

Many philosophers take such structural requirements at face value, granting that practical reason is rightly governed by and responsive to these wide-scope demands. But what is the relation between structural requirements of this kind and our reasons for action and belief? One view, held in common by Humeans and by some Kantian constructivists (see Section Naturalism and Normativity), is that reasons are fundamentally derivative from rational requirements. What one has reason to do, on this view, is what one would desire or intend to do if one was fully rational (i.e., fully in compliance with the wide-scope structural requirements that govern one's attitudes in combination).

For those who do not share this reductionist view, however, the status of rational requirements becomes more puzzling. One might hold that practical reason is ultimately answerable to two different kinds of constraints: to rational requirements on the one hand, and to independent facts about what one has reason to do on the other hand. But this position is potentially unstable. Once the independence of structural requirements from normative reasons is made clear, it is no longer obvious why we should care about whether our attitudes do or do not comply with the structural requirements (Raz, 2005). On this view, there is nothing wrong with failing to take the necessary means to your end, unless the end itself is one that you have compelling reason to pursue. More generally, it has been argued that there are no independent requirements of structural rationality at all, and that the appearance of such requirements within the deliberative point of view can be explained by substantive features of the reasons to which both practical and theoretical reason are ultimately and properly responsive (Kolodny, 2005).

Humean proponents of structural approaches to practical reason have attempted to make room for the rational criticism of individual ends, without departing from the spirit of *Zweckrationalität*, by expanding their view to encompass the totality of an agent's ends. Thus, even if there are no values that are ultimately independent of an agent's given ends, it would appear possible to criticize particular intrinsic desires by reference to others in the agent's subjective motivational set. An agent's desire for leisure, for instance, might be subordinated insofar as its satisfaction would frustrate the realization of other goals that are subjectively more important to the agent. Practical reasoning, it might be suggested, is a holistic enterprise, concerned not merely with the specification of means to the realization of individual ends, but with coordinated achievement of the totality of an agent's ends.

Many philosophers view this holistic perspective as the most fruitful way of conceptualizing the tasks of practical reasoning. It defines an important and difficult problem for practical reasoning to solve, without departing from the metaphysically modest assumption that there is no court of appeal for the rational criticism of an agent's ends that is independent of those ends themselves. The most sophisticated and influential expression of this holistic perspective is the *maximizing* conception of rationality. According to the maximizing conception, the fundamental aim of practical reasoning is to determine which course of action would optimally advance the agent's complete set of ends. Thus it has been suggested that the rational action is the one whose subjective expected *utility* – reflecting both the utility of possible outcomes, from the agent's point of view, and the agent's beliefs about the probability of those outcomes – is the highest.

The maximizing conception of rationality finds its most complete expression in *decision theory* and in the theory of *rational choice* (as developed, for instance, in economics). These disciplines articulate with mathematical precision the assumption that one is acting rationally to the extent one does what is likely to bring about the best state of affairs, given both one's preferences and one's beliefs about the probable consequences of the alternatives that are open to one. Proponents of these theories often claim for them the additional advantage of empirical adequacy, contending that they are flexible enough to accommodate the full range of behaviors that human agents engage in, both within the marketplace and outside of it. Especially if one operates with the notion of 'revealed preferences' – preferences, that is, that are inferred solely on the basis of actual behavior – then virtually anything an agent might choose to do could be interpreted as an attempt to maximize expected utility. Decision theory, on this account of it, becomes an all-encompassing framework for the interpretation of human behavior, according to which agents are always striving to produce outcomes that would be optimal, relative to their

preferences and beliefs (*see* Rational Choice Explanation: Philosophical Aspects).

If decision theory is interpreted in this way, however, then its relevance to the understanding of practical reasoning begins to seem correspondingly tenuous. Maximizing rationality initially looks to be a norm of practical reflection. In this guise, its appeal lies in the idea that there can be rational requirements on action, stemming from the totality of an agent's preferences and beliefs, even if it is not assumed that there are independent, substantive standards for the critical assessment of individual ends. But this normative understanding of maximizing rationality is tenable only if it is at least conceivable that individual agents might occasionally fail to comply with its requirements – an 'ought' that it is not so much as possible to flout is not really an 'ought' at all. Proponents of the maximizing approach need to be attentive to the latent tension between the empirical and the normative aspirations of their theory, for maximizing rationality cannot claim both perfect empirical adequacy and normative authority for practical deliberation.

Assuming this tension can be resolved, further questions arise about the plausibility of the normative requirement to maximize expected utility. Doubts have been raised, for instance, in regard to the assumption that it is necessarily irrational to fail to select that action that would be optimal, relative to one's preferences and beliefs. Perfectly rational agents often seem to be content with states of affairs that are 'good enough,' from the perspective of their aims and desires, even when they are aware of alternatives that promise a higher return; they 'satisfice,' rather than seeking to maximize the value of the outcomes that they bring about (Slote, 1989). They also treat their past intentions and plans as fixed constraints on their decision-making, rather than attempting to maximize subjective utility anew in every situation they confront (Bratman, 1987). Some contend that maximizing rationality is flexible enough that it can accommodate alleged counterexamples of these kinds (Pettit, 1984). If not, however, there may be grounds for doubting that it represents a fundamental and uncontroversial norm for reasoning about action.

Specification, Expressive Rationality, and Moral Reasons

If maximizing rationality is not the unproblematic requirement of practical reasoning that it initially seemed to be, what are the alternatives to it? Consider the assumption that critical assessment of an agent's individual ends is off-limits. This apparently truistic assumption has been questioned by some philosophers, who point out that a great many of one's aims in life are rather inchoate; people want, for instance, to be successful in their careers, and loyal to their friends, without being entirely clear about what these ends require of them. To the extent one's ends are indefinite in this way, they will not provide satisfactory starting points for instrumental, maximizing, or even satisficing reflection. One will need to specify such ends more precisely before one can begin to think about which means they require one to pursue, or to generate from them a rank-ordering of possible outcomes. Here is a possible task for practical reasoning that does not fit neatly into the categories of *Zweckrationalität*, however broadly construed (Wiggins, 1987; Richardson, 1994).

Practical reflection about ends is not an easy or well-defined activity. There are no straightforward criteria for success in this kind of reflection, and it is not even clear when it has been brought to a satisfactory conclusion. These considerations are no doubt partially responsible for the assumption – especially widespread in the social sciences – that there is no reasoning about final ends. On the other hand, how is one supposed to clarify one's largest and most important ends, if not by reasoning about them in some way? Rather than exclude such reflection because it does not conform to a prior and possibly inappropriate paradigm of proper reasoning, perhaps the conception of practical reasoning should be expanded to make room for clarificatory reflection about the ends of action. To do so would be to acknowledge that practical reasoning can be both objective and subjective at the same time; by working out the meaning and implications of such commitments as loyalty or success, for instance, agents also help to get clear the values that define who they are (Taylor, 1985; *see* Hermeneutics).

A different point at which to question received models of practical reasoning is the theory of value to which they are committed. Maximizing and satisficing models alike seem committed to a consequentialist account of the relation of action to value: the thesis that value inheres ultimately in states of affairs, and that actions are to be judged rational to the extent they produce valuable states of affairs. This thesis seems at odds, however, with intuitive ways of thinking about value. It appears a distorted understanding of friendship, for instance, to say that what friends value fundamentally are states of affairs (involving, say, joint activities with the friend); what people value as friends are rather the concrete persons with whom they are befriended. This suggests a different way of conceptualizing the rationality of action. What makes actions rational, it has been claimed, is not the production of valuable states of affairs, but rather their success in expressing the attitudes that it is rational to adopt toward the true bearers of intrinsic value: people, animals, and things (Anderson, 1993). A supposed advantage of this approach is its ability to explain the rationality of behaviors that seem intuitively sensible, but that are hard to fit into the consequentialist scheme (such as commitments deriving from one's past involvement in an activity or project, which can look like an irrational weighing of 'sunk costs' to the consequentialist; compare Nozick, 1993).

Morality provides an especially fertile source of examples and problems for the theory of practical reasoning. One of the defining questions of moral philosophy is the question of the rational authority of moral norms: to what extent, and under what conditions, are people rationally required to comply with the demands of conventional morality? Reflection on this question has produced some of the most far-reaching and illuminating philosophical work on the nature of practical reasoning. Two divergent tendencies within this body of work should be singled out. Some approaches to

moral reasoning proceed by relating it to patterns of reflection familiar from other, non-moral domains, particularly the maximizing patterns canvassed in Section Instrumental and Maximizing Rationality. Thus it has been argued that, though morality imposes constraints on the direct pursuit of individual utility, these constraints can be justified in terms of the values expressed by ordinary economic rationality; a strategy of constrained maximization is recommended on grounds of enlightened self-interest, and this in turn accounts for the authority of moral norms to govern the practical reflection of individuals (Gauthier, 1986). Other philosophers have sought to make sense of morality as a source of rational norms by assimilating it more directly to the maximizing conception. For instance, proponents of *consequentialism* as an ethical theory – the theory that agents are morally required to maximize the good, impartially conceived – consider it an advantage of this account that it represents moral requirements in terms of the maximizing conception of rationality familiar from other contexts (Harsanyi, 1982; *see* Consequentialism Including Utilitarianism; Economics and Ethics).

A different approach stresses the discontinuities between moral and non-moral patterns of reasoning. According to this approach, morality is a source of constraints (such as prohibitions on murder and deception) that cannot be represented accurately within the framework of maximizing rationality (e.g., Scanlon, 1998). If this is correct, then those who wish to make sense of moral requirements, as norms that appropriately govern the reflection of individual agents, will need to expand their conception of the forms and possibilities of practical reasoning (*see* Social Work Ethics; Free Will and Action).

See also: Consequentialism (Including Utilitarianism); Irrationality: Philosophical Aspects; Naturalism; Norms.

Bibliography

Anderson, E., 1993. Value in Ethics and Economics. Harvard University Press, Cambridge, MA.

Blackburn, S., 1998. Ruling Passions. Clarendon Press, Oxford.

Bratman, M., 1987. Intention, Plans, and Practical Reason. Harvard University Press, Cambridge, MA.

Broome, J., 1999. Normative requirements. Ratio 12, 398–419.

Gauthier, D., 1986. Morals by Agreement. Clarendon Press, Oxford.

Gibbard, A., 1990. Wise Choices, Apt Feelings. Harvard University Press, Cambridge, MA.

Hampton, J., 1998. The Authority of Reason. Cambridge University Press, Cambridge.

Harman, G., 1986. Change in View. MIT Press, Cambridge, MA.

Harsanyi, J.C., 1982. Morality and the theory of rational behavior. In: Sen, A., Williams, B. (Eds.), Utilitarianism and Beyond. Cambridge University Press, Cambridge.

Kolodny, N., 2005. Why be rational? Mind 114, 509–563.

Korsgaard, C., 1996. Skepticism about practical reason. In: Korsgaard, C. (Ed.), Creating the Kingdom of Ends. Cambridge University Press, Cambridge.

Korsgaard, C., 1997. The normativity of instrumental reason. In: Cullity, G., Gaut, B. (Eds.), Ethics and Practical Reason. Clarendon Press, Oxford.

Moran, R., 2001. Authority and Estrangement: An Essay on Self-Knowledge. Princeton University Press, Princeton, NJ.

Nozick, R., 1993. The Nature of Rationality. Princeton University Press, Princeton, NJ.

Parfit, D., 1997. Reasons and Motivation, vol. 77. The Aristotelian Society (Suppl.), 99–130.

Pettit, P., 1984. Satisficing Consequentialism, vol. 58. The Aristotelian Society (Suppl.), 165–176.

Raz, J., 2005. The myth of instrumental rationality. Journal of Ethics and Social Philosophy 1. http://www.jesp.org/PDF/6863_Raz-vol-1-no-1-rev.pdf.

Richardson, H.S., 1994. Practical Reasoning about Final Ends. Cambridge University Press, Cambridge.

Scanlon, T.M., 1998. What We Owe to Each Other. Harvard University Press, Cambridge, MA.

Schroeder, M., 2007. Slaves of the Passions. Oxford University Press, Oxford.

Slote, M., 1989. Beyond Optimizing. Harvard University Press, Cambridge, MA.

Smith, M., 1987. The humean theory of motivation. Mind 96, 36–61.

Taylor, C., 1985. What is human agency? In: Taylor, C. (Ed.), Human Agency and Language. Cambridge University Press, Cambridge.

Wallace, R.J., 1999. Three conceptions of rational agency. Ethical Theory and Moral Practice 2, 217–242.

Wiggins, D., 1987. Deliberation and practical reason. In: Wiggins, D. (Ed.), Needs, Values, Truth. Blackwell, Oxford.

Williams, B., 1981. Internal and external reasons. In: Williams, B. (Ed.), Moral Luck. Cambridge University Press, Cambridge.

Practice: An Anthropological Approach

Jane Fajans, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Kemper, volume 17, pp. 11945–11948, © 2001, Elsevier Ltd.

Abstract

Practice is a methodological and theoretical frame for investigating the relations between social individuals and the social and cultural milieux in which they live and work. 'Practice' is activity performed to produce an effect that is known or envisioned in advance by the performer. Practice is not the rote reproduction of social and cultural life. Praxis, rather than practice, offers a new dimension by examining how individuals are socialized into the roles and practices of their environment but also adapt the practices they have learned to the situations they confront. Such a perspective recognizes social and cultural change. Practice theory enables a dynamic and complex analysis of multiple social and cultural situations.

Introduction

Practice is a methodological and theoretical frame for investigating the relations between social individuals and the social and cultural milieux in which they are born, live, and work. It examines the actions and interactions that people engage in and how these activities produce results in the social world while simultaneously engendering the skills and subjectivity of the actor. The cumulative effect of these outcomes becomes the ongoing manifestations of what anthropologists and sociologists call social structure and culture.

For years, anthropology and anthropologists have studied the social activities comprising the minutiae of everyday life, as Malinowski has said, in order to abstract the principles of social organization, cultural beliefs and ideas, and the roles and statuses of community participants. The importance of such observations has never really been questioned, but the importance of the phenomena that were being observed was not always appreciated in its fullest dimensions. Anthropologists engaged in trying to understand not just the forms of culture and society but also the processes that produce and maintain them have looked at all manners of social and cultural phenomena to investigate these processes. These diverse activities have been credited with transforming individuals into social actors and binding these actors into social networks.

A theory of practice goes beyond observation and description of the activities that people perform, both daily and on special occasions, to look at "the integration of social organization, social action, and the production of meaning – the ways in which social processes are turned by practical activity into cultural forms and in turn inform the improvisation of social practices" (Calhoun and Sennett, 2007: p. 5). Thus, practice investigates how people's behavioral practices constitute productive processes that simultaneously and dialectically produce both the actors and the social and cultural networks in which they act. This intertwined realm includes construction and enrichment of both the individual aspects of motivation and desire, and the social complements of those features such as social values and cultural desires. In this way, practice theory seeks to overcome the mind–body dichotomy as well as the individual–collective divide. In both instances, the divides are both mediated and joined by their cocreation through ongoing activity.

'Practice' is activity performed in order to produce an effect known or envisioned in advance by the performer. The performance of the action in question serves reflexively as the objectification of this intention. In other words, the proper performance of the activity is the purpose, as for example in practicing the playing of a musical instrument. Such an act may be conceived as serving to improve one's instrumental technique or one's ability to perform more perfectly a particular piece of music with the instrument in question. Such practice has a dual effect. The performance of the act in the prescribed manner is essential to its purpose, but each enactment reflexively changes the way the musician performs the piece in an embodied and subjective way. At the same time, the rendition of the piece is a phenomenon that stimulates new understanding, aspirations, and desires in both the actor and the potential audience.

Just as earlier anthropological studies focused on an enormous array of social phenomena, practice-oriented anthropology looks at widely disparate types of phenomena in which actors perform processes that place them in social and cultural situations and networks and affect the outcomes of events in the world. Contemporary theorists of practice theory in anthropology have applied the insights to an enormous variety of social and cultural phenomena spanning subjects such as gardening and socializing children (Fajans, 1997), weaving (Greenfield), investment banking (Ho), cooking (Sutton) cheese making (Paxson), chocolate making (Terrio), producing the state (Byrnes), and surviving poverty in Oaxaca, Mexico (Mahar).

This convergence of one theoretical application to such a wide variety of topics is not really surprising because anthropology has always investigated a plethora of subjects, and many theorists have argued that these approaches are generally applicable to a large array of phenomena. The study of the social and cultural dimensions of practice is therefore part of the genealogy of anthropology over more than a century. The theoretical implications of these newer studies, however, have provided an added dimension.

Bourdieu and the Theory of Practice

The flowering of interest in practice theory arose after the publication of Pierre Bourdieu's *Outline of a Theory of Practice* in 1972 (English translation, 1977). What made Bourdieu's project different was that he made practice part of a theoretical analysis rather than just a component of the observation and participation that previous anthropologists had engaged in. He sought to find regularities in people's behavior that could be described both as part of a statistical analysis and as phenomena that went beyond what was attributed to be either in people's heads or in the world outside. He saw practices as uniting what is customary and strategic, as well as what is routinized and improvisatory. Individuals grow up in cultural settings where certain behaviors are taken for granted; new members learn them through implicit rather than explicit learning. Such activities are undertaken in (relatively) unconscious ways and are understood as the necessary foundations of social life. Bourdieu calls such routinized and unquestioned behavior and beliefs 'doxa.' Doxa is by definition unquestioned. What turns doxa into either orthodoxy or heterodoxy is the comparison of one implicit habitus with another. Bourdieu does not have a means of introducing change – even small adaptive change – into his notion of habitus. However, he believed such activities have to be understood in a temporal framework, taking into account variations in performance over seasons, life cycles, and changing social situations (Jenkins, 1992: p. 68). Bourdieu looks at patterns and practices over time to see how some aspects of the habitus vary in cyclical or oscillating patterns that differ from night to day, summer to winter, male to female, or old to young, or in many other ways. In these variations, he suggests, may be found the seeds of change and the grounds of adaptation to new practices.

Despite attributing greater importance to everyday activities as the building blocks of social and cultural life, practice theory often presupposes that actors perform their activities in relatively unconscious (or not wholly conscious) ways. This aspect of the theory is embedded in the idea of habitus. Other anthropologists have used the term 'habitus,' especially Mauss and Durkheim, but Bourdieu elaborated on the notion of habitus as a concept to surpass the mind–body dualism of received philosophy, and to act as a bridge between individual free will and socially determined structures of conformity. Habitus is a set of bodily practices, which Bourdieu calls "an acquired set of generative schemes objectively adjusted to the particular conditions in which it is constituted" (1977: p. 95). These schemes are a system of dispositions, which are both mental and bodily practices because they are in the heads of people but are also enactments of dispositions. When followed, these acts or practices have effects in the world that essentially produce a socially recognized and constructed world.

Bourdieu begs the question of how the central patterns of *habitus* are learned and, once learned, become motivated, so that the acting subject is led to repeat them in the same form. How this process of reproduction occurs remains unspecified, with terms like internalization, incorporation, and so on. But to move from statistically regular patterns to socially and culturally motivated ones requires an understanding of social meaning and social value. These qualities motivate actors to participate as social beings in cultural and social milieux and reciprocally make this participation meaningful. Action in this sense mediates between the subjective individual and the external social and cultural relations, structures, and activities that form the visible manifestation of both culture and social organization. Motivation in this conceptualization is not intrinsically linked to competition for status or dominance, but to the fulfillment of needs and desires that have culturally specific meanings. Bourdieu addresses these culturally specific values through his terms 'social capital' and 'symbolic capital,' but he is less clear on how subjects come to value and desire the particular capital of their social group. These values are established through repeated (although not necessarily identical) actions and interactions in complex social encounters. Since most activities are social, the practices most frequently examined by anthropologists construct the nexus of interpersonal social relations as they also produce an acting subject.

The significance of these social building activities is not always recognized. In much of Bourdieu's work, subjectivity, agency, consciousness, cultural symbolism, and meaning are left as abstract phenomena, primarily attributed as reflexive determinations of a self-reproducing social structure and the competition for relative distinction within it. The intricate specifics of the practices and actors in real situations are lost in the theoretical abstraction. This absence leaves the theory seeming more tautological than engaged. As Bourdieu says, doxa is the quasiperfect fit between the objective structures and internalized structures of agents that makes the structures in question appear natural and unquestioned.

Such a perspective cannot account for change. Although Bourdieu speaks of improvisation, he gives little evidence of how improvisation emerges or whether it can effect change. Others try to account for change between a subject and the external world by analyzing multiple embodied processes such as the senses that are both cognitive and bodily activities through which actors respond to their environment. In most cases, such cognitive and sensory responses cause minor alterations in social and behavioral relations, but these changes and adjustments open the possibility of more radical change and progressive evolution. The idea of social stasis is no longer an acceptable assumption, but neither is the idea that all things are possible. Cultures evolve and develop, and the expectations of behavior within them are socially valued and judged.

A More Inclusive Theory

An interest in the relationship between human action and social transformation can be traced back quite a long way, certainly back to Marx and Weber. But since then, oscillations between overly structural and overly actor-focusing frameworks have dominated perspectives (Ortner, 1989: p. 15). Practice has too frequently been associated with the rote reproduction of social and cultural life through the dialectic of action as production. Praxis, rather than practice (Marx, 1947), offers a new dimension to this more mechanical model. Praxis examines the way that activity is not just externally directed but also simultaneously part of the production of the subjectivity and consciousness of the actor (as in the

musician practicing an instrument) as it manifests an effect in the external world as well. Contemporary theories attempt to add a variety of theoretical insights to this theory. In particular, to understand the ways that embodied practices are not just rote and repetitive actions, theorists look to two other theoretical schools: Soviet and Eastern European activity theory and Piagetian social psychology. These add dimension to the development of practice theory.

The theoretical side of practice 'theory' shows strong convergence with activity theory. Activity theory, developed in the former Soviet Union and other Eastern European nations, derives from Marx's notion of *praxis* (Vygotsky, 1962, 1978). The theory of praxis from which this theoretical school developed uses a different definition of praxis than that most commonly used in the United States and Western Europe (cf. Sahlins, 1981; Levi-Strauss, 1966).

Praxis in this derivation refers to the form of action that is built upon a dialectical interaction between a subject and an external phenomenon, object, person, or idea. In this view, action is prior to either the subject or the object, which are considered abstractly or separately (prior in every sense: epistemological, experiential, developmental, practical, historical, and theoretical). Each action is reflexive: in acting on the object, the subject reflexively acts upon itself, is acted upon, and is affected by the object. Thus, praxis is simultaneously subjective and objective, material and ideal. It involves a subjective purpose and an objective constraint. Praxis refers to processes that are objectively and materially existent but are also the expression of subjective needs, ideas, and purposes. It is neither solely subjective nor wholly objective; rather, it is cultural because it is symbolically mediated. This may mean that the forms of the subject or object are constituted as cultural categories, but they are also mediated by special symbolic language (e.g., language, money, or special forms of exchange). These symbolic forms of praxis are the basis of social consciousness and cultural forms.

The boundaries usually constructed around the concepts of subject and object, internal and external, and entity and action are in this context part of a continuous relationship in which each part produces and in turn is produced by the others. In this relationship, activity holds the determinative role. The activity mediates internal-subjective states and motivations and external outcomes and relationships.

This dialectal relationship is expressed in other ways by other theorists. In particular, Jean Piaget builds both a psychological and social theory on the notion of schema. A schema is the form of a relationship between an acting subject and a portion of an objective external domain. It is based on a subject recognizing stimuli such as touch, language, events, and social interactions and learning appropriate responses to such events. The process of receiving and recognizing these stimuli is what Piaget calls 'assimilation.' The actor responds to such experiences in evolving ways, learning a more and more acceptable pattern of response. Piaget calls this response 'accommodation.' The response is a learned behavior, but it is not an instinctive one. When stimuli change and internal responses vary, so does the reaction of the subject. In this way, schemas contain the capacity to change and adapt to new contexts and situations. The concept of activity is broken into several different components, condensed as action and reaction, in which each is part of a feedback relationship that is fluid and potentially changing. The relation as a whole is the schema, and the actors and actions form a constantly evolving dynamic interface.

Change in both this and other senses has been a problem for many social scientists to account for or explain. The more successfully they describe the relationship between actors and their sociocultural context, the more difficult it is to imagine change arising from within this dialectical system. Practice theorists such as Ortner confront the way that cultural schemas sustain a 'kind of durability' (1989: p. 61). For this reason, many social theories have fallen back on change as requiring external forces. Ortner and other practice theorists have tried to refute the necessity for such outside intervention, believing and illustrating how culturally constituted engagements have a way of 'escaping' those intentions (Ortner, 1989: p. 89). Thus social variability, which is ever present in such common forms as age, class, and gender as well as ritual and mythic framings of subjectivity and motivation, can, in the context of differential experiences, initiate change.

As Lave and others (Lave and Wenger, 1991) have shown and argued in discussions of learning in both formal and informal settings, people do not learn in isolation but in the context of relations with others, so learning is social and depends on the situated contexts in which it occurs. The context of participation in both learning and practicing skills and knowledge is one of both reproduction and potential creative production. Lave is paradoxically trying to create a theoretical conception of learning without engaging in a process of abstraction.

This paradox is a key to thinking about practice as a fundamentally different understanding of the world: one in which activity is both the content and the structure of the phenomenon and is both the evidence and the theory of which it consists. Lave and Wenger "shift the focus from the individual as learner to learning as participation in the social world, and from the concept of cognitive process to the more encompassing view of social practice" (Lave and Wenger, 1991: p. 43). Viewing learning in this social context evolves from the theories of the school of social psychology led by Vygotsky. Vygotsky developed an approach to learning that "is characterized by the distance between the problem-solving abilities exhibited by a learner working alone and that learner's problem-solving abilities when assisted or collaborating with more-experienced people" (Lave and Wenger, 1991: p. 48). They call this a 'scaffolding approach,' in which learning is not so much a process of internalization as a process of social transformation: "In contrast with learning as internalization, learning as increasing participation in communities of practice concerns the whole person acting in the world" (Lave and Wenger, 1991: p. 49). In contrast with other approaches, this perspective focuses on a "continually evolving, continuously renewed set of relations" (Lave and Wenger, 1991: p. 50). "Because of the contradictory nature of collective social practices, and because learning processes are part of the working out of these contradictions in practice, social reproduction implies the renewed construction of resolutions to underlying conflicts" (Lave and Wenger, 1991: p. 58).

Practice includes the activities that people perform on both a daily and occasional basis. However, the deeper study of practice goes beyond the mere acts of daily life to look at the motivations, desires, and values that actors hold in relation to their acts and the ways in which their performance further instills such beliefs and values.

As Calhoun and Sennett say, "As practice, culture is neither Bagehot's 'hard cake of custom' nor the realm of pure creativity idealized in notions of genius. It is made and remade in almost imperceptible small ways as well as occasional large bursts of innovation. As practice, culture is an achievement, not simply an environment" (Calhoun and Sennett, 2007: p. 7).

To attend to practice is indeed to attend to the ways in which human actions are embedded in social organization. Its effectiveness is made possible, as Bourdieu shows, partly by the internalization of social relationships and beliefs through experience. New actions are produced in structured improvisations, informed by intuitions about current social circumstances, habits and capabilities drawn from previous ones, and cultural schemas. Culture, in this sense, appears as an abstract system of norms or knowledge and also as implicit, often embodied understandings of how things work. Practice examines the nexus of individuals, social groups, communities, and institutions and cannot be fully understood by looking at one element in isolation. Even monolithic institutions are composites of the workers who enact the practices everyday.

Anthropologists must look at the ways that activity is embedded in social situations as well as how such activities also produce social situations and relations. One way to see such processes in their most evident manifestations is to look at the processes of acquiring skills and knowledge. Such contexts of learning span a vast array of activities, from gardening to weaving, and from cheese making to investment banking. Many studies emphasize that learning is a complex and dialectical process that involves a good deal more than just knowing the 'rules of the game.' Although Bourdieu is particularly interested in these relatively unconscious patterns of learning which are a part of habitus, many analysts feel that Bourdieu's concentration on unconscious practices ignores many of the important aspects of social life. Often, the process of acquiring these skills evolves from knowing how to follow a set of rules to being able to improvise to achieve an end even when steps have to be varied or eliminated in the process. Even more significant is the ability to read the environment to judge the next step in processes as different as chopping wood, making cheese, or making stock trades. As Piaget proposes, the subject assimilates information and experience through a series of life encounters and uses these experiences to construct a schema of appropriate actions and reactions. When put in a situation requiring response, the subject-actor uses these experiences to act appropriately and adjust his or her actions to the particularities of the situation. In this frame, individuals are socialized into the roles and practices of their environment and also adapt the practices they have learned to the situations they confront. Situated practice is the ability to adjust to new situations in ways appropriate not just to the end desired but also to the community at large (Lave). The other side of this adaptive ability is the skill to alter the situation to fit the actor's desired outcomes. In this way, the study of practice and activity offers a way to understand both how individuals come to fit the cultural models and expectations of a person and how factors such as age, gender, class, race, and ethnic modes of behavior modify these expectations. The sum of such activity perpetuates the structure of the community in which they live, but at the same time opens the possibility of (subtle) change and alteration that emerges from each instance of practice.

Bourdieu's focus is on the visible world of social practice, but more recent practice theorists have looked at the processes in less material realms, including the analysis of myth and history (Ortner, 1989) and finance (Ho, 2009). Others have examined the activities of petty bureaucrats as the building blocks in the construction of large state apparatuses (Byrnes). Ho uses Bourdieu's ideas of disposition and habitus to discuss the practices of investment banking culture and the ways that these, "plus ways of being, inclination, and predisposition of the body constitute a collective habitus and system of dispositions, which organize action, produce practices, and construct social structures and worlds" (Bourdieu in Ho, 2009: p. 11). In this context, the dispositions go beyond what earlier analysis investigated and examine the production of the subjectivity and habitus of investment bankers and how they perform the acts of being elite and powerful. Power gives them the sense that they can be risk takers. Cumulatively, these actions affect the larger industry culture.

As Ortner says, practice is an activity in a political sense and is linked to structure: "The structural questions here normally concern the 'internalization' (if one takes an actor-centered perspective) or 'reproduction' (if one takes a system-centered perspective) of structure through practice" (1989: p. 194).

Most practice theorists see these processes from the actor's point of view, but they are not rules in the conventional sense. As Bourdieu insists, these activities are conceptually bound patterns of action formulated from the actor's subjective point of view. If one takes a broader view, however, activities are parts of schemas that connect subjects with the objective world through activity. They are schemas as defined by Piaget (1962) and Fajans (1997).

This more 'bird's-eye' view also allows a temporal perspective. As Ortner says, "a theory of practice is a theory of history. It is a theory of how social beings with their diverse motives and diverse intentions make and transform the world in which they live. It is a theory that allows social and cultural analysts to put all their various methodological tools to work – ethnographic and historical research; structural, interpretive, and 'objectivist' analytical approaches – in ways that enhance and enrich the effectiveness of each" (1989: p. 193).

The Many Forms of Activity

To understand how anthropologists approach and use all of these tools in practice theory, one must examine the range of contexts and analyses that emanate from this theoretical perspective.

Fajans applies practice theory to uncover the underlying schema of the Baining of Papua New Guinea. She looks at the many aspects of Baining social life that are premised on an overarching set of values that the Baining achieve through a range of activities. Although appearing as disparate activities,

these different processes produce similarly valued ends. These activities include the transformations that human work produces in such diverse contexts as swidden gardening, feeding children, and constructing dance masks. Each of these activities is, for them, a transformative process in which natural entities, the forest, an unsocialized child, and unconnected leaves and bark are transformed into a social product that itself has productive capacities (Fajans, 1997).

In a quite different social context, Paxson (2012) describes the many techniques of making artisanal cheese. Although the basic steps of cheese making are routinized, the cheese makers she describes are constantly monitoring and reading the environment of the cheese room and the progress of the transformation from milk to cheese. Paxson follows the cheese makers' own descriptions of their activity, which they understand as both art and science. These producers are constantly reading the environment of their dairy and adjusting their activities according to temperature, humidity, the speed of curd formation, and so on. The cheese makers know the process but adjust the minutiae of the activity in accordance with the actual development of the curd, which they come to know both cognitively and bodily. Each day might be subtly different from the day before and after. They have a feel for the game, but they are open to shifts in practice. Sometimes, their acts produce something quite new and different. This can be an exciting new creation or a problematic failure, but either way they are modifying and changing the practice in a fairly continuous way. This understanding of practice instantiates the notion of skill proposed by Ingold (1983) in which skill is performed as an extension of the mind and body in interaction with an environment that changes in response to the actions of the subject. The art and science of cheese making comprise a good example of practice as habituated by adaptive and nonrote activity as described here.

In a related example, Sutton (2012: p. 302) studies how people learn to cook. Cooking in much of the world is an activity that occurs frequently, in close proximity to many other activities, and in spaces frequented by members of a household. In such contexts, it is hard to separate a context in which learning to cook is separate from other social engagements and activities. In this respect, cooking forms part of the habitus, an accepted and relatively unconscious aspect of day-to-day life. Although this is an activity in which individuals are immersed from early childhood, young members of the household are only gradually instilled with the knowledge not only to understand what others are doing but also to do it themselves. In many contexts, the actual practice of the art of cooking may be restricted to particular genders, classes, or professions. Such learning may be subtle, with the assignment of progressively harder chores, and it may appear effortless, but it may also occur with specific instruction and potential criticism when not done properly. Much of this learning and teaching is a form of apprenticeship, learning from one who knows the skills to be acquired. Apprenticeship can range from quasiformal instruction in a task to much more subtle acquisition of a skill through long-term proximity. As Lave and Wenger (1991) describe, learning in a social setting often stimulates students to achieve greater mastery than they would on their own. Much traditional learning is social in this way. In cooking, however, as in cheese making, the end product of the activity must have certain characteristics. Learning to cook requires knowing how particular foods or dishes should look, smell, and taste (Sutton, 2012). Just performing the skills of cooking may not produce the desired ends.

Cooking shares with more formal activities the need to learn certain basic skills and procedures, but it also requires the ability to improvise and adapt. In many apprenticeship situations, however, the standard requires the novice to first learn the skill as practiced by the more authoritative members of the group, and only after proving that ability would improvisation be permitted. Terrio (2000) describes the rigorous requirements to become a chocolatier in France, but once initiated into the guild, practitioners vie to outdo their peers in their production. Again, the presence of a community stimulates performance. Even in this case, however, there are constraints. Actors have to conform to certain cultural and social criteria that define authentic and proper behavior.

Practice theory has also been used to examine the ways that urban squatters adjust to and make sense of their lives in new settings, that is, how they reconfigure the relations between structures and habitus. Anthropologist Cheleen Mahar examines these adaptations by looking at how migrants use memory and imagination to interpret their social conditions and subjective experience. Because memories are both individual and collective, the uses they make of memory are both subjective and social. In addition, the stories they tell and the events they experience are perceived through a process that is both social and corporeal (Mahar, 2010: p. 4).

Recently, anthropologists have looked at activity in more abstract terms. Analyses of social actors who perform various tasks like stock transactions and petty bureaucratic jobs offer a new understanding of the ways that day-to-day activities produce larger social formations and structures that appear external to the actions of individuals but are dialectically imbricated in their actions. Byrnes has elucidated how regional representatives of the state bureaucracy for monitoring manufacturing and industry embody and enact the state in their encounters with local workers and businesspeople. In these instances, the seemingly small and somewhat insignificant interactions nonetheless instantiate the relations of power and authority in the region. The monolithic state is continuously produced through such mundane interactions.

Similarly, Ho (2009) analyzes the Wall Street banking industry through the practices and habitus of its investment bankers. These employees enact a particular habitus of hard work, knowledgeable understanding, and intelligent risk taking that, when combined, defines a set of values that distinguish the banking sector in which they work. They simultaneously perform these roles and are motivated to embody these qualities in order to appear potent in their jobs.

Conclusion

Practice is both a form of action and a means of analysis; it encompasses both description and theory. Attending to practice means examining the coproduction of individual and sociocultural domains through multiple acts of speech,

performance, kinship, and so on. In each of these, the activity is a manifestation of an intention or desire, an effect in social relations, and a means of transforming one's relation within the intersubjective world in which one participates. Activity in this sense is a productive behavior, an expression of a relationship, and a moment in an ongoing dynamic of production. Practice theory enables a dynamic and complex analysis of multiple social and cultural situations.

See also: Apprenticeship, Anthropological Aspects; Craft Production, Anthropology of; Habitus: History of a Concept; Knowledge, Anthropology of; Performance: Anthropological Aspects; Personhood, Anthropology of; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development.

Bibliography

Bourdieu, Pierre, 1977. Outline of a Theory of Practice (R. Nice, Trans.). Cambridge University Press, Cambridge, MA.

Calhoun, Craig J., Sennett, Richard, 2007. Introduction. In: Practicing Culture. Routledge, London.

Fajans, Jane, 1997. They Make Themselves: Work and Play among the Baining of Papua New Guinea. University of Chicago Press, Chicago, IL.

Ho, Karen, 2009. Liquidated: An Ethnography of Wall Street. Duke University Press, Durham, NC.

Ingold, Tim, 1983. Being Alive: Essays on Movement, Knowledge, and Description. Routledge, London.

Jenkins, Richard, 1992. Pierre Bourdieu. Routledge, London.

Lave, Jean, Wenger, Etienne, 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, MA.

Levi-Strauss, Claude, 1966. The Savage Mind. University of Chicago Press, Chicago.

Mahar, Cheleen Ann-Catherine, 2010. Reinventing Practice in a Disenchanted World: Bourdieu and Urban Poverty in Oaxaca Mexico. University of Texas Press, Austin, TX.

Marx, Karl, 1947. Capital: A Critical Analysis of Capitalist Production. International Publishers, New York.

Ortner, Sherry B., 1989. High Religion: A Cultural and Political History of Sherpa Buddhism. Princeton University Press, Princeton, NJ.

Paxson, Heather, 2012. The Life of Cheese: Crafting Food and Value in America. University of California Press, Berkeley, CA.

Piaget, Jean, 1962. Play, Dreams, and Imitation in Childhood. Norton, New York.

Sahlins, Marshall, 1981. Historical Metaphors and Mythical Realities: The Early History of the Sandwich Islands. University of Michigan Press, Ann Arbor.

Sutton, David, 2012. Cooking Skills, the Senses, and Memory: The Fate of Practical Knowledge (Original work published 2006). In: Counihan, Carole, Van Esterik, Penny (Eds.), Food and Culture: A Reader, third ed. Routledge, London.

Terrio, Susan J., 2000. Crafting the Culture and History of French Chocolate. University of California Press, Berkeley, CA.

Vygotsky, L.S., 1962. Thought and Language. MIT Press, Cambridge, MA.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Pragmatic Disorders

Brigitte Stemmer, Université de Montréal, Montreal, QC, Canada; Brock University, St. Catherines, ON, Canada; and CRIUGM, Chemin Queen Mary, Montreal, QC, Canada

Published by Elsevier Ltd.

Abstract

Not being able to appropriately tell a story, have a conversation with someone, make a request, and understand metaphors, ironic statements or jokes despite an unimpaired language system are examples of pragmatic disorders. They can occur in childhood as well as in adulthood when the brain is afflicted by a stroke or some other injury or disease process, and can be devastating and affect our psychosocial well-being, mental health, and occupational success. This contribution addresses questions such as what is linguistic pragmatics, when do pragmatics disorders occur, and how are they explained.

Introduction

The impact of disorders of pragmatics may be at least as devastating as disorders with the structural use of language. Both disorders can influence a person's mental health status, psychosocial well-being, and occupational status and success. Unlike structural and semantic disorders of language, however, disorders of pragmatics are usually less obvious and more difficult to recognize as a handicap (for a discussion see Cummings, 2011). For a long time, the description and evaluation of language disorders focused on phonetic, morphological, syntactic, and semantic aspects of language. This, however, does not adequately explain the richness of meanings that arise in the complex interaction between people. Consider the following question, "What do Barack Obama, Angela Merkel, and Marilyn Monroe have in common?" You may not know the answer, neither do I, but by now you probably guessed that the utterance is the beginning of a joke and whatever the answer, it will play on some characteristics of these people, what you know about them or situations surrounding them. And you did not make this inference solely based on the words strung together by some grammatical rules but additionally on what was not said, or your pragmatic knowledge. The way language is used and interpreted while taking into consideration the characteristics of the interlocutors and the effects of contextual, situational, and social variables is subsumed under the notion of linguistic pragmatics. Linguistic pragmatics refers to the way language is used in interaction for the purpose of communication. It is the impairment of this aspect of language that is the focus of this article. To understand what these impairments entail, it is useful to discuss in more detail linguistic pragmatics first.

Language as Social Action

Looking at the way language is used rather than simply at the words and grammar is credited to philosophers such as Ludwig Wittgenstein, John Austin, John Searle, and H. Paul Grice. Work in this tradition summarized key features of language as social action and analytic frameworks were developed for understanding how language works. Discourse scholars later applied these analytic frameworks to sequences of text and talk in context. The classic pragmatic models – usually focusing on isolated utterances – were thus expanded and modified to incorporate influences from other approaches such as discourse pragmatics (for a summary see Blum-Kulka and Hamo, 2011).

An analytic framework that has been used frequently to detect pragmatic impairments is the one developed by Austin and Searle for speech acts (such as convincing or inviting someone, making a promise, complaining to someone, and so on) (Austin, 1962; Searle, 1975). What we actually 'say', that is the linguistic means we use to utter speech acts, may be different from the content we intend to communicate. For example, an utterance such as "I'll go ice climbing next weekend" may be an invitation to someone to join in or information that I won't be home next weekend. "It's cold in here" may be information about the temperature in the room, or it may be an indirect request for closing the window. Under normal circumstances, we are pretty good at understanding what the utterance is supposed to mean – although misunderstandings may occur. How is this possible? One influential theory suggests that we perform reasoning steps to recognize what a speaker wants to imply or communicate (so-called implicatures); (Grice, 1975, 1978). We do so by assuming cooperation between the interlocutors, which is guided by following a set of maxims such as the maxim of quality (truthfulness: do not say that which you believe to be false or that for which you lack adequate evidence), the maxim of quantity (informativeness: make your contribution as informative as is required, but do not contribute more information than is required), the maxim of relation (relevance: make your contribution relevant to the exchange), and the maxim of manner (clarity: be brief and orderly, avoid ambiguity and obscurity of expression).

Although Grice's theory had and still has an immense influence in pragmatic studies, it has been challenged on theoretical and empirical grounds. One criticism concerns the assumption that the comprehension process is guided by predictable expectations that are derived from principles and maxims. Some researchers have argued that when people communicate they always have something in mind that is relevant to them and thus the maxim of relevance always applies (and similar arguments have been advanced for the other maxims). What is important is that the message is 'relevant', that it is meaningful and makes sense to the speaker and hearer in a 'mutual cognitive environment' (Sperber and Wilson, 1988). In their relevance theory Sperber and Wilson

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54027-6

(1988, 2007) thus subsumed the Gricean maxims as one supermaxim called 'Relevance.' At the same time, they introduced a psychological component to the verbal comprehension process.

Another criticism is based on the argument that pragmatics is not only implicated in what is communicated but also in what is said. During normal language understanding different aspects of pragmatic knowledge are accessed such as deep, default, or stereotypical background knowledge that provides an interpretation of what speakers say (primary pragmatic knowledge) (Gibbs, 1999; Recanati, 1993). Another aspect involves specific, local contextual factors that guide the interpretation of what speakers implicate or communicate (secondary pragmatic knowledge). It is, however, not clear what criteria to use to determine what counts as primary and what as secondary pragmatic knowledge and the usefulness of distinguishing primary and secondary pragmatic knowledge has been questioned. Some pragmatic knowledge might be more salient than other depending on the knowledge and experience we have accumulated during our life time, the knowledge and experience that is activated, and the affective states we are in. In fact, it has been suggested that it is the degree of salience rather than inferences made on the basis of the linguistic or nonlinguistic context that determines the meaning of an expression (for a summary of the graded salience hypothesis see Giora, 2003).

The Notion of Pragmatic Disorders

Although there is currently no agreement on how we arrive at the proper interpretation of an utterance or a discourse, there is agreement that this ability can be impaired in children as well as in adults. Already in 1983, in their attempt to classify developmental language disorders, Rapin and Allen (1983) described a syndrome of language impairment, which they termed the 'semantic-pragmatic deficit syndrome.' Children suffering from this syndrome were described as being overly verbose, having poor turn-taking skills in conversations, poor discourse and narrative skills, and having problems with initiating, maintaining, and terminating discourse topics. The term was later modified to 'semantic-pragmatic disorder' by Bishop and Rosenbloom (1987). A debate ensued around the validity of using the term as a diagnostic label for developmental language disorders. Some researchers (e.g., Aarons and Gittens, 1993) argued that the children described with these disorders also had cognitive deficits typical for autism and that 'semantic-pragmatic disorders' was a disorder belonging to the autistic spectrum. This was supported by reports that did not find differences between semantic-pragmatic syndromes or disorders and autism (Bishop, 1989; Boucher, 1998; Gagnon et al., 1997). Subsequently, semantic-pragmatic disorders were mostly viewed as part of the autistic spectrum rather than as a developmental language disorder. The modern day successor of the term semantic-pragmatic disorders is pragmatic language impairment (PLI), pragmatic disorders, or clinical pragmatics with the latter term sometimes viewed as a subcategory of pragmatic disorders.

Theorists and clinicians frequently distinguish between primary and secondary pragmatic disorders. Whereas primary pragmatic disorders describe deficiencies that are not a consequence of linguistic problems, secondary pragmatic disorders describe an individual who suffers from deficiencies of structural language, which affects his or her pragmatic competence. For example a person who is not able to produce a grammatically correct sentence (**Drive home*) may be unable to appropriately produce a request (*Could you give me a lift home?*), open or take turns in a conversation (*I would like to pick-up your previous argument; May I interrupt you here?*), produce narratives, and so on. In such cases, we say that the pragmatic impairment is secondary to a deficient linguistic structure. This is different from someone whose language is intact but who does not know or understand how to appropriately open a conversation or to produce speech acts according to the context and situation. For example, saying to a friend "My car broke down this morning. Give me a lift home tonight, ok?" would be appropriate if addressed to a friend but not to a colleague at work where it might be more appropriate to say something like "My car broke down this morning. It would be really nice if you could give me a lift home. It's not much out of your way." The distinction between primary and secondary pragmatic disorders is, however, problematic. Many individuals that show pragmatic inappropriateness also exhibit cognitive problems (such as attention, memory, or executive functions) and it is not always clear what was first, the chicken (pragmatic disorder) or the egg (cognitive problems).

Pragmatic Disorders in Children and Adults

Pragmatic disorders have been described in children during development and in adults subsequent to some type of brain injury or brain abnormality. Although studies of pragmatic impairments emerged more than 30 years ago, many issues are still controversial or require clarification. An important and still controversial question is whether pragmatic disorders exist as an independent entity or whether they are associated with some specific disease or subsequent to mental or cognitive impairments. In addition, if structural language problems exist, it is not always easy to determine to what degree these problems influence or lead to pragmatic deficiencies. Another issue concerns the assessment of pragmatic impairments. On the positive side, there is agreement that not only structural aspects of language but also pragmatic skills need to be evaluated and many assessment instruments that target classic language skills nowadays also include pragmatic components. On the negative side, assessment instruments targeting pragmatic disorders are frequently not standardized, lack real-life application (ecological validity), rely primarily on reports by friends or family members, lack cultural adaptation, or require too much time (or expertise) to make them clinically useful. Scientific studies of pragmatic disorders often focus on specific aspects of pragmatic disorders, on specific patient groups and use different analytical frameworks. All this makes it difficult to establish whether pragmatic profiles exist and if so whether they are common or distinct between patient populations and within a patient group.

Developmental Pragmatic Disorders

When evaluating pragmatic impairments in children, the developmental process also needs to be considered. It may not always be possible to determine whether pragmatic deficiencies are due to a language system that is not yet fully developed or due to immature pragmatic skills, or whether we are dealing with 'pure' pragmatic disorders (assuming they exist). Another question that arises is whether a specific disease (such as autistic spectrum disorders, ASDs) is characterized by features that overlap with pragmatic skills such as social aspects of communication.

Autism Spectrum Disorders (ASD)

Autism spectrum disorders (ASDs) are characterized by impairments in the domains of communication and social skills, and by restricted, repetitive patterns of behavior, interests, or activities. Pragmatic impairments are considered a salient feature of these disorders. Studies have demonstrated deficiencies with making inferences and comprehending speech acts and figurative language such as metaphor and irony, evaluating humor, or detecting violations of or adhering to Grice's maxims. Individuals with ASD do not make use of context to appropriately explain nonliteral utterances such as sarcasm and jokes. They also show abnormalities in conversational behavior such as talking either too much or too little and having problems with taking turns. They also fail to follow politeness rules and make irrelevant remarks. The most commonly used explanations for these deficiencies are the theory of mind (ToM) and the central coherence weaknesses hypothesis, or both. Weak central coherence refers to a tendency to interpret utterances in isolation without integrating information from many sources. ToM weaknesses refer to difficulties with inferring the goals, emotions, and beliefs of other people on the basis of their mental states (for summaries see Cummings, 2009; Loukusa and Moilanen, 2009; Tager-Flusberg and Caronna, 2007). However, not all researchers adhere to these views. It has been argued that the central coherence hypothesis does not distinguish between contextual processing and processing at a local level (such as within a text). These are considered independent abilities and individuals with ASD do not have problems with processing at a local level (Martin and McDonald, 2003). Others have questioned the validity of ToM as an explanation for pragmatic difficulties of individuals with ASD. The argument is that most of these individuals also have executive-function problems that require planning, flexibility, and working memory abilities, and these abilities are related to performance of ToM tasks. Further, not all individuals with ASD show problems with ToM tasks. In addition, different brain regions seem to be activated in individuals with ASD compared to healthy individuals (for a summary see Tager-Flusberg and Caronna, 2007).

Specific Language Impairment (SLI)

Specific language impairment (SLI) is a developmental language disorder of unknown etiology. These children exhibit 'specific' language problems, and it was traditionally assumed that motor, sensory, and nonverbal intellectual abilities are unimpaired. However, studies showing that SLI children frequently demonstrate reduced processing speed, working memory, and attention shifting impairments have questioned the traditional view. Frequently described language problems are morphosyntactic errors (e.g., past tense marking) and lexical semantic errors alongside a large range of other verbal difficulties. Traditionally, it was assumed that SLI children are unimpaired in pragmatic skills or, if they occurred, they were viewed as secondary to the structural language problems. Some authors have, however, argued that a subset of the SLI population show pragmatic impairments that cannot be explained by their structural language problems (for a summary see Cummings, 2009). Among the pragmatic deficiencies described are inappropriate responses in conversations, decreased ability to use context to comprehend implied meanings, and decreased performance on inferential comprehension tasks (for summaries see Bishop, 2000; Cummings, 2009, 2010).

Other Developmental Disorders

Whereas there is a large literature investigating pragmatic impairments in individuals with ASD and SLI, fewer studies have investigated pragmatic impairments in other developmental disorders. A diverse range of pragmatic impairments that are difficult to reconcile have been described in children with genetic disorders such Down's syndrome (47 chromosomes, XX), Klinefelter syndrome (47 chromosomes, XXY) Williams syndrome (spontaneous deletion of genetic material from chromosome 7), or fragile X syndrome. Pragmatic impairments have also been reported in about 70% of children with emotional and behavioral disorders (reviewed by Benner et al., 2002) and in children with ADHD (Cummings, 2013). Some studies have described pragmatic language impairment in children outside of any developmental disorder while others have raised doubts that pragmatic language impairment exists without autism (Boucher, 1998; Gagnon et al., 1997; Gibson et al., 2013).

DSM-5: Social (Pragmatic) Communication Disorder (SCD)

Before ending this section, a brief note on the new diagnostic category introduced in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) is warranted. The DSM-5 published in 2013 (American Psychiatric Association, 2013) has added 'social (pragmatic) communication disorder' as a new diagnostic category to "ensure that the unique needs of affected individuals are met." ASD must be ruled out for SCD to be diagnosed. DSM-5 characterizes the disorder as follows:

> SCD is characterized by a persistent difficulty with verbal and nonverbal communication that cannot be explained by low cognitive ability. Symptoms include difficulty in the acquisition and use of spoken and written language as well as problems with inappropriate responses in conversation. The disorder limits effective communication, social relationships, academic achievement, or occupational performance. Symptoms must be present in early childhood even if they are not recognized until later when speech, language, or communication demands exceed abilities.

The new diagnostic category has raised much controversy and criticism. Some argue that SCD does not have reliability

or validity as a distinct category of neurodevelopmental disorder. It is argued that SCD does not exist as an independent entity but seems to be associated with other disorders such as autism, SLI, or ADHD (see previous discussion; for a summary see Norbury, 2013; Helen Tager-Flusberg, 2013). It has further been criticized that the currently available assessment instruments either lack standardization, which are not ecologically valid, or rely too much on parent or teacher report measures (Helen Tager-Flusberg, 2013). SCD is also prone to cultural bias or interpretation. It thus remains to be seen whether SCD as a new diagnostic category will survive DSM-5 updates.

Acquired Pragmatic Disorders

Pragmatic impairments can also occur outside the developmental phase such as when injuries to or disease of the brain occurs. Such acquired pragmatic disorders have been reported in patients after a stroke to the left or right hemisphere, in patients with traumatic brain injury (TBI), in patients with a tumor, or in patients with neurodegenerative disease (such as dementia or Parkinson's disease). Some of these patients may not show any structural or semantic language impairment and yet present with severe pragmatic impairments (for reviews of pragmatic disorders in various patient populations see Cummings, 2009, 2013; Stemmer, 2008b).

Cerebrovascular Disease: Right Hemisphere Damage (RHD) and Left Hemisphere Damage (LHD)

The most common reason for right- or left hemisphere damage is an interruption of blood supply to the brain such as a blood clot in an artery (ischemic stroke) or a rupture of a blood vessel (hemorrhagic stroke). A stroke damaging neural circuits in regions of the left frontal or temporal lobes of the brain is frequently associated with language production or comprehension problems such as Broca's or Wernicke's aphasia. Damage to neural circuits in the right hemisphere of the brain have been associated with disorders of pragmatics including affective prosody, aspects of discourse (e.g., conversational and narrative style, coherence, speech acts) or figurative language (e.g., metaphor, irony, joke). Consider the transcript of a right hemisphere damaged (RHD) patient's attempt to summarize a story about a plane crash that had been read to him (see **example 1**). Although his linguistic skills are unimpaired and neuropsychological testing did not reveal attention or memory problems, it is obvious that he diverges from the main story and introduces new elements irrelevant to the summary and which were in fact not in the story. Some (but certainly not all) RHD patients display a somewhat odd narrative style, which has been described as embellishing, rambling, tangential, noninformative, irrelevant to the central topic and confabulatory. They show problems with discourse organization reflected in story structure, cohesion and logical coherence difficulties. They may jump from topic to topic leaving the listener with gaps of information, miss the overall point of a conversation or show difficulties extracting the relevant themes of a story. Besides these discourse problems, RHD patients have also demonstrated problems with using prosody to interpret the emotional content of discourse.

Example 1

(^ indicates a pause)

Patient: Well ^ yes ^ Laura was on her way to ^^^ hm ^^ I forgot ^^^ anyway she was on a plane and then there was a terrible noise and the plane kept falling and falling and it must have really been scary ^^^ yeah ^^^ I remember there was this plane crash last year in the US ^^^ I forgot exactly but many people died although some survived ^^ it's a real miracle ^^ you know ^^ can you imagine being in a plane that crashes ^ I mean ^ hm ^ yeah ^ actually I read somewhere compared to car accidents plane crashes are very very rare ^^^ yeah ^^^ ^^^

Therapist: ok ^^ you said that Laura was on a plane and the plane kept falling.

Patient: hmm ^^ oh yes ^^ right ^^ and there was this thunderstorm and the tent was on this cliff with people screaming ^^^ being in a tent when it rains is no fun …

Other pragmatic difficulties include problems at the level of integrating various pieces of information to build new situation or mental models. In Example 1, the patient was able to make global inferences. Although it was never explicitly mentioned in the text, the patient inferred that the plane had crashed, that there was a thunderstorm and that Laura was in a tent. The text explicitly mentioned that the person in the tent was having a dream. Nevertheless, the patient was not able to 'combine' the pieces of information to come up with the correct story in which Laura was in a tent during a thunderstorm and dreaming of a plane crash.

Similar to narratives, using and understanding figurative or nonliteral language (such as nonconventional indirect requests) requires the ability to draw inferences and integrate a wide range of information sources. A large literature has documented the difficulties of RHD patients with figurative and nonliteral language that requires such integration processes. For example, an RHD patient might associate the metaphor "He had a heavy heart" with a picture of a person stumbling under the burden of carrying a large heart. But he may not be able to draw on his experience and knowledge of the world and integrate these pieces of information to come up with the right interpretation.

(For summaries of pragmatic disorders in RHD patients see, for example, Joanette and Brownell, 1990; Champagne-Lavau et al., 2006; Cummings, 2009, 2013; Stemmer, 1999, 2008a,b; Winner et al., 1998).

Aphasia is a language disorder usually provoked by a left hemisphere lesion, most frequently (but not exclusively) caused by a stroke. Like patients with RHD, aphasic patients are a heterogeneous group and, based on their linguistic performance, several aphasic syndromes have been described. Depending on the aphasic syndrome displayed, these patients are differentially impaired in their ability to comprehend and/or produce language. Studying pragmatic abilities in these patients can thus be challenging and it is difficult to determine whether the pragmatic difficulties are due to the language impairment or if these are discourse problems per se. A lack of initiating discourse and topic changes, a reduced query production or a reduced variety in the use of types of speech acts, less informative discourse and less specificity, and

accuracy of the message have been described in aphasic patients. Most researchers attribute pragmatic difficulties in LHD patients to their linguistic impairment. There are, however, a few studies that seem to suggest that these difficulties may be independent of linguistic impairment. The issue is currently unresolved.

Traumatic Brain Injury (TBI)

Car and sport accidents, violent crimes, and so on can cause injury to the head and lead to what is called TBI. Patients suffering from TBI are a particularly heterogeneous group because a trauma can affect all parts of the brain and frequently more than one lesion site is involved (and thus various brain networks affected). This makes it very difficult to compare research findings and arrive at generalized conclusions. Many TBI patients show normal lexical and grammatical skills. Some TBI patients, especially those with damage to the frontal lobes of the brain, exhibit pragmatic impairments similar to patients with right hemisphere lesions such as in formulating nonconventional indirect requests despite preserved sensitivity to social factors. TBI patients with frontal lobe involvement frequently display impairments of executive functions, that is a constellation of disorders including difficulties with problem solving and mental flexibility, a failure to abstract, perseverative behavior, and a loss of control and monitoring functions. Executive impairments can severely interfere with narrative or conversational discourse production. TBI patients have been reported to exhibit poor topic management and to produce disorganized, repetitive, and over- or underdetailed discourse. Other pragmatic impairments documented in TBI patients are difficulties with using or detecting figurative or nonliteral language such as the use of nonconventional indirect requests or sarcasm and deficiencies in using politeness markers.

Neurodegenerative Disorders

Neurodegenerative disorders include various types of dementia but also patients with Parkinson's disease, Huntington's disease, multiple sclerosis and so on. Studies of patients with Alzheimer's disease have demonstrated difficulties with opening a conversation, comprehension of figurative language, and ToM tasks. Alzheimer patients have also demonstrated lexical and syntactic impairment, which makes it difficult to judge to what degree the linguistic impairment influenced the pragmatic deficiencies. Patients with frontotemporal dementia exhibited problems with organizing discourse and ToM tasks and showed poor performance in moral reasoning, emotion processing, and showing empathy. Patients with Parkinson's disease have been reported to show poor performance with the comprehension of figurative language (irony, metaphor) and speech acts as well as with turn taking in conversations.

Psychiatric Disease

Among patients with a psychiatric disease, schizophrenia is the one that has been most frequently investigated in terms of pragmatic disorders. Impairment in social functioning has been a hallmark of schizophrenia and a series of studies has shown that schizophrenia patients display problems with ToM tasks. At the discourse level, patients with schizophrenia demonstrate difficulties adhering to a topic, distinguishing relevant from nonrelevant content and have difficulties with taking turns in conversation (for a summary see Meilijson et al., 2004). Schizophrenia patients also exhibit difficulties with interpreting nonconventional indirect requests (such as "It is raining outside" intended to mean "Give me a lift home") and figurative language such as metaphors, irony, idioms, or jokes.

Explaining Pragmatic Disorders

The previous discussion has shown that pragmatic language disorders can be identified in patient populations that are rather heterogeneous. Many of the pragmatic 'symptoms' overlap and do not seem to be specific for any patient population. What has emerged from lesion studies is that pragmatic disorders can occur independently of language disorders such as in RHD patients and that language disorders may lead to pragmatic deficiencies. These observations are supported by neuroimaging studies indicating that communicative abilities (including pragmatic abilities) and language abilities rely on different cortical systems (for a review see Willems et al., 2010).

Several suggestions to explain or find the underlying cause for pragmatic deficiencies have been made. Prominent ones are the ToM hypothesis, the inferencing hypothesis and the mental model hypothesis (for a summary and discussion see Martin and McDonald, 2003; Stemmer and Joanette, 1998; Stemmer, 2008a). Consider the ToM hypothesis: If you are not able to infer and thus represent in your own mind another person's mental state (such as her hopes, beliefs, beliefs about beliefs, moods, desires, intentions), then you will not be able to employ such representations to interpret, predict, and judge the utterances and behavior of the other person. The ToM hypothesis is closely related to the inferencing hypothesis in that the ability to make inferences about other people's mental states is central. The inferencing hypothesis suggests that the pragmatic impairments observed are due to difficulties generating inferences at local and global levels such as making inferences from a text you read in the newspaper, from a statement made by your friend, or from the position of cars involved in an accident. Testing this hypothesis has, however, produced diverging results possibly due to a lack of properly defining what was meant by inferencing, a neglect in distinguishing between different types of inferencing and also a lack of a theoretical framework such as inference generation models. The mental model hypothesis has been viewed as an orverarching theory encompassing the ToM and inferencing hypothesis. In this view, we use the information available to us to construct mental models (or modify or integrate them), in order to understand, explain, and predict phenomena of the environment, and to act accordingly. It has, for example, been suggested that the problems observed in the RHD patient in example 1 and in other patients with pragmatic problems was due to an inefficiency in manipulating or updating mental models they had constructed. Although the mental model hypothesis seems to account best for many of the pragmatic impairments described, none of the hypotheses discussed here is able to fully explain the many different facets of pragmatic impairments. Reasons may be that pragmatic impairments are a conglomerate of very diverse 'symptoms' that may have

different and/or multiple causes that are difficult to capture by one theory.

There is some empirical support for the claim that pragmatic disorders may occur without other cognitive systems (such as executive functions including attention and working memory; monitoring and control systems, language) or emotional components being impaired. Some doubt has, however, been raised whether such cases indeed exist as the evaluations used in the studies to exclude cognitive impairments may not have been sufficiently exhaustive or the instruments used may not have been specific enough to capture the impairment. This issue is still controversial.

Pragmatic Disorders and the Brain

Numerous studies have associated the frontal lobes of the brain with pragmatic functions. The frontal lobes cover a large area of the brain, are subdivided into several anatomical subregions and have rich connections to other parts of the brain. It is important to note that the frontal lobes are not a unitary system as they are implicated in a variety of functions such as language production (Broca's region), executive, inhibiting, and control mechanisms (see Figure 1). The frontal lobes also play a role in self-awareness and consciousness, social and personality development and various memory functions. It is thus not surprising that damage to the frontal lobes and the prefrontal cortex (PFC) in particular can lead to pragmatic impairments as has been described in various patient populations. It is also not surprising that – due to the functional fractionation of the frontal system and its rich relationship to other neural networks in the brain – damage to different frontal lobe systems and their interconnections may affect pragmatic performance in different ways. Most clinical studies have not considered the functional fractionation of the frontal lobe system in a systematic way nor investigated its effect on different facets of pragmatic behavior. Another issue that needs to be solved is the observation that patients with lesions outside the frontal lobes have also shown pragmatic impairments.

(For summaries see Ferstl et al., 2008; Ferstl, 2010; Stemmer, 2008a, 2008b, 2013) (*see* Lateralization of Language as Demonstrated by Brain Imaging Procedures).

Considering the difficulties of investigating clinical populations, the hope has been that modern neuroimaging techniques (such as functional magnetic resonance imaging (fMRI), event-related potentials (ERPs) and so forth) would shed light on these issues as these techniques can also be used with healthy populations and thus eliminate the problems inherent in patient studies. Neuroimaging studies investigating discourse and figurative language in healthy populations only partially confirm the findings of behavioral studies with patient populations. Although early behavioral studies implicated the right hemisphere in high-level discourse and nonliteral/figurative language studies, later behavioral studies were more controversial on these issues. There is general agreement from neuroimaging studies that figurative and nonliteral language processing does not necessarily involve the right hemisphere but does so at times. The exact circumstances under which the right hemisphere is implicated still need further clarification. Structural discourse as well as figurative and nonliteral

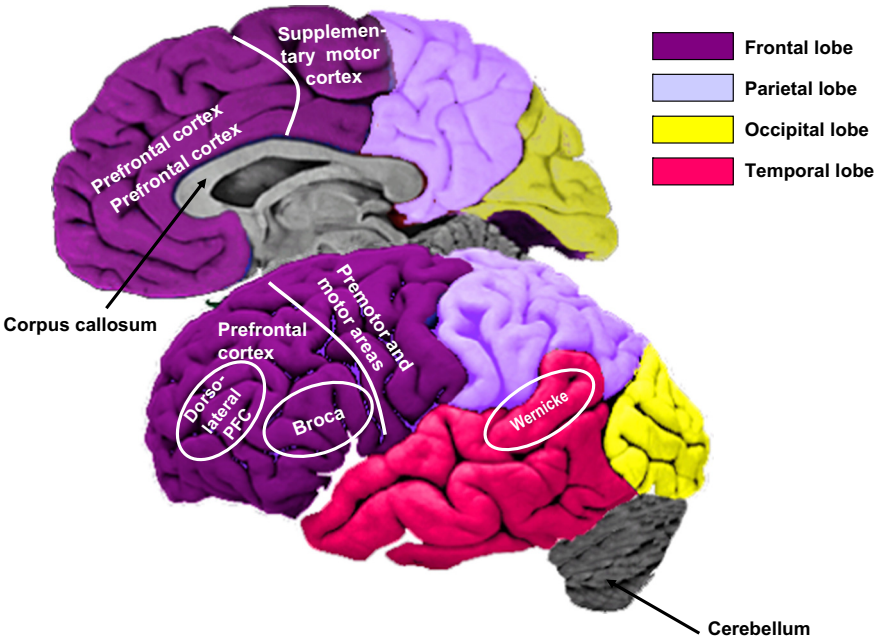

Figure 1 The bottom hemisphere shows the lateral view of the brain (i.e., looking at the brain from the side) and the top hemisphere the medial view (i.e., the surface that shows when cutting the brain vertically into two halves). The white line roughly separates the premotor and motor area from the PFC. The corpus callosum is a large fiber bundle connecting the two hemispheres. In the left hemisphere, Broca's region has classically been associated with language production and Wernicke's region with language comprehension (*see* Lateralization of Language as Demonstrated by Brain Imaging Procedures).

language implicate the same language regions that are active during word and sentence processing. Additional brain regions are recruited such as the anterior regions of the temporal lobe possibly reflecting semantic and episodic memory integration processes. The dorsomedial region of the PFC is also activated which has been interpreted as reflecting conscious strategic processes. Integrating text information seems to additionally involve regions of the anterior temporal and parietal cortex.

Although neuroimaging studies on discourse, nonliteral, and figurative language and aspects of social cognition have accumulated, reconciliation of the findings is still difficult as the studies use different stimuli, tasks, terminology, and imaging techniques. The picture that emerges so far is still rough, needs further work and confirmation and thus should be considered rather a basis for discussion than the established truth.

See also: Aphasia; Functional Brain Imaging of Language processes; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures.

Bibliography

Aarons, M., Gittens, T., June 1993. Semantic-pragmatic disorder (or a little bit autistic?). College of Speech and Language Therapists Bulletin 18.

Austin, J.L., 1962. How To Do Things with Words. Clarendon Press, Oxford.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Publishing, Arlington, VA.

Benner, G.J., Nelson, J.R., Epstein, M.H., 2002. Language skills of children with EBD: a literature review. Journal of Emotional and Behavioral Disorders 10, 43–56.

Bishop, D.V.M., 1989. Autism, Asperger's syndrome and semantic-pragmatic disorder: where are the boundaries? British Journal of Disoders of Communication 24 (2), 107–121.

Bishop, D.V.M., 2000. Pragmatic language impairment: a correlate of SLI, a distinct subgroup, or part of the autistic continuum? In: Bishop, D.V.M., Leonard, L.B. (Eds.), Speech and Language Impairments in Children: Causes, Characteristics, Intervention and Outcome. Psychology Press, Hove, East Sussex, pp. 99–113.

Bishop, D.V.M., Rosenbloom, L., 1987. Childhood language disorders: classification and overview. In: Yule, W., Rutter, M. (Eds.), Language Development and Disorders. MacKeith Press, London, pp. 16–41.

Blum-Kulka, S., Hamo, M., 2011. Discourse pragmatics. In: van Dijk, T.A. (Ed.), Discourse Studies: A Multidisciplinary Introduction. Sage, London, pp. 143–164.

Boucher, J., 1998. SPD as a distinct diagnostic entity: logical considerations and directions for future research. International Journal of Language and Communication Disorders 33 (1), 71–108.

Champagne-Lavau, M., Stip, E., Joanette, Y., 2006. Social cognition deficit in schizophrenia: accounting for pragmatic deficits in communication abilities? Current Psychiatry Reviews 2, 309–315.

Cummings, L., 2009. Clinical Pragmatics. Cambridge University Press, Cambridge, UK.

Cummings, L., 2010. Pragmatic disorders. In: Stone, J.H., Blouin, M. (Eds.), International Encyclopedia of Rehabilitation. Available online: http://cirrie.buffalo.edu/encyclopedia/en/article/363/.

Cummings, L., 2011. Pragmatic disorders and their social impact. Pragmatics and Society 2 (1), 17–36.

Cummings, L., 2013. Clinical Pragmatics. In Oxford Handbooks. Online. From: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199697960.001.0001/oxfordhb-9780199697960-e-001 (retrieved 30.12.13.).

Ferstl, E.C., 2010. Neuroimaging of text comprehension: where are we now? Italian Journal of Linguistics 22 (1), 61–88.

Ferstl, E.C., Neumann, J., Bogler, C., von Cramon, D., 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Human Brain Mapping 29 (5), 581–593.

Gagnon, L., Mottron, L., Joanette, Y., 1997. Questioning the validity of the semantic pragmatic syndrome diagnosis. Autism 1 (1), 37–55.

Gibbs, R.W., 1999. Interpreting what speakers say and implicate. Brain and Language 68, 466–485.

Gibson, J., Adams, C., Lockton, E., Green, J., 2013. Social communication disorder outside autism? A diagnostic classification approach to delineating pragmatic language impairment, high functioning autism and specific language impairment. Journal of Child Psychology and Psychiatry 54 (11), 1186–1197.

Giora, R., 2003. On Our Mind: Salience, Context, and Figurative Language. Oxford University Press, Oxford.

Grice, H.P., 1975. Logic and Conversation. Academic Press, New York.

Grice, H.P., 1978. Further Notes on Logic and Conversation. Academic Press, New York.

Joanette, Y., Brownell, H.H., 1990. Discourse Ability and Brain Damage: Theoretical and Empirical Perspectives. Springer, New York.

Loukusa, S., Moilanen, I., 2009. Pragmatic inference abilities in individuals with Asperger syndrome or high-functioning autism. A review. Research in Autism Spectrum Disorders 3 (4), 890–904.

Martin, I., McDonald, S., 2003. Weak coherence, no theory of mind, or executive dysfunction? Solving the puzzle of pragmatic language disorders. Brain and Language 85 (3), 451–466.

Meilijson, S.R., Kasher, A., Elizur, A., 2004. Language performance in chronic schizophrenia: a pragmatic approach. Journal of Speech, Language and Hearing Research 47, 695–713.

Norbury, C.F., 2013. Practitioner review: social (pragmatic) communication disorder conceptualization, evidence and clinical implications. Journal of Child Psychology and Psychiatry. http://dx.doi.org/10.1111/jcpp.12154. Online publication.

Rapin, I., Allen, D., 1983. Developmental language disorders: nosologic considerations. In: Kirk, U. (Ed.), Neuropsychology of Language, Reading, and Spelling. Academic Press, New York, pp. 155–184.

Recanati, F., 1993. Direct Reference: From Language to Thought. Blackwell, Cambridge.

Searle, J., 1975. Indirect speech acts. In: Cole, P., Morgan, J.L. (Eds.), Syntax and Semantics, vol. 3: speech act. Academic Press, New York, pp. 59–82.

Sperber, D., Wilson, D., 2007. Pragmatics, the Oxford Handbook of Contemporary Philisophy. Online publication. 10.1093/oxfordhb/9780199234769.003.0018.

Sperber, D., Wilson, D. (Eds.), 1988. Relevance. Communication and Cognition. Harvard University Press, Cambridge, MA.

Stemmer, B., 1999. Discourse studies in neurologically impaired populations: a quest for action. Brain and Language 68, 402–418.

Stemmer, B., 2008a. Neuropragmatics. In: Ball, M.J., Perkins, M., Mueller, N., Howard, S. (Eds.), The Handbook of Clinical Linguistics. Blackwell Publishing, Malden, pp. 61–78.

Stemmer, B., 2008b. Neuropragmatics: disorders and neural systems. In: Stemmer, B., Whitaker, H.W. (Eds.), Handbook of the Neuroscience of Language. Academic Press, London, pp. 175–188.

Stemmer, B., 2013. Neuropragmatics, Oxford Handbooks. Online. From: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199697960.001.001/oxfordhb-9780199697960-e-003 (retrieved 31.12.13.).

Stemmer, B., Joanette, Y., 1998. The interpretation of narrative discourse of brain-damaged individuals within the framework of a multilevel discourse model. In: Beeman, M., Chiarello, C. (Eds.), Right Hemisphere Language Comprehension. Perspectives from Cognitive Neuroscience. Lawrence Erlbaum, Mahwah, NJ, pp. 329–348.

Tager-Flusberg, H., 2013. Evidence Weak for Social Communication Disorder. Online publication. From: http://sfari.org/news-and-opinion/specials/2013/dsm-5-special-report/evidence-weak-for-social-communication-disorder (retrieved 30.05.13.).

Tager-Flusberg, H., Caronna, E., 2007. Language disorders: autism and other pervasive developmental disorders. Pediatric Clinics of North America 54 (3), 469–481.

Willems, R.M., Hagoort, P., Casasanto, D., 2010. Body-specific representations of action verbs: neural evidence from right- and left-handers. Psychological Science 21 (1), 67–74.

Winner, E., Brownell, H., Happé, F., Blum, A., Pincus, D., 1998. Distinguishing lies from jokes: theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. Brain and Language 62, 89–106.

Pragmatics, Linguistic

Jef Verschueren, University of Antwerp, Antwerp, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.C. Levinson, volume 17, pp. 11948–11954, © 2001, Elsevier Ltd.

Abstract

Linguistic pragmatics is defined, for the purposes of this article, as the science of language use. Its roots in linguistics, psychology, anthropology, and especially semiotics and the philosophy of language are briefly described. Then some basic notions of pragmatics are reviewed: deixis and indexicality, forms of implicit meaning such as presupposition and conversational implicature, speech acts, and conversation. Major considerations for a definition of the field are introduced: the view of language as action, the role of cooperation and coordination, the importance of common ground and context, the distinction between coded and inferred meaning. The major approaches and traditions are briefly described, as well as interdisciplinary interfaces linking pragmatics to other fields in the humanities and social sciences. The article concludes with a cursory look at current trends and perspectives.

Within linguistics, pragmatics is most easily defined as the science of language use. Its deeper roots go back as far as Graeco–Roman antiquity, with its theory of moods (relating grammatical categories such as indicative, interrogative, and imperative with types of sentence usage in the form of statements, questions, and commands) and with Aristotelian principles of rhetoric. Its more recent history can be traced to a multiplicity of nineteenth and early twentieth century theories of language in various European countries and in the United States (cf Nerlich and Clarke, 1996). I restrict myself to mentioning just a few – largely early- to mid-twentieth century – precursors. The German psychologist and linguist Karl Bühler (1879–1963), for instance, developed an integrated theory of language as system and language in use. He is best known for his 'organon model' of the linguistic sign: in relation to the world, it is a symbol performing a representational function; in relation to the speaker's consciousness, the sign is a symptom or index with an expressive function; in relation to the addressee it is signal with an appellative function. In France, Émile Benveniste (1902–76) developed his theory of enunciation (*énonciation*). He attributed to language, the property of being foundational for human thought, society, and subjectivity. In his view, therefore, meaning is generated in the very act of enunciating or using language. In England, the Polish anthropologist Bronisław Malinowski (1884–1942) emphasized the socio-cultural functioning of language, introducing notions such as 'phatic communion' and 'context of situation,' the latter playing an important role in the development of John Rupert Firth's (1890–1960) contextualist and functionalist linguistics. For Firth, meaning, context, and human behavior were inseparably related, a fact to be taken into account for the analysis of language at all levels of structure, from the phonetic to the sociolinguistic. Meanwhile, Ludwig Wittgenstein (1989–51) developed his later philosophical ideas, focusing on the constitutive role of 'language games' and the importance of looking at the meaning of utterances in relation to their use. In the United States, pragmatism emerged as a philosophical method aimed at the clarification of the meaning of signs and the facilitation of communication. The earlier pragmatist work by, i.a., Charles Sanders Peirce (1939–14) and William James (1842–1910) eventually led to Charles Morris' (1901–79) classical definition, in the context of a general theory of signs or semiotics, of pragmatics as the study of the relation between signs and their users or interpreters (as distinct from syntax, studying the relationships of signs with other signs, and semantics, investigating the relation between signs and the objects to which they are applicable).

If the deeper roots of pragmatics are multidisciplinary, so is its development within linguistics from the early stages in the 1960s and 1970s. Two major sources of inspiration were philosophical: John Austin's (1962) theory of speech acts, further elaborated by John Searle (1969), and Paul Grice's (1989) theory of meaning, and in particular of conversational implicature – a concept to be explained below. Emphasizing historical continuity, it is interesting to know that the seminal contributions by Austin and Grice were first presented in 1955 and 1967, respectively, as William James lectures at Harvard. Further links with other disciplines will be pointed out along the way in the following pages.

After introducing some of the basic notions that consistently appear throughout the pragmatics literature, a few paragraphs will be devoted to the question of how to define the field. Different approaches or traditions in linguistic pragmatics will then be presented, as well as some of the main interdisciplinary interfaces. This article ends with a cursory look at current trends and future perspectives.

Basic Notions in Linguistic Pragmatics

Though principles relating language structure to language use, i.e., pragmatic principles, entered various traditions of linguistic scholarship much earlier, it was not until the 1960s and 1970s that systematic attention to them gathered strength in linguistics and started to adopt the 'pragmatics' label proposed in semiotics by Morris (1938). A strong impulse came from generative semanticists (George Lakoff, Robin Tolmach Lakoff, James D. McCawley, John R. Ross) who countered the Chomskyan generative grammar view of a purely

syntactic deep structure. They argued that contextual information had to be brought in to explain the functioning of linguistic structures. A typical example of how this could be done was provided by Gordon and Lakoff (1971) in their formulation of 'conversational postulates' for deriving the indirect meaning of an utterance from its literal meaning: if someone says *Can you help me?* in a context in which this cannot be simply a question for information (because, e.g., the hearer's ability to help is obvious), then the interpretation *I request you to help me* is generated by an inference rule that says that stating or questioning a felicity condition on a speech act (the assumed ability to comply with a request being a felicity condition on requests) in a context that blocks the literal interpretation of that statement or question (e.g., because of its being obvious), counts as a performance of that act. These kinds of considerations rapidly gained ground in a significant segment of the professional world of linguists, resulting in the straightforward pursuit of typically pragmatic topics. Some of the basic notions that structured the development of the field include deixis, conversational implicature, presupposition, speech acts, and conversational structure (each one, not incidentally, forming the basis for a chapter in Levinson, 1983).

Deixis

A first notion connecting linguistic form to contexts of use is 'deixis,' the label for linguistic devices (called 'indexicals') pointing at the way in which a given utterance relates to dimensions of time, space, society, and discourse (i.e., 'indexicality'). Temporal deixis, e.g., marked with verbal tenses and temporal adverbs (*now*, *yesterday*), locates an utterance in time, taking the moment of speaking or writing, the time of utterance, as the deictic center. Similarly, the contextual dependency of utterances on location in space, or spatial deixis, is carried by adverbs (*here*, *there*), demonstratives (*this*, *that*), directional verbs (*come*, *go*), and the like. Social deixis comprises first of all person deixis, indicated (in English) with personal pronouns, but also attitudinal deixis, as when indexical expressions (including elaborate systems of honorifics in certain languages) signal aspects of social status or forms of respect. In Amharic (an Ethiopian language), e.g., second-person singular male pronouns show four degrees of respect, from insulting, over familiar and intermediate, to highly respectful. Sometimes a distinction is made between relational social deixis (bearing on interactional positioning: speaker–addressee *I–you*, speaker–referent *I–he/she/they*, speaker–bystander *I–we*, where *we* refers to the speaker plus at least one other person present who is included as the source of what is said) and absolute social deixis (referring to relatively stable properties of personhood such as gender). We speak of discourse deixis, finally, when a form of expression (such as *This article will explain*…) refers to earlier, simultaneous, or later discourse, which is then often metapragmatically characterized as a specific type of discourse activity (*article*, *explain*).

Though the phenomenon was studied much earlier by Charles Sanders Peirce, Otto Jespersen, Karl Bühler, Émile Benveniste, and Roman Jakobson, to name just a few, a seminal contribution that placed it at the core of much pragmatic research, was Fillmore (1975).

Implicit Meaning

The reinterpretation of *Can you help me?* as *I request you to help me*, arrived at by means of an inferential process described as a conversational postulate by Gordon and Lakoff, exemplifies a large class of phenomena consisting of aspects of meaning that are not directly or not completely encoded in the form of linguistic expressions, but that are nevertheless communicated thanks to reliance on prior or accessible knowledge on the part of the interlocutors. Utterances do not only anchor language use in context (the case of deixis), but they also make assumptions about it.

A clear case of implicit meaning is that of 'entailment' or 'logical implication.' The definite article *the* in *John is the vice-chancellor* logically implies or entails that John is *a vice-chancellor*. This is a truth-conditional phenomenon: A entails or logically implies B if and only if every situation that makes A true also makes B true. This is different from what Grice called 'conventional implicature,' which does not involve truth-conditionality: the two utterances *John is smart and reliable* and *John is smart but reliable* have the same truth conditions, but the *but* in the second utterance conventionally implicates that there is a contrast and possible incompatibility between being smart and being reliable.

Another type of implicit meaning is 'presupposition.' Presuppositions, or meanings that must be assumed, understood, taken for granted in order for an utterance to make sense, are triggered by a vast range of linguistic forms and structures, from definite descriptions (e.g., *the king of France* presupposing that the king of France exists) to factive verbs (e.g., *to regret*, presupposing that what is regretted is a fact, or actually happened), cleft constructions (e.g., *It was John who left early* presupposing that someone left early), and counterfactuals (e.g., *If only the weather were better* presupposing that the weather is bad). A basic property of presupposition is that it remains constant under negation: the existential presupposition of *The king of France is bald* remains unchanged in *The king of France is not bald*.

A type of implicit meaning which much attention has been devoted to in the linguistic pragmatic literature, though – like implications and presuppositions – originally discussed in philosophy, bears the label of 'conversational implicature.' Paul Grice (see Grice, 1989), whose term this is, suggested that conversations are regulated by a principle of cooperation (according to which conversational contributions should be made in such a way as to satisfy the interactional purposes and needs of the moment) and a number of maxims of conversation (such as, do not say less or more than is required, say only what you believe to be true, be relevant). If a maxim appears to be violated, the assumption that the speaker is still adhering to the principle of cooperation will generate extra meaning beyond what is literally said. Thus *There is a garage around the corner* in response to *I am out of petrol* conversationally implicates that the speaker assumes the garage is open and has petrol for sale; otherwise the answer would cease to be relevant. A conversational implicature is called 'particularized' if and only if an utterance triggers it "only by virtue of specific contextual assumptions that would not invariably or even normally obtain," while it is considered 'generalized' if and only if it is normally inferred from an utterance "*unless* there are unusual specific contextual assumptions that defeat

it" (Levinson, 2000, p. 16). Thus *Some of the guests are already leaving*, uttered in response to *What time is it?*, gives rise to the generalized conversational implicature that not all of the guests are already leaving, and to the particularized conversational implicature that it must be relatively late (otherwise it would not be sufficiently relevant in the given context of the question that is asked). The same utterance would, in response to *Where is John?*, trigger the same generalized conversational implicature (since *some of* generally implicates *not all*, irrespective of the context, unless very specific circumstances arise), while a particularized conversational implicature (on the assumption of the relevance of the utterance in the context) would be the suggestion that perhaps John has already left. Generalized conversational implicatures consist in implicit meaning that can be conventionally inferred from forms of expression in combination with assumed standard adherence to Grice's conversational maxims, while particularized (occasion-specific) conversational implicatures constitute a type of implicit meaning inferred from the apparent flouting of a conversational maxim in combination with assumed adherence to Grice's cooperative principle.

Speech Acts

The idea that speaking is a form of social action has a long and complex history. But the notion of speech as action, as picked up in linguistic pragmatics, found its crystallization in Austin's (1962) and especially Searle's (1969) formulation of speech act theory. Very briefly summarized, Austin's attention was first attracted by a class of utterances with which speakers do not simply say something (e.g., *John will come to your party*, a constative in his terminology) but with which they clearly perform an action (such as the act of promising in *I promise to come to your party*, a so-called performative, in this case more specifically an explicit performative – in contrast to the primary performative *I will come to your party* – because the act performed is also named). This type of utterance had already been the focus of attention for a number of linguists. But for Austin further reflection on this phenomenon led to a rejection of the simple dichotomy between constatives (in which something is said that can be evaluated along the dimension of truth or falsity) and performatives (in which something is done that can only be evaluated in terms of so-called felicity, i.e., appropriateness in relation to aspects of context). Constatives, after all, can also be infelicitous (e.g., when the speaker does not have adequate evidence for what is said), and performatives usually have a propositional content. Therefore, he concluded that all utterances are sayings and doings at the same time, containing 'locutions' (acts *of* saying what is said, the proposition), 'illocutions' (what is done *in* saying something, such as making a statement, making a promise, giving an order), and 'perlocutions' (what is done *by* saying something, i.e., the effect that is scored, such as imparting a belief).

Ignoring the perlocutionary aspect, Searle transformed this approach into his speech act formula F(p), where F stands for illocutionary force, the action side of every speech act, and p for the content or proposition (consisting of a reference and a predication). Searle also developed a systematic way of describing speech acts in terms of various types of felicity conditions, or, as conceived in his theory, necessary and sufficient conditions. For the speech act of promising, for instance, one can formulate a propositional content condition (p specifies a future action by the speaker), preparatory conditions (the speaker is capable of doing what is promised, and under normal circumstances it is not obvious that the action specified would be undertaken), a sincerity condition (the speaker intends to do what is promised), and an essential conditions (the speaker's utterance counts as a commitment to a future course of action).

On the basis of this descriptive apparatus, from which three focal points are distilled (psychological or mental state, the 'direction of fit' between words and world, and the 'illocutionary point') five categories of speech acts are distinguished: assertives (e.g., statements; expressing a belief, making words fit the world, and committing the speaker to the truth of what is asserted), directives (e.g., requests; expressing a wish, making the world fit the words, and counting as an attempt to get the hearer to do something), commissives (e.g., promises; expressing an intention, making the world fit the words, and counting as a commitment for the speaker to do something in the future), expressives (e.g., apologies; expressing a variety of psychological states, having no direction of fit between worlds and words, and simply counting as the expression of a mental state), and declarations (e.g., baptizing; not expressing any psychological state, making both words fit the world and the world fit the words, and bringing about a change in – usually institutional – reality).

Special attention was also paid to a category of indirect speech acts: acts with a communicative force different from the one typically associated with the form of expression. An example is the question *Can you help me?* with which we opened this section on basic notions, and the subsection on implicit meaning in particular.

Conversation

Actual language use is not as neat as speech act theory makes it look. For one thing, most language use may be organized around units and structures that cannot so readily be described in terms of individual, mostly sentence-based, speech acts. In the development of pragmatics after Levinson (1983), a vast proportion of the research that does not deal explicitly with written language, has concentrated on forms or aspects of conversational interaction, whether or not in the framework of 'conversation analysis' (or CA) in the strict sense.

In contrast with the other basic notions, with a firm basis in linguistics, semiotics, and philosophy, the interest in conversation entered linguistics from sociology. A strong impulse came from the work of Harvey Sacks in the 1960s (as collected in Sacks, 1992), whose project was to investigate levels of social order as revealed in the structural organization of everyday conversational language use. Sacks was himself inspired by Garfinkel's (1967) ethnomethodological interest in practical reasoning and Goffman's (1963) focus on the structural significance of aspects of face-to-face interaction. Garfinkel's program was directly influenced by Husserlian phenomenology the implications of which were applied to the social sciences by Alfred Schuetz, in whose view social life is mainly structured around shared meanings. It is those shared

meanings that ethnomethodology extracts from primary phenomena such as commonplace activities and everyday social interaction.

Parallel to, and in interaction with, the development of CA, anthropologists such as Dell Hymes and John Gumperz developed an ethnography of speaking, a theory of communicative competence, and eventually a form of interactional sociolinguistics using conversations as primary data (Gumperz and Hymes (Eds.), 1972; Gumperz, 1982).

Defining the Field

The above notions have more in common with each other than the somewhat artificially separated manner of presentation would suggest. Spelling out all the conceptual connections is not possible in the context of this article. But we cannot avoid the question as to what exactly binds all these topics in such a way as to warrant their treatment under the unifying flag of linguistic pragmatics. There are several definitional threads that intertwine in a variety of ways, giving rise to different approaches and traditions. I just list them here.

- The primordial idea underlying pragmatics is the view of 'language as action.' There is no language without language use. Language does not simply consist of structures that can be systematically related to abstract meanings. It is fundamentally an intricately structured instrument to communicate or elicit information, to persuade, to provoke actions, to make commitments, to express one's feelings, to bring about institutional changes. Such functionalities, to which the structures of the instrument are well-adapted, underlie an endless range of activities such as buying bread, answering the phone, exchanging gossip, searching the internet, interviewing a politician, banning arms sales, delivering a campaign speech.
- The actions or activities presented as instances of language use are all forms of 'social action.' In other words, they involve some kind of 'interaction' between different people, even if this is not immediately transparent. It is this aspect of language use that gives rise to the centrality of the notion of 'cooperation' – the basis of a second definitional thread. Avoiding an exclusively harmony-oriented connotation of the term, it is possible to emphasize an aspect of 'coordination.' Even the most noncooperative communicative exchange, or the most unharmonious form of impoliteness, requires some sort of coordination or, in Clark's (1996) terms, 'joint action.' A more recent terminological alternative is 'co-participation' (Hanks, 2012), which allows for mutuality and reciprocity as well as antagonism and challenge. The exact nature of the interactional dynamics that is involved, and the implications of agency or the role of individual intentionality, remain some of the most crucial points to be elucidated in a theory of pragmatics.
- Coordination, cooperation, or joint actions all require interlocutors who share some kind of 'common ground.' Meaningful interaction between people takes as its starting point implicit assumptions about the extent to which knowledge, goals, and attitudes are shared or fail to be shared, and such assumptions can be adjusted in the course of a verbal exchange. Assumed common ground goes further than mutually presupoposed knowledge facilitating the recognition of intentions. What is involved, ultimately, is everything that has come to be discussed under the labels of 'context' and 'contextualization' – the core of a third definitional thread we must mention. In fact, one of the most common ways of distinguishing pragmatics from semantics has been to define semantics as the study of meaning without context and pragmatics as the study of meaning in context. Though this is a complicated and controversial issue, there is no doubt about the centrality of contextual considerations for pragmatics. A useful starting point for a theorization of context is to be found in the phenomenon of indexicality.

While these three appear, in one form or another, in most pragmatic traditions, there are three more definitional threads, focusing on units of analysis, the role of speaker and hearer in the meaning generation process, and specific cognitive processes, that underlie specific approaches.

- As to units of analysis, it is common practice to distinguish between sentences and utterances. When this distinction is made, it is generally suggested that pragmatics must deal with utterances, i.e., actually produced sentences in their context of production. There is a French pragmatic tradition, which is entirely focused on the corresponding notion of *énonciation*. Extending the idea, discourse – less associated with smaller sentence-length structures – is often identified as the true unit of analysis for pragmatics. The sentence–utterance distinction triggered the corresponding notions of sentence meaning (the province of semantics) and utterance meaning (the domain of pragmatics), to which the more specific notion of speaker meaning can be added. The latter is essentially Gricean 'nonnatural meaning' focusing on the speaker's communicative intentions, which also play a crucial role in orthodox speech act theory.
- Moving away from an exclusive focus on the contribution a speaker's intentions make to the meaning of an utterance, a distinction has been introduced between meaning that is 'coded,' as opposed to meaning that must be 'inferred.' Coded meaning, meaning that is directly attached to the linguistic code, is then regarded as the domain of semantics, while pragmatics studies meaning that cannot be read from the linguistic form directly but that is arrived at through processes of interpretation or inferencing. This is the definitional criterion handled, for instance, by Ariel (2010). The coded versus inferred dichotomy is also reflected in theories of pragmatics that place implicitness at its definitional core.
- Principles of language use adduced to define the field of pragmatics have often been formulated in cognitive terms. While aspects of cognition are involved in everything we do with language, widely accepted attempts have been made to reduce all pragmatic explanation to relatively straightforward principles such as 'relevance' and 'economy.' Keeping things simple for the sake of this quick overview, Sperber and Wilson's (1986) relevance theory, for instance, restricts pragmatics to the investigation of inferential computations

of meaning that handle relevance as the primordial principle (thus streamlining the Gricean approach, while clearly building on it), further supported by a principle of cognitive economy that allows language users to invest precisely the amount of cognitive processing effort that is needed to achieve contextually desirable communicative effects.

Clearly, the enumerated properties of and perspectives on language use lead to a wide range of different approaches and traditions, some of which are reviewed below.

Major Approaches and Traditions

The main divide in the pragmatic landscape is between an orientation that narrows down the field to what is felt to be a linguistically manageable and relatively clearly demarcated area, and an orientation that broadens the field to everything of relevance that can be said about using language as a human activity. The distinction is often described, inaccurately, as one between an Anglo-American and a Continental European tradition. In Yan Huang's words, in the Anglo-American conception "pragmatics is defined as the systematic study of meaning by virtue of, or dependent on, language use" (Huang, 2007, p. 4), focusing on topics such as implicature, presupposition, speech acts, and deixis, and positioning pragmatics in contrast to phonetics, phonology, morphology, syntax, and semantics as an additional component of a theory of language, quite distinct from the interdisciplinary areas of anthropological linguistics, sociolinguistics, and the like. For the Continental European tradition, on the other hand, "pragmatics is defined in a far broader way, encompassing much that goes under the rubric of sociolinguistics, psycholinguistics, and discourse analysis" (Huang, 2007, p. 4), and representing a general functional perspective on every possible aspect of linguistic behavior, in line with Morris' (1938) all-inclusive definition of pragmatics as the study of whatever relations there are between signs and their users or interpreters.

Without going into questions concerning the accuracy of the labels, there is indeed a contrast between a narrower view of pragmatics, treating it as a component of linguistics directly related and opposed to semantics, and a broader view that handles it as a general usage-oriented perspective on language. The basic idea behind what came to be called the perspective view (cf Verschueren, 1999) was very simple. Pragmatics can briefly be characterized as the science of language use, as I did in the first line of this article. When using language, all levels of linguistic structure are involved. Therefore, there are pragmatic aspects, describable and explainable in terms of usage conditions and processes, to be situated in relation to every unit of linguistic analysis, no matter whether they have already been assigned to the realm of one of the traditional components of a linguistic theory, from phonetics to semantics. Similarly, using language involves cognition, social structures, and cultural context. Therefore, much of what finds its way into the hyphenated, interdisciplinary fields of linguistics, must also be relevant to pragmatics.

Within the narrower version of pragmatics, significant developments have taken place in the past few decades. The classical Gricean theory of conversational implicature, for instance, has given birth to a number of so-called neo-Gricean approaches (Horn, 1984; Levinson, 2000; Huang, 2007); it is in this context that Levinson's theory of generalized conversational implicature must be situated. But there is also a tradition that is sometimes called postGricean. Here we encounter relevance theory, which, in Carston's (2009) account, draws a different distinction between implicit and explicit meaning. If Max's question *How was the party*? is responded to by Amy with *There wasn't enough drink and everyone left early*, there is not only the fully explicit linguistically encoded meaning (there was not enough drink + everyone left early), but also a level of 'explicatures': there was not enough ALCOHOLIC drink to satisfy people + everyone WHO CAME TO THE PARTY left early. These meaning enrichments, according to this account, must still be regarded as part of the explicit meaning because the truth or falsity of the utterance will be judged in terms of them. This contrasts with meaning that is wholly pragmatically inferred: given what one knows about parties, Amy's utterance means that the party was not a success. (For a further discussion of such issues, see Recanati, 2010.)

As to the broader perspective view of pragmatics, we find less coherent lines of theory formation, simply because of the open-endedness of the endeavor. If using language consists in the continuous making of linguistic choices, they are never made in isolation: cognition, society, and culture are involved. Some choices are made automatically, some consciously, but all choices are at least potentially subject to reflexivity, i.e., metalinguistic or metapragmatic awareness. Some are made for language-internal structural reasons (the possibilities and limitations of the available resources), others are driven by real or perceived aspects of the circumstances of use, motivated by beliefs, desires, intentions, expectations, or by those attributed to others in the context of interaction. Choices have to be made at all levels of structure, from the choice of a language, code or genre to phonetic/phonological, morphological, syntactic, lexical, semantic details. They may range over variety-internal options, or they may involve regionally, socially, or functionally distributed types of variation. Speakers do not only choose forms; they also choose strategies (e.g., deference), which in turn have an impact on structural choices. Furthermore, choices are made both in producing and in interpreting an utterance.

In order to make sense of this complex process, Verschueren (1999) proposes three key notions for pragmatics: variability, negotiability, and adaptability. Variability is the property of language that defines the range of possibilities from which choices can be made. This range is itself dynamic, as it is not only subject to diachronic change or instant innovation, but also to variable degrees of accessibility for different speakers or for the same speaker at different moments. Moreover, at any given moment in the course of interaction, a choice may rule out alternatives or create new ones for the current purposes of the exchange – though these effects can always be renegotiated, which brings us to our second key notion. Negotiability is the property of language responsible for the fact that, in spite of conventions and default options (cf Levinson, 2000; Jaszczolt, 2005), choices are not made mechanically or according to strict rules or fixed form–function relationships, but rather on the basis of highly flexible principles and strategies. Negotiability, which implies

both rationality and reflexivity, is also necessary because of significant degrees of indeterminacy on the side of the language producer (who is under an obligation to make choices even when the ideal ones are not directly accessible at a specific moment in the exchange) and the interpreter (whose understanding is based on choices that do not necessarily exclude their alternatives from the world of interpretation). Language use as choice-making from such a variable range of options in such a negotiable manner would not be understandable without positing a high degree of adaptability as the property of language that makes all this possible in such a way as to meet – to a satisfactory degree – human communicative needs.

This notion of adaptability is clearly related to the biologically adaptive emergence of human language itself (cf Verschueren and Brisard, 2009). Without going into any of the details here, the origins of language may be related to the twin phenomena of reflexivity and theory of mind, which are human ways of handling social complexity. The concept of 'reflexiveness' or 'reflexivity' (playing a prominent part in social psychology ever since Mead, 1934), refers to the capacity of the mind to bend back upon itself, to be aware of its own experiences as residing in a self that is situated in a social context of interaction. That social context consists of other selves which, by a transformation of reflexiveness which is commonly called 'theory of mind,' are perceived as minds comparable to the self. It is this step that is identified by Tomasello (1999) as the distinguishing trait that allowed *Homo sapiens* to engage in a rapid process of cumulative cultural learning: the basic human biological adaptation compared with other primates is one for living socially. It is a moral step, because by recognizing the other as related to oneself, a subject is itself introduced into a community of 'persons,' conscious agents engaged in reading each other's minds, which enables them to set up common goals. It is this adaptation for living socially that has produced the 'sociality' recognized in Enfield and Levinson (Eds.) (2006) as the key to understanding not only language, but the essence of cognition, culture, and human interaction generally.

Interdisciplinary Interfaces

Needless to say that pragmatics is characterized by plenty of intralinguistic interfaces. Especially the relationships between pragmatics and grammar or syntax, or between pragmatics and semantics speak to the imagination. Bridging syntax and semantics in relation to pragmatics, there are numerous pragmatic studies of information structure. Also the connection with speech sounds, and in particular prosody, claims a lot of attention, and fields have developed such as morphopragmatics. But in the context of this encyclopedia, interdisciplinary connections seem more pertinent.

Irrespective of the position one takes, pragmatics does not exist without its interfaces with differently oriented endeavors in the study of language. As shown above, historically formative traditions that have shaped pragmatics are to be found in semiotics, philosophy, psychology, anthropology, and sociology. During the past few decades, fruitful interfaces have been established, re-established, or developed, with fields ranging from psychotherapy to artificial intelligence. Here I can just hint at a few noteworthy tendencies.

A first tendency that should not escape notice in the interface between pragmatics and neighboring disciplines, with a crucial impact on what language use is supposed to be as an object of investigation, is the forceful re-establishment of the connection between pragmatics and semiotics. While the linguistic pragmatics of the 1970s (roughly) had a tendency to grow away from the larger family of students of sign systems, today some kind of reintegration has become inevitable as a result of the growing awareness that language use itself cannot be fully understood as a separable mode of communication and expression, i.e., outside of a context in which also nonverbal aspects of communication play a vital role: facial expressions, gaze, gesture, proxemics, position in and movement through space, and the like. The importance of these aspects of what has come to be referred to as 'multimodality' has been strengthened in recent decades by the incremental availability of new communication tools such as mobile phones and computers, which affect basic modes of communication such as speech sounds or still and moving images in conjunction with verbal output.

Second, the identity of academic fields dealing with cognition, society, and culture – the traditional points of reference for those who accept the fundamental interdisciplinarity of pragmatics – has undergone significant changes. In particular, they have themselves interdisciplinarily grown together. Vygotsky's early twentieth-century recognition of the social grounding of the cognitive functioning of the human mind (see Mead, 1934; Vygotsky, 1978) – forgotten or discarded by many cognitivists – has been revived as a clear awareness of the cultural basis of human cognition and the social nature of the brain (Tomasello, 1999). Even truth has recently been described as a matter of socio-cultural practices (Dupret, 2011). Maybe the strongest expression of the merging of foci is to be found in the newly foregrounded notion of human sociality (mentioned above, cf Enfield and Levinson (Eds.), 2006), introduced as a cover term for socially oriented and culturally shaped cognition, capable of reflexively monitoring and hypothesizing other people's mental states (perceptions, knowledge, intentions) and their hypotheses about our own, and explaining the fundamental properties of language use in social interaction, whether in face-to-face contexts or at the level of large and complex social and political structures. In line with this is also renewed interest in the ritual functions of speech (Senft and Basso (Eds.), 2009) and in linguistic relativity, i.e., the idea that language structures and ways of looking at the world are not independent of each other (e.g., Gumperz and Levinson (Eds.), 1996). At the present state of research, pragmatics can no longer pretend to be the science of language use without taking the full complexity of human capacities and practices into account. There is no language use without cognition, culture and society, no cognition without culture and society, nor culture and society without cognition and language use. It is not surprising, then, that overlaps and interfaces with sociolinguistics, psycholinguistics, and anthropological linguistics are abundant, and that they sometimes lead to newly labeled endeavors such as cognitive pragmatics and intercultural pragmatics.

Finally, there is a large number of necessarily interdisciplinary areas of application. Thus pragmatics offers insights for forensic work (Coulthard and Johnson, 2007) and the treatment of patients with language disorders (Perkins, 2007). Pragmatically oriented linguists contribute to the discussion of socially and politically relevant issues. Sometimes these efforts bear the label 'critical discourse analysis' (Wodak and Meyer (Eds.), 2009). Sometimes they take the form of straightforward pragmatic analyses of ideological processes in language use (Verschueren, 2012). Ideology is a particularly appealing area since it mainly concerns meaning that is taken for granted and therefore carried along implicitly in sociopolitically relevant discourse – implicit meaning being one of the core phenomena studied by pragmaticians.

Trends and Perspectives

During the past 10–20 years, the study of language use has been revolutionized by technological developments, with an immediate impact on traditions of research in pragmatics. Clearly, investigations of interactional details, as in CA, would not be possible without audio- and videorecordings, the latter enabling close attention to aspects of multimodality as well. While these tools have been well-established for quite a while, developments in computer software are facilitators we could no longer do without. Computers, moreover, now make it possible to do research on vast corpora. This does not only affect the amount of data that can be studied, but also the way in which certain types of investigation are conducted. Needless to say that there are gains and losses in these developments. There are types of meaning generation processes – the stuff pragmaticians should ultimately be interested in – which are not easily operationalized in a way detectable by computer. But on the other hand, computational pragmatics gives access to phenomena that would otherwise remain hidden. The coming years will no doubt witness further developments that, if used with the necessary care, will keep improving or facilitating the way of working.

In addition to the burgeoning fields of corpus pragmatics and historical pragmatics, increasingly experimental methods are used to investigate meaning-generating processes (Noveck and Sperber (Eds.), 2004). Thus the extent to which speakers incorporate what is 'implicated' into their understanding of an utterance can be tested; for instance, whether the statement 'Max has three children' in response to 'How many children does Max have?' is understood that he has exactly three, can be tested by asking native speakers to judge its truth or falsity in a context where Max has four children. Experimental pragmatics faces the challenge of coping with two levels of verbal communication: instructions on how to perform a task, and the verbal task itself. Moreover, there is something paradoxical in the idea of an experimental pragmatics: pragmatics should keep on open mind for the widest possible range of variables that are involved in language use, while experimentation requires close control over a limited set of variables. Keen awareness of this, however, leads researchers to ever more refined experimental methods.

There are not only technologically facilitated and methodologically innovative trends and perspectives. We currently also see a growing awareness (as in an emerging tendency toward 'emancipatory pragmatics'; cf Hanks et al., 2009) of the limiting, predominantly western, cultural basis of thinking about language use. Conscious efforts to include different perspectives in pragmatic theorizing may lead to unexpected new insights in the years to come.

See also: Code Switching: Linguistic; Corpus Linguistics; Deixis; Discourse and Indentity; Evidentials in Linguistics; Gesture in Linguistics; Language and Society; Linguistic Anthropology; Linguistic Presupposition; Politeness and Language; Pragmatic Disorders; Tense, Aspect, Mood and Evidentiality, Linguistics of.

Bibliography

Ariel, M., 2010. Defining Pragmatics. Cambridge University Press, Cambridge.
Austin, J.L., 1962. How to Do Things with Words. Oxford University Press, Oxford.
Carston, R., 2009. The explicit/implicit distinction in pragmatics and the limits of explicit communication. International Review of Pragmatics 1, 35–62.
Clark, H.H., 1996. Using Language. Cambridge University Press, Cambridge.
Coulthard, M., Johnson, A., 2007. An Introduction to Forensic Linguistics: Language in Evidence. Routledge, London.
Dupret, B., 2011. Practices of Truth: An Ethnomethodological Inquiry into Arab Contexts. John Benjamins, Amsterdam.
Enfield, N.J., Levinson, S.C. (Eds.), 2006. Roots of Human Sociality: Culture, Cognition and Interaction. Berg, Oxford.
Fillmore, ChJ., 1975. Santa Cruz Lectures on Deixis, 1971. Indiana University Linguistics Club, Bloomington.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, New Jersey.
Goffman, E., 1963. Behavior in Public Places. Free Press, New York.
Grice, H.P., 1989. Studies in the Way of Words. Harvard University Press, Cambridge, Massachusetts.
Gordon, D., Lakoff, G., 1971. Conversational postulates. In: Papers from the Seventh Regional Meeting of the Chicago Linguistic Society, pp. 63–84.
Gumperz, J.J., 1982. Discourse Strategies. Cambridge University Press, Cambridge.
Gumperz, J.J., Hymes, D.H. (Eds.), 1972. Directions in Sociolinguistics: The Ethnography of Communication. Holt, Rinehart & Winston, New York.
Gumperz, J.J., Levinson, S.C. (Eds.), 1996. Rethinking Linguistic Relativity. Cambridge University Press, Cambridge.
Hanks, W.F., 2012. Modalities of co-participation. Journal of Pragmatics 44 (5), 563–565.
Hanks, W.F., Ide, S., Katagiri, Y. (Eds.), 2009. Towards an emancipatory pragmatics.Special Issue, Journal of Pragmatics 41, 1.
Horn, L.R., 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Schiffrin, D. (Ed.), Meaning, Form, and Use in Context: Linguistic Applications. Georgetown University Press, Washington, DC, pp. 11–42.
Huang, Y., 2007. Pragmatics. Oxford University Press, Oxford.
Jaszczolt, K.M., 2005. Default Semantics: Foundations of a Compositional Theory of Acts of Communication. Oxford University Press, Oxford.
Leech, G.N., 1983. Principles of Pragmatics. Longman, London.
Levinson, S.C., 1983. Pragmatics. Cambridge University Press, Cambridge.
Levinson, S.C., 2000. Presumptive Meanings: The Theory of Generalized Conversational Implicature. The MIT Press, Cambridge, MA.
Mead, G.H., 1934. Mind, Self, and Society: From the Standpoint of a Social Behaviorist. The University of Chicago Press, Chicago.
Mey, J.L., 2001. Pragmatics: An Introduction, second ed. Blackwell, Oxford.
Morris, C., 1938. Foundations of the theory of signs. In: Foundations of the Unity of Science: Towards an International Encyclopedia of Unified Science, vols. I, 2. The University of Chicago Press, Chicago.
Nerlich, B., Clarke, D.D., 1996. Language, Action, and Context: The Early History of Pragmatics in Europe and America, 1780–1930. John Benjamins, Amsterdam.
Noveck, I.A., Sperber, D. (Eds.), 2004. Experimental Pragmatics. Palgrave Macmillan, London.
Östman, J.-O., Verschueren, J. (Eds.), 1995. Handbook of Pragmatics. John Benjamins, Amsterdam. Annual installments since 1995, also online at: http://www.benjamins.com/online.

Perkins, M., 2007. Pragmatic Impairment. Cambridge University Press, Cambridge.
Recanati, F., 2010. Truth-conditional Pragmatics. Oxford University Press, Oxford.
Sacks, H., 1992. Lectures on Conversation, 2 vols. Basil Blackwell, Oxford.
Searle, J.R., 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.
Senft, G., Basso, E.B. (Eds.), 2009. Ritual Communication. Berg, Oxford.
Sperber, D., Wilson, D., 1986. Relevance: Communication and Cognition. Blackwell, Oxford.
Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard university Press, Cambridge, MA.
Verschueren, J., 1999. Understanding Pragmatics. Edward Arnold, London.
Verschueren, J., 2012. Ideology in Language Use: Pragmatic Guidelines for Empirical Research. Cambridge University Press, Cambridge.
Verschueren, J., Brisard, F., 2009. Adaptability. In: Verschueren, J., Östman, J.-O. (Eds.), Key Notions for Pragmatics. John Benjamins, Amsterdam, pp. 28–47.
Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.
Wodak, R., Meyer, M. (Eds.), 2009. Methods of Critical Discourse Analysis, second ed. Sage, London.

Pragmatism: Philosophical Aspects

Peter Godfrey-Smith, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Pragmatism is a philosophical movement which began in the late nineteenth century in the USA and exerted significant influence until around the time of John Dewey's death in 1952. Pragmatism fell from favor for much of the latter part of the twentieth century, but is now seeing a revival. The 'classical' pragmatists are C.S. Peirce, William James, and John Dewey. Their philosophies differed a good deal, but all defended broadly empiricist views of thought and knowledge which emphasize the role of thought in guiding action and reject various traditional empiricist psychological positions. Pragmatists also tend to reject correspondence theories of truth and attempt to ground human knowledge on a special 'foundational' set of beliefs. Prominent English-speaking philosophers associated with pragmatism in more recent years include W.V.O. Quine, Richard Rorty, Hilary Putnam, and Robert Brandom.

Pragmatism began in the late nineteenth century with the work of C.S. Peirce, William James, and John Dewey. Pragmatism's influence subsided after the death of Dewey in 1952, but is seeing a sustained revival. This article will first give a rough summary of characteristic themes in pragmatist philosophy, and then look more closely at Peirce, James, and Dewey. The final section will discuss the transformation of pragmatism in the late twentieth century and new forms it is taking in the twenty-first.

'Classical' Pragmatism

Charles S. Peirce, William James, and John Dewey are often referred to as the 'classical' pragmatists. All three were born and worked in the USA, and all combined their philosophical work with strong interests in other disciplines. The views of the 'classical pragmatists' have common themes but are also dissimilar in many ways. They developed views of thought, knowledge, and related topics which are empiricist in a broad sense but reject much of the psychological picture usually associated with empiricism. The empiricist tradition focused on the connections between thought and sensory experience. The classical pragmatists did so too, but in a way that treated thought in the context of its overall role in the life of a human agent, especially in the guidance of action. The pragmatists also thought that traditional empiricism has too passive and too atomistic a view of the mind.

Pragmatist philosophers generally reject attempts to understand human knowledge by appeal to some special set of 'foundational' beliefs that support all the others. For pragmatism, both specific beliefs and general methods of inquiry should be judged by their consequences, by their usefulness in achieving human goals. Although pragmatists stress the role of knowledge in guiding action and solving practical problems, it is an error to see pragmatism as claiming that the direction of research, or choices between rival theories, should be guided substantially by practical or commercial demands. A detailed account of the relationship between theoretical work and practical goals was only worked out successfully by Dewey, though, as discussed below.

The nature of truth has been a central topic for pragmatists, and a source of much trouble for them in the classical period. Pragmatists generally reject 'correspondence' theories of truth, theories which claim that a true belief or statement is one which represents the world as it really is. After rejecting correspondence, pragmatists have had a difficult time devising an alternative view of truth.

An additional unifying thread in classical pragmatism is more subtle and elusive. The classical pragmatists held that there is an important place in the universe for human choice and initiative (a theme more prominent in James and Dewey than in Peirce). In different ways they opposed philosophical systems which regard the world as 'finished,' 'complete,' or impervious to the effects of choice. Pragmatists also tend to hold 'humanistic' perspectives on matters involving morals and values, opposing both nihilistic views that reject all moral assessment as founded on illusion, and views that locate moral and other evaluative facts outside the everyday world of human striving and well-being.

Charles Sanders Peirce (1839–1914)

Peirce's career began well but ended very badly, without him ever having held a long-term academic position. His writings are very ambitious but often disorganized. He evaded obscurity, and at times sheer expiration, largely through James's support.

Peirce's most important works discuss the relations between doubt, inquiry, belief, and action. His central texts on these topics are two papers which appeared in the *Popular Science Monthly*, *The Fixation of Belief* in 1877 and *How to Make Our Ideas Clear* in 1878. Peirce also wrote a good deal about logic, measurement, the theory of signs (which he called 'semiotic'), probability, evidence, and some speculative cosmological ideas (Buchler 1955; Misak 2013).

Peirce argued that inquiry always begins with 'real and living doubt' which he distinguished from the feigned doubt of Descartes. Doubt prompts inquiry, which aims at belief. Peirce claimed that the essence of belief is the establishment of a habit of action. Alexander Bain had made this suggestion in

the mid-nineteenth century, within the framework of an associationist psychology. Peirce and other pragmatists made this link between belief and action central to their philosophies. Peirce believed that science is the most effective method for relieving doubt and acquiring useful habits of action. His defense of science emphasized the cumulative, long-term, and social properties of scientific inquiry, especially the tendency of scientific testing to produce a 'convergence' of belief among people who start from different positions.

In his paper *How to Make Our Ideas Clear*, Peirce argued for a principle which later became famous as 'the pragmatic maxim.' Peirce said that to work out what the meaning of a 'conception' is, we should work out "what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object." This principle is obscure and was later paraphrased by James, Peirce, and others in a number of conflicting ways. One possible interpretation is to see the famous 'maxim' as expressing a fairly familiar empiricist view – that our concept of any object is just our concept of its possible effects on our experience. Then the novelty in Peirce's paper lies in his attempt to derive this empiricist principle from an unusual starting point – a theory of the role of belief in guiding action. Some later discussions in Peirce and James treat the maxim itself as a claim about how the meaning of a belief or set of beliefs (a 'conception') is found in its effects on the believer's action.

In the same paper, Peirce also gave a definition of truth. He said that what we mean by 'the truth' is the opinion which is "fated to be ultimately agreed to by all who investigate," or in another formulation, the opinion that will be held at the limit of scientific inquiry. Both here and elsewhere, Peirce treated topics in epistemology and the philosophy of science in ways that placed much emphasis on the social context of testing and investigation.

William James (1842–1910)

James made pragmatism famous, especially in his 1907 book *Pragmatism*, the most widely read of all pragmatist writings. James had a brilliant and influential career based at Harvard, in both philosophy and psychology.

James wrote a number of important works before he began calling himself a 'pragmatist' and aligning his views with those of Peirce. In his earlier philosophical works, James argued that we have the right to actively choose sides on momentous, unresolved philosophical issues that will affect how we live our lives (James, 1897/1956). Choices made on these issues will be strongly affected by our individual temperaments; we should not expect all individuals to find the same philosophies appealing. The issues James applied this doctrine to included the existence of God, the freedom of the will, and the reality of moral facts. In this earlier period, James also published a landmark work in psychology, his two-volume *Principles of Psychology* (1890/1950).

From 1898, James began to embed his characteristic themes within an overtly 'pragmatist' framework. He gave credit to Peirce's role in discussions in Cambridge in the early 1870s, in a social circle called the 'Metaphysical Club' (Menand, 2001). To a greater extent than Peirce and Dewey, James took pragmatism to be continuous with the tradition of English-speaking empiricist philosophers such as David Hume and J.S. Mill. James publicized pragmatism with great energy, and transformed it in ways that induced unease in Peirce. James presented pragmatism as a way to avoid the errors of the two key rival philosophies of his day – overly scientistic and materialist empiricism on one side, and sentimental, over-optimistic religious idealism on the other. James endorsed a version of the pragmatic maxim, described earlier, taking it to be a method for finding the real content in philosophical doctrines and disagreements. James argued that when the maxim is applied to many standard philosophical debates, it does *not* typically show that these lack content; instead the maxim reveals hidden and substantial issues that are at stake. These issues, for James, often concern the long-term direction being taken by the universe, and the role of human choice in affecting how things turn out. James saw pragmatism as a way of distilling the genuine human significance from obscure philosophical theories and debates.

James was a less scientistic thinker than Peirce, and more individualistic than Peirce and Dewey. James talked of assessing ideas in terms of their 'cash value,' an unfortunate phrase which led many to interpret pragmatism as vulgar and anti-intellectual. James also invited a stream of criticism through his tendency to discuss the nature of truth in a vague and simplistic way. He said at various times that true ideas are just those that can be verified, or those that are useful. Some of the criticisms of James rested on misinterpretations, but there is no question that he wrote about this topic imprecisely. Despite this, James's writings have endured. His discussions of the role of individual temperament in philosophical choices never seem dated, for example, and James's attractive personality, imagination, and boundless energy radiate from the pages of his work.

John Dewey (1859–1952)

Dewey's work is the culmination of the 'classical' period in the pragmatist tradition. In the course of a remarkably long and productive career, Dewey published in virtually all areas of philosophy and in psychology, education, and politics. He wrote both popular, accessible works and dense, uncompromising ones. Dewey spent his most important years first at the University of Chicago, where he worked in philosophy, education, and psychology, and then at Columbia, where he focused more on philosophy and was also active in politics.

Dewey's work went through several phases. Early in his career he accepted an idealist philosophy of the type influenced by Hegel, and also had Christian interests. Around the turn of the century, he turned toward a more scientific approach to philosophy, influenced by Darwin and by James's *Principles of Psychology*. Around this time, Dewey also dropped his commitment to Christianity. Dewey came to reject all supernatural forms of religion, and did not share James's sympathy for mystical ideas, but he continued to see value in some kinds of religious experience (Dewey, 1969–72, 1976–83).

The scientifically oriented philosophy Dewey developed after 1900 was initially fairly similar to James's pragmatism,

and some of this work influenced and was acknowledged by James. In later decades, Dewey reworked pragmatist themes within what he called a 'naturalistic' framework – a framework which sets out from a biological description of living organisms and their relations to their environments, and emphasizes also the importance of social interaction in human life. Dewey argued that intelligence is a means for humans to transform their environments in order to deal with the problems posed by uncertainty and change in natural events.

In *Experience and Nature* (1925) perhaps his greatest work, Dewey defended his naturalistic view of mind and knowledge, and criticized the philosophical tradition for its postulation of false divides or 'gulfs' – between mind and matter, thought and object, theoretical and practical. The philosophical tradition is plagued by 'dualisms' which lead to pseudoproblems, problems of establishing contact between realms that should never have been set against each other in the first place. The source of these dualisms is a 'split in being' established by the ancient Greeks, a split between the 'perfect, permanent, self-possessed' and the 'defective, changing, relational.' Dewey sought to replace these dualisms with a view based on various kinds of 'continuity' – between mind and nature, between organism and environment, and between cognition and simpler biological capacities. For Dewey, these natural continuities also provide the material needed to resolve oppositions between fact and value. Much of Dewey's later work is a mixture of careful system-building based on assertions of 'continuity' and sweeping historical surveys, tracing the histories of crucial philosophical errors back and forward through thousands of years (Dewey, 1981–90).

Dewey defended a social theory of mind, claiming that thought only exists within a language-using community. He argued that the primary role of science is to help human societies deal with problems and control their environments. But Dewey thought that it was unscientific for scientists to direct their work according to *specific* practical problems. Rather, science is the study of a special subset of properties of natural affairs – relations and connections, called by Dewey 'instrumental' properties. Science is most successful in expanding our capacities for problem-solving and transforming our environment when it is directed on the study of the instrumental features of nature in an open-ended way, unconstrained by immediate practical applications.

Dewey rejected what he called the 'spectator theory of knowledge,' the view that the genuine knower is someone who registers what is going on but does not intervene. Instead, the proper role of knowledge is to enable humans to transform their situations in beneficial ways. This also led Dewey into some of the same arguments that James had with defenders of correspondence or 'copy' theories of truth. Dewey thought that theories of truth as correspondence were aimed at inventing magical relations between thought and the world in order to overcome what he saw as a nonexistent problem – the problem of how the mind and the external world could have any contact with each other. He also thought correspondence theories belong with a spectator view of the role of knowledge. He did not take seriously the idea that 'copying' the world can be a *means* to transforming it; he saw these inevitably as rivals.

Dewey wrote often of the role played by knowledge in 'transforming' or 'reconstructing' the world, but this 'transformation' should not be understood as implying an idealist metaphysical view. Dewey thought that knowledge is a factor in changing things in the world because knowledge has a role in guiding *action*, which transforms things by means of physical relationships. I said earlier that a second thread in the classical pragmatists, in addition to their treatment of thought and action, is a resistance to a 'closed' conception of the universe and an assertion of the significance of human choice. In Dewey's work, this theme, which was developed in a vague and overly cosmic way in James, is given a more down-to-earth treatment. Dewey insisted on the reality of the progressive reshaping of the world by human choice, though this reshaping is limited in scope, contingent in its effects, and draws on craft skills as well as ideas.

Dewey was a significant figure in US political and social thought (Westbrook, 1991). He defended a version of liberalism and wrote extensively about the proper structure of a democratic society. At various times he was attacked by both the right and the left. He hoped for a more 'democratic' economic order, but kept communism at arm's length. Dewey's best known political activity was his chairing of an international inquiry into Stalin's trial of Trotsky. The inquiry entitled its report 'Not Guilty.'

At Chicago, Dewey worked closely with George Herbert Mead, a founder of social psychology in the USA. Mead (1934) argued that human individuality is a product of social embedding; that both cognition in general and specific mental developments such as a sense of self emerge as products of various kinds of symbolic behavior within a community. Dewey also wrote about art and took a lifelong interest in education. When he was at Chicago he established an experimental school. Dewey favored problem-solving as an approach to all types of learning, and his ideas remained influential in education during decades when his name had almost dropped out of English-speaking philosophy.

After Dewey

Dewey's ideas, and pragmatism in general, subsided from philosophical discussion for several decades after Dewey's death in 1952. A primary reason for this was the growing focus on formal logic and the philosophy of language in US philosophy after World War II, prompted especially by the immigration of European philosophers and logicians who defended a rigorous 'analytic' style of philosophy. After Peirce, the pragmatists had not taken an interest in formal logic. Dewey also had idiosyncratic and rather opaque ideas in the philosophy of language, which rapidly came to seem dated.

One of the leaders of the new analytic philosophy, W.V.O. Quine, had significant effects on the perceived viability of pragmatism, and on its interpretation, through some brief remarks in famous works, especially *Two Dogmas of Empiricism* (1951). There, Quine, in response especially to Rudolf Carnap, endorsed what he saw as a pragmatist view of belief change. Quine's emphasis was on conceptual flexibility; he held that even the most basic parts of our network of beliefs (including logical principles) could rationally be modified in response to unexpected experiences. Quine's influence helped keep the idea of a pragmatist epistemology alive, but in a way that

foregrounded some pragmatist themes and backgrounded others in a new way (Godfrey-Smith, 2013). Quine treated beliefs as tools for prediction – in this way, he was closer to the mainstream empiricist tradition. He did not give the link between thought and action anything like the role it had in classical pragmatism.

The most important figure in the revival of pragmatism has been Richard Rorty (1931–2007), who also carried the influence of pragmatism outside of philosophy into neighboring disciplines. Rorty viewed himself as carrying on the tradition of Dewey, and to a lesser extent James. Rorty called himself an 'antiessentialist' about traditional philosophical concepts such as truth, knowledge, and justice. Antiessentialism in this sense is the denial that there is anything general to say, from a philosophical point of view, about the *nature* of truth, the *nature* of justice, and so on. It is possible and worthwhile to give a theory of how the word 'true' functions in ordinary discourse, but pointless to give a theory of truth that goes beyond this (Rorty, 1982).

The continuity between Rorty's ideas and Dewey's is a matter of controversy. Dewey's mature work, as outlined earlier, includes a mixture of sweeping historical stories aimed at dissolving philosophical problems, and careful system building of his own. Rorty endorses the dissolving but not the system building. However, a strong case can be made that Dewey's critical points and his attempts to 'get over' standard debates usually depend on his positive philosophical theories. Moreover, James, as discussed earlier, used pragmatism to breathe new life into metaphysical problems, not to deflate them. As Rorty's views on traditional philosophical questions are so focused on dissolving and deflating, it is in some ways more appropriate to associate Rorty with the later work of Ludwig Wittgenstein rather than with pragmatism. Rorty, however, made a substantial contribution to the development of pragmatist ideas by claiming that the right approach to the problem of truth for the pragmatist philosopher is not to identify the true with the useful, or with the verifiable, but to give a 'deflationary' view of truth (Rorty, 1991). Deflationism, which was developed in the late twentieth century in other parts of the philosophy of language, holds that, to say that a sentence is true is not to say that the sentence has a special correspondence relation to the world, or that it has a special kind of usefulness, but is just to *assert* the sentence, perhaps in a way sensitive to the context of the conversation. We have the word 'true' in our language because it allows us to abbreviate, give credit, and generalize – 'Everything you said about Toyotas is true.' The word 'true' is no more than a conversational tool of that kind.

Many other philosophers in the late twentieth and early twenty-first centuries who have allied themselves with pragmatism have been largely concerned with problems of language, meaning, and truth. These philosophers have also taken on board the emphasis on flexibility in belief change and the rejection of foundationalist views in epistemology that were present in the classical pragmatists and foregrounded by Quine. In some cases, though by no means all, a link to pragmatism is seen as part of an attempt to dissolve or 'get over' traditional debates about mind and knowledge.

Hilary Putnam was for some years a leading defender of a strongly 'realist' set of views about the relation between thought and the world, but he changed his mind, and developed a position he called 'internal realism,' which rejects the correspondence theory of truth and which Putnam allied to pragmatism. Putnam argued that it is not possible for a theory to pass every possible test, and command consensus when all the evidence is in, and yet be false; this view resembles Peirce's theory of truth. Putnam has recently reassessed these issues yet again. Robert Brandom (2010) defends a project he calls 'analytic pragmatism.' This is an approach to language that focuses on the *use* of words and linguistic forms in discourse, as contrasted with making the *meaning* of words primary. This emphasis on use owes much to Ludwig Wittgenstein. Huw Price (2003) also defends a use-based approach to language, rejects the correspondence theory of truth, and sees his pragmatism as strongly deflationary of metaphysical issues, advocating a 'quietist' approach – a resolutely nontheoretical stance – toward many traditional debates. Philip Kitcher (2012) also wants philosophy to move, with the aid of pragmatism, past standard debates in epistemology and metaphysics but is less deflationary about social and moral philosophy, where more substantial work is to be done. Other contemporary philosophers with significant connections to pragmatism include Simon Blackburn, Susan Haack, Cheryl Misak, and Arthur Fine.

In my outline of the 'classical' pragmatists, I emphasized their treatment of the link between belief and action. This, for the classical pragmatists, was an innovation that took them beyond the mainstream empiricist tradition. For better or worse, from Quine onward this theme subsided in 'pragmatist' philosophy. The emphasis instead has been on conceptual flexibility (Quine), rejection of the correspondence theory of truth and representation-based approaches to language (Rorty, Brandom, Price), and finding ways to get beyond sterile philosophical debates (Rorty, Price, Kitcher, Fine). But as the early sections of this article may also have made clear, wholesale opposition to philosophical theorising had little role in classical pragmatism. James, in particular, thought that the significance of standard philosophical debates was, if anything, *under* estimated. Dewey's criticisms of the philosophical tradition were not offered in a free-standing form, isolated from positive views on philosophical topics; the critical and constructive sides of his thought were tied tightly together. So the themes of recent pragmatists differ a fair amount from those of the originators of the movement, but at all stages, including the 'classical,' pragmatism has been a diverse and mutable collection of ideas.

See also: Dewey, John (1859–1952); Empiricism, History of; Irrationality: Philosophical Aspects; James, William (1842–1910); Mead, George Herbert (1863–1931); Social Science, The Idea of; Wittgenstein, Ludwig (1889–1951).

Bibliography

Brandom, R., 2010. Between Saying and Doing: Towards an Analytic Pragmatism. Oxford University Press, Oxford.

Buchler, J. (Ed.), 1955. Philosophical Writings of Peirce. Dover, New York.

Dewey, J., 1925. Experience and Nature vol. 1 of Dewey (1981–90).

Dewey, J., 1969–72. In: Boyston, J.A. (Ed.), The Early Works. Southern Illinois University Press, Carbondale, IL.

Dewey, J., 1976–83. In: Boyston, J.A. (Ed.), The Middle Works. Southern Illinois University Press, Carbondale, IL.
Dewey, J., 1981–90. In: Boyston, J.A. (Ed.), The Later Works. Southern Illinois University Press, Carbondale, IL.
Godfrey-Smith, P., 2013. Quine and pragmatism. In: Harman, G., Lepore, E. (Eds.), A Companion to W.V.O. Quine. Wiley-Blackwell, Hoboken, NJ.
James, W., 1890/1950. Principles of Psychology. Dover, New York.
James, W., 1897/1956. The Will to Believe and Other Essays in Popular Philosophy. Dover, New York.
James, W., 1907. Pragmatism: A New Name for Some Old Ways of Thinking. Longmans, Green, and Co, New York.
Kitcher, P.S., 2012. Prelude to Pragmatism. Oxford University Press, Oxford.
Mead, G.H., 1934. In: Morris, C.W. (Ed.), Mind, Self, and Society. University of Chicago Press, Chicago.
Menand, L., 2001. The Metaphysical Club: A Story of Ideas in America. Farrar, Strauss, and Giroux, New York.
Misak, C., 2013. The American Pragmatists. Oxford University Press, Oxford.
Price, H., 2003. Truth as convenient friction. Journal of Philosophy 100, 167–190.
Putnam, H., 1990. Realism with a Human Face. Harvard University Press, Cambridge, MA.
Quine, W.V.O., 1951. Two dogmas of empiricism. Philosophical Review 60, 20–43.
Rorty, R., 1982. Consequences of Pragmatism: Essays 1972–1980. University of Minnesota Press, Minneapolis, MN.
Rorty, R., 1991. Objectivity, Relativism and Truth: Philosophical Papers, vol. I. Cambridge University Press, Cambridge, UK.
Westbrook, J., 1991. John Dewey and American Democracy. Cornell University Press, Ithaca, NY.

Pragmatist Social Thought, History of

Hans Joas, Humboldt University, Berlin, Germany; and University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Pragmatism' is the name of an originally American philosophical movement; its classical figures are C.S. Peirce, W. James, J. Dewey, and G.H. Mead. The basic idea of pragmatism is that all cognition is ultimately grounded in practical problem situations. Action is thus conceived as based on 'habits,' but in need of constant creative reorganization. Pragmatism emphasizes a nonfoundationalist understanding of philosophy, the fallible nature of truth, the social nature of the self, the importance of inquiry in all fields of human life, and metaphysical pluralism (R. Bernstein). Pragmatism contributed among other things to the development of semiotics, functionalist psychology, the phenomenology of moral and religious experience, social psychology, and an antideterminist metaphysics. In the writings of Dewey, a radical version of democracy became articulated as the normative core of pragmatism. Although influential in a wide range of social scientific disciplines, pragmatism was marginalized in academic philosophy from the 1950s to the 1970s. In recent years, it has experienced a considerable revival (Rorty, Putnam, Habermas).

'Pragmatism' is the name of an originally American philosophical movement; its classical figures are C.S. Peirce, W. James, J. Dewey, and G.H. Mead. The basic idea of pragmatism is that all cognition is ultimately grounded in practical problem situations. Action is thus conceived as based on 'habits,' but in need of constant creative reorganization. Pragmatism emphasizes a nonfoundationalist understanding of philosophy, the fallible nature of truth, the social nature of the self, the importance of inquiry in all fields of human life, and metaphysical pluralism (R. Bernstein). Pragmatism contributed among other things to the development of semiotics, functionalist psychology, the phenomenology of moral and religious experience, social psychology, and an antideterminist metaphysics. In the writings of Dewey, a radical version of democracy became articulated as the normative core of pragmatism. Although influential in a wide range of social scientific disciplines, pragmatism was marginalized in academic philosophy from the 1950s to the 1970s. In recent years, it has experienced a considerable revival (Rorty, Putnam, Habermas).

Pragmatism is the name usually given to the classical philosophical movement in the United States. It emerged in the last decades of the nineteenth century and achieved a certain hegemony in the country's intellectual life during, and immediately after, the so-called Progressive Era (1896–1914). From the 1930s, and even more after 1945, it was largely displaced by other currents of thought in philosophy, the social sciences, and public political discourse. American pragmatism also attracted considerable attention in Europe, especially around 1910, although exposition and critical discussion of pragmatism displayed widespread misunderstandings and the tendency to depreciate it by reducing it to an expression of alleged American national characteristics.

Pragmatism Defined

The everyday sense of the term 'pragmatic' certainly contributed to these misunderstandings, connoting as it does a kind of 'muddling through' that is oriented to immediate requirements, disregards theoretical or moral principles, and treats the given features of the situation simply as part of a calculation. The term pragmatism has the same Greek roots as praxis, practical, etc. The founder of pragmatism, Charles Sanders Peirce (1839–1914), coined the term as a result of his reflections on Kant's use of the terms 'pragmatic' and 'practical.' Peirce's lectures and writings around 1878 are now regarded as the original documents of pragmatism. These first became known only to narrow circles of intellectuals in Cambridge, Massachusetts, and did not achieve wider notice until some 20 years later, when William James (1842–1910) referred to them in some lectures, but above all through his lectures on 'pragmatism' published in 1907. Peirce came to distance himself from James' pragmatism and to call his own philosophy 'pragmaticism.'

Along with Peirce and James, the inner core of pragmatism is usually taken to include John Dewey (1859–1952) and George Herbert Mead (1863–1931). The differences between all these thinkers were so evident, not least to the 'pragmatists' themselves, that their subsumption into a single school or movement has constantly been questioned. The most widely accepted view, however, has been that despite the differences there is enough of a core of shared views to justify speaking of 'pragmatism' as a distinct philosophical orientation.

What are the basic themes of pragmatism? Peirce's approach starts from his critique of Descartes' methodological principle of radical doubt and the resultant program of making the thinking ego's certainty of itself into the firm foundation of a new philosophy. The pragmatist questions the meaningfulness of Cartesian doubt, not in order to defend unquestionable authorities against the emancipatory claim of the thinking ego, but arguing for a more substantial doubt, that is, the grounding of cognition in real problem situations. In pragmatism, the guiding idea of the doubting ego is replaced by the idea of a cooperative search for truth in order to overcome real problems of action. Real doubt occurs in action, conceived as a cyclical succession of phases. Thus, every perception of the world and every action in it are anchored in an unreflective belief in self-evident conditions and successful habits. But these habitual ways of acting constantly come up against the resistance of the world, which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03192-5

is seen to be the source of the destruction of unreflective expectations.

The resultant phase of real doubt leads to a reconstruction of the interrupted context. Perception must grasp new or different aspects of reality; action must attach itself to other elements of the world or reorganize its own structure. This reconstruction is a creative accomplishment by the actor. If he or she succeeds, via a change in perception, in acting differently and thus getting going again, then something new has come into the world: a new way of acting that can be institutionalized and itself become an unreflectively followed routine. Thus, pragmatists see all human action in the opposition between unreflective habits of action and creative accomplishments. This also means that creativity is seen here as characterizing achievements in situations that demand a solution, rather than the unconstrained creation of something new without a constitutive background in unreflective habits.

From this basic model of pragmatism, in which action and cognition are combined in a particular way, we can derive the other central claims of pragmatism. In the metaphysics of pragmatism, reality is not deterministic: rather, it enables and demands creative action. In the epistemology of pragmatism, knowledge is not the reproduction of reality but an instrument for dealing with it successfully. The semantics of pragmatism locates the meaning of concepts in the practical consequences for action resulting from their use or their difference from other concepts. Thus in pragmatism's theory of truth, the truth of sentences can only be determined by way of a process of agreement on the success of action based on them, and not on, say, their correspondence with an uninterpreted reality. The misunderstanding of pragmatism as principally a movement aiming at the destruction of the ideal of true knowledge resulted primarily from the isolation of individual sentences, such as some of William James's statements about truth, from the whole complex of pragmatist thought.

Main Variants of Historical Pragmatism

The leading representatives of pragmatism contributed to different fields of inquiry. Peirce was primarily interested in developing a general theory of scientific knowledge and a broadly conceived theory of signs of semiotics. His multifaceted work, which does not deal with questions of social theory, includes important thoughts on the intersubjective use of signs and on the creative production of hypotheses ('abduction'). This theory of signs, particularly in its emphasis on the 'discourse' of scientists in an experimental community, has become very influential.

William James at first worked mainly in psychology, in which he saw the prospect of a way out of the dilemma between a religiously based belief in the free will of the moral agent and the scientific image of the world as a universe governed by causal processes. The solution, for James, lay in the functionality for the survival of the human organism in its environment, which he saw in the human capacity to pay deliberate attention to perceptual impressions and to choose between alternative courses of action. A 'functionalist' psychology could proceed by understanding all mental achievements in terms of their function for the organism's active mastery of its environment.

James elaborated his antideterminist stance in an influential article on *The Will to Believe*, in which he defended the legitimacy of religious experience against scientistic skepticism. This essay also became the point of departure for his wide-ranging study of the varieties of religious experience, in which he founded the empirical study of religion not on religious institutions or theological doctrines, but on the phenomenology of religious experience. In his last years, he began to use the phenomenology of experience as a guideline for the elaboration of a pluralist, nondeterminist metaphysics.

John Dewey, whose thought at first developed rather independently of the first pragmatists, abandoned his earlier neo-Hegelian approach during the 1890s and linked up with the logical and psychological aspects of pragmatism as developed by Peirce and James. He constructed a philosophy that extended the core ideas of pragmatism not only into all the traditional domains of philosophy (metaphysics, logic, ethics, aesthetics) but also, and in particular, into the fields of educational theory and social and political philosophy. He gave more explicit emphasis than the others to a radical version of democracy as the normative core of pragmatism. Taking part in a great number of political controversies, he became one of the most influential American intellectuals ever and *the* representative thinker of the United States.

George Herbert Mead, a friend of Dewey, did most to develop James' strategy of the translation of pragmatist themes into the program of a biologically based empirical social science. His decisive contribution to social theory consists of his theory of the specific features of human communication and hence the attempt to thematize the constitution of personality structures in the dynamics of interpersonal relationships. Mead attacked the assumption of a presocial substantial self and replaced it by a theory of the genesis of the self in which even the interaction of a person with himself or herself is conceptualized to be the result of social structures. Mead also pursued this line of thought in the direction of problems of cognitive development, such as the constitution of permanent objects in experience and the constitution of structures of time. His ethical theory and antideterminist philosophy of history remain rather neglected.

Thus it can be said that pragmatism has never been a clearly delimited school of thought with a fixed research program; it shows, however, the characteristics of an intellectual movement or, at least, constellation with the common elements of antifoundationalism, the fallible nature of truth, the social nature of the self, the importance of inquiry in all fields of human life, and (metaphysical) pluralism (see Bernstein, 1992, 2010). During the first decades, pragmatism exerted a considerable influence on a variety of disciplines in the United States (functionalist psychology, institutionalist economics, educational reform, law, theology, political theory, the Chicago school of sociology). After the 1930s, this influence clearly waned, reaching a nadir in the 1950s and the 1960s (An indicator of this development is probably the fact that the 1968 edition of this encyclopedia did not contain an entry on pragmatism). One of the few currents that kept the heritage of pragmatist social thought alive was the school of symbolic interactionism in sociology, which produced a rich amount of empirical work

mostly on interpersonal relationships, processes of emergence and definition of deviant and subcultural behavior, social movements, and the division of labor in professional organizations. There were other main figures of American sociology at that time in whose work pragmatist orientations can easily be detected (C. Wright Mills, Philip Selznick).

Contemporary Pragmatist Thinking

Since the late 1970s, a renaissance of pragmatist philosophy has been underway. Apart from thinkers like Richard Bernstein and John Smith, who had continuously pursued a pragmatist agenda, two thinkers have probably contributed most to this renaissance. One was Richard Rorty who had for a long time tried to articulate pragmatist motifs within the tradition of analytical philosophy before breaking (more or less) with it and proposing a variant of postmodern thinking that he calls pragmatism. Although Rorty was a perceptive social critic, he had no social theory and actively repudiated the need for it. His philosophy has sparked a vast critical literature and reappropriation of the historical pragmatists; his main rival in these debates was Hilary Putnam in whose work one can find a thorough and significant restatement of pragmatist philosophy.

The other main originator of this renaissance is Jürgen Habermas. Together with Karl-Otto Apel, these two German thinkers gave a new twist to the winded history of the European reception of pragmatism. From the beginnings of the twentieth century, very diverse schools of European thinking had used elements from American pragmatism for their own purposes (e.g., French Catholic Modernism, Italian proto-Fascism, German Philosophical Anthropology, and others). Other European thinkers came close to pragmatism in some aspects of their work (e.g., Max Scheler, Arnold Gehlen, Martin Heidegger, and Ludwig Wittgenstein) without spelling out their relationship with it. Peirce's ideas about the 'discourse' of scientists has been of decisive importance for the 'discourse ethics' developed by Apel and Habermas. Mead's understanding of symbolically mediated interaction has been crucial for the 'Theory of Communicative Action' by Habermas; he declares Mead the main inspiration for 'the paradigm shift from purposive to communicative action,' which he proposes in sociological theory.

Habermas' faithfulness to the pragmatist heritage has been disputed almost as much as Rorty's. There are obviously other possibilities for pragmatist social thought than his combination of a theory of communication with systems theory or than his restriction of the tasks of moral theory to the program of discourse ethics. One can claim (Joas, 1996, 2000) that the most fundamental contribution of pragmatist social thought lies not in its emphasis on communication, but in its specific understanding of the creativity of human action and the emergence of value commitments. This understanding distinguishes pragmatist social thought from approaches focusing on either rational action or normatively oriented action, or from a systems theory or structuralism, which disconnect the dynamics of human action from the development of social wholes. On the basis of pragmatism understood as a theory of the 'situated creativity' of human action, it becomes possible to reconceptualize rationality and normativity and to develop a theory of macrosocial processes and social change (Ansell, 2011) prepared by the political theory and philosophy of history in the works of the classical pragmatists.

See also: Critical Theory; Dewey, John (1859–1952); Frankfurt School: Institute for Social Research; James, William (1842–1910); Mead, George Herbert (1863–1931); Political Theory, History of; Postmodernism: Philosophical Aspects; Sociology, History of.

Bibliography

Ansell, C.K., 2011. Pragmatist Democracy: Evolutionary Learning as Public Philosophy. Oxford University Press, Oxford.
Bernstein, R.J., 1971. Praxis and Action. Duckworth, Philadelphia, PA.
Bernstein, R.J., 1992. The resurgence of pragmatism. Social Research 59, 813–840.
Bernstein, R.J., 2010. The Pragmatic Turn. Polity, Cambridge.
Blumer, H., 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.
Dewey, J., 1967–1987. Early/Middle/Late Works. Southern Illinois University Press, Carbondale, IL.
Dickstein, M. (Ed.), 1998. The Revival of Pragmatism: New Essays on Social Thought, Law, and Culture. Duke University Press, Durham, NC.
Habermas, J., 1984/1987. The Theory of Communicative Action, 2 vols. Beacon Press, Boston, MA (English translation by T. McCarthy).
Habermas, J., 1992. Postmetaphysical Thinking. MIT Press, Cambridge, MA.
James, W., 1987/1992. Writings 1878–1899/1902–1910. Library of America, New York.
Joas, H., 1996. The Creativity of Action. University of Chicago Press, Chicago, IL (English translation by J. Gaines, P. Keast).
Joas, H., 2000. The Genesis of Values. Polity Press, Cambridge, UK (English translation by G. Moore).
Mead, G.H., 1934. Mind, Self, and Society from the Standpoint of a Social Behaviorist. University of Chicago Press, Chicago, IL.
Mead, G.H., 1964. Selected Writings. Bobbs-Merrill, Indianapolis, IN.
Peirce, C.S., 1931–1960. Collected Papers. Harvard University Press, Cambridge, MA.
Rorty, R., 1979. Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.
Selznick, P., 1992. The Moral Commonwealth: Social Theory and the Promise of Community. University of California Press, Berkeley, CA.
Thayer, H.S., 1981. Meaning and Action: A Critical History of Pragmatism. Hackett, Indianapolis, IN.

Prefrontal Cortex

Bryan Kolb, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.M. Fuster, volume 17, pp. 11969–11976, © 2001, Elsevier Ltd.

Abstract

The prefrontal cortex is the region in the cerebral cortex that includes the tissue that has projections from the mediodorsal nucleus of the thalamus and/or the amygdala and that has extensive projections to the basal ganglia and the rest of the cerebral cortex. Primates have an expanded prefrontal cortex and may have subregions not present in other mammals. The prefrontal cortex participates in extended neural networks with other cortical regions, and especially the posterior parietal cortex. The prefrontal cortex supports cognitive functions that are necessary to organize behavior in time and in context. This includes a form of intelligence, sometimes referred to as fluid intelligence. There are five general categories of symptoms associated with disease of the prefrontal cortex: disturbances of motor function; loss of divergent thinking; environmental control of behavior; temporal memory; and emotion and social behavior.

Introduction: What Is Prefrontal Cortex?

The term 'prefrontal' has an uncertain origin dating back to at least the early 1900s and over time has meant different things to different people. One way to define cortical regions is by considering their thalamic afferents. Thus, the lateral geniculate nucleus defines primary visual cortex, whereas the medial geniculate nucleus defines primary auditory cortex. Using this logic, Rose and Woolsey (1948) proposed that the cortical projection of the mediodorsal nucleus of the thalamus (MD) would define the 'orbitofrontal' cortex of mammals. These authors saw their definition as a way of solving the problem that the frontal region of primates beyond the motor and premotor cortex has a prominent layer IV, and thus was called frontal granular cortex, whereas in nonprimates layer IV is indistinct or absent. Because the MD projection was constant across the mammals they examined and the granular cortex was not, they believed that their definition of orbitofrontal cortex solved the apparent conflict.

One major problem with the Rose and Woolsey definition is that MD is not a homogenous structure and shows large differences in organization and connectivity across mammalian species. The differences in connectivity are not just in the afferents to MD but also in the efferents from different regions of MD. Just because MD projection cortex is present across mammalian phylogeny, it does not mean that the cortex has the same function. We thus must use other criteria in parallel with the Rose and Woolsey definition.

Reep (1984) proposed a different solution to the definitional problem. He proposed that the connections with the amygdala and certain brainstem regions such as the dopaminergic input from the ventral tegmentum, as well as cytoarchitecture be used to make cross species comparisons. Using these criteria he identified five different subregions (see **Figure 1**) present in monkeys, cats, and rats: (1) a medial infralimbic region receiving input from the amygdala but not MD; (2) a medial prelimbic region (Brodmann's areas 25 and 32) receiving input from MD, amygdala, and dopaminergic brainstem; (3) a region of agranular cortex overlying the claustrum receiving input from MD, amygdala, and dopaminergic brainstem; (4) the anterior cingulate region receiving input from MD and the anterior thalamus; and (5) a region of cortex lying between prelimbic and insular areas that receives input from MD but not the amygdala. He noted that other regions, such as area 9/46 of the monkey and human, do not appear to follow any obvious patterns across species and thus may be unique to one or a few orders of mammals.

Wise (2008) has argued forcefully that the unique presence of granular cortex in primates implies that primates have frontal regions that are unique and not found in other species. Pushing this idea further we can argue the common ancestor of living mammalian orders had a prototypical 'prefrontal' cortex, which has been modified and expanded as the brains grew larger. The expansion of prefrontal areas likely began by subdividing the prototypical regions and the eventual development of granular cortical regions may indeed have led to novel functions. Pandya and Yeterian (1985) proposed that the expansion of prefrontal subareas is related directly to the increase in sensory areas and the increase in posterior parietal cortex. Primates are also unique in the expansion of the posterior parietal regions (*see* Cortical Areas Engaged in Movement: Neuroimaging Methods) so it is not a leap to conclude that the prefrontal granular and posterior parietal regions have unique functions.

The granular areas in primates include dorsolateral prefrontal cortex (Brodmann's area 46), dorsal and medial prefrontal areas (different parts of area 9), ventral prefrontal cortex (areas 12, 44, 45, and 47), frontal pole cortex (area 10), and rostral orbitofrontal cortex (parts of areas 11, 13, and 14) (see **Figure 2**). These regions can be functionally divided into four regions namely dorsolateral prefrontal cortex, dorsomedial prefrontal cortex, ventromedial prefrontal region, and orbital prefrontal region, in addition to the adjacent anterior cingulate cortex, as illustrated in **Figure 3**.

The frontal cortex as a whole has a special relationship with the basal ganglia because although virtually all of the cerebral cortex projects to the basal ganglia, the frontal cortex receives most of the return projections via the thalamus (Middleton and Strick, 2000). These projections are present in species as diverse as monkeys and rats and, in addition, in both species there are a number of parallel, apparently functionally segregated circuits (Alexander et al., 1990; Middleton and Strick, 2000; Uylings et al., 2003). For example, there are cortico-basal

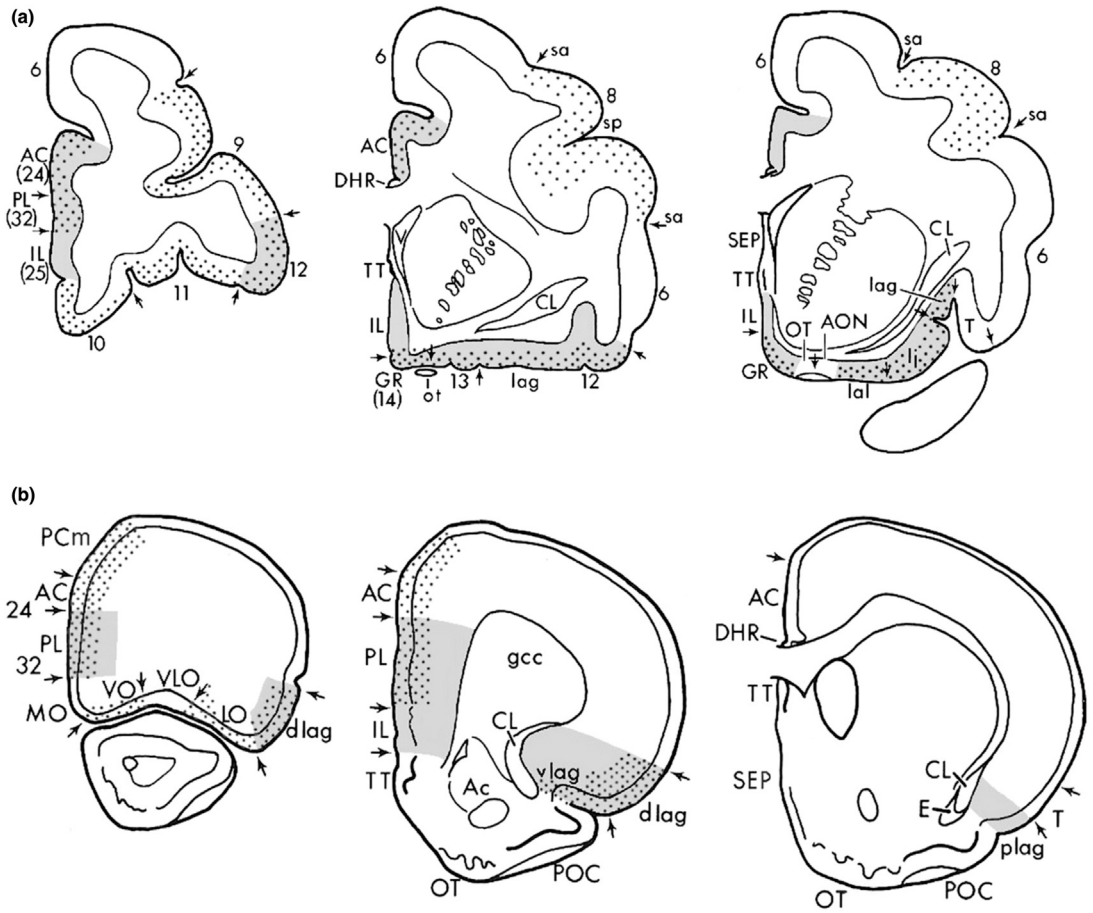

Figure 1 The prefrontal cortex. (a) Serial sections through a rhesus monkey brain showing different cytoarchitectonic regions (Brodmann numbers). Dotted areas receive projections from MD; gray areas receive projections from the amygdala. (b) Serial sections through a rat brain showing architectonic areas and MD, and amygdala projection areas as in the monkey. Both species have similar regions of overlap in MD and amygdala projections as well as areas of no overlap. Abbreviations: AON, accessory olfactory nucleus; AC, anterior cingulate; Ac, anterior commissure; CL, claustrum; gcc, genu of the corpus callousm; IL, infralimbic; lag, lateral agranular; MO, medial orbital; PL, prelimbic; plag, posterior lateral agranular; sa, arcuate sulcus; sp, principal sulcus; SEP, septum; TT, tania tecta; vlag, ventral lateral agranular; VLO, ventral lateral orbital; VO, ventral orbital. Modified and adapted from Reep, R., 1984. Relationship between prefrontal and limbic cortex: a comparative anatomical review. Brain, Behavior and Evolution 25, 1–80.

ganglia circuits to motor cortex, occulomotor cortex, anterior cingulate/medial orbitofrontal, lateral orbitofrontal, and dorsolateral regions in both species.

However, there are also possibly unique prefrontal–striatal connections from prefrontal granular regions. Wise (2008) notes that if the granular prefrontal cortex is unique, it likely projects to parts of the striatum that other species also lack, possibly found in the more dorsal striatal regions.

In sum, the definition of prefrontal cortex remains inconclusive as no single criterion will suffice. The best current definition is that the prefrontal cortex is the region in the cerebral cortex that includes the tissue that has projections from MD and/or amygdala and that has extensive projections to the basal ganglia and rest of the cerebral cortex. Given the extensive period of modification of cerebral regions in mammalian evolution, we should not be surprised that there is considerable variation in the details of prefrontal organization and function and that although many prefrontal regions are ubiquitous across mammals, there are some uniquely primate regions as well.

Why Is There a Prefrontal Cortex?

A key concept in answering this question is to remember that it is behavior, not the cortex that is selected. Hence, the question becomes "What behaviors were selected that led to the development of the prefrontal cortex?" Warren and Kolb (1978) argued that although the details of behavior vary, in general mammals must solve many of the same problems in their daily lives. For example, all mammals detect and interpret sensory stimuli, relate this information to past experience, and generate behavioral strategies. Although the details of what the animals do vary with species and likely have been selected to promote survival in a specific niche (i.e., species-typical behaviors), there are general behavioral capacities demonstrable in all

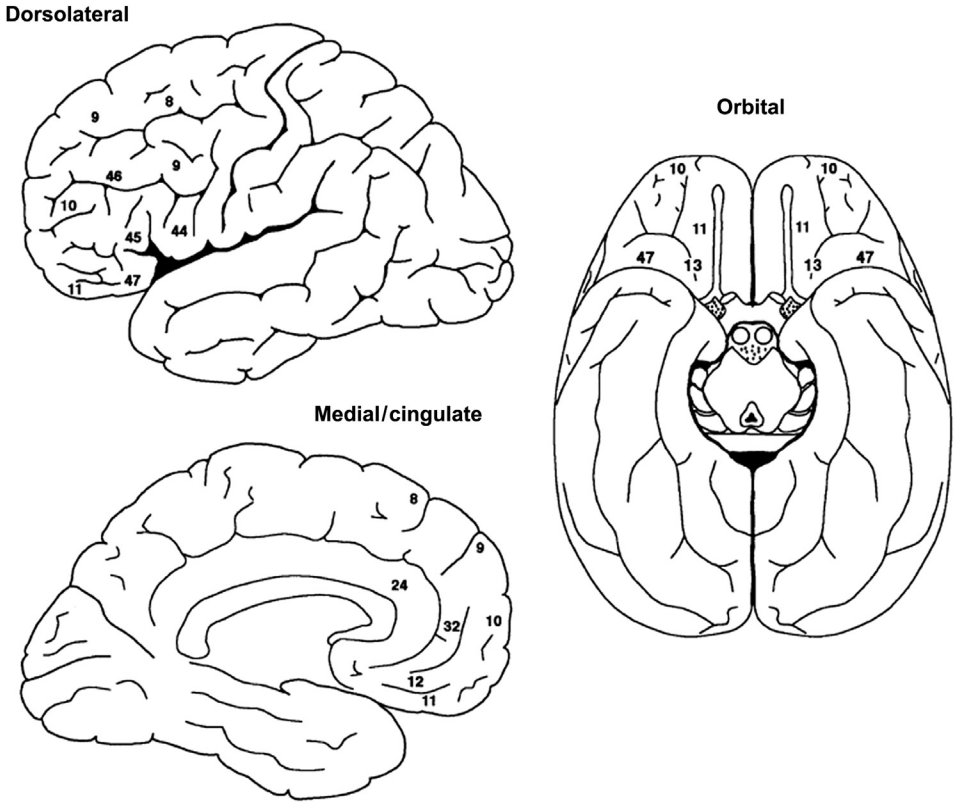

Figure 2 The prefrontal cortex in the human, with cytoarchitectonic areas numbered according to Brodmann's map.

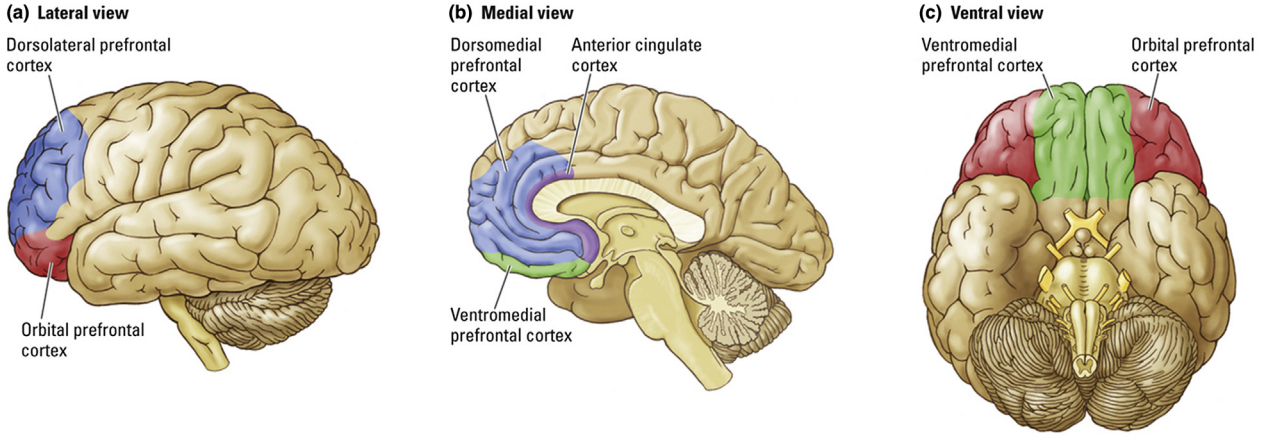

Figure 3 Functional regions of the prefrontal cortex and the adjacent anterior cingulate cortex. Reproduced from Kolb, B., Whishaw, I.Q., 2014. Introduction to Brain and Behavior, fourth ed. Worth, New York, from permission from Worth Publishers.

mammalian species and could be designated as class-common behaviors.

But what is the class-common function(s) of the prefrontal cortex? This is a difficult question, but a review of theories of frontal-lobe function produces a common theme that the prefrontal cortex is involved in the temporal organization of behavior (e.g., Fuster, 1998; Goldman-Rakic, 1987). The general idea is that the prefrontal cortex supports cognitive functions that are necessary to organize behavior in time and in context. The temporal/contextual control of behavior requires at least six components. First, there must be an ongoing record of sensory experience that has occurred recently. This allows behavior to be disconnected from ongoing sensory stimulation and related to other stimulation that may have occurred previously. Second, there must be an ongoing record of what movements are being produced by the brain at any given moment, a record known as reafference. Reafference allows the prefrontal cortex to modify ongoing behavior in response to changes in the environment. Third, there must be selection of incoming experiential information, a process often referred to

as attention. The importance of particular sensory information obviously varies depending upon the environmental demands at any given moment. Fourth, there must be inhibition of some motor impulses and excitation of others to produce appropriate behavioral sequences. Fifth, behavior must be flexible with respect to both internal and external environments. It is one of the characteristics of mammals that we are able to adapt behavioral patterns to changing contexts. The importance of context-dependent behavior is clear in the complex social interactions of mammals. The make-up of the social group at any given time dictates the behavior of each animal. Species as diverse as apes and horses, for example, have a clear pecking order but the details of this social order varies with the precise grouping. Given the presence and position of certain animals a given ape (or horse) may be bold and relaxed whereas the presence of a single additional ape may lead to the same animal being quiet and nervous. It is likely no accident that the prefrontal cortex has grown so large in primates, an order known to be highly social. Sixth, there must be an ongoing monitoring or the consequences of behavior. If reward is associated with a particular set of behaviors, but not with another, then the association between movements and consequences needs to be made.

The Connectome and the Prefrontal Cortex

Continuing research into charting human brain connectivity and the connectome reveals that frontal-lobe regions are central to many cortical networks (van Essen and Ugurbil, 2012; www.humanconnectome.org). The most-studied network, called the brain's *default network*, links a set of far-flung brain regions active in participants who are resting rather than engaged in specific cognitive tasks (Buckner, 2013). Buckner emphasizes, however, that 'default network' is a misnomer, because this network is also active during such directed tasks as thinking about one's past (autobiographical memory), thinking about the future, or when the mind is 'wandering.'

A related salience network is seen in correlated activity among the anterior cingulate cortex, supplementary motor cortex, and anterior insular cortex. Bonnelle et al. (2012) suggest that the salience network is most active when a behavioral change is needed and that it operates to modulate other networks' activities. For example, if the salience network is not functioning properly, the default network shows excessive activity, leading to lapses in attention.

The prefrontal cortex is also a major participant in many cortical networks involved in emotional behaviors (e.g., Roy et al., 2012). The ventromedial prefrontal region plays an especially active role in these networks, and mood disorders result from abnormal activity in these systems (Price and Drevets, 2012).

Intelligence and the Frontal Lobes

The large evolutionary expansion of the prefrontal cortex suggests that increased human intelligence relative to other primates is related to prefrontal size. But until recently, the general consensus held that intelligence does not reside in the frontal lobes (e.g., Hebb, 1949), because large frontal injuries, including frontal lobotomies, have little effect on intelligence as measured by standard intelligence tests. Yet for more than a century psychologists have searched for other measures of intelligence, so perhaps intelligence and the frontal lobe can be related to measurements other than IQ score.

Nearly a century ago, Spearman (1927) proposed that some general factor (the *g* factor) contributes to all cognitive activity. Although his idea has been widely criticized, in fact, a near-universal positive correlation among cognitive tests suggests that, indeed, something like *g* does exist. The best single tests of *g* usually involve problem-solving tasks, which engage fluid intelligence, the ability to see abstract relationships and draw logical inferences. Fluid intelligence can be contrasted with crystalized intelligence, the ability to retain and use knowledge acquired through prior learning and past experience, which is highly related to Weschler IQ score.

Brain imaging studies link tests of fluid intelligence to a characteristic pattern of activity, both in dorsolateral and medial prefrontal cortex and in posterior parietal cortex. Key structures include posterior parietal and prefrontal regions, which the researchers propose, are related to processes of integration and abstraction (posterior parietal) and problem solving, evaluation, and hypothesis testing (prefrontal) (see reviews by Colom et al., 2010; Deary et al., 2010).

Wise (2008) argued that the prefrontal granular cortex is the location of this expanded intellectual capacity and it functions to store knowledge about how to behave under a wide range of nonroutine and complicated circumstances, all of which require attention to our actions and the diverse behavioral contexts encountered over a lifetime. He argues further that this expanded capacity in primates confers only a subtle advantage, but it has enormous consequences. Specifically, he proposes that when we primates recognize a rare risk in a complex pattern of events or an outstanding opportunity knocks, we have advantage because we are so good at knowing what to do and what is likely to happen in the event.

Passingham and Wise (2012) propose that the unique primate capacity makes a qualitative leap in humans. They argue that new prefrontal and posterior parietal areas emerged during the evolution of anthropoid apes, eventually leading to disproportionately large prefrontal and parietal regions in the modern human brain, which laid the groundwork for prefrontal–parietal networks for the expansion of fluid intelligence (general problem solving) in humans (Genovesio et al., 2014). Genovesio et al. argue that as our ancestors' brains enlarged, larger metabolic demands placed a premium on identifying productive foraging locations that might differ in food quality, volume, and danger in getting to the food. According to this hypothesis, foraging strategies that supported reduced risk taking and hominid survival also selected for the superior problem-solving capacities that evolved in humans and thus increased our intelligence. The risk-taking behavior frontal-lobe patients display illustrates what happens when the newly evolved frontal system dysfunctions. Further, the new problem-solving capacities would not be tapped by standard IQ tests, which is consistent with the clinical observations of frontal-lobe patients.

Symptoms of Prefrontal Cortex Injury

There are five general categories of symptoms associated with disease of the prefrontal cortex. We do not mean to imply that the brain respects these categories but rather that the categories provide a conceptual framework within which to consider the symptoms.

Disturbances of Motor Function

In a classic paper, Karl Lashley (1960) asked how movements are put together in a particular order. How is it, he asked, that a violinist can play an arpeggio so quickly and flawlessly? Clearly, each note is not 'thought of' separately. And how is it that, in a tennis game, a player can move much too fast to have considered each movement by itself? The selection and programming of movements is impaired by injury to both the supplementary motor cortex (*see* Premotor Cortex) and prefrontal cortex. For example, patients with prefrontal lesions have deficits in the copying of sequences of hand and limb or facial movements. Although the patients can make the required movements, they have difficulty in placing them in the correct order.

Another form of movement disorder is in the generation of facial and tongue movements in speech. Patients with injury to Broca's area and related prefrontal regions on the left are impaired at the production of speech and, as well, have deficits in using verbs and producing appropriate grammar. Verbs describe movements and grammar requires the correct sequencing of words.

Loss of Divergent Thinking

As noted earlier, frontal-lobe lesions do not affect performance on standard IQ tests, but they do affect performance on tests of problem solving. Problem solving requires an ability that has been referred to as divergent thinking by Guilford (1967). He noted that traditional intelligence tests appear to measure what can be called convergent thinking – that there is just one correct answer to each question. Definitions of words, questions of fact, arithmetic problems, puzzles, and block designs all require single correct answers that are easily scored. In contrast, divergent thinking requires the generation of multiple answers to a single question rather than a single correct answer. An example is asking for a list of the possible uses for a brick.

Frontal-lobe injury interferes with the intelligence required by divergent thinking. One example is reduced behavioral spontaneity. Frontal-lobe patients are unable to generate many answers to questions such as "write down as many words starting with 'S' as you can think of" or "make as many different doodles as you can think of" in 5 min. This deficit is not just seen in paper and pencil tests but in other tests as well, such as in a test requiring patients to "make as many different movements of the hand and arm as you can think of in 5 minutes"

Patients with frontal-lobe lesions are also especially impaired at developing novel cognitive plans or strategies for solving problems. This is especially true when they encounter a novel challenge where well-practiced strategies will not suffice. For example, when patients are asked questions that require reasoning based on general knowledge and for which no immediate strategy is obvious, frontal-lobe patients perform poorly and often gave bizarre responses.

Environmental Control of Behavior

Perhaps frontal-lobe patients' most commonly observed trait is difficulty in using environmental cues (feedback) to regulate or change their behavior. This difficulty manifests itself in myriad ways. For example, patients consistently perseverate on a variety of test situations, particularly those with changing demands. The best example is observed in the Wisconsin Card Sorting Test, a standard clinical test of frontal-lobe injury. In this test, a subject is presented with four stimulus cards bearing designs that differ in color, form, and number of elements. The subject's task is to sort the cards into piles in front of one or another of the stimulus cards. The only help given the subject is to be told whether the choice is correct or incorrect.

The test works on the following principle: the correct solution is, first, color; when the subject has figured out this solution, without warning the correct solution then becomes form. The subject must now inhibit classifying the cards on the basis of color and shift to form. When the subject has succeeded at selecting by form, the correct solution again changes unexpectedly, this time to the number of elements. It will later become color again, and so on.

Shifting response strategies is particularly difficult for people with frontal lesions. They may continue responding to the original stimulus (color) for as many as 100 cards until testing is terminated. Throughout this period, they may comment that they know that color is no longer correct. They nevertheless continue to sort on the basis of color. One person stated (correctly): "Form is probably the correct solution now so this [sorting to color] will be wrong, and this will be wrong, and wrong again."

Temporal Memory

Temporal memory (or working memory) deficits were the first to be demonstrated quantitatively in both laboratory animals and humans, and these deficits have strongly influenced many theories of prefrontal function (e.g., Fuster, 1998; Goldman-Rakic, 1987). The first demonstrations were seen in a test known as delayed response. If the location of reward is signaled by a cue (such as a light), both human and nonhuman laboratory animals cannot remember where the cue was after only a few seconds. Although this was initially interpreted as evidence of a short-term memory deficit, it can equally be attributed to a disorder of attention (see Wise, 2008).

Emotion and Social Behavior

Changes in effect and social behavior are obvious and striking effect of prefrontal cortex damage. The most publicized example of personality change subsequent to frontal-lobe lesions is that of Phineas Gage, first reported by John Harlow in 1868. Gage was a dynamite worker who survived an explosion that blasted an iron tamping bar through the front of his head. The bar, was about 1 m long and 3 cm wide at its widest point.

After the accident, Gage's behavior changed completely. He had been of average intelligence and was "energetic and persistent in executing all of his plans of operation" according to Harlow, but afterward he was "fitful, irreverent, indulging at times in the grossest profanity, manifesting but little deference to his fellows, impatient of restraint or advice when it conflicts with his desires, at times pertinaciously obstinate, yet capricious and vacillating, devising many plans of operation, which are no sooner arranged than they are abandoned in turn for others appearing more feasible."

Disorders Affecting the Frontal Lobe

Many people with psychiatric or neurological disorders present symptoms characteristic of frontal-lobe injury. Disturbance of frontal function likely contributes significantly to the behavioral symptoms of each disease. In addition, chronic stress alters prefrontal neurons in ways that affect some memory and behavioral functions.

In schizophrenia, there is an abnormality in the mesolimbic dopamine projection that terminates largely in the prefrontal cortex, a decrease in blood flow to the frontal lobes, and possible frontal-lobe atrophy. Schizophrenia patients perform poorly on all tests of frontal-lobe function, but they perform normally on tests of parietal-lobe function. Similarly, patients with Korsakoff's syndrome, a metabolic disorder of the central nervous system often associated with chronic alcoholism that results in degeneration of MD, also perform poorly on tests of prefrontal function.

Addictive drugs change the structure of neurons in both the orbitofrontal and the medial prefrontal regions in rats (for a review, see Robinson and Kolb, 2004). Drug addiction is likely related to abnormalities in prefrontal structure and function that are associated with the maladaptive decision-making characteristic of addictive behavior.

Prolonged stress is correlated with functionally significant changes in the structure of neurons in the prefrontal cortex that affect temporal memory and goal-directed behaviors (see review by McEwen and Morrison, 2013). Laboratory animal studies have shown that chronic stress in male rats reduces synaptic space in neurons in the medal prefrontal cortex but increases synaptic space in the orbitofrontal cortex. In contrast, similar stress in female rats produces the opposite effect in medial PFC regions. In young adult rats these changes are at least partially reversible with periods of reduced stress, but aged animals show no such recovery. We can speculate that age-related decline in cognitive abilities could partly reflect the chronic effects of stress throughout a lifetime.

See also: Attention, Neural Basis of; Cortical Areas Engaged in Movement: Neuroimaging Methods; Dysexecutive Syndromes; Parietal Lobe; Premotor Cortex; Working Memory, Neural Basis of.

Bibliography

Alexander, G.E., Crutcher, M.D., DeLong, M.R., 1990. Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, 'prefrontal' and 'limbic' functions. In: Uylings, H.B.M., Van Eden, C.G., De Bruin, J.P.C., Feenstra, M.P.G. (Eds.), The Prefrontal Cortex: Its Structure, Function and Pathology, Progress in Brain Research, vol. 85. Elsevier, Amsterdam, pp. 119–146.

Bonnelle, V., Ham, T.E., Leech, R., Kinnunen, K.M., Mehta, M.A., Greenwood, R.J., Sharp, D.J., 2012. Salience network integrity predicts default mode network function after traumatic brain injury. Proceedings of the National Academy of Sciences of the United States of America 109, 4690–4695.

Buckner, R., 2013. The brain's default network: origins and implications for the study of psychosis. Dialogues in Clinical Neuroscience 15, 351–358.

Colom, R., Karama, S., Jung, R.E., Haier, R.J., 2010. Human intelligence and brain networks. Dialogues in Clinical Neuroscience 12, 489–501.

Deary, I.J., Penke, L., Johnson, W., 2010. The neuroscience of human intelligence differences. Nature Review Neuroscience 11, 201–211.

van Essen, D.C., Ugurbil, K., 2012. The future of the human connectome. NeuroImage 62, 1299–1310.

Fuster, J., 1998. The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe, third ed. Raven Press, New York.

Genovesio, A., Wise, S.P., Passingham, R.E., 2014. Prefrontal-parietal function: from foraging to foresight. Trends in Cognitive Sciences 18, 72–81.

Goldman-Rakic, P.S., 1987. Circuitry of the primate prefrontal cortex and regulation of behavior by representational memory. In: Plum, F. (Ed.), Handbook of Physiology: The Nervous System, Vol. 5, Part 1 Higher Functions of the Brain. American Physiological Society, Bethesda, MD.

Guilford, J.P., 1967. The Nature of Human Intelligence. McGraw-Hill, New York.

Hebb, D.O., 1949. Organization of Behavior. McGraw-Hill, New York.

Kolb, B., Whishaw, I.Q., 2014. Introduction to Brain and Behavior, fourth ed. Worth, New York.

Lashley, K.S., 1960. The problem of serial order in behavior. In: Beach, F.A., Hebb, D.O., Morgan, C.T., Nissen, H.W. (Eds.), The Neuropsychology of Lashley. McGraw-Hill, New York, pp. 529–543.

McEwen, B.S., Morrison, J.H., 2013. The brain on stress: vulnerability and plasticity of the prefrontal cortex over the life course. Neuron 79, 16–29.

Middleton, F.A., Strick, P.L., 2000. Basal-ganglia 'projections' to the prefrontal cortex of the primate. Cerebral Cortex 12, 926–935.

Pandya, D.N., Yeterian, E.H., 1985. Architecture and connections of cortical association areas. In: Peters, A., Jones, E.G. (Eds.), Cerebral Cortex, vol. 4. Plenum, New York, pp. 3–61.

Passingham, R.E., Wise, S.P., 2012. The Neurobiology of the Prefrontal Cortex: Anatomy, Evolution, and the Origin of Insight. Oxford University Press, New York.

Price, J.L., Drevets, W.C., 2012. Neural circuits underling the pathophysiology of mood disorders. Trends in Cognitive Sciences 16, 61–71.

Reep, R., 1984. Relationship between prefrontal and limbic cortex: a comparative anatomical review. Brain, Behavior and Evolution 25, 1–80.

Robinson, T.E., Kolb, B., 2004. Structural plasticity associated with drugs of abuse. Neuropharmacology 47 (Suppl. 1), 33–46.

Rose, J.E., Woolsey, C.N., 1948. The orbitofrontal cortex and its connections with the mediodorsal nucleus in rabbit, sheep and cat. Association for Research in Nervous and Mental Disorders 27, 210–232.

Roy, M., Shohamy, D., Wager, T.D., 2012. Ventromedial prefrontal-subcortical systems and the generation of affective meaning. Trends in Cognitive Sciences 16, 147–156.

Spearman, C., 1927. The Abilities of Man. Macmillan, New York.

Uylings, H., Groenewegen, H., Kolb, B., 2003. Does the rat have a prefrontal cortex? Behavioural Brain Research 146, 3–17.

Warren, J.M., Kolb, B., 1978. Generalizations in neuropsychology. In: Finger, S. (Ed.), Brain Damage, Behavior and the Concept of Recovery of Function. Plenum Press, New York.

Wise, S.P., 2008. Frontal forward fields: phylogeny and fundamental function. Trends in Neurosciences 31, 599–608.

Prefrontal Cortex Development and Development of Cognitive Function

Robbin Gibb and Bryan Kolb, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Diamond, volume 17, pp. 11976–11982, © 2001, Elsevier Ltd.

Abstract

The prefrontal cortex (PFC) is the anterior-most portion of the cerebrum and functions to plan and direct behavior in varying contexts over time. It supports important socioemotional and cognitive skills, also known as executive function, that are associated with both academic achievement and life success. The PFC is comprised of many subareas that work together and with other cortical and subcortical regions of the brain to produce the complex cognitive and emotional capabilities associated with human behavior. The PFC is the last brain region to mature and development in this area proceeds into the fourth decade of life. The slow timeline for development of the PFC makes it more sensitive to environmental influences. This allows the acquisition of convention learning through experience but leaves the PFC vulnerable to external insults that may result in suboptimal development and neuropsychiatric disease.

Introduction

The prefrontal cortex (PFC) is the anterior area of the frontal cortex and it accounts for approximately 25–30% of the cortical mantle. It is the association cortex of the frontal lobe and is located anterior to the premotor and motor cortices. The major anatomical features of the PFC are large pyramidal neurons in layer III and granularity of layer IV although the granularity is less pronounced in rodents than in humans, nonhuman primates, and carnivores (Kostovic et al., 2008). Although a universally accepted definition of PFC does not exist, it has historically been thought that input from the medial dorsal nucleus (MD) of the thalamus to rostral regions of the frontal cortex is a unifying feature. The PFC receives input from all other cortical areas and its main function is to plan and direct behavior of all types (i.e., motor, cognitive, emotional, and social) across time. In fact, the PFC supports what is termed 'executive function (EF)' and thus is considered the executive of the brain. EF refers to important socioemotional and cognitive skills that are known to be highly correlated with both academic and life success. EF is blanket term that is considered to include creativity, self-regulation, organization and planning, problem solving, and working memory. The PFC has a protracted developmental timeline that starts in the prenatal period and finishes in the fourth decade of life when myelination is complete (Kolb et al., 2012). It is the last area of the brain to mature and as a result is sensitive to a variety of experiences that shape the underpinnings of complex cognitive and emotional capabilities. The PFC's prolonged period of maturation also makes it more sensitive to adverse experiences that can result in abnormal PFC development (often manifested as neuropsychiatric disorders.)

Subregions of the PFC

The PFC is considered multimodal association cortex that receives and integrates highly processed information from sensory association areas (granular cortex) as well as information from limbic areas (agranular and dysgranular cortex). The PFC is not a homogeneous area and as such is generally divided into subregions using location, cytoarchitecture, and connectivity as defining features (*see* Prefrontal Cortex for a figure showing prefrontal cortical regions). The PFC can be divided into three major subregions:

1. *Lateral prefrontal cortex* (*LPFC*), Brodmann's areas 9 and 46, mainly make reciprocal connections with the posterior parietal areas and the superior temporal sulcus. The dorsolateral cortex has extensive connections with regions to which the posterior parietal cortex also projects, including the cingulate cortex, basal ganglia, and superior colliculus.
2. *Orbitofrontal cortex* (*OFC*), areas 47 and lateral portions of 11, 12, and 13, gain input from all sensory modalities. The OFC's main cortical afferents project from the temporal lobe, including the auditory regions of the superior temporal gyrus, the visual regions of the inferotemporal cortex (area TE), S2 (somatosensory area 43), gustatory cortex in the insula, and olfactory regions of the pyriform cortex.
3. *Medial prefrontal cortex* (*MPFC*), areas 10, 14, and 25, as well as medial parts of areas 11, 12, and 13, and the anterior part of 32, receives cortical connections from the dorsolateral PFC (DLPFC), posterior cingulate cortex, and medial temporal cortex.

In addition to these three prefrontal regions, another region called the anterior cingulate cortex, is closely associated with the PFC. This region includes Brodmann's area 24 and part of 32 and makes extensive bidirectional connections with motor, premotor, prefrontal, and insular cortex.

Function and Connectivity of PFC Subregions

The subregions of the PFC are highly interconnected, which suggests that prefrontal functions are subserved by multiple areas of the PFC. The PFC has reciprocal connections with sensory cortical areas that may be a result of its use of integrated sensory information to guide top-down attentional demands. The PFC is also highly connected to the parietal and temporal lobes of the brain, which support spatial and memory

processes. Finally, the PFC is highly interconnected with subcortical structures including the hippocampus, basal ganglia, thalamus, hypothalamus, and the amygdala. Connections of some subregions of the PFC to limbic areas may reflect the use of affective information to guide behavior based on social information and environmental contingencies.

Medial Prefrontal Cortex

The MPFC is involved in the processing and integration of affective and social information (Grossmann, 2013). It is also involved in memory and decision-making. Based on the myriad of functions that seem to be supported by the MPFC, Euston et al. (2012) propose that the MPFC with support from the hippocampus is responsible for recalling the best course of action or emotional response to specific events in specific contexts.

Ventral Areas

The ventromedial portion of the MPFC functions to regulate emotional responses in a manner consistent with changing environmental contexts. It has diverse neural connections that include sensory input from all modalities and output to areas that contribute to emotional regulation and expression (e.g., amygdala, hypothalamus, ventral tegmental area, and medial temporal cortex). It integrates social and affective information and thus contributes to a wide range of social cognitive abilities. The ventromedial PFC has recently been shown to be critical in the regulation of amygdala activity (Motzkin et al., 2014). The rostral portion of the PFC, while the largest subarea of the PFC is perhaps the least well understood. It is connected to areas in the frontal, parietal, and temporal lobes (Gonen-Yaacovi et al., 2010) and has been shown to be activated by pleasant touch through its connections with reward-related regions. The rostral PFC has low cell density that may reflect more space available for intrinsic and extrinsic connections and it has an extraordinarily high number of dendritic spines per cell suggestive of an important role in integration or coordination of incoming information. It has been proposed that the rostral PFC supports prospective memory or the ability to carry out an intentional act on the basis of previously acquired information and mentalizing (Theory of Mind) or development of the understanding of mental states in both self and others (Dumontheil et al., 2008). The rostral PFC has also been implicated in higher order decision-making associated with exploitation of available resources (Wunderlich et al., 2009).

Dorsal Areas

The dorsal MPFC is recognized to play a role in decision-making, conflict monitoring, and reward-guided learning (Botvinick, 2007; Rushworth and Behrens, 2008). It receives input from the insula, limbic regions, and sensory association cortices (Cavada and Reinoso-Suarez, 1985), and higher order motor areas in the frontal lobe (Matsuzaka et al., 2012). Recently, the dorsomedial PFC has been additionally implicated in understanding mental states in others and empathy (Waytz et al., 2012; Kang et al., 2013) whereas the rostral PFC shows greater activation when understanding the mental state of self (Kang et al., 2013; D'Argembeau, 2013).

Lateral Prefrontal Cortex

The LPFC is primarily involved in the processes of cognitive control and memory and is strongly connected to the temporal cortex (Takahashi et al., 2006).

Ventral Areas

The ventrolateral regions of the PFC are involved in active retrieval and selection of information (first-order executive processes (Tanji and Hoshi, 2008)). In addition, the right ventrolateral PFC has been identified as active in tasks of self-control under a variety of conditions including response inhibition, thought suppression, and delayed gratification (Cohen and Lieberman, 2010). It has connections with the amygdala, superior temporal polysensory area, and auditory areas.

Dorsal Areas

The DLPFC is involved in motor planning and organization and as such has connections with the basal ganglia and ventral anterior nucleus of the thalamus (Rektorova et al., 2008; Morris et al., 2014). It is also commonly associated with working memory processes (Cieslik et al., 2013). It has connectivity with the dorsal (how) pathway and the ventral (what) pathway of the visual system as well as the posterior parietal cortex (Olson and Luciana, 2008; O'Reilly, 2010). It also has connectivity with the OFC and plays a contributory role to intelligence and higher order thought processes (Barbey et al., 2012).

Ontogenetic Development of the PFC

Prenatal Development

Although neurons destined for the PFC are produced in the ventricular zone from neuroepithelial cells at 7 weeks postconception, the cytoarchitectonic hallmarks of the PFC are not visible until 36 weeks postconception. Afferents from monaminergic (dopamine and serotonin) areas of the brain stem arrive in the frontal areas of the cortex at approximately 9 weeks postconception whereas acetylcholine innervation of the frontal areas does not appear until about 13 weeks postconception. Fibers arising from the MD of the thalamus invade the cortical plate between 24 and 28 weeks postconception and callosal connections arise approximately 32 weeks postconception and are refined through elimination into the postnatal period of life (Kostovic et al., 2008).

Postnatal Development

In the neonatal period there is rapid formation of dendritic spines in the PFC with the number peaking at about 2.5–4 years, depending upon the region and cortical layer (Petanjek et al., 2011). Spine development occurs in association with the acquisition and refinement of language and related cognitive processes. The process of synaptogenesis occurs largely as a result of intense interactions with environment and implicates experience as an important factor in the stabilization and pruning of PFC synapses (Nelson et al., 2000). A recent study on synaptic spine formation in the PFC over the lifespan has confirmed that spine density in childhood

is two- to threefold higher than the spine density observed in adults and shows a downward decline during puberty, which continues into the third decade of life (Petanjek et al., 2011).

Giedd et al. (1999) showed that in frontal areas, gray matter volume peaks between 11 and 13 years followed by a prolonged period where gray matter volume is reduced. Although decreases in gray matter volume have commonly been attributed to synaptic pruning, Dumontheil et al. (2008) caution that such changes may additionally arise from increases in myelination in the PFC areas during this period.

When brain growth is assessed using a 'distance from center' measure to examine the radial expansion of exterior cortical surface, the rostral PFC shows one of the highest rates of brain growth between the ages of 5 and 11 years and continues to show growth throughout the period of adolescence (Sowell et al., 2004). It has been proposed that the prolonged development of this PFC area along with its unique anatomical characteristics support the notion that the function of the rostral PFC is to support the cognitive behavior that uniquely distinguishes humans from other species (Dumontheil et al., 2008).

Myelination, which speeds the rate of information flow between neurons, begins largely in the postnatal period for the PFC and continues into the fourth decade of life. While it is tempting to suppose that the maturational timeline for myelination of the PFC is genetically controlled there is clear evidence that environmental influences play an integral role in this process. For example, studies of pianists show that intensive training that occurs over many years is associated with increased myelination (Bengtsson et al., 2005).

Not surprisingly the subregions of the PFC have different rates of maturation. The most protracted development of the PFC appears in the rostral PFC and the DLPFC (Olson and Luciana, 2008).

Sex differences also exist in PFC development on the basis of maturation and are area dependent. For example, the ventromedial PFC appears to develop the earliest among the areas of the PFC and this maturation occurs earlier in males than in females. Conversely, the myelination process that requires decades to complete in the rostral PFC is finished earlier in females than in males.

Functional Development of PFC

When considering the development of the brain, the concept of 'sensitive periods' or periods of time when experiences have stronger influences functional and anatomical outcomes comes in to play. In sensory brain areas the sensitive periods largely occur in the first year of life but in the PFC the sensitive period appears to extend into late childhood or early adolescence. It has been proposed (Thompson-Schill et al., 2009) that delayed maturation of PFC, at a cost of functional performance in infancy and early childhood, is advantageous for human learning. The theory outlined by Thompson-Schill and colleagues makes a case for 'cognition without control.' The basic tenet of this theory is an immature brain is better positioned to master cultural and linguistic conventions, or convention learning. As an example, the authors highlight the mental processes associated with 'creativity.' They note that when faced with a challenge to create new uses for a familiar object, adults often struggle, in part because they have a wealth of experiences with the object that shape their expectations about its use. Conversely, young children are not hindered by this 'functional fixedness' and are able to create a myriad of uses for such an object. Children are also better at language acquisition than adults and this talent arises as a benefit of an immature PFC. A mature PFC provides adults with the ability to think flexibly and to capably use contextual information to optimize behavioral output or performance. Thompson-Schill and colleagues argue that the PFC optimizes performance of an individual and if it were to mature earlier, broader aspects of learning subserved by subcortical areas would be preempted.

PFC Function in Infants

Despite the prolonged period of maturation of the PFC in humans, significant changes occur in the first year of life and function to support cognitive advances seen in infants at this time. Infants are born with the behavioral biases to attend to certain stimuli in their environment, i.e., faces and voices. These biases may provide the foundation on which social cognition is built. Using near infrared spectroscopy or NIRS researchers discovered that newborns show greater activation in the MPFC to their mother reading a story using infant-directed speech versus adult-directed speech (Saito, 2007). Infants at 4–13 months of age when tested with a strange female voice and their mother's voice only showed activation of the MPFC when their mother's voice was heard (Naoi et al., 2012). This work, along with behavioral work shows that by 7–9 months of age infants show strong preferences for their primary caregiver and this is supported through changes in PFC function. Further studies using NIRS showed a simple smile that was preceded by eye contact activates the PFC in 4-month-old babies (Grossmann et al., 2008) and 9- to 12-month-old infants showed significantly greater activation of the MPFC to their smiling mother when compared to a stranger (Minagawa-Kawai et al., 2009). Interestingly, mothers of the babies exhibited a similar PFC response when looking at their own infants smiling. Grossmann et al. (2013) used NIRS to demonstrate that 5-month-old infants when gazing at an object in their environment were able to detect when their adult social partners followed their gaze to that object. The detection was demonstrated by activation in the left PFC. Using NIRS, Kida and Shinohara (2013) demonstrated that 10-month-old babies show activation of the rostral PFC, similar to adults, when touching velvet but this response is absent in 3- and 6-month-old babies. This suggests that maturation of the rostral PFC is required to process activation in the tactile affective system and the critical point of maturation is between the sixth and tenth postnatal month. Neuropsychological tests on infants and children have demonstrated that working memory or the ability to hold important information in mind begins to develop at about 8 months of age (Diamond, 1985). This suggests that the DLPFC, the area supporting this ability, is beginning to show maturational changes as well.

PFC Function in Children

Children as young as 1 year of age begin to show signs of attentional control, an EF associated with control of

emotional behavior by the ventromedial PFC. In contrast, the development of metacognitive forms of EF (planning and verbal fluency) supported by the DLPFC does not begin until about 3 years of age (Ardila, 2013). Perlman et al. (2014) used NIRS in 3- to 5-year-old children to identify the PFC locations activated during 'winning' and 'frustration.' Their results show that winning activated the MPFC whereas the LPFC was activated during frustration. A recent study of the maturation of planning abilities in girls and boys was conducted on 4- and 5-year-olds with a re-test interval of 1 year (Unterrainer et al., 2013). The Tower of London task was employed to assess how planning strategies matured in these children over the course of 1 year. It was shown that girls mature in advance of boys in their development of planning at this age and this appeared to be related to differences in maturational rates of the DLPFC in girls and boys. Takeuchi et al. (2013) examined the effects of TV viewing on brain structure in children using structural magnetic resonance imaging (MRI). They gathered preexperimental data from children between the ages of 5 and 18 years and then re-studied them 2 years later. They demonstrated that duration of TV viewing was positively correlated with both gray matter and white matter volume in the rostral PFC and MPFC areas. They also confirmed that duration of TV viewing was negatively correlated with IQ in these children. This study demonstrates the powerful impact that environment wields over maturational processes in the PFC.

PFC Function in Adolescence

Adolescence is defined as the transitional period of life that spans late childhood to adulthood. Early adolescence is characterized by puberty, which is associated with surging gonadal hormones. Interestingly, these gonadal hormones confer a high degree of neuroplasticity in certain brain areas and one area affected is the PFC. During adolescence the PFC undergoes extensive reorganization and it is particularly sensitive to environmental influences. The involvement of epigenetics in the mechanisms underlying the dynamic changes in the PFC at this age has been largely understudied. Epigenetic mechanisms are highly responsive to the environment and as such may also contribute to the vulnerability of the PFC to negative experience during adolescence (Morrison et al., 2014).

The emergence of self-consciousness was examined in child, adolescent, and adult populations using fMRI (Somerville et al., 2013). Participants viewed cues that lead them to believe that a peer was actively watching them and this caused increases in autonomic arousal that was particularly prevalent in the adolescent group. This behavioral pattern was accompanied by emergent striatal and MPFC connectivity during adolescence that has been thought to promote social evaluation of contexts.

In a study aimed at evaluating the development of affective theory of mind, children of age 12–14 years and adults (19–25 years) were recruited to undergo fMRI while evaluating video clips of actors displaying various affective mental states. The ventromedial PFC showed stronger activation in adolescents than in adults while viewing the videos, a finding that suggests this area may be involved in the development of affective theory of mind processing in adolescents.

Recently, a longitudinal study (participants aged 9–23 years) was conducted to examine the brain regions that support inhibitory control (Ordaz et al., 2013). fMRI data from over 100 participants were gathered and specific brain regions within circuits that support motor response control, executive control, and error processing were studied. The data revealed that mean growth curves of activation in motor control regions did not change with age, executive control regions showed decreased activation with age until adolescence, and processing activation of the dorsomedial PFC and anterior cingulate increased into adulthood. The results suggest that these processes possess distinct maturational timelines that contribute to the unique behaviors exhibited by teens.

Risk-taking is a behavioral trait that is commonly exhibited by adolescents. It has been proposed that poor executive control and delayed PFC maturation contribute to risk-taking behavior. A review of MRI, behavioral studies, and epidemiological studies suggest that incentive motivation arising from limbic areas shows a substantial increase from childhood to adolescence and then declines. Immature PFC function at the same time may make this motivation hard to subdue and thus risk-taking increases (Luciana, 2013).

Environmental Influences on PFC Development

Although brain development proceeds along a basic template that is genetically determined, it is clear that experiential factors contribute to and alter developmental trajectories and behavioral outcomes. Two characteristics of PFC organization are important for advancing our understanding of how environment has such a dynamic impact on this cortical area. First, dendritic spine density, a proxy for excitatory synapses, is two- to threefold higher in the PFC of children than in adults. Second, the overproduction of dendritic spines in development is highest and the rate of synapse elimination is lowest in the PFC. Using rodent models to study behavioral changes mediated by anatomical changes in the PFC, a wide variety of environmental factors (experiences) have been identified that change the development of the circuitry in the PFC. These include sensorimotor experience, exposure to stress, parent–infant relationships, peer–peer relationships, and exposure to psychoactive drugs (Kolb et al., 2012). Interestingly, the timing of the experience (age at exposure), the sex of the subject, and previous environmental exposures all contribute to the impact of the experience on the behavioral and anatomical outcomes. It is clear that maternal experience during pregnancy can influence behavior and brain development in the offspring and it has recently been shown that the same is true for father experience in the preconception period (Mychasiuk et al., 2012, 2013).

Neurodevelopmental Disorders and the PFC

Attention Deficit Hyperactivity Disorder

In a recent experiment, Shaw et al. (2007) used MRI to examine cortical thickness in children ranging in age from

preschoolers to young adults. Participants were scanned at least twice in 3-year intervals. The researchers demonstrated that children with attention deficit hyperactivity disorder (ADHD) show normal maturation of the brain but the maturational timeline lagged behind children without the disorder by about 3 years. Development of the MPFC was delayed by up to 5 years in youth with ADHD compared to undiagnosed controls. The motor cortex was the only area examined that showed accelerated development in the ADHD-affected children. This mismatch in developmental timelines of the PFC and motor cortex may account for inattention, restlessness, and fidgeting in affected children.

Autism

In 2002, Carper and colleagues conducted an MRI study on 2- to 11-year-old children diagnosed with autism. Their data revealed an increase in cortical thickness in the frontal lobe of autistic children compared to controls with the biggest differences observed in 2- and 3-year olds. In a postmortem study of PFC tissue derived from autistic individuals and controls, Courchesne et al. (2011) discovered that the tissue derived from autistic individuals (aged 2–16 years) had 79% more cells in the DLPFC and 29% more cells in the MPFC compared to normally developing controls. These results suggest that the autism phenotype may result from exuberant growth in the PFC and a failure to prune unnecessary cells and connections.

Schizophrenia

Over the past two decades, schizophrenia has increasingly been regarded as a neurodevelopmental disorder. It was previously believed that negative events in the prenatal period of life derailed brain function two decades later in patients affected with schizophrenia. Now that PFC has been shown to undergo an extended period of development, the notion that negative events in postnatal life, childhood, and early adolescence may contribute to the disease is gaining more acceptance. In late adolescence or early adulthood when a diagnosis of schizophrenia is made, fMRI studies show abnormal PFC activation during performance of working memory tasks. Specifically, patients show greater activation of PFC in association with reduced performance on working memory tasks (Callicott et al., 2000). What is not clear is how these changes arise. In 2013, Catts and colleagues published work that suggested the best way to understand the neuropathology associated with schizophrenia is to better understand the processes underlying normal brain development.

Metaplasticity and the PFC

The experiences an individual has during the period of PFC development have the potential to alter neuroplasticity in this area. This altered circuitry then responds differently to experiences that occur later in life, a phenomenon called metaplasticity. Consider the following example. Using a rodent model, the effect of cognitive training on a two-way active avoidance (TWA) paradigm during infancy was examined on later cognitive performance after training in adulthood. This was compared to pretraining on the TWA followed by training in adulthood (Bock et al., 2014). The authors were surprised to discover that the infant training improved avoidance learning in adulthood even though the infant rats were unable to learn a successful avoidance strategy. In addition, training in infancy was more effective than pretraining on the same task in adulthood. These results support the notion that infant cognitive training alters developmental organization and primes PFC neuronal circuits for adult learning.

Training EF to Support Academic Achievement and Life Success

One of the most intriguing aspects of PFC development in childhood is the notion that training can enhance EF and improvements in EF will lead to better success in school and in life. A number of tests of EF have been developed to assess this ability in children as young as two in order to determine if remediation is warranted so it can be delivered in a timely fashion such that a child starting kindergarten will have the necessary EF skills in place to ensure better academic success. This notion is appealing to researchers, educators, and parents alike and could make an important difference in life outcomes of children. A word of caution comes to mind based on the theory proposed by Thompson-Schill and colleagues (discussed above).

If delayed maturation of the PFC is truly a benefit for learning in the developing brain, will earlier maturation truncate learning possibilities and force a premature emphasis on optimized performance of brain function without the necessary underpinnings of convention learning?

See also: Brain Development: Ontogenetic Neurobiology of; Cerebral Cortex; Cognitive Development in Infancy, Neural Mechanisms of; Prefrontal Cortex.

Bibliography

Ardila, A., 2013. Development of metacognitive and emotional executive functions in children. Applied Neuropsychology Child 2 (2), 82–87. http://dx.doi.org/10.1080/21622965.2013.748388.

Barbey, A.K., Colom, R., Solomon, J., Krueger, F., Forbes, C., Grafman, J., 2012. An integrative architecture for general intelligence and executive function revealed by lesion mapping. Brain 135 (Pt 4), 1154–1164. http://dx.doi.org/10.1093/brain/aws021aws021 [pii].

Bengtsson, S.L., Nagy, Z., Skare, S., Forsman, L., Forssberg, H., Ullen, F., 2005. Extensive piano practicing has regionally specific effects on white matter development. Nature Neuroscience 8 (9), 1148–1150. http://dx.doi.org/10.1038/nn1516 pii:nn1516.

Bock, J., Poeggel, G., Gruss, M., Wingenfeld, K., Braun, K., 2014. Infant cognitive training preshapes learning-relevant prefrontal circuits for adult learning: learning-induced tagging of dendritic spines. Cerebral Cortex 24 (11), 2920–2930. http://dx.doi.org/10.1093/cercor/bht148. Epub 14 Jun 2013.

Botvinick, M.M., 2007. Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. Cognitive, Affective, & Behavioral Neuroscience 7 (4), 356–366.

Callicott, J.H., Bertolino, A., Mattay, V.S., Langheim, F.J., Duyn, J., Coppola, R., Weinberger, D.R., 2000. Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited. Cerebral Cortex 10 (11), 1078–1092.

Carper, R.A., Moses, P., Tigue, Z.D., Courchesne, E., 2002. Cerebral lobes in autism: early hyperplasia and abnormal age effects. Neuroimage 16 (4), 1038–1051. doi:S1053811902910992 [pii].

Catts, V.S., Fung, S.J., Long, L.E., Joshi, D., Vercammen, A., Allen, K.M., Shannon Weickert, C., 2013. Rethinking schizophrenia in the context of normal neurodevelopment. Frontiers in Cellular Neuroscience 7, 60. http://dx.doi.org/10.3389/fncel.2013.00060.

Cavada, C., Reinoso-Suarez, F., 1985. Topographical organization of the cortical afferent connections of the prefrontal cortex in the cat. Journal of Comparative Neurology 242 (3), 293–324. http://dx.doi.org/10.1002/cne.902420302.

Cieslik, E.C., Zilles, K., Caspers, S., Roski, C., Kellermann, T.S., Jakobs, O., et al., 2013. Is there "one" DLPFC in cognitive action control? Evidence for heterogeneity from co-activation-based parcellation. Cerebral Cortex 23 (11), 2677–2689. http://dx.doi.org/10.1093/cercor/bhs256bhs256 [pii].

Cohen, J., Lieberman, M., 2010. The common neural basis of exerting self-control in multiple domains. In: Trope, Y., Hassin, R., Ochsner, K. (Eds.), Self-Control in Society, Mind, and Brain. Oxford University Press, Oxford, pp. 141–160.

Courchesne, E., Mouton, P.R., Calhoun, M.E., Semendeferi, K., Ahrens-Barbeau, C., Hallet, M.J., Pierce, K., 2011. Neuron number and size in prefrontal cortex of children with autism. JAMA 306 (18), 2001–2010. doi:10.1001/jama.2011.1638306/18/2001 [pii].

D'Argembeau, A., 2013. On the role of the ventromedial prefrontal cortex in self-processing: the valuation hypothesis. Frontiers in Human Neuroscience 7, 372. http://dx.doi.org/10.3389/fnhum.2013.00372.

Diamond, A., 1985. Development of the ability to use recall to guide action, as indicated by infants' performance on AB. Child Development 56 (4), 868–883, 026.x.

Dumontheil, I., Burgess, P.W., Blakemore, S.J., 2008. Development of rostral prefrontal cortex and cognitive and behavioural disorders. Developmental Medicine & Child Neurology 50 (3), 168–181. doi:10.1111/j.1469-8749.2008.02026.xDMCN02026 [pii].

Euston, D.R., Gruber, A.J., McNaughton, B.L., 2012. The role of medial prefrontal cortex in memory and decision making. Neuron 76 (6), 1057–1070. doi:10.1016/j.neuron.2012.12.002S0896-6273(12)01108-7 [pii].

Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., Zijdenbos, A., Rapoport, J.L., 1999. Brain development during childhood and adolescence: a longitudinal MRI study. Nature Neuroscience 2 (10), 861–863. http://dx.doi.org/10.1038/13158.

Gonen-Yaacovi, G., Gilbert, S., Burgess, P., 2010. Functional connectivity between distinct subregions within rostral prefrontal cortex (BA 10) and other brain areas: a cluster analysis the 20th Annual Rotman Research Institute conference. The Frontal Lobes. http://dx.doi.org/10.3389/conf.fnins.2010.14.00120.

Grossmann, T., 2013. The role of medial prefrontal cortex in early social cognition. Frontiers in Human Neuroscience 7, 340. http://dx.doi.org/10.3389/fnhum.2013.00340.

Grossmann, T., Johnson, M.H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., Csibra, G., 2008. Early cortical specialization for face-to-face communication in human infants. Proceedings of the Royal Society B: Biological Sciences 275 (1653), 2803–2811. doi:10.1098/rspb.2008.0986Q6RW78322710W7N8 [pii].

Grossmann, T., Lloyd-Fox, S., Johnson, M.H., 2013. Brain responses reveal young infants' sensitivity to when a social partner follows their gaze. Developmental Cognitive Neuroscience 6, 155–161. doi:10.1016/j.dcn.2013.09.004S1878-9293(13)00056-X [pii].

Kang, P., Lee, J., Sul, S., Kim, H., 2013. Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. Frontiers in Human Neuroscience 7, 686. http://dx.doi.org/10.3389/fnhum.2013.00686.

Kida, T., Shinohara, K., 2013. Gentle touch activates the anterior prefrontal cortex: an NIRS study. Neuroscience Research 76 (1–2), 76–82. http://dx.doi.org/10.1016/j.neures.2013.03.006S0168-0102(13)00082-5 [pii].

Kolb, B., Mychasiuk, R., Muhammad, A., Li, Y., Frost, D.O., Gibb, R., 2012. Experience and the developing prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America 2 (Suppl. 109), 17186–17193. doi:10.1073/pnas.1121251109121251109 [pii].

Kostovic, I., Judas, M., Petanjek, Z., 2008. Structural development of the human prefrontal cortex. In: Nelson, C., Luciana, M. (Eds.), Handbook of Developmental Cognitive Neuroscience, second ed. MIT Press, Cambridge, MA.

Luciana, M., 2013. Adolescent brain development in normality and psychopathology. Development and Psychopathology 25 (4, Pt 2), 1325–1345. doi:10.1017/S0954579413000643S0954579413000643 [pii].

Matsuzaka, Y., Akiyama, T., Tanji, J., Mushiake, H., 2012. Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. Proceedings of the National Academy of Sciences of the United States of America 109 (12), 4633–4638. doi:10.1073/pnas.1119971109119971109 [pii].

Minagawa-Kawai, Y., Matsuoka, S., Dan, I., Naoi, N., Nakamura, K., Kojima, S., 2009. Prefrontal activation associated with social attachment: facial-emotion recognition in mothers and infants. Cerebral Cortex 19 (2), 284–292. doi:10.1093/cercor/bhn081bhn081 [pii].

Morris, R.W., Dezfouli, A., Griffiths, K.R., Balleine, B.W., 2014. Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. Nature Communications 5, 4390. doi:10.1038/ncomms5390ncomms5390 [pii].

Morrison, K.E., Rodgers, A.B., Morgan, C.P., Bale, T.L., 2014. Epigenetic mechanisms in pubertal brain maturation. Neuroscience 264, 17–24. doi:10.1016/j.neuroscience.2013.11.014S0306-4522(13)00954-8 [pii].

Motzkin, J.C., Philippi, C.L., Wolf, R.C., Baskaya, M.K., Koenigs, M., 2014. Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. Biological Psychiatry. http://dx.doi.org/10.1016/j.biopsych.2014.02.014 pii:S0006-3223(14)00109-7.

Mychasiuk, R., Harker, A., Ilnytskyy, S., Gibb, R., 2013. Paternal stress prior to conception alters DNA methylation and behaviour of developing rat offspring. Neuroscience 241, 100–105. doi:10.1016/j.neuroscience.2013.03.025S0306-4522(13)00247-9 [pii].

Mychasiuk, R., Zahir, S., Schmold, N., Ilnytskyy, S., Kovalchuk, O., Gibb, R., 2012. Parental enrichment and offspring development: modifications to brain, behavior and the epigenome. Behavioural Brain Research 228 (2), 294–298. doi:10.1016/j.bbr.2011.11.036S0166-4328(11)00835-7 [pii].

Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., Yamamoto, J., Kojima, S., 2012. Cerebral responses to infant-directed speech and the effect of talker familiarity. Neuroimage 59 (2), 1735–1744. doi:10.1016/j.neuroimage.2011.07.093S1053-8119(11)00906-2 [pii].

Nelson, C.A., Monk, C.S., Lin, J., Carver, L.J., Thomas, K.M., Truwit, C.L., 2000. Functional neuroanatomy of spatial working memory in children. Developmental Psychology 36 (1), 109–116.

Olson, E., Luciana, M., 2008. The Development of Prefrontal Cortex Functions in Adolescence: Theoretical Models and a Possible Dissociation of Dorsal versus Ventral Subregions.

Ordaz, S.J., Foran, W., Velanova, K., Luna, B., 2013. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. Journal of Neuroscience 33 (46), 18109–18124. doi:10.1523/JNEUROSCI.1741-13.201333/46/18109 [pii].

O'Reilly, R.C., 2010. The what and how of prefrontal cortical organization. Trends Neuroscience 33 (8), 355–361. http://dx.doi.org/10.1016/j.tins.2010.05.002S0166-2236(10)00076-7 [pii].

Perlman, S.B., Luna, B., Hein, T.C., Huppert, T.J., 2014. fNIRS evidence of prefrontal regulation of frustration in early childhood. Neuroimage 85 (Pt 1), 326–334. doi:10.1016/j.neuroimage.2013.04.057S1053-8119(13)00404-7 [pii].

Petanjek, Z., Judas, M., Simic, G., Rasin, M.R., Uylings, H.B., Rakic, P., Kostovic, I., 2011. Extraordinary neoteny of synaptic spines in the human prefrontal cortex. Proceedings of the National Acadamey of Science of the United States of America 108 (32), 13281–13286. http://dx.doi.org/10.1073/pnas.1105108108105108108 [pii].

Rektorova, I., Sedlackova, S., Telecka, S., Hlubocky, A., Rektor, I., 2008. Dorsolateral prefrontal cortex: a possible target for modulating dyskinesias in Parkinson's disease by repetitive transcranial magnetic stimulation. International Journal of Biomedical Imaging 2008, 372125. http://dx.doi.org/10.1155/2008/372125.

Rushworth, M.F., Behrens, T.E., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nature Neuroscience 11 (4), 389–397. doi:10.1038/nn2066nn2066 [pii].

Saito, Y., Aoyama, S., Kondo, T., Fukumoto, R., Konishi, N., Nakamura, K., et al., 2007. Frontal cerebral blood flow change associated with infant-directed speech. Archives of Disease in Childhood – Fetal and Neonatal Edition 92 (2), F113–116 adc.2006.097949 [pii]10.1136/adc.2006.097949.

Shaw, P., Eckstrand, K., Sharp, W., Blumenthal, J., Lerch, J.P., Greenstein, D., Rapoport, J.L., 2007. Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. Proceedings of the National Academy of Sciences of the United States of America 104 (49), 19649–19654. http://dx.doi.org/10.1073/pnas.0707741104 pii:0707741104.

Somerville, L.H., Jones, R.M., Ruberry, E.J., Dyke, J.P., Glover, G., Casey, B.J., 2013. The medial prefrontal cortex and the emergence of self-conscious emotion in adolescence. Psychological Science 24 (8), 1554–1562. doi:10.1177/0956797613475633095679761347563309567976134756 33 [pii].

Sowell, E.R., Thompson, P.M., Leonard, C.M., Welcome, S.E., Kan, E., Toga, A.W., 2004. Longitudinal mapping of cortical thickness and brain growth in normal children. Journal of Neuroscience 24 (38), 8223–8231. doi:10.1523/JNEUROSCI.1798-04.200424/38/8223 [pii].

Takeuchi, H., Taki, Y., Hashizume, H., Asano, K., Asano, M., Sassa, Y., Kawashima, R., 2013. The impact of television viewing on brain structures: cross-sectional and longitudinal analyses. Cerebral Cortex. http://dx.doi.org/10.1093/cercor/bht315 pii:bht315.

Tanji, J., Hoshi, E., 2008. Role of the lateral prefrontal cortex in executive behavioral control. Physiological Reviews 88 (1), 37–57. doi:10.1152/physrev.00014.200788/1/37 [pii].

Thompson-Schill, S.L., Ramscar, M., Chrysikou, E.G., 2009. Cognition without control: when a little frontal lobe goes a long way. Current Directions in Psychological Science 18 (5), 259–263. http://dx.doi.org/10.1111/j.1467-8721.2009.01648.x.

Unterrainer, J.M., Ruh, N., Loosli, S.V., Heinze, K., Rahm, B., Kaller, C.P., 2013. Planning steps forward in development: in girls earlier than in boys. PLoS One 8 (11), e80772. doi:10.1371/journal.pone.0080772 PONE-D-13–28508 [pii].

Vetter, N.C., Weigelt, S., Dohnel, K., Smolka, M.N., Kliegel, M., 2014. Ongoing neural development of affective theory of mind in adolescence. Social Cognitive and Affective Neuroscience 9 (7), 1022–1029. doi:10.1093/scan/nst081nst081 [pii].

Waytz, A., Zaki, J., Mitchell, J.P., 2012. Response of dorsomedial prefrontal cortex predicts altruistic behavior. Journal of Neuroscience 32 (22), 7646–7650. doi:10.1523/JNEUROSCI.6193-11.201232/22/7646 [pii].

Wunderlich, K., Rangel, A., O'Doherty, J.P., 2009. Neural computations underlying action-based decision making in the human brain. Proceedings of the National Academy of Sciences of the United States of America 106 (40), 17199–17204. doi:10.1073/pnas.0901077106 [pii].

Prehistoric Art

Margaret W Conkey, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 17, pp. 11982–11986, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

In this article, the notion of 'prehistoric art' is shown to be a difficult one, with many implications and problems, and it is suggested that the term could be replaced by 'archaeological arts.' Several issues are raised for the study of prehistoric arts: the problems of (differential) preservation, the importance of archaeological context, the utility of materials science, and the techniques of chronometric dating. Among the major research questions in this field, two questions are highlighted: what are the 'origins' or the 'birth' of art, and how 'meaning' can be inferred from so-called objects of art.

People have been making objects, forms, and images that we consider to be 'art' for at least 35 000 years and perhaps even longer (Tribolo et al., 2006). Some claim that evidence of red ochre, as a kind of coloring material, indicates human engagement with artistic activities (such as body painting) for more than 200 000 years in Africa. Others have pointed to individual objects, dated to even earlier than that, which they are convinced were modified by ancient humans with an intent to represent something, suggesting that 'art' has long been a part of the cultural repertoire of humans, even of different biological species than we anatomically modern humans (Burdukiewicz, 2014; Zilhão, 2007). When most lay people learn that they are talking to an archaeologist, they inevitably ask, 'What is the most fantastic thing you have ever found?,' hoping we will be able to tell them about a gold decorated necklace, a carved figurine, or magnificent painted jar: they want to know about 'prehistoric art.'

The very term, 'prehistoric art,' is difficult to define (e.g., Palacio-Peréz, 2012). Technically speaking, it refers to the 'art' made by prehistoric peoples; that is, by people in those societies without, and usually before the development of, written texts. There have in recent times been human societies without writing systems (nonliterate) and with what might be called 'the arts,' but these are more likely to be what we call 'ethnographic arts' because they are still 'living' arts and can be described ethnographically. There are some problems with the way in which the term 'prehistory' is used, including the implication that, without writing systems, prehistoric groups have no history, and experienced little change. But we know this is not the case at all, and that there are myriad ways other than in writing by which 'history' is reckoned, transmitted, and experienced. Also, in relation to the idea of 'prehistoric art,' there are some archaeologically known societies, such as the Maya or Egyptians, who had writing systems, but whose 'arts' are sometimes referred to as prehistoric or, more often, as ancient art. In fact, it is perhaps better to think of the arts that we are interested in as the archaeological arts, rather than prehistoric ones. With such a term, we realize we are dealing with the art(s) from past human societies that we study by using the wide range of archaeological methods, theories, and techniques.

Then there is the term 'art,' which is itself notoriously difficult to define. Furthermore, it is a term that has had differing meanings even within the last two centuries of Western scholarship and culture. Since the nineteenth century, 'we' in the Western tradition have associated the term, 'art,' with aesthetics and aesthetically pleasing attributes; as something that is usually produced by specialists who are often considered to be gifted individuals with admirable degrees of creativity, and who are often separated from real-life, everyday activities. Our present-day notions of art include the idea that art is to be held in awe while, at the same time, the production and products of art are embedded in the capitalist system, with economic as well as cultural value; art objects function in their own domain, that of high culture and aesthetics.

However, all this is unlikely to be applicable to the art-making and art images and forms that we are concerned with in prehistoric and archaeological contexts. It is a well-known truism in anthropology that in most cultures which have been studied ethnographically, there is rarely a word, or even a concept, equivalent to our word and concept, 'art.' It is not really possible to try to separate out 'art' from other aspects of cultural life, such as art as something distinct from religion, social relations, and/or politics. Yet there are images, forms, and cultural practices in other cultural contexts, both ethnographic and archaeological, that we acknowledge unproblematically as 'art,' mostly because, if they were produced by our own culture, we would classify them readily as art: cave and rock paintings, sculpted figurines or statues, decorated bronze goblets, and the like.

Is there a better term: Visual culture? Material culture? Images? Representations? Material symbols? The term or label we use, such as 'art,' is not itself the real problem; what is more of an issue is that if we use such a term in our own sense, without recognizing its historical specificity, then what we mean by it is often left unexamined, and the analysis or interpretation of the phenomenon is led in certain directions because of the label. If what we have is 'art,' then there must be artists and aesthetic meanings. However, as Morphy (1999) has so eloquently noted: we need to develop a concept of art that 'neither distances the objects of other societies from Western cultures nor appropriates the object to European categories.'

Books that treat the topic of 'prehistoric art' as a single entity are rare; most book-length treatments are concerned with the art and imagery of specific regions (e.g., Keyser, 1992), time periods (e.g., Leroi-Gourhan, 1967; White, 2003), or in specific

media (Dowson, 1992; Clottes, 2002). In 1966, Powell authored an overview of prehistoric art, which was divided into four chapters: the art of the hunters, that of the cultivators, that of the metalworkers, and what he called the 'art of a barbarian nation,' referring to the Celts of Europe (Powell, 1966). His coverage was limited to Europe, and he treated each 'kind' of art by apparent categories of 'lifestyle,' but it is just as much a book about the 'progress' of art from the Paleolithic period (the hunters), through the Neolithic (the cultivators), the Bronze and Iron Ages (the metalworkers), to the 'archaic state' (Celts). As such, he presents a now-problematic evolutionist view, a view of 'progress' through different chronocultural periods. Indeed, for much of the nineteenth and twentieth century, the ethnographic (and archaeological) arts have been embedded in a progressivist notion (Errington, 1998); we used to talk unquestioningly of 'primitive art' (Price, 1989). Twenty-first century scholars, however, have become increasingly critical of the twentieth-century paradigm, instead demanding a broader conception of art that expands beyond Eurocentric boundaries placed on the concept (Moro-Abadía and González-Morales, 2008).

More recently, Bahn (1998) has authored an illustrated history of prehistoric art, but it is almost exclusively about 'rock art,' with a short chapter on 'body art' and one on 'objets d'art,' or portable art (items that could circulate). He returns again and again to a baseline of the Ice Age imagery of Europe, although this book is an important source for appreciating that rock art – paintings, engravings, and petroglyphs on rock surfaces, rock shelters, and in caves – is truly global in its distribution. Bahn's map of sites is a map of the world, while Powell's is one of Europe. This is indicative of a real shift of scope in our understanding and treatments of 'early art' and/or rock art from all times. As we will see later, the search for the earliest art has been one of the major preoccupations of many who study 'prehistoric art.'

How to Study Prehistoric Art

First, it is always important to recognize that the corpus of materials that we consider as 'prehistoric art' is certainly only a sample of what was made in the human past. There is always the problem of preservation, and that what we have in the present is only a sample of past artistic production. It is no coincidence that rock art, ceramic sculpture and decorated pots, metal objects, and images made out of ivory, bone, antler, shell, and beads of various sorts are some of the primary forms of prehistoric art. The wood, hide, basketry, cloth, netting and twining, feathers, and body art are rarely, if ever, preserved. If the arts of ethnographic groups are any indication, we have an impoverished archaeological sample of prehistoric visual and material culture.

But given the preservation of at least some portion of prehistoric artistic production, how do we then go about studying it? First, we need some theoretical or conceptual frameworks within which even to think about what the images and forms might be about, especially if we cannot directly apply our own notions of 'art,' artists, and aesthetics. Much of this has to come from related fields, such as the anthropology of art, the ethnographic arts, and art history. We need to draw on understandings of what visual and material representation is all about (e.g., Gombrich, 1961; Bryson, 1983), of symbolism (e.g., Geertz, 1983), and of humans as materialists and symbolists (Conkey, 1993; Gell, 1998); of how technology and materials are integral to the art forms (e.g., Lechtman, 1984), of style (Conkey and Hastorf, 1990); and of the social contexts and production of art (e.g., Wolff, 1981), and of the 'social life of things' (e.g., Appadurai, 1986). Studies of 'art' in ethnographic contexts (e.g., Morphy, 1991), general overviews of the anthropology of art (e.g., Layton, 1981), and ethnoarchaeological studies (*see* Ethnoarchaeology) provide important resources.

Second, we need to recognize the importance of the archaeological context of the so-called art forms, and we need to take context into consideration when we try to interpret or understand the meanings, uses, or even the making of the art. Do the preconceptions that female figurines – from about 7000 years ago from sites in such places as southeast Europe – are objects of worship and a female-centered world change when we recognize that some, if not many, of the figurines have been recovered from trash deposits, often broken? Are the Ice Age paintings of southwestern Europe found deep in caves, away from daylight, comparable to those found on daylit shelters? Is there a difference in the uses and meanings between objects found in everyday places and living sites, and those found in architectural structures that are so different in size, form, and location that they must be other-than-'houses'? The individual images and objects may well take their specific meanings from the locations where they were made and/or found, or from the other artifacts or structures with which they are associated.

We can use the most up-to-date techniques and methods, especially those of materials science, to make inferences about the images and objects: we can identify the sources of the raw materials, such as clay, stones, or pigments, which in turn can inform on trade, exchange, circulation of goods, and the social relations that effect such practices. We can identify different ways in which pigments were processed by Ice Age cave 'artists' in Europe (e.g., Clottes, 1993), and therefore suggest the social relationships that must have been obtained in order to have images in different sites made by the same pigment 'recipes.' We can show how certain fibers were used to apply pigment in Australian rock paintings (e.g., Cole and Watchman, 1992). We can infer the stages of manufacture of metal earspools (e.g., Lechtman, 1984) and how these stages resonate with other cultural principles and practices. We can use infrared photography to 'see' color and images that have long since faded away. We can use high-powered microscopic analysis to detect the precise strokes of a stone tool on the surface of a bone object; from such analyses have come the suggestion that lunar calendars may have been made at least 15 000 years ago (Marshack, 1991). These are just a few examples of how our increasingly sophisticated methods of materials science that are common in the analysis of other archaeological materials can be applied to the study of prehistoric art images and objects.

Both of these approaches – knowing the archaeological contexts of the art and the application of scientific techniques – have featured prominently in the increasing ability to date many objects and images (David et al., 2013; Pettitt and Pike,

2007; Tribolo et al., 2006). It has long been the case that objects that are found in stratigraphic context – that is, in the geological/archaeological layers of an archaeological site – can, if excavated and recorded carefully, be attributed to at least a relative date (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology). They can also be given more secure dates if other artifacts or materials in the same stratigraphic layer are suitable for some of the many techniques of archaeological or chronometric dating (*see* Archaeometry and Chronology, Stratigraphy, and Dating Methods in Archaeology). However, many art forms, such as rock art or objects that are unfortunately out of context, cannot be dated in this associative manner. Many of these images themselves cannot be dated at all, although recent advances have changed this somewhat. Earlier generations of scholars, for example, dated all the cave art of Europe or the varying rock art styles of Australia by means of stylistic criteria, often based on assumptions as to what 'ought' to be earlier image-making and what 'ought' to be later. At the same time, many researchers were reluctant even to consider the art as a subject worthy of archaeological study, primarily because it could not be dated (*see* Whitley and Loendorf, 1994).

That much art cannot be dated has, in combination with the fact that the interpretation of art is never straightforward (if only because it is the product of symbolic cultural practices that are not easily comprehended through the usual archaeological methods), led many archaeologists to dismiss art and images as viable archaeological data. This has been the case especially when it comes to objects and images from many thousands of years ago, where there are no real historic links between the archaeological 'cultures' and those known ethnohistorically or ethnographically, where there is no direct historical method that can be applied (*see* Archaeology, Theory in). Some have suggested that the recent explosion of interest in the global study of rock art is due, in part, to new dating methods that have allowed researchers to begin to date imagery with more security. For this genre of prehistoric art – rock art – the most important new technique has been the development of accelerator mass spectrometry carbon-14 dating, which can be done using a very minute sample of material, thus not destroying a painting or image, although it can only be used if there are organic materials in the pigments, such as charcoal (*see* Archaeometry).

Some Major Questions

At least two major questions come to mind when we sit down to address the subject of prehistoric art: first there is the question of the origins or the 'birth' of art; second, there is the challenging issue of how we can interpret the art and approach what it might mean. Until recently, it was thought that the cave paintings and other arts of the Upper Paleolithic period (*c*.30 000–10 000 years ago) in Europe comprised the 'first' art, even though images from ancient cultures had been found all over the globe (see Bahn, 1998 on the 'discovery' of prehistoric art). This was a compelling idea because there were many different caves with images and many other objects and art forms; because it seemed to be associated with the earliest evidence for anatomically modern humans – our own species, *Homo sapiens*; and because it was in Europe, which allowed for the idea of a continuous tradition of art – a crucial essence of what it means to be human – from deep prehistory up into the twenty-first century.

However, work during the last few decades of the twentieth century has undermined seriously the idea that there is any single 'birthplace' of art: both southern Africa (Tribolo et al., 2006) and Australia (David et al., 2013) have yielded rock art images that can be attributed to the same general time period as the European Paleolithic imagery. There are those who point to more isolated finds (see Bahn, 1998: 86–87), although many are debated hotly as being 'art' or intentional representations at even earlier dates. We now recognize that anatomically modern humans have a deeper history, perhaps as much as 200 000 years longer, than previously thought. However, the exact nature of the relationship between being anatomically modern and the appearance of representational and/or symbolic imagery is hardly resolved, but the fact that this kind of imagery appears in multiple locations, in varying forms and amounts, calls for a more contextual approach: what is it about life and social relations at these times that set a context within which 'art' – at least in more durable forms that have been preserved – became a part of cultural practices? While the search for 'the origins' of many phenomena in the human past is always a seductive question, as far as prehistoric art is concerned, it is now clear that there was no single 'center' for art-making, but that the preserved evidence around the globe attests the engagement of ancient humans with materials, images, and technologies (of working stone, of coloring materials, etc.) for cultural purposes that still elude us.

Thus, the final, but most difficult question is the interpretive one: how can we infer some (if any) of the meanings of prehistoric art? This question is both the most difficult and most complex question, and what everyone wants to know. Since all of archaeology is interpretive, while drawing on as much empirical evidence as possible, and since human behavior is so multivocal, with so many potential meanings and ambiguities, it is unlikely to expect a single answer, a single interpretation. Even in known or ethnographic contexts, the so-called meaning of art-making or of the art itself is ever-shifting, contextual, and multiple, dependent upon viewers, users, makers, and situations. Do we assume that prehistoric peoples were somehow 'simpler' and therefore a (simple) single account can suffice?

Despite the complexity of the challenge, there are some things that we can say about interpreting prehistoric art. In some instances, we can infer information clearly about clothing from human depictions, about activities as depicted in 'scenes,' preferred symbols or forms, and what kinds of object were traded or used in certain contexts. In other instances, we can advance general theories, such as the role of shamans in making images that derive from their visions (e.g., Lewis-Williams and Dowson, 1989; Turpin, 1994), such as the ways in which ritual landscapes were created with art forms and megaliths, such as the idea that some arts appear to be 'about' interpersonal social relations, whereas other arts appear to be 'about' maintaining the corporate entity (e.g., Layton, 1985). While we are unlikely ever to fix, with certainty, what a particular picture or object 'meant,' recent past decades have witnessed enormous gains, thanks to the application of scientific

techniques, the innovative and expanded theories for interpretation, and a more widespread recognition that the archaeological arts are just as crucial to our understanding of past societies as are other kinds of archaeological data.

See also: Archaeometry; Arctic Archaeology and Prehistory; Art, Anthropological Aspects of; Chronology, Stratigraphy, and Dating Methods in Archaeology; Mesoamerica, Archaeology of; Ritual and Symbolism, Archaeology of.

Bibliography

Appadurai, A., 1986. The Social Life of Things. Cambridge University Press, Cambridge, UK.

Bahn, P., 1998. The Cambridge Illustrated History of Prehistoric Art. Cambridge University Press, Cambridge, UK.

Bryson, N., 1983. Vision and Painting: The Logic of the Gaze. Yale University Press, New Haven, CT.

Burdukiewicz, J., 2014. The origin of symbolic behavior of Middle Palaeolithic humans: recent controversies. Quarternary International 326–327, 398–405.

Clottes, J., 1993. Paint analyses from several Magdalenian caves in the Ariège region of France. Journal of Archaeological Science 20, 223–235.

Clottes, J., 2002. World Rock Art. Getty Conservation Institute, Los Angeles.

Cole, N., Watchman, A., 1992. Painting with plants: investigating fibers in aboriginal rock paintings at Laura, North Queensland. Rock Art Research 9, 27–36.

Conkey, M., 1993. Humans as materialists and symbolists. In: Rasmussen, D.T. (Ed.), The Origins and Evolution of Humans and Humanness. Jones and Bartlett, Boston, MA, pp. 95–118.

Conkey, M., Hastorf, C. (Eds.), 1990. The Uses of Style in Archaeology. Cambridge University Press, Cambridge, UK.

David, B., Geneste, J.-M., Petchey, F., Delannoy, J.-J., Barker, B., Eccleston, M., 2013. How old are Australia's pictographs? A review of rock art dating. Journal of Archaeological Science 40, 3–10.

Dowson, T., 1992. Rock Engravings of Southern Africa. Witwatersrand University Press, Johannesburg, South Africa.

Errington, S., 1998. The Death of Authentic Primitive Art and Other Tales of Progress. University of California Press, Berkeley, CA.

Geertz, C., 1983. Art as a cultural system. In: Geertz, C. (Ed.), Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books, New York.

Gell, A., 1998. Art and Agency: An Anthropological Theory. Clarendon Press, Oxford.

Gombrich, E., 1961. Art and Illusion: A Study in the Psychology of Pictorial Representation, second ed. Pantheon, New York.

Keyser, J.D., 1992. Indian Rock Art of the Columbia Plateau. University of Washington Press, Seattle, WA.

Layton, R., 1981. The Anthropology of Art. Columbia University Press, New York.

Layton, R., 1985. The cultural context of hunter-gatherer rock art. Man (N.S.) 20, 434–453.

Lechtman, H., 1984. Andean value systems and the development of prehistoric metallurgy. Technology and Culture 25 (1), 1–36.

Leroi-Gourhan, A., 1967. Treasures of Prehistoric Art. Harry Abrams, New York.

Lewis-Williams, J.D., Dowson, T., 1989. Images of Power: Understanding Bushman Rock Art. Southern Book Publishers, Johannesburg, South Africa.

Marshack, A., 1991. The Roots of Civilization. McGraw-Hill, New York.

Moro-Abadía, O., González-Morales, M., 2008. Paleolithic art studies at the beginning of the twenty-first century: a loss of innocence. Journal of Anthropological Research 64, 529–552.

Morphy, H., 1991. Ancestral Connections: Art and an Aboriginal System of Knowledge. University of Chicago Press, Chicago.

Morphy, H., 1999. Traditional and modern visual art of hunting and gathering peoples. In: Lee, R.B., Daly, R. (Eds.), The Cambridge Encyclopedia of Hunters and Gatherers. Cambridge University Press, Cambridge, UK, pp. 441–448.

Palacio-Peréz, E., 2012. The origins of the concept of 'palaeolithic art': theoretical roots of an idea. Journal of Archaeological Method and Theory 20, 682–714.

Pettitt, P., Pike, A., 2007. Dating European palaeolithic cave art: progress, prospects, problems. Journal of Archaeological Method and Theory 14, 27–47.

Powell, T.G.E., 1966. Prehistoric Art. Praeger, New York.

Price, S., 1989. Primitive Art in Civilized Places. University of Chicago Press, Chicago.

Tribolo, C., Mercier, N., Selo, M., Valladas, H., Joron, J.-L., Reyss, J.-L., Henshilwood, C., Sealy, J., Yates, R., 2006. TL dating of burnt lithics from Blombos cave (South Africa): further evidence for the antiquity of modern behavior. Archaeometry 48, 341–357.

Turpin, S. (Ed.), 1994. Shamanism and Rock Art in North America. Rock Art Foundation, San Antonio, TX.

White, R., 2003. Prehistoric Art: The Symbolic Journey of Mankind. Harry N. Abrams, New York.

Whitley, D., Loendorf, L., 1994. New Light on Old Art: Recent Advances in Hunter-Gatherer Rock Art Research. Monograph No. 36, UCLA Institute of Archaeology, Los Angeles, CA.

Wolff, J., 1981. The Social Production of Art. St. Martin's Press, New York.

Zilhão, J., 2007. The emergence of ornaments and art: an archaeological perspective on the origin of 'behavioral modernity'. Journal of Archaeological Research 15, 1–54.

Prejudice and Discrimination

Thomas F Pettigrew, University of California Santa Cruz, Santa Cruz, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Intensive research throughout the world has led to the broadening of social science conceptions of both intergroup prejudice and discrimination. Prejudice is now seen as occurring at varying levels of depth. In fact, it can be so subtle that people are often unaware of their prejudices even though these attitudes influence their behavior. Situational effects and social norms also strongly influence the expression of prejudice. Similarly, discrimination is now recognized as largely a complex web of institutional arrangements that produces inequalities between groups. Indeed, the complexity of the phenomenon makes it extremely difficult to study and measure precisely. Discrimination is usually normative. Its structural web outlives its initiators and may not reflect current attitudes. Moreover, discrimination feeds upon itself; it is typically cumulative and self-perpetuating. These changing conceptions have uncovered the reality that the relationship between prejudice and discrimination is not nearly as simple as previously believed. Given the power of long-established and accepted discriminatory norms, many relatively tolerant people routinely discriminate against outgroups without animosity. In turn, anti-discriminatory norms can cause the highly prejudiced to resist acting on their beliefs. To be sure, prejudice is related to discrimination, but not in any simple one-to-one fashion. These considerations have direct implications for social policy designed to combat group discrimination. Structural efforts that increase favorable intergroup contact, such as affirmative action and school desegregation programs, are most likely to diminish both prejudice and discrimination.

Social science conceptions of prejudice, discrimination, and the relationship between the two concepts have changed and become more complex in recent years. Throughout much of the twentieth century, the study of these topics was centered on North America and its racial and ethnic tensions. But increasingly sophisticated twenty-first-century research conducted throughout the world has required broader definitions and theoretical approaches.

Prejudice

Consider first the definition of prejudice. The word 'prejudice' derives from the Latin word *praejudicium* – literally 'before judgment.' And many European languages retain this emphasis on prejudgment today: e.g., *vooroordeel* (Dutch), *Vorurteil* (German), *pregiudizio* (Italian), *preconcello* (Portugese). But modern social science employs the term as signifying not just prejudgment but emotion as well.

Gordon Allport's (1954: p. 9) seminal volume, *the Nature of Prejudice*, offered a straightforward definition that guided research and theory for decades. Ethnic prejudice, he wrote, is "an antipathy based upon a faulty and inflexible generalization. It may be felt or expressed. It may be directed toward a group as a whole, or toward an individual because he is a member of that group." This working definition served social science, and especially social psychology, well because it combined the two key components of the phenomenon – one affective ('antipathy') and the other cognitive ('faulty and inflexible generalization' – better known as a stereotype). Allport wisely chose to omit any link to actual behavior in his definition.

But for some years only one component of Allport's definition received attention. During psychology's 'cognitive revolution' (1960–85), the cognitive features of prejudice were studied intensively by social psychologists while the equally important role of emotion was virtually ignored. Considerable progress was made in understanding stereotypes – gross characterizations of whole human groups. But ignoring affect – hate, envy, threat – rendered an incomplete understanding of prejudice during this period.

Further study, however, uncovered a critical missing element in Allport's approach to prejudice – the situational context in which prejudice is or is not triggered. Consequently, social psychologists now argue for a broader and more precise definition of prejudice. Alice Eagly (Eagly and Diekman, 2005) and her colleagues maintain that a complete definition must consider the social context. Focusing on sexism, Eagly advocates a *role-congruency model*. Women in traditional roles that 'fit' with the female stereotype (e.g., mother, secretary) do not threaten the status quo. Indeed, they typically evoke favorable, although sometimes patriarchal and patronizing, responses. Negative prejudice arises when women occupy 'nonfitting' roles that conflict with the dominant stereotype of females (e.g., business leader, mathematician). These women directly threaten traditional gender norms.

This perception of role incongruence takes place within the larger framework of societal norms. For instance, when enough women rise to head such major American corporations as Lockheed, International Business Machines, and Hewlett-Packard, normative acceptance in the workplace rises even if there remains for sexist individuals a sense of a lack of fit with their stereotypes of women. Discrimination declines for normative and structural reasons although the decline in sexist attitudes may lag in time. Normative change is primarily driving the reduction in discrimination rather than any immediate lessening of prejudice.

Effective measures of sexism reveal this dual view of women (Glick and Fiske, 1996). These scales tap both approval of women in traditional roles and disapproval of them when they seek high-status, nontraditional roles. This phenomenon is called 'ambivalent sexism.' Note that this more subtle and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25086-1

complex form of prejudice acts to resist change in two ways – rewarding those women who do not press for change and punishing those who do. A comparable analysis can be made for racial prejudice.

Consequently, the basic definition of prejudice requires extension. Prejudice will be activated when outgroups threaten the status quo by assuming nontraditional roles that violate established group stereotypes.

A second broadening from Allport's classic definition involves the depth with which prejudice is held. Allport had hinted at this with his assertion that prejudice 'may be felt or expressed.' But research, using an array of different methods, has now shown that prejudice can take a variety of subtle forms – indeed, so subtle that many people are truly unaware of their intolerance and negative outgroup stereotypes. As progress has been made against group discrimination and prejudice in many countries, the older blatant forms of bigotry have receded (but not disappeared) while the more symbolic and subtle forms have increased.

An example of these more subtle forms consists of the withdrawal of positive emotions toward the outgroup without any particular increase in negative feelings (Greenwald and Pettigrew, 2014). Two items have demonstrated this tendency among both European and North American survey samples: "How often have you felt sympathy (or 'admiration') for [outgroup members] who grew up in slums and poverty?" The subtlety of these sympathy and admiration items is indicated by the fact that they are typically *not* seen as reflecting prejudice by most respondents. Yet in seven independent European samples, they not only correlated with other measures of prejudice against immigrants but also predicted prodiscrimination beliefs. Those who reported that they rarely or never felt sympathy or admiration for an immigrant minority in their country were significantly more likely than others to support sending home those immigrants who commit crimes or have no immigration papers.

Another form of these less blatant types of prejudice has been labeled *infrahumanization*. This concept refers to the tacitly held belief that one's ingroup is more human than an outgroup (Leyens et al., 2000). Infrahumanization arises when people reserve 'human essence' for the ingroup and deny it to the outgroup. This process denies the outgroup such secondary emotions as love and attachment. Secondary emotions are what we feel after we first experience such primary emotions as anger, disgust, fear, and happiness. Primary emotions are shared with animals; secondary emotions are thought to be distinctively human. This phenomenon probes the deeper recesses of prejudice that psychoanalytic writers have stressed.

The most studied and surprising of these new methods is the Implicit Association Test (IAT; Banaji and Greenwald, 2013). The basic idea of the test is that associations not accessible to awareness can influence our attitudes and behavior. The IAT, a computer-based measure, requires that users rapidly categorize two target concepts with an attribute (e.g., the concepts 'black' and 'white' with the attribute 'smart'), such that easier pairings (faster responses) are interpreted as more strongly associated in memory than more difficult pairings (slower responses). Thus, if White subjects connect 'white' with 'smart' faster than they manage with 'black' with 'smart,' the IAT records this as a deeply subtle indication of prejudice. The strength of such a measure is that it subverts such widespread problems of explicit self-report methods as the lack of awareness or social desirability bias.

Critics have argued that these new procedures are not assessing prejudice. But consistently these subtle measures have been shown to possess both internal and external reliability and boast correlations with other variables comparable to those of traditional scales of prejudice. The common manner in which those who resist the idea that they are prejudiced is to oppose changes that would benefit stigmatized outgroups – such as affirmative action and desegregated facilities – for ostensibly nonprejudicial reasons.

Another consistent finding of the literally thousands of research studies of prejudice concerns the surprising degree of similarity in findings across different types of prejudice targets and in different nations around the globe. With remarkably few exceptions, prejudice against stigmatized outgroups is typically found to be greater among the elderly, the poorly educated, political conservatives, those with little outgroup contact, and those who score highly on measures of such personality syndromes as authoritarianism and social dominance orientation. The profile of the relatively tolerant is just the opposite: the young, the well educated, political liberals, and those with outgroup contact and low scores on authoritarianism and social dominance orientation (Pettigrew and Tropp, 2011)

Discrimination

A similar broadening of the definition has taken place for understanding intergroup discrimination. The term is derived from the Latin word *discriminatio*, which means to perceive distinctions among phenomena or to be selective in one's judgment. Cognitive psychology retains the first of these meanings, and social science retains the second. But there is also a difference in definitions between the micro- and macrosocial sciences. At the microlevel, discrimination is often viewed as individuals acting to restrict unfairly the resources of others. But at the macro-level, discrimination is employed to describe complex social systems that produce intergroup inequities in social outcomes. It is this second definition, widely used in economics, political science, and especially sociology, that has proven more useful and is the primary focus of this article.

Here again, recent research demonstrates that what initially seemed to be an uncomplicated phenomenon is in reality quite complex. Given the obvious hate involved in the vicious and well-known acts that were perpetrated by White American lynch mobs against African-Americans and Nazis against Jews, it is little wonder that many believe extreme prejudice is normally involved in discrimination. But actually much discrimination occurs without prejudice or even awareness. Discrimination is not, as popularly thought, simply the accumulation of individual acts of unfairness between members of different groups. It consists as well of an elaborate web of institutional arrangements that produces group inequalities across racial, ethnic, gender, or other group lines. Studied throughout the social sciences, these inequalities can involve political power, economic resources, and cultural access.

Standard definitions of discrimination at the macrolevel typically focus on group differences with varying types of qualifications. In sociology, Robin Williams (1947: p. 39; italics added) held that discrimination exists "… to the degree that individuals of a given group who are otherwise *formally qualified* are not treated in conformity …" with how ingroups are treated. In economics, Gary Becker (1968: p. 81; italics added) wrote that economic discrimination occurs "… against members of a group whenever their earnings fall short of the amount *warranted by their abilities.*" But how are we to determine who is 'formally qualified' and what is 'warranted by their abilities'? These significant qualifications to definitions of discrimination raise difficult problems for studying and understanding the phenomenon.

Viewing discrimination from an institutional web perspective casts a wide net, and such definitions leave room for substantial ambiguity. Just what constitutes discrimination and how do you measure it? These questions have both significant policy and legal implications. Disagreement about what constitutes discrimination stems from two sources – one ideological and political and the other empirical. First, because discrimination violates the law and professed values in many countries, a judgment that unequal outcomes reflect discrimination can initiate a demand for unpopular change and costly remedies. Resistance to such change and costs often activates heated debate about the extent of discrimination. Second, deficiencies in analysis and evidence limit the ability of social scientists to trace precisely the dynamic system of effects triggered by discrimination over time.

Consider the political issue. The broadest definitions of discrimination assume that no groups have inherent characteristics that warrant inferior social outcomes. Thus, virtually *all* inequality becomes a legacy of discrimination and a social injustice to be remedied. Favored by the political left, this view appears to most social scientists as too-sweeping. There are obvious reasons for group disparities that most observers would not label as discrimination as revealed by the qualifications made in virtually all the basic definitions of the term.

By contrast, political conservatives favor a narrow definition that limits the concept's scope by including only actions consciously intended to restrict a group's life chances and choices. For several reasons, most social science specialists also reject this formulation. An intentionality criterion returns the concept to the realm of psychology and individual actions while it deflects attention from the critically important restraining social structure. Moreover, the invisibility of intentions creates formidable obstacles for substantiating discrimination. That those responsible for outgroup discrimination deliberately intended to discriminate is exceedingly difficult to prove in a court of law. Finally, the effects of discrimination are the same whether or not the causal institutional mechanisms were intended to produce group inequalities.

This rejection of the intentionality criterion leads to an important distinction between *direct and indirect discrimination*. Direct racial discrimination occurs at points where inequality is generated, often intentionally. When decisions are based explicitly on race, ethnicity, gender, or some other group distinction, discrimination is direct. Indirect discrimination occurs when the inequitable results of direct discrimination serve as a basis for later decisions. Hence, discrimination is indirect when an ostensibly neutral criterion becomes a proxy for the outgroup in determining social outcomes. Indirect discrimination magnifies and perpetuates the original injury into the future. And the older definitions of the phenomenon typically do not allow for the pernicious effects of indirect discrimination.

Wage discrimination illustrates this critical distinction. Direct discrimination exists when employers pay equally qualified members of two groups different rates for the same work. Indirect discrimination exists when the employers pay the two groups unequally because prior discrimination in employment, education, or housing created apparent differences in qualifications or channeled the groups into better- and worse-paying jobs. This direct/indirect distinction resembles the distinction often made by legal systems between disparate treatment and disparate impact. While intentional direct discrimination may have triggered the causal chain, the original injury is perpetuated and magnified by largely unwitting accomplices. To be sure, these accomplices typically benefit from the discriminatory arrangements and have little incentive to alter them. But the important point is that these beneficiaries of the discriminatory practices did not initiate them and may not even be aware of their operation. The discriminatory norms were formed earlier by others, and later generations merely follow the entrenched and long-accepted intergroup norms.

This conceptual complexity carries over to impede the empirical study of discrimination. What may appear as rigorous quantitative analyses often camouflage the crucial issues. An examination of wage *decomposition analysis* illustrates the point. Assessments of racial discrimination produced by decomposing gross group differences in social outcomes are common in both sociology and economics. This approach first assigns one segment of the gross group differential to 'qualifications' and other factors deemed by the researcher as legitimate determinants of social rewards. The residual segment not demonstrably linked to these presumed 'legitimate' determinants of the outcomes is then presented as the estimate of discrimination. But without better information than is usually available and closer agreement on what constitutes discrimination, no unique estimate of discrimination is possible.

This daunting conclusion is the result of the fact that the choice of control variables to index 'legitimate' determinants of social outcomes essentially shapes the answers. Indeed, each such study essentially redefines discrimination with its unique choice of control variables. Any appearance of scientific certitude is an illusion. Not surprisingly, discrimination estimates from this approach using the same basic data vary enormously.

Think about this simple hypothetical example. Two European researchers, using the same data set, each wish to estimate the degree of employment discrimination against first- and second-generation immigrants. The first investigator obtains the employment difference between natives and immigrants and then controls for age, education, gender, and fluency in the national language. The second investigator obtains the same employment differential and controls for the same four variables but adds an additional control for prior work experience (an indicator of indirect discrimination). Obviously, the two investigators have defined discrimination differently. The first researcher, with one less control variable,

will necessarily obtain a larger estimate of discrimination than the second researcher.

Empirical research on group discrimination must mirror the phenomenon in its variety and complexity. The regression decomposition approach is useful but limited. Regression analyses could provide more pertinent information if based on more homogeneous job groups (Conway and Roberts, 1994) and on structural equation models that test reciprocal causation. Most important, if the aim is to guide policy, a framework far more complex than the dichotomous discrimination-or-not approach is required. Research that traces the actual processes of institutional discrimination is essential. Also needed is more attention to victims' perceptions of discrimination and to the changes generated by antidiscrimination efforts.

Two summary propositions emerge from this discussion. First, discrimination is largely normative. Its structural web outlives its initiators and may not reflect current attitudes. Second, discrimination feeds upon itself; it is typically cumulative and self-perpetuating. To be effective, remedies must intervene and reverse its 'vicious circle.'

How Does Prejudice Relate to Discrimination?

Given the complexities involved with both concepts, it is hardly surprising that the relationship between prejudice and discrimination is also complex. One popular view is that there is a simple relationship between the two phenomena: namely, prejudiced individuals of the ingroup discriminate against members of the outgroup while unprejudiced ingroup members do not discriminate. Such a view ignores such important social dynamics as the power of social norms to guide behavior.

On this question, the macro- and micro-social sciences agree (DiTomaso, 2012; Greenwald and Pettigrew, 2014). Sociologists stress the importance of societal structures in producing intergroup discrimination. Such historical realities as past slavery and immigrant poverty are inevitably associated with current differences in employment and wealth and in turn with residential segregation. The resulting lack of intergroup contact provides a breeding ground for perceived difference that can take the form of stereotypes, intergroup threat, and wariness if not outright dislike (Pettigrew and Tropp, 2011). Widespread belief in a 'just-world' can lead to perceiving impoverished outgroups as deserving of their low status, having brought their disadvantages on themselves. In this way, the residue of past discrimination can produce future discrimination as part of a vicious circle (Myrdal, 1944).

Sociologists appeal to *norms* – widely shared understandings of what constitute acceptable social behavior. Both formal norms (e.g., a posted 65-mph automobile speed limit) and informal norms (keeping to the right if driving below the speed limit) powerfully guide behavior (Pettigrew, 1991). Also important are rewards and punishments that support the norms (e.g., an enforced 75-mph speed limit on the same highways). Norms shape intergroup interaction and provide a common meaning to the interaction.

Moreover, discriminatory norms are characteristically cumulative and self-perpetuating. As noted above, normative persistence means that their discriminatory effects typically outlive the initiators of discriminatory practices. The initiators may well have been highly prejudiced; but later generations follow these norms that benefit their ingroup without necessarily harboring the animosity that led to the creation of the norms. As a result, norms are likely to remain unchanged, even while attitudes are shifting markedly. At present, in the midst of rapid formal change in intergroup relations in Northern Ireland, South Africa, and the southern United States, old norms are slowly receding while new equalitarian norms have yet to develop fully. In all three societies, reticence and awkwardness characterize intergroup interaction together with widespread intergroup avoidance and informal discrimination that is not necessarily related to prejudice.

Norms can be strong enough so that people think and behave in conflicting ways in different social contexts without awareness of the inconsistency. At an Indiana Steel Mill in the 1950s, Black and White Americans were members of the same racially desegregated union and worked well together (Reitzes, 1953). Only 12% of the White workers reported low acceptance of African-Americans on the job. Those most involved in the union were the strongest supporters of the union's pro-desegregation norms. But these same White workers also lived in racially segregated, all-White neighborhoods and many belonged to activist pro-segregation neighborhood groups. Indeed, 84% of those who accepted African-Americans at work were highly resistant to Blacks living in their neighborhoods. Those most involved with their neighborhoods were the most resistant. The behavior in each setting favored the setting's prominent ingroup – union solidarity at work, racial solidarity at home.

Social psychologists agree and have extended this sociological analysis. They consider conformity an essential concept for understanding prejudice. Conformity is the link that accounts for how societal forces eventuate into patterns of acceptance or discrimination by individuals. Relatively unprejudiced White Americans generally follow their ingroup's norms. Such norms typically favor their ingroup. If these norms involve preferential treatment of the ingroup, most people – like the Indiana Steel workers – will conform without personal animus toward the outgroup.

Together with conformity, ingroup favorability operates to create discrimination. Quite often ingroup favoritism is hidden even from those who practice it. Consider the way much job recruitment occurs. Good workers are asked frequently to seek out others for job openings. Because of extensive ('de facto') racial segregation in residences, schools, and workplaces, this practice often leads to White workers to draw necessarily on virtually all-White acquaintanceship networks, and thus they recommend only other Whites for the vacancies (Reskin, 1998). By drawing heavily on ingroup ties, this seemingly harmless process can sustain or exacerbate discrimination without the involvement of hostility toward the outgroup.

Thus, relatively unprejudiced people can routinely engage in discrimination, while prejudiced people often do not. Both are following the social norms under which they live. For instance, unprejudiced Whites in the Southern United States of the 1930s had to follow the dominant norms of racial discrimination or face negative consequences in their community. Similarly, highly prejudiced White workers

employed by affirmative action firms in the United States today follow the nondiscriminatory norms of their employer or risk losing their jobs.

In short, different social circumstances can elicit different reactions and override personal attitudes and preferences. But it is important to note that this phenomenon does not deny the potential of prejudice to influence discriminatory behavior. When relevant social norms are not operating, prejudice is likely to relate more closely to discrimination. And even when the norms are operating, the prejudiced and unprejudiced will reveal their differences in the degree of fervor with which they follow the norms. Southern American lynchers and Nazis wreaked their havoc with the support of societal norms; in these instances, individual prejudice and the norms combined with tragic results.

Both experiments and surveys reveal just how strong yet restrained ingroup bias and favoritism can be (Greenwald and Pettigrew, 2014). Blatantly negative stereotypes of the outgroup (e.g., stupid, lazy) are often rejected, but positive stereotypes (e.g., smart, ambitious) are typically assigned more to the ingroup. In effect, the illogic is that 'they' are not worse than 'we,' are but 'we' are better than 'they' are. In addition, equity norms are often compromised. For instance, in competitive game studies, subjects will typically not punish outgroup members, but they will reward ingroup members more if there is some ostensibly 'fair' rationale – such as calling the extra gains a 'bonus.'

The same type of subtle ingroup favorability effects are found in the enormous amount of public opinion survey data gathered over the years on White American attitudes toward racial change. Some of these attitudes have changed dramatically. Surveys of White Americans' racial attitudes over the past 50 years show that Whites have steadily increased their support for policies that provide housing and educational rights for African-Americans. These surveys also reveal White Americans' increasing support for racial intermarriage, equal opportunity in employment, and the acceptability of an African-American candidate for US President. The last of these trends finally came to fruition in 2008.

Yet, White Americans support for policies that provide direct governmental assistance to minorities has been largely unchanged during the past half century. Thus, White Americans have steadily opposed enhanced social services to minorities and affirmative action for under-represented minorities in hiring or college admissions (Schuman et al., 1997).

In other words, national American surveys since 1950 have found that most White Americans accept the basic principles of equal opportunity but still resist the implementation of policies that would bring about equality by directly helping outgroups (Pettigrew, 1979). Nevertheless, ingroup favorability, however, leads a majority of White Americans to favor many government programs that do provide direct assistance largely to themselves – from farm subsidies to tax loopholes and low tax rates on the wealthy.

To sum up, prejudice and discrimination are related, but not in the simple manner as is popularly thought. The prejudiced are more likely to favor and practice discrimination against outgroups. But group discrimination typically arises from long-established norms that most people conform to without hostility or even awareness.

Implications for Social Policy

These major shifts in social science thinking about prejudice and discrimination have significant implications for social policy. The two phenomena are closely related, but not in the simple one-to-one fashion often assumed. Laws and institutional practices that directly counter discriminatory norms are far more likely to reduce intergroup discrimination than attempts limited to reducing prejudice alone. Indeed, the new norms themselves are likely to lessen prejudice as well as discrimination.

Social psychologists have long shown that behavioral changes alter attitudes more easily than do attitudinal changes alter behavior. New norms and new laws can and do change behavior. But, of course, such societal transformations will be resisted when intergroup attitudes are extremely hostile.

Recent advances in the civil rights of gays and lesbians in Europe and North America illustrate this intricate process that weaves prejudice reduction with structural change. In the 1980s, the gay community began to make special efforts to 'come out,' to publicly announce their group status. 'Straights' suddenly discovered that they knew many gays, even friends and family members they did not know previously were gay. Intergroup contact between gays and straights had been widespread, but not until group status was recognized could this contact affect attitudes. Homophobia began to recede, and only then were governments emboldened to legalize same-sex marriage. And this change in the law in turn has led to further reductions in anti-gay prejudice.

Critical to this process is the role of intergroup contact. Social psychologists in many countries have found intergroup contact to be one of the most effective means of reducing prejudice in all of its forms – subtle as well as blatant. If extreme threat is not involved, intergroup contact is surprisingly effective. And a meta-analysis of more than 500 studies conducted in 38 nations shows it is especially effective in diminishing homophobia (Pettigrew and Tropp, 2011).

This process indicates that institutional alterations that further intergroup contact under positive conditions are the most likely to succeed. Desegregated schools and other public facilities as well as affirmative action programs in North America, the Netherlands, and other nations all lead to greater intergroup contact.

Where such programs have taken root, we can expect further decreases in both prejudice and discrimination.

See also: Authoritarian Personality; Desegregation; Intergroup Relations; Racism: Social Psychological Perspectives; Sexism; Social Dominance Orientation; Stereotypes in Social Psychology.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Addison Wesley, Reading, MA.

Banaji, M.R., Greenwald, A.G., 2013. Blindspot: Hidden Biases of Good People. Delacorte Press, New York.

Becker, G., 1968. Economic discrimination. In: Sills, D.J. (Ed.), International Encyclopedia of the Social Sciences. Macmillan, New York.

Conway, D.A., Roberts, H.V., 1994. Analysis of employment discrimination through homogeneous job groups. Journal of Econometrics 61, 103–131.

DiTomaso, N., 2012. The American Non-dilemma: Racial Inequality without Racism. Russell Sage, New York.

Eagly, A.H., Diekman, A.B., 2005. What is the problem? Prejudice as an attitude-in-context. In: Dovidio, J.F., Glick, P., Rudman, L. (Eds.), On the Nature of Prejudice: Fifty Years after Allport. Blackwell, Malden, MA, pp. 19–35.

Glick, P., Fiske, S.T., 1996. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology 70 (3), 491–512.

Greenwald, A., Pettigrew, T.F., 2014. With malice toward none and charity for some: ingroup favoritism enables discrimination. American Psychologist 69 (8).

Leyens, J. Ph, Paladino, M.P., Rodriguez, R.T., Vaes, J., Demoulin, S., Rodriguez, A.P., Gaunt, R., 2000. The emotional side of prejudice: the attribution of secondary emotions to ingroups and outgroups. Personality and Social Psychology Review 4 (2), 186–197.

Myrdal, G., 1944. An American Dilemma. Harper & Row, New York.

Pettigrew, T.F., 1979. Racial change and social policy. Annals of the American Academy of Political and Social Science 441, 114–131.

Pettigrew, T.F., 1991. Normative theory in intergroup relations: explaining both harmony and conflict. Psychology and Developing Societies 3 (1), 3–16.

Pettigrew, T.F., Tropp, L., 2011. When Groups Meet: The Dynamics of Intergroup Contact. Psychology Press, New York.

Reitzes, D.C., 1953. The role of organizational structures: union versus neighborhood in a tension situation. Journal of Social Issues 9, 37–44.

Reskin, B.F., 1998. The Realities of Affirmative Action in Employment. American Sociological Association, Washington, DC.

Schuman, H., Steeh, C., Bobo, L., Krysan, M., 1997. Racial Attitudes in America. Harvard University, Cambridge, MA.

Williams, R.M., 1947. The Reduction of Intergroup Tensions. Social Science Research Council, New York.

Prejudice in Society: Psychological Perspectives

Jacques-Philippe Leyens, Catholic University of Louvain, Louvain-la-Neuve, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prejudice went through different waves of research. First, it was considered a trouble of personality. Second, it was seen as the result of normal processes. Third, it implied implicit and explicit ways of reacting, and showed that people are ambivalent. Like aversive racists who believe not to be racists, their prejudice surfaces in ambivalent ways. Many prejudiced reactions are automatic, but control can appear in disguised ways. In the last decade, dehumanization increasingly has been investigated and links with the theory of mind have been established. Possibilities of reducing prejudice are examined at the group and at the individual level.

Prejudice is defined as unfair beliefs, affects, and discriminatory behaviors toward members of a group as a result of their membership to this group. Prejudiced groups may be ethnic or sexual, for instance, and prejudice will take the name of racism and sexism. By definition, prejudice implies generalizations, and many authors defend the idea that the generalizations are irrational or even immoral. Such a view implies that objective truth is attainable. It also overlooks the fact that different ideologies influence the sophisticated ways that prejudice takes.

Prejudice as Individual Differences

Authoritarian personality was the first explanation of prejudice, which derived from an attempt to understand the fascist establishment and ideology (Adorno et al., 1950). This theory based on right-wing socialization was complemented by a more general one focusing on dogmatism (Rokeach, 1960). Recent theorizing gave rise to Altemeyer's (1988) Right-wing authoritarianism and social dominance (Sidanius and Pratto, 1999). Although based on different theoretical premises, intraindividual versus sociostructural questionnaires designed to tap these individual differences often show high correlations.

Prejudice as a Normal Process

In the early 1970s, Tajfel et al. (1971) imagined a minimal group situation that would serve as a baseline to which variables would be added to obtain discrimination between groups. To their surprise, mere categorization sufficed to lead to ethnocentrism, also called ingroup favoritism bias. Although this bias is not necessarily linked to outgroup derogation, researchers interpreted the phenomenon as discriminatory. A misunderstanding of the minimal group situation thus led people to consider that prejudice was a normal and inevitable process that would provide a positive social identity to group members. Tajfel (1981) proposed the social identity theory that became a main worldwide influx to work on intergroup relations and especially on prejudice from the point of view of dominant and dominated groups. With the variables of permeability (vs impermeability), stability (vs instability), and legitimacy (vs illegitimacy), Tajfel offered precise hypothesis concerning individual mobility and social (intergroup) change.

Beliefs in normality were reinforced by uncontrollable behaviors. All people know the negative stereotypes held by different groups and many do not share them. When they are overloaded by a demanding environment and in need of cognitive capacities, however, they may let prejudice leak. Such perspective makes a clear distinction between beliefs or stereotypes, on the one hand, and prejudice, on the other (Devine, 1989). It should be noted, however, that most books on prejudice (or racism) associate it with stereotypes and discrimination.

Prejudice as Ambivalence

Little by little, Myrdal's (1944) *An American Dilemma: The Negro Problem and Modern Democracy* was recognized by the Americans. They recognized that besides their ideas that all humans are equal, they discriminated against blacks. Until now, ambivalent behaviors and theorizing followed.

- Katz et al. (1986) showed that white Americans behaved ambivalently toward blacks. Depending on the salience of ideology, egalitarianism, or meritocracy (also called Protestant work ethics), they thought blacks were equal to them or believed that blacks did not work sufficiently well and did not deserve credit. They behaved accordingly with black targets. It is ambivalent amplification racism.
- Dovidio and Gaertner (1986) described many whites as aversive racists. They sincerely believe to be nonracists but, unconsciously, they cannot but behave as racists. These people are fundamentally ambivalent.
- For instance, when interacting with a black person, aversive racists are able to control their language and come out happy of the exchange. Their partners, however, are not satisfied and treat the whites as hypocritical because their nonverbal behaviors, which are difficult to control, showed signs of uneasiness instead of kindness. Stated otherwise, aversive racists' behaviors are ambivalent.
- Targets of prejudice may be ambivalent (Steele, 1997). Bright black students were asked to answer to a difficult test on verbal fluency. They were instructed that the task was either part of an exercise or a test to gauge their ability. Compared with the performance of white students, the exercise led to similar results, whereas the test induced a poorer performance. In other words, when asked to disconfirm the negative stereotype about blacks,

these bright students confirmed the stereotype. Apparently, the poor performance was due to interfering thoughts while trying to solve the difficult test. The students normally were able to solve the test; however, when asked to disconfirm the reputation of their group, they thought of possible failure. In other words, targets of prejudice were ambivalent.

- Sexism can be an ambivalent theme. So much so that authors (Glick and Fiske, 1996) working on the topic called their theory 'ambivalent sexism.' One might consider that the theory is doubly ambivalent. Many men are hostile toward women as reports of sexual abuse in newspapers daily convince the readers. Ambivalently, the same men may show benevolence toward women. Where the theory is doubly ambivalent is that benevolence is also considered sexist. Rather than measuring routine politeness or gallantry, benevolent sexism includes paternalism, gender differentiation, and intimate heterosexuality. A questionnaire of ambivalent sexism has been tested successfully in an impressive number of countries. Not unsurprisingly, benevolent sexism attracted most attention and the overall results are convincing.
- Stereotypes circulate about any group and may elicit prejudice. Fiske (2011) found a simple stereotype content model that is able to categorize any group. Besides categorization and content, the model also predicts emotions and behaviors toward the groups. The model is based on the two dimensions that are considered essential, and sufficient, for person perception. The two dimensions are competent versus incompetence, and warmth versus cold. The first dimension refers to status and the second one depends on the cooperation – competition relation between groups. Although the two dimensions are not completely orthogonal, they can be treated as such. Four quadrants result. The warm and competent quadrant includes the ingroup and reference groups. It elicits admiration and provokes both active (helping) and passive (associating) protection. The cold and incompetent quadrant includes drug addicts, homeless persons, and Arabs. The emotion is disgust and the behavior is passive (exclusion) and active (attacking) harm. The other two quadrants are ambivalent; they are high on one dimension and low on the other one. The competent but cold quadrant contains rich people, Asians, and Jews. It leads to envy and behaviorally to passive facilitation (uniting) but active harm (fighting). Finally, warm but incompetent groups are pitied and elicit active facilitation (helping) but passive harm (neglecting). Not much is to be said about the warm and competent group. In accordance with social identity theory, it facilitates ingroup favoritism. The other extreme outgroups, incompetent and cold, are special and will receive more attention in the section devoted to dehumanization. Ambivalence describes the last two quadrants and is recognized by the *but* link between the dimensions. They deal with status that divides people. Their main function is to direct comparisons that may be prejudicial as they provoke feelings of envy and scorn. In sum, most outgroups are represented ambivalently.
- Social interactions between members of different groups have been neglected up to recent investigations and theory (Shelton and Richeson, 2006). Research has adopted an individualistic approach. People have beliefs about one another, beliefs about how the other person views them, and beliefs about the interaction. Stress and coping have been the main sources of inspiration. It has been shown that whites and blacks have similar beliefs, although they are afraid to cope for different reasons. They all share the belief that the interaction protagonist is not interested in their encounter, but whites are frightened to be perceived as racists and blacks are scared to be discriminated against. Racial minorities also seek to be respected and to be seen as competent more than do whites, whereas whites seek to be liked and seen as moral more than racial minorities do. These different aspirations lead to misunderstandings, and all of these misinterpretations show the ambivalence that guides these interactions.
- Color-blindness and color-consciousness are two ideologies that provoke a double ambivalence. Color-blind people believe all individuals are equal and, thus, similar. Color-conscious people think that differences between groups must be recognized. Some decades ago, social psychology was unanimously color-blind. Categories were banned. More recent studies have concluded that color-consciousness, favored by minorities, was more efficient when conducting racially mixed encounters. Especially in the United States, people have ambivalent views about these two ideologies. Educated people state that they are in favor of color-consciousness, probably because research says so, but they continue behaving in color-blind ways. Groups (whites, blacks, women, for instance) should not be singled out in a conversation; minorities' strengths (music, sport, sensitivity) become politically incorrect because they are widely spread within a dominated group. Even mere thinking about a stereotype is considered a sin. Color-conscious people are convinced that they perceive equality among targets. Thus, they have ambivalence in the perception of these two different ideologies. The second ambivalence lies within each perspective. Color-blindness contains two strategies. On the one hand, the world should be inhabited only by individuals who are equal. On the other hand, color-blindness implies that minorities assimilate to the majority and, consequently, is a prejudiced strategy. Color-consciousness also includes two strategies. Either group boundaries are accepted and respected, or they are recognized without respect. These two strategies represent integration or explicit racism, but they both send clear signals to their targets allowing a good performance when having to collaborate. Majorities prefer assimilation, whereas minorities favor integration. Curiously, showing signs of group membership upsets people who defend an individualistic world (Maquil et al., 2009). The general field of research on interactions between majorities and minorities is also ambivalent because it is conducted without reference to the domain of the contact hypothesis, and inversely, contact research is conducted without any links with interaction studies.

Automaticity and Social Control

During the decade preceding this century, automatic (unconscious, implicit, nonconscious) process was pitted against

controlled (conscious, explicit, self-reported) process. Theories and methods considered both processes to be mutually exclusive. Theories followed the cognitive view that the two processes were to be given specific attention. Methods measured separately automatic versus controlled responses. Priming, masking priming, and speed were used to obtain automatic responses nearly impossible to avoid. Control was considered time consuming and effortful in contrast to automaticity.

The new century suggested that no response is process pure (Devine and Sharp, 2009). Theories and methods addressed simultaneous attention to the mix of automaticity and control. New research showed limits in the inevitability of stereotypes activation as well as evidence in favor of the malleability of automatic intergroup biases. It also revealed that control does not necessarily need much intentional effort.

Several variables moderate automatic responding. It is the case of restricted attention. When responses are not processed in a meaningful way because of attentional restrictions, automaticity does not seem to play a role. The environment is another variable; viewing people in front of a church or within a prison leads to different responses. Social roles and appearance of targets are a third variable; for instance, Negroid versus less Negroid face features color the responses. Motivation is a fourth variable. Neural activity reveals that physicians do not respond to pain like laypeople. Physicians inhibit empathy, but it remains a neurosciences question whether this inhibition is conscious or unconscious (Decety, 2011).

Automaticity is malleable, because of practice and counterstereotypic examples (e.g., Nelson Mandela). These variables interfere with the deeply rooted socialization that made some stereotypes almost inevitable. The result of this is that malleability is limited. For instance, it is hard to erase the idea that anger is typical of blacks.

Gaunt et al. (2002) showed the interplay of automatic and controlled processes. Using Jacoby's (1991) process dissociation, they proportioned the role of automatic and familiar responses from the one of control when having to memorize liked and disliked racial stereotypes. The same procedure has been utilized by other researchers to verify, for instance, whether race affected the shooting of black and white criminal suspects.

Neurosciences have attempted to provide more powerful evidence to the interplay between automatic and controlled responses. Readers should realize that this evidence is particularly descriptive and that, up to now, it does not contribute much to theoretical advances. The role of median prefrontal cortex (mPFC) in social cognition will be alluded to later. A study by Cunningham et al. (2004) is famous for showing neural activity of automaticity and control. The latter processes were activated by showing black and white faces at subminimal and supraliminal speeds. Participants' task was to locate the faces. When rapid, automatic responses were activated, there was greater activity of the amygdala for blacks than for whites. The amygdala is known to be a sign of fear. When there was control, the dorsal anterior cingulated cortex and lateral prefrontal cortex (LPFC) were particularly active when viewing black faces. This means mere attention. Moreover, the ventral region of LPFC was also activated more for black faces, which suggests inhibition of a given response. The different regions when the speed was low may show a double response: attention and inhibition.

Dehumanization

To find what is human, one may look at what is uniquely human (e.g., secondary emotions such as love, admiration, contempt, or shame) or what is typically human (e.g., personality traits common to humans, such as curious, friendly, impatient, or jealous). Ingroup members attribute more positive and negative secondary emotions to their group than to (many) outgroups. Primary emotions, shared with animals, should not differ between ingroup and outgroup (Leyens et al., 2001, 2007). Because ingroup members are perceived as human, stimuli linked to them are also humanized. This phenomenon has been called infrahumanization (Leyens et al., 2000) or animalistic dehumanization; it is mainly an intergroup process. It complements a mechanistic dehumanization (Haslam, 2006) on the basis of typically human characteristics. Such a name implies that people lacking typical attributes become like robots. It is an individual and intergroup process. Individuals and ingroups give more negative typical features to them or to their group than they give to ingroup members and to outgroups. A third approach to humanity derives from the theory of mind. The mind has two dimensions: agency and experience (Gray et al., 2007). Infrahumanization is linked especially to agency, but it also relates to experience. Mechanistic dehumanization works mainly with experience and shows some degree of agency too.

Research on infrahumanization has shown that the ingroup may be particularly humanized, that outgroups are dehumanized, or that both processes work simultaneously. Conflict and dislike are unnecessary to obtain infrahumanization, but ingroup members must be identified with their group. Status and power of the ingroup moderate the effects. High-status groups or powerful groups infrahumanize more than do low-status or dominated groups.

The more adjectives that were estimated to be descriptive of people in the cold and incompetent quadrant of the stereotype model, the more inhuman they were rated. Such a result relates more to dehumanization than with infrahumanization (Vaes and Paladino, 2010). Moreover, exemplars of groups of drug addicts and homeless people were shown to people in a scanner who had to say which emotion they felt. Exemplars of groups belonging to the other quadrants were also displayed with the instruction to reveal the emotion. Functional magnetic resonance imaging was taken. The mPFC was proven to react to social cognition. Whereas mPFC was activated for regular people, it did not respond to members of cold and incompetent groups. In fact, these members showed neural activation linked to disgust (Harris and Fiske, 2006).

Reduction of Prejudice

Contact Hypothesis

Many authors claim that the best predictor of prejudice reduction is intergroup contact. According to all researchers, this contact must meet conditions spelled out by Allport (1954): equal group status within the situation, common goals, intergroup cooperation, and the support of institutional authorities. It was also considered that the contact situation must provide opportunity to become friends. US scholars

favor decategorization to meet these conditions. Europeans, accustomed to many countries, defend the idea that boundaries have to be respected. Stated otherwise, US researchers plead for diminished group saliency, whereas European authors privilege increased group saliency. This opposition evokes the notion of color-blindness versus color-consciousness. Pettigrew (1998) saw a solution in taking time into account. First, essential contact conditions and individual differences should lead in a first phase to decategorization. When trust is settled between groups, there could be salient categorization, also called harmonious differentiation. In a third phase, groups could envisage uniting and forming a new group, called recategorization.

This causal sequence with contact leading to reduced prejudice is still problematic. Prejudice also elicits unpleasant contact. This possibility explains the importance of friendship while the contact is established and allows for the performance of common tasks. The US Armed Forces are known to constitute an example of success concerning racial contacts. With little doubt, this success is due to the conjunction of the essential facilitating conditions. By contrast, difficulties in school desegregation showed the lack of ability to meet these same conditions.

The contact hypothesis does not work similarly in high-status and low-status groups (Tropp and Pettigrew, 2005). Meta-analysis reveals that the relationships between contact and prejudice tend to be weaker among members of minority status groups than among members of majority status groups. This difference may be due the perception of contact by minorities. Dominated groups enter less easily into contact because they view it negatively. It is not really a question of anxiety as minorities perceived a significant social distance in intergroup contact (Binder et al., 2009).

Individualistic Models

In addition to the categorical models evoked in the contact hypothesis (decategorization, harmonious differentiation, and recategorization), an approach relies mostly on the intellectual (and somewhat emotional) abilities of individuals to consider evidence contradictory to their prejudice. Three models have been proposed as possible ways to reduce prejudice.

The Conversion Model

According to the conversion model, which could also be called the Desmond Tutu or Nelson Mandela model, an encounter with an atypical member of a disliked group could provoke a conversion that would generalize to the whole category. Empirical evidence is scarce.

The Bookkeeping Model

In the bookkeeping model, people integrate different pieces of information that they receive from members of stigmatized groups and that contradict their beliefs. Reduction of prejudice would be proportionate to the contradicting information. Empirical evidence exists but is not overwhelming.

The Subtyping Model

When confronted with contradicting information, people try to subcategorize the exceptions to maintain their representations of the whole category intact. A strategy often is used in which representations are made such as, "These women are unemotional; but they are not real women, they are businesswomen."

Reconciliation

Intergroup conflicts have attracted much attention since the Cold War between the United States and the former Soviet Union. The realist approach takes into account rational actors who bargain. Several scholars have recommended initiating cooperative gestures. The opponent feels the need to respond in the same perspective, and bargains may lead to agreements (Nadler et al., 2008). Recent research looks at reconciliation in the path of the Trust and Reconciliation Commission in South Africa. Intergroup reconciliation now investigates how to reduce emotional barriers blocking potential solutions. A minimum of trust is supposed to exist between the victims and perpetrators. Trust, however, does not suffice. Dominated groups suffer because they are represented as powerless; they thus come to discussions with a focus on conflicting matters and the wish to regain some power. Dominating groups, on the other hand, experience the feeling of lacking morality. They will take part in meetings with dominated members with the hope that they will be recognized as moral actors. To this aim, they propose compromising matters. As a result, conflicting points are opposed to conciliation efforts. This distinction is parallel to the search for power and for morality. Time is necessary to expose respective ethnocentrism. Studies suggest that these different ethnocentrisms need to be respected rather than risk having these groups remove their trust.

See also: Affirmative Action, Sociology of; Authority, Social Theories of; Communitarianism, Sociology of; Discrimination: Racial; Ethnic, Racial, and Nationalist Movements; Movements, Social; Nationalism, Sociology of; Racism, Sociology of; Social Psychology, Sociological.

Bibliography

Adorno, T., Frenkel-Brunswik, E., Levinson, D., Sanford, R., 1950. The Authoritarian Personality. Harper, New York.

Allport, G., 1954. The Nature of Prejudice. Addison Wesley, Reading, MA.

Altemeyer, B., 1988. Enemies of Freedom: Understanding Right-Wing Authoritarianism. Jossey-Bass, San Francisco, CA.

Binder, J., Zagefka, H., Brown, R., Funke, F., Kessler, T., Mummendey, A., Maquil, A., Demoulin, S., Leyens, J.-P., 2009. Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis amongst majority and minority groups in three European countries. Journal of Personality and Social Psychology 96, 843–856.

Cunningham, W.A., Raye, C.L., Johnson, M.K., 2004. Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. Journal of Cognitive Neurosciences 16, 1717–1729.

Decety, J., 2011. The neuroevolution of empathy. Annals of the New York Academy of Sciences 1231, 35–45.

Devine, P., 1989. Stereotypes and prejudice: their automatic and controlled components. Journal of Personality and Social Psychology 56, 5–18.

Devine, P.G., Sharp, L.B., 2009. Automaticity and control in stereotyping and prejudice. In: Nelson, T.D. (Ed.), Handbook of Prejudice, Stereotyping, and Discrimination. Psychology Press, New York.

Dovidio, J.F., Gaertner, S.L., 1986. Prejudice, discrimination, and racism: historical trends and contemporary approaches. In: Dovidio, J.F., Gaertner, S.L. (Eds.), Prejudice, Discrimination, and Racism. Academic Press, New York, pp. 1–34.

Fiske, S.T., 2011. Envy up, Scorn down: How Status Divides Us. Russell Sage Foundation, New York.

Gaunt, R., Leyens, J.-P., Demoulin, S., 2002. Intergroup relations and the attribution of emotions: controllability of memory for secondary emotions associated to ingroup versus outgroup. Journal of Experimental Social Psychology 38, 508–514.

Glick, P., Fiske, S.T., 1996. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology 70, 491–512.

Gray, H.M., Gray, K., Wegner, D.M., 2007. Dimensions of mind perception. Science 315, 619.

Harris, L.T., Fiske, S.T., 2006. Dehumanizing the lowest of the low: neuro-imaging responses to extreme out-groups. Psychological Science 17, 847–853.

Haslam, N., 2006. Dehumanization: an integrative review. Personality and Social Psychology Review 10, 252–264.

Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.

Katz, I., Wackenhut, J., Hass, R.G., 1986. Racial ambivalence, value duality, and behavior. In: Dovidio, J.F., Gaertner, S.L. (Eds.), Prejudice, Discrimination, and Racism. Academic Press, Orlando, FL, pp. 35–59.

Leyens, J.-P., Demoulin, S., Vaes, J., Gaunt, R., Paladino, M.P., 2007. Infra-humanization: the wall of group differences. Social Issues and Policy Review 1, 139–172.

Leyens, J.-P., Paladino, M.P., Rodríguez-Torres, R., Vaes, J., Demoulin, S., Rodríguez-Pérez, A., Gaunt, R., 2000. The emotional side of prejudice: the attribution of secondary emotions to ingroups and outgroups. Personality and Social Psychology Review 4, 186–197.

Leyens, J.-P., Rodriguez, A.P., Rodriguez, R.T., Gaunt, R., Paladino, P.M., Vaes, J., Demoulin, S., 2001. Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. European Journal of Social Psychology 31, 395–411.

Maquil, A., Demoulin, S., Dovidio, J.F., 2009. Strategies for prejudice reduction: the norms of non-discrimination. In: Demoulin, S., Leyens, J.-P., Dovidio, J.F. (Eds.), Intergroup Misunderstandings: Impact of Divergent Social Realities. Psychology Press, New York, NY.

Myrdal, G., 1944. An American Dilemma: The Negro Problem and Modern Democracy. Harper & Bros, New York.

Nadler, A., Malloy, T.E., Fisher, J.D., 2008. The Social Psychology of Intergroup Reconciliation. Oxford University Press, New York.

Pettigrew, T.F., 1998. Intergroup contact theory. Annual Review of Psychology 49, 65–85.

Rokeach, M., 1960. The Open and Closed Mind. Basic Books, New York.

Shelton, J.N., Richeson, J.A., 2006. Interracial interactions: a relational approach. Advances in Experimental Social Psychology 38, 121–181.

Sidanius, J., Pratto, F., 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. University Press, New York.

Steele, C.M., 1997. A threat in the air: how stereotypes shape intellectual identity and performance. American Psychologist 52, 613–629.

Tajfel, H., Billig, M., Bundy, R.P., Flament, C., 1971. Social categorization and intergroup behaviour. European Journal of Social Psychology 1, 149–177.

Tajfel, H., 1981. Human Groups and Social Categories: Studies in Social Psychology. Cambridge University Press, Cambridge.

Tropp, L.R., Pettigrew, T.F., 2005. Relationships between intergroup contact and prejudice among minority and majority status groups. Psychological Science 16, 951–957.

Vaes, J., Paladino, M.P., 2010. The uniquely human content of stereotypes. Group Processes and Intergroup Relations 13, 23–39.

Prejudice in Society: Sociological Perspectives

Richard T Schaefer, DePaul University, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prejudice is a negative attitude that rejects an entire group; discrimination is behavior that deprives a group of certain rights or opportunities. Prejudice does not necessarily coincide with discrimination, as is made apparent by a typology developed by sociologist Robert Merton. Several theories have been advanced to explain prejudice: scapegoating, exploitation, and the normative approach. Although widespread expression of prejudice has declined, color-blind racism allows the status quo of racial and ethnic inequality to persist. Prejudice is not limited to the dominant group; members of subordinate groups often dislike one another. The mass media seem to be of limited value in reducing prejudice and may even intensify ill feeling. Equal-status contact and the shared-coping approach may reduce hostility between groups, but data show few friendships cross racial lines. In response to increasing diversity in the workplace, corporations and organizations have mounted diversity training programs to increase organizational effectiveness and combat prejudice.

Prejudice is a negative attitude toward an entire category of people. The two important components in this definition are attitude and entire category. Prejudice involves attitudes, thoughts, and beliefs, not actions.

A prejudiced belief leads to categorical rejection. Prejudice is not disliking someone you meet because you find his or her behavior objectionable. It is disliking an entire racial or ethnic group, even if you have had little or no contact with that group. A college student who requests a room change after 3 weeks of enduring his or her roommate's sleeping all day, playing loud music all night, and piling garbage on his or her desk is not prejudiced. He or she is displaying prejudice, however, if he or she requests a change on arriving at school and learning from his or her roommate's luggage tags that his or her new roommate is of a different nationality.

Distinguished from Discrimination

Prejudice is related and sometimes confused with discrimination. The sociological perspective of prejudice distinguishes it from discrimination. Prejudice and discrimination are related concepts but are not the same. Prejudice is a belief or attitude; discrimination is action. Discrimination involves behavior that excludes all members of a group from certain rights, opportunities, or privileges. Like prejudice, it must be categorical. If an individual refuses to hire as a typist an Italian American who is illiterate, it is not discrimination. If an individual refuses to hire any Italian Americans because he or she thinks they are incompetent and does not make the effort to see if an applicant is qualified, it is discrimination.

Merton's Typology

Prejudice does not necessarily coincide with discriminatory behavior. Sociologist Robert Merton (1949, 1976), in exploring the relationship between negative attitudes and negative behavior, identified four major categories. The label added to each of Merton's categories may more readily identify the type of individual being described. These are as follows:

Type I: The unprejudiced nondiscriminator: all-weather liberal
Type II: The unprejudiced discriminator: reluctant liberal
Type III: The prejudiced nondiscriminator: timid bigot
Type IV: The prejudiced discriminator: all-weather bigot

Liberals, as the term is employed in Types I and II, are committed to equality among people. The all-weather liberal believes in equality and practices it. Merton was quick to observe that all-weather liberals may be far removed from any real competition with subordinate groups such as African-Americans or women. Furthermore, such people may be content with their own behavior and may do little to change themselves. The reluctant liberal is not even this committed to equality among groups. Social pressure may cause such a person to discriminate. Fear of losing employees may lead a manager to avoid promoting women to supervisory capacities. Equal-opportunity legislation may be the best way to influence the reluctant liberals.

Types III and IV do not believe in equal treatment for racial and ethnic groups, but they vary in their willingness to act. The timid bigot, Type III, will not discriminate if discrimination costs money or reduces profits or if he or she is pressured not to by peers or the government. The all-weather bigot unhesitatingly acts on the prejudiced beliefs that he or she holds.

LaPiere Study

Merton's typology points out that attitudes should not be confused with behavior. People do not always act as they believe. Close to a century ago, Richard LaPiere (1934, 1969) exposed the relationship between racial attitudes and social conduct. From 1930 to 1932, LaPiere traveled throughout the United States with a Chinese couple. Despite an alleged climate of intolerance of Asians, LaPiere observed that the couple was treated courteously at hotels, motels, and restaurants. He was puzzled by the good reception they received; all the conventional attitude surveys showed extreme prejudice by Whites toward the Chinese.

Was it possible that LaPiere had been fortunate during his travels and consistently stopped at places operated by the

tolerant members of the dominant group? To test this possibility, he sent questionnaires asking the very establishments at which they had been served if the owner would "accept members of the Chinese race as guests in your establishment." More than 90% responded no, even though LaPiere's Chinese couple had been treated politely at all the establishments. How can this inconsistency be explained? People who returned questionnaires reflecting prejudice were unwilling to act based on those asserted beliefs: they were timid bigots.

The LaPiere study is not without flaws. First, he had no way of knowing whether the respondent to the questionnaire was the same person who had served him and the Chinese couple. Second, he accompanied the couple, but the questionnaire suggested that the arrival would be unescorted (and in the minds of some, uncontrolled) and perhaps would consist of many Chinese individuals. Third, personnel may have changed between the time of the visit and the mailing of the questionnaire (Deutscher et al., 1993).

Attitudes are not important in small matters, they are important in other ways: Lawmakers legislate and courts may reach decisions based on what the public thinks. This is not just a hypothetical possibility. Legislators in the United States are often persuaded to vote in a certain way by what they perceive as changed attitudes toward immigration, affirmative action, school busing, abortion, national health care, and prayer in public schools.

We examine the theories of why prejudice exists from the sociological perspective, which would fall in to Merton's Types III and IV above and discuss the content and extent of prejudice today.

Theories of Prejudice

Prejudice is learned. Friends, relatives, newspapers, books, movies, and television all teach it. Awareness begins at an early age that there are differences among people that society judges to be important. Several theories have been advanced to explain the rejection by individuals of certain groups in a society.

Exploitation Theory

Racial prejudice is frequently used to justify keeping a group in a subordinate position such as a lower social class. Conflict theorists, in particular, stress the role of racial and ethnic hostility as a way for the dominant group to keep intact its position of status and power. Indeed, this approach maintains that even the less-affluent White working class uses prejudice to minimize competition from upwardly mobile minorities.

This exploitation theory is clearly part of the Marxist tradition in sociological thought. Karl Marx emphasized the exploitation of the lower class as an integral part of capitalism. Similarly, the exploitation or conflict approach explains how racism can stigmatize a group as inferior so that the exploitation of that group can be justified. As developed by Oliver Cox (1942), exploitation theory saw prejudice against Blacks as an extension of the inequality faced by the entire lower class.

Exploitation theory does not necessarily explain prejudice in all its forms. First, not all minority groups are exploited economically to the same extent. Second, many groups that have been the victims of prejudice have not been persecuted for economic reasons – for example, the Quakers. Nevertheless, as Gordon Allport (1979) concludes, the exploitation theory correctly points a finger at one of the factors in prejudice, that is, the rationalized self-interest of the upper classes.

Scapegoating Theory

Scapegoating theory says that prejudiced people believe that they are society's victims. Exploitation theory maintains that intolerant individuals abuse others, whereas scapegoaters feel that they are being abused themselves. The term scapegoat comes from a biblical injunction telling the Hebrews to send a goat into the wilderness to symbolically carry away the people's sins. Similarly, the theory of scapegoating suggests that an individual, rather than accepting guilt for some failure, transfers the responsibility for failure to some vulnerable group. In the major tragic twentieth-century example, Adolf Hitler used the Jews as the scapegoat for all German social and economic ills in the 1930s. This premise led to the passage of laws restricting Jewish life in pre–World War II Germany and eventually escalated into the mass extermination of Europe's Jews. Scapegoating of Jews persists. A national survey in 2009 showed that one out of four people in the United States blame 'the Jews' for the recent financial crisis. *Anti-Semitism* – anti-Jewish prejudice and discrimination – remains a very real phenomenon (Malhotra and Margalit, 2009).

Like exploitation theory, scapegoating theory adds to our understanding of why prejudice exists but does not explain all its facets. For example, scapegoating theory offers little explanation of why a specific group is selected or why frustration is taken out on the real culprit when it is possible. Also, both the exploitation and the scapegoating theories suggest that every individual sharing the same general experiences in society would be equally prejudiced, but that is not the case. Prejudice varies among individuals who would seem to benefit equally from the exploitation of a subordinate group or who have experienced equal frustration.

Normative Approach

Although personality factors are important contributors to prejudice, normative or situational factors must also be given serious consideration. The normative approach takes the view that prejudice is influenced by societal norms and situations that serve to encourage or discourage the tolerance of minorities.

Analysis reveals how societal influences shape a climate for tolerance or intolerance. Societies develop social norms that dictate not only what foods are desirable (or forbidden) but also what racial and ethnic groups are to be favored (or despised). Social forces operate in a society to encourage or discourage tolerance. The force may be widespread – for example, the pressure on White Southerners to oppose racial equality while there was slavery or segregation. The influence of social norms may be limited – for example, one male who finds himself becoming more sexist as he competes with three females for a position in a prestigious law firm.

Social psychologist Thomas Pettigrew (1958, 1959) collected data that substantiated the importance of such social norms in developing a social climate conducive to the expression of prejudice. Pettigrew found that Whites in the South were more anti-Black than Whites in the North, and that Whites in the United States were not as prejudiced as Whites in the Republic of South Africa.

The Content of Prejudice: Stereotypes

An important approach by sociologists and other social scientists to the understanding of prejudice has been the role that stereotypes play. Numerous scientific studies have been made of these exaggerated images.

The systematic study of stereotypes began with David Katz and Kenneth Braly's (1933) use of the *checklist approach*. College students were presented with a list of 84 adjectives such as *sly*, *cruel*, *neat*, and so on. They were asked to list which traits they considered most characteristic of 10 groups: Germans, Italians, Irish, English, Blacks, Jews, Americans, Chinese, Japanese, and Turks. The students' selection of traits for each group consistently agreed with one another, especially for Blacks and Jews. Several other researchers have confirmed this technique.

Scholarship on stereotyping has acknowledged that the stigmatizing of subordinate groups (or out-groups) can take different forms. Social psychologist Susan Fiske and her associates have developed a stereotype content model (SCM) that considers stereotypes along two dimensions. The 'warmth' dimension examines whether the prejudice regards members of other groups as not competitive for scarce resources and therefore warmly or are competitive and therefore with less warmth. The other dimension is 'competence,' which considers whether the object or prejudice is of high social status or not.

Like Merton's approach, this leads to four types. Examples of groups within the combinations of high and low levels of warmth and competence are as follows:

Type I: Competitive, low-status groups: the paternalistic stereotype (for example: the elderly and people with disabilities)

Type II: Noncompetitive, low-status groups: the contemptuous stereotype (for example: the poor and welfare recipients)

Type III: Noncompetitive, high-status groups: the admiration stereotype (for example: other members of one's own group and close allies)

Type IV: Competitive, high-status groups: the envious stereotype (for example: Asians and Jews and people who are wealthy)

This model has allowed a more detailed analysis of the content of prejudice as well as the different forms of expressions such as contempt toward Type II and resentment toward Type IV. Also, stereotypes may be ambivalent as with Type I where one may feel members of the out-group (for example, the elderly) are nice but incompetent (Fiske et al., 2002; Cuddy et al., 2008).

The SCM has held up in understanding prejudice in societies outside the United States. In one study conducted in seven European and three East Asian nations, SCM continued to prove its utility. The one exception was that in the more collectivist cultures of Hong Kong, Japan, and South Korea, there was less willingness to endorse the admiration stereotype (Type III) toward people like themselves even though the respondents did negatively stereotype out-groups (Cuddy et al., 2009).

Labels take on such strong significance that people often ignore facts contradicting their preformed beliefs. People who believe many Italian Americans to be members of the Mafia disregard law-abiding Italian Americans. Muslims are regularly portrayed in a violent, offensive manner that contributes to their being misunderstood and distrusted. Gradually, the mass media – movies, television, newspapers, and periodicals – are presenting a more accurate, evenhanded portrayal of racial, ethnic, and religious groups; however, there is significant room for improvement.

Although explicit expressions of racial and ethnic stereotypes in the media are becoming less common, it is much too soon to write the obituary of racial and ethnic stereotypes. In addition, stereotyping is not limited to racial and ethnic groups. Other groups are subjected to stereotyping. Probably easiest to see in daily life and the mass media is sexism. Sexism is the ideology that one sex is superior to the other. Images and descriptions of women and even girls often reinforce sexism. *Homophobia*, the fear of and prejudice toward homosexuality, is present in every facet of life: the family, organized religion, the workplace, official policies, and the mass media. Like the myths and stereotypes of race and gender, those about homosexuality keep gay men and lesbian women oppressed as a group and may also prevent sympathetic members of the dominant group, the heterosexual community, from supporting them.

In the wake of the 9/11 attacks on the World Trade Center, there has been renewed interest in racial profiling and the degree to which it relies on stereotypes. Profiling by law enforcement officers, including customs officials, airport security, and police, assumes that people fitting certain descriptions are likely to be engaged in something illegal.

Racial profiling persists despite overwhelming evidence that it is not a predictive approach toward identifying potential troublemakers. Whites are more likely to be found with drugs in the areas in which minority group members are disproportionately targeted. A federal study made public in 2005 found little difference nationwide in the likelihood of being stopped by law enforcement officers, but African-Americans were twice as likely to have their vehicles searched, and Latinos were five times more likely. A study of New York City police officers describing some 4.43 million stops between 2004 and mid-2012 found that Blacks and Latinos accounted for 83% of people who were stopped and frisked, and a related study found that Whites were 50% more likely to be carrying weapons (Center for Constitutional Rights, 2011; Goldstein, 2013; Herbert, 2010; Tomaskovic-Devey and Warren, 2009).

Are stereotypes held only by dominant groups about subordinate groups? The answer is clearly no. White Americans even believe generalizations about themselves, although admittedly these are rather positive. Subordinate groups also hold exaggerated images of themselves. Studies

before World War II showed a tendency for Blacks to assign to themselves many of the same negative traits assigned by Whites. Today, African-Americans, Jews, Asians, and other minority groups largely reject stereotypes of themselves, although subordinate groups will to some degree stereotype each other.

The Extent of Prejudice

Interest in developing theories of prejudice or studying its concept has been exceeded only by interest in measuring it. From the outset, efforts to measure prejudice have suffered from disagreement over exactly what constitutes intolerance and whether there is such a phenomenon as no prejudice at all. Add to these uncertainties the methodological problems of attitude measurement, and the empirical study of prejudice becomes an undertaking fraught with difficulty.

The extent of prejudice can be measured only in relative, rather than absolute, differences. We cannot accurately say, for example, that prejudice toward Puerto Ricans is four times greater than that toward Portuguese Americans. We can conclude that prejudice is greater toward one group than toward the other; we just cannot quantify how much greater. The social distance scale is especially appropriate to assess differences in prejudice.

The Social Distance Scale

Robert Park and Ernest Burgess (1921: 440) first defined *social distance* as the tendency to approach or withdraw from a racial group. Emory Bogardus (1968) conceptualized a scale that could empirically measure social distance. His social distance scale is so widely used that it is frequently referred to as the Bogardus scale.

The scale asks people how willing they would be to interact with various racial and ethnic groups in specified social situations. The situations describe different degrees of social contact or social distance. The seven items used, with their corresponding distance scores, follow. People are asked if they would be willing to admit each group:

To close kinship by marriage (1.00)
To my club as personal chums (2.00)
To my street as neighbors (3.00)
To employment in my occupation (4.00)
To citizenship in my country (5.00)
As only visitors to my country (6.00)
Would exclude from my country (7.00)

A score of 1.00 for any group would indicate no social distance and therefore no prejudice. The social distance scale has been administered to many different groups in other countries as well. Despite some minor flaws and certain refinements needed in the scale, the results of these studies are useful and can be compared.

The data from social distance studies over several generations have shown a decided decline in the expression of prejudice even if distinctions remain – that is, people still place groups at different degrees of social distance from themselves. This is a part of a larger trend in sociological research that shows a marked decline in expression of prejudice III in social surveys. However, scholars have cautioned that more tolerance toward direct questions seeking to tap racism may exaggerate the degree of true change in social attitudes (Parrillo and Donoghue, 2013).

Color-Blind Racism

The sociological approaches of exploitation, scapegoating, and normative were developed by measuring people's fairly explicit expressions of prejudice such as, 'I would not vote for a woman for President' or 'I would not want to have Roma or Gypsies living in my community.' Over the last three generations, surveys have consistently shown declining amounts of explicit prejudice being expressed. Yet, casual observations still show deep-seated distrust and intolerance, which often explode in to violence or what have been termed hate crimes.

Color-blind racism refers to the use of race-neutral principles to defend the racially unequal status quo. Yes, 'no discrimination for college admission' should exist, yet the disparity in educational experiences means that formal admissions criteria will privilege White high school graduates. 'Health care is for all,' but if you do not have workplace insurance, you likely cannot afford it.

Color-blind racism has also been referred to as laissez-faire, postracialism, or aversive racism, but the common theme is that notions of racial inferiority are rarely expressed and that proceeding color-blind into the future will perpetuate inequality. In the post-civil rights era and with the election of President Barack Obama, people are more likely to assume discrimination is long past and express views that are more proper – that is, lacking the overt expressions of racism of the past.

An important aspect of color-blind racism is the recognition that race is rarely invoked in public debates on social issues. Instead, people emphasize lower social class, the lack of citizenship, or illegal aliens; these descriptions serve as proxies for race. Furthermore, the emphasis is on individuals failing rather than recognizing patterns of groups being disadvantaged. This leads many White people to declare that they are not racist and that they do not know anyone who is racist. It also leads to the mistaken conclusion that more progress has been made toward racial and ethnic equality and even tolerance than has really taken place.

Economically less-successful groups such as African-Americans and Latinos have been associated with negative traits to the point that issues such as urban decay, homelessness, welfare, and crime are viewed as race issues, even though race is rarely mentioned explicitly. Besides making it harder to resolve difficult social issues, this is another instance of blaming the victim. These perceptions come at a time the willingness of the government to address domestic ills is limited by increasing opposition to new taxes and continuing commitments to fight terrorism here and abroad. The color line remains, even if more people are unwilling to accept its divisive impact on everyone's lives (Ansell, 2008; Bonilla-Silva, 2006; Bonilla-Silva and Seamster, 2011; Kang and Lane, 2010; Winant, 2004: 106–108).

Mood of the Subordinate Group

Sociologists have noted that approaches to prejudice often focus solely on intolerance of the powerful toward out-groups. However, this is an incomplete picture, as one should also consider prejudice directed among out-groups toward each other and even prejudice that people direct inward (Shelton, 2000).

Sociologist William E.B. Du Bois relates an experience from his youth in a largely White community in Massachusetts. He tells how, on one occasion, the boys and girls were exchanging cards, and everyone was having a lot of fun. One White girl, a newcomer, refused his card as soon as she saw that Du Bois was Black. He wrote that he was "shut out from their world by a vast veil" (Du Bois, 1903: 2). In using the image of a 'veil,' Du Bois describes how members of subordinate groups learn they are being treated differently. In his case and that of many others, this leads to feelings of contempt toward all Whites that continue for a lifetime.

Prejudice is as diverse as the nation's population. It exists not only between dominant and subordinate peoples but also among specific subordinate groups. Unfortunately, until recently, there was little research on this subject except for a few social distance scales administered to racial and ethnic minorities. A recent study documents that prejudice among Korean and Dominican immigrants and Black Americans (Roth and Kim, 2013).

Reducing Prejudice

The obvious way to eliminate prejudice is to eliminate its causes: the desire to exploit, the fear of being threatened, and the need to blame others for one's own failure. But how can one on a societal level strive to reduce prejudice?

Much research has been done to determine how to change negative attitudes toward groups of people. The most encouraging findings point to education, mass media, intergroup contact, and workplace training programs.

Education

Most research studies show that well-constructed educational programs have a positive effect on reducing prejudice, at least temporarily. The reduction is rarely as much as one might want, however. The difficulty is that a single program is insufficient to change lifelong habits, especially if little is done to reinforce the program's message once it ends. Persuasion to respect other groups does not operate in a clear field, because, in their ordinary environments, people are still subjected to situations that promote prejudicial feelings (Allport, 1979).

Studies document that increased formal education, regardless of content, is associated with racial tolerance. Research data show that highly educated people are more likely to indicate respect and liking for groups different from themselves. Why should more education have this effect? It might promote a broader outlook and make a person less likely to endorse myths that sustain racial prejudice. Formal education teaches the importance of qualifying statements such as "even though they have lower test scores, you need to remember the neighborhoods from which they come." Education introduces one to the almost indefinite diversity of social groups and the need to question rigid categorizations, if not reject them altogether. Colleges increasingly include a graduation requirement that students complete a course that explores diversity or multiculturalism. Another explanation is that education does not reduce intolerance but instead makes people more careful about revealing it. Formal education may simply instruct people in the appropriate responses. Despite the lack of a clear-cut explanation, either theory suggests that the continued trend toward a better-educated population will contribute to a reduction in overt prejudice (Schaefer, 1986, 1996).

Mass Media

Mass media, like schools, may reduce prejudice without requiring specially designed programs. Television, radio, motion pictures, newspapers, magazines, and the Internet present only a portion of real life, but what effect do they have on prejudice if the content is racist or antiracist, sexist or antisexist? As with measuring the influence of programs designed to reduce prejudice, coming to strong conclusions on mass media's effect is hazardous, but the evidence points to a measurable effect.

The absence of racial and ethnic minorities in television and motion pictures is well documented. They are less likely to play recurring roles and are far underrepresented in key decision-making positions such as directors, producers, and casting agents. Television series are only part of the picture. Whites do news broadcasting predominantly, and local news emphasizes crime, often featuring Black or Hispanic perpetrators; print journalism is nearly the same (Media Matters for America, 2013; Writers Guild of America West, 2013).

Equal-Status Contact

An impressive number of research studies have confirmed the *contact hypothesis*, which states that intergroup contact between people of equal status in harmonious circumstances causes them to become less prejudiced and to abandon previously held stereotypes.

The key factor in reducing hostility, in addition to equal-status contact, is the presence of a common goal. If people are in competition, as already noted, contact may heighten tension. However, bringing people together to share a common task has been shown to reduce ill feelings when these people belong to different racial, ethnic, or religious groups (Fine, 2008; Paluck and Green, 2009).

The limited amount of intergroup contact because of segregated neighborhoods, schools, and workplaces, is of concern given the power of the contact hypothesis. If there is no positive contact, then how can we expect a decrease in prejudice? National surveys show prejudice directed toward Muslim Americans, but social contact bridges that hatred. In a 2006 survey, 50% of people who were not acquainted with a Muslim favored special identification for Muslim Americans, but only 24% of those who knew a Muslim embraced that same view.

Similarly, people personally familiar with Muslims are more than one-third less likely to endorse special security checks just for Muslims and are less nervous to see Muslim men on the same flight with themselves. Although negative views are common toward Muslim Americans today, they are much less likely to be endorsed by people who have had intergroup contact (Saad, 2006).

As African-Americans and other subordinate groups slowly gain access to better-paying and more-responsible jobs, the contact hypothesis takes on greater significance. Usually, the availability of equal-status interaction is taken for granted; yet, in everyday life, intergroup contact does not conform to the equal-status idea of the contact hypothesis. Furthermore, as we have seen, in a highly segregated society such as the United States, contact tends to be brief and superficial, especially between Whites and minorities. The apartheid-like friendship patterns prevent us from learning firsthand not just how to get along but also how to revel in interracial experiences (Bonilla-Silva and Embrick, 2007; Miller, 2002).

Corporate Response: Diversity Training

Prejudice carries a cost. This cost is not only to the victim but also to any organization that allows prejudice to interfere with its functioning. Workplace hostility can lead to lost productivity and even staff attrition. Furthermore, if left unchecked, an organization – whether a corporation, government agency, or nonprofit enterprise – can develop a reputation for having a 'chilly climate.' If a business has a reputation that it is unfriendly to people of color or to women, qualified people are discouraged from applying for jobs there and potential clients might seeking products or services elsewhere.

It is difficult to make broad generalization about the effectiveness of diversity training programs because they vary so much in structure between organizations. At one extreme are short presentations that seem to have little support from management. People file into the room feeling it is something they need to get through quickly. Such training is unlikely to be effective and may be counterproductive by heightening social tensions. At the other end of the continuum is a diversity training program that is integrated into initial job training, reinforced periodically, and presented as part of the overall mission of the organization, with full support from all levels of management. In such businesses, diversity is a core value, and management demands a high degree of commitment from all employees.

Remarkably, the prevalence of any diversity programs in organizations remains slow (10–30%), even after more than 30 years since the diversity-management paradigm was first widely viewed as good for business. Even inexpensive steps are not widely adopted. Unfortunately, corporations with lower representation of women and minorities are less likely to embrace diversity programs.

Research into different corporate policies has found two that are particularly effective. Diversity task forces that bring together people from different departments to brainstorm about opening up hiring opportunities appear to eventually increase the diversity in upper management. A second successful policy is the diversity mentoring programs designed for aspiring women and minorities, as well as White men, to achieve their career goals. Unfortunately, research suggests that if White men perceive African-Americans are the primary organizers of such efforts, networking progress can actually have a negative impact (Dobbin et al., 2011).

Overview

Several theories by sociologists have tried to explain why prejudice exists. Some emphasize economic concerns (the exploitation and scapegoating theories), whereas other approaches stress normative factors. No one explanation is sufficient. Surveys conducted in the United States over the past 60 years point to a reduction of prejudice as measured by the willingness to express stereotypes or maintain social distance. Survey data also show that many Whites and Blacks are still intolerant of each other. Prejudice involving Hispanic groups, Asian-Americans, and relatively large recent immigrant groups such as Arab Americans and Muslim Americans is well documented. Issues such as immigration and affirmative action reemerge and cause bitter resentment. Furthermore, ill feelings exist among subordinate groups in schools, in the streets, and in the workplace.

Equal-status contact may reduce hostility among groups. However, in a highly segregated society defined by inequality, such opportunities are not typical. The mass media and workplace diversity programs can be of value in reducing discrimination but have not done enough and may even intensify ill feeling by promoting stereotypical images. While strides are being made in increasing the appearance of minorities in positive roles in television and films, one would not realize how diverse our society is by sampling advertisements, programs, or movie theaters.

See also: Affirmative Action, Sociology of; Discrimination: Racial; Prejudice in Society: Psychological Perspectives; Racial Relations; Racism, Sociology of.

Bibliography

Allport, G.W., 1979. The Nature of Prejudice. 25th Anniversary. Addison-Wesley, Reading, MA.

Ansell, A.E., 2008. Color blindness. In: Schaefer, R.T. (Ed.), Encyclopedia of Race, Ethnicity, and Society, vol. 1. Sage, Thousand Oaks, CA, pp. 320–322.

Bogardus, E., 1968. Comparing racial distance in Ethiopia, South Africa, and the United States. Sociology and Social Research 52, 149–156.

Bonilla-Silva, E., 2006. Racism without Racists, second ed. Rowman & Littlefield, Lanham, MD.

Bonilla-Silva, E., Embrick, D.G., 2007. "Every place has a Ghetto…" the significance of Whites' social and residential segregation. Symbolic Interaction 30, 323–345.

Bonilla-Silva, E., Seamster, L., 2011. The sweet enchantment of color blindness in Black face: explaining the 'miracle,' debating the politics, and suggesting a way for hope to be 'for real' in America. Political Power and Social Theory 22, 139–175.

Center for Constitutional Rights, 2011. Stop-and-Frisks of New Yorkers in 2010 Hit All-Time High at 600, 601; 87 Percent of Those Stopped Black and Latino. Available at: http://ccrjustice.org (accessed 02.03.11.).

Cox, O.C., 1942. The Modern Caste School of Social Relations. Social Forces (December) 21, 218–226.

Cuddy, A.J.C., Fiske, S.T., Glick, P., 2008. Warmth and competence as universal dimension of social perception: the stereotype content model and the BIAS map. Advances in Experimental Social Psychology 40, 61–149.

Cuddy, A.J.C., et al., 2009. Stereotype content model across cultures: toward universal similaritites and some differences. British Journal of Social Psychology 48, 1–33.
Deutscher, I., Pestello, F.P., Pestello, H.F., 1993. Sentiments and Acts. Aldine de Gruyter, New York.
Dobbin, F., Kim, S., Kalev, A., 2011. You can't always get what you need: organizational determinants of diversity programs. American Sociological Review 76, 386.
Du Bois, W.E.B., 1903. The Souls of Black Folks: Essays and Sketches (reprint). Facade Publications, New York, 1961.
Fine, G., 2008. Robber's cave. In: Schaefer, R.T. (Ed.), Encyclopedia of Race, Ethnicity, and Society, vol. 3. Sage, Thousand Oaks, CA, pp. 1163–1164.
Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, Jun, 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82 (6), 878–902.
Goldstein, J., 2013. Judge rejects New York's Stop-and-Frisk policy. New York Times, A1–A16 (12 August).
Herbert, B., 2 February 2010. Jim Crow policy. New York Times. A27.
Kang, J., Lane, K., 2010. Seeing through colorblindness: implicit bias and the law. UCLA Law Review 58, 465–520.
Katz, D., Braly, K., 1933. Racial stereotypes of one hundred college students. Journal of Abnormal & Social Psychology 28, 280–290.
LaPiere, R.T., 1934. Attitudes vs actions. Social Forces, 230–237.
LaPiere, R.T., 1969. Comment of Irwin Deutscher's looking backward. American Sociologist 4, 41–42.
Malhotra, N., Margalit, Y., 2009. State of the nation: anti-semitism and the economic crisis. Boston Review (May/June). Accessible at: http://bostonreview.net/BR34.3/malhotra_margalit.php.
Media Matters for America, 2013. Diversity in Evening Cable News in 13 Charts (13 May). Accessible at: mediamatters.org.
Merton, R.K., 1949. Discrimination and the American creed. In: MacIver, Robert M. (Ed.), Discrimination and National Welfare. Harper & Row, New York, pp. 99–126.
Merton, R.K., 1976. Sociological Ambivalence and Other Essays. Free Press, New York.
Miller, N., 2002. Personalization and the promise of contact theory. Journal of Social Issues 58, 387–410.
Paluck, E.L., Green, D.P., 2009. Prejudice reduction: what works? A review and assessment of research and practice. Annual Review of Psychology 60, 339–367.
Park, R.E., Burgess, E.W., 1921. Introduction to the Science of Sociology. University of Chicago Press, Chicago.
Parrillo, V., Donoghue, C., 2013. The national social distance study: ten years later. Sociological Forum 28, 597–614.
Pettigrow, T.F., 1958. Personality and socio-cultural factors in intergroup attitudes: a cross-national comparison. Journal of Conflict Resolution 2, 29–42.
Pettigrow, T.F., 1959. Regional differences in anti-Negro prejudice. Journal of Abnormal and Social Psychology 59, 28–36.
Roth, W.D., Kim, N.Y., 2013. Relocating prejudice: a transnational approach to understanding immigrants' racial attitudes. International Migration Review 47, 330.
Saad, L., 2006. Anti-Muslim sentiments fairly commonplace. The Gallup Poll (10 August).
Schaefer, R.T., 1986. Racial prejudice in a capitalist state: what has happened to the American creed? Phylon 47, 192–198.
Schaefer, R.T., 1996. Education and prejudice: unraveling the relationship. The Sociological Quarterly 37, 1–16.
Shelton, J.N., 2000. A reconceptualization of how we study issues of racial prejudice. Personality and Social Psychology Review 4, 374–390.
Sherif, M., Sherif, C., 1969. Social Psychology. Harper & Row, New York.
Tomaskovic-Devey, D., Warren, P., 2009. Explaining and eliminating racial profiling. Contexts 8, 34–39.
Winant, H., 2004. The New Politics of Race: Globalism, Difference, Justice. University of Minnesota Press, Minneapolis.
Writers Guild of America West, 2013. Diversity on TV Writing Staffs: Writers Guild Releases Latest Research Findings (26 March). Accessible at: http://www.wga.org.

Relevant Websites

http://www.adl.org/ – Anti-Semitism.
http://www.fbi.gov/about-us/investigate/civilrights/hate_crimes – Hate Crimes.
http://www.teachersagainstprejudice.org/index2.php – Teaching against Prejudice.
http://www.splcenter.org/what-we-do/teaching-tolerance – Teaching Tolerance.

Premotor Cortex

Giacomo Rizzolatti, Università di Parma, Parma, Italy; and IIT Brain Center for Social and Motor Cognition, Parma, Italy
Giuseppe Luppino, Università di Parma, Parma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The frontal lobe of primates consists of two large regions, a rostral region involved in cognitive functions (prefrontal lobe) and a caudal region mostly devoted to motor functions (agranular frontal cortex). This last region is formed, posteriorly, by the primary motor cortex and, anteriorly, by a mosaic of distinct areas called premotor areas. Premotor areas can be subdivided into two main classes: caudal premotor areas and rostral premotor areas. The caudal premotor areas elaborate sensory information for action. The rostral premotor areas appear to play a control role on the caudal areas, relaying to them information from the prefrontal and cingulate cortex on individuals' motivation and long-term plans. In this way, potential motor actions formed in the caudal premotor areas can be executed or not according to the individual's needs.

The frontal lobe of primates is formed by two large regions: a rostral region involved in cognitive functions, called prefrontal lobe (*see* Prefrontal Cortex Development and Development of Cognitive Function), and a caudal one mostly devoted to motor functions. Histologically, the caudal region is characterized by the absence of granular cells. It is also referred to as the agranular frontal cortex.

The agranular frontal cortex consists of two main sectors: area 4 (the primary motor cortex) and area 6. Area 6 is neither anatomically nor functionally unitary, but is formed by a *mosaic* of distinct areas (see Picard and Strick, 2001; Rizzolatti and Luppino, 2001). In the present article, we will review the general organization of the premotor areas in monkeys and discuss in detail the operations of one of them. A short comparison between the monkey and human premotor cortex will be made in the last section.

Classes of Premotor Areas and Their Cortical Connections

The agranular frontal cortex is formed by seven distinct areas (areas F1–F7, see **Figure 1**). Area F1 corresponds to the Brodmann area 4 (the 'primary motor cortex') (*see* Motor Cortex);

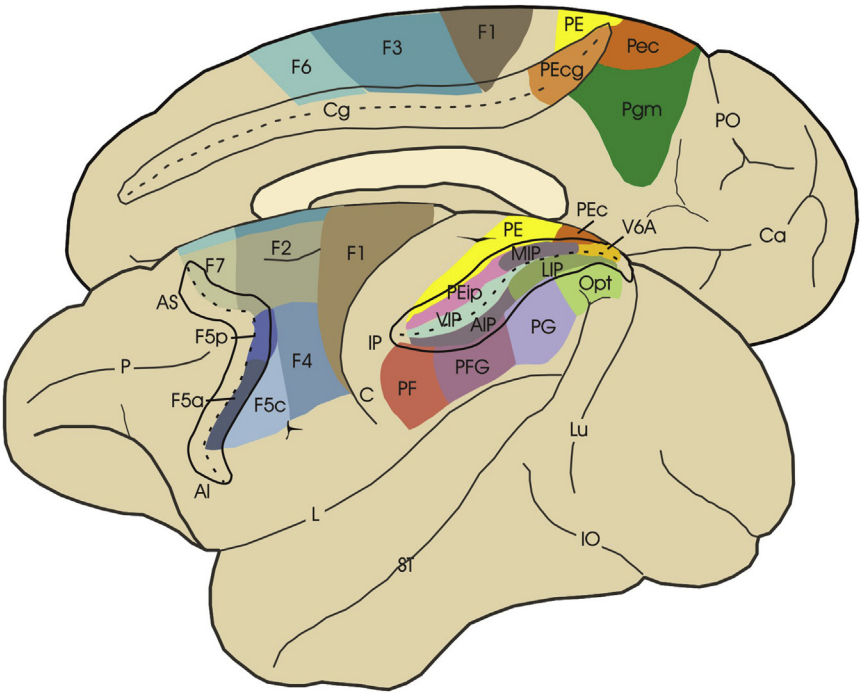

Figure 1 Mesial (top) and lateral (bottom) views of the macaque brain showing the subdivision of the agranular frontal and posterior parietal cortex. The arcuate, cingulate, and intraparietal sulci are drawn unfolded to show the areas buried within their banks. For the nomenclatures and definition of the areas, see text. Abbreviations: AI = inferior arcuate sulcus; AS = superior arcuate sulcus; C = central sulcus; Ca = calcarine fissure; Cg = cingulate sulcus; IO = inferior occipital sulcus; IP = intraparietal sulcus; L = lateral fissure; Lu = lunate sulcus; P = principal sulcus; PO = parieto-occipital sulcus; ST = superior temporal sulcus.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55046-6

the other areas are located within the Brodmann area 6. They are referred to as 'premotor areas.' Recently, one of them – area F5 – has been found to be formed by three distinct sectors: F5a, F5c, and F5p (**Figure 1**).

Each premotor area has a specific pattern of anatomical connections. According to the characteristics of to this pattern, the premotor areas can be subdivided into two main classes: the caudal and the rostral premotor areas (Rizzolatti and Luppino, 2001). The caudal premotor areas are *F2*, *F3*, *F4*, *F5p*, and *F5c*, while the rostral ones are *F5a*, *F6*, and *F7*.

Connections with the Spinal Cord and Intrinsic Motor Connections

The organization of the corticospinal and corticobulbar projections from the premotor areas is consistent with their subdivision into two classes. The corticospinal projections originate, besides the primary motor area, from those sectors of the *caudal* premotor areas where forelimb, trunk, or hind-limb movements are represented (see He et al., 1993). The sectors of these areas where face/mouth movements are represented are the source of corticobulbar projections targeting the facial nucleus (see Morecraft et al., 2001). It is interesting to note that F2, F3, F4, F5p, and F5c display somatotopically organized connections with F1, as well as, with each other. These data indicate that although F1 is the only area with direct access to the spinal cord motor neurons, the caudal premotor areas have also access to the spinal cord. Thus, these areas can be involved in the generation and control of movements both trough F1 or in parallel with its output.

The organization of the rostral premotor areas is radically different. These areas do not project to the spinal cord, their descending input terminating in various parts of the brain stem. Furthermore, the rostral premotor areas are not connected with F1 and, especially in the case of F6, have widespread connections with other motor areas. An exception is the dorso-rostral part of F7, corresponding to the supplementary eye field (SEF), which has very poor connections with other motor areas, but is strongly connected with the frontal eye field (FEF). These data indicate that the rostral premotor areas are involved in the generation of motor behavior only indirectly, through their subcortical relays or their connections with the caudal premotor areas.

Connections with Parietal, Prefrontal, and Cingulate Areas

The connections with the *parietal cortex* are the major source of input to the agranular frontal cortex and especially to caudal premotor areas and area F1. Similar to the premotor region the posterior parietal cortex is also formed by a mosaic of distinct areas (**Figure 1**). Each of them deals with specific aspects of sensory information and with the control of specific effectors. Note that the classical distinction between inferior parietal lobule (IPL), related to both visual and somatosensory information, and superior parietal lobule (SPL), exclusively related to the somatosensory modalities, is incorrect. Both IPL and SPL receive somatosensory and visual input.

Studies on the organization of the parietofrontal connections indicate that premotor and parietal areas are connected with each other in a very specific way (Rizzolatti et al., 1998). Accordingly, it is possible to identify a series of anatomical circuits formed by parietal and motor areas linked one to another by predominant connections. In general, these circuits connect SPL with dorsal and mesial premotor areas and IPL with ventral premotor areas. Functional evidence, when available, indicates that parietal and frontal areas forming each of these circuits share common functional properties. Therefore, the functional correlate of this anatomical organization is that each of these circuits is involved into particular aspects of sensory motor transformations. The essence of this operation is the transformation of a description of the stimuli in sensory terms into their description in motor terms. Thus, these parietofrontal circuits, and not the frontal motor areas in isolation, should be considered the functional units of the cortical motor system (Rizzolatti et al. 1998).

Prefrontal projections to the motor cortex are primarily directed to rostral premotor areas (Rizzolatti and Luppino, 2001; Gerbella et al., 2011). Prefrontal input to F7 (except for the SEF) originates from the dorsal part of the lateral prefrontal cortex (DLPF), that to F6 and the SEF originates from both DLPF and the ventral (VLPF) part of the lateral prefrontal cortex, and, finally, that to F5a originates only from the VLPF. In addition, F6 and the SEF are also the targets of strong afferents originating from the rostral cingulate cortex (area 24c). Cingulate connections to F7 and F5a are considerably weaker, while F5a is densely connected with rostral opercular frontal areas.

Our knowledge of the organization of the prefrontal and cingulate cortex is much less detailed than that of the parietal cortex. It is, however, widely accepted that these regions are not involved in the mere analysis of sensory stimuli, but rather play a role in 'higher order' functions such as working memory, temporal planning of actions, and motivation. These data indicate that caudal and rostral premotor areas play a different functional role in motor control. Caudal premotor areas, together with parietal areas, are involved in transformation of sensory information into potential motor actions. Rostral premotor areas appear to play a hierarchically higher role. These areas receive information concerning agent's goals, motivation, and memorized spatial and nonspatial information from the prefrontal, cingulate, and opercular frontal cortex. This information is then conveyed to the caudal premotor areas determining when and which potential actions, coded in these areas, might become actual actions.

Functional Organization of the Premotor Areas

To illustrate the information processing carried out in the premotor cortex, we will describe the activity of an area (area F5) in detail. As far as the other premotor areas are concerned only their fundamental properties will be reviewed.

Area F5

Area F5 occupies the rostroventral part of the agranular frontal cortex (**Figure 1**). Its electrical stimulation evokes hand and mouth movements. The firing of neurons in F5 correlates with specific motor acts (e.g., grasping) or segments of them (e.g., hand opening, hand closure during grasping) rather than with

the individual movements that form a motor act. Grasping, holding, tearing, and manipulating are the motor acts most represented.

F5 was originally described as a single entity. Recent multiarchitectonic data showed that F5 is not homogeneous but formed by three discrete sectors (Belmalih et al., 2009). One of them, designated as F5 'convexity' (F5c), is located on the convexity of the postarcuate cortex adjacent to the inferior arcuate sulcus. The other two, dubbed as 'posterior' and 'anterior' F5 (F5p and F5a, respectively), lie within the postarcuate bank at different anteroposterior levels (**Figure 1**).

F5c is a motor field related to hand, face, and mouth actions. A substantial proportion (about 20%) of its neurons form a particular class of visuomotor neurons, called *mirror neurons* (see Rizzolatti and Craighero, 2004). The distinctive feature of these neurons is that they discharge *both* when the monkey *observes* another individual (a human being or another monkey) performing a hand or mouth action and when it *executes* the same or a similar action. Mirror neurons do not discharge in response to the simple presentation of food or other objects of interest to the monkey.

The fundamental function of mirror neurons of F5 is that of transforming observed actions, done by others, into corresponding motor representations in the brain of the observer. By eliciting these motor representations, which are identical to those that the observer uses to produce voluntary movements, mirror neurons allow the observer to understand what the agent is doing.

It may sound bizarre that in order to understand an action done by another, one should activate the motor system. Actually, this is not so strange. A mere visual perception, without involvement the motor system only provides a description of the visible aspects of the movements. It does not give information on what really implies doing it, on the links between the observed actions and other actions related to them or on how to imitate it. This can be achieved only if the observed action is mapped onto the motor system of the observer. Thus, the activation of the mirror neurons is essential to provide the observer with an 'experiential' comprehension of the observed action.

There is evidence that, especially in humans, besides action understanding the mirror mechanism is involved also in other cognitive functions like imitation.

F5c is connected with area SII and areas AIP and PFG located in the IPL. These last areas are, in turn, richly connected with sectors of the superior temporal sulcus (STS) known to code complex motor acts (Perrett et al., 1989). Mirror neurons have been found in both areas PFG and AIP, while neurons of STS appear to be devoid of motor properties.

Other studies have evidenced that neurons of F5c are involved in the evaluation and comparison of current and remembered sensory information for perceptual decisions (Lemus et al., 2009) and in the evaluation of the outcomes of these decisions for organizing future behavior.

F5p is essentially a hand motor field, encoding specific goal-directed motor acts. A proportion of F5p neurons are responsive to the observation of 3D objects ('canonical' neurons, see Rizzolatti and Luppino, 2001). These neurons are involved in the selection of hand motor acts appropriate for object interactions based on the object physical characteristics (Jeannerod et al., 1995). F5p is tightly connected with the hand field of the primary motor area. These connections enable F5p to generate object-oriented hand actions through the modulation of motor output of primary motor cortex (Prabhu et al., 2009). F5p also project to the brain stem and to the cervical spinal cord. This suggests a direct role of this area in the execution of hand motor acts.

As F5c, F5p is densely connected with areas AIP, PFG, and SII. Area AIP is a hand-related field located in the rostral part of the lateral bank of the intraparietal sulcus. It contains a large number of neurons that discharge during the execution of specific grasping movements.

Sakata et al. (1995) subdivided the neurons of AIP into three groups: 'motor-dominant,' 'visual and motor,' and 'visual-dominant' neurons. About half of 'visual and motor' and 'visual-dominant' neurons discharge not only during grasping execution, but also when objects are simply fixated. In these neurons, the visual responses appear to reflect the intrinsic properties of the observed object. The other half of 'visual and motor' and 'visual-dominant' neurons is not active during the passive fixation of objects. These neurons fire in association with the observation of hand or hand/object interactions. Some of these neurons are mirror neurons.

AIP receives information on physical properties of the objects from the dorsal visual stream. Recently, it has been shown that AIP is also connected with the ventral bank of STS and the inferotemporal cortex (Borra et al., 2008; Nelissen et al., 2011). This last connection suggests that AIP has information on object semantics. Thus, AIP and F5p appear involved in visuomotor transformation for grasping based not only on the physical characteristics of the object but also on their meaning.

The third area heavily connected with F5c and F5p is area SII. This area, located in the parietal operculum, is a somatosensory area of higher order involved in tactile recognition of objects and in coding tactile expectancies about impending contact with objects. SII may provide F5 with tactile feedback or memory-based somatosensory expectations, instrumental for selection of grasping motor programs and/or the control of fingertip forces for grasping stability.

F5a has basically agranular structure, but some features proper of the granular frontal cortex. Among them, the presence, although moderate, of granular cells, the rather homogeneous and relatively prominent layer V, and the relatively low myelin content. F5a has some parietal connections similar to those of the other F5 sectors, but also show some features absent in those sectors. More specifically, F5a is the only sector of area F5 connected with ventrolateral prefrontal areas (area 12, and ventral 46). Furthermore, it displays connections with rostral opercular frontal sectors much stronger that those of the other F5 sectors.

Functionally, F5a has been less studied than the other sectors of area F5. However, on the basis of its connections, it is plausible that F5 is a site of integration of sensory motor signals originating from the parietal cortex with higher order information originating from prefrontal and frontal opercular areas. These last inputs could give F5p information on objects shapes, texture, and weight related to past experience as well as

information related to agent's intention. This integration could be instrumental for the generation and control of actions.

Caudal Premotor Areas (F2, F3, F4)

Area F2 forms the caudal part of the dorsal premotor cortex. Area F2 is electrically excitable and has a somatotopic organization with a leg and an arm representation located dorsal and ventral to the superior precentral dimple, respectively (Rizzolatti and Luppino, 2001). Functional studies of F2 have been mostly focused on its motor properties, by employing visually instructed motor delay tasks. Some F2 neurons discharge in association with movement onset (movement-related activity), others become active during the delay period when the monkey is waiting for the go signal (set-related activity), others discharge at the appearance of the instruction stimulus (signal-related activity). These data indicate that F2 is not simply involved in motor execution, but plays a role also in some aspects of motor preparation (Wise et al., 1997). Furthermore, F2 neurons appear to be involved in coding action instructions associated with a symbolic cue or retrieving motor instructions at an abstract level even when the location of actual motor targets is not yet available. Finally, there is evidence that F2 neurons activity can reflect covert mental rehearsal of arm reaching movements elicited by the mere observation of their sensory outcomes (see Rizzolatti and Kalaska, 2013).

F2 contains neurons predominantly responding to somatosensory and in particular proprioceptive stimuli. It receives predominant projections from areas PEc and PEip. However, the ventro-rostral part of F2 hosts neurons responding to visual stimulation and is predominantly connected with areas MIP and V6A.

All the various parietal areas connected to F2 host neurons active during the execution of arm-reaching movements. They are considered to be part of a parietofrontal system crucially involved in the sensory guidance of arm reaching movements. There is evidence that in MIP and V6A there are neurons responsive to both visual and somatosensory stimuli. V6A also contains neurons related to the execution of wrist and hand movements, suggesting a role of V6A-F2 (ventro-rostral part) circuit in the control of reaching to grasp movements.

Area F3 (or SMA proper) is located on the mesial surface of the agranular frontal cortex (**Figure 1**). Its main input originates from somatosensory parietal areas (mostly mesial area PE). F3 contains a complete representation of body movements, with a prevalence of proximal, arm and leg, and axial movements. Electrical stimulation of F3 frequently evokes complex movements, involving both proximal and distal joints. Single joint movements are rarely evoked (Luppino et al., 1991). Neurons of F3 fire in association with active movements.

It has been proposed that F3 encodes limb position in order to set the posture for the incoming movements. According to this view, F3 plays a fundamental role in *predictive* postural adjustments. Some neurons of F3 were found to fire in relation to specific sequences of movements. It has been, therefore, suggested that F3 is involved in sequential organization of learned multiple movements (Tanji, 2001).

Area F4 forms the caudal part of PMv. Its electrical stimulation elicits arm, neck, face, and mouth movements (Gentilucci et al., 1988). Head avoidance, as in the presence of threatening stimuli, has been also described (Graziano, 2006). Neurons of F4 become active during the execution of goal-directed motor acts, such as reaching for food or bringing it to the mouth, others discharge during oro-facial movements. Based on their sensory responses, F4 neurons can be subdivided into two categories: unimodal (mostly tactile) neurons and bimodal (visual and tactile) neurons (Fogassi et al., 1996). In both categories, tactile receptive fields (RFs) are rather large and predominantly located on the face, arm, and the upper part of the body. Bimodal neurons have three-dimensional visual RFs in register with the tactile fields and limited to the space around the body (peripersonal or reaching space). In most of these neurons, the position of the visual RFs does not change when the gaze or the body part, where the tactile RFs is located, moves. These neurons play a fundamental role in transforming the position of objects located in the peripersonal space into appropriate motor acts directed toward or away those objects (Fogassi et al. 1996).

F4 is tightly connected with areas PF and VIP of IPL (**Figure 1**). Area PF is a possible source of somatosensory input related to the face and the mouth (Rozzi et al., 2008). Area VIP is the target of dorsal visual stream areas involved in the analysis of motion perception (e.g., MT, MST, FST) and of somatosensory areas (see Colby, 1998; Rizzolatti et al. 1998). This area hosts visual neurons and bimodal (visual and tactile) neurons displaying functional properties similar to those of bimodal F4 neurons (see Colby, 1998). Area VIP represents, therefore, an initial step in the elaboration of visual information for building the peripersonal space (Rizzolatti et al., 1997, 1998; Colby, 1998).

Rostral Premotor Areas (F6, F7)

Area F6 (or pre-SMA) is located rostral to F3 on the mesial sector of the agranular frontal cortex. It corresponds to the anterior part of the classical SMA described by Woolsey and Penfield. Motor responses can be evoked from it only with rather high current intensities. Typically, these responses consist of slow, rather complex arm movements (Luppino et al. 1991). Visual responses are common in F6, while somatosensory responses are rare. F6 neurons, recorded during active arm movements, show frequently a long leading activity preceding movements (Tanji, 1994).

Studies carried out in naturalistic conditions, where monkeys were allowed to interact with objects presented close to or far from them, showed that F6 neurons exert a global control on arm actions, rather than a control limited on specific, distal (like F5), or proximal (like F4) motor acts. Some neurons of F6 discharge at the presentation of graspable objects independent of their location, increasing their firing as the stimulus moves toward the monkey. Others, on the contrary, are inhibited at the presentation of the graspable objects, but start to discharge as soon as an object is brought close to the monkey. In short, there are several facilitation/inhibition combinations that characterize different F6 neurons. What appears to be constant is the modulation of the neuron discharge according to whether an object can or cannot be grasped (see Rizzolatti et al., 1998).

Recent data showed that F6 plays an important role in the early stages of motor learning, precisely when the cognitive

aspect of the procedure to be learned is acquired. This conclusion is based on the behavior of single neurons and on inactivation experiments showing that the number of errors for learning new sequences increases after muscimol (a selective $GABA_A$ agonist) injections in F6. In contrast, the performance of already learned sequences is not affected by F6 (Hikosaka et al., 1999).

Summing up, F6 appears to play a double role in motor control. First, it controls the activity of the posterior premotor areas. Normally, even when the posterior areas are activated, movement does not start. Only when external contingencies and motivational factors allow it, the movement occurs. F6 is the area that renders the movement onset possible. A second role of F6 is more cognitive. It consists in the control of visuomotor associations underlying the declarative phase of motor learning.

Area F7 forms the rostral part of the PMd. Functionally area F7 is not homogeneous. The dorsal part of it is not related to motor control of body parts, but contains an oculomotor field, called the 'SEF.' Anatomically, the SEF is strongly connected with the FEF, but is also the target of strong projections from both dorsal and ventral areas of the lateral prefrontal cortex. Furthermore, the SEF receives afferents from the rostral cingulate cortex. These anatomical data suggest that the SEF could be involved in the control of eye movements at a higher hierarchical level than the FEF. In line with this hypothesis are functional data showing that the SEF is involved in encoding saccades relative to an object-centered frame of reference (Olson and Gettner, 1995) as well as in monitoring the consequences of eye movements (Stuphorn et al., 2000).

The remaining part of F7 has been the target of relatively few functional studies. Some F7 neurons show visual responses even when the stimulus is not instructing a subsequent movement. Others show visual responses when the location of the stimulus matches the target of the arm movement (Vaadia et al., 1986).

F7 receives strong projections from dorsal area 46 and area 8b of the prefrontal cortex (Luppino et al., 1993), i.e., from areas involved in the working memory of spatial positions. Minor inputs to F7 originate from parietal areas PGm and V6A (Matelli et al., 1998). These data could be at the basis of the possible functional role of F7-non SEF in coding locations of sensory stimuli, or arbitrary sensory information, for the guidance of arm movements. This role is suggested by results showing that neurons in this area are related to eye and arm movements (Fujii et al., 2000), are involved in coding spatial localization of external stimuli for reaching movements (Vaadia et al., 1986), and have visual responses, even when the stimulus is not instructing a subsequent movement (di Pellegrino and Wise, 1991). Lesions including F7-non SEF and F2vr (both targets of DLPFd) severely affect the performance of the animal in previously learned motor association tasks and prevent new learning (see Passingham, 1993).

The Premotor Cortex of Humans

As in the monkey, the premotor cortex of humans is formed by a series of distinct areas, some located on the mesial wall of the frontal lobe, others on its convexity.

The large supplementary motor cortex described by Penfield and Woolsey has been found to be formed by two anatomically and functionally distinct areas: area SMA proper and pre-SMA. Functionally, they are considered to be the homologues of monkey area F3 and area F6, respectively. The fMRI data indicate that in both monkeys and humans, the border between the two medial areas is located at the stereotaxic level AP = 0.

The dorsolateral part of the human premotor cortex is formed by two major sectors: the dorsal (PMd) and ventral (PMv) premotor cortex. Unlike in the monkey, an anatomical marker separating PMd from PMv is absent in humans. However, on the basis of embryological data and on location of the FEF it has been suggested that the border between human PMd and PMv should lie approximately at Z level 50 in Talairach coordinates (Rizzolatti et al., 2002). This border has been recently confirmed by diffuse tensor imaging studies (Tomassini et al., 2007).

There is evidence for a rostro-caudal heterogeneity of PMd that most likely parallel the monkey subdivision. In fact, fMRI studies showed that movement tasks activate the caudal sector of PMd, while more cognitively demanding tasks tends to activate its rostral sector.

A large number of brain imaging studies showed that PMv contains the mirror mechanism. Mirror responses were found in PMv but they also to extend rostrally in the posterior part of the inferior frontal gyrus. Some studies reported mirror responses also in PMd.

Although, it is very likely that human PMv, as that of the monkey, is formed by different areas, a clear, consistent parcellation of human PMv has not yet been established. Similarly, the issue of what might be the monkey homologue of area 44 is matter of debate.

In conclusion, the comparison of human and monkey premotor regions clearly indicate that the basic functional organization of this region has been maintained in evolution. The most dramatic modifications have been those due to the appearance of speech and related evolution of areas mediating this function.

Acknowledgments

This article was supported by EU grant 'Cogsystem' to GR and by Compagnia di San Paolo (Neuroscience Program).

See also: Cortical Areas Engaged in Movement: Neuroimaging Methods; Motor Cortex; Parietal Lobe; Visual Streams: Dorsal and Ventral.

Bibliography

Belmalih, A., Borra, E., Contini, M., Gerbella, M., Rozzi, S., Luppino, G., 2009. Multimodal architectonic subdivision of the rostral part (area F5) of the macaque ventral premotor cortex. Journal of Comparative Neurology 512, 183–217.

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., Luppino, G., 2008. Cortical connections of the macaque anterior intraparietal (AIP) area. Cerebral Cortex 18, 1094–1111.

Colby, C.L., 1998. Action-oriented spatial reference frames in cortex. Neuron 20, 15–24.

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., Rizzolatti, G., 1996. Coding of peripersonal space in inferior premotor cortex (area F4). Journal of Neurophysiology 76, 141–157.

Fujii, N., Mushiake, H., Tanji, J., 2000. Rostrocaudal distinction of the dorsal premotor area based on oculomotor involvement. Journal of Neurophysiology 83, 1764–1769.

Gentilucci, M., Fogassi, L., Luppino, G., Matelli, M., Camarda, R., Rizzolatti, G., 1988. Functional organization of inferior area 6 in the macaque monkey. I. Somatotopy and the control of proximal movements. Experimental Brain Research 71, 475–490.

Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., Luppino, G., 2011. Cortical connections of the anterior (F5a) subdivision of the macaque ventral premotor area F5. Brain Structure and Functions 216, 43–65.

Graziano, M., 2006. The organization of behavioral repertoire in motor cortex. Annual Review of Neuroscience 29, 105–134.

He, S.Q., Dum, R.P., Strick, P.L., 1993. Topographic organization of corticospinal projections from the frontal lobe: motor areas on the lateral surface of the hemisphere. The Journal of Neuroscience 13, 952–980.

Hikosaka, O., Sakai, K., Nakahara, H., Lu, X., Miyachi, S., Nakamura, K., Rand, M.K., 1999. Neural mechanisms for learning of sequential procedures. In: Gazzaniga, M.S. (Ed.), The New Cognitive Neurosciences. MIT Press, Cambridge, pp. 553–572.

Jeannerod, M., Arbib, M.A., Rizzolatti, G., Sakata, H., 1995. Grasping objects: the cortical mechanisms of visuomotor transformation. Trends in Neuroscience 18, 314–320.

Lemus, L., Hernandez, A., Romo, R., 2009. Neural encoding of auditory discrimination in ventral premotor cortex. Proceedings of the National Academy of Sciences 106, 14640–14645.

Luppino, G., Matelli, M., Camarda, R., Rizzolatti, G., 1993. Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey. Journal of Comparative Neurology 338, 114–140.

Luppino, G., Matelli, M., Camarda, R.M., Gallese, V., Rizzolatti, G., 1991. Multiple representations of body movements in mesial area 6 and the adjacent cingulate cortex: an intracortical microstimulation study in the macaque monkey. Journal of Comparative Neurology 311, 463–482.

Matelli, M., Govoni, P., Galletti, C., Kutz, D.F., Luppino, G., 1998. Superior area 6 afferents from the superior parietal lobule in the macaque monkey. Journal of Comparative Neurology 402, 327–352.

Morecraft, R.J., Louie, J.L., Herrick, J.L., Stilwell-Morecraft, K.S., 2001. Cortical innervation of the facial nucleus in the non-human primate: A new interpretation of the effects of stroke and related subtotal brain trauma on the muscles of facial expression. Brain 124, 176–208.

Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.A., 2011. Action observation circuits in the macaque monkey cortex. Journal of Neuroscience 31, 3743–3756.

Olson, C.R., Gettner, S.N., 1995. Object-centered direction selectivity in the macaque supplementary eye field. Science 269, 985–988.

Passingham, R.E., 1993. The Frontal Lobe and Voluntary Action. Oxford University Press, Oxford.

Perrett, D.I., Harries, M.H., Bevan, R., Thomas, S., Benson, P.J., Mistlin, A.J., Chitty, A.J., Hietanen, J.K., Ortega, J.E., 1989. Frameworks of analysis for the neural representation of animate objects and actions. Journal of Experimental Biology 146, 87–113.

di Pellegrino, G., Wise, S.P., 1991. A neurophysiological comparison of three distinct regions of the primate frontal lobe. Brain 114, 951–978.

Picard, N., Strick, P.L., 2001. Imaging the premotor areas. Current Opinion in Neurobiology 11, 663–672.

Prabhu, G., Shimazu, H., Cerri, G., Brochier, T., Spinks, R.L., Maier, M.A., Lemon, R.N., 2009. Modulation of primary motor cortex outputs from ventral premotor cortex during visually guided grasp in the macaque monkey. Journal of Physiology 587, 1057–1069.

Rizzolatti, G., Craighero, L., 2004. The mirror neuron system. Annual Review of Neuroscience 27, 169–192.

Rizzolatti, G., Fadiga, L., Fogassi, L., Gallese, V., 1997. The space around us. Science 277, 190–191.

Rizzolatti, G., Fogassi, L., Gallese, V., 2002. Motor and cognitive functions of the ventral premotor cortex. Current Opinion in Neurobiology 12, 149–154.

Rizzolatti, G., Kalaska, J.F., 2013. Voluntary movements: the premotor and parietal cortex. In: Kandel, E.E., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J. (Eds.), Principle of Neuroscience. McGraw-Hill, pp. 865–893.

Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.

Rizzolatti, G., Luppino, G., Matelli, M., 1998. The organization of the cortical motor system: new concepts. Electroencephalography and Clinical Neurophysiology 106, 283–296.

Rozzi, S., Ferrari, P., Bonini, L., Rizzolatti, G., Fogassi, L., 2008. Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. European Journal of Neuroscience 28, 1569–1588.

Sakata, H., Taira, M., Murata, A., Mine, S., 1995. Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. Cerebral Cortex 5, 429–438.

Stuphorn, V., Taylor, T.L., Schall, J.D., 2000. Performance monitoring by the supplementary eye field. Nature 408, 857–860.

Tanji, J., 1994. The supplementary motor area in the cerebral cortex. Neuroscience Research 19, 251–268.

Tanji, J., 2001. Sequential organization of multiple movements: Involvement of Cortical Motor Areas. Annual Review of Neuroscience 24, 631–651.

Tomassini, V., Jbabdi, S., Klein, J.C., Behrens, T.E.J., Pozzilli, C., Matthews, P.M., Rushworth, M.F.S., Johansen-Berg, H., 2007. Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. Journal of Neuroscience 27, 10259–10269.

Vaadia, E., Benson, D.A., Hienz, R.D., Goldstein, M.H., 1986. Unit study of monkey frontal cortex: active localization of auditory and visual stimuli. Journal of Neurophysiology 56, 934–952.

Wise, S.P., Boussaoud, D., Johnson, P.B., Caminiti, R., 1997. Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. Annual Review of Neuroscience 20, 25–42.

Prenatal and Infant Development: Overview, Current Trends, Future Directions

Lewis P Lipsitt and Jack Demick, Brown University, Providence, RI, USA
Carolyn Rovee-Collier[†]**,** Rutgers University, Piscataway, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L.P. Lipsitt, volume 17, pp. 11994–11997, © 2001, Elsevier Ltd.

Abstract

Section Overview of this article provides an overview of milestones in prenatal and infant development (from conception through 2 years of age and the emergence of language). Section Current Trends discusses current trends in the field with emphases on newer theories (e.g., dynamic system theories) and research, notably on the 'ecologies of infancy' (diversity of prenatal and infant development including the influences of culture and socioeconomic status, the effects of parenting, and nonparental care of infants) and on neurological brain assessments of infants. Section Future Directions concludes with consideration of future directions for the development of an integrated infant science.

Overview

Since 1980 the pace of new scientific knowledge about prenatal and infant development has increased sharply, yielding a picture of an organism actively engaged in extracting information from and interacting with the world. Even the fetus is responsive to stimulation in all sensory modalities and has the capability of learning – a capacity that expands in new and increasingly sophisticated ways after birth. The effects of early experiences are both cumulative and enduring throughout infancy (*see* Human Development, Theories of). Studies with prenates have revealed that neurally mediated behavior following birth is continuous with behavior during the prenatal period, finally shifting from principally subcortical to increasingly cortical control between 2 and 4 months of age. This shift ushers in a rapid increase in social and other learned behaviors that, in relatively short order, will position the infant on the threshold of childhood, armed with the motoric, social, and cognitive tools necessary to cross over. Thus, this article considers the mileposts of the organism's prenatal and postnatal achievements, recent theory and research, and future prospects for prenatal and infant science.

Infancy extends from birth to 2 years of age. It has long been regarded as a special period of life in which numerous important transformations of structure take place, and the essential processes of information exchange with the world are laid down. Over this period, the infant is transformed from a condition of utter dependence on the caregiver for survival to a condition of relative independence or life-preserving agency – a transformation accompanied by marked physical, cognitive, and behavioral changes. While no characterization of this transition is entirely satisfactory, several elements set the infant apart from children of other ages. Despite disclaimers that what happens very early in life can be reversed or improved upon, most developmental psychologists celebrate the very considerable and enduring influence of early experience (*see* Infancy and Childhood: Emotional Development; Socialization in Infancy and Childhood). The following section paints a broad picture of infant development, its prenatal antecedents, and its relation to the behavior of older children and adults.

Prenatal Development

Physical and Cognitive Development

The 'prenatal period' extends from conception to birth. During the 'germinal stage' (fertilization–2 weeks) and 'embryonic stage' (2–8 weeks), the fertilized egg rapidly divides and attaches to the uterine wall, the major organs and systems of the body differentiate, and interconnections form between the developing components of the central and peripheral nervous systems. During the fetal stage (8 weeks–birth), the organism grows rapidly and the physical features that are identified as human become increasingly apparent.

During the prenatal period, the mechanisms that will facilitate the organism's adaptation postnatally become functionally mature. Infants born prematurely, for example, who readily learn responses that increase the amount of quiet sleep, reveal this anticipatory maturation. The characteristics of the mother's voice are also learned during the prenatal period. After birth, hungry infants orient and root in the direction of their mother's voice, thereby increasing their efficiency of breastfeeding irrespective of whether the mother is holding them on her left or right side. Newborns will also learn to activate a tape recording of their mother's voice – but not of another woman's voice – by sucking nonnutritively on a modified nipple. Moreover, 'prenatal learning' is highly specific: newborns will suck harder to listen to their mother reading a Dr Seuss passage that she had read aloud daily during the last 6 weeks of gestation than to hear her reading a novel Dr Seuss passage (DeCasper and Spence, 1986). This specificity continues postnatally throughout the first half-year. Similar learning prenatally familiarizes organisms with the linguistic environment into which eventually they will be born.

The Jacksonian Principle

The development and dissolution of prenatal motor function is described by the 'Jacksonian principle,' which states that the

[†] Deceased.

last reflexes to develop are the first to disappear when the organism undergoes demise, whereas the first to develop are the last to disappear (Jackson, 1884, cited in Taylor, 1958). The contralateral (avoiding) reflex, which moves a stimulated site away from a tactile stimulus, is supplanted weeks later by an ipsilateral reflex (Hooker, 1952). This sequence is reversed by fetal anoxia. The reversion occurs because reflexes which develop first are well practiced and require less oxygen to be initiated than more recently developed reflexes. Similarly, the order in which sensory functions develop (tactile, vestibular, thermal, chemical, auditory, and visual) is subsequently reversed during injury, disease, or aging.

Mileposts of Infant Development

From 0 to 6 Months

Birth marks the organism's transition from a parasitic, waterborne existence in a thermally constant environment with constant access to nutrition to an independent, land-dwelling existence in a desiccating and cold environment with discontinuous access to food. The occasion of birth involves a transition important primarily for triggering the onset of the 'regulatory mechanisms' (respiration, blood flow, digestion) critical for this transition. These regulatory arrangements and their interactions are the major constraints on the newborn's behavior and the environmental conditions it can tolerate (Adolph, 1968). The evolutionary requirement of all newborn mammals is to grow, which necessitates maximizing energy intake and minimizing energy expenditure (*see* Lifespan Development: Evolutionary Perspectives). Not surprisingly, therefore, newborns primarily eat and sleep. Early postnatal learning also reflects this necessity: newborns readily learn where and when food is accessible. Also, because full-term newborns do not physiologically thermoregulate before 4–9 weeks of age, they do not require energetically costly responses which can compete with behavioral thermoregulation.

Over the first 3 months of life, the birth weight of the infant doubles. During this period, brain development is very rapid and influenced appreciably by the kinds of experience to which the infant has been subjected. While most information about dendrite proliferation, the directionality of neuronal connections, cell death, and the effects of deprivation on neural and behavioral development comes from mammals other than humans, the universality of the effects described by neuroanatomists and neuropsychologists is impressive.

In the period following birth, much of the infant's behavior is reflexive, that is, specific stimuli call forth specific reactions (e.g., touches to the mouth elicit rooting and mouthing, abrasions of the skin elicit retraction of the injured body part). This early behavior is principally under the control of the subcortical portions of the brain, and the elicited behaviors are typically swift and obligatory.

Between 2 and 4 months of age, cortical maturation – especially dendritic proliferation – is increasingly evident, and neonatal reflexes disappear or weaken. Also during this period, activity undergoes a major transition from elicited to voluntary. The developmental sequence in which reflexes emerged during the prenatal period is recapitulated during this transitional period, and the sequence in which the postnatal reflexes appear is similarly reversed when the organism is challenged (Humphrey, 1969). The postnatal repetition of fetal activity sequences is again due to the lower oxygen requirement of more primitive, hence more practiced, reflexes. Practicing neonatal reflexes reduces their oxygen requirement and speeds the transition.

The transitional process has been implicated in the 'sudden infant death syndrome' (SIDS). McGraw (1943) characterized the period when reflexes are neither reliably elicited nor under voluntary control as a period of 'disorganized behavior.' The transitional period of the respiratory occlusion reflex corresponds to the time when SIDS is most likely to occur (95% of all SIDS or 'crib death' cases occur between 2–5 months). By 6 months, the occlusion reflex has come under cortically mediated voluntary control, and the infant is once again able to defend successfully against respiratory occlusion (Lipsitt, 1979, 2003; *see* Infant and Child Mortality in Industrialized Countries). Also, over the first 3 months of life, the behavioral repertoire of the infant sharply expands.

Infants can acquire both classically conditioned and operantly conditioned responses at birth and can remember what they have learned for a surprisingly long time (Lipsitt, 1969). An eyeblink reflex that is classically conditioned at 10 days of age, for example, is remembered for at least 10 days. Newborns can habituate to a repeatedly presented stimulus and discriminate it from others that they have not previously encountered. Imitation, a particularly efficient form of learning, can be demonstrated within hours of birth. Not only can newborns imitate an adult's facial gestures immediately after they are modeled but, by 6 weeks of age, infants can imitate facial gestures after a 24-h delay. Piaget (1962) described early facial imitation as stimulus-bound and reflexive and hypothesized that it disappeared along with many other neonatal reflexes. However, it was later shown not to disappear (e.g., Meltzoff and Moore, 1994). Infants' early imitation of facial expressions foreshadows their later dyadic interactions with adults and siblings.

By 3 months of age, infants can acquire arbitrary categories, learn the serial order of an arbitrary list, and pick up incidental information about events in their visual surround – including information about correlated features ('what goes with what'). Although some aspects of visual processing (e.g., accommodation and convergence, and feature search) are adultlike by 3–4 months, others (e.g., visual acuity and conjunction search) do not achieve adult levels for 1 year. Infants can detect depth by 2–3 months but do not fear depth until they begin to crawl. They can also recognize faces of live adults after 24 h by 6 weeks of age and can recognize them from photos by 5 months.

Observers often remark that infants blossom into 'social creatures' between 2 and 4 months of age, cooing and smiling widely at individuals who interact with them, reaching to other persons, and intensely watching the behavior of others. By their fourth month, infants follow conversations between adults with their eyes, differentiate social from nonsocial objects (which they previously did not), and imitate vocalizations. Not surprisingly, social stimuli are highly effective in establishing and maintaining behavior at this time. The features that define a stimulus as social, however, change with age. Although newborns differentiate between facial expressions and prefer to look at faces than at other visual stimuli, this preference is not based on a social dimension but on contrast and movement – critical features of the human face

that influence infants' scanning patterns over the first 8 weeks. By 4 weeks of age, however, infants display greater pupillary dilation to social than nonsocial stimuli; by 6 weeks, they display differential cardiac responses to mothers and strangers; and by 12 weeks, they display differential pupillary responses to them and smile more frequently at the mother than at a stranger.

From 6 to 12 Months

During the second half-year, a normal infant becomes a walker, a speaker, and a social being who responds emotionally to parents and differentially to familiar persons and strangers. The simultaneous appearance of strong attachment to caregivers and wariness of unfamiliar individuals at 9–10 months of age is one of the most consistent benchmarks of this period (Ainsworth, 1979). At the same time, infants begin to look at the target of a pointing finger rather than at the finger per se and begin to exhibit social referencing – a strategy by which infants assess the emotional valence of a new situation or person by checking how the caregiver has responded and responding in kind.

For a half-century, another developmental benchmark was attainment at 10 months of object permanence – the concept that an unseen object or person continues to exist. In the classic object search task, infants younger than 10 months consistently failed to retrieve an object they had just watched an adult hide. New paradigms using looking instead of reaching measures, however, have now revealed that even 3-month-old infants can locate where an object was hidden.

The types of learning of which infants are capable and their understanding of the physical world rapidly expand over the last half-year. Infants can individuate objects, solve problems requiring causal reasoning, form learning sets ('learning how to learn'), comprehend some spoken speech, repeat a couple of syllables and perhaps produce a word, plan and execute a detour, and track objects through occlusion – a finding that has fueled heated debate over whether infants can count. Before infants begin to crawl (7–9 months), they have already learned what happens in what places. Once infants can locomote, they begin to form a cognitive map or representation of the spatial relationships between these places, which facilitates navigation from one place to another.

From 12 to 24 Months

Over this period, infants' linguistic, motoric, and cognitive skills continue to improve and broaden. Between 12 and 18 months of age, for example, infants can use landmarks to locate objects instead of using their own body as a reference point and, between 18 and 24 months, they can use geometric cues in the environment to reorient. Although 18 months was set as the minimum age for deferred imitation of actions on an object (Piaget, 1962), the use of tasks particularly suited to younger infants has lowered this age to 6 months. Still, infants' ability to imitate televised demonstrations lags well behind their ability to imitate live demonstrations. The same live demonstration that 6-month-old infants can imitate after 24 h cannot be imitated from television before 18 months, although 15-month-old infants can imitate some televised actions after 24 h if the viewing conditions are more naturalistic. Generally speaking, the pervasiveness of reports that infants' cognitive and motoric performance is superior in natural settings calls into question most if not all minimum age limits that have been imposed by laboratory tests.

Within the last 5 years, research on the ontogeny of memory has revealed that retention increases linearly over the infancy period, and reminders protract retention even longer. At all ages, a single reactivation or reinstatement reminder, for example, doubles the life of a memory – a remarkable result given that memories of older infants are initially so long. The finding that a few reinstatements can maintain a memory acquired at 6 months through 2 years of age reveals that early experiences can be enduring and raises new questions about infantile amnesia (adults' inability to recollect experiences from early childhood). The two memory systems (declarative, nondeclarative) found in older children and normal adults are also functional by 3 months (Rovee-Collier et al., 2001).

Risk Factors and Individual Differences

In recent years, advances in the successful delivery and survival of very small and premature babies, often born in jeopardy, have fueled new debate over the importance of early experience. Because the lower threshold for survival is very new, it is inevitable that time must pass before the numerous longitudinal studies, now under way, will yield critical information about the relation between specific prenatal deficits and neonatal hazards, on the one hand, and particular developmental outcomes, on the other. Such investigations are unlikely to rule out the possibility that some affected infants will be indistinguishable from their age-matched peers in developmental competencies. It is a question of proportions and resiliency.

Although infants born 38 weeks after conception are, on average, less mature than those born at 40 weeks are, the level of physiological maturity can vary among infants born at 40 weeks by as much as 3–4 weeks. Sex differences are also present at birth, with girls being 1–2 weeks ahead of boys in neurological maturation. In normal infants, some of this variability disappears by the end of the first year, but many individual differences persist, the full range and significance of which are not yet fully appreciated.

Early attempts to predict later intellectual performance from infant behavior focused on sensory and motor abilities and were unsuccessful. Modest-to-good correlations have now been found between measures of information processing beginning at 5–6 months (e.g., habituation rate, novelty preference, 1-week retention ratios) and scores on various intelligence tests at 3–5 years. The Fagan Test of Infant Intelligence (Fagan and Detterman, 1992), which correlates later intelligence with infant novelty preference scores, is sensitive to early exposure to environmental toxins, substances of abuse, nutritional deficits and supplements, maternal HIV infection, intraventricular hemorrhage, bronchopulmonary dysplasia, genetic anomalies, and various neurological abnormalities.

Early Origins

The origins of research on infant development can be traced to baby diaries published by Darwin and Preyer in the late

nineteenth century (*see* Lifespan Development: Evolutionary Perspectives). Over the first half of the twentieth century, detailed diaries of their own children's behavior stimulated Morgan, Stern, Guillaume, Valentine, and Piaget to propose theoretical accounts of early cognitive development (Wallace et al., 1994). At the same time, Pavlov's seminal work spawned longitudinal studies of classical conditioning with infants in Russia and other Eastern Bloc countries. These early studies provided the first systematic data on infant learning and documented the sequential onset of sensory function from the age at which stimuli from different modalities became effective conditioned stimuli. During this period, Western researchers produced only scattered reports of infant conditioning and retention, emotional development, activity, and color vision.

In 1958, federal funding of a long-term prospective study of infant development ushered in the era of modern infancy research. Using sophisticated apparatus and innovative methods, researchers at 12 sites across the United States began systematically to document the infant's basic sensory and learning abilities. These initial inquiries sparked subsequent generations of research on all aspects of perceptual, cognitive, and behavioral development across the entire span of infancy. By the end of the twentieth century, a large amount of developmental data were amassed on infant attention, perception, cognition, motor skills, and the sequellae of many early risk conditions and prematurity such as poverty.

Current Trends

Theory

Dynamic Systems Theories

Given the complexity of the issues at hand, the last 20 years or so have seen changes in some of the theories in which infant developmentalists are interested. For example, many have turned to transactional approaches, which treat the "… person's behaving, including his most advanced knowings as activities not of himself alone, nor even primarily his, but as processes of the full situation of organism–environment" (Dewey and Bentley, 1949: p. 104). In perhaps the most general version of dynamic systems theory based on an elaboration and extension of Werner's (1957) comparative developmental theory (Demick, 2011), the organism (here, infant), the environment, and their respective aspects (the physical/biological, psychological/intrapersonal, and sociocultural aspects of the organism together with the physical, interorganismic, and sociocultural aspects of the environment) mutually define, and cannot be considered independent of, one another.

The implications of such conceptualization are far-reaching in that: (1) the appropriate unit of analysis becomes the infant-in-environment system whereby the infant is viewed as a system embedded in a variety of larger environmental systems; (2) a change in one aspect of the infant-in-environment system impacts the system as a whole, which is assumed to function in a state of dynamic equilibrium; (3) unidirectional socialization effects (parent-as-cause, child-as-effect) and bidirectional socialization effects (behaviors usually interpreted as the causes of correlated child behaviors may be parental reactions to children's constitutionally based traits such as temperament) are alternatively seen as transactional effects (infants and parents mutually and cumulatively impact one another in a continual process of development and change); and (4) by extension, infants' inherent characteristics are shaped by experience and this multidirectional influence of heredity and environment continues throughout life. Although there are subtle differences, others have termed theories employing such notions as 'systems,' 'ecological,' or 'contextual' theories.

Dynamic Systems Theory Applied to Infant Development

Some infant developmentalists now view development as dynamic insofar as the organization of the system as a whole changes as the infant matures and acquires new experiences; as one subsystem emerges, that change is accompanied by new experiences that influence and are influenced by comprehensive changes in related component systems (Thelen and Smith, 2006). For example, higher-level behavioral organization, development of self-produced locomotion, emerges from the coordination of the infant's component processes (visual–vestibular adaptation, visual attention, social referencing) and the change in a single process, self-produced locomotion, affects a wide array of his or her psychological skills (height perception, spatial localization of hidden objects, concept formation, emotional expression). Other developmentalists see inner biological factors as embedded in the individual, who is embedded in a family, embedded in a community, embedded in a culture. Thus, as the dynamics of a family change, changes in the relationship between parents, and between each of the parents and the infant, are also expected to change. The 'goodness-of-fit' or 'match' between infants and aspects of their developmental contexts also influence their long-term outcomes (Lerner et al., 2011).

Research

With the contemporary appearance of newer developmental theories with different emphases and/or of novel research methods (e.g., neurological assessment), the last 20 years have also seen changes in some of the research problems in which infant developmentalists are interested. The following represents a partial list.

Ecological Studies

As a function of the transactional approach and of ecological theories, infant researchers have more recently examined the many ecologies of infancy by posing at least four general research questions as follow (Bornstein et al., 2014; *see* Ecology of Aging).

- To what extent do parents' interactions with each other influence infant development?

Parents exert direct effects on their infants, for example, through heritability (cf epigenesis) and parenting (e.g., with infant physical, cognitive, and psychosocial competencies related to parental warmth, responsiveness, stimulation, and nonrestrictiveness; *see* Socialization in Infancy and Childhood). However, indirect effects (e.g., quality of parental relationships) can have just as significant impacts on infant development as direct effects. Supporting Bowlby's (1951) early observation,

when the relationship between parents is warm and supportive, parents are more attentive and sensitive to their infants. Paternal support has less effect on mothers (society expects mothers, regardless of their psychological states, to be devoted to the caretaking of their infants) than maternal support has on fathers (societal expectations of fathers' behavior toward infants are more varied, which impact their behaviors in marital relationships differentially). Partner support leads to more positive parenting quality especially when families are experiencing stressors (e.g., those associated with the transition to parenthood, difficult child temperament, financial hardship). Both younger and older motherhood present challenges to infant rearing. Young mothers are less responsive and attentive to their infants, this being related to less knowledge and realistic expectations about infant development as a consequence of less or poorer schooling; older mothers are often not prepared for the physical demands of infant rearing. Same-sex couples (lesbians more so than gay males, however) typically have experiences similar to heterosexual couples although this does not hold for single parents who must often deal with financial hardship and/or lack of social support. Poverty and economic failure are associated with punitive parenting, child abuse, and child neglect. Infants' exposure to domestic violence elevates their distress in response to conflict later in life.

- To what extent do sibling and peer interactions influence infant development?

Supporting Harlow's (1960) classic findings that infant monkeys learn how to play, fight, relate to, and communicate with others through peer interactions, older siblings engage in more assertive and directive behaviors than their younger siblings, including infants. With great interest, infants follow older siblings around, who in turn influence infants' cognitive and psychosocial skills through both direct teaching and modeling. There is remarkable stability across time in the extent of interaction between infants and their preschool age siblings. The quality of sibling relationships is also moderated by infant temperament (sociable infants receive more attention), sex (same-sex infants and siblings relate better than opposite-sex siblings), maternal management style (that fosters cooperation), parent–child relationships (securely attached infants exhibit less negative behavior when mother plays with siblings alone), and parent–parent relationships (positive parent–parent relationships lead to more prosocial and less aggressive sibling relationships).

Infant–peer relationships develop through a series of stages: (1) although very young babies exhibit an interest in peers, social interaction between infants is infrequent and unsustained during the first 6 months but becomes more complex from 6 to 12 months (including looks, smiles, vocalizations, imitation, and responsiveness to peer overtures but still with very little physical contact); (2) by 2 years, infants (who typically develop empathic and prosocial behavior between 12 and 18 months) no longer respond contagiously to their peers' emotions but rather observe their peers' emotions carefully and respond appropriately with prosocial (helping, giving) behaviors. Positive infant–peer relationships that continue into childhood are also moderated by family harmony and secure attachment to parents.

- To what extent does day care influence infant development?

Within the United States, 58% of married mothers of infants younger than 12 months work outside of the home (with higher rates for unmarried mothers and married mothers with infants older than 12 months). The rates are much lower in other countries because new mothers are often allowed paid leave from work for at least 1 year (vs 3 months in the United States). Nonetheless, longitudinal studies often subsidized by the governments of many countries indicate that high quality of care (characterized by adequate attention to each infant, encouragement of sensorimotor and language development, safeguards for health and safety, and warm, responsive professional caregivers) is crucial to infant development regardless of whether it takes place in a family day care or a center day care setting. In addition to the need for high-quality day care, its impact may vary from infant to infant with less difficulty exhibited by girls versus boys, by infants with easier versus more difficult temperaments, and by infants from families with more versus less income and education. Positive benefits of infant day care include infants' increased cognitive, language, and to a lesser extent social development as well as increased family finances and decreased parental depression, which are related to more adaptive children. Further, secure attachments to day care providers neither threaten nor displace secure attachments to parents. However, infant day care is detrimental when the mother is insensitive and the infant spends more than 20 h per week in a poor-quality program with too many children per group (NICHD Early Child Care Research Network, 2005).

- To what extent do socioeconomic status (SES) and culture influence infant development?

Fueled by changes in the demographic composition of our contemporary society, researchers have routinely begun to explore racial/ethnic differences in aspects of infant as well as child, adolescent, and adult development. The literature is thorny because investigators have often not differentiated between SES (typically based on education and occupation) and race/ethnicity (categorization of individuals based on outward appearance, which some have argued is merely a 'social construction' providing some with power because, e.g., dark-skinned people with African ancestry are more genetically diverse than light-skinned people with European ancestry). Thus, SES and race/ethnicity must routinely be considered independent factors that often interact with one another (cf Silva and Demick, 2013).

With respect to the surprisingly small number of studies on the influence of SES on infant development, researchers have replicated Labov's (1969) classic work documenting that, even though their infants are not yet talking, middle-class US mothers speak to them more (attempting to engage them in conversation to expand their communicative abilities) than lower-class mothers (who use more directives and corrections to foster compliant behavior). In line with this, there are also socioeconomic differences in maternal and paternal holding, maternal carrying, and paternal caregiving with extreme economic disadvantage leading to parents' decreased ability to respond effectively to their infants.

With respect to cultural differences, the motor skills (e.g., head control, limb extension, sitting, standing, walking) of infants in Uganda, for example, are more advanced than the motor skills of American infants, which is thought to be related to differential child rearing practices (whereby Ugandan mothers extensively carry their infants on cradleboards located on their backs so as not to interrupt their work or other daily chores, which in turn provides infants with increased vestibular, visual, and proprioceptive stimulation necessary for motor skill development). Further, some researchers (e.g., Bornstein, 2012) have interpreted cultural differences in terms of the now-classic distinction between functioning in individualistic versus collective societies, while others have generated complex models of the multidirectional route of many influences (social position variables, racism, segregation, promoting/inhibiting environments, adaptive culture, child characteristics, family) on minority children's development (e.g., Garcia Coll et al., 1996).

Studies of General Developmental Processes in Infancy: The Case of Body Representations

One area of recent interest to infant developmentalists is the early development of body representations. This research is the focus of the thirteenth installment in the *Cambridge Studies in Cognitive and Perceptual Development* series, edited by Hatano and Fischer. As the editors of this volume (Slaughter and Brownell, 2012) note, "Because we engage with the world and each other through our bodies and bodily movements, being able to represent one's own and others' bodies is fundamental to human perception, cognition and behavior. This edited book brings together, for the first time, developmental perspectives on the growth of body knowledge in infancy and early childhood and how it intersects with other aspects of perception and cognition …" (p. i).

This work is notable for several reasons. Although body image is one of the most widely researched constructs within psychology and related disciplines (Cash and Smolak, 2012), there has been little theoretical and/or empirical work on this problem by infant developmentalists. Thus, this work fills a much needed gap that will allow insights into body experience in infant development as well as into infant precursors of body experience development in children, adolescents, and adults.

Although the field of development is no longer dominated by monolithic theories, the contributions conform to Werner's (1957) cognitive-developmental analysis of body experience as developing first within the sensorimotor level (body action), then within the perceptual level (body perception), and finally within the conceptual level (body conception), all necessary precursors of the development of the next level, self-conception. Further, the organization – namely, the development of representations of the bodily self, of the bodies of others, and of the integration between self and other body representations – supports another of Werner's main themes of an increase in self-object differentiation with development. In these ways, this beginning work supports the generalization that there may be a synergistic relationship between grand developmental theories and problem areas previously unexplored (Demick, 2013).

This research area also has numerous implications. For example, early development of body representations helps us to understand better infant perception and representation of human form, imitation, understanding of biological motion, intentional understanding, action production, and self-representation. Most notably, the role of body representation for sensorimotor integration during visual perception also sheds light on its relationship to more general processes such as symbolic development. Further, in presenting appropriate theory to clarify terminology, this research sheds light on longstanding issues in infant development such as whether babies are born with or without a sense of self (broadly defined to include a bodily self). The proposed notion that infants possess a basic experiential, embodied self-unity at birth (body schema) but develop a more conceptual embodied self-unity (body image) over time can suggest that the newborn's body is for him or her, a differentiated entity among other entities in the environment (Dondi et al., 1999). Finally, in contrast to the qualitative methodologies typically employed in body image research on older individuals, the use of rigorous experimental methodologies appropriate for infants will ultimately help developmentalists assess the conditions under which cause–effect relationships in body-self interdependence occur in ontogenesis from its earliest beginnings.

Studies of Individual Differences in Infancy: The Case of Resilience

In line with the previous concern for individual differences (e.g., premature vs normal gestation infants), another hot-button topic for developmentalists in general is resilience, typically defined as "good outcomes in spite of threats to adaptation or development" (Masten, 2001: p. 228). Although this definition is clear enough, the construct of resilience has met both theoretical and methodological obstacles. Some researchers (e.g., Lipsitt and Demick, 2011, 2012) have acknowledged that the relationships among 'resilience' and earlier proposed constructs such as 'invulnerability,' 'stress resistance,' and 'protective factors' as well as the most efficacious way to measure resilience have been debated for some time but never resolved.

Lipsitt's (2003) work on SIDS as a biobehavioral phenomenon has direct applicability to the study of resilience in infancy. His approach assumes that the behavioral seeds of resilience inhere in the predisposing capabilities of the newborn infant. Viable human infants have reflexes, in some more intense than others, and these approach and avoidance behaviors are of lifesaving merit. The congenital reflexes of the newborn are 'coaxed' by environmental stimuli concordant with the growth of neuromyelin tissue and dendrite proliferation. 'Resilient infants' may be regarded as those in whom the migration to cortically mediated learned behavior from subcortical reflexes is successful, while nonresilient infants who succumb to SIDS during the critical 2- to 5-month period, when 90–95% of SIDS deaths occur, are those in whom the transition has been unsuccessful or incomplete. Nonresilient infants who succumb to SIDS during the critical 2- to 5-month period are those in whom the transition has been incomplete or unsuccessful. These nonresilient infants have not acquired the learned psychomotor strategies which, in resilient infants, can save the child's life from respiratory occlusion. Thus both researchers argue that future studies on resilience would be strengthened by investigations of antecedent–consequent

relations between prior adversities like familial violence and poverty, on the one hand, and later recovery or adaptation as mediated by a variety of coping strategies, on the other.

Lipsitt's discussion of the infant precursors of resilience sets the stage for Demick's (2014) contextual view of resilience in later stages of life as consisting of the individual's combination of the psychological traits of strength (e.g., hardiness, perseverance) and flexibility (e.g., mobility of cognitive style) with his or her capacity to shift back and forth between different subject–object conditions depending on the demands of the situation. Thus, both researchers argue that future studies on resilience would be strengthened by heightened laboratory investigations on antecedent–consequent relations between prior adversities (e.g., families, violence, poverty) and later recovery or adaptation as mediated by various coping strategies and fortuitous events.

Transdisciplinary Studies

Given the transdisciplinary nature of this encyclopedia, it is noteworthy that there are more and more research problems on which infant developmentalists and those in allied basic and applied professions are collaborating. Several of these are outlined below:

- Brain–behavior relationships in infant development

In the concluding section of our article in the first edition of this volume, we predicted that cognitive scientists and behavioral neuroscientists would soon join forces to relate developmental changes in infant cognition to specific changes in the developing brain, and that this effort would require the refinement of existing techniques and development of less invasive techniques for use with very young infants. We were correct on both counts.

Brain–behavior relationships in infant development have been assessed through the use of challenging methodologies that examine either infants' gross cortical electrical activity (measured by cortical event-related potentials, ERPs) or neurological anatomy (assessed by functional magnetic resonance imaging, fMRI). Use of these techniques is problematic because of cost and difficulty keeping infants in still positions. Nonetheless, based on these methods and the emerging evidence from them, it is clear that infants' memories are viable at very young ages and are capable of remarkable feats of deferred imitation and acting on goals established through previous experience, all in advance of presumptions from Piaget's (1962) important work on stages of development.

- Economics of infant development

Psychologists have joined forces with economists, political scientists, educators, and sociologists to collaborate on another problem of mutual interest, namely, the economics of infant development. Recent estimates suggest that a middle-class family can expect to spend over $15 000 simply bringing an infant from the hospital to a home equipped with the necessary supplies. A middle-class family can also expect that it will spend approximately another $225 000 for food, shelter, and other expenses (not including costs associated with college attendance and inflation). However, based on a cost–benefit analysis of early programming for pregnant mothers and young children relative to other social programs (e.g., increased numbers of high school personnel, adult job programming), Heckman (2006) argues that successful early intervention attempts like infant cognitive skill building, family support programs yield positive results such as lower adolescent pregnancy rates and less public assistance usage, and have a better rate of return. Such policy analyses of developmental data from other countries have also been conducted (e.g., Case et al., 2004).

- Parental development

Fueled by the publication of Bornstein's five-volume *Handbook of Parenting* and his initiation of a new journal *Parenting: Science and Practice*, researchers from a variety of disciplines (e.g., psychology, sociology, anthropology, medicine) have begun to explore the problem area of parental development. Although Bornstein's efforts have their origins in the emergence of ecological studies, investigators have taken a number of classical 'stage' theoretical approaches to this problem ranging from psychoanalytic theory (in the first stage, parents must relive and/or renegotiate their own oral conflicts when attempting to help their infants negotiate the same conflict) to cognitive-developmental theory (in the first stage, parents exhibit an 'egoistic orientation' and perceive their infants simply as projections of themselves) to sociological theory (the first stage, transition to parenthood, is a period of crisis for new parents that inevitably leads to decreased marital satisfaction). However, very few researchers have adopted transactional theories, which as noted have become increasingly popular and which led to Bornstein's new initiatives.

As exception, Demick (2002) argues that, whenever there is a perturbation to the person-in-environment system (e.g., cognitive disequilibrium in parents related to their infants' or children's behavior), parents must restore cognitive equilibrium to the system by reorganizing their self–world relationships. Such developmental processes occur with great frequency in transacting with one's infants or children on a daily basis. Although the particular issues posed by one's infants or children may change from one moment to the next (even if some appear with great regularity), parental reactions do not necessarily change. Parents are constantly faced with restoring cognitive equilibrium (or equilibrium to the person-in-environment system) in a dialectical process that they may experience similarly every time they are faced with a novel (or not so novel) stimulus from their infants or children. That is, parents are continually faced with attempting to return their differentiated and conflicted person-in-environment system states to more differentiated and integrated ones.

Future Directions

The next 20 years will witness extensive research that elaborates on present knowledge and which may challenge our current beliefs about infant development. Although advances have been made by cognitive and behavioral neuroscientists to relate developmental changes in infant cognition and, more recently, infant affect to specific changes in the developing brain, we again stress that our knowledge of infant brain–behavior relationships is still in its 'infancy.' New noninvasive techniques (e.g., similar to ERPs and fMRIs) are developed for use with very young infants. Methodological and technological advances should

also allow for new insights into prenatal development and its contribution to postnatal development.

Given the Zeitgeist that has contributed to making 'context' the buzzword for developmental psychologists and 'brain' the mantra for cognitive- and neuroscientists, future work should begin with these spirits of the time toward delineating and examining the structural and functional interrelationships between and among levels of organization (biological, psychological, sociocultural). Inroads into such conceptualization and research have already been made by some investigators.

Triggered by excitement surrounding the integration of previously independent lines of study (social developmental and neuroscience research), Bernier et al. (2012) have examined prospective links between quality of the infant caregiving environment and children's subsequent executive functioning (higher-order cognitive processes such as set-shifting, planning, and impulse control leading to successful goal-directed and self-regulated action, assumed located in the brain's frontal lobe). Given significant brain plasticity (whereby experience largely determines which synaptic connections persist or are pruned related to lack of use) and brain development (protracted postnatal development of the prefrontal cortex), the researchers assessed families on maternal interactive behavior, paternal interactive behavior, and child attachment security on three occasions (when infants were between 1 and 2 years of age) and child executive functioning on two occasions (at 2 and 3 years of age). They found that scores of parental behavior, including maternal sensitivity, maternal autonomy, and quality of father–infant interactions, as well as infant attachment status, are subsequently related to children's performance on executive functioning tasks (specifically, the exhibition of both strong cognitive flexibility and working memory). However, they also uncovered that infants' attachment status is the best predictor of their later executive functioning, leading to the conclusion that attachment security plays a meaningful role in the development of executive control. More studies such as this are sorely needed since they have the potential to advance theory, research, and practice as well as transdisciplinary collaboration.

Finally, before transdisciplinary collaboration might take place, researchers will benefit from comparing and contrasting the assumptions underlying their respective theoretical approaches. We are confident that the continuation of the strident pursuit of clarifying research on infant behavior and development, and the inevitable theoretical progress that will result, will reveal more subtle but profound commonalities from behavioral and dynamic system approaches (Lipsitt and Demick, 2012). We will not only become closer to understanding one another better but also to developing further the science of infancy in all of its real world complexity.

See also: Cognitive Psychology: History; Concept Formation in Childhood: Cross-Cultural Perspectives; Conscious Control During Childhood, Development of; Ecology of Aging; Human Development, Theories of; Infancy and Childhood: Emotional Development; Infant and Child Mortality in Industrialized Countries; Lifespan Development: Evolutionary Perspectives; Socialization in Infancy and Childhood.

Bibliography

Adolph, E.F., 1968. Origins of Physiological Regulations. Academic Press, New York.

Ainsworth, M.D.S., 1979. Attachment as related to mother-infant interaction. In: Rosenblatt, J.B., Hinde, R.H., Beer, C., Bushnell, M. (Eds.), Advances in the Study of Behavior, vol. nine. Academic Press, New York, pp. 29–40.

Bernier, A., Varlson, S.M., Deschênes, M., Matte-Gagné, 2012. Social factors in the development of early executive functioning: a closer look at the caregiving environment. Developmental Science 15 (1), 12–24.

Bornstein, M.H., 2012. Cultural approaches to parenting. Parenting: Science and Practice 12 (2–3), 212–221.

Bornstein, M.H., Arterberry, M.E., Lamb, M.E., 2014. Development in Infancy: A Contemporary Introduction, fifth ed. Psychology Press, New York.

Bowlby, J., 1951. Maternal Care and Mental Health. World Health Organization, Geneva.

Case, A., Paxson, C., Ableidinger, J., 2004. Orphans in Africa: parental death, poverty, and school enrollment. Demography 41 (3), 483–508.

Cash, T.F., Smolak, L., 2012. Body Image: A Handbook of Science, Practice, and Prevention, second ed. Guilford Press, New York.

DeCasper, A.J., Spence, M.J., 1986. Prenatal maternal speech influences newborns' perception of speech sounds. Infant Behavior and Development 9, 133–150.

Demick, J., 2014. A Revised Manual for Witkin, Oltman, Raskin, and Karp's (1971) Group Embedded Figures Test (GEFT). Mind Garden, Menlo Park, CA.

Demick, J., 2013. "You're no-body till some-body loves you:" A review of V. Slaughter & C.A. Brownell's early development of body representations. PsycCRITIQUES 58 (25).

Demick, J., 2011. Effects of children on adult development and learning: 50 years of theory and research. In: Hoare, C. (Ed.), The Oxford Handbook of Reciprocal Adult Development and Learning, second ed. Oxford University Press, New York,pp, pp. 365–380.

Demick, J., 2002. Stages of parental development. In: Bornstein, M. (Ed.), Handbook of Parenting, Being and Becoming a Parent, second ed., vol. three. Erlbaum Associates, Mahwah, NJ, pp. 389–413.

Dewey, J., Bentley, A.F., 1949. Knowing and the Known. Beacon Press, Boston.

Dondi, M., Simion, F., Caltran, G., 1999. Can newborns discriminate between their own cry and the cry of another newborn infant. Developmental Psychology 35 (2), 418–426.

Fagan, J.F., Detterman, D.K., 1992. The fagan test of infant intelligence: a technical summary. Journal of Applied Developmental Psychology 13 (2), 173–193.

Garcia Coll, C., Lamberty, G., Jenkins, R., McAdoo, H.P., Crnic, K., Wasik, B.H., Garcia, H.V., 1996. An integrative model for the study of the developmental competencies in minority children. Child Development 67 (5), 1891–1914.

Harlow, H.F., 1960. Primary affectional patterns in primates. American Journal of Orthopsychiatry 30, 676–684.

Heckman, J.J., 2006. Skill formation and the economics of investing in disadvantaged children. Science 312, 1900–1902.

Hooker, D., 1952. The Prenatal Origin of Behavior. University of Kansas Press, Lawrence, KS.

Humphrey, T., 1969. Postnatal repetition of human prenatal activity sequences with some suggestions of their neuroanatomical basis. In: Robinson, R.J. (Ed.), Brain and Early Behavior: Development in the Fetus and Infant. Academic Press, London, pp. 43–84.

Labov, W., 1969. Contraction, deletion, and inherent variability of the English copula. Language 45 (4), 715–762.

Lerner, R.M., Lewin-Bizan, S., Warren, A.E.A., 2011. Concepts and theories of human development. In: Bornstein, M.H., Lamb, M.E. (Eds.), Developmental Science: An Advanced Textbook, sixth ed. Psychology Press, New York, pp. 3–49.

Lipsitt, L.P., 1969. Learning capacities of the human infant. In: Robinson, L.J. (Ed.), Brain and Early Behavior: Development in the Fetus and Infant. Academic Press, London, pp. 227–245.

Lipsitt, L.P., 1979. Critical conditions in infancy: a psychological perspective. American Psychologist 34 (10), 973–980.

Lipsitt, L.P., 2003. Crib death: a biobehavioral phenomenon? Current Directions in Psychological Science 12 (5), 164–170.

Lipsitt, L.P., Demick, J., 2012. Theory and measurement of resilience: views from development. In: Ungar, M. (Ed.), The Social Ecology of Resilience: A Handbook of Theory and Practice. Springer Science, New York, pp. 43–51.

Lipsitt, L.P., Demick, J., 2011. Resilience science comes of age: old age, that is. Review of P.S. Fry & C.L.M. Keyes' New frontiers in resilient aging: life-strength and well-being in later life. PsycCRITIQUES 56 (26).

Masten, A.S., 2001. Ordinary magic: resilient processes in development. American Psychologist 56 (3), 227–238.

McGraw, M.B., 1943. The Neuromuscular Maturation of the Human Infant. Columbia University Press, New York.

Meltzoff, A.N., Moore, M.K., 1994. Imitation, memory, and the representation of persons. Infant Behavior and Development 17, 83–99.

NICHD (National Institute of Child Health and Human Development) Early Child Care Research Network, 2005. Child Care and Child Development: Results from the NICHD Study of Early Child Care and Youth Development. Guilford Press, New York.

Piaget, J., 1962. Play, Dreams, and Imitation in Children. Routledge and Kegan Paul, London.

Rovee-Collier C.K., Barr R., 2001. Infant cognition. In: Pashler, H. (Series Ed.), Wixted, J. (vol. Ed.), Stevens' Handbook of Experimental Psychology, vol. 4: Methodology, third ed. Wiley, New York, pp. 693–791.

Rovee-Collier, C.K., Hayne, H., Colombo, M., 2001. The Development of Implicit and Explicit Memory. Benjamins Publishing, Amsterdam.

Silva, J., Demick, J., 2013. Body experience, self-experience, and nutrition in middle adolescence: sex, race, socioeconomic status. In: Paper Presented at Annual Meetings of the American Academy of Pediatrics (AAP) National Conference. Orlando, FL.

Slaughter, V., Brownell, C.A. (Eds.), 2012. Early Development of Body Representations. Cambridge University Press, Cambridge, UK.

Smotherman, W.P., Robinson, S.R., 1988. Behavior of the Fetus. Telford Press, Caldwell, NJ.

Taylor, J.E. (Ed.), 1958. Selected Writings of John Hughlings Jackson, vol. 2. Hodder and Stoughton, London.

Thelen, E., Smith, L.B., 2006. Dynamic systems theories. In: Lerner, R.M. (Ed.), Damon, W. (Series Ed.), Handbook of Child Psychology, vol. one: Theoretical Models of Human Development. Wiley, Hoboken, NJ, pp. 258–312.

Wallace, D.B., Franklin, M.B., Keegan, R.T., 1994. The observing eye: a century of baby diaries. Human Development 37 (1), 1–29.

Werner, H., 1957. Comparative Psychology of Mental Development. International Universities Press, New York.

Relevant Websites

http://www.afit.org – Alliance for Infants and Toddlers.

http://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/infants.html – Centers for Disease Control and Prevention (CDC).

http://www.developingchild.harvard.edu – Center on the Developing Child at Harvard University.

http://www.childdevelopmentinfo.com – Child Development Institute.

http://www.nlm.nih.gov/medlineplus/infantandnewborndevelopment.html – Medline Plus.

http://www.nacd.org – National Association for Child Development (NACD).

http://www.pbs.org/parents/child-development/ – P.B.S. Parents.

http://www.srcd.org – Society for Research in Child Development (SRCD).

http://www.waimh.org – World Association for Infant Mental Health.

http://www.zerotothree.org – Zero to Three: National Center for Infants, Toddlers, and Families.

Multimedia

Center on the Developing Child at Harvard University: www.developingchild.harvard.edu. Designed to translate science into policy for multiple audiences including the public, this site contains multimedia materials mainly of interest to professionals and academics. Included are brief videos (e.g., 'In Brief: The Science of Neglect'), interactive features in the form of easy-to-follow slideshows on key concepts in early childhood development science (e.g., 'How Early Experiences Alter Gene Expression and Shape Development') and program evaluation (e.g., 'Early Childhood Program Evaluations: A Decision-Maker's Guide'), scientific publications, and working papers. The site also provides descriptions of the Center's initiatives including its National Scientific Council on the Developing Child and the National Forum on Early Childhood Policy and Programs.

Center on the Social and Emotional Foundations for Early Learning: http://csefel.vanderbilt.edu/resources/training_infant.html. This Vanderbilt University Center offers video infant/toddler training modules on promoting social and emotional competence. The modules are based on input about useful types and content of training from focus groups consisting of program administrators, early infant and childhood educators, and family members. All modules are designed to be consistent with evidence-based practices in the scientific literature. Each module comes with a PowerPoint presentation and all modules, including archived ones, are available in Spanish.

Centers for Disease Control and Prevention (CDC): http://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/infants.html. This site contains multimedia materials for families, healthcare providers, and early childhood educators. For parents, there are videos (e.g., on how to look for developmental milestones) and Public Service Announcement (PSA) Video Kits on "Learn the Signs, Act Early" for them to distribute to local television and radio stations. For providers, the site supplies developmental screening tools and educational materials to provide to parents. For educators, there is a Childcare Provider Resource Kit including a CD-ROM with fact sheets and growth charts on child development for distribution to parents and posters for classroom display.

Education Portal: http://education-portal.com/academy/course/index.html. This Web site, which offers free college courses that may be transferred into accredited institutions, contains several lessons on Piaget's theory of cognitive development. Unlike many other online lessons that begin with Piaget's preoperational stage in early childhood, this site has a complete module on his sensorimotor stage of cognitive development that occurs in infancy with a focus on the construct of object permanence. This and other modules on Piaget (as well as all modules offered on this site) consist of written text, video, definitions, extensive examples, and a quiz (along with recommendations for other essential videos on infant development more generally). Although Piaget's theory has not been fully supported by research data, his work was groundbreaking particularly insofar as he provided us with an early glimpse into the mind of the infant.

Endowment for Human Development (EHD): www.ehd.org/movies.php. This is the official Web site of EHD, a nonprofit organization dedicated to improving health education and public health by equipping educators, clinicians, and governments with information to appreciate, apply, and communicate about the science of health and human development. The sight offers numerous selection of movies and videos on both normal and abnormal child development; some are free while others are available for purchase at low prices. The site also includes the best free video of the 'biology of prenatal development' that we were able to find on the net. The site also offers fact sheets (e.g., prenatal summary, prenatal time, prenatal overview), recommended readings on a variety of topics (including prenatal development), study guides, and a regular newsletter.

Head Start: An Office of the Administration for Children and Families Early Childhood Learning & Knowledge Center (ECLKC): https://eclkc.ohs.acf.hhs.gov. This Web site for Head Start, the best known early intervention program of its kind, provides a continually updated set of webcasts and videos for both lay and professional audiences. For example, the video 'Making Learning Meaningful' helps parents discover strategies for making learning relevant to children's everyday lives. The webinar 'A-133 Audit Training Webcast' aids professionals in understanding the importance of the annual program audit and the ways in which it can be used to strengthen an organization. Relevant here, there is an extensive collection of webcasts focusing on infants and toddlers from a description of Head Start's Family Child Care Program Option to those detailing lessons about infants, for example, building with blocks or playing with a ball. These latter broadcasts provide parents with information on what to expect from infants when they explore new materials as well as to help parents learn how simple activities may introduce infants to math and other concepts (e.g., exploration of the ball's roundness).

National Science Foundation (NSF): http://www.nsf.gov/news/special reports/science nation/babyface.jsp. This site highlights an NSF-funded study at the University of Miami on how infants' smiles provide clues to healthy development. An interdisciplinary team consisting of developmental scientists, machine vision experts, and face perception researchers are capturing on video cameras short structured playtimes between infants and their mothers. The videos are analyzed with a software program that registers and tracks the movements of the faces of the infants and their mothers as they interact in time toward illustrating the ways in which infants develop, for example, a sense of shared social emotion. In a companion study, the researchers are employing the same software program for examining the same types of interaction between mothers and their infants at risk for autism spectrum disorder (ASD) because they have an older sibling with a diagnosed ASD. Not only does this site describe funded research by one of our top scientific organizations but it also provides a video demonstrating the study's new software, which may transform how other questions about infant development are addressed.

Presidency, United States

John Hart, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of the U.S. presidency has traditionally been broad in scope, disparate, eclectic, and unconstrained by fixed theoretical and conceptual frameworks. During the past 30 years' considerable introspection and self-criticism among presidency scholars have resulted in a more rigorous methodological approach to presidency scholarship and greater concern about its theoretical and conceptual underpinnings. No single approach dominates the political science of the U.S. presidency and current scholarship represents a mix of traditional and new approaches. In recent years, rational choice theorists have staked a claim for the comprehensiveness of their approach and have suggested that it may offer presidency scholarship the analytical and theoretical coherence it has lacked. Future research is likely to be driven in part by scholarly professional imperatives and, while rational choice and new institutionalism are at the cutting edge of presidency scholarship, that approach has not gone unchallenged. It remains to be seen whether the parsimony of rational choice theory can deal with the many and complex variables relevant to the study of the presidency.

Meaning

The term 'presidency' is applied to several distinct perspectives on the office of the President of the United States. Most commonly, and in its narrowest sense, it is employed as a label to refer to the incumbency of one particular President, as in the presidency of Abraham Lincoln or the presidency of Franklin Roosevelt. More broadly, the term applies to the historical development and legacy of a succession of presidencies since the office was established in 1789. In a more precise constitutional sense, the presidency is that branch of government established by Article II of the Constitution of the United States encompassing a set of constitutional and statutory powers and duties of the President together with additional 'implied' or 'inherent' powers and responsibilities not specifically enumerated in the Constitution or statute law. Finally, the term 'presidency' is used to denote the institutional structure of the office of President embracing the Executive Office of the President, a collection of staff units, originating from proposals contained in the Brownlow Report of 1937, that provide the President with the capacity to manage the work of the executive branch of government and to liaise with the principal constituents of the President, namely the Congress, the media and interest groups.

These different usages of the term present no significant analytical difficulties. In the literature on the U.S. presidency, the context is usually sufficient to clarify the sense in which the term is being employed and in none of its various forms is it conceptually or theoretically problematic. This is partly because the academic study of the presidency has not been as encumbered by rigid theoretical or conceptual frameworks as many other areas of the study of politics – a matter of concern to some presidency scholars who wish it were so, and a relief to others who see value in the broad, disparate, and eclectic approach that has characterized the study of the presidency to date.

Intellectual Context

The academic study of the U.S. presidency has mainly been the preserve of historians and political scientists. Historians have generally confined their attention to individual presidents and presidencies whereas political scientists have traditionally concentrated on the constitutional and political power of the presidency and its institutional structure and development over time. Few historians, with some notable exceptions (see, e.g., Schlesinger, 1973; McDonald, 1994), have focused on the broader institutional development of the American presidency and only a handful of political scientists (see, e.g., Milkis, 1993; Skowronek, 1993) have borrowed from the discipline of history to further what political science has had to say about the presidency. Given the vast amount of scholarship on U.S. presidents and the presidency produced by historians and political scientists, there has been little significant cross-fertilization between the two disciplines and it would not be unfair to say that the study of the U.S. presidency, as distinguished from the study of U.S. presidents, has been dominated by political scientists.

However, one should not ignore the contribution to literature on the presidency derived from the discipline of psychology, triggered principally by the publication of James David Barber's *Presidential Character* in 1972 (see, e.g., Simonton, 1987; George and George, 1998), nor, during the past decade, the participation of legal scholars in expanding the debate over the nature of executive power consequent upon the controversies generated by the presidency of George W. Bush (see, e.g., Yoo, 2005; Posner and Vermeule, 2010; Goldsmith, 2012; Fisher, 2013).

Until the publication in 1960 of Richard Neustadt's seminal work, *Presidential Power*, scholarship on the U.S. presidency had primarily focused on the constitutional and statutory powers of the President, the various roles and functions of the President, and its relationship to the other branches of government as exemplified in the classic texts of Edward S. Corwin (1957), Clinton Rossiter (1960), and Wilfred Binkley (1962). There has also been a strong normative tradition in the literature the presidency, but this has produced various and divergent diagnoses of the weaknesses of the presidency and no uniform prescription for change, although the British system of cabinet government or variants of it has tended to be among the more popular reform proposals.

Neustadt's *Presidential Power* brought the study of the presidency into contact with the behavioral revolution in postwar political science. Neustadt defined power in terms of the president's persuasive or bargaining ability derived from his 'professional reputation' and 'public prestige.' Asserting that 'powers are no guarantee of power,' he proceeded to show that the leadership ability of presidents was not inherent in the office but was highly personal and dependent on the President's own capabilities as a seeker and wielder of effective influence upon others involved in governing the country. Neustadt's immediate impact was to shift the concerns of presidency scholars away from the formal, legal, and institutional boundaries of presidential power toward a more behavioral and theoretical effort to locate the reality of power and leadership in the White House.

While Neustadt's work was a major break with the tradition of presidency scholarship it did not represent a departure from the basic values of the leading presidency scholars. Indeed, it reinforced those who were firm in their belief that the presidency was and ought to be the center of action in American politics and who advocated a strong and dominant presidency as the only institution capable of responding to the social and political problems facing the country. They represented a generation reared during the presidency of Franklin D. Roosevelt and who made FDR, or what was often referred to as 'the New Deal presidency,' the normative model for their scholarship.

The Changing Focus of Presidency Scholarship

Notwithstanding Neustadt's impact, the political science of the American presidency lagged behind the behavioral revolution for much of the 1960s and 1970s and not until the 1980s did the community of presidency scholars begin to exercise the kind of introspection and self-criticism on which behavioralism in America had thrived. That introspection produced a clear but belated appeal for the study of the presidency to connect with the basic tenets of the behavioral revolution – a more rigorous scientific methodology, the need for quantitative measures, and a greater concern for the conceptual and theoretical underpinnings of presidency scholarship (see, e.g., Edwards and Wayne, 1983). That appeal did result in a redirection of the focus of presidency scholars and, over the past 30 years or so, the political science of the U.S. presidency has taken a quantum leap theoretically, methodologically, and conceptually. A leading contemporary scholar of the presidency, Terry Moe, has described this development as a 'revolution' and 'fundamentally a revolution in theory' (Moe, 2009).

The fundamental values of presidency scholars also changed around the same time. The combined effect of the Vietnam War, Watergate, and the response to 9/11 led to a significant reassessment of the New Deal model of presidential power and even those who continued to advocate strong presidential leadership became more sensitive to notions of accountability and restraint. By the end of the century, presidency scholarship had begun an assault on what Mark Peterson called the 'presidency-centered perspective' on American government (see Peterson, 1990; Jones, 1994; Fisher, 2012).

Current Theory and Research

The epistemological transformation of presidency scholarship from the mid-1980s could not properly be called a revolution because it did not succeed in displacing older, established approaches entirely. The 'public law' tradition, for example, has flourished during the past two decades, in part due to the highly regarded contributions of Louis Fisher (see, particularly, Fisher, 2013) and triggered by the practical relevance of issues such as the war power, the legislative veto, impoundments, military tribunals, and impeachment. Scholars following in the footsteps of Corwin have made significant contributions to the study of the presidency that have not been dependent on theoretical sophistication or quantitative methodology. Similarly, the 'institutional' school has also been reinvigorated by political developments in recent times. Watergate and the Iran-contra affair, for example, acted as a catalyst for substantial research on the role of the presidential staff and the structure and organization of the Executive Office of the President (see Burke, 1992; Hart, 1995; Walcott and Hult, 1995) and, although much of this work is theoretically and conceptually grounded, it also builds on the descriptive and empirical approach that underlies the 'institutional' tradition in presidency scholarship.

The breadth of current research on the U.S. presidency would be difficult to categorize. In addition to those areas mentioned above, it encompasses the study of presidential personality and leadership style, decision-making processes, popularity, professional reputation, bargaining and persuasion, personnel management and the bureaucracy, media management, presidential rhetoric, the impact of public opinion on presidents, the impact of presidents on public policy, and much more. The current scope of the study of the U.S. presidency is well illustrated in the almost-800 pages of *The Oxford Handbook of the American Presidency* (Edwards and Howell, 2009).

Current research also embraces a variety of approaches and techniques which, when combined with the range of subject matter, makes presidency scholarship a broad, diverse, and pluralistic area of study open to the criticism that it lacks coherence in focus and methodology. But it would also be true to say that much current research on the presidency builds upon and extends the seminal research of the past and, while there may not be a single, coherent approach to the political science of the American presidency, there is continuity of research concerns coexisting with a willingness to utilize new methodological and theoretical developments.

Rational choice theorists, however, have staked a claim for the comprehensiveness of their approach and suggested that it may offer presidency scholarship the analytical and theoretical coherence it presently lacks (see Moe, 2009). Applications of rational choice theory to the study of the presidency are growing in number and recent work by rational choice theorists, particularly in the field of presidential–congressional relations, has made a significant impact (see especially Krehbiel, 1998; Cameron, 2000; Beckmann, 2010) and laid the foundations for an extension of that approach to other hitherto less theoretically receptive aspects of the presidency, such as presidential–bureaucratic relations (see Lewis, 2008).

Methodological and Theoretical Problems

Although the cutting edge of presidency research is in the hands of the rational choice theorists and their close companions the new institutionalists, their approach has been contested by other presidency scholars who are less convinced of its general utility across the spectrum of presidency scholarship and who question the desirability of imposing a conceptual and methodological uniformity on future research. The debate is exacerbated by the tendency of the new institutionalists to distinguish between the 'institutional' and the 'personal' presidency, rejecting the latter in the interests of parsimony and theoretical clarity. While it is true that behavioral studies of the 'personal' aspects of the presidency, as exemplified by James David Barber's pioneering work on presidential personality (Barber, 1972), revealed how difficult it is to theorize on the basis of a large number of complex variables, that has yet to be accepted as a sufficient argument for ignoring the complexity of the U.S. presidency.

A more longstanding methodological problem for presidency scholars is the lack of access to their subject of study. Presidents of the United States generally do not give interviews to political scientists, nor do they respond to questionnaires, nor for the most part can they be observed at work, neither do their memoirs or papers usually provide the kind of data that presidency scholars require. This is a principal reason why the study of the presidency has lagged behind other branches of the U.S. government in terms of scientific rigor and why the gaps in the state of knowledge about the presidency are still quite substantial.

Future Directions

During the past 30 years, research on the presidency has been driven in large measure by two principal forces – the behavior of presidents themselves and the events associated with their presidencies, and the imperatives of the political science profession. In all probability, these two forces will likely determine the direction of presidency research in the foreseeable future.

Presidency scholars have often been caught unprepared by the behavior of recent presidents and the events surrounding their presidencies. The Vietnam War was a catalyst for substantial research on the war power. The growth and expansion of the presidential staff system was barely studied before Watergate. The weakness of Jimmy Carter as President provided the impetus for Polsby's (1983) landmark study of the reformed presidential nomination process and its impact on the nature of the presidency. Ronald Reagan's public popularity prompted numerous studies of 'going public' strategies of presidential leadership (see especially Kernell, 1986). As long as occupants of the White House continue to surprise scholars of the presidency they are likely to influence the range and nature of future research.

Similarly, the professional emphasis on theoretical sophistication and scientific rigor is likely to determine the priorities of future research on the presidency as it has done in the past. The danger is that such imperatives may overwhelm and marginalize those aspects of presidency scholarship where theory and quantitative methods are less applicable and less suitable to the object of study.

See also: Bounded Rationality; Executive Government; Institutionalism; Leadership; Presidential Government.

Bibliography

Barber, James D., 1972. The Presidential Character: Predicting Performance in the White House. Prentice-Hall, Englewood Cliffs, NJ.
Beckmann, Matthew N., 2010. Pushing the Agenda: Presidential Leadership in US Lawmaking 1953–2004. Cambridge University Press, New York.
Binkley, Wilfred E., 1962. President and Congress. Random House, New York.
Burke, John P., 1992. The Institutional Presidency. Johns Hopkins University Press, Baltimore, MD.
Cameron, Charles M., 2000. Veto Bargaining: Presidents and the Politics of Negative Power. Cambridge University Press, New York.
Corwin, Edward S., 1957. The President: Office and Powers, fourth rev. ed. New York University Press, New York.
Edwards, George C., Wayne, Stephen J. (Eds.), 1983. Studying the Presidency. University of Tennesse Press, Knoxville, TN.
Edwards, George C., Howell, Williams G. (Eds.), 2009. The Oxford Handbook of the American Presidency. Oxford University Press, New York.
Fisher, Louis, 2012. Teaching the presidency: idealizing a constitutional office. PS: Political Science and Politics 45 (1), 17–31.
Fisher, Louis, 2013. The Law of the Executive Branch: Presidential Power. University Oxford University Press, New York.
George, Alexander L., George, Juliette L., 1998. Presidential Personality and Performance. Westview Press, Boulder, CO.
Goldsmith, Jack, 2012. Power and Constraint: The Accountable Presidency after 9/11. W.W. Norton, New York.
Hart, John, 1995. The Presidential Branch: From Washington to Clinton. Chatham House, Chatham, NJ.
Jones, Charles O., 1994. The Presidency in a Separated System. Brookings Institution, Washington, DC.
Kernell, Samuel, 1986. Going Public: New Strategies of Presidential Leadership. CQ Press, Washington, DC.
Krehbiel, Keith, 1998. Pivotal Politics: A Theory of U.S. Lawmaking. Chicago University Press, Chicago.
Lewis, David E., 2008. The Politics of Presidential Appointments: Political Control and Bureaucratic Performance. Princeton University Press, Princeton, NJ.
McDonald, Forrest, 1994. The American Presidency: An Intellectual History. University Press of Kansas, Lawrence, KS.
Milkis, Sidney M., 1993. The President and the Parties: The Transformation of the American Party System Since the New Deal. Oxford University Press, New York.
Moe, Terry, 2009. The revolution in presidential studies. Presidential Studies Quarterly 39 (4), 702.
Neustadt, Richard E., 1960. Presidential Power: The Politics of Leadership. John Wiley, New York.
Peterson, Mark A., 1990. Legislating Together: The White House and Capitol Hill from Eisenhower to Reagan. Harvard University Press, Cambridge.
Polsby, Nelson W., 1983. Consequences of Party Reform. Oxford University Press, New York.
Posner, Eric A., Vermeule, Adrian, 2010. The Executive Unbound: After the Madisonian Republic. Oxford University Press, New York.
Rossiter, Clinton, 1960. The American Presidency, rev. ed. Harcourt, Brace & World, New York.
Schlesinger, Arthur M., 1973. The Imperial Presidency. Houghton Mifflin, Boston.
Simonton, Dean K., 1987. Why Presidents Succeed: A Political Psychology of Leadership. Yale University Press, New Haven.
Skowronek, Stephen, 1993. The Politics Presidents Make: Leadership from John Adams to George Bush. Harvard University Press, Cambridge, MA.
Walcott, Charles E., Hult, Karen M., 1995. Governing the White House: From Hoover through LBJ. University Press of Kansas, Lawrence, KS.
Yoo, John, 2005. The Powers of War and Peace: The Constitution and Foreign Affairs after 9/11. University of Chicago Press, Chicago.

Presidential Government

Juan J Linz, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from previous edition, volume 17, pp. 12000–12006, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article analyzes presidentialism and semipresidentialism in the context of a comparative study of forms of democratic government. It highlights the consequences of a dual democratic legitimacy of presidents and congress, and enumerates the possible patterns of conflict and/or compromise. The election of presidents and congress for fixed time periods and their mutual independence generates a distinctive rigidity compared to parliamentary regimes. The essay summarizes the negative consequences, basically derived from the above-mentioned defining characteristics, particularly divided government. It also presents the arguments in favor of presidentialism. A number of less essential or variable characteristics are noted, like different forms of election and the implications for the party system and public policy, as well as for the relation to the bureaucracy and the military. A reference is made to the debate about the instability of presidential democracies. In a second section semipresidentialism is defined. The proliferation of such systems in recent times, in many cases dubiously democratic, and the need to distinguish the 'formal' and the 'material' constitution is noted. Formally semipresidential regimes actually function as parliamentary. Particular attention is given to the case of the Fifth French Republic, as the pragmatic case and to the pattern of 'cohabitation' when the president is elected by one party (or coalition) and the prime minister depends on the confidence of a parliament with a majority of a different party or parties. The success in France should not hide situations in which no prime minister enjoys parliamentary support, like at the end of the Weimar Republic in Germany where the president appointed cabinets that could not find support in the Reichstag, a pattern that led to the appointment of Hitler as chancellor.

Democratic governance takes two basic forms: parliamentarism and presidentialism; some people would add semipresidentialism, leaving aside the very unique Swiss institutional structure. Each of them presents a variety of forms that lead to question the dichotomy. There are a few defining characteristics from which many consequences can be derived. Each is identified with a paradigmatic case: the British Westminster parliamentarism, US presidentialism and the semipresidentialism of France. However, as Sartori (1994) notes, parliamentarism takes a variety of forms like the German Federal Republic – the Kanzler democracy – and the assembly parliamentarism of the Third and Fourth French Republics, and contemporary Italy.

Presidential systems are based on an executive with considerable powers in the constitution, and generally with full control of the composition of the cabinet and the administration, elected by the people (directly or by an electoral college elected for that purpose) for a fixed period of time, and are not dependent on a formal vote of confidence by the democratically elected representatives in a parliament. The president is not only the holder of executive power, but is also the symbolic head of state who cannot be dismissed between elections except in rare cases of impeachment.

Two features stand out in presidential systems:

a. Both the president, who controls the executive and is elected by the people (or an electoral college elected by the people for that sole purpose), and an elected legislature (unicameral or bicameral) enjoy democratic legitimacy. It is a system of 'dual democratic legitimacy.'
b. Both the president and congress are elected for a fixed term, the president's tenure in office is independent of the legislature, and the survival of the legislature is independent of the president. This leads to what we characterize as the 'rigidity' of the presidential system.

Most of the characteristics and problems of presidential systems flow from these two essential features. Some other nondefining features of presidentialism are often associated with it, such as term limits or no re-election, automatic succession by a vice-president, more or less extensive rule making, and emergency powers. One characteristic, so normal that it can be included in the definition, is that the presidency is a unipersonal office. There have only been two cases of directly elected pluripersonal 'presidencies': the two-person Cypriot administration (1960–63) and the Uruguayan Colegiado (which governed twice – 1918–33 and 1952–67).

The basic characteristic of the president is the full claim to democratic legitimacy. The claim very often has strong plebiscitary components, although sometimes it is based on fewer popular votes than are received by many prime ministers in parliamentary systems heading minority Cabinets. The presidency gives the incumbent, who combines the qualities of head of state representing the nation and the powers of the executive, a very different aura and self-image, and creates very different popular expectations than those redounding to a prime minister with whatever popularity he might enjoy after his party receiving the same number of votes.

The most striking fact is that in a presidential system, the legislators, particularly when they represent well-organized, disciplined parties that constitute real, ideological, and political choices for the voters, also enjoy a democratic legitimacy, and it is possible that the majority of a legislature might represent a different political choice from that of the voters supporting a president. Who, on the basis of democratic principles, is better legitimated to speak the name of the people: the

president, or the congressional majority that opposes his policies? Since both derive their powers from the vote of the people, a conflict is always latent and sometimes erupts dramatically; there is no democratic principle to resolve it, and the mechanisms that might exist in the constitution are generally complex, highly technical, legalistic, and therefore of doubtful democratic legitimacy for the electorate. It is no accident that in some of those situations the military intervenes as 'poder moderador.'

It could be argued that such conflicts are normal in the USA and have not led to serious crisis. It would exceed the limits of this essay to explain the uniqueness of American political institutions and practices that have limited the impact of such conflicts, including the unique characteristics of the American political parties that lead many American political scientists to ask for a more responsible, disciplined ideological party system.

In a unipersonal presidency there is only one winner who gains full control of the executive. It has been argued that in a Westminster parliamentary system a party winning an absolute majority is also (or even more) a 'winner take all' while in a presidential system it is less likely that a president will have an absolute majority in congress for the entire period of his mandate. However, it is a president, one person that wins control of the executive rather than a party, and that person gains that power irrespective of the size of his plurality or in a runoff in a narrowed competition in which necessarily one candidate will have a majority. The loser in a presidential election ends with no public office in contrast to the leader of a defeated party in a parliamentary election likely to hold a seat in Parliament. Part of the unipersonal election is, in many presidential systems, the succession by a vice-president chosen by the nominee using criteria that would not have been those of the electorate or a party leadership.

The plurality rule remains the most common method for electing presidents. However, the fear of a president being elected by a small plurality has led to introduce a majority runoff election among the two leading candidates (when no candidate has received a majority). The plurality rule (Shugart and Carey, 1992) favors a two-party competition or at least the formation of broad pre-election coalitions and discourages candidates from minor parties. The runoff election, on the other hand, encourages a larger number of candidates in the first run, either in the hope to make it to the runoff (even with a modest number of votes) or place themselves in the condition to bargain for their support by one or the other of the two front runners.

Each procedure can have negative consequences. The plurality rule in a fragmented party system can lead to the election with the support of a small minority. The runoff in addition to encouraging party fragmentation in the first run, can create false coalitions (which will not hold after the election), or mortgages the candidate with allies which may exact a price or distort his image. It also generates in the winner the feeling that he represents the majority of the people, though in the first run he had weak support and his opponent might have been ahead.

In a multiethnic society the fear that a candidate from a populous state could be elected against those from other ethnic-religious communities, has led to a complex system of concurrent pluralities, to assure the election of a president with appeal across ethnic groups (as in Nigeria).

There are a few instances in which the popular vote, in the absence of a decisive majority, does not determine the winner. In Chile, before 1973, the Congress chose the president in the event that no candidate received a majority of popular votes among the top two contenders. Traditionally the one with the largest plurality was elected. In the USA, when in the non-deliberative electoral college no candidate obtains the vote of the majority of electors at the first count, the decision goes to the House of Representatives.

The fixed term of office of both presidents and congress introduces a particular rigidity. A failed president cannot be (normally) forced to relinquish power, but is obliged to continue governing, and he cannot (constitutionally) dismiss a congress committed to oppose his policies. A failed parliamentary prime minister can lose a vote of confidence, be replaced by another leader of his party or coalition, and if he feels that the opposition is wrong, call for an election hoping to discipline his party or coalition and/or revalidate his authority. The rigidity and the length of some presidential mandates leads to the frequent rule of no-reelection, or of no continuous reelections.

The dual legitimacy resulting from separate election of president and the legislature and their mutual independence can lead to 'divided government.' The congress and the presidency may be controlled by different parties (or factions of the same party) creating an impasse. How to govern?

Ackerman (2000) has developed three scenarios. The first emphasizes accommodation, based on bargaining, they may be called the Madisonian hope. Fiorina (1996) argues that American voters desire a divided government because they are uncertain about endorsing an ideological direction offered by either party. Mayhew (1991) argues that divided government has not impeded major legislation. The second involves constitutional breakdown, in which one or another power assaults the system and installs itself as a single lawmaker. A situation of this type is in Latin America, where traditionally the military as 'poder moderador' intervened temporarily, or more recently assuming complete power. Referring to my analysis (Linz, 1994), Ackerman calls this breakdown the 'Linzian nightmare.' A third less dramatic alternative is backbiting, mutual recrimination, and partisan deadlock, which he calls 'crisis of governability.' To the use of constitutional or extraconstitutional decree powers, emergency legislation, patterns described in detail by Carey and Shugart (1998) and their collaborators in a chapter significantly entitled 'Calling Out the Tanks or Filling Out the Forms,' I would add a fourth alternative that explains the absence of recorded conflicts: the application of the 'rule of anticipated reactions.' The avoidance of conflict by not initiating policies is likely to lead to conflict even when desirable, necessary, and proposed by the president at the time of his election.

The disjunction between the image of the voters and the president himself as an initiator of major policies, reforms, and change, and the gridlock, mutual blaming, and/or of a frustrated president or congress confronted with *decretazos*, an avalanche of decrees as *faits accomplis*, reducing the legislature to impotence, does not contribute to political

efficacy and the legitimacy of either or both institutions. The breakdown of democracy and the turn to unconstitutional rule or authoritarianism is only one possible outcome. 'Formal' democratic continuity, particularly in recent crises in Latin America (or the ex-USSR countries) is no proof that the system works.

The situation that Ackerman calls of full authority, when the presidency, the House and Senate are controlled by the same party and there is a favorable majority in the Supreme Court is rare and may last a short time. The result is a burst of legislation, law-making initiatives that will be strong in symbolic statement and difficult to change afterwards, with less concern about implementation. In a parliamentary system, where laws can more easily be changed with electoral defeat, governments are more interested in statutes framed in mid-sized concepts keyed to operational realities, to make them work.

Although the separation of executive and the legislative power, more accurately separate institutions sharing powers, implies that the bureaucracy (and the military) are the domain of the presidency, the fact is that they have two masters. Congress is perhaps increasingly adept to intervene in details of public administration and defense thanks to the prerogative of 'oversight,' of financing departments and programs, the long-standing freedom to investigate. A result is the dependence on both the presidency and the congress and indirectly the separation from both, and subordination to neither. A response to this situation has been 'politicized professionalism': the growing number of political appointees, the large number of in-and-outers, and the short tenure of those appointed who prefer to move to the private sector (and consequently the lack of accumulated experience and continuity in policies). The president turns to 'loyalists' rather than experienced civil servants who would be concerned about their relations with powerful congressmen or women.

In the ongoing debate over the success or failure of presidentialism it is important to distinguish:

a. The basic theoretical issues derived from the defining characteristics of presidentialism (Linz, 1994; Shugart and Carey, 1992).
b. The problems for presidentialism not derived from the basic characteristics; like no reelection, tensions with vice-presidents, mid-term congressional elections, etc.
c. The problems with presidentialism and the actual breakdowns of presidential compared to parliamentary systems. A comparison plagued by a number of methodological problems; time period covered, regional comparisons, level of economic development (OECD vs developing countries), and increasingly formal stability with 'defective' democracies (some actually authoritarian regimes) (Stepan and Skach, 1994; Shugart and Carey, 1992; Thibaut, 1996).
d. The absence of a possibility of a transition to parliamentarism and therefore the limited relevance of a theoretical comparison and the need to consider 'improvements,' modifications of presidentialism that correct some features contributing to some of its failures.
e. The need to keep the problems of semi-presidentialism, offered as a more viable alternative to presidentialism, separate. In this case, the lack of consensus on which countries are semipresidential and the small number of cases (at least before the new postcommunist regimes), makes an analysis like (c) difficult.

The basic argument for the separation of powers is that it provides mutual checks, enhancing freedom, preventing the concentration of power enjoyed by a British prime minister. Mutual checks, the greater difficulty in enacting policies, the opportunity for all interests to be heard and taken into account, the slowing down of the process, and the opportunity for vetoes, generate weak executives and block many of their initiatives, and this is presumably what the voters want when they split their vote for congress and the president. Perhaps this is an advantage in a society that works well independently of public policies, in a wealthy market economy with no major need for redistributive policies, or with independent nonpolitical agencies, but is unlikely to fit societies facing serious problems of underdevelopment and social injustice, requiring more effective government action.

Among the positive aspects of presidentialism its advocates list:

The separation of powers of congress and the executive allows the 'representatives' to act as representatives of interests, particularly of those with a geographical basis, since they are not bound to support the policies of the executive and no party discipline is needed to support a government. In the absence of proportional representation, they enjoy considerable independence from the party leadership and are more likely to represent interests rather than broad policy or ideological alternatives. All politics become local. As Sartori (1994, p. 84) writes: three factors "ideological unprincipledness, weak and undisciplined parties, and locality centered politics, contribute to the working of US presidentialism."

The separate election, particularly mid-term elections, and the limited influence of the president in many cases in the nomination or the making of lists of candidates (especially in federal systems) and in their electoral success, makes the representatives independent. Even their own partisans will owe them almost nothing and often doubt their usefulness in the elections. The selection by primaries has probably reinforced that tendency. This makes members of congress, compared to parliamentarians, better representatives of interests of their constitutents but less reliable in supporting a party (and presidential) policy program. In fact, that 'representative' function is seen (Shugart and Carey, 1992) as one of the positive aspects of presidentialism.

One presumed advantage is the identifiability of presidents, compared to the incertitude of who will head the government in some parliamentary systems based on coalition-making among multiple parties after the election. However, in many parliamentary systems the voter knows well who will head the government if the party he votes for wins. Although presidential candidates identify themselves with certain policy goals, they are less bound to a program than a party and the prime minister it supports, or minor parties that make their coalition support dependent on certain demands (for example, national minority parties).

Shugart and Carey (1992, pp. 44–45) consider the personal identifiability the basis for greater accountability. This might be true, except when reelection is barred and for the blaming game

between the president and congress (particularly in the case of divided government).

Presidential elections allow the citizen to choose one individual to put his trust in – on account of real or perceived qualities and commitments without the mediation of complex institutions: a party, a legislature, or a plurality of actors like parties forming a coalition. That personalization simplifies the choice. It also permits outsiders, not identified with a party, who may not have held any office before, to appeal personally and directly to the voter: Collor, Fujimori, Tyminski, Perrot. That appeal requires access to the media, which in turn may involve access to funding – in some cases personal wealth – and considerable misunderstanding of what is necessary to govern. It may also appeal to 'being above parties,' perceived as divisive, somewhat like a monarch. The election of military leaders: Mannerheim, Hindenburg, De Gaulle, Eanes, even Eisenhower, reflects this desire. TV is particularly congruent with these patterns.

Latin American presidents greatly vary in terms of their constitutional powers, both reactive (veto power, for example) and proactive (decree legislation) and in terms of their partisan support in congress. The possible combinations (Mainwaring and Shugart, 1997, pp. 40–54) are too complex to summarize here.

The difficulties with presidentialism in Latin America have generated two opposite responses: attempts to curtail the powers of the president out of fear of the abuse of power and authoritarian predispositions and efforts to strengthen the presidency as an answer to the weakness of the presidency, particularly in relation to congress.

Many presidential systems have a vice-president, with limited functions except when he has to step in to succeed the president for reasons of death, resignation, impeachment, or incapacity. He generally is not elected independently but on a ticket with the presidential candidate, chosen by him, often for electoral or personal reasons without concern about his or her qualifications. The frequent destructive conflicts between presidents and vice-presidents have led to recent constitutions to not include that office.

Semipresidentialism

Semipresidentialism as a distinctive form of organizing democratic political life, gained importance as a constitutional model and political reality with the consolidation of the French Fifth Republic defined by Sartori (1994, p. 132):

a. The head of state (President) is elected by popular vote – either directly or indirectly – for a fixed term of office.
b. The head of state shares the executive power with a Prime Minister, thus entering a dual authority structure whose three defining criteria are (c) to (e).
c. The President is independent from parliament, but is not entitled to govern alone or directly and therefore his will must be conveyed and processed via his government.
d. Conversely, the Prime Minister and his Cabinet are president-independent in that they are parliament-dependent: they are subject to either parliamentary confidence or no-confidence (or both), and in either case need the support of a parliamentary majority.
e. The dual authority structure of semi-presidentialism allows for different balances and also for shifting prevalences of power within the executive, under the strict condition that the 'autonomy potential' of each component unit of the executive does subsist.

Shugart and Carey (1992) prefer to call this system 'premier presidentialism' and distinguish it from 'presidential parliamentary' (which others include under semipresidentialism) on the basis of the constitutional position of the president.

Since its inception in the constitution of the Weimar Republic (1918) and the Finnish Constitution, it has been the object of considerable dispute. Different terms were used to characterize it: bipolar executive, parliamentary presidential, and Duverger established the term semipresidentialism. The ambiguities surrounding it lead some authors to treat it as a variant of presidentialism, while others see it as a form of parliamentarism. This author, as Sartori, considers it a form *sui generis* distinct from both presidentialism and parliamentarism.

Bartolini (1984) has shown that semipresidentialism was introduced in countries that achieved independence and sought a symbol for the new nation (Ireland, Iceland, and earlier in Finland). In Portugal, it was an attempt to overcome the uncertainties of a transition to democracy.

The French semipresidentialism of the Fifth Republic after the introduction in 1962 of direct presidential election is a two-headed system. It makes possible the alternation – according to some authors – between presidential and parliamentary phases depending on the party or parties of the president having a majority in the chamber. Sartori, rightly, prefers to speak of 'oscillation,' since, in the case of split majorities – described by the French as 'cohabitation' – the president retains some specific powers, particularly in foreign affairs and is not reduced to the status of a president head of state in a parliamentary republic. Nor can a president cum government, that has to govern with and through another body, be assimilated to a 'pure' president that governs alone, that is the government. The oscillation assures considerable flexibility and until now 'cohabitation' has worked relatively well.

Elgin (1999, p. 14) lists countries that he considers semi-presidential: 6 in Western Europe: France, Finland, Portugal, Austria, Iceland, and Ireland; 5 in Central and Eastern Europe: Poland, Bulgaria, Romania, Croatia, Macedonia, and Slovenia; 11 in the former USSR: Russia, Ukraine, Lithuania, Moldavia, and Belarus (of them in the Asian belt: Armenia, Azerbaijan, Kazakstan, Kyrgyzstan, and Uzbekistan; 5 in Asia and the Middle East: Lebanon, Maldives, Mongolia, South Korea, and Sri Lanka; 3 in America: Dominican Republic, Guyana, and Haiti; and 11 in Africa. A total of 42 countries, among which only 18 in 1999 were rated from 1 to 3 on the Freedom House index of political rights; 6 in Western Europe, 4 in Central and Eastern Europe, 1 (Lithuania) in the ex-USSR; 2 in Asia; 1 in the Americas; and 4 in Africa.

The conclusion about regimes that from a formal constitutional perspective and definition can be classified as semi-presidential is that the material constitution (the actual working over time) varies enormously. In some, the president, although popularly elected and with considerable legal powers, acts more like a president in parliamentary regimes. One

difficulty is that a large number of countries that have adopted semipresidential institutions are new democracies, and are politically unsettled (and doubtful democracies), and it is premature to characterize those polities. Moreover, much depends on the personality and political goals of the founding presidents and their successors.

Though all conceptual classification of forms of government cover a wide range of variations, this is even truer about semipresidential given the ambiguities in the constitutions about the powers of presidents and prime ministers, the role of leadership and personality, and the fact that the extended list of such regimes includes many countries dubiously democratic.

On the basis of the long-time actual working of formally semipresidential regimes, Linz (1994) has excluded Austria, Ireland, and Iceland from the list and he and Sartori (1997) would also exclude Portugal after 1982.

The dual legitimacy resulting from the separate direct election of president and parliament persists in semipresidential regimes, but the principle that a prime minister requires the confidence of a majority in the parliament reduces – perhaps even eliminates – the conflicts derived from 'divided government.' The French constitutional practice in the situations of 'cohabitation' supports that interpretation, though one might still ask to what extent it reflects a long tradition of parliamentarism (in addition to other favorable conditions).

We might add to the complexity (and perhaps confusion) a distinction between formally semipresidential regimes in which the president exercises significant powers (and is expected to do so) and those where he has not done so (and is not expected to do so). We obviate the argument that the formal institutional structure is not irrelevant, and that in fact it might at some point become politically relevant. The formally but not materially semipresidential regimes on that account are not parliamentary though they may work as parliamentary. More complicated is the issue – not considered in the earlier studies of this regime type – of the president enjoying such formal and actual powers that the prime minister and his cabinet are fully dependent on the president and the parliament, though formally empowered to question that dominance, is unable to do so. Should such regimes be considered presidential rather than semipresidential? The question can be raised about Russia and Sri Lanka. Again, the complex constitutional rules cannot be ignored but so can the actual working of the system under Yeltsin not be fitted into a 'semipresidential mode.' With such a narrowing down we still would have a number of working stable pure semipresidential regimes.

Presidents generally are identified with a political party, a leader of a party, or after winning the nomination of a party become its leader (yet when defeated lose that position). However, in Russia, Yeltsin, though elected with the support of the parties, opted for not identifying with a party and the constitution formally forbids the president from being a party member. The same is true for other post-communist presidential or semipresidential constitutions.

The idea of president above party, not identified with a party and not campaigning for a party, is likely to make it more difficult for a president to have party support, and does not contribute to the building of strong parties.

Semipresidentialism can work well when the presidential majority and the majority in parliament are of the same party or coalition and when, without being the same, there is a majority in the chamber, supporting a prime minister of a different party (the French 'cohabitation'). The system becomes more problematic as Kalteefleiter (1970) has analyzed in detail when the parliament is unwilling or unable to provide support to the choice of prime minister or of the president but is also unable to support someone of its own choice. This can be the case of a highly fractioned parliament or of a polarized pluralism in which the extremes make it impossible for the center to govern. Parties might be ready to tolerate weak presidential cabinets (abdicating their responsibility) and the president governing by decree. The president, on the other hand, having the power of dissolution, may be tempted to call for elections to solve the impasse and construct a majority. Repeated elections in a climate of crisis might deepen the crisis and contribute to strengthen the antisystem disloyal or semiloyal oppositions. These patterns developed in Germany under the presidency of Hindenburg and his presidential cabinets, contributed to the breakdown of the Weimar Republic.

Editor's Note

Juan Jose Linz was emeritus professor of political science at Yale University and an acknowledged expert on theories of totalitarian and authoritarian systems of government. In addition to his many contributions to political science, he also was a significant figure in post-War political sociology. Linz died October 1, 2013. His aritcle on Presidential Government is being reprinted as a "Legacy entry."

See also: Electoral Systems; Executive Government; Leadership; Parliamentary Government; Parliaments, History of; Party Systems; Presidency, United States.

Bibliography

Ackerman, B., 2000. The new separation of powers. Harvard Law Review 113 (3), 634–729.

Carey, J.M., Shugart, M.S. (Eds.), 1998. Executive Decree Authority. Cambridge University press, Cambridge, UK.

Duverger, M., 1978. Echec on Rev. Albin Michel, Paris.

Duverger, M., 1980. A new political system model: semi-presidential government. European Journal of Political Research 8 (2), 165–187.

Elgie, R. (Ed.), 1999. Semi-presidentialism in Europe. Oxford University Press, Oxford, UK.

Fiorina, M., 1996. Divided Government. Allyn and Bacon, Boston, MA.

Kalteefleiter, W., 1970. Die Funktiones des Staatsoberhauptes in der parlamentcnschen Demokratie. Verlag, Westdeutscher, Köln.

Linz, Juan J., 1994. Presidential or parliamentary democracy: does it make a difference? In: Juan J. Linz, Arturo Valenzuela (Eds.), The Failure of Presidential Democracy. Johns Hopkins University Press, Baltimore, MD.

Linz, J., Stepan, A. (Eds.), 1978. The Breakdown of Democratic Regimes. Johns Hopkins University Press, Baltimore.

Mainwaring, S., 1990. Presidentialism in Latin America. Latin American Research Review 25, 157–179.

Mainwaring, S., 1993. Presidentialism, multipartism, and democracy: The difficult combination. Comparative Political Studies 26 (2), 198–228.
Mainwaring, S., Shugart, M.S. (Eds.), 1997. Presidentialism and Democracy in Latin America. Cambridge University Press, Cambridge, UK.
Mayhew, D.R., 1991. Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–1990. Yale University Press, New Haven, CT.
Sartori, G., 1994. Comparative Constitutional Engineering: An Inquiry into Structures, Incentives, and Outcomes. New York University Press, New York.
Serraforo, M.D., 1999. El poder y su sombra. los vice presidentes. Editorial de Belgrono, Buenos Aires.
Shugart, M.S., Carey, J.M., 1992. Presidents and Assemblies: Constitutional Design and Electoral Dynamics. Cambridge University Press, New York.
Stepan, A.C., Skach, C., 1994. In: Linz, Valenzuela (Ed.), Presidentialism and parliamentarism in comparative perspective, p. 36.
Thibaut, B., 1996. Prësidentialismus und Demokratie in Lateinamerika. Leske and Budrich, Opladen, Germany.

Pretend Play and Cognitive Development

Mark Nielsen, The University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

When children pretend, their behavior is driven by their imagination and capacity for seeing the world in a way that is contrary to how really it is. They use objects in ways that are inconsistent with their intended purpose, attribute items with properties they do not have, and act as someone they are not. Doing so is a unique human universal that is associated with a range of core social cognitive achievements. Here, evidence is outlined for the cultural universality of pretense alongside explication of the links between pretending and other developmental domains.

> … a man does not learn through any kind of instruction or study in later life anything like so much as the child learns in the first four years of his careless existence, through the perceptions and ideas acquired in his play.
> – Preyer (1893)

Every family of mammals shows some evidence of play in some species (Burghardt, 2005). In humans, play emerges early in ontogeny, and in its initial expression, it is characterized by the functional treatment of objects where infants attend to cause-and-effect relations with objects, focusing on literal and intended uses, such as pushing a toy on wheels or placing building blocks together. Toward the end of the second year, infants begin to demonstrate an interest in applying their knowledge base to symbolically manipulate objects and their properties, and to allow their imagination, rather than the stimulus itself, to dominate their behavior: They begin to pretend (Harris, 2000; Leslie, 1987).

When children pretend, their play becomes characterized by the symbolic, nonliteral transformation of the here-and-now whereby the world, in the context of a play sequence, gets treated contrary to reality. A banana is talked into as if it is a telephone. An empty cup gets 'sipped' from as if it contains hot tea. A red block becomes a poisonous fish (see **Figure 1**). With development, children's pretense incorporates ever more complex and cognitively demanding routines, as simple behaviors directed toward the self translate to more sophisticated actions involving sequences and object substitution. During this time, the meanings that are associated with pretend play become augmented by a differentiation from real-life conditions and situations, persons, and objects that are immediately present.

Many developmental psychologists are motivated to search for the ontogenetic origins of behaviors that define the human condition. The less culturally specific a behavior is and the more it has demonstrable, replicable links to other core domains of development, the more likely it is to be considered in this light. Pretend play is rare in that it satisfies both criteria. In the following sections, I first outline evidence for the cultural universality of pretend play and then follow by detailing aspects of young children's social and cognitive development that are intimately tied to their engagement in and expression of pretense.

Pretend Play in a Cross-Cultural Context

As a discipline, psychology has been criticized for making claims about human behavior that are staked in ignorance of the vast majority of the world's population (Arnett, 2008; Henrich et al., 2010). To some degree, this is true of the pretend-play literature, where cross-cultural studies remain a minority. There is, however, a sufficiently broad base from which valid generalizations can be made. Foremost among these is that engaging in pretense can be genuinely considered a human universal. Children of !Ko and !Kung hunter-gatherers in Southern Africa will pretend to be animals and the adults who hunt them; melons will be used to make dolls that are treated as babies (Sbrzesny, 1976, as cited in Konner, 2010). Among the agricultural Kpelle of Liberia, children act out roles they see adults engaged in, such as being a blacksmith (Lancy, 1996). Martini (1994) reported that children living in the French Polynesian Marquesas Islands

Figure 1 Typical pretend act of young Western children. In this sequence, a 3-year-old boy (left) explains to his 6-year-old brother that the red plastic blocks are 'poisonous fish.'

pretended to paddle canoes, hunt, and fish, while Roopnarine et al. (1994) noted that children in India will act out traditional celebrations and folktales. Indeed, wherever evidence of pretense has been sought, it has been found. There are differences in what children do and in the frequency and duration with which they do it, but what is beyond question is that no matter the environment in which they grow, children spend a considerable portion of their early years pretending.

Pretend Play in Five Distinct Brazilian Communities

Providing clear examples of cross-cultural similarities and differences, Gosso et al. (2007) indexed the naturalistic free play of 4–6 year olds living in five distinct cultural communities in Brazil: Parakanã Indians (from Paranowaona in the Amazon region in the north), seashore children (from Ubatuba, a small town of the southeast), and urban groups of low, mixed, and high socioeconomic status (SES) from São Paulo, a metropolis of the southeast.

Critically, in terms of the proportion of time spent playing, pretense was among the highest of all play types for children in all cultural groups. Children in all groups also engaged in more acts of symbolic modification (this included acts in which the self or another was attributed an identity different from reality, using an object to stand in for another, attributing properties to objects that the children do not have, and pretending to be in a noncurrent physical state, e.g., being sick) than symbolic creation (imagining something is present when it is not, e.g., pretending to be in a spaceship or taking a shower with soap in the absence of any substratum). Finally, when engaged in socio-dramatic play, all children mainly assumed *the identity of family members* (e.g., a 5-year-old seashore girl pretended to be the mother of another girl, ordering her to buy milk, candy, chocolate, and cake at the grocery store). The second girl then went to a bench, pretended it was the grocery store, and returned showing an imaginary bag, asking, ("Was it just that that I was supposed to buy?"), followed by *animals* (e.g., a 5-year-old Indian girl said she was a turtle, 'jumping' into the 'river' and 'swimming') and *fantastic characters* (e.g., a 5-year-old mixed-SES urban boy put an orange cloth on his shoulders and invited a 4-year-old peer to play vampire, saying, "I have been wearing this cloth for more than a thousand years. I could borrow it to you").

Along with these similarities, clear differences were found. For example, the high-SES children showed a greater proportion of symbolic modifications than seashore, Indian, and low-SES children, with the mixed-SES group showing a higher frequency than Indians. The high-SES urban children also engaged in more 'creation' than low-SES children and Indians, with the mixed-SES children falling in between. Among Indians, object property modification (e.g., using a toy wooden boat for a real boat) was more frequent in boys, whereas object identity (e.g., using a lemon in place of soap or a leaf for a washbasin) predominated in girls. A similar pattern was observed among seashore and mixed-SES urban children, but character identity modification (e.g., pretending to be a mother, child, horse, or dog) was more common in this group than object identity modification. Gosso et al. (2007) also found that transportation themes featured prominently among Indian children. Interestingly, although they are common and readily available for all members of the community to use, canoes were rarely included during pretend-acts by these children. In contrast, they often pretended to drive, pickup trucks and motorboats, both of which are status symbols being needed to make long-distance travels. Similarly, driving a car was a common theme for children living in the low-SES urban community, where having a car is an indication of heightening of social class.

Pretend Play in Three Geographically Distinct Communities

In a more recent study, 1–3-year-old children's expression of and engagement in pretend play were investigated in three distinct communities: in two traditional, small-scale cultural settings in rural Peru and rural India, and in a typical Western, middle-class cultural setting in Canada (Callaghan et al., 2011). As in Gosso et al. (2007), similarities and differences were found. Asked when their children began to engage in pretense, most mothers suggested at around 2–3 years of age, regardless of their cultural heritage. Mothers also reported that when their children started to pretend, social role-play was the most common form. However, whereas the Canadian and Peruvian mothers said their children often pretended with others and would commonly use one object to stand in for another, Indian mothers said their children usually only pretend when on their own and rarely engage in object substitution.

To directly test for pretend abilities, Callaghan et al. (2011) gave children 2 min to play with a range of everyday items as well as sets of toys considered either conventional (e.g., a cloth doll with clothes, crockery and cutlery, and replica cars and animals) or nonconventional (e.g., a stuffed sock, cork coasters, shower mounting unit, and door stop). Once children had played with one set of toys, the experimenter demonstrated two symbolic actions with them (e.g., giving the doll a drink of juice and dancing the doll around the table, or smashing two cars together and having a close encounter with another car at an intersection) and encouraged them to play again. Children were then given the second set and encouraged once again to play the 'car' or 'doll' game with the new set of toys. Irrespective of background, children's rates of symbolic actions were higher for conventional than nonconventional toy types. However, in terms of overall rates of production, the Canadian children produced symbolic-acts with the nonconventional toys approximately five times more frequently and with the conventional toys about 2.5 times more frequently than children in India or Peru. With the nonconventional toys, over 40% of the Canadian children produced at least one symbolic act, whereas less than 10% of the children from the rural groups did so.

Callaghan et al. (2011) also investigated whether or not children would respond differently to an experimenter's attempts at either pouring from a container or eating a food item when these actions were couched as 'pretend' (where the experimenter adopted a pretense attitude while modeling the actions) or as 'trying' (where the experimenter demonstrated the same actions but conveyed effort rather than pretense).

For the trying trials, in terms of simple imitations of the experimenter's actions (e.g., trying to pour water from the closed container into the cup) and actions that went beyond those demonstrated (e.g., trying to open the lid of the container before pouring), children in all cultural settings performed at a high level. Notably, though, Peruvian children were moderately lower in their performance relative to Canadian and Indian children. However, in the pretend trials Canadian children performed considerably more actions, whether replications of what the experimenter modeled (e.g., pretending to bite the orange) or actions that went beyond those demonstrated (e.g., pretending to 'pour' into a container and then giving a 'drink' to a teddy bear) than Indian and Peruvian children. Moreover, the average age at which they produced at least one target pretend act was lower for the Canadian children (34.4 months) than the Indian (46.5 months) and Peruvian (45.8 months) children.

Summary

The aforementioned studies highlight how children's cultural background can have an important influence on their pretense, influencing factors such as how play is engaged in, who is played with, and what is used to play with. The differences in pretense outlined throughout this section are telling, but they are likely due to environmental factors. In explaining the lowered rates of pretense that they found in the Indian and Peruvian children, Callaghan et al. (2011) suggest that this is likely due to the almost absent input from supportive adults. They note that their interview data suggest that only the Canadian children engaged in frequent, rich, interactive pretend-play episodes with their mothers. In this context, cultural differences are attributed to differences in scaffolding, not competence or mental ability. Similarly, Gosso et al. (2007) argue that differences in the expression of pretense are environmentally bound and that "there are no differences in cognitive functioning of the children of various cultural backgrounds" (see also Göncü et al., 2000).

It has been argued that the emergence of childhood as a novel life stage is one of the most critical steps in the evolution of our species (Bogin, 1990; Locke and Bogin, 2006; Nielsen, 2012). Regardless of the veracity of this argument, what is clear is that irrespective of the environment in which they flourish, childhoods, the world over are, at least in part, spent pretending. To some, pretense may appear to be the pastime of a mind unencumbered by more serious, more vital pursuits. But its global pervasiveness suggests otherwise, something borne out by a large body of literature documenting links between the exhibition of pretense and a range of important social-cognitive behaviors.

Developmental Correlates of Pretend Play

In order to pretend, children must imagine something that is currently not true, then behave as if it were (e.g., by 'drinking hot tea' from an empty cup). When placed in the context of an extended scenario (e.g., having a 'tea party with friends'), the imagined world, along with the possible consequences of any behavior, has to be considered and maintained. If a cup is tipped over, whatever is 'spilled' should be 'cleaned up.' By pretending, children thus develop a capacity to generate and reason with novel suppositions and imaginary scenarios. The ability to pretend is thus afforded a special role in development, a role emphasized by associations between children's engagement in pretend play and the emergence and establishment of a host of other critical skills. This section outlines some of these.

Social Competence

In a large-scale study of the social and cognitive associations with young children's pretend play, Connolly and Doyle (1984) observed 91 children aged between 35 and 69 months engaged in free play at their child care centers. The occurrence and complexity of pretend-acts exhibited while interacting with others were correlated with (1) teacher ratings of children's competence in interacting with peers, (2) teacher ratings of how popular children were with other students, (3) children's ability to correctly identify how characters in a series of emotional stories might feel (e.g., a child's bag of candy is taken by another child), and (4) the amount of time spent in social interactions. Children's engagement in pretense did not correlate with a range of less socially oriented measures, such as (1) their competence in meeting the demands and regulations of the classroom, (2) their ability to select from an array of items those that would make appropriate gifts for different individuals (mother, father, teacher, and opposite-sex peer), and (3) the type of affective mood evident (rated as positive, negative, or neutral) while interacting with others. Whilst cautioning against interpreting causation from correlation, Connolly and Doyle suggest their findings which indicate that social pretend play is the most predictive of the skills and competencies that are apparent in children's social interactions with their peers, and that pretending provides a framework for the development of social skills (see also Taylor et al., 2004).

Research by Uren and Stagnitti (2009) explored pretend-play as a facilitator for social competence and involvement in school activities. Children were assessed with various measures to determine their engagement in pretend play, interactions with peers, and involvement in school activities. They reported that a higher complexity in pretend play (i.e., the ability to perform detailed and highly elaborate pretend interactions with conventional and nonstructured toys) was related to greater engagement with peers and better social abilities. Alternatively, children with poor pretend-play skills showed more disruptive behaviors in their interactions with peers. Children with complex pretend-play skills were also more involved in their classroom activities and showed greater cognitive abilities during play. These abilities included the introduction and extension of ideas and stories throughout play, and a higher level of verbalization.

These studies highlight the link between children's developing social skills and their engagement in pretend play. While the studies provide no foundation for assuming causality, intervention programs have exploited this link (e.g., O'Connor

and Stagnitti, 2011), using play as a successful means to improve young children's social skills.

Language

According to Piaget (Piaget, 1962), during the majority of the sensorimotor period, infants operate at a level of pre-representation, whereby behavior based on the evocation of an absent object (i.e., representation in thought) is not observed. However, at the end of the sensorimotor period (about $1^1/_2$ –2 years of age), the ability to represent something by means of a 'signifier' emerges. This provides infants with a medium for going beyond their current veridical perceptions and permits the appearance of a series of significant symbolic skills, including the ability to treat the real world in a nonliteral manner (i.e., pretend play). Werner and Kaplan (1963) saw the development of language in a similar light, conceiving language and play as outcomes of the general progress of the semiotic function.

In keeping with Piagetian theory, a number of studies have documented an association between pretend play and language development in childhood (e.g., Bates and Thal, 1991; Eisert and Lamorey, 1996; Lewis et al., 2000; McCune, 1995; Ungerer and Sigman, 1984). Some of these studies reported associations between pretend play and language comprehension prior to the middle of the second year (Lyytinen et al., 1997; Tamis-LeMonda and Bornstein, 1993, 1994). Hence, it may be that language comprehension scaffolds the pretense abilities of infants and young toddlers by providing them with a means of better understanding the overtures of others. Indeed, Kavanaugh et al. (1983) reported that, while playing with their infants aged between 12 and 21 months, carers almost exclusively initiated pretense episodes by verbally attributing properties to objects (see also Rakoczy et al., 2005; Striano et al., 2001).

This idea of attributing properties to objects relates to what is termed 'object substitution,' a core domain of pretense in which an object is treated in a manner other than its normal use (e.g., a banana is used as a telephone) or is designated as being something that it is not (e.g., a wooden block is labeled a 'fast car'). This type of pretense emerges in the second half of the second year of life and is initially constrained by the substituted objects being relatively simple in shape, having minimal surface details and being geometrically similar to the replaced object (Smith and Jones, 2011). A wooden block may be a 'fast car,' but a book cannot. To study whether this form of pretense is linked to language development through an emerging capacity to identify object properties, Smith and Jones (2011) tested children aged 18–30 months for productive vocabulary, object recognition (where they were asked to identify which of the three objects were named by an experimenter), and object substitution abilities (where children had to choose, from an array of four items, one that could be used in place of another – e.g., using a block of wood as a 'bed' for a doll). For the object tasks, children were tested using both rich and sparse (minimally representational) objects. Smith and Jones found a strong positive relation between children's sparse object recognition scores and the numbers of target object substitutions they produced, and that changes in visual object recognition were related to noun vocabulary size. For Smith and Jones, these associations indicate that while object substitutions are not causally related to language delay, their absence may be an easily detected signal of a problem in language acquisition.

Theory of Mind

The purported link between pretense and language was subject to increased empirical scrutiny following a proposal by Leslie (1987) in which he argued that children's ability to pretend was associated with the development and elaboration of the mental state lexicon and the subsequent emergence of a theory of mind. The term 'theory of mind' refers to a kind of 'commonsense understanding' of the world that involves the appreciation that oneself and others are beings who possess a range of mental states, such as thoughts, beliefs, and desires, and that it is these mental states that determine behavior. We act toward the world, not as it really is, but how we perceive or believe it to be. I may open my umbrella because I think it is about to rain. Whether it really is or is not about to rain is irrelevant – my behavior is driven by my belief that it will rain.

Tasks used to assess understanding of 'false beliefs' in others have become the standard tool used to index the child's developing 'theory of mind.' For example, a child might be shown a box that would typically contain a popular treat (e.g., an M&Ms box). When given the box to explore, the child discovers that it contains something unexpected (e.g., pencils). They are then asked what someone else who has never looked inside the box would expect to find. Children at around 3 years of age suggest that someone else would expect to find what they themselves know is inside the box (e.g., pencils), whereas by 5 years of age most children answer that someone else would reasonably expect to find contents reflecting what is depicted on the cover (e.g., M&Ms). Success on these tasks depends on the child's ability to distinguish between a mental representation (e.g., the belief of 'X' that the contents of a box reflect what is advertised on the packaging) and the actual state of the world (the box, unbeknown to 'X,' contains something unexpected).

Leslie (1987) maintained that to perform the nonliteral mental transformations that characterize pretense, children must first be able to simultaneously appreciate two alternative and contradictory models of reality such that the literal world is prevented from being corrupted by what occurs in the fantasy world. In a similar way, the idea of mental states cannot be discovered through either language learning (where conceptual understanding must necessarily precede linguistic expression) or perceptual evidence (given that mental states are unobservable). For Leslie, through the ability to appreciate the dual representations that are first evident in pretense, the pre-school child begins to relate expressions to the language of thought. Furthermore, it is this capacity for representational reasoning that the child subsequently exploits when constructing a theory of mind. The critical set of abilities contributing to the appreciation of belief and, hence, to the development of a mature theory of mind typically emerge between 3 and 4 years of age (Wellman meta-analysis). However, as noted in this article, children typically begin to engage in pretend acts from 18 months of age onward. In response to this apparent

asynchrony in development, Leslie (1987, 1988) has claimed that pretend play is in fact an early manifestation of the requisite abilities necessary for the development of a theory of mind.

In line with Leslie (1987), multiple studies have documented positive correlations between various indices of pretend play and children's success on false-belief tasks (for a summary, see Smith, 2010, p. 186, Table 9.1). For example, Taylor and Carlson (1997) assessed 3- and 4-year-old children's theory of mind development (as indexed by a series of 13 tasks, including false belief) and their involvement in pretend play. They found that variance in 4 year olds' performance on the theory of mind tasks was explained by both the creation of an imaginary character and a composite score derived from 20 fantasy play measures. Both results were maintained after the effects of chronological age and verbal intelligence were controlled for. However, of the fantasy measures, variance in theory of mind was best accounted for by three variables – the performance of imaginary-object pantomimes, the use of blocks as props in pretend play rather than as functional items, and the ability to impersonate another person. Furthermore, Taylor and Carlson comment that children who pretended to be a machine performed worse on theory of mind tasks than children who did not, thus proposing that gaining an appreciation of the causality of mental states obtained through impersonation occurs only in role-play involving animate beings.

Creativity

On 20 July 1969, in what must be counted as one of the crowning technological achievements of our species, Neil Armstrong and Buzz Aldrin stepped onto the surface of the moon. This occurred less than a century after the first manned heavier-than-air craft successfully took flight. During this same half century, the nearest that chimpanzees, our closest living animal relatives, got to flight was when they swung through the trees. While we invented tools to build spacecraft, their tool use has not radically changed from what was employed some 7 million years ago when we shared a common ancestor. But before we could fly to the moon, we needed the idea of rockets that could fly to the moon. We needed a mind capable of remarkable imagination and creativity – a mind that, for a number of theorists, develops through the pretense that occurs in children's play, where they get to practice the creative process that underpins innovation in adulthood (Carruthers, 2002, 2006; Pellegrini et al., 2007).

The idea that pretense and innovation are linked is not new (Almy, 1967; Bruner, 1972; Lieberman, 1977; Piaget, 1962; Sutton-Smith, 1967). Testing this purported link Dansky and Silverman (1973) gave a group of 4- to 6-year-old children 10 min to play with a set of objects that included paper towels, a screwdriver, empty matchboxes, and paper clips. An experimenter showed other children in another condition what could be done with these objects (e.g., turning screws with the screwdriver or putting small sticks into empty matchboxes). These children were subsequently asked to copy what the experimenter had done. Children in a third condition were given a box of crayons and four pictures to color in. Regardless of what condition they were in, each child was subsequently asked to suggest alternate uses for the four stimulus objects. Responses were scored according to whether they were standard (i.e., encompassing what the object had been primarily designed for, e.g., suggesting that the paper towel could be used to wipe up something) or nonstandard (i.e., suggestions that would not be typically associated with the object's design function, such as using a matchbox to 'make a TV' or using the screwdriver as 'a merry-go-round'). Children in the play condition produced significantly more nonstandard responses for each object, whereas there was no difference in the mean number of standard responses given (for replication, see Dansky, 1980; Dansky and Silverman, 1975; Li, 1978; for criticism, see Smith and Whitney, 1987).

In a study by Mullineaux and DiLalla (2009), the relationship between pretend play and later creativity was examined by comparing realistic role-play (e.g., pretending to be another person) at age 5 to divergent thinking at ages 10–15. This longitudinal design consisted of a free-play session for the initial measure of pretend-play role-playing, and two measures of creativity (i.e., divergent thinking tasks) at follow-up. The two creativity measures involved having the participant list various functions for eight objects and finish an incomplete drawing. It was reported that more realistic role-play at the initial observation was related to a greater ability to list alternative uses for items at follow-up. However, this relationship was not present for the drawing task.

Indirect support for pretense being the bedrock of creativity comes from the literature on children with autism who display impairments in both domains (Craig and Baron-Cohen, 1999; Jarrold, 2003). Evidence also comes from children who grow up in Old Order Mennonite communities, where a traditional rural existence is embraced and modern technology and innovations are shunned. In stark contrast to children raised in less conservative religious environments, the pretense of these children predominantly consists of acting out real-life roles with little integration of imaginary objects (Carlson et al.,1998). Finally, children with imaginary companions show a strong inclination towards engaging in fantasy-themed play and show higher levels of creativity than children who do not have imaginary companions (Hoff, 2005; Schaeffer, 1969; Taylor, 1999). Adults who recall having an imaginary companion as a child have also been shown to outperform those who do not on a variety of creativity indices (Gleason et al., 2003; Kidd et al., 2010; Taylor et al., 2003).

Summary

Pretend play is a pivotal activity for young children that can aid their cognitive, social, physical, and emotional development (Ginsburg, 2007). Indeed, engaging in pretend play has been linked to a range of emerging skills that go beyond those noted here, such as group work, self-advocacy, leadership, self-regulation, and negotiating tactics (Erikson, 1985; Hurwitz, 2002; Vieillevoye and Nader-Grosbois, 2008). Moreover, there are few behaviors that have returned such consistent, replicable, multilab associations as pretend play

has in regard to the wide range of key developmental domains. Of course, arguments about causation must be treated with extreme caution. But this should not undermine the importance of pretending to young children's development.

Conclusion

Instinctively, children like to play. They have an innate curiosity and a healthy sense for adventure, exploration, and excitement. Play is the early classroom, where children learn how to socialize, show initiative, negotiate conflicts, learn about rules, and communicate clearly. When their play begins to be populated by pretense, they are embarking on a journey with children the world over, a journey no other animal takes. Its expression may differ from child to child, its themes and structural components may change from community to community, but for many of the world's children pretending remains pervasive, intuitive, and engaging; and its absence (or limited expression) constitutes a marker of disrupted or atypical development. There are numerous domains or behaviors that are thought to be core components of young children's social and cognitive development (e.g., language and theory of mind). A claim can be staked that pretending is among the most crucial.

See also: Cognitive Development: Child Education; Cross-Cultural Research Methods in Psychology; Early Emotional Development and Cultural Variability; Evolutionary Developmental Psychology; Imitation During Infancy and Early Childhood; Intentionality in Language and Communication, Emergence of; Mind, Theories of; Personality and Creativity; Play During Infancy and Early Childhood: Cultural Similarities and Variations; Play, Anthropology of; Scientific Concepts: Development in Children; Self-Development in Childhood and Adolescence; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Almy, M., 1967. Spontaneous play: an avenue for intellectual development. Young Children 22, 265–276.
Arnett, J.J., 2008. The neglected 95%: why American psychology needs to become less American. American Psychologist 63, 602–614.
Bates, E., Thal, D., 1991. Associations and disassociation in language development. In: Miller, J. (Ed.), Research on Child Language Disorders: A Decade in Progress. Pro-Ed, Austin, TX, pp. 145–168.
Bogin, B., 1990. The evolution of human childhood. Bioscience 40, 16–25.
Bruner, J., 1972. Nature and uses of immaturity. American Psychologist 27, 687–708.
Burghardt, G.M., 2005. The Genesis of Animal Play. MIT Press, Cambridge, MA.
Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., et al., 2011. Early social cognition in three cultural contexts. Monographs of the Society for Research in Child Development 76 (2, Serial No. 299).
Carlson, S.M., Taylor, M., Levin, G.R., 1998. The influence of culture on pretend play: the case of Mennonite children. Merrill-Palmer Quarterly 44, 538–565.
Carruthers, P., 2002. Human creativity: its cognitive basis, its evolution, and its connections with childhood pretence. British Journal for the Philosophy of Science 53, 225–249.
Carruthers, P., 2006. The Architecture of the Mind: Massive Modularity and the Flexibility of Thought. Oxford University Press, New York.
Connolly, J.A., Doyle, A.B., 1984. Relation of social fantasy play to social competence in preschoolers. Developmental Psychology 20, 797–806.
Craig, J., Baron-Cohen, S., 1999. Creativity and imagination in autism and Asperger syndrome. Journal of Autism and Developmental Disorders 29, 319–326.
Dansky, J.L., 1980. Make-believe: a mediator of the relationship between play and associative fluency. Child Development 51, 576–579.
Dansky, J.L., Silverman, I.W., 1973. Effects of play on associative fluency in preschool-aged children. Developmental Psychology 9, 38–43.
Dansky, J.L., Silverman, I.W., 1975. Play: a general facilitator of associative fluency. Developmental Psychology 11, 104.
Eisert, D., Lamorey, S., 1996. Play as a window on child development: the relationship between play and other developmental domains. Early Education and Development 7, 221–235.
Erikson, R.J., 1985. Play contributes to the full emotional development of the child. Education 105, 261–263.
Ginsburg, K., 2007. The importance of play in promoting healthy child development and maintaining strong parent-child bonds. Pediatrics 119 (1), 182–192.
Gleason, T.R., Jarudi, R.N., Cheek, J.M., 2003. Imagination, personality, and imaginary companions. Social Behavior and Personality 31, 721–738.
Göncü, A., Mistry, J., Mosier, C., 2000. Cultural variations in the play toddlers. International Journal of Behavioral Development 24, 321–329.
Gosso, Y., Morais, M., Otta, E., 2007. Pretend play of Brazilian children: a window into different cultural worlds. Journal of Cross-Cultural Psychology 38, 539–558.
Harris, P.L., 2000. The Work of the Imagination. Blackwell, Oxford.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. Most people are not WEIRD. Nature 466, 29.
Hoff, E., 2005. Imaginary companions, creativity, and self-image in middle childhood. Creativity Research Journal 17, 167–180.
Hurwitz, S.C., 2002. To be successful: let them play! Child Education 79, 101–102.
Jarrold, C., 2003. A review of research into pretend play in autism. Autism 7, 379–390.
Kavanaugh, R.D., Whittington, S., Cerbone, M.J., 1983. Mother's use of fantasy speech to young children. Journal of Child Language 10, 45–55.
Kidd, E., Rogers, P., Rogers, C., 2010. The personality correlates of adults who had imaginary companions in childhood. Psychological Reports 107, 1–10.
Konner, M., 2010. The Evolution of Childhood. The Belknap Press of Harvard University Press, London.
Lancy, D.F., 1996. Playing on the Mother-Ground: Cultural Routines for Children's Development. Guilford Press, New York/London.
Leslie, A., 1987. Pretence and representation: the origins of "theory of mind". Psychological Review 94, 412–426.
Leslie, A., 1988. Some implications of pretence for mechanisms underlying the child's theory of mind. In: Astington, J.W., Harris, P.L., Olson, D.R. (Eds.), Developing Theories of Mind. Cambridge University Press, New York, pp. 19–46.
Lewis, V., Boucher, J., Lupton, L., Watson, S., 2000. Relationships between symbolic play, functional play, verbal and non-verbal ability in young children. International Journal of Language and Communication Disorders 35, 117–127.
Li, A.K.F., 1978. Effects of play on novel responses in kindergarten children. Alberta Journal of Education Research 24, 31–36.
Lieberman, J.N., 1977. Playfulness: Its Relationship to Imagination and Creativity. Academic Press, New York.
Locke, J.L., Bogin, B., 2006. Language and life history: a new perspective on the development and evolution of human language. Behavioral and Brain Sciences 29, 259–325.
Lyytinen, P., Poikkeus, A.M., Laakso, M.L., 1997. Language and symbolic play in toddlers. International Journal of Behavioral Development 21, 289–302.
Martini, M., 1994. Peer interactions in Polynesia: a view from the Marquesas. In: Roopnarine, J.L., Johnson, J.E., Hooper, F.H. (Eds.), Children's Play in Diverse Cultures. State University of New York Press, Albany, pp. 73–103.
McCune, L., 1995. A normative study of representational play at the transition to language. Developmental Psychology 31, 198–206.
Mullineaux, P.Y., DiLalla, L.F., 2009. Creative preschool play behaviors and early adolescent creativity. Journal of Creative Behavior 43, 41–57.
Nielsen, M., 2012. Imitation, pretend play and childhood: essential elements in the evolution of human culture? Journal of Comparative Psychology 126, 170–181.
O'Connor, C., Stagnitti, K., 2011. Play, behaviour, language and social skills: the comparison of a play and a non-play intervention within a specialist school setting. Research in Developmental Disabilities 32, 1205–1211.
Pellegrini, A.D., Dupuis, D., Smith, P.K., 2007. Play in evolution and development. Developmental Review 27, 261–276.
Piaget, J., 1962. Play, Dreams, and Imitation in Childhood. Norton, New York.
Preyer, W., 1893. Mind of the Child. In: Development of the Intellect, vol. II. Appleton, New York.

Rakoczy, H., Tomasello, M., Striano, T., 2005. On tools and toys: how children learn to act on and pretend with "virgin objects". Developmental Science 8, 57–73.

Roopnarine, J.L., Hossain, Z., Gill, P., Brophy, H., 1994. Play in the East Indian context. In: Roopnarine, J.L., Johnson, J.E., Hooper, F.H. (Eds.), Children's Play in Diverse Cultures. State University of New York Press, Albany, pp. 9–30.

Schaeffer, C.E., 1969. Imaginary companions and creative adolescents. Developmental Psychology 1, 747–749.

Smith, L.B., Jones, S.S., 2011. Symbolic play connects to language through visual object recognition. Developmental Science 14, 1142–1149.

Smith, P.K., 2010. Children and Play. Wiley-Blackwell, Chichester, UK.

Smith, P.K., Whitney, S., 1987. Play and associative fluency: experimenter effects may be responsible for previous positive findings. Developmental Psychology 23, 49–53.

Striano, T., Tomasello, M., Rochat, P., 2001. Social and object support for early symbolic play. Developmental Science 4, 442–455.

Sutton-Smith, B., 1967. The role of play in cognitive development. Young Children 22, 361–370.

Tamis-LeMonda, C.S., Bornstein, M.H., 1993. Play and its relations to other mental functions in the child. New Directions for Child Development 59, 17–28.

Tamis-LeMonda, C.S., Bornstein, M.H., 1994. Specificity in mother-toddler language-play relations across the second year. Developmental Psychology 30, 283–292.

Taylor, M., 1999. Imaginary Companions and the Children Who Create Them. Oxford University Press, New York.

Taylor, M., Carlson, S.M., 1997. The relation between individual differences in fantasy and theory of mind. Child Development 68, 436–455.

Taylor, M., Carlson, S.M., Maring, B.L., Gerow, L., Charley, C.M., 2004. The characteristics and correlates of fantasy in school-age children: imaginary companions, impersonation, and social understanding. Developmental Psychology 40, 1173–1187.

Taylor, M., Hodges, S.D., Kohanyi, A., 2003. The illusion of independent agency: do adult fiction writers experience their characters as having minds of their own? Imagination, Cognition, and Personality 22, 361–380.

Ungerer, J.A., Sigman, M., 1984. The relation of play and sensorimotor behavior to language in the second year. Child Development 55, 1448–1455.

Uren, N., Stagnitti, K., 2009. Pretend play, social competence and involvement in children aged 5–7 years: the concurrent validity of the Child-Initiated Pretend Play Assessment. Australian Occupational Therapy Journal 56, 33–40.

Vieillevoye, S., Nader-Grosbois, N., 2008. Self-regulation during pretend play in children with intellectual disability and in normally developing children. Research in Developmental Disabilities 29, 256–272.

Werner, H., Kaplan, B., 1963. Symbol Formation. Wiley, New York.

Prevention of Crime and Delinquency

Kristine Levan and David A Mackey, Plymouth State University, Plymouth, NH, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K.J. Moore, volume 5, pp. 2910–2914, © 2001, Elsevier Ltd.

Abstract

Crime prevention is the act of stopping or reducing criminal activities, by either known or unknown offenders. This article explores various aspects of crime prevention, including the theoretical basis, as well as efforts by various criminal justice agencies, including law enforcement and correctional systems. Specific areas of interest, including substance abuse prevention and the role of firearms in crime, are also addressed.

Introduction

There is no singular definition of crime prevention, nor is there a singular means by which to prevent crime. One of the most concise definitions given by Steven Lab is, "Crime prevention entails any action designed to reduce the actual level of crime and/or the perceived fear of crime" (2006: p. 682). This definition would mean that crime prevention is inclusive of all crime reduction efforts. This may include those at the level of the criminal justice agencies, as well as by institutions, communities, and individuals. Any discussion of crime prevention should begin with the theoretical basis behind the concept. Additionally, the discussion also focuses on tactics by law enforcement, as well as those centering on sentencing and punishment, as well as specific efforts directed at firearms, drugs, and alcohol.

Theory of Crime Prevention

Crime prevention policies and programs draw from traditional criminological theories, such as learning theories and choice-based theories. Choice-based theories include deterrence, routine activity theory, and situational crime prevention. According to rational choice theory, individuals have free will to choose between criminal or noncriminal behaviors. Deterrence theory, the most prominent choice-based theory, is based on the idea that people engage in crime when the anticipated benefits of committing a crime outweigh the likely costs of crime. Conversely, the person is deterred from crime when the costs of crime outweigh the expected or predicted gains (Pratt et al., 2006). The quality of formal punishments is a key aspect of deterrence theory. In deterrence theory, there is an emphasis on swift, certain, and severe sanctions for criminal behavior. Deterrence presents a solid theoretical foundation for some policies and programs, and it offers a fairly straightforward solution to the problem of crime specifically to increase the likelihood of detecting a crime and strengthening the punishments associated with criminal behavior.

There are two forms of deterrence: general and specific. General deterrence seeks to prevent crime among the general population through the public awareness of both crimes and punishments. Specific deterrence seeks to prevent crime committed by those individuals who have already committed a previous offense. In theory, these offenders would not commit another offense because they experienced the pains of punishment meeting a threshold level of swift, certain, and severe penalties. For the prevention of crime, deterrence theory has a common sense appeal. Most people seek to avoid the painful experiences and variety of costs associated with formal sanctions by the criminal justice system.

Routine activity theory and situational crime prevention are more contemporary aspects of choice theory. Routine activity theory was developed by Cohen and Felson in 1979. The theory states that, a crime is more likely to take place when a victim motivated offender, and the absence of capable guardians converge in time and place. It is closely connected to rational choice since offenders still engage in decision-making processes whether to engage in crime but in a very specific context. Offender propensity to commit a crime may be constant yet specific places and contexts vary in terms of the temptations and controls present in the setting (Felson and Boba, 2010). The key component of the theory is the idea that crime can be controlled by manipulating the opportunity for crime by the strategic use of place managers who would have the responsibility for supervision of specific locations. Handlers provide additional supervision for those at risk of becoming an offender while guardians provide supervision for potential victims; a mentor provides an example for each category. Another example is the negative correlation found between the prosocial activities youths are involved in at the community level and their involvement in criminal activity. By engaging in law-abiding behavior, these youths are less likely to be both victims and offenders (McCall et al., 2012).

Likewise, situational crime prevention seeks to apply choice theory principles to very specific opportunities for crime (Clark, 1995). Crime prevention applications utilizing situational crime prevention emphasizes increasing the effort to commit the offense, increasing the risk of detection, reducing the rewards associated with the act, reducing the provocations to escalate a course of action, and remove excuses for offending (Center for Problem-Orientated Policing, 2011). Crowe (1991) has outlined the principles of crime prevention through environmental design (CPTED) to include the strategies of natural access control, natural surveillance, and territorial reinforcement. The strategies seek to reduce

opportunities for crime and also increase the risks of detection for offenders. Crowe (1991) notes that CPTED can be applied to problems associated with very specific settings such as restrooms in schools and convenience store layout and design. Using CPTED for the design of communities can have a positive impact on traffic patterns and related nuisance behaviors. Oscar Newman (1996) coined the term defensible space in the analysis related to the failure of the 2700-unit Pruitt–Igoe Housing Project.

Wilcox et al. (2013) provide a detailed discussion of the relationship between criminality as an individual propensity and the opportunities available to commit crime. They contend that the distribution of crime is driven by the perceived opportunities for criminal conduct in a specific context. Some places experience higher levels of crime attributed to the less guarded interactions of potential offenders and potential victims, the physical layout of a location with features that diminish the risks of detection and apprehension, as well as areas featuring lower levels of informal control. These factors operate independently and possibly interactively with one another, thus producing more opportunities for crime and less restraints in a more cumulative fashion.

In addition to traditional criminological theories, recent crime prevention strategies have adapted the medical model borrowed from the public health field. While deterrence theory has provided a historical foundation for criminal justice policies and crime prevention strategies, the medical model represents a fundamental shift from a reliance on responding to criminal events after their occurrence to a response strategy based on a true preventive strategy. The public health model emphasizes three approaches to crime prevention: primary, secondary, and tertiary prevention.

The medical model is best illustrated by society's potential responses to public health risks such as heart attacks, diabetes, or hypertension. Farrington (2000) notes that the public health model seeks to prevent the occurrence of these above-mentioned health ailments with a variety of strategies. A focus on emergency room treatment capacity, technology, and time to treatment intervention to improve the response to heart attack victims surely has a positive impact on life expectancy and outcomes although an exclusive reliance on such an approach fails to take into consideration other ways to reduce a person's initial risk of experiencing a heart attack. Individuals may benefit from making changes to their lifestyle, such as diet, exercise, and tobacco use, with a goal to reduce their improving their risk of health issues. Likewise, individuals who are at an increased risk to experience a heart attack may lower this risk with medication as well as additional lifestyle modifications.

Primary prevention takes a proactive stance to reduce the likelihood or severity of the targeted outcome prior to its development. Shader (2003) notes that primary prevention focuses on strengthening factors associated with abstaining from or avoiding criminal behavior. These factors are also known as resiliency factors. Primary prevention also seeks to reduce or minimize factors associated with an individual's likelihood to engage in crime; these factors are also known as risk factors. Farrington (2000) notes that a sufficient body of research permits policy makers to determine the relative cost-benefit impact of strengthening resiliency factors or address risk factors on the level of subsequent offending. This cost-benefit analysis allows policy makers to target factors with research supporting its effectiveness and also with the greatest potential cost savings.

Primary prevention emphasizes preventative efforts well before the onset of crime and delinquency. Regoli et al. (2010) identify prenatal programs, parenting programs, and preschool programs as model primary prevention efforts. Prenatal programs identify resource deficiencies for expecting mothers and target the delivery of services to improve the physical well-being of the mother and unborn child. These efforts seek to avoid or minimize negative factors for the baby such as low birth weight, neurological impairments, and exposure to various environmental toxins, which are factors associated with developmental and educational impairments. These factors are linked to poor school achievement and potential higher likelihood for involvement in delinquency in subsequent years.

While primary prevention programs impact a larger swath of the public, secondary prevention efforts target a narrower population. Secondary prevention focuses on individuals and places at an increased risk for crime and delinquency. Regoli et al. (2010) note the focus of secondary prevention is to reduce the prevalence, seriousness, or duration of crime and delinquency. Illustrating the concept of risk factors, Farrington (2000) identified nine risk factors for criminality among boys aged 8 to 10 years in separate studies in London and in Pittsburgh. These variables were "hyperactivity, poor concentration, low achievement, an antisocial father, large family size, low family income, a broken family, poor parental supervision, and parental disharmony" (Farrington, 2000: p. 5). The Juvenile Mentoring Program was developed to counter the development of risk factors among youths many of whom had attending risk factors (Novotney et al., 2000).

Tertiary prevention efforts seek to prevent future crime by those individuals who have already committed a crime. Prison-based programs best exemplify tertiary prevention programs. As is the case with primary and secondary prevention programs, not all tertiary programs produce their intended outcomes. The Scared Straight targeted youths who were seen by teachers, parents, and juvenile probation officers as likely to engage in more serious delinquency if not diverted from their current life trajectory. The Scared Straight program was made popular by a television documentary in 1978 narrated by Peter Falk. Finckenauer et al. (1999) contend that youths who experienced Scared Straight had worse outcomes in terms of delinquency than their counterparts in the control group. Klenowski et al. (2010) examined the research on juvenile awareness programs, the definition of which includes programs like Scared Straight. They examined eight evaluation studies of sufficient rigor and concluded that juvenile awareness programs are not effective in deterring juveniles. Critically, they note that youths who participate in these programs have worse outcomes than their peers who do not participate.

Police and Crime Prevention

Police agencies have incorporated crime prevention since the inception of policing in London in 1829. Sir Robert Peel

envisioned a mission for the London Metropolitan Police Department that centered on crime prevention. When crafting the mission of the London Metropolitan Police Department, Peel crafted nine principles detailing their mission and philosophy with crime prevention as a focal concern. These principles were

> The basic mission for which the police exist is to prevent crime and disorder.
>
> The ability of the police to perform their duties is dependent upon the public approval of police actions.
>
> Police must secure the willing co-operation of the public in voluntary observation of the law to be able to secure and maintain the respect of the public.
>
> The degree of co-operation of the public that can be secured diminishes proportionately to the necessity of the use of physical force.
>
> Police seek and preserve public favour not by catering to public opinion, but by constantly demonstrating absolute impartial service to the law.
>
> Police use physical force to the extent necessary to secure observance of the law or to restore order only when the exercise of persuasion, advice, and warning is found to be insufficient.
>
> Police, at all times, should maintain a relationship with the public that gives reality to the historic tradition that the police are the public and the public are the police; the police being only members of the public who are paid to give full-time attention to duties which are incumbent upon every citizen in the interests of community welfare and existence.
>
> Police should always direct their action strictly towards their functions, and never appear to usurp the powers of the judiciary.
>
> The test of police efficiency is the absence of crime and disorder, not the visible evidence of police action in dealing with it. (Nazemi, 2009)

Preventive patrol is the primary technique used by the police to prevent crime. This focus, which started with the formation of early police departments, has incorporated motorized patrol, foot patrol, improved dispatch, and differential response to calls to aid in maximizing police visibility and response. The reliance on preventative patrol would remain untested until the Kansas City Patrol Study challenged the basic assumption of the effectiveness of preventive patrol (Kelling et al., 1974), which concluded increased levels of preventive patrol had no effect on crime and victimization.

Subsequent research by Sherman and Weisburd (1995) sought to address a methodological issue present in the Kansas City Patrol Study. They note there was a built-in bias toward a finding of no difference between the levels of patrol since comparisons across only 15 police precincts. To remedy this issue, Sherman and Weisburd used 110 hot spots in the city of Minneapolis, which generated a disproportionate amount of calls for service in the city. The 110 locations were divided randomly into an experimental group and a control group. In the experimental group of hot spots, there was a target goal of 3 h of total police presence between the hours of 11 a.m. to 3 a.m. The hot spots in the experimental group received roughly double the amount of police presence as the control group. Officers could engage in a variety of activities while at the hot spot but the primary goal was for officers to stay at the hot spot for a relatively short period of time and return frequently to the location.

Sherman and Weisburd (1995) report a modest deterrent effect using directed police patrols in crime hot spots. They describe the impact as microdeterrence effect since the impact had a small geographic application and the displacement of crime was not tested. Displacement of crime refers to the movement of crime from one location, in this case the physical area receiving a fairly significant dosage of police presence, to another area. In effect, there is potential for motivated offenders to seek criminal opportunities in other areas with less official supervision and presence. Sherman and Weisburd note the optimal length of the police presence in the directed patrol is about 12 min. They note the impact of directed police patrol on social disorder. The results indicated that social disorder was present in 1 of every 50 encounters on the street in the experimental group while social disorder was present in one of every 25 encounters on the street in the control group.

Smith and Scott (2013) provide an extensive discussion of the historical progress of crime prevention strategies of police. They note that modern crime prevention efforts can be grouped into three broad categories of activities: community policing, problem-orientated policing, and intelligence-led policing. They note that community policing has become ubiquitous but the proliferation of the related tactics associated with community policing has come at the expense of clarity of the concept. Enduring concepts of community policing include improved collaboration between the police and the public in developing police priorities and an emphasis on problem solving. Problem-orientated policing and intelligence-led policing likewise focus on solving problems but can be used within the traditional style of policing.

Sentencing and Punishment

At the very core of criminal sentencing philosophies is the idea of crime prevention. When an offender is given a sentence, it is believed that not only will he or she not commit crime while serving their prescribed punishment, but beyond that time frame as well. A wide range of punishments can be given, depending on the severity of the offense, ranging from a citation to capital punishment.

Intermediate sanctions are those that are the most commonly used, and often include some level of community supervision to prevent the offenders from recidivating. Community supervision includes probation and parole, as well as additional methods of control, such as electronic monitoring or home confinement. At year end in 2011, the United States had approximately 4 814 200 individuals on community supervision (Maruschak and Parks, 2012). Because the likelihood of recidivism is the greatest immediately upon release from jail or prison, the most intensive forms of

community supervision generally occur at the beginning of their supervision period (Petersilia, 2003).

In the United States, offenders sentenced to serve a term of incarceration are given either indeterminate or determinate sentences, depending on the state in which they live. Most states have moved toward the determinate sentencing system, which allocates a specified length of time to be served for each crime, as opposed to the indeterminate sentencing system, which offers a range of time to be served, depending on factors such as efforts toward rehabilitation. Some states also prescribe mandatory minimum sentences. Mandatory minimums are designed to deter potential offenders due to the severity of the punishment. For instance, some states have instated three strikes laws, which mandate that after three felonies, offenders are given mandatory life sentences. While most states stipulate these felonies must be violent, California does not differentiate between violent and nonviolent felonies.

At the end of 2011, there were 2 239 800 offenders serving time in either jails or prisons (Glaze and Parks, 2012). While incarcerated, some inmates are provided opportunities for rehabilitation, education, and vocational training, all of which are designed to prevent recidivating upon their release. However, due to increasing jail and prison populations and lack of financial resources, many of these recidivism reduction programs have been either drastically reduced or eliminated (Petersilia, 2003).

Capital punishment is the most severe penalty that can be levied against an offender. Worldwide, 58 nations list death as a potential punishment, including the United States, where it is reserved for murder with aggravating circumstances. Although these circumstances vary by state, some commonalities include murder of a law enforcement officer, murder of young victims, murder for remuneration, and murder during the commission of other felonies.

As a crime prevention tool, specific deterrence is achieved by capital punishment. Once an offender has been executed, he or she no longer has the opportunity to commit future crimes. General deterrence is more difficult to achieve. The basic tenets of general deterrence (the punishment must be severe, certain, and swift) cannot be adequately achieved under the current system. Inarguably, capital punishment is a severe penalty. However, the possibility of not being apprehended, plea bargains, and options for lesser sanctions reduce the certainty of the punishment. The swiftness of the death penalty is reduced by the lengthy trial process, as well as the time spent on death row, which averages about 15 years (Time on Death Row).

Guns and Crime

The relationship between firearms and crime is a tenuous and complex one. The United States has high rates of gun ownership, and correspondingly high rates of gun-related crime. Firearm ownership is the most prevalent in the southern regions of the United States and among white males (Cook and Ludwig, 1995; Spano et al., 2008; Wyant and Taylor, 2007).

Firearms are used in about 8% of violent crimes, including 28% of robberies and 5% of assaults (Truman and Rand, 2010). Handguns are the most prevalent weapon used in homicide, followed by other guns, knives, other weapons, and blunt objects (Cooper and Smith, 2011). An argument or assault is more likely to be fatal if a firearm is present (Hepburn and Hemenway, 2004).

These and similar statistics are often cited to increase gun control policies. However, some argue that because many of the firearms used to commit crime are not obtained legally, gun control is simply more successful at keeping firearms out of the hands of the noncriminogenic population. For instance, the Brady Bill of 1993 attempts to prevent the sale of the firearms to those who are the most likely to use them for commission of a crime. The main tenets of the bill instituted a background check for all firearm purchases, as well as a 5-day waiting period.

Most firearms that are used to commit a crime are not purchased legally (Piquero, 2005; Wright and Rossi, 1986). Many are obtained through friends or family, or through 'straw purchasers,' who legally purchase the guns and then sell them on the black market. These illegal transactions often occur quickly after the legal purchase, especially in the case of semiautomatic and large caliber guns (Brandl and Stroshine, 2011). Additionally, guns may be stolen during household burglaries or other property crimes. Annually, an average of 232 400 firearms are reported stolen (Langton, 2012). Many of these firearms will be used to commit crime, or be sold on the black market to those who cannot purchase a gun legally.

Thus, it seems that those who want to obtain firearms will do so, through either legal or illegal means. Some scholars, such as John Lott (2000) argue that by increasing the prevalence of firearms among the noncriminogenic population, crime can be better prevented. This is especially true if the weapons are concealed, as the criminals will be unaware of who is armed, theoretically deterring them from crime regardless of whether their proposed target is armed.

Other initiatives have been implemented to decrease the presence of firearms among criminogenic populations. For instance, gun buyback programs provide gift cards in small amounts in exchange for firearms from neighborhood residents. These programs had little impact, if any, on gun violence (Callahan et al., 1994). Project Exile is another initiative, which gives felons in possession of a firearm a mandatory minimum sentence. There have been mixed reports on the effectiveness of this initiative.

Drugs and Alcohol

According to the National Institute on Drug Abuse (2012), in 2011, about 22.5 million Americans aged 12 years and older used illicit drugs in the past month. Marijuana is the most commonly used drug, followed by psychotherapeutics, cocaine, hallucinogens, inhalants, and heroin. Additionally, more than half of those surveyed reported being current alcohol drinkers, with about $^{1}/_{4}$ of respondents reporting binge-drinking behaviors.

Prevention of drugs and alcohol is difficult for several reasons. Alcohol is legal, and is controlled in terms of creating laws in limiting the sale to minors. Marijuana is legal in a few states, but is illegal in most. Prescription drugs are legal to those who they are prescribed. The partial legality of these substances

make them difficult to monitor and control, especially as black markets are created for them.

The use of most substances, legal or not, generally occurs in private residences, making it difficult to prevent or monitor the individuals using them. Moreover, with the exception of drunk or drugged driving, or other crimes occurring in conjunction with these offenses, they are generally considered victimless crimes. In the absence of a victim, not only are they unlikely to be reported, but they also may be low on the priority list for law enforcement, especially if there is a relatively large number of violent or property crimes with which they must contend.

Despite these obstacles, there are initiatives in place to try to prevent drug- and alcohol-related offenses. For instance, to reduce Driving Under the Influence (DUI) or Driving While Intoxicated (DWI) incidents, most states enforce a blood alcohol content level of less than .08. In order to deter individuals from driving under the influence, police checkpoints are often set up at roadblocks.

Other methods involve preventive education. Perhaps the most well known of these is Drug Abuse Resistance Education (DARE), which began in the 1980s. Generally, DARE brings law enforcement officers into elementary or middle schools to provide drug education to students, who in turn make a pledge to desist from drugs. Many studies have cited DARE's ineffectiveness to decrease drug use among participants. Some studies have indicated that substance abuse among DARE participants actually increased after being exposed to the program (Sherman et al., 1998). The DARE program recently adopted the 'keepin' it REAL' program. This revision includes teaching effective communication strategies, as well as considering diverse cultural and ethnic audiences (Hecht et al., 2010).

Another tactic used for prevention is the use of drug treatment programs. Because of increased sentences directed at substance abusers, they comprise approximately 56% of federal and 20% of state inmates (West and Sabol, 2010). Although only a small number of those who need treatment will receive it, primarily due to financial restraints (Petersilia, 2003), less expensive alternatives, such as those delivered in group or therapeutic community settings, have had positive results in preventing future substance abuse (Mitchell and MacKenzie, 2003).

Conclusion

Just as there are many causes of crime, there are many ways in which crime can be prevented. Preventing crime and delinquency often involved multiple agencies and individuals. Criminal justice agencies, such as police, courts, and corrections, all play an important role in prevention. Communities also play an important role by being proactive with efforts such as neighborhood watch engaging in collective efficacy. Individuals are also responsible for crime prevention. This can be accomplished by making their physical spaces less susceptible to crime (for example, by installing security systems or locking the doors to their homes), and avoiding situational crimes (for example, by not being in high-crime areas). Despite the numerous efforts put forth in order to prevent crime, it should be remembered that not all crime is preventable.

See also: Communities and Crime; Guns and Criminals; Policing; Recidivism; Routine Activities, Criminal Opportunities, Crime and Crime Prevention.

Bibliography

Brandl, S.G., Stroshine, M.S., 2011. The relationship between gun and gun buyer characteristics and firearm time-to-crime. Criminal Justice Policy Review 22 (3), 285–300.

Callahan, C.M., Rivara, F.P., Koepsell, T.D., 1994. Money for guns: evaluation of the Seattle buy-back program. Public Health Reports 109 (4), 472–477.

Center for Problem-Orientated Policing, 2011. Twenty Five Techniques of Situational Crime Prevention. Retrieved: http://www.popcenter.org/25techniques/.

Clark, R.V., 1995. Situational crime prevention. In: Tonry, M., Farrington, D.P. (Eds.), Building a Safer Society: Strategic Approaches to Crime Prevention, vol. 19. University of Chicago Press, Chicago, pp. 91–150.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–605.

Cook, P.J., Ludwig, J., 1995. Guns in America: Results of a Comprehensive National Survey on Firearms Ownership and Use. Police Foundation, Washington, DC.

Cooper, A., Smith, E.L., 2011. Homicide Trends in the United States, 1980–2008 (NCJ 236018). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Crowe, T.D., 1991. Crime Prevention through Environmental Design: Applications of Architectural Design and Space Management Concepts. Butterworth-Heinemann, Boston.

Farrington, D.P., 2000. Explaining and preventing crime: the globalization of knowledge – The American Society of Criminology 1999 presidential address. Criminology 38 (1), 1–24.

Felson, M., Boba, R., 2010. Crime and Everyday Life, fourth ed. Sage Publications, Thousand Oaks, CA.

Finckenauer, J.O., Gavin, P.W., Hovland, A., Storvall, E., 1999. Scared Straight: The Panacea Phenomenon Revisited. Waveland Press, Prospect Heights, IL.

Glaze, L.E., Parks, E., 2012. Correctional Populations in the United States, 2011 (NCJ 239972). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Hecht, M.L., Colby, M., Miller-Day, M., 2010. The dissemination of keepin' it REAL through DARE America: a lesson in disseminating health messages. Health Communication 25 (6/7), 585–586.

Hepburn, L.M., Hemenway, D., 2004. Firearm availability and homicide: a review of the literature. Aggression and Violent Behavior 9, 417–440.

Kelling, G.L., Pate, D., Dieckman, D., Brown, C.E., 1974. The Kansas City Preventive Patrol Experiment: A Summary Report. Police Foundation, Washington, DC. Retrieved from: http://www.policefoundation.org/pdf/kcppe.pdf.

Klenowski, P.M., Bell, K.J., Dodson, K.D., 2010. An empirical examination of juvenile awareness programs in the United States: can juveniles be "Scared Straight"? Journal of Offender Rehabilitation 49, 254–272.

Lab, S.P., 2006. Crime prevention, politics, and the art of going nowhere fast. Justice Quarterly 21 (4), 681–692.

Langton, L., 2012. Firearms Stolen during Household Burglaries and Other Property Crimes, 2005–2010 (NCJ 239436). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Lott, J.R., 2000. More Guns, Less Crime: Understanding Crime and Gun Control Laws. University of Chicago Press, Chicago, IL.

Maruschak, L.M., Parks, E., 2012. Probation and Parole in the United States, 2011 (NCJ 239686). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

McCall, P.L., Land, K.C., Dollar, C.B., Parker, K.F., 2012. The age structure-crime rate relationship: solving a long-standing puzzle. Journal of Quantitative Criminology 29 (2), 167–190.

Mitchell, O., MacKenzie, D.L., 2003. Crime prevention via prison-based drug treatment: a systematic review and assessment of the research. In: Kury, H., Obergfell-Fuchs, J. (Eds.), Crime Prevention: New Approaches. Weisser Ring, Weberstrabe, Germany.

Nazemi, S., 2009. Sir Robert Peel's Nine Principals of Policing. Retrieved: http://www.lacp.org/2009-Articles-Main/062609-Peels9Principals-SandyNazemi.htm.

Newman, O., 1996. Creating Defensible Space. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC. Retrieved: http://www.huduser.org/publications/pdf/def.pdf.

Novotney, L.C., Mertinko, E., Lange, J., Kelly Baker, T., 2000. Juvenile Mentoring Program: A Progress Review. U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention. Retrieved: http://www.ncjrs.gov/pdffiles1/ojjdp/182209.pdf.

Petersilia, J., 2003. When Prisoners Come Home: Parole and Prisoner Reentry. Oxford University Press, New York, NY.

Piquero, A.R., 2005. Reliable information and rational policy decisions: does gun control fit the bill? Criminology and Public Policy 4 (4), 779–797.

Pratt, T.C., Cullen, F.T., Blevins, K.R., Daigle, L.E., Madensen, T.D., 2006. The empirical status of deterrence theory: a meta-analysis. In: Cullen, F.T., Wright, J.P., Blevins, K.R. (Eds.), Taking Stock: The Status of Criminological Theory. Advances in Criminological Theory, vol. 15, Transaction Publishers, Washington, DC, pp. 367–395.

Regoli, R.M., Hewitt, J.D., DeLisi, M., 2010. Delinquency in Society, eighth ed. Jones and Bartlett Publishing, Sudbury, MA.

Results from the 2011 National Survey on Drug Use and Health: Summary of National Findings, 2012. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Shader, M., 2003. Risk Factors for Delinquency: An Overview (NCJ 207540). Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, U.S. Department of Justice, Washington, DC. Retrieved from: http://www.ncjrs.gov/pdffiles1/ojjdp/frd030127.pdf.

Sherman, L.W., Gottfredson, D.C., MacKenzie, D.L., Eck, D., Reuter, P., Bushway, S.D., 1998. Preventing Crime: What Works, What Doesn't, What's Promising. U.S. Department of Justice, National Institute of Justice, Washington, DC.

Sherman, L.W., Weisburd, D., 1995. General deterrent effects of police patrol in crime "hot spots": a randomized controlled trial. Justice Quarterly 12 (4), 625–648.

Smith, T.R., Scott, J., 2013. Policing and crime prevention. In: Mackey, D.A., Levan, K. (Eds.), Crime Prevention. Jones and Bartlett Learning, Burlington, MA.

Spano, J., Freilich, J.D., Bollard, J., 2008. Gang membership, gun carrying and employment: applying routine activities theory to explain violent victimization among inner-city, minority, youth living in extreme poverty. Justice Quarterly 25 (2), 381–410.

Time on Death Row, n.d. Death Penalty Information Center. Retrieved from: www.deathpenaltyinfo.org/time-death-row.

Truman, J.L., Rand, M.R., 2010. Criminal Victimization, 2009 (NCJ 231327). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

West, H.C., Sabol, W.J., 2010. Prisoners in 2009 (NCJ 231675). U.S. Department of Justice, Bureau of Justice Statistics.

Wilcox, P., Miller Gialopsos, B., Land, K.C., 2013. Multilevel criminal opportunity. In: Cullen, F.T., Wilcox, P. (Eds.), The Oxford Handbook of Criminological Theory. Oxford University Press, New York.

Wright, J.D., Rossi, P.H., 1986. Armed and Considered Dangerous: A Survey of Felons and Their Firearms. Aldine Publishing Co., Hawthorne, NY.

Wyant, B.R., Taylor, R., 2007. Size of household firearm collections: implications for subcultures and gender. Criminology 45 (3), 519–546.

Prevention Research

Jean L Hill, New Mexico Highlands University, Las Vegas, NM, USA
Bret Kloos, University of South Carolina, Columbia, SC, USA
Elizabeth Thomas, Rhodes College, Memphis, TN, USA
Abe Wandersman, University of South Carolina, Columbia, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of the history of prevention science from its conceptual foundations and definitions through its development as a well-established and extremely productive discipline. Effective approaches for program dissemination and evaluation are presented. Other topics covered include the relationships between risk and resiliency, prevention versus promotion, and program implementation versus program innovation. Current efforts in assessing cost-effectiveness, the cultural relevancy of prevention programs, and the role of prevention science in informing public policy are explored. Examples of effective prevention and promotion programs are included throughout the article.

In 1962, David Weikart, the director of special education for the Michigan, Ypsilanti Public School District, came up with an ambitious plan. A large portion of the children in his district came from disadvantaged homes, and many did not start school until the first grade. The director knew that these children were faced with an elevated risk of school failure and other negative outcomes compared to children born in more affluent circumstances. His plan was to prevent those negative outcomes by providing the children with a high-quality, academically based preschool program. This was the start of the Perry Preschool Program, one of the longest running primary prevention programs in the United States.

Over the next 5 years, 123 children were recruited into the project when they were 3 or 4 years old. Half of the children attended the preschool program; the other half did not. The initial evaluations of the program showed higher intellectual performance among preschool participants compared to the control group. These effects, however, diminished somewhat over time, suggesting that the positive effects of the program were only short-lived. But at age 10, the researchers found something surprising, the positive effects seemed to be amplifying rather than diminishing over time. Program participants were significantly less likely to have been held back a grade or placed in special education than children in the control group. At age 15, program participants demonstrated higher academic achievement than those in the control group, and by ages 19 and 27, program participants were showing significantly higher educational achievement, employment rates, earnings, and rates of home ownership, and lower rates of criminal behavior and single parenthood than the comparison group. Later evaluations found that these positive effects continued until the participants were 40 years old and concluded that the general public gained $12.90 in return for every $1 spent on the program (Schweinhart et al., 1993, 2005).

The results of this study have had an immense effect not only on the public recognition of the benefits of high-quality preschool programs, but also on the field of prevention science as a whole. In this article we will give a brief overview of the history of prevention science, focusing on the extensive development of the field in the past 50 years. In particular, we will address current theory and research considering: (1) the relationship between the concepts of prevention and promotion; (2) effective approaches for disseminating programs, including issues of fidelity and context-based innovation; (3) the central role of program evaluation in implementation; and (4) assessing cost-effectiveness.

Conceptual Foundations of Prevention Approaches

Contemporary approaches to prevention draw upon a range of intellectual backgrounds and disciplines. In this section we briefly review major schools of thought and leading advocates for prevention.

Public Health Models for Prevention

Systematic investigation and intervention to control the spread of contagious diseases, such as cholera, polio, or tuberculosis, date back to the mid-nineteenth-century England and physician John Snow (Kloos et al., 2012). Identification of risk factors (e.g., using water from a specific neighborhood well) and changing environmental conditions (e.g., removing a pump handle, better sanitation) have become a cornerstone of public health approaches to prevention. These early efforts established principles common to most prevention efforts. First, it is assumed that the problem of concern will be distributed among a specific population or specific area. This distribution needs to be understood as a rate of exposure or vulnerability. New cases of the problem can be tracked as the *incidence* of a problem and compared across time and space. The overall rate of those who have the problem establishes the *prevalence* of that problem in a specific region. Second, it is assumed that science can identify factors that increase an individual's risk to developing the problem. These factors may be environmental, individual, or result from an interaction of individuals in particular environments. When risks are not well understood, more study is needed. Third, once risk can be identified, broad-scale efforts are needed to identify those who are at greater risk. Fourth, effective interventions need to be

developed to help people avoid the risk through changing environmental conditions (e.g., draining swamps to reduce exposure to mosquitoes with malaria) or boosting individuals' capacities to fend off risk (e.g., a policy of early vaccination). From a public health perspective, individual treatment may be inefficient to address large areas of need, not timely enough, inefficient, or too costly. Although it may not be possible to completely avoid new cases of a problem in a population, there is value in greatly reducing the number of people affected.

Introducing Prevention to Mental Health

Mental health treatment had long included goals of reducing stress, but a conceptual framework for systematic prevention was largely lacking until the mid-twentieth century. Public health and community psychiatry helped to bridge these notions and achieve goals for which progress had been so limited. Nicolas Hobbs (1964) proposed that a 'third mental health revolution' occurred when prevention of dysfunction was introduced as a goal of mental health systems. As conceived by Hobbs and his contemporaries, community mental health would provide services in the local contexts where people spent their daily lives, and if possible, to avoid problems becoming episodes of mental illness that required hospitalization. Hobbs' revolution was built, in part, upon two observations in the late-1950s. First, one of the influential reports from the U.S. Joint Commission on Mental Illness and Health documented that we will never be able to train enough mental health professionals to meet the expressed need for traditional individual-oriented mental health services (Albee, 1959). Second, persons of lower socioeconomic class were more likely to receive hospitalization (e.g., psychotropic drugs or electroshock) and less access to effective outpatient care (Hollingshead and Redlich, 1958). Thus, new models of service were needed to prevent problems where possible. As proposed by Hobbs, a strategy of prevention included both pragmatic grounds (i.e., meeting the expressed need for mental health services) and a values-based concern about providing greater access to effective mental health care, especially to those who did not have resources to seek their own private treatment. This blending of pragmatics and values leads to an ideology of prevention that seeks to avoid illness rather than only ameliorate it.

Articulating Strategies for Prevention

Gerald Caplan's (1964) *Principles of Preventive Psychiatry* created an elegant framework for promoting prevention in mental health that was both parsimonious and generative. He provided practical examples of how ideas from sociological and epidemiological analyses of disadvantage and risk could be translated into mental health practices.

Tertiary Prevention

This level of prevention targets people who already have an illness or problem. The focus of tertiary prevention is to reduce the rate of existing disorders in a community, notably by targeting the duration and associated conditions that can exacerbate and prolong episodes of illness. Like public health models, Caplan maintained that the "community is the client." Prevention must include large-scale programs to change social conditions for large programs of persons with a problem (e.g., housing and employment programs as policy for all persons with disabilities). Caplan distinguished this from rehabilitation, which he defined as efforts to work with an individual client on her/his deficits. In mental health, tertiary prevention guides efforts to go beyond custodial care to address social conditions and quality of services to keep people functioning in communities. An early example of a tertiary prevention program is the Community Lodge program, which is designed to prevent the rehospitalization of adults living with serious mental illness (Fairweather, 1979).

Secondary Prevention

This approach seeks to lower the number of cases of a problem for a population determined to be at increased risk. Typically, this is accomplished by early identification of risk and interventions to alter the influence or exposure to factors that put people at greater risk (e.g., building individuals' capabilities, removal of stressors). As prevention has been practiced, this level often focuses on children and youth, identifying situations and vulnerabilities that could lead to problems later in life. Alternatively, early detection of risk can mean early screening and diagnosis of premorbid illness (e.g., colon cancer screening, mental health screenings in work places or schools). This early identification of heightened risk does not necessarily lead to individual treatment. Rather, these approaches often involve teaching skills or applying additional resources to promote functioning. This approach to prevention requires reliable identification of risk factors and potent interventions to reduce risk.

Parenting programs for families with young, highly oppositional children are an example of secondary prevention. The children in these families are already displaying behaviors that are not only considered problematic in themselves, but are also predictive of more serious problems later in development (Patterson et al., 2010; Webster-Stratton, 1984).

Primary Prevention

While the first two levels of Caplan's approach to prevention could be viewed as extensions of traditional mental health services, primary prevention introduced radically new ways of thinking about promoting mental health and well-being. "(Primary prevention) seeks to reduce the risk for the whole population, so that, although some may become ill, their number will be reduced. It thus contrasts with individual patient-oriented psychiatry, which focuses on a single person and deals with general (community) experiences only insofar as they are combined in his unique experience" (Caplan, 1964, p. 26). This emphasis on addressing risks to health of a whole population *before* signs of risk for individuals are identified moved mental health services away from primarily focusing on individuals' deficits or potential deficits. For Caplan, primary prevention included assessment of health and conditions that promote health and well-being.

Programs to increase women's access to and use of prenatal care (Edgerley et al., 2007), efforts to ensure that children do not begin smoking cigarettes (Jason et al., 1999), and programs to increase the use of safe sex practices to prevent sexually

transmitted disease (Lykes et al., 2007) are all examples of primary prevention programs.

Applying Models for Prevention beyond Mental Health

While Caplan primarily focused on mental health, his ideas for prevention have been influential outside of psychiatry and related professions. His insights into primary prevention, in particular, have led to efforts to identify conditions that put people at risk for a variety of life problems and attempts to prevent them for entire populations (e.g., substance abuse, teen pregnancy, exposure to violence).

Barbara Snell Dohrenwend (1978) and George Albee (1982) advanced the field by proposing heuristics to conceptualize both potential risk and resources to avoid the development of problems. First, they helped expand how risk was conceptualized. Risk factors could be either proximal or distal to persons' experiences. Furthermore, risks could be physiological or psychological, social, or environmental. Similarly, potential protective factors could be identified in each of these domains. Summarized in Albee's prevention equation, cumulative risk for an individual developing a particular problem could be viewed as the numerator of an equation where the denominator represents collective resources to resist developing the problem.

$$\frac{\text{Stress} + \text{Physical vulnerability}}{\text{Coping skills} + \text{Social support} + \text{Self-esteem}}$$

These frameworks identify several potential targets of intervention to reduce risk or to promote coping and competence. Maurice Elias (1987) applied this framework to address conditions associated with the development of behavioral, emotional, or social disorders in populations rather than just in individuals. This extended prevention formula targets changing environmental conditions (e.g., families, schools, organizations, neighborhoods, and societies) rather than focusing solely on changing individual behavior or exposure to risk. These models helped expand the range of preventive strategies that might be used for intervention (e.g., health promotion, skills training, environmental control, social policy) and create opportunities to apply preventive strategies to a wide range of social problems as well as health concerns.

Pandisciplinary Approach to Prevention

Thirty years after Hobbs called for prevention to be integral to mental health care, the Institute of Medicine (IOM) issued a consensus statement promoting prevention approaches across disciplines. This report introduced a new framework for classifying preventive interventions: universal, selective, and indicated prevention (Mrazek and Haggerty, 1994). This framework seeks to promote prevention of health problems widely and to promote collaboration across disciplines. Below, we briefly describe each of these three approaches.

Universal preventive interventions are utilized for everyone in a given population group to promote well-being or avoid problems. They are usually implemented for populations without specification of special risk but avoid risks common to all (e.g., using barriers to blood-born pathogens, washing hands). Universal prevention is similar to Caplan's model of primary prevention. *Selective preventive interventions* are intended for those at above-average risk for developing disorders. The risk may be based on exposure to environmental stressors (e.g., neighborhood crime, pollution) or individual factors (e.g., cognitive vulnerabilities, health conditions). In selective prevention, indicators of risk have been documented as being related to etiologies of disorders but are not criteria for diagnoses of particular disorders. *Indicated preventive interventions* are directed toward individuals who are determined to be at high risk for developing a disorder. Unlike selective prevention, indicated prevention is used when persons show early symptoms of the disorder but their condition does not yet meet diagnostic criteria. Caplan's concept of secondary prevention appears to encompass both selective prevention and indicated prevention as proposed by the IOM.

While the IOM model has promoted prevention as a strategy to address health concerns in a variety of disciplines, it has been critiqued as too narrowly focusing on strategies to address individual deficits. For example, preventive efforts centered on preventing negative childhood health outcomes will not encourage positive development or promote normal development of children (Weissberg et al., 2003). This individual focus can overlook environmental conditions and raises the potential to overlook environmental factors in health problems (Wandersman and Florin, 2003). Although the foci and strategy of prevention may be debated by those from different disciplines, it is clear that prevention as a strategy for intervention and prevention science was well-established at the start of the twenty-first century.

The Development of Prevention Science

Prevention science encompasses a number of different areas of research, including research regarding the etiologies of specific physical, behavioral, or emotional problems and the development of effective interventions for particular disorders. An important emerging area of research focuses on understanding the science of implementing preventive interventions. As with many fields, there are debates about the merits of different emphases, with some arguing that additional research into etiologies is not necessary to develop effective interventions, or even that there has been a wealth of evidence-based interventions developed and the current critical need is to better understand the science of implementing those successful interventions.

Two important areas of discourse in prevention science are the concepts of prevention of disorders versus the promotion of wellness and the relationship between risk and resiliency.

Prevention and Promotion

While a range of prevention and promotion strategies are generally acknowledged in prevention science and prevention policy, the relative emphases on the prevention of disorder and the promotion of wellness and competence can conflict when considering the commitment of resources to each. Advocates of each view are ardent in their preferences for the focus of prevention science, with those focusing on prevention arguing that we have developed a great deal of knowledge regarding

how to prevent specific disorders and those focusing on wellness promotion arguing that the absence of a specific disorder does not ensure psychological well-being. These advocates point to an extensive knowledge base that has developed regarding the promotion of sound health and social competence (Kloos et al., 2012).

As this discussion illustrates, the goals of preventing specific disorders and promoting wellness and competence are not necessarily mutually exclusive. The implementation and effects of these strategies may be very similar in particular circumstances. For example, the prevention of obesity and the promotion of physical activity may use similar interventions and have similar outcomes.

Adding to the complexity of this debate are the related recent findings that the complete results of prevention efforts cannot be known without the collection of longitudinal data, stretching over decades. While these data are difficult to collect, there is evidence of the results of some prevention programs amplifying over time, such as in the Perry Preschool Project, or even cascading to produce benefits entirely unpredicted by the original program designers.

For example, recent research regarding the long-term effects of parent-training programs has found results as unexpected as improved educational, occupational, and income levels for mothers (Patterson et al., 2010). These cascading beneficial effects make sense theoretically when you consider the relationship between prevention and promotion. Parent-training programs such as the one developed by Patterson and his colleagues are designed to prevent the long-term problems associated with the early onset of conduct disorders in children. Extensive research has shown that these problems can be prevented by training parents to interact with their children in new ways, eliminating the coercive patterns of interaction generally found in those families. Patterson et al. hypothesize that teaching mothers to relate in positive noncoercive ways with their children also results in healthier (and happier) relationships with romantic partners, friends, colleagues, and supervisors. In other words, the same skills that help mothers to prevent serious behavior problems in their children also serve to promote positive relationships in other parts of their lives.

A related discussion in prevention science is the increased understanding of the relationship between risk and resiliency.

Risk and Resiliency

The concept of resiliency has gained immense importance in prevention science since it was first introduced by Emmy Werner as the central finding in her 40-year longitudinal study on the island of Kauai. Werner and her colleagues followed all 698 children born on the island in 1955. Of those children, 30% were exposed to one or more risk factors in early infancy, including complications surrounding the pregnancy and birth, poverty, family disruption, parental psychopathology, and mothers with limited education. Two-thirds of the children exposed to four or more of these risk factors experienced significant difficulties in childhood or adolescence. The negative effect these risk factors had on child development was clearly evident in the research findings.

But that was not the finding that was most surprised and impressed Werner and her colleagues. They were struck by that remaining one-third, the children who were exposed to four or more risk factors but still developed into 'competent, confident, and caring adults' (Werner, 1996). Subsequent research has identified numerous factors associated with resiliency, which can be usefully grouped into categories of individual differences, relationships, and community resources and opportunities (Masten and Powell, 2003).

Risk factors in child development had been identified and debated for decades before an understanding of resiliency factors began to emerge. It is only after solid research on both risk and protective factors became available that it was possible to begin the development of comprehensive approaches to promote positive development. Examples of such approaches include The Search Institute and Communities That Care. While there are some differences between the programs, both work on a community level, on the basic assumption that a community that actively reduces risk factors in child development and strengthens protective factors will support healthy development among all its children. These programs explicitly work to prevent problems through the promotion of wellness in child development.

Each program works by educating community leaders about risk and protective factors, and then helping them to assess the presence or absence of those factors in their communities, investigate model programs for reducing risk factors and increasing protective factors, and implement and evaluate those programs. These programs have been demonstrated to be quite successful. A cost-effectiveness evaluation of community-based programs in Pennsylvania implemented largely through the Communities That Care model demonstrated returns of up to $25 for every dollar invested by the Commonwealth. A single program could generate up to $130 million dollars in savings (Jones et al., 2008).

Methodological Challenges

Developing, implementing, and assessing prevention programs present several methodological challenges for the field of prevention science. While assessing the effectiveness and cost-effectiveness of the prevention efforts can be difficult, these goals can be largely met through established methodologies. That is not necessarily true for the challenges inherent in effectively implementing and disseminating programs to communities.

Assessing Effectiveness

Proverbs and fables from around the world testify to our belief in the effectiveness of prevention efforts. But this deep cultural belief in the effectiveness of prevention efforts is no substitute for empirical evidence. Providing that evidence presents several methodological challenges, both theoretical and pragmatic in nature.

The basic challenge is central to the whole concept of prevention. Prevention programs are intended to prevent something from occurring and it is difficult to demonstrate that an event that would have occurred without intervention was

prevented from occurring. Often the event to be prevented and the intervention designed to prevent it occur years or even decades apart, such as the prevention of adult criminal behavior through the provision of high-quality day care.

At this point there is a great deal of evidence demonstrating that effective prevention programs have been developed for a wide variety of topic areas. For example, meta-analyses of prevention programs aimed at children have shown efficacy for programs aimed at teen pregnancy, school bullying, childhood obesity, childhood and adolescent depression, child and adolescent drug abuse, programs implemented in infancy or early childhood, mentoring programs, and school-based programs to enhance social–emotional learning skills (Durlak and Wells, 1997, 1998; DuBois et al., 2011; Weissberg and Durlak, 2006).

The state of the field is also such that there is excellent information regarding best practices to ensure that a prevention program will be effective. Programs should be theory driven and evidence based. They should be comprehensive, including multiple interventions in multiple settings to achieve interrelated goals. They should be appropriately timed, of a sufficient dosage to have a lasting effect, and be socioculturally relevant. Effective programs include a strong behavioral component, teaching relevant skills and providing opportunities to practice those skills. They should focus on developing positive relationships, and instituting second-order change by fostering planned changes in institutions, communities, and policies. Finally, effective prevention programs provide significant support for the staff members implementing the program and include a planned program evaluation component (Nation et al., 2003).

Assessing Cost-Effectiveness

Many prevention programs are quite inexpensive to implement, such as mentoring programs that rely largely upon volunteers. Even these inexpensive programs still come with some monetary costs, and for some programs those costs are considerable. For example, Nurse Home-Visiting programs that begin prenatally have been shown to increase prenatal care, decrease low-birth-weight babies, improve parenting, and have long-term positive effects on various aspects of child development (Howard and Brooks-Gunn, 2009). But these programs are quite expensive to implement, since they utilize nurse practitioners and other professional or paraprofessional support staff.

These differences in cost raise two important questions for prevention science. First, if there is more than one prevention intervention or policy that has demonstrated effectiveness in the area of interest, which will achieve the program goals for the lowest cost? Second, at what point do the monetary benefits of a prevention program exceed the costs of the program? The first question can be answered through a cost-effectiveness analysis and the second through a cost-benefit analysis.

Cost-effectiveness analyses compare multiple prevention programs to determine which program results in the greatest benefit for the least cost. An early example is the analysis by the RAND Corporation comparing California's three-strikes law to four evidence-based programs that prevented criminal behavior. The analysis found that two of the four programs were more cost-effective than the three-strikes law (Greenwood et al., 1998).

Cost-benefit analyses compare the cost of providing a prevention program with the cost savings that are ultimately achieved by that program. Essentially, it is estimating the return on the investment in a prevention program. Multiple studies demonstrate that even relatively expensive programs can demonstrate significant returns on the public investment, in the range of $10 returned for every dollar spent and higher (Jones et al., 2008; Reynolds et al., 2011; Schweinhart et al., 2005).

These analyses face many challenges. While estimating the cost of programs may be relatively simple, estimating the savings generated from those programs is not. Many of the cost savings come from decreased use of the criminal justice system, special education services, and income support programs. But in addition to those savings, there are also returns to society in increased employment, income, and income tax revenue. Predictive models, which estimate these returns to society going forward, are obviously much more tentative than longitudinal studies that derive data from the actual life experiences of their participants. While these types of longitudinal studies are difficult, time-consuming, and rare, it is exactly these studies that are documenting some of the highest rates of return for prevention programs (Reynolds et al., 2011; Schweinhart et al., 2005).

Program Implementation and Dissemination

There is a major gap between prevention science research and the programs that are commonly occurring in schools and community settings. For example in 2011, a report of a national survey of 2500 public school districts and 5847 schools found that many schools offered crime prevention and substance-abuse prevention programs (U.S. Department of Education, 2011). However, only a small proportion (7.8%) was evidence-based programs and only a minority was implemented with minimal standards of fidelity. Therefore, only 3.5% were both research based *and* implemented with fidelity. This clearly illustrates the importance of implementing with quality and disseminating information about evidence-based programs with appropriate support for implementing with quality. In 2008, Wandersman and colleagues collaborated with colleagues at the Centers for Disease Control and Prevention (CDC) to develop the Interactive Systems Framework for Dissemination and Implementation (ISF) (Wandersman et al., 2008) to help bridge research and practice. The ISF describes the three systems listed below, and their interconnections, that are crucial for bridging research and practice.

Synthesis and Translation System

There are numerous articles in many journals about a topic area (e.g., childhood obesity). For the information to be widely used in practice, there needs to be a synthesis of the literature (e.g., literature reviews, meta-analyses). However, the synthesis articles that result rarely come with 'how-to's' for the practitioner. Therefore, it is necessary to have user-friendly translations of the materials.

The Delivery System

The Delivery System consists of the organizations and collaborations that provide services to residents, consumers, students, etc. (e.g., schools, mental health agencies, Boys and Girls Clubs). These organizations have general capacity (e.g., number and type of personnel, organizational structure, fiscal, and technical resources) and must acquire the capacity to implement an intervention with quality. The premise is that the Synthesis and Translation System will provide evidence-based information about what to do (and not to do) in order to achieve positive health and human service outcomes. However, as the 2011 study suggests – evidence-based programs implemented with quality are a small minority of the programs (U.S. Department of Education, 2011). It is hypothesized, and there is evidence (e.g., Joyce and Showers, 2002) that support in the form of training and technical assistance/coaching helps lead to better quality implementation.

The Support System

The third system in the ISF is the Support System. Often, there are individuals or organizations that provide support (e.g., tools, training, technical assistance) to assist the Delivery System in developing the capacity to implement programs with quality. Working with the idea that if it is important to be evidence-based about programs, Wandersman et al. (2012) describe the importance of delivering support in an evidence-based/evidence-informed way.

Emerging Directions

In addition to the ongoing issues discussed above, there are several emerging topics in the field of prevention science that warrant exploration. Three of them will be discussed here: the relative emphasis on program fidelity versus program innovation, the importance of incorporating cultural knowledge and values into prevention programming, and the relationship between prevention science and public policy.

Program Fidelity versus Program Innovation

A major discussion in the field of prevention science is the dynamic tension between the need for fidelity in the implementation of programs that have demonstrated effectiveness, and the benefits of modifying model programs to better fit local context. The importance of program fidelity has been well documented and the push for evidence-based programming has a solid empirical basis. But an insistence on strict adherence to every aspect of model programs has its own drawbacks. Must every aspect of a program be faithfully reproduced in order to ensure the replication of beneficial effects? Research suggests that this is not the case, and that in fact there are *core components* of interventions that drive the effectiveness and must be carefully replicated (Elias et al., 1997).

Research identifying those key components is a particularly important goal for prevention science since there are also excellent reasons to encourage innovation in program implementation (Durlak and DuPre, 2008). Programs that are adapted to address the specific contexts in which they are implemented are more likely to be successful, and these types of innovations, if done carefully and with documentation, can lead to important advancements in prevention science. The Interactive Systems Framework for Dissemination and Implementation described earlier provides one avenue to ensure that programs are implemented with a high level of quality, even when those implementing the program choose to make targeted adaptations.

One area of particular importance for program innovation has to do with cultural context.

Cultural Diversity

In the field of prevention science, there is a growing emphasis on the adaptation of preventive interventions to diverse communities. The American Psychological Association's recently adopted Guidelines for Prevention (2013) encourages, for example, the use of "socially and culturally relevant preventive practices adapted to the specific context in which they are implemented" in its set of nine guidelines, a recommendation that is second only to the use of evidence-based interventions. Research has demonstrated that effective prevention programs attend to relevant cultural and contextual factors, and that successful implementations of evidence-based programs frequently include significant adaptations (Durlak and DuPree, 2008). Implementation of evidence-based interventions in different cultural contexts is understood by many in the field to be the next research frontier (O'Connell et al., 2009).

In countries around the world, populations are becoming increasingly diverse. Societies are also growing more aware of social inequities and how they lead to educational and health disparities marked by race, gender, social class, culture, ethnicity, sexual orientation, ability/disability, age, and religion. Diversity within and between communities is an important part of understanding when and where particular preventive interventions may be most effective. Attention to diversity also helps prevention researchers and practitioners to acknowledge their own cultural worldviews, including individually held values and collective professional values. What one chooses to study, what one believes needs preventing, and how one chooses to intervene depend upon values, assumptions, and definition of the problem to be prevented. Problem definition also affects how successful program outcomes are defined and who decides what counts in determining program success.

Grappling with these issues of problem definition and valued outcomes in a successful cultural adaptation requires collaboration with community members affected by the target intervention. Community collaboration also improves prevention research and evaluation – enhancing research questions, helping to address research challenges that emerge, and enhancing the usefulness of results. Community involvement significantly enhances the chance of sustainability and long-term success of interventions as well. Some scholars suggest immersing oneself in the cultural community to become a more sensitive partner; others suggest participatory research models aimed at empowering and building capacity at the local level. At the foundation of different contemporary approaches, however, is an orientation toward dialogue, with

researchers and other stakeholders explicitly stating values, assumptions, and priorities to avoid a prevention focus that is not relevant or salient – or even worse, furthers inequalities or oppression.

Meaningful adaptation of prevention programs reflects the integration of cultural values at a deep structural level rather than just on a surface level (Nation et al., 2003). As noted in the previous section, prevention researchers are currently focused on finding a successful balance of meaningful adaptation with program fidelity, emphasizing the retention of core program components and an adequate strength or 'dosage' of a program across settings (Kumpfer et al., 2002). One of the key challenges in the field is to figure out the essential 'core' ingredients for a particular intervention to be successful as it is adapted across diverse cultural contexts. With this emphasis on finding the right blend of program fidelity and adaptation, it is becoming increasingly clear that program implementation in any specific context is necessarily also a process of program innovation.

Another challenge for the field is to work across levels of analysis, examining not just processes and outcomes for individual-level change, but collecting evidence of organizational and systems change. As discussed previously, preventive interventions often work to identify risk factors within communities that have high rates of poverty and/or people of color. Thus, researchers have noted the need to identify culturally relevant strengths at individual and community levels to implement relevant and valued interventions and to avoid pathologizing entire communities (Kenny and Hage, 2009). A health-promotion framework for prevention enables thinking about interventions that build on strengths within minority communities (National Black Child Development Institute, 2013) and working with majority groups to promote economic and social inclusion within and across communities (American Psychological Association, Presidential Task Force on Preventing Discrimination and Promoting Diversity, 2012). For example, new efforts to prevent interpersonal and institutional racism work across multiple levels of analysis, not just focusing on individuals and groups, but also focusing on social institutions and systems that perpetuate inequality and exclusion. They include interpersonal interventions drawing on social psychology's contact theory to prevent the formation of racist attitudes and beliefs, as well as public policy advocacy as an organizational and systems-level prevention strategy (Buhin and Vera, 2009).

Prevention and Public Policy

Earlier we cited research demonstrating the cost-effectiveness of community-based prevention programs in Pennsylvania (Jones et al., 2008). The research found that the programs generated up to $25 in savings for every dollar invested by the Commonwealth. Much of these savings came from decreased costs to the criminal justice system. Home-visiting programs, high-quality day care, mentoring programs, and parenting training programs have all been shown to reduce crime and crime-related costs to the point where the programs more than pay for themselves. Yet, in the United States there has been an immense increase in public funding for incarceration in recent decades, at the same time that funding for prevention programming has been cut.

This disconnect between the findings of prevention science and public policies is widespread and a source of frustration to many prevention scientists. The goal is not just to advocate for the adoption of specific programs. Rather, their goal is to provide broad education to policy makers regarding the benefits of prevention approaches. The importance of this work is illustrated by the Guidelines for Prevention in Psychology, adopted by the American Psychological Association. Guideline 9 specifically states that "Psychologists are encouraged to inform the deliberation of public policies that promote health and well-being when relevant prevention science findings are available" (American Psychological Association, 2012). As we hope we have illustrated in this article, a wealth of prevention science exists. Perhaps the major challenge faced by the field is to ensure that our institutions, communities, and public policies are informed by that knowledge.

Summary

In this article we reviewed the growth of prevention from a cultural value to specific areas of study across many disciplines. The future holds promise as we look to develop more systematic approaches to engaging communities to address their needs and more effectively promote policies that are both cost-effective and promote health and well-being for various groups in our communities.

See also: Drugs: Illicit Use and Prevention; Epidemiology, Concepts and Methods of; Health Education and Health Promotion; Prevention of Crime and Delinquency; Resilience; Social Epidemiology.

Bibliography

Albee, G.W., 1959. Mental Health Manpower Trends. Basic Books, New York.

Albee, G.W., 1982. Preventing psychopathology and promoting human potential. American Psychologist 37, 1043–1050.

American Psychological Association, Prevention Guidelines Work Group, 2012. Guidelines for Prevention in Psychology. American Psychological Association, Washington, DC.

American Psychological Association, Presidential Task Force on Preventing Discrimination and Promoting Diversity, 2012. Dual Pathways to a Better America: Preventing Discrimination and Promoting Diversity. American Psychological Association, Washington DC.

Buhin, L., Vera, E., 2009. Preventing racism and promoting social justice: person-centered and environment-centered interventions. Journal of Primary Prevention 30, 43–59.

Caplan, G., 1964. Principles of Preventive Psychiatry. Basic Books, New York.

Dohrenwend, B.S., 1978. Social stress and community psychology. American Journal of Community Psychology 6, 1–14.

DuBois, D., Portillo, N., Rhodes, J., Silverthorn, N., Valentine, S., 2011. How effective are mentoring programs for youth? A systematic assessment of the evidence. Psychological Science in the Public Interest 12, 57–91.

Durlak, J.A., Wells, A.M., 1997. Primary prevention mental health programs for children and adolescents: a meta-analytic review. American Journal of Community Psychology 25, 115–152.

Durlak, J., Wells, A., 1998. Evaluation of indicated preventive intervention (secondary prevention) mental health programs for children and adolescents. American Journal of Community Psychology 26, 775–802.

Durlak, J.A., DuPre, E.P., 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting the implementation. American Journal of Community Psychology 41, 327–350.

Edgerley, L., El-Sayed, Y., Druzin, M., Kierman, M., Daniels, K., 2007. Use of a community mobile health van to increase early access to prenatal care. Maternal and Child Health Journal 11, 235–239.

Elias, M.J., 1987. Establishing enduring prevention programs: advancing the legacy of Swampscott. American Journal of Community Psychology 15, 539–553.

Elias, M.J., Zins, J., Weissberg, R.P., Frey, K., Greenberg, M., Haynes, N., Kessler, R., Schwab-Stone, M., Shriver, T., 1997. Promoting Social and Emotional Learning: Guidelines for Educators. Association for Supervision and Curriculum Development, Alexandrian, VA.

Fairweather, G.W., 1979. Experimental development and dissemination of an alternative to psychicatric hospitalization. In: Munoz, R., Snowden, L., Kelly, J.G. (Eds.), Social and Psychological Research in Community Settings. Jossey-Bass, San Francisco, pp. 305–342.

Greenwood, P., Model, K., Rydell, P., Chiesa, J., 1998. Diverting Children from a Life of Crime: Measuring Costs and Benefits. Rand Corporation. http://www.rand.org/pubs/monograph_reports/MR699-1.html.

Hobbs, N., 1964. Mental health's third revolution. American Journal of Orthopsychiatry 34, 822–833.

Hollingshead, A., Redlich, F., 1958. Social Class and Mental Illness: A Community Study. Wiley, New York.

Howard, K., Brooks-Gunn, J., 2009. The role of home-visiting programs in prevention child abuse and neglect. The Future of Children 19, 119–146.

Jason, L.A., Berk, M., Schnopp-Wyatt, D.L., Talbot, B., 1999. Effects of enforcement of youth access laws on smoking prevalence. American Journal of Community Psychology 27, 143–160.

Jones, D., Bumbarger, B., Greenberg, M., Greenwood, P., Kyler, S., 2008. The Economic Return on PCCD's Investment in Research-based Programs: A Cost-Benefit Assessment of Delinquency Prevention in Pennsylvania. The Prevention Research Center for the Promotion of Human Development, The Pennsylvania State University.

Joyce, B.R., Showers, B., 2002. Student Achievement through Staff Development. ASCD.

Kenny, M.E., Hage, S.M., 2009. The next frontier: prevention as an instrument of social justice. The Journal of Primary Prevention 30, 1–10.

Kloos, B., Hill, J., Thomas, E., Wandersman, A., Elias, M.J., Dalton, J.H., 2012. Community Psychology: Linking Individuals and Communities, third ed. Wadsworth/Cengage, Belmont, CA.

Kumpfer, K.L., Alvarado, R., Smith, P., Bellamy, N., 2002. Cultural sensitivity and adaptation in family-based prevention interventions. Prevention Science 3, 241–246.

Lykes, C., Kay, L., Crepaz, N., Herbst, J., Passin, W., Kim, A., Rama, S., Thadiparthi, S., DeLuca, J., Mullis, M., 2007. Best-evidence interventions: findings from a systematic review of HIV behavioral interventions for US populations at high risk, 2000–2004. American Journal of Public Health 97, 133–143.

Masten, A., Powell, J., 2003. A resilience framework for research, policy, and practice. In: Luthar, S. (Ed.), Resilience and Vulnerability: Adaptation in the Context of Childhood Adversities. Cambridge University Press, New York, pp. 1–25.

Mrazek, P., Haggerty, R., 1994. Reducing Risks for Mental Disorders: Frontiers for Preventive Intervention Research. National Academy Press, Washington, DC.

Nation, M., Crusto, C.,A., Kumpfer, K.L., Seybolt, D., Morrissey-Kane, E., et al., 2003. What works in prevention: principles of effective prevention programs. American Psychologist 58, 449–456.

National Black Child Development Institute, 2013. Being Black Is Not a Risk Factor: A Strengths-Based Look at the State of the Black Child.

O'Connell, M.E., Boat, T., Warner, K.E. (Eds.), 2009. Preventing Mental, Emotional, and Behavioral Disorders among Young People: Progress and Possibilities. National Academies Press, Washington, DC.

Patterson, G., Forgatch, M., DeGramo, D., 2010. Cascading effects following intervention. Development and Psychopathology 22, 949–970.

Reynolds, A., Temple, J., White, B., Ou, S., Robertson, D., 2011. Age 26 cost-benefit analysis of the child-parent center early education program. Child Development 82, 379.

Schweinhart, L.J., Barnes, H.V., Weikart, D.P., 1993. Significant Benefits: The HighScope Perry Preschool Study through Age 27 (Monographs of the HighScope Educational Research Foundation, 10). HighScope Press, Ypsilanti, MI.

Schweinhart, L.J., Montie, J., Xiang, Z., Barnett, W.S., Belfield, C.R., Nores, M., 2005. Lifetime Effects: The HighScope Perry Preschool Study through Age 40 (Monographs of the HighScope Educational Research Foundation, 14). HighScope Press, Ypsilanti, MI.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011. Prevalence and Implementation Fidelity of Research-Based Prevention Programs in Public Schools: Final Report, Washington, DC.

Wandersman, A., Chien, V.H., Katz, J., 2012. Toward an evidence-based system for innovation support for implementing innovations with quality: tools, training, technical assistance, and quality assurance/quality improvement. American Journal of Community Psychology 50, 445–459.

Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K., Stillman, S., Blachman, M., Dunville, R., Saul, J., 2008. Bridging the gap between prevention research and practice: the Interactive Systems Framework for Dissemination and Implementation. American Journal of Community Psychology 41, 171–181.

Wandersman, A., Florin, P., 2003. Community interventions and effective prevention. American Psychologist 58, 441–448.

Weissberg, R.P., Durlak, J., 2006. Meta-analysis of the Effect of Social-Emotional Learning and Positive Youth Development Programs on Academic Achievement and Problem Behaviors. Available from: www.CASEL.org.

Weissberg, R.P., Kumpfer, K., Seligman, M., 2003. Prevention that works for children and youth: an introduction. American Psychologist 58, 425–432.

Webster-Stratton, C., 1984. A randomized trial of two parent-training programs for families with conduct-disordered children. Journal of Consulting and Clinical Psychology 52, 666–678.

Werner, E., 1996. Vulnerable but invincible: high risk children from birth to adulthood. European Journal of Child and Adolescent Psychiatry 5, 47–51.

Relevant Websites

http://www.colorado.edu/cspv/blueprints/ – Blueprints for Healthy Youth Development- The Center for the Study and Prevention of Violence, University of Colorado Boulder.

http://www.cdc.gov/ – Centers for Disease Control and Prevention.

http://www.communitiesthatcare.net – Communities That Care.

http://www.preventviolence.info/evidence_base.aspx – Centre for Public Health, Liverpool John Moores University.

https://www.crimesolutions.gov/ – Crime Solutions.

http://www.nrepp.samhsa.gov/ – The National Registry of Evidence-Based Programs and Practices- The Substance and Mental Health Services Administration.

Perhaps the strongest evidence of the growth of prevention science is the number of online databases promoting the implementation of evidence-based prevention programs. These databases are specifically designed to give community members and policy makers a searchable collection of prevention and promotion programs with at least some scientific evidence of effectiveness.

http://www.search-institute.org/ – The Search Institute.

Primate Behavior: Significance for Understanding Humans

Roberto A Delgado, University of Southern California, Los Angeles, CA, USA
Elizabeth A Fox, Duke University, Durham, NC, USA
Jeanne Altmann, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.A. Fox, J. Altmann, volume 18, pp. 12017–12023, © 2001, Elsevier Ltd.

Abstract

Into the twenty-first century, primate biology remains a critical area of research for understanding the evolutionary ecology of human social behavior. In this article, we briefly outline the history of primate behavioral studies within the field of physical, now biological, anthropology, discussing past uses of primatology for building hypotheses for human behavioral evolution. We then discuss changes in the theoretical approach to the study of primate behavior and their effects on the understanding of social systems. We conclude by presenting current issues in primatology as they relate to the study of human sociality and behavioral evolution.

Introduction

Into the twenty-first century, primate biology remains a critical area of research for understanding the evolutionary ecology of human social behavior. The modern anthropological subfield of primatology began with anatomical comparisons of nonhuman primates to reconstruct the morphological evolution of humans, and interest in the behavioral similarities between human and nonhuman primates can be traced to Huxley's (1863) assertion of a phylogenetic relationship between humans and African apes. Since their inception, primate studies have been incorporated into diverse fields of inquiry beyond anthropology that include biology, ecology, and psychology. Physical anthropologists were interested principally in the degree to which primates and their behavior could be utilized as a model for reconstructing the behavior and cognitive capacities of ancestral humans. Contemporary primate studies continue to inform a broad array of fields in both the human and nonhuman behavioral sciences, and this integrative approach contributes to our understanding of both human evolution and the behavior of extant humans (Washburn, 1973). In this article, we briefly outline the history of primate behavioral studies within the field of physical, now biological, anthropology, discussing past uses of primatology for building hypotheses for human behavioral evolution. We then discuss changes in the theoretical approach to the study of primate behavior and their effects on the understanding of social systems. We conclude by presenting current issues in primatology as they relate to the study of human sociality and behavioral evolution.

Early Primate Models of Ancestral Hominin Behavior

Early primate models for the behavior of ancestral humans were based on either (1) assumptions of similarity in ecological adaptation between extant savannah-living primates and ancestral humans, or (2) shared behavioral traits within the clade of African hominoids. Raymond Dart, the discoverer of the late Pliocene *Australopithecus* specimen at Taung, in South Africa, popularized the hypothesis of a saltatory shift to human bipedalism, tool use, and hunting as a result of moving from forested habitats to savannahs (Dart, 1957). Specifically, Dart drew explicit parallels to baboons as contemporary savannah dwellers that might share behavioral similarities with ancestral humans. However, subsequent, though brief, field studies of wild baboons revealed distinct differences from modern humans. Nonetheless, early descriptions of baboon sociality (DeVore and Washburn, 1963), which emphasized male dominance and aggression as important influences on social structure, were used to typify the human condition (Kinzey, 1987). The historical prejudices and methodological weaknesses that gave rise to these models have received considerable critical scrutiny (Cartmill, 1993) and, although both the conceptual approaches and practical challenges to fieldwork were anticipated (Carpenter, 1945), these early studies planted the seeds for future primate field studies by asking the following questions: (1) In what way do ecological variables shape social systems? (2) What were the relevant ecological factors that shaped early human evolution? and (3) What can extant primates tell us about human sociality and behavioral evolution?

Ecological affinities led to the baboon model, whereas phylogenetic affinities between humans and nonhuman apes prompted field studies of chimpanzees, gorillas, orangutans, and bonobos (Hamburg and McCown, 1979). Taking a comparative approach, Louis Leakey sponsored well-publicized field studies of three great ape species. Anthropologists focused most attention on the common chimpanzee because of their close anatomical, behavioral, and genetic affinities with humans (Sibley and Alquist, 1987). Behaviors observed in both humans and chimpanzees were assumed to be phylogenetically conservative and therefore likely to have been expressed in our last common ancestor. For example, tool use and manufacture; meat eating, with cooperative hunting and sharing; and the mixed savannah/woodland landscape of chimpanzees at Gombe, in Tanzania, reinforced the importance of this suite of characteristics to the evolution of modern humans (Kinzey, 1987).

However, contrary to expectation, these studies revealed that neither baboon nor chimpanzee models necessarily have strong explanatory power in reconstructing the behavior of ancestral humans. First, extant species evolved under unique selective pressures and are not satisfactory typological specimens of ancestral species. Second, there is significant intraspecific variation in the behaviors of chimpanzees and baboons, with no adequate criteria for selecting which behaviors are applicable to models of ancestral humans. Third, if behaviors common to chimpanzees and humans are assumed to be phylogenetically conservative, there are significant problems in interpreting from the fossil record morphological evidence that is not shared by either modern chimpanzees or humans (Potts, in Kinzey, 1987).

These criticisms apply to the assumptions and methods employed by these modeling approaches and not to the usefulness of primate behavioral studies per se in elucidating ancestral and current human behavior. Past mistakes have served to refine the approaches by which primate studies are conducted and applied. At the beginning of the twenty-first century, few primate studies now make explicit use of typological models. Instead, contemporary studies of primate behavior emphasize a comparative approach in which the functions of behaviors are investigated under varying ecological and social conditions. The next section presents important conceptual changes in the study of primate behavior and their influence on our understanding of human behavior.

Conceptual Shifts in Primate Behavior Studies

During the same time that primate behavior was being used to construct models of ancestral hominines, it was also an area of active interest in the zoological sciences, particularly in the subfield of behavioral ecology. Behavioral ecology investigates behavioral adaptations to varying ecological and social conditions. Although behavioral ecologists directly observe phenotypic traits and behavioral strategies, they use these observations to test falsifiable predictions generated within an evolutionary framework in which natural selection favors individuals who act in ways that increase their inclusive fitness. For example, studies testing optimal foraging theory use economic cost–benefit models to generate predictions of behavioral responses to ecological conditions that affect an individual's energy budget. Other studies, particularly those that focus on reproductive behavior such as mating and offspring care, examine an individual's decisions when situated in a particular social milieu (Krebs and Davies, 1991). This approach to the study of primate behavior is distinctly different from that which was used previously in two important ways: (1) its goal is to test and further develop general evolutionary principles for behavior rather than create models of specific extinct species, and (2) comparative outgroups can be taken from all animal taxa rather than those restricted to the primate order.

Within the field of anthropology, human behavioral ecologists applied evolutionary concepts to test economic models of the ecology of contemporary human hunter-gatherers. Human behavioral ecologists focused on contingencies that had been found to be important in nonhuman primates: predation, reproductive opportunities, and the spatiotemporal distribution and abundance of resources. However, the application of primate studies to the interpretation of human behavior encountered significant difficulties following the publication of E. O. Wilson's *Sociobiology* (1975). Criticism of early sociobiological theory is well documented and expressed two main concerns. First, the effusion of speculations about the application of sociobiological principles to human behavior was untestable and therefore unfalsifiable (Caro and Borgerhoff Mulder, 1987). The second body of criticism, which objected to sociobiological depictions of humans as inflexible executants of the interests of selfish genes, stemmed from a poorly explained differentiation between genotypic determinism and the phenotypic expression of behavioral flexibility. Suites of behaviors resulting from complex socioenvironmental conditions (e.g., rape) were incorrectly treated as simple units of evolutionary analysis (Sunday and Tobach, 1985). Disagreement about the application of sociobiological theory to the study of human behavior resulted in a division between cultural anthropologists, who emphasized human uniqueness, and those biological (formerly physical) anthropologists who studied primate behavior with an eye toward understanding that of humans. Although it is a truism that genetic makeup influences behavior, and genetic relatedness influences to whom particular behaviors may be expressed, genetic determinism has no place in behavioral ecology.

The contemporary use of primate behavior within anthropology has shifted from the use of primates as typological models of human origins to field studies and experiments that test predictions developed within a behavioral ecology framework (Rodman, 1999). Significant methodological changes have accompanied this theoretical shift. First, standard methods for the collection of quantifiable data facilitate interpretable statistical tests of predictions and comparison across primate studies (Altmann, 1974). Second, more primate species are studied, thereby providing larger and more diverse samples for comparison. Third, long-term field studies provide information on seasonal and interpopulational variability in behavior, the ontogeny of and change in relationships among individuals, and the causes and consequences of demographic processes among long-lived species. Fourth, a more diverse representation of scientists expands the field of inquiry to include aspects of social life previously assumed ancillary to social organization, particularly female reproductive strategies and the energetics of offspring care (Altmann, 1980; Hrdy, 1999). Analyses of the resulting data sets demonstrate rich intra- and interspecific variabilities in behavior and relationships that parallel the intraspecific variability observed in humans.

Current Primate Studies in Anthropology

Social Behavior

Two general principles are commonly invoked to explain group size and composition in birds and mammals. First, female reproductive success is limited primarily by the availability of ecological resources, particularly food. With whom females form groups, and the competitive regime between females

within those groups, is a function of the distribution of ecological resources in space and time (Wrangham, 1980; van Schaik, 1989). Male distribution, on the other hand, is patterned upon that of females (Andelmann, 1986). The diversity of grouping patterns observed among primates across a range of environments supports this general framework.

Whether of the scramble or contest type – a reflection of resource monopolization – competition for resources underlies this ecological model (Krebs and Davies, 1991). Competition, however, can be expressed in multiple ways. Aggression between group members certainly occurs, and individuals obtain inclusive fitness benefits by aiding kin in contests, but not all competition results in direct conflict. Unrelated individuals also engage in cooperative and affiliative behaviors, and reconciliation often occurs after conflicts (de Waal, 1983). Indeed, several benefits accrue from affiliative interactions with nonkin. First, individuals who engage in postconflict reconciliation experience more rapid reductions in stress responses than those that do not (Aureli, 1992). Second, initiating affiliative behaviors garners favor with powerful group members and develops relationships with potential future allies (de Waal, 1983). Third, males who invest in nonsexual affiliative behavior toward females obtain future matings with their 'friend' (Smuts, 1985). Peaceful interactions are essential to the social cohesion of groups and provide alternative means of exerting influence and eliciting support. Long life spans and the opportunity for repeated encounters between known individuals provide the conditions under which individuals can predicate present behavior related to anticipated future social interactions (Trivers, 1971; Byrne and Whiten, 1988).

The study of behavioral ecology in humans is likewise concerned with the tactics employed by modern humans to survive and reproduce within a particular socioecological context. Using principles developed in the field of animal behavior, human behavioral ecologists collect data from modern hunter-gatherers to test predictions derived from ecological optimization models related to food acquisition, time allocation, and habitat use (Winterhalder and Smith, 2000). Human behavioral ecologists are able to measure directly the economic results of behavioral traits thought to be important to human evolution, such as altruistic cooperation and food sharing, the intersexual division of labor, long juvenile dependency periods, and females' lengthy postreproductive life spans. The results of such studies have led to important challenges to the primacy of hunting in human history, suggesting instead that the economics of female provisioning and care of offspring provide a more parsimonious explanation for the evolution of human social systems (Hawkes et al., 1997; O'Connell et al., 1999).

Despite the general applicability of behavioral ecology models, there is strong evidence that primate studies in particular are more informative for understanding human sociality and behavioral evolution than research with other animal taxa. While studies of other taxa, particularly social mammals, serve to test theories that may be relevant to humans regarding the causes and consequences of different social systems, the complex nature of social relationships observed in nonhuman primates make them particularly useful as comparative species for understanding the function of human social relationships. In addition, the use of nonhuman primates as comparison species controls for traits more phylogenetically conservative than behavior that may influence alternative strategies (e.g., functional morphology, life history parameters, and physiology). Confining the study of behavioral ecology to humans within the field of anthropology limits the perspective from which to understand human behavior. For example, cross-cultural studies can demonstrate intraspecific variation, but they do nothing to reveal which traits or behaviors are uniquely human. Interspecific comparative studies among primates reveal the extent and degree to which certain behavioral patterns are confined to humans. Hence, primate studies both focus and expand the field of vision in the study of human behavior.

Functional Morphology

Anatomists have a rich tradition of inferring the behavior of extinct species based on fossil skeletons, and tooth morphology figures prominently in the reconstruction of hominin diets (Jolly, 1970). Interpretation of fossils is possible, however, only with studies of analogous structures and their function in extant species, but researchers must also be careful in their phylogenetic interpretations because shared skeletal morphology may not always mean shared evolutionary history (Wood and Harrison, 2011). Recently, anatomists have teamed up with primate behavioral ecologists to combine direct observations of primate feeding behavior with dental morphology. This partnership helps to refine the techniques used to determine ancestral diet and feeding patterns. For example, field studies of the feeding ecology of four sympatric Sumatran primates allowed Ungar (1996) to test whether incisor size corresponds with diet. His finding that incisor size and shape best reflect the mode of ingestion rather than diet per se resolves some conflicts in interpretations of the fossil record. Dental microwear analyses can reveal more detailed information about diet than can analysis of gross dental morphology. Recent field studies examine primate feeding behavior, measure the physical properties of foods, and cast the teeth of wild primates so as to better understand the relationship between food type, mode of ingestion, and microwear (Teaford, 1994). Similar studies are being conducted on humans (Ungar and Spencer, 1999). With further refinement of this new technique, analysis of dental microwear can play an important role in the reconstruction not only of the diet of ancestral hominines but the use of teeth for nonfeeding functions as well (Teaford, 1994).

Studies of positional behavior investigate the relationship between postcranial musculoskeletal anatomy and locomotion. The means by which wild great apes travel in their natural environments reveal the function of bone and muscle in the limbs (Fleagle, 1999), and the study of comparative anatomy between great apes and fossil hominines provides a basis for inferring locomotor behavior in hominines. This approach has shown that bipedal hominines descended from an arboreal primate that moved primarily by climbing rather than brachiating (Fleagle, 1999). Likewise, *Australopithecus*

afarensis, the hominin most likely responsible for the famous Laetoli footprints (Leakey, 1987), probably spent a good proportion of the day moving arboreally (i.e., from tree to tree) as well as bipedally (Stern and Susman, 1983). Field observations also provide clues to possible sex differences in behavior. In apes, sexual dimorphism corresponds with differential use of arboreal and terrestrial substrates for moving, with males spending more time on the ground than females (Doran, 1996). If the sexually dimorphic *A. afarensis* moved on the ground and in the trees, sex differences in substrate use may likewise have occurred.

Intelligence and Cognition

The comparative study of intelligence and cognitive abilities among nonhuman primates in the laboratory spans more than half a century (Russon et al., 1996). The discovery of tool manufacture and use in wild chimpanzees (Goodall, 1968) put to rest concerns that the development of the cognitive capacities necessary to use tools by laboratory primates was an artifact of intensive interaction with humans. Wild orangutans also manufacture and use tools with considerable complexity (Fox et al., 1999). In contrast, the proficiency with which capuchin monkeys use tools in both field and laboratory settings is significantly lower than that of great apes (Visalberghi and Trinca, 1989). Great apes have crossed the line into a domain of intelligence once postulated to be uniquely human; furthermore, the intelligence required for ape-level tool use and manufacture evolved before the orangutan–African ape split (van Schaik et al., 1996).

Tool use behavior, however, is only one demonstration of intelligence. The decision-making processes required for large-bodied apes to travel arboreally is another (Povinelli and Cant, 1995). Complexity in social interactions such as deception evinces cognitive skills as well (Byrne and Whiten, 1988). Primate studies provide empirical proof that many purported aspects of human uniqueness, such as intelligence and culture, are unique to humans only insofar as they are defined as such (Cartmill, 1990). The corpus of work in this field demonstrates that, broadly defined, intelligence has evolved by degrees rather than as a result of a single selective pressure on early hominids. Diverse levels and types of cognitive capacities in ecological and social domains for an array of primates elicit two questions about humans: (1) Is the higher intelligence demonstrated in humans an added degree of intelligence or is it a difference of kind, or both? and (2) What selected for human-level intelligence? Language, cumulative culture, and altruistic cooperation reflecting prosocial tendencies are obvious difference between humans and nonhuman primates, and they will undoubtedly figure prominently in answering these questions.

Future Directions in Research

Although most primate species live in cohesive social units with permanent male–female associations, groups also comprise individuals with interests that often differ from those of their groupmates. In addition, those interests change as individuals grow to maturity, reproduce, and then, in some species, shift into a postreproductive phase of life. Building upon earlier research, future studies are needed to investigate the complexity of factors that influence individual decisions within the context of group life, in particular, (1) the strategies that individuals employ to satisfy their reproductive interests, even when those interests conflict with other group members; and (2) reproductive decisions that incorporate not just matings or fertilizations, but offspring survival as well.

Male aggression, or the threat of aggression, toward females is observed in most species of primates including humans (Smuts, 1992; Smuts and Smuts, 1993). Some of this conflict is with current or potential mating partners. However, significant variation in rates of sexual aggression occurs between and within species (Arnqvist and Rowe, 2005). For example, woolly spider monkey females copulate without aggression from their mates or other males, while male sexual aggression in rhesus macaques occurs frequently. In baboons, particular males develop 'friendships' with females while others do not (Smuts and Smuts, 1993). There is even variation in individual males: some human males with pair-bonded mates force copulations with other females, and some pair-bonded males alternately exhibit affiliative and aggressive behavior toward their mates (Ellis, 1989). Despite the seemingly widespread occurrence among primate species, however, extreme forms of direct sexual aggression in the form of forced copulations are uncommon in primates other than humans and orangutans. Some hypotheses to explain this striking contrast have been proposed and deserve further exploration.

What is the function of male sexual aggression? Under what conditions does it occur? Feminist scholars and biological and social scientists have long recognized that conflict between female and male reproductive interests influence individual behavior and shape the social systems in which individuals live (Hrdy, 1977; MacKinnon, 1989; Arnqvist and Rowe, 2005). Theoretical treatments of intersexual mating conflict in human and nonhuman primates provide a basis for predicting the conditions under which male sexual aggression, and conversely female counterstrategies to thwart males, occur (Smuts, 1992; Smuts and Smuts, 1993). Recent studies provide some preliminary basis for comparison (Smuts and Smuts, 1993), but to strengthen the comparative method, more studies conducted across a greater spectrum of primates are necessary. Of particular importance is quantifying the costs of sexual coercion, such as reduced foraging efficiency, in addition to infant or female mortality. New noninvasive hormonal assay techniques for field-collected samples are starting to be used (Whitten et al., 1998) and will allow researchers to measure endocrine response by females who are subjected to male sexual aggression. Using these techniques, research can reveal the degree to which male aggression produces physiological effects that may negatively impact female fecundity.

Relatively little is understood regarding why females resist mating with some males but not others. Equally little is known about female tactics to circumvent male attempts to restrict female reproductive choice. More studies are needed to determine the degree to which females select mates based on variations in genetic quality or resource-holding potential among males (direct mate choice) or whether other factors, including

male aggression, sufficiently constrain females and lead to facultative, or indirect, female mate choice (Gowaty, 1992; Wiley and Poston, 1996). Whether for genetic quality, direct benefits, or constraints placed by males, with whom a female mates has an important reproductive outcome. Females are therefore predicted to exercise tactics, including sperm selection, that optimize their abilities to retain control over reproductive choice and the timing of reproduction. Techniques for genetic analysis using noninvasive sampling methods have been developed (Borries et al., 1999) and will in the future facilitate evaluation of the success of both male and female reproductive strategies.

Variability in male reproductive decisions is also important. In contrast to the previously pervasive view within evolutionary studies and behavioral ecology, a recent shift recognizes that the acquisition of multiple mates is not the sole factor motivating male reproductive decisions. Whether they are excluded by other males, fail to attract mates, or are constrained by the need to allocate time to nonmating activities, male primates are limited in the number of females with whom they can mate and produce viable young. Males, therefore, optimize reproductive effort within existing constraints. Limitations on male mating opportunities have traditionally been viewed in terms of intrasexual competition and mating aggression toward females, but this is not always the case. Male choosiness for mates is likely to be one important tactic in species that have long periods of infant and juvenile dependency (Altmann, 1997). Differences in female fertility, for example, may guide males away from young females as mates. Furthermore, conception and birth are merely the first steps to reproductive success; offspring must survive into adulthood to reproduce because reproductive success is most accurately measured in the second generation (Clutton-Brock, 1988). Offspring survival is critically dependent on maternal competence, including mothers' abilities to obtain access to resources necessary to rear young. Males are, therefore, expected to prefer those proven females who are older, higher ranking, and reproductively experienced (Altmann, 1997).

Females with insufficient resources to rear young may abandon current offspring to invest in future reproductive opportunities. Alternatively, to meet the demands of rearing offspring, mothers in some species share the task of offspring care. Fathers sometimes significantly assist females, and the need for paternal care constrains male promiscuity as it does in many species of birds (Clutton-Brock, 1991). Likewise, related or unrelated individuals sometimes assist mothers. Related individuals gain inclusive fitness benefits resulting from the survival of grandchildren, sisters, and cousins, while both related and unrelated individuals may gain needed experience necessary for the survival of their own future offspring (Hrdy, 1999). Long juvenile dependency and the need for assistance in rearing offspring play important roles in the lives of extant hunter-gatherers, and it figures prominently in models of human social evolution (O'Connell et al., 1999). The function of extended juvenile dependency, how mothers meet the demands of extensive infant care, and the consequences for reproductive success are among the most fascinating questions of human and nonhuman primate life histories. Yet because primate life spans are so long, these questions are also among the hardest to study. As data from uninterrupted long-term field studies become available, questions concerning the fitness advantages of variations in parental and alloparental care will be addressed.

See also: Comparative Method in Evolutionary Studies; Human Behavioral Ecology; Kin Selection; Primate Groups and Their Correlates; Primates, Evolution of.

Bibliography

Altmann, J., 1974. Observational study of behavior: sampling methods. Behaviour 49, 227–267.

Altmann, J., 1980. Baboons Mothers and Infants. Harvard University Press, Cambridge, MA.

Altmann, J., 1997. Mate choice and intrasexual reproductive competition: contributions to reproduction that go beyond acquiring more mates. In: Gowaty, P.A. (Ed.), Feminism and Evolutionary Biology. Chapman and Hall, New York.

Andelmann, S.J., 1986. Ecological social determinants of cercopithecine mating patterns. In: Rubenstein, D.I., Wrangham, R.W. (Eds.), Ecological Aspects of Social Evolution: Birds and Mammals. Princeton University Press, Princeton, NJ.

Arnqvist, G., Rowe, L., 2005. Sexual Conflict. Princeton University Press, Princeton.

Aureli, F., 1992. Post-conflict behaviour among wild long-tailed macaques (*Macaca fascicularis*). Behavioral Ecology and Sociobiology 31, 329–337.

Borries, C., Laundhardt, K., Epplen, C., Epplen, J.T., Winkler, P., 1999. DNA analyses support the hypothesis that infanticide is adaptive in langur monkeys. Proceedings of the Royal Society London 266, 901–904.

Byrne, R.W., Whiten, A. (Eds.), 1988. Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans. Clarendon Press, Oxford, UK.

Caro, T.M., Borgerhoff Mulder, M., 1987. The problem of adaptation in the study of human behavior. Ethology and Sociobiology 8, 61–72.

Carpenter, C.R., 1945. Concepts and problems of primate sociometry. Sociometry 8 (1), 56–61.

Cartmill, M., 1990. Human uniqueness and theoretical context in paleoanthropology. International Journal of Primatology 11, 173–192.

Cartmill, M., 1993. A View to a Death in the Morning: Hunting and Nature Through History. Harvard University Press, Cambridge, MA.

Clutton-Brock, T.H. (Ed.), 1988. Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems. University of Chicago Press, Chicago, IL.

Clutton-Brock, T.H., 1991. The Evolution of Parental Care. Princeton University Press, Princeton, NJ.

Dart, R.A., 1957. The osteodontokeratic culture of *Australopithecus prometheus*. Transvaal Museum (Pretoria) Memoir 10, 1–105.

DeVore, I., Washburn, S.L., 1963. Baboon ecology and human evolution. In: Howell, F.C., Bourlievre, F. (Eds.), African Ecology and Human Evolution. Aldine, Chicago, IL.

de Waal, F., 1983. Chimpanzee Politics: Power and Sex among Apes. Harper and Row, New York.

Doran, D.M., 1996. Comparative positional behavior of the African apes. In: McGrew, W.C., Marchant, L.F., Nishida, T. (Eds.), Great Ape Societies. Cambridge University Press, Cambridge, UK.

Ellis, L., 1989. Theories of Rape: Inquiries into the Causes of Sexual Aggression. Hemisphere, New York, NY.

Fleagle, J.G., 1999. Primate Adaptation and Evolution, second ed. Academic Press, San Diego, CA.

Fox, E.A., Sitompul, A.F., van Schaik, C.P., 1999. Intelligent tool use in wild Sumatran orangutans. In: Parker, S.T., Miles, H.L., Mitchell, R.W. (Eds.), The Mentalities of Gorillas and Orangutans. Cambridge University Press, Cambridge, UK.

Goodall, J., 1968. The behavior of free-living chimpanzees in the Gombe Stream Reserve. Animal Behaviour Monographs 1, 161–311.

Gowaty, P.A., 1992. Evolutionary biology and feminism. Human Nature 3, 217–249.

Hamburg, D.A., McCown, E.R. (Eds.), 1979. The Great Apes. Benjamin_Cummings, Menlo Park, CA.

Hawkes, K., O'Connell, J.F., Blurton Jones, N.G., 1997. Hadza women's time allocation, offspring provisioning, and the evolution of long post-menopausal life spans. Current Anthropology 38, 551–577.

Hrdy, S.B., 1977. The Langurs of Abu: Female and Male Strategies of Reproduction. Harvard University Press, Cambridge, MA.
Hrdy, S.B., 1999. Mother Nature. Pantheon, New York.
Huxley, T.H., 1863. Evidence as to Man's Place in Nature. Williams and Norgate, London.
Jolly, C.J., 1970. The seed eaters: a new model of hominid differentiation based on a baboon analogy. Man 5, 5–26.
Kinzey, W. (Ed.), 1987. The Evolution of Human Behavior: Primate Models. State University of New York Press, Albany, NY.
Krebs, J.R., Davies, N.B., 1991. Behavioural Ecology. Blackwell Scientific, Oxford, UK.
Leakey, M.D., 1987. Introduction (to the hominid footprints). In: Leakey, M.D., Harris, J.M. (Eds.), Laetoli: A Pliocene Site in Northern Tanzania. Clarendon, Oxford, UK.
MacKinnon, C.A., 1989. Toward a Feminist Theory of the State. Harvard University Press, Cambridge, MA.
O'Connell, J.F., Hawkes, K., Blurton Jones, N.G., 1999. Grandmothering and the evolution of *Homo erectus*. Journal of Human Evolution 36, 461–485.
Povinelli, D.J., Cant, J.G.H., 1995. Arboreal clambering and the evolution of self-conception. Quarterly Review of Biology 70, 393–421.
Rodman, P.S., 1999. Whither primatology? The place of primates in contemporary anthropology. Annual Review of Anthropology 28, 311–339.
Russon, A.E., Bard, K.A., Parker, S.T. (Eds.), 1996. Reaching into Thought: The Minds of Great Apes. Cambridge University Press, Cambridge, UK.
Sibley, C.G., Alquist, J.E., 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. Journal of Human Evolution 26, 99–121.
Smuts, B.B., 1985. Sex and Friendship in Baboons. Aldine, New York.
Smuts, B.B., 1992. Male aggression against women: an evolutionary perspective. Human Nature 3, 1–44.
Smuts, B.B., Smuts, R.W., 1993. Male aggression and sexual coercion of females in non-human primates and other mammals: evidence and theoretical implications. Advances in the Study of Behavior 22, 1–63.
Stern, J.T., Susman, R.L., 1983. The locomotor anatomy of *Australopithecus afarensis*. American Journal of Physical Anthropology 53, 69–80.
Sunday, S.R., Tobach, E. (Eds.), 1985. Violence Against Women: A Critique of the Sociobiology of Rape. Gordian Press, New York.
Teaford, M.F., 1994. Dental microwear and dental function. Evolutionary Anthropology 3, 17–30.
Trivers, R.L., 1971. The evolution of reciprocal altruism. The Quarterly Review of Biology 46 (1), 35–57.
Ungar, P.S., 1996. Relationship of incisor size to diet and anterior tooth use in sympatric Sumatran anthropoids. American Journal of Primatology 38, 145–156.
Ungar, P.S., Spencer, M.A., 1999. Incisor microwear, diet, and tooth use in three Amerindian populations. American Journal of Physical Anthropology 109, 387–396.
van Schaik, C.P., 1989. The ecology of social relationships amongst female primates. In: Standen, V., Foley, R.A. (Eds.), Comparative Socioecology: The Behavioural Ecology of Humans and Other Mammals. Blackwell Scientific, Oxford, UK.
van Schaik, C.P., Fox, E.A., Sitompul, A.F., 1996. Manufacture and use of tools in wild Sumatran orangutans. Naturwissenschaften 83, 186–188.
Visalberghi, E., Trinca, L., 1989. Tool use in capuchin monkeys: distinguishing between performing and understanding. Primates 30, 511–521.
Washburn, S.L., 1973. The promise of primatology. American Journal of Physical Anthropology 38, 177–182.
Whitten, P.L., Brockman, D.K., Stavisky, R.C., 1998. Recent advances in noninvasive techniques to monitor hormone-behavior interactions. Yearbook of Physical Anthropology 41, 1–24.
Wiley, R.H., Poston, J., 1996. Indirect mate choice, competition for mates, and coevolution of the sexes. Evolution 50, 1371–1381.
Wilson, E.O., 1975. Sociobiology. Harvard University Press, Cambridge, MA.
Winterhalder, B., Smith, E.A., 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. Evolutionary Anthropology 9, 51–72.
Wood, B., Harrison, T., 2011. The evolutionary context of the first hominins. Nature 470, 347–352.
Wrangham, R.W., 1980. An ecological model of female-bonded primate groups. Behaviour 15, 262–300.

Primate Groups and Their Correlates

Pascal Gagneux, University of California San Diego, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most primates are highly social and live in groups ranging from small reproductive units to large troops counting hundreds of individuals. Spatial and temporal distribution of food resources, competition with other species, and threat of predation are important determinants of primate group size. Many female primates and their dependent offspring forage largely independently of males, and most primate groups are female bonded. Furthermore, mating and breeding occur in a large variety of ways and may include individuals from outside existing groups. Like their closest living relatives, the chimpanzees and bonobos, humans live in multimale, multifemale groups. However, humans differ from all other primates by commonly forming strong pair bonds within such groups. Cultural norms for group affiliation and mating patterns can shape the genetic structure of human populations.

The Diversity of Primates, Their Social Organization, and Their Mating Systems

Primates include over 300 species of prosimians, monkeys, apes, and humans. These range in size from mouse-sized prosimians and pigmy marmosets weighing less than 100 g to orangutans and gorillas where males weigh over 200 kg. Most primates are tree dwelling, but some, such as gorillas and baboons, are mostly terrestrial. While some of the prosimians (a group of relatively 'primitive' primates including galagos, pottos, lorises, and lemurs) are nocturnal and mostly solitary, the vast majority of primates are diurnal and extremely social. In fact, sociality and sexuality are key primate adaptations. Biologists use the term 'mating system' to describe the patterns of behavior related to mate choice, reproduction, and dispersal. Such mating systems are intimately linked to the social system the animals live in, but they are not one and the same. Mating partners include, but need not be restricted to, social partners. A key characteristic of females in many primate species is their monthly ovulation. Breeding seasons do exist in some primate species, but overall, periods of female fertility are more frequent in primates than in most other large mammals. Accordingly, sexual behavior is frequent, and it has acquired functions in addition to reproduction.

Primate mating systems represent a wide spectrum ranging from polyandry, where a single female mates with several males (South American tamarins); to monogamous or serially monogamous pairs (owl monkeys and gibbons); to multimale, multifemale groups (chimpanzees, spider monkeys, red colobus monkeys, and many baboons); to single-male, multifemale groups, where the male has exclusive reproductive access to multiple females during his tenure (gorillas and mandrills). Single-male groups can experience temporal visits by groups of bachelor males (guenons); conversely, solitary males and females can temporarily gather when resources are plenty (orangutans).

Size differences between males and females (sexual dimorphism) seem to correlate with the degree to which males of a given species are polygynous (i.e., compete for reproductive access to females). Males and females of the monogamous gibbons or owl monkey (*Aotus*) are the same size, but males of the polygamous gorillas and mandrills can weigh more than twice as much as conspecific females. Male primates are often in intense competition for access to resources and fertile females. In some primates, including our closest living relatives, the chimpanzees, males collaborate in defending resources (they are strongly territorial) and females.

Primate mating systems are far more flexible and dynamic than previously thought. Changes in resource distribution or social changes can cause animals to alter the ways in which they find and choose mating partners. The same species may show different mating systems in different locations or at different times in the same location (Dunbar, 1988; Dixson, 2012).

Reproductive Strategies of Male and Female Primates

Sexual reproduction refers to the process by which male and female germ cells (sperm and egg) fuse and produce individuals with unique (and ephemeral) genetic combinations. The (haploid) germ cells carry half of the reshuffled chromosomes that the parents inherited from the grandparents. Female germ cells (ova) are large, costly to produce, and limited in number (one per month for three or four decades at most). In contrast, male germ cells (sperm) are tiny, much cheaper to produce, and are produced in huge numbers (hundreds of millions per day) throughout life by mature males. In animals with internal fertilization and gestation, this discrepancy in reproductive investment is exaggerated as females invest substantial resources into the developing fetus during gestation.

Female mammals bear an even more disproportionate cost of reproduction because of lactation, the physiological cost of which is higher than that of gestation itself. Prolonged parental care in primates adds to that burden as it is mostly provided by mothers. Male investment may end with successful insemination, after which males can look for further mates. The average number of offspring born to male and female of a given primate species is roughly the same, but the variance is much higher in males, where some highly successful males sire large numbers of offspring and many others none at all.

Biologists have repeatedly pointed at the imbalance between male and female reproductive costs to explain why males have a natural tendency to be promiscuous while females tend to be 'coy' and choosier (Hrdy and Williams,

1983). This caricature of male emphasis on quantity and female emphasis on quality is in part responsible for the fact that students of animal behavior have long overlooked the importance of female choice and multiple paternity (Lackman-Ancrenaz et al., 1992). Females, much rather than just being passive 'resources' for which males compete, are pursuing their own reproductive interests very actively, often within the constraints of male physical dominance and female–female competition. Their strategies are often less conspicuous and may involve long-term patterns, which explains why they have so often been overlooked.

Male primates can undoubtedly increase their reproductive success by maximizing the number of matings with fertile females. In multimale groups, this can occur by (mostly) physical dominance of one male over others. Subordinate males in such groups can achieve reproduction utilizing different approaches, such as 'consortships' in chimpanzees, where a male and female temporarily leave the social group.

There may be strong advantages to multiple matings for females as well. Primates have long periods of development, and infants are very dependent on parental assistance. Parental care, mostly by mothers but in many cases also involving allo-mothering by other group members including males, is crucial to primate reproduction. The tamarins, small primates of the Amazon forest, even have cooperative paternal care for the twins born to the female. The offspring of many primate species benefit from the protection and care provided by males with whom their mothers associate. Infant baboons and infant gibbons are often carried by males. In the monogamous owl monkeys, the infants are carried mostly by the male, and chimpanzee males are known to adopt orphaned male infants. Wherever male help is needed with the rearing of offspring, and there are several males present in the group, incertitude of paternity can cause males to assist offspring that they did not sire, resulting in a strong motivation for females to mate with many different males. On the other hand, too much incertitude about paternity is potentially dangerous, as males can be a threat to offspring that they perceive as not their own. Infanticide by males is well documented in several primate species, including gorillas and chimpanzees.

Genetic Consequences of Mating Patterns

The unique genetic identity of each individual can be directly characterized by genetic methods based on the identification of highly variable genetic loci (stretches of DNA located on particular chromosomes). Sampling of wild individuals can be based on small samples of tissue, blood, hair, or feces. These can be collected noninvasively or with minimum discomfort to the animals using small biopsy darts. There is much information from dozens of long-term research sites about the genetic composition of social groups, paternity within the group, and genetic differentiation between males and females and between different social groups and geographically distinct populations (Di Fiore, 2003).

A dozen or so highly variable loci can be amplified from samples using the polymerase chain reaction technique and will provide a unique genotype for each individual, defined by the combination of two genetic variants (alleles) found at each locus. The most commonly used type of loci are simple sequence repeats (also called simple tandem repeats or microsatellites), short stretches of repeated sequence motifs that have a very high mutation rate and thus exist in a number of states (alleles with a distinct number of repeats) within each population. In addition, short sequence reads from the clonally inherited mitochondrial DNA and Y chromosomes can help determine the maternal and paternal ancestry of each individual. These pieces of DNA are inherited without recombination and only through one parent, and can thus be used to measure male- or female-mediated gene flow.

Studies of genotypes in a large number of species of wild primate populations have yielded genetic evidence for several aspects of reproductive and social behavior. Female-bonded baboons show higher relatedness among females (these share more genetic variants when compared to the general population) but not males, which are known to disperse from their natal groups after puberty (Altmann et al., 1996). Studies of gorilla groups have confirmed that a single resident male, and sometimes also his brother, sire most of the offspring with the group's females (Nsubuga et al., 2008). Male orangutans have been found to employ their two different mating strategies with success: both large, flanged males chosen by females as well as small, nonflanged males that coercively mate with females sire offspring (Goossens et al., 2006). In many species with multiple males and females, higher male rank correlates with higher reproductive success. This is true for chimpanzees as well, but alternative mating strategies by lower ranking males combined with female preferences contribute to substantial reproductive success for lower ranking males as well (Wroblewski et al., 2009).

A surprising finding is that males from the same chimpanzee communities are not more closely related than are females, despite the fact that chimpanzees are mostly patrilocal (females disperse around sexual maturity and males remain in their native groups). A possible explanation for this finding is that the rate of extragroup paternity might be elevated enough (6–10%) to introduce sufficient male gene flow between social groups (Schubert et al., 2011). This indicates that social units and reproductive units need not be completely identical.

Sex as a Social Tool

In bonobos, apes known for their frequent sexual interactions, sex is directly used to reduce tension in the group and is often exchanged for food. Chimpanzee females exchange sex for meat that is usually controlled by males, which do most of the hunting and use meat as a social currency mostly among themselves. A female chimpanzee will copulate more than 6000 times with a very large number of males (10–30) while she can maximally produce only six surviving offspring. Similarly, the high frequency of same-sex sexual contact in bonobos plays important social roles. In fact, bonobos appear to be the one species of primates where frequent female–female sex facilitates the existence of strong female alliances, alliances strong enough to keep male aggression and dominance in check.

Advertised Ovulation in Multimale Groups

Females in many species with multimale groups exhibit conspicuous pink swellings of their genital area during several days around the time of their ovulation. These swellings are known as estrous swellings, and they represent a formidable 'tool' of female primates. Different species of primates have independently evolved these signals, including gelada baboons, where the swelling is on the female's chest rather than around her rear end.

A female with a swelling becomes the focus of attention of all males in the group. Her social status increases, and males will compete for access to her. Several females can have estrous swellings at the same time, thus making it virtually impossible for any given male to monopolize access to all of them. Estrous swellings appear to play an important role for species in which females disperse and join new social groups after they become sexually mature. Swellings are a kind of 'social passport' that greatly facilitates a female's acceptance by unfamiliar males.

There is evidence for certain amounts of disconnect between the advertising of fertility and the actual time of ovulation. Such a disconnect could work to the advantage of the female because she could escape the controlling efforts of certain males and exert a freer choice of paternity for her offspring. The topic of 'concealed ovulation' takes on considerable importance as many scientists have suggested that humans have attained a stage of perfectly concealed ovulation (concealed even from a large proportion of women themselves). To what degree the inability of human males to detect the moment of peak fertility in women has played a role in leading to the pair bonding in our species is a much-debated issue.

Advertised Male Quality

Female criteria for choosing their mates remain poorly understood. In gelada baboons, male chest coloration is directly proportional to social status (Bergman et al., 2009). Females can certainly learn about future social partners through their year-long interactions with males, but for detecting 'good genes' they must rely on indirect signs such as body symmetry and relative absence of disease and parasites.

Determinants of Sociality in Primates

The distribution of resources and the presence of predators play key roles for the distribution of individuals across space. When resources are clumped, groups of animals can live around a particular cluster of resources and defend a certain territory. Such groups can form cohesive social units. When resources are few and very dispersed, no single cluster of resources can be defended even by large groups of animals. Such environments tend to harbor animals living in much smaller groups or as dispersed individuals. Even highly social populations of chimpanzees are known to scatter into much smaller subgroups during periods of low food abundance. The solitary orangutans, on the other hand, can gather in large numbers during periods of plenty (mast fruiting).

There are many good reasons for living socially. More eyes can see better, and more ears can hear better. Living socially facilitates the search for food and the detection of predators. Another advantage is the phenomenon of the 'selfish herd,' where the presence of other individuals spreads the risk of predation as any of the other group members is as likely to be preyed upon as oneself. Also, cooperation in antipredator defense is possible only in groups. Obviously, social life facilitates finding mates, even if it will increase the competition for them. Animals in groups can cooperate in defending resources against competitors. Such cooperative resource defense, when done by males, may have the added benefit of attracting females to the group, which may represent unwelcome competition to the females already resident in the group. Finally, the presence of individuals possessing different capabilities (such as hunting, raiding beehives, and remembering the time and place certain foods can be harvested) in a group will benefit all group members.

Unique individual experience and history shape large-brained long-lived organisms such as primates. The mating strategies within a social group can vary dramatically between individuals. Some males in multimale groups, for example, may not show any ambition with respect to the dominance hierarchy. They spend their whole life as subordinate males. Their friendship with certain females in the group may in the end translate into numerous offspring. Certain females may be much more successful at forming social bonds with dominant males. Sheer force can allow a male to ascend in the dominance hierarchy, but the majority of individuals in his group may not follow him. The cognitive capacities of primates appear to have coevolved with their tendency for sociality. Long-lasting relationships with many group members require good memory and benefit greatly from social intelligence.

Costs of Living Socially

There are multiple costs to sociality. Competition between males can create social instability and conflict. Such conflict is often violent and can lead to harm to bystanders, or even infanticide by males that hope to remove the offspring of competitors and gain rapid access to fertile females (the loss of an infant causes lactating females to resume ovulatory cycles).

Competition between females also occurs, for access to males, access that may lead to support (shared resources) and male protection during social conflicts. Competition between females may be the factor that precludes the formation of strong female coalitions against males in many species.

Infanticide perpetuated by females has been observed in some wild primate groups and may represent the darkest side of female–female competition for resources. The study of female dominance hierarchies in natural primate populations is still in its infancy. Such hierarchies tend to be more subtle; nevertheless, female social rank can play an important role for reproductive success and change the way females invest in sons or daughters.

A definite cost of living socially is that it can include high levels of social stress. Such stress derives from harassment by other group members (Sapolsky, 2005). Other costs of living as a social group include higher risks of infection by pathogens

and parasites and elevated risk of being detected by predators due to the noise and vocalizations associated with social interactions.

Most Primate Groups Are Female Bonded

In most primates, females and their female relatives remain in their native groups while most males disperse to other groups when they reach sexual maturity. Some notable exceptions to female residence are found in colobus and spider monkeys and in our closest living relatives, chimpanzees and bonobos, where females disperse upon reaching sexual maturity and males remain in their natal communities.

The driving force behind dispersal is thought to be inbreeding avoidance and/or competition with animals of the same sex (Moore and Rauf, 1984). Sexual mimicry in white-cheeked gibbons where adolescent females look like males and in red colobus where young males have pseudoswellings indicates that same-sex competition is important in those species.

The mechanisms underlying inbreeding avoidance are poorly understood. Whether primates are able to recognize their paternal kin by features inherited from the father, or whether they simply avoid mating with individuals from the same age cohort, is still unclear (Alberts, 1999). In humans, close proximity at an early age seems to be the crucial mechanism that triggers avoidance as sexual partners.

Do Males Follow Female Spacing?

The way females deploy themselves across time and space tends to determine how males can space themselves around females. Only if the ecosystem is capable of feeding a large number of females with enough tolerance for each other can males attempt to control such groups. A single male can control a harem, or a group of males can cooperate in female and territory defense and share the reproductive access to females. All multifemale groups create the potential for female–female competition over access to resources, and access to males and the support these may provide.

Competition between potential mates may create the opportunity to enhance the visibility of positive traits or weaknesses. Because competition involves stress and stress is known to impact an animal's immune system, competition can exaggerate the visible differences in male quality.

If female choice incites increased male competition, then female choice may indirectly lead to an increased size difference between males and females. Because strong and assertive males can be more successful at guarding resources that they share with females, females may favor the evolution of such males. Females need male protection from the unwanted attention of other males, in the extreme case even forced copulation, as well as against the threat of male infanticide; both are part of male mating strategy in many primate species (Smuts, 1993). Strong and protective males, by the same token, will be more successful at restricting the independence of females. Permanent male vigilance and mate guarding of a single female are likely to strongly reduce a male's capacity to pursue other females, and thus represent one of the ways pair bonding could come about.

The Benefit of Multiple Paternity to Females

Mating with multiple males can be advantageous for a female in various ways. As mentioned in this article, confusing the paternity of her offspring can effectively protect them from male infanticide. The female may gain support from her mating partners, both social support during conflicts with other group members and material support in the form of shared resources.

More recently, the possibility of a genetic benefit has been proposed, and some evidence is starting to accumulate (Zeh and Zeh, 1996, 1997). Because the primate fetus develops inside the mother, the female and the fetus need to achieve a type of 'immunological détente.' That is, the fetus must not be rejected by the mother's powerful immune system, but at the same time, the demands of the fetus must be kept in check, not allowing it to exact too high a toll on maternal resources.

Multiple paternity will produce more diverse offspring. Such diversity may have much to do with defense against disease. The vertebrate immune system is an immensely sophisticated and very costly system that allows long-lived organisms such as primates to survive in a world in which very short-lived, rapidly evolving parasites and pathogens (ranging from worms and protozoa to bacteria and viruses) are a constant threat. The immune system relies on genetic variation that is present in all populations as well as such generated anew in each individual in order to produce a flexible repertoire of molecular probes (antibodies and T-cell receptors), which detect telltale molecules of specific intruding pathogens. It seems critical that each fetus inherits genes coding for two rather different immunological 'toolkits' in order for the offspring to be well equipped for its countless future encounters with pathogens. Conceiving offspring with different fathers would allow females to produce sets of offspring with more variation in their immune repertoires, strongly reducing the risk of losing all or most offspring to the same disease. The same obviously applies to males as well.

Furthermore, paternal and maternal genes have to work in harmony in order to produce healthy offspring. Some combinations may simply fail to work, and yet others may work but put the offspring at a disadvantage from the start. The detection of embryos with strongly limited immune repertoires (i.e., two almost identical sets of multihistocompatibility complex (MHC) alleles, also called HLA in humans) inherited from both parents can cause spontaneous abortions in many mammal species, including our own. Because of the importance of compatibility between genomes, the most promising genes for each female are found in different males. Finally, multiple paternity may allow females to reduce inbreeding where females cannot recognize close kin.

The Cost of Multiple Paternity for Females

The immune challenges to the mother are increased with each additional pregnancy by another father. Each new father

confronts the female's immune system with a new set of foreign molecules. Data on pregnancy-associated hypertension (preeclampsia, a disease unique to humans) indicate that the risk for this immune-mediated, potentially fatal imbalance in maternal immune tolerance of the fetus is higher if the pregnancy is from another male than the previous one. A further cost to multiple paternity is sexually transmitted infections. The risk of infection is proportional to the number of different mating partners. The cost of infection ranges from a short disease episode to abortion, congenital malformation of the offspring, and permanent sterility.

Molecules and Behavior

Ova and sperm engage in intense molecular dialogue, a process that is important for creating reproductive barriers between species but that is also suspected to involve an additional level of (postmating) female choice. Female primates are known to produce a large quantity of antibodies against sperm in their genital tracts, and they also unleash huge numbers of immune cells after mating. These processes strongly favor diversity in paternity, as the antibodies would be formed against the sperm of the 'usual' mating partner, leaving sperm from unknown males at an advantage.

The role of chemoreception (pheromones) in regulating some primate behavior remains very poorly understood. The fact that primates may be able to smell important immune molecules (MHC) points toward the potentially important roles of such molecules and their detection in mate choice and sexual behavior (Wedekind and Füri, 1997). Recent genetic studies support the notion that MHC variants underlie mate choice in some human populations (Chaix et al., 2008).

Postmating Female Choice and Sperm Competition

Certain sperm may be more likely to be taken up by the egg as they approach it. Such cryptic female choice would represent a choice of individual sperm, possibly based on the genes carried by that sperm and expressed on its surface. Certain genes of crucial importance in early embryonic development are expressed on the surface of sperm.

Males can compete for fertilization of females via a process called 'sperm competition.' When females mate with many males within a short period of time, the sperm of different males will be available for fertilizing her egg. This exerts pressure on males to produce larger quantities of sperm in order to outcompete other males. In chimpanzees and bonobos, where multiple mating with many males gives rise to strong sperm competition, males have conspicuously large testicles and produce large amounts of sperm (several billion per ejaculation versus ~200 million in humans). They even produce 'copulatory plugs,' a gelatinous mass formed by seminal fluid that blocks the cervix. In gorillas, in contrast, where silverback males face very little competition in their groups, males have very small testicles and inseminate females with rather modest numbers of sperm. Gorillas and humans alike do not form copulatory plugs.

Genomic Conflicts

Because different males can sire subsequent offspring, a conflict of interest can arise between the paternal and the maternal genes. A female will spread her total reproductive efforts across all her subsequent offspring. Males are likely to have sired offspring with other females and thus are spreading their reproductive effort across several females. The larger the contrast between the number of mating partners between males and females, the more acute this conflict becomes. Paternal genes in the embryo are likely to cause each embryo to be overdemanding from the point of view of the mother. Such paternal genes will increase the male's reproductive success at a detriment to each female, who will have invested too much in that particular offspring, as opposed to evening out her investment across all her pregnancies.

It is this conflict that seems to be at the root of a phenomenon called genomic imprinting. Genomic imprinting occurs when, between the two variants of each gene present in the embryo (the paternal and maternal alleles), only one of the two is expressed. Often, only the allele inherited from the mother is expressed, while the paternal allele appears silenced. Imprinting may have several functions, but one of them may be female control of overdemanding genes from the father (Mochizuki et al., 1996; Spencer et al., 1998; Crespi, 2010).

Factors Determining Primate Mating Systems

Primate mating systems represent the outcome of a delicate balance between various constantly changing forces. These include (1) ecological factors such as the distribution of resources and pressure by predators; (2) female foraging and mating strategies, including the need to find multiple fathers of high genetic quality and good genetic compatibility, and one or more supportive social partners; (3) male mating strategies that may entail access to the largest number of fertile females, often in severe competition with other males; and (4) the need for parental care of infants and juveniles.

Compromises between male and female strategies seem to occur at every level, from gene expression to behavior. The existence of lifelong friendships between adult primates of both sexes introduces an additional dimension, that of emotional attachment and trust, between individuals.

Reflections on Human Groups and Families

The origin of the human family has long fascinated biologists and philosophers and continues to do so. Humans live in the largest groups of all primates yet primarily live in pair bonds.

Humans are unusual in forming pair bonds within larger groups of multiple males and females. Anthropological data suggest that less than a quarter of traditional human societies were traditionally monogamous, with the vast majority having mild or widespread polygyny (Murdock and White, 1969; White et al., 1988). Most foraging societies have some polygyny, but the vast majority of men are monogamous across these societies (Marlowe, 2003).

With the development of agriculture and the possibility of amassing surplus resources, polygyny became more feasible for rich and powerful males. More food also meant more humans and larger social groups. Large social groups with either monogamous pairs or polygynous units with a single male provide settings in which females can mate with males other than their predominant social partner or husband. Studies of extrapair paternity in humans are difficult to carry out and have yielded very different estimates. Some societies, such as the traditional Kalahari Desert foragers the !Kung, have been shown to have a extrapair paternity rate of 2%. For others, such as the natural fertility populations such as the Himba of Namibia, the extrapair paternity rate is 17% (Scelza, 2011). Modern Western societies can have relatively high levels of spousal infidelity, but many of these societies show very moderate rates of extrapair paternity of around 2% (Anderson, 2006).

Pair bonding appears to be an important characteristic of most human societies, and it has been argued that it was a key phenomenon in the evolution of our species (Chapais, 2008). Only prolonged pair bonding can lead to true biparentality and father–offspring recognition, which in turn would make possible the establishment of kinship networks that extend across social groups.

How old is this phenomenon? Some paleontologists have suggested that the relative lack of sexual dimorphism in possibly ancestral hominids such as the 4.5-million-year-old *Ardipithecus ramidus* strongly indicates monogamous behavior and pair bonding (Lovejoy, 2009). Others put the origin of pair bonding in the lineage leading to modern humans only at a much later time point, with the origin of the genus *Homo* and possibly the use of fire and development of cooking about 2 million years ago (Wrangham et al., 1999).

In humans, pair bonding can result from mutual infatuation and love, but in many societies, spouses are chosen by the parents and arranged marriages are important tools for forming alliances and managing wealth. Humans are the only species that have marriage ceremonies. Not all societies share these, but most have rituals that celebrate the union of a new husband and wife. These rituals often involve the participation of both families, local elders, and/or priests, and included gift giving for the newlyweds. In addition to biological incest avoidance, most human societies have strict rules against incest. The degree of relatedness considered to form incest varies strongly between societies. Many traditional societies actively encourage marriage between cousins or between uncles and nieces. Such cultural practices leave recognizable patterns in the genomes of humans in those societies, such as much larger segments of the genome without recombination than in more outbred populations (homozygosity by descent; Leutenegger et al., 2011).

Despite the fact that most traditional human societies tolerated or even encouraged polygamy, most Western societies have outlawed polygamy. In the past, exceptionally powerful male individuals such as Genghis Khan were able to produce offspring in thousands, effectively leaving genetic imprints on entire populations that can now be detected by genetic methods (such as strong biases in Y-chromosome frequencies; Zerjal et al. 2003). It has recently been suggested that the rapid spread of the monogamous norm across modern societies, most of which banned polygamy in the twentieth century, might have been facilitated by cultural evolution involving competition between monogamous and polygynous societies (Henrich et al., 2012).

More recently, human families have taken on yet other forms. High divorce and remarriage rates in Western societies bring with them the existence of many families with stepparents. There is some evidence that growing up with a stepparent can be dangerous for children due to the much higher risks of abuse perpetrated mostly by stepfathers (Daly and Wilson, 1985).

In some Western societies, same-sex couples can form legally recognized unions. These couples can adopt children or have children with the help of assisted reproduction. There is now a generation of children growing up with two fathers or two mothers.

Among heterosexual families in the industrialized world, more and more couples resort to assisted reproductive technologies due to problems with fertility. In the United States, 1–2% of all babies born are conceived via assisted reproductive technologies. These technologies can include donor sperm from a man other than the husband, donor eggs and surrogate motherhood, or even a third genetic parent in the case of nuclear transfer of an egg nucleus into the egg of a younger woman (where the mitochondria are from a different woman than the chromosomes).

Some sperm donors have achieved reproductive success of well over 100 offspring. There are attempts at limiting such numbers, and several countries have set limits to the number of offspring per donor to ranges of 2–25.

Surrogate mothers often are not the donors of the eggs, meaning that this modern method of reproduction includes a woman gestating a child with whom she is not related at all, often for financial gain. Finally, the delay in reproduction observed in many industrial societies' amounts to a large-scale genetic experiment, as it is becoming clear that the mutational load increases with parental age. Whole genome sequencing of family trios reveals that fathers, who contribute most mutations to the offspring, pass on one additional mutation with every year that they delay fatherhood (Kong et al., 2012).

Future Directions

The genetics of mating behavior are still very limited. The future will surely bring more insight into the biological basis for species-specific propensities toward one mating system or the other. Exciting findings in the regulatory region of the mammalian vasopressin receptor have been interpreted as directly causing monogamous behavior in mammals (Hammock and Young, 2005). Phylogenetic and genomic comparison of this region across a large number of rodents, however, indicates that mammalian monogamy is not controlled by a single gene (Fink et al., 2006; see also Young et al., 1999). The rapid spread of the human monogamous norm is a clear indication that social and cultural forces can rapidly mold the predominant mating system in our species. Not only is understanding primate mating systems essential if one wants to understand human evolution, but also, as many primates are currently facing extinction, such an

understanding is also crucial for conservation efforts of nonhuman primate species, in the wild and in captivity.

See also: Adaptation, Fitness, and Evolution; Comparative Method in Evolutionary Studies; Domestic Violence: Legal Perspectives; Empirical Legal Studies; Gene–Culture Coevolution; Genetics, Disease, and Reproduction; Human Behavioral Ecology; Human Cooperation, Evolution of; Human Mate Choice, Evolution of; Human Mate Choice; Incest Prohibition, Origin and Evolution of; Kin Selection; Kinship, Evolution of; Marriage Systems, Evolution of; Primate Behavior: Significance for Understanding Humans; Primates, Evolution of; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Sexual Orientation: Genetic Aspects.

Bibliography

Alberts, S.C., 1999. Paternal kin discrimination in wild baboons. Proceedings of the Royal Society of London B: Biological Sciences 266, 1501–1506.

Altmann, J., Alberts, S.C., Haines, S.A., Dubach, J., Muruthi, P., Coote, T., Geffen, E., Cheesman, D.J., Mututua, R.S., Saiyalel, S.N., Wayne, R.K., Lacy, R.C., Bruford, M.W., 1996. Behavior predicts genes structure in a wild primate group. Proceedings of the National Academy of Sciences of the United States of America 93 (12), 5797–5801.

Anderson, K.G., 2006. How well does paternity confidence match actual paternity? Current Anthropology 47 (3), 513–520.

Bergman, T.J., Ho, L., Beehner, J.C., 2009. Chest color and social status in male geladas (*Theropithecus gelada*). International Journal of Primatology 30, 791–806.

Chaix, R., Cao, C., Donnelly, P., 2008. Is mate choice in humans MHC-dependent? PLoS Genetics 4 (9), e1000184.

Chapais, B., 2008. Primeval Kinship: How Pair-Bonding Gave Birth to Human Society. Harvard University Press, Cambridge.

Crespi, B.J., 2010. The origins and evolution of genetic disease risk in modern humans. Annals of the New York Academy of Sciences 1206, 80–109.

Daly, M., Wilson, M.I., 1985. Child abuse and other risks of not living with both parents. Ethology and Sociobiology 6, 197–210.

Di Fiore, A., 2003. Molecular genetic approaches to the study of primate behavior, social organization, and reproduction. American Journal of Physical Anthropology Suppl. 37, 62–99.

Dixson, A.F., 2012. Primate Sexuality. Oxford University Press, Oxford, UK.

Dunbar, R.I.M., 1988. Primate Social Systems. Croom Helm, London.

Fink, S., Excoffier, L., Heckel, G., 2006. Mammalian monogamy is not controlled by a single gene. Proceedings of the National Academy of Sciences of the United States of America 103 (29), 10956–10960.

Goossens, B., Setchell, J.M., James, S.S., Funk, S.M., Chikhi, L., Abulani, A., Ancrenaz, M., 2006. Philopatry and reproductive success in Bornean orang-utans (*Pongo pygmaeus*). Molecular Ecology 15 (9), 2577–2588.

Hammock, E.A., Young, L.J., 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308 (5728), 1630–1634.

Henrich, J., Boyd, R., Richerson, P.J., 2012. The puzzle of monogamous marriage. Philosophical Transactions of the Royal Society of London B: Biological Sciences 367 (1589), 657–669.

Hrdy, S.B., Williams, G.C., 1983. Behavioral biology and the double standard. In: Wasser, S.K. (Ed.), Social Behavior of Female Vertebrates. Academic Press, New York, pp. 3–17.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W.S., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U., Stefansson, K., 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488 (7412), 471–475.

Lackman-Ancrenaz, I., Bruford, M.W., Gowaty, P.A., 1992. Evolutionary biology and feminism. Human Nature 3, 217–249.

Leutenegger, A.L., Sahbatou, M., Gazal, S., Cann, H., Génin, E., May 2011. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? European Journal of Human Genetics 19 (5), 583–587.

Lovejoy, C., 2009. Reexamining human origins in light of *Ardipithecus ramidus*. Science 326 (5949), 74.

Marlowe, F.W., 2003. The mating system of foragers in the standard cross-cultural sample. Cross-Cultural Research 37, 282–306.

Mochizuki, A., Takeda, Y., Iwasa, Y., 1996. The evolution of genomic imprinting. Genetics 144, 1283–1295.

Moore, J.J., Rauf, A., 1984. Are dispersal and inbreeding avoidance related? Animal Behaviour 32, 94–112.

Murdock, G.P., White, D., 1969. Standard cross-cultural sample. Ethnology 8, 329–369.

Nsubuga, A.M., Robbins, M.M., Boesch, C., Vigilant, L., March 2008. Patterns of paternity and group fission in wild multimale mountain gorilla groups. American Journal of Physical Anthropology 135 (3), 263–274.

Sapolsky, R.M., 2005. The influence of social hierarchy on primate health. Science 308 (5722), 648–652.

Scelza, B.A., 23 December 2011. Female choice and extra-pair paternity in a traditional human population. Biology Letters 7 (6), 889–891.

Schubert, G., Stoneking, C.J., Arandjelovic, M., Boesch, C., Eckhardt, N., Hohmann, G., Langergraber, K., Lukas, D., Vigilant, L., 2011. Male-mediated gene flow in patrilocal primates. PLoS One 6 (7), e21514.

Smuts, B.B., 1993. Male aggression and sexual coercion of females in non-human primates and other mammals: evidence and theoretical implications. In: Slater, P.J.B., Rosenblatt, J.S., Milinski, M., Snowdon, C.T. (Eds.), Advances in the Study of Behavior. Academic Press, New York, pp. 1–63.

Spencer, H.G., Feldman, M.W., Clark, A.G., 1998. Genetic conflicts, multiple paternity and the evolution of genomic imprinting. Genetics 148, 893–904.

Wedekind, C., Füri, S., 1997. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? Proceedings of the Royal Society of London B: Biological Sciences 264, 1471–1479.

White, D.R., Betzig, L., Borgerhoff, M., Mulder, Chick, G., Hartung, J., Irons, W., Low, B.S., Otterbein, K.F., Rosenblatt, P.C., Spencer, P., 1988. Rethinking polygyny: co-wives, codes, and cultural systems. Current Anthropology 29 (4), 529–572.

Wrangham, R.W., Jones, J.H., Laden, G., Pilbeam, D., Conklin-Brittain, N., 1999. The raw and the stolen: cooking and the ecology of human origins. Current Anthropology 40 (5), 567–594.

Wroblewski, E.E., Murray, C.M., Keele, B.F., Schumacher-Stankey, J.C., Hahn, B.H., Pusey, A.E., 2009. Male dominance rank and reproductive success in chimpanzees, *Pan troglodytes schweinfurthii*. Animal Behaviour 77 (4), 873–885.

Young, L.J., Nilsen, R., Waymire, K.G., MacGregor, G.R., Insel, T.R., 1999. Increased affiliative response to vasopressin in mice expressing the V1a receptor from a monogamous vole. Nature 400, 766–768.

Zeh, J.A., Zeh, D.W., 1996. The evolution of polyandry I: intragenomic conflict and genetic incompatibilty. Proceedings of the Royal Society of London B: Biological Sciences 263, 1711–1717.

Zeh, J.A., Zeh, D.W., 1997. The evolution of polyandry II: post-copulatory defences against genetic incompatibilty. Proceedings of the Royal Society of London B: Biological Sciences 264, 69–75.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M.E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S.Q., Tyler-Smith, C., 2003. The genetic legacy of the Mongols. American Journal of Human Genetics 72 (3), 717–721.

Primates, Evolution of

Robert D Martin, The Field Museum, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Primates, an order of mammals containing some 400 extant and 500 fossil species, document a striking adaptive radiation. Extant species fall into five geographically restricted 'natural groups,' each with a distinct ancestor: lemurs, lorises, and bushbabies, tarsiers, New World monkeys, and Old World monkeys and apes. Tree-shrews, once included in Primates, are now allocated to the order Scandentia. With tree-shrews excluded, primates share many features distinguishing them from other mammals and reflecting adaptations in their common ancestor. Those features generally reflect adaptation for arboreal activity in tropical or subtropical forests. Current evidence indicate that primates may have diverged from other mammals in the Late Cretaceous.

Primates, one of 18 extant orders of placental mammals, are of special interest because they include the human species (*Homo sapiens*). Tree-shrews, generally included in Primates for some 60 years, now are widely thought to be only distantly related to primates and are accordingly allocated to their own order Scandentia. Even after excluding tree-shrews, almost 400 living primate species can now be recognized (updated from Groves, 2005), partly due to discovery of entirely new species but mainly because new information, especially evidence from chromosomal and DNA analysis, has led to subdivision of existing species. This is a dramatic increase over the 150 species listed in a leading handbook in the 1960s (Napier and Napier, 1967). The order Primates also includes more than 500 fossil species (also continually increasing as more discoveries are made), documenting the evolution of this order of mammals.

The evolutionary history of primates is a striking example of adaptive radiation, in which descendants from an original common ancestor with a particular set of features diverged and adapted to occupy different geographic areas and ecological niches. Following the advent of Darwinian evolutionary theory, this adaptive radiation was at first reconstructed almost entirely on the basis of morphological (anatomical) features. But over the years, the comparative framework gradually has expanded to include physiological, behavioral, ecological, chromosomal, and molecular evidence.

Living Primates

The logical starting point for reconstruction of primate evolution is the array of extant species. This is partly because recognition of the group began with the study of living representatives and partly because many important features (notably chromosomal and molecular characters) are not fossilized. Known fossil relatives of living primates can be included as a second step in the overall framework, and they bring with them the unique benefit of an accompanying geological time scale. Later integration of fossil species in a preliminary phylogenetic tree based on extant forms also permits use of paleontological evidence to test its validity.

Extant nonhuman primates can be divided into five 'natural groups,' each with a restricted geographic range (**Figure 1**). As geographic isolation plays a fundamental part in evolutionary divergence, these groups are inherently likely to reflect basic subdivisions within the primate tree. Indeed, they are immediately recognizable in all widely accepted phylogenetic reconstructions and classifications. The first natural group contains the Madagascar lemurs (infraorder Lemuriformes). These are the only primates present on that island, and they bear witness to a substantial local radiation (Tattersall, 1982). More than 60 extant species can now be recognized and the modern array also includes at least 17 subfossil species that died out over the past few millennia, following the first human colonization of Madagascar. The second group contains the lorises and the bushbabies (infraorder Lorisiformes). They are distributed widely in Africa, South Asia, and Southeast Asia, but the total number of species (about 28) is markedly smaller than for lemurs and represents a more limited radiation. An even smaller radiation is evident with the third group, containing about a dozen closely related tarsier species inhabiting the Southeast Asian archipelago (infraorder Tarsiiformes). The fourth group contains the monkeys of South and Central America (infraorder Platyrrhini), which are the only primates now present in the New World. This is a relatively large group of some 128 species, divided into claw-bearing marmosets and tamarins (family Callitrichidae) and true monkeys (family Cebidae). As with the lemurs of Madagascar, this once again documents a major local radiation (Kinzey, 1997). The fifth group, containing the Old World monkeys and apes of Africa, South Asia, and Southeast Asia (infraorder Catarrhini) is the largest, with approximately 132 species. In fact, humans also belong to this group, but expanded far beyond this original geological range after the genus *Homo* emerged. Within this fifth group, which is most directly relevant to human evolution, there is a clear division between Old World monkeys (superfamily Cercopithecoidea) and apes and humans (superfamily Hominoidea), although the two subgroups overlap extensively in geographic distribution.

The first three natural groups of primates (lemurs, lorises, bushbabies, and tarsiers) have remained relatively primitive and have been labeled collectively prosimians (literally meaning 'before the monkeys'). They are comparatively small (modal body weight approximately 500 g) and the majority of species are nocturnal. Diurnal species (certain relatively large-bodied members of the families Lemuridae and Indridae) are

 http://dx.doi.org/10.1016/B978-0-08-097086-8.81030-2

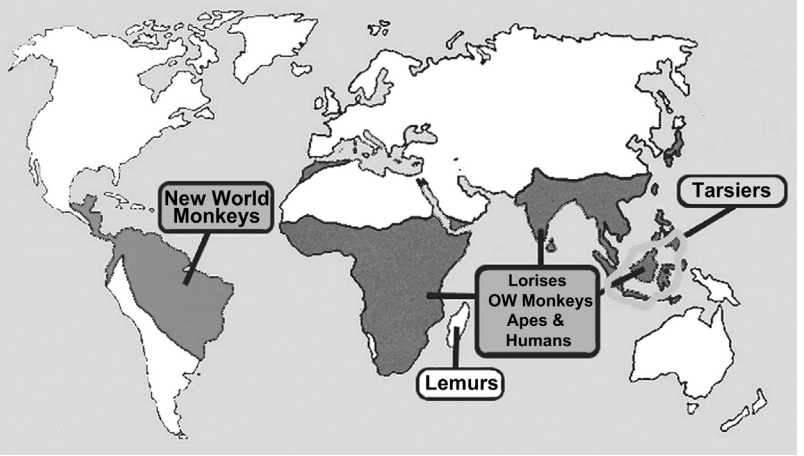

Figure 1 Geographic distribution of living nonhuman primates (excluding tree-shrews). Shaded areas indicate the approximate ranges of the five natural groups, with the exception of the loris group (which occupies most but not all of the range occupied by Old World monkeys and apes).

few in number and occur only among the lemurs. In fact, some lemurids (species of *Eulemur* and *Hapalemur*) show a highly unusual mixture of nocturnal and diurnal activity that has been termed 'cathemeral activity' (Tattersall, 1988). This is just one example of the great diversity of unusual adaptations found among the lemurs of Madagascar.

Members of the fourth and fifth natural groups of primates (the higher primates: monkeys, apes, and humans) contrast with the prosimians in sharing a distinctive set of advanced features in their teeth, jaws, brain, and reproductive system, and have been collectively labeled simians. They are comparatively large in size (modal body weight approximately 5 kg), and almost all species are diurnal. Indeed, the only nocturnal simians are the New World owl monkeys (eight species of *Aotus*). The 10-fold difference in modal body size between simians and prosimians – 5 kg versus 500 g – is undoubtedly connected with the general distinction between nocturnal and diurnal habits. This is just one example of the pervasive importance of body size in primate biology. So the availability of a compilation of body weights for almost all primate species, very largely derived from wild animals, merits special mention (Smith and Jungers, 1997). Overall, living primates cover a 300-fold range of body sizes, extending from 40 g for the smallest of the lesser mouse lemurs (a species of *Microcebus*) to 120 kg for an adult male gorilla. Differences in body size account for many of the differences seen between primate species.

Definition of Living Primates

It is often stated that living primates represent a graded series, ranging from relatively primitive to very advanced, and that therefore it is difficult to demarcate them from other orders of mammals (notably Insectivora, which often have been seen as the most primitive placental mammals alive in the twenty-first century). This view, however, was influenced heavily by the inclusion of tree-shrews in the order Primates. Detailed reexamination of primates to the exclusion of tree-shrews revealed that all extant species in fact share a substantial set of characteristics. That shared set of features distinguishes primates fairly clearly from other placental mammals, as is indicated by the following definition (Martin, 1990, pp. 639–640):

> Primates are typically arboreal inhabitants of tropical and subtropical forest ecosystems. Their extremities are essentially adapted for prehension, rather than grappling of arboreal supports. A widely divergent hallux (big toe) provides the basis for a powerful grasping action of the foot in all genera except *Homo*, while the hand usually exhibits at least some prehensile capacity. The digits typically bear flat nails rather than bilaterally compressed claws; the hallux always bears a nail. The ventral surfaces of the extremities bear tactile pads with cutaneous ridges (dermatoglyphs) that reduce slippage on arboreal supports and provide for enhanced tactile sensitivity in association with dermal Meissner's corpuscles. Locomotion is hindlimb-dominated, with the center of gravity of the body located closer to the hindlimbs, such that the typical walking gait follows a diagonal sequence (forefoot precedes hindfoot on each side). The foot is typically adapted for tarsi-fulcrimation, with at least some degree of relative elongation of the distal segment of the calcaneus, commonly resulting in reverse alternation of the tarsus (calcaneo-navicular articulation). The visual sense is greatly emphasized. The eyes are relatively large and the orbits possess (at least) a postorbital bar. Forward rotation of the eyes ensures a large degree of binocular overlap. Ipsilateral and contralateral retinofugal fibres are approximately balanced in numbers on each side of the brain and organised in such a way that the contralateral half of the visual field is represented. Enlargement and medial approximation of the orbits is typically associated with ethmoid exposure in the medial orbital wall (though there are several exceptions among the lemurs). The ventral floor of the well-developed auditory bulla is formed predominantly by the petrosal. The olfactory system is unspecialized in most nocturnal forms and reduced in diurnal forms. Partly because of the increased emphasis on vision, the brain is typically moderately enlarged, relative to body size, in comparison to other living mammals. The brain of living primates always possesses a true Sylvian sulcus (confluent with the rhinal sulcus) and a triradiate calcarine sulcus. Primates are unique among living mammals in that the brain constitutes a significantly larger proportion of fetal body weight at all stages of gestation. Male primates are characterised by permanent precocial descent of the testes into a postpenial scrotum; female primates are characterised by the absence of a urogenital sinus (i.e. urinary and reproductive tracts entirely separate). In all primates, involvement of the yolk-sac in placentation is suppressed, at least during the latter half of gestation. Primates have long gestation periods, relative to maternal body size, and produce small litters of precocial neonates. Fetal

growth and postnatal growth are characteristically slow in relation to maternal size. Sexual maturity is attained late and life-spans are correspondingly long relative to body size. Primates are, in short, adapted for slow reproductive turnover. The dental formula exhibits a maximum of 2.1.3.3/2.1.3.3. The size of the premaxilla is very limited, in association with the reduced number of incisors, which are arranged more transversely than longitudinally. The cheek teeth are typically relatively unspecialized, though cusps are generally low and rounded and the lower molars possess raised, enlarged talonids.

An additional distinctive feature should be added to this definition (Sussman, 1999): All living primate species live in fairly elaborate social networks. Initially, nocturnal primate species were labeled 'solitary' because many of them tend to move around alone at night. But countless field studies have since shown that they all have well-established patterns of social interactions based on shared ranges, communal nesting, and regular encounters over the course of the night. Although there are no immediately recognizable social groups of the kind found in all diurnal primates except the orangutan, there are nonetheless long-lasting associations between individuals. The pronounced social tendency now recognized for all primates is surely connected with the high level of investment in individual offspring (including intensive parental care) and relatively long life spans that characterize primates.

The distinctive set of characteristics shared by living primates reflects a suite of adaptations developed in their common ancestor. Taken in combination with various primitive features that can be inferred for the ancestral condition, such as relatively small body size and nocturnal habits (Martin, 1990), they permit a plausible reconstruction of the earliest primates. They were probably small-bodied, nocturnally active arboreal creatures that were particularly well adapted for exploitation of narrower branches away from main tree trunks (the fine branch niche), where they fed on a mixed diet of insects and fruit. Inclusion of fruits in the diet accounts for certain modifications of their cheek teeth. Leaping and running among the narrow-diameter terminal branches of trees and among lianas would account for their typical nail-bearing, grasping extremities, notably the prehensile foot. It also provides an explanation for emphasis of the visual sense, with large, forward-facing eyes and specialization of the brain for highly effective visual orientation. Predation on small animals in the network of fine branches doubtless also influenced the conspicuous development of the visual system that characterizes the primates (Cartmill, 1992; Cartmill et al., 2007). Development of prehensile extremities also involved a shift toward use of the hands rather than the snout in grasping food items, thus explaining shortening of the snout and the associated reduction in number of incisors. Adaptations for low reproductive turnover and high investment in individual offspring can be interpreted as responses to life under the relatively stable and predictable conditions of tropical forests combined with comparatively low predation levels among fine branches.

Phylogenetic Relationships among Living Primates

Encouragingly, pioneering analyses based on chromosomal and molecular data broadly confirmed reconstructions of phylogenetic relationships of primates based on morphological evidence, leading to the emergence of a general consensus (Martin, 1990). Although certain relationships in the primate tree remained controversial, notably with respect to early divergences, this confirmation, stemming from largely independent analyses of new data, provided valuable support for earlier phylogenetic reconstructions based exclusively on morphological evidence. Because very deep divergences are involved, however, the pattern of relationships between primates and other mammals remained uncertain and therefore controversial. Over the past two decades, there has been a dramatic increase in information from molecular comparisons, especially because of burgeoning results from direct sequencing of DNA, in some cases extending to the sequencing of entire genomes. As far as primates are concerned, this has led to further confirmation of the general pattern of phylogenetic relationships among them and to resolution of several aspects that had remained questionable. As far as mammals in general are concerned, the molecular evidence has led to the construction of overall trees that have yielded an entirely new perspective on the relationship between primates and other mammals.

Taking primates first, the substantial body of evidence now available clearly confirms that the five natural groups of extant primates (lemurs, lorises and bushbabies, tarsiers, New World monkeys, Old World monkeys, apes, and humans) are phylogenetically distinct subunits within the tree (**Figure 2**). In other words, each of the five natural groups is monophyletic (derived from a specific common ancestor within the tree). As far as lemurs and lorises are concerned, the question of monophyly remained controversial for some time. This was because certain cranial features led several authors to suggest that mouse and dwarf lemurs in the family Cheirogaleidae are more closely related to lorises and bushbabies than to other lemurs. In fact, a review of chromosomal and molecular evidence eventually provided quite convincing evidence that the lemurs form a monophyletic group distinct from a second monophyletic group containing lorises and bushbabies (Yoder, 1997). Since that time, numerous comparative analyses of DNA sequences have confirmed the monophyly of lemurs and lorises, notably in a study that included every extant primate genus (Goodman et al., 2005). Together, lemurs, lorises, and bushbabies constitute a more inclusive monophyletic group (**Figure 2**). Members of this group are known as strepsirrhines, because of their common retention of a rhinarium (a naked, moist area of skin surrounding the nostrils in the primitive mammalian condition).

The position of the tarsiers in the primate tree has long been a controversial issue. Traditionally, tarsiers were included in the suborder Prosimii along with lemurs and lorises. This largely reflected the fact that prosimian primates have retained numerous primitive features and naturally will resemble one another for that reason alone. Comparative morphological studies increasingly revealed that tarsiers share a number of significant derived features with higher primates (traditionally allocated to the suborder Anthropoidea). In fact, the earliest evidence for an evolutionary link between tarsiers and higher primates came from comparative studies of placentation at the end of the nineteenth century. Lemurs, lorises, and bushbabies uniformly have a noninvasive (epitheliochorial) type of placentation, contrasting starkly with the highly invasive

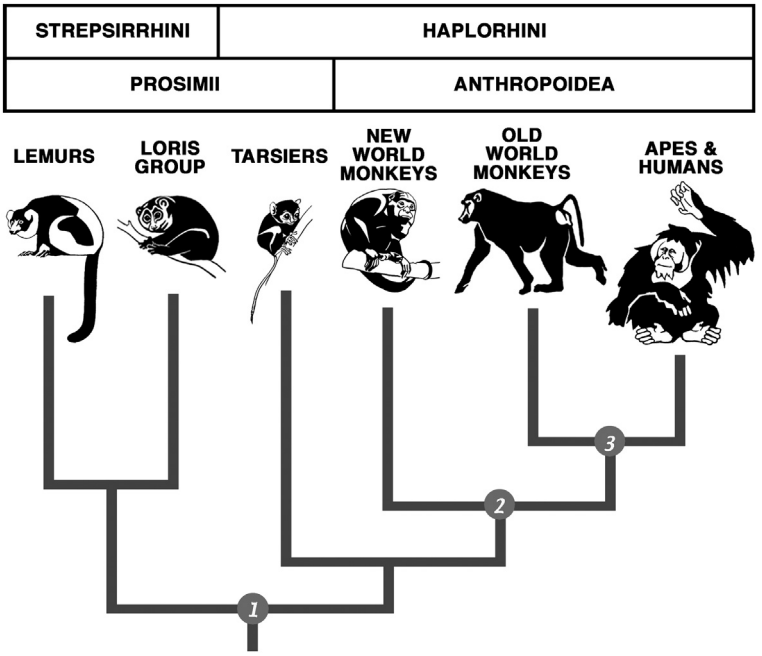

Figure 2 Outline phylogenetic tree of extant primates (excluding tree shrews) showing the likely pattern of relationships among the five monophyletic natural groups, which are all descended from the last common ancestor of primates (node 1). Traditional classifications distinguish between the generally more primitive prosimians (Prosimii) and the simians or higher primates (Anthropoidea), which are characterized by a set of more advanced features. Because there is now convincing evidence that tarsiers are related to higher primates , however, many classifications now have the two suborders Strepsirrhini (lemurs, lorises, and bushbabies) and Haplorhini (tarsiers and higher primates). It is now generally accepted that the higher primates are monophyletic, with a distinct common ancestor (node 2) and that Old World monkeys, apes, and humans form a lower level monophyletic group with a more recent common ancestor (node 3).

(hemochorial) placentation characterizing tarsiers and higher primates.

Evidence from placentation was progressively complemented with the identification of striking similarities between tarsiers and higher primates in the olfactory system, the visual system, and the brain. But many authors remained unconvinced by all of these similarities. In this case, too, the controversy was resolved gradually by accumulating molecular evidence, beginning with immunological comparisons, passing through studies of amino acid sequences in proteins and eventually culminating in analyses of DNA sequences. Before the advent of DNA sequence information, an overall review of evidence (Martin, 1990) indicated that tarsiers are, indeed, the sister group of higher primates. But doubt persisted thereafter because mitochondrial DNA (mtDNA) sequences failed to support a link between tarsiers and higher primates. This aberrant outcome has been shown to be the result of an unusual pattern of evolutionary change in the mtDNA of tarsiers (Schmitz et al., 2002). There is now overwhelming evidence from the studies of nuclear DNA sequences that tarsiers are, indeed, the sister group of higher primates. A study using vast amounts of sequence data from the nuclear genome (Jameson et al., 2011) clearly separated tarsiers from strepsirrhines and linked them to higher primates instead. In other words, tarsiers and higher primates together constitute a higher level monophyletic group of primates. They are collectively known as haplorhines because (uniquely among mammals) they share complete suppression of the rhinarium.

In the realm of molecular comparisons, particularly valuable information has emerged from studies of retroposons in primates and other mammals. These are DNA sequences, often of retroviral origin, that are inserted into the nuclear genome. Many are relatively short sequences known as SINEs (short interspersed elements). As far as is known, retroposons are inserted at random sites in the genome and subsequently are not eliminated. Insertion sites hence serve as identifiable characters that are virtually free of convergence or reversal. One prominent group of SINEs found only in the primate genome is constituted by Alu sequences. Detailed investigation of Alu insertions in primates has revealed that tarsiers and higher primates share several unique retropositions that are lacking in strepsirrhines (Zietkiewicz et al., 1999; Schmitz et al., 2005). When this evidence from site-specific insertions is seen in conjunction with that from numerous studies of other nuclear DNA sequences, it now seems virtually certain that tarsiers and higher primates are sister groups.

The overall pattern of relationships between the major groups of primates has now become relatively clear. On one side of the primate tree, it has generally emerged that Lemuriformes and Lorisiformes are sister groups (i.e., their individual common ancestors were derived from an earlier common ancestor). On the other side, it has also generally emerged that tarsiers are linked to higher primates and that the latter (monkeys, apes, and humans) together constitute a distinct monophyletic group.

Before the availability of molecular evidence, there was in fact some doubt regarding a monophyletic origin of higher primates. On morphological grounds, several authors proposed that New

World monkeys on the one hand and Old World monkeys and apes on the other descended separately from prosimians that lacked certain defining features of higher primates. In other words, it was suggested that key similarities between modern New World and Old World higher primates arose through parallel evolution. From the outset, however, molecular evidence has consistently and unequivocally indicated that higher primates belong to a monophyletic group. Although molecular evidence has not permitted identification of the morphological characters present in the common ancestor of higher primates, it does indicate the existence of a long branch between the divergence of tarsiers and the emergence of that common ancestor. Molecular evidence consistently has indicated that New World and Old World higher primates are monophyletic sister groups. Moreover, within the Old World cluster, Old World monkeys and apes (including humans) have also been identified as lower level monophyletic groups in every study that has been conducted.

Molecular evidence has been crucial for the identification of higher level relationships among mammals. Although many attempts have been made to identify such relationships from morphological evidence, no clear consensus developed. Indeed, in some cases in which several morphological studies began to suggest a consensus interpretation, notably in linking bats to colugos (Volitantia), molecular evidence has yielded directly conflicting results. In the course of a little more than a decade, large-scale molecular studies across mammals have led to a fairly clear consensus interpretation regarding relationships among the 18 extant orders of placental mammals. In particular, it was quite soon recognized from analyses of DNA sequences that there are four main clusters of placental mammals (Murphy et al., 2001; Springer et al., 2004). Subsequent studies, notably including analyses of SINEs (Nishihara et al., 2006), repeatedly have confirmed this basic phylogenetic pattern, as indicated in a seminal synthesis of abundant available DNA evidence (Bininda-Emonds et al., 2007). The four superordinal clusters that have consistently emerged are (1) Afrotheria (aardvarks, afrosoricid insectivores, elephants, elephant-shrews, hyraxes, and sirenians); (2) Euarchontoglires (colugos, lagomorphs, primates, rodents, tree-shrews); (3) Laurasiatheria (artiodactyls with cetaceans, bats, carnivores, eulipotyphlan insectivores, perissodactyls, and pangolins); and (4) Xenarthra (a single order including anteaters, armadillos, and sloths). Among other things, this arrangement plainly rules out any phylogenetic link between bats and colugos, as the former are included in the superorder Laurasiatheria, while the latter belong to the superorder Euarchontoglires. Although the superordinal cluster Euarchontoglires has emerged consistently from molecular studies across mammals, relationships within that cluster have remained somewhat labile. The consensus of available evidence indicates that colugos are most likely to be the sister group of primates within a cluster known as Archonta, while lagomorphs and rodents are sister groups within a second cluster known as Glires. Several studies have linked tree-shrews and lagomorphs as sister groups. Regardless of the actual pattern of relationships within Euarchontoglires, the inclusion of colugos is a striking innovation. It has rarely been suggested on morphological grounds that colugos have any relationship to primates, and they have generally been completely ignored in discussions of primate evolution. Clearly, this must change.

The Primate Fossil Record

The known record of reliably identified fossil primates dates back to the beginning of the Eocene epoch, about 55 mya (Hartwig, 2002). An assemblage of earlier fossil forms mainly from the Paleocene epoch (65–55 mya), collectively labeled 'archaic primates' (Plesiadapiformes), traditionally has been allocated to the order Primates. These early fossil forms (some of which survived into the Eocene) show few, if any, of the features in the definition of primates given above, and their connection with primates remains questionable. Even if plesiadapiformes did have some early connection with primates, they undoubtedly diverged before the origin of 'primates of modern aspect' (euprimates), which include extant primates and their direct fossil relatives. Because plesiadapiformes are a side-group that diverged at such an early stage, they are of little relevance to the later adaptive radiation of euprimates.

Excluding Plesiadapiformes, the currently documented primate fossil record is divided fairly sharply into two segments separated by a gap of at least 6 million years in the middle of the Oligocene epoch (35–25 mya). Approximately equal numbers of fossil primate species (about 250) are known from each half of the record. Initially, primates of the Eocene epoch (55–35 mya) were documented only by early prosimians from Europe and North America. On both continents, they can be divided fairly clearly into two families, generally larger lemurlike Adapiformes (e.g., *Adapis* in Europe, *Notharctus* in America), and generally smaller tarsierlike Omomyiformes (e.g., *Necrolemur* in Europe, *Shoshonius* in America). Many authors have linked adapiforms to strepsirrhines and omomyiforms to haplorhines. Almost the entire skeleton is known for *Notharctus*; it is lemurlike and certainly demonstrates that many defining features of modern primates were present at that stage. Unfortunately, the postcranial skeleton of omomyiforms is poorly documented in the known fossil record, although several associated elements of *Shoshonius* have provided useful information regarding skeletal morphology.

This long-established record from America and Europe has been radically modified by more recent fossil finds from North Africa, Asia, and Southeast Asia. In North Africa, fossils from the latest Eocene and early Oligocene, notably from the Fayum site in Egypt, have documented the presence of a variety of early simians (e.g., *Aegyptopithecus* and *Apidium*) along with some potential relatives of modern strepsirrhines and tarsiers. Adding to the picture, Chinese middle Eocene fossil deposits have yielded evidence of early simians (*Eosimias*) and a putative direct relative of modern tarsiers (*Tarsius eocaenus*), along with adapids and omomyids. A well-preserved skull of an early Eocene omomyiform, *Teilhardina*, has also been reported from China. Moreover, late Eocene sites in Myanmar and Thailand have yielded several fossil simians (e.g., *Amphipithecus* and *Siamopithecus*), although they are almost exclusively known from jaws and teeth. Overall, the fossil record now shows that both prosimians and simians were present in the Eocene.

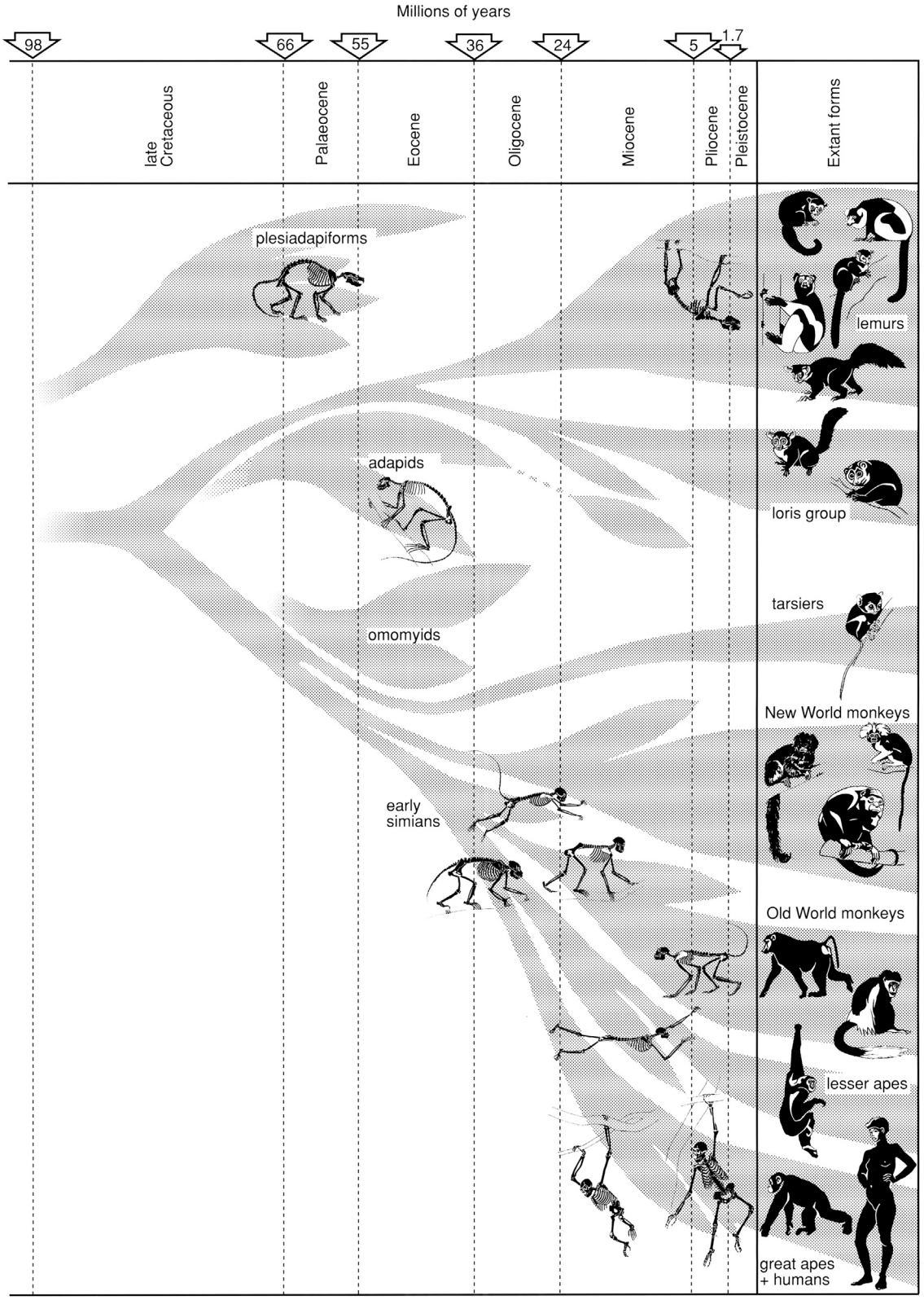

Figure 3 Overall phylogenetic tree of primates (excluding tree-shrews) showing the likely pattern of relationships among the five monophyletic natural groups of living primates. This tree also includes major groups of fossil representatives. Note that no direct connection between plesiadapiformes and 'primates of modern aspect' is shown because evidence for such a link is questionable. The initial date of divergence for primates of modern aspect is set at about 80 mya. Reproduced from: Martin, R.D., 1993. Primate origins: plugging the gaps. Nature 363, 223–234; original illustration by Lukrezia Bieler-Beerli.

Things are entirely different in the second half of the primate fossil record, extending from the late Oligocene to the present day. Direct fossil relatives of New World monkeys are known from South America, beginning in the latest Oligocene (*Branisella*) and extending into the Miocene (e.g., *Homunculus*). In the Old World (notably Africa), early Miocene deposits have yielded fossil relatives of bushbabies and lorises (e.g., *Progalago*) and of monkeys and apes (e.g., *Victoriapithecus* and *Proconsul*). A few fragments of tarsiers have been reported from Miocene deposits in Thailand. Thus, four of the five extant natural groups of primates can be traced back to about 25 mya, and one (tarsiers) can be tentatively traced back to 40 mya. Nevertheless, there are also notable gaps in the known fossil record. The most striking gap is in Madagascar, where no single fossil from the major radiation of lemurs is known prior to the recently extinct subfossils. Similarly, in the New World, no single fossil form can be linked reliably to the marmosets and tamarins, a substantial proportion of the modern species.

Gaps in the fossil record influence directly the inferred date for the origin of primates. The standard practice has been to assume that primates diverged during the Paleocene epoch, perhaps 60–65 mya, just a few million years before the earliest primates known from the fossil record. Direct inference from the known fossil record inevitably leads to serious underestimation if there are substantial gaps in the fossil record, as is undoubtedly true of primates (Martin, 1993). Several studies that calibrated molecular phylogenies for mammals using various dates for fossils of well-documented groups outside the primates have all concluded that primates diverged from other mammals about 90 mya. Allowing for some elapse of time between this initial divergence from other mammals and initiation of the adaptive radiation leading to the array of extant primates, this would imply that their last common ancestor existed about 80–85 mya (Figure 3). Two analyses of the primate fossil record that explicitly allowed for the existence of gaps have yielded a very similar date (Tavaré et al., 2002; Wilkinson et al., 2011).

See also: Adaptation, Fitness, and Evolution; Body, Evolution of; Brain, Evolution of; Comparative Method in Evolutionary Studies.

Bibliography

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446, 507–512.

Cartmill, M., 1992. New views on primate origins. Evolutionary Anthropology 1, 105–111.

Cartmill, M., Lemelin, P., Schmitt, D., 2007. Primate gaits and primate origins. In: Ravosa, M.J., Dagosto, M. (Eds.), Primate Origins: Adaptations and Evolution. Springer, New York, pp. 403–435.

Goodman, M., Grossman, L.I., Wildman, D.E., 2005. Moving primate genomics beyond the chimpanzee genome. Trends in Genetics 21, 511–517.

Groves, C.P., 2005. Order primates. In: Wilson, D.E., Reeder, D.M. (Eds.), Mammal Species of the World: A Taxonomic and Geographic Reference, vol. 1. Johns Hopkins University Press, Baltimore, pp. 111–184.

Hartwig, W.C. (Ed.), 2002. The Primate Fossil Record. Cambridge University Press, Cambridge.

Jameson, N.M., Hou, Z.-C., Sterner, K.N., Weckle, A., Steiper, M.E., Wildman, D.E., 2011. Genomic data reject the hypothesis of a prosimian primate clade. Journal of Human Evolution 61, 295–305.

Kinzey, W.G. (Ed.), 1997. New World Primates: Ecology, Evolution, and Behavior. Aldine de Gruyter, New York.

Martin, R.D., 1990. Primate Origins and Evolution: A Phylogenetic Reconstruction. Princeton University Press, Princeton, NJ.

Martin, R.D., 1993. Primate origins: plugging the gaps. Nature 363, 223–234.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.

Napier, J.R., Napier, P.H., 1967. A Handbook of Living Primates. Academic Press, London.

Nishihara, H., Hasegawa, M., Okada, N., 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proceedings of the National Academy of Sciences USA 103, 9929–9934.

Schmitz, J., Ohme, M., Zischler, H., 2002. The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Molecular Biology and Evolution 19, 544–553.

Schmitz, J., Roos, C., Zischler, H., 2005. Primate phylogeny: molecular evidence from retroposons. Cytogenetics and Genome Research 108, 26–37.

Smith, R.J., Jungers, W.L., 1997. Body mass in comparative primatology. Journal of Human Evolution 32, 523–559.

Springer, M.S., Stanhope, M.J., Madsen, O., de Jong, W.W., 2004. Molecules consolidate the placental mammal tree. Trends in Ecology & Evolution 19, 430–438.

Sussman, R.W., 1999. Primate ecology and social structure. In: Lorises, Lemurs and Tarsiers, vol. 1. Pearson Custom Publishing, Needham Heights, MA.

Tattersall, I., 1982. The Primates of Madagascar. Columbia University Press, New York.

Tattersall, I., 1988. Cathemeral activity in primates: a definition. Folia Primatologica 49, 200–202.

Tavaré, S., Marshall, C.R., Will, O., Soligo, C., Martin, R.D., 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416, 726–729.

Wilkinson, R.D., Steiper, M.E., Soligo, C., Martin, R.D., Yang, Z., Tavaré, S., 2011. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Systematic Biology 60, 16–31.

Yoder, A.D., 1997. Back to the future: a synthesis of strepsirrhine systematics. Evolutionary Anthropology 6, 11–22.

Zietkiewicz, E., Richer, C., Labuda, D., 1999. Phylogenetic affinities of tarsier in the context of primate Alu repeats. Molecular Phylogenetics and Evolution 11, 77–83.

Primitive Society

Alexander F Robertson, University of California Santa Barbara, Santa Barbara, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12046–12050, © 2001, Elsevier Ltd.

Abstract

'Primitive' is a term from the evolutionary vocabulary of the nineteenth century that in popular understanding still identifies the interests of the modern discipline of anthropology. Anthropologists have sought to come to respectful terms with their 'Primitive' subject matter by redefining it in terms of a distinctive social structure, culture, spatial distribution, and type of livelihood. But despite efforts to set aside disparaging contrasts between primitive society and modern civilization, evolutionary assumptions about human progress remain embedded in both scientific and humanistic approaches to society.

In popular usage, 'primitive society' distinguishes ourselves from other people who have made little progress toward what we understand as civilization. There are very few such people left today: they live in scattered communities in deserts or rainforests, and they interest us mainly because we think of them as the living relics of our own distant ancestors. The comparison is usually unflattering. Until the beginning of the nineteenth century 'primitive' simply meant 'first' or 'earliest,' but as the word was applied to the original inhabitants of territories colonized by European and American states, it acquired the connotations of inferior, backward, rude.

In the evolutionary social science which developed toward the end of the nineteenth century, 'primitive society' came to be regarded as the defining subject matter of the emerging discipline of anthropology. The new natural science of biology provided the basic inspiration, and physical anthropology remains one of the four subfields of the discipline. However, although physical (or 'biological') anthropologists are much concerned with comparisons among primates they no longer distinguish human bodies as more or less primitive. We vary from one population to another in average body size, skin color, hair type, etc., and people within each community (neighbors or family members, young and old) look unalike, but unless we wish to be regarded as 'racist' we studiously avoid implying that any of these differences connotes inferiority or even some sort of adaptive advantage. Physically we are neither less 'primitive' nor more 'modern' than our ancestors 50–100 000 years ago.

It is, therefore, somewhat ironic that Darwinian hypotheses of the 'ascent of man' by natural selection have provided such a tenacious metaphor for distinguishing among human societies and cultures. For most social scientists for most ofthe twentieth century, being 'primitive' was a relative judgment about how people think and behave, not about their genes, brains, or bodies. Nineteenth-century explanations of social evolution construed Western civilization not just as the most successful human adaptation, but also as an inevitable and morally necessary path of progress.

While metabolizing or copulating might have changed little over the millennia, there were certainly more and less civilized ways of dining or marrying. And it was significant that only civilized modern man had the capacity to understand and explain the difference.

The first generation of modern anthropologists sought a general, empirically valid definition of the 'primitive,' and used the word without inhibition. Tylor's article defining anthropology in the famous 11th edition of *Encyclopedia Britannica* (1911) may represent the apogee of the technical use of the term. Thereafter, 'primitive' continued to appear in the title of ethnographic monographs, more as a generic label for subject matter than as an analytical concept. By the middle of the twentieth century, even the label had gone out of favor in the social sciences. The dissolution of the European empires and the emergence of new, transnational political-economic interests brought a shift in intellectual attitudes. Ideas of 'underdevelopment' as a natural, primitive condition which would be relieved by technically-enhanced progress toward modernity, soon yielded to ideas of 'underdevelopment' as a political-economic predicament inflicted on citizens of the 'Third World' by the expansion of industrial capitalism from its metropolitan centers. Peoples in the new states who were once classified as 'primitive' were now identified as the least privileged members of a widely distributed 'Fourth World,' which included the impoverished citizens of 'developed' countries.

The only entry under 'primitive' in the 1968 counterpart of this *Encyclopedia* referred to naive styles in art. 'Primitive society' has resurfaced in the present work as a historical curiosity rather than as current terminology. The phrase says more about changing interests, attitudes, and theories in anthropology than about the peoples to whom it once referred. In trying to come to terms with their designated subjects during a century of radical social change, anthropologists have buried 'primitive' society in layers of definition. These include specification of a distinctive sort of social structure, mode of livelihood, spatial distribution, and mentality, all defined (tacitly or explicitly) by a contrast with, and presumed progress toward, civilization as we understand it. Very occasionally an anthropologist will resort to 'primitive' as a portmanteau term: Hallpike (1976) uses it 'simply as a convenient abbreviation for "non-literate, small-scale, face-to-face-societies with subsistence economies." Although the pejorative implications are resisted sternly, there is still an evident need to refer contrastively to people who have not yet been absorbed within the ambit of 'literate, large-scale, impersonal societies with industrial economies.' Rejecting one word around which a discipline was constructed is not enough to dismiss the assumptions about evolutionary progress which linger on in our understanding of the differences between 'Us' and the sorts of people who are 'Our' anthropological subjects.

Livelihood

The simplest and least obviously offensive definition of the primitive has been the material circumstances of their lives. Anthropology as a discipline is still organized around a sequence of social types based on technical developments in the means of production. The prototypical human population of 'hunter-gatherers' lived in nature, taking what they needed day-by-day, whereas subsistence farmers were obliged to work on nature with tools and skills, taking possession of land and other resources, and thereby setting in motion the larger political and economic organization of the state. Technical complexity increased the division of labor, which produced societies in which larger numbers of people depended on more specialized functional relationships with each other. This was especially true of the second great technical revolution which began with the industrialization of cloth production in northwest Europe toward the end of the eighteenth century. With capitalism as its social vector, industrialism swept out to transform the world, creating among other things a new and palpable contrast between the modern and the primitive. It also produced the first generation of social scientists whose role in the ever-expanding division of labor was to explain what was going on.

Social Structure

In the pattern of evolutionary contrasts which emerged in the writing of Morgan, Tylor, Rivers, and the other early anthropologists, the social world of the primitive was the small, technically, and socially 'simple' community. In this evolutionary formulation, 'primitive society' was almost a contradiction in terms, "society" being seen as the multiplication of roles and relationships in expanding populations, with heterogeneous individuals bound by new forms of rational contract into complex entities like the modern nation state. The primitive, by contrast, was a 'society of clones' (Kuper, 1988), held together 'mechanically' (Durkheim's term) by the relations of reproduction. Social anthropologists pieced together the logic of systems of descent and marriage which apparently gave these 'homogeneous' communities their cohesion. For example, the simple rule 'reckon that you are the child of your father rather than your mother' was enough to tease a whole population out into mutually-exclusive (patri-) lineages, conferring distinct political identities on people who might never actually meet each other. Debates raged about whether paternity or maternity was the more primitive mechanism of social solidarity. The 'elementary structures of kinship' (clans, demes, moieties, etc.) conferred statuses and roles by birth, but these 'ascribed' relationships were less useful as people were drawn into much larger fields of interaction in technically more advanced societies where survival and progress depended on geographical movement and the exercise of choice and imagination.

Governed by the biologically-derived moral mechanisms of kinship, the primitive tribe could manage without officials and rulers. From an imperial perspective, these emerging definitions of the 'tribe,' boosted by romantic writings of the period, were both intellectually attractive and administratively useful. The tribe suggested a prototypical democracy, a predominantly male political structure, and a paradigm of nationalism – the great seductive and destructive passion of the twentieth century. After two world wars which revealed the barbarity of civilization and cast doubts on who was more primitive than whom, the natural logic of the 'tribe' (as expounded, e.g., by Evans-Pritchard for *The Nuer* (1940) of the southern Sudan) was appealing. In the 1968 *Encyclopedia*, 'Primitive society' was broken down into seven cross-references: Anthropology, Culture, Evolution, Hunting and Gathering, North American Indians, and – the principal entry – Tribal Society. However, the 'Tribal' label could not stick. The year 1968 also brought the revival of the intellectual left, and the association of 'tribalism' with the ideologies and practices of imperialism quickly made it taboo. Anthropologists were now being accused of complicity in the task not just of identifying but inventing the tribes which had been the units of European and North American colonial government.

At this time, the other substitution for 'primitive' was traditional, the preferred counterpart for modern in the evolutionary scheme. But very soon this too was found to be pejorative, mainly because the innumerable books juxtaposing the two words in their titles seemed to presume an ethnocentric convergence on a Euro-American paradigm of modernity. 'Traditional societies' were scattered and diverse, stuck in the inertia of custom until history drew them inexorably into the global monotype of 'modern society.' In the 1980s such teleologies of modernization were rejected by scholars who defined themselves philosophically as postmodern. They wrote critically of our historic and ideological self-centeredness and in praise of a diverse category of 'The Other,' which stood mostly against the hegemony of (Western) scholarship itself. An emblematic study was Clastres's (1977) account of the Tupi-Guarani indians of the Amazon, whose profoundly primitive social structure supposedly set up antibodies for the development of inequality, leadership, capitalism, the state, and the other vices of modern civilization. Such writing paved the way for a revival of sympathetic interest in the 'primitive.' Postmodern characterizations of 'Otherness' invoked a much older assumption about primitive society: its spatial isolation. Earlier anthropologists had pursued the prototypical primitive to the ends of the earth, especially Australia and Oceania, and ethnographic accounts of these tiny encapsulated communities in turn fleshed out the definition of 'primitive.' Postmodern anthropologists pointed out that the (relatively) concerted forces of worldwide capitalism had tended to marginalize people physically remote from the metropolitan centers, making 'peripheralness' a common denominator of what had hitherto been called traditional, tribal, or primitive. The desire to see 'Others' entirely in their own terms as just 'different' not only meant abandoning pernicious comparisons with 'Us' but also the assumption that the 'Others' had anything specific in common among themselves – some generic sort of 'traditional' culture or 'primitive' social structure.

Most problematically, it was not even safe to assume that everyone had essentially the same sort of body, because biology or physiology themselves were only historically specific sets of ideas for coming to terms with the human condition. This deliberate affront to science has exacerbated a split in anthropology, set in motion by the invention of 'sociobiology' in the

1970s. With 'evolutionary psychology' as the link, a new genre of anthropology is pursuing genetically endowed physical mechanisms which shape culture, for example, through the processes of language acquisition or mate selection. From this point of view, 'primitive society' as represented by today's vanishing hunter populations are living testimony to the 'Environment of Evolutionary Adaptedness' in which modern man was physically constructed and mentally endowed. Virtual banking or an appreciation of Beethoven is fundamentally as 'primitive' as gnawing bones. These exciting speculations have captivated public attention, putting the denser, relativistic theories of idealist anthropologists at a disadvantage.

However, it is ironic that the radically Cartesian post-modern approaches to society have taken the psychic unity of human beings entirely for granted: it is not our minds which differ, but the way they are culturally made up. This returns us to some very old and persistent mental assumptions underlying the meaning of 'primitive' in its various transmutations.

Mentality

In drawing analytical metaphors from biology or botany the social sciences and humanities have not always kept pace with theoretical advances in the natural sciences themselves. The nineteenth century evolutionist idea of recapitulation was absorbed by the emerging social sciences, and lingers as a latent assumption about human behavior in forms which biological scientists would now barely recognize. The original formula, as specified by Ernst von Haeckel, was that the life history of the individual organism repeats, from the embryo onward, the stages of its species evolution ('ontogeny recapitulates phylogeny'). Transposed to the social domain, this fostered the notion that children are primitive – and, by a facile converse, that primitives are children. Thus, Lubbock (1875) was fascinated by 'The similarity of savages and children,' and struggled to prove that the childish trait of phonetic reduplication (words like *bonbon)* was an infantile characteristic of primitive languages, absent from civilized English, French, German, and Greek. For social philosophers like Henry Sumner Maine, primitive society was 'the infancy of man,' and as late as 1953 Robert Redfield commended the human race for its tendency 'to grow up.'

The metaphoric transfer of this early biological concept to society and culture continues to create much confusion. Although it is implausible that a modern child could be reliving the traumatic experiences of ancestral adults, the idea that children relive primitive urges remains at the heart of Freudian psychology. Indeed, in his study of *Totem and Taboo* Freud (1938[1919]) pursued the argument to 'resemblances between the psychic lives of savages and neurotics.' The logic of recapitulation leads inexorably to the conclusion that the metamentality of 'culture' is, for primitive people, childish and crazy, and that their world is a kindergarten which can only benefit from some civilized parenting – a good excuse for colonization and imperialism. This elaborate equation of the mad, the childish, and the primitive is extraordinarily tenacious, despite many attempts to sort out the confusion. Margaret Mead, for example, sought to demonstrate that primitive Melanesian children were actually 'more rational' than their 'animistic' parents. This subverts Enlightenment assumptions that rationality increases as a function of both of individual and species development. But for such a devout idealist and anti-evolutionist as Mead, it also raises disturbing question about how children everywhere come ready-equipped with this sort of scientific rationality, and think straight until 'enculturation' fills their primitive minds with superstitious rubbish.

Anthropologists have always traded on public fascination with the magic and superstition we associate with primitive (and childish) thought. Given its persistence in modern society, religion itself has not been an acceptable marker of the primitive. The discriminating variable here is, once again, scale: the universalistic 'World Religions' (Christianity, Judaism, Islam) deal with the mass of individual souls, while the narrow-minded particularistic religions of primitive society connect people spiritually to community and locality ('animism,' 'fetishism') often through the biosocial idiom of descent ('ancestor worship,' 'totemism'). These peculiarities of primitive thought have interested French anthropologists particularly, immersed as they are in the idealist tradition. Inspired by Franz Boas, Kroeber, Mead, Benedict and other American anthropologists tried very explicitly to reject the evolutionary premises, and pursued a relativistic interpretation of culture (rather than 'society') which sought to avoid the prejudices of a superior western civilization. However, the central conundrum of relativism seems inescapable: how can we distinguish one culture from another if we have no agreed definition (some sort of 'metaculture') of what culture is in the first place? Simply to insist that all human beings have the mental capacity to 'do' culture evades the issue of how that capacity is physically embodied, how it constrains behavior, and (the evolutionary question) how and why we need it.

The most fundamental belief, which very few modern anthropologists would dare challenge, is of psychic unity, a conception of the human mind as inhabiting a standard brain which may determine how but not what we think. In the idealist tradition, what we think is determined outside of ourselves in an integral, self-reproducing system of ideas – culture, with a 'mind' of its own. The trouble is how to reconcile the enormous differences in culture with the primitive sameness of our minds. A symptom of lingering evolutionism has been the association of this primitive mentality with a closeness to nature.

This is captured in the old term savage, a synonym for primitive which appeared in anthropological titles well into the twentieth century, though increasingly in ironic quotation marks (Lévi-Strauss (1966) *The Savage Mind*; Goody (1971) *The Domestication of the Savage Mind*; LaFontaine (1981) *The Domestication of the Savage Male*, etc.). Although in French 'sauvage' still has a more benign, Rousseauian ring, the English connotations offierceness and intemperateness latterly made it at least as stigmatic as 'primitive.'

Seeking a general and objective axis of intellectual difference, many anthropologists have settled on literacy as a key technical distinction between primitive and modern mentality. Thus, in the early 1930s, Margaret Mead uses the word 'primitive' to refer to those peoples who are completely without the benefits of written tradition, and in 1953 Robert Redfield declared 'I shall use "primitive" and "preliterate" interchangeably' (he also used a range of cognate terms including 'folk' and

'precivilized'). By the 1970s pre- or nonliterate had become the favored term for the societies traditionally discussed by anthropologists, and it remains a key universal classification for those working in the idealist tradition.

Nostalgia

A final reason for the persistence of notions of the primitive within and beyond anthropology is plain old-fashioned nostalgia – the old image of the noble savage as a counterpoint to the vices of civilization. Anthropologists are notorious for their populist commitment to the people among whom they have their main (usually first) ethnographic experience. Indeed, it has often been said that anthropologists are people seeking a special sort of redemption in the notion that the moral virtue lacking in so-called civilized society may be rediscovered among simple people who are more in tune with nature. Defense of primitive peoples, usually in conjunction with the protection of their forest or desert environments, has become a respectable political commitment for many anthropologists. Despite Marx's rage against populist utopianism, his own material laws of history imagined a return to the perfect altruism of primitive society. This image of our natural selves before we were corrupted by the rise of capitalism was inspired by the ethnographic vision of 'Ancient Society' in North America constructed in the mid-nineteenth century by one of the founding fathers of anthropology, Lewis Henry Morgan. Marxists have been rebuked for urging the violent recovery of this old society without offering any plausible account of its structure. That it should somehow combine the benefits of the scale, technical expertise, occupational specialization and intellectual achievements of modern industrial society directly contradicts most of what has been learned about primitive society.

Today, the idea of the primitive lives on in a schizophrenia of admiration and disgust. On the one hand, popular writers like Robert Kaplan have promised the return of the modern world to barbarism, while anthropologists like Paul Richards explain that what appears as primitive lapses in Sierra Leone or Cambodia is a wholly modern circumstance – the global failures of industrial capitalism. In whatever guise, the concept of the primitive seems irrepressible in our imagination of ourselves, lurking in the shadows as we struggle to detach our vile, ancient bodies from the dream of a transcendent modern mind.

Bibliography

Clastres, P., 1977. Society Against the State. Blackwell, Oxford, UK.

Durkheim, E., 1933. The Division of Labor in Society. Free Press, Glencoe, IL.

Evans-Pritchard, E.E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Clarendon Press, Oxford, UK.

Freud, S., 1938[1919]. Totem and Taboo: Resemblances Between the Psychic Lives of Savages and Neurotics. Penguin, Harmondsworth, UK.

Goody, J., 1990. The Oriental, the Ancient and the Primitive: Systems of Marriage and the Family in the Pre-industrial Societies of Eurasia. Cambridge University Press, Cambridge, UK.

Hallpike, C.R., 1976. Is there a primitive mentality? Man 11, 253–270.

Kaplan, R.D., 1994. The coming anarchy: how scarcity, crime, overpopulation, and disease are rapidly destroying the social fabric of our planet. Atlantic Monthly February 273 (2), 44–76.

Kuper, A., 1988. The Invention of Primitive Society: Transformations of an Illusion. Routledge, London.

Lewis, I.M., 1968. Tribal Society. International Encyclopedia of the Social Sciences. Free Press, Glencoe, IL. 146–151.

Lowie, R., 1920. Primitive Society. Boni and Liveright, New York.

Lubbock, J., 1875. The Origin of Civilisation and the Primitive Condition of Man: Mental and Social Conditions of Savages, third ed. Longmans, Green, London.

Maine, H.J.S., 1893[1874]. Lectures on the Early History of Institutions, sixth ed. Murray, London.

Mead, M., 1932. An investigation of the thought of primitive children, with special reference to animism. Journal of the Royal Anthropological Institute 62, 173–190.

Redfield, R., 1953. The Primitive World and its Transformations. Cornell University Press, Ithaca, NY.

Stocking Jr., G.W., 1982. Race, Culture, and Evolution: Essays in the History of Anthropology, Phoenix edn. University of Chicago Press, Chicago.

Terray, E., 1972. Marxism and 'Primitive' Societies. Monthly Review Press, New York.

Tylor, E.B., 1913[1871]. Primitive Culture, fifth ed. Murray, London.

Printing as a Medium

Adrian Johns, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12050–12055, © 2001, Elsevier Ltd.

Abstract

This article first defines printing as the mechanical reproduction of texts using a press. It then examines the history of the practice from Gutenberg to the industrial era, identifying three levels on which society sought to shape its development: the levels of domestic morality (within the printing house itself), craft organization (within the guild of master printers), and the polity (within the church and state). It outlines the regulatory mechanisms (patents, licenses, etc.) that shaped early printing, before briefly showing how those mechanisms, in collapsing, gave rise to modern legal regimes, especially copyright. The article then moves on to appraise social–scientific interpretations of printing's cultural impact. It describes the arguments of scholars including Jürgen Habermas, Marshall McLuhan, Elizabeth Eisenstein, Robert Darnton, and Roger Chartier. It concludes by suggesting that in future, as the unquestioned cultural dominance of print finally comes to an end, we may be more prepared to see its effects as deriving as much from our practices of reading as from a cultural logic that might have been imposed on us by the press itself.

Printing is a technique for producing all kinds of paper-based records (books, periodicals, musical scores, ephemera, etc.) by means of ink, a press, and reusable characters. It permits the mechanical production of many identical copies, thereby eliminating the corruption inevitable in hand copying. Printing has had profound consequences in all spheres of social life. In particular, it has often been ascribed a major role in the transition to modernity that took place in the Western world after the Renaissance.

The History and Impact of Printing

The Invention of Printing

Techniques for the mechanical reproduction of texts are not a uniquely Western invention. Such techniques existed in Korea and China hundreds of years before they did in the West. Yet, it is in the West that the press had its greatest impact. It is customary to divide the history of printing there into two periods: the hand-press era, lasting from the mid-fifteenth century to the early nineteenth; and the industrial era, beginning shortly after 1800 with the invention of the steam press and lasting into the present. Social scientists' arguments about the effects of printing as a medium tend to concentrate on the former, in an effort to help explain modernization in terms of a putative 'printing revolution.' But, as will be suggested below, they would often be more aptly applied to the latter.

There is no plausible route by which Asian technologies may have influenced Johann Gutenberg (*c*.1398–1468), a goldsmith from the Rhineland town of Mainz who built the first European press, so he may legitimately be called its inventor. At first, his intent was anything but revolutionary. When he constructed his first press, Gutenberg did so as a calculated intervention in the world of commercial manuscript production that already existed. His earliest printed books were designed to resemble manuscripts, and he intended to sell them as such. He also swore his workmen to confidentiality about his new technique, apparently, in the manner of apprentices in any late-medieval craft. But the strategy did not last. In 1462 Mainz was besieged and sacked, and Gutenberg's printing house broke up. His workers went to other towns and set up their own operations. Before long, printers had appeared in a score of German, Dutch, Italian, and French cities, often fostering legends that printing had been invented locally. Venice, Avignon, Strasbourg, and Haarlem created the most plausible of such histories, as a result of which determining the true origins of the press remained a notoriously intractable problem well into the nineteenth century. William Caxton set up his printing house in London at the end of this first wave of expansion. By that time, in 1480 or so, a new European craft was firmly established.

The Practice of Early Printing

That craft centered on the press – a simple, largely wooden machine, capable of being made by semiskilled workmen and, if small enough, amenable to concealment and transport from place to place. This press would normally be operated in a place called a printing house. Here, a compositor would set type into a frame called a form, which would then be placed in the press itself, which two pressmen would operate to make up to 1000 impressions per day. The resulting sheets would then be dried, and collated to make books. It sounds, and in essence it was, a simple process. But in practice, it was hard to sustain. Most printing houses were severely undercapitalized, and depended for their day-to-day survival not on the printing of books, but on that of ephemera and pamphlets – anything, in fact, that furnished regular and frequent influxes of cash. Almost all of these products were destroyed soon after manufacture. The population of books that has lasted to the present day to fill our great research libraries therefore represents an extreme distortion of the reality of early modern print culture.

This being so, printing raised an immediate question: whether culture could be allowed to be defined by commercial necessity. Early modern Europe attempted to answer that question by constraining printing on three social levels: the household, the craft, and the polity (where the polity might be a religious one, such as the Catholic Church).

In the first place, a printing house, like most places of work in early modern Europe, was normally supposed to

be a residence as well. The master printer generally lived in the same building – as, quite frequently, did his journeymen and apprentices. There was a reason why this was held to matter. It was conventional to believe that the moral structure of the patriarchal household largely guaranteed the legitimacy of practices pursued within that household, and printing was to be subjected to the same moral oversight.

In the second place, printing was also organized into guilds, one of which existed in every major city. To be a master printer was to be a respected member of such a group. Each maintained its own conventions, expected to be honored by all members of the book trade. Violations were identified and remedied by book-traders themselves, generally behind closed doors, so that for outsiders the image of an intrinsically harmonious order of print was sustained. Again, there was nothing strange about this; it was how all early modern crafts operated.

And in the third place, such guilds operated in fragile harmony with regulations imposed by the church and the state. Chief among these were 'privileges' (patents) and licenses. Privileges were grants of literary property, to use an anachronistic phrase, given out by the monarch or an equivalent authority. They conflicted in principle with titles allocated by the guilds internally, which was one reason why they were eventually replaced by copyright conventions. A license, on the other hand, was a permission to print (in Latin, imprimatur, 'it may be printed'), granted by an authorized individual to the printer or bookseller. In principle, all works had to be vetted by such an official before publication, but in practice this was rather less severe than the censorship sometimes portrayed by modern historians. Most books were never licensed, and only a tiny minority were pursued for lack of an imprimatur. The Counter-Reformation Catholic Church extended this principle of licensing into a systematic Index of Prohibited Books, which listed titles the ecclesiastical licensers had decreed too unsound for general readers. Again, while this undoubtedly affected the availability of certain books and authors (Machiavelli and Copernicus, for example) in Catholic lands, protestant publishers quickly realized that appearing on the Index virtually guaranteed best-seller status in their own countries, and they rushed to print large impressions.

In these ways, then, the social order of early modern Europe sought to accommodate the potentially revolutionary new craft. It was never an entirely stable arrangement, and by the late seventeenth century, it was breaking down. In England, first, licensing lapsed in 1695. With it went much of the authority of the local guild, called the Stationers' Company. Nobody wanted outright deregulation, however, and the major result was a hard-won new compromise between state, authorial, and trade interests. It was articulated in the so-called 'Copyright Act' (1710 – the term copyright never actually appears in it). This act became the basis for the entire subsequent history of copyright and intellectual property law. By the late eighteenth century, agitation for reform of printing and literary property on similar lines had occurred across Europe, and in the years around 1800 broadly similar copyright provisions were established in all major Western states; they extended beyond Europe and the colonies in the nineteenth century. The modern conventions of printed authorship thus owe their origin to the existence and instability of early modern attempts to define and regulate printing: to licensing, patents, and the craft practices of printers themselves (Chartier, 1992).

Industrial Printing

The technology of printing remained virtually unchanged from soon after Gutenberg's invention until around 1800. At that time, however, change came rapidly. Not only did the legal and conventional practices governing the industry alter once and for all; so did the technology of printing itself. First came Stanhope's metal press. This was soon followed by the steam press, which at a stroke increased production rates by an order of magnitude. Stereotyping finally made the dream of truly identical copies a reality. And at the same time, the technology of papermaking changed dramatically with the success of Henry and Sealy Fourdrinier's machines in Hertfordshire. Although it never had the iconic status of the steam press, the papermaking machine had as great a role in the changes of this period, and Brunel himself hailed it as 'one of the most splendid inventions of our age.' With the advent of all these new devices occurring within a generation, between 1800 and 1840 printing underwent its own industrial revolution.

The first commercial steam printing was done at the office of The Times in London. It vastly increased the print run of the newspaper, quite literally overnight (the management had the machines installed in secrecy to prevent unrest among the soon-to-be-redundant workers, and unveiled them only when they produced their first day's issue). Nor did the improvements cease there: by 1827 The Times had new presses that were producing 5000 copies an hour. Steam printing was understandably slower to take hold in smaller publishing markets, since it required a large capital outlay – Cambridge University Press, for example, was unusually avant-garde among specialist houses in adopting the technology, yet still did not do so for more than a generation. And in general, the book publishing business was less immediately affected by industrialization than the newspaper and periodical press. But the advent of industrial technology nonetheless changed printing forever, and vastly increased its social impact.

In fact, many of the attributes that historians and social scientists attribute to the hand press may more properly be assigned to the steam press, in conjunction with mechanical papermaking and stereotyping. Typographical fixity in particular was never as clear a component of hand printing as we tend, anachronistically, to assume – in an edition of perhaps 300–1000 copies, any number of variations would occur across different copies, as correction would be done on the fly. This kind of variability now ceased. Print runs 10 or 100 times as great as had previously been viable also represented a quantitative leap as great as that associated with Gutenberg. They made possible, indeed in economic terms virtually necessitated, the first national daily press. The development of mass literacy, too, was really much more a phenomenon of the nineteenth and especially twentieth centuries than of the sixteenth and seventeenth; mass education of the kind parodied by Dickens in Hard Times relied on steam printing. At the same time, since industrial printing required a great deal more concentration of capital than had hand printing, it encouraged agglomeration

and uniformity within the publishing industry itself. The press barons of the twentieth century, with their incalculable influence on social and political life, owed their baronies to the printing techniques introduced in the nineteenth century. In effect, printing had become a branch of factory culture.

A Printing Revolution?

Gutenberg's invention had profound social and cultural consequences, and modern historians and social scientists have tried hard to discover what they were. First and foremost, the press produced texts in unprecedented quantities. Lucien Febvre and Henri-Jean Martin (1958), in the first modern work to focus on the history and impact of printing, calculated that the number of printed books soon exceeded the number ever produced in manuscript up to that time. The sheer quantity of this order had its own consequences. Books were suddenly available in unprecedented numbers, in more places, and at lower costs. As they spread through social ranks, people – especially town-dwellers – made these new objects their own. Literacy increased markedly, and that alone constitutes one of the major turning points in the development of modern social order. But mere numbers cannot explain the more specific cultural consequences of the press. Major qualitative changes were also brought about by print. These have been described most exhaustively by Elizabeth Eisenstein (1979). They were of two major kinds: transformations in the making and appearance of the page, and innovations in the uses of books by writers and readers.

After Gutenberg's brief experiment with imitation, the printed page soon began to look very different from its manuscript predecessor. Typefaces and layouts became standardized. Cross referencing, indexes, errata, and notes came into use, according to conventions that themselves grew more uniform. Woodcuts and engravings allowed the use of repeatable images, permitting diagrams, charts, and scientific illustrations to be communicated with a security never before possible (Ivins, 1953; McKenzie, 1986). Hitherto, it had been very difficult to express sophisticated claims in pictorial form, so rapidly did images degrade in the hands of copyists. The simple ability to juxtapose and compare reliable representations of competing claims revolutionized the possibilities for learned work.

This textual 'fixity,' as Eisenstein calls it, affected a number of cultural realms, among which religion and science may be singled out. It is not a coincidence that Luther's reformation lasted, while previous attempts at reform (such as those of Hus, or the Lollards) had withered or been wiped out. The permanence and cornucopic plenty of print made Luther's program impossible to suppress. Indeed, it is in the furore over Lutheranism that some of the most characteristic forms of mass communication were invented, such as the pamphlet and the polemical cartoon (Scribner, 1981). And, most fundamentally, only such a machine could truly hope to permit a 'priesthood of all believers' by making the Bible available as an object to innumerable readers.

In the sciences too the advent of print permitted great upheavals. The contemplative natural philosophy of the schoolmen gave way before a newly active interrogation of natural processes, modeled, allegedly, on the practices of craftsmen. Print not only circulated this new natural knowledge but also helped to create it, since printers' homes became 'polyglot households' where scientists of all nationalities could meet and correspond. Print thus helped forge the norm of open communication that is so central to science. And men like Henry Oldenburg secured the status of the new science in novel forms of printed communication. Oldenburg's Philosophical Transactions, launched in 1665, survived to rank today as the oldest of all scientific journals.

But printed texts alone did not change the world; printed texts plus new ways of using them did. Not least, the sheer mass of printed matter required new classification techniques – in effect, a science of bibliography – to master a worldwide 'library without walls.' But more profound was the change in reading practices that print facilitated. From its early days, Philosophical Transactions was the central plank of a new, transnational readership of naturalists, gentlemen, mathematicians, and scholars – the ancestor of today's scientific community. The common readership of such periodicals was rapidly recognized as uniting a novel kind of collective, called the republic of letters or, rather later, the public sphere. This 'public' was defined by its engagement with print. And with Baconian ambition, it took all knowledge to be its province (Habermas, 1962). It marked a decisive break from traditional and absolutist societies. In Habermas' very influential account, it therefore laid the foundations for true sociopolitical transformation, most notably in the form of the French revolution. In that sense, the 'printing revolution' ushered in by Gutenberg warranted its name.

Theories about Print

It was in the context of the Jacobins that printing was first recognized as an explicitly revolutionary force. True, Francis Bacon's salute to the press (along with gunpowder and the mariner's compass) was much cited; yet it was little more than an aphorism. The first fully articulated and contextual interpretation of printing's impact was that of the Marquis de Condorcet (1795), who was seeking to explain the overthrow of the French monarchy. Condorcet sketched out a history of 'the human spirit' in a series of stages, with the invention of the press constituting a major turning point and effectively making the deposition of the Bourbons into 'the revolution that the discovery of printing must bring about.' In composing this sequence, he became the first writer to chart a trajectory of print-based modernization that saw its effects manifested first in the sciences, and only later in general social life. That trajectory has retained its persuasive power in modern times, but its meaning has been construed in very different ways.

Heirs to Condorcet, modern historians, and social scientists have continued to concentrate their interpretative efforts on the typographical ancient régime that came to an end with the dual inventions of copyright and the steam press. Compared with this, the period of industrial printing has received comparatively slight attention. There have been many empirical histories of printing, publishing, and reading in the nineteenth and twentieth centuries, but nothing as influential in social–scientific terms as Habermas or Eisenstein. Indicative is the

reception of Habermas' work on the public sphere: most Anglophone readers neglect its second half, which tells a depressing story of the replacement of Enlightenment public culture with industrial mass culture. His story is far more elegiac than celebratory, but one would not realize this to read most of the Anglo-American commentators on *The Structural Transformation* of the Public Sphere (1962).

Where they do not ignore the later history of printing, most scholars simply conflate it with the earlier history, and speak of 'print' as though it were a single, uniform entity. Probably the most famous and notorious of construals along these lines has been that of Marshall McLuhan (1911–80), the Canadian literary critic and scholar. McLuhan's *The Gutenberg Galaxy* (1962) and *Understanding Media* (1964) set the pattern for a series of works published in the 1960s that made extravagant claims for the cultural and even psychological impact of print. His works used what he called 'probes' – unsystematic, radical aphorisms delivered scattergun fashion to jolt readers out of the formulaic thinking that, he believed, typography inevitably instilled. The probes worked, in at least one sense. McLuhan himself became a cult figure: Tom Wolfe remarked that he sounded like "the most important thinker since Newton, Darwin, Freud, Einstein, and Pavlov."

What McLuhan achieved was to make 'media' the subject of analysis in their own right. He argued that these media, of which print was the archetype, brought the world together, dissolving boundaries. The most important of these boundaries were sociopsychic in character. Indeed, McLuhan's very definition of a medium was that it acted as an 'extension of man,' transcending the material constitution of the human frame. In effect, he believed that the state of media determined not only how humans lived, but what humans were. This he presented as a matter of evolution. In particular, McLuhan portrayed what he called 'typographic man,' namely the kind of being who lived according to the cultural logic of print. Typographic man thought in terms of linear logic and objectivity, because his reference points were fixed texts, in what later became the Eisensteinian sense of fixity. Typographic man was both individualist (he could be sure who he was) and nationalist (he could see the bounds of his community, and tell the difference consistently between it and others). There had been no such being before about 1450. At that point, typographic man had begun to displace 'tribal man,' who had been restricted to local contacts and communication by hand. And in his turn typographic man was now becoming extinct, to be replaced by something else that McLuhan himself left nameless, but who could easily be called 'electronic man.'

McLuhan remarked that it was in the nature of electronic media to integrate with the nervous system itself, establishing humans as nodes on a worldwide network. In effect, the skin itself would dissolve as a social constraint, as neurology became inseparable from sociology. Written in the early 1960s, before the first computer network, it is easy to see why McLuhan has recently been rediscovered by the digerati and labeled a 'prophet of the Web.'

Yet McLuhan's arguments were ultimately too faddish in their tone, and too determinist in their content, to remain convincing beyond their bare outlines. Criticisms soon started to mount (an influential example is Jonathan Miller's short 1971 introduction, which leaves little doubt where its author stands). There are few social scientists or historians today who would openly acknowledge McLuhan as a leading influence on their representations of print and its consequences. But that influence is nonetheless real. For the most part, it is refracted through the historical analysis of Eisenstein. Eisenstein's careful arguments about the impact of the hand press gave empirical weight to McLuhan's aphorisms. Through her, his proposals have gained academic respectability. An example is Benedict Anderson's widely praised account of nationalism (1983), which rests on an explicitly Eisensteinian argument linking print to time reckoning, and thence to consciousness of national identity. Similarly once removed arguments of this kind can be found in many of the analyses of 'print culture' to have appeared in the 1980s and 1990s.

Yet there is one important respect in which current trends in sociological and historical research are starting to depart consciously from this approach. This is the move toward an empirical history of the practice of reading. The immediate origins of this trend lie in France, where the postwar enterprise of the history of the book was born. The spur came from the responses of both history and sociology to the dominant Annales school. In both fields, a realization took hold in the later 1970s that quantitative social–scientific accounting failed to capture something fundamental about the cultural impact of print. At the same time, Pierre Bourdieu (1979) was emphasizing the importance of active cultural appropriation by readers of the press rather than passive cultural reception by those readers (Hoggart (1957) had earlier made similar arguments for an English readership). By the early 1980s it looked as though attention to the varied ways of using books, building on Bourdieu's approach, might stand a chance of seeing what the quantitative Annalistes were clearly missing. Since then the history of reading has become a burgeoning field in its own right.

The leading proponents of the history of reading have been cultural historians like Roger Chartier in France and Robert Darnton in the United States, who have pioneered the reinterpretation of the French Revolution in these terms. Chartier in particular objects to the determinism about print that is implicit in the work of Habermas, and would direct attention more to changes in reading practice in the eighteenth century (Chartier, 1990). In particular, he argues that it is difficult to understand how the scatological and pornographic literature that admittedly flooded eighteenth-century France was intrinsically more devastating to the priesthood, say, than the rich seam of libels that appeared in Luther's day. Instead, Chartier would point to a newly skeptical and extensive practice of reading that was devoted to these materials, making them far more damaging in use. This practice contributed substantially to the desacralization of kingship. It hence made regicide possible, and with it the origins of modernity. In this light, the printing revolution needs to be redefined as part of a broader reading revolution (Cavallo and Chartier, 1999).

As a result, research into printing as a medium is today leaving behind quasideterminist accounts of print exerting some kind of cultural 'logic' on societies, and toward empirical and historical research into the different ways in which societies have made use of the technology and its products. Important examples include the national histories of the book now being published in Britain, the United States, and other countries (Chartier and Martin, 1982–1986; Amory

and Hall, 1999; Hellinga and Trapp, 1999). D.F. McKenzie's 1984 study of a decisive cultural encounter mediated by print in his native New Zealand shows how this modern approach can influence anthropological understandings as well as historical. In McKenzie's terms, excessively schematic talk of "print logic" is being replaced by a properly contextual "sociology of texts" (McKenzie, 1986). It is interesting to speculate whether, with print logic thus banished, the use of 'print culture' as a catch-all explanatory device may eventually fall into abeyance too. And that may happen, not coincidentally, just as the advent of electronic media signals the end of half a millennium in which print was the predominant communicative medium.

See also: Computer Mediated Communication; Literacy and Illiteracy, History of; Mass Media, Political Economy of; Media Effects; Media and History; News: General; Radio as a Medium; Television: General.

Bibliography

Amory, H., Hall, D.D. (Eds.), 1999. A history of the Book in America, Volume 1: The Colonial Book in the Atlantic World. Cambridge University Press, Cambridge, UK.

Anderson, B., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Bourdieu, P., 1979. La Distinction: Critique Sociale du Jugement. Éditions de Minuit, Paris (Also published as 1986 Distinction: A Social Critique of the Judgement of Taste. Routledge, London).

Cavallo, G., Chartier, R. (Eds.), 1999. A History of Reading in the West. Polity Press, Cambridge, UK.

Chartier, R., 1990. Les Origines Culturelles de la Révolution Franaise. Seuil, Paris (also published as The Cultural Origins of the French Revolution. Duke University Press, Durham, NC (1991).).

Chartier, R., 1992. L'Ordre des Livres: Lecteurs, Auteurs, Bibliothques en Europe entre XIVe et XVIIIe Sicle. Alinea, Paris (also published as The Order of Books. Polity, Cambridge, UK (1994).).

Chartier, R., Martin, H.-J. (Eds.), 1982–1986. Histoire de l'Edition Franaise, 4 vols. Promodis, Paris.

Condorcet, M.J.A.N., Caritat, Marquis de, 1795. Esquisse d'un Tableau Historique des Progrs de l'Esprit Humain. Agasse, Paris (also published as Sketch for a Historical Picture of the Progress of the Human Mind. Noonday Press, New York (1955).).

Darnton, R., 1982. The Literary Underground of the Old Regime. Harvard University Press, Cambridge, MA.

Eisenstein, E.L., 1979. The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe. Cambridge University Press, Cambridge, UK.

Febvre, L.P.V., Martin, H.J., 1958. L'Apparition du Livre. Michel, Paris (also published as The coming of the Book: the Impact of Printing 1450–1800. Verso, London (1984).).

Habermas, J., 1962. Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft. Luchterhand, Berlin (also published as The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. MIT Press, Cambridge, MA (1989).).

Hellinga, L., Trapp, J.B. (Eds.), 1999. The Cambridge History of the Book in Britain: 1400–1557. Cambridge University Press, Cambridge, UK; New York.

Hoggart, R., 1957. The Uses of Literacy: Aspects of Working Class Life with Special Reference to Publications and Entertainments. Chatto and Windus, London.

Ivins, W.M., 1953. Prints and Visual Communication. Harvard University Press, Cambridge, MA.

McKenzie, D.F., 1984. The sociology of a text: orality, literacy and print in early New Zealand. The Library 6 (6), 333–365.

McKenzie, D.F., 1986. Bibliography and the Sociology of Texts. British Library, London.

McLuhan, M., 1962. The Gutenberg Galaxy: The Making of Typographic Man. Routledge and Kegan Paul, London.

McLuhan, M., 1964. Understanding Media: The Extensions of Man. Routledge and Kegan Paul, London.

Miller, J., 1971. Marshall McLuhan. Viking, New York.

Scribner, R.W., 1981. For the Sake of Simple Folk: Popular Propaganda for the German Reformation. Cambridge University Press, Cambridge, UK.

Prison Life, Sociology of

Paul Rock, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12055–12059, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

A distinctive sociology of prison life emerged when the prisons of North America and Europe became newly accessible to sociological and anthropological researchers in the 1940s and 1950s. It was to be characteristically informed by the structural-functionalism that was then prevalent, dwelling in particular on the apparent functional interdependence between prisoners and staff and between formal and informal social relations that seemed to sustain stable patterns of accommodation in authoritarian, coercive institutions. Its subsequent development was shaped by a succession of qualifications and elaborations that stressed how prisons differed from one another in their scale, regime, composition, and conceptions of prisoners' rights. Research in the area eventually atrophied as prison administrators became more defensive and less inclined to encourage close observation and as penologists and criminologists became intrigued by political economy and the broad historical writings of Michel Foucault.

The sociology of prison life concentrates on thick descriptions of the actions and accounting practices that make up the distinctive social organization of a prison. Chief amongst its preoccupations have been how administrators and staff strive to maintain an authority which can never be wholly legitimate in the eyes of their captives; how, in turn, those captives contend individually and collectively with conditions of deprivation and adversity, including the adversities inflicted by their fellow inmates (Morris and Morris, 1966, pp. 168–9); how a negotiated social order can emerge out of the exchanges between the two; and how, as a result, prisons manage to persist at all.

The proper study of prison had to await the emergence of universities and other institutions that could sustain social research, deploy some form of fieldwork as part of its methodology, and appreciate that the prison is a strange institution which 'challenges the anthropological imagination' (Sparks et al., 1996, p. 33). It was not until the 1940s and beyond that those three strands converged in what are commonly held to be the first significant pieces of writing about the sociology of prison life, Clemmer's (1940) *The Prison Community* and, perhaps more important, Sykes's (1958) *The Society of Captives* and Goffman's (1961) *Asylums*. Because they set a pattern and foil for much of what was to come, it is useful to recite their principal arguments.

The Prison Community reported research conducted in the 1930s in a penitentiary for adult felons by a sociologist attached to the mental health staff. Clemmer used the methods of anthropology to study what he called the 'facts of social life,' the groupings, patterns of leadership, controls, leisure pursuits, sex, and work of prison, and to illustrate the complex relations, and the informal relations in particular, of a predatory, suspicious and atomized world of inmates banded publicly against authority but privately riddled by informers and mistrust.

The Society of Captives depicted inmates suffering 'the pains of imprisonment' and striving for the goals of better food, more rewarding work, fuller information, access to sex and pornography, security, status, and autonomy, formally denied them by an administration that purported to treat all prisoners more or less alike and certainly completely distinct from staff. It was in this sense that prison was held to engender anomic pressures that led, in their turn, to the production of illegitimate stratagems, informal social structures and a code of conduct that could satisfy a number of the inmates' aspirations. Inmates were enjoined not to interfere in one another's business, not to 'lose their head,' not to exploit one another, not to weaken, and not to trust the guards. Like many another code, it was largely exhortatory, a form of 'figurative action' rather than a reliable predictor of behavior and its accompanying social relations generated neither 'a perfect solidarity … [nor] a warring aggregate' (Sykes, 1958, p. 83). In practice, some prisoners were forced or encouraged to service the sexual demands of others; some asserted their social standing by violence, defiance or conspicuous example; some supported illegal markets in the provision of goods; and overlapping hierarchies emerged.

The ensuing social order was ostensibly clandestine and based on a rejection of the rejecters, and the staff were supposed to suppress it, but the staff confronted their own difficulties. They themselves wielded great power in principle but little in effect, they were constrained by rules and the practical exigencies and limits of control, they were outnumbered, troubled by the precariousness of their authority, and discouraged from drawing their weapons lest they be disarmed by inmates. Their reply was to replace a formal equality of treatment by a *de facto* recognition and rewarding of the stronger inmates and inmate groupings, and they did so in exchange for a devolved social order in which inmates effectively regulated one another. The consequence was what Sykes called 'functional corruption,' but it could as readily be seen in functionalist language as a drift towards system integration, equilibrium, or homeostasis, an exemplification of the doctrine that the latent functions of action can work in seemingly paradoxical fashion to promote stability in social systems.

Goffman (1961) worked as a participant–observer in a mental hospital in Washington, DC, to construct a model of asylums, prisons, and like organizations as 'people-processing'

'total institutions' that were substantially cut off from the larger world by symbolic and physical boundaries; a continuous round of activity lived publicly within the walls; a relative lack of internal differentiation of space, role, and temporal schedule; an unbridgeable divide between staff and inmates; and, through what he called a 'mortification of the self,' the discarding of old identities and the assumption of new, degraded, and formally homogeneous identities peculiar only to the institution.

Five main processes have shaped the character and development of much of the work that followed. First has been the adoption of a somewhat critical stance. It was unavoidable that criminologists, penologists, and sociologists came motivated to the study of prison life, sometimes wishing to reform prisons, sometimes to abolish them, and sometimes – more rarely – to analyze them with the dispassion that they might extend to any other social institution. But it has been a strong distaste for prisons that has colored much of what they have written, and a number have been avowedly partisan and, indeed, polemical in their work. That stance has, in its turn, led to methodological difficulties in winning access to, and trust from, a number of key groups in prison, and prison officers especially.

Work has, second, selectively incorporated interpretations of transformations occurring in the political, administrative, and demographic conditions of penal institutions. In the United States, for instance, the sociology of prison life was born in descriptions of autocratic, *laissez-faire* regimes and the resistance that was waged by conservative inmates who feared they might lose power in the face of administrative reform. It moved on through a discussion of American penology's flirtation with rehabilitation in the early 1960s and the dilemmas that arose when the particularistic ethos of treatment confronted the universalism of disciplinary control. It looked at the radical politicization of prisoners, and particularly of black prisoners, in the late 1960s and early 1970s. It analyzed the efforts at legalistic reform, the organizational instabilities, fragmentation of inmate communities, and new waves of riots of the 1970s and early 1980s. And it has culminated most recently in an analysis of the new, racialized politics of mass incarceration and incapacitation that, in Simon's phrase, is the consequence of 'governing through crime' (Simon, 1993).

Internally, American prisons were to become 'Balkanized' into antagonistic racial camps based in some measure on a prior membership of street gangs. Externally, Simon has claimed, the logic of the penal system was to be extended out from the prisons through parole, policing, and probation to subject large blocks of the population, and particularly the young, black male population of the cities of the United States, to a tight control based on actuarial and risk-driven calculations of propensities to offend. Prison life has, in effect, spread beyond the walls of the American penitentiary to the ghetto.

Third, like the sociology of almost any other substantive area, the study of prison life has tended to edit and import some of the more general and evolving theoretical preoccupations of sociology at large as they have appeared to illuminate, and be illuminated by, those structural changes. It has already been observed, for instance, that work was launched inside a context set by the structural-functionalism and anomie theory of the 1940s and 1950s, a context suited to the study of willful, autocratic repression. Current work focuses on issues of resistance and authority, gender and power, 'governmentality' and 'structuration,' and other matters now engaging a number of the wings of sociology.

Fourth, like any other method, and aided by those transformations, interpretations and theoretical preoccupations, its focus has advanced restlessly from one phenomenon to another, continually introducing new 'facts' with which to re-interpret prison life. Above all, it has produced what is in effect a running series of footnotes or qualifications to the foundational models of the 1950s and 1960s.

It was easy enough at the beginning to overlook some of the finer and more nuanced details of structure. The social anthropologist in the American penologist would have been fascinated by the inmate community's vividly phrased argot and roles. What was less often reported – because it was less obtrusive and less interesting at first sight – was the presence of those who were relatively disengaged and played no part in the clandestine life of the prison, the 'retreatists' and others: who tried to serve their time uneventfully and discreetly and who eventually became interesting precisely because of their earlier neglect. What was also partially obscured at first was the lives of those apart from the inmate. Sociologists moved away from the inmate and towards chaplains, medical officers, and others: for example, the male prison officer beset by problems of violence, and the female prison officer and her gendering of control.

Studies report not only the features that were neglected in the earliest work, but also new features that have emerged since that work was done, but much remains unexplored, for instance, the structural repercussions of the social class of prisoners, aging and the age composition of prison populations, the world of prison administration, and such groupings as gay prisoners and the prison untouchables (the 'nonces' or nonsenses of English prison argot) imprisoned for sexual and violent offences against children or the elderly.

Most important, the political control of discreditable information and the absence of a sense of anthropological curiosity and of a robust ethnographic tradition in much of the world have ensured that there have been almost no studies of prison life outside Britain, North America and a few small pockets of northern Europe. Sociologists began to emphasize how the early models had tended to extrapolate from one form of time-bound regime alone, the American, authoritarian, maximum security prison for men. It is a model that works well enough where power is arbitrary and formally unchecked, controllers are outnumbered and resistance is fragmented, and it applies broadly to German concentration camps and Soviet labor camps as well as to some of the prisons of the United States. But doubts have been raised about how far or strictly it should otherwise be generalized. It is clearly a limited model. Take Ketchum (1965), who described how in Ruhleben, an internment prison camp established in Berlin for the citizens of countries hostile to Germany in the First World War, there could be seen almost none of the informal responses to incarceration listed by Clemmer, Goffman or Sykes. To the contrary: the inmate population behaved cooperatively, peacefully, and without mutual exploitation, and its conduct invites questions about the impact of class, shame, and pre-existing patterns of organization on the social construction of prison life.

Again, take research on a number of therapeutic or rehabilitative regimes outside the United States. Sparks (1994) studied Barlinnie, a small unit for long-term prisoners in a Scottish prison, and Genders and Player (1995) examined Grendon Underwood, an English therapeutic prison to which serving prisoners in other establishments could elect to go. The three authors declared that neither Barlinnie nor Grendon was critically troubled by the legitimation crises or conflicts of other prisons, but both institutions did benefit from being regimes within regimes – inmates could always be transferred back to the larger and more coercive system from which they had come – and it would be wrong to assume that therapy alone is a proof against disorder or confrontation. So it was that Mathieson's (1965) *Defences of the Weak* described a medium-secure Norwegian prison, organized on therapeutic lines, whose disunited and cynical inmates challenged those who guarded them in a language of 'censoriousness' that turned therapeutic ideology into a medium of criticism, defense, and attack. Theirs was a response very dissimilar to that of other models.

A second, allied form of qualification to the foundational model attacked the proposition that prisons are total institutions which sever biographies and create new men and women within an indigenous culture (Sparks et al., 1996, p. 44). Instead of the fractured identities described by Goffman, study after study has revealed that inmates actually bring with them beliefs, associations, and styles of behavior from the outer world.

The importation into prison of social forms from the street gangs and antagonistic racial politics of America has already been noted, and there are other borrowings that have presumably been colored by local and national traits in the world outside the walls. Thus, Irwin and Cressey (1962) talked in America of *thieves* who carried a wider criminal culture into prisons, treating imprisonment as a career contingency; *convicts* whose behavior was a localized adaptation to the special travails of prison life; and *square Johns* who had an allegiance to neither convict nor thief but to a show of obedience to authority. Also, Cohen and Taylor (1972) talked about the affinities between the manner in which long-term inmates in Durham Prison in England managed problems of serving time and the offenses for which they had been convicted, armed robbers worrying about physical deterioration and concentrating on body-building exercises, and fraudsters engaging in copious attempts at litigation.

The discovery of the gendered character of imprisonment is as revealing. The very first studies dwelt on institutions for men. Whilst it was recognized from the outset that male inmates confronted largely male problems stemming from a loss of liberty, status, and heterosexual contact, gender itself was not treated as especially problematic or interesting. It was only when ideas about sexual identity were applied in explanations of the behavior of women inmates in the 1960s, women who were said to form themselves into pseudo-families in response to a loss of emotional sustenance, that the influence of gender itself became notable. Other sociologists have subsequently exextended and modified that early work, but the thesis that women combine together differently remains. Later, ideas about masculinity were themselves turned back on the social organization of men's prisons to translate them into places ordered by a 'hegemonic masculinity.' Male prisons, it was argued, must be understood as institutions that are dominated by a very masculine exercise of violence, reliance on hierarchy, ritualized combativeness, and the suppression of displays of emotion and vulnerability.

Last, it must be stressed how closely, although never absolutely, the study of prison life has been bound up with the use of ethnographic methods (Terence and Pauline Morris said of their base in Pentonville Prison, London, that it was 'like the anthropologist's hut on the village street' (Morris and Morris, 1966, p. 8)). Its history has, by extension, roughly paralleled the rise and fall of sociological ethnography itself. It was probably at its most vigorous between the 1950s and early 1970s, when, according to Janowitz, it 'contain[ed] some of the most outstanding research monographs in sociology' (Foreword to Jacobs, 1977, p. ix). Its decline may be linked in part to the seemingly inevitable half-life enjoyed by any body of theory and research, and to problems of access and publication that began to surface in the 1970s when prisons were turbulent and their managers became politically defensive and defeatist. But it may also be attributed to the ascent to influence of Michel Foucault (1979), whose political economy of discipline prompted scholars to turn *en bloc* to reflections about what they called 'penality' in a markedly more abstract and empirically disengaged manner.

Foucault took Bentham's plans for a concentric, panopticon prison to be an especially telling representation of the new technologies of control that were emerging in society at large. Prisoners in the brightly lit rim could never know whether they were being observed by guards in the shadowy inspection tower at the center, and the result was supposed to be that they would discipline themselves, colluding in their own subordination in a decentralized, self-sustaining, economical and dispersed system of power. The panopticon was never finally constructed in England, and Foucault's history was flawed, but so seductive and wide-reaching was his metaphor that it became an archetype for a large new family of ideas, and sociologists came to adopt Foucault rather than Clemmer, Sykes or Goffman as their mentor. Jonathan Simon (1999, p. 4) observed that, 'the great tradition of studies of prison social organization that begins with Clemmer and Sykes … has largely ceased in the last twenty years … Much of the research now is conducted from afar [and] … the new distance introduces important interpretive problems … .'

Editor's Note

The study of life behind bars has continued to evolve since the original publication of this article, of course. A flavor for recent developments from a variety of methodological, disciplinary and geographical perspectives can be obtained from the following: Alison Liebling et al. (2012), Dominique Moran (2013), Yvonne Jewkes (2012), Meredith Huey Dye et al. (2014), Crewe (2011), and Rod Earle and Coretta Phillips (2012).

See also: Control, Social; Courts and Adjudication; Crime and Class; Crime, Race and Ethnicity; Crime, Sociology of; Criminal Justice, Sociology of; Delinquency, Sociology of; Foucault,

Michel (1926–84); Goffman, Erving (1922–1982); Imprisonment: Sociological Aspects; Law: Economics of Its Public Enforcement; Pestalozzi, Johann Heinrich (1746–1827); Police: A Sociology of Knowledge Approach; Prisons and Imprisonment; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Rehabilitation of the Offender.

Bibliography

Clemmer, D., 1940. The Prison Community. Holt, Rinehart and Winston, New York.

Cohen, S., Taylor, L., 1972. Psychological Survival. Penguin, Harmondsworth, UK.

Crewe, Ben, 2011. Soft power in prison: implications for staff–prisoner relationships, liberty and legitimacy. European Journal of Criminology 8 (6), 455–468.

Dye, Meredith Huey, Aday, Ronald H., Farnet, Lori, Jordan, Raley, 2014. "The Rock I Cling To:" religious engagement in the lives of life-sentenced women. The Prison Journal. Available online prior to publication: http://dx.doi.org/10.1177/0032885514537605.

Earle, R., Phillips, C., 2012. Digesting men? Ethnicity, gender and food: Perspectives from a 'prison ethnography.' Theoretical Criminology 16 (2), 141–156.

Foucault, M., 1979. Discipline and Punish. Penguin, London.

Genders, E., Player, E., 1995. Grendon: A Study of a Therapeutic Prison. Oxford University Press, Oxford, UK.

Goffman, E., 1961. Asylums. Doubleday Anchor, Garden City, NY.

Irwin, J., Cressey, D., 1962. Thieves, convicts and the inmate culture. Social Problems 10, 145–155.

Jacobs, J., 1977. Stateville: The Penitentiary in Mass Society. University of Chicago Press, Chicago.

Jewkes, Y., 2012. Autoethnography and emotion as intellectual resources: doing prison research differently. Qualitative Inquiry 18 (1), 63–75.

Ketchum, J., 1965. Ruhleben: A Prison Camp Society. Oxford University Press, Oxford, UK.

Liebling, A., Hulley, S., Crewe, B., 2012. Conceptualizing and measuring the quality of prison life. In: Gadd, David, Karstedt, Susanne, Messner, Steven F. (Eds.), The Sage Handbook of Criminological Research Methods. Sage, Thousand Oaks, CA, pp. 358–372 (Chapter 24).

Mathieson, T., 1965. Defences of the Weak. Tavistock, London.

Moran, D., 2013. Between outside and inside? Prison visiting rooms as liminal carceral spaces. GeoJournal 78 (2), 339–351.

Morris, T., Morris, P., 1966. Pentonville: A Sociological Study of an English Prison. Routledge, London.

Simon, J., 1993. Poor Discipline: Parole and the Social Control of the Underclass. University of Chicago Press, Chicago.

Simon, J., 1999. The 'society of captives' in the era of hyper-incarceration. Unpublished paper.

Sparks, R., Bottoms, A., Hay, W., 1996. Prisons and the Problem of Order. Clarendon Press, Oxford.

Sparks, R., 1994. The Barlinnie Special Unit as prison and escape. Prison Service Journal 95, 2–6.

Sykes, G., 1958. The Society of Captives. Princeton University Press, Princeton, NJ.

Prison Life, Sociology of: Recent Perspectives from the United Kingdom

Deborah H Drake, The Open University, Milton Keynes, UK
Sacha Darke, University of Westminster, London, UK
Rod Earle, The Open University, Milton Keynes, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prison life both fascinates and repels. As with many aspects of punishment, it attracts the interest of both academics and the general public. In this short and accessible article, the principal issues of prison life are presented in a historical context that traces the emergence of focused academic study of the way people live, and die, in prison. The most influential theoretical perspectives are clearly set out alongside a discussion of their influence on research and analysis in the UK and beyond. Questions of women's experience and that of black and minority ethnic prisoners are explored before a consideration of postcolonial prison studies is introduced. These studies of prison life beyond the axis of Europe and North America challenge some of the accumulated academic wisdom of Anglophone and European studies of prison life, indicating the potential of novel developments to come in an era which, unfortunately, shows no signs of declining to produce more and more prisons.

The sociology of prison life, briefly defined, encompasses the systematic study of prisoner societies, prison staff culture, and/or prisons as organizations. With continued and increasing uses of the prison as a criminal justice disposal, academic interest in prison life has been sustained and has grown across many countries across the globe. Contemporary studies of prison life form a diverse body of work that variously examine the prison world through concepts of space, place, architecture, gender, ethnicity, law, political economy, and national and global governance. Contemporary developments within the sociology of prison life, however, continue to build upon and extend a knowledge base that has developed cumulatively since 1940 and thus the first academic sociological study of prison life.

Foundations

In *The Prison Community* (1940), Donald Clemmer presented the first sociological study of prison life. Based on his research in the Menard Branch of the Illinois State Penitentiary, Clemmer put forward and established the notion of 'prisonization.' This concept can be understood as the impact of the prison experience on prisoners, and is thus a descriptor of the process by which those who enter prison take on the "folkways, mores, customs, and general culture of the penitentiary" (1940: 299). Clemmer argued that prisonization may occur to greater or lesser degrees depending on a range of factors (length of sentence, level of association with peer groups within prison, strength of ties outside prison, inter alia). Some of the factors associated with the process result from the peculiarities and constraints of the prison environment, including the acceptance (or resistance) to the powerless position of being a prisoner and the conditions under which prisoners must eat, dress, work, sleep, and so forth. However, Clemmer suggested that relational factors also influenced the process. Chief among them was the extent to which a prisoner integrated into a primary group among the prisoner community. Belonging to a primary group required adherence to an 'inmate code' or system of norms that demanded loyalty to the prisoner group and opposition to staff. The result of the process, according to Clemmer, was a (further) rejection of societal norms and a continued or deepened commitment to a value system outside the laws and conventions of mainstream society. Clemmer's findings and theories about prison life and prisoner experience dominated academic understandings and thinking about the prison for over 20 years and, to some extent, his ideas remain influential even today.

In the late 1950s and early 1960s, further studies of the prison began to emerge. Three further early sociological studies of the prison are worthy of note. In 1958, Gresham Sykes published *The Society of Captives: A Study of a Maximum Security Prison*, based on research in the New Jersey State prison. In this work and in Sykes and Messinger (1960), aspects of the 'inmate code' and its influence on prison social life and prisoner experience are examined. Sykes and Messinger argued that Clemmer's theory of prisonization was incomplete because it did not explain how or why prisoner culture formed in the first place. Sykes argued that the prison environment exerted pressures or 'pains' that altered the relational parameters between prisoners. Thus, prisoners needed to adapt to various demands of the environment and their fellow inmates in order to endure the ordeal of prison life. The pains of imprisonment, Sykes argued, included several deprivations: liberty, goods and services, heterosexual relationships, autonomy, and security. These deprivations, in turn, exercised a destabilizing influence that could result in competing prisoner hierarchies, clandestine activities and illegal markets, and a variety of other informal structures and *sub rosa* codes of conduct.

Around the same time as Sykes' study of the prisoner society, another alternative consideration of the experience of institutional confinement was put forward. Drawing on a participant-observation study of a mental hospital in Washington, DC, Chicago School sociologist, Erving Goffman published *Asylums* in 1961. Goffman coined the term 'total institution,' which he defined as institutions that have an

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45035-X

"encompassing or total character, [as] symbolised by the barrier to social intercourse with the outside … that is often built right into the physical plant, such as locked doors, high walls, barbed wire … The central feature of total institutions can be described as a breakdown of the barriers ordinarily separating [the] three spheres of life [sleep, play and work]" (1961: 16–18). Goffman utilized a symbolic interactionist (or interpretivist) approach to consider the experience of confinement and argued that through the inherently artificial and degrading environment of the total institution, inmates experience a 'mortification of the self.' Their previous identities become disrupted or jettisoned, and new institutional identities take shape. However, these new identities form in response and are peculiar to the particular institution in which the inmate is confined. The struggle to cope with and maintain a sense of identity in the face of staff and institutional pressures to erode the prisoner's sense of self were identified by Goffman as the main interactional dynamics that characterize total institutions.

In 1962, John Irwin and Donald Cressey published an important theoretical article entitled: *Thieves, Convicts and the Inmate Culture*. In this paper, they argued that too much emphasis had been placed on the influence of the prison environment on the development of inmate culture, and that not enough attention had been focused on the external cultural influences and behavioral patterns of prisoners. Irwin and Cressey argued that prisoners 'imported' certain subcultural values, beliefs, and conventions into the prison world that had held meaning for them outside. Irwin and Cressey's paper was the first to consider the social roles that prisoners occupied prior to their imprisonment and the extent to which this 'latent culture' influenced in-prison culture. They suggested that the prisoner community corresponded directly with current and changing societal trends and that prisoner societies were more chaotic, diverse, and less rule-bound than previous studies had suggested.

Dominant Theoretical Models

The main theoretical models on prison social life that have dominated and, to some extent, continue to feature in the research literature have their origins in the above mentioned foundational studies. Thus, theories on the sociology of prison life tend to be based on either a functionalist model, an 'importation' model, or some integrated version of these two broad approaches.

Functionalist models take a problem-solving view to account for prisoner culture. From a functionalist perspective, it is thought that prisoner culture develops as it does *in response or as an adaptation to* the frustrations and deprivations of imprisonment. The purest branch of the functionalist model argues that any differences in the organization of prison establishments are irrelevant because the experience of total institutions is so coercive that it homogenizes prisoner responses to their environment (see Clemmer, 1940; Sykes, 1958; Goffman, 1961). By contrast, a situational functionalist model argues that prisoners' responses to prison life are situationally contingent on institutional characteristics (Grusky, 1959; Wilson, 1968). Broadly speaking, functionalist perspectives are known in the prison literature as 'deprivation' models because they seek to explain prison life and prisoners' adaptations to it as responses to the pains of imprisonment and the deprivations associated with the loss of liberty.

In response and in contrast to functionalist perspectives, explanations that favor an importation model argue for the importance of preprison social roles in shaping prisoner culture. Importation models are underpinned by a 'negative selection' view of the prisoner group. That is, involvement in criminal activity is the one common variable among prisoners, which is indicative of varying degrees of opposition to conventional norms. Thus, the social environment of prison life is shaped not only by the constraints of the prison environment, but also by the imported value systems, customs, and the various subcultures to which prisoners belong prior to their entry into prison.

Further studies of prison life have proposed slight deviations from these models. For example, Jacobs (1977), DiIulio (1987), and Useem and Kimball (1989) identified the importance of administrative approaches, organization, and control on influencing the prisoner society. However, much contemporary prison literature tends to present integrated versions of prison social life, which draw on aspects of both importation and deprivation models (Tittle, 1972; Liebling and Arnold, 2004; Crewe, 2009).

Mainstream Studies of Prisons and Prison Life in the UK

Subsequent to the above foundational studies, a number of in-depth examinations of the prison have been undertaken in a variety of penal setting, many emerging from within the UK (for example, inter alia Morris and Morris, 1963; King and Elliott, 1977; Carlen, 1983; Sparks et al., 1996; Liebling and Arnold, 2004; Crewe, 2009). The first sociological study of prison life in Britain was conducted by Terence and Pauline Morris in Pentonville prison in 1963. Their research provided a great deal of descriptive, ethnographic detail about Pentonville and drew comparisons to the work of Clemmer and Sykes. However, among prison administrators, the work was received as a polemical reading of the prison and was thus viewed as a somewhat controversial contribution to the prison literature.

Since the 1960s, a steady stream of sociological commentary, research, and scholarship on prisons in the UK has emerged from the Institute of Criminology, University of Cambridge. Initially, the relationship between the Institute of Criminology and prison administrators was forged by the first director of the Institute, Sir Leon Radzinowicz. Radzinowicz acted as a frequent advisor to senior civil servants and prison administrators, and also sat on the Advisory Council on the Penal System (1966–78). Through this council, Radzinowicz advised on a number of matters associated with penal and prison policy, including chairing the committee, which set up the system of maximum-security prisons in England and Wales (known as the 'dispersal' prison system, see Radzinowicz, 1998).

The Institute of Criminology has continued to work closely with prison administrators in the UK and elsewhere, specializing in sociological studies of the prison. In 1996, Richard Sparks, Anthony Bottoms, and Will Hay published *Prisons and*

the Problem of Order. This book, based on ethnographic research in two maximum-security prisons (Albany and Long Lartin) in the late 1980s, considered the factors that influenced the orderliness (or otherwise) of a prison regime. Sparks et al. argued for the importance of legitimacy in the administration of prisons. Although a significant sociological study of prison order and administration, this work is better described as an examination of prisons as organizations, rather than a study of prison life per se. Thus it might be differentiated from other works in prison sociology, which focused more directly on prisoner societies and social roles.

In 2000, the Prisons Research Centre at the Institute of Criminology, University of Cambridge, was established under the directorship of Professor Alison Liebling. Liebling has established a formidable reputation as an authority on prison quality of life, having conducted several large studies on various aspects of prison life, including prisoner suicide, self-harm, and safer custody (1992; Liebling et al., 2005), staff–prisoner relations (2011), the 'moral performance of prisons' (Liebling and Arnold, 2004), inter alia. Liebling's work on 'what matters' in prisons has been influential in shaping the way prison administrators measure and conceptualize prison quality. Liebling's concept of 'moral performance' draws administrative attention to several aspects of humane custody, such as respect, trust, safety, dignity, and well-being. In addition to Liebling's own work, a steady stream of researchers and doctoral students working at the Prisons Research Centre have produced numerous sociological and psychosocial studies of prison life, prisons as organizations, and prison administration (e.g., Bosworth, 1999; Crewe, 2009; Drake, 2012).

Consideration of Differing Perspectives

One of the main difficulties faced by sociological prison researchers and theorists concerned with examining the inner life of prisons has been untangling the complex relationships that influence it. As with society in general, there is an extreme complexity of attitudes, opinions, beliefs, and values of any given prison population. Although many studies of the prison have ignored the diversity of prisoner experiences and the heterogeneity of prisoners' viewpoints on their own lives and their reflections on the experience of imprisonment, there have been some valuable contributions that have applied interpretive sociological theory to ethnographic or interview data. These works have illustrated the connections between human agency, intersubjective meanings, and environmental or structural conditions of prison life (see Goffman, 1961; Irwin and Cressey, 1962; Cohen and Taylor, 1972; Bosworth, 1999; Jones and Schmidt, 2000; Rowe, 2011). There are, however, a range of structural positions or standpoints from which prison life might be considered, including the perspective of women prisoners, LGBT prisoners, black and binority ethnic prisoners, or civilian versus disciplined prison staff.

Sociology of Prison Life in Women's Prisons

Like many areas of sociology, understandings of the human experience of prison life were initially androcentric, i.e., established from the standpoint of men's experiences. Though there are some very early examinations of women's imprisonment from a sociological perspective (see for example Selling, 1931), the first large-scale studies of women's imprisonment were conducted in the mid-1960s by David Ward and Gene Kassebaum (1965) in California and Rose Giallombardo (1966) in West Virginia (but see also Heffernan, 1972). These early studies did not fully examine women's prisons or women's experiences of imprisonment as separate areas of inquiry from men's prisons. Instead, the findings from studies of prisoner societies made up of women were compared and contrasted with previous findings from studies of men in prison. Ward and Kassemaum, for example, commented on how women's experiences of imprisonment differed noticeably from men's, that there was not a strong 'inmate code' among women. They argued that women prisoners tended to form 'families' or 'cliques' through which contraband might be distributed and shared, rather than the entrepreneurial, black market that characterized men's imprisonment. These early researchers argued that women's gender roles were imported into the prison with them, while the deprivations and context of institutional life shaped and dictated the way these roles were fulfilled. As Candace Kruttschnitt and Rosemary Gartner (2005: 22) have noted "Important sources of variation in adjustment – institutional characteristics and characteristics of the prisoner population – were left largely unexplored."

A great deal of the sociology of women's imprisonment has focused on women's intimate relationships in prison (Norris, 1974; Propper, 1981; Owen, 1998). However, more recent developments in the sociological literature on prison life in women's prisons has moved on to consider a range of other aspects of women's imprisonment, including the way women are constituted in penal discourse and are subjected to a gendered form of control (Carlen, 1983); how experiences of imprisonment are shaped by characteristics of the institution (Rock, 1996); the influence of macro-sociological and political factors on women's experiences of imprisonment (Kruttschnitt and Gartner, 2005); and the way women construct identities, engage in resistance, and negotiate power in prison (Bosworth, 1999; Rowe, 2011).

Convict Criminology

Much insight has been gained into prison experiences from 'insider' perspectives, including autobiographical and fictionalized accounts from authors with experience of incarceration (e.g., Serge, 1930; Davis, 1971). In particular, the work of 'convict criminologists' (those who have experienced prison life first hand, as prisoners, and who now have academic careers) has provided the sociology of prison life with vivid and reflective accounts of prison life, and studies of prison that could not have been provided from a conventional perspective. Convict criminology is a branch of critical criminology started in the late 1990s by academics in the USA who were prisoners, exprisoners, or exprison staff. It is concerned with challenging existing definitions of crime, the overuse of prison as a penal sanction, and exposing the shortcomings of the criminal justice system. Convict criminology emerged in the USA as a response to the continued failure of mainstream criminologists and criminal justice practitioners to appreciate the perspective of the convict. This failure, according to convict criminology, is

evidenced in the persistent representation of prisoners and exprisoners as distinctive, 'inferior' types of human being. These perceptions foster a damaging indifference toward a proliferating and self-sustaining criminal justice industry that is both inhumane and unnecessary (see in the USA, Irwin, 1970; Jones et al., 2009; in the UK Aresti 2012; Earle, 2011).

Black and Minority Ethnic Prisoners

That prisons tend to gather the kinds of people that Loic Wacquant (2001) describes as 'the wretched of the city' is no particular revelation to anyone with even the most passing familiarity with who gets sent to prison. However, any visitor to a US or British prison will be quickly confronted by visual evidence that this is a dark-skinned population. Much more than in the general population, prisoners tend to be drawn from black and minority ethnic groups. White people are still a majority in these country's prisons, but their chances of being incarcerated are far less than their black or minority ethnic counterparts. In the USA, analysts such as Angela Davis (2003) and Michelle Alexander (2010) identify the continuities of American racism that translate from the days of slavery, the plantation system, through the legal machinations of 'Jim Crow' to the current 'prison-industrial' complex.

Loic Wacquant (2001) recognizes the interplay of race and class dynamics in the declining American cityscapes as generating 'a deadly symbiosis' in which black ghettos feed the prisons, who in turn recycle their damaged black populations back into the decrepit housing and grinding poverty that produced them in the first place. Controversially, Wacquant describes this relationship as entirely functional to a largely white neoliberal political class as it serves to separate and distinguish a black 'underclass,' reproducing white anxiety and simultaneously offering containment of its racialized threat.

Few other white commentators provide such an urgent or powerful analysis of the racial configurations of the US penal nightmare. His analysis is rich in detail, broad in sweep, and grand in gesture. However critics (Newburn, 2010) find a tendency to erase differences and variation within the US carceral complex that do not align so closely with the racialized dynamics he identifies.

The variable and shifting dynamics of race, the state, and prison on the opposite side of the Atlantic are examined in a detailed British empirical study (Phillips, 2012). Coretta Phillips' study is invaluable in analyzing the specific configurations of 'the multicultural prison.' Phillips finds 'enduring trialities' in the combination of race, class, and gender that provide the two prisons she studied with a diverse population. Attending to the detailed stories gathered through sustained ethnographic immersion, she provides a nuanced account that engages as fully with the 'pains of racism' as it does with the pains of incarceration. In place of simplistic dyads of race, Phillips (2012: 203) presents 'the contorted inflections of racism.' Amid the continuing evidence of systemic, institutionalized racism that selects and sieves the black and minority ethnic population through every stage of criminal justice procedure from arrest through to incarceration, Phillips also identifies changing patterns of resistance, reconciliation, resentment, and racialization. The multicultural prison reflects multicultural realities in being both convivial and conflicted.

The global diversity of prison has distinctive patterns of racialization, but the sheer scale and longevity of the disproportionality in the US and British prison systems demand focused attention. Determining the precise mechanisms that produce such features of prison populations is a complex task and neglecting to do so seems to be implicated in their endurance.

Prison Studies Grounded in Sociological Critique

Considerations of prison life that have focused heavily on either administrative concerns or the factors that account for differing prisoner societies have tended to avoid questions associated with the broader social meaning of imprisonment as a form of punishment or the extent to which prisoner experiences serve the assumed purposes of imprisonment. However, there is much sociological and critical criminological literature that problematizes the use, operation, and proliferation of prisons. In 1972, Stanley Cohen and Laurie Taylor published *Psychological Survival: The Experience of Long Term Imprisonment*. This book considers the lived experience of a maximum-security wing of a prison in the North of England. It exposes the personal and psychological struggles of prisoners attempting to cope with isolation, heavy control measures, and long-term prison sentences. The book is one of the first to offer an account of prison life from the standpoint of prisoners. Cohen and Taylor achieved a degree of authenticity that had eluded previous prison researchers by developing their ideas for the book in collaboration with the prisoners who they met through a teaching-based relationship, rather than during a formalized research project per se. The book continues to influence critical prison scholars due to its poignancy in capturing and analyzing aspects of the experience of long-term imprisonment.

A number of critical studies and writings about prison life have aimed to draw attention to the chronic failure of prisons or to question the extent to which experiences of imprisonment fulfill the supposed or official purposes of prisons as a social institution. For example, British prisons in the 1970s and 1980s were troubled by prisoner riots and disturbances, thus critical scholars were particularly concerned by these problems. In 1991, Phil Scraton, Joe Sim, and Paula Skidmore published *Prisons Under Protest* (1991), which examined the problem of prison protests through the study of Peterhead prison in Scotland. The book argued, unequivocally, that any appearance of order in prisons was purely the result of enforced compliance or coercion and that "Life in most British prisons is an unrelenting imposition of authority" (1991: 62). Moreover, one of the most influential critical studies of the prison is found in Thomas Mathiesen's *Prison on Trial: A Critical Assessment* (1990). In this volume, Mathiesen systematically outlines the many ways in which prisons fail to fulfill their supposed purposes (e.g., deterrence, preventing reoffending, etc). He declares the prison a fiasco, but exposes some of the underlying structural and political economic forces that perpetuate its use.

Similarly, radical sociological writing and activism against the use of the prison has a long history in the USA. Angela Y. Davis has perhaps been the most vocal opponent of the use of the prison in the USA and the expansion of the 'prison industrial complex.' The prison industrial complex is a term

grounded in political economic analysis of the prison system represents the idea that the expansion of the prison population is attributable to the political and economic influence of both private prison companies and the various ancillary agencies and corporations that supply goods and services to maintain the prison system. Davis has thus drawn academic and public attention to some of the structural social factors that influence the use of the prison (and, in turn, prison life), including the problem of the disproportionate use of imprisonment for people of color.

Prison Life Beyond the North

We have seen that the dominant theoretical models in the sociology of prison life literature were developed mostly in the light of situated practices in just two countries, the US and the UK. With the expansion of the use of the prison around the world, however, historical, ethnographic, and first-hand examinations of prison life outside of Anglo and Northern societies have recently begun to emerge, much in the English language (see, inter alia, Aguirre, 2005; Bandyopadhyay, 2010; Dikötter and Brown, 2007; Piacentini, 2004; Focaal, 2014; Miller and Campbell, 2014; Prison Service Journal, 2014; Salvatore and Aguirre, 1996). An essential starting point for this literature is that Southern prisons are poorly resourced in comparison to the prisons of Western Europe and North America. Important for current purposes, Southern prisons are not only more overcrowded and austere than prisons in the US and the UK, for example, but also they are often characterized by acute staff shortages. In Brazil, for instance, it is quite normal to find just one guard on duty per 100, even 200 inmates (Darke, 2013). Under these conditions, officers depend on inmates to collaborate in the administration of everyday prison routines, for instance, working as office clerks, medical auxiliaries, cleaners, cooks, and porters, and on inmates' families and voluntary sector groups to provide essential goods and services such as food, toiletries, medicines, and doctors. Equally significant, in many countries prison officers only enter the wings to lock and unlock cells. As a result, Southern prisoners are largely left to their own devices, not only to organize everyday routines on the wings, but also to maintain discipline (Darke and Karam, forthcoming; Garces et al., 2013).

This recent vibrancy in academic engagement with prisons in the postcolonial world raises significant questions over the extent to which the dominant theoretical models in the sociology of prison life literature (regarding prisonization, mortification, latent culture, and prisons as organizations) that we outlined earlier in this article are applicable beyond the global North. For instance, despite the obvious material deprivations and inability of officers to provide security, researchers have noted that Southern prisons are not necessarily as disorderly or inmate hierarchies as competing and predatory as our theories might predict. To make sense of this apparent contradiction, studies of Southern prison have focused on the ways in which prison staff and inmates may respond to the inhumane conditions in which they find themselves living and working through negotiation rather than conflict. Mahuya Bandyopadhyay (2010) explains that bad prison conditions give rise to social bonds as well as social tensions. When deprivation becomes a shared experience, as it is in prisons across the global South, it appears that humans have a natural capacity for solidarity and mutual aid.

Finally, and closely related, the impoverished nature of postcolonial prisons leads to a need to reconsider the universal applicability of Goffman's (1961) concept of the total institution. The first point that needs to be made here is that the barriers between prisons and communities are generally more permeable than in the West. For instance, prisoners typically have greater contact with their families, who we have noted, are often relied upon to make up for shortfalls in state provision. Further, with the growing phenomena of gang culture in the guardless prison wings and the likewise inadequately policed ghettoes of Latin America and other parts of the developing world (Hinton and Newburn, 2008), prison and community life are becoming increasingly linked. For instance, in large parts of Brazil, prisoners are allocated to particular prisons and prison wings according to the gang in control of the area that they come from. Involvement in gang activity itself is not a requirement. Outside the prison, gangs provide material support such as food parcels to prisoners' families. Gang leaders arriving in prison are likely to be fast-tracked as inmate leaders; imprisoned gang leaders are likely to return to positions of authority in the community upon release.

The work of prison researchers beyond the West casts new light on the sociology of imprisonment. The enduring thematic issues associated with prison life, such as order, security, control, and staff–prisoner relationships, take on a different character, offering new insights when examined in differing cultural, social, and political contexts. Christie (2004) has argued that prisons expose key aspects of the machinery of the states that they represent. Thus, by considering prisons from different social and political contexts alongside one another, there is much understanding to be gained both about practices of imprisonment and the constitution of state powers. As the use of the prison continues to proliferate around the world, findings and observations from sociological prison studies from all corners of the globe will form a vital part of the continuing development of our understanding of the sociology of prison life.

Bibliography

Aguirre, Carlos, 2005. The Criminals of Lima and Their Worlds: The Prison Experience, 1850–1935. Duke University Press, Durham.

Alexander, Michelle, 2010. The New Jim Crow – Mass Incarceration in the Age of Colorblindness. The New Press, New York.

Aresti, Andy, 2012. Developing a convict criminology group in the UK. Journal of Prisoners on Prison 21 (1 & 2).

Bandyopadhyay, Mahuya, 2010. Everyday Life in a Prison: Confinement, Surveillance, Resistance. Orient Blackswan, New Delhi.

Bosworth, Mary, 1999. Engendering Resistance: Agency and Power in Women's Prisons. Ashgate Dartmouth, Brookfeild, VT.

Carlen, Pat, 1983. Women's Imprisonment: A Study in Social Control. Routledge & Kegan Paul, London.

Christie, N., 2004. A Suitable Amount of Crime. Routledge, London.

Clemmer, Donald, 1940. The Prison Community. Holt, Rinehart and Winston, New York.

Cohen, Stanley, Taylor, Laurie, 1972. Psychological Survival: The Experience of Long-Term Imprisonment. Penguin, Harmondsworth.

Crewe, Ben, 2009. The Prisoner Society: Power, Adaptation and Social Life in an English Prison. Clarendon Press, Oxford.

Darke, Sacha, 2013. Inmate governance in Brazilian prisons. Howard Journal of Criminal Justice 52, 272–284.
Darke, Sacha, Karam, Maria L. Latin American prisons. In: Jewkes, Y., et al. (Eds.), Handbook on Prisons. Routledge, Abington, forthcoming.
Davis, Angela Y., 1971. If They Come in the Morning. Orbach and Chambers Ltd., London.
Davis, Angela Y., 2003. Are Prisons Obsolete? Seven Stories Press, New York.
Dikötter, Frank, Brown, Ian (Eds.), 2007. Cultures of Confinement: A History of the Prison in Africa, Asia and Latin America. Cornell University Press, Ithaca, NY.
Dilulio, John J., 1987. Governing Prisons: A Comparative Study of Correctional Management. Collier Macmillan, London.
Drake, Deborah H., 2012. Prisons, Punishment and the Pursuit of Security. Palgrave, Basingstoke.
Earle, Rod, 2011. Prison and university: a tale of two institutions. In: British Society of Criminology 2010 Conference Proceedings, vol. 11, pp. 20–37. Available at: http://www.convictcriminology.org/bcc.htm.
Everyday prison governance in Africa. Prison Service Journal 202, 2014.
Focaal, 2014. Prison climates in the global South 68.
Garces, Chris, Martin, Tomas, Darke, Sacha, 2013. Informal prison dynamics in Africa and Latin America. Criminal Justice Matters 91 (1), 26–27.
Giallombardo, Rose, 1966. Society of Women: A Study of a Women's Prison. Wiley, New York.
Goffman, Erving, 1961. On the characteristics of total institutions. In: Cressey, D. (Ed.), The Prison: Studies in Institutional Organization and Change. Holt, Rinehart and Winston, New York, pp. 15–67.
Grusky, Oscar, July 1959. Organisational goals and the behaviour of informal leaders. American Journal of Sociology 65, 59–67.
Heffernan, Esther, 1972. Making It in Prison: The Square, the Cool and the Life. Wiley-Interscience, New York.
Hinton, Mercedes S., Newburn, Tim (Eds.), 2008. Policing Developing Democracies. Routledge, London.
Irwin, John, 1970. The Felon. Prentice-Hall, Englewood Cliffs.
Irwin, John, Cressey, Donald R., 1962. Thieves, convicts and the inmate culture. Social Problems 10, 145–147.
Jacobs, James B., 1977. Stateville: The Penitentiary in Mass Society. University of Chicago Press, Chicago.
Jones, Richard, S., Ross, Jeffrey Ian, Richards, Stephen C., Murphy, Daniel S., 2009. The first dime: a decade of convict criminology. Prison Journal 89 (2), 151–171.
Jones, Richard S., Schmidt, Thomas J. (Eds.), 2000. Doing Time: Prison Experience and Identity Among First-Time Inmates. Jai Press, Stamford.
King, Roy D., Elliott, Kenneth W., 1977. Albany: Birth of a Prison. Routledge & Kegan Paul, London.
Kruttschnitt, Candace, Gartner, Rosemary, 2005. Marking Time in the Golden State: Women's Imprisonment in California. Cambridge University Press, Cambridge.
Liebling, Alison, 1992. Suicides in Prisons. Routledge, London.
Liebling, Alison (Assisted by Arnold, H.), 2004. Prisons and Their Moral Performance. Clarendon Press, Oxford.
Liebling, Alison, Tait, Sarah, Durie, Linda, Stiles, Annick, Harvey, Joel (Assisted by Rose, Gerry), 2005. An Evaluation of the Safer Locals Programme. Institute of Criminology. University of Cambridge, Cambridge.
Liebling, Alison, Price, David, Shefer, Guy, 2011. The Prison Officer, second ed. Willan, Cullompton.
Mathiesen, Thomas, 1990. Prison on Trial. Waterside Press, Winchester.
Miller, Vivien, Campbell, James (Eds.), 2014. Transnational Penal Cultures: New Perspectives on Discipline, Punishment and Desistance. Routledge, London.
Morris, Terence, Morris, Pauline, 1963. Pentonville. Routledge and Kegan Paul, London.
Newbold, Greg, 1989. Punishment and Politics: The Maximum-Security Prison in New Zealand. Oxford University Press, New York.
Newburn, Tim, 2010. Diffusion, differentiation and resistance in comparative penality. Criminology and Criminal Justice 10 (4), 341–352.
Norris, Linda, 1974. Comparison of two groups in a southern state women's prison: homosexual behavior versus nonhomosexual behavior. Psychological Reports 34, 75–78.
Owen, Barbara, 1998. In the Mix: Struggle and Survival in a Women's Prison. State University of New York Press, Albany.
Piacentini, Laura, 2004. Surviving Russian Prisons: Punishment, Economy and Politics in Transition. Willan, Cullompton.
Phillips, Coretta, 2012. The Multicultural Prison: Ethnicity, Masculinity and Social Relations Among Prisoners. Oxford University Press, Oxford.
Propper, Alice M., 1981. Prison Homosexuality: Myth and Reality. Lexington Books, Lexington, MA.
Prison realities: views from around the world. South Atlantic Quarterly 113 (3), 2014.
Radzinowicz, Leon, 1998. Adventures in Criminology. Routledge, London.
Rock, Paul, 1996. Reconstructing a Women's Prison: The Holloway Redevelopment Project, 1968–1988. Clarendon Press, Oxford.
Ross, Jeffrey I., Richards, Stephen C., 2003. Convict Criminology. Wadsworth, Belmont.
Rowe, Abigail, 2011. Narratives of self and identity in women's prisons: stigma and the struggle for self-definition in penal regimes. Punishment and Society 13 (5), 571–591.
Salvatore, Ricardo D., Aguirre, Carlos, 1996. The Birth of the Penitentiary in Latin America. University of Texas Press, Texas.
Selling, Lowell, 1931. The pseudo-family. American Journal of Sociology 37, 247–253.
Serge, Victor, 1930. Men in Prison. Doubleday, New York.
Scraton, Phil, Sim, Joe, Skidmore, Paula, 1991. Prisons under Protest. Open University Press, Milton Keynes.
Sparks, Richard, Bottoms, Anthony E., Hay, Will, 1996. Prisons and the Problem of Order. Clarendon Press, Oxford.
Sykes, Gresham M., 1958. The Society of Captives: A Study of a Maximum Security Prison. Princeton University Press, Princeton.
Sykes, Gresham M., Messinger, Sheldon M., 1960. The inmate social system. In: Cressey, D.R., Grosser, G.H., McCleery, R., Ohlin, L.E., Sykes, G.M., Messinger, S. (Eds.), Theoretical Studies in Social Organisation of the Prison. Social Science Research Council, New York, pp. 11–13.
Tittle, Charles R., 1972. Society of Subordinates – Inmate Organisation in a Narcotic Hospital. Indiana University Press, Bloomington.
Useem, Bert, Kimball, Peter, 1989. States of Siege: US Prison Riots, 1971–1986. Oxford University Press, Oxford.
Wacquant, Loic, 2001. Deadly symbiosis: when ghetto and prison meet and mesh. In: Garland, D. (Ed.), Mass Imprisonment: Social Causes and Consequences. Sage, London.
Ward, David A., Kassebaum, Gene G., 1965. Women's Prison: Sex and Social Structure. Aldine, Chicago.
Wilson, Thomas P., 1968. Patterns of management and adaptations to organisational roles: a study of prison inmates. American Journal of Sociology 74, 146–157.

Prisoner's Dilemma, One Shot and Iterated

Diego Ríos, CONICET – UNTREF – UNSAM, Buenos Aires, Argentina
Eleonora Cresto, CONICET – Universidad de Buenos Aires, Buenos Aires, Argentina

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Schick, volume 18, pp. 12059–12062, © 2001, Elsevier Ltd.

Abstract

The Prisoner's Dilemma (PD) is a game whose Nash equilibrium is not Pareto efficient. The sheer perversity of the interaction involved in the PD has bewildered game theorists and social scientists alike. Despite the fact that players could be better off by jointly cooperating, they have individually a dominant strategy for defection that carries out a suboptimal outcome for everyone. This article explores and assesses this result, contrasting one-shot and repeated interactions.

Introduction

The Prisoner's Dilemma (henceforth, 'PD') figures prominently in game theoretic literature. It was first devised and discussed by Merrill Flood and Melvin Dresher in 1950, in the context of studies pursued by the RAND Corporation on strategic interaction (see Poundstone, 1992). Princeton mathematician Albert Tucker is responsible for the story usually related to introduce the PD; it appeared in print for the first time in Luce and Raiffa (1957). One possible version of it recounts how two criminals are arrested and placed in isolated cells. Each of them only cares about his own personal welfare, and not about the well-being of his accomplice; they both wish to minimize the number of years they remain in jail. The sheriff talks to them separately and offers the following proposal: "You can confess or remain silent. If both of you confess, you will get 10 years in jail. If you confess and your accomplice remains silent, you will get a one-year jail sentence and your accomplice will remain in prison for 25 years. If both of you remain silent, you and your accomplice will be jailed for 3 years."

Given the sheriff's proposal, each prisoner can reason thus: "If my partner remains silent, it is better for me to confess, because one-year in jail is better than three. If my partner does confess, it is also better for me to confess, because 10 years in jail is better than 25. In either case, irrespective of what my accomplice does, it is better for me to confess than to remain silent." Both players reason the same way, and they get 10 years each. The surprising result is that they could have obtained a mutually beneficial sentence of 3 years, had they both remained silent.

The PD has acquired a pervasive importance in mainstream social sciences and has been the subject of innumerable articles and books. Phenomena as diverse as the provision of public goods, the overexploitation of natural resources, the existence of price wars between firms, the escalation in arms races between superpowers and the overexpenditure in political party campaigns have been employed as putative candidates for analysis in terms of a PD. A note of precaution must be raised at this stage. Social phenomena are very rich. Game theoretic modeling requires impoverishing reality to focus only on a few abstract parameters, so it is never obvious whether a certain real-world situation satisfies the mathematical structure of a PD. We need to be careful not to confuse the target phenomenon with its game theoretic *representation*. Further discussion is given in Section From Social Phenomena to Game Theory.

The following section analyzes the structure of a two-person PD that is played just once; Section *N*-Persons PD addresses the problems that arise when we consider extending the two-person case to a game with *n*-players; Section Iterated PD, Two Persons deals with iterated versions of the game; and Section Evolutionary PD examines the so-called evolutionary PDs.

Two-Person, One-Shot PD

Figure 1 models the PD. Payoffs are represented by cardinal numbers. To keep it simple, assume that the years of imprisonment linearly amount to negative units of utility. Assume our prisoners are 'Row' and 'Column,' respectively. Each of them has to choose between options C and D, where C (the 'cooperative' choice) is the option to remain silent, while D ('defection') amounts to confession.

Information on Figure 1 can be generalized. For a PD matrix to arise it is sufficient to have payoffs a, b, c, and d for Row, and payoffs a^*, b^*, c^*, and d^* for Column, as shown in Figure 2, such that $c > a > d > b$, and $c^* > a^* > d^* > b^*$, with both $a > (c + b)/2$ and $a^* > (c^* + b^*)/2$ (Rapoport and Chammah, 1965; Axelrod, 1984). If ties are allowed the strict inequality could be relaxed by either $c \geq a$ or $d \geq b$, but not both; the same considerations apply to Column's values. If Row and

		Column	
		C	D
Row	C	−3, −3	−25, −1
	D	−1, −25	−10, −10

Figure 1

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63065-9

		Column	
		C	D
Row	C	a, a^*	b, b^*
	D	c, c^*	d, d^*

Figure 2

Column have the same values, the matrix is said to be symmetric (as in Figure 1).

Mainstream game theory usually assumes, for any game, that players have common knowledge of their options and payoffs, as well as of the fact that they are rational – meaning, not only does each player know that they are both rational, but each player also knows that the other player knows that they are both rational; each player knows that the other player knows that she knows this, etc. Note, however, that common knowledge of rationality is not necessary to account for the prisoners' reasoning as mentioned in Section Introduction. Each player has a strictly dominant strategy (D), and hence the game has a unique Nash equilibrium (D, D); if ties are allowed, D is weakly dominant for each agent. In either case, given that defection dominates cooperation, standard game theory has it that players can solve the game by just looking into their own payoffs and ignoring the payoffs of their opponent; the players cannot do better by unilaterally departing from their dominant strategy (but see Rationality Considerations for attempts to question the use of dominance reasoning in this or similar settings).

Still, insofar as both $a > (c + b)/2$ and $a^* > (c^* + b^*)/2$ hold, the pair of strategies (C, C) is the only Pareto-optimal point of the game, and hence both agents will be worse off by following the dominant strategy. (By contrast, if the inequalities are not satisfied, the agents will be able to reach Pareto-optimal points by using independent mixed strategies – hence, the injunction to defect might not be perceived as perplexing as in a bona fide PD.) The most interesting feature of the PD then is the existence of a unique Nash equilibrium (D, D) that is Pareto suboptimal. The explanation of why the players end up with a Pareto-inferior outcome is not hard to grasp. As a matter of fact, each player distinguishes between what is and what is not under his control. Row player compares payoffs along Columns – the only variables that are open to him – and Column player compares payoffs along Rows, for the same reason. By contrast, nobody compares payoffs along the diagonal, because diagonal payoffs are not under the *exclusive* control of any of the two players.

It has been argued that a PD models, with disarming simplicity, a possible conflict between individual and collective rationality. Collectively, everyone would be better off cooperating than defecting, but, individually, defection is the dominant strategy for each player. Incidentally, notice that the putative conflict between individual and collective rationality need not be read as a clash between egoistic and altruistic motives. Two agents can have altruistic preferences and yet be in a PD. By way of illustration, suppose Row and Column are placed in separate rooms and are given the option to press either a C-button or a D-button. They both know that, if Row presses the D-button while Column presses the C-button, \$4 will accrue to Column and nothing to Row; *mutatis mutandis* for Column pressing the D-button while Row presses the C-button. If they both go for the D-button, they receive \$1 each; if they both go for the C-button, they both get \$3. Suppose each agent only cares about the well-being of his partner; to simplify things, assume that Row obtains one unit of utility for each dollar received by Column and vice versa. The situation can then be described by the matrix shown in Figure 3, which is just a special case of Figure 2, and thus it has the structure of a PD, despite the fact that both agents have altruistic motivation.

As we have indicated, standard game theoretic rationality excludes the possibility of cooperation in a canonical, one-shot PD; the structure of interaction is such that only defection is sanctioned as rational. Of course one can always seek to modify the matrix in order to make defection more costly, for instance by introducing punishment (see Section From Social Phenomena to Game Theory). But this move should be carefully interpreted. The payoffs are supposed to factor every relevant information; they already include all there is to consider, and not just visible compensations such as money or years of imprisonment. Therefore, if the payoffs are changed, quite simply the game is changed. Once that happens, there is no point in claiming to have found a cooperative solution to the PD, because the game that is being played is not a PD any more (Binmore, 1994).

At this stage, it is crucial to distinguish between a PD *matrix* and a PD *situation*. The PD matrix is a mathematical structure. The PD situation is simply the target phenomenon that one wishes to model. In the following subsection, we elaborate on the significance of this distinction. We shall examine three different ways in which agents facing the very same scenario might avoid the fate of a PD: they can try to change the circumstances, so as to end up with a game defined by a less problematic matrix; they can understand the circumstances differently (or equivalently, they can seek to change their preferences) in order to become involved in a more benign game; or they can refuse to conceive themselves as playing a game in the first place, among other possibilities.

From Social Phenomena to Game Theory

Once the players believe they are dealing with a PD situation, they can attempt to modify their incentives so as to promote cooperation. The presence of third-party punishers, for example, helps to mitigate defection by forcing players to enter

		Column	
		C	D
Row	C	3, 3	0, 4
	D	4, 0	1, 1

Figure 3

into mutually advantageous exchanges. Legislation, for instance, is devised to enable people to cooperate by converting PD situations into interactions with cooperative solutions. Note that this leaves what we have stated concerning the logic for defection in the original PD *matrix* completely intact.

External intervention is not the only way a PD situation can be modified. Cooperation can be generated *endogenously*. By way of illustration, consider Oliver Williamson's analysis of trade practices during the Middle Ages, between merchants located in distant places (Williamson, 1983). The lack of an enforcing mechanism posed a serious problem of credibility at the moment of exchanging commercial promises. Both merchants would be better off if they could cooperate; but they were both afraid of the opportunistic behavior of the other party; they preferred not to cooperate rather than cooperate without being reciprocated. The solution was to allow for the exchange of asymmetrically valued goods. One merchant would send his own son to the other merchant to add credibility to his promise of future payment; the other would send the merchandise; once payment became effective by the first merchant, the second would proceed to liberate the son who was retained as credit security. This mechanism endogenously circumvents a situation of mutual distrust which could lead to a PD matrix, but without assuming exogenous agents charged to enforce contractual agreements.

Such examples of both exogenous and endogenous mechanisms being used to enhance cooperation can be described as cases in which agents find themselves in a PD situation, but where a PD matrix is nonetheless avoided due to changes in the relevant circumstances. Even more radically, notice that different agents under the same circumstances can disagree on whether they are, or are not, involved in a PD situation. Clearly, if the agents do not consider themselves as dealing with a PD situation, they will just not have a PD matrix to worry about; in general, matrices are obviously underdetermined by real-world phenomena. We can call it *the problem of interpretation*: whether players are or are not trapped in a dilemma depends on the way they interpret the relevant facts and on how they structure their preferences.

Thus, rational agents can use game theoretic considerations to arrive at quite different conclusions. Consider a modified version of Tucker's story; the sheriff's proposal remains as before (so in a sense, the external facts have not changed), but both prisoners attribute great value to cooperation in itself. Suppose their utilities are as shown in Figure 4.

Then the rankings from Figure 2 no longer hold; rather, we now have $a > c \geq b > d$; Column values are symmetric. We have just obtained what is often referred to as a *Stag Hunt* game. It is paradigmatically illustrated with the story of two friends who go on a hunt; they can choose to hunt either a stag or a hare, and neither knows what the other will choose to do. In order for Row to succeed in hunting a stag, Column must choose to hunt a stag too. Furthermore, each of them can hunt a hare without help, although a hare is not as worthy as a stag. Notice that here we have two Nash equilibria, (C, C) and (D, D), though of course only (C, C) is Pareto-optimal. It is a 'game of trust,' or a 'game of assurance': if Row has great confidence that Column will choose C, she should choose C as well. In short, depending on how the agents involved perceive things, a potential PD situation might well turn into a Stag Hunt matrix – or into any other game, for that matter (Skyrms, 2003).

A more radical twist on the interpretation problem can be found in Schick's analysis (2003). Schick has argued that agents might disagree on whether they face a true dilemma even when their options and preferences *are* actually just as described by the PD matrix; the reason is they might not think of themselves as playing a game in the first place. To this effect, Schick distinguishes between *Plights* and *Dilemmas*. Row and Column are in a Prisoner's Plight every time their options and preferences are as exhibited by what has thus far been called a standard PD matrix, and they know them to be so. Suppose by way of concreteness their payoffs are as shown in Figure 3. To be in a true PD, in addition, they should be *in a game*, namely, they should each see the other as now ready to make a choice from among the available options. But this further condition need not be fulfilled. Consider, for example, propositions A and B, where A is 'Column will do as I do' (or better: 'whichever of Cooperate/Defect describes what I do, the same describes what my partner does'), and B is 'Column will do the opposite' (better yet: 'whichever of that pair describes what I do, the other describes what my partner does'). Row might then organize things along the payoff structure shown in Figure 5.

Row does not know what Column will do, so Row does not know whether A or B is true; moreover, she supposes A and B to be probabilistically independent of her own choices (for alternative perspectives that question this assumption, see Section Rationality Considerations). Notice that Column cannot think of A or B as actions that are open to her; hence, this is no longer a game setting, but a one-person decision problem. As there is no longer a dominant option, it may well be rational for Row to cooperate (i.e., to remain silent). Column, in turn, may also organize things in terms of A/B, in

		Column	
		C	D
Row	C	4, 4	0, 3
	D	3, 0	3, 3

Figure 4

		Column	
		A	B
Row	C	3, 3	0, 4
	D	1, 1	4, 0

Figure 5

which case she will also be dealing with a one-person decision problem. The upshot is that people are not always playing games; sometimes they do not think in terms of the other's choosing and acting on their options. Once again, this line of reasoning is not meant to offer a way out of the dilemma. But it does point to the fact that there might be fewer dilemmas in real life than what we could have first assumed.

Rationality Considerations

Thus far we have taken the standard game theoretic analysis of the PD for granted; the generic matrix in **Figure 2** could be interpreted as embodying a *dilemma* because the theory sanctions as rational a Pareto-inferior outcome. In the light of this, it might be contended that our pre-theoretical concept of rationality is not fully determined or that there is more than one such concept at stake. This, in turn, has led some authors to question the legitimacy of the standard analysis and the use of dominance rules to solve the PD matrix.

According to Davis (1977, 1985), for instance, once we assume that the players have common knowledge of rationality as well as of the fact that they face a perfectly symmetrical predicament, we are bound to conclude that they cannot but choose alike; outcomes in which one of them defects and the other cooperates are just not possible. Hence each player can only choose between defecting while the other also defects, and cooperating while the other also cooperates. The only rational thing to do for both players is then to cooperate (for an extensive criticism of this argument, see Binmore, 1994). Interestingly, Davis does not question the standard game theoretic analysis when players are not assumed to have common knowledge of the fact that they are perfectly rational agents.

More generally, it has been suggested that PDs might be a symptom of a deeper conflict between Bayesian intuitions and equilibria concepts from game theory. Consider a PD-like matrix in which the choices of Row are not probabilistically independent of those of Column. Suppose Row judges the probability that Column will choose C conditional on Row's choosing C to be close to one; *ditto* for the probability that Column will choose D conditional on Row's choosing D. In this case, Bayesian rationality seems to indicate that Row should choose C, as C's expected utility is greater than D's. The same argument applies to Column; both Row and Column can judge each other to be expected utility maximizers, and hence, they ought to make choices that fail to yield a Nash equilibrium (Levi, 1997). Thus, as long as each prisoner has reasons to perceive her partner as sufficiently similar to herself (so as to be very likely to make the same choices), we have a PD scenario which, arguably, does not lead to a true dilemma.

These and similar suggestions have been criticized on the grounds that, in a PD matrix, each agent gets to choose independently of what her partner does. Therefore, even if there were a probabilistic dependency between the options, there is no *causal* dependency; therefore, in this case weighting each agent's utilities with conditional probabilities is bound to yield the wrong verdict.

Proposals about what a rational agent should do when probabilistic and causal dependency come apart, however, remain controversial. Discussions on this topic have been fueled by Robert Nozick's description of *the predictor's paradox*, better known as *Newcomb's problem* (Nozick, 1969; see Lewis, 1979, for an argument to the effect that Newcomb's problem is just a PD). Suppose an agent can choose between (1) the content of an opaque box; or (2) the content of both the opaque box and a transparent one, which is seen to contain $1000. Obviously, (2) appears to be a dominant option. As it happens, however, the opaque box contains 1 million dollars if and only if a very reliable predictor of human behavior has already predicted that the agent will go for option (1), and nothing otherwise. Notice that, in the usual reading of the problem, the probability of there being a million dollars in the opaque box conditional on the agent's choosing (1) is very large; the probability of there being a million dollars in the opaque box conditional on the agent's choosing (2) is very low. This is so even though, of course, the agent's choices cannot cause the money to appear in the box; either the money is already there, or it is not. Assuming utilities are linear with money, the expected utility of option (1) is close to one million, and that of option (2) is close to 1000. Thus, as in the PD scenario enriched with probabilities that were described in the earlier paragraphs, in Newcomb's setting there appears to be a conflict between choosing the dominant option (2), and choosing the one with the largest expected utility (1) (see **Figure 6**).

To account for this conflict, in the wake of Newcomb's paradox, some authors have sought to develop a *causal* brand of decision theory. In some of its versions, causal decision theory advises us to calculate expected utility using probabilities of counterfactual conditionals (which are meant to be sensitive to causal links) rather than standard conditional probabilities. Hence, every time there is causal independence between acts and possible states of the world, we will obtain those options with the highest expected utility, thus calculated, to coincide with dominant options (Gibbard and Harper, 1978; Lewis, 1981). The fact remains, however, that agents who choose the dominated option in a Newcomb's setting are on average at a better advantage than those who stick to the dominant one – just as, in a PD, a pair of cooperative agents end up individually better off than a pair of defectors. Once again, we could take this fact as evidence of a clash between different underlying notions of rationality; alternatively, it has also been contended that, if a setting is such that *irrationality* is rewarded, then there is nothing our theory can (or should) do about it. (See the collection of papers in Campbell and Snowden, 1985 for a broader perspective on this discussion, including articles arguing that

	The opaque box contains 1 million dollars	The opaque box does not contain 1 million dollars
(i) One box	1 000 000	0
(ii) Two boxes	1 001 000	1 000

Figure 6

morals drawn from Newcomb's setting should be taken *cum grano salis*; for a sympathetic and thorough presentation of causal decision theory, see Joyce, 1999).

N-Persons PD

Mainstream social sciences have identified many instances of conflict between individual and collective rationality. Mancur Olson argues that the mismatch between individual incentives and collective benefits might hamper collectively beneficial actions (1965). Russell Hardin claims that the same discontinuity between individual interests and collective benefits accounts for the depletion of common resources (Hardin, 1971). This kind of phenomena involves the participation of many individuals. A natural question to consider is whether these situations can be modeled as multipersons PDs.

Russell Hardin was among the first to conceptualize collective action problems as *N*-person PDs (1971, 1982). Dawes did the same with the tragedy of the commons (1980). Collective action problems occur when it becomes necessary to mobilize many agents – eventually all of them – in order to provide a nonexcludable good that is beneficial for the group but that requires costly behavior on the part of the individuals. The tragedy of the commons occurs when players fail to form a self-restrained coalition to avoid the depletion of common resources.

The distinction between situations and matrices that we have introduced in earlier sections is relevant here as well. As we have said, Hardin and Dawes attempted to show that collective action problems and the tragedy of the commons are essentially PD situations that can be translated into a PD matrix. Michael Taylor adds a note of caution. He argues that some cases of collective action can be modeled as assurance games involving many players (Taylor, 1987). The issue is not without theoretical and practical significance. Hardin claims that, given that Olson's collective action problems are essentially *N*-person's PDs, any analysis prescribing a solution for the PD must prescribe a similar solution for the collective action problem (Hardin, 1982). Naturally, this will not be possible if, as Taylor holds, there are competing game matrices for the same collective action situation.

Philip Pettit extends the issue still further (1986). He distinguishes between two types of many-persons PD. *Free-rider problems* are paradigmatic cases of productive cooperation, in which defectors enjoy the benefits of the cooperative behavior of the others. They parasitize the efforts of the cooperative individuals and they share the benefits of the outcome. In *foul-dealing problems*, on the other hand, defectors get a benefit at the *expense* of the other players. Defectors obtain a benefit that is superior to the benefit that everyone would obtain in case of universal cooperation (Pettit, 1986). The provision of public goods is an example of the free-riding problem. The Hobbesian logic of the state of nature is an illustration of the foul-dealing case. Consider for instance the case of mutual disarm. Defectors not only enjoy the advantage of universal peace, but they also end up in a preeminent position – being armed in a disarmed world at the detriment of others.

Iterated PD, Two Persons

Consider now a repeated PD. In games of repeated interaction, the possibility of future contact becomes crucial for deciding whether or not to cooperate. There are two possible scenarios. In one of them, the game is finitely iterated and the players know the number of encounters. In the other case, iteration is infinite or of unknown length. Mainstream game theory treats these two situations differently. There are, however, empirical data that challenge the adequacy of this analysis.

Finitely Iterated PD

The analysis of finitely repeated games is similar to that of one-shot games: rational players should not cooperate. The last move in a finitely repeated PD *is* a one-shot PD having defection as a dominant strategy. Assuming common knowledge of rationality and common knowledge of the length of the game, defection propagates by backward induction to the earlier moves. For the sake of concreteness and without loss of generality, in **Figure 7** we have represented an iterated PD in extended form. The game has two moves for each player, and the red line divides the first instance of the game from the second one. The game is played simultaneously by each one of the players, starting by player 1. Payoffs are represented with ordinal numbers.

Backward induction reasoning mandates thinking from the end of the game to the beginning. Thus, player 1, before choosing whether to cooperate or to defect, calculates what player 2 will do in each possible ending of the game. In the upper branch of the tree, for instance, player 2 will need to choose between cooperating (3,3,3,3) or defecting (3,3,1,4). In case the game arrives to that point, rationality requires for player 2 to defect. Player 1 will continue doing these calculations *for each ending point*, before deciding his own behavior at the starting of the game. By backward reasoning, player 1 is able to deduce player 2 ultimate choices, restricting his own available options. This reasoning delivers universal defection (a result that is represented by the lowest line in the graphic). Note, however, that the equilibrium outcome (2,2,2,2) is evidently Pareto-inferior with respect to universal cooperation (3,3,3,3).

Experimental data provide a less conclusive picture. In theory, rational players must always defect. In practice, however, a well-established empirical literature shows that players tend to behave cooperatively at the beginning of the game and defect when the end of the game is approaching (Selten and Stoecker, 1986; Andreoni and Petrie, 1993). This finding casts doubts on the psychological plausibility of the backward induction method and it blurs the distinction between finitely and indefinitely repeated PDs.

An interesting literature has emerged to scrutinize this issue (Pettit and Sugden, 1989; Binmore, 1997; Kreps et al., 1982). Kreps et al. (1982) argue that small doses of incomplete information can account for the experimental findings showing cooperation in the early stages of the game. Suppose that one of the players believes that there is some chance for the other to behave irrationally; in that case, it is also reasonable for her to behave irrationally and cooperate. The argument applies in reverse mode to the other player,

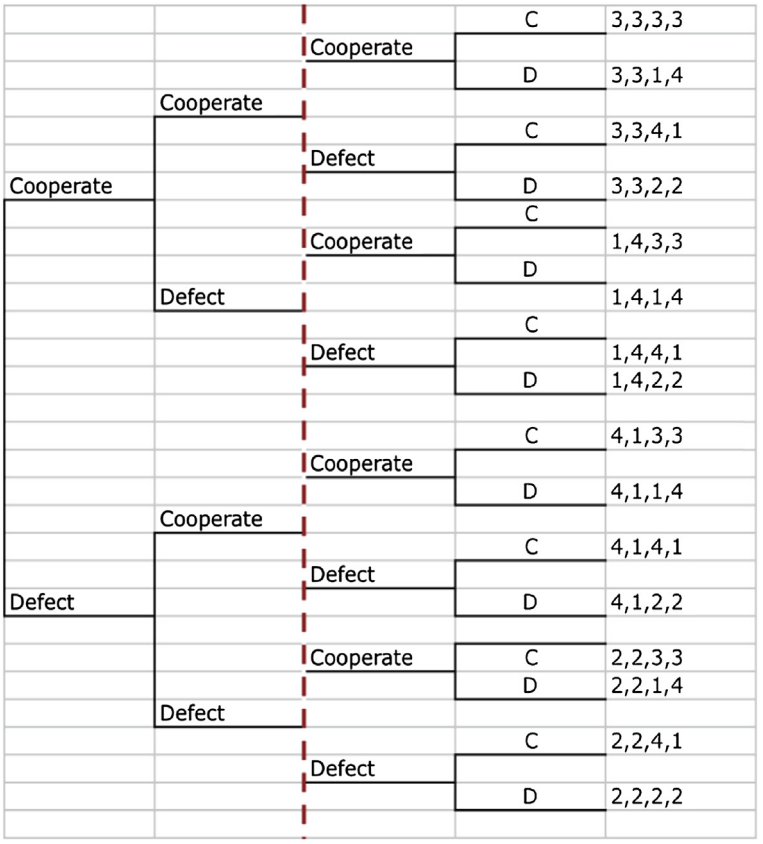

Figure 7

reinforcing the belief that he has an irrational opponent. Relaxing the assumptions of common knowledge of rationality and allowing for incomplete information generates drastic changes in the way the game is supposed to be played.

Indefinitely Repeated PD

Indefinitely repeated PD involves cases where the players ignore the length of the game. Most models assume a probability measure indicating the chances of future encounters. A discount factor could also be included in order to calculate the present value of future payoffs. The relevance of these assumptions depends, among other elements, on the nature of the encounters – on whether agents are involved in pairwise interactions or simply playing the field.

Game theorists analyzed a wealth of contingent strategies. So-called nice strategies are never the first to defect. An example is GRIM, which starts cooperating, but when the opponent defects, it defects as well, and continues defecting irrespectively of the behavior of the other player. Another example is TIT-FOR-TAT (henceforth, 'TFT'). TFT always starts by cooperating, and then does what the other player did on the previous move.

TFT has a number of interesting features: (1) As stated, it is a nice strategy: it will not defect before its opponent does; (2) it is also retaliating: it will respond to a defection by defecting as well; (3) it is forgiving: if the opponent, after having defected, returns to cooperation, TFT will respond by cooperating again; and (4) TFT is non-envious: it just attempts to do as much as it can, irrespective of what the opponent is doing.

In the early 1980s, Robert Axelrod invited researchers to submit computer programs implementing strategies to be pitted against each other in a PD repeated 200 times. Devised by Anatol Rapoport, TFT entered the scene as one of the 14 programs in the competition (Axelrod, 1984). TFT, which obtained the greatest sum of payoffs over all plays, was declared the winner. Sometime later Axelrod repeated the experiment, allowing for new strategies to enter the tournament; this time he used a probabilistic termination rule to avoid end-game effects. At that point 63 strategies took part of the competition. Researchers were aware of the previous results as well as of the score obtained by TFT. Some new strategies were thus specially devised to counter TFT, such as TIT-FOR-TWO-TATS, sponsored by John Maynard Smith. TFT won again.

These results need to be carefully interpreted. To begin with, notice that TFT is not the best strategy *simpliciter*; it is only the best in the sense that it obtained overall the greatest sum of payoffs. Several strategies were able to defeat TFT on one-to-one competitions. The success of TFT was possible because it added the benefits of cooperation *across* competing strategies. The nature of the tournament was decisive for this success: in a tournament that matched strategies pairwise and only

allowed the winner of each pair to continue the competition, TFT could have been eliminated early (Axelrod, 1984). Second, the fact that the tournament was played by automata was significant because it ruled out mistakes. In a competition where mistakes or trembles are possible, TFT might be provoked by another TFT by mistake, triggering an escalating process of mutual defection. TFT does not have any instruction to stop such escalations (Axelrod, 1984). Last but not least, the results obtained by Axelrod depended largely on the original setting. The starting point of the tournament only included a limited number of strategies; many other possibilities were excluded for no good reason. The advantages of TFT are then dependent on the existence of alternative strategies actually available in the population (Binmore, 1998).

Evolutionary PD

The PD can also be interpreted in evolutionary terms. In this context, large populations adjust their strategies in agreement with the strategies available in the environment. Payoffs express the fitness of specific phenotypes; they reflect the differential spread of strategies in the environment. Note that they are conceived of as true automata: they do not *choose* to cooperate or defect, they are rather wired to behave in certain ways. For the sake of concreteness, imagine a population with two different phenotypes: defectors and tit-for-tatters. We assume the typical replication dynamics that describes the change in the long run of the relative frequency of a set of phenotypes under the condition that they produce error-free copies of themselves. In the replication dynamics framework, behaviors with above-average fitness increase their frequency in the population. The focus is specifically on the role of selection; mutations are not explicitly modeled, but it is assumed that they disturb the system, albeit in an isolated manner. We assume, moreover, that interactions occur on random pairs n times, and that the payoffs for each interaction are cardinally represented as in **Figure 1**.

If two defectors meet each other, they each obtain -10 multiplied by the number of plays; to wit: $-10n$. If two tit-for-tatters meet each other they obtain $-3n$ each. The case in which a defector meets a tit-for-tatter is more interesting: in this scenario the tit-for-tatter obtains -25 and the defector -1. However, the tit-for-tatter will retaliate all along the remaining moves of the game, obtaining in the second move -10 for himself and -10 for the defector. This can be represented as $-10(n-1)$. When playing against a tit-for-tatter, the defector gets $-1+(-10)(n-1)=-10n+9$, and the tit-for-tatter gets $-25+(-10)(n-1)=-10n-15$.

We introduce now the information concerning the proportion of each phenotype in the population. Let us represent the proportion of tit-for-tatters in the population by x. The expected payoff for a defector is

$$x(-10n+9)+(-10n)(1-x). \qquad [1]$$

Whereas the expected payoff for the tit-for-tatter is

$$x(-3)n+(1-x)(-10n-15). \qquad [2]$$

We now combine [1] and [2] into the following inequality:

$$[x(-3)n+(1-x)(-10n-15)>x(-10n+9) + (-10n)(1-x)]. \qquad [3]$$

Solving for x, we obtain all the values at which tit-for-tat is fitter than defection:

$$x>15/(7n+6). \qquad [4]$$

Notice the importance of the value of n in [4]: the greater the value of n, the lower the value of x. This means that the longer the play of game, the easier for tit-for-tatters to spread, because x and n are inversely related. At the exact point in which the right side and the left side of [3] equate, there is a balancing point; at that point the population is polymorphic with exact proportions of tit-for-tatters and defectors coexisting side by side. This co-existence, however, is not stable: any mutant that modifies the proportion in the strategies available in the environment tips the balance and pushes the population toward one extreme or the other (all tit-for-tatters or all defectors). If the new entrant is a tit-for-tatter, the population evolves toward a situation where every phenotype is a tit-for-tatter. If the new entrant is a defector, the whole population ends up being composed of defectors. These two extreme situations (all phenotypes are tit-for-tatters or all are defectors) represent two possible monomorphic equilibria. These equilibria are stable; they are robust enough to resist the entrance of new exemplars of the opposite phenotype.

This analysis is simplified in different ways. To begin with, note that we are comparing only two strategies: defection and TFT. If more strategies are allowed into the picture, the analysis becomes more complex (Nowak, 2006). Second, as we have already said, we modeled the game using the traditional replication dynamics in an infinite population setting where strategies are randomly paired. This model might not take into account important features. Paring strategies randomly do not capture geographical effects that might be of paramount importance for cooperation (Axelrod, 1984).

Geographical effects can emerge in different ways. One possible line of analysis consists in making local interaction derive from exogenous factors. Imagine a population where some members tend to interact with individuals that are like themselves according to some arbitrary phenotypic trait. If this trait is correlated to cooperativeness, we should expect clusters of cooperators that meet each other using that specific trait as clue. Neighboring effects can be also endogenously generated. Although the original interaction is random, subsequent interactions can be modeled as a Bayesian process in which individuals *adjust* the frequency of their interactions according to prior successful interactions with specific individuals (Skyrms, 2004). Other models use cellular automata that exclude memory, relying exclusively on the state of the neighbors' neighbors (Nowak and May, 1992). These models have the potential to account for the formation of clusters of cooperativeness in a way that the process of random pairing might neglect.

See also: Evolution, Natural and Social: Philosophical Aspects; Evolutionary Approaches to Human Behavior: Philosophical

Aspects; Individualism versus Collectivism: Philosophical Aspects; Intentionality and Rationality; Irrationality: Philosophical Aspects; Joint Action; Methodological Individualism: Philosophical Aspects; Pareto, Vilfredo (1848–1923); Rational Choice Explanation: Philosophical Aspects; Trust: Philosophical Aspects; Von Neumann, John (1903–57).

Bibliography

Andreoni, J., Petrie, R., 1993. Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence. Economic Journal 103, 507–585.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Binmore, K., 1994. Game Theory and the Social Contract: Playing Fair, vol. 1. The MIT Press, Cambridge, MA.

Binmore, K., 1997. Rationality and backward induction. Journal of Economic Methodology 4, 23–41.

Binmore, K., 1998. Game Theory and the Social Contract: Just Playing, vol. 2. The MIT Press, Cambridge, MA.

Campbell, R., Snowden, L., 1985. Paradoxes of Rationality and Cooperation. University of British Columbia Press, Vancouver.

Davis, L., 1977. Prisoners, paradox and rationality. American Philosophical Quarterly 14. Reprinted in Campbell and Snowden, 45–58.

Davis, L., 1985. Paradoxes of rationality and cooperation. In: Campbell, R., Sowden, L. (Eds.), Is the Symmetry Argument Valid? Campbell and Snowden, Vancouver, pp. 255–262.

Dawes, R., 1980. Social Dilemmas. Annual Review of Psychology 31, 169–193.

Gibbard, A., Harper, W.L., 1978. Counterfactuals and two kinds of expected utility. In: Hooker, C., Leach, J.J., McClennen, E. (Eds.), Foundations and Applications of Decision Theory, vol. 1. Reidel Publishing Company, Dordrecht, pp. 125–162.

Hardin, R., 1971. Collective action as an agreeable N-prisoners' dilemma. Behavioral Science 16, 472–481.

Hardin, R., 1982. Collective Action. John Hopkins Press, Baltimore.

Joyce, J., 1999. The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge.

Kreps, D., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoner's dilemma. Journal of Economic Theory 27, 245–252.

Kuhn, S., 2007. Prisoner dilemma. Stanford Encyclopaedia of Philosophy.

Levi, I., 1997. Prediction, deliberation, and correlated equilibrium. In: The Covenant of Reason: Rationality and the Commitment of Thought. Cambridge University Press, Cambridge.

Lewis, D., 1979. Prisoner's dilemma is a Newcomb problem. Philosophy and Public Affairs 8, 235–240.

Lewis, D., 1981. Causal decision theory. Australasian Journal of Philosophy 59, 5–30.

Luce, R.D., Raiffa, H., 1957. Games and Decisions. John Wiley & Sons, New York.

Nowak, M., 2006. Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press, Cambridge, MA.

Nowak, M., May, R., 1992. Evolutionary games and spatial chaos. Nature 359, 826–827.

Nozick, R., 1969. Newcomb's problem and two principles of choice. In: Rescher, N. (Ed.), Essays in Honor of Carl G. Hempel. Reidel, Dordrecht, pp. 114–146.

Olson, M., 1965. The Logic of Collective Action: Publics Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.

Pettit, Ph, 1986. Free riding and fool dealing. Journal of Philosophy 83 (7), 361–379.

Pettit, Ph., Sugden, R., 1989. The backward induction paradox. Journal of Philosophy 86 (4), 169–182.

Poundstone, W., 1992. Prisoner's Dilemma. Doubleday, New York.

Rapoport, A., Chammah, A., 1965. Prisoner's Dilemma. The University of Chicago Press, Ann Arbor.

Schick, F., 2003. Ambiguity and Logic. Cambridge University Press, Cambridge.

Selten, R., Stoecker, R., 1986. End behavior in sequences of finite prisoner's dilemma supergames: a learning theory approach. Journal of Economic Behaviour and Organization 7, 47–70.

Skyrms, B., 2004. The Stag Hung and the Evolution of Social Structure. Cambridge University Press, Cambridge.

Taylor, M., 1987. The Possibility of Cooperation. Cambridge University Press, Cambridge.

Williamson, O., 1983. Credible commitments: using hostages to support exchange. American Economic Review 73 (4), 519–540.

Prisons and Imprisonment

Matthew DeMichele, RTI International, Research Triangle Park, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Tonry, volume 18, pp. 12055–12059, © 2001, Elsevier Ltd.

Abstract

Outside of execution, incarceration is the most severe form of punishment used by modern societies. This article describes the origins of prisons and imprisonment, offers comparative statistics on incarceration rates, discusses theories of punishment, highlights unanticipated consequences of incarceration, and suggests potential ways to make punishment more effective.

Origins: How Did Prisons Develop?

Prisons are used to house individuals convicted for serious crimes. However, few people are aware of how prisons developed. In fact, early nineteenth century American prisons were not thought of as punishment. The idea of incarcerating someone as punishment was unheard of. Prisons were developed as a liberal reform to stem abusive punishments. Locking people in prisons was believed to be costly and overly sympathetic. Why did the prison become the central form of punishment? What changed to bring about incarceration? What goals were prisons to achieve? How does punishment reflect broader social features? This article addresses these questions.

Prisons are not a natural feature of society. Rather, for much of Western development societies relied on banishment, physical punishment, and executions. Some common punishments included cutting off portions of an ear or finger, whipping, branding, and the wheel. These punishments were carried out in front of large crowds, and sent a message to onlookers of what happens to criminals. Additionally, these punishments left visible marks to stigmatize individuals. This way, if an individual showed-up in a town missing part of his/her ear or finger, others would know that he/she was a criminal.

Prisons are a construction of nineteenth century social, political, and economic developments. Rusche and Kirchheimer (1968) demonstrated that physical punishments slowly gave way to incarceration as new economic forms emerged. During the nineteenth century, urban industrial capitalism replaced agrarian production and mercantile exchanges. Industrial production required large populations of able-bodied workers to operate factories. Urbanization required people to leave their small rural communities and migrate to central cities. Prior to this time, perceptions of crime and criminals were much different. In Rothman's (1971) historic treatise, *The Discovery of the Asylum: Social Order and Disorder in the New Republic*, he explained the shift from accepting crime, poverty, and insanity as natural elements of the community to the emergence of scientific understandings and interpretations of these issues as social problems. Colonial Americans saw crime and other forms of deviance (e.g., poverty, insanity) as natural parts of social organization that needed to be controlled, deterred, and contained, but never thought they would be eliminated (Rothman, 1971). Urban congestion, separation from family members, and lack of strong social ties to community, required larger governmental responses to crime, insanity, and poverty. No longer could people rely on informal social networks. This created a need for new kinds of formal social control and assistance, in which the asylum, prisons, and welfare supports emerged simultaneously. Urban industrialization did not cause these responses to social problems, per se, but this reconfiguration of community necessitated new forms of control and assistance.

The prison was developed with the ultimate goal of fixing or rehabilitating inmates. Michele Foucault (1977) characterized this shift in goals for punishment as a move from 'punishing the body' to 'punishing the soul.' The initial forays into incarceration date back to the sixteenth century with jails, poorhouses, and workhouses, but these were sparse and typically locally operated facilities in which individuals were housed in horrid conditions. Foucault (1977) depicted a movement spanning from around the 1770s to the 1860s in which prison developers became focused on changing inmates' behavior. By punishing the soul, Foucault referred to the religious instruction, vocational training, and inclusion of psychiatric sciences in prisons to adjust the way inmates thought and acted. This orientation explains why prisons were referred to as houses of correction, and, currently, US prisons are part of state-based Departments of Corrections. Behavioral change became a central stated goal for prison administrators, policy makers, and reformers. But, many continue to question how well prisons have achieved this goal.

The United States was an early adopter of imprisonment with prisons built throughout the nineteenth century. In New York, the Auburn prison, opened in 1819, and, in Pennsylvania, prisons opened in 1826 and 1829, in Pittsburgh and Philadelphia, respectively (Rothman, 1971). The Auburn system was known as the congregate system in which inmates slept in individual cells, but worked together during the day in silence. The Pennsylvanian prisons were more concerned with separation and introspection, including inmates never seeing other inmates. Inmates worked in solitude in their cells and only spoke with guards and chaplains. Silence, separation, and introspection were seen as tools to teach prosocial values and compensate for the poor upbringings and lack of discipline that were believed to cause criminal behavior. Prisons were to instill in inmates the needed discipline that their childhoods' lacked. Rothman (1971: p. 97) quotes Tocqueville and Beaumont after their visit to US prisons, in 1831: "Everything passes in the most

profound silence, and nothing is heard in the whole prison but the steps of those who march, or sounds proceeding from the workshops … We felt as if we traversed catacombs; there were a thousand living beings, and yet it was a desert solitude."

These US penal models quickly spread to other states and countries. The prison was seen as replacing the horrible conditions of the jails and other crude forms of confinement. Reformers concentrated on psychiatric approaches to train inmates to return to society as prosocial individuals. The prison has vacillated between emphasizing punishment and retribution vs rehabilitation and reentry since its inception. Currently, penal administrators and policy makers grapple with ways to balance these contradictory goals.

Comparative Incarceration Rates

Social scientists use comparisons to explain similarities and differences across units of analysis. The most prevalent measure of the size of incarcerated populations is reported as a rate per 100 000 people in the country. This measure gives the relative number of people in prisons and allows for making some comparisons across countries and within countries over time. When analyzing incarceration rates, it is obvious that there is tremendous variation in incarceration rates. Comparative research is hampered by making sure that the countries being compared are similar in several important ways. For instance, it is difficult to make comparisons between the United States and small nonindustrial countries or between countries with totalitarian regimes. Instead, researchers compare the United States to other advanced capitalist democracies.

The United States incarcerates between seven and ten times the proportion of adults as other Western democratic countries. There are currently more than 2 million adults incarcerated in the United States, which results in an incarceration rate of 719 per 100 000 (World Prison Brief, 2013). This incarceration rate reflects a break with previous times. From 1925 to 1975, the US incarceration rate fluctuated between 90 and 180 per 100 000. Then, in the mid-1970s, there were a series of changes to the criminal justice system that fostered a steady increase in incarcerated populations. Some of these changes include replacing indeterminate sentencing with determinate sentencing, the drug war, three-strike laws, zero tolerance ordnances, more police, and more aggressive prosecution of cases. These changes reflect a new orientation toward imprisonment that emphasized punishment and retribution. Although there are several explanations for this change, many scholars point to the publication of a research article in which it was argued 'nothing works' in correctional rehabilitation (Martinson, 1974). The nothing works report provided academic support for the notion that prisons do not actually 'correct' people. And, if rehabilitation is impossible, then, agencies should no longer focus on behavior change.

How do the US incarceration rate trends fit with other countries? Typically, comparative researchers compare advanced capitalist countries found in the West. This research groups countries according to politics and culture such that Anglo-American, Continental European, and Scandinavian countries fit together on a series of policy outcomes (e.g., economics, incarceration, welfare spending) (Cavadino and Dignan, 2006; Esping-Anderson, 1990). For instance, in Scandinavia, throughout the post–WWII period, incarceration rates fluctuate between 45 and 70 per 100 000. One exception of this is Finland's incarceration rate, which, throughout the 1960s, averaged 145 per 100 000. However, this rate reflected Finland's cultural attachment to Russia and included many political prisoners. As Finland became more incorporated in Scandinavian culture, politics, and law, policy makers embarked on a decarceration movement that included inmate amnesties and early releases. Finland now incarcerates about 60 per 100 000 population (World Prison Brief, 2013).

Continental European incarceration rates are slightly higher, on average, than Scandinavian countries. These rates fluctuate between 65 and 110 per 100 000 throughout the second half of the twentieth century. This is not to suggest complete homogeneity among these countries. Austria, for instance, had an average incarceration rate, in the 1960s, of 113 per 100 000 and in the early 2000s, the average rate was 88 per 100 000. Belgium, France, Italy, and the Netherlands, however, all experienced incarceration growth. The Netherlands provides an interesting case because the average incarceration rate was 25 per 100 000 through the 1960s, but, in the early 2000s, the rate jumped to 89 per 100 000. David Downes (2001) explained this incarceration rate growth as a product of social and criminal justice changes, namely, prosecutors became more aggressive and prison waiting lists were eliminated. Downes' research highlights the fluid nature of imprisonment as a product of policy decisions.

Although the United States is the most aggressive punisher among the West, the Anglo-American countries stand out compared to other Western countries. These countries have average incarceration rates that fluctuate between 73 and 700 per 100 000 from 1960 to 2002. Australia, Canada, Ireland, New Zealand, and the United Kingdom all had increasing incarceration rates throughout this time frame. Australia, New Zealand, and the UK incarceration rates more than doubled from averages ranging between 65 and 73 per 100 000 to 129 and 152 per 100 000. These advanced capitalist countries display different ranges of acceptable levels of punishment in which Anglo-American countries have the greatest tolerance for punishment and Scandinavian countries have the lowest thresholds.

Explanations of Comparative Incarceration Rate Differences

Several criminologists have tried to explain cross-national incarceration rate differences and temporal changes. These explanations typically build on two classic theoretical perspectives: (1) Emile Durkheim's cultural thesis and (2) Marxian conflict perspectives. Durkheim suggested that punishment reflects cultural norms, and the act of punishment furthers cohesion around such norms. He argued that crime and punishment are functional societal features. Marxian perspectives differ by emphasizing the role of punishment to maintain class boundaries and, essentially, punish the working classes as a central weapon in class conflict. The most compelling contemporary explanations of incarceration rates integrate the functional and conflict perspectives by

recognizing that punishment solidifies cultural norms and it also disproportionately impacts the most disadvantaged in society.

The leading comparative explanations of incarceration rate variation capture structural and policy differences. Garland (2001) suggested that the United States and the United Kingdom have cultural preconditions making them more susceptible to harsh criminal justice responses to crime and deviance. Cavadino and Dignan (2006) and Downes and Hanson (2006) found that incarceration rates are inversely related to welfare provisions such that some countries emphasize inclusive, as opposed to exclusive, approaches to addressing crime, deviance, and poverty. There is a growing body of research that emphasizes political–economic differences such that the influence of conservative political institutions – referred to as neo-liberalism – increase incarceration rates, compared to more egalitarian politics – referred to as corporatism that are less reliant on incarceration (for review, see Jacobs and Jackson, 2010). Greenberg (1999), Jacobs and Kleban (2003), and Sutton (2000, 2004) have all found similar results in which the left-oriented corporatist policies found in Scandinavia and Continental Europe support lower incarceration rates, compared to more conservative politics found in Anglo-American countries.

Other criminologists have extended these political–economic perspectives to suggest that there are specific legal differences contributing to incarceration rates. Savelsberg (1994) and Whitman (2003) found the influence of organizational differences between European and Anglo-American criminal justice agencies consequential for punishment outcomes. Namely, in the former, practitioners are considered trained civil servants and have the ability to make decisions without concerns of public disappointment. Conversely, Anglo-American criminal justice practitioners work in legal environments – i.e., common law – in which public participation and social interests are more easily reflected. Features such as case-based law development, independent juries, and broad prosecutorial discretion are suggested to make it easier to respond to public desires for harsher punishments. Lacey (2008) extended this argument when she pointed out that political dynamics are important, and argued that researchers need to consider differences in how practitioners are selected, trained, and operate throughout their career. Advancing her perspectives, DeMichele and Janoski (2010) showed that differences in lay participation, judicial power, and different rules of evidence are consequential for incarceration rates. The consensus among these explanations is that politics and legal differences are essential to understanding comparative incarceration rate variations.

Unanticipated Consequences of Incarceration

Social policies have the potential for unanticipated negative consequences. The notion of unanticipated negative consequences was identified by Robert Merton (1936: p. 894) as an important area for sociological study. Unanticipated consequences are often related to policy makers lacking knowledge of causal mechanisms, simply making mistakes, following a priori beliefs, and incorrect predictions about outcomes (Merton, 1936; Payne and DeMichele, 2011). Several criminologists have called attention to unanticipated negative consequences stemming from US mass incarceration. Simply, despite the planning and preparation put into correctional policies, mass incarceration has resulted in several deleterious outcomes for inmates, neighborhoods, and society.

Correctional officials and policy makers too often develop policies in the aftermath of highly publicized outrageous crimes. Roger Roots (2004) demonstrated how governmental responses to some social problems are developed in such a hurry, and, as such these policy makers have little chance of thinking through potential unanticipated consequences. Roots identified times when well-intentioned laws may have actually 'backfired' to exacerbate the plight of the poor, single women, elderly, and others for which social policies were directed to help. Criminologists have identified several unanticipated consequences related to mass incarceration. Most of these consequences make it more difficult for the formerly incarcerated to return to society.

Much as premodern forms of punishment sought to 'mark' offenders with physical signs of their criminality, contemporary practices stigmatize offenders in many ways that prevent them from fully participating in society. For instance, felony voting restrictions prevent millions of Americans from voting. Sex offender registration and housing requirements prevent individuals from living in many communities and increase the likelihood of such individuals hiding from treatment parole agents (i.e., absconding). Similarly, having a drug conviction prevents people from acquiring financial aid to attend college. In many states, any felony conviction can permanently block individuals from receiving welfare support, and now many forms of employment require disclosure of convictions and compliance with a background check.

The problem with these sorts of collateral consequences is that they exacerbate the criminogenic effects of incarceration. That is, such policies do not reduce criminal propensities, but rather they create a context in which offenders have blocked opportunities to engage in prosocial lifestyles. The final section of this article introduces readers to ways in which crime control policy makers can incorporate more effective strategies to alleviate the potential for unanticipated negative consequences of incarceration.

Potential Improvements

Correctional policy makers have the opportunity to institute approaches that improve the chances for offenders to succeed in the community. Although Martinson (1974) argued that nothing works in correctional rehabilitation, other researchers have identified principles of effective corrections. MacKenzie (2006) provides one of the more thorough reviews of effective correctional strategies. She focuses on strategies because this is a broader concept including different programs, policies, and interventions. The emphasis of her book is to call attention to how correctional strategies can focus on achieving crime and recidivism reductions, not just punishing offenders. The agenda is to use scientific research

to understand the effect of correctional strategies to foster crime reductions. What sort of things should correctional administrators and policy makers concentrate on to reduce crime? MacKenzie (2006: p. 59) identifies several criminogenic needs – i.e., malleable individual traits related to crime – to be addressed: procriminal attitudes, procriminal associates, impulsivity, weak socialization, risk taking, and deficits in educational, vocational, and employment skills. Research finds that focusing attention on these types of offender characteristics can bring about the greatest impact on crime reductions.

Adopting effective correctional strategies requires administrators, policy makers, and the public to move away from traditional concerns for punishment alone. The mass incarceration movement has fostered overcrowded prisons, removed treatment and training from prisons, and created an environment rife with violence and gang membership. MacKenzie (2006) makes clear that punishment alone is not enough. Instead, research finds that strategies that rely on swift and certain responses to criminality are more effective than simply inflicting pain on offenders. It is ironic that prisons developed as a way to improve the horrible conditions of jails and workhouses, and to end the brutality of physical punishments. But, at least since the mid-1970s, US prisons have shifted away from concerns with *correcting people* to adopt strategies that focus on managing populations of the most disadvantaged in society and warehousing them in environments that make personal improvements nearly impossible. Does anyone suspect that sending people to these overcrowded, violent, and no-frills prisons will result in good citizens? Research makes clear that locking people up and throwing away the key is an ineffective strategy that diverts from the original intentions of imprisonment.

See also: Criminal Justice, Sociology of; Deterrence: Sanction Perceptions; Imprisonment: Sociological Aspects; International and Comparative Criminology; Mass Imprisonment and Its Consequences; Prison Life, Sociology of: Recent Perspectives from the United Kingdom; Punishment: Comparative Politics of; Re-entry and Reintegration After Incarceration; Rehabilitation of the offender.

Bibliography

Cavadino, M., Dignan, J., 2006. Penal Systems: A Comparative Approach. Sage, London.

DeMichele, M., Janoski, T., 2010. A Theory of Cross-National Incarceration Rate Variation: Three Worlds of Western Punishment, 1960–2002. American Sociological Association, Atlanta, Georgia.

Downes, D., 2001. The "macho" penal economy: mass incarceration in the United States – a European perspective. Punishment and Society 3 (1), 61–80.

Downes, D., Hansen, K., 2006. Welfare and imprisonment in comparative perspective. In: Armstrong, S., McAra, L. (Eds.), Perspectives on Punishment. Oxford University Press, Oxford.

Esping-Anderson, G., 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton, NJ.

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Random House, New York.

Garland, D., 2001. The Culture of Control: Crime and Social Order in Contemporary Society. University of Chicago Press, Chicago.

Greenberg, D., 1999. Punishment, division of labor, and social solidarity. In: Laufer, W.S., Adler, F. (Eds.), The Criminology of Criminal Law. Transaction, New Brunswick, pp. 283–361.

International Centre for Prison Statistics, World Prison Brief. http://www.prisonstudies.org/info/worldbrief/index.php (accessed 12.02.13.).

Jacobs, D., Kleban, R., 2003. Political institutions, minorities, and punishment: a pooled cross-national analysis of imprisonment rates. Social Forces 82, 725–755.

Jacobs, D., Jackson, A., 2010. On the politics of imprisonments: a review of systematic findings. Annual Review of Law and Social Science 6, 129–149.

Lacey, N., 2008. The Prisoners' Dilemma: Political Economy and Punishment in Contemporary Democracies. Cambridge University Press, New York.

MacKenzie, D., 2006. What Works in Corrections: Reducing the Criminal Activities of Offenders and Delinquents. Cambridge University Press, New York.

Martinson, R., 1974. What works? Questions and answers about prison reform. The Public Interest 35, 22–54.

Merton, R., 1936. The unanticipated consequences of purposive social action. American Sociological Review 1, 894–904.

Payne, B., DeMichele, M., 2011. Sex offender policies: a case of unanticipated consequences. Aggression and Violent Behavior 16 (3), 177–187.

Rothman, D., 1971. The Discovery of the Asylum: Social Order and the Disorder in the New Republic. Little, Brown, and Co, Boston, MA.

Roots, R., 2004. When laws backfire: unintended consequences of public policy. American Behavioral Scientist 47 (11), 1376–1394.

Rusche, G., Kirchheimer, O., 1968. Punishment and Social Structure. Columbia University Press, New York.

Savelsberg, J., 1994. Knowledge, domination, and criminal punishment. American Journal of Sociology 99, 911–943.

Sutton, J., 2000. Imprisonment and social classification in five common law democracies, 1955–1985. American Journal of Sociology 106 (2), 350–386.

Sutton, J., 2004. The political economy of imprisonment in affluent western democracies, 1960–1990. American Sociological Review 69 (2), 170–189.

Whitman, J., 2003. Harsh Justice: Criminal Punishment and the Widening Divide between American and Europe. Oxford University Press, New York.

Privacy of Individuals in Social Research: Confidentiality

Robert F Boruch, University of Pennsylvania, Philadelphia PA, USA
Ellen A Donnelly, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Boruch, volume 18, pp. 12070–12072, © 2001, Elsevier Ltd.

Abstract

This article covers basic definitions and frameworks for assuring the confidentiality of individuals' records that are generated or acquired in social research. The latter includes observational studies, surveys, and randomized experiments. Good professional practice, ethics codes, and law are all considered. To the extent possible, experience and practice in different countries is also handled.

Introduction

This article begins with basic definitions of privacy, confidentiality, and security. It covers the legal and ethical obligations that researchers have in some countries to establish and maintain responsible 'data stewardship' practices that include assuring individual privacy and the confidentiality of research records. Different ways to minimize degradation of privacy, recognize individuals' rights, and enhance confidentiality of records with identifiable information are described. This article then outlines strategies for making records accessible for statistical analysis while preserving privacy protections. Examples are drawn heavily from experience in the United States, but we recognize related important work in other countries.

We define social research here as efforts to develop generalizable scientific knowledge about a phenomenon or group. We distinguish this activity from research dedicated to making judgments about particular individuals or groups as in the case of administrative decisions in the courts or social support systems, for example.

Definitions

For contemporary and earlier analysts, the terms privacy, confidentiality, and security reflect an amalgam of personal and socio-cultural values. As Acquisti (2004) observes, the fundamental concept of privacy will mean "different things to different people, including the scholars who study it, and raises different concerns at different levels" (p. 150). Privacy in this article broadly refers to a person's right to control the information about himself or herself (e.g., Boruch and Cecil, 1979; Duncan et al., 2011). The idea of privacy engenders respect for an individual's autonomy in sharing personal attitudes or experiences with others and his or her freedom from excessive intrusion.

While privacy concerns individuals' autonomy, confidentiality refers to the status given to information that individuals provide to the researcher. This status involves regulating or prohibiting the disclosure of personal data on an identifiable individual. A recipient of such personal data is usually bound by legal agreement, obligation, or promise to not release information about data providers without the individual's consent unless certain exceptional conditions are met.

Security here pertains to safeguarding physical or electronic records so as to provide protection against unauthorized access and misuse. Security measures might include physical controls (e.g., locks or guards), personnel policies (e.g., background checks), data-system controls (e.g., passwords or encryption), or other special procedures (Goldman and Mulligan, 1996).

An Ethical and Legal Context

Participants in social surveys, randomized field trials, or other research in the behavioral and social sciences entrust personal information to researchers. In contemporary social research, administrative records on identifiable individuals may also be disclosed to researchers in the interest of assisting an organization in improving its functioning and achieving its mission. Generally, researchers or recipients of such data have ethical and legal responsibilities associated with their becoming stewards of these data. The concept of 'data stewardship' refers to a researcher's managing of personal data in an effort to assure individuals' privacy, confidentiality of information, record security, and beneficent use of the data for advancing societal understanding of a phenomenon.

Concerns for responsible data stewardship are reflected in international and professional codes of ethics. For instance, Principle 6 of the United Nation's Fundamental Principles of Official Statistics formalizes the idea that all individual data collected by statistical agencies or organizations must be kept confidential and used exclusively for statistical purposes. The European Commission's Community Research and Development Information Service offers comparable guidance on privacy and confidentiality. In the United States, the Census Bureau distinguishes data stewardship as a formal process that 'goes beyond the law' to ensure respect for privacy and release of high-quality information about the nation's population (U.S. Census Bureau, 2012). Commitments to ethical data stewardship are also reflected and reified in U.S. federal government regulations such as the Federal Policy for the Protection of Human Subjects, known as the 'Common Rule,' and institutionalized review processes such as Institutional Review Boards. In Canada, among other countries,

comparable review institutions are called Research Ethics Boards.

Laws may also govern data stewardship practices and techniques. For example, the U.S. Privacy Act of 1974 establishes procedures for the collection, use, and dissemination of personally identifiable information provided to federal agencies. Since 2002, the Confidential Information Protection and Statistical Efficiency Act extended federal agencies' disclosure protections to all data obtained under a pledge of confidentiality. Similar statutes dictate confidentiality and data access policies for specific agencies like the Census Bureau (e.g., Section 9, title 13 of the U.S. Code) and the National Center for Health Statistics (e.g., Public Health Service Act). The National Center for Education Statistics and education research initiatives must adhere to at least four federal laws guiding student and family information usage (Duncan et al., 2011). Most such laws (or associated legal regulations) stipulate that disclosure of personally identifiable records without an individual's consent is prohibited except under clearly specified conditions such as imminent danger to the individual or to others.

A number of other developed countries have codified government-wide privacy and confidential guarantees. Sweden's Secrecy Act of 2009 and Data Act of 1973 include penalties for failure to meet data stewardship obligations. Wallgren and Wallgren (2007) provide additional references to contemporary pertinent regulations issued by government statistical agencies in Canada, Denmark, the Netherlands, and Finland. Boruch and Cecil (1979) discuss some early history of these reforms.

Assuring Privacy and Confidentiality in Surveys, Randomized Field Trials, and Other Data Collection Efforts

Ethical codes in the social sciences date at least from 1948 with the Code of Marketing and Social Research Practice issued by the European Society for Opinion and Market Research. Analogous codes continue to be produced and periodically revised by statistical, educational, psychological, anthropological, and other social science societies in many developed countries. Such codes have also been adopted or tailored by some for profit and not-for-profit research entities that do international work, e.g., American Institutes for Research.

To satisfy ethical codes as well as legal demands for responsible data stewardship, a researcher must consider privacy and confidentiality issues at each stage in the process of a study. When first designing a study, for instance, the researcher must seek to minimize the information that is asked of a study participant and tailor questions to both meet a study's objectives and minimize intrusiveness. A researcher may also limit the number of people who are engaged in a study so as to achieve a minimum necessary sample size, a professional standard that coincidentally meets an ethical concern about reducing the extent to which people are asked to disclose personal information.

Under many codes of ethics and certain legal regulations, individuals being recruited for or becoming involved with a study must be told whether their participation in the research is voluntary. People usually must be informed of the purpose of the research and the disposition or functions of the data that they provide. Typically, they must be assured that the information they provide will remain confidential, that is, their responses will be used only for research purposes and personally identifiable information will not be disclosed to anyone else.

During and after a study's completion, a researcher is usually obligated to cooperate with courts and other government agencies if such organizations demand information on identifiable individuals who happen to be involved in research. In certain circumstances, however, one can assure data providers of the confidentiality of the information they provide despite government threat or legal action. For instance, the National Institutes of Health requires grantees in the United States who research delinquent activity, illicit substance use, or certain other sensitive behaviors to issue documents called Certificates of Confidentiality. These certificates protect investigators legally against 'involuntary disclosures' of names or identifiable information of their research subjects by judicial or administrative subpoena. That is, the researcher who collects sensitive information cannot be forced to disclose identifiers. The role of such Certificates is discussed in the context of research on children, for example, in reports issued by the Institute of Medicine in the United States (Field and Behrman, 2004) and by Boruch et al. (2012) more generally. We are aware of no similar statutory protection in countries other than the United States, and the Certificates apply only to records maintained in the United States.

Certificates of Confidentiality, however, cannot be treated as ironclad guarantees. In the wake of the 11 September 2001 attacks, the U.S. government created new powers to access personal information. The U.S. PATRIOT Act, for instance, permits the U.S. Attorney General to petition courts for *ex-parte* orders for the release of Department of Education data that could inform investigation or prosecution of domestic or international terrorism, i.e., who is a student in what course of studies. In light of major policy-making events, keeping track of current 'shield laws' protecting investigators and research subjects as well as documenting episodes when assured confidentiality could not be legally guaranteed is important. Boruch et al. (2012) give recent information on vulnerability in the context of the U.S. judiciary.

Similarly, Certificates of Confidentiality do not protect against voluntary disclosures by researchers. When presented with evidence of child abuse, reportable communicable disease, or a subject's threatened violence against one's self or others in a study, researchers may especially wish to convey such information to law enforcement, health and human service departments, or other governmental agencies (e.g., Takas et al., 1995). According to the U.S. Office for Human Research Protection's (2003) guidelines, researchers should indicate in their subjects' consent forms that the researcher may indeed choose to or be required by law to disclose information concerning these subjects. This practice has been widely embraced by the Center for Disease Control and Prevention, National Institute on Drug Abuse, National Institute on Alcohol Abuse and Alcoholism, and other institutions devoted to examining sensitive topics.

Inventive statisticians, economists, and others have also developed useful ways that are useful at times to elicit information from individuals in personal interviews and that do not degrade privacy. As a consequence, they also protect the confidentiality of the information that is provided. One such approach falls under the technical rubric of 'randomized response.'

Randomized response and related approaches involve collecting information that discloses nothing to the researchers about the state of the particular subject or the subject's behavior like use of addictive substances or violations of law. Nevertheless, these strategies, augmented with some elementary statistical arithmetic, produce credible estimates of the incidence of a sensitive trait or behavior in the study. In the simplest cases, each respondent in a sample is asked to randomly choose one of two questions to answer, one of which is innocuous and the other being sensitive. This may be probabilistically based like rolling a die. Which question is answered is not disclosed to the interviewer, nor is the result of the die roll. Rather, the respondent merely gives a 'Yes' or 'No' response. Given simple probability theory and aggregate (not individual) information on incidence for the innocuous question, it is an easy matter to compute the proportion of respondents who have the sensitive characteristic. But there is no way to determine individual status with certainty.

The basic approach and arithmetic of randomized response is given in epidemiology texts such as Levy and Lemeshow (2008). In the Netherlands, Lensvelt-Mulders et al., 2005 summarize recent evidence on its value. Statisticians from Cuba and Spain have also handled the topic (e.g., Bouza et al., 2010). Boruch and Cecil (1979) further review early empirical evidence on the approach's value.

Data Sharing, Privacy, Confidentiality, and Researchers' Use of Statistical Databases

Making local, regional, or national data available to other researchers presents additional challenges to ensuring privacy and confidentiality. Researchers may make their statistical data available so as to permit secondary analysis, confirmation or disconfirmation of the original findings, and for other scientific purposes. Codes of ethics enunciated by the American Educational Research Association (AERA), and the American Psychological Association (APA), among others, have even encouraged the practice of such sharing in the interest of advancing scientific understanding. Yet, sharing data may put two ethical standards in tension with each other. It may be publicly beneficial to release data for statistical analysis. But availability of such data may compromise confidentiality and introduce new possibilities for unauthorized, non-research use of personal records. Section 12 of the AERA (2011) Code and the APA's (2012) Task Force Report recognize the tension and provide advice on its handling.

Developing and using statistical databases benefit society at times. To the extent that the same dataset can be put to work frequently to benefit the public, the cost/effectiveness ratio decreases. For instance, in educational research, scholars have undertaken numerous reanalyses of Tennessee's Project STAR class size experiment data. These studies subsequently informed policy debates about reducing the size of classroom in 20 American states. No confidentiality of records issues emerged over the 20-year course of this activity.

While potentially enriching and cost-effective, new statistical databases are blurring the boundaries between publicly useful and traditionally private information. An increasing number of databases are unrelated to social scientific inquiry and help private companies and corporations gather information about consumers. They may nonetheless be used also by researchers. Duncan et al. (2011) discuss a case in point: Iceland.

In the late 1990s, Iceland's parliament passed legislation permitting a bio-pharmaceutical company to construct a centralized health database of all Icelandic people's genealogical, genetic, and health information. The database linked nearly 80% of Icelanders to a national genealogy going back over 1000 years (Hlodan, 2000).

Throughout the early 2000s, genetic biobank projects proliferated in countries as diverse as Taiwan, Estonia, Tonga, and the United Kingdom. The U.S. National Institute of Health also floated a proposal for a large U.S. genomic databank, but budgetary pressures and strong political opposition have halted rigorous debate on this issue (Greely, 2007). These events underscore the fact that as new types of biological, geospatial, audio and video, and social networking data emerge, politicians and researchers must continually reevaluate the boundaries of public and private information, the characteristics of legitimate social science research, and standing policies that protect the individual's rights to privacy.

Learning how to make individual records available for statistical analysis in ways that do not jeopardize the privacy of the individuals who are the source of the record information is not easy. Since the 1970s, federal statistical agencies in the Nordic countries, North America, and many others have sought to reconcile the competing demands of privacy and of data access. Although most government agencies concern themselves with surveys rather than randomized trials on the effects of social programs, lessons learned from agencies' survey experiences help to inform the sharing data strategies for randomized experiments.

The approaches that have been developed to assure researchers' access to existing datasets without compromising privacy concerns can be classified into two categories. The first involves simple, non-legal procedures that focus on assuring that records cannot be used for learning about an individual. The second involves legal arrangements, including statutes, regulations, and contractual agreements discussed briefly below, Boruch et al. (2012) give more detail.

Consider first statistical approaches to making data available and simultaneously protecting individual privacy. Typically, the data steward who releases a record file for the public's use must delete the direct and clear identifiers of individuals from data files. That is, addresses and other unique information are expunged. The custodian may also choose to release anonymized data on a sample of records, rather than on a complete assembly, so as to enhance confidentiality.

Further, because individual records are often detailed and external collateral information is available, sophisticated statistical approaches may be used to eliminate the possibility of deductive disclosure. Such techniques include deleting of

certain variables from the file, eliminating certain micro-records that allow disclosure on small samples of particular individuals, reducing the precision of entry by making the reporting categories crude, introducing controlled random perturbation to true values, and systematically swapping data entries to sever relationships between records and data providers. These technical approaches are often designated as forms of statistical disclosure (deductive disclosure) limitation technology. Advances in this technology have been made by statisticians in the Netherlands, the United Kingdom, Canada, Israel, and Luxemburg and other countries. References and guidance are given in Duncan et al. (2011), Fienberg and Willenborg (1999); Wallgren and Wallgren (2007).

Other individual protections involve legal arrangements at the institutional level. 'Data enclaves' are organizations created to (1) permit intensive sophisticated analyses by people other than the original data custodian and (2) protect the confidentiality of micro-records with the same rules that govern the original custodian. At some academic and research institutions, for instance, enclaves have been established that permit researchers to analyze data that would be accessible ordinarily only to federal civil servants. The rules that govern the data's use, including security procedures and the threat of criminal and civil sanctions, apply to the researchers as they do to the civil servants who are the formal custodians.

A variant on the enclave idea involves licensing agreements. A specific example of this arrangement, developed by the National Center for Education Statistics in the United States, is a license granted to particular individuals to analyze specific datasets that would not otherwise be accessible on privacy grounds. The justifications for such privilege must be made. The licensing agreement includes requirements for security of the micro-records, sanctions, and other rules for disclosure.

Privacy, Confidentiality, and the Use of State Databases in the United States

In the United States, researchers might desire to access the administrative record systems and databases of state governments. Each of the nation's 50 states has its own laws and regulations governing third party access to such data. In particular, state laws may differ appreciably in their provisions for permitting a researcher (and certain others) to acquire and use individually identifiable micro-records. Within each state, data access policies can also differ across systems like schools, hospitals, child welfare and adult welfare agencies, crime and justice agencies, runaway shelters, or other social service institutions.

Currently, no national or state-based compendiums exist for identifying which administrative data systems are legally open or closed to researchers and specifying the conditions of receiving access. The development of these compendiums for various policy spheres serves as one crucial step in helping researchers understand how to properly request and use sensitive governmental data in their studies. In recent decades, some compendiums chronicling the availability of administrative data have been constructed for special purposes. These compendiums describe the structure of particular states' record systems and summarize access rules for nongovernmental actors. For instance, the Office of Juvenile Justice and Delinquency Prevention Research (1995) details state guidelines for procuring administrative records during searches for missing or abducted children.

The Bureau of Justice Statistics (BJS) has been comprehensive in documenting procedures for obtaining criminal history information. The agency has produced 12 reports since 1972; more recent reports supplant earlier ones. BJS' (2002) compendium gives an overview of each state's laws and regulations regarding (1) requirements for state agencies to maintain criminal processing records and (2) policies for disseminating and using criminal history information for noncriminal justice purposes (p. 2). The full text of these laws and regulations can be found in an electronic repository managed by SEARCH, the National Consortium for Justice Information and Statistics. The compendium further outlines the policies of each state on 29 dimensions.

One such dimension is whether the state has enacted laws or regulations concerning the availability of criminal history records to legitimate researchers. About 36 states have provisions governing research access. The BJS cites this compendium assists administrators and policymakers develop fair and effective information management practices. In its own words, BJS proclaims, "The compendia are intended to promote the evolution of enlightened privacy and information policy," (2003: p. 1). For this reason, replication efforts by other agencies or jurisdictions that detail available governmental data and their regulations are critical and valuable.

Privacy, Record Linkage, and Integrated Data Systems

Record linkage, or combining two or more sets of records referring to the same person from different databases in the interest of research, poses another set of privacy problems. Records are often matched according to individual identifiers like names, addresses, Social Security Numbers, dates of birth, or other sensitive information. Interest in record linkage projects has grown as computer technology has made record-keeping easier and policymakers have increasingly relied on rigorous statistical analyses to evaluate the effects of social programs. Organizations like the MacArthur Foundation have begun supporting several initiatives involving integration of administrative records in education, health, and juvenile justice. Similarly, Actionable Intelligence for Social Policy of the University of Pennsylvania has begun to distinguish areas of best practice with respect to integrated data in its commissioned papers.

According to the U.S. General Accounting Office (U.S. GAO, 2001), several types of privacy issues arise from record linkage projects. First, data custodians must deliberate whether consent to linkage is required from their data subjects. Consent may be required in some circumstances, as record linkage typically falls under the purview of data disclosure laws like US Privacy Act of 1974. If consent to linkage is not necessary or cannot be obtained, data custodians might notify data providers of linkage activity. Agreement on consent, disclosure, and notification procedures, however, may be complicated by differences in data policies between agencies or organizations and nations. This may be especially true for countries with governments that are decentralized.

Most importantly, a linked dataset itself may contain more sensitive or detailed information. The concentration of such

information can present greater security risks. Even after removing clear identifiers, a linked dataset may possess enough sufficient information to make re-identification (also called deductive disclosure) possible. For these reasons, public opinion tends not to favor record linkage. Loss of public support can be detrimental to research initiatives. In 2000, the exposure of a relatively unknown governmental database linking the records of 33.7 million in Canada inspired a public outcry that ultimately led to the demise of the county's largest record linkage project (also see Privacy Commissioner of Canada, 2000).

Nevertheless, the case for linkage has to be made in that researchers must weigh potential benefits of record linkage to society against potential costs. An early example of an effort to discover the benefits and costs of linkage for administrative purposes is given in the U.S. GAO (1986a,b). This study determined that certain administrative linkages are not worth the effort. Recent examples, more to the point for social scientists, are offered in the U.S. GAO (2001).

The Future of Privacy and Confidentiality

Independent of governmental interests in assuring confidentiality of personally identifiable research records and social researchers' interests in that aim, private organizations continue to develop massive integrated data systems. The evolution of Facebook, LinkedIn, and similar internet-based activity presents remarkable challenges. For the researcher, understanding how to assure that people are informed about what he or she plans to do with the public data available on such sites, in the face of people's naïve beliefs that the information they post is private, is a complicated matter. Similarly, researchers who wish to understand how to improve online academic courses by accessing test scores have related challenges if they wish to link test scores with background information on the students' or on the teachers' credentials. The issues, as of this writing, are inchoate. Options for resolving them are currently unclear.

Peer-reviewed journals in relevant disciplines are a continuing resource for learning about issues and options discussed briefly above. For instance, *The Journal of Empirical Research on Human Research Ethics* issues articles on confidentiality protection and people's perceptions of it, on informed consent, in the context of privacy and related matters. *The Journal of Official Statistics*, published under the auspices of Statistics Sweden, covers some technical matters as do others. Short courses and workshops developed for researchers are necessary at times to assure that researchers keep abreast of the state of the art in this sector.

Because laws and codes can prevent valuable research, it is essential to understand how and when to balance interests in confidentiality with the potential or actual benefits of research. The advancement of social science knowledge depends on achieving the balance. The latter embodies the challenges and the future prospects.

See also: Confidentiality and Statistical Disclosure Limitations; Ethical Dilemmas, Research, and Access to Treatment; Ethical Guidelines and Codes of Conduct in Social and Behavioral Research; Informed Consent of Research Subjects; Research Ethics, Cross-Cultural Dimensions of.

Bibliography

Acquisti, A., 2004. Security of personal information and privacy: technological solutions and economic incentives. In: Camp, J., Lewis, R. (Eds.), The Economics of Information Security. Kluwer, Dordrecht, pp. 179–186.

American Educational Research Association, 2011. Code of ethics. Educational Researcher 40 (3), 145–156.

American Psychological Association, 2011. Task Force Report on Data Sharing and Data Linking. American Psychological Association, Washington, DC.

Boruch, R.F., Cecil, J.S., 1979. Assuring the Confidentiality of Social Research Data. University of Pennsylvania, Philadelphia.

Boruch, R.F., Cecil, J.S., Victor, T., Turner, H., Hyatt, J., 2012. Resolving ethical issues in randomized controlled trials. In: Bowne, E., Brown, S. (Eds.), Perspectives on Evaluating Criminal Justice and Corrections: Advances in Program Evaluation, vol. 13. Emerald Group Publishing, United Kingdom, pp. 95–127.

Bureau of Justice Statistics, 2002. Criminal History Record Information: Compendium of State Privacy and Security Legislation: Current Status by Category and State. US Department of Justice, Washington, DC.

Bouza, C.N., Herrera, C., Mitra, P., 2010. A review of randomized responses procedures: the qualitative variable case. Revista Investigación Operacional 31 (3), 240–247.

Cecil, J.S., Wetherington, G.T. (Eds.), 1996. Court ordered disclosure of academic research: a clash of values of science and law. Law and Contemporary Problems, vol. 59 (3), pp. 1–7.

Department of Health and Human Services, 2011. Human subjects research protections: enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators, proposed rule. Federal Register 76 (143), 44512–44531.

Donaldson, M.S., Lohr, K. (Eds.), 1994. Health Data in an Information Age: Use, Disclosure, and Privacy. National Academy Press, Washington, DC.

Duncan, G.T., Elliot, M., Salazar-Gonzalez, J., 2011. Statistical Confidentiality. Springer, New York.

Field, M.J., Behrman, R.E. (Eds.), 2004. Ethical Conduct of Clinical Research Involving Children. Institute of Medicine, Washington, DC.

Fienberg, T., Willenborg, L.C.R.J. (Eds.), 1999. Disclosure limitation methods for protecting the confidentiality of statistical data. Journal of Official Statistics, vol. 14 (4), 337–335.

Greely, H., 2007. The uneasy ethical legal underpinnings of large-scale genomic biobanks. Annual Review of Genomics and Human Genetics 8, 343–367.

Goldman, J., Mulligan, D., 1996. Privacy and Health Information Systems: A Guide to Protecting Patient Confidentiality. Center for Democracy and Technology, Washington, DC.

Hlodan, O., 2000. For Sale: Iceland's Genetic History. http://www.actionbioscience.org/genomics/hlodan.html (accessed 13.05.05.).

Lensvelt-Mulders, G.J.L.M., Hox, J.J., van der Heijden, P.G.M., Maas, C.J.M., 2005. Meta-analysis of randomized response research: thirty-five years of validation. Sociological Methods & Research 33 (3), 319–348.

Levy, P.S., Lemeshow, S., 2008. Sampling of Populations: Methods and Applications, fourth ed. John Wiley & Sons, New York.

National Research Council, 14–15 October 1999. Committee on National Statistics Workshop on Confidentiality of and Access to Research Files. Work Shop Papers. National Research Council, Washington, DC.

Privacy Commissioner of Canada, 2000. Privacy Commissioner Annual Report, 1999–2000. Minister of Public Works and Government Services, Canada, Ottawa.

Soecken, K.L., 1987. Randomized response methodology in health research. Evaluation Health Professions 10 (1), 58–66.

Takas, M., Bass, D. (Eds.), 1995. Addressing Confidentiality of Records in Searches for Missing Children: Final Report. Office of Juvenile Justice and Delinquency Prevention, Washington, DC.

U.S. Census Bureau, 2012. Data Protection and Privacy Policy (accessed 13.05.05.). http://www.census.gov/privacy/data_stewardship/.

U.S. General Accounting Office, 1986a. Computer Matching: Factors Influencing the Agency Decision-making Process (PEMD-87-3BR). USGAO, Washington, DC.

U.S. General Accounting Office, 1986b. Computer Matching: Assessing Its Costs and Benefits (PEMD-87-2BR). USGAO, Washington, DC.

U.S. General Accounting Office, 2001. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information (GAO-01-126SP). USGAO, Washington, DC.

U.S. Office for Human Research Protections, 2003. Guidance on Certificates of Confidentiality (accessed 13.05.05.). http://www.hhs.gov/ohrp/policy/certconf.html.

Wallgren, A., Wallgren, B., 2007. Register-based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Hoboken.

Privacy: Theoretical and Legal Issues

Kirsty Hughes, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Garison, volume 18, pp. 12067–12070, © 2001, Elsevier Ltd.

Abstract

Privacy is a subject that has captured the interest of scholars across a wide range of disciplines resulting in a significant and rich body of literature. Yet despite these efforts we still struggle to understand privacy. It is a nebulous concept with a strong subjective element. There are three central and contested issues which underpin most debates. They are what is the scope of privacy; why do we value privacy; and critiques of privacy. Understanding the complexity of these issues is vital to the development of effective privacy laws.

The Developing Field

Privacy is a subject that has captured the interest of scholars across a wide range of disciplines, resulting in a significant and rich body of literature. Yet despite these efforts we still struggle to understand privacy. It is a nebulous concept with a strong subjective element: we disagree as to when, where, and why we want or need privacy; thus it seems unlikely that there will ever be a universally accepted precise definition of privacy.

Any realistic attempt to understand privacy cannot be limited to purely legal or philosophical 'definitions'. Indeed in 2007 the New Zealand Law Commission noted that "it is regrettable that the concept of privacy has not been treated as multi-disciplinary (encompassing fields of philosophy, political theory, policy studies, historical and cultural studies) as much as it ought to have been in some legal literature" (New Zealand Law Commission, 2007). Whilst there is considerable work still to be done, there is a huge body of cross-disciplinary literature on privacy emerging. Every year numerous conferences bring together researchers from a whole range of disciplines (including economics, computer science, law, philosophy, political science, communication studies, and media studies), as well as policy-makers and practitioners. Such conferences play a key role in facilitating cross-disciplinary discussion as well as comparative and global analysis. This cross-fertilization is crucial. Lawyers cannot operate in a vacuum; they need to know how people behave and why they behave that way, the consequences of interferences with privacy, and the costs of such interferences – the answers to these questions lie beyond the confines of the law. Moreover, modern technology means that the consequences of an interference with privacy may ripple well beyond the boundaries of a particular jurisdiction; for example, where unlawfully obtained photographs are published in overseas publications or on the Internet, such developments raise serious problems for enforcement which render comparative and international research particularly valuable.

Discussions relating to privacy deal with numerous questions, but there are three central and contested issues which underpin most debates. They are what is the scope of privacy; why do we value privacy; and critiques of privacy.

Scope of Privacy

To develop effective privacy laws we need to be able to identify privacy-related interests. Scholars have offered a number of different means of identifying the scope of privacy and privacy-related interests. Traditionally life has been separated into spheres of public life and private life. This is a complex idea that is linked to theories of the proper role of the state. The public/private divide was first used to distinguish between politics which took place in the public sphere and the family which existed in the private sphere – for discussion, see Beate Rössler (2005). The idea that life can be divided into public life and private life is also reflected in privacy scholarship. This separation of public and private spheres emerges in a number of ways. It is found in theories based on private places, private information, and private decisions. Naturally, within each of these models the location of the dividing line between what is public and what is private is contested. Moreover, the very idea that we can distinguish between what is public and what is private is itself heavily contested. Partly as a response to the perceived limitations of these models, broader theories have been developed, which allow for a more subjective element. Rather than simply categorizing interests as 'public' or 'private' these theories have emphasized the subjective nature of privacy, they acknowledge that individuals have different privacy needs and that we do not have universally shared views about what is private and what is not. This sort of approach is often found in access-based models, which are based on the need to be left alone and to control access to ones self. Thus the privacy literature can be broadly categorized into the following types of theories: those that protect private places; those that are based upon managing private information; those that encompass limiting access to self; and those that seek to establish a realm of private decision-making. Scholars may seek to accommodate one or more of these approaches into their conception of privacy. For a critique of these different approaches, see Solove (2002).

One approach is to designate particular locations as private and to identify activities which occur in those places as private and deserving of privacy. This restricts privacy to private spaces. Clearly one of the difficulties with this approach is identifying what constitutes a 'private space'; the home is often regarded as the paradigm private place, but where else is private? A related

objection is that there are some activities which occur in public places which should be regarded as private (Moreham, 2006). Alan Westin identified four types of privacy: solitude, intimacy, anonymity, and reserve; these states of privacy range from experiences of privacy when one is alone to experiences of privacy in public (Westin, 1967). Even if we can identify objectively designated 'private places' the idea that we only seek or experience privacy in those spaces is unrealistic and does not accord with how we behave in practice (Westin, 1967; Altman, 1975; Moreham, 2006; Hughes, 2012).

Decision-making privacy is a model which also derives from the public/private divide and the liberal tradition. Some models of privacy designate the realms of life into which the state should not interfere. This means that there are decisions which one is entitled to take without being dictated to or influenced by the state, and are therefore called decision-making privacy. This type of analysis has been used in the United States where it has developed out of judicial interpretations of the Fourth Amendment, *Olmstead v United States* (1928) 277 US 438; and *Katz v United States* (1967) 389 US 347. For example, in *Griswold v Connecticut* (1965) 381 US 479, the Supreme Court found that prohibiting contraception violated the marital right to privacy and in *Roe v Wade* (1973) 410 US 113, the Supreme Court found that laws prohibiting the right to abortion violated a constitutional right to privacy. Decision-making privacy is attractive for some scholars because privacy becomes a tool for social change. However, decision-making privacy is problematic because it is highly subjective. It may be possible to achieve a degree of consensus within a particular sector of society or within a given society but universal agreement may not be possible. Moreover, even within a society, individuals have different perspectives on the role of the state and on decisions that should be protected as private. It should also be noted that some scholars have suggested that the issues raised by decision-making are more directly related to autonomy than privacy (Gross, 1971; Gavison, 1980).

Coupled with the home, private information is the most readily accepted aspect of privacy. It is generally acknowledged that privacy involves controlling access to, and the use of, private information. The problem is usually not whether private information falls within the scope of privacy per se, but rather what constitutes 'private' information, and when it should be protected. This has become increasingly complex and problematic as new forms of technology have provided numerous ways to obtain, collate, store, and search for information. Information is often designated 'private' by looking at its 'type' or 'form'. Types of information that may be regarded as private include details of an individual's sex life, medical treatment, or salary. Forms of information that may be regarded as private include e-mails, medical records, or diaries. Where things tend to get more tricky is when an individual asserts that other types or forms of information are private because the information is 'personal', meaning that the information relates to an aspect of him or herself that is important to him or her. People reasonably disagree as to what information relates to the individual, and what information it is reasonable to want to withhold. A further problem is determining when information should be protected and when it is no longer private. Can information be considered private after it has been voluntarily published on social networking web sites or blogs? Are the web sites that we browse on the Internet private information? If information has been in the public domain can it be protected as private information in the future? Can public figures expect that their information will be protected as private?

Most of these ideas contribute something useful to the literature. The identification of private places and private information is beneficial as there clearly are some types of information and some places that can usefully be categorized as private. For example, we generally regard our medical records as private. Similarly the home is regarded as the paradigmatic space for privacy and we expect to be able to exclude the state and other people from it , if we so desire. However, these universally shared ideas about privacy do not exhaust our needs for privacy. We are not always concerned with preventing the disclosure of information, nor do we stop needing privacy when we leave our homes. This lacuna is addressed by access-based theories.

The limited capacity of an information-based approach to privacy was highlighted by an example used by Richard Parker (Parker, 1974). Parker described a situation in which an astronaut's bodily functions are monitored by ground control. Parker suggests that even if this technology means that ground control knows exactly what the astronaut is doing, one could still imagine the astronaut switching off the camera whilst he was evacuating feces. Parker suggests that the astronaut's ability has very little effect on his ability to control information about himself, but it would still be a severe loss of privacy if he could not turn off the camera. It is difficult to explain the privacy-related interests which arise in this scenario if one regards privacy as restricted to our interests in controlling access to information. An information-based theory of privacy may seek to explain this example by suggesting that the image of the astronaut evacuating is a form of private information. However, that would involve a dubious use of the term 'private information', which would raise questions about definitional coherency and breadth. More significantly, this type of analysis would neglect the vulnerability of the astronaut. The astronaut's vulnerability stems from the visual element of the intrusion and the feeling of exposure. This sensation is not reflected in an information-based approach to privacy. Nicole Moreham has explored this issue. She suggests that something "less tangible and more 'spiritual' than 'information' is obtained in these situations and an approach based entirely on 'information' therefore fails to accommodate them adequately" (Moreham, 2005).

The difference between information-based approaches and access-based approaches is also crucial when privacy runs up against other competing rights and interests. Privacy often has to be defended against those who argue that "one has nothing to fear if they have nothing to hide"; for a critique of this view, see Solove (2007). Characterizing privacy as a right to control private information renders it vulnerable to these challenges. These debates often involve one party operating on the presumption that the privacy-seeker would only wish to keep information private if the dissemination of that information would have adverse consequences for the privacy-seeker. This presumption is wrong. It assumes that what we are seeking is secrecy and that secrecy is sought to facilitate deception.

However, many of our concerns about privacy have nothing to do with secrecy or deception. We simply wish to have control over our lives. It is not simply that we fear the consequences of exposure. It is that we object to being subjected to the control of others.

Access-based theories recognize that we may legitimately object to activities such as surveillance, even if those activities do not lead to the gathering of harmful information (or in fact any information) and even if they occur in public places. Objections to these types of intrusions do not relate to the private nature of the location, nor do they stem from concerns over protecting the disclosure of private information. This need for privacy is recognized in theories based on controlling access to self; however, those theories have been criticized for their breadth and for their inability to distinguish privacy from other forms of unwanted access.

There are a number of variations of the access-based model. Warren and Brandeis advocated 'the right to be let alone' in an article that has shaped the scope of US privacy law and scholarship (Warren and Brandeis, 1890). Their concerns were triggered by the development of the portable camera and its use by the 'yellow press'. This technological development meant that individuals were vulnerable in ways that they had not been before. Scholars have developed this idea of privacy as the 'right to be let alone' (Allen, 1988; Bok, 1983; Gross, 1967; Van Den Haag, 1971). For example, Ruth Gavison suggests that an invasion of privacy occurs where an individual is subjected to unwanted physical proximity (Gavison, 1980). Nicole Moreham suggests a model based on desired access (Moreham, 2005). These theories of privacy reflect many of our intuitions about privacy. However, there is a danger that we end up with a conception of privacy that is too broad. As Daniel Solove explains "certainly not all access to the self infringes upon privacy, only access relating to specific dimensions of the self or to particular matters and information" (Solove, 2008). Solove is right in that these theories of privacy struggle to identify privacy scenarios. However, it is questionable whether his approach is any better. He has suggested that instead of trying to establish an over-arching theory of privacy, we should identify privacy problems. He has developed a taxonomy which encapsulates the range of privacy problems which have been identified in the existing literature. The taxonomy includes four groups of 'harmful activities': (1) information collection; (2) information processing; (3) information dissemination; and (4) invasion. He breaks down this last category into two types of activities: (1) intrusion; and (2) decisional interference. His description of 'intrusions' describes the type of problems that would emerge from the unwanted access approach. However, Solove offers no means of identifying these 'intrusions'. Hughes suggests that one solution to this problem might be to examine the way in which we experience privacy, and in particular the role that barriers play in establishing privacy, (Hughes, 2012; see also Rickless, 2007; and Altman, 1975).

Value of Privacy

Why do we want to or why should we want to protect privacy? What value or purpose does it serve? The value of privacy is as contested an issue as its scope. The arguments break down to four, sometimes competing and sometimes overlapping, points of view. First there are those that believe that whilst privacy consists of a number of interests, there is nothing special that justifies uniting these interests into a single concept and that privacy is therefore redundant as a concept (reductionist stance); there are others who believe that privacy is underpinned by a single value; still others believe that privacy is underpinned by a cluster of values which benefit the individual; and finally there are those who believe that privacy is beneficial for society.

Judith Jarvis Thomson offered the classic reductionist critique of the right to privacy (Jarvis Thomson, 1975). Reductionist critiques of privacy argue that privacy can be reduced to a number of other interests and that there is no special characteristic that unites these interests. As a result, in any scenario where one might wish to claim a right to privacy, there is actually a more suitable interest or right available and privacy is superfluous. Thomson's argument has been criticized because it requires us to accept the existence of a number of other rights; and a key weakness is her failure to explain why her cluster of rights is preferable to a right to privacy (Scanlon, 1975; Reiman, 1976; Rössler, 2005). Following Thomson's article, a number of scholars sought to respond by identifying 'the' value which underpins privacy. For example Rachels (1975) argued that privacy was essential for intimacy and Reiman (1976) argued that privacy is connected to personhood. However, many of these values appear to have some relationship to privacy, e.g., intimacy, autonomy, dignity, personhood, etc. Thus other theorists have responded to this issue by suggesting that a cluster of values underpins privacy.

Alan Westin identified a cluster of 'functions' of privacy. Those functions were personal autonomy; emotional release; self-evaluation; and limited and protected communication (Westin, 1967). Daniel Solove uses Wittgenstein's family resemblance concept to describe the complexity of privacy and its relationship to a cluster of values (Solove, 2002, 2008). Wittgenstein explained his family resemblance concept by examining the ways in which we use the term 'games'. If we consider different types of games, we see that there is no single characteristic that identifies all games: some use a ball, a board, or pieces; some are played in teams or by individuals; some involve winning or losing; but none of these characteristics is prevalent in all scenarios where we might term something a game (Wittgenstein, 1958). Solove draws upon this idea to argue that the quest for a common denominator leads to greater confusion. Other scholars have also used a cluster approach: see Feldman's human flourishing approach (Feldman, 1994, 1997).

Unlike a number of other rights, for example the right to freedom of expression, privacy has traditionally been seen as an individualistic and even selfish right which affords few (if any) social benefits. A number of scholars have suggested that this approach is too narrow and that the social and group benefits involved in privacy have been rendered invisible as a result. These scholars have highlighted the social value of privacy (Schoeman, 1992; Regan, 1995; Feldman, 1997; Rössler, 2005). The social value of privacy may be harder to comprehend than the social value of other rights and interests; however we need to strive to understand its role.

Critiques of Privacy

Critics call into question the idea that privacy is desirable. They argue that there are other needs and interests which are adversely affected, that the abused child or partner does not want privacy, that in fact what they really want is dignity, autonomy, and freedom from oppression, and that these are the very rights and interests that privacy precludes.

One of the most powerful parts of the literature is the feminist critique of privacy. The feminist critique of privacy highlights the fact that employing concepts such as 'public' and 'private' to identify privacy-related interests is problematic because these concepts are so deeply entrenched in political ideology. These objections are summed up by the feminist slogan 'the personal is political' which reflects the idea that issues which are characterized as 'private' are in fact, and should be seen as, political. This is perhaps most apparent when these concepts are used to claim a sphere of private decision-making, for discussion, see Rössler (2005). Critics claim that the concept of a private life is employed to prevent women from fully engaging in the public sphere and leaves women vulnerable to oppression in the private sphere. These issues are reflected in feminist opposition to the designation of the home as private. Identifying the home as private, shields the home and activities which occur within it from the scrutiny of others, opening up possibilities for domestic violence and other forms of abuse (Allen, 1988; MacKinnon, 1987, 1989; Suk, 2009a,b). Some feminists argue that the private realm is both idealized and undervalued: it is idealized when it is stereotyped as a realm of affection, acceptance, and freedom from pressure; it is undervalued as most of the work conducted in the private realm – child rearing and house work – is unappreciated and is in fact invisible in society (Lacey, 1998). Furthermore, the liberal idea that the private realm is free from state influence is also false as the state often structures private life.

Other scholars have also attacked the value of privacy from a range of perspectives. For example, Amitai Etzioni offers the classic communitarian critique of the right to privacy (Etzioni, 1999). Etzioni does not reject privacy entirely; instead his criticism is targeted at what he perceives as the privileging of privacy over the common good. A further critique of privacy is offered by Richard Posner who has attacked the protection of personal information on the grounds of economic inefficiency (Posner, 1977–78, 1981).

Legal Protection of Privacy

Privacy is protected using a range of different types of laws including criminal law, tort law, data protection laws, constitutional rights, and human rights. Privacy problems may therefore require analysis across a number of different areas of law drawing upon torts, statutory provisions, rights, and regulatory regimes. They may also raise issues across different jurisdictions, for example, where the information is obtained in one jurisdiction and used in another jurisdiction, or where information is published globally.

Privacy laws reflect the broad scope of the concept and may be relevant to a whole range of activities such as stop and search techniques, the use of surveillance technology, the use of DNA and retention of DNA data, access to e-mail and internet usage data, the publication of news stories and photographs, harassment, and the right to contraception or to have an abortion, amongst many others.

The right to privacy receives international recognition under Article 12 of the Universal Declaration of Human Rights, Article 17 of the International Covenant on Civil and Political Rights, and Article 8 of the European Convention on Human Rights (ECHR), which protect the right to respect for private life.

The European Convention on Human Rights is an international treaty which has been ratified by the members of the Council of Europe. Individuals can bring proceedings against their state before the European Court of Human Rights (based in Strasbourg). The Court issues judgments; it can find that the state has violated a Convention right, and it may require the state to pay non-pecuniary damages and costs under just satisfaction.

The Convention does not contain an express right to privacy. Article 8 ECHR protects the right to respect for private life, family life, home, and correspondence. However, as Article 8 ECHR was designed to implement Article 12 of the Universal Declaration of Human Rights (which does refer to privacy), it seems to have been intended that Article 8 ECHR would deal with privacy rights. The Court has also referred to a right to privacy (see, for example, *Armonienė v Lithuania* (App. No. 36 919/02) (2009) 48 EHRR 53); and a number of Parliamentary Assembly resolutions expressly state that Article 8 ECHR guarantees a right of privacy (see, for example, Resolution 1165 (1998) of Parliamentary Assembly of the Council of Europe on the Right to Privacy and Resolution 428 (1970) of Parliamentary Assembly of the Council of Europe a Declaration on Mass Communication Media and Human Right. See also Resolution 1843 (2011) of Parliamentary Assembly of the Council of Europe on the Protection of Privacy and Personal Data on the Internet and Online Media).

Under the Court's Article 8 ECHR jurisprudence, it has held that wiretapping (*Malone v United Kingdom* (1985) 7 EHRR 14); the power to stop and search citizens in the street (*Gillan and Quinton v United Kingdom* (2010) 50 EHRR 45); a state DNA database (*S and Marper v United Kingdom* (2009) 48 EHRR 50); and the installation of a GPS tracking device to a suspect's car, engage the right to privacy (*Uzun v Germany* (2011) 53 EHRR 24). The right to privacy is not absolute under the Convention, thus even where the right is engaged the Court will then have to consider whether the interference may be justified under Article 8(2) ECHR. Under that provision, an interference is justified if the interference is prescribed by law, is in pursuit of a legitimate aim, and is necessary in a democratic society. The legitimate aims are set out in the text of the Convention right itself as follows: national security, public safety, or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Article 8 ECHR also affects the protection afforded by the state when the intruder is a private individual or organization. Thus the European Court of Human Rights has held that the state failed to protect privacy where a magazine published photographs of the applicant (see *Von Hannover v Germany* (2005) 40 EHRR 1) and where a photographer took photographs of a child without the child's parents' consent (see

Reklos and Davourlis v Greece [2009] EMLR 16). It should be noted though that in these cases the European Court of Human Rights, and the domestic courts will have to balance Article 8 ECHR against Article 10 ECHR (the right to freedom of expression). The European Court of Human Rights has laid out a set of criteria for striking an appropriate balance between those two rights in *Von Hannover v Germany* (*No. 2*) (2012) 55 EHRR 15 and *Axel Springer Germany* (2012) 55 EHRR 6. Those criteria are as follows: (1) whether the publication contributes to a debate of general interest; (2) how well known is the person concerned and what is the subject of the report; (3) prior conduct of the person concerned; (4) content, form, and consequences of the publication; and (5) the circumstances in which the photographs were taken.

In the United States privacy has received protection by virtue of a number of provisions, including the Fourth Amendment and the Fourteenth Amendment, as well as a number of privacy torts: the American Law Institute Restatement of Torts followed Prossers classification of four privacy torts, Prosser (1960). Section 8 of the Canadian Charter of Rights and Freedoms has a remit similar to the United States Fourth Amendment. Most Western countries offer some protection to privacy whether it is in the form of constitutional or human rights or through other legal regimes.

One of the key tools in privacy law is the reasonable expectation of privacy test. The reasonable expectation of privacy test was first developed in the United States in the landmark Fourth Amendment phone tapping case – *Katz v United States* (1967). Since *Katz* the test has been the touchstone for Fourth Amendment jurisprudence. In Canada the test is used in Section 8 of the Canadian Charter of Rights and Fundamental Freedoms jurisprudence. The test has also been used in the jurisprudence of the European Court of Human Rights as well as the English law of misuse of private information. The test is used in different contexts in different jurisdictions (Hughes, 2012) and its capacity to deal with the problems that we face is contested (Solove, 2010; Kerr, 2007, 2008; Hughes, 2012).

One of the biggest differences between legal regimes is the balance struck between privacy and competing rights and interests, in particular the right to freedom of expression. Whilst the United States provides strong protection to freedom of expression as a result of the First Amendment, the European Convention on Human Rights and a number of European countries such as the United Kingdom, Germany, and France offer a more balanced approach to the two rights. The tension between these two rights most often arises in a media context, and in particular when deciding whether an injunction should be issued to restrict the publication of a news story, which contains private information or photographs.

Many jurisdictions currently face a number of key challenges including, how to respond to new technology which provides novel means of interfering with privacy, whether legal regimes should be based on an opt-in or opt-out model, to what extent individuals, and in particular children, should be allowed to waive their right to privacy, and how to enforce legal protection in the digital world, particularly against anonymous intruders. To address these concerns we will need to return to the difficult questions examined by privacy theory. Moreover we will need to consider whether existing conceptions of privacy provide us with the necessary tools to address current and future problems or whether these challenges necessitate a reconsideration of how we conceptualize privacy.

See also: Civil Liberties and Human Rights; Domestic Violence: Legal Perspectives; Feminist Legal Theory; Human Rights: Legal Aspects; Law and Technology; Law as an Instrument of Social Change; Law, Sociology of; Law: The Socio-Legal Perspective; Rights: Legal Aspects; Torts in Law.

Bibliography

Allen, Anita, 1988. Uneasy Access: Privacy for Women in a Free Society. Rowman and Littlefield, Totowa, NJ.
Altman, I., 1975. The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding. Brooks/Cole Publishing Co, Pacific Grove, CA.
Bok, Sissela, 1983. Secrets: On the Ethics of Concealment and Revelation. Pantheon Books, New York, NY.
Etzioni, A., 1999. The Limits of Privacy. Basic Books, New York.
Feldman, David J., 1994. Secrecy, dignity, or autonomy? Views of privacy as a civil Liberty. Current Legal Problems 47 (2), 41.
Feldman, David J., 1997. Privacy-related rights: their social value. In: Birks, P. (Ed.), Privacy and Loyalty. Clarendon Press, Oxford, p. 21.
Gavison, Ruth, 1980. Privacy and the Limits of the Law. Yale Law Journal 89 (3), 421.
Gross, Hyman, 1967. The concept of privacy. New York University Law Review 43, 34.
Gross, Hyman, 1971. Privacy and autonomy. In: Pennock, J. Rowland, Chapman, J.W. (Eds.), Nomos XIII: Privacy. Atherton, New York, NY.
Hughes, Kirsty, 2012. A behavioural understanding of privacy and its implications for privacy law. Modern Law Review 75 (5), 806.
Kerr, Ian, 2007. Emanation, snoop dogs and reasonable expectation of privacy. Criminal Law Quarterly 52, 392.
Kerr, Ian, 2008. Tessling on my brain: the future of lie detection and brain privacy in the criminal Justice System. Canadian Journal of Criminology and Criminal Justice, 367.
Lacey, Nicola, 1998. Unspeakable Subjects: Feminist Essays in Legal and Social Theory. Hart Publishing, Oxford.
Levin, Hannah A., Askin, Frank, 1977. Privacy in the courts: law and social reality. Journal of Social Issues 33 (3), 138–139.
MacKinnon, Catherine A., 1987. Privacy v equality: beyond roe v wade. In: MacKinnon, Catherine A. (Ed.), Feminism Unmodified: Discourses on Life and Law. Harvard University Press, Cambridge, MA.
MacKinnon, Catherine A., 1989. Towards a Feminist Theory of the State. Harvard University Press, Cambridge, MA.
Moreham, Nicole A., 2005. Privacy in the common law: a doctrinal and theoretical analysis. Law Quarterly Review 121, 628, 650–651.
Moreham, Nicole A., 2006. Privacy in Public Places. Cambridge Law Journal, 606–635.
New Zealand Law Commission, 2007. A Conceptual Approach towards Privacy. NZLC MP19, Wellington.
Parker, R., 1974. A definition of privacy. Rutgers Law Review 275.
Posner, Richard, 1977–1978. The right of privacy. Georgia Law Review 12, 393.
Posner, Richard, 1981. The economics of privacy. The American Economic Review 71 (2), 405.
Prosser, William L., 1960. Privacy. California Law Review 48 (3), 383.
Rachels, James, 1975. Why privacy is important. Philosophy & Public Affairs 4, 323–33.
Regan, P., 1995. Legislating Privacy: Technology, Social Values and Public Policy. University of North Carolina, Chapel Hill, NC.
Reiman, Jeffrey H., 1976. Privacy, intimacy and personhood. Philosophy & Public Affairs 6 (1), 26.
Rickless, Samuel, 2007. The right to privacy unveiled. San Deigo Law Review 44, 773.
Rössler, Beate, 2005. The Value of Privacy. Polity Press, Cambridge.
Scanlon, Thomas, 1975. Thomson on privacy. Philosophy & Public Affairs 4, 315.
Schoeman, F., 1992. Privacy and Social Freedom. Cambridge University Press, Cambridge.
Solove, Daniel J., 2002. Conceptualizing privacy'. California Law Review 90 (4), 1087–1156.
Solove, Daniel J., 2005–2006. A taxonomy of privacy. University of Pennsylvania Law Review 154, 477.
Solove, Daniel J., 2007. "I've got nothing to hide" and other misunderstandings of Privacy. San Diego Law Review 44, 745.

Solove, Daniel J., 2008. Understanding Privacy. Harvard University Press, Cambridge, MA.
Solove, Daniel J., 2010. Fourth amendment pragmatism. Boston College Law Review 51, 1511.
Suk, Jeannie, 2009a. Is privacy a Woman? Georgetown Law Journal 97, 485.
Suk, Jeannie, 2009b. At Home in the Law: How the Domestic Violence Revolution Is Transforming Privacy. Yale University Press.
Thomson, Judith J., 1975. The right to privacy. Philosophy and Public Affairs 5 (4), 295.
Van Den Haag, Ernest, 1971. On privacy. In: Rowland Pennock, J., Chapman, J.W. (Eds.), Nomos XIII: Privacy. Atherton, New York, NY.
Wacks, Raymond, 1980. The poverty of "privacy". Law Quarterly Review 96, 73.
Warren, Samuel D., Brandeis, Louis D., 1890. The right to privacy. Harvard Law Review, 193.
Westin, Alan F., 1967. Privacy and Freedom. Atheneum, New York, NY.
Wittgenstein, Ludwig, 1958. Philosophical Investigations (G.E.M. Anscombe, Trans.). Basil Blackwell, Oxford.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 19

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 19

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 19 *xliii*

VOLUME 19

Proactive Interference 1
Ian Neath and Aimée M Surprenant

Probabilistic Grammars and their Applications 9
Stuart Geman and Mark Johnson

Probabilistic Thinking 16
Robyn M Dawes

Probability and Chance: Philosophical Aspects 23
Christopher Hitchcock

Probability Density Estimation 29
David W Scott

Probability Theory: Formal 33
J Michael Steele

Probability Theory: Interpretations 37
Teddy Seidenfeld

Probation, Offenders and Risk 43
Peter Raynor

Problem Selection in the Social Sciences: Methodology 49
Wout Ultee

Problem Solving and Reasoning: Case-Based 56
David Leake

Problem Solving and Reasoning, Psychology of 61
Philip N Johnson-Laird

Problem Solving during Infancy and Early Childhood, Development of 68
Hiromitsu Miyata

Procrastination 73
Piers Steel and Katrin B Klingsieck

Procrustes Analysis 79
John C Gower

Pro-Environmental Behavior 82
Deniz S Ones, Brenton M Wiernik, Stephan Dilchert, and Rachael Klein

Professional Development 89
Liz Beddoe

Professions and Professionalization, History of 95
Hannes Siegrist

Professions in Organizations 101
CR (Bob) Hinings

Professions, Sociology of 107
Andrew Abbott

Profits 111
Tony Aspromourgos

Program Evaluation 117
Christiane Spiel, Barbara Schober, and Evelyn Bergsmann

Progress: History of the Concept 123
Dorothy Ross

Projective Methods in Psychology 129
George Stricker and Stephen J Lally

Promotions 135
Michael Bognanno and Moritz Ritter

Propensity Scoring 140
MH Clark

Proper and Common Names: Impairments of Anomia 147
Carlo Semenza

Property: Anthropological Aspects 153
Chris Hann

Property Crime: Victimization and Trends 160
Jan Van Dijk

Property: Legal Aspects 167
Peter S Menell

Property: Legal Aspects of Intergenerational Transmission 173
Ronald Chester

Property Ownership 177
Frank Decker

Property Rights and Urban Planning 184
Barrie Needham

Property Rights, History of 190
Frank Decker

Prophetism 196
Malcolm Hamilton

Prophylaxis in Psychiatry 200
David Castle

Proportional Representation 205
Jack H Nagel

Prosecution 210
Kay Levine and Malcolm Feeley

Prosocial Behavior and Empathy 216
John F Dovidio and Jillian C Banfield

Prosocial Behavior during Adolescence 221
Aerika S Brittian and Marisha L Humphries

Prosocial Behavior during Childhood and Cultural Variations 228
Stuart I Hammond, Whitney Waugh, Emma Satlof-Bedrick, and Celia A Brownell

Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations 233
Jesse Drummond, Whitney E Waugh, Stuart I Hammond, and Celia A Brownell

Prosocial Behavior, Cultural Differences in 238
Daniel J Hruschka and Joseph Henrich

Prosocial Behavior, Effects of Parenting and Family Structure on 244
Liat Hasenfratz and Ariel Knafo

Prosopagnosia 250
Beatrice de Gelder and Jan Van den Stock

Prosopography (Collective Biography) 256
Christophe Charle

Prospect Theory 261
Floris Heukelom

Prospective Memory 270
Michael K Scullin, Hillary G Mullet, Gilles O Einstein, and Mark A McDaniel

Prostitution and Sex Work 279
Jennifer Musto, Crystal A Jackson, and Elena Shih

Prototype Theory in Linguistics 286
John R Taylor

Psychiatric Assessment: Negative Symptoms 290
Hans-Jurgen Möller

Psychiatric Care, Economics of 295
Martin Knapp

Psychiatric Social Work in the United Kingdom 301
Colin Pritchard

Psychiatry and Religion 307
Kate M Loewenthal

Psychiatry: Anthropological Aspects 313
Alfred Kraus

Psychiatry: Informed Consent and Care 320
Laura W Roberts

Psychoanalysis, History of 326
Elke Mühlleitner

Psychoanalysis in Sociology 330
Anthony Elliott

Psychoanalysis: Current Status 336
Allan Tasman

Psychodynamic Therapy: Short-Term 344
Jacques P Barber and Richard F Summers

Psychohistory 349
Rudolph Binion

Psycholinguistics: Overview 353
Brian MacWhinney

Psychological Anthropology 359
Claudia Strauss

Psychological Contracts 366
Simon LD Restubog, Kohyar Kiazad, and Christian Kiewitz

Psychological Treatment, Effectiveness of 372
Michael J Lambert and Kara L Harmon

Psychological Treatments Delivered Nontraditionally 379
Marion K Jacobs and Jim Rue

Psychological Treatments, Empirically Supported 382
William C Sanderson

Psychological Treatments: Randomized Controlled Trials 386
William S Agras

Psychologism, History of 391
Martin Kusch

Psychology of Inferences 394
Janice M Keenan

Psychology of the Real Self: Psychoanalytic Perspectives 400
Roger MK Ng

Psychology, History of (Early Period) 406
Tracy B Henley

Psychology, History of (Twentieth Century) 412
Spencer A McWilliams

Psychometrics 418
Denny Borsboom and Dylan Molenaar

Psychometrics: Classical Test Theory 423
James Algina and Hariharan Swaminathan

Psychometrics: Combinatorial Data Analysis 431
Michael Brusco and Douglas Steinley

Psychometrics: Item Response Theory 436
David Thissen

Psychometrics: Preference Models with Latent Variables 440
Ulf Böckenholt

Psychoneuroimmunology 445
Adam J Massey and Kavita Vedhara

Psychopathy 451
Stuart F White and James R Blair

Psychopharmacotherapy, Costs of 457
J Fritze

Psychopharmacotherapy: Side Effects 462
Nitasha Shetty and Handan Gunduz-Bruce

Psychophysical Theory and Laws, History of 470
Stephen W Link

Psychosocial Work 477
Timothy Sim

Psychosomatic Medicine 484
Santosh K Chaturvedi and Soumya Parameshwaran

Psychotherapy and Hermeneutics 491
Gabriele Chiari

Psychotherapy and Pharmacotherapy 495
Richard Balon and Michelle B Riba

Psychotherapy, History of: Psychiatric Aspects 502
Sidney Bloch and Edwin Harari

Psychotherapy Integration 509
John C Norcross

Psychotherapy Process Research 515
David Orlinsky, Erkki Heinonen, and Armin Hartmann

Public Administration: Organizational Aspects 521
Judy Johnston

Public Broadcasting 526
Michael Tracey

Public Bureaucracies 531
Francis E Rourke

Public Choice Economics 534
Michael C Munger

Public Goods: International 540
Adrienne Héritier

Public Health as Social Science 545
Brian Oldenburg, Fiona Cocker, and Benjamin Schüz

Public Interest 552
Mark Blitz

Public Interest Law: The United States and Beyond 555
Scott L Cummings

Public Opinion: Social Attitudes 562
Edward P Freeland

Public Policy 569
Anke Hassel

Public Policy Schools 576
John W Ellwood and Eugene Smolensky

Public Policy: Ethics 580
Andrzej Klimczuk

Public Psychiatry and Public Health Mental Systems 586
Sokratis Dinos

Public Reason 591
Gerald Gaus and Chad Van Schoelandt

Public Sector Organizations 597
Renate E Meyer and Stephan Leixnering

Public Sphere and the Media 603
Slavko Splichal

Public Sphere: Eighteenth-Century History 610
R Chartier

Public Sphere: Nineteenth and Twentieth-Century History 615
C Calhoun

Publishing as Medium 620
Beth Luey

Publishing: Academic and University Presses 624
Peter J Dougherty

Punishment 631
Claire EM Hamilton

Punishment: Comparative Politics of 636
Robert C Fried

Punishment: Social and Legal Aspects 655
Jonathan Simon

Putnam, Hilary (1926–) 662
Maria Baghramian

Qualitative Methodology in Developmental Psychology 668
Carolin Demuth and Günter Mey

Qualitative Methods, History of 676
John R Hall

Qualitative Methods in Geography 682
Isabel Dyck

Qualitative Research 685
Joseph A Maxwell and L Earle Reybold

Quality of Government 690
Bo Rothstein and Jan Teorell

Quantification in History 695
Konrad H Jarausch and Peter A Coclanis

Quantification in the History of the Social Sciences 700
Theodore M Porter

Quantifiers 707
David Gil

Quantile Regression 712
Roger Koenker

Quantitative Cross-national Research Methods 719
Gosta Esping-Andersen and Adam Przeworski

Quantitative Linguistics 725
Stefan Th Gries

Quantitative Methods and Crime 733
David McDowall

Quasi- and Field Experiment 736
Alex Díaz, María Jiménez-Buedo, and David Teira

Quasi-Experimental Designs 742
Kelly Hallberg and Jared Eno

Queer Anarchism 747
Jamie Heckert, Deric Shannon, and Abbey Willis

Queer Anthropology 752
Cymene Howe

Queer Geography 759
Andrew Gorman-Murray and Scott McKinnon

Queer Theory and Intersectionality 765
Notisha Massaquoi

Questionnaires: Cognitive Approaches 771
Norbert Schwarz

Questionnaires: Collection of Subjective Expectations Data 776
Charles F Manski

Quetelet, Adolphe (1796–1874) 781
Ron Lesthaeghe

Queues 784
Carmen Armero and Susie Bayarri

Race and Academic Motivation 790
Sandra Graham

Race and Racism in the Twenty-First Century 796
Yolanda T Moses

Race and Racism, Geography of 801
James A Tyner and Joshua Inwood

Race and the Law 807
Richard Ford

Race: Ethnicity and Health 812
Nicelma J King and Carolyn B Murray

Race: Ethnicity and Urban Planning 820
June M Thomas

Race: Genetic Aspects 825
Guido Barbujani

Race Identity 833
Clifford L Broman

Race: Prespectives and Crime 837
Michael J Leiber and Jennifer H Peck

Racial Relations 845
Joe R Feagin

Racism, History of 852
GM Fredrickson

Racism, Sociology of 857
Matthew Clair and Jeffrey S Denis

Racism: Social Psychological Perspectives 864
Martha Augoustinos and Danielle Every

Radical Social Work 870
Hilde Bloemen

Radio as a Medium 877
Joy E Hayes

Random Assignment: Implementation in Complex Field Settings 883
Robert G St. Pierre

Rape and Sexual Coercion 887
Peggy Reeves Sanday

Rational Choice and Organization Theory 892
Mary A Zey

Rational Choice Explanation: Philosophical Aspects 896
Russell Hardin

Rational Choice Theory in Sociology 901
Vincent Buskens

Rationality in Society 907
Andreas Flache and Jacob Dijkstra

Rawls, John (1921–2002) 913
Elaine Kellman

CONTRIBUTORS TO VOLUME 19

Andrew Abbott
University of Chicago, Chicago, IL, USA

William S. Agras
Stanford University, Stanford, CA, USA

James Algina
University of Florida, Gainesville, FL, USA

Carmen Armero
Universitat de València, Spain

Tony Aspromourgos
School of Economics, The University of Sydney, Sydney, NSW, Australia

Martha Augoustinos
School of Psychology, University of Adelaide, Adelaide, SA, Australia

Maria Baghramian
School of Philosophy, University College Dublin, Dublin, Ireland

Richard Balon
Wayne State University School of Medicine, Detroit, MI, USA

Jillian C. Banfield
Yale University, New Haven, CT, USA

Jacques P. Barber
Adelphi University, Garden City, NY, USA

Guido Barbujani
University of Ferrara, Ferrara, Italy

Susie Bayarri
Universitat de València, Spain

Liz Beddoe
School of Counselling, Human Services and Social Work, Faculty of Education, University of Auckland, Auckland, New Zealand

Evelyn Bergsmann
University of Vienna, Vienna, Austria

Rudolph Binion
Brandeis University, Waltham, MA, USA

James R. Blair
National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Mark Blitz
Claremont McKenna College, Claremont, CA, USA

Sidney Bloch
St Vincent's Hospital, University of Melbourne, Melbourne, VIC, Australia

Hilde Bloemen
KHLeuven, Belgium

Ulf Böckenholt
Northwestern University, Evanston, IL, USA

Michael Bognanno
Temple University, Philadelphia, PA, USA

Denny Borsboom
University of Amsterdam, Amsterdam, The Netherlands

Aerika S. Brittian
University of Illinois at Chicago, Chicago, IL, USA

Clifford L. Broman
Michigan State University, East Lansing, MI, USA

Celia A. Brownell
University of Pittsburgh, Pittsburgh, PA, USA

Michael Brusco
Florida State University, Tallahassee, FL, USA

Vincent Buskens
Utrecht University, Utrecht, The Netherlands

C. Calhoun
New York University, New York, NY, USA

David Castle
University of Melbourne, Melbourne, VIC, Australia

Christophe Charle
CNRS, Paris, France

R. Chartier
Ecole des Hautes Etudes en Sciences Sociales, Paris, France

Santosh K. Chaturvedi
National Institute of Mental Health and Neurosciences, Bangalore, India

Ronald Chester
New England School of Law, Boston, MA, USA

Gabriele Chiari
School of Specialization in Constructivist Psychotherapy, CESIPc, Florence, Italy

Matthew Clair
Harvard University, Cambridge, MA, USA

M.H. Clark
University of Central Florida, Orlando, FL, USA

Fiona Cocker
Melbourne School of Population and Global Health, University of Melbourne, VIC, Australia

Peter A. Coclanis
University of North Carolina at Chapel Hill, NC, USA

Scott L. Cummings
UCLA School of Law, Los Angeles, CA, USA

Robyn M. Dawes[†]
Carnegie Mellon University, PA, USA

Frank Decker
Sydney, NSW, Australia

Beatrice de Gelder
Brain and Emotion Laboratory Leuven (BELL), Leuven, Belgium; Faculty of Psychology and Neuroscience, Maastricht Brain Imaging Centre M-BIC, Maastricht University, Maastricht, The Netherlands; and Cognitive and Affective Neuroscience Laboratory, CoRPS—Center of Research on Psychology in Somatic diseases, Tilburg University, Tilburg, The Netherlands

Carolin Demuth
Aalborg University, Aalborg, Denmark

Jeffrey S. Denis
McMaster University, Hamilton, ON, Canada

Alex Díaz
UNED, Madrid, Spain

Jacob Dijkstra
University of Groningen, Groningen, The Netherlands

Stephan Dilchert
Baruch College, The City University of New York, New York, NY, USA

Sokratis Dinos
School of Health, BPP University, London, UK

Peter J. Dougherty
Princeton University Press, Princeton, NJ, USA

John F. Dovidio
Yale University, New Haven, CT, USA

Jesse Drummond
University of Pittsburgh, Pittsburgh, PA, USA

Isabel Dyck
Queen Mary University of London, London, UK

Gilles O. Einstein
Furman University, Greenville, SC, USA

Anthony Elliott
University of South Australia, Adelaide, SA, Australia

John W. Ellwood
University of California, Berkeley, CA, USA

Jared Eno
American Institutes for Research, Chicago, IL, USA

Gosta Esping-Andersen
Universitat Pompeu Fabra, Barcelona, Spain

Danielle Every
Appleton Institute, CQUniversity, Wayville, SA, Australia

Joe R. Feagin
Texas A&M University, College Station, TX, USA

Malcolm Feeley
University of California, Berkeley, Berkeley, CA, USA

Andreas Flache
University of Groningen, Groningen, The Netherlands

Richard Ford
Stanford Law School, Stanford, CA, USA

G.M. Fredrickson
Stanford University, Stanford, CA, USA

Edward P. Freeland
Survey Research Center, Princeton University, Princeton, NJ, USA

Robert C. Fried
University of California, Los Angeles, Los Angeles, CA, USA

[†] Deceased.

J. Fritze
Association of Private Health Insurers, Cologne, Germany

Gerald Gaus
University of Arizona, Tucson, AZ, USA

Stuart Geman
Brown University, Providence, RI, USA

David Gil
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Andrew Gorman-Murray
University of Western Sydney, Penrith, NSW, Australia

John C. Gower
The Open University, Milton Keynes, UK

Sandra Graham
University of California, Los Angeles, CA, USA

Stefan Th Gries
University of California, Santa Barbara, CA, USA

Handan Gunduz-Bruce
Yale University School of Medicine, New Haven, CT, USA

John R. Hall
University of California, Davis, Davis, CA, USA

Kelly Hallberg
American Institutes for Research, Chicago, IL, USA

Claire E.M. Hamilton
Queen's University Belfast, Belfast, Northern Ireland, UK

Malcolm Hamilton
University of Reading, Reading, UK

Stuart I. Hammond
University of Pittsburgh, Pittsburgh, PA, USA

Chris Hann
Max Planck Institute for Social Anthropology, Halle, Germany

Edwin Harari
St Vincent's Hospital, University of Melbourne, Melbourne, VIC, Australia

Russell Hardin
New York University, New York, NY, USA

Kara L. Harmon
Grand Junction VAMC, Grand Junction, CO, USA

Armin Hartmann
Universitätsklinikum Freiburg, Freiburg, Germany

Liat Hasenfratz
The Hebrew University of Jerusalem, Jerusalem, Israel

Anke Hassel
Hertie School of Governance, Berlin, Germany

Joy E. Hayes
University of Iowa, Iowa City, IA, USA

Jamie Heckert
Anarchist Studies Network, UK

Erkki Heinonen
National Institute for Health and Welfare, Helsinki, Finland

Tracy B. Henley
Texas A & M University – Commerce, Commerce, TX, USA

Joseph Henrich
University of British Columbia, Vancouver, BC, Canada

Adrienne Héritier
Robert Schuman Centre for Advanced Studies, European University Institute, Italy

Floris Heukelom
Nijmegen School of Management, Radboud University Nijmegen, Nijmegen, The Netherlands

C.R. (Bob) Hinings
University of Alberta, Edmonton, AB, Canada

Christopher Hitchcock
California Institute of Technology, Pasadena CA, USA

Cymene Howe
Rice University, Houston, TX, USA

Daniel J. Hruschka
School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

Marisha L. Humphries
University of Illinois at Chicago, Chicago, IL, USA

Joshua Inwood
University of Tennessee at Knoxville, Knoxville, TN, USA

Crystal A. Jackson
John Jay College of Criminal Justice-CUNY, New York, NY, USA

Marion K. Jacobs
University of California, Los Angeles, CA, USA

Konrad H. Jarausch
University of North Carolina at Chapel Hill, NC, USA

María Jiménez-Buedo
UNED, Madrid, Spain

Philip N. Johnson-Laird
Princeton University, Princeton, NJ, USA; and New York University, New York, NY, USA

Mark Johnson
Brown University, Providence, RI, USA

Judy Johnston
Centre for Management and Organization Studies, University of Technology, Sydney, NSW, Australia

Janice M. Keenan
University of Denver, Denver, CO, USA

Elaine Kellman
King's College London, London, UK

Kohyar Kiazad
Faculty of Business and Economics, Monash University, Caulfield, VIC, Australia

Christian Kiewitz
School of Business Administration, University of Dayton, Dayton, OH, USA

Nicelma J. King
University of California, Davis, Davis, CA, USA

Rachael Klein
Korn Ferry, Minneapolis, MN, USA

Andrzej Klimczuk
Warsaw School of Economics, Warszawa, Poland

Katrin B. Klingsieck
Faculty for Arts and Humanities, University of Paderborn, Paderborn, Germany

Ariel Knafo
The Hebrew University of Jerusalem, Jerusalem, Israel

Martin Knapp
Personal Social Services Research Unit, London School of Economics and Political Science, London, UK

Roger Koenker
University of Illinois, Urbana, IL, USA

Alfred Kraus
Centre for Psychosocial Medicine, University Hospital for General Psychiatry, Heidelberg, Germany

Martin Kusch
University of Vienna, Wien, Austria

Stephen J. Lally
American School of Professional Psychology, Argosy University, Washington, DC, USA

Michael J. Lambert
Brigham Young University, Provo, UT, USA; and Queensland University, Brisbane, QLD, Australia

David Leake
Indiana University, Bloomington, IN, USA

Michael J. Leiber
Department of Criminology, University of South Florida, Tampa, FL, USA

Stephan Leixnering
WU Vienna University of Economics and Business, Vienna, Austria

Ron Lesthaeghe
Free University, Brussels, Belgium

Kay Levine
Emory Law School, Atlanta, GA, USA

Stephen W. Link
University of California, San Diego, La Jolla, CA, USA; and McMaster University, Hamilton, ON, Canada

Kate M. Loewenthal
Royal Holloway, University of London, Egham, UK; and Glyndwr University, Wales, London, UK

Beth Luey
Arizona State University, Fairhaven, MA, USA

Brian MacWhinney
Carnegie Mellon University, Pittsburgh, PA, USA

Charles F. Manski
Northwestern University, Evanston, IL, USA

Notisha Massaquoi
Women's Health in Women's Hands Community Health Centre, Faculty of Social Work, Ryerson University, Toronto, ON, USA

Adam J. Massey
Faculty of Medicine and Health, School of Clinical Sciences, Queen's Medical Centre, University of Nottingham (UK), Nottingham, UK

Joseph A. Maxwell
George Mason University, Fairfax, VA, USA

Mark A. McDaniel
Washington University in St. Louis, St. Louis, MO, USA

David McDowall
School of Criminal Justice, University at Albany—SUNY, Albany, NY, USA

Scott McKinnon
University of Western Sydney, Penrith, NSW, Australia

Spencer A. McWilliams
California State University San Marcos, San Marcos, CA, USA

Peter S. Menell
University of California, Berkeley, CA, USA

Günter Mey
University of Applied Sciences, Hochschule Magdeburg-Stendal, Hansestadt Stendal, Germany

Renate E. Meyer
WU Vienna University of Economics and Business, Vienna, Austria

Hiromitsu Miyata
Okanoya Emotional Information Project (OEIP), Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency (JST), Nagoya, Japan; and Graduate School of Information Science, Nagoya University, Nagoya, Japan

Dylan Molenaar
University of Amsterdam, Amsterdam, The Netherlands

Hans-Jurgen Möller
Ludwig Maximilian University, Munich, Germany

Yolanda T. Moses
University of California Riverside, Riverside, CA, USA

Elke Mühlleitner
Giessen, Germany

Hillary G. Mullet
Duke University, Durham, NC, USA

Michael C. Munger
Duke University, Durham, NC, USA

Carolyn B. Murray
University of California Riverside, Riverside, CA, USA

Jennifer Musto
Wellesley College, Wellesley, MA, USA

Jack H. Nagel
University of Pennsylvania, Philadelphia, PA, USA

Ian Neath
Memorial University of Newfoundland, St. John's, NL, Canada

Barrie Needham
Radboud University, Nijmegen, The Netherlands

Roger M.K. Ng
Kowloon Hospital, Kowloon, Hong Kong

John C. Norcross
University of Scranton, Scranton, PA, USA

Brian Oldenburg
Melbourne School of Population and Global Health, University of Melbourne, VIC, Australia

Deniz S. Ones
University of Minnesota, Minneapolis, MN, USA

David Orlinsky
University of Chicago, Chicago, IL, USA

Soumya Parameshwaran
Kasturba Medical College, Mangalore, India

Jennifer H. Peck
School of Criminology and Criminal Justice, Florida Atlantic University, Boca Raton, FL, USA

Theodore M. Porter
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Colin Pritchard
Bournemouth University, Bournemouth, UK; and University of Southampton, Southampton, UK

Adam Przeworski
New York University, New York, NY, USA

Peter Raynor
Swansea University, Swansea, UK

Simon L.D. Restubog
Research School of Management, College of Business and Economics, The Australian National University, Canberra, ACT, Australia

L. Earle Reybold
George Mason University, Fairfax, VA, USA

Michelle B. Riba
University of Michigan, Ann Arbor, MI, USA

Moritz Ritter
Temple University, Philadelphia, PA, USA

Laura W. Roberts
University of New Mexico, Albuquerque, NM, USA

Dorothy Ross
Johns Hopkins University, Baltimore, MD, USA

Bo Rothstein
The Quality of Government Institute, University of Gothenburg, Gothenburg, Sweden

Francis E. Rourke
Johns Hopkins University, Baltimore, MD, USA

Jim Rue
Laguna Beach, CA, USA

Peggy Reeves Sanday
University of Pennsylvania, Philadelphia, PA, USA

William C. Sanderson
Hofstra University, Hempstead, NY, USA

Emma Satlof-Bedrick
University of Pittsburgh, Pittsburgh, PA, USA

Barbara Schober
University of Vienna, Vienna, Austria

Benjamin Schüz
School of Psychology, University of Tasmania, TAS, Australia

Norbert Schwarz
Department of Psychology, University of Southern California, Los Angeles, CA, USA

David W. Scott
Rice University, Houston, TX, USA

Michael K. Scullin
Emory University School of Medicine, Atlanta, GA, USA

Teddy Seidenfeld
Carnegie Mellon University, Pittsburgh, PA, USA

Carlo Semenza
Università di Padova, Padova, Italy; and I.R.C.C.S. Ospedale S. Camillo, Lido di Venezia, Italy

Deric Shannon
Emory University, Oxford, GA, USA

Nitasha Shetty
Yale University School of Medicine, New Haven, CT, USA

Elena Shih
Brown University, Providence, RI, USA

Hannes Siegrist
Universität Leipzig, Leipzig, Germany

Timothy Sim
The Hong Kong Polytechnic University, The Hong Kong Special Administrative Region, People's Republic of China

Jonathan Simon
School of Law, UC Berkeley, Berkeley, CA, USA

Eugene Smolensky
University of California, Berkeley, CA, USA

Christiane Spiel
University of Vienna, Vienna, Austria

Slavko Splichal
University of Ljubljana, Ljubljana, Slovenia

Robert G. St. Pierre
Abt Associates Inc., Breckenridge, CO, USA

Piers Steel
Haskayne School of Business, University of Calgary, Calgary, AB, Canada

J. Michael Steele
University of Pennsylvania, Philadelphia, PA, USA

Douglas Steinley
University of Missouri-Columbia, Columbia, MO, USA

Claudia Strauss
Pitzer College, Claremont, CA, USA

George Stricker
American School of Professional Psychology, Argosy University, Washington, DC, USA

Richard F. Summers
University of Pennsylvania, Philadelphia, PA, USA

Aimée M. Surprenant
Memorial University of Newfoundland, St. John's, NL, Canada

Hariharan Swaminathan
University of Connecticut, Storrs, CT, USA

Allan Tasman
School of Medicine, University of Louisville, Louisville, KY, USA

John R. Taylor
University of Otago, Dunedin, New Zealand

David Teira
UNED, Madrid, Spain

Jan Teorell
Lund University, Lund, Sweden

David Thissen
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

June M. Thomas
University of Michigan, Ann Arbor, MI, USA

Michael Tracey
University of Colorado at Boulder, Boulder, CO, USA

James A. Tyner
Kent State University, Kent, OH, USA

Wout Ultee
Radboud University Nijmegen, Nijmegen, The Netherlands

Jan Van den Stock
Brain and Emotion Laboratory Leuven (BELL), Leuven, Belgium

Jan Van Dijk
Tilburg University, Tilburg, The Netherlands

Chad Van Schoelandt
University of Arizona, Tucson, AZ, USA

Kavita Vedhara
Faculty of Medicine and Health, School of Medicine, University of Nottingham (UK), Nottingham, UK

Whitney Waugh
University of Pittsburgh, Pittsburgh, PA, USA

Stuart F. White
National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Brenton M. Wiernik
University of Minnesota, Minneapolis, MN, USA

Abbey Willis
University of Connecticut, Storrs, CT, USA

Mary A. Zey
University of Texas, San Antonio, TX, USA

This page intentionally left blank

Proactive Interference

Ian Neath and Aimée M Surprenant, Memorial University of Newfoundland, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Proactive interference (PI) is when something learned earlier impairs the ability to remember something learned more recently. PI can be observed in all subject populations and in all memory systems, including the perceptual representation system, procedural memory, semantic memory, long-term episodic memory, and short-term/working memory. One current theoretical account of PI attributes its cause to cue overload: the more items a cue subsumes, the less effective the cue will be. Other accounts suggest that items learned earlier can reduce the distinctiveness of current items, making successful retrieval less likely.

Proactive interference (PI) occurs when something learned earlier impairs a person's ability to remember something learned more recently. For example, if you change phone numbers, memory for your previous number can make remembering your new number more difficult. A related phenomenon, retroactive interference, is similar except that it is when something learned recently impairs a person's ability to remember something learned earlier. For example, learning your new telephone number can make remembering your previous number more difficult.

Historical Overview

PI has been well known since the beginnings of scientific research on memory. For example, in 1889, Hugo Münsterberg described an example based on his own observations after changing the pocket in which he carried his watch. When asked for the time, he would reach in the old pocket for the watch, not the new pocket. In 1894, Müller and Schumann formally defined PI as follows: "When any two items, as *A* and *B*, have been associated, it is more difficult to form an association between either and third item, *K*" (McGeoch, 1942: p. 402).

The Müller-Schumann Law of Associative Inhibition received substantial empirical support as part of what became known as Interference Theory (for a review of the theory and supporting data, see Crowder, 1976; *see also* Retroactive Interference). The prototypical design is outlined in Table 1 and has two phases each of which uses paired associates learning. In this task, the subject is asked to learn a particular novel pairing of two stimuli, such as two nonsense words or two real words. For example, in Phase I, a person might see the word pairs DOG-ROOF, CAR-DESK, HOUSE-MUG, and so on. The stimuli in Phase I are often represented by the term A-B, with A standing for the first item, or cue, and B standing for the second item, or target. At test, the subjects are given the first item (A) and are asked to provide the second item (B). PI is observed in Phase II, when people are asked to learn more items. The control group is asked to learn a second list that contains entirely new items, such as PHONE-COFFEE, LAMP-WALL, referred to as C-D pairs. The use of the letters C and D indicate that the lists are different from those in Phase I. The experimental group is asked to learn a list that reuses the A items but pairs them with new, D, items such as DOG-COFEE, CAR-WALL. Note that in Phase I, DOG was paired with ROOF but in Phase II, DOG is now paired with COFFEE. At test, the experimental group does worse than the control group because the B items studied in Phase I interfere with memory for the D items in Phase II.

Table 1 A typical AB–AD design to produce proactive interference (PI). Recall of the D items is worse in the experimental group than in the control group

Group	*Time 1*	*Time 2*	*Test*
Control	A–B	C–D	C–D
Experimental	A–B	A–D	A–D

Of course, PI can be observed in more simple designs, as noted by Underwood (1957). He pointed out that in most memory experiments, subjects are given a list, are asked to recall it, and are then given another list, which they recall; this sequence is repeated many times. What if we were to plot percent correct on a single list, as opposed to averaged over lists, and also indicate how many previous lists had been experienced? Figure 1 reproduces his results based on numerous studies: As the number of prior lists increased, the level of recall decreased. (In Underwood's (1957) original figure, each study was plotted separately. In the figure shown here, when there are two or more studies for a particular number of previous lists, the mean of those values is used.)

PI and Multiple Memory Systems

Currently, the dominant theoretical account of memory divides it into multiple different systems, with the most popular formulation being the five memory systems described by Tulving and his colleagues (e.g., Tulving, 1985; for a review, see Surprenant and Neath, 2009). These are often further classified as either nondeclarative memory systems (the perceptual representation system and procedural memory) or declarative memory systems (semantic, episodic, and short-term/working memory). Although there is no question that PI is observable in the first four of the hypothesized memory systems, there is substantial disagreement about whether it occurs in certain

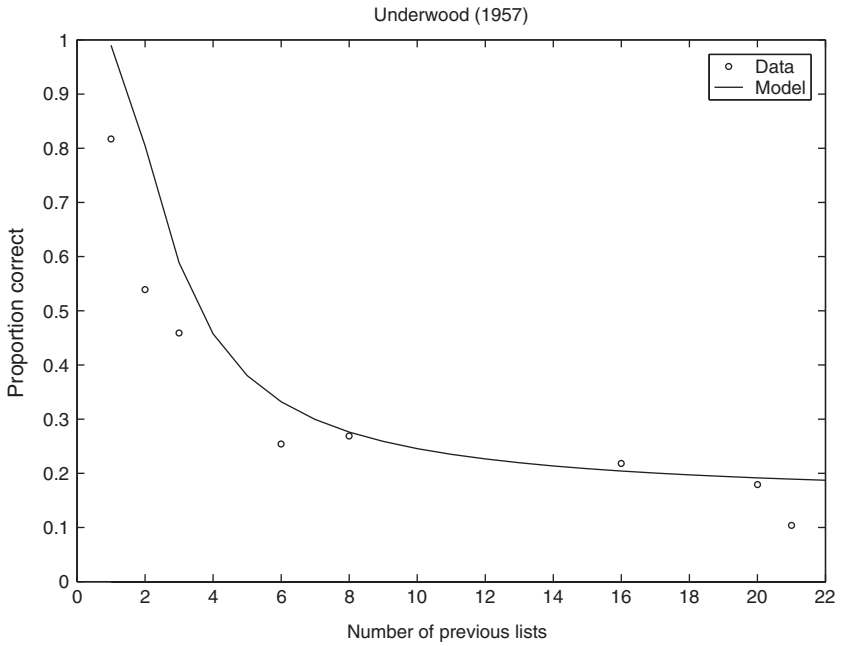

Figure 1 Proportion correct recall of a list of items as a function of the number of previous trials. The data points are from Underwood (1957), the line shows the prediction of the model SIMPLE, described below. Source: Brown, G.D.A., Neath, I., Chater, N., 2007. A temporal ratio model of memory. Psychological Review 114, 539–576. Adapted with permission of APA.

types of working memory. Below we describe representative studies that demonstrate PI in each memory system in turn.

PI and Nondeclarative Memory

The first major division of memory is between nondeclarative Non declarative (procedural) memory and declarative memory systems. The former is concerned with knowing 'how' to do something (for example, how to ride a bicycle), whereas the latter is concerned with knowing 'that' something is the case (for example, that George Washington was the first president of the United States). There are two types of nondeclarative memory, the perceptual representation system and procedural memory. PI can be observed in both.

PI in the Perceptual Representation System

The perceptual representation system consists of a collection of domain-specific modules that operate on perceptual information about the form and structure of words and objects. Much of the data used to argue for the existence of the perceptual representation system as a separate memory system comes from studies of priming, in which certain stimuli are seen at Time 1 and then their influence is measured by subsequent performance at Time 2. Importantly, the typical task uses incidental learning (there is usually a cover task so that the subject is unaware that there will be a later or subsequent test of the material) and also uses indirect measures of memory (there is no direct mention that the test is related to the learning episode). This type of memory is often referred to as implicit memory (*see* Implicit Memory).

Lustig and Hasher (2001) demonstrated PI using a word-fragment completion task, a task commonly used to evaluate the perceptual representation system. The baseline group of subjects saw word fragments, such as *a _ l _ _ g y*, and were asked to complete them to form a valid word. This provides a baseline of how many times a particular fragment was completed with a particular word. There were two additional groups, both of which participated in two phases. In Phase I, the subjects saw a list of words and were asked to indicate the number of vowels that occurred in each word. This task induces processing of individual letters, but is unlikely to induce the subjects to actively memorize the stimuli for a subsequent test. There was then a pause, during which several additional tasks unrelated to the experiment occurred. The purpose was to make it less likely that the subjects would link Phase I and Phase II. Finally, in Phase II, the key tests were presented. Both experimental groups received word fragments and were asked to complete them. The difference between the interference and noninterference groups was the relation between the fragments shown at test and the words seen earlier in Phase I. For example, the noninterference group might receive the word fragment *a _ l _ _ g y* at test, and would also have seen the word *allergy* in the original list. The interference group also saw *allergy* in the original list, but in addition, they saw a potentially interfering word such as *analogy*. Although both groups showed priming (they completed the fragment more often than the baseline group), there was less priming in the interference condition than in the noninterference condition.

It should be noted that the stimuli need not necessarily be verbal in order to observe PI in the perceptual representation system; for example, Köster et al. (2002) demonstrated PI in implicit memory for odors.

PI in Procedural Memory

Procedural memory is a memory system responsible for basic associations, motor learning, and simple cognitive skills. PI is easily demonstrated in tasks thought to tap this system. For example, Cook (1941) used a mirror drawing task in which the subject is asked to trace a shape shown in a mirror. One block of trials had the mirror reflecting one way, whereas the second block of trials changed the reflection. Cook found that after practicing five trials, performance (in terms of accuracy) dropped substantially when the mirror was changed. Similar results have been found for other motor skills. One of the more notable is the study by Fitts and Jones (1961) analyzing crashes in World War II aircraft. They found that accidents were especially likely when pilots switched between the B-25, C-47, and C-82 aircraft due to differences in the motor movements needed to operate the aircraft.

Declarative Memory

Declarative memory is divided into two major types, semantic memory (*see* Semantic Memory) and episodic memory (*see* Episodic Memory). One key distinction between these two systems is whether the remembering person is aware of the learning episode. For example, many people know that George Washington was the first president of the United States, but they are unaware of when they learned this and are unaware of any other details from the learning episode. In contrast, when people recall a list of words immediately after hearing them, they are fully aware of the episode in which the target items were experienced. In contrast to semantic memory, then, "episodic memory affords the additional capability of acquisition and retention of knowledge about personally experienced events and their temporal relations in subjective time and the ability to mentally 'travel back' in time" (Tulving, 1985: p. 387). Episodic memory is further divided into long-term and short-term (or working) memory.

PI in Semantic Memory

There is a very large volume of literature showing that information in semantic memory can cause PI in a wide range of tasks. For example, Waszak et al. (2004) asked subjects to alternate between a word-reading task and a picture-naming task. In the picture-naming task, the subject saw a word printed on a picture, such as the word *throat* printed on a picture of a foot, and was asked to say the name of the picture; in this example the response would be *foot*. Then, three different word-reading tasks could follow. The key here is that there were two types of priming: identity priming (i.e., the item itself, similar to the priming discussed in the perceptual representation system example with word fragment completion) and semantic priming (i.e., a related item, in which responses to the word *doctor* are faster if the word *nurse* has been processed previously). In one condition, the task featured the same stimuli and therefore invoked only identity priming (but keep in mind the subject would now be asked to respond with the word, i.e., *throat* rather than the name of the picture *foot*). The second condition had only semantic priming; for example, the word *thumb* printed on a picture of a leg. The third condition had novel items, such as the word *ship* printed on an airplane, which involves neither identity nor semantic priming. The main variable of interest was how long it took to read the word following a switch from the picture-naming task. Waszak et al. (2004) found that it took longer to read the word *thumb* when the prior trial consisted of naming a picture of a foot than when the prior trial was naming a picture of an airplane. That is, PI occurred because a semantically related item on an earlier trial interfered with the ability to read a word on the next trial.

One area of particular interest is the effect of prior knowledge on the acquisition of new knowledge. For example, people have experienced PI when moving from one type of problem and then trying to solve a second one (Luchins, 1942), when moving from one text editor to another (Singley and Anderson, 1989), and when learning a second language (MacWhinney, 2005). Roediger and Neely (1982) describe a number of additional cases of what they term 'retrieval blocks' in semantic memory; in each, the block can be attributed to the role of prior information or PI.

PI in Long-term Episodic Memory

Most laboratory-based studies use episodic tasks: a series of items are presented, and then a test session occurs in which memory for the just-presented information is assessed. Episodic memory has often been subdivided into long-term and short-term memory. PI is readily observed in long-term memory. For example, Greenberg and Underwood (1950) had subjects learn lists on four successive days; on each day, the subjects first recalled the list learned the previous day. It is this 24-h retention interval that makes it a test of long-term memory. With no prior lists, the mean percent correct was approximately 70%, but with three prior lists, performance dropped to approximately 25% correct.

The other type of episodic memory is variously known as short-term or working memory (*see* Short Term Memories, Theories of). There has been considerable debate over whether PI occurs in this system. It is worthwhile taking the time to examine how the debate has changed over the years, as the question is intimately related to the issue of what causes forgetting.

PI in Short-term/Working Memory

Does PI occur in short-term memory? The question originated with the first studies on short-term memory and concentrated on two main areas, the Brown–Peterson paradigm and investigations into a subsystem of working memory called the focus of attention.

PI in the Brown–Peterson Paradigm

In the Brown–Peterson paradigm (named after two sets of researchers who independently developed the task), the subject hears or sees a stimulus (e.g., three consonants in a row, such as DBX), and is then given a 3-digit number. The subject's task is to count backward by 3 (or 4 or 7) from the 3-digit number for a certain amount of time. At the end of this period, the subject is asked to recall the three consonants in order. The purpose of

counting backward is to prevent rehearsal while minimizing overt interference: the digits are sufficiently different from the letters that they should not interfere. The number of items to recall is small enough that capacity limits should not be a problem. The main results of interest are shown in Figure 2: Peterson and Peterson (1959) varied how long the subjects counted backward, including conditions of 3, 6, 9, 12, 15, and 18 s. What is most noteworthy is that after as little as 18 s of counting backward, subjects could recall only about 10% of the items.

This result has been replicated a large number of times, and was initially described as showing the rapid loss of information in short-term memory due to decay. However, a number of studies quickly questioned this conclusion.

Underwood (1957) had shown that recall decreases as the number of prior lists increase. What happens if performance on the first trial of a Brown–Peterson task is examined? Keppel and Underwood (1962) did just this and found no difference in forgetting on the first trial as a function of distractor duration. That is, it did not matter whether the subject counted backward for 3 s or for 30 s, the performance was equivalent. Differences as a function of the distractor duration appeared only on subsequent trials, a finding consistent with the idea that PI builds up over trials. This result has been replicated a number of times (see Crowder, 1976; for a review).

Wickens et al. (1963) came up with a simple test of the interference account. If PI builds up over trials because similar stimuli are used on each trial, changing the type of stimuli should release the subject from the negative effects of PI. To test this idea, the researchers presented consonants on the first three trials; on the fourth trial, they presented numbers. Because numbers were now being used as to-be-remembered stimuli, a different distractor task was used: naming colors. As shown in Table 2, there were actually four groups: one switched from letters to numbers, another switched from numbers to letters, a third had only letters on all four trials, and the fourth had only numbers on all four trials. Performance in the switched groups was much better on the fourth trial than in the control groups. This effect became known as the build-up and release from PI, and has also been replicated numerous times with many different kinds of stimuli. For example, build-up and release from PI can be observed when changing taxonomic category, presentation modality, or even part of speech. One particularly interesting release concerns what is known as the semantic differential. If all the words on the first three trials have similar meanings, such as something pleasant (sunny, enjoy, happy), but the words on the last trial mean something different (kill, death, illness), PI will build up and then be released. Crowder (1976) provides a summary of much of this literature.

Build-up and release from PI has also been observed with stimuli including television news (Gunter et al., 1980); prose passages (Blumenthal and Robbins, 1977); and motor learning tasks (Dickinson and Higgins, 1977). Dempster (1985) has shown PI can build up in situations like those encountered with standard academic tests, and Bunting (2006) has observed PI in measures typically used to assess and predict cognitive performance, including intelligence.

Note that the decay view initially proposed as the cause of forgetting in this situation has two difficulties accounting for the data. First, why is performance equivalent on the first trial regardless of whether the delay is 3 or 30 s? Second, why does performance improve on the fourth trial in switched group when the delay is exactly the same as in the unswitched group? Forgetting in the Brown–Peterson paradigm, then, is best accounted for by PI rather than decay.

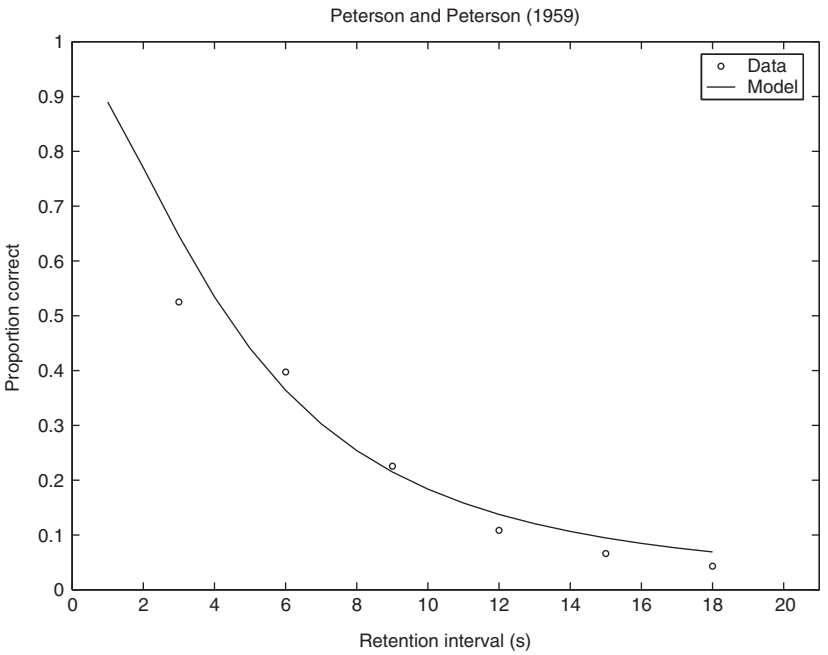

Figure 2 The proportion of trigrams correctly recalled as a function of the duration of the distractor task. The data points are from Peterson and Peterson (1959), the line shows the prediction of the model SIMPLE, described below. Source: Brown, G.D.A., Neath, I., Chater, N., 2007. A temporal ratio model of memory. Psychological Review 114, 539–576. Adapted with permission of APA.

Table 2 Design to show build-up and release of PI. The two control groups will show decreasing performance over all four trials. The two experimental groups will show decreasing performance through Trial 3, but will show enhanced performance on Trial 4

Group	*Trial 1*	*Trial 2*	*Trial 3*	*Trial 4*
Control 1	Digits	Digits	Digits	Digits
Experimental 1	Digits	Digits	Digits	Letters
Control 2	Letters	Letters	Letters	Letters
Experimental 2	Letters	Letters	Letters	Digits

PI in the Focus of Attention

There is another form of temporary storage that is held to be immune from PI, and it is usually referred to as the focus of attention. In most theoretical accounts of working memory, the focus of attention is exactly what it sounds like: it contains those items that are currently the focus of attention and, crucially, its capacity is very limited. The maximum number of items that can simultaneously be held in the focus of attention is approximately 3–4 (see Cowan, 2000). The prediction is that in general, tasks with fewer than four items will not show PI because those items can all fit in the focus of attention. However, tasks with more than four items will show PI, because in addition to the focus of attention, other parts of short-term or working memory are needed to retain the items.

Consistent with this view, there are a large number of studies that have failed to find PI with short lists but have observed PI with longer lists. The typical study manipulates semantic similarity. For example, on Trial 1, four items are shown, followed by a target item. The subject's task is to indicate whether the target was one of the items just seen. Usually, performance is close to perfect, so the main variable of interest is response time. On Trial 2, items from the same semantic category are used again. For example, Trial 1 might have four types of metal, and Trial 2 might have four additional types of metal. At some point, the category changes; for example, it might change from types of metal to types of vegetables. In studies that use lists with six or more items on each trial, PI builds up; that is, response time to correctly determine whether the target was on the list increases. When a category change occurs, response time decreases. In contrast, in studies that use lists of only four items, there is no evidence of PI (Halford et al., 1988).

One problem with this type of study is that it may be the case that people do not process items in terms of their meaning if there are very few items to remember. For example, for a list of only three items, it might be more economical (cognitively speaking) to just repeat the items over and over. If this is the case, then one can make the following prediction: If people are doing more superficial processing with shorter lists than with longer lists, PI should be observable on the shorter lists if the potentially interfering item is similar on the more superficial level. That is, PI will be seen if the type of items is similar. For example, if the first few trials all use digits, then PI should build up. A number of studies have confirmed this prediction. For example, if the first few trials each have three digits, but at some point the type of stimuli changes to letters, build-up and release of PI is observed even though the number of to-be-remembered items is less than the presumed capacity of the focus of attention (see Surprenant and Neath, 2009).

PI has also been observed in other tasks that also tap the focus of attention. For example, Ruusuvirta et al. (2008) used a two-tone pitch comparison task in which the subject was asked to judge whether the second tone was lower, the same as, or higher in pitch than the first tone. Given that there are only two items, it is difficult to argue that the capacity of the focus of attention has been exceeded. Nonetheless, PI was observed.

PI in Different Populations

A detailed account of PI in all potential populations is beyond the scope of the current article. However, it is clear that PI has been observed in all populations where it has been directly tested. For example, children typically show more PI than adults, with PI decreasing during late child development (Kail, 2002). Older adults (*see* Aging and Memory) also show PI, with some data indicating that PI increases for older adults as compared with younger adults (Bowles and Salthouse, 2003). Another intriguing finding concerns amnesia (*see* Amnesia: General): By definition, patients with amnesia have difficulty remembering information. There is evidence, however, that some patients with amnesia have memory difficulties because they cannot forget: in at least some cases, they are suffering from massive PI (Hamilton and Martin, 2007).

Theoretical Accounts of PI

Most explanations of PI are very similar to the description of the basic phenomenon: earlier items interfere with memory for later or subsequent items (e.g., Brown et al., 2007; Henson, 1998; Mensink and Raaijmakers, 1988). The differences between the various accounts have to do with exactly how the earlier items are processed at the time of test. Initially, Interference Theory attributed PI to response competition; Crowder (1976) provides a detailed review of the data that first supported and then challenged this account. Here, we focus on current theoretical accounts.

A Cue-Overload Account of PI

One explanation of PI invokes cue overload (Watkins and Watkins, 1975). Cue overload refers to the finding that the more items a cue subsumes, the less effective the cue will be and has been used as an explanation for a number of different memory effects including the list length effect, the fan effect, and even levels of processing (see Surprenant and Neath, 2009; for a review). According to a cue overload interpretation, the category of the to-be-remembered items can be a significant retrieval cue. It is as if the subject were asking at test, 'What were the three letters I just saw?' As more trials occur, the number of potential responses increases. The cue, in this case, letters, becomes less and less helpful, as there are more and more letters to choose from. The number of items subsumed under the category name increases and

performance decreases. When the type of category changes, as in the build-up and release of PI studies, a new cue is used that subsumes fewer items and thus performance improves. It is as if the subject were now asking at test, 'What were the three digits I just saw?' Note that this explanation is not necessarily incompatible with the distinctiveness account outlined below.

A Discrimination Account of PI

Baddeley (1976) offered an explanation for PI based on distinctiveness, and the idea is easy to appreciate through an analogy with visual perception. If you are looking at a set of evenly spaced telephone poles, the closer poles are easy to distinguish from one another, but as the poles get further away from the viewer, the poles start to blend together. This idea was originally developed by Koffka in 1935, and has been invoked by a number of different theorists (see Surprenant and Neath, 2009).

Baddeley's (1976) focus was on the build-up and release of PI in the Brown-Peterson paradigm. Imagine four groups of subjects in a standard Brown-Peterson task. All groups receive the to-be-remembered stimuli, and are then asked to count backward for 9 s. The difference between the groups is how much time elapses between the end of one trial and the beginning of the next trial; this is known as the intertrial interval. For the 0 s group, the next trial starts immediately, so they have a 0 s intertrial interval. For the 15 s group, the next trial starts 15 s after the end of the first, so the intertrial interval is 15 s. The final two groups have intertrial intervals of 30 and 60 s. Loess and Waugh (1967) found that the longer the intertrial interval, the better performance, going from .41 to .47 to .52 to .57 for the 0, 15, 30, and 60 s groups, respectively. (In a second experiment, they included an intertrial interval of 2 min, and did not observe any PI at all.) Assuming it took approximately 20 s to conduct one trial (9 s to present the to-be-remembered items and just over 10 s allowed for recall), we can represent the sequence of events graphically, as shown in Figure 3. Relatively speaking, the four dots representing the four trials are easier to discriminate from one another in the 60 s intertrial interval group (the bottom dots) than in the 0 s intertrial interval group (the upper dots). If this account is correct, then it makes an interesting prediction, outlined below.

Baddeley (1976) suggested that it is not the absolute amount of time but rather the relative amount of time that is important and calculated a ratio that predicts performance levels. To illustrate his ratio, consider another Brown-Peterson experiment reported by Turvey et al. (1970). For ease of exposition, we focus on three of the conditions, which differed only in how long the subjects performed the distracting activity, and we ignore the time taken to recall. One group of subjects counted backward for 10 s, one for 15 s, and one for 20 s. PI built up over trials, as would be expected. On the fifth trial, Turvey and associates had every group perform the distracting activity for 15 s; thus, one group's retention interval increased, one group's remained the same, and the third group's retention interval decreased. What Turvey and colleagues found was the three groups performed differently: On this fifth trial, the 10 s group got worse than on the fourth trial, the 15 s group stayed the same, and the 20 s group got better! Note once again that this pattern is inconsistent with a decay explanation: the same delay does not result in the same level of performance.

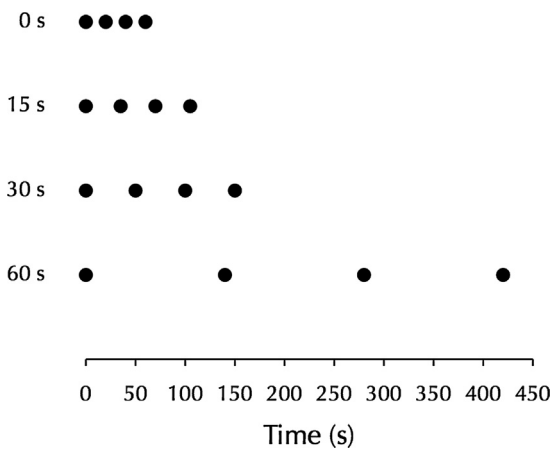

Figure 3 Representation of four conditions in the Loess and Waugh (1967) experiment. Each dot represents a complete Brown-Peterson trial and the gaps between the dots show the duration of the intertrial interval.

Baddeley (1976) offered an explanation that builds on the ideas in Figure 4. To simplify a little, we focus on Trial 4 versus Trial 5. The time of presentation of the to-be-remembered items on Trial 4 is represented as P4, and is set as an arbitrary zero point in Table 3. For the 10 s group, retrieval of the items occurs after 10 s (not shown in the table), and this is followed by presentation of the to-be-remembered item on Trial 5 after 10 s, which is represented as P5. Because this is Trial 5, the retrieval attempt is made 15 s later, so R5 occurs at the 25 s marker in the table below. For the 15 and 20 s groups, the same details are shown, but note how P5 for each group occurs later than for the 10 s group. The last column in the table shows what Baddeley called the 'Discrimination Ratio' which is the time between presentation of the fourth item and presentation of the fifth item (the numerator), and the time between presentation of the fifth item and recall of the fifth item (the denominator). For the three conditions, the ratios are 10/15 or .67; 15/15 or 1.00; and 20/15 or 1.33. As the ratio increases, recall increases even though the time to remember the item is identical.

A Distinctiveness Account of PI

The idea described above has been expanded and incorporated into a relative distinctiveness model of memory called Scale Independent Memory, Perception, and Learning or SIMPLE (Brown et al., 2007; *see also* Temporal Distinctiveness Models of Memory). In SIMPLE, items are assumed to be represented as points along a dimension (or as points in multidimensional space). In episodic tasks, the main dimension is temporal with the zero point being the time of recall. The temporal values undergo a log transformation, and then a value is calculated for each item indicating how similar it is to every other item. Importantly, closer neighbors have a larger effect than more distant neighbors. When additional items for

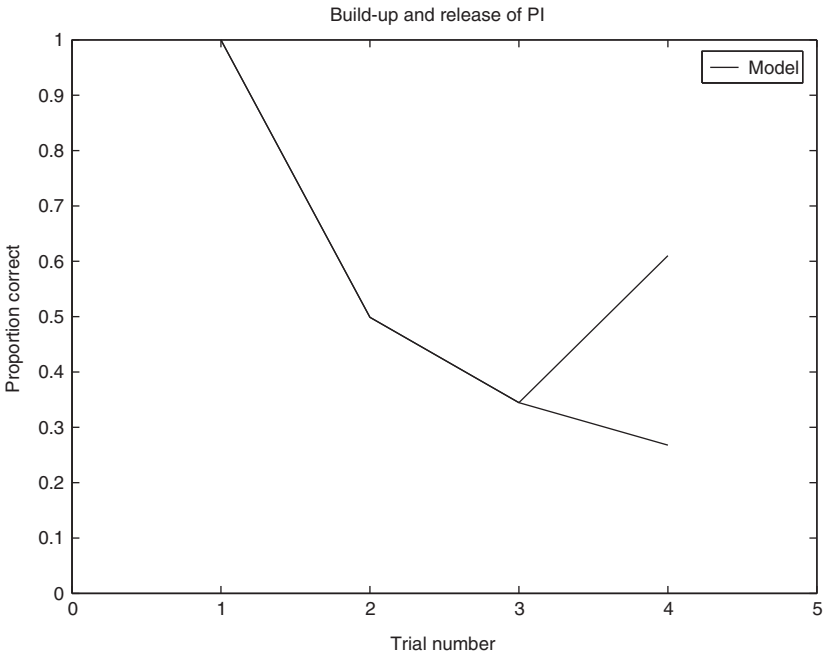

Figure 4 Idealized results showing the build-up of PI and, in the experimental condition, the release from PI when the type of to-be-remembered information is changed.

Table 3 Illustration of how to calculate Baddeley's (1976) discrimination ratio: Later items that are separated more from earlier items will have a larger ratio and a smaller effect of PI

Group	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*P4–P5*	*P5–R5*	*Ratio*
10	P4		P5			R5			10	15	.667
15	P4			P5			R5		15	15	1.000
20	P4				P5			R5	20	15	1.333

previous lists are similar, the relative distinctiveness of the correct item decreases, other factors being equal. If items differ on two dimensions (e.g., time and semantic), the effects of PI are diminished because the second dimension allows for more accurate discrimination. Brown et al. (2007) describe in detail how SIMPLE produces the lines shown in Figures 1, 2, and 4. Although SIMPLE has not been applied to nondeclarative memory, a similar type of explanation can be advanced.

See also: Aging and Memory; Amnesia: General; Episodic Memory; Implicit Memory; Non Declarative (Procedural) Memory; Retroactive Interference; Semantic Memory; Short Term Memories, Theories of; Temporal Distinctiveness Models of Memory.

Bibliography

Baddeley, A.D., 1976. The Psychology of Memory. Basic Books, New York.

Blumenthal, G.B., Robbins, D., 1977. Delayed release from proactive interference with meaningful material: how much do we remember after reading brief prose passages? Journal of Experimental Psychology: Human Learning and Memory 3, 754–761.

Bowles, R.P., Salthouse, T.A., 2003. Assessing the age-related effects of proactive interference on working memory tasks using the Rasch model. Psychology and Aging 18, 608–615.

Brown, G.D.A., Neath, I., Chater, N., 2007. A temporal ratio model of memory. Psychological Review 114, 539–576.

Bunting, M., 2006. Proactive interference and item similarity in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 32, 183–196.

Cook, T.W., 1941. Mirror position and negative transfer. Journal of Experimental Psychology 29, 155–160.

Cowan, N., 2000. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences 24, 87–185.

Crowder, R.G., 1976. Principles of Learning and Memory. Erlbaum, Hillsdale, NJ.

Dempster, F.N., 1985. Proactive interference in sentence recall: topic-similarity effects and individual differences. Memory & Cognition 13, 81–89.

Dickinson, J., Higgins, N., 1977. Release from proactive and retroactive interference in motor short-term memory. Journal of Motor Behavior 9, 61–66.

Fitts, P.M., Jones, R.H., 1961. Analysis of factors contributing to 460 "pilot error" experiences in operating aircraft controls. In: Sinaiko, H.W. (Ed.), Selected Papers on Human Factors in the Design and Use of Control Systems. Dover, New York, pp. 332–358.

Greenberg, R., Underwood, B.J., 1950. Retention as a function of stage of practice. Journal of Experimental Psychology 40, 452–457.

Gunter, B., Berry, C., Clifford, B.R., 1980. Release from proactive interference with television news items: evidence for encoding dimensions within televised news. Journal of Experimental Psychology: Human Learning and Memory 6, 216–223.

Halford, G.S., Maybery, M.T., Bain, J.D., 1988. Set-size effects in primary memory: an age-related capacity limitation? Memory & Cognition 16, 480–487.

Hamilton, A.C., Martin, R.C., 2007. Proactive interference in a semantic short-term memory deficit: role of semantic and phonological relatedness. Cortex 43, 112–123.
Henson, R.N.A., 1998. Short-term memory for serial order: the Start-End Model. Cognitive Psychology 36, 73–137.
Kail, R., 2002. Developmental change in proactive interference. Child Development 73, 1703–1714.
Keppel, G., Underwood, B.J., 1962. Proactive inhibition in short-term retention of ingle items. Journal of Verbal Learning and Verbal Behavior 1, 153–161.
Köster, E.P., Degel, J., Piper, D., 2002. Proactive and retroactive interference in implicit odor memory. Chemical Senses 27, 191–206.
Loess, H., Waugh, N.C., 1967. Components of short-term proactive interference. Journal of Verbal Learning and Verbal Behavior 6, 455–460.
Luchins, A.S., 1942. Mechanization in problem solving: the effect of Einstellung. Psychological Monographs 54, i–95.
Lustig, C., Hasher, L., 2001. Implicit memory is vulnerable to proactive interference. Psychological Science 12, 408–412.
MacWhinney, B., 2005. A unified model of language acquisition. In: Kroll, J.F., De Groot, A.M.B. (Eds.), Handbook of Bilingualism: Psycholinguistic Approaches. Oxford University Press, New York, pp. 49–67.
McGeoch, J.A., 1942. The Psychology of Human Learning: an Introduction. Longmans, Green and Company, New York.
Mensink, G.J., Raaijmakers, J.G., 1988. A model for interference and forgetting. Psychological Review 95, 434–455.
Peterson, L.R., Peterson, M.G., 1959. Short-term retention of individual verbal items. Journal of Experimental Psychology 58, 193–198.
Roediger III, H.L., Neely, J.H., 1982. Retrieval blocks in episodic and semantic memory. Canadian Journal of Psychology 36, 213–242.
Ruusuvirta, T., Wikgren, J., Astikainen, P., 2008. Proactive interference in a two-tone pitch-comparison task without additional interfering tones. Psychological Research/ Psychologische Forschung 72, 74–78.
Singley, M.K., Anderson, J.R., 1989. The Transfer of Cognitive Skill. Harvard University Press, Cambridge, MA.
Surprenant, A.M., Neath, I., 2009. Principles of Memory. Psychology Press, New York.
Tulving, E., 1985. How many memory systems are there? American Psychologist 40, 385–398.
Turvey, M.T., Brick, P., Osborn, J., 1970. Proactive interference in short-term memory as a function of prior-item retention interval. Quarterly Journal of Experimental Psychology 22, 142–147.
Underwood, B.J., 1957. Interference and forgetting. Psychological Review 64, 49–60.
Waszak, F., Hommel, B., Allport, A., 2004. Semantic generalization of stimulus-task bindings. Psychonomic Bulletin & Review 11, 1027–1033.
Watkins, O.C., Watkins, M.J., 1975. Buildup of proactive inhibition as a cue-overload effect. Journal of Experimental Psychology: Human Learning and Memory 104, 442–452.
Wickens, D.D., Born, D.G., Allen, C.K., 1963. Proactive inhibition and item similarity in short-term memory. Journal of Verbal Learning and Verbal Behavior 2, 440–445.

Probabilistic Grammars and their Applications

Stuart Geman and Mark Johnson, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12075–12082, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Formal grammars are widely used in speech recognition, language translation, and language understanding systems. Grammars rich enough to accommodate natural language generate multiple interpretations of typical sentences. These ambiguities are a fundamental challenge to practical application. Grammars can be equipped with probability distributions, and the various parameters of these distributions can be estimated from data (e.g., acoustic representations of spoken words or a corpus of hand-parsed sentences). The resulting probabilistic grammars help to interpret spoken or written language unambiguously. This article reviews the main classes of probabilistic grammars and points to some active areas of research.

Introduction

Natural language processing is the use of computers for processing natural language text or speech. Machine translation (the automatic translation of text or speech from one language to another) began with the very earliest computers (Kay et al. 1994). Natural language interfaces permit computers to interact with humans using natural language, for example, to query databases. Coupled with speech recognition and speech synthesis, these capabilities will become more important with the growing popularity of portable computers that lack keyboards and large display screens. Other applications include spell and grammar checking and document summarization. Applications outside of natural language include compilers, which translate source code into lower-level machine code, and computer vision (Fu 1974, 1982).

Most natural language processing systems are based on formal grammars. The development and study of formal grammars is known as computational linguistics. A *grammar* is a description of a language; usually it identifies the sentences of the language and provides descriptions of them, for example, by defining the phrases of a sentence, their interrelationships, and perhaps also aspects of their meanings. *Parsing* is the process of recovering a sentence's description from its words, while *generation* is the process of translating a meaning or some other part of a sentence's description into a grammatical or well-formed sentence. Parsing and generation are major research topics in their own right. Evidently, human use of language involves some kind of parsing and generation process, as do many natural language processing applications. For example, a machine translation program may parse an input language sentence into a (partial) representation of its meaning, and then generate an output language sentence from that representation.

Modern computational linguistics began with Chomsky (1957), and was initially dominated by the study of his 'transformational' grammars. These grammars involved two levels of analyses, a 'deep structure' meant to capture more-or-less simply the meaning of a sentence, and a 'surface structure' which reflects the actual way in which the sentence was constructed. The deep structure might be a clause in the active voice, 'Sandy saw Sam,' whereas the surface structure might involve the more complex passive voice, 'Sam was seen by Sandy.'

Transformational grammars are computationally complex, and in the 1980s several linguists came to the conclusion that much simpler kinds of grammars could describe most syntactic phenomena, developing Generalized Phrase-Structure Grammars (Gazdar et al. 1985) and Unification-based Grammars (Kaplan and Bresnan 1982, Pollard and Sag 1987, Shieber 1986). These grammars generate surface structures directly; there is no separate deep structure and therefore no transformations. These kinds of grammars can provide very detailed syntactic and semantic analyses of sentences, but even today there are no comprehensive grammars of this kind that fully accommodate English or any other natural language.

Natural language processing using handcrafted grammars suffers from two major drawbacks. First, the syntactic coverage offered by any available grammar is incomplete, reflecting both our lack of understanding of even relatively frequently occuring syntactic constructions and the organizational difficulty of manually constructing any artifact as complex as a grammar of a natural language. Second, such grammars almost always permit a large number of *spurious ambiguities*, that is, parses which are permitted by the rules of syntax but have unusual or unlikely semantic interpretations. For example, in the sentence 'I saw the boat with the telescope,' the prepositional phrase 'with the telescope' is most easily interpreted as the instrument used in seeing, while in 'I saw the policeman with the rifle,' the prepositional phrase usually receives a different interpretation in which the policeman has the rifle. Note that the corresponding alternative interpretation is marginally accessible for each of these sentences: in the first sentence one can imagine that the telescope is on the boat, and in the second, that the rifle (say, with a viewing scope) was used to view the policeman.

In effect, there is a dilemma of coverage. A grammar rich enough to accommodate natural language, including rare and sometimes even 'ungrammatical' constructions, fails to distinguish natural from unnatural interpretations. But a grammar sufficiently restricted so as to exclude what is unnatural fails to accommodate the scope of real language. These observations

led, in the 1980s, to a growing interest in stochastic approaches to natural language, particularly to speech. Stochastic grammars became the basis of speech recognition systems by outperforming the best of the systems based on deterministic handcrafted grammars. Largely inspired by these successes, computational linguists began applying stochastic approaches to other natural language processing applications. Usually, the architecture of such a stochastic model is specified manually, while the model's parameters are estimated from a *training corpus*, that is, a large representative sample of sentences.

As explained in the body of this article, stochastic approaches replace the binary distinctions (grammatical vs. ungrammatical) of nonstochastic approaches with probability distributions. This provides a way of dealing with the two drawbacks of nonstochastic approaches. Ill-formed alternatives can be characterized as extremely low probability rather than ruled out as impossible, so even ungrammatical strings can be provided with an interpretation. Similarly, a stochastic model of possible interpretations of a sentence provides a method for distinguishing more plausible interpretations from less plausible ones.

The introduction sets out formally various classes of grammars and languages. Probabilistic grammars are introduced in Section Grammars and Languages, along with the basic issues of parametric representation, inference, and computation.

Grammars and Languages

The formal framework, whether used in a transformational grammar, a generalized phrase-structure grammar, or a more traditionally styled context-free grammar, is due to Chomsky (1957) and his co-workers. This section presents a brief introduction to this framework. But for a thorough (and very readable) presentation we highly recommend the book by Hopcroft and Ullman (1979).

The concept and notation of a formal grammar is perhaps best introduced by an example:

Example 1: Define a *grammar*, G_1, by $G_1 = (T_1, N_1, S, R_1)$, where $T_1 = \{\text{grows, rice, wheat}\}$ is a set of words (a 'lexicon'), $N_1 = \{\text{S, NP, VP}\}$ is a set of symbols representing grammatically meaningful strings of words, such as clauses or parts of speech (e.g., S for 'Sentence,' NP for 'Noun Phrase,' VP for 'Verb Phrase'), and $R_1 = \{\text{S} \rightarrow \text{NP VP, NP} \rightarrow \text{rice, NP} \rightarrow \text{wheat, VP} \rightarrow \text{grows}\}$ is a collection of rules for rewriting, or instantiating, the symbols in N_1. Informally, the nonterminal S rewrites to sentences or clauses, NP rewrites to noun phrases and VP rewrites to verb phrases. The *language*, L_{G_1}, generated by G_1 is the set of strings of words that are reachable from S through the rewrite rules in R_1. In this example, $L_{G_1} = \{\text{rice grows, wheat grows}\}$, derived by $\text{S} \Rightarrow \text{NP VP} \Rightarrow \text{rice VP} \Rightarrow \text{rice grows}$, and $\text{S} \Rightarrow \text{NP VP} \Rightarrow \text{wheat VP} \Rightarrow \text{wheat grows}$.

More generally, if T is a finite set of symbols, let T^* be the set of all strings (i.e., finite sequences) of symbols of T, including the empty string, and let T^+ be the set of all nonempty strings of symbols of T. A language is a subset of T^*. A *rewrite grammar* G is a quadruple $G = (T, N, S, R)$, where T and N are disjoint finite sets of symbols (called the *terminal* and *nonterminal* symbols respectively), $S \in N$ is a distinguished nonterminal called the *start* or *sentence symbol*, and R is a finite set of productions. A *production* is a pair (α, β) where $\alpha \in N^+$ and $\beta \in (N \cup T)^*$; productions are usually written $\alpha \rightarrow \beta$. Productions of the form $\alpha \rightarrow \varepsilon$, where ε is the empty string, are called *epsilon productions*. This entry restricts attention to grammars without epsilon productions, that is, $\beta \in (N \cup T)^+$, as this simplifies the mathematics considerably.

A rewrite grammar G defines a *rewriting relation* $\Rightarrow_G \subseteq (N \cup T)^* \times (N \cup T)^*$ over pairs of strings consisting of terminals and nonterminals as follows: $\beta A \gamma \Rightarrow \beta \alpha \gamma$ iff $A \rightarrow \alpha \in R$ and $\beta, \gamma \in (N \cup T)^*$ (the subscript G is dropped when clear from the context). The reflexive, transitive closure of $\Rightarrow$ is denoted $\Rightarrow^*$. Thus $\Rightarrow^*$ is the rewriting relation using arbitrary finite sequences of productions. (It is called 'reflexive' because the identity rewrite, $\alpha \Rightarrow \alpha$, is included.) The language generated by G, denoted L_G, is the set of all strings $w \in T^+$ such that $S \Rightarrow^* w$.

Rewrite grammars are traditionally classified by the shapes of their productions. $G = (T, N, S, R)$ is a *context-sensitive* grammar if for all productions $\alpha \rightarrow \beta \in R$, $|\alpha| \leq |\beta|$, that is, the right-hand side of each production is not shorter than its left-hand side. G is a *context-free* grammar if $|\alpha| = 1$, that is, the left-hand side of each production consists of a single non terminal. G is a *left-linear* grammar if G is context-free and β (the right-hand side of the production) is either of the form Aw or of the form w where $A \in N$ and $w \in T^*$; in a *right-linear* grammar β always is of the form wA or w. A right or left-linear grammar is called a *regular* grammar. G_1, in Example 1, is context sensitive and context free.

It is straightforward to show that the classes of languages generated by these classes of grammars stand in strict equality or subset relationships. Specifically, the class of languages generated by right-linear grammars is the same as the class generated by left-linear grammars; this class is called the *regular* languages, and is a strict subset of the class of languages generated by context-free grammars, which is a strict subset of the class of languages generated by context-sensitive grammars, which in turn is a strict subset of the class of languages generated by rewrite grammars.

It turns out that context-sensitive grammars (where a production rewrites more than one nonterminal) have not had many applications in natural language processing, so from here on we will concentrate on context-free grammars, where all productions take the form $A \rightarrow \beta$, where $A \in N$ and $\beta \in (N \cup T)^+$.

An appealing property of grammars with productions in this form is that they induce tree structures on the strings that they generate. And, as Section Probability and Statistics shows, this is the basis for bringing in probability distributions and the theory of inference. We say that the context-free grammar $G = (T, N, S, R)$ generates the labeled, ordered tree ψ, iff the root node of ψ is labeled S, and for each node n in ψ, either n has no children and its label is a member of T (i.e., it is labelled with a terminal) or else there is a production $A \rightarrow \beta \in R$ where the label of n is A and the left-to-right sequence of labels of n's immediate children is β. It is straightforward to show that w is in L_G if G generates a tree ψ whose *yield* (i.e., the left-to-right sequence of terminal symbols labeling ψ's leaf nodes) is w; ψ is called a *parse tree* of w (with respect to G). In what follows, we define Ψ_G to be the set of parse trees generated by G, and $\mathscr{Y}^G(\cdot)$ to be the function that maps trees to their yields.

Example 1 (continued): The grammar G_1 defined above generates the following two trees, ψ_1 and ψ_2.

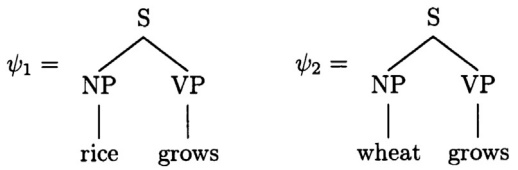

In this example, $\mathscr{Y}(\psi_1) =$ 'rice grows' and $\mathscr{Y}(\psi_2) =$ 'wheat grows.'

A string of terminals w is called *ambiguous* if w has two or more parse trees. Linguistically, each parse tree of an ambiguous string usually corresponds to a distinct interpretation.

Example 2: Consider $G_2 = (T_2, N_2, S, R_2)$, where $T_2 = \{$I, saw, the, man, with, telescope$\}$, $N_2 = \{$S, NP, N, Det, VP, V, PP, P$\}$ and $R_2 = \{$S $\rightarrow$ NP, VP, NP $\rightarrow$ I, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ NP, PP, N $\rightarrow$ man, N $\rightarrow$ telescope, VP $\rightarrow$ V, NP, VP $\rightarrow$ VP, PP, PP $\rightarrow$ P, NP, V $\rightarrow$ saw, P $\rightarrow$ with$\}$. Informally, N rewrites to nouns, Det to determiners, V to verbs, P to prepositions and PP to prepositional phrases. It is easy to check that the two trees ψ_3 and ψ_4 with the yields $\mathscr{Y}(\psi_3) = \mathscr{Y}(\psi_4) =$ 'I saw the man with the telescope' are both generated by G_2. Linguistically, these two parse trees represent two different syntactic analyses of the sentence. The first analysis corresponds to the interpretation where the seeing is by means of a telescope, while the second corresponds to the interpretation where the man has a telescope.

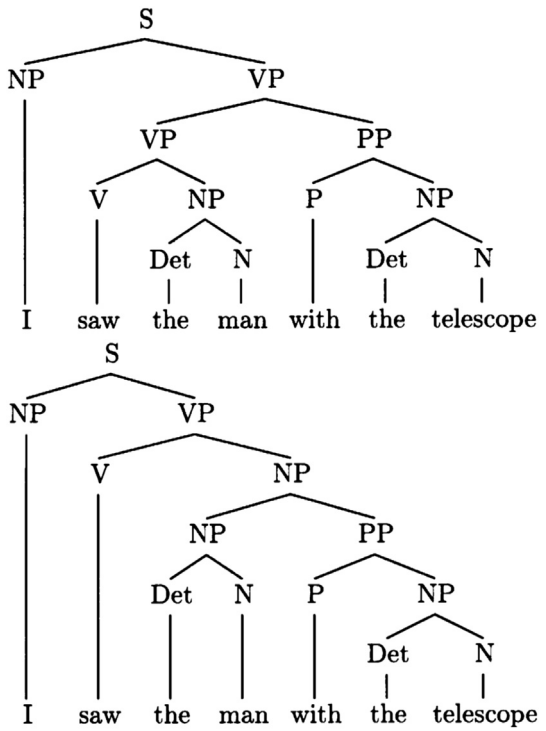

Probability and Statistics

Obviously broad coverage is desirable – natural language is rich and diverse, and not easily held to a small set of rules. But it is hard to achieve broad coverage without massive ambiguity (a sentence may have tens of thousands of parses), and this of course complicates applications like language interpretation, language translation, and speech recognition. This is the dilemma of coverage that we referred to earlier, and it sets up a compelling role for probabilistic and statistical methods.

We briefly review the main probabilistic grammars and their associated theories of inference. We begin in Section Regular Grammars with probabilistic regular grammars, also known as hidden Markov models (HMM), which are the foundation of modern speech recognition systems—see Jelinek (1997) for a survey.

In Section Context-Free Grammars we discuss probabilistic context-free grammars, which turn out to be essentially the same thing as branching processes. Finally, in Section Gibbs Distributions, we take a more general approach to placing probabilities on grammars, which leads to Gibbs distributions, a role for Besag's pseudolikelihood method (Besag 1974, 1975), various computational issues, and, all in all, an active area of research in computational linguistics.

Regular Grammars

We will focus on right-linear grammars, but the treatment of left-linear grammars is more or less identical. It is convenient to work with a *normal form*: all rules are either of the form $A \rightarrow bB$ or $A \rightarrow b$, where $A, B \in N$ and $b \in T$. It is easy to show that every right-linear grammar has an equivalent normal form in the sense that the two grammars produce the same language. Essentially nothing is lost.

Probabilities

The grammar G can be made into a probabilistic grammar by assigning to each nonterminal $A \in N$ a probability distribution p over productions of the form $A \rightarrow \alpha \in R$: for every $A \in N$

$$\sum_{\substack{\alpha \in (N \cup T)+ \\ \text{s.t. } (A \rightarrow \alpha) \in R}} p(A \rightarrow \alpha) = 1 \qquad [1]$$

Recall that Ψ_G is the set of parse trees generated by G (see Section Grammars and Languages). If G is linear, then $\psi \in \Psi_G$ is characterized by a sequence of productions, starting from S. It is, then, straightforward to use p to define a probability P on Ψ_G: just take $P(\psi)$ (for $\psi \in \Psi_G$) to be the product of the associated production probabilities.

S $\rightarrow$ aS	$p = 0.80$
S $\rightarrow$ bS	$p = 0.01$
S $\rightarrow$ bA	$p = 0.19$
A $\rightarrow$ bA	$p = 0.90$
A $\rightarrow$ b	$p = 0.10$

Example 3: Consider the right-linear grammar $G_3 = (T_3, N_3, S, R_3)$, with $T_3 = \{a, b\}$, $N_3 = \{S, A\}$ and the productions (R_3) and production probabilities (p):

The language is the set of strings ending with a sequence of at least two *b*s. The grammar is ambiguous: in general, a sequence of terminal states does not uniquely identify a sequence of productions. The sentence *aabbbb* has three parses (determined by the placement of the production S → bA), but the most likely parse, by far, is S → aS, S → aS, S → bA, A → bA, A → bA, A → b ($P = 0.8 \cdot 0.8 \cdot 0.19 \cdot 0.9 \cdot 0.1$), which has a *posterior* probability of nearly 0.99. The corresponding parse tree is shown below.

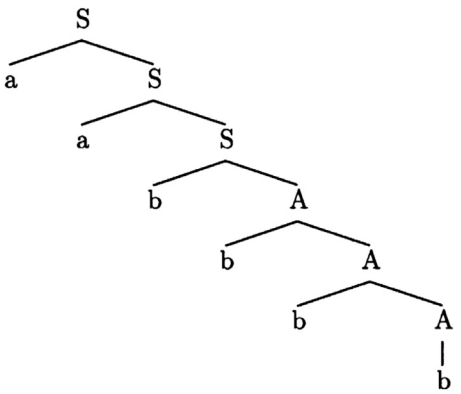

Inference

The problem is to estimate (see Estimation: Point and Interval) the transition probabilities, $p(\cdot)$, either from parsed data (examples from Ψ_G) or just from sentences (examples from L_G). Consider first the case of parsed data ('supervised learning'), and let $\psi_1, \psi_2, \ldots, \psi_n \in \Psi$ be a sequence taken iid according to P. If $f(A \to \alpha; \psi)$ is the counting function, counting the number of times transition $A \to \alpha \in R$ occurs in ψ, then the likelihood function (see Likelihood in Statistics) is

$$L = L(p; \psi_1, \ldots, \psi_n) = \prod_{i=1}^{n} \prod_{A \to \alpha \in R} p(A \to \alpha)^{f(A \to \alpha; \psi_i)} \qquad [2]$$

The maximum likelihood estimate is, sensibly, the relative frequency estimator:

$$\widehat{p}(A \to \alpha) = \frac{\sum_{i=1}^{n} f(A \to \alpha; \psi_i)}{\sum_{i=1}^{n} \sum_{\beta s.t} {}_{A \to \beta \in R} f(A \to \beta; \psi_i)} \qquad [3]$$

The problem of estimating p from sentences ('unsupervised learning') is more interesting, and more important for applications. Recall that $\mathscr{Y}(\psi)$ is the yield of ψ, that is the sequence of terminals in ψ. Given a sentence $w \in T^+$, let Ψ_w be the set of parses which yield $w : \Psi_w = \{\psi \in \Psi : \mathscr{Y}(\psi) = w\}$. Imagine a sequence $\psi_1, \ldots, \psi_n$, iid according to P, for which only the corresponding yields, $w_i = \mathscr{Y}(\psi_i) \quad 1 \le i \le n$, are observed. The likelihood function is

$$L = L(p; w_1, \ldots, w_n) = \prod_{i=1}^{n} \sum_{\psi \in \Psi_{w_i}} \prod_{A \to \alpha \in R} p(A \to \alpha)^{f(A \to \alpha; \psi_i)} \qquad [4]$$

As is usual with hidden data, there is an EM-type iteration for climbing the likelihood surface – see Baum (1972) and Dempster et al. (1977):

$$t + 1(A \to \alpha) = \frac{\sum_{i=1}^{n} E_{\hat{p}_i}[f(A \to \alpha; \psi) \psi \in \Psi_{w_i}]}{\sum_{A \to \beta \in R} \sum_{i=1}^{n} E_{\hat{p}_i}[f(A \to \beta; \psi) \psi \in \Psi_{w_i}]} \qquad [5]$$

Needless to say, nothing can be done with this unless we can actually evaluate, in a computationally feasible way, expressions like $E_{\hat{p}}[f(A \to \alpha; \ \psi) | \psi \in \Psi_w]$. This is one of several closely related computational problems that are part of the mechanics of working with regular grammars.

Computation

A sentence $w \in T^+$ is *parsed* by finding a sequence of productions $A \to bB \in R$ which yield w. Depending on the grammar, this corresponds more or less to an interpretation of w. Often, there are many parses and we say that w is ambiguous. In such cases, if there is a probability p on R then there is a probability P on Ψ, and a reasonably compelling choice of parse is the most likely parse:

$$\arg\max_{\psi \in \Psi_w} P(\psi) \qquad [6]$$

This is the *maximum a posteriori* (MAP) estimate of ψ—obviously it minimizes the probability of error under the distribution P. (Of course, in those cases in which Eqn. [6] is small, P does little to make w unambiguous.)

What is the probability of w? How are its parses computed? How is the most likely parse computed? These computational issues turn out to be more-or-less the same as the issue of computing $E_{\hat{p}}[f(A \to \alpha; \psi) | \psi \in \Psi_w]$ that came up in our discussion of inference. The basic structure and cost of the computational algorithm is the same for each of the four problems – compute the probability of w, compute the set of parses, compute the best parse, compute $E_{\hat{p}}$. In particular, there is a simple dynamic programming solution to each of these problems, and in each case the complexity is of the order $n \cdot |R|$, where n is the length of w, and $|R|$ is the number of productions in G – see Jelinek (1997), Geman and Johnson (2000). The existence of a dynamic programming principle for regular grammars is a primary reason for their central role in state-of-the-art speech recognition systems.

Context-Free Grammars

Despite the successes of regular grammars in speech recognition, the problems of language understanding and translation are generally better addressed with the more structured and more powerful context-free grammars. Following our development of probabilistic regular grammars in the previous section, we will address here the interrelated issues of fitting context-free grammars with probability distributions,

estimating the parameters of these distributions, and computing various functionals of these distributions.

The context-free grammars $G = (T, N, S, R)$ have rules of the form $A \to \alpha$, $\alpha \in (N \cup T)^+$, as discussed previously in Section Grammars and Languages. There is again a normal form, known as the Chomsky normal form, which is particularly convenient when developing probabilistic versions. Specifically, one can always find a context-free grammar G', with all productions of the form $A \to BC$ or $A \to a$, $A, B, C \in N$, $\alpha \in T$, which produces the same language as G: $L_{G'} = L_G$. Henceforth, we will assume that context-free grammars are in the Chomsky normal form.

Probabilities

The goal is to put a probability distribution on the set of parse trees generated by a context-free grammar in Chomsky normal form. Ideally, the distribution will have a convenient parametric form, that allows for efficient inference and computation.

Recall from Section Grammars and Languages that context-free grammars generate labeled, ordered trees. Given sets of nonterminals N and terminals T, let Ψ be the set of finite trees with:

a. root node labeled S;
b. leaf nodes labeled with elements of T;
c. interior nodes labeled with elements of N;
d. every nonterminal (interior) node having either two children labeled with nonterminals or one child labeled with a terminal.

Every $\psi \in \Psi$ defines a sentence $w \in T^+$: read the labels off of the terminal nodes of ψ from left to right. Consistent with the notation of Section Regular Grammars we write $\mathscr{Y}(\psi) = w$. Conversely, every sentence $w \in T^+$ defines a subset of Ψ, which we denote by Ψ_w, consisting of all ψ with yield $w(\mathscr{Y}(\psi) = w)$. A context-free grammar G defines a subset of Ψ, Ψ_G, whose collection of yields is the language, L_G, of G. We seek a probability distribution P on Ψ which concentrates on Ψ_G.

The time-honored approach to probabilistic context-free grammars is through the production probabilities $p : R \to [0, 1]$, with

$$\sum_{\substack{\alpha \in N^2 \cup T \\ \text{s.t.}(A \to \alpha) \in R}} p(A \to \alpha) = 1 \qquad [7]$$

Following the development in Section Regular Grammars, we introduce a counting function $f(A \to \alpha; \psi)$, which counts the number of instances of the rule $A \to a$ in the tree ψ, i.e., the number of nonterminal nodes A whose daughter nodes define, left-to-right, the string α. Through f, p induces a probability P on Ψ:

$$P(\psi) = \prod_{(A \to \alpha) \in R} P(A \to \alpha)^{f(A \to \alpha; \psi)} \qquad [8]$$

It is clear enough that P concentrates on Ψ_G, and we shall see shortly that this parameterization, in terms of products of probabilities p, is particularly workable and convenient. The pair, G and P, is known as a probabilistic context-free grammar, or PCFG for short. (Notice the connection to branching processes – Harris (1963): Starting at S, use R, and the associated probabilities $p(\cdot)$, to expand nodes into daughter nodes until all leaf nodes are labeled with terminals – i.e., with elements of T.)

Inference

As with probabilistic regular grammars, the production probabilities of a context-free grammar, which amount to a parameterization of the distribution P on Ψ_G, can be estimated from examples. In one scenario, we have access to a sequence $\psi_1, \ldots, \psi_n$ from Ψ_G under P. This is 'supervised learning,' in the sense that sentences come equipped with parses. More practical is the problem of 'unsupervised learning,' wherein we observe only the yields, $\mathscr{Y}(\psi_1), \ldots \mathscr{Y}(\psi_n)$.

In either case, the treatment of maximum likelihood estimation is essentially identical to the treatment for regular grammars. In particular, the likelihood for fully observed data is again Eqn. [2], and the maximum likelihood estimator is again the relative frequency estimator Eqn. [3]. And, in the unsupervised case, the likelihood is again Eqn. [4] and this leads to the same EM-type iteration given in Eqn. [5].

Computation

There are again four basic computations: find the probability of a sentence $w \in T^+$, find a $\psi \in \Psi$ (or find *all* $\psi \in \Psi$) satisfying $\mathscr{Y}(\psi) = w$ ('parsing'); find

$$\arg\max_{\substack{\psi \in \Psi_{\text{s.t.}} \\ \mathscr{Y}(\psi) = w}} P(\psi)$$

('maximum *a posteriori*' or 'optimal' parsing); compute expectations of the form $E_{\hat{p}_t}[f(A \to \alpha; \psi) | \psi \in \Psi_w]$ that arise in iterative estimation schemes like (5). The four computations turn out to be more-or-less the same, as was the case for regular grammars (Section Computation) and there is again a common dynamic-programming-like solution – see Lari and Young (1990, 1991), Geman and Johnson (2000).

Gibbs Distributions

There are many ways to generalize. The coverage of a context-free grammar may be inadequate, and we may hope, therefore, to find a workable scheme for placing probabilities on context-sensitive grammars, or perhaps even more general grammars. Or, it may be preferable to maintain the structure of a context-free grammar, especially because of its dynamic programming principle, and instead generalize the class of probability distributions away from those induced (parameterized) by production probabilities. But nothing comes for free. Most efforts to generalize run into nearly intractable computational problems when it comes time to parse or to estimate parameters.

Many computational linguists have experimented with using Gibbs distributions, popular in statistical physics, to go beyond production-based probabilities, while nevertheless preserving the basic context-free structure. Next we take a brief look at this particular formulation, in order to illustrate the various challenges that accompany efforts to generalize the more standard probabilistic grammars.

Probabilities

The sample space is Ψ_G, the set of trees generated by a context-free grammar G. Gibbs measures are built from sums of more-or-less simple functions, known as 'potentials' in statistical physics, defined on the sample space. In linguistics, it is more natural to call these features rather than potentials. Suppose, then, that we have identified M linguistically salient features $f_1, \ldots, f_M$, where f_k: $\Psi_G \to R$, through which we will characterize the fitness or appropriateness of a structure $\psi \in \Psi_G$. More specifically, we will construct a class of probabilities on Ψ_G which depend on $\psi \in \Psi_G$ only through $f_1(\psi), \ldots, f_M(\psi)$. Examples of features are the number of times a particular production occurs, the number of words in the yield, various measures of subject-verb agreement, and the number of embedded or independent clauses.

Gibbs distributions have the form

$$P_\theta(\psi) = \frac{1}{Z}\exp\left\{\sum_{i=1}^{M}\theta_i f_i(\psi)\right\} \qquad [9]$$

where $\theta_1, \ldots, \theta_M$ are parameters, to be adjusted 'by hand' or inferred from data, $\theta = (\theta_1, \ldots, \theta_M)$, and where $Z = Z(\theta)$ (known as the 'partition function') normalizes so that $P_\theta(\Psi) = 1$. Evidently, we need to assume or ensure that $\sum_{\psi\in\Psi_G}\exp\{\sum_1^M\theta_i f_i(\psi)\} < \infty$. For instance, we had better require that $\theta_1 < 0$ if $M = 1$ and $f_1(\psi) = |\mathscr{Y}(\psi)|$ (the number of words in a sentence), unless of course $|\Psi_G| < \infty$.

Relation to Probabilistic Context-Free Grammars

Gibbs distributions are much more general than probabilistic context-free grammars. In order to recover PCFG's, consider the special feature set $\{f(A \to \alpha; \psi)\}_{A\to\alpha\in R}$: The Gibbs distribution Eqn. [9] takes on the form

$$P_\theta(\psi) = \frac{1}{Z}\exp\left\{\sum_{A\to\alpha\in R}\theta_{A\to\alpha}\mathrm{f}(A\to\alpha;\psi)\right\} \qquad [10]$$

Evidently, then, we get the probabilistic context-free grammars by taking $\theta_{A\to\alpha} = \log_e p(A \to \alpha)$, where p is a system of production probabilities consistent with Eqn. [7], in which case $Z = 1$. But is Eqn. [10] more general? Are there probabilities on Ψ_G of this form that are not PCFGs? The answer turns out to be no, as was shown by Chi (1999) and Abney et al. (1999): given a probability distribution P on Ψ_G of the form of Eqn. [10], there always exists a system of production probabilities p under which P is a PCFG.

Inference

The feature set $\{f_i\}_{i=1,\ldots,M}$ can be accommodate arbitrary linguistic attributes and constraints, and the Gibbs model Eqn. [9], therefore, has great promise as an accurate measure of linguistic fitness. But the model depends critically on the parameters $\{\theta_i\}_{i=1,\ldots,M}$, and the associated estimation problem is, unfortunately, very hard. Indeed, the problem of unsupervised learning appears to be all but intractable.

Let us suppose, then, that we observe $\psi_1, \psi_2, \ldots, \psi_n \in \Psi_G$ ('supervised learning'), iid according to P_θ. In general, the likelihood function, $\prod_{i=1}^n P_\theta(\psi_i)$, is more or less impossible to maximize. But if the primary goal is to select good parses, then perhaps the likelihood function asks for too much, or even the wrong thing. It might be more relevant to maximize the likelihood of the observed parses, *given the yields* $\mathscr{Y}(\psi_1), \ldots, \mathscr{Y}(\psi_n)$:

$$\prod_{i=1}^{n} P_\theta(\psi_i|\mathscr{Y}(\psi_i)) \qquad [11]$$

Maximization of Eqn. [11] is an instance of Besag's remarkably effective pseudolikelihood method (Besag 1974, 1975), which is commonly used for estimating parameters of Gibbs's distributions. The computations involved are generally much easier than what is involved in maximizing the ordinary likelihood function. Take a look at the gradient of the logarithm of Eqn. [11]: the θ_j component is proportional to

$$\frac{1}{n}\sum_{i=1}^{n} f_j(\psi_i) - \frac{1}{n}\sum_{i=1}^{n} E_0\left[f_j(\psi)|\mathscr{Y}(\psi) = \mathscr{Y}(\psi_i)\right] \qquad [12]$$

and $\mathrm{E}_\theta[f_j(\psi)|\mathscr{Y}(\psi)]$ can be computed directly from the set of parses of the sentence $\mathscr{Y}(\psi)$. (In practice there is often massive ambiguity, and the number of parses may be too large to feasibly consider. Such cases require some form of pruning or approximation.)

Thus gradient ascent of the pseudolikelihood function is (at least approximately) computationally feasible. This is particularly useful since the Hessian of the logarithm of the pseudolikelihood function is non-positive, and therefore there are no local maxima. What is more, under mild conditions pseudolikelihood estimators (i.e., maximizers of Eqn. [11]) are consistent; (Chi 1998).

Bibliography

Abney, S., McAllester, D., Pereira, F., 1999. Relating probabilistic grammars and automata. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Morgan Kaufmann, San Francisco, pp. 542–609.

Baum, L.E., 1972. An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. Inequalities 3, 1–8.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society, Series D 36, 192–236.

Besag, J., 1975. Statistical analysis of non-lattice data. The Statistician 24, 179–195.

Chi, Z., 1998. Probability models for complex systems. Brown University PhD thesis.

Chi, Z., 1999. Statistical properties of probabilistic context-free grammars. Computational Linguistics 25 (1), 131–160.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.

Cohen, Shay B., Smith, Noah A., 2010. Covariance in Unsupervised Learning of Probabilistic Grammars. The Journal of Machine Learning Research 11, 3017–3051.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Fu, K.S., 1974. Syntactic Methods in Pattern Recognition. Academic Press, New York.

Fu, K.S., 1982. Syntatic Pattern Recognition and Applications. Prentice-Hall, Englewood Cliffs, NJ.

Gazdar, C., Klein, E., Pullum, G., Sag, I., 1985. Generalized Phrase Structure Grammar. Blackwell, Oxford, UK.

Geib, C.W., Goldman, R.P., 2009. A Probabilistic Plan Recognition Algorithm Based on Plan Tree Grammars. Artificial Intelligence 173 (11), 1101–1132.

Geman, S., Johnson, M., 2000. Probability and statistics in computational linguistics, a brief review. Division of Applied Mathematics. Brown University Internal publication.

Getoor, L., Taskar, B., 2007. Introduction to statistical relational learning. The MIT press, Cambridge, MA.
Harris, T.E., 1963. The Theory of Branching Processes. Springer, Berlin.
Hopcroft, J.E., Ullman, J.D., 1979. Introduction to Automata Theory. Languages and Computation. Addison-Wesley, Reading, MA.
Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press, Cambridge, MA.
Kaplan, R.M., Bresnan, J., 1982. Lexical functional grammar: A formal system for grammatical representation. In: Bresnan, J. (Ed.), The Mental Representation of Grammatical Relations. MIT Press, Cambridge, MA, pp. 173–281. Chap. 4.
Kay, M., Gawron, J.M., Norvig, P., 1994. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI Press, Stanford, CA.
Lari, K., Young, S.J., 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language 4, 35–56.
Lari, K., Young, S.J., 1991. Applications of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language 5, 237–257.
Pollard, C., Sag, A., 1987. Information-Based Syntax and Semantics. In: CSLI Lecture Notes Series No.13. University of California Press, Stanford, CA.
Shieber, S.M., 1986. An Introduction to Unification-Based Approaches to Grammar. In: CSLI Lecture Notes Series. University of California Press, Stanford, CA.

Probabilistic Thinking

Robyn M Dawes[†], Carnegie Mellon University, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12082–12089, © 2001, Elsevier Ltd.

Abstract

While games of chance have been played for thousands of years, and while experienced players often had good intuitive ideas of the relative frequencies of various outcomes, the concept of probability as referring to the ratio of favorable outcomes to total outcomes of a 'fair' gambling device emerged only 500 or so years ago in Western societies. That allowed evaluation of new games not yet played and later led to a more abstract conception of probability as a measure satisfying certain conditions ('axioms'). Only in the past 150 years or so has probability had a major role in science and only in the last 50 years or so in everyday life – as, for example, in evaluating medical outcomes or technological risks or manufacturer liability. Thus, especially in attempting to assess probability in everyday situations, our intuitions are often deficient – more so than those concerning quantity, space, and time. The major problem is not, however, one of making random errors in probability judgments but of making systematic ones that result from a number of well-known and research systematic cognitive biases and heuristics, which are described in this article. These systematic departures from rational assessment do not imply that coherent and accurate probabilistic thinking is (anywhere near) impossible, but that when departures from coherence and accuracy occur, they tend to follow a predictable pattern.

Introduction

Some people have enjoyed playing games that involved at least a component of chance ever since civilizations – as we know them – emerged in human history. Many of these people apparently had a fairly good intuitive idea of what we might now term the 'probabilities' or 'odds' of various outcomes associated with these games. Alternatively, at least some people (whom in hindsight we might now label 'superstitious') believed that the outcomes of these games could indicate something about the favorability of the world in general toward the gambler (a belief that may exist even today – at least on an implicit basis – for chronic gamblers), specifically the favorability or unfavorability of a particular god that was involved with a particular game.

In the last 600 years or so, people have proposed systematic ways of evaluating probabilities – e.g., in terms of equally likely outcomes in games of chance or in terms of subjective belief. In the last 150 years or so, scientific endeavor has started to make specific use of probabilistic thinking, and in the last 100 years or so, probabilistic ideas have been applied to problems of everyday life (see Gigerenzer et al., 1989). Currently, for example, jurors are often asked to determine manufacturer or company liability on the basis of differential rates of accidents or cancers; in such judgments, jurors cannot rely on deterministic reasoning that would lead to a specification of exactly which negative outcome is due to what nor can they rely on 'experience' in these contexts in which they have had none. Thus, it is extraordinarily important to understand – if, indeed, generalizations about such thinking are possible – how it is that people think about and evaluate probabilities, and, in particular, the problems that they (we) have. The major thesis presented here is that both the thinking and the problems can be understood in terms of *systematic deviations* from the normative principles of probabilistic reasoning that have been devised over the past five or so centuries. These deviations are not inevitable; it is certainly possible for people to think about probabilities coherently (at least on occasion); nevertheless, *when* deviations occur, they tend to be of the systematic nature to be described here; they occur most often when people are engaged in what is termed 'automatic' or 'associational' thinking, as opposed to 'reflective' and 'comparative' thinking. (These terms will be made more precise later in this article.)

A Brief History (of Thought)

Beginning in early infancy, we develop intuitions about space, time, and number (at least to the extent of 'one, two, three, many'). These intuitions are essentially correct. They remain throughout our lives basically unaltered except for small variations, and they form the foundation of far more sophisticated mathematical systems that have been developed over the past thousands of years. Moreover, their importance in our lives has been clear throughout those millennia.

In contrast, probabilistic thinking is often unintuitive, and it generally can be taught only after we are familiar with fractions, so that analogs to relative frequency are clear. Moreover, our intuitions about probabilities are often just plain wrong. For example, belief in the 'gambler's fallacy' (negative recency) can even be explicitly rejected, but nevertheless implicitly influence decision making – such as when people demand a greater amount of money in preference to gambling on an outcome that is 'due' (e.g., a tail after five straight heads) than in preference to gambling on one that is not (e.g., an additional head); see Gold (1997).

The fairly recent, and somewhat 'iffy,' use of probabilistic judgment does not imply, however, that people did not gamble throughout the centuries. As mentioned previously, the outcomes of games of chance could provide not just monetary winners and losers but 'signals' from 'the gods' about

[†] Deceased.

favorability toward the gambler; this belief is well captured by the narrative of the novelist Robert Graves (1943) about Caligula's gambling shortly before his assassination. This tyrant is both watching gladiator games, often putting to death people he has bribed to lose, and rolling a form of dice called 'astragali' – which are four-sided dice made from pig or lamb bones.

Caligula is playing with his Uncle Claudius, who in order to avoid winning (a dangerous outcome for Claudius) hands the emperor beautiful new dice loaded to come up with a 'Venus roll,' which consists of four separate sides (generally with the numbers 1, 3, 4, and 6). Caligula's subsequent success convinces him that Venus is favorable toward him that day, and he does not take the usual precautions upon leaving the games. Throughout his life, Claudius feels guilty that the loaded dice sent the wrong message to Caligula about how Venus felt about him that day – an inappropriate intervention between a god and a man. (According to Hacking, 1975, the interpretation of such events as 'signals' from 'on high' is common up to the Renaissance, even after people in Europe had given up polytheism for monotheism; the reader might also note how often in the *Old Testament* people decide to 'cast lots' to decide what to do – apparently in the belief that God would determine the outcome.)

Of course, not everyone believed in such intervention. For example, Cicero is quoted as claming that "Nothing is so unpredictable as a throw of the dice [modern translation]," and that "both now and then indeed he will make it [the Venus roll] twice and even thrice in succession. Are we going to be so feebleminded then as to aver that such a thing happened by the personal intervention of Venus rather than by pure luck?" (see David, 1962: p. 25).

Among those not so superstitious (such as Cicero, who was himself assassinated), the intuitive understanding gained through experience of which bets are good ones and which bets are bad ones led to fairly good implicit estimates of probability. Even now, however, Wagenaar (1988) finds chronic gamblers to be more reliant on intuition and experience than on mathematical analysis.

In contrast, explicit probabilistic thinking goes beyond experience. We can, for example, indicate what are good or bad bets in gambles we have just now constructed, with which we have had no previous experience. Probabilistic thinking can also be extended to 'personal belief' about uncertain events, it can be used to evaluate risks, and it can serve as a guide for social policy.

Systematic, probabilistic thinking appears to have started with questions about gambling (in particular about gambles that have not yet been played multiple times). Despite the fact that people gambled a lot, however, the formal theory did not begin until the Italian Renaissance, specifically in correspondence and argumentation between Geronimo Cardano (1501–76) and his associates. The approach that has been ascribed to Cardano involved *comparative* rather than associational thinking. He proposed that the likelihood of success of a gamble could be *computed*, or at least approximated, by comparing the number of outcomes of the device favorable to the gambler divided by the total number of outcomes. In situations where the gambling device is set up to make each possible outcome equally likely, this ratio assessment of probability is used today. The ratio assessment leads to certain general principles, often termed 'axioms,' which, in turn, provide a more abstract notion of probability, and allow different interpretations of probability to be developed – just provided they lead to measures consistent with the axioms. For example, Ramsey (1931) demonstrated that when explicit probabilities are used to determine fair gambling odds, then the gambler will follow the basic principles of probability theory if and only if it is not possible to make a 'Dutch Book' (see de Finetti, 1974–75: p. 87) such that the gambler always loses; that is, the gambler will be consistent if and only if it is impossible for a highly intelligent opponent to make a series of bets such that the opponent always wins.

After Cardano, the idea of probabilistic inference developed quite rapidly. Many French and Dutch intellectuals corresponded with each other about what constituted better or worse bets – not just in games that had been played chronically, but in games of their own invention. Interestingly, when they discussed probability, they often did so in terms of the expected value of the game, and questions were often posed as to which of the two games had a higher expected value. Then there was rapid development of probability theory – for example, Bernoulli's 'Law of Large Numbers,' which specified the relationship between the probability of a single outcome and the probability distribution of multiple outcomes of the same type (for example, the relationship between the probability a fair coin comes up heads and the probability that it comes up heads at least three times in 10 trials); the 'law' itself derived the limit of the proportion of heads as the number of trials becomes indefinitely large. To jump ahead to recent times, von Neumann and Morgenstern published their classic work *Theory of Games and Economic Behavior* (1944/1947), in which again the question was one of choosing between different gambles, but in terms of expected utility rather than expected value. This work was quite consistent with the earliest work, in which the question of probability was secondary to the question of choice between alternative courses of action.

Probability Assessments

All probabilistic analysis is based on the idea that (suitably trained and intelligent) people can at least *recognize* good probabilistic arguments presented by someone else, or discovered or thought of by themselves, but not necessarily generate good assessments. The very fact that there was correspondence about the gambles – and occasionally some disputes about them – indicated that people do not automatically assess probabilities in the same way, or accurately (e.g., corresponding to relative frequencies, or making good gambling choices).

The von Neumann and Morgenstern work, however, does involve some *psychological* assumptions that people can engage in 'good' probabilistic thinking. First, they must do so implicitly, so that the choice follows certain 'choice axioms' that allow the construction of an expected utility model – i.e., a model that *represents* choice as a maximization of implicit expected utility, that, in turn, requires that probabilities at the very least follow the standard axioms of probability theory. It also implies considering conditional probabilities in a rational

manner, which is done only when implicit or explicit conditional probabilities are consistent with Bayes' theorem. Thus, the von Neumann and Morgenstern work required that people be 'Bayesian' in a consistency sense, although that term is sometimes used to imply that probabilities should at base be interpreted as degrees of belief. Another way in which probability assessment must be 'good' is that there should be some at least reasonable approximation between probabilities and long-term relative frequencies; in fact, under particular circumstances (of interchangeability and indefinitely repeated observations), the probabilities of someone whose belief is constrained by Bayes' theorem must approximate relative frequencies (Dawid, 1982).

Are people always coherent? Consider an analogy with swimming. People do swim well, but occasionally we drown. What happens is that there is particular systematic bias in attempting to swim that makes it difficult. We want to hold our heads above water. When, however, we raise our heads to do so, we tend to assume a vertical position in the water, which is one of the few ways of drowning (aside from freezing or exhaustion or being swept away in rough waters).

Just as people drown occasionally by trying to hold their heads above water, people systematically deviate from the rules of probabilistic thinking. Again, however, the emphasis is on 'systematic.' For example, there is now evidence that people's probabilistic judgments are 'subadditive' – in that when a general class is broken into components, the judgmentally estimated probabilities assigned to disjoint components that comprise the class sum to a larger number than the probability assigned to the class. That is particularly true in memory, where, for example, people may recall the frequency with which they were angry at a close friend or relative in the last month and the frequency with which they were angry at a total stranger, and the sum of the estimates is greater than an independent estimate of being angry period (even though it is possible to be angry at someone who is neither a close friend or relative nor a total stranger; see Mulford and Dawes, 1999). The clever opponent will then bet against the occurrence of each component but on the occurrence of the basic event, thereby creating a Dutch Book.

The Heuristics and Biases Approach to Problems of Probabilistic Thinking

The basic conception is that we all use a series of 'heuristics' in our everyday estimation (explicit or implicit) of probabilities, and these heuristics lead to biases. These heuristics provide a type of 'bounded rationality' (Simon 1955, 1956, 1979) – in that they substitute for a very detailed or logical analysis, an analysis that may be beyond our 'computational capacity.' Just as associational thinking serves us well in many contexts, however, so do the heuristics. Occasionally, however, they lead to grossly erroneous – systematically erroneous – probabilistic thinking. The conception that heuristics have some type of 'generalized validity,' but that they often lead us astray, was well argued in a classic *Science* article by Tversky and Kahneman (1974); unfortunately, the principle has been lost on some critics, who can point out many situations in which these heuristics 'work.' Naturally, they do. If they did not have some sort of generalized utility, they would not have survived over generations of people trying to make probabilistic judgments.

One of the most important of these heuristics is termed the 'representativeness' heuristic, which was first proposed by Tversky and Kahneman in an article concerning *Belief in the Law of Small Numbers* (1971). The idea supported in that article is that we expect very small samples to 'represent' the populations from which they were drawn, where the 'we' includes even people with a fair amount of formal training in statistics. That expectation leads to judgment of probability on the basis of the degree to which an instance or a sample *matches* a population or general idea, rather than on a judgment of how it can be generated. For example, when tossing a coin, a result consisting of six heads appears to be much less likely than one of a head followed by two tails, followed by two heads, and finally a tail. But, of course, each of the sequences is equally likely; each occurs with probability 1/64. In fact, the sequence of a head followed by two tails followed by two heads and a tail is seen as more likely than a sequence of strict alternation of heads and tails – even though, in general, people believe that there is more than a chance amount of alternation when they judge what is truly 'random' (and in fact are most apt to pick as random the sequence that has a transition probability of 2/3 rather than 1/2). This incorrect expectation of alternation leads to such erroneous beliefs as that in the 'hot hand' in basketball when observing clusters of similar outcomes (e.g., successful shots). See, for example, Gilovich et al. (1985).

Another heuristic is termed as availability one. Briefly, we estimate frequencies on the ease with which instances 'come to mind,' or on our judgment of the ease with which we can recall or generate instances. It is, in fact, generally true that when there are more instances as opposed to fewer in the world, we find it easier to recall or generate more as opposed to fewer. The problem is, however, that there are determinants of ease of recall or generation that are independent of – or sometimes even in conflict with – actual relative frequency in our environment. For example, vivid instances are more readily recalled than pallid ones. (See again the result on 'subadditivity' in memory.) Moreover, instances that are easily imaginable (even if we have not actually experienced them) are more readily 'brought to mind' than are those difficult to imagine.

The availability heuristic creates a bias when we are unaware of the factors other than frequency that can result in ease of recall or generation, but use it as a way of estimating actual frequency. Sometimes, the bias is so strong that we even ignore instances that we *cannot* sample – a 'structural availability' problem. For example, clinical psychologists often claim that no child sexual abusers stop on their own without therapy. When asked to substantiate this claim, they will cite many of the child abusers with whom they have had contact, often mandated by a judge to enter therapy with them. The problem is, however, that they would not have had contact with any who had stopped on their own, and by definition they were seeing people only in a therapy setting, hence could not observe any of them stopping 'without therapy.' Nevertheless, the experience with such people is so salient and compelling that even while acknowledging the logical problem with the conclusion, many of these clinical psychologists simply revert to the belief a day

or a week after having publicly recognized the availability bias (see Dawes, 1988: p. 102).

Following the ideas of Fischhoff and Beyth-Marom (1983) as elaborated by Dawes (1998), the standard cognitive biases and heuristics can be understood by alluding to simple forms of Bayes' theorem. Consider, for example, the relationship between a symptom S and a disease D, and suppose a diagnostician observes this symptom S. If the probability of the disease is assessed on the basis of a pure matching or association between the symptom and the disease, independent of considerations of conditional probability, there is no normative structure to which the judgment corresponds. More often, however, the judgment will be made on the basis of the conditional probabilities – a normatively correct judgment if it is based on the conditional probability of the disease given the symptom, $P(D/S)$, but a representative judgment if it is the probability of the symptom given the disease, $P(S/D)$. Unfortunately, there is a lot of evidence that the judgment is often made on the basis of the latter relationship – when conditional probabilities are considered at all.

The relationship between these two probabilities is given by rewriting Bayes' theorem as:

$$P(D/S) = \frac{P(S/D)P(D)}{P(S)} \qquad [1]$$

It can also be written in terms of the 'ratio rule':

$$\frac{P(D/S)}{P(S/D)} = \frac{P(D)}{P(S)} \qquad [2]$$

Because the ratio rule follows from the very definition of conditional probability, it is simply incoherent to equate the probability of the disease given the symptom with the probability of the symptom given the disease in the absence of a simultaneous belief that the probability of the disease and the probability of the symptom are identical. (Again, if someone were to make a series of bets based on 'fair betting odds' in such a belief, an opponent could make a Dutch Book against that person.) Classic representative thinking – e.g., 'she gave a typical schizophrenic response, therefore she must be schizophrenic' – embraces such an identity. What is critical, of course, are the *base rates* of D and S, yet extensive research has shown that people underutilize such base rates in their intuitive judgments – or at least fail to incorporate them to a sufficient extent into their prior beliefs, given that they are often seen as 'not relevant.' (Why, people often ask, should the *general* probabilities of these symptoms and diseases be relevant to a particular judgment about a particular person with a particular symptom? Such an argument is sometimes even followed by bland assertions to the effect that 'statistics do not apply to the individual' – a wholly erroneous conclusion, as can be attested to by the heavy cigarette smoker who develops lung cancer.)

People often do not ignore these base rates *entirely* in the judgments, particularly if the probabilistic nature of the judgment is made clear (see Gigerenzer et al., 1989). Nevertheless, the 'underutilization' of such base rates is ubiquitous. In the psychological literature it was first decried by Meehl and Rosen (1955), who noted that in case conferences they attended, clinical psychology and psychiatry judges tended to ignore base rates completely; later, underutilization has been established in a variety of experimental situations using a wide variety of subjects, especially by Kahneman and Tversky (1972, 1973). For example, when asked to judge whether a description is of a particular engineer or a particular lawyer, subjects make the same judgment whether the experimenter states that the description was drawn from a pool of 70% engineers and 30% lawyers, or vice versa. As Bar-Hillel (1990) points out, the 70/30 split does not appear to be 'relevant,' but it's relevance would most likely be clear if the split were 99/1, and certainly if it were a 100/0. (Again, we can hypothesize that people can recognize probabilistic arguments – although it's possible to be fooled on occasion – but often not make them spontaneously.) Thus, for example, someone who is said to be interested in sailing and carpentry and sharing model train sets with a child is judged to be much more likely to be an engineer than a lawyer, whether this person was drawn from a pool of 70% engineers or 70% lawyers. Making the sampling procedure absolutely 'transparent' does, however, reduce that tendency.

The use of Bayes' theorem to understand the relationship between symptoms and disease leads naturally to its use to consider the general relationship between any hypotheses and any bit of evidence. Let e refer to some bit of evidence about whether a hypothesis h is or is not true; for example, a symptom may be regarded as a bit of evidence, and having a particular disease a hypothesis; or the hypothesis may concern who is going to win the election in the United States in the year 2020 and a particular bit of evidence may be the margin of victory or defeat of the potential candidate in a more local election. Bayes' theorem expressed in terms of evidence and hypotheses is presented as

$$P(h/e) = \frac{P(e/h)P(h)}{P(e)} \qquad [3]$$

This form of Bayes' theorem, while true, often leads to complications when trying to evaluate the probability of the evidence. There is, however, a way to avoid having to estimate that probability. What we can do is to consider *odds* that the hypothesis is true. These odds are the probability the hypothesis is true given the evidence divided by the probability that this hypothesis is false given the evidence – that is, by the ratio $P(h/e)/P(-h/e)$. Once we know the odds, it is trivial to compute the probability. The advantage of considering odds is that the denominator in Bayes' theorem cancels out when we compute these odds. That is

$$\frac{P(h/e)}{P(-h/e)} = \frac{P(e/h)P(h)}{P(e/-h)P(-h)} \qquad [4]$$

Now we are in a position to categorize the most common forms of representative thinking. Pseudodiagnosticity refers to making an inference about the validity of hypothesis h on the base of evidence e without considering hypothesis $-h$. Another way of stating pseudodiagnosticity is that it involves considering only the numerator of the first term on the right-hand side in eqn [4]. People do that when they state that a bit of evidence is 'consistent with' or 'typical of' some hypothesis without concerning themselves about alternative hypotheses, or in particular the negation of the hypothesis. For example, in attempting to diagnose whether a child has been sexually abused, alleged experts in court often refer to

symptoms 'typical' of such abuse – without concerning themselves with how typical these symptoms are of children who have *not* been abused, or the frequency with which children have been sexually abused.

Another type of representative thinking involves considering the probability of the evidence given the hypothesis or hypotheses without looking at the second terms in eqn [4], that is, the prior odds. The reason that these odds are important is that they yield the *extent* of the evidence and hypotheses considered. For example, being a poor speller may be evidence of dyslexia, in the sense, that it has a very high probability given dyslexia. However, in order to diagnose dyslexia on the basis of poor spelling, we have to know something about the base rates of poor spelling and the base rates of dyslexia in the population from which the person whom we wish to diagnose was chosen.

Dilution effects occur when evidence that does not distinguish between hypotheses in fact influences people to be less convinced of a most likely hypothesis; the problem is similar to that found in pseudodiagnosticity in that people do not realize that evidence that is not very likely given the hypothesis may be equally unlikely given the negation of that hypothesis, but may in fact believe a hypothesis less as a result of collecting evidence that is unlikely if it is true. Dilution is simply the converse of pseudodiagnosticity.

Finally, it is possible to describe *availability* biases as well by eqn [4]. People believe that they are sampling evidence given the hypothesis, when, in fact, they are sampling this evidence given the hypothesis *combined with* the manner in which they are sampling. For example, when private practice clinical psychologists claim that they are sampling characteristics of people who fall into a certain diagnostic category on the basis of their experience, what they are really sampling is people who fall in that category *and who come to them*. Similarly, when we sample our beliefs about 'what drug addicts are like,' most of us are sampling on the basis of how the media present drug addicts – both in 'news' programs and in dramatizations (Dawes, 1994b). Doctors often sample on the basis of their contact with addicts when these addicts are ill, perhaps gravely so, while police are often sampling on the basis of their experience with these same addicts during arrests and other types of confrontation. It is not surprising, then, that doctors are in favor of a 'medical' approach to drug addiction including such policies as sterile needle exchanges, while police are in favor of much more punitive policies. Both are sampling evidence *combined with* their exposure to it.

All these anomalies have been demonstrated experimentally. It is important to point out, however, that the investigation of these anomalies did not spring simply from understanding Bayes' theorem and then from creating very clever experimental situations in which subjects' inferences systematically violate it. The motive to study these anomalies of judgment arises from having observed them on a more informal basis outside the experimental setting – and trying to construct an experimental setting, which yields a greater measure of control, that will allow them to be investigated in a systematic manner. The reader is referred, for example, to the essay of Meehl (1977) or the book of Dawes (1994a) for descriptions of these biases in the clinical judgment of professional psychologists and psychiatrists who rely on their own 'experience' rather than (dry, impersonal) 'scientific' principles for diagnosis and treatment.

Analyzing an Important Example

This section ends with a discussion of pseudodiagnosticity, a bias that is prevalent in 'expert' proclamations encouraging 'child sex abuse hysteria' (Dawes, 2001). The main problem here is that hypotheses are not compared: instead, single hypotheses are evaluated in terms of the degree to which evidence is 'consistent with' them; in addition, evidence is often *sought* in terms of its consistency with or inconsistency with 'favorite hypotheses' – rather than in terms of its ability to distinguish *between* hypotheses. This type of pseudodiagnosticity has made its way into the legal system, where experts are allowed to testify that a child's recanting of a report of sexual abuse is 'consistent with' the 'child sexual abuse accommodation syndrome' (Summit, 1983). The finding is that many children who are *known* to have been sexually abused deny the abuse later (recant), therefore, the probability of the evidence (the child recants) given the hypothesis (actual abuse) is not as low as might be naively assumed; therefore, recanting is 'consistent with' having been abused (i.e., can be considered part of a 'syndrome' of accommodation to abuse).

The problem with this pseudodiagnostic reasoning is that it compares the probability of the evidence given the hypothesis to the probability of the *negation of the evidence* given the hypothesis, whereas a rational comparison is of the probability of the evidence given the hypothesis to the probability of the evidence given the *negation of the hypothesis*. When considering this latter comparison, we understand immediately that the recanting would be diagnostic of actual abuse only if the probability of recanting given actual abuse were *higher* than the probability of recanting given such abuse had not occurred – a highly implausible (not to mention paradoxical) conclusion.

Such pseudodiagnostic reasoning is involved in other disasters as well, for example, the explosion of the US Space Shuttle Challenger on 22 January 1986. When considering the possibility that cold temperatures might cause the O-rings to malfunction, "the worried engineers were asked to graph the temperatures at launch time for the flights in which problems have occurred" (Russo and Schoemaker, 1989: p. 197). Thus, these engineers attempted to evaluate the hypothesis that there would be O-ring problems by asking what the temperature was given the existence of these problems. As indicated in **Figure 1(a)** (Figure 6 in Russo and Schoemaker 1989: p. 198), there appeared to be no relationship. When, however, the temperature of launches in which there were *no* problems were considered, a very clear relationship between problems and temperature is evident (see **Figure 1(b)**, from Russo and Schoemaker 1989: Figure 7, p. 198.)

Moreover, people actively seek evidence compatible or incompatible with a hypothesis rather than evidence that distinguishes between hypotheses, as has been extensively studied by Doherty and his colleagues (see, for example Doherty et al., 1979). Consider, for example, subjects who are asked to play the role of medical students (and actual medical students prior to specific training, see Wolf et al., 1985) who must

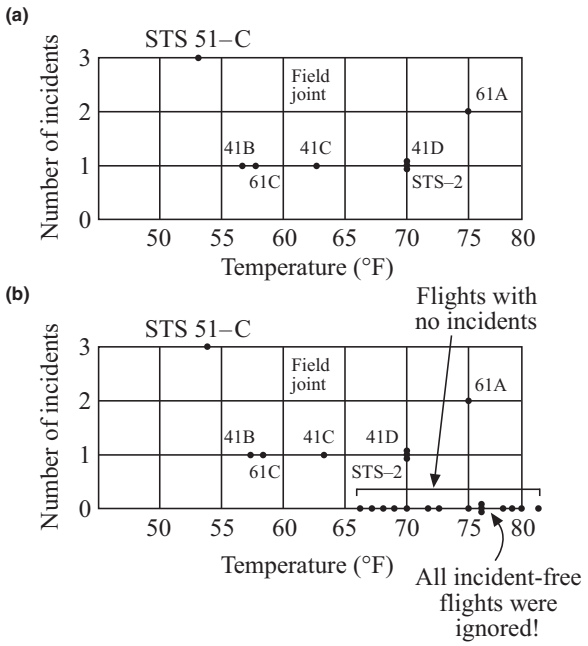

Figure 1 Partial and total analyses of the relationship between (a) data reviewed prior to launch and (b) the full story from past flights. Source: Russo J.E., Schoemaker P.J.H., 1989. Decision Traps: Ten Barriers to Brilliant Decision Making and How to Overcome Them. Simon and Schuster, New York, pp. 197–198.

determine whether patients have one of two conditions. The conditions are explained, and subjects are told that a patient has two symptoms. The judges are then presented with the conditional probability of one of these symptoms given one of the diseases (e.g., that it is very likely that the patient has a high fever given the patient has meningitis). They can then choose to find out the probability of the other symptom given the same disease, the probability of the first symptom given the second disease, or the probability of the second symptom given the second disease. While virtually no subjects choose a different symptom and a different disease, the majority choose to find the probability of the *second* symptom given the first disease (the 'one that is focal' as a result of being told the first symptom). But for all these subjects know, of course, the first symptom may be more or less typical of the *alternative disease*, as may the second symptom. Finding out that the probability of these two symptoms given only one disease in no way helps to distinguish between that disease and some other. *Of course, in real medical settings doctors may have prior knowledge of these other disease/symptom relationships,* but because the diseases are not even identified in the current setting, such knowledge would be of no help.

Occasionally, pseudodiagnosticity has been confused with what has been termed a 'confirmation bias' – that is, a desire to search for evidence that is very probable given a focal hypothesis. As Fischhoff and Beyth-Marom (1983), Trope and Bassok (1982), and Skov and Sherman (1986) point out, however, such a search does not necessarily imply pseudodiagnosticity (or any other particular irrationality) because it may be possible that the evidence is *not* found – that is, that instead of finding the anticipated evidence, the diagnostician finds its negation. In fact, Skov and Sherman (1986) demonstrate that in searching for evidence – at least in their experimental context – subjects often became quite rational *when they are informed of the conditional probabilities*. The irrationality occurs in the search for which probabilities to seek, and in 'jumping' to a conclusion on the basis of only one probability (i.e., on the basis of the first term in the numerator in the odds form of Bayes' theorem).

The cognitive explanation of pseudodiagnosticity and the other heuristic-based biases covered in the section is simple. People often think by association; in fact, whole computer programs have been built to simulate human judgment on the basis of 'matching' characteristics to schemas. The point of pseudodiagnosticity is that only one 'schema' (i.e., hypothesis) is considered, whereas *coherent diagnostic judgment is always comparative*. In fact, all the biases covered here are based on associational thinking, whereas clear normative probabilistic thinking always involves comparisons.

Summary and Implications

When we deal with probabilistic situations in everyday life, we can often 'muddle through,' but occasionally not appreciating the comparative nature of valid probabilistic thinking can lead to judgmental disasters. Similarly, experience with games of chance can lead to some pretty good intuitive estimations of the likelihood of various payoffs, but once again there can be systematic biases, such as belief in the gambler's fallacy. Explicit probabilistic thinking has developed in our cultures only in the last 500 years or so, and in the last 150 years or so it has become increasingly important in both scientific reasoning and in our day-to-day living – and even in legal contexts (where as jurors we may be asked to judge the liability of a particular company by comparing the probability of developing cancer among those exposed to it or its product as opposed to the 'base rate' probability, or may be asked to evaluate what an expert witness states about the relationship between a symptom and a diagnosis). Here, if we rely on our intuitive grasp of probabilities, we often find that we have systematic biases, much as novice swimmers may attempt to keep their heads above water and thereby drown. These biases – which have been briefly described and cataloged in this article – involve explicitly defined deviations from the conclusions of valid probabilistic reasoning, which follow from the simplest forms of Bayes' theorem. They are not random deviations, which could be modeled as analogous to an engineering context involving a 'signal' plus 'noise.'

Of course, we are also capable of precise probabilistic thinking – just as swimmers are capable of assuming a horizontal position in the water and consequently not drowning. The point is, however, that when we deviate from the rules ('axioms') of probability theory or from coherent judgments, these deviations follow the patterns described here – as has been documented in everyday observation and studied experimentally by those advocating the 'heuristics and biases' approach to such judgment. For a greater elaboration of this approach than is found here, see the chapters in Kahneman et al.'s book *Judgment Under*

Uncertainty: Heuristics and Biases (1982). These systematic departures themselves in turn may be connected to thinking in terms of associations, whereas good probabilistic judgment always requires comparative thinking.

See also: Information Theory; Information, Economics of; Probability Theory: Interpretations; Probability and Chance: Philosophical Aspects.

Bibliography

Bar-Hillel, M., 1990. Back to base-rates. In: Hogarth, R.M. (Ed.), Insights in Decision Making: A Tribute to Hillel J. Einhorn. University of Chicago Press, Chicago.

Dawes, R.M., 1988. Rational Choice in an Uncertain World. Harcourt Brace Jovanovich, San Diego, CA.

Dawes, R.M., 1994a. House of Cards: Psychology and Psychotherapy Built on Myth. Free Press, New York.

Dawes, R.M., 1994b. AIDS, sterile needles, and ethnocentrism. In: Heath, L., Tinsdale, R.S., Edwards, J., Posavac, E., Bryant, F.B., Henderson-King, E., Suarez-Balcazar, Y., Myers, J. (Eds.), Social Psychological Applications to Social Issues, Applications of Heuristics and Biases to Social Issues, vol. III. Plenum, New York, pp. 31–44.

Dawes, R.M., 1998. Behavioral decision making and judgment. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), The Handbook of Social Psychology, vol. 11. McGraw-Hill, Boston, MA, pp. 497–548.

Dawes, R.M., 2001. Everyday Irrationality: How Pseudoscientists, Lunatics and the Rest of Us Systematically Fail to Think Rationally. Westview Press, Boulder, CO.

David, F.N., 1962. Games, Gods, and Gambling. Hafner Pub. Co., New York (Chapter 4).

Dawid, A.P., 1982. The well-calibrated Bayesian. Journal of the American Statistical Association 77 (379), 605–610.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., Schiavo, M.D., 1979. Pseudodiagnosticity. Acta Psychologica 43, 111–121.

de Finetti, B., 1974–75. Theory of Probability: A Critical Introductory Treatment, vol. 1. John Wiley and Sons, London.

Fischhoff, B., Beyth-Marom, R., 1983. Hypothesis evaluation from a Bayesian perspective. Psychological Review 90, 239–260.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Kruger, L., 1989. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press, Cambridge, UK.

Gilovich, T., Vallone, R., Tversky, A., 1985. The hot hand in basketball: on the misperception of random sequences. Journal of Personality and Social Psychology 17, 295–314.

Gold, E., 1997. The Gambler's Fallacy, Doctoral Dissertation Submitted in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy. Carnegie Mellon University.

Hacking, I., 1975. The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference. Cambridge University Press, New York.

Kahneman, D., Tversky, A., 1972. Subjective probability: a judgment of representativeness. Cognitive Psychology 3, 430–454.

Kahneman, D., Tversky, A., 1973. On the psychology of prediction. Psychological Review 80, 237–251.

Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK.

Meehl, P.E., 1977. Why i do not attend case conferences. In: Meehl, P.E. (Ed.), Psychodiagnosis: Selected Papers. W.W. Norton, New York, pp. 225–302.

Meehl, P.E., Rosen, A., 1955. Antecedent probability in the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin 52, 194–216.

Mulford, M., Dawes, R.M., 1999. Subadditivity in memory for personal events. Psychological Science 10 (1), 47–51.

von Neumann, J., Morgenstern, O., 1944/1947. The Theory of Games and Economic Behavior, second ed. Princeton University Press, Princeton, NJ.

Ramsey, F., 1931. Truth and probability. In: Ramsey, F.P. (Ed.), The Foundations of Mathematics and Other Logical Essays. Harcourt, Brace and Co., New York, pp. 156–198.

Russo, J.E., Schoemaker, P.J.H., 1989. Decision Traps: Ten Barriers to Brilliant Decision Making and How to Overcome Them. Simon and Schuster, New York.

Simon, H., 1955. A behavior model of rational choice. Quarterly Journal of Economics 69, 99–118.

Simon, H., 1956. Rational choice and the structure of the environment. Psychological Review 63, 129–138.

Simon, H., 1979. Models of Thought, vol. 1. Yale University Press, New Haven, CT.

Skov, R.B., Sherman, S.J., 1986. Information-gathering processes: diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. Journal of Experimental Social Psychology 22, 93–121.

Summit, R.C., 1983. The child sexual abuse accommodation syndrome. Child Abuse and Neglect 7, 177–193.

Trope, Y., Bassok, M., 1982. Confirmatory and diagnosing strategies in social information gathering. Journal of Personality and Social Psychology 43 (1), 22–34.

Tversky, A., Kahneman, D., 1971. Belief in the law of small numbers. Psychological Bulletin 76, 105–110.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Wagenaar, W., 1988. Paradoxes of Gambling Behavior. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wolf, F.M., Gruppen, L.D., Billi, J.E., 1985. Differential diagnosis and the competing hypothesis heuristic – a practical approach to judgment under uncertainty and Bayesian probability. Journal of the American Medical Association 235, 2858–2862.

Probability and Chance: Philosophical Aspects

Christopher Hitchcock, California Institute of Technology, Pasadena CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12089–12095, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The central philosophical issue involving probability concerns the interpretation of ordinary probability claims such as 'there is a 30% probability that it will rain today.' *Epistemic* interpretations of probability understand this to be a claim about belief under uncertainty: it describes the extent to which a certain type of rational agent would (or ought to) believe that it will rain, given the available evidence. Epistemic interpretations include the *Classical*, *Subjective*, and *Logical* Interpretations. *Objective* interpretations of probability read the claim as describing some quantifiable feature of the objective world, sometimes called 'chance.' The concept of chance is thought by some to be intimately connected with *indeterminism*. A system is indeterministic if its later states are not determined by its earlier states in conjunction with the relevant *laws* of evolution. Many objective interpretations of probability invoke the notion of *relative frequency*: the probability of an outcome on a trial of a particular type is thought to be related to the relative frequency with which outcomes of that type do in fact occur on trials of that type. Objective interpretations of probability include the Finite Relative Frequency, Hypothetical Limiting Relative Frequency, Propensity, Best System, and Theoretical Role interpretations.

The central philosophical issue involving probability concerns the interpretation of ordinary probability claims, such as 'there is a 30% probability that it will rain today.' *Epistemic* interpretations of probability understand this to be a claim about belief under uncertainty: it describes the extent to which a certain type of rational agent would (or ought to) believe that it will rain, given the available evidence. *Objective* interpretations of probability read the claim as describing some quantifiable feature of the objective world, sometimes called 'chance.' The concept of chance is thought by some to be intimately connected with *indeterminism*. A variety of epistemic and objective interpretations of probability will be discussed below.

The Problem of the Interpretation of Probability

The problem of the interpretation of probability can be posed using a familiar example: what does a meteorologist *mean* when he or she says that there is a 30 percent probability of rain on a given day? To give an account of what a proposition means is to say what must be the case in order for the proposition to be true. If the meteorologist says that it *will* rain, this assertion is true if and only if it does rain. Whether it rains or not, however, neither outcome by itself renders the *probabilistic* claim true. So what must be the case in order for this claim to be true?

An adequate solution to this problem must provide some way of making sense of the numerical values that appear in probability claims (and must do so in such a way that those values are in accord with the basic principles of probability – see Probability Theory: Formal). The claim that 30 percent of the balls in a given urn are blue has a well-understood meaning: the number of blue balls in the urn, divided by the total number of balls in the urn, is equal to 0.3. In the case of the weather report, however, no comparable formula is immediately available.

Proposed interpretations of probability tend to fall within one of two broad camps, *objective* (or *empirical*) and *epistemic* (or *subjective*) interpretations. Objective interpretations interpret probability claims as describing some feature of the world. Epistemic interpretations of probability, by contrast, interpret probability claims as claims about the extent to which the claim is or ought to be believed. Note that this standard terminology is somewhat misleading: some epistemic interpretations (such as the logical interpretation) take probabilities to be objective in the sense that they do not vary from person to person. The various interpretations of probability canvassed below are often presented as *rivals*, but it may be more appropriate to think of them as complementary. Most researchers who have grappled with interpretive issues have been of the opinion that the concept of probability has both an objective and an epistemic dimension. Accordingly, philosophers often talk of 'objective probabilities' and 'epistemic probabilities.' Objective probabilities are sometimes called 'chances.'

Philosophers (as well as many philosophically minded statisticians) have become interested in the problem of interpretation of probability in part because of the ubiquity of probability claims in science as well as ordinary life, and in part because of the utility of probability theory in analyzing a variety of concepts of interest to philosophers, such as causation (see Causation: Physical, Mental, and Social), confirmation of hypotheses by evidence (see Atomism and Holism: Philosophical Aspects), and rationality (see Rational Choice Explanation: Philosophical Aspects).

This article will focus on the philosophical, rather than the mathematical, dimensions of the problem of interpreting probability. Readers interested in this topic would also do well to read the article on *Probability: Interpretations*. Other surveys of this topic include Salmon (1967), Sklar (1993), and Howson (1995).

Determinism and Chance

In a famous thought experiment, the French physicist Pierre Simon de Laplace (1819) imagined a demon who could

observe the exact position and momentum of every particle in the universe at a given time. Such a demon could plug this information into Newton's equations of motion and predict the state of the universe at all future times. In this thought experiment, Laplace was articulating the concept of *determinism*. A system is *deterministic* if its state at a given time, together with the laws of evolution for that system (see Causes and Laws: Philosophical Aspects), uniquely determine the state of the system at later times. For Laplace, determinism meant that all *objective* probabilities would have to have values of zero or one. Laplace thus thought that probabilistic claims, such as the meteorologist's prognostication described above, reflect human ignorance about which outcome is determined to occur.

At present, there is good reason to believe that certain physical systems described by *quantum mechanics* behave indeterministically. For example, if a vertically polarized photon strikes a polarizer that is oriented with an angle of θ, quantum theory entails that the photon has a probability of $\cos^2\theta$ of passing through. According to the orthodox interpretation of quantum theory (which is not uncontroversial), this probability is *irreducible*: there is no further fact about the photon that determines what it will do on this occasion. If this is correct, then the probability value $\cos^2\theta$ does not merely reflect human ignorance about what the photon will do: it is an objective feature of the world. (For a thorough discussion of philosophical issues involving determinism, see Earman (1986).) One important philosophical issue is whether only such irreducible probabilities are genuine chances. If so, then it would be safe to infer that *none* of the probability claims made within the social sciences describe genuinely objective probabilities.

The word 'chance' is also sometimes used as a singular noun to refer to a chance that is not equal to zero or one. For example, the thesis of determinism is sometimes expressed by saying that there is no genuine chance in the world; 'the USA soccer team has a chance of winning the next world cup' says that the chance of their winning is greater than zero and less than one; to 'take a chance' is to embark on a venture that has a chance of success that is greater than zero but less than one; and so on.

Epistemic Probability

Laplace claimed that probability claims do not describe objective features of the world, but the state of human knowledge (and ignorance). Epistemic interpretations of probability all share this general feature.

The Classical Interpretation

If asked the probability of rolling an odd number on a six-sided die, most people would unhesitatingly respond with the answer one-half: there are six possible outcomes of the die roll, three of which are odd. This illustrates the *classical interpretation* of probability, often associated with Laplace (1819). According to this interpretation, the possible outcomes of a particular trial are divided into 'equally possible' cases, and the probability of outcome F on a particular trial is equal to the number of such cases in which F is true, divided by the total number of possible cases. Which cases are equally possible is determined by the *Principle of Indifference*, which states that two cases are equally possible in the absence of any reason to prefer one outcome over the other.

There are a number of difficulties with the principle of indifference. When the weather reporter claims that the probability of rain is 0.3, there seem to be at least two different types of outcome (rain and no rain) but it is hard to see what the 10 equally possible outcomes (three of which involve rain) could be. Even in much simpler cases, the principle of indifference is not precise enough to specify the equally possible cases. An urn contains both red balls and blue balls; some of the red balls are light red, and some are dark red. What is the probability that the next ball drawn is blue: are there two equally possible cases (red and blue) or three (light red, dark red, and blue)? This problem is further exacerbated when there are infinitely many possible outcomes.

Subjective Interpretations

According to the subjective interpretation of probability (whose proponents are variously called subjectivists' personalists, or Bayesians), probability measures the degrees of belief or *credences* of a certain type of idealized rational agent. Subjectivists regard probability theory as analogous to *formal logic*. Logic alone never tells an agent whether to believe a proposition, but it does require that an agent's beliefs be consistent with one another. For example, the rules of logic forbid an agent to believe 'it will rain today,' 'if it rains today, it will not snow today,' and 'it will snow today' at the same time. According to subjectivists, probability theory is an extension of logic that allows an agent to have degrees of confidence in a proposition that range in value between zero and one. Probability theory alone will not tell the agent how much confidence to place in the proposition 'it will rain today,' but rather imposes a consistency-like constraint (usually called *coherence*) on the totality of the agent's degrees of belief.

The numerical value of an agent's degree of belief in some proposition is connected with the agent's evaluation of bets on the truth of that proposition. For example, if an agent has a degree of belief of 0.3 for the proposition 'it will rain today,' that means that he or she would find a bet on rain at 3:7 odds – a bet that costs the agent three units, and buys a chance to win 10 units (for a profit of seven units) if it rains – to be fair. This connection between degrees of belief, utility, and evaluation forms the foundation of a theory of rational choice wherein an agent always acts so as to maximize expected utility (see Rational Choice Explanation: Philosophical Aspects; Elicitation of Probabilities and Probability Distributions).

This connection between degrees of belief and evaluation is often deployed in arguments that rational degrees of belief conform to the formal rules of probability (see Probability Theory: Formal). When an agent's degrees of belief conform to the probability calculus, then his or her overall system of evaluations will have a number of desirable features. For example, Ramsey (1926) showed that having one's degrees of belief conform to the rules of probability is a necessary condition for having one's overall system of *preferences* satisfy a set of plausible constraints. One such constraint is transitivity: if the agent prefers A to B and B to C, then the agent

prefers A to C. A second class of results comprises the so-called *Dutch Book* theorems. (See, for example, de Finetti, 1964 and Kemeny, 1955.) These results show that if and only if an agent's degrees of belief violate the rules of probability, the agent will be vulnerable to a 'Dutch Book.' A Dutch Book is a system of bets such that the agent finds each individual bet fair, but where the agent is guaranteed of suffering a net loss no matter what the outcome. For example, a bet on heads at 2:1 odds together with a bet of equal stakes on tails at 2:1 odds would constitute a Dutch Book, for the agent purchasing the two bets would be guaranteed of losing 2 units on one bet, and winning 1 unit on the other. (For more details, *see* Probability: Interpretations.)

The requirement that rational degrees of belief satisfy the rules of probability theory is a *synchronic* constraint on one's degrees of belief: it concerns an agent's degrees of belief at one particular time. Some subjectivists impose a further *diachronic* constraint on allowable methods for changing one's degrees of belief over time. Suppose that a rational agent's degrees of belief at some particular time are represented by the probability function P. If the agent then acquires some new piece of information I, then the agent's new degree of belief in any proposition A is equal to the *conditional probability* $P(A/I)$ (see *Probability: Formal* for the definition of conditional probability). This rule of updating degrees of belief by *conditionalizing* can also be motivated by a form of Dutch Book argument (proved by David Lewis and reported in Teller (1973)).

This rule for changing degrees of belief has been championed by so-called *Bayesians* as a theory of confirmation. According to this theory, a hypothesis H is confirmed or supported by a piece of evidence E just in case conditionalizing on E increases the probability of H. This probabilistic approach to confirmation theory has proven to be very fruitful. see also Howson and Urbach (1993).

The various theorems invoked by subjectivists are *formal* results, and there is considerable debate over their interpretation. In particular, there is disagreement about how, or even whether, the Dutch Book theorems really show that coherence or updating by conditionalization are canons of rationality. Critics of subjectivism point out that even an incoherent agent would be capable of following the proofs of these theorems, and hence of foreseeing the dangers of Dutch Books; thus coherence and updating by conditionalization are not really necessary for avoiding financial loss (*see* Howson and Urbach, 1993 for further discussion).

A second focus of criticism is the *subjective* nature of degrees of belief. Rational agents may differ in the degree to which they believe a proposition such as 'it will rain today'; indeed their degrees of belief may range all the way from zero to one. Moreover, the constraints of synchronic coherence plus updating by conditionalization do not rule out the possibility that one's degrees of belief are *dogmatic* in a certain respect. It is consistent with these constraints that an agent have degree of belief of 0.99 that each toss of a given coin will land heads, and after observing a thousand tails in a row, *still* believe that the *next* toss will almost certainly be heads. Subjectivists have been able to prove many interesting results about how an agent's degrees of belief *would* evolve in the face of new evidence *if* the agent's original degrees of belief (or *prior probabilities*) were nondogmatic in various ways (de Finetti's, 1964 representation theorem is one important example), but there is nothing in subjectivism that would *require* the agent to have such nondogmatic prior probabilities. For many, this latitude makes subjective degrees of belief ill-suited as a foundation for theories of rational choice or rational inference.

Logical Interpretations

Logical interpretations aim to avoid the problems of subjectivist interpretations by imposing further constraints on an agent's prior probabilities. Prior probabilities are determined by the logical structure of the space of outcomes, much as they are in the classical interpretation. Unlike the classical interpretation, however, probability assignments can reflect the bearing of evidence that discriminates among the possible outcomes. Thomas Bayes (1763) and Laplace (1819), who are commonly associated with the subjectivist and classical approaches (respectively), both endorsed a version of the logical approach to probability. Twentieth century proponents include Carnap (1950), Jeffreys (1961), and Jaynes (1973).

Logical interpretations of probability inherit many of the difficulties of the classical interpretation discussed above: logical considerations alone do not determine a unique prior probability distribution. Moreover, many of the most plausible candidates for logical probabilities – such as one in which a sequence of coin tosses is *independent and identically distributed* (see Probability Theory: Formal or Distributions, Statistical: Special and Discrete) – turn out to be dogmatic in just the manner described in Section Subjective Interpretations

The Need for Objective Interpretations

Although few would doubt that our ordinary concept of probability has an epistemic dimension, it seems implausible that *all* probability claims can be interpreted epistemically. Consider the claim that a photon has a certain probability of passing through a polarizer; it would be very odd indeed if the truth of this claim (and hence of quantum mechanics) depended on the extent to which some agent(s) may believe that the photon will pass through or on the extent to which available evidence confirms this hypothesis. At best, epistemic interpretations of probability will have to be supplemented by an objective interpretation of at least some probability statements.

Objective Interpretations of Probability

In practice, probabilities are often inferred from frequency data. To estimate the probability that a given coin will land heads when tossed, for example, one could toss it many times and observe the relative frequency with which the coin lands heads. Such relative frequencies are *objective* features of the world. Hence relative frequencies provide a natural resource for objective interpretations of probability claims.

Finite Relative Frequency Interpretation

Perhaps the simplest interpretation of a probability claim would be to understand it as a claim about an actual finite

relative frequency. For example, if a coin were tossed exactly 10 times and then destroyed, resulting in 'heads' on six of those 10 tosses, then the probability that the coin lands heads when tossed would be 0.6. The central difficulty with this definition of probability is that it rules out unrepresentative samples by definition. For example, if a coin is fair, i.e., has a probability of 0.5 of landing heads when tossed, that should not be taken to rule out the possibility of it landing hands on six of the 10 occasions on which it is tossed. According to the finite frequency definition of probability, however, if the probability of heads is 0.5, then *by definition* the coin will land heads on exactly half of its tosses. Thus, while finite relative frequencies provide indispensable evidence for probabilistic claims, probabilities cannot be identified with finite relative frequencies.

Limiting Relative Frequency Interpretation

In order to avoid these problems with finite frequencies, some philosophers have suggested that probabilities be identified with frequencies in an *infinite* sample. If the first n tosses of a coin yields $f(n)$ heads, then the relative frequency of heads in the first n tosses is $f(n)/n$. The *limiting relative frequency* of heads in the infinite sequence of tosses is the limit as n goes to infinity of $f(n)/n$. This approach to interpreting probability was pioneered by Venn (1866); leading twentieth century proponents of this interpretation (hereafter called *frequentists*) include Richard von Mises (1957) and Hans Reichenbach (1949).

One important problem with limiting relative frequencies is that they seldom, if ever, exist: no coin is actually tossed an infinite number of times. Thus the limiting relative frequencies that determine the probability of an outcome are really *hypothetical limiting relative frequencies*: to say that the probability of heads on a toss of a certain coin is 0.5 is to say that *if* the coin were to be tossed an infinite number of times, the limiting relative frequency of heads *would be* 0.5. This is a hypothetical or *counterfactual* claim; by itself, it says nothing about what features of the *actual world* make the probabilistic claim true. The problem of interpretation of counterfactuals is itself a thorny philosophical problem (*see* Counterfactual Reasoning, Qualitative: Philosophical Aspects; Counterfactual Reasoning, Quantitative: Philosophical Aspects).

One important issue that divides frequentists is whether it is possible to ascribe a probability value to the outcome of a single trial. One tradition, following von Mises (1957), views probability as a property of a sequence of trials, or a 'collective.' Any single coin toss belongs to infinitely many different sequences: the sequence of all coin tosses, the sequence of all tosses of that particular coin, the sequence of all tosses of that coin by imparting a particular force from a particular height above the floor, and so on. The limiting relative frequency of 'heads' may be different in each of these sequences. Thus, there is no unique probability of 'heads' on a given toss, but only a probability for 'heads' in a collective of tosses. While frequentists in this tradition deny that there is one privileged sequence of coin tosses, not just any sequence of tosses is fit to bear probability values. For example, if an infinite sequence of coin tosses were to yield an alternating sequence of 'heads' and 'tails,' the relative frequency of 'heads' would be 0.5. The regularity of the pattern of outcomes would nonetheless render this sequence inappropriate for characterizing a probability value. Rather, probabilities are limiting relative frequencies in a *random* sequence of trials. There is no universally accepted mathematical characterization of randomness, although this is currently an area of active research (see Earman 1986, Chap. VIII). For a more detailed exposition of this version of the limiting relative frequency interpretation, see Probability: Interpretations.

One feature of this approach is that it allows chances to take on values other than zero or one even if determinism is true. It may be that on any particular toss of a coin, the linear and angular momentum imparted to the coin together with its height above the floor and so on determine the outcome of the coin toss. Nonetheless, there may be a random sequence of coin tosses where the limiting relative frequency is not equal to zero or one, for the tosses comprising the collective need not be identical with respect to such determining factors.

The relegation of probabilities to properties of collectives would seem to be a drawback of this approach to probability. When deciding whether to bet on a coin toss, or to postpone a picnic, one wants to know the probability of heads *on that particular toss*, or the probability of rain *on this particular day*. Some frequentists, following Reichenbach, believe that it is possible to assign probability values to the outcomes of individual trials. This requires that a choice be made from among all of the sequences to which the trial in question belongs. For example, a particular day belongs in an infinite number of sequences of days, and the limiting relative frequency of rain need not be the same in all of them, so which sequence determines the probability of rain on the day in question? Presumably not just the sequence of all days, for then the probability of rain would not vary from day to day. A natural thought is that what is wanted is a sequence of days *just like this one*. No two days will be exactly alike in *all* respects, however; the best that can be hoped for is a sequence of days that are alike in all *relevant* respects. The key problem, then, is to say what makes some respects of similarity between days relevant and others irrelevant. One detailed proposal is developed in Salmon (1984, Chap. 3).

On this approach, determinism is compatible only with chances of zero or one. Suppose, for example, that on a given day, climactic conditions determine that it will rain. Then the appropriate sequence will presumably consist of days in which the same climactic conditions prevail, and hence the limiting relative frequency of rain within that sequence will be one.

Propensity Interpretation

The propensity interpretation of probability, first introduced by Karl Popper (1959), is motivated by some of the difficulties with the limiting relative frequency interpretation. According to this interpretation, a coin, in conjunction with the conditions in which it is tossed, has a certain *disposition* or *propensity* to land heads. The probability that the coin will land heads is a measure of the strength of this propensity. This propensity is a property of the particular experimental set-up, hence, it is appropriate to assign a probability to an individual trial. Moreover, this propensity provides grounds for the truth of claims about what the coin *would do* if tossed infinitely many times in identical (or relevantly similar) conditions.

While this is certainly a useful way of thinking about single case probabilities within a frequentist framework, critics wonder whether anything has really been added. A coin is about to be tossed: there is some property of this experimental set-up in virtue of which the coin *would* land heads with a certain limiting relative frequency if this experiment were repeated infinitely often; calling this property a 'propensity' of a certain strength does little to indicate just what this property is. Moreover, if propensities are Pnot simply identified with hypothetical limiting relative frequencies, it is hard to see what is supposed to ground the claim that propensities have the mathematical structure of probabilities.

Best System Interpretation

The philosopher David Lewis (1986) has proposed an interpretation according to which an objective probability claim is true when it is entailed by a probabilistic *law* (*see* Causes and Laws: Philosophical Aspects). On this view, objective probability is intimately connected with indeterminism. A probabilistic claim is a law if it belongs to the set of propositions that provides the best overall systematization of the world. In order to be the best systematization, a set of propositions must obtain the optimal balance between informativeness, simplicity, and fit with data. For example, consider the claim that a photon will pass through a polarizer with a probability of $\cos^2\theta$. Although the frequency with which photons do pass through when the experiment is performed may diverge from this number, the probability value is firmly grounded in the principles of quantum mechanics, which overall provides an excellent fit with experimental data, and is a powerful and elegant theory.

Unlike the frequency interpretations, however, the truth conditions for probabilistic claims provided by the best system interpretation are not very precise. It is not clear what makes one set of propositions simpler than another, nor how simplicity is to be balanced against informativeness and goodness of fit.

Theoretical Role Interpretations

One further line of interpretation is influenced by the philosophical school of pragmatism (*see* Pragmatism: Philosophical Aspects). In this approach, objective probability claims are to be understood in terms of their role in our overall conceptual system. In particular, to understand a probability claim is to know what kinds of evidence would support or undermine that claim, and how to act accordingly if one accepts that claim as true. This approach has been developed by Mellor (1971), Skyrms (1980), and Lewis (1986).

Consider an agent whose degrees of belief conform to the rules of probability (see Section Subjective Interpretations), and let *Ch* be a function that assigns numerical values to propositions. The agent takes *Ch* to be objective chance precisely if his or her degrees of belief are guided by beliefs about the values of *Ch*. For example, suppose that the agent believes *Ch* ('it will rain today') to be 0.3. If the agent understands *Ch* to be objective chance, then his or her degree of belief in the proposition 'it will rain today' will also be 0.3. This degree of belief, together with the agent's utility function, can then be used to decide whether or not to postpone a picnic. In this way, then, the agent's beliefs about objective chances influence his or her decisions. Moreover, since objective probability claims are now propositions about which the agent has degrees of belief, the apparatus of Bayesian confirmation theory shows how those degrees of belief will change in the face of new evidence.

This approach has a number of virtues, not least of which is that it explains why objective chances have the formal properties of probability theory: in order for beliefs about chances to guide degrees of belief in the appropriate way, objective chances must have the same formal structure as rational degrees of belief. There is, however, an important drawback to this approach. In analyzing the meaning of objective probability claims in terms of their theoretical role instead of their truth conditions, this account does not answer what many have taken to be the central question of objective probability: what must be the case in the world in order for the (objective) probability of rain to be 0.3?

Conclusion

Interest in the philosophical aspects of probability dates back to the origins of the formal theory of probability in the seventeenth century. Epistemic and frequency interpretations of probability have been around for well over a century. In the second half of the twentieth century, interest in objective interpretations of single-case probabilities has dramatically increased the range of interpretive options available. Given the current lack of consensus, and the increasingly widespread use of probabilistic concepts in both philosophy and everyday life, it is safe to predict that interest in the problem of the interpretation of probability will continue into the twenty-first century.

See also: Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Probabilistic Thinking; Probability Theory: Formal; Probability Theory: Interpretations; Problem Solving and Reasoning, Psychology of.

Bibliography

Achinstein, Peter, 2001. The book of evidence. Oxford University Press, New York.
Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society 53, 370–418.
Carnap, R., 1950. Logical Foundations of Probability. University of Chicago Press, Chicago.
Colyvan, Mark, 2004. The Philosophical Significance of Cox's Theorem. International Journal of Approximate Reasoning 37 (1), 71–85.
de Finetti, B., 1964. Foresight: its logical laws, its subjective sources. In: Kyburg, H.E., Smokler, H.E. (Eds.), Studies in Subjective Probability. Wiley, New York.
Earman, J., 1986. A Primer on Determinism. Dordrecht: Reidel. Chaps. II, VIII.
Howson, C., Urbach, P., 1993. Scientific Reasoning: The Bayesian Approach, 2nd edn. Open Court, Chicago.
Howson, C., 1995. Theories of probability. British Journal for the Philosophy of Science 46, 1–32.
Jaynes, E.T., 1973. The well-posed problem. Foundations of Physics 3, 477–493.

Jeffreys, H., 1961. Theory of Probability, 3rd edn. Clarendon Press, Oxford, UK.
Kemeny, J.G., 1955. Fair bets and inductive probabilities. Journal of Symbolic Logic 20, 263–273.
Laplace, P.S.de, 1819. Essai Philosophique sur les Probabilités. Mme. Ve Courcier, Paris.
Lewis, D., 1986. Philosophical papers, vol. II. Oxford University Press, Oxford, UK. see especially the introduction and Chap. 19.
Mellor, Hugh, David, 2005. Probability: A philosophical introduction. Routledge, New York.
Mellor, D.H., 1971. The Matter of Chance. Cambridge University Press, Cambridge, UK.
von Mises, R., 1957. Probability, Statistics, and Truth, 2nd rev. English edn. Allen and Unwin, London.
Popper, K., 1959. The Propensity interpretation of probability. British Journal for the Philosophy of Science 10, 25–42.
Ramsey, F., 1926. Truth and probability. In: Mellor, D.H. (Ed.), Philosophical Papers. Cambridge University Press, Cambridge, UK.
Reichenbach, H., 1949. The Theory of Probability. University of California Press, Berkeley, CA.
Salmon, W.C., 1967. The Foundations of Scientific Inference. University of Pittsburgh Press, Pittsburgh, PA. see especially Chap. V.
Salmon, W.C., 1984. Scientific Explanation and the Causal structure of the world. Princeton University Press, Princeton, NJ. see especially Chap. 3.
Sklar, L., 1993. Physics and Chance. Cambridge University Press, Cambridge, UK. see especially Chap. 3.
Skyrms, B., 1980. Causal Necessity. Yale University Press, New Haven, CT.
Teller, P., 1973. Conditionalization and observation. Synthese 26, 218–258.
Venn, J., 1866. The Logic of Chance. Macmillan, London.

Probability Density Estimation

David W Scott, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12095–12099, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The probability density function is useful for summarizing and exploring a set of data. When the parametric form is unknown, then a nonparametric estimator such as the histogram is appropriate. This article illustrates the presentation of data by the histogram compared to the empirical distribution function. The steps required to construct a histogram are described, and data-based methods to accomplish this task are presented. Finally, the more efficient kernel density estimator is introduced and illustrated with real data.

The probability density function is useful for summarizing and exploring a set of data. When the parametric form is unknown, then a nonparametric estimator such as the histogram is appropriate. This article illustrates the presentation of data by the histogram compared to the empirical distribution function. The steps required to construct a histogram are described, and data-based methods to accomplish this task are presented. Finally, the more efficient kernel density estimator is introduced and illustrated with real data.

Associated with each random variable, X, measured during an experiment is the cumulative distribution function (cdf), $F(x)$. The value of the cdf at $X = x$ is the probability that $X \leq x$. Features in a set of random data may be computed analytically if the cumulative distribution function is known. The probability that a random sample falls in an interval (a, b) may be computed as $F(b) - F(a)$. This formula assumes the values of X are continuous. Discrete random variables taking on only a few values are easy to handle; thus, this article focuses on continuous random variables with at least an interval scale.

The same information may be obtained from the probability density function (pdf), $f(x)$, which equals $\Pr(X = x)$ if X takes on discrete values, or the derivative of $F(x)$ if X is continuous. If either $F(x)$ or $f(x)$ is known, then any quantity of interest such as the variance may be computed analytically or numerically. Otherwise, such quantities must be estimated using a random sample $x_1, x_2, \ldots, x_n$. The relative merits of estimating $F(x)$ vs. $F(x)$ are an important part of this article.

Density estimation takes two distinct forms – parametric and nonparametric – depending on prior knowledge of the parametric form of the density. If the parametric form $f(x|\vec{\theta})$ is known up to the p parameters $\vec{\theta} = (\theta_1, \ldots, \theta_p)$, then the parameters may be estimated efficiently by maximum likelihood or Bayesian algorithms. When little is known of the parametric form, nonparametric techniques are appropriate. The most commonly used estimators are the histogram and kernel methods, as discussed below.

Histograms and CDFs

Grouping and counting the relative frequency of data to form a histogram may be taught at the beginning of every statistics course, but the concept seems to have emerged only in 1662 when John Graunt studied the Bills of Mortality and constructed a crude life table of the age of death during those times of plague (Westergaard, 1968). Today the histogram and its close relative, the stem-and-leaf plot, provide the most important tools for exploring and representing structure in a set of data. Studying the shape of a well-constructed histogram can verify an assumption of normality, or reveal departures from normality such as skewness, kurtosis, or multiple modes. The recent study of the theoretical properties of histograms has led to concrete methods for computing well-constructed figures as discussed below.

With only a random sample of data, the cdf may be estimated by the empirical cdf (ecdf), denoted by $F_n(x)$. The value of the ecdf at x is the fraction of the sample with values less than or equal to x, that is, $\{\#x_i \leq x\}$. Now the probability any one of the x_i is less than x is given by the true cdf at x. Hence, on average, the value of $F_n(x)$ is exactly $F(x)$, and $F_n(x)$ is an unbiased estimator of $F(x)$. The inequality $\{X_i \leq x\}$ is a binomial random variable, so that the variance of $F_n(x)$ may be computed as $F(x)(1 - F(x))/n$. In fact, there is no other unbiased estimator with smaller variance.

The graph of the empirical cdf looks like a staircase function. **Figure 1** displays the ecdf of the treatment times estimated by 33 physicians when presented with a hypothetical patient suffering from extreme migraine headaches (Hebl and Xu, 2000). The dotted line is the cdf of the fitted normal density $N(31.3, 9.9^2)$, which appears to fit reasonably well. **Figure 2** displays a histogram of the same data, which does not appear to be normal after all. Social scientists often are taught to favor the ecdf over the histogram, but in fact aside from computing probabilities, the ecdf is not well-suited for the easy understanding of data.

The probability density function, $f(x)$, carries the same information as the cdf. Since the derivative of the empirical cdf $F_n(x)$ is not well-defined, estimators that rely on grouping have been devised. A convenient interval, (a, b), containing the data is chosen, and subdivided into m 'bins' of equal width $h = (b - a)/m$. The value h is called the bin width (and is often called the smoothing parameter). A 'well-constructed' histogram must specify appropriate values for a, b, h, and m. Given such parameters, the number of samples falling into each of the

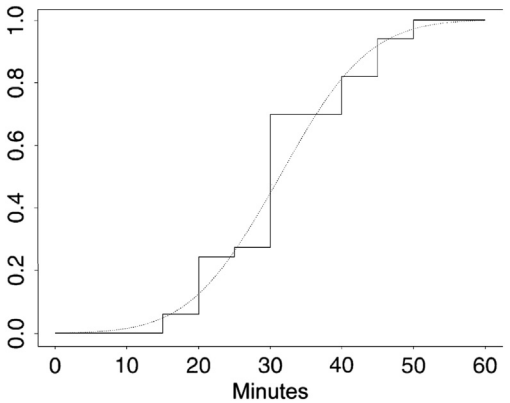

Figure 1 Empirical cumulative distribution function of physicians' estimated treatment times for a patient suffering migraine headaches. The dotted line is the cdf of the normal fit.

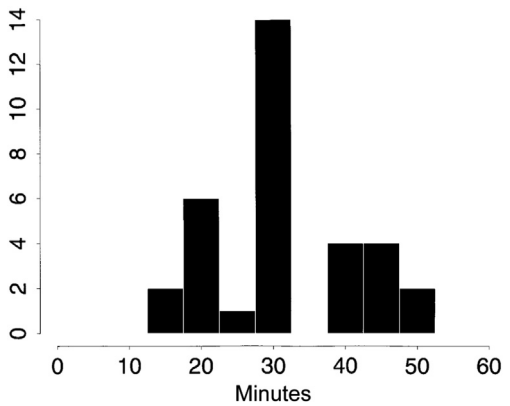

Figure 2 Frequency histogram of treatment times.

m bins is tallied. Label the m bin counts $\nu_1, \nu_2, \ldots, \nu_m$ from left to right. If the interval (a, b) contains all the samples, then $\sum_{k-1}^{m} \nu_k = n$. A *frequency histogram* is simply a plot of the bin counts, as in Figure 2. Although the values are all multiples of 5 min, the data may be treated as continuous. However, the bin width should only be a multiple of 5 and care should be exercised to choose a and h so that no data points fall where two bins adjoin. Here, $a = 2.5$, $b = 57.5$, and $h = 5$. The histogram suggests a departure from normality in the form of three modes. Re-examine the ecdf in Figure 1 and try to identify the same structure.

A true probability density is a non-negative function that integrates (sums) to 1. Hence, the *density histogram* is a simple normalization of the frequency histogram in the kth bin,

$$\hat{f}(x) = \frac{\nu_k}{nh} \qquad [1]$$

The normalization is important when the data are heavily skewed and unequal bin widths are constructed. In such cases, the interval (a, b) is partitioned into m bins of width $h_1, h_2, \ldots, h_m$. The bin counts $\nu_1, \nu_2, \ldots, \nu_m$ should *not* be plotted as they are not relatively proportional to the true density histogram, which is now given by $f(x)=\nu_k/(nh_k)$ in the kth bin.

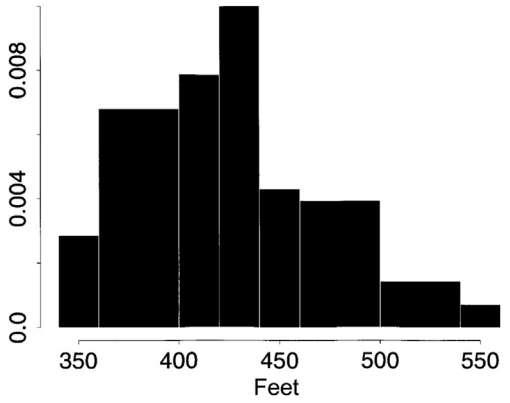

Figure 3 Density histogram of the distances of Mark McGwire's 70 home runs in the 1998 season.

Figure 3 displays a density histogram of the distances of the 70 home runs hit by Mark McGwire during his record-setting 1998 American baseball season. A bin width of 20 ft was chosen, and then several adjacent bins were collapsed to make the histogram look unimodal. The long right tail reflects the booming home runs for which McGwire is known.

Bin Width Construction

Rules for determining the number of bins were first considered by Sturges in 1926. Such rules are often designed for normal data, which occur often in practice. Sturges observed that the mth row of Pascal's triangle, which contains the combinatorial coefficients $\binom{m-1}{0}, \binom{m-1}{1}, \ldots, \binom{m-1}{m-1}$, when plotted, appears as an ideal frequency histogram of normal data. (This is a consequence of the central limit theorem for large m.) The sum of these 'bin counts' is 2^{m-1} by the binomial expansion of $(1+1)^{m-1}$. Solving the equation $n=2^{m-1}$ for the number of bins, we obtain Sturges' rule

$$m = 1 + \log_2 n \qquad [2]$$

This rule is widely applied in computer software.

The modern study of histograms began in Scott (1979) and Freedman and Diaconis (1981). The optimization problem formulated was to find the bin width h to make the average value of the 'distance' $\int [\hat{f}(x) - f(x)]^2 \, dx$ as small as possible. The solution for the best bin width is

$$h^* = \left[\frac{6}{\int f'(x)^2 dx}\right]^{1/3} n^{-1/3} \qquad [3]$$

While the integral of the derivative of the unknown density is required to use this formula, two useful applications are available. First, the *normal reference rule* gives

$$h^* = 3.5\sigma n^{-1/3} \qquad [4]$$

since the value of the integral for a normal density is $(4\sqrt{\pi}\sigma^3)^{-1}$. The second application was given by Terrell and Scott (1985) who searched for the 'easiest' smooth density

(which excludes the uniform density) and found that the number of bins in a density should always exceed $\sqrt[3]{2n}$ or that the bin width should always be less than

$$h^* \leq 3.73\sigma n^{-1/3} \qquad [5]$$

These choices are called *oversmoothed rules*, as the optimal choices will never be more extreme for other less smooth densities. Notice that the normal reference rule is very close to the upper bound given by the oversmoothed rule. For the McGwire data, the oversmoothed rule gives $h^* \leq 42'$.

More sophisticated algorithms for estimating a good bin width for a particular data set are described in Scott (1992) and Wand (1997). For example, a cross-validation algorithm (Rudemo 1982) that attempts to minimize the error distance directly leads to the criterion

$$\mathrm{CV}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)n^2 h}\sum_{k=1}^{m} v_k^2 \qquad [6]$$

The bin counts are computed for a selection of bin widths (and bin origins), the criterion computed, and the minimizer chosen, subject to the over-smoothed bound. The criterion tends to be quite noisy and Wand (1997) describes alternative plug-in formulae that can be more stable.

All of these rules give bin widths which shrink at the rate $n^{-1/3}$ and give many more bins than Sturges' rule. In particular, most computer programs give histograms which are *oversmoothed*. The default values should be overridden, especially for large datasets. Similarly, the error of the best histogram decreases to zero at the rate $n^{-2/3}$, which can be improved to $n^{-4/5}$ as described below. However, the histogram remains a powerful and intuitive choice for density estimation and presentation.

Advanced Algorithms

Density estimators with smaller errors are continuous functions themselves. The easiest such estimator is the frequency polygon, which connects the midpoints of a histogram. A separate error theory exists. For example, the oversmoothed bin width for a histogram from which the frequency polygon is drawn is given by

$$h^* \leq 2.33\sigma n^{-1/5}$$

Note that this bin width is wider than the rules for the histogram alone.

Many other estimators exist, based on splines, wavelets, Fourier series, or local likelihood, but all are closely related to the kernel estimator of Rosenblatt (1956) and Parzen (1962), which is described here. The kernel can be any probability density function, but the normal density is a common and recommended choice. Here the smoothing parameter h is the standard deviation of the normal kernel. If the normal density $N(\mu,\sigma^2)$ at a point x is denoted by $\phi(x|\mu, \sigma)$, then the kernel estimator is

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n}\phi(x|x_i, h)$$

This form shows that the kernel estimator is a mixture of normal densities which are located at the data points. The oversmoothed smoothing parameter value is

$$h^* \leq 1.14\sigma n^{-1/5}$$

An example is given below.

Visualization

The primary uses of density estimation are to display the structure in a set of data and to compare two sets of data. The histogram is adequate for the first task, but overlaying two histograms for comparison is difficult to interpret. The continuity and improved error properties of the kernel estimators excel for this purpose.

In the great home run race of 1998, Sammy Sosa finished second with 66 home runs. Figure 4 displays the empirical cdfs of the distances for the two home run kings (Keating and Scott 1999). This graph clearly shows that McGwire hits longer home runs than Sosa at every distance. In Figure 5, kernel estimates of the McGwire and Sosa data with h=10 are overlaid. The possible structure in these data is reinforced by the similarity of the two densities. Each has two primary modes at 370 ft and 425 ft, with smaller feature around 475 ft. In addition, McGwire's density clearly shows the right tail of his longest homers. In contrast, the modes in the original histogram of Figure 3 did not appear convincing, and were eliminated by locally adjusting the bin widths. The continuity of the kernel and other estimators makes the assessment of modes much easier.

Multivariate Densities

Histograms can be constructed in the bivariate case and displayed in perspective, but cannot be used to draw contour plots. Kernel methods are well-suited for this purpose. Given a set of data $(x_1, y_1), \ldots, (x_n, y_n)$, the bivariate normal kernel estimator is given by

$$\hat{f}_K(x,y) = \frac{1}{n}\sum_{i=1}^{n}\phi(x|x_i, h)\phi(y|y_i, h)$$

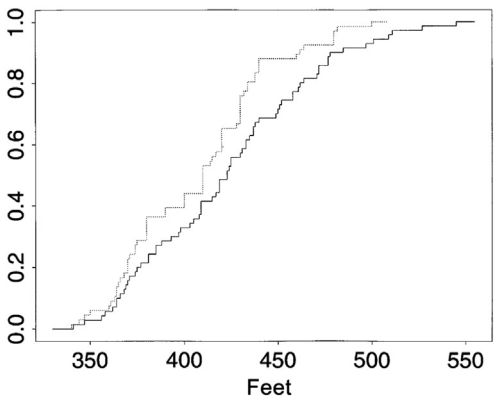

Figure 4 Empirical cumulative distribution functions of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in 1998.

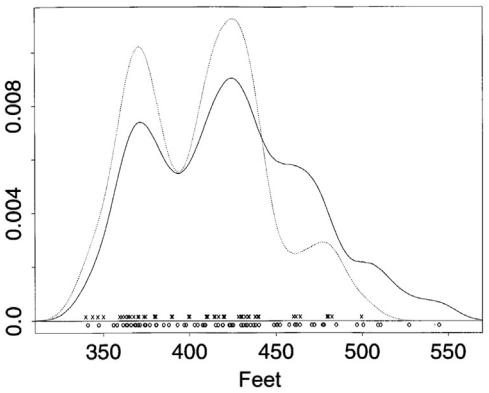

Figure 5 Kernel density estimates of the distances of Mark McGwire's 70 (solid line) and Sammy Sosa's 66 (dotted line) home runs in the 1998 season. McGwire's data are plotted below the horizontal line and Sosa's above.

A different smoothing parameter can be used in each variable.

For three- and four-dimensional data, the kernel estimator can be used to prepare advanced visualization renderings of the data. Scott (1992) provides an extension set of examples.

Other Materials

A number of other monographs exist that delve into various aspects of nonparametric approaches. A thorough survey of all topics by top researchers is contained in Schimek (2000). Early works popularizing these ideas include Tapia and Thompson (1978) and Silverman (1986). Wand and Jones (1994) and Simonoff (1996) focus on kernel methods. Fan and Gijbels (1996) and Loader (1999) focus on local polynomial methods. General smoothing ideas are discussed by Härdle (1991) and Bowman and Azzalini (1997). Many other excellent references may be followed in these books.

See also: Probability Theory: Formal.

Bibliography

Botev, Z.I., Grotowski, J.F., Kroese, D.P., 2010. Kernel density estimation via diffusion. The Annals of Statistics 38 (5), 2916–2957.

Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus. Clarendon Press, Oxford.

Elgammal, Ahmed, Duraiswami, Ramani, Harwood, David, Davis, Larry S., 2002. Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance. Proceedings of the IEEE 90 (7), 1151–1163.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

Fraley, Chris, Raftery, Adrian E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97 (458), 611–631.

Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57, 453–476.

Härdle, W., 1991. Smoothing Techniques: With Implementation in S. Springer, New York.

Hebl, M., Xu, J., 2000. In: Lane, D. (Ed.), Rice Virtual Lab in Statistics. http://onlinestatbook.com/rvls/index.html.

Keating, J.P., Scott, D.W., 1999. A primer on density estimation for the great home run race of ' 98. Stats 25, 16–22.

Loader, C., 1999. Local Regression and Likelihood. Springer, New York.

Parzen, E., 1962. On estimation of probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27, 832–837.

Rudemo, M., 1982. Empirical choice of histogram and kernel density estimators. Scandinavian Journal of Statistics 9, 65–78.

Schimek, M.G. (Ed.), 2000. Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons, New York.

Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Tapia, R.A., Thompson, J.R., 1978. Nonparametric Probability Density Estiamtion. Johns Hopkins University Press, Baltimore, MD.

Terrell, G.R., Scott, D.W., 1985. Oversmoothed non-parametric density estimates. Journal of the American Statistical Association 80, 209–214.

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. Chapman and Hall, New York.

Wand, M.P., 1997. Dated-based choice of histogram bin width histogram. The American Statistician 51, 59–64.

Westergaard, H., 1968. Contributions to the History of Statistics. Agathon, New York.

Probability Theory: Formal

J Michael Steele, University of Pennsylvania, Philadelphia, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12099–12104, © 2001, Elsevier Ltd.

Abstract

Probability theory is a branch of mathematics that evolved from the investigation of social, behavioral, and physical phenomena that are influenced by randomness and uncertainty. This article begins its survey of probability theory with a discussion of the impact of A.N. Kolmogorov's framework for the mathematical understanding of probability and the role that his axioms had in transforming probability from a modeling art to a mathematical science. The survey then discusses the important notion of a random variable, and the tools for describing such variables, including means, variances, distribution functions, densities, and the fundamental property of independence. The article also discusses the law of large numbers – the mathematical result that underlies the popular 'law of averages' and the central limit theorem – the mathematical result that suggests why the 'bell curve' is present in many social and physical phenomena. The article concludes by reviewing sources of further information on probability theory.

Probability theory is a branch of mathematics that has evolved from the investigation of social, behavioral, and physical phenomena that are influenced by randomness and uncertainty. For much of its early life, probability theory dealt almost exclusively with gambling games, and, even though there were important contributions made by such distinguished mathematicians as Pierre de Fermat, Blaise Pascal, and Pierre-Simon de Laplace, the field lacked respectability and failed to attract sustained attention.

One cause of the slow development of probability theory in its early days was the lack of a widely accepted foundation. Unlike the geometry of Euclid, or even the analytical investigations of Newton and Leibnitz, the theory of probability seemed to be eternally tied to the modeling process. In this respect, probability theory had greater kinship with the theories of heat or elasticity than with the pristine worlds of geometry or algebra.

Over time, foundations for probability were proposed by a number of deep thinking individuals including von Mieses, de Finetti, and Keynes, but the approach that has come to be the most widely accepted is the one that was advanced by Andrey Nikolayevich Kolmogorov (1933).

Kolmogorov's approach to the foundations of probability theory developed naturally from the theory of integration that was introduced by Henri Lebesgue and others during the first two of decades of the twentieth century. By leaning on the newly developed theory of integration, Kolmogorov demonstrated that probability theory could be simply viewed as another branch of mathematics. After Kolmogorov's work, probability theory had the same relationship to its applications that one finds for the theory of differential equations. As a consequence, the stage was set for a long and productive mathematical development.

To be sure, there are some philosophical and behavioral issues that are not well addressed by Kolmogorov's bare-bones foundations, but over the years, Kolmogorov's approach has been found to be adequate for most purposes. The Kolmogorov axioms are remarkably succinct, yet they have all the power that is needed to capture the physical, social, and behavioral intuition that a practical theory of probability must address.

Kolmogorov's Axiomatic Foundations

Central to Kolmogorov's foundation for probability theory was his introduction of a triple $(\Omega, \mathscr{F}, P)$ that is now called a probability space. The triple's first element, the 'sample space' Ω, is only required to be a set, and on the intuitive level, one can think of Ω as the set of all possible outcomes of an experiment. For example, in an experiment where one rolls a traditional six-faced die, then one can take $\Omega = \{1, \ldots, 6\}$.

The second element of Kolmogorov's probability space is a collection $\mathscr{F}$ of subsets of Ω that satisfy three basic consistency properties that will be described shortly. On the intuitive level, one can think of the elements of $\mathscr{F}$ as 'events' that may occur as a consequence of the experiment described by $(\Omega, \mathscr{F}, P)$. Thus, to continue with the example of rolling a die, the set $A = \{1, 3, 5\} \subset \Omega$ would correspond to the event that one rolls an odd number.

Two of the three consistency properties that Kolmogorov imposes on the $\mathscr{F}$ are quite trivial. First, $\mathscr{F}$ is required to contain Ω. Second, $\mathscr{F}$ must be closed under complementation; so, for example, if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$ where $A^c = \{\omega : \omega \in \Omega$ and $\omega \notin A\}$. This third condition is only a bit more elaborate; the collection $\mathscr{F}$ must be closed under countable unions. Thus, if $A_k \in \mathscr{F}$ for $k = 1, 2, \ldots$, then one requires that the union $A_1 \cup A_2 \cup \ldots$ of all the events in the countable set $\{A_k : 1 \leq k < \infty\}$ must again be an element of $\mathscr{F}$.

The most interesting element of Kolmogorov's triple $(\Omega, \mathscr{F}, P)$ is the 'probability measure' P. Formally, P is just a function that assigns a real number to each of the elements of $\mathscr{F}$, and, naturally enough, one thinks of $P(A)$ as the probability of the event A. Thus, one can specify a probability model for the outcome of rolling a fair die, by 'defining' P for $A \subset \Omega$ by $P(A) = 1/6|A|$, where $|A|$ denotes the number of elements of the set A.

For sample spaces Ω that are not finite, more care is required in the specification of the probability measures P. To deal with general Ω, Kolmogorov restricted his attention to those P that satisfy three basic axioms:

Axiom 1. For all $A \in \mathscr{F}$, one has $P(A) \geq 0$.
Axiom 2. $P(\Omega) = 1$.

Axiom 3. For any countable collection $\{A_k \in \mathscr{F}: 1 \leq k < \infty\}$ with the property that $A_j \cap A_k = \Phi$ for all $j \neq k$, one has

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Axioms 1 and 2 just reflect the intuitive view that $P(A)$ measures the frequency with which A occurs in an imagined sequence of repetitions of the experiment $(\Omega, \mathscr{F}, P)$. For most mathematicians, axiom 3 also just reflects the most basic intuition about the way probabilities should behave. Nevertheless, axiom 3 is more subtle because it deals with an infinite collection of sets, and in some ways, such collections are outside of our direct experience. This has lead some philosophers to examine the possibility of avoiding Kolmogorov's third axiom, and over the years various attempts have been made to replace Kolmogorov's third axiom with the simpler assumption of 'finite additivity.'

Random Variables

In most applications of probability theory, the triple $(\Omega, \mathscr{F}, P)$ that gives life to probability as a rigorous mathematical subject is almost invisible. In practice, builders of probability models take advantage of various shortcuts that have the effect of keeping the probability triple at a respectful distance. The most important of these shortcuts is the notion of a 'random variable.'

On the intuitive level, a random variable is just a number that depends on a chance-driven experiment. More formally, a random variable X is a function from Ω to the real numbers with the property that $\{\omega: X(\omega) \leq t\} \in \mathscr{F}$ for all t. What drives this definition is that one inevitably wants to talk about the probability that X is less than t, and for such talk to make sense under Kolmogorov's framework, the set $\{\omega: X(\omega) \leq t\}$ must be an element of $\mathscr{F}$. Random variables make the modeler's job easier by providing a language that relates directly to the entities that are of interest in a probability model.

The Distribution Function and Related Quantities

There are several ways to specify the basic probabilistic properties of a random variable, but the most fundamental description is given by the 'distribution function' of X, which is defined by $F(t) = P(X \leq t)$. Knowledge of the distribution function tells one everything that there is to know about the probability theory of the random variable. Sometimes it even tells too much.

The knowledge one has of a random variable is often limited, and in such cases it may be useful to focus on just part of the information that would be provided by the distribution function. One common way to specify such partial information is to use the 'median' or the 'quantiles' of the random variable X. The median is defined to be a number m for which one has $P(X < m) \leq 1/2 \leq P(X \leq m)$. Roughly speaking, the median m splits the set of outcomes of X into two halves so that the top half and the bottom half each have probability that is close to one-half.

The quantile x_p does a similar job. It splits the possible values of X into disjoint sets, so that one has $P(X < x_p) \leq p \leq P(X \leq x_p)$. When $p = 1/2$, the quantile x_p reduces to the median, and, when $p = 1/4$ or $p = 3/4$, then x_p is called the 'lower quartile,' or the 'upper quartile,' respectively.

Mass Functions and Densities

If a random variable X only takes values from a finite or countable set S, then X is called a 'discrete' random variable. In such cases, one can also give a complete description of the probabilistic properties of X by specification of the 'probability mass function,' which is defined by setting $f(x) = P(\{x\})$ for all $x \in S$. Here, one should note that the probability mass function f, permits one to recapture the distribution function by the relation

$$F(t) = \sum_{x \leq t, x \in s} f(x) \quad \text{for all } t$$

There are many important random variables whose values are not confined to any countable set, but for some of these random variables, there is still a description of their distribution functions that is roughly analogous that provided for discrete random variables. A random variable X is said to be 'absolutely continuous' provided that there exist a function f such that the distribution function of X has a representation of the form

$$F(t) = \int_0^t f(x)\mathrm{d}x \quad \text{for all } t$$

The function f in this representation is called the 'density' of X, and many random variables of practical importance have densities. The most famous of these are the 'standard normal' (or, 'standard Gaussian') random variables that have the density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for all } -\infty < x < \infty$$

which has a key role in the Central Limit Theorem that will be discussed shortly.

Much of the theory of probability can be developed with just the notion of discrete or absolutely continuous random variables, but there are natural random variables that do not fit into either class. For this reason, the distribution function remains the fundamental tool for describing the probabilities associated with a random variable; it alone is universal.

Mathematical Expectation: A Fundamental Concept

If X is a discrete random variable, then its 'mathematical expectation' (or just expectation, for short) is defined by

$$E(X) = \sum_{x \in s} x f(x)$$

where f is the probability mass function of X. To illustrate this definition, one can take X to be the outcome of rolling a fair die, so that $f(x) = 1/6$ for all $x \in S = \{1, 2, \ldots, 6\}$ and

$$E(X) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + \cdots + 6\cdot\frac{1}{6} = \frac{7}{2}$$

In a way that is parallel – yet not perfectly so – the expectation of an absolutely continuous random variable X is defined by the integral

$$E(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x$$

Here one needs to work a bit harder to illustrate the definition. Consider, for example, an experiment where one chooses a number at random out of the unit interval [0, 1]. From the intuitive meaning of the experiment, one has $P(A) = a$ for $A = [0, a]$, and, from this relationship, one can calculate that the density function for X is given by $f(x) = 1$ for $x \in [0, 1]$ and by $f(x) = 0$ otherwise. For the density f, one therefore finds that

$$E(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \int_{0}^{1} x\cdot 1\ \mathrm{d}x = \frac{1}{2}$$

From this formula, one sees that the expected value of a number chosen at random from the unit interval is equal to one-half, and for many people this assertion is perfectly reasonable. Still, one should note that the probability that X equals one-half is in fact equal to zero, so the probabilistic use of the word 'expected value' differs modestly from the day-to-day usage.

The probability distribution function and the expectation operation provide almost all of the language that is needed to describe the probability theory of an individual random variables. To be sure, there are several further notions of importance, but these may all be expressed in terms of the distribution or the expectation. For example, the most basic measure of dispersion for the random variable X is its 'variance,' which is defined in terms of the expectation by the formula

$$\mathrm{Var(X)} = \mathrm{E(X} - \mu)^2 \quad \text{where } \mu = \mathrm{E(X)}$$

Finally, the 'standard deviation' of X, which is defined to be the square root of Var(X), provides a useful measure of scale for problems involving X.

Introducing Independence

The world of probability theory does not stop with the description of a single random variable. In fact, it becomes rich and useful only when one considers collections of random variables – especially collections of 'independent' random variables.

Two events A and B in $\mathscr{F}$ are said to be independent provided that they satisfy the identity,

$$P(A \cap B) = P(A)\,P(B)$$

What one hopes to capture with this definition is the notion that the occurrence of A has no influence on the occurrence of B – and vice versa. Nevertheless, the quantities that appear in the defining formula of independence are purely mathematical constructs, and any proof of independence ultimately boils down to the proof of the defining identity in a concrete model $(\Omega, \mathscr{F}, P)$.

Consider, for example, the experiment of rolling two fair dice, one red and one blue. The sample space Ω for this experiment may be taken to be the 36 pairs (j, k) with $1 \le j \le 6$ and $1 \le k \le 6$, where one thinks of j and k as the number rolled on the red and blue die, respectively. Here, for any $A \subset \Omega$ one can set $P(A) = |A|/|\Omega|$ in order to obtain a model for a pair of fair dice. Under this probability model, there are many pairs of events that one can prove to be independent. In particular, one can prove that the event of rolling an even number on the blue die is independent of rolling an odd number on the red die. More instructively, one can also prove that the event of rolling an even number on the blue die is independent of the parity of the sum of the two dice.

Extending Independence

The concept of independence can be extended to random variables by defining X and Y to be independent provided that the events $\{X \le s\}$ and $\{Y \le t\}$ are independent for all real s and t. One easy consequence of this definition is that for any pair of monotone functions ϕ and ψ the random variables $\phi(X)$ and $\psi(Y)$ are independent whenever X and Y are independent. This mapping property reconciles nicely with the intuitive assertion: if X and Y have no influence on each other, then neither should $\phi(X)$ and $\psi(Y)$ have any influence on each other.

The importance of independence for probability theory will be underscored by the theorems of Section The Law of Large Numbers, but first the notion of independence must be extended to cover more than just pairs of random variables. For a finite collection of n random variables $X_1, X_2, \ldots, X_n$, the condition for 'independence' is simply that

$$\begin{aligned} &P(X_1 \le t_1, X_2 \le t_2, \ldots, X_n \le t_n) \\ &= P(X_1 \le t_1)P(X_2 \le t_2)\cdots P(X_n \le t_n) \end{aligned}$$

for all real values $t_1, t_2, \ldots, t_n$. Finally, an infinite collection of random variables $\{X_s : s \in S\}$ is said to be independent provided that every finite subset of the collection is independent.

The Law of Large Numbers

Any mathematical theory that hopes to reflect real-world random phenomena must provide some rigorous interpretation of the intuitive 'law of averages.' Kolmogorov's theory provides many such results, the most important of which is given in the following theorem.

Theorem 1 (*Law of Large Numbers*). Suppose that $\{X_i : 1 \le i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathscr{F}(\cdot)$, so $P(X_i \le t) = \mathscr{F}(t)$ for all $1 \le i < \infty$ and all real t. If the expectation $\mu = E(X_1)$ is well-defined and finite, then the random variables

$$\frac{1}{n}\{X_1 + X_2 + \cdots + X_n\}$$

converge to their common mean μ with probability one. More precisely, if one lets

$$A = \left\{\omega : \lim_{n\to\infty} \frac{1}{n}\{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)\} = \mu\right\}$$

then the event A satisfies $P(A) = 1$.

The Central Limit Theorem

The second great theorem of probability theory is the famous Central Limit Theorem. Although it is not tied as tightly to the meaning of probability as the Law of Large Numbers, the Central Limit Theorem is key to many of the practical applications of probability theory. In particular, it often provides a theoretical interpretation for the 'bell curve' that emerges in countless empirical investigations.

Theorem 2 (*Central Limit Theorem*). Suppose that $\{X_i{:}1 \le i < \infty\}$ is a sequence of independent random variables with a common distribution function $\mathscr{F}$. If these random variables have a finite variance $\mathrm{Var}(X_i) = \sigma^2$, then

$$\lim_{n\to\infty} P\left(\frac{1}{\sigma\sqrt{n}}\{X_1 + X_2 + \cdots + X_n - n\mu\} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\,\mathrm{d}u$$

Stochastic Processes

The most fundamental results of probability theory address the behavior of sums of independent random variables, but many applications of probability theory lead to sequences of random variables $\{X_n{:}1 \le n < \infty\}$ that may not be independent. Such sequences are called 'stochastic processes,' and they are of many different types.

The simplest nontrivial stochastic process is the 'Markov chain,' which is used to model random phenomena where X_{n+1} depends on X_n, but, given X_n, the value of X_{n+1} does not depend on the rest of the past $X_{n-1}, X_{n-2}, \ldots, X_1$. To construct such a process, one most often begins with an $n \times n$ matrix $T = \{p_{ij}\}$ with entries that satisfy $0 \le p_{ij} \le 1$ and with row sums that satisfy $p_{i1} + p_{i2} + \cdots + p_{in} = 1$. One then generates the Markov chain by making sequential selections from the set $\mathscr{S} = \{1, 2, \ldots, n\}$ in accordance with the rows of the 'transition matrix' T. Specifically, if $X_n = i$, then X_{n+1} is obtained by choosing an element of $\mathscr{S}$ in accordance with the probabilities (p_{ij}) given by the ith row of T.

A second group of stochastic processes that is of considerable practical importance is the set of 'martingales.' Roughly speaking, these are stochastic process that have the property that the expectation of X_{n+1} given the values of $X_n, X_{n-1}, \ldots, X_1$ is just equal to X_n. One reason that martingales are important is that they provide a model for the fortune of an individual who gambles on a fair game. Such gambling games are relatively unimportant in themselves, but many economic and financial questions can be reframed as such games. As a consequence, the theory of martingales has become an essential tool in the pricing of stock options and other derivative securities.

The theory of stochastic processes is not confined to just those sequences $\{X_n{:}1 \le n < \infty\}$ with the discrete index set $\{1 \le n < \infty\}$ and, in fact, almost any set $\mathscr{S}$ can serve as the index set. When one takes $\mathscr{S}$ to be the set of non-negative real numbers, the index is often interpreted as time, and in this case one speaks of continuous time stochastic processes. The most important of these are the 'Poisson process' and 'Brownian motion.' Brownian motion is arguably the most important stochastic process.

Directions for Further Reading

For a generation, Feller (1968) has served as an inspiring introduction to probability theory. The text only assumes a modest background in calculus and linear algebra, yet it goes quite far. The text of Dudley (1989) is addressed to more mathematically sophisticated readers, but it contains much material that is accessible to readers at all levels. In particular, Dudley (1989) contains many informative historical notes with careful references to the original sources.

For an introduction to the theory of stochastic processes, the text by Çinlar (1975) is recommended, and for an easy introduction to Brownian motion, martingales, and their applications in finance, one can consult Steele (2000).

For background on the early development of probability theory, the books of David (1962) and Stigler (1986) are useful, and the article by Doob (1994) helps make the link to the twenty-first century. Finally, for direct contact with the master, anyone does well to read Kolmogorov (1933).

See also: Probability Density Estimation; Probability Theory: Interpretations; Probability and Chance: Philosophical Aspects.

Bibliography

Çinlar, E., 1975. Introduction to Stochastic Processes. Prentice Hall, Englewood Cliffs, NJ.

David, F.N., 1962. Games, Gods, and Gambling: The Origins and History of Probability from the Earliest Times to the Newtonian Era. Griffin, London.

Doob, J.L., 1994. The development of rigor in mathematical probability (1900–1950). In: Pier, J.-P. (Ed.), Development of Mathematics 1900–1950. Birkhauser-Verlag, Basel, Switzerland.

Dudley, R.M., 1989. Real Analysis and Probability. Wadsworth-Brooks/Cole, Pacific Grove, CA.

Feller, W., 1968. An Introduction to Probability and Its Applications, third ed., vol. 1. Wiley, New York.

Kolmogorov, A.N., 1933. Grundbegriffe der Wahrscheinlich-keitsrechnung. Springer-Verlag, Berlin (Morrison, N., 1956. Foundations of the Theory of Probability. Chelsea, New York, Trans.).

Steele, J.M., 2000. Stochastic Calculus and Financial Applications. Springer, New York.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge, MA.

Probability Theory: Interpretations

Teddy Seidenfeld, Carnegie Mellon University, Pittsburgh, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12104–12110, © 2001, Elsevier Ltd.

Abstract

This article is divided into two sections. The first reviews the formal theories of quantitative and qualitative probability. Conditions are given when a quantitative probability exists that agrees with (or almost agrees with) a qualitative probability. The second section reviews four kinds of semantics for the quantitative theory: the Subjective, the Classical (or Logical), the Objective, and the Fisher–Kyburg interpretation. The Subjective interpretation is presented through deFinetti's Dutch Book argument. Variants of Logical interpretations are described as these relate to LaPlace's *Principle of Insufficient Reason*. Two varieties of Objective probability are considered: limiting frequency theory and chance (or propensity) theory. Last, the Fisher–Kyburg theory is sketched in relation to fiducial probability.

Qualitative and Quantitative Probability

Introduction

It is entirely unremarkable that the commonplace word 'probability' and its cognates display a wide variety of distinct meanings in ordinary language. It is used as a classificatory concept, as in "It is *probable that* it will snow this winter" or "It is *improbable that* a fair coin will land heads 10 times in succession." It is used as a (qualitative) comparative relation, as in "It is *more probable that* someone will be killed by lightning *than* by a falling meteorite next year." And, of course, it is used as a quantitative concept throughout science, e.g., Mendel's law, "The probability is 0.75 that the offspring of a cross between simple hybrids will show the dominant trait." What is worth noting, however, is that despite longstanding controversies over the legitimate interpretation of 'probability,' these diverse usages can nonetheless be linked easily to a few formal theories of mathematical probability.

The received formal theory of quantitative probability, due to Kolmogorov, is elegant. Mathematical probability is a (real-valued) function, $\boldsymbol{P}$, with values between 0 and 1 (inclusive), defined over a field of events closed under the Boolean operations of union, intersection, and complementation. The sure event Ω is certain, $\boldsymbol{P}(\Omega) = 1$, and $\boldsymbol{P}$ is additive over unions of disjoint events. The qualitative formal theory of probability uses a binary relation $\langle$ also defined on an algebra of events: $E \langle F$ means that event F is more probable than event E. The qualitative formal theory is presented through axioms that mimic the quantitative one.

The outline for this article is as follows. Sections The Formal Theory of Quantitative Probability and The Formal Theory of Qualitative Probability review relevant details of the received, formal (mathematical) theories of quantitative and qualitative probability, and some basic connections between them. (This section includes some technical results that the reader might bypass initially, to return to on a second reading.) Section Interpretations of Quantitative Probability surveys several interpretations of the quantitative theory, including Sections Subjective Probability, Logical (or Classical) Probability, Objective Probability, and an approach (Section The Fisher–Kyburg Interpretation) relating to Fisher's enigmatic fiducial probability. Additional readings are suggested in Section Additional Readings.

The Formal Theory of Quantitative Probability

Begin with Kolmogorov's (1956) mathematical theory for quantitative probability, dating from the 1930s. A measure space $\langle \Omega, \mathscr{F}, \boldsymbol{P} \rangle$ is triple comprised by:

1. a set of possibilities, Ω, whose elements, ω, are points of the space. A field $\mathscr{F} = \{E \text{ of subsets of } \Omega\}$. Elements of $\mathscr{F}$ are the events of the space. It is assumed that $\Omega \in \mathscr{F}$. Being a field, $\mathscr{F}$ is closed under finite unions, intersections, and complementation with respect to Ω. When $\mathscr{F}$ is infinite, usually it is required
2. that $\mathscr{F}$ is a σ-field, closed under countable unions and intersections.
3. A probability function $\boldsymbol{P}$ over $\mathscr{F}$ satisfying these axioms:
   3.1. $0 \leq \boldsymbol{P}(E) \leq 1$
   3.2. $\boldsymbol{P}(\Omega) = 1$
   3.3. $\boldsymbol{P}(E \cup \mathrm{F}) = \boldsymbol{P}(\mathrm{E}) + \boldsymbol{P}(\mathrm{F})$, whenever $\mathrm{E} \cap \mathrm{F} = \varnothing$. (finite additivity)
   3.4. $\boldsymbol{P}(\cup E_i) = \Sigma_i \boldsymbol{P}(E_i)$, whenever $E_i \cap E_j = \mathrm{F}\,\varnothing$ for $i \neq j$ (σ-additivity).

Moreover, when $\boldsymbol{P}(F) > 0$, a conditional probability given F, $\boldsymbol{P}(\cdot|F)$, is well defined over $\mathscr{F}$ by the equation

4. $\boldsymbol{P}(\cdot|F)\boldsymbol{P}(F) = \boldsymbol{P}(\cdot \cap F)$

Kolmogorov's mathematical theory provides a basis for each of the two nonquantitative uses of probability, illustrated above.

The classificatory concept of probability may be explicated as:

'E is *probable*' provided that $\boldsymbol{P}(E) > c > 0.5$, for some value c, determined by the context of the assertion. Then, 'E is *improbable*' means that the complementary event E^c is probable.

The (qualitative) comparative concept of probability may be explicated as:

'E is *more probable than* F' provided that
$$\boldsymbol{P}(E) > \boldsymbol{P}(F).$$

The Formal Theory of Qualitative Probability

A mathematical theory of qualitative probability is readily obtained from the quantitative one. The binary relation $\langle$ over events is said to be a (qualitative) comparative probability relation if it satisfies the following three axioms that parallel the first three axioms of Kolmogorov's theory.

5.1. $\langle$ is an acyclic, transitive relation over $\mathscr{F}$. That is, if $E \langle F$ and $F \langle G$ then $E \langle G$, and it is never the case that $E \langle E$. The qualitative relation that E and F are equiprobable events, denoted $E \approx F$, is defined by the condition that neither $E \langle F$ nor or $F \langle E$. Then $\approx$ is transitive and reflexive, i.e., $E \approx E$ always obtains.
5.2. $\varnothing \langle \Omega$ and for no event E, $E \langle \varnothing$.
5.3. Provided that $E \cap (F \cup G) = \varnothing$, $F \langle G$ if and only if $(E \cup F) \langle (E \cup G)$.

Say that a quantitative probability $\boldsymbol{P}$ agrees with the comparative probability relation $\langle$ on the condition that, $\boldsymbol{P}(E) < \boldsymbol{P}(F)$ if and only if $E \langle$ F. deFinetti (1931) asked whether every comparative probability relation $\langle$ has an agreeing quantitative probability. Kraft et al. (1959) established that the answer is no, even when $\mathscr{F}$ is a finite space. However, they also showed that for finite spaces, the following (stronger) additivity condition is both necessary and sufficient for an agreeing quantitative probability to exist.

5.4. Let $(E_1, \ldots, E_m)$ and $(F_1, \ldots, F_m)$ be two sequences of events from $\mathscr{F}$ such that, for each point ω in Ω, the same number of events from each sequence contain ω. If it is not the case that $F_i \langle E_i$ $(i = 1, \ldots, m-1)$, then it is not the case that $E_m \langle F_m$.

Aside: It is interesting to note that the very same condition, expressed for a quantitative probability, serves as a necessary and sufficient condition for extending the domain of a quantitative probability, Horn and Tarski (1948).

A different relationship between quantitative and qualitative probabilities is also worth consideration. Say that a quantitative probability $\boldsymbol{P}$ almost agrees with the comparative probability relation $\langle$ on the condition that $\boldsymbol{P}(E) < \boldsymbol{P}(F)$ only if $E \langle F$. That is, quantitative strict inequalities entail the corresponding qualitative (strict) inequality. However, when $\boldsymbol{P}$ almost agrees with $\langle$, it may be that $\boldsymbol{P}(E) = \boldsymbol{P}(F)$ and $E \langle F$. Kraft et al. (1959) provide an example of a finite space equipped with a qualitative probability relation that admits a continuum of almost agreeing quantitative probabilities but no agreeing quantitative probability. For finite spaces, the analog to (5.4) is the following necessary and sufficient condition for a qualitative probability to admit an almost agreeing quantitative probability.

5.5. Let $(E_1, \ldots, E_m)$ and $(F_1, \ldots, F_m)$ be two sequences of events from such that, for each point ω in Ω, fewer events from the first sequence contain ω than do events from the second sequence. If it is not the case that $F_i \langle E_i$ $(i = 1, \ldots, m-1)$, then $E_m \langle F_m$.

On infinite spaces, the situation is more complex and more interesting. The following illustrates how a qualitative probability may make finer discriminations than any quantitative probability, i.e., the qualitative relation $\langle$ is non-Archimedean.

Example

Let $\mathscr{F}$ be the set of all subsets of the positive integers $\{1, 2, \ldots\} = \Omega$. Let $\langle$ be a qualitative probability subject to the following two constraints:

1. Each integer, i.e., each point is as probable as any other. That is, for each pair of integers i and j, $i \approx j$.
2. The qualitative probability relation respects proper subsets. That is, whenever $E \subset F$ (but $E \neq F$) then $E \langle F$.

(*Aside*: To establish the existence of a qualitative probability satisfying these constraints requires an appeal to the Axiom of Choice.) Condition (1) ensures that no countably additive probability agrees with $\langle$, since any agreeing probability must assign each point of Ω probability 0. Condition (2) insures that no finitely additive probability agrees with $\langle$, because, by the fact just noted, $\boldsymbol{P}\{i, j\} = \boldsymbol{P}\{i\} + \boldsymbol{P}\{j\} = 0$. But by (2) $\{i\} < \{i, j\}$. There exists an almost agreeing (uniform) finitely additive probability for $\langle$, where $\boldsymbol{P}\{i\} = 0$ for each integer i. Hence, this qualitative probability ordering is finer than any quantitative probability can be.

Interpretations of Quantitative Probability

Subjective Probability

The contemporary theory of subjective (or personal) probability has its roots in the writings by Ramsey (1926) and deFinetti (1937), from the 1920s and 1930s. Savage (1954) gives what to many is a definitive account of this line of reasoning. The subjective interpretation of the mathematical theory treats probability as a (rational) agent's assessment of uncertainty about a truth-value bearing proposition, E. (Here, equate the proposition asserting E and the event E of the mathematical theory.) The agent's uncertainty about E is expressed as her/his degree of belief in E, and under appropriate assumptions may be measured by the agent's disposition to wager-on-E (*see* Bayesian Statistics).

The so-called 'Dutch Book' argument offers a prudential account of why the rational agent who is willing to engage in betting (under conditions made explicit, below) will have degrees of belief that conform to the formal probability calculus. Suppose that betting is in monetary units.

Definitions

A bet on E, at odds of $r{:}(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case E obtains, and with a loss of $-r(S)$ in case E fails to obtain.

A bet against E, at odds of $r{:}(1-r)$, with total stake $S > 0$ pays off to the agent with a loss of $-(1-r)S$ in case E obtains, and with a gain of $r(S)$ in case E fails to obtain.

The agent's betting odds of $r{:}(1-r)$ for E are fair just in case the agent is indifferent between betting on and betting against E at those odds, at least for moderate size stakes.

A (finite) set of bets constitutes a Dutch Book against the agent if the agent suffers a net loss from these bets regardless which state of affairs obtains.

The agent's fair betting odds are coherent if no Dutch Book is possible against them.

An elementary version of the 'Book' argument is captured in the following result:

The Dutch Book Theorem (deFinetti and Ramsey)

Suppose that the rational agent offers fair odds for each event E in the space $\mathscr{F}$ and for moderate size stakes is willing to accept finite combinations of bets made at her/his fair odds. Then the agent is coherent, that is, immune to having a Dutch Book made if and only if her/his fair odds satisfy the axioms of mathematical (finitely additive) probability (3.1)–(3.3).

Conditional probability, too, can be made the subject of the Dutch Book criterion by introducing called-off bets, e.g., a bet on E given F, which is annulled if F fails to occur.

Definition

A called-off bet on E given F, at odds of $r{:}(1-r)$, with total stake $S > 0$ pays off to the agent with a gain of $(1-r)S$ in case $E \cap F$ obtains, pays off with a loss of $-r(S)$ in case F obtains but E fails to obtain, and results in no change (0 payoff) in case F fails to obtain.

The concept of a called-off bet against E given F is likewise defined in parallel with the respective concepts for betting against an event, as is the condition of fair called-off odds. When called-off bets are included among those to be wagered on, the following improved result ensues.

Dutch Book Theorem for Conditional Probability

Assume that the agent offers fair odds and fair called-off odds, and for moderate size stakes is willing to accept finite combinations of wagers at these fair odds. Then the agent is coherent if and only if the fair odds satisfy the axioms of mathematical (finitely additive) probability (3.1)–(3.3), and the definition of conditional probability (4).

- *Remark*: The reader is alerted to the sad fact that when $\boldsymbol{P}(F) = 0$, condition (4) does not establish that $\boldsymbol{P}(\cdot|F)$ is a probability over $\mathscr{F}$, i.e., it may fail axioms (3.1)–(3.3). Hence, the Dutch Book argument for conditional probability does *not* establish that, given an event F of probability 0, the coherent agent's called-off fair odds given F constitute a probability.
- *Remark*: Some writers use a dynamic interpretation of the conditional probability, so that $\boldsymbol{P}(\cdot|F)$ designates the agents' subsequent degrees of belief upon learning that event F obtains. The Dutch Book argument for conditional probability does not underwrite this dynamic interpretation. It is part of the static constraints on a rational agent's degrees of belief at a fixed time. The called-off bets may, however, serve as a model of the agent's hypothetical reasoning at the fixed time.

Shimony (1955) introduces a more restrictive standard of coherence, which he calls strict coherence. An agent's fair odds are strictly coherent provided that there is no possibility of making a 'book' against her/him where the agent stands no chance of winning on net, but risks losing for at least one point of Ω.

Shimony establishes that an agent's fair odds are strictly coherent if and only if they are coherent and only the impossible event $\varnothing$ receives probability 0. Despite the desirability of having one's odds be strictly coherent, the resulting standard precludes continuous probability distributions, unfortunately.

Savage (1954) gives an elegant version of personal probability in which the agent's probability is reduced to her/his rational preference over acts. In the briefest outline, Savage's axioms include the assumption that the agent's preferences reveal a qualitative probability, as follows. Let b and w be prizes and assume that the agent prefers receiving the better prize b to the worse one w. For an event E, let the *wager-on-E* be the act that rewards the agent with prize b if E occurs and with w if E fails. Thus, these wagers are generalization of the bets used in deFinetti's Dutch Book argument. Savage postulates that the agent strictly prefers the wager-on-F to the wager-on-E if and only if the qualitative probability relation $E \langle \mathrm{F}$ obtains. Savage introduces additional postulates to insure that exactly one (finitely additive) quantitative probability agrees with this qualitative probability. That quantitative probability is offered as the agent's personal probability over events.

Logical (or Classical) Probability

Dating to Laplace's (1951) eighteenth-century work, at least, is the view that quantitative probability is determined once the space $\langle \Omega, \mathscr{F} \rangle$ is fixed. In its simplest version, LaPlace's *Principle of Insufficient Reason* states that the points of Ω determine equiprobable states. However, this principle leads quickly to a variety of inconsistencies, which are worth noting.

For one, LaPlace's principle makes the probability of an event depend upon which partition of states is used to describe it. A die's landing with the 1-spot up is made equiprobable with its not landing that way when Ω is just the binary partition; in which case, by LaPlace's principle, the probability is 1/2 for each outcome. However, if the state space is partitioned into six points depending upon how many spots land facing upwards, LaPlace's principle yields a probability of 1/6 for the outcome with 1-spot showing uppermost. We may continue to refine the state space, to include the information that the sum of the visible side faces on the die (as seen from a given perspective) is either less than, or is equal to, or is greater than the face showing uppermost. Then there are 14 distinct possibilities, but only one of these has a 1-spot uppermost. Then, by LaPlace's principle, the probability of a 1-spot uppermost is 1/14.

When the state space is continuous, the problem of applying LaPlace's principle grows more serious as a transformation to an equivalent random variable may lead to a different probability. For example, if Ω is the set of real numbers in the unit interval [0,1] and $\mathscr{F}$ is the space of Borel measurable subsets of the identity variable $X(\omega) = \omega$, then LaPlace's principle yields a uniform distribution over X. However, if we transform the space so that $\Omega' = [0, 1]$ with $\omega' = \omega^3$ and define the equivalent random variable $Y(\omega') = \omega'$, then Laplace's principle yields a uniform distribution when applied to the equivalent random variable Y. But a uniform probability distribution over X is evidently inconsistent with a uniform probability distribution over Y.

In an early-twentieth-century variant of LaPlace's idea, Keynes (1921) strove to revitalize the principle by applying it to

a formalized language in the fashion of Russell and Whitehead's *Principia Mathematica*. Moreover, Keynes's theory has the added subtlety that takes qualitative probability as primitive, and that relation is only partially defined: Not all propositions can be compared even by a qualitative probability relation in Keynes's system. Following Keynes, Carnap (1950) developed quantitative probability for a logical calculus. At first, Carnap sought to determine a single, well-defined quantitative probability for the monadic predicate calculus. But soon enough his approach, Carnap and Jeffrey (1971), became one best described as subjectivist, favoring symmetry conditions for atomic propositions similar to deFinetti's concept of exchangeability. Papers by Scott and Krauss (1966) and Gaifman and Snir (1982) continue this tradition.

A more productive alternative than appeal to a formalized language to revive LaPlace's principle is the use of mathematical symmetries as a basis for fixing the necessary equiprobable states. Most notable is Jeffreys's (1961) theory of *Invariance*, dating from the 1930s. Jeffreys's sophisticated Bayesian treatment appeals to transformational invariants of parametric statistical models of the observed data to determine a 'prior' probability for the parameter. For example, a uniform 'prior' probability is the invariant distribution for a location parameter, e.g., for the mean n of data from a normal $\mathbf{N}[\mu, 1]$ distribution. Fraser (1968) uses mathematical groups to determine the necessary invariances for fixing a probability, e.g., Haar measure. Each of these theories provides a *local*, not a *global*, interpretation of probability. That is, these theories provide necessary conditions for probability defined with respect to a specific inference problem. When multiple inference problems are considered at once, e.g., when the same theoretical quantity is a location parameter for one data set and a scale parameter for a second data set, difficulties ensue. These theories then generate what must be seen as either inconsistent or nonunique solutions. Furthermore, the invariant probabilities are not always countably additive. So called 'improper' priors, e.g., a uniform distribution over a real-valued parameter, correspond to a purely finitely additive probability by assigning equal probability to each unit interval for the parameter, which intervals constitute the elements of a countable partition. Dawid et al. (1973) provide an important litany of 'paradoxes' for these theories, involving both transformations of parameters and multiple, simultaneous inferences.

Yet another approach for reviving LaPlace's principle is found in the original work of E.T. Jaynes (1983), who uses entropy as a guide to uncertainty. His principle of *Maximum Entropy* dictates that, subject to a set of constraints (i.e., expectations for bounded random variables), a rational agent should fix probability over the space by choosing that probability that simultaneously satisfies the constraints while maximizing entropy. (For discrete spaces this requires choosing the probability $\boldsymbol{P} = \{p_i{:}i = 1, \ldots\}$ to maximize $-\sum_i \log(p_i)\ p_i$ subject to the constraints.) For example, when Ω is finite and the constraint set is vacuous, the uniform distribution over Ω maximizes entropy. Unfortunately, the Maximum Entropy principle inherits the problems of LaPlace's principle, as it is sensitive both to repartitioning the space Ω and, in the continuous case, equivalent transformations.

Objective Probability

Limiting and Finite Frequency Interpretations

Since Venn's (1876) writings in the nineteenth century, at least, it has been popular to attempt to link probability with limiting frequency in infinite sequences of repeated trials, thereby underwriting an objective interpretation. von Mises' (1931, 1939) work in the early twentieth century offered a mathematical program for this approach. For von Mises, probability applies to a collective, which is a suitably random sequence of repeated trials. For example, a (hypothetically infinite) sequence of repeated flips with a particular coin that can land either only heads-up or tails-up forms a collective and supports the fact that the objective probability of heads up is 1/2 provided that:

1. the limiting frequency in the sequence of the outcome heads up is 1/2.
2. the sequence of outcomes is random, i.e., this limiting frequency for heads up is invariant over selections of infinite subsequences.

Several important mathematical problems attend this approach. As is well known, the set of events that have a limiting frequency in a given sequence may fail to form a field (let alone a σ-field). Moreover, when Ω is infinite, even when limiting frequencies do exist they may fail to be σ-additive. However, the formal aspect of von Mises' program that has been given most scrutiny is the theory of randomness.

Evidently, for a sequence to pass muster as random in von Mises's sense, not all infinite subsequences are relevant. That is, assuming the limiting frequencies of the binary events {heads up, tails up} exist and are not either 0 or 1, trivially, there will be, e.g., the infinite subsequence consisting of outcomes with only heads up, whose limiting frequency for that event is 1, not 1/2. To sidestep this reduction of the randomness concept, von Mises proposed that only certain families of subsequences are relevant to objective randomness. These subsequences are determined by combinations of place selection rules, e.g., choose every kth trial, and tests for after-effects, e.g., consider the subsequence of outcomes preceded by two-consecutive 'heads up' outcomes. One intuition von Mises offered for these restrictions is that the random sequence should be shown to be immune to sure profits from any feasible gambling system that might be used against it. Wald (1937) showed that a formally consistent theory was possible if the tests for randomness were limited to a countable infinity of rules. Church (1940) suggested using the denumerable class of recursive functions, in an attempt to make the concept of a random sequence language invariant. However, Ville (1939) showed that collectives based on denumerably many selection rules fail to validate some of the basic 'strong law' theorems of Kolmogorov's theory, e.g., the law of the iterated logarithm fails. Also, by appeal to martingale theory, Ville showed that such collectives fail von Mises's intuitive criterion for randomness: they do not preclude successful gambling systems.

Reichenbach (1938, 1949) proposed a rival frequency theory that avoided von Mises's randomness conditions, even permitting its application to finite sequences. In doing so, however, the resulting theory depends ever more heavily on epistemic considerations for its warranted application. For example, a (sequence) that alternates successively between

heads up on odd numbered flips and tails up on even numbered flips has a limiting frequency of 1/2 for each outcome. But, given that fact about the order of outcomes, we would be hard pressed to find a good use for this 'objective' probability when making inferences about such a process.

A more interesting finite-frequency interpretation is based on Kolmogorov's (1963, 1968, 1984) idea for interpreting randomness as informational complexity in finite sequences. A proper account of this approach is beyond the scope of this article. For first-rate discussions, see Fine (1973) and van Lambalgen (1987).

Chance and Propensity Interpretation

The philosophers Braithwaite (1953), Hacking (1965) and Levi (1980) propose and defend an objective interpretation of probability based on the idea that an objective chance is a theoretical, quantitative disposition for a process or kind of trial to respond when activated. That is, consider a process or kind of trial described as flipping this coin by hand. The chance of its landing heads up on such a process is taken to be an unobserved, quantitative disposition for the coin to respond when flipped. The propensity interpretation is thin, on purpose, in that chances satisfy the mathematical theory of probability by stipulation!

What is particularly novel in these approaches about making chance a theoretical concept is that probability is *not* reduced to observed frequencies, contrary to the frequency interpretation. Rather, the meaning of the theoretical term *chance* is given by rules for testing chance claims (Braithwaite) or by rules linking chance with subjective probability (Hacking and Levi). In the latter case, central to the meaning of chance is a so-called *Direct Inference* principle. Put simply, that principle asserts: the personal probability $\boldsymbol{P}$ for a specific outcome O of the kind of the trail occurring is p, given the information that the chance is p for outcome O on that kind of trial. To wit: $\boldsymbol{P}$ (coin lands heads up on a toss | chance is p that the coin lands heads up on a toss) $= p$.

This treatment of chance as a theoretical term, with it meaning supplied by rules for application of the concept, accords with the Pragmatic turn in philosophy as proposed by Peirce (1932) in the nineteenth century, though Peirce himself favored a frequency interpretation.

The Fisher–Kyburg Interpretation

Fisher (1956) and Kyburg (1961) introduce an original semantics for probability, focused on the single-case assertion, e.g., (*). The probability is 1/2 that the next flip of this coin lands heads up.

For this theory, probability is fundamentally epistemological, grounded on both knowledge and ignorance of class-frequencies. The statistician who asserts (*) is required to have:

1. *knowledge* of frequencies of outcomes of heads in the population of flips with this coin, a population that includes the next flip.
2. *ignorance* of the statistics of outcomes in a smaller statistical population of coin flips that includes the next flip.

In this sense, their theory takes instances of direct inference (based on frequency information about populations) as the definition of probability.

For both Fisher and Kyburg, this interpretation of probability underwrites an important statistical analysis, e.g., Fisher's fiducial probability. For example, let the random variable X take values from a population with known frequencies approximated by a normal distribution, $\mathbf{N}(\mu, 1)$, but with parameter μ unknown. Then the pivotal quantity $V = (\mu - X)$ has a known distribution in this population. Its values are distributed according to a normal $\mathbf{N}(0, 1)$ distribution. Suppose $X = 7$ is observed on a particular trial. The Fisher–Kyburg theory then asserts that on this same trial, given $X = 7$, the statistician's probability distribution for V is $\mathbf{N}(0, 1)$. This claim rests on the assertion of ignorance on the part of the statistician: the statistician knows of no smaller population that includes this trial with statistics for V different from those of the $\mathbf{N}(0, 1)$ population. The upshot is a fiducial $\mathbf{N}(7, 1)$ probability distribution for μ, given $X = 7$. When a random variable X has a continuous, parametric distribution $F(X, \theta)$, then F itself may be treated as the pivotal quantity, having a known uniform U[0, 1] distribution. Fiducial inference about F, given the observation that $X = x$, yields probability for the unknown parameter satisfying Fisher's enigmatic formula: $P(\theta|X) = -\partial F/\partial\theta|_{X=x}$. Kyburg's theory has the added generality over Fisher's in that he uses interval-valued frequency information, leading to a theory with lower and upper probability for single-case assertions. Despite the considerable novelty of this approach, it is challenged severely by nuisance parameters. That is, this interpretation of probability provides only a limited account of marginal probability, as its theory of conditional probability is not well defined, at least not as required by Bayes's theorem.

Additional Readings

Nagel's et al. (1936) and Nagel (1939) essays on quantitative probability are excellent reviews, still, of philosophical issues relating to rival interpretations. In particular, one important issue not addressed here is whether objective probability requires a commitment to indeterminism in science. Nagel makes clear why there is no conflict between objective probability and deterministic laws in science. A more elementary presentation of rival interpretations of quantitative probability is found in Kyburg's (1970) textbook.

Fine's (1973) book gives a good overview of complexity theories and randomness, relating in particular to Kolmogorov's theory from the 1960s.

Fishburn's (1986) article offers a wealth of insights on both philosophical and formal questions relating to qualitative and quantitative concepts of probability. The bibliography there is particularly noteworthy. Also valuable is the discussion of comparative probability given in Walley and Fine (1979).

See also: Probabilistic Thinking; Probability Density Estimation; Probability Theory: Formal; Probability and Chance: Philosophical Aspects.

Bibliography

Braithwaite, R.B., 1953. Scientific Explanation. Cambridge University Press, Cambridge, UK.

Carnap, R., 1950. Logical Foundations of Probability. University of Chicago, Chicago.
Carnap, R., Jeffrey, R. (Eds.), 1971. Studies in Inductive Logic and Probability, vol. 1. University of California Press, Berkeley, CA.
Church, A., 1940. On the concept of a random sequence. Bulletin of the American Mathematical Society 46, 130–135.
Dawid, P., Stone, M., Zidek, J.V., 1973. Marginalization paradoxes in Bayesian and structural inference. Journal of the Royal Statistical Society 35, 189–233 (with discussion).
deFinetti, B., 1931. Sul significato suggettivo della probabilita. Fund Math 17, 298–329.
deFinetti, B., 1937. Foresight: Its Logical Laws, Its Subjective Sources (Kyburg, S. Eds. and Trans. (1964)). Robert E. Kreiger Publishing Co., Huntington, NY.
Fine, T., 1973. Theories of Probability. Academic Press, New York.
Fishburn, P.C., 1986. The axioms of subjective probability. Statistical Science 1, 335–358 (with discussion).
Fisher, R., 1956. Statistical Methods and Scientific Inference. Hafner Press, New York.
Fraser, D.A.S., 1968. The Structure of Inference. John Wiley, New York.
Gaifman, H., Snir, M., 1982. Probabilities over rich languages, testing and randomness. Journal of Symbolic Logic 47, 495–548.
Hacking, I., 1965. Logic of Statistical Inference. Cambridge University Press, Cambridge, UK.
Horn, A., Tarski, A., 1948. Measures in Boolean algebras. Transactions of the American Mathematical Society 64, 467–497.
Jaynes, E.T., 1983. In: Rosenkrantz, R. (Ed.), Papers on Probability, Statistics, and Statistical Physics. D. Reidel Publishing, Dordrecht, The Netherlands.
Jeffreys, H., 1961. Theory of Probability, third ed. Oxford University Press, London.
Keynes, J.M., 1921. Treatise on Probability. Macmillan, London.
Kolmogorov, A.N., 1956. Foundations of Probability. Chelsea Publishing, New York.
Kolmogorov, A.N., 1963. On tables of random numbers. Sankhya 25, 369–376.
Kolmogorov, A.N., 1968. Three approaches to the definition of the concept of 'amount of information'. Selected Translational Mathematical Statistics and Probability 7.
Kolmogorov, A.N., 1984. The logical basis for information theory and probability theory. IEEE Transactions on Information Theory 14, 662–664.
Kraft, C., Pratt, J., Seidenberg, A., 1959. Intuitive probability on finite sets. Annals of Mathematical Statistics 30, 408–419.
Kyburg, H.E., 1961. Probability and the Logic of Rational Belief. Wesleyan University Press, Middleton, CT.
Kyburg, H.E., 1970. Probability and Inductive Logic. Collier-Macmillan, London.
Kyburg, H.E., Smokler, H.E. (Eds.), 1964. Studies in Subjective Probability. John Wiley, New York.
Laplace, P.S., 1951. A Philosophical Essay on Probabilities. Dover, New York.
Levi, I., 1980. The Enterprise of Knowledge. MIT Press, Cambridge, MA.
van Lambalgen, M., 1987. Random Sequences. University of Amsterdam, Amsterdam.
von Mises, R., 1931. Wahrscheinlichkeitsrechnung und ihre An-wendungen in der Statistik und theoretischer Physik. Deuticke, Leipzig.
von Mises, R., 1939. Probability, Statistics, and Truth. Macmillan, New York (Translation of Wahrscheinlichkeit, Statistik, und Wahrheit. Wien (1928).).
Nagel, E., 1939. Principles of the Theory of Probability. University of Chicago Press, Chicago.
Nagel, E., Margenauf, L., Ducasse, C.J., Wilks, S.S., 1936. The meaning of probability. Journal of the American Statistical Association 31 (193), 10–30.
Peirce, C.S., 1932, 1936. Collected Papers, Vols. 2 and 6. Harvard University Press, Cambridge, MA.
Ramsey, F.P., 1926. Truth and probability. In: Braithwaite, R.B. (Ed.), The Foundations of Mathematics 1950. Humanities Press, New York, pp. 63–92.
Reichenbach, H., 1938. Experience and Prediction. University of Chicago Press, Chicago.
Reichenbach, H., 1949. Theory of Probability. University of California Press, Berkeley, CA.
Savage, L.J., 1954. The Foundations of Statistics. John Wiley, New York.
Scott, D., Krauss, P., 1966. Assigning probabilities to logical formulas. In: Hintikka, J., Suppes, P. (Eds.), Aspects of Inductive Logic. North-Holland, Amsterdam.
Shimony, A., 1955. Coherence and the axioms of confirmation. Journal of Symbolic Logic 20, 644–660.
Venn, J., 1876. The Logic of Chance, second ed. Macmillan, London.
Ville, J., 1939. Etude critique de la notion de collectif. Gauthiers-Villars, Paris.
Wald, A., 1937. Die Widerspruchsfreiheit des Kollektivsbegriffs der Wahrscheinlichkeitsrechnung. Ergebnisseeins Mathematisches Kolloquiums 8, 38–72.
Walley, P., Fine, T., 1979. Varieties of modal (classificatory) and comparative probability. Synthese 41, 321–374.

Probation, Offenders and Risk

Peter Raynor, Swansea University, Swansea, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the history and development of probation services and their gradual evolution from a focus on the welfare of offenders to a focus on risk, public safety, and the reduction of reoffending. Probation in England and Wales is presented as a particularly clear example of how probation has had to adapt to both new research findings and political fashions. The conflict between these different influences is explored, and possible scenarios for future development are discussed.

Introduction: Origins and Principles

This article describes the origins of probation and distinctive features that have traditionally characterized its theory and practice, the threats and challenges posed by evaluation research, the revival of rehabilitation in the 1990s as an evidence-based approach to managing and reducing the risk of offending, and contemporary developments that point to new challenges and possible threats. At each stage in this story, probation has been shaped by its social and political environment as well as its own professional traditions, and some environments have been more favorable than others.

The origins of probation are complex. Recent historical research points to several sources. In the United Kingdom, the Police Court Mission of the Church of England Temperance Society traced its origin to a donation by the shoemaker Frederick Rainer in 1876. Other obvious precursors can be found in the mid-nineteenth century voluntary work of John Augustus in Boston, Massachusetts (US), who befriended and helped offenders placed under his supervision by the Courts. At the same time, the innovations of Matthew Davenport Hill, Recorder in Birmingham, England, released selected offenders on recognizance when suitable volunteer supervisors could be found in the community. The first legal enactment of a recognizable form of probation was in Massachusetts in the United States in 1873. England and Wales persisted with a fully voluntary system until early in the next century, when a Probation of Offenders Act was passed in 1907. (England and Wales will be mentioned together in this article, because for criminal justice purposes, these two countries within Britain form a single jurisdiction. Scotland has a separate legal system and its own arrangements.) Vanstone (2008) dates the major period of international expansion as lasting from 1878 to 1920 and lists 43 jurisdictions to which probation spread between those dates. There was variation: the administering agency could be judicial or could be part of central or local government; punishment could be conditionally withheld or imposed and then suspended subject to compliance with supervision; sometimes probation would be an alternative to prosecution rather than a sentencing option, and supervision was not always a requirement. However, the 'Anglo-American model' (Vanstone, 2008) was widely adopted, with probation used instead of punishment, supervision by an officer of the Court, and the option of returning to Court for reconsideration of sentence if the requirements of probation were breached. These requirements normally included reporting to a supervising officer when required, receiving visits at home, 'leading an industrious life,' and avoiding further offenses. In the earlier decades of the twentieth century, probationers could be required to abstain from alcohol.

The roots of probation can be found in a mixture of religious missionary commitment and nineteenth century charitable endeavor, showing concern for the welfare of poor people and a desire to improve their morals and behavior, particularly where drink was concerned. Fuller accounts of the history are available in Vanstone (2004), Raynor and Robinson (2009), and, for recent periods, Canton (2011). By the early twentieth century, most industrial countries were responding to demands that the state should play a greater role in safeguarding the welfare of citizens, and activities such as social work and probation were drawn into this process. For example, the 1907 British Probation of Offenders Act was part of a period of Government-led social reform that laid the foundations for the later welfare state (Vanstone, 2004). Consequently, during the first half of the twentieth century, probation in many countries moved from the periphery of charitable and religious endeavor to become a small but important part of the system of state-funded welfare services and an integral part of the criminal justice system. By the early 1950s, the criminologist Max Grünhut, a refugee from Nazi Germany who established the study of criminology at Oxford University, reflected a professional consensus when he described the essential features of probation as "a combination of two things, conditional suspension of punishment, and personal care and supervision by a court welfare officer" (Grünhut, 1952: p. 168).

By this time, professional thinking and literature were emphasizing the process of individual supervision and importance of the supervisory relationship. Offenders under supervision would report to probation offices, often weekly, for regular interviews with their probation officer. Originally influenced by what would now be considered a strange mixture of ideas about the psychology of offending, including Lombrosian theories and an early interest in eugenics (Vanstone, 2004), probation officers and people who ran their training courses increasingly adopted many ideas from psychodynamic casework, which had been enthusiastically embraced by the social work profession. Many probation officers trained on social work courses and saw themselves as

social workers, albeit with a special role and expertise in the criminal justice system and the Courts.

Two other early and durable developments in probation work are worth noting at this stage. The first concerns the provision of information and advice about individual defendants to sentencing judges or magistrates before sentence is passed. This was a development from the practice of early voluntary workers such as John Augustus and Police Court Missionaries who, if they had interviewed a defendant who they thought could be saved from imprisonment, would often directly address the Court to plead for release under supervision. Over the years, this became formalized as a written document, known as a Social Enquiry (or Inquiry) Report, and more recently a Pre-Sentence Report. Such reports have become a major part of the work of most probation services, and much has been written about their role and contents: for example, early controversies concerned whether reports should make an explicit recommendation as to sentence or simply provide a history and leave the sentencing decision to the Court; whether they should contain a full social history, as a social casework file might, or whether they should concentrate only on matters relevant to the current offense; whether they should be understood as a document supporting the defendant or as neutral professional advice; and whether they have any significant influence on sentencing decisions. Evidence from England and Wales suggests that they do, but only if carefully prepared and well written (Gelsthorpe and Raynor, 1995). Other issues raised from the 1980s onward concern the contribution of stereotyped language and thinking in presentence reports to sentencing outcomes that discriminate against women (see Dominelli, 1984; Eaton, 1986; Gelsthorpe, 1992; Hedderman and Gelsthorpe, 1997) and minority ethnic groups (Whitehouse, 1983; Moxon, 1988; Dominelli et al., 1995; Hudson and Bramhall, 2005), and how such outcomes might be avoided. Originally little more than a plea for mercy and an offer to supervise, presentence reports have become sophisticated documents informed by detailed risk assessments, often following a standard format, sometimes provided as a template on probation service computers.

The second early accretion to the probation officer's role in most countries was responsibility for the supervision of some offenders released from custodial sentences. Known in Britain as aftercare, and later as through-care or resettlement (or, in the United States, reintegration or reentry), this also has roots in nineteenth century charitable work. Voluntary organizations (for example, Discharged Prisoners' Aid Societies) formed to provide support and assistance on release to prisoners deemed as having paid their debt to society, or simply as unfortunate, and in need of help to reestablish themselves in the community. Religious groups organized Prison Gate Missions, and as in the Police Court Missions, encouraged abstention from alcohol. During the twentieth century, many custodial sentences included periods of supervision after release, with the possibility of recall to prison for failure to comply with license requirements. Increasingly, the supervision of offenders subject to such requirements came to be undertaken by probation service staff rather than by voluntary organizations. By the end of the 1960s, in England and Wales, a substantial part of the Probation Service's workload was made up of aftercare cases, mainly young offenders but including some older offenders on parole after early release from long sentences, and several thousand 'voluntary aftercare' cases where there was no requirement to supervise but a voluntary agreement to provide and receive advice and support. These voluntary cases had almost completely disappeared from the probation workload in England and Wales by the end of the twentieth century, but were an important part of the job in the 1960s and 1970s.

The statutory aftercare caseload, consisting of enforceable licenses, grew as prison populations rose and more categories of prisoners became subject to various forms of conditional supervised release. These developments emphasized the controlling side of the probation officer's work: released prisoners were supervised on behalf of the penal system rather than the Courts, and this helped to both integrate probation more closely into the structure of the criminal justice system and focus its work increasingly on adult offenders.

The Response to 'Nothing Works'

The era of comfortable identification with the social work profession, and confidence that good results were being achieved, came to an end in the mid-1970s, when a series of studies of the outcomes of rehabilitative work with offenders indicated that no sentence, including probation, was significantly affecting reoffending rates. The best known and most vivid expression of these conclusions was an American article by Robert Martinson that argued that there was "very little reason to hope that we have in fact found a sure way of reducing recidivism through rehabilitation" (Martinson, 1974: p. 49). This was summarized and simplified by the press as 'nothing works,' and similar studies were conducted in other countries and in relation to other aspects of social work. These studies often had sophisticated comparative designs, including random allocation to treatment and comparison groups, but their careful measurement of outcomes was not matched by a similar clarity over inputs. It was correctly argued, for example, by Fischer (1976), that such studies often failed to distinguish between good and bad practices, but simply documented the aggregate absence of effect that came from mixing the two together. However, research results that questioned the effectiveness of spending on welfare were popular with some politicians at a time when increasing welfare budgets in Western countries were coming under pressure, and in some countries 'nothing works' became an orthodoxy that went largely unchallenged for a decade (McGuire, 2002).

Probation's response to this kind of challenge can best be illustrated by examining what happened in England and Wales, which then had one of the oldest and most sophisticated Probation Services and could be regarded as a paradigm case of how probation reshapes itself in response to changing expectations and wider policies. Briefly, the apparent discrediting of the 'treatment model' led to new thinking about what probation could offer as a sentencing option within the criminal justice system. It was demonstrably much cheaper than prison, and if nothing works, cheaper is better. Probation's new goal became not so much to change offenders as to change the decisions made about them, by offering supervised alternatives to custodial sentences under various guises: 'diversion,'

'intermediate sanctions,' and various 'intensive probation' schemes designed to appear demanding enough to be a realistic sentence for offenders who would otherwise have gone to prison. Additionally, new community-based sentencing options appeared, including Community Service (later known as Community Punishment, then Unpaid Work, then Community Payback). This change of direction toward providing alternatives to custody was confirmed by the Home Office (which then directed criminal justice policy in England and Wales) in a 'Statement of National Objectives and Priorities' in 1984, and for nearly a decade, this strategy was relatively successful (Raynor and Robinson, 2009). Furthermore, this was paradoxically a creative period for practitioners interested in developing more effective ways of supervising offenders. Intensive forms of probation meant more time under supervision, which needed to be filled with meaningful content. Probation officers developed day centers, and many different forms of group work, sometimes involving structured programs and sometimes anticipating later developments by teaching social skills, self-management, and problem solving.

More fundamentally, 'alternatives to custody' could not compete seriously with prison unless they could be shown not to expose the public to significantly greater risks of victimization by offenders. Thus the ideas of public protection and risk management, which had previously been minor themes in probation practice, became more salient. At this early stage, little attempt was made to distinguish between the risk of reconviction and the risk of serious harm to victims; this distinction was to become increasingly important during the next few years. The probability of reconviction for people on probation in England and Wales is normally around 37% within a year of the start of the probation order, but the risk of serious harm is much lower and, being a relatively rare event, much harder to predict. Probation staff during the era of diversion had been used to estimate another kind of risk, namely, the risk of receiving a custodial sentence, and had been used to offer more intensive supervision when custody seemed likely, so the principle of offer more intervention when risk was higher was already known before the focus shifted from the offender's risk of a custodial sentence to the potential victim's risk of suffering harm. Once that shift was under way, the purpose of probation was redefined again so that the beneficiaries were understood to be no longer the probationers or the Courts but the potential future victims of their offenses (Garland, 1997). The emergence of a strong victims' support movement influenced perceptions of the aims of sentencing (Rock, 1991). It is also important to note that despite the increasing amount of risk talk in probation, no reliable or consistent method of assessing risk was made generally available to probation officers until around the end of the century.

Meanwhile, researchers in Canada and the United States, where many psychologists and researchers had never been convinced that 'nothing works,' were beginning to produce results that became very important for the next stage in probation's evolution. In particular, the new statistical technique of meta-analysis was being used to combine the findings of large numbers of studies to increase the reliability of conclusions, and meta-analytic studies of offender rehabilitation were clearly beginning to refute the dogma that 'nothing works' (Lösel, 2012).

The Second Challenge: 'Prison Works' and the Case of England and Wales

The early 1990s began what was to prove an era of turbulence for probation in England and Wales. Many similar developments can be found, for similar reasons, in other countries (see Kalmthout and Durnescu, 2008) but they appeared particularly strongly in England and Wales due to increased centralization of the management of probation and increasing political involvement in its development. The Probation Service entered the decade as a decentralized, locally controlled service guided by committees of local magistrates, equipped with a policy of diverting offenders from custodial sentences, the beginning of some new evidence-based ideas about reducing reoffending, and a traditional mission, set out in official documents, to 'advise, assist, and befriend' the people it supervised. By 2001, it was a national organization, controlled by central government, with an official mission that included 'the proper punishment of offenders' and a standard strapline on its publications, which read 'Enforcement, rehabilitation and public protection.' Although some commentators also point to continuities in the concerns and professional culture of the Probation Service (see Raynor, 2012), the change was quite dramatic and worth exploring in some detail.

In 1993, an unpopular Conservative Government replaced the minister responsible for prisons and probation, who had supported the policy of diversion, with a more right-wing politician, Michael Howard. He shortly declared in a major speech that "prison works," and set off an increase in the prison population that has continued almost without a break until the present. Probation needed a new brand and a new purpose in a suddenly hostile political climate, in which the training of probation officers was deliberately separated from that of social workers. Probation's leaders turned to the emerging research on 'what works' to support a new mission of reducing reoffending and managing risk. A survey by the Probation Inspectorate, carried out in the mid-1990s but not published until a new Government was in power (Underdown, 1998), found that much existing practice failed to incorporate the principles of effectiveness, which new research supported. A key influence at this time was the work of Canadian researchers, often summarized as the risk–need–responsivity model (RNR), which, with strong supporting evidence, argues that supervision should concentrate on higher risk offenders where it will make more difference; that it should address those features of offenders' circumstances and behavior that increase the risk of reoffending (known as 'criminogenic needs'); and that it should use methods appropriate to the learning styles of the people under supervision (with 'general' responsivity based on the common characteristics of offenders, and 'specific' responsivity based on distinctive needs of particular groups or individuals) (see Andrews and Bonta, 2010).

These developments led to a considerable shift in how rehabilitation was understood, and in the specific practices associated with it (McGuire, 2002; Raynor and Robinson, 2009). The new research pointed to methods based on

learning theory and particularly to cognitive behavioral approaches, often structured into 'programs' that followed a prescribed curriculum laid out in a manual. Moreover, the new focus on risk (both of reoffending and harm to potential victims) became too important to be left to the individual subjective judgment of practitioners. As risk management and risk reduction became central justifications for having a Probation Service, research was carried out on various structured approaches to risk assessment, and by the end of the decade, the Home Office had begun developing its own Offender Assessment System (OASys) for use throughout the prison and probation systems (now grouped together as 'correctional services').

Unfortunately, OASys proved to be so complex and burdensome that its potential benefits were never fully realized, and much of the information that should have been gathered for research was incomplete (Debidin, 2009). The fact that OASys took up so much time, so that probation officers were spending more time with their computers than with the people they were supposed to supervise, contributed to a generally negative response to risk assessment. For example, some people felt that it emphasized negative aspects of offenders' lives. Some evidence suggested that risks could be overpredicted, especially for women (Raynor, 2007) and that this contributed to inappropriate risk-driven styles of supervision (for some examples, see Carlen, 2002; for some alternative approaches, see Sheehan et al., 2007, 2011). In addition, some commentators associated the rise of risk assessment with changes in probation culture and more punitive attitudes toward offenders. The reality seems to be that these things can happen, particularly where unpopular innovations are introduced in an already punitive political climate, but some other studies also show that risk assessment methods that identify needs and targets for assistance can substantially improve effectiveness and support rehabilitation, when applied in a more favorable context (Raynor and Miles, 2007).

The focus on risk also led to the development of new forms of intensive surveillance and supervision for offenders assessed as dangerous or at high risk of doing serious harm. Collaboration between all relevant agencies was developed and organized through 'Multi-Agency Public Protection Arrangements' in which regular meetings of senior personnel organized supervision and monitored the behavior of offenders who caused particular concern. This relatively successful approach led, for example, to effective shared supervision by probation staff and police officers, which would have been unlikely in previous decades, and similar approaches to prolific property offenders who account for a disproportionate share of all crime. One major problem, however, is that the prediction of dangerous behavior is never as accurate as the prediction (itself only approximate) of general reoffending. There is always some risk left unmanaged and some serious offences committed by people under supervision, leading to criticism of all organizations involved in public protection (see, for example, Kemshall, 2008).

While some developments of the 1990s, such as the expansion of electronic tagging provided by private companies under contract to the Government, were primarily about surveillance and low-cost control outside prisons, a good deal of effort and resources also went into delivering cognitive behavioral group programs to probationers with a medium to high risk of reoffending. By the time a new Government was elected in 1997, and the Probation Inspectorate's Report on 'Strategies for Effective Offender Supervision' was published (Underdown, 1998), enough was known to allow a bid for significant Government funding for 'pathfinder' programs under a national Crime Reduction Program, which ran from 1999 to 2003. The aim was to roll out programs to all probation areas and deliver them effectively. The project of developing and evaluating such a large operation in only 3 years proved overambitious, and targets were missed in spite of the recruitment of unsuitable offenders to programs in an attempt to meet them. Not surprisingly, the outcomes revealed by research were not as promising as had been hoped. The general research verdict on the 'pathfinder' programs was that promising initiatives had been frustrated by poor and rushed implementation (Raynor, 2004).

While these substantial changes and developments were being generated within the professional world of probation, other rather different influences were brought to bear from outside, especially from the political realm. Politicians of all major parties in Britain were very conscious (perhaps over-conscious) of public concern about crime, which had risen rapidly during a period of mass unemployment in the 1980s. The politicians' response was to seek opportunities to present themselves as 'tough on crime.' Michael Howard was an early example, but many followed his lead, including the 'New Labour' Prime Minister Tony Blair. Now that the Probation Service was under the control of London-based politicians, they thought it simply did not look tough enough. Probation officers had always been required to take enforcement action if people under supervision did not comply with their probation order or aftercare license; for probation and other community sentences, enforcement took the form of prosecution for breach of probation and return to Court for possible resentencing. Some official reports suggested that this was not happening consistently, or enough, and politicians acted on this. A series of 'National Standards' documents starting in 1992 under the Conservative Government made enforcement requirements stricter and progressively removed the discretion of supervising officers over questions of when to take enforcement action. Breach rates increased and helped to swell the prison population. Furthermore, politicians thought that probation should be made to look more punitive: hence the insertion of 'punishment' into mission statements, the use of more menial physical tasks in Community Service, the visible identification of offenders doing Community Service by making them wear bright orange jackets, and the change of name to Community Punishment. Ambitious managers whose future careers depended on nimble self-alignment to the political climate helped to accelerate these processes, and the official branding of probation became more explicitly punitive.

So did the Probation Service's efforts to endure its own survival 'work'? By 2010, it was costing about £1 billion a year, serving a population of about 56 million people in England and Wales, with over 15 000 employees; it was supervising nearly a quarter of a million offenders on any given day, and preparing about 200 000 reports for the Courts in a year. Recent

research on accredited programs had begun to show much better results: when actual reconviction rates are compared to those statistically predicted for particular groups of offenders, the differences can be used to estimate the effect that a particular sentence or intervention is having, and by 2007, probationers required to undergo programs were reconvicting 10.3% less than the expected rate. For those who actually completed the programs (completion rates had risen substantially since the 'pathfinder' phase), the difference was 25.8% of the predicted rate (see Raynor and Robinson, 2009). However, other measures show a less encouraging picture. The use of both imprisonment and community sentences has increased in England and Wales since the early 1990s, while the use of fines has decreased. Therefore, if there is still much diversion going on, it is more likely to be diversion from fines than from prison.

Where Are We Now? Policy Choices and the Future of Probation

As comparative studies of probation show, England and Wales are not alone in these difficulties. In most countries, probation systems are looking for the right balance between a helping role and a law enforcement role, and locate themselves at different points along this continuum. England and Wales have an unusually strong emphasis on punishment, but the main policy dilemmas are similar in many countries (Kalmthout and Durnescu, 2008). Other developments are also widespread: for example, new technologies such as electronic tagging, including tracking tags, and even experimentation with 'kiosk' probation, which allows offenders to report to a machine rather than a person (*The Observer*, 29 April 2012). What these innovations lack is the element of personal guidance and help. Other developments concern the management and governance of probation services. For example, in England and Wales, the term 'offender manager' is now preferred to probation officer, and the Government is developing plans to allow the private and voluntary sectors to take over 70% of the Probation Service's current work (Ministry of Justice, 2013). In theory, these new providers will be paid according to how far they achieve reductions in reoffending. Many details of the policy remain to be worked out, but it is likely to represent the first major shrinkage in the scope and remit of the Probation Service in over a century. Probation services have always worked with the voluntary sector, and the private sector has built and operated a small number of prisons in Britain, but the wholesale privatization of the majority of probation work is a radical new departure that has led to concern about the legitimacy and accountability of new private sector providers (see the special issue of the *Probation Journal*, December 2012; Burke, 2012).

Some other developments, particularly in probation research, are more encouraging for people working in the field. The work of desistance theorists is proving influential (see McNeill et al., 2010): they study the process of desisting from crime, and promote approaches to supervision that support and facilitate desistance. Other ideas related to this perspective include an emphasis on probationers' own strengths and positive goals, to balance what some commentators see as an overemphasis on needs and deficits in the RNR model. While evidence to support these new approaches is limited, several studies have shown that skilled and efficient individual supervision makes a significant contribution to the reduction of reoffending below expected levels (see McNeill et al., 2010). This both helps to challenge overreliance on group programs and underlines the importance of appropriate individual contact and support for those who are engaged in programs.

Other developments have focused on the needs and experiences of particular groups of offenders in a diverse population. Research on 'what works' was originally based largely on majority white male populations in North American or European penal systems. It needs to be informed by further research and service users' views when applied to disadvantaged minority ethnic groups who are likely to have experienced discrimination and prejudice (Lewis et al., 2006; Bhui, 2009), or women, who constitute roughly half the population but a much lower proportion of offenders (Worrall, 1990; McIvor, 2004; Sheehan et al., 2007). There is evidence (reviewed by Hedderman, 2004) that risk factors and 'criminogenic needs' of women offenders overlap to some extent with those of male offenders, but with important differences. Patterns of offending are also different, typically involving a high proportion of minor property crimes such as shoplifting and resulting in lower rates of recidivism than those of male offenders. Recent studies of 'what works' with women offenders (for example, Sheehan et al., 2007, 2011) have emphasized the need for support and motivation, and for practical assistance with a wider range of personal and social difficulties than is typically found among male offenders. Such research supports approaches to supervision that are informed by an understanding of individual differences and socially imposed disadvantage. Overall, our level of knowledge about offender rehabilitation and the quality of the research base are both still improving, and both are arguably higher than at any time in the past. For example, the research network CREDOS, the Collaboration of Researchers for the Effective Development of Offender Supervision, set up in 2007, has active researcher members in 11 countries (see McNeill et al., 2010, which includes papers from the CREDOS conferences). The limitations and frustrations in the development of offender rehabilitation are often due to poor implementation of good ideas, or to policy and practice contexts distorted by external political pressures.

Conclusion

The future shape of probation seems likely to be determined not by developments within the field of probation itself, however promising, but by political choices reflecting the divergent political cultures of different jurisdictions. Countries with neoliberal governments and a neoliberal political culture tend to 'responsibilize' individuals and hold them accountable and culpable for most of the difficulties and problems they encounter in their lives: social causation is underplayed, and individual choice is all. Criminal justice systems in neoliberal

political cultures tend to emphasize punishment rather than rehabilitation or reintegration, and often have high rates of imprisonment (Cavadino and Dignan, 2006). Where more social democratic political cultures exist and social development and welfare are seen as a partnership between state support and individual initiative, assumptions about collective responsibility for managing social problems will often sustain a more rehabilitative and inclusive approach to criminal justice (Downes and Hansen, 2006). These might also be good environments for developments based on restorative justice, which is already the preferred option for young offenders in some jurisdictions, for example, Northern Ireland (see Jacobson and Gibbs, 2009). The choice between the two scenarios outlined in this paragraph ultimately rests with politicians, those who elect them, and other opinion formers within society. Probation staff and other professionals can prepare themselves and their organizations for future challenges by continuing research into what helps offenders to desist from offending, and communicating the results.

See also: Community Supervision and Diversion in the United States; Criminal Justice and Social Work Practice; Criminal Justice, Sociology of; Deterrence Theory: Crime; Imprisonment: Sociological Aspects; Prisons and Imprisonment; Punishment; Recidivism; Rehabilitation of the offender; Women and Criminal Justice.

Bibliography

Andrews, D.A., Bonta, J., 2010. The Psychology of Criminal Conduct. Anderson Publishing, New Providence, NJ.
Bhui, H.S. (Ed.), 2009. Race and Criminal Justice. Sage, London.
Burke, L. (Ed.), 2012. Special Issue: Opening the Criminal Justice Market. Probation Journal 59 (4), 317–417.
Canton, R., 2011. Probation – Working with Offenders. Willan, Abingdon.
Carlen, P. (Ed.), 2002. Women and Punishment. Willan, Cullompton.
Cavadino, M., Dignan, J., 2006. Penal Systems: A Comparative Approach. Sage, London.
Debidin, M., 2009. A Compendium of Research and Analysis on the Offender Assessment System (OASys) 2006–2009. Research Series 16/09. Ministry of Justice, London.
Dominelli, L., 1984. Differential justice: domestic labour, community service and female offenders. Probation Journal 31 (3), 100–103.
Dominelli, L., Jeffers, L., Jones, G., Sibanda, S., Williams, B., 1995. Anti-racist Probation Practice. Arena, Aldershot.
Downes, D., Hansen, K., 2006. Welfare and Punishment: The Relationship between Welfare Spending and Imprisonment. Briefing 2. Crime and Society Foundation, London.
Eaton, M., 1986. Justice for Women? Family, Court and Social Control. Open University Press, Milton Keynes.
Fischer, J., 1976. The Effectiveness of Social Casework. C. C. Thomas, Springfield.
Garland, D., 1997. Probation and the reconfiguration of crime control. In: Burnett, R. (Ed.), The Probation Service: Responding to Change. Proceedings of the Probation Studies Unit First Colloquium: Probation Studies Unit Report No. 3. University of Oxford Centre for Criminological Research, Oxford.
Gelsthorpe, L., 1992. Social inquiry reports: race and gender considerations. Home Office Research Bulletin 32, 17–22.
Gelsthorpe, L., Raynor, P., 1995. Quality and effectiveness in probation officers' reports to sentencers. British Journal of Criminology 35 (2), 188–200.
Grünhut, M., 1952. Probation in Germany. Howard Journal 8 (3), 168–174.
Hedderman, C., 2004. The 'criminogenic' needs of women offenders. In: McIvor, G. (Ed.), Women Who Offend. Jessica Kingsley, London, pp. 227–244.
Hedderman, C., Gelsthorpe, L., 1997. Understanding the Sentencing of Women. Home Office Research Study 170. Home Office, London.
Home Office, 1984. Probation Service in England and Wales: Statement of National Objectives and Priorities. Home Office, London.
Hudson, B., Bramhall, G., 2005. Assessing the 'other': constructions of 'Asianness' in risk assessments by probation officers. British Journal of Criminology 45 (5), 721–740.
Jacobson, J., Gibbs, P., 2009. Making Amends: Restorative Youth Justice in Northern Ireland. Prison Reform Trust, London.
Kalmthout, A. Van, Durnescu, I. (Eds.), 2008. Probation in Europe. Wolf Legal Publishers, Nijmegen.
Kemshall, H., 2008. Understanding the Community Management of High Risk Offenders. Open University Press, Maidenhead.
Lewis, S., Raynor, P., Smith, D., Wardak, A. (Eds.), 2006. Race and Probation. Willan, Cullompton.
Lösel, F., 2012. Offender treatment and rehabilitation: what works? In: Maguire, M., Morgan, R., Reiner, R. (Eds.), The Oxford Handbook of Criminology, fifth ed. Oxford University Press, Oxford, pp. 986–1016.
Mair, G., Burke, L., 2012. Redemption, Rehabilitation and Risk Management: A History of Probation. Routledge, Abingdon.
Martinson, R., 1974. What works? Questions and answers about prison reform. The Public Interest 35 (Spring), 22–54.
McGuire, J. (Ed.), 2002. Offender Rehabilitation and Treatment: Effective Programmes and Policies to Reduce Re-offending. Wiley, Chichester.
McIvor, G. (Ed.), 2004. Women Who Offend. Jessica Kingsley, London.
McNeill, F., Raynor, P., Trotter, C. (Eds.), 2010. Offender Supervision: New Directions in Theory, Research and Practice. Willan, Abingdon.
Ministry of Justice, 2013. Transforming Rehabilitation: A Strategy for Reform. Ministry of Justice, London.
Moxon, D., 1988. Sentencing in the Crown Court. Home Office Research Study 103. HMSO, London.
Observer, 2012. Probation officers to be replaced by electronic kiosks in pilot scheme. The Observer. April 29, 2012.
Raynor, P., 2004. The probation service 'pathfinders': finding the path and losing the way? Criminal Justice 4 (3), 309–325.
Raynor, P., 2007. Risk and need assessment in British probation: the contribution of LSI-R. Psychology, Crime and Law 13 (2), 125–138.
Raynor, P., 2012. Community penalties, probation, and offender management. In: Maguire, M., Morgan, R., Reiner, R. (Eds.), The Oxford Handbook of Criminology, fifth ed. Oxford University Press, Oxford, pp. 928–954.
Raynor, P., Miles, H., 2007. Evidence-based probation in a microstate: the British Channel Island of Jersey. European Journal of Criminology 4 (3), 299–313.
Raynor, P., Robinson, G., 2009. Rehabilitation, Crime and Justice, second ed. Palgrave Macmillan, Basingstoke.
Rock, P., 1991. Helping Victims of Crime. Oxford University Press, Oxford.
Sheehan, R., McIvor, G., Trotter, C. (Eds.), 2007. What Works with Women Offenders. Willan, Cullompton.
Sheehan, R., McIvor, G., Trotter, C. (Eds.), 2011. Working with Women Offenders in the Community. Willan, Abingdon.
Underdown, A., 1998. Strategies for Effective Offender Supervision: Report of the HMIP What Works Project. Home Office, London.
Vanstone, M., 2004. Supervising Offenders in the Community: A History of Probation Theory and Practice. Ashgate, Aldershot.
Vanstone, M., 2008. The international origins and initial development of probation: an early example of policy transfer. British Journal of Criminology 48 (6), 735–755.
Whitehouse, P., 1983. Race, bias and social enquiry reports. Probation Journal 30 (3), 43–49.
Worrall, A., 1990. Offending Women. Routledge, London.

Problem Selection in the Social Sciences: Methodology

Wout Ultee, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociological methodology has rules for testing hypotheses more severely, and for the formulation of theories rich in content. This article presents rules for problem selection by surveying cases from sociology's past and present. Among these rules are treating a seemingly small question as a subproblem of a larger one, expanding existing problem structures by adding new subquestions, revising a problem structure by a new overarching question, picking out contradictions between theories and research findings, and replacing false assumptions behind questions.

Social methodology textbooks state techniques of proper empirical analysis. For instance, they tell how partial coefficients of association are to be computed, adding that these coefficients provide a more severe test of multifactor explanations than bivariate associations. Books on theory building in sociology set standards for determining whether certain hypotheses are better than others. Logically unfalsifiable ones are out; future propositions should provide more information than current ones. Any science starts with problems. Are there precepts for selecting problems too? Yes, and this article draws upon principles intimated by sociologists (Lenski, 1966: p. 20; Blau and Duncan, 1967: pp. 9–10, 194–199, 402), proposed by philosophers of science (Bunge, 1967: pp. 165–221), and applied by sociologists to sociology (Merton, 1959; Ultee, 1980). These standards are elucidated by cases from sociology's past and present.

Overarching Problems and Subproblems

Durkheim 'was not interested in suicide at all,' so says Collins' (1994) textbook on sociological traditions. But then, why did Durkheim dub his 1897 study *Le Suicide*? Weber in 1910 grumbled that his critics understood him to explain the rise of capitalism fully from the ethos of Protestantism, and retorted that he only dealt with the question of whether this ethos fitted the spirit of entrepreneurial capitalism. Yet in the 1920 preface to his collected papers on world religions, Weber's interest was rationalization processes, including the rise of capitalism. These changes had gone furthest in the West, and Weber wondered why.

If these two stories about sociology's questions carry a moral, the lesson is not that persons nowadays regarded as sociology's founders, did not know what they were up to. The message amounts to a methodological command on meritorious problems. Durkheim grasped that seemingly small questions become pertinent when taken as subproblems of larger ones, and Weber understood that it is wise to divide mammoth problems into more modest ones, and those in their turn into even more tractable questions. Durkheim and Weber entered sociology's hall of fame because they pointed toward a bunch of smaller questions falling under one and the same heading. The subtitle of *Le Suicide* runs *Étude de sociologie*, lifting a phenomenon from psychology into that science which studies societies – and the extreme case of leaving them.

Problem Structures and Their Dynamics

For Durkheim, a society's suicide rate tells about its cohesion; the higher this rate, the less cohesive a society. Weber observed that a society's inhabitants do not always use their economic freedoms, for instance, those of enterprise and labor, and supplemented the problem of how these rights got included in a society's code of law, with the problem of the strength of a person's motivation to use these liberties. Both questions were part of the question of why entrepreneurial capitalism arose in the West. In its turn, the question of the rise of capitalism, including bankers' as well as entrepreneurial capitalism, was a subquestion of the even larger question of the speed of rationalization processes in the West during the last couple of centuries.

In Durkheim's *oeuvre*, the question of a society's degree of cohesion also comprises long-term trends in the prevalence of economic exchanges when comparing Ancient Israel, the Roman Empire, and medieval Burgundy, and the degree to which assemblies of hunters and gatherers are religious in nature. Weber's question of rationalization led to less grand ones on the rise of science and the bureaucratization of states. Any methodology of problem selection should contain the rule that questions should not be judged on their own, but according to their place in a structure of overarching problems and subproblems. However small a problem seems, it has more appeal if it is part of an overarching problem. And however towering a question appears, it grows less scary if it is broken down into subquestions.

The Durkheimian Problem Structure

Figure 1 depicts the Durkheimian problem structure. The novelty of *Le Suicide* was not simply that Durkheim turned the question of suicide into a subquestion of the larger one of societal cohesion. Hobbes (1651) took order and its opposite violence as the master problem of those studying the body politic. For him, civil wars and wars between sovereigns were part of the problem of order and violence. Next to it, Durkheim placed the problem of the presence and absence of ties between a society's members. The counterpart of the question 'why do people not always live in peace?' is not only 'why do they sometimes use violence?' but also 'why do they sometimes live relatively isolated from each other?' For Durkheim, suicide was the supreme example of severing the ties between persons and the society in which they live.

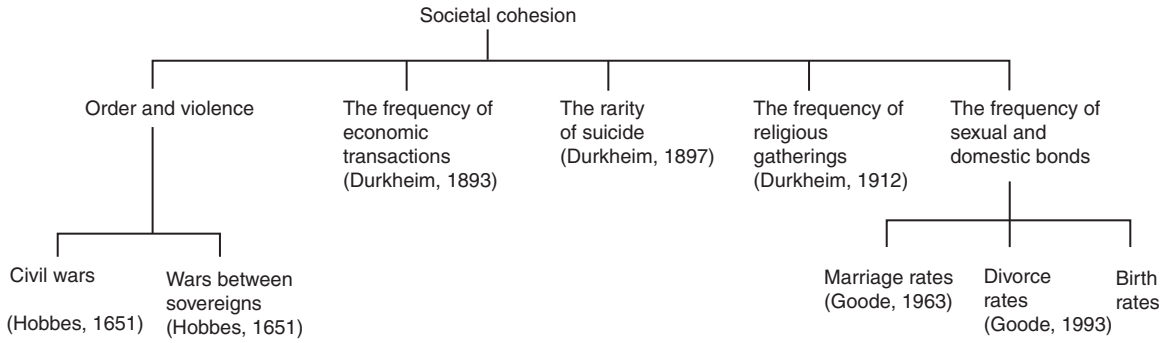

Figure 1 The Durkheimian problem structure and its dynamics.

Apart from *Le Suicide*'s question, Figure 1 encompasses the questions guiding Durkheim's two other main studies. *De la division du travail social* from 1893 focuses the frequency of economic transactions between a society's members at various points in history. It narrows this question down to that of the degree to which cooperative law (civil law regulating economic contracts by restitution) is more widespread than criminal law (penal law involving repressive sanctions). *Les formes élémentaires de la vie religieuse* from 1912 notes that Australian aboriginals disperse when hunting and gathering their food, and asks to what extent their leisurely meetings celebrate a totem and turn them into members of one society. Durkheim's questions involve a shift from questions on economic ties, by way of questions on coliving, to questions on religious ties, yet they always pertain to societal cohesion.

Given the Durkheimian problem structure, later scholars can raise questions on types of attachment to which Durkheim paid limited or no attention. Examples are questions on changes in the rates of marriage, birth, and divorce in industrial societies. A wedding creates a sexual and domestic bond between two members of the opposite sex, divorce dissolves it, and a higher birth rate amounts to more parent–child relations in a society. focuses on the making and undoing of sexual and domestic ties. These monographs amount to a new branch of the Durkheimian problem structure.

The Weberian Problem Structure

The Weberian problem structure and its dynamics are given in Figure 2. The overarching problem is that of rationalization, and Merton (1938) pursued the subproblem of the rise of science in the West, and Blau (1974) that of bureaucratization. The art historian Gombrich (1960) studied questions on the skills of painters to furnish a faithful representation.

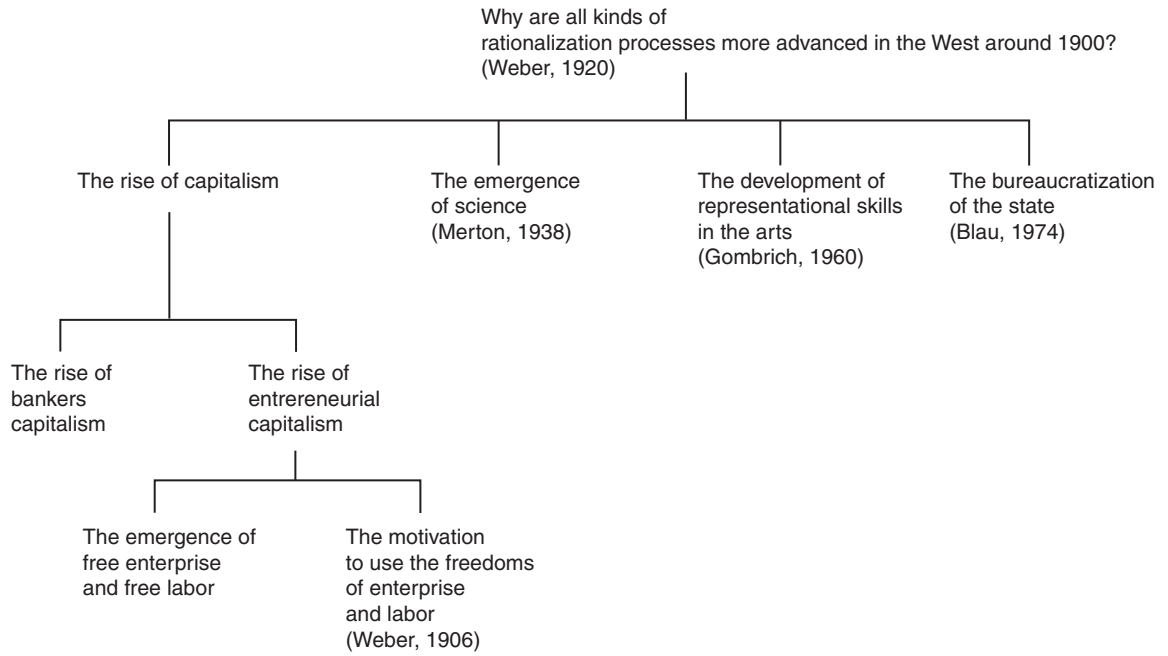

Figure 2 The Weberian problem structure and its dynamics.

The relation between problems as depicted in Figures 1 and 2 are logical in nature. An overarching problem, usually by way of uncontroversial auxiliary assumptions, implies subproblems. It is another, and less important, matter whether a person raising a subproblem of an overarching problem of an earlier scholar studied under that master or read a book by her or him. Merton and Blau took the lead from Weber; Gombrich was not influenced by him. Weber invokes Heinrich Wölfflin once when discussing the question of representational skills; Gombrich quotes Wölfflin frequently.

As stated, the step from an overarching question to a subquestion involves largely undisputed assumptions. Few would argue that capitalism is less efficient than slavery. Yet this proposition was belied by Fogel and Engerman's (1974) test for the American South shortly before slavery was abolished. A scholar working on a question can make progress by replacing these auxiliary assumptions by better approximations. Weber did not do so, but Durkheim did. In countries prescribing suicide under certain circumstances, suicide is not an instance of low cohesion, but of strong solidarity. In that case, Durkheim spoke of altruistic suicide.

Behind Weber's distinction between bankers' and entrepreneurial capitalism lurks evidence speaking against the hypothesis that the rise of capitalism is always accompanied by the growth of banking. Protestant businessmen did not take out loans. Their religion put a premium on frugality, and contributed to reinvestment of profits in their enterprise. The background of Weber's distinction between laws granting people economic freedoms and the use of these freedoms by a society's members forms the observation that people sometimes do not use their rights, as well as the hypothesis that they exercise these rights only if they are motivated to do so. Classical economics assumed that every person had this drive; Weber held that Protestantism strengthened it. A wise rule is to state the background of a question as fully as seems necessary for bringing out a problem structure.

The Weberian Problem Structure Radically Revised

The division of an overarching problem into subproblems may drastically alter. Studies by Smith (1776), Elias (1939), Samuelson (1954), Van den Doel (1979), and De Swaan (1988) imply the revision of the Weberian problem in Figure 3. Weber never made clear what state bureaucracies were doing. States possess the legitimate monopoly on the means of violence; Elias added that states wield the legitimate monopoly on taxation. Were all taxes spent on armies, police, and prisons? This obviously is not the case. Here, the hypothesis is useful that markets make for the optimal production of individual goods (Smith), but not for the optimal production of collective goods (Samuelson). The production of collective goods approaches optimality in a state with universal suffrage and a bureaucracy that executes its laws (Van den Doel).

De Swaan listed collective bads and laws contributing to their elimination and to the production of collective goods. First, poverty is a collective bad. Poor people are more likely to trespass the law, and crimes form a general threat to life and property. Following De Swaan, this danger is alleviated by state insurance against disability and unemployment. Contagious diseases amount to a second collective bad. If local authorities provide each household with piped water and an outlet on the sewers, public health improves. Third, compulsory schooling lowers the percentage of children who cannot read and write, and makes for a more optimal functioning of markets and bureaucracies. Weber's conception of the state is saved by the assumption that armies, police, and prisons avert collective bads. Armies lessen a threat to a state's independence, police patrols contribute to safety on the streets, and prisons deter.

Problem Structures and Theories

As the saying goes, a question well put is a question half answered. If it is clear that some problem is part of a larger problem, and if a solution for the overarching problem is at hand, a possible solution to the smaller problem is available too. But then, if an answer to a particular smaller question of a larger one turns out to be false, the opportunities to disestablish a solution to another subproblem of the larger one increase. Additionally, the notion that problem structures exist and have a dynamic makes the project of sociology have more continuity than suggested by Boudon (1979) and Gove (1995).

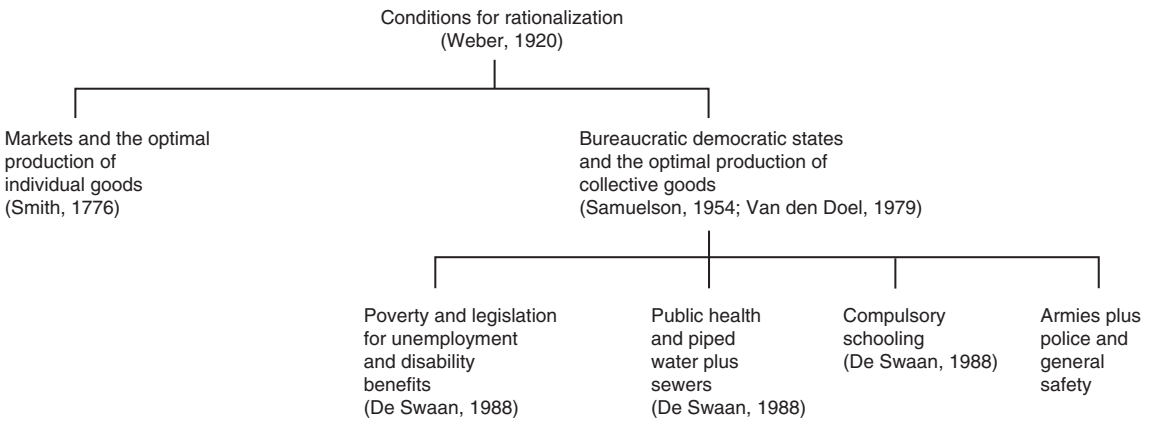

Figure 3 The Weberian problem structure radically revised.

They accepted that sociology's questions to outsiders show a bewildering variety, and bypassed the possibility that one and the same larger question may harbor a host of seemingly disparate questions.

Contradictions and Issues

One way of formulating a problem in a more useful way is to ascertain its place in a problem structure. Are there much injunctions for problem choice? Sociologists debating sociology have argued about the primacy of theory versus the supremacy of research. In contrast, Popper (1962) defended that every science, including sociology, most fruitfully starts with contradictions between general hypotheses and specific research findings.

Contradictions – or, to use (Kuhn's 1962) terms, puzzles and anomalies – contribute to sociology's growth. For Merton's (1973) sociology of science, the prevalence of priority disputes in a field was surprising in several respects. Its explanation was not the egotism of the scientists having made discoveries, since their part in the disputes was small. Nor was it to be accounted for by the increasing competition between scientists, since the frequency and intensity of priority disputes in the West decreased over time. A rule of method runs: single out contradictions.

Incompatibilities between hypotheses also invite new research and theorizing. They sometimes are called issues, but that term also refers to false alternatives. An example is the question of whether intelligence is determined by social or by genetic factors. Intermediate positions are possible, the list of social factors may be incomplete, and the location of intelligence on genes may be unknown. Popular science states issues in a metaphorical way; are we born with a brain hardwired with rules for language, or is the brain a big sponge, capable of creating rules for language by mopping up what we hear?

The phrasing of a contradiction problem sometimes contains seeds for its solution. For instance, Huizinga (1941) asked how it was possible that painting and other forms of culture flourished in the Netherlands during the seventeenth century, so shortly after this society became independent from Spain. His answer was that Dutch culture blossomed because the Netherlands was a new state. In states that have recently become independent, the upper strata have risen from the lower and the more open a country's elite, the more its culture will flower. Trevor-Roper (1976) wondered how it was possible that an emperor as preoccupied with expanding his territory as Charles V (1500–58) furthered the art of painting so much. In the end, Trevor-Roper surmised that Charles V commissioned paintings to the extent that, and precisely because, they legitimated his political power.

Apart from the rule for theory building to mobilize the information in a contradiction problem, there is the precept to turn the problem around. Rather than asking why the number of unruly passengers on airlines is increasing, it might be worthwhile to ponder why this number until now did not increase much more strongly. A more academic example is as follows. Marxism noted that a lot of workers refrain from joining unions and strikes, and raised the question of how workers come to have a false consciousness. Olson (1965) held that unions and strikes produce collective rather than private goods, applied the prisoners' dilemma from game theory in mathematics, predicted that rational workers would not strike, and raised the question of why people participate in social movements. However, it should be clear that any theory with some substance to it should explain that some people join labor unions and certain others do not.

Strings of Questions

The terms problem and question sometimes are used interchangeably, referring to queries of any specificity. However, everyday usage of expressions like object of investigation, task, question, and problem suggests that some inquiries are more focused than others. What does the articulation of problems involve?

Problem Articulation

If it is maintained that sociology is about societies (and not about human beings taken on their own), it sometimes is said that the object of research has been stipulated. Yet no single feature of this unit has been singled out for explanation. If this property is taken to be, say, the 'functioning' of a society, the object of inquiry turns into what has been called a task. Yet the circumscription of this aspect remains limited, since the term does not focus any particular way of functioning. If a distinction is made between optimal and suboptimal functioning, the study involves a question, with that term being taken in a narrow sense. Of course, other types of functioning may be distinguished too. Questions in their turn do not allude to possible causes. If research on some sample of societies determines the extent to which less stability in a society's environment goes together with a less optimal functioning of this society, the matter at hand is articulated so fully that the problem now is whether some logically falsifiable hypothesis is true. Putting a question mark behind a falsifiable proposition amounts to posing a problem in some narrow sense of that word.

The example of a string of increasingly articulated queries just presented neglects any difficulty in obtaining measurements for the degree of stability in a society's environment, and the extent to which the functioning of a society approaches optimality. It has been held that questions should refer to directly observable phenomena only, and in the human sciences therefore to outward behavior and not to mental states. This rule is a leftover from a wrongheaded empiricist philosophy of science. All terms to some extent are theoretical, and any proposition whatsoever, whether about acts or thoughts of persons or processes and situations within of societies, never can be definitely proved.

From Descriptive to Explanatory Questions

There are strings of questions of another kind. A researcher's early questions often are descriptive: how many of the people

living in 1972 Britain and born to working-class parents had moved up to the higher classes? A follow-up question might be whether this probability to climb equals the chance to descend from the upper to the lower classes. A yet later question presupposes the answers to both earlier questions, and seeks to ascertain whether upward and downward mobilities were as prevalent in 1949 as in 1972. By raising additional descriptive questions, narratives become richer.

At some point, researchers leave descriptive questions behind them, to raise explanatory questions. The explanatory question following up on the previous string of descriptive questions ran as why was during the 1970s upward mobility less widespread in Britain than in Sweden, and more widespread than in France? The obvious answer to that question was that left-wing governments make for more upward mobility (Heath, 1981; Goldthorpe, 1987). And after a comparison of upward mobility in several industrial societies at the same point in time, the follow-up question may be raised how these societies compare over longer periods (Erikson and Goldthorpe, 1992; Breen, 2004).

From Superficial to Deeper Explanatory Questions

Some explanations 'scratch the surface only,' others delve deeper. Why would particular ideologies make for more upward mobility? Just like individuals, governments cannot always attain their goals. By which means then do social-democrats attain the goal of more upward mobility? Does Sweden have more upward mobility than Britain, because social-democrats make children branch out at a later age to schools that demand more or less of pupils? Or is this the case because social-democrats pursued a policy of full employment? And then, if the answer to the first question is 'no' and that to the second one 'yes,' why would there be more upwardly mobile persons in a society without unemployed persons?

There is no obvious point at which the hunt for deeper explanations should stop. Some theoreticians in effect defend that 'the buck stops here.' No one nowadays seriously defends that propositions about societies are logically independent from propositions about individuals. However, whether rock-bottom explanations of the features of societies always should refer to the motives and aversions of a society's members, as well as the opportunities they have and the barriers they face, is controversial. In the discussion on 'methodological individualism' (O'Neill, 1973), the borderline between methodological rules and substantive requirements was violated. Perhaps it is wise to no longer state this debate in terms of theories explaining societal phenomena, but to settle it by recognizing that short-hand statements of sociology's questions refer to societal features only, but that these questions, if they are stated fully, always refer to the persons making up a society's population, that is, to individuals and societies (Ultee, 1998).

From Meager to Thick Descriptive Questions

Some descriptive questions – and their ensuing explanatory questions – can be meager, whereas others aim for lavishness. Thus, it often has been ascertained that no society ever has been and ever will be completely egalitarian. As Lenski (1966: p. 20) maintained, in the debate between the conservative and the radical tradition in sociology, it is possible to make headway by transforming problems stated in a categorical way into problems phrased in variable terms. It may indeed be the case that full equality never has been observed in any society; however, is the lowest degree of inequality observed until now the lowest that is possible? And do long-standing differences between highly industrialized countries in their ruling ideology explain differences in their degree of income inequality? Questions about equality that demand 'yes' or 'no' are less precise than questions about degree of inequality.

Following Lenski, progress also can be made by breaking down the compound concepts contained in problems into their constituent elements. Thus, the question of whether power – in the singular – determines privilege – in the singular – will not do. Rather, the question is to what extent differences in the consumption of a certain good in a particular society depend on the control of various resources. When answering this question, Lenski distinguished the skills of hunters and gatherers, the size of gardens in horticultural and of fields in agrarian societies, the possession of metal weapons, the private ownership of machines, the technical skills of the people designing machines, the range of legitimate command accruing to a political position, and the right to vote in elections. Lenski also suggested that the change from agriculture to industry was accompanied by a shift in the right to vote. Once elections were infrequent with only the propertied classes having the right to vote, whereas nowadays there is universal suffrage for periodic elections. By breaking down compound problems into their constituents, thick descriptions and detailed explanations are obtained.

Erroneous and Well-Stated Questions

If descriptive questions are logically prior to explanatory questions, explanatory questions may be wrong. Answers to questions can turn out false, questions themselves may be mistaken. The question of why in the 1970s social mobility was more widespread in France than in Sweden is erroneous, since in that decade mobility occurred more often in Sweden. The Olson question of exactly which selective incentive makes persons protest against the hunting of seals assumes that some selective incentive is always at work, which may not be the case. This example suggests that if a rock-bottom explanation slips into a question, that problem turns into leading question.

Methodological rules that stipulate conditions banishing erroneous questions and guaranteeing well-stated ones seem doomed. It is wise to review the presupposed description, but checking has to stop somewhere. In this respect, it is interesting that the answer to an explanatory question always implies that the original question may not have been fully right. When Durkheim asked why Protestantism goes together with more suicide than Catholicism, he devised the hypothesis that more tightly knit religions show a lower suicide rate. To derive what he wished to explain, Durkheim assumed that Catholics are more integrated than Protestants. He knew that this assumption was off the mark for England, since the Anglican Church

was a state church. Yet in this way, he could account for England as the deviant case to the regularity he wished to explain. Malewski (1964) showed with examples from social psychology that every explanation of a regularity predicts conditions under which exceptions occur to the regularity to be explained.

Duncan held that the question of factors influencing the chances that the members of a society move up or down poses the issue poorly. The prime factor is father's occupation. However, the proposition that the lower the occupational prestige of a person's father, the higher this person's chances of upward mobility forms a truism. After all, the hypothesis thrives on bottom and ceiling effects. To test more interesting hypotheses, according to Duncan, questions on change in a person's life should not be stated in terms of some difference score, but as questions on the strength of the relation between a person's earlier and later score, and on the factors influencing that relation.

Undertheorized Questions

Schütz (1932) raised the question of why human beings understand one another, and did so with a certain urgency by phrasing his question as 'how is it possible that people understand each other?' This query sounds like a contradiction, although casual observation suggests that some people at some places and in some times indeed understand one another. However, no particular hypothesis has been derived from a specific theory holding that it is impossible for human beings to understand each other. Thus, the theory rendering the question compelling is lacking. The Parsonian problem of how order is possible (Parsons, 1937) is undertheorized in a similar way. Any derivation of the hypothesis that society is violent, from the principle that persons follow their interests, involves additional assumptions. These can be wrong, which makes the question of how order is possible less pressing than it sounds. Indeed, a full statement of the question 'how is order possible' implies the conditions under which order occurs, and those under which it does not.

Research of the past decades not infrequently addressed the question of whether the association observed in a cross-sectional sample of a society's inhabitants between their year of birth and a particular characteristic is to be accounted for as an age, cohort, or period effect. As methodologists point out, this question is undertheorized too. The available hypotheses do not predict some age effect, some cohort effect, and some period effect. Those hypotheses are so vague that their conjunction is almost if not fully logically unfalsifiable. The hypotheses at hand have more substance; if people marry, grow up in lean times, and live through a war, then their values will be less postmaterialist (Inglehart, 1977). Those variables can be independently ascertained, making statistical models identifiable.

Conclusion

What are the prospects for more attention to problem selection in scholarly studies on human societies and in social methodology books? Since 1998, there is an English language textbook aiming to state (and answer) sociology's questions: *Sociology, The Central Questions* by Kornblum. One of its questions runs as "Stratification and social class, are they inevitable, or can there be a classless society?" This question may be taken as erroneous, yet the text following it makes clear that students are taught to think in gradual instead of categorical terms about inequalities. Nevertheless, this type of textbook does not seem to have caught on. As to review articles, some now distinguish generations of research. They do so not only by pointing toward more advanced techniques for data analysis and new theories, but also by invoking problem shifts. Ultee (2006) did so in research on the theme of welfare states and inequalities, and Knutsen (2007) in research on class voting. But this type of review articles has not become common yet.

Lieberson (1985) subtitled one of his social methodology books 'the improvement of social research and theory,' but pointed toward the prevalence of 'unanswerable questions,' and he pleaded for social researchers to avoid them. Abbott (2004) sought to complement rules for testing hypotheses by stating rules for thinking them up. For this, he took his inspiration from Pólya's (1957) *How to Solve It*, with it being a problem. But in the final chapter, Abbott, fortunately, was concerned with 'how to pose it,' that is, how to arrive at interesting problems, pointing toward puzzles. Firebaugh (2008) presented seven rules for good social research. At first sight, none of them refers to problems. But upon studying the argument, it turns out that Firebaugh repeatedly points toward erroneous questions. His prime example is the question of why income disparities on a worldwide scale have been growing for the past few decades, whereas they have been declining (although they have been rising in most countries of the world taken on their own). Upon closer inspection, this example also illustrates the rule that a question well put is a question half answered (the rise in average income has been strongest in the poorest and most populous countries).

In everyday life, researchers consider some problems difficult, others significant, and yet others exciting. These notions are difficult to put into formal rules. However, a more difficult problem might be taken as an explanatory question to which several solutions have been proposed, none of which worked or led to closer approximations. Thus, difficult problems are specific kinds of Popperian contradictions. Also, a structure consisting of overarching problems and subquestions goes some way toward formalizing the intuition that a problem is significant. Perhaps a problem area is called exciting, if, after many a failed attempt, the solution of an overarching problem seems near, because progress is being made on subproblems.

Of course, it may be argued, with Ziman (1987), that the question of how scientists choose their problems is erroneous, since most researchers do not have a choice. Their problems are pregiven by the organization that employs them. This observation does not do away with the question of problem choice, but makes the question of the formal rules for problem choice by the higher authorities of these organizations more pressing. The methods outlined above should be taken not only as principles for researchers who have a choice, but also as precepts for institutes run on a budget, and for foundations that award subsidies.

See also: Explanation: Conceptions in the Social Sciences; Generalization: Conceptions in the Social Sciences; Grounded Theory: Methodology and Theory Construction; Historical Explanation, Theories of: Philosophical Aspects; Kuhn, Thomas S. (1922–96); Logical Positivism and Logical Empiricism; Merton, Robert K (1910–2003); Objectivity: Philosophical Aspects; Popper, Karl Raimund (1902–94); Social Properties (Facts and Entities): Philosophical Aspects; Weber, Max (1864–1920); Weberian Social Thought, History of.

Bibliography

Abbott, A., 2004. Methods of Discovery, Heuristics for the Social Sciences. Norton, New York.

Agassi, J., 1975. Science in Flux. Reidel, Dordrecht.

Blau, P.M., 1974. On the Nature of Organizations. Wiley, New York.

Blau, P.M., Duncan, O.D., 1967. The American Occupational Structure. Wiley, New York.

Boudon, R., 1979. La Logique du Social. Hachette, Paris (1981. The Logic of Social Action. Routledge, London).

Breen, R. (Ed.), 2004. Social Mobility in Europe. Oxford University Press, Oxford.

Bunge, M., 1967. Scientific Research I. The Search for System. Springer, Berlin.

Collins, R., 1994. Four Sociological Traditions. Oxford University Press, Oxford.

De Swaan, A., 1988. In Care of the State. Polity, Cambridge.

Durkheim, É., 1893. De la Division du Travail Social. Alcan, Paris (1933. The Division of Labour in Society. MacMillan, New York).

Durkheim, É., 1897. Le Suicide, Étude de Sociologie. Alcan, Paris (1951. Suicide. A Study in Sociology. Free Press, Glencoe).

Durkheim, É., 1912. Les Formes Élémentaires de la Vie Religieuse. Alcan, Paris (1995. The Elementary Forms of Religious Life. Free Press, New York).

Elias, N., 1939. Über den Prozess der Civilization. Franke, Bern (1994. The Civilizing Process. Basil Blackwell, Oxford).

Erikson, R., Goldthorpe, J.H., 1992. The Constant Flux, a Study of Class Mobility in Industrial Societies. Clarendon Press, Oxford.

Firebaugh, G., 2008. Seven Rules for Social Research. Princeton University Press, Princeton.

Fogel, R.W., Engerman, S.L., 1974. Time on the Cross. The Economics of American Negro Slavery. Wildwood, London.

Goldthorpe, J.H., 1987. Social Mobility and Class Structure in Modern Britain, second ed. Clarendon, Oxford.

Gombrich, E.H., 1960. Art and Illusion: A Study in the Psychology of Pictorial Representation. Phaidon, London.

Goode, W.J., 1963. World Revolution and Family Patterns. Free Press, New York.

Goode, W.J., 1993. World Changes in Divorce Patterns. Yale UP, New Haven, CT.

Gove, W.R., 1995. Is sociology the integrative discipline in the study of human behavior? Social Forces 72, 1197–1206.

Heath, A., 1981. Social Mobility. Fontana, Glasgow.

Hobbes, T., 1651. Leviathan (1968 Penguin, Armondsworth).

Huizinga, J., 1941. Nederland's Beschaving in de 17de Eeuw. Tjeenk Willink, Groningen (1969. Dutch Civilization in the 17th Century. Harper & Row, New York).

Inglehart, R., 1977. The Silent Revolution. Princeton UP, Princeton.

Knutsen, O., 2007. The decline of social class. In: Dalton, J.R., Klingemann, H. (Eds.), The Oxford Handbook of Political Behavior. Oxford University Press, Oxford, pp. 457–480.

Kornblum, W., 1998. Sociology, the Central Questions. Harcourt Brace, Fort Worth.

Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lieberson, S., 1985. Making it Count, the Improvement of Social Research and Theory. University of Californai Press, Berkeley.

Lenski, G.E., 1966. Power and Privilege. McGraw-Hill, New York.

Malewski, A., 1964. Generality levels of theories concerning human behavior. In: Zetterberg, H.L., Lorenz, G. (Eds.), A Symposium on Theoretical Sociology. Bedminster, Totowa.

Merton, R.K., 1938. Science, Technology and Society in Seventeenth-Century England. Saint Catherine Press, Bruges.

Merton, R.K., 1959. Notes on problem finding in sociology. In: Merton, R.K., Bloom, L., Cottrell Jr, L.S. (Eds.), Sociology Today: Problems and Prospects. Basic Books, New York.

Merton, R.K., 1973. The Sociology of Science. University of Chicago Press, Chicago.

Olson, M., 1965. The Logic of Collective Action. Harvard UP, Cambridge, MA.

O'Neill, J. (Ed.), 1973. Modes of Individualism and Collectivism. Heinemann, London.

Parsons, T., 1937. The Structure of Action. McGraw-Hill, New York.

Pólya, G., 1957. How to Solve It. Double Day, Garden City.

Popper, K.R., 1962. Die logik der sozialwissenschaften. Kölner Zeitschrift für Soziologie und Sozialpsychologie 14, 233–248 (1976. The logic of the social sciences. In: Adorno, T.W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., Popper, K.R. (Ed.), The Positivist Dispute in German Sociology. Harper & Row, New York).

Samuelson, P., 1954. The pure theory of public expenditure. Review of Economics and Statistics 36, 387–389.

Schütz, A., 1932. sinnhafte Aufbau der sozialen Welt. Springer, Wien.

Smith, A., 1776. The Wealth of Nations (1970 Penguin, Harmondsworth).

Trevor-Roper, H., 1976. Princes and Artists. Thames and Hudson, London.

Ultee, W., 1980. Fortschritt und Stagnation in der Soziologie. Luchterhand, Darmstadt.

Ultee, W., 1996. Do rational choice approaches have problems? European Sociological Review 12, 167–179.

Ultee, W., 1998. Bringing individuals back into sociology. In: Blossfeld, H.P., Prein, G. (Eds.), Rational Choice Theory and Large-Scale Data Analysis. Westview, Boulder.

Ultee, W., 2006. Problem shifts in the study of welfare states and societal inequalities. Research in Social Stratification and Mobility 24, 377–386.

Van den Doel, H., 1979. Democracy and Welfare Economics. Cambridge UP, Cambridge.

Weber, M., 1920. Gesammelte Aufsätze zur Religionssoziologie. Mohr, Tübingen (1930. The Protestant Ethic and the Spirit of Capitalism. Allen & Unwin, London).

Winckelmann, J. (Ed.), 1978. Maw Weber. Die Protestantische Ethik II. Kritiken und Antikritiken. Gütersloh, Mohn.

Ziman, J.M., 1987. The problem of "problem choice". Minerva 25, 92–106.

Zuckerman, H., 1978. Theory choice and problem choice in science. In: Gaston, J. (Ed.), Sociology of Science. Jossey-Bass, San-Francisco.

Problem Solving and Reasoning: Case-Based

David Leake, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Case-based reasoning (CBR) is a paradigm of artificial intelligence and stance for cognitive science research. CBR models the reasoning process as primarily memory based, taking retrieved prior cases as a starting point for solving new problems and interpreting new situations. Fundamental problems include how to guide retrieval of relevant cases and how to adapt them to new needs. The CBR perspective is shaped by studies of human cognition, but the resulting approach has proved effective for developing applied intelligent systems in a wide range of domains. New directions include efforts to harness experiential knowledge on the Internet.

Case-based reasoners solve new problems by retrieving stored 'cases' describing similar prior problem-solving episodes and adapting their solutions to fit new needs. As case-based reasoners process each problem, they learn by storing records of their problem solving as new cases to guide future reasoning. Case-based reasoning (CBR) research studies the CBR process both as a model of human cognition and as an approach to building intelligent systems. Principles from CBR research serve as a foundation for applied computer systems for tasks such as supporting human decision making, aiding human learning, and facilitating access to electronic information repositories.

CBR as Cognitive Model

A major motivation for initial research on CBR was the observation of how humans remember, reason, and learn from experiences. Studies of story understanding by Schank and others noted that people who encounter surprising events are often reminded of similar past events, which they use to help them understand and respond to the new situations. Studies of reminding events investigated human memory organization, giving rise to dynamic memory theory (Schank, 1982), which included theories of how episodes are organized and indexed in human memory to be retrieved across different contexts, and the circumstances under which reminding events are triggered. Reminding events are often failure driven, occurring when expectations are violated. For example, a doctor whose patient's condition plummets after taking a medicine might be reminded of a previous patient's similar reaction.

Failure-driven reminding events can facilitate learning by providing additional data as a basis for generating and testing candidate explanations. They may also provide information about past solutions; remembering how the past emergency was treated may suggest a treatment for the new patient. Even if the past solution does not apply directly, as when the new patient is allergic to the drug used in the previous case, it may suggest a class of drugs to use. The flexibility to apply prior cases to solve new problems arises from flexible retrieval – the ability to retrieve cases that are relevant, even if they reflect different circumstances – and adapting their solutions as needed to fit the new situation.

Being reminded of prior problem-solving failures can be useful as well. When a doctor is reminded of a mistake or failed treatment, remembering the outcome of the previous treatment helps to predict the new outcome and warns of potential pitfalls to avoid. Thus CBR supports learning from either successes or failures.

Psychological studies provide evidence supporting human CBR for problem solving in a wide range of tasks such as learning programming, performing mathematical problem solving, explaining anomalous events, and decision making (see Leake (1998) for examples of this work).

CBR also plays an important role in human cognition for classification and interpretation. For example, in legal reasoning in the American legal system, prior cases are the basis for legal judgments: lawyers build arguments about how to interpret rules of law by comparing and contrasting precedent cases to the current situation (see Ashley and Rissland, 1988).

Research on CBR as a cognitive model studies the knowledge and processes that underlie human CBR, addressing topics such as memory organization, similarity assessment, and analogical mapping (Kolodner, 1993; Leake, 1998). Computational models developed in this work illuminate functional constraints on reasoning processes and can provide testable hypotheses about human reasoning for psychological investigation.

Understanding of human CBR process is closely related to understanding of similarity and analogical reasoning and learning. A summary of the relationship of CBR to psychological research on analogy is included in Mantaras et al. (2005).

CBR as an Intelligent Systems Approach

Apart from cognitive science, CBR has been studied as a knowledge-based systems approach to developing intelligent systems. The use of CBR in intelligent systems is motivated both by considerations for system development and for user interactions. Some key motivations are to facilitate knowledge acquisition, to facilitate knowledge maintenance, to increase problem-solving efficiency and solution quality, and to benefit user acceptance. Knowledge acquisition is a classic problem for rule-based reasoning. Some domains are hard to codify in

terms of the general rules assumed by rule-based systems; for example, domains in which rules would require too many exceptions. In general, acquiring rules from experts is difficult, as they find it hard to characterize their reasoning in rule-based form.

However, experts find it natural to share their 'war stories' of experiences, and those experiences correspond to cases. Such cases need not capture the entire problem-solving structure of the situation, but instead need only to include the information needed to determine applicability and to enable adaptation of the old solution to new circumstances. This is not to say that CBR avoids all knowledge acquisition issues: in CBR, traditional rule acquisition issues are replaced by issues of capturing and representing reasoning episodes, selecting similarity and indexing criteria, and capturing case adaptation knowledge. However, in practice, these may be more tractable.

In addition, the initial case knowledge of a CBR system need not provide complete coverage of the problem space. Systems can be fielded with a limited set of 'seed cases,'which are augmented by acquiring missing cases when existing cases prove inadequate during problem solving. Only the parts of the problem space in which problems arise in practice need to be covered.

CBR can provide efficiency benefits when problems are time consuming to solve from scratch, but stored cases provide a sufficiently close starting point to be easily adaptable to new needs. For example, generating a new home design may be expensive, but customizing that design to reflect customer preferences (e.g., by changing kitchen appliances) may be simple. Quality benefits arise from the benefits of experience. The successful prior use of a previous case attests to its quality; the debugging needed to apply it in the past is automatically incorporated, avoiding repeating past mistakes.

User acceptance is furthered by the ability of CBR systems to provide natural explanations of decisions by accounting for decisions in terms of actual previous examples, rather than in terms of rules that users may not understand or trust.

Steps in Case-Based Problem Solving

Computer CBR systems have been developed for numerous tasks including planning, design, and explanation. Regardless of the specific reasoning task, any case-based problem-solving system follows the same basic steps, referred to as the CBR cycle. In the basic CBR cycle, the system receives a problem, generates indices for retrieval and retrieves candidate cases, assesses their similarity to the new problem, adapts their lessons to fit the new situation, evaluates the result (possibly repeating adaptations as needed until the result is satisfactory), applies the result, and retains a record of the problem, solution, and outcome, as a new case to guide future reasoning. The retention of cases at the end of the cycle means that in CBR, learning is a by-product of every problem-solving episode. Thus computer models of CBR must account for four main processing steps: retrieval/similarity assessment, adaptation, evaluation/application, and storage. The process is sometimes characterized by the 'four REs': retrieve, reuse, revise, and retain (Aamodt and Plaza, 1994).

Retrieval

In order to facilitate retrieval of relevant cases, stored cases are organized by indices characterizing circumstances in which the cases are likely to be relevant. For example, in a medical CBR system to suggest treatments for illnesses, treatment cases could be indexed by the diseases they treat and the characteristics of patients (e.g., age) that tend to determine whether a given treatment is appropriate. The indices for a given domain may be determined by knowledge engineering or learned, for example, by induction or by generating explanations to determine feature importance. Cases capturing structured information (e.g., plans or workflows) may be indexed by substructures as well as content-based flat indices.

Given a new problem to solve, a case-based reasoner first performs situation assessment to analyze the input problem and describe it in terms of the indices used to organize cases. For example, in a medical context, rather than trying to diagnose based only on the patient's current vital signs, it may be important to derive more abstract features such as 'deteriorating condition,' based on consideration of the patient's history.

Retrieval algorithms use the description to retrieve the cases in memory expected to be most relevant. For example, nearest neighbor retrieval algorithms select the most similar prior cases according to a metric that weights features by their relative importance. Alternatively, case libraries may be organized into discrimination networks to enable retrievals based on only a small number of feature comparisons. After candidate cases have been retrieved, a finer grained similarity assessment process is often used to identify the best candidates and to determine differences between the old and new situations, for which the old solution must be adapted.

Sometimes a complete problem description is not available at the start of problem solving. For example, for diagnosis, the symptoms that are relevant may be unknown until the problem is diagnosed. Such problems can be addressed by conversational case-based reasoning (CCBR) (Aha et al., 2006). CCBR is an incremental process, which begins with the user providing an initial input, from which the system identifies a preliminary set of relevant cases and determines features most relevant to discriminating between those cases. The user is presented with a list of questions about those features, ordered by their expected usefulness, and can choose which of those questions to answer. The user can also examine any of the presented cases. Each time the user answers a question, the system revises the list of questions and cases, taking into account the new information to narrow down candidate cases. The cycle terminates when only one case remains (i.e., when a unique case matches the problem description), when cases are exhausted (i.e., when no case sufficiently matches the problem description), or when the user is satisfied with one of the cases from the set of candidates and chooses not to continue the dialog. CCBR question selection algorithms are designed to minimize the expected length of the dialog, given the system's cases. Question selection may also reflect information about user characteristics and contextual factors, such as the difficulty of providing answers to particular questions. For example, in a medical diagnosis system, the system might favor easy-to-answer questions over those that can only be answered by invasive tests.

Case Adaptation

The case adaptation step revises the lessons of the retrieved case (or revises and combines the lessons of multiple retrieved cases) to generate a new solution. Adaptation may involve operations such as transformation of the structure of a solution or substitution of components. For example, a doctor might combine elements of prior treatments to address multiple complaints, or might adapt the treatment used for an adult to a treatment for a child by substituting a medication approved for children. Adaptation may also require operations such as numerical adjustments, for example, to reduce an adult dosage to suit a child.

CBR systems often use rule-based methods to transform prior solutions, applying general adaptation rules. Model-based reasoning has been applied to guide adaptation based on causal structure, for example, when adapting device design cases. Constraint satisfaction methods have been used to assure that adaptations respect problem constraints. CBR has been applied to the case adaptation problem itself, to capture and reuse adaptations performed by a user and for speedup in learning from adaptations that the CBR system generates by other methods. An alternative approach to case adaptation is to use derivational analogy to account for new circumstances by replaying a prior solution generation process rather than revising a prior solution (Veloso, 1994).

In interpretive CBR, the case-based reasoner does not modify the retrieved case, but instead develops justifications for why the prior interpretation applies (or does not apply) to the new situation. For example, in legal reasoning, a reasoner aiming to apply a precedent may develop justifications for why apparent differences are not important. Justifications may also account for why particular features should take precedence. For example, if CBR is being used to predict whether a person will find a particular movie enjoyable, justifications could compare the movie to one the person enjoyed previously, assessing not only the similarities and differences but also the relative importance of those aspects.

Evaluation/Application

Tentative results of case adaptation are evaluated to assess the appropriateness of the new solution and to determine whether additional changes are needed, possibly resulting in a cycle of repeated adaptation and evaluation. Once a solution is judged acceptable, it is applied to the real problem. If the solution fails, it may be repaired, reevaluated, and reapplied.

Storage

Each processing episode of a CBR system, whether resulting in success or failure, is stored to guide future reasoning. Cases containing successful solutions suggest solutions for similar future problems, while cases storing failures provide warnings of problems to avoid and information on how to avoid them. The storage step packages information about the problem, solution, and its results in a new case. It also assigns indices for the case and places it in the case library (also called the 'case base') for future use. CBR systems retain cases in an ungeneralized form, enabling storage of the case itself with minimal processing. However, to enable stored cases to be reused in a wide range of circumstances, CBR systems often generalize the indices under which cases are stored. The storage step may involve decision making about whether it is worthwhile to retain the new case (e.g., whether it provides new coverage).

Research issues

Retrieval and Similarity Assessment

Effective case retrieval depends on appropriate retrieval algorithms, well-organized case bases, and indices that are useful for the current task. In case-based problem solving, cases are indexed by information about the problems they solve. For example, a case-based planning system may index plan cases by the goals they achieve and constraints they satisfy. Indices should be predictive of the relevance of cases and should involve features of problems that are likely to be known to the reasoner; both abstract and concrete features may be useful. A crucial research issue in CBR is developing appropriate indexing vocabularies for particular reasoning tasks. In order to retrieve useful cases for novel situations, methods are also needed for reformulating or transforming indices to widen search when no directly relevant cases are available.

CBR research has also explored replacing traditional measures of similarity with similarity criteria that directly reflect pragmatic needs. A basic tenet of CBR is that similar problems will have similar solutions; this premise is expected to lead to similar cases being a better starting point for generating new solutions. This is normally viewed as reflecting a commonsense belief about the ease of adaptation: the solution for a similar problem will require less adaptation to generate a new solution than one that is less similar. However, it is not necessarily true that the reasoner's similarity criteria (generally based on semantic similarity) will actually ease adaptation, which has led to studies of how to link retrieval directly to adaptability (Smyth and Keane, 1998).

Research on retrieval algorithms investigates methods for achieving rapid retrievals from large case bases (e.g., by exploiting parallel algorithms). From the perspective of cognitive science, there has been much interest in the types of features that are relevant to human retrieval and the processes by which analogies are mapped to new situations.

As CBR has been explored in areas such as recommender systems, retrieval criteria beyond similarity have become important. For example, in product recommendation, it may be valuable to provide a varied collection of options, making it important to balance similarity and diversity. Likewise, different users may have different preferences, making it important to personalize the information presented to the user.

Case Adaptation

Developing case adaptation rules is often difficult because of knowledge acquisition problems: the needed knowledge may not be available or may be difficult to codify. Consequently, improving case adaptation methods is a fundamental challenge for CBR. Research areas include applying machine learning techniques to automatically generate case adaptation rules; applying CBR to a CBR system's internal processing, to adapt

cases based on prior adaptation experiences rather than rules; and using data mining and web mining methods to generate new adaptation knowledge or to support the adaptation process on the fly. It is important to note that sufficiently powerful case adaptation can enable cases to be used in very novel ways; research on flexible retrieval and adaptation processes has given rise to CBR models of creative problem solving.

Case-Base Maintenance and CBR System Maintenance

As CBR systems receive long-term use, keeping the case base accurate becomes an important problem. Likewise, as cases are accumulated, it may be necessary to reorganize or update the case base in order to improve its coverage of important problems and to avoid slow retrievals. This has led to considerable interest in methods for improving case quality and retrieval efficiency by strategic deletion, revision, and addition of cases to a case base. A major focus of this effort has been based on the idea of competence-based deletion, first proposed by Smyth and Keane, in which a CBR system reasons about which cases can be deleted with minimal risk of lost coverage (see Leake et al. (2001) for a sampling of this area). Because cases are increasingly generated by multiple distributed sources (e.g., different offices of an organization), additional issues are arising on how to update and access distributed case information.

In long-lived CBR systems, the case base is not the only knowledge container that must be maintained. For example, indices may need to be adjusted, similarity criteria may need to be changed, and source selection criteria for external knowledge sources may need to be revised, each giving rise to a current of research.

Integration with Other Methods

CBR systems may be integrated with other systems and approaches – both artificial intelligence systems and other systems – to leverage off the strengths of each. For example, case-based and rule-based reasoning may be integrated with each other, with cases used to guide interpretation of rules or to suggest targets for rule-based reasoning. CBR may also be used for speedup in learning, with initial rule-based problem solving building up a case library that can then be used instead of rules, bypassing costly reasoning from scratch, or to avoid errors by storing exceptions to rules. Conversely, rule-based reasoning may be useful to augment the case library and extend the range of problems that a CBR system can solve.

CBR in the Large

A growing focus of CBR research concerns harnessing Internet-based knowledge sources. Such sources include formal ontologies, data sets available as linked open data, knowledge that can be mined from text in general-purpose unstructured Web sources such as Wikipedia, and the vast range of experiential knowledge now available on the Web. For example, databases of reviews provide a rich source of cases generated by many individuals, which could potentially be a source for CBR. More generally, stories are available in many forms, and contain rich case information.

An important enabler for applying CBR to unstructured Web sources is textual CBR, which aims at applying CBR processes to textual sources. Web-based sources are also being exploited for roles such as providing the knowledge needed to support case adaptation.

Another active focus is exploiting structured records of processes, as captured on the Internet and elsewhere, to generate structured cases. This has spurred interest in process-oriented CBR and trace-based reasoning, which examine how CBR can be applied to business workflows, e-Science provenance traces of scientific computing workflows, and other types of raw traces, to exploit episodic knowledge to support future tasks. For example, in e-Science, provenance capture tools have been developed to automatically capture traces of the derivations of data by scientific computing processes. The process information captured by provenance traces can be mined to generate cases, which become a source for CBR systems to use to aid nonexperts at constructing workflows for future experiments.

CBR Applications

Applications of CBR include both autonomous and interactive systems. As observed by Kolodner (1993), human users are often adept at evaluating the relevance of prior cases and applying their lessons, but may lack the experience needed to acquire a sufficient set of cases. On the other hand, case adaptation and evaluation are difficult for CBR systems, but practical methods exist for accessing large case libraries. These complementary strengths have led to increasing focus on case-based 'retrieve and propose' systems that support human reasoners by providing them with relevant prior cases to aid their decision making. For example, 'help desk' systems applying CCBR have been widely used to aid telephone support personnel who diagnose customer problems. Commercial CBR 'shells', which provide core CBR functionality for managing user-supplied cases, similarity information, and adaptation knowledge, are available to facilitate the development of such applications. As novel problems and solutions are added to the case base during use, a fielded CBR system builds up a sharable 'corporate memory' of cases, aiding corporate knowledge management. This has led to the use of CBR applications in many task areas.

Two illustrations of active CBR task domains are medical reasoning and electronic commerce. CBR has received attention in many medical reasoning contexts, for tasks such as diagnosis, decision support, interpretation of test results, and prescriptive systems to advise patients; bioinformatics applications are receiving growing attention as well. In electronic commerce, case-based support systems are used during online purchases of computers, to help buyers configure their systems by presenting similar past configurations as a starting point for them to modify. This helps users find products fitting their needs, even if initially they do not know all the important features to request.

In addition, the CBR cognitive model and CBR principles are guiding the design of educational systems by, for example, suggesting when useful learning is most likely to occur and how to retrieve examples to support that learning. Likewise, CBR is

being applied to accessing hypermedia and online repositories of information – providing indexing vocabularies and example-based query interfaces – and is being used to focus on information retrieval, in order to extract task-relevant information from large-scale information sources. Another active research area, textual CBR, develops methods for facilitating knowledge acquisition by automatically capturing and indexing cases in textual form.

Relationship to Other Artificial Intelligence Methods

Much artificial intelligence research focuses on methods for generating new solutions from scratch by chaining together general rules. CBR is based primarily on reasoning from specific experiences rather than general principles. In addition, learning is an integral part of CBR; much knowledge is acquired as a by-product of problem solving rather than precoded.

The learning done by CBR systems strikes a middle ground between knowledge-poor and knowledge-rich learning approaches. Unlike knowledge-poor inductive learning methods, CBR exploits domain knowledge when available. For example, many CBR systems use explanations to identify causally relevant features, guide indexing, and case adaptation. However, CBR does not depend on a complete and correct domain theory; unlike knowledge-rich methods such as explanation-based generalization, CBR can still be used in the absence of a rule-based domain theory.

In its emphasis on building up knowledge by learning, CBR research is in the same spirit as research on neural networks. However, CBR contrasts in storing and reasoning from examples rather than storing only learned generalizations. In addition, CBR performs 'lazy learning.' Rather than immediately generalizing training instances and discarding the examples, as in, for example, explanation-based and neural network learning, CBR stores cases with minimal processing and processes them further only if and when adaptation is needed to solve new problems. In addition, unlike neural networks, CBR systems can change behavior immediately in response to new examples: a single correct example of dealing with a situation, and no further training, is all that is needed for the system to process that situation correctly in the future.

Ian Watson observed that the use of a CBR approach does not require the use of any specific technology, but rather, requires the use of a reasoning process based on retrieval and adaptation of prior experiences, which may be realized in many different ways. For example, a CBR system could rely on a database system to store cases, and a rule-based system to adapt them.

Conclusion and References

A case-based problem solver solves new problems by retrieving stored problem cases and adapting their lessons to fit new circumstances; it learns by storing new episodes for future use. This process can provide robust reasoning and learning even in domains that are hard to codify in rule-based forms. As illustrated in the previous sections, CBR research has provided computational cognitive models of human information processing, artificial intelligence techniques, and the basis of information technology for a wide range of task domains.

The CBR cognitive model is discussed by Kolodner (1993) and Leake (1998). A volume edited by Leake (1996) begins with an introduction and tutorial, and then surveys the foundations of the field with chapters on classic issues and applications in tasks such as decision aiding, knowledge navigation, design, diagnosis, and education; Kolodner (1993) provides an in-depth examination of key issues; and a volume edited by Riesbeck and Schank (1989) includes chapters on important early CBR research projects and 'microversions' of the computer programs those projects developed. The process of developing CBR applications is described by Watson (2003). Mantaras et al. (2005) provide an extensive survey of numerous approaches to key CBR areas, as part of a journal special issue with commentaries on many facets of CBR (Aha et al., 2005).

See also: Deductive Reasoning Systems; Problem Solving and Reasoning, Psychology of.

Bibliography

Aamodt, A., Plaza, E., 1994. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Communications 7 (1), 39–52.
Aha, D., Marling, C., Watson, I., 2005. Case-based reasoning commentaries: introduction. Knowledge Engineering Review 20 (3), pp. 201–202.
Aha, D., McSherry, D., Yang, Q., 2006. Advances in conversational case-based reasoning. The Knowledge Engineering Review 20 (3), 247–254.
Ashley, K., Rissland, E., 1988. A case-based approach to modelling legal expertise. IEEE Expert 3 (3), 70–77.
Kolodner, J., 1993. Case-Based Reasoning. Morgan Kaufmann, San Mateo, CA, pp. 21.
Leake, D. (Ed.), 1996. Case-Based Reasoning: Experiences, Lessons, and Future Directions. AAAI Press/MIT Press, Menlo Park, CA.
Leake, D., 1998. Cognition as case-based reasoning. In: Bechtel, W., Graham, G. (Eds.), A Companion to Cognitive Science. Blackwell, Oxford, pp. 465–476.
Leake, D., Smyth, B., Wilson, D., Yang, Q. (Eds.), 2001. Maintaining Case-Based Reasoning Systems. Blackwell special issue of Computational Intelligence, 17 (2).
Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M., Cox, M., Forbus, K., Keane, M., Aamodt, A., Watson, I., 2005. Retrieval, reuse, revision, and retention in CBR. Knowledge Engineering Review 20 (3).
Riesbeck, C., Schank, R., 1989. Inside Case-Based Reasoning. Lawrence Erlbaum, Hillsdale, NJ.
Schank, R., 1982. Dynamic Memory: A Theory of Learning in Computers and People. Cambridge University Press, Cambridge, UK.
Smyth, B., Keane, M., 1998. Adaptation-guided retrieval: questioning the similarity assumption in reasoning. Artificial Intelligence 102 (2), 249–293.
Veloso, M., 1994. Planning and Learning by Analogical Reasoning. Springer Verlag, Berlin.
Watson, I., 2003. Applying Knowledge Management: Techniques for Building Corporate Memories. Morgan Kaufmann, San Mateo, CA.

Relevant Websites

Extensive case-based reasoning information is available from the CBR Wiki, http://cbrwiki.fdi.ucm.es.

David Leake, Indiana University.

Problem Solving and Reasoning, Psychology of

Philip N Johnson-Laird, Princeton University, Princeton, NJ, USA; and New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.A. Simon, volume 18, pp. 12120–12123, © 2001, Elsevier Ltd.

Abstract

This article reviews what is known about how humans reason and how these processes underlie problem solving. The orthodox view was that deductive reasoning depends on formal rules akin to logic, but the evidence implies instead that it depends on mental simulation: individuals rely on mental models of the possibilities to which premises refer. This account applies to inductions of conclusions and to abductions of explanations. The three sorts of reasoning – deduction, induction, and abduction – underlie the solution of complex problems. But, problem solving can also demand creative insights, and the article concludes with an analysis of their nature.

Introduction

Psychologists have investigated reasoning and problem solving for over a century. Their studies of reasoning at first took for granted that reasoners relied on some sort of logic. Their studies of problem solving were scattershot investigations but began to be integrated in unified theories after the development of computer programing. This article begins with the psychology of the three main sorts of reasoning. In deduction, the aim is to draw a conclusion that must be true granted the truth of the premises, that is, to make a *valid* inference. In induction, the aim is to draw a plausible conclusion that goes beyond the given information, and so there is no guarantee of validity. In abduction, the aim is to induce an explanation of or hypothesis about the given information. Hence, abduction is a special case of induction, but one that aims to go beyond, say, a mere generalization of some facts in order to explain them. These three sorts of reasoning often underlie the solutions of problems, but problem solving may also depend on insights – a form of creative thinking. The article accordingly concludes with a description of how people solve problems.

Reasoning

Deduction

Problem 1: Either the fault is in the cable or the fault is in the software. The fault is not in the cable. So, what follows?

The answer depends on deductive reasoning, and the traditional arbiter of deduction is logic, which stipulates that an inference is *valid* if its conclusion is true in any case in which its premises are true. A valid solution to problem 1 is that the fault is in the software. Logicians distinguish between formal systems for deriving the proofs of inferences and semantic systems for establishing the validity of inferences. Sentential logic deals with inferences that hinge on sentential connectives, such as idealized versions of 'if,' 'or,' and 'and.' Its formal system can be framed using rules of inference for each of the connectives, for example:

- A or B.
- Not-A.
- Therefore, B.

The logical form of the premises in problem 1 matches this rule granted that *A* denotes 'the fault is in the cable, '*not-A* denotes its negation, and *B* denotes' the fault is in the software.' So, a single application of the rule allows us to derive the conclusion that the fault is in the software. In contrast, the semantic system for sentential logic makes use of truth tables, which display the truth of a proposition as a function of the truth or falsity of its constituent propositions. So, the truth table for *A* or *B*, or *both* is as follows:

A	*B*	*A or B, or both*
True	True	True
True	False	True
False	True	True
False	False	False

This meaning of disjunction is inclusive because it allows that if both *A* and *B* occur, the disjunction is true, as shown in the first row in the truth table. For an exclusive disjunction, *either A or else B, but not both*, the first row is false. The first premise of problem 1 corresponds to the truth table above. The second premise, however, states that the fault is not in the cable, *not-A*, and so this proposition eliminates the first two rows of the truth table. The only remaining row that is true is one in which the fault is in the software, *B*, and so the inference is valid: its conclusion is true in the only case in which the two premises are true. The same inference follows from the interpretation of the disjunction as exclusive. Of course, a complete account of the sentential calculus, which itself is only a part of logic, calls for much more than a rule for disjunction. But, the present account suffices to illustrate the difference between formal rules and semantics.

Psychologists have followed in the steps of logicians. At first, they developed theories depending on formal rules of inference, such as the rule for disjunction above (e.g., Rips, 1994). They argued that the difficulty of an inference depends on the number of steps in its proof, and so the preceding disjunctive inference should be easy because it can be carried out in a single step. But, they allowed that rules in the full set for logical inferences differed in the difficulty of

their application, and they had some success in fitting the theory to experimental data.

Could reasoning depend instead on a semantic approach akin to truth tables? At first sight, the answer seemed to be: no. The reason is illustrated in the example:

Problem 2: Either the fault is in the cable or it's in the software and it's in the printer. It's not in the cable. So, what follows?

The conclusion that the fault is in the software and it's in the printer is readily derived from the preceding formal rule given that this conjunction is mapped into *B*. In contrast, the truth table for the disjunction calls for eight rows, because each of its three simple propositions can be true or false. In general, each time a simple proposition is added to an inference the size of the corresponding truth table doubles. Yet, the inference does not double in difficulty, just as problem 2 is not twice as hard as problem 1.

During the last 30 years, an alternative semantic theory has been developed based on mental simulations (e.g., Byrne, 2005; Johnson-Laird, 1983, 2006). Its essence is that when people understand the meaning of an assertion, the interpretative system builds mental models of each of the main possibilities to which it refers. A conclusion that holds in all the models of the premises is necessary; a conclusion that holds in most of the models is probable – the theory extends to numerical probabilities, and a conclusion that holds in one of the models is at least possible. The inclusive disjunction in problem 1 has three mental models, which are shown in separate rows in this diagram:

fault in cable	
	fault in software
fault in cable	fault in software

Mental models are not made of words, which are used here for simplicity, but from representations of the world. As far as possible, these representations are 'iconic,' in the sense that their structure corresponds to the structure of what they represent. Most importantly, they represent only what is true unless explicit reference is made to what is false. One consequence is that the number of mental models does not double with each additional simple proposition in an inference, and so models are much more feasible than truth tables as a foundation for human reasoning.

Evidence tends to corroborate models as simulations on which naive reasoners rely, that is, those individuals who have not mastered logic. A simple prediction concerns the difference in reasoning between inclusive and exclusive disjunctions. The rule of inference above applies to both sorts of disjunction, and so formal theories do not predict that inferences based on them should differ in difficulty. In contrast, an inclusive disjunction has three mental models, whereas an exclusive disjunction has only two mental models. For example, the exclusive disjunction, *either the fault is in the cable or else it's in the software, but not both*, has only two mental models:

fault in cable	
	fault in software

Fewer models are less of a load on working memory, and so the theory predicts that problem 1 should be easier with exclusive disjunctions than with inclusive disjunctions. Experiments have corroborated this prediction (Johnson-Laird, 2006; Chapter 3).

The theory makes a more striking prediction: reasoners should make predictable errors in reasoning, because mental models represent only what is true. Such errors should occur when individuals evaluate whether several assertions can all be true at the same time, that is, whether they are consistent. This task relates to logic, because one method of proof is to add the negation of a conclusion to the premises and then to prove that the resulting set of assertions is inconsistent. Consider the following problem in which both disjunctions are exclusive:

Problem 3: Either the pie is on the table or else the cake is on the table. Either the pie isn't on the table or else the cake is on the table. Can both of these assertions be true at the same time?

Most individuals respond, 'yes,' because they assume that when the cake is on the table both assertions are true. After all, their respective mental models each has a model of this possibility. But, this compelling inference is an illusion. It occurs because models represent only what is true. If instead each model represents both what is true and what is false, then the first exclusive disjunction has these 'fully explicit' models of what is on the table:

pie	not-cake
not-pie	cake

This diagram uses 'not-cake' to represent that it is false that the cake is on the table. Likewise, the second exclusive disjunction has these fully explicit models of what is on the table:

not-pie	not-cake
pie	cake

For both disjunctions to be true at the same time, there needs to be a possibility that they have in common. But, as the fully explicit models show, no such possibility exists. Ergo, the two assertions cannot both be true at the same time.

The preceding illusory inference is just one of a whole swathe of different illusions depending on various sorts of connective. They occur both in drawing conclusions and in assessing consistency (for a review, see Johnson-Laird, 2006; Chapter 11). (A set of assertions is said to be consistent if, and only if, they can all be true at the same time.) These illusions are a litmus test for reasoners' use of mental models, because no other current theory of reasoning predicts their occurrence, and it is not easy to see how to extend theories based on formal rules of inference to predict them. At present, such theories rely on rules that are correct, and the introduction of incorrect rules, which would be needed to predict the illusions, would yield an inconsistent system. In the case of mental models, however, the illusions are

a consequence of their failure to represent what is false. For most inferences, this assumption is benign. Its surprising consequences emerge only for some inferences, and they were discovered in the output of a computer program implementing the theory.

Dual Systems of Reasoning

Many psychologists distinguish between a rapid intuitive system of reasoning and a slower deliberative system (e.g., Evans, 2007). Individuals differ in the degree to which they submit their intuitions to a deliberative test (see, e.g., Stanovich, 1999). The same dual-process account also applies to judgments (Kahneman, 2011). The model theory distinguishes between the computational power of the two systems: intuitions rely on a primitive system that can represent only a single mental model and that can carry out only a small finite number of operations, whereas deliberations depend on working memory, and so they can construct alternative models and carry out loops of operations on them until they exhaust the capacity of working memory. One sign of the use of intuitions occurs when people have to draw conclusions from pairs of premises, such as:

Problem 4: Some of the authors are bakers. Some of the bakers are caterers. What follows?

Inferences of this sort are known as syllogisms, and their logic was formulated first by Aristotle, the founder of Western logic. Twelve different psychological theories of syllogisms exist to explain systematic errors. One common error, for example, is to draw the conclusion for problem 4:

Some of the authors are caterers.

This conclusion might hold given the premises, but the inference isn't valid, because the bakers who are authors need not be the same bakers who are caterers. One appealing theory is that naïve individuals draw probable conclusions rather than valid conclusions (Oaksford and Chater, 2007). Other theories are based on formal rules of inference (Rips, 1994), and still others on various intuitive heuristics. A recent meta-analysis showed that none of these theories is satisfactory (Khemlani and Johnson-Laird, 2012). What is needed is a dual-process theory in which the intuitive system constructs a plausible model of a set of individuals, such as, for the premises above:

author	baker	caterer
author	baker	caterer
author		
	baker	
		caterer

Unlike the previous diagrams, this diagram represents a single mental model of five different individuals, and it embodies the heuristic that repeated terms, such as 'bakers,' refer to the same individuals. The model supports the common but erroneous conclusion that some of the authors are caterers. Those reasoners who deliberate about this conclusion may succeed in constructing an alternative model of the premises in which they separate the bakers who are authors from the bakers who are caterers, so the first two individuals in the preceding model become two individuals of this sort:

author	baker	
	baker	caterer

The resulting model refutes the intuitive conclusion and leads to the response that no definite conclusion about the relation between artists and caterers follows from the premises. This dual-process theory gives a satisfactory account of syllogistic reasoning.

Induction

Inductions are inferences that use knowledge or beliefs to go beyond the semantic information in the premises. Because knowledge and beliefs are fallible, inductions are invalid, though they can yield true conclusions. Some philosophers, such as Karl Popper, have argued that inductions play no role in science. But, clear cases to the contrary have occurred in astronomy, meteorology, and medicine. For example, Kepler's first law of planetary motion describes the orbits of the planets as ellipses with the sun at one of the two foci. He studied Tycho Brahe's observations of the positions of planets and induced his laws from them. Likewise, in daily life, everyone makes inductions.

A common misunderstanding, attributable to Aristotle, is that inductions lead from the particular to the general. But, they can also lead from one generalization to another; for example, A poodle's jaw is strong enough to bite through wire; therefore, a German shepherd's jaw is strong enough to bite through wire. Individuals who know that German shepherds are bigger and stronger than poodles are likely to accept this inference. Inductions can also lead to particular conclusions, for example, the master of *The Herald of Free Enterprise*, a car ferry, inferred that the bow doors had been closed when it left Zeebrugge in Belgium on March 6, 1987. In fact, they had not been closed, the sea rushed in, the vessel capsized, and over 100 people drowned. Inductions can be risky.

Here are two similar inferences:

Problem 5: If the client makes an order, then the goods are shipped. The goods are shipped. What follows?

Most people infer: The client made the order.

Problem 6: If the client makes an order, then the goods are shipped. The client makes an order. What follows?

Most people infer: The goods are shipped. A subtle difference between the two conclusions is that the first one is in the past tense, whereas the second one is in the present tense, which often refers to an event in the future, because English has no future tense. The difference in tense reflects the timing of the two events in the conditional: first, the client makes an order, and then the goods are shipped. Reasoners take the categorical premise in each inference to refer to the present, and so naturally an event that occurs

earlier is in the past, and an event that occurs later is in the future. They unconsciously choose the appropriate tense to refer to these timings.

The difference between the two inferences reflects what is known as 'modulation.' That is, the meanings of clauses, what they refer to, and knowledge can modulate the core meanings of sentential connectives. In logic, connectives have a fixed meaning; in language, they have core meanings corresponding to their meanings in logic, but modulation can alter these meanings. One effect is the temporal modulation that occurs in the preceding examples: knowledge can introduce various relations between events referred to in the clauses governed by the connective, and one such relation is a temporal order, which, in turn, can be reflected in a reasoner's choice of tense. Another effect is to block the construction of models of possibilities. Consider the disjunction: *Either she doesn't play a musical instrument or she doesn't play the piano*. The core meaning of an inclusive disjunction, *not-A or not-B*, refers to three possibilities: *not-A & not-B*, *not-A & B*, and *A & not-B*. But, the present disjunction refers only to two possibilities:

not-instrument	not-piano	(not-A & not-B)
instrument	not-piano	(A & not-B)

The model of the possibility:

not-instrument	piano	(not-A & B)

is blocked because if she plays the piano then of course she plays a musical instrument. In other words, the disjunction – rather unusually – yields the definite conclusion that she doesn't play the piano.

Knowledge and beliefs are fallible, and so the use of the present and past tenses in the earlier inferences is not deductively valid – the inferences are inductions. One corollary of modulation is that no general form of inference, such as *not-A or not-B*, *B*; therefore, *not-A*, is guaranteed to yield a valid deduction whatever the contents of *A* and *B*. Validity, therefore, has to be decided on a case-by-case basis. Inferences are valid not in virtue of their logical forms, which match rules of inference, but in virtue of their meanings. A variety of experiments have corroborated the occurrence of modulations that add relations and that block the construction of models of possibilities. Induction accordingly increases semantic information: it eliminates possibilities, either by adding elements to mental models or by blocking the construction of a mental model altogether.

Human reasoners have an inductive skill that far surpasses any known algorithm, whether based on mental models or formal rules. It is the ability to formulate explanations. If an induction yields a generalization or a scientific law, then scientists, in turn, may be able to explain it on the basis of more fundamental ideas. For instance, Einstein's general theory of relativity explained planetary orbits as a result of the sun's mass curving space–time. Following the nineteenth century logician, C.S. Peirce, inferences yielding hypotheses or explanations are known as 'abductions.'

Abduction

Abduction is a special case of induction. As an illustration consider the problem:

Problem 7: If a pilot falls from a plane without a parachute, the pilot dies. This pilot did not die, however. Why not?

In both the United States and East Asia, some individuals respond with a valid deduction: This pilot did not fall from a plane without a parachute. But, others respond with explanations, such as:

- The pilot fell into a deep snowdrift.
- The plane was on the ground.
- The pilot was already dead.

This propensity to formulate explanations is extraordinary. It seems to depend on mental simulations, but no one has been able to figure out quite how it is done. Humans can offer putative explanations for almost anything. The following problem is based on two sentences selected at random from separate stories also selected at random, but one word in the second sentence was changed – the opening noun was replaced with a pronoun referring back to the previous sentence:

Problem 8: Celia made her way to a shop that sold TV sets. She had recently had her ears pierced. What's going on?

Individuals readily provide such explanations as:

- Celia was getting reception in her earrings and wanted the TV shop to investigate.
- She wanted to see her new earrings on closed circuit TV.
- She had won a bet by having her ears pierced and was spending the money on a TV set.

Occasionally, reasoners are unable to formulate an explanation, and it happens more often when the two sentences do not refer to the same individual (see Johnson-Laird, 2006; Chapter 14).

Problem 9: If your friend has gone to get the car then she'll be back to pick you up in less than 10 min. Your friend has gone to get the car. But, she has not come back in 20 min. What do you infer?

Logic can tell you that there is an inconsistency, but it cannot tell you how to answer the question. And it does not imply that you should withdraw your conclusion that your friend will be back in 10 min even though she hasn't returned. Logic is 'monotonic,' that is, with each additional premise, further valid inferences can be made, and no additional premise ever calls for the withdrawal of a previous conclusion. In contrast, human reasoning is 'non-monotonic,' that is, when a beautiful conclusion collides with an ugly fact, individuals tend to retract the conclusion. Students of artificial intelligence have devised many systems of non-monotonic reasoning, but none of them offers a psychologically plausible answer to problem 9.

One answer going back to William James is that when individuals discover an inconsistency between the consequences of their beliefs and the facts, they should revise their beliefs in a minimal way to restore consistency. So, you should cease to believe that your friend has gone to get the car. But, if you found yourself in the circumstances of problem 9, you are much more likely to try to explain what has happened. Perhaps your friend had difficulty in starting the car, or perhaps she ran into a one-way system that forced her to make a long detour. If you can simulate a plausible explanation, then it will have the

side effect of leading you to abandon at least one of your beliefs. But, inconsistencies elicit explanations, not minimal changes to your beliefs, because explanations help you to decide what to do.

Evidence corroborates this account. Experiments have examined inconsistencies, such as:

Problem 10: If someone pulled the trigger, then the gun fired. Someone pulled the trigger, but the gun did not fire. Why not?

Participants responded with causal explanations that resolved the inconsistency, for example: *someone emptied the gun and there were no bullets in its chamber*. In a further study, participants assessed the probability of putative explanations for such inconsistencies. On each trial, they assessed the probabilities of a cause and its effect (as in the response above), the cause alone, the effect alone, and various control assertions. They ranked as the most probable explanation, a cause and its effect (Johnson-Laird, 2006). Because these conjunctions were ranked as more probable than their constituent propositions, the assessments violated the probability calculus – they are an instance of the 'conjunction' fallacy in which a conjunction is considered as more probable than one of its constituents (Kahneman, 2011). They show that the most plausible resolutions of inconsistencies are explanations, not minimal changes to beliefs.

Problem Solving

Problems are of many, many sorts. This article has already illustrated 10 different sorts of problems in reasoning. In fact, there are so many problems that, as the theory of computability shows, there are not enough solutions to go round, that is, many problems are provably insoluble (including the problem of determining whether or not a problem is soluble in principle). In science, mathematics, and technology, problems can be complex and unique, such as the problem of determining how cholera was transmitted from one person to another, the problem of breaking the Nazi Enigma code, and the problem of constructing a controllable heavier-than-air craft. A close examination of how individuals solved these particular problems shows that they relied on all three sorts of reasoning (Johnson-Laird, 2006). But, they also relied on other sorts of thinking, including analogies and creative insights (see Holyoak and Morrison, 2012).

Many sorts of problem are not unique but come in variations that each call for thought. To complete a tax form, for example, is a problem that most people have to solve over and over. When they first tackle such a problem, they may flail around trying one move and then another with only a few constraints, if any, governing their choices. With experience, however, they acquire a strategy for solving the class of problem, that is, a set of systematic constraints that guide their generation of elementary tactical moves leading to a solution.

A major advance in the psychology of problem solving occurred in Newell and Simon's theory, and its computer implementation (see Newell, 1990; Newell and Simon, 1972). In some cases, humans already know the strategy for solving a problem. Otherwise, they have to use weak strategies applicable to problems of many sorts. A natural strategy is to work forward from the initial state of a problem toward its solution, constrained by available knowledge. Another strategy, which goes back to Plato, is to start with the required solution and then to work backward toward the initial state. This strategy is known as 'means-ends' analysis because it depends on knowledge of the means available to achieve the ends. A problem solver assesses the difference between the current state of the solution and the required goal state. As long as there's a difference between the two, the solver searches for a move that reduces the difference. If no such move is available, the solver sets up a subgoal to find such a move, and means-ends analysis is applied recursively to this subgoal. The size of the space of possible sequences of moves – the 'problem space' – does not always predict the difficulty of a problem.

The means-ends strategy has been influential both in the study of how humans solve problems and in the development of hypotheses about the overall architecture of the mind (Newell, 1990). But, it is not applicable to certain problems, such as:

Problem 11: Given the following 6 squares, remove 5 pieces to leave exactly 3 squares with no loose ends, i.e., pieces that are not connected to any other piece at one or other end.

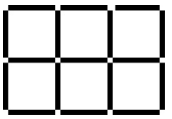

The goal is described, but not in the form of an explicit solution that allows a solver to compare it with the current state of the solution. What happens as individuals solve a series of such problems is that they gradually learn the different moves that can be made and deduce their consequences. One move is to remove the two pieces at right angles to each other that make up an external corner: the effect is to eliminate one square. Another move is to remove a single piece, such as the middle piece in the bottom row, which also eliminates one square. The removal of the two top corners and the piece in the middle of the bottom row solves the problem: five pieces are removed to leave three squares. With no experience of these problems, individuals tend to start out by trying out one move or another in a process of 'trial-and-error.' They generate a move and deduce its consequences. Even though the move may be futile and have to be undone later, they have learned something useful. A key assumption is that the problem-solving system can shift knowledge from the evaluative stage of the process to the generative stage. Hence, knowledge that individuals infer from deducing the consequences of a move can shift so that it starts to act as a local constraint on their generation of moves. With further experience, many individuals realize that a global constraint on each move in the problem is the current ratio of the number of pieces to be removed to the number of squares to be eliminated. They can then deduce the set of pertinent moves to be considered next, for example, if the ratio is much greater than one, then an appropriate move should, if possible, remove many pieces but only one square.

Another strategy in solving certain problems alternates between working forward and working backward. Individuals start working forward from the initial state and then switch to

working backward from the goal, especially if they get stuck. They alternate in this way until they either solve the problem or give up. A theoretical goal is to categorize all the feasible human strategies for solving problems. The difficulty is to take into account all possible sorts of problems (see Holyoak and Morrison, 2012; Part 4).

Insight

Sometimes when you are trying to solve a problem, you have a sudden insight and the solution pops into your mind seemingly from nowhere. When Archimedes was in the bath thinking about how to measure the volume of a king's crown, he suddenly realized – perhaps as he lay down and water slopped over the edge of the bath – that he could easily measure how much water the crown displaced. Legend has it that he cried, 'Eureka,' the Greek for 'I have it,' leapt from the bath, and ran naked down the street. When you have an insight, at one moment you are at an impasse unable to solve a problem, and at the next moment you are aware of the solution. Some sorts of problems depend on an insight for their solution whereas other sorts of problems do not. But, the occurrence of an insight is a subjective matter. It depends on whether an individual solved the problem, and on whether it was a result of a discontinuity in their thinking yielding a significant change of which they were aware.

Psychologists have studied problems whose solutions seem to depend on insights:

Problem 12: In the following simple addition, each letter has a unique value, and D = 5, so what are the values of the remaining letters?

```
  DONALD
+ GERALD
--------
  ROBERT
```

It follows at once that T = 0. Individuals can then try substituting different integers for letters using trial and error. But, this procedure soon leads nowhere: there are too many possibilities. As you study the problem, a sudden insight may hit you about the column: O + E = O. This pattern can occur only if E = 9, and there is a carry of 1 from the previous column. Such insights are essential to solve the problem.

Insight seems different from the normal process in which you make steady progress toward solving a problem. Skeptics doubt whether it depends on a special sort of thinking. If you search for the number that opens a combination lock, its discovery is bound to be sudden, but it does not depend on a process that differs from the one that hitherto failed. Experimental evidence, however, shows that insight is different (Holyoak and Morrison, 2012; Part 4). In routine problem solving, you have a sense that you are getting 'warmer' as you approach a solution. This sense does not occur prior to an insight. The difference is even evident in the regions of the brain that become active. So, what are the mental processes underlying an insight?

At least three hypotheses are plausible. The first is that individuals become fixated on an inappropriate method, but a period in which they think about other matters allows the misleading cues to wane, and so the insight – a switch to the correct cues – is more likely. Such perseverations certainly occur, but the problem with this hypothesis is its assumption that individuals know the correct method but can't access it. The claim may hold for some problems, but insights sometimes lead to the discovery of new methods. The second hypothesis, which is based on the idea of a problem space, is that insight depends on switching from a search in the current problem space to a search in the meta-level space of possible problem spaces for a new representation of the problem. In a case of insight, this new representation yields the solution to the problem. For example, most people are at first defeated by the problem of the 'mutilated' chessboard. The goal is to cover a chessboard with dominoes, or to prove that the task is impossible, but the board has been mutilated by the removal of a square at one corner and a square of the same color at the diagonally opposite corner. Individuals who start laying out dominoes on the board get nowhere. They have the wrong representation of the problem. The key insight depends on constructing a new representation: each domino covers one white square and one adjacent black square, but the diagonally opposite corners are of the same color. Hence, after they are removed, the remaining squares do not include equal numbers of black and white squares, and so the task is impossible.

A third hypothesis is that insight depends on a change in the constraints governing the problem-solving strategy. The current constraints yield an impasse. The abductive system registers the impasse, and so it sets up an attempt to modify the constraints. The process is initiated consciously, but it proceeds unconsciously. Modified constraints may, or may not, yield a solution. In the case of a false insight, you think that you have a solution, but, despite the change in constraints, you subsequently deduce that you were wrong. A genuine insight is accordingly an initial failure to solve a problem, followed by a change in constraints, followed by a solution. Once an insight has occurred, it may be used henceforth to constrain the generation of moves. A change in constraints is bound to be discontinuous, but it may be preceded by useless changes. A special case of a change in constraints yields a new representation of a problem. But, other insights depend merely on relaxing the existing constraints.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Decision and Choice: Heuristics; Deductive Reasoning Systems; Problem Solving and Reasoning: Case-Based.

Bibliography

Byrne, Ruth M.J., 2005. The Rational Imagination: How People Create Alternatives to, Reality. MIT Press, Cambridge, MA.

Evans, Jonathan St. B.T., 2007. Hypothetical Thinking: Dual Processes in Reasoning and, Judgement. Psychology Press, Hove, UK.

Holyoak, Keith, Morrison, Robert (Eds.), 2012. The Oxford Handbook of Thinking and Reasoning. Oxford University Press, New York.

Johnson-Laird, Philip N., 1983. Mental Models. Harvard University Press, Cambridge, MA.

Johnson-Laird, Philip N., 2006. How We Reason. Oxford University Press, Oxford, UK.
Kahneman, Daniel, 2011. Thinking Fast and Slow. Farrar, Strauss, & Giroux, New York.
Khemlani, Sangeet, Johnson-Laird, Philip N., 2012. Theories of the syllogism: a meta-analysis. Psychological Bulletin 138, 427–457.
Newell, Allen, 1990. Unified Theories of Cognition. Harvard University, Press, Cambridge, MA.
Newell, Allen, Simon, Herbert A., 1972. Human Problem Solving. Prentice Hall, Englewood Cliffs, NJ.
Oaksford, Michael, Chater, Nick, 2007. Bayesian Rationality: The Probabilistic Approach to Human Reasoning. Oxford University Press, Oxford.
Rips, Lance J., 1994. The Psychology of Proof. MIT Press, Cambridge, MA.
Stanovich, Keith E., 1999. Who Is Rational? Studies of Individual Differences in Reasoning. Erlbaum, Mahwah, NJ.

Relevant websites

http://mentalmodels.princeton.edu/ – Mental Models.
http://shelf1.library.cmu.edu/IMLS/MindModels/ – Mind Models of Newell and Simon.

Problem Solving during Infancy and Early Childhood, Development of

Hiromitsu Miyata, Okanoya Emotional Information Project (OEIP), Exploratory Research for Advanced Technology (ERATO), Japan Science and Technology Agency (JST), Nagoya, Japan; and Graduate School of Information Science, Nagoya University, Nagoya, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Problem solving forms an essential part of daily intellectual activities, not only for adults but also for infants and young children. Planning, or the internal processes that mediate an intended action, is a crucial cognitive component for efficient problem solving, with its earliest precursor observed in infants below 1 year of age. This article reviews representative empirical research in the field, which involves various behavioral tasks including tool use such as spoons, tube problems, and maze tasks. Mental processes underlying these behaviors develop from simple prospection of events to organizing different event sequences to face novel problems.

Problem solving undoubtedly forms an essential part of our lives, not only for everyday events but even for more global and political problems. Psychologists have asked how developing humans become capable of problem solving. Traditionally, Jean Piaget suggested a renowned four-stage model of logical thinking, beginning with the sensorimotor stage at 0–2 years before proceeding to the preoperational, concrete operational, and formal operational stages toward adolescence (Piaget, 1985). Piaget's stage-based models and research methods, which largely depended on the linguistic reports of children, came to be criticized with the advent of the cognitive approach in the 1970s (Fujimura, 2011). That is, researchers have come to place greater emphasis on the internal processes underlying problem solving, which develop in continuous ways while involving multiple, mutually related mental systems.

Based on such cognitive perspectives, efficient solution of complex problems should be mediated by sophisticated forms of planning, which is deemed as the internal process of formulating an organized method about one's future behavior (Friedman and Scholnick, 1997; see also Friedman et al., 1987). Planning ahead should be beneficial for human infants and children, because it should enhance efficient decision making when the agent encounters various physical and social events. Literature on developmental psychology has suggested that planning ahead in order to achieve a goal is present during the second year of life, with its earliest precursor starting during the first year. Planning-based problem solving is deemed especially important in that it essentially involves storing information in the brain from the outer world and then internally transforming and organizing multiple elements of future actions in an appropriate sequence (Miyata and Fujita, 2011). Such active representational systems are regarded as a hallmark of human thinking. Humans' uniqueness in this respect appears relevant despite the fact that comparative cognition studies have recently insisted that even some nonhuman species, including primates and birds, may possess planning capabilities at a certain limited level (for an overview, see Emery and Clayton, 2004; Miyata and Fujita, 2011).

This article reviews representative literature on empirical research relevant to problem-solving capabilities in infants and young children. It especially focuses on how modern, cognitive perspectives understand problem solving at different developmental stages. Evidence from each stage is considered in terms of the levels and complexity of the underlying internal cognitive processes. Also raised are the issues that require elaborated work in future, including metacognitive processes and neural underpinnings of problem solving.

Prospective Control in Infants

Prospective control of motor behavior is found in very young infants, even though this process should be distinguished from problem solving. Even for infants less than 1 year of age, object properties such as size and orientation elicit preparatory sizing of grip and adjustments in hand orientation (Lockman et al., 1984; Newell et al., 1989; von Hofsten and Ronnqvist, 1988). A number of studies also suggest that these infants may be using mental imagery to guide motor behavior. Clifton et al. (1991) presented small and large objects in the dark, which elicited a one-handed grasp and two-handed grasps, respectively, to 6-month-olds. Before the testing, participants had been shown these objects in the light, which were paired with unique sounds. The infants successfully chose the appropriate grip configuration for each object's size. McCarty et al. (2001a) presented 9-year-old infants with rods oriented either vertically or horizontally out of the infant's reach, before making the room dark and bringing the rod within reach of the participant. Infants exhibited appropriate hand orientation for these invisible rods, in comparable ways as when the rods were visible. Both these studies suggest that infants made motor adjustments on the basis of the memory for the object they had seen earlier.

Another example comes from a moving object paradigm. von Hofsten (1980) presented infants with a moving toy whose trajectory brought it within reach of participants. Analyzing how infants reached out to grasp the toy, von Hofsten (1980) argued that infants anticipated the future point in space at which they would intercept the toy. Not only motor behaviors but also kinematic measures have suggested use of mental imagery when reaching invisible objects. Clifton et al. (1994)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23074-2

presented 7-month-old infants with sounding objects in the dark and found that the moving speed of participants' hands decreased after approaching the object, in comparable ways as they do when reaching visible objects. This was suggested to reflect use of mental imagery regarding the object's location. These studies are suggestive, although not conclusive, that infants below 1 year of age may use some mental imagery about forthcoming events. However, these prospective behaviors are not likely to reflect problem solving guided by planning, which should involve active manipulation of internal representations to achieve a goal.

How do infants move from simple prospective behavior to planning-based performance? Planning in its simplest form should require at least two steps; that is, first reaching an object and then using the object that is in hand to achieve a certain goal. Motor adjustment during a goal-directed behavior would provide one key to this stage. Studies involving adults have contrasted reaching and grasping of an object when they performed subsequent divergent actions. Johnson-Frey et al. (2004) reported that approach kinematics in adults was influenced by the participants' intention to replace the object to a novel location, in contrast to simply holding and lifting the object. Marteniuk et al. (1987) showed that the adults' approach to an object was slower, had lower peak speed, and had a longer deceleration as the hand approached the object, when precise perceptual-motor action was subsequently required. Berthier et al. (1996) found that the same kinematic pattern was observed when approaching a small, delicate object, as opposed to a larger object. Do infants at a certain developmental stage show comparable goal-directed behavior? Claxton et al. (2003) presented 10-month-old infants with a ball and encouraged the participants to either throw it down or fit it into a hole. The latter case required a more precise and detailed action than the former. Both peak and average speed of the approach for the object was lower in the latter condition, suggesting that infants cognitively planned the subsequent action while approaching the ball by hand. Such motor adjustment could involve prospective control while looking beyond what is being seen and considering what to do after grasping the object, and thus could be a precursor of complex problem solving that appears at later stages.

Formulating Sequential Actions

Toddlers in the second year of life come to perform fairly complex, sequential actions. For example, they perform on simple maze and puzzle tasks (Miyata et al., 2009). At around 18–24 months of age, they come to build block towers (Bayley, 1969). At the beginning, they can balance just two or three blocks, but soon the number of blocks increase to five or even more to form a tall tower. Building a tall block tower requires precise sequential movements of picking up and stacking pieces of blocks. That is, it should involve cognitive abilities such as planning sequential actions, anticipating in advance the outcomes of these actions, as well as having a representation of the correct goal state. It also requires sufficient motor skill to put successive blocks on top of the tower. Parents are often impressed by both the cognitive capacities and motor skills achieved by the toddlers at this stage.

Chen et al. (2010) investigated individual differences in the motor skills when toddlers between 18 and 21 months of age built block towers, by examining the kinematic measures of their movements. They asked whether high-tower builders show speed profile when approaching the tower and releasing the blocks different from that of the low-tower builders. Toddlers were presented with two tasks: one was the precision task of performing tower building, while the other was the imprecise task of 'cleaning up' the blocks by placing them into a wide, open container. Point of peak velocity, which specifies the point during the movement at which the hand reached its fastest speed, was used as a kinematic measure showing individual differences in motor skills. Toddlers who built high towers reached peak speed earlier in their movements compared to those who built low towers. These patterns of reaching peak velocity early in the movements were comparable to data from adults (Berthier et al., 1996; Johonson-Frey et al., 2004). Retesting the same children when they are around 3 years old no longer replicated comparable individual differences, when all children had sufficiently mastered the task. These findings by Chen et al. (2010) appear important in that toddlers below 2 years of age show good capabilities for complex sequential planning and motor skills. However, the task still does not require hierarchical organization of multiple components or movements, as observed in the tool use contexts.

Development of Action Planning

Friedman and Scholnick (1997) described the sequence of problem solving as first representing the problem, then devising and carrying out a solution, monitoring its effectiveness, and finally making necessary corrections. This organized schema is true for young children's behaviors when learning to use a tool, such as picking up a spoon. McCarty et al. (1999) identified three grips when 9- to 19-month-old children grasped and used spoons and other toys. A radial grip occurred when the handle was grasped with the thumb toward the goal end of the item (i.e., the bowl of the spoon). An ulnar grip occurred if the handle was grasped with the thumb away from the goal end of the item. A goal-end grip occurred if the end of the item was directly grasped. The ulnar grip and the goal-end grip both indicate lack of planning, and the radial grip was deemed to be the only correct and efficient strategy suggesting action planning. Nine-month-olds persisted in primarily using their preferred hand, disregarding the orientation of the item. This resulted in the awkward ulnar or goal-end grips when the spoon was in the difficult orientation. The infants often incorrectly placed the handle of the spoon directly in the mouth. This tendency to reach with the preferred hand was also observed in the 14-month-olds. However, they made corrections when necessary, so that they placed only the bowl of the spoon in the mouth and never its handle. The 19-month-olds tended to inhibit reaching with their preferred hand. They alternated hands on the basis of the orientation of the spoon, reaching with their handle-side hand to achieve a radial grip. In this way, food could be transported efficiently and smoothly to the mouth.

Based on these findings, McCarty et al. (1999) proposed a model suggesting four action planning strategies used by children learning to pick up a spoon. The first is called the 'feedback-based strategy,' in which the child simply follows his or her own preferences until a situation occurs where the child can recognize that the goal has been achieved. The second is the 'partially planned strategy,' in which the child plans the transport of the spoon to the mouth, but does not control for the grasping strategies. The third is termed the 'fully planned strategy,' in which the orientation of the spoon is first taken into account, so that the child can grasp the handle of the spoon with an efficient, radial grip. The final stage is referred to as the 'habitual solution,' where grasping of the spoon with a radial grip has become a heuristic and no longer requires solution of a problem on every trial.

Interestingly, this developmental trend toward an efficient grip was observed only for the spoons and not for the other toys, such as rattles and bells. One interpretation could be that children may have recognized the consequence of their planned actions. That is, correct use of a spoon resulted in having food as reward, whereas the goals of grasping a toy seemed more diverse and undetermined. Another reason could be that the spoon had great familiarity and involved a self-directed, as opposed to other-directed, action. McCarty et al. (2001b) addressed the latter hypothesis by comparing the use of radial grips by 9- to 24-month-olds between different tools. Specifically, the hammer and the magnet involved other-directed actions, whereas the familiar spoon and the less familiar hairbrush could be either self-directed or other-directed. Results clearly revealed that, for both familiar and unfamiliar items, self-directed actions elicited more correct, radial grips than did the other-directed ones. Age difference was also found in that the older groups tended to use the radial grips more frequently than the younger groups. Claxton et al. (2009) further used a waterwheel task that involved both self- and other-directed actions, in which the salience of the outcomes between a radial grip and an ulnar grip was quite enhanced. Again, the self–other difference in planning the correct action was observed. These data suggest that self-directedness of the actions still plays a crucial role in terms of planning an efficient action strategy at the end of the second year.

To what extent does the progress to higher levels of action organization as observed for spoons effectively generalize to other situations? Chen and Siegler (2000) found that children around 2.5 years of age were able to select an appropriate rake among several ones differing in length, after which they successfully transferred this knowledge to new rakes when superficial aspects of their physical appearances were altered. On the other hand, Metevier (2006) showed that 36-month-old children had great difficulty in coordinating two different tasks in an appropriate sequence. That is, most children failed to use a rake to retrieve a rod and then to use that rod to expel the toy from a tube, even though they performed well on these separate problems. These data suggest that the efficiency with which children around 3 years of age can transform problem-solving strategies obtained in one task to wider domains differs greatly depending on factors such as self-directedness, task complexity, and so on.

Encouraging children's Problem Solving

An effective way to enhance problem solving for young children could be to give verbal or other instructions to encourage the use of efficient strategies. Adults are well capable of effectively using visual images to help solve a number of problems (Kosslyn, 1975; Wohldmann et al., 2007). Are comparable visualization techniques available for young children? Rieser et al. (1994) found that 3-year-old children were capable of using visual imagery to imagine spatial locations of objects in another room. Richards and Sanderson (1999) showed that children 3 years of age or older effectively used imagery when solving syllogistic reasoning problems. A more recent example comes from a 'gravity error' paradigm. In the original task, Hood (1995) used a tube problem in which children were asked to find a ball that was dropped into one of the three tubes. The tubes were all opaque and were sometimes interwoven to form a visuospatial maze. Three-year-old children made errors by consistently searching in the location directly beneath the point at which the ball was dropped. This was termed a gravity error, reflecting the children's erroneous assumption that an object should drop straight down from where they saw it disappear. Hood (1995) found that the gravity error was persistent even after attempting to train 3-year-olds on how to solve the problem. Joh et al. (2011) further devised verbal instructions when 3-year-olds solved comparable tube problems, by encouraging children to imagine the ball rolling down the opaque tube. Children given this instruction performed better than the control participants, who just waited without imagining the movements of the ball. This suggests that children used the experimenter's explicit encouragement by instructions to imagine the paths of the ball.

More direct instructions on how to solve the problem could also be helpful for children of these ages. Bascandziev and Harris (2010) used a tube paradigm comparable to Hood (1995), in which the experimenter gave verbal instructions on how the tube constrained the movement of the ball. Children were instructed to follow the tube with their eyes to its end, so that they could determine where the ball would reappear. Three-year-old children, although not 2-year-olds, successfully benefited from this instruction to improve searching performance. Even though eye movements of the children were not recorded, it appears plausible that 3-year-olds effectively scanned the apparatus with their eyes during the invisible displacement. The results thus suggest that these preschoolers are capable of using visual imagery when solving physical problems and that the use of more sophisticated visualization strategies could be enhanced by adults giving effective testimony or encouragement. These emerging mental capacities should be the crucial keys that later leads to more complex planning and problem solving during childhood, which is to be integrated into more abstract representational systems and executive functions.

Higher Stages of Complex Problem Solving

Internal processes underlying problem solving during the preschool ages move from simple action or event planning to more organized and structured mental strategies to face the

complex world. Hudson and Fivush (1991) proposed four levels in the development of event planning at this stage. Level 0: simple anticipation of a sequence of actions for familiar events; level 1: single goal event planning, or the process of mentally considering different options for the sequence of actions to achieve a single goal; level 2: multiple event goal planning, or taking into account alternatives for sequential actions having two or more goals in mind; level 3: coordinated event planning, or mentally decomposing event sequences into multiple subgoals and hierarchically organizing subgoals from different events; level 4: novel event planning, or decomposing event sequences into multiple actions and subgoals, then modifying and joining these components to construct novel action sequences. These different levels imply that children become better aware of the relationships between actions and goals, while using the goal state to construct more complex plans.

Hudson et al. (1995) examined the ability of 3-, 4-, and 5-year-old children to construct plans about familiar events such as going to the beach and going grocery shopping, by analyzing the verbal reports of participants. They showed that these children are capable of constructing advance plans, that is, representation of an expected sequence of actions planned in advance of an action. The researchers further found that the ability to construct more complex plans – such as remedy plans, that is, plans to remedy unexpected mishaps that may occur in the process of carrying out advance plans, and prevention plans, that is, advance plans that include actions intended to prevent mishaps from occurring – develop with age. Hudson et al. (1995) suggested that 3- and 4-year-olds' plans would have reflected level 1: single goal event planning, such that they could at most use knowledge about familiar events to anticipate event sequences. By contrast, 5-year-olds were suggested to exhibit level 2: multiple event goal planning, and level 3: coordinated event planning, such that these children could now flexibly adapt their plans to achieve novel goals.

Can these developmental courses be observed not only for verbal problems but also for nonverbal cognitive tasks? To consider this issue, maze tasks should provide one fine example. Detour or maze problems have been used frequently in development psychology. Piaget (1954) traditionally argued that infants toward the end of the sensorimotor stage become capable of taking indirect routes that go beyond direct perception of the disappeared object during three-dimensional detour navigation. Lockman (1984) showed that infants made reaching detours before corresponding locomotor detours, and that they were less likely to make detours around transparent than opaque obstacle barriers. Diamond (1990) suggested that infants in the latter half of their first year start to employ detour reaches to retrieve objects placed in a transparent box with an opening to one side.

What about maze navigation in young children? Miyata et al. (2009) examined the performance of 3- to 4-year-old children on computerized maze problems by using a touch-sensitive screen. These mazes were based on a navigation task originally developed to test pigeons' (*Columba livia*) maze-solving performance and planning (Miyata et al., 2006), to move a target square to a goal square by making sequential responses on the touch monitor with fingers. In the first experiment, children around the age of 3 years were proficient at making a detour to reach the goal while avoiding an L-shaped barrier. Participants took direct routes more frequently than indirect routes. In the second experiment, a plus-shaped maze task was used, and the goal occasionally jumped to the end of another corner when the target arrived at the center of the maze. Parallel to pigeons (Miyata and Fujita, 2008), the 3- and 4-year-old children exhibited both correct responses toward a new goal and errors toward the old goal location immediately after the goal shift, with reaction times longer when they made correct responses than when they made errors. In addition, 3-year-olds had greater difficulty immediately adjusting their previously planned strategies after the change of goal locations, whereas 4-year-olds more frequently adjusted their response toward a new goal. This seems consistent with the use of higher level plans such as remedy plans and prevention plans as the children become older, testified by using a rigid nonverbal behavioral paradigm. More recently, Miyata and colleagues modified the computer-assisted navigation task so that it required 3- to 5-year-old children to visit two or three identical goals in a sequence, a task that represented a traveling salesperson problem (unpublished data). Preliminary analyses suggested that older children used more efficient route selection strategies by shortening the overall lengths of the traveling paths, consistently with the development in the levels of planning.

Future Perspectives

We have so far seen how primitive forms of prospective control during infancy subsequently lead to action planning, which then becomes more sequential and organized during early childhood to attain flexible problem-solving performance. Modern cognitive perspectives collectively have drawn decent sketch on the developmental course of mental systems underlying problem solving, whereas future research may involve a number of promising issues that still remain to be explored. One topic to be covered concerns the development of metacognition involved in problem solving. Frese and Stewart (1984), for example, regarded metacognitive planning as the most sophisticated among different levels of planning, in terms of the degree of abstraction. According to Frese and Stewart (1984), a metacognitive plan entails the person's knowledge of general problem-solving strategies, which should enable the person to generate novel plans in any area. Despite the assumed large adaptive value of this capability when facing novel situations, little empirical work so far has addressed its developmental course. For example, before starting to solve a problem, one could determine whether or not to start the plan if one has metacognition on whether he or she knows the solution to it. This could be examined by analyzing whether hint-seeking behavior is more frequent before solving difficult problems than before solving easy ones (Miyata and Fujita, 2011). Not just verbal reports by children but also more systematic, computerized settings could be effectively used to examine nonverbal metacognitive processes.

Another important perspective for future work concerns the neural correlates of problem solving in infants and young children, the issue relevant to the emerging field of cognitive developmental neuroscience. A number of neuroimaging techniques including functional magnetic resonance imaging and positron emission tomography, which require immobilization

during testing, have been difficult to be applied to infants and young children. However, recent progress in near-infrared spectroscopy (NIRS) has enabled measurement of brain activities in infants younger than 1 year (Minagawa-Kawai et al., 2008). NIRS noninvasively measures changes in hemoglobin concentration and its oxygenation level accompanying regional brain activities, by emitting and detecting continuous near-infrared lasers with two wavelengths. It also enables measurement of cortical activities in live situations of problem solving. For example, Miyata et al. (2011) applied the plus-shaped maze task comparable to Miyata et al. (2009) to adult humans using NIRS. In addition to large prefrontal activities during maze performance, right inferior prefrontal cortex also showed significant activations after making error responses, suggesting use of additional cognitive resources for adjustment of movements. Application of these behavioral paradigms to infants and young children using NIRS should promisingly uncover differentiated brain activities associated with different stages of planning and problem solving. Efforts inspired by these emerging perspectives should help to draw novel and more sophisticated pictures on the developmental course of problem solving, in a way that could be effectively connected to a broader literature on humans' cognitive development.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Conscious Control During Childhood, Development of; Executive Functions During Childhood, Development of; Motor Skills: Development in Infancy and Early Childhood; Piaget's Theory of Cognitive Development; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Bascandziev, I., Harris, P.L., 2010. The role of testimony in young children's solution of a gravity-driven invisible displacement task. Cognitive Development 25, 233–246.

Bayley, N., 1969. Manual for the Bayley Scales of Infant Development. The Psychological Corporation, New York, USA.

Berthier, N.E., Clifton, R.K., Gullipalli, V., McCall, D.D., Robin, D.J., 1996. Visual information and object size in the control of reaching. Journal of Motor Behavior 28, 187–197.

Chen, Y., Keen, R., Rosander, K., von Hofsten, C., 2010. Movement planning reflects skill level and age change in toddlers. Child Development 81, 1846–1858.

Chen, Z., Siegler, R., 2000. Across the great divide: bridging the gap between understanding of toddlers' and other children's thinking. Monographs of the Society for Research in Child Development 65 (i–vii), 1–96.

Claxton, L.J., Keen, R., McCarty, M.E., 2003. Evidence of motor planning in infant reaching behavior. Psychological Science 14, 354–356.

Claxton, L.J., McCarty, M.E., Keen, R., 2009. Self-directed action affects planning in tool-use tasks with toddlers. Infant Behavior and Development 32, 230–233.

Clifton, R.K., Rochat, P., Litovsky, R.Y., Perris, E.E., 1991. Object representation guides infants' reaching in the dark. Journal of Experimental Psychology: Human Perception and Performance 17, 323–329.

Clifton, R.K., Rochat, P., Robin, D.J., Berthier, N.E., 1994. Multimodal perception in the control of infant reaching. Journal of Experimental Psychology: Human Perception and Performance 20, 876–886.

Diamond, A., 1990. Developmental time course in human infants and infant monkeys, and the neural bases of inhibitory control in reaching. In: Diamond, A. (Ed.), The Development and Neural Bases of Higher Cognitive Functions. New York Academy of Sciences, New York, USA, pp. 637–669.

Emery, N., Clayton, N., 2004. Comparing the complex cognition of birds and primates. In: Rogers, L., Kaplan, G. (Eds.), Comparative Vertebrate Cognition: Are Primates Superior to Non-primates? Kluwer Academic/Plenum Publishers, New York, USA, pp. 3–55.

Frese, M., Stewart, J., 1984. Skill-learning as a concept in life-span developmental psychology: an action-theoretic approach. Human Development 27, 145–162.

Friedman, S.L., Scholnick, E.K., 1997. The Developmental Psychology of Planning: Why, How, and when Do We Plan? Erlbaum, Hillsdale, NJ.

Friedman, S.L., Scholnick, E.K., Cocking, R.R., 1987. Blueprints for Thinking: The Role of Planning in Cognitive Development. Cambridge University Press, Cambridge, England.

Fujimura, N., 2011. Ronri-teki sikou no hattatsu (Development of logical thinking). In: Ujiie, T., Chin, S. (Eds.), Hattatsu-shinrigaku Gairon (An Introduction to Developmental Psychology). NHK Publishing Co., Ltd, Tokyo, Japan.

Hood, B.M., 1995. Gravity rules for 2- to 4-year-olds? Cognitive Development 10, 577–598.

Hudson, J.A., Fivush, R., 1991. Planning in the preschool years: the emergence of plans from general event knowledge. Cognitive Development 6, 393–415.

Hudson, J.A., Shapiro, L.R., Sosa, B.B., 1995. Planning in the real world: preschool children's scripts and plans for familiar events. Child Development 4, 984–998.

Joh, A., Jaswal, V., Keen, R., 2011. Imagining a way out of the gravity bias: preschoolers can visualize the solution to a spatial problem. Child Development 82, 744–750.

Johnson-Frey, S.H., McCarty, M.E., Keen, R., 2004. Reaching beyond spatial perception: effects of intended future actions on visually guided prehension. Visual Cognition 11, 371–399.

Kosslyn, S.M., 1975. Information representation in visual images. Cognitive Psychology 7, 341–370.

Lockman, J.J., 1984. The development of detour ability during infancy. Child Development 55, 482–491.

Lockman, J.J., Ashmead, D.H., Bushnell, E.W., 1984. The development of anticipatory hand orientation during infancy. Journal of Experimental Child Psychology 37, 176–186.

Marteniuk, R.G., MacKenzie, C.L., Jeannerod, M., Athenes, A., Dugas, C., 1987. Constraints on human arm movement trajectories. Canadian Journal of Experimental Psychology 41, 365–378.

McCarty, M.E., Clifton, R.K., Collard, R.R., 1999. Problem solving in infancy: the emergence of an action plan. Developmental Psychology 35, 1091–1101.

McCarty, M.E., Clifton, R.K., Ashmead, D.H., Lee, P., Goubet, N., 2001a. How infants use vision for grasping objects. Child Development 72, 973–987.

McCarty, M.E., Clifton, R.K., Collard, R.R., 2001b. The beginnings of tool use by infants and toddlers. Infancy 2, 233–256.

Metevier, C.M., 2006. Tool-Using in Rhesus Monkeys and 36-Month-Old Children: Acquisition, Comprehension, and Individual Differences, Ph.D. Thesis, The University of Massachusetts, Amherst, USA.

Minagawa-Kawai, Y., Mori, K., Hebden, J., Dupoux, E., 2008. Optical imaging of infants' neurocognitive development: recent advances and prospects. Developmental Neurobiology 68, 712–728.

Miyata, H., Fujita, K., 2008. Pigeons (*Columba livia*) plan future moves on computerized maze tasks. Animal Cognition 11, 505–516.

Miyata, H., Fujita, K., 2011. Planning abilities in non-human animals: new findings in primate and birds. The Japanese Journal of Animal Psychology 61, 69–82 (in Japanese).

Miyata, H., Itakura, S., Fujita, K., 2009. Planning in human children (*Homo sapiens*) assessed by maze problems on the touch screen. Journal of Comparative Psychology 123, 69–78.

Miyata, H., Ushitani, T., Adachi, I., Fujita, K., 2006. Performance of pigeons (*Columba livia*) on maze problems presented on the LCD screen: in search for preplanning ability in an avian species. Journal of Comparative Psychology 120, 358–366.

Miyata, H., Watanabe, S., Minagawa-Kawai, Y., 2011. Two successive neurocognitive processes captured by near-infrared spectroscopy: prefrontal activation during a computerized plus-shaped maze task. Brain Research 1734, 90–99.

Newell, K.M., Scully, D.M., McDonald, P.V., Baillargeon, R., 1989. Task constraints and infant grip configurations. Developmental Psychobiology 22, 817–831.

Piaget, J., 1954. The Construction of Reality in the Child. Free Press, New York, USA.

Piaget, J., 1985. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. University of Chicago Press, Chicago, USA.

Richards, C.A., Sanderson, J.A., 1999. The role of imagination in facilitating deductive reasoning in 2-, 3-, and 4-year-olds. Cognition 72, B1–B9.

Rieser, J.J., Garing, A.E., Young, M.F., 1994. Imagery, action, and young children's spatial orientation: It's not being there that counts, it's what one has in mind. Child Development 65, 1262–1278.

von Hofsten, C., 1980. Predictive reaching for moving objects by human infants. Journal of Experimental Child Psychology 30, 369–382.

von Hofsten, C., Ronnqvist, L., 1988. Preparation for grasping an object: a developmental study. Journal of Experimental Psychology: Human Perception and Performance 14, 610–621.

Wohldmann, E.L., Healy, A.F., Bourne, L.E., 2007. Pushing the limits of imagination: mental practice for learning sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition 33, 254–261.

Procrastination

Piers Steel, Haskayne School of Business, University of Calgary, Calgary, AB, Canada
Katrin B Klingsieck, Faculty for Arts and Humanities, University of Paderborn, Paderborn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Procrastination is a pervasive and pathological delay, where we put off despite expecting to be worse off. It is associated with other forms of self-regulatory failure, rooted in our neurobiology, though can be exacerbated by environmental features. Though several effective treatments for it are available, procrastination is self-perpetuating as procrastinators will often focus on emotional coping strategies, putting off dealing with the source of their delays.

Part 1. Definition and Types of Procrastination

Terms and references that can be ascribed to procrastination occur throughout the historical record, dating back to ancient Egypt's middle of the 18th dynasty, under the reign of Thutmosis III (i.e., "Friend, stop putting off work and allow us to go home in good time"). It appears in most major religious texts, including Islam, Hinduism, Buddhism, and Christianity. It is also increasingly common, with approximately 25% of the modern population confessing that they typically procrastinate. Given its prevalence, it is unsurprising that the term has been used in various ways to describe different forms of delay. However, among all these descriptions, common themes do emerge.

Definitions of Procrastination

Procrastination is typically taken as an irrational or a self-defeating delay, to be worse off for putting off. After a historical review of definitional efforts, Steel (2007) integrates previously established key components to produce, "to voluntarily delay an intended course of action despite expecting to be worse off for the delay." Further building on this effort, Klingsieck (2013a) refines this definition to, "the voluntary delay of an intended and necessary and/or [personally] important activity, despite expecting potential negative 3 procrastinated, it should be: i) intended, ii) voluntarily delayed, iii) foreseeably pathological." However, there is tension among all these components as they all can be considered continuous dimensions. An 'intended' task may range from a flippant notion to a proclaimed goal. Circumstances may make goal pursuit easy, meaning any delay is clearly voluntarily, to extremely difficult verging on the impossible, making delay effectively involuntary. The problematic consequences of the delay can range from the obvious to the unapparent in both likelihood and magnitude. Though there are some delays that are clearly procrastination and some that are not, it can be debated when one becomes the other.

Differences between Procrastination and Strategic Delay

Procrastination requires delay but is not equivalent to delay. However, at times the notion of 'active' or 'positive' procrastination will arise, referring to a strategic self-serving delay (Chu and Choi, 2005). This is not compatible with the dominant use of the term or procrastination's strong association with self-regulatory failure. However, there are forms of positive or purposeful delay that superficially appear similar to procrastination. In particular, purposely leaving a manageable portion of a project until near the deadline, taking advantage of increased motivation that occurs then, can be beneficial. In contrast, involuntarily leaving excessive amounts of work until just before the deadline because one cannot find motivation earlier typically is not. In the former, the person can choose to work earlier but prefers not to. In the latter, the person may prefer to work earlier but cannot find the motivation. Helping to differentiate the two is the question "How useful would it be for you to have more motivation earlier, well before the deadline?." While everyone might find more motivation desirable, for the procrastinator this will be especially so.

Types of Procrastination

Typologies of procrastination can be organized by life domain (e.g., academic or health), by self-reported reason or justification (e.g., rebellion or arousal), and by cause (e.g., self-efficacy or value driven). Since procrastination appears to require some sort of impulsive weakness of the will, where we act against our own best interests, it is facilitated by proximity to temptations or environmental cues. Personality traits are predictive of what temptations procrastinators tend to be susceptible to, resulting in different phenomenologies of procrastination (Schouwenburg, 2004). People high in the personality trait of extraversion, for example, will ascribe their procrastination to their need to socialize while conversely those low in extraversion (i.e., introverts) will attribute their procrastination to their need for solitude. Most studied among these is the neurotic or perfectionist procrastinator, who is most likely to seek clinical treatment and has given rise to the belief that perfectionism causes procrastination in general. Though there is almost no association between perfectionism and procrastination, impulsive perfectionists should personally experience their perfectionism as being the cause of their procrastination and this particular typology may respond well to techniques for managing anxiety.

Demographic Differences

Though procrastination is found within all societies and all time periods, there are areas of prevalence (Steel and Ferrari, 2013). To begin with, it is associated with a modernized society as greater access to temptation increases the incident rates of procrastination. Sensibly, the easier and more enjoyable it is to put off tasks, the more people procrastinate. Cultural values also may influence the rate of procrastination among nations, with Ireland reporting the most procrastination among English-speaking countries. Procrastination, being self-defeating, is associated with lower levels of income and education and less relationship success. In short, procrastinators are more likely to be unemployed, less educated, and single. Finally, procrastination is seen more in men than women and in the younger rather than older. Notably, sex differences in procrastination partially explain women's rise both academically and in the workforce.

Part 2. Measurement of Procrastination

As per the definition of procrastination, there are certain inherent problems in identifying when procrastination occurs. It may not be entirely clear to the observer or even the individual. The degree of intentionality of the task is not always evident and possibly malleable in retrospect. Obstacles may intervene hindering speedy completion of the task that are as seen insurmountable by the individual but only inconvenient by an observer. Also, the consequences of the delay may not be evident, allowing the overly optimistic, who truly believes there is plenty of time left, to appear to be a procrastinator to a dispassionate observer. Finally, standards and desires differ so the exact same consequences for delay can be procrastination by one, who is worse off, but not for another (Silver and Sabini, 1981).

Self-Report

Procrastination, like personality measures, can be measured as an overall general trait or by specific life domain. Typically, improved measurement can be achieved by using as specific a measure as the context will allow. By 2006, there have been at least 10 procrastination measures used in research (Steel, 2007), with several more developed since.

General Measures

Among the most popular of the general measures of procrastination are the General Procrastination Scale (GPS), the Adult Inventory of Procrastination (AIP), Tuckman Procrastination Scale, and the Decisional Procrastination Scale (DPS). Notably, the first three of this list is often characterized as representing behavioral procrastination, the delay of a task, while the last of these represents decisional procrastination, the delay of choice. While there is support for goal choice and goal pursuit to represent distinct stages, the DPS does show strong associations with the other behavioral measures and even contains an item "Even after I make a decision I delay acting upon it." It is questionable whether the decision procrastination, as measured, is a separate construct.

More recently, there appears to be criterion contamination problems among many of these measures (Steel, 2010). They assess procrastination by not only asking if the person delay irrational (e.g., "Putting off until the last minute has cost me money in the past"), but also if the person is prompt (e.g., "I generally return phone calls promptly") or if the person simply delays (e.g., "I find myself running later than I would like to be"). While the last two components are related to procrastination, they are not procrastination itself. Failing to be prompt is not the same as procrastinating and simply being busy or overcommitted can cause delay.

In response, the Pure Procrastination Scale has been developed that focuses purely on irrational delays, comprised of items from the DPS, the GPS, and the AIP that load together and reflect the traditional definition of procrastination (Steel, 2010). It shows improved associations with key outcome variables, such as subjective well-being, and proves to be functionally equivalent to the Irrational Procrastination Scale, which is designed to assess procrastination as an irrational delay.

Specific Measures

With life domains ranging from the marital to the financial, there are numerous potential ways to focus procrastination. Though some other specific measures have been used, such as procrastination regarding household safety, only one area has proven dominant: academic. In particular, Solomon and Rothblum's (1984) seminal measure, the Procrastination Assessment Scale – Student is often employed in this context, though somewhat basic in construction. It asks students to rate 'school activities in general' as well as five specific tasks (e.g., writing a term paper, keeping up with weekly readings) in term of how much they procrastinate, whether their procrastination is a problem, and do they want to stop procrastinating. In addition, there is a separate 'reasons for procrastination scale,' which provides 26 possible explanations for student's procrastination.

Given that procrastination is equally or even more of concern in many other life domains, including financial, such as failing to put enough aside for retirement, and health, such as failing to seek treatment, more specific measures for other domains should be entertained.

Observed

Difficulties in observing procrastination have not prevented several attempts. To the extent that the outcome of delay is clearly foreseeable, adverse, and there is task commitment, observed measures of procrastination are likely valid, especially in the aggregate. Like self-report measures, academia has proven to be of particular interest.

Academic Measures

Difficulties in observing procrastination are assuaged somewhat in an academic setting. Goals can be assigned, such as quizzes or essays, ensuring that a majority of students have a degree of intention. Problems of procrastination are well known in this setting and early starts are repeatedly counseled. The effect of delay is foreseeable and adverse, at least in many cases. Consequently, academic self-reports of procrastination and observed delays in this context are substantively associated (Steel et al., 2001). Specific

assignments typically used include the following: class attendance, course grade, quiz grade, and meeting a deadline for an assignment. Other less common academic aspects assessed include getting a PhD, publication rate, getting tenure, and timing of experimental participation. In particular, a computerized personal system of instruction, often now used in massive open online courses, provides an almost ideal venue to study academic procrastination. The pace of assignment completion, which is among the best inconspicuous measures of observed procrastination, can be automatically assessed for an extremely large number of people. Using this as a measurement foundation, a comprehensive set of interventions and diagnostics can be explored.

Other Measures

Other attempts to assess procrastination behavioral are far more limited. Interesting examples include when Christmas shopping is done, a gift receipt or check is cashed, or a letter is mailed. To some extent, an open critique is that the delay may not be irrational but due to the participant simply being busy or uninterested. One exception to this is by Reuben et al. (2007), who gave MBA students the option of either cashing a smaller check now or a larger check later. Though most students choose the smaller but immediate check, they did not cash it until after they would have received the larger but later one.

Looking forward, more success may be achieved by creating observed or behavioral measure at the group, neighborhood, or nation level. It would be constructed by identifying common, foreseeable, and almost universally beneficial tasks required for most people. For example, this may include the degree they pay their bills on time, the number of canceled dentist appointments, the degree of preparation for retirement, or the ratio of gym memberships to gym attendance. While individual exceptions to these self-regulatory tasks may occur, especially if there are financial constraints, in aggregate it would reveal if a community was indeed prone to procrastination.

Self-Deception

Because procrastination has ambiguous aspects, it leaves the possibility for self-deception. We can retrospectively recast our behavior as unintentional (e.g., "I wasn't really planning to do that anyway"), as unforeseeable (e.g., "I thought I had enough time!"), or as beneficial (e.g., "It is probably best I didn't get that done"). This can be considered an emotional coping strategy, where people look for explanations that are pleasing or self-serving (Sirois and Pychyl, 2013). This strategy is common in impulse-related disorders, where people sacrifice larger but later benefits for short-term rewards. Consequently, procrastinators may tend to congregate and compare themselves to other procrastinators to facilitate downward counterfactuals, where we contrast our situation with those even worse off. Denial, aside from making procrastination difficult to assess, also perpetuates the condition. Taking a task-oriented strategy, rather than an emotional coping strategy, tends to reduce the degree of procrastination. At times, procrastinators must accept and forgive themselves for their procrastination before they can move from an emotional to a task-oriented coping style (Wohl et al., 2010).

Part 3. Research Perspectives

Concerning the field of psychology, theoretical approaches to procrastination can be grouped into several perspectives (Klingsieck, 2013a). These perspectives are differential psychology, the perspective of motivational and volitional psychology, the perspective of clinical psychology, the situational perspective, and the perspective of neurobiology.

Perspective of Differential Psychology

Differential psychology understands procrastination as a personality trait. Studies in this perspective have built up an extensive nomological network of procrastination and have focused on studying the relationship of procrastination with other trait and traitlike variables (Klingsieck, 2013a).

The Big Five

The big five is a common personality taxonomy used to study procrastination (Steel, 2007; van Eerde, 2003). Increased procrastination is mainly associated with decreased conscientiousness while the factors of extraversion, openness to experience, and agreeableness are not significantly related to procrastination. Neuroticism has an intermittent relationship with procrastination, depending on the measure, which disappears once conscientiousness is controlled for. In addition, some studies have examined the relationship between intelligence and procrastination, but have found no correlation (cf Steel, 2007).

Perfectionism

Early works in the field viewed procrastination as primarily caused by irrational beliefs, such as perfectionism (Burka and Yuen, 2008). Correlational results do not support this notion, indicating a relatively minor contribution compared to other constructs, especially self-efficacy and impulsiveness. However, the role of perfectionism continues to be debated. One reason for the mixed results might be the different approaches to operationalize perfectionism. Within some articles it is understood as a one-dimensional construct, while other authors understand perfectionism as a multidimensional construct (e.g., procrastination associated with socially prescribed perfectionism vs procrastination associated with self-oriented perfectionism). Alternatively, perfectionism's perceived relationship may be due to an increased likelihood of neurotic procrastinators seeking treatment – a self-selection effect – combined with any other major personality trait, high or low, will be experienced by an individual as being instrumental in causing their procrastination (Schouwenburg, 2004). However, this does not preclude that treating pathological perfectionism beliefs could help the *neurotic* procrastinator.

Self, Self-Esteem, and Self-Handicapping

Procrastination is also related to low self-esteem and certain identity aspects such as a diffuse ego identity and greater discrepancies between actual, ought, and desired self. Whether this is a cause or consequence of procrastination is debatable. To the extent that low self-esteem is related to low self-efficacy, it can be viewed as a cause. To the extent that low esteem is due to the compromised performance resulting from procrastination, it is an effect. On balance, the mostly widely

adopted position in the procrastination field is the former of these, that it is a cause, especially with regards to self-handicapping. Delaying tasks until it becomes difficult to determine a person's true capability can be a self-esteem protecting strategy. On the other hand, it can be argued that self-handicapping is a more calculated act and does not reflect the impulse-driven behavior typically associated with procrastination. In either case, they both represent self-defeating delays, but may respond to different treatments. Self-handicappers would best be treated by challenging underlying irrational beliefs. Procrastinators would best be treated by reducing the preference reversal effects of impulsiveness.

Time, Time Orientation, and Time Preferences

Concerning time-related personality styles, procrastination is associated with circadian rhythms known as the 'eveningness' type (Digdon and Howell, 2008). With regard to time orientations, procrastinators seem to display a reduced focus on the future, a fatalistic focus on the present (only avoidance procrastination) or a hedonistic focus on the present (only arousal procrastination), and negative attitudes about the past (Jackson et al., 2003).

Perspective of Motivational and Volitional Psychology

The motivational and volitional psychology perspective understands procrastination as a failure in motivation and/or volition (Klingsieck, 2013a), leading to the intention–action gap (Steel, 2007).

Motivational Aspects

Concerning motivational aspects, procrastination is less likely to occur in the case of intrinsically, self-determined, or flow-inducing activities. Furthermore, low procrastination is related to a mastery goal orientation and internal locus of control. Self-efficacy is often studied with regard to academic procrastination. However, studies display mixed results concerning the role of self-efficacy. Some studies understand and investigate test and self-efficacy as individual contributors to procrastination, while others focus on the mediating role of self-efficacy between perfectionism and procrastination. However, meta-analytically it has on average a strong connection (Steel, 2007).

Behavioral Economics and Temporal Motivation Theory

The phenomenon of procrastination violates the mainstream economic assumption of rationality, being a purposeful and self-defeating delay. Consequently, the field of behavioral economics, which relaxes the rationality assumption, has shown more interest in the phenomenon. Typically, economists adopt a version of the behavioral psychology's matching law, which has similarities to classical economics' subjective expected utility equation, except that rewards are hyperbolically discounted. The concept of hyperbolic discounting or time discounting allows for an interdisciplinary bridge between psychology and economics in the explanation of procrastination. Within this realm, procrastination is related to picoeconomics and the rational choice framework, as well as the concept of present-biased preferences.

The Temporal Motivation Theory (Steel and König, 2006) explains procrastination from a time-discounting perspective. It is an integrative theory that incorporates the best-validated aspects of the core motivational theories. Results were that the constructs of expectancy (e.g., self-efficacy), value (e.g., task aversiveness), and time sensitivity (e.g., impulsiveness) are the major predictors of procrastination. These constructs are organized into an equation consistent with other major theories, particularly Expected Utility Theory, Hyperbolic Discounting, and the Matching Law. Mathematically expressed, $\text{motivation} = (\text{expectancy} \times \text{value})/(1 + \text{impulsiveness} \times \text{delay})$. Motivation increases as the expectancy of an outcome and the size or value of an outcome increase. Motivation decreases as the delay before this outcome and the impulsiveness increases. As per Figure 1, procrastination occurs due to preference reversal, where a proximally or immediately available temptation pulls or distracts away from a former intention. According to this theory, procrastination is more likely to occur if the outcome of an unpleasant activity offers rewards in the distant future.

Volitional Aspects

Concerning volitional aspects, procrastination is associated with decreased self-regulation, decreased self-control, decreased action control, or volitional problems in general. More specifically, procrastinators seem to lack time management strategies, seem to be easily distracted by attractive action opportunities, and display a less frequent use of learning strategies.

Perspective of Clinical Psychology

The clinical psychology perspective focuses on the conditions of and interventions for the clinically relevant extent of procrastination (Klingsieck, 2013a). In explaining procrastination, these theoretical approaches often draw on psychoanalysis, cognitive behaviorism, and neuropsychology (for a summary, see Steel, 2012). Of all the perspectives, it is the clinical one that puts the negative consequences and correlates of procrastination intensely into focus because these determine whether procrastination is clinically relevant or not.

Affective and Cognitive Aspects: Fear of Failure, Test Anxiety, and Depression

Self-help books often state that the most prominent causes of procrastination are irrational beliefs, such as fear of failure or

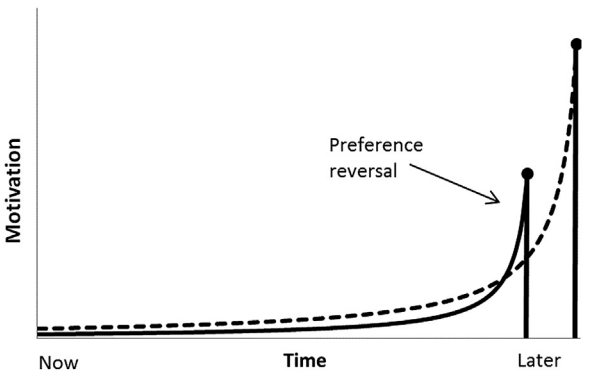

Figure 1 Preference reversal due to temporal discounting.

test anxiety. Thus, procrastination is often investigated within a framework of coping styles and anxiety (e.g., Burka and Yuen, 2008).

Depression is considered both a cause (through the mechanism of diminished energy) and a consequence (through the mechanism of reduced performance) of severe procrastination. However, procrastination is related to depression only for clients diagnosed with depression, being that procrastination is a symptom of the condition.

Behavioral Aspects: ADHD

In realm of procrastination's link to impulsiveness, the relationship between procrastination and attention deficit hyperactivity disorder (ADHD) is discussed but rarely studied. The Axis 1 disorder ADHD is identified by procrastination among other symptoms, such as inattentiveness, difficulty in getting work done, and organizational problems. However, like depression, procrastination is related to ADHD only for persons diagnosed with ADHD, it being a symptom of the condition. For individuals who are not diagnosed with ADHD, procrastination is not associated with attention deficits. However, procrastinators are more likely to be prone to boredom than nonprocrastinators and, like those with ADHD, seek more stimulating situations (Ferrari, 2010).

Situational Perspective

Although intuitively obvious, especially given the volitional problems of procrastinators, few studies have been conducted that focus on environment characteristics that promote procrastination. Thus, the situational perspective encompasses only a small subset of theoretical ideas and empirical studies on procrastination. The situational perspective understands procrastination as a phenomenon that is evoked by certain environmental characteristics (Klingsieck, 2013a). The features associated with procrastination are task difficulty and attractiveness, plausibility of the assignment, autonomy, teachers' or supervisors' characteristics, and the proximity and saliency of temptations and distractions.

Perspective of Neurobiology

Research endeavors in the field of neurobiology have been scarce, although the dualistic theory of mind provides a promising model to explain procrastination (cf Steel, 2012). Referred to by various terms, such as hot/cool, passion/reason, or system 1/system 2, it indicates that the prefrontal cortex often makes or maintains intentions. However, when the mechanisms of the prefrontal cortex are exhausted or compromised, decision making relocates to the limbic system. The limbic system is much more susceptible to short-term temptations and distractions. Consequently, we make intentions to act but find it difficult to follow through despite wishing we would. Drawing on historical records, functional magnetic resonance image research, genetic studies, comparative psychology, developmental psychology, and evolutionary psychology, Steel (2012) finds considerable support for this model. There is also empirical evidence that procrastination can be largely explained by problems with the prefrontal cortex role as the executive function (Rabin et al., 2011).

Part 4. Impact in Different Life Domains

Procrastination research has investigated procrastination in a variety of different life domains, however, studies that compare procrastination's characteristics across different domains are yet to be conducted (Klingsieck, 2013a; for a first endeavor see Klingsieck, 2013b). The life domains studied are education (i.e., academic procrastination), work and financial, health, and subjective well-being. Each of them is briefly discussed below.

The procrastination of study-related activities in a student population is often studied under the term academic procrastination. While procrastination also focuses on procrastination in other life domains, this domain is the most studied within procrastination research. The research mainly focuses on reasons and consequences of procrastination showing that procrastination can have a detrimental impact on achievement and on subjective well-being of students.

Research on procrastination with regard to job-related aspects has concentrated on procrastination of job search behaviors and workplace performance. Procrastination has been shown to be negatively related to implementation intention formation, but to positively predict intention–behavior discrepancies and the extent to which job seekers engage in other activities instead of performing planned job search activities. Moreover, it is also negatively related to the process quality of job seeking. Research concerning procrastination in the workplace is surprisingly scarce. This is striking given that Steel (2012) reviews estimates that put procrastination at over 25% of most people's work time, potentially reducing productivity by $1.25 trillion in the United States alone. Studies have included identifying what job characteristics exacerbate procrastination (i.e., the situational perspective) as well as the consequences of procrastination. For instance, procrastination is frowned upon in the workplace. Interestingly, although procrastination is often understood as a deficit in self-management, it does not seem to be related to deficits in managing others. Concerning the procrastination of important financial aspects, studies have shown that the majority of people by their own standards feel that they have procrastinated savings for their retirement too long. The same holds true for procrastinating on taxes.

The health domain is one of the more procrastination relevant area, speaking to the speed of seeking treatment or 'patient compliance' in treatment programs. Constructs studied in the health domain are often the same as for academic procrastination (e.g., self-efficacy, self-control), indicating that this is largely the same phenomenon and the results are transferable. Within the health area, procrastination presents itself as a self-regulation deficit leading to the intention–action gap with regard to dieting, physical training, and preventive medical checkups. For example, colon cancer is among the deadliest despite being the most treatable, simply because people delay getting colonoscopies.

Procrastination has been shown to have various negative effects on subjective well-being. For example, it has been shown to result in augmented stress and regret. Procrastinators often report having feelings of personal failure due to their procrastination and frustration regarding their inability to start earlier. Furthermore, studies have shown that procrastination can also pertain to enjoyable activities, where people put off pleasurable pursuits. Although procrastination might function

as a short-term mood repair, in the long run it has negative consequences on subjective well-being.

Part 5. Treatments

The negative consequences of procrastination have led to the development of various intervention programs and to the publication of numerous self-help books to overcome procrastination. One of the oldest treatment ideas is the idea of behavior control. Most interventions have focused on either teaching self-management strategies (e.g., goal setting, time management, planning, monitoring, creating the right environment for studying) or implementing therapeutic strategies to tackle negative affect and cognitions (e.g., cognitive restructuring).

The number of these self-help books concerning procrastination has grown steadily in recent years. The works written by counseling psychologists emphasize irrational beliefs, such as perfectionism and low self-esteem, by procrastinators (e.g., Burka and Yuen, 2008). The works written by research psychologists (e.g., Ferrari, 2010; Steel, 2012) indicate that procrastination is primarily a self-regulation failure and an impulse-related issue. Consequently, they focus on improving self-regulation skills and situational or stimulus control.

Therapeutic strategies in overcoming procrastination pertain to techniques of cognitive restructuring (to tackle negative cognitions), relaxation exercise (to tackle negative affect), time restriction, and, paradoxical interventions. For example, a special issue of the *Journal of Rational-Emotive & Cognitive-Behavior Therapy* (REBT; Dryden, 2012) came out in which different approaches concerning REBT are delineated. Anxiety-based treatments are likely to be most effective for the perfectionist or neurotic procrastinator. On the other hand, the outpatient clinic at the University of Münster (Germany) has built a special outpatient clinic for procrastinators focusing on cognitive-behavioral therapy that may be more broadly applicable. The methods they employ include realistic goal setting, stimulus control, and self-motivation.

Most trainings aiming at overcoming procrastination foster strategies concerning self-management such as goal setting, time management, planning, monitoring, and creating the right environment for studying. For a summary of these interventions, see Schouwenburg et al. (2004).

See also: Anxiety and Anxiety Disorders; Behavior Therapy: Background, Basic Principles, and Early History; Decision Making, Psychology of; Decision Making: Nonrational Theories; Depression; Emotional Regulation; Executive Functions During Childhood, Development of; Expectancy-Value-Cost Model of Motivation; Future Time Perspective and Motivation; Grit; Health Self-Regulation, Motivational and Volitional Aspects of; Motivation and Actions, Psychology of; Motivation in Clinical Interventions; Personality and Economics; Personality and Values at Work; Self-Efficacy; Self-Regulated Learning; Subjective Wellbeing, Psychology of; Wisdom, Psychology of; Work Motivation.

Bibliography

Burka, J.B., Yuen, L.M., 2008. Procrastination: Why You Do It, What to Do about It. Da Capo, Reading, MA.

Chu, A.H.C., Choi, J.N., 2005. Rethinking procrastination: positive effects of "active" procrastination behavior on attitudes and performance. The Journal of Social Psychology 145, 245–264.

Digdon, N.L., Howell, A.J., 2008. College students who have an eveningness preference report lower self-control and greater procrastination. Chronobiology International 25, 1029–1046.

Dryden, W., 2012. Dealing with procrastination: the REBT approach and a demonstration session. Journal of Rational-Emotive & Cognitive-Behavior Therapy 30 (4), 264–281.

van Eerde, W., 2003. A meta-analytically derived nomological network of procrastination. Personality and Individual Differences 35, 1401–1419.

Ferrari, J.R., Johnson, J.L., McCown, W.G., 1995. Procrastination and Task Avoidance: Theory, Research, and Treatment. Plenum, New York.

Ferrari, J.R., 2010. Still Procrastinating: The No-Regrets Guide to Getting It Done. Wiley, Hoboken, NJ.

Jackson, T., Fritch, A., Nagasaka, T., Pope, L., 2003. Procrastination and perceptions of past, present, and future. Individual Differences Research 1, 17–28.

Klingsieck, K.B., 2013a. Procrastination: when good things don't come to those who wait. European Psychologist 18, 24–34.

Klingsieck, K.B., 2013b. Procrastination in different life-domains: Is procrastination domain specific? Current Psychology 32, 175–185.

Rabin, L.A., Fogel, J., Nutter-Upham, K.E., 2011. Academic procrastination in college students: the role of self-reported executive function. Journal of Clinical and Experimental Neuropsychology 33 (3), 344–357.

Reuben, E., Sapienza, P., Zingales, L., 2007. Procrastination and Impatience (No. W13713). National Bureau of Economic Research, Cambridge.

Rothblum, E.D., Solomon, L.J., Murakami, J., 1986. Affective, cognitive and behavioral differences between high and low procrastinators. Journal of Counseling Psychology 33, 387–394.

Schouwenburg, H.C., 2004. Procrastination in academic settings: general introduction. In: Schouwenburg, H.C., Lay, C.H., Pychyl, T.A., Ferrari, J.R. (Eds.), Counseling the Procrastinator in Academic Settings. American Psychological Association, Washington, DC, pp. 3–17.

Schouwenburg, H.C., Lay, C.H., Pychyl, T.A., Ferrari, J.R. (Eds.), 2004. Counseling the Procrastinator in Academic Settings. American Psychological Association, Washington, DC.

Silver, M., Sabini, J., 1981. Procrastinating. Journal for the Theory of Social Behavior 11 (2), 207–221.

Sirois, F., Pychyl, T., 2013. Procrastination and the priority of short-term mood regulation: consequences for future self. Social and Personality Psychology Compass 7 (2), 115–127.

Solomon, L.J., Rothblum, E.D., 1984. Academic procrastination: frequency and cognitive-behavioral correlates. Journal of Counseling Psychology 31 (4), 503–509.

Steel, P., 2007. The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychological Bulletin 133, 65–94.

Steel, P., 2010. Arousal, avoidant and decisional procrastinators: do they exist? Personality and Individual Differences 48, 926–934.

Steel, P., 2012. The Procrastination Equation: How to Stop Putting Things Off and Start Getting Things Done. Pearson Education, Harlow, UK.

Steel, P., Ferrari, J., 2013. Sex, education and procrastination: an epidemiological study of procrastinators' characteristics from a global sample. European Journal of Personality 27 (1), 51–58.

Steel, P., König, C.J., 2006. Integrating theories of motivation. Academy of Management Review 3, 889–913.

Steel, P., Brothen, T., Wambach, C., 2001. Procrastination and personality, performance, and mood. Personality and Individual Differences 30, 95–106.

Wohl, M., Pychyl, T., Bennett, S., 2010. I forgive myself, now I can study: how self-forgiveness for procrastinating can reduce future procrastination. Personality and Individual Differences 48 (7), 803–808.

Procrustes Analysis

John C Gower, The Open University, Milton Keynes, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Procrustes analysis is concerned with matching two or more configurations. It has its origins in factor analysis, where the problem is to match configurations that represent different sets of factor loadings. More recent applications are concerned with matching multidimensional configurations arising from forms of multidimensional scaling or from direct measurement. Emphasis is on matching the shapes of the configurations independently of their position. Orthogonal rotations are commonly used, but more general transformations may be allowed. When simultaneously matching several configurations, their average configuration gives an important summary, analogous to the group-average of individual differences scaling and some forms of generalized canonical correlation.

Procrustes analysis is concerned with matching two, or more, configurations of points. Imagine two maps of a region, each giving the positions of the same towns. The two maps are oriented differently, are of different sizes, differ in their accuracies, and use different map projections. At first glance, it may be difficult to see much resemblance between the two maps. Procrustes analysis addresses such problems, usually taking one of the configurations as a target and fitting the other to it, optimally, by using appropriate transformations. The 'maps' may be derived in many ways and may be in several dimensions but, for the concept of matching to be meaningful, it is important that the configurations refer to two representations of the same things: the same towns, the same cases (in multidimensional scaling (MDS)), the same variables (in factor analysis), the same products (in marketing), and the same landmarks (describing biological organisms).

The name comes from Greek mythology where Procrustes was an innkeeper who had a bed (sometimes two). Travelers were stretched or limbs chopped to make them fit the bed; with two, adjustment was ensured. The pejorative implications are no worse in Procrustes analysis than in other forms of data analysis, where, misguidedly, data may be fitted into the straightjacket of an unrealistic model.

Factor Analysis

The origins of Procrustes analysis (see Hurley and Catell, 1962) are in factor analysis, where loadings $\mathbf{X}_1$ of k factors $\mathbf{f}$ occur in the form $\mathbf{X}_1\mathbf{f}$ and so may be replaced by $(\mathbf{X}_1\mathbf{T})(\mathbf{T}^{-1}\mathbf{f})$, showing that they are determined only up to the arbitrary $k\times k$ matrix $\mathbf{T}$. By choosing a suitable $\mathbf{T}$, we may match $\mathbf{X}_1$ 'the traveler' to a previously determined or hypothetical set of loadings given in a matrix $\mathbf{X}_2$ 'the bed'; $\mathbf{T}$ representing the chopping/stretching transformation. Usually, the degree of match is judged by least squares, minimizing $\|\mathbf{X}_1\mathbf{T}–\mathbf{X}_2\|$ the sum of the squared distances between like points of the target and transformed configurations. Without further constraint, this is the multivariate multiple-regression problem with well-known solution.

$$\mathbf{T} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$$

Two of the most important constraints on $\mathbf{T}$ are

1. $\mathbf{T}$ must be orthogonal (denoted here by $\mathbf{T}=\mathbf{Q}$) and
2. $\mathbf{T}$ must be a direction cosine matrix (denoted here by $\mathbf{T}=\mathbf{C}$)

Both choices relate to the interpretation that $\mathbf{f}$ is expressed relative to its orthogonal system of coordinate axes and the columns of $\mathbf{T}$ give direction cosines of new axes relative to the original orthogonal axes. When $\mathbf{T}=\mathbf{Q}$, a new set of orthogonal axes is being sought that gives a better match to $\mathbf{X}_2$; when $\mathbf{T}=\mathbf{C}$ we imagine $\mathbf{X}_2$ as being referred to oblique axes. The orthogonal Procrustes least-squares solution is $\mathbf{Q}=\mathbf{V}\mathbf{U}'$. where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$ is the singular value decomposition of $\mathbf{X}_2'\mathbf{X}_1$. With $\mathbf{C}$, minimization is intrinsically more difficult and is less well defined than in the orthogonal case. This is because there are at least two ways of expressing the position of a point relative to oblique axes and also because there are good arguments for replacing ordinary least squares by a weighted form that measures distance in the metric appropriate to the chosen system of oblique axes, giving four variants in which $\mathbf{T}$ is either $\mathbf{C}$, $\mathbf{C}'$, $\mathbf{C}^{-1}$, or $(\mathbf{C}')^{-1}$. The simplest case was first discussed by Mosier (1939) but not finally solved until Browne (1967) gave a practical numerical method for solving the normal equations. Special algorithms have been proposed by Browne and Kristof (1969) and Gruvaeus (1970) for the case $\mathbf{T}=\mathbf{C}^{-1}$. However, all problems may be tackled uniformly by proceeding as if the directions of all but one of the axes are known and then estimating the optimal direction of the remaining axes, cycling through all the axes iteratively until a stable solution is found for all the directions.

Rather than handling the indeterminacy of factor loading as above, the modern tendency is to use an internal criterion of simple structure such as Varimax (Kaiser, 1958) that has no reference to an external target $\mathbf{X}_2$. Alternatively, $\mathbf{X}_2$ may be coded in binary form where $x_{ij}=1$ indicates a desired high loading of i the variable on the jth factor, and $x_{ij}=0$ indicates a low loading. This latter form remains a type of Procrustes problem, but least squares is no longer appropriate and is replaced by choosing $\mathbf{T}$ to maximize the ratio of the sum of squares of the squared high loadings to the sum of squares of the low loadings (see Lawley and Maxwell, 1971).

Variants and Later Developments

The problems outlined above are referred to as two-sets Procrustes problems because two configuration matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are involved. There are many variants:

1. Alternative constraints on $\mathbf{T}$ such that $\mathbf{T}$ is a rectangular column-orthonormal matrix, or is a permutation matrix
2. Two transformation matrices $\|\mathbf{X}_1\mathbf{T}_1–\mathbf{X}_2\mathbf{T}_2\|$
3. Isotropic scaling: minimize $\|\rho\mathbf{X}_1\mathbf{T}–\mathbf{X}_2\|$
4. Anisotropic scaling: minimize $\|\mathbf{X}_1\mathbf{S}_R\mathbf{T}–\mathbf{X}_2\|$, or $\|\mathbf{X}_1\mathbf{T}\mathbf{S}_R–\mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{T}–\mathbf{X}_2\|$, $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_R\mathbf{T}–\mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\ \mathbf{T}\mathbf{S}_R–\mathbf{X}_2\|$, or $\|\mathbf{S}_L\mathbf{X}_1\mathbf{S}_{R_1}\mathbf{T}\mathbf{S}_{R_2} - \mathbf{X}_2\|$ for unknown diagonal $\mathbf{S}_R$ and $\mathbf{S}_L$ where the suffices R and L denote whether the scaling matrices occur on the right or left of $\mathbf{X}_1$
5. Replace least squares by, e.g., L_1-norm, inner product of trace$(\mathbf{X}_2'\mathbf{X}_1\mathbf{T})$ or a robust estimation criterion
6. Two-sided Procrustes problems $\|\mathbf{R}\mathbf{X}_1\mathbf{T}_1–\mathbf{X}_2\|$ where $\mathbf{R}$ and $\mathbf{T}$ are to be estimated, usually under specified constraints
7. Weighted forms – minimize trace $\mathbf{N}(\mathbf{X}_1\mathbf{T}–\mathbf{X}_2)\mathbf{W}(\mathbf{X}_1\mathbf{T}–\mathbf{X}_2)'$ where $\mathbf{N}$ and $\mathbf{W}$ are given matrices, one of which may be ignored
8. Handling missing values

Brief comments on the above must suffice; Gower and Dijksterhuis (2004) give details. It should be clear that $\mathbf{X}_1$ and $\mathbf{X}_2$ no longer represent factor loadings. Anisotropic scaling allows for differences in size; the more esoteric versions of anisotropic scaling verge on the more pejorative interpretations of the Procrustes legend: permitting real violence to the data. The product of two orthogonal transformation matrices is equivalent to one and the inner product and least-squares criteria, then, give the same result. Weights $\mathbf{N}$ are useful to cover replication or missing rows of $\mathbf{X}_1$ and need not be diagonal, although they usually are. Note that weights are given matrices, whereas anisotropic scaling is expressed in terms of unknown diagonal matrices requiring estimation. However, sometimes $\mathbf{W}$ may be a function of the unknown $\mathbf{T}$, as with oblique Procrustes analysis in a metric where $\mathbf{W}$ is $\mathbf{C}'\mathbf{C}$ or its inverse. Albers and Gower, 2010 give an overview of the many complications that can arise when estimating missing values in Procrustes analysis.

Changes in orientation are given by an orthogonal matrix $\mathbf{Q}$, which includes rotations and reflections, leaving intact the essential distance and angular properties of the configurations. A $p\times q$ orthonormal matrix $\mathbf{T} = \mathbf{P}$ with q columns represents a projective from p to q dimensions. Projections may be interesting because

1. $\mathbf{X}_2$ has fewer dimensions than $\mathbf{X}_1$
2. A match may be required in some specific number of dimensions, say $q = 2$

When $q < p$ we may still match in p dimensions by rotating in the higher dimensional space; e.g., we may rotate three points lying on a line to fit three points lying in a plane. All we have to do is to augment $\mathbf{X}_2$ by $p < q$ zero columns. That $\|\mathbf{X}_1\mathbf{P}–\mathbf{X}_2\|$ can be less than $\|\mathbf{X}_1\mathbf{Q}–\mathbf{X}_2\|$ (where $\mathbf{X}_2$ is augmented by zero columns) has been seen as a justification for preferring projections to rotations, but this is not comparing like with like. I believe that projection Procrustes analysis is rarely justified.

Nowadays, applications of Procrustes analysis are more often than not more to do with matching configurations rather than factor loadings (Commandeur, 1991). For example $\mathbf{X}_1$ and $\mathbf{X}_2$ may be the two sets of points given by different MDSs of the same data, or they may refer to MDS solutions based on different proximity/distance matrices measured between the same p objects, perhaps based on different variables. Fruitful applications have been found in the food sciences, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are derived from scores given by two judges on traits of several products (e.g., Dijksterhuis and Gower 1991/1992). Applications in the biological sciences include the analysis of shape where the configurations refer to the directly observed coordinates of landmark features on organisms (see e.g., Bookstein, 1991, Dryden and Mardia, 1998). In shape analysis the matching of one configuration to another is often termed registration.

Generalizations

Given K configurations $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_K$ we may match them in pairs by doing the corresponding two sets of Procrustes analyses, minimizing $s_{ij} = \|\mathbf{X}_1\mathbf{T}_1–\mathbf{X}_2\mathbf{T}_2\|$ for all pairs (i,j). Each s_{ij} may then be assigned to a matrix $\mathbf{S}$ and analyzed by some form of MDS (see Gower, 1971 for an example from physical anthropology).

Alternatively, we may minimize $\sum_{k=1}^{K}\|\mathbf{X}_k\mathbf{T}_k - \mathbf{G}\|$ or $\sum_{k\leq k'}^{K}\|\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_{k'}\mathbf{T}_{k'}\|$ where $\mathrm{G} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{X}_i\mathbf{T}_i$ often termed the group average. For the orthogonality criterion, both are equivalent to minimizing $\sum_{k=1}^{K}\|\mathbf{T}_k'\mathbf{X}_k'\mathbf{G}\|$. Generalized Procrustes analysis may be viewed as a three-mode analysis, with modes for rows, columns, and configurations (see Kroonenberg, 2008 for three-mode analysis). For some information on algorithms, see Ten Berge (1977). To help interpretation, each dimension of the analysis may be summarized in an analysis of variance, partitioning the total into terms for the group average and for departures from the group average. Some dimensions may be identified with noise and some with signal, further subdivided into displayed and undisplayed parts (Dijksterhuis and Gower 1991/1992).

It has been suggested that maximizing the group average of Procrustes analysis should be a main objective of Procrustes analysis. For orthogonal $\mathbf{T}$, both approaches are equivalent, but for projections, maximizing the group average is not the same as minimizing the criteria given above. The situation gets very convoluted, especially when the variants listed in Section 2 are admitted. This is because when focusing on the group average, the transformations, isotropic scaling, and anisotropic scaling may be regarded as being attached to $\mathbf{G}$ rather than to the $\mathbf{X}_i$, giving two versions of every variant (see Gower and Dijksterhuis, 2004 for further details).

The reason for focusing on maximizing the group average may be influenced by the important part played by the concept of a group average in several psychometric methods (see, e.g., Individual Differences Scaling) as a useful summary of K data sets. Ultimately, the range of such methods may be so divergent from orthodox Procrustes procedures that the Procrustes terminology becomes questionable. However, here we shall single out mention of generalized canonical correlation analysis, both

because of its superficial resemblance to Procrustes analysis and because of its general importance. There are several ways of generalizing canonical correlation analysis, but one of the most useful is based on the equivalent Procrustean forms $\sum_{k=1}^{K}\|\mathbf{X}_k\mathbf{T}_k - \mathbf{G}\|$ or $\sum_{k\leq k'}^{K}\|\mathbf{T}_k'\mathbf{X}_k'\mathbf{X}_{k'}\mathbf{T}_{k'}\|$, where now $\mathbf{T}_k$ is a general matrix whose columns conform with the rows of $\mathbf{X}_k$ but where $\mathbf{T}_k$ has r columns independently of k. Now the columns of $\mathbf{X}_k$ are interpreted as variables and not as coordinates of configurations. When $K=2$, minimization coincides with classical two-sets canonical correlation analysis, thus satisfying one of the primary conditions for generalization. The generalization to $K>2$ satisfies the generalized (two-sided) eigenvector problem. The columns of $\mathbf{X}_k\mathbf{T}_k$ are the canonical variables, and the Procrustean form of the least-squares criterion shows that the mean (i.e., the form taken by the group average) of the K canonical variables in each dimension matches the group average. The sum of squares of the correlations between the K leading canonical variables and the group average have a correlational interpretation, so linking the generalization with two-group canonical correlation.

The block matrix with elements $\mathbf{Z}=\{\mathbf{X}_k'\mathbf{X}_{k'}\}$, central to generalizing canonical correlation analysis, is also the basis of the STATIS method, usually after some initial standardization of each $\mathbf{X}_k$. The leading eigenvector of the matrix with (k,k')th element trace$\|\mathbf{X}_k'\mathbf{X}_{k'}\|$ estimates isotropic scaling factors s_k, say; then $\sum_{k=1}^{K} s_k\mathbf{X}_k$ may be interpreted as a group average. This and similar links between Procrustes analysis and other methods that involve a group average, are likely to be a continuing area for research in the immediate future. Research relating to biological landmark data and shape analysis remains active and will continue (see Kendall et al., 1999). Other research will focus on algorithms, the effects of estimation criteria and the difficult area of inference.

See also: Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Mathematical Psychology: History; Multidimensional Scaling I; Multidimensional Scaling II.

Bibliography

Albers, C.J., Gower, J.C., 2010. A general approach to handling missing values in Procrustes analysis. Advances in Data Analysis and Computing 4, 223–237.

Bookstein, F.I., 1991. Morphometric Tools for Landmark Data. Cambridge University Press, Cambridge, UK.

Browne, M.W., 1967. On oblique Procrustes rotation. Psychometrika 32, 125–132.

Browne, M.W., Kristof, W., 1969. On oblique Procrustes rotation of a factor matrix to a specified pattern. Psychometrika 34, 237–248.

Commandeur, J.J.F., 1991. Matching Coefficients. DSWO Press, Leiden, The Netherlands.

Dijksterhuis, D.B., Gower, J.C., 1991/1992. The interpretation of generalized Procrustes analysis and allied methods. Food Quality Preference 3, 67–87.

Dryden, I.L., Mardia, K.V., 1998. Statistical Shape Analysis. Wiley, Chichester, UK.

Gower, J.C., 1971. Statistical methods for comparing different multivariate analyses of the same data. In: Hodson, J.R., Kendall, D.G., Tautu, P. (Eds.), Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh Scotland.

Gower, J.C., Dijksterhuis, G.B., 2004. Procrustes Problems. Oxford University Press, Oxford.

Gruvaeus, G.T., 1970. A general approach to Procrustes pattern rotation. Psychometrika 38, 493–505.

Hurley, J.R., Catell, B.B., 1962. The Procrustes program: producing direct rotation to test a hypothesized factor structure. Behavioural Science 7, 258–262.

Kaiser, H.F., 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23, 187–200.

Kendall, D.G., Barden, D., Carne, T.K., Le, H., 1999. Shape and Shape Theory. Wiley, Chichester, UK.

Kroonenberg, P.M., 2008. Applied Multivariate Data Analysis. Wiley, Hoboken.

Lawley, D.N., Maxwell, A.E., 1971. Factor Analysis as a Statistical Method, second ed. Butterworths, London.

Mosier, C.I., 1939. Determining a simple structure when loadings for certain tests are know. Psychometrika 4, 149–162.

Ten Berge, J.M.F., 1977. Orthogonal Procrustes rotation for two or more matrices. Psychometrika 42, 267–276.

Pro-Environmental Behavior

Deniz S Ones and Brenton M Wiernik, University of Minnesota, Minneapolis, MN, USA
Stephan Dilchert, Baruch College, The City University of New York, New York, NY, USA
Rachael Klein, Korn Ferry, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Environmental sustainability is one of the most pressing issues facing societies today. Employees play a key role in contributing to organizational environmental performance. This article describes the domain of employee green behaviors (pro-environmental behaviors at work), distinguishes them from related constructs, provides an overview of determinants, and reviews interventions designed to support them.

Environmental problems are escalating. Research from scientific disciplines converges on the conclusion that the degradation of our natural environment has hastened and that the capability of natural systems to cope with such decline has been compromised (IPCC, 2013, 2014). Globally, societies are becoming increasingly aware of the negative consequences of environmental degradation on human health, ecosystems, and sustainability of life on earth.

Environmental sustainability, defined as "living within the regenerative capacity of the biosphere" (Wackernagel et al., 2002: p. 9266), must be one of society's most important goals in order to ensure the long-term viability of humanity on this planet. Two approaches to achieving environmental sustainability are mitigating current environmental harm and adapting human activities to function within the limits of the natural environment to limit future damage to the biosphere.

Economic activity has the largest impact on natural resources depletion, water use, land use, greenhouse gas emissions, and biodiversity loss. Businesses contribute to, but are also affected by, environmental degradation, which introduces uncertainty into supply chains and core business operations. In addition, consumers are demanding environmentally responsible business practices and products. Thus, business organizations, especially large multinational corporations, are increasingly concerned with understanding and contributing to environmental sustainability. Ultimately, organizations recognize that the viability of their continued operation depends on practicing *corporate social responsibility* and embracing a 'triple bottom line' that acknowledges the importance of social equity and environmental sustainability in addition to economic prosperity.

It is important to note that while corporate social responsibility (CSR) and corporate environmental performance are related, these two aspects of organizational performance are conceptually distinct, have different goals, and derive from distinct antecedents and interventions. CSR aims to improve the well-being of people directly (e.g., through charitable donations, safe working conditions, or by providing training and education). Environmental performance focuses on ensuring the sustainability of the natural environment, in order to protect *all future life* on this planet, by responsibly using natural resources. This theoretical distinction is especially important when managing employees, as the knowledge, skills, and motivation needed to support organizational, social, and environmental goals differ dramatically. Moreover, CSR and corporate environmental performance show different relationships with other variables such as corporate financial performance (Orlitzky et al., 2003).

For organizations, environmental sustainability means responsible management of natural resources in their operations, to sustain the natural environment for organizational economic viability, as well as present and future life on earth. According to a current survey of more than 1600 CEOs, resource scarcity and climate change are among the top five trends that are transforming business today (SAM and PwC, 2011). The same survey also found high CEO support for (1) protecting the interests of future generations (76%) and (2) measuring and reducing organizations' environmental footprints (80%). However, both goals are sufficiently broad to prompt a closer look at what business organizations are actually doing to support their environmental sustainability goals.

Organizational Pro-Environmental Initiatives

One prominent way that organizations support environmental sustainability is by enacting and implementing *organizational pro-environmental initiatives*, which refer to organizational level programs, actions, and interventions that enhance environmental sustainability. Functionally, they can be directed at minimizing, mitigating, and neutralizing environmental impact or at bringing about positive environmental change. Nearly all large organizations are implementing pro-environmental initiatives to some degree, and companies which are environmental leaders have implemented several dozen initiatives and embedded environmental sustainability into their core organizational identities and business strategies (Bansal and Roth, 2000).

Most proposed solutions to emerging environmental threats, including most organizational pro-environmental initiatives, are structural and material changes, such as the creation of new sustainable products, processes, practices, and technologies to replace ecologically harmful ones. Such advances are being made, but environmental scholars agree that the rate is slower than necessary to curb further damage. Organizations are also recognizing that slow to emerge technological and structural solutions to environmental crises

are unlikely to sufficiently protect their current and near-future economic interests. Environmental degradation can be considered a form of maladaptive human behavior (Maloney and Ward, 1973). Organizations do not consume resources, generate waste, or release pollutants. Such decisions are made, acted on, and demanded by leaders, employees, and consumers. Conversely, organizations do not behave in ways that promote sustainability, whereas individual employees do. Preventing and reversing environmental harm thus must involve changing unsustainable human behavior in general. In organizational settings, environmental sustainability goals are achievable only through the efforts of employees at all levels of the organization (Dilchert and Ones, 2012a).

Using Human Resources to Enable and Support Organizational Environmental Sustainability

Despite the key role that employees play in bringing about environmental outcomes, the impact of individual-level behaviors on environmental sustainability has been primarily examined in household settings (Dilchert and Ones, 2012b). The environmental contributions individuals make in their work lives may be more important because, through their actions, employees can impact the environmental performance of organizations as a whole. Hough and Muschewske (2012) astutely observed that "major innovations leading to increased levels of environmental sustainability will come from an engaged and committed workforce responding to a compelling vision communicated by business leaders that new ways to achieve shareholder value through improved environmental sustainability methodologies must be found." Employees are the building blocks on which organizational environmental performance is founded.

Individual Pro-Environmental Behaviors

The general domain of individual behaviors that contribute to environmental sustainability is referred to as *pro-environmental behaviors*, which have been defined as "individual behaviors contributing to environmental sustainability (such as limiting energy consumption, avoiding waste, recycling, and environmental activism)" (Mesmer-Magnus et al., 2012: p. 160). These behaviors may be public (for example, taking mass transit, participating in a rally for an environmental cause) or private (for example, composting, not using home air conditioning on a hot day). The positive environmental behaviors individuals engage in as part of their personal lives are volitional actions rooted in their own initiative. Although societal structures, such as the presence of a public transportation system or recycling program in one's city, may support or hinder pro-environmental behaviors, acting in ways that benefit the environment is ultimately a personal choice.

When environmental behaviors are performed in work organizations or as part of performance in individuals' work roles, they have been referred to as *employee green behaviors*. Employee green behaviors are defined as "scalable actions and behaviors that employees engage in that are linked with and contribute to or detract from environmental sustainability" (Ones and Dilchert, 2012a: p. 87). Besides the setting in which they take place (at work), these employee green behaviors are also distinguished from general pro-environmental behaviors by the degree of control that organizations exert over them. Whereas pro-environmental behaviors are wholly based on individual volitional choice, employee green behaviors are under greater scrutiny by the employer and might even be required by organizations as part of individuals' job duties. Their primary target of impact is the environmental performance of the organization.

In one regard, employee green behaviors are a more comprehensive concept, because – analogous to other performance constructs (see below) – they are defined to include behaviors with both a positive and negative impact on the natural environment. That is, positive employee green behaviors improve organizational environmental performance, whereas negative employee green behaviors detract from organizational environmental performance. Refraining from positive environmental behaviors is not the same as actively harming the environment. The former can be thought of as 'green behavior inhibition' and the latter as 'ungreen behavior activation.' Overall then, the definition of employee green behaviors distinguishes the concept from that of general pro-environmental behavior in terms of (1) organizational control over the behaviors, (2) their target (organizational environmental performance), and (3) their spectrum (positive/negative behavioral activation/inhibition). Although it would be fitting to refer to the domain as pro-environmental behaviors at work, we prefer the term employee green behaviors to highlight its distinctiveness from pro-environmental behavior more generally defined. When employee green behaviors are used as partial index of the employee's relative worth to the organization, they may be referred to as *employee environmental performance*.

The content and targets of employee green behaviors have been studied in a wide variety of organizations, jobs, and countries by researchers at the University of Minnesota and the City University of New York (Ones and Dilchert, 2012a; 2013). Critical incidents studies and careful consideration of several thousand specific behaviors were used to identify 16 broad types of employee green behaviors. These fall into five even broader main content categories. These categories, called the Green Five, are:

1. *Conserving*. This category includes behaviors aimed at avoiding waste and preserving resources. That is, conservation by reducing use, reusing, recycling, and repurposing resources is the behavioral content of this category. Frugality and thrift form the category's functional core.
2. *Avoiding harm*. This category includes behaviors aimed at reducing impact and mitigating environmental damage. Central to this category are behaviors that prevent pollution (or, at the negative pole, actions that pollute the environment). This category also includes monitoring one's environmental impact or footprint and taking steps to recover or restore the environment after damage has been done (such as planting trees, picking up litter, or cleaning an oil spill). The psychological core of most avoiding harm behaviors

(at the positive end) may be characterized by cautiousness, self-control, and responsibility.

3. *Transforming*. This category involves adapting and changing to be more sustainable. It includes creating innovative solutions, such as changing how work is done to be more sustainable and creating new sustainable products and processes, as well as adopting innovations made by others, such as choosing responsible alternatives (e.g., green products, renewable energy, durable materials) and embracing technologies that are better for the environment. Adaptability and openness to change are core to transforming behaviors.
4. *Influencing others*. This category includes social behaviors that support or enable other individuals' sustainable behaviors. It includes providing individuals with education and training to develop the knowledge and skills necessary to perform employee green behaviors. This category also includes behaviors that motivate, encourage, or support green behaviors, such as providing incentives for recycling, coordinating a carpool system, or praising individuals who come up with new green ideas. Leadership, sociability, and altruism are important in influencing behaviors.
5. *Taking initiative*. This category involves behaviors that break the mold or reject the status quo. In performing these behaviors, a level of personal risk is accepted to advance environmental sustainability. This category includes initiating new environmental programs or policies, lobbying and activism for the environment, and putting environmental interests first. Entrepreneurship, risk-taking, and self-sacrifice are hallmarks of these behaviors.

The behaviors within each of these categories are similar in the kinds of actions individuals take and in their psychological foundations, even if they are targeted at different sustainability goals or utilize different resources. For example, reducing water use, reducing energy use, and reducing food waste are all conserving behaviors involving the preservation of resources. These behaviors are psychologically more similar to each other than are reducing energy use, purchasing solar power, and designing a more energy efficient LED (behaviors all involving energy, but requiring divergent behavioral categories of conserving and transforming). Even though the behaviors in the second set are all targeted toward energy, the psychological mechanisms that impel them are quite different. When conceptualizing, measuring, and influencing employee green behaviors, it is important to acknowledge the multitude of behaviors that can contribute to a particular sustainability goal and that multiple targets can be affected through the same intervention.

Jobs and Pro-Environmental Behaviors

As economic activities evolve to incorporate capital and technological resources to address environmental sustainability needs of societies, labor demands also evolve. Embedding pro-environmental behaviors into jobs requires an understanding of job performance dimensions. Job performance is a multidimensional construct with a substantive general factor that spans across its higher order dimensions of task performance, organizational citizenship performance, and avoidance of counterproductive work behaviors (CWB; Viswesvaran and Ones, 2000). Rather than conceptualizing employee green behaviors as yet another performance dimension, we argue that environmental performance at the employee level can in fact span all of these dimensions. The degree to which such behaviors are considered task performance, citizenship performance, or counterproductive behaviors depends on their direction (positive/negative), target, and the nature of employees' job family and formal job duties.

Green jobs refer to new and emerging occupations that arise to address environmental sustainability needs, particularly in the energy production industry (e.g., wind turbine technician). For these jobs, job performance includes green task performance as part of task performance, green citizenship behaviors as part of citizenship performance, and the avoidance of environmentally harmful behaviors or green CWB. Thus, in these jobs, employee green behaviors span across all three general performance dimensions.

Greening jobs refer to occupations for which enhancements to skills are required to meet new sustainability-related demands (e.g., architects who need to acquire knowledge and skills for designing energy efficient buildings). For these jobs, green task performance demands are increased in level and/or variety. Citizenship behaviors directed at environmental sustainability and avoidance of CWB in the sustainability arena are also conceptualized as part of job performance.

Greening economy jobs are traditional occupations for which demand is increasing to meet expanding green industry needs as economies are undergoing environmental sustainability transformations (e.g., clerical workers in recycling companies). For these jobs, no (or few) components of task performance are expected to focus on green task performance. However, there may be increased demand on employees to perform green citizenship behaviors and avoid green CWB, given green organizational focus and concerns.

Finally, even for occupations that are not part of these taxonomic categories, the economic and cultural milieu of organizations focused on the triple bottom line mean that employee green behaviors increasingly are included as part of individual job performance, most notably in domains other than task performance. An important takeaway is that employee green behaviors can function as task performance, citizenship performance, or CWB depending on the occupation and industry. The content of the Green Five (summarized above) can permeate all job performance dimensions.

Determinants of Employee Green Behaviors

Like all work performance behaviors, employee green behaviors are directly determined by three individual characteristics: knowledge, skill, and motivational choices of where, at what level, and how long to direct one's effort. Other individual factors, such as values, attitudes, personality characteristics, as well as various organizational interventions, contribute to employee green behaviors either by influencing choices and directing effort or by providing employees with the knowledge and skill needed to perform the relevant behaviors.

The vast majority of research on environmental interventions and education has focused on how to raise awareness of environmental problems and increase individuals' knowledge about how to solve them. Knowledge, skill, and awareness are key enablers of employee green behaviors. Awareness of environmental problems and the consequences of individual behaviors are preconditions for pro-environmental action, and a lack thereof is a major barrier (Hansla et al., 2008). Such awareness can lead individuals to develop moral norms about the need for environmental protection, as well as feelings of guilt in events of norm violation (Bamberg and Möser, 2007). However, more important than simple awareness of problems is knowledge about their causes and the actions necessary to solve them (Hines et al., 1986/1987). "It is important to know that something needs to be done, but it is even more important to know what and how to do it" (Dilchert and Ones, 2012b: p. 195).

Awareness, knowledge, and skill are necessary but not sufficient conditions for employee green behaviors. While an employee cannot act sustainably if he/she does not know how; simply knowing how does not mean that the employee actually *chooses* to perform green behaviors. For this reason, a great deal of psychological research on pro-environmental behaviors has focused on individual characteristics that influence environmental motivation – individuals' choices about which environmental behaviors they actually engage in.

A large body of research has examined the impact of demographic characteristics on environmental behavior (see Klein et al., 2012). Stereotypes of the 'environmental consumer' are frequently portrayed and discussed in popular media outlets. Individuals who are concerned about the environment and who perform pro-environmental behaviors are characterized as being young, often female, highly educated, and affluent. Research has shown only modest support for these stereotypes (Hines et al., 1986/1987). Education and income are only weakly related to environmental behavior. Pro-environmental behaviors are not the province of the wealthy, and general education does not necessarily lead to increased knowledge, skill, or motivation to perform pro-environmental behaviors. Gender is also only weakly related to pro-environmental behavior performed in individuals' personal lives, though women tend to perform more employee green behaviors at work than men. Age shows more complex relationships with pro-environmental behavior. Wiernik et al. (2013), based on their meta-analysis of general population samples, reported that although most age–behavior relationships were negligible, small to moderate relationships were found for engaging with nature, avoiding harm, and conserving resources. Older individuals appear to be more likely to engage with nature and avoid environmental harm. Perhaps more interestingly, although prior to 1995, younger individuals tended to perform more conserving behaviors than older individuals, this trend appears to have reversed. Based on data from the last two decades, older individuals engage in more conserving behaviors. Relationships between age and green behaviors in work settings are similar; older employees are somewhat more likely to perform conserving, avoiding harm, and influencing others' green behaviors (Wiernik et al., 2011).

Beyond demographic characteristics, environmental attitudes and values are key drivers of pro-environmental behaviors. Environmentally relevant values (e.g., altruistic, egoistic, anthropocentric, traditional; Dunlap and Van Liere, 1978; Stern and Dietz, 1994), connectedness to nature (Mayer and Frantz, 2004), and environmental perspective taking (Schultz, 2000) are all linked with pro-environmental behavior. Individuals are more likely to perform pro-environmental behaviors when they feel that the environment is worth protecting and that those pro-environmental behaviors will have an important positive environmental impact (Bamberg and Möser, 2007). This link between positive attitudes and behavior is especially strong when the behaviors are difficult or challenging (Kaiser and Schultz, 2009). When a behavior demands a great deal of effort, only those individuals who experience perceived value will bother to perform it.

One form of attitude that is vitally important for employee green behaviors is *environmental commitment*. Environmental commitment is defined as "the extent to which an individual is dedicated to environmental sustainability and is willing to engage in pro-environmental behaviors" (Mesmer-Magnus et al., 2012: p. 159). Individuals who are committed to the environment incorporate sustainability into their personal identities and act in ways that support this identity. When organizations adopt the environment as a key stakeholder in their operations, employees' commitment to the *organization*, the extent to which employees are bonded to and linked with the organization, also serves to support the environment. By combining individual environmental commitment with employee commitment to environmentally responsible organizational goals, strong motivation to perform employee green behaviors can be formed.

Recently, researchers have also begun to examine the influence of broader psychological characteristics and personality traits on individuals' pro-environmental behaviors. Broad personality traits show strong relationships with environmentally relevant behaviors. Wiseman and Bogner (2003) report that individuals low on agreeableness and conscientiousness tend to exploit environmental resources more than others. Furthermore, lower impulsivity and lower emotional stability (i.e., higher neuroticism) relate to conserving behaviors and preservation. Agreeableness and conscientiousness have also been found to relate to self-reported electricity conservation (Milfont and Sibley, 2012). Finally, openness to experience, conscientiousness, and extraversion appear to contribute to pro-environmental behaviors among students, and openness correlates with general conservation and adopting green habits and behaviors (Markowitz et al., 2012). In work settings, recent empirical research has also found that openness, agreeableness, and conscientiousness predict employee green behaviors (Dilchert and Ones, 2012b). Individuals who are responsible, considerate of others, and who place value in beauty, ideas, and experiences tend to act in ways that support environmental well-being.

An important method for examining individuals' environmentally relevant choices is to consider the reasons individuals cite for engaging in environmentally beneficial or harmful behaviors; what psychological functions do such behaviors

serve? These psychological functions are captured by *environmental motives*. Klein (2014) conducted a series of studies examining the motives individuals gave for their employee green behaviors. These studies identified 18 distinct functional motives grouped into four categories. *Prosocial motives* include performing green behaviors in order to benefit the environment, to help other people, or out of a sense of social responsibility. *Enabling capabilities* involve conditions that increase the ease or capability of sustainable actions, including performing green behaviors simply because the individual knows how to, out of habit or personal preference, or because environmentally responsible options were available or convenient. *Extrinsic motives* involve performing employee green behaviors in order to obtain some other benefit not relevant to the environment, such as performing an environmentally beneficial or harmful behavior for its health or safety benefits or out of financial considerations. Finally, *image motives* involve performing green behaviors in order to foster one's own or the organization's reputation, in order to conform with organizational or cultural norms, or because such behaviors are required. Of the four categories of motives, the intrinsically oriented motives (prosocial motives and enabling capabilities) showed stronger relationships with employee green behaviors than did extrinsic or image motives.

Improving Employee Green Behaviors

Methods to improve employees' environmental performance can take two broad pathways – focusing on individuals or focusing on interventions.

Focusing on individuals involves recruiting or selecting candidates who are likely to perform employee green behaviors with higher frequency. Environmental sustainability can be incorporated into organizational recruitment efforts, in order to attract environmentally committed individuals. This requires developing materials emphasizing organizational environmental goals, training recruiters to address environmental questions and issues, and targeting sources likely to provide environmentally passionate job applicants. Incorporating environmental messages into recruitment efforts can also lead to higher rates of applicant attraction and job offer acceptance (Grolleau et al., 2012; Jones et al., 2014).

Organizational staffing efforts can also target environmental sustainability by selecting employees using psychological characteristics that predict employee green behaviors. Many of the individual characteristics described above could be used to select employees, including environmental values and environmental commitment. Broad personality traits, such as conscientiousness, agreeableness, and openness, are especially promising as selection criteria for improving environmental performance because these characteristics are also notably linked to other important job performance dimensions, including task performance, organizational citizenship behaviors, avoiding counterproductive behaviors, and creativity and innovation. It is frequently believed that promoting environmental sustainability must come at the cost of other individual and organizational goals, especially financial goals. This is not the case. The same personality characteristics that lead individuals to be environmentally responsible also lead them to be productive, helpful, rule-following, and innovative employees (Ones and Dilchert, 2012b).

In contrast to focusing on bringing environmentally responsible individuals into organizations, an intervention-focused approach to improving environmental performance seeks to increase existing employees' green behaviors by increasing their knowledge, skills, and motivation. Research on environmental interventions in organizations is in its infancy, but many studies on promoting pro-environmental behavior have been reported in the broader social psychological literature. These interventions have been reviewed by Osbaldiston and Schott (2012). Pro-environmental behavior interventions can be grouped into four categories based on the psychological methods used to promote behavior:

1. *Salience interventions* increase individuals' motivation for green behavior by drawing attention to sustainable options and making performing such behaviors easier. These interventions include increasing convenience (e.g., placing compost bins in cafeterias), providing prompts (e.g., posting signs reminding drivers not to idle their engines), offering appeals to individuals' sense of responsibility (e.g., "Only you can prevent forest fires"), or enforcing requirements for green behaviors.
2. *Informational interventions* increase individuals' knowledge and skills to perform green behaviors. They range in complexity from providing mere justifications (why individuals should perform pro-environmental behaviors), to provide instructions (what steps need to be followed), to providing more detailed information about the consequences of environmental behavior, and finally to providing comprehensive education about environmental systems and the impact humans have on them.
3. *Monitoring and feedback interventions* involve tracking individuals as they perform environmental behaviors and tailoring the intervention based on their particular needs and behaviors. These interventions can involve providing feedback on how individuals are performing and where they can improve, as well as providing penalties, rewards, or incentives designed to lead individuals to choose environmentally beneficial actions over harmful ones.
4. *Psycho-social interventions* involve leveraging social, psychological, and perceptual processes to influence environmental behavioral choices. These interventions can include asking individuals to make formal commitments and set environmental goals, placing social models of effective green behavior in individuals' work environments, or improving individuals' self-efficacy for performing green behaviors. One form of psycho-social intervention that appears to be particularly effective is asking individuals to evaluate whether their environmentally relevant behaviors match their espoused environmental attitudes and values. The cognitive dissonance generated when individuals realize that their behaviors do not support their attitudes can lead individuals to consider and adopt specific alternatives that will better match their stated environmental values.

Each of these classes of interventions can be effective for promoting pro-environmental behaviors to some degree.

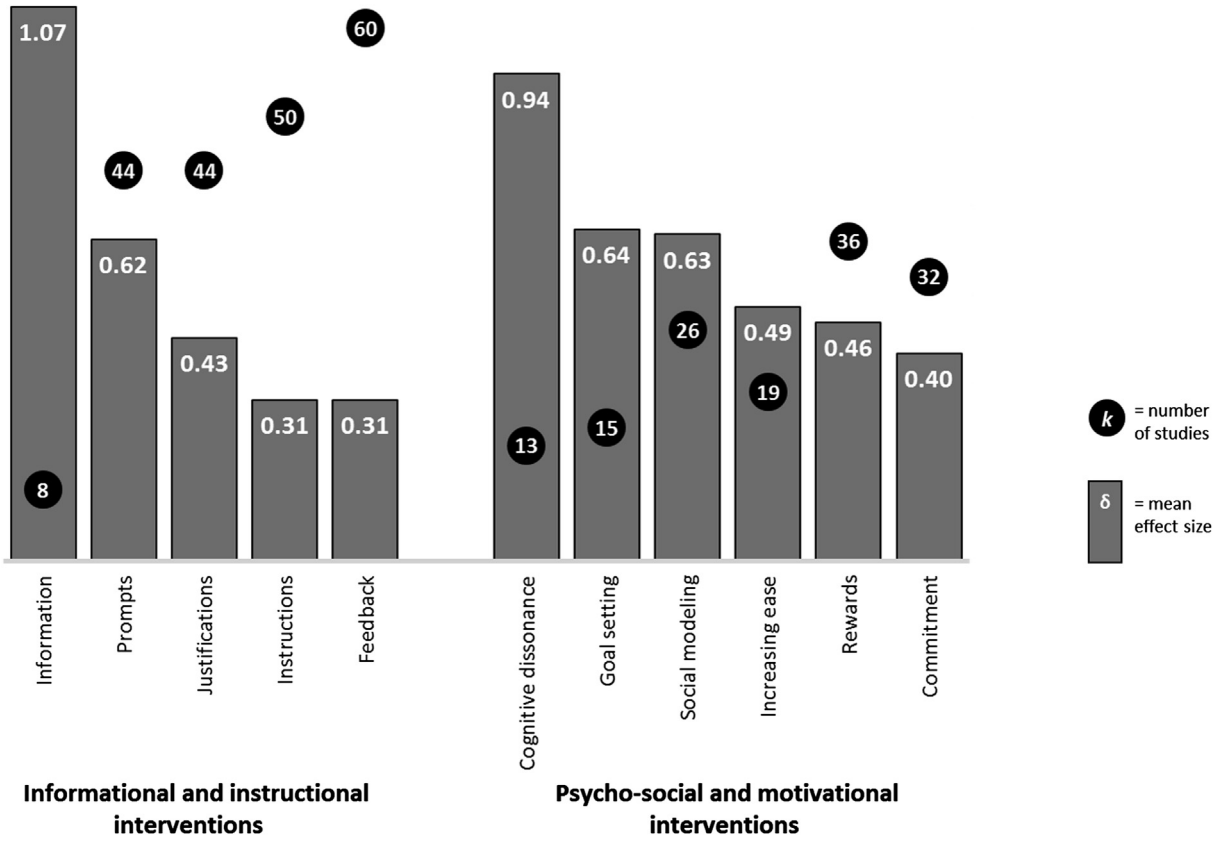

Figure 1 Impact of interventions on pro-environmental behaviors.

Figure 1 summarizes the relative efficacy of various interventions based on meta-analytic evidence previously summarized in Ones and Dilchert (2013, Tables 7.3 and 7.4). The most effective single interventions across settings appear to be placing prompts encouraging pro-environmental behaviors in strategic locations, cognitive dissonance exercises, and providing information on social norms (e.g., "You are using more energy than 50% of your neighbors"). More effective, however, are interventions that combine several approaches in order to support pro-environmental behaviors using multiple psychological pathways.

Conclusions

Environmental sustainability is a pressing societal issue, and the importance of developing our economies within the limits of the natural environment will only continue to grow in the future. Individual actions, especially the actions of employees in organizations, play a vital role in achieving, or failing to achieve, environmental sustainability. Understanding the diversity of behaviors in this domain, the individual characteristics that enable and motivate these behaviors, and the interventions that can be used to support and promote them, is of paramount importance for organizations hoping to achieve lasting organizational viability.

See also: Environmental Attitudes and Behavior: Measurement; Environmental Cognition, Perception, and Attitudes; Globalization and Transnational Corporations; Green Jobs; Sustainable Practices.

Bibliography

Bansal, P., Roth, K., 2000. Why companies go green: a model of ecological responsiveness. Academy of Management Journal 43, 717–736.

Bamberg, S., Möser, G., 2007. Twenty years after Hines, Hungerford, and Tomera: a new meta-analysis of psycho-social determinants of pro-environmental behavior. Journal of Environmental Psychology 27, 14–25.

Dilchert, S., Ones, D.S., 2012a. Environmental sustainability in and of organizations. Industrial and Organizational Psychology: Perspectives on Science and Practice 5, 503–511.

Dilchert, S., Ones, D.S., 2012b. Measuring and improving environmental sustainability. In: Jackson, S.E., Ones, D.S., Dilchert, S. (Eds.), Managing Human Resources for Environmental Sustainability. Jossey-Bass/Wiley, San Francisco, CA, pp. 187–221.

Dunlap, R.E., Van Liere, K.D., 1978. The "New Environmental Paradigm": a proposed measuring instrument and preliminary results. Journal of Environmental Education 9, 10–19.

Grolleau, G., Mzoughi, N., Pekovic, S., 2012. Green not (only) for profit: an empirical examination of the effect of environmental-related standards on employees' recruitment. Resource and Energy Economics 34, 74–92.

Hansla, A., Gamble, A., Juliusson, A., Gärling, T., 2008. The relationships between awareness of consequences, environmental concern, and value orientations. Journal of Environmental Psychology 28, 1–9.

Hines, J.M., Hungerford, H.R., Tomera, A.N., 1986/1987. Analysis and synthesis of research on responsible environmental behavior: a meta-analysis. Journal of Environmental Education 18, 1–8.

Hough, L.M., Muschewske, R.C., 2012. Implications for HR and OD professionals in the future: can I-O principles and practices contribute to the achievement of corporate environmental sustainability? In: Jackson, S.E., Ones, D.S., Dilchert, S. (Eds.), Managing Human Resources for Environmental Sustainability. JosseyBass/Wiley, San Francisco, CA, pp. 407–415.

IPCC, 2013. Climate Change 2013: The Physical Science Basis (Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change). Cambridge University Press, Cambridge, United Kingdom and New York.

IPCC, 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects (Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change). Cambridge University Press, Cambridge, UK and New York.

Jones, D.A., Willness, C.R., Madey, S., 2014. Why are job seekers attracted by corporate social performance? Experimental and field tests of three signal-based mechanisms. Academy of Management Journal 57, 383–404.

Kaiser, F., Schultz, P.W., 2009. The attitude-behavior relationship: a test of three models of the moderating role of behavioral difficulty. Journal of Applied Social Psychology 39, 186–207.

Klein, R.M., May 2014. Employee Motives for Engaging in Environmentally Sustainable Behaviors: A Multi-study Analysis (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Klein, R.M., D'Mello, S., Wiernik, B.M., 2012. Demographic characteristics and employee sustainability. In: Jackson, S.E., Ones, D.S., Dilchert, S. (Eds.), Managing Human Resources for Environmental Sustainability. Jossey-Bass/Wiley, San Francisco, CA, pp. 117–154.

Maloney, M.P., Ward, M.P., 1973. Ecology: let's hear from the people. An objective scale for the measurement of ecological attitudes and knowledge. American Psychologist 28, 583–586.

Markowitz, E.M., Goldberg, L.R., Ashton, M.C., Lee, K., 2012. Profiling the "pro-environmental individual": a personality perspective. Journal of Personality 80, 81–111.

Mayer, F.S., Frantz, C.M., 2004. The connectedness to nature scale: a measure of individuals' feeling in community with nature. Journal of Environmental Psychology 24, 503–515.

Mesmer-Magnus, J., Viswesvaran, C., Wiernik, B.M., 2012. The role of commitment in bridging the gap between organizational sustainability and environmental sustainability. In: Jackson, S.E., Ones, D.S., Dilchert, S. (Eds.), Managing Human Resources for Environmental Sustainability. Jossey-Bass/Wiley, San Francisco, CA, pp. 155–186.

Milfont, T.L., Sibley, C.G., 2012. The Big Five personality traits and environmental engagement: associations at the individual and societal level. Journal of Environmental Psychology 32, 187–195.

Ones, D.S., Dilchert, S., 2012a. Employee green behaviors. In: Jackson, S.E., Ones, D.S., Dilchert, S. (Eds.), Managing Human Resources for Environmental Sustainability. JosseyBass/Wiley, San Francisco, CA, pp. 85–116.

Ones, D.S., Dilchert, S., 2012b. Environmental sustainability at work: a call to action. Industrial and Organizational Psychology: Perspectives on Science and Practice 5, 444–466.

Ones, D.S., Dilchert, S., 2013. Measuring, understanding, and influencing employee green behaviors. In: Huffman, A.H., Klein, S.R. (Eds.), Green Organizations: Driving Change with I-O Psychology. Routledge, New York, pp. 115–148.

Orlitzky, M., Schmidt, F.L., Rynes, S.L., 2003. Corporate social and financial performance: a meta-analysis. Organization Studies 24, 403–441.

Osbaldiston, R., Schott, J.P., 2012. Environmental sustainability and behavioral science: meta-analysis of proenvironmental behavior experiments. Environment and Behavior 44, 257–299.

SAM, PriceWaterhouseCoopers, 2011. 2011 Sustainability Yearbook. Retrieved from: https://www.pwc.ch/user_content/editor/files/publ_adv/pwc_sustainability_yearbook_2011_e.pdf.

Schultz, P.W., 2000. Empathizing with nature: the effects of perspective taking on concern for environmental issues. Journal of Social Issues 56, 391–406.

Stern, P.C., Dietz, T., 1994. The value basis of environmental concern. Journal of Social Issues 50, 65–84.

Viswesvaran, C., Ones, D.S., 2000. Perspectives on models of job performance. International Journal of Selection and Assessment 8, 216–226.

Wackernagel, M., Schulz, N.B., Deumling, D., Linares, A.C., Jenkins, M., Kapos, V., Randers, J., 2002. Tracking the ecological overshoot of the human economy. Proceedings of the National Academy of Sciences 99, 9266–9271.

Wiernik, B.M., Ones, D.S., Dilchert, S., 2013. Age and environmental sustainability: a meta-analysis. Journal of Managerial Psychology 28, 826–856.

Wiernik, B.M., Ones, D.S., Dilchert, S., Biga, A., April 2011. Age Differences in Green Work Behaviors across 11 Countries (Poster presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL).

Wiseman, M., Bogner, F.X., 2003. A higher-order model of ecological values and its relationship to personality. Personality and Individual Differences 34, 783–794.

Professional Development

Liz Beddoe, School of Counselling, Human Services and Social Work, Faculty of Education, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In social work as in many professions, the learning process does not end after initial training leading to qualification or licensing. Professional development (PD) is a career-long process within which, via diverse learning activities, professionals refine and augment their knowledge and skills and undertake personal professional supervision to enhance critically reflective practice. PD is held to be vital in practice remaining competent, ethical, and optimistic. These elements of professional practice while seemingly virtuous are not innocent of the political and organizational contexts in which they occur. This article critically reviews PD, reflective practice, supervision in social work, and the role of higher education and research.

Continuing Professional Development

It is generally held that all professions need to ensure their members remain well informed about new knowledge and policy and are able to demonstrate continuing capability in core practice requirements. Some form of professional development (PD) or continuing professional education (CPE) is considered a vital mechanism to ensure that the profession is able to respond to the challenges of rapid change. In social work, such change includes the influence of world and local events on social problems, the challenges of international mobility, and the needs of new and changing communities.

In many nations, increased emphasis on quality assurance legislation has provided impetus to develop mandates for ongoing professional education and development, for example, where registration and licensing bodies set requirements for supervision and PD linked to eligibility for renewal of an annual practicing certificate. These legislative features encourage employing organizations to pay closer attention to the ongoing development of their professional workforce. While social workers are registered or licensed as individuals, the pragmatic world of health and social services compliance criteria, dictated by purchase agreements with funders, will ensure that some provision will be made. These legislative mandates may provide support for resources for continuing development but also shift the impetus for PD and CPE away from the individual practitioner (and her or his personal professional accountability) toward a more managed approach within organizational settings. This shift renders PD and CPE part of managerial control; ongoing education becomes an outcome, not a process (Beddoe, 2009).

CPE is both an individual and a corporate investment, and economic policy has promulgated the imperative of lifelong learning and its broad social and economic benefits (Beddoe, 2009). Current requirements of the quality assurance and regulation add pressure on employers to provide resources to support for social workers' continuing professional development (CPD). Table 1 summarizes the many elements of time and focus across PD and CPE.

A functionalist account of PD and CPE, focusing on compliance, provides only a partial and inadequate explanation of the significance of career development to many social workers and their organizations. In particular, higher educational qualifications have symbolic value for social workers as they help to mark out status within the profession (Beddoe, 2013). Jarvis (1996) argues that one of the consequences of the greater demarcation of professional workers is the need for recognition of being up-to-date which "connotes evolving expertise in a world in which practitioner knowledge is quickly rendered out of date by the fast pace of research and scholarship" (Dirkx, 2006: p. 37). The possession of a higher qualification may 'add value' for what it symbolizes rather than what it entails. Similarly, as continuing education becomes more closely linked to workplaces, many higher degrees are now available for part-time enrollment and are linked back to

Table 1 Time and focus of professional development (PD) and continuing professional education (CPE)

	Elements
Time	PD follows the period of preservice education undertaken by professionals in which they achieve a first qualification with eligibility for entry to their profession
	PD extends professionals' learning in the workplace
	PD is ideally ongoing throughout professional careers
	PD enables professionals to update with new knowledge and new practices over the course of a career
	PD supports professionals' progress from beginning to mature practitioners
	CPE may be integral to career development and promotion
Focus	PD maintains and enhances practitioners' individual competence
	PD includes participation in professional (clinical) supervision
	PD may be structured or unstructured, formal, and informal
	PD is generally separated from the main activities of work
	PD is often located within the workplace
	PD often delivers codified knowledge through 'best practice' guides
	CPE is generally found in an external institution
	CPE encourages intellectual engagement with knowledge utilization and creation
	CPE in higher education fosters a spirit of critical inquiry

the work context through practical work-based components (Jarvis, 1996). The process of credentialing can influence individual practitioners as they consider their value in the market. In the current era, social work is not dissimilar to other professions – it is subject to the impact of globalization, shortages or gluts of particular kinds of professionals, international markets for education and professional skills, and increasing specialization within broad occupational categories (Jarvis, 1996).

Consideration of PD for social work cannot occur without recognition of the complex social, political, and organizational contexts in which the profession is practiced. We live in an era where social work has been challenged in terms of its methods and foundations of knowledge. Lorenz (2012: p. 496) argues that the current ideological context in which social work struggles to retain legitimacy:

> is characterized by two core criteria: efficiency of procedures (exemplified by the rise of managerial approaches to service delivery systems) and individualization in the form of privatizing social responsibility and activating individual members of society to secure their own livelihood and future.

It is within the managed professional environment that Lishman (2009) suggests that there are two contrasting contemporary models influencing PD in social work. The first is a competency model, which is frequently developed in a top–down approach and reflects political and organizational rather than professional imperatives and values. Competency models are frequently dominated by instrumental and codified procedural knowledge and thus deliver what is felt to be known 'best practice' to practitioners rather than the skills for independent and critical inquiry. In the second approach, Lishman suggests a learning-based reflective practice approach (Fook and Gardner, 2007; Redmond, 2004), which incorporates "evidence-based and research-minded practice" (Lishman, 2009: p. 374). The problem is that the drivers for reflective practice and evidence-based practice (EBP) are quite different. The drive toward a positivist approach to support for social work activity remains highly contested and is frequently held to be linked to the contemporary agenda of efficiency and steering the privatization of responsibility for welfare highlighted by Lorenz above. For social work, the call for EBP has shaped much of the debate around the ongoing intellectual and PD of social workers.

Within the neoliberal regime, which characterizes many Western nations, social work is called "to prove its willingness to play a part in the restructuring of social solidarity systems from an essentially collectivist perspective to an individualized approach" in order to "secure recognition as a useful management tool" (Lorenz, 2012: p. 497). Lorenz argues that if social work aims to retain a critical stance toward the restructuring of the welfare state "then it becomes unwise to meet this challenge with the epistemological instruments derived from the positivist tradition that, from the outset, excludes references to the political context" (p. 497).

CPD in social work is thus not innocent of ideological elements and can be a site of struggle between the aspirational claims of the profession (including education) and the more pragmatic preoccupations of policy makers and managers (Beddoe, 2009, 2011, 2013). The increasing regulation of social work often brings these tensions to the foreground and it has been argued that regulation can be a Trojan horse, leading to more governmental shaping of social work education and supplanting the critical analysis that underpins social work in higher education (van Heugten, 2011).

This article will explore several significant aspects of contemporary PD in social work, such as reflective practice, the purpose and benefits of supervision in social work, and the role of continuing education in fostering social work scholarship and practitioner contributions to research. This journey will commence with a brief exploration of reflective practice, which has been widely adapted by many service professions as an essential component in ensuring professional competence (Kilminster et al., 2010) and has become central to the practice of supervision.

Reflective Practice

Kilminster et al. (2010) note that although reflective practice was initially developed as a way to address problems in professional practice, it has been adapted by educators and professional bodies as a mainstream practice. They argue that this process has pulled reflective practice away from the margins and in doing so reduced its critical and transformative dimensions. The development of 'critical reflection' to be briefly explored below is an important countering response within continuing education and development for social work.

The seminal contributions of Donald Schön (1983) underpin reflective practice. Schön stated that the knowledge base of a profession has four important properties: "it is specialized, firmly bounded, scientific and standardized" (1983: p. 23). The focus is on problem-solving. However, in the often quoted "swampy lowlands" (p. 42) of messy human service practice, Schön notes that the nature of the problems, the 'who' and 'what' of problem definition, and decision-making are highly contentious (p. 41). Effective practice thus requires more than top–down 'expert' intervention. In his influential work 'The Reflective Practitioner' (1983), he reported on an exploration of how professional practitioners worked in action, noting that those with reflective ability shared some common attributes. They were able to see each individual service user and his or her circumstances as unique, were accessible and flexible, and able to respond to unexpected concerns or behaviors in the moment or 'in-action.' Highly skilled professionals were not dominated by reliance on codified knowledge and rubrics based in what he called 'technical rationality' – ways of knowing grounded in science and procedure (1983: p. 21). Domination of practice by expert knowledge and tools can alter the balance of art and science now commonly understood as necessary in the helping professions. More reflective practice is flexible and essentially collaborative based in sensitive and enabling relationships. This work and the contributions of more recent scholars of reflective practice underpin much of current practice wisdom in social work.

Fook and Gardner (2007) have made a major contribution to reflective practice in clearly linking reflective practice to the core aims of social work as a profession grounded in social

justice and human rights, via the development of critical reflection. Clearly differentiated from more neutral notions of reflection, with 'critical reflection,' Fook and Gardner (2007) see a reflective approach as affirming in practice other ways of knowing, such as the role of personal experience and its interpretation in supporting a holistic understanding. Fook and Gardner (2007: p. 25) note the complexity of experience that practitioners encounter in their day-to-day work and the significance of a holistic approach that focuses on "the whole experience and the many dimensions involved: cognitive elements, feeling elements, meanings and interpretations from different perspectives." Such an approach facilitates the discovery of the kinds of knowledge relevant to the unpredictability and anxieties of modern-day practice. Fook and Gardner's critical reflection approach (p. 27) is "a model for improving practice," which involves: "A bottom-up understanding of theory and practice; exposing problematic gaps between espoused and enacted theory," closing the gap and placing value on "intuition, artistry, and creativity," and adopting a "contextual, holistic, and experiential approach." Such an approach seeks to uncover and address unhelpful assumptions and in doing so contribute to critical practice knowledge.

As reflective practice has become central in professional activities focused on the development and maintenance of competent practice, supervision is often identified as a site where it is fostered and practiced.

Supervision

Supervision is a core component of PD for social workers. Supervision is generally held to provide a major forum for facilitating reflective practice and learning. It can be defined as an interactive dialogue between at least two people, one of whom is a supervisor. This dialogue

> shapes a process of review, reflection, critique and replenishment for professional practitioners. Supervision is a professional activity in which practitioners are engaged throughout the duration of their careers regardless of experience or qualification. It is accountable to professional standards and defined competencies and to organisational policy and procedures.
>
> Davys and Beddoe, 2010: p. 21

It is generally held that professional supervision has been a core social work practice since the development of social work during the latter part of the nineteenth century with an initial focus on the administrative aspects (Tsui, 1997). There is a broad agreement about the three main functions of supervision being education, administration, and support. Hughes and Pengelly (1997: p. 42) proposed a triangular model where the three functions of supervision are managing service delivery, facilitating practitioners' PD, and focusing on practitioners' work. This model demonstrates the tensions inherent within supervision in the turbulent social work environment, where conflicts may be experienced as these functions compete for time and attention. A significant feature of Hughes and Pengelly's model compared with earlier functional models is the rejection of support as a function of supervision, with the authors asserting rather that support is a core condition. It is the supervision relationship in which support and trust are nurtured that creates the environment for honest and critical exploration of practice challenges.

In the 1990s, concern was expressed that supervision had been reduced to an audit and compliance tool, in an increasingly performative culture in public services social work (Payne, 1994). It was not uncommon for supervision to virtually disappear from its roots in an educative, reflective practice. Supervision has been restored to a central place in social work practice in recent times as it is frequently identified in major reviews of social work as contributing to good practice outcomes and ensuring ongoing practitioner well-being and retention.

Supervision in the Organizational Context

The status of supervision encompasses elements of paradox in the present era. Recently, supervision has entered a new phase of prominence as three main factors have led to changes in the wider social work context : first, the neoliberal preoccupation with performance management and second, and interrelated, the impact of 'the risk society' and the associated public challenges to professional practice (Davys and Beddoe, 2010). Fear of service failure, concern for public safety, and a deep fear of public criticism on the part of both regulatory bodies and governments have encouraged supervision to be harnessed as a compliance measure. The third factor, the promotion of supervision as a mechanism for securing social workers' well-being and retention in the profession, reflects the other side of the surveillance coin, the specter of organizational anxiety about the safety of workers.

While there are many aspects of the retrieval of supervision from a purely managerial space, along with recent expansions of function as a well-being measure, supervision seems ostensibly a benign practice; however, it remains a contested space. Adamson (2011: p. 185) argues that supervision is "not 'politically innocent' and subject to wider influences." Adamson (2011: p. 186) argues for a broad distinction between 'micro' and 'macro' issues in supervision with the former focused on "intra- and interpersonal concerns, the ethics of relationships, and quality-control issues of best practice." The 'macro' focus of supervision practice is on workforce and organizational concerns.

Centrality of the contextual dynamics of supervision will always be vulnerable to policy swings and practice trends and thus subject to different, unreliable interpretations by practitioners and managers. The linking of supervision and managerial surveillance is not entirely put to rest.

Over recent decades, new public management practices have had considerable impact on health and social work services with high-profile disasters and tragedies ushering in a plethora of performance measures, audits, and quality assurance systems. An impact of the preoccupation on quality assurance has been an undue emphasis on the administrative functions of supervision and a consequential shift away from the in-depth review and reflective critique of practice. Jones (2004: p. 12) described it as a 'trade-off' between managerial and professional concerns. Within a risk-averse political environment in social work, the oversight of practice

decisions can create stress and anxiety. Supervisors may fear simple mistakes that threaten reputations, to the specter of serious harmful events for service users, patients, colleagues, or the public (Beddoe, 2010). In highly surveillant, risk-averse environments, anxiety can stifle innovation and diminish creative and collaborative practice.

Within the third factor identified above, organizational concern about the well-being of workers has provided a strong argument to support more creative, flexible, and less compliance-driven approaches to supervision. In the past 10 years, supervision has become subject to increased scrutiny to examine its effectiveness. Mor Barak et al. (2009: p. 25) conducted a *meta*-analysis of research on supervision and noted that various dimensions of supervision "may have protective, proactive, or preventive roles in ensuring a positive work environment that can contribute to worker effectiveness." These dimensions include supervisory task assistance, social and emotional support, and supervisory interpersonal interaction (Mor Barak et al., 2009: p. 26). In the United Kingdom, a recent review of the research on supervision by Carpenter et al. (2012) found evidence that supervision is associated with job satisfaction and protects against stress, and "practitioners should insist that good supervision be provided by their employers" (p. 16). This review confirmed support for supervision that goes beyond tasks and compliance.

Adamson (2011) argues that the best responses to complex practice situations are found in attention to reflective practice, structural analysis, critical thinking, and ethically sensitive behavior. Supervision can create a space to create a less certain space, where assumptions – moral, personal, professional, or political can be challenged and alternative explanations and possible solutions developed. Such deeper exploration in turn may promote more effective management of emotions by making space for positive reappraisal of complex situations and workers' emotional responses (Adamson, 2011).

Supervision and Reflective Practice

As supervision has been revitalized within the profession, it has become subject to greater scholarship and development. Accordingly, supervision is promoted again as a core reflective practice within the profession's mission, bringing personal, professional, organizational, and political elements into a reflective space. Supervision contributes to the strengthening of professional identity in social work, to 'convey the mission and vision' of social work (Tsui, 2005: p. 11) and can bring the examination of culture, gender and class politics, discrimination, and institutional racism to the foreground (Hair and O'Donoghue, 2009). Excellent supervisors do not merely tick boxes in compliance reporting but are thinkers who have the ability to hold and explore the anxieties of practice while retaining a critical edge in stimulating supervision. Supervision is best conceptualized as an intellectual activity. How then to achieve all these lofty aims in supervision? At the 'nuts-and-bolts' level, a process needs to be learned.

An example of a reflective learning model of supervision combines an experiential learning approach with a contractual process of exploration within the supervision process (Davys and Beddoe, 2010). This reflective learning model is a cyclic process within a set structure with tasks for the supervisor and supervisee. It consists of four main stages, which the supervisor and supervisee go through when working on an issue brought to supervision. The session itself is bounded by beginning and ending processes that enable safe learning, the valuing of the process, and the transfer of learning into subsequent practice. The four main stages encompass the experiential learning cycle and are named 'the event'; 'exploration, including impact and implications'; 'experimentation'; and finally 'evaluation' (Davys and Beddoe, 2010: pp. 93–105). **Table 2** summarizes the key tasks and processes of this model.

Thoughtful and reflective supervision, well resourced and honored with education and specific PD, has a significant continuing contribution to make within the profession. As a practice, supervision needs to be further researched and developed so it becomes more of a reality and less aspirational.

Higher Education

Social work education varies widely around the globe. In some countries, it has been offered almost exclusively at undergraduate level with 3- or 4-year specialist degree as the standard path to entry to the profession. However, in many nations over the past decade or so, there has been shift toward a much greater provision of social work education at postgraduate level, as has been common in the United States. It is argued that social work must be a research-led profession and thus postgraduate study is needed, achieved by the inclusion of an honors year of study, including application of research methods and a dissertation. The graduate school model is also growing in strength where graduates in social sciences and humanities enter a master's program that embeds theory, clinical practice, and research. The increased focus on research literacy is laudable, but the focus remains squarely on educating social workers to be consumers and translators of research rather than producers. This is not a healthy state for any profession; given that the creation of a body of knowledge for practice is an essential component of professional status. Higher degrees by research lead to a greater emphasis on knowledge production and scholarly dissemination.

It seems probable that PD for social work will increasingly focus on research, whether this encompasses 'in-service' training to update people with distilled research findings or requiring new graduates to enter postqualifying education reasonably early in their careers. This latter possibility will largely depend on governments' willingness to invest in social work. Major reviews of practice in England, for example, have recommended that higher education plays a major role in CPE for practitioners including the development of postgraduate programs in child protection studies, but Moriarty and Manthorpe (2013) note that changes in the regulatory body of social work may mean a shift away from the idea of a nationally regulated framework of higher qualifications. Instead, a new approach will encourage employers and universities to reach agreements about the content and delivery of CPD and CPE. In many countries in the current economic climate, access to higher education becomes more expensive and many practitioners are reliant on the willingness of employers to support their undertaking higher degrees. In this climate, signals support the continuation of fairly pragmatic

Table 2 The reflective learning model for supervision

Beginnings	The start of the session involves two key tasks. First, the creation of the space for supervision and ensure both parties are fully present and able to attend the work of supervision. Second, the agenda is set and items prioritized.
Stage 1: The event	In this stage, the experience or event is identified and the story surrounding it retold, which will help clarify the goal for the session. Once explicit and realistic goals are determined, the session can move on to the next stage of exploration. This process encourages the supervisee to be an active participant and take responsibility for the supervision outcome. During supervision, the supervisor will assist to retain purpose and direction toward reaching the goal.
Stage 2: Exploration	During this stage, the task to explore the issue brought to supervision. This is not a time for the supervisor to prove her/himself by providing the answers. Rather clarification of an issue is strengthened by ownership on the part of the supervisee. The role of the supervisor is to assist in the exploration of the issue from different perspectives, some of which have not previously been considered and will help the supervisee arrive at a new understanding or decision. In order to balance the emotional and personal elements of the supervisor issue with cognition, the model divides stage 2 into two substages: *Impact-reflection*: Here the event is explored in relation to its impact on the personal 'self' – feelings, actions, beliefs, values, and links to previous behaviors. It is here that a supervisee may come to identify the emotional impact of the work and there is opportunity to vent and explore emotion in a safe environment. *Implications – conceptualization*: Following an adult-learning framework, emotional expression needs to be balanced with exploration of the cognitive aspects of the role. This involves application of an intellectual framework that assists the supervisee to integrate theory, standards of practice, ethics, service policies, and protocols with their professional self. By doing so, new understandings and strategies may be discovered.
Stage 3: Experimentation	This is the stage to test out whether solutions arrived at during the exploration are viable, examine potential challenges, and develop a contingency plan. This stage is vital to the success of implementing the changes or strategies, which have been identified into practice.
Stage 4: Evaluation	This relates to evaluating whether the specific issue has been addressed and if the reflective learning process has achieved its purpose. The stage does not signify the end of the entire supervision session. The cycle may be repeated for additional supervision issues.
Conclusion	Finally, the supervision session will end with time for the session to be considered as a whole, both the content and the reflective learning process. It is an opportunity for giving and receiving feedback between the supervisor and the supervisee and to make arrangements for future sessions.

top–down-driven CPD, rather than transformative education that is driven by critical inquiry (Beddoe, 2009).

In the international arena, despite some discouraging trends, there has simultaneously been growing emphasis on EBP in social work, which may offer a positive pathway to greater social work scholarship. In many Western countries, this development can be linked to the 'modernization' agenda of governments demanding standardized interventions and greater professional accountability. The EBP movement has advanced a strong argument that social work practice be informed and developed from the results of empirical research. This trend is not without critique and again can focus on distillation and delivery of codified knowledge rather than fostering scholarly work by practitioners. We should not be seduced by uncritical claims that EBP will solve all the problems of social work and its service users. Criticism of EBP includes argument that it privileges quantitative research; is based on a flawed understanding of social science; silences service users; and fails to adequately account for the ambiguities and complexities of social work practice (Beddoe, 2011). Social workers, it is argued, can be much more context-sensitive and aligned to the moral mission of social work and thus report the unique feature of service users' worlds. Participatory action research and collaborative research undertaken by practitioners can also bring new knowledge to policy development and associated practice improvement.

Research and Scholarship in Social Work Practice

Bearing in mind the cautions raised above, perhaps one of the most positive features in the contemporary social work PD landscape is the discussion taking place about the importance of practitioner research. This is happening in many countries around the world, and Lunt et al. (2011) report that there are many attempts to promote practitioner research, which extends further than practitioner studies carried out by individuals but includes collaborations and projects supported by academic institutions and organizations. While there may be many diverse aims behind these developments, a common goal is to encourage social workers to move beyond passive relationships with research and take up active roles in generating and utilizing research in practice contexts (Lunt et al., 2011).

It is hard to quantify practitioner research as much may not find its way into traditional published outlets. Practitioner research is more likely to involve small, highly local studies of an informal nature. Social work practitioners are often motivated to study existing data that are collected routinely in service organizations with a view to better understand the needs and experiences of service users in order to improve the responsiveness and quality of services offered. Lunt et al. (2011: p. 188) note that practitioner projects are often time-limited with a variety of research designs and methodology with a focus on evaluation and that such characteristics are common, irrespective of whether the drive arises from 'tertiary study, agency imperative, or practitioners themselves.'

A survey of the literature identifies many challenges to social workers in participating in more active forms of research activity (Beddoe, 2011). Most commonly cited barriers are matters of time, confidence, knowledge, and skill; skepticism that social workers need to do more than just process

information; access to extant research and literature and the practical resources needed to conduct projects. It is frequently noted that practitioners struggle to set aside and maintain time for research, as metric systems of allocation and timeliness of casework demand their attention.

The most difficult barriers, however, are likely to be those that have their foundations in those arising from the relatively fragile position of social work in professional hierarchies and the concomitant weak knowledge claim (Beddoe, 2011, 2013). Lunt et al. noted that professional "self-image may act as a barrier to practitioner research with social workers viewing themselves as helpers rather than intellectuals" (2011: p. 188).

It is an ethical and moral imperative for social work to make a contribution to the development and utilization of knowledge. This requires the development of social work scholarship and research especially in relation to the many enduring problems of our age. Critical inquiry can enable social workers to make use of knowledge gained in everyday practice to advocate both for service improvements and the removal of the numerous obstacles to full economic and social participation for peoples and communities. Practitioner scholarship utilizes the well-honed social work skills of observation, analysis, intervention, and review already demonstrated by highly competent practitioners to provide a window on the experience of many communities.

Conclusion

What then is the likely future of social work professional development? In the current recessionary era, it seems likely that social worker involvement in higher degrees by research will grow rather slowly, but as a profession in the grip of the professional project, social workers will seek to raise their professional capital through achieving higher degrees and undertaking scholarship (Beddoe, 2013). Advances in information and communication technology continue to grow and social media is likely to play a much greater role in supporting CPD but research and investment in such development is in its infancy in many countries. Further research is needed to explore whether social workers will find social media helpful (Moriarty and Manthorpe, 2013). The current debates about access to peer-reviewed research are highly relevant to social work. Much social work research is locked away behind commercial publishers' 'pay walls' and practitioners are often unable to access journals without their employer's assistance. Open-access journals, blogs, Web-based curated content already demonstrate great potential linked through specialist groups in social media. A highly information-literate graduate workforce will undoubtedly lead the charge for continuing access to up-to-date resources and opportunities to participate in research and development.

See also: New Managerialism and New Public Sector Management; Research and Theory Building in Social Work; Social Work Education; Social Work in Organizations.

Bibliography

Adamson, C., 2011. Supervision is not politically innocent. Australian Social Work 65 (2), 185–196.

Beddoe, L., 2009. Creating continuous conversation: social workers and learning organizations. Social Work Education 28 (7), 722–736.

Beddoe, L., 2010. Surveillance or reflection: professional supervision in 'the risk society'. British Journal of Social Work 40 (4), 1279–1296. http://dx.doi.org/10.1093/bjsw/bcq018.

Beddoe, L., 2011. Investing in the future: social workers talk about research. British Journal of Social Work 41 (3), 557–575.

Beddoe, L., 2013. Continuing education, registration and professional identity in New Zealand social work. International Social Work. http://dx.doi.org/10.1177/0020872812473139 (published online on 5 June 2013).

Carpenter, J., Webb, C.M., Bostock, L., Coomber, C., 2012. Effective Supervision in Social Work and Social Care. Research Briefing 43. Social Care Institute for Excellence, London.

Davys, A., Beddoe, L., 2010. Best Practice in Supervision: A Guide for the Helping Professions. Jessica Kingsley, London.

Dirkx, J.M., 2006. Studying the complicated matter of what works: evidence-based research and the problem of practice. Adult Education Quarterly 56 (4), 273–290.

Fook, J., Gardner, F., 2007. Practising Critical Reflection: A Resource Handbook. Open University Press, Maidenhead, UK.

Hair, H.J., O'Donoghue, K., 2009. Culturally relevant, socially just social work supervision: becoming visible through a social constructionist lens. Journal of Ethnic and Cultural Diversity in Social Work 18 (1), 70–88.

van Heugten, K., 2011. Registration and social work education: a golden opportunity or a Trojan horse? Journal of Social Work 11 (2), 174–190.

Hughes, L., Pengelly, P., 1997. Staff Supervision in a Turbulent Environment – Managing Process and Task in Front-line Services. Jessica Kingsley, London.

Jarvis, P., 1996. Continuing education in a post-modern or global society: towards a theoretical framework of comparative analysis. Comparative Education 32 (2), 233–244.

Jones, M., 2004. Supervision, learning and transformative practices. In: Gould, N., Baldwin, M. (Eds.), Social Work, Critical Reflection and the Learning Organisation. Ashgate, Aldershot, pp. 11–22.

Kilminster, S.M., Zukas, M., Bradbury, H., Frost, N., 2010. Introduction. In: Bradbury, H., Frost, N., Kilminster, S.M., Zukas, M. (Eds.), Beyond Reflective Practice: New Approaches to Professional Lifelong Learning. Routledge, Abingdon, UK, pp. 1–14.

Lishman, J., 2009. Personal and professional development. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates, third ed. Palgrave, Basingstoke, pp. 370–385.

Lorenz, W., 2012. Response: hermeneutics and accountable practice: lessons from the history of social work. Research on Social Work Practice 22 (5), 492–498.

Lunt, N.T., Ramian, K., Shaw, I., Fouché, C., Mitchell, F., 2011. Networking practitioner research: synthesising the state of the 'art'. European Journal of Social Work 15 (2), 185–203.

Mor Barak, M.E., Travis, Dnika J., Pyun, H., Xie, B., 2009. The impact of supervision on worker outcomes: a meta-analysis. Social Service Review 83 (1), 3–32.

Moriarty, J., Manthorpe, J., 2013. Post-qualifying education for social workers: a continuing problem or a new opportunity? Social Work Education, 1–15.

Payne, M., 1994. Personal supervision in social work. In: Connor, A., Black, S.E. (Eds.), Performance Review and Quality in Social Care. Jessica Kingsley, London, pp. 43–58.

Redmond, B., 2004. Reflection in Action: Developing Reflective Practice in Health and Social Services. Ashgate, Aldershot.

Tsui, M.S., 1997. Empirical research on social work supervision: the state of the art (1970–1995). Journal of Social Service Research 23 (2), 39–54.

Schön, D., 1983. The Reflective Practitioner. Temple Smith, London.

Tsui, M.S., 2005. Social Work Supervision:Contexts and Concepts. Sage, Thousand Oaks.

Relevant Websites

http://www1.uwindsor.ca/criticalsocialwork/ (accessed 14.07.13.) – Critical Social Work – Open Access Journal.

http://ifp.nyu.edu/ (accessed 14.07.13.) – Information for Practice.

http://www.jswve.org/ (accessed 14.07.13.) – Journal of Social Work Values and Ethics.

http://www.scie-socialcareonline.org.uk (accessed 14.07.13.) – Social Care Institute for Excellence: Social Care on Line.

Professions and Professionalization, History of

Hannes Siegrist, Universität Leipzig, Leipzig, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Professions are characterized by special functions, strategies, structures, forms of knowledge, and sociocultural styles. The history of the professions investigates the ways in which particular knowledge-based occupations are institutionalized and organized in different times and contexts. The professions include the occupations of doctor, lawyer, professor, and engineer, among others. Professionalization refers to programs, projects, and processes affecting the social and symbolic construction of professions. The comparative history of the professions shows that paths, cycles, and types of professionalization may vary. Teleological understandings of professionalization and static conceptions of the professions are thus relativized.

Introduction

Professions are characterized by special functions, strategies, structures, forms of knowledge, and sociocultural styles. Professionalization refers to programs, strategies, and processes affecting the social and symbolic construction of occupation and status. In modern historical scholarship, the professions include the occupations of doctor, lawyer, professor, and engineer, among others. The history of the professions investigates the ways in which functional groups and status groups are institutionalized and organized in different times or historical constellations and contexts; and how the pattern of a profession is diffused, emulated, and adapted in processes of modernization, nationalization, and globalization. It shows how and why professions, in the restricted sense of the term, differ from the status groups and occupations of entrepreneur, merchant, artist, clerk, skilled manual worker (crafts, *métiers*) and the so-called semi-professions. Professionalization involves the formation of an occupation, on the one hand, and interrelated developments regarding the social division of labor, structures of authority, and sociocultural inequality, on the other.

Historical research on professionalization concentrates on such issues as the distribution of knowledge, influence, wealth, and power at a particular point in time; the social definition of behavioral prerogatives; the regulation of central values and functions; and forms of social and cultural exclusion. It is concerned with the motives, interests, and strategies of actors who either promote or hinder processes of professionalization. Recent historical research on professionalization has tended to oscillate between individualizing historical perspectives and systematic social science perspectives. The former emphasizes the peculiarities of particular professions or the special features of the professions in a particular historical context. The latter concentrates on general features, similarities, and convergences in time and space. International comparative research shows that while the professions are often concerned with the same or similar functions, problems, and tasks, these may be interpreted in different ways and institutionalized in different forms. Similarly, depending on the characteristics of a particular profession and on the particular historical circumstances, the paths, cycles, and types of professionalization and deprofessionalization may vary. By taking these points into consideration, historians have relativized both teleological understandings of professionalization and static or functionalist conceptions of the professions.

Definitions of Profession and Professionalization

Abstract models and temporally undifferentiated concepts of the professions have a heuristic value in historical research. Historians of the professions usually begin with a historical explication of the concept and then ask how ideas and sociocultural patterns are articulated in a specific context with regard to the relevant dimensions in particular domains. The individual dimensions and processes are more or less closely linked in concrete historical cases. The result is that the professions and professionalism find expression in a graded series of variations.

The professions are characterized ideally by the following series of interrelated dimensions: Capabilities and skills are justified scientifically or systematically. Professional knowledge is described in such terms as 'exclusive,' 'more profound,' 'inaccessible to lay persons,' and 'not easily understandable.' It is acquired in special institutions such as universities, professional schools, and internships with professionals. It forms the basis of the superiority of the professional over lay persons and clients. Also included in this body of knowledge are rules and attitudes which govern its application in a way which promotes trust between the professions, their clients, and their social environment. These include formal procedures, rituals, titles, a professional habitus of studiousness, collegiality, and altruism, and a general orientation toward the common good. The professions claim for themselves a specific style of earning a living that is characterized by the principle of remuneration 'befitting ones rank' and through specific forms of contract and recompense. The professions demand for themselves higher prestige. They develop specific organizations for collective integration, self-administration, self-control, and for the representation of their interests, such as the 'chartered professions' in England, the disciplinary councils and professional chambers on the European continent, and, more generally, professional associations of various kinds. The professions strive for 'autonomy.' Viewed historically, this means, first, exiting from relations of patronage with traditional secular and

clerical elites and, second, the reduction of external control and dependence on the state, on the market, on the rich and powerful, on clients or beneficiaries, and on employers.

Professionalization is a social project, which articulates itself in discourses on science, performance, honor, public welfare, and society and in the ideology of professionalism. It is intended to secure for professionals a high or midrange status in the system of social and cultural inequality.

The historical literature distinguishes between professionalization that is externally controlled and professionalization that is controlled by the occupational groups themselves. Similarly, it distinguishes between heteronomous and autonomous professions. In the classical research on the 'free' or 'liberal' professions, especially in the Anglo-Saxon literature, professionalization is understood largely from the perspective of members of the occupational groups themselves. In recent research, however, the analytical concept of professionalization is being used increasingly for professionals who are occupied in state and private administrations and enterprises as well (Burrage and Torstendahl, 1990; Perkin, 1989; Siegrist, 1988). Previously, and especially on the European continent, scholarly interest in professions was often guided by concepts such as 'higher education,' 'meritocracy,' 'bureaucracy,' and 'civil servants.' Recent research about artists, authors, and creative workers in the modern cultural industry and information business, discusses whether they conceive of themselves as professionals, emulate professionalistic patterns, and rely on strategies of formal professionalization (McClelland, 2003; Müller-Jentsch, 2011).

The historical connection between occupational role and social status is articulated concretely in concepts such as the 'gentlemanly professions,' the 'learned professions,' or, more simply, the 'professions' in England and the United States; *professions libérales*, *bourgeoisie à talents*, *capacités* in France; *gelehrte Stände*, *höhere Berufe*, *bürgerliche Berufe*, *mittelständische Berufe*, *Akademiker*, and *Bildungsbürgertum* in the German-speaking lands of Europe; *professioni maggiori*, *professioni nobili*, and *borghesia unmanistica* in Italy; 'intellectuals' in Western Europe; and *intelligentsia* in Eastern Europe (Siegrist, 1988, 1996; Freidson, 1986; Conze and Kocka, 1985; Malatesta, 1995; Charle, 1996; McClelland et al., 1995). In the twentieth century, professionals express claims to membership in the middle classes through discourses on the 'loss of middle-classness' and the 'threat of proletarianization.' Under National Socialism and Fascism, the meaning of the 'free professions' was hollowed out, as professionals were redefined as militant 'thinking workers,' who took part in the 'struggle of the volk and the master race.' In the Communist world, the concept of the socialist intelligentsia implied the transcendence of class relations, the end of occupational privileges, and the subordination of professionals to the state and the party. Modern western concepts such as 'technocrat' and 'expert' define the role and position of professionals independently from social or economic class.

Conceptions and History of Professionalization and Professions

The history of the professions begins in the high and late Middle Ages and in the early modern era with the autonomous corporate colleges of the higher professions (lawyers, physicians, university professors) and the lower professions such as surgeons and notaries (Prest, 1986; Betri and Pastore, 1997). The various 'premodern' professions have specific orders of knowledge and a relatively stable social position and legal rank in the society of estates and among the urban elites and middle strata. They enjoy privileges with regard to education, access to the professions, the practice of the professions, and ethical monitoring. The specific type of nonmanual work and the way of earning a livelihood distinguishes the higher professions from handicrafts and commercial occupations. The institutional pattern and the mentalities of the old European corporate professional culture varied according to the degree of the abstraction of knowledge and social proximity to the nobility, the patriciate, and the urban bourgeoisie. As a rule, the right to practice a profession was limited to members of a specific sociolegal category and to a specific territory.

The traditional culture of the professions came under pressure already in the rise of the premodern territorial state. It began to dissolve in the eighteenth and early nineteenth centuries, when authoritarian and bureaucratic societies of Europe reformed professions and brought them under the control of the state. At the same time, in radical–liberal, democratic, and egalitarian societies such as France in the revolutionary era, the United States and Switzerland in the mid-nineteenth century, the professions were largely deregulated or were established as free trades (Ramsey, 1984; Siegrist, 1988, 1996). The general orientation of the different professions varied between (1) loyalty to and dependence on the state, (2) a progressive–liberal worldview, and (3) the desire to defend or revitalize traditional professional cultures and corporate mentalities. Different professions displayed many similarities with regard to the level of education, the orientation toward science and scholarship, and the social position of their members. Nevertheless the various occupational groups often competed with one another for cultural hegemony, social prestige, and political power in the modern states and nations.

Institutionalization of Professional Training

The specific way in which education and training are institutionalized plays a central role in processes of professionalization. The states of the European continent led the way in the creation of theoretically and scientifically based training programs and in the transfer of such programs to (reformed) universities and elite institutions of higher learning such as the French Grandes Écoles. On the other hand, it is also true that (complementary) training through practice survived and even retained a central importance (trainees in state agencies, mandatory internships with practitioners, or, for example, when the first professional years serve, in effect, as a sort of apprenticeship). In Anglo-Saxon countries, the transition from 'training on the job' or 'shop culture' to 'school culture' has intensified since the late nineteenth century (Lundgreen, 1990). A comparison of the length of time required for professional training in different countries reveals significant differences, even for the same occupation. Some of these differences may be attributed to variations in the contents of training programs, while others depend on factors such as tradition,

restrictive state policies affecting the labor market, or exclusive strategies of the professions, and social elites themselves.

The Institutionalization of Professions

Beginning in the eighteenth century, it is important to distinguish between the legal institutionalization of the professions by the state, on the one hand, and the formalization and regulation that is intended and implemented by the occupational groups themselves, on the other. With regard to the latter, one must also distinguish among (1) the continual development of a pre-modern profession, which, in principle, was never dissolved (such as the English Bar); (2) the restructuring of professions, which had been temporarily deinstitutionalized (as was the case with physicians and lawyers in the United States); and (3) the formation of new professions by members of functional and occupational groups that first arose in the nineteenth century (especially engineers but also a number of other 'new' professions). All of these involve long processes, which typically culminate in accreditation through the state or the legislation.

In much of Europe, the state or the government or legislation took the initiative to regulate and institutionalize the professions (Siegrist, 1996; Ramsey, 1984; McClelland et al., 1995). The state, as the guarantor of the legal order, the promoter of culture, and as an intervening force in society, needed professional services and provided the professions with knowledge, rights, and duties. In this way, the state distinguished the professions symbolically from the occupations and functional roles of business, commerce, and agriculture, which had been deregulated through the abolition of guilds and the liberalization of the trades. Henceforth, the 'professions' in the restricted sense included only those vocations that were occupied with central goods and values or with securing property, honor, health, public order, morality, and science. In this new configuration of the professions, the 'old' European conceptions of profession, guild, and vocation were combined with 'modern' concepts of progress, liberal order, freedom, science, rationality, national culture, and civilization.

In contrast to liberal and democratic societies, monarchical, bureaucratic, and statist societies tend to define professionals as civil servants and to establish a specific hierarchy of the professions: at the pinnacle of this hierarchy are professionals in the civil service, who are followed by state-appointed professionals, then by highly qualified members of the 'free professions' and finally by those professionals whose training and practice are not prescribed exactly. In liberal societies, the hierarchy is less pronounced, and the free professions stand at the pinnacle (Siegrist, 1996). Since the late nineteenth century, professionals have been increasingly employed in private enterprises, such as industrial firms, banks, insurance companies, and research laboratories. The occupational and status groups of professionals who are employees develop their self-understanding with reference to the social types of the civil servant, the free professional, and the entrepreneur.

The professional and social types of the nineteenth century may be characterized as follows:

1. Professionals who are 'civil servants' are to a great extent bound in the service of the king or the state. If necessary, the individual lawyer, physician, engineer, or professor must disregard professional criteria and follow the directives of his superior. Due, however, to the increasing recognition of professional competence and of the indefinite tenure of higher civil servants, specific ideas and forms of professional autonomy also come into play in the judiciary, in the administration, and in higher education. Professions, such as the engineers who were educated in the elite technical colleges in France and then entered into the civil service, developed a special esprit de corps (Schweitzer, 1999);
2. Members of 'state-appointed professions' are holders of public office. Professionals such as these enjoy a kind of protected market (numerus clausus), but they are also monitored closely. Political appointment strengthens the authority of the expert in society but limits his autonomy, impedes the formation of a self-defined collective consciousness of the profession, and makes the organization of professional associations more difficult. Examples include the Prussian, Bavarian, and Austrian lawyers well into the late nineteenth century, the French and Italian court lawyers (*avoués*, *procuratori*), the Prussian county physicians, and the community physicians of Italy. In the late nineteenth century, some of these state-appointed professions were able to emancipate themselves from state protection, which had come to be viewed as patronizing. Linking discourses on political freedom and the free market with discourses on professional autonomy, they expressed the desire to become 'truly free professions.' Since clients increasingly made similar demands, many states changed the corresponding laws and regulations in the last third of the nineteenth century;
3. The 'free professions' are characterized by more elaborate forms of training and examination and distinguished titles. Credentials are derived mostly from the state, but the free professional must be able to adapt to one or more segments of the market for professional services. Minimal standards for both professional and ethical behavior are often regulated by law, at least in principle. In practice, such regulations are enforced either by professional courts of honor (disciplinary bodies with the authority of public law or executive boards of associations with informal authority) or by ordinary courts and clients; and
4. In those areas where professional functions are exercised in the form of free trade, the constraints of the market and the increasing authority of science and scholarship encourage informal processes of professionalization. Professionals are themselves motivated to improve their own qualifications. A professional elite defines professional qualifications, founds an association, and attempts to marginalize or discredit so-called semiprofessional practitioners.

Cognitive and Emotional Orientations and Strategies

In the eighteenth and nineteenth centuries, the contents, forms, and meanings of professional knowledge changed. As was the case with other societal elites, professionals formulated new scientific discourses and new models for professional conduct. The physician achieved greater recognition as a scientifically and practically trained healer and health expert (Huerkamp, 1985; Léonard, 1981), while the lawyer came to be viewed as a legal

expert and 'priest of justice,' who was responsible for the realization of the rule of law and for securing the rights, the property, and the honor of the citizens (Siegrist, 1996). The engineer was the chief proponent of scientific and technical progress, and was long considered to be an expert for administration, supervision, and organization (Schweitzer, 1999). The university professor and the teacher in the gymnasium or *lycée* were the experts for education, esthetics, and morality. Finally, the Protestant minister was the educated theologian, whose legitimacy was no longer derived from faith and piety alone.

The degree of the systematization of knowledge increased over the course of the nineteenth century but the superiority of the professions and the efficiency of professional conduct remained controversial for an extended period of time. This is evident in the debates over the competence of university-trained 'physicians,' practical 'surgeons,' 'herbalists,' 'quacks,' nuns, midwives, and orderlies, and, more generally, over 'professional knowledge' compared with the knowledge of 'lay persons.' The professions had enduring difficulties in their efforts to prove their greater efficiency and to legitimize their autonomy with reference to science. Therefore, they attempted to support their claims to professional autonomy with supplementary appeals to values and attitudes that they shared with most of their middle class clients and employers. The leading concepts in these efforts were independence, responsibility, impartiality, altruism, public service, acting 'in good faith,' and 'delicacy' (in the sense of tact and discretion). These values and attitudes were often codified in laws, professional regulations, codes of ethics, and professional manuals. In many cases, such moral standards applied to the whole person of the professional, i.e., not only to his occupational activities but also to his behavior as a private person, as a member of the social elites or middle classes, and as a good citizen as well.

Compulsion and persuasion served to bridge the gap between professional culture and lay culture. Professionalization was embedded in the historical development of power relations, social hierarchies, and disciplinary processes, which emphasized the compulsive character of the social action of professionals. With the spread of medical, legal, technical, moralizing, and educational practices, professional culture as a combination of knowledge and power penetrated all aspects of modern life and thought (Goldstein, 1984, 1987). The professions are said to have used their knowledge in order to secure for the elites and themselves social and economic advantages. On the other hand, empirical, social, and cultural–historical studies, in which concrete interactions between experts and lay persons are investigated, often conclude that experts were hardly able to realize their cultural projects in a pure form. For quite some time, lay persons avoided contact with professionals by seeking advice with regard to health from pastors or priests and by seeking legal council from union secretaries.

The historical record shows that the professions that were close to the state or involved with legally regulated cases and procedures were, as a rule, more successful in compelling recognition of their competence. In contrast, clients who made a short visit to the doctor's office, to a lawyer, or to a priest were not exposed to experts to the same degree. On the whole, the free professions had to rely more heavily on persuasion. This explains why symbolic performances, clothing, decoration of work spaces, rituals, and rhetoric are such an essential part of professional culture.

Professions, Markets, and Economic Situation

The rise of the professions in the nineteenth and twentieth centuries was due largely to the increased and increasingly differentiated demand of private and public customers for goods and services. To a degree, professionals were able to stimulate this demand as well. They fulfilled a continually growing need for the explanation of the world, the formulation and attribution of meanings, for edification, and for entertainment. Through their influence, questions of law and property, and of health and death were viewed less fatalistically. Many new opportunities arose for them in the construction trade, commerce, industry, banking and insurance, research, higher education, and also in agriculture. The professions profited from a general increase in prosperity and in the cultural needs of the middle classes. In the twentieth century, there was a rising demand for medical, social, and cultural services on the part of lower social strata, which had previously been neglected. Due to developments in the fields of education and insurance in the welfare state, and because of a general rise in income levels, these strata developed into a new mass market for professional services. The changing profile of their clients led to a revision of the traditional attitudes and practices of the professions.

Professionalization is often conceived as a strategy for market closure, cartellization and monopolization. Indeed, professional projects aim at reducing competition and stabilizing or increasing the incomes of the members of an exclusive group. Yet, comparative historical research about the bundle of professional aims and strategies, show that the relative weight of economic interests and the forms in which they are expressed, vary quite a lot. The economic situation of the professions has at no time and in no place been homogeneous. The spectrum ranges from the modest conditions of the lower middle class (*petite bourgeoisie*) to those of the upper middle class (*haute bourgeoisie*). Income and wealth vary according to stages in career development. They depend upon individual abilities and performance, on regulations governing fees and honoraria, on conditions in various markets, and on general economic cycles. In the nineteenth and early twentieth centuries, it was not uncommon for professional incomes to be augmented by supplementary, nonprofessional activities, by dowries and inheritance, by earnings from real estate, or by income from stocks and bonds. This was typical of the bourgeois and petit bourgeois milieus to which the professionals belonged through birth, marriage, or social relations. In the twentieth century, the income of professionals was still significantly higher than that of workers and the national average. Due to increasing competition, increasing recruitment from families of humbler background, and to economic and political crises, professionals have often become convinced that their heyday is behind them and that they are threatened with decline (Jarausch, 1990). Under these circumstances, there is often an intensification of conflict within the professions between protectionists and those favoring a more aggressive market orientation.

Beginning in the late nineteenth century, as the model of the free professions continued to spread, all of the professions became increasingly dependent upon the market. In reaction to

this, there were many debates among professionals over the 'overcrowding' of the market and the closure of the various professions. In such constellations, forms of a defensive professionalism and protectionist strategies became more popular in the professional middle classes. There was a general trend toward lengthening training programs and making them more difficult, in order to stem the influx of young aspirants. The pattern of discrimination against female candidates was widespread. In principle, women had been admitted to higher education since *c.*1900, but they were effectively hindered in practicing the classical professions for quite some time (Jacobi, 1994). In Central and Eastern Europe, members of religious and ethnic 'minorities,' who had been emancipated in the nineteenth century, still experienced discrimination when they attempted to gain admission to the professions. Due to their supposed 'over-representation' in universities and academic occupations, Jews were often made responsible for the worsening of the working conditions and the standard of living among professionals (Jarausch, 1990; McClelland et al., 1995). It was not until the second half of the twentieth century that discrimination on the basis of religion, ethnicity, and gender gradually came to be forbidden.

Particularism vs Universalism

Depending on varying perspectives among both contemporaries and historians, the modern history of the professions may be understood in terms of the nationalization of culture and society or in terms of the universalization of knowledge and the rise of an European or even global professional culture. In the process of nation-building, members of the professions became part of the elites, which supported the development of national society, politics, science, and culture. The national system provided them with opportunities and assigned rights and duties to them. As functional and cultural elites of the nation, they had a common interest in maintaining the interconnection of power, status, knowledge, and earnings (Léonard, 1981; Siegrist, 1996; Pue and Sugarman, 2003; Pérez-Perdomo, 2006). In both the public arena and in more exclusive spheres, they represented themselves through professional and bourgeois values and symbols; these included historical accounts, myths, and *lieux de mémoire* as well as cultural artifacts such as busts, statues, and portraits of famous professionals in public buildings. Professionals have tended to view their own structures, practices, and status as being nationally specific, even in those cases, where, in objective terms, there is a dominant international pattern or a transnational epistemic community, which is subject only to slight regional variations and nuances.

Comparative historical research and the history of institutional and cultural transfer relativize national interpretations and lead to significant revisions of national views of history. Such approaches have shown how states and national professions imitate and vary foreign symbol systems and institutions, when they are convinced that this will lead to an improvement in their performance and in their status in national and international competition (Ramsey, 1984; Lundgreen, 1990; Siegrist, 1996; McClelland et al., 1995). In some cases, occupational patterns that were originally imposed by European empires were maintained after the end of imperial domination. French models for advocates and for engineers were imposed on other lands during the Napoleonic era; but subsequently they were adopted voluntarily, because they were found to be compelling. The Habsburg monarchy influenced the structure of the professions in Southeastern Europe. Prussian prototypes exerted a strong influence in the development of the civil service and the professions in Northern Central Europe. During the colonial period, the model of the English and French professions diffused worldwide, and they have retained their attraction in the postcolonial era as well (Pernau, 2008). The Soviet model of the socialist intelligentsia spread after 1945 to the lands of the Eastern Bloc. In Europe, American patterns have competed with national ones at least since the 1980s.

Since the late nineteenth century, the European or universal character of knowledge and professional culture has often been celebrated at international congresses, but national patterns survived into the twentieth century. In Europe, these have only begun to weaken in recent decades, in the wake of liberalization, the integration of the European market, and globalization (Evetts, 1998). This is a clear indication of the fundamental difference between the professions and other occupations, which were liberalized at an earlier date and more thoroughly.

Historiography and Emphases in Current Research

Traditional historiography tended to be focused on a single profession. It served primarily to affirm a tradition and to establish an identity, to further the understanding of ones own profession and even to celebrate it, to delimit the profession vis-à-vis other occupations and semi-professions, and to justify the status quo and lend support to demands made in the public arena (Burrage, 1990; Siegrist, 1996). The oldest genres in the historiography of the professions include works which defend traditions, privileges, and customs in the face of threatening political, social, and legal developments or alternatively, they are polemical works by outsiders and civil servants who want to reform the professions. Since the nineteenth century, there has been a proliferation of biographies and memoirs of exemplary professionals, of *festschriften* on the occasion of anniversaries, celebrations, and congresses, of memoranda and surveys with historical aspects, and of learned historical treatises in encyclopedias on the general development and various special aspects of the professions. More ambitious historical monographs on particular professions first appeared in the late nineteenth century. The authors were typically either academic or leading members of the profession in question with an interest in history. Until recently, academic historiography has only showed a moderate interest in the history of the professions. Prior to the 1960s, both amateur and academic historians were more interested in individual professions than in typical aspects of the various professions and rarely took the approaches and theses of the sociology of the professions into account.

Precursors of the systematic historical comparison (and sociology) of the professions include works which, beginning around 1800, compared local and regional occupational ordinances and occupational conditions with foreign equivalents.

This sort of comparison and historical reflection was often conducted in connection with the preparation of national laws pertaining to professions. Around 1900, representatives of the new discipline of sociology began to take abstract and systematic approaches to individual professions, to groups of professions, and to specific aspects of the processes of professionalization. In the efforts of sociologists to discover the 'essence' of the professions or to develop a 'general theory' of the professions, historical aspects and cultural peculiarities were neglected. In the 1960s, the sociology and history of class and social conflict and pioneering contributions to international comparison, reintroduced a historical perspective to the study of the professions (see the historical overviews on the sociology of professions by Freidson, 1986; Burrage, 1990). These works inspired the new research on the professions in social history, in which the development of the professions is historicized and recontextualized, without, however, sacrificing the goals of systematizing, conceiving of historical types, and contributing to the formulation of a general theory. Today, the history of the professions is an established and dynamic field of social history, cultural history, the history of the sciences, and the history of education. Historical approaches to gender have shown why professionalization has, until very recently, remained a male phenomenon and how professionalization and gender have entered into specific kinds of symbiotic relationships.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Ethical Codes, Professional: Business Codes; Social History.

Bibliography

Betri, M.L., Pastore, A. (Eds.), 1997. Avvocati, Medici, Ingegneri: Alle Origini Delle Professioni Moderne. Clueb, Bologna, Italy.

Burrage, M., Torstendahl, R. (Eds.), 1990. Professions in Theory and History. Sage, London.

Burrage, M., 1990. The professions in sociology and history. In: Burrage, M., Torstendahl, R. (Eds.), Professions in Theory and History. Sage, London.

Charle, C., 1996. Les intellectuels en Europe au 19e siècle. Éditions du Seuil, Paris.

Conze, W., Kocka, J. (Eds.), 1985. Einleitung. Bildungsbürgertum im 19. Jahrhundert. Teil 1: Bildungssystem und Professionalisierung in internationalen Vergleichen. Klett-Cotta, Stuttgart, Germany.

Evetts, J., 1998. Professionalism beyond the nation-state: international systems of professional regulation in Europe. Journal of Sociology and Social Policy 18, 47–64.

Freidson, E., 1986. Professional Powers: A Study of the Institutionalization of Formal Knowledge. University of Chicago Press, Chicago.

Goldstein, J., 1984. Foucault among the sociologists: the 'disciplines' and the history of the professions. History and Theory 23, 170–192.

Goldstein, J., 1987. Console and Classify. The French Psychiatric Profession in the Nineteenth Century. Cambridge University Press, Cambridge, UK.

Huerkamp, C., 1985. Der Aufstieg der Ärzte im 19. Jahrhundert. Vom gelehrten Stand zum professionellen Experten. Vandenhoeck & Ruprecht, Göttingen, Germany.

Jacobi, J. (Ed.), 1994. Frauen zwischen Familie und Schule. Professionalisierungsstrategien bürgerlicher Frauen im internationalen Vergleich. Böhlau, Köln, Germany.

Jarausch, K.H., 1990. The Unfree Professions: German Lawyers, Teachers and Egineers, 1990–1950. Oxford University Press, New York.

Jessen, R., Vogel, J. (Eds.), 2002. Wissenschaft und Nation in der europäischen Geschichte. Campus, Frankfurt am Main, Germany.

Léonard, J., 1981. La médecine entre les savoirs et les pouvoirs. Histoire intellectuelle et politique de la médecine française au 19e siècle. Aubier, Paris.

Lundgreen, P., 1990. Engineering education in Europe and the USA, 1750–1930: the rise to dominance of school culture in the engineering profession. Annals of Science 47, 33–75.

Malatesta, M., 1995. Society and the Professions in Italy, 1860–1914. Cambridge University Press, Cambridge, UK.

McClelland, C.E., Merl, S., Siegrist, H. (Eds.), 1995. Professionen im Modernen Osteuropa/Professions in Modern Eastern Europe. Duncker & Humblot, Berlin, Germany.

McClelland, C.E., 2003. Prophets, Paupers, or Professionals? A Social History of Everyday Visual Artists in Modern Germany 1850-Present. Lang, Bern.

Müller-Jentsch, W., 2011. Die Kunst in der Gesellschaft. VS Verlag für Sozialwissenschaften, Wiesbaden.

Pérez-Perdomo, R., 2006. Latin American Lawyers: A Historical Introduction. Stanford University Press, Stanford, California.

Perkin, H., 1989. The Rise of Professional Society in England since 1880. Routledge, London.

Pernau, M., 2008. Bürger mit Turban. Muslime in Delhi im 19. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen, Germany.

Prest, W.R., 1986. The Rise of the Barristers: A Social History of the English Bar 1590–1640. Clarendon Press, Oxford, UK.

Pue, W.W., Sugarman, D. (Eds.), 2003. Lawyers and Vampyres: Cultural Histories of Legal Professions. Hart, Oxford-Portland Oregon.

Ramsey, M., 1984. The politics of professional monopoly in nineteenth-century medicine: the French case and its rivals. In: Geison, G.L. (Ed.), Professions and the French State 1700–1900. Pennsylvania University Press, Philadelphia.

Schweitzer, S., 1999. Der Ingenieur. In: Frevert, U., Haupt, H.-G. (Eds.), Der Mensch des 19. Jahrhunderts. Campus, Frankfurt am Main, Germany.

Siegrist, H. (Ed.), 1988. Bürgerliche Berufe. Zur Sozialgeschichte der freien und akademischen Berufe im internationalen Vergleich. Vandenhoeck & Ruprecht, Göttingen, Germany.

Siegrist, H., 1996. Advokat, Bürger und Staat. Sozialgeschichte der Rechtsanwälte in Deutschland, Italien und der Schweiz (18.–20. Jh.). Klostermann, Frankfurt am Main, Germany.

Professions in Organizations

CR (Bob) Hinings, University of Alberta, Edmonton, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12160–12166, © 2001, Elsevier Ltd., with an updated Further Reading section supplied by the Editors.

Abstract

The study of professionals in organizations is about the intersection of two areas of enquiry that have quite distinct histories: the study of professions and the study of organizations. There have been times when they have been related strongly and other times when they have gone their distinct ways. In fact, the study of professions in organizations has gone through three, overlapping, phases since the mid-twentieth century. The first, started in the 1950s and particularly strong in the 1960s, was that of a critical examination of the role of professionals in bureaucratic organizations. The second, begun in the 1970s and continuing, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The third, which has its origins in the 1960s and 1970s, blossomed since the 1980s, emphasizing the dynamic nature of professionals in organizations and acknowledged the variety of organizational locations. This article outlines these phases, dealing with the central concepts and findings, examining the questions that each one raises, and pointing the way to the next phase of research in this rich field.

Professionals and Bureaucracy–1950s Onwards

The origins of the initial discussion of the relationship between professionals and organizations lies in the sociology of the professions and the discussions of defining characteristics of a profession. There are a number of contributions, but one that is often seen as central is that of Greenwood (1957) who suggested that professions have: a basis in systematic theory; prolonged training and certification; recognized authority; community sanction and legitimation; professional culture and codes of ethics. In essence, he came up with a set of traits or components that, when present in their entirety, defined the complete profession. This trait approach was echoed in a variety of ways by Wilensky (1964) and Hickson and Thomas (1969). There were two thrusts to these discussions. The first was to establish the characteristics that make up a profession. The second was to determine empirically which occupations were 'truly' professional and which were 'partly' professional and which 'aspired' to the status of a profession.

Out of these discussions came a relatively agreed upon set of characteristics of a profession that had structural and value aspects. The structural aspects were such things as a full-time occupation, a training and certification system, a professional association, and an ethical code. But, more importantly for the study of professionals and organizations, was the extension of Greenwood's notion of a professional culture into a set of values which encompassed peer reference, vocation, public service, self-regulation, and autonomy. When this work in the sociology of the professions connected with the sociology of organizations it was in terms of the relationship between the values held by professionals and those promulgated by the modern, bureaucratic organization. This relationship was seen primarily as antagonistic (cf Hall, 1968).

Through the 1950s and 1960s the great influence on the sociological study of organizations was Weber (1949). Weber's principal contribution to this field was his characterization of organizations in terms of the authority relations within them (systems of imperative coordination) and the ways in which such systems had evolved historically. Each ideal type of authority, charismatic, traditional, and rational–legal, had its associated organizational form.

Weber saw the rational–legal authority system with its organizational form of bureaucracy as the dominant institution of modern society. The authority system is rational because means are designed expressly to achieve certain goals and it is legal because authority is exercised through an office with its associated rules and procedures. For Weber the bureaucratic organization was technically the most efficient form of organization possible.

In terms of Weber's evolutionary history of social systems, bureaucracies represent the final stage in depersonalization. Offices (jobs) are held by experts who are arranged in a hierarchy. Rules and procedures provide predictability and consistency. Information is recorded and stored. Personal and business affairs are separated. There is 'the methodical attainment of a definitely given and practical end by means of an increasingly precise calculation of means.' Starting in the economic sphere, bureaucracy is such a potent method of organizing that it becomes characteristic of all areas of society such as education, government, politics, religion, and so on. Much of organizational analysis of the 1950s, 1960s and 1970s was essentially concerned with establishing, extending, and critiquing Weber's notion of bureaucracy (e.g., Pugh et al., 1969)

As Abbott (1988) in his influential work has pointed out, this 1950s and 1960s view of professions had an underlying assumption of professionalization as a natural process; that is, a regular sequence by which an occupation moved to 'full' professionalism (Wilensky, 1964). Essentially, this view, together with 'the search for bureaucracy' produced a unitary view of both professional occupations and organizations. Professionals, it was suggested, were socialized into occupations where the key values were autonomy, peer control, and vocation. For example, Hall (1968) operationalized

professional values as the professional organization as a reference, a belief in service to the public, a belief in self-regulation, a sense of calling to the field, and feelings of autonomy. Bureaucracy was operationalized as hierarchy of authority, division of labor, rules, procedures, impersonality, and technical competence. However, his conclusion was 'an assumption of inherent conflict between the professional or the professional group and the employing organization appears to be unwarranted.'

Of course, both sides of this coin, the professional and the organization, could be, and were opened up to scrutiny. On the organizational side of the coin, Scott (1965) had put forward the notion of Autonomous and Heteronomous Professional Organizations. Pugh et al. (1969) distinguished between Full Bureaucracies, Workflow Bureaucracies, and Personnel Bureaucracies as well as Implicitly Structured organizations. The point of these taxonomies and typologies was to open up the possibility of there being organizational forms in which the values espoused by individual professionals would be embraced and protected, thus potentially transforming the nature of the debate beyond an inherent conflict.

On the other side of the coin, the changing nature of professions was being stressed together with their organizational locations. Two issues are important. First, there was a trend of increasing professionalism within bureaucratic organizations. Occupations were arising that claimed professional status, such as social workers, nurses, and managers. These were bureaucratically produced and located occupations. Second, Hastings and Hinings (1970) suggested, from their work on chartered accountants, that there were established professions that carried out tasks that were centrally located in bureaucratic organizations and that, as a result, there was little or no conflict between professional and bureaucratic values in these situations.

This 'opening up' of both organization and profession led to new developments in understanding professionals in organizations. However, this theme of potential conflict between certain forms of organization and professionalism continues today. For example Raelin (1991) has examined the management of professionals in organizational contexts, writing about 'the clash of cultures' when generic managers are responsible for professionals. In the UK, and Europe more generally, there has been a concern with 'The New Public Management' and its organizational impact on professionals employed in healthcare, government at all levels, and education. The concern or thrust is still that there are incompatibilities between professional ways of working and values and certain kinds of organizational principles and practices. It continues to be a fruitful area of study.

Professionals in the Labor Market–1970s Onward

The second phase, begun in the 1970s and again, continuing until today, critically examined the orthodox view of professionals and gave a more power-based, labor market analysis of their position and how it affected organizational labor markets. The basis for this was laid down in critiques of the 'professionals in organizations' approach in the previous decades. The previous research questions had been anchored firmly in issues about the structure and values of professionalism with notions of an underlying process of professionalization arising from the nature of work in modern economies (Cooper and Robson, 2006).

The problems that were raised continued the opening up of professional occupations away from a homogeneous view of them. The uneven nature of professionals was stressed; that is, they could not and should not be considered as a unitary set of occupations. Similarly, the rise of the bureaucratic, organizationally based professions both in the context of the state and the private sector was recognized more fully. Abbott (1988) calls this the 'monopoly school' which attributes developments in professionalization 'to a desire for dominance or authority' (p. 15). The occupational/professional project is to control the conditions of work (the labor process and the labor market) and to increase and maintain occupational status and power (Johnson, 1972; Larson, 1977; Friedson, 1986).

The issues raised by this approach 'dislocate' the relationship between professionals and organizations both because it makes our understanding of professional occupations more complex, and also because it shifts the locus of debate away from values and structure to power and control. Indeed, the connection of this critique of the functions of professions is not as strongly related to the theme of professionals in organizations as the theories being critiqued. The emphasis becomes recentered on the nature of a profession and the professional project. In particular, the claim to professional status by an occupational group, for example, social workers, is seen as a way to control entry to the occupation and the demand for membership, and to establish an authoritative position within the labor market.

In a sense, the connection to organizations is primarily through an analysis of the rise of new professions and a continuation of the theme of 'the professionalization of everyone' (Wilensky, 1964). Because a key question to be answered is why new occupations make the claim to be professional, that is, the professional project, the examination is, in part, one of organizationally based occupations. Thus, the issue of professionals in organizations is not one of 'traditional,' autonomous professionals coming to terms with the rise of the bureaucratic organization, but the establishment of functionally and bureaucratically derived new occupations as professional. The issue for the new professional in this organizational setting is to acquire authority and control the conditions of their occupation's existence.

Thus, the professional/organizational relationship is not one of adaptation between two different sets of structure and values, but of an occupational group that is functionally defined and organizationally embedded acquiring the legitimacy and dominance that comes from professional status. The professional/organizational relationship is one of power plays, control of the labor process and fights over legitimacy that are played out both in the organizational setting and the institutional/regulatory setting within a particular society. This approach continues to be part of the discussion of professions in organizations and is to be particularly found in discussions of the healthcare sector (Friedson, 1986; Larson, 1977) and of the new public management. One of the issues that is raised by this stream of theorizing is that of the 'semiprofessions.' This was an attempt to distinguish between fully professionalized

occupations and other occupations that aspired to professional status. These latter occupations tend to be primarily those that are organizationally produced and located and where the individual occupation members do not work as independent practitioners. Johnson (1972) sees this distinction (and the use of other descriptors such as 'quasi,' 'marginal,' or 'limited' professions) as an illustration of power relations between occupational groups. For him, professions retain positions of authority vis-à-vis the semiprofessions, whose claim to professional status is contested. As he puts it, 'the pervasiveness of the ideology (of professionalism) is not indicative of the extent of professionalism (p. 59).

The theoretical and research issue of competition between groups for professional status continues to be an important topic. Several writers have observed the phenomenon of stratification within the professions. Abbott (1988) observes that interprofessional competition is a fundamental fact of professional life with jurisdictional disputes between occupational groups over claims to particular expertise and knowledge bases. And with more and more occupational groups that lay claim to such expertise and knowledge being strongly based in organizations, then the dynamic between professionals and organizations changes, on both sides.

Professionals in Professional Organizations–the 1980s Onwards

The third phase, which has its origins in the 1960s and 1970s, and blossomed in the 1980s and 1990s, emphasized the dynamic nature of professionals in organizations and acknowledged the variety of both professional and organizational locations. The issues of the variety in professional settings had been increasingly raised and, at the level of the professional system, Abbott (1988) has pointed out that there is continuous competition and jurisdictional disputes between professional groups as they dispute claims to expertise. Similarly, on the organizational side, the typologies and taxonomies of organizations both differentiated between types of bureaucracy and introduced new organizational forms.

This discussion has emerged in two different forms. The first is setting the discussion within the changes occurring in national and international economies. As thinking about 'the new economy' and 'globalization' has developed (Aharoni and Nachum, 2000), so two issues have been emphasized which bring together much of the previous theorizing about professions. On the one hand there is the rise of new occupations which claim professional status. For example, in the software development industry there are computer engineers, software engineers, and systems analysts who are highly educated but operate in paraprofessional environments across boundaries of time and space. Financial services have grown rapidly, as has the whole sector of business services, and along with this have come investment analysts, financial planners, and mergers and acquisitions specialists whose professional status is unclear although one sees the same processes as were written about in the 1960s and 1970s, namely those of educational qualifications, professional associations, examinations, and designations and the ideological claim to professional status.

On the other hand there is the rise of the 'knowledge-based organization' (Alvesson, 1995) together with the development of concepts such as the flexible organization and the learning organization. The suggestion is that organizations of the new economy are increasingly different from the various kinds of bureaucracy of the old economy. One of the reasons for this is that they are more service-based and that those services require highly educated employees who, in value terms, are much more like the 'old' professions, being committed to autonomy, vocation, and collegiality. These new organizational forms are seen as much more friendly and hospitable to professionals. Also, it is suggested that these forms are spreading into old economy organizations as they respond to the rigors of global competition.

The second form of the discussion is a renewed interest in the professional service firm, for two different reasons, each connected to new organizations and new occupations. Peters (1992) and Alvesson (1995) have suggested that the professional service firm is an example of the flexible, knowledge-based organization; that there is a wealth of experience in these organizations from which the newly developing knowledge-based firms can learn. But, at the same time, with the explosion of knowledge-based services, and, in particular, the tremendous expansion of business services in OECD economies, there is much more competition over professional jurisdictions. If one takes the example of the 'Big 5' accounting firms (formerly the Big 8 and the Big 10), there has been a continuous expansion of services within these firms and more and more integration across services and across nations (Greenwood et al., 1998).

The same processes are at work within other professional jurisdictions and professional service firms. In law, for example, there are strong internationalization pressures. Also, there is much discussion over the multidisciplinary practice, allowing nonlawyers to have a controlling interest in law firms. In 2000, the Law Society of New South Wales (calling itself 'the most progressive legal jurisdiction in the world') voted to allow multidisciplinary practices in law firms and also to allow them to incorporate.

Part of the process of defining these issues and also a result of them being articulated has been a rise in interest in the professional service firm (PSF). There were classic studies by Smigel (1964) and Montagna (1968) that examined the ways in which professionals operated in the professional service firm. But because of the concerns of theorists with issues of the rise of the bureaucratic organizational form and the response of professionals to that, and the concern with the monopoly critique, there was little or no interest in studying the professional service firm *per se*, even though it was the implicit, professionally hospitable setting contrasted with the bureaucratic form. It is this which has become the focus of much interest and research in the past two decades and in its own way it reflects issues of professionalization and bureaucratization, although the latter has been recast as managerialism and located in the context of knowledge-based organizations and globalization.

The theoretical and empirical location of studies of professional organization has been, primarily, in either the professional association and system or nonprofessional organizations. Very little study had been done of professionals in

professional organizations, organizations that can be described as PSF. One important dimension in understanding professions in organizations is the extent to which they are primarily located in professionally dominated organizations or in other kinds of organizations, the so-called bureaucratic professions. For example, practicing lawyers are employed primarily in law firms that are organized as professional partnerships; only a small proportion are employed outside legal practices; for example, as corporate counsel or as lawyers within the public defender justice system. Accountants tend to be split about 50/50 between private practice and employment in private and public sector organizations. However, almost all accountants are trained in private practice. Engineers are employed overwhelmingly in private and public sector organizations with only a small proportion of the profession being employed in engineering consulting firms. And engineers receive their postuniversity professional certification training in a wide variety of settings.

There are three points to be taken from this. First, professionals work in a variety of organizational settings, something that has been ignored in much previous work. Second, there are some professionals who primarily work in bureaucratic settings and experience little or no conflict. Hastings and Hinings (1970) showed that chartered accountants working in industrial settings experienced little role conflict and they argued that part of the activity of accountants is to bureaucratize others. Similarly, the debate over the semi-professionals such as social workers, teachers, and nurses centered around the fact that they were located entirely in bureaucratic organizations and did not have independent, autonomous practices. But, third, the arena of professional work that had been ignored in particular was that of the professional service firm in spite of the fact that implicit in many analyses of professionals, their location in independent professional organizational settings was considered to be very important as the archetypal setting for an autonomous professional practitioner.

Much of the work of the 1980s and 1990s has been concerned with understanding the nature of the professionally based, private practice organization (Nelson, 1988; Brock et al., 1999). As this work developed it also embraced the theme of bureaucratization although in a different guise, as managerialism.

This initial work was concerned with elaborating the nature of the traditional professional service firm, which Greenwood et al. (1990) typified as the P^2 form. The P^2 stands for the two key elements of professionalism and partnership. The PSF is an organization whose core workers are professionals. Traditionally they are from one dominant profession such as accounting or law. And the governance system of the organization is a partnership where the senior members of the professional group jointly and severally take responsibility for the direction of the firm.

There are a number of organizational consequences to these twin defining features. In human resource terms, one is the 'up or out' system or tournament. Young professionals either make it to partner or leave the firm. Organizationally, they are relatively loose structures, representative and collective decision-making with management being something that is done after dealing with clients. Issues of management and methods of organizing are not issues that are seen as important to discuss in the P^2 firm. The emphasis is on the individual professional and their ability to do good professional work. There are clearly strong echoes in this of the kinds of discussions that have arisen in the 1980s and 1990s about the knowledge-based firm with its employment of highly qualified individuals who are applying new, and sometimes esoteric knowledge, and, it is suggested, require freedom and flexibility to achieve results. In the P^2 organization, interesting work and the freedom and flexibility to do it are prized.

However, the nature of the professional firm has been changing due to changes in the institutional and market contexts of professional work. Because of the development of the new economy and the rise of services generally and business services in particular, there have been increasing pressures on the established professions such as law, accounting, and medicine. For more than a decade now, Western economies have been in a period of deregulation with the ideology of the market being dominant. Thus, jurisdictional boundaries have blurred and the level of *inter*-professional competition has increased. The rapid rise in the numbers of management consultants, financial services advisors, software development engineers, information technology consultants, media and marketing occupations, etc. has changed the dynamics of professional organizations. Similarly, there have been substantial changes in markets. On the one hand the demand for these professional business services has increased markedly, making it one of the most rapidly growing sectors in many economies. But the demands for these services are from increasingly sophisticated clients, operating on a global scale.

It has been suggested that a major consequence of these pressures and changes is that the professional, P^2 organizational form is in decline (Brock et al., 1999), and that it is being replaced by forms that place a greater emphasis on managerialism (Cooper et al., 1996). Cooper et al., (1996) named this new organizational form in professional firms the Managed Professional Business (MPB). In such an organization professionalism becomes centered on developing lines of business, hierarchies develop, strategic planning is introduced, targets are more binding, and decision-making rests in the hands of management teams rather than the partnership at large. So, the argument goes, the terms and conditions of work in the MPB move towards those of the regular corporation. However, this can only go so far because of the knowledge-based work of professionals (Rose and Hinings, 1999).

What this suggests is convergence for professionals in organizations, but of a new kind. It is not convergence around either professionalism or managerialism (bureaucratization), *per se*, but around a combination of these which is organized around the increasing role of knowledge in 'postindustrial,' service-based organizations. As yet, we have not sufficiently tracked these kinds of developments to write authoritatively about them.

Conclusion

So what can we conclude about research in this rich field and the status of our current understanding? There are a number of things. A very important one is the continuing theme

of the potential conflict between professionalism and bureaucratization and/or managerialism. That is, the notion of a professional encompasses values and cultural statements that are inimical to the forms and cultures of hierarchically organized, profit-driven corporations. However, this continuing theme has been modified in two important ways, both of which require much more study.

The first is the opening up of the idea of the professional. There is still a tendency to work with ideas of 'true,' 'full,' or 'traditional' professions that are seen as archetypal or at least worked with as ideal types. But, much of the original work on the professional–bureaucratic relationship, and, in particular, the ideas of the monopoly school, suggested that there are a range of professions exhibiting varying degrees of compatibility with bureaucratic working. More research is needed which starts from the *diversity* between professionals, rather than the assumption of homogeneity. There are still incompatibilities between some professional ways of working and their associated values and some organizational principles and practices. But whether those professions that have evolved from bureaucratic organizations have more compatibility with those forms is still very much an open question, and it needs to be connected to issues of claims to professional status and competition between occupational groups for professional status.

The issue that Abbott (1988) sees as central, namely, interprofessional competition and jurisdictional disputes between occupational groups, should be related to the types of organizations within and between which such competition and disputes take place. The healthcare setting is a particularly good one for examining these kinds of issues because of the numbers of professional groups involved.

A second modification to the professional–bureaucratic opposition thesis lies in the variety of organizational settings and their changing nature. While the death of bureaucracy is much over-stated, there is no doubt that new organizational forms are emerging that emphasize knowledge acquisition, use, and dissemination, together with flexibility and autonomy in organizational working. We might argue that the organic, flexible design is the 'new' organizational form. But there is a range of established and emerging forms within which there are differing responses to expertise and professionalism (Van de Ven et al., 2013).

A further development of these themes comes from our increased understanding of the role and nature of the professional service firm. Within the 1990s more and more interest and research has been centered on, *inter alia*, accounting firms, law firms, management consulting firms, and engineering consulting firms. These are, of course, the archetypal sites for the study of professionals within settings which should be the most accommodating because they are organized around professional work and managed by the professionals themselves. But they are also operating in changing institutional and market environments and face intra- and interprofessional competition. More research needs to be done on the way in which these professional fields are being restructured, new occupational grouping emerging, new organizational forms developing, and the organizational–professional relationship being changed. Part of this is also that there are important differences *within* as well as *between* professionals in terms of organizational setting.

Finally, an important and intriguing topic to be examined is the possible convergence of professionals in organizations and organizations of professionals. That is, on the one hand, many 'nonprofessional' organizations are becoming more knowledge-based and are experimenting with more flexible and autonomy-based organizational forms (e.g., Duska, 2005). On the other hand, professional organizations are moving into new markets, facing more competition, changing their knowledge base, and growing in size. In doing this they are becoming more managed, more strategic, and possibly decreasing some aspects of autonomy and flexibility. What they can offer is challenging work with major clients. This suggests a new kind of convergence, combining professionalism and managerialism (bureaucratization), due to the increasing role of knowledge in 'postindustrial,' service-based new economies. These changes present a real challenge for research.

See also: Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Organization: Overview; Organizational Culture, Anthropology of; Organizations, Sociology of; Professions and Professionalization, History of; Professions, Sociology of; Weber, Max (1864–1920).

Bibliography

Abbott, A., 1988. The System of Professions. University of Chicago Press, Chicago.
Aharoni, Y., Nachum, L., 2000. Globalization of Services: Some Implications for Theory and Practice. Routledge, London.
Alvesson, M., 1995. Management of Knowledge-intensive Companies. De Gruyter, New York.
Brock, D., Powell, M., Hinings, C.R. (Eds.), 1999. Restructuring the Professional Organization: Accounting, Healthcare and Law. Routledge, London.
Cooper, D.J., Robson, K., 2006. Accounting, professions and regulation: Locating the sites of professionalization. Accounting, Organizations & Society 31 (4/5), 415–444.
Cooper, D.J., Hinings, C.R., Greenwood, R., Brown, J.L., 1996. Sedimentation and transformation in organizational change: The case of Canadian law firms. Organization Studies 17, 623–647.
Duska, R.F., 2005. The autonomy of the profession: Whose agent are you? Journal of Financial Service Professionals 59 (3), 25–28.
Friedson, E., 1986. Professional Powers. University of Chicago Press, Chicago.
Greenwood, R., Hinings, C.R., Brown, J.L., 1990. The P2-Form of strategic management: Corporate practices in the professional partnership. Academy of Management Journal 33, 725–755.
Greenwood, R., Rose, T., Cooper, D.J., Hinings, C.R., Brown, J.L., 1998. The global management of professional services: The example of accounting. In: Clegg, S., Ibarra, E., Bueno, L. (Eds.), Theories of Management Process: Making Sense Through Difference. Sage, Thousand Oaks, CA.
Greenwood, E., 1957. Attributes of a profession. Social Work 2, 44–55.
Hall, R.H., 1968. Professionalization and bureaucratization. American Sociological Review 33, 92–104.
Hastings, A., Hinings, C.R., 1970. Role relations and value adaptation: A study of the professional accountant in industry. Sociology 4, 353–366.
Hickson, D.J., Thomas, M., 1969. Professionalization in Britain. Sociology 3, 37–54.
Johnson, T., 1967. Professions and Power. Macmillan, London.
Larson, M.S., 1977. The Rise of Professionalism. University of California Press, Berkeley, CA.
Montagna, P., 1968. Professionalization and bureaucratization in large professional organizations. American Journal of Sociology 74, 138–145.
Nelson, R., 1988. Partners with Power: The Social Transformation of the Large Law Firm. University of California Press, Berkeley, CA.
Peters, T., 1992. Liberation Management. Macmillan, London.
Pugh, D.S., Hickson, D.J., Hinings, C.R., 1969. An empirical taxonomy of structures of work organizations. Administrative Science Quarterly 14, 115–126.

Raelin, J., 1991. The Clash of Cultures: Managers Managing Professionals. Harvard Business School Press, Boston.

Rose, T., Hinings, C.R., 1999. Global clients' demands driving change in global business advisory firms. In: Brock, D., Powell, M., Hinings, C.R. (Eds.), Restructuring the Professional Organization: Accounting, Healthcare and Law. Routledge, London.

Scott, W.R., 1965. Reactions to supervision in a heteronomous professional organization. Administrative Science Quarterly 10, 65–81.

Smigel, E.O., 1964. The Wall Street Lawyer. Free Press, New York.

Van de Ven, A.H., Ganco, M., Hinings, C.R., 2013. Returning to the frontier of contingency theory of organizational and institutional designs. Academy of Management Annals 7 (1), 393–440.

Weber, M., 1949. Theory of Economic and Social Organization (A. Henderson, Trans.). Free Press, Glencoe, IL.

Wilensky, H., 1964. The professionalization of everyone? American Journal of Sociology 70, 137–158.

Professions, Sociology of

Andrew Abbott, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12166–12169, © 2001, Elsevier Ltd.

Abstract

The sociology of professions studies the organization of expert work. Until quite recently, such work has been characterized, above all in Anglophone countries, by a particular social form – the profession – uniting individual practice, collective association, specialized education and examination, and corporate discipline. Recent historical changes have brought the organization of experts into a more uniform pattern worldwide, as this older 'professional' form of organization has broken down under the pressure of salaried employment, rising state dominance, and an increased 'capitalization' of expert work. (At the same time, however, the English meaning of the word 'profession' seems to be becoming standard.) Studies of professions have reflected these developments. Early twentieth century work was typological, a style that gave way to functionalism in the 1950s. The 1970s brought a critical, 'professional power' school, which in turn gave way in the 1990s to a concern with professional work and with the impact of state and capital on the professions. An important strand of the literature has considered the professional intelligentsia as a 'new class.'

The sociology of professions is a branch of the sociology of work concerned with the analysis of expert occupations. It analyzes their patterns of organization, types of work, and social status. Because the sociology of professions first arose in the UK and the USA, it took its name and much of its early conceptual apparatus from the pattern of expert work characteristic of those countries. In recent years, however, the field has become more international, a change that is transforming both its empirical base and its theoretical superstructure.

Expert Occupations and the Sociology of Professions

The sociology of professions first arose to study the particular social form taken on by expert occupations in England and the USA in the course of the nineteenth century. This form was characterized by a combination of independent individual practice with some type of collective association. It usually involved specialized education, examination or licensure, service fees, and some form of autonomous discipline. By the early twentieth century, such a form was characteristic of medicine, law, accounting, architecture, and various other fields in the Anglophone world.

The first systematic analysis of these 'professions' was by Carr-Saunders and Wilson in 1933, who set forth a typological analysis of occupations resting on various 'traits' that were taken to characterize the professions. The Carr-Saunders and Wilson analysis had a hard time with the military and the clergy, which were certainly 'professions' in the English sense of the term. Moreover, it made no effective contact with the quite different organization of expert work on the continent.

The French, for example, have no word with the denotation of the Anglo-American 'profession.' The French do distinguish *professions libérales*, but these include only experts in free individual practice, not all experts of a certain type. *Profession* itself has much broader purview than the English 'profession,' as does the parallel *métier*. Moreover, the organization of French experts has historically been far more state-dependent than has that of Anglo-American ones. Nor was France the only problem. The word and concept of profession were no closer to the German culture of expertism than to the French. Indeed, the characteristic pattern for organizing expert workers on the continent has more often been through quasi-civil service arrangements than through autonomous 'professional' ones. The sociology of professions thus began as a study of a particular kind of expert occupation. Only gradually have historical and comparative study made the limiting character of that approach clear.

Historical Development of the Field

From the Carr-Saunders book of the 1930s up into the 1970s, typology remained the dominant strategy of the sociology of professions. Millerson (1964) presented a comprehensive summary of all the various qualities that had been used to define 'professionhood.' The typological approach eventually strangled inquiry, as investigators worried whether or not various groups really were professions. Indeed, a substantial literature grew up in this period on the 'semiprofessions,' usually felt to include teaching, librarianship, nursing, and social work. These fit the typological model in some ways (advanced education, licensing) but missed it in others (independent practice). The so-called semiprofessions were all largely female occupations, a fact that would later become more emphasized, but engineering, although largely male, shared many of the same characteristics. Within typological thinking, salaried employment of professionals presented a problem. A small literature in the 1950s and 1960s considered this problem under the heading of 'professionals in bureaucracies,' viewing such professionals as being under strong role strain because of the conflicting demands of profession and organization.

The typological model enabled the sociology of professions to come to a rapprochement with structural functionalism. In a celebrated essay, Parsons (1954) argued that professions were

rationalist (as opposed to traditional), that they exercised functionally specific authority (as opposed to undifferentiated authority), and that they were universalist (as opposed to particularist). The later Parsonian tradition saw the professions as squarely anchoring the pattern maintenance sectors of crucial social subsystems. The typologists' focus on the organization rather than the content of professional work accorded better with the early Parsons, however, which kept them apart from the later structural functional interest in the professions.

Throughout the early postwar period, students of Everett Hughes and others in the Chicago tradition continued to amass case studies of professions and semiprofessions. Many of these fit the typological model badly. Their dependent nature often mocked the high role structural functionalists assigned to professions. One way around this problem was to argue that professionalization was a characteristic process that many occupations had not yet fully traversed. In the 1960s and 1970s, sociologists of professions argued for a characteristic professionalizing process, a sequence of stages through which all expert occupations developed (Wilensky in 1964 provided the most famous analysis of such an argument).

In the later 1960s, the sociology of professions, like many other subfields, turned toward skeptical interpretations. Major reinterpretations came from Johnson (1967), Freidson (1970) and Larson (1977). This new 'professional power' interpretation saw the professions in a different light. Freidson brought a generation of detailed studies of medical workplaces and medical knowledge together with a theory of professional organization to argue that the structures of professional organization were hardly the direct outcomes of rational and functionally specific universalism. The power of doctors over nurses suddenly seemed less a matter of advanced training than an enforced local division of labor. At a much more general level, Johnson recast professionalism as one among a number of modes of occupational control. This argument was carried further by Larson's reinterpretation of professions as collective mobility projects in which expert groups sought rewards through control of certain markets for services.

The rise of the power school coincided with a wide new interest in professions among historians, particularly in the USA. Scattered works had previously considered some major professions – English law and US medicine, for example. But the 1970s and early 1980s brought a flood of works covering various professions in the USA, England, and to a lesser extent France and Germany. These various works covered periods from the early twentieth century as far back as the seventeenth and eighteenth centuries. Although there were (and remain) important lacunae, theorists of professions began to work with a new empirical base. Perhaps more important, the historians turned to the sociological literature for conceptual guidance and in many cases found little of use to them.

A number of other trends helped recast study of professions in the 1970s and 1980s. Most important was the steady drift of professionals away from independent practice and into corporate settings. On the one hand, this was simply a continuation of the trend toward larger and larger professional firms. In law, accounting, and architecture, sizable firms dated from the first half of the twentieth century, at least in the USA. Legal restrictions generally prevented this development both in the UK and on the continent. On the other hand was the development of multiprofessional firms, a pattern that had begun in architecture and now spread into areas like accounting, finance, and consulting. These largely professional firms were complemented by a much larger body of salaried professionals scattered throughout commercial and nonprofit organizations in the USA. By 1980, few US professions indeed were more than 50% made up of independent practitioners.

The rise of salaried professional work in the USA accomplished three things. First, it brought the US professional world much closer to the European pattern. The pattern of large-scale salaried employment by professionals had fascinated the French particularly, who had developed, in the concept of cadres, the notion of a technocratic intelligentsia as a new mode of occupational structure (see Boltanski, 1982). Second, the new salaried professional work reawakened the interest in the 'professionals in bureaucracies' problem of role conflict. Third, the new pattern led to new concepts of 'deprofessionalization' and 'proletarianization of professionals.' Although the latter term obviously derived from the attempt to develop a Marxist analysis of professions, both concepts attempted to come to terms with the notion – an absurdity under the 'professionalization' concept – that occupations could lose the perquisites and possibly even the form of professions once these had been gained. A wide variety of studies appeared in the 1970s and 1980s examining the degree to which professions and professionals were losing control of work and cultural authority as well as enduring more mundane changes in salaries and tenure.

By the mid-1980s, then, the sociology of professions looked quite different. It had lost nearly all touch with structural functionalism. There was confusion over which professions would develop in which directions at which times. A large new body of research opened to theorizing not only a wide variety of quasi-professional occupations in the USA and England, but also the expert occupations of the continent. In many cases, this new data had extraordinary historical depth (see, e.g., the comparative work in Cocks and Jarausch, 1990). In 1988, Abbott attempted to synthesize the entire area by refocusing the sociology of professions on actual professional work (both cultural and practical), turning away from the strong emphasis on the social structures of professionalism that had become the center of the literature. He envisioned a world of interprofessional competition over jurisdiction of work. Changes flowed into this world both from within professions and from beyond them, recasting some professions, privileging others, destroying yet others. Professions attempted to claim control of work in the workplace, before the public, and within the state, and the course of professions' histories was determined by the ensemble of this melee of interaction. Abbott explicitly attacked the still-continuing typological tradition for eliding large areas of work. For him a profession was anything that competed like one.

Although Larson and others had attempted some international comparisons, Abbott's heavy reliance on European cases, particularly French ones, presaged what became the major development in studies of professions in the 1990s. This was the trend toward examining the relation of professions and the state. Much of the most interesting work came from Europe, for example, Jarausch's (1990) comparative analysis of lawyers, teachers, and engineers in Germany in the first half

of the twentieth century. Krause (1994) has made the most general attempt at analyzing relations between the state and professions (and also capital), studying the USA, the UK, France, Germany, and Italy by means of case studies of law, medicine, engineering, and academics in each country. Krause's analysis focuses on the relative decline of professional autonomy and power in all these societies in the years after 1970. Both state and capital have increased their authority over professions.

The relation of state and professions has been underscored as an empirical question by the vast complications of creating standardized professional credentials in the European Union. The future of the sociology of professions will be greatly shaped by the theoretical developments necessary to analyze this process of internationalization of professional life. There is also a 'relations with capital' side to internationalization, encompassed in the largely unstudied complexities of the transnational corporations' use of professional services. There has been some study of 'offshoring' of professional services (e.g., of engineering services in telecommunications), but this particular capitalist attack on professional organization has not been studied in detail. In general, there is a startlingly small amount of work on expert occupations in the third world, with the possible exceptions of lawyers and, of course, the military.

Current Problems

The central present issue for the sociology of professions, then, is the question of how the professions will be reshaped by the economic reorganization that arose out of the crisis of the 1970s. Many of the recent transformations have long antecedents, particularly the drift to salaried employment with its questioning of professions' cultural authority. Other trends, however, are new or newly important because of the economic conjuncture. Internationalization has been crucial. By the late 1980s the restrictions on professional firm size began to fall in Europe, largely under the onslaught of US firms, particularly law firms and multiprofessional consulting firms, which were the only firms able to manage some of the huge legal and financial projects characteristic of the new international economy. The new gigantism has transformed relations both within the professions, and, more important, between the professions and the state. It is in the international context, too, that the question of the profession–state relation is most clearly posed.

But there have been important internal effects to rethink as well. The decline of medical power from the pinnacle portrayed by Freidson has given new meaning to the phrase 'proletarianization of professionals.' This decline, driven by reorganizations of insurers and regulators, requires revising general theory about how expertise is organized in society. In that sense, the sociology of professions is again confronted with a functionalist question: how is expertise to be produced, controlled and reproduced in modern societies?

The issue of state and capitalist control is only one side of this question. On the other is the issue of commodification of professional knowledge. It is at present much too early to know the effects of computer-aided design systems, medical diagnosis systems, and other expert systems on the future of professions. But it is clear that new theory is required. It is also clear that commodification will be used to further reduce the status of professionals (see Abbott, 1991).

Indeed, the field faces again the boundary question that has so long obsessed it. Earlier in this century, the boundary issue was 'what is a profession?' Given the new structure of employment relations, even in such highly professionalized surroundings as technological research firms and law firms, it is not clear what kinds of workers ought to be the subject of the sociology of professions. If the pattern of associational professionalism that gave rise to the field itself is either dead or transformed, then the field must retheorize its objects of scrutiny. Abbott's test – a profession is an occupation that competes by retheorizing others' work – presupposes fixed and organized occupations of a kind that simply may not exist under modern conditions of employment.

While the main focus of the professions literature has been on organizational and structural levels, some have insisted on a focus at the individual level, on the professionals themselves. The most general theme in this area has been a recurrent attempt to resolve the problem professionals (or expert workers) represent to Marxist theories of work. Originating in Djilas's book, *The New Class* (1957), the theme was continued in Bell's *Coming of Post-Industrial Society* (1973), with its predictions of a knowledge world of the future. Attempting to synthesize this literature Brint (1994) recognizes the intense pressures brought on older structures of professionalism by modern economies. He sees a decline of 'social trustee professionalism' as professionals become mere private citizens with particular interests. Perkin's (1989) much broader analysis argues that the idea of profession became a model for modern societies generally in the middle 50 years of this century. The ideal of an individual with personal 'knowledge capital,' combined with the notion of just reward, was in Perkin's view fundamental to the conception of the modern welfare state. In that state's fall, argued Perkins, fell the professional ideal as well.

Perkins' argument provides a useful conclusion. In his article on 'professions' in the prior edition of this encyclopedia, Talcott Parsons (1968) closed with the confident assertion that "the professional complex has already not only come into prominence but has even begun to dominate the contemporary scene in such a way as to render obsolescent the primacy of old issues of political authoritarianism and capitalistic exploitation." Now that history has shown him wrong, the sociology of professions must develop a theory that can stand the test of major historical change.

Bibliography

Abbott, A., 1988. The System of Professions. University of Chicago Press, Chicago.
Abbott, A., 1991. The future of professions. Research in the Sociology of Organizations 8, 17–42.
Bell, D., 1973. The Coming of Post-Industrial Society. Basic, New York.
Boltanski, L., 1982. Les cadres. Editions de Minuit, Paris.
Brint, S., 1994. In an Age of Experts. Princeton University Press, Princeton, NJ.
Carr-Saunders, A.P., Wilson, P.A., 1933. The Professions. Oxford University Press, Oxford, UK.
Cocks, G., Jarausch, K.H., 1990. German Professions. Oxford University Press, New York.

Djilas, M., 1957. The New Class. Praeger, New York.
Faulconbridge, J.R., Muzio, D., 2012. "Professions in a globalizing world" towards a transnational sociology of the professions. Internaitonal Sociology 27 (1), 136–152.
Freidson, E., 1970. Profession of Medicine. Dodd Mead, New York.
Gorman, E.H., Sandefur, R.L., 2011. 'Golden age,' quiescence, and revival: how the sociology of professions became the study of knowledge-based work. Work and Occupations 38 (3), 275–302.
Jarausch, K.H., 1990. The Unfree Professions. Oxford University Press, New York.
Johnson, T.J., 1967. Professions and Power. Macmillan, London.
Krause, E.A., 1994. Death of the Guilds. Yale, New Haven, CT.
Larson, M.S., 1977. The Rise of Professionalism. University of California Press, Berkeley, CA.
Millerson, G., 1964. The Qualifying Associations. Routledge, London.
Muzio, D., Brock, D.M., Suddaby, R., 2013. Professionals and institutional change: towards an institutionalist sociology of the professions. Journal of Management Studies 50 (5), 699–721.
Parsons, T., 1954. The professions and social structure. In: Essays in Sociological Theory. Free Press, New York, pp. 34–49.
Parsons, T., 1968. Professions. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 12. Macmillan, New York, pp. 536–547.
Perkin, H., 1989. The Rise of Professional Society. Routledge, London.
Riska, E., 2014. Gender and the Professions. The Wiley-Blackwell Encyclopedia of Health, Illness, Behavior and Society. John Wiley and Sons, New York.
Svensson, L.G., Evetts, Julia (Eds.), 2010. Sociology of the Professionas: Continental and Anglo-saxon Traditions. Daidalos, Sweden.
Wilensky, H., 1964. The professionalization of everyone? American Journal of Sociology 70, 137–158.

Profits

Tony Aspromourgos, School of Economics, The University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the main currents in the economic theory of profits from the eighteenth to the twentieth century – with particular reference to classical economics, marginalist or 'neoclassical' economics, the significance of profits for entrepreneurship, and Keynesian economics. It also examines how theories have connected the determination of rates of profit on capital, shares of profits in revenue, and absolute levels of profits.

Introduction

From the standpoint of the history of economic theory, the development of attempts at systematic explanation of the determination of profits has its beginnings in the mid-eighteenth century, most notably, with Richard Cantillon (1680?–1734) and François Quesnay (1694–1774) – the latter, the key figure of the French Physiocratic school. The object of their analyses is a more or less decentralized market economy, but not yet a social economy that could be regarded as genuinely capitalist. These are early contributors to what later comes to be called 'classical economics,' from William Petty (1623–87) to Karl Marx (1818–83), discussed in the next section. From the late-nineteenth century, there is the rise of the 'marginalist' school, commonly called neoclassical economics, which becomes dominant in academic economics after World War II, discussed in the following section. The significance of profits for theories of entrepreneurship is then examined. The penultimate section considers how theories have connected the determination of rates of profit on capital, shares of profits in gross revenue or value added, and absolute levels of profits.

But what of earlier Western thought? Is there anything that could be called 'theory' of profits to be found there? The systems of premodern Western thought concerning society, polity, and economy are first and foremost normative systems rather than descriptive systems. The most salient of these is medieval Catholic Scholasticism, a synthesis of the Judeo-Christian religious tradition and the ancient Greek philosophical tradition (as well as Roman law), most famously personified in the thought of Thomas Aquinas (*c*.1225–74). Perhaps the most striking theme of Scholastic thought that contingently touches upon profits is its strictures against money-lending for a return in the form of interest payments ('usury'). But there is no essential connection between these doctrines about interest and the subject of profits, since the former could apply to a world of only consumption loans (as distinct from loans for business purposes).

In any case, our purpose here is to elucidate descriptive doctrines; explanations of the existence of the phenomenon of profits, rather than ethical doctrines concerning the moral legitimacy or otherwise of profits. Nevertheless, in the course of the development of primarily ethical doctrines, elements of what may reasonably be called descriptive economic theory do arise in premodern thought. For example, Scholastic concepts of 'the just price' of a commodity gave rise to explanations of actual market prices, albeit crude; but little about the determination of profits that might be associated with those market prices. The concept of profits that arises in the eighteenth century belongs to a world of capitalism in which the systematic protection of individuals' property rights and enforcement of private contracts, together with free competition (in particular, the free mobility of capital and labor between different economic activities), establishes profits *as a systematic phenomenon with structure and a certain logic*, within a form of generalized liberal market economy. These institutions and forms of market organization only exist in fragmentary form in premodern times.

Classical Economics

The theorizing of profits in classical economics is focused upon rates of profits on capital 'advanced' or invested in production; that is to say, the ratio between the absolute level of profits and the value of capital invested, per time period. From the standpoint of profit maximization, this ratio is the magnitude that the owners or investors of capital are ultimately interested in: the proportion between their investment and their return. Anne Robert Jacques Turgot (1727–81) and Adam Smith (1723–90) are the earliest writers to clearly conceptualize profit rates so understood; but we concentrate on Smith, David Ricardo (1772–1823), and Marx in this section. The previous writers mentioned above, Cantillon and Quesnay, gave profits a systematic role in their economic theories, but not conceptualized in terms of the ratio of profits to capital invested. In Cantillon, profits appear as primarily a return for risk-bearing by 'entrepreneurs.' In Quesnay, they are the source from which capital accumulation by entrepreneurial farmers is financed, driving economic growth; but their persistence over time is difficult to rationalize in his theoretical framework.

Turgot and Smith, as well as the subsequent contributors to the classical tradition, decompose profit rates into risk premia that differ across different capital investments and a kind of 'pure' rate of return on capital investment as such (our term). See, for example, Smith (1976[1776]): book I, chapter 10. Hence, on the presumption of risk averse behavior by capital owners or capitalists (at least on average), there will be a kind of equilibrium structure of profit rates, even under competitive conditions, reflecting higher returns for riskier investments, although Smith has some doubts as to the predominance of

risk averse behavior. (Restrictions on competition – notably, restrictions on the free mobility of capital between different activities – will also cause profit rate differentials, with higher than competitive profit rates earned in activities less exposed to competition.) Smith is also careful to establish that, conceptually, remunerations for the functions of supervision and management of production, even if commonly referred to as profits, are, in functional terms, really a species of wages (Smith, 1976[1776]: pp. 128–130). This can be obscured when one and the same person is both the provider of capital to an enterprise and the manager of the enterprise. Smith and Turgot are also clear that if an enterprise employs borrowed funds to finance capital expenditures, some proportion of the profits will be distributed to the lenders as interest payments (e.g., Smith, 1976[1776]: pp. 112–113; Turgot, 1977[1769–70]: pp. 76–77, 85–88).

At the most abstract level of theorizing, it is common for the classical economists, as well as later economic theory, to focus upon a uniform or 'general' rate of profit in the economy, thereby leaving aside (to be treated separately) differential risk and the associated profit rate differentials under competitive conditions. As to what determines this underlying, uniform or general rate of profit on capital as such, Smith treats profits as an element of the realization, at the level of revenues or incomes, of the economy-wide production of a social surplus. This classical concept of surplus refers to that part of the gross product of the economic system which is available for free disposal, after replacement of the necessary inputs used up in the production of that gross output, including among those inputs the customary subsistence consumption of the labor employed. The question then is how the surplus is distributed or realized in the form of incomes. For Smith, the surplus is realized in the income forms of land rents, pure profits, and a part of wages, depending on the balance of bargaining power around the labor contract. But his theory of the actual magnitude of the general rate of profit rests on a highly unsatisfactory notion of 'competition of capitals.' Smith supposes that the greater the extent of capital accumulation, the greater the degree of competition between enterprises, and hence downward pressure on the general rate of profit (Smith, 1976[1776]: book I, chapter 9).

This does not amount to a convincing, determinate theory of the profit rate; and indeed, a generation or so after Smith, Ricardo demonstrates – albeit under rather restrictive conditions – that given the production methods in use in an economy, the general rate of profit is determined by, and inversely related to, the level of real wages (Ricardo, 1951 [1821]: chapter 6). The restrictive conditions Ricardo utilizes to arrive at this fundamental result involve, in particular, a labor theory of exchange value: It is supposed that the relative prices of commodities are explained by the quantities of labor time directly and indirectly required in the production of those commodities. Subsequently, it has been rigorously demonstrated that the inverse relationship holds under quite general conditions (e.g., Kurz and Salvadori, 1995). From the analytical point of departure of this inverse relationship, Ricardo's focus is the course of the profit rate over time, resulting from change in production methods over time. On the one hand, declining output from capital investment on poorer quality land as population expands will place downward pressure on the profit rate, given the real wages of labor. On the other hand, technical innovation associated with human invention can enable a rising profit rate, given the level of real wages. If the former factor dominates the latter, the general rate can fall to such a low level as to extinguish the motive to capital accumulation, bringing about a stationary state (zero economic growth; Ricardo, 1951[1821]: chapter 21).

In Marx's economics, considerably derivative from Ricardo's, profits are explained by 'exploitation' of labor, in the descriptive sense that profits are attributed to labor time worked by the average worker being greater than the labor time directly and indirectly required for the production of the worker's own consumption. The level of the rate of profit is then determined by reference to the proportion between the surplus labor time worked by the workforce (over and above the labor time required to produce its own consumption) and the labor time embodied in the capital invested in production (Marx, 1967: vol. 1 (1867)). (Thus expressing the profit rate, in terms of surplus labor time, does not seem to add any insight additional to the simple observation that in a capitalist economy the workers as a whole, as workers, do not receive the entire net product of the economy.) In the so-called 'transformation problem,' Marx seeks to translate economic variables (such as the profit rate), which he first expresses in terms of labor time, into the monetary values associated with competitive prices ('prices of production'). Marx also famously supposed a tendency for the general rate of profit to decline in the course of capitalist economic development, connected with the likelihood of crises under capitalism. But even within Marx's own theoretical framework, this is a contingent result. To translate it into more familiar modern language, in effect, Marx is expressing the general rate of profit as a ratio between profits per employed worker and capital per employed worker. The basis for a tendency toward a declining rate of profit is then the supposition of a rising capital-to-labor ratio, relative to any rising profits per worker. The possibility that any such trend toward a rising capital-labor ratio could be offset, or more than offset, by rising profits per worker is evident (Marx, 1967: vol. 1 (1867): pp. 304–305. vol. 3 (1894): pp. 232–240, 250–259).

Marginalist Economics

The key founding figures of this new approach to economic theory are William Stanley Jevons (1835–82), Alfred Marshall (1842–1924), Carl Menger (1840–1921), and Léon Walras (1834–1910). Rather than attempting to outline the particulars of the various efforts toward a theory of profits in these writers, and in those who followed their lead in subsequent decades of the twentieth century, we here seek to distil the essence of the theory of profits that arises out of this framework.

The fundamental point of departure is a conception of human psychology in terms of behavior governed by a pleasure/pain calculus. From this vantage point, economic behavior is explained in terms of the individual's maximizing net benefit ('utility' net of 'disutility'). Hence, for example, the desire to work and consume can be theorized along these lines. Suppose a person whose only marketable resource is his labor and who wishes to consume. He can supply labor, with a view to acquiring revenue, in order to enable consumption. On the

supposition that work generates disutility and consumption generates utility, the individual will have a desired labor supply and associated desired consumption demand that maximizes net utility – at least on the further supposition that there is increasing 'marginal disutility' from work and decreasing 'marginal utility' from consumption. Assuming there is sufficient labor demand and sufficient supply of consumption goods to satisfy the individual's desired labor supply and consumption demand, more time worked increases income, enabling higher utility from consumption; but notionally, as time worked and consumption increase, the increment of 'pain' from additional labor is rising and the increment of 'pleasure' from additional consumption is falling, so that at some point there is an 'optimal' combination of labor supply and consumption demand for the individual, where the rising marginal disutility from labor is just offset by the diminishing marginal utility from the consumption enabled by labor income.

The underlying characterization of human psychology here is certainly open to question (although 'utility' need not be supposed as cardinally measurable, but rather, merely as 'ordinal,' expressing individuals' preference rankings). But even if, for the sake of argument, one accepts the validity of that psychology, how can it apply to explaining profits from capital investment? With respect to capital, what possible analogue could there be to the supposed disutility of labor? Capital goods are produced means of further production; they are reducible to the labor, natural resources, and capital goods used to produce them; and those latter capital goods are in turn reducible to labor, natural resources, and yet other capital goods; and so on. So that, in the limit, capital is reducible to a 'dated' series of labor and natural resource inputs. What then is the painful contribution to production for which capital provision must be compensated by way of profits? The fundamental marginalist answer is the pain of *waiting*.

The provision of capital is understood to result from saving: abstinence from current consumption, or deferral of consumption into the future. It is the disutility of not consuming, or of saving, which must be balanced against the utility of future consumption resulting from current capital investment. Greater saving enables more capital investment and hence, higher consumption goods output in the future; but notionally, as saving and investment increase the increment of disutility from additional saving is rising and the increment of utility from additional capital (the 'marginal product' of capital) is falling. That is to say, supposedly analogous to the rising marginal disutility of labor and diminishing marginal utility of consumption (discussed above), there is rising marginal disutility of saving and diminishing marginal productivity of capital as the capital intensity (or 'roundaboutness') of production increases. At the level of the social economy as a whole, the general rate of profit (renamed 'the rate of interest' in the marginalist framework) is supposed to be determined so as to balance or equilibrate the aggregate supply of capital (saving) and the aggregate demand for capital (or investment, in flow terms). The equilibrium rate of interest induces the required saving. In this marginalist theoretical framework, there is no classical surplus – in the sense of a net product available for free disposal – to be found. Under competitive conditions, *all* remunerations to all 'factors of production' (notably, labor and capital) are 'necessary,' at least at the margin, in the sense that they are functional to ensuring the equilibration of the system.

This theory of the general rate of interest (profit) is embedded in the marginalist framework within a larger system of simultaneous determination of commodity prices, distributive variables (prices of factors of production), and economic activity levels. The economics of Knut Wicksell (1851–1926), who adopts the 'Austrian' approach to capital theory of Eugen von Böhm-Bawerk (1851–1914), provides the best illustration of the logic of profits within this larger theoretical system, from among what may be called the second generation of marginalist economists (e.g., Wicksell, 1934[1901], 1935[1906]). It is one thing to propose a kind of 'parable' in which an individual balances the disutility of work against the utility of consumption, or the disutility of saving (of abstinence from current consumption) against the utility of future consumption. It is another to demonstrate that this vision can be applied to explain outcomes in a mass social economy in which the plans of a myriad of maximizing individuals have to be coordinated so that the aggregated planned demands and supplies for each and every commodity and factor of production (notably, labor and capital) are brought into balance (a 'general equilibrium'). In particular, with regard to our focus here, this requires demonstrating how the rate of interest (along with all the other interacting variables of the complete theory) can coordinate the planned saving and planned demand for capital (or planned investment, in flow terms) of a myriad of individuals, with the saving and investment in general undertaken by different people.

The plausibility of this notion of the rate of interest (profit) as determined by the balancing or equilibration of the supply of and demand for capital (or of planned saving and investment) crucially rests upon the idea of demand functions for factors of production (in particular, capital), in which demand is a negative function of the relative factor price (for capital, the rate of interest relative to other factor prices). This construct has been subject to a decisive critique since the 1950s, although with alarmingly little impact on marginalist orthodoxy (Garegnani, 1990; Kurz and Salvadori, 1995: pp. 427–467). That critique turns upon a certain circularity of reasoning resulting from the heterogeneity of capital goods, which entails that the rate of interest (profit) is already presupposed in the prices of the capital goods that make up any measure of the quantity of aggregate capital demanded. It is to be noticed also that the hoped for marginalist 'general equilibrium' under competitive conditions entails full employment of the factors of production (a balance of demand and supply for factors of production as well as commodities). John Maynard Keynes (1883–1946) attacked this aspect of the marginalist theory in particular. In Keynes's (1936) alternative theory of aggregate economic activity levels in a capitalist economy, based upon his principle of effective demand, rather than the aggregate activity levels and the demand for factors of production adapting to the supply of factors of production (the potential productive capacity of the economy), activity levels and aggregate labor employment are regulated by aggregate effective demand for commodities, with persistent labor unemployment possible if not likely. It follows also from this that, *vis-à-vis* marginalism,

the equilibration of planned saving and investment is turned on its head, so to speak: rather than saving being the prerequisite for investment, investment determines saving via a 'multiplier' mechanism (Garegnani, 1978–79).

Entrepreneurship

In both the classical and marginalist traditions of economic theory, the fundamental object of analysis with respect to profits is the rate of return for the provision of capital in production, as such, notwithstanding very different theories of the profit (interest) rate in these two frameworks. What about the figure of the 'entrepreneur' in capitalism, and her remuneration? There is a loose sense of the term, commonly employed, in which the entrepreneur is merely the primary decision-maker in the firm or economic enterprise: the agent who organizes production and sale and so on. There is also the aspect of risk-bearing in production and sale. In the classical tradition, these two sorts of economic phenomena can be reduced to a species of labor and of capital provision respectively: managerial and organizational functions can be conceived of as a particular kind of labor earning wages, and the theory of profit rates on capital can conceptualize profits as proportional to both the quantity of capital invested and the degree of exposure to risk – risk of business failure or default in particular. (There is also a differential illiquidity aspect to different forms of capital, also generating profit rate differentials under competitive conditions.) The same kind of reduction can be accommodated within the marginalist framework.

But it is evident that in both the classical and marginalist traditions, a positive rate of profit (interest) is not primarily attributed to 'risk,' although risk and differential risk can be incorporated as additional factors in gross rates of return on capital in both approaches. In the former, net profits are fundamentally a realization of part of the distribution of the social surplus, depending in particular upon the forces regulating real wages; in the latter, net profits are a remuneration for 'waiting' – for abstention from consumption (saving). From the standpoint of both theories, an economy subject to complete certainty with respect to economic outcomes – e.g., a steady-state growing economy or a stationary economy, with all the economic agents' expectations of the future completely fulfilled – in general will exhibit a positive, equilibrium general rate of profit (interest).

But there is a stronger sense of entrepreneurship – which is not reducible to some combination of managerial labor and risk exposure – conceptualized in somewhat different ways in seminal contributions by Joseph Schumpeter (1883–1950; 1934[1911]) and Frank Knight (1885–1962; 1933[1921]). In the former, the focus is upon the entrepreneur as the economic agent driving innovation (e.g., technical progress); in the latter, it is exposure to 'uncertainty' of pecuniary outcome in *unprecedented* economic situations – conditions of uncertainty (rather than mere risk) in the radical sense that no well-defined probability distribution over all possible outcomes of the enterprise exists. This uncertainty cannot be reduced to calculable risk for which an ex ante required remuneration could be specified. It is economic action or decision-making in these kinds of conditions or situations, which constitutes the substantial sense of entrepreneurship.

What the contributions of Schumpeter and Knight point to is that the notion of a normal or equilibrium rate of remuneration for entrepreneurship, in that substantial sense of the term – remuneration that would be akin to normal or equilibrium rates of remuneration for labor or capital (under competitive conditions) – does not really make sense. Rather, the entrepreneur, properly understood in this substantial sense, is the 'residual claimant' to the net revenues of an innovative or unprecedented enterprise – the gross revenues net of the remuneration of all other contributors to the enterprise. And more often than not, the residual will be nonpositive; that is to say, the majority of 'entrepreneurs' (properly understood) actually fail. 'Profit,' in some indeterminate sense as to magnitude, might be a motivation (or even the sole motivation) to entrepreneurial activity; but there can be no definite rate of return to entrepreneurship, which can be conceived of as bringing forth an equilibrium supply of entrepreneurship. Entrepreneurship in this strong sense might better be thought of as akin to a scarce natural (human) resource, which may earn a kind of scarcity rent, even if such remuneration is commonly called profits.

Rates, Shares, and Levels

In the discussion of classical and marginalist economics above the focus was upon rates of profit (interest) on capital. What about the share of profits in revenue and the absolute level of profits – whether at the firm, industry, or economy-wide level? If we suppose we have a sound or plausible theory of the general rate of profit (interest) – whether the classical surplus approach, the marginalist approach, or some other – and of equilibrium profit rate differentials, are there equilibrium profit shares and equilibrium absolute levels of profits, associated with the equilibrium profit rates?

Conceptually at least, it is straightforward to derive shares and levels from rates. The share of profits in the gross or net revenue of a firm, an industry, or an economy will be equal to the relevant profit rate times the capital-output ratio of the firm, industry, or economy. (For the share in gross revenue, the 'output' measure will be gross revenue from sales; for the share in net revenue, the output measure will be value added.) Then one can arrive at the absolute level of profits as the relevant profit share times the level of output of the firm, industry, or economy. But are there equilibrium capital-output ratios and equilibrium output levels, analogous to equilibrium profit rates?

The answers are more clear-cut for the marginalist approach. The general equilibrium character of the theory, with all variables simultaneously determined, means that capital-output ratios are simultaneously determined with profit rates, as a consequence of the supposed existence of demand functions for capital, in which capital-output ratios are a negative function of the general rate of interest (profit). In the modern reconstruction of the classical approach, arising out of the seminal work of Piero Sraffa (1898–1983; 1960), it can also be shown that – given the level of real wages or the general rate of profit – the profit-maximizing

capital-output ratios are determinate (whether or not firms and industries have a spectrum of alternative production methods available to them). But there are no demand functions for factors of production, as in the marginalist framework, entailed by, or required for, those results in the modernized classical framework. Furthermore, it is not necessary that *every* firm or enterprise in an industry should exhibit the profit-maximizing capital-output ratio: only the 'dominant' production method employed in an industry will be associated with that capital-output ratio; and firms employing that dominant production method will only exhibit that equilibrium capital-output ratio, and associated equilibrium profit share, when the level of output produced enables 'normal' or desired rates of utilization of the firms' capital stocks.

With regard to absolute levels of profits, also in the marginalist framework economic activity levels are simultaneously determined with profit rates and capital-output ratios. Hence, absolute levels of profits are fully determined in the general equilibrium. These equilibrium activity levels are associated with full employment of factors of production due to the 'supply-side' determination of activity levels in that approach. In classical economics, the issue of absolute levels is more complicated since there is not really a theory of aggregate activity levels to be found there (although Marx has interesting insights on the subject). However, a Keynesian demand-side determination of activity levels and aggregate outputs can be synthesized with the classical approach to income distribution and prices (Garegnani, 1983). A classical-Keynesian synthesis along these lines enables a determination of equilibrium outputs and profits, radically different from the marginalist approach. But this proposition is subject to a caveat, derivative from the qualification above concerning dominant production methods: a demand-side, Keynesian determination of activity levels does not necessarily entail that all firms in the industries making up the economy exhibit equilibrium capital-output ratios, nor that firms are operating at normal or desired rates of utilization of capital.

Independently of Keynes, and at about the same time, Michal Kalecki (1899–1970) also formulated a demand-side approach to explaining economic activity levels – the principle of effective demand – along essentially the same lines as Keynes. From the point of view of the theory of profits, what is of particular interest about Kalecki's formulation of the theory is that the multiplier mechanism, whereby planned investment spending determines planned saving, is shown to be also a principle for determining the absolute level of aggregate profits in the economy as a whole (e.g., Kalecki, 1937). To see this logic, suppose a simple case in which the aggregate income of the economy is divided between wages and profits; 'households' earn the wages and spend them entirely on consumption; and 'firms' earn the profits and engage in spending on investment, the latter spending being determined independently of the level of profits. Aggregate revenue to the firms will be governed by the sum of consumption and investment spending; and their aggregate profits will be that spending net of aggregate wages paid. Hence, since aggregate wages equal consumption spending, aggregate profits are determined by the level of investment spending. The same essential logic continues to apply under more general assumptions (notably, saving by wage earners, consumption by profit recipients, the existence of a public sector, and foreign trade and financial flows).

Conclusion

In the view of the present writer, the modern reconstruction of the classical approach, combined with Keynes's approach to economic activity levels, provides a more robust and plausible framework for the theory of profits than the marginalist approach. As a matter of fact, subsequent to the faltering of the marginalist general equilibrium research program of the immediate decades after World War II, in marginalist economics there is not much interest any more in addressing the 'big question' of the determination of the general rate of return on capital as such (although the notion is more or less entailed, or at least implied, by conventional macroeconomic models). In mature liberal capitalism, with sophisticated financial systems (in particular, largely autonomous credit creation by banks), it is possible for theory in a classical-Keynesian framework to affirm that profit rates on capital in production and interest rates on financial instruments are brought into alignment by competition (not necessarily strict equality), but to turn the traditional causation (from profit rates to interest rates) on its head, so to speak. Instead of the general rate of profit being the independent variable determining the general level of interest rates, the latter – independently determined by monetary policy interacting with financial markets – regulates profit rates. Both Keynes (1936: pp. 202–204) and Sraffa (1960: p. 33), from rather different perspectives, pursued this fruitful idea. A confirmation of the validity of that conception of the determination of the profitability of capital would give transparency to the ultimately arbitrary or conventional character of the magnitude of pure profits (rates, shares, and absolute levels) – profits as the appropriation of a part of the social surplus, rather than a necessary supply price of capital in the manner of marginalism.

See also: Capital: History of the Concept; Capitalism: The History of the Concept; Labor, Division of; Value: History of the Concept.

Bibliography

Garegnani, P., 1978–79. Notes on consumption, investment and effective demand. Cambridge Journal of Economics 2, 335–353, 3, 63–82.

Garegnani, P., 1983. Two routes to effective demand. In: Kregel, J. (Ed.), Distribution, Effective Demand and International Economic Relations. Macmillan, London, pp. 69–80.

Garegnani, P., 1990. Quantity of capital. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: Capital Theory. Macmillan, London, pp. 1–78.

Kalecki, M., 1937. A theory of the business cycle. Review of Economic Studies 4, 77–97.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Macmillan, London.

Knight, F.H., 1933. Risk, Uncertainty and Profit. London School of Economics and Political Science, London (first ed. 1921).

Kurz, H.D., Salvadori, N., 1995. Theory of Production: A Long-Period Analysis. Cambridge University Press, Cambridge.

Marx, K., 1967. Capital: A Critique of Political Economy, 3 vols. International Publishers, New York (first German edns 1867, 1885, 1894).

Ricardo, D., 1951[1821]. On the Principles of Political Economy and Taxation, third ed. In: Sraffa, P. (Ed.), The Works and Correspondence of David Ricardo, vol. 1. Cambridge University Press, Cambridge (first ed. 1817).

Schumpeter, J.A., 1934. The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Harvard University Press, Cambridge, MA (R. Opie, Trans., from the second German ed. 1926; first German ed. 1911).

Smith, A., 1976[1776]. An Inquiry into the Nature and Causes of the Wealth of Nations. In: Campbell, R.H., Skinner, A.S., Todd, W.B. (Eds.), The Glasgow Edition of the Works and Correspondence of Adam Smith, vol. 2. Clarendon, Oxford.

Sraffa, P., 1960. Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory. Cambridge University Press, Cambridge.

Turgot, A.R.J., 1977[1769–70]. Reflections on the formation and distribution of wealth. In: Groenewegen, P.D. (Ed.), The Economics of A.R.J. Turgot. Martinus Nijhoff, The Hague, pp. 43–95 (Trans. of Réflexions sur la Formation et la Distribution des Richesses, in Éphémérides du Citoyen, Paris, November 1769, December 1769, January 1770 (three instalments), from the 1788 ed.).

Wicksell, K., 1934. Lectures on Political Economy, vol. 1 (General Theory). George Routledge & Sons, London (E. Classen, Trans., first Swedish ed. 1901).

Wicksell, K., 1935. Lectures on Political Economy, vol. 2 (Money). George Routledge & Sons, London (E. Classen, Trans., first Swedish ed. 1906).

Program Evaluation

Christiane Spiel, Barbara Schober, and Evelyn Bergsmann, University of Vienna, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Evaluation systematically investigates characteristics and merits of programs and can be done in all phases of a program. The purpose of evaluation is to provide information on the effectiveness of programs so as to optimize outcomes, quality, and efficiency. The article gives an introduction to program evaluation by considering the following four questions: What is program evaluation? At what levels can programs be evaluated? What standards can be set to enhance the quality and fairness of evaluation? How can benefits to users be achieved?

Evaluations are, in a broad sense, concerned with the effectiveness of programs. While common sense evaluation has a very long history, evaluation research, which relies on the scientific method, is a young discipline that has nevertheless grown massively in recent years. In the last few decades, program evaluation has made important contributions to various social domains, e.g., to education, health, economics, and environment. This article gives an introduction to program evaluation by considering the following four questions: What is program evaluation? At what levels can programs be evaluated? What standards can be set to enhance the quality and fairness of evaluation? How can benefits to users be achieved?

What Is Program Evaluation?

Definition, Distinction from Other Activities, Users and Doers of Evaluations

Evaluation, broadly construed, has a very long history. For example, in ancient Rome, tax policies were altered in response to observed fluctuations in revenues, and many thousands of years ago, hunters developed complicated procedures for fishing and trapping. In recent years, common sense program evaluation has evolved into evaluation research, which relies on scientific research methods. This article focuses on this last type of evaluation (for introductory texts on evaluation research see, e.g., Berk and Rossi, 1999; Fink, 2005; Mathison, 2005; Mertens and Wilson, 2012; Rossi et al., 2004; Stufflebeam and Shinkfield, 2007; Wholey et al., 2010).

In what follows, the terms evaluation, program evaluation, and evaluation research are used synonymously. Concern for the first large-scale national evaluations of social programs in the USA first arose as part of the War on Poverty, and methods for monitoring programs are still evolving rapidly. Evaluation has proven its usefulness to such an extent that most federal and state agencies require evaluations of new programs.

Programs are the objects of evaluation, sometimes also called projects or interventions. Program evaluation derives from the idea that social programs should have demonstrable merits and explicit aims by which success or failure may be empirically judged. For example, literacy programs for high school students should lead to measurable improvements in their reading skills. Introductory texts on evaluation describe programs as systematic activities that are provided on a continuing basis to achieve preplanned purposes e.g., a national program to train teachers to work in rural and underserved areas, a campaign to vaccinate all children in a school district (e.g., Fink, 2005; Joint Committee on Standards for Educational Evaluation, 2011).

Evaluation systematically investigates characteristics and merits of programs. It is the purpose of evaluation to provide information on the effectiveness of programs so as to optimize outcomes, quality, and efficiency. Program evaluations are conducted by applying many of the same methods that social researchers rely on to gather reliable and valid evidence. These include formulating the study questions, designing the evaluation, setting standards to assess the program impact, conducting ongoing monitoring of the program functioning, collecting and analyzing data, assessing program effectiveness relative to costs, and reporting results. For examples of evaluation in various fields of clinical and applied psychology see, e.g., Bickman et al. (1995), Gueron and Pauly (1991), Hope and Foster (1992), Maggin et al. (2011), Spiel et al. (2006), or Wong et al. (2008).

While a few authors (e.g., Scriven, 2002) take the view that theories are a luxury for the evaluator and not essential to explanations, the majority of researchers call for injections of theory into every aspect of evaluation design and analysis (e.g., Berk and Rossi, 1999; Coryn et al., 2011; Jacobs et al., 2012; Mathison, 2005). This implies that program evaluation capitalizes on existing theory and empirical generalizations from the social sciences, and is, therefore, after all, applied research.

By relying on the scientific method, program evaluation differs from activities such as newspaper reporting and social commentary. Program evaluation differs from the social sciences in its goals and audience. Evaluation goals are defined by the client, that is, the individual, group, or organization that commissions the evaluator(s), e.g., policy makers. However, often programs and clients do not have clear and consistent goals that can be evaluated for effectiveness. Therefore, to meet evaluation standards, program evaluation should be conducted by professional evaluators who are experts in the social sciences, using a broad repertoire of concepts and methods, but who also have expertise in project and staff management and have at least basic knowledge of the evaluated domain.

While the impetus for evaluation rests in policy, information from evaluations are used by the *stakeholders*, all individuals, or groups that may be involved in or affected by a program evaluation. These are at least six groups (Fink, 2005): (1) federal, state, and local government, e.g., a city,

the National Institute of Health; (2) program developers, e.g., health care service providers, the curriculum department in a school; (3) communities, e.g., a geographically contiguous area, people with a shared health-related problem; (4) policy maker, e.g., a state subcommittee on human services; (5) program funders, e.g., the Department of Education; (6) social researchers, e.g., in government, public agencies, and universities. Therefore, program evaluation is also an interdisciplinary task where evaluators and stakeholders work together (Cousins and Chouinard, 2012).

Key Concepts and Assignment Criteria

Evaluation can and should be done in all phases of a program. *Baseline data* collected before the start of the program are used to describe the current situation, e.g., the need for an intervention. In addition, baseline data are used to monitor and explain changes. For example, if the emergency medical services attempt to reorganize their service to better care for children, they may collect interim data to answer questions such as: In what cases did the emergency medical services fail to meet the needs of critically ill children? What are the differences between their needs and the needs of adults? What is the current performance of medical staff?

Before the start of the program a *prospective evaluation* can be employed to determine the program's potential for realization, its effectiveness, and its impact, that is, the scope of its effects.

In *formative evaluation*, interim data are collected after the start of a program but before its conclusion. The purpose of formative evaluation is to describe the progress of the program and, if necessary, to modify and optimize the program design.

A *process evaluation* is concerned with the extent to which planned activities are executed and therefore is nearly always useful. For example, process evaluation can accompany implementation of a community-oriented curriculum in medical schools.

Outcome evaluation deals with the question of whether programs achieve their goals. However, as mentioned before, goals are often stated vaguely and broadly and therefore *program effectiveness* is difficult to evaluate. In practice, the concept of effectiveness must always address the issue 'compared to what?' Often, it is difficult to distinguish program effects from chance variation and from other forces affecting the outcome. Berk and Rossi (1999) distinguish between marginal effectiveness, relative effectiveness, and cost-effectiveness. *Marginal effectiveness* concerns the dosage of an intervention measure. For example, to prove the assumption that a better student–teacher ratio would improve student performance, the classroom size has to be reduced. In this case, two or more different student–teacher ratios are the doses of intervention and their outcomes are compared.

Relative effectiveness is evaluated by contrasting two and more programs, or a program and the absence of the program. For example, lecture-based learning can be contrasted with problem-based learning in its effects on students' practical performance.

Cost-effectiveness considers effectiveness in monetary units. Evaluation is usually concerned with the costs of programs and their relationship to effects and benefits. For example, in the evaluation of a physical education program, one can ask how much gain in monetary units can be expected due to reduced use of health services.

The entire impact of a program e.g., the extent of its influence in other settings is very difficult, maybe even impossible, to assess. Here, *impact evaluations* are applied, that is, historical reviews of programs are performed after the programs have been in operation for some period of time. If a new medical education program was implemented, an improvement in the quality of the entire health system might be expected as an impact of the program. However, the evaluation of such an improvement might be extremely difficult.

At What Levels Can Programs Be Evaluated?

Programs can be evaluated not only in different phases but also on different levels. Kirkpatrick (1998; Kirkpatrick and Kirkpatrick, 2010; see also Spiel, et al., 2010a) defined four sequential levels; each level is important and has an impact on subsequent levels. The evaluation process becomes more difficult and time-consuming at higher evaluation levels, but it also provides more valuable information. In the following paragraphs, the four levels are briefly described using a parenting program as an example (see Spiel and Strohmeier, 2012).

Level 1 – Reactions

Evaluation on this level measures how program participants react to the program. Therefore, on this level customer satisfaction is measured. For instance, when parents are asked how much they have liked attending a course on parenting styles, this kind of evaluation would tap the level of reactions.

Level 2 – Learning

Learning can be defined as the extent to which participants change attitudes, improve knowledge, and/or increase skills and competencies as a result of attending the program. That means, for example, that this kind of evaluation would investigate whether parents have improved their competencies after attending a course on parenting styles. Such an evaluation could apply questionnaires or tests before and after the course.

Level 3 – Behavior

Evaluation on this level investigates whether changes in behavior have occurred because of attending a program. Thus, the question of whether the learning can be transferred into everyday life is crucial for this kind of evaluation. When using the example of the parents attending a course on parenting styles, the question of whether they would apply what they have learned to their daily parenting behavior is the main goal of an evaluation study focusing on this level. Observations or interviews with the children might be the methods of choice in that specific case.

Level 4 – Results

Here, the focus of the evaluation is on the final results occurring because of participation in the program. The final results can

include, for instance, increased production, improved quality, increased sales, decreased costs, or higher profit. Taking the parenting styles course as an example, an evaluation on this level would measure the adaptation outcomes of the children of participating parents in the long run. However, population and sample selection have to be carefully defined for such an evaluation.

What Standards Can Be Set to Enhance the Quality and Fairness of Evaluation?

Because of the increasing number of evaluations and the high impact that decisions based on results of evaluations may have, there has been an ongoing discussion since the early 1970s about guidelines for effective evaluations. In the absence of any clear definition of what constitutes a reasonable program evaluation, the *Joint Committee on Standards for Evaluation Research* (JCSEE) in cooperation with evaluation and research specialists compiled knowledge about program evaluation and established a systematic process for developing, testing, and publishing evaluation standards. The Joint Committee began its work in 1975. In 1989, the progress for developing standards was adopted by the American National Standards Institute, and became available worldwide.

Originally, the standards were developed to guide the design, implementation, and assessment of evaluations of educational programs, projects, and materials. Since then, the *Program Evaluation Standards* (JCSEE, 2011) have expanded to include applications of the standards in such other settings as medicine, nursing, the military, business, law, government, and social service agencies. The standards are published for people who commission and conduct evaluations as well as for people who use the results of evaluations.

The Joint Committee defined an *evaluation standard* as a principle that is mutually agreed upon by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation (JCSEE, 2011). The standards provide advice on how to judge the adequacy of evaluation activities, but they do not present specific criteria for such judgments. The reason is that there is no such thing as a routine evaluation, and the standards encourage the use of a variety of evaluation methods. The 30 standards are organized around the five important attributes of sound and fair program evaluation: *utility*, *feasibility*, *propriety*, *accuracy*, *and evaluation accountability* (JCSEE, 2011).

Utility standards guide evaluations so that they will be informative, timely, and influential. They require evaluators to keep the stakeholders' needs in mind in all phases of the evaluation. The eight standards included in this category are Evaluator Credibility, Attention to Stakeholders, Negotiated Purposes, Explicit Values, Relevant Information, Meaningful Processes and Products, Timely and Appropriate Communication and Reporting, Concern for Consequences, and Influence.

Feasibility Standards call for evaluations to be realistic, prudent, diplomatic, and economical. They recognize that evaluations usually are conducted in natural settings. The four standards included in this category are Project Management, Practical Procedures, Contextual Viability, and Resource Use.

Propriety Standards are intended to facilitate protection of the rights of individuals affected by an evaluation. They urge evaluators to act lawfully, scrupulously, and ethically. In this category, seven standards are included: Responsive and Inclusive Orientation, Formal Agreements, Human Rights and Respect, Clarity and Fairness, Transparency and Disclosure, Conflicts of Interests, and Fiscal Responsibility.

Accuracy Standards are intended to ensure that an evaluation will reveal and transmit accurate information about the program's merits. They determine whether an evaluation has produced sound information. The eight accuracy standards include Justified Conclusions and Decisions, Valid Information, Reliable Information, Explicit Program and Context Descriptions, Information Management, Sound Designs and Analyses, Explicit Evaluation Reasoning, and Communication and Reporting. *Validity* refers to the degree to which a piece of information or measure assesses what it claims to assess, while *reliability* refers to the degree to which a piece of information is free from 'measurement error,' that is, the exactness of the information.

Evaluation Accountability Standards propose the need to investigate, document, and improve evaluation quality and accountability. The three standards included in this category are Evaluation Documentation, Internal Metaevaluation, and External Metaevaluation. *Metaevaluation* is the evaluation of an evaluation. Internal Metaevaluations are done by the commissioners of the evaluation, whereas External Metaevaluations are conducted by evaluators who are not involved in the evaluation at hand.

For each standard, the Joint Committee has summarized guidelines for application and common errors. In addition, one or more illustrations of the standard's application are given (JCSEE, 2011). However, the Joint Committee acknowledges that the standards are not all equally applicable in every evaluation. Professional evaluators must make a selection, identifying those that are most applicable in a particular context. Evaluation standards are also provided by other committees and associations, such as Guiding Principles for Evaluators (American Evaluation Association, 2004; for further information on different standards see Patton, 2012).

How Can Benefits to Users Be Achieved?

Types of Evaluation Utilization

The utility standards present different approaches to enhance the utility of an evaluation. But there is not just one point of view concerning evaluation utilization; different types of evaluation utilization have emerged in recent decades (Contandriopoulos and Brousselle, 2012; Fleischer and Christie, 2009). Rossi et al. (2004) mentioned three types of utilization of evaluations: instrumental, conceptual, and persuasive/symbolic (see also Balthasar, 2006; Johnson, 1998). Balthasar and Johnson added process use/process-related use, a type that was originally suggested by Patton (1997; see also Alkin, 2005; Weiss et al., 2005). Weiss et al. (2005) added the so-called imposed use to the four aforementioned types. As a matter of course, imposed use is not yet an established type of use. **Table 1** presents descriptions of these five types of evaluation utilization (see also Bergsmann and Spiel, 2009).

Table 1 Types of evaluation utilization

Types of evaluation utilization	*Explanation*
Instrumental use	Indicates "instances where respondents in the study could document the specific way in which the social science knowledge had been used for decision-making or problem-solving purpose" (Alkin, 2005: p. 435).
Conceptual use	Indicates "instances in which the knowledge about an issue influences a policy maker's thinking although the information was not put to a specific, documentable use" (Alkin, 2005: p. 435).
Symbolic use	Indicates instances in which evaluations are used to "justify preexisting preferences and actions" (Weiss et al., 2005: p. 13).
Process use	"The evaluation process itself could have impact; that is, it could lead to utilization … the act of conducting an evaluation and the interactions between the evaluator and various stakeholders can lead to changes in the individuals or in the organisation. Thus it is not the findings of the evaluation that contribute to use but the process itself" (Alkin, 2005: p. 435).
Imposed use	"It is a type of use that comes about because of pressure from outside" (Weiss et al., 2005: p. 16).

User-Oriented Evaluations

The future of evaluation is tied to the future effectiveness of programs (Patton, 1997). However, in order to improve programs, users' perspectives have to be integrated into program evaluation. Moreover, according to Patton (2012), the specific use of an evaluation should be defined by the users. Consequently, user-oriented evaluation is recommended because it requires and promotes systematic and continual critical reflection and feedback. Participants have the opportunity to learn the logic of evaluation and the discipline of evaluation reasoning. Prominent evaluation concepts in this field are utilization-focused (Patton, 1997), empowerment (Fetterman and Wandersman, 2005), participatory (King and Cousins, 2007), and collaborative evaluations (Cousins and Shulha, 2008).

Patton (1997; see also 2012) proposed the concept of utilization-focused evaluation. According to him, the evaluator and the users work together to reach the utilization defined by the users. Patton (2012) divides the evaluation into several steps (see **Figure 1**); each step should contribute to the specific and intended use.

The first step is to start with stakeholder analysis: After identifying potential intended users of the evaluation, the primary intended users should be determined. They are then invited to clarify and prioritize the primary purposes and intended uses of the evaluation. Determining purposes and intended uses of the evaluation is a further crucial step in the utilization-focused evaluation process. This step, including the prioritization of the evaluation questions, is expatiated in

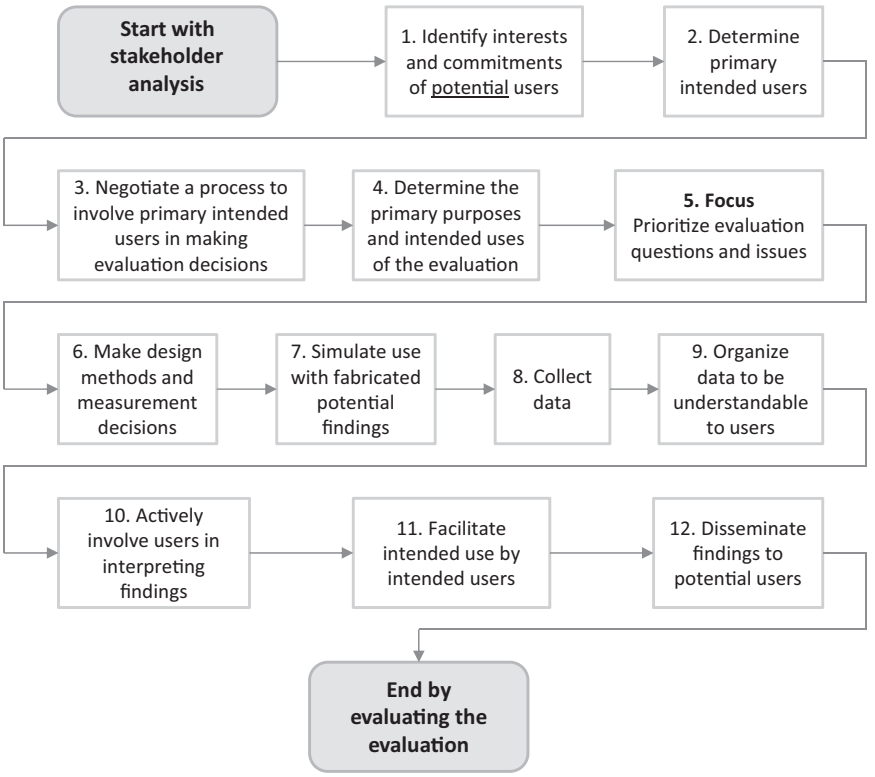

Figure 1 Concept of utilization-focused evaluation.

detail in the following section. The following steps of the utilization-focused evaluation process (Figure 1) are similar to scientific research. However, users should be actively involved in all steps. The final step of the process is a metaevaluation of the evaluation (see also JCSEE, 2011).

Empowerment evaluation is another model that aims at improving the evaluated program, mostly through the development of the skills and autonomy of selected participants (Fetterman, 2005; Fetterman and Wandersman, 2005). It is generally conducted with a pool of program staff members and stakeholders who will be actively involved in identifying evaluation questions, collecting and analyzing data, and formulating next steps.

Goal Explication as an Important Prerequisite for User Orientation

Evaluation goals, including the intended uses, are defined by the clients who commission the evaluation. However, clients often do not have clear and consistent goals – either for their programs or for the evaluation – that can be evaluated for effectiveness. Consequently, they have not developed a kind of working model that relates actions within the program to the program goals. In such cases, professional evaluators should systematically work with the clients and representatives of all stakeholder groups to answer the following central questions (Spiel et al., 2010a; see also Spiel et al., 2012): (1) What are the goals of the program? (2) How should these goals be achieved? (3) How could program effectiveness (goal attainment) be recognized? (4) Are the needed resources available?

A workshop on goal explication is commonly organized to answer these four questions (see Spiel et al., 2010b). Using a wide range of facilitation techniques, the clients and stakeholders are supported in precisely describing their evaluation goals and prioritizing them (Question 1). Very often, the congruence between goals and activities is rather weak. Evaluators help the clients identify how the intended activities might work and how they can contribute to reaching the intended goals (Question 2). This also includes the systematic discussion of whether the resources needed (money and staff, but also the stakeholders' willingness to participate in the program) are available (Question 4). Answering the third question is typically very difficult for the clients of evaluation. They have to identify measureable indicators of effectiveness, labeled 'operationalization' in scientific research (Vogt and Johnson, 2011). Consequently, answering the third question also includes defining the goals of the evaluation. Again, evaluators use facilitation techniques to support the process of identifying the indicators for effectiveness. If possible, criteria for effectiveness are also defined (Fink, 2005). Applying such a procedure should also contribute to establishing an 'evaluative attitude' in the clients and stakeholders in the long run.

Concluding Remarks

In the last few decades, the demands for sound program evaluation have continued to grow in such a way that some authors (e.g., Pawson and Tilley even in 1997) fear it may become mandatory that everything be evaluated. Therefore, it is essential that quality control procedures and standards are introduced for all facets and tasks of evaluation and that the evaluation itself is evaluated against pertinent standards. Furthermore, the user orientation should be systematically considered. Several authors have made recommendations on how to recognize and consider the users' needs. However, there is also the risk of oversimplifying and only using a checklist of standards. But there is no fixed recipe of how to run a successful evaluation.

See also: Alcohol Interventions: Disease Models vs. Harm Reduction; Economic Evaluations and Electoral Consequences; Educational Evaluation: Overview; Empowerment Evaluation; Implementation Science; Performance Appraisal and Evaluation; Social Experiments; Social Psychology: Research Methods.

Bibliography

Alkin, M., 2005. Utilization of evaluation. In: Mathison, S. (Ed.), Encyclopedia of Evaluation. Sage, Thousand Oaks, CA, pp. 434–436.

American Evaluation Association, 2004. Guiding Principles for Evaluators. American Evaluation Association, Fairhaven, MA.

Balthasar, A., 2006. The effects of the institutional design on the utilization of evaluation: evidenced using qualitative comparative analysis (qca). The International Journal of Theory, Research and Practice 12 (3), 1–16.

Berk, R.A., Rossi, P.H., 1999. Thinking about Program Evaluation. Sage, Thousand Oaks, CA.

Bergsmann, E., Spiel, C., 2009. Context, standards and utilisation of evaluation in Austrian federal ministries. In: Fouquet, A., Méasson, L. (Eds.), L'évaluation des politiques publiques en Europe : Cultures et futurs/Policy and Programme Evaluation in Europe: Cultures and Prospects. L'Harmattan, Grenoble, pp. 187–198.

Bickman, L., Guthrie, P.R., Foster, M., Lambert, W.E., Summerfelt, T., Breda, C., Heflinger, C., 1995. Managed Care in Mental Health: The Fort Bragg Experiment. Plenum, New York.

Contandriopoulos, D., Brousselle, A., 2012. Evaluation models and evaluation use. Evaluation 18 (1), 61–77.

Coryn, C.L.S., Noakes, L.A., Westine, C.D., Schroter, D.C., 2011. A systematic review of theory-driven evaluation practice from 1990 to 2009. American Journal of Evaluation 32 (2), 199–226.

Cousins, B., Chouinard, J.A., 2012. Participatory Evaluation Up Close: An Integration of Research-based Knowledge. Information Age Publishing, Charlotte, NC.

Cousins, B., Shulha, L., 2008. Complexities in setting program standards in collaborative evaluation. In: Smith, N., Brandon, P. (Eds.), Fundamental Issues in Evaluation. Guilford, New York, pp. 139–158.

Fetterman, D.M., 2005. Empowerment evaluation. In: Mathison, S. (Ed.), Encyclopedia of Evaluation. Sage, Thousand Oaks, CA.

Fetterman, D.M., Wandersman, A., 2005. Empowerment Evaluation: Principles in Practice. Guilford Press, New York.

Fink, A., 2005. Evaluation Fundamentals: Insights into Outcome, Effectiveness and Quality of Health Programs. Sage, Thousand Oaks, CA.

Fleischer, D.N., Christie, C.A., 2009. Evaluation use: results from a survey of U. S. American evaluation association members. American Journal of Evaluation 30 (2), 158–175.

Gueron, J., Pauly, E., 1991. From Welfare to Work. Russell Sage, New York.

Hope, T., Foster, J., 1992. Conflicting forces: changing the dynamics of crime and community on 'problem' estate. British Journal of Criminology 32 (4), 488–504.

Jacobs, W.J., Sisco, M., Hill, D., Malter, F., Figuerodo, A.J., 2012. Evaluation theory-based evaluation: information, norms, and adherence. Evaluation and Program Planning 35 (3), 354–369.

Johnson, R.B., 1998. Toward a theoretical model of evaluation utilization. Evaluation and Program Planning 21 (1), 93–110.

JCSEE – Joint Committee on Standards for Educational Evaluation, 2011. The Program Evaluation Standards, third ed. Sage, Thousand Oaks, CA.

King, J.A., Cousins, J.B., 2007. Making sense of participatory evaluation: framing participatory evaluation. New Directions for Evaluation 2007 (114), 83–105.

Kirkpatrick, D., 1998. Evaluating Training Programs, second ed. Berrett-Koehler, San Francisco, CA.
Kirkpatrick, D.L., Kirkpatrick, J.D., 2010. Evaluating Training Programs: The Four Levels, third ed. Berrett-Koehler, San Francisco, CA.
Mathison, S. (Ed.), 2005. Encyclopedia of Evaluation. Sage, Thousand Oaks, CA.
Maggin, D.M., Chafouleas, S.M., Goddard, K.M., Johnson, A.H., 2011. A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. Journal of School Psychology 49 (5), 529–554.
Mertens, D.M., Wilson, A.T., 2012. Program Evaluation Theory and Practice: A Comprehensive Guide. Guildford Press, New York.
Patton, M.Q., 1997. Utilization-focused Evaluation: The New Century Text, third ed. Sage, Thousand Oaks, CA.
Patton, M.Q., 2012. Essentials of Utilization-focused Evaluation. Sage, Thousand Oaks, CA.
Pawson, R., Tilley, N., 1997. Realistic Evaluation. Sage, London.
Rossi, P.H., Lipsey, M.W., Freeman, H.E., 2004. Evaluation: A Systematic Approach, seventh ed. Sage, Thousand Oaks, CA.
Scriven, M., 2002. Evaluation Thesaurus, fourth ed. Sage, Newbury Park, CA.
Spiel, C., Gradinger, P., Lüftenegger, M., 2010a. Grundlagen der Evaluationsforschung [Evaluation research basics]. In: Holling, H., Schmitz, B. (Eds.), Handbuch Statistik, Methoden und Evaluation [Handbook Statistics, Methods, and Evaluation]. Hogrefe, Göttingen, pp. 223–232.
Spiel, C., Lüftenegger, M., Gradinger, P., Reimann, R., 2010b. Zielexplikation und Standards in der Evaluationsforschung [Goal explication and standards in evaluation research]. In: Holling, H., Schmitz, B. (Eds.), Handbuch Statistik, Methoden und Evaluation [Handbook Statistics, Methods, and Evaluation]. Hogrefe, Göttingen, pp. 252–260.
Spiel, C., Schober, B., Reimann, R., 2006. Evaluation of curricula in higher education: challenges for evaluators. Evaluation Review 30 (4), 430–450.
Spiel, C., Schober, B., Reimann, R., 2012. Modelling and measurement of competencies in higher education: the contribution of scientific evaluation. In: Zlatkin-Troitschanskaia, O., Blömeke, S. (Eds.), Modeling and Measurement of Competencies in Higher Education. SensePublishers, Rotterdam, pp. 183–194.
Spiel, C., Strohmeier, D., 2012. Evidence-based practice and policy: when researchers, policy makers, and practitioners learn how to work together. European Journal of Developmental Psychology 9 (1), 150–162.
Stufflebeam, D.L., Shinkfield, A.J., 2007. Evaluation Theory, Models, & Applications. Jossey-Bass, San Francisco, CA.
Vogt, W.P., Johnson, R.B., 2011. Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences. Sage, Thousand Oaks, CA.
Weiss, C., Murphy-Graham, E., Birkeland, S., 2005. An alternate route to policy influence: how evaluations affect D.A.R.E. American Journal of Evaluation 26 (1), 12–30.
Wholey, J.S., Hatry, H.P., Newcomer, K.E., 2010. Handbook of Practical Program Evaluation, third ed. Wiley, San Francisco, CA.
Wong, V.C., Cook, T.D., Barnett, W.S., Jung, K., 2008. An effectiveness-based evaluation of five state pre-kindergarten programs. Journal of Policy Analysis and Management 72 (1), 122–154.

Relevant Websites

http://www.eval.org/ – American Evaluation Association (AEA).
https://netforum.avectra.com/eWeb/StartPage.aspx?Site=APPAM&WebCode=HomePage – Association for Public Policy and Measurement (APPAM).
http://www.degeval.de/ – DeGEval – Gesellschaft für Evaluation.
http://www.europeanevaluation.org/ – European Evaluation Society (EES).
http://www.europeanevaluation.org/community/nese.htm – Network of Evaluation Societies in Europe (NESE).
http://www.evaluation.org.uk/ – UK Evaluation Society (UKES).

Progress: History of the Concept

Dorothy Ross, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Extension of the progress of science into the realm of government and morals was central to Enlightenment belief in historical progress. During the nineteenth century, progress provided the framework and central problem of social science theory, and belief in the progress of the social sciences became the ideology of the social science disciplines. Discontents with progress had been present from the start and the conviction that progress was inevitable collapsed in the twentieth century, although in mitigated forms, it waxed and waned with historical expectations. At the end of the century, postmodern sensibilities challenged the progress both of history and the social sciences.

The idea of progress – that "civilization has moved, is moving, and will move in a desirable direction" (Bury, 1932) – was integral to the project of creating social sciences in eighteenth century Europe, and the relationship has continued since. During the nineteenth century, the idea of progress formed the historical framework and central problematic of social science theory and practice. One component, the idea of the progress of the social sciences figured prominently as the ideology of the social science disciplines. From its origin, however, the idea of progress was accompanied by recognition that progress created new problems and by doubts that the modern world coming into existence would be superior to the past or present. The belief that progress was inevitable collapsed in the disasters of the twentieth century, although belief in the possibility of progress continued in mitigated forms, waxing and waning with the portent of historical events, the expectations of different national cultures, and the character of the social sciences.

This article embraces a theoretical tension that runs through recent studies of the human and social sciences. The project of creating social sciences was both a shared idea, aspiration, or program aimed at social improvement by means of science; and a framework, a mode of understanding the world within which certain kinds of ideas and practices rather than others came into being. In this latter sense, the program of science aimed at social improvement was a set of possibilities and constraints within which the social sciences formed. In the spirit of this duality, studies of the social sciences have emphasized the importance of language as both speech act and the medium in which meanings are produced, especially those integrated bodies of language called discourses that form from extended conversations around a shared problem. A discourse provides a bounded language for varied semantic use. Variant discourses of progress were the medium within which social scientists formed their purposes and within which they represented those purposes as true and legitimate (Porter and Ross, 2003).

Eighteenth-Century Origins

The discourse of progress and the social science project emerged in early modern Europe as part of a new understanding of history as a realm of human construction, propelled ever forward in time and taking new qualitative forms. Modern society was decisively different from its feudal and ancient forerunners, engaged upon a novel historical course. Replacing earlier cyclical and supernatural views of historical time, historicism instantiated in human history the teleological movement and universal scope that governed the Christian drama of redemption. During the eighteenth century, enlightened thinkers argued that the advance of reason was moving history in a progressive direction, even to a utopian end. The emergence of the social sciences was itself a linchpin of progress. Science was understood as the most advanced form of reason, and the extension of the progress of science into the realm of government and morals – barely begun – was central to the expectation of civilizational advance. The array of human differences across the globe also sharpened for European thinkers the progressive character of their own society and the stages by which they had reached modernity (Meek, 1976). The task of the social sciences was to understand the forces that governed the modern European world and to guide it into a better future.

The eighteenth-century discourse of progress drew upon, and was often coincident with, the early formulations of social science. Associationist psychology explained the ability of human mental capacities to improve over time, while the studies of historians and economists and the observations of travelers set the trajectory of progress toward greater complexity and refinement with increasing social differentiation and intercourse came improving manners, morals, knowledge, and rationality. The guiding role of providence in this process was still visible, but it was understood to work through nature. Particularly in Britain and France, natural principles assured the existence and continuation of progress.

Several major theories of progress emerged within this discourse. Scottish 'philosophical historians' like Adam Smith and Adam Ferguson developed four- or five-stage theories of history in which the economic mode of production at each stage shaped political organization, customs, and moral temper. Economic growth, fueled by the expansion of population, the division of labor, and technological advance, emerged within a complex social process, for history was often the result of unintended consequences. In contrast, the philosophers of the French Enlightenment like Condorcet focused on reason as the overriding factor in historical progress.

The rational acts of enlightened individuals, and particularly of utilitarian science, moved humanity forward, ignorance and war held it back. In Germany, Herder and Kant made historical progress the work of philosophical and moral reason. History was the progressive fulfillment of humanity's moral and spiritual destiny, realized through the unique development of each historical people.

Emerging in contested political arenas, the early theories of progress staked out ideological positions of long-standing influence in politics and the social sciences. Designedly, Smith's social historical view lent itself to a liberal presumption against state action; Condorcet's, to a justification of the authority of social scientists in the guidance of society; Kant's, to a statecraft subordinated to idealist ends. Moreover, if the reliance on progressive reason initially assigned all evil to reactionary enemies, Scottish republicans and German idealists recognized from the start that progress necessarily involved losses. By the end of the eighteenth century, Rousseau's contention that progress in reason involved instinctual loss; Ferguson's analysis of the political, social, and psychological costs incurred by economic and social progress; Schiller's summation of these complaints into a 'classic diagnosis of the ills of modern society in terms of division, fragmentation, isolation, conflict, and distortion'– all drove a wedge between material and moral progress and set the foundation for a long line of variants on the discourse of historical advance that made progress itself responsible for the problems of modern society (Abrams, 1971: p. 211).

The Nineteenth-Century Triumph of Progress

Despite these misgivings, the American and French revolutions aroused millennial hopes throughout Europe and when the American Revolution succeeded and the French Revolution was destroyed from within and without, those hopes were projected into the secular future. The rapid advance of science and its technological adjuncts began to remake the Atlantic world, heralding an ever-improving future. Progress became the dominant view of history among the educated classes of Europe and North America, while contesting versions of progress proliferated. Responding to rapid capitalist development and the continuing possibility of political and social revolution, theorists of political economy, sociology, and anthropology recast the eighteenth-century theories of progress into social scientific discourses that specified the factors that controlled the course of history.

Classical political economy, drawing on the Scottish theorists, focused on the natural laws of economic growth and, of equal importance, on how the expanded product would be distributed. The question of whether the average workman would be better off, despite growing inequality, became urgent as egalitarian, utopian versions of progress appeared. The Reverend Thomas Malthus argued that food supply would keep wages at subsistence level. His political economy, informed by an Anglican reading of history, supported a cautious view of progress, incorporating cyclical movement and the necessity of suffering. The theories of David Ricardo were often pressed into this mold, though most practicing economists, including Ricardo, maintained a more hopeful account of economic growth based on capital accumulation. By midcentury, recognition of the role of technology in increasing productivity shifted the weight of economic opinion toward more optimistic predictions of future well-being (Berg, 1990).

American exceptionalists and socialist radicals transformed this liberal political economy in opposite directions. Henry C. Carey argued that the uncultivated American continent would abrogate altogether the action of Malthusian law, vindicating America's utopian historical course (Ross, 1991). Socialists, charging the capitalist economy with the degradation of workmen and the destruction of community, made capitalism an intermediate stage, to be superceded by socialism. In Karl Marx's formulation, history progressed by a process of historical contradiction between the forces and social relations of production: capitalism produced within itself the revolutionary proletariat that would destroy it and the technological basis of abundance that would support socialism and ultimately, communism (Cohen, 1978).

Auguste Comte, inheriting Condorcet's identification of historical progress with progress of the human mind, made social science its central agent. Civilization moved by necessary law from the theological to the metaphysical to the ultimate stage, the scientific, still struggling to be born. Although this trajectory was inevitable, action in accord with historical law could ease the transition. A new science of society, sociology, would objectively examine the world as it is and, based thereon, set out a plan for the rational ordering of society, opinion, and feeling (Lenzer, 1975). Herbert Spencer's version of sociology, influenced by liberal political economy and Scottish philosophical history, removed the central planning role from sociology, but gave it the same power to discern the laws of historical progress to which successful social action must conform. More complacent about his own era, Spencer believed that the industrial age would encourage moral, intellectual, and social improvement. Drawing on biological notions of adaptation to the environment, the inheritance of acquired characteristics, and an embryological model of differentiation, Spencer turned historical progress into an aspect of social and cosmic evolution (Peel, 1971).

Charles Darwin's *The Origin of Species* (1859) reinforced both the idea of progress and the use of biological analogy. Despite the radically nonprogressivist implications of natural selection, Darwin's theory, with its suggestion of the continuity of animal and human life, was often read in this era of faith in progress – even by Darwin himself – as proof of the progressive character of both evolution and history. The contemporaneous recognition of the long age of the earth created a space for 'prehistory,' solidifying the uniquely advanced position of European civilization (Bowler, 1989). Social scientists developed a 'comparative method' that sorted peoples and races on an evolutionary grid, assigning them to stages of a single evolutionary process by Euro-centered hierarchical standards.

Nineteenth-century Romantic critics of the Enlightenment discourse of progress also worked within its progressive secular conception of historical time. Even as they argued that analytical reason and the conventional society being created by economic development were destroying humanity's instinctual, aesthetic, and spiritual resources, they retained a vision of utopian possibility, available in the present to genius and in the future to all humanity through the creative imagination.

The Problem of Progress, 1880–1914

By the turn of the twentieth century, however, historical events and cultural criticism placed increased burdens on the belief in progress. The modernist heirs of romanticism often presented a dire picture of modernity as a desiccated world of mechanism, instrumental reason, and alienation. Even for those who continued to believe that the benefits of science and economic growth far outweighed their drawbacks, the problems created by rapid industrialization, mass democracy, and imperialism were multiplying. The social sciences formed specialized, university-based disciplines in large part to define and grapple with the problems of modernity. In that context, Enlightenment hopes for modernity took chastened forms. At the same time, more stringent conceptions of science outlawed deterministic laws of history like those of Comte and Spencer as metaphysical constructions and made the moral valuation inherent in the concept suspect. By the early twentieth century, the rediscovery of Mendelian genetics attenuated the link between evolutionary biology and historical change, and critics, notably anthropologist Franz Boas, attacked the comparative method and its assumption of a single line of evolutionary advance. Progress no longer appeared to social scientists as an inevitable, universal process, guaranteed by natural or historical law, but the assumption that history moved in a progressive direction remained deeply woven into the discourses of the social science disciplines.

In economics, marginalism shifted the focus of attention from economic growth to allocation. Those who joined the mixed neoclassical discourse grounded capitalism in laws analogous to the laws of physical nature and retained optimistic assumptions of economic growth and its central role in social advance, but those assumptions no longer framed their specialized and dehistoricized discourse. Other economists who sought a qualitative change in capitalism worked within a self-consciously historical or institutional framework, like the socialist Thorstein Veblen who asked in 1899 "Why Is Economics Not an Evolutionary Science?" (Ross, 1991).

The classic sociological theorists of this period took as their problematic the integration of individual and society and reduced the trajectory of history to two or three stages of modal social organization. In Germany, where antimodernist attitudes flourished, sociologists were most pessimistic. For Ferdinand Tönnies, traditional face-to-face community (*Gemeinschaft*) and modern capitalist society (*Gesellschaft*) marked polar modes of social organization. The impersonality and loss of spirituality of modern society were inevitable consequences of its organizing principles. Max Weber saw the positive advantages of the instrumental rationality and secularism that characterized modernity, yet termed the framework it built for modern society an 'iron cage' (Liebersohn, 1988).

Emile Durkheim, like American sociologists, retained the more hopeful nineteenth-century formulation that placed the present troubled society in transition to a more harmonious modernity. Durkheim argued that a new kind of functional and normative social integration was forming; the anomie of individuals thrown outside the regulative norms of society by the upheavals of social transformation would subside as modern society developed new forms of social solidarity. Edward A. Ross posited a transition from the natural social control of face-to-face community to the more fragile, 'artificial' controls of modern society, similar to the German model of *Gemeinschaft*/*Gesellschaft*, but he too expected that a more harmonious adjustment would be reached (Ross, 1991).

It was primarily in thinking about their own role in modern society that social scientists explicitly reaffirmed the central role of science in progress. In his presidential address to the American Psychological Association in 1899, John Dewey expressed both the anxieties of modernity and the solution favored by his colleagues. 'With tremendous increase in control of nature,' he noted, "we find the actual realization of ends, the enjoyment of values growing unassured and precarious." But, he added, "the entire problem is one of the development of science and of its application to life." When the social sciences are further developed, "we can anticipate no other outcome than increasing control in the ethical sphere" comparable to "the revolution that has taken place in the control of physical nature" (Dewey, 1976: pp. 149–150). That science itself progressed, that it produced increasing control over nature and social life, and that such control worked to the benefit of humanity remained unquestioned. This faith served as the ideological basis for the social science disciplines.

Progress Dethroned, 1914–50

The modern West's confidence in progress could not survive the massive destruction of the two world wars, the onset of economic depression, and the rise of totalitarian politics. Science and technology, hitherto proof of the progress of reason, were deeply implicated in these disasters. At the same time, a sharp sense of historical discontinuity, breaking off the new world of mass production, mass consumption, and mass politics from the past, destroyed the sense of cumulative gain in civilization on which progress depended. Historians mark World War I as the end of progress as an unquestioned public faith and as an idea that social scientists could presuppose or hope to demonstrate.

The toll on discourses of progress was greatest in Europe, which bore the brunt of the devastation and upheaval. Many threads of progressive theory were enfeebled or drawn into the orbit of totalitarian politics, as in Soviet Russia's Marxist–Leninist ideology of the revolutionary state and the 'new' man and woman. Other liberal and socialist variants of the discourse of progress survived in the Scandinavian forging of welfare states and in Austria's brief melding of empirical social science and socialist meliorism. Two notable efforts to sustain the possibility of progress despite dire views of the present were to have substantial influence during the second half of the century, particularly in the United States. In Austria and England, Friedrich Hayek – deploring the socialist direction of interwar economies – laid the foundation for neoliberalism in his *Road to Serfdom* (1944). And a group of mostly Jewish social theorists at the Frankfurt Institute for Social Research combined Marx's narrative of socialist progress and the modernist critique of capitalist society and culture. Exiled after 1933 to the United States, they found the basic source of modern evils in the radical separation of human beings from nature effected by modern science. The Enlightenment's objectification of nature objectified humanity itself, turning the human world into

a field for manipulation and political domination. Their vision of utopia no longer marked an historical future, but served only as a flickering standard by which to criticize the present (Jay, 1996).

The possibility nonetheless remained open, particularly in the United States, of constructing limited lines of progress and maintaining the hope that limited gains worked toward the larger progress of history. Thus the American sociologist William F. Ogburn explained that social scientists had dropped the term progress and were instead studying *social change* as a 'term free from dogmatic or moral implications.' Social change is caused by inventions, Ogburn argued, but inventions occur unevenly, creating 'lags' until society 'adjusts.' In his analysis, however, adjustments can always be made and adjustments always resulted in improvements. Progress was not so much abandoned as forced underground (Ogburn, 1930).

The point at which it remained still vibrantly alive, especially in the United States, was the faith in science. The neopositivist understanding of science itself suggested that science's pragmatic relation to reality allowed it to reconstruct the world to human purposes (Porter, 1994). While Ogburn was moderate in his hopes for prediction and control, other American social scientists of the 1920s, like the behaviorist psychologist John B. Watson and the political scientist Charles Merriam, expressed utopian faith in science and technology (Ross, 1991). Although a more sober tone emerged throughout the West in the 1930s, the possibility of historical progress through science remained a central feature of American social science theory and ideology.

Progress as Americanization, 1950–70

In the decades after World War II, the possibility of a progressive modern society guided by social science gained energy and urgency from the defeat of fascism, the disintegration of colonial empires, and the threat of communism. As the strongest power to emerge from the war, the United States embraced that possibility with a mixture of self-confidence and anxiety. In what was experienced as the 'American Moment' of an 'American Century,' the country already appeared to stand at the summit of world history, ready to embody the values its exceptionalist history promised. At the same time, Cold War challenges and the uncertainties of international leadership created palpable anxiety.

The discourse of progress continued to define the social science disciplines' hopes for modernity and their own central role. Both were explicitly embodied in modernization theory, which set out the distance the underdeveloped world still needed to traverse to achieve the Western, or more specifically American, norms of economic growth, democracy, and individualism (Gilman, 2003). Defined as a 'process of social change that seeks to govern itself by rational policy planning,' modernization declared social scientists' role in that process 'indispensable' (Lerner, 1968). During the same decades, University of Chicago economist Milton Friedman turned Hayek's complex arguments against state control of the economy into a simplified neoliberal ideology that promised a virtually utopian outcome to the organization of economy and society on free market principles – a prescription that would later become influential within and outside the United States (Burgin, 2012).

Considerable complexity and uncertainty were nonetheless visible in postwar social science. Economists Alexander Gerschenkron and Albert O. Hirschman developed more sophisticated theories of economic development than the modernizationists' linear model. The leading social science theories of the postwar decades modeled functionalist, cybernetic, and equilibrium systems that expressed the desire for stability as well as the sense of progress achieved. In the work of David Riesman and studies of the 'authoritarian personality' that drew on the personnel and ideas of the Frankfurt school, social scientists found a 'soft' totalitarianism of conformism in the United States parallel to the 'hard' totalitarianism of Nazi Germany and the Soviet Union, although these darker fears did not close off the possibility of progress.

As part of its international leadership role, America worked to export its models of social science theory and research. In Europe, American influence peaked by 1970 as native traditions revived and new theorists analyzed the possibility of progress with greater circumspection than in the United States. The *Annales* historians disaggregated the flow of historical time and turned their attention to the structure of *la longue durée*. In German sociology, the dehumanizing character of modern society remained a powerful, and sometimes dominant, theme. One of the most hopeful German theorists, Jürgen Habermas blended Marxist traditions with American theories of pragmatism and democracy to project not a utopian future, but a utopian ideal against which modern society might measure its defects (Skinner, 1985).

The idea of progress and its drawbacks also played a conspicuous role in the development of the social sciences in the non-Western world. As an arm of secular, Western-style development since the nineteenth century, the social sciences had been a major site of debate over whether or how far to follow Western models of economic growth or revive traditional cultures. After World War II, the developing world also served as a major source of critical revision of Western conceptions of progress as modernization: Latin American economists recast modernization as economic dependency; Japanese social theorists linked the country's rapid economic growth to its blending of capitalist forms and traditional feudal culture. Proliferating formulations of modernity redefined tradition and modernity as mutually transformative rather than diametrically opposed processes (Gaonkar, 2001).

Progress and Postmodernity, 1970–2010

Since the 1970s, long-standing discontents with modernity and the global reach of developmental processes have opened the way for a variety of conflicting discourses on progress. The critique of the positivist conception of scientific progress by Thomas Kuhn's *The Structure of Scientific Revolutions*, 1962, launched the most consequential challenge for the social sciences. Although Kuhn argued that his work did not undermine the rationality of science nor destroy the idea of scientific progress, it did radically remove science from positivism's universalistic narrative of the cumulative, progressive discovery

of truth and resituate it in a social historical world of shifting contexts and revolutionary change (Skinner, 1985). Kuhn's work presented the social sciences with a double challenge. He was skeptical that most parts of the social sciences possessed a paradigm – a shared model of scientific practice – and hence qualified as mature sciences. And by historicizing science, he suggested that the positivist premises of mainstream social science were misguided; the often-marginalized interpretive models of social knowledge based on historicist premises could legitimately claim a central place in the social sciences.

Although Kuhn's critique of the positivist conception of scientific progress had roots in a scientific philosophy allied with positivism, it was soon joined to the mounting philosophical criticism of positivism and to the broader attack on the discourse of progress by postmodern theorists (Isaac, 2012; Bernstein, 1985). The most radical postmodern challenge came from the home of the rationalist Enlightenment. Francois Lyotard linked the idea of progress, with its universalist and rationalist premises, to the most oppressive, totalitarian impulses of modern Western society. Michel Foucault held the social sciences and professions responsible for the knowledge/power that disciplined the unruly masses of modern society in the name of enlightened rationality. Animated by pluralistic and libertarian values, this critique gained power from the proliferation of postcolonial nations and by opposition to the increasing reach of Western capitalism into every area of the world. In the postmodern reading, globalization was producing not a single narrative of universal progress, but a pluralistic, hybrid, decentered world of many historical narratives. If, as the term postmodern suggests, this vision does not entirely escape modernity and can itself be read as a liberationist narrative of progress drawn from Romantic rather than Enlightenment sources, it nonetheless called into question the universalist and rationalist premises upon which discussions of progress had been conducted in the social sciences since the eighteenth century (McGowan, 1991).

In those segments of social science identified most closely with mathematics and natural science, like economics, Kuhn and postmodernism have had little influence. The framework of historical progress held strong in economics, where it remained implicit in the economists' assumption of the beneficent operation of the ideal market and explicit in the neoliberal gloss on free market economics. The economists' own ideology of scientific progress also held firm in continued positivist understandings of science and claims to possession of a consensual paradigm (Backhouse, 1997).

Postmodern currents have more powerfully undermined the discourse and ideology of progress in social science disciplines other than economics, where the political conflicts and social movements of the 1960s and 1970s were more keenly felt. As a result, divisions between positivist and historicist/hermeneutic fields within these disciplines multiplied and deepened (Backhouse and Fontaine, 2010). Anthropology was most deeply shaken by the postmodern assertion that an objective Western science of peoples classified as 'others' was impossible and immoral. In sociology, the historical trajectory of progress from traditional to modern society which had guided social theory through the nineteenth and twentieth centuries appeared less relevant to the issues raised by globalization. As sociological theories built on that scaffolding collapsed, sociologists' visions of historical advance shifted to the multiracial, multicultural, multigendered world valued by postmodernists (Calhoun, 2007).

The power of the economics discipline, together with the fall of communism, continued globalization, and the rise of neoliberalism, revived the discourse of progress in the public sphere. In the United States, the narrative of American exceptionalism and the view of globalization as the Americanization of the world returned. Social scientists formulated and debated more complex narratives of universal progress toward the present form of capitalist society, such as Frances Fukuyama's 'end of history' in the Euro-American model of liberal democracy, David Landes' attribution of Western economic growth to its cultural virtues, and Jared Diamond's geographical explanation for the dominance of the modern societies of the Eurasian landmass. Despite the postmodern critique of science and the many fears of science's destructive capacities, the scientific and technological conquest of nature often remained the chief motor and guarantee of progress in these narratives and in the self-conception of the social sciences themselves. Still, the context of fragmentation and critique forced some social scientists to abandon, and others to reconsider, the ability of their disciplines to meet the criteria of scientific progress (Rule, 1997; Smelser, 2005). If the eighteenth-century formulation of historicism launched the discourse of progress and gave the social sciences their charter, the critical historicism of the late twentieth century had called both progress and that charter into question.

See also: Enlightenment: Impact on the Social Sciences; Frankfurt School: Institute for Social Research; Historicism; Kuhn, Thomas S. (1922–96); Linguistic Turn and Discourse Analysis in History; Modernization and Modernity in History; Positivism, History of; Social Science, The Idea of.

Bibliography

Abrams, M.H., 1971. Natural Supernaturalism: Tradition and Revolution in Romantic Literature. W.W. Norton, New York.

Backhouse, R.E., 1997. Truth and Progress in Economic Knowledge. Edward Elgar, Cheltenham, UK.

Backhouse, R.E., Fontaine, P. (Eds.), 2010. The History of the Social Sciences Since 1945. University of Cambridge Press, Cambridge, UK.

Berg, M., 1990. Progress and providence in early nineteenth-century political economy. Social History 15, 365–375.

Bernstein, R.J., 1985. Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis. University of Pennsylvania Press, Philadelphia, PA.

Bowler, P.J., 1989. The Invention of Progress: The Victorians and the Past. Blackwell, Oxford, UK.

Burgin, A., 2012. The Great Persuasion: Reinventing Free Markets Since the Depression. Harvard University Press, Cambridge, MA.

Bury, J.B., 1932. The Idea of Progress. Macmillan, New York.

Calhoun, C. (Ed.), 2007. Sociology in America: A History. University of Chicago Press, Chicago.

Cohen, G.A., 1978. Karl Marx's Theory of History. Princeton University Press, Princeton, NJ.

Dewey, J., 1976. Psychology and social practice. In: Boydston, J. (Ed.), John Dewey, The Middle Works, vol. 1. Southern Illinois University Press, Carbondale, IL, pp. 131–150.

Gaonkar, D.P., 2001. Alternative Modernities. Duke University Press, Durham, NC.

Gilman, N., 2003. Mandarins of the Future: Modernization Theory in Cold War America. Johns Hopkins University Press, Baltimore, MD.

Isaac, J., 2012. Working Knowledge: Making the Human Sciences from Parsons to Kuhn. Harvard University Press, Cambridge, MA.

Jay, M., 1996[1973]. The Dialectical Imagination: A History of the Frankfurt School and the Institute for Social Research 1923-1950. University of California Press, Berkeley, CA.

Lenzer, G., 1975. Auguste Comte and Positivism. Harper & Row, New York.

Lerner, D., 1968. Modernization: social aspects. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences. Macmillan and the Free Press, New York.

Liebersohn, H., 1988. Fate and Utopia in German Sociology, 1870–1923. MIT Press, Cambridge, MA.

McGowan, J., 1991. Postmodernism and Its Critics. Cornell University Press, Ithaca, NY.

Meek, R.L., 1976. Social Science and the Ignoble Savage. Cambridge University Press, Cambridge, UK.

Ogburn, W.F., 1930. Change, social. In: Seligman, E.R.A. (Ed.), Encyclopaedia of the Social Sciences, vol. 3. . Macmillan, New York, pp. 330–334.

Peel, J.D.Y., 1971. Herbert Spencer: The Evolution of a Sociologist. Basic Books, New York.

Porter, T.M., 1994. The death of the object: Fin de siecle philosophy of physics. In: Ross, D. (Ed.), Modernist Impulses in the Human Sciences 1870–1930. Johns Hopkins University Press, Baltimore, MD.

Porter, T.M., Ross, D. (Eds.), 2003. The Modern Social Sciences: The Cambridge History of Science, vol. 7. Cambridge University Press, Cambridge, UK.

Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, Cambridge, UK.

Rule, J.B., 1997. Theory and Progress in Social Science. University of Cambridge Press, Cambridge, UK.

Skinner, Q., 1985. The Return of Grand Theory in the Human Sciences. Cambridge University Press, Cambridge, UK.

Smelser, N.J., 2005. The questionable logic of 'mistakes' in the dynamics of knowledge growth in the social sciences. Social Research 72, 237–262.

Projective Methods in Psychology

George Stricker and Stephen J Lally, American School of Professional Psychology, Argosy University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Stricker and K. Somary, volume 18, pp. 12181–12185, © 2001, Elsevier Ltd.

Abstract

Projective techniques is a term that refers to any test or procedure designed to provide information about individuals by allowing them to respond freely to ambiguous stimuli. This article reviews the history of these techniques and the unique insights they provide. Additionally, the article focuses on four of the most frequently used specific projective tests, the Rorschach, the Thematic Apperception Test, projective drawing tests, and sentence completion tests. These tests' administration is reviewed and their psychometric strengths and limitations are listed. Finally, the article concludes by identifying the place of projective methods within the larger sphere of psychological testing.

Projective techniques is a term that encompasses any test or procedure designed to increase insight into individuals by allowing them to respond freely to ambiguous stimuli. The underlying assumption is that, when faced with unstructured or ambiguous stimuli, people will reveal aspects of their personality in their attempt to structure the material. In other words, the vague quality of the stimulus evokes free fantasy play and triggers projection on the part of the subject while simultaneously revealing aspects of psychological functioning in the nature of their responses. Many experts believe that everything said or written in response to a stimulus has dynamic meaning and that people reveal conflicts, anxieties, attitudes, beliefs, and needs in their attempt to structure or complete an incomplete stimulus. Projective techniques tend to have brief, general instructions that, when coupled with ambiguous stimuli, tend to elicit less defensive responses than more structured testing. It should be noted that a concern about conflicts, anxieties, attitudes, beliefs, and needs, and the usual attempts to blend these into a picture of a person, is most closely associated with one approach to personality, the psychoanalytic approach. A preference for psychodynamic theory is not required in order to use projective techniques, but it is frequently found. Perhaps this is because one of the hallmarks of the psychoanalytic approach is the prominent role assigned to unconscious processes, and projective techniques are designed with the intention, satisfied or not, of eliciting unconscious material.

Historical Overview

Historically, the turn of the twentieth century was a time when psychological testing was applied predominantly for the study of intelligence. The trend was toward finding data or norms descriptive of groups of people (nomothetic data). At this time, personality measures tended to examine only a single trait at a time (e.g., introversion/extroversion, flexibility). The scope began to broaden about two decades later when interest surfaced in tests that yielded specific data about an individual (idiographic data). In the 1920s and 1930s, a variety of projective testing instruments emerged that relied in part or whole on the concept of projection. The notion of projection with these tests gave birth to the progression of projective techniques throughout the rest of the twentieth century and the belief that ambiguous stimuli can be used to gain insight into the dynamics of the respondent. Frank (1939) coined the term 'projective methods' to describe a variety of techniques used by clinicians with ambiguous stimuli that pull for projection. He postulated a projective hypothesis, namely, that the world is structured according to internal needs and external demands, so that the ambiguous external stimuli will elicit structuring based on internal factors. Projective tests varied on the nature of the stimuli and the tasks required; some require association to inkblots (e.g., Rorschach), others completion of incomplete sentence stems (e.g., sentence completion tests), and still others construction either of drawings (e.g., Draw-A-Person (DAP) test) or stories (Thematic Apperception Test (TAT)). The 'Projective Movement' grew tremendously during the 1940s and 1950s, as clinicians moved away from the nomothetic approach to studying their patients toward the more idiographic approach. More recently, because of psychometric considerations, and possibly because of a reduction in the number of clinicians trained in psychodynamic approaches, the use of projective techniques has waned, leaving an open question and considerable controversy over the value of the projective approach.

There is also disagreement about how, or how much, to classify a test as projective. Arguably, the leading modern-day researcher on the Rorschach (Exner, 2003) has argued that only a minority of the variables coded on the test involve the process of projection. More recently, some have argued (Meyer and Kurtz, 2006) that classifying tests as either projective or objective (objective is usually defined as the tests in which there is a limited set of answers to questions and the answers are scored to an objective preexisting key) tends to give rise to misleading connotations. In particular, it may be assumed that the objective tests are more reliable and valid; and projective tests are more subjective, less reliable, and even less valid. Free response and performance-based tests have been suggested as alternative terms to refer to projective tests as these terms do not appear to imply problematic psychometric qualities.

Psychometric Considerations

A psychological test or assessment device is subject to evaluation on standard dimensions and, regardless of how compelling

a rationale may be, must demonstrate that it meets the usual psychometric criteria before it can be considered to be useful. Foremost among these criteria are reliability and validity.

Reliability

Reliability refers to the extent to which a test provides a consistent measure of what it is measuring. If a test lacks reliability, it is akin to using a rubber ruler, providing different measurements each time it is applied, and there is little credence that can be given to any of those measurements. The first approach to the measurement of reliability is internal consistency, which assesses the extent to which every item in a test is measuring the same characteristic. The second approach is through test–retest, the idea that if the construct that is being assessed is stable then a test should come up with a similar score or measure when the test is given on two separate occasions. The third approach is whether there is reliability between different raters of a test. If different individuals are looking at the same sample of behavior that is measured on the test (i.e., an answer to a question, a drawing, or a response to an inkblot), they should be able to code that behavior in a similar fashion.

Although the most projective techniques employ many items, the items often are designed to be responsive to different dimensions, and so, a projective test may often be seen as a collection of disparate single items rather than a collection of redundant items. Unfortunately, in measurement, redundancy is often desirable, and it is well demonstrated that single items are subject to far more unreliability than collections of items, in which errors may cancel each other out. If there are indeed collections of single items, it can be argued that the test really is a collection of tests, each of which measures a single dimension (and therefore requires a separate assessment of reliability and validity). In such a situation, it can be argued that it serves to reduce the reliability of each. Proponents of projective tests would counter that the object being measured is personality, which contains several dimensions, and that the items should converge to produce a clear picture of the multidimensionality of the respondent. On either side of the argument, item consistency is typically not a measure of reliability evaluated for projective tests. More typically, researchers and test developers examine the reliability in terms of the stability of the scores (test–retest) and consistency of coding (interrater reliability). To the degree these measures of reliability are wanting, the validity of the test is negatively impacted.

Validity

Reliability is necessary but not sufficient for a test. An unreliable test cannot be valid, but a reliable test may not necessarily be valid. Validity is an assessment of the extent to which a test measures what it is intended to measure. This leads to a very interesting problem in assessing the validity of projective techniques. Most projective techniques purport to measure many different aspects of the person, and it is necessary to indicate the validity of every inference that a test draws. Thus, rather than asking, for example, whether the Rorschach is valid, it is necessary to ask, for each inference drawn from the Rorschach, whether or not that inference is valid. One statistical estimate cannot represent the validity of the test, and the general question cannot be answered (Stricker and Gold, 1999).

It is not unusual for a research project to compare the validity of projective techniques with more objective, self-report, and measures of personality (e.g., McClelland et al., 1989). The usual finding of such studies is that the two approaches do not correlate with each other. Some people have interpreted this as indicating a failure in the validity of the projective approach, a conclusion that assumes the validity of objective tests (Lilienfeld et al., 2000). Others (e.g., Hiller et al., 1999) have determined that this lack of correspondence indicates that each approach is valid, but that they are measuring different aspects of the person. This view suggests that objective tests are superior at measuring objective criteria, such as behavior, whereas projective approaches are superior at measuring more covert, unconscious aspects of the person (Hiller et al., 1999). This interpretation is consistent with the theoretical position that usually underlies each of the approaches to measurement.

It is also reasonable to ask whether the incorporation of a projective approach to a test battery adds anything to the process of assessment (incremental validity). If the conclusion about measuring different aspects of the person is accepted, then projective approaches do contribute incremental validity insofar as the assessor is interested in less overt aspects of the respondent. Given the length of time necessary for the administration and interpretation of projective techniques, the question of cost-effectiveness also can be raised. The answer to this question probably is dependent on the individual circumstance. It should be clear, however, that validity, incremental validity, and cost-effectiveness are separate questions, and the answer to one does not determine the answer to the others. In the review of the most commonly used projective tests, along with a description of the tests and their administration, both their reliability and validity will be discussed.

Most Frequently Used Projective Tests

Rorschach

The Rorschach test or, as it is sometimes referred to by the public, the inkblot test was published in 1921 by Hermann Rorschach (1921/1942). The use of inkblots was not a new concept in the 1920s. In fact, a popular game among children at the time was Blotto, in which children made or bought premade inkblots and gave interpretations to the blots. Rorschach's clinical interest in the use of blots was sparked in 1911, while working with Eugen Bleuler, his professor and supervisor. Bleuler coined the term schizophrenia and investigated ways to differentiate schizophrenia from other forms of psychosis. Rorschach happened to observe in his work with patients suffering from schizophrenia that they tended to respond differently in Blotto than patients with other forms of psychosis or dementia. He decided to study this phenomenon more systematically and by 1921 had administered the test often enough to believe in its usefulness as a diagnostic aid in identifying schizophrenia and published his findings in a monograph (Rorschach, 1921/1942). In the monograph, in addition to the results concerning schizophrenia, he noted

that certain kinds of responses were consistent with certain personality traits or styles. The monograph included a set of 10 inkblots and described the administration and the initial coding system. Rorschach died in 1922, but his successors saw the potential of his tool as an idiographic, projective measure and worked over the next two decades to develop the scoring system for responses to the inkblots. Although other inkblot sets have been developed (e.g., Holtzman Inkblot Test; Holtzman et al., 1961), Rorschach's set of inkblots are the most widely used and studied.

Samuel Beck was a psychology graduate student who attempted, shortly after Rorschach's death, to do a standardization study on the Rorschach as a dissertation topic. He began his work at the Institute for Child Guidance in New York and later took positions at both Boston Psychopathic Hospital and Harvard Medical School. Over the years, he gained extensive experience with the use of the inkblots and became acutely aware of the need for more intensive work toward standardization (Beck, 1950).

Around the same time, Bruno Klopfer came to Zürich to work as a technician at the Zürich Psychotechnic Institute where Rorschach inkblots were routinely used in the screening of candidates for employment positions. After working there for several years, he immigrated to the United States in 1934 and began working at Columbia University. The psychology graduate students at Columbia, who were frustrated by the lack of formal training available in the use of inkblots, were eager to have him share his experience with them. Klopfer began giving informal seminars to students and, in his work with these students, perceived the need to develop new scoring codes beyond those articulated by Rorschach. Thus, by 1936, Klopfer was working hard to expand the existing coding system (Klopfer et al., 1954).

At the same time, Beck was busy writing papers that criticized the development of new scoring systems for their lack of empirical soundness. He was specifically critical of Klopfer's attempts to expand beyond Rorschach's published system. Over the next 20 years, the arguments continued and both Beck and Klopfer developed different scoring systems. Despite this controversy over scoring, the method of using the Rorschach continued to burgeon. By 1957, five very different scoring systems had emerged.

John Exner conducted a comprehensive review of the five scoring systems in what was supposed to be a journal article, but after several years ended up as a book (Exner, 1969). He found that the administration rules varied across the systems, each collected data differently, and the coding varied across the different systems. Even when the codes included Rorschach's original categories, they were defined differently so that the codes were not equivalent. In the end, he concluded that there were five uniquely different Rorschach tests that used a common set of inkblots. In 1968, he founded the Rorschach Research Foundation to answer which system had the best empirical foundation and which system had the greatest clinical utility. In surveys of clinicians between 1968 and 1972, he found that most clinicians were scoring the tests in a haphazard manner. Many did not score it at all, and those who did tended to use features of different scoring systems. Only a small minority used one system to score an entire test. Exner concluded that he needed to develop a system that integrated features of the available scoring systems that had a strong empirical foundation. In addition to a reliable system with good validity, Exner wanted a system that was not wedded to any single personality theory. Based on a large pool of Rorschach protocols, Exner articulated a standard administration and scoring of the test. He drew on the five existing systems to develop what he labeled as the Comprehensive System. Variables were only included if they could be reliably scored, and the system continued to be reviewed and revised from its first appearance in 1974 to the latest edition (Exner, 2003). Although at one time, many clinicians treated the Rorschach purely as a projective device and simply interpreted the content of the responses according to their understanding of personality with the publication of the Comprehensive System and a sizable body of research supporting its use, the use of the Comprehensive System has become widespread and the accepted standard for interpretation.

Starting in the mid-1990s and continuing for several years (Grove et al., 2002; Wood et al., 1996), a number of criticisms were leveled against the Rorschach and the Comprehensive System. Among the criticisms were questions raised about whether the Comprehensive System's norms overpathologized individuals as well as about the reliability and validity of a number of variables in the Comprehensive System. These criticisms fueled a debate in both, the professional and public literature, about the use of the Rorschach. Many of the criticisms were addressed (e.g., Meyer and Archer, 2001; Society for Personality Assessment, 2005); however, there still remained concerns about the adequacy of the Comprehensive Systems norms (Shaffer et al., 2007) and questions about the validity of some of the variables (Mihura et al., 2012). These concerns, combined with Exner's death in 2006 and the resulting inability to use the Comprehensive System title, led to the development of the Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011). The R-PAS is based on the Comprehensive System and four of its five authors were part of Exner's research group. The R-PAS slightly modified administration to optimize the number of responses, eliminated some variables that either had poor reliability or questionable validity, based interpretations on norms developed from 13 countries, and plotted variables using standard scoring (i.e., t-scaling) to aid interpretation. It holds promise to further solidify the Rorschach's reliability and validity. Initial results indicate good to excellent interrater reliability with the scoring of the R-PAS (Viglione et al., 2012) and good support for some of the composite variables included in the R-PAS (see Meyer and Eblin, 2012, for an overview of the changes and the data supporting the R-PAS variables).

Thematic Apperception Test

In 1935, Christiana Morgan and Henry Murray introduced the Thematic Apperception Test (Murray, 1943), a series of provocative pictures used to explore the respondent's dynamics. The stimuli are more highly structured than inkblots and respondents are required to give more meaningfully organized verbal responses. Respondents are asked to describe the scene in the picture, the apparent emotional content and thoughts of the people involved, what

led up to the scene, and what the outcome will be. There are 31 cards, with some cards designed to be given to all respondents and others to be used with just males or females, or just with adults or children. The original instructions were to have respondents provide stories to a set of 10 cards, and then a week later respond to a second, different set of 10 cards. In actual practice, clinicians vary on which cards and how many are administered, and rarely is the test administered in a second session. This test was developed on the premise, shared by other projective techniques, that people reveal something of their own personality (i.e., their motivational and emotional condition), as well as their cognitive structure, when ascribing thoughts and feelings to ambiguously drawn characters. Other sets of pictures have been developed as stimuli for telling stories, including a set that has animals engaged in human activity for use with children (Children's Apperception Test; Bellak and Abrams, 1997), and a set of photographs with adolescents confronting developmental challenges (Tasks of Emotional Development; Cohen and Weil, 1975). In response to concerns that the TAT oversamples certain issues (e.g., dysphoria) and the lack of ethnic and racial diversity in the drawings' characters, the tell-me-a-story (TEMAS) (Malgady et al., 1984) was developed. However, in spite of the many sets of pictures that have been developed for using the apperceptive approach, the TAT remains the most widely used.

A number of different scoring systems have been developed, but perhaps the most popular scoring method for this test was devised by David McClelland (McClelland et al., 1953). McClelland's work was devoted to measuring achievement motivation and he scored the TAT responses across seven dimensions: (1) stated need for achievement (someone in the story expresses a desire to achieve a goal); (2) instrumental activity (whether the outcome of the instrumental activity is successful, doubtful, or unsuccessful); (3) anticipatory goal status (whether the person in the story anticipates achieving the goal or being frustrated by it); (4) obstacles or blocks (whether the goal in the story is blocked by an obstacle); (5) nurturant press (whether there are nurturing forces aiding the characters in goal achievement); (6) affective states (emotions present associated with achieving or being frustrated by the goal); and (7) achievement themes (when achievement imagery is the focus of the story). Using McClelland's scoring criteria, researchers have been able to achieve a high level of interrater reliability and have been able to show significant correlations between the test scores as predicted indicators of achievement (Vane, 1981). However, the validity has been found to be affected by story length (i.e., poorer validity with individuals giving brief, unelaborated stories). Other scoring systems measuring such personality characteristics and coping methods have shown similar results with very good reliability and results supportive of the test's validity (Hibbard and Porcerelli, 1998; Westen et al., 1990; also see Cramer, 1999, for an overview of other scoring systems and their reliability and validity). However, most clinicians do not use formal scoring systems to interpret the TAT but instead simply read the responses, look for common themes, treat the stories as psychologically meaningful communications, and consider them as projective tools for understanding the patient's underlying conflicts and needs. The use of this intuitive method, along with the variability across clinicians in the cards administered, likely leaves the test's reliability and validity in actual practice much lower than what has been achieved in the published research.

Projective Drawings

In 1949, Karen Machover described the figure drawings by respondents that clinicians could use to identify personality dynamics. In her DAP test, respondents are asked to draw a full person (gender not specified ahead of time), and then to draw a person of the opposite sex of the first drawing. They are then asked to make up stories for each drawing. The absolute and relative size of the persons are examined as well as details such as the position on the page, the quality of the lines, the sequence of parts drawn, the stance, position of the arms, and grounding effects. A similar constructive projective test, the House–Tree–Person test, was developed at the same time by John Buck (1948). There have been a number of variations to these original constructive tests (e.g., Draw-A-Family, Kinetic Family Drawing, Draw-A-Person in the Rain).

The manuals describing the administration, scoring, and interpretation of these tests varies from fairly minimal with little psychometric information (e.g., Buck, 1948) to fairly comprehensive with detailed information on scoring and reliability (e.g., Draw a Person: Screening Procedure for Emotional Disturbance; DAP:SPED; Naglieri et al., 1991). There are three different methods that clinicians have used to interpret the drawings. The first is the use by clinicians of their global impressions and associations to the drawings to develop hypotheses about the artist's self-perception, motor expression, anxieties, or needs. The second interpretation method is to link specific characteristics of the drawing with particular aspects of the artist's personality or diagnosis (e.g., small figures are associated with low self-esteem). A number of authors (e.g., Buck, 1992; Burns, 1987) have produced glossaries that provide interpretations for these specific signs. The third method does not link specific signs with a specific interpretation but rather focuses on the frequency of indicators of pathology in the renderings. This method originated with Elizabeth Koppitz (1968), and the DAP:SPED is a commercially produced example of this method. The interrater reliability of the global impressions method has generally been low and the ability to identify pathology poor (Kahill, 1984). This has been explained by one of the proponents (Hammer, 1997) of this method as due to the interpretation being more "an Art than a Science…half of the psychologists are adept at this art, and about half are not" (p. 377). The interrater reliability of the specific signs approach has been variable depending on the level of training of the raters (Kahill, 1984), but the evidence to support the validity of the interpretations has generally been poor (Kahill, 1984; Swenson, 1968). Both the test–retest and interrater reliability are good for the overall ratings method such as found in the DAP:SPED (Naglieri et al., 1991), and such methods have been able to successfully differentiate emotionally disturbed children from nondisturbed children (Matto, 2002; McNeish and Naglieri, 1993; Naglieri and Pfeiffer, 1992). However, it is not clear that most clinicians use the overall rating method for interpretation, and even with the

use of that method there are concerns about the strength of the conclusions and their value over other tests (Lally, 2001).

Sentence Completion Tests

Another type of projective test is sentence completion tests where respondents are asked, orally or in writing, to complete incomplete sentence stems (e.g., What worries me…; Women…; My mother…). Their origin as a projective test is derived from Carl Jung's use of the word association test, where individuals were given a word and instructed to reply with the first word that they can think of. These word associations were supposed to provide insight into individuals' unconscious ideas and motivations (Jung, 1910). The use of sentence stems was seen as an improvement over just the single word, and starting in the 1930s there have been a number of standardized written sentence completion tests developed. The most widely used (Holaday et al., 2000) of these standardized tests is the Rotter incomplete sentence blank (RISB; Rotter and Rafferty, 1950; Rotter et al., 1992). It has 40 sentence stems that can be self- or orally administered to high school students and adults, and is designed to measure both adjustment and maladjustment. The manual (Rotter et al., 1992) provides an objective scoring system and data indicating acceptable short-term test–retest reliability (but poorer long-term – 6 months to 3 years – test–retest reliability) and good interrater reliability. Using the formal scoring system, the RISB has shown the ability to differentiate maladjusted versus well-adjusted individuals. However, as with the TAT and projective drawings, most examiners (Holaday et al., 2000) appear to use their intuitive clinical judgment to interpret the test rather than a formal scoring system. The reliability and validity of such an approach is unknown. As with the TAT and other apperceptive measures, the value of the task often is dependent on the particular set of items that is employed, and the more closely the items approximate the target being studied, the more valuable the responses can be shown to be (Moretti and Rossini, 2003).

Concluding Comments

It can be concluded, with some certainty, that the original promise of the projective hypothesis (Frank, 1939) has not been sustained, and that there are many limitations on the conclusions that can be drawn from the responses to ambiguous stimuli, especially when psychologists' interpretations are guided by just their associations. However, there is evidence that such stimuli can produce valid inferences, and inferences that cannot be drawn from other approaches to measurement. Traditionally, the inferences drawn from projective tests have been seen as dependent on the theoretical approach and goals of the examiner, whose skillfulness and clinical acumen were important variables in considering the utility of the conclusions drawn from projective methods (e.g., Hammer, 1997). Although such an approach may be valid for particular individuals, it does lend itself to ready verification for the individual and does not allow the test to demonstrate psychometric capabilities. However, projective tests are not limited to solely being interpreted by clinicians' intuition. Because projective techniques do show the promise of providing valid information that is different than that produced by objective measures, it is worth pursuing methods of interpreting these tests that have good reliability and validity so that clinicians are able to gain a more complete understanding of the individuals they assess.

See also: Classical (Psychometric) Test Theory; Equivalence and Transfer Problems in Cross-Cultural Research; Neuropsychological Testing; Reliability and Validity Assessment: New Approaches; Reliability: Measurement.

Bibliography

Beck, S.J., 1950. Rorschach's Test. I. Basic Processes, second ed. Grune & Stratton, New York City, NY.
Buck, J.N., 1948. The H-T-P. Journal of Clinical Psychology 4, 151–159.
Buck, J.N., 1992. House-Tree-Person Projective Drawing Technique: Manual and Interpretive Guide. rev. ed.; revised by Warren, W.L. Western Psychological Services, Los Angeles, CA.
Burns, R.C., 1987. Kinetic-House-Tree-Person Drawings (K-H-T-P). Brunner/Mazel, New York City, NY.
Bellak, L., Abrams, D.M., 1997. The T.A.T., the C.A.T., and the S.A.T. in Clinical Use, sixth ed. Allyn & Bacon, Needham Heights, MA.
Cohen, H., Weil, G.R., 1975. Tasks of Emotional Development: Manual. T.E.D. Associates, Brookline, MA.
Cramer, P., 1999. Future directions for the Thematic Apperception Test. Journal of Personality; Assessment 72, 74–92.
Exner, J., 1969. The Rorschach Systems. Grune & Stratton, New York City, NY.
Exner, J., 2003. The Rorschach: A Comprehensive System, fourth ed., vol. 1. Wiley, New York City, NY.
Frank, L.K., 1939. Projective methods for the study of personality. Journal of Psychology 8, 389–413.
Grove, W.M., Barden, C., Garb, H.N., Lilienfeld, S.O., 2002. Failure of Rorschach-Comprehensive-System-based testimony to be admissible under the Daubert-Joiner-Kumho standard. Psychology, Public Policy, and Law 8, 216–234.
Hammer, E.F., 1997. Editor's comment. In: Hammer, E.F. (Ed.), Advances in Projective Drawing Interpretation. Thomas, Springfield, IL, p. 377.
Hibbard, S., Porcerelli, J., 1998. Further validation for the Cramer Defense Mechanism Manual. Journal of Personality Assessment 70, 460–483.
Hiller, J.B., Rosenthal, R., Bornstein, R.F., Berry, D.T.R., Brunell-Neuleib, S., 1999. A comparative meta-analysis of Rorschach and MMPI validity. Psychological Assessment 11, 278–296.
Holaday, M., Smith, D.A., Sherry, A., 2000. Sentence completion tests: a review of the literature and results of a survey of members of the Society for Personality Assessment. Journal for Personality Assessment 74, 371–383.
Holtzman, W.H., Thorpe, J.S., Swartz, J.D., Herron, E.W., 1961. Inkblot Perception and Personality: Holtzman Inkblot Technique. University of Texas Press, Austin, TX.
Jung, C.G., 1910. The association method. American Journal of Psychology 31, 219–269.
Kahill, S., 1984. Human figure drawings in adults: an update of the empirical evidence, 1967–1982. Canadian Psychology 25, 269–292.
Klopfer, B., Ainsworth, M.D., Klopfer, W.G., Holt, R.R., 1954. Developments in the Rorschach Technique: Technique and Theory, vol. 1. World, New York City, NY.
Koppitz, E., 1968. Psychological Evaluation of Children's Human Figure Drawings. Grune & Stratton, New York City, NY.
Lally, S.J., 2001. Should human figure drawings be admitted into court? Journal of Personality Assessment 76, 135–149.
Lilienfeld, S.O., Wood, J.M., Garb, H.N., 2000. The scientific status of projective techniques. Psychological Science in the Public Interest 1, 27–66.
Machover, K., 1949. Personality Projection in the Drawing of the Human Figure. Thomas, Springfield, IL.
Malgady, R.G., Costantino, G., Rogler, L.H., 1984. Development of a Thematic Apperception Test (TEMAS) for urban Hispanic children. Journal of Consulting and Clinical Psychology 52, 986–996.
Matto, H.C., 2002. Investigating the validity of the Draw A Person: Screening Procedure for Emotional Disturbance: a measurement validity study with high-risk youth. Psychological Assessment 14, 221–225.

McClelland, D.C., Atkinson, J., Clark, R., Lowell, E., 1953. The Achievement Motive. Appleton-Century-Crofts, New York City, NY.

McClelland, D.C., Koestner, R., Weinberger, J., 1989. How do self-attributed and implicit motives differ? Psychological Review 96, 690–702.

McNeish, T.J., Naglieri, J.A., 1993. Identification of the seriously-emotionally disturbed with the Draw A Person: Screening Procedure for Emotional Disturbance. Journal of Special Education 27, 115–121.

Meyer, G.J., Archer, R., 2001. The hard science of Rorschach research: what do we know and where do we go? Psychological Assessment 78, 219–274.

Meyer, G.J., Eblin, J.J., 2012. An overview of the Rorschach Performance Assessment System (R-PAS). Psychological Injury and Law 5, 107–121.

Meyer, G.J., Kurtz, J.E., 2006. Advancing personality assessment terminology: time to retire "objective" and "projective" as personality test descriptors. Journal of Personality Assessment 87, 223–225.

Meyer, G.J., Viglione, D.J., Mihura, J.L., Erard, R.E., Erdberg, P., 2011. Rorschach Performance Assessment System: Administration, Coding, Interpretation, and Technical Manual. Rorschach Performance Assessment Systems, LLC, Toledo, OH.

Mihura, J.L., Meyer, G.J., Dumitrascu, N., Bombel, G., 2012. The validity of individual Rorschach variables: systematic reviews and meta-analyses of the comprehensive system. Psychological Bulletin 139, 548–605.

Moretti, R.J., Rossini, E.D., 2003. The Thematic Apperception Test (TAT). In: Hilsenroth, M., Segal, D. (Eds.), Personality Assessment. In: Hersen, M. (Editor-in-Chief), Comprehensive Handbook of Psychological Assessment, vol. 2. John Wiley & Sons, Hoboken, NJ, pp. 356–371.

Morgan, C.D., Murray, H.A., 1935. A method for investigating fantasies: the Thematic Apperception Test. Archives of Neurology and Psychiatry 34, 289–306.

Murray, H.A., 1943. Thematic Apperception Test Manual. Harvard University Press, Cambridge, MA.

Naglieri, J.A., McNeish, T.J., Bardos, A.N., 1991. Draw A Person: Screening Procedure for Emotional Disturbance Examiner's Manual. PRO-ED, Austin, TX.

Naglieri, J.A., Pfeiffer, S.I., 1992. Validity of the Draw A Person: Screening Procedure for Emotional Disturbance with socially-emotionally disturbed sample. Psychological Assessment: A Journal of Consulting and Clinical Psychology 4, 156–159.

Rorschach, H., 1921/1942. Psychodiagnostics: A Diagnostic Test Based on Perception, fourth ed. Grune & Stratton, New York.

Rotter, J.B., Lah, M.I., Rafferty, J.E., 1992. Rotter Incomplete Sentences Blank. Harcourt Brace, San Antonio, TX.

Rotter, J.B., Rafferty, J.E., 1950. Manual: The Rotter Incomplete Sentences Blank. Psychological Corporation, New York City, NY.

Shaffer, T.W., Erdberg, P., Meyer, G.J., 2007. Introduction to the JPA special supplement; on international reference samples for the Rorschach Comprehensive System. Journal of Personality Assessment 89 (S1), S2–S6.

Society for Personality Assessment, 2005. The status of the Rorschach in clinical and forensic practice: an official statement by the Board of Trustees of the Society for Personality Assessment. Journal of Personality Assessment 85, 219–237.

Stricker, G., Gold, J.R., 1999. The Rorschach: toward a nomothetically based, idiographically applicable configurational model. Psychological Assessment 11, 240–250.

Swenson, C.H., 1968. Empirical evaluations of human figure drawings. Psychological Bulletin 70, 20–44.

Vane, J.R., 1981. The Thematic Apperception Test: a review. Clinical Psychology Review 1, 319–336.

Viglione, D.J., Blume-Marcovici, A.C., Miller, H.L., Giromini, L., Meyer, G., 2012. An inter-rater reliability study for the Rorschach Performance Assessment System. Journal of Personality Assessment 94 (6), 607–612.

Westen, D., Ludolph, P., Lerner, H., Ruffins, S., Wiss, C., 1990. Object relations in borderline adolescents. Journal of the American Academy of Child and Adolescent Psychiatry 29, 338–348.

Wood, J.M., Nezworski, M.T., Stejskal, W.J., 1996. The Comprehensive System of the Rorschach: a critical examination. Psychological Science 7, 3–10.

Promotions

Michael Bognanno and Moritz Ritter, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the existing empirical findings in the economics literature on promotions in hierarchies. First, the basic facts concerning promotion are listed. Second, the promotion outcomes, focusing on wages, health and happiness, and promotion determinants are discussed. Last, the prominent theories of promotions and the relevant empirical evidence are summarized.

The typical organizational structure of the firm is hierarchical with a diminishing number of positions at higher levels. Most positions in large organizations are filled through promotion rather than hiring from outside the firm. Although outside hiring is conducted at all levels of the hierarchy in large US corporations, over 75% of the executives in the top four reporting levels were promoted into their positions and spent the majority of their working lives within their current firm (Bognanno, 2001). The pay structure in firm hierarchies is convex, exhibiting increasing interrank pay differences. An employee's advancement within the firm is referred to as a promotion. The predominant use of internal promotion to fill vacant positions in the hierarchy and the convexity of the pay structure serve to create strong promotion incentives for workers within the firm. This article first summarizes the empirical findings on promotion outcomes and determinants, focusing on wages, health and happiness and then presents the prominent theories of promotions and discusses the relevant empirical evidence. (For a general review of the personnel economics literature see Lazear and Oyer (2012). For a comprehensive review of the theory and evidence in internal labor markets see Waldman (2012).)

The notion of employee advancement is broad enough for promotion to be defined in different ways. Advancement in responsibilities, reporting level, pay grade or job title all constitute a promotion in some sense. However, the magnitude of the pay consequences and the frequency of the event will differ with the definition. Large firms tend to have more job titles than pay grades and more pay grades than reporting levels. It is possible for one of these aspects to change without a corresponding change in the others. While promotions have been defined in different ways in the literature, a widely used definition is a job change accompanied by upward movement in the hierarchy or in pay.

Annually in US corporate hierarchies, with promotions defined as a job change accompanied by a higher pay grade, roughly 90% of upper level executives are stationary, 9% are promoted or move laterally, and less than 1% suffer demotion. At higher levels in the firm hierarchy, the rate of promotion decreases, while the rewards to promotion increase. The rate of growth in pay also increases at higher levels. Not surprisingly, changes in pay are most favorable for the promoted and least favorable for the demoted. The full financial rewards of a promotion are not realized immediately; in the years subsequent to a promotion, pay increases remain elevated but dissipate with time. (Belzil and Bognanno (2008) find the rate of pay growth to be higher at more senior levels in the firm. The annual percentages of executives promoted, remaining in position and demoted are taken from this article. Their data support the findings summarized in the paragraph. Evidence of higher rates of promotion in lower hierarchical levels was noted in Baker et al. (1994b) and Belzil and Bognanno (2008).)

Promotion Outcomes

Promotion has been established as important to wage growth for employees in organizations. Through the analysis of personnel records from individual firms, both Lazear (1992) and Baker, Gibbs and Holmstrom (henceforth BGH (1994a,b)) demonstrated the importance of promotion on within-firm wage growth. Lazear (1992) used 13 years of personal records to study the influence of job assignment on wages and turnover for full-time workers at a large durable-goods manufacturer. Job change was linked to wage growth. Workers remaining more than 7 years in the same job suffered declining real wages. Most of the variation in pay was found to exist between jobs and not within jobs. BGH (1994a,b) used 20 years of personnel records in a detailed examination of the internal labor market of a large firm. Promotion was necessary to wage growth in the firm, as real pay fell with increasing position tenure. Hierarchical level in the firm alone explained about 70% of the variance in pay (BGH, 1994b). Controlling for age and tenure, Bognanno (2001) finds that increasing time in a position reduces the pay of all executives except Chief Executive Officers (CEOs) in an analysis of an 8-year panel consisting of hundreds of the largest US firms. Age, education, and level in the firm explained 43% of the variation in pay. With the addition of controls for the worker's firm tenure, job tenure, whether the executive was externally hired into the position and firm fixed effects, 69% of the variation in pay was explained.

Interestingly, research focusing on the nonmonetary aspects of promotion indicates that promotion may be a mixed blessing. Higher ranking positions are associated with higher levels of responsibility and stress. Consequently, the effect of a promotion on a worker's overall wellbeing might be negative. Unfortunately, the overall effect of a promotion is not easy to assess in a simple cross-section of workers. Workers in lower ranked positions are potentially different from workers in higher positions in ways other than level. The innate

characteristics that contribute to high performance and promotion may in themselves be related to stress levels, independent of the level of the job.

A recent paper (Johnston and Lee, 2013) addresses this problem by employing the Household, Income and Labor Dynamics in Australia survey to assess how a promotion affects a worker's overall wellbeing. On the positive side, they find that a promotion produces increased job satisfaction in the short run (approximately 3 years) and increased income permanently. However, while happiness and physical health outcomes remain unchanged, there is a lasting negative effect on mental health and a lasting increase in hours of work. Workers report a significant deterioration in their mental health due to job stress approximately 2 years after a promotion. The authors conclude that promotions do not increase workers' wellbeing and potentially have a negative mental health effect.

Similar findings are reported by Boyce and Oswald (2012) using the British Panel Household Study. They report that promotion leads to a significant deterioration in the psychological wellbeing of the promoted workers. Kosteas (2011) also reports a short term positive effect of promotions on job satisfaction after controlling for wage increases in the NLSY79. Additionally, workers who believed promotion was possible in the near future had higher job satisfaction and lower quit rates, supporting the idea of a beneficial 'anticipation effect.'

Promotion Correlates and Determinants

It is challenging to accurately predict promotions from year to year when the large majority of executives are stationary. However, promoted executives differ on average in some respects from those not promoted and there are factors that make promotion more likely. In terms of differences between those advancing and those not, promoted executives tend to earn more than others in their reporting level prior to promotion and the percentage of promoted executives with above average earnings prior to promotion increases with level (Bognanno, 2001). Additionally, promoted executives tend to receive large pay raises prior to promotion (BGH, 1994a; Belzil et al., 2012; Gibbs, 1995; Dohmen, 2004). In the year prior to a change in the CEO, the highest paid non-CEO in the firm is selected as the new CEO about 80% of the time (Bognanno, 2001). Belzil, Bognanno and Poinas (henceforth BBP, 2012) found that promoted executives tend to be younger, more educated, and have less firm tenure than others in their job who were not promoted. BGH (1994a,b) also find education positively correlated with promotion probabilities. A positive correlation between the employee's performance ratings in the firm and the probability of promotion has also been found in the literature. (See Medoff and Abraham (1980), Dohmen (2004), and DeVaro and Waldman (2012).)

Several studies have reported evidence consistent with the notion of a 'fast track.' (See Gibbs and Hendricks (2004) and references therein.) The idea is that some executives exhibit a faster rate of promotion and appear to be on a 'fast track.' The finding of a positive correlation between an executive's promotion in 1 year and promotion in the next, consistent with 'fast tracks,' appears in BBP (2012). Promoted executives in their data have about a 75% greater likelihood of promotion in the year following a promotion than nonmovers and about a 70% greater likelihood 2 years later. Additionally, prior to promotion, the soon-to-be promoted executives were younger and had below average position tenure in comparison to nonmovers in the same job and firm.

A careful model of the determinants of promotion cannot use pre-promotion relative wage position or wage growth as exogenous variables because the process of wage determination and promotion is arguably closely linked. BBP (2012) consider a model of promotion probability that conditions on unobservable individual heterogeneity, allows for an endogenous initial condition, and considers the explanatory power of exogenous individual and firm characteristics. They find that the most influential determinant of promotion is unobserved individual heterogeneity. Persistent unobserved individual heterogeneity explains 70% of the variation in promotion probabilities. This factor could represent executive ability, of which there was no measure in the data.

Promotion determinants of lesser importance in BBP (2012) are past career advancement speed, human capital (age and education), firm variables (profits, sales and size), tenure and newcomer status at the firm, and reporting level in the firm. Education and firm tenure have a weak positive influence on promotion probabilities. Age, newcomer status and level have a negative effect. Age, education, tenure, newcomer status in the firm and unemployment upon hire combined account for only 8.3% of the variation in promotion probabilities. Firm variables are even less influential. Sales, profits and employment levels accounted for only 3% of the variation in promotion probabilities. The contemporaneous national unemployment rate had a very slight dampening effect on promotion, while the unemployment rate upon hire had no effect.

BBP (2012) found promotion probabilities to differ by functional area. Marketing services, management areas, and manufacturing appear to have an advantage in promotion. The three weakest functional areas for promotion were employee relations, legal and management information systems. The functional areas account for 5.4% of the variation in promotion probabilities.

Theories and Empirical Support

The Tournament Theory of Promotions

The application of tournament theory to promotion in hierarchies is appealing in part because the theory provides an explanation for the increasing interrank pay spreads observed in organizations. As such, it provides a rationale for high levels of CEO pay, a topic that receives considerable attention. It is also consistent with the high rate of internal promotions in organizations to fill position vacancies and with the diminishing number of positions at higher ranks in the organization.

The tournament model (Lazear and Rosen, 1981; Rosen, 1986) views promotions as a 'prize' for which workers compete – promotions serve as an instrument to align the incentives of the workers with the interests of the employer. In the model, managers are in a competition to climb a hierarchical ladder; the pay increase associated with

climbing to the next rung, combined with uncertainty over who will be promoted, induces them to provide effort.

Consider a firm with many workers. Worker output is a function of effort and luck. Output is observable to the owner but effort is not. To induce the workers to exert effort, the owner promises a prize in the form of a bonus (promotion) to the workers who produce the most output. At the end of the period, the best or a top share of workers are promoted to the next level and receive the accompanying pay increase. Providing an incentive in the form of such a rank-order tournament, in which the workers are rewarded based on their relative performance, has three distinct advantages over piece-rate contracts (payments based on the amount of output produced). First, it might be easier to evaluate the relative performance of workers than measure their absolute performance in certain settings. Second, in the tournament model, the owner has no incentive to be dishonest about the level of output workers produced; the owner is committed to promoting the best performing workers, independent of their level of output. Third, if worker output is subject to a common random shock that impacts the output of all workers, paying them based on absolute output (a piece-rate) exposes workers to risk to which they may be averse. The rank order comparisons used to allocate prizes in tournaments filter out this common shock.

There are tradeoffs faced when operating a promotion tournament if the objective includes both selecting the best person for a position and maximizing the effort exerted when workers are of unequal ability. The greater the difference in ability, the less likely it is that the outcome of the competition will be determined by the level of worker effort and the weaker the promotion incentives become. The firm can counteract this by handicapping the participants to create an even competition and improve the incentives to exert effort but doing so reduces the likelihood that the more able participant will win. The firm cannot both select the most able worker and maximize incentives.

Another aspect of the tournament model is that a worker's current pay is not directly tied to current productivity and does not serve as an incentive to the worker receiving the pay. However, it does serve as an incentive to the workers below who are motivated by the prospect of potentially receiving this pay upon promotions in the future. This incentive property of the hierarchical pay structure also implies that the pay raises upon promotion must be large enough to induce the agents to exert effort, i.e., there must be discrete jumps between pay grades. Moreover, the size of the pay increase is inversely related to the probability of a promotion. It is more difficult to be promoted to the level of a CEO than to be promoted from shift worker to supervisor of a small group of workers, so the promotion to CEO should carry a much larger pay increase. As well, the pay increase upon promotion to CEO is larger because that is that last promotion one can receive. Promotion at lower levels carries the immediate reward and the option value of remaining in the tournament and in contention for subsequent promotions.

The widespread use of tournaments is noted in Baker et al. (1988) in a study summarizing the empirical evidence on the firm's internal incentive structure. It reports the "overwhelming use of promotion-based incentive systems" rather than pay-for-performance systems, especially in settings where performance evaluation is subjective.

Moreover, there is a substantial literature testing individual predictions of the tournament theory. First, as discussed above, the pay increase associated with a promotion should be positively related to the number of workers competing for that promotion. This is supported by Leonard (1990), who finds that pay differentials are higher toward the top of the hierarchy and that pay increases are greater if the promotion rate is lower. All employing panel data including more than 200 firms, Main et al. (1993), Bognanno (2001), and Eriksson (1999) find support for the tournament theory hypothesis that the reward for promotion to CEO should increase with the number of competitors for the position.

Also supportive of tournament theory, DeVaro (2006) presents evidence from the Multi-City Study of Urban Inequality, a cross-sectional survey of 3000 establishments in four metropolitan between 1992 and 1995, suggesting that promotions are determined by workers' relative performance. He structurally estimates a model of promotion tournaments in which worker performance, the wage spread, and promotions are endogenous. The results support the tournament model in that worker effort is increasing in the wage spread and decreasing in the random component (luck) of promotion. Firms, in turn, increase the wage spread in the presence of a large random component to maintain incentives.

As the filling of vacant positions via promotion is pervasive and promotions bring financial rewards, promotion incentives are created within firms. However, there is still the question of whether firms intentionally manage them at the level of the individual. Some findings in the literature conflict with the notion that firms manage promotion incentives in accord with tournament theory. Gibbs (1995) found that those waiting longer for promotion received a smaller pay raise from promotion. Workers who go long periods without promotion suffer a decrease in the lifetime value of a promotion because they are closer to their retirement age and also have a reduced probability of being promoted at all given they have been passed over previously. Gibbs (1995) found no evidence that the firm altered short-term incentives in response to changes in these promotion incentives.

A similar result is found in Bognanno (2001), where an additional year in the same position prior to a promotion in level reduced the pay increase from the promotion. One explanation of this finding based in the actual practice of compensation administration is that executives who go longer without a promotion are more likely to have risen higher in their pay grade prior to promotion and promotion to the bottom of the next of the next pay grade constitutes a smaller increase for them. Bognanno (2001) documents that, relative to other workers in the same reporting level, promoted workers earn more than average before promotion and less than average afterward.

Asymmetric Learning and Job Assignment

In the basic tournament model described above, positions do not differ, except in pay. Promotions result purely as an incentive device for workers to provide effort. An alternative (or additional) rationale for promotions is provided by

assignment models. Firms are a collection of positions, each fulfilling different tasks and hence requiring different abilities and skills. Over time, the firm and the worker learn about the position that is best suited for the worker, so promotions lead to a better assignment of workers to positions, while the higher pay is intended to retain the worker in the firm. In this class of models, a promotion signifies an improvement in the 'match' between the worker and the job.

A common feature of assignment models is the presence of asymmetric information about the worker's ability. In Waldman (1984), the current employer learns the worker's ability after one period of employment, whereas other employers do not learn the worker's ability and only observe current and past job assignments. This asymmetry creates a strategic element in the current employer's decision of whether or not to promote a worker. A promotion of a high-ability worker improves job sorting and increases output, but also sends a signal to the job market that this worker is of high ability. In order to avoid having the worker poached after a promotion, the current employer must provide a sufficiently large pay increase; wages do not depend on the worker's productivity in the current job assignment, but rather the value of the worker as perceived by outside employers. As a result, the current employer promotes fewer than the socially optimal number of workers.

Asymmetric information about workers' ability can also explain the aforementioned fast tracks. Serial correlation in promotions can arise for two reasons. First, more able workers accumulate the human capital necessary for promotion more quickly and, hence, exhibit a faster rate of promotion at each level of the hierarchy than less able workers. Second, by promoting a worker in the past, the current employer has already signaled that the worker is of high ability. The worker is then already paid accordingly, making subsequent promotions less costly. This, in turn, implies that the employer will choose to strategically delay a promotion for other workers in order to avoid having to adjust the wage rate.

Bernhardt (1995) shows that the signaling effect of promotion has important implications for the probability of promotion for workers of different observed skill levels. As the firm has to pay a worker with high observed skill (e.g., education) more than a worker with low observed skill, the high-skill worker is more likely to be promoted, even if both workers are of equal ability. The resulting wage increase upon promotion is smaller for the more skilled worker, which implies that a worker with low observed skill must be of very high ability in order to be promoted. However, after the initial promotion of the low observed skill worker, the worker's high ability is revealed, which implies that the observed skill level will matter less for future promotions. One implication of this is that the higher a worker is in the hierarchy, the less important observed characteristics become.

One line of empirical research on the assignment model looks for evidence of a promotion signal. It is natural that promotion brings about a pay increase; however, if the pay increase is larger for promotions that reveal more positive information to the labor market about an employee, support for the notion of promotion signaling is found. More information would be revealed when those with less observed skill (e.g., education) are promoted or when younger workers with little labor market history are promoted. Findings on the potential signaling role of promotions appear in DeVaro and Waldman (2012) and Bognanno and Melero (2012).

Using the data from the large US manufacturing firm analyzed by BGH (1994a,b), DeVaro and Waldman (2012) look for evidence in support of two promotion signaling implications that they draw from Bernhardt (1995). Because it less costly to the firm to promote someone who is already thought of as highly able in the external labor market, more educated workers should have an advantage in being promoted over the less educated, performance held constant. As well, the wage increase upon promotion should be greater for those who are promoted among the less educated. They find supportive evidence for both implications. In a more recent article using data from the British Household Panel Survey for the period 1991–2005, Bognanno and Melero (2012) also find support for the implication that those with less education, who are nevertheless promoted, receive larger promotion rewards. Younger workers also receive larger pay increases upon promotion when promotion is defined as the attainment of new managerial responsibilities.

Another line of empirical research on the assignment model focuses specifically on fast tracks. The source of fast tracks has been hypothesized to result from a promotion signaling effect and also simply from the persistent benefit of an executive having higher ability. In an attempt to determine whether in-sample promotions yield an inherent benefit in subsequent promotion, BBP (2012) estimate a model of promotion that conditions on unobserved individual heterogeneity (controlling for persistent unobserved differences in ability). They find no evidence that in-sample promotions increase subsequent promotion probabilities, indicating that promotions do not yield an inherent benefit in subsequent promotion for upper-level management. Hence, the notion of fast tracks resulting from the persistent benefit of being of high ability is supported over the notion that fast tracks result from the signaling effect of promotion. Despite this finding, it may be the case that early career promotions for younger, lower ranking employees have a signaling effect that provides a benefit in subsequent promotions.

Human Capital and Symmetric Learning

An additional facet to the theoretical consideration of promotion is the role of human capital acquisition by the worker. This human capital can be specific to the firm, specific to the task the worker is performing, or general in nature. A promotion can serve as an incentive to accumulate firm-specific human capital. Prendergast (1993) illustrates how a firm can use the prospect of a promotion to a higher paid position to induce the worker to invest in firm-specific human capital. In the absence of a second position, a firm always has the option of reneging on the promise to increase pay if the worker's specific human capital is not verifiable by an outside party and, hence, not contractually enforceable. However, the presence of a higher paying second position, that also produces more output for the firm, aligns worker and firm incentives: the worker has an incentive to accumulate specific human capital in order to be promoted and the firm has an incentive to promote the worker since the new position generates a larger surplus for the firm.

Bayo-Moriones and Ortín-Ángel (2006) use data from interviews at 653 Spanish industrial plants in 1997 and find that use of internal promotion for blue-collar workers is higher at plants in which those workers acquire more specific human capital. However, Sullivan (2010) notes that, after controlling for occupation and industry tenure, tenure with one employer does not lead to significant income gains, suggesting that employer-specific human capital may play a minor role in promotions for most workers. In fact, he finds that the returns to tenure are at the industry level for managers.

A more comprehensive model is presented in Gibbons and Waldman (1999). This model integrates job assignment, symmetric learning and on-the-job human capital accumulation into a unified framework and delivers predictions which are broadly consistent with the empirical evidence in BGH (1994a,b): (1) large wage increases upon promotion, (2) serial correlation in wage increases and fast tracks, (3) frequent real wage decreases, and (4) infrequent use of demotions. These results arise because workers who are promoted quickly are those for whom the employer has revised innate ability upward the most. Higher innate ability causes productivity to rise faster with experience. Thus, an early promotion leads to a large increase in pay, results in large future pay increases, and reduces the time necessary before the next promotion. For other workers, the expected ability is revised downward and, hence, their real wage falls. The worker, however, is not demoted unless the downward revision is very large, making demotions infrequent.

Gibbons and Waldman (2006) extended their model to include schooling and task-specific human capital. This enriched model is able to account for the increased promotion probability for higher educated workers because more educated workers begin their careers with more human capital and continue to accumulate human capital more quickly because schooling is assumed to be positively correlated with the worker's ability to learn on the job. Thus, more educated workers advance more quickly through the job hierarchy. Task-specific human capital produces a so-called cohort effect, in which a cohort that is initially hired at a relatively higher average wage continues to earn a relatively higher average wage many years later. Cohort effects have been found in the empirical literature (see BGH (1994b) and Oreopoulos et al. (2012)).

The theoretical and empirical literature relating to promotion is evolving, resulting in a better understanding of the wage and promotion dynamics with in firms. At the same time that the importance of promotion has been made clear for earnings, new evidence questions whether promotions lead to an increase in the employee's long-term wellbeing.

See also: Agency Theory; Happiness and Work; Hierarchies and Markets; Job Displacement Insurance.

Bibliography

Baker, G., Gibbs, M., Holmstrom, B., 1994a. The wage policy of a firm. Quarterly Journal of Economics 109, 881–919.

Baker, G., Gibbs, M., Holmstrom, B., 1994b. The internal economics of the firm: evidence from personnel data. Quarterly Journal of Economics 109, 921–955.

Baker, G., Jensen, M.C., Murphy, K.J., 1988. Compensation and incentives: practice vs theory. Journal of Finance 43, 593–616.

Bayo-Moriones, A., Ortín-Ángel, P., 2006. Internal promotion versus external recruitment in industrial plants in Spain. Industrial and Labor Relations Review 59, 451–470.

Belzil, C., Bognanno, M., 2008. Halo effects, and the earnings dynamics of American executives. Journal of Labor Economics 26, 287–310.

Belzil, C., Bognanno, M., Poinas, F., 2012. Promotion Determinants in Corporate Hierarchies: An Examination of Fast Tracks and Functional Area. IZA Discussion Papers 6948.

Bernhardt, D., 1995. Strategic promotion and compensation. Review of Economic Studies 62, 315–339.

Bognanno, M., 2001. Corporate tournaments. Journal of Labor Economics 19, 290–315.

Bognanno, M., Melero, E., 2012. Promotion Signals, Age and Education. IZA Discussion Papers 6431.

Boyce, C.J., Oswald, A.J., 2012. Do people become healthier after being promoted? Health Economics 21, 580–596.

DeVaro, J., 2006. Internal promotion competition in firms. Rand Journal of Economics 37, 521–542.

DeVaro, J., Waldman, M., 2012. The signaling role of promotions: further theory and empirical evidence. Journal of Labor Economics 30, 91–147.

Dohmen, T.J., 2004. Performance, seniority, and wages: formal salary systems and; individual earnings profiles. Labour Economics 11, 741–763.

Eriksson, T., 1999. Executive compensation and tournament theory. Journal of Labor Economics 17, 262–280.

Gibbons, R., Waldman, M., 1999. A theory of wage and promotion dynamics inside firms. Quarterly Journal of Economics 114, 1321–1358.

Gibbons, R., Waldman, M., 2006. Enriching a theory of wage and promotion dynamics inside firms. Journal of Labor Economics 24, 59–107.

Gibbs, M., 1995. Incentive compensation in a corporate hierarchy. Journal of Accounting and Economics 19, 247–277.

Gibbs, M., Hendricks, W.E., 2004. Do formal salary systems really matter? Industrial and Labor Relations Review 58, 71–93.

Johnston, D.W., Lee, W.-S., 2013. Extra status and extra stress: are promotions good for us? Industrial and Labor Relations Review 66 (1), 32–54.

Kosteas, V.D., 2011. Job satisfaction and promotions. Industrial Relations 50, 174–194.

Lazear, E., 1992. The job as a concept. In: Bruns, W. (Ed.), Performance Measurement, Evaluations, and Incentives. Harvard University Press, Boston, pp. 183–215.

Lazear, E., Oyer, P., 2012. Personnel economics. In: Gibbons, R., Roberts, J. (Eds.), The Handbook of Organizational Economics. Princeton University Press, pp. 479–519.

Lazear, E., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. Journal of Political Economy 89, 841–864.

Leonard, J.S., 1990. Executive pay and firm performance. Industrial and Labor Relations Review 43, 13–29.

Main, B.G.M., O'Reilly III, C.A., Wade, J., 1993. Top executive pay: tournaments or; teamwork? Journal of Labor Economics 11, 606–628.

Medoff, J., Abraham, K., 1980. Experience, performance, and earnings. Quarterly Journal of Economics 95, 703–736.

Oreopoulos, P., von Wachter, T., Heisz, A., 2012. The short- and long-term career effects of graduating in a recession. American Economic Journal: Applied Economics 4 (1), 1–29.

Prendergast, C., 1993. The role of promotion in inducing specific human capital acquisition. Quarterly Journal of Economics 108, 523–534.

Rosen, S., 1986. Prizes and incentives in elimination tournaments. American Economic Review 76, 701–715.

Sullivan, P., 2010. Empirical evidence on occupation and industry specific human capital. Labour Economics 17, 567–580.

Waldman, M., 1984. Job assignments, signaling, and efficiency. Rand Journal of Economics 15, 255–267.

Waldman, M., 2012. Theory and evidence in internal labor markets. In: Gibbons, R., Roberts, J. (Eds.), The Handbook of Organizational Economics. Princeton University Press, pp. 520–574.

Propensity Scoring

MH Clark, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A propensity score is the conditional probability that a participant will be assigned to a treatment condition given a set of observed covariates. These scores are used as a single, aggregated covariate to make statistical adjustments through matching, weighting, stratifying or covariate adjustments to reduce selection bias in nonrandomized designs. Assuming that propensity scores are computed from all relevant covariates, propensity score adjustments should balance the treatment and control groups in an experiment and provide an unbiased estimate of the treatment effect.

Quasi-Experiments and Selection Bias

Although few would argue that randomized control trials are the most effective designs when making strong, causal inferences, many social and behavioral scientists find that their field experiments do not allow for random assignment to conditions. In such cases, researchers frequently opt for quasi-experiments instead. Quasi-experiments are very similar to randomized experiments except that they lack random assignment to conditions. Assignment to conditions may be based on self-selection, sequential selection, or preexisting groups. While nonrandom assignment may be more realistic or convenient, it also weakens the internal validity of an experiment, making it more difficult to make causal conclusions. According to John Stuart Mill, there are three criteria that establish our ability to make causal conclusions: (1) there is a significant relationship between the cause (independent variable) and effect (dependent variable), (2) the cause precedes the effect, and (3) we can rule out all other causal explanations – that is, the cause really is what is responsible for the treatment effect or change in the dependent variable (as cited by Shadish et al., 2002). In theory, random assignment balances all characteristics of those assigned to each treatment condition. Therefore, when the independent variable is manipulated, the only reasonable causal explanation for a difference in the treatment and control groups is that the effect was caused by the manipulated variable. In the absence of random assignment, we are no longer able to rule out alternate causal explanations for the cause and effect relationship, since it is very likely that other factors may be responsible for the difference between the conditions.

Although well-designed quasi-experiments can avoid many threats to internal validity, they are inherently at risk of selection bias. Selection bias is an imbalance of one or more factors between the experimental and comparison groups that influences the outcome in an experiment. It occurs when the reasons for selection are unknown, such as when participants self-select into a treatment condition for reasons unknown to the experimenter. For example, if alcoholics enroll themselves into a rehabilitative treatment program, they may be systematically different (i.e., more motivated to recover) than those who do not seek treatment. Therefore, the motivation to recover may be what is responsible for their recovery as much as (or instead of) the intervention itself.

Methods for Reducing Selection Bias

Using quasi-experiments does not prevent us from examining causal relationships, but they do limit our confidence when making causal inferences. Therefore, to increase the validity of these claims, we must modify our research designs, analyses, or both. When circumstances allow it, researchers can modify the research design by adding design elements (Shadish et al., 2002). Design elements include observations, comparison groups, and other factors that may or should not be related to the dependent variables. Adding observations, such as pretests, posttests, or follow-up tests, allow researchers to (1) test groups for similarities prior to an intervention, (2) examine long-term effects of a treatment, or (3) if other events coincide with the treatment, we can observe whether dependent variables continue to change with the presence or absence of our treatment or the other events. Adding comparison groups, such as a no-treatment control, placebo, or alternative or abbreviated treatments, can also permit researchers to identify what specific procedures are responsible for a change in outcome. Including a placebo will allow researchers to distinguish whether the true treatment (e.g., an antidepressant pill) is responsible for an improved outcome over the mere belief that one is being treated (e.g., by taking a sugar pill). Similarly, researchers may be able to see which specific components of an intervention are necessary by comparing a full treatment (e.g., diet, exercise, and reduced stress) to a partial treatment program (e.g., diet alone). Finally, researchers may also want to examine how variables other than the dependent variable are affected by the independent variable to see if the independent variable is what is responsible for the change in the dependent variable. If a researcher finds that a variable that correlates with the dependent variables is also affected by the intervention, it improves the likelihood that the treatment is the cause. However, if a variable that is unrelated to the dependent variable (or is not likely to be influenced by the intervention) changes after the intervention, it is possible that the intervention is not responsible for the change in either outcome.

Although modifying designs is more likely to control for selection bias than statistical adjustments, statistically controlling for known covariates that may bias outcomes is a popular alternative. Statistical adjustments are frequently performed after data have been collected or in cases when designs cannot

be altered. However, for statistical adjustments to be effective, researchers must have sufficient data to make these adjustments. Some of the more common methods of statistical adjustments use pretest observations as covariates in an analysis of covariance (ANCOVA) or to create change scores. Therefore, if treatment and control groups are different at pretest, these adjustments can account for these differences when comparing the posttest scores between the groups. Other methods of statistical adjustments include matching, stratification, weighting, and adding covariates to regression models (as opposed to an analysis of variance). These methods, which are described in greater detail below, can accommodate a variety of covariates. Sometimes they are used with a single covariate, such as a pretest; but often they are used with multiple covariates, such as when researchers want to control for several demographic characteristics at once. More recently, researchers combine the simplicity of using a single covariate with the thoroughness of using multiple covariates by using aggregate covariates. An aggregate covariate is a single measure that considers the variance of several variables concurrently. While there are several ways of aggregating measures, the propensity score is one of the most commonly used aggregate covariates.

What Is a Propensity Score?

A propensity score, often notated as $e(x)$, is the conditional probability that a participant would be assigned to a particular study condition based on a set of observed or measured covariates (Rosenbaum and Rubin, 1983). In most cases, they are estimated as the probability that a person would be assigned (or self-select) into the treatment condition, $\mathrm{pr}(z = 1|x)$, where z is the study condition ($z = 1$ specifies the treatment condition) and x is a set of observed covariates. As probabilities, propensity scores range from 0 to 1, where scores above 0.5 predict that a participant will be in the treatment group and those below 0.5 predict that a participant will be in the comparison group. These scores can then be used to equate groups on the observed covariates using statistical adjustment methods commonly used with single covariates, such as matching, stratifying, weighting, or covariate adjustments using ANCOVAs or multiple regressions (Stuart, 2010). In theory, propensity score adjustments (PSAs) should balance the treatment and control groups on all of the observed covariates used to compute the propensity scores, thereby reducing the bias created by nonrandom assignment. Assuming that all variables that influenced the selection into conditions and outcome variables are included as covariates, the adjusted treatment effects should be unbiased (Rosenbaum and Rubin, 1985).

When to Use Propensity Scores

Since their introduction by Rosenbaum and Rubin in 1983, propensity scores have been used to adjust treatment effects in nonrandomized experiments in a variety of social science fields (Casciano and Massey, 2011; Gunter and Daly, 2012; LaFrance et al., 2012), and their use has increased exponentially within the past 30 years (Bai, 2011). Unfortunately, there are adverse effects that come with that increased popularity, namely, that they may be used inappropriately. As with any statistic, there are certain conditions that must be met for the statistical results to be valid. While the continued testing of propensity scores helps to define the appropriate data parameters and sampling conditions (Pan and Bai, in press), it is also important to understand the designs that propensity scores can improve. While propensity score adjustments can be applied to a variety of nonrandomized studies, they were intended to be used with quasi-experiments in which the assignment method is unknown and there are several available covariates that relate to selection. PSAs were created to improve causal inference and internal validity; therefore, they are only appropriate when used with causal studies that have distinct interventions. If the assignment to treatment conditions is based on a known criterion, such as when students are assigned to reading programs based on a pretest of reading proficiency, a regression discontinuity design may be just as effective, and easier to use, than PSAs. It is also important that researchers using PSAs are able to compute propensity scores from several covariates related to both selection into conditions and the dependent variable. If researchers decide to use propensity scores after data have already been collected, they may be limited to a few demographic covariates that do not sufficiently model the selection process. Therefore, it may be preferable to consider the covariates that could influence selection before data are collected so that these can be measured.

Computing Propensity Scores

Selecting Covariates for Propensity Scores

Covariates used to estimate the propensity scores should be related to both the assignment condition (treatment or control) and the dependent variables of the experiment (Austin et al., 2007). Traditional regression adjustments often consider only how covariates relate to the dependent variable. However, if covariates are selected only for their correlations with the outcomes, the resulting propensity scores will not improve the unobserved balance between treatment groups. Since the propensity score is intended to predict selection into conditions, this needs to be the first consideration when selecting covariates. If we are able to account for all factors that influence selection into conditions, the adjusted treatment effects should replicate those from random experiments (Rubin, 2008). While this may be sufficient to achieve balance, some of these covariates may be unnecessary if they are not also related to the outcome. For example, if the treatment group has proportionally more males than the control group, the groups may be unbalanced with respect to biological sex. However, this is only a problem in estimating the treatment effect if sex is correlated with the dependent variable. If so, sex is a nonignorable covariate and should be included as a covariate (Rosenbaum and Rubin, 1983). If not, it may be excluded from the propensity score estimation with little change to the adjusted treatment effects.

However, a bigger concern in selecting covariates is that researchers measure and include enough nonignorable covariates when computing propensity scores. One of the assumptions when using propensity scores is that all nonignorable covariates are included in the analyses. If a covariate, such as

biological sex, is related to selection and the outcome, but is not included in the propensity score estimation, the treatment effects will still be biased. Steiner et al. (2010) found that when covariates were limited to only a few commonly available demographic covariates (i.e., age, ethnicity, sex, and marital status), less than half of the selection bias was removed. Although it was not always the case that adding more covariates reduced more bias, it was important that the most influential covariates (i.e., pretests) were included in the propensity score models. In addition to selecting covariates that are linearly related to conditions and outcomes, it may also be beneficial to include nonlinear covariates or interactions of covariates (Dehejia and Wahba, 1999; Lee et al., 2010).

Although it is important to determine the statistical relationships between each covariate and the assignment condition and outcome, it is not necessary that the covariates are *significantly* related (at $p < 0.05$) to be included in the propensity score model. Even when using common univariate tests (e.g., *t*-tests, chi-squares, odds ratios, Pearson correlations) to assess whether or not there are relationships between individual covariates and assignment condition and outcome, some researchers use a less stringent p value (i.e., $p < 0.20$; Clark and Cundiff, 2011; Diamond and Sekhon, 2013). Other researchers use effect sizes, such as a standardized bias (SB) estimate, to measure the difference between the treatment and control groups for each covariate.

$$\mathrm{SB} = \frac{M_T - M_C}{\sqrt{\frac{s_T^2 + s_C^2}{2}}} \times 100 \qquad [1]$$

where M_T is the mean of the treatment group, M_C is the mean of the control group, s_T^2 is the variance of the treatment group, and s_C^2 is the variance of the control group (Bai, 2013; Rosenbaum and Rubin, 1985). When the covariates are dichotomous, odds ratios can be converted to SB values using a formula similar to Hasselblad and Hedges's (1995):

$$\mathrm{SB} = \frac{\ln\left(OR\right)\sqrt{3}}{\pi} \times 100 \qquad [2]$$

SB values above 5 suggest that the covariates are not balanced between the treatment and control groups and are nonignorable. Unfortunately, just as using hypothesis tests with $p < 0.05$ are probably too liberal for assessing bias, using the above formulae with $\mathrm{SB} > 5$ may be too stringent.

Procedures for Computing Propensity Scores

Logistic regression is the most common statistical procedure for computing propensity scores. In this procedure, all of the selected covariates are concurrently included in a logistic regression model to predict the assignment condition, and the propensity scores are the resulting predicted probabilities for each unit. *Classification* or *regression trees* are used to predict assignment through a sequence of hierarchical, binary splits. Each split is determined by the probability that a participant will select into each condition based on a single covariate. The splitting process continues for each subsequent covariate until a specified number of nodes (or branches) is obtained or until all covariates are included in the model. The resulting binary tree has terminal nodes representing groups of participants who have the same predicted probability for being in the treatment condition (propensity score), even though each node may have estimated the same propensity score from different predictors (Westreich et al., 2010).

Bagging, also known as bootstrap aggregating, averages the results of many classification trees from a series of bootstrapped samples (Lee et al., 2010). In this procedure, cases of participants are randomly sampled (with replacement) from the available data set to estimate propensity scores using the classification tree procedure described above. Cases that are not sampled are used to "estimate the misclassification rate" (Luellen et al., 2005, p. 544). A new classification tree is estimated for each bootstrapped sample, and the resulting trees are aggregated to create a more stable propensity score for each participant. *Boosted modeling* averages multiple propensity score models that were created using different combinations of covariates. Each model is weighted based on its ability to predict the assignment conditions and aggregated to form a single, stable model. Unlike bagging, it uses all available cases from the original sample for each model, and it weights cases based on how difficult they are to classify. Each individual prediction model can be estimated with either logistic regression or classification trees (Lee et al., 2010).

In a study that compared three of these procedures for estimating propensity scores, Luellen et al. (2005) found that logistic regression tended to create propensity scores that were able to reduce bias better than classification trees or bagging. Although propensity scores created through bagging with 100 bootstrapped samples were able to reduce as much bias as those created using a logistic regression, bagging using fewer samples did not yield good propensity scores or consistent results.

Predicting Assignment for Multivalued Treatments

Although propensity scores were originally created to predict the probability that participants would select into one of two conditions, researchers frequently use quasi-experiments that test more than two levels of treatment. In such designs, estimating the propensity score becomes more complicated. When predicting several nominal treatment conditions, researchers can estimate propensity scores by creating separate propensity score models for each possible pair of conditions. Therefore, an experiment with three conditions (i.e., control, medication, and psychotherapy) will have three possible pairs of conditions and will have three different propensity score models. These multiple paired comparisons can be estimated using logistic regression (Falconer et al., 2011; Rubin, 1997) or multinomial or nested logit models (Feng et al., 2011; Imbens, 2000). Although having to account for multiple propensity scores is more analytically cumbersome than accounting for a single covariate, this method maintains a traditional approach to computing propensity scores while accommodating multiple factor levels. Ordinal predictions permit researchers to use an ordinal logistic regression to predict ordered dose-like levels of conditions (i.e., no treatment, partial treatment, full treatment) and results in a single propensity score (Joffe and Rosenbaum, 1999; Park and Russell, 2012). When the independent variable is continuous (i.e., number of cigarettes smoked), Imai and van Dyk (2004)

demonstrated that propensity scores can be created using a linear regression.

Assessing Propensity Score Balance

It is important to keep in mind that the primary goal of the propensity score is not to optimally predict assignment to conditions, but to balance the distributions of covariates over conditions (Rosenbaum and Rubin, 1984). Since the propensity score is an aggregate measure of multiple covariates, it stands to reason that if the propensity scores are balanced between the treatment and control groups, then adjustments using the propensity scores will balance the groups on the individual covariates. Therefore, before using the propensity scores to make statistical adjustments, they should be tested to see whether or not the propensity score distributions are similar for each treatment level. Although there are several statistical tests to determine balance, it is not clear which one is best; therefore, it may be useful to use more than one. The simplest method is to graph propensity score distributions for the treatment and control groups and visually inspect the extent to which they overlap (Bai, 2013; Shadish et al., 2008). For a more objective estimate, a Kolmogorov–Smirnov test can be used to test for a difference between the distributions (Diamond and Sekhon, in press). Finally, Rubin (2001) suggests that propensity scores are balanced if (1) the difference between the treatment and control group means of the propensity scores is less than half a standard deviation, (2) the ratio of the treatment and control group variances of the propensity scores is close to one, and (3) the ratio of the group variances of the residual errors of each covariate predicting the propensity scores is close to one. If the propensity scores are not balanced, rescaling the propensity scores using a logit transformation will normalize the distributions and improve linear modeling (Rubin, 2001).

Adjustments Using Propensity Scores

Matching

Matching identifies and pairs or groups similar cases from experimental and control groups based on the proximity of their propensity scores. Common methods for matching include *exact matching*, in which cases are paired on exactly the same scores; *nearest neighbor* or *greedy matching*, which selects the closest matches first and then progressively farther matches; *caliper matching*, which limits the distance of matches on a specified criterion; and *optimal matching* or *net costmatching*, which finds the smallest cumulative distance from all possible pairs. Although exact matching provides the most precise pairing, it is often unrealistic to use with continuous data unless the range is very small (i.e., age of fourth graders). **Table 1** illustrates the paired results for the last three types of matching.

Although the first few matches made with nearest neighbor matching are very good, the last two matches are poor. Setting a limit to the distance of the match will prevent nearest neighbor matching from having these poor matches. Assuming that the ratios of the variances of the propensity scores in the treatment and control group are close to 1, using a caliper that is four times the pooled standard deviation (s_p) of the propensity scores can remove up to 95% of the selection bias. However, some researchers choose smaller calipers, such as 0.25 s_p. If the pooled standard deviation of the propensity scores is 0.24, then the caliper width is 0.06 (using the second caliper standard). Therefore, the notreatment case can be matched with a control case if the distance between the two propensity scores is farther than 0.06. Since the last two matches using the nearest neighbor method have distances of 0.12 and 0.31, these four cases would not have matches and would be dropped from the analyses that would estimate the treatment effect. Optimal matching does not necessarily drop cases, but arranges pairs in a way so they have the smallest global distance. Although the closest match to the treatment case with PS = 0.63 is a control case with PS = 0.64, this match does not benefit the sample as a whole. Optimal matching considers all possible matches concurrently. Therefore, the summed distances between the propensity scores in the treatment and control groups drop from 0.53 using nearest neighbor matching to 0.51 using optimal matching. Because the poor matches are removed in caliper matching, the total distance between all matched pairs is only 0.1 using the sample data in **Table 1**. Bai (2013) has shown that caliper matching generally tends to reduce bias better than other matching methods. Unfortunately, it drops several cases in both the treatment and control groups, which reduces the sample size and generalizability of the results.

To limit the number of treated participants dropped from the analyses, some researchers recommend having proportionately more participants in the control group to match with those in the treatment group (Rubin and Thomas, 1996;

Table 1 Propensity score matching methods

Unmatched		*Nearest neighbor*		*Caliper*		*Optimal*	
Treatment	*Control*	*Treatment*	*Control*	*Treatment*	*Control*	*Treatment*	*Control*
0.25	0.13	0.63	0.64	0.63	0.64	0.38	0.36
0.38	0.36	0.38	0.36	0.38	0.36	0.41	0.45
0.41	0.45	0.74	0.71	0.74	0.71	0.63	0.58
0.63	0.58	0.41	0.45	0.41	0.45	0.74	0.64
0.74	0.64	0.25	0.13			0.25	0.13
0.89	0.71	0.89	0.58			0.89	0.71

Note: The values in the table are the propensity scores on which participants are matched.

Stuart, 2010). Researchers can also use *index matching* or *ratio matching*, which matches each treatment participant with multiple control participants (Shadish et al., 2002; Stuart, 2010). For example, instead of pairing one person in the treatment group to one person in the control group, we might match each person in the treatment group with three people in the control group. *Full matching* not only permits several control cases to be matched with a single treatment case, but also allows several treated cases to be matched with a single control case. Once treatment and control cases are matched, adjusted treatment effects can be made using a paired samples *t*-test if matching a treatment case to a single control case (Austin, 2008), a generalized estimating equation (GEE) in which the matches are random effects (Clark and Cundiff, 2011), or a McNemar's test if the dependent variable is dichotomous (Falconer et al., 2011).

Stratification

Stratification (also known as subclassification) classifies all of the participants into several categories or strata based on their propensity scores so that each stratum includes participants from the treatment and control groups. Although the number of strata and methods for classifying participants into a specific stratum (i.e., within a certain percentile interval or propensity score interval) can vary (Luellen et al., 2005), Cochran (1968) proposed that five strata of equal sample sizes could account for 90% of the bias from the observed covariates. Once strata are created, there are several ways to compute the adjusted effects. When the dependent variable is continuous, the most common method is to compute a between-subjects analysis, such as an independent measures *t*-test, for each stratum and average effect sizes across strata (Austin and Mamdani, 2006; Shadish et al., 2008). Another method is to use a treatment-by-strata factorial ANOVA that accounts for the propensity scores in the error variance when estimating the main effect for treatment (Clark, in press). If the dependent variable is dichotomous, chi-squares or other contingency table approaches should be used.

Although stratification usually permits the inclusion of all cases, some cells may contain few or no units. Reducing the number of strata or changing the cut-points that define the strata may reduce this problem, but determining ideal cut-points for strata can be difficult (Luellen et al., 2005). While stratification is not as commonly used as matching, the statistical procedures are often less complicated.

Covariate Adjustment

Covariate adjustment uses propensity scores as a covariate in an ANCOVA or a multiple regression and removes bias by accounting for the correlation between the propensity scores and the dependent variable (Rosenbaum and Rubin, 1983; Shadish et al., 2008). Although some studies have indicated that covariate adjustment can be an effective method for removing bias (Austin and Mamdani, 2006; Clark, in press), other researchers do not recommend this method unless covariates or propensity scores are linearly related to the outcome (Rosenbaum, 2001) and variances among covariates are homogeneous (Rubin, 2001). However, in some cases, covariate adjustments using propensity scores may not sufficiently reduce covariate bias even when these conditions are met (Shadish et al., 2008).

Weighting

Weighting attempts to balance treatment and control groups by multiplying observations by the inverse of the propensity score. Two commonly used propensity score weighting estimators are the *inverse probability of treatment weighted* (*IPTW*) *estimator* and the *standard mortality ratio* (*SMR*) *weighted estimator* (Kurth et al., 2006). The IPTW estimator weights observations in both the treatment and control groups by either the inverse of the propensity score or one minus the propensity score. The adjusted mean difference between the treatment and control groups can be found using the following equation by Hirano and Imbens (2001):

$$\widehat{\tau} = \frac{\sum_{i=1}^{N} \frac{r_i z_i}{e(x)_i}}{\sum_{i=1}^{N} \frac{z_i}{e(x)_i}} - \frac{\sum_{i=1}^{N} \frac{r_i(1-z_i)}{1-e(x)_i}}{\sum_{i=1}^{N} \frac{1-z_i}{1-e(x)_i}} \qquad [3]$$

where r_i is the dependent variable, z_i is the treatment condition, and $e(x)_i$ is the propensity score. The treatment effect can then be found by substituting the mean difference in an independent measures *t*-test with this adjusted mean difference.

The SMR weighted estimator only weights the observations for the control group by the odds of the propensity score. The adjusted mean for the control group (M_C) can be found by the equation by McCaffrey et al. (2004):

$$M_C = \frac{\sum_{i=1}^{N} r_i(1-z_i)\left(\frac{e(x)_i}{1-e(x)_i}\right)}{\sum_{i=1}^{N} (1-z_i)\left(\frac{e(x)_i}{1-e(x)_i}\right)} \qquad [4]$$

A single sample *t*-test could be used to determine the difference between this adjusted mean and the unadjusted mean of the treatment group. Although some studies have found that propensity score weighting is effective in reducing bias (Austin, 2010; Kurth et al., 2006), it may not be effective when many propensity scores are very close to one or zero (Rubin, 2001).

Assessing Covariate Balance

Once statistical adjustments are made to the observations, it is important to assess how well the propensity scores balanced the individual covariates between the treatment and control groups. Although univariate tests are commonly used to test for significant differences between groups for individual covariates and omnibus tests, such as Hotelling's T or factorial ANOVAs, have been used to test for imbalance between the covariates as a set, Ho et al. (2007) argue that hypotheses tests are inappropriate, since "balance is a characteristic of the observed sample, not some hypothetical population" (p. 221). Therefore, in lieu of hypotheses tests, SB estimates (eqns [1] and [2]) and measures of bias reduction (Bai, 2013; Shadish et al., 2008) may be used to determine how well PSAs reduced selection bias. If the SB estimate for a given covariate is less than 5 after the PSA, then the adjustment sufficiently

Table 2 Covariate balance before and after propensity score adjustments

Covariate	SB_{unadj}	SB_{adj}	*PBR*
Age	19.50	2.25	88.46%
High School GPA	70.25	12.75	81.85%

balanced the treatment and control groups on that covariate. Even if the SB estimate did not reach this benchmark, we can determine the percent of bias reduction (PBR) from

$$\mathrm{PBR} = \frac{|\mathrm{SB}_{\mathrm{unadj}}| - |\mathrm{SB}_{\mathrm{adj}}|}{|\mathrm{SB}_{\mathrm{unadj}}|} \times 100 \qquad [5]$$

where $\mathrm{SB}_{\mathrm{unadj}}$ is the SB estimate of the covariate before a PSA and $\mathrm{SB}_{\mathrm{adj}}$ is the bias estimate of the covariate after the adjustment (Bai, 2013; Cochran and Rubin, 1973). For example, **Table 2** shows that a PSA was able to reduce the SB for 'Age' from 19.5 to 2.25, which suggests a sufficient reduction in bias. Although 'High School GPA' did not meet the same criterion, the PSA was still able to reduce more than 80% of the observed bias.

Computer Programs That Compute PSAs

There are currently several computer programs and macros that support PSAs. Any statistical package that computes a logistic regression will allow researchers to estimate propensity scores. 'MatchIt' in R (Ho et al., 2011), 'psmatch_multi' in SAS (Fraeman, 2010), 'PS Matching' (Thoemmers, 2012) and Painter's macro (2004) in SPSS, and 'teffects psmatch' and 'psmatch2' in Stata (Leuven and Sianesi, 2003; StataCorp, 2013) all allow researchers to implement propensity matching using a variety of matching methods. While weighting, covariate adjustment and stratifying can also be done in a variety of statistical packages; there are fewer that offer adjustment methods specific to propensity scores (Lanehart et al., 2012; Holmes, 2014).

Limitations of PSAs

Although many studies have shown that PSAs reduce bias as well as or better than traditional covariate adjustments (Shah et al., 2005; Stürmer et al., 2006), other studies have found that traditional or alternative adjustment methods are more effective in reducing bias (Austin, 2010; Kurth et al., 2006; Shadish et al., 2008). This suggests that the specific conditions under which propensity scores are used affects their ability to reduce bias. Statistical adjustment methods, sampling methods, sample sizes, and available covariates all influence the quality of the propensity score and its ability to sufficiently reduce bias (Austin and Mamdani, 2006; Bai, in press; Steiner et al., 2010). Even when propensity scores can sufficiently balance observed covariates, they cannot adjust for hidden bias, which occurs when relevant covariates are not included in the propensity score model. Without including all covariates that contribute to selection bias in the propensity score model, PSAs cannot provide unbiased treatment effects.

See also: Causal Inference: Overview; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Instrumental Variables in Statistics and Econometrics; Optimal Experimental Design; Quasi-Experimental Designs.

Bibliography

Austin, P.C., 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in Medicine 27, 2037–2049.

Austin, P.C., 2010. The performance of different propensity score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Statistics in Medicine 29, 2137–2148.

Austin, P.C., Grootendorst, P., Anderson, G.M., 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statistics in Medicine 26, 734–753.

Austin, P.C., Mamdani, M.M., 2006. A comparison of propensity score methods: a case study estimating the effectiveness of post-AMI statin use. Statistics in Medicine 25, 2084–2106.

Bai, H., 2011. Using propensity score analysis for making causal claims in research articles. Educational Psychology Review 23, 273–278.

Bai, H., 2013. A bootstrap procedure of propensity score estimation. The Journal of Experimental Education 81, 157–177.

Bai, H. Current issues on propensity score matching: matching with/without replacement, common support, and sample ratio. In: Pan, W., Bai, H. (Eds.), Propensity Score Analysis: Fundamentals, Developments and Extensions. Guilford Press, New York, NY, in press.

Casciano, R., Massey, D.S., 2011. School context and educational outcomes: results from a quasi-experiment. Urban Affairs Review 48, 180–204.

Clark, M.H., Cundiff, N.L., 2011. Assessing the effectiveness of a college freshman seminar using propensity score adjustments. Research in Higher Education 52, 616–639.

Clark, M.H. A comparison of propensity score adjustment methods for reducing selection bias. In: Pan, W., Bai, H. (Eds.), Propensity Score Analysis: Fundamentals, Developments and Extensions. Guilford Press, New York, NY, in press.

Cochran, W.G., 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24, 205–213.

Cochran, W.G., Rubin, D.B., 1973. Controlling bias in observational studies: a review. Sankhya, Series A 35, 417–446.

Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association 94, 1053–1061.

Diamond, A., Sekhon, J.S., 2013. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics 95, 932–945.

Falconer, M.K., Clark, M.H., Parris, D., 2011. Validity in an evaluation of healthy families Florida: a program to prevent child abuse and neglect. Children and Youth Services Review 33, 66–77.

Feng, P., Zhou, X.H., Zou, Q.M., Fan, M.Y., Li, X.S., 2011. Generalized propensity score for estimating the average treatment effect of multiple treatments. Statistics in Medicine 31, 681–697.

Fraeman, K.H., 2010. An introduction to implementing propensity score matching with SAS. Proceedings of the NorthEast SAS Users Group (online). Available from: http://www.nesug.org/Proceedings/nesug10/ad/ad05.pdf (01.10.13.).

Gunter, W.D., Daly, K., 2012. Causal or spurious: using propensity score matching to detangle the relationship between violent video games and violent behavior. Computers in Human Behavior 28, 1348–1355.

Hasselblad, V., Hedges, L.V., 1995. Meta-analysis of screening and diagnostic tests. Psychological Bulletin 117, 167–178.

Hirano, K., Imbens, G.W., 2001. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Services and Outcomes Research Methodology 2, 259–278.

Ho, D.E., Imai, K., King, G., Stuart, E.A., 2007. Matching as a nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15, 199–236.

Ho, D.E., Imai, K., King, G., Stuart, E.A., 2011. MatchIt: nonparametric preprocessing for parametric causal inference. Journal of Statistical Software 42, 1–28.

Holmes, W.M., 2014. Using Propensity Scores in Quasi-Experimental Designs. Sage, Thousand Oaks, CA.

Imai, K., van Dyk, D.A., 2004. Causal inference with general treatment regimes: generalizing the propensity score. Journal of the American Statistical Association 99, 854–866.

Imbens, G.W., 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87, 706–710.

Joffe, M.M., Rosenbaum, P.R., 1999. Propensity scores: invited commentary. American Journal of Epidemiology 150, 327–333.

Kurth, T., Walker, A.M., Glynn, R.J., Chan, K.A., Gaziano, J.M., Berger, K., Robins, J.M., 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. American Journal of Epidemiology 163, 262–270.

LaFrance, J.P., Rahme, E., Iqbal, S., Elftouh, N., Vallee, M., Laurin, L.P., Ouimet, D., 2012. Association of dialysis modality with risk for infection-related hospitalization: a propensity score-matched cohort analysis. Clinical Journal of the American Society of Nephrology 7, 1598–1605.

Lanehart, R.E., Rodriguez de Gil, P., Kim, E.S., Bellara, A.P., Kromrey, J.D., Lee, R.S., 2012. Propensity score analysis and assessment for propensity score approaches using SAS procedures. Proceedings of the SAS Global Forum (online). Available from: http://support.sas.com/resources/papers/proceedings12/314-2012.pdf (01.10.13.).

Lee, B.K., Lessler, J., Stuart, E.A., 2010. Improving propensity score weighting using machine learning. Statistics in Medicine 29, 337–346.

Leuven, E., Sianesi, B., 2003. PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing (online) Available from: http://ideas.repec.org/c/boc/bocode/s432001.html (01.10.13.).

Luellen, J.K., Shadish, W.R., Clark, M.H., 2005. Propensity scores: an introduction and experimental test. Evaluation Review 29, 530–558.

McCaffrey, D.F., Ridgeway, G., Morral, A.R., 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods 9, 403–425.

Pan, W., Bai, H. (Eds.). Propensity Score Analysis: Fundamentals, Developments and Extensions. Guilford Press, New York, NY, in press.

Painter, J., 2004. Propensity Score Matching via SPSS (online). Available from: http://www.unc.edu/~painter/SPSSsyntax/propen.txt (01.10.13.).

Park, J., Russell, L.B., 2012. Effect of non-leisure physical activity on mortality in U.S. adults: does propensity score matching make a difference? Annals of Epidemiology 22, 575–580.

Rosenbaum, P.R., 2001. Observational studies. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioral Sciences. Elsevier Science Limited, Oxford, pp. 10808–10815.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P.R., Rubin, D.B., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79 (387), 516–524.

Rosenbaum, P.R., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician 39, 33–38.

Rubin, D.B., 1997. Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine 127 (8), 757–763.

Rubin, D.B., 2008. For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2, 808–840.

Rubin, D.B., 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology 2, 169–188.

Rubin, D.B., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52, 249–264.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA.

Shah, B.R., Laupacis, A., Hux, J.E., Austin, P.C., 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. Journal of Clinical Epidemiology 58, 550–559.

Shadish, W.R., Clark, M.H., Steiner, P.M., 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. Journal of the American Statistical Association 103, 1334–1344.

StataCorp, 2013. Stata Treatment-Effects Reference Manual: Potential Outcomes/Counterfactual Outcomes (Release 13). StataCorp LP, College Station, TX.

Steiner, P.M., Cook, T.D., Shadish, W.R., Clark, M.H., 2010. The differential role of covariate selection and data analytic methods in controlling for selection bias in observational studies: results of a within-study comparison. Psychological Methods 15, 250–267.

Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., Schneeweiss, S., 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariate methods. Journal of Clinical Epidemiology 59, 437–447.

Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 25, 1–21.

Thoemmers, F., 2012. Propensity Score Matching in SPSS (online) Available from: http://arxiv.org/ftp/arxiv/papers/1201/1201.6385.pdf (01.10.13.).

Westreich, D., Lessler, J., Funk, M.J., 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. Journal of Clinical Epidemiology 63, 826–833.

Proper and Common Names: Impairments of Anomia

Carlo Semenza, Università di Padova, Padova, Italy; and I.R.C.C.S. Ospedale S. Camillo, Lido di Venezia, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Proper names essentially refer to individual entities while common names refer to categories. This fact makes the link proper names have with their reference particularly fragile. As a consequence, processing proper names seems to require special cognitive and neural resources. Neuropsychological dissociations demonstrate that proper names and common names follow functionally and anatomically distinct neural networks. These networks might have been shaped by precise evolutionary demands. Their precise location and working, depending on specific tasks, is still partly unknown. They are, moreover, differently sensitive to focal or generalized brain damage, cognitive changes with age, or lack of organic resources. The sensitivity of proper names to brain damage makes them a useful clinical tool in the early diagnosis of mental deterioration.

Common and proper names have a different type of reference. This is a fundamental distinction that has been shown to impinge on the kind of processing each of the two categories undergoes at the neural level. Its biological value, as it will later be argued, may have been shaped by precise evolutionary demands. Research in neuroscience has thus contributed to the understanding of how proper names are processed in the brain with respect to common names. Past and current philosophical and linguistic theories on the nature of these two lexical categories have directly inspired neuropsychological research, receiving, in turn, empirical support.

The main theoretical idea, originated in philosophy, is that, unlike common names, proper names carry reference, but not (except perhaps in minor role) sense. In other words, they may be considered 'pure referring expressions' (e.g., Mill, 1893; Kripke, 1980) or, as some might prefer (e.g., Searle, 1969), as loci of identifying 'descriptions.' These conclusions are not straightforward, and, indeed they conceal issues that are hotly debated. While aware of this, the neuropsychologist's position is, however, that neuropsychological findings support a common name/proper name distinction in one version or another, the different versions depending on the amount and type of descriptive resources associated with a name. What is sufficient to account for neuropsychological findings is the view that proper names, as opposed to common names, denote individual entities but do not essentially rely on, or rely on only to a limited extent, sets of attributes. Proper names thus essentially refer to individuals (or individual groups) while common names refer to categories. Another useful way of expressing this distinction is that proper names have only 'token' reference while common names carry 'type reference' (Jackendoff, 1983).

As Semenza et al. (1998) observed, a name designating a category applies to a set of attributes that overlap or interact with each other via high-probability connections. In the set of attributes labeled by a proper name, instead, attributes combine together incidentally; they relate to each other only by virtue of belonging to entities that are unique. Moreover, changing features over time does not change the name.

The distinction in the function of common and proper names thus reminds, in some respects, that made between semantic and episodic memory, but with an important difference: the mechanisms distinguishing proper and common names are more peripheral and operate at the interface with the lexical level. Naming individuals with respect to categories does not only imply different connections among defining attributes within semantic memory: it also implies different mechanisms in the activation of the phonological word form. These mechanisms may be understood as types of information-collecting nodes and pointers. In theoretical models proposed in the late 1980s, the main representational and processing differences between proper names and common names have been thus identified in the activation of the phonological form. A review by Cohen and Burke (1993) concluded that, using essentially overlapping concepts, all models emphasized the difference in connectivity serving the token/type distinction. In virtually all these models, only a single connection is envisaged from the individual entity labeled by a proper name and the name. Connections between the components of a concept and a common name are instead multiple and, as a result, are on the whole stronger.

The token/type distinction may not be entirely sufficient to explain all neuropsychological findings, however. Other properties of semantic memory must come into play. The token/type distinction, for example, does not easily explain the case of deictic pronouns (e.g., 'I' or 'you'). Surely these pronouns make token reference. However, no patient with a selective proper name deficit has ever been shown to have problems with deictic pronouns. Clearly, there is an overwhelming difference in word frequency in favor of pronouns. Moreover, 'I' applies to any person pronouncing the word, and this differentiates 'I' from names of persons. The large majority of what are intuitively considered proper names include names of persons, individual animals, specific objects or events, buildings, brands, or geographical places. Some of these subcategories seem to behave in the brain differently from one another. Other referential expressions may pose problems for using the token/type distinction in explaining all phenomena observed in neuropsychology, for example, 'temporal' names: those religious festivities (Easter, Yom Kippur, Ramadan), days, or months. The names of religious festivities have not so far been investigated in patients with deficits relative to proper names and no patient was shown to have problems with the names of days or months. Word frequency may be a factor, but additional properties of temporal names may contribute in keeping keep them apart in the brain from other proper names.

In conclusion, with these provisos, the distinction between individual and general semantics goes a long way in explaining available neuropsychological data.

Neuropsychological Dissociations

In neuropsychology, interesting cases of dissociation were discovered whereby, as a result of a brain lesion, patients were affected by selective anomia for proper names, with normal retrieval of common names. The complementary dissociation, proper name selective sparing, has also been reported.

Anomia for proper names, in general, affects retrieval in all testing conditions and comes in at least four different varieties. The first, and most common, variety, called 'pure proper name anomia' is that originating with a deficit at a postsemantic level. It has been interpreted as a failure to access the name's phonological form from an intact semantic system (e.g., Semenza and Zettin, 1988, 1989; Lucchelli and De Renzi, 1992; Hittmair-Delazer et al., 1994). The phonological form itself seems to be spared in these cases, since the ability to read aloud irregularly spelled proper names is preserved. Two different subtypes have been reported: one concerns all proper names, while the other is limited to the names of persons. It is still unclear whether proper names other than the ones of persons are easier than the latter or rather have particular properties (see above) that may help in circumventing the deficit. Most of these cases are not sensitive to phonemic as well as to semantic cueing, but a combination of both types of cue has been shown to yield some improvement (Semenza and Zettin, 1988). Other patients showing a significant improvement when aided with a phonological cue have been reported (e.g., Lucchelli and De Renzi, 1992; Otsuka et al., 2005).

A second variety of proper name anomia derives from a problem within the semantic system. In this case (e.g., Miceli et al., 2000), information about individual people is selectively lost with respect to information about other entities. In a third variety, instead, the deficit seems to consist in the isolation of information on individual entities from both the general semantic system and the output lexicon (Van der Linden et al., 1995; Semenza et al., 1998). Thus, while some patients cannot retrieve the name but have no problem showing they know everything else about the individual, other less frequently observed patients seem to be selectively impaired in the semantic store containing all the information about that individual, including the corresponding name. In 'isolation' cases, the information was available but could be retrieved by the patient only if given the proper name itself. A last variety is that of 'prosopanomia' (Semenza et al., 2003b). This appears to be a face-specific optic aphasia: while unable to retrieve the names of people looking at their faces, the patient can do so on definition. Unlike in prosopagnosia, the selective inability to recognize faces, there is no trouble in recognizing the face, but like in optic aphasia, the deficit seems to be instead to retrieve enough semantic information from visual stimuli (in this case only from faces) to activate the name.

Anomias for proper names of the first type are also important for their concomitant symptoms. At least in the purest cases, where anomia seems the only reported symptom, careful neuropsychological assessment systematically uncovers a constellation of associated deficits. Semenza and Zettin (1989) thus showed that their proper name anomic could not retrieve a member of an arbitrarily related word pair given the other; likewise, the same patient and similar ones could not retrieve 'token' information arbitrarily connected with other information, such as known phone numbers of people, titles of wordless pieces of music, dates of known events, etc. This finding directly supports theories viewing proper names as pure referring expressions, whose link with the entity they represent is 'arbitrary,' nondescriptive (i.e., not implying attributes), and thus devoid of sense.

Cases of selective sparing of proper names are generally found in otherwise very severely affected patients. These cases are also more complicated than proper name anomias. Severity, on the one hand, has seriously hampered the possibility of testing the patients in the desirable level of detail. Moreover, the conditions in which proper names are selectively spared seem to differ widely from case to case. Unlike in cases of the most common (postsemantic) proper name anomia, which all look rather similar, no single reported case of proper name sparing closely resembles any of the others. The first such case was observed by McKenna and Warrington (1978) and concerned only the names of nations. In the case of Cipolotti et al. (1993), the patient could be tested only in writing, where he could retrieve only names of countries and famous people. Semenza and Sgaramella's (1993) patient could not retrieve any name: the patient's selective preservation of proper names emerged only in the middle of a gibberish jargon, and, more clearly, after phonemic cueing in picture naming. Cipolotti's (2000) patient showed a proper name superiority just for countries, only in oral naming and reading aloud. Schmidt et al. (2004) reported sparing of proper names only in the written condition, in a patient whose articulation was totally impaired. This patient's semantic knowledge about items whose (common) name he could not retrieve was preserved. He never committed paragraphias; these facts led the authors to suggest that the functional deficit could be located in accessing, from an intact semantic level, an intact orthographic lexicon. Unfortunately, this patient's articulation problems prevented assessing the status of his phonological lexicon. In the case described by Lyons et al. (2002), the patient missed not only common nouns but also the correspondent semantic information: in contrast, the patient showed a normal ability to retrieve people's names and the corresponding biographical information. The case of Lyons et al. thus mirrors that of Miceli's et al. (2000). Finally, a selective sparing of proper names was also found by Pavão Martins and Farrajota (2007). Their case, a problem in accessing the phonological form of common names from an intact semantic system mirrors Semenza and Zettin's 'pure proper name anomia' cases (1988; 1989). Importantly, moreover, the same article reports the complementary side of the dissociation: another case of pure proper name anomia, matched for age and education and tested on exactly the same material as the sparing case.

In summary, a double dissociation was found in the production of proper and common names. This double dissociation was found at two different levels, at least. Pavão Martins and Farrajota (2007) showed the dissociation in accessing an intact phonological form from an intact semantic system. Combined together, the cases of Miceli et al. (2000)

and of Lyons et al. (2002) provide a double dissociation at the level of semantic memory. These findings cannot be explained as resulting from a generic difficulty factor, at least as far as the access to phonological form is concerned, since the two patients reported by Pavão Martins and Farrajota (2007) were tested on exactly the same material. Thus, what was difficult for one patient was easy for the other patient and vice versa. These findings support overall the idea that processing proper and common names follows at least partially independent processing pathways in the brain. A theoretical model, based on neuropsychological findings and capitalizing on the separation of individual semantics from general semantic memory and on separate access of this information to the phonological output lexicon, has been proposed by Semenza (2006, 2009). As shown in **Figure 1**, this model accounts for the observed patterns of anomia for proper names.

The Problem of Difficulty

The above-mentioned double dissociation does not, in itself, allow the inference that the two independent tasks are of equal difficulty, however. In fact, the problem of whether the independent pathways used to process proper and common names pose the same burden to the cognitive system is a different problem, independent from the functional and anatomical separateness of the two categories. One pathway, i.e., the proper name pathway, has been traditionally thought to be intrinsically more difficult to use, but an experimental confirmation of this hypothesis has been difficult to obtain (Cohen and Burke, 1993). Pictures of faces (for proper names) and pictures of objects (for common names) are very hard to match for perceptual difficulty. Diary studies, collecting participants' naming failures in a given period of time, may be hampered by the fact that one often hardly notices missing a common name, while missing a proper name may cause serious embarrassment. The best available evidence for proper names being indeed more difficult to process rests mainly on two types of observations. The first concerns the so-called 'baker–baker' paradox: it is easier to learn that a face belongs to a baker than it is to learn that the same face belongs to a Mr Baker (McWeeny et al., 1987). This effect cannot be attributed to differences in the phonological form or frequency of occurrence of occupations versus proper names. Another methodology, repetition of supraspan lists of words, first used in Hittmair-Delazer et al. (1994), shows a significantly weaker priority effect when the list is composed of proper names than when it is composed of common nouns matched to proper names for frequency, length, and phonological complexity. This effect widens with aging, and comes to an extreme with Alzheimer's disease (AD) (Semenza et al., 1996). In young people, a weakening of the priority effect for proper names results from exposure to high altitude (Pelamatti et al., 2003). Overall, these results suggest that the linking of proper names with their reference is more fragile than that of common nouns with their reference and may even require a larger amount of metabolic resources.

Anatomical Differences

The double dissociation shown by clinical cases between naming with common names and naming with proper names suggests that their processing happens in at least partially separate neural structures. Notwithstanding clear neurological dissociations, localizing proper name production with respect

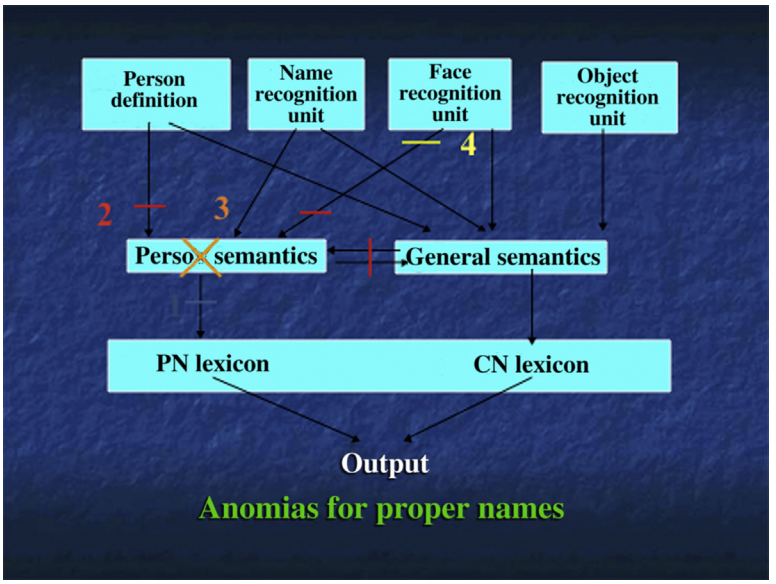

Figure 1 A model of name production distinguishing proper and common names. Proper names anomia derive from disconnections 1–4: (1) 'pure proper name anomia' (anomia in accessing an intact phonological lexicon from an intact individual semantic representations), (2) isolation of individual semantics, (3) loss/degradation of individual semantics, and (4) prosopanomia. In this picture, 'person semantics' stands for individual semantics. The box representing the (output) lexicon distinguishes proper and common names for the sake of clarity, but it does not imply that proper and common names occupy a different space. PN, proper name; CN, common name.

to common names proved to be a hard task. Findings obtained with the classic clinical anatomical correlation method have been further corroborated by lateralized tachistoscopic presentation, event-related potentials, and neuroimaging. The conclusions are not unequivocal, however.

Lesions found to cause selective proper name anomia or proper name sparing, always in the left hemisphere, are, in fact, manifold. They encompass the temporal lobe, the posterotemporal/occipital lobe, the left basal ganglia (including amygdala), and the left thalamus. Available reviews of these cases (whose details may be found in Semenza et al., 1995; Semenza, 1997, 2006, 2009; Yasuda et al., 2000) allow the following conclusions:

1. In proper name anomia, the left temporal lobe is often damaged;
2. Several proper name anomia cases are determined by lesions clearly outside the left temporal lobe (e.g., Semenza and Zettin's, 1988, patient);
3. The left temporal lobe is most often damaged in selective sparing of proper names;
4. At least two cases of selective sparing of proper names follow a lesion in the left temporal pole (LTP).

Findings in single case studies thus contradict a very influential view, originally forwarded by Damasio et al. (1996), in a study combining neuroimaging (positron emission tomography (PET)) with a group study, according to which proper name processing takes place in the LTP. In later studies (Damasio et al., 2004; Tranel, 2006), this view has been considerably moderated. But the same method seemed to confirm the role of the LTP to naming of famous landmarks (Tranel, 2006). The best data in favor of a pivotal role for the LTP come, however, only from group studies in face naming but when proper names are prompted by a definition, the results may be different. Zamarian et al. (2008) thus compared patients with left and right temporal lobe epilepsy on both picture naming and naming of famous people on definition. Left temporal lobe epileptics were worse than right temporal lobe epileptics on naming pictures, but no difference emerged in naming on definition.

The best evidence against the role of the LTP in naming comes, however, from the two cases reported by Pavão Martins and Farrajota (2007). In this study, the LTP was damaged in a patient with sparing of proper names, but undamaged in a second patient who suffered proper name anomia. These two patients had the same age and education and came from the same milieu. Importantly, as mentioned above, they were tested using the same material, thus this double dissociation cannot result from any unbalance between the tests tapping each of the two name categories. Unlike in the case of group studies, in each single case, the dissociation between proper names and common nouns is clearly demonstrated in a series of tasks, in particular those tapping the lexical retrieval level. Such dissociation is not guaranteed in the group study approach. A detailed discussion about the possible role of the LTP in proper name processing can be found in Semenza (2011). The LTP is indeed likely to play an important role (see below), but clinical evidence accumulated over the past years suggests that proper name processing is not confined to this brain location.

Surgical removal of the frontal part of the left uncinate fasciculus, the structure connecting the LTP to the orbitofrontal cortex, results in selective difficulties in naming people (Papagno et al., 2011). This finding confirms the coworking of the LTP and frontal areas in processing proper names. But anomia for proper names can occur without frontal lobe damage. The contribution of the frontal lobe to proper name processing may thus be limited to specific tasks and conditions.

Clinical investigations were paralleled by neuroimaging studies. Using PET, Damasio et al. (1996) found that two main areas of significant activation distinguished naming faces from a baseline task requiring judging the orientation of unfamiliar faces presented either upright or upside down: the LTP and the right temporal pole (RTP). While activation of the LTP confirmed the results of the clinical investigation, activation of the RTP was interpreted as reflecting the face recognition process. A first follow-up study (Grabowski et al., 2001) obtained the same results as far as naming faces was concerned as well as when naming was required of pictures of famous landmarks. This finding was interpreted as evidence that the LTP is critical for retrieving the names of unique entities such as persons and landmarks. A further PET investigation by Damasio et al. (2004), however, yielded similar but more complex results. In this study, the areas of activation in naming faces (as compared to animals and tools) were the RTP, the left anterior sector of the superior temporal sulcus, the ventromedial prefrontal cortices, and the anterior cingulate. The authors argued that regions related to the processing of emotions (i.e., the ventromedial prefrontal cortices and anterior cingulate cortex) would be required because unique familiar faces, for most subjects, are emotionally salient stimuli. The activation of the LTP, however, only reached a marginal level of significance.

An activation of the LTP in face naming (and other tasks involving proper names, e.g., matching proper name with a face on multiple choice), although most often in connection with other areas, was shown in many functional magnetic resonance imaging studies (Gorno Tempini et al., 1998; Leveroni et al., 2000; Tsukiura et al., 2001, 2003; Rotshtein et al., 2005; Pourtois et al., 2005). The most original contribution of this group of studies has been to show that, while the left anterior temporal region is important for the retrieval of people's names irrespective of their familiarity, the right superior temporal and bilateral prefrontal areas are crucial for associating newly learned faces and names (Tsukiura et al., 2003).

Converging evidence about the role of the anterior temporal lobe and connected areas in proper name processing come from electrophysiological studies. Proverbio et al. (2001) showed that the tacit retrieval of a proper name's phonological form elicits a strong activation of left anterior temporal and left centrofrontal areas; the same task showed greater involvement of occipitotemporal areas with common names. A study conducted via direct cortical stimulation (Giussani et al., 2009) also underlines the role of inferior frontal areas. In this study, face-naming specific sites were found to be more extended with respect to those shown by neuroimaging, including the superior, medial, and inferior frontal lobe gyri, as well as the anterior part of the superior and medial temporal gyri.

The nature of the task employed, and, possibly, several, so far not completely investigated, factors may be the reason of

different empirical findings. In conclusion, as Tranel (2006) argues, "depending on factors such as the exact nature of the stimulus, task demands, and the nature of a particular individual's abilities, the same task might engage different convergence zones across relatively comparable individuals." Thus, as Semenza (2009) proposes, various aspects of proper name retrieval may be sustained by different structures differently participating in a complex network, which remains, at the present time, poorly understood.

Two Systems: Why?

Why did different name processing systems develop in the evolution of the human brain? The philosophical and linguistic theories mentioned in the introduction make clear that proper and common nouns serve two fundamentally different functions: categorization and individuation. Semenza (2006, 2009) argued that, in the primitive world it might have become increasingly important, for the sake of survival, not only to be able to categorize surrounding entities but also to label categories in a precise fashion. Using a precise name, e.g., 'lions' or 'swamps,' for a kind of ferocious cat and for a kind of dangerous place, might have, for example, saved conspecifics' lives by allowing quicker and more efficient descriptions and warning. The ability to single out, and call by a proper name, individual entities ('F' is a friend, 'E' is an enemy, 'P' is a dangerous place) might have served similar warning aims. Distinct functions may allow quicker and more precise processing. Those humans who more efficiently and unambiguously distinguished categories and individual entities might have had a greater chance of being winners in the scheme of natural selection.

These considerations remind of findings in the animal literature. For example, chicks discriminate with the right hemisphere between known and new, while they assign category with the left hemisphere (Vallortigara et al., 2008). Two systems can thus be envisaged as stemming from evolution: one for top-down control, taking care of category assignment and routine behavior, and one for bottom-up control, working for individual recognition and socially relevant stimuli. When language was acquired by humans, this distinction was recruited for the lexicon.

In the particular case of naming humans, Semenza (2006, 2009) further argued, additional social purposes, also useful for reproduction, might have determined an advantage for the individual gifted in this respect. This idea was further developed and given some empirical support. Simmons and Martin (2009) suggested that the LTP is specifically involved in supporting socially relevant conceptual knowledge, in particular by mediating associations between names and person-related semantic information. Using both conjunction and region of interest analyses, Simmons et al. (2010) found person-selective but not domain-general responses in both the left and the right anterior temporal lobe when subjects encoded semantic facts. These structures are also engaged in social conceptual processing. The anterior temporal lobes would work in unison with a wider social cognition system to support learning facts about other people. The fact that the anterior temporal lobe is involved in processing social conceptual knowledge generally may be accommodated with the fact that it also acts in processing information about unique entities, in particular information about faces, and, to a lesser extent, landmarks.

The social aspect of proper name processing, as mentioned above, may also, interact with age. Tsukiura et al. (2011) showed that the anterior temporal lobe showed greater activity for names of faces than for job titles, but there was no age difference in the activity in this region; functional connectivity between hippocampal and anterior temporal lobe activities in both retrieval tasks was, however, higher for young than for older adults. Ischebeck et al. (2011) showed that activation of the LTP in encoding proper names looked significantly stronger in older subjects than in a younger group. Semenza (2011) thus argued that social interaction apparently contributes to convey proper names processing toward areas closer to those supporting social cognition. Consistently with Simmons and Martin's (2009) view, the LTP may accumulate both social conceptual knowledge and information about unique entities, important for social interaction acquired in life. Its close and privileged connections with the frontal lobe and with the wider social cognition system would thus be reinforced throughout life. Relying on this system may help the elderly in overcoming the higher demands of proper name processing relative to processing of common names.

Proper Names in Clinical Practice

Naming with proper names has been used in clinical practice. Anomia disproportionately affecting proper names is a rather common early finding in AD. Semenza et al. (2003a) tested suspect patients, who did not yet meet the commonly accepted criteria for a diagnosis of AD on proper and common name retrieval. Seven months later, a subgroup of these patients was found to have developed the disease. This subgroup was compared with the remaining patients on neuropsychological assessment batteries that were administered the first time. Only proper name retrieval was found to be more deficient in patients who later developed AD. Delazer et al. (2003) showed that proper name anomia in AD is due to semantic deficits, to a degradation of phonological representations, as well as to problems in accessing the phonological representation. It is only this last cause that contributes to the problems in proper name retrieval experienced by normal elderly people. Access to semantic knowledge about famous people has also been shown to be vulnerable in patients in stages of AD (among many others, see Ahmed et al., 2008; Beeson et al., 1997).

Specific investigations were conducted to see whether the difficulty in naming people in minimal cognitive impairment (MCI) reflects problems associated with lexical access or with semantic access (Juncos-Rabadán et al., 2011, 2013). Semantic access was not impaired. These studies concluded that failure in phonological access may be the first step in the progression from MCI to AD as regards naming difficulties. The tip-of-the-tongue phenomenon for proper names may effectively discriminate MCI patients from normal controls. MCI thus seems to involve a greater deficit in transmission from semantic to phonological representations than that produced by normal aging.

See also: Aphasia; Bilingual Aphasia; Lexical Consolidation; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Names and Naming; Prosopagnosia; Semantics; Syntax–Semantics Interface; Temporal Lobe.

Bibliography

Ahmed, S., Arnold, R., Thompson, S.A., Graham, K.S., Hodges, J.R., 2008. Naming of objects, faces and buildings in mild cognitive impairment. Cortex 44, 746–752.

Beeson, P.M., Holland, A.L., Murray, L.L., 1997. Naming famous people: an examination of tip-of-the-tongue phenomena in aphasia and Alzheimer's disease. Aphasiology 11, 323–336.

Cipolotti, L., 2000. Sparing of country and nationality names in a case of modality-specific oral output impairment: implications for theories of speech production. Cognitive Neuropsychology 17, 709–729.

Cipolotti, L., McNeill, J.E., Warrington, E.K., 1993. Spared written naming of proper nouns: a case report. Memory 1, 289–311.

Cohen, G., Burke, D.M., 1993. Memory for proper names: a review. Memory 1, 249–264.

Cohen, L., Bolgert, F., Timsit, S., Chermann, J.F., 1994. Anomia for proper names after left thalamic infarct. Journal of Neurology, Neurosurgery and Psychiatry 57, 1283–1284.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.

Damasio, H., Tranel, D., Grabowski, T.J., Adolphs, R., Damasio, A.R., 2004. Neural systems behind word and concept retrieval. Cognition 92, 179–229.

Delazer, M., Semenza, C., Reiner, M., Hofer, R., Benke, T., 2003. Anomia for people names in DAT – evidence for semantic and post-semantic impairments. Neuropsychologia 41, 1593–1598.

Giussani, C., Roux, F.E., Bello, L., Lauwers-Cances, V., Papagno, C., Gaini, S.M., Puel, M., Demonet, J.F., 2009. Who is who: areas of the brain associated with recognizing and naming famous faces. Journal of Neurosurgery 110, 289–299.

Gorno Tempini, M.L., Price, C.J., Josephs, O., Vandenberghe, R., Cappa, S.F., Kapur, N., Frackowiak, R.S.J., 1998. The neural system sustaining face and proper name processing. Brain 121, 2103–2118.

Grabowski, T.J., Damasio, H., Tranel, D., Ponto, L.L.B., Hichwa, R.D., Damasio, A., 2001. A role for the left temporal pole in the retrieval of words for unique entities. Human Brain Mapping 13, 199–212.

Hittmair-Delazer, M., Denes, G., Semenza, C., Mantovan, M.C., 1994. Anomia for people's names. Neuropsychologia 86, 345–358.

Ischebeck, A., Steindl, B., Reishofer, G., Koschutnig, K., Semenza, C., Ebner, F., Neuper, C., 2011. The effect of age on the memory for proper names. An fMRI-Study. In: Presented at the 29th European Workshop on Cognitive Neuropsychology, Bressanone, Italy.

Jackendoff, R., 1983. Semantics and Cognition. MIT Press, Cambridge, MA.

Juncos-Rabadán, O., Rodríguez, N., Facal, D., Cuba, J., Pereiro, A.X., 2011. Tip-of-the-tongue for proper names in mild cognitive impairment: semantic or, post-semantic impairments? Journal of Neurolinguistics 24, 636–651.

Juncos-Rabadán, O., Facal, D., Lojo-Seoane, C., Pereiro, A.X., 2013. Does tip-of-the-tongue for proper names discriminate amnestic mild cognitive impairment? International Psychogeriatrics 25 (4), 627–634.

Kripke, S., 1980. Naming and Necessity. Basil Blackwell, Oxford.

Leveroni, C.L., Seidenberg, M., Mayer, A.R., Mead, L.A., Binder, J.R., Rao, S.M., 2000. Neural systems underlying the recognition of familiar and newly learned faces. Journal of Neuroscience 20, 878–886.

Lucchelli, F., De Renzi, E., 1992. Proper name anomia. Cortex 28, 221–230.

Lyons, F., Hanley, J.R., Kay, J., 2002. Anomia for common names and geographical names with preserved retrieval of names of people: a semantic memory disorder. Cortex 38, 23–35.

McKenna, P., Warrington, E.K., 1978. Category-specific naming preservation: a single case study. Journal of Neurology, Neurosurgery and Psychiatry 41, 571–574.

McWeeny, K.H., Young, A., Hay, D.C., Ellis, A.W., 1987. Putting names to faces. British Journal of Psychology 78, 143–144.

Miceli, G., Capasso, R., Daniele, A., Esposito, T., Magarelli, M., Tomaiuolo, F., 2000. Selective deficit for people names following left temporal damage: an impairment of domain-specific conceptual knowledge. Cognitive Neuropsychology 17, 489–516.

Mill, J.S., 1893. A System of Logic. Longmans, Green and Co., London.

Otsuka, Y., Suzuki, K., Fujii, T., Miura, R., Endo, K., Kondo, H., Yamadori, A., 2005. Proper name anomia after left temporal subcortical hemorrhage. Cortex 41, 39–47.

Papagno, C., Miracapillo, C., Cesarotti, A., Romero Lauro, L., Castellano, A., Falini, C., Casaceli, C., Bello, L., 2011. What is the role of the uncinate fasciculus? Surgical removal and proper name retrieval. Brain 134, 405–414.

Pavão Martins, I.P., Farrajota, L., 2007. Proper and common names: a double dissociation. Neuropsychologia 45, 1744–1756.

Pelamatti, G., Pascotto, M., Semenza, C., 2003. Verbal free recall in high altitude: proper names vs common names. Cortex 39, 97–103.

Proverbio, A.M., Lilli, S., Semenza, C., Zani, A., 2001. ERP indexes of functional differences in brain activation during proper and common names retrieval. Neuropsychologia 39, 815–827.

Pourtois, G., Schwartz, S., Segheir, M.L., Lazeyras, F., Villaumier, P., 2005. View-independent coding of face identity in frontal cortices is modulated by familiarity: an event-related fMRI study. Neuroimage 24, 1214–1224.

Rotshtein, P., Henson, R.N.A., Treves, A., Driver, J., Dolan, R.J., 2005. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nature Neuroscience 8, 107–113.

Schmidt, D., Buchanan, L., Semenza, C., 2004. Sparing of proper names and dates in aphasia. Brain and Language 91, 9–10.

Searle, J., 1969. Speech Acts. An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

Searle, J., 2000. Speech Acts. Cambridge University Press, Cambridge.

Semenza, C., Zettin, M., 1988. Generating proper names: a case of selective inability. Cognitive Neuropsychology 5, 711–721.

Semenza, C., Zettin, M., 1989. Evidence from aphasia for proper names as pure referring expressions. Nature 342, 678–679.

Semenza, C., Sgaramella, T.M., 1993. Production of proper names: a clinical case study of the effects of phonemic cueing. Memory 1, 265–280.

Semenza, C., Mondini, S., Zettin, M., 1995. The anatomical basis of proper name processing: a critical review. Neurocase 1, 183–188.

Semenza, C., Nichelli, F., Gamboz, N., 1996. The primacy effect in the recall of lists of common and proper names: a study on young, elderly, and Alzheimer disease subjects. Brain and Language 5, 45–47.

Semenza, C., 1997. Proper-name-specific aphasias. In: Goodglass, H., Wingfield, A. (Eds.), Anomia: Neuroanatomical and Cognitive Correlates. Academic Press, San Diego.

Semenza, C., Zettin, M., Borgo, F., 1998. Names and identification: an access problem. Neurocase 4, 45–53.

Semenza, C., Borgo, F., Mondini, S., Pasini, M., Sgaramella, M.T., 2003a. Proper names in patients with early Alzheimer's disease. Neurocase 9, 64–70.

Semenza, C., Sartori, G., D'Andrea, J., 2003b. He can tell which master craftsman blew a Venetian vase, but he cannot name the Pope: a patient with a selective difficulty in naming faces. Neuroscience Letters 352, 73–75.

Semenza, C., 2006. Retrieval pathways for common and proper names. Cortex 42, 884–891.

Semenza, C., 2009. The neuropsychology of proper names. Mind and Language 24, 347–369.

Semenza, C., 2011. Naming with proper names: the left temporal pole theory. Behavioural Neurology 24, 277–284.

Simmons, W.K., Martin, A., 2009. The anterior temporal lobes and the functional architecture of semantic memory. Journal of the International Neuropsychology Society 15, 645–649.

Simmons, W.K., Reddish, M., Bellgowan, P.S.F., Martin, A., 2010. The selectivity and functional connectivity of the anterior temporal lobes. Cerebral Cortex 20, 813–825.

Tranel, D., 2006. Impaired naming of unique landmarks is associated with left temporal polar damage. Neuropsychology 20, 1–10.

Tsukiura, T., Fujii, R., Fukatsu, T., Otsuki, T., Okuda, J., Umetsu, A., Suzuki, K., Tabuchi, M., Yanagawa, I., Nagasaka, T., Kawashima, R., Fukuda, H., Takahashi, S., Yamadori, A., 2001. Neural basis of the retrieval of people's names: evidence from brain-damaged patients and fMRI. Journal of Cognitive Neuroscience 14, 922–937.

Tsukiura, T., Namiki, T., Fujii, T., Iijima, T., 2003. Time-dependent neural activations related to recognition of people's names in emotional and neutral face-name associative learning: an fMRI study. Neuroimage 20, 784–794.

Tsukiura, T., Sekiguchi, A., Yomogida, Y., Nakagawa, S., Shigemune, Y., Kambara, T., Akitsuki, T., Taki, Y., Kawashima, R., 2011. Effects of aging on hippocampal and anterior temporal activations during successful retrieval of memory for face-name associations. Journal of Cognitive Neurosciences 23, 200–213.

Vallortigara, G., Snyder, A., Kaplan, G., Bateson, P., Clayton, N.S., Rogers, L.J., 2008. Are animals autistic savants? PLoS Biology 6 (2), e42. http://dx.doi.org/10.1371/journal., pbio.0060042.

Van der Linden, M., Bredart, S., Schweich, M., 1995. Developmental disturbance of access to biographical information and people's names: a single case study. Journal of the International Neuropsychological Society 1, 589–595.

Yasuda, K., Nakamura, T., Beckman, B., 2000. Brain processing of proper names. Aphasiology 14, 1067–1089.

Zamarian, L., Semenza, C., Trinka, E., Walser, G., Bonatti, E., Benke, T., Delazer, M., 2008. Naming famous persons in patients with temporal lobe epilepsy. Epileptic Disorders 10, 373–391.

Property: Anthropological Aspects

Chris Hann, Max Planck Institute for Social Anthropology, Halle, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of property has been deployed in such specific ways in Western political philosophy, law, and economics that some anthropologists question its usefulness for comparative analysis. This article argues for its retention and the continued investigation of both 'real' and intangible property relations. Anthropological approaches challenge abstract models by exploring the sociocultural contexts in which property objects are held and exchanged, including the ways in which legal rights mesh with relations of power. In the era of neoliberal capitalism, anthropologists also need to reengage with world history and connect the specific claims of property to wider frameworks of citizenship.

Introduction

Property, a term used in everyday speech to refer to owned objects, is conventionally defined by scholars in terms of the rights held by persons (or groups of persons) in material and immaterial things, including other sentient beings and even other humans (notably when owned as slaves). It involves legal claims and has consequences for economic efficiency. Property is more than a technical institution, more than a relationship between persons and objects, and a great deal more than the right of a single owner to possess, use and alienate an object exclusively, though private property of this kind has been the standard of recent centuries. It expresses the fundamental norms and values of a society, such as equality, freedom and justice. Property thus has a long history in philosophy. The concept and the basic approaches to this institution in modern disciplines such as psychology, sociology, and economics have been shaped by particular intellectual traditions in Western Eurasia, and by the dominance of North Atlantic capitalist societies in recent centuries. For the proponents (or ideologists) of these societies, private property is central to liberty as well as to economic efficiency. For critics, such as Karl Marx, private ownership of the means of production is the basis of class inequality, which only collective ownership can overcome.

The power of this Marxist critique waned as social scientists studying industrial societies came to recognize that property was just one source of social inequality, and not necessarily the most important or enduring. Scholars have come to appreciate the dangers in applying either bourgeois or socialist models outside the regions and historical epochs for which they were developed. Some have come to question the value of the term 'property.' However, it seems safe to assure that all human societies regulate and reflect on their relations with different categories of things in their environments and transmit these understandings over time. In this sense, property is universal, a total social institution in the Maussian sense, 'embedded' in every social order (Hann, 1998). This article will explore what light the discipline of sociocultural anthropology has shed on property, past and present. Although this anthropology is a modern subject with a relatively short history, I shall organize my argument historically, initially by distinguishing between the evolutionist speculations of the nineteenth century and the ethnographic perspectives, which dominated in the twentieth. For much of this history, the principal focus of property research was land tenure ('real property'), but in recent decades, the study of intellectual property has gained in salience. I shall then turn to my own research in rural Eastern Europe to illustrate some of the central issues in more detail. With propertization invading ever new domains of social life and world economic institutions increasingly unstable in the first decades of the twenty-first century, I shall conclude by arguing that anthropologists are now well placed to renew their theoretical critique of a particular Western configuration.

Evolutionist Approaches

Property was a central concern of the legally trained scholars who played a key role in the emergence of anthropology (ethnology) as a distinctive field of inquiry in the nineteenth century. In England, the challenge was to make sense of how property was held in the most important of the colonies, India. In his *Ancient Law* (1861), Sir Henry Maine contributed the concept of a 'bundle of rights,' an invitation to disaggregation that has remained extremely useful to legal anthropologists and continues to be refined (von Benda-Beckmann et al., 2006). He emphasized that property rights were rights against other persons, i.e., they could only be understood in social context; at the same time, he was in no doubt that the rise of private property rights went hand in hand with the rise of civilization.

No less important was Lewis Henry Morgan's answer to the challenge of the New World, where the European colonists found societies more radically different from their own. The interpretation of property rights was central to these encounters. Morgan's sympathies were with the Iroquois and other native Americans, whose rights had been usurped by conquerors. In his magnum opus *Ancient Society* (1877), Morgan argued that the entire history of humanity could be traced as the history of property. He followed Jean-Jacques Rousseau in seeing private property as the original source of social inequality, rejecting the Lockean argument that property ownership could be justified in terms of the human labor invested.

The political conclusions of Morgan's work were drawn out by the early intellectuals of the socialist movement, above all

by Engels (1884). His evolutionary scheme was enshrined into the dogmas of historical materialism and eventually into Soviet anthropological science. Scholars in this school were expected to classify societies according to a unilineal scheme, in which the form of property holding was the key diagnostic. This was considered a rigorous science, an infinite improvement on Proudhon's facile claim that all property was theft. Ethnographers and historians were charged with documenting the degeneration from a pristine 'primitive communism,' in order to make credible the utopian ideology that promised a return to such egalitarianism in the communist future.

Not all evolutionist scholars accepted the postulates of the socialists. In Britain, James Frazer argued that *tapu* prohibitions in the Pacific and similar forms of 'superstition' all over the world had facilitated the evolution of private property, for him as for Maine a measure of human progress (Frazer, 1909). In Germany, Karl Bücher insisted that an 'individual search for food' preceded the emergence of any collectivist institutions (Spittler, 2008). Bücher was primarily an economic historian and not a fieldworker. His lectures in Leipzig in 1908 were attended by the young Bronisław Malinowski, who recognized that the debate about ancient forms of property could not be settled on the basis of the evidence available. Speculation about origins had to give way to scientific ethnography. This set the agenda for the rest of the twentieth century.

Ethnography and Diversification

Fieldworkers in the new paradigm undermined the simple binary between collective and individual, which had dominated so much earlier work. By looking more closely at what was going on in the major productive activities in the Trobriand Islands (Malinowski, 1935), or in the exchange of valuables (Malinowski, 1922; cf Mauss, 1924 [1990]), it was possible to recognize that both individuals and groups had rights, but that no party had a bundle of rights approximating private property in a highly commoditized economy based on the market and contract. Following this recognition, the critique of contemporary capitalist society faded into the background. Thus, the conceptual scheme applied by Max Gluckman (a political radical) in his celebrated study of Barotseland (1965) can be seen as a refinement of the framework developed by Malinowski (a political conservative) for the Trobriand Islands. Rather than resolving the contradictory conjectures of their nineteenth century predecessors, the anthropologists of the mid-twentieth century documented the contemporary complexity of land tenure. Their results were useful to the last generation of colonial administrators (Meek, 1946). Gluckman proposed that, while the rights to produce on specific plots were generally vested in particular households or persons, other rights were dispersed in a 'hierarchy of estates' of administration, typically ranging from a village headman to a paramount chief. None of these were owners in the Western sense – and friction was sure to arise if this resource, the basis of life, were to be treated as an alienable commodity. Around the same time, following a different intellectual route, Karl Polanyi argued that land (together with labor and money) was a 'fictitious commodity' (Polanyi, 1944).

The most important bridge between the new ethnography and the older concerns of the evolutionists was constructed by Jack Goody (1962, 1976). Working in the early 1950s in Northern Ghana, then the Gold Coast, Goody became interested in the way property of different types was transmitted, eventually distinguishing sub-Saharan African modes of 'devolution' from the more individualized forms of inheritance and 'heirship' that were characteristics of the plow economies of Europe and Asia (Eurasia). The dichotomy was to be seen in the principal form of marriage payment ('horizontal' bride-wealth in Africa versus 'vertical' dowry in Eurasia). It was also expressed in social structures (Eurasia was more hierarchical) and many aspects of culture (cuisine, flowers, etc.). Goody supplemented his own and other ethnographic studies with historical evidence and data from the Human Relations Area Files. However, this form of macrocomparison over an extended time frame attracted few followers. Goody's emphasis upon modes of property transmission paid insufficient attention to the relations of production for the taste of the neo-Marxist anthropologists who rose to prominence in the 1970s. He remained isolated thereafter, as the mainstream of the discipline came to emphasize the interpretations made possible through ethnographic encounters, and to shy away from anything, which smacked of evolutionism.

Theoretical trends in anthropology were reinforced by other factors, such as an apparent decline in the significance of inherited property in the era of welfare states, and the recognition that *access* and managerial *control* were as important as formal ownership in explaining the workings of mature capitalism. With the contraction of the agricultural sector, the study of real property lost its earlier overriding significance. Pioneering studies of hunter-gatherers that provided some ethnographic support for notions such as 'primitive communism' were duly noted, but they did not have the impact they would surely have had in the earlier epoch (Lee, 1979). A partial exception was the attention that was paid to contemporary challenges in managing 'common pool resources' (Bardhan and Ray, 2008). Whereas the ideologists of capitalism argued that a failure to institutionalize strong private property rights would lead inexorably to abuses and a 'tragedy of the commons,' anthropologists joined other scholars in showing that local communities were perfectly capable of working out communal solutions that were more efficient as well as more equitable than those generated by the market and private property. What was true for land also held in the marine sector, despite the greater difficulty in specifying sharp boundaries in this sector (Acheson, 2003).

These anthropological contributions tended to draw theoretically on the 'new institutional economics.' They were generally oriented toward solving problems of rational coordination in the present rather than pursuing long-term evolutionist arguments. These approaches have remained marginal to anthropological theorizing. Economists have remained implicitly or explicitly evolutionist, assuming that property regimes change whenever the benefits (in terms of allocative efficiency) are calculated to exceed the cost. Stephen Gudeman has vigorously challenged this reductionist perspective on property in the course of putting forward an

alternative general model, one which pays equal attention to the social bases of human communities or what he terms *mutuality* (2008: pp. 79–84). According to Gudeman, this mutuality, together with its pendant, the principle of the market, is present in all human societies, including our own.

If sociocultural anthropologists have cooperated with economists and political scientists to explore ways of regulating access to the material environment, they have consolidated their old alliances with lawyers and historians in other sectors. Much attention has been paid to intellectual property. If patents and trademarks emerged historically to serve a purpose in the real economy, the function of a prominent consumer brand in the twenty-first century often seems entirely detached from any such moorings. It is explicable only in another sphere, that of *symbolic value*. Yet precisely because of their importance in shaping personal and collective identities, these brands have massive consequences for productive systems and the distribution of income.

The study of intangible property rights is not new in anthropology. Earlier researchers had pointed to the importance of symbols and secret, sacred knowledge among remote, nonliterate peoples. Under neoliberal capitalism, the descendants of these peoples are likely to find themselves obliged to register such symbols as cultural heritage and thus the exclusive property of their group. Bowing to the hegemonic legal forms, no matter how alien to traditional norms, is often the only way for vulnerable groups to defend their interests. This 'propertization' of cultural knowledge has major implications for the holders of rights at both group and individual levels (Brown, 2003). Anthropologists are generally disposed to supporting the claims of indigenous groups. However, ethical issues may arise: should group members who have permanently moved away retain their rights and be able to bequeath them? What about long-term residents who do not belong to the indigenous group but may be equally poor and marginalized: are they to be excluded from title and the ensuing income streams?

Propertization here refers to the accelerating process of extending the logic of exclusive property rights to new kinds of 'thing,' ranging from body parts to the genome and 'indigenous scientific knowledge.' If Polanyi were alive today, he would surely add human body parts to his list of fictitious commodities (see B. Parry, 2004). He might also be tempted to include the scientific knowledge on which new medical technologies depend. New possibilities of surrogacy are enabling for some but they may be exploitative for others. Progress in genetic knowledge has generated contestation over ownership at multiple levels in the unique case of Iceland (Pálsson and Harðardóttir, 2002). The internet has become a key forum both for the implementation of new forms of property and for creative countercurrents in the spirit of a new commons. For example, copyright claims in the field of popular music, have been countered 'from below' through innovative use of the Net, just as the excluded have always subverted the formal property rights of the powerful throughout history (Hann, 2007). But this may be too sanguine an interpretation of current trends. The capitalist private property configuration continues to dominate and to spread relentlessly, legitimized by schools in economic thought which attach the highest importance to strong property rights. Flexible systems of customary landholding or marine tenure are routinely dismissed in favor of new fences or, where fences are impossible, by 'individual transferable quotas' of fish, based on property dogma rather than a substantive concern to manage resources efficiently and equitably (Minnegal and Dwyer, 2011).

In the light of these trends, one might expect to find property reassuming the central role it played for the anthropological vision of Morgan in the nineteenth century. This has not happened so far. The reasons may have to do with increased specialization, which makes it hard for contemporary anthropologists to investigate property as a total social fact. Thus anthropologists with an economics background have concentrated on questions of allocative efficiency (Gilman and Hunt, 1998), while colleagues with a legal background are more likely to emphasize the negotiation of claims, persuasion, and symbolic communication (Rose, 1994). Legal anthropologists have recently distinguished four 'layers' in formulating a comprehensive framework for the investigation of property: the wider cultural tradition, legal regulations, social relations, and finally 'property practices' (von Benda-Beckmann et al., 2006). However, this apparatus is explicitly advanced as a heuristic and not as a contribution to theory. The authors venture no claim as to which layers are more likely to initiate changes in a property system, or the circumstances in which layers may coexist despite evident disharmony between them. Among the great majority of anthropologists, who have no training in either law or economics, it is interesting to note that the term property seems to be fading. This trend is illustrated in the work of Marilyn Strathern, who is suspicious of property because of its Western baggage (1999). For her, Melanesia is so radically different from the West that it makes no sense to persist with the established terms. The Melanesian person is a 'dividual' rather than a rights-holding Western individual. Others have argued in a similar vein that not even 'things' or 'relations' are valid terms cross-culturally, and suggested that 'value' be a more appropriate focus of anthropological research (Verdery and Humphrey, 2004; Verdery, 2003). Alternatives such as ownership and appropriation are held to be less contaminated by Western history (Strang and Busse, 2011).

(Post-)Socialist Transformations of 'Real Property'

At this point let me step back from these general developments and illustrate how one anthropologist found himself obliged to address issues of property in the course of longitudinal studies of land tenure in rural Eastern Europe during and after socialism. The collectivization of land in the former Soviet Union and allied states was a world-historical event that is frequently invoked by the proponents of private property ideologies. Soviet law did not need the category 'real estate,' since it declared that land was owned by the state and not available for alienation (though it could be transferred between socialist enterprises). In addition to the brutality and loss of life occasioned by collectivization in the USSR of the 1930s, critics have alleged that the new institutions of cooperative (*kolkhoz*) and state (*sovkhoz*) farms exemplified the inefficiency of collective forms of property holding. Only

on their own personal (household) plots did villagers have an incentive to work efficiently because they themselves would reap the benefits, either in terms of subsistence goods, or through sales on local markets, and frequently both. According to the stereotype, all things outside this domestic sphere were considered fair game for pilfering and no one had any incentives to use them efficiently; this was obviously detrimental to overall economic performance and led eventually to the collapse of the entire bloc.

In my fieldwork in post-collectivization Hungary, I often heard similar negative opinions expressed by villagers; but my analysis led me to qualify these antisocialist clichés (Hann, 1980). The community of Tázlár is located in a region of the Great Hungarian Plain, which was resettled in the late nineteenth century. Capitalism was developing in this era, but there were not enough industrial jobs for an expanding rural population. As a result, regions of poor natural endowment were colonized as a kind of 'internal frontier.' Immigrants purchased parcels of land as private property. Most of them built their cottages on their plot; a nuclear center was not consolidated until the socialist era. Property boundaries were unambiguous (aided by the late Habsburg cadastral survey) and there were no complex 'hierarchies of estates.' In a region that remained economically backward, the acreage one owned was the prime determinant of both economic capacity and social status.

The significance of real property was temporarily inverted in the Stalinist years, when so-called *kuláks* (rich peasants) became the targets of political campaigns and leveling was imposed by force. Mass collectivization was eventually implemented around 1960. Most villagers signed up for one of the new cooperatives with great reluctance. Those with above-average holdings were devastated to find themselves in an association on a par with fellow villagers who had previously possessed little or nothing. However, not all property rights were eliminated. After joining one of the new cooperatives, villagers retained title to their plots. These rights continued to be transmitted through inheritance, though few took the trouble to visit the necessary offices to update the registers. These legal niceties seemed irrelevant in the face of Communist Party power. Tázlár diverged from the standard *kolkhoz* pattern in the extent to which villagers were able to continue farming their own plots as a household unit. This was a pragmatic concession due to the scattered settlement pattern. However, here too there was a commitment to amalgamating fields eventually, to facilitate more rational, mechanized cultivation. Population decline as some villagers left permanently for new jobs in socialist towns facilitated such *remembrement*. Where plots were appropriated, their owners were always offered alternative fields elsewhere. In short, the old property regime was not suppressed completely, but private transactions of land were precluded. Power was in the hands of the cooperative, the leaders of which were in practice compelled to follow policies laid down by authorities external to the village, controlled ultimately by the Communist Party. We could theorize this in Gluckman's terms as a socialist 'hierarchy of estates' (Humphrey, 1983).

With the introduction of the 'New Economic Mechanism' in 1968, the socialist cooperatives were encouraged to take further steps to support the household farming of their members. The mechanized sector under collective cultivation specialized in the production of grain and fodder; in Tázlár it benefited from state subsidies paid to areas of poor natural endowment. Much of the output was sold to households to support the labor-intensive raising of animals. This was a productive symbiosis. In 1974, the separate cooperatives merged to form a single association. Despite the emergence of a stratum of professional leaders, policies in the 1980s continued to support the household sector. Presocialist property inequalities in terms of land ownership ceased to be of much significance, but new ones appeared as successful households built bathrooms and two-storey houses and purchased cars. Urban intellectuals sniffed at what they considered to be a frenzy of accumulation in the last decades of socialism (I did too). Yet for many, this symbiosis between household and cooperative offered an escape from poverty, and the unequal rewards could be justified in terms of unequal labor inputs. They did not depend on unequal access to the means of production. Arguably, this was a pragmatic balance of the main 'forms of integration' identified by Karl Polanyi, a native of Hungary (Polanyi, 1944). The demotion of property respected the above-noted dictum of his substantivist school that land was a 'fictitious commodity.' At the same time, the impressive economic expansion of the last decades of socialism refuted the dogmas of bourgeois property ideologies: property rights in the field of production were decidedly fuzzy, yet output increased impressively (just as it did later in China following that country's opening to the market). Villagers worked out pragmatic ways to combine large-scale and small-scale production, and found it perfectly rational to work long hours to accumulate private property in the realm of consumption. In this way, the socialist hierarchy of estates of administration was softened and became responsive to local preferences. Yet this did not mean that villagers warmed to their new institutions and abandoned their historically formed attachment to an ideal of full private ownership.

I contrasted this Hungarian situation with what I found in a follow-up study of a community in rural Poland (it was located in the center of the former province of Galicia, where Malinowski was born and grew up). For political reasons, following the turning point of 1956, the socialist government in Poland was unable to implement collectivization. The old pattern of family farming persisted. However, with the state committed to socialist goals in the industrial sector, modernization of the rural sector was hardly a priority. As a result, farming technologies remained backward, productivity levels low, and the opportunities for private accumulation very limited. Polish farmers had somewhat more secure rights of possession than their Hungarian counterparts, but they too were not free to expand their holdings, since the land market was strictly controlled. In short, the non-collectivized farmers in Poland were a lot worse off than the villagers who benefited from the pragmatic compromises struck in post-collectivization Hungary (Hann, 1985).

The collapse of socialism was immediately accompanied by a radical change in the property system, which affected all sectors of the economy. The whole world is familiar with the

rapid emergence of a new class of Russian billionaires in the wake of privatization in the energy sector. Anthropologists have documented the social and cultural consequences of the 'unmaking' of the Soviet economy more generally (Humphrey, 2002). Legal scholars have drawn attention to the enormous difficulty of creating effective private 'property bundles' from the mass of former state or cooperative assets. For example, many store windows in Moscow in the 1990s remained empty because of a 'tragedy of the anticommons' (Heller, 1998). This problem, the mirror of the more familiar tragedy of the commons, was one of underutilization rather than excessive exploitation of a natural resource. The suboptimal use of urban real estate resulted from the dispersal of property rights to multiple stakeholders (generally persons of influence under the old regime). More concentration of core rights in a single owner would, according to Heller, have led to economically more efficient outcomes.

Anthropologists have tended to base their fieldwork in the countryside, where the property bundles are different. In most of the former Soviet Union, as in Eastern Germany, land is still farmed in large units based on the old *kolkoz* and *sovkoz* (Hann et al., 2003). Legal privatization has not led to a structure of family farmers. Outcomes are shaped primarily by local power relations and many new 'owners' do not even bother to identify the boundaries of their holdings (Yalcin-Heckmann, 2010). Elsewhere, in many parts of Eastern Europe, plots were returned to their former owners, respecting original boundaries. This led to enormous social problems as well as economic irrationalities (Hann, 1993; Verdery, 2003). Hungary opted for a 'compensation' path that fell short of guaranteeing restitution of original holdings, in order to avoid losing the economic benefits of amalgamated fields. Even so, possession of title deeds suddenly reassumed the significance it had lost under socialism. Many in Tázlár felt emotionally compelled to claim the plots farmed by their forebears, even when this seemed irrational in purely economic terms (given poor soils, the withdrawal of state subsidies, the collapse of the supportive cooperative and of product markets in other ex-socialist countries). In short, an emphasis on property rights displaced the late socialist emphasis on expanding output; production has declined considerably since the halcyon days of the cooperative; large areas of Tázlár are no longer cultivated at all.

After Hungary joined the EU, a minority of the new private farmers has been able to benefit from this new source of subsidy. The EU supports specific branches of production and it also pays owners to maintain rural environments, which may entail leaving land fallow. Some high value crops, notably vineyards, are nowadays tended by their private owners with a dedication that is hard to imagine on the part of a socialist brigade leader (even if this leader had material incentives to improve his stewardship). Yet many private owners have difficulty disposing of their wine and bemoan the lack of a cooperative to assist them in standardizing a product to facilitate marketing. Some who were most critical of the pragmatic form of cooperative, which prevailed in this region of Hungary before 1990 would now like to reinvent it. They recognize that their rights as private owners (a powerful ideology, which prevented them from positively embracing socialist hierarchy) are now constrained by new 'hierarchies of estates' reaching beyond Budapest to Brussels. This 'tragedy of the privates' (Hann, 2006) differs from the above-mentioned 'tragedy of the anticommons.' According to the property lawyer Heller's analysis of the latter, the solution lies in more complete bundles of private property rights. However, in the case of rural real estate, I argue for the superiority of the socialist system, which brought land into *de facto* communal ownership, managed by a collective institution that was gradually brought under local control. The property lawyer and the anthropologist can agree that it is insufficient to insist on the 'clarity' of property rights and to blame the failures of 'transition' on factors such as corruption and inadequate property registers and legal procedures. But for Heller, the problem is to specify the optimal content of the property bundle. He takes it for granted that a system based on private, exclusive rights will be an economic improvement on Soviet law, in which the category of real estate was lacking. The anthropologist questions this initial bias of the Western lawyer, just as Gudeman (2008) questions the reductionism of the Western economists, and instead probes more deeply into the social consequences of the destruction of socialist institutions at the local level.

This is thus a complex situation, arguably as complex as land tenure among the Trobriand Islanders, studied by Malinowski. Anthropologists have struggled to place these post-socialist transformations in the context of world history; the history of their own discipline in the twentieth century has not been much help in this. According to the theories of the anthropologists in the wake of Malinowski and Gluckman, the 'individual versus collective' dichotomy was a misconceived Western imposition on preindustrial land tenure systems. Actually, existing property systems took the form of hierarchies of rights, combining the personal, the familial (household), the village community, etc. But tell this to an East European villager whose collective farm is broken up as fences are (re)built along the old boundaries, and he might well respond: 'claptrap.' The fact is that these fields have been privatized. The contrast in Tázlár is even clearer to the villagers than the imposition of collectivization in an earlier generation. Here, as elsewhere, for a small minority private ownership of the means of production has again become lucrative, while the majority has paid a high price. When pursuing these questions in Tázlár more than a decade after the end of socialism, many villagers answered with reference to wider frameworks, including their *entitlements* as citizens. For many, the formal restitution of some rights to land had brought no benefit at all, especially if, as a result of the demise of the cooperative, the household no longer had any source of regular income. Often, a minimal state pension was now the key to survival, analogous to the ownership of a parcel of land for their ancestors (Hann, 2006).

As with the 'tragedy of the commons' thesis discussed above, collective property does not have to mean fuzziness and confusion. In Tázlár in the 1970s, everyone knew the boundaries of the new collective fields, and respected the power of the cooperative to redistribute land that had been abandoned by urban migrants. Some of the collective activities were not well organized and political factors interfered from time to time – the cooperatives had many imperfections. But it needs to be recalled that socialist property systems were designed to be different from the Western configuration, and not to simulate

the economic outcome of a system which prioritized private ownership. In practice, few anthropologists have been as charitable toward this socialist alternative as they have toward different ontologies in remote Melanesia. There is no evidence that centrally planned economies had higher levels of corruption than their Western capitalist equivalents, with their well-oiled machinery of bribes and kickbacks. Productivity per unit of labor may have been inferior, but the other side of this coin was the absence of unemployment and high levels of mutuality at the workplace and in the community. Social inequalities increased considerably in the later socialist decades, as more and more 'things' could be owned as private property and passed on to the next generation. But so long as the great bulk of the means of production and financial institutions were owned and controlled collectively, inequalities were held in check.

Conclusion

There was an unfortunate disjuncture between the trends in anthropological theorizing, as summarized in section III, and the trend which set in during the last two decades of the twentieth century in global political economy, accentuated by the end of the Cold War and the extension of capitalism into the former Soviet bloc. In anthropology, the neo-Marxist energies were largely spent when Ronald Reagan and Margaret Thatcher inaugurated the era of neoliberalism, with policies that emphasized the privatization of industries, which had come under some form of collective or state ownership. Social inequalities increased dramatically, in part due to reductions in inheritance taxes. Property was transparently at the heart of social clearages. Yet anthropologists have not seized the opportunity. Some, in partnership with legal scholars, have investigated new forms of intangible propertization, but by and large anthropologists have left the big picture of unequal distribution to other disciplines.

I have argued that they have nonetheless made a useful contribution, above all by insisting on the historicity of Western property theories, and on the dangers in deploying (even within Western societies) simple dichotomies such as that between collective and individual. The case for complexity can readily be strengthened by extending the notion of property beyond its familiar referents, material, and immaterial. It is possible, for example, to justify a range of 'natural' inequalities if we take the view that celebrities such as Clint Eastwood *own* their remarkable bodies and are thus entitled to the income streams that being Clint Eastwood generates. The resulting inequalities can be countered by laws that provide all citizens with a range of entitlements, e.g., in the event of their falling sick or being unable to find employment. For many, to extend the term property to such general entitlements is to go too far beyond its established use. Yet ordinary people are likely to make exactly this sort of comparison when comparing their well-being under different regimes. At the very least, a right that is not vested in all citizens but in particular families, such as the right of a man to pass on his job at the Bhilai steelworks to his son (Parry, in press), should be considered as a property right, because it functions in the same way that rights to inherit real property used to function before the coming of industry.

In conclusion, we must remember that ideas about property have been of considerable significance in many intellectual traditions for many thousands of years. Sociocultural anthropologists began by engaging with European speculations about origins, but they have not fulfilled the early *historical* aspiration to take property institutions as a prism for understanding human social evolution. They have done rather better when it comes to documenting the diversity of property institutions and *cultural* variation in the very definition of its components (person, thing, etc.) as well as in the configuration. But this can be pushed too far: to emphasize unique local ontologies and the sheer inapplicability of any Western categories can be disabling. The modern anthropologists' success, based on micro, ethnographic studies, may have hindered their ability to mount a *sociopolitical* critique of property in terms of the stark, brutal inequalities it continues to sustain. In this sense, we need to recover the simplicity of property. Taken together, the works of the last century and a half have made substantial contributions, but it is time to develop a new synthesis that would draw the historical, the cultural, and the sociopolitical dimensions of property together.

See also: Capitalism: The History of the Concept; Capitalism; Citizenship, Anthropology of; Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Citizenship: Political; Property Ownership; Property Rights and Urban Planning.

Bibliography

Acheson, James, 2003. Capturing the Commons: Devising Institutions to Manage the Maine Lobster Industry. University Press of New England, Hanover.

Bardhan, Pranab, Ray, Isha, 2008. Economists, anthropologists, and the contested commons. In: Bardhan, Pranab, Ray, Isha (Eds.), The Contested Commons: Conversations between Economists and Anthropologists. Blackwell, Oxford, pp. 1–24.

Benda-Beckmann, Franz von, Benda-Beckmann, Keebet von, Wiber, Melanie, 2006. The properties of property. In: Benda-Beckmann, Franz von, Benda-Beckmann, Keebet von, Wiber, Melanie (Eds.), Changing Properties of Property. Berghahn, New York, pp. 1–40.

Brown, Michael, 2003. Who Owns Native Culture? Harvard University Press, Cambridge, Mass.

Engels, Frederick, 1884 [1972]. The Origin of the Family, Private Property, and the State. Pathfinder Press, New York.

Frazer, James G., 1909. Psyche's Task: A Discourse Concerning the Influence of Superstition on the Growth of Institutions. Macmillan, London.

Gilman, Antonio, Hunt, Robert C. (Eds.), 1998. Property in Economic Context. University Press of America, Lanham, MD.

Gluckman, Max, 1965. The Ideas in Barotse Jurisprudence. Yale University Press, New Haven.

Goody, Jack, 1962. Death, Property and the Ancestors. Stanford University Press, Stanford.

Goody, Jack, 1976. Production and Reproduction. Cambridge University Press, Cambridge.

Gudeman, Stephen, 2008. Economy's Tension. Berghahn, New York.

Hann, Chris, 1980. Tázlár: A Village in Hungary. Cambridge University Press, Cambridge.

Hann, Chris, 1985. A Village without Solidarity: Polish Peasants in Years of Crisis. Yale University Press, New Haven.

Hann, Chris, 1993. From production to property: decollectivization and the family-land relationship in contemporary Hungary. Man 28 (3), 299–320.

Hann, Chris, 1998. Introduction: the embeddedness of property. In: Hann, Chris M. (Ed.), Property Relations. Renewing the Anthropological Tradition. Cambridge University Press, Cambridge, pp. 1–47.

Hann, Chris, 2006. Not the Horse We Wanted: Postsocialism, Neoliberalism. Eurasia, LIT, Berlin.
Hann, Chris, 2007. A new double movement? Anthropological perspectives on property in the age of neoliberalism. Socio-Economic Review 5 (2), 287–318.
Hann, Chris, the "Property Relations" Group, 2003. The Postsocialist Agrarian Question: Property Relations and the Rural Condition. LIT, Münster.
Heller, Michael E., 1998. The tragedy of the anticommons: property in the transition from Marx to markets. Harvard Law Review 3 (3), 622–688.
Humphrey, Caroline, 1983. Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm. Cambridge University Press, Cambridge.
Humphrey, Caroline, 2002. The Unmaking of Soviet Life: Everyday Economies after Socialism. Cornell University Press, Ithaca.
Humphrey, Caroline, Verdery, Katherine, 2004. Introduction: raising questions about property. In: Verdery, Katherine, Humphrey, Caroline (Eds.), Property in Question: Value Transformation in the Global Economy. Berg, Oxford, pp. 1–25.
Lee, Richard B., 1979. The !Kung San: Men, Women and Work in a Foraging Society. Cambridge University Press, Cambridge.
Maine, Henry, 1861. Ancient Law. J. Murray, London.
Malinowski, Bronislaw, 1922. Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagos of Melanesian New Guinea. G. Routledge and Sons, London.
Malinowski, Bronislaw, 1935. Coral Gardens and Their Magic: A Study of the Methods of Tilling the Soil and of Agricultural Rites in the Trobriand Islands. Allen & Unwin, London.
Mauss, Marcel, 1990 [1924]. The Gift: The Form and Reason for Exchange in Archaic Societies. Routledge, London.
Meek, Charles K. (Ed.), 1946. Land Law and Custom in the Colonies. Oxford University Press, Oxford.
Minnegal, Monica, Dwyer, Peter, 2011. Appropriating fish, appropriating fishermen: tradable permits, natural resources and uncertainty. In: Strang, Veronica, Busse, Mark (Eds.), Ownership and Appropriation. Berg, Oxford, pp. 197–215.
Morgan, Lewis H., 1877. Ancient Society. Henry Holt, New York.
Pálsson, Gísli, HarÐardóttir, Kristín E., 2002. For whom the cell tolls: debates about biomedicine. Current Anthropology 43 (2), 271–301.
Parry, Bronwyn, 2004. Bodily transactions: regulating a new space of flows in "bio-information". In: Verdery, Katherine, Humphrey, Caroline (Eds.), Property in Question: Value Transformation in the Global Economy. Berg, Oxford, pp. 29–48.
Parry, Jonathan, in press. The embourgeoisement of a 'proletarian vanguard'. In: Jhodka, Surinder (Ed.) Interrogating India's Modernity. Democracy, Identity and Citizenship. Oxford University Press, Delhi.
Polanyi, Karl, 1944. The Great Transformation. Rinehart, New York.
Rose, Carol, 1994. Property and Persuasion. Westview, Boulder, CO.
Spittler, Gerd, 2008. Founders of the Anthropology of Work. LIT, Berlin.
Strang, Veronica, Busse, Mark (Eds.), 2011. Ownership and Appropriation. Berg, Oxford.
Strathern, Marilyn, 1999. Property, Substance and Effect: Anthropological Essays on Persons and Things. Athlone, London.
Verdery, Katherine, 2003. The Vanishing Hectare: Property and Value in Postsocialist Transylvania. Cornell University Press, Ithaca.
Yalcin-Heckmann, Lale, 2010. The Return of Private Property: Rural Life after Agrarian Reform in the Republic of Azerbaijan. LIT, Berlin.

Property Crime: Victimization and Trends

Jan Van Dijk, Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The best sources of information on property crimes are victimization surveys. Using the results of the International Crime Victims Survey (ICVS) conducted once or more in over 80 countries from all world regions, an overview is given on rates of victimization by burglary, other personal theft, vehicle-related theft, and consumer fraud, including e-fraud. Distributions of victimization are explained as resulting from differential opportunities for crimes across regions. Results of the ICVS provide evidence of universal drops in property crimes in Western countries over the last 10 years. These drops are interpreted as resulting from improved security against property crimes as illustrated by the proven impact of security measures on the falls in household burglaries and car thefts.

Introduction

Property crime is and always has been one of the most pervasive categories of crime. It can loosely be defined as crimes directed at the illegal appropriation of other people's material goods. It includes different types of theft such as vehicle theft, pickpocketing and shoplifting, as well as various forms of fraud, including e-fraud and burglaries in households or commercial premises. In this article, we will discuss the most common types of property theft but exclude property crimes involving the use of violence against persons (robberies) and vandalism (destruction of property).

Figures of crime recorded by the police depend on the number of crimes reported by victims to the police. As known, a large proportion of crimes are not reported to the police or not recorded by them (generally known as the 'dark numbers' of crime). This holds also true for a large proportion of property crimes. In this article, the volume and trends in property crimes will therefore be described using results of victimization surveys among the public rather than official figures of crimes recorded by the police. Victimization surveys take a sample of the population (usually householders) to ask them if in the recent past they have been a victim of crime, whether or not they reported what happened to the police. Several studies have shown that survey-based estimates can give a more reliable picture of the volume and trends of crime than police figures. The most important early victimization survey is the National Crime Victimization Survey (NCVS) in the USA. Victimization surveys have since been conducted in most other Western countries and many Latin American countries as well (Aebi and Linde, 2012). In 1988, a standardized victimization survey was launched, called the International Crime Victims Survey (ICVS). This survey has been repeated in five major sweeps in a selection of over 80 countries selected from all world regions (Van Dijk et al., 2007; Van Dijk, 2008). Its results are acknowledged as the most authoritative source of comparative information on levels of crime (Lynch, 2006). The ICVS results will be used in this article to describe the volume and trends in victimization by main types of property crime across world regions. We will first present data on the prevalence of victimization by burglaries, other thefts, consumer frauds including e-fraud, vehicle-related crimes, and crimes against businesses. Next we will discuss trends in property crimes over time. Finally we will discuss the key determinants of the distribution and trends of these crimes.

The Prevalence of Burglary Victimization

Burglary in a dwelling is a universally frequently occurring type of property crime with a high impact on the victims (Mawby, 2001). The consequences of burglary in terms of monetary value may vary widely in different contexts. While in the industrialized world burglars frequently steal objects of a high value, such as audio/video equipment, jewelry, or computers, burglary in developing countries is often aimed at stealing cattle, food, house appliances, linen, or cutlery. Regardless of the economic losses, victims regard burglaries as very serious since it is a violation of their privacy in the domestic sphere.

The ICVS questionnaire asks respondents whether anyone did actually get into their house or flat without permission and steal or try to steal something. **Figure 1** shows the regional, 1-year victimization rates for domestic burglary in urban areas.

ICVS results show that on average one in four households in urban areas is visited by a burglar at least once annually. Households in urban Africa, Latin America, and Oceania (Australia, New Zealand) are most at risk. The European, Asian, and North American regions showed risks below the global average. The high prevalence in Africa is confirmed by regional studies (Afrobarometer, 2004; UNODC, 2005). Police-recorded burglary as recorded by the UN Crime Surveys, was highest in Australia, while the ICVS results indicate that the most affected regions were, as said, Africa and Latin America, which in turn showed the lowest rates in police data. In our view, rates of police-recorded burglaries give no useful information on the actual prevalence of burglary across regions or countries.

Other Personal Thefts

The ICVS contains a question on experiences with personal theft such as theft of purses, wallets, clothing, or equipment. Victims of such thefts are asked whether they were carrying

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45075-0

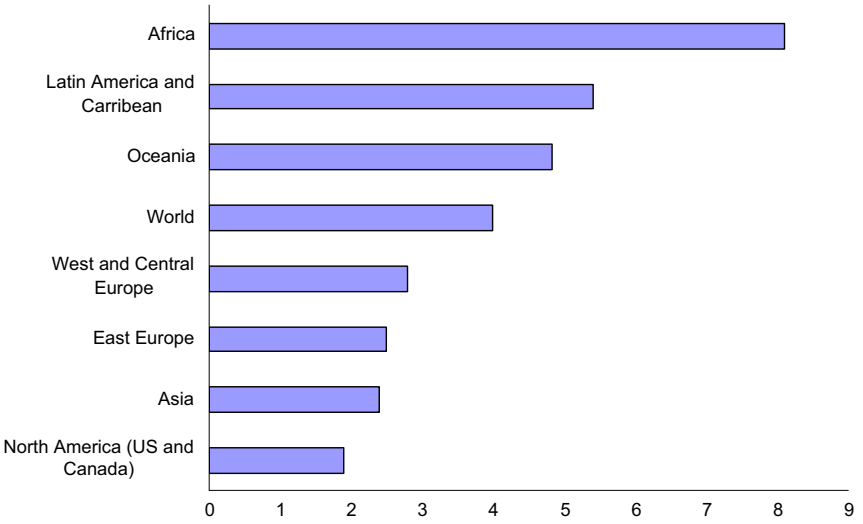

Figure 1 Percentages of public in urban areas victimized by household burglary during the past 12 months, by world region. Source: ICVS, 1996–2005; Van Dijk, J.J.M., 2008. The World of Crime: Breaking the Silence on Problems of Security, Justice, and Development across the World. SAGE Publications, Los Angeles.

what was stolen. Those answering in the affirmative are classified as victims of pickpocketing. Figure 2 shows the regional rates.

The distribution of regional rates of pickpocketing is somewhat exceptional. The crime of pickpocketing seems comparatively rare in North America and Australia/New Zealand. The countries of Eastern Europe find themselves in an unusual high place, at the same level as Latin America, Asia, and Africa. Pickpocketing is three times more common in Eastern Europe than in Western Europe with rates in Central Europe lying somewhere in between. This high position of Eastern Europe differs from the more modest positions of the region on the world rankings for other types of property crimes. A possible explanation for the special distribution of pickpocketing is differences in the common modes of mass transportation. In regions and countries that rely heavily on public transportation such as trains and city trams, there are more opportunities for pickpocketing than in countries where cars are the dominant form of transportation.

Consumer Fraud

The ICVS asks whether the respondent had been the victim of a consumer fraud over the past year, focusing on cheating in commerce or the services industry. The most common frauds occurred in shops or during construction or repair work or in a garage. Consumer fraud shows higher average rates and more variation across regions than any other type of crime. Thirty-eight percent of the respondents from Eastern Europe actually experienced a consumer fraud of this type, followed by 30% in Asia and 26% in Africa. In Australia and New Zealand consumer fraud is almost unknown. The comparatively high rates in Eastern Europe have been interpreted as a side effect of

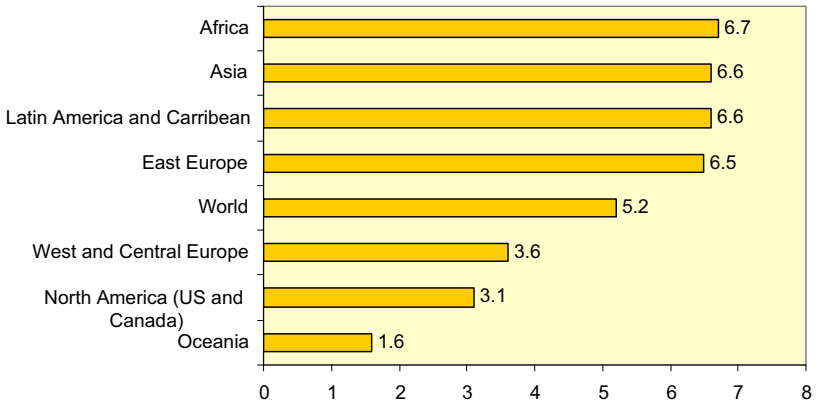

Figure 2 Percentages of public in urban areas victimized by pickpocketing during the past 12 months, by region. Source: ICVS, 1996–2000; Van Dijk, J.J.M., 2008. The World of Crime: Breaking the Silence on Problems of Security, Justice, and Development across the World. SAGE Publications, Los Angeles.

the transition from a planned economy – coupled by a huge black market – toward a regulated market economy (Zvekic, 1998). The extent of consumer fraud appears to be higher in newly emerging market economies where consumer protection is underdeveloped and regulation of markets is weaker.

Victims of consumer fraud were asked in which type of situation the incident had taken place, including while shopping on the Internet. At country level, 45% of victims said the fraud had taken place in a shop. Eleven percent mentioned either building or construction work or a garage. Nine percent mentioned shopping online. This implies that 1% of the national respondents of the participating countries have been victimized by a fraud on Internet. Among city inhabitants the victimization rate was 1.5%. The 2003 NCVS of the USA reported a national victimization rate of credit card theft of 2.4%. Recent surveys in the United Kingdom showed rates went up from 3.7% in 2009 to 6.4% in 2009 (Walker et al., 2009). Internet-based frauds have now apparently reached prevalence rates that are similar to those of common crimes as car theft or pickpocketing.

Car Theft and Joyriding

While a number of cars are stolen for the purpose of 'joyriding' or to commit other crimes, others are either dismantled or sold on domestic or transnational markets for stolen vehicles. In Australia and North America around 80% of the stolen cars are eventually recovered. In these regions most cases of car theft comprise 'joyriding.' In most European and Asian countries only on average around half of stolen cars are recovered. In many cases cars recovered after being stolen are seriously damaged. In the judgment of victims, both car theft and joyriding are judged to be among the most serious forms of conventional crimes covered by the ICVS (Van Dijk, 1999). Apparently many owners are much attached to their cars and resent loosing them or finding them damaged due to a criminal act.

Figure 3 shows the regional rates of victimization by car theft, including joyriding.

Regional rates for car theft show a distribution across regions that is almost the mirror image of victimization by burglary. Rates in North America and Oceania are the highest. Lowest rates are found in Africa, Eastern Europe, and Asia. Obviously the level of car theft as of other car-related crimes is strongly dependent on national rates of car ownership. Where more households own cars, the prevalence of such crimes is higher. Car crimes are a prime example of how affluence-related opportunities of crime drive up the volume of property crime. To give a better picture of the risks for owners, rates are often calculated of the percentages of owners victimized by these crimes in the course of a year. Car owners' victimization rates for car theft are, of course, higher than the general population's victimization rates for car theft. The differences are most marked in regions where car ownership is comparatively low (Africa and Latin America).

Other Vehicle Theft

In several European and Asian cities, bicycles are among the most important modes of transportation. Bicycle theft provides yet another example of how the levels and nature of volume crime are shaped by the dominant modes of transportation in a country. Bicycle theft, as car theft, appears to be clearly driven by ownership levels, with the Netherlands in top position. Even in otherwise low-crime countries such as China and Japan, the general availability of bicycles generates high rates of bicycle theft as 'crimes of expediency.' The analysis reveals a clear *exponential* relationship between bicycle ownership and theft. If bicycle theft surpasses critical levels, a chain reaction seems to be set into motion by victims committing themselves to stealing – or to knowingly buying from fences – to replace the stolen bicycle. This phenomenon of 'compensatory stealing' of

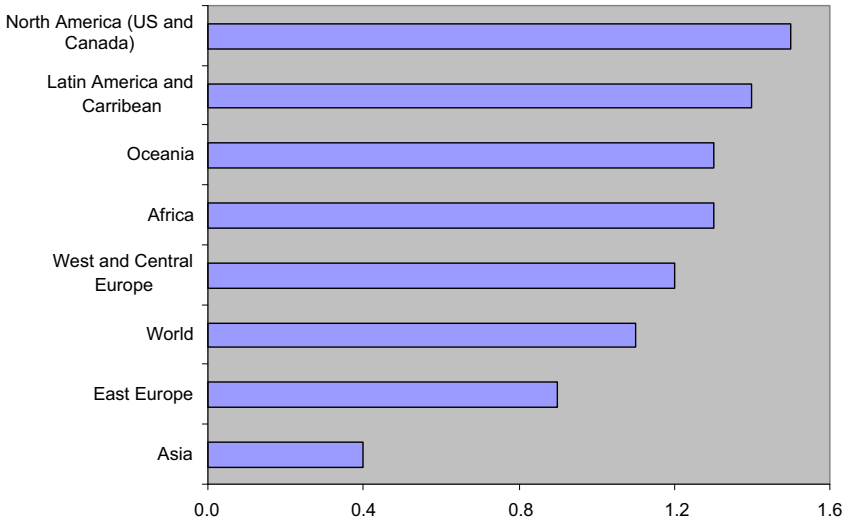

Figure 3 Percentages of the public in urban areas victimized by car theft or joyriding during the past 12 months per region. Source: ICVS, 1996–2005; Van Dijk, J.J.M., 2008. The World of Crime: Breaking the Silence on Problems of Security, Justice, and Development across the World. SAGE Publications, Los Angeles.

bicycles has been well documented among student populations in the Netherlands (Felson, 1997).

Equally strong relationships have been demonstrated between motorcycle ownership and motorcycle theft. Motorcycle theft also provides an interesting example of how the introduction of new safety regulations and ensuing changes in behavioral patterns yield an unexpected crime prevention bonus. In the 1970s, several countries introduced safety laws requiring motorcyclists to wear helmets. Offenders wanting to steal a motorcycle now had to go equipped with a helmet or they would be spotted quickly. The enactment of the new helmets legislation has led to dramatic declines in motorcycle thefts in several countries (Mayhew et al., 1989).

The Heavy Crime Burden of the Business Sector

Victimization through common crimes is usually associated with the public at large. In reality the business sector is the prime target group of many forms of property crime. Crimes committed against businesses are not accurately reflected in police figures for the same reason as other crimes. To collect better information on 'commercial victimization' across countries special victimization surveys among businesses have been conducted. The International Crime Business Survey was carried out in 1992 among, initially, samples of retailers in nine European countries, Australia, and South Africa (Walker, 1994; Van Dijk and Terlouw, 1996; Naudee et al., 1999). The questionnaire was modeled after the ICVS and contained a similar range of conventional crimes. The results show that most retail businesses are regularly victimized by conventional crimes. The corporate victimization percentages are many times higher than those of private persons. This is not only true for theft (e.g., shoplifting) but for burglary as well. Studies of the total costs of crime have indicated that the total damages resulting from property crime of the corporate sector surpass those of the public at large (Mayhew, 2003).

Recently, international accounting firms have taken up an interest in crime surveying. A victimization survey of companies across the world indicates that 45% of all respondents had suffered from some form of economic crime in the course of the past 2 years. These global results indicate an increase on previous such surveys which holds across regions (PriceWaterhouseCoopers, 2005). Prevalence rates remained the highest by far among companies in Africa. A similar survey of global companies in 2004 found that just under one half of respondents (47%) had experienced a significant fraud in the last year (Ernst and Young, 2004). According to these two surveys half or more of the perpetrators of economic crimes are company insiders (managers or staff).

Determinants of Property Crime

Urbanization and Crime

A consistent finding in traditional analyses of police-recorded crime rates was that recorded crime is more prevalent in large cities than in rural areas (Beirne, 1993). In many European countries, for example, the number of police-recorded crimes is at least twice as high per 1000 inhabitants in the larger cities as in the countryside. Data from NCVSs have confirmed the urban–rural crime divide. The fairly consistent relationship between urbanization and levels of volume crime at both individual and country level calls for further theoretical reflection. According to 'routine activity theory,' crime rates are determined by the confluence in space and time of pools of motivated offenders and viable opportunities to commit crimes (Cohen and Felson, 1979). This theoretical perspective can help to explain the urban–rural crime divide. While poverty by itself may not be a major cause of crime – in Asia rates of property crimes remain low in even the poorest countries – the propensity to offend is higher among young, disenfranchised males than among other groups. Other factors possibly contributing to higher levels of offending in large cities are weakened family ties and other forms of social cohesiveness and the concentration of people from different cultures. At the same time, and perhaps more importantly, targets of crime such as cars and other luxury goods are abundantly available in large cities while the daily routines of urban dwellers may leave these targets of crime less 'socially well guarded.' Typical for large cities, for example, are shopping malls and large-scale school complexes where 'natural surveillance' is weak. City life generally brings people into contact more often with strangers, including potential offenders. This confluence of criminogenic push and pull factors in large cities goes some way to explain the higher crime rates in big cities (Felson, 1997).

Demographics and Crime

Thefts, burglaries, assaults, and other forms of volume crime are to a large extent youth-related phenomena. In all countries, the majority of offenders as well as a disproportionately large proportion of victims are males between 15 and 30 years of age (Adler et al., 1998). The ICVS database allows comparisons of the victimization rates of various age groups for different types of crime. Universally, younger people (24 years of age or less) are markedly more at risk than older ones, defying the stereotypical notion of offenders preying on the elderly. Findings on offending show the same age-related patterns and profiles. Among 12–18 year olds in many countries, engagement in forms of juvenile delinquency is very common among teenagers, especially boys. Self-reported delinquency surveys, another method for collecting information on crime independently from the police, show that the peak of young people's criminal activities lies between 15 and 18 years (Bartelet, 2004). Overrepresentation of young males among offenders is attributed by criminologists to weak social integration during the transition from childhood to adulthood (Wilson and Herrnstein, 1985). Many studies report steep declines in criminal involvement after the age of 25 for property crimes and 30 for violent crime (called 'aging out' phenomenon).

Analyses of ICVS data on victimization by common crime confirm that overall victimization rates are significantly higher in societies with more young males such as most African and Latin American countries. Previous analyses of global ICVS victimization data have shown that the macro factors of urbanization and age structure combined explain half of the variation in national victimization rates (Van Dijk, 2008). Put

differently, if all world populations were roughly similar in degree of urbanization and age structure, national differences in levels of crime would be reduced by half.

Affluence and Crime

From an economic perspective, it seems self-evident that for those living in extreme poverty, the commission of profitable offenses presents a tempting alternative source of income. This line of thinking, known as anomia or strain theories, has dominated the twentieth-century criminology. Empirical studies do not confirm, however, any straightforward relationship between poverty and crime at the macrolevel (Belknap, 1989; Van Wilsem, 2003). For a proper understanding of crime rates the prevalence of poor or unemployed people in a population is not the only relevant factor. The prevalence of potential victim populations and of situational factors facilitating crime must also be taken into consideration (Van Dijk, 1994). Some of these factors are positively related to affluence. One such factor is, as mentioned above, the availability of suitable targets of property crime. Affluence brings into mass circulation goods that fit the acronym CRAVED: Concealable, Removable, Available, Valuable, Enjoyable, and Disposable (Clarke, 1999). Examples of such craved goods driving up crime rates in modern affluent societies are mobile phones (Harrington and Mayhew, 2001). Affluence is in many parts of the world accompanied by urbanization and ensuing weakening of social cohesiveness. These factors drive up levels of property crimes, regardless of the income situation among local pools of potential offenders.

According to conventional criminological theories, fewer people will be motivationally inclined to engage in offending if societies become more prosperous. However, crime trends in developed countries over the past 50 years have not been in line with this optimistic model of crime prevention through development. Trends in crime over time in the Western world are positively related to levels of affluence. All developed countries experiencing 'economic miracles' in the 1960s after recovering from the World War II saw their crime rates surging, with crime rates multiplying fivefold or more in some countries (Laycock, 2001). In conclusion, the crime reducing effects of affluence through a reduction of the motivation to offend for economic reasons are more than offset by the increased availability of opportunities to commit profitable crimes.

Trends in Property Crime over Time

The first ICVS was carried out in 1989 and repeated four times till 2005 in a large group of countries. For Johannesburg, Buenos Aires, Australia, Canada, and 11 European countries data are available from all five sweeps. The results show that levels of conventional crime have in most parts of the world continued to rise up to the mid-1990s. This rise in crime has started in the 1960s or 1970s. After reaching peaks between 1995 and 2000, national crime rates have fallen significantly across the Western world. Figure 4 shows results for victimization rates by any crime (largely determined by property crime).

Since the third sweep in the mid-1990s victimization by crime has dropped significantly nearly everywhere in the participating nations. The falls in crime are by no means a unique American phenomenon as previously claimed. The

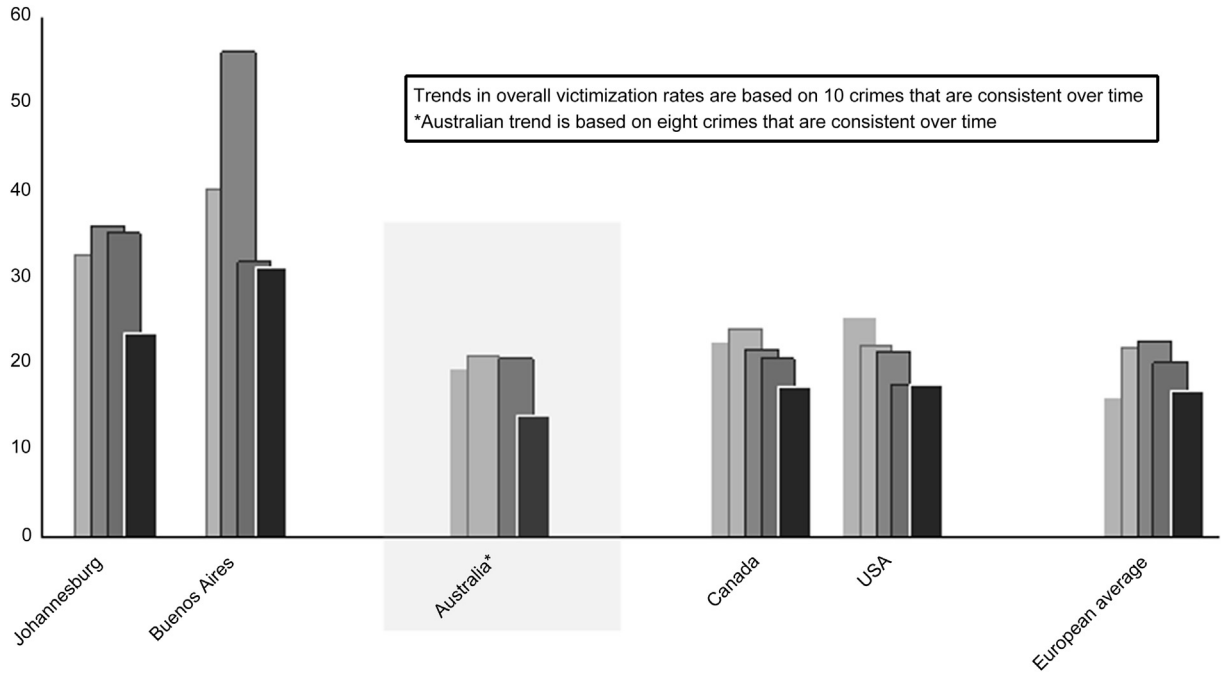

Figure 4 Trends in crime experienced by national or city populations during 1988–2005; overall victimization rates in Argentina (Buenos Aires), South Africa (Johannesburg), Australia, USA, Canada, and Europe (nine countries). Source: ICVS, 1988–2005.

falls are near universal in the industrialized nations and have also happened in the cities of Johannesburg and Buenos Aires.

Many different explanations have been put forward for the drop in crime in the USA (Blumstein and Wallman, 2006; Zimring, 2006). Factors that are special for the USA such as zero-tolerance policing, mass incarceration, or the end of the crack cocaine epidemic cannot explain the universal drops in crime across the Western world (Van Dijk et al., 2012). The universal drops in the two most common types of property crime, car theft and household burglary, seem primarily to have been the result of improved security as will be illustrated below.

Levels of car theft have gone up universally in Western nations from 1960 onward in tandem with rates of car ownership. Although car ownership levels are stable or still rising, thefts of cars have gone down considerably in all Western nations, sometimes by more than 50%. In Germany, rates of car theft have already started to fall by almost 50% in 1961. This drop tracked federal legislation that made high-quality steering column locks mandatory in all cars. Car thefts dropped already the same year (Mayhew, 1992). In the USA and Britain similar legislation was passed 10 years later but limited to the fitting of steering column locks in newly sold cars. As was to be expected, the impact of these regulations was much more delayed than the German. In Britain it took almost 10 years before over 80% of cars were fitted with steering column locks. It was indeed around that time that rates of car theft per 100 owners started to stabilize. In the USA the impact of the new regulation came faster, which appears related to a faster renewal of the vehicle population than in Britain. The critical penetration rate of 80% was reached earlier and car thefts rates started to fall accordingly much sooner. Studies into the impact of state-of-the-art security measures against car theft have been conducted in many countries in more recent years, often using data from victimization surveys (Farrell et al., 2011a). Electronic immobilizers were found to be the most effective to prevent theft of cars. In Australia, the USA, Europe, and Canada the falls in car theft have tracked increases in the penetration rate of anti-car theft security among cars (Fujita and Maxfield, 2012). Electronic immobilizers were made mandatory in 1998 for all newly sold cars within the European Union. Within 10 years after the regulation took effect, it had reduced car thefts by 70% in the Netherlands and by 80% in Britain, according to the analyses of Van Ours and Vollaard (2012). Across the Western world improved car security has substantially reduced nonprofessional car theft (joyriding) (Van Dijk, 2012). Over the last years the introduction of parts marking seems to have driven down professional car theft as well in both the USA and Taiwan.

The second example of security-driven falls in crime is household burglary. In the Netherlands the installment of basic household security measures has been actively promoted by the government since the mid-1980s. In 1999, basic household security was included in the Building Regulations and since then such security is mandatory for all newly built houses. Vollaard and Van Ours (2011) analyzed results of the Dutch national victimization surveys to determine the impact of the new building regulations upon burglary victimization rates of owners of newly built and older houses. Their analysis shows that risks to be burgled of houses built after 1999 are reduced by 50%. According to the authors the new building regulations had been responsible for almost a fifth of the total drop in burglaries in the Netherlands in recent years.

Western countries show considerable variation in levels of household security. Levels of home security are by far the highest in the Netherlands and the UK and the lowest in Estonia, Switzerland, and Denmark. The repeat of the ICVS in 2010 in eight Western countries allows a test of the hypothesis that trends in burglary victimization between 2004 and 2010 are related to the level of security in 2004. The results revealed that among the eight participating countries such relationship is indeed in evidence (Van Dijk and Vollaard, 2012; Van Dijk, 2012). Burglary rates dropped in the Netherlands and the UK and went up in Denmark and Estonia and remained stable in the four countries with medium levels of security. When measures of home security are added up, the correlation between level of home security in 2004 and the trends in burglary victimization between 2005 and 2010 is near perfect. The fact that the UK and the Netherlands resemble Denmark and Switzerland on key economic and social indicators makes the relationship the more telling.

The links between improved security and drops in car theft and burglary suggest that security has acted as the main driver of the universal drops in property crime in a more general sense. If levels of crime have been propelled, as supposed by Felson and others, by expanding pools of suitable targets of crime during the postwar economic booms, improved security might eventually have driven crime rates down (Farrell et al., 2011b). Important examples of improved security across the board are the huge increases in private security guards and alarm centers. In 'Western' countries private security guards now outnumber police officers (Van Steden and Sarre, 2007; de Waard, Berghuis, 2012). But such 'responsive securitization' is not limited to human surveillance by security guards. Measures to prevent crime have become ubiquitous in all corners of modern society. Harnessing new technology, security provisions have been built into homes, cars, stores and parking lots, public transport and public/social housing, schools and hospitals, offices and other work places, entertainment venues and sports stadiums, airports and seaports, and to warehouses and transportation terminals (Clarke and Newman, 2006). Other specific examples of effective security are situational measures against robberies of post offices and banks (Laycock, 2001). The security hypothesis is especially pertinent to understanding emerging trends in new forms of e-fraud. Relatively high levels of identity theft in the USA, for example, are explained by levels of mandatory security below those on the European continent (Van der Meulen, 2011).

As security, crime prevention, and criminal justice reach sufficient levels, the costs of offending will increase and crime rates are likely to start falling. Increased affluence leads to a 'crime harvest' in the medium term that can last for decades. Subsequently measures are put in place by producers of vulnerable products or potential victims to counter increased losses from crime, and crime rates begin to stabilize at a lower level and then to fall.

Whether the current security-driven international drops in property crime will be sustainable in future years remains to be seen. Now that levels of many types of property crime have been more than halved, the motivation to apply security

measures may wane and opportunities of burglary, theft, and fraud may widen again. New products and services may once again ample offer opportunities for the commission of profitable crimes. In the USA the first signs of a new upswing in rates of property crimes have been observed through the national victimization surveys. In 2012, victimization by both violent and property crime went up considerably for the second consecutive year (Langton et al., 2013). A new crime epidemic might be in the making, posing a challenge to both governments and potential victims.

Bibliography

Adler, F., Mueller, G.O.W., Laufer, W.S., 1998. Criminology. McGraw-Hill, Boston.

Aebi, M., Linde, A., 2012. The growth of victimization surveys in Latin America. In: Groenhuijsen, M., Letschert, R., Hazenbroek, S. (Eds.), Liber Amicorum J.J.M. Van Dijk. Wolf Legal Publishers, Nijmegen, pp. 9–21.

Afrobarometer, March 2004. Afrobarometer Round 2: Compendium of Comparative Results from a 15-Country Survey. Working paper No. 34.

Blumstein, A., Wallman, J., 2006. The Crime Drop in America, revised ed. Cambridge University Press, Cambridge.

Beirne, P., 1993. Inventing Criminology. Essays on the Rise of 'Homo Criminalis'. State University of New York Press, Albany/New York.

Bartelet, R., 2004. Self-reported Juvenile Delinquency in England and Wales, The Netherlands and Spain. European Institute for Crime Prevention and Control, Helsinki.

Belknap, J.E., 1989. The economics-crime link. Criminal Justice Abstracts 21 (1).

Blumstein, A., Wallman, J., 2000. The Crime Drop in America. Cambridge University Press, New York.

Clarke, R.V., 1999. Hot Products, Understanding, Anticipating and Reducing Demand for Stolen Goods. Police Research Series. Home Office, London.

Clarke, R.V., Newman, G., 2006. Outsmarting the Terrorists. Praeger, Westport.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Ernst & Young, 2004. Fraud, the Unmanaged Risk, 8th Global Survey. Online: www.ey.com/south africa.

Farrell, G., Tseloni, A., Tilley, N., 2011a. The effectiveness of vehicle security devices and their role in the crime drop. Criminology and Criminal Justice 11, 21–35.

Farrell, G., Tseloni, A., Tilley, N., Mailley, J., 2011b. The crime drop and the security hypothesis. Journal of Research in Crime and Delinquency 48, 147–175.

Felson, M., 1997. Crime and Everyday Life, second ed. Pine Forge Press/SAGE, Thousand Oaks.

Felson, M., Clarke, R.V., 1998. Opportunity makes the thief. practical theory for crime prevention. In: Police Research Series, Paper 98. Policing & Reducing Crime. Home Office, London.

Fujita, S., Maxfield, M., 2012. Security and the drop in car theft in the United States. In: Van Dijk, J., Tseloni, A., Farrell, G. (Eds.), The International Falls in Crime. PalgraveMacMillan.

Harrington, V., Mayhew, P., 2001. Mobile Phone Theft. Home Office Research Study 235. Home Office, London.

Langton, L., Planty, M., Truman, J., 2013. Criminal Victimization 2012. Bureau of Justice Statstics, Washington.

Laycock, G., 2001. Scientists or Politicians – Who Has the Answer to Crime?. Lecture delivered on 26 April, 2001 University College London, The Jill Dando Institute of Crime Science, London.

Lynch, J.P., 2006. Problems and promise of victimization surveys for cross-national research. In: Crime and Justice, vol. 34. University of Chicago, Chicago.

Mayhew, P., 1992. Steering column locks and car theft. In: Clarke, R.V. (Ed.), Situational Crime Prevention: Successful Case Studies. Harrow and Heston, Albany, NY, pp. 52–65.

Mayhew, P., April 2003. Counting the costs of crime in Australia. In: Trends & Issues in Crime and Criminal Justice, vol. 247. Australian Institute of Criminology.

Mayhew, P., Clarke, R.V., Elliot, D., 1989. Motorcycle theft, helmet legislation and displacement. Howard Journal 28, 1–8.

Mawby, R.I., 2001. Burglary. Crime and Society Series. William Publishing, Devon, UK.

Naudee, C.M.B., Prinsloo, J.H., Martins, J.H., 1999. Crimes against the South African Business Sector. University of South Africa, Pretoria.

PriceWaterhouseCoopers, 2005. Global Economic Crime Survey 2005, in Collaboration with Martin Luther University. Economy and Crime Research Center.

United Nations Office on Drugs and Crime (UNODC), 2005. Crime and Development in Africa. Vienna.

Van Dijk, J.J.M., 1994. Understanding crime rates: on the interactions between the rational choices of victims and offenders. British Journal of Criminology 34, 105–121.

Van Dijk, J.J.M., 2012. Closing the doors; highlights of the International Crime Victims Survey 1987-2012 (valedictory lecture 7 December), Tilburg University, The Netherlands.

Van Dijk, J.J.M., Terlouw, G.J., 1996. An international perspective of the business community as victims of fraud and crime. Security Journal 7, 157–167.

Van Dijk, J.J.M., 1999. The experience of crime and justice. In: Newman, G. (Ed.), Global Report on Crime and Justice. United Nations Office for Drug Control and Crime Prevention, Centre for International Crime Prevention. Oxford University Press, New York.

Van Dijk, J.J.M., 2008. The World of Crime: Breaking the Silence on Problems of Security, Justice, and Development across the World. SAGE Publications, Los Angeles.

Van Dijk, J.J.M., Mayhew, P., Killias, M., 1990. Experiences of Crime across the World. Key Findings from the 1989 International Crime Survey. Kluwer Law and Taxation Publishers, Deventer.

Van Dijk, J.J.M., Van Kesteren, J., Smit, P., 2007. Criminal Victimization in International Perspective: Key Findings from the 2004–2005 ICVS and EUICS. Boom Juridische Uitgevers, Den Haag.

Van Dijk, J.J.M., Tseloni, A., Farrell, G. (Eds.), 2012. The International Drop in Crime, New Directions in Research. PalgraveMacMillan.

Van Dijk, J.J.M., Vollaard, B., 2012. Self-limiting crime waves. In: Van Dijk, J., Tseloni, A., Farrell, G. (Eds.), The International Drop in Crime; New Directions in Research. Palgrave/Macmillan, Hampshire, UK.

Van der Meulen, N., 2011. Fertile Grounds; the Facilitation of Financial Identity Theft in the USA and the Netherlands. Wolf Legal Publishers, Nijmegen.

Van Ours, J.C., Vollaard, B., 2012. The Engine Immobilizer; a Non Starter for Car Thieves. Department of Economics, Tilburg University.

Van Steden, R., Sarre, R., 2007. The Growth of Privatized Policing: Some Cross-national Data and Comparisons. Free University of Amsterdam.

Van Wilsem, J., 2003. Crime and Context. The Impact of Individual, Neighborhood, City and Country Characteristics on Victimization (Doctoral Dissertation, Katholieke Universiteit Nijmegen, 2003). ICS.

Vollaard, B.A., Van Ours, J.C., 2011. Does regulation of built-in security reduce crime? Evidence from a natural experiment. The Economic Journal 121, 485–504.

Waard, de, Berghuis, B., July 2012. Trends in de private veiligheidszorg: de stand van zaken in 2012. Secondant 3/4, 8–15.

Walker, A., et al., July 2009. Crime in England and Wales 2008/2009. In: Findings from the British Crime Survey and Police-Recorded Crime, vol. 1. Home Office, London.

Walker, J., 1994. The First Australian National Survey of Crimes against Businesses. Australian Institute of Criminology, Canberra.

Wilson, J.Q., Herrnstein, R.J., 1985. Crime and Human Nature. Simon and Schuster, New York.

Zimring, F., 2006. The Great American Crime Decline. Oxford University Press, Oxford.

Zvekic, U., 1998. Criminal Victimization in Countries in Transition. UNICRI, Rome.

Property: Legal Aspects

Peter S Menell, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Property encompasses the governance of resources – real property (land), natural resources, and personal property (chattels and intangibles, including copyrights, patents, stocks, claims of right). Private property finds support in Locke's labor theory, as well as libertarian, utilitarian, and personhood philosophical traditions. Distributive justice, civic republicanism, and ecological theories take more qualified stances on the protection of private property, while Marxist ideology rejects private property as a vestige of exploitation and alienation of the worker class. Modern conceptions of property trace their roots to Blackstone's inviolate, unified bundle of rights (see following). As societies moved beyond agrarian modes of production, and as legal relations evolved from status to contract, the bundle of rights gradually came to be seen as malleable instruments to be tailored by legislatures, courts, government agencies, and markets. By the twentieth century, zoning and other governmental regulations significantly qualified the rights of property owners, and scholars developed frameworks for analyzing the economic, distributive, and other moral implications of different systems of resource governance. A new conception of property integrating the many institutional structures governing resources, the cultural contingency of governance regimes, and the dynamism of governance institutions has emerged. This broader framework provides the basis for tackling the new frontiers of resource governance.

Property encompasses the governance of resources – real property (land), natural resources, and personal property (chattels and intangibles, including copyrights, patents, stocks, and claims of right). Natural law theory, beginning with Aristotle and perhaps best articulated in the writings of John Locke (1698), posits that "life, liberty, and property" exist naturally as a matter of fundamental principles of justice, irrespective of the state. Echoing Locke's formulation, the United States Constitution provides that a person may not "be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation." By contrast, legal positivists, such as Jeremy Bentham (1789), assert that property rights arise only to the extent that they are formally recognized by the state: "Property and law are born together, and die together. Before laws were made there was no property; take away laws, and property ceases." Both views give content to what we understand as property. Legal positivism captures the practical reality that legislatures and courts define property rights in most societies, and that property rights have practical meaning only to the extent that they are enforceable. Nonetheless, natural law (and other philosophical frameworks) informs the content of property rules by providing the justification for state-created rights.

Philosophical Justifications for and Critiques of Private Property Rights

Waldron (1988) suggests that property is best viewed as a "system of rules governing access to and control of material resources." These rules are often characterized as a bundle of rights – in the case of land, the rights to possess (or occupy), to exclude, to use, to transfer, to enjoy fruits, and to alter or destroy – that can be allocated in different ways. Particular bundles represent different conceptions of property, which derive from one or more philosophical justifications.

Labor Theory

John Locke (1698) offered a strong natural rights justification for private property, which remains a central pillar of property theory today. Beginning with the proposition that all humans possess property in their own 'person,' Locke argued that

> [t]he 'labour' of his body and the 'work' of his hands, we may say, are properly his. Whatsoever, then, he removes out of the stat[e] that Nature hath provided and left it in, he hath mixed his labor with it and joined to it something th[a]t is hi[s] own and thereby makes it his property. It being by him removed from the common state Nature placed it in, it hath by this labour something annexed to it that excludes the common right of other men. For this 'labour' being the unquestionable property of the labourer, no man but he can have a right to what that is once joined to, at least where there is enough and as good left in common for others.

Locke's theory is widely viewed as justifying first possession as a basis for acquiring property rights (Rose, 1985; Epstein, 1979). While useful as a basis for establishing rights in land and raw materials in a world of abundant unclaimed resources, Locke's theory significantly qualifies the laborer's claim in a world of scarcity – requiring 'enough and as good left in common for others.' Waldron (1988), Rose (1987), Ryan (1984), Becker (1977), and Schlatter (1951) elaborate and critique Locke's justification for private property rights.

Libertarianism

Drawing upon Lockean principles, libertarians emphasize individual freedom to control and alienate their private property. Nozick (1974) asserts that the state should play no role in distributing or redistributing property acquired properly, apart from respecting the voluntary transfers of property owners. Epstein (1985) argues for a broad interpretation of the government's obligation to compensate owners of private

property for regulation and other governmental decisions interfering with private property rights.

Utilitarianism

First articulated by Jeremy Bentham (1789), utilitarianism judges the allocation of resources and the design of governance institutions on the basis of the utility or individual satisfaction that they produce. In applying this theory to property rights, utilitarian scholars have focused principally upon the inefficiency that can result from incomplete or shared resource ownership and external effects of property use (such as pollution) on neighboring parcels. In a classic work, Hardin (1968) describes the "tragedy of the commons," whereby unregulated access to a common pasture (or any other common resource such as air, water, or roads) results in overuse. Hardin suggested that private property or government regulation is needed to govern the commons.

Yet private ownership can produce other pathologies. Heller (1998, 2008) explains that excessive fragmentation of property interests can produce a "tragedy of the anti-commons," whereby the cost of determining rights and negotiation paralyze resource use and development. Merrill and Smith (2000) and Smith (2012) explain how standardization and modularity of property units can promote efficient property development and use and ameliorate anti-commons problems.

Menell and Meurer (2013) show that the claiming of resources – such as newly discovered land, mineral rights, or most saliently today, intellectual resources (e.g., patents) – can produce 'notice externalities.' In economic parlance, an externality arises when an actor imposes costs on other parties – for example, by discharging pollutants that adversely affect a downstream fishery. A notice externality arises when property claimants obscure their claims – for example, by seeking patent claims with amorphous boundaries or hiding their identities. This raises the costs of those seeking to develop 'neighboring' properties (or in the case of patents, 'neighboring' technologies), who must expend additional resources on surveying (or in the case of patents, search, and opinion letters) and bear the risk and cost of later disputes about encroachment or infringement. Once their investments are sunk, the property claimant can pounce opportunistically, with the object of holding up the neighbor's investment. Had the property claim been clear from the outset, the neighbor could have confronted the issue prior to sinking the investment.

Economists generally prefer private property institutions (Posner, 1998), although the efficacy of private property depends critically upon the costs of creating, monitoring, and enforcing such rights (Smith, 2002; Lueck, 1995; Ellickson, 1993; Anderson and Hill, 1975; Demsetz, 1967). Government regulation often entails two costs: (1) direct costs to comply with the regulations imposed; and (2) influence costs attributable to rent-seeking (or lobbying) – ex ante efforts to secure economic advantage through favorable government decisions by those interests affected by regulation. Although private property regimes often ameliorate the tragedy of the commons, shared governance institutions including informal social norms can maximize utility in a range of circumstances depending upon the cohesion and history of the community affected and the nature of the resources in question (Libecap, 1994; Ellickson, 1991; Ostrom, 1990; Rose, 1986).

Personhood

The personhood justification for private property, derived from Kant's Philosophy of Law and Hegel's Philosophy of Right, has been elaborated in modern legal discourse in the work of Radin (1982, 1993). "The premise underlying the personhood perspective is that to achieve proper development – to be a person – an individual needs some control over resources in the external environment. The necessary assurances of control take the form of property rights" (Radin, 1982). The personhood justification for private property emphasizes the extent to which property is personal as opposed to fungible: the justification is strongest where an object or idea is closely intertwined with an individual's personal identity and weakest where the 'thing' is valued by the individual at its market worth. The personhood justification for private property is limited and highly subjective (Schnably, 1993).

Distributive Justice

Theories of distributive justice seek to allocate society's resources on the basis of just principles. The process of determining such principles is the focus of considerable debate. Many theories incorporate utilitarian and Lockean principles. In perhaps, the most influential work on distributive justice of the past century, Rawls (1971) offers an 'ideal contractarian' theory of distributive shares in which a just allocation of benefits and burdens of social life is determined by what rational persons would choose from behind a 'veil of ignorance,' which prevents them from knowing what abilities, desires, parentage, or social stratum they would occupy. Rawls concludes that people behind such a veil would adopt what he calls the 'difference principle': 'primary goods' – not only wealth, income, and opportunity, but also bases of self-respect – would be distributed to the maximal advantage of a representative member of the least advantaged social class. Rawls (1993) emphasizes that the difference principle does not call for simple egalitarianism but rather measures to assure that "the basic needs of all citizens can be met so that they can take part in political and social life."

Civic Republicanism

Republicanism rejects the view that individuals are wholly autonomous and that the public interest is just the aggregation of individual preferences. Republicanism stresses civic virtue and the role of participation in public discourse as essential to the realization of the common good. Civic republicanism relies upon the propertied independence of its citizens. Such ownership of capital affords citizens the competence to engage in meaningful discourse and protects against the subversion of democracy by a wealthy oligarchy or an aggressive state (Simon, 1991).

Marxist Theory

Karl Marx viewed human history as an evolutionary process of class struggle leading, ultimately, to a classless society

(communism) in which the state will wither away. In the capitalist phase of this evolution, private property provides the basis for the most important social relation: ownership of the means of production. According to Marx, the class owning the means of production (bourgeoisie) acquires the property (capacity to work) of the labor class (proletariat) through the employment relation, thereby obtaining exclusive control of and hence power over the worker class. Such subordination, exploitation, and alienation sow the seeds of revolution, leading to the establishment of a proletarian dictatorship, socialism (productive resources owned by workers through participatory institutions), and eventually communism, a classless society without exploitation or inequalities in which property in land and rights of inheritance are abolished (Marx and Engels, 1848).

Ecological Theories

As understanding of the connections between technology, land development, and the environment has deepened, many environmental scholars have increasingly viewed private property rights more skeptically (Sax, 1983). Traditional reformist theories – which are anthropocentric in nature and draw heavily upon utilitarian and other traditional philosophical frameworks – favor internalization of the adverse impacts of development (Menell and Stewart, 1994; Ellickson, 1977, 1973). Building upon Leopold's (1949) call for a land ethic that seeks to "preserve the integrity, stability, and beauty to the biotic community," Freyfogle (1996), Sprankling (1996), and Sax (1993) propose a shift away from private property toward greater regulation and coordination of land and natural resource development. At the other extreme, Anderson and Leal (1991), drawing upon property rights and public choice literatures, advocate complete reliance on private property and free market principles as the best institutional choice to address substantially all environmental problems. Menell (1992) highlights the institutional limitations of both extremes and urges careful attention to the strengths and limitations of the broad spectrum of institutional alternatives in governing resources.

Pluralist and Hybrid Theories of Property Rights

The insights of the many philosophical perspectives have led some scholars to synthesize multiple strands into new and comprehensive theories for justifying and evaluating property. Munzer (1990), for example, constructs a theory of property built upon principles of utility and efficiency, justice and equality, and rewards based on labor. Simon (1991) synthesizes civic republicanism and socialism to advocate an intermediate form of property emphasizing enterprise and community-based democracy (cf. Hansmann, 1996), offering a utilitarian explanation for worker-owned and managed firms and organizations.

The Emergence of Private Property Rights

Although no single philosophical theory provides a complete explanation for private property rights, utilitarianism has emerged as the dominant framework for understanding property rights systems in capitalist economies. In a seminal article, Demsetz (1967) posited that private property regimes develop to internalize externalities. As the benefits of internalizing externalities rise (e.g., where a common resource becomes more valuable) or the cost of internalizing externalities fall (e.g., as technology reduces the costs of fencing or monitoring property rights), private property systems emerge to promote efficient management of the resource. As support for this thesis, Demsetz cites anthropological evidence from the Americas following the arrival of the Europeans and the origins of the fur trade. The fur trade, by spurring the hunting of beavers, created the need for property rules to internalize and minimize the effects of resource exploitation. In response, Native Americans established private hunting territories.

Demsetz's theory explains the emergence of private property institutions across a broad range of contexts, including air (sulfur oxide emissions in the United States), water allocation, fishing, the telecommunications spectrum, wetlands, roads (tolls), trademarks (causes of action for diluting famous marks), and publicity (commercial exploitation of a person's image, likeness, or voice). Nonetheless, by focusing exclusively upon private property institutions, Demsetz's theory overlooks much of the richness of property institutions and the complex social processes that affect their evolution. Smith (2002), Libecap (1994), Ostrom (1990), Ellickson (1991), Rose (1986), and others explore the broad spectrum of property institutions lying between the poles of pure private property and the commons, where informal and shared governance emerge and succeed in allocating resources. The emergence and nature of property rights in a given place and time are also likely to reflect aspects of that society's political economy, including features of the legislative process and of government agencies that administer property rights (Buchanan and Tullock, 1962). Interest groups tend to form in those areas where government policies create concentrated benefits or impose concentrated costs (Olson, 1965). Laws and agencies creating or affecting property rights (e.g., through regulation) attract substantial rent-seeking activity. Thus, the design of property regimes often reflects political, as well as economic and social, considerations.

Evolution of Jurisprudential and Analytical Conceptions of Property

Although property rights are as old as human civilization, the modern jurisprudential conception of property traces its roots to the agrarian and early mercantilist economies that developed during the decline of feudalism. Sir William Blackstone (1765), author of the first widely accepted legal treatise on English law, regarded property as an absolute right

> which consists in the free use, enjoyment, and disposal of all acquisitions, without any control or diminution, save only by the laws of the land …

> So great moreover is the regard of the law for private property, that it will not authorize the least violation of it; no, not even for the general good of the whole community … .

Although some have questioned Blackstone's rigor and cogency (Burns, 1985), few doubt his influence upon the jurisprudence of property in England and America. Maine (1861) reinforced the superiority of private property institutions through his historical account of the progress of civilization from primitive forms of shared ownership among kinship groups (status) to individual ownership and alienability of property interests (contract).

The industrial revolution and the shift from an agrarian economy to urban and other community structures changed the ways in which land and other resources were owned and exploited. The desire to restrict land through contract pushed courts to liberalize constraints on non-possessory interests, presaging the disintegration of the unitary bundle of sticks (French, 1982; American Law Institute, 1936). Property came to be seen increasingly as a malleable concept that could be tailored to serve utilitarian purposes.

In the early twentieth century, Hohfeld (1913) developed a detailed descriptive framework for capturing the richness of legal rights, reconceptualizing property as *relationships* among people relative to things. Recognizing four primary legal entitlements – rights, privileges, powers, and immunities – Hohfeld postulated four basic legal relations ('correlatives'): 'right–duty,' 'privilege–absence of right,' 'power–liability,' and 'immunity–disability.' Hohfeld provided the vocabulary and structure for representing the complexity of property relations – bundles of rights (and relations) – as well as the conceptual menu from which legislatures and courts could craft property rights.

Blackstone's conception of property as an inviolate, unified bundle of rights came under increasing assault with the proliferation of zoning and other public land use controls upon private property rights and the expansion of the role of government at all levels following the New Deal in the United States. Property rights, like much of the law governing legal relations, came increasingly within the purview of political institutions, and accordingly gave way to the instrumentalist (utilitarian) nature of politics.

In a seminal article, Coase (1960) developed a new, instrumentalist framework for analyzing property rights. In much the way that a frictionless surface enables physicists to isolate and study the laws of physics, Coase hypothesized a world of zero transaction costs (costless bargaining and enforcement of agreements), which provided valuable insights into the interplay of legal and market institutions. Coase demonstrated that in a world of costless transactions, an efficient allocation of resources will result regardless of how property rights are initially allocated. The desire of conflicting property owners to maximize their joint profit will lead them to bargain to an efficient allocation of resources. The initial allocation of entitlements will affect the distribution of wealth, but not the total level of wealth.

As with the frictionless surface, Coase recognized that costless transactions do not exist in the real world. His article fueled a deluge of scholarship exploring the implications of costly transactions for the allocation and design of property rights. Building upon Coase's construct, Calabresi and Melamed (1972) developed a framework that divides the analysis of property rights into two inquiries: (1) how should property rights be allocated; and (2) how should such entitlements be enforced – by a property rule (enabling property owners to enjoin interference with their entitlement and requiring consent for violation or transfer of the entitlement), a liability rule (allowing property owners court-determined damages for invasion of their property right, as in the case of eminent domain), or by a rule of inalienability (preventing transfer of a property right, as with prohibitions on the sale of body parts).

Calabresi and Melamed propose that entitlements and enforcement rules be based on economic efficiency, distributional goals, and other considerations of justice. Where transactions are costly, entitlements should be allocated so that the party who is best able to minimize harm (the least cost avoider) bears responsibility for the harm in order to promote economic efficiency. With regard to the enforcement of entitlements, Calabresi and Melamed demonstrate that the use of liability rules can minimize the costs of making an erroneous decision about the allocation of entitlements where transactions are costly. Subsequent scholars have refined and extended the framework and applied it in a range of contexts (Symposium, 1997; Kaplow and Shavell, 1996; Merges, 1996; Radin, 1996; Krier and Schwab, 1995; Rose-Ackerman, 1985; Polinsky, 1980; Ellickson, 1977, 1973).

In a provocative essay published in the early 1980s, Grey (1980) argued that the field of property "ceases to be an important category in legal and political theory" as a result of the 'disintegration' of the conception of property. In his view, the dividing of property ownership into a 'bundle of rights' and the development of differing conceptions of property across disciplines has 'fragmented' property theory 'into a set of discontinuous usages.' Numerous scholars have questioned Grey's claims on a range of bases (Merrill and Smith, 2001 (contending that it is essential to view the institution of property as an *in rem* ('thing') concept), Smith, 2012; Epstein, 2010; Munzer, 1990), yet the observation that provided the point of departure for Grey's essay, the 'disintegration' of property, resonates today (Menell, 2011; Dwyer and Menell, 1998). Property has lost the simplicity of Blackstone's conception – unitary, exclusive, and omnipotent control – but a new conception of property integrating the many institutional structures governing resources, the cultural contingency of governance regimes, and the dynamism of governance institutions has emerged.

Property as a Triadic Relation

Dwyer and Menell (1998) present property as a dynamic system comprising three components: resources, institutions, and culture. Resources may be governed by background legal institutions (such as common law rules, or in civil law regimes, codes), social norms, markets (contractual agreements supported by the background legal institutions), political institutions (e.g., zoning), or some combination thereof. At any point in time, the mix of governance rules and institutions reflects the nature of the resources governed and the society or culture (knowledge, values, cohesion, technology) in which the resources are situated. Thus, in smaller, more cohesive social settings, we observe greater use of social norms or shared governance structures, whereas the

same resources tend to be governed through formal private property institutions in larger, more heterogeneous communities. The triadic framework captures the economic and cultural dependency of resource governance as well as the diversity and richness of property regimes.

The triadic conception of property also incorporates the dynamic nature of resource governance. As reflected in Figure 1, changes in the society (such as the value attributed to the protection of endangered species or technological advances reducing the costs of monitoring private property rights in particular resources) generate changes in the design and mix of governance institutions. Such changes, by affecting the resource base, can generate further reactions within the overall system through various feedback mechanisms. The system can also be affected by external shocks, such as changes in the larger political atmosphere. Demsetz's (1967) positive account represents a special case of this dynamic process, but the triadic relation generalizes the dynamism of institutional evolution to account for a broader mix of governance institutions (see also Smith, 2002).

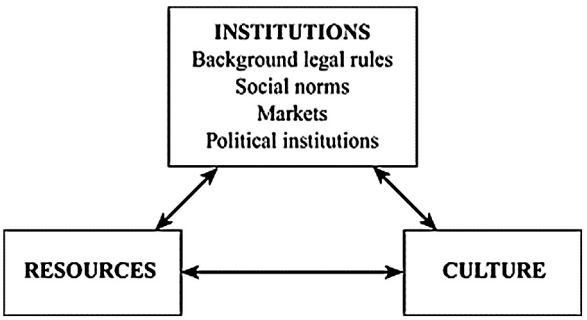

Figure 1 Property.

New Frontiers

Notwithstanding the settlement of substantially all significant land masses on the planet, the frontiers of property continue to expand and evolve. Among the most important resources in today's economy are intangibles – ideas, genetic code, cyberspace, telecommunication spectrum (Merges et al., 2012). The Internet has significantly expanded opportunities to distribute expressive works, but has also threatened the ability of copyright owners to control their works (Menell, 2002, 2011). This new medium has reshaped perceptions about rights to reproduce and distribute expressive works. Advances in bioscience have spawned disputes over rights and control of human genetic code, cell lines, and indigenous plants. And even in the more conventional realm of land management, increasing pressures on natural resources, greater knowledge of ecosystem effects, and changes in attitudes toward the environment continue to alter the balance among private property rights, public ownership, and regulation (Dwyer and Menell, 1998; Menell and Stewart, 1994; Rieser, 1997). Public rights in artistic treasures, historical documents, architectural works, archeological discoveries, and scholarly research present a further set of challenges for a system of governance built significantly upon a foundation of private property rights (Sax, 1999). Just as the advent of rights in government benefits expanded our conception of property four decades ago (Reich, 1964), advances in technology and changing social values continue to expand and alter our conception of property (Rose, 1998).

See also: Intellectual Property: Law in Context; Law and Geography; Legal Positivism; Natural Law; Property: Legal Aspects of Intergenerational Transmission; Socialist Law.

Bibliography

American Law Institute, 1936. Restatement of the Law of Property. American Law Institute Publishers, St Paul, Minnesota, MN.
Anderson, T.L., Hill, P.J., 1975. The evolution of property rights: a study of the American West. Journal of Law and Economics 12, 163–179.
Anderson, T.L., Leal, D.R., 1991. Free Market Environmentalism. Pacific Research Institute for Public Policy, San Francisco.
Becker, L., 1977. Property Rights: Philosophic Foundations. Routledge & Kegan Paul, London.
Bentham, J., 1789. An Introduction to the Principles of Morals and Legislation.
Blackstone, W., 1765. Commentaries.
Buchanan, J., Tullock, G., 1962. The Calculus of Consent: Logical Foundations of Constitutional Democracy. University of Michigan Press, Ann Arbor, MI.
Burns, R.P., 1985. Blackstone's theory of the 'absolute' right of property. University of Cincinnati Law Review 54, 67–86.
Calabresi, G., Melamed, A.D., 1972. Property rules, liability rules, and inalienability: one view of the cathedral. Harvard Law Review 85, 1089–1128.
Coase, R., 1960. The problem of social cost. Journal of Law and Economics 3, 1–28.
Demsetz, H., 1967. Toward a theory of property rights. American Economic Review Papers and Proceedings 57, 347–358.
Dwyer, J.P., Menell, P.S., 1998. Property Law and Policy: A Comparative Institutional Perspective. Foundation Press, Westbury, NY.
Ellickson, R.C., 1973. Alternatives to zoning: covenants, nuisance rules, and fines in land use controls. University of Chicago Law Review 40, 681–781.
Ellickson, R.C., 1977. Suburban growth controls: an economic and legal analysis. Yale Law Journal 86, 385–511.
Ellickson, R.C., 1991. Order without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.
Ellickson, R.C., 1993. Property in land. Yale Law Journal 102, 1315–1400.
Epstein, R.A., 1979. Possession as the root of title. Georgia Law Review 12, 1221–1243.
Epstein, R.A., 1985. Taking: Private Property and the Power of Eminent Domain. Harvard University Press, Cambridge, MA.
Epstein, R.A., 2010. The disintegration of intellectual property? A classical liberal response to a premature obituary. Stanford Law Review 62, 455–521.
French, S.F., 1982. Toward a modern law of servitudes: reweaving the ancient strands. Southern California Law Review 55, 1261–1319.
Freyfogle, E.T., 1996. Ethics, community, and private land. Ecology Law Quarterly 23, 631–661.
Grey, T., 1980. The disintegration of property. Nomos 22, 69–85.
Hansmann, H., 1996. The Ownership of Enterprise. Harvard University Press, Cambridge, MA.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.
Heller, M., 1998. The tragedy of the anticommons: property in the transition from Marx to markets. Harvard Law Review 111, 621–688.
Heller, M., 2008. The Gridlock Economy: How Too Much Ownership Wrecks Markets, Stops Innovation, and Costs Lives. Basic Books.
Hohfeld, W.N., 1913. Fundamental legal conceptions as applied in judicial reasoning. Yale Law Journal 23, 16–59.
Kaplow, L., Shavell, S., 1996. Property rules versus liability rules: an economic analysis. Harvard Law Review 109, 713–790.
Krier, J., Schwab, A., 1995. Property rules and liability rules: the cathedral in another light. New York University Law Review 70, 440–483.
Leopold, 1949. A Sand County Almanac. Oxford University Press, New York.
Libecap, G., 1994. The conditions for successful collective action. Journal of Theoretical Politics 6, 569–598.
Locke, J., 1698. Second Treatise of Government.

Lueck, D., 1995. The rule of first possession and the design of the law. Journal of Law and Economics 38, 393–436.
Maine, H.S., 1861. Ancient Law.
Marx, K., Engels, F., 1848. The Communist Manifesto.
Menell, P.S., 1992. Institutional fantasylands: from scientific management to free market environmentalism. Harvard Journal of Law and Public Policy 15, 489–510.
Menell, P.S., 2002. Envisioning copyright law's digital future. New York Law School Law Review 46, 63–199.
Menell, P.S., 2011. Governance of intellectual resources and disintegration of intellectual property in the digital age. Berkeley Technology Law Journal 27, 1523–1559.
Menell, P.S., Meurer, M., 2013. Notice failure and notice externalities. Journal of Legal Analysis 1.
Menell, P.S., Stewart, R.B., 1994. Environmental Law and Policy. Little, Brown, Boston, MA.
Merges, R.P., 1996. Contracting in liability rules: intellectual property and collective rights organizations. California Law Review 84, 1293–1393.
Merges, R.P., Menell, P.S., Lemley, M.A., 2012. Intellectual Property in the New Technological Age, sixth ed. Aspen Law & Business, New York.
Merrill, T.W., Smith, H.E., 2000. Optimal standardization in the law of property: the numerus clausus principle. Yale Law Journal 110, 1–70.
Merrill, T.W., Smith, H.E., 2001. What happened to property in law and economics? Yale Law Journal 111, 357–398.
Munzer, S.R., 1990. A Theory of Property. Cambridge University Press, Cambridge, UK.
Nozick, R., 1974. Anarchy, State, and Utopia. Blackwell, Oxford, UK.
Olson Jr., M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.
Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, UK.
Polinsky, A.M., 1980. Resolving nuisance disputes: the simple economics of injunctive and damage remedies. Stanford Law Review 32, 1075–1112.
Posner, R., 1998. Economic Analysis of Law, fifth ed. Aspen Law & Business.
Radin, M.J., 1982. Property and personhood. Stanford Law Review 34, 957–1015.
Radin, M.J., 1993. Reinterpreting Property. University of Chicago Press, Chicago, IL.
Radin, M.J., 1996. Contested Commodities. University of Chicago Press, Chicago, IL.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Rawls, J., 1993. Political Liberalism. Harvard University Press, Cambridge, MA.
Reich, R., 1964. The new property. Yale Law Journal 73, 733–787.
Rieser, A., 1997. Property rights and ecosystem management in US fisheries: contracting for the commons? Ecology Law Quarterly 24, 813–832.
Rose, C., 1985. Possession as the origin of property. University of Chicago Law Review 52, 73–88.
Rose, C., 1986. The comedy of the commons: custom, commerce, and inherently public property. University of Chicago Law Review 53, 711–781.
Rose, C., 1987. Enough, and as good of what? Northwestern University Law Review 81, 417–442.
Rose, C., 1998. The several futures of property: of cyberspace and folk tales, emission trades and ecosystems. Minnesota Law Review 83, 129–182.
Rose-Ackerman, S., 1985. Inalienability and the theory of property rights. Columbia Law Review 85, 931–969.
Ryan, A., 1984. Property and Political Theory. Blackwell, Oxford, UK.
Sax, J.L., 1983. Some thoughts on the decline of private property. Washington Law Review 58, 481–496.
Sax, J.L., 1993. Property rights and the economy of nature: understanding. Lucas v. South Carolina Coastal Council. Stanford Law Review 45, 1433–1455.
Sax, J.L., 1999. Playing Darts with a Rembrandt: Public and Private Rights in Cultural Treasures. University of Michigan Press, Ann Arbor, MI.
Schlatter, R., 1951. Private Property: The History of an Idea. George Allen & Unwin.
Schnably, S.J., 1993. Property and pragmatism: a critique of Radin's theory of property and personhood. Stanford Law Review 45, 347–407.
Simon, W.H., 1991. Social republican property. UCLA Law Review 38, 1335–1413.
Smith, H.E., 2002. Exclusion versus governance: two strategies for delineating property rights. The Journal Legal Studies 31, S453–S487.
Smith, H.E., 2012. Property as the law of things. Harvard Law Review 125, 1691–1726.
Sprankling, J.P., 1996. The antiwilderness bias in American property law. University of Chicago Law Review 63, 519–589.
Symposium, 1997. Property rules, liability rules, and inalienability: a twenty-five year retrospective. Yale Law Journal 106, 2083–2213.
Waldron, J., 1988. The Right to Property. Clarendon Press, Oxford, UK.

Property: Legal Aspects of Intergenerational Transmission

Ronald Chester, New England School of Law, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Inheritance is a legal and customary institution that not only is universal but also has been so for centuries. This article discusses both historical and current differences in the transmission of wealth between generations, as well as the relative importance of these differences. It also analyzes future trends in the transmission of property between generations and considers possible changes to the current legal regimes.

Property is transmitted at death by succession, will, will substitute, or a combination of these methods. While the right to transmit property by will is seen by some as a natural right, more often it is viewed as a positivist creation of law for the convenience of society. One of the chief issues in intergenerational wealth transmission has long been the extent to which a decedent could, by use of a will, vary in customary or legal succession patterns protecting the family. The ability to change these succession patterns by will has increased, particularly among English-speaking peoples, as societies have evolved from feudalism to modern free-market capitalism. The Anglo-American concept of a trust, which separates legal from equitable title and can provide managed income for a beneficiary without ever-vesting full title in that individual, complements, or may even supplant, the will in transmitting property between generations.

The attempted abolition of inheritance by the Bolsheviks in 1918 failed in the Soviet Union and was never attempted by the Communist Chinese. Both nations today have inheritance laws that generally mimic those of the civil law systems of Western Europe, where family protection through forced-succession laws tends to limit testamentary freedom. Allowing large amounts of capital to pass between generations without significant taxation is problematic for countries (like the United States) that profess a belief in equal opportunity. Continental nations are generally more active than the United States in taxing such wealth. Such countries as France, Spain, and Germany limit testator freedom not only in this way but also by demanding the family protection mentioned here. Whether the United States will join these and other countries in their approach to intergenerational wealth transfer or continue to protect the rights of individual decedents over the needs both of blood relatives and of society is an open question as the twenty-first century unfolds.

Wills and Succession throughout History

Property exists essentially in two forms: movable (including intangible) and immovable (land). An heir's right of succession to property by law or custom is an ancient and powerful right rooted in kinship ties. For example, early European feudal societies allowed family members to share that portion of the land vacated by a kinsman's death (see Chester, 1982). By the twelfth century in England and in many parts of Western Europe, lands were not partitioned but descended to the eldest son of the landowner under the concept of primogeniture.

The right of a beneficiary to take property under a will is dependent on the legal power or right of the decedent to transmit this property at death to that beneficiary. This right was recognized relatively recently in Western thought and has often been regarded, at least by the positivist tradition, as having been created for society's convenience. Natural rights advocates, by contrast, might consider the right to bequeath an aboriginal attribute of property itself that may not be subject to curtailment (see, e.g., Ascher, 2011; Chester, 1998b, 2007).

In Western society, the legitimization of the right of bequest symbolized the shift from feudal to individualistic conceptions of property. Unknown to the early law of Greece, India, Egypt, Babylon, and Israel, the will was by no means a universal institution. The testamentary conception built upon by Europeans had its roots in Roman law and was resurrected after feudalism by ecclesiastical courts applying Romano-canonical law (see McMurray, 1919).

In England, the customary right of testamentary disposition was recognized in very early times, but land – the primary source of wealth – could not be willed until after the Norman invasion and the establishment of feudal tenures. Finally, in 1540, the Statute of Wills declared the right of an individual to leave real property to whomever the individual pleased by means of a duly executed will. This statute confirmed a customary right long recognized for personal property (Chester, 1982).

Once established, the right of an individual to leave property outside the family had to be reined in to protect certain family members from disinheritance. For example, in England, primogeniture protected the eldest sons of landed decedents until well into the nineteenth century. Common law rights of dower and curtesy for surviving wives and husbands provided them with interests in the land of the decedent at decedent's death (see Chester, 1982). 'Dower' refers to the portion to which a surviving wife is entitled, while 'curtesy' refers to what a surviving husband may claim.

Even though a married English woman of the sixteenth, seventeenth, and eighteenth centuries could not will property (except, possibly, with her husband's consent) or own it in her own name, it could descend through her. However, married women's assets were protected to some extent by courts of equity by means of trusts and other devices, especially toward the end of that period.

Children who were omitted from wills were not protected directly, except those benefiting under primogeniture. However, the wills of their parents were strictly scrutinized to ensure that these children had not been left out unintentionally. One English rule derived from Roman law even stated that to cut off an heir 'without a shilling' raised a presumption that the testator was legally incompetent to make a will. Thus, even in England, the family's right to inherit property was protected not only when the decedent died without a will (intestate) but also to some extent when the decedent did leave such a will (see Chester, 1982).

The entrenchment of the right of testation in England was associated with the movement from feudalism, with its emphasis on family and kinship, to free-market capitalism, with its emphasis on the individual. By the mid-nineteenth century, the adult, male, Anglo-American owner of both realty and movable property had nearly unlimited power of testamentary disposition. In contrast, countries such as France, Spain, and Germany on the European continent made the move to capitalism with a right of testation that, even when recognized, was much more sharply curtailed than in Anglo-America by the requirement of fixed shares for the surviving spouse and children.

In the eighteenth and nineteenth centuries, the principal difference between the Continental and Anglo-American systems of succession was the different attitude each displayed toward the right of family members to succeed to the decedent's property. Although the Spanish principle of forced succession appeared in theory to be a significant curb on the Anglo-American concept of testamentary freedom, in practice the differences were not so marked. This reality stems from two separate phenomena: (1) in the majority of instances, free testamentary power among the Anglo-Americans was exercised in favor of those who would be forced heirs under a mandatory system such as that in Spain; and (2) even where testamentary power appears absolutely undiluted, far more people die without wills than with them, resulting in the passage of their property to those who, under Spanish law, for example, would be designated forced heirs (see McKnight, 1996).

Spanish lawyers of the eighteenth and nineteenth centuries were trained in the Germanic tradition, which accepted only an extremely narrow testamentary power. Thus, while Anglo-American lawyers thought of both testate and intestate successions, the basic distinction for the Spanish was between those who were forced heirs and those who were not.

What is more, married women in the Spanish system not only had a much more extensive right to property than did those in Anglo-America, but also had will-making power, which married women did not have in Anglo-America. Married American women, for example, did not gain the power of will–making until the passage of the Married Women's Property Acts of the late nineteenth century.

In Spain, the surviving spouse was also entitled to half of the gains of the marriage under what is today termed 'the community property system.' Most of the property that was the decedent's alone passed to descendant forced heirs, despite any contrary testamentary provisions, and if there were no descendants, ascendants took as forced heirs, although in smaller shares than descendants. If the decedent died without a will, the estate passed to the forced heirs, or, if none, to other heirs as dictated by law (see McKnight, 1996).

French law also maintained forced heirship under the concept of legitim. During the French Revolution, this concept was viewed as a means of eliminating the feudal principle of primogeniture in order to break up feudal estates. There was some relaxation in the strict adherence to forced heirship by the time of the Napoleonic Code, which announced the general French view of succession that is held even today. Under the Code, half of the testator's estate (after marital property has been distributed to the surviving spouse) was freely disposable if there was one child, one-third if there were two children, and one-fourth if there were three or more children (see McKnight, 1996).

In general, since the early fifteenth century, the difference between Continental and Anglo-American views of succession, seen most strikingly in the Hispanic tradition, is that the Spanish adhered to the traditional Germanic tribal concept of familial right to property, whereas the Anglo-Americans took an individualistic view of who should succeed to the decedent's property (see McKnight, 1996). By the mid-nineteenth century, this English individualistic notion had triumphed throughout the United States, with the exception of Louisiana, which followed the French law of legitim. The Anglo-American system was, however, tempered by the concept of community property in states with a Spanish tradition and by dower and curtesy (see section Wills and Succession throughout History) in the others. Community property rules gave each spouse an immediate half interest in any property acquired during the marriage, while dower provided a surviving wife with a life interest in one-third of her dead husband's real property, and curtesy gave a surviving husband a life interest in all of the wife's real property.

Trusts

A particularly ingenious device that could be used for the intergenerational transfer of wealth was developed in Anglo-American courts of equity. Called the trust, it has sometimes been copied in other legal systems, such as that of modern Japan (see *International Law Digests*, 1999). This form of property holding enabled the decedent to pass legal title to a third party who held the property for the benefit of its intended beneficiary. This third-party trustee was subject to the equitable interest of the beneficiary, who generally received the income from the property for a period of time up to life. At the end of the designated period, the equitable title of that beneficiary would cease, transferring absolute ownership to that beneficiary or another chosen by the decedent.

The temporary division of legal and equitable title to property in trust allows, among other things, professional management of trust property assets. It also allows decedents to delay the actual vesting of full title to the property for a period of time up to and including a generation or more, during which period a beneficiary may receive a managed stream of income. Trusts in the United States that pay income to a decedent's children during their lives, and upon their deaths pay the principal to the grandchildren, are called

generation-skipping trusts. Estate tax advantages of these trusts (achieved because the children never received full title) were cut back by the Tax Reform Act of 1986 and subsequent amendments (which imposed a generation-skipping transfer tax; see Johnson and Eller, 1998; Dukeminier et al., 2009).

Wills and Succession among Nations in the Twentieth Century

The United States took freedom of testation one step further than did England and the countries of the British Commonwealth. While, in varying degrees, these nations developed flexible systems of family maintenance to protect surviving spouses and children (at least adult, competent children) from disinheritance, the United States maintained the right of the testator to disinherit his or her children completely, even those children who were minors and/or had need. Thus, while surviving spouses in the United States are protected either by the elective share (in 'common law' states following the English tradition) or by community property laws (in those states following the Spanish tradition), the protection of surviving children theoretically is left to the whims of the parent.

Although this disparity has been attacked, it has also been noted that flexible concepts such as lack of testator competence, undue influence, and 'unintentionally' omitted children indirectly protect surviving children to some extent in most 'common law' states (see Chester, 1997, 1998a).

That the United States's focus on the individual rights of the decedent, even to the exclusion of children and other descendants, is unusual can also be seen by examining the systems of the Soviet Union (before its dissolution) and China. Traditionally, the Soviet law of inheritance did not differ substantially from that of the civil law countries of Western Europe (see Dukeminier et al., 2009). Although the Bolsheviks attempted to change this during the early years of their takeover of power, this attempt failed. Nearly unique in the modern history of humankind, a 1918 Soviet Union law read: "inheritance, testate and intestate, is abolished. Upon the death of the owner his property (moveable and immovable) becomes the property of [the state]" (1 Sob. Uzak., RSFSR, No. 34, item 456, 26 April 1918). This attempt to abolish inheritance proved so unpopular with the people that inheritance was reestablished within four years (see Foster-Simons, 1985).

At present, inheritance law in Russia and the republics of the former Soviet Union generally follows the civil law views of Western Europe. In this way, it is similar to the law in modern China (see Dukeminier et al., 2009). In that country, if first-order heirs (spouse, children, and parents) inherit, then no second-order heirs (brothers, sisters, and grandparents) will inherit. Following a kind of community property system, all property in China acquired during marriage is joint property, unless the marital partners have executed a contract to the contrary (see *International Law Digests*, 1999). Since much of China is still rural and nonindustrial, custom plays an important part in the actual distribution of estates (see Foster-Simons, 1985), just as it does, for example, in tribal Africa. This distribution depends partly on the law and the terms of any will and partly on the social and economic facts relevant to the intestate, or the testator and his or her heirs (see *International Law Digests*, 1999).

Free Testation and Capitalism

It is tempting to view the relatively free testation of the United States as coincident with that country's status as the exemplar of capitalism. Except for the elective share provided for surviving spouses (e.g., one-third of the estate), testation in the United States is predicated on protecting the free will of the testator. While England and the Commonwealth countries are certainly capitalistic, their laws show, on the whole, more of a socialist's regard for protection of society (and family) than do those of the United States. While family protection is achieved in these countries by reposing a great deal of discretion in probate judges to ensure family maintenance (as distinguished from the use of fixed shares in the Continental system), this still provides a strong check on the whims of the testator. The governmental intrusion into testation is, of course, even more pronounced on the Continent, where fixed shares for the family are the norm.

In its 1987 decision in *Hodel v. Irving* (481 US 704, 1987), the US Supreme Court even seemed ready to protect *constitutionally* the right of an individual to transmit property at death in any manner that he or she might choose. While that decision has been criticized and not extended beyond its facts (which involved government confiscation of noneconomic land parcels descending at the death of American Indians), it symbolizes the United States' fierce protection of individualism, even after the individual's death (Chester, 1995, 1998b, 2007). Interestingly, the occasion for that decision was the government's attempt to save Indians from the disastrous economic effects of the Anglo-American individual property system that had been forced on the various tribes. Without imposition of the notion of individually held property in the first place, functioning Indian tribes, much like the tribes in sub-Saharan Africa, might have continued to treat property, which is chiefly in the form of land, as a family (and, by extension, community) asset that was never the individual's to transmit at death in the first place.

Implications of Inheritance and Taxation Policy for Equal Opportunity and the Distribution of Wealth

Allowing large amounts of capital to pass between generations without significant taxation is problematic for countries, such as the United States, that profess a belief in equal opportunity. Taxation of estates and gifts is widely seen as an intrusion on the freedom to transmit property at death – a freedom that Americans particularly seem to cherish. While the American estate and gift tax has high nominal rates, its high exemption levels (as of this writing, $5 000 000 per decedent, $10 000 000 per married couple, plus cost-of-living adjustments) and numerous loopholes, including a deduction of up to 100% for all property left to a surviving spouse, may make it in effect a 'voluntary tax' (see Cooper, 1979).

Countries such as France and Spain have both an inheritance tax and an annual tax on accumulated wealth. Perhaps not coincidentally, these countries do not have so strong a tradition as does the United States of allowing testators to do almost anything they want to do with their money at death.

Commentators (e.g., see Ascher, 1990, 2011; Chester, 1976, 1982, 1998a, 1998b, 2007) have noted that the policies of free wealth transmission and equality of opportunity coexist very awkwardly in the United States. Although legal systems cannot control differences in natural ability or the luck of being born into a family that values, for example, culture and education, they *can* curb that form of luck represented by inheritance. If equality of opportunity is truly a key goal, then children fortunate enough to have been reared and educated by wealthy parents need not be allowed the additional luck of freely inheriting their parents' property. Strong and effective taxation of inheritances would contribute significantly to this goal of equal opportunity.

Whatever the logic of this point of view, in the United States there appears to be little public support for more effective taxation on postmortem (or even lifetime) transmission of wealth to heirs. In fact, at the beginning of this twenty-first century, bills in Congress relating to the estate tax generally call for its softening or its abolition. Those supporting such bills argue that low or nonexistent taxation of inherited wealth is good for the economy because it preserves large blocks of investment capital in private rather than government hands. However, it might also be said that untrammeled inheritance saps the initiative both of those who receive it and of those who cannot possibly compete on equal terms with those who do (see Chester, 1976, 1982).

The Future

In looking toward the future of intergenerational wealth transmission in the twenty-first century, one might begin by considering the changing nature of the family itself. Particularly in advanced capitalist nations such as the United States, the traditional family is showing signs of disintegration. Traditional heterosexual marriages often end in divorce, and several forms of nontraditional relationships, such as same-sex marriage, are being legitimized. Since advanced capitalist nations offer many forms of wealth that are obtainable by individuals, there is no longer a critical need for children to succeed to a parent's land. Thus, the strict family protection practiced by the inheritance systems of many capitalist nations outside the United States may seem outdated to some.

American law, unlike that of many nations, does not directly protect children from disinheritance by their parents. By contrast, other nations may choose not to aggravate certain social trends, like the atomization of family and society, which many of them are already experiencing. Their common judgment may be that, even in the twenty-first century, some homage to the family, traditional or otherwise, still needs to be paid in the field of intergenerational wealth transfer. Whether the restraints on individualism that such a judgment entails will remain the norm internationally and whether they will become the norm in the United States remain open questions.

See also: Children and Law; Family Law in the United States; Gender and the Law; Intellectual Property: Law in Context; Intergenerational Justice.

Bibliography

Ascher, M.L., 1990. Curtailing inherited wealth. Michigan Law Review 89, 69–151.
Ascher, M.L., 2011. But I thought earth belonged to the living. Texas Law Review 89, 1149–1177.
Chester, R., 1976. Inheritance and wealth in a just society. Rutgers Law Review 30, 62–101.
Chester, R., 1982. Inheritance, Wealth, and Society. Indiana University Press, Bloomington, IN.
Chester, R., 1995. Is the right to devise property constitutionally protected? The strange case of *Hodel v. Irving*. Southwestern University Law Review 24, 1195–1213.
Chester, R., 1997. Should American children be protected against disinheritance? Real Property, Probate, and Trust Journal 32, 405–453.
Chester, R., 1998a. Disinheritance and the American child. Utah Law Review, 1–35.
Chester, R., 1998b. Inheritance in American legal thought. In: Miller Jr., R.K., McNamee, S.J. (Eds.), Inheritance and Wealth in America. Plenum, New York, pp. 23–43.
Chester, R., 2007. From Here to Eternity? Property and the Dead Hand. Vandeplas Publishing, Lake Mary, FL.
Cooper, G., 1979. A Voluntary Tax? Brookings Institution, Washington, DC.
Dukeminier, J., et al., 2009. Wills, Trusts, and Estates, eighth ed. Aspen Publishers, New York.
Foster-Simons, F., 1985. The development of inheritance law in the Soviet Union and the People's Republic of China. American Journal of Comparative Law 32, 33–62.
International Law Digests, 1999. Martindale-Hubbell, Summit, NJ.
Johnson, B.W., Eller, M.B., 1998. Federal taxation of inheritance and wealth transfers. In: Miller, R.K., McNamee, S.J. (Eds.), Inheritance and Wealth in America. Plenum, New York, pp. 61–90.
McKnight, J.W., 1996. Spanish legitim in the USA – its survival and decline. American Journal of Comparative Law 44, 75–107.
McMurray, O.K., 1919. Liberty of testation and some modern limitations thereon. Illinois Law Review 14, 96–123.
Rheinstein, M., Glendon, M.A., 1995. Inheritance and succession. In: The New Encyclopedia Britannica (Macropaedia), vol. 21. Encyclopedia Britannica, Inc., Chicago, IL. pp. 638–647.

Property Ownership

Frank Decker, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Property ownership can be understood as the ultimate residual right that has been left with the owner after other property rights such as possession have been granted to others. Providing the legal basis for the core operations of a credit-based monetary economy, property ownership allows individuals to enter into economic dealings by issuing claims to their property, while retaining possession or control of their assets. Historically, claims to property took the form of book debts, debt instruments and money and were enforced through elaborate legal systems. Networks of creditor–debtor relationships then began to regulate production, distribution and consumption via interest-bearing loan contracts and became the characteristic form of economic organization in ancient and later modern ownership-based economies around the world.

Introduction

Property ownership refers to the right of ownership over property. In contrast to possession, which governs the physical use, occupation and control of an object, ownership can be understood as the ultimate residual right left with the owner after other property rights, such as possession have been granted to others (Goode, 2004). The occupier of a house may, for example, retain possession of the house but transfer ownership to a banker as security for a loan. This conceptual separation of ownership and possession in turn provides the underlying legal basis for the core operations of any credit-based monetary economy such as credit sales, lending, leasing, the taking of loan security and the creation of money (*see* Property Rights and Urban Planning.).

Property ownership as a characteristic element of society is believed to have originated in the Greco-Roman world (Pipes, 1999; Bethell, 1998; Heinsohn and Steiger, 2013). After the fall of the Roman Empire, property ownership reemerged on a significant scale in the fourteenth century in England and – in line with the transformation and success of the English economy – was transplanted or adopted in many parts of the world. Unlike tribal and command systems which apply customary rules or coercive mechanisms, the members of ancient and modern ownership-based societies enter into contractual creditor–debtor relationships that regulate production, distribution and consumption and in which property ownership is the dominant feature (*see* Property Rights and Urban Planning).

In the wake of the Communist Manifesto, the concept of property ownership fell into intellectual disrepute and for many decades was viewed disparagingly by Western intellectuals (Bethell, 1998). As a result of this peculiar history, property ownership has suffered from a long period of theoretical neglect. It was not until the fall of communism made it clear that more than deregulated markets and foreign investment was required to build successful economies, that the economic significance of property ownership was fully recognized and placed within coherent theoretical framework (Heinsohn and Steiger, 1996, 2013; Stadermann and Steiger, 2001; Stadermann, 2006; Steiger, 2006; Decker, 2010; see also Bethell, 1998; Pipes, 1999; De Soto, 2000).

This article deals with the economic consequences of property ownership, which are the development of a system where claims to property (including money), ownership premium and interest become the core elements underpinning economic organization. The first section shows that, following an initial distribution of former feudal land holdings to its members, ownership-based societies were dominated by the issue of debt and debt enforcement. The second section explores the question of how claims to property were made quantifiable and ultimately settled, describing the emergence of money of account and settlement assets including coinage. The third section outlines how owners deployed their assets by creating claims to their property, documented in the form of book accounts, various debt-related instruments and importantly money. The fourth section describes how property ownership gives rise to the ownership premium and interest, outlining why modern ownership-based systems suffer from periodic crises but also tend to innovate and grow.

Claims to Property and Their Enforcement

Both the Roman civitas and the Greek polis are regarded as the prototypes of ownership-based forms of social and economic organizations that emerged in the Mediterranean and in many parts of the Near East. Ownership represented a radical, discontinuous and permanent shift from the preceding feudal organization (Finley, 1981; Heinsohn, 1984). It has been argued (Heinsohn, 1984) that following a series of cataclysmic events, revolutionaries divided the former feudal land holdings and created a new and free society, a 'government without a king and a democracy' (Plutarch, Theseus XXIV, 2–3). Each citizen now stood 'on an equal footing' (Plutarch, Theseus XXIV, 2–3) and outside the collective social safety networks of tribe (reciprocity) or feudal seigneurie (emergency rations). This meant that the giving of credit and the entering into debt became a necessary means of dealing with bad harvests, variability of yields and other emergencies, which could affect one land allotment but not another. This new type of credit, now occupying a central role in the ancient economy, was no longer based on the rules of reciprocity between helping friends or neighbors (Heinsohn and Steiger, 2013). In contrast, debts now became legally enforceable and the newly established systems of law

provided the creditor with powerful claims to the general property of the debtor or to the debtor as a person.

It is telling that the earliest legislation in ancient Greece and Rome extensively deals with the issue of credit and debt. Both Athens and Rome were confronted at an early stage with debt-bondage crises that 'involved a substantial section of the citizenry' (Finley, 1981: 160). Debt bondage or 'nexum' has been interpreted by some as a premonetary form of interest, where a debtor bound himself to a creditor to provide labor services during the period of the loan (Heinsohn and Steiger, 1996). If the loan was not repaid when due, the debtor remained in debt bondage and was bound to perform the same services until the eventual repayment of the loan (Finley, 1981). Solon's reforms as ruler of Athens (*c*.594 BC) were reported to have included the freeing of debt-bondsmen, the abolishion of debt bondage in the future and the cancellation of existing debts (Aristotle, Ath. Pol. 2.2; Finley, 1981). The latter was signified by the removal of mortgage stones ('horoi'). In Rome, the law of the 12 Tables (*c*.451/50 BC) set a maximum interest rate (Tables 8 and 18). Evidence indicating the periodic cancellation of debts and the release of people who were enslaved for debt also provide the prime examples of legislation in early Mesopotamia (Westbrook, 2003).

The Greco-Roman procedures for enforcing debts and claiming property were comprehensive. For example, in the law of Athens (*c*.435–322 BC), an action could be brought by the owner to claim property from another person. On non-performance an action for ejectment ('dike exoules') could be brought to seize the property due to him or other property of equal value. Hence, a debtor's property was always subject to a potential forfeiture until a debt was repaid (MacDowell, 1978). The Roman law of the 12 Tables (3, 1–6) outlines a treatment which was particularly harsh by modern standards ('manus iniectio'). If judgment was given against the debtor and payment had not been made within the required time, creditors could kill the debtor or alternatively sell him as a slave abroad (Kaser, 1996), thereby providing the debtor with a strong incentive to transfer ownership of his property assets or his services to the creditor. In later times, the execution against the debtor's property was introduced as an additional option, with that property being subsequently sold by public auction to pay for an unpaid debt or transferred to the creditor in settlement (Kaser, 1996; Crook, 1967).

Debt enforcement also remains of central importance in modern ownership-based systems. For instance, English law allowed creditors to imprison insolvent debtors until the debt was satisfied (the writ of 'capias ad satisfaciendum'), a practice that was only abolished (with some notable exceptions) by the Debtors Act of 1869. Thereafter, the seizure of the debtor's assets and the distribution of sales proceeds to creditors in proportion ('pari passu') to the quantity of their claims superseded the imprisonment of debtors as a result of private actions.

An alternative way to safeguard the recovery of property from defaulting or insolvent debtors is the taking of loan security. This typically gives the secured creditor priority over other unsecured creditors and the power to sell the asset identified as loan security and be compensated from the proceeds. Security rights are well evidenced in ancient Greece. About 250 'horoi' (mortgage stones) have been found and documented (Millett, 1985). These were placed on the land, inscribed with the particulars of the loan and provided public notice that the land was hypothecated. It is believed that possession of the property stayed with the debtor and that the creditor acquired ownership or a right to enter the property on default (Finley, 1951; Harrison, 1968). The law of Athens also used the pledge, where a thing (usually a movable) is handed over to the creditor as security for a loan (Harrison, 1968). Roman law security rights were the 'pignus' (pledge), where possession of the loan security was transferred to the creditor; the 'fiducia' (mortgage), where ownership was transferred to the creditor; and the 'hypotheca' (hypothec or charge), where possession and ownership remained with the debtor, but the creditor acquired a charge against the asset, which could be enforced against third parties (Kaser, 1981). In turn, Roman law has become the model for modern security law (*see* Property Rights and Urban Planning; Security: Commercial Transactions.).

Claims to Property: Their Quantification and Settlement

Entering into debt and the enforcement of debts as claims against the general property of a debtor raises two important questions. First, how can debts and consequential claims to inherently different property assets be quantified and be reduced to a common economic measure? Second, how can claims to property be ultimately settled?

The question of measurement of property is reflected in the development of abstract units for debts or monies of account (Steuart, 1767). Historically, these were often related to weight units as, for example, in the Babylonian system of 'shekel' (of c. 8.4 g), 'mina' (of approximately 502 g = 60 shekel) and 'talent' (at 60 'minas'). In the Greek system, the money of account was based on the system of 'obol' (derived from 'spit') and 'drachma' ('handful' of six spits = 6 'obols'). Both terms also meant weights. 'Drachmas' were then related to the 'mina' ('mna', 100 'drachmas' = 1 'mina') and 'talent' (=60 'mina') of the Babylonian system (Chantraine, 1979a,b,c).

When weight units are the basis of a money of account, the specification and settlement of a debt can be expressed in terms of weighing out a specific quantity of a commodity. This principle is also reflected in the Roman 'mutuum' (cash loan) requiring the weighing, counting or measuring of fungibles (Roby, 1902). For instance, an obligation of 1 'shekel' is settled with 8.4 g of silver, as a consequence of which the price of 8.4 g silver is set to one (silver) 'shekel'. The money of account is (silver) 'shekel', with the price represented by the number of units ('1') and not, as is commonly assumed, the quantity of silver (Stadermann, 2006; Decker, 2012). As in the case of the silver commodity itself, all other property assets are then assigned a price in numbers of (silver) 'shekels', found for example, in Ur III (*c*.2150–2000 BC) merchant balance sheets (Postgate, 1992). Silver attained a special role only in so far as it became the preferred asset to specify and settle debts ('settlement asset', Decker, 2010, 2012). Other commodities were used for this purpose, such as, copper in ancient Rome. In this way, claims to inherently different property assets were able to be specified as nominal obligations with an absolute price expressed in the designated money of account.

Final settlement could occur by delivery of a defined quantity of a commodity. However, the common money of account opened up the possibility of other modes of discharge. Mutual obligations could now be set-off or payment could occur by delivery of an alternative good with the same price if this was accepted by the creditor. Both forms of discharge could be affected in account books and did not require a physical settlement asset or a physical exchange of goods.

The final discharge of obligations was later refined by the use of coins instead of commodities as settlement assets (Decker, 2010). The invention of coinage around the end of the seventh century BC is attributed to the Lydians (at the time under Greek influence) and was rapidly adopted by Greek city states (Seaford, 2004). Coins can be defined as shaped objects with a sign or symbol indicating or implying the number of units in the money of account. Originally believed to have been issued by individuals, coinage later came under state control (Chantraine, 1979d). Obligations and prices continued to be expressed in the same money of account (e.g., 'drachmas'), but the discharge now occurred by counting out a corresponding number of coins (Seaford, 2004; for Rome see Harris, 2008b) rather than by weighing out a certain quantity of a commodity (e.g., silver). This reflects the 'fiduciary' nature of coinage: a coin is legally accepted at its stated unit expressed in money of account, even though its intrinsic (bullion) price may deviate from it (Seaford, 2004). Obligations could now be discharged based on the principles of 'nominalism' (Knapp, 1905), which means with designated assets which 'if added together according to the nominal value indicated thereon, produce a sum equal to the amount of the debt […] regardless of both their intrinsic and their functional value' (Mann, 1971: 76).

While the reasons for the introduction of coinage are considered 'irrecoverable' (Seaford, 2004; Harris, 2008a), state-issued coinage appears to have addressed situations where shortages in supply of silver caused periodic monetary crises. These could now be mitigated by the introduction of token coinage (where the price of the material is insignificant) or debased silver coinage, both issued by the state (Heinsohn, 1984; Heinsohn and Steiger, 2013; Seaford, 2004). State-issued fiduciary or token coinages represent an obligation of the state and are issued in payment of state expenditure and received back through taxes, fines and levies. Hence they are backed by the state's power to raise taxes. There is evidence of iron and bronze coinages, coins made from silver mixed with other base metals. However, most Greek coins of the Archaic and Classical period had a relatively high silver content. While circulating locally at a premium over bullion, a high silver content would have made them acceptable outside the boundaries of the polis, where they are assumed to have traded at the price of bullion (Kraay, 1976; Seaford, 2004; Heinsohn and Steiger, 2013).

Claims to Property as Book Accounts, Debt Instruments and Money

Settlement assets such as commodities or state-issued coins allow the final discharge of an obligation. State-issued assets are attractive because the final settlement between private individuals always requires an asset 'that is not a liability of the same debtor' (Decker, 2010: 184). The advantage is the finality of such transactions; the disadvantage is that property owners have to find ways to acquire the physical settlement asset, which may be in short supply or which requires the corresponding sale of an asset. Hence, in many instances property owners avoid transactions in physical settlement assets altogether by creating and documenting claims to their property, as is evidenced in the use of book accounts, debt-related instruments and money (Decker, 2010).

Claims to property or obligations are promises to pay a sum in the designated settlement asset specified in the money of account. These are backed by the general property of the issuer and rely on the separation of ownership and lesser rights such as possession. The property owner becomes liable for the issued claim under debt recovery law, which places a temporary burden on his property, most importantly his land allotment. Hence, the claim is enabled by ownership and leaves the physical possession of the property unaffected. For example, a debtor receives a loan of barley seeds, which the creditor and debtor document in an account book or as an acknowledgment of debt with a sum specified in X units of the designated money of account. The debtor remains in possession and continues with the agricultural use of the land. The loan is then repaid at harvest time by the delivery of a quantity of barley sufficient to yield a price of X units of the specified money of account plus the agreed amount of interest. This is again documented as an account entry (or as an acknowledgment of debt) offsetting the original obligation and removing the burden on the property. No coinage changes hands in this transaction.

At a general level, the majority of economic dealings does not produce one-off transactions but create circular, self-liquidating payment flows, where one transaction is likely to be matched by another off-setting transaction. This is the reason why claims to property are particularly useful in economic dealings. Consider, for example, the situation of a merchant who supplies producers with materials, but at the same time purchases their finished products. Similarly, a producer A, who has an account with merchant B and enters into an obligation with supplier C, can pay C by ordering B to create a book entry in favor of C, assuming C also has dealings with B (Decker, 2010).

Claims to property are also the underlying basis for the creation of money. Money, here defined as a special type of claim to property, is created when a creditor (typically a bank) issues or documents claims to its property (per note or book entry) as part of a credit contract with a debtor and frequently issues and redeems these claims (Heinsohn and Steiger, 2013). Hence a producer A, after receiving book credit with his banker B as part of a secured or unsecured loan can pay his supplier C by ordering B to create a book entry in favor of C, assuming C also has a bank account with B. Money in this definition originates from private transaction needs and is distinguished from state-issued coinage or debt. The latter have been characterized as 'state-issued means of payment' and are primarily created to meet the transaction needs of the state (Stadermann, 2006).

A common and important form of credit between Roman property owners utilized account entries. Since the early Republic, Roman citizens kept account books, the 'codex

accepti et expensi', also referred to as 'tabulae' (Kaser, 1971). Obligations were documented by making entries in the account book of the creditor ('litteris obligatio' or obligation by book entry). As additional evidence, a written acknowledgment by the debtor could be set up and the transaction could be passed through the accounts of intermediaries, who in this way acted as witnesses (Kaser, 1971). The resulting obligations were referred to as 'nomina' (names of debtors). 'Nomen facere' means to make an entry and consequently to make a loan. 'Nomen' refers to the name of the debtor (Roby, 1902). 'Nomina' could be transferred and were used as a means of payment; they were the most common mechanism for large purchases and could be the subject of a pledge (Harris, 2008b; Kaser, 1981). There was a market for 'nomina' (Harris, 2008b). Courts treated law suits on book entries as it treated suits on actual cash loans and 'nomina', like movables, animals and land were seized from insolvent debtors (Roby, 1902). At a more general level, obligations could be transferred and payment delegated ('delegatio'). Evidence showing the use of written acknowledgments of debt ('chirographa') can also be found (Jolowicz and Nicholas, 1972; Mitteis, 1894). To give two examples; an inscription that preserved the Ephesian Debt Law of 85 BC refers to 'cheirographa' ('notes of hand', Wankel, 1979); and fragmentary tablets were recovered from Herculaneum which refer to chirographa in the context of account books (Arangio-Ruiz, 1948). In later periods of the Roman Empire, written debt notes became the standard form of documenting loan contracts ('cautio', Kaser, 1981). Consequently, Roman property owners could advance credit or make payments per book entry or by written acknowledgments of debt and the resulting claims could be used to discharge or set-off other obligations.

This system was further enhanced by the development of banking, which began to occupy a central role in the ancient Greek and Roman economies. For example, Athenian bankers ('trapezites' who appeared in the fifth century BC; Bogaert, 1994) made payments on behalf of their customers and documented these transactions in their book accounts (Bogaert, 1994). The volume of bank lending is believed to have been extensive and bank accounts were widely held (Cohen, 2008). Commercially active customers are known to have held accounts with the same bank, increasing the likelihood of book entry set-offs occurring. For instance, the maritime traders at Athens had accounts with the bank of Pasion, Athen's largest bank (Cohen, 2008). Banking was also widespread in Rome. Roman bankers (in evidence from 318 BC; Andreau, 1999) documented operations that effected an account in a register ('rationes'), which was required by law to be produced at court when a client was involved in a law suit (Andreau, 1999). Bank clients could order their bankers to make payment to other parties and checks existed in some areas (Andreau, 1999). Bank advances created by book entry could be secured by mortgage ('fiducia') (Roby, 1902). There is evidence that ancient banks practiced 'fractional reserve banking', i.e., bank funds created through lending by book entry clearly exceeded the physical coin reserves (Cohen, 2008; Mrozek, 1985; Harris, 2008b; Bogaert, 1994). Consequently, it has been argued (Decker, 2010) that the book credits created in ancient economies by individuals and increasingly bankers as part of their lending activities represent one of the early instances of (ownership-based) money.

While there is still some debate about the economic role and significance of book accounts, debt instruments and bank deposits in the ancient economy (Harris, 2008b), overwhelming evidence of their importance is available from early modern Europe. Like their Roman predecessors, private individuals kept book accounts and periodically set-off debts against debts (Kerridge, 1988). Accounts were complemented by claims to property known as 'negotiable instruments' (Goode, 2004) including the bill of exchange (an order to another party to pay a sum of money) and the promissory note (a promise to pay a sum of money). Negotiable instruments were designed to be transferable to other parties. The underlying basis of bills and notes is property ownership: on nonpayment a holder can sue the parties to the bill or note and in the worst case have the parties' assets seized and sold through a judicial process and then be compensated from the proceeds.

Use of, and reliance on, book debt and negotiable instruments was extensive. For early modern England it has been estimated that the ratio of coinage to notes, bills and book debt was at least 1:20 (Kerridge, 1988), with manufacturers, shopkeepers and tradesmen 'involved in complex webs of debt and credit' (Wrightson, 2002: 293). It was common practice to set off debts against each other, with only the remaining balances settled in coin or with payment being even further deferred by writing out a promissory note (Decker, 2010). Out of this practice evolved a system that relied on factors and middlemen, for precisely the reason that it greatly increased the opportunities for these mutual set-offs to occur (Rogers, 1995). Bills, notes and book accounts were also extensively relied upon in the Colonies of the British Empire (see, for example, Decker, 2010, 2011; Michener and Wright, 2006). They also provided the much needed elasticity of funds in the take-off phase of the industrial revolution in Europe (North, 1994). A review of nineteenth century civil law cases in England, New South Wales and Connecticut show that at the time about 90% of all cases were debt related (Kercher, 1996).

In modern ownership-based societies, money predominately reemerged in the form of bank notes (promissory notes payable to bearer on demand) and bank deposits. These were issued or created by private entities, often merchants or industrialists who, based on their central role in the payment system, developed into 'bankers' (Pressnell, 1956; Lipson, 1956; Decker, 2010, 2011). The typical nineteenth century private note-issuing bank such as the Bank of England, the numerous English country banks (Pressnell, 1956), colonial banks (Decker, 2011) or the banks of New York (Gibbons, 1858) created money by discounting bills of exchange and promissory notes (i.e., the purchase of bills and notes at a discount to their face value) and by making secured and unsecured loans. The debtor received the discount or loan proceeds in the form of the bank's notes or as a deposit entry in the bank's book accounts, which became the standard means of payment. Bank capital was held to ensure that notes could be redeemed on demand and to provide a buffer for unforeseen loan losses. Consequently, the soundness of the private note-issuing bank was based on its capital and the quality of its loan assets, which were underpinned by the property of its debtors (Heinsohn and Steiger, 2013). Broadly

the same principles apply to modern central banking systems, where money is created via commercial banks that enter into secured loan contracts or repurchase agreements with a central bank (Heinsohn and Steiger, 2013).

Bank money becomes of central importance for an ownership-based system in three ways. First, debtors are freed from the obligation to redeem their own notes on demand; an obligation which is now performed by specialized bankers who are compensated with interest (Heinsohn and Steiger, 2013 see already Steuart, 1767). Second, bank notes allow the ready monetization of property assets, which turns these into generally accepted purchasing power without requiring the sale of the underlying assets (see already Steuart, 1767). Third, banks can set and stabilize property prices as measured in their own notes. This follows from the fact that in each discount or loan operation, a bank must specify a price of the underlying property assets in its own notes (Heinsohn and Steiger, 2013; Decker, 2013; Stadermann, 2006). By regulating the amount of bank money created and the rate of interest, a bank safeguards the purchasing power of its notes, thereby creating a price anchor for nominal obligations, which is indispensible in any ownership-based system. For example, nineteenth century banks throughout the world set the price of gold in their currency units, by purchasing and selling gold and commercial bills (Stadermann, 2002, 2006). Absolute prices for all other property assets follow by comparison from the bank determined prices for assets underlying loan and discount operations (Heinsohn and Steiger, 2013).

Ownership Premium and Interest

Following the initial distribution of property allotments, members of ownership-based societies use claims to their property to enter into economic dealings that regulate production, distribution and consumption. While property is subject to such claims – as for instance when assets are used as security for a loan or to back 'nomina', promissory notes, bills of exchange or bank notes – the potential claims of third parties place a burden on the property. The owner, while still in physical possession or control, is no longer able to freely dispose of the asset. A very important consequence of property ownership is therefore that any property asset that is unburdened and free, must carry a premium: the ownership premium (Heinsohn and Steiger, 2013). In turn, the loss of ownership premium must be compensated. The compensation takes the form of interest, which is a characteristic element of both ancient and modern ownership-based systems but is absent in tribal and command systems (Heinsohn and Steiger, 2013).

When a note-issuing bank creates notes for a debtor as part of granting a loan, the bank has to underwrite the issued notes with its capital. Capital is required as the bank remains liable for the full nominal amount of the issued notes and, in order to protect its solvency, must be able to absorb even unidentified and improbable loan losses. For the same reason, the debtor must provide loan security to the bank or provide an unsecured claim to his property (enforced by court order if needed) in return for the loan. Both parties lose ownership premium in this process, as the bank has to set aside (burden) its own capital to back the note issue, while the debtor has burdened his assets by way of loan contract. The bank's loss of ownership premium is compensated by interest from the debtor. The loss of ownership premium experienced by the debtor is compensated by the liquidity premium attached to the money notes, because they can discharge obligations at any time. Interest therefore compensates the loss of ownership premium and not, as is commonly assumed, forgone consumption or a loss of liquidity, which neither creditor nor debtor experience in this transaction (Heinsohn and Steiger, 2013).

The existence of interest implies that the amount of money repaid must exceed the money advanced as capital. Therefore, for the economy in aggregate 'it is always the case that less money is lent than owed' (Heinsohn and Steiger, 2013: 113). The need to create money via interest-bearing loan contracts is therefore extorting an ongoing pressure to create a surplus, by way of production or otherwise, that can be converted into additional money to service the interest on outstanding loans. Assume, for instance, that a set of loan contracts is rolled over for a second period, as producers seek to maintain their level of economic activity. In this case, the volume of money at the beginning of the second period, as is reflected in the sum of outstanding money-creating loan contracts, must be the sum of the original loan principals plus interest. Because the money required for the interest payments must also be created in loan contracts (here assumed to run over the same period), the money supply increases in each period by the aggregated amount of interest.

It has long been argued that an interest-induced increase in the money supply was a material issue for the Roman economy and led to periodic liquidity crises (Frederiksen, 1966; Mrozek, 1985). Because the supply of money in the form of 'nomina' and bank deposits was constantly growing (at a typical interest rate of 12% p.a. the money stock would double in about 6 years) an ever increasing gap occurred between the volume of money and the volume of state-issued settlement assets in the form of coins (Mrozek, 1985, 2001). As coinage was needed for the periodic settlement of balances, their increasing scarcity combined with uncertainty created by war, political crises or otherwise, periodically led to a collapse of credit markets and falling land prices. For instance, debt and liquidity crises impacted Rome in 193–192 BC, the 80s BC, in 63 BC, in 49 BC and in AD 33. Responses by Roman authorities included the cancellation of debts, staggering dates for repayment, interest-free loans, settlement in property rather than coin, measures to limit the hoarding of coins and measures to compel Romans to invest in Italic tribal land (Andreau, 1999; Mrozek, 2001; Frederiksen, 1966).

Liquidity crises continued to be a feature of modern ownership-based systems. In the nineteenth and twentieth century, the adverse impact of the elasticity limitations inherent in state-issued payment instruments led to the creation of central banks as lenders of last resort. This made money issued by central banks in loans secured by eligible property assets the leading settlement asset (Decker, 2010). The underlying principle of last resort lending is that everyone with sound loan security should have access to liquidity in a crisis. This principle emerged in response to the English crises of 1793, 1811 and 1825/26, where the Bank of England had to relax its standards for eligible bills and advanced its

notes on exchequer bills (secured by merchandise) and later directly on merchandise to stay the panic (Bagehot, 1873). The implication of last resort lending is that all principal property classes of an economy need to be convertible into central bank money in a crisis, once again emphasizing the property-based nature of money.

The interest-induced pressure to return a larger nominal amount of money than borrowed is also interlinked with the factors that drive economic development, growth, productivity and technical progress. In the case of a producer, the realized price of the production output, i.e., quantity times unit price, must be larger than the money advance and result in a profit (Heinsohn and Steiger, 2013). Because obligations are nominal, a surplus (for the economy as a whole) can be realized by a higher volume of production outputs or higher prices. The resulting inherent upward pressure on prices must be counteracted by a restrictive monetary policy (Stadermann, 2002). Consequently, producers must compete for bank money, while being exposed to fixed repayment obligations and uncertain sale proceeds. They will seek to outperform their competition by lower prices or the introduction of new and innovative products (Heinsohn and Steiger, 2013).

A special characteristic of the modern ownership-based system is the dominance of free wage labor. In contrast to the chattel slave, who was part of the production assets and could be mortgaged and sold, a wage labourer obtains a money wage. For the producer, the money advance required to pay wages must be raised beforehand and cannot, unlike equipment or slaves, be financed by using the wage laborers as collateral (Decker, 2013). Money expended on wages is also lost if the production outputs cannot be sold at the required price. Consequently, it has been argued that producers are constantly trying to reduce the level of wage labor, thereby creating a state of constant technical innovation, which has become a unique characteristic of modern ownership-based systems (Heinsohn and Steiger, 2013). This inherent tendency to generate productive growth has been reflected in unparalleled economic growth and prosperity in developed countries around the world. In contrast, the lack of access to formal ownership and security rights for the majority of people in developing countries has been identified as the key reason for their relative lack of development (De Soto, 2000). It follows from this that property ownership and the associated monetary system of production is the main factor for triggering economic development and not as is often assumed the inefficiencies arising from common as opposed to private control over resources.

While the ownership premium creates the demand for interest, it also reflects the readiness to burden property by entering into new loan contracts and therefore the willingness to invest. The ownership premium will rise sharply during a credit crisis or in an economic downturn as, for example, following a period of 'over-investment' due to an industry-wide innovation cycle, such as revolutions in transport, energy and information technology (Heinsohn and Steiger, 2013). When the ownership premium exceeds the expected returns on capital, owners will reduce the amount of property that is made available for the purposes of money creation. As owners try to protect themselves from asset losses, good loan security becomes rare in a crisis and central bank action is hampered because it cannot create the missing property required for debtors to obtain fresh loans (Heinsohn and Decker, 2010). In a deep downturn a redistribution of property ownership – at times facilitated through temporary state ownership, bail-outs and subsidies – may become the only way out of the crisis.

See also: Ancient Greece and Rome, History of; Capitalism; Colonization and Colonialism, History of; Feudalism; Property Rights, History of; Slaves and Slavery, History of.

Bibliography

Andreau, J., 1999. Banking and Business in the Roman World. Cambridge University Press, Cambridge.
Arangio-Ruiz, V., 1948. Les Tablettes D'Herculanum. Revue Internationale des Droits de L'Antiques I, 9–25.
Bagehot, W., 1873. Lombard Street. A Description of the Money Market. 1962. Richard D. Irwin Inc., Homewood, IL.
Bethell, T., 1998. The Noblest Triumph. Property and Prosperity through the Ages. St. Martin's Griffin, New York.
Bogaert, R., 1994. Banking in the ancient world. In: Bogaert, R., Kurgan-Van Hentenryk, G., Van Der Wee, H. (Eds.), A History of European Banking. Fonds Mercator, Antwerp.
Chantraine, H., 1979a. Drachme. In: Der Kleine Pauly, Band 2. Deutscher Taschenbuch Verlag, München. Col. 155–156.
Chantraine, H., 1979b. Gewichte. In: Der Kleine Pauly, Band 2. Deutscher Taschenbuch Verlag, München. Col. 791–793.
Chantraine, H., 1979c. Mna. In: Der Kleine Pauly, Band 3. Deutscher Taschenbuch Verlag, München. Col. 1368–1369.
Chantraine, H., 1979d. Münzwesen. In: Der Kleine Pauly, Band 3. Deutscher Taschenbuch Verlag, München. Col. 1447–1452.
Cohen, E.E., 2008. The elasticity of the money-supply at Athens. In: Harris, W.V. (Ed.), The Monetary Systems of the Greeks and Romans. The Oxford University Press, Oxford, pp. 66–83.
Crook, J., 1967. Law and Life in Rome. Cornell University Press, New York.
De Soto, H., 2000. The Mystery of Capital. Why Capitalism Triumphs in the West and Fails Everywhere Else. Bantam Press, London.
Decker, F., 2010. The Emergence of Money in Convict New South Wales. Metropolis, Marburg.
Decker, F., 2011. Bills, notes and money in colonial new South Wales 1788–1822. Financial History Review 18 (1), 71–90.
Decker, F., 2012. Settlement Assets: The Missing Economic Category in the Theory of Money. SSRN Working paper. http://papers.ssrn.com/abstract=2007637.
Decker, F. (Ed.), 2013. Ownership Economics. Routledge, London.
Finley, M.I., 1951. Studies in Land and Credit in Ancient Athens 500-200 B. C. Rutgers University Press, New Brunswick, NJ.
Finley, M.I., 1981. Economy and society in ancient Greece. In: Shaw, B.D., Slader, R.P. (Eds.). Penguin Books, Harmondsworth.
Frederiksen, M.W., 1966. Caesar, Cicero and the problem of debt. The Journal of Roman Studies 56 (1 and 2), 128–141.
Gibbons, J.S., 1858. The Banks of New York, Their Dealers, the Clearing House, and the Panic of 1857. Edition of 1870. D. Appleton & Co, New York.
Goode, R., 2004. Commercial Law, third ed. Penguin Books, London.
Harris, W.V., 2008a. Introduction. In: Harris, W.V. (Ed.), The Monetary Systems of the Greeks and Romans. The Oxford University Press, Oxford, pp. 1–11.
Harris, W.V., 2008b. The nature of Roman money. In: Harris, W.V. (Ed.), The Monetary Systems of the Greeks and Romans. The Oxford University Press, Oxford, pp. 174–207.
Harrison, A.R.W., 1968. The Law of Athens. At the Clarendon Press, Oxford.
Heinsohn, G., 1984. Privateigentum, Patriarchat, Geldwirtschaft: Eine sozialtheoretische Rekonstruktion zur Antike. Suhrkamp, Frankfurt am Main.
Heinsohn, G., Decker, F., 2010. A property economics explanation of the global financial crisis. In: Kolb, R.W. (Ed.), Lessons from the Financial Crisis, Causes, Consequences, and Our Economic Future. John Wiley & Sons, Hoboken, NJ.
Heinsohn, G., Steiger, O., 1996. Eigentum, Zins und Geld, sixth ed. 2009. Metropolis, Marburg.

Heinsohn, G., Steiger, O., 2013. Ownership economics. On the foundations of interest, money, markets, business cycles and economic development. In: Decker, F. (Ed.), Ownership Economics. (Trans. and ed. with comments and additions by F. Decker) Routledge, London.
Kaser, M., 1971. Das Römische Privatrecht. Erster Abschnitt. Das altrömische, das vorklassische und das klassische Recht, second ed. C.H.Beck, München.
Kaser, M., 1981. Römisches Privatrecht. Juristische Kurz-lehrbücher. C.H.Beck, München.
Kaser, M., Hackl, K. (Eds.), 1996. Das Römische Zivilprozessrecht, second ed. revised. C.H.Beck, München.
Kercher, B., 1996. Debt, Seduction & Other Disasters, the Birth of Civil Law in Convict New South Wales. The Federation Press, Sydney.
Kerridge, E., 1988. Trade and Banking in Early Modern England. Manchester University Press, Manchester.
Knapp, G.F., 1905. Staatliche Theorie des Geldes, fourth ed. 1923. Duncker & Humblot, Leipzig.
Kraay, C., 1976. Archaic and Classical Coins. Methuen, London.
Lipson, E., 1956. The Economic History of England, sixth ed., vol. III. Adam & Charles Black, London.
MacDowell, D.M., 1978. The Law in Classical Athens. Paperback edition 1986. Cornell University Press, Ithaca.
Mann, F.A., 1971. The Legal Aspect of Money, third ed. At the Clarendon Press, Oxford.
Michener, R., Wright, R.E., 2006. Development of the US monetary union. Financial History Review 13.1, 19–41.
Millett, P., 1985. The Attic Horoi reconsidered in the light of recent discoveries. In: Finley, I.M. (Ed.), Studies in Land and Credit in Ancient Athens 500-200 B. C. Transactions Inc., New Brunswick, NJ.
Mitteis, L., 1894. Reichsrecht und Volksrecht. Teubner, Leipzig.
Mrozek, S., 1985. Zum Kreditgeld der frühen Kaiserzeit. Historia: Zeitschrift für alte Geschichte 34 (3) (3rd Qtr.), 310–323.
Mrozek, S., 2001. Faenus. Studien zu Zinsproblemen zur Zeit des Prinzipats. Franz Steiner Verlag, Stuttgart.
North, M., 1994. Das Geld und seine Geschichte. C.H. Beck, München.
Pipes, R., 1999. Property and Freedom. Alfred A. Knopf, New York.
Postgate, J.N., 1992. Early Mesopotamia. Routledge, London.
Pressnell, L.S., 1956. Country Banking in the Industrial Revolution. Clarendon Press, Oxford.
Roby, H.J., 1902. Roman Private Law, vol. II. At the University Press, Cambridge.
Rogers, J.S., 1995. The Early History of the Law of Bills and Notes. First paperback edition 2004. Cambridge University Press, Cambridge.
Seaford, R., 2004. Money and the Early Greek Mind. Cambridge University Press, Cambridge.
Stadermann, H.-J., 2002. Das Geld der Ökonomen. Mohr Siebeck, Tübingen.
Stadermann, H.-J., 2006. Nominalökonomie. In: Stadermann, H.-J., Steiger, O. (Eds.), Allgemeine Theorie der Wirtschaft, Zweiter Band. Mohr Siebeck, Tübingen.
Stadermann, H.-J., Steiger, O., 2001. Schulökonomie. In: Stadermann, H.-J., Steiger, O. (Eds.), Allgemeine Theorie der Wirtschaft, Erster Band. Mohr Siebeck, Tübingen.
Steiger, O., 2006. Property economics versus new institutional economics: alternative foundations of how to trigger economic development. Journal of Economic Issues XL (1), 183–208.
Steuart, J., 1767. An Inquiry into the Principles of Political Economy. Printed for A. Millar and T. Cadell, in the Strand, London.
Wankel, H. (Ed.), 1979. Die Inschriften von Ephesos. Teil 1a. Nr. 1–47 (Texte). Rudolf Habelt Verlag, Bonn.
Westbrook, R. (Ed.), 2003. A History of Ancient Near Eastern Law. Brill, Leiden.
Wrightson, K., 2002. Earthly Necessities, Economic Lives in Early Modern Britain, 1470–1750. Penguin Books, London.

Property Rights and Urban Planning

Barrie Needham, Radboud University, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A property right gives someone the right to use a 'thing' – this case, land and buildings – in a particular way. The right is a legal relationship between the recognized user and all others who have to respect that use. Property rights are created in order to regulate the traffic between legal persons, and in that function they affect the way in which cities are formed and used. Urban planning necessarily restricts the way in which people may use their property rights, and for that reason is subject to political and economic analysis.

The Relevance of Property Rights for Urban Studies and Planning

People and organizations can own property rights in land and buildings. How they exercise those rights can affect the form of the city and the way in which it is used. It follows that the nature of the property rights and the legal context in which they are exercised is important for the life of the city. In addition, when public authorities practice urban planning, they usually have to restrict the exercise by the property owners of their rights. That makes urban planning a political activity, for people value the freedom which property rights can offer.

For those reasons, a good understanding of the nature of property rights is essential for urban studies and planning. Here, attention is paid to how property rights are created and can be changed, how they are embedded in a country's legal system, why they are important not just to individuals but also to the wider society (cities could not function without property rights), and the many ways in which they are restricted (and not just by urban planning). Moreover, it is shown that urban planning can be more effective and sensitive if it is based on a thorough understanding of property rights.

The Nature of Property Rights

There are many things which can be used in a variety of nonexclusive ways by different people. A house, for example, can be used for living in by someone who rents it from someone who uses it as a source of income and who, in turn, pays another person for the use of the ground on which the house stands. It is possible that the occupier sublets a room in the house to yet another person, or a shed in the garden to someone for use as a garage. Also possible is temporary use of the house, for example, just for the summer season. If one or a collection of those uses is protected by law – the person who exercises the use can by force of law prevent others from interfering with that exercise – that use is called a property right. The example shows that there can be more than one property right in one 'thing.' There are various types of 'thing' over which property rights can be exercised: immovable property (land and buildings), movable property (chattels), intellectual property (e.g., authorship of a book), and water (e.g., extracting water from a river). For this article, we restrict ourselves to rights in immovable property, alternatively called landed property, real estate, or real property.

It follows from the above definition that a property right is not a relationship between a person (the holder of the right) and a thing, but between persons: between the holder of the right and all others who are required to respect that right. This was stressed by Hohfeld (1913, 1917) when he pointed out that whenever there is a person enjoying a right, there is another person with an associated duty, namely to permit the first person to enjoy the right.

If this power is to be exercised effectively, it must be possible for the holder of the right to use force against someone who transgresses it. If, for example, a tenant has a formal right to occupy a room in a house and the landlord wishes to eject him for reasons outside the contract, the tenant must be able to seek protection against the landlord. In a state with a rule of law, the only entity entitled to use force is the state. So the plaintiff (the one whose right is transgressed) cannot himself force another to respect his right: instead, he is able to ask the court to take action against the defendant (the transgressor). This is crucial: an interest in a thing is a right only when it is protected by law. If the right is transgressed, it is for the holder of the right to take the initiative get it upheld. This is the essence of civil, or private, law. A plaintiff asks the court to take action against a defendant. A property right consists, therefore, not only of a description of the use which may be exercised, but also how it may be defended in law. The latter requires also that the nature of the use and the person of the user must be registered (for example, through a notary or in a cadastral register).

It follows from this that property rights are a social creation: there can be no property rights without agreement on what they entail (the task of the legislature) and without the possibility of enforcing them legally (the task of the courts). It is not that something is protected because it is a property right, but that something becomes a property right by being protected (Bromley, 1998: 25). Or "It is not property which is owned, but ownership which creates property" (Denman, 1978: 26).

In this respect, a right can be distinguished from a lesser interest in land. I have an interest in firm foundations under my house, and in many countries the state considers that interest to be so important that it gives me the 'right of support': if my neighbor digs a hole in her garden so big that it endangers the stability of my house, I can ask the court to stop her. I might have an interest also in the color in which my neighbor paints

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74032-3

her house: I find, for example, vivid pink disturbing. But the state might not consider that interest to be big enough to give me the legal right to proceed against my neighbor: that property right does not exist.

The Social Function of Property Rights

Property rights are created in order to regulate the 'traffic' between legal persons. Before someone builds a house on a plot of land, he wants to know precisely what his rights are to that land: for example, has another person a right of way over it, the right to repossess it after 50 years, or the right to discharge water over it in case of excess water upstream? Before a bank takes a mortgage on a house in return for a loan, it wants to know whether the mortgagee has given tenancy rights to another person. Clear and enforceable property rights reduce uncertainty, foster trade and good relationships, create a stable framework for long-term investments, and give continuity to a society. It is for this reason that throughout the ages states have created – or recognized existing de facto – property rights.

Although all states have created property rights, those rights are not the same in all states. In particular, most states recognize and protect the right of citizens and other private legal persons to own 'things,' but there are big differences between states in what the citizen is entitled to own. Socialist and communist states, for example, often restrict (or have restricted) the right of citizens to own land, or a house, or more than one house, or the means of production beyond tools and equipment for own use. In the Soviet Union, the right of ownership was defined in almost the same way as in market states: the big difference was in what could be owned (Honoré, 1961).

The Creation of Property Rights

There are no property rights without recognition and enforcement by law: property rights are created. It has been suggested that this creation follows certain regularities. Ellickson (1993) put forward the 'efficiency thesis' that 'land rules within a close-knit group evolve as to minimize its members' costs.' Demsetz (1967) put forward the thesis that when conditions change, resulting in valuable external effects which cannot be traded, property rights are changed so that those external effects can be internalized and thus realized. However, more recent work has cast doubt on such evolutionary laws. North (1990), for example, argues that property rights are made and changed by those with power, and that those people often want to preserve their privileges. He argues also that path dependency strongly influences the way in which legal rules change. Barzel (1989) argues, nevertheless, that economic analysis can help to explain why property rights change.

A country has often developed its formal legal system in close connection with its system of property rights: the wish to define and enforce property rights has influenced the legal system. Alternatively, a society may reform its legal system deliberately in order to change the existing system of property rights (as happened after the French Revolution and in the United States after the declaration of independence). Path dependency will influence the creation of those legal rights, and informal practices will influence the formal laws and how they are applied. It is for such reasons that it is common to talk of a country having a land system, which is an amalgam of formal rules and informal practices. In a mature country, this makes a consistent and coherent whole with the other aspects of the legal and political system. It follows that the land system and the property rights contained within it cannot be changed easily, apart from small modifications.

Those small modifications are part of the continuous adaptions to the law. For example, laws to regulate the relationship between landlord and tenant are often changed. Or it is ruled that a 'presumptive right,' such as the right to a view or the right to make a loud noise, should be taken away. Radical changes in property rights in land, such as land reforms, are rare because they would disrupt deep social and economic relationships. (It should be noted that the above applies to changes in the private law regulating property rights. The way in which property rights may be exercised can be changed by changes in the public law, such as for land-use planning or environmental policy. This is discussed later: restrictions on the exercise of property rights.)

When property rights are changed, the nature of the legal system affects the way in which that happens. Under English Common Law, for example, private parties might agree voluntarily on a new type of use. If one of the parties breaks that agreement, the other party might take the case to court. If the court judges in favor of the plaintiff, then a new property right has been created. (It is in this way that in some countries 'squatters' rights' have been created.) If, on the other hand, the law is codified, a new property right can be created only by changing statute law: it might be that the occasion for doing that is changing social practices, but those practices become legally recognized only after a legislative change. (For example, it was not possible in the Netherlands to acquire the transferable right to own an apartment: there was strong social pressure to introduce that; it was created by law in 1951.)

Varieties of Property Rights

There is an enormous variety of property rights in existence: apparently societies have found it useful to create such differentiation.

In some countries, there is the right to full ownership, which includes all possible uses of that thing. This is the Roman law concept of dominium, the right to use something, enjoy it, and dispose of it 'de la manière la plus absolue' (in the terms of the Code Civil Napoleon, 1804, article 54.4). This finds its expression in Dutch law also, where ownership is defined as 'the most comprehensive right that a person can have over a thing' (Civil Code, article 5.1). (It should be noted that the idea of making an exhaustive list of all possible uses is explicitly rejected, for additional possible uses might arise, which would nevertheless fall within the right of full ownership.) In addition, there are partial rights, which give the right to use the thing in just one or a limited number of ways. The owner of the full right can split

off partial rights and dispose of them. He who exercises a partial right is then not the owner, but the holder, of that right: he exercises it with the permission of the owner.

In countries with an Anglo-Saxon legal system, the starting point is not the full right of ownership, but partial rights: there are many separate property rights in the thing, each of which can be owned by a different person. There is full ownership if certain of these rights (a particular 'bundle of rights') are combined in such a way that they can be exercised by one person. This is fee simple, in Anglo-Saxon terms, and the rights included need not be comprehensive: certain partial rights can be owned and exercised by others.

The particular set of rights which constitute 'full ownership' varies from country to country (e.g., Kruse, 1939; Honoré, 1961). Nevertheless, it has been suggested that full ownership of a thing always includes:

- the right to exclude others from that thing (*usus*);
- the right to dispose of (to alienate) that thing (*abusus*);
- the right to the 'fruits' derived from that thing (*usus fructus*).

(The Latin words in brackets are from Roman law on property rights.)

Legal ownership of a 'thing' can be more limited than full ownership. It must include the right to dispose of it (perhaps only after a partial right has expired: it is a reversionary right) and it must be for an indefinite period.

Partial rights can give use of landed property in very many different ways (and see Stake, 2000):

- The right to a particular use of the land, where that use can coexist with other uses of the same land. This is an 'easement.' A fishing right is an example of this, as is a right of way.
- The right to use, temporarily, the whole parcel of land or part of it, to the exclusion of all others. If the right is to a part, that part can be separated from the rest horizontally (e.g., a field on a farm) or vertically (e.g., mining rights, or air rights). Leasehold falls under this type of right. If the duration is limited, it is specified to whom the right (the future interest) will revert at the end of the term. The duration can be 'in perpetuity,' in which case there is no future interest.
- The right to the 'fruits' of the land. The fruits might be income, or 'in natura' such as crops, or occupation.
- The right to 'alienate' the land or a right on it.
- The right to attach conditions on how the new owner/holder of the land may use it. A covenant is an example. Restrictive covenants (specifying what the new owner may not do) are more common than positive covenants (specifying what the new owner must do). Home owners' associations use positive covenants: someone buying a house is required to become a member of the home owners' association and to delegate to that association certain powers for managing the housing estate.
- The right to use, but only for a specified period. Easements are often in perpetuity, leaseholds usually for a limited period, as is 'usus fructus.'
- The right to split off one or more of the possible uses and to transfer this to someone else. An example is the right of a tenant to sublet part of the premises.
- The right is to be shared and jointly exercised with others (co-ownership). For example, it can be arranged that spaces which are used by all those living on a housing estate (a car park or green areas) be owned jointly by all the residents.
- The right to exercise the use without interference from owners/holders of rights on other pieces of land. (This might be realized differently, under generic law concerning nuisance or good neighborliness.)
- The right is settled on the land or on the person holding/owning the land. An important distinction, derived from Roman law is between material rights, or 'rights in rem,' and contractual rights, or 'rights in persona.' The former are attached to the thing irrespective of the (legal) person exercising the rights: they 'run with the land.' Therefore, they can be freely traded. The latter are not attached to the thing but may be exercised only by the (legal) person specifically named in connection with that right: they cannot, therefore, be traded. For example, in some countries a tenancy lease can be traded, but a rental agreement cannot. A contractual right usually ends when the named person leaves voluntarily, or dies, or at the end of the contractual period.

Often, another type of distinction is made, namely between:

- the rights of private ownership;
- the rights of shared ownership; and
- the rights of public ownership.

This, however, is not useful. All property rights are 'private' in the sense that the holder can go to court to protect his right. A municipal authority, for example, has a 'private' right to occupy the offices it uses, and in principle all legal persons can hold any property right, unless the law of the country forbids it. For these reasons, it is better to talk of different ways of holding, or managing, property rights. Bromley (2004) calls these 'property regimes.' He uses Hohfeld's concepts (see earlier text) and distinguishes between:

State property: Individuals have the duty to observe the use and access rules determined by the controlling or managing state agency. The agency has the right to determine the use and access rules.

Private property: Individual owners have the right to undertake socially acceptable uses, and have the duty to refrain from socially unacceptable uses. Others (nonowners) have the duty to refrain from preventing socially acceptable uses and the right to expect that only socially acceptable uses will occur.

Common property: The management group (the owners) has the right to exclude nonmembers, and nonmembers have the duty to abide by the exclusion. Individual members of the management group (the co-owners) have both rights and duties with respect to use rates and maintenance of the thing owned.

Nonproperty: No defined group of users or owners, and the benefit stream is open to anyone. Individuals have privileges and no rights with respect to use rates and maintenance of the thing. That latter is an open-access resource.

The great advantage of such a classification is that helps to avoid much careless thinking about the 'tragedy of the commons' (Hardin, 1968). Nonproperty – open-access

resources – can indeed by overexploited. But there is no reason why common property should be, if it is well managed by its members. Ostrom (1990) has done particularly useful work in analyzing the property rights necessary for managing 'common pool resources' (as she calls them) to great social advantage.

Restrictions on the Exercise of Property Rights

The exercise of property rights is always limited. The person with full ownership of land might wish that 'it is my land and I can do with it what I like,' but in practice the legislature always restricts that exercise. Again we take the example of Dutch law (Civil Code, article 5.1) where the definition of full ownership quoted above is supplemented by:

- "the owner is entitled freely to use the thing and to exclude others from that thing, as long as this use does not infringe the rights of others, nor the constraints imposed by written and unwritten law" and
- "the owner of the thing is owner also of the benefits arising from that thing, apart from any rights that others may have."

The restrictions arise partly from the contents of the right itself. Apart from the right of full ownership (which might include all possible uses, known and unknown – see earlier text), all other property rights are partial and therefore restricted. Sometimes these restrictions are entered into voluntarily. For example, if the owner of a freehold interest in land sells a ground lease on it, he has limited himself contractually: he may not use his land in other ways during the period of the ground lease, and the lessee may use the land only in accordance with the terms of the lease. Such restrictions can be enforced under civil, or private, law. Other restrictions which can be enforced under civil law are not entered into voluntarily, such as nuisance or 'good neighbor' laws, including certain laws about water extraction or discharge.

Another type of restriction is imposed by the state. For example, the owner/holder of a right on land (even the right of full ownership) might not be allowed to discharge poisonous fumes or polluted water. Under planning legislation, the restriction is often imposed that the land owner may change the use of that land only in certain ways: for example, the owner of a dwelling house may change its use to a small professional office, but not to a café. Under building legislation, there is usually the restriction that nothing may be built without first acquiring a building permit.

The most drastic restriction on property rights imposed by the state is expropriation: the state may acquire the full ownership, or a partial right (e.g., nationalization of development rights in England and Wales in 1947) against the wishes of the owner. It is in this connection that the famous statement in the French Declaration of Human Rights must be understood: that property is 'inviolable et sacré.' This was intended to protect property owners against arbitrary expropriation and against expropriation without proper compensation.

The restrictions imposed by the state are regulated differently from those included in the terms of the property right. In the latter case, a transgression of the right (a breaching of the restriction) is enforced only if a legal person asks the court to take action: otherwise, the state does not intervene. Where the restrictions are imposed by the state, it is the state itself which takes the initiative if the restrictions are breached. Such restrictions are imposed and enforced under public, or administrative, law.

When the state imposes a restriction on the exercise of a property right, this takes away some of the rights of the owner/holder. Then the politically important question arises: Should the owner be compensated for that 'taking'? (Note that the question of compensation for taking does not arise when restrictions are entered into voluntarily under private law, not even when restrictions are imposed under private law, such as nuisance or good neighbor laws.) The considerations are the following. If I own a firearm, my use of it is restricted: I am not allowed to shoot people with it. Similarly, if I own a car, my use is restricted, for I may not drive at more than 50 km h^{-1} in a built-up area. In both cases, some of my rights have been 'taken away' by the state. But no one would suggest that I should be compensated for that loss. However, if the state takes all my rights (expropriation) then compensation should be paid (unless there is a legally justified 'confiscation'). But what if the state takes away just some of my rights?

In the United States this is discussed with the concept of 'police powers': these powers can be used in the interests of 'public health, safety and morals' without any compensation to those who are restricted. In that country, it has been decided that, when the owners of landed property are restricted in the exercise of their powers for land-use or environmental reasons, this is in the interest of 'public health, safety and morals.' As a result, those restrictions are not a 'taking.' The exception is when such restrictions go 'too far,' such as removing all economically viable use. That latter is a taking under the terms of the US Constitution, fifth amendment, which stipulates "nor shall private property be taken for public use, without just compensation." Under English law, compensation for owners of landed property can be demanded when restrictions mean that there is no more 'reasonable beneficial use.' In the Netherlands, a similar rule applies. There in addition, under the principle that public regulations which bear harder on some than on others should be compensated – égalité devant les charges publiques – owners can demand compensation when a plan is changed and this causes downzoning.

The Politics of Property Rights

Because property rights are a social creation (see earlier text), they can be evaluated socially, with a view to changing them in some desired way. Because this requires the legislature to change the law, this is a task for politicians.

This is not a new insight: property rights have always attracted political attention. One of the first and best-known justifications for private ownership of property was made by John Locke (1632–1704): one had a right, he said, to ownership of that which one had appropriated by one's own labors. As a corollary, he said that the task of government was to protect private property. This idea had a great

influence on those who drafted the US Constitution. "The moment the idea is admitted into society that property is not as sacred as the laws of God, and that there is not a force of law and public justice to protect it, anarchy and tyranny commence," said John Adams in 1790 (quoted in Jacobs, 2012: 74–75). The total rejection of this attitude – the anarchic 'property is theft,' so should be owned collectively, put forward famously by Proudhon (1809–65) – is not held widely, for it contradicts the centuries-old experience that property rights are useful for regulating social and economic relationships.

Although justifications for property ownership have changed greatly since those classical statements (see for example Becker, 1977; Munzer, 1990), the social value of private ownership of property is still regarded as important. It is now incorporated in the Protocol of the European Convention of Human Rights (article 1): "Every legal person is entitled to the peaceful enjoyment of his possessions. No one shall be deprived of his possessions except in the public interest and subject to the conditions provided for by law and by the general principles of international law." But this is far from giving an absolute right. That article continues: "The preceding provisions shall not, however, in any way impair the right of a State to enforce such laws as it deems necessary to control the use of property in accordance with the general interest or to secure the payment of taxes or other contributions or penalties."

In the current debate on the politics of property rights, the extreme positions can be described as: the right to full ownership should be exercised only by private legal persons and that right should be restricted only by civil/private law, and; property rights should be owned and exercised as much as possible by public authorities.

Those who argue that property rights should be owned and exercised as much as possible by private legal persons and with as little interference as possible by public authorities (that is, without restrictions imposed under public law) are moved by a distrust in government and by a belief that if property rights can be freely exchanged this will be good for economic welfare. Also, these proponents hold high the liberal value of individual freedom: people should be able to use and transfer property rights voluntarily (under minimal conditions of the law of property, tort, and contract – Hayek, 1989). They see private ownership of property as a protection against misuse of state powers. This latter argument is put forward not only by those with much property to defend but sometimes also by those with little: if 'the little man' owns defensible property rights, this can protect him against abuses of power by the rich and the powerful. This latter consideration was strong in the revolt against the 'ancien regime' in France at the end of the eighteenth century, and in the decision by many landless peasants in Europe to migrate to the United States in order to acquire land there.

Those who argue that property rights should be owned and exercised as much as possible by public authorities do so partly as a reaction against the abuse by some big private landowners of their powers (as in England, Scotland, and Ireland in the nineteenth century, also concerning the haciendas in Spain and parts of South America). There is still much misery caused by the exclusionary effects of property rights on land with, in the extreme, people without such rights being unable to get food, shelter, sometimes simple living space (Davy, 2012). In addition, there is an argument from welfare economics, namely that a free market will not produce the highest welfare if there are imperfections, such as external effects, natural monopolies, public goods: that there are very many such imperfections inherent in land and buildings; and that a public authority is able to remove or correct for those 'market failures.' The classical exposition of this argument was made by Pigou (1932). A public body can intervene in the market for property rights by, for example, restricting the types of uses to which the owner may put his land (in order to reduce negative external effects), by expropriating the ownership right (where there is a natural monopoly given by location), and by public ownership to provide public goods such as parks and open space.

This argument for public intervention derived from welfare economics has been weakened considerably by experience with large-scale public ownership of land and buildings, in particular in communist and socialist states. This has not always realized the benefits hoped for: instead there has been inertia, lack of incentives, corruption, gross mistakes, and the citizens who should have been most helped – the poorest and weakest – have not always been favored. Proponents of free markets sometimes talk of this in terms of 'public failures' when public authorities try to correct for market failures (see Wolf, 1979).

The argument that a public authority should correct for market failures has been weakened also by theoretical work. It is clear that the market in property rights for land and buildings does not always produce what people want: for example, who is to build roads which connect the parcels of land in private use, if the government cannot get hold of the necessary land? But perhaps such failures can be tackled without direct public intervention. In 1960, Coase argued that market failures arise mainly because of transaction costs and because property rights are not clearly delineated. Incomplete separation of rights creates common property, i.e., in the public domain (Barzel, 1989) and, if these rights are valued, resources are spent on capturing them. This can lead to rent-seeking and 'common property dissipation.' According to this argument, the problem of market failures can be tackled by not only Pigovian measures but alternatively by public measures to enable markets in property rights to function well, to reduce transaction costs, and to delineate exclusive property rights (Webster and Lai, 2003).

Pursuing this line of argument further, it is not useful to think in terms of the contrast between a free market in property rights on the one hand, and a strong public regulation and exercise of those rights on the other. Instead, we recognize the social arguments for strong property rights exercised by private legal persons (the classical liberal standpoint), also that exercise cannot be unconstrained or exclusive, and we look for government measures which enable the market in property rights to work better. This – which has been called proprietary governance (Pennington, 2002), or the property rights model – can be done by structuring markets in property rights in such a way that private actors working through those markets involuntarily produce socially desirable results. This approach has received considerable attention, especially as an

alternative to zoning land use (e.g., Ellickson, 1993; Needham, 2006). If structuring markets in that way cannot produce the required results (because, for example, transaction costs are high and cannot be reduced sufficiently, as is the case with roads in urban areas), then direct public actions can be justified.

The Economics of Property Rights

When a property right can be traded (and this does not apply to rights which are inalienable, nor to 'rights in persona' – see earlier text), it will have a market value. Economic theory can, thus, be applied to explain the prices of, and transactions in, property rights. The application to rights in real estate is common: real estate economics (see e.g., Harvey, 1987; Jaffe and Sirmans, 1989).

A more radical combination of economic theory and property rights is when existing economic theory is modified by incorporating into it property rights (as institutions). The work of Barzel (1989) is a good example of this. His starting point is that most exchanges of goods and services can be interpreted as the exchange of property rights over an asset. However, property rights are rarely full delineated, and obtaining the missing information costs time and money: the transaction costs can be high. Barzel introduces institutions into Walrasian (classical, micro-) economic theory. This he calls "the property rights transaction costs model" which can give a "better understanding of the allocation of resources and of the interaction of this allocation with the legal system" (op. cit. p. 117). In this way, the economics of property rights has become one strand in New Institutional Economics (see, e.g., Cooter and Ulen, 2004).

See also: Law and Economics; Old and New Institutionalism in Economics; Real Estate Development; Transaction Costs and Property Rights.

Bibliography

Barzel, Y., 1989. Economic Analysis of Property Rights. Cambridge University Press, Cambridge, MA.

Becker, L.C., 1977. Property Rights: Philosophical Foundations. Routledge and Kegan Paul, Boston, MA.

Bromley, D.W., 1998. Rousseau's revenge: the demise of the freehold estate. In: Jacobs, H.M. (Ed.), Who Owns America: Social Conflict over Property Rights. University of Wisconsin Press, Madison, pp. 19–28.

Bromley, D.W., 2004. Property rights: Locke, Kant, Peirce and the logic of volitional pragmatism. In: Jacobs, H.M. (Ed.), Private Property in the 21st Century: The Future of an American Ideal. Edward Elgar, Cheltenham, UK, pp. 19–30.

Coase, R.H., 1960/1988. The problem of social cost. In: Coase, R.H. (Ed.), The Firm, the Market and the Law. University of Chicago Press, Chicago, pp. 95–156.

Cooter, R., Ulen, T., 2004. Law and Economics, fourth ed. Addison-Wesley, Reading, MA.

Davy, B., 2012. Land Policy: Planning and the Spatial Consequences of Property. Ashgate, Farnham.

Demsetz, H., 1967. Toward a theory of property rights. American Economic Review 57, 347–350.

Denman, D.R., 1978. The Place of Property. Geographical Publications Ltd, Berkhamstead.

Ellickson, R.C., 1993. Property in land. The Yale Law Journal 102, 1315–1400.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Harvey, J., 1987. Urban Land Economics: The Economics or Real Property, second ed. Macmillan, London.

Hayek, F., 1989. Spontaneous (grown) order and organized (made) order. In: Modlovsky, M. (Ed.), Order – with or Without Design. Centre for Research into Communist Economies, London, pp. 101–123.

Hohfeld, W.N., 1913. Some fundamental legal conceptions as applied in judicial reasoning. The Yale Law Journal 23, 16–59.

Hohfeld, W.N., 1917. Fundamental legal conceptions as applied in judicial reasoning. The Yale Law Journal 26, 710–770.

Honoré, A.M., 1961. Ownership. In: Guest, A.G. (Ed.), Oxford Essays in Jurisprudence. Clarendon Press, Oxford, pp. 107–147.

Jacobs, H.M., 2012. Talking about property rights over tea: discourse and policy in the US and Europe. In: Hartmann, T., Needham, B. (Eds.), Planning by Law and Property Rights Reconsidered. Ashgate, Farnham, pp. 71–96.

Jaffe, A.J., Sirmans, C.F., 1989. Fundamentals of Real Estate Investment. Prentice Hall, Engelwood Cliffs, NJ.

Kruse, V., 1939. The Right of Property (P.T. Federspiel, Trans. from Danish). Oxford University Press, Oxford.

Munzer, S.R., 1990. A Theory of Property. Cambridge University Press, Cambridge, MA.

Needham, B., 2006. Planning, Law, and Economics: The Rules We Make for Using Land. Routledge, London.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge, MA.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, MA.

Pennington, M., 2002. Liberating the Land: The Case for Private Land-Use Planning. Hobart Papers No. 143. Institute of Economic Affairs, London.

Pigou, A.C., 1932. The Economics of Welfare, fourth ed. Macmillan, London.

Stake, J.E., 2000. Decomposition of property rights. In: Bouckhaert, B., de Geest, G. (Eds.), Encyclopedia of Law and Economics. Edward Elgar, Cheltenham, pp. 32–61.

Webster, C., Lai, L.W.C., 2003. Property Rights, Planning and Markets. Edward Elgar, Cheltenham.

Wolf, C., 1979. A theory of nonmarket failure: framework for implementation analysis. The Journal of Law and Economics 22, 107–139.

Property Rights, History of

Frank Decker, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Property is the class of things that is capable of being owned. Of greatest significance historically were land, cattle, and humans. The most important property rights are ownership, possession, and security. While possession is encountered in various forms throughout history, ownership only emerged in the Greco-Roman world and is absent from tribal communities and feudal/socialist command systems. After the fall of the Roman Empire, an ownership-based system on a significant scale began first to reemerge in England in the fourteenth century and was adopted through colonization, conquest or reform across large parts of the world.

Introduction

Commercial transactions in any modern developed economy with a legal system grounded in common law or civil law are concerned with dealings involving property and property rights. The sale of goods, the lease of equipment or land, the issue and transfer of negotiable instruments, the grant of security in support of loans, the creation of central bank money, and the enforcement of debts are all based on property rights, which create the underlying basis of economic activity and business operations.

The term 'property' is ambiguous but it may be explained as the class of those things ('res' in Roman law) that can be subject to property rights and consequently are capable of being owned. In turn, the term 'property rights' can be used to identify the relationship between an individual and a thing and – most importantly – the effect of that relationship on the rest of the world including the state (Bridge, 2002).

There are a large number of different property rights that can exist with respect to a thing and the details of how such rights are implemented vary across legal systems. The most important examples of property rights are possession, ownership, and security (e.g., hypothecation) (Goode, 2004). Property rights are distinguished from contractual rights in that they can be enforced not just against a particular individual but against the world at large and furthermore they survive the insolvency of the counterparty (Goode, 2004). Different property rights can coexist in respect of the same thing and may be held by different people: the ownership of a house can remain with A while hypothecated to bank B as a consequence of B having lent funds to A to purchase the house. At the same time, possession (occupation and control) of A's house may be with C, who has rented the house from A and lives in it during the period of the loan.

This article is structured as follows. The first part explains the four building blocks of a developed property rights system – property, possession, ownership, and security. The concepts require some detailed discussion, as a persistent confusion over their legal significance exists in the literature. The terms 'property,' 'ownership,' and 'possession' are often used interchangeably and undue weight is given to the rights of use and control, which are associated with possession. Security – of critical economic importance – is rarely discussed. Hence, this part places a particular emphasis on clarifying the legal terminology, drawing out the distinction between ownership and possession and explaining security rights. The second part of the article provides a discussion of the origins and historical development of property rights, covering tribal communities, feudal/socialist command systems and ownership-based systems, including Greco-Roman and modern systems. It reveals that ownership-based systems have only existed for a relatively short period of time and that ownership – contrary to common belief – is not a universal characteristic of human organization.

Property (Things)

The term 'property' is derived from the Latin word 'proprietas,' the Roman law term for 'dominium' or ownership (Kaser, 1981; Pollock and Maitland, 1898). Under English property law, the term 'property' can however signify both the right of ownership (Roman law 'dominium') and the class of things capable of being owned (Roman law 'res'). This twofold meaning in terminology introduces considerable complexity, making it very important to distinguish between each of the meanings.

Property, in the sense of things that can be owned, typically includes a wide range of objects, of which historically the most important were land, slaves, and cattle. The term also includes other movable tangible objects such as ships, goods, plant and equipment, legal tender bank notes, and coinage as well as intangible property such as debts and shares in a company. However, not all things can be subject to property rights. In modern economies humans are, for example, no longer considered property and hence are incapable of being owned, sold, or used to secure loans. Other examples are the air and the stars. It is not always easy to determine what individual legal jurisdictions regard as property, particularly when the property is intangible in nature.

Important classes of property in classical Athens (around 400 BC) were land, houses, coinage, and slaves. Also capable of being owned were cattle, ships, vehicles, clothes, jewelry, and the like (MacDowell, 1978). In Roman law, an important category of property (classified as 'res mancipi') consisted of Italic land, slaves, beasts of draught, and burden (oxen, horses, asses, mules), as well as rights of way and water (rustic praedial servitudes). These had to be formally conveyed (in the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62150-5

presence of witnesses) when there was a change in owner; mere delivery of the thing was ineffective. In addition, most tangible objects such as grain, oil, bricks, food, fuel, coinage, clothes, animals, and agricultural produce were property (Nicholas, 1975; Kaser, 1981).

English law (Bridge, 2002; Goode, 2004) classifies property into real and personal property. Real property is land while personal property comprises all things capable of being owned once land is excluded. Tangible, movable things like goods, wares, and merchandise are commonly referred to as chattel property or chattels (derived from the word cattle, Pollock and Maitland, 1898). Historically, slaves were deemed chattel property in the British Colonies; slaves in the West Indian plantations could, for example, be sold and mortgaged in the nineteenth century. Another important class of personal property is intangible property. Examples are debts, shares in a company, promissory notes, bills of exchange, and intellectual property. Intangible property cannot be physically possessed. Nonetheless, documentary intangibles (that is, debts embedded in a document) such as promissory notes and bills of exchange have some of the characteristics of a chattel because they can for practical purposes be equated with tangible goods and can be subject to security rights that require transfer of possession. An important example is negotiable instruments, which can be pledged (see below).

Possession

The term 'possession' is derived from the Latin word 'possessio,' a Roman law term, which in early Roman times is believed to have signified land on which someone is 'sitting' (Kaser, 1981). Legally, possession is a right very distinct from ownership, whereas in common language possession is frequently used when ownership is the more appropriate term. As the eminent English jurists Pollock and Maitland (1898, Vol. 2. Chapter IV: Ownership and Possession) pointed out: "When a man says 'I possess a watch,' he generally means 'I own a watch.'" Similarly, property in the sense of things capable of being owned is often misleadingly referred to as 'possessions.' Conversely, the terms 'ownership' and 'property' are frequently used when the underlying subject matter is occupation, use and control, and reference should rather be made to possession. This confusion between ownership and possession is widespread in the social science literature, in particular economics (Steiger, 2006).

Although possession is incapable of a precise legal definition, it can be understood as the factual occupation, physical control, use, and enjoyment of property with the intent to exercise this control against others (Goode, 2004). Examples include not only the owner and occupier of land and the tenant of a building but also the thief in control of a stolen handbag. While this concept of possession may appear straightforward, the underlying details are of surprising complexity. The two main issues are, on the one hand, the extent of actual occupation or physical control required to obtain or maintain possession and, on the other, the identification of the circumstances in which the person who actually physically holds the thing is not considered legally as the possessor.

What amount of occupation or physical control constitutes possession depends on the circumstances of the actual case, the degree of control the nature of the thing permits, and the degree of control a possessor would customarily be expected to exercise over the thing. Legally, possession can still be maintained even though the thing is no longer within the actual immediate physical control of the possessor. Under English law, a person is considered to have possession of those movables, which not only are within his sight or reach, but also beyond his immediate control such as the books in the room next door and implements left in a field. A person remains in possession of the tables and chairs inside the house he occupies whether he is in the house or not. The same can be said of the chair left outside the house on the lawn (see the examples given by Pollock, 1888). Similarly, someone who finds a purse of gold and leaves it in his country house while serving a prison sentence a hundred miles away is considered as being in possession until a burglar takes the purse and by this act acquires possession (Bridge, 2002; Holmes, 1881).

Conversely, a person who actually holds the thing may legally not be the possessor. Examples are servants, agents, or employees, who have the use or custody of things but not possession. In a situation where A gives his coat to his servant B to take it to the tailor for alteration, it turns out that although B has physical control, A is still in possession (Pollock, 1888). Similarly, the employee sitting at the office desk using the office computer is legally not considered the possessor of the desk and computer, which remain in the possession of the employer (Bridge, 2002). Under German law, a person resting on a park bench is not considered in possession of the bench. Similarly, a dinner guest is not considered in possession of the chair or the cutlery he uses (Schapp and Schur, 2002; Wolff and Raiser, 1957). Another example is the relationship between landlord and tenant. Under English and German laws a tenant has possession, reflecting the actual occupation and control of the leased property in question; in Roman and French law, a tenant is merely the 'détenteur' and the landlord remains in possession (Nicholas, 1975; Goode, 2004).

Legislation can further complicate these relationships. An insightful example is the *Lien on Wool and Stock Mortgage Act (NSW)* 1843, introduced during the 1843-New South Wales depression and in force until 2005. The Act enabled possession over the wool clip growing on a sheep's back, which was in most instances not yet in existence, to be assigned to the lienee, typically a bank or a merchant advancing funds to the lienor (pastoralist). The lienor retained the actual control of the sheep (Decker, 2008).

Possession of chattels is generally transferred by delivery or taking. Delivery, in contrast to taking, is the voluntary transfer of possession from one person to another. Typically, this requires the physical delivery of the thing. However, in certain situations symbolic delivery is also sufficient. One example is the handing over of possession of the contents of a locked room through the delivery of the keys to the room (Bridge, 2002) or the transfer of timber logs in an open forest by simple agreement (Wolff and Raiser, 1957). Examples of gaining possession by taking include the acquisition of lost or abandoned goods, the capture of wild animals (also referred to as first possession), and the theft of articles by a wrongdoer. Possession of land is transferred by entry or ejectment.

Property law in turn protects possession. It provides the possessor with powerful remedies against any third party who unlawfully interferes with the thing in question as well as in situations where the possessor is dispossessed. In the interest of public peace, this protection is also given to the wrongdoer. Hence a possessor is allowed to exclude third parties and has remedies if dispossessed. Possession, being a property right, also survives the insolvency of the counterparty. For example, if A has leased a vehicle from B, A may retain possession for the duration of the agreement even if B goes into liquidation (Goode, 2004; Bridge, 2002).

Ownership

'Ownership' is one English term for 'dominium' (with 'property' being another term). It is a concept based on Roman property law, with 'dominium' reflecting the complete command that the head of the Roman household, the paterfamilias, wielded over its members (his wife and children) and its property (which included slaves). It encompassed unfettered power over life and death.

Consequently, ownership represents the 'greatest possible interest in a thing which a mature system of law recognizes' (Honore, 1961: p. 108), providing the owner with, for example, the power of occupation and use; the ability to exclude others; consumption; the right to the fruits and profit; destruction; free disposition through sale, gift and exchange; and the burdening of the thing through the granting of security rights (see below) and servitudes to others (Wolff and Raiser, 1957). Servitudes are property rights including rights of way and water, rights of light and support for buildings, and usufruct, being the right to take the fruits and profits from someone's property (Nicholas, 1975; Wolff and Raiser, 1957).

Considerable complexity is introduced by the fact that ownership as the fullest possible interest in a thing includes possession. Hence, as far as exclusive occupation, use and control, and the protection from interference by others are concerned, an owner who is also a possessor is in a similar position to the mere possessor (Holmes, 1881). However, the essence of ownership is not the exclusivity of use and the protection from interference. As Pollock and Maitland identified a long time ago, the pursuit of a thief by the person whose goods were taken "is an assertion of possession rather than of ownership" (Pollock and Maitland, 1898, Vol. 2. Chapter IV: Ownership and Possession). Consequently, ownership is "the ultimate residual right in a thing" (Nicholas, 1975: p. 154; Goode, 2004), designed so that fractions of the right of ownership such as possession or servitudes can be separated and surrendered to others without surrendering ownership itself. It is this capacity to separate ownership and possession that distinguishes modern credit-based monetary societies from tribal and feudal/socialist systems.

An important case in point is a transaction where the owner has surrendered the possession of a chattel for a period of time or at will (that is, being entitled to its return on request). Under English law such an arrangement is technically referred to as a bailment. Examples are an arrangement where the thing is provided to another person under a lease contract for his use, or where the thing is delivered to a creditor as security for a loan under a security arrangement. In both situations, ownership is retained, and the owner has remedies against third parties even though he is not in possession and has also parted with the immediate right to possession for the period of the contract. For example, if the bailee B in breach of contract sells the thing to a third party C, the owner A (bailor) can proceed against C (Pollock and Maitland, 1898; Bridge, 2002).

The critical point is that ownership allows possession to be transferred to other parties while legal rights against the (nonowner) possessor and against third parties interfering with the thing are retained. This separation of ownership and possession in turn enables the operations of a credit-based economy, which could not function without leases, the taking of nonpossessory loan security (see below) and the reservation of ownership over goods supplied but not yet fully paid (Bridge, 2002).

Proof of ownership without possession is difficult. Two important methods of assisting proof are prescription, which requires only possession and the lapse of time ('usucapio'), and registration of title in a public register (Nicholas, 1975). In Roman law a possessor could become the owner of a thing by 'usucapio,' if he, subject to a series of conditions, had had possession of the thing for a period of time. This was set at 2 years for land and 1 year for movables (Nicholas, 1975; Kaser, 1981). Modern societies based on property rights have developed effective registration systems. Ownership and any transfer of ownership are documented in public registers, of which the most important relate to land. Ownership is evidenced by an extract from the register.

Based on Karl Marx's analysis (Marx, 1867), it has become commonplace to contrast private ownership ('Privateigentum') and collective ownership ('gesellschaftliches kollektives Eigentum'). Care has to be taken when using these concepts. Legally, private ownership implies that an individual can hold and exercise ownership rights. In Roman times, those individuals would be the paterfamilias, and in societies with common or civil law systems any person (including corporations) with sufficient legal capacity to enter into contracts. In contrast, collective ownership of a thing within an ownership-based society refers to state ownership, a situation where the right of ownership is held by the state or a government agency (Wolff and Raiser, 1957). However, Marx and Engels envisioned the abolition of ownership altogether and its replacement with the 'collective *use* of all means of production and the distribution of all goods according to joint agreement' (Engels, 1914: p. 71, emphasis added). Hence, their primary concept is possession in the sense of physical control instead of ownership as the basis of relations between creditors and debtors.

Security

Any lender must ensure that loan losses will not exceed the lender's capital, even in unlikely events. One way to contain this exposure and to equalize the risk between lender and borrower is the taking of security (Heinsohn and Steiger, 2013, *see* Security: Commercial Transactions). Security fulfills two purposes. First, it enables the lender to more quickly recoup the amounts owed if the borrower fails to make repayments of loan principal and interest according to the terms of the loan contract. Second, as a property right rather then a mere contractual right, it provides protection to the lender against

competing claims of other unsecured creditors on the borrower's insolvency (McCracken and Everett, 2004). For example, A holds a hypothec over B's land as security for a loan. If B becomes insolvent, A is entitled to have recourse to the hypothecated assets, which are not available to the unsecured creditors. This is because A's right represents a relationship with the asset (thing) taken as security and not only a contractual relationship with the insolvent debtor B.

The taking of security has a long history. It is well evidenced in ancient Greece (Finley, 1951; Harrison, 1968) and was fully developed and codified in Roman law (Kaser, 1981). Security rights can be classified as possessory and nonpossessory. The most important possessory security right is the pledge (Roman 'pignus'). Here the borrower delivers the thing used as security to the lender. The lender acquires possession but the borrower retains ownership. Should the borrower default, the lender has the right to sell the thing, discharging all or part of the obligation with the sale proceeds.

In contrast, nonpossessory security rights are those that leave the possession of the thing used as security (also referred to as loan collateral) with the borrower. These are economically the most important as the borrower can raise funds from a lender and can continue with the occupation and use of his things, which are typically production assets (Decker, 2013). Nonpossessory security rights can be implemented in two ways. First, the borrower retains possession but transfers the ownership of the thing taken as security to the lender under an agreement to reconvey ownership on performance of the loan contract. The transfer of ownership provides the lender with the powers and rights of an owner, subject to legal restrictions and the terms of the loan agreement. Most importantly, ownership confers the power to sell the hypothecated asset if the terms of the loan agreement have been breached. Examples of this type of security right are the Roman law 'fiducia' (Kaser, 1981; Nicholas, 1975); the legal mortgage under English law (where possession may be transferred back to the borrower) and the security transfer ('Sicherheitsübertragung') for chattels in German law (Wolff and Raiser, 1957). Second, the borrower retains ownership and possession but the lender is provided with a 'charge' against the thing taken as security. The charge provides the lender with the powers specified in the loan agreement, typically the power to insure, the power to appoint a receiver and the power to sell (McCracken and Everett, 2004). Examples of this type of security are the pledge without possession, 'pignus obligatum' or 'hypotheca' in Roman law (Kaser, 1981; Nicholas, 1975), the equitable charge (including the fixed and floating charge) under English law and the 'Hypothek' under German law (Wolff and Raiser, 1957).

Nonpossessory security rights are often documented in public registers that identify the thing subject to the security arrangement. Examples are land, ships, motorcars, cattle, and wool. Registration often helps to regulate the priority of claims and allows other creditors to assess whether there are existing claims against the borrower's assets in question. As far back as ancient Greece, the registration of security rights was known. It was customary to place a mortgage stone ('horos') on the land inscribed with the particulars of the loan. Approximately 250 'horoi' have been found and documented (Millet, 1985).

Security rights play a critical economic role as the creation of money relies on the ability to encumber property by way of security for loans (*see* article Property Ownership and Heinsohn and Steiger, 2013). Central bank money is created when commercial banks enter into a loan contract with the central bank that must be secured by eligible central bank collateral. Furthermore, the function of the central bank as the lender of last resort is based on security rights. The underlying principle is that anyone able to furnish sound loan security must be able to obtain liquid central bank funds in a crisis (Heinsohn and Steiger, 2013). In turn, commercial banks create deposits through loans, the majority of which are directly or indirectly secured by land, houses, ships, cars, cattle, agricultural products or plant, and equipment.

Historical Development

Humans have been organized in tribal communities for the majority of human history. It has been estimated that communities of hunters and gatherers account for 90% of that period (Lee and Daly, 2005). Anthropological studies confirm that tribal communities knew a form of possession protected by force or customary rules. There is evidence that tribes laid exclusive claims to territories that were defended against intruders. Land could be allotted to a family household who had a use right; game that had been killed and left lying to be fetched later was marked with branches as tabu signs; and the manufacture of articles created a personal claim to their possession (Sahlins, 1974; Thurnwald, 1932). However, the difference between possession and ownership as used in the context of creditor and debtor, that is, the "distinction between the actual holding of a thing and the legal claim to it […] is unknown" (Thurnwald, 1932: p. 192). Or, as the eminent anthropologist Malinowski notes, the "term 'own' as we use it is meaningless, when applied to a native society" (Malinowski, 1922: p. 117). Consequently, while a form of possession is known, ownership as a residual right exercised independently of possession is unknown to tribal communities and therefore is unlikely to have played a role throughout most of human history.

Another dominant form of social organization throughout history has been command systems such as Mycenaean Greece, the Inca Empire, the medieval European feudal seigneuries and state socialism. Command systems are organized through coercive mechanisms where a ruling class extracts dues and services from serfs or serflike workers (Heinsohn and Steiger, 2013). As in the case of tribal communities, it has been observed that property law in command systems – if one can use the term – was limited to possession. For example, medieval European feudalism was governed by the concept of 'seisin' (possession, Pollock and Maitland, 1898). Land in feudal systems was held by a sovereign lord and individuals had access and use rights in return for dues and services. To use "the word 'ownership' […] would have been almost meaningless […] [f]or all land and a great many human beings were burdened at this time with a multiplicity of obligations" (Bloch, 1961: pp. 115–116). This is consistent with the fact that the English word 'ownership' is only evidenced as recently as 1583 (Pollock and Maitland, 1898), a time when England was no longer a feudal society and its transition to an ownership-based economy governed by relations between creditors and

debtors was well underway. The notion of possession as the leading concept also holds true in relation to the treatment of chattels in medieval law. It has been noted that the "punishment of thieves and the restoration of chattels to those from whom they have been stolen were the main objects which the law had set itself to attain" (Pollock and Maitland, 1898, Vol. 2. Chapter IV: Ownership and Possession). In relation to the lord, the serf was without formal rights. An English lord might beat or imprison his serf, seize his chattels, and eject him at his discretion (Pollock and Maitland, 1898; Lipson, 1956). The serf was also unprotected against the world at large. If a third party ejected him, it was the lord who was considered disseised (dispossessed) (Pollock and Maitland, 1898).

There also exists no right of ownership in the former and failed system of state socialism with its system of central planning, although the term 'ownership' was used. While use rights were allocated to individuals over, for example, rural land, these were not full ownership rights, but limited rights stripped of essential powers such as the right to divide, sell, lease, or hypothecate the land. Consequently, the German Federal Court in a landmark dispute over land rights granted by the former socialist authorities in East Germany concluded that these socialist use rights no longer fulfilled the essential meaning of ownership according to the German civil code (Heinsohn and Steiger, 2013).

The situation regarding the ancient economies of Greece and Rome (1000 BC–AD 500, Finley, 1973) is very different. It is within those economies that property rights and a concept of ownership as separate from possession in the modern sense are believed to have originated (Pipes, 1999; there is also some evidence from Mesopotamia, see Westbrook, 2003; Heinsohn, 1984) and within which private ownership became the characteristic feature of social organization (Finley, 1973). In Greece, the property law of Athens recognized the distinction between ownership and possession. There is a strong body of evidence confirming the existence of security rights, including those rights that left possession of the collateral with the debtor. One example was the most common form of real security, 'prasis epi lysei' (sale subject to redemption, Finley, 1951). Evidence in texts is confirmed and supplemented by numerous mortgage stones (Harrison, 1968; Millett, 1985).

It was, however, the Romans who developed property law into 'an elaborated articulated system of principles' (Nicholas, 1975: p. 1), providing mankind for the first time with a fully fledged system of property rights, including ownership as a right distinct from possession and security rights encompassing nonpossessory rights such as the hypothec. In turn, Roman law became the model for property law systems in Europe, forming the basis of the French civil code (Code Napoléon of 1804) and the German civil code of 1896 (Nicholas, 1975). An exception perhaps is England, which has retained its own common law. Yet even the development of the common law was not without Roman law influence. Most legal systems currently operating in the world are directly or indirectly derived either from English common law, French civil law, or German civil law (Berkowitz et al., 2003).

How did the right of ownership first emerge? This is a topic of considerable uncertainty. The Bronze Age, with its feudal command systems such as the Mycenaean civilization, is reported to have come to a rather abrupt end with archaeological evidence revealing cataclysms resulting in the widespread destruction of palaces and fortress complexes (Finley, 1981; Welwei, 1983). Invasions and large-scale natural disasters have been considered as possible causes (Finley, 1981; Welwei, 1983; Heinsohn, 1984). In Greece, a completely new society was then built by those who remained (Finley, 1981). Hence, it has been argued (Heinsohn, 1984) that the emergence of societies based on ownership, like the Roman civitas and the Greek polis, was not a gradual evolutionary process but was triggered by exceptional historical circumstances. In the wake of natural disasters, revolutionaries such as the legendary Greek hero Theseus, who is reported to have promised a new society – a 'government without a king' (Plutarch) – overturned their feudal rulers and established a free society based on the distribution of former feudal land holdings to their members. According to Varro, Romulus – the legendary founder of Rome – allotted an equal amount of land (the 'heredium') to each citizen. This created, for the first time, a society based on private ownership of land (*see* article Property Ownership) (Heinsohn, 1984; Heinsohn and Steiger, 2013). The societies created in this way were in sharp contrast to feudal societies as they were regulated not by rules of command but instead by relationships between creditors and debtors, who were bound by property law and who sold, mortgaged, and leased their assets. At the same time, the newly formed ownership-based societies became the birthplace of freedom and democracy (Pipes, 1999; Steiger, 2005). At a more general level, it has been argued that tribal, command and ownership-based systems represent three distinct archetypes of human organization, which impose radically different rules of production, distribution, consumption, and accumulation. In contrast to possession, ownership is only present in one of the three systems (Heinsohn and Steiger, 2013; Steiger, 2006).

The fall of the Roman Empire ended the relatively short regime of ownership-based systems. It was not until the fourteenth century that a new type of ownership-based society began to reemerge on a significant scale in England. In the mid-fourteenth century the demographic disaster of the Black Death swept away substantial parts of the European population. Demographic changes of this magnitude weakened and undermined the existing social structures throughout Europe. The reaction was either a tightening (Eastern Europe) or a disintegration (Western Europe) of feudal structures (Brenner, 1976). England, whose population decline from the Black Death (1348/49) has been estimated at 30–50% and which experienced a violent and widespread peasant rebellion in 1381, witnessed a complete disintegration of the old feudal order within a century of the Black Death (Lipson, 1956). With labor scarce and the position of peasants strengthened, English landlords could no longer rely on compulsory labor services to cultivate their demesne (the portion of the seigneurie, where the product was taken directly by the lord). Having lost their serf-labor while retaining the land, landlords began to rent out their demesne to substantial tenants who could sublet the land. In addition, the service obligations owed by serfs were commuted to money rents and traditional holdings – providing unlimited use rights when the tenure was hereditary – were converted into shorter term leases when an opportunity arose (Wrightson, 2002). This allowed landlords to commence the economic management not only of their demesne, but also the rest of the seigneurie after evicting customary tenants (Nabholz, 1941).

The introduction of widespread tenant farming raised the legal question of who had possession: the lord or the tenant/leaseholder (lessee). The answer, already developed in the early thirteenth century, was the 'theory of double seisin' (Pollock, 1888: p. 48). The "lessee had a kind of seisin and yet the lessor [lord] did not lose the seisin which he had before" (Pollock, 1888: p. 49). It subsequently became customary to confine the term seisin to the lord's right over his holdings and the Roman term 'possession' was introduced as a technical term to describe the seisin of the land exercised by the lessee, who occupied the land. The lessee, unlike the medieval serf, had now possessory rights against third parties. This development was paralleled by the evolution of bailment of chattel property. When bailment became more important, the bailor acquired rights against third parties and not just against the bailee (Pollock and Maitland, 1898). Hence, a conceptual separation analogous to that of ownership and possession was recreated in English common law underpinning the reemergence of an ownership-based economic system built on tenant farming, which transformed England into the leading economic power of the world (*see* article Property Ownership).

In line with the transformation and success of the English economy, English property law was transplanted in the colonies of the British Empire. As part of this process, former aboriginal land was distributed on a vast scale to European settlers through purchase or land grants to companies and individuals. In addition, schemes to distribute public lands in the colonies – not dissimilar to the original land distribution attributed to Romulus – were set up. Land was granted to any eligible individual, subject to occupancy and improvement conditions and either for free (for example, the United States Homestead Act of 1862 and the Canadian Dominium Lands Act of 1872) or on generous credit terms (for example, the New South Wales Crown Lands Alienation Act 1861 and the Victorian Lands Act of 1862). From the first half of the nineteenth century, many countries introduced major property and land reforms to modernize their economies in order to keep up with the developments in England. Prominent examples are the Stein–Hardenberg reforms of 1807 in Prussia, the 1861-emancipation edict by Tsar Alexander II, and the reforms in Japan under the Emperor Meiji in the second half of the nineteenth century. In 2007, China, the country with the world's largest population, introduced the Property Law of the People's Republic of China, formalizing a development path taken by England six centuries earlier (Heinsohn and Steiger, 2013; Steiger, 2006).

See also: Ancient Greece and Rome, History of; Capitalism; Colonization and Colonialism, History of; Feudalism; Marx, Karl (1818–83); Meiji Revolution, The; Middle Ages, The; Plagues and Diseases in History; Property Ownership; Security: Commercial Transactions; Slaves and Slavery, History of.

Bibliography

Berkowitz, D., Pistor, K., Richard, J.F., 2003. Economic development, legality, and the transplant effect. European Economic Review 47, 165–195.
Bloch, M., 1961. Feudal Society. Volume 1 – The Growth of Ties of Dependence (L.A. Manyon, Trans.). The University of Chicago Press, Chicago.
Brenner, R., February 1976. Agrarian class struggle and economic development in pre-industrial Europe. Past and Present 70, 75–93.
Bridge, M., 2002. Personal Property Law, third ed. Oxford University Press, Oxford.
Decker, F., 2008. The legal and economic history of the Lien on Wool and Stock Mortgage Act 1843 (NSW). Legal History 12 (2), 151–175.
Decker, F., 2013. Ownership economics: an introduction. In: Heinsohn, G., Steiger, O. (Eds.), Ownership Economics. Translated and edited with comments and additions by F. Decker. Routledge, London, pp. xvii–xxxv.
Engels, F., 1914. Grundzüge des Kommunismus. Reclam edition 1977. Reclam, Stuttgart.
Finley, M.I., 1951. Studies in Land and Credit in Ancient Athens 500–200 B.C. Rutgers University Press, New Brunswick.
Finley, M.I., 1973. The Ancient Economy. Chatto & Windus, London.
Finley, M.I., 1981. Early Greece: The Bronze and Archaic Ages, New and revised edition. Chatto & Windus, London.
Goode, R., 2004. Commercial Law, third ed. Penguin Books, London.
Harrison, A.R.W., 1968. The Law of Athens. At the Clarenden Press, Oxford.
Heinsohn, G., 1984. Privateigentum, Patriarchat, Geldwirtschaft: Eine sozialtheoretische Rekonstruktion zur Antike. Suhrkamp, Frankfurt am Main.
Heinsohn, G., Steiger, O., 2013. Ownership Economics. Translated and edited with additions and comments by F. Decker. Routledge, London.
Holmes, O.W., 1881. In: Pereira, P.J.S., Beltran, D.M. (Eds.), The Common Law, 2011 Electronic Edition. University of Toronto Law School, Toronto.
Honore, A., 1961. Ownership. In: Guest, A.G. (Ed.), Oxford Essays in Jurisprudence. Oxford University Press, Oxford.
Kaser, M., 1981. Römisches Privatrecht. Juristische Kurz-Lehrbücher. C.H. Beck, München.
Lee, R.B., Daly, R., 2005. Foragers and others. In: Lee, R.B., Daly, R. (Eds.), The Cambridge Encyclopedia of Hunters and Gatherers, Paperback edition. Cambridge University Press, Cambridge.
Lipson, E., 1956. The Economic History of England. In: The Middle Ages, eleventh ed., vol. 1. Adam & Charles Black, London.
MacDowell, D.M., 1978. The Law in Classical Athens, Paperback edition 1986. Cornell University Press, Ithaca.
Malinowski, B., 1922. Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagos of Melanesian New Guinea, Paperback edition 1961. Dutton & Co, New York.
Marx, K., 1867. Das Kapital: Kritik der politischen Ökonomie. Erster Band. In: Karl Marx-Friedrich Engels-Werke, Band 23. Dietz Verlag, Berlin, 1969.
McCracken, S., Everett, A., 2004. Everett and McCracken's Banking and Financial Institutions Law, sixth ed. Law Book Company, Sydney.
Millett, P., 1985. The Attic horoi reconsidered in the light of recent discoveries. In: Finley, I.M. (Ed.), Studies in Land and Credit in Ancient Athens 500–200 B.C. Transactions Inc., New Brunswick.
Nabholz, H., 1941. Medieval agrarian society in transition. In: Clapham, J.H., Power, E. (Eds.), The Cambridge Economic History of Europe. From the Decline of the Roman Empire, The Agrarian Life of the Middle Ages, vol. 1. At the University Press, Cambridge.
Nicholas, B., 1975. An Introduction to Roman Law. Clarendon Press, Oxford.
Pipes, R., 1999. Property and Freedom. Alfred A. Knopf, New York.
Pollock, F., 1888. An essay on possession in the common law, parts I & II. In: Pollock, F., Wright, R.S. (Eds.), An Essay on Possession in the Common Law, Parts I, II & III. Clarendon Press, Oxford.
Pollock, F., Maitland, F.W., 1898. In: The History of English Law before the Time of Edward I. Reprint of 2nd edition, with a Select Bibliography and Notes by Professor S.F. Milsom, 2 vols. Liberty Fund, 2010, Indianapolis. Accessed from: http://oll.libertyfund.org/title/2314/219571 (on 10.06.12.).
Sahlins, M., 1974. Stone Age Economics. Tavistock Publications, London.
Schapp, J., Schur, W., 2002. Sachenrecht, third ed. Verlag Vahlen, München.
Steiger, O., 2005. Eigentum und Recht und Freiheit: Eine Triade und 66 Thesen. In: Krieg, W., Galler, K., Stadelmann, P. (Eds.), Richtiges und gutes Management: vom System zur Praxis – Festschrift für Fredmund Malik. Haupt, Bern, pp. 153–178.
Steiger, O., 2006. Property economics versus new institutional economics: alternative foundations of how to trigger economic development. Journal of Economic Issues XL (1), 183–208.
Thurnwald, R., 1932. Economics in Primitive Economies. Oxford University Press, Oxford.
Welwei, K.-W., 1983. Die Griechische Polis. Kohlhammer, Stuttgart.
Westbrook, R., 2003. A History of Ancient Near Eastern Law. Brill, Leiden, Boston.
Wolff, M., Raiser, L., 1957. Sachenrecht von Martin Wolff. In: Enneccerus, L., Kipp, T., Wolff, M. (Eds.), Lehrbuch des Bürgerlichen Rechts, Dritter Band, Tenth ed. Tübingen, J. C. B. Mohr (Paul Siebeck).
Wrightson, K., 2002. Earthly Necessities. Economic Lives in Early Modern Britain, 1470–1750. Penguin Books, London.

Relevant Website

http://oll.libertyfund.org/ – The Online Library of Liberty, A project of Liberty fund, Inc.

Prophetism

Malcolm Hamilton, University of Reading, Reading, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12208–12212, © 2001, Elsevier Ltd.

Abstract

This article examines the phenomenon of prophetism that occurs widely in many religious traditions and cultures in many periods of their history. After discussing the nature and definition of prophetism and of the central figure in prophetic currents and movements – the prophet – the article traces the sociological foundations of the concept in the work of Max Weber and his stress on the charismatic nature of prophecy. The way Weber differentiated between prophets and other religious specialists is discussed. Weber's analysis of Biblical prophecy is then examined followed by a discussion of later critiques of Weber and of recent work on Biblical prophecy. The article concludes with an examination of the role of the prophet in religious movements, and in particular millennial movements.

Prophetism may be defined as a religious current or movement promoted by the activities of a prophet. A prophet in turn may be defined as a leader of or key participant in a movement or group who proclaims a religious mission on the basis of personal revelation. While many prophets have been founders of religious movements they are not necessarily so but may be significant figures within a revitalization of a religious group or simply play a central role in the religious life of a group or community by pronouncing upon religious matters through revelatory insights or experiences. Neither is the foretelling of future events a necessary component of prophecy but is not uncommonly associated with it. A prophet may or may not convey a message from the gods or be the mouthpiece of divinity. They may express the word of God or the gods or may simply convey religious truths to their followers, which are not claimed to stem from such supernatural sources. Buss (1980) puts the point usefully in the following terms: 'The role of a prophet or seer can be characterised as the receiving and transmitting of communications not available to ordinary conscious sensitivity which are held to come from a source (the divine) or through a form of perception transcending normal spatio-temporal limitations.' Prophets may or may not claim spiritual or supernatural powers themselves and may or may not receive revelation through the experience of states of trance or possession.

What is essential to the notion of prophecy, then, is the relatively emotional proclamation of a religious message that has been revealed personally in a nonroutine and non-mundane manner. Prophets are called to their mission and compelled to follow it whether voluntarily or, as is not uncommon, very reluctantly.

The category of prophet as a religious specialist or type of religious figure is one which owes much to Weber who gives it a central role in his sociology of religion (see Religion, Sociology of). For Weber a prophet is 'a purely individual bearer of charisma, who by virtue of his mission proclaims a religious doctrine or divine commandment' (1965). Weber distinguishes the prophet from the category of priest. The priest commands religious authority on the basis of a position within a sacred tradition or religious organization and dispenses salvation by virtue of holding an office while the prophet commands religious authority on the basis of personal charisma and recognition of a personal calling to a mission through revelation. The term 'charisma' is used by Weber to refer generally to extraordinary powers essentially of a magical or non-mundane nature (see Charisma and Charismatic). The priest for Weber is thus the upholder and guardian of tradition while the prophet is a religious innovator or at least a radical revitalising force.

Weber is careful to distinguish the prophet from a number of other types of religious specialist. While the authority and following of prophets may be based upon magical charisma they are not primarily magicians or workers of miracles. The prophet is concerned primarily with doctrine and commandment. It is this that distinguishes the prophet from the mystagogue whose following is based purely upon attribution of magical powers. Yet it is rare that prophets are not expected to give proof of their particular gifts and such proof has more often than not taken the form of possession of magical abilities. Reflecting the fact that charisma may be either inherent or acquired through undergoing extraordinary experiences or practices, these capacities may be inherent in the person or may be acquired through, e.g., ascetic regimes or cultivation of ecstatic states.

Again, while prophets engage, in many instances, in divinatory activities, as was the case, e.g., with the Biblical *nebiim* (singular *nabi*), they may be distinguished from diviners on the basis that it is normal for the latter to be economically recompensed for their services. Prophets provide their spiritual revelations gratuitously and, indeed, often unbidden. To do otherwise would be to undermine their distinctive claim to religious and spiritual influence.

Likewise, Weber distinguishes the prophet from related or similar religious specialists such as legislators, arbitrators, social reformers, orators and teachers on the grounds that lacking in the case of all of them is the proclamation of a religious truth on the basis of personal revelation. Later usages of the term 'prophet' are often less precise and looser than Weber's but in general at the root of them is this specific emphasis upon a purely personal revelatory impulse and calling.

Weber distinguished between two fundamentally different types of prophecy. In the 'ethical' or 'emissary' type, exemplified most clearly by figures such as Zoroaster or Muhammad, the prophet is an instrument for the proclamation of the message or commandments of a god. The doctrines proclaimed by such a prophet take the form of ethical duties. In the

'exemplary' type, most clearly exemplified by the Buddha, the prophet proclaims a soteriological message by personal example and demonstration of the means or path to salvation which others are free to follow to their personal benefit if they wish or ignore to their personal detriment.

In both cases the prophetic revelation acts to promote a meaningful understanding of life in relation to the whole of reality to which conduct must be brought into alignment if salvation is to be attained. It involves the construction of a cosmos as a meaningfully ordered totality. As Parsons (1965) says 'the prophet is above all the agent of the process of breakthrough to a higher, in the sense of more rationalised and systematised, cultural order, an order at the level of religious ethics, which in turn has implications for the nature of the society in which it becomes institutionalised.' While there is such a stress in Weber's characterization of the prophet upon the construction of a more rationalized and systematized meaningful order, prophecy, Parsons is careful to point out, is not fundamentally an intellectual or theological enterprise but frequently takes emotional and sometimes even violent forms.

Weber's most extensive analysis of specific prophetic activity and prophetic movements occurs in his study of the development of Judaism, in his *Ancient Judaism* (1952), a development in which the Biblical prophets are given a crucial role (see Judaism). Weber traces the emergence of a new type of prophet towards the end of the period of the confederacy that he distinguishes from the earlier ecstatics or *nebiim* and the *roeh* or oracle givers. These new prophets saw themselves as called upon to transmit the will of Yahweh to the people and to the rulers. Early prophets of this type did not pronounce unless solicited but progressively they spoke independently and increasingly critically of the political authorities. From the time of Elijah onwards prophets spoke as much to the masses as to the rulers culminating in a situation in which the prophets became the intellectual leaders of the opposition against all that monarchy was perceived to stand for – centralization of power and government, heavy taxation, and social injustice.

Not that Weber considered social justice to be their primary concern. It was the destiny of the state and of the people that motivated their essentially religious yet inextricably political message. It was their purely religious concerns with fulfilling the commandments of Yahweh that brought upon them the opposition of the authorities. For Weber the Biblical prophets represented the values and ethos of the presettlement confederacy.

Classical Biblical prophecy in Weber's analysis is characterized by independence, spontaneity, and disregard for popularity among any sector of Israelite society. It was indifferent to economic, material, or other reward in contrast to other types dependent for their livelihood upon prophecy and prediction. They were at best tolerated and not infrequently persecuted.

While not primarily diviners and seers the classical Biblical prophets nevertheless stressed the dire fate that Israel would suffer if it did not keep the terms of its covenant with God. Such was the precarious situation of the Israelite states in Palestine that gloomy predictions were not uncommonly fulfilled. The religious conduct of the Israelite people was linked intimately by the activities of the prophets to their material and political fate and to that of the state.

The conditions which promoted the development and impact of prophetic currents of this type were those of military and political insecurity and unpredictability. Weber considers that new religious insights rarely occur in the main centers of civilization but rather in areas on their periphery which are subject to their influence and pressure. Those who live in such regions repeatedly are given reason to question the meaning of events that affect them. Unlike the inhabitants of long stable, politically secure, and settled regions who tend to take the world for granted, those outside these areas constantly are puzzled by the turn of events; 'the possibility of questioning the meaning of the world presupposes the capacity to be astonished about the course of events' (Weber, 1952).

Subsequent research on Biblical prophecy has often been critical of aspects of Weber's analysis and by implication of his analysis of prophecy and prophetism in general. Berger (1963) and Turner (1974) have questioned whether the Biblical prophets were the independent force based upon pure charismatic authority and motivated entirely by religious concerns that Weber portrays them as being. In fact, the classical Biblical prophets are more likely, according to some recent research, to have occupied official positions in the cult of Yahweh. The conclusion that writers such as Turner draw from this, that it requires us to reconsider Weber's understanding of prophetic charisma as an autonomous and personal force and to recognize that it arises from the incumbency of positions in an established social institution is, perhaps, rather extreme. It is likely that there were several different types of prophet in ancient Israelite society. Emmet (1956), who considers that Weber equates prophecy with charismatic authority which she finds too restrictive, usefully differentiates a number of types of prophet. First, there is the institutional prophet who fulfils a traditional and established role. Second, corresponding to Weber's charismatic prophecy, there is the leader of a movement which seeks radical change. Finally, there are prophets who do not attract a personal following but who act as independent critics, advocates of reform or who provide moral leadership. While the prophets associated with the Temple in Jerusalem, whose function was an institutionalized one within the Temple cult, were perhaps more numerous, the classical Biblical prophets with which Weber was concerned may be considered to be of a different type. Whether they enjoyed greater autonomy and independence stemming from personal charisma is still, however, open to question.

Weber's stress upon the charismatic authority of the classical Biblical prophets has been challenged by Petersen (1979) on the grounds that there is no evidence that they were leaders with a group of followers, which he claims is implied by Weber's notion of charismatic authority. Neither is it useful, in his view, to characterize Biblical prophets as office holders. The dichotomy between charismatic leader and office holder is, in any case, he argues, far too strictly drawn. We are led, therefore, to question whether the Israelite prophets could have provided the breakthrough that Weber attributes to them. In a later work Petersen (1981) argues that we may get a better understanding of the Biblical prophets if we see them as having a specific role. He applies role theory in an analysis of their position in Israelite society based upon a close examination of the use of a variety of terms in Hebrew used to refer to such figures. The results show that there was more than one type of prophetic

role in this society. Some held established positions in the official cult and operated also by offering a personal service for a fee. Others, termed central morality prophets, pronounced as individuals at times of crisis but in order to uphold established values and institutions. Still others, peripheral prophets, operated outside institutional contexts and established values.

Also doubtful of the simple dichotomy of functionary versus autonomous charismatic is Lang (1983). The prophetic tradition, he argues, was largely an institution of the land-owning nobility that enabled them to express publicly their political views. He attempts to show that the prophets were largely from a land-owning, noble background. While they were not chosen by formal procedures but simply emerged, they nevertheless, according to Lang, had to learn how to be an acceptable prophet and to observe the rules governing appropriate behavior for prophets. They were closer, however, to Emmet's third type, in Lang's view, and not wholly established functionaries within the cult. The contrast between charisma and tradition in discussions of prophetism is indeed often overstated. Lang points out that they are often bound up with one another in complex ways such that prophetic charisma can be seen as a force for renewal which emerges within traditional structures at certain times. It is a force that changes and reforms without destroying.

While the Weberian view of Biblical prophecy has, in recent decades, been questioned by some, it has for others stood the test of time and deeper scrutiny rather more robustly. Williams' analysis of sources (1969), e.g., supports the view that many of the classical prophets were independent of the official cult. A detailed study by Wilson (1980) is alert to the complexities and differences between periods and regions. He contrasts the premonarchical Ephraite tradition of the North with the monarchical period in both the Northern and Southern kingdoms and also different periods and circumstances in the South. In the premonarchical North, prophetic behavior was expected to conform to familiar patterns regarding the stereotypical linguistic style in which pronouncements were made and to customary behavioral patterns. This clearly suggests that these prophets occupied a defined position and role within the central religious institutions of the society. In the monarchical period prophets seem to have become figures of less central significance but rather more peripheral to the established religious institutions and rather more concerned with producing change. In Judah, however, political circumstances and alliances differentially affected the type of role and significance that prophets had. Some, standing outside the established cult, were more critical of the status quo while others exercised a more supportive function. This reminds us that prophetism, at least of this type and in these circumstances, may not simply refer to a movement or development led by a prophetic figure but to a dynamic situation of struggle to define or lay claim to the accepted interpretation of a religious tradition or interpretation. Lang (1983) points out that, like many prophets in various traditions and cultures, Jeremiah engaged in conflict with rival prophets opposed to his policy of accommodation with the Babylonians and who advocated total resistance.

Such debates highlight the extent to which Biblical prophecy, and it is an observation applicable to all prophecy, is associated with crises which, while differing extensively in kind, generally call forth a questioning of the status quo and of established institutions and patterns of power and authority. This point has been made by Carroll (1979) in the case of ancient Palestine and Weber's analysis. The religious-political factions that tend to emerge in such circumstances may understandably become aligned with pro-regime and anti-regime types of prophecy.

Not that this necessarily means that prophetism is really a form of politics under the guise of religion. As Albrektson (1972) notes, Weber may well be correct in attributing to the Biblical prophets motives which were primarily religious. Circumstances often made their pronouncements political in tone but at root they were essentially utopian and idealistic rather than pragmatic or designed to stir up political opposition for its own sake.

The political element in prophetism is a factor that comes to the fore most notably in studies of those forms of prophetism that can be described as millennial. It is in the context of the quite extensive studies of such movements that prophetism has been most systematically examined (see Millennialism). Not all prophetic movements are millennial and not all millennial movements are prophetic, though most are. It is prophetic millennial movements, however, that have received most attention (see, e.g., Wilson, 1975). The role of the prophet in millennial movements varies. Such movements may be initiated by the visions, revelations, and activities of a prophet. Once underway they may call forth a prophet who focuses relatively disordered and uncoordinated activities of participants and who gives coherent expression to previously vague, inchoate, and diverse ideas. The appearance of the prophet transforms millennial ferment into a millennial movement.

The role of the prophet in the millennial movement generally has been found to reflect Weber's emphasis upon prophetism being associated with social and political crisis or radical social change. On the other hand, it is rarely a complete break with the past and more often a process of bridging past and present, the traditional and the new. The prophetic message in such movements tends to reinterpret old understandings in ways appropriate for an altered situation while at the same time adding new elements and opening up the imagination to new possibilities. The prophet gives concrete expression to what are often inarticulate feelings and ideas and channels them into a program of action promising redemption from current ills. The prophet helps those who are confused by circumstances of change, uncertainty, and instability to comprehend and grasp the meaning of their situation by offering insight derived from special revelation which through understanding points the way forward for remedying what is unsatisfactory. Prophetic movements frequently are found in circumstances of disaffection in which a population is subject to foreign or colonial rule or have little comprehension of the mechanisms and structures of power or how to operate within them in order to ameliorate their situation. The prophet provides the missing knowledge and means of finding solutions. It is significant that many prophetic leaders of such movements are more knowledgeable than their followers as a result of wider experience through travel, limited education, or access to information. The prophet, by providing a focus, is often able to bring together otherwise fragmented communities and diverse groups previously unable to operate in

concert in the pursuit of common aims. What is promised, however, is not simply material but also usually, and more significantly, moral and emotional redemption through the recovery of lost personal and cultural dignity and sense of self-worth.

See also: Charisma and Charismatic; Charisma, Social Aspects of; Millennialism; Monasticism: West and East; Religion, History of; Religious Organizations; Weber, Max (1864–1920); Weberian Social Thought, History of.

Bibliography

Albrektson, B., 1972. Prophecy and politics in the Old Testament. In: Biezais, H. (Ed.), The Myth of the State. Almqvist and Wiksell, Stockholm, Sweden.

Berger, P., 1963. Charisma and religious innovation: the social location of Israelite prophecy. American Sociological Review 28, 940–950.

Buss, M.J., 1980. The social psychology of prophecy. In: Emerton, J.A. (Ed.), Prophecy. W. de Gruyter, Berlin.

Carroll, R.P., 1979. When Prophecy Failed: Cognitive Dissonance in the Prophetic Traditions of the Old Testament. Seabury, New York.

Emmet, D., 1956. Prophets and their societies. Journal of the Royal Anthropological Institute 86, 13–23.

Lang, B., 1983. Monotheism and the Prophetic Minority: An Essay in Biblical History and Sociology. Almond Press, Sheffield, UK.

Parsons, T., 1965. Introduction. In: Weber, M. (Ed.), The Sociology of Religion. Methuen, London.

Petersen, D.L., 1979. Max Weber and the sociological study of ancient Israel. In: Johnson, H.M. (Ed.), Religious Change and Continuity. Jossey Bass, San Francisco.

Petersen, D.L., 1981. The Role of Israel's Prophets. Journal for the Study of the Old Testament, Supplement Series 17.

Turner, B.S., 1974. Weber and Islam. Routledge, London.

Weber, M., 1952. Ancient Judaism. Free Press, Glencoe, IL.

Weber, M., 1965. The Sociology of Religion. Methuen, London.

Williams, J.G., 1969. The social location of Israelite prophecy. Journal of the American Academy of Religion 37, 153–165.

Wilson, B.R., 1975. Magic and the Millennium. Paladin, St Albans, UK.

Wilson, R., 1980. Prophecy and Society in Ancient Israel. Fortress Press, Philadelphia, PA.

Prophylaxis in Psychiatry

David Castle, University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Levine, S. Gershon, volume 18, pp. 12212–12216, © 2001, Elsevier Ltd.

Abstract

Prophylaxis in psychiatry can be considered as a series of phase-specific constructs, targeting different stages of mental illnesses. Types of intervention encompass medications as well as broader psychological and psychosocial constructs. In this article, we discuss three psychiatric disorders as illustrations of the different prophylactic paradigms; these disorders are schizophrenia, bipolar disorder, and major depression. We then address the important aspect of physical health problems – notably cardiovascular issues such as obesity, diabetes, and hypercholesterolemia – that are often associated with major mental illnesses and which may be compounded by certain psychiatric medications. Finally, we provide an overview of issues pertinent to adherence to treatment interventions.

Schizophrenia

Schizophrenia is a disorder that appears to have its origins in abnormal neurodevelopment, albeit it usually manifests in terms of 'obvious' symptoms such as delusions (false beliefs) and hallucinations (abnormal perceptions), collectively called 'positive symptoms,' in the late teens in males and early to late 1920s in females (Castle et al., 1998). The major known risk factors for schizophrenia are genetic, with a complex polygenetic multifactorial inheritance compounded by epigenetic and gene–environment interaction effects. Known 'environmental' risk factors include maternal exposure to influenza type A and (possibly) toxoplasmosis; maternal starvation, anemia, and low vitamin D levels; and pregnancy and birth complications (Castle and Buckley, 2015). There is also evidence that cannabis consumption in the teens can increase the likelihood of later schizophrenia. The problem with any of these factors as targets for preventing schizophrenia is their low associated risk (mostly of the order of a twofold increase in risk) and their lack of specificity. This is not to say that attempting to reduce such risk factors is a bad idea, but simply that the 'yield' at a population level is likely to be low. For example, it has been estimated that the population attributable risk of cannabis in schizophrenia is of the order of 8% (Arseneault et al., 2004), suggesting that elimination of cannabis from the face of the globe would reduce the incidence of schizophrenia by around 8%; not a trivial amount, but small in terms of return for such a massive endeavor.

A more parsimonious approach could be to target women at high genetic risk for having a child with schizophrenia (e.g., if she has schizophrenia herself) and offering them the very best in obstetric and perinatal care. All of the maternal risk factors quoted above could be addressed, as well as the drug and alcohol consumption which is all too often associated with schizophrenia and which can also impact the developing brain. Assisting the mother to bond with her infant could also assist, expressly as affective expression is often limited in people with schizophrenia. Such approaches have the challenge that many pregnant women with schizophrenia are suspicious of and avoid health services, often through fear of having their baby removed from them (Hearle and McGrath, 2000).

As already stated, schizophrenia usually only manifests itself definitively in later adolescence or older, but certain 'markers' might be evident even in childhood. These include delays in language acquisition and motor milestones, as well as impairments in socialization. However, any such factors are very nonspecific and most children with such 'markers' do not develop schizophrenia, making them an impossible target for preventative interventions. However, there has been increasing interest in whether certain individuals at supposed 'ultra high risk (UHR)' for the disorder might be effectively identified and offered prophylactic treatment. Various approaches to determining UHR status have been researched, but most include some combination of psychotic symptoms, such as paranoia or grandiosity that do not reach the intensity required to meet diagnostic criteria for the disorder as such; brief limited psychotic episodes or BLIPS; and having a family history of schizophrenia or schizotypy and showing psychosocial decline. Initial studies were very promising in the sense that around 40% of people identified in high-risk specialist services who met such criteria 'transitioned' to full psychosis. However, more recent studies using broader sampling frames have had a much lower yield (of the order of 8–10%) (Fusar-Poli et al., 2014). Extrapolation to a general population shows that in individuals who are not 'help-seeking,' the positive predictive value of putative UHR-like symptoms is very low (around 1%) (Werbeloff et al., 2012). A major problem for this approach is the lack of specificity of such symptoms (psychosis-like symptoms occur in 10–15% of the population) and also the lack of specificity of potential outcomes (people with such symptoms are also at higher risk of developing mental disorders other than schizophrenia).

The refinement of the UHR approach, by adding other symptoms and potentially neurobiological markers for schizophrenia (e.g., Lawrie et al., 1999; Shimushi et al., 2000) is a worthy endeavor, but even if this reaps dividends in terms of enhancing the positive predictive value, there remains a question as to what prophylactic interventions could be offered to prevent 'transition' to a psychotic illness. A number of approaches have been tested, including antipsychotic medication and cognitive behavior psychotherapy or combinations thereof, but unfortunately the more recent,

methodologically robust studies have largely been negative (Castle, 2012). One very intriguing study showed a dramatic effect for the simple intervention of fish oil (omega-3 fatty acid) for 3 months, but this study has not been replicated. In sum, the UHR approach, while potentially very exciting, has thus far not delivered on its early promise. No specific intervention can at this stage be recommended, and most clinicians would eschew exposing putative UHR individuals to antipsychotic medications, expressly in light of the side effects associated with such agents (see Section Physical Health Prophylaxis in Psychiatry). Psychological interventions do hold some promise, not so much in preventing transition to psychotic illness, but in helping the individual deal with distressing symptoms.

Moving on from the UHR approach, there has been worldwide interest and excitement about the potential for early intervention using a series of established treatments (including antipsychotic medications, psychological treatments such as cognitive behavior therapy and family treatment) in ameliorating the long-term course of schizophrenia. Part of this approach has been to try to reduce the so-called 'duration of untreated psychosis (DUP),' as longer DUP has been associated with worse outcomes in schizophrenia. A number of trials have suggested that the delivery of comprehensive treatment packages by a specialist team or service can indeed improve some outcomes (e.g., reduced hospitalizations) but the effects are modest as best and largely disappear once the specialist intervention halts (Bosanac et al., 2010). One conclusion is that good treatment is beneficial for schizophrenia irrespective of the phase of illness, and that there is nothing particularly special about the first episode; but other commentators are more generous to the early intervention field, believing that the earlier the intervention, the better the longer-term outcome. Certainly one should not only withhold treatment from people early in the course of an established psychotic illness, but arguably such interventions should be offered to the individual (and their family, where possible) for as long as is required.

Irrespective of the efficacy of otherwise of UHR and early psychosis services, in remains the sad case that many people with schizophrenia continue to relapse and/or experience functional decline and social and vocational dislocation. Maintenance therapy in schizophrenia is provided by long-term treatment. Based on accumulated experience, it was suggested that people with schizophrenia could be looked at as two groups. In first group, maintenance therapy seems to play a 'suppressive role,' and its discontinuation leads to a rapid reappearance of psychotic symptoms. The second putative group comprises individuals in whom the treatment has more of a 'prophylactic' role. In this latter group, discontinuation of treatment leads to a more gradual and remote-in-time recurrence of symptoms (see Gelenberg et al., 1991).

It is estimated that about 10% of remitted medication-free patients with schizophrenia will relapse in any given month. That means that after 6 months, roughly about half of these patients will relapse, whereas after 12 months about 70% will do so, and only a small portion will not relapse at the end of 2 years. Several strategies of maintenance treatment have been suggested for these patients. The first strategy consists of continuous maintenance treatment. This strategy applies relatively lower doses of antipsychotic treatment as compared with that used for acute exacerbation. The second strategy includes 'drug holidays,' that is, treatment periods interrupted by drug-free periods. This strategy was suggested in order to lower the incidence of side effects while retaining drug efficiency. However, such a strategy is not recommended nowadays since research studies could not support its effectiveness. The third strategy suggests 'intermittent' treatment or 'targeted dose' for chronic schizophrenia patients. In this case, patients are treated during the acute episode only, while monitored for the appearance of prodromal symptoms (first symptoms anticipating the appearance of an exacerbation). Once prodromal symptoms appear, pharmacological treatment is initiated. Again, this strategy does not have general support in the literature or clinical practice.

In general, pharmacological treatment in schizophrenia includes the use of a group of drugs collectively called 'antipsychotics.' These drugs have the ability to block neurons using the neurotransmitter dopamine to transmit messages to other neurons (dopamine D2 receptors). While effective for the positive symptoms of schizophrenia, these agents are not terribly effective for the negative symptoms (withdrawal, apathy, restriction of affect) and also tend to cause extrapyramidal side effects (distressing motoric problems such as stiffness, tremor, and severe restlessness) and also raised plasma prolactin.

In the 1980s, a second generation of drugs came into use. The prototype was clozapine (actually introduced initially in the 1970s but withdrawn because of hematological problems in most countries, to be reintroduced in the late 1980s with various safety requirements). Clozapine is a drug that, besides its ability to block some types of dopaminergic receptors, can also exert a blocking effect on certain serotonergic receptors – labeled as 5-HT2 neurons – in the central nervous system. This drug was shown to have a lower side-effect profile compared to the more traditional drugs; especially it has minimal extrapyramidal side effects and does not elevate plasma levels of prolactin. However, clozapine also has several distressing side effects including the lowering of white blood cell count that may expose these subjects to severe infections; myocarditis and cardiomyopathy, weight gain and diabetes, and profound sedation, to name a few. Clozapine seems to ameliorate not only the positive illness symptoms, but also the negative symptoms of schizophrenia (a set of symptoms characterized by impaired volition and motivation, apathy, and paucity of speech), and it may also improve some neurocognitive impairments demonstrated in this disease (see Kane et al., 1988).

Beside clozapine, other second-generation or 'atypical' antipsychotic drugs have also been developed, including risperidone, olanzapine, quetiapine, paliperidone, and lurasidone. Other atypical antipsychotics with unique mechanisms of action are amisulpride (a selective dopamine D2/3 blocker) and aripiprazole (a partial D2 antagonist). There has been much debate about whether these newer agents are truly more effective or better tolerated than the older (and much cheaper!) typical drugs, but most Western countries have embraced them and they have become the mainstays of prevention of relapse in people with schizophrenia (Castle et al., 2013). A number of problems remain, however, including side effects such as weight gain and diabetes with some of these agents (see Section Physical Health Prophylaxis in Psychiatry) as well as problems with medication adherence (addressed below).

Also, it goes without saying that medications are not the sole effective interventions for people with schizophrenia; indeed, an holistic multidisciplinary approach dealing with psychological, vocational, and family issues is required to achieve the best outcomes for people with schizophrenia. This is also consonant with the so-called 'Recovery Movement,' which aims to empower people with a mental illness gain better control over their lives in a broader sense, beyond simply focusing on symptoms of illness. A number of innovative interventions are trying to deliver on this: some include using people with the 'lived experience' of mental illness to act as peer support workers. A particular additional issue in many jurisdictions is that people with schizophrenia use illicit substances to excess, and there is good evidence that this is associated with worse outcomes and more relapses; again, concerted attempts are required to engage people with schizophrenia in reducing such harmful drug use.

Maintenance in Bipolar Illness

Bipolar disorder is characterized by the appearance of manic episodes with or without the occurrence of depressive episodes. It is estimated that the majority of patients who had a manic episode will have one or more recurrences. Since both manic and depressive episodes may be associated with progressive deterioration of interpersonal and occupational functioning, and may also carry the risk of harming oneself or others, an adequate maintenance treatment for bipolar illness is of utmost importance.

In terms of medication, several mood stabilizers – that is, drugs designed for maintenance of the euthymic (normal) state while preventing the appearance of both manic and depressed episodes – are available. The first mood stabilizer to be used was lithium. Other commonly used mood stabilizers are valproic acid and its derivatives, and carbamazepine. The anticonvulsant lamotrigine has an established place in prophylaxis for bipolar depression. Antidepressants are generally not favored as prophylaxis in people with bipolar disorder, as they can make mood unstable: certainly they should not be used in the absence of a mood stabilizer. The atypical antipsychotics (see Section Schizophrenia) are increasingly used for people with bipolar disorder, both in acute phase as well as maintenance (Malhi et al., 2010).

Lithium bears particular attention as it was the first mood stabilizer, and is arguably still the 'gold standard' treatment. Three early open trials suggested that lithium may be effective in the prophylaxis of bipolar disorders (Baastrup, 1964; Hartigan, 1963; Baastrup and Schou, 1967). During the 1970s, a series of double blind studies were conducted to examine the possible efficacy of lithium in the prophylaxis of bipolar disorder. Goodwin (1994) summarized these studies calculating the recurrence rates over a 1-year period for placebo (a nonactive treatment) to be 81%, but only 34% for lithium. These reports included two types of studies: prospective studies treating patients with either lithium or placebo; and discontinuation studies where one group of patients continues with lithium treatment for a given period of time, whereas the other discontinues treatment for a similar time period.

From the 1980s and on, more prospective and discontinuation studies were conducted. These studies basically confirmed lithium's efficacy as maintenance treatment for bipolar disorder. They also suggested that there may be an advantage in keeping maintenance lithium blood levels at about 0.6 meq/l, as it may be associated with fewer side effects and improved adherence (Coppen et al., 1983), and that lithium plasma levels of 0.8–1.0 meq/l seem to be associated with lower rates of relapse (Gelenberg et al., 1989). In general, although many authors still suggest that lithium is an efficient drug for bipolar illness, especially in the classic bipolar patient (Grof, 1999), some experts now point to the fact that there is a group of patients who are less responsive to this drug or that demonstrate a decreased response throughout the years. There is also the problem of potential renal dysfunction in conjunction with lithium use.

One important issue associated with long-term maintenance therapy is the finding of predictors for a beneficial, or lack of, response for a given drug. Several predictors of lithium response were offered. Grof et al. (1983a,b, 1999) suggested a better response to lithium treatment in a group of patients demonstrating a course of illness with periods of euthymia (normal range of mood), followed by mania, depression, and then again euthymia. These authors also suggested that the absence of rapid cycling (four and more illness episodes per year), the absence of personality disorder, and the diagnosis of primary bipolar disorder (i.e., not secondary to street drugs or organic cause) predict a good prophylactic response to lithium. Faedda et al. (1991) in agreement with Grof suggested that mania–depression–euthymia course is associated with a good response to lithium. Others suggested that good response to lithium in relatives and a family history of mood disorders carry a good response. Aagaard and Vestegaard (1990) suggested that bipolar patients tend to demonstrate decreased response to lithium treatment when they have a lengthy episode, they are young or of female gender. Abou-Saleh and Coppen (1990) suggested that positive prophylactic response to lithium over 6–12 months predicts further good response in the future. Patients with dysphoria and mixed mania (manic episode associated with some depressive symptomatology) were suggested to respond less to lithium treatment, and some studies suggest this is also true for rapid cycling bipolar disorder. Hopefully, a knowledgeable use of predictors may help the physician to choose selected subgroups of patients, who will benefit from lithium maintenance treatment and who will require other medications or combinations thereof. For a recent pragmatic review of the pharmacological treatment of bipolar disorder, the reader is directed to Malhi et al. (2010).

Again, it is important to appreciate that even with excellent medication adherence, people with bipolar disorder remain very vulnerable to the recurrence of both manic and depressive episodes. There is increasing excellent evidence to support the adjunctive use of targeted psychosocial interventions as prophylaxis. Strategies include psychoeducation, mood monitoring, mapping early warning signs, cognitive behavior therapy, and interpersonal and social rhythm therapy. The interested reader is referred to the clinical review by Lauder et al. (2010).

Maintenance Treatment in Major Depression

Depression is a serious illness, with substantial suffering and disability and a high suicide risk. Recurrence rates in people with major depression have been estimated to be as high as 80%. Two types of studies report the efficacy of antidepressants to prevent relapse in the first year after remission from a depressive episode. The first is a crossover design, while the second employs a continuation design. In the first type of study, patients are first brought into remission with a given medication and are then maintained on that medication for a period of 2–3 months. Patients who respond are then entered into a double blind phase for up to another year, in which they receive randomly either the same antidepressant or placebo. In the continuation design, acutely depressed patients are assigned randomly to either active drug or placebo. Patients who respond continue with the same treatment. Eighteen studies compared a group of antidepressants collectively called tertiary amine tricyclics (this name describes the common molecular structure characterizing this group of drugs) finding that while 50% relapsed on placebo, only 23% relapsed on the active drugs. Three studies compared another group of newer antidepressants collectively called selective serotonin reuptake inhibitors (this name describes the hypothesized mechanism of action of this group of drugs) finding that the relapse rate for the active drug was 13–26% whereas for placebo it was 46–57% (see Janicak et al., 1997). More recent studies have shown prophylactic efficacy for the newer antidepressants, including the serotonin–noradrenaline reuptake inhibitors, the presynaptic anticholinergic receptor antagonist mirtazapine, and the melatonin M1 receptor agonist agomelatine.

In general, it is agreed that prevention of relapse in major depressive illness is of major importance and every effort must be taken to address this. Contrary to the early reports that supported the use of low doses of medication in the prophylactic treatment of major depression, several authors now recommend the use of full dose; that is, a similar dose to that used for the acute treatment of a major depressive episode.

In general, it is recommended to continue maintenance treatment after a first episode of major depression for about 6 months and then to taper the drug gradually. If another episode occurs, the recommendation is to continue treatment for several years, and after a third episode lifetime maintenance treatment is recommended.

Physical Health Prophylaxis in Psychiatry

It is increasingly recognized that people with mental illnesses often suffer an undue burden of underrecognized and undertreated physical morbidities. In people with schizophrenia, for example, life expectancy is some 15–20 years lower than that of the general population. Most of this excess in early death is due to cardiovascular illness. In a recent national survey of people with psychotic disorders conducted across seven sites in Australia, around a quarter of respondents were considered at high cardiovascular disease risk, with high rates of undertreatment (Galletly et al., 2012).

These distressing statistics are compounded by the fact that a number of the medications used to manage the symptoms of mental illness over protracted periods cause or exacerbate risk factors such as obesity, diabetes, and hypercholesterolemia. A number of the atypical antipsychotics are particularly implicated in this regard, but valproic acid and lithium also carry the risk of weight gain and other medical problems.

In considering prophylaxis in psychiatry, then, it is imperative to ensure adequate psychoeducation as well as effective monitoring and targeted interventions for physical health problems. Various strategies have been looked at in this regard, with the most successful being where physical health is imbedded into mainstream mental health services (e.g., Organ et al., 2010), with agreed care pathways and referral networks to other health professionals (including dieticians, exercise physiologists, and medical specialists such as endocrinologists, as required). Primary care providers can play a key role in optimizing the physical health of the mentally ill.

Adherence and Its Importance in Prophylaxis in Psychiatry

Adherence to drug treatment is defined as adherence to an agreed treatment schedule between a physician and a patient. It is estimated that not more than 50% of patients fully adhere to their treatment and it is suspected that the magnitude of non-adherence may be even greater in the case of long-term maintenance treatment. Various factors influence adherence, including the nature of the patient–physician relationship, the patient's set of beliefs and expectations, as well as the occurrence of side effects. Issues of particular relevance to adherence in psychiatric disorders are (1) the occurrence of delusions and severe cognitive distortions about the index drug and its effects and side effects; (2) the patient's cognitive functioning and motivation, which may be impaired in certain psychiatric disorders; and (3) the patient's general attitude toward mental illness. Especially important in this context is the level of stigma felt by the patient: the greater the stigma, the less adherent the patient is likely to be.

Several strategies can be offered to enhance adherence and these should be tailored according to the patient's condition and level of understanding. First, an empathic approach, one that builds basic trust with the patient and their family is of utmost importance. Second, whenever possible, the patient with the guidance of their physician should be an active participant in decisions made for the treatment plan. This does not mean that the patient decides or dictates such decision making, but that they are presented with several treatment alternatives and an educated decision is agreed upon with the guidance of the physician. Third, the patient should be educated as to side effects and ways to cope with those once they appear. Treatment diaries and involvement of family members whenever necessary can also be of help. Since maintenance treatment is long term, it is our opinion that a responsible physician should renew such a drug treatment 'agreement' at least every several months, discussing it with the patient, and perhaps with relevant family and community members if needed. The physician should also actively ask in each meeting with the patient about side effects as well as examine whether there is any change in the patient's set of beliefs about the drug and its action that may need renewed education.

Future Directions for Prophylaxis in Psychiatry

Genetic epidemiological studies have shown that genes play a major role in several psychiatric disorders including schizophrenia, bipolar illness, and major depression. It is now believed that these disorders are not related to a single gene, but rather are determined polygenetically. In general, it seems that gene–environment interactions are involved in the causation of many psychiatric disorders. Interestingly, the vulnerability to environmental influences may also be genetically determined (see Owen et al., 2000, for review).

The human genome project, now completed, is expected to contribute to the finding of genes responsible for the susceptibility, occurrence, and maintenance of these disorders. The sorting out of genetic variations in the population, which are related to these psychiatric disorders, may open new avenues to genetic medical interventions prior to the manifestation of these diseases. Such interventions may alter the course or even prevent these diseases from occurring. Although this project is still far off, one cannot ignore it at the dawn of this new era, which started with the recent completion of the decoding of the human genome.

See also: Antidepressant Drugs; Bipolar (Manic-Depressive) Disorder; Depression; Patient Adherence to Health Care Regimens; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Public Psychiatry and Public Health Mental Systems; Schizophrenia, Treatment of; Schizophrenia.

Bibliography

Aagaard, J., Vestegaard, P., 1990. Predictors of outcome in prophylactic lithium treatment: a two-year prospective study. Journal of Affective Disorders 18, 259–266.

Abou-Saleh, M.T., Coppen, A.J., 1990. Predictors of long term outcome of mood disorder on prophylactic lithium. Lithium 1, 27–35.

Arseneault, L., Cannon, M., Witton, J., Murray, R., 2004. Cannabis as a potential causal factor in schizophrenia. In: Castle, D.J., Murray, R.M. (Eds.), Marijuana and Madness. Cambridge University Press, Cambridge, pp. 101–118.

Asarnow, J.R., 1988. Children at risk for schizophrenia: converging lines of evidence. Schizophrenia Bulletin 14, 613–631.

Baastrup, P.C., 1964. The use of lithium in manic depressive psychosis. Comprehensive Psychiatry 5, 396–408.

Baastrup, P.C., Schou, M., 1967. Lithium as a prophylactic agent. Archives of General Psychiatry 16, 162–172.

Bosanac, P., Patton, G.C., Castle, D.J., 2010. Early intervention in psychotic disorders: faith before facts? Psychological Medicine 40 (3), 353.

Castle, D., 2012. Is it appropriate to treat people at high risk before the onset of psychosis? Medical Journal of Australia 196, 557.

Castle, D.J., Buckley, P.F., 2015. Schizophrenia, second ed. Oxford University Press, Oxford.

Castle, D.J., Sham, P., Murray, R.M., 1998. Differences in distribution of ages of onset in males and females with schizophrenia. Schizophrenia Research 33, 179–183.

Castle, D.J., Keks, N., Newton, R., Schweitzer, I., Copolov, D., Paoletti, N., Burrows, G.D., Tiller, J., 2013. Pharmacological approaches to the management of schizophrenia: Ten years on. Australasian Psychiatry 21, 329–334.

Coppen, A., Abou-Saleh, M., Milln, P., Bailey, J., Wood, K., 1983. Decreasing lithium dosage reduces morbidity and side effects during prophylaxis. Journal of Affective Disorders 5, 353–362.

Erlenmeyer-Kimling, L., Cornblatt, B.A., Rock, D., Roberts, S., Bell, M., West, A., 1993. The New York high risk project: anhedonia, attentional deviance and psychopathology. Schizophrenia Bulletin 19, 141–153.

Faedda, G.L., Baldessarini, R.J., Tohen, M., Strakowski, S.M., Waternaux, C., 1991. Episode sequence in bipolar disorder and response to lithium treatment. American Journal of Psychiatry 148, 1237–1239.

Fusar-Poli, P., Yung, A.R., McGorry, P., van Os, J., 2014. Lessons learned from the psychosis high-risk state: towards a general staging model of prodromal intervention. Psychological Medicine 44 (1), 17–24.

Galletly, C.A., Foley, D.L., Waterreus, A., Watts, G.F., Castle, D.J., McGrath, J.J., Morgan, V.A., 2012. Cardiometabolic risk factors in people with psychotic disorders: the second Australian national survey of psychosis. Australian and New Zealand Journal of Psychiatry 46 (8), 753–761.

Gelenberg, A.J., Bassuk, E.L., Schoonover, S.C. (Eds.), 1991. The Practitioner's Guide to Psychoactive Drugs, third ed. Plenum, New York, pp. 170–176.

Gelenberg, A.L., Kane, J.M., Keller, M.B., Lavori, P., Rosenbaum, J.F., Cole, K., Lavelle, J., 1989. Comparison of standard and low serum levels of lithium for maintenance treatment of bipolar disorder. New English Journal of Medicine 321, 1489–1493.

Goodwin, G.M., 1994. Recurrence of mania after lithium withdrawal. British Journal of Psychiatry 164, 149–152.

Grof, P., 1999. Excellent lithium responders people whose lives have been changed by lithium use. In: Birch, W.J., Gallictio, V.S., Becker, K.W. (Eds.), Lithium: 50 Years of Psychopharmacology. Weidner Publishing Group, Cheshire, CT, pp. 36–51.

Grof, P., Alda, M., Grof, E., Fox, D., Cameron, P., 1999. The challenge of predicting response to stabilizing lithium treatment. British Journal of Psychiatry 163 (21), 16–19.

Grof, P., Hux, M., Grof, E., Arato, M., 1983a. Prediction of response to stabilizing lithium treatment. Pharmacopsychiatry 16, 195–200.

Grof, P., Lane, J., Hux, M., 1983b. Predicting the response of long term lithium treatment: interim results of a prospective study, in the prediction of lithium response. In: Dufour, H., Pringuey, D., Milech, T. (Eds.), Proceedings from Symposium in Marseille, France; June 2–3, 1980. France University, Marseille, pp. 43–53.

Hartigan, G.P., 1963. The use of lithium salts in affective disorders. British Journal of Psychiatry 109, 810.

Hearle, J., McGrath, J., 2000. Reproductive, preconceptual and antenatal needs of women with schizophrenia. In: Castle, D.J., McGrath, J., Kulkarni, J. (Eds.), Women and Schizophrenia. Cambridge University Press, Cambridge, pp. 79–95.

Janicak, P.G., Davis, J.M., Preskorn, S.H., Ayd, F.J., 1997. In: Principles and Practice of Psychopharmacotherapy, second ed. Williams & Wilkins, Baltimore, MD, pp. 295–301.

Kane, J., Honigfeld, G., Singer, J., Meltzer, H., 1988. Clozaril collaborative study group: clozaril for the treatment-resistant schizophrenic: a double-blind comparison with chlorpromazine. Archives of General Psychiatry 45, 789–796.

Lauder, S.D., Berk, M., Castle, D.J., Dodd, S., Berk, L., 2010. The role of psychotherapy in bipolar disorder. Medical Journal of Australia 193 (Suppl. 4), S31–S35.

Lawrie, S.M., Whalley, H., Kestelman, J.N., Abukmeil, S.S., Byrne, M., Hodges, A., Rimmington, J.E., Best, J.J., Owens, D.G., Johnstone, E.C., 1999. Magnetic resonance imaging of brain in people at high risk of developing schizophrenia. Lancet 353 (9146), 30–33.

Malhi, G.S., Adams, D., Berk, M., 2010. The pharmacological treatment of bipolar disorder in primary care. Medical Journal of Australia 193 (Suppl. 4), S24–S30.

Mednick, S.A., Parnas, J., Schulsinger, F., 1987. The Copenhagen high-risk project, 1962–1986. Schizophrenia Bulletin 13 (3), 485–495.

Organ, B., Nicholson, E., Castle, D., 2010. Implementing a physical health strategy in a mental health service. Australasian Psychiatry 18 (5), 456–459.

Owen, M.J., Cardo, A.G., O'Donovan, M.C., 2000. Psychiatric genetics: back to the future. Molecular Psychiatry 5, 22–31.

Phillips, L., Yung, A., Hearn, N., McFarlane, C., Hallgren, M., McGorry, P., 1999. Preventive mental health care: accessing the target population. Australian and New Zealand Journal of Psychiatry 33, 912–917.

Shimushi, T., Kurachi, M., Kurokawa, K., Yotsctsugi, T., Urehara, T., Itch, H., Saitoh, O., 2000. Plasma homovanillic acid in the prodromal phase of schizophrenia. Biological Psychiatry 47, 428–433.

Weintraub, S., 1987. Risk factors in schizophrenia: the stony brook high-risk project. Schizophrenia Bulletin 13 (3), 439–450.

Werbeloff, N., Drukker, M., Dohrenwend, B.P., Levav, I., Yoffe, R., van Os, J., Weiser, M., 2012. Self-reported attenuated psychotic symptoms as forerunners of severe mental disorders later in life. Archives of General Psychiatry 69 (5), 467–475.

Yung, A.R., Phillips, L.J., McGorry, P.D., McFarlane, C.A., Francey, S., Harrigan, S., Patton, G.C., Jackson, H.J., 1998. Prediction of psychosis. A step towards indicated prevention of schizophrenia. British Journal of Psychiatry 172 (Suppl.), 14–20.

Proportional Representation

Jack H Nagel, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Proportional representation (PR) refers to electoral systems designed to approximate the ideal of proportionality in converting citizens' votes into legislative seats. All PR systems require multimember constituencies. In the most common form, list PR, electors vote for lists of candidates designated by parties. In a hybrid variant, mixed-member proportional, about half the representatives are elected from single-member districts and half from party lists, but seats won in districts are subtracted from list allocations so they do not diminish overall proportionality. The least common version of PR, the single-transferable vote, permits voters to rank individual candidates, crossing party lines if they wish. The degree to which a PR scheme approaches proportionality depends on the number of representatives elected per constituency, the formula used to convert votes into seats, the compensatory use of higher tiers of representation above the basic constituencies, and the application of vote thresholds for representation. PR methods were first adopted for national elections by Belgium in 1899. They are now employed in many countries around the world and are dominant in Europe and Latin America.

As a standard for making political evaluations, 'proportional representation' (PR) implies that representatives of a group (R_g) participate in a legislature or other governmental entity in numbers that bear the same ratio to the total membership of that body (R) as persons belonging to or identifying with the group (P_g) have to the total population governed (P) – thus, $R_g/R = P_g/P$. The groups in question may be residents of territorial subdivisions; persons sharing identities based on race, ethnicity, religion, language, or gender; or adherents of ideological or policy tendencies; but most often the relevant groups are political parties. 'PR' also, and more commonly, refers to methods of electing representatives designed to achieve an approximation to that standard in converting citizens' votes into seats in a legislature or council. In this article, 'PR' will indicate such electoral systems, while the ideal at which they aim will be referred to as 'proportionality.'

Although the concept of proportionality is simple, PR electoral systems are highly diverse, varying with choices made from a menu of design elements, some of which are complex and nonobvious. One feature that all PR systems require is the use of multimember districts – constituencies from which more than one representative is to be chosen. Therefore, except in the limiting sense that the winner should have more votes than any loser, PR does not apply to elections that choose a single winner, such as contests for executive posts (e.g., a president, governor, or mayor) or for legislative seats from single-member districts (SMDs). In the study of legislatures, a bifurcation is thus often made between PR and SMD systems, the former sometimes called 'proportional' and the latter 'majoritarian,' although design choices within those two broad families affect the degree to which either label is actually appropriate.

Three Types of PR Electoral Systems

Despite great and often idiosyncratic variation, PR electoral systems can be divided into three broad types, of which the first is by far the most common.

List PR

In the pure form of list PR, the 'closed list,' political parties designate before the election a rank-ordered list of candidates. Electors vote for a party, not for individual nominees. Each party wins seats proportional to the votes cast for its list. For example, if 10 representatives are to be elected from a constituency, a party winning 20% of the vote should be awarded two seats; these will go the candidates ranked first and second on the party's list. List PR systems are employed for national parliamentary elections in most countries of continental Europe and Latin America, as well as some countries in other regions.

The Single-Transferable Vote

Single-transferable vote (STV), also known as the quota-preferential system, depends on a preferential ballot, which asks voters to rank individual candidates (not parties) from most to least preferred. A quota of votes, Q, is established. All candidates receiving Q or more first preferences are elected. To determine the remaining winners, two types of transfers are carried out. First, 'surplus' votes (those above Q) received by winning candidates are 'transferred' to voters' second (or, if necessary, subsequent) choices. If seats remain unfilled after all surplus votes have been reassigned, then the candidate ranking last in votes is eliminated, and ballots giving first preferences to him or her are transferred to the voters' second (or subsequent) choices. These transfers are iterated until quotas are attained by the required number of candidates. Perhaps because of its complexity, both in balloting and in the procedure for selecting winners, STV is employed in national elections by just a few countries – Ireland, Malta, and (for the Senate only) Australia.

Mixed-Member Proportional

In an effort to combine the virtues of proportionality and local representation, West Germany after World War II pioneered the hybrid 'mixed-member proportional' (MMP) system. In MMP, some representatives (half in Germany) are elected

from single-member districts, and the rest are elected from party lists. Electors cast two votes – one for an individual constituency representative and one for a party. Seats that parties win from constituencies are subtracted from their list allocations to make overall delegations proportional to party votes. This crucial 'compensatory' feature makes MMP a truly proportional system.

Semiproportional Systems

Several systems that may be described as 'semiproportional' are intended to offer some representation to minority parties or groups, but offer less likelihood of approaching true proportionality, compared with the three main types of PR.

Parallel Systems

A second mixed type, easily confused with MMP, is the 'parallel' system (also known as mixed-member parallel and mixed-member majoritarian). Like MMP, parallel systems elect some legislators from single-member constituencies and some from PR lists. Parallel systems diverge from MMP by not compensating for constituency victories in allocating seats to party lists. In effect, the legislature is chosen by an amalgamation of two separate electoral systems, only one of which aims at proportionality. If, as is usually the case, seats chosen by PR comprise no more than half the assembly, then the parallel system is appropriately classified as semiproportional. Japan and Russia both adopted parallel systems in 1993, but Russia switched to list PR beginning with its 2007 election.

Limited and Cumulative Voting

The term 'semiproportional' is also applied to electoral systems that are intended to achieve a degree of proportionality, but do not employ PR formulas to do so. These systems use multimember districts, voting for individual candidates, and the plurality formula. The opportunity for more proportional outcomes arises from their unusual balloting rules. The 'limited vote' requires electors to cast fewer votes than the number of representatives to be elected (M). In the extreme version, known as the 'single-nontransferable vote' (SNTV), electors may vote for only one candidate, even though $M > 1$ will be elected. Japan employed SNTV until 1993, and Afghanistan adopted SNTV early in the twenty-first century for elections to the House of the People. 'Cumulative' voting gives electors M votes but allows them to 'cumulate' votes, by casting more than one of the votes for a favored candidate or candidates. Cumulative voting is not used for any national legislature, but it is employed in some U.S. local elections. Limited voting and cumulative voting, which have similar underlying strategic properties, enable well-coordinated minorities to win seats. However, because parties must avoid dispersing votes among too many candidates, these systems tend to reduce competition, and such competition as does occur tends to be within parties or voting blocs, causing internal conflict and factionalism.

Design Elements of PR Systems

In establishing any system of PR, choices must be made about various essential elements. These decisions affect the degree of proportionality that elections can deliver, the ability of smaller parties to win representation, and other aspects of how the system will function in practice.

District Magnitude

The degree to which a PR electoral scheme can approach perfect proportionality depends most strongly on the number of representatives elected per constituency, or 'district magnitude' (M). For example, if Parties A, B, and C receive 50, 40, and 10% of the vote respectively, and $M = 2$, then most PR formulas will give A and B one seat each, while shutting out C entirely – a very crude approximation to proportionality. In contrast, if $M = 100$, A can receive 50 seats, B 40, and C 10, resulting in perfect proportionality. The greatest magnitudes occur when the entire nation constitutes a single constituency, as in Israel and the Netherlands, where M equals 120 and 150, respectively. In most PR systems, however, the country is divided into multiple constituencies with lesser magnitudes, to ensure representation for different geographic areas.

Formulas

Although perfect proportionality is easy to achieve in contrived examples like the one above, in real elections the numbers seldom work out so evenly. For example, if $M = 10$ and Party C wins 15% of the vote, should it receive one or two representatives? To solve such problems, PR systems in practice employ mathematical formulas or algorithms that allocate seats to lists. These formulas employ one of two basic mechanisms – divisors or quotas. In divisor (or highest average) methods, seats are assigned sequentially. Each seat goes to the party which, at that point in the allocation, has the largest quotient after its votes are divided by a divisor (D) that varies with the number of seats the party has already received (S). Well-known divisors include the D'Hondt formula, $D = S + 1$, thus producing the sequence of divisors 1, 2, 3, …; and the Sainte-Laguë formula, $D = 2S + 1$, resulting in the sequence 1, 3, 5, …. In quota (or largest remainder) formulas, parties receive an initial allocation of seats equal to the number of quotas achieved in the vote for their list. The quota (Q) is determined by dividing total votes by M (Hare or natural quota), $M + 1$ (Droop quota), or $M + 2$ (Imperiale quota). Any remaining seats are assigned according to largest remainders – that is, by the size of parties' leftover votes after subtracting the quotas they have already used. Quotas are also employed to award seats in STV.

Over the years, PR formulas have inspired much political controversy, because different divisors and quotas favor large, medium, or small parties – for example, D'Hondt tends to give more seats to larger parties, whereas Sainte-Laguë approximates proportionality more closely, thus benefiting smaller parties. Nevertheless, district magnitude has a far more powerful effect on parties' fortunes; small parties generally do better as M rises.

Besides their use to determine the winners of legislative elections, PR formulas have two other important applications.

They can be used for 'apportionment' – the assignment of seats to districts (e.g., states) in proportion to population, as is required for the U.S. House of Representatives following each decennial census. In fact, several of the major PR formulas were first devised by American statesmen (Jefferson, Hamilton, and Webster) to apportion House seats prior to their later, independent invention in Europe to allocate seats based on electoral results (Balinski and Young, 2001). More recently, the D'Hondt divisor rule has been employed to structure the sequential allocation of cabinet portfolios to parties in Northern Ireland as a crucial part of the 1998 Good Friday peace accords. Similar procedures are used to assign executive responsibilities to council members in some Danish municipalities and for choosing committee chairs in the European Parliament (O'Leary et al., 2005).

Tiers

The use of relatively low-magnitude constituencies offers better geographic representation but may result in considerable departures from proportionality when votes and seats are compared nationwide. As a corrective, some PR systems include a second, national 'tier' of representatives (sometimes called 'additional' or 'supplementary' members) who are selected in a compensatory fashion. That is, parties that are underrepresented after assignment of constituency seats receive more of the second-tier seats, so as to establish a closer approximation to proportionality across the country as a whole. Except for Finland, all the Scandinavian countries use a second tier of seats for this purpose. The concept is similar to MMP, except that tiered PR systems generally ask citizens to cast only one vote, not two as in MMP; and MMP does not attempt proportionality at the constituency level. As a result, compared with two-tier list PR, MMP systems require a higher percentage of compensatory seats to achieve reasonable overall proportionality.

An implicit second tier is also involved in a novel proposal that, at first glance, appears to defy the usual axiom that PR requires multimember constituencies. That is the plan for 'fair majority voting' devised by Balinski (2008) as a way to neutralize partisan gerrymandering in the United States, where a federal statute mandates that all states use single-member districts for elections to the House of Representatives. Under Balinski's plan, elections would be conducted in each district, just as they are under the current system. However, votes for district candidates would also be tallied at the state level and credited to the candidates' parties. If necessary to achieve greater proportionality statewide, certain plurality winners at the district level might be denied seats, which would be assigned instead to trailing candidates whose parties' allocations would otherwise be less than proportionate. Thus the state serves as a *de facto* second tier, but the lack of supplementary members at the state level necessitates the counterintuitive step of giving seats to district losers.

Thresholds

PR systems with high-magnitude districts, nationwide lists, or compensatory upper tiers can enable numerous small parties to win parliamentary seats. Critics of such 'permissive' PR systems condemn proliferation of small parties, contending that it confuses voters, complicates coalition formation, renders governments less stable, and obscures accountability for governmental performance. Perhaps the most influential objection is that lack of barriers to entry can award legislative seats to extremist, antidemocratic parties. Thus Hermens (1941) maintained that Weimar Germany's excessively permissive version of PR enabled the Nazi Party to gain a foothold in the Reichstag, which they exploited to establish a totalitarian regime. When democracy was restored in West Germany after World War II, the new MMP system included a requirement that a party must win at least 5% of the party vote to seat candidates from its party list. Such formal 'thresholds' are now common in list PR and MMP systems; they range from 0.67% in the Netherlands to 10% in Turkey. Low-magnitude systems, including STV, have no need of a formal threshold, because any PR formula applied to a small M imposes a high *de facto* threshold. The MMP systems of Germany and New Zealand also include an alternative threshold – parties can win list seats either by attaining the threshold percentage of the party vote or by winning a specified number of constituency seats (three in Germany and one in New Zealand). In New Zealand, this 'back-door' threshold has frequently enabled very small parties to receive list seats (Nagel, 2012).

When small parties or voters have not fully adapted to thresholds – the former by amalgamating and the latter by refraining from voting for parties likely to fall short of the requirement – then the imposition of a threshold can result in many 'wasted votes' and substantial departures from proportionality in an ostensibly PR system. Such outcomes are more likely during the early stages of democracy, as in several East European countries during the 1990s (Birch, 2003).

Choices in Balloting

PR systems also differ in the choices they allow voters to make as they cast ballots. 'Closed-list' PR requires voters to designate only their preferred party – effectively, a vote for its entire list in the preference order established by the party. In contrast, 'open-list' PR systems permit (or sometimes require, as in Brazil and Finland) voters to indicate preferences for individual candidates on their chosen party's list. In the 'panachage' system of Switzerland and Luxembourg, electors may even distribute votes for candidates across different party lists. In the STV version of PR, votes are cast for individuals, and electors may move back and forth across parties while ranking candidates. Closed-list PR assigns control over the fate of individual candidates to parties' internal decision processes (which determine their position on the list and thus their prospects for election), whereas STV reserves that power for voters, and open-list PR occupies an intermediate position. Consequently, closed-list PR strengthens party organization and promotes party discipline in the legislature, whereas STV and (usually to a lesser degree) open-list PR are conducive to intraparty electoral competition and individualistic behavior among legislators.

In practice, however, these differences may be blurred. In STV systems, unless district magnitude is kept low (on the order of three to five representatives per constituency), the task of ranking numerous candidates becomes so difficult that most voters will follow instructions given by their parties. That is the case in STV elections for the Australian Senate, where M equals 6 or sometimes 12, and voters must submit a complete ranking

of all candidates, who frequently total 30 or more. Thus, if it imposes excessive cognitive demands, a system intended to maximize the autonomous power of individual voters may end up functioning rather like closed-list PR.

Historical Development

Because PR methods are nonobvious and complex, their development stands as one of the key intellectual achievements in the history of democracy. The Marquis de Mirabeau enunciated the principle of proportionality in 1789, but methods designed to achieve it were not refined until well into the nineteenth century. In 1834 Victor Considérant devised a list PR system that Antoine Morin of Switzerland developed further in 1862. In 1885 an international conference on PR in Antwerp recommended the list PR method proposed a few years earlier by a Belgian, Victor D'Hondt. In 1899, Belgium adopted D'Hondt's system and thus became the first country to use list PR. By 1920 most countries of western Europe had adopted versions of PR (Carstairs, 1980).

Beside the intellectual and moral appeal of proportionality, two practical motives spurred the turn to PR in Europe. In societies divided along linguistic or religious lines, such as Switzerland, Belgium, and the Netherlands, PR offered all groups a stake in democracy. For countries about to expand the franchise, PR assured ruling groups and conservative parties that they would not be shut out of parliament, as might have occurred under existing plurality and majority electoral systems once previously disenfranchised citizens became a majority of voters. These motives also influenced later decisions to adopt PR, including the notable cases of South Africa in 1994 and Iraq in 2005.

Perhaps because of their long history of elections predating the development of PR, the United Kingdom and the United States (as well as some countries they influenced, including India and Canada) have clung to the older, simpler method of single-member districts and plurality rule (also known as 'first-past-the-post'). Reformers who promote PR in English-speaking countries have usually favored STV, reflecting its invention by the Englishman Thomas Hare in 1857 and John Stuart Mill's enthusiastic advocacy. That tradition was broken in 1993 when New Zealand replaced first-past-the-post with MMP. Subsequently, MMP was adopted for the new national assemblies in Scotland and Wales; but efforts in the early twenty-first century to establish MMP in the Canadian provinces of British Columbia, Québec, and Ontario fell short of approval. The major gain for PR in that period was Russia's switch from a parallel system to list PR.

Agreement on a list PR system was often a key ingredient in democratizations during the late twentieth century, as in Portugal, Spain, South Africa, and Indonesia. However, in the 1990s the global trend shifted to mixed electoral systems, both the semiproportional parallel version and fully proportional MMP. By 2013, of the world's 213 countries and territories with direct elections to the national legislature, 92 (43.2%) used PR systems (81 list PR, 8 MMP, and 2 STV) to elect the principal chamber. Another 25 countries used semiproportional systems (23 parallel and 2 SNTV) (International IDEA). In many nations, PR methods are also used for elections to upper chambers, regional and local councils, or supranational authorities, such as the European Parliament. However, on a population basis, SMD plurality (first-past-the-post) remains the most widely used method for electing national legislatures, mainly because it continues to be employed by the two most populous democracies, India and the United States (Reynolds et al., 2005: pp. 29–33).

Effects and Evaluations

The obvious intended benefit of PR systems is to establish what advocates consider an essential condition of electoral fairness – a close correspondence between parties' votes and seats. The use of a PR formula does not preclude significant deviations from proportionality when thresholds are high or district magnitudes are low, especially if parties have not yet adapted strategically to the parameters of the electoral system. However, PR systems as a group do succeed in delivering substantially more proportional outcomes than their chief rivals.

Although the immediate effect of list PR and MMP is to promote proportionality for political parties rather than underlying population groups, most parties will compete for votes of minority groups by placing members in electable list positions. If they do not, groups large enough to surpass the threshold can win representation by organizing their own parties. Consequently, these two versions of PR usually deliver improved representation to minorities. The same incentives apply to electing women. Where they have become politically active and organized, women generally win more legislative seats under list PR and MMP than under SMD systems; there is less evidence about whether STV helps women.

Because PR is less likely than SMD systems to waste votes cast for smaller parties, it usually results in multiparty parliaments. The association between PR and multipartyism must be regarded as permissive rather than automatic. Within the constraints imposed by the electoral framework (district magnitude, formula, tiers, and threshold), the number of parties depends on the number of societal cleavages and on the willingness and ability of politicians to exploit them by organizing parties. Proportionality and multipartyism together make one-party majorities rare, so parliaments elected by PR typically form coalition governments or minority governments that must get support from other parties to pass legislation. Again, the contrast with SMD systems is not absolute, because such outcomes are not uncommon under first-past-the-post, especially where smaller parties win seats from regional strongholds, as in Canada and India.

Critics of PR usually concede its representational advantages, but they contend that it produces a fragmented party system, unstable coalition governments, and excessive influence by extremist parties. Their favorite negative examples include Italy (before 1993 reforms), Israel, and Weimar Germany. In response, some advocates of PR recommend 'moderate PR' with districts of intermediate magnitude or (equivalently) significant thresholds, rather than the permissive forms those three cases exemplify. For example, Carey and Hix (2011) propose that 'effective magnitudes' (incorporating the effect of thresholds) between 4 and 8 offer a 'sweet spot,' because they deliver almost as much proportionality as higher magnitude systems while producing less fragmented party systems and simpler

governments. In contrast, McGann (2013) argues that low barriers to entry by new parties help protect against corruption and complacency among established parties.

Scholars' empirical evaluations of electoral systems tend to group them into two large families, 'proportional' and 'majoritarian.' Following the influential work of Lijphart (2012), the former is often identified with 'consensus' politics. PR is indeed an essential component – along with super-majoritarian decision rules, federalism, bicameralism, and other checks on majority rule – of the 'consociational democracy' model that Lijphart, O'Leary, and others prescribe for deeply divided societies. However, some of the most successful countries using PR are unicameral parliamentary systems that make legislative decisions by majority rule. These systems require frequent interparty negotiation to form majorities, but such agreements often fall far short of consensus. Moreover, evidence shows that countries within the so-called 'majoritarian' camp – especially those that actually decide elections by plurality rather than majority rule – less consistently form governments based on a majority of voters than do PR systems (Powell, 2000; McDonald and Budge, 2005). For these and other reasons, some scholars (e.g., Nagel, 2000; McGann, 2006) believe that PR elections to a majority-rule legislature provide the best guarantee of majoritarian outcomes.

A recent statistical literature attempts to link economic outcomes to electoral systems. Some of these studies have associated PR with greater income equality, more spending on welfare and pensions, higher governmental spending overall, larger deficits, and higher inflation (Persson and Tabellini, 2004; Rogowski and Kayser, 2002). However, such analyses usually rely on the crude dichotomy between proportional and majoritarian electoral systems, thus not taking into account the considerable variation within each family. Moreover, investigations that are confined to advanced economies (e.g., OECD members) have trouble separating the effect of electoral systems from the cultural, historical, and ideological gulf between Anglo-American and continental European democracies. Research with more extensive data bases includes many countries in Latin America, Africa, and Eastern Europe that have PR legislatures and strong directly elected presidents, a combination that some leading political scientists do not recommend (e.g., Lijphart, 1991).

Although long-standing controversies over the merits of PR continue, evidence suggests that most experts on electoral systems have high regard for PR. Of the nine alternatives presented to 170 experts in a survey, PR variants won four of the top five places. MMP finished a strong first, with 52 first preferences, followed by STV and open-list PR. Closed-list PR was edged out for fourth place by the alternative vote, an SMD system that uses majority rather than plurality rule. SMD plurality, the traditional rival to PR, finished a poor sixth in the evaluation, with only 21 of the 170 experts considering it the best. Overall, nearly 70% of the experts ranked one or another of the PR systems as their most preferred choice (Bowler et al., 2005).

See also: Consociationalism; Elections: Apportionment; Electoral Systems; Majoritarianism and Majority Rule; Parliamentary Government; Party Systems; Political Representation; Voting: Turnout.

Bibliography

Balinski, M.L., 2008. Fair majority voting (or how to eliminate gerrymandering). American Mathematical Monthly 115, 97–113.

Balinski, M.L., Young, H.P., 2001. Fair Representation: Meeting the Ideal of One Man, One Vote, second ed. Brookings Institution Press, Washington.

Birch, S., 2003. Electoral Systems and Political Transformation in Post-Communist Europe. Palgrave Macmillan, Basingstoke, UK.

Bowler, S., Farrell, D.M., Pettitt, R.T., 2005. Expert opinion on electoral systems: so which electoral system is "best"? Journal of Elections, Public Opinion and Parties 15, 3–19.

Bowler, S., Grofman, B. (Eds.), 2000. Elections in Australia, Ireland, and Malta under the Single Transferable Vote. University of Michigan Press, Ann Arbor.

Carey, J., Hix, S., 2011. The electoral sweet spot: low – magnitude proportional electoral systems. American Journal of Political Science 55, 383–397.

Carstairs, A.M., 1980. A Short History of Electoral Systems in Western Europe. Allen & Unwin, London.

Colomer, J.M. (Ed.), 2004. Handbook of Electoral System Choice. Palgrave Macmillan, Houndmills, UK.

Cox, G.W., 1997. Making Votes Count. Cambridge University Press, Cambridge, UK.

Farrell, D.M., 2001. Electoral Systems: A Comparative Introduction. Palgrave, Houndmills, UK.

Gallagher, M., Mitchell, P. (Eds.), 2008. The Politics of Electoral Systems, Paperback ed. Oxford University Press, Oxford.

Grofman, B., Lijphart, A. (Eds.), 2002. The Evolution of Electoral and Party Systems in the Nordic Countries. Agathon, New York.

Hermens, F.A., 1941. Democracy or Anarchy? A Study of Proportional Representation. University of Notre Dame Press, Notre Dame, Indiana.

Lijphart, A., 1991. Constitutional choices for new democracies. Journal of Democracy 2 (1), 72–81.

Lijphart, A., 1994. Electoral Systems and Party Systems. Oxford University Press, Oxford.

Lijphart, A., 2012. Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries, second ed. Yale University Press, New Haven.

McDonald, M.D., Budge, I., 2005. Elections, Parties, Democracy: Conferring the Median Mandate. Oxford University Press, Oxford.

McGann, A., 2006. The Logic of Democracy: Reconciling Equality, Deliberation, and Minority Protection. University of Michigan Press, Ann Arbor.

McGann, A., 2013. Fairness and bias in electoral systems. In: Nagel, J.H., Smith, R.M. (Eds.), Representation: Elections and Beyond. University of Pennsylvania Press, Philadelphia.

Nagel, J.H., 2000. Expanding the spectrum of *democracies*: reflections on proportional representation in New Zealand. In: Crepaz, M.M.L., Koelble, T.A., Wilsford, D. (Eds.), Democracy and Institutions: The Life Work of Arend Lijphart. University of Michigan Press, Ann Arbor, pp. 113–127.

Nagel, J.H., 2012. Evaluating democracy in New Zealand under MMP. Policy Quarterly 8 (2), 3–11.

O'Leary, B., Grofman, B., Elklit, J., 2005. Divisor methods for sequential portfolio allocation in multi-party executive bodies: evidence from Northern Ireland and Denmark. American Journal of Political Science 49, 198–211.

Persson, T., Tabellini, G., 2004. Constitutions and economic policy. The Journal of Economic Perspectives 18, 75–98.

Powell, G.B., 2000. Elections as Instruments of Democracy: Majoritarian and Proportional Visions. Yale University Press, New Haven.

Reynolds, A., Reilly, B., Ellis, A., 2005. Electoral System Design: The New International IDEA Handbook. International Institute for Democracy and Electoral Assistance, Stockholm.

Rogowski, R., Kayser, M.A., 2002. Majoritarian electoral systems and consumer power: price-level evidence from the OECD countries. American Journal of Political Science 46, 526–539.

Shugart, M.S., Wattenberg, M.P. (Eds.), 2001. Mixed-Member Electoral Systems: The Best of Both Worlds?. Oxford University Press, Oxford.

Taagepera, R., Shugart, M.S., 1989. Seats and Votes. Yale University Press, New Haven.

Relevant Websites

http://www.ifes.org/ – International Foundation for Electoral Systems (IFES), Washington, DC.

http://www.idea.int/ – International Institute for Democracy and Electoral Assistance (IDEA), Stockholm.

Prosecution

Kay Levine, Emory Law School, Atlanta, GA, USA
Malcolm Feeley, University of California, Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prosecutors have been called the 'gatekeepers of criminal justice' because without their actions, judicial intervention and sanctions could not occur. However, a study of prosecution should not be limited to assessments of the prosecutor's role in individual cases. One must instead consider the ways in which countries deal with competing needs and limited resources when they establish and operate their prosecution systems. Because prosecution contributes substantially to the definition and implementation of criminal justice policy, any comprehensive examination of prosecutors' functions must include an analysis of their relationship to the state.

This article considers six of the most salient features of prosecution systems: (1) the degree to which prosecution is handled by public officials or private parties; (2) the degree to which the prosecutor's office is centralized or decentralized; (3) the independence and political accountability of the prosecutor; (4) the prosecutor's relationship to and with the police; (5) the prosecutor's discretion in charging; and (6) the prosecutor's role in court. The content of this article will be limited to the role of prosecution in democratic countries, but it ends with commentary about prosecution in a globalized criminal justice system that must deal with terrorism and war crimes in nations across the globe.

In its broadest sense, 'prosecution' signifies an aggrieved party's initiation and pursuit of a legal case through the court system. It begins with the filing of a complaint and continues through trial, awarding of remedies, and appeal. While this term technically applies to both civil and criminal matters, it is most commonly used in the criminal law context. In a criminal case, the term 'prosecution' refers to pursuing a criminal complaint against someone accused of a crime, usually by an official called a prosecutor, but occasionally by the complaining victim directly. This article focuses on the most common form of prosecution, a case brought by a public official against one charged with a crime. The authors draw on comparative studies to illuminate differences and similarities among prosecution systems throughout the world.

Prosecutors have been called the 'gatekeepers of criminal justice' because without their actions, judicial intervention and sanctions could not occur (Di Federico, 1998). However, a study of prosecution should not be limited to assessments of the prosecutor's role in individual cases. One must instead consider the ways in which countries deal with competing needs and limited resources when they establish and operate their prosecution systems. Because prosecution contributes substantially to the definition and implementation of criminal policy, any comprehensive examination of prosecutors' functions must include an analysis of their relationship to the state.

This article considers six of the most salient features of prosecution systems: (1) the degree to which prosecution is handled by public officials or private parties; (2) the degree to which the prosecutor's office is centralized or decentralized; (3) the independence and political accountability of the prosecutor; (4) the prosecutor's relationship to and with the police; (5) the prosecutor's discretion in charging; and (6) the prosecutor's role in court. The content of this article will be limited to the role of prosecution in democratic countries; for a discussion of criminal courts in nondemocratic nations (*see* Trials; Nazi Law; Socialist Law; Political Trials; War Crimes Tribunals, Crimes against Humanity and Transitional Justice; Law and Courts in Authoritarian States).

Public or Private Prosecution

In many legal systems, criminal law was once a branch of what scholars now call 'torts' (*see* Torts in Law), whose prosecution was initiated and pursued by the injured parties. With the rise of the modern state, crime became a distinct offense against the state itself, and responsibility for prosecution was assumed by the state (Langbein, 1974). This trend emerged earlier and more sharply in Europe than in England, where the system of privately initiated and financed prosecution remained in place well into the nineteenth century.

Prosecutors as Public Officials

In Europe and indeed in most places in the world, public prosecutors play a much larger role in the criminal process than they do in the United States and England. Continental prosecutors control the investigation of crime, often having direct supervisory powers over the police, and typically oversee the execution and disposition of almost all criminal arrests. This tradition derives largely from the institution of the *Ministère Publique* from the Napoleonic Code. Furthermore, some civil law countries combine the role of prosecutor and judge, allowing officials to move back and forth between the two positions at different points in their careers and sometimes within the same case. A number of non-Western countries that have based their criminal justice systems on German or French models (such as Japan and Korea) also permit role flexibility to varying degrees (*see also* Trials; Judges; Criminal Procedure: Adversarial and Inquisitorial Legal Systems). Recent developments in France and Italy, however, suggest the use of more clearly defined boundaries between criminal justice system actors.

Common law countries reject this expansive conception of the prosecutor's office and formally prohibit any degree of fluidity in roles, regarding such interchangeability as in conflict with due process and separation of powers principles (*see also* Due Process in Law). In common law countries such as the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86099-7

United States, prosecutors unilaterally (with very few exceptions) make filing and case-processing decisions. The investigative work is typically left to the police, and judging is the domain of elected or appointed members of the judiciary. (Of course, in actual practice, the discretion with which each of these actors is vested allows some flexibility in task performance; *see also* Plea Bargaining: A Comparative Analysis and Policing for examples. Moreover, state trial court judges often begin their careers as state prosecutors.) In England and Wales, the Crown Prosecution Service (CPS), established in 1985, comprises exclusively professional prosecutors under the authority of a national Director of Public Prosecutions, but like their American counterparts, British prosecutors have no real supervisory authority over the local police.

Elements of Private Prosecution

Since the mid-nineteenth century, public prosecutors everywhere have become more powerful and independent. However, since the 1970s, with the growth of the victims' rights movement in many countries, legislatures have begun to provide victims with a greater voice in criminal proceedings. The degree of victim empowerment varies tremendously among countries. For example, Belgian law recognizes that certain crimes (e.g., defamation) require a formal complaint by the victim in order for the public prosecutor to initiate a criminal case. Even in that event, however, the public prosecutor is not obligated to file charges. In Germany, a victim may act as a private prosecutor if the public prosecutor does not file charges, or may join the public prosecution as a joint plaintiff and support the prosecutor's case. A French crime victim who feels ignored by the authorities may file a civil case based on the criminal behavior. This action automatically initiates a criminal case, thereby forcing an otherwise reluctant public prosecutor to become involved.

American crime victims generally have no formal right to participate directly in criminal cases. Nonetheless, victim willingness to give evidence may greatly influence the prosecutor's decision to proceed with a case. Cooperative victims are not a prerequisite to a successful prosecution, but prosecutors often make their decisions based on victim proclivities (Vera, 1981). While a lay person cannot control the criminal case, a victim of crime in the United States can file a civil lawsuit against the defendant based on the same conduct at issue in the criminal case; the remedies available through the lawsuit are limited to civil damages, however. Interestingly, the role of private interests in prosecution has made a comeback in the United States, as scholars have recently debated the legitimacy of allowing private victims to finance investigations or prosecutions that are conducted by public officials (Kennedy, 1997).

Centralized or Decentralized Office

Scholars distinguish between highly centralized systems that place prosecutors in a national civil service hierarchy controlled by a ministry of justice, and decentralized systems where prosecutors are selected locally. Overall, these structural and institutional features are likely to be closely related to and nested in larger structures of government and to make significant differences (Damaska, 1981). Needless to say, variations exist between these extremes, and even within each category, according to hiring practices and filing controls.

At one extreme is the American experience. The vast majority of criminal cases – both serious and minor – in the United States are filed and processed at the state level. However, 'state control' is something of a misnomer, because most American prosecutors are county officials who enforce state law. The chief prosecutor of each county (usually called the District Attorney) is locally elected, while the deputies or assistants may be either career civil servants or employees of a specific administration. State attorneys general have little control over their behavior or filing practices, handling mostly training and appeals (Flemming et al., 1992). Each county operates its own hiring system and typically recruits graduates from local law schools. There is little advance training for the job; on-the-job training occurs locally, within the office or in short courses run by the state or statewide prosecutors' association. As a result, each office sets its priorities according to the demands of its electorate and according to the interests of its leaders; this can produce great diversity in policies among prosecutors' offices, even within the same state (Utz, 1978; Di Federico, 2005; Levine and Wright, 2012; but see Levine, 2006 for an example of statewide convergence around a set of norms).

At the other end of the continuum lie countries like France and Japan, where the procuracy is tightly controlled and hierarchical. For example, at the apex of the Japanese system sits the Supreme Public Prosecutors Office, followed by 8 High Public Prosecutors Offices and then 50 District Public Prosecutors Offices, which in turn are overseen by the Supreme Court. At the base are 452 Local Public Prosecutors Offices. The local offices prosecute relatively minor cases, while the district offices handle all serious crimes. Moreover, Japanese prosecutors are an elite; they graduated from top law schools, passed a rigorous bar exam, and completed 2 years of training at the Supreme Court's Legal Training Institute. Because they are overseen by the Supreme Court, Japanese prosecutors have close association with judges; indeed, prosecutors and judges can and often do move between positions. Due to this hierarchical oversight, Japanese prosecutors value consistency and follow well-understood and consistent patterns of decision making (Johnson, 1998, 2002). A similar set of practices holds for France as well (Hodgson, 2005).

A high degree of centralization tends to limit the impact of public opinion on individual prosecutors, and on the national prosecutors' office as a whole. Prosecutors in more decentralized systems – and the United States is the extreme – have vastly more discretion but are (at least in theory) subject to shifts in public opinion owing to the election of the local district attorney (Wright, 2009). These differences are not ironclad; some modern prosecutors in Italy, Spain, France, Portugal, and elsewhere are sensitive to public opinion, and at times actively court the media in order to generate public interest and support for their cases (Di Federico, 2005). But even so, without an electoral process, they have less need than their American counterparts to be attuned to public opinion.

Lastly, prosecutors in civil law countries that subscribe to the centralized, professionalized model tend to view their positions as careers; in contrast, many American prosecutors work for a few years and then leave government for private practice. These patterns have their pros and cons. Frequent

turnover can hinder development of a sense of confidence, professionalism, and independence, but the hierarchical service system has been criticized for producing insular prosecutors – individuals and local offices who pursue their own agendas, often reflective of personal idiosyncrasies, or political or social biases (Di Federico, 1998). Political oversight may be the answer. Yet where political loyalties run deep and are reinforced by family and clan association, political oversight would invariably result in political agenda setting within prosecutors' offices and undermine the independence (Nelken, 2011). Such politicization is not hypothetical; it is pervasive in many of the new democracies in Latin America and Eastern Europe, where party, class, and group loyalties undermine the civil service and rule-of-law norms (Brinks, 2008).

Independence and Accountability to the Public

Because prosecutors exercise vast powers, one of the greatest challenges for institutional design and one of the central concerns of both normative and empirical scholarship on prosecutors is the degree to which they are both independent and accountable (Barkow, 2009). How does one prevent selective prosecution of one's political opponents and make sure that prosecutors do not run roughshod over others' rights? Prosecutors must be both independent and accountable, but from whom and to whom? And to what extent? Complete independence suggests no boundaries, the freedom to pick and choose as one likes. Substantial accountability may mean responsible oversight, but it can mean that one binds officials to external criteria of questionable merit.

One solution to this dilemma is to adopt a policy of full prosecution; every offense known to officials for which there is probable cause must be prosecuted to the full extent of the law (Davis, 1969). This is the doctrine found in many criminal codes, both in civil law and common law countries, and a standard response prosecutors give to those inquiring about how they exercise the powers. However, even the most ardent of civil law systems recognize that 'full prosecution' does not mean full; rather, it means that there is a systematic and well-developed policy with respect to bringing or declining to bring charges that is followed consistently (Nelken, 2011).

Still other institutional devices have been adopted to assure prosecutorial independence and accountability. Most common throughout the democratic world is the establishment of a centralized and professionalized prosecutorial system, complete with civil service protections, hierarchical control within the office, professional advancement, and oversight by ministries of justice. This may produce a strong prosecutors' office, as in several Northern European countries and Japan, instilling in prosecutors a high sense of duty and obligation to professional norms. In countries with traditions of a strong civil service, this model works well (Barkow, 2009; Johnson, 2002; Langbein and Weinreb, 1977).

Other countries face considerable challenges attempting to enforce this degree of control over prosecutorial decision making. Historically, the United States has had a weak tradition of civil service and its people have been leery of those exercising governmental power. Accordingly, in the early nineteenth century, the states adopted the practice of electing both judges and chief prosecutors at the local (county) level. With the experience of colonial judges and prosecutors appointed in London still in mind, the objective was not to 'politicize' these offices, but to try to assure that those who exercised these powers did not do the bidding of distant political authorities. It was a way of assuring both independence and accountability. Of course, the situation has changed considerably. Since the early 1900s if not before, research has consistently revealed that American prosecutors, both elected (state) and appointed (federal) are highly subject to political whim and consistently support 'get tough' policies (Wright, 2009). Furthermore, the office of chief prosecutor at both the local and the federal level is often a springboard for seeking electoral office, and given the near continuous 'war on crime,' this almost always means that prosecutors with an eye toward future electoral office are likely to support expanded criminalization, harsh sentencing laws, and vigorous prosecution. This connection exacerbates the politicization of crime, and by itself goes a long way to explaining why the United States is distinctive in the harshness of its justice (Stuntz, 2012; Davis, 2007; Flemming et al., 1992). In contrast, in Germany, Japan, the Scandinavian countries, and elsewhere, crime policy is viewed as a technical issue usually best left in the hands of the experts, which include prosecutors (Johnson, 2002; Weigand, 1999).

Relationship to and with Police

Is a prosecutor a justice official or a member of law enforcement? While Langbein (1974) reports that English justices of the peace served as both investigators and prosecutors in the fourteenth and fifteenth centuries, most countries today have formally separated the roles of police and prosecutor. Nonetheless, prosecutors vary in the degree to which they supervise or participate in police investigations. A prosecutor may bring a high level of control to police activities by formally supervising a police investigation (before or after bringing charges), by interviewing witnesses, or by attending the execution of search warrants. In contrast, a prosecutor may manifest low levels of control by simply filing charges in response to documents and evidence collected and submitted by investigating police officers.

In code-based countries with strongly centralized prosecution agencies, greater participation and supervision has long been the norm, at least in theory. Police in those nations may be statutorily required to notify prosecutors of an ongoing investigation or the detention of suspects, and the prosecutor can directly interrogate suspects or collect evidence. Systematic empirical literature on prosecutor–police relations is sparse, despite the lively debate over the degree to which prosecutorial oversight and investigative authority may serve as a useful check on police power. Some suggest it has substantial impact (Johnson, 2002; Frase, 1990, 1999), while others maintain that it is largely myth (Hodgson, 2005). Stefan Braum, for example, has argued that, although the European Court of Human Rights considers excessive police power to pose a risk to security and human rights, many European countries these days have public prosecution services that exert only weak control over the police. That is, with the integration of Europe into the

European Union and with the rise of terrorist activity in the twenty-first century, the international law enforcement agency Europol has grown in strength, subject to neither prosecutorial nor judicial control of any meaningful sort (Braum, 2012).

Low levels of prosecutorial control over police have long characterized common law countries like the United States; police typically handle the investigation and arrest phases of a criminal matter and notify the prosecutor only after arrest. While federal level prosecutors often become far more involved in the investigatory phases, coordinating wiretaps, search warrants, and witness interviews, most state prosecutors do not interrogate suspects or assist in the collection of evidence (Leo, 1996). Hence, police often control the flow of information on which prosecutors must rely in order to decide whether and what crimes to charge. This has led some commentators to conclude that prosecutors do little more than ratify police decisions (Feeley and Lazerson, 1983), rendering the arrest itself far more significant than the prosecutor's filing decision. The same tendency has been observed in Canada (Grosman, 1969).

The division of labor between police and prosecutors is not always clear-cut or unproblematic, as police behavior can influence both the onset and eventual disposition of a criminal case (Skolnick, 1975). These tensions are manifest in the British experience. The police were Britain's public prosecutors from 1829 to 1986, due to the absence of a suitable alternative; during this period, they developed a vested interest in controlling all stages of criminal cases – from investigation to arrest to prosecution (Fionda, 1995). The CPS was created to separate police and prosecutorial functions in order to remedy a perceived lack of police objectivity in prosecutorial duties, such as the filing and disposing of cases. However, in the years since the creation of the CPS, police officers have fought for control over filing decisions and have been accused of manipulating files and witnesses to influence prosecutors' options (Nduka-Eze, 1995). (*See* Crime Policy in the United States.)

Prosecutorial Discretion at Charging

Perhaps the central scholarly question about prosecutors is how they exercise their vast discretion to bring or decline a case, and if they opt for the former, what crime(s) to charge. Over the past 100 years in the United States and England, an extensive literature has addressed the issue of prosecutorial discretion and its implications for the just administration of the criminal law. Until the 1980s, scholarship on prosecutors in most civil law countries maintained that prosecutors exercise no or very little discretion since they operate under a system of full and mandatory prosecution; this claim at times has engendered a lively debate between those who maintained that even under compulsory prosecution schemes there is considerable discretion and plea bargaining (Goldstein and Marcus, 1977a,b) and those who believed there is not (Langbein and Weinreb, 1977). Whatever once may have been the case, scholars generally concede that at least since the 1980s, prosecutorial discretion in bringing charges and in negotiating charges and sentence recommendations has expanded tremendously (Luna and Wade, 2012).

Several factors account for this enlargement of the prosecutor's power in nations across the globe. First, more governments have sought to expedite court proceedings by relying more heavily on oral presentations and less on dossiers. Because prosecutors control the oral presentation of evidence and argument, the power over courtroom proceedings has shifted markedly from judges to prosecutors in jurisdictions following this trend (Caianiello, 2012; Kyprianou, 2009).

More importantly, however, we have seen the emergence of plea bargaining in countries that previously prohibited it (at least as an official matter) (Turner, 2009; Bradley, 2012) (*see also* Legal Issues: Public Opinion). While the reasons for adoption and forms of the practice vary, it seems to have taken hold across national borders. For example, some countries (like Italy) have introduced plea bargaining in order to secure finality of process; others, in their efforts to clamp down on drug conspiracies and organized crime, have embraced the time-honored Anglo-American practice of trading leniency for information implicating more important actors in criminal organizations (known as the little fish–big fish phenomenon). Still others have statutorily granted discounts to those who plead guilty (Caianiello, 2012). One can also find examples of countries, such as Russia and Bulgaria, adopting plea bargaining but imposing limits on how that practice is conducted, such as by prohibiting plea bargaining in serious violent crimes and prohibiting charge bargaining (Turner, 2009: p. 140). Notably, exceptions to these global trends can be found in Asia – while India began the practice in 2006 and Taiwan did so in 2004, China and Japan have so far refused to pursue plea bargaining as a formal policy (Turner, 2009: p. 171).

Political developments since the 1980s have expanded prosecutor discretion and hence power in plea negotiations. The near worldwide 'war on crime,' which began in the early 1980s, produced a 'culture of control' in many countries (Garland, 2001), a culture that is characterized at least in part by substantially harsher legislative sentencing ranges and public commitment to more vigorous prosecution. Harsher mandatory minimum sentences, three strikes or career criminal laws, sentence guidelines, and other legislative efforts to 'structure' sentences have resulted in more discretionary power for prosecutors, because they can circumvent legislatively 'mandated' draconian outcomes in exchange for pleas of guilty to offenses carrying lesser but still harsh penalties. Such developments have occurred worldwide, but the effects are most evident in the United States (Garland, 2001).

At the same time, legislatively created 'alternatives' to prosecution and supervised probation have expanded, and prosecutors have seized on these programs to gain still more discretionary power. Against the threat of substantial criminal charges, prosecutors can successfully press those charged with lesser offenses to agree to a variety of conditions for noncustodial supervision and surveillance, perhaps even reduced due process rights, in exchange for the government's agreement to drop or suspend prosecution (Wade, 2006). Moreover, those on probation who are arrested on new charges (often obtained because of probationers' diminished due process rights) are frequently offered a deal: waive the right to a probation revocation hearing and the new charges will be dropped. If accepted, the deal allows law enforcement to send a person back to prison

almost immediately and with only a peremptory appearance in court (Greenspan, 2014).

In hierarchically organized prosecution systems, one is likely to find national prosecutorial policies governing these exercises of discretion, but in decentralized systems, policies emanate from supervisors in individual offices, leading to troubling disparities across jurisdictions that are all subject to the same formal law. For example, in the United States, many states have statutes authorizing the death penalty for certain forms of murder, but the rate of filing capital charges for similar crimes varies tremendously among the individual counties (American Bar Association, 2006). Even with this variation in filing practice, prosecutors in all jurisdictions regularly rely on plea bargaining to secure guilty pleas in exchange for not filing capital charges (Thaxton, 2013). This is not just a capital murder phenomenon; what holds for these occasional high-stakes, high-visibility cases also holds for much more frequent 'ordinary' cases.

The power of the prosecutor is most evident in the declining rate of criminal trials. In the United States and England, for example, the rate of criminal trials, never high since the late nineteenth century, has plummeted to nearly zero (Galanter, 2004; Mulcahy, 2013; Wright, 2005). Indeed, since the 1990s, commentators have begun to describe the American criminal justice system as an administrative system driven by prosecutors rather than presided over by judges (Lynch, 1998), leading for calls to reform what they see as a dysfunctional and broken system of justice (Stuntz, 2012; Barkow, 2008). What holds most clearly in the United States is found elsewhere as well (Guarneri and Pederzolo, 2001).

Roles in Court

The final point of comparison concerns the prosecutor's role during the trial of a criminal case. The prosecutor's role will largely be determined by the type of trial system in which he or she works (adversarial, inquisitorial, or some mixture thereof; *see* Trials; Criminal Procedure: Adversarial and Inquisitorial Legal Systems). The prosecutor may be an active adversarial participant striving to win a conviction, or a more passive, impartial actor charged with ensuring that justice results.

When the trial depends largely on the contents of a written dossier (in countries such as Japan, Argentina, France, and Germany), the prosecutor is generally more passive; he or she submits the dossier containing the investigative reports to the judge (or to the panel of judges) before trial, but generally relinquishes control over questioning of witnesses. In many of these systems, the prosecutor is obligated to ask the judge for an acquittal following trial if the evidence merits this result, and the prosecutor typically has the right to appeal both the outcome of the case and the sentence.

In systems that require oral testimony (including Italy, Great Britain, Israel, and the United States), the prosecutor is the principal player in the courtroom. The prosecutor may sit near the fact finders or judges to establish rapport; controls the majority of the witnesses and the presentation of evidence; and argues the case both first and last, providing both the initial opening statement and the final closing argument. For this reason, prosecutors in oral-trial systems may feel more personally invested in the outcome of a case and often place tremendous pressure on themselves to win convictions (Heilbroner, 1990; Baker, 1999). In some countries, they have no right to appeal acquittals.

In the United States, where prosecutors are more autonomous than perhaps anywhere else in the world, some prosecutors have experimented with new roles, which Levine has labeled 'the new prosecution' and 'problem-solving prosecution' (Levine, 2005, 2007). Prosecutors have always exercised considerable discretion to drop and reduce charges, but she has found that some American prosecutors have experimented with community-based efforts to 'solve' problems that, while technically criminal, are better handled (they believe) with a social services approach – albeit with the threat of full prosecution in the background. Many such prosecutors have joined with judges and defense attorneys to develop problem-solving courts for arrestees with substance abuse problems, mental health problems, or domestic problems. Or on a more ad hoc basis, they may use the threat of prosecution to extract agreement from a father to provide financial support for his child or to seek treatment for alcoholism (Levine, 2005).

Conclusion: New Issues in the Twenty-First Century

Up until this point, the article has focused exclusively on prosecution policies and practices in domestic fora. But in the twenty-first century, as crimes against humanity increasingly seem to transcend national borders, prosecutors everywhere have had to deal with a globalized criminal justice system. This challenge is manifest in the International Criminal Court, with prosecutions of genocide and other war crimes, as well as in prosecutions of terrorist organizations in domestic and military tribunals (*see* War Crimes Tribunals, Crimes against Humanity and Transitional Justice; Law and Terrorism in the US and Beyond; Human Rights: Legal Aspects; Torture: Legal Perspectives; International Law and Treaties). Some commentators are even exploring the possible creation of a European Public Prosecutor's Office to handle transborder criminal disputes (Ligeti, 2013).

Even more so than their domestic law counterparts, prosecutors in globalized criminal justice systems have faced questions about the quality of evidence necessary to mount a successful prosecution, which investigative techniques are acceptable means to extract information, and the advisability of plea bargaining in these extremely serious crimes (particularly whether to give immunity to lower level players in order to obtain their testimony against the leaders of these organizations). Moreover, prosecutors handling terrorism and war crimes cases must keep in mind a range of goals. Beyond conviction of the guilty, identification of the innocent, deterrence of future wrongdoing, and provision of voice and closure to victims, globalized prosecutions of terrorists and war criminals can facilitate long-term peace in a troubled region and shed light on horrific events that might otherwise remain secret. Prosecutors cannot unilaterally achieve these aims, but their support for certain approaches (or, likewise, their resistance to certain approaches) can give preference to one set of goals over another. In essence, as Luna and Wade (2012) have observed, "(Like) their domestic colleagues, international prosecutors must become thoughtful policymakers and case managers in light of the lofty, if not unachievable, goals placed on imperfect systems of justice" (Luna and Wade, 2012: p. 287).

See also: Common Law; Courts and Adjudication; Criminal Justice, Sociology of; Criminal Procedure: Adversarial and Inquisitorial Legal Systems; Death Penalty; Empirical Legal Studies; Law: Economics of Its Public Enforcement; Plea Bargaining: A Comparative Analysis; Police: A Sociology of Knowledge Approach; Policing; Punishment: Social and Legal Aspects; Trials.

Bibliography

Alschuler, A.W., 1968. The prosecutor's role in plea bargaining. University of Chicago Law Review 36, 50–112.

American Bar Association, 2006. Evaluating Fairness and Accuracy in State Death Penalty Systems: The Georgia Death Penalty Assessment Report. American Bar Association, Washington, DC.

Baker, M., 1999. D.A.: Prosecutors in Their Own Words. Simon and Schuster, New York.

Barkow, R., 2008. The ascent of the administrative state and the demise of mercy. Harvard Law Review 121, 1332–1381.

Barkow, R., 2009. Institutional design and the policing of prosecutors: lessons from administrative law. Stanford Law Review 61, 869–915.

Barkow, R., 2011. Federalism and criminal law: what the feds can learn from the states. Michigan Law Review 109, 519–561.

Bradley, C., 2012. The Prosecutor's role: plea bargaining and evidentiary exclusion. In: Luna, E., Wade, M. (Eds.), The Prosecutor in Transnational Perspective. Oxford University Press, Oxford, pp. 91–101.

Braum, S., 2012. Prosecutorial control of investigations in Europe: a call for judicial oversight. In: Luna, E., Wade, M. (Eds.), The Prosecutor in Transnational Perspective. Oxford University Press, Oxford, pp. 67–81.

Brinks, D., 2008. The Judicial Response to Police Killings in Latin America: Inequality and the Rule of Law. Cambridge University Press, New York.

Caianiello, M., 2012. The Italian public prosecutor: an inquisitorial figure in an adversarial proceeding? In: Luna, E., Wade, M. (Eds.), The Prosecutor in Transnational Perspective. Oxford University Press, Oxford.

Corso, P., 1993. Italy. In: Van Den Wyngaert, C. (Ed.), Criminal Procedure Systems in the European Community. Butterworths, London, pp. 223–260.

Damaska, M., 1981. The reality of prosecutorial discretion: comments on a German monograph. American Journal of Comparative Law 29, 119.

Davis, A., 2007. Arbitrary Justice: The Power of the American Prosecutor. Oxford University Press, New York.

Davis, K.C., 1969. Discretionary Justice. Louisiana State University Press, Baton Rouge, LA.

Di Federico, G., 1998. Prosecutorial independence and the democratic requirement of accountability in Italy. British Journal of Criminology 38, 371–387.

Di Federico, G. (Ed.), 2005. Recruitment, Professional Evaluation and Career of Judges in Europe: Austria, France, Germany, Italy, The Netherlands and Spain. Lo carabeo, Bologna.

Feeley, M.M., 1979. The Process is the Punishment. Russell Sage, New York.

Feeley, M.M., Lazerson, M.H., 1983. Police-prosecutor relationships: an inter-organizational perspective. In: Boyum, K., Mather, L. (Eds.), Empirical Theories about Courts. Longman Press, New York, pp. 216–243.

Felstiner, W.L., 1979. Plea contracts in West Germany. Law and Society Review 13, 309–326.

Fionda, J., 1995. Public Prosecutors and Discretion: A Comparative Study. Clarendon Press, Oxford.

Flemming, R.B., Nardulli, P.F., Eisenstein, J., 1992. The Craft of Justice: Politics and Work in Court Communities. University of Pennsylvania Press, Philadelphia.

Frase, R.S., 1990. Comparative criminal justice as a guide to American law reform: how do the French do it, how can we find out, and why should we care? California Law Review 8, 539–683.

Frase, R.S., 1999. France. In: Bradley, C.M. (Ed.), Criminal Procedure: A Worldwide Study. Carolina Academic Press, Durham, NC, pp. 143–184.

Galanter, M., 2004. The vanishing trial. Journal of Empirical Legal Studies 1, 459–570.

Garland, D., 2001. The Culture of Control. Oxford University Press, New York.

Goldstein, A., Marcus, M., 1977a. The myth of judicial supervision in three 'inquisitorial' systems: France, Italy, and Germany. Yale Law Journal 87, 240–283.

Goldstein, A., Marcus, M., 1977b. Comment on continental criminal procedure. Yale Law Journal 87, 1570–1577.

Greenspan, R., 2014. The Transformation of Criminal Due Process in the Administrative State. Quid Pro LLC, New Orleans (Classic Dissertation Series).

Grosman, B., 1969. The Prosecutor: An Inquiry into the Exercise of Discretion. University of Toronto Press, Toronto.

Guarneri, C., 1996. Prosecution in two civil law countries: France and Italy. In: Nelken, D. (Ed.), Comparing Legal Cultures. Aldershot Press, Brookfield, VT.

Guarneri, C., Pederzolo, P., 2001. The Power of the Judge. Oxford University Press, Oxford.

Harnon, E., Stein, A., 1999. Israel. In: Bradley, C.M. (Ed.), Criminal Procedure: A Worldwide Study. Carolina Academic Press, Durham, NC, pp. 217–244.

Heilbroner, D., 1990. Rough Justice: Days and Nights of a Young D.A. Pantheon Books, New York.

Herrmann, J., 1992. Bargaining justice – a bargain for German criminal justice. University of Pittsburgh Law Review 53, 755–776.

Hodgson, J., 2005. French Criminal Justice. Hart Publishing, Oxford.

Johnson, D., 1998. The organization of prosecution and the possibility of order. Law and Society Review 32, 247–308.

Johnson, D., 2002. The Japanese Way of Justice. Oxford University Press, New York.

Kennedy, J., 1997. Private financing of criminal prosecutions and the differing protections of liberty and equality in the criminal justice system. Hastings Constitutional Law Quarterly 24, 665.

Kritzer, B., 2004. Disappearing trials? A Comparative Perspective Journal of Empirical Legal Studies 1, 735–754.

Kyprianou, D., 2009. Adversarial vs inquisitorial prosecution systems in Europe: commonalities and diversities. In: Jaishankar, K. (Ed.), International Perspectives on Crime and Justice. Cambridge Scholars Press, Newcastle.

Langbein, J.H., 1974. Prosecuting Crime in the Renaissance: England, Germany, France. Harvard University Press, Cambridge, MA.

Langbein, J.H., Weinreb, L.L., 1977. Continental criminal procedure: 'myth' and reality. Yale Law Journal 87, 1549–1569.

Leo, R., 1996. Inside the interrogation room. Journal of Criminal Law and Criminology 86 (2), 266–303.

Levine, K., 2005. The new prosecution. Wake Forest Law Review 4, 1135–1214.

Levine, K., 2006. The intimacy discount: prosecutorial discretion, privacy and equality in the statutory rape caseload. Emory Law Journal 55, 691.

Levine, K., 2007. Can prosecutors be social workers? Studies in Law, Politics, and Society 40, 127–154.

Levine, K., Wright, R., 2012. Prosecution in 3-D. The Journal of Criminal Law and Criminology 102, 1119–1180.

Ligeti, K. (Ed.), 2013. Toward a Prosecutor for the European Union, vol. 1. Hart Publishing, Oxford.

Luna, E., Wade, M., 2012. Prosecution in exceptional contexts and nondomestic fora. In: Luna, E., Wade, M. (Eds.), The Prosecutor in Transnational Perspective, pp. 279–291.

Lynch, G., 1998. Our administrative system of criminal justice. *Fordham* Law Review 66, 2117–2146.

Montana, R., 2009. Paradigms on judicial supervision and co-ordination between police and prosecutors: the Italian case in a comparative perspective. European Journal of Crime, Criminal Law and Criminal Justice 20, 99–120.

Mulcahy, L., 2013. The collective interest in private dispute resolution. Oxford Journal of Legal Studies 33, 31–57.

Nduka-Eze, C., 1995. The CPS and independence from the police. New Law Journal, 1843–1844.

Nelken, D., 2011. Can prosecutors be too independent? An Italian case study. In: Daems, T., Snaken, S., Van SylSmit, D. (Eds.), European Penology. Hart Publishing, Oxford.

Skolnick, J., 1975. Justice without Trial: Law Enforcement in Democratic Society, second ed. Wiley, New York.

Stuntz, W., 2012. The Collapse of the American Criminal Justice System. Harvard University Press, Cambridge.

Thaxton, S., 2013. Leveraging death. Journal of Criminal Law and Criminology 103, 475–552.

Turner, J.I., 2009. Plea Bargaining across Borders. Wolters Kluwer, Austin, TX.

Utz, P., 1978. Settling the Facts: Discretion and Negotiation in Criminal Court. Lexington Books, New York.

Vera Institute of Justice, 1981. Felony Arrests: Their Prosecution and Disposition and New York City's Courts. Longman, New York.

Wade, M., 2006. The power to decide – prosecutorial control, diversion and punishment in European criminal justice systems today. In: Jehle, J.M., Wade, M. (Eds.), Coping with Overloaded Criminal Justice: The Rise of Prosecutorial Power across Europe. Springer, Berlin.

Weigand, T., 1999. Germany. In: Bradley, C.M. (Ed.), Criminal Procedure: A Worldwide Study. Carolina Academic Press, Durham, NC, pp. 187–211.

Wright, R., 2005. Trial distortion and the end of innocence. Pennsylvania Law Review 154, 79–156.

Wright, R., 2009. Why prosecutor elections fail us. Ohio State Journal of Criminal Law 6, 691–741.

Prosocial Behavior and Empathy

John F Dovidio and Jillian C Banfield, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Both cognitive empathy (perspective taking) and emotional empathy (affective reactions to another person's need) are critical determinants of adult prosocial behavior. Empathy can elicit altruistically motivated helping, reflecting desire primarily to improve another person's welfare. In general, people experience more empathy and exhibit greater prosocial behavior toward others who are in greater need, are more similar to the potential helper, and are members of the helper's social group. Evoking cognitive or emotional empathy toward a member of another group not only promotes prosocial behavior, but also improves attitudes toward the other group as a whole.

Prosocial Behavior and Empathy (in Adulthood)

Prosocial behavior represents a broad category of acts, such as helping, sharing, comforting, donating, or volunteering, and cooperation, that are intended to benefit others. Research on prosocial behavior has addressed not only the antecedents and consequences of these actions, but also the different motivations that may underlie these behaviors. Prosocial actions can be either altruistically or egoistically motivated (Batson, 2011). Altruistic motivation has the ultimate goal of increasing another person's welfare; egoistic motivation has the primary objective of improving one's own welfare. In addition, prosocial behaviors are shaped by both cognitive (e.g., assessments of costs and rewards) and affective (e.g., arousal and emotion) processes.

Empathy is a response that 'stems from another's emotional state or condition, is congruent with the other's emotional state or condition, and involves at least a minimal degree of differentiation between the self and the other' (Eisenberg and Fabes, 1990: p. 132). It also involves both cognitive and emotional processes. Empathy can be measured using self-reports, nonverbal expressions, or psychophysiological indices. Socialization influences the extent to which people experience empathy for others. Children whose mothers are empathic, warm, good at perspective taking, and comforting are more likely to become highly empathic themselves. Women report higher levels of empathy than men, even when they physiologically show similar levels of arousal.

Empathy also appears to have biological and genetic components. Empathic responses are found in most species and are present among human infants shortly after birth. Brain-imaging studies support the idea that humans are hard-wired to be empathic. The human brain contains mirror neurons, which are cells that become activated when observing others' behaviors in ways as if observers themselves were engaging in that behavior. Although there is some controversy about when and how mirror neurons function, they do represent a structure in the brain that bridges the experiences of others with people's own experience (Zaki and Ochsner, 2012). In addition, empathy is a genetically heritable trait: Identical twins have more similar levels of empathy than fraternal twins or siblings. These findings indicate that empathy is a basic human capacity and may represent a basic mechanism for translating genetic prosocial predispositions, which may have evolutionary benefits, into action.

There is substantial empirical evidence that people are fundamentally empathic and emotionally responsive to the needs of others. People are aroused physiologically and subjectively by the distress of others. This reaction appears even among very young children and occurs across cultures. In addition, people show greater empathy toward others who are closer and more similar to them.

Different kinds of empathy produce different types of responses. Cognitive empathy represents the process by which individuals are able to adopt the perspective of another person, imagining what the other person thinks and feels. When people feel that they can relate to others and adopt their perspective, they show patterns of neurological activity similar to those that occur when they personally experience such an event. Emotional empathy relates to the affective response that stems from recognizing another person's need. It is associated with psychophysiological arousal, such as increased heart-rate. How people experience that arousal emotionally – whether as sadness, personal distress, or empathic concern – depends upon the magnitude of the arousal, the closeness with the person in need, the nature of the situation, and dispositional characteristics of the observer. Some people tend to experience empathic arousal more strongly, some have more difficulty regulating their arousal once it occurs, and some people are more inclined to experience one emotion over other emotions. Although cognitive empathy typically arouses some degree of emotional empathy, people can take the perspective of others without having a significant emotional reaction. In addition, people can experience emotional empathy spontaneously, without consciously adopting the perspective of the other person. Both cognitive and emotional empathy can motivate prosocial behavior.

Research Trends

Research in the area of prosocial behavior and empathy has developed in several stages. The work of the early and mid-1960s typically focused on norms, such as social responsibility and reciprocity, which seemed to govern prosocial behavior. By the end of that decade and into the early 1970s,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24024-5

investigators, stimulated by public outrage at bystander apathy, investigated factors that reduced the likelihood of intervening in crisis and emergency situations. As researchers also explored the reasons why people *do* engage in prosocial activities, they began to consider more fully the role of empathy and developmental influences. In the 1980s, interest in prosocial actions declined somewhat and the questions moved primarily from *when* people help to *why* people help. Researchers frequently attempted to understand fundamental motivational processes, considering how different affective consequences of empathy could produce either egoistic or altruistic motivation.

Research in the 1990s provided a clearer link between prosocial motivations and general personal, social, and intergroup orientations. In the 1990s, the research also expanded its focus from spontaneous helping between individuals to more sustained and collective prosocial actions, such as donating (e.g., blood or money) or volunteering to help others, even when there is no opportunity to meet those who benefit from such actions (Wilson, 2000). Over the past decade, researchers have turned their attention to more microlevel processes, such as the brain structures and functions underlying empathy, and more macrolevel processes, such as the role of prosocial behavior in managing intergroup relations and contributing to social coordination within groups (Penner et al., 2005).

When Do People Help?

Research has identified a range of social and situational factors that influence helping and other prosocial actions (Dovidio et al., 2006). In terms of social factors, people are more likely to engage in prosocial activities for others with whom they are more closely related, are more similar, and share group membership. People are less likely to respond prosocially when others are perceived as more responsible for their plight or otherwise undeserving of assistance. With respect to situational factors, people are more likely to help in situations that are more serious and clear. They are less likely to help when they believe that others are present and will take action, which relieves a bystander from having to assume personal responsibility for intervention.

Both cognitive (e.g., norms and perceived costs and rewards) and affective (e.g., arousal and emotion) processes underlie prosocial behavior. In terms of cognitive influences, whether a person helps depends upon the outcomes of a series of decisions (Latané and Darley, 1970). People must recognize that help is needed, assume personal responsibility for intervening, decide how to help, and then take action. This model has also been applied to nonemergency situations.

Alternatively, a cost–reward analysis of prosocial action assumes an economic view of human behavior. In a potential helping situation, a person analyzes the circumstances, weighs the probable costs and rewards of alternative courses of action, and then arrives at a decision that will result in the best personal outcome. Current research is consistent with this central tenet of the cost–reward approach. In addition to cognitive processes, the arousal and affect associated with experiencing empathy for another person is a critical factor in determining when and explaining why people help others.

Empathy and Prosocial Motivation

Empathy is central to the processes that motivate prosocial action. Cognitive empathy sensitizes people to the needs of others and to the situational forces that may have produced these needs. There is a general positive association between dispositional cognitive empathy and a broad range of prosocial actions (Davis, 1994). In addition, engaging regularly in helping, such as through sustained volunteering or blood donation, can create a self-image of being empathic, which leads people to be even more prosocial in the future. Cognitive empathy in the form of perspective taking not only influences the degree of empathic arousal that people experience but also, in combination with other cognitive processes, shapes the nature of the empathic emotion that is experienced. Cognitive empathy can facilitate self-focused emotions, such as sadness or distress, or other-oriented emotions, such as sympathy or empathic concern.

Although most researchers agree that empathic arousal is an important determinant of prosocial action, there is less agreement about the nature of this emotion and how it actually motivates prosocial behavior. Empathy can produce different emotions depending on the context. In severe emergency situations, bystanders may become upset and distressed; in less critical, less intense problem situations, observers may feel sad, tense, or concerned and sympathetic (Batson, 2011; Dovidio et al., 2006).

How empathic arousal is interpreted, in turn, elicits either egoistic or altruistic motivation. Empathic arousal interpreted as personally directed emotions, such as sadness or personal distress, promotes egoistically motivated prosocial behavior, with the primary goal of relieving this unpleasant state. In contrast, empathic arousal experienced as empathic concern produces altruistic motivation, with the primary goal of improving the other person's welfare.

Empathy and Egoism

Models of egoistic motivation posit that helping is directed by the primary goal of improving one's own welfare; the anticipated benefit to others is secondary. From this perspective, empathy elicits self-oriented emotions (e.g., sadness or distress) that are experienced as unpleasant and motivate actions, such as helping, that are perceived to relieve them. Two of these approaches are the negative-state relief model and the arousal: cost–reward model.

According to the negative-state relief model (Cialdini et al., 1997), feelings of guilt or sadness motivate people to engage in behaviors that will improve their mood. Helping represents one such behavior, because people learn through socialization and experience that helping is self-rewarding. Three fundamental implications of the negative-state relief model have received support (see Dovidio et al., 2006). First, a variety of negative states, including guilt from having personally harmed a person and sadness from simply observing another person's unfortunate situation, can motivate helping. Second, other events besides helping (e.g., receiving praise) may just as effectively make people feel better, and exposure to these events can thus relieve the motivation to help caused by negative states. Third, negative moods motivate helping only if people

believe that their moods can be improved by helping. Negative feelings will *not* promote helping if people are led to believe that these feelings cannot be relieved or if, as with younger children, the self-rewarding properties have not yet developed.

Affective empathy can produce other emotional reactions, such as distress and upset, particularly among bystanders in emergencies. According to the arousal: cost–reward model (Piliavin et al., 1981), empathic arousal is generated by witnessing the distress of another person. When the bystander's empathic arousal is attributed to the other person's distress, it is emotionally experienced by the observer as unpleasant. The observer is therefore motivated to reduce it. The person then weighs various costs and rewards for helping or not helping, and chooses a course of action that relieves the unpleasant arousal while minimizing net costs and maximizing rewards. One normally efficient way of reducing this arousal is by helping to relieve the other's distress. Thus, from this perspective, arousal motivates a bystander to take action, and the cost–reward analysis shapes the direction that this action will take.

Supportive of the arousal: cost–reward model, empathic arousal attributed to the other person's situation motivates helping. Facial, gestural, and vocal indications of empathically induced arousal, as well as self-reports of empathically induced anxiety, are all positively related to helping (see Dovidio et al., 2006). Consistent with the hypothesized importance of attributing this arousal to the other's situation, people are more likely to help when arousal from extraneous sources such as exercise, erotic films, and aggressive films is attributed to the immediate need of another person. People are less likely to help when arousal generated by witnessing another person's distress is associated with a different cause (e.g., misattributed to a pill). In addition, extreme empathic overarousal or the inability to regulate empathic arousal, which may interfere with the attribution process, can reduce helpfulness.

Also consistent with the arousal: cost–reward model, people donate more money when confronted with a single, identifiable victim than a group of victims. It has been suggested that this effect may occur because people experience 'psychic numbing' (i.e., less sympathy) when they are confronted with the needs of a group of people rather than an individual. Recent evidence, though, suggests that when people do not expect to have to help the victims, they report more compassion for a group than for an individual. However, when there are financial or emotional costs to helping, people actively try to reduce their feelings of compassion for the victims (Cameron and Payne, 2011). Thus, when prosocial action becomes too demanding, people 'downregulate' their emotions and experience less emotional empathy.

Empathy and Altruism

In contrast to egoistic models of helping is the empathy–altruism hypothesis (Batson, 2011). Although this framework acknowledges that egoistically motivated helping occurs, it also posits that true altruism also exists. The primary mechanism in the empathy–altruism hypothesis is the emotional reaction to another person's problem. Empathy that is experienced emotionally as compassion and concern (i.e., as 'empathic concern') produces altruistic motivation. Because altruistic motivation has the primary goal of improving the other person's welfare, an altruistically motivated person will help if (1) helping is possible, (2) helping is perceived to be ultimately beneficial to the person in need, and (3) helping personally will provide greater benefit to the person in need than would assistance from another person also able to offer it.

Numerous experiments have produced impressive empirical support for the empathy–altruism hypothesis (Batson, 2011). Participants who experience relatively high levels of empathic concern show high levels of helpfulness even when it is easy to avoid the other person's distress, when they can readily justify not helping, when helping is not apparently instrumental to improving the benefactor's own mood, and when mood-improving events occur prior to the helping opportunity. In addition, because it motivates people primarily to improve the welfare of others, people experiencing greater empathic concern are more likely to withhold direct assistance from another person if helping (e.g., providing the answer to a problem) will interfere with the other person's longer term best interests (e.g., learning how to solve the problem themselves).

However, several researchers have proposed alternative explanations that challenge the contention that helping can be altruistically motivated. These explanations have focused on (1) how feelings of empathic concern may be associated with special costs for not helping or rewards for helping, (2) how feelings of sadness that are aroused along with empathic concern actually are the primary determinants of helping, and (3) how manipulations used to induce empathic concern for another person also create a greater sense of self-other overlap, or 'oneness,' so that helping may also have direct and primary benefits for the helper (see Cialdini et al., 1997). Nevertheless, despite the critiques and controversies about aspects of the empathy–altruism hypothesis, the preponderance of evidence from more than 30 years of experimentation on this question strongly suggests that truly altruistic motivation may exist and that all prosocial behavior is not necessarily egoistically motivated.

Empathy, Prosocial Behavior, and Intergroup Relations

Recently, researchers have focused on understanding prosocial behavior that occurs between members of different groups. In contrast to helping members of one's own group, helping members of other groups may be inhibited by intergroup biases. In general, people feel more positively about members of their own group than members of other groups, and they empathize, cognitively and emotionally, more with them. Also, people are more willing to help members of their own group than members of a different group.

For empathy to occur, people need to feel a sense of attachment to the person in need. People more readily feel a greater sense of attachment to members of their own group than members of other groups. In addition, empathy more strongly predicts helping behavior when the recipient and giver of help belong to the same group than when they belong to different groups. Instead, attraction and liking are better predictors of helping toward members of other groups (Sturmer et al., 2006).

Although the research shows that, because of intergroup bias, people generally spontaneously experience less cognitive and emotional empathy for members of another group than of their own group, and consequently are less likely to engage in prosocial behavior to benefit them, empirical evidence also reveals that *evoking* empathy for outgroup members can decrease intergroup bias. Cognitive empathy (perspective taking) reduces stereotyping of members of another group, reduces blame for their disadvantage, and increases recognition of discrimination faced by members of the group (Todd et al., 2012). Recognizing discrimination is an important step in motivating action to redress injustice. Thus, facilitating recognition of discrimination through perspective taking increases support for policies, like affirmative action, that benefits the members of another group.

Emotional empathy also improves attitudes toward members of another group. When Whites focus on the emotions of Blacks, rather than being objective, as they view episodes of racial discrimination, they experience greater moral outrage and exhibit reduced intergroup bias and increased willingness to take prosocial action for Blacks. In addition, focusing on the emotions of members stigmatized groups as they describe their needs and problems arouses emotional empathy in the form of empathic concern, which in turn improves attitudes toward the entire group (Batson, 2011).

Cognitive and emotional empathy also play critical roles in how intergroup contact increases positive attitudes toward members of other groups. The contact hypothesis, which proposes that intergroup group under specified conditions (e.g., cooperative interaction and common goals) can improve attitudes toward other groups, is one of the most widely researched frameworks for intergroup relations. A *meta-analysis* of hundreds of studies indicated that intergroup contact improved attitudes toward other groups largely because it increased both perspective taking and feelings of emotional empathy toward these groups and their members (Pettigrew and Tropp, 2008).

Helping Behavior and Intergroup Power Relations

Whereas empathy promotes prosocial actions to benefit recipients, helping behavior motivated by other factors can be manipulative, improving the benefactor's status and decreasing the longer term welfare of the recipient. According to the intergroup helping as status relations model (Nadler, 2002), members of higher status groups are motivated to maintain their group's status, whereas members of lower status groups seek to improve their group's status. These motivations affect willingness to give and receive help. Specifically, members of high-status groups may help lower status groups, in part, to maintain the status hierarchy. They are likely to offer assistance, particularly when they feel that their group's superior status is insecure, that increases the dependency of the other group on their group (e.g., providing complete answers for task) but not empowering forms of help (e.g., offering a hint that will help others learn the answer on their own).

Members of low-status groups, who strive to enhance the position of their group, tend not to seek dependency-oriented help from the high-status group and react negatively when such unsolicited assistance is offered. Members of the lower status group may resist such help because they do not wish to reinforce their group's inferiority and dependence on the higher status group. Thus, whereas empathy typically motivates people to think more positively about, and act prosocially to benefit, members of other groups, helping exchanges in the absence of empathy can, in some cases, operate to decrease the welfare of the recipient's group while enhancing the power and superior position of the helper's group. Helping behavior is not always synonymous with prosocial behavior.

Conclusion

Empathy is a basic human capacity that is one of the most important motivators of prosocial behavior, which is the behavior that intended to benefit others. Cognitive empathy involves adopting the perspective of another person, whereas emotional empathy relates to the affective response that occurs when people perceive another person's need. Both forms of empathy are prevalent across all human cultures. Individual differences in empathy are determined by genetic, biological, environmental, and socialization processes.

The social context, the nature of the situation, the characteristics of the person in need, and the personality of the potential helper not only affect assessments of costs and rewards and the decisions about whether to engage in prosocial acts, but also shape empathic responses. People display higher levels of prosocial behavior toward others who are similar to them, are members of their group, or who are in greater need, and they do so because they experience higher levels of empathy under these conditions. In addition, how empathy is experienced emotionally is fundamental; empathy can produce either egoistic motivation, intended to improve one's own or their group's welfare, or altruistic motivation, with the goal of improving others' welfare. Experiencing cognitive and emotional empathy improves intergroup relations, but helping motivated by factors other than empathy can reinforce power disparities and intergroup bias. Thus, cognitive and emotional empathy plays critical roles in creating and maintaining positive, productive, and mutually beneficial interpersonal and intergroup relations.

See also: Desegregation; Emotion, Evolution of; Emotions and Intergroup Relations; Empathy During Early Childhood Across Cultures, Development of; Helping Intergroup Relations; Intergroup Relations; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Prosocial Behavior, Cultural Differences in; Prosocial Behavior, Effects of Parenting and Family Structure on; Social Dilemmas, Psychology of.

Bibliography

Batson, C. Daniel, 2011. Altruism in Humans. Oxford University Press, New York.

Cameron, C. Daryl, Payne, B. Keith, 2011. Escaping affect: how motivated emotion regulation creates insensitivity to mass suffering. Journal of Personality and Social Psychology 100, 1–15.

Cialdini, Robert B., Brown, Stephanie L., Lewis, Brian P., Luce, Carol, Neuberg, Steven, L., 1997. Reinterpreting the empathy–altruism relationship: when one into one equals oneness. Journal of Personality and Social Psychology 73, 481–494.

Davis, Mark H., 1994. Empathy: A Social Psychological Approach. Brown and Benchmark, Madison, WI.

Dovidio, John F., Piliavin, Jane A., Schroeder, David A., Penner, Louis A., 2006. The Social Psychology of Prosocial Behavior. Lawrence Erlbaum, Mahwah, NJ.

Eisenberg, Nancy, Fabes, Richard A., 1990. Empathy: conceptualization, measurement, and relation to prosocial behavior. Motivation and Emotion 14, 131–149.

Latané, Bibb, Darley, John M., 1970. The Unresponsive Bystander: Why Doesn't He Help? Appleton-Century-Crofts, New York.

Nadler, Arie, 2002. Inter-group helping relations as power relations: maintaining or challenging social dominance between groups through helping. Journal of Social Issues 58, 487–502.

Penner, Louis A., Dovidio, John F., Piliavin, Jane A., Schroeder, David A., 2005. Prosocial behavior: multilevel perspectives. Annual Review of Psychology 56, 365–392.

Pettigrew, Thomas F., Tropp, Linda R., 2008. How does intergroup contact reduce prejudice? Meta-analytic test of three mediators. European Journal of Social Psychology 38, 922–934.

Piliavin, Jane A., Dovidio, John F., Gaertner, Samuel L., Clark, Russell D. III, 1981. Emergency Intervention. Academic Press, New York.

Sturmer, Stefan, Snyder, Mark, Kropp, Alexandra, Siem, Birte, 2006. Empathy-motivated helping: the moderating role of group membership. Personality and Social Psychology Bulletin 32, 943–956.

Todd, Andrew R., Bodenhausen, Galen V., Galinsky, Adam D., 2012. Perspective taking combats the denial of intergroup discrimination. Journal of Experimental Social Psychology 48, 738–745.

Wilson, John, 2000. Volunteering. Annual Review of Sociology 26, 215–240.

Zaki, Jamil, Ochsner, Kevin N., 2012. The neuroscience of empathy: progress, pitfalls, promise. Nature Neuroscience 15, 675–680.

Prosocial Behavior during Adolescence

Aerika S Brittian and Marisha L Humphries, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prosocial behaviors are an aspect of adolescents' positive development that has gained greater attention in the developmental literature since the 1990s. In this article, the authors review the literature pertaining to prosocial behaviors during adolescence. The authors begin by defining prosocial behaviors as prior theory and empirical studies have done. They describe antecedents to adolescents' prosocial behaviors with a focus on two primary factors: *socialization* and *cultural orientations*. Accordingly, the authors review prior literature on prosocial behaviors among different ethnic/cultural groups throughout this article. As limited studies have examined prosocial behaviors among some specific ethnic groups, the authors conclude with recommendations for future research.

Adolescence is a period of human development marked by several biological, cognitive, and social transitions. Physical changes, such as the onset of puberty and rapid changes in body composition (e.g., height, weight, and sex characteristics) prompt adolescents to engage in greater self-exploration (McCabe and Ricciardelli, 2003). Enhanced cognitive abilities permit adolescents to engage in more symbolic thinking and to contemplate abstract concepts, such as the self and one's relationship to others (Kuhn, 2009; Steinberg, 2005). Furthermore, adolescence is marked with increased responsibilities at home and in the school context, opportunities for caregiving within the family, and mutuality in peer relationships (American Psychological Association, 2008). Moreover, society demands a greater level of psychosocial maturity and expects greater adherence to social norms from adolescents compared to children (Eccles et al., 2008). Therefore, adolescence presents itself as a time of major life transitions. In light of these myriad transitions, adolescents are further developing *prosocial behaviors*.

Although the emergence of prosocial behaviors (e.g., expressed behaviors that are intended to benefit others) begins in early childhood, the developmental transitions described above allow adolescents to become active agents in their own developmental process. Behavior that is motivated by adolescents' concern for others is thought to reflect optimal social functioning or prosocial behaviors (American Psychological Association, 2008). While the early literature focused primarily on prosocial behaviors among young children (e.g., Garner, 2006; Garner et al., 2008; Iannotti, 1985) there are several reasons to track prosocial development into adolescence. First and foremost, individuals develop cognitive abilities that allow them to better phenomenologically process and psychologically mediate life experiences that may facilitate (e.g., completing household chores and caring for siblings) or hinder (e.g., interpersonal conflict and perceptions of institutional discrimination) prosocial development (e.g., Brown and Bigler, 2005). Adolescents express more intentionality in which activities they will engage in and become selective in where they choose to devote their energies (Mahoney et al., 2009). Finally, adolescents are afforded more opportunities to express helping behaviors in other social spheres beyond the family context, such as in schools, communities, and civic society (Yates and Youniss, 1996).

Origins and Definitions of Prosocial Behaviors

Since the turn of the twenty-first century, there has been growing interest in understanding the relationships that exist between the strengths of individuals and resources within communities (e.g., person ↔ context) in order to identify pathways for healthy development, or to understand how adolescents' thriving can be promoted. This line of thinking is commonly described as the positive youth development perspective (e.g., Lerner et al., 2009). Although the adolescent literature still predominantly focuses on problematic development (e.g., delinquency and risk-taking behaviors), studies on adolescents' prosocial development have increased substantially since the 1990s (Eisenberg et al., 2009a), paralleling the paradigm shift from a deficit-based model of development to one focusing on positive attributes of youth (e.g., Benson et al., 2006; Lerner, 2005).

Generally described as the expression of voluntary behaviors with the intention to benefit others (Carlo, 2006; Eisenberg, 2006; see full review by Eisenberg et al., 2009a), prosocial behavior is one aspect among others of positive adolescent development that is gaining greater attention in the literature. Theory on prosocial development is rooted in the literature on moral development, which includes cognitive aspects of moral reasoning (e.g., how individuals decide between moral dilemmas; Kohlberg, 1978), moral behaviors (e.g., expression of behaviors that benefit society; Eisenberg and Fabes, 1998), and emotions (e.g., empathy; Eisenberg and Fabes, 1990).

Empirical studies on adolescents' prosocial development have found that different types of prosocial behaviors may exist. For example, Carlo and colleagues (e.g., Carlo et al., 2010; Carlo and Randall, 2002) found six types of *prosocial tendencies* (intentions to help others): *compliant*, *dire*, *emotional*, *altruistic*, *anonymous*, and *public*. Compliant helping refers to an individual's intent to assist when asked. Emotional helping refers to helping in emotionally evocative situations (e.g., witnessing another individual crying). Dire helping refers to

intentions to assist in emergency situations. Altruistic helping refers to a person's intention to help with no expectation of reward. Anonymous has been described as helping without anyone knowing. Finally, public helping has been described as intentions to help when others are watching. Research on these types of prosocial tendencies indicates that some behaviors are more related to each other than others. For example, Calderón-Tena et al. (2011) found that compliant, emotional, and dire helping may be related positively to one another, while other studies suggest that altruistic helping may be different from other types of prosocial tendencies (e.g., Brittian et al., 2013).

Research indicates that adolescents evaluate the situational context when determining whether to engage in prosocial behavior (e.g., Tisak et al., 2002). Some youth who engage in high levels of public prosocial behavior may do so out of self-interest needs rather than for the benefit of others. For instance, in a study of largely Caucasian adolescents, *proactive prosocial behavior* (e.g., behaving prosocially to benefit the self) was the only type of prosocial behavior that was related to aggressive behavior. Specifically, adolescents who reported positive or normative beliefs about engaging in aggressive behaviors were more likely to engage in proactive prosocial behaviors, and were less likely to engage in altruistic helping and reactive aggressive behaviors (Boxer et al., 2004). Proactive prosocial behavior requires more socially sophisticated and manipulative conceptualizations of behaviors (Boxer et al., 2004). In this particular study, adolescents who endorsed retaliatory aggressive beliefs were more likely to engage in proactive prosocial behavior. This suggests that prosocial behaviors that are self-serving may be facilitated by motivations that are similar to those that facilitate aggressive behavior. If this is true, then perhaps the label prosocial behavior is inadequate or inappropriate for these types of self-serving behaviors since the behavior is not based on helping others, but rather is motivated by satisfying the needs of the self. More importantly, what does it mean to engage in prosocial behaviors for the benefit of the self? Does it yield differential effects for adolescents? Do these types of behaviors still serve as a protective factor against challenging or problematic experiences? These questions not only have research implications, but applied implications as well. For instance, educators and providers have to consider how and what types of prosocial behaviors are promoted through programs, in schools, and by parents and caregivers.

The expression of prosocial behaviors also varies by demographic and developmental characteristics, such as gender and age. In general, adolescent girls are more likely to exhibit prosocial behaviors compared to adolescent boys (e.g., Hastings et al., 2007). However, some studies have observed that boys engage in more public helping (e.g., helping when one is being observed; Carlo and Randall, 2002). More helping behaviors among girls may reflect that girls are socialized to be more caring and nurturing and to exhibit greater levels of concern for others (Eisenberg et al., 2009). Divergent from other studies on prosocial behavior that found gender differences, McMahon et al. (2006) reported that African-American boys exhibited more prosocial behaviors than African-American girls. Therefore, gender differences in prosocial behaviors may vary among US adolescents based on cultural differences (Eisenberg and Morris, 2004) as well.

The literature on prosocial behaviors and ontogeny has yielded mixed age difference findings depending on the type of prosocial behavior assessed. Some research (Carlo and Randall, 2002; Fabes et al., 1999) contends that older adolescents engage in more prosocial behaviors while other research has noted mixed findings. In general, older children have more advanced prosocial skills typically due to increased empathetic and perspective taking abilities, and increased and nuanced opportunities for social interactions (Kokko et al., 2006). In addition, some types of social interactions run counter to each other. For example, individuals on average tend to become more prosocial and less competitive and individualistic across the life span (Van Lange et al.,1997). Differential findings have shown that prosocial behaviors decrease in late adolescents, while other researchers have not found age differences among adolescents. For example, there is some research that revealed a declining trend in prosocial behavior when children are assessed longitudinally using the same forms of measurement (Kokko et al., 2006; Côté et al., 2002). Therefore, it is unclear whether the decline in prosocial behaviors observed was due to developmental changes as adolescents mature or were attributed to issues of measurement (i.e., measurement of prosocial behavior across time is not robust enough to detect developmental changes). Due to mixed findings, future research needs to explore gender and developmental age changes in a more sophisticated way that examines the intersectionality of these two demographic issues (gender and age) along with the role of culture.

Antecedents to Prosocial Behavior

Adolescents' prosocial behaviors do not develop void of contextual influence. Youth are part of a larger system, involving family, peers, school, societal values, and the historical context (Bronfenbrenner, 2005). Indeed, research describes two major contextual factors that foster prosocial behaviors: *socialization* and *cultural orientations*.

Socialization

Socialization is the process through which individuals acquire beliefs, values, social norms, and practices that allow them to successfully interact with society (Göncü and Gauvain, 2012; Rogoff, 2003). Parents represent a key socializing agent through which adolescents learn to express prosocial behaviors (Carlo et al., 2007). One direct way that socialization relates to prosocial behaviors is through parenting practices. A study by Carlo et al. (2011) found that parents' socialization, measured as parental inductions (e.g., degree of positive reasoning and explanations that parents use with their adolescent) played a significant role in promoting sympathy among adolescents, consequently promoted prosocial tendencies indirectly; however, no significant ethnic differences in these associations were found between Mexican-American and Caucasian adolescents who participated in the research. A multicultural study by Shen

et al. (2013) indicated that inductive parenting practices were positively related to moral development indirectly through empathy among Caucasian, Mexican-American, and Taiwanese adolescents. Interestingly, the aforementioned study reported that parenting practices were related to sympathy but not perspective-taking among Mexican-American and Taiwanese adolescents compared to Caucasian adolescents. Researchers noted that more studies are needed across cultural groups to understand why these cultural differences were observed.

In addition to direct socialization (e.g., talking to adolescents about helping others), adolescents may acquire prosocial behaviors through informal social learning (Lave and Wenger, 1991), where parents model behaviors and youth observe how parents interact with other members of society. Certainly, some research indicates that adolescents are likely to engage in helping behaviors with parents, such as volunteering and community service (e.g., Fletcher et al., 2000). These activities may or may not continue throughout adolescence, as youth form their own goals and interests.

Peers represent another prominent social context for adolescents' prosocial behaviors. Wentzel et al. (2007) found that peers played an influential role in adolescents' prosocial behavior. Specifically, adolescents' perceptions of their peers' expectations regarding prosocial behavior was significantly related to the adolescents' engagement in actual prosocial behaviors at school. Important note, researchers indicated that the racial demographics of the school may have influenced this effect. In this case, peer influence on adolescents' behavior appeared to be influenced by whether the student attended a school where his or her racial group represented the majority of the student population. Peer influences are typically discussed in terms of negative or antisocial influences; however, this research provides initial support for the positive influence that peers can have on adolescents' developmental outcomes, mainly prosocial behaviors. Although further research is needed to unpack this finding, these findings do highlight the positive influential role of peers in adolescents' development.

Socialization of prosocial behaviors can occur in more indirect ways than social modeling. Mainly, adolescents' perceptions of their parents and peers' perspectives on prosocial behavior can also have a significant impact on adolescents' prosocial behaviors. For example, adolescents' perceptions of parental and peer prosocial behavioral norms, assessed as parents and peers approval or disapproval of their participation in problem behaviors, emerged as protective factors in reducing risk factors associated with American Indian youth's engagement in violence in an urban context (Bearinger et al., 2005). While the literature on cultural variations of socialization and adolescents' prosocial behavior is limited, there is evidence to suggest that the environment that parents create and parents' expectations for service, as well as peer culture, are directly and indirectly related to the development of adolescents' prosocial behaviors.

Cultural Orientations

It has been well documented that culture influences children and adolescents' development (Rogoff, 2003); however, less is known about the relationship between culture and positive aspects of adolescent development, such as prosocial behaviors. Systematic research on the development of prosocial behaviors among youth from diverse cultural backgrounds is underdeveloped and rarely explored (Humphries and Jagers, 2009). The use of a cultural framework or perspective to understand adolescents' prosocial development acknowledges that human development is situated in and is influenced positively by cultural values, traditions, and institutions (American Psychological Association, 2008).

This section reviews research in two main contexts where studies elucidate how culture intersects with prosocial behaviors among adolescents: (1) communal/collective orientations and (2) religious/spiritual orientations. Both cultural orientations have been examined in relation to prosociality. First, the authors will discuss the literature on communalism/collectivism, followed by religiosity/spirituality.

Communal/Collective Orientations

Research on collective and communal orientations and prosocial behaviors among minority youth dates back to the 1970s. Knight and Kagan (1977) observed that Mexican-American children (ages 5–9 years) expressed more prosocial behaviors compared to Anglo-American children who exhibited higher levels of individualism and competitiveness. Researchers currently posit that collaboration, a group or communal orientation, is a central feature of Mexican-American cultural orientation that promotes prosocial behaviors (Knight and Carlo, 2012). Several contemporary studies have sought to understand how cultural factors relate to the development of prosocial tendencies among Latino adolescents (e.g., Armenta et al., 2011). For example, a study by Schwartz et al. (2007) found that an orientation toward Hispanic culture was positively related to Latino early adolescents' prosocial behaviors. Calderón-Tena et al. (2011) found that Mexican-American cultural values, assessed as *familism* (e.g., respect for elders and family obligations), were positively related to Mexican-American adolescents' prosocial ideas. Moreover, Armenta et al. (2011) observed that Mexican-American cultural values (assessed as types of familism values) were positively related to several types of prosocial tendencies (e.g., compliant, emotional, dire, and anonymous) compared to mainstream American values (assessed as self-reliance, material success, and personal achievement) that related to more public helping and less altruistic helping.

Ward (1995) contended that the dominant culture in the United States negatively impacts African-American adolescents' positive development. Mainly, she asserted that the excessive focus on autonomy of the individual and the preoccupation with and competition for consumption of materialism is thought to erode prosocial behaviors among African-American adolescents. Moreover, she argued that American mainstream cultural values of *individualism* and *materialism* actually run counter to the cultural orientations of African-Americans, which includes *communalism* and *religiosity/spirituality*.

Communalism implies an awareness of the fundamental interdependence of others with a premium being placed on

social bonds and group obligations (Boykin et al., 1997). This specific cultural orientation places the needs of the group before those of the individual. Studies on African-Americans have found that communalism is related positively to correlates of African-American adolescents' prosocial behaviors, including empathy (Humphries et al., 2000; Jagers et al., 2007) and moral reasoning (Humphries et al., 2000; Woods and Jagers, 2003). Humphries et al. (2000) observed that communalism emerged as a significant predictor of moral reasoning among African-American fifth and eighth grade boys, but not among African-American girls. This gender by culture finding was surprising given that African-American girls in the study reported higher communalism scores and seemed to resonate to a communal orientation more so than their male counterparts. Therefore, more research is needed to understand the intersection between aspects of culture, such as communalism, and gender among African-American youth.

This is not to suggest that a communal or collective orientation is uniquely associated with African-Americans. Research indicates that other cultural groups that utilize a collective or community-focused frame of reference exhibit similar associations with prosocial behaviors. For example, similar patterns of association have been found among Native American youth. Reservation living, even in the context of intergenerational poverty, may provide Native American youth with a source of support and protection facilitating their positive development. Specifically, LaFramboise et al. (2006) noted that community support, maternal warmth, and enculturation (orientation toward one's culture of origin) emerged as protective factors for Native American early adolescents living on a reservation. Notably, the majority of youth in this study (over 60%) were characterized as prosocial while avoiding problem behaviors (LaFramboise et al., 2006).

Religious/Spiritual Orientations

Many studies have acknowledged the positive association between religious/spiritual orientation and prosocial behaviors (e.g., Furrow et al., 2004; Jagers and Mock, 1993; Johnson, 2008; King and Benson, 2005). Some research indicates that religious orientation serves as a type of social control by discouraging risky behaviors among adolescents, including substance use and delinquency (e.g., Brittian et al., 2013; Wallace and Forman, 1998). Most religions encourage adolescents to think beyond themselves, bolster concern for others well-being, and promote development of a broader worldview (Furrow et al., 2004; King, 2003). Moreover, it has been suggested that many religions are associated with prosocial behaviors due to an emphasis on helping others, service to the community, and promoting a collective orientation (Saroglou et al., 2004; Eisenberg et al., 2011). Furthermore, a spiritual or religious orientation requires the belief in a higher being and a conceptualization that connects the individual to this higher being. Therefore, this relationship requires perspective taking, a skill that is necessary for optimal social development and implementation of prosocial ideas and behaviors.

Research on adolescent populations from various cultural groups has yielded findings that support the association between religious/spiritual orientation and prosocial behaviors. In a diverse sample of youth, involving largely Latino adolescents, religious identity was associated with prosocial beliefs and attitudes (Furrow et al., 2004). A significant association was found between religiosity and certain types of prosocial behavior for Caucasian adolescents (Hardy and Carlo, 2005). Specifically, Caucasian adolescents who reported higher levels of religiosity were more likely to engage in altruistic, compliant, and anonymous forms of prosocial behaviors. Eisenberg et al. (2009b) found that cross-religious friendships were a significant factor for Indonesian adolescents' socioemotional functioning and prosocial behaviors at school. Moreover, French et al. (2008) observed that religious involvement was positively related to prosocial behaviors among eighth and ninth grade Indonesian Muslim youth, and researchers attributed this positive association to the collectivistic nature of religion. Similarly, a longitudinal study of Italian adolescents indicated that religious coping was predictive of more prosocial behaviors (Eisenberg et al., 2011).

Finally, spirituality and religion are important cultural dimensions that are present in African-American culture and the everyday lives of African-American people (Mattis and Jagers, 2001; Taylor et al., 2004). A longitudinal study by Smetana and Metzger (2005) found that African-American adolescents' religious and spiritual beliefs were longitudinally predictive of their community involvement and engagement in prosocial behaviors 3 years later. This overall body of research provides further evidence for the need of future studies to examine particular types of prosocial behaviors and attitudes as they relate to religiosity and spirituality among adolescents in order to gain a better understanding of the associations between this cultural orientation and positive development.

Methodological Issues and Future Directions for Research

One persistent question that remains unanswered in the literature is around intentions to engage in helping behaviors. Some research attempts to address this question. For example, prior literature indicates that, in general, youth are more likely to assist close friends than acquaintances or strangers (Tisak et al., 2002). In addition, some studies have found that an adolescent's willingness to engage in prosocial behaviors may be influenced by the cultural background of the 'other.' Specifically, adolescents may be more willing to direct prosocial actions toward those who share their same cultural background; this has been found among studies with children (e.g., Zimmerman and Levy, 2000; Zinser et al., 1981). Furthermore, this issue may be complicated by whether the adolescent is in an environment where his or her cultural group is in the majority or minority (e.g., Eisenberg et al., 2009). Therefore, future studies should consider how the target of one's prosocial behavior influences intentions to help in various social situations.

In addition, some critical areas around conceptualization (e.g., definition) and measurement of prosocial behaviors as they relate to cultural groups exist. For example, some studies may assess volunteerism or civic engagement among adolescents (e.g., Hart et al., 2007), but may not specifically describe

these actions as prosocial behaviors. For example, Sherrod and Lauckhardt (2009) posit that the literature on civic engagement has yielded mixed findings with regard to minority youth. It appears that culture and ethnicity influence adolescents' prosocial behaviors but in complex ways (e.g., Carlo et al., 1999). In addition, it may be difficult to distinguish how ethnicity and race independently influence adolescents' prosocial development since these characteristics are often confounded with social class (Sherrod and Lauckhardt, 2009). In this case, more within-cultural group investigation is needed. Furthermore, when studies account for the availability of opportunities to engage in prosocial behaviors, research findings become more complicated (Balsano, 2005). For this reason, it is possible that some of the differences observed in previous studies can be attributed to definition and measurement of prosocial behaviors. Therefore, more research is needed to understand how culture intersects with other demographic characteristics (e.g., income) to influence prosocial behaviors among adolescents.

Emerging literature indicates that negative social experiences, such as perceived interpersonal and institutional discrimination, pose a significant threat to prosocial development (e.g., Brittian et al., 2013; see also Sherrod and Lauckhardt, 2009). However, other research suggests that a strong sense of fidelity (e.g., the ability to maintain prosocial values in spite of societal contradictions; Erikson, 1965) may lead to prosocial behaviors, if adolescents commit to social groups and institutions. For example, if a young person feels strongly and positively about his or her cultural group and perceives that his or her group is treated unfairly in society (e.g., perceived institutional discrimination), he or she may choose to engage in social and political arenas in order to improve the conditions for his or her cultural group (e.g., O'Leary and Romero, 2011). This type of prosocial behavior may be described as a precursor to engagement in social justice among adolescents (e.g., Ginwright, 2007) or a form public helping. Brittian et al. (2013) found that perceptions of ethnic discrimination were indirectly related to public helping among Mexican-American adolescents through a direct relationship with Mexican-American cultural values. O'Leary and Romero (2011) found instantiations of increased civic engagement in response to an antiethnic studies government bill among Mexican descent students' and this type of 'public' helping served as a protective factor for students' mental health. Therefore, future studies should further investigate factors that impede prosocial behaviors among minority adolescents and continue to explore conditions through which prosocial behaviors occur in spite of these perceived challenges.

Conclusions

Although research on adolescents' prosocial behaviors has burgeoned significantly over the past 20 years, there are several areas that the literature can be advanced. Despite the relative growth in research examining adolescents' prosocial behaviors in general, research on cultural groups is scarce, with notable exceptions described in this article. In addition, several areas were elucidated that promote adolescents' prosocial behaviors (e.g., parent and peer socialization, and cultural orientations) that warrant further investigation.

See also: Eisenberg's Theory of Prosocial Reasoning; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Prosocial Behavior, Effects of Parenting and Family Structure on.

Bibliography

American Psychological Association, Task Force on Resilience and Strength in Black Children and Adolescents, 2008. Resilience in African American Children and Adolescents: A Vision for Optimal Development. American Psychological Association, Washington, DC.

Armenta, B.E., Knight, G.P., Carlo, G., Jacobson, R.P., 2011. The relation between ethnic group attachment and prosocial tendencies: the mediating role of cultural values. European Journal of Social Psychology 41, 107–115.

Balsano, A., 2005. Youth civic engagement in the United States: understanding and addressing the impact of social impediments on positive youth and community development. Applied Development Science 9, 188–201.

Bearinger, L.H., Perringell, S., Resnick, M.D., Skay, C.L., Potthoff, S.J., Eichorn, J., 2005. Violence perpetration among urban American Indian youth: can protection offset risk? Archives of Pediatrics & Adolescent Medicine 159, 270–277.

Benson, P.L., Donahue, M.J., Erickson, J.A., 1993. The faith maturity scale: conceptualization, measurement, and empirical validation. Research in the Social Scientific Study of Religion 5, 1–26.

Benson, P.L., Scales, P.C., Hamilton, S.F., Sesma Jr., A., 2006. Positive youth development: theory, research, and applications. In: Lerner, R.M. (Ed.), Handbook of Child Psychology: Vol. 1 Theoretical Models of Human Development, sixth ed. John Wiley, Hoboken, NJ, pp. 894–941.

Boxer, P., Tisak, M.S., Goldstein, S.E., 2004. Is it bad to be good? an exploration of aggressive and prosocial behavior subtypes in adolescence. Journal of Youth and Adolescence 33, 91–100.

Boykin, A.W., Jagers, R.J., Ellison, C., Albury, A., 1997. Communalism: conceptualization and measurement of an Afrocultural social ethos. Journal of Black Studies 27, 409–418.

Brittian, A.S., Lewin, N., Norris, S.A., 2013. South African adolescents' perceptions of religion in time of social change. Journal of Adolescent Research 28, 642–663.

Brittian, A.S., O'Donnell, M., Knight, G.P., Carlo, G., Umaña-Taylor, A.J., Roosa, M.W., 2013. Associations between adolescents' perceived discrimination and prosocial tendencies: the mediating role of Mexican American values. Journal of Youth and Adolescence 42, 328–341.

Bronfenbrenner, U., 2005. Making Human Beings Human: Bioecological Perspectives on Human Development. Sage, Thousand Oaks, CA.

Brown, C.S., Bigler, R.S., 2005. Children's perceptions of discrimination: a developmental model. Child Development 76, 533–553.

Calderón-Tena, C.O., Knight, G.P., Carlo, G., 2011. The socialization of prosocial behavior tendencies among Mexican American adolescents: the role of familism values. Cultural Diversity and Ethnic Minority Psychology 17, 98–106.

Carlo, G., 2006. Care-based and altruistically based morality. In: Killen, M., Smetana, J.G. (Eds.), Handbook of Moral Development. Erlbaum, Mahwah, NJ, pp. 551–579.

Carlo, G., Randall, B.A., 2002. The development of a measure of prosocial behaviors for late adolescents. Journal of Youth and Adolescence 31, 31–44.

Carlo, G., Fabes, R., Laible, D., Kupanoff, K., 1999. Early adolescence and prosocial/moral behavior II: the role of social and contextual influences. Journal of Early Adolescence 19, 132–133.

Carlo, G., Knight, G.P., McGinley, M., Hayes, R., 2011. The roles of parental inductions, moral emotions, and moral cognitions in prosocial tendencies among Mexican American and European American early adolescents. Journal of Early Adolescence 31, 757–781.

Carlo, G., Knight, G.P., McGinley, M., Zamboanga, B.L., Jarvis, L.H., 2010. The multidimensionality of prosocial behaviors and evidence of measurement equivalence in Mexican American and European American early adolescents. Journal of Research on Adolescence 30, 334–358.

Carlo, G., McGinley, M., Hayes, R., Batenhorst, C., Wilkinson, J., 2007. Parenting styles or practices? Parenting, sympathy, and prosocial behaviours among adolescents. Journal of Genetic Psychology 168, 147–176.

Côté, S., Tremblay, R.E., Nagin, S., Zoccolillo, M., Vitaro, F., 2002. The development of impulsivity, fearfulness, and helpfulness during childhood: patterns of consistency and change in the trajectories of boys and girls. Journal of Child Psychology and Psychiatry 43, 609–618.

Eccles, J.S., Brown, B.V., Templeton, J., 2008. A developmental framework for selecting indicators of well-being during the adolescent and young adult years. In: Brown, B. (Ed.), Key Indicators of Child and Youth Well-being. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 197–236.

Eisenberg, N., 2006. Prosocial behavior. In: Bear, G.G., Minke, K.M. (Eds.), Children's Needs III: Development, Prevention, and Intervention. National Association of School Psychologists, Washington, DC, pp. 313–324.

Eisenberg, N., Fabes, R.A., 1990. Empathy: conceptualization, measurement, and relation to prosocial behavior. Motivation and Emotion 14, 131–149.

Eisenberg, N., Fabes, R.A., 1998. Prosocial development. In: Damon, W., Eisenberg, N. (Eds.), Handbook of Child Psychology, Vol. 3: Social, Emotional, and Personality Development, fifth ed. John Wiley, New York, pp. 701–778.

Eisenberg, N., Fabes, R.A., 1999. Emotion, emotion-related regulation, and quality of socioemotional functioning. In: Balter, L., Tamis-LeMonda, C. (Eds.), Child Psychology: A Handbook of Contemporary Issues. Psychology Press, New York, pp. 318–335.

Eisenberg, N., Morris, A.S., 2004. Moral cognitions and prosocial responding in adolescence. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. Wiley, Hoboken, NJ, pp. 155–188.

Eisenberg, N., Castellani, V., Panerai, L., Eggum, N., Cohen, A., Pastorelli, C., et al., 2011. Trajectories of religious coping from adolescence into early adulthood: their form and relations to externalizing problems and prosocial behaviour. Journal of Personality 79 (4), 841–873.

Eisenberg, N., Morris, A.S., McDaniels, B., Spinrad, T.L., 2009a. Moral cognitions and prosocial responding in adolescence. In: Steinberg, I., Lerner, R. (Eds.), Handbook of Adolescent Psychology, third ed. Wiley, Hoboken, NJ, pp. 229–265.

Eisenberg, N., Sallquist, J., French, D.C., Purwono, U., Suryanti, T.A., Pidada, S., 2009b. The relations of majority-minority group status and having an other-religion friend to Indonesian youths' socioemotional functioning. Developmental Psychology 45, 248–259.

Erikson, E.H., 1965. Youth: fidelity and diversity. In: Erikson, E.H. (Ed.), The Challenge of Youth. Anchor Books, New York, pp. 1–28.

Fabes, R.A., Carlo, G., Kupanoff, K., Laible, D., 1999. Early adolescence and prosocial/moral behavior I: the role of individual processes. Journal of Early Adolescence 19, 5–16.

Fletcher, A.C., Elder Jr., G.H., Mekos, D., 2000. Parental influences on adolescent involvement in community activities. Journal of Research on Adolescence 10, 29–48.

French, D.C., Eisenberg, N., Vaughan, J., Purwono, U., Suryanti, T.A., 2008. Religious involvement and the social competence and adjustment of Indonesian Muslim adolescents. Developmental Psychology 44, 597–611.

Furrow, J.L., King, P.E., White, K., 2004. Religion and positive youth development: identity, meaning, and prosocial concerns. Applied Developmental Science 8, 17–26.

Garner, P.W., 2006. Prediction of prosocial and emotional competence from maternal behavior in African American preschoolers. Cultural Diversity and Ethnic Minoirty Psychology 12, 179–198.

Garner, P.W., Dunsmore, J.C., Southam-Gerrow, M., 2008. Mother-child conversations about emotions: linkages to child aggression and prosocial behavior. Social Development 17, 259–277.

Ginwright, S.A., 2007. Black youth activism and the role of critical social capital in black community organizations. American Behavioral Scientist 51, 403–418.

Göncü, A., Gauvain, M., 2012. Sociocultural approaches to educational psychology: theory, research, and application. In: Harris, K., Graham, S., Urdan, T., McCormick, C.B., Sinatra, G.M., Sweller, J. (Eds.), Handbook of Educational Psychology. Vol. 1: Theories, Constructs, and Critical Issues. American Psychological Association, Washington, pp. 123–152.

Hardy, S.A., Carlo, G., 2005. Religiosity and prosocial behaviors in adolescence: the mediating role of prosocial values. Journal of Moral Education 34, 231–249.

Hart, D., Donnelly, T., Youniss, J., Atkins, R., 2007. High school community service as a predictor of adult voting and volunteering. American Educational Research Journal 44, 197–219.

Hastings, P.D., Utendale, W.T., Sullivan, C., 2007. The socialization of prosocial development. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization: Theory and Research. Guilford Press, New York, pp. 638–664.

Humphries, M.L., Jagers, R.J., 2009. Culture: a possible predictor of morality for African American adolescents. Journal of Research on Adolescence 19, 205–215.

Humphries, M.L., Parker, B.L., Jagers, R.J., 2000. Predictors of moral reasoning among African American children: a preliminary study. Journal of Black Psychology 26, 51–64.

Iannotti, R.J., 1985. Naturalistic and structured assessments of prosocial behavior in preschool children: the influence of empathy and perspective taking. Developmental Psychology 21, 46–55.

Jagers, R.J., Mock, L.O., 1993. Culture and social outcomes among inner-city African-American children: an Afrographic exploration. Journal of Black Psychology 19, 391–405.

Jagers, R.J., Sydnor, K., Mouttapa, M., Flay, B.R., 2007. Protective factors associated with preadolescent violence: preliminary work on a cultural model. American Journal of Community Psychology 40, 138–145.

Johnson, B.R., 2008. A tale of two religious effects: evidence for the protective and prosocial impact of organic religion. In: Kline, K.K. (Ed.), Authoritative Communities: The Scientific Case for Nurturing the Whole Child. Springer, New York, pp. 187–226.

King, P.E., 2003. Religion and identity: the role of ideological, social, and spiritual contexts. Applied Developmental Science 7, 197–204.

King, P.E., Benson, P.L., 2005. Spiritual development and adolescent well-being and thriving. In: Roehlkepartain, E.C., King, P.E., Wagner, L., Benson, P.L. (Eds.), The Handbook of Spiritual Development in Childhood and Adolescence. Sage, Thousand Oaks, CA, pp. 384–398.

Knight, G.P., Carlo, G., 2012. Prosocial development among Mexican American youth. Child Development Perspectives 6, 258–263.

Knight, G.P., Kagan, S., 1977. Development of prosocial and competitive behaviors in Anglo-American and Mexican-American children. Child Development 48, 1385–1394.

Kohlberg, L., 1978. Revisions in the theory and practice of moral development. New Directions for Child Development 2, 83–88.

Kokko, K., Tremblay, R.E., Lacourse, E., Nagin, D.S., Vitaro, F., 2006. Trajectories of prosocial behavior and physical aggression in middle childhood: links to adolescent school dropout and physical violence. Journal of Research on Adolescence 16, 403–428.

Kuhn, D., 2009. Adolescent thinking. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, Vol. 1: Individual Bases of Adolescent Development, third ed. Wiley, Hoboken, NJ, pp. 152–228.

LaFramboise, T.D., Hoyt, D.R., Oliver, L., Whitbeck, L.B., 2006. Psychological impact of biculturalism: evidence and theory. Psychological Bulletin 114 (3), 395–412.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, England.

Lerner, R.M., 2005. Foreword: promoting positive youth development through community and after school programs. In: Mahoney, J.L., Larson, R.W., Eccles, J. (Eds.), Organized Activities as Contexts of Development: Extracurricular Activities, After-School and Community Programs. Erlbaum, Mahwah, NJ, pp. ix–xii.

Mahoney, J.L., Vandell, D.L., Simpkins, S., Zarrett, N., 2009. Adolescent out-of-school activities. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology: Contextual Influences on Adolescent Development, third ed. Wiley, Hoboken, NJ, pp. 228–269.

Mattis, J.S., Jagers, R.J., 2001. A relational framework for the study of religiosity and spirituality in the lives of African Americans. Journal of Community Psychology 29, 519–539.

McCabe, M.P., Ricciardelli, L.A., 2003. Sociocultural influences on body image and body changes among adolescent boys and girls. Journal of Social Psychology 143, 5–26.

McMahon, S.D., Wernsman, J., Parnes, A.L., 2006. Understanding prosocial behavior: the impact of empathy and gender among African American adolescents. Journal of Adolescent Health 39, 135–137.

O'Leary, A.O., Romero, A., 2011. Chicana/o students respond to Arizona's anti-ethnic studies bill, SB 1108. Aztlán: A Journal of Chicano Studies 36, 9–36.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, New York, NY.

Saroglou, V., Delpierre, V., Dernelle, R., 2004. Values and religiosity: a meta-analysis of studies using Schwartz's model. Personality and Individual Differences 37, 721–734.

Schwartz, S., Zamboanga, B., Jarvis, L.H., 2007. Ethnic identity and acculturation in Hispanic early adolescents: mediated relationships to academic grades, prosocial behavior, and externalizing symptoms. Cultural Diversity and Ethnic Minority Psychology 13, 364–373.

Shen, Y.-L., Carlo, G., Knight, G.P., 2013. Relations between parental discipline, empathy-related traits, and prosocial moral reasoning: a multicultural examination. Journal of Early Adolescence 33 (7), 994–1021.

Sherrod, L.R., Lauckhardt, J., 2009. The development of citizenship. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology. Vol. 2: Contextual Influences on Adolescent Development, third ed. Wiley & Sons, Hoboken, NJ, pp. 372–408.

Smetana, J.G., Metzger, A., 2005. Family and religious antecedents of civic involvement in middle class African American late adolescents. Journal of Research on Adolescence 15, 325–352.

Steinberg, L., 2005. Cognitive and affective development in adolescence. Trends in Cognitive Sciences 9, 69–74.

Taylor, R.J., Chatters, L.M., Levin, J., 2004. Religion in the Lives of African Americans: Social, Psychological, and Health Perspectives. Sage, Thousand Oaks, CA.

Tisak, J., Maynard, A.M., Tisak, M.S., 2002. AIRA: measurement of adolescents' judgments regarding intentions to respond to physical and verbal aggression. Aggressive Behavior 28, 207–223.

Van Lange, P.A.M., Otten, W., de Bruin, E.M.N., Joireman, J.A., 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. Journal of Personality and Social Psychology 73 (4), 733–756.

Wallace, J.M., Forman, T.A., 1998. Religion's role in promoting health and reducing risk among American youth. Health, Education & Behavior 25, 721–741.

Ward, J.V., 1995. Cultivating a morality of care in African American adolescents: a culture based model of violence prevention. Harvard Educational Review 65, 175–188.

Wentzel, K.R., Filisetti, L., Looney, L., 2007. Adolescent prosocial behavior: the role of self-processes and contextual cues. Child Development 78, 895–910.

Woods, L.N., Jagers, R.J., 2003. Are cultural values predictors of moral reasoning in African American adolescents? Journal of Black Psychology 29, 102–118.

Yates, M., Youniss, J., 1996. A developmental perspective on community service in adolescence. Social Development 5 (1), 85–111.

Zimmerman, B.J., Levy, G.D., 2000. Social cognitive predictors of prosocial behavior toward same and alternate race children among white pre-schoolers. Current Psychology 19, 175–193.

Zinser, O., Rich, M.C., Bailey, R.C., 1981. Sharing behavior and racial preference in children. Motivation and Emotion 5, 179–187.

Prosocial Behavior during Childhood and Cultural Variations

Stuart I Hammond, Whitney Waugh, Emma Satlof-Bedrick, and Celia A Brownell, University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Children's prosocial behaviors include acts such as sharing, helping, comforting, cooperating, and can come to involve complex behaviors such as lying to protect someone's feelings. Because prosocial behaviors are performed for diverse reasons, both moral and nonmoral, research findings on prosociality in children are complex, although clear age trends are present. A variety of theories attempt to explain children's prosocial development, each providing insights into certain aspects. Although cultural influences are evident throughout the development of prosocial behavior, future research will need to untangle the true significance of these influences.

Understanding Prosocial Behavior

Defining Prosocial Behavior

Children's prosocial behavior encompasses a diverse range of behaviors that can include mundane acts like sharing toys with friends, remarkable acts like helping someone who is hurt or incapacitated, and even some very sophisticated acts like lying to protect another's feelings. The commonality that binds these behaviors together is that they are performed with the intention to benefit another person (e.g., Eisenberg et al., in press). However, beyond this point of cohesion, children's prosocial behaviors are marked by tremendous diversity, in type, motivation, and cause.

Researchers have long recognized that both moral motivations, such as an altruistic act produced solely for the good of another, and nonmoral motivations, such as positive acts to please a parent, to gain social status, or to maintain a friendship underlie prosocial behavior (Bierhoff, 2002). Prosocial behaviors are also diverse in their mechanisms. To the extent that prosocial behavior in children is thought to reflect a single unified construct, the findings of the larger body of research on prosocial behavior can seem disappointingly inconclusive or even baffling. For example, in younger children, some prosocial tendencies are correlated with aggressive tendencies, a seeming paradox (e.g., Persson, 2005). Considered more closely, these findings may reflect the fact that some children are skilled in using both instrumental aggression and instrumental prosocial acts to their benefit, in order to get other children to like them or do what they want. In yet other cases, it can be that children with poor impulse control, or even those who are more assertive, may be more likely to both help and harm others. Similarly, in terms of gender, although girls are thought to be more prosocial than boys, and do seem to have higher levels of self-reported compassion for others, actual findings in this area are mixed (see review in Eisenberg et al., in press). Differences tend to be stronger in survey data (self- and other-reported) and much more modest in observational or experimental data. These findings could reflect factors such as a greater assertiveness in boys, and thus a greater likelihood to engage in certain forms of prosociality.

Prosocial Development

A further complexity in understanding prosociality in the childhood years centers on the issue of development and transformation. Children's prosocial behavior lies at the intersection of early-developed action patterns and emerging social cognitive and social emotional capacities (Eisenberg et al., in press). Infants begin to display prosocial behaviors soon after their first birthday, such as sharing food or toys with a parent (Hay and Cook, 2007). As children go on to develop greater motor and physical competency, advance in social, cognitive, and affective understanding, encounter novel social situations such as in school and other community contexts, and become subject to new expectations from parents, peers, and teachers, their prosociality is further transformed (Eisenberg et al., in press).

Although at a general level prosocial behavior becomes more frequent and more reliable in older children, across cultures, developmental changes may vary the frequency of particular prosocial acts. For example, some prosocial behaviors may diminish as the motivation driving them declines or changes. Infants may help a parent with 'fun' activities such as setting the table because they are interested in taking part and because they enjoy spending time with the parent (Rheingold, 1982). With development, this activity might be seen increasingly as a boring chore, and children accordingly become less likely to help spontaneously. However, children may develop new motivations for these same prosocial behaviors, such as a desire to please their parent, receive an allowance, or because they perceive the chore as their responsibility. Or perhaps parents will no longer invite them to do this particular task, instead assigning a task such as caring for a pet.

As they develop, children also begin to view certain behaviors as inappropriate or unnecessary, whether due to recognition of social norms or to more accurate social perception. For example, older children become less likely to aid a stranger spontaneously, perhaps out of social propriety or because they can more accurately judge that a person does not always need help. Whereas infants' 'help' in the home is often viewed with irritation and exasperation by parents, because of the low competency of their prosocial behavior (Rheingold, 1982), children may come to learn prosocial behavior 'by omission,' recognizing that they have a responsibility to keep their rooms clean or keep out of their parents' way (Klein et al., 2009).

With the emergence of more advanced social capacities, children also begin to display novel forms of prosocial behavior. For example, although children are typically socialized to tell the truth but sometimes lie to hide wrongdoings,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23182-6

they may also come to learn that the truth may actually be hurtful, such that in some instances, such as receiving a disappointing gift, lying may actually be considered prosocial (e.g., Williams et al., 2013). With development, children become more competent prosocial actors, but the meaning and causes of their actions are also transformed, rendering it somewhat difficult to identify simple developmental trends.

Research and Practice

Researchers have focused the bulk of their empirical efforts on the relation of prosociality to more positively viewed psychological characteristics, particularly ones typically regarded as moral, such as caring for others and altruism, and less on the relation of prosocial behavior to factors such as status-seeking. This reflects in part an everyday valuing of morally driven prosociality, as well as the practical concern of promoting prosocial behaviors in applied and educational contexts, both as a means of promoting social harmony and good citizenship and as a means of reducing antisocial behaviors, such as bullying and aggression (e.g., Jennings and Greenberg, 2009). Despite these well-intended efforts, developmental trends raise challenging questions about the status of prosocial behavior and its relation to morality. Although prosocial behavior is often taken as an unalloyed good, and associated with positive development, the true developmental picture is more complicated. Some prosociality is associated with undesirable social and psychological characteristics and conversely, while some prosocial behavior, such as prosocial lying, may be prosocial in intent, its overall moral status is sometimes difficult to gauge.

Theories of Prosocial Development

There is no single dominant theoretical account of prosocial behavior. Rather, the major theories of prosocial development each make important contributions to different aspects of our understanding of children prosocial behavior and different areas of prosocial development.

Cognitive Developmental

The cognitive developmental moral theories of prominent figures such as Piaget, Kohlberg, and Turiel (Lapsley, 1996) are generally more concerned with the development of moral reasoning and moral motivation than with prosocial action. Traditional cognitive developmental theory has placed a heavy emphasis on the development of autonomous moral reasoning, in large part because Piaget's (1965) seminal work grew out of a reaction to the sociologist Durkheim's view of the imposition of moral norms on children by society. As a legacy of this reaction, prosocial behavior that originates in the commands and directives of a parent is regarded as somewhat detrimental to the child's moral development, although this is as much an issue of power relations in communication and parents can adopt more cooperative communicative styles. Conversely, cooperative interactions among peers, such playing games together, are thought to occur in the context of equality and mutual respect and to foster the child's autonomous reasoning and moral development.

Cognitive-developmental theories do not, in and of themselves, map out a clear pathway to the development of prosocial behavior and at least in the case of Piaget's theory, the primary question is the inverse: how moral reasoning emerges out of early prosocial interactions (Carpendale, 2009). Nevertheless, numerous studies have shown that children's prosocial behavior is positively, though modestly, associated with more advanced levels of moral reasoning (Eisenberg et al., in press).

Empathy

Empathy, an affective response that arises from comprehending the emotional state or condition of another, is thought to be a major factor in children's prosocial behavior. Research on empathy fixes motives for prosociality in other-oriented emotionality. Through the process of empathy, when children see another's emotion, usually a negative emotion such as sadness, they come to feel that emotion and are thereby motivated to help the other person (Hoffman, 2000). With development, children are able to empathize with broader features of another's situation, including imagining what someone might be feeling (e.g., after hearing about a natural disaster in another country) and becoming motivated to help (e.g., collecting donations).

Empathy is sometimes distinguished from sympathy, or showing concern for another without necessarily comprehending their emotional state and personal distress, or distress caused by the sight of another's suffering. Both empathy and sympathy are positively correlated with prosociality, and appear to become more so with age (Eisenberg et al., 2013). However, the route from a shared emotional state to actual prosocial action has long bedeviled researchers (Penner et al., 2005). One reason for this difficulty is an overwhelming focus on negative emotions such as sadness and pain, and correspondingly on prosocial behavior such as helping those who are upset are hurt. These situations are probably in the minority as compared to the amount of prosocial behaviors such as sharing and cooperation that occurs in positive and happy interactions (Brownell et al., 2002). Nevertheless, recent research on empathy has renewed interest in the role of emotions, emotion understanding, and other social cognitive capacities in the development of prosocial behavior.

Biological

Biological theories of prosocial behavior come in two principal forms. The first explores prosocial behaviors as a function of human and mammalian evolution. Earlier strands of this research generally focused on the paradox between selfish and self-maximizing behavior, necessary for reproductive success, and the existence of altruism, which reduces one's own reproductive success. Sacrificing or risking one's own life or well-being for another's means fewer of one's own (altruistic) genes survive in the gene pool, making the existence of altruism in a population difficult to explain. However, the paradox was ultimately resolved by theory and empirical findings showing that organisms are more likely to aid kin than nonkin.

More recent research, has attempted to demonstrate empirically that altruistic forms of prosociality are widespread in both humans and other mammals, and to provide plausible theories

about how prosocial behavior could have evolved (e.g., Tomasello and Vaish, 2013). To the extent that this research has aimed to show that prosociality is innate or unlearned, it has tended to focus on infants, who are, at least to a certain line of thought, not yet influenced by society or culture.

A second class of biological theories explains prosocial behavior with reference to neuropsychological development, and how individual differences in biological factors such emotional reactivity and empathy may be related to prosocial development (Eisenberg et al., in press). There does not seem to be a particular brain region associated with prosocial behavior, which given its diverse causes and motives, is perhaps unsurprising.

Evolutionary and biological theories are generally not geared toward explaining the developmental changes in prosociality that occur during childhood, although some recent studies have attempted to explore how children move from early cooperation and helping to more norm-driven forms of prosociality, such as making sure others follow rules (Tomasello and Vaish, 2013).

Social Learning

Social learning theories explore the social and interactional factors that play a role in prosocial development. Social learning theories place a heavy emphasis on the actors besides the child, particularly parents and peers, but also teachers, as well as larger social systems, such as schools and communities (Spivak and Howes, 2011). Nevertheless, parents remain an important focus of the literature, and particularly how parenting style, parental discipline techniques, and other socialization techniques are related to children's prosocial behavior (Grusec et al., 2013).

Although early social learning theory has in the past favored a unilateral transmission model of prosociality from parent and the larger society to the child, converging in some ways with the sociological model that Piaget sought to reject, more recent research has stressed a bidirectional developmental relation between the child and others. For example, as children display certain prosocial behaviors such as cleaning up in the context of play and spontaneous interaction, their parents may (or may not) come to expect these behaviors of the child. However, the child may also respond to the parent's bids in various ways, such as by disobeying parental requests for these acts, or differentially responding based on how the parent solicits cooperation (Grusec et al., 2013).

Cultural Factors in Prosocial Development

Framing Cultural Issues

Recent research on children's prosocial development has sought to explore cross-cultural differences. Cultural influences can arise at many points in children's prosocial behavior, not only in the way adults socialize prosocial behavior directly and indirectly, but also in the way that cultural tools and technologies afford particular behaviors. For example, societies with formal school systems radically change the types of interactions children have on a day-to-day basis, hence the opportunities, expectations, and demands for prosocial behavior.

Culture can be viewed as an additional influence on prosocial development that might influence differences in the frequency or timing of particular prosocial behaviors, either accelerating or delaying them, although the developmental pathways and mechanisms remain the same. Historically, the cognitive developmental approach to moral and prosocial development has framed it in terms of universal moral norms (e.g., respect for others) that originate in reason, and has argued against strong cultural relativism, which proposes moral thought and behavior unique to particular cultures. However, to the extent that these theories are constructvist, they clearly allow for individual differences. To some extent, the dichotomy between universality and relativism is also recreated in biological theories, albeit with common descent and heredity rather than reason as the framing point for universality.

Alternatively, culture can be seen as the medium in which all development, prosocial and otherwise, occurs (Rogoff, 2003). Cultural approaches may further invoke wholly separate cultural routes to prosocial development, involving different motivations and mechanisms by which children come to assist, comfort, and cooperate with others. For example, Kärtner et al. (2010) propose that German infants' prosociality develops out of a mentalistic self-other distinction, socialized by their mothers, whereas Indian infants additionally seem to display situational prosociality, which is oriented toward cooperative tasks and joint action as culture-specific characteristics. Arguably, this theoretical position can also be universalist in holding that human social functions, such as assisting others, develop universally, albeit through different pathways, i.e., individualistic versus collectivist norms (Greenfield et al., 2003). Of the major theoretical approaches to prosocial development, social learning theory is the most flexible in terms of examining and expecting broad cultural differences in prosociality.

Prosociality at Home

Rogoff's (2003) work on cultural development suggests that children get involved in different forms of prosocial behavior in the home environment across cultures. In some tribal cultures, such as the Efe in the Democratic Republic of Congo, even very young children are allowed to use machetes to cut fruit, whereas in typical North American households, kitchen knives, and sometimes even the kitchen itself, are considered off limits to young children, and only gradually sanctioned for use by older children. Thus, assisting others with meal preparation begins much earlier in the former. In many societies, children aged 5 and older begin to take care of other children, whereas in the UK, this is illegal if the older child is under 14.

In the urban United States, expectations for children's prosocial involvement at home appear substantially lower than for other societies and typically involve tasks such as cleaning up the house or one's bedroom and meal preparation (Klein et al., 2009). However, this cultural generality can vary from family to family. For example, children of single parents may be provided with more responsibilities to care for others and to take on more substantial household tasks (Rehberg and Richman, 1989). Conversely, families with paid domestic help demand substantially less prosocial behavior from their children (Klein et al., 2009). This suggests that social class,

although outside the purview of this article, is another important factor to consider in children's prosocial behavior.

Sharing

Rochat et al. (2009) examined sharing in children in five cultures (the US, China, Peru, Fiji, and Brazil), and sampled different social classes in Brazil. Children were asked to distribute candies and snacks between themselves and an experimenter over several trials. In one trial, children were asked to divvy up a 'perfect share' after which the experimenter would then decide the recipient. Across cultures, 3-year-old children were more likely to self-hoard (give all candies to oneself), but also to 'other-hoard' (giving all candies to the other), suggesting that young children may have had a different understanding of the task. In contrast, 5 year olds generally distributed candies more equally. At a cultural level, poor children from Recife in Brazil were the most likely to self-hoard, and Peruvian and Fijian children the most likely to other-hoard and also more likely to distribute fairly. US children were as likely to self-maximize their share of candy as some of the unschooled poor children.

Caring for Others

De Guzman et al. (2005), in a naturalistic study conducted in the rural Philippines and a small town in the United States, found that children in both settings were more likely to act prosocially toward infant relatives than toward child and adult relatives. However, children in the United States were more likely to act prosocially toward nonkin children, whereas children from the Philippines were more likely to help adult and infant relatives. The authors remark that their findings reflect in part vastly different social ecologies, particularly as older American children go to school, and thereby come into contact with far more same-age peers and are in infrequent contact with infants.

In an experimental study, Trommsdorff et al. (2007) found that when German, Israeli, Malaysian, and Indonesian children were playing with an adult whose balloon popped, it was the German and Israeli children who were the most likely to offer to comfort the adult. However, other-focused distress after the adult lost the balloon was high in all cultural groups. The experimenters hypothesized that in the Asian cultures face-saving may have been a factor in why the children did not offer comfort and care to the adult.

Modesty and Politeness

In research related to the concept of prosocial lying, Heyman et al. (2011) examined how children in Japan and America evaluated a child who committed a positive deed, either in public or in private. Japanese children evaluated truth-telling less positively than American children, and lying about one's good deeds less negatively, particularly when the question about the deed was asked in public. Ma et al. (2011) found similar patterns in Chinese children. Although both studies reflected children's judgments about a hypothetical situation rather than observed prosocial behavior, the researchers attribute this pattern of findings to a cultural norm of modesty in East Asian cultures.

Race and Ethnicity

In addition to class, race and ethnicity can contribute to subcultures within larger cultures. Spivak and Howes (2011) found that African-American children were slightly more likely to be prosocial than White or Latino children. Harris and Graham (2007) note that in some African-American neighborhoods values such as respect may emerge as prosocial, and that there may be fewer differences in prosocial behavior across racial and ethnic groups than within one's in-group.

Conclusions and Future Directions

From current studies of children's prosocial behavior, it seems clear that many prosocial behaviors occur at different rates across cultures, at least at an aggregate level. However, the data do not suggest that these cultural differences are absolute, such that some prosocial behavior in one culture is considered wholly wrong to members of another culture. It is also unclear how to group cultural differences in prosocial behavior, as there is often a great deal of variation even within supposedly similar societies (Rogoff, 2003). In the study by Heyman et al. (2011) and others, Asian cultures contrast with the US in terms of modesty. However, in the study of sharing by Rochat et al. (2009), Peruvian children seem more similar to children from Fiji than to children in neighboring Brazil, who behave more similarly to American children.

Although larger theoretical ideas about cultural influences underlie these findings, clearly a great deal more research needs to be conducted on prosocial behavior and development cross-culturally, including assessing children on multiple forms of prosocial behavior, before more decisive conclusions can be drawn, and this research must also consider social class, and race and ethnicity as part of the cultural fabric, particularly in heterogeneous societies.

In addition to longitudinal studies on prosocial development, current research on children's prosocial behavior, which is generally restricted to relatively discrete episodes, such as a child helping a parent or peer in distress, should also focus on children's longer-term prosocial commitments, which may emerge in new social and cultural contexts such as schools, community centers, and even the Internet.

See also: Eisenberg's Theory of Prosocial Reasoning; Empathy During Early Childhood Across Cultures, Development of; Moral Development, Cultural Differences In; Mother-Child Communication: Cultural Differences; Parenting Attitudes and Beliefs across Cultures; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Self-Development in Childhood and Adolescence.

Bibliography

Bierhoff, H.W., 2002. Prosocial Behaviour. Psychology Press, New York.

Brownell, C.A., Zerwas, S., Balaram, G., 2002. Peers, cooperative play, and the development of empathy in children. Behavioral and Brain Sciences 25 (1), 28–29.

Carpendale, J.I.M., 2009. Piaget's theory of moral development. In: Müller, U., Carpendale, J.I.M., Smith, L. (Eds.), The Cambridge Companion to Piaget. Cambridge University Press, Cambridge, pp. 270–286.

De Guzman, M.R.T., Edwards, C.P., Carlo, G., 2005. Prosocial behaviors in context. Applied Developmental Psychology 26, 542–558.

Eisenberg, N., Spinrad, T.L., Morris, A., 2013. Empathy-related responding in children. In: Killen, M., Smetana, J.G. (Eds.), Handbook of Moral Development, second ed. Psychology Press, New York, pp. 184–207.

Eisenberg, N., Spinrad, T.L., Knafo, A. Prosocial Development. In: Lerner, R.M. (Ed.), Handbook of Child Psychology and Developmental Science, seventh ed. Wiley, New York, in press.

Greenfield, P.M., Keller, H., Fuligni, A., Maynard, A., 2003. Cultural pathways through universal development. Annual Review of Psychology 54 (1), 461–490.

Grusec, J.E., Chaparro, M.P., Johnston, M., Sherman, A., 2013. The development of moral behaviour and conscience from a socialization perspective. In: Killen, M., Smetana, J.G. (Eds.), Handbook of Moral Development, second ed. Psychology Press, New York, pp. 113–134.

Harris, Y.R., Graham, James A., 2007. The African American Child: Development and Challenges. Springer, New York.

Hay, D.F., Cook, K.V., 2007. The transformation of prosocial behavior from infancy to childhood. In: Brownell, C.A., Kopp, C.B. (Eds.), Socioemotional Development in the Toddler Years. Guilford, New York, pp. 100–131.

Heyman, G.D., Itakura, S., Lee, K., 2011. Japanese and American children's reasoning about accepting credit for prosocial behavior. Social Development 20 (1), 171–184.

Hoffman, M.L., 2000. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, New York.

Jennings, P.A., Greenberg, M.T., 2009. The prosocial classroom: teacher social and emotional competence in relation to student and classroom outcomes. Review of Educational Research 79 (1), 491–525.

Kärtner, J., Keller, H., Chaudhary, N., 2010. Cognitive and social influences on early prosocial behavior in two sociocultural contexts. Developmental Psychology 46 (4), 905–914.

Klein, W., Graesch, A.P., Izquierdo, C., 2009. Children and chores: a mixed-methods study of children's household work in Los Angeles families. Anthropology of Work Review 30 (3), 98–109.

Lapsley, D.K., 1996. Moral Psychology. Westview Press, Boulder, Colo.

Ma, F., Xu, F., Heyman, G.D., Lee, K., 2011. Chinese children's evaluations of white lies: weighing the consequences for recipients. Journal of Experimental Child Psychology 108 (2), 308–321.

Newcomb, A.F., Bagwell, C.L., 1995. Children's friendship relations: a meta-analytic review. Psychological Bulletin 117 (2), 306–347.

Penner, L.A., Dovidio, J.F., Piliavin, J.A., Schroeder, D.A., 2005. Prosocial behavior: multilevel perspectives. Annual Review of Psychology 56 (1), 365–392.

Persson, G.E.B., 2005. Developmental perspectives on prosocial and aggressive motives in preschoolers' peer interactions. International Journal of Behavioral Development 29 (1), 80–91.

Piaget, J., 1965. The Moral Judgment of the Child. Free Press, New York.

Rehberg, H.R., Richman, C.L., 1989. Prosocial behaviour in preschool children: a look at the interaction of race, gender, and family composition. International Journal of Behavioral Development 12 (3), 385–401.

Rheingold, H.L., 1982. Little children's participation in the work of adults, a nascent prosocial behavior. Child Development 53 (1), 114–125.

Rochat, P., Dias, D.G., Liping, G., Broesch, T., Passos-Ferreira, C., Winning, A., Berg, B., 2009. Fairness in distributive justice by 3- and 5-year-olds across seven cultures. Journal of Cross-Cultural Psychology 40 (3), 416–442.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, Incorporated, New York.

Spivak, A.L., Howes, C., 2011. Social and relational factors in early education and prosocial actions of children of diverse ethnocultural communities. Merrill-Palmer Quarterly 57 (1), 1–24.

Tomasello, M., Vaish, A., 2013. Origins of human cooperation and morality. Annual Review of Psychology 64, 231–255.

Trommsdorff, G., Friedlmeier, W., Mayer, B., 2007. Sympathy, distress, and prosocial behavior of preschool children in four cultures. International Journal of Behavioral Development 31 (3), 284–293.

Williams, S.M., Kirmayer, M., Simon, T., Talwar, V., 2013. Children's antisocial and prosocial lies to familiar and unfamiliar adults. Infant and Child Development 22 (4), 430–438.

Zahn-Waxler, C., Radke-Yarrow, M., King, R.A., 1979. Child rearing and children's prosocial initiations toward victims of distress. Child Development 50 (2), 319–330.

Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations

Jesse Drummond, Whitney E Waugh, Stuart I Hammond, and Celia A Brownell, University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In their first years of life, infants exhibit a wide range of prosocial behaviors, including helping, sharing, comforting, and cooperating. Over the course of infancy, these behaviors become more sophisticated and socially appropriate. Although consistent age trends are demonstrated, prosocial behavior in infancy is best considered a multidimensional category consisting of distinct behaviors rather than a single unified construct. Moreover, the expression of prosocial behavior in infancy and its development are moderated by context, type of prosocial behavior, relationships with the others involved, and culture. Future research is needed to examine the effects of each of these factors on development and individual differences in early emerging prosocial behavior.

Prosocial Behavior in Infancy

Prosocial behavior in infancy is distinguished by great diversity and marked developmental change. The earliest manifestations of prosocial behaviors, traditionally defined as any intentional action that benefits another person (e.g., Grusec et al., 2002; Eisenberg et al., 2006), appear near the end of the first year of life (Hay and Cook, 2007), although some nascent forms of prosociality may appear earlier. These acts become progressively more coordinated, socially skilled, and often more frequent by the end of the third year (Hay and Cook, 2007) in concert with developing social, cognitive, language, and motor abilities (Moore, 2007).

Initial work on infants' prosociality focused on empathic responding, including both concerned attention and attempts to understand another's distress, and efforts to comfort the distressed other (e.g., Zahn-Waxler et al., 1979; Zahn-Waxler et al., 1992; Dunn, 1988; Rheingold, 1982; see Eisenberg et al., 2006 for a review). In the decades that followed these seminal works, researchers have uncovered a wide range of prosocial behaviors that infants exhibit in varied contexts and with different people (e.g., parents, siblings, peers, unfamiliar adults). Included on this spectrum are helping, sharing, cooperating, and comforting, as well as less prototypical behaviors such as informative pointing and participating in household chores. These behaviors are ubiquitous, often spontaneous, and enthusiastic.

In contrast to the broad picture of infant prosociality painted by this earlier research, recent investigations have explored its more fine-grained aspects and revealed some puzzling patterns. First, among toddlers, different prosocial behaviors are often uncorrelated with one another (Dunfield et al., 2011; Hay and Cook, 2007), suggesting that young children are not uniformly prosocial. Moreover, early prosocial behavior has been positively associated with aggression (Hay, 2006), and young children frequently ignore another's distress or exacerbate the victim's plight rather than exhibit concern or offer comfort (Dunn, 1988). These findings stand in stark contrast to the notion that prosocial behavior is a stable, early appearing, unified construct that governs much of early social interaction.

Knowledge of cultural variations in infant prosocial responding may contribute to our understanding of its development as well as its diversity. Although few in number, extant cross-cultural examinations of infant behavior reveal cultural differences in rates of prosocial responding in infants (Callaghan et al., 2011), how parents socialize prosociality in their infants (Mosier and Rogoff, 2003), and the mechanisms that motivate prosocial behavior (Kärtner et al., 2010).

In the following sections, we first describe the different types of prosocial behaviors children exhibit in infancy and how they change over the first few years of life. We then reflect on the developmental trends and multidimensionality of prosocial behavior. Finally, we consider cultural variations in early prosocial responding.

Describing Prosocial Behavior

In describing early developments in prosociality, we group behaviors into the three domains identified by Hay and Cook (2007): *feeling for others*, *ministering to others*, and *working and playing with others*. These domains, while not completely orthogonal, show some evidence of distinct patterns of internal coherence and provide a useful first step in understanding prosocial development.

Feeling for Others

The domain with the longest philosophical tradition, rooted in the theories of empathy and sympathy from thinkers such as Adam Smith, is feeling for others. Defined as "an affective response more appropriate to another's situation than one's own" (Hoffman, 2000: p. 4), empathy has long been considered a potent motivator of prosocial behavior and a necessary component of prosocial development in adults and children alike (Batson, 1991; Eisenberg et al., 2006). According to Hoffman, an infant's ability to empathize with another first manifests in reactive contagious crying in the first few days of life and gradually develops into veridical empathic concern in the third year. At this latter age, a child can empathize with another in distress, take another's perspective, and reflect on

others' needs. Although the demonstration of empathic concern for another in distress has itself been considered a prosocial response, the process of empathizing with another may act as a mechanism for all kinds of prosocial behaviors (Davis, 2005).

Empathic concern is traditionally operationalized as facial and/or vocal expressions of concern, including sad or worried looks (e.g., plaintive face with furrowed brow and downturned mouth) and sympathetic statements (e.g., "I'm sorry") (Zahn-Waxler et al., 1992). In a longitudinal examination of infants' empathic concern in response to naturally occurring and simulated distresses in the home, Zahn-Waxler and colleagues found that infants aged 13–15 months showed very low levels of empathic concern, whether they had caused the distress or not; by 23–25 months, infants were showing empathic concern significantly more frequently. This pattern of findings suggests that infants are capable of feeling for another at the beginning of their second year, although they rarely act on this ability. Their propensity to empathize with another, or at least their demonstration of this propensity, grows significantly with age.

Recent work has experimentally examined situational differences in children's empathic responding. Spinrad and Stifter (2006) examined 18-month olds' concerned attention to their mother, a stranger, and a crying baby doll. Infants showed more concerned attention toward the baby than toward the stranger or the mother, and more concerned attention toward the stranger than toward the mother. Nichols et al. (2009) showed similar situational differences: infants between 12 and 24 months of age showed significantly more empathic concern to a crying baby than to their mother.

In sum, feeling for others, measured by empathic concern in response to others' distress, is evident early in the second year and becomes more prevalent as children reach the third year. Infants exhibit empathic concern toward many different others, while also showing situational specificity in their empathic responding. Somewhat paradoxically, however, the ability to feel for others may also result in personal distress in response to another's distress, which is quite prevalent in young infants and may inhibit some of the more overt prosocial behaviors described below (Eisenberg et al., 2006).

Ministering to Others

In the second year, infants respond to the needs of others in specific goal-directed and instrumental ways that reveal some level of understanding of the other's current situation and need (Hay and Cook, 2007). This domain of actions, referred to as ministering to others, includes comforting, nurturing, and helping.

Infants show a remarkable ability to comfort others in distress, even at very young ages. They will hug distressed victims, bring them objects to soothe their distress, find someone else to help them, or give them physical assistance (Dunn, 1988; Vaish et al., 2009; Zahn-Waxler et al., 1979). In concert with developments in empathic concern, comforting behaviors grow in frequency and complexity over the second and third years. At the beginning of their second year, infants tend to comfort others infrequently and use predominantly physical means (e.g., hugging or patting; Zahn-Waxler et al., 1992). At this age they also tend to inappropriately overgeneralize specific actions to new and different situations; for example, a child who pats his mother on the head after she has hit it on a cabinet and then does the same to a sibling who has fallen and scraped his knee (Dunn, 1988). In their third year, toddlers comfort distressed others more frequently and do so in more elaborate ways, including using distraction or humor and protecting the other from further harm (Zahn-Waxler et al., 1992; Dunn, 1988). Toddlers at this age also tend to persist in their comforting efforts, trying a number of different comforting techniques (Dunn, 1988). Importantly, they also begin to respond to conditions of need even in the absence of overt emotional distress (Dunn, 1988; Vaish et al., 2009).

Many of the comforting behaviors captured in these studies occurred in response to a younger sibling in distress (Dunn, 1988). This finding, when coupled with those described above with respect to concerned attention to a crying baby, suggests that infants in their second year are able, and at times willing, to nurture others. Although it has been documented in many different societies (Hay and Cook, 2007), nurturing as an early form of prosocial responding has yet to be studied directly.

Infants also show a propensity to help others. An early form of helping involves completing goal-directed actions to remedy another's specific problem (e.g., fetching an out-of-reach object that an adult needs to finish a task). This 'instrumental helping' behavior emerges just after the first birthday at about 14 months of age, consistent with an infant's developing ability to understanding and act upon an adult's desire or need (Warneken and Tomasello, 2007; Svetlova et al., 2010). By midway through their second year, infants show similar goal-directed helping in more difficult tasks, such as opening a cabinet for an adult who is holding several heavy books or helping him stack objects (Warneken and Tomasello, 2006). Moving into the third year, infants show an increasing propensity to help in these goal-directed ways. When given the opportunity to fetch out-of-reach objects, 24- (Dunfield et al., 2011) and 30-month-old infants (Svetlova et al., 2010) help significantly more frequently than 18-month-olds. Indeed, in a recent analysis of helping behavior in infancy, almost all 30-month-old infants helped on every occasion while less than half the 18-month-old infants did so, illustrating the growing ability to understand and alleviate another's need (Svetlova et al., 2010).

However, infants find more complex helping scenarios challenging. When an adult needs an object to alleviate a more abstract emotional need (e.g., a toy to make him happy after he begins to cry), infants up to 30 months of age help significantly less frequently than they do when the adult needs an object to complete a concrete goal-directed action (e.g., a clothespin to clip a cloth to a clothesline); their performance decreases still more when it is the infant's own toy that would alleviate the adult's emotional need (Svetlova et al., 2010). Nevertheless, 30-month-old infants show a much greater ability to help a sad adult by bringing him a toy than do their 18-month-old counterparts (Svetlova et al., 2010), illustrating their greater perspective-taking ability and emotion understanding.

In sum, early in their second year, infants begin to comfort and help others in distress. This ministering behavior becomes

more frequent, varied, and appropriate over time. By the end of the third year of life, infants can quite readily grasp the needs of another in distress, whether action-based or emotional, and aid in addressing those needs.

Working and Playing with Others

In addition to feeling for and ministering to others, infants also begin to work and play cooperatively with others. They explore their social world, participating in and often facilitating social interaction with others. Beginning in the second year of life, they spontaneously cooperate in joint actions, help others with chores, and share what they have and what they know with those around them (Brownell, 2011; Hay and Cook, 2007).

In everyday interactions with siblings, peers, and parents, infants are spontaneous and enthusiastic cooperators. Infants will eagerly fetch a car that has run off the track so her sibling can run it down the track again, or bring her sibling more puppets to play with if he is putting on a puppet show (Dunn, 1988). These actions are often unprompted and coupled with positive affect and joy (Dunn, 1988). Such cooperative exchanges have been reported in infants as young as 12 months, when they mostly consist of showing and handing interesting objects to parents (Hay, 1979); by 14 months some infants will reengage an adult who has interrupted a joint activity, illustrating their enjoyment of joint participation and possibly demonstrating rudimentary understanding of cooperation toward a joint goal (Warneken and Tomasello, 2007). As children develop through the second and into the third year, these cooperative interactions become more frequent and more explicitly coordinated (Dunn and Munn, 1986; Dunn, 1988; Warneken et al., 2006; Brownell et al., 2006).

Beginning in the second year, infants also begin to participate with their parents in routine household tasks. In a seminal study, Rheingold (1982) found that most infants between 18 and 30 months of age eagerly began to participate together with their parents, and in some circumstances an unknown adult, in a series of household tasks designed to resemble chores that parents do around the house (e.g., sweeping, setting the table). Toddlers often began to participate as soon as the parent stated her intent and did so with enthusiasm and positive affect. They frequently coupled their actions with vocalizations reflecting an understanding of the goal, such as those stating an intention (e.g., 'I can do messy papers') and signaling completion (e.g., 'all clean') (Rheingold, 1982: p. 119). The frequency of these behaviors and vocalizations increased with age, and by 30 months many of the infants spontaneously used the word 'help' to describe their participation. While the intention to help is clear, proficiency is another matter; parents report that their child's participation is often unhelpful or even counterproductive, and some parents will refrain from doing chores until their child is asleep.

In addition to participating in these joint activities with others, infants begin to share with others during the second year. This is first evident in infants' informative pointing, which emerges around the first birthday or soon thereafter. For example, they share information they themselves have about the location of an object that another has lost and is seeking (Lizkowski et al., 2008). In the middle of the second year, infants begin to share tangible resources such as toys or food with others. These early sharing behaviors require considerable communicative support about the potential recipient's need, and young infants often share only after a direct request, but with age they share more quickly (Brownell et al., 2013; Dunfield et al., 2011). Interestingly, other analyses of sharing behavior have found that the frequency of sharing may level off or even decrease over the preschool years as children become more selective in choosing when and with whom to share (Hay and Cook, 2007).

In sum, starting around their first birthday, infants begin to cooperate with others, participate with others in everyday tasks, and share both information and material resources with others. These prosocial behaviors gradually become more spontaneous, frequent, and sophisticated over the course of infancy, preparing infants to enter childhood armed with a considerable array of prosocial skills and grounding the further development of social competence over childhood and adolescence.

Multidimensionality of Prosocial Behavior

Despite general patterns of growth common to feeling for others, ministering to others, and working and playing with others, prosocial behavior is not a single unified construct, nor is it motivated by the same underlying altruistic tendencies; indeed, different prosocial behaviors have distinct correlates (Dunfield et al., 2011), and evidence for relationships among different forms of prosocial behavior is mixed. Goal-directed helping in the second year is generally unrelated to other forms of prosocial behavior, such as sharing (Dunfield et al., 2011) and cooperating (Warneken and Tomasello, 2007). Concerned attention to another in distress is significantly related to prosocial interventions in some situations, but not in others (Zahn-Waxler et al., 1992; Spinrad and Stifter, 2006). Similarly, certain prosocial behaviors show consistency across some contexts, but not others (Nichols et al., 2009; Spinrad and Stifter, 2006). Recent findings have also revealed distinct neurocognitive mechanisms for different prosocial behaviors: Paulus et al. (2013) found that instrumental helping was associated with greater right than left cortical temporal activity, while comforting was associated with greater left than right frontal activity.

In addition, prosocial behavior is often exhibited concurrently with antagonistic behaviors. In the first year, sharing is positively correlated with aggression (Hay, 2006; Hay and Cook, 2007). Children at this age can be simultaneously comforting and violent, as in the case of a young girl who upon seeing a friend cry after his ball was taken by another child, bit the child who took the ball and returned it to her friend (Dunn, 1988). The victim in distress is not immune from these inconsistencies, as illustrated by a 30-month-old who patted his crying brother, tried to distract him, spoke affectionately to him, and, when all bids failed, threatened to hit him if he did not stop crying (Dunn, 1988). Prosocial behavior also occurs at fairly low rates, and children quite routinely respond to distress by ignoring the victim, laughing at the victim, or even exacerbating the victim's distress (Hay and Cook, 2007; Dunn, 1988).

Prosocial behavior, then, is best understood as a multidimensional construct consisting of distinct behaviors that vary

across different situations and that may be governed by distinct mechanisms. As a corollary, the findings suggest that the same behavior may be driven by different mechanisms for different infants (Kärtner et al., 2010): for one child, the ability or propensity to empathize with an adult's perceived distress may motivate his retrieving of a dropped object; for another, a strong desire to affiliate with others may be the driver. Complicating matters is that these motivations are not mutually exclusive and may operate simultaneously within any one child.

So how do we make sense of prosocial behavior in infancy? The first step may be to acknowledge that what unifies these behaviors is their social orientation. As described by Hay and Cook (2007), "early prosocial tendencies may best be seen as a basic impulse to engage positively with other people that only gradually becomes *selective*, *socially appropriate*, *self-regulated*, and *morally informed* activity" (original emphasis) (p. 102). Future work is needed to establish coherent patterns among these different behaviors at different ages and to examine how this gradual change occurs.

Cultural Variation in Prosocial Responding

There has been relatively little research devoted to cultural differences in prosocial behavior across infancy and early childhood, in part due to prevailing theory that early prosocial behavior is universal and evolutionarily selected, and thus culturally invariant (e.g., Warneken and Tomasello, 2006). Consistent with this view is the assertion that prosocial behavior only becomes rule-governed as children leave the infancy period, suggesting that cultural norms only exert their influence later in development as older children become more directly engaged in cultural institutions (Hay and Cook, 2007).

Still, there is reason to believe that early prosocial behavior may vary by culture. For example, Canadian and Peruvian infants, aged 10–13 months, direct an adult's attention toward interesting stimuli more frequently than age-matched Indian infants; 2-year-old infants in Canada exhibit higher rates of instrumental goal-based helping with adults than do infants in India and Peru; and infants in India and Peru collaborate more frequently with an adult than do infants in Canada (Callaghan et al., 2011).

Parenting styles in both play and disciplinary contexts may influence the transmission of prosocial behavior and values beginning in infancy (Hay and Cook, 2007; Pettygrove et al., 2013). Empathic caregiving, for example, has been associated with prosocial responses to distress in the second year (Zahn-Waxler et al., 1979) and certain parenting techniques (e.g., guilt induction) are considered crucial for early prosocial and empathic development (Hoffman, 2000). An infant's sense of responsibility and agency in moments of distress may be reduced by frequent parent interventions (Hay and Cook, 2007). These findings become culturally relevant when we consider that parenting styles, values, and techniques vary dramatically by culture.

The consequences of parenting style for prosocial development in distinct cultures are illustrated by the work of Mosier and Rogoff (2003) with American and Guatemalan families. When conflict arose between infants in their second year and their older siblings, Guatemalan mothers tended to bestow a privileged status upon their infants, allowing them to do whatever they pleased; in contrast, American mothers expected their infants to abide by the same rules as their older siblings. Guatemalan mothers rarely, if ever, forced their infant to cooperate, instead attempting to persuade and ultimately acquiescing to the infant's desire. This privileged status of Guatemalan infants was also endorsed by their older siblings, aged 3 and 5 years, even when the mother was absent. Accordingly, more infant–sibling interactions among Guatemalan children included offering objects, helping others with objects, and jointly working with objects, whereas more infant–sibling interactions among American children included conflict over objects.

Mirroring this finding in other cultures, Kärtner et al. (2010) found that parents in Delhi, India, rated obedience-related socialization goals (e.g., learning to obey parents and older people) as more important than did parents from Berlin, Germany. The proportions of children showing prosocial and tentative prosocial responses varied marginally between cultures, but in both cultures the degree to which parents emphasized obedience-related socialization goals was associated with their infant's prosocial responses to distress. In addition, Kärtner et al. (2010) found that mirror self-recognition, a measure of self-other differentiation, was related to prosocial behavior only in the Berlin sample. That infants in Delhi responded prosocially whether or not they possessed a theoretically important socio-cognitive prerequisite illustrates that different developmental pathways, some emotion-based and others action-based, may exist for the same behaviors (Kärtner et al., 2010; also see Svetlova et al., 2010). Thus, mechanisms of prosocial development apart from parenting may also vary by culture.

Conclusions and Future Directions

In their first few years, infants develop a remarkable capacity for acting in myriad ways to benefit others. These behaviors are often unrelated to one another and show relatively little consistency across different types of tasks. Further, the expression of prosocial behavior in infancy, and its development over time, is moderated by context, type of prosocial behavior, relationships with the others involved, and culture. While the motivations for prosocial behavior in such young children are still largely unknown, and may be multiple and overlapping, the role of social interaction in its development and expression is unquestionable (Hay and Cook, 2007).

Research findings imply that prosocial behavior is multidimensional and includes different elements that may be qualitatively distinct. The influences of corollary factors on prosocial development in infancy and early childhood, when prosocial behaviors are first emerging, warrant further examination. The role of culture is especially elusive given the relative paucity of investigations of cultural differences in early prosocial responding, but the recent finding of culturally variant pathways to prosocial development (Kärtner et al., 2010), offers a promising new perspective on the contributors to developmental change.

See also: Early Emotional Development and Cultural Variability; Eisenberg's Theory of Prosocial Reasoning; Emotional Development, Effects of Parenting and Family Structure on; Empathy During Early Childhood Across Cultures, Development of; Infancy and Human Development; Moral Development, Cultural Differences In; Moral Development, Theories of; Prosocial Behavior During Childhood and Cultural Variations; Prosocial Behavior, Effects of Parenting and Family Structure on; Self in Culture: Early Development; Social and Emotional Development in the Context of the Family.

Bibliography

Batson, C.D., 1991. The Altruism Question: Toward a Social-Psychological Answer. Erlbaum, Hillsdale, NJ.

Brownell, C.A., Ramani, G.B., Zerwas, S., 2006. Becoming a social partner with peers: cooperation and social understanding in one- and two-year-olds. Child Development 77 (4), 803–821.

Brownell, C.A., 2011. Early developments in joint action. Review of Philosophy & Psychology 2, 193–211.

Brownell, C.A., Svetlova, M., Anderson, R., Nichols, S., Drummond, J., 2013. Socialization of early prosocial behavior: parents' talk about emotions is associated with sharing and helping in toddlers. Infancy 18 (1), 91–119.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., Tomasello, M., 2011. Early social cognition in three cultural contexts. Monographs of the Society for Research in Child Development 76 (2), 1–142.

Davis, M.H., 2005. A 'constituent' approach to the study of perspective taking. In: Malle, B.F., Hodges, S.D. (Eds.), Other Minds. Guilford Press, New York, pp. 44–55.

Dunfield, K., Kuhlmeier, V.A., O'Connell, L., Kelley, E., 2011. Examining the diversity of prosocial behavior: helping, sharing, and comforting in infancy. Infancy 16 (3), 227–247.

Dunn, J., Munn, P., 1986. Siblings and the development of prosocial behavior. International Journal of Behavioral Development 9, 265–284.

Dunn, J., 1988. The Beginnings of Social Understanding. Harvard University Press, Cambridge, MA.

Eisenberg, N., Fabes, R.A., Spinrad, T.L., 2006. Prosocial development. In: Eisenberg, N. (Ed.), Handbook of Child Psychology, Social, emotional, and personality development, vol. 3. Wiley, Hoboken, NJ, pp. 646–718.

Grusec, J.E., Davidov, M., Lundell, L., 2002. Prosocial and helping behavior. In: Smith, P.K., Hart, C.H. (Eds.), Blackwell Handbook of Childhood Social Development. Blackwell Publishing, Malden, MA, pp. 457–474.

Hay, D.F., 1979. Cooperative interactions and sharing between young children and their parents. Developmental Psychology 15 (6), 647–653.

Hay, D.F., 2006. Yours and mine: toddler's talk about possessions with familiar peers. The British Journal of Developmental Psychology 24, 39–52.

Hay, D.F., Cook, K.V., 2007. The transformation of prosocial behavior from infancy to childhood. In: Brownell, C.A., Kopp, C.B. (Eds.), Socioemotional Development in the Toddler Years. Guilford, New York, pp. 100–131.

Hoffman, M.L., 2000. Empathy and Moral Development: Implications for Caring and Justice. Cambridge University Press, New York.

Kärtner, J., Keller, H., Chaudhary, N., 2010. Cognitive and social influences on early prosocial behavior in two sociocultural contexts. Developmental Psychology 46 (4), 905–914.

Lizkowski, U., Carpenter, M., Tomasello, M., 2008. Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. Cognition 108, 732–739.

Moore, C., 2007. Understanding self and others in the second year. In: Brownell, C.A., Kopp, C.B. (Eds.), Socioemotional Development in the Toddler Years. Guilford, New York, pp. 43–65.

Mosier, C.E., Rogoff, B., 2003. Privileged treatment of toddlers: cultural aspects of individual choice and responsibility. Developmental Psychology 39 (6), 1047–1060.

Nichols, S.R., Svetlova, M., Brownell, C.A., 2009. The role of social understanding and empathic disposition in young children's responsiveness to distress in parents and peers. Cognition, Brain, Behavior 8 (4), 449–478.

Paulus, M., Kühn-Popp, Licata, M., Sodian, B., Meinhardt, J., 2013. Neural correlates of prosocial behavior in infancy: different neurophysiological mechanisms support the emergence of helping and comforting. NeuroImage 66, 522–530.

Pettygrove, D.M., Hammond, S.I., Karahuta, E.L., Waugh, W.E., Brownell, C.A., 2013. From cleaning up to helping out: parental socialization and children's early prosocial behavior. Infant Behavior and Development 36, 843–846.

Rheingold, H.L., 1982. Little children's participation in the work of adults, a nascent prosocial behavior. Child Development 53 (1), 114–125.

Spinrad, T.L., Stifter, C.A., 2006. Toddlers' empathy-related responding to distress: prediction from negative emotionality and maternal behavior in infancy. Infancy 10 (2), 97–121.

Svetlova, M., Nichols, S.R., Brownell, C.A., 2010. Toddlers' prosocial behavior: from instrumental to empathic to altruistic helping. Child Development 81 (6), 1814–1827.

Vaish, A., Carpenter, M., Tomasello, M., 2009. Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. Developmental Psychology 45 (2), 534–543.

Warneken, F., Chen, F., Tomasello, M., 2006. Cooperative activities in young children and chimpanzees. Child Development 77 (3), 640–663.

Warneken, F., Tomasello, M., 2006. Altruistic helping in human infants and young chimpanzees. Science 311 (5765), 1301–1303.

Warneken, F., Tomasello, M., 2007. Helping and cooperation at 14 months of age. Infancy 11 (3), 271–294.

Zahn-Waxler, C., Radke-Yarrow, M., King, R.A., 1979. Child rearing and children's prosocial initiations toward victims of distress. Child Development 50, 319–330.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., Chapman, M., 1992. Development of concern for others. Developmental Psychology 28 (1), 126–136.

Prosocial Behavior, Cultural Differences in

Daniel J Hruschka, School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA
Joseph Henrich, University of British Columbia, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Humans frequently act in ways that benefit others or contribute to their society. Despite this common human capacity for prosocial behavior, populations can differ dramatically in the propensity to engage in various kinds of helping, sharing, and cooperation. In this article, we review empirical findings of cross-population diversity in prosocial behavior, outline current theories for these population differences, and identify future areas for research.

Introduction

The odds of a stranger coming to your aid on the street depend in important ways on the location of that street. In one international study, residents of Rio de Janeiro and Lilongwe (Malawi) nearly always helped strangers – assisting blind people at a street corner, alerting strangers to dropped pens, and helping handicapped pedestrians gather their dropped belongings (86–93% of the time). Meanwhile, residents of New York City and Kuala Lampur helped strangers in these same tasks less than half the time (40–45%) (Levine et al., 2001). Researchers have observed comparable cross-population variation in other prosocial behaviors, including how people voluntarily share a pot of money with an anonymous other (Henrich et al., 2005, 2010a), contribute to a public good (Buchan et al., 2009; Herrmann et al., 2008; Fearon et al., 2009), punish norm violators (Henrich et al., 2006), follow an impartial rule, and help a friend in the face of countervailing social norms (Trompenaars and Hampden-Turner, 1997). In many cases, educated members of western industrialized countries are at the far end of the human spectrum (Henrich et al., 2010b).

Scholars have formulated a number of potential explanations for cross-population variation in prosocial behavior, including population size and density, market integration, religious affiliation, connection with the global community, quality of institutions, and pathogen stress. In this article, we review the existing evidence from psychology, anthropology, and economics for cross-population variation in how people engage in prosocial behavior. We then outline prominent hypotheses for these differences, as well as recent attempts to test these hypotheses.

Operationalizing Prosocial Behavior for Cross-Population Studies

Intuitively, prosocial behavior can include donating money to charity, volunteering time at a soup kitchen, helping a friend with homework, following rules of good behavior, protecting the weak, and keeping the peace. Psychology has a rich tradition of studying prosocial behavior, and researchers have proposed a diverse set of definitions for the concept (Batson and Powell, 2003; Eisenberg et al., 2007; Penner et al., 2006). Most definitions concur that a prosocial act benefits others, but they differ in more fine-grained requirements about intention, motivation, and the cost to the actor. Was the benefit intended? Was the behavior voluntary or intrinsically motivated? Did the actor incur a cost? Here, we consider prosocial behavior in its broadest sense – acts that benefit others – understanding that it may be important to consider these more fine-grained questions when studying any specific behavior.

Within this broad set of behaviors, Batson also outlines three major classes: (1) aiding another person, (2) contributing to a group, and (3) following a socially valued rule, such as being honest or respecting another's property (Batson and Powell, 2003). Researchers have used several methods for assessing and comparing these different kinds of prosocial behaviors across populations, including asking people how they would respond in hypothetical scenarios (Trompenaars and Hampden-Turner, 1997; Gelfand et al., 2011), examining allocations of real money and other goods in behavioral experiments (Henrich et al., 2005, 2010a; Buchan et al., 2009; Herrmann et al., 2008; Rustagi et al., 2010), and observing how people help when confronted with experimenter-created situations in the real world (Levine et al., 2001). In this article, we focus primarily on studies of real rather than hypothetical behavior, which are currently limited to a few kinds of prosocial acts: (1) helping a stranger (Levine et al., 2001), (2) sharing a pot of money or goods with an anonymous stranger (Henrich et al., 2005, 2010a), (3) sacrificing to punish a cheater (Henrich et al., 2010a, 2006), (4) contributing to a public good (Herrmann et al., 2008; Rustagi et al., 2010; Buchan et al., 2009), (5) following an explicit rule for allocating goods (Hruschka et al., 2014), and (6) reciprocating a transfer of goods (Vollan, 2012).

The Scale of Cross-Population Variation

Population differences in prosocial behavior emerge at many scales – between communities from different countries (Levine et al., 2001; Herrmann et al., 2008), between urban and rural populations in a country (Steblay, 1987), between cities within countries (Levine et al., 1994), between communities from different small-scale societies (Henrich et al., 2005,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24103-2

2010a), and between communities from the same ethno-linguistic group (Henrich et al., 2005, 2010a; Lamba and Mace, 2011; Rustagi et al., 2010; Vollan, 2012; Fearon et al., 2009). Researchers frequently refer to variation between ethno-linguistic groups or between countries as 'cross-cultural,' a habit which implies that such variation is due to variation in culture – socially learned modes of thought and behavior. However, it is entirely possible that such cross-population differences result instead from common responses to a shared ecological or social environment or even to genetic differences (Eisenberg and Hayes, 2011; Chiao and Blizinsky, 2010). Conversely, variation within ethno-linguistic units has been inferred to result from environmental or ecological factors since culture must be constant within such groups (Lamba and Mace, 2011). This too is a premature inference (Henrich et al., 2012), since communities within the same ethno-linguistic unit can also differ in terms of socially learned modes of thought and behavior. Modern evolutionary approaches to 'culture' have long dispensed with essentializing notions that attribute 'cultures' to entire ethno-linguistic groups. For these reasons, we describe observed differences simply as 'cross-population differences' until the underlying causes and mechanisms are identified (following Henrich et al., 2010b).

Explanations for Cross-Population Variation in Prosocial Behavior

Many explanations for population variation in prosocial behavior hinge on differences in values, such as collectivism, embeddedness, or self-expression, which are hypothesized to motivate helping and sharing in different ways (Knafo et al., 2009; Levine et al., 2001). For example, some cross-cultural psychologists have proposed that individuals in societies with greater collectivism will be more generous to in-group members but either equally or less generous to out-group members (Triandis, 1989). Others have argued that region-specific values, such as *simpatico* in Spanish and Latin American cultures may account for population differences in helping (Levine et al., 2001).

Value-based arguments outline a link between cognition and behavior, but they leave open the question of how these value differences arise. One possibility is that these values simply reflect particular cultural histories of different populations. According to this argument, over the course of human history, some groups became more collectivist and others more individualist in an arbitrary process of cultural drift. Another possibility is that prosocial values and behaviors respond to or coevolve with specific social and material situations in predictable ways. Here we review seven prominent explanations which propose general processes of coadaptation with the environment.

Before reviewing these explanations, it is important to consider a key difference in how they view behavior interacting with the environment. Some theories treat the physical and social environment as an exogenous driver of behavior (e.g., greater population density leads to less helping of strangers or greater pathogen stress leads to increased favoritism toward in-group members). Other theories posit that physical and social environments can themselves arise from patterns of prosocial behavior. For example, infectious disease risk, which is proposed by some theories to be a driver of prosocial behavior, is not simply an exogenous element of the environment. Rather it has changed in response to the emergence of public health institutions, which were in turn the outcome of early prosocial attempts to improve others' health. Such feedback between environments and behavior can lead to significant coevolutionary trajectories. According to some theories, these trajectories may account for the modern-day emergence and maintenance of cultural systems such as markets and world religions (see following section). If such feedbacks exist, it is also possible that the processes described in the following section may interact in important ways, rather than being independent and mutually exclusive explanations.

Population Density and Size

A number of theories predict that population size or density influence prosocial behavior. Milgrim argued that living in cities had a number of cognitive consequences, including a reduction in prosocial behavior toward strangers. Specifically, with increasing population density and frequency of interactions among strangers, individuals cannot plausibly concern themselves with everyone they meet. Rather, they develop norms whereby they selectively attend to known individuals, but effectively ignore strangers. According to this hypothesis, as people interact more frequently with a larger number of individuals, we would expect them to become less helpful toward strangers (Milgrim, 1970). Darley and Latane propose an alternative hypothesis that individuals are less likely to offer help when more potential helpers are present at the immediate location as this makes any single person less responsible for helping. This explanation based on a bystander effect predicts that individuals will be less helpful when more people are around (Latane and Nida, 1981; Darley and Latane, 1968). A third line of reasoning from economics and sociology focuses on the size of communities, rather than population density. It predicts that individuals will be less willing to help, share, or engage in collective action in the context of larger communities, because it is more difficult to enforce the eventual return of that prosocial behavior (Olson, 1965). A fourth argument proposes instead that larger communities cannot function without indoctrinating members with internalized norms of generalized sharing, helping, and cooperation. This indoctrination among larger groups might counteract or even reverse the effect of strategic decisions to defect and might even increase levels of prosociality in larger communities (Henrich et al., 2010a). Consistent with the first two explanations, a study of helping behavior in 36 US cities found that increasing population density (rather than size) was the strongest predictor of decreasing willingness to help strangers in a number of situations (Levine et al. 2001). In other studies of sharing and contributions to collective efforts, community or group size has shown mixed results (Henrich et al., 2010a; Lamba and Mace, 2011; Henrich et al., 2012). Indeed, consistent with the fourth argument, a recent multisite study showed that one kind of prosocial behavior, anonymous prosocial punishment, actually *increased* with increasing community size (Henrich et al., 2010a).

Market Integration

The market integration hypothesis proposes that market norms emphasizing fair treatment of anonymous others have culturally evolved to sustain mutually beneficial exchanges in contexts demanding frequent interaction with strangers. In turn, as individuals increasingly interact with markets, they adopt and internalize these norms (Henrich et al., 2010a). Reciprocally, markets spread more successfully in places with such norms. According to this hypothesis, individuals with greater market exposure will be more likely to have adopted or internalized these norms and thus will treat anonymous others more fairly. This hypothesis has been tested, replicated, and extended in two separate projects covering 24 different societies scattered across the globe, from Siberia to New Guinea. The populations involved hunter-gatherers, slash-and-burn horticulturalists, subsistence farmers, and wage workers. Overall, more market integrated societies tend to make more even splits of a pot of money, independent of the threat of punishment, income, wealth, education, community size, sex, and age (Henrich et al., 2005, 2010a). Since such equitable tendencies emerge even when punishment is not possible, and anonymity is assured, the authors argue they arise from internalized local norms. Supporting the internalization hypothesis, recent work shows that 'giving' by westerners in such experiments is the automatic response (Rand et al., 2012) and involves activation of the brain's reward circuitry (Harbaugh et al., 2007).

Two important concerns with Henrich et al.'s projects revolve around the many sources of variation in such a large, global sample (many factors probably are not controlled for), and the direction of causality – market integration in these studies – cannot be identified as an exogenous cause. Both concerns have been addressed recently. Among 57 communities in a single ethno-linguistic herding group in Ethiopia, Rustagi et al. (2010) measured people's willingness to cooperate with strangers in an economic game. Since these groups are tied to their land by customary rights, people's distance to markets can be used as a proxy for market integration that is exogenous (not subject to reverse causality). Herders who live closer to markets were more cooperative, even after statistically controlling for sex, wealth, literacy, land, and many other variables. Here market contact seems to have caused people to be more cooperative in anonymous situations. Perhaps most importantly, groups with more cooperators were better able to make, monitor, and enforce local agreements that resulted in measurable improvements in forest conservation (Rustagi et al., 2010). That is, the experience of markets changes our sociality to make it easier to construct and enforce certain kinds of institutions – increasing institutional quality. Voors et al. (2012) have similar findings from Burundi.

Religion

Many religious traditions emphasize the importance of prosocial behavior, and thus enculturation in specific religions may influence prosocial behavior. Numerous studies in psychology have now shown that 'priming' (unconsciously bringing to mind) thoughts of god causes theists, but not atheists, to give more in behavioral games (Norenzayan and Shariff, 2008). One current theory holds that modern world religions, such as Christianity and Islam, were able to spread precisely because they effectively enculturated norms of prosocial behavior which galvanized large-scale cooperation among relatively anonymous strangers (Atran and Henrich, 2010). According to this view, followers of modern world religions, such as Christianity and Islam, will be more likely to have internalized these norms of prosocial behavior and will thus treat anonymous others with greater fairness and generosity. Findings from the cross-society studies described earlier are also consistent with this hypothesis (Henrich et al., 2010a).

Institutional Quality Hypothesis

Variants of the institutional quality hypothesis propose that informal and formal institutions change the costs and benefits of prosocial behavior, which in turn shape the social norms and behavior. Public services, global markets, and social safety nets that mitigate material threats may render investments in an expansive network of kith and kin less necessary as alternative forms of social insurance (Inglehart and Welzel, 2005; Hruschka, 2010). Legal mechanisms guaranteeing relatively safe exchange and interaction in largely anonymous settings lower the costs of interacting with and cultivating beneficial relationships with strangers. Finally, to fully participate in these institutions, individuals must follow norms of fairness and civil interaction (Hruschka and Henrich, 2013). Based on this hypothesis, when individuals have access to functioning formal institutions, they are more likely to follow institutional rules, such as honest communication and equitable treatment of others. Moreover, reduced barriers to and increased benefits of cultivating beneficial relationships with strangers will lead to greater trust with anonymous individuals and reduced parochialism in prosocial behavior. This hypothesis is consistent with a recent study comparing trusting behavior among Nama pastoralists across two countries varying in institutional quality (Namibia and South Africa) (Vollan, 2012). It is also consistent with findings from an international study of rule following and parochialism across eight sites which varied in terms of institutional quality (Hruschka et al., 2014).

Globalization

This explanation proposes that as people increasingly participate in the network of global, social, and cultural relations, they are less likely to think in terms of in-groups and out-groups and more likely to imagine humankind as a 'we' where there are no 'outsiders' (Buchan et al., 2009). Thus, individuals with greater interactions with global communication (e.g., cell phones, televisions, and employment in transnational firms) will be more inclined to engage in collective action with individuals outside of their immediate in-group. Consistent with this hypothesis, Buchan et al. (2009) found that contribution to global public goods increases when people identify with the world rather than their own nation.

Pathogen Stress

Recently, several evolutionary researchers have proposed that preferential prosociality toward in-group members constitutes a form of behavioral immune system against the spread of pathogens. According to this hypothesis, in regions with high risk of infection by dangerous pathogens, individuals will preferentially interact with in-group members in a way that insulates them from infection by out-group members (Fincher and Thornhill, 2012; Schaller and Murray, 2010). Although originally predicting xenophobia (negative out-group attitudes and behaviors), the theory has been extended to account for in-group favoritism (positive in-group attitudes and behaviors) as well (Fincher and Thornhill, 2012). This hypothesis differs crucially from other hypotheses by positing that the adaptive mechanisms responsible for this effect are specific to pathogen risk and were designed to impede the spread of pathogens. Emerging experimental evidence suggests that people do indeed adjust social motivations and behaviors (i.e., conformism) to specific cues of pathogen threats over and above generalized threats (Murray and Schaller, 2012). However, existing cross-population analyses suggest that broad population variation in prosociality is more likely a result of other factors, such as institutional quality or extrinsic risk (Hackman and Hruschka, 2013, Hruschka and Henrich, 2013).

Experience of War

Evolutionary researchers have also proposed that humans should be sensitive to the threat or experience of intergroup conflict. Work with senior Israeli citizens before, during, and after the 2008 conflict with Hezbollah (Gneezy and Fessler, 2012), shows that people's willingness to punish unfairness increased during the war, but returns to prewar level soon after. Consistent with this, research in both the Republic of Georgia (Caucuses) and Sierra Leone (West Africa) converges to indicate that the experience of war during a developmental window from roughly age 7 to 20 increases people's 'parochial egalitarianism' – their tendency to favor more equal sharing within their in-group, relative to an out-group (that is not the opponent). This effect endures at least a decade post conflict (Bauer et al., 2012). Convergent lines of experimental research have emerged from conflict studies in Burundi (Voors et al., 2012). Since different populations have different recent experiences of war, we should expect global variation to arise from this factor.

Future Directions – Prosocial Trade-Offs and Social Norms

In addition to identifying population variation in behavior, cross-population studies can lead to better theory and research by productively challenging common assumptions about prosocial behavior. Consider the case of raiding parties among East African Turkana pastoralists. On occasion, Turkana males assemble large war parties, raid neighboring ethnic groups, and steal their victims' herds. Despite the high potential costs of participating in a raid, many warriors resist the urge to desert their party, and instead assist in the theft. If instead a warrior chooses to shirk and desert his companions, he faces a number of potential consequences. Community members may verbally rebuke him or they may demand that he sacrifices an animal from his herd. In especially egregious cases, the deserter's age-mates may tie him to a tree and beat him with a whip (Mathew and Boyd, 2011). Participating in cattle theft and exacting retributive justice are not actions that most readers would associate with prosociality. Yet these acts – assisting one's comrades and punishing a cheater – are intended to benefit others and to contribute to the well-being of one's group, two key criteria of prosocial behavior (Mathew and Boyd, 2011).

The Turkana example illustrates two important points. First, any given social situation can invoke many potentially contradictory prosocial motives – helping companions, protecting one's community and its territory, or respecting the life, property, and happiness of others. In this thicket of competing motives and outcomes, humans can disagree, sometimes dramatically, about what motives, behaviors, and outcomes are most desirable and thus what counts as prosocial or antisocial. Even an act as apparently prosocial as excess generosity can be considered antisocial if it comes off as flaunting one's advantage, as an attempt to bring another person into one's debt, or simply as violating a norm of moderation (Herrmann et al., 2008). To address this ambiguity, social norms often emerge in frequently encountered contexts to establish the appropriate way to behave. Even within a single society, different contexts can call for different behaviors. In the USA, helping a friend is good, but helping a friend make money on the stock market by leaking company secrets is not. Among Samburu pastoralists of north-central Kenya, individuals are more likely to evenly split a pot of money if the activity is framed in terms of traditional meat sharing than as a decontextualized sharing game (Lesorogol, 2007). Work examining population and contextual variation in what counts as the appropriate or prosocial way to act in specific contexts will shed important light on how social norms of prosociality are formed and how they guide behavior.

Second, humans live in complex webs of social relationships with frequently conflicting allegiances and obligations. Sacrificing for one person or group may mean injuring or neglecting another (Reicher et al., 2012). Thus, while the willingness to sacrifice for others or for one's group appears to be a recurring aspect of human psychology, human populations differ markedly in the allegiances and obligations seen as most important and in the behaviors most appropriate when acting in the service of others, one's group, or on a moral principle. Consider a pedestrian who walks past an injured bicyclist without offering aid. Perhaps that pedestrian is being selfish with her time. Or, perhaps she is trying to catch a bus to comfort her friend in the hospital. Alternatively, consider allocations in the dictator game, where an individual can divide $10 between himself and an anonymous stranger. The money kept for self is usually considered the selfish option. However, if the person turns around and spends that money on food to be shared with others, the 'selfish' decision simply reflects a different preference for prosocial sharing. These prosocial outside options can be difficult to control and measure in scientific studies. However, given potential cross-population

differences in such outside options, they should be carefully considered when interpreting cross-population variation in observed behavior. More generally, by clearly specifying alternative prosocial options, researchers can gain insight into how people make tough decisions between apparently incommensurate kinds of 'good' behavior.

A key problem in understanding both how prosocial norms arise and how people negotiate prosocial trade-offs is to understand the time scales of change that give rise to cross-population variation (Schaller and Murray, 2010). On very short time scales, people likely have some ability to make facultative calculations about the relative merits of investing in different prosocial options, as we saw when individuals punished more during a war. On developmental timescales, early life experience may influence one's willingness to engage in certain kinds of prosocial behavior toward others, as we see in Sierra Leone and Georgia (Bauer et al., 2012). On generational timescales, people acquire culturally transmitted beliefs, norms, motivation, habits, practices, values, reasoning styles, or strategies by using adaptive learning heuristics. Cultural traits that permit individuals or groups to successfully manage beneficial social interactions can spread and persist through a variety of cultural evolutionary processes (Henrich et al., 2010a). At the deepest timescale, if cultural niches are sufficiently stable, natural selection may favor genes that foster different amounts of in-group investment (Chiao and Blizinsky, 2010; Eisenberg and Hayes, 2011). Integrating study designs across these various timescales – including behavioral and field experiments and longitudinal and cross-population analyses – will be crucial for identifying how these population differences in prosocial behavior arise and persist (Hackman and Hruschka, in press).

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Psychology; Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Helping Intergroup Relations; Prosocial Behavior During Adolescence; Prosocial Behavior During Childhood and Cultural Variations; Prosocial Behavior During Infancy and Early Childhood: Developmental Patterns and Cultural Variations; Prosocial Behavior and Empathy; Prosocial Behavior, Effects of Parenting and Family Structure on.

Bibliography

Atran, S., Henrich, J., 2010. The evolution of religion: how cognitive by-products, adaptive learning heuristics, ritual displays and group competition generate deep commitments to prosocial religions. Biological Theory 5, 18–30.

Batson, C.D., Powell, A.A., 2003. Altruism and prosocial behavior. In: Handbook of Psychology.

Bauer, M., Cassar, A., Chytilova, J., Henrich, J., 2012. Warfare Experience during Ontogeny Increases Egalitarian and Parochial Motivations. United Nations University, Maastricht, Netherlands.

Buchan, N., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., Foddy, M., 2009. Globalization and human cooperation. Proceedings of the National Academy of Sciences 106, 4138–4142.

Chiao, J.Y., Blizinsky, K.D., 2010. Culture–gene coevolution of individualism–collectivism and the serotonin transporter gene. Proceedings of the Royal Society B: Biological Sciences 277, 529–537.

Darley, J.M., Latane, B., 1968. Bystander intervention in emergencies: diffusion in responsibility. Journal of Personality and Social Psychology 8, 377–383.

Eisenberg, D.T.A., Hayes, M.G., 2011. Testing the null hypothesis: comments on 'Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene'. Proceedings of the Royal Society B: Biological Sciences 278 (1704), 329–332.

Eisenberg, N., Fabes, R.A., Spinrad, T.L., 2007. Prosocial development. In: Handbook of Child Psychology.

Fearon, J.D., Humphreys, M., Weinstein, J.M., 2009. Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia. American Economic Review 99, 287–291.

Fincher, C., Thornhill, R., 2012. Parasite stress promotes in-group assortative sociality and heightened religiosity. Behavioral and Brain Sciences 35, 61–79.

Gelfand, M.J., et al., 2011. Differences between tight and loose cultures: a 33-nation study. Science 332, 1100–1104.

Gneezy, A., Fessler, D.M.T., 2012. Conflict, sticks and carrots: war increases prosocial punishments and rewards. Proceedings of the Royal Society B 279, 219–223.

Hackman, J., Hruschka, D., 2013. Fast life histories, not pathogens, account for state-level variation in homicide, child maltreatment, and family ties in the US. Evolution and Human Behavior 34 (2), 118–124.

Harbaugh, W.T., Mayr, U., Burghart, D.R., 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. Science 316, 1622–1625.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., Mcelreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J.Q., Tracer, D., 2005. "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences 28, 795–855.

Henrich, J., Boyd, R., Mcelreath, R., Gurven, M., Richerson, P.J., Ensminger, J., Alvard, M., Barr, A., Barrett, H.C., Bolyanatz, A., Camerer, C., Cardenas, J.C., Fehr, E., Gintis, H., Gil-White, F., Gwako, E., Henrich, N., Hill, K., Lesorogol, C., Patton, J.Q., Marlowe, F., Tracer, D., Ziker, J., 2012. Culture does account for variation in game behavior. Proceedings of the National Academy of Sciences 102, E32–E33.

Henrich, J., Ensminger, J., Mcelreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2010a. Markets, religion, community size, and the evolution of fairness and punishment. Science 327, 1480–1484.

Henrich, J., Heine, S.J., Norenzayan, A., 2010b. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–83.

Henrich, J., Mcelreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, J., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767–1770.

Herrmann, B., Thoni, C., Gachter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.

Hruschka, D.J., 2010. Friendship: Development, Ecology, and Evolution of a Relationship. University of California Press, Berkeley, CA.

Hruschka, D.J., Hackman, J. When are cross-group differences a product of a human behavioral immune system? Evolutionary Behavioral Sciences. in press.

Hruschka, D.J., Henrich, J., 2013. Institutions, parasites and the persistence of in-group preferences. PLOS ONE 8 (5), e63642.

Hruschka, D.J., Efferson, C., Jiang, T., Falletta-Cowden, A., Sigurdsson, S., McNamara, R., Sands, M., Munira, S., Slingerland, E., Henrich, J., 2014. Impartial institutions, pathogen stress, and the expanding social network. Human Nature 25 (4).

Inglehart, R., Welzel, C., 2005. Modernization, Cultural Change, and Democracy: The Human Development Sequence. Cambridge University Press, Cambridge.

Knafo, A., Schwartz, S.H., Levine, R.V., 2009. Helping strangers is lower in embedded cultures. Journal of Cross-Cultural Psychology 40, 876–879.

Lamba, S., Mace, R., 2011. Demography and ecology drive variation in cooperation across human populations. Proceedings of the National Academy of Sciences 108, 14426–14430.

Latane, B., Nida, S., 1981. Ten years of research on group size and helping. Psychological Bulletin 89, 308–324.

Lesorogol, C., 2007. Bringing norms in: the role of context in experimental dictator games. Current Anthropology 48, 920–926.

Levine, R.V., Martinez, T.S., Brase, G., Sorenson, K., 1994. Helping in 36 U.S. cities. Journal of Personality and Social Psychology 67, 69–82.

Levine, R.V., Norenzayan, A., Philbrick, K., 2001. Cross-cultural differences in helping strangers. Journal of Cross-Cultural Psychology 32, 543–560.

Mathew, S., Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. Proceedings of the National Academy of Sciences 108, 11375–11380.

Milgrim, S., 1970. The experience of living in cities. Science 167, 1461–1468.

Murray, D.R., Schaller, M., 2012. Threat(s) and conformity deconstructed: perceived threat of infectious disease and its implications for conformist attitudes and behavior. European Journal of Social Psychology 2, 180–188.

Norenzayan, A., Shariff, A.F., 2008. The origin and evolution of religious prosociality. Science 322, 58–62.
Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.
Penner, L.A., Dovidio, J.F., Piliavin, J.A., Schroeder, D.A., 2006. The Social Psychology of Prosocial Behavior. Erlbaum, Mahwah, NJ.
Rand, D.G., Greene, J.D., Nowak, M.A., 2012. Spontaneous giving and calculated greed. Nature 489, 427–430.
Reicher, S., Haslam, S.A., Smith, J.R., 2012. Working toward the experimenter: reconceptualizing obedience with the Milgram paradigm as identification-based followership. Perspectives on Psychological Science 7, 315–324.
Rustagi, D., Engel, S., Kosfeld, M., 2010. Conditional cooperation and monitoring explain success in forest commons management. Science 330, 961–965.
Schaller, M., Murray, D.R., 2010. Infectious disease and the evolution of cross-cultural differences. In: Schaller, M. (Ed.), Evolution, Culture, and the Human Mind.
Steblay, N.M., 1987. Helping behavior in rural and urban environments: a meta-analysis. Psychological Bulletin 102, 346–356.
Triandis, H.C., 1989. The self and social behavior in cultural contexts. Psychological Review 96, 506–520.
Trompenaars, F., Hampden-Turner, C., 1997. Riding the Waves of Culture: Understanding Diversity in Global Business. McGraw-Hill, New York.
Vollan, B., 2012. Weird reciprocity? A 'within-culture across-country' trust experiment and methodological implications. Journal of Institutional Economics 8, 371–397.
Voors, M.J., Nillesen, E.M., Verwimp, P., Bulte, E.H., Lensink, R., Soest, D.P., 2012. Violent conflict and behavior: a field experiment in Burundi. American Economic Review 102, 941–964.

Prosocial Behavior, Effects of Parenting and Family Structure on

Liat Hasenfratz and Ariel Knafo, The Hebrew University of Jerusalem, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is widely acknowledged that parents from different cultural communities use different techniques to socialize prosocial behaviors and these differences are associated with differences in quantity and quality of prosocial behaviors. There is, however, a wide disagreement regarding the exact role that culture has in explaining observable phenomenological differences and thus relatively little systematic research on parenting and the development of prosocial behavior in a cross-cultural context. The authors, thus, use individual existing studies to create an overview of the cultural variety in parenting and prosocial behavior and to outline how culture, parenting and prosocial behavior interact in concert to create diversity in forms of human cooperation and compassion.

Introduction

This article aims to contribute to an understanding on the relationship between parenting and children's prosocial behavior from a culturally informed point of view.

The socialization of prosocial behavior has been studied almost entirely among Western white middle class populations. Theories on the socialization of prosocial behavior have largely been developed by researchers in North America (e.g., Grusec and Goodnow, 1994; Hoffman, 1988). In contrast to the well-researched relationship between parenting and prosocial behavior, there is very little research on parenting and prosocial behavior in a cross-cultural context.

Across the world, the particular ways by which parents interact with their children is influenced by the particular cultural beliefs about the human nature and its development, and beliefs about the role of parents in promoting adaptive social development (i.e., parental ethnotheories, Harkness and Super, 2006). It is widely acknowledged that parents from different cultural communities use different techniques to socialize prosocial behaviors and these differences are associated with differences in quantity and quality of prosocial behaviors (Eisenberg et al., 2006). There is, however, a corresponding wide disagreement among researchers regarding the exact role that culture has in explaining observable phenomenological differences, and this is probably one of the reasons for which culture has played such a minimal role in the explanation of the relation between parenting and the development of prosocial behavior.

Moreover, the existing research is rather heterogeneous, both with regard to the underlying theoretical assumptions about the role of culture and with regard to the definitions and operationalizations of parenting and of the prosocial behaviors observed. Thus, whether parents differ in their behaviors and preferences for socialization strategies and whether these strategies have similar associations with children's prosocial behavior across cultures remain open questions.

In the present article, the authors use individual existing studies to create an overview of the cultural variety in parenting and prosocial behavior. Viewed in such a context, the often unrelated findings come together to suggest two general ways of cultural influence.

The authors begin with an outline of the generally accepted ideas about parenting and prosocial behaviors and then describe each of the constructs in a cultural perspective. Finally, they outline how culture, parenting, and prosocial behavior interact in concert to create diversity in forms of human cooperation and compassion.

What Is Prosocial Behavior and How Does It Develop?

Prosocial behavior is defined as 'voluntary behavior intended to benefit another' (Eisenberg et al., 2006). It is characterized by acts of kindness, compassion, and helping behaviors, which many consider to be one of the finest qualities of human nature.

Typical examples of prosocial behavior include volunteering, sharing toys, treats or food with friends, instrumental help (e.g., helping a peer with school assignments), costly help (e.g., risking own life to save others), and emotionally supporting others in distress (e.g., comforting a peer following a disappointing experience or caring for a person not feeling well). Empathy is considered as the emotion that provides both the foundation for prosocial development and the mechanism for social influence over behavior (Hastings et al., 2007).

According to both Hoffman's theory (2000) and ample empirical findings, prosocial behavior and empathy emerge early in life and develop until at least early adulthood. At its beginning, prosociality is assumed to be primarily self-oriented. It has been argued (Hoffman, 2000) that infants feel self-distress in reaction to the distress of others because they are incapable of differentiating their own experiences from those of others. Although recently evidence has accumulated showing that empathy emerges earlier than previously thought (Roth-Hanania et al., 2011), there is an overall agreement that prosocial behavior becomes increasingly other-oriented over the course of development. Gradually, self-distress is replaced by other-oriented concern, requiring some understanding of others' mental states (Hoffman, 2000). Several studies have documented the enactment of prosocial behaviors such as sharing, helping, and showing compassion as early as the second year of life and find the maximum growth in maturity and frequency of prosocial behaviors in the toddler and

preschooler period (Eisenberg et al., 2006). From childhood to adolescence, further increases are found in sharing, and less so in helping or providing emotional support. The boost in prosocial behavior with age is attributed to developmental increases in cognitive abilities associated with detecting others' needs and determining ways to help, in empathy-related responding, and in the moral understanding of the importance of helping others (Eisenberg et al., 2006).

However, because of its rather broad definition, it has been difficult to determine the exact developmental course of prosocial behavior. For example, Eisenberg and Fabes' (1998) meta-analysis found that prosocial behavior increases with age, but that the increases vary in size depending on the methodological aspects of each study. In a similar vein, investigators have not always found correlations between indices of prosociality. Prosocial behavior can include a variety of observable behaviors (e.g., helping, sharing, comforting, cooperating, and showing compassion) that may be driven by rather different motivations (e.g., the wish to benefit the other, self-interest, or practical concerns); they may be accompanied by different emotions (e.g., empathy, sympathy, or distress), and based on different cognitions (i.e., differing interpretations of situational cues). These different motivations and manifestations may reflect important distinctions for understanding prosocial behavior and its development, but relatively little socialization research has addressed this issue. There are various indications that, even within a culture, measuring prosocial behavior and development is moderated by factors such as age, sex, and the specific operationalizations involved. For example, when studying the development of children's willingness to make costly prosocial choices, it matters whether the observed child population comes from a high or a low SES background (Benenson et al., 2007). The willingness to share or to cooperate varies with the age and the sex of the person in need of it. Similarly, helping or sharing with an unknown other or with an acquaintance (Fehr et al., 2008) has neither theoretically nor practically the same meaning.

Classic operationalizations to study such voluntary prosocial acts guided by internalized values and goals rather than by the expectation of concrete or social rewards or the avoidance of punishment, include testing children's self-initiated willingness to share their resources such as food (i.e., chocolate) with an unknown stranger or to help a distressed person, who has dropped, broken, or lost something.

Such operationalizations are useful measures of spontaneous selfless acts, but admittedly they do not capture the various forms of human prosociality. Moreover, as elaborated later, neither the above-mentioned conceptualization of prosocial behavior nor the ways in which it is studied bear up in a cross-cultural comparison.

What Role Does Parenting Play in Bringing about Prosocial Behavior?

From the pioneering work of Baumrind (e.g., 1971) and Hoffman (e.g., 1970) to the present, parenting styles and techniques of discipline have been the most prominent explanatory frameworks for the development of prosocial behavior and have inspired ample research. Most of this research is based on single-time correlational studies within the well-studied postindustrialized Western information societies. Nevertheless, taken together, these studies have created a widely accepted profile of child rearing that typifies socialization of more prosocial children (see Eisenberg et al., 2006 for a review of this research).

The authors focus here on the construct of parenting styles because it represents a whole system of parent–child interaction that entails also more general aspects such as parental beliefs and more specific interaction patterns.

Research has classified the various interaction patterns according to the extent to which they reflect responsiveness and demandingness (Maccoby and Martin, 1983), or warmth and strictness (García and Garcia, 2009; Sears et al., 1957). Authoritative parents are balancing reasonable exertions of control and consistent expectations for maturity with flexibility and responsiveness to children's desires and needs. Authoritarian parents, in contrast, score higher on demandingness than on responsiveness; they "value obedience and restrict autonomy" (Baumrind, 1966: p. 890). Indulgent parents express affection to their children but show low levels of behavioral control.

In line with the idea that children will learn best socially appropriate behavior from those adults whom they have accepted as role models (Grusec and Goodnow, 1994), an authoritative parenting style has been shown to make children more receptive to efforts to foster concern for others (Hastings et al., 2000) – and identified as ultimately the best way to promote prosocial behavior. Parents of prosocial children model prosocial behavior themselves, encourage such behavior from their children and provide explanations for these expected behaviors (Hastings et al., 2007). In support of this assumption, the other parenting styles have been found to be less related to prosocial behavior and at times even as undermining the prosociality of their offspring.

The authoritarian style, for instance, has been described as believing in strict adherence to parental rules, the use of punishment (also corporal punishment) and unlikely to discuss rules with the child, behaviors that may lead children to reject instead of accept parental socialization efforts and undermine children's prosocial behavior by modeling a lack of concern for the needs of others (Hastings et al., 2006). Parental permissiveness, or indulgence, on the other hand, is viewed as leaving children without a clear sense of parental and social expectations, unsuitable to instill in children prosocial values as moral imperatives (Baumrind, 1971; Steinberg et al., 1994; Maccoby and Martin, 1983).

The notion that a democratic parenting style will be most effective in promoting prosocial behaviors is based on the idea that on the one hand too much external control will contravene internally regulated action because it impedes autonomy, and that on the other hand too little external guidance and demands will lead to undirected and aimless action. For example, self-determination theory (Deci and Ryan, 1985) has postulated that because humans share the fundamental psychological need for autonomy, children will internalize societal values and behaviors (i.e., make them their own) when parents support children's need for autonomy and foster an internalization of social rules and norms rather than

apply external pressure. Accordingly, the theory predicts less adaptive outcomes, whenever and wherever the cultural values are overly discordant with the intrinsic human nature (Grolnick et al., 1997). Indeed, the use of either physical or psychological control have been shown to have either no or negative effects on teaching prosocial behavior (Hastings et al., 2006 for a review of this research).

However, the relationship between the ways of parenting and prosocial behavior maybe culture dependent. Cross-cultural and anthropological studies have shown that cultural groups differ with regard to the emphasis they place on supporting autonomy or on fraternal submission and filial piety, which in turn will influence how much parents will focus on promotion of autonomy in their socialization practices. For example, in contrast to white middle class North Americans, who typically value assertiveness and independence in their children, parents in countries with more collectivist orientations, such as traditional China, tend to emphasize the importance of behavioral inhibition and subjugation of individual wants for the good of the family (Rothbaum et al., 2000a).

Cultures differ with regard to how they construe the self in relation to the social group, and these differing constructions create different frameworks for shaping on the one hand socialization goals and on the other hand the ideas about the ways in which these goals may be achieved (Keller et al., 2005; Keller et al., 2002; Harkness and Super, 2006) – ultimately shaping and demanding different parenting styles.

What Is Culture and How Is It Shaping Parenting?

Culture describes the 'shared symbolic system' of a community (Boesch, 1991). These systems are made up of shared language, knowledge, and networks of social relations and contain customs, laws, and beliefs. Cultural psychologists have long argued that developmental psychologists have tended to assume, if only by their restricted use of study populations, that the psychological functions they are studying have a biological basis and are therefore shared by all members of the human species (Norenzayan and Heine, 2005). This is in spite of the fact that different ecocultural environments (Berry, 1976), different socioeconomic social structures (Morelli and Tronick, 1991), or different types of community (Raval and Martini, 2009) result in different cultural models with different meaning systems that are forming the contexts for culture-specific parent–child interaction and ultimately demand different socialization styles.

In all cultures, parents interact with their children in ways that reflect their beliefs about socially appropriate behavior. Parental beliefs about child behavior and appropriate socialization goals and strategies come in part from folk knowledge about human nature and the structure of society (LeVine and White, 1987). According to Harkness and Super, parental ethnotheories are the "nexus through which elements of the larger culture are filtered" and represent central components of child development and socialization practices (Harkness et al., 2001: p. 9). A growing body of empirical studies has demonstrated that parents' beliefs about child behavior and development have important direct and indirect influences on children's cognitive, emotional, and social competence (Goodnow and Collins, 1990).

For example, the psychological assumptions about the course and pace of prosocial development and the role of democratic parenting in promoting socially adaptive development that has been reviewed above seems at least in part to be rooted in the ideas about the human nature prevailing in Western postindustrialized culture. In Western white middle class communities, prosocial socialization begins at earliest ages, is viewed as the sole responsibility of the immediate nuclear family, and as the result of a successful application of effective socialization techniques. Those beliefs and the child rearing practices that go along with them are in line with the Puritan notion of a child's nature as inherently wicked, ruthless, and selfish and the belief that only extensive socializing will "foster salvation of their souls by enforcing habits of righteousness," constituting the Protestant heritage of many Western cultures (Mosier and Rogoff, 2003).

Not all cultures believe that the child's inborn nature is selfish and that he or she has to be 'taught generosity and altruism.' The Maya in Guatemala view prosocial behavior as inherent to the social order and its development is the result of a process of gradual understanding and maturation (Mosier and Rogoff, 2003). Thus, in contrast to the notion of a need for an extensive socialization prevailing in Euro-American societies described above, children of Mayan parents do not have to be 'extensively socialized' into behaving prosocially. Similarly, in traditional Chinese families, a child's socialization is viewed as the responsibility of the community, not only of the parents, and an individual, who fails to develop adequate prosocial behaviors, is considered to be the failure of the community as a whole (Xu et al., 2007).

However, there have been relatively few studies systematically comparing parental belief systems from a cross-cultural perspective, and existing studies have focused primarily on children's cognitive development (e.g., Miller and Goodnow, 1995) rather than on social development.

The few existing studies that have examined parental beliefs about prosocial behaviors in a cross-cultural context indicate that there are significant cultural differences in the techniques used to socialize children into becoming (pro)social creatures. For example, in comparison with US mothers, Japanese mothers of young children have been found to be much less direct and forceful in their disciplinary strategies (Conroy et al., 1980; Zahn-Waxler et al., 1996). Similarly, in a cross-national comparative study of early development and socialization, Japanese mothers tended to expect obedient, self-controlled child behavior at an earlier age than US parents, whereas US parents tended to value assertive, socially skilled behavior in their young children (Hess et al., 1980).

To summarize, cultural differences with regard to customary beliefs about the human nature and its development inform culture-specific assumptions about parental strategies and mechanisms to promote children's socially adaptive development. Theorists and researchers have been arguing about the consequences of such a cultural diversity in parenting. Some researchers argue that a universal human capacity for prosocial emotions and prosocial behavior exists as part of human altruism (Hoffman, 2000), and that culture is an adaptive system of rules and practices shared by a community

(Matsumoto and Juang, 2004). Thus, culture-specific parenting styles may be just as adaptive as the authoritative style is for the well-researched White Western middle class families, as long as parenting provides the right composition of warmth and demandingness. The Nso of Cameroon, for instance, who socialize children toward obedience, respect for elders, and responsibility for others, view good parenting as the one in which parents take the lead by controlling and training the child (Yovsi et al., 2009).

Others, however, have been arguing that the idea of universality of parenting dimensions and/or of prosocial behavior is insufficient as it can neither explain the well-documented differences in prosocial behavior among cultures – which may vary both in kind and/or in amount (e.g., Knafo et al., 2009; Eisenberg et al., 2006; Whiting and Whiting, 1973) – nor can it predict individual differences in prosocial behavior within cultures (Park et al., 2012).

Does Culture Influence Prosocial Behavior?

According to Trommsdorff et al. (2007), a universal ability for altruism "cannot predict universality of prosocial behavior in different cultural contexts because (1) such behavior includes evaluations based on cultural values, self-construal, and moral beliefs; (2) cultural differences in prosocial behavior are documented (Whiting and Whiting, 1973); and (3) the concept of prosocial behavior consists of a variety of forms that makes a generalization difficult even at the individual level."

Indeed, Knafo et al. (2009) found that people from cultures high on embeddedness (e.g., Malaysia and Singapore), with a high regard for the family or extended in-group norms, were more reluctant to help strangers than people from cultures low on embeddedness (e.g., Sweden and Costa Rica) that view the autonomous individual as the key social unit. On the other hand, cultures with a high regard for the family tend to demand more and earlier of children to take responsibility for the welfare of family members and to assign chores and responsibilities in the family. The early assignment of chores and taking on responsibility for welfare of family members and the family's economic well-being appears to provide children with many more opportunities to practice and enact altruism. Several researchers have suggested that children and youth from traditional rural and semiagricultural communities and from relatively traditional subcultures are therefore more cooperative than children from urban or Westernized cultures (see Eisenberg et al., 2006 for an overview and discussion of this research). In support of this notion, Knight et al. (1982) found Mexican-American children to be more cooperative than Anglo-American children, who were found to be rather competitive. Rao and Stewart (1999) found Chinese and Indian children were more generous than middle class Western preschool-aged children. These findings appear to support the above-made postulation, but does helping a needy stranger cross the street have the same meaning as helping with the household chores or sharing food with an unknown stranger in a theoretical allocation task?

Making cross-cultural comparisons can be difficult because prosocial behavior is always measured within the context of what constitutes prosocial behavior within the context of a certain community. Using the same operationalization across cultures, such as the amount of observable helping behavior, can tell us of cross-cultural differences in the quantitative expressions of a certain type of behavior, but it might fail to capture the meaning and value the behavior has within the particular cultural system. Cultures differ considerably in their valuing of different types of prosocial action. Hindu Indians, for instance, view prosocial behavior performed because of reciprocity considerations as more moral than American adults do (Miller and Bersoff, 1994). There is reason to assume that, although sharing with an unfamiliar peer seems a good operationalization of altruistic prosocial behavior, it might not be a useful tool to study unselfish behavior in cultures in which sharing with an anonym is very unusual. Trommsdorff et al. (2007) found that children from Western cultures (i.e., Israeli and German) were significantly more willing to console a distressed adult experimenter spontaneously than were children from South Asian cultures (i.e., Indonesia and Malaysia). Such a result suggests that children from Western cultures are more prosocial than children from South Asian cultures, but it is equally possible that instead of or at least in addition to measuring spontaneous prosocial behavior, the experimental set-up also tapped cross-cultural differences in customary child–adult interactions.

In a similar vein, the prevailing theoretical view is that comforting and helping are emotion-related prosocial behaviors assumed to be strongly related with empathy. Giving and sharing, in contrast, are two qualities of prosocial behavior that are not primarily evoked by emotional reactions but are associated with a perspective of fairness and equality, which implies a more sophisticated cognitive development (Hoffman, 2000). If, however, cultures value and promote the development of different kinds of prosocial behaviors, then this conceptualization can hardly bear up in a cross-cultural comparison (Trommsdorff and Kornadt, 2003).

Studying parenting and children's behavioral outcomes in a cross-cultural context represents a key to the examination of human prosociality in its various cultural expressions.

Parenting and Prosocial Behavior in a Cross-Cultural Context: An Integration

Theorists have suggested two possible nonexclusive effects of culture on parenting and prosocial behavior.

First, the parenting styles and practices have different meaning for children's prosocial development in different cultural contexts (Kağitçibaşi, 1996). The same behavior on the part of parents conveys different meanings to children in different cultures.

Although the above description of the Western-Protestant idea about the child's inborn nature as 'selfish' and the human 'need for autonomy' is overly schematic, it is used here to exemplify how cultural ideas about the human nature are shaping theoretical as well as applied approaches to parenting and prosocial development. For example, the notion that we are born selfish has implication for when and how we will expect prosocial behaviors to develop (Hoffman, 1975, 1982) and when we will begin to teach empathy and altruism as described earlier. On the other hand, the focus on the need for

autonomy specifies that one may not enforce rules and norms if one wants to be successful in making children compassionate and kind. The implicit idea that humans are born selfish and as autonomous individuals not only implies which will be the most appropriate parenting style, but also prescribes the kind of prosocial behaviors one may observe and the motivations underlying them, and it even can serve to explain and predict the developmental course of prosocial behavior (i.e., from a self-oriented to an other-oriented prosociality) we have been observing.

Taking culture into account can help explain when Western white middle class parents, who model prosocial behavior and explain why rules are necessary rather than impose them, have children who are more prosocial than parents who use power-assertive means of socialization (e.g., Krevans and Gibbs, 1996).

In contrast to what has been found to promote and explain prosocial behavior among Western middle class offspring, research has consistently shown that Chinese parents are directive and controlling in their parenting style, rooted in the belief that rigorous and responsible teaching is reflecting high involvement with and love for their children (e.g., Chao, 1994; Wu et al., 2002). Chao (1994) has explained two culturally based concepts of parental control: 'chiao shun' and 'guanc.' 'Chiao shun' means 'training children in appropriate behaviors' (Chao, 1994). 'Guan' literally means 'to govern.' 'Guan' has a very positive connotation for the Chinese because it also can mean 'to love' or 'to care for.' Children interpret parents' firm control as an expression of their parents' care and concern. This explains why Chinese children of parents with an authoritarian parenting style may develop to share at least as willingly as American children raised with a democratic parenting style (Rao and Stewart, 1999).

Mayan parents in Mexico allow their children to decide how much to eat and sleep, whether to take medicine, and whether to go to school (Greenfield, 1999), which would be classifiable as indulgent or permissive parenting style according to Baumrind's (1971) typology, but according to Mosier and Rogoff (2003), this does not result in the above-mentioned confusion regarding societies' norms and expectations. Already by the age of 5 years, Mayan Mexican children share and care for their younger siblings, something that European-American kindergarteners are developmentally not even expected to do, and they do so in the absence of the guiding eyes of their mothers.

Thus, parental beliefs about the nature of development seem to be both the foundation and the initiator of many parent–child outcomes and they play an important role in explaining parental socialization practices and outcomes (Harkness and Super, 2006). Understanding parents' ethnotheories about their children is key to understanding the strategies parents use to help their children grow up to become successful members of their communities.

The second effect of culture on the relationship between parenting and prosocial behavior may be that cultural differences in parental socialization are the means by which cultural differences in prosocial tendencies arise (Knight et al., 1982; Whiting and Whiting, 1973). Cross-cultural perspectives on development assume that cultural emphases on particular values and socialization toward these values have an accelerating influence on the development in specific domains (Greenfield et al., 2003; Keller, 2007; Rothbaum et al., 2000b). It has been reported that in some cultures such as in West Africa, prosocial behavior is encouraged as early as infancy (e.g., infants are offered objects and then encouraged to return the gifts) to foster sharing and exchange norms believed to bind the social group together (Nsamenang, 1992).

Westerners may value prosocial acts that appear to be based on endogenous motivation more than do people from traditional cultures, whereas people from traditional cultures value prosocial actions that reflect responsiveness to others' stated needs and reciprocal obligations (Jacobsen, 1983), and this may result in different motivations and forms of prosocial behaviors.

Culture has not necessarily an influence on the existence of universal psychological and developmental phenomena but it influences their expression. Keller and Kaertner (2012) explain "Human behavior and experience are profoundly influenced by species-specific, panhuman psychological mechanisms that evolved as adaptations to solve particular challenges that our ancestors faced" (p. 66).

The profound human capacity to adapt to environmental challenges and constraints implies that development cannot be universal across different contexts. Moreover, as illustrated in the case of prosocial behavior, universal psychological processes are insufficient to explain the variety of forms in which human prosociality can be encountered. Culture is the framework that may allow us to predict the culture-specific expression of this 'panhuman' relationship.

See also: Behavioral Problems, Effects of Parenting and Family Structure on; Emotional Development, Effects of Parenting and Family Structure on; Empathy During Early Childhood Across Cultures, Development of; Empathy: A Social Neuroscience Perspective; Human Development, Theories of; Parenting Attitudes and Beliefs across Cultures; Prosocial Behavior During Childhood and Cultural Variations; Socialization in Infancy and Childhood.

Bibliography

Baumrind, D., 1971. Current patterns of parental authority. Developmental Psychology 4 (1, Pt. 2), 1–103. http://dx.doi.org/10.1037/h0030372.

Benenson, J.F., Pascoe, J., Radmore, N., 2007. Children's altruistic behavior in the dictator game. Evolution and Human Behavior 28 (3), 168–175. http://dx.doi.org/10.1016/j.evolhumbehav.2006.10.003.

Berry, J.W., 1976. Cross-Cultural Research and Methodology Series: III. Human Ecology and Cognitive Style: Comparative Studies in Cultural and Psychological Adaptation. Sage, Oxford, England.

Boesch, E.E., 1991. Symbolic Action Theory and Cultural Psychology. Springer-Verlag Publishing, New York, US.

Chao, R.K., 1994. Beyond parental control and authoritarian parenting style: understanding Chinese parenting through the cultural notion of training. Child Development 65 (4), 1111–1119. http://dx.doi.org/10.2307/1131308.

Conroy, M., Hess, R.D., Azuma, H., Kashiwagi, K., 1980. Maternal strategies for regulating children's behavior: Japanese and American families. Journal of Cross-cultural Psychology 11 (2), 153–172. http://dx.doi.org/10.1177/0022022180112002.

Deci, E.L., Ryan, R.M., 1985. Self-determination. Wiley Online Library.
Eisenberg, N., Fabes, R.A., 1998. Prosocial development handbook of child psychology. In: Social, Emotional, and Personality Development, fifth ed., vol. 3. John Wiley & Sons Inc., Hoboken, NJ, US. pp. 701–778.
Eisenberg, N., Fabes, R.A., Spinrad, T.L., 2006. Prosocial development. In: Eisenberg, N., Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Social, Emotional, and Personality Development, sixth ed., vol. 3. John Wiley & Sons Inc., Hoboken, NJ, US, pp. 646–718.
García, F., Gracia, E., 2009. Is always authoritative the optimum parenting style? Evidence from Spanish families. Adolescence 44 (173), 101–131.
Goodnow, J.J., Collins, W.A., 1990. Development According to Parents: the Nature, Sources, and Consequences of Parents' Ideas. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, England.
Greenfield, P.M., 1999. Historical change and cognitive change: a two-decade follow-up study in Zinacantan, a Maya community in Chiapas, Mexico. Mind, Culture, and Activity 6 (2), 92–108. http://dx.doi.org/10.1080/10749039909524717.
Greenfield, P.M., Keller, H., Fuligni, A., Maynard, A., 2003. Cultural pathways through universal development. Annual Review of Psychology 54, 461–490. http://dx.doi.org/10.1146/annurev.psych.54.101601.145221.
Grolnick, W.S., Deci, E.L., Ryan, R.M., 1997. Internalization within the family: the self-determination theory perspective. In: Grusec, J.E., Kuczynski, L. (Eds.), Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory. John Wiley & Sons Inc., Hoboken, NJ, US, pp. 135–161.
Grusec, J.E., Goodnow, J.J., 1994. Impact of parental discipline methods on the child's internalization of values: a reconceptualization of current points of view. Developmental Psychology 30 (1), 4–19. http://dx.doi.org/10.1037/0012-1649.30.1.4.
Harkness, S., Super, C.M., 2006. Themes and variations: parental ethnotheories in Western cultures. In: Rubin, K., Chung, O.B. (Eds.), Parenting Beliefs, Behaviors, and Parent-Child Relations: a Cross-cultural Perspective. Psychology Press, New York, US, pp. 61–79.
Harkness, S., Super, C.M., Axia, V., Eliasz, A., Palacios, J., Welles-Nyström, B., 2001. Cultural pathways to successful parenting. International Society for the Study of Behavior Development Newsletter 38, 9–11.
Hastings, P.D., Utendale, W.T., Sullivan, C., 2007. The socialization of prosocial development. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization: Theory and Research. Guilford Press, New York, US, pp. 638–664.
Hastings, P.D., Zahn-Waxler, C., Robinson, J., Usher, B., Bridges, D., 2000. The development of concern for others in children with behavior problems. Developmental Psychology 36 (5), 531–546. http://dx.doi.org/10.1037/0012-1649.36.5.531.
Heine, S.J., Lehman, D.R., Markus, H.R., Kitayama, S., 1999. Is there a universal need for positive self-regard? Psychological Review 106 (4), 766–794. http://dx.doi.org/10.1037/0033-295X.106.4.766.
Hess, R.D., Kashiwagi, K., Azuma, H., Price, G.G., Dickson, W.P., 1980. Maternal expectations for mastery of developmental tasks in Japan and the United States. International Journal of Psychology 15 (4), 259–271. http://dx.doi.org/10.1080/00207598008246996.
Hoffman, M.L., 1970. Conscience, personality, and socialization techniques. Human Development 13 (2), 90–126. http://dx.doi.org/10.1159/000270884.
Kağitçibaşi, Ç., 1996. Cross-cultural psychology and development. In: Pandey, J., Sinha, D., Bhawuk, D.P.S. (Eds.), Asian Contributions to Cross-cultural Psychology. Sage Publications, Inc., Thousand Oaks, CA, US, pp. 42–49.
Keller, H., 2007. Cultures of Infancy. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.
Keller, H., Kärtner, J., Gelfand, M., Chiu, C., Hong, Y. Development–The Cultural Solution of Universal Developmental Tasks, vol. 3. Oxford University Press, NY, in press.
Keller, H., Voelker, S., Yovsi, R.D., 2005. Conceptions of parenting in different cultural communities: the case of West African Nso and Northern German women. Social Development 14 (1), 158–180. http://dx.doi.org/10.1111/j.1467-9507.2005.00295.x.
Keller, H., Yovsi, R.D., Voelker, S., 2002. The role of motor stimulation in parental ethnotheories: the case of Cameroonian Nso and German women. Journal of Cross-cultural Psychology 33 (4), 398–414. http://dx.doi.org/10.1177/00222102033004003.
Knafo, A., Schwartz, S.H., Levine, R.V., 2009. Helping strangers is lower in embedded cultures. Journal of Cross-cultural Psychology 40 (5), 875–879. http://dx.doi.org/10.1177/0022022109339211.
Knight, G.P., Kagan, S., Buriel, R., 1982. Perceived parental practices and prosocial development. The Journal of Genetic Psychology: Research and Theory on Human Development 141 (1), 57–65.
Krevans, J., Gibbs, J.C., 1996. Parents' use of inductive discipline: relations to children's empathy and prosocial behavior. Child Development 67 (6), 3263–3277.
LeVine, R.A., White, M., 1987. Parenthood in social transformation. In: Lancaster, J.B., Altmann, J., Rossi, A.S., Sherrod, L.R. (Eds.), Parenting across the Life Span: Biosocial Dimensions. Aldine Publishing Co., Hawthorne, NY, US, pp. 271–293.
Maccoby, E.E., Martin, J.A., 1983. Socialization in the context of the family: parent-child interaction. Handbook of Child Psychology 4, 1–101.
Matsumoto, D., Juang, L., 2004. Culture and Psychology, third ed. Wadsworth/Thomson Learning, Belmont, CA, US.
Miller, J.G., Bersoff, D.M., 1994. Cultural influences on the moral status of reciprocity and the discounting of endogenous motivation. Personality and Social Psychology Bulletin 20 (5), 592–602. http://dx.doi.org/10.1177/0146167294205015.
Miller, P.J., Goodnow, J.J., 1995. Cultural practices: toward an integration of culture and development. In: Goodnow, J.J., Miller, P.J., Kessel, F. (Eds.), Cultural Practices as Contexts for Development. Jossey-Bass, San Francisco, CA, US, pp. 5–16.
Mosier, C.E., Rogoff, B., 2003. Privileged treatment of toddlers: cultural aspects of individual choice and responsibility. Developmental Psychology 39 (6), 1047–1060. http://dx.doi.org/10.1037/0012-1649.39.6.1047.
Norenzayan, A., Heine, S.J., 2005. Psychological universals: what are they and how can we know? Psychological Bulletin 131 (5), 763–784. http://dx.doi.org/10.1037/0033-2909.131.5.763.
Nsamenang, B.A., 1992. Perceptions of Parenting among the Nso of Cameroon Father–Child Relations: Cultural and Biosocial Contexts. Aldine de Gruyter, Hawthorne, NY, US, 321–343.
Park, S.-Y., Trommsdorff, G., Lee, E.G., 2012. Korean Mothers' Intuitive Theories Regarding Emotion Socialization of Their Children. Bibliothek der Universität Konstanz.
Rao, N., Stewart, S.M., 1999. Cultural influences on sharer and recipient behavior: sharing in Chinese and Indian preschool children. Journal of Cross-cultural Psychology 30 (2), 219–241. http://dx.doi.org/10.1177/0022022199030002005.
Raval, V.V., Martini, T.S., 2009. Maternal socialization of children's anger, sadness, and physical pain in two communities in Gujarat, India. International Journal of Behavioral Development 33 (3), 215–229. http://dx.doi.org/10.1177/0165025408098022.
Rothbaum, F., Morelli, G., Pott, M., Liu-Constant, Y., 2000a. Immigrant-Chinese and Euro-American parents' physical closeness with young children: themes of family relatedness. Journal of Family Psychology 14 (3), 334–348. http://dx.doi.org/10.1037/0893-3200.14.3.334.
Rothbaum, F., Pott, M., Azuma, H., Miyake, K., Weisz, J., 2000b. Trade-offs in the study of culture and development: theories, methods, and values. Child Development 71 (5), 1159–1161. http://dx.doi.org/10.1111/1467-8624.00220.
Roth-Hanania, R., Davidov, M., Zahn-Waxler, C., 2011. Empathy development from 8 to 16 months: early signs of concern for others. Infant Behavior & Development 34 (3), 447–458. http://dx.doi.org/10.1016/j.infbeh.2011.04.007.
Sears, R.R., Maccoby, E.E., Levin, H., 1957. Patterns of Child Rearing. Row, Peterson and Co., Oxford, England.
Steinberg, L., Lamborn, S.D., Darling, N., Mounts, N.S., Dornbusch, S.M., 1994. Over-time changes in adjustment and competence among adolescents from authoritative, authoritarian, indulgent, and neglectful families. Child Development 65 (3), 754–770. http://dx.doi.org/10.2307/1131416.
Trommsdorff, G., Friedlmeier, W., Mayer, B., 2007. Sympathy, distress, and prosocial behavior of preschool children in four cultures. International Journal of Behavioral Development 31 (3), 284–293. http://dx.doi.org/10.1177/0165025407076441.
Trommsdorff, G., Kornadt, H.-J., 2003. Parent-Child Relations in Cross-cultural Perspective. Bibliothek der Universität Konstanz.
Whiting, J.W.M., Whiting, B.B., 1973. Altruistic and egoistic behavior in six cultures. Cultural Illness and Health, 56–66.
Wu, P., Robinson, C.C., Yang, C., Hart, C.H., Olsen, S.F., Porter, C.L., Wu, X., 2002. Similarities and differences in mothers' parenting of preschoolers in China and the United States. International Journal of Behavioral Development 26 (6), 481–491. http://dx.doi.org/10.1080/01650250143000436.
Xu, A., Xie, X., Liu, W., Xia, Y., Liu, D., 2007. Chinese family strengths and resiliency. Marriage & Family Review 41 (1–2), 143–164. http://dx.doi.org/10.1300/J002v41n01_08.
Yovsi, R.D., Kärtner, J., Keller, H., Lohaus, A., 2009. Maternal interactional quality in two cultural environments: German middle class and Cameroonian rural mothers. Journal of Cross-cultural Psychology 40 (4), 701–707. http://dx.doi.org/10.1177/0022022109335065.
Zahn-Waxler, C., Friedman, R.J., Cole, P.M., Mizuta, I., Hiruma, N., 1996. Japanese and United States preschool children's responses to conflict and distress. Child Development 67 (5), 2462–2477. http://dx.doi.org/10.2307/1131634.

Prosopagnosia

Beatrice de Gelder, Brain and Emotion Laboratory Leuven (BELL), Leuven, Belgium; Faculty of Psychology and Neuroscience, Maastricht Brain Imaging Centre M-BIC, Maastricht University, Maastricht, The Netherlands; and Cognitive and Affective Neuroscience Laboratory, CoRPS–Center of Research on Psychology in Somatic diseases, Tilburg University, Tilburg, The Netherlands
Jan Van den Stock, Brain and Emotion Laboratory Leuven (BELL), Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.W. Young, volume 18, pp. 12233–12236, © 2001, Elsevier Ltd.

Abstract

Prosopagnosia refers to impaired recognition of faces. The article first describes the different etiological factors and clinical phenotypes of prosopagnosia: developmental and acquired prosopagnosia for the former and apperceptive, associative, amnestic, and multimodal prosopagnosia for the latter. This includes a review of the face and function specificity of the deficit. Subsequently, the results from neuroimaging studies are discussed. The article ends with a paragraph on neurodegeneration and prosopagnosia.

Introduction

Faces are attention-getters, and this is one of the main reasons why, in social interaction and in our thinking about it, the face occupies center stage. There may also be other, quantitative, reasons why the face plays such a central role. Indeed, very few, if any, natural objects carry as much information as does the face. Whether an observer needs to know identity or gender, is wondering about age or attractiveness, is listening to speech, or is making an evaluation of trustworthiness, the gaze turns to the other's face. If only for these reasons, one understands that there is more to learn from the face than from any other source of social signals.

Two assumptions are underlying the many claims about face specificity that are at the forefront of discussions on prosopagnosia: (1) that faces trigger perceptual processes, which have category-specific markers, and (2) that faces occupy a neurofunctional niche of their own, such that neural face representation coexists with (but does not overlap with) object representation – a view that in one sense or another, it is linked to modularity or more generally to the notion of special functional units in the brain. Clinical cases constitute critical tests for theoretical models, and patients suffering from deficits in face recognition (Bodamer, 1947) have long served as an important touchstone for models of face processing.

Acquired and Developmental Prosopagnosia

Imagine that if every time you encounter the face of an acquaintance, it seemed no more familiar to you than that of a total stranger. This peculiar symptom is known as prosopagnosia – and not surprisingly, it takes a heavy toll on the social life of the patients who suffer from it. Prosopagnosics can even have difficulties recognizing people whom they are very close to, such as immediate family members, and they have to rely on other cues to identify a person – clothing, for instance, or voice. In less-radical cases, prosopagnosics are helped by the context in which they routinely encounter individuals, but they fail to recognize them when they encounter them outside the usual context – for example, when they meet a colleague in the market.

Prosopagnosia may occur after a neurological antecedent, in which case it is known as acquired prosopagnosia (AP), and is often associated with lesions in the occipito-temporal brain regions, especially in the right hemisphere (Barton et al., 2002; Damasio et al., 1982; Landis et al., 1986; Levine and Calvanio, 1989; Meadows, 1974; Sergent and Signoret, 1992; Marotta et al., 2001; Sorger et al., 2007; Steeves et al., 2006; de Gelder et al., 2003) or following a head trauma but without evident brain lesions showing up on standard imaging (Hadjikhani and de Gelder, 2003; de Gelder et al., 2003).

Prosopagnosia can also be present without any neurological history. In this case, it is known as developmental prosopagnosia (DP). However, this does not imply the absence of structural brain abnormalities. For instance, there are strong indications that congenital hypoplasia of the cerebellar vermis can be associated with face recognition difficulties (Van den Stock et al., 2012b). Patients with DP usually report lifelong problems in recognizing people by the face. DP sometimes occurs in different generations of a single family, suggesting a genetic component, more particularly an autosomal dominant mode of inheritance (Grüter et al., 2007). The term 'congenital prosopagnosia' is sometimes used to refer to prosopagnosia without any neurological antecedents. However, the congenital nature of DP is a matter of debate, and there is no conclusive evidence that all of these cases have a genetic basis. The term 'developmental prosopagnosia' has a more descriptive connotation and stresses (next to the genetic hypothesis) the possibility of a developmental disorder – for instance, in the maturation and structural connectivity of different cortical areas, such as the fusiform gyrus (Behrmann et al., 2007; Thomas et al., 2009). The pattern of deficits in AP and DP is similar as far as some basic aspects of face processing are concerned, but different in other ways (de Gelder and Rouw, 2000a).

The prosopagnosia symptom centers on recognition of personal identity but not of facial expression, and such

a dissociation has long been the cornerstone of the models of face processing in the neuropsychological literature of the last two decades, and is at the basis of the face recognition model of Bruce and Young (1986). Until recently, this dissociation was generally accepted – and as a consequence, researchers rarely investigated facial expression recognition itself, nor was much attention paid to the possibility of an interaction between the face identity and the facial expression recognition system. When problems with faces are defined as problems with identity recognition, then understanding face recognition may or may not be seen as a subspecies of object recognition. Substantial effort has gone into finding evidence for dissociations between face- and object recognition. In the 1980s and 1990s, the gold standard for supporting this kind of special or modular cognitive ability argument was to find a double dissociation. Patients with object-recognition deficits that were not prosopagnosic (as opposed to prosopagnosics without object-recognition problems) perfectly fit this bill. However such pure cases have been proved extremely rare.

Indeed, while a number of cases of prosopagnosia have been reported over the last hundred years, only very few provide evidence for pure or genuine prosopagnosia, affecting face perception in the complete absence of any other object perception deficit. Regarding the question whether prosopagnosia is exclusively a face-identity recognition deficit or whether the face recognition deficit is simply the most dominant deficit, the debate has not been concluded yet. As shown in a well-known review of the literature published at the end of the 1980s, almost all of the reported cases have associated deficits in object recognition (Farah, 1990). Lively debates will undoubtedly continue as long as new ways of testing face specificity are being developed that reflect increasingly better understanding of face recognition.

Face and Function Specificity of Prosopagnosia

Standard procedures in prosopagnosia assessment must include at least a basic screening of the visual object-recognition abilities with standard clinical test batteries such as the Birmingham Object Recognition Battery (Riddoch and Humphreys, 1993) or the Visual Object and Space Perception Battery (Warrington and James, 1991). It is not uncommon that prosopagnosics score outside of the normal range on at least one of the tests (e.g., Duchaine, 2000). Standard clinical tests for evaluating face identity matching and memory of face identity include the Benton Facial Recognition Test (Benton et al., 1983) and the Warrington Recognition Memory Test (Warrington, 1984). The scores on these tests should thus be interpreted cautiously, especially when no data are available on reaction time (as is the case for the older studies).

More modern tests and batteries are available now, targeting specific aspects of face- and object processing (de Gelder et al., 1998; de Gelder and Rouw, 2000b; Duchaine and Nakayama, 2006; de Gelder and Van den Stock, 2011; de Gelder and Bertelson, 2009). These provide a more detailed analysis of the intact and anomalous aspects of face- and object processing. The Face Emotion Action Stimulus Test (de Gelder and Van den Stock, in preparation), for example, targets a range of specific abilities including face detection, emotional face identity memory, neutral face identity memory, facial emotion matching, facial identity matching, object identity matching, facial part-to-whole matching, object part-to-whole matching, and configural face and object processing.

Since normal face processing involves multiple hierarchical and parallel processes, impairments in different processes will result in different types of behavioral and neuroanatomical correlates. In developmental disorders like DP, heterogeneity is rather the rule than the exception and this is also reflected in the behavioral findings. Extensive evaluations of single cases or small groups of prosopagnosics have found different patterns of impairment on multiple aspects of face processing, including configural processing, modulation of the N170, within-object spatial relations, recognition of emotions, gender discrimination, recognition of famous faces, holistic face processing, and facial attractiveness ratings.

There have been some attempts at making a taxonomy of the heterogeneity in clinical phenotypes associated with prosopagnosia. An older and often ignored effort at getting underneath the surface of a behavioral deficit was to understand prosopagnosia as a form of agnosia using a historical distinction between two types of agnosia, described by Lissauer (1890): (1) apperceptive agnosia reflects impairment in consciously perceiving and recognizing stimuli and (2) associative agnosia refers to a relatively preserved ability to perceive objects, but an inability to interpret what is seen or match it with its associated meaning. Apperceptive prosopagnosia therefore relates to an inability to perceive a face as an invariant face and patients suffering from this type will be mostly impaired in tasks tapping into face *detection* and face *discrimination*; whereas, associative prosopagnosia is in the first place associated with the recognition of individual faces and involves a deficit in the kind of encoding that allows telling individual exemplars apart and accessing stored representations. It is the latter type that has been mainly reported and investigated as it is assumed that prosopagnosia arises at a higher-order recognition and cognitive level. Questions about intact face detection typically are not raised in prosopagnosia and intact face detection is typically not seen as possibly a separate and earlier stage in face perception (but see de Gelder and Rouw, 2000a; de Gelder et al., 2003; Righart et al., 2010). Formal neuropsychological evaluation of associative prosopagnosia requires the use of 'semantically unique items,' i.e., stimulus exemplars that carry stimulus-specific semantic associations. In other words, the stimulus needs to have unique semantic properties that are not present in any other exemplar of the same stimulus category. Typically, famous faces figure as the preferred stimulus category for this purpose, although idiosyncratically familiar (Van den Stock et al., 2012a) or experimentally familiarized (Stollhoff et al., 2011) faces have also been used.

A less known third subtype has been proposed: amnestic prosopagnosia or prosopamnesia, which refers to the impaired ability to learn new faces. In case of acquired face recognition deficits, diagnostic criteria for a 'pure' variant would include preserved face perception, preserved memory for nonface visual material, and preserved recognition of premorbidly familiar faces (Tippett et al., 2000); whereas, for the developmental variant, the deficit would consist of abnormal processing of repeated unfamiliar faces, in combination with normal processing of familiar faces and familiar and unfamiliar nonface

visual materials (Williams et al., 2007). However, little is known about face familiarization and there is limited evidence for a clear distinction between associative and amnestic DP.

Finally, associative and apperceptive prosopagnosia are to be distinguished from impaired person recognition. There is anecdotal evidence of person recognition deficits that are not face selective and extend to voices and names (Tyrrell et al., 1990; Evans et al., 1995; Gentileschi et al., 2001; Gainotti et al., 2003). The modality specificity of the person recognition deficit would be among the main differentiating characteristics between prosopagnosia and person recognition symptoms. A multimodal person recognition deficit argues against a pure prosopagnosia. However, the majority of the studies on face recognition difficulties have only investigated the visual modality (Gainotti, 2013). In addition, cases of phonagnosia have been reported, showing impaired voice recognition, which may or may not be modality specific (Garrido et al., 2009; Hailstone et al., 2010; Van Lancker and Canter, 1982; Van Lancker et al., 1988; Van Lancker et al., 1989).

Functional Imaging Findings

The major focus of face perception research continues to be the so-called configuration-based processing or configural processing in short (Tarr, this volume). Configural processing refers to the mechanism of object perception that goes beyond the mere analysis of features or parts, but includes the analysis of spatial relations between the constituting features or parts of the object. Configural processing is most often measured by the inversion effect, which counts as the signature of intact face processing skills and it is commonly tested for in case of face recognition deficits. A few studies have reported that when a face perception disorder is present, the person or the patient does not show the normal inversion effect (de Gelder and Rouw, 2000a), but this phenomenon is also still a matter of debate (Huis in 't Veld et al., 2012). What is less clear is whether an anomaly in perceiving the face configuration and in showing the inversion effect, when confronted with neutral faces, does in fact generalize to faces where it is not the identity but the expression that matters. The available fMRI studies focusing on face perception in prosopagnosics have revealed inconsistent results (see Van den Stock et al., 2008 for an overview). In the first fMRI study with prosopagnosics, we studied activation in the occipital face area (OFA) and fusiform face area (FFA) of three prosopagnosia patients. None of them showed the normal higher activity to faces compared to objects in either region (or elsewhere). Furthermore, in two patients, there was similar activation for faces and objects in OFA and FFA (see Figure 1).

Few studies included the emotional dimension provided by a face. We investigated whether adding an emotional expression would normalize the face processing style of prosopagnosics with respect to the inversion effect. We presented neutral and emotional faces to patients with AP with lesions in

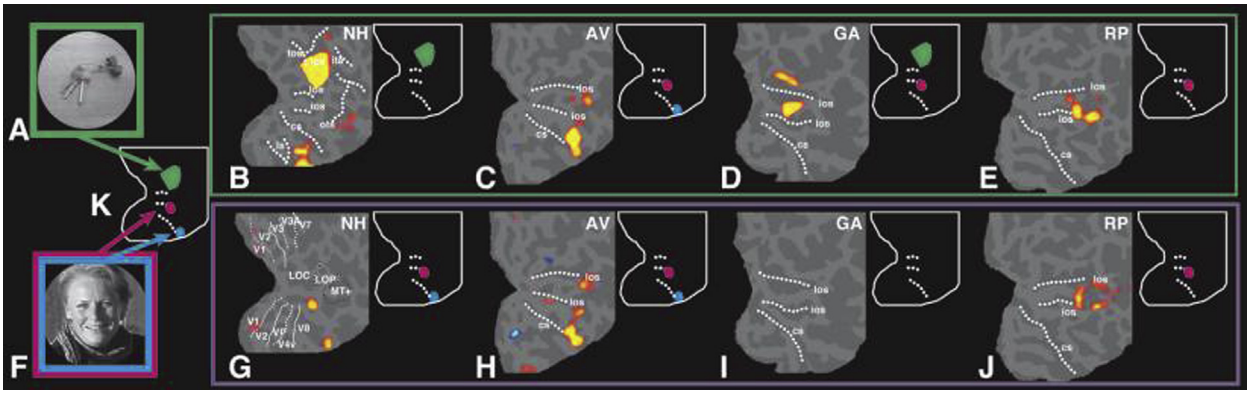

Figure 1 Activation obtained for objects (B–E) and faces (G–J) compared to control stimuli in the right hemisphere of a normal subject (B, G) and in three prosopagnosic patients (C–E, H–J). Data are represented on a flattened representation of the occipital cortex, with gyri appearing in light gray, whereas sulci are in darker gray. The dotted lines in (B) represent the major sulci (tos, transverse occipital sulcus; ips, intraparietal sulcus; its, inferior temporal sulcus; ots, occipito-temporal sulcus; los, lateral occipital sulcus; ios, inferior occipital sulcus; cs, collateral sulcus; ls, lingual sulcus). G shows the location of the different visual areas obtained by retinotopic mapping. (A) Object and (F) face show an example of the stimuli used in the experiment, and (K) is a diagram of the activation obtained for objects (in green) compared to scrambled objects and faces compared to scrambled faces (in blue and magenta) in a normal representative subject. On the upper right part of each panel, a diagram indicates the areas activated in each subject in comparison with the normal subject (K). In normal subjects, the viewing of objects compared to scrambled objects (B) elicited activation principally in the lateral occipital cortex (LO, in green in the diagram), whereas faces compared to scrambled faces activated two areas: the anterior part of the collateral sulcus and fusiform gyrus (fusiform face area (FFA), in blue) and the inferior occipital gyrus and sulcus (inferior occipital gyrus (IOG), in magenta). In patient AV, the developmental prosopagnosic, viewing both objects compared to houses (C) and faces compared to houses (H) activated the same region of the cortex, in the anterior part of the collateral sulcus and the fusiform gyrus (FFA). Objects did not produce activation in the expected LO region, and faces failed to activate the IOG area. In patient GA, injured at 18 months, faces failed to elicit any activation in the occipital cortex compared to scrambled faces (I), whereas objects compared to scrambled objects activated both the LO and the IOG area (D). Finally, patient RP, injured at 7 years, showed similar activation to both faces compared to houses (J) and objects compared to houses (E), situated in the IOG region. No activation was seen in the FFA area. None of the prosopagnosic patients showed activation in both the FFA and the IOG to face presentation. Instead they showed activation in either one or the other, or none at all.

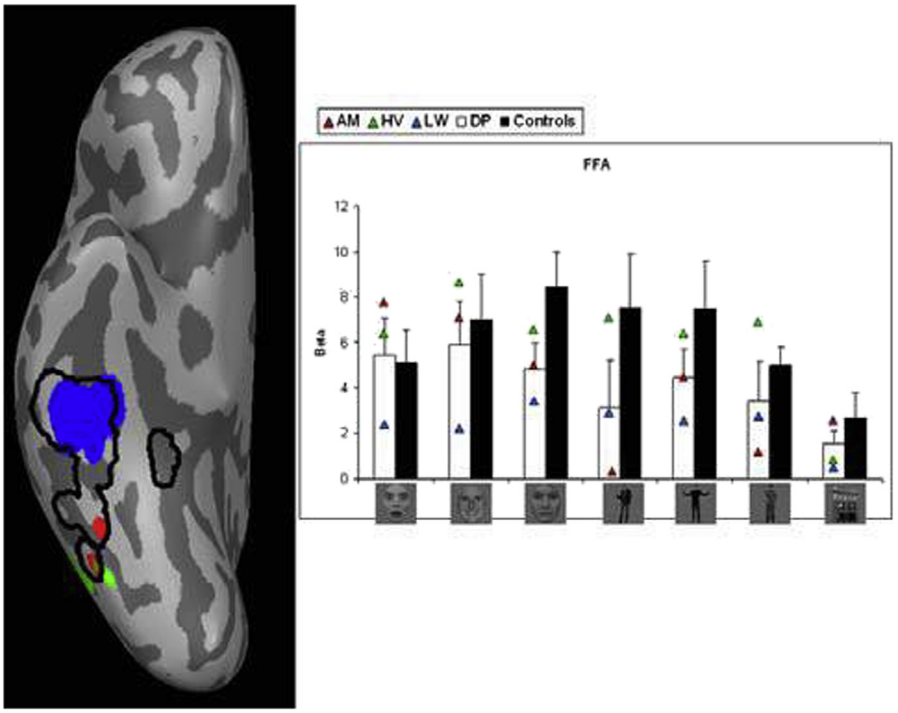

Figure 2 Face-specific activation in the right fusiform gyrus when comparing faces (fearful/happy/neutral) with houses. Left: Areas are shown on an inflated right hemisphere. Activation maps of the control subjects are collapsed and displayed by the black contours. Activation of the individual developmental prosopagnosia subjects (DPs) is plotted in color. Right: beta values by condition, group, and DP subject. Error bars represent one standard error of the mean. Conditions represent from left to right: fearful faces, happy faces, neutral faces, fearful bodies, happy bodies, neutral bodies, and houses. White columns display the average value of the three patients. Black columns show the average value of the controls. Triangles represent the individual values of the DPs.

FFA, inferior occipital gyrus, or both. The results showed that the patients more strongly activated other face-sensitive areas like the superior temporal sulcus or amygdala when they perceive facial expressions (de Gelder et al., 2003). Furthermore, they were more accurate and faster in processing emotional faces compared to neutral faces, consistent with other studies (Duchaine et al., 2003; Jones and Tranel, 2001; Nunn et al., 2001). Since the patients in de Gelder et al. (2003) had lesions in the ventral occipito-temporal cortex, we wondered how these brain areas respond to emotional information in DP patients. Therefore, we presented a group of DPs and matched controls with fearful, happy, and neutral faces (and bodies). The results showed normal activation in FFA for emotional faces (fearful and happy), but lower activation for neutral faces, compared to controls (Van den Stock et al., 2008) (see **Figure 2**).

Prosopagnosia and Neurodegeneration

Neurodevelopmental syndromes can reveal important insights into normal cognitive functioning as selective deficits may arise in the course of neurodevelopmental syndromes, including neurodegeneration. In addition, the change in symptomatology associated with increased regional brain atrophy following progression of the neurodegenerative disease process can provide information regarding the regional involvement in cognitive functioning. While the results from studies targeting face perception in, for instance, autism spectrum disorders are heterogeneous (Barton et al., 2004), there have been recent significant advances in the characterization and categorization of dementia syndromes, for instance, concerning both language (Gorno-Tempini et al., 2011) and behavioral (Rascovsky et al., 2011) variants of frontotemporal lobar degeneration. The behavioral and neuroanatomical findings from neurodegenerative syndromes grossly mirror the apperceptive-associative-amnestic-multimodal classification of prosopagnosia. For instance, there is evidence that the neurodegenerative syndrome, known as 'posterior cortical atrophy,' is associated with apperceptive prosopagnosia and neurodegeneration of primarily occipito-temporal regions (Meek et al., 2013). There is consensus that semantic dementia, a syndrome characterized by progressive semantic deficits and anterior temporal lobe abnormalities, is associated with associative prosopagnosia (Snowden et al., 2004). A recent structural neuroimaging study in a group of neurodegenerative patients with progressive atrophy of primarily the frontal and temporal lobes provided additional evidence for the role of the anterior temporal lobes in famous face recognition: the results revealed an association between left anterior temporal lobe atrophy and famous face naming, whereas famous face recognition was related to bilateral anterior temporal lobe atrophy (Gefen et al., 2013). These results are extended by the selective progressive prosopagnosia symptomatology of an anatomical

subtype of frontotemporal neurodegeneration, characterized by predominantly right temporal lobe abnormalities as evidenced by several case reports (Evans et al., 1995; Gainotti et al., 2003; Joubert et al., 2003; Grossi et al., 2012).

However, there have been an increasing number of case reports on atypical presentations of neurodegenerative syndromes, including face perception symptomatology. For instance, we recently reported two atypical cases of Alzheimer's disease. One patient (MV) showed a mirror sign (inability to recognize one's own image in the mirror) (Van den Stock et al., 2012) and her performance on in-house custom-designed face and object processing tests was indicative of face-selective hyperfamiliarity processing with relatively intact identity recognition. The second patient (IS) displayed a particular delusion: she carried around a toy doll, of which she was convinced it was her own living grandchild (Van den Stock et al., 2013) and her neuropsychological face and object processing abilities also indicated face-selective hyperfamiliarity processing but also impaired identity recognition. These findings illustrate the heterogeneity in clinical phenotypes regarding different aspects of face and object processing.

As there is an increasing body of evidence on reports of prosopagnosia, in both scientific and nonscientific media (Della Sala and Schweinberger, 2013), the future decade holds promise regarding a better understanding of normal and abnormal face perception. But this will require putting the debate on face perception in the broader context of object recognition on the one hand, and of affect perception on the other. It is fair to conclude that for the time being, specialized studies on the localization of one, two, or more face areas in the brain have not yet moved forward our understanding of prosopagnosia.

Bibliography

Bodamer, J., 1947. Die prosop-Agnosie. Archiv fur Psychiatrie und Nervenkrankheiten 179, 6–53.

Barton, J.J., Cherkasova, M.V., Hefter, R., Cox, T.A., O'connor, M., Manoach, D.S., 2004. Are patients with social developmental disorders prosopagnosic? Perceptual heterogeneity in the Asperger and socio-emotional processing disorders. Brain 127, 1706–1716.

Barton, J.J.S., Press, D.Z., Keenan, J.P., O'Connor, M., 2002. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. Neurology 58, 71–78.

Behrmann, M., Avidan, G., Gao, F., Black, S., 2007. Structural imaging reveals anatomical alterations in inferotemporal cortex in congenital prosopagnosia. Cerebral Cortex 17, 2354–2363.

Benton, A.L., Sivan, A.B., Hamsher, K., Varney, N.R., Spreen, O., 1983. Benton facial Recognition Test. Oxford University Press, New York.

Bruce, V., Young, A., 1986. Understanding face recognition. British Journal of Psychology 77, 305–327.

Damasio, A.R., Damasio, H., Van Hoesen, G.W., 1982. Prosopagnosia: anatomic basis and behavioral mechanisms. Neurology 32, 331–341.

Della Sala, S., Schweinberger, S.R., 2013. Face blindness and person misidentification in non-scientific parlance. Cortex 49, 2276–2280.

Duchaine, B., Nakayama, K., 2006. The Cambridge face memory test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia 44, 576–585.

Duchaine, B.C., 2000. Developmental prosopagnosia with normal configural processing. Neuroreport 11, 79–83.

Duchaine, B.C., Parker, H., Nakayama, K., 2003. Normal recognition of emotion in a prosopagnosic. Perception 32, 827–838.

Evans, J.J., Heggs, A.J., Antoun, N., Hodges, J.R., 1995. Progressive prosopagnosia associated with selective right temporal lobe atrophy. A new syndrome? Brain 118 (Pt 1), 1–13.

Farah, M.J., 1990. Visual Agnosia: Disorders of Object Vision and what They Tell Us about Normal Vision. MIT, Cambridge, MA.

Gainotti, G., 2013. Is the right anterior temporal variant of prosopagnosia a form of 'associative prosopagnosia' or a form of 'multimodal person recognition disorder'? Neuropsychology Review 23, 99–110.

Gainotti, G., Barbier, A., Marra, C., 2003. Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. Brain 126, 792–803.

Garrido, L., Eisner, F., Mcgettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: a selective deficit of vocal identity recognition. Neuropsychologia 47, 123–131.

Gefen, T., Wieneke, C., Martersteck, A., Whitney, K., Weintraub, S., Mesulam, M.M., Rogalski, E., 2013. Naming vs knowing faces in primary progressive aphasia: a tale of 2 hemispheres. Neurology 81, 658–664.

de Gelder, B., Bachoud-Levi, A.C., Degos, J.D., 1998. Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces. Vision Research 38, 2855–2861.

de Gelder, B., Bertelson, P., 2009. A comparative approach to testing face perception: face and object identification by adults in a simultaneous matching task. Psychologica Belgica 42, 177–190.

de Gelder, B., Frissen, I., Barton, J., Hadjikhani, N., 2003. A modulatory role for facial expressions in prosopagnosia. Proceedings of the National Academy of Sciences of the United States of America 100, 13105–13110.

de Gelder, B., Rouw, R., 2000a. Configural face processes in acquired and developmental prosopagnosia: evidence for two separate face systems? Neuroreport 11, 3145–3150.

de Gelder, B., Rouw, R., 2000b. Paradoxical configuration effects for faces and objects in prosopagnosia. Neuropsychologia 38, 1271–1279.

de Gelder, B., Van den Stock, J., 2011. The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. Frontiers in Psychology 2, 181.

de Gelder, B., Van den Stock, J., The facial Expressive action stimulus test (FEAST), in preparation.

Gentileschi, V., Sperber, S., Spinnler, H., 2001. Crossmodal agnosia for familiar people as a consequence of right infero polar temporal atrophy. Cognitive Neuropsychology 18, 439–463.

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M., Grossman, M., 2011. Classification of primary progressive aphasia and its variants. Neurology 76, 1006–1014.

Grossi, D., Soricelli, A., Ponari, M., Salvatore, E., Quarantelli, M., Prinster, A., Trojano, L., 2012. Structural connectivity in a single case of progressive prosopagnosia: the role of the right inferior longitudinal fasciculus. Cortex.

Grüter, M., Grüter, T., Bell, V., Horst, J., Laskowski, W., Sperling, K., Halligan, P.W., Ellis, H.D., Kennerknecht, I., 2007. Special issue: original article hereditary prosopagnosia: the first case series. Cortex 43, 734–749.

Hadjikhani, N., de Gelder, B., 2003. Seeing fearful body expressions activates the fusiform cortex and amygdala. Current Biologyogy 13, 2201–2205.

Hailstone, J.C., Crutch, S.J., Vestergaard, M.D., Patterson, R.D., Warren, J.D., 2010. Progressive associative phonagnosia: a neuropsychological analysis. Neuropsychologia 48, 1104–1114.

Huis in 't veld, E., Van den Stock, J., De Gelder, B., 2012. Configuration perception and face memory, and face context effects in developmental prosopagnosia. Cogn Neuropsychol 29, 464–481.

Jones, R.D., Tranel, D., 2001. Severe developmental prosopagnosia in a child with superior intellect. Journal of Clinical and Experimental Neuropsychologyogy 23, 265–273.

Joubert, S., Felician, O., Barbeau, E., Sontheimer, A., Barton, J.J., Ceccaldi, M., Poncet, M., 2003. Impaired configurational processing in a case of progressive prosopagnosia associated with predominant right temporal lobe atrophy. Brain 126, 2537–2550.

Landis, T., Cummings, J.L., Christen, L., Bogen, J.E., 1986. Are unilateral right posterior cerebral lesions sufficient to cause prosopagnosia? Clinical and radiological findings in six additional patients. Cortex: A Journal Devoted to the Study of the Nervous System and Behavior 22, 243–252.

Levine, D.N., Calvanio, R., 1989. Prosopagnosia: a defect in visual configural processing. Brain and Cognition 10, 149–170.

Lissauer, H., 1890. Ein Fall von Seelenblindheit nebst einem Beitrage zur Theorie derselben.[A case of visual agnosia with a contribution to theory.]. Archiv fuer Psychiatrie und Nervenkrankheiten 21, 222–270.

Marotta, J.J., Genovese, C.R., Behrmann, M., 2001. A functional MRI study of face recognition in patients with prosopagnosia. Neuroreport 12, 1581–1587.

Meadows, J.C., 1974. The anatomical basis of prosopagnosia. Journal of Neurology Neurosurgery and Psychiatry 37, 489–501.

Meek, B.P., Locheed, K., Lawrence-Dewar, J.M., Shelton, P., Marotta, J.J., 2013. Posterior cortical atrophy: an investigation of scan paths generated during face matching tasks. Frontiers Human Neuroscience 7, 309.

Nunn, J.A., Postma, P., Pearson, R., 2001. Developmental prosopagnosia: should it be taken at face value? Neurocase 7, 15–27.

Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J., Van Swieten, J.C., Seelaar, H., Dopper, E.G., Onyike, C.U., Hillis, A.E., Josephs, K.A., Boeve, B.F., Kertesz, A., Seeley, W.W., Rankin, K.P., Johnson, J.K., Gorno-Tempini, M.L., Rosen, H., Prioleau-Latham, C.E., Lee, A., Kipps, C.M., Lillo, P., Piguet, O., Rohrer, J.D., Rossor, M.N., Warren, J.D., Fox, N.C., Galasko, D., Salmon, D.P., Black, S.E., Mesulam, M., Weintraub, S., Dickerson, B.C., Diehl-Schmid, J., Pasquier, F., Deramecourt, V., Lebert, F., Pijnenburg, Y., Chow, T.W., Manes, F., Grafman, J., Cappa, S.F., Freedman, M., Grossman, M., Miller, B.L., 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 134, 2456–2477.

Riddoch, M.J., Humphreys, G.W., 1993. BORB: The Birmingham Object Recognition Battery. Erlbaum UK, Hove.

Righart, R., Andersson, F., Schwartz, S., Mayer, E., Vuilleumier, P., 2010. Top-down activation of fusiform cortex without seeing faces in prosopagnosia. Cerebral Cortex 20, 1878–1890.

Sergent, J., Signoret, J.L., 1992. Functional and anatomical decomposition of face processing: evidence from prosopagnosia and PET study of normal subjects. Philosophical Transactions of the Royal Society of London B Biological Sciences 335, 55–61 discussion 61–62.

Snowden, J.S., Thompson, J.C., Neary, D., 2004. Knowledge of famous faces and names in semantic dementia. Brain 127, 860–872.

Sorger, B., Goebel, R., Schiltz, C., Rossion, B., 2007. Understanding the functional neuroanatomy of acquired prosopagnosia. Neuroimage 35, 836–852.

Steeves, J.K., Culham, J.C., Duchaine, B.C., Pratesi, C.C., Valyear, K.F., Schindler, I., Humphrey, G.K., Milner, A.D., Goodale, M.A., 2006. The fusiform face area is not sufficient for face recognition: evidence from a patient with dense prosopagnosia and no occipital face area. Neuropsychologia 44, 594–609.

Stollhoff, R., Jost, J., Elze, T., Kennerknecht, I., 2011. Deficits in long-term recognition memory reveal dissociated subtypes in congenital prosopagnosia. PLoS One 6, e15702.

Thomas, C., Avidan, G., Humphreys, K., Jung, K.J., Gao, F., Behrmann, M., 2009. Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nature Neuroscience 12, 29–31.

Tippett, L.J., Miller, L.A., Farah, M.J., 2000. Prosopamnesia: a selective impairment in face learning. Cognitive Neuropsychology 17, 241–255.

Tyrrell, P.J., Warrington, E.K., Frackowiak, R.S., Rossor, M.N., 1990. Progressive degeneration of the right temporal lobe studied with positron emission tomography. Journal of Neurology Neurosurgery and Psychiatry 53, 1046–1050.

Van den Stock, J., de Gelder, B., De Winter, F.L., Van Laere, K., Vandenbulcke, M., 2012a. A strange face in the mirror. Face-selective self-misidentification in a patient with right lateralized occipito-temporal hypo-metabolism. Cortex 48, 1088–1090.

Van den Stock, J., Vandenbulcke, M., Zhu, Q., Hadjikhani, N., de Gelder, B., 2012b. Developmental prosopagnosia in a patient with hypoplasia of the vermis cerebelli. Neurology 78, 1700–1702.

Van den Stock, J., Van De Riet, W.A., Righart, R., de Gelder, B., 2008. Neural correlates of perceiving emotional faces and bodies in developmental prosopagnosia: an event-related fMRI-study. PLoS One 3, e3195.

Van den Stock, J., De Gelder, B., Van Laere, K., Vandenbulcke, M., 2013. Face-selective hyper-animacy and hyper-familiarity misperception in a patient with moderate Alzheimer's disease. J Neuropsychiatry Clin Neurosci 25, E52–E53.

Van Lancker, D.R., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain and Cognition 1, 185–195.

Van Lancker, D.R., Cummings, J.L., Kreiman, J., Dobkin, B.H., 1988. Phonagnosia: a dissociation between familiar and unfamiliar voices. Cortex 24, 195–209.

Van Lancker, D.R., Kreiman, J., Cummings, J., 1989. Voice perception deficits: neuroanatomical correlates of phonagnosia. Journal of Clinical and Experimental Neuropsychology 11, 665–674.

Warrington, E.K., 1984. Recognition Memory Test. Nelson, Windsor, NFER.

Warrington, E.K., James, M., 1991. The Visual Object and Space Perception Battery. Thames Valley Test Company, Bury St Edmunds, UK.

Williams, M.A., Berberovic, N., Mattingley, J.B., 2007. Abnormal FMRI adaptation to unfamiliar faces in a case of developmental prosopamnesia. Current Biology 17, 1259–1264.

Prosopography (Collective Biography)

Christophe Charle, CNRS, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12236–12241, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Prosopography, or collective biography, has developed considerably since the 1960s, expanding in the fields of modern and contemporary history, after the tool was invented and applied primarily in ancient and medieval history. The method's principle is simple: defining a population on the basis of one or several criteria and designing a relevant biographical questionnaire containing a range of variables or criteria, which serve to describe it in terms of social, private, public and/or cultural, ideological, or political dynamics. Once the documentation has been compiled, the data can be evaluated with the help of number of techniques, both quantitative and qualitative, including manual counting or computerized processing, statistical tables or factor analyses. The proliferation of studies in contemporary history and historical sociology using the prosopographic method is due to intellectual trends and techniques developed since about 1970. An overview of these is provided in this article.

Prosopography, or collective biography, has developed considerably since the 1960s, expanding in the fields of modern and contemporary history, after the tool was invented and applied primarily in ancient and medieval history. The method's principle is simple: defining a population on the basis of one or several criteria and designing a relevant biographical questionnaire containing a range of variables or criteria which serve to describe it in terms of social, private, public and/or cultural, ideological, or political dynamics. Once the documentation has been compiled, the data can be evaluated with the help of number of techniques, both quantitative and qualitative, including manual counting or computerized processing, statistical tables or factor analyses. The proliferation of studies in contemporary history and historical sociology using the prosopographic method is due to intellectual trends and techniques developed since about 1970. The concomitant decline of global statistical methods and of the intellectual impact of Marxism, and the desire, typical of societies focusing more and more on the individual, to come to grips with individual experience and the diversity of social patterns have inclined historians of recent times to emphasize social microhistory, which gives priority to prosopography or collective biography. Thanks to this technique, since the 1970s, historians have increasingly set their sights on gaining intimate knowledge of small problem-related constituencies. Some historians, such as Lawrence Stone, have criticized the method for its elitist and abstract view of society; others think that it is limited to specific privileged groups. In fact, the 1990s registered an extension to middle and lower strata of society and even a new comparative ambition, so that collective biography seems to have a fruitful prospect.

Prosopography, or collective biography, has developed considerably since the 1960s, expanding in the fields of modern and contemporary history, after the tool was invented and applied primarily in ancient and medieval history. The method's principle is simple: defining a population on the basis of one or several criteria and designing a relevant biographical questionnaire containing a range of variables or criteria, which serve to describe it in terms of social, private, public and/or cultural, ideological, or political dynamics, depending on the population under scrutiny and the questionnaire that is being used. In the words of Lawrence Stone: "Prosopography is the investigation of the common background characteristics of a group of actors in history by means of a collective study of their lives" (Stone, 1971: p. 46). Once the documentation has been compiled, which is the most time-consuming aspect of the work, the data can be evaluated with the help of a number of techniques, both quantitative and qualitative, including manual counting or computerized processing, statistical tables or factor analyses, depending on the scope and sophistication of the questionnaire and the sample.

To appreciate the success of this method and demonstrate its extension to different branches of history, several historiographic references are cited in Section Historiographic References. To gain insight into its merits and limitations, it is necessary to analyze a few examples of its application, taken primarily from recent periods see Section Merits and Limitations where other social history methods can be used, permitting us to weigh the advantages and disadvantages of both alternatives. An attempt is made to define the heretofore untapped potential of the method in the area of comparative studies in Section Toward Comparative Prosopography. This possibility has been the least explored to date.

Historiographic References

Ancient history scholars claim to be the fathers of prosopography (Chastagnol, 1970; Nicolet, 1970). The term was conspicuously used in the modern sense by T. Mommsen in the announcement of the *Prosopographia Imperii Romani*, published in 1897 (Groag and Stein, 1897; Lalouette, 1999). This serial was imitated for other periods of ancient history throughout the twentieth century, both in collective works and in individual monographs. Prime examples are *The Prosopography of the Later Roman Empire* by A.H.M. Jones and J.R. Martindale (1971–92) and *Prosopographie chrétienne du Bas-Empire*, of which two volumes have appeared, edited by André Mandouze

(1982) and Charles Pietri and Luce Pietri (1982). Countless monographs contain prosopographies of elites of antiquity. One of the earliest is Jean Kirchner's *Prosopographia attica* (1901), followed by Paul Poralla's 1913 work on the Lacedaemonians and the classics by Friedrich Münzer (1920), Ronald Syme (1939), and Claude Nicolet (1966), etc. The application of prosopography to medieval times started to develop between the world wars. Studies focus on administrative personnel and the entourage of the sovereigns of the principal Western European states. Since the 1960s, the body of literature has became quite copious, as various surveys (Bulst, 1986; Griffiths, 1986) reveal. The same goes for modern history, where the initiator was Sir Lewis Namier (1929, 1930), who studied English political elites. His works inaugurated an ongoing series of studies on the House of Commons from the Middle Ages to 1832 (Cruickshanks, 1986). French and Italian investigation has concentrated on diverse selections of officials taken from public administration and the judiciary as well as on ecclesiastical, intellectual, financial, and commercial elites (see Charle et al., 1980; Nagle, 1986).

In the area of contemporary history, work began during the second half of the 1960s. Most of it was inspired by the perspective on elites introduced by sociologists, particularly American, influenced by the theses of Pareto and Mosca beginning in the 1950s. In France, one of the issues that stimulated research of this kind was the debate on the French Revolution opened by the theses of F. Furet and D. Richet (Richet, 1969; discussed by Vovelle, 1974), proposing an alternative to the Marxist interpretation then dominant in France, founded on prosopographic study of elites before and after the upheaval. This is the background of the investigation by L. Bergeron and G. Chaussinand-Nogaret on notables of the Napoleonic Era (1979). Similar works on prominent figures of the nineteenth century and the bourgeoisie since the 1960s have relied on the prosopographic method. This approach was inspired by Ernest Labrousse (Tudesq, 1964) (for an overall survey, see Charle, 1992). French sociology of education, influenced by the works of Pierre Bourdieu (Bourdieu and Passeron, 1970; Bourdieu, 1989), has also been the source of numerous intellectual and academic prosopographies (Karady, 1972, 1973). In Britain, Germany, and the USA and more recently in Italy, Spain, and Switzerland, collective biographies of political, economic, religious, intellectual, artistic, and militant elites have proliferated. They are an outgrowth of great national historiographic debates on subjects such as the foundation of political rifts in the British parliament, the evolution and function of English nobility, the distinguishing characteristics of German, Swiss, and Italian bourgeoisies (Kocka and Frevert, 1988; Tanner, 1990; Augustine, 1994; Malatesta, 1995), the formation and reproduction of American elites (Baltzell, 1958; Jaher, 1973), and so on. In contemporary history, the method has also been applied to larger groups than elites, which are the sole focus of prosopography in the spheres of ancient, medieval, and modern history because records on other segments of the population are not available. The history of women, the middle classes, lower classes, and even marginals have also availed themselves of collective biography in recent decades thanks to abundant printed sources, oral history, and new assessments of large biographic files, which already existed (Dreyfus et al., 1996; Omnès, 1997).

Merits and Limitations

The proliferation of studies in contemporary history and historical sociology using the prosopographic method is due to intellectual trends and techniques developed since the 1970s. The concomitant decline of global statistical methods and of the intellectual impact of Marxism, which featured social classes with huge populations, and the desire, typical of societies focused more and more on the individual, to come to grips with individual experience and the diversity of social patterns have inclined historians of recent times to emphasize social microhistory, which gives priority to prosopography or collective biography. Finally, the widespread advance of computerization has facilitated complex processing of bigger and bigger databases (Stone, 1971; Millet, 1985). Whereas in former times the lack of mass documentation required patient gathering of data on the individual level before any collective generalizations could be made, in recent periods resorting to prosopography has been construed as a change of perspective from traditional social history based on serial documentation.

Thanks to this technique, since the 1970s, historians have increasingly set their sights on gaining intimate knowledge of small problem-related constituencies. Thus, S.W. Serman's study of officers of the Second Republic and the Second Empire traced the long-lasting confrontation in nineteenth-century France between the aristocratic ideal and the rising democratic tide (Serman, 1978). Investigations of elites (high-ranking personalities of the July Monarchy (Tudesq, 1964; Girard et al., 1976), the business leaders of the Second Empire (Plessis, 1985; Barjot, 1958), ruling circles of the Third Republic (Estèbe, 1982; Charle, 1987), and deputies to the national assemblies in Paris and Frankfurt in 1848–49 (Best, 1990)) were attempts to understand the evolution of models of reproduction dominant in one regime toward those prevailing in other systems, the disparity between myths justifying the social order and sociological mechanisms, the possible social impact of political revolutions, reasons for the unique French road to industrialization, the peculiar German path to an incomplete parliamentary system, and the relations between English entrepreneurs and previously dominant elites (Berghoff, 1991). In short, these collective biographies allow us to revamp our answers to some major questions as well as to conduct an in-depth reassessment of cherished tenets of social history without confining ourselves to the preconceived structure of serial and quantifiable sources. In these new approaches, groups define themselves with reference to relational characteristics or their reciprocal images or even their capacity to impose an image of themselves on others and on the majority of their members. The notion of construction of the object thus becomes decisive in this microhistory based on collective biographies.

This is precisely the point that, since 1971, has triggered premature criticism of prosopography, summed up by Lawrence Stone with reference to modern history studies inspired by Sir Lewis Namier: the biases of official sources on which the biographies rely lead to a partial view of reality, the delimitation of populations is largely arbitrary, and criteria applied for the biographies are often restrictive. All this contributes to an elitist, cynical, and conformist perspective on the spheres of leadership and their relations with society at large (Stone, 1971). These

criticisms are not valid unless the historians confuse the method with its objective and forget that they can never consider more than a fraction of reality as determined by the sources, which they select and the limitations of their biographical questionnaires. They are less likely to fall into these traps for recent periods, where the findings of collective biographies can be confronted with other sources, where partial prosopographies can be compared and merged, and where the diversity of groups readily available for investigation goes well beyond the scope of the elites to which earlier studies were confined.

The prosopographical historian navigates permanently between two reefs: the undefined biography of individuals (with the risk of losing sight of the collective dimension) and the opposite: the expansion of prolific samples to the scale of the entire society (with the risk of reducing the questionnaire to its lowest common denominator). To escape from this dilemma, collective investigation by means of large research teams, adopting common principles, has been applied with varying success in different countries. Examples include the ongoing study of the British House of Commons (Namier and Brooke, 1964), the study of mayors in France (Agulhon et al., 1986), the investigation of members of parliament of the French Third Republic (Corbin and Mayeur, 1995), the collective biographies of legislators by the Zentrum für Historische Sozialforschung in Cologne (Schröder et al., 2000; Best and Cotta, 2000). These surveys require a long-term institutional financial commitment, which has become more and more difficult to find since profitability criteria were introduced in academic research. They also assume long-term involvement of loyal research personnel, which is less and less compatible with normal levels of staff turnover. This is why the principle of coordinated monographs, being more flexible and decentralized, appears to be more realistic in the contemporary university environment. It also overcomes the need to sacrifice the originality of each individual's work in the anonymity of a hierarchic endeavor. However, to avoid the risk of dissipation or noncomparability of partial results, it posits a minimum of harmonization and entails meetings to fine-tune the common questionnaires, coding systems, and evaluations. In exploiting collective biography, one is not only forced to gauge the conditions of sample validity. One must also recognize the limits of the categories applied a priori to the data. Since the responses to the questions that are included involve social stakes, the process of interpretation of prosopographic data is characterized, more so than in other methods, by historians who consciously or unconsciously put their mark on the results at all stages of the study: sampling, data gathering, coding, and processing.

Toward Comparative Prosopography

In France and increasingly in most other developed countries, nearly all groups that lend themselves to collective biography have been adopted by a biographer: aristocracies, dignitaries, urban bourgeoisies, administrative elites, financial, commercial and entrepreneurial elites, intellectual and university elites, artists, middle-class professionals (physicians, jurists, journalists, professors), students, officeholders at all levels, labor leaders and militant feminists, even marginals, etc. Is a decline of the method of collective biography imminent? Will it be stopped in the name of the unyielding law of diminishing returns and as a result of the resurgence of historiographic topics, which are less amenable to this approach (cultural history, history of memory, history of collective sensibilities)? Three considerations allow us to reject this premature diagnosis of a decline in the advantages of a method, which has been so fruitful in the area of recent social history.

First, apart from elites, other social groups, which are beginning to revise their social history thanks to insiders' collective biographies (middle classes and, increasingly, lower classes, not only through the eyes of their militant elites, but also in the form of life stories of privileged witnesses) (Ozouf et al., 1992; Pudal, 1991; Gribaudi, 1987) are far from being thoroughly investigated. Application of the method to these new fields requires new reflection on the composition of samples (which can no longer be exhaustive) and on the relations between individual paths and morphology: the internal diversity of the configurations is directly proportional to the size of the target population. This again raises the issue of sample representativeness, which had disappeared when investigators employing prosopography of elites resorted to exhaustive polls. All these groups are transitory, while elites, by definition at the top of the heap, are target groups; hence characteristics of the individuals of which such groups are composed cannot be viewed in isolation, but must be understood as vectors of multiple strategies.

Second, European historiographers are acquainted with the adjustments required to enable them to merge complementary studies to arrive at a homogeneous scrutiny of comparable groups from different societies. Thus, the comparative research coordinated by Jürgen Kocka on European bourgeoisies has demonstrated that not all of the difficulties in analyzing this group at an international level are due to the unequal advancement of studies; above all they arise from the collateral persistence of divergent investigations – some inspired by classical methods of social history, particularly in Germanic and Anglo-Saxon countries, others, particularly in France, primarily involving research based on collective biographies, which result in different approaches to different problems, and at times to inappropriate comparisons (Kocka and Frevert, 1988; Charle, 1990).

Collective biography, therefore, still has a nearly unexploited field at its disposal: that of comparative prosopography. The principal, eternal objection to comparative historical study is the necessity, given the abundance of material, to conduct secondhand research, using monographs prepared by other scholars, with the compounded risks of hasty generalization of conclusions taken out of their unique context and of the permanent lack of correspondence in the compared data, particularly in social history, on account of the recurring coding problem. Despite these obstacles, German researchers have already forged ahead, and certain French historians and specialists in England have followed suit since the 1990s (Best, 1990; Berghoff and Möller, 1994; Charle, 1994; Siegrist, 1995; Cassis, 1997; Ruggiu, 1997). Collective biography is an extraordinarily demanding methodology in terms of research time and rigorous organization of data. Comparative prosopography multiplies the time it takes to assemble the data by at least two – for two countries – or more, if one wants to compare a greater number of cases. The problem is compounded by the

extra time needed to design the research and evaluate the data. In order to test the initial hypothesis of feasibility and fruitfulness of a comparative project, it is convenient to start in a domain where one of the two countries involved already has a critical mass of usable data allowing researchers to define an analogous scope of investigation in the other selected country, which must possess a comparable, evident interest in the chosen elite. This explains the fact that comparative prosopographies deal primarily with the best-documented and most frequently studied elites (in big business, politics, higher education, and high society), as in the cases mentioned previously. Collective biographies at a national level have permitted us to better understand the internal rifts in the various groups, which have been studied as well as their social and generational dynamics. In certain favorable cases, it has been possible to relate these rifts to political, ideological, or religious stances or apply them to better understand certain failures or successes. Comparative prosopography should allow us to go even further in this interpretive analysis by relativizing correlations, which were taken for granted in a given national or social setting. They should account for common trends that transcend borders and the peculiarities of one moment in time, of one milieu or one nation. For instance, the generally accepted notion that the expansion of higher education translates into an opening of university teaching positions to a wider segment of society is disproved if one compares the collective biographies of the professors of the University of Berlin (Arts and Sciences) with their counterparts at the Sorbonne at the turn of the nineteenth and twentieth centuries. The major divergence accounts for the misunderstandings between the two academic communities even when they attempt to cooperate, and goes a long way toward explaining their widely varying relations with other elites during political crises (Charle, 1994).

The goal of comparative prosopography, as with any kind of prosopography, is to employ collective biographies to portray the real social behavior of the institutions or the milieus in which the individual subjects of the study interact (Genet and Lottes, 1996). The main danger in applying comparative prosopography is therefore, as with many erudite ventures, that one will discover nothing but well-established or generally acknowledged truths in each country concerned. One usually arrives at this impasse or this disappointing result when one deals with samples that are too vast to lend themselves to controlled comparisons, which would allow the study to qualify any dominant feature or put it into perspective. In light of this danger it is essential to make a particularly judicious choice of variables to be compared. We face a series of common methodological issues, which transcend historical periods. Their resolution would allow us to launch a wider debate on the possible foundations of a comparative sociocultural history of European society.

See also: Ancient Greece and Rome, History of; Biography and Society; Biography: Historical; Comparative History; Ethnoarchaeology; Historical Thought and Historiography: Current Trends; History: Representation Within and Functions of; History: Theories and Methods; Middle Ages, The; Quantification in History; Social History; Social Inequality in History (Stratification and Classes).

Bibliography

Agulhon, M., et al., 1986. Les Maires en France du Consulat à nos jours. Publications de la Sorbonne, Paris.

Augustine, D.L., 1994. Patricians and Parvenus. Wealth and High Society in Wilhelmine Germany. Berg, Oxford.

Autrand, F. (Ed.), 1986. Prosopographie et genèse de l'Etat moderne. L'Ecole normale supérieure de jeunes filles, Paris.

Baltzell, E.D., 1958. Philadelphia Gentlemen. The Making of a National Upper Class. Free Press, Glencoe, IL.

Barjot, D. (Ed. in chief), 1958. Les patrons du Second Empire. Cenomane, Picard, Le Mans, Paris.

Bergeron, L., Chaussinand-Nogaret, G., 1979. Les Masses de Granit. EHESS, Paris.

Berghoff, H., 1991. Englische Unternehmer 1870–1914. Eine Kollektivbiographie führender Wirtschaftsbürger in Birmingham, Bristol und Manchester. Vandenhoeck & Ruprecht, Gottingen, Germany.

Berghoff, H., Möller, R., 1994. Tired pioneers and dynamic newcomers? A comparative essay on English and German entrepreneurial history, 1870–1914. Economic History Review XLVII (2), 262–287.

Best, H., 1990. Die Mnner von Bildung und Besitz. Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich, 1848/49. Droste Verlag, Düsseldorf, Germany.

Best, H., Cotta, M. (Eds.), 2000. Parliamentary Representatives in Europe, 1848–2000. Legislative Recruitment and Careers in Eleven European Countries. Oxford University Press, Oxford, New York.

Bourdieu, P., Passeron, J.C., 1970. La Reproduction. Minuit, Paris.

Bourdieu, P., 1989. La Noblesse d'Etat, grandes écoles et esprit de corps. Minuit, Paris. Paris.

Bulst, N., Genet, J.-Ph (Eds.), 1986. Medieval Lives and the Historian. Studies in Medieval Prosopography, Kalamazoo, MI.

Bulst, N., 1986. La recherche prosopographique récente en Allemagne (1250–1650). Essai d'un bilan. In: Autrand, F. (Ed.), Prosopographie et genèse de l'Etat moderne. L'Ecole normale Supérieure de jeunes filles, Paris, pp. 35–52.

Cassis, Y., 1997. Big Business: The European Experience in the Twentieth Century. Oxford University Press, Oxford, New York.

Charle, C., et al., 1980. Prosopographie des élites franaises (XVIè–XXè siècles). Guide de recherche. CNRS, Institut d'histoire moderne et contemporaine, Paris.

Charle, C., 1987. Les Élites de la République (1880–1900). Fayard, Paris.

Charle, C., 1990. A la recherche des bourgeoisies européennes. Le Mouvement Social 153, 91–97.

Charle, C., 1992. The present state of research on the social history of elites and the bourgeoisie: a critical assessment on middle class and elites. Contemporary European History 1 (1), 99–112.

Charle, C., 1994. La République des universitaires (1870–1940). Le Seuil, Paris.

Charle, C., 1996. Du bon usage de la biographie sociale comparée. In: Dreyfus, M., Pennetier, C., Viet-Depaule, N. (Eds.), La part des militants, Biographie et mouvement ouvrier: Autour du Matron, Dictionnaire biographique du mouvement ouvrier franais. Editions de l'Atelier, Paris, pp. 51–72.

Chastagnol, A., 1970. La prosopographie méthode de recherche sur l'histoire du Bas Empire. Annales ESC 25 (5), 1229–1235.

Corbin, A., Mayeur, J.M. (Eds.), 1995. Les Immortels du Sénat. Publications de la Sorbonne, Paris.

Cruickshanks, E., 1986. Multi-biographical analysis as an approach to parliamentary history. In: Autrand, F. (Ed.), Prosopographie et genèse de l'Etat moderne. L'École normale supérieure de jeunes filles, Paris, pp. 335–344.

Cameron, A., British Academy, 2003. Fifty Years of Prosopography: The Later Roman Empire, Byzantium and Beyond. Published for the British Academy by Oxford University Press, Oxford.

Dreyfus, M., Pennetier, C., Viet-Depaule, N. (Eds.), 1996. La part des militants, Biographie et mouvement ouvrier: Autour du Matron, Dictionnaire biographique du mouvement ouvrier franais. Editions de l'Atelier, Paris.

Estèbe, J., 1982. Les Ministres de la République (1871–1914). Presses de la Foundation Nationale des Sciences Politiques, Paris.

Genet, J.-Ph, Lottes, G. (Eds.), 1996. L'État moderne et les élites XIIIè–XVIIIè siècles. Apports et limites de la méthode prosopographique. Publications de la Sorbonne, Paris.

Girard, L., Serman, W., Gossez, R., 1976. La Chambre des députés en 1837–39. Publications de la Sorbonne, Paris.

Gribaudi, M., 1987. Itinéraires ouvriers. Espaces et groupes sociaux Turin au début du Xxème siècle. Editions de l'EHESS, Paris.

Griffiths, R., 1986. Bureaucracy and the English state in the later middle ages. In: Autrand, F. (Ed.), Prosopographie et genèse de l'Etat moderne. L'Ecole normale supérieure de jeunes filles, Paris, pp. 52–65.
Groag, E., Stein, A. (Eds.), 1897. Prosopographia Imperii Romani Saec I. II III Pars I. Walter de Gruyter, Berlin and Leipzig, Germany.
Jaher, F.C. (Ed.), 1973. The Rich, the Well Born, and the Powerful. Elites and Upper Classes in History. University of Illinois Press, Urbana, IL.
Jones, A.H.M., Martindale, J.R. (Eds.), 1971, 1980, 1992. The Prosopography of the Later Roman Empire. Cambridge University Press, Cambridge.
Karady, V., 1972. Normaliens et autres enseignants la Belle Epoque. Note sur l'origine sociale et la réussite dans une profession intellectuelle. Revue franaise de sociologie 13 (1), 35–58.
Karady, V., 1973. L'expansion universitaire et l'évolution des inégalités devant la carrière d'enseignant au début de la IIIème République. Revue franaise de sociologie 14 (4), 443–470.
Kocka, J., Frevert, U. (Eds.), 1988. Bürgertum im 19. Jahrhundert. Deutsche Taschenbuch Verlag, Munich, Germany.
Kirchner, J., 1901. Prosopographia attica. Reimer, Berlin, Germany.
Keats-Rohan, K.S.B., 2007. Prosopography Approaches and Applications: A Handbook. Unit for Prosopographical Research, Linacre College, University of Oxford, Oxford.
Lalouette, J., 1999. De l'exemple à la série: histoire de la prosopographie. Siècles, cahiers d'histoire des entreprises et des communautés 10, 3–20.
Malatesta, M. (Ed.), 1995. Society and the Professions in Italy, 1860–1914. Cambridge University Press, Cambridge.
Mandouze, A. (Ed.), 1982. Prosopographie chrétienne du Bas-Empire. Prosopographie de l'Afrique chrétienne (303–533), vol. 1. Editions du CNRS, Paris.
Millet, H. (Ed.), 1985. Informatique et prosopographie. Editions du CNRS, Paris.
Münzer, F., 1920. Römische Adelsparteien und Adelsfamilien. J.B. Metzlersche Verlagsbuchhandlung, Stuttgart, Germany.
Nagle, J., 1986. Prosopographie et histoire de l'État: la France moderne XVIè–XVIIIè siècles. In: Autrand, F. (Ed.), Prosopographie et genèse de l'État moderne. L'Ecole normale supérieure de jeunes filles, Paris.
Namier, L., 1929. The Structure of Politics at the Accession of George III. Macmillan, London.
Namier, L., 1930. England in the Age of the American Revolution. Macmillan, London.
Namier, L., Brooke, J., 1964. The House of Commons, 1754–1790. Published for the History of Parliament Trust by her Majesty's Stationery Office, London.
Nicolet, C., 1966. L'Ordre équestre à l'époque républicaine (312–43 avant J.-C). De Boccard, Paris.
Nicolet, C., 1970. Prosopographie et histoire sociale: Rome et l'Italie l'époque républicaine. Annales ESC 25 (5), 1209–1228.
Omnès, C., 1997. Ouvrières parisiennes. Marchés du travail et trajectoires professionnelles au 20è siècle. ditions de l'EHESS, Paris.
Ozouf, J., et al., 1992. La République des instituteurs. Le Seuil, Paris. Gallimard.
Pietri, C., Pietri, L. (Eds.), 1982. Prosopographie chrétienne du Bas-Empire. Prosopographie de l'Italie chrétienne (303–533), vol. 2. Ecole franaise de Rome, Rome.
Plessis, A., 1985. Régents et gouverneurs de la Banque de France (1852–1870). Droz, Geneva, Switzerland.
Poralla, P., 1913. Prosopographie der Lakedaimonier bis auf die Zeit Alexanders des Grossen. Kommissions-Verlag von Max, Breslau, Germany.
La prosopographie, problèmes et méthodes, 1988. Mélanges de l'Ecole franaise de Rome, Moyen Age. Temps Modernes 100 (1), 5–284.
Pudal, B., 1991. Prendre parti, pour une sociologie historique du PCF. Presses de la Foundation Nationale des Sciences Politiques, Paris.
Richet, D., 1969. Autour des origines idéologiques lointaines de la Révolution franaise. Annales ESC 24 (1), 1–23.
Ruggiu, F.J., 1997. Les élites des villes moyennes en France et en Angleterre XVIIè–XVIIIè siècles. L'Harmattan, Paris.
Schröder, W.-H., Weege, W., Zech, M., 2000. Historische Parlamentarismus-, Eliten- und Biographieforschung, Forschung und Service am Zentrum für Historische Sozialforschung, Köln, Zentrum für Historische Sozialforschung. Historische Sozialforschung Supplemente/Beiheft 11.
Serman, S.W., 1978. Le corps des officiers sous la Deuxième République et le Second Empire. Atelier de reproduction des thèses de Lille III, Lille, France.
Siegrist, H., 1995. Advokat, Staat und Bürger. V. Klostermann, Frankfurt, Germany.
Stone, L., 1971. Prosopography. Daedalus 100 (1), 46–79.
Syme, R., 1939. The Roman Revolution. Clarendon Press, Oxford.
Tanner, A., 1990. Aristokratie und Bürgertum in der Schweiz im 19. Jahrhundert: Verbürgerlichung der 'Herren' und aristokratische Tendenzen im Bürgertum. In: Brndli, S., et al. (Eds.), Schweiz im Wandel. Studien zur neueren Gesellschaftsgeschichte. Helbing & Lichtenhahn, Basel & Frankfurt, pp. 209–228.
Tudesq, A.J., 1964. Les grands notables en France (1840–49). Etude historique d'une psychologie sociale. Presses Universitaires de France, Paris.
Vovelle, M., 1974. L'élite ou le mensonge des mots. Annales ESC 29 (1), 49–72.

Prospect Theory

Floris Heukelom, Nijmegen School of Management, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Daniel Kahneman and Amos Tversky's Prospect theory (1979) is to be situated in the history of probability and rational choice theory, as well as in a twentieth century American context of social engineering. As such, it is part of a vast literature encompassing psychology, economics, and mathematics, which seeks to criticize or elaborate the axiomatization of behavior as commenced by von John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* and Jimmie Savage's *The Foundations of Statistics* (1954). Nevertheless, and despite an important precursor in the work of Maurice Allais (1953a,b,c, Allais and Hagen, 1979), Prospect theory was the first attempt to produce a complete descriptive theory of human decision making under uncertainty, as it deviates from the norm of expected utility theory. Since 1979, Prospect theory has seen a gradually rising recognition of theories, experiments, and discussions that further elaborate on its arguments. That being said, Prospect theory has been particularly influential in Management & Organizational Behavior and Marketing & Advertising (Note: This article draws on Heukelom (2014)).

History and Context

Two historical contexts are important for Prospect theory, that of probability theory and that of the social engineering of society. Decision theory, or the theory of risk, or rational choice theory, goes back to the second half of the seventeenth century when scholars started to investigate how to calculate mathematically the optimal decision in uncertain situations. The mathematics that came out of these and similar questions was probability theory and rational choice theory (e.g., Daston, 1988; Hacking, 1975).

A major problem confronting decision theory was what became known as the St. Petersburg Paradox, invented by Nicholas Bernoulli in 1713. Nicholas Bernoulli demonstrated that gambles could be constructed for which decision theory computed a maximum willingness to pay that was clearly at odds with intuition. The most famous solution to the St. Petersburg Paradox was offered by his cousin Daniel Bernoulli in 1738 (Bernoulli, 1954; Jorland, 1987). Daniel Bernoulli argued that individuals base their choice between two or more (uncertain) outcomes not on the monetary or objective value of the outcomes, but on their individual, subjective valuation of these outcomes. Bernoulli labeled this individual valuation 'moral value' and supposed it decreases with increasing monetary outcome, as in the curve below (**Figure 1**).

Up until the early twentieth century the literature on mathematical theories of decision making under uncertainty consisted predominantly of attempts to solve this and similar paradoxes (Edwards, 1954; Jorland, 1987).

Between the 1920s and the 1950s a number of ideas were introduced that thoroughly reshaped the way decision theorists thought about decision making under uncertainty. (This paragraph briefly indicates a few points in a large literature. Useful overviews include Eriksson and Hajek (2007), Hajek (2007), and von Plato (1994).) Authors such as Bruno de Finetti and Frank Ramsey introduced the idea that probability theory could not only be applied to objective uncertainties out there in the world, such as the probability that a coin falls heads and the probability that the sun rises tomorrow, but also to subjective probabilities, that is uncertainties inside the individual of the sort "how uncertain am I that it will rain tomorrow?" (de Finetti, 1949; Ramsey, 1931). In a related development, authors such as John Maynard Keynes and Rudolf Carnap extended the theory of logic to include uncertain propositions, that is propositions with a degree of probability that is less than 1 (Carnap, 1950; Keynes, 1921). In this logical probability approach, uncertainty stems from the subject's personal belief in the occurrence of an event. The difference between objective and subjective probability is that objective probability is a probability obtained on the basis of available information and mathematical theory, a probability that by definition is the same for all human beings. Subjective probability, on the other hand, is a number attached to the personal belief of an individual. Subjective probabilities of the same event may thus differ across individuals.

The subjectivist probability theory commenced by de Finetti and Ramsey and the logical probability approach of Keynes, Carnap, and others made statistics a part of decision theory. Decision theory was no longer only about which decision we

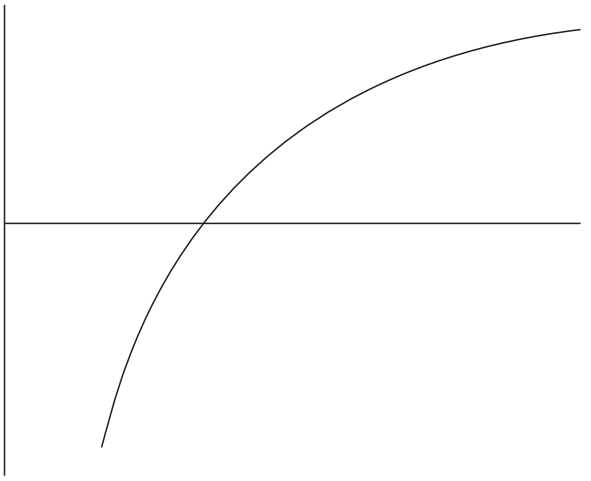

Figure 1 Relationship between monetary value (horizontal axis) and moral value (vertical axis; Bernoulli, D., 1954(1738). Exposition of a new theory on the measurement of risk. Econometrica 22 (1), 26; letters referred to in text Bernoulli have been removed).

as human beings should make given our preferences and the objective probability of different states of the world, it was now also about which conclusion should be inferred by statisticians from statistical data. Decision theory had incorporated statistics and was now an all-encompassing theory of human decision making under uncertainty.

However, the uncertainty employed in decision theory had always been an uncertainty external to human beings, even if individuals might assess and respond to this uncertainty in different ways. In 1944, John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* added a new chapter to the history of decision theory by constructing a theory that explains human decision making in relation to the uncertainty stemming from other individuals' strategic behavior. More strongly, one could argue that von Neumann and Morgenstern started the new field of decision theory which subsequently led to the (re)discovery of a few earlier publications and developments that could be reconstructed as the field's long history.

While von Neumann was working with Morgenstern on the *Theory of Games* in Princeton in 1941, he hired a bright new research assistant from the University of Michigan, Jimmie Savage. Some 10 years later Savage published *The Foundations of Statistics* (1954) in which he tried to ground all the above developments into one overlapping theory of human decision making. Savage (1954) quickly became the central point of reference for all those working in decision theory and related areas, and for all those critical of it.

Savage's most outspoken critic during the twentieth century was Maurice Allais from the École Nationale Supérieure des Mines in Paris. Over a lunch break during the "Colloque International sur le Risque," held in Paris from 12 to 17 May 1952, Allais employed some well-chosen hypothetical choices between lotteries, which induced Savage to choose in ways that violated his own theory (Heukelom, forthcoming). The most famous pair would become the following.

Which of the following two gambles do you prefer?

1. A 100% chance at 100 million francs
2. A 89% chance at 100 million francs
   A 1% chance at nothing
   A 10% chance at 500 million francs. Savage took the exchange rate at the time to be 350 francs to the dollar, which makes the 100 million francs equal roughly 285 000 1952 dollars. Similarly, 500 million francs equals about 1 400 000 1952 dollars. Most people choose (1), which does not maximize the expected pay-off, but which may be explained as resulting from risk aversion. Now, consider the next choice.

Which of the following two gambles do you prefer?

3. A 89% chance at nothing
   A 11% chance at 100 million francs
4. A 90% chance at nothing
   A 10% chance at 500 million francs
   (Allais, 1953a: p. 527)

Here, most people choose (4), which maximizes expected pay-off and is the riskier gamble of the two. This implies that individuals maximize and nonmaximize expected monetary pay-off, and are both risk seeking and risk averse. As a result, these outcomes were considered to be a serious problem for decision theory.

Rather than admitting his theory was wrong, Savage introduced a new distinction between a normative and an empirical interpretation of the theory. Under the normative interpretation scientists and all other normal, healthy individuals determined whether these were indeed the rules to be followed, should one desire to act rationally in choice situations under uncertainty. Within the empirical domain psychologists and other empirical scientists investigated when, how, and why actual, real-world decision making behavior by individuals is in accordance with decision theory.

In turn, American psychologists who worked in a psychometric approach to psychology that was strongly influenced by nineteenth century German experimental psychology, interpreted this distinction between a normative and an empirical interpretation to be the same as their use of a normative versus a descriptive psychology. The normative and descriptive have always been closely related in psychology. In Freudian psychoanalysis as well as in military recruits' IQ testing during World War One, the object was as much to describe and classify human behavior as it was to steer it in more desirable directions. In the words of James McKeen Cattel (1860–1944), the first professor of psychology in the United States at the University of Pennsylvania, the aim of psychology was "to describe, to understand, and to control human conduct" (Cattel, 1930: p. 31).

Kahneman and Tversky

Amos Tversky received a B.A. in the human sciences from Hebrew University in Jerusalem, Israel, in 1961. His teacher Israel Gutman recommended him to mathematical psychologist Clyde Coombs, after which Tversky entered graduate school at the University of Michigan. In 1965, Tversky obtained his PhD under the supervision of Ward Edwards and Clyde Coombs, combining Coombs' mathematical psychology with Edwards' behavioral decision research. The methodological tension Tversky was struggling with in his work of the late 1960s involved the question how Savage's theory, considered to be true on the basis of introspective, or intuitive arguments, could be combined with experimental results that pointed in many directions, but only occasionally in the direction of the intuitive theory. Tversky could not proceed as an experimentalist who seeks to test whether a theory is right or wrong. As Savage's axioms were understood as introspective or intuitive truths, they could ultimately only be proven wrong on the basis of introspective or intuitive reasoning. Normative decision theory was simply not understood and employed as a theory that could be falsified experimentally. At the same time Tversky wanted to do justice to the experimental results he had obtained. Sticking to the axioms of decision theory would have implied that whenever a systematic behavioral deviation was observed in the experiments, there was something wrong with the experiment. This would mean that the majority of Tversky's experiments were invalid. Tversky was thus effectively caught between the intuitive truth of the axioms of measurement theory and decision theory, and the behavioral deviations that surfaced in his experiments.

It was Daniel Kahneman who solved Tversky's problem. Kahneman obtained a B.A. from Hebrew University in 1956 while working as a psychologist in the Israeli army. In 1958 he moved to the United States and obtained a PhD from the University of California at Berkeley in 1961 under the supervision of Susan Ervin. After completing his PhD Kahneman returned to the Psychology Department at Hebrew University where he would remain until 1978. In the meantime, however, he was among others a lecturer in the graduate program of the University of Michigan in 1968–69. While teaching at Michigan Kahneman invited his younger colleague Tversky to lecture a class on recent developments in judgment and decision making (Kahneman, 2002). In the 1960s, Kahneman's research was about semantic differentials, optometry, vision research, and related themes. Kahneman's research in this period was unrelated to mathematical psychology, unrelated to decision theory, unrelated to behavioral decision research, and despite some retrospective hints of Kahneman to the contrary, entirely unrelated to economics.

To Kahneman it was and is a given fact of life that human beings often make cognitive errors. However, science could help in two ways. First, scientists could set out what the correct way of behaving is for each situation. For a truck driver missing the traffic light, it is obvious what the correct behavior is, but for a flight instructor seeking to optimize his responses to recruits' flying performance it may not be intuitively clear what the correct way of reasoning is. Scientists can therefore help to establish the correct way of reasoning. Second, scientists, and in particular psychologists, could help by investigating when, how, and in what way human beings make cognitive errors and thus provide a basis for designing tools or education to help human beings correct these cognitive errors.

In 1969 Kahneman and Tversky started a collaboration that would result in 21 papers and 2 coedited books. They continued to cooperate on different projects until Tversky's death in 1996, but the most productive and creative period was from 1969 to 1979, including the widely cited 1974 *Science* and 1979 Prospect theory article in *Econometrica*. The cooperation was initiated by Kahneman, who was looking for new ways to experimentally test his intuition that an individual's cognitive apparatus systematically fails, and who tried to find a theory that might account for these cognitive errors. But the ensuing program stemmed as much from Tversky's growing doubts concerning Savage and Edwards' assumption that, generally speaking, individuals decide according to the normative rules of logic, Bayesian statistics, and expected utility theory (Savage, 1954; Edwards, 1954, 1961).

Prospect Theory

Kahneman and Tversky's *Prospect Theory: an Analysis of Decision under Risk* (1979) was initially entitled 'Value theory' when first submitted to *Econometrica* in 1975. Eventually, the name Prospect theory was chosen because that term did not refer to any other theory or phenomenon in economics and psychology (Kahneman, 2002). Prospect theory has often been presented as being different from Kahneman and Tversky's earlier widely cited Heuristics and Biases program (Kahneman, 2002; Tversky and Kahneman, 1974). But while it is certainly true that Prospect theory brought the different heuristics into a new overarching framework and focused on the norm of expected utility theory instead of statistics, the foundation still was the idea that human beings rely on a set of heuristics for their decision making and that the use of these heuristics sometimes leads to systematic deviations from the normatively correct decision. The 4 years between 1975 and 1979 were used to tweak a for the most part finished argument to fit an audience of not only psychologists, but also, and in particular, economists, decision theorists, mathematical psychologists, and other mathematically oriented behavioral scientists (Kahneman, 2002).

Despite an important precursor in the work of Maurice Allais (1953a,b,c; Allais and Hagen, 1979), Prospect theory was the first attempt to produce a complete descriptive theory of human decision making under uncertainty, as it deviates from the norm of expected utility theory. As in Heuristics and Biases, Kahneman and Tversky based their argument on a series of hypothetical questions they had presented to experimental subjects, in this case psychology students at Hebrew University. The problems subjects were presented with were decision problems, involving different material outcomes and different probabilities. Most of the questions were reformulations or variants of Allais' decision problems. One example of Kahneman and Tversky's use of an Allais-type approach is in the question where subjects were asked to state which of the following lottery options they preferred.

$$\text{A}: \ (4000, \ .80) \quad \text{or} \quad \text{B}: \ (3000)$$

That is, they were asked whether they preferred 4000 shekel with a probability of .8, or 3000 shekel for certain (at the time of the experiment, 4000 shekel was about one-third of the modal monthly Israeli income). Most of the subjects in this case chose B. This implied that they did not maximize the expected monetary outcome. However, opting for choice B could be explained by assuming that the decision maker was risk averse. Subsequently, subjects were asked which of the following two lottery options they preferred.

$$\text{C}: \ (4000, \ .20) \quad \text{or} \quad \text{D}: \ (3000, \ .25)$$

In this case, most of the subjects chose C and hence maximized the expected monetary outcome. This was considered problematic in combination with the first choice as it implied that subjects were sometimes risk averse, but on other occasions maximized the expected monetary outcome and hence were not risk averse. Note that the second choice is equal to the first with probabilities divided by four. With these and similar examples, Kahneman and Tversky illustrated that despite its normative status, expected utility theory as a descriptive theory was invalidated. In specific circumstances people systematically deviated from the norms of expected utility theory. A new descriptive, 'alternative account of individual decision making under risk,' was therefore required (Kahneman and Tversky, 1979: p. 274).

According to Prospect theory, a human decision maker first employs a number of heuristics to make a decision problem manageable. This process was called the editing phase. Complicated decisions are broken down into different simpler decisions, different decisions are lumped together into one big decision, a benchmark is set with which the decision is

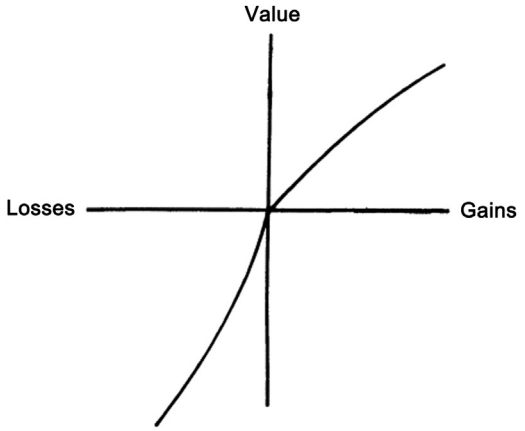

Figure 2 A hypothetical value function. Reproduced from Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 279.

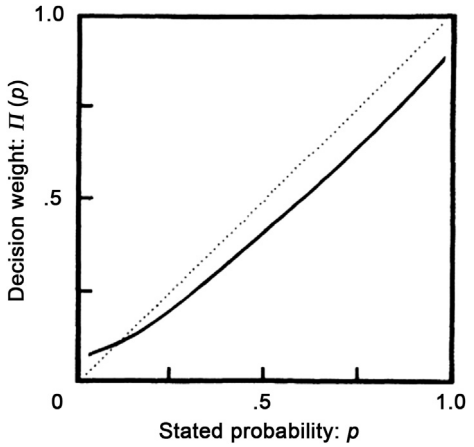

Figure 3 A hypothetical weighting function. Reproduced from Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 283.

compared, and so on. The purpose of this editing phase is to make the decision manageable. After this, the decision is evaluated in what was referred to as the evaluation phase. The evaluation phase has the same structure as the maximization of expected utility, but instead of the objective values of the material pay-off and probability, it uses the individual's subjective perception of the material pay-off and probability. The subjective perception of the material pay-off was referred to as value (denoted v) and the subjective perception of probability was referred to as decision weight (denoted π). In expected utility theory, a subject who is faced with a choice between outcome x that occurs with probability p and outcome y that occurs with probability q derives utility according to the following function.

$$U(x, p; y, q) = p \cdot u(x) + q \cdot u(y) \qquad [1]$$

in which utility u is a subjective valuation of the outcome according to the axioms of von Neumann and Morgenstern (1944) and Savage (1954). In other words, it defines how an individual values an outcome given its preferences if it behaves according to the normative rules of rational decision making. Furthermore, in the expected utility theory of eqn. [1] the individual perceives the probabilities of the outcomes as what they objectively are. In prospect theory, by contrast, a subject that following the editing phase faces the exact same choice values this choice according to this function.

$$V(x, p; y, q) = \pi(p)v(x) + \pi(q)v(y) \qquad [2]$$

in which v is similar to u, but based on empirical observations in experiments rather than axiomatically defined utilities; constructed with respect to an individual reference point, rather than to an objectively defined benchmark; and with a risk-seeking character in the loss domain. The experimentally derived individual perceptions of the von Neumann–Morgenstern–Savage defined utility, then, yielded the following curve (**Figure 2**).

In a similar way, and based on the same questionnaire experiments, individual's perception of the actual (i.e., objective) probabilities was inferred to be as follows (**Figure 3**).

In a clever way the opening sentences of Prospect theory alluded to frameworks of both psychologists, mathematical decision theorists, and economists.

> Expected utility theory has dominated the analysis of decision making under risk. It has been generally accepted as a normative model of rational choice, and widely applied as a descriptive model of economic behavior. Thus it is assumed that all reasonable people would wish to obey the axioms of the theory and that most people actually do, most of the time.
>
> Kahneman and Tversky, 1979: p. 263

To psychologists and decision theorists these sentences restated a well-known normative – descriptive framework that went back to Savage (1954) and Edwards (1954) in particular, and signaled a contribution to an established field of research. On the other hand, economists in the line of Friedman (1953) (in)famous positive–normative dichotomy might have raised their eyebrows at the injunction of 'normative,' but would certainly have agreed that reasonable people wish to obey the axioms of expected utility theory and that they actually do so, or at least most of the time. Note, furthermore, that Kahneman and Tversky carefully avoided the term 'rational,' and used 'reasonable' instead. Evoking the term 'rational' might have suggested that this was an article in the line of critique of economics. The use of 'rational' would certainly have induced some economists to think that these two psychologists had the same research program as Simon, who had won the Nobel memorial prize in economics the year before. From the start, prospect theory was carefully constructed so as to be able to broaden the scope to economists especially.

That said, it seems unnecessarily confusing to label the actual, descriptive valuation of the outcome of a choice by the individual 'value,' instead of utility. This was first of all part of a broader confusing use of concepts such as value and utility by psychologists and economists. At the same time, it was a clever deviation from other meanings and definitions. By making 'value' instead of 'utility' the term that referred to the

subjective attractiveness of a choice option, Kahneman and Tversky were able to design a framework that could be accepted by both psychologists and economists. To psychologists, the value framework matched the Kurt Lewinian valence framework; to economists it preserved the traditional framework of expected valuation by the individual, while at the same time allowing for the possibility that individual economic behavior deviates from the von Neumann–Morgenstern expected utility through the replacement of 'utility' with 'value.' In addition, the term 'value' was historically a central concept in economics that could be used very well to denote the pleasure an economic agent derives from choosing a specific option.

Assessment

Kahneman and Tversky's research is part of a vast literature encompassing psychology, economics, and mathematics, which seeks to criticize or elaborate the axiomatization of behavior as commenced by von Neumann and Morgenstern (2004) and Savage (1954) (Gigerenzer and Murray, 1987; Goldstein and Hogarth, 1997; Hogarth and Reder, 1987; Guala, 2005; Fishburn, 1981). Some of these contributions have been more successful than others, but arguably, few, if any, have been as successful as have been Kahneman and Tversky – as exemplified by Kahneman's 2002 Nobel memorial prize in economics and the number of times Kahneman and Tversky's work has been cited (e.g., Laibson and Zeckhauser, 1998). At the same time, Kahneman and Tversky's research has been met with substantial criticism, part of which can probably be explained by Kahneman and Tversky's success and influence itself.

The most often repeated criticism has been that Kahneman and Tversky believed people to be irrational. The argument was that if human beings can send people to the moon and return them safely, they cannot be that irrational. One of the most remarkable exponents of this view was Edwards. Edwards never elaborated upon his reservations with respect to the work of Tversky and Kahneman, perhaps from fear of losing the image of coherence of his program, yet it is no secret that he disagreed with their work (e.g., Phillips and Von Winterfeldt, 2006; Krantz – Interview, 2008; Dawes – Interview, 2008). The 1980s and 1990s gave rise to a whole surge of criticism regarding Kahneman and Tversky's approach to human decision behavior. The most prominent philosophical critique was provided by Cohen (1981). The most extensive criticisms from within experimental psychology came from Gerd Gigerenzer who questioned the set-up and interpretation of Kahneman and Tversky's experiments, and who wondered how people manage to travel to the moon if they are so bad at assessing uncertainties (e.g., Gigerenzer, 1991, 1993, 1996; Gigerenzer and Murray, 1987; Sedlmeier and Gigerenzer, 1997, 2000; Hertwig and Gigerenzer, 1999). Gigerenzer has been the only critic to whom Kahneman and Tversky have explicitly responded (Kahneman and Tversky, 1996). Other critics include Lopes (1991) and Cosmides and Tooby (1996). The most salient question pertaining to the work of the two psychologists is therefore why they were successful where others were less successful, despite the broad range of critique they received. In my opinion, three aspects stand out in particular. (See Hands, 2011 for further explanations.)

Friendly Criticism and the Normative–Descriptive Shuffle

In prospect theory Kahneman and Tversky employed the standard normative – descriptive distinction of experimental psychology, behavioral decision research, and mathematical psychology, and assumed that economists would employ the very same distinction. This partly reflected a standard understanding of economics in behavioral decision research and psychology. One possible conclusion therefore is that Kahneman and Tversky were not familiar with economists' use of positive and normative, and understood these terms to mean the same thing as the descriptive and normative of behavioral decision research. But Kahneman and Tversky's use of normative and descriptive can also be seen as a very clever way of trying to convince economists of the relevance of Prospect theory for economics. In Prospect theory, Kahneman and Tversky did not tell the economists that their theory was complete nonsense or useless. Instead, they claimed to understand economics as using one theory to cover both the normative and the descriptive realm. For the normative part they fully agreed with economists, which fitted in neatly with practice in behavioral decision research. But, Kahneman and Tversky argued, economists had been mistaken in using that same theory in the descriptive domain.

Put differently, according to Kahneman and Tversky there was nothing wrong with economists' theory of expected utility maximization. It was only that this was the normative theory, and not an accurate description of actually observed human behavior. Economists did not need to abandon the theory of expected utility maximization, but instead they should seek a proper descriptive counterpart to this normative theory. A first explanation for Kahneman and Tversky's success among economists is that they provided a relatively friendly criticism which allowed expected utility theory a life under the rubric of the normative theory. Moreover, they offered an alternative for the descriptive realm in the form of prospect theory that closely resembled expected utility theory.

The Role of the New Type of Experiments

A second reason for the success of Prospect theory's perspective of human decision making were the type of experiments employed. In their collaborative research during the 1970s, Kahneman and Tversky conducted different kinds of experiments than the two had done individually in the 1960s. From the early 1970s onward, the experiments consisted of questionnaires with hypothetical questions which mainly students were asked to fill out. These questionnaires could be distributed anywhere, to participants at conferences, to students during a course at the university, and in a shopping mall on a Saturday afternoon. These questionnaire experiments provided only copies of the questionnaires and pens to fill them out. No separate laboratory space was needed and no payment of the experimental subjects was required, filling out the questionnaires took a few minutes at the most. In the 1979 Prospect theory article, Kahneman and Tversky defended their new method of experimentation and contrasted it with two other

methods of investigation. First they distinguished the possibility of 'field studies,' which used 'naturalistic or statistical observation' that could yield important insights when a new area of research was opened, but in the end could only provide "crude tests of qualitative predictions, because probabilities and utilities cannot be adequately measured" (Kahneman and Tversky, 1979: p. 265). Second, they recognized the method of laboratory experiments, which had the disadvantages that stakes could only be relatively small and that for financial reasons it was difficult to conduct an experiment in which the subjects were asked to choose between, say, $300 for certain or a .8 chance at obtaining $400. The third method, which they argued proved to be the best for solving the problem they were investigating, was 'the method of hypothetical choices.' It solved the problem of the absence of control of field studies, while at the same time solving the external validity problem of laboratory experiments. It did however rely on the assumption that people "know how they would behave in actual situations of choice" and that they "have no special reason to disguise their true preferences" (Kahneman and Tversky, 1979: p. 265).

There is not a direct link between Prospect theory, and the new, more relaxed standards of the experimental method. That is to say, Prospect theory could, in theory, have been developed without the new means of conducting experiments. But the new standards of experimental control facilitated Kahneman and Tversky's research in fundamental ways. It is safe to say that, without the method of hypothetical questions they could not have developed heuristics and biases and prospect theory. Thus, although there is not a direct link between Kahneman and Tversky's Prospect theory and their experimental method, it is difficult to see how they could have been tested and developed without this new method. The success of Kahneman and Tversky's new program is hence in part explained by the new experimental methodology they employed.

Intuitively Appealing Examples

A third reason for Kahneman and Tversky's success is their effective use of intuitively appealing examples. Typically, the argument in Kahneman and Tversky's research was made not so much by giving theoretical explanations for why such and such was a good theory or account of observed behavior, but by supplying examples of the experimental questions subjects had been asked which were meant to give the reader an intuitive understanding of the point they were trying to make. The appeal of the examples employed gradually developed over time. The best known query among the many experimental questions of Kahneman and Tversky became the so-called 'Linda problem,' no less because it was often used by Kahneman and Tversky's adversaries. The Linda problem runs as follows:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which of the following two alternatives is more probable:

1. Linda is a bank teller.
2. Linda is a bank teller and active in the feminist movement (Tversky and Kahneman, 1983: pp. 297, 299).

On average, there was a strong bias toward judging (2) to be more probable than (1), this despite the fact that (2) is logically contained in (1). Because this bias is an illustration of the failure to see that the probability of the conjunction of two or more events can never exceed the probability of one of the events, this bias was labeled the conjunction fallacy.

The appeal and intuition of these examples provide a third explanation for Kahneman and Tversky's success, because the readers of their articles could relate the examples to their own everyday decision making. It was also via these intuitive examples that Kahneman and Tversky received some strong opposition in the 1980s and 1990s. Kahneman and Tversky's adversaries typically proceeded by deconstructing their examples and illustrations, or by giving counterexamples. It was for instance shown that a different conclusion could be inferred from the observed behavior, or that responses to another set of hypothetical questions falsified Kahneman and Tversky's conclusions (e.g., Gigerenzer, 1991, 1993, 1996; Hertwig and Gigerenzer, 1999; Lopes, 1991).

Influence in Behavioral and Social Sciences

The term 'Prospect theory' is only associated with the theory by Kahneman and Tversky. As a result, a first, rough indication of its rise and spreading after its introduction in 1979 is a standard Google ngram search. Figure 4 presents the

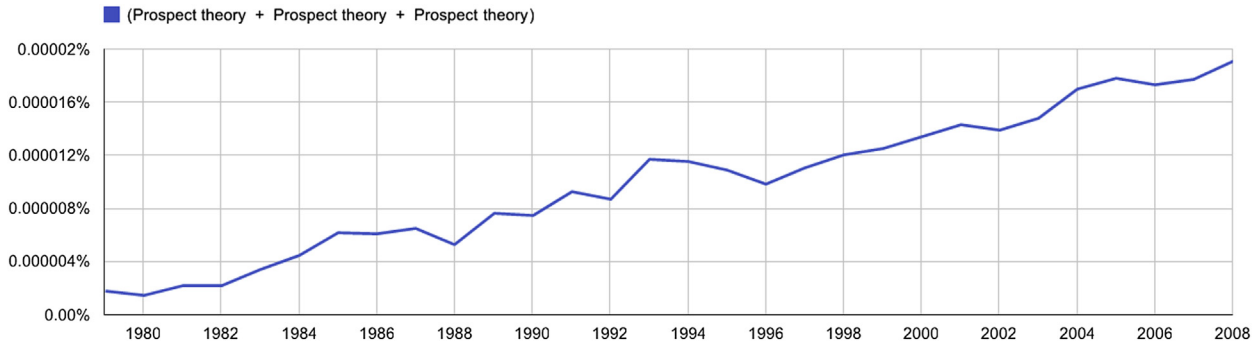

Figure 4 Google ngram moving average of 2 years.

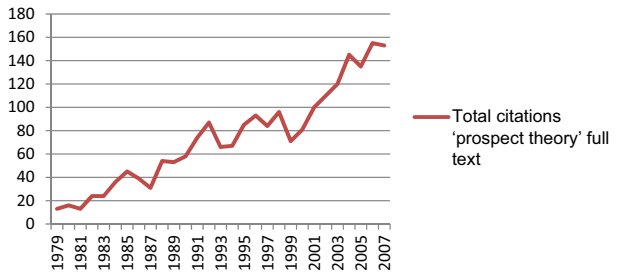

Figure 5 Total citations 'prospect theory' full text.

results for the combined appearance in the Google Books project of either prospect theory, Prospect theory, or Prospect Theory (ngram is case sensitive), with a moving average of 2 years.

Because Google ngram reports appearance of terms as a percentage of total available Google books, the outcome is not sensitive to a general rise in the number of books over the years. The results indicate a gradually rising recognition of prospect theory and of theories, experiments, and discussions that further elaborate on its arguments.

An alternative quantitative suggestion at the rise and spreading of prospect theory may be obtained through Jstor – which includes the vast majority of scientific articles, books, pamphlets, reviews, and other science-related publications. A minor drawback is that while Jstor offers the contents of some journal until early 2013 (at the time of writing), most are only available through 2010, while a substantial number does not run further than 2009 or 2008. To be on the safe side regarding the general trend, it seems better not to look beyond 2007. A full-text search of all published material available in Jstor up to 2007, then, yields the following result (Figure 5).

Unlike Google ngram, a Jstor search does not discount for a general rise in published and/or available sources over the period investigated. Nevertheless, it seems safe to conclude the results indicate a trend similar to that of Figure 4.

Table 1 Citations 'prospect theory' Jstor, full text, 1979–2007. Due to overlap, the sum of the subcategories listed is larger than the total number of citations

Category	*Number titles*	*Number citations*	*Citations/ titles*
Marketing and advertising	14	199	14.2
Management and organizational behavior	31	418	13.5
Business	235	1346	5.7
Psychology	18	95	5.3
Finance	31	100	3.2
Economics	173	543	3.1
Political science	152	406	2.7
Law	97	202	2.1
Sociology	128	233	1.8
Public policy and administration	38	53	1.4
Philosophy	99	56	0.6
Sum Subcategories	1016	3651	3.6
Total	1695	2128	1.3

In addition, Jstor permits us to derive some indication of the relative impact of Prospect theory in various scientific fields. A difficulty here is that the categories defined by Jstor include some overlap. The journal *Econometrica*, for instance, is both listed in the category of Economics and in the category of Mathematics. As a result, while the total references to Prospect theory in Jstor from 1979 to 2007 is 2128, the sum of the citations in the Jstor-subcategories with the most references to Prospect theory equals 3651 – as reported in the bottom two rows of Table 1 below. A second nuance is that the number of titles included in each category varies widely, from 235 in Business to 14 in Marketing and Advertising. This may be solved by calculating the citations per journal in each category. However, the number of issues a journal publishes per year, the number of articles per journal, and the number of citations per article may differ between categories. These variations cannot be controlled for. Finally, Jstor does not indicate how well it represents each discipline.

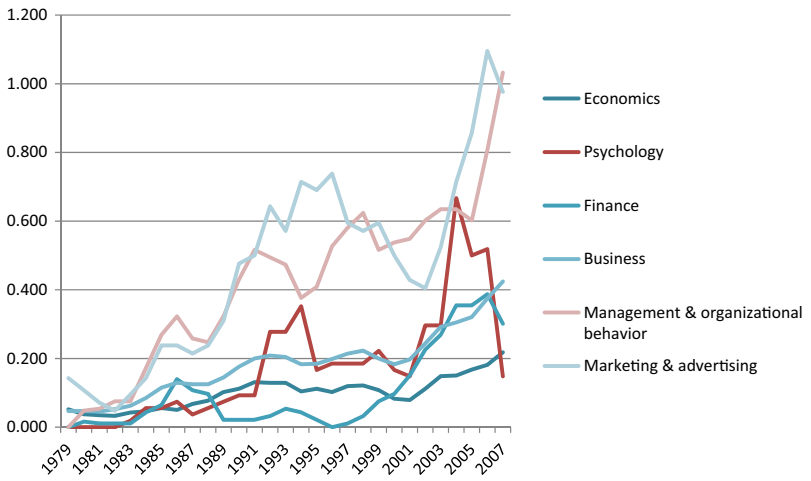

Figure 6 Citations/title per Jstor category, 1979–2007, moving average 3 years, except 1980 (2 years) and 1979 (1 year).

That being said, a subdivision of Jstor Prospect theory full-text appearances offers some hints as to the general pattern. First of all, Figure 3 shows Prospect theory has had a wide impact in the social sciences and related fields. Even relatively remote fields such as Philosophy and Sociology have produced substantial numbers of references to Prospect theory. But, secondly, the table also indicates the impact of Prospect theory to vary widely between fields. Judged by citations per title, a field such as Philosophy has produced the occasional discussion of and reference to Prospect theory, while Prospect theory has become part of the standard vocabulary in Management and Organizational Behavior and Marketing and Advertising. The contrast with Psychology and Economics is striking. Kahneman and Tversky were psychologists, who published in psychological and economic journals. Moreover, in 2002 Kahneman received the Nobel prize for Economics. Yet, while both Psychology and Economics produce substantial numbers of citations of Prospect theory, it had its largest impact outside of those two fields. The case apart of Business confirms these signals. The number of citations in Business is the largest by far, but is the product of 235 titles. As a result, it seems not unlikely that reasonable subcategories of Business could be defined with varying levels of Prospect theory citations. Given that parts of Business are closely related to Management and Organizational Behavior and Marketing and Advertising, the numbers for Business to further emphasize the relative impact of Prospect theory in these fields (Figure 6).

Finally, tracing the citations per title over time, as in Table 1, does not yield unambiguous results. The graphs seem to suggest that the popularity of Prospect theory in psychology is past its peak, while its popularity in other fields continues to rise. A final assessment, however, will have to await the third edition of this encyclopedia.

See also: Decision and Choice: Heuristics; Heuristics in Social Cognition; Heuristics, History of; Rational Choice Theory in Sociology; Rational Choice and Organization Theory; Tversky, Amos (1937–96); Von Neumann, John (1903–57).

Bibliography

Non-Published Sources

David Krantz, Interview with the Author, 20 June 2008. Columbia University, New York City.

Robyn Dawes, Interview with the Author, 23 June 2008. Carnegie Mellon University, Pittsburgh.

Published Sources

Allais, M., 1953a. Fondements d'une Théorie Positive des Choix Comportant un Risque et Critique des Postulats et Axioms de L'Ecole Americaine. Econometrie XL, 257–332.

Allais, M., 1953b. La Psychologie de L'homme Rationnel devant le risque: La Théorie et L'experience. Journal de la société statistique de Paris 94, 47–73.

Allais, M., 1953c. Le comportement de l'homme Rationnel devant de le risque: critique des postulats et axioms de l'école americainne. Econometrica 21, 503–546.

Allais, M., Hagen, O. (Eds.), 1979. Expected Utility Hypotheses and the Allais Paradox. D. Reidel Publishing Company, London.

Bernoulli, D., 1954[1738]. Exposition of a new theory on the measurement of risk. Econometrica 22 (1), 23–36.

Carnap, R., 1950. Logical Foundations of Probability. University of Chicago Press, Chicago.

Cattel, M., 1930. Psychology in America. In: Psychological Review Ninth International Congress of Psychology, Proceedings and Papers, pp. 12–32.

Cohen, L.J., 1981. Can human irrationality be experimentally demonstrated? The Behavioral and Brain Sciences 4, 317–370.

Cosmides, L., Tooby, J., 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions form the literature on judgment under uncertainty. Cognition 58, 1–73.

Daston, L., 1988. Classical Probability in the Enlightenment. Princeton University Press, Princeton.

Edwards, W., 1954. The theory of decision making. Psychological Bulletin 51, 380–417.

Edwards, W., 1961. Behavioral decision theory. Annual Review of Psychology (12) 473–498.

Eriksson, L., Hajek, A., 2007. What are degrees of belief? Studia Logica 86 (2), 185–215.

de Finetti, B., 1949. Le vrai et le probable. Dialectica 3, 78–93.

Fishburn, P.C., 1981. Subjective expected utility: a review of normative theories. Theory and Decision 13, 139–199.

Gigerenzer, G., Murray, D.J., 1987. Cognition as Intuitive Statistics. Lawrence Erlbaum, Hillsdale.

Gigerenzer, G., 1991. How to make cognitive illusions disappear: beyond "heuristics and biases". European Review of Social Psychology 2, 83–115.

Gigerenzer, G., 1993. From metaphysics to psychophysics and statistics. The Behavioral and Brain Sciences 16 (1), 139–140.

Gigerenzer, G., 1996. On narrow norms and vague heuristics: a reply to Kahneman and Tversky (1996). Psychological Review 103 (3), 592–596.

Goldstein, W.M., Hogarth, R.M. (Eds.), 1997. Research on Judgement and Decision Making: Currents, Connections, and Controversies. Cambridge University Press, Cambridge.

Guala, F., 2005. The Methodology of Experimental Economics. Cambridge University Press, Cambridge.

Hacking, I., 1975. The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference. Cambridge University Press, London, New York.

Hajek, A., 2007. Interpretations of probability. Winter 2007 Edition. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy.

Hands, D.W., 2011. Normative Rational Choice Theory: Past, Present, and Future SSRN.

Hertwig, R., Gigerenzer, G., 1999. The 'conjunction fallacy' revisited: how intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making 12, 275–305.

Heukelom, F., 2014. Behavioral Economics: A History. Cambridge University Press, Cambridge.

Heukelom, F. A History of the Allais Paradox. British Journal for the History of Science, forthcoming.

Hogarth, R.M., Reder, M.W. (Eds.), 1987. Rational Choice: The Contrast between Economics and Psychology. University of Chicago Press, Chicago.

Jorland, G., 1987. The Saint Petersburg paradox 1713–1937. In: Kruger, L., Daston, L.J., Heidelberger, M. (Eds.), The Probabilistic Revolution, vol. 1. MIT Press, Cambridge, pp. 157–190.

Kahneman, D., 2002. Autobiography. From http://nobelprize.org/economics/laureates/2002/kahneman-autobio.html.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–292.

Kahneman, D., Tversky, A., 1996. On the reality of cognitive illusions. Psychological Review 103 (3), 582–591.

Keynes, J.M., 1921. Treatise on Probability. Macmillan & Co, London.

Laibson, D., Zeckhauser, R., 1998. Amos Tversky and the ascent of behavioral economics. Journal of Risk and Uncertainty 16, 7–47.

Lopes, L.L., 1991. The rhetoric of irrationality. Theory and Psychology 1, 65–82.

von Neumann, J., Morgenstern, O., 2004[1944]. Theory of Games and Economic Behavior. Princeton University Press, Princeton.

Phillips, L.D., Von Winterfeldt, D., 2006. Reflections on the Contributions of Ward Edwards to Decision Analysis and Behavioral Research. LSEOR, London, LSE.

Porter, T.M., 1986. The Rise of Statistical Thinking 1820–1900. Princeton University Press, Princeton.

von Plato, J., 1994. Creating Modern Probability. Cambridge University Press, Cambridge.
Ramsey, F., 1931. The Foundations of Mathematics and Other Logical Essays. Routledge, London.
Savage, L.J., 1954. The Foundations of Statistics. John Wiley & Sons, New York.
Sedlmeier, P., Gigerenzer, G., 1997. Intuitions about sample size: the empirical law of large numbers. Journal of Behavioral Decision Making 10, 33–51.
Sedlmeier, P., Gigerenzer, G., 2000. Was Bernoulli wrong? On intuitions about sample size. Journal of Behavioral Decision Making 13, 133–139.
Simon, H.A., 1955. A behavioral model of rational choice. Quarterly Journal of Economics 69 (1), 99–118.
Tversky, A., Kahneman, D., 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review 90, 293–315.
Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 1124–1131.
Wald, A., 1950. Statistical Decision Functions. Wiley, New York.

Prospective Memory

Michael K Scullin, Emory University School of Medicine, Atlanta, GA, USA
Hillary G Mullet, Duke University, Durham, NC, USA
Gilles O Einstein, Furman University, Greenville, SC, USA
Mark A McDaniel, Washington University in St. Louis, St. Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Prospective memory refers to the ability to remember to perform delayed intentions such as remembering to take medication. Prospective memory is dissociable from other forms of memory by its future focus, the need for planning, and a reliance on self-initiated retrieval. Current theoretical approaches examine whether prospective memory is supported by effortful monitoring (anterior prefrontal cortex), relatively automatic spontaneous retrieval (hippocampus), or both. Conditions that have an impact on these processes (and supporting neural structures), such as substance abuse, neurodegenerative diseases, and aging, are associated with prospective memory impairments. Forming specific plans, avoiding interruptions, and relying on salient external cues can improve prospective remembering.

The ability to remember is one of the brain's most important functions, and thus memory research has been at the forefront of psychological science for over 125 years. Historically, most memory research has focused on retrospective memory, or memory for past events, such as remembering a baseball game you went to with your father as a child. Whereas retrospective memory captures our ability to recollect what has occurred in the past, researchers have begun to recognize the many other adaptive uses to which memory is put, such as enabling individuals to recall at the correct moment an intention that they need to perform (e.g., silencing a cell phone when entering a classroom). The ability to remember to perform delayed intentions is referred to as prospective memory.

Prospective memory is ubiquitous in our daily lives. We rely on prospective memory to meet social needs such as remembering to wish a friend a happy birthday; to perform daily chores such as remembering to start the dishwasher before going to bed; to excel in the workplace by remembering to follow through on assigned tasks that cannot immediately be performed; and to maintain health by remembering to take medications with dinner. Because a great deal of each day is spent forming intentions and acting on those intentions, it is no surprise that failures in prospective memory constitute at least half of everyday forgetting (Crovitz and Daniel, 1984).

Failures in prospective memory can have critical, even life-threatening consequences. One epidemiological study found that forgetting to take one's blood pressure medication at least one time significantly increased the risk of heart attack and death (Nelson et al., 2006). Prospective memory is also important in the workplace. For example, in the intensive care unit, nurses have many health-relevant prospective memory demands. Failing to remember to deliver medications, to redress wounds, or to properly chart changes in vital signs can result in poor health outcomes for the patient. Moreover, in aviation, aircraft pilots must remember to perform several actions sequentially prior to take-off and landing and failure to remember to perform any of these actions may result in injury or death. Although airline accidents are rare, Dismukes (2006) observed that almost one-fifth of major airline accidents can be attributed to prospective memory failures. Thus, prospective memory is not only ubiquitous in our lives, but it is also important to maintaining health and safety (Dismukes, 2012).

The Empirical Study of Prospective Memory

Early prospective memory research assessed the frequency of failures in prospective memory relative to other kinds of memory. These were typically diary studies in which participants recorded all of their memory failures during a set period of time. In one diary study, Terry (1988) observed that 70% of daily forgetting errors were related to prospective memory, indicating that prospective memory has been underrepresented in memory research.

The identification of prospective memory as an under-recognized topic helped to spark several objective research studies. The first objective studies of prospective memory took a seminaturalistic approach, asking participants to mail postcards (Meacham and Singer, 1977) or to call the laboratory (Moscovitch, 1982) at predetermined times. These studies effectively captured some of the naturalistic components of prospective memory tasks and stimulated interest in factors thought to influence prospective remembering (e.g., the age of the participants), but they were limited by a lack of experimenter control over the use of external aids such as calendars.

Laboratory Approaches

The introduction of laboratory paradigms (Einstein and McDaniel, 1990; Kvavilashvili, 1987) for studying prospective memory greatly promoted research by allowing experimenters the ability to control and manipulate variables hypothesized to have influence on prospective remembering. In the typical laboratory paradigm, participants are first busily engaged in an ongoing decision-making task such as rating the pleasantness of words. After receiving ongoing task instructions, participants receive instructions to perform an unrelated task at some predetermined point in the future – for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51055-1

example, to press the *Q* key when the word *coffee* appears during the pleasantness-rating trials. The prospective memory action (e.g., press the *Q* key) is usually designed to be simple and easy to remember. By keeping the contents of the intention simple (i.e., the retrospective component of the prospective memory task), researchers are better able to isolate the components that are unique to prospective memory.

Distinguishing Characteristics of Prospective Memory

There are three primary distinguishing features of a prospective memory task. First, whereas retrospective memory tasks are past-focused, prospective memory tasks are future-focused. Second, prospective memory tasks (e.g., press *Q* for *coffee*) are always embedded within an ongoing activity (e.g., pleasantness rating), and they require the interruption of that ongoing activity. The purpose here is to simulate real-life prospective memory situations in which people are actively engaged in an ongoing activity (e.g., driving home), but must suspend that activity in order to perform a prospective memory intention (e.g., stopping at the post office). The third, and perhaps most important, distinguishing feature of prospective memory tasks is that retrieval of the prospective memory intention is self-initiated. When the appropriate cue to perform the intended action occurs (e.g., the word *coffee*), there is no explicit prompt to perform the prospective memory task. Unlike in cued recall (retrospective memory), where the experimenter gives participants a cue word and specifically asks the participant to retrieve the associated word, in a prospective memory task, participants must switch (without external prompting) from seeing the prospective memory cue as another ongoing task stimulus to a cue to initiate the intended action. Therefore, prospective memory is often referred to as the ability to 'remember to remember.'

Categories of Prospective Memory Tasks

Although all prospective memory tasks are future-focused, embedded within an ongoing activity, and require self-initiated retrieval, subcategories of prospective memory tasks have emerged. In 'event-based' tasks, an environmental cue indicates that the prospective memory action should be performed (in the above example, the word *coffee* is an event-based cue). Real-world event-based prospective memory tasks include remembering to deliver a message (intention) when you see your friend (cue) and remembering to turn off to go to the grocery store (intention) when driving by a particular intersection (cue). Prospective memory tasks may also be 'time-based.' In time-based prospective memory tasks, one must remember to perform the prospective memory response after a certain length of time has elapsed. For example, to chill a can of soda, one might place it in the freezer with the intention of removing it in approximately 20 min (failing to remember would result in a sticky mess to clean up).

The Stages of Prospective Memory

Like retrospective memory, prospective memory also can be described in terms of three stages: encoding, storage, and retrieval. Research has focused on understanding the differences between retrospective memory and prospective memory at each of these stages. These differences are illustrated in Figure 1.

Encoding

One of the key functions of episodic memory is planning for future events (Klein et al., 2010). Yet, planning is rarely incorporated into traditional retrospective memory tests. By contrast, encoding a prospective memory intention may involve a high degree of planning (Figure 1). For example, one must plan how and when to accomplish the intention – such information is typically not necessary to consider when encoding information in retrospective memory. The precise amount of planning involved, however, varies greatly across prospective memory tasks. For example, remembering to remove pasta from the cooktop after 10 min may require little planning when that is the only item that you are preparing for dinner, but fulfilling the same intention may require more planning when you are also busily preparing vegetables, fish, and dessert.

Research on the planning phase of prospective memory has yielded two general conclusions. First, there are clear individual differences in the quality and complexity of the plans forged to complete intentions (Kliegel et al., 2000). Second, though individuals commonly form general prospective memory intentions, such as "I want to exercise more," these general intentions are often never initiated. A far more effective prospective memory encoding strategy, which is referred to as implementation intention encoding, is to form a very specific plan (e.g., "I will go to the gym to use the treadmill at 2 p.m. after my Chemistry class"; Gollwitzer, 1999). Implementation intention research is described more fully in the Improving Prospective Memory section.

Storage

The storage of prospective memories is in some ways similar to the storage of retrospective memories (Figure 1). For example, nocturnal sleep benefits the consolidation, or the long-term storage, of both prospective and retrospective memories (Scullin and McDaniel, 2010). Slow-wave sleep may be particularly important for consolidation of both retrospective and prospective memories (Diekelmann et al., 2013). Despite this similarity, research on the shorter-term storage of memories suggests that prospective memory storage has at least two characteristics that distinguish it from retrospective memory storage. Whereas retrospective memory has long been known to decline across increasing retention intervals (Ebbinghaus, 1885/1964), prospective memory may decline, remain stable, or even increase with longer retention intervals (Hicks et al., 2000; Martin et al., 2011). In addition, studies of levels of memory activation suggest that prospective memory intentions are stored at a higher state of activation than other memory contents. Research shows that, following encoding of a prospective memory intention, participants are faster to respond to words that are associated with their intention during recognition tasks (Goschke and Kuhl, 1993) and lexical decision tasks (Marsh et al., 1998). These 'intention superiority' effects are only observed when the individual will personally perform the intention, and not when the individual is expecting to watch someone else perform the intention, or when the

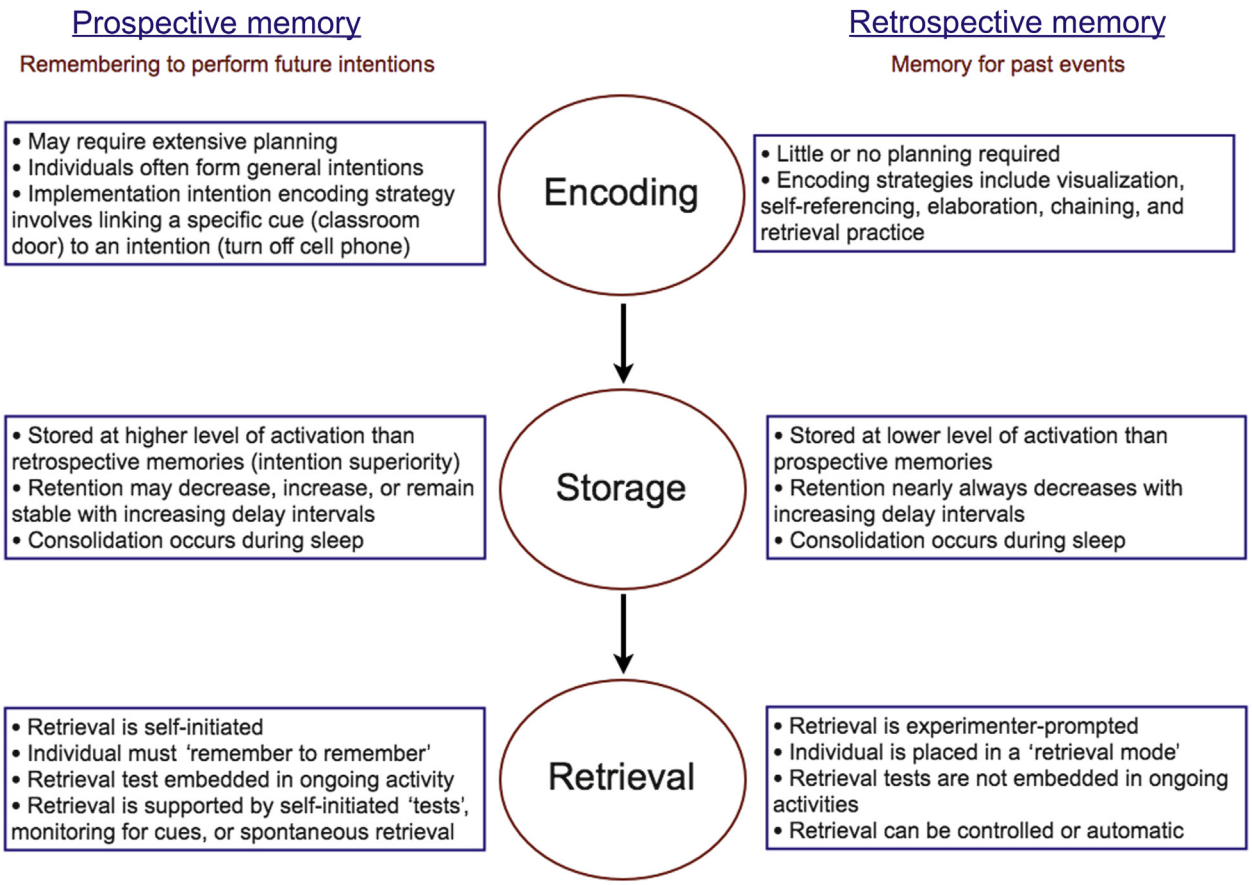

Figure 1 The stages of prospective memories and retrospective memories. Many factors distinguish the encoding, storage, and retrieval stages of prospective memories from retrospective memories.

intention has already been completed (Scullin et al., 2009). Thus, to-be-performed prospective memory intentions exist at a heightened level of activation, and this intention superiority allows intentions to come to mind quickly in response to relevant environmental cues (Goschke and Kuhl, 1993).

Retrieval

The most critical distinguishing feature of prospective memory tasks is that retrieval is self-initiated (Figure 1). Accordingly, much research has focused on understanding the processes operating at prospective memory retrieval. One possibility is that, after forming a prospective memory intention, people begin to engage attentional resources to monitor the environment for the appropriate time/cue to carry out the action. The Test-Wait-Test-Exit model (Harris, 1984; Harris and Wilkins, 1982) contends that people periodically 'test' whether it is the appropriate time to carry out their intention. In this scenario, an individual self-initiates prospective memory retrieval and determines ('tests') whether it is time to perform the intended action. If a 'test' reveals that it is too early to perform the intended action, then normal activities are resumed until another 'test' is initiated. If the 'test' reveals that it is time to perform the intended action, then the action is performed and the intention is disengaged ('exited'). In time-based prospective memory tasks, participants may engage in periodic clock-checking behavior to determine whether it is the appropriate time to perform the action (e.g., "Is it time to remove the soda from the freezer?"). People typically engage in more frequent clock-checking behavior as the time to perform the intention approaches, and this strategic increase in clock checking is associated with better prospective memory performance (Harris and Wilkins, 1982). The Test-Wait-Test-Exit model has received empirical support for time-based prospective memories. The primary limitation of the Test-Wait-Test-Exit model is that it does not specify a mechanism for how people self-initiate the 'test.'

Preparatory Attentional and Memory Processes Theory

One possibility is that the cognitive system actively searches for cues that signal that the prospective memory intention should be performed. For example, the preparatory attentional and memory processes (PAM) theory (Smith, 2003), which may be conceptualized as an extrapolation of the Test-Wait-Test-Exit model (Harris and Wilkins, 1982) to event-based prospective memory, contends that people will periodically recollect their plans (cf 'test'), which can subsequently lead the individual to monitor their environment for cues that signal that the prospective memory intention should be performed. Consider the concrete example of remembering to buy a loaf of bread on the way home. The PAM theory assumes that we periodically self-initiate retrieval of this plan

throughout the day (e.g., during breaks when we consider "what we later have to do"), and if we later self-initiate retrieval of the plan when we are driving home then we will begin to monitor the environment for signs of the grocery store. An important component of the PAM theory is that prospective memory retrieval cannot occur in the absence of monitoring and that failures in prospective memory are due either to lapses in devoting attentional resources to monitoring or to a retrospective memory loss of the contents of the intention (e.g., forgetting that it is bread that needs to be purchased from the grocery store).

Support for the PAM theory of prospective memory monitoring comes from studies showing that trying to remember to perform a prospective memory intention can detract attention from other ongoing activities. The argument is that monitoring is effortful and requires attentional resources that could otherwise be devoted toward efficiently making ongoing task decisions. Therefore, if one is monitoring for the grocery store on the way home, then one's driving performance might be impaired. Consistent with this argument, participants who have a prospective memory demand often perform an ongoing task more slowly or less accurately than participants who perform the same ongoing task alone (e.g., Marsh et al., 2003; Park et al., 1997; Smith, 2003).

Multiprocess Theory

Like the PAM theory, the multiprocess theory (McDaniel and Einstein, 2000) assumes that prospective memory retrieval is sometimes accomplished via attentionally demanding monitoring processes. An important difference between these theories, however, is that the multiprocess theory also assumes that relatively automatic, 'spontaneous retrieval' processes can support many instances of prospective remembering. This view arises from the observation that people often have multiple prospective memory demands, many of which may not be executed for hours or days (e.g., remembering to call your grandmother on her birthday), and it is often difficult to precisely anticipate when a prospective memory action can be completed (e.g., delivering a message to a colleague at some point while at work). For these reasons, it seems unlikely that people would (or could) engage in attentionally demanding monitoring processes to remember all of their prospective memory intentions all of the time. But, if people can still remember to perform prospective memory intentions even if they do not constantly monitor, then there must also exist a less-effortful retrieval mechanism. The multiprocess theory contends that in the absence of monitoring, environmental cues that are strongly associated with a prospective memory intention can stimulate the retrieval of the associated intention (spontaneous retrieval). For example, after forming the intention to invite a friend to an upcoming party, passing by that friend in the hallway may cause the associated intention to spontaneously 'pop' into conscious awareness (for an analog in the autobiographical memory literature, see Berntsen, 2009).

The question of whether the occurrence of a target event (e.g., the grocery store) that has been associated with an intention can spontaneously trigger retrieval of an episodic memory (e.g., the intention to stop at the store and buy bread) is theoretically interesting. On the one hand, stimuli in our environment rarely distract us from the task at hand: When enjoying a conversation with a friend in a restaurant, the objects and events in our surroundings usually do not elicit extraneous episodic memories (except for individuals with posttraumatic stress disorder; Berntsen, 2010). Based partially on this observation, Tulving (1983) proposed the interesting idea that retrieval does not occur unless we are in a 'retrieval mode.' From this retrieval-mode perspective, you have to be trying to remember (e.g., engaging preparatory attentional processes) in order to retrieve an intention in response to processing a target event. By contrast, the multiprocess theory proposes that being in a retrieval mode is not a condition for episodic memory retrieval and that cues associated with intentions can spontaneously trigger retrieval of those intentions.

Recent research has provided strong evidence that spontaneous retrieval processes support many instances of prospective remembering. The overall approach in these studies is to create conditions of no monitoring and to examine whether successful prospective remembering occurs in the absence of monitoring. In a series of experiments that are illustrated in Figure 2, Scullin et al. (2010) eliminated monitoring by requiring participants to perform a block of over 500 ongoing task trials. Because monitoring is effortful, people are typically unable to sustain monitoring over long blocks of trials, especially when monitoring is unreinforced (i.e., when there are no opportunities to perform the prospective memory task). By the end of the block monitoring had dissipated completely, yet when a strong ('focal') prospective memory cue was presented on Trial 501, 73% of participants still remembered to perform the prospective memory action. The participants who successfully remembered to perform the prospective memory action showed no evidence for monitoring proximal to Trial 501. Therefore, monitoring is not always required for successful prospective remembering; in the absence of monitoring, seeing the sign for the grocery store can spontaneously trigger retrieval of the associated intention to pick up bread.

Given the findings that prospective memory is sometimes supported by monitoring and other times supported by spontaneous retrieval, much research has been devoted to identifying and examining factors that determine reliance on these retrieval processes (e.g., Einstein et al., 2005). For example, people are more likely to monitor when the importance of the prospective memory task is emphasized and when the prospective memory task must be performed often. Conversely, reliance on spontaneous retrieval is greater when there is a long delay between the formation of the intention and the opportunity to complete that action, and also when the ongoing task encourages focal processing of the prospective memory cues (i.e., processing of the relevant features of the cues that were also processed at encoding), which is demonstrated in Figure 2. The most recent development of the multiprocess theory contends that monitoring and spontaneous retrieval processes interact dynamically to support prospective memory in naturalistic settings (Scullin et al., 2013).

Neural Correlates of Prospective Memory

Much of the existing research on prospective memory has been behavioral, but there is now growing interest in understanding the neural underpinnings of prospective memory. Such

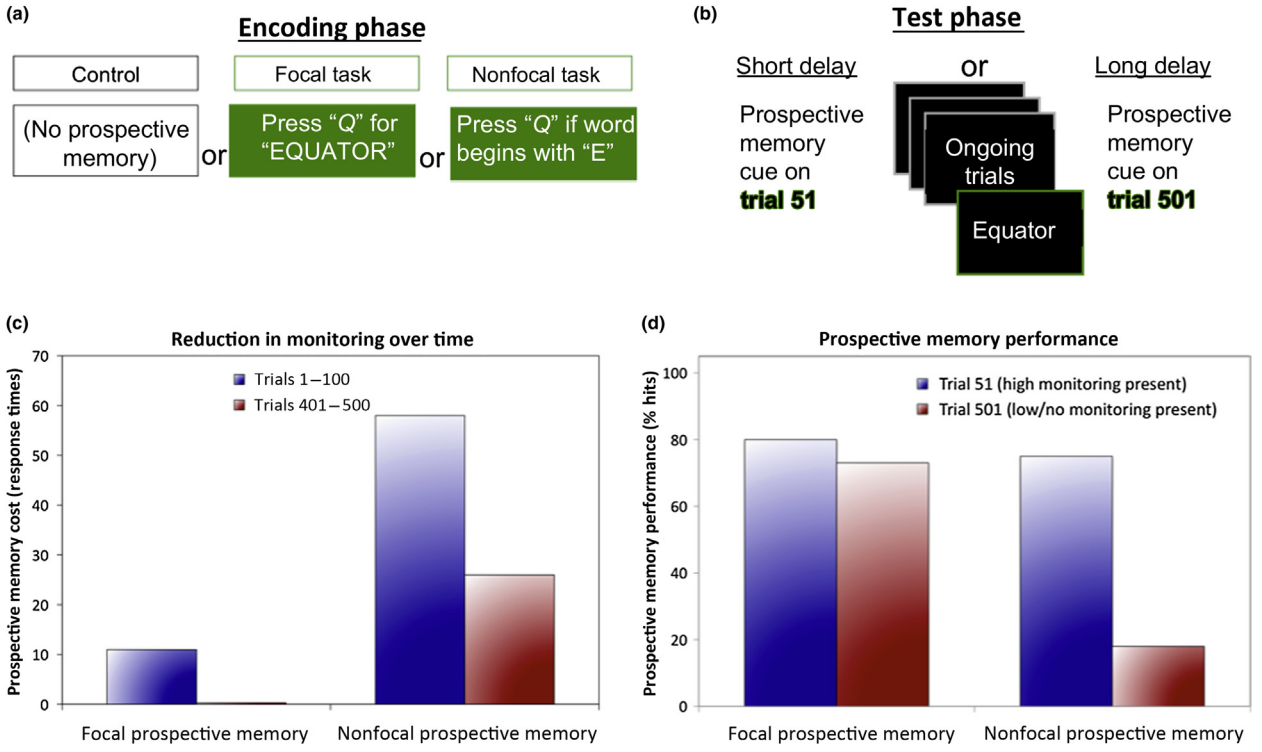

Figure 2 (a) Participants encoded either a focal or a nonfocal prospective memory task (or no prospective memory task, as in the control condition) and then (b) were tested following a short or long retention interval. (c) The results showed that because monitoring is resource-consuming it wanes over time when the prospective memory intention is not actively being performed (zero cost indicates no monitoring). (d) When monitoring is absent (focal condition, Trial 501), spontaneous retrieval processes can still support prospective memory retrieval in response to good retrieval cues (focal cues). However, prospective memory performance in response to poor retrieval cues (nonfocal cues) drops dramatically when monitoring is reduced. Adapted from Scullin, M.K., McDaniel, M.A., Shelton, J.T., Lee, J.H., 2010. Focal/nonfocal cue effects in prospective memory: monitoring difficulty or different retrieval processes? Journal of Experimental Psychology: Learning, Memory, and Cognition 36, 736–749, (Experiments 3 and 4).

research has primarily focused on the retrieval phase of prospective memory rather than on encoding or storage stages. The literature suggests that two structures, the anterior prefrontal cortex and the medial temporal lobe, may be critical for monitoring and spontaneous retrieval, respectively.

Anterior Prefrontal Cortex

The anterior prefrontal cortex has been implicated in executive control functions, which thereby makes it a candidate for supporting effortful prospective memory monitoring processes. Case studies of prospective memory have supported this idea. For example, Burgess et al. (2000) found that patients with lesions in areas including the anterior prefrontal cortex demonstrated more prospective memory impairments in a task that required effortful monitoring than age- and intelligence-matched controls.

Human neuroimaging studies have further supported a role for the anterior prefrontal cortex in prospective memory. Using positron emission tomography (PET), Burgess et al. (2001) found that the anterior prefrontal cortex demonstrated greater activation during intervals in which an individual was monitoring for a prospective memory intention than during a baseline interval in which there was no prospective memory intention. In addition, using functional magnetic resonance imaging, Reynolds et al. (2009) also demonstrated that the anterior prefrontal cortex is engaged to support prospective memory monitoring, and that the neural signature of prospective memory monitoring is distinct from other cognitive constructs such as vigilance and working memory. All existing functional neuroimaging prospective memory studies have been conducted on event-based prospective memory. Whether time-based prospective memories are supported by the same anterior prefrontal cortex system is an open question.

Medial Temporal Lobe

Neuropsychological studies of prospective memory have also investigated the role of the medial temporal lobe in supporting spontaneous retrieval of intentions. The medial temporal lobe, which includes the hippocampus, has previously been implicated as a structure that supports reflexive-associative retrieval (Moscovitch, 1994). Evidence for a role of the medial temporal lobe in spontaneous retrieval of prospective memories arises from studies of clinical populations. Individuals who show a very early stage of dementia of the Alzheimer's type are also highly likely to show medial temporal lobe declines, but perhaps not prefrontal cortex declines, relative to age-matched controls (Buckner, 2004). Therefore, such patients provide a useful model for testing the role of the medial temporal lobe in prospective memory. As illustrated in

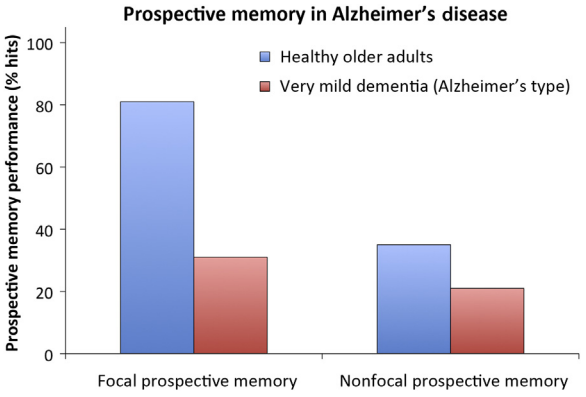

Figure 3 Very mild dementia of the Alzheimer's type is associated with declines in focal prospective memory (spontaneous retrieval) but not nonfocal prospective memory (monitoring). This figure was modified from McDaniel, M.A., Shelton, J.T., Breneiser, J.E., Moynan, S., Balota, D.A., 2011. Focal and nonfocal prospective memory performance in very mild dementia: a signature decline. Neuropsychology 25, 387–396.

Figure 3, McDaniel et al. (2011) found that patients with very mild Alzheimer's disease performed worse than healthy control participants on a prospective memory test that relied on spontaneous retrieval (focal prospective memory), but not on a test that relied on monitoring (nonfocal prospective memory). These results were consistent with the idea that the medial temporal lobe is a critical structure for supporting spontaneous retrieval of prospective memories.

Gordon et al. (2011) provided additional evidence for an association between the medial temporal lobe and spontaneous retrieval. Using structural MRI, they compared the integrity of regions of the medial temporal lobe to the ability to spontaneously retrieve prospective memories. They found a robust association between spontaneous-retrieval-dependent prospective memory performance and the structural volume of the medial temporal lobe (particularly the hippocampal subregion). Therefore, both the anterior prefrontal cortex and the medial temporal lobe are critical for supporting prospective memory.

Conditions Associated with Impaired Prospective Memory

Everyone experiences prospective memory failures, but some individuals appear to have worse prospective memory than others, and researchers have attempted to understand the underlying conditions that can lead to prospective memory impairments. Many studies have now been conducted on prospective memory in special populations that are at risk for having impairments in their anterior prefrontal cortex or hippocampal regions (i.e., the neural structures implicated in supporting prospective memory).

Substance Abuse

3,4-Methylenedioxymethamphetamine (MDMA) (Ecstasy)

Substance abuse, including the misuse or overuse of illegal drugs and alcohol, has long been known to affect neural and cognitive functioning. MDMA ('Ecstasy') use, for example, is linked to hippocampal and prefrontal cortex abnormalities (Cowan, 2007), and therefore one might anticipate MDMA users to demonstrate prospective memory impairments. Indeed, Rendell et al. (2007) found that MDMA users – even relatively infrequent users – show worse event-based and time-based prospective memory performance than nonusers.

Cannabis (Marijuana)

A second substance that is illegal in most countries and is commonly implicated in hippocampal impairments is cannabis (marijuana). Large-scale studies have found more frequent reports of everyday prospective memory failures in cannabis users than in nonusers (e.g., Cuttler et al., 2012). In addition, Montgomery et al. (2012) found that cannabis users performed much more poorly on naturalistic event-based tasks (e.g., remembering to turn on a coffee machine when a meeting begins) than nonusers, and that the impairments were exacerbated with higher frequency and doses of cannabis use.

Alcohol

Beyond illegal substances, other drugs such as alcohol have been linked to prospective memory impairments. For example, college-aged adults show more failures on time-based and event-based prospective memory tasks while intoxicated, even with relatively modest doses in which the participants do not subjectively perceive themselves to be intoxicated (Montgomery et al., 2011). Chronic alcohol abuse, which has been linked to impairments in hippocampal and prefrontal cortex functioning, may also have detrimental effects on prospective memory such that prospective memory functioning is impaired even when the individual is not intoxicated. Consistent with this idea, Heffernan et al. (2002) found that reported prospective memory failures were more frequent in chronic heavy alcohol users (i.e., individuals who ingest 30–60 units of alcohol per week for at least the last 5 years) than in low-dose/alcohol-free control individuals. The severity of chronic alcohol abuse may also be linked to prospective memory impairments. Korsakoff's Syndrome, which typically results from chronic alcohol abuse, is associated with profound frontal lobe impairments. Such frontal lobe impairments may explain why patients with Korsakoff's syndrome perform even worse than heavy users of alcohol on prospective memory tasks that rely on effortful monitoring (Brunfaut et al., 2000).

Neurodegenerative Diseases

Because prospective memory is critical for supporting health and for sustaining independent living (e.g., remembering to take medication, attending physician appointments, etc.) researchers have evaluated whether prospective memory is impaired in patient populations (Pirogovsky et al., 2012). Prospective memory impairments have been observed in several special populations including patients with schizophrenia, multiple sclerosis, attention-deficit disorder, AIDS due to human immunodeficiency virus (HIV) infection, and hypertension. However, the greatest amount of research has focused on prospective memory impairments in Alzheimer's disease and Parkinson's disease (see Figures 3 and 4,

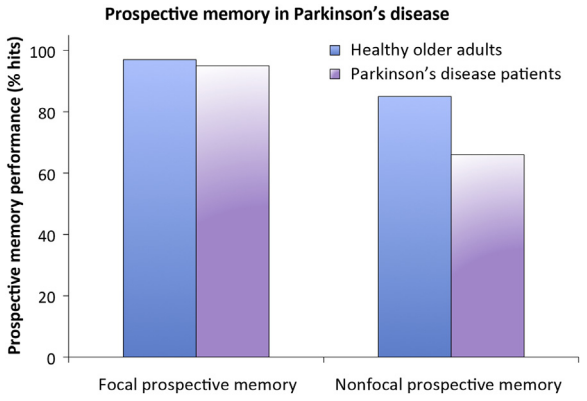

Figure 4 Parkinson's disease is associated with declines in nonfocal prospective memory (monitoring) but not focal prospective memory (spontaneous retrieval). This figure was adapted from Foster, E.R., McDaniel, M.A., Repovš, G., Hershey, T., 2009. Prospective memory in Parkinson disease across laboratory and self-reported everyday performance. Neuropsychology 23, 347. The difference between this study and the McDaniel et al. (2011) study in healthy older adult nonfocal prospective memory performance relates to changes in some experimental features.

for a neurological condition by prospective memory task dissociation).

Alzheimer's Disease

Alzheimer's disease, which is known for its impact on the hippocampus as well as other neural structures (Buckner, 2004), has been widely diagnosed based on poor retrospective memory functioning. However, one Memory Clinic reported that 40% of their patients believed their primary memory issues to be related to remembering to perform delayed intentions (i.e., prospective memory; Kliegel and Martin, 2003). Consistent with patients' reported experience, several studies using objective measures of prospective memory have reported profound prospective memory impairments in Alzheimer's disease patients (Figure 3). Though retrospective memory failures (i.e., not remembering the prospective memory instructions) can partially explain prospective memory failures in this population, prospective memory deficits are observed even after controlling for retrospective memory deficits (e.g., McDaniel et al., 2011). Researchers have recommended that prospective memory tests be used alongside retrospective memory tests for the early detection of Alzheimer's disease (Duchek et al., 2006).

Parkinson's Disease

Prospective memory functioning has also been studied in Parkinson's disease patients. Parkinson's disease is typically known for its motor symptoms (e.g., bradykinesia, tremor, etc.), which are caused by the death of substantia nigra dopaminergic cells, but this disease has also been associated with cognitive impairments due to impaired frontal lobe functioning. Prospective memory impairments are robust in Parkinson's disease patients, and as illustrated in Figure 4, these deficits are more likely to be observed during prospective memory tests that rely on monitoring (nonfocal prospective memory) than those that rely on spontaneous retrieval (focal prospective memory; Foster et al., 2009; Kliegel et al., 2011). Some prospective memory impairments in Parkinson's disease may be partially reduced by administration of dopaminergic medication, which benefits frontal lobe functioning (Costa et al., 2008).

Normal Aging

One of the most common topics of research on memory concerns the evaluation of how memory changes with normal aging. The field of prospective memory has been no different. The effect of aging on prospective memory was one of the earliest questions in the field (e.g., Einstein and McDaniel, 1990), and aging has remained such a focus of the field that there are now multiple metaanalyses published on the topic (e.g., Henry et al., 2004; Kliegel et al., 2008). Though early research suggested that prospective memory is spared in normal aging, subsequent research suggests that this is not always the case. In both time-based and event-based laboratory prospective memory tasks, younger adults often outperform older adults. Age differences emerge especially when the laboratory prospective memory task challenges effortful monitoring processes. Such results might be attributable to the neurophysiological changes observed in the anterior prefrontal cortex with increasing age (McDaniel and Einstein, 2011).

Age differences in prospective memory are particularly pronounced in situations when an intention is successfully retrieved but the individual must postpone executing the intention. For example, an older adult might retrieve the intention to take medication but then postpone this intention in order to answer a ringing telephone; the question is whether the older adult still remembers to take the medication after answering the phone. Einstein et al. (2000) captured this prospective memory challenge in the laboratory using the delay-execute paradigm. As in the typical prospective memory laboratory paradigm, participants are told to perform a prospective memory action (e.g., press *Q*) in response to an environmental cue (e.g., the word *coffee*). However, participants must delay executing the prospective memory response for 5–15 s following the environmental cue. The results are demonstrated in Figure 5. Relative to a baseline condition in which the prospective memory response could be executed immediately, younger adults demonstrated minimal declines in prospective memory performance when they had to temporarily delay intention execution. Older adults, on the other hand, demonstrated dramatic declines in prospective memory performance, even with a 5-second delay interval (McDaniel et al., 2003). Therefore, interrupting the ability to immediately execute a prospective memory intention might be one of the most important factors that lead to prospective memory impairments in older adults.

Though age differences are often observed in laboratory prospective memory tasks, older adults commonly perform as well as, if not better than, younger adults in naturalistic prospective memory studies. The typical explanation for this age – prospective memory paradox (Rendell and Thomson, 1999) – is that in naturalistic settings older adults more often rely on external cues, or even on strategies such as reminders and avoiding interruptions, to bolster their prospective memory performance. The section on Improving Prospective

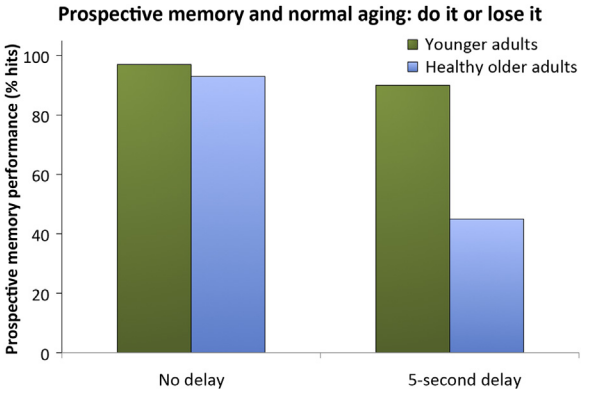

Figure 5 Results of the delay-execute paradigm. If an intention must be maintained in consciousness for 5 s prior to execution then prospective memory performance will drop dramatically in older adults but not in younger adults. The figure, and the phrase "do it or lose it" are adapted from McDaniel, M.A., Einstein, G.O., Stout, A.C., Morgan, Z., 2003. Aging and maintaining intentions over delays: do it or lose it. Psychology and Aging 18, 823 (Experiments 1 and 2b).

Memory elaborates on the approaches that have been found to be most viable for improving prospective remembering.

Improving Prospective Memory

Prospective memory failures can be embarrassing, such as forgetting to bring to class a reading assignment (perhaps even one on prospective memory) or failing to silence your phone before class and receiving a call. Failures in prospective memory can also be detrimental to health (e.g., failing to remember to take medication), and in extreme cases, they can even be fatal (e.g., a pilot who fails to remember to set the flaps during flight landing). Therefore, it is in every individual's best interest to know how to improve prospective memory.

Avoiding Interruptions

One line of research has focused on avoiding interruptions (cf deficits seen in Figure 5; Einstein et al., 2000). In the workplace environment, and especially healthcare environments such as the intensive care unit, nurses are constantly interrupted while attempting to perform prospective memory intentions (e.g., administering medication, recording vital signs). Grundgeiger et al. (2010) found that avoiding interruptions by insisting on completing an intention prior to focusing on the interrupting task is an optimal strategy for reducing prospective memory errors in healthcare.

Implementation Intention Encoding

Another line of research has attempted to boost prospective remembering at the encoding stage. Gollwitzer (1999) noted that people typically encode intentions in a very general manner (e.g., 'I need to call my mother'), and that general intentions are likely to go unfulfilled. The solution, referred to as implementation intention encoding (Gollwitzer, 1999), involves specifying a cue for when the intention should be executed (e.g., "When I finish my calculus class at 2 p.m., then I will call my mother") while visualizing performing that intention in one's mind. Benefits of implementation intention encoding on subsequent prospective memory performance have been demonstrated in several laboratory experiments (e.g., McDaniel et al., 2008). Furthermore, implementation intention encoding has produced benefits for real-world prospective memory intentions such as remembering to monitor blood glucose levels (Liu and Park, 2004), remembering to get an influenza vaccination (Milkman et al., 2011), and remembering to take Vitamin C supplements (Sheeran and Orbell, 1999).

Reminders and Smartphone Technology

A third line of research that focuses on improving prospective memory aims to remove the demanding cognitive components from prospective memory tasks. For example, using sticky notes (e.g., 'Go to the grocery store!') removes the 'memory' from 'prospective memory.' With advances in digital technology, reminders may be even more effective. Smartphones that are kept near their user and produce a sound or vibration concomitant with written text have the potential to act as reminders of prospective memory intentions. The benefit of smartphone applications and other digital reminder systems (e.g., Google calendar) for facilitating prospective remembering in everyday settings has been observed even in patients with traumatic brain injury and severe amnesia (McDonald et al., 2011; Svoboda and Richards, 2009). Such results provide promise for the future of improving prospective memory functioning in patient populations, the workplace, and everyday life.

See also: Aging and Memory; Alzheimer's Disease, Neural Basis of; Autobiographical Memory; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Declarative Memory, Neural Basis of; Episodic Memory; Future Thinking; Sleep and Memory.

Bibliography

Berntsen, D., 2009. Involuntary Autobiographical Memories: An Introduction to the Unbidden Past. Cambridge University Press.
Berntsen, D., 2010. The unbidden past involuntary autobiographical memories as a basic mode of remembering. Current Directions in Psychological Science 19, 138–142.
Brunfaut, E., Vanoverberghe, V., d'Ydewalle, G., 2000. Prospective remembering of Korsakoffs and alcoholics as a function of the prospective-memory and on-going tasks. Neuropsychologia 38, 975–984.
Buckner, R.L., 2004. Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. Neuron 44, 195–208.
Burgess, P.W., Quayle, A., Frith, C.D., 2001. Brain regions involved in prospective memory as determined by positron emission tomography. Neuropsychologia 39, 545–555.
Burgess, P.W., Veitch, E., de Lacy Costello, A., Shallice, T., 2000. The cognitive and neuroanatomical correlates of multitasking. Neuropsychologia 38, 848–863.
Costa, A., Peppe, A., Brusa, L., Caltagirone, C., Gatto, I., Carlesimo, G.A., 2008. Dopaminergic modulation of prospective memory in Parkinson's disease. Behavioural Neurology 19, 45–48.
Cowan, R.L., 2007. Neuroimaging research in human MDMA users: a review. Psychopharmacology 189, 539–556.
Crovitz, H.F., Daniel, W.F., 1984. Measurements of everyday memory: toward the prevention of forgetting. Bulletin of Psychonomic Society 22, 413–414.
Cuttler, C., McLaughlin, R.J., Graf, P., 2012. Mechanisms underlying the link between cannabis use and prospective memory. PLoS One 7, e36820.

Diekelmann, S., Wilhelm, I., Wagner, U., Born, J., 2013. Sleep to implement an intention. Sleep 36, 149–153.

Dismukes, R.K., 2006. Prospective memory in aviation and everyday settings. In: Kliegel, M., McDaniel, M.A., Einstein, G.O. (Eds.), Prospective Memory: Cognitive, Neuroscience, Developmental, and Applied Perspectives. Erlbaum, Mahwah, NJ.

Dismukes, R.K., 2012. Prospective memory in workplace and everyday situations. Current Directions in Psychological Science 21, 215–220.

Duchek, J.M., Balota, D.A., Cortese, M., 2006. Prospective memory and apolipoprotein E in healthy aging and early stage Alzheimer's disease. Neuropsychology 20, 633–644.

Ebbinghaus, H., 1964. Memory: A Contribution to Experimental Psychology (H.A. Ruger and C.E. Bussenius, Trans.). Dover, New York (Original work published 1885).

Einstein, G.O., McDaniel, M.A., 1990. Normal aging and prospective memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 717–726.

Einstein, G.O., McDaniel, M.A., Manzi, M., Cochran, B., Baker, M., 2000. Prospective memory and aging: forgetting intentions over short delays. Psychology and Aging 15, 671–683.

Einstein, G.O., McDaniel, M.A., Thomas, R., Mayfield, S., Shank, H., Morrisette, N., Breneiser, J., 2005. Multiple processes in prospective memory retrieval: factors determining monitoring versus spontaneous retrieval. Journal of Experimental Psychology: General 134, 327–342.

Foster, E.R., McDaniel, M.A., Repovš, G., Hershey, T., 2009. Prospective memory in Parkinson disease across laboratory and self-reported everyday performance. Neuropsychology 23, 347.

Gollwitzer, P.M., 1999. Implementation intentions: strong effects of simple plans. American Psychologist 54, 493–503.

Gordon, B.A., Shelton, J.T., Bugg, J.M., McDaniel, M.A., Head, D., 2011. Structural correlates of prospective memory. Neuropsychologia 49, 3795–3800.

Goschke, T., Kuhl, J., 1993. The representation of intentions: persisting activation in memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 1211–1226.

Grundgeiger, T., Sanderson, P., MacDougall, H.G., Venkatesh, B., 2010. Interruption management in the intensive care unit: predicting resumption times and assessing distributed support. Journal of Experimental Psychology: Applied 16, 317–334.

Harris, J.E., 1984. Remembering to do things: a forgotten topic. In: Harris, J.E., Morris, P.E. (Eds.), Everyday Memory, Actions, and Absent-Mindedness. Academic Press, London, pp. 71–92.

Harris, J.E., Wilkins, A.J., 1982. Remember to do things: a theoretical framework and an illustrative experiment. Human Learning 1, 123–136.

Heffernan, T.M., Moss, M., Ling, J., 2002. Subjective ratings of prospective memory deficits in chronic heavy alcohol users. Alcohol and Alcoholism 37, 269–271.

Henry, J.D., MacLeod, M.S., Phillips, L.H., Crawford, J.R., 2004. A meta-analytic review of prospective memory and aging. Psychology and Aging 19, 27–39.

Hicks, J.L., Marsh, R.L., Russell, E.J., 2000. The properties of retention intervals and their effect on retaining prospective memories. Journal of Experimental Psychology: Learning, Memory, and Cognition 26, 1160–1169.

Klein, S.B., Robertson, T.E., Delton, A.W., 2010. Facing the future: memory as an evolved system for planning future acts. Memory and Cognition 38, 13–22.

Kliegel, M., Altgassen, M., Hering, A., Rose, N.S., 2011. A process-model based approach to prospective memory impairment in Parkinson's disease. Neuropsychologia 49, 2166–2177.

Kliegel, M., Jäger, T., Phillips, L.H., 2008. Adult age differences in event-based prospective memory: a meta-analysis on the role of focal versus nonfocal cues. Psychology and Aging 23, 203–208.

Kliegel, M., Martin, M., 2003. Prospective memory research: why is it relevant? International Journal of Psychology 38, 193–194.

Kliegel, M., McDaniel, M.A., Einstein, G.O., 2000. Plan formation, retention, and execution in prospective memory: a new approach and age-related effects. Memory and Cognition 28, 1041–1049.

Kvavilashvili, L., 1987. Remembering intention as a distinct form of memory. British Journal of Psychology 78 (4), 507–518.

Liu, L.L., Park, D.C., 2004. Aging and medical adherence: the use of automatic processes to achieve effortful things. Psychology and Aging 19, 318–325.

Marsh, R.L., Hicks, J.L., Bink, M.L., 1998. Activation of completed, uncompleted, and partially completed intentions. Journal of Experimental Psychology: Learning, Memory, and Cognition 24, 350–361.

Marsh, R.L., Hicks, J.L., Cook, G.I., Hansen, J.S., Pallos, A.L., 2003. Interference to ongoing activities covaries with the characteristics of an event-based intention. Journal of Experimental Psychology: Learning, Memory, and Cognition 29, 861–870.

Martin, B.A., Brown, N.L., Hicks, J.L., 2011. Ongoing task delays affect prospective memory more powerfully than filler task delays. Canadian Journal of Experimental Psychology 65, 48–56.

McDaniel, M.A., Einstein, G.O., 2000. Strategic and automatic processes in prospective memory retrieval: a multiprocess framework. Applied Cognitive Psychology 14, S127–S144.

McDaniel, M.A., Einstein, G.O., 2011. The neuropsychology of prospective memory in normal aging: a componential approach. Neuropsychologia 49, 2147–2155.

McDaniel, M.A., Einstein, G.O., Stout, A.C., Morgan, Z., 2003. Aging and maintaining intentions over delays: do it or lose it. Psychology and Aging 18, 823.

McDaniel, M.A., Howard, D.C., Butler, K.M., 2008. Implementation intentions facilitate prospective memory under high attention demands. Memory and Cognition 36, 716–724.

McDaniel, M.A., Shelton, J.T., Breneiser, J.E., Moynan, S., Balota, D.A., 2011. Focal and nonfocal prospective memory performance in very mild dementia: a signature decline. Neuropsychology 25, 387–396.

McDonald, A., Haslam, C., Yates, P., Gurr, B., Leeder, G., Sayers, A., 2011. Google calendar: a new memory aid to compensate for prospective memory deficits following acquired brain injury. Neuropsychological Rehabilitation 21, 784–807.

Meacham, J.A., Singer, J., 1977. Incentive effects in prospective remembering. The Journal of Psychology 97, 191–197.

Milkman, K.L., Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C., 2011. Using implementation intentions prompts to enhance influenza vaccination rates. Proceedings of the National Academy Of Sciences, USA 108, 10415–10420.

Montgomery, C., Ashmore, K.V., Jansari, A., 2011. The effects of a modest dose of alcohol on executive function and prospective memory. Human Psychopharmacology 26, 208–215.

Montgomery, C., Seddon, A.L., Fisk, J.E., Murphy, P.N., Jansari, A., 2012. Cannabis-related deficits in real-world memory. Human Psychopharmacology 27, 217–225.

Moscovitch, M., 1982. A neuropsychological approach to memory and perception in normal and pathological aging. In: Craik, F.I.M., Trehub, S. (Eds.), Aging and Cognitive Processes. Plenum, New York, pp. 55–78.

Moscovitch, M., 1994. Memory and working with memory: evaluation of a component process model and comparisons with other models. In: Schacter, D.L., Tulving, E. (Eds.), Memory Systems. MIT Press, Cambridge, MA, pp. 269–310.

Nelson, M.R., Reid, C.M., Ryan, P., Willson, K., Yelland, L., 2006. Self-reported adherence with medication and cardiovascular disease outcomes in the second Australian national blood pressure study (ANBP2). The Medical Journal of Australia 9, 487–489.

Park, D.C., Hertzog, C., Kidder, D.P., Morrell, R.W., Mayhorn, C.B., 1997. Effect of age on event-based and time-based prospective memory. Psychology and Aging 12, 314.

Pirogovsky, E., Woods, S.P., Filoteo, J.V., Gilbert, P.E., 2012. Prospective memory deficits are associated with poorer everyday functioning in Parkinson's disease. Journal of the International Neuropsychological Society 18, 1–10.

Rendell, P.G., Gray, T.J., Henry, J.D., Tolan, A., 2007. Prospective memory impairment in "ecstasy" (MDMA) users. Psychopharmacology 194, 497–504.

Rendell, P.G., Thomson, D.M., 1999. Aging and prospective memory: differences between naturalistic and laboratory tasks. Journal of Gerontology: Psychological Sciences 54B, P256–P269.

Reynolds, J.R., West, R., Braver, T., 2009. Distinct neural circuits support transient and sustained processes in prospective memory and working memory. Cerebral Cortex 19, 1208–1221.

Scullin, M.K., Einstein, G.O., McDaniel, M.A., 2009. Evidence for spontaneous retrieval of suspended but not finished prospective memories. Memory and Cognition 37, 425–433.

Scullin, M.K., McDaniel, M.A., 2010. Remembering to execute a goal: sleep on it! Psychological Science 21, 1028–1035.

Scullin, M.K., McDaniel, M.A., Shelton, J.T., 2013. The Dynamic Multiprocess Framework: Evidence from prospective memory with contextual variability. Cognitive Psychology 67, 55–71.

Scullin, M.K., McDaniel, M.A., Shelton, J.T., Lee, J.H., 2010. Focal/nonfocal cue effects in prospective memory: monitoring difficulty or different retrieval processes? Journal of Experimental Psychology: Learning, Memory, and Cognition 36, 736–749.

Sheeran, P., Orbell, S., 1999. Augmenting the theory of planned behavior: roles for anticipated regret and descriptive norms. Journal of Applied Social Psychology 29, 2107–2142.

Smith, R.E., 2003. The cost of remembering in event-based prospective memory: investigating the capacity demands of delayed intention performance. Journal of Experimental Psychology: Learning, Memory, and Cognition 29, 347–361.

Svoboda, E., Richards, B., 2009. Compensating for anterograde amnesia: a new training method that capitalizes on emerging smartphone technologies. Journal of the International Neuropsychological Society 15, 629–638.

Terry, W.S., 1988. Everyday forgetting: data from a diary study. Psychological Reports 62, 299–303.

Tulving, E., 1983. Elements of Episodic Memory. Clarendon Press, Oxford, pp. 56–62.

Prostitution and Sex Work

Jennifer Musto, Wellesley College, Wellesley, MA, USA
Crystal A Jackson, John Jay College of Criminal Justice-CUNY, New York, NY, USA
Elena Shih, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by Davidson O'Connell, volume 18, pp. 12244–12247, © 2001, Elsevier Ltd.

Abstract

Prostitution is a gendered, sexualized, and racialized labor system, one that typically involves the exchange of sexual services for money, goods, or other benefits. Sex work encompasses different types of intimate arrangements that blur the boundaries between erotic, emotional, and economic labor. Sex work is also part of an industry and commercial market that is global in reach and diverse in its spatial, legal, and occupational organization. This article investigates the multifaceted dimensions of prostitution and sex work in a twenty-first century context. In it, we explore how sex work as an intimate form of labor has evolved alongside other social and economic shifts, such as technology innovations and the growth of neoliberal globalization. We further examine the divergent contexts in which sex work takes place, how scholars have differentially approached its study and how states' employment of disparate regulatory strategies and antiprostitution policies have, in many instances, stigmatized sex workers.

Prostitution, Sex Work, and the Commodification of Sexual Activity

Accounting for twenty-first century manifestations of prostitution or 'sex work' – a term used interchangeably throughout this article – is a cumbersome exercise. This is because prostitution is neither a universal empirical phenomenon, nor does it express streamlined meanings or definitions (Ditmore, 2006). To examine prostitution within a social science framework invites consideration of the divergent environments in which it takes place. It further demands attention to the varied ways in which sex workers' gender, race, class, sexual identity, and citizenship status intersect in shaping their experiences of sexual labor and the uneven degrees of oppression, discrimination, and criminalization they may face (Kempadoo, 2001; Kempadoo and Doezema, 1998; Kempadoo et al., 2005; Vanwesenbeeck, 2001; Wahab, 2005; Weitzer, 2005).

Broadly defined, prostitution is a gendered, sexualized, and racialized labor and economic arrangement that commonly involves the exchange of sexual services for money or goods (Peppet, 2013: 10, Showden and Majic, 2014: xiii). Sex work may also encompass a range of intimate labors (Constable, 2009: 50) such as stripping, bar work, massage, camming, tourist encounters, and "the girlfriend experience" (Bernstein, 2007) – among others – and include more than "direct sex-for-money exchanges" (Hoang, 2010: 256). This article focuses on physical acts of sexual intimacy that are traded in some form. Although erotic dance and porn are commonly included under the umbrella of sexual labor, we exclude them here for space and specificity.

The examples we discuss here aim to illuminate key themes under discussion; however, there is undoubtedly variance across national contexts that were not explored in depth. Due to space constraints, certain topics are also left unexamined, including research on customers, in-depth country histories, specific workplace practices, the experiences of transgender and male sex workers, sex-trade involved youth, and the role that market mediators, managers and 'pimps' play in facilitating sexual labor.

Precise estimates of the scope and scale of prostitution are difficult to track (Brents and Hausbeck, 2010: 13; Weitzer, 2007; Vanwesenbeeck, 2001). Complicating assessments of prostitution further still are frequent scholarly and policy references to data based on imprecise and inflated numbers (Wagenaar et al., 2013: 26, 87). Sometimes framed under the umbrella of a "global sex trade" (Kempadoo, 2001), prostitution and its attendant markets are invariably localized even while its local emergence is shaped by migration patterns, neoliberal globalization (Bernstein, 2007; Chin, 2013), and domestic and intergovernmental policies. The state plays a central role in how individuals experience sex work and any corresponding violence and stigma they encounter (Bernstein and Schaffner, 2005; Kelly, 2008; Sanders et al., 2009).

Legal, Spatial, and Sociotechnical Dimensions of Sex Work

Prior research has examined workers' motivations for entering sex work (Hardy et al., 2010; Abel et al., 2010; Sanders, 2005; Katsulis, 2010) and explored the physical, aesthetic, and emotional dynamics that accompany sexual labor (Brents and Jackson, 2013; Brewis and Linstead, 2000; Chapkis, 1997; Sanders, 2005). These studies begin to offer classifications of varying sex industries that include everything from brothels to independent escort services. They also consider how place, space and class shape intimate exchanges, suggesting that the meaning, terms, and empirical reality of prostitution is contingent and contextually specific.

Scholars have also focused on geographies of sex work (Hubbard, 1999, 2013; Bernstein, 2007; Katsulis, 2010; Kingston, 2014; Wonders and Michaloski, 2001), interrogating its spatial and legal dimensions. Sex work may take place in 'offline' outdoor and street-based settings or indoor venues

such as clubs, hotels, brothels, and residential areas. Whether focused on rural or urban settings or the specific locations where clients are secured and exchanges occur, the spatial organization of commercial sex is significant in that it impacts workers' experiences of sexual labor and additionally shapes community responses to it (Hubbard, 2001, 2004).

The growth and expanded use of online, mobile, and network technologies signals that sex work is increasingly mediated by and through technology. For instance, mobile phones and online classified ad sites enable sex workers and third party actors to advertise sexual services to prospective customers at low costs (Peppet, 2013). Technologies also create new niche markets and expand the possibilities for different kinds of commodified sexual encounters. Camming, for example, involves models performing web-based sex shows (Richtel, 2013) that are accessible by viewers from all over the globe. On one hand, this may give adult consensual sex workers or those who coordinate their services the ability to screen potential clients (Peppet, 2013: 2008–2010) and wield greater control over their working environments. On the other hand, online ad postings and mobile communication create new mechanisms for law enforcement to surveil the online, mobile, and social media activities of voluntary sex workers and as well as individuals deemed at risk for forced prostitution and "sex trafficking." (Musto and boyd, 2014). Such surveillance may lead to heightened criminalization and the entrapment of immigrant and domestic sex workers alike, thus underscoring that though technology stands to reconfigure how sex work is organized and facilitated, sexual labor remains highly stigmatized and closely monitored by the state.

Countries today are grappling with new and different approaches to regulating the sale of sex (Dewey and Kelly, 2011; Goodyear and Weitzer, 2011; Kempadoo et al., 2005; Weitzer, 2012). In addition to controlling the sale of sex itself, policies focused on other phenomena such as immigration, public health, morality, and sex trafficking also impact how prostitution is regulated.

Goodyear and Weitzer (2011) argue that government regulation falls into two general categories: "(1) intensified *criminalization* designed to eradicate it; and (2) *liberalization* based on principles of harm reduction, labor rights, and civil regulation of commerce" (16, italics in original). The Netherlands, a country where prostitution is legally permitted has introduced municipally orchestrated licensing schemes (Outshoorn, 2014: 174). Canada also regulates sex work (Van der Meulen et al., 2013), but through a combination of licensing requirements and criminalization (Auger, 2014: 101). Other countries have regulated prostitution through different forms of criminalization. In 1999, the Swedish government moved to criminalize the purchase of sex (Levy and Jakobsson, 2014: 5). Proponents of the Nordic 'End Demand' strategy viewed it as a way to reduce prostitution by criminalizing the clients of sex workers (Levy and Jakobsson, 2014: 7). While the law aimed to abolish prostitution by reducing demand, scholars find that it did not accomplish this goal but instead moved sex work from visible street locations to indoor locations or to "neighboring countries" (Chu and Glass, 2013: 105, Levy and Jakobbson, 2014). Though some scholars find that sex workers' safety has been compromised following the law's passage (Levy and Jakobbson, 2014), the 'Swedish Model,' as it is known, has become popular among some policy-makers, activists, and academics, evidenced by its adoption or desired adoption across countries in Europe, under review in Canada, and certain states in the United States.

In the United States, both the sale and purchase of sex is criminalized – save rural brothels in the state of Nevada (Brents et al., 2010). Efforts to 'clean up' and manage street-based prostitution through law enforcement antiprostitution raids have been extended indoors and online (Bernstein, 2007; Musto and boyd, 2014) and expanded as a response to concerns about sex trafficking. U.S. sex worker rights organizers have challenged criminalization and agitated for legal and cultural change (Chateauvert, 2014; Jenness, 1990, 1993; Majic, 2013; Showden and Majic, 2014; Weitzer, 1991). In New Zealand, prostitution was decriminalized in 2003. The New Zealand Prostitutes' Collective worked with a range of state and key actors to facilitate the change in law. Though the impact of decriminalization is still being studied, some research indicates that it has led to better health and other outcomes for many workers, though some – youth, for example – continue to struggle to access services and work safely (Abel et al., 2010).

Irrespective of whether countries criminalize, decriminalize, or legalize commercial sex, it is an industry that is highly, albeit unevenly, regulated. Moreover, though countries may adopt what appear to be different approaches to its regulation, the effects of laws and the kinds of spatial control sex workers experience bear some similarity. Certain broad observations can be made. First, policies aimed at reducing the demand of prostitution, or removing street-based prostitution from public view, may not reduce it per se but shift it to off-street and online locations. Second, cultural, religious, and societal expectations that sex workers are 'fallen women' to be pitied, or exploited victims in need of saving (Doezema, 2001), have exacerbated the stigma associated with sex work even in countries where prostitution is legalized or decriminalized. Third, the magnitude with which prostitution is regulated and policed highlights how the criminal justice system is used as a form of social control over poor women and other marginalized and vulnerable populations such as undocumented immigrants (Chapkis, 2003; Ditmore, 2005), youth of color, and gender nonconforming individuals. Fourth, a disconnect commonly exists between how laws are written and how they are implemented on the ground, which can promote 'unintended consequences' (Wagenaar et al., 2013: 88) that detrimentally impact sex workers. Finally, widespread moral panics surrounding "sex trafficking" (Weitzer, 2007; Mahdavi, 2014) in diverse national contexts have shored up antiprostitution sentiments where prostitutes are seen as victims in need of protection – yet many sex workers, in reality, are treated as criminal offenders subject to carceral intervention (Ditmore, 2009; Bernstein, 2010; Musto, 2013). Here it is not just the state that has sought to intervene on the behalf of sex workers. Transnational faith-based efforts to abolish prostitution have aligned different groups like conservative evangelical Christians with radical, antiprostitution-minded feminists (Bernstein, 2007). These interventions have subjected sex workers to heightened religious and

humanitarian oversight (Bernstein, 2010) sometimes by outside transnational actors including global governance institutions like the United Nations, International Labor Organization, and multinational lending institutions like the International Monetary Fund when punishments for noncompliance to antitrafficking efforts are accompanied by international sanctions (Shih, 2014b; Mahdavi, 2014). They have also brought nonstate sponsored activist campaigns, shelters, and alternative employment schemes to countries with disparate histories for managing commercial sex (Shih, 2013).

Sex Work as Violence, Exploitation, Labor, and Choice

Disagreement exists as to whether sex workers wield control over the conditions and environments in which they operate. For example, one can find in the literature hotly contested debates and blurred lines between voluntary consensual sex work and forced prostitution (e.g., sex trafficking) (see Jeffreys, 2009; Raymond, 2004 for concerns about the sale of sex as violence against women; for critiques see O'Connell Davidson, 2006; Weitzer, 2007). Scholars routinely observe that sex work and sex trafficking are empirically distinct phenomena. Yet since the 1990s, sustained attention and persistent contestation about prostitution have resuscitated longstanding feminist concerns about power, sex, gender, oppression, violence, and the state and amplified international and domestic policy attention to sex trafficking.

The terms of these debates and the discursive tensions they have produced are widely documented (Chuang, 2010; Limoncelli, 2009; Anderson and Andrijasevic, 2008; Agustín, 2007; Saunders, 2005; Parreñas, 2011). Some activists have expressed concern that sex workers and trafficking victims have been inadequately represented within feminist discussions and that sex workers' voices have been discounted. Even in light of these valuable critiques, a brief discussion of feminist theoretical treatment of prostitution is warranted, if only to signal how competing feminist orientations to the topic continue to frame mainstream discussions and focus policy considerations.

For over a half a century, feminists across the globe have vigorously debated the political salience of prostitution, particularly as it relates to the experiences of women. As Anderson and Andrijasevic (2008) note, 'abolitionist feminists' view prostitution as akin to 'sexual slavery,' an extension of 'patriarchal domination over female sexuality,' and understand it as a 'gender crime' (Anderson and Andrijasevic, 2008: 139). Comparing prostitution to a gender crime has gained traction in many countries following the UN's adoption of the Palermo Protocol in 2000, an instrument that frames antitrafficking efforts through the lens of controlling crime (Anderson and Andrijasevic, 2008: 136). Abolitionist ideologies lend support to criminal justice antiprostitution efforts, what Bernstein critiques as a form of 'carceral feminism' (Bernstein, 2010). This is a feminist inspired antitrafficking approach that relies on the state to protect prospective sex trafficked victims from (perceived) threats and violence, resulting in a 'rescue industry' (Agustín, 2007) of government agencies, nonprofits, and NGOs whose funding and existence relies on identifying, rescuing, and rehabilitating sex trafficking victims/survivors. Bernstein's (2010) observations about the fusion of abolitionist sentiments with criminal justice efforts demonstrate how abolitionist ideologies have been bolstered by heightened attention to sex trafficking globally.

Scholars have documented the problems with the abolitionist approach. In a U.S. context, antiporn crusaders of the 1980s arguably found new cause in antitrafficking efforts that began in the late 1990s. In so doing, they abandoned broader sociolegal concerns about migration, global economic inequality, and labor rights across a range of informal labors (e.g., domestic work, construction) in favor of a narrow focus on sex trafficking of women and girls (Agustín, 2005; Ditmore, 2005; Doezema, 2002, 2010; Kempadoo et al., 2005; Sanghera, 2005; Schaeffer-Grabiel, 2010). The violence against women framework thus neglects to grapple with histories of racism, colonialism, imperialism, and the global economic conditions that shape twenty-first century prostitution (Kempadoo, 2001: 34; 37–38). The abolitionist feminist perspective also blunts the development of effective strategies to secure sex workers' health, occupational safety, and human rights (Kempadoo, 2001).

Others have instead adopted a 'sex workers' rights' based approach to the topic (Anderson and Andrijasevic, 2008: 139; Delacoste and Alexander, 1998; Ditmore et al., 2010). Arguing against the premise that all prostitution is coerced and highlighting the problematic conflation of prostitution with violence against women (Saunders, 2005), prorights feminists suggest that sex work is a form of labor (Chateuavert, 2014: 193), and that people can and do voluntarily choose to engage in prostitution, not only amid constrained economic circumstances but as a well-informed 'advancement strategy' (Brennan, 2002; as cited in Saunders, 2005: 353). Prorights and transnational feminists draw attention to structural factors (Hoang, 2010) that render women in general and women of color and *transwomen* in particular vulnerable to economic precarity.

Disagreement as to whether voluntary prostitution is possible (Dworkin, 2004; Farley and Barkan, 1998; Jeffreys, 2009; Raymond, 2004) or whether the regulation of sex work contributes to sex trafficking are not merely philosophical nor are these debates exclusive to feminists. Rather, these debates contribute to questions about policy and raise a number of critical questions: Should countries decriminalize or legalize the sale of sex to best protect workers and prevent trafficking? Or should prostitution be further prosecuted and penalized? What kinds of laws, policies, and activities support sex workers and prevent against coercion, exploitation, and workplace abuses within the sex trade? What experiences are privileged by the violence against women framework and what experiences and circumstances does it leave unacknowledged? Levy and Jakobsson (2014) find that laws that criminalize the purchase of sex in Sweden advance a narrow rendering of prostitution as violence against women (2014: 3). This framing ignores transgender and men's experiences, in addition to ignoring the voices of women within the sex trade and arguably reflects more broadly held antiprostitution sentiments that do not complicate the gender essentialist, heteronormative, and transphobic assumptions on which they are built. Debates surrounding prostitution ultimately hinge on how different interest groups, whether feminists, sex workers, public health

officials, legal experts, economists, and policymakers differentially define harm, violence, and protection and the kinds of laws, policies, and activities deemed best equipped to support sex workers and prevent against coercion, exploitation, and workplace abuses within the sex trade.

Sex Work and the Nation State

A discussion of sex work also invites consideration of the structural conditions and geopolitical context in which different kinds of sexual labor manifest. This kind of analysis proves essential to understanding the disparate and uneven experiences of individuals who engage in sexual commerce in a late modern, globalized landscape. Global economic restructuring, income inequality, international sex tourism (Hall and Ryan, 2001; Ryan, 2000; Wonders and Michalowski, 2001), and the unprecedented scale of internal and international migration (Agustín, 2003) situate prostitution at the nexus of global capital, mobility, and the nation state.

Prostitution is sometimes framed within the context of state-sponsored decision-making and corporate expansion, such as the state-sponsored brothels in the capital city of Chiapas, Mexico (Kelly, 2008). Others trace the emergence of contemporary forms of prostitution in South World cities to free trade zones (Kempadoo, 2001), areas occupied by military and peacekeeping forces (Cheng, 2011), "global cities" (Sassen, 2004, Chin, 2013: 9) and sex tourism as a strategy of economic development (Padilla, 2007; O'Connell Davidson and Sánchez Taylor, 2005; Da Silva and Blanchette, 2009). Sex work is not wholly dissimilar from other low wage and informal work opportunities migrants may seek. However, the constraints and stigma migrant sex workers experience are distinct and typically more pronounced both because of their citizenship status and as a result of law enforcement oversight (Chapkis, 2003; Agustín, 2005; Luibhéid, 2002).

Laws that make prostitution illegal often contradict economic development policies that indiscriminately seek profit off the bodies of working class people. For instance, while prostitution is illegal in Thailand, the government has supported unbridled tourist-industry expansion as a primary engine of economic development. Following the arrival of tens of thousands of American military service to Thailand for Rest and Recreation (R and R) during the Indochina Wars, Thailand's sex tourist economy became the largest in the world. In the present day, sex tourists from across the globe – including a large number of clients from within Thailand and across Asia – arrive in Thai cities seeking different forms of intimate exchange, including diverse companionship arrangements (Skrobanek et al., 1997; Manderson and Jolly, 1997; Jeffery, 2002; Wilson, 2004; Empower, 2011). For instance, in research on professional girlfriends in Cambodia, Hoefinger describes these unique relationships as 'transactional partnerships; (Hoefinger, 2013). The diverse clientele of sex workers has also been specifically addressed by Hoang (2014), whose research in Vietnam suggests that the aesthetic desires of different male clients illuminate new sex worker strategies for embodiment vis-à-vis race, nation, and modernity. The specificities of sexual commerce and the constellation of relationships that emerge from such arrangements suggest that the moral panics around sex tourism are yet another instance of over-inflated vigilance and stigmatization of sex work.

Everywhere from South Korea, China, Mexico, Brazil, France, Italy, Japan, to the U.S., among other destinations of migration, global panics and antisex trafficking discourses and politics have fostered a migration and policy environment in which the depiction of sex work as violence overshadows highly exploitative manual labor conditions in the service and manufacturing sectors. For instance, Shih (2013; 2014a) work on vocational training programs for former sex workers in Beijing and Bangkok argues that the typical gendered rehabilitative labors of sewing, jewelry making, and garment work offer nonautonomous and exploitative working conditions of manual labor (Shih, 2014b). Plambech's (2014) critique of the larger cyclical repatriation/deportation industry for Nigerian sex workers to Italy reveals the undeniably punitive effects of global sex worker rescue efforts (Plambech, 2014). Exclusive attention to rescuing and rehabilitating sex workers overshadows attention toward improving safety and accountability within sex work. Rather than rescue, sex worker rights organizations have asked for increased attention to violence, protection against police violence, greater employer accountability, secure payment mechanisms, and the right to unionize.

In May 2014, months before the launch of the World Cup in several cities throughout Brazil, local government and police participated in a sanitization campaign that illegally arrested over 100 Rio-based sex workers and subjected them to numerous rights violations. Earlier that year, Putin's "security ring of steel" around the Olympic town of Sochi conducted similar operations, reminding observers of similar policies to ban prostitutes from Moscow during the 1980 Olympics (Vostocable, 2014). Similar policies existed just prior to the 2008 Beijing Olympics: mass-scale 'deportations' of rural migrant sex workers to their hometowns (Nyiri, 2010). The links between large-scale internationally televised sporting events and prostitution (Hennig et al., 2007, Richter et al., 2010) have also been highlighted by a recent spate of reporting on the link between sex trafficking and the U.S. Super Bowl (Peterson, 2011). Yet studies find that sporting events like the World Cup do not lead to increased sex trafficking but rather increased police attention (Gould, 2010; Richter et al., 2010; Tavella, 2007)

Sex Workers as Knowledge and Cultural Producers

Sex workers have led local and transnational efforts to advocate for sex workers' health, occupational safety, and human rights. Across the globe, sex workers demand the right to work safely and free from violence, and they agitate for basic human rights to work, family, and citizenship. Some engage in research, such as the online *Research for Sex Work* journal of the Global Network of Sex Work Projects that has produced scholarship since 1998. Works like *Sex Work: Writing by Women in the Sex Industry* (Delacoste and Alexander, 1998), *Playing the Whore: The Work of Sex Work* (Gira Grant, 2014), and *Hos, Hookers, Call Girls, and Rent Boys: Professionals Writing on Life, Love, Money, and Sex* (Sterry and Martin, 2009) are U.S.-based

cultural interventions into the narratives told about sex workers by academics and political stakeholders.

In Thailand, the EMPOWER Foundation, founded in 1984, is the nation's earliest sex worker rights organization, staffed by and in service of sex workers. It provides a variety of outreach, research, and advocacy programs including language training in English and Thai, rights awareness, and health education, and has implemented the nation's first sex worker owned and operated bar, *The Can-Do Bar*, which operates in compliance with local labor laws (Empower, 2012; Kinney, 2014). In 2012, Empower joined several sex worker rights organizations throughout the Southeast Asian region to found 'SWASEAN' (Sex Workers of ASEAN), building on the Association of Southeast Asian Nations (ASEAN), the primary political and economic agreement in the region. SWASEAN has organized the exhibition of sex worker-produced works of art, ranging from films, music, visual and performance art, in different venues throughout the region. Alongside their cultural work, Empower frequently publishes their own books and manuals, and in 2012 published a landmark research report, 'Hit and Run,' that provided extensive documentation of how the antihuman trafficking movement has inflicted harm on sex workers (Empower, 2012).

Several well-known sex worker rights organizations have sought to make a large-scale impact, often integrating local outreach and direct services to sex workers with national global research, advocacy and policy making agendas. In India, the Durbar Mahila Samanwaya Committee (DMSC), founded in 1992, is a collective of over 65 000 sex workers who organize and work in the Sonagachi district in Kolkata. DMSC leads 'community-based structural interventions (CBSI),' pioneering one of the largest community development models of interventions to HIV/AIDS, forging a "multisectoral response to structural drivers of HIV/AIDS" (Swendeman, 2013). The New Zealand Prostitutes Collective developed a draft that became the Prostitution Reform Bill of 2003, effectively decriminalizing sexual labor in that country. Arguably these achievements are due to the collaboration between sex workers and allies from public health sectors and policymakers, as well as other interested agencies and organizations (Abel et al., 2010).

While sex workers have incorporated new and innovative strategies for cultural activism, these perspectives are still often marginalized by mainstream, sensationalized portrayals of sex as an 'iconic, unequivocal, and perpetual human rights violation' (Shah, 2013). These accounts – a genre that Carole Vance calls a 'melomentary' (Vance, 2012) – includes television, film, journalistic and literature accounts and represent ripe artifacts of the socially constructed – and often contested – terrains of sex work.

Concluding Thoughts

In this late modern, postindustrial landscape in which sexual commerce is organized, the boundaries between erotic, emotional, and economic exchanges are invariably blurred (Bernstein, 2007; Sanders, 2005; Zelizer, 2005) and individuals' motivations and experiences with it wide-ranging. Scholarly debates surrounding prostitution continue to be fiercely waged, particularly as concerns about sex trafficking have fostered antiprostitution responses. This article offers snapshots of some of the broad terms, discursive tensions, and overarching questions that foreground contemporary scholarly treatment of prostitution and sex work. As prostitution and sex work continue to evolve, future research by sex workers and scholars is needed to map the salience of these changes and gauge its impact on the people most directly impacted.

See also: Crime Policy in the United States; Employment and Labor, Regulation of; European Union Law; Gender and the Law; Immigration, Emigration, and Citizenship; International Law and Treaties; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; South Asian Legal Traditions.

Bibliography

Abel, Gillian, Fitzgerald, Lisa, Healy, Catherine, Aline, Taylor (Eds.), 2010. Taking the Crime Out of Sex Work: New Zealand Sex Workers' Fight for Decriminalisation. Policy Press, Bristol, UK.

Anderson, Bridget, Andrijasevic, Rutvica, 2008. Sex, slaves and citizens: the politics of antitrafficking. Soundings 40, 135–145.

Auger, Cheryl, 2014. Criminalized and licensed: local politics, the regulation of sex work, and the construction of 'Ugly bodies.' In: Showden, C., Majic, S. (Eds.), Negotiating Sex Work: Unintended Consequences of Policy and Activism. University of Minnesota Press, Minneapolis.

Agustín, Laura María, 2003. A migrant world of services. Social Politics 10 (3), 377–396.

Agustín, Laura María, 2005. New research directions: the cultural study of commercial sex. Sexualities 8 (5), 618–631.

Agustín, Laura María, 2007. Sex at the Margins: Migration, Labour Markets and the Rescue Industry. Zed Books, New York.

Bernstein, Elizabeth, 2007. Temporarily Yours: Intimacy, Authenticity, and the Commerce of Sex. University of Chicago Press, Chicago.

Bernstein, Elizabeth, 2010. Militarized humanitarianism meets carceral feminism: the politics of sex, rights, and freedom in contemporary antitrafficking campaigns. Signs 36 (1), 45–71.

Bernstein, Elizabeth, 2012. Carceral politics as gender justice? The "traffic in women" and neoliberal circuits of crime, sex, and rights. Theory and Society 41 (3), 233–259.

Bernstein, Elizabeth, Schaffner, Laurie (Eds.), 2005. Regulating Sex: The Politics of Intimacy and Identity. Routledge, New York.

Brennan, Denise, 2002. Selling sex for visas: sex tourism as a stepping-stone to International Migration. In: Ehrenreich, Barbara, Hochschild, Arlie Russell (Eds.), Global Woman Nannies, Maids and Sex Workers in the New Economy. Granta, London, pp. 154–168.

Brents, Barbara G., Kathryn, Hausbeck, 2010. Sex work now: what the blurring boundaries around the sex industry mean for sex work, research and activism. In: Hope Ditmore, Melissa, Levy, Antonia, Willman, Alys (Eds.), Sex Work Matters: Power and Intimacy in the Sex Industry. Zed Books, London.

Brents, Barbara, Jackson, Crystal A., 2013. Gender, emotional labour and interactive body work: negotiating flesh and fantasy in sex workers' labor practices. In: Wolkowitz, C., Cohen, R., Sanders, T., Hardy, K. (Eds.), Body/Sex/Work: Intimate, Embodied and SexualisedLabour. Palgrave MacMillan, Basingstoke, UK, pp. 77–93.

Brents, Barbara G., Jackson, Crystal A., Hausbeck, Kathryn, 2010. The State of Sex: Tourism, Sex and Sin in the New American Heartland. Routledge, New York.

Brewis, Joanna, Linstead, Steven, 2000. 'The worst thing is the screwing': context and career in sex work. Gender, Work and Organization 7, 168–180.

Chapkis, Wendy, 1997. Live Sex Acts: Women Performing Erotic Labor. Routledge, New York.

Chapkis, Wendy, 2003. Trafficking, migration, and the law: protecting innocents, punishing immigrants. Gender & Society 17 (6), 923–937.

Chateauvert, Melinda, 2014. Sex Workers Unite: A History of the Movement from Stonewall to SlutWalk. Beacon Press, Boston.

Cheng, Sealing, 2011. On the Move for Love: Migrant Entertainers and the US Military in South Korea. University of Pennsylvania Press, Philadelphia.

Chin, Christine B.N., 2013. Cosmopolitan Sex Workers: Women and Migration in a Global City. Oxford University Press, Oxford.

Chu, Sandra Ka Hon, Glass, Rebecca, 2013. Sex work law reform in Canada: considering problems with the nordic model. Alberta Law Review 51, 101–201.

Chuang, Janie, 2010. Rescuing trafficking from ideological capture: prostitution reform and antitrafficking law and policy. University of Pennsylvania Law Review 158, 1655–1728.

Constable, Nicole, 2009. The commodification of intimacy: marriage, sex, and reproductive labor. Annual Review of Anthropology 38, 49–64.

Da Silva, Ana Paula, Blanchette, Thaddeus Gregory, 2009. Sexual tourism and social panics: research and intervention in Rio de Janeiro. Souls 11 (2), 203–212.

Delacoste, Frederique, Alexander, Priscilla (Eds.), 1998. Sex Work: Writings by Women in the Sex Industry, second ed. Cleis Press, San Francisco.

Dewey, Susan, Kelly, Patty (Eds.), 2011. Policing Pleasure: Sex Work, Policy, and the State in Global Perspective. New York University Press, New York and London.

Ditmore, Melissa Hope, 2005. Trafficking in lives: how ideology shapes policy. In: Kempadoo, K., Sanghera, J., Pattanaik, B. (Eds.), Trafficking & Prostitution Reconsidered: New Perspectives on Migration, Sex Work, and Human Rights. Paradigm, Boulder, CO, pp. 107–126.

Ditmore, Melissa Hope, 2006. In: Encyclopedia of Prostitution and Sex Work, vol. 1. Greenwood, Westport, CT.

Ditmore, Melissa Hope, 2009. The Use of Raids to Fight Trafficking in Persons. Urban Justice Center, New York.

Ditmore, Melissa, Levy, Antonia, Willman, Alys, 2010. Sex Work Matters. Zed Press, London.

Doezema, Jo, 2001. Ouch! Western Feminists' 'Wounded Attachment' to the 'Third World Prostitute.' Feminist Review 67, 16–38.

Doezema, Jo, 2002. Who gets to choose? Coercion, consent, and the UN trafficking protocol. Gender and Development 10 (1), 20–27.

Doezema, Jo, 2010. Sex Slaves and Discourse Masters: The Construction of Trafficking. ZedBooks, London.

Dworkin, Andrea, 2004. Pornography, prostitution, and a beautiful and tragic recent history. In: Stark, Christine, Whisnant, Rebecca (Eds.), Not for Sale: Feminists Resisting Prostitution and Pornography. Spinifex Press, North Melbourne, Australia, pp. 137–148.

Empower Foundation, 2011. Bad Girls of Lanna: Our Story of Sex Work in Chiang Mai. Empower University Press, Thailand, Bangkok.

Empower Foundation, 2012. Hit and Run: The Impact of Antihuman Trafficking Policy and Practice on Sex Workers' Human Rights in Thailand. Empower University Press, Bangkok.

Empower Foundation, 2013. The Last Rescue of Siam. http://www.empowerfoundation.org/sexy_en.php.

Farley, Melissa, Barkan, Howard, 1998. Prostitution, violence, and posttraumatic stress disorder. Women & Health 27 (3), 37–49.

Feldman, Valerie, 2014. Sex work, politics and the internet: carving out political space in the blogosphere. In: Showden, C., Majic, S. (Eds.), Negotiating Sex Work: Unintended Consequences of Policy and Activism. University of Minnesota Press, Minneapolis.

Gira Grant, Melissa, 2014. Playing the Whore: The Work of Sex Work. Verso Press, Brooklyn.

Goodyear, Michael, Weitzer, Ronald, 2011. International trends in the control of sexual services. In: Dewey, S., Kelly, P. (Eds.), Policing Pleasure: Sex Work, Policy, and the State in Global Perspective. New York University Press, New York and London, pp. 16–30.

Gould, Chandre, 2010. Moral panic, human trafficking and the 2010 Soccer World Cup. Agenda: Empowering Women for Gender Equity 24 (85), 31–44.

Hall, M.C., Ryan, C., 2001. Sex Tourism: Marginal People and Liminalities. Routledge, New York.

Hardy, Kate, Kingston, Sarah, Sanders, Teela (Eds.), 2010. New Sociologies of Sex Work. Ashgate, Surrey, UK.

Hennig, Jana, Craggs, Sarah, Larsson, Fred, Laczko, Frank, 2007. Trafficking in human beings and the 2006 world cup in Germany. MRS No. 29 International Organization for Migration.

Hoang, Kimberly Kay, 2010. Economies of emotion, familiarity, fantasy, and desire: emotional labor in Ho Chi Minh City's sex Industry. Sexualities 13 (2), 255–272.

Hoang, Kimberly, 2014. Competing technologies of embodiment Pan-Asian Modernity and Third World dependency in Vietnam's contemporary sex industry. Gender & Society 28 (4), 513–536. http://dx.doi.org/10.1177/0891243214523122.

Hoefinger, Heidi, 2013. Sex, Love and Money in Cambodia: Professional Girlfriends and Transactional Relationships. Routledge, New York.

Hubbard, Phil, 1999. Sex and the City: Geographies of Prostitution in the Urban West. Ashgate, London.

Hubbard, Phil, 2001. Sex zones: intimacy, citizenship and public space. Sexualities 4 (1), 51–71.

Hubbard, Phil, 2004. Cleansing the metropolis: sex work and the politics of zero tolerance. Urban Studies 41 (9), 1687–1702.

Hubbard, Phil, 2013. Cities and Sexualities. Routledge, New York.

Jeffery, Leslie Ann, 2002. Sex and Borders: Gender, National Identity and Prostitution Policy in Thailand. University of Hawai'i Press, Honolulu.

Jeffreys, Elaine, 2013. China, Sex and Prostitution. Routledge, New York.

Jeffreys, Sheila, 2009. The Industrial Vagina: The Political Economy of the Global Sex Trade. Taylor & Francis, London.

Jenness, Valerie, 1990. From sex as sin to sex as work: COYOTE and the reorganization of prostitution as a social problem. Social Problems 37 (3), 403–420.

Jenness, Valerie, 1993. Making it Work: The Prostitutes' Rights Movement in Perspective. De Gruyter, New York.

Katsulis, Yasmina, 2010. Sex Work and the City: The Social Geography of Health and Safety in Tijuana, Mexico. University of Texas Press, Austin, TX.

Kempadoo, Kamala, 2001. Women of color and the global sex trade: transnational feminist perspectives. Meridians 1 (2), 28–51.

Kempadoo, Kamala, Sanghera, Jyoti, Pattanaik, Bandana (Eds.), 2005. Trafficking & Prostitution Reconsidered: New Perspectives on Migration, Sex Work, and Human Rights. Paradigm, Boulder, CO.

Kempadoo, Kamala, Doezema, Jo (Eds.), 1998. Global Sex Workers: Rights, Resistance, and Redefinition. Routledge, New York.

Kelly, Patty, 2008. Lydia's Open Door: Inside Mexico's Most Modern Brothel. University of California Press, Berkeley.

Kingston, Sarah, 2014. Prostitution in the Community: Attitudes, Action, and Resistance. Routledge, New York.

Kinney, Edith, 2014. Raids, rescues, and resistance: women's rights and Thailand's response to human trafficking. In: Showden, C., Majic, S. (Eds.), Negotiating Sex Work: Unintended Consequences of Policy and Activism. University of Minnesota Press, Minneapolis.

Levy, Jay, Jakobsson, Pye, 31 March 2014. Sweden's abolitionist discourse and law: effects on the dynamics of Swedish sex work and on the lives of Sweden's sex workers. Criminology and Criminal Justice, 1748895814528926. http://crj.sagepub.com/content/early/2014/03/31/1748895814528926.full.pdf+html.

Limoncelli, Stephanie, 2009. The trouble with trafficking: conceptualizing women's sexual labor and economic human rights. Women's Studies International Forum 32 (4), 261–269.

Luibhéid, Eithne, 2002. Entry Denied: Controlling Sexuality at the Border. University of Minnesota Press, Minneapolis.

Mahdavi, Pardis, 2014. From Trafficking to Terror: Constructing a Global Social Problem. Routledge, New York.

Majic, Samantha, 2013. Sex Work Politics: From Protest to Service Provision. University of Pennsylvania Press, Philadelphia, PA.

Majic, Samantha, 2014. Participation despite the odds: examining sex workers' political engagement. New Political Science 36 (1), 76–95.

Manderson, Lenore, Jolly, Margaret (Eds.), 1997. Sites of Desire/Economies of Pleasure: Sexualities in Asia and the Pacific. University of Chicago Press, Chicago.

Musto, Jennifer, 2013. Domestic minor sex trafficking and the detention-to-protection pipeline. Dialectical Anthropology 37 (2), 257–276.

Musto, Lynne, Jennifer, boyd, danah, 2014. The trafficking technology nexus. Social politics: international studies in gender, state, and society. Social Politics 21 (3), 461–483.

Nyíri, Pál, 2010. Mobility and Cultural Authority in Contemporary China. University of Washington Press, Seattle.

O'Connell Davidson, Julia, 2006. Will the real sex slave please stand up? Feminist Review 83, 4–22.

O'Connell Davidson, Julia, Taylor, Jacqueline Sánchez, 2005. Travel and taboo: heterosexual sex tourism in the Caribbean. In: Bernstein, Elizabeth, Schaffner, Laurie (Eds.), Regulating Sex: The Politics of Intimacy and Identity. Routledge, New York, pp. 83–99.

Outshoorn, Joyce, 2014. The contested citizenship of sex workers: The case of the Netherlands. In: Showden, C., Majic, S. (Eds.), Negotiating Sex Work: Unintended Consequences of Policy and Activism. University of Minnesota Press, Minneapolis, pp. 171–194. http://www.upress.umn.edu/book-division/books/negotiating-sex-work.

Padilla, Mark, 2007. Caribbean Pleasure Industry: Tourism, Sexuality, and AIDS in the Dominican Republic. University of Chicago Press, Chicago.

Parreñas, Rhacel, 2011. Illicit Flirtations: Labor, Migration, and Sex Trafficking in Tokyo. StanfordUniversity Press, Stanford, CA.

Plambech, Sine, 2014. Between 'Victims' and 'Criminals' : rescue, deportation, and everyday violence among Nigerian migrants. Social Politics 21 (3), 382–402.
Peppet, Scott R., 2013. Prostitution 3.0? Iowa Law Review 98, 1989–2060.
Peterson, Matt, 2011. Efforts of sex trafficking task force led to 133 Arrests in Run-up to super bowl XLV," Dallas Morning news (accessed 05.06.14). http://crimeblog.dallasnews.com/archives/2011/02/efforts-of-sex-trafficking-tas.html.
Raymond, Janice G., 2004. Prostitution on demand: legalizing the buyers as sexual consumers. Violence Against Women 10 (10), 1156–1186.
Richtel, Matt, 21 September 2013. Intimacy on the Web, with a Crowd. New York Times. http://www.nytimes.com/2013/09/22/technology/intimacy-on-the-web-with-a-crowd.html?pagewanted%25C2%25BCall&_r%25C2%25BC0.
Richter, Marlise L., Chersich, Matthew F., Scorgie, Fiona, Luchters, Stanley, Temmerman, Marleen, Steen, Richard, February 2010. Sex work and the 2010 FIFA world cup: time for public health imperatives to prevail. Globalization and Health 6 (1).
Ryan, C., 2000. Sex tourism: paradigms of confusion. In: Clift, S., Carter, S. (Eds.), Tourism and Sex: Culture, Commerce and Coercion. Cassell, London, pp. 23–40.
Sanders, Teela, 2005. Sex Work: A Risky Business. Willan, Cullompton, UK.
Sanders, Teela, O'Neill, Maggie, Pitcher, Jane, 2009. Prostitution: Sex Work, Policy and Politics. Sage, London.
Sanghera, Jtoyi, 2005. Unpacking the trafficking discourse. In: Kempadoo, K., Sanghera, J., Pattanaik, B. (Eds.), Trafficking & Prostitution Reconsidered: New Perspectives on Migration, Sex Work, and Human Rights. Paradigm, Boulder, CO, pp. 3–24.
Sassen, Saskia, 2004. Global cities and survival circuits. In: Ehrenreich, Barabara, Hochschild, Russell (Eds.), Global Woman: Nannies, Maids, and Sex Workers in the New Economy. Holt Paperbacks, New York, pp. 254–274.
Saunders, Penelope, 2005. Traffic violations determining the meaning of violence in sexual trafficking versus sex work. Journal of Interpersonal Violence 20 (3), 343–360.
Schaeffer-Grabiel, Felicity, 2010. Sex trafficking as the 'new slave trade'? Sexualities 13 (2), 153–161.
Shah, Svati, 2013. Brothels and big screen rescues: producing the idea of 'prostitution in India' through documentary film. Interventions 15 (4), 549–566.
Shih, Elena, 2013. Globalising rehabilitation regimes: the moral economy of vocational training in after-trafficking work. In: Yea, Sallie (Ed.), Forcing Issues: Rethinking and Rescaling Human Trafficking in the Asia-Pacific Region. Routledge, New York, pp. 64–78.
Shih, Elena, 2014a. The antitrafficking rehabilitation complex. Contexts 13 (1), 16–25.
Shih, Elena, 2014b. Freedom Markets: Moral and Political Economies of The Human Trafficking Movement in China, Thailand and The United States (Doctoral dissertation). University of California, Los Angeles.
Showden, Carisa, Majic, Samantha (Eds.), 2014. Negotiating Sex Work: Unintended Consequences of Policy and Activism. University of Minnesota Press, Minneapolis.
Skrobanek, Siriporn, Boonpakdi, Nattaya, Janthakeero, Chutima, 1997. The Traffic in Women: Human Realities of the International Sex Trade. Zed Books, London.
Sterry, David Henry, Martin Jr., R.J., 2009. Hos, Hookers, Call Girls, and Rent Boys: Professionals Writing on Life, Love, Money, and Sex. Soft Skull Press, Berkeley, CA.
Swendeman, Dallas, 2013. Community-led Structural Interventions as a Model of Empowerment For Health and Development: The Sonagachi/Durbar Intervention with Sex Workers in India, 141st APHA Annual Meeting (2–6 November 2013).
Tavella, Anne Marie, 2007. Sex trafficking and the 2006 world cup in Germany: concerns, actions and implications for future international sporting events. Northwestern University Journal of International Human Rights 6 (1).
Vance, Carole S., 2012. Innocence and experience: melodramatic narratives of sex trafficking and their consequences for law and policy. History of the Present 2 (2), 200–218.
Van der Meulen, Emily, Durisin, Elya M., Love, Victoria, 2013. Selling Sex: Experience, Advocacy, and Research on Sex Work in Canada. UBC Press, Vancouver.
Vanwesenbeeck, Ine, 2001. Another decade of social scientific work on sex work: a review of research 1990–2000. Annual Review of Sex Research 12, 242–289.
Vostocable, 2014. Send Them Away—the Olympic Banishment of Prostitutes. Vostocable (accessed on 05.06.14.). http://vostokcable.wordpress.com/2014/04/15/send-them-away-the-olympic-banishment-of-prostitutes/.
Wagenaar, Hendrik, Altink, Sietske, Amesberger, Helga, 2013. Final report of the International Comparative Study of Prostitution Policy: Austria and the Netherlands. Den Haag. Available at: http://lastradainternational.org/lsidocs/Prostitution_policy_report.pdf.
Wahab, Stephanie, 2005. Guest editor's introduction. Journal of Interpersonal Violence 20 (3), 263–269.
West, Jackie, 2000. Prostitution: collectives and the politics of regulation. Gender, Work, and Organization 7 (2), 106–118.
Weitzer, Ron, 1991. Prostitutes' rights in the United States: the failure of a movement. The Sociological Quarterly 32, 23–41.
Weitzer, Ron, 2005. New directions in research on prostitution. Crime, Law & Social Change 43, 211–235.
Weitzer, Ron, 2007. The social construction of sex trafficking: Ideology and institutionalization of a moral crusade. Politics & Society 35, 447–475.
Weitzer, Ron, 2012. Legalizing Prostitution: From Illicit Vice to Lawful Business. New York University Press, New York and London.
Wilson, Ara, 2004. The Intimate Economies of Bangkok: Tomboys, Tycoons, and Avon Ladies in the Global City. University of California Press, Berkeley.
Wonders, Nancy, Michalowski, Raymond, 2001. Bodies, borders, and sex tourism in a globalized world: a tale of two Cities—Amsterdam and Havana. Social Problems 48 (4), 544–571.
Zelizer, Viviana, 2005. The Purchase of Intimacy. Princeton University Press, Princeton, NJ.

Prototype Theory in Linguistics

John R Taylor, University of Otago, Dunedin, New Zealand

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition volume 13, pp. 8954–8957, © 2001, Elsevier Ltd.

Abstract

Prototype theory, as developed by Rosch, has had repercussions in two main areas of linguistics: lexical semantics and syntax. Word meanings are the names of categories, and the meanings of many words display characteristic prototype effects (fuzziness of category boundaries, degrees of neo-Kantians of category members). Further areas of application have been semantic change and the structure of polysemy networks. The prototype approach does, however, encounter problems in connection with theories of semantic compositionality. Linguistic constructs, such as syntactic and lexical categories, also display prototype effects. The application of prototype theory to the study of parts of speech and syntactic constructions has been especially fruitful.

Prototype theory, as developed by Rosch, has had repercussions in two main areas of linguistics: lexical semantics and syntax. Word meanings are the names of categories, and the meanings of many words display characteristic prototype effects (fuzziness of category boundaries, degrees of representativity of category members). Further areas of application have been semantic change and the structure of polysemy networks. The prototype approach does, however, encounter problems in connection with theories of semantic compositionality. Linguistic constructs, such as syntactic and lexical categories, also display prototype effects. The application of prototype theory to the study of parts of speech and syntactic constructions has been especially fruitful.

Prototype theory is a theory of categorization. Categorization is a central issue in linguistics, for several reasons. In the first place, words can be regarded as names for categories, and using a word involves an act of categorization (Labov, 1973). In stating that The cup is in the saucer, a speaker is categorizing the figure object as a 'cup,' rather than as a mug, bowl, vase, or other kind of object; the reference object is being categorized as a 'saucer,' not as a dish or a plate; while the location of the cup vis-à-vis the saucer is being categorized as an 'in'-relationship, rather than as an 'on'-relationship, or whatever. Knowing these words involves knowing how to apply the words appropriately to entities that fall within the categories that they designate.

There is another way in which categorization impinges on linguistic theory. The moment one pursues even the most elementary linguistic analysis, language itself becomes an object of categorization. Like any intellectual discipline, linguistics requires its practitioners to set up categories for the description of the object of study. In any linguistic description, one will encounter such entities as words and morphemes, phrases and clauses, nouns and verbs, subjects and objects, clitics and particles, phonemes and allophones. 'Doing linguistics' consists, in large measure, in setting up categories such as these, defining them, studying their properties and interrelationships, working out criteria for determining their membership, and applying them to the data at hand. The ontological status of the categories themselves may also become an issue. Where do categories such as 'noun' and 'verb' come from? Are they inherent in the object of study? Are they mere descriptive conveniences? Are they constructs of linguistic theory?

Two Approaches to Categorization

Two major approaches to categorization are the 'classical' approach and the 'prototype' approach. On the classical approach, a category is defined in terms of a set of necessary and sufficient conditions for category membership. The set of conditions specifies the essence, as it were, of the category. Any entity meeting all the conditions is, by definition, a member of the category; failure to meet any of the conditions excludes an entity from the category. On this view, category boundaries are discrete (an entity either is, or is not, a member), and the category lacks internal structure (all members are of equal status).

The simplicity of the classical view makes it intuitively appealing. A major problem is that many categories are simply not amenable to classical definition. What, for example, are the defining features of 'chair,' 'vegetable,' and 'love,' or, for that matter, of 'cup,' 'saucer,' and 'in'? In fact, classical categorization probably works best only for 'expert' categories, that is, categories that have been created and explicitly defined (in classical terms!) by experts, whether these be scientists, bureaucrats, administrators, or whoever. For most words of everyday speech, the classical theory is quite inappropriate.

On the prototype approach, a category is understood, primarily, in terms of its good examples. Entities are associated with the category to the extent that they resemble the good examples. Often, the boundary of a prototype category will be fuzzy – it might not be clear whether some entities belong in the category or not. Moreover, prototype categories have an internal structure, in that some entities count as more central members than others.

Prototype categorization came to the attention of linguists largely due to the work of the psychologist, Eleanor Rosch. In a series of experiments, summarized in Rosch (1978), it emerged that the meanings of several very ordinary words, such as bird, toy, vehicle, are structured around good examples. People readily agree that a robin is a better example of 'bird' than a penguin; they can determine the truth of A robin is a bird

faster than A penguin is a bird; when asked to name members of a category, they name more central examples before more marginal examples; and when drawing inferences about a category, people tend to draw on knowledge of central members, rather than on knowledge about the category as a whole. Rosch observed prototype effects both with 'natural kind' categories – that is, categories such as 'bird,' which are, as it were, given by nature – and with categories that are understood relative to human concerns, such as 'vehicle,' 'furniture,' and 'toy.'

Rosch's discoveries were eagerly taken up by linguists and extended by a wide range of linguistic data (Lakoff, 1987; Kleiber, 1990; Tsohatzidis, 1990; Geeraerts et al., 1994; Taylor, 1995). An especially fruitful application has been the study of semantic change (Geeraerts, 1997). The prototype notion has also been widely applied to the study of the categories of linguistic description (Langacker, 1987; Croft 1991).

Problems and Refinements

Despite its success in lexical semantics and elsewhere, prototype categorization raises a number of problems.

While not denying the reality of prototype effects, Fodor (1998) argues that concepts cannot be equated with prototypes. The fish prototype (a herring?) and the pet prototype (a poodle?) contribute nothing at all to the complex concept 'pet fish,' or to the prototype of 'pet fish' (a goldfish?). Generally, as a concept gets specified in greater detail, prototype effects not only shift, they also tend to get effaced. Consider the diminishing role of the 'nurse' prototype in one's understanding of nurse, male nurse, retired male nurse, etc. On the other hand, prototype effects show up even on categories that arguably are subject to classical definition, such as 'odd number.' Armstrong et al. (1983) report that subjects readily agreed that 3 was a better example of 'odd number' than 447, even though, obviously, 447 is no less of an odd number than 3.

On one interpretation of these findings, prototype effects are extrinsic to word meanings as such and arise only when a person applies a concept to the world. This view draws a distinction between the 'essence' of a category (the 'real' meaning of a word) and 'recognition procedures' invoked when speakers attempt to categorize things and eventualities in the world. On this view, the fuzziness of category boundaries results from the fact that eventualities in the world may not lend themselves to clear-cut categorization, even though the words of a language do have clear-cut meanings. Although it might not be clear whether a particular utterance counts as a 'lie,' the meaning of lie might nevertheless be clear-cut.

Degree of Membership versus Representativity

Before endorsing this exclusion of prototypes from word meaning proper, it is important to distinguish two aspects of prototype categories: fuzziness of category boundaries (a question of category membership) and degree of representativity within a category (a question of category structure).

It is uncontroversial that certain categories are inherently fuzzy, in that it is simply not possible to make a clean distinction between what goes into the category and what does not. This is the case with many adjectival concepts such as 'old,' 'tall,' and 'bald.' While there is little doubt what a (prototypical) bald head looks like, also what it means to be not bald, it would be futile to draw a boundary between 'bald' and 'not bald,' e.g., in terms of the number of hairs on a person's head.

However, prototype effects also show up on categories that arguably do have clear boundaries. Although turkeys might not be good examples of the bird category, they are birds nonetheless, albeit not very representative ones. Representativity, to be sure, can be understood in different ways – in terms of frequency of occurrence, cultural or environmental salience, or statistical averaging over category instances. Representative examples are typically invoked in default contexts. Given only the mention of 'a bird,' with no further specification, one would assume a representative example. People who wish they were a bird would surely be disappointed if they were transformed into a turkey! The internal structure of a category cannot reasonably be excluded from a person's knowledge of word meanings.

Prototypes and Background Knowledge

Categories, in order to be viable, need to be conceptually coherent (Murphy and Medin, 1985). A category is coherent to the extent that it is based on theories of how the world is, and how it links up with a person's encyclopedic knowledge. A category is not simply a conjunction of an ad hoc set of features, but the category has to make sense.

Lakoff (1987) discusses this aspect in terms of his notion of the idealized cognitive model (ICM). The concept 'bachelor' is understood against an ICM of marriage practices (which in turn is based in ICM's of sexuality, procreation, and family). One aspect of the ICM is the idea that each person, after reaching the 'marriageable age,' sooner or later marries. Within the ICM, bachelor may well be definable in classical-like terms, i.e., as 'unmarried,' 'adult,' and 'male.' Prototype effects arise, because the ICM may not fit the world perfectly. The Pope is not a good example of 'bachelor,' not because of uncertainties about whether the Pope is unmarried, adult, or male, but because Catholic clergy are outside the scope of the ICM.

Prototypes and Features

On the classical theory, categories are assembled, as it were, out of their features; the features preexist the category, are evaluated independently of the category, and must be simpler than the category they define. A prototype, in contrast, can be perceived holistically, as a gestalt; features are no longer the building blocks of categories; they emerge, as dimensions of similarity, through acquaintance with category members. It is unlikely that people construct their bird category through conjunction of features such as 'having feathers' and 'laying eggs'; knowledge of feathers and egg-laying emerges from acquaintance with birds, not vice versa. Likewise with linguistic categories. One might propose that ability to inflect for tense is a defining characteristic of verbs (for English, at least). Yet, the very notion of tense inflection rests on a prior acquaintance with members of the verb category.

Given this revised understanding of features, a feature-based view of prototype categories becomes possible through the

differential weighting of the features. Some features may be better cues to category membership than others; some features may be virtually indispensable; others are able to be overridden. The prototype will be a representation that maximizes feature weights; an entity exhibiting highly weighted features will automatically count as a category member; an entity exhibiting less highly weighted features might still count as a member, though a more marginal one. Moreover, it is possible for two entities to belong to a category, even though they have no category-specific features in common. The way also opens up for a reconciliation of classical and prototype categorization – a classical category is the limiting case of a prototype category, i.e., a category understood solely in terms of essential features.

Prototypes and Polysemy

It is useful to distinguish between the prototype structure of a concept and the polysemy structure of a word. There is a clear difference, in principle, between applying the word fruit to different kinds of fruit and the use of the word in such expressions as the fruit of his labors. The latter usage represents a semantic extension of the word, in a way that is different from the use of the word to refer, say, to olives.

Take the preposition over (Lakoff, 1987). Fly over the Atlantic, the lamp hanging over the table, put your hands over your face, come over here, there is water all over the floor, the party is over, do it over again – these exemplify different senses of the word, rather than more or less representative examples of one and the same sense. To the extent that these senses can be clearly individuated, it is possible to construct a network of sense relations, which links up the different senses according to their similarity. Such an array invites a prototype analysis, with one of the senses selected as central and the others radiating out from it.

There are, however, different criteria for selecting the central sense, and these do not always deliver the same result. On a diachronic perspective, the basic sense is the historically oldest one, from which the others have been derived over time. On a developmental perspective, the basic sense is the one that children acquire first. Alternatively, the descriptive linguist (or lexicographer) might select as basic the most frequent sense, or the one to which the others can be most economically related. The real question, however, concerns the speaker's mental organization of polysemy. Currently, this is a lively topic in polysemy research (Taylor, 1995).

Prototypes and Linguistic Categories

The prototype concept may be applied, not only to the study of word meanings, but to the very categories of linguistic description. Many linguistic concepts (including the very notion of 'language' and 'a language') turn out to have a prototype structure, in that they exhibit degrees of representativity, and often have fuzzy boundaries. Uncertainty as to what goes into a category and what does not pertains even to such basic notions as what constitutes a word (as opposed to a bound morpheme, clitic, or phrase), the part-of-speech categories (whether a particular item is a noun, verb, adjective, etc.), constituency (whether a particular string of words is a syntactic constituent), and syntactic relations (whether a particular constituent is a modifier or adjunct, the subject of a clause, or whatever).

Some linguistic theories propose to decide such issues by fiat. In practice, however, the linguist's strategy often reflects a prototype conception, even though this might not be explicitly acknowledged. When deciding the category status of a linguistic item, it is usual to apply a set of tests (Croft, 1991). The tests are set up on the basis of what, intuitively, count as good examples of the category in question, whereby each of the tests diagnoses a property typical of the good examples. An item that passes all the tests is ipso facto a member of the category; an item that fails the tests is not a member. Sometimes, an item passes only some of the tests; it will have to be regarded as a marginal example, or even as an item of uncertain status.

Consider the parts of speech, perhaps better called 'lexical categories' such as 'noun,' 'verb,' 'adjective,' 'preposition' (*see* Parts of Speech). These can be characterized either in semantic or in formal terms. Semantically, nouns prototypically designate categories of things, i.e., concrete, time-stable, spatially bounded objects (chain, cup, etc.); verbs designate events, involving rapid changes in state (explore, arrive); adjectives designate fairly stable properties of things (hot, young); while prepositions designate a relation, typically a spatial relation, between things (on, at). It is commonly agreed that lexical category cannot be reliably predicted from a word's semantics. In spite of their reference to rapid changes in state, explosion and arrival are nouns, just as good nouns, in fact, as chair and cup. This is because the words exhibit the full range of formal properties typical of the noun category, i.e., they pass all the syntactic tests for nounhood – they pluralize, they take the full range of determiners, they can be modified by an adjective, they can head a subject noun phrase, and so on. There are words that fail to exhibit the full range of properties. Groceries, in I bought some groceries, looks like the plural of a noun grocery – yet there is no such noun (*I bought a grocery), neither can groceries take numeral quantifiers (*five groceries). Syntactically, groceries is a somewhat marginal noun (though still a noun). However, there are words whose status is genuinely unclear. Consider saying, in without my saying a word. Being the complement of a preposition, my saying a word looks like a noun phrase, headed by saying, and saying looks like a noun, in that it takes possessive my. Other determiners, though, are not possible (without *the/*that/*a saying a word). Moreover, saying is followed by what looks like a direct object (a word), and the ability to take a direct object is a distinctly verb-like property, not at all a property of nouns.

Both semantic and formal criteria, then, fail to deliver clear-cut lexical categories. Nevertheless, there is an important sense in which the semantic prototypes have priority. Any language has to provide its speakers with the means to refer to 'things' and 'events,' 'properties' and 'relations,' and the semantic prototypes of the lexical categories, in English and other languages, are understood in such terms. Indeed, without the semantic prototypes, there would be no basis for recognizing the categories 'noun' and 'verb' across the different languages of the world.

See also: Lexical Semantics; Lexicology and Lexicography; Lexicon.

Bibliography

Armstrong, S.L., Gleitman, L.R., Gleitman, H., 1983. What some concepts might not be. Cognition 13, 263–308.

Croft, W., 1991. Syntactic Categories and Grammatical Relations. University of Chicago Press, Chicago.

Fodor, J.A., 1998. Concepts: Where Cognitive Science Went Wrong. Clarendon Press, Oxford, UK.

Geeraerts, D., 1997. Diachronic Prototype Semantics: A Contribution to Historical Lexicology. Clarendon Press, Oxford, UK.

Geeraerts, D., Grondelaers, S., Bakema, P., 1994. The Structure of Lexical Variation: Meaning, Naming, and Context. Mouton de Gruyter, Berlin.

Kleiber, G., 1990. La sémantique du prototype: Catégories et sens lexical, first ed. Presses Universitaires de France, Paris.

Labov, W., 1973. The boundaries of words and their meanings. In: Bailey, C.J., Shuy, R. (Eds.), New Ways of Analysing Variation in English. Georgetown University Press, Washington, DC.

Lakoff, G., 1987. Woman, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago University Press, Chicago.

Langacker, R.W., 1987. Foundations of Cognitive Grammar, vol. 1. Stanford University Press, Stanford.

Murphy, G., Medin, D., 1985. The role of theories in conceptual coherence. Psychological Review 92, 289–316.

Rosch, E., 1978. Principles of categorization. In: Rosch, E., Lloyd, B.B. (Eds.), Cognition and Categorization. Lawrence Erlbaum, Hillsdale, MI.

Taylor, J., 1995. Linguistic Categorization: Prototypes in Linguistic Theory. Clarendon Press, Oxford, UK.

Tsohatzidis, S. (Ed.), 1990. Meanings and Prototypes: Studies in Linguistic Categorization. Routledge, London.

Psychiatric Assessment: Negative Symptoms

Hans-Jurgen Möller, Ludwig Maximilian University, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Clinicians differentiate between positive and negative symptoms. Positive symptoms (e.g., of schizophrenia) are feelings or behaviors seen in the mentally ill that are not usually present in healthy humans, for example delusions and hallucinations. Negative symptoms, likewise, are feelings or behaviors normally present in healthy people that are lacking in the unwell, such as flattening of affect, poverty of speech and social withdrawal. They occur in different mental disorders, especially in schizophrenia. In schizophrenia, they are the phenomenological counterpart of positive symptoms. Both classes of symptoms might be related to different underlying pathophysiologies. The standardized assessment of negative symptoms is not without problems. In the last decade, efforts have been made to improve the assessment of instruments to overcome these diagnostic limitations.

Negative symptoms broadly comprise blunted affect, apathy, attentional impairment, poverty of thought or speech, incoherence, loosening of association, and social withdrawal. The origins of the positive/negative terminology are usually attributed to Jackson, a British neurologist of the last century, who drew the distinction in order to delineate primary from secondary neurological phenomena.

Today, the concept of negative symptoms is especially associated with schizophrenia/schizophrenic disorders; negative symptoms were described as an important part of the clinical symptomatology of schizophrenia already at the beginning of this century by Bleuler and Kraepelin, when the concept of schizophrenic disorders was developed. However, negative symptoms not only are specific for schizophrenia/schizophrenic disorders, but also found, for example, in affective disorders. In the last decades, efforts have been made to improve the standardized assessment of negative symptoms. Especially the introduction of several novel/atypical antipsychotics, which were thought in the last decade of the twentieth century to have better efficacy in negative symptoms compared to classical antipsychotics, increased the interest in clinical research on negative symptoms. The positive/negative distinction may represent different manifestations of a common underlying process. However, it may also suggest the existence of different pathophysiologies in schizophrenia/schizophrenic disorders.

Definition of Negative Symptoms

Negative symptoms in schizophrenia comprise broadly blunted affect, apathy, attentional impairment, poverty of thought or speech, incoherence, loosening of association, thought blocking, and social withdrawal. However, there is no unanimous agreement on the concept of negative symptoms in schizophrenia: some authors adhere to a broad interpretation (Meltzer and Zureick, 1989; Rösler and Hengesch, 1990), while others prefer a narrow approach (Crow, 1985; Strauss et al., 1974; Wing, 1989). The controversy centers on incoherence, loosening of association, and thought-blocking since these were also attributed to positive symptoms. This disparity in the selection of items also manifests itself in the outcome of patient testing with different rating scales. Different authors use different concepts and definitions for the type and number of symptoms implicated.

These problems of definition can also be identified in the rating scales used for measuring negative symptoms.

History of the Term 'Negative Symptoms'

The origins of the positive/negative distinction are usually attributed to Jackson (1887), the British neurologist who drew the distinction in order to delineate primary from secondary neurological phenomena. Accordingly, positive symptoms are defined as the presence of a behavioral function that is otherwise not normally present in an individual. Negative symptoms are defined as the absence of a behavioral function that is usually present in the normal individual. Negative symptoms all appear to be characterized by the loss of an important cognitive or behavioral function, while positive symptoms represent an exaggeration or distortion of a normal function.

Standardized Assessment and Dimensional Differentiation of Negative Symptoms

To improve the reliability of diagnostic assessment of negative symptoms, several scales have been developed, mostly in the last two decades (SANS, Andreasen, 1981; negative subscore of the PANSS, Kay et al., 1987; NSA, Alphs et al., 1989; CAINS, Horan et al., 2011). Standardization of the assessment of negative symptoms through rating scales had a positive effect on inter-rater reliability. For example, the interclass correlations for the Scale for the Assessment of Negative Symptoms (SANS) and the Scale for the Assessment of Positive Symptoms (SAPS) with complex global ratings range between 0.70 and 0.88. Principally, the negative symptoms ratings have a somewhat lower inter-rater reliability than the inter-rater reliability of positive symptoms. The most recently developed scales claimed an increased inter-rater reliability, even in an international perspective, and improved validity.

Rating scales applied in the past decades included, among others, the anergia score of the Brief Psychiatric Rating Scale

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27041-4

(BPRS), the Manchester Scale and Crow's modification of this scale, the negative score of the Positive and Negative Scale (PANSS), the SANS, and Abrams and Taylor's Rating Scale for Emotional Blunting. Nowadays, the SANS and the PANSS are the most frequently used rating scales. It is not predictable whether these standard scales will be replaced more and more by the recently developed scales.

In nearly all studies in which a broad spectrum of negative symptoms was investigated by a factor analytical approach, two or three or sometimes even more dimensions of negative symptoms were found. This emphasizes the inadequacy of a unidimensional concept of negative symptoms in the complex phenomenology of this domain (Meltzer and Zureick, 1989; Horan et al., 2011).

Symptomatic Overlap with Depression and Parkinsonism

The differentiation between depression and negative symptoms is difficult on the basis of both clinical phenomenology and rating scales, and therapeutic effects on depressive symptoms might be misinterpreted as successful treatment of negative symptoms and vice versa. Gaebel (1989) reported a correlation of $r = 0.34 - 0.57$ at different measurement points for the BPRS subscores depression and anergia. In a follow-up study on schizophrenic patients, the correlation between the scores for depression and retardation/apathy in the Inpatient Multidimensional Psychiatric Scale (IMPS) amounted to $r = 0.35$ (Möller and von Zerssen, 1986). To overcome this problem, the Calgary Depression Scale was developed, which diminishes the overlap between negative symptoms and depressive symptoms.

Another problem in the interpretation of rating scale data from studies in this field is that the negative symptoms score correlates positively with scales measuring the neuroleptic-induced Parkinson syndrome. Furthermore, there is a correlation between negative symptoms and hyperkinetic neurological symptoms such as akathisia and dyskinesia. One could therefore conclude that drugs without extrapyramidal side effects, in general, may appear to be more advantageous in treating negative symptoms.

Negative Symptoms in Schizophrenia

Both Kraepelin and Bleuler described a core of schizophrenic symptoms corresponding to the modern concept of negative symptoms (**Table 1**). Later, this position was gradually displaced by an emphasis on productive features as being pathognomonic of schizophrenia, in fact, as being a prerequisite for its diagnosis. This is exemplified in the diagnostic criteria for schizophrenia put forward by Schneider 1959, in the research diagnostic criteria (Feighner et al., 1972), and in the *Diagnostic and Statistical Manual DSM-III*, *DSM-III-R*, and *DSM-IV* criteria (Spitzer et al., 1978).

The need to incorporate both negative and positive features into a model of schizophrenia was brought out by the factor analytic studies of the 1960s and in particular by the research of Strauss et al. (1974), who adapted Jackson's (1887) terminology for classifying different profiles in schizophrenia.

Table 1 Relationship between the concepts of basic and negative symptoms

Kraepelin (1904)	*Bleuler (1911)*	*Wing and Brown (1970)*	*Strauss et al. (1974)*
Cognitive decline ('stultification,' 'dementia')	Association disorder	Schizophrenic thought disorders, slowness of thought, and poverty of speech	Certain kinds of formal thought disorders, such as blocking
Flatness of affect	Affective disorder		
	Autism	Flatness of affect	Blunting of affect
Loss of intentional power	Ambivalence abulia, lack of initiative	Social withdrawal, low motivation	
Loss of energy	Lack of interest	Lack of initiative, underactivity	Apathy

Reproduced from Klosterkötter, J., 1990. Minussymptomatik and kognitive Basissymptome [Minus symptoms and basic cognitive symptoms]. In: Möller, H.J., Pelzer, E. (Eds.), Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik [Newer Approaches to the Diagnosis and Treatment of Schizophrenic Minus Symptoms]. Springer, New York, Berlin, Heidelberg, pp. 15–24.

Whereas Crow's (1980b) model viewed the positive and negative symptoms as annulated facets of schizophrenia that may reflect separate pathological processes, Andreasen (1982) proposed instead a typological model, which described positive and negative symptoms as opposing features that characterize different subtypes.

Which symptoms of schizophrenia are considered positive or negative varies across different authors. Nearly all authors include flat affect and poverty of speech among the negative symptoms and hallucinations and delusions among the positive symptoms. Most regard anhedonia, apathy, avolition, and abolia as negative symptoms. Thought disorder, bizarre behavior, and inappropriate affect, however, are variably classified as positive or negative.

Negative Symptoms in Other Psychiatric/Mental Disorders

Anergia, anhedonia, poverty of speech, and affective flattening, which occur as part of the negative symptoms in schizophrenia, are also observed in patients severely ill with depressive disorder. Although negative symptoms of schizophrenia may be different from depressive symptoms in depression, they do overlap, thus making negative symptom complexes and psychomotor retardation, lack of energy, anhedonia, and other symptoms of depression significantly associated. On the other hand, it was observed that a distinct hypokinetic syndrome involving anhedonia, depressed mood, and cognitive impairment existed not only in negative schizophrenia and retarded depression but also in Parkinson's disease, suggesting that reduced dopamine turnover might play a role in the pathophysiology of each.

Primary and Secondary Negative Symptoms

For theoretical and practical reasons, it is very important to differentiate between different types of negative symptoms (Möller, 2007). Especially the differentiation between primary and secondary negative symptoms, as proposed by Carpenter et al. (1985), seems of great relevance. While primary negative symptoms are part of a morbogenic deficit state, secondary negative symptoms are the consequence of

1. positive symptoms (e.g., social withdrawal because of paranoid ideas);
2. extrapyramidal effects (e.g., motor retardation as an indicator of neuroleptogenic akinesia);
3. depressive symptoms (e.g., postpsychotic or pharmacogenic depression); or
4. social understimulation (e.g., hospitalization).

According to this distinction, the treatment of negative symptoms has to be differentiated concerning the different causes.

Angst et al. (1989) underlined the necessity of discriminating negative symptoms according to different stages in the course of schizophrenia, i.e., the prepsychotic period, the productive psychotic period (characterized by a mixture of positive and negative symptoms, whereby negative symptoms may be hidden at this stage), and the postpsychotic period (characterized by a residuum with or without positive symptoms, postpsychotic endogenous, or reactive depression, postremissive exhaustion).

Crow (1980a) postulated the stability of a specific subtype of schizophrenia with positive symptoms (type I) and a subtype with negative symptoms (type II), explaining this by different biological causes. Most other authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally and cannot be taken as indicators for a specific biological or phenomenological subtype of schizophrenia.

Relationship Between Positive and Negative Symptoms in Schizophrenia

The positive/negative distinction may represent different manifestations of a common underlying process. It may also suggest the existence of different pathophysiologies in schizophrenia. Hints about the inter-relationships of these processes may be gleaned from the intercorrelations among their final common expressions as symptomatic phenomenology. An initial effort by Andreasen (1982) showed that the positive and negative symptoms clusters were significantly correlated within clusters and inversely correlated between clusters, suggesting a bipolar construct with positive and negative symptoms at the opposite ends of a continuum. This was replicated by another group. However, later investigations showed little or no relationship between positive and negative symptoms, strongly suggesting that they are independent processes within schizophrenia.

Since both positive and negative symptoms are frequent in schizophrenia, they are likely not to be completely independent (Möller et al., 2010). However, some findings have been controversial. For example, Rosen et al. (1984) failed to find a positive/negative symptom correlation.

Course of Negative Symptoms in Schizophrenia

Data on the course of negative symptoms are mostly not from spontaneous/natural course, but from therapeutic intervention studies, among others randomized control group studies, observational studies with defined treatments or follow-up studies with insufficient information about treatments. In acute first episode or remanifestation cases, negative symptoms are tightly linked to the presence of other manifestations of psychoses, especially to positive symptoms (Möller and von Zerssen, 1986). In several studies on the effect of neuroleptics, also called *antipsychotics*, in acute schizophrenic episode, antipsychotics reduced the negative symptoms, together with the positive symptoms, significantly (Möller, 2003).

Overall, negative symptoms appear to be more stable in subacute/subchronic patient samples. Pogue-Geile and Harrow (1985) recorded high and stable levels of anhedonia 1 and 3 years after episodes of schizophrenia. Biehl et al. (1986) found that negative symptoms had significant retest stability between years 2 and 5 in a sample from the International Pilot Study of Schizophrenia. In a 5-year follow-up study of schizophrenia carried out in Munich (Möller and von Zerssen, 1986), a decrease was observed in the mean score of positive symptoms and also in the syndrome retardation as an indicator of negative symptoms. The correlation for the retardation score between discharge and follow-up was 0.36. In a more recent 15-year follow-up study on first hospitalized patients with schizophrenia, a decrease of negative symptoms was found between admission and discharge and an increase between discharge and follow-up. At discharge the score of negative symptoms was higher than the score of positive symptoms, indicating that under long-term conditions they are more relevant than positive symptoms (Möller et al., 2010).

Many authors came to the conclusion that, in contrast to the views of Crow (1980b), the negative subtype and the positive subtype of schizophrenia are not stable conditions during the short- or long-term course of the disease. Most authors stressed the point that negative and positive symptoms can be combined cross-sectionally or longitudinally and cannot be taken as indicators of a specific biological or phenomenological subtype. To make the picture even more complicated, there is also a relatively close association with cognitive disturbances which together with the negative symptoms are interpreted as core symptoms of schizophrenia. Very important is that negative symptoms are closely associated with poor social functioning and therefore better treatments are urgently needed.

Biological Correlates of Negative Symptoms in Schizophrenia

The most prominent biochemical finding in schizophrenic patients with negative symptoms appears to be the reduction in central dopaminergic, serotonergic, and noradrenergic activities, and, if at all, a general common pattern can be described.

This decrease in amine activity tends to be associated with structural brain abnormalities, i.e., cortical atrophy or enlarged ventricles. There are indications that neuroleptics/antipsychotics reduce at least those negative symptoms of schizophrenia that are secondary to positive symptoms when the positive symptoms are effectively treated. However, negative symptoms of schizophrenia that occur independently of positive symptoms may also be reduced, especially through treatment with atypical antipsychotics, also called second generation antipsychotics, or by comedication with antidepressants or other compounds, e.g., glutamatergic compounds (Möller, 2003, 2004). The efficacy of neuroleptic treatment of negative symptoms seems to be related to their pharmacological profile, i.e., their interference with dopaminergic, noradrenergic, and serotonergic receptor systems. As to atypical antipsychotics, among others, the 5-HT_{2A} blockade was hypothesized as being a relevant mechanism for the efficacy of atypical antipsychotics in treating negative symptoms. More recently, a glutamatergic hypothesis came into the focus of neurobiological and psychopharmacological research.

As mentioned above, there is a considerable overlap in the symptomatology of psychiatric disorders with respect to negative symptoms. Also, different types of negative symptoms in schizophrenia have to be considered. In addition, the lack of homogeneity in the definition and assessment of negative symptoms in schizophrenia can be a problem. All these difficulties have consequences for research and explain some of the inconsistencies in biochemical and pharmacological schizophrenia research. The results of biochemical investigations as well as of positron emission tomography and magnetic resonance imaging studies on negative symptoms warrant further research efforts in this field to find, among others, a more rational neuroleptic drug treatment in negative symptoms of schizophrenia.

Treatment of Negative Symptoms in Schizophrenia

There has been a strong debate in the literature on whether neuroleptics can really reduce negative symptoms, and especially whether they can reduce primary negative symptoms and not just secondary negative symptoms. Some earlier reviews found evidence that some antipsychotics, especially some atypical antipsychotics, can reduce negative symptoms. However, since most of the trials were carried out in patients with negative and positive symptoms during acute exacerbations of schizophrenic psychoses, the question arises whether only those negative symptoms were reduced that were secondary to positive symptoms. In addition, the overlap between negative symptoms of schizophrenia and extrapyramidal side effects and depression has to be taken into account. As an alternative hypothesis to the direct effect of a drug on negative symptoms, the possibility that a drug reduces depression, akinesia, sedation, etc., and in this way also reduces negative symptoms, must be considered seriously.

Clearer results of drug treatment of negative symptoms may be obtained if more sophisticated designs are used for psychopharmacological trials on negative symptoms.

A European collaboration group, therefore, made methodological suggestions as early as the 1980s (Möller et al., 1994) focusing on patients with predominantly negative symptoms (Table 2), which were later adopted by the European drug authority, EMEA.

One possibility of differentiating between direct and indirect drug effects on negative symptoms is to use special statistical procedures. Formerly, only two studies had used an elementary statistical approach to account for the possibility of indirect effects on negative symptoms. Meltzer (1985) used a covariance analysis and found that the neuroleptic effect on negative symptoms could not be explained via a neuroleptic effect on positive symptoms alone. van Kammen et al. (1987) used a regression analysis which showed that the neuroleptic effect on negative symptoms could be explained by the effect on positive symptoms. More complex statistical approaches were not used until the mid-1990s. Möller et al. (1995) applied a complex statistical model, a path analysis, to the data of the North American Risperidone Study (Marder and Meibach, 1994), including effects on positive, negative, and extrapyramidal symptoms, to determine whether the atypical antipsychotic risperidone had a more potent direct effect on negative symptoms than haloperidol. The aim was to determine whether the (direct) effect on negative symptoms remains after the effects which are possibly caused through the reduction of positive and extrapyramidal symptoms have been excluded. The estimate from the path model showed that even after statistical control for indirect effects on secondary negative symptoms, risperidone improved negative symptoms to a greater extent than haloperidol.

While after the advent of the atypical antipsychotics, also called second generation antipsychotics, there was some enthusiasm that these compounds might have a much better efficacy in negative symptoms than the first generation antipsychotics, in the following years this view was challenged by

Table 2 Suggestions of the working group on negative symptoms in schizophrenia

Aim: evaluating treatment effects on negative symptoms in schizophrenia
Patient selection
Positive symptoms not dominating the actual clinical picture, e.g., PANSS negative type
Duration of negative symptoms >6 months
Stable condition of the schizophrenic illness >6 months
Flat affect and poverty of speech as core symptom of negative symptoms
No/low depression score
Design: double-blind comparison to placebo or active drug
Efficacy parameters: BPRS or SANS or PANSS
Other scales: depression scales, EPS scales
Statistical analyses
End-point comparison
Interaction with productive symptoms, depression, EPS

BPRS, brief psychiatric rating scale; EPS, extrapyramidal symptoms; PANSS, positive and negative syndrome scale; SANS, scale for the assessment of negative symptoms.

Reproduced from Möller, H.J., van Praag, H.M., Aufdembrinke, B., Bailey, P., Barnes, T.R., Beck, J., Bentsen, H., Eich, F.X., Farrow, L., Fleischhacker, W.W., Gerlach, J., Grafford, K., Hentschel, B., Hertkorn, A., Heylen, S., Lecrubier, Y., Leonhard, J.P., McKenna, P., Maier, W., Pedersen, V., Rappard, A., Rein, W., Ryan, J., Sloth Nielsen, M., Stieglitz, R.D., Wegener, G., Wilson, J., 1994. Negative symptoms in schizophrenia: considerations for clinical trials. Psychopharmacology 115, 221–228.

a lot of studies which did not prove such a significant superiority (Tandon et al., 2008).

Currently there is some hope that new compounds based on other pharmacological mechanisms like glutamatergic effects might have stronger effect on negative symptoms. Beside psychopharmacological treatment, also psychotherapeutic approaches were developed, especially kinds of behavior therapy.

See also: Depression; Differential Diagnosis in Psychiatry; Nosology in Psychiatry; Schizophrenia, Treatment of; Schizophrenia.

Bibliography

Andreasen, N.C., 1981. Scale for the Assessment of Negative Symptoms. University of Iowa, Iowa City.

Andreasen, N.C., 1982. Negative vs. positive schizophrenia. Definition and validation. Archives of General Psychiatry 39, 789–794.

Angst, J., Stassen, H.H., Woggon, B., 1989. Effect of neuroleptics on positive and negative symptoms and the deficit state. Psychopharmacology 99, S41–S46.

Alphs, L.D., Summerfeld, A., Lann, H., Muller, R.J., 1989. The negative symptom scale: a new instrument to assess negative symptoms of schizophrenia. Psychopharmacology Bulletin 25, 159–163.

Bandelow, B., Muller, P., Gaebel, W., Kopcke, W., Linden, M., Muller-Spahn, F., Pietzcker, A., Reischies, F.M., Tegeler, J., 1990. Depressive syndromes in schizophrenic patients after discharge from hospital. European Archives of Psychiatry and Clinical Neuroscience 240, 113–120.

Biehl, H., Maurer, K., Schubart, C., Krumm, B., Jung, E., 1986. Prediction of outcome and utilization of medical services in a prospective study of first onset schizophrenics. Results of a prospective 5-year follow-up study. European Archives of Psychiatry and Clinical Neuroscience 236, 139–147.

Bleuler, E., 1911. Dementia Praecox oder die Gruppe der Schizophrenien. [Dementia Praecox and the Group of Schizophrenias.]. Deuticke, Leipzig.

Carpenter, W.T., Heinrichs, D.W., Alphs, L.D., 1985. Treatment of negative symptoms. Schizophrenia Bulletin 11, 440–452.

Crow, T.J., 1980a. Molecular pathology of schizophrenia: more than one disease process? British Medical Journal 280, 66–68.

Crow, T.J., 1980b. Positive and negative schizophrenic symptoms and the role of dopamine. British Journal of Psychiatry 137, 383–386.

Crow, T.J., 1985. The two-syndrome concept: origins and current status. Schizophrenia Bulletin 11, 471–486.

Feighner, J.P., Woodruff-, R.A., Winokur, G., Robins, E., Guze, S.B., Munoz, R., 1972. Archives of General Psychiatry 26, 57–63.

Gaebel, W., 1989. Indikatoren und Prädiktoren schizophrener Krankheitsstadien und Verlaufsausgänge [Thesis: Indicators and Predictors of Stages in Schizophrenic Illness and Course Outcomes]. Free University of Berlin, Berlin.

Gaebel, W., 1990. Erfassung und Differenzierung schizophrener Minussymptomatik mit objektiven verhaltensanalytischen Methoden [Recording and differentiation of schizophrenic negative symptoms with objective behavior analytical methods]. In: Möller, H.J., Pelzer, E. (Eds.), Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik [Newer Approaches to the Diagnosis and Treatment of Schizophrenic Minus Symptoms]. Springer, New York, Berlin, Heidelberg, pp. 79–92.

Horan, W.P., Kring, A.M., Gur, R.E., Reise, S.P., Blanchard, J.J., 2011. Development and psychometric validation of the clinical assessment interview for negative symptoms (CAINS). Schizophrenia Research 132, 140–145.

Jackson, J.H., 1887. Remarks on evolution and dissolution of the nervous system. Journal of Mental Science 23, 25–48.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophrenia Bulletin 13, 261–276.

Klosterkötter, J., 1990. Minussymptomatik and kognitive Basissymptome [Minus symptoms and basic cognitive symptoms]. In: Möller, H.J., Pelzer, E. (Eds.), Neuere Ansätze zur Diagnostik und Therapie schizophrener Minussymptomatik [Newer Approaches to the Diagnosis and Treatment of Schizophrenic Minus Symptoms]. Springer, New York, Berlin, Heidelberg, pp. 15–24.

Kraepelin, E., 1904. Psychiatrie: Ein Lehrbuch [Psychiatry: A Textbook], 7th ed. Barth, Leipzig.

van Kammen, D.P., Hommer, D.W., Malas, K.L., 1987. Effect of pimozide on positive and negative symptoms in schizophrenic patients: are negative symptoms state dependent? Neuropsychobiology 18, 113–117.

Marder, S.R., Meibach, R.C., 1994. Risperidone in the treatment of schizophrenia. American Journal of Psychiatry 151, 825–835.

McCreadie, R.G., 1989. The Scottish first episode schizophrenia study. VII. Two-year follow-up. Acta Psychiatrica Scandinavica 80, 597–602.

Meltzer, H.Y., 1985. Dopamine and negative symptoms in schizophrenia: Critique of the type I-II hypothesis. In: Alpert, M. (Ed.), Controversies in Schizophrenia. Guilford, New York, pp. 110–136.

Meltzer, H.Y., Zureick, J., 1989. Negative symptoms in schizophrenia: a target for new drug development. In: Dahl, S.G., Gram, L.F. (Eds.), Clinical Pharmacology in Psychiatry. Springer-Verlag, Berlin, pp. 68–77.

Möller, H.-J., 2003. Management of the negative symptoms of schizophrenia: new treatment options. CNS Drugs 17, 793–823.

Möller, H.-J., 2004. Non-neuroleptic approaches to treating negative symptoms in schizophrenia. European Archives of Psychiatry and Clinical Neuroscience 254, 108–116.

Möller, H.-J., 2007. Clinical evaluation of negative symptoms in schizophrenia. European Psychiatry 22, 380–386.

Möller, H.-J., von Zerssen, D., 1986. Der Verlauf schizophrener Psychosen unter den gegenwärtigen Behandlungsbedingungen [The Course of Schizophrenic Psychoses under the Current Treatment Conditions]. Springer-Verlag, Berlin, Heidelberg, New York.

Möller, H.J., van Praag, H.M., Aufdembrinke, B., Bailey, P., Barnes, T.R., Beck, J., Bentsen, H., Eich, F.X., Farrow, L., Fleischhacker, W.W., Gerlach, J., Grafford, K., Hentschel, B., Hertkorn, A., Heylen, S., Lecrubier, Y., Leonhard, J.P., McKenna, P., Maier, W., Pedersen, V., Rappard, A., Rein, W., Ryan, J., Sloth Nielsen, M., Stieglitz, R.D., Wegener, G., Wilson, J., 1994. Negative symptoms in schizophrenia: considerations for clinical trials. Psychopharmacology 115, 221–228.

Möller, H.-J., Müller, H., Borison, R.L., Schooler, N.R., Chouinard, G., 1995. A path-analytical approach to differentiate between direct and indirect drug effects on negative symptoms in schizophrenic patients. A re-evaluation of the North American risperidone study. European Archives of Psychiatry and Clinical Neuroscience 245, 45–49.

Möller, H.-J., Jäger, M., Riedel, M., Obermeier, M., Strauss, A., Bottlender, R., 2010. The Munich 15-year follow-up study (MUFUSSAD) on first-hospitalized patients with schizophrenia or affective disorders: comparison of psychological course and outcome and prediction of chronicity. European Archives of Psychiatry and Clinical Neuroscience 260, 367–384.

Pogue-Geile, M.F., Harrow, M., 1985. Negative symptoms in schizophrenia: their longitudinal course and prognostic importance. Schizophrenia Bulletin 11, 427–439.

Ragin, A.B., Pogue-Geile, M., Oltmanns, T.F., 1989. Poverty of speech in schizophrenia and depression during in-patient and post-hospital periods. British Journal of Psychiatry 154, 52–57.

Rosen, W.G., Mohs, R.C., Johns, C.A., Small, N.S., Kendler, K.S., Horvath, T.B., Davis, K.L., 1984. Positive and negative symptoms in schizophrenia. Psychiatry Research 13, 277–284.

Rösler, M., Hengesch, G., 1990. Negative symptom in AMDP-system [Negative symptoms in the AMDP system]. In: Baum, U., Fähndrich, E., Stieglitz, R.D., Woggon, B. (Eds.), Veränderungsmessung in Psychiatrie und klinischer Psychologie [Measurement of Changes in Psychiatry and Clinical Psychology]. Profil, Munich, pp. 329–339.

Schneider, K., 1959. Clinical Psychopathology. Gune & Stratton, New York.

Spitzer, R.L., Endicott, J.E., Robins, E., 1978. Research Diagnostic Criteria (RDC) for a Selected Group of Functional Disorders, third ed. New York State Psychiatric Institute, New York.

Strauss, J.S., Carpenter Jr, W.T., Bartko, J.J., 1974. The diagnosis and understanding of schizophrenia. Part III. Speculations on the processes that underlie schizophrenic symptoms and signs. Schizophrenia Bulletin 11, 61–69.

Tandon, R., Belmaker, R.H., Gattaz, W.F., Lopez-Ibor, J.J., Okasha, A., Singh, B., Steind, D.J., Olie, J.P., Fleischhacker, W.W., Moeller, H.J., 2008. World Psychiatric Pharmacopsychiatric Section statement on comparative effectiveness of antipsychotics in the treatment of schizophrenia. Schizophrenia Research 100, 20–38.

Wagman, A.M.I., Heinrichs, D.W., Carpenter Jr, W.T., 1987. Deficit and nondeficit forms of schizophrenia: neuropsychological evaluation. Psychiatry Research 22, 319–330.

Wing, J.K., 1989. The concept of negative symptoms. British Journal of Psychiatry 155, 10–14.

Wing, J.K., Brown, G.W., 1970. Institutionalism and Schizophrenia. Cambridge University Press, Cambridge.

Psychiatric Care, Economics of

Martin Knapp, Personal Social Services Research Unit, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Tyutyulkova, S.S. Sharfstein, volume 18, pp. 12267–12272, © 2001, Elsevier Ltd.

Abstract

The primary aim of a health care system is to improve the health of patients and families, and if available resources can be used more efficiently then greater health improvements can be achieved. Economic evaluations offer important insights. This article describes cost-effectiveness and related analyses, how they measure outcomes and costs, and how decision-makers can be supported in making trade-offs when better health outcomes can only be achieved at greater cost. Examples of mental health evaluations are used to illustrate these approaches.

Introduction

It sometimes feels as if it took a devastating global macroeconomic shock to wake some people up to the fact that resources are scarce. In fact, scarcity is the perennial state of the world: there are never enough resources to meet everyone's needs or to satisfy everyone's wants. Consequently, those individuals who control, manage, or allocate resources have to make tough choices about how to use them. This affects not only decision-makers at national strategic level, such as government ministers and chief executives of major corporations, but also almost every purchaser, provider, and professional within health, social care, housing, and other systems. It affects individuals, families, and communities too. Indeed, anyone juggling their own or their organization's budget – or managing nonmonetary resources such as their own or staff members' time, in-patient beds, therapy hours, or medicines – will be acutely aware that resources are almost never sufficient to respond positively to every demand or request for them.

If we assume that each of these decision-makers is reasonably rational, then they are likely to want to use their available resources to achieve the best outcomes. In the health field, those outcomes might be gauged in terms of how well needs are met, or symptoms alleviated, or pain reduced, or normal functioning restored, or quality of life improved. This is where economics can be helpful: a core aim of economic analysis is to help decision-makers to achieve the goal of using available resources so as to achieve the best possible outcomes. In this article I will look at economic evaluation – particularly cost-effectiveness analysis and some of its variants – and how it can support decision making. I will describe what cost-effectiveness analysis aims to do, and what information it needs, offering some illustrations from the psychiatric field.

Decisions-makers face tough choices between competing uses of scarce resources. How does a hospital decide how many resources should be allocated to each medical specialty, or what the balance should be between in-patient and out-patient services? How does a clinician or clinical team decide which individuals should be treated and in what priority order? When offering treatment, which medication or psychological therapy or other approach should be chosen for which patients? Looking more broadly, what proportion of available resources should be spent addressing today's needs and what proportion should be invested in strategies that could prevent the emergence of new needs? Broader still, how does a national or regional government decide how much to allocate to health rather than (say) education, housing, or the armed forces? None of these questions is solely or even primarily an economic issue, and none of them will be answered satisfactorily only with economic evidence. But each and every one of these questions has an economics element, because whatever decision is taken will undoubtedly have economic implications. Evidence from an economic evaluation can therefore help the decision-maker with those difficult choices. At the same time, these decision-makers will need and want to weigh up other considerations such as fairness, equity of access, dignity, environmental sustainability, and so on. These are all valid and important aims, but I will not be covering them here.

Economic Evaluation

Among the key questions that need to be addressed when considering whether to license, recommend, prescribe, purchase, or use a particular intervention, two are particularly pertinent here. (I will use the term 'intervention' generically in this article, covering services, treatments, preventive strategies, or wider policy initiatives.) The first question is 'Does it work?': is the intervention effective in reducing or removing psychiatric symptoms, or improving health-related quality of life, or achieving other outcomes? If the answer to this first question is likely to be yes – that is, if the intervention appears to be effective – then the second question to be addressed must be 'Is it worth it?': is the cost of the intervention worth incurring in order to achieve that level of effectiveness? This second question is what an economic evaluation such as a cost-effectiveness analysis tries to answer. However, neither of these two questions can be answered adequately without reference to some comparator, even if that comparator is 'doing nothing' (placebo).

This then suggests that there are basically five steps for an economic evaluation of the cost-effectiveness kind:

1. the question that needs to be answered by the evaluation needs to be made clear;
2. two or more interventions are then compared;

3. the outcomes of each intervention are assessed;
4. the associated costs of each intervention are measured; and
5. the trade-offs between better outcomes and higher costs are then considered.

Economic evaluations can be employed to answer quite broad and also relatively narrow questions, and it is important to be clear what the purpose of the study is; I discuss this in Section The Evaluative Question to Be Answered. The best way to make a comparison might be to carry out a randomized controlled trial, although other approaches might be suitable in some circumstances. I will not rehearse the arguments for and against randomized trials here, but I will later describe how simulation modeling can supplement or extend trial data (Section Modeling). The outcomes to be assessed will depend crucially on what question the evaluation is seeking to address; I briefly discuss the options in Section Effectiveness and later introduce two types of outcome measure that health economists have often employed: utility (Section Utility Measurement) and monetary benefits (Section Monetary Benefits). Cost measurement is discussed in Section Costs, not just in terms of health system resources used to deliver the interventions, but also the wider resource implications to health care and other systems and to patient and families. The final element of an economic evaluation – bringing the outcome and cost findings together – is considered in Section Making Trade-Offs. Some of the challenges and issues raised by economic evaluation are discussed in Section Some Issues and Challenges, before a brief conclusion (Section Conclusions).

The Evaluative Question to Be Answered

There are a few different 'types' or approaches to economic evaluation, as described below, but they share some common features. In particular, they share a common approach to the conceptualization, definition, and measurement of costs, as outlined in the next section. The main differences between the approaches are with respect to how they define and assess outcomes; these differences arise primarily because different approaches are needed to answer slightly different questions.

If the question addressed by an economic evaluation relates to patients with a particular clinical need (such as severe depression or Alzheimer's disease) and is concerned with the economic case for one therapy over another, then the outcomes that are most relevant are those specific to the diagnosis (such as alleviation of depressive symptoms or slowing down of cognitive decline) as well as some more general indicators (such as quality of life and personal functioning). In this case, health economists would carry out a form of evaluation called cost-effectiveness analysis. It is the most commonly conducted type of economic evaluation in the health field, with hundreds of new studies published each year, quite a few of them in the area of psychiatry. In this evaluative approach, outcomes are measured in exactly the same way as the clinical researcher would want to measure them, and costs are measured with greater or lesser breadth, depending on the context (see Section Costs). A cost-effectiveness analysis tells decision-makers what course of action (e.g., what treatment) most efficiently addresses a specific set of clinical needs. Some examples are given below.

Some decision-makers face broader questions. For example, they may need to decide whether to treat more people with depression rather than spending their available funds elsewhere in the health system, say on treating coronary heart disease. If that is the case, they not only need to know the costs of the different options, but also some common measure of outcome that would allow them to compare the impact of depression treatment with the impact of coronary heart disease treatment. Because diagnosis-specific symptoms are quite different between depression and coronary heart disease, a generic outcome measure is needed that has relevance in both clinical areas. Economists have developed the 'utility' measure, often operationalized in terms of quality-adjusted life years (QALYs), which I will come back to in Section Utility Measurement. Another generic measure is the disability-adjusted life year, more commonly used in evaluations in low-income countries, but also widely employed as a generic way to compare levels of 'disability burden' across disease areas and countries (e.g., Murray et al., 2012). Using a generic outcome measure like a QALY, it is possible to conduct a cost-utility analysis. This form of evaluation tells the (strategic) decision-maker where they will get the biggest impact – across a range of clinical areas – in terms of health-related quality of life relative to the resources expended.

The evaluation could be broader still, especially if the decision-maker needs to choose how to allocate resources between (say) expenditure on health care, broadly defined, and expenditure in another arena, such as education or social care. Once again, the economic evaluation would need to compare costs between the various alternatives, which is usually straightforward, but now outcomes would need to be measured to allow sensible comparison. One way to do this is to value those outcomes in monetary terms; this is what health economists call a cost-benefit analysis (see Section Monetary Benefits).

We can therefore see how the policy or practice question that the evaluation needs to address will influence the type of economic evaluation that should be carried out. However, as shown in some of the illustrative studies set out later, a single study can usually support more than one evaluative approach if the right measures are employed.

Costs

Some costs are directly associated with the treatment of a disorder, such as the money spent on medications and psychiatrists' salaries. Other costs are indirect, such as lost productivity because of ill-health, which can be very high in the mental health area (Organization for Economic Cooperation and Development, 2012), or the family cost of unpaid care and support, which can be even higher, for example in the case of dementia (Alzheimer's Disease International, 2010). Productivity losses arise because of absenteeism, reduced performance while at work ('presenteeism'), early retirement, reduced opportunities for career development, days 'out of role' for people not in paid work, and reduced lifetime productivity due to premature mortality. Costs for unpaid carers stem from similar productivity losses, as well as out-of-pocket expenses (payments for treatment or to transport the

person with mental health needs no services) and psychological strain.

An economic evaluation would therefore need data on the services individuals use and how frequently they are used and the unit (average) costs of each of them (Knapp and Beecham, 1990). Other data would be needed on the impact of poor mental health (and of effective treatment) on productivity, such as absenteeism or unavailability for work, and on the associated costs, usually approximated by a typical or average national wage rate (under the assumption that wage equals marginal productivity). If an individual receives support from a family member or other unpaid carer, then the amount of such support provided (e.g., the number of hours in a week) would need to be estimated and again a cost attached to those hours. Some researchers would value these hours of unpaid care at a typical wage rate, and some have used a replacement cost approach (taking a value equal to the cost of employing a homecare worker, for example).

Effectiveness

Arguably the most intuitive type of economic evaluation to a psychiatrist or mental health decision-maker is cost-effectiveness analysis (CEA), which measures costs as set out above and outcomes along dimensions that would be recognized by clinicians and other service professionals, such as changes in symptoms, behavior, functioning, and health-related quality of life. A CEA can then help decision-makers choose between interventions aimed at specific health needs, such as two different treatments for depression. For example, change in depressive symptoms could be measured using the Hamilton rating scale (Hamilton, 1960), and then the evaluation would compute the difference in costs between two interventions over a specified period of time (say 6 months) and the difference in the change in Hamilton score over that same period. If one intervention achieves a greater change in Hamilton score than the other, and is also the less costly of the two, then it would be straightforward to see that it is more cost-effective – it improves outcomes and saves money.

An example can be given from a study comparing two treatments for people with severe depressive episodes (unipolar or bipolar): electroconvulsive therapy (ECT) with repetitive transcranial magnetic stimulation (rTMS). Outcomes were assessed at the end of treatment, and then 3 and 6 months later, and costs calculated during the treatment period and over the 6-month follow-up. On the primary outcome – the Hamilton rating scale – ECT was significantly more effective at the end of the treatment period, but by the 6-month follow-up point there was no difference between the treatment groups. On the secondary outcome measures – mood self-ratings on the Beck Depression Inventory (BDI; Beck et al., 1996) and the Brief Psychiatric Rating Scale (Overall and Gorham, 1962) – patients randomized to ECT were doing better at the 6-month follow-up point (Eranti et al., 2007). Costs were significantly lower for the ECT group than the rTMS group over the 6-month period, and rTMS was clearly more cost-effective than ECT (Knapp et al., 2008).

A cost-effectiveness analysis does not have to be confined to a single outcome measure (in which case it is sometimes called a cost-consequences analysis). Now the evaluation looks at the cost differences and a range of effectiveness differences – one for each relevant outcome dimension. The economic evaluation of ECT and rTMS not only had Hamilton as the primary outcome but also calculated QALY gains over the evaluation period. These were based on the SF-6D measure (Brazier et al., 2002), with societal weights attached. The economic evaluation could therefore compute incremental cost-effectiveness ratios for each outcome measures in turn. There was no difference in QALY gain between the interventions.

This cost-consequences approach has the advantage of breadth, but poses a challenge if one intervention is found to be comparatively better on one outcome measure, but comparatively worse on another. The preferred option in this case is not clear and decision-makers must weigh up the strength of evidence, perhaps on the basis of value judgments about the relative importance of the respective outcome dimensions.

Making Trade-Offs

An intriguing and difficult challenge in an economic evaluation – and obviously in the very real world of decision making – is the scenario mentioned just now: when one intervention is more effective than another but also has higher costs. Which of the two interventions then represents the best use of resources? The answer hinges on the extent to which the decision-maker is prepared to trade-off better outcomes for lower costs. In other words, what is the decision-maker's willingness to pay for the better outcomes? Not surprisingly, there is no single or simple answer: it is a matter of value judgment.

The usual way to present evidence from an economic evaluation to illustrate this trade-off is to calculate the incremental cost-effectiveness ratio. This ratio divides the extra cost associated with a new intervention by its additional effect. For example, in a study of computerized cognitive behavioral therapy (CCBT) compared to treatment as usual for people with depression or anxiety, it was found that CCBT was better in terms of alleviating symptoms and improving work and social functioning over an 8-month follow-up period (Proudfoot et al., 2004), but it was also more expensive to the health service (McCrone et al., 2004). The cost of achieving an incremental improvement in symptoms as measured by the BDI – i.e., the difference in cost between CCBT and usual care, divided by the difference in BDI score at the follow-up point – was calculated to be £21, equivalent to a cost of £2.50 per depression-free day over the follow-up period. The decision-maker must decide if that additional cost is worth paying.

Health economists often use cost-effectiveness acceptability curves (CEACs) to show the trade-off that a decision-maker will need to make in order to decide whether higher costs are justified by better outcomes (Petrou and Gray, 2011). In the case of the CCBT evaluation, the plotted CEACs showed that, even if the value placed by the decision-maker on a unit reduction in the BDI was as low as £40, there was still an 81% probability that CCBT would be cost-effective. Similarly, assigning a societal value of just £5 to each additional depression-free day would result in an 80% probability that CCBT would be cost-effective (McCrone et al., 2004).

Utility Measurement

Some decisions about resource allocations require comparisons to be made that are beyond a specific diagnostic context, such as when setting departmental budgets within a hospital or taking strategic decisions about which health needs to prioritize. As mentioned earlier, economists have developed a single, over-arching measure called utility, usually operationalized as a preference-weighted, health-related quality-of-life measure, with the QALY being the best-known and most widely used (Petrou and Gray, 2011). Values are elicited from the general public on the quality of life associated with different states of health (through a series of robust field studies), with a value of 1 being assigned to a state of perfect health and a value of 0 assigned to death. The extra years of life that flow from an intervention can then be weighted by the quality of each of those years.

The study of CCBT introduced earlier included an assessment of QALYs. These were approximated from the BDI: the cost per QALY was £2190, the interpretation of which will be discussed later. It is more common to use a specific QALY-generating tool; the EuroQol (EQ-5D) is one such most widely used measure (EuroQol Group, 1990). An example of an evaluation that took this approach looked at a manual-based coping strategy intervention for family carers of people with dementia, comparing the new program to support as usual in a randomized controlled trial (Knapp et al., 2013b). Carers in the intervention group were found to have better outcomes as measured in terms of both affective symptoms (Livingston et al., 2013) and QALY gains. Costs were actually slightly higher for the carers who received the coping strategy intervention, compared to those who were supported as usual, with the cost-effectiveness ratio being £6000 per QALY gained. How is this ratio to be interpreted?

In England and Wales, the National Institute for Health and Care Excellence (NICE) employs a framework to help the groups developing its clinical, public health, and social care guidelines to decide whether better outcomes are 'worth' the higher costs sometimes necessary to achieve them. Their approach uses the generic QALY measure of health outcome and suggests a threshold value for cost-effectiveness: a medication or other intervention that costs more than £20 000 per QALY would generally not be considered 'worth' it since – NICE would argue – the resources (represented by the cost) could probably be better spent somewhere else in the health system. This threshold is only there to provide guidance – it is certainly not a rigid rule – but has the signal virtue of reminding everyone – doctors and nurses, patients and carers, taxpayers, and voters – that resources are not limitless, and that difficult choices have to be made about how to deploy them (Appleby et al., 2007; Cerri et al., 2014). Some other countries have now established similar processes to guide allocation decisions (McDaid and Cookson, 2003).

Almost every health economic evaluation would use a QALY measure nowadays, but always alongside symptom or similar 'clinical' measures of outcome. Although it is the most widely used QALY-generating tool, the EQ-5D does not always perform as sensitively as needed in research on treatments for some severe mental health problems or for the youngest or oldest patients (Brazier, 2010). This difficulty has encouraged the development of disease-specific tools to measure QALYs, such as the DEMQOL-U for studies of people with dementia (Mulhern et al., 2012).

Monetary Benefits

Cost-benefit analysis is a useful form of economic evaluation. Because costs and outcomes (benefits) are now measured in the same monetary units it is possible to see whether the benefits exceed costs, that is, whether the evaluation provides support for the intervention to go ahead. With two or more interventions to be compared, the one with the greatest net benefit (the difference between benefits and costs) would be seen as the most attractive from the efficiency standpoint. Cost-benefit analyses are also useful when decision-makers have to allocate resources across different sectors (e.g., health care, housing, education, and so on). But they are also intrinsically difficult to do, because it is generally not easy to convert a clinical outcome measure such as an improvement in symptoms or quality of life into a monetary value.

One of the few cost-benefit analyses carried out in the mental health area evaluated Individual Placement and Support (IPS) for people with severe mental health problems who are unemployed but would like to work, compared to standard vocational rehabilitation. IPS has a lot of promise as an approach to establish people in paid employment (Bond et al., 2012), but almost all studies have been carried out in a single setting. A randomized controlled trial was conducted to examine the economic case for IPS within and between six European countries (Burns et al., 2007). Two primary outcomes were chosen: additional days worked in competitive settings and additional percentage of individuals who worked at least one day. The former could be converted to a monetary measure by attaching a value equal to average gross wage (assumed to be equal to the value of productivity added to the economy). The study found a large difference in net benefit in favor of IPS, representing a more efficient use of resources than standard vocational rehabilitation (Knapp et al., 2013a).

Modeling

A randomized controlled design is undoubtedly the most robust approach to adopt in order to answer a range of questions, such as when deciding which medication is better for the treatment of a particular illness, and the examples above offer illustrations of how such a design can easily accommodate an economic evaluation. But a randomized design might not always be feasible in evaluating whether, say, some preventive strategies are effective or cost-effective. It might also be infeasible, and possibly unethical, to ask people to remain in randomly allocated groups for many years while the research team gathered data to assess long-term consequences of different interventions. Moreover, decision-makers today often want evidence somewhat faster than would normally be possible with a study that is collecting new information (primary data).

An approach that can sometimes be used is some form of mathematical modeling such as a simple decision tree or

a Markov approach, populated with data from previous trials, observational studies, or routine management information systems (secondary data). Models are representations of what might happen in reality, for example, tracing pathways through care for individuals with particular characteristics or needs; estimating the associated outcomes and costs; and then comparing them in order to gauge whether one is more effective, cheaper, or more cost-effective than the other (Briggs et al., 2006). Models are more generalizable and flexible than studies that collect primary data, and they can also generate findings much earlier. But they are clearly simplifications of reality and can only be as good as the data available to populate them, so that their limitations do need to be borne in mind.

Some examples can be offered. There has been extensive research on parenting programs: Bonin et al. (2011) modeled the economic pay-offs from parenting programs where there is a child in the family with a conduct disorder. Based on data gathered by others in around 20 randomized control trials, they examined the economic case for parenting programs over 25 years by tracing through the possible consequences for service use (health, social care, special education), contacts with the criminal justice system (particularly in adolescence and early adulthood), and patterns of employment. The cost of a parenting program averaged around £1200 per family in 2008–09 prices, but the economic return over a 25-year period was calculated to be between 2.8 and 6.1 times greater than this intervention cost.

A similar modeling approach was employed in investigating the economic case for 15 mental health promotion and mental illness prevention initiatives (Knapp et al., 2011). For each intervention, there was already well-established evidence of effectiveness. The aim was to examine whether there were economic pay-offs from these interventions in terms of direct (immediate or longer-term) cash savings to the public sector, to employers, or to the wider society (e.g., through crimes averted). Investigating the wider impacts was important, given the extensive impact of many mental health problems. The research gave national and local decision-makers information on whether interventions they were keen to promote on the grounds of health improvement and quality-of-life gains were also at least affordable or perhaps even generated significant economic gains.

Some Issues and Challenges

Economic evaluations can be carried out from a range of perspectives: looking only at the economic impacts on the health system perhaps, or on the government sector, or on the whole of society. Different perspectives for a study are driven by a need to answer slightly different policy or practice questions. The perspective adopted will have implications for how broadly costs are measured, and it might also have implications for whether or not the intervention under study is found to be cost-effective.

Some interventions might lead to (relative) cost savings, which could be attractive to purchasers or policy-makers. However, sometimes those savings do not accrue immediately or they might be achieved in budgets other than the one from which the intervention was funded. Some savings might not be easily 'cashable,' such as when an intervention reduces hospital admissions but does not (immediately or perhaps ever) close beds or save expenditure as a result. Another saving that is not 'cashable' is when the burden of unpaid care from families is reduced. It is generally harder to argue the economic case for an intervention when its main impacts are some years away rather than immediately when any savings or other economic benefits are enjoyed in other sectors and when impacts cannot readily be turned into 'real' savings.

A related challenge is when an intervention does not actually reduce expenditure (i.e., it is not cost-saving), but it still appears to be a good use of resources because the outcomes it generate are considered to be worth it. As noted earlier, if one intervention has better outcomes and also higher costs, then a trade-off has to be made. At a time when budgets are under considerable pressure, it can be hard to persuade decision-makers to spend more than they are currently, even if the clinical benefits experienced by patients are impressive. The purpose of a health system, of course, is to improve health rather than to reduce expenditure, so an intervention that can substantially improve health ought to be given proper consideration. The difficulty is that the decision-maker only has so many dollars or euros to spend.

Conclusions

Decision-makers within health systems across the world need to balance their budgets. They need to do so while pursuing their primary objective, which is to maximize the impact that services and treatments can have on the health and quality of life of patients and families. This is a tough task. Resources are scarce relative to the demands for them, and also scarce relative to the needs of the population. The upshot is that careful choices always have to be made about to utilize those resources. Economics evidence should not dominate decision making in health, social care, or other systems, but it does need to be stirred into the mix along with evidence on what needs are or are not being met, what other outcomes are achieved, and how well care and support arrangements match up to the preferences of patients and families.

See also: Child and Adolescent Psychiatry, Principles of; Cognitive Behavioral Therapy; Dementia: Psychiatric Aspects; Depression; Evidence-Based Psychiatry; Mental Disorders, Epidemiology of; Psychopharmacotherapy, Costs of; Social Intervention in Psychiatric Disorders.

Bibliography

Alzheimer's Disease International, 2010. World Alzheimer's Report 2010: The Global Economic Impact of Dementia. Alzheimer's Disease International, London.

Appleby, J., Devlin, N., Parkin, D., 2007. NICE's cost-effectiveness threshold. British Medical Journal 335, 358–359.

Beck, A.T., Steer, R.A., Brown, G.K., 1996. Beck Depression Inventory, Second Edition Manual. Psychological Corporation, San Antonio Texas.

Bond, G.R., Drake, R.E., Becker, D.R., 2012. Generalizability of the individual placement and support (IPS) model of supported employment outside the US. World Psychiatry 11, 32–39.

Bonin, E., Stevens, M., Beecham, J., Byford, S., Parsonage, M., 2011. Costs and longer-term savings of parenting programmes for the prevention of persistent conduct disorder: a modelling study. BMC Public Health 11, 803.

Brazier, J., 2010. Is the EQ-5D fit for purpose in mental health? British Journal of Psychiatry 197, 348–349.

Brazier, J., Roberts, J., Deverill, M., 2002. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 21, 271–292.

Briggs, A., Claxton, K., Sculpher, M., 2006. Decision Modelling for Health Economic Evaluation. Oxford University Press, Oxford.

Burns, T., Catty, J., Becker, T., et al., 2007. The effectiveness of supported employment for people with severe mental illness: a randomised controlled trial in six European countries. The Lancet 370, 1146–1152.

Cerri, K., Knapp, M., Fernandez, J.-L., 2014. Decision making by NICE: examining the influences of evidence, process, and context. Health Economics, Policy and Law 9, 119–141.

Eranti, S., Mogg, A., Pluck, G., et al., 2007. A randomised controlled trial with six-month follow-up of repetitive transcranial magnetic stimulation and electroconvulsive therapy for severe depression. American Journal of Psychiatry 164, 73–81.

EuroQol Group, 1990. EuroQoL: a new facility for the measurement of health-related quality of life. Health Policy 16, 199–208.

Hamilton, M., 1960. A rating scale for depression. Journal of Neurology, Neurosurgery and Psychiatry 23, 56–62.

Knapp, M., Beecham, J., 1990. Costing mental health services. Psychological Medicine 20, 893–908.

Knapp, M., Romeo, R., Mogg, A., et al., 2008. Cost-effectiveness of transcranial magnetic stimulation vs. electroconvulsive therapy for severe depression: a multi-centre randomised controlled trial. Journal of Affective Disorders 109, 273–285.

Knapp, M., Patel, A., Curran, C., et al., 2013a. Supported employment: cost-effectiveness across six European sites. World Psychiatry 12, 60–68.

Knapp, M., King, D., Romeo, R., et al., 2013b. The START study (STrAtegies for RelaTives): cost-effectiveness of a manual-based coping strategy programme in promoting the mental health of family carers of people with dementia: a pragmatic randomised controlled trial. British Medical Journal 347, f6342.

Knapp, M., McDaid, D., Parsonage, M., 2011. Mental Health Promotion and Mental Illness Prevention: The Economic Case. Department of Health, London.

Livingston, G., Barber, J., Rapaport, P., et al., 2013. The START study (STrAtegies for RelaTives): clinical effectiveness of a manual based coping strategy programme in promoting the mental health of family carers of people with dementia: a pragmatic randomised controlled trial. British Medical Journal 347, f6276.

McCrone, P., Knapp, M., Proudfoot, J., et al., 2004. Cost-effectiveness of computerised cognitive behavioural therapy for anxiety and depression in primary care: randomised controlled trial. British Journal of Psychiatry 185, 55–62.

McDaid, D., Cookson, R., 2003. Evaluating health care interventions in the European Union. Health Policy 63, 133–139.

Mulhern, B., Rowen, D., Brazier, J., et al., 2012. Development of DEMQOL-U and DEMQOL-Proxy-U: generation of preference-based indices from DEMQOL and DEMQOL-Proxy for use in economic evaluation. Health Technology Assessment 17, 1–160.

Murray, C., Vos, T., Lozano, R., et al., 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990—2010: a systematic analysis for the global burden of disease study 2010. The Lancet 380, 2197–2223.

Organization for Economic Cooperation and Development, 2012. Sick on the Job? Myths and Realities about Mental Health and Work. OECD, Paris.

Overall, J.E., Gorham, D.R., 1962. The brief psychiatric rating scale. Psychological Reports 10, 799–812.

Petrou, S., Gray, A., 2011. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. British Medical Journal 342, d1766.

Proudfoot, J., Ryden, C., Everitt, B., et al., 2004. Clinical efficacy of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trail. British Journal of Psychiatry 185, 46–54.

Psychiatric Social Work in the United Kingdom

Colin Pritchard, Bournemouth University, Bournemouth, UK; and University of Southampton, Southampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Mental health services still face the dilemma of 'care versus control,' reflecting societal anxiety about mental illness. Psychiatric Social Work in adult services work as Approved Social Workers predominantly concerned with formal admission. In the Child and Adolescent Mental Teams the role is more integrative. Recent research indicates the importance of the interactive bio-psycho-social model to address the increases in Attention Deficit Hyperactivity Disorder (ADHD) and Autism and recognizes the psychiatric–child protection interface to prevent extremes of abuse, assisting children of psychiatric parents to deal with the impact of mental illness – thus making children safer and contributing to the reduction of subsequent problems in their adulthood.

Introduction: Socio-Historical and Policy Perspectives

Psychiatrist social work is a specialism within the profession and has developed differently in diverse parts of the world, although there will be recognizable elements, particularly those concerned with the stigma usually associated with mental ill health, that cross borders. In this article, I consider the development of psychiatric social work primarily in the UK. As I have been a psychiatric social worker for most of my professional career, the history of this specialism in this country is intertwined with my own journey within it.

The social policy analyst Richard M. Titmuss (1965: p. 4) said, "Reality starts with history." Britain's provision for people with mental health problems emerged from medieval ideas and beliefs about 'possession.' This then assumed the need for a priestly intervention, until influenced by the 'Enlightenment.' This resulted in a slow movement toward an effort of 'caring' for the 'deranged,' albeit with strong control elements because of fears of "moon-struck madness, moping melancholia and demonic frenzy" (Milton, 1674: p. 485).

Psychiatric Social Work (PSW) in the United Kingdom over the past 50 years coincides with my career, because in 1957 I admitted my first compulsory admission under the 1890 Lunacy Act. This was to the local county mental hospital, which in the 1950s typically held almost 2000 patients, of whom less than 20% were 'Voluntary,' having been admitted under the 'modernizing' Mental Health Act 1930. Thus we struggled with the issue of care versus control. An example of this dilemma was seen in a high-profile tragedy where a European Senator killed his mother. Due to his status, the case was referred personally to the country's President, who decreed:

> if he was insane when he killed his mother and not pretending to be insane, you are not to be concerned about punishment because insanity is punishment enough … [however] … You tell me he was in the care of friends and guardians when the murder took place, therefore you should inquire whether there is a case that one or more were guilty of neglect and you should deal with them accordingly.

Students are surprised to learn that the 'President' in question was the second-century Roman Emperor Marcus Aurelius. Yet even this philosopher-Emperor asked if there was someone to blame – so, there is little that changes!

One early impetus for change followed the attack on the British King George III, by a delusional Mr McNaughton. However, the King showed some fellow sympathy and so McNaughton's Rules entered British case law. According to this law, mentally sick persons could not be held criminally guilty because of diminished responsibility, as their delusions make them incapable of understanding right from wrong (Brown, 2009). The 1890 Lunacy Act was concerned with safety; so the magistrates (and not 'doctors of the mad') admitted people to 'lunatic asylums.' Nothing in this Act mentioned treatment or discharge, so apart from response to families making an appeal, persons committed to the asylums were kept there and this law resulted in overemphasis on keeping people secure. What was intended to be a civilized refuge became repressive; and the asylums (created in 1813) became full and fuller.

A seminal attitudinal change took place following the First World War and the recognition of 'shell shock' (understood today as Post Traumatic Stress Disorder). Middle-class men of sensitivity, such as the war poets Wilfred Owen and Siegfried Sassoon, gave the wider public some idea of the pity and horrors of war.

Developments in the new medical specialty of psychiatry, and the realization that a mental breakdown need not be permanent, led to the 1930 Mental Treatment Act, which eliminated the term 'Lunatic Asylum' and facilitated short-term admission into 'Mental Hospitals', so that 'voluntary' admission' became an option for patients.

After the Second World War, under the reforming Labour Government, the Welfare State was introduced. In 1948 the Poor Law of 1601 was abolished. Social work developed along two lines: either as 'social' or as 'psychological' within the then various welfare services. For example, adult or elderly care (disability, homelessness) was more social and financial, whilst mental health and children's services were more psychological.

The modest changes implemented under the Mental Health Act 1930 however, did little to reduce the number of people incarcerated and this did not change until the seminal 1959 Mental Health Act (MHA), which placed the emphasis upon 'treatment'; and ensured that whenever possible, patients were admitted with minimal restrictions. The Psychiatric Social Workers' input was as Mental Welfare Officers (MWOs) who

increasingly were trained; and although they had a role in compulsory admissions, they were charged to seek minimum restriction for the individual and were responsible for the support of the patients in the community, seeking to make an alliance between patient and family. The success of the act was seen in the closure of the large psychiatric hospitals that had produced 'institutionalized' patients and staff.

The Mental Health Act of 1959 abolished the difference between general and psychiatric hospitals. All hospitals today have a psychiatric unit; and the only separate psychiatric service are the forensic units, concerned with people considered dangerous.

So, within 10 years, instead of most patients being 'certified,' nearly 90% had become 'voluntary,' Alongside a 'care-in-the-community' emphasis, the large mental hospitals were closed and patients went to psychiatric departments in general hospitals, reflecting Nye Bevan's great vision of integrating health and social care at the core of the National Health Service (NHS) Act of 1948. This is still ostensibly the aspiration for the NHS.

This period of progressive community care was linked to the new psychopharmacological medicines, giving confidence that people could be cared for and controlled in the community. However, there was the danger of the 'chemical cosh' and fears of 'soviet type psychiatry,' led by critics such as British psychiatrist RD Laing and Thomas Szasz from the USA, who challenged the authenticity of the concept of 'mental illness' (Pritchard, 2006).

The Development of Independent Social Work Departments

One problem that emerged following the establishment of the British Welfare State was that social work was spread across many different agencies, both public and voluntary. There was an overlap of services; whilst central government departments were responsible for national policy, the actual services were provided and delivered by the Local Authorities (LA). Thus the Children's Department, and Child, Adolescent, and Adult Mental Health, were the responsibility of the Departments of Health and Probation, the Home Office, the Educational Welfare Officer, the Department of Education and Housing, and Welfare Services under the Local Authority; it meant that the same family could be visited by six different Local Authority–based social workers and possibly other professionals from various voluntary agencies including the NSPCC, Moral Welfare, etc.

In part this was because of the impact of poverty, upon families, which even today has a deleterious impact on health and education (Pritchard and Williams, 2011; Pritchard et al., 2014), resulting in families occupying different life pathways. Thus, a mother (who might herself have been in child care) with poor parenting skills, giving rise to concerns about her children, may have had a partner, who could have been a previous or current offender in touch with the Police, Probation, and the Benefits Office. There could have been rent arrears and involvement with the Housing Department; the Education Welfare Officer (EWO) would have been concerned by the children's truancy; and the mother might have been referred to the mental health team for treatment of depression, by her GP (general practitioner). In this situation there was repetition and duplication of services; confusion and fragmentation occurred as different agencies focused only on their particular part of the problem. This was resolved following the *Seebohm Report* of 1968, leading to the creation of the integrated Local Authority Social Services Department, designed as a generic 'one-stop' agency, to bring all services together. However, Probation stayed separate under the Home Office.

Specialist social workers like PSWs (psychiatric social workers) quickly realized that Seebohm's message of a generic 'one-stop' inclusive agency was misunderstood. Thus specialist social workers, especially those offering mental health services, were discouraged by the local authorities, because there was concern that this new service would be independent of medical control (Pritchard, 2006).

These tensions coincided with the anti-psychiatry movement of the 1960s and 1970s, led by British and American critics epitomized by RD Laing and Thomas Szasz. Laing postulated that mental illness was a rational response to the modern world and parental oppression, which suppressed the person's 'real self.' He argued strongly against medicalization and overdependency on psychopharmacology, which led to neglect of the psychosocial dimensions.

Szasz was more philosophical, arguing that in order to differentiate between mental and physical illness and treat some behavioral/mental disorders "as if they were a physical illness," biological evidence should be required; he believed that "mental illness was a myth." Paradoxically, there is now a greater understanding of the constant bio-psycho-social interaction in the etiology of conditions recognized as 'psychiatric.' Descartes' division of mind and body is false; modern technology, (such as the brain scan) shows that there is a constant feedback loop and interactions between biological, psychological, and social stimuli (Baron-Cohen, 2011; Buckingham and Daniolos, 2013; Koss et al., 2014; Palmer et al., in press).

One dilemma in Western psychiatry is that some wish to be accepted by mainstream medicine while others are more psychosocially orientated. However, pressure upon resources makes it easier to go for 'the right pill.' This is a big mistake; for example, when patient-related outcome measures were integrated into 'high tech' neurosurgery and treatment plans, there were measurable fiscal and psychosocial benefits (Pritchard et al., 2011). Similar results have been found for mental health (Kingdon and Turkington, 2000; Pritchard, 2006).

The public's perception of mental disorder has always been ambivalent at best. But the 'anti-psychiatry' movement was given a boost by the film *One Flew Over the Cuckoo's Nest*, which portrayed psychiatric care in hospitals as controlling via Orwellian misuse of language concerned only with control. The public's view combined with a political perspective led to the more legalistic Mental Health Act of 1983, in which the Psychiatric Social Worker (PSW) in adult psychiatry became the 'Approved Social Worker' (ASW), whose role was to be a potential brake and safeguard for patients and where possible, to minimize compulsory admissions. In addition, Community Psychiatric Nurses (CPNs) were brought into play; they were NHS-based, whereas ASWs were affiliated to the Local Authority. Over the next 20 years, there was a subtle shift

between the two roles. ASWs became more 'legalistic' and less supportive in the community and their role increasingly was taken up by the CPNs.

It is reasonable to assert that often public health policy is determined by a high profile case. For example in child protection, the tragedies of Victoria Climbe, (2001), Baby Peter Connelly (2007), and Daniel Pelka (2013), give the impression of a totally failed child protection service despite empirical evidence that child abuse-related deaths have never been lower in the UK since records began (Pritchard et al., 2013a). But the concept "naught is good or ill but as thinking makes it so" (Shakespeare, 1978: Hamlet, Act 2, lines 248–249) can be transposed in a modern world to "naught is good or ill but as the media makes it so"! The equivalent high profile case for mental health was Christopher Clunis, suffering from paranoid schizophrenia, who stabbed to death a total stranger, Jonathan Zito, on the London Underground in 1992. This triggered a wide media response, fueled the latent public prejudice about mentally ill people, and urged greater control of the mentally ill in the community. The media gave wide coverage to Mrs Zito who founded the charity The Zito Trust, campaigning for closer control and treatment of mentally ill patients in the community. This led to the 2007 Mental Health Act, which introduced 'Community Treatment Orders' (CTOs), which in effect compels mentally ill people in the community to take their medication, which Mr Clunis was not doing. Since this Act, the Zito Trust has been wound up, having achieved its objective (Bowcott, 2009). The hope remains of ensuring that the statistically very rare Clunis-Zito tragedy will happen even less frequently. This is not denying that violence occurs disproportionately more often with those who have psychiatric disorders. Yet, mentally ill people, (in part because of homelessness) are 4.9 times more likely to be homicide victims than people who do not suffer from this condition (Webb et al., 2013). Furthermore, psychiatric patients are at considerably higher risk of suicide and account for between 50 to 60% of the UK's annual average of 4200 suicides, i.e., psychiatric patients account for between 2100 and 2800 suicides a year (suicides greatly exceed the tragic toll of 9/11).

Overall, the MHA of 2007 (introducing the CTOs), has led to a more restrictive ethos. While there is value in seeking to ensure that people take necessary medication and improve contact with services, the increase in compulsion and its potential misuse remains an ethical dilemma. Conversely, mentally ill people have their lives shortened not just by self-inflicted violence, but also (in both adults and adolescents) by the accumulation of the side effects of psychopharmacological medication taken over years (e.g., Koelch et al., 2008; Gissler et al., in press).

The author recently worked with a grieving family whose daughter had died from cardiac and respiratory complications, aged 43, after years of medication for schizophrenia. Before her breakdown aged 22, she was tall and weighed less than 57 kilos. However, the latest new drug to better 'manage' her symptoms, increased her weight from 83 to 118 kilos in less than 18 months. That the coroner boldly included on her death certificate that she died of cardiac-respiratory failure to which the psychotropic drug had contributed; is a sign that the problem beginning to be recognized (e.g., Choong et al., 2012; Gissler et al., in press).

I am not advocating an anti-psychopharmacological position. There is real evidence of the lifesaving and enhancing benefits of appropriate, carefully monitored medication. Rather, I am highlighting the need for those working with people with mental disorders to understand the effects of the constant interaction of bio-psycho-social factors and to consider more actively the long-term effects of drugs.

In respect to CTOs there is some evidence from Australian studies showing that patients who have maintained better contact with social services have lowered readmissions or shorter lengths of stay because of earlier admissions (Kisely et al., 2013). However, an earlier Cochrane Review, Kisely et al. (2011) had concluded that patients on CTOs showed no significantly better use of service or improved social functioning. Yet, an important finding was that people on CTOs were less likely to be victims of violent crime (Kisely et al., 2011). More long-term research is required to be confident that the benefits of CTO outweigh the ethical dilemmas of compulsory treatment. The issue is to ascertain for whose benefit the CTOs are administered – patients, family, or society.

The social care component of Britain's multidisciplinary Community Mental Health Teams (CMHTs) has seen major changes. A CMHT aims to support people in the community, but these have twice as many CPNs than social workers, apparently because resources are restricted. The rate of ASWs per 100 000 population has doubled over the past decade (Huxley et al., 2005), which inevitably makes greater demands, especially during an economic recession. This might partially explain the ASWs low morale and sense of being undervalued, with higher levels of work-related stress (Evans et al., 2006). Furthermore, compared with the specialist indepth mental health training previously received by PSWs, current ASWs complain of the need for better training, especially in dealing with more severe cases (Walid and Wright, 2004).

Finding employment is a major issue for those with mental illness wishing to live independently. A recent study examined the use of the Individual Placement and Support (IPS) Model, to find that whilst IPS patients doubled their employment rate, but it still stood at only 22% (Heslin et al., 2011). This is a reminder of the vital importance of the 'social' element in the bio-psycho-social dimension that impacts the lives of vulnerable people.

Psychiatric Social Work in Child and Adolescent Mental Health

One area in which the counseling and supportive roles of PSWs continue in Britain is in Child and Adolescent Mental Health (CAMH) teams. These teams are usually led by a Child Psychiatrist, with Mental Health Social Workers (MHSWs) and Child Psychologists completing the interdisciplinary team. They work closely with schools and Educational Social Workers to provide more intensive psychosocial support services for children, especially those who have been neglected and abused and are Looked-After-Children (LAC), i.e., children requiring an alternative to family care. There is a long-standing concern about the psychosocial development of LACs, particularly if they were victims of neglect and abuse. Successive Acts have enabled these children to receive ongoing support up to

age 21; for those with special needs, support may continue for a longer time. There is considerable research about the poorer outcome of children leaving care (Stein, 2013), and CAMH teams are very active in supporting such young people (Stanley et al., 2005; Simkiss et al., 2013). Unsurprisingly, the outcomes for former LAC are significantly poorer than children who did not require local authority care. However, when former looked-after young men (aged 16–24 years) were compared with another group of relatively deprived children, who were former Excluded-from-School (EFS), the previous LAC men had considerably better outcomes than the EFS youth, who, unlike those leaving care, had no automatic social work support. It was the unsupported former EFS children who had significantly higher criminality, a murderer rate more than a 1000 times higher, and a suicide rate 100 times higher than their peers in the general populace. There were no murderers or suicides in the social work–supported LAC group, which is another indication that good social work works (Pritchard and Williams, 2009; Pritchard et al., 2014).

Two other big issues currently face CAMH social workers and society: Attention-Deficit-Hyperactivity-Disorder (ADHD) and Autism. Regarding ADHD, I worked in the three child guidance clinics covering the largest county in England, the West Riding of Yorkshire in the 1960s. Annually, there were fewer than two such children per clinic. Space precludes an adequate examination of ADHD, other than to say that it has inevitable disruptive psychosocial consequences, involving many child-related services. An examination of the online research database Medline shows the first paper related to ADHD, (then designated Minimal Cerebral Dysfunction), was in 1964. The hundredth paper was published in 1972. In October 2013, over a 1000 (1073) had been published, totaling about 16 737 papers since 1964. The debate still rages as to whether ADHD is linked to family disruption or underlying organic conditions. Meanwhile, drug companies have been very successful in offering a 'simple' chemical-based cure; they tend to ignore the *causes*, which are complicated by how families respond to their ADHD children (Dalsgaard et al., 2013). There is strong indicative evidence that changes in lifestyles and multiple environmental pollution contribute to ADHD. The related evidence comes from studies on the changing pattern of adult mortality. For example, whilst Western nations have been relatively successful in reducing aged-related adult cancer deaths, this hides the rising incidence of cancer across all age bands (Pritchard and Hickish, 2011). Another strong indication is the alarming increase in neurological morbidity and deaths, with dementia beginning a decade earlier, raising serious etiological questions about multiple environmental pollution, which is being increasingly linked to rises in the etiology of neurological morbidity as well as in ADHD and Autism (Pritchard et al., 2013b).

Could environmental pollution be a part of multifactorial causes of psychosocial problems in children? In 1968, Rachel Carson's seminal *Silent Spring* predicted that there would be a rise in 'irritant diseases.' These along with asthma and other immune disorders, with their associated psychosocial problems, are affecting Western children at an astonishing rate (Pritchard et al., 2011). This may seem unrelated to PSW but the importance of the bio-psycho-social model reveals that disciplinary boundaries need to be crossed to understand current problems and to challenge seemingly easy, short-term 'chemical' solutions.

In relation to Autism, Professor Leo Kanner first described the condition in 1944, suggesting it was an early form of psychosis caused by 'attachment failure' by parents, whom he described as 'refrigerator parents.' This provided an observation that attributed cause to untestable Freudian psychology, proving to be one of the worst misattributions in modern medicine – blaming parents without evidence (Pritchard, 2006). Even in the late 1960s, questions were being asked about Autism's possible neurological etiology. It is now recognized that the Autism spectrum is far more complicated. Broadly, it is a genetic-neuro-biological disorder that has cognitive and communication problems impacting upon the child's psychosocial development, and may be linked to an early manifestation of schizophrenia (e.g., Rutter and Sandberg, 1985; Howlin et al., in press).

The Psychiatric–Child Protection Interface

The third and most important aspect of the modern PSW is to understand the evidence-based concept of the child protection–psychiatric interface. This remains somewhat controversial. Early researchers emphasized that child abuse was an ubiquitous problem, crossing social classes and cultures; they largely ignored the psychiatric dimension, even as late as 1998 (Pritchard, 2004). For example, in the national confidential audit of violence and mental disorder, no mention was made of the age or relationship of the victims to the assailant (Webb et al., 2013), either because it was not initially reported or to avoid further stigmatization of mental illness being linked to child abuse. Yet, at the fatal extremes of abuse, the vast majority of assailants are predominately in a 'parental role' with many suffering from mental disorder. Consequently, with a few exceptions, the element of psychiatric child protection is unacknowledged (Falkov, 1996; Reder and Duncan, 1999), and this undermines effective treatment and child protection.

The psychiatric–child protection interface is far wider than just dealing with the potential extremes of 'abuse' and should focus upon the impact a parental psychiatric disorder has upon the child's development – an approach totally compatible with Britain's progressive *Every Child Matters* agenda. This demands that services relevant to children ensure that the five crucial aims for the child "being healthy, staying safe, enjoying and achieving, making a positive contribution, and economic well-being" are met (DCSF, 2005: p. 3). A greater awareness of the child protection–psychiatric interface could make the child safer and reduce the innate threat of removing a child just because there is a psychiatrically ill parent (Seeman, 2014). A better case-specific analysis could be made by recognizing potential risks – one psychiatric ill mother in 3000 goes on to kill her child (Pritchard et al., 2014).

Manfred Bleuler influenced my practice when he spoke of 'Children of Sick Parents.' I saw that older and teenaged children often had to 'care' for mentally ill parents, a fact that has been ignored until fairly recently. Fortunately, recent research is recognizing the children's need for help to cope with the "burden of mental disorder." Otherwise, without preventative services, they themselves are more likely to develop problems

as adults (Jensen et al., 2014; Seltzer et al., 2013; Viana et al., 2013).

A further element in the psychiatric–child protection interface is the link between child neglect and disproportionate levels of subsequent mental disorder as adults; neglecting parents themselves are often victims of an intergenerational cycle of neglect (Martorell and Zongrone, 2012; Buckingham and Daniolos, 2013; Sellers et al., 2013).

Social workers seeking to avoid being 'judgemental' seldom use a psychiatric diagnosis, in particular the concept of 'personality disorder,' which includes the judgemental sounding labels – psychopath or sociopath. Personality disorder is a useful concept because it highlights 'the persistent and characteristic' behavior of the parent who crucially does not appear able to care, lacking the core need to be altruistic and empathetic (Baron-Cohen, 2011; Brandon et al., 2012; Buckingham and Daniolos, 2013). Children with harsh parental behavioral patterns are at higher risk of becoming 'personality disordered,' perhaps best summed up by Macbeth's 2nd Murderer, "I am one whom the buffets and blows of this vile world hath so incensed I am reckless what I do to spite the world" (Shakespeare, 1978: Macbeth, Act 3, lines 106–109).

It is selfe-vident that we must not forget the vital importance of socioeconomic factors, as there is an overlap between relative poverty and psychopathology found throughout the Western world. Child poverty and poorer educational, health, and criminal outcomes, which are worse amongst children of poorer parents, compound the psychiatric situation (Palmer et al., in press; Pritchard et al., 2014).

Extremes of the Psychiatric–Child Protection Interface

After rare but tragic Child-Abuse-Related-Deaths (CARDs), there is massive negative media coverage, which raises the question, 'Who kills children?' In a study of a decade of CARD assailants who killed children from a 4% sample of the UK population, the initial work had sought to avoid further stigmatization of the mentally ill by focusing upon the differential levels of risk of the assailants, rather than on the frequency of those who killed the children (Pritchard, 2004). The latter focus showed that the majority – 12 of 18 (67%) – of assailants were within-family, and most had psychiatric problems (Pritchard et al., 2013). The study also revealed that all the blood fathers (4) committed suicide after the murder, as did two mothers; two other mothers had severe suicidal behaviors, quite apart from other mental health problems. The highest risk came from men with previous convictions for violence, both 'within-family' (4) and 'extra-family' (5), whose convictions heuristically classified them as Violent-Multi-Criminal-Child-Sex-Abusers. Their persistent, characteristic behavior over the years warranted the attribution of the 'severe personality disordered' tag (Pritchard et al., 2013).

Synthesis

Psychiatric social work is 'generic' in that people's mental health problems span all agencies and human behavior is a constant bio-psycho-social interaction. Edward Wilson has begun to integrate biology with the social sciences (Wilson, 2013) and has postulated, with evidence, the concept of a 'social gene.' This growing synthesis between what initially appeared to be contradictory theories is seen in Simon Baron-Cohen's seminal work (2011) that provides a superb evidence-based explanation of severe personality disorders demonstrating a bio-psychosocial etiology for Bowlby's Attachment Theory and explaining the absence of empathy in abusers.

Recent and current research highlights the need for an integrated bio-psycho-social understanding of, and preventative intervention for, people with mental health problems; and a better comprehension of the pediatric–child protection–psychiatric interface. Perhaps for the first time, at the national child protection level in UK, as mental health problems figure large in Serious Case Reviews, this message is beginning to be recognized (Brandon et al., 2012). Moreover, neuroscience research linking biological measurements, such as cortisol levels, with a range of environmental factors, (PTSD, maternal and socioeconomic disadvantage) has exploded in the past 10 years (e.g., Ehlert, in press; Koss et al., 2014; Palmer et al., in press), providing a biological understanding of problems, as well as recognition of the vital importance of child–parent attachment and constant bio-psycho-social feedback. The practice and policy implications appear self-evident:

1. Social work and the social sciences should rid themselves of the archaic notion of dualisms, such as mind versus body, nature versus nurture, dualisms.
2. Integrate the bio-psycho-social development model into human services as a range of preventative potential interventions. This could include biological measurements, such as cortisol levels, to measure stress reductions (Koss et al., 2014; Palmer et al., in press) that would improve psychosocial risk assessments and chart progress.
3. The policy implications are that pediatricians should overcome their recent reluctance to engage in potentially conflictual relationships with parents; accept their collaborative child protection role alongside children's services, education, and police; and recognize the interface with psychiatry.

Conclusion

In hindsight, while examining the case files of a decade of child homicides (Pritchard, 2004), if the adult psychiatrist had thought 'children – what is happening to them,' and if the child protection social workers had thought 'mental health' – then half the children who died would have been alive. This thinking should underpin the role of the modern mental health social worker, the professonal heir to the PSW. Encouraged by UK child-abuse-related-deaths being lower than ever (Pritchard et al., 2014), child-related disciplines could work even better together and be more preventative, not only in reducing child neglect, but in ameliorating the impact of parental mental disorder on children. By recognizing the importance and potential of the pediatric–child protection–psychiatric interface, within a family's socioeconomic context professionals can respond to the long-standing needs of 'children of sick parents.'

See also: Child Protection; Mental Health; Social Work Theory.

Bibliography

Baron-Cohen, S., 2011. Zero Degrees of Empathy. Allan Lane, London.

Bowcott, O., 2009. Jane Zito: Why Its Time to End the Campaign. Observer, 17th May, 28.

Brown, R., 2009. The Approved Mental Health Professional's Guide to Mental Health Law, second ed. Learning Matters, London.

Brandon, M., Sidebotham, P., Bailey, S., 2012. New Learning from G Serious Case Reviews: A Two-year Report for 2009-1. Department of Education, London.

Buckingham, E., Daniolos, P., 2013. Longitudinal outcomes of victims of child abuse. Current Psychiatric Reports 15, 342–345.

Choong, E., Bondolfi, G., Etter, E., et al., 2012. Psychotropic drug induced weight gains and other metabolic complications in a Swiss psychiatric population. Journal Psychiatric Research 46, 540–546.

Dalsgaard, S., Nielsen, H., Simonsen, M., 2013. Five-fold increase in national prevalence rates of ADHA medications for children, and adolescents with autism spectrum disorder. Journal Child Adolescence Psychopharmacology 23, 432–439.

Department of Children, Schools & Families, 2003. Every Child Matters. Cm 5860. HMSO, London.

Ehlert, U. Enduring psychobiological effects of childhood adversity. PMID 38, 1850–1857.

Evans, S., Huxley, P., Webber, M., 2006. Mental health, burnout, job satisfaction among mental health social workers in England & Wales. British Journal Psychiatry 188, 75–80.

Evens, S., Huxley, P., Baker, C., 2012. The social care component in multi-disciplinary mental health teams: a review and national survey. British Medical Journal 344, 1776.

Falkov, A., 1996. Study of Working Together: Part 8 Reports of Fatal Abuse and Parental Psychiatric Disorder. An Analysis of a 100 Area Child Protection Committees. HMSO, London.

Gissler, M., Munk, L.O., Obsy, U., et al., Patterns in mortality among people with severe mental disorders across birth cohorts: a register based study of Denmark and Finland 1982–2006. Public Health 13 (11), 834–836.

Heslin, M., Howham, L., Leese, M., McCrone, P., Huxley, P., Thornicroft, P., 2011. World Psychiatry 10, 132–137.

Howlin, P., Savage, S., Moss, P. Cognitive and language skills in adults with autism: a 40-year follow-up. Journal of Child Psychology Psychiatry 55, 49–58.

Huxley, P., Evens, S., Webber, M., et al., 2005. Staff shortages in the mental health workforce: the case of the disappearing approved social worker. Health Social Care Community 13, 504–513.

Jensen, S., Dumontheil, I., Barker, E., 2014 Development inter-relations between early maternal depression, contextual risks and interpersonal stress and their effect on later chid cognitive functioning. Depression Anxiety. http://dx.doi.org/10.1002/da.22147.

Jones, R., Everson-Hock, E., Papaioannou, D., 2011. Factors associated with outcomes for looked-after-children and young people: a correlates review of the literature. Child Care Health 37, 613–622.

Kingdon, D., Turkington, D., 2000. Cognitive behaviour therapy for general psychiatrists in the management of patients with psychoses. British Journal Psychiatry 177, 101–106.

Kisely, S., Campbell, L., Preston, N., 2011. Compulsory community and involuntary outpatient treatment for people with severe mental disorders. Chochrane Database Systematic Review. CDO004408.

Kisely, S., Preston, N., Jianquo, X., 2013. Reducing all-cause mortality among patients with psychiatric disorders: a population based study (CPO). Canadian Medical Association Journal 185, 50–56.

Koelch, M., Schnoor, K., Fegert, J.M., 2008. Ethical issues in psychopharmacology use in children and adolescents. Current Opinion Psychiatry 21, 598–605.

Koss, K.J., George, M.R., Cummings, E., 2014. Asymmetry in children' salivary cortisol and alpha analysis in the context of marital conflict: Links to children's emotional security and adjustment. Developmental Psychobiology 56, 836–849.

Martorell, R., Zongrone, A., 2012. Intergenerational influences on child growth and under-nutrition. Paediatric Perinatal Epidemiology 26, 302–314.

Milton, J., 1674. Paradise Lost. Bk 11, second ed. S. Simmons, London.

Palmer, F., Anan, K., Graff, J., et al., Early adversity, socio-economic development and stress in 1 year old urban children. Journal Paediatrics 163, 1733–1739, 3476.

Pritchard, C., 2004. The Child Abusers: Research & Controversy. Open University Press, Buckingham.

Pritchard, C., 2006. Mental Health Social Work: Evidence Based Practice. Routledge, Abingdon & New York.

Pritchard, C., Hickish, T., 2011. Comparing cancer mortality rates in England & Wales with other major developed countries 1979–2006. British Journal Cancer 105, 1788–1994.

Pritchard, C., Williams, R., 2009. Does social work make a difference? A controlled study of former looked-after-children and excluded-from-school adolescents now men aged 16–24: subsequent offences, being victims of crime and suicide. Journal Social Work 9, 285–307.

Pritchard, C., Williams, R., 2011. Poverty and child m(0-14) mortality in the USA and other Western countries as an indicator of "how well a nation meets the needs of its children" (UNICEF). International Journal of Adolescent Health and Medicine 23, 251–255.

Pritchard, C., Davey, J., Williams, R., 2013a. Who kill children? Re-examining the evidence. British Journal of Social Work, 1–36. Advanced access May 3rd.

Pritchard, C., Mayers, A., Baldwin, D., 2013b. Changing patterns of neurological mortality in the 10 major developed countries 1979–2010. Public Health 127, 357–368.

Pritchard, C., Williams, R., Wallace, M., 2014. Child mortality and poverty in the western nations 1980–2010: are English-Speaking-Countries children disadvantaged? Childhood. http://dx.doi.org/10.1177/09075682135/35/3308.

Reder, P., Duncan, S., 1999. Lost Innocence – Follow up Study of Fatal Child Abuse. Routledge, London.

Rutter, M., 2013. Changing concepts and findings on autism. Journal Autism Developmental Disorders 43, 1749–1757.

Rutter, M., Sandberg, S., 1985. Epidemiology of child psychiatric disorder: methodological issues and some substantive findings. Child Psychiatry Human Development 15, 209–233.

Seeman, M., 2014. Intervention to prevent child custody loss in mothers with schizophrenia. Schizophrenia Research & Treatment. http://dx.doi.org/10.1155/20212/796793.

Sellers, R., Collishaw, S., Rice, F., et al., 2013. Risk of psychopathology in adolescent offspring of mothers with psychopathology and recurrent depression. British Journal Psychiatry 202, 108–114.

Seltzer, L., Ziegler, T., Connelly, M., et al., 2013. Stress induced elevation of oxytocin in maltreated children: evolution, neurodevelopment and social behaviour. Psychoneuro-endocrinology 38, 1850–1857.

Shakespeare, W., 1978. The Complete Plays. Oxford University Press (original work published year, 1978).

Simkiss, D., Stallard, N., Thorogood, M., 2013. A systematic literature review of the risk factors associated with children entering public care. Child Care Health 39, 628–642.

Stanley, N., Rirdan, D., Alaszewski, H., 2005. The mental health of Looked-after-children: matching response to need. Health Social Care Community 13, 239–248.

Stein, M., 2013. Young People Leaving Care. Jessica Kinglsey Publishers, London.

Titmuss, R.M., 1965. Social Policy. George Allen Unwin, London.

Viana, M., Gruber, M., Shaly, V., et al., 2013. Family burden related to mental health and physical disorders in the world: results from WHO mental health surveys. Review Brazil Psychiatry 35, 115–125.

Walid, K., Wright, D., 2004. A national survey of approved social workers in the UK: information, communication and training needs. British Journal Social Work 34, 663–675.

Webb, R., Shaw, J., Appleby, L., 2013. Risk of people with mental illness dying by homicide. British Medical Journal 346, 1336–1339.

Wilson, E., 2013. The Social Conquest of Earth. Liverright Publishing Corporation, New York.

Psychiatry and Religion

Kate M Loewenthal, Royal Holloway, University of London, Egham, UK; and Glyndwr University, Wales, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.T. Harding IV, volume 19, pp. 13096–13102, © 2001, Elsevier Ltd.

Abstract

This article examines the changing relations between psychiatry and religion. It examines views of the causes and cures of mental illness in some religious traditions. Some psychiatric views of religion and its place in mental illness are described. Current psychiatric practice is examined, including areas such as ascertaining the patient's religious history, understanding cultural variations in religious behavior, and religious barriers to seeking psychiatric help. There is a brief overview of attempts to alleviate problems, including an outline of steps taken by professional organizations to accommodate the issues raised for psychiatrists by religious factors.

The relationship between psychiatry and religion has been difficult and variable, and has developed over the two centuries of psychiatry's official existence.

Psychiatry is the branch of medicine dealing with mental, emotional, or behavioral disorders (Merriam-Webster, 2013). For present purposes, it is significant that 'psych' is a stem of Greek origin often translated as 'soul' ('-iatry' refers to medical treatment). The term was first used in the early nineteenth century: this timing is significant since it reflects a developing medical interest in insanity, for which earlier there had been little or no treatment other than confinement under more or less (usually less) humane conditions. Current psychiatric views on the causes and cures of psychiatric disorder involve a huge range of detailed suggestions involving biological, social, and psychological levels of analysis.

Religion has many definitions and methods of assessment (e.g., Brown, 1987; Hill and Hood, 1999; Loewenthal, 2000). Major religions agree on the existence of nonmaterial (i.e., spiritual) reality, and see the purpose of life as increasing harmony in the world by doing good and avoiding evil. Monotheistic religions hold that the source of existence (i.e., G-d) is the source of moral directives. Religious traditions offer a range of views on the causes and cures of psychiatric disorders, and these often appear sharply different from the views offered in psychiatry, focusing chiefly on spiritual factors.

How have psychiatry and religion related to each other?

There has been a generally fraught history. Foskett (1996) describes it as a history of discord. This article will first examine how insanity and its causes and cures have been seen in selected religious traditions, and will then look at how religion has been seen through some psychiatric eyes. Finally, the current state of the relations between psychiatry and religion will be viewed, along with an overview of attempts to alleviate some of the problems.

Beliefs About Mental Illness in Different Religious Traditions

Psychiatry developed in the Western world at a time when Christianity was the dominant religion in Europe, the United States, and numerous colonial countries. Twentieth-century social and other changes have increased the numbers of Muslims, Jews, Hindus, Buddhists, and members of other (including new) religious movements in the Western countries. Religions have many different forms or denominations, with considerable variations in beliefs and practices. All have been and are in a constant process of development and change.

Christianity

The doctrine of the trinity (Father, Son, and Holy Spirit) is central to Christianity, as is the idea that the death of Jesus atoned for the sins of humanity. Important for mental health are Christian dogmas regarding sin. Suffering is the result of sin. Salvation involves justification, the removal of sin and its effects by one or more of repentance, penance, indulgence, confession, absolution, and forgiveness (e.g., Stagg, 1962). These doctrines may be associated with pejorative views of sufferers from illness and misfortune, including sufferers from mental illness (Lerner, 1980). It has been reported that Christians suffering from mental illness may be made to feel that their illness and suffering are the result of their sins (Stanford, 2007). However, although suffering is not seen as a desirable end in itself, it can be seen more positively as a gateway to renewal and rebirth. Some historians of psychiatry have suggested that a widespread Christian view of mental illness, particularly during the fifteenth to seventeenth centuries, is that it is the result of demonic possession (Lipsedge, 1996).

Islam

In Arabic, Islam means submission to the will of G-d. There is a clear core of religious duties, including belief in G-d and the prophets, prayer, giving away a proportion of one's goods, fasting in the month of Ramadan, and pilgrimage. In Islam, sin results from pride and self-sufficiency, leading to forgetfulness of divine unity and wrong moral/religious choices. An important Islamic view of the mentally ill is that they are the 'afflicted of Allah.' Some contemporary observers have noted that Jinn possession may be seen among Muslims as a cause of insanity (Dein et al., 2008). It is noteworthy that the earliest recorded psychiatric institutions – established over a 1000 years ago – were in Moslem countries (Loewenthal, 1994).

Judaism

Judaism is the oldest of the monotheistic religions, and a central tenet is the belief in the unity of G-d. There are many religious commandments, governing more or less the total lifestyle and practiced to varying extents: diet, sexual behavior, work, business ethics, and worship are among the areas of religious law. Madness is mentioned in the Jewish biblical texts without clear attribution of its causes, though there is some suggestion of a distinction between spirit possession and insanity. Later, Jewish sources of the Talmudic period distinguish a range of psychiatric conditions resembling those found in contemporary practice (Miller, 1972). The legal status of the insane with regard to their civil and religious obligations is a matter of discussion in works on Jewish law: for example, whether the insane are valid witnesses (probably not) and whether insanity is grounds for divorce (probably yes, but the situation is complicated). The Hebrew term for madness *(choli nefesh*) means sickness of the soul, and there are rabbinic figures in Jewish communities who carry out counseling and therapeutic activities based on views that the rectification of spiritual–moral failings – such as pride – will improve mental health. Misfortune is seen as warning to the individual to improve and as a divine test of the individual, also as part of an overall divine plan in which everything is for the ultimate good.

Hinduism

Hinduism is the religion of India and has many varied manifestations, tolerating a wide range of beliefs and practices. There is an underlying monotheism, with lesser divinities as aspects of one G-d. This infinite principle is truly the sole reality and the ultimate cause and goal. There is a religiously sanctioned caste system, said to be in some decline. Religious worship is carried out in a shrine in the home, usually by women. Transmigration of souls and reincarnation are important aspects of Hindu belief. The ultimate goal is infinity (G-d), and the attainment of this goal is prevented by *karma* (rebirth); following death and a sojourn in heaven or hell, the soul is reborn into a physical form determined by actions in the previous incarnation. This process of rebirth (*samsara*) is seen as potentially endless, and not progressive in any way. Misfortunes are an aspect of *karma*, and can be escaped by *marga* – emancipation – of which the main types are duty, knowledge, and devotion. Bhugra (1996) has described the importance of Ayurvedic medicine in the treatment of mental illness in India, and in one study, nearly half the psychiatric patients (81% Hindu) questioned had consulted a religious healer (Campion and Bhugra, 1997). Craissati (1990) states that mental illness in India carries a major stigma and that consulting a psychiatrist and going to the hospital are regarded as last and desperate resorts. In rural communities, where most of India's population (80%) lives, tolerance for bizarre behavior is very high. Psychotic symptoms are seen as the result of spirit possession or black magic, or as the consequences of a previous life. A mentally ill person will be taken to a healer within the community. If this is not successful, it is regarded as the family's duty to bear with the person. Craissati says that patients are only brought to the 'Western' psychiatric hospital where there is serious threat to the norms governing family life and sexual behavior.

Buddhism

The fundamental teaching of Buddhism involves viewing an attachment to the world and its pleasures as the cause of pain. Self-mortification is also an extreme to be avoided; the founder of Buddhism, Gautama Buddha taught the 'middle path.' Life is fundamentally a process of suffering. As in Hinduism, transmigration and rebirth are not seen as progressive, and the central aim of religious belief and practice is to liberate. The eight-fold path to freedom from suffering includes right thought, speech, action, and mental attitudes. These lead to the cessation of pain, and to enlightenment, and Nirvana (wherein the soul will not be reborn to further suffering). Buddhist psychology is popular in the West (Valentine, 1989). It is suggested that religious practices and thoughts will promote the relief of many if not all forms of mental illness and forms of meditation have been incorporated into contemporary cognitive behavior therapy (Segal et al., 2002).

Overall Religious Views on Mental Illness

This summary of views from selected religious traditions suggests that some similarities across religions can be seen. Beliefs about mental illness and its treatment may be tied to beliefs about sin and suffering, though malign spiritual beings may also be implicated. Mental illnesses are seen to result from separation from the divine or possession by evil. Psychiatry may be mistrusted, except as a last resort, and 'folk' healers may employ a wide range of symbolically loaded rituals and cognitive behavioral techniques designed to facilitate spiritual healing, leading (hopefully) to an improvement in mental health.

Some empirical work indicates that the views of religiously active people on psychiatric disorder will reflect religiously taught views – although in practice most of those questioned have a foot in both cultures, supporting both contemporary views on stress and biological factors, while simultaneously acknowledging the importance of spiritual factors (Cinnirella and Loewenthal, 1999; Leavey et al., 2007; Dura-Vila et al., 2011; Bayes and Loewenthal, 2013). Sin, however, is not a popular explanation among religious laity. In the Western world, the older generation of minority group immigrants may see orthodox health services and psychiatry as having little to offer. For example:

> … in the older generation if you've got a mental illness, they think you've got a bad spirit in you (and) that's why you hear all these voices and that's why this person (voice) is telling you to do things you've been doing … they'll just think, 'oh, go to a holy person and they will cure you.'
>
> Hindu informant, quoted in Cinnirella and Loewenthal, 1999.

Beliefs About Religion among Psychiatrists

Attitudes to psychiatric disorder change with time. In Western Europe and the United States, for example, one view of

disorder up until the eighteenth century was that it was caused by demonic possession, perhaps a result of the preoccupation with witchcraft from the fifteenth century onward, although stress was often seen as an important factor (Lipsedge, 1996). Fear, neglect, isolation, and ill treatment of the mentally ill were socially and religiously sanctioned. Bedlam in London and the Bicetre in Paris chained up their unruly inmates. The aim of such institutions was simply to isolate the unruly from society. Sometimes, the insane would be auctioned off to be 'cared for' (used as slaves) by farmers (Loewenthal, 1994). By the nineteenth century, medical opinion calling for more humane treatments was growing. Tuke, a Quaker merchant, founded the York retreat at the end of the eighteenth century. Here, prayer and religious devotion were seen as central to healing. Pinel and Connolly pushed the movement to see the insane as individuals suffering from sickness, requiring treatment. Pinel abolished chains from the Bicetre, and in the 1840s, Dorothea Dix began a campaign to improve conditions in institutions for the insane in the United States. Religious practice was often (but not always) acknowledged as an important feature of psychiatric care. In Britain the Lunacy Act of 1890 ordered a church in every asylum; inmates had to attend twice a day. In France, by contrast, Pinel would not allow the mentally ill to be exposed to religious practices: these might encourage delusions and hallucinations. Psychiatric hospitals in most Western countries nowadays normally maintain a chaplaincy service. This does not always cater for minority religions; chaplains are sometimes unpaid, and are normally overloaded.

Major changes in psychiatric thinking include the introduction of psychoanalysis in the early twentieth century, the anti-medical-model movement, the introduction of cognitive behavioral and other 'talk' therapies in the later part of the century, as well as the enormous growth in the neuroscientific and biochemical understanding of psychiatric disorder.

Psychoanalysis, spearheaded by Freud, carried as part of its baggage the general idea that religion was unnecessary and damaging to mental health. Four of Freud's books (1927, 1928, 1930, 1939) analyzed religion, reflecting and helping to form a growing *zeitgeist* that viewed religion as a collection of primitive superstitions, neurotic rituals, and an illusion, comforting at one time, but no longer necessary in these enlightened times. Freud described religion as a 'universal obsessional neurosis,' which has succeeded because it spares the individual the labor of developing his or her own neurosis. Freud described G-d as a projection of the image of the father, and "a system of wishful illusions together with a disavowal of reality, such as we find nowhere else but in amentia" (Freud, 1907, 1927).

Although Bettelheim (1983) attempted to see Freud's ventures as essentially spiritual in nature, and it is certainly true that religion was a major focus of Freud's attentions, Freud's writing has epitomized negative twentieth-century views on the role of religion in mental health. Other twentieth-century writers have contributed to this *zeitgeist*, perhaps most notably Albert Ellis in his earlier writings (e.g., Ellis, 1958). He was an outspoken participant in public debate, arguing that the influence of religion was to increase guilt and anxiety over wrongdoing: religion harmed mental well-being. His views on religion became more benign in later life (Ellis, 2004).

In contrast to Freud, Jung (1958) saw the psychological quest for wholeness and healing as a journey that is essentially spiritual. Jung's views were developed by others during the twentieth century, and were welcomed by those responsible for training clergy in pastoral work.

Contrary to Freud's and Ellis's claims, thousands of pieces of empirical work have by now shown that religiosity is normally associated with better mental health (Koenig et al., 2012). Nevertheless, there are still some lacunae: for example, the suggestion that religious ritual is compelled by anxiety has meant that very limited attention has as yet been given to the positive experiences and mental health benefits of the many religious rituals engaged in (Dein and Loewenthal, 2013).

The existence of a chaplaincy service can be a compelling reason for psychiatrists to feel that religious issues are for the chaplains to deal with: religious issues are outside the psychiatrist's competence and so are ignored (Neeleman and Persaud, 1995). Nevertheless, possibly related to the rise in popularity of the concept of spirituality, alongside the growing volume of work suggesting the importance of religious and spiritual factors in mental health, psychiatrists have taken a growing interest in religious and spiritual issues (Bhugra, 1996; Dein et al., 2011). Spirituality has been suggested as a possible contemporary alternative to religion (Heelas and Woodhead, 2005; King and Leavey, 2010; Dein et al., 2012), and contemporary psychiatric and social-scientific writing on mental health often feature discussions on its relations with spirituality as well as or instead of its relations with religion (e.g., Cook et al., 2009; Lewis et al., 2011; Paloutzian and Park, 2013; Pargament et al., 2013).

Views and practices about the place of religion in psychiatric theory practice have changed over the nineteenth and twentieth centuries. Generally, there have been growing numbers of attempts to integrate the role of religion into the understanding of the patient's psychiatric history, and to acknowledge the role of religion in treatment planning. The authors turn now to look at some specific aspects of contemporary practice.

Religion in Contemporary Psychiatric Practice

Some early signs of serious psychiatric interest in the role of religion in psychiatric care included conferences and publications dating from the 1990s (e.g., Bhugra, 1996; Loewenthal, 1994, Littlewood and Lipsedge, 1997; Greenberg and Witztum, 2001; Koenig et al., 2001). Some of this work was culturally limited (to practice in Christian-based Western cultures), but some emphasized the importance of acknowledging the importance of cultural factors and variations in the roles played by religion in different cultural religious groups.

There have been a growing number of claims that knowing the patient's religious history is important in psychiatric practice. Empirical work has suggested that many patients would welcome inquiries from mental health professionals about the place of religious and spiritual issues in their psychological battles, and there are useful published guidelines about how this may be done (Pargament, 2007; Koenig,

2008; Culliford and Eagger, 2009). Some of these suggestions have proved controversial. Notably, Koenig (2008) suggested that psychiatrists should take a spiritual history, support healthy religious beliefs and challenge unhealthy beliefs, pray with patients (in 'highly selected cases'), and work in liaison with trained clergy. The suggestion that psychiatrists are able to pass judgment on the healthiness of different religious beliefs and on the appropriateness of praying with their patients has aroused significant concerns (Cook et al., 2011; Poole and Higgo, 2011).

Barriers to seeking professional help for psychiatric problems are of major concern. These barriers often involve religious factors. Delay in help seeking can involve significant worsening of psychiatric illness.

These barriers chiefly involve fears of misjudgment, concerns about religiously inappropriate treatment, and fears of stigmatization.

Psychiatrists are less likely to be religiously active than their clients and this can heighten clients' mistrust (Bergin and Jensen, 1990; Cook, 2011), involving fears of professional misjudgment of religious behaviors, of which there are many examples (Loewenthal, 2007). Religious practices may give the appearance of bizarre behavior, which may be interpreted as mad, unreasonable, neurotic, selfish, or heartless. For example:

> The concept of a 'good death' is very important to Hindus. A good death involves making spiritual preparation for the soul leaving the body, and settling ones worldly affairs. One of the religious customs which many Hindus feel should be observed is that when death occurs, the person should be lying on the ground, preferably in a certain orientation with respect to the north. This has led to many difficulties with Hindu patients in British hospitals. Many will make efforts to be at home if and when they feel the end is approaching. When the dying person is in hospital, he or she may manage to climb or roll out of their bed in order to be on the floor for the moment of death. This obviously requires tremendous effort, but it is often wasted, especially if there are no relatives or friends to explain the meaning of the behaviour. The horrified nurses rush over to replace the nearly moribund patient safely back in bed – much to his/her distress. Their sufferings are increased when their pleas and protestations are not effective, and every time they manage to assume the religiously proper position on the floor, they are firmly bundled back onto the bed. The first time the patient appeared on the floor, the nurses might have thought it was a mishap of some kind, but after the third or fourth time, they thinks/he is deluded, mad, or deliberately being a nuisance.
>
> Firth, 1997

The fear of misjudgment constitutes a barrier to the seeking of professional help for mental health difficulties. Judgment biases have sometimes been demonstrated empirically. For example Yossifova and Loewenthal's (1999) vignette study involved clinical judgments of case histories identical in every detail except that some patients were described as religiously active, while others were not. Religious people were more likely to be seen as suffering from obsessive illness than nonreligious people with the same symptoms.

A further barrier is the fear that treatment recommendations may be seen as religiously inappropriate. Some religious groups will avoid psychotherapy if there is a risk of physical contact between the sexes (Loewenthal, 2006). Those from collectivist religious cultural groups may feel that professional suggestions to alleviate distress and psychiatric symptoms by leaving abusive relationships may be rejected on the grounds that family and community solidarity is more important and religiously valuable than individual well-being (Gilbert et al., 2004; Loewenthal, 2013).

An associated, stronger barrier is the extent to which mental illness is stigmatized, and this effect may be particularly marked in tightly knit social groups, of which religious groups are an outstanding examples:

- Our people do not go to the doctor (when depressed), in fact they hide it, because they think that if people know about it they will not accept them and they'll be laughed at and would be completely shut off because there is this prejudice (Muslim, in Cinnirella and Loewenthal, 1999).
- The one thing Black people hate is for anybody to find out there is any form of mental illness in their families … what they try to do is shut that person away and deal with it by themselves as opposed to going through all the networks and being exposed. (Black Christian, in Cinnirella and Loewenthal).
- I wonder what type of families need this (kind of help)? Is it just those who can't cope? I might feel ashamed to ask for such help (Orthodox Jewish, in Loewenthal and Brooke-Rogers, 2004).

Examples of recent work in the field give further indications of some of the issues preoccupying investigators. Regan et al. (2013) systematically reviewed the literature showing how religion influences pathways into care for dementia patients in minority ethnic groups. The main conclusions were that religion hinders access to the traditional care for dementia, which may increase the burden on relatives and other carers. For example an African-American woman, a Pentecostal Christian, filled with the Holy Ghost in a nursing home became excitable and fainted. Her behavior was misinterpreted as pathological and she was unnecessarily restrained. Reports of such incidents lead to fear of misunderstanding, and delay care seeking. However, Regan et al. conclude that religion, as in other areas, assists carers in positive coping.

Another recent study (Dura-Vila et al., 2013) examined the accounts of coping by contemplative nuns who had been sexually abused by priests. Trauma was "transformed into a symbolic religious narrative" in lengthy processes of solitary introspection involving forgiveness, sacrifice, and salvation, which reshaped their core beliefs and allowed the abused women to regain their sense of spiritual well-being: for example, the postabuse contemplation eventually led to an increase in the sense of spiritual purity. This work contributes to the growing literature on posttraumatic spiritual growth, and on positive psychology, important elements in the psychiatry–religion interface.

Improving Practice

From within religious groups, clients see therapists from their own religious group as more likely to understand them.

However, they can be worried that this makes it more likely that their problems will become known in the group (Cinnirella and Loewenthal, 1999; Loewenthal and Rogers, 2004). One solution is the provision of own-group help in discreet locations, in which the client feels less concern about being seen by people they know. Further, clients may select own-group helpers belonging to a social subgroup different from their own. Attempts have been made to train own-group professionals, particularly in counseling skills – clients may still have concerns about the professionalism and confidentiality, but on the whole such training schemes have worked well and have filled a gap in service provision (Loewenthal and Rogers, 2004).

Attempts to improve awareness of mental health issues, and of ways to manage them, include dissemination of information via various media including the Internet (for example, the American Psychiatric Association Web site (2013) offers a wide range of mental health information for the lay public, as does the UK Royal College of Psychiatrists Web site (2013)), offering 'Mental health information for all.' These attempts may help to reduce the strength of some religiously related barriers to professional help seeking.

Professional organizations have been active in improving training to increase awareness and knowledge of norms of relevant religious behavior. An important step taken by the American Psychiatric Association was to introduce the V code into the fourth edition of its internationally used diagnostic manual; this was a significant attempt to depathologize religious behavior, and to alert professionals to the possibilities and methods of distinguish normative from pathological religious behavior (Scott et al., 2003; Dein et al., 2011). Significant books contribute to awareness of this issue (e.g., Bhugra, 1996; Peteet et al., 2011). The UK Royal College of Psychiatrists Special Interest Group focusing on spiritual issues in psychiatry, the UK National Spirituality and Mental health Forum, and other bodies organize conferences and publications (Royal College of Psychiatrists, 2013; National Spirituality and Mental Health Forum, 2013), with significant focus on the importance of religious issues in diagnosis and treatment. There are several academic journals featuring research and thinking on the psychiatry–religion interface (notably *Mental Health Religion and Culture, Transcultural Psychiatry, and The International Journal of Social Psychiatry*). In related disciplines, Division 36 of the American Psychological Association focuses on the psychology–religion interface, and several journals contribute to the understanding of social-scientific perspectives on mental health issues (for example, *Anthropology and Medicine, Journal of Religion and Health, International Journal for the Psychology of Religion, and Social Science and Medicine*).

There has been a useful growth in the volume of work designed to improve understanding of the interface between psychiatry and religion, for both professionals and laypeople. There is scope for more, perhaps particularly for further attempts to improve understanding of mental disorder among laypeople, to reduce the stigma associated with this disorder, to reduce the negative impact of religious factors, and to allow scope for the positive roles that can be played by religion in coping with psychiatric disorders.

See also: Concepts and Methods of Cognitive Therapies; Counseling and Psychotherapy: Ethnic and Cultural Differences; Diagnostic Processes in Clinical Psychology and Psychotherapy; Religion and Health: Sociological Considerations; Religion, Psychology of; Religion: Psychiatric Aspects; Spirituality.

Bibliography

American Psychiatric Association, 2013. Mental Health. http://www.psychiatry.org/mental-health. doi: 6/10/13.
Bayes, J.E.B., Loewenthal, K.M., 2013. How do Jewish teachings relate to beliefs about depression in the strictly orthodox Jewish community? Mental Health, Religion and Culture 16, 852–862.
Bergin, A.E., Jensen, J.P., 1990. Religiosity of psychotherapists: a national survey. Psychotherapy: Theory, Research, Practice, Training 27, 3–7.
Bettelheim, B., 1983. Freud and Man's Soul. Chatto and Windus, London.
Bhugra, D., 1996. Hinduism and ayurveda: implications for managing mental health. In: Bhugra, D. (Ed.), Psychiatry and Religion: Context, Consensus, and Controversies. Routledge, London.
Brown, L.B., 1987. The Psychology of Religious Belief. Academic Press, London.
Campion, J., Bhugra, D., 1997. Experiences of religious healing in psychiatric patients in South India. Social Psychiatry and Psychiatric Epidemiology 32, 215–221.
Cinnirella, M., Loewenthal, K.M., 1999. Religious influences on beliefs about mental illness in minority groups: a qualitative interview study. British Journal of Medical Psychology 72, 505–524.
Cook, C.H., 2011. The faith of the psychiatrist. Mental Health, Religion and Culture 14, 9–18.
Cook, C.H., Powell, A., Sims, A. (Eds.), 2009. Spirituality and Psychiatry. Royal College of Psychiatrists Publications, London.
Cook, C.H., Powell, A., Sims, A., Eagger, S., 2011. Spirituality and secularity: professional boundaries in psychiatry. Mental Health, Religion and Culture 14, 35–42.
Craissati, J., 1990. Mental health care in India. The Psychologist 3, 19–22.
Culliford, L., Eagger, S., 2009. Assessing spiritual needs. In: Cook, C., Powell, A., Sims, A. (Eds.), Spirituality and Psychiatry. Royal College of Psychiatrists Publications, London, pp. 16–38.
Dein, S., Alexander, M., Napier, D.A., 2008. Jinn, psychiatry and contested notions of misfortune among East London Bangladeshis. Transcultural Psychiatry 45, 31–55.
Dein, S., Lewis, C.A., Loewenthal, K.M., 2011. Psychiatric views on the place of religion in psychiatry: an introduction to this special edition of mental health, religion and culture. Mental Health, Religion and Culture 14, 1–8.
Dein, S., Cook, C.H., Koenig, H., 2012. Religion, spirituality, and mental health: current controversies and future directions. Journal of Nervous and Mental Disease 200 (10), 852–855.
Dein, S., Loewenthal, K.M., 2013. The mental health benefits and costs of Sabbath observance among orthodox Jews. Journal of Religion and Health 52, 1382–1390.
Dura-Vila, G., Hagger, M., Dein, S., Leavey, G., 2011. Ethnicity, religion and clinical practice: a qualitative study of beliefs and attitudes of psychiatrists in the United Kingdom. Mental Health, Religion and Culture 14, 53–64.
Dura-Vila, G., Littlewood, R., Leavey, G., 2013. Integration of sexual trauma in a religious narrative: transformation, resolution and growth among contemplative nuns. Transcultural Psychiatry 50, 21–46.
Ellis, A., 1958. Sex without Guilt. Hillman, New York.
Ellis, A., 2004. The Road to Tolerance: The Philosophy of Rational Emotive Behavior Therapy. Prometheus Books, Amherst, NY.
Firth, S., 1997. Dying, Death and Bereavement in a British Hindu Community. Peeters, Leuven.
Foskett, J., 1996. Christianity and psychiatry. In: Bhugra, D. (Ed.), Psychiatry and Religion: Context, Consensus, and Controversies. Routledge, London.
Freud, S., 1907. Obsessive Acts and Religious Practices. Collected Papers 1907/1924. Hogarth Press, London.
Freud, S., 1927. The Future of an Illusion. Hogarth Press, London.
Freud, S., 1928. Totem and Taboo: Resemblances between the Psychic Lives of Savages and Neurotics. Dodd, New York.
Freud, S., 1930. Civilisation and Its Discontents. Hogarth Press, London.
Freud, S., 1939. Moses and Monotheism. Hogarth Press and the Institute of Psychoanalysis, London.

Gilbert, P., Gilbert, J., Sanghera, J., 2004. A focus group exploration of the impact of *izzat*, shame, subordination and entrapment on mental health and service use in South Asian women living in Derby. Mental Health, Religion and Culture 7, 109–130.

Greenberg, D., Witztum, E., 2001. Sanity and Sanctity: Mental Health Work among the Ultra-Orthodox in Jerusalem. Yale University Press, New Haven and London.

Heelas, P., Woodhead, L., 2005. The Spiritual Revolution: Why Religion Is Giving Way to Spirituality. Blackwell, Oxford.

Hill, P.C., Hood, R., 1999. Measures of Religiosity. Religious Education Press, Birmingham, Alabama.

Jung, C.G., 1958. Psychology and Religion: East and West. Routledge and Kegan Paul, London.

King, M., Leavey, G., 2010. Spirituality and religion in psychiatric practice: why all the fuss? The Psychiatrist 34, 190–193.

Koenig, H., 2008. Religion and mental health: what should psychiatrists so? Psychiatric Bulletin 32, 201–203.

Koenig, H.B., McCullough, M.E., Larson, D.B., 2001. Handbook of Religion and Health. Oxford University Press, Oxford.

Koenig, H., King, D., Carson, V. (Eds.), 2012. Handbook of Religion and Health. Oxford University Press, Oxford and New York.

Leavey, G., Loewenthal, K.M., King, M., 2007. Challenges to sanctuary: the clergy as a resource for mental health care in the community. Social Science and Medicine 65, 548–559.

Lerner, 1980. The Belief in a Just World: A Fundamental Delusion. Plenum, New York.

Special issue: psychiatrists' views on the place of religion in psychiatry. In: Lewis, C.A., Dein, S., Loewenthal, K. (Eds.), Mental Health, Religion and Culture 14 (1).

Lipsedge, M., 1996. Religion and madness in history. In: Bhugra, D. (Ed.), Psychiatry and Religion: Context, Consensus, and Controversies. Routledge, London, pp. 23–50.

Littlewood, R., Lipsedge, M., 1997. Aliens and Alienists: Ethnic Minorities and Psychiatry, third ed. Oxford University Press, London.

Loewenthal, K.M., 1994. Religion and Mental Health. Chapman & Hall, London.

Loewenthal, K.M., 2000. The Psychology of Religion: A Short Introduction. Oneworld/Penguin, Oxford reprinted 2004.

Loewenthal, K.M., 2006. Strictly orthodox Jews and their relations with psychotherapy and psychiatry. World Cultural Psychiatry Research Review 1. Special Issue on Culture, Spirituality and Mental Health.

Loewenthal, K.M., 2007. Religion, Culture and Mental Health. Cambridge University Press, Cambridge.

Loewenthal, K.M., 2013. Religion, spirituality and culture. In: Pargament, K.I., Exline, J., Jones, J., Mahoney, A., Shafranske, E. (Eds.), APA Handbook of Psychology, Religion and Spirituality. American Psychological Association, Washington, DC.

Loewenthal, K.M., Rogers, B.M., 2004. Culture sensitive support groups: how are they perceived and how do they work? International Journal of Social Psychiatry 50, 227–240.

Merriam-Webster, 2013. Merriam-Webster 2013.com/dictionary/psychiatry doi: 25.09.13.

Miller, L., 1972. Mental illness. In: Roth, C. (Ed.), Encyclopedia Judaica. Keter Publishing, Jerusalem.

National Spirituality and Mental Health Forum, 2013. http://mhspirituality.org.uk/about.html doi: 06/10/13.

Neeleman, J., Persaud, R., 1995. Why do psychiatrists neglect religion? British Journal of Medical Psychology 68, 169–178.

Paloutzian, R.F., Park, C.L. (Eds.), 2013. Handbook of the Psychology of Religion and Spirituality, second ed. Guilford Press, New York.

Pargament, K.I., 2007. Spiritually Integrated Psychotherapy: Understanding and Addressing the Sacred. Guilford, New York.

Pargament, K.I., Exline, J., Jones, J., Mahoney, A., Shafranske, E. (Eds.), 2013. APA Handbook of Psychology, Religion and Spirituality. American Psychological Association, Washington, DC.

Peteet, J.R., Lu, F.G., Narrow, W.E. (Eds.), 2011. Religious and Spiritual Issues in Psychiatric Diagnosis: A Research Agenda for DSM-V. American Psychiatric Publishing, Arlington, VA.

Poole, R., Higgo, R., 2011. Spirituality and the threat to therapeutic boundaries in psychiatric practice. Mental Health, Religion and Culture 14, 19–30.

Regan, J.L., Bhattacharyya, S., Kevern, P., Rana, T., 2013. A systematic review of religion and dementia care pathways in black and minority ethnic populations. Mental Health, Religion and Culture 16, 1–15.

Royal College of Psychiatrists, 2013. Mental Health Information for All. http://www.rcpsych.ac.uk/healthadvice.aspx. doi: 06/10/13.

Scott, S., Garver, S., Richards, J., Hathaway, W.L., 2003. Religious issues in diagnosis: the V-Code and beyond. mental health. Religion and Culture 6, 161–173.

Segal, Z., Teasdale, J., Williams, M., 2002. Mindfulness-Based Cognitive Therapy for Depression. Guilford, New York.

Stagg, F., 1962. New Testament Theology. Broadman Press, Nashville, TN.

Stanford, M.S., 2007. Demon or disorder: a survey of attitudes toward mental illness in the Christian church. mental health. Religion and Culture 10, 445–449.

Valentine, E.R., 1989. A cognitive psychological analysis of meditation techniques and mystical experiences. Ethical Record, 9–20.

Yossifova, M., Loewenthal, K.M., 1999. Religion and the judgement of obsessionality. Mental Health, Religion, and Culture 2, 145–152.

Psychiatry: Anthropological Aspects

Alfred Kraus, Centre for Psychosocial Medicine, University Hospital for General Psychiatry, Heidelberg, Germany

Published by Elsevier Ltd.

Abstract

Psychiatry today understands itself empirically to a far-reaching extent, as a science of facts of various methods. In its medical practice, being related to the whole mentally ill human, the question is put as to how far scientific methods are related to the whole being of a human, that is, to anthropological aspects and in which way these are guiding scientific research.

Various kinds of anthropologies have implications for medicine, particularly for psychiatry and psychotherapy. Most of these anthropologies are latent and not always easy to determine. These are, however, also anthropological approaches that are related to a certain kind of pioneering anthropology.

Along with the separation of natural and cultural sciences, we have to discriminate between anthropology in the sense of natural sciences and philosophical as well as cultural anthropologies. Biological anthropology, identical with physical anthropology, or human biology, is concerned with the biological nature of humans. In contrast to or supplementary to this, philosophical anthropology refers to statements about the personal or spiritual essence of humans. Cultural anthropology, mostly interchangeable with ethnology comparing different cultures, is concerned with the question, 'What is variable or constant in human beings in different cultures?' Transcultural psychiatry relates to cultural anthropology. All of these anthropologies can have a relationship to psychiatry in one way or another. Some approaches in psychiatry can conceptually have a stronger and explicit relationship to a certain kind of anthropology. We will be concerned here mainly with these approaches in psychiatry.

From a methodological point of view, we can speak of two main perspectives in psychiatry: a natural–scientific one and a hermeneutical one (hermeneutical not in the narrower sense). In some way, both these perspectives can be conceptually and methodologically opposite or supplementary to each other. The natural–scientific perspective relates to laws, and the other perspective relates to reasons (e.g., such as motives). The natural–scientific perspective limits the experience to what is repeatable, measurable, and calculable. It strives to eliminate the subjectivity of the patients. The hermeneutical perspective is mainly concerned with the subjectivity of the patient, his/her personal situation and history, and with his/her unique inner experiences. A basic task of every kind of anthropology is, or should be, to determine in which respect humans are different from any other kind of being, particularly from animals, and how far anthropological statements are universally valid. All these questions are fundamentally important for all concepts of mentally ill persons.

The most important influences on psychiatry have emanated from different orientations of philosophical anthropology. Questions such as 'What mental illness is in its various manifestations?,' 'Which methods of investigation and which kinds of therapies are adequate to mentally ill patient?,' and 'Which method is ethically justifiable?' were in different cultures and in different historical epochs, as well as they are now, to a high degree dependent on the respective conception of human beings. In different ways, the various approaches determine the border between the pathological and the normal state, between illness and health. This is to a very high degree dependent *on* basic philosophical assumptions about the nature of humans, as, for example, the relationship of body and soul, of biological and sociological determinations, of human heteronomy and autonomy in general. Because psychiatry as a science as well as a therapeutic practice latently always implies an anthropology, it cannot avoid reflecting on these premises. The respective anthropology has an influence upon what in psychiatry can become a scientific object and by which methods this object is accessible. Thus, basic anthropological assumptions on the one hand not only promote the development of certain approaches of psychiatry but on the other hand they may also prevent the development of others.

Considering a certain mental disorder under anthropological aspects, or under the aspect of the relationship of certain dimensions of human beings, can have consequences for a widening development and refinement of our modes and methods of making experiences with our patients. The experiences of the patients might be very strange to us, and make it difficult for us to emphasize with them.

From a philosophical point of view, the most fundamental question of anthropology is, 'To what extent is it possible to start with a previously given essence of a human being and in which way is the human being incomplete, "an open possibility" (Jaspers 636) in what he/she is?' This question became one of the main topics of existential philosophy: 'How far can it be determined what man is in his essence?'

Some anthropological approaches tend to a comprehensive foundation of medicine, sometimes with a danger to idealize certain approaches in medicine (Schmidt-Degenhard, 2002). In contrast, anthropological approaches can be very reductive, and empirically lead to nothing but theory.

What Jaspers said about philosophy in general is also valid for anthropology. He stresses its significance for psychiatry; it does not only free the inner space for new possibilities of knowledge but it can also open up horizons that are more adequate to a better understanding of the patient. A switching between anthropological and empirical research (Mundt and Spitzer, 1999) has a decisive significance for innovation and thus for the development of psychopathology as well as psychiatry in its therapeutic and rehabilitative practice.

Immanuel Kant ([1798]1978) was one of the first philosophers who explicitly formulated a philosophical anthropology. He differentiated between a physical and a pragmatic anthropology. Although the first one is concerned with what humans are by nature, the second one, identical with philosophical anthropology, asks what they can make out of themselves. Kant's philosophy became important for psychiatry not only because he developed interesting ideas about mental illness, seeing its general features in a loss of communal sense (sensus communis), that is, the loss of the ability of a person to be corrected by others (§ 53, p. 117), but also especially because he has influenced psychiatry with respect to its methodology. Showing the dependency of experience on a priori ideas and thus overcoming the traditional splitting of subject and object, Kant's philosophy ([1781/87]1965) stood for an opposition of idealism to realism, which maintains the independence of recognizing the subject from the object. This opposition has had a constant effect not only on many philosophical schools of thought but also on psychiatry. With his concept of the transcendental subject, he paved the way for Husserl's phenomenology and for existential approaches of philosophy. These new concepts and methods on their part made possible approaches such as *Daseinsanalyse* and existential psychoanalysis, existential psychiatry, and cognitive approaches. Because above all in psychoses the comprehension of the world by the experiencing subject has changed, these psychiatric approaches claim to provide a better understanding of the cognitive processes in these patients.

The following authors of a philosophical anthropology have become important in psychiatry: J.G. Herder (1744–1803), for emphasizing the human as a being gifted with language; G.W.F. Hegel (1770–1831), for his historical philosophical critique of anthropology; M. Scheler (1874–1928), for being the founder of modern anthropology in his work *The Place of Man in the Universe*; H. Plessner (1892–1985), for speaking of the eccentric position of human beings, who have to realize themselves what they are; and A. Gehlen (1904–67) for his determination of human beings as being deficient because of their lack of adjustment to their surroundings and their reduction of instinct. At the same time, an openness to the world corresponds to this lack.

For a psychiatry that calls itself anthropological, the following philosophers were above all influential: S. Kierkegaard (1813–55), who initiated the development of existential philosophy; E. Husserl (1859–1938), who was the main founder of the phenomenological approach in philosophy; M. Heidegger (1889–1976), who put forward Daseinsanalysis; J.P. Sartre (1905–80), who proposed existential psychoanalysis; and M. Merleau-Ponty (1908–61), who put forward his phenomenology of the body. However, there has also been an influence of medicine, especially of psychiatry, on philosophical anthropology: here, we could mention Heinroth (1773–1843), V. von Weizsäcker (1886–1957), and L. Binswanger (1881–1966).

Multiple influences on psychiatry also arose from biological anthropology (including paleoanthropology, human genetics, and research on human bodily constitution), from social and cultural anthropology, from cognitive and evolutional anthropology. Social anthropology and role theory (Kraus, 1980) had a strong impact on the development of social psychiatry and antipsychiatry, followed by many psychopathology concepts and therapeutic measures.

Cultural anthropology examines the cultural variation of the human being compared to human constants independent of culture and epoch and very clearly shows the double significance of the concept of anthropology already mentioned above. It shows the variability of the actual being and behavior of human beings and at the same time marks the boundaries of anthropology as a universally valid theory of the essence of humankind. Cultural anthropology has cultivated a special approach of a transcultural or cultural psychiatry, which investigates the variability and frequency of symptoms of disturbance entities, nowadays defined in glossaries, as well as of these entities themselves in different cultures. It has also led to the question whether regarding the classical entities of disturbances such as schizophrenia and manic-depressive illness we are actually dealing with entities of natural diseases, or whether these are entities coming from concepts of understanding. Transcultural or crosscultural psychiatry is thus fundamentally involved in the contemporary crisis of nosology in psychiatry.

With the notion of a cognitive anthropology, Emrich and Schiefenhövel (2001) sum up different approaches of psychological, physiological (Singer, 1994; Roth, 1994), and philosophical constructivism. In cognitive anthropology, the leading idea is the general quality of constructivity ascribed to systems generating consciousness (Roth, 1994). The progress of modern evolutional biology could in the future enable a phylogenetic and ontological understanding of mental disturbances.

Heinroth can be looked at as the founder of an anthropological medicine in its true sense in the context of psychiatry, not to be confused with a medical anthropology. Many anthropological approaches in general medicine, especially those of Weizsäcker and Wyss, have also had effects on psychiatry.

Janzarik (1988) has presented an approach of his own to anthropological psychiatry. In his structural–dynamic concept, he sees psychopathological phenomena in their neurobiological (dynamic) foundation as coherently connected with biographical, structurally bound significances.

Kick (2004, 2005, 2009) has dealt with ethical and anthropological questions, and has compared therapy and marketing of healthiness, such as material goods, by seeing the therapist as a therapeutic manager and the patient as a client. He also tested dilemmata of anthropological premises in borderline situations such as prevention of suicide and medical assisted suicide.

Schmitt (2003) investigated the old anthropological model of diet and ethical questions with respect to the progress of research. Schmitt (2010) stressed also reconciliation as the aim of psychotherapy.

Schmidt-Degenhard (1997, 2002) has given several instructive overviews on the phenomenological–anthropological approach in psychiatry. Schmidt-Degenhard (2000) contributed much to the differentiation of the illusion of the imagination, fiction, and the oneiroid.

Mundt (1989) pleaded for an alternation between a hermeneutic–holistic view and an empirical–alienating objectivation. Mundt and Spitzer (1999) opined that cognitive neurosciences overcame the dualism of these methods by integrating

biological function and subjective experience. Mundt (1989) and Mundt with Spitzer (1999) gave overviews on contemporary psychopathology. Applying the notion of intentionality, Mundt (1984, 1991) developed a new understanding of syndromes of deficiency and other syndromes in schizophrenics. Mundt and Fuchs (2002) dealt with an interaction of objectifying methods and holistic experiences under the aspect of anthropological phenomenology.

Under the heading of phenomenological–anthropological psychiatry, different approaches can be summarized, which have a phenomenological method and an anthropological orientation in common. Psychiatrists, psychologists, and philosophers of this approach are among others: In Germany: v. Baeyer, Blankenburg, Bürgy, Feldmann, Fuchs, v. Gebsattel, Häfner, Kisker, Kraus, Lang, Müller-Suhr, Schmidt-Degenhard, Tellenbach, and Wyss. In France: Charbonneau, Naudin, Pringuey, and Tatossian. In the United States: Hersch, Schwartz, Straus, and Wiggins. In Italy: Ballerini, Callieri, Calvi, and Stanghellini. In Denmark: Parnas and Zahavi. In Switzerland: Holzhey-Kunz. In Chile: Dörr-Zegers and Figaroa. These psychiatric approaches of an eidetic phenomenology of essence and of a phenomenology of constitution as well as of daseinsanalysis and existence analysis relate to different authors of philosophical anthropology in its broader sense such as E. Husserl, M. Heidegger, J.P. Sartre, and M. Merleau-Ponty.

Even if the understanding of phenomenology and phenomenological in these different approaches is not always the same, the notion of phenomenon aims at that 'which-shows-itself-in-itself' and as such is different from the notion of the symptom, which points to something that does not just show itself directly, namely, to the illness underlying it. An approach strongly oriented to the phenomenon enables us to describe the original individual subjective experience as well as the respective life world of the patient more adequately and in more detail than is otherwise possible.

'Anthropological' signifies on the one hand what is common to all people and on the other hand what in contrast to the animal is specifically human. In the context of anthropological psychiatry with this term, the relationship of psychiatric disturbances to basic human structures is meant, as, for instance, consciousness and subjectivity, intersubjectivity and individuality, history and temporality, freedom and corporality. These approaches not only take into consideration human essence because of ethical or humanistic reasons but above all they consider methodological reasons. The aim is to understand psychiatric disturbances as regular variations of these basic structures that means of human essence itself. This is done not so much to investigate the factual conditions for the development of psychopathological phenomena but to investigate the conditions originating from human essence, which made them possible. For instance, hallucination can be understood out of the normal act of perception as an uncoupling of spontaneous imaginative acts from a normal interaction with receptivity (Blankenburg, 1980a,b).

Binswanger's (1955) concept of 'anthropological proportion' has turned out to be especially fertile in psychiatry. The partly oppositional dimensions of height and breadth in human existence, of self-realization and world realization, of individuation and relationship to community, of need for continuity and openness to innovation, of tolerance and intolerance of ambiguity, of overidentification and underidentification normally stand in a certain proportion to each other. These proportions are shifted in psychiatric disturbances. Relating to this alteration of the respective proportions, some pictures of illness are oppositional. Thus, an overidentification with their respective being in melancholics can be opposed to an underidentification in hysterics (Kraus, 1996a). Overidentification with the social role, for example, with the environment, can understandably precipitate a melancholic phase in the case of their loss. One speaks of an empty-nest depression, if a parent becomes depressed when children leave home. Other examples are depression in the situation of retirement or of relocation. Hysterics, on the contrary, tend rather not to identify with their respective being or their actual feelings. They are striving to be more and something different than they actually are, or to show feelings they do not have. The abnormal as well as the normal can in this way be located in a polar field that is structured through ideal norms that oppose one another in a field where 'deviations' from at least two sides are possible, as Blankenburg (1980a) put it. The pathological element can in this way be traced back to healthiness. On the other hand, healthiness shows itself as something that always holds the possibility of becoming pathological by an alteration of proportions.

Such a concept of mental disturbances offers a special possibility of identification with the patient in the therapeutic process. Beyond a purely descriptive psychopathology in the sense of Jaspers (1965), it is thus possible to develop a dynamic phenomenology of psychiatric disturbances that has a dialectic character. Jaspers in his subjective psychopathology emphasized the special experience of the patient as opposed to a mere behavioral description in the sense of an objective psychopathology. Assigning Jaspers to the phenomenological–anthropological approach is sometimes disputed, because on the one hand he was one of the most important promoters of this approach with many of his ideas and concepts, but on the other hand, he was very critical of the application of philosophical anthropology in psychiatry and had a very restrictive understanding of phenomenon, identifying it with symptom. Later on, he was more sympathetic with this approach.

An eidetic essence-oriented phenomenology described above goes a step further in trying to grasp the essence of altered experiences and behavior in a so-called eidetic reduction in the sense of Husserl. In a process of free imaginative variation, one tries here to pass from a fact to the essence of a fact, from a sensual to a categorical presentation. Categorical presentation means, for instance, to recognize a particular species of tree in an actual tree, or in different kinds of behavior and experiences of hysterics the element of underidentification, of the pretense of being instead of actual being. Such an eidetic reduction can also be carried out in relation to Heidegger's notion of '*In-der-Welt-sein*' or Merleau-Ponty's '*Être-au-monde*' to show the special relationship of patients to themselves and to the world, which is expressed in their experiences and behavior. A constitutive or transcendental phenomenology in the sense of Husserl should be differentiated from an eidetic phenomenology. Instead of passing from the fact to the essence of a fact as

in eidetic phenomenology, constitutive phenomenology examines the constitutive achievements of consciousness, which lets us perceive the concrete as it appears to us.

This approach of a constitutive phenomenology and the following philosophy of existence became important for psychiatry, because in the perception of psychotic patients – but not only for them – the world and its significances constitute themselves in another way than in healthy people. In psychiatry, this resulted not only in making the experience of objects, for instance, classical symptoms, into a research topic, but also the preobjective experience, that is, the constitutive processes out of which self and world develop. The particular method of constitutive phenomenology so called epoché, that is, putting all judgments of reality in brackets by refraining from it. This is particularly important for the diagnostician and the therapist for being open as much as possible for the often extremely strange world in which psychotic people live and to transcend in their world. Not only to judge patients insofar as they deviate from the norms of normality but from their being-in-the-world is one of the most important advantages of the phenomenological–anthropological diagnostics.

The preobjective experience is especially important in the intuitive diagnostics, for instance, in the so-called precox feeling. Such preobjective experiences are, for instance, the spatialization, temporalization, intersubjectivity, and corporality of a patient. Classical symptoms and concepts, such as delusion, hallucination, phobic anxieties, and obsessive–compulsive symptoms, assessed by a psychopathology oriented to the object turned out to be constructed on such preobjective experiences.

Phenomenology of constitution, as Husserl put it, has recently entered into a manifold relationship with cognitive theory and cognitive neuroscience (Wiggins and Spitzer, 1997; Gallagher, 1997). Varela (1996) even proposed a neurophenomenology. The main problem of such attempts is that these very different methods have a completely different understanding of what the notions of evidence in both methods mean. Evidence in the phenomenological–anthropological approach is not identical merely with plausibility. Several authors of the phenomenological–anthropological approach showed how in the contents of delusion and hallucination versus the psychotic tendency of dissolution of their world patients try again to give sense to their new world, even if a distorted one (Feldmann and Schmidt-Degenhardt, 1997; Müller-Suur, 1980; Kraus, 1994a).

Husserl's concept of an intersubjectively constituted life world has become increasingly important for psychiatry, above all for the understanding of schizophrenic modes of existence (Callieri and Castellani, 1981; Blankenburg, 1979). People with schizophrenia are characterized by a damaged rootedness in the life world, connections of significances and relevancies becoming confused, which shows itself in extravagance and oddity and also in the delusions of these patients. Blankenburg (1971) speaks of a lack of natural self-evidence in schizophrenics. Although schizophrenics often lack an inner standard that would allow them a behavior appropriate to the social norms, melancholics – being too adapted to these norms – are too strongly fixed and to their lifeworld to common sense (Blankenburg, 1979; Stanghellini, 1997; Kraus, 1996b). This is already shown premorbidly by their hypernomic behavior, which is characterized by a lack of distance to normative expectations (Kraus, 1996b). It is also shown by the fact that they find their identity above all in their social roles. For this reason, they are especially in danger of becoming depressed in the case of the loss of a social role (Kraus, 1996b; see below).

Binswanger has developed his 'Daseinsanalyse' (not to be confused with '*Daseinsanalytik*') above all under the influence of Heidegger's '*Daseinsanalytik*' in *Being and Time* (Heidegger, 1963), which, against the ontological intention of the author, Binswanger understood as an anthropology. With 'Daseinsanalyse' we are concerned with a hermeneutic of the inner life history, which is asked for its underlying individual concept of the world determining the concrete experience and behavior of a person.

Sartre's (1962) existential psychoanalysis has also contributed to a better understanding of certain pathological modes of existence such as those of hysterics, melancholics, and patients with anxiety and obsessive–compulsive disorders. In this way, we can establish certain existential types, which can have an impact on the classification of these disorders (Kraus, 1994b).

Partly as a reaction to the philosophy of consciousness of Husserl, and also following the late Husserl, the philosophy of the twentieth century was very much determined by the topics of a phenomenology of the life world and of the body (e.g., Merleau-Ponti (1965), H. Schmitz (1964–78), B. Waldenfels (2000)). For Husserl, the life world became particularly important for the foundation of scientific knowledge in everyday life experiences, that is, in pregiven matters, in sedimented experiences of generations. This became important, for example, in the understanding of schizophrenia under the aspect of an so called 'loss of naturalness' (Blankenburg, 1971).

It was particularly Merleau-Ponti (1965) who showed how our relationship to the world is determined by our bodilyness, that we have world only through our bodilyness. Thus, our world is always already a world interpreted by our bodilyness. Our body has so its own valid sense, its own understanding of the world already before we are an ego. This becomes particularly apparent in the kind of schizophrenia characterized only by certain bodily symptoms, in an so called coenesthetic schizophrenia (Kraus, 2011).

Here as well as in other psychiatric disorders, for example, in melancholia, Dörr-Zegers and Tellenbach (1980); Fuchs (2001); Kraus (2002) spoke of a corporealization, a coming to the fore of the physical corporeal body, which is normally as a lived body more or less tacitly transcended (Sartre, 1962). In the melancholic state, the body loses its capability of resonance in front of the environment (Fuchs, 2001, 2005). Because of this corporealization, a general kind of depersonalization (Fuchs, 2001; Kraus, 2002) may appear. Phenomenological–anthropological psychiatry is not so much interested in the deviant behavior as such of patients, but with the possibility to behave in one or another way. Thus, freedom is a prominent topic of this approach. Because forensic psychiatry has mainly to do with the question of an incapacity to behave in accordance with social norms, the phenomenological–anthropological approach is also of significance in this field. Tatossian (1996) was right in saying that

operational criteria are not sufficient, when it comes to assessing the capacity of a patient to act freely.

There are two fields of present psychiatry where anthropological therapy and contributions are particularly important. These are diagnostics and classification as well as psychotherapeutic treatment and rehabilitation. Already the notion of illness in psychiatry is so different from that in somatic medicine that it may be asked, whether it has to be anthropologically complemented (Wyss, 1973, 1982) or reformulated. According to Jaspers (1965), in psychiatry, we have to discriminate between disorders with a known somatic cause, called species (Gattung) and such disorders without a known somatic cause that Jaspers called ideal types (Idealtypen). Ideal types are constructs won by thinking intuition as an ideal typical coherent whole. Schneider (1967) took this notion of type from Jaspers saying that with schizophrenia and cyclothymia, we do not have to do a diagnosis in the medical sense but only with two types of mental states. Because symptoms and disorders here have a different logical structure, he spoke of criteria instead of symptoms. The glossaries of diagnosis and classification of Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Statistical Classification of Diseases and Related Health Problems (ICD) did not follow this way of Jaspers and Schneider but returned to Kraepelins' concept of these two psychoses as somatic illnesses, the notion of criterion here having a different, only an operational meaning.

The mentioned present manuals of diagnosis and classification were on the one hand seen as a great success, but on the other hand they have also been criticized. A biological psychiatrist wanted the criteria of the respective entities of disorders in their description to be more functionalized (van Praag, 1996; Gaebel, 1996), so as to let the classification entities appear as disturbance entities of some functions.

A team of the World Psychiatric Association called International Guidelines for Diagnostic Assessment (Mezzich and Ustum, 2002) proposed that the amendment of the manuals should be more oriented to the whole person of the patient, instead of seeing him only as the bearer of an illness. Different from the present reductionist manuals, these should be combined with a standardized, multiaxially personalized ideographical recording of the respective patient. Humanistic and ethical demands should be taken into account, and a prototypical categorization is requested.

All these aims are very similar to those of a phenomenological–anthropological psychiatry. From this point of view, a phenomenological–anthropological diagnostics and classification could be developed, complementing the symptomatological–criteriological one, as we proposed in some articles (Kraus, 1994b, 2003). The last one is mainly oriented to the morbus of the patient (in the sense of having a disease, e.g., having schizophrenia), the phenomenological one is holistically oriented to the person of the patient and his lifeworld (in the sense of being ill, e.g., of being schizophrenic). Consequently, in the symptomatological–criteriological diagnostics, we have to do with symptoms in the phenomenological one with phenomena. Phenomena are then not understood as identical with symptoms but mainly in the sense of eidetic and constitutive phenomenology and daseinsanalysis.

In the diagnostic manuals, disease entities are conceived as existing in reality, as being ontological realistic in nature. Here, disease entities and symptoms are taken as facts. It is not considered that these facts are loaded with meanings and values that incorporate cultural bias toward these entities and their symptoms (Fulford, 1989, 2005; Moij, 2012). The manuals conceive of a mental disorder as a direct consequence of a deficient or disturbed cerebral structure (Moij, 2012) that causes some dysfunction. The criteria of the disease entities are principally elementaristic, without a unifying sense. Behavioral aspects, not the experience of the patient, are in the foreground. The reliability of the criteria is taken as being more important than the validity. In the phenomenological–anthropological diagnostics and classification the classificatory entities can be seen as existential types (Kraus, 1994b) of modifications, or basic existential structures of being-in-the-world and being-with-others. Daseinsanalytic authors would speak of an orientation to the ontologically meant 'existentials' of Heidegger. Here, the methodological focus is on the individuality, subjectivity, historicity, temporality, spatiality, the embodiment of the schizophrenic, the manic, and the melancholic patient. All these phenomena are unified in a certain structure of the respective existential type in the sense of a psychopathological organizer, as Rossi Monti and Stanghellini (1996) put it. Thus, for example, schizophrenia, mania, and melancholia are seen as certain modifications of the being-in-the-world and in this way anchored in the essence of a human. Even if the existential types are not primarily oriented to the morbus entity, but to the being-in-the-world of the patient, they can be of important heuristic value for the finding of disease entities. Their particular value is that the guidelines and norms of psychotherapy and rehabilitation result immediately from the respective existential type, that is, from the diagnostic process itself.

In the following, we will deal with the procedure of a phenomenological–anthropological diagnostics and its significance for the therapeutic measure in the case of patients with a manic-depressive disorder, which we see as a kind of identity disorder. Before we can do this, we have to start from a concept of human identity that should not be confused with that of the identity of things. Instead of a concept of identity as a substance with certain qualities, we need a concept of identity that allows for developing oneself, that is, for identifying oneself with certain representatives of identity and also for giving them up, dissolving these identities in the course of life. Such a kind of identity is, for example, provided by Mead (1980), which we take over.

We will demonstrate how manic depressives have great difficulties with this kind of what we call a dialectic identity. They tend to establish an identity structure overidentify themselves with their respective social roles, with the spatiotemporal conditions of their everyday life, with their possessions, their dwelling, etc., being representatives of what we call object identities. Their overidentification is also shown by their hypernomic behavior that is a relationship to social norms without distance to social norms that would allow them to interpret the norms according to their respective situation. Thus, manic depressives tend to fulfill their duties in social roles in an excessive manner. They are also overidentified with their roles in

intimate partnerships in all these cases the object identities (the me in the sense of Mead) are normally counterbalanced by an ego identity better called a subject identity (the I in Mead's understanding), which allows for a certain distance to our object identities to react upon these by the subject identity. This is particularly important in the case of change or loss of one's object identities. Already in the premorbid state, the hypernomic behavior is characterized by a lack of ego achievements and consequently of lack of ego identity. This lack makes these patients particularly sensitive to situations such as retirement, chance of dwelling, any kind of role loss, which precipitate depression. The symptomatology of melancholic depression shows also this structure of overidentification. Because of the identity crises that result from a real, feared, or imagined loss of object identity, which cannot be compensated by a sufficient ego identity or subject identity, the patients cling to their object identities even more, which is expressed in the themes of their delusional ideas. According to the four fundamental kinds of basic dimensions of human identity, there are mostly four kinds of delusional topics to be found in melancholia: hypochondriacal delusion, corresponding to the identity we find in our lived body (Leib); delusion of impoverishment, corresponding to the identity we find in our possessions; delusion of guilt, corresponding to the identity we find in the evaluating judgments of others or of ourselves. Although the first three delusional themes relate to the clinging to or flight into object identities, delusion of insuffiency relates directly to the lack of ego identity of ego achievements.

In mania, compared to in depression, we find a precise inversion of the kind of being. Here, the patients struggle to overcome every kind of being determined by their object identities. In a willfulness to define themselves autonomously, they attend to assert their subject-I, negating their object identity. But these attempts fail because they cannot go without their object identities; melancholics are imprisoned in their overidentification with their object identities, but the manics feel themselves forced to use every opportunity to free themselves from every attachment to their object identities.

In the case of melancholia, the symptomatological–criteriological diagnostics would see a general inhibition and in mania a general loss of inhibition, whereas the phenomenological–anthropological diagnostics does not only point merely to the kind of behavior in these states, isolated from the person of the patient, but also to the respective change of his being, his change of identity, which secondarily results in the respective kind of behavior. Psychotherapy in these cases would also not only be oriented to the behavior but also to the identity of the patient in the sense of an identity therapy (Kraus, 1995). The particular value of a phenomenological–anthropological diagnostics is that the guidelines and norms of psychotherapy and rehabilitation result immediately from the diagnostic process itself, seeing a certain kind of identity structure as a particular existential type. Thus, the respective anthropological understanding of the illness has serious implications for the kind of therapy we prefer and has also consequences for the significance and special kind of rehabilitation.

Anthropology has promoted a series of therapeutic measures. We only mention Daseinsanalysis in the sense of Biswanger (1957, 1961), Blankenburg (1982), Boss (1951), Condrau (1963), Holzhey-Kunz (1994), Lang and Weiss (1992), and Lang (1992, 2011), existential psychoanalysis and existential psychotherapy in the wake of Sartre and other philosophers (Cannon, 1991; Hersch, 2003; Kraus, 1995; Mooji, 2012; Yalom, 1980), and the anthropological-integrative psychotherapy of Wyss (1992). Some of these authors and other psychotherapists (Kraus, 1995; Mundt, 1996; Schmitt, 2010) are eclectic in their psychotherapeutic measures but with an anthropological orientation.

See also: Anthropology, History of; Cognitive Anthropology; Health and Illness: Mental Representations in Different Cultures; Mentally Ill, Public Attitudes towards; Mind–Body Dualism; Normality and Mental Health; Psychological Anthropology.

Bibliography

Binswanger, L., 1955. Vom anthropologischen Sinn der Verstiegenheit. In: Ausgewählte Vortrage und Aufsätze, vol. 2. Franke, Bern, Germany, 235–242.
Binswanger, L. Schizophrenie, 1957. Neske, Pfullingen, Germany.
Binswanger, L., 1960. Melancholie und Manie. Neske, Pfullingen, Germany.
Binswanger, L., 1961. Über die daseinsanalytische Forschungsrichtung in der Psychiatrie. In: Ausgemachte Vorträge und Aufsätze. Bd. 1. Francke, Bern.
Binswanger, L., 1992. Drei Formen missglückten Daseins. In: Herzog, M. (Ed.), Ludwig Binswangers ausgewählte Werke, vol. I. Asanger, Heidelberg, Germany, pp. 1–443.
Binswanger, L., 1957. Schizophrenie. Neske, Pfullingen.
Blankenburg, W., 1971. Der Verlust der natürlichen Selbstverständlichkeit. Enke, Stuttgart, Germany.
Blankenburg, W., 1978. Was Heißt Anthropologische Psychiatrie? Lit, Münster, Hamburg, London.
Blankenburg, W., 1979. Phänomenologische Epoche und Psychopathologie. In: Sprondel, M., GrathofT, R. (Eds.), AlfredSchütz und die Idee des Alltags in den Sozialwissenschaften. Enke, Stuttgart, Germany, pp. 125–139.
Blankenburg, W., 1980a. Anthropologisch orientierte psychiatrie. In: Peters, U.W. (Ed.), Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts, vol. 10. Kjndler, Zürich, pp. 182–197.
Blankenburg, W., 1980b. Phenomenology and psychopathology. Journal of Phenomenological Psychology 11, 50–78.
Blankenburg, W., 1982. Die Daseinsanalyse. In: Eicke, D. (Ed.), Kindlers Psychologie des 20. Jahrhunderts. Tiefenpsychologie. Band 4: Individual psychologie und Analytische Psychologie. Betz, Weinheim, Basel.
Blankenburg, W., 1983. Die Psychotherapie Schizophrener als Ort psychoanalytisch-daseinsanalytischer Konvergenz. Nervenarzt 54, 144–149.
Blankenburg, W., 1989. Der Krankheitsbegriff in der Psychiatrie. In: Brennpunkte der Psychiatrie der Gegenwart. Bd 9. Springer, Berlin, Heidelberg, pp. 119–140.
Boss, M., 1951. Beitrag zur daseinsanalytischen Fundierung des psychiatrischen Denkens. Schweiz Archives of Neurology and Psychiatry 67 (1), 15–19.
Callieri, B., CastcIlani, A., 1981. On the psychopathology of the life-world. In: Bello, A.A. (Ed.), Analecta HusseHiana. Reidel, Dordrecht, Germany, pp. 172–202.
Cannon, B., 1991. Sartre & Psychoanalysis. An Existentialist Challenge to Clinical Metatheory. University Press of Kansas, Lawrence, KS.
Condrau, G., 1963. Daseinsanalytische Psychotherapie. Huter, Bern.
Dörr-Zegers, O., Tellenbach, H., 1980. Differentialphänomenologie des depressiven Syndroms. Über melancholische und epileptische Depressivität. Nervenarzt 51, 113–118.
Emrich, H.M., Schiefenhövel, W., 2001. Philosophical Anthropology: Basic Science of Psychiatry. Springer Heidelberg. NY, London.
Feldmann, H., Schmidl-Degenhard, M., 1997. SIrukturale Affinität des Unverständlichen im schizophrenen Wahn. Nervenarzt 68, 226–230.
Fuchs, T., 2000a. Psychopathologie von Leib und Raum. Steinkopf, Darmstadt.
Fuchs, T., 2001. Melancholia as a desynchronisation towards a psychopathology of interpersonal time. Psychopathology 34, 179–186.

Fuchs, T., 2005. The phenomenology of body, space and time in depression. Comprendre 15, 108–121.
Fulford, K.W.M., 1989. Moral theory and medical practice. Cambridge University Press, Cambridge.
Fulford, K., 2005. Values in psychiatric diagnosis: developments in policy, training, and research. Psychopathology 38, 171–176.
Gaebel, W., 1996. Objektivierende Psychopathologiel in der biologisch-psychiatrischen Forschung. In: Saß, H. (Ed.), Psychopathologische Methoden und psychiatrische Forschung. Jena Stuttgart: Gustav Fischer, Germany.
Gallagher, S., 1997. Mutual enlightment: recent phenomenology in cognitive science. Journal of Consciousness Studies 4 (3), 195–214.
Gehlen, A., 1966. Der Mensch, eighth ed. Athenäum, Frankfurt, Germany.
Heidegger, M., 1963. Sein und Zeit. Niemeyer, Tübingen, Germany.
Hersch, E.L., 2003. From Philosophy to Psychotherapy. University of Toronto Press, Toronto, Buffalo, London.
Holzhey-Kunz, 1994. Leiden Am Dasein. Die Daseinsanalyse und die Aufgabe einer Hermeneutik psychopathologischer Phänomene Passagen Verlag.
Janzarik, W., 1988. Strukturdynamische Grundlagen der Psychiatrie. Enke, Stuttgart, Germany.
Jaspers, K., 1965. Allgemeine Psychopalhologie. Springer, Berlin.
Kant J., [1781/87]1965. Critique of Pure Reason (N. Kemp Smith, Trans.). St. Martin's Press, New York.
Kant J., [1798]1978. Anthropology (V.L. Dowdell, Trans., rev. and ed. Rudnick, H.H.). Southern Illinois University Press, Carbondale, IL.
Kick, H.A., 2004. Lebensende, Sterbehilfe, Suizid: ethische Dilemmata als Herausforderung zur Klärung menschenbildlicher Prämissen. In: Kick, H., Taupitz, J. (Eds.), Ethische Orientierung in Grenzsituationen des Lebens. Lebenskunst und Sterbekunst, Sterbebegleitung und Trauerarbeit. Münster.
Kick, H.A., 2005. Patienten oder Kunden – Therapeutische Offerte oder Marketing: anthropologische Unterscheidung und ethische Konsequenzen. In: Kick, A., Taupitz, J. (Eds.), Gesundheitswesen zwischen Wirtschaftlichkeit und Menschlichkeit. Litt, Münster Hamburg, London, pp. 151–178.
Kick, H.A., 2009. Grenzsituationen und Wertebildung – eine prozessdynamische Interpretation nach Karl Jaspers. In: Engelhardt, D.von, Gerigk, H.-J. (Eds.), Karl Jaspers im Schnittpunkt von Zeitgeschichte, Psychopathologie. Literatur und Film, Heidelberg, pp. 73–88.
Kraus, A., 1980. Bedeutung und Rezeption der Rollentheorie in der Psychiatrie. In: Peters, U.H. (Ed.), Ergebnisse für die Medizin, 2: Psychiatrie. Die Psychologie des 20. Jahrhunderts, vol. 10. Kindler, Zürich.
Kraus, A., 1994a. Phenomenology of the technical delusion in schizophrenics. Journal of Phenomenological Psychology 25 (1), 51–69.
Kraus, A., 1994b. Phenomenological and criteriological diagnosis: different or complementary. In: Sadler, J., Wiggings, O.P., Schwartz, M. (Eds.), Philosophical Perspectives on Psychiatric Diagnosis and Classification. Johns Hopkins University Press, Baltimore, MD, pp. 148–160.
Kraus, A., 1995. Psychotherapy based on identity problems of depressives. American Journal of Psychotherapy 49, 197–212.
Kraus, A., 1996a. Identitätsbildung Melancholischer und Hysterischer. In: Seidler, G.H. (Ed.), Hysterie Heute. Ferdinand Enke, Stuttgart, Germany, pp. 103–110.
Kraus, A., 1996b. Role performance, identity structure and psychosis in melancholic and manic-depressive patients. In: Mundt, C., Goldstein, M.J., Hahlweg, K., Fiedler, P. (Eds.), Interpersonal Factors in the Origin and Course of Affective Disorders. Gaskell, London, pp. 31–47.
Kraus, A., 1999. Phänomenologisch-anthropologische psychiatrie. In: Helmchen, H., Henn, F., Lauter, H., Sartorius, N. (Eds.), Grundlagen der Psychiatrie. Springer, Berlin.
Kraus, A., 2002. Melancholie: eine Art von Depersonalisation? In: Fuchs, T., Mundt, C. (Eds.), Affekt und affektive Störungen. Schoningh, Paderborn, München, Wien, Zürich, pp. 169–186.
Kraus, A., 2003. How can the phenomenological-anthropological approach contribute to diagnosis and classification in psychiatry? In: Fulford, B., et al. (Eds.), Nature and Narrative. Oxford Press, Oxford, New York, pp. 199–216.
Kraus, A., 2005. Psychotherapy based on identity problems. American Journal of Psychotherapy 49, 197–212.
Kraus, A., 2011. Schmerzen und andere Leibgefühlsstörungen bei zönaesthetischer Schizophrenie. In: Schiltenwolf, M., Herzog, W. (Eds.), Die Schmerzen. Königshausen & Neumann, Würzburg, pp. 221–236.
Lang, H., 1982. Phenomenologisch-anthropologisch orientierte Psychiatrie und Psychologie im französischen Sprachraum. In: Eicke, D. (Ed.), Tiefenpsychologie, Individualpsychologie und Analytische Psychologie, vol. 4. Weinheim, Basel, Switzerland.
Lang, H., Weiss, H., 1992. Interdiziplinäre Anthropologie. Königshausen und Neumann, Würzburg.
Lang, H., 2011. Die strukturale Triade und die Entstehung früher Störungen. Klett-Cotta, Stuttgart.
Mead, G.H., 1980. Geist, Identität und Gesellschaft. Suhrkamp, Frankfurt a.M.
Merleau-Ponti, M., 1965. Phänomenologie der Wahrnehmung. De Gruyter, Berlin.
Mezzich, J.E., Ustun, T.B., 2002. International classification and diagnosis in Psychiatry. Critical review and future directions. Psychopathology (special issue).
Monti, E.P., Stanghellini, G., 1996. Psychopathology: an edgeless razor? Comprehensive Psychiatry 37 (3), 196–204.
Mooij, A., 2012. Psychiatry as a Human Science. Phenomenological, Hermeneutical and Lacanian Perspectives. Rodopi, New York, Amsterdam.
Müller-Suur, H., 1980. Das Sinn-Problem in der Psychose. Hofgreve, Göttingen, Germany.
Mundt, C., 1984. Der Begriff der Intentionalität und die Defizienzlehre von den Schizophrenen. Nervenarzt 55, 582–588.
Mundt, C., 1989. Concepts of intentionality and their application to the psychopathology of schizophrenia. A critique of the vulnerability model. In: Spitzer, M., Maher, B.A. (Eds.), Philosophy and Psychopathology. Springer, New York, London, pp. 85–95.
Mundt, C., 1991. Constituting reality – its decline and repair in the long-term course of schizophrenic psychoses: the intentionality model. In: Marneros, A., Andreasen, N.C., Tsuang, M.T. (Eds.), Negative versus Positive Schizophrenia. Springer, Heidelberg, pp. S96–S108.
Mundt, C., 1996. Zur Psychotherapie des Wahns. Nervenarzt 67, 515–523.
Mundt, C., Spitzer, M., 1999. Psychopathologie heute. In: Psychiatrie der Gegenwart, fourth ed., Vol. I. Springer, Heidelberg, Germany, pp. 4–44.
Mundt, C., Fuchs, T., 2002. Anthropologische Phänomeologie heute. Ethik. Wechselspiel zwischen sinnverbindenden, objektivierenden Methoden und ganzheitlich-intuitiver Wahrnehmung. Schöningh, Paderborn.
Praag van, H.M., 1996. Functional psychopathology: an essential diagnostic step in biological psychiatric research. In: Ackenbeil, M., Bondy, B., Engel, R., Armann, M., Nedopil, N. (Eds.), Implications of Psychopharmacology to Psychiatry. Springer, Berlin, Heidelberg.
Roth, G., 1994. Gehirn und Geist. In: Schiefenhovel, W., Vogel, C., Vollmer, G., Opolka, U. (Eds.), Vom Affen Zum Halbgott. Thieme, Stuttgart, Germany.
Sartre, J.P., 1962. Das Sein und das Nichts. Rowohlt, Hamburg, Germany.
Schmidt-Degenhard, M., 1992. Die oneiroide Erlebnisform. Zur Problemgeschichte und Psychopathologie des Erlebens fiktiver Wirklichkeiten. Springer Berlin Heidelberg, New York, Tokio.
Schmidt-Degenhard, M., 1997. Zur Standortbestimmung einer anthropologischen Psychiatrie. Fortschritte der Neurologie Psychiatrie 65, 473–480.
Schmidt-Degenhard, M., 2000. Imaginative Aspekte des Wähnens. In: Hermann Faller (Hg) Angst, Zwang und Wahn. königshausen und Neumann, Würzburg.
Schmidt-Degenhard, M., 2002. Die Bedeutung der phanomenologisch-anthropologische Forschung für die Psychiatrie der Gegenwart. In: Affekt und affektive Störungen. Schoningh Paderborn.
Schmitt, W., 2003. Diätethik als anthropologisches Modell und ethische Leitlinie angesichts medizinischer Forschungsfortschritte. In: Kick, H.A., Taupitz, J. (Eds.), Handeln und Unterlassen. Ethik und Recht in den Grenzbereichen von Medizin und Psychologie, Berlin, Heidelberg, pp. 9–16.
Schmitt, W., 2010. Versöhnung als Therapieziel. Aspekte einer anthropologischen Psychotherapie. In: Kick, H., Dietz, G. (Eds.), Trauma und Versöhnung. Heilungswege in Psychotherapie, Kunst und Religion. LIT Verlag, Berlin.
Schmitz, H., 1964–1978. System der Philosophie. Bouvier, Bonn.
Schneider, K., 1967. Klinische Psychopathologie (8. Aufl.). Thieme, Stuttgart, Germany.
Singer, W., 1994. Gehirn und Bewusstsein. Spektrum, Heidelberg, Germany.
Stangheilini, G., 1997. For an anthropology of vulnerability. Psychopathology 30, 1–11.
Tatossian, A., 1996. La phénoménologie: une épistémologie pour la psychiatrie. Confrontations Psychiatriques 37.
Tellenbach, H., 1975. Die Begründung psychiatrischer Erfahrung und psychiatrischer Methoden in philosophischen Konzeptionen vom Wesen des Menschen. In: Gadamer (Ed.), Neue Anthropologie. Philosophische Anthropologie, 1. Teil, Stuttgart, pp. 138–181.
Tellenbach, H., Dörr-Zegers, O., 1980. Differentialphänomenologie des depressiven Syndroms. Über melancholische und epileptische Depressivität. Nervenarzt 51, 113–118.
Varela, F., 1996. Neurophenomenology. A methodological remedy for the hard problem. Journal of Consciousness Studies 4, 330–349.
Waldenfels, B., 2000. Das Leibliche Selbst. Suhrkamp, Frankfurt am Main.
Wiggins, O.P., Spitzer, M., 1997. Cognitive science. In: Embree, L., Behnke, E.A., Carr, D., et al. (Eds.), Encyclopedia of Phenomenology. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 101–104.
Wyss, D., 1973. Beziehung und Gestalt. Vandenhoeck & Ruprecht, Göttingen.
Wyss, D., 1982. Der Kranke als Partner. In: Vandenhoeck und Ruprecht, vols. I and 2. Göttingen, Germany.
Wyss, D., 1992. Von der Psychoanalyse Sigmund Freuds zur anthropologisch-integrativen Psychotherapie. In: Lang, H., Weiß, H. (Eds.), Interdisziplinäre Anthropologie. Königshausen & Neumann, Würzburg.
Yalom, J.D., 1980. Existential Psychotherapy. Basic Books, New York.

Psychiatry: Informed Consent and Care

Laura W Roberts, University of New Mexico, Albuquerque, NM, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12296–12302, © 2001, Elsevier Ltd.

Abstract

Informed consent is a philosophical and legal doctrine that has come to serve as the cornerstone of ethically sound clinical care in Western society. This doctrine requires that individuals truly understand and freely and knowledgeably make choices regarding their treatment. Informed consent is an ideal that emphasizes respect for the individual and places primacy on autonomous decision-making. Informed consent, furthermore, is inherently relational, that is, it occurs in the context of a professional relationship in which the fundamental values of the medical profession find expression. Fulfilling the ideal of informed consent in practice is an elusive aim, however. As suggested by intuition, clinical phenomenology, and empirical evidence, it is particularly challenging when consent is sought from individuals who suffer from mental illness. For this reason, it is important to consider the specific issues arising in informed consent in the context of mental health care. This article provides a clinically oriented overview of informed consent in mental illness by: (1) characterizing the core conceptual elements of informed consent; (2) applying informed consent concepts in mental health clinical practice; and (3) giving a brief overview of key empirical data related to mental and physical illness in clinical consent decisions.

Informed consent is a philosophical and legal doctrine that has come to serve as the cornerstone of ethically sound clinical care in Western society. This doctrine requires that individuals truly understand and freely make decisions regarding their treatment. Informed consent is an ideal that emphasizes respect for the individual and places primacy on autonomous decision-making. Furthermore, it is inherently relational, that is, it occurs in the context of a professional relationship in which the fundamental values of the medical profession find expression. Fulfilling the ideal of informed consent in practice is an elusive aim, however. As suggested by intuition, clinical phenomenology, and empirical evidence, it is particularly challenging when consent is sought from individuals who suffer from mental illness. For this reason, it is important to consider the specific issues arising in informed consent in the context of mental health care. This article provides a clinically oriented overview of informed consent in mental illness by: (1) characterizing the core conceptual elements of informed consent; (2) applying informed consent concepts in mental health clinical practice; and (3) giving a brief overview of key empirical data related to mental illness in clinical consent decisions.

Elements of Informed Consent

Informed consent does not occur in isolation or in a moment, nor is it reducible to a signature on a legal document. It is a shared, significant process between people – one of whom faces a series of evolving choices critical to his or her mental and physical health, the other of whom is a professional person entrusted with helping to define these choices astutely and present them truthfully, respectfully, and compassionately. In clinical care, promoting the health and alleviating the suffering of the individual patient will be the basis of any treatment considered in the consent process. Consent in this situation focuses on the patient's personal values, is based on the overall goals for care, and hinges on trust in the clinician–patient relationship.

Information

The sharing of appropriate information is the first of three core elements of informed consent. The patient should be provided with accurate and balanced information in understandable language. This information should include the anticipated risks and benefits of a given treatment, the procedures involved, and its alternatives and their expected outcomes. Alternatives may include implementation of another treatment or additional treatment, or it may involve no intervention upon the course of illness at all. While every minute detail cannot be reviewed in this discussion, rare but serious risks should be characterized in a manner that conveys their relevance and reality. Furthermore, the omission of information in the disclosure process should not be intended to deceive or mislead the person undertaking the decision.

Ideally information related to a matter requiring consent will be disclosed in a manner that fosters dialogue, allows for clarification of complex issues and the identification of misunderstandings, and keeps opportunities open for subsequent discussion as new issues emerge. In obtaining informed consent for treatment of mild depression, for instance, the clinician should discuss the variable course of depressive illnesses, including the potential for symptom resolution in the absence of treatment and the chance of recurrence without intervention. The clinician should review the rationale for a proposed antidepressant agent given the symptoms of the patient and the patient's attitudes and values. The discussion should cover the medication's known possible side effects, potential adverse effects, and their relative likelihood. The clinician should explicitly anticipate the choices that will be faced by the patient in the future in light of the therapeutic alternatives discussed and the nature of the illness. This discussion should be documented by the clinician as an important element of the care process.

Consent forms that accurately and completely describe a treatment and outline patient rights represent an important tool in the consent process. Consequently, consent forms should be developed carefully in terms of their understandability (e.g., simple but precise language), readability (e.g., size of font), content (e.g., purpose, risks and benefits, rights, responsibilities of the clinician/researcher), and related issues. Consent forms, furthermore, signify the importance of a new decision for the patient, and they give specific focus to the dialogue between the professional and the patient. Even general consent forms such as those for hospitalization or for initiating medication – which may or may not confer legal protection when a medico-legal conflict occurs – have this value. From a legal perspective, notes written by the clinician reflecting the disclosure dialogue with the patient may be far more salient and can carry greater weight.

Decisional Capacity

Decisional capacity, the second element of informed consent, represents a sophisticated clinical assessment of an individual's global and specific cognitive abilities within a particular context and in relation to a particular choice. It is distinct from 'competence' which is a legal determination, i.e., a formal assessment by a judge or other officer of the court regarding a person's ability to function within a general domain of one's life such as financial or medical decisions (Appelbaum and Gutheil, 1991). The presence of a serious symptom (e.g., delusions, memory loss, suicidal ideation, diminished cognitive functioning) or of a serious diagnosis (e.g., schizophrenia, dementia, major depression, developmental disability) does not mean that an individual lacks decisional capacity. Similarly, being a young adolescent or a very elderly person does not, *a priori*, preclude one from being decisionally capable. In the context of mental illness, decisional capacity assessment may be subtle, complex clinical work that entails constructing a cohesive sense of the individual's strengths and deficits in light of a number of factors. The type of decision to be made, the nature, time course, and severity of the person's symptoms, the patient's situation, and past experiences, personal values, psychological defenses or coping 'style,' and other factors should all be considered.

Decisional capacity itself has four principal components, each of which should be viewed on a continuum and not as 'all or none' phenomena (Appelbaum and Grisso, 1988). The first component is the individual's ability to express a preference (i.e., capacity for expression or communication). On the extreme end of the continuum will be severely ill psychiatric patients who are mute or catatonic, or who manifest complete language and thought disorganization, and therefore may be intermittently 'incapable' of communication. In the medical context, a patient who is comatose due to metabolic disturbances, aphasic secondary to a stroke, delirious, or unconscious after a head injury will also lack this most fundamental ability. There are times when this component of decisional capacity is very difficult to assess, such as in the spinal cord injured patient who communicates only through eye blinks, or the dysphoric person with Broca's aphasia who is frustrated by the process of generating words but has the capacity to understand and indicate choices. For these reasons, while the ability to communicate is often obvious, this assessment in many situations will require careful interactions with patients and astute observational and listening skills on the part of the clinician.

The second component of decisional capacity is the individual's ability to understand relevant information (i.e., capacity for comprehension). In talking with the clinician, the patient with this ability will be able to take in factual information regarding the nature of the illness, the need for intervention, the kinds of treatments under consideration, their relative risks and benefits, and the possible consequences of not intervening. Comprehension is influenced by intelligence, educational level, and prior experience. In assessing this dimension, it is critical that the patient state his or her understanding of the situation and the choices at hand in their own words. Otherwise it is difficult to know if the patient is using terms and phrases of the clinicians without truly knowing their meaning.

Third is the individual's ability to think through the decision in a way that is 'rational' and deliberate (i.e., capacity for reasoning or rational thinking). This component of decisional capacity relates to the person's analytical abilities given a specific set of facts or data and, like comprehension, is affected by intelligence and educational level. For instance, the 'rational' person, all other factors being equal, will consistently choose a medication that has a 75 % likelihood of treating the illness effectively over one that has a 50 % chance. Similarly, the 'rational' person will consistently choose a medication that poses fewer risks and promises fewer adverse side effects than another medication. In essence, this dimension of decisional capacity has to do with the ability to use objective information.

Fourth, and often most important, is the person's ability to make meaning of the decision and its consequences within the context of his or her personal history, values, emotions, and life situation (i.e., capacity for appreciation). Deeply dependent upon psychological insight, this is the most sophisticated component of decisional capacity. It can be very difficult to assess clinically. In some individuals with early dementia, for instance, a person's social graces and cognitive skills may be only superficially affected and yet these individuals may be unable to think through the meaning of an important treatment decision and its repercussions. In some patients suffering from either anxiety or affective disorders, the negative cognitive distortions inherent to their illnesses may interfere with their abilities to interpret information within a personal context, although other components of decisional capacity may be intact. Importantly, in people with severe mental illnesses such as schizophrenia, the simple presence of psychotic symptoms such as hallucinations or delusions may *not* interfere with the individual's overall decisional capacity, but the cognitive and insight deficits associated with the illness may interfere with their capacity for appreciation of the situation at hand. Alternatively, even the most gravely ill individuals may have difficulty expressing themselves and an incomplete sense of the factual information but may truly appreciate the severity of the situation and possess a deep and consistent sense of who they would like to make decisions for them.

Decisional capacity is therefore a rich construction, not simply reducible to an individual choosing a predetermined 'right answer' when making a decision of importance. It is the

person's ability to gather, assimilate, and use information, to truly see the reasons for and against a given option, and to develop a meaningful decision in the larger context of one's life. The seriousness and complexity of any specific decision will govern how rigorously these criteria are applied. In the context of serious mental illness, a person may be fully capable of deciding whether or not to begin a new psychotropic medication but, on a given day, may be incapable of deciding whether to enroll in a clinical protocol involving a medication-free interval and potential randomization to placebo. The standard for decisional capacity is also robust when a very ill patient refuses treatment that is highly recommended because of the benefit it is likely to bring to him (see discussion in the section Clinical Application of Informed Consent Concepts). This is not to say that patients cannot refuse treatment; when the risks of refusal are great, however, there is a greater responsibility on the caregiver to assure that the patient is genuinely capable of this decision and to manage its consequences.

Voluntarism

Beyond information disclosure and decisional capacity, the final element of informed consent is voluntarism or autonomy. Autonomy represents the expression of the authentic preferences of the individual in the absence of coercive pressures. Voluntary preferences may be tempered by the realities of the situation such as the presence of a serious diagnosis and the lack of 'curative' treatments for many mental illnesses, but should reflect the person's sustained and genuine wishes as evidenced in their expressed values and prior life decisions. Autonomous wishes may be deeply influenced by social and cultural expectations and psychological history. For instance, a patient who has been in the military may be oriented to hierarchical power relationships and may view a clinician's recommendation as an 'order' to be obeyed. Similarly, a seriously mentally ill person who has been institutionalized involuntarily in the past may see themselves as no longer possessing the right to their own choices. Illness processes may also affect patient preferences as well as the ability to express autonomous wishes. A person who suffers deeply either psychologically or physically may feel desperate and desolate, phenomena that can generate inherent internal coercive pressures. A person who has been the victim of violence may see themselves as relatively powerless. Finally, clinical syndromes that involve apathy, alexithymia, amotivation, or cognitive slowing may render ill individuals less able to identify and express their true wishes.

Clinical Application of Informed Consent Concepts

Translation of informed consent concepts to clinical practice is affected by the clinical status and situation of the patient. In emergencies, for instance, there may be little time for information sharing and mutual decision-making and the standard for consent is less rigorous than in other contexts. Acting beneficently, i.e., to save the life or preserve the health of the patient in an acute situation, is the clinician's duty in such circumstances. This is held to be critical ethically and justifiable legally; it is presumed that in such emergency situations the patient wishes to be helped and that the clinician has permission to act with the patient's well-being and best interests at heart. Often the patient, if alert and conscious, will explicitly assent to treatment in these circumstances, i.e., voluntarily accept the choice proffered by the clinician at the time. If the patient is unconscious or is unable to communicate at all, the clinician acts with the 'implied consent' of the patient.

When the patient refuses treatment in an emergent situation, however, a very rigorous standard for consent must be applied because of the likelihood of an adverse outcome and because there are many clinical factors that may be adversely affecting the judgment of the patient. Undiagnosed illness, delusional thinking, severe physical pain, emotional distress, and interpersonal pressure all may undermine the ability of the patient to make a fully informed, autonomous choice in an acute situation. Evaluation, hospitalization, and treatment, even against the apparent wishes of the patient, may be essential in such circumstances. Consider two illustrative patient cases:

Case Example 1

A man diagnosed with a serious mental illness characterized by psychotic and mood symptoms was brought to a psychiatric crisis unit after being found wandering on the street near his home. It was discovered that he had superglued his eyelids shut, and it was imperative that he receive medical attention immediately. He was outwardly calm, did not speak, answered questions through nodding, writing, and hand gestures, but actively refused to get into an ambulance to be taken to the medical emergency room. The clinician on duty that evening knew the patient well and was able to deduce that the patient's peculiar behavior was linked to a recurrent delusion he had when entering a manic period. In this case, the patient's ability to make a fully informed decision was impaired by his delusional belief that if he looked directly at people, they would be instantly destroyed by powerful lasers that had been implanted in his eyes. Once the clinician initiated a conversation about these beliefs with the patient, it was possible to clarify that the patient wanted help and that he was willing to accept treatment, but was terrified that he would hurt people or himself be killed. The patient thus had some decisional ability and could offer assent for emergency care. The clinical team opted to pursue a legal process that would allow involvement of a time-limited treatment guardian due to concerns over the patient's compromised insight and judgment and the severity of his delusions.

Case Example 2

An agitated man demanded discharge from a surgical unit two days after an emergency appendectomy. He had elevated heart rate, blood pressure, and temperature, had signs of abdominal infection, and yet was insistent that he leave the hospital. The medical team was conflicted about granting the patient an 'AMA discharge' ('Against Medical Advice') – the nurses and residents believed this to be the patient's 'right to refuse treatment' but the attending clinician did not agree. On initial psychiatric evaluation, the patient showed irritability and very

subtle indications of diminished attention, was oriented except to day of the week, and had no other apparent abnormalities on an abbreviated mental status examination. He was noted to be tremulous. He was placed on involuntary 'hold' status. Within an hour, the patient was agitated, disoriented, and expressed disorganized thoughts. Additional history was then obtained that the patient drank one to two pints of vodka daily prior to admission. Postsurgical infection in combination with alcohol withdrawal had triggered a significant delirium, an early indication of which, ironically, was his treatment refusal and discharge demand. His most distressing symptoms and disorientation resolved rapidly with treatment, and he withdrew his request to leave the hospital.

When a patient refuses treatment there may be a significant clinical issue that requires close attention, and involuntary treatment may be imperative. Clarification of underlying clinical considerations in such circumstances is the caregiver's primary responsibility – a duty grounded in the bioethics principles of respect for persons, clinical standards of care, fidelity, and beneficence. Careful evaluation of the patient's true strengths and abilities is important so that he or she may exercise choice in every area possible. The intent is not to undermine the patient's autonomy. Rather, it is to empower the clinician only insofar as to help advance the health and well-being of the patient and to assure that the patient is optimally able to express his or her true wishes to accept or decline a given therapeutic approach. In sum, in situations where clinical decisions may have very serious or irreversible health consequences, including death, informed refusal of treatment must fulfill the most stringent standards.

The goal of informed, voluntary decision-making may be more clearly fulfilled in the context of care for chronic illness. Over the course of time, the clinician may come to know the patient – their symptoms, their values and concerns, their responsiveness to treatment interventions, and their personal supports. In turn, the patient may have a greater sense of trust, express themselves with greater openness, and gather more complete information about the illness and potential therapeutic interventions. The patient with a chronic illness also lives out the consequences of side effects, adverse reactions, and benefits of treatments in a manner different from the patient with a time-limited health problem. In this context, the patient may better appreciate how their illness and treatment affect other aspects of their life, and it is possible to come to know their preferences and concerns with greater certainty. Expressing autonomous wishes and deciding upon specific treatment approaches and the overall goals of therapy through mutual understanding and dialogue are more intrinsically possible in the context of chronic illness. The standard for consent in chronic illness, thus, is very rigorous in terms of the patient's level of information, opportunities to enhance decision-making, and creating a circumstance in which patients may indicate their wishes and hopes.

In clinical practice, psychiatric advance directives may be valuable in translating concepts of informed consent to patient care decisions. Advance directives articulate patient preferences for future care. Advance directives are often formal written papers created and signed by the patient; they may also consist of a body of documentation developed by the patient, family members, and clinicians regarding his or her wishes and values concerning personal health care. Advance directives are helpful in that they serve as a reference for the clinician if the patient becomes decisionally incapable or compromised. For example, knowing that they had a pattern of developing compromised judgment during periods of mania, a patient with bipolar disorder completed an advance directive requesting that treatment be instituted despite their (anticipated) protests. Advance directives also provide evidence of the patient's past personal healthcare preferences that may result in controversial clinical decisions, e.g., when patients refuse treatment based on deeply and consistently held personal convictions such as the refusal of blood transfusions by a patient who is a Jehovah's witness. Ideally, these issues will not be addressed for the first time in a moment of great crisis. They should be anticipated, discussed, and documented carefully beforehand. Advance directives do not, under any circumstances, replace the wishes expressed by a patient who is currently decisionally capable. Moreover, there should be opportunities to adapt, or even reverse, earlier decisions as one cannot always tell prospectively how one will respond in a crisis situation. For these reasons, advance directives may guide but can never supplant prudent, reflective clinical decision-making.

In surrogate decision-making, healthcare choices are made for the decisionally impaired patient by another person. This person may be a spouse, sibling, adult child, friend, or a court-appointed guardian. The surrogate decision-maker has the responsibility of serving the best interests and well-being of the patient while seeking to make decisions that are faithful to the personal values and prior wishes of the patient. This can be difficult. In some situations, the patient's wishes are not known. In other situations, the patient is too young or too cognitively compromised to have demonstrated their personal values. In yet other situations, these values and wishes are known and appear to conflict with accepted standards of clinical care. Consequently, in working with surrogate decision-makers, it is important for clinicians to clarify the goals and value system governing the choices made on behalf of the patient. It is also crucial that the patient's decisional capacity be assessed accurately and, in most circumstances, the patient's symptoms be treated aggressively so that they can contribute to the choices made to whatever extent possible.

Empirical Studies on Consent in Clinical Care

Nearly 400 studies performed since 1970 suggest that informed consent is an ideal that is often difficult to achieve in clinical and research practice involving people with physical and mental illness (Sugarman et al., 1999). These data indicate, however, that issues faced in clinical care for mental illness are not radically different in degree or kind from those encountered in the treatment of physical illness (Stanley et al., 1988). Moreover, an extensive body of evidence on consent consistently demonstrates the poor readability, understandability, and ineffectiveness of formal consent documents in clinical care (Grossman et al., 1994). An important recent advance in the informed consent field has been the development of the MacArthur Competence Assessment Tool for Treatment (MacCAT-T) (Grisso et al., 1997). This is a structured interview designed to assess patients' clinical care decision-making

abilities based on empirical work conducted over 10 years by Appelbaum and N Grisso. While this instrument is best utilized as a tool for approaching decisional capacity evaluation, not as a definitive 'test' of competence, it will introduce greater consistency of consent standards to both scientific and clinical arenas.

Evidence on consent in the care of mental illness reveals many key challenges for clinicians (Roth et al., 1982). Studies of psychotic, affective, and dementia disorders, for instance, have shown that symptoms of mental illness may adversely impact information-based or more strictly cognitive aspects of consent (Roberts, 1998). In the 1995 MacArthur Treatment Competence Study (Appelbaum and Grisso, 1995), the decisional abilities of patients with schizophrenia or major depression, patients with ischemic heart disease, and non-patient community volunteers (total n=498) were assessed. When acutely ill, the psychiatric patients in this study had greater difficulties with decision-making on several formal cognitive measures than did the medically ill patients and community participants. The problems were more severe for individuals with psychotic symptoms than for depressed individuals. Clinical treatment led to significant improvement of these deficits, resulting in closer resemblance across the three groups.

Smaller studies echo these findings. An early study by Schachter et al. (1994) involving 38 schizophrenia patients revealed that the presence of conceptual disorganization, delusions, and hallucinations was associated with patients' poor performance on the information elements of consent. Appelbaum et al. (1981) assessed 50 people admitted to a mental health inpatient unit and found that most were unable to consent to hospitalization. Later this team evaluated 100 voluntarily hospitalized psychiatric patients, finding that a significant minority did not understand information presented to them on the purpose of hospitalization and discharge (Appelbaum et al., 1998). Beck (1988) evaluated the capacities of 56 psychiatric patients on antipsychotic medications, revealing that they had limited understanding and appreciation of their treatment. Stanley et al. (1988) compared the reasoning ability of elderly individuals suffering from major depression (n=45) and dementia (n=38) with elderly control subjects (n=20). The cognitively impaired dementia patients and, to a lesser extent, the depressed patients had significant deficits on a number of decision-making items. Similarly, Marson et al. (1995) used a clinical vignette-based instrument to measure the level of competence exhibited in medical treatment decisions by individuals with dementia and healthy comparison subjects. Not unexpectedly, this work revealed that the severity of cognitive impairment in dementia correlated with lack of ability to formulate logical reasons for treatment choice, i.e., addressing the 'rational reasons' dimension of decisional capacity. Interestingly, global measures of dementia severity such as the Mini-Mental Status Exam were not predictive of clinical decision-making capacity.

Studies of advance directives and surrogate decision-making also have important implications for mental health care. Several survey and interview studies have revealed that advance directives are underutilized (Bradley et al., 1998) and that patient preferences are rarely identified prospectively (Seckler et al., 1991). In terms of surrogate decision-making, several studies have shown that most elderly patients would prefer that a family member such as a spouse or adult child serve as an alternative decision-maker (High, 1989). Studies of hypothetical decision-making by proxies suggest that they may not assess patients' wishes accurately, however Sachs et al. (1994) studied 42 decisionally capable dementia patients, 64 proxy decision-makers, and 60 well elderly individuals and found that the decision-makers' predictions of patient preferences were overall discordant with the patients' own stated wishes. Similarly, Warren et al. (1986) studied participation choices made by proxy decision-makers and 168 decisionally capable nursing home patients with respect to hypothetical, minimal risk research projects. The results demonstrated that the surrogate decision-makers used criteria other than the patient's predicted preferences in making these choices.

Clear evidence is also emerging that the barriers to consent may be amenable to intervention, offering significant hope (Roberts et al., 2000). Interventions in which information is presented verbally and in writing, using clear and understandable language, have been shown to improve patients' understanding of consent decisions (Carpenter et al., 2000). The timing of consent information and the provision of information in a relaxed manner or on repeated occasions is important to patients' abilities to integrate and apply this data (Roberts, 1998). Treatment of symptoms, ranging from hallucinations to physical pain, has been observed to improve the quality of patients' decision-making. Finally, clinician behaviors that reflect how the consent process is valued, e.g., by giving time to the consent dialogue, have been found to be valued by patients and surrogate decision-makers (Roberts and Roberts, 1999). The observation that educational efforts, symptom treatment, and attitudinal obstacles can be addressed suggests a number of constructive approaches to consent in the context of clinical care for the mentally ill.

See also: Civil Rights; Ethical Dilemmas, Research, and Access to Treatment; Ethics and Psychiatry; Free Will and Action; Informed Consent of Research Subjects; Normality and Mental Health; Nosology in Psychiatry; Privacy of Individuals in Social Research: Confidentiality; Responsibility: Philosophical Aspects; Social Work Ethics; Sociology and Psychiatry.

Bibliography

Appelbaum, B.C., Appelbaum, P.S., Grisso, T., 1998. Competence to consent to voluntary psychiatric hospitalization: a test of a standard proposed by APA. American Psychiatric Association. Psychiatric Services 49, 1193–1196.

Appelbaum, P.S., Grisso, T., 1988. Assessing patients' capacities to consent to treatment. New England Journal of Medicine 319, 1635–1638.

Appelbaum, P., Grisso, T., 1995. The MacArthur Competence Study I, II, III. Law and Human Behavior 19, 105–174.

Appelbaum, P., Gutheil, T., 1991. Clinical Handbook of Psychiatry and the Law, second ed. Williams & Wilkins, Baltimore, MD.

Appelbaum, P., Mirkin, S., Bateman, A., 1981. Empirical assessment of competency to consent to psychiatric hospitalization. American Journal of Psychiatry 138, 1170–1176.

Beck, J., 1988. Determining competency to consent to neuroleptic drug treatment. Hospital Community Psychiatry 39, 1106–1108.

Bradley, E.H., Peiris, V., Wetle, T., 1998. Discussions about end-of-life care in nursing homes. Journal of the American Geriatric Society 46, 1235–1241.

Carpenter Jr, W.T., Gold, J.M., Lahti, A.C., et al., 2000. Decisional capacity for informed consent in schizophrenia research [see comments] Archives of General Psychiatry 57, 533–538.

Grisso, T., Appelbaum, P.S., Hill-Fotouhi, C., 1997. The MacCAT-T: a clinical tool to assess patients' capacities to make treatment decisions. Psychiatric Services 48, 1415–1419.

Grossman, S.A., Piantadosi, S., Covahey, C., 1994. Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? Journal of Clinical Oncology 12, 2211–2215.

High, D.M., 1989. Standards for surrogate decision making: what the elderly want. Journal of Long Term Care Administration 17, 8–13.

Kaufmann, C.L., Roth, L.H., Lidz, C.W., Meisel, A., 1981. Informed consent and patient decisionmaking: the reasoning of law and psychiatry. International Journal of Law and Psychiatry 4, 345–361.

Lambert, P., Gibson, J.M., Nathanson, P., 1990. The values history: an innovation in surrogate medical decision-making. Law, Medicine, and Health Care 18, 202–212.

Lidz, C., 1984. Informed Consent: A Study of Decisionmaking in Psychiatry. Guilford Press, New York.

Marson, D.C., Cody, H.A., Ingram, K.K., Harrell, L.E., 1995. Neuropsychologic predictors of competency in Alzheimer's disease using a rational reasons legal standard. Archives of Neurology 52, 955–959.

Roberts, L.W., 1998. The ethical basis of psychiatric research: conceptual issues and empirical findings. Comprehensive Psychiatry 39, 99–110.

Roberts, L., Roberts, B., 1999. Psychiatric research ethics: an overview of evolving guidelines and current ethical dilemmas in the study of mental illness. Biological Psychiatry 46, 1025–1038.

Roberts, L.W., Fraser, K., Quinn, D., McCarty, T., Severino, S.K., 1999. Persons with mental illness. In: Bennahum, D. (Ed.), Managed Care: Financial, Legal, and Ethical Issues. Pilgrim Press, Cleveland, OH.

Roberts, L.W., Warner, T.D., Brody, J.L., 2000. Perspectives of patients with schizophrenia and psychiatrists regarding ethically important aspects of research participation. American Journal of Psychiatry 157, 67–74.

Roth, L.H., Lidz, C.W., Meisel, A., Soloff, P.H., Kaufman, K., Spiker, D.G., Foster, F.G., 1982. Competency to decide about treatment or research: an overview of some empirical data. International Journal of Law and Psychiatry 5, 29–50.

Sachs, G.A., Stocking, C.B., Stern, R., Cox, D.M., Hougham, G., Sachs, R.S., 1994. Ethical aspects of dementia research: informed consent and proxy consent. Clinical Research 42, 403–412.

Schachter, D., Kleinman, I., Prendergast, P., Remington, G., Schertzer, S., 1994. The effect of psychopathology on the ability of schizophrenic patients to give informed consent. Journal of Nervous and Mental Disorders 182, 360–362.

Seckler, A.B., Meier, D.E., Mulvihill, M., Paris, B.E., 1991. Substituted judgment: how accurate are proxy predictions? [see comments] Annals of Internal Medicine 115, 92–98.

Stanley, B., Stanley, M., Guido, J., Garvin, L., 1988. The functional competency of elderly at risk. Gerontologist 28 (Suppl.), 53–58.

Sugarman, J., McCrory, D.C., Powell, D., Krasny, A., Adams, B., Ball, E., Cassell, C., 1999. Empirical research on informed consent. An annotated bibliography. Hastings Center Report 29, S1–S42.

Warren, J.W., Sobal, J., Tenney, J.H., Hoopes, J.M., Damron, D., Levenson, S., Deforge, B.R., Muncie Jr, H.L., 1986. Informed consent by proxy. An issue in research with elderly patients. New England Journal of Medicine 315, 1124–1128.

Psychoanalysis, History of

Elke Mühlleitner, Giessen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sigmund Freud's biography, the development and interpretation of psychoanalysis constitute an important chapter in twentieth-century cultural and intellectual history both as a body of theory and as cultural practice. An ongoing attempt to determine the meaning of Freud and psychoanalysis has led to the publication of hundreds of volumes, and the historiography of psychoanalysis has found its own place within the history of science. Historical literature about psychoanalysis ranges from the self-representations and autobiographical works of Freud to the first official biography, from moderate revisionism and contextualized approaches to the harsh, and at times hostile, clinical, feminist, epistemological, hermeneutical, and psychobiographical critiques. A general history of psychoanalysis as a single science can be cataloged in three ways: (1) as a history of the psychoanalytic theories from Freud to the present; (2) as a history of the psychoanalytic movement, its institutions and practices; and (3) as a self-reflection on the historiography of psychoanalysis itself.

Sigmund Freud's biography, the development and interpretation of psychoanalysis constitute an important chapter in twentieth-century cultural and intellectual history both as a body of theory and as cultural practice. Each culture and time has absorbed Sigmund Freud (1856–1939) and psychoanalysis differently. A general history of psychoanalysis can be cataloged in three ways: (1) as a history of the psychoanalytic theories from Freud to the present; (2) as a history of the psychoanalytic movement, its institutions and practices; and (3) as a self-reflection on the historiography of psychoanalysis itself.

A History of the Psychoanalytic Theories from Freud to the Present

The development of Freud's theory of the unconscious and his psychotherapeutic method can be understood on the basis of his professional development in the context of a process of differentiation and specialization, in particular, medicine and philosophy, in the late nineteenth century. Historical works of the kind, for example, studies of the scientific and social context and the institutional structure of the field, have created a methodologically sophisticated and broad view of psychoanalysis. Various types of hermeneutical studies of psychoanalysis itself, that is, visions of it as a system of symbols and as a method for culture-decoding rather than as a type of natural science, have broadened the view of the history of psychoanalysis. Jean Laplanche and Jean-Bertrand Pontalis's *Vocabulaire de la Psychanalyse* (1967), the first conceptual dictionary of psychoanalysis, is based on the techniques of the history of philosophy and the French tradition of the epistemology of the sciences. As the basic concepts of the psychoanalytic theory have undergone changes that reflect the development of the sociopolitical climate, the authors have incorporated the later concepts of Freud and his followers as well as critics who created new theories and schools of psychoanalysis (e.g., Melanie Klein, Jacques Lacan). Their attempts have stimulated a whole universe of separate hermeneutical and epistemological studies, and especially, with the centenary of the birth of psychoanalysis various encyclopedic works and summaries have been published (e.g., Erwin, 2002; Skelton, 2006).

Freud's psychoanalysis with its concepts *subjectivity* and the *scientific empirical method* has been interpreted as a characteristic product of Viennese modernity, and has been seen as a parallel to phenomena like Zionism and Austromarxism. Bourgeois society and European modernity experienced significant crises in Vienna of the *fin de siècle*, where psychoanalysis was founded in the last decades of the Austro-Hungarian Monarchy. During the period between 1890 and 1910, extraordinary achievements were made in the areas of music, literature, art, and science. These were fundamentally influential for new developments in observation and thought in the twentieth century, and the psychoanalytic theory enjoyed its heyday in the years before World War I. The following factors were generative forces behind the *fin de siècle* creativity: the rise of new cultural founding figures of the liberal bourgeoisie, the influence of the Jewish intelligentsia, and the upward mobility of immigrants from the Austro-Hungarian Empire; the resistance against the political control, bureaucracy and the repression of the Roman Catholic Church and censorship; the cultural networks established with European centers and circles abroad; the import of foreign ideas; and the intellectual exchange between the submetropoles of the countries of the monarchy (Nautz and Vahrenkamp, 1993).

From 1859 until 1938, Sigmund Freud lived and worked in Vienna, and the significance of Vienna as the site of psychoanalysis' foundation and development is evident. Socio-politico-cultural studies have dealt with Freud's *fin de siècle* (see Schorske, 1981; McGrath, 1986), viewing psychoanalysis within Viennese culture, for example, at the juncture of a specific time and place. Special interest has focused on Freud's Jewish identity and the relationship between Judaism and psychoanalytic theory. Psychoanalytic theory and Freud's biography were and in many ways still are intertwined in a unique way. Therefore, the historian is in the position of having to treat psychoanalytic concepts as both strictly personal documents and as documents connected to a broader theoretical system. In Freud's early texts, the evolution of psychoanalytic theory is presented linearly, as though it was fundamentally the same as any other scientific field, but there are serious historiographical difficulties entailed in approaching psychoanalysis in this fashion. Freud's self-analysis, his

autobiographical writings, and self-created image of psychoanalysis as a science and various theoretical concepts of psychoanalysis (e.g., seduction theory, dream interpretation, Oedipal conflict, resistance, repression, transference, and countertransference) have become prominent in the field of psychoanalytic historiography, including Freud's own changes of theory throughout his life, as well as post-Freudian developments (e.g., drive-psychology, ego-psychology, object-relations theory, attachment theory, the development from defense to adaptation, from instinctual to motivational diversity, the development of the neurosciences).

A History of Freud's Biography, the Psychoanalytic Movement, and Its Institutions and Practices

In the historiography of psychoanalysis, the genre of Freud studies has been differentiated from the historical examination of the psychoanalytic movement in its local national centers e.g., Berlin, London, or Paris and within special cultures e.g., Japan or India. Biographical studies dominate over sociohistorical works, and the focus has been for a long time on the struggle over legacy and legitimacy rather than on a social history of the field.

Freud Biographies

Three groups of historical literature about Freud have been described by historians: (1) publications of primary sources, for example, Freud's correspondences; (2) publications of historical materials concerning case studies and theoretical works of Freud; and (3) critical-academic studies about psychoanalysis. The history of Freud biography and the psychoanalytic movement began with Freud's autobiographical works *On the History of the Psychoanalytic Movement* (1914) and *An Autobiographical Study* (1925), and in his own historical papers, Freud – in his one-sided but coherent story – suggests that the history of psychoanalysis is basically the history of refusals, resistances, and the regroupings of the orthodox Freudian ideas, institutions and people. History of psychoanalysis also appears at times to be synonymous with the history of splits, schisms, and dissidents. But, just as the proliferation of schools and the consequent battles among and splittings within psychoanalytic institutes stimulated abundant polemics, these upheavals also stimulated reworkings of almost every important item in the history of Freud biography.

Dreams from *The Interpretation of Dreams*, patients from *The Studies on Hysteria* and later case studies have been analyzed and reanalyzed by scholars from new perspectives; all the schisms and quarrels in the history of psychoanalysis were rediagnosed by the latter-day schismatics and schism-haters. Based on newly accessed archival material and historical information, early speculations and widely differing interpretations have been reconsidered. Freud biographies have addressed almost every aspect of Freud's life (profession, social status, political views, culture, religion, affection-emotion), sometimes in a general fashion, sometimes in a very specific manner, and reflecting the time and the standpoint of their authors. To forestall future falsifications in popular biographies and early attempts at studying Freud's biography by his pupils and followers, Freud's youngest daughter Anna Freud, a psychoanalyst herself, felt prompted to authorize Ernest Jones to write a first full-scale biography (Young-Bruehl, 1998). Freud studies properly began with the pioneering research of Siegfried Bernfeld; his scholarship was invaluable to Ernest Jones when putting together his trilogy *The Life and Work of Sigmund Freud* (1953–1957). As a member of the first generation of loyal followers, Jones' monumental work became the precipitant for a third type of historical accounts: critiques of Jones's biography, or what might be called second-order biographies.

For a long time, Freud's prominence nurtured the illusion of an isolated progression of psychoanalysis apart from sociopolitical developments as well as medicoscientific trends at the time. A cultural–historical approach to Freud became more intense as the 1960s unfolded. With Henri Ellenberger's *The Discovery of the Unconscious* (1970), a mode of Freud biography contextualizing Freud began to flourish, taking the precursors of Freud's ideas more into account than Jones did.

The Psychoanalytic Movement

Interest in supporters, students, and contributors of Freud has long been marginal or limited to only the most well-known representatives. With the publication of *The Minutes of the Vienna Psychoanalytic Society* (Nunberg and Federn, 1962–1975), a new focus was put on the institutional framework, the form in which these scientists cooperated, and how they organized their professional and personal exchanges. Works on Freud's friendships and professional relationships include the dynamical and sociological aspects of the formation of groups, the production of scientific knowledge, the structure of mastery and slavery (often basic to the formation of psychoanalytic institutions), as well as more contextually embedded accounts of specific psychoanalytic cultures. Further, *Freud and his Followers* (Roazen, 1975) stimulated studies of the family networks and subnetworks; as a consequence, biographies of early pupils and later analysts have emerged. Their adherent or schismatic relations to Freud figured largely in these accounts. A special body of historical literature has focused on the role and influence of female patients and analysts. The founding of new psychoanalytic schools has created a new intellectual–historical approach, including the professionalization and institutionalization of the discipline in different cultures and at different times. From the early days of psychoanalysis, the internationalization of psychoanalysis had its own motor and effect on theory and practice (e.g., Kutter, 1992, 1995; Makari, 2008). Taking geographical and temporal particularities into account, a Western (European and North American) perspective has been criticized.

Economic and political circumstances, as well as cultural and sociohistorical connections, have all influenced the theoretical, practical, and organizational developments of psychoanalysis as a discipline and are part of the production of psychoanalysis as a corpus of knowledge. Fenichel (1998) has observed the impacts of social and political reality on the production of psychoanalytic theory and practice during the years when Central European Jewish psychoanalysts, candidates in training and their families were forced to flee National Socialist Germany and Austria and the occupied

countries in Europe. The production of scientific findings and the transformation and transfer of knowledge as social processes in which internal and external factors are in a constant exchange does not minimize the role of the founder of the science of the unconscious in its beginning, but it destroys the illusion of a single isolated discovery. The shifts in the history of psychoanalysis can be seen and interpreted as a complex, dialogical, multiauthor enterprise and depend on the studies of the special scientific tradition and the study of psychoanalytic microcultures (see Hale, 1995; Lockot, 1985; Roudinesco, 1982–1986).

A Self-Reflection on the Historiography of Psychoanalysis itself, and the Problems of Methodology

The history of psychoanalysis has been a rapidly growing, controversy-ridden, and attention-attracting area. Over the past 40 years, each of the historical visions has inspired and/or provoked a revision, and each revision a counterrevision. The history of psychoanalysis has to take into account the peculiar, complex, and powerful interplay among psychoanalysis, historiography, and ideology; because this has not been recognized, many widely disparate accounts of psychoanalytic history have been proffered. Neighboring fields, such as the history of medicine, neuroscience, philosophy, psychology and science, women's studies, the history of law, crime and deviance, the history of professions, the history of sexuality, the history of the body, and cultural studies, literature, queer/gay/gender studies, have had their impacts on the historiography of psychoanalysis as well.

Studies in the area of Freud biography and the pre- and early history of psychoanalysis have received more attention. Because connections with later periods seem less powerful and convincing, the fixation on Freud himself and, even more particularly, on the first decades of Freud's discoveries have meant that later developments during and especially since Freud's lifetime are achieving less attention.

The various schools of psychoanalysis have fashioned for themselves idealized genealogies, lines of intellectual filiation between great minds of the past and contemporary practitioners, tracing a suitable line of evolution, and have been selective of their usable past. They may also assist in sharpening contested interdisciplinary or intradisciplinary boundaries in which 'positive' episodes are memoralized, taken as characteristic of the essence, while 'negative' ones are omitted, serving to legitimate particular theories and methodologies and delegitimizing others. Today, scholars have come to the opinion that historical accounts of psychoanalytic theory and practice should include thoughts and speculations about the special acquisition and transmission of psychoanalytic knowledge.

A 'coherent' picture of Freud and the psychoanalytic movement has undergone a profound change. In the historiography of psychoanalysis, at least two different schools of thought have been described: the 'official' history of Freud, which has been largely though not exclusively produced and consumed by practicing psychoanalysts, and a 'revisionist' approach that has emerged in the last decades and has been mainly produced by people working outside of institutional psychoanalysis (e.g., historians, psychologists, philosophers, sociologists, political scientists). For a long time, the two groups have debated separately – with a trend to specialization and fragmentation thereby repeating the fragmentation of the object they are analyzing – and have tried to avoid other findings. But lately, a dialogue between historians and clinicians – both interested in history – concerning the history has arisen, and there have been attempts to overcome prejudices. One outcome of this dialogue has been the formation of local and international organizations and discussion groups and the establishment of special journals and periodicals as well as various new platforms on the Internet. Still, the issues of language, including the problems of translation and the questions concerning authorship and audience remain central, as in other scientific disciplines. Language questions have been correlated with speculative histories on Freud and the psychoanalytic movement.

After the height of the professional prestige of psychoanalysis, when there was remarkably little criticism from historians of psychiatry, a revisionist historiography of psychoanalysis was stimulated by e.g., historians, sociologists, philosophers, literary scholars, feminist authors, cultural scientists, and public intellectuals; traditional biography and the general longitudinal narrative were followed by the study of institutions introducing the ideal of objective, systematic, and quantitative analysis, having in mind methodologically sophisticated, empirically substantial, and sociologically oriented scholarship. On the other hand, as scholars have argued, the demystifications of hagiographical idealizations have often only generated 'heroic' neo-Marxist and Foucauldian remystifications. Revisionist and hagiographical literature have been diagnosed by contemporary scholars as both reductionistic and lacking self-reflection.

Psychoanalysis as a theory and a profession emphasizes the value and virtue of self-awareness and in its therapeutical practices asks for an introspective mode of inquiry nevertheless, with a few exceptions, it has remained resolutely unintrospective about the methodological and ideological conditions under which it writes the narrative of its own past. That is all the more peculiar in light of the past decade's extensive and sophisticated discussion concerning different conceptual, methodological, and epistemological strategies for writing the history of science generally. Two approaches have been observed, the internalist approach, stressing cognitive development, and the externalist approach, focusing on the social, economic, political, and professional determinants of scientific theory and practice. All history writing has been diagnosed perspectively as the expression of a particular historical environment, culture, background, and self-consciousness among cultural-critics. LaCapra (1988) emphasizes the Freudian point that any historian's relationship with the past, and through that with the present, is a transference relationship, and that, in this transference, repetition and denial come more easily than working through and mourning (Smith and Morris, 1992; Young-Bruehl and Schwartz, 2011)

Jones' idealization of Freud has been rejected, and, particularly in the medium of psychobiography, a much more contextualized, conflicted, and complicated founder of psychoanalysis has emerged in the historical literature. Jones

did not interpret blanks in his documentation as secrets of resistances, but nowadays psychologically informed biographers consider the fact that many of the most important things in people's lives are never recorded.

Contrary to the sophisticated histories of Freud's unfolding discoveries, historians have lamented the difficulties of psychobiography, and some of the key methodological problems concerning childhood experiences, the workings of the mind, and the character of thought, have been widely recognized – although such recognition has not yet informed a major historiographical work on psychoanalytic biography and history writing.

Despite the progressive de-Freudianization of psychiatry and psychology in general, Freud's place as one of the supreme minds of the twentieth century remains secure for most historians. Psychoanalysts have altered psychoanalytic theory over the course of time, their theoretical and methodological findings depending on their analysands, their social context, and the historicopolitical changes in general. Historiography has altered accordingly, considering and tolerating the complexity instead of simplistic and popular versions of psychoanalysis.

See also: Cognitive Psychology: History; Freud, Sigmund (1856–1939); Gender and Feminist Studies in Geography; Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Personality Disorders; Personality Theory and Psychopathology; Psychoanalysis in Sociology; Psychologism, History of; Psychotherapy, History of: Psychiatric Aspects; Social Psychology; Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts; Unconscious: History of the Concept.

Bibliography

Ellenberger, H.F., 1970. The Discovery of the Unconscious. The History and Evolution of Dynamic Psychiatry. Allen Lane, London.

Erwin, E. (Ed.), 2002. The Freud Encyclopedia. Theory, Therapy, and Culture. Routledge, New York, London.

Fenichel, O., 1998. In: Mühlleitner, E., Reichmayr, J. (Eds.), 119 Rundbriefe Europa (1934–1938), Amerika (1938–1945). Stroemfeld Verlag, Basel, Frankfurt am Main.

Forrester, J., 1994. A whole climate of opinion: rewriting the history of psychoanalysis. In: Micale, M.S., Porter, R. (Eds.), Discovering the History of Psychiatry. Oxford University Press, New York, Oxford.

Freud, S., 1914. Zur Geschichte der psychoanalytischen Bewegung. In: Gesammelte Werke X. S Fischer, Frankfurt am Main [1957. On the history of the psychoanalytic movement. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud XIV. Hogarth Press, London].

Freud, S., 1925. Selbstdarstellung. In: Gesammelte Werke XIV. S Fischer, Frankfurt am Main [1959. An autobiographical study. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud XX. Hogarth Press, London].

Grubrich-Simitis, I., 1993. Zurück zu Freuds Texten. Stumme Dokumente sprechen machen. S Fischer, Frankfurt am Main [1996. Back to Freud's Texts: Making Silent Documents Speak. Yale University Press, New Haven, CT].

Hale Jr., N.G., 1995. The Rise and Crisis of Psychoanalysis in the United States: Freud and the Americans, 1917–1985. Oxford University Press, New York.

Haynal, A., Falzeder, E. (Eds.), 1994. 100 Years of Psychoanalysis: Contributions to the History of Psychoanalysis. H. Karnac, London.

Jones, E., 1953–1957. The Life and Work of Sigmund Freud, first ed., 3 vols. Basic Books, New York.

Kearns, K., 1997. Psychoanalysis, Historiography, and Feminist Theory. The Search for a Critical Method. Cambridge University Press, Cambridge, UK.

Kutter, P. (Ed.), 1992, 1995. Psychoanalysis International: A Guide to Psychoanalysis throughout the World. Frommann-Holzboog, Stuttgart, Bad Cannstadt.

LaCapra, D., 1988. History and psychoanalysis. In: Meltzer, F. (Ed.), The Trail(s) of Psychoanalysis. University of Chicago Press, Chicago.

Laplanche, J., Pontalis, J.-B., 1967. Vocabulaire de la Psychanalyse. Presses Universitaires de France, Paris [1973. The Language of Psychoanalysis. Hogarth Press and the Institute of Psycho-Analysis, London].

Lockot, R., 1985. Erinnern und Durcharbeiten: Zur Geschichte der Psychoanalyse und Psychotherapie im Nationalsozialismus. Fischer Taschenbuch Verlag, Frankfurt am Main.

Lorenzer, A., 1984. Intimität und Soziales Leid. Archäologie der Psychoanalyse. S Fischer, Frankfurt am Main.

Loewenberg, P., 1995. Fantasy and Reality in History. Oxford University Press, New York.

Makari, G., 2008. Revolution in Mind. The Creation of Psychoanalysis. HarperCollins, New York.

McGrath, W.J., 1986. Freud's Discovery of Psychoanalysis. Cornell University Press, Ithaca, NY.

Mühlleitner, E., 1992. Biographisches Lexikon der Psychoanalyse. Die Mitglieder der Psychologischen Mittwoch-Gesellschaft und der Wiener Psychoanalytischen Vereinigung, 1902–38.

Nautz, J., Vahrenkamp, R. (Eds.), 1993. Die Wiener Jahrhundertwende. Einflüsse, Umwelt, Wirkungen. Böhlau, Wien.

Nunberg, H., Federn, E., 1962–1975. The Minutes of the Vienna Psychoanalytic Society, 4 vols. International Universities Press, New York.

Roazen, P., 1975. Freud and His Followers. Knopf, New York.

Roudinesco, E., 1982–1986. La bataille de cent ans. Histoire de la psychoanalyse en France, 2 vols. Editions Ramsay, Paris [1990. Jacques Lacan & Co: A History of Psychoanalysis in France 1925–1985. University of Chicago Press, Chicago].

Schorske, C.E., 1981. Fin-de-Siècle Vienna. Vintage Books, New York.

Skelton, R. (Ed.), 2006. The Edinburgh International Encyclopedia of Psychoanalysis. Edinburgh University Press, Edinburgh.

Smith, J., Morris, H. (Eds.), 1992. Telling Facts. History and Narration in Psychoanalysis. Johns Hopkins University Press, Baltimore, MD.

Young-Bruehl, E., 1998. Subject to Biography: Psychoanalysis, Feminism, and Writing Women's Lives. Harvard University Press, Cambridge, MA.

Young-Bruehl, E., Schwartz, M., 2011. Warum die Psychoanalyse keine Geschichte hat [Why psychoanalysis has no history]. In: Psyche Jg, vol. 65.

Psychoanalysis in Sociology

Anthony Elliott, University of South Australia, Adelaide, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In what follows the appropriation of psychoanalysis in sociological theory is charted, with four key sociological traditions considered: first, several early and influential articulations of psychoanalysis and sociology developed in Europe (particularly as represented in the work of the Frankfurt school of sociology) and North America (especially in the writings of Talcott Parsons); second, structuralist and poststructuralist appropriations of psychoanalysis as refracted in sociological debate; third, feminist and postmodern social theory; and finally, more contemporary developments within European psychoanalytic traditions that have influenced the social sciences and humanities.

This article addresses the complex, and often ambivalent, relationship that psychoanalysis has had with sociology. The psychoanalytic tradition, beginning with Freud, is critically examined, primarily in terms of the impact of Freud's theory of the unconscious for theorizing both individual and social structural aspects of the contemporary epoch in sociological research. The appropriation of psychoanalysis in sociological theory is then charted, with four key sociological traditions considered: first, several early and influential articulations of psychoanalysis and sociology developed in Europe (particularly as represented in the work of the Frankfurt school of sociology) and North America (especially in the writings of Talcott Parsons); second, structuralist and poststructuralist appropriations of psychoanalysis as refracted in sociological debate; third, feminist and postmodern social theory; and finally, more contemporary developments within European psychoanalytic traditions that have influenced the social sciences and humanities.

Psychoanalysis, as developed by Sigmund Freud and his followers, has had a major impact upon sociological theory and modern sociology. Freud's central discoveries – the unconscious, sexual repression, the Oedipus complex, and the like have been deployed by sociologists to interpret and discuss the self and human subjectivity, gender and sexuality, the family and socialization, language and ideology, as well as the formation of cultural identities and forms of political domination. Notwithstanding this impact, sociology as a discipline has long had a difficult, and indeed fraught, relationship with psychoanalysis. Sociologists have criticized psychoanalytical theory on the grounds of its methodology, epistemology, and ontology. Despite these criticisms, many contemporary sociologists remain engaged with, and some strongly committed to, the psychoanalytic tradition in order to conceptualize the relation between the individual and society, especially the complex, contradictory ways that human subjects acquire and reshape the ideas, values, symbols, beliefs, and emotional dispositions of the wider society. This has been particularly evident over recent decades, in which Freudian themes and psychoanalytic motifs have been used to analyze sexual politics, issues of identity and lifestyle, as well as the debate over modernity and postmodernism.

Freud and the Sociological Imagination

Freudian psychoanalysis shares with sociology a primary preoccupation with the fate of the individual or self in the context of social relationships and the wider cultural process. While Freud's own writings were primarily derived from his clinical work with patients, and to that degree are at variance with the core methodologies of mainstream social science, his characterization of human personality and self-identity has clear parallels with, say, Thomas Hobbes' theory of human nature or Karl Marx's account of the self-seeking individual in organized capitalist society. Yet whereas both Hobbes and Marx stress the impact of social forms in the constitution of the self, Freud's methodological starting point is the individual psyche, principally the instinctual impulses and libidinal longings that shape the human imagination. The self for Freud is radically fractured or divided, split between consciousness of identity and the repressed unconscious. The biographical trajectory of the self, according to Freud, is carried on against the backdrop of this radical otherness of the unconscious, a domain of the psyche that infuses three agencies of the psyche: id, ego, and superego. The id, lying at the root of unconscious desire, is that which cannot be symbolized yet constantly strives for expression in our daily lives, manifesting itself in dreams, daydreams, slips of the tongue, and the like. It is a hidden area of the self, which knows no reality, logic, or contradiction; the unconscious is at the root, says Freud, of how people, for example, can simultaneously love and hate their parents. The superego, as internal anchoring of cultural prohibition, is founded in this id. Desire, according to Freud, infiltrates all human intentions, ideals, and imperatives. Similarly, Freud sees parts of the ego as interwoven with the force of the id, the self arising as a product of the unconscious. While Freud's writings on human personality and the constitution of the self have proven of interest to sociologists, it is his account of the relations between self and society – primarily his late writings on civilization – that has most influenced the development of modern sociology. In his late writings, Freud comes to see human beings living under the destructive force of a terrifying death drive, as well as strict cultural prohibitions on sexual desire and enjoyment. These themes are set out in his

magisterial books *Beyond the Pleasure Principle* (1920) and *Civilization and its Discontents* (1930). Civilization, writes Freud, is repressive. Society imposes severe psychic demands upon individuals to achieve cultural conformity, demands that produce intense personal misery and neurotic suffering. Some sociologists have interpreted this aspect of Freud's sociology as at one with the conservative social thought of Hobbes, likening it to the theorem that cultural reproduction demands the overcoming of chaotic passion. But it is clear that this involves a radical misunderstanding of Freud's position. The dualisms that Freud uses to analyze modern civilization (consciousness/unconscious, desire/repression, the pleasure principle/the reality principle, Eros/Thanatos) suggest a disjunction between self and world. According to Freud, ambivalence is at the core of the individual's relation to the self, to others, and to society.

Freud's writings on the fate of the self in contemporary culture have strongly influenced sociological debates – from Herbert Marcuse to Michel Foucault. Too much repression, Freud says, leads to intense unconscious anguish, hostility, and rage. At such a point, the intensification of unconscious desire can release the 'mental dams' of sexual repression in a far-reaching way. The issue of the subjective seeds of social and political transformation are thus at the heart of Freud's theoretical contribution to sociology and social theory.

The Integration of Psychoanalysis and Sociology

With regard to social analysis and sociological theory, Freudian ideas have loomed large in the sociological conceptualization of human subjectivity and interpersonal relationships, sexuality and modern culture, as well as the mix of reason and irrationality in politics and history. There have been three key sociological developments in this connection: first, there has been several early and influential articulations of psychoanalysis and sociology developed in Europe (particularly as represented in the work of the Frankfurt school of sociology) and North America (especially in the writings of Talcott Parsons); second, structuralist and poststructuralist appropriations of psychoanalysis as refracted in sociological debate; and finally, feminist and postmodern social theory has been significantly shaped by psychoanalysis.

Early Disciplinary Articulations

In Europe, Herbert Marcuse and Theodor Adorno – leading members of the Frankfurt school of critical theory – turned to Freud in order to reconceptualize the relation between self and society. The political motivation prompting this turn to Freud had its roots in Marcuse's and Adorno's attempts to conceptualize the rise of fascism, Nazism, and also the spread of bureaucratic capitalism. From Freud's theory of the Oedipus complex, and particularly his theorem concerning the repression of infantile sexuality, Marcuse and Adorno developed the notion of 'the authoritarian personality'. Driven by a desire for conformity and clear rules, the authoritarian personality was viewed by Marcuse and Adorno as a character type strongly prevalent in the German middle classes, a character type who hungered for strong leadership, social order, and regulation. Not only in Nazi Germany, however, was this personality type to be found; in the advanced liberal societies of the West, tendencies toward authoritarianism and conformism are increasingly evident. Marcuse thus spoke of the emergence of 'one-dimensional man'.

Marcuse's radical Freudianism, in particular, won a wide audience in the 1960s – not only in social science circles, but among student activists and sexual liberationists. Arguing that the so-called sexual revolution of the 1960s did not seriously threaten power structures of the established social order, Marcuse sought to show how demands for freedom were routinely rechanneled for commercial interests. The core of his analysis rested upon the distinction he drew between basic and surplus repression. Basic repression he defined as the minimum level of psychological renunciation demanded by the social structure and cultural order. Repression that is surplus, by contrast, refers to the intensification of self-restraint demanded by asymmetrical relations of power. Marcuse describes the 'monogamic-patriarchal' family, for example, as a site of surplus repression. Interestingly, while Marcuse saw signs of surplus repression increasingly everywhere in late capitalist society, he remained remarkably optimistic about the possibilities for social and cultural change.

On the other side of the Atlantic, a different approach to the integration of psychoanalysis and sociology was fashioned. The core thematic of this approach concerned social order, socialization, and the reproduction of the social system. The grand theorist of American sociology, Talcott Parsons, employed Freudian ideas to understand how basic symbols and values are internalized by human subjects throughout the socialization process. According to Parsons's appropriation of Freud for social theory, the structure of human personality is an outcome of an internalization of desired objects, role relationships, and ethico-cultural values that make up the broader social network. In this approach, it is the linkage of personality structure, the social system, and the cultural system that is stressed. Unlike Marcuse's and Adorno's emphasis on the social manipulation of the unconscious, Parsons finds a kind of preestablished harmony between the individual and society.

Structuralist and Poststructuralist Psychoanalytic Sociology

For many years, the integration of psychoanalysis and social analysis developed by the Frankfurt school was commonly regarded as the most sophisticated and important work in this subfield of modern sociology. From the late 1960s onward, however, the impact of French theory, particularly structuralist and poststructuralist philosophy, became increasingly influential in terms of theorizing the social dimensions of psychoanalysis. The key figure in this connection was Freud's French interpreter, Jacques Lacan. Seeking to rework the core concepts of psychoanalysis in the light of modern linguistics, Lacan argued that the unconscious exemplifies key features of language; as Lacan famously argues, "the unconscious is structured like a language". The subject or 'I', according to Lacan, is not self-transparent, but is rather located in a system of signification from which identity is fashioned. For Lacan, intersubjectivity is at the center of psychological functioning and its disturbances; distortions or pathologies at the level of the self are, says Lacan, located in 'the discourse of the other'.

It is perhaps Lacan's essay "The mirror stage as formative of the function of the I" (1949) that has come to exert most influence on contemporary sociological theory. In the essay, Lacan conceptualizes the small infant's initial recognition of itself in a mirror or reflecting surface, and of how this generates a sense of identity. Through the mirror, says Lacan, the infant makes an imaginary identification with its reflected image, an identification that the infant reacts to with a sense of jubilation and exhilaration. But the mirror image of the self for Lacan is, in fact, a distortion; the mirror lies. The mirror stage is radically 'imaginary', in Lacan's theorization, since the consolingly unified image of selfhood that it generates is diametrically opposed to the bodily fragmentation and lack of coordination of the child. These imaginary traps and distortions are a universal and timeless feature of self-organization, and Lacan sees such illusions as directly feeding into and shaping pathologies of the self in contemporary culture.

Lacan was not especially interested in the social applications of psychoanalysis; it was one of his followers, the French Marxist political philosopher Louis Althusser, who brought Lacanian theory into the center of key debates in sociology. In his important essay *Ideology and Ideological State Apparatuses* (1984[1971]), Althusser analyzed ideology in terms of the process by which individuals come to understand and relate to themselves in a manner, which supports dominant class relations. According to Althusser, ideology provides an imaginary identity, an imagined map for locating oneself in the wider social network. Echoing Lacan, Althusser uses the notion of the mirror stage to deconstruct ideology. There is a duplicate mirror structure at the heart of ideology says Althusser, a structure that grants to the self an ideological mirror in which it can recognize itself and other people. Althusser calls this process 'interpellation' – the capturing of the individual within the net of received social meanings.

The Lacanian/Althusserian account of the decentering of the subject has been highly influential in recent sociological theory, and has impacted upon debates concerning agency, structure, class, social fragmentation, and cultural order. The social-theoretical work of Paul Hirst, Barry Hindess, Stuart Hall, Etienne Bailbar, Pierre Machery, Fredric Jameson, and Slavoj Zizek is, in differing ways, indebted to the Lacanian/Althusserian theory of the subject and its ideological subjection.

Feminism, Postmodernism, and Psychoanalytic Sociology

It is not only in studying politics and social change that psychoanalysis has become an important theoretical tool for sociology, but also in debates concerning gender and sexual politics Freudian ideas have been incorporated into social theory. In the sociology of sexuality, the sociology of the family, and especially the social theory of gender transformation, Freudian psychoanalysis has played a vital role in expanding sociologists' understanding of the subjective and affective components of human social relationships.

The psychoanalytic perspective made a forceful entry into contemporary feminist sociological theory in Juliet Mitchell's pioneering book, *Psychoanalysis and Feminism* (1974). Mitchell deployed Freudian and Lacanian psychoanalytic ideas as a means of connecting a discussion of gender power with an Althusserian/Marxist theory of late capitalist society. Against this theoretical backdrop, she asserted that definitions of masculinity and femininity are framed through linguistic and historical structures – with man as a self-determining, autonomous agent, and woman as a lacking other. Such gender dualism, according to Mitchell, is highly conducive to capitalist social regulation – the split between private and public, the pathologies of the familial life, and the like. Mitchell's ideas, while criticized in some feminist sociological circles, had a lasting impact on psychoanalytically oriented feminist sociology. Especially in terms of Lacanian feminist approaches, the writings of authors such as Julia Kristeva, Luce Irigaray, Jacequline Rose, and Judith Butler have significantly influenced sociological debate in recent years.

In the United States, the feminist theories of Nancy Chodorow, Jessica Benjamin, and Jane Flax have been influential in contemporary sociology. These feminist authors draw from the psychoanalytic perspective, but rather than turn to Lacan and French psychoanalysis their work selectively incorporates the insights of Freudian and post-Freudian (especially object-relational) theory. In Chodorow's work, it is part of an attempt to understand the psychic components of female and male socialization, especially in terms of the unconscious forces that shape gender roles. In Benjamin's work, psychoanalysis is deployed to rethink the dynamics of domination and submission within the wider frame of gender, society, and history. In Flax's discussion of psychoanalysis, feminism, and postmodernism, it is primarily a set of philosophical observations about the development of gender relations and the sociology of sexuality and intimacy.

Related to the intertwining of psychoanalysis and feminism, postmodernist appropriations of psychoanalysis have also been influential in recent sociological interventions. This has been especially true for the contested modernity/postmodernity debate, in which sociologists as diverse as Anthony Giddens, Alaine Touraine, and Zygmunt Bauman have drawn from Freud to analyze anew the self and self-identity. In the writings of Michel Foucault, Jean-Francois Lyotard, and Jacques Derrida, which have also had a strong influence on sociological theory, there has been much debate about the fate of the individual or 'death of the subject' in postmodernist culture. In all of these discussions, psychoanalysis has provided sociology with conceptual tools for questioning and deconstructing the enlightenment and rationality of progress.

Current Trends in European Psychoanalysis and Psychoanalytic Sociology

Let us now turn to consider, finally, some current trends in European psychoanalysis, trends that have impacted in various ways upon the development and deepening of psychoanalytic sociology.

Cornelius Castoriadis: Radical and Social Imaginaries

The late European psychoanalyst and social theorist Cornelius Castoriadis has articulated a theory of imagination – at once psychic and social – that has powerfully impacted upon some versions of sociology and social theory. Castoriadis's reflections on the imaginary principally concern the complex ways in which a world (at once emotional and social) somehow or

other comes to be ordered and organized from groundlessness or chaos. In *The Imaginary Institution of Society* (1987), Castoriadis argues that fantasy is a site of multiple, fractured and contradictory positionings of the individual in relation to self, to other people and to society and history. He claims that the psyche is continually elaborating representations and fantasies; as the flow of representations are produced, so new positionings of self and other are defined, which in turn leads to newer forms of fantasy, identification and cultural association. There is for Castoriadis a delicious indeterminacy at the heart of the Freudian unconscious, such that the regulative hierarchies of self, sexuality, gender, and power are constantly rearranged and sometimes transformed, at least partially as a consequence of this ceaseless psychic flux.

At its simplest Castoriadis's emphasis on the creative nature of the imagination underscores the permutation of fantasies and identifications that selves produce endlessly in relation to society and history. We insert ourselves, through the psychic flux of imagination, at one and the same moment as both creator and created, self and other, identity and difference; we draw on existing social institutions and cultural conventions to produce new images of self and society, which in turn feeds back into the cycle of representations. In all this, Castoriadis's central theme is creativity – of the individual self and the broader society. Underlining creativity, Castoriadis's theoretical position is a far cry from the insipid, commercially constructed notion of the 'ever-new' in popular culture. What distinguishes Castoriadis's position from popular understandings of creativity is his stress on the open ended and ambivalent nature of psychic representation and cultural production, and it is this stress, which necessarily involves reflecting on the more distressing aspects of violence, aggression, and destruction in contemporary culture. "Creation", writes Castoriadis, "does not necessarily – nor even generally – signify 'good' creation or the creation of 'positive values'. Auschwitz and the Gulag are creations just as much as the Parthenon and the *Principia Mathematica*" (Cornelius Castoriadis, *Philosophy, Politics, Autonomy*, Oxford: Oxford UP, 1991: pp. 3–4). It is hard – says Castoriadis – to grasp, and harder to understand, that sociopolitical paths or fields of imagination stretch all the way from progressive politics to fanaticism and fascism. But the search for alternative futures and the search for autonomy and justice are among the creations in Western history that people value highly and judge positively; the practice of critique, of putting things into question, forms a common starting point for a radical challenge to received social and political meanings.

Castoriadis (1989), following Freud, contends that originary fantasy and representation underlies the capacity of the subject for critical self-reflection and autonomy. As he says of psychoanalysis itself: 'the possibility of representing *oneself as* representational activity and of putting oneself in question as such is not just a philosophical subtlety; it corresponds to the *minimum* we require of every patient when we try to lead him or her to discover that X is not Y but that it is very much so for *his* or *her* own representational activity and that there may be reasons for this' (1989: p. 27). Representation, unconscious flux, and originary fantasmatization: these are, says Castoriadis, necessary conditions for the possibility of reflectiveness in the individual human subject.

Julia Kristeva: Imagination Reassessed

Julia Kristeva, Professor of Linguistics at the University of Paris VII, did her psychoanalytic training with Lacan. Already in her earliest theoretical work, however, she indicated a determination to move beyond the conceptual terrain of Lacanianism narrowly defined. For example, her doctoral dissertation, which was subsequently published as her first book *Revolution in Poetic Language*, blends linguistics and psychoanalytical theory to advance a novel account of how preverbal experience – infantile, maternal, and poetic – enters into, shapes, distorts, and disrupts language through processes of art, literature, and psychoanalysis. In her more recent work, Kristeva has turned away from Lacan and back to classical Freudianism, and especially toward Kleinian psychoanalysis.

According to Kristeva, the psychic work of representation is a universal feature 'present in all psychic activities and behaviors', and is by no means restricted to the therapeutic relationship, nor to particular aspects of mental functioning, such as the standard psychoanalytic menu of daydreaming or erotic imaginings. Rather, it is our ordinary experiences – from sporting activities to the practicalities of learning and education – that are saturated with this originary imagination. All psychic activity, says Kristeva, is 'impregnated with fantasies'. What is clear in Kristeva's account of fantasy, is that this imaginary domain is inextricably interwoven with the motions of pleasure and unpleasure, the most primitive impulses of desire and aggressiveness which bring a world of subjectivity into being in the first place. Freud astutely captured the theatrical dynamics of sensational life in terms of the logics of dreaming; and it is these affective processes (the dream work), which for Kristeva dominate the mental apparatus from start to finish. Yet what might Kristeva be gaining by drawing attention to the imaginary resilience – the creative representational refashioning of the senses – of everyday life? And what, we might ask, is gained by thinking of what happens to our wishes (inseparable from figure and fantasy) in categories that emphasize the prolinguistic: fantasy life is "expressed in and dealt with by *mental processes far removed from words* and conscious relational thinking".

Kristeva conceptualizes what she refers to as the 'proto-fantasy' as a kind of oscillation of the imagination, with the human subject internally divided, split between infantile narcissism and the other's lack. Strictly speaking, if representation is an 'active presence of fantasy scenes', this is because desire, for Kristeva as for Lacan, is the desire of the Other. To desire the Other is a kind of fashioning, an imagining of what the other dreams, an imitating, an identification with the other's desire. Notwithstanding that it is the inescapability of imaginary misrecognition that leaves the human subject to impute an imaginary fullness to the other's desire which, in fact, pertains only to the representation (i.e., the imaginary plenitude that the subject itself desires), the point is there would be no meaning, not to say anything of the possibility for self-knowledge, without these imaginative fashionings.

Kristeva has written in great depth about the length people will go to in creating obstacles to pleasure; in doing so, she has reformulated Freud's account of Oedipal desire as a general theory of the constitution of the subject and its Baroque imaginings. What is significant to sociology here is that

Kristeva's reflections on the unconscious or preconscious fantasy captures something important about the imaginary making and taking of pleasure in daily life. In Kristeva's reckoning, the psychoanalytic theory of fantasy is about the human subject's imaginings that inform, say, perceptions of the body, imaginings about how one sounds and speaks, imaginings about one's sporting prowess, imaginings about pop stars and celebrities, imaginings about educational advancement and intellectual recognition, and imaginings about where one is headed or what may be wrong with one's life.

More than any other psychoanalyst, perhaps more than Freud's foundational insights, Kristeva captures the complex ways people use their imagination to make life meaningful. For Kristeva, individuals are captured by, and in thrall to, their unconscious fantasies – these radically strange, foreign social dreams. Such fantasies, in addition to constituting intercourse between unconscious dreams and practical life, are the very imaginings of imagination.

Jean Laplanche: Imaginary Seductions

Jean Laplanche trained with Lacan, and his early writings indicate a strong conceptual debt to his former analyst. Indeed, the book for which Laplanche is perhaps best known in the Anglo-American world is *The Language of Psychoanalysis* (1973), coauthored with J.B. Pontalis – an encyclopedic coverage of core psychoanalytic concepts through the lens of French Freudianism. Further works of psychoanalytic exposition and critique followed, including the influential tract *Life and Death in Psychoanalysis* (1976), in which Laplanche struggled to remain faithful to the Lacanian modifications to psychoanalysis, principally through expressing his general suspicion of structural theory. It seems likely, however, that Laplanche's lasting contribution to psychoanalysis derives from his reflections on otherness in the formation of human subjectivity, as developed in his 'general theory of seduction', set out in the five-volume *Problématiques* (1980–87).

Laplanche was one of the first post-Lacanians to write of the strange transformations – the condensations, displacements, and reversals – of unconscious repression, which results in the formation of an internal foreign other, of what Freud called a thing-presentation, or the depths of imagination itself. He has been one of the few major psychoanalytic thinkers, period, to focus on the irreducible creativity of unconscious work, by which he means specifically the field of symbolizing activity. His psychoanalytic work, to a considerable degree, represents a series of reflections on the ontology of determinism within Freudianism. Laplanche's account of the internal otherness – an unconscious of strange, foreign bodies – is foregrounded as at the core of psychic life. As he writes:

> What guarantees the alien-ness of the other? Can one affirm here, with Lacan, the priority of language? If for my part, I speak rather of a "message", this is for at least two well-defined reasons: *firstly*, the message can just as easily be non-verbal as verbal; and for the baby it is principally non-verbal. *Secondly*, emphasising "language" effaces the alterity and individuality of the other in favour of transhistorical structures.
>
> (Laplanche, 1997: p. 660)

Laplanche, like Kristeva, rejects the linguistic imperialism of Lacanian doctrine: "the message can just as easily be nonverbal as verbal". So too, like Kristeva, Laplanche distances himself from a concern with 'transhistorical structures' (phylogenesis, language) in favor of the essential uniqueness and individuality of human imagination. In shifting away from Lacan and back to Freud – returning to prelinguistic psychical representatives or fantasmatic constructions made of images and split from words – Laplanche will emphasize that in the act of *psychic translation* the singular individual *creates* in the strongest sense of the term.

It could be said that Laplanche is out to provide a social theory of our *struggle for representation* in the field of symbolizing activity – which, in a sense, has been the subject of all psychoanalytic theories since Freud unearthed the unconscious logics of the dream. For in his preoccupation with the problem of translation – by which is meant the psychic force field of representations, resemblances, contiguities, condensations, and reversals – Laplanche's work plays ingeniously on a subtle, but definite, relation between human subjects in the context of symbolic and social formations. For Laplanche, it is essential to grasp that the infant is, from the beginning of life, presented with what he calls 'messages' (both verbal and nonverbal) by parents, messages which the infant is ill-equipped to adequately deal with or understand on an emotional plane. It makes perhaps less difference what the soft caresses of a mother actually signify as regards the self-understandings of the adult, though part of Laplanche's interest turns on the way parents always convey far more than they consciously intend. What matters in Laplanche's scheme is that the infant has been addressed or called with a message, a message that is at once exciting and mystifying.

The striking feature of Laplanche's theorization of the message as enigmatic is its sheer open-endedness. His account of the psychosexual development of the individual subject in terms of the ongoing emotional work of translation a retranslation would make no sense were it not for the recognition that, because of the small infant's initially limited ways of trying to emotionally process proffered messages, psychic life is always, necessarily, imaginative, creative, and inventive. Unlike the iron determinism of the early Lacan's emphasis on the Symbolic subjection of the subject, it is the mystifying element of the message that for Laplanche sparks imaginative associations in the child. What is inescapable for the infant – and then subsequently for the adult – is that such mystifying messages demand continual psychic work are in need of continual translation. Indeed, Laplanche himself has acknowledged that he came up with the concept of 'message', with all this implies of the need for translation, in order to overcome the rigid determinism of psychoanalysis in France since Lacan.

Concluding Remarks

Castoriadis, Kristeva, and Laplanche are not alone among those theorists of the contemporary age who have wrestled with the question of imagination as well as individual and social transformations affecting imaginary life. Fortunately, for academic social science but also for the demands of sociology, there have been a growing number of voices raising pressing

political issues about the conditions and consequences of our imaginative interpersonal relations in a postmodern world of fragmentation and fracture. Such authoritative voices today include, to mention only a few: Judith Butler, Drucilla Cornell, Slavoj Zizek, Homi Bhabha, Christopher Bollas, Lynne Segal, Fredric Jameson, Charles Lemert, Luce Irigaray, Elizabeth Grosz, Nancy Chodorow, Jane Flax, Charles Spezzano, Thomas Ogden, and Jessica Benjamin. Each has drawn from psychoanalysis to develop a particular angle on the changing relations between self and society in the contemporary epoch. Each has focused on specific problematics of current social conditions – from feminism to postmodernism, from psychotherapy to literature – in rethinking the terms of both individual and collective imaginaries. Psychoanalysis, it transpires, continues to provide vital food for thought to sociology in the early years of the twenty-first century.

See also: Critical Theory: Contemporary; Frankfurt School: Institute for Social Research; Freud, Sigmund (1856–1939); Gellner, Ernest (1925–95); Parsons, Talcott (1902–79); Postmodernism in Sociology; Psychoanalysis, History of; Unconscious: History of the Concept.

Bibliography

Althusser, L., 1984[1971]. Ideology and Ideological State Apparatuses. Essays on Ideology. Verso, London.
Castoriadis, C., 1987. The Imaginary Institution of Society. Polity, Cambridge, MA.
Castoriadis, C., 1989. Done and be done. In: Curtis, D. (Ed.), The Castoriadis Reader. Blackwell, Oxford, pp. 361–417.
Castoriadis, C., 1991. Philosophy, Politics, Autonomy. Oxford University Press, Oxford.
Chodorow, N., 1978. The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender. University of California Press, Berkeley, CA.
Elliott, A., 1999. Social Theory and Psychoanalysis in Transition: Self and Society from Freud to Kristeva. Free Association Books, London and New York.
Elliott, A., 2004. Social Theory Since Freud. Routledge, London and New York.
Freud, S., 1935–1974. The Standard Edition of the Complete Psychological Works of Sigmund. (J. Stachey, Trans.). Hogarth Press, London.
Kristeva, Julia, 1984. Revolution in Poetic Language. Columbia University Press, New York.
Lacan, J., 1977. Ecrits: A Selection. Tavistock Press, London.
Lacan, J., 2006. The mirror stage as formative of the function of the I. In: Lacan, J., Miller, J.A. (Eds.), Écrits [1949]. W.W. Norton, New York, pp. 75–81.
Laplanche, J., 1976. Life and Death in Psychoanalysis. Johns Hopkins University Press, Baltimore, MD. Laplanche, J. (1980–1987) Problematiques.
Laplanche, J., 1997. Theory of seduction and the problem of the other. International Journal of Psychoanalysis 78, 653–666.
Laplanche, J., Pontalis, J.B., 1973. The Language of Psychoanalysis. Hogarth Press, London (original publication in French, 1967).
Marcuse, H., 1956. Eros and Civilization. Ark, London.
Mitchell, J., 1974. Psychoanalysis and Feminism. Allen Lane, London.

Psychoanalysis: Current Status

Allan Tasman, School of Medicine, University of Louisville, Louisville, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by O.F. Kernberg, volume 18, pp. 12308–12315, © 2001, Elsevier Ltd.

Abstract

Psychoanalysis was first developed by Sigmund Freud beginning in the late 1890's in Europe as a theory of mental development, a theory of the nature of pathological mental processes, and a method of treatment for psychiatric disorders. Freud's theoretical work was followed by a body of psychoanalytic theory contributions referred to as ego psychology which was then followed by bodies of work known as object relations and self-psychology. Freud believed that psychopathology has its origins in mental processes which are generally outside of the individual's awareness as a result of the vicissitudes of development. Freud believed that these clues about the nature of the unconscious pathological processes could be gained from a variety of sources including the analysis of the patient's dreams and of the patient's transference with the psychoanalyst. The transference was described as a reawakening of early developmental conflicts in the relationship with the psychoanalyst. Further the two major schools of modern psychotherapy, psychoanalytic psychotherapy and cognitive behavioral psychotherapy are based on the same important premise that psychopathology has its origins in developmental and interpersonal processes and that the pathological functions operate in an unconscious manner. This chapter highlights important aspects of psychoanalytic theories of both development and treatment with a particular focus on the use of transference within the treatment process.

Psychoanalysis is (1) a personality theory, and, more generally, a theory of psychological development and functioning that focuses particularly on unconscious mental processes; (2) a method for the investigation of psychological functions based on the exploration of the patent's verbal and nonverbal productions within a special therapeutic setting; and (3) a method for treatment of a broad spectrum of psychopathological conditions, which at one time or the other over the second half of the twentieth century was prescribed for the symptomatic neuroses (anxiety states, depression, obsessive–compulsive disorder, conversion hysteria, and dissociative hysterical pathology), sexual inhibitions and perversions ('paraphilias'), and the personality disorders. Psychoanalysis has also been applied, mostly in modified versions, in psychoanalytic psychotherapies, to the treatment of many psychiatric illnesses named above as well as severe personality disorders, psychosomatic conditions, and certain psychotic conditions, particularly a subgroup of patients with chronic schizophrenic illness.

At the present time, psychoanalysis is not considered a first-line treatment for most psychiatric conditions, in large part due to the cost and time requirements of psychoanalysis, although the major psychotherapeutic approaches in use today, namely psychoanalytic and cognitive–behavioral therapies, are based on psychoanalytic concepts. Both psychoanalytic psychotherapies and cognitive–behavioral psychotherapies are based on a version of the vicissitudes of both normal and pathological mental development embodied in psychoanalytic theories of development. And, both approaches are based on an identical assumption that psychopathology has its origins in developmental and interpersonal processes, and that the pathological functions operate essentially in an unconscious manner. While the *language* that describes both the unconscious processes and the techniques used in treatment also differ between these two approaches, and while there is some difference in emphasis on cognitive and affective processes, the goals and actual techniques of both treatment approaches are parallel to psychoanalysis, which are to bring the pathological unconscious processes into awareness so they can be remediated. The goals of psychoanalytic treatment, although, are the most far reaching, both in breadth and depth of mental processes that are the subject for understanding and remediation through treatment.

Freud's Theory of the Mental Apparatus

Freud's (1953a) starting point was his study of hysterical patients and the discovery that, when he found a way to help these patients piece together a coherent account of the antecedents of their conversion symptoms, dissociative phenomena, and pathological affective dispositions, all these psychopathological phenomena could be traced to traumatic experiences in their past that had become unconscious. In the course of a few years, Freud abandoned his early efforts to recover repressed material by means of hypnosis, and replaced hypnosis with the technique of 'free association,' an essential aspect of psychoanalytic technique until the present time.

Practicing this method led Freud to several lines of discovery: to begin, he conceptualized unconscious mechanisms of defense that opposed the recovery of memories by free association. He described these mechanisms to include repression, negation, isolation, projection, introjection, transformation into the opposite, rationalization, intellectualization, and most important, reaction formation. The latter involves overt, chronic patterns of thought and behavior that serve to disguise and disavow consciously experienced conflicted emotions by experiencing the opposite emotion consciously. The discovery of reaction formations led Freud to the psychoanalytic study of character pathology and normal character formation, and still constitutes an important aspect of

the contemporary psychoanalytic understanding and treatment of personality disorders (for practical purposes, character pathology and personality disorders are synonymous concepts).

A related line of development in Freud's theories was the discovery of the differential characteristics of conscious and unconscious thinking. This set of concepts came to be known as the topographic theory. Freud differentiated conscious thinking, the 'secondary process,' invested by 'attention cathexis' and dominated by sensory perception and ordinary logic in relating to the psychosocial environment, from the 'primary process' of the 'dynamic unconscious.' That part of the unconscious mind he referred to as 'dynamic' exerted constant pressure or influence on conscious processes, against the active barrier constituted by the various defensive operations, particularly repression. The dynamic unconscious, Freud proposed, presented a general mobility of affective investments, and was ruled by the 'pleasure principle' in contrast to the 'reality principle' of consciousness. The 'primary process' thinking of the dynamic unconscious was characterized by the absence of the principle of contradiction and of ordinary logical thinking, the absence of negation and of the ordinary sense of time and space, the treatment of a part as if it were equivalent to the whole, and a general tendency toward condensation of thoughts and the displacement of affective investments from one to another mental content.

Finally, Freud proposed a 'preconscious,' an intermediate zone between the dynamic unconscious and consciousness. It represented the storehouse for retrievable memories and knowledge and for affective investments in general, and it was the seat of daydreaming, in which the reality principle of consciousness was loosened, and derivatives of the dynamic unconscious might emerge. Free association, in fact, primarily tapped the preconscious, as well as the layer of unconscious defensive operations which oppose the emergence of material from the dynamic unconscious.

This model of the mind as a 'place' with unconscious, preconscious, and conscious 'regions' thus constituted the topographic model, although Freud eventually supplanted it with the 'structural theory,' namely the concept of three interacting psychic structures, the ego, the superego, and the id (Freud, 1961).

The Id

The id is the mental structure that contains the mental representatives of the 'drives'; that is, the ultimate intrapsychic motivations that Freud (1955) described in his final, 'dual drive theory' of libido and aggression, or metaphorically, the sexual or life drive and the destruction or death drive. In exploring unconscious mental processes, what at first appeared to be specific traumatic life experiences turned out to reflect surprisingly consistent, repetitive, intrapsychic experiences of a sexual and aggressive nature.

Freud (1953a) was particularly impressed by the regularity with which his patients reported the emergence of childhood memories reflecting seductive and traumatic sexual experiences on the one hand, and intense sexual desires and related guilt feelings, on the other. He discovered a continuity between the earliest wishes for dependency and being taken care of (the psychology, as he saw it, of the baby at the mother's breast) during what he described as the 'oral phase' of development; the pleasure in exercising control and struggles around autonomy in the subsequent 'anal phase' of development (the psychology of toilet training); and, particularly, the sexual desire toward the parent of the opposite gender and the ambivalent rivalry for that parent's exclusive love with the parent of the same gender. He described this latter state as characteristic of the 'infantile genital stage' (from the third or fourth to the sixth year of life) and called its characteristic constellation of wishes and conflicts the positive Oedipus complex. He differentiated it from the negative Oedipus complex, that is, the love for the parent of the same gender, and the corresponding ambivalent rivalry with the parent of the other gender. Freud proposed that oedipal wishes came to dominate the earlier infantile hierarchy of oral and anal wishes, becoming the fundamental unconscious realm of desire.

Powerful fears motivated the repression of awareness of infantile desire: the fear of loss of the desired person (the individual toward whom the emotional impulse, or drive, is directed is called the 'object' in the original English translation), and later of the loss of the object's love was the basic fear of the oral phase, directed against libidinal wishes to possess the breast; the fear of destructive control and annihilation of the self or the object was the dominant fear of the anal phase directed against libidinal wishes of anal expulsion and retentiveness, and the fear of castration, 'castration anxiety,' the dominant fear of the oedipal phase of development, directed against libidinal desire of the oedipal object. Unconscious guilt was a dominant later fear, originating in the superego as that set of functions developed around the second year of age and generally directed against drive gratification (see Section The Superego). Unconscious guilt over sexual impulses unconsciously equated with oedipal desires constituted to Freud a major source of many types of pathology, such as sexual inhibition and related character pathology.

Freud believed the infants biological status and its changes served as the impetus for the development of mental functions, and described the oral phase as essentially coinciding with the infantile stage of breast feeding, the anal phase as coinciding with struggles around sphincter control, and the oedipal stage as developing gradually during the second and through the fourth years, and culminating in the fourth and the fifth years of life. This latter phase would then be followed by more general repressive processes under the dominance of the installation of the superego, leading to a 'latency phase' roughly corresponding to the school years, and finally, to a transitory reactivation of all unconscious childhood conflicts under the dominance of oedipal issues during puberty and early adolescence. Erik Erikson later expanded Freud's description of developmental phases through adolescence to a series of adult developmental phases, and his contribution met with general acceptance.

The drives represent for human behavior what the instincts constitute for the animal kingdom, that is, the ultimate biological motivational system. The drives are constant, highly individualized, developmentally shaped motivational systems. Under the dominance of the drives and guided by the primary process, the id exerts an ongoing pressure toward gratification, operating in accordance with the pleasure principle. Freud

initially equated the drives with primitive affects. After discarding various other models of unconscious motivation, he ended up with the dual drive theory of libido and aggression.

He described the libido or the sexual drive as having an 'origin' in the erotogenic nature of the leading oral, anal, and genital bodily zones; an 'impulse' expressing the quantitative intensity of the drive by the intensity of the corresponding affects; an 'aim' reflected in the particular act of concrete gratification of the drive; and an 'object' consisting of displacements from the dominant parental objects of desire.

The introduction of the idea of an aggressive or 'death' drive, arrived at later in Freud's writing, stemmed from his observations of the profound self-destructive urges particularly manifest in the psychopathology of major depression and suicide, and of the 'repetition compulsion' of impulse driven behavior that frequently seemed to run counter to the pleasure principle that supposedly governed unconscious drives. He never spelled out the details of the aggressive drive as to its origins. Freud described drives as intermediate between the body and the mind; the only thing we knew about them, Freud (1957a) suggested, were 'representations and affects.'

The Ego

While the id is the seat of the unconscious drives, and functions according to the 'primary process' of the dynamic unconscious, the ego, Freud (1961) proposed, is the seat of consciousness as well as of unconscious defense mechanisms that, in the psychoanalytic treatment, appear as 'resistance' to free association. The ego functions according to the logical and reality-based principles of 'secondary process,' negotiating the relations between internal and external reality. Guided by the reality principle, it exerts control over perception and motility; it draws on preconscious material, controls 'attention cathexes,' and permits motor delay as well as selection of imagery and perception. The ego is also the seat of basic affects. Freud particularly focused his attention on anxiety, which he viewed as an alarm signal against the psychologically experienced danger of emergence of unconscious, repressed, and conflicted impulses. This alarm signal may turn into a disorganized state of panic when the ego is flooded with external perceptions that activate unconscious desire and conflicts, or with overwhelming, traumatic experiences in reality that resonate with such repressed unconscious conflicts, and overwhelm the particularly sensitized ego in the process.

Freud originally equated the 'I,' that is, the categorical self of the philosophers, with consciousness. Later, once he established the theory of the ego as an organization of both conscious and unconscious functions, he at times treated the ego as if it were the subjective self, and at other times, as an impersonal organization of functions. Out of this ambiguity evolved the contemporary concept of the self within modern ego psychology as well as in British and American object relations and cultural psychoanalytic contributions (Kernberg, 1984).

An alternative theory of the self was proposed by Kohut (1971), the originator of the self-psychology approach within contemporary psychoanalysis. Kohut's work was controversial within psychoanalysis in part because he discarded the structural theory on which psychoanalysis had been historically based. He replaced the three 'structures' that what he called the 'bipolar self.' The self, to Kohut, emerged from two primitive sets of functions that served to maintain psychological equilibrium. The first was called the idealized parental imago, a set of internal representations of the omnipotent and all caring parent on whom security, and even life, relied. The second was the grandiose self, representing an internalization of the idealized parent into the self. Kohut believed these two structures transformed as development proceeded into adult capacities and psychological abilities that maintained psychological equilibrium, self-cohesion he called it, self-awareness and reflection, and a wide range of adaptive functions. Another controversial view of Kohut was his later diminution of the importance of the oedipal phase of development in health and psychopathology in favor of an emphasis on pre-oedipal development.

Kohut's work drew heavily on the school of thought that came to be known as object relations theory, first promulgated by Melanie Klein in the 1930s. Klein and those who followed her and elaborated on her theories emphasized to a much greater extent than Freud the importance of the years before the oedipal phase for psychological development and the origins of psychopathology. Freud emphasized the role of trauma in derailing normal early developmental processes, but the object relations theorists emphasized a much broader perspective, positing that early psychological development was immensely influenced by the nature of the mother–infant psychological attachment (in the present, one would say not only the attachment to the mother but to anyone playing a substantial role in early nurturance and care) and the resulting internalized perceptions of the self and parent. In this view, not only trauma, but also the vagaries of the typical and individually unique attachment exerted a powerful and determinative effect on the infant's psychological development, interacting with the inborn characteristics of the infant (such as introversion/extroversion, risk aversion/risk taking, and passivity/aggressiveness).

Nowadays, an integrated concept of the self as the seat of subjectivity is considered an essential structure of the ego, and the concept of 'ego identity' refers to the integration of the concept of the self. Because developmental processes in early infancy and childhood are now better understood, an integrated self-concept usually goes hand in hand with the capacity for an integrated concept of significant others. An unconscious tendency toward primitive dissociation or 'splitting' of the self-concept and of the concepts of significant objects runs counter to such integration.

Character, or personality, from a psychoanalytic perspective, may be defined as constituting the behavioral aspects of ego identity (the self-concept) and the internal relations with significant others (the internalized world of 'object relations'). The sense of personal identity and of an internal world of object relations in turn reflects the subjective side of character. It was particularly the ego psychological approach – one of the dominant contemporary psychoanalytic schools which emerged in the 1930's – that developed the analysis of defensive operations of the ego, and of pathological character formation as a stable defensive organization that needed to be explored and resolved in the psychoanalytic treatment. In the process, ego psychology importantly contributed to the

psychoanalytic understanding of mental functioning and had a significant impact on psychoanalytic treatment. This school of thought posited that all conscious mental processes, including free association were influenced by ego actions, and that the understanding of ego functions could be both revealed in the patient's verbal productions, and well as serve as a focus for the therapeutic action of psychoanalysis.

Freud (1957c) conceptualized narcissism as the libidinal investment of the ego or self, in contrast to the libidinal investment of significant others ('objects'). In proposing the possibility of a withdrawal of libidinal investment from others with an excessive investment in the self as the basic feature of narcissistic pathology, he pointed to a broad spectrum of psychopathology, and thus first stimulated the contribution of Abraham (1919), and later those of Grunberger (1979), Jacobson (1964), Kernberg (1975), Klein (1957), Kohut (1971), and Rosenfeld (1987). Thus crystallized the description of the narcissistic personality as a disorder derived from a pathological integration of a grandiose self as a defense against unbearable aggressive conflicts, particularly around primitive envy.

The Superego

In his analysis of unconscious intrapsychic conflicts between drive and defense, Freud regularly encountered unconscious feelings of guilt in his patients, reflecting an extremely strict, unconscious infantile morality, which he called the superego. This unconscious morality could lead to severe self-blame and self-attacks, and particularly to abnormal depressive reactions, which he came to regard as expressing the superego's attacks on the ego. It was, particularly, in studying normal and pathological mourning that Freud (1957c) arrived at the idea of excessive mourning and depression as reflecting the unconscious internalization of the representation of an ambivalently loved and hated lost object. In unconsciously identifying the self with that object introjected into the ego, the individual now attacked his own self in replacement of the previous unconscious hatred of the object; and the internalization of aspects of that object into the superego reinforced the strictness of the individual's preexisting unconscious infantile morality.

Freud traced the origins of the superego to the overcoming of the Oedipus complex via unconscious identification with the parent of the same gender: in internalizing the oedipal parent's prohibition against the rivalry with him or her and the unconscious death wishes regularly connected with such a rivalry, and against the incestuous desire for the parent of the other gender, this internalization crystallized an unconscious infantile morality. The superego, thus based upon prohibitions against incest and parricide, and a demand for submission to, and identification with the oedipal rival, became the guarantor of the capacity for identification with moral and ethical values in general.

Freud also described the internalization of the idealized representations of both parents into the superego in the form of the 'ego ideal.' He suggested that the earliest sources of self-esteem, derived from mother's love, gradually fixated by the baby's and small child's internalizations of the representations of the loving mother into the ego ideal, led to the parental demands also becoming internalized. In other words, normally self-esteem is maintained both by living up to the expectations of the internalized idealized parental objects and by submitting to their internalized prohibitions. This consideration of self-esteem regulation leads to the clinical concept of narcissism as normal or pathological self-esteem regulation, in contrast to the theoretical concept of narcissism as the libidinal investment of the self.

Psychoanalytic Treatment

The Psychoanalytic Theory of Psychopathology

This theory proposes that the clinical manifestations of the symptomatic neuroses, character pathology, perversions, sexual inhibitions, and selected types of psychosomatic and psychotic illness reflect unconscious intrapsychic conflicts between drive derivatives following the pleasure principle, defensive operations reflecting the reality principle, and the unconscious motivations of the superego. Unconscious conflicts between impulse and defense are expressed in the form of structured conflicts between the agencies of the tripartite structure. There are ego defenses against impulses of the id; the superego motivates inhibitions and restrictions in the ego; at times, the repetitive, dissociated expression of id impulses ('repetition compulsion') constitutes an effective id defense against superego pressures. The resolution of unconscious conflicts implies the analysis of all these intersystemic conflicts.

All these conflicts are clinically expressed by three types of phenomena: (1) inhibitions of normal ego functions regarding sexuality, intimacy, social relations, and affect activation; (2) compromise formations between repressed impulses and the defenses directed against them; and (3) dissociative expression of impulse and defense. This latter category implies a dominance of the splitting mechanisms referred to before. These have acquired central importance in the understanding of severe character pathology as reflected in contemporary psychoanalytic thinking. The concept of developmental deficit has also been widely used to describe the view that it may be not only a conflict, but also a lack of developmental maturation, even in the absence of major conflict, which may also both contribute to psychopathology and serve as a focus for remediation through treatment.

The Structural Formulation of the Psychoanalytic Method

Psychoanalytic treatment consists, in essence, in facilitating the reactivation of the pathogenic unconscious conflicts in the treatment situation by means of a systematic analysis of the defensive operations directed against them. In so doing, the treatment may address barriers to normal development, which deficits may also be addressed as issues for attention on their own (see following section below for further discussion). This leads to the gradual emergence of repressed impulses, with the possibility of elaborating them in relation to the analyst (through the emergence of transference), and their eventual adaptive integration into the adult ego. Freud (1953c) had described the concept of 'sublimation' as an adaptive transformation of unconscious drives: drive derivatives, converted into a consciously tolerable form, are

permitted gratification in a symbolic way, while their origin remains unconscious. The result of this process is an adaptive, nondefensive compromise formation between impulse and defense. In analysis, the gradual integration into the patient's conscious ego of unconscious wishes and desires from the past and the understanding of the fantasized threats and dangers connected with them facilitates their gradual elaboration and subliminatory expression both in the consulting room and in everyday life.

The Object Relations Theory Formulation of Psychoanalytic Treatment

In the light of contemporary object relations theory, the formulation based upon the structural theory (resolution of unconscious conflicts between impulse and defense) has changed, in the sense that all unconscious conflicts are considered to be imbedded in unconscious internalized object relations. Such internalized object relations determine both the nature of the defensive operations and of the impulses against which they are directed. These internalized object relations constitute, at the same time, the 'building blocks' of the tripartite structure of id, ego, and superego. Object relations theory proposes that the gradual analysis of intersystemic conflicts between impulse and defense (structured into conflicts among ego, superego, and id) decomposes the tripartite structure into the constituent conflicting internalized object relations. These object relations are reactivated in the treatment situation in the form of an unconscious relation between self and significant others replicated in the relation between patient and analyst; that is, the 'transference.' Transference is not, as described in the common parlance, a unitary response, but a multilayered and multifocal set of responses and perceptions.

The transference is the unconscious repetition in the 'here and now' of unconscious, conflicting pathogenic relationships from the past. The transference reflects the reactivation of the past conflict not in the form of a memory, but in the form of a repetition. This repetition provides essential information about the past, but constitutes at the same time a defense in the sense that the patient repeats instead of remembering. Therefore, transference has important informative features that need to be facilitated in their development, and defensive features that need to be therapeutically resolved once their nature has been clarified. Transference analysis is the fundamental ingredient of the psychoanalytic treatment.

The Psychoanalytic Treatment Process

Gill (1954), in a classical definition that is still relevant today, proposed the definition of psychoanalysis as a treatment that facilitates the development of a 'regressive transference neurosis' and its resolution by means of interpretation alone, carried out by the analyst from a position of technical neutrality.

The concept of a regressive transference neurosis refers to the gradual gathering into the relationship with the analyst of the patient's most important past pathogenic experiences and unconscious conflicts. The concept of a regressive transference neurosis has been largely abandoned in practice because, particularly in patients with severe character pathology, transference regression occurs so early and consistently that the gradual development of a regressive transference neurosis is no longer a useful concept.

Gill's proposal that the resolution of the transference be achieved 'by interpretation alone' refers to 'interpretation' as a set of the psychoanalyst's interventions that starts with 'clarification' of the patient's subjective experiences communicated by means of free association, expands with the tactful 'confrontation' of aspects of the patient's patterns of behavior that are expressed in a dissociated or split-off manner from his subjective awareness, and thus complements the total expression of his intrapsychic life in the treatment situation, and finally evolves into 'interpretation per se.' Interpretation per se implies the formulation of hypotheses regarding the unconscious meanings in the 'here and now' of the patient's material, and the relation of these unconscious meanings with the 'there and then' of the patient's unconscious, past pathogenic experiences.

The concept of 'technical neutrality' refers to the analyst's impartiality regarding both impulse and defense, with a concerned objectivity that provides a helpful collaboration with the patient's efforts to come to grips with his intrapsychic conflicts.

This definition of the nature of psychoanalytic treatment needs to be complemented with the contemporary concepts of 'transference,' 'countertransference,' 'acting out,' and 'working through.'

An Object Relations Theory Model of the Transference and Countertransference

Modern object relations theory proposes that, in the case of any particular conflict around sexual or aggressive impulses, the conflict is imbedded in an internalized object relation, that is, in a repressed or dissociated representation of the self ('self-representation') linked with a particular representation of another who is a significant object of desire or hatred ('object representation'). Such units of self-representation, object representation, and the dominant sexual, dependent, or aggressive affect linking them are the basic 'dyadic units,' whose consolidation will give rise to the tripartite structure. Internalized dyadic relations dominated by sexual and aggressive impulses will constitute the id; internalized dyadic relations of an idealized or prohibitive nature the superego, and those related to developing psychosocial functioning and the preconscious and conscious experience, together with their unconscious, defensive organization against unconscious impulses, the ego.

These internalized object relations are activated in the transference with an alternating role distribution, that is, the patient enacts a self-representation while projecting the corresponding object representation onto the analyst at times, while at other times projecting his self-representation onto the analyst and identifying with the corresponding object representation. The impulse or drive derivative is reflected by a dominant, usually primitive affect disposition linking a particular dyadic object relation; the associated defensive operation is also represented unconsciously by a corresponding dyadic relation between a self-representation and an object representation under the dominance of a certain affect state.

The concept of countertransference, originally coined by Freud as the unresolved, reactivated transference dispositions of the analyst is currently defined as the total affective states of the analyst in response to the patient and his or her transference, shifting from moment to moment, and providing important data of information to the analyst. The countertransference, thus defined, may be partially derived from unresolved problems of the analyst, but stems as well from the impact of the dominant transference reactions of the patient, from reality aspects of the patient's life, and sometimes from aspects of the analyst's life situation, that are emotionally activated in the context of the transference developments. In general, the stronger the transference regression, the more the transference determines the countertransference; thus the countertransference becomes an important diagnostic tool. The countertransference includes both the analyst's empathic identification with a patient's central subjective experience ('concordant identification') and the analyst's identification with the reciprocal object or self-representation ('complementary identification') unconsciously activated in the patient as part of a certain dyadic unit, and projected onto the analyst (Racker, 1957). In other words, the analyst's countertransference implies identification with what the patient cannot tolerate in himself or herself, and must dissociate, project, or repress.

At this point, it is important to refer to certain primitive defensive operations that were described by Klein (1952) and her school as well as Kohut (1971) and his school in the context of the analysis of severe character pathology. Primitive defensive operations are characteristic of patients with severe personality disorders, and emerge in other cases during periods of regression. They include splitting, projective identification, denial, omnipotence, omnipotent control, primitive idealization, and devaluation (contempt). All these primitive defenses center on splitting; that is, an active dissociation of contradictory ego (or self) experiences as a defense against unconscious intrapsychic conflict. They represent regression to the phase of development (the first 2–3 years of life) before repression and its related mechanisms mentioned are established.

Primitive defensive operations present important behavioral components that tend to induce behaviors or emotional reactions in the analyst, which, if the analyst manages to 'contain' them, permit him to diagnose in himself projected aspects of the patient's experience. Particularly 'projective identification' is a process in which (1) the patient unconsciously projects an intolerable aspect of self-experience onto (or 'into') the analyst; (2) the analyst unconsciously enacts the corresponding experience ('complementary identification'); (3) the patient tries to control the analyst, who now is under the effect of this projected behavior; and (4) the patient meanwhile maintains empathy with what is projected. Such complementary identification in the countertransference permits the analyst to identify himself through his own experience with the aspects of the patient's experience communicated by means of projective identification. This information complements what the analyst has discovered about the patient by means of clarification and confrontation, and permits the analyst to integrate all this information in the form of a 'selected fact' that constitutes the object of interpretation. Interpretation is thus a complex technique that is very much concerned with the systematic analysis of both transference and countertransference.

Contemporary Trends of the Psychoanalytic Method

Contemporary psychoanalytic technique can be seen as having evolved from a 'one person psychology' into a 'two person psychology,' and then into a 'three person psychology.' The concept of 'one person psychology' refers to Freud's original analysis of the patient's unconscious intrapsychic conflicts by analyzing the intrapsychic defensive operations that oppose free association. The 'two person psychology' refers to the central focus on the analysis of transference and countertransference. In the views of the contemporary intersubjective, interpersonal, and self-psychology psychoanalytic schools, the relationship between transference and countertransference is mutual, in the sense that the transference is at least in part a reaction to reality aspects of the analyst, who therefore must be acutely mindful of his contribution to the activation of the transference. The so-called 'constructivist' position regarding transference analysis assumes that it is impossible for the analyst to achieve a totally objective position outside the transference/countertransference bind.

In contrast, the contemporary 'objectivist' position, represented by the 'three person psychology' approaches of the Kleinian school, the French psychoanalytic mainstream, and significant segments of contemporary ego psychology proposes that the analyst has to divide himself between one part influenced by transference and countertransference developments, and another part that, by means of self-reflection, maintains himself or herself outside this process, as an 'excluded third party,' who, symbolically, provides an early triangulation to the dyadic regression that dominates transference developments. This triangulation in the treatment situation becomes particularly important in the treatment of severe personality disorders.

The 'repetition compulsion' as a resistance of the id is very likely a form of acting out as a defense against emotional containment of an extremely painful or traumatic set of experiences. 'Working through' refers to the repeated elaboration of an unconscious conflict in the psychoanalytic situation. It is a major task for the analyst, who has to be alert to the subtle variation in meanings and implications of what on the surface may appear to be an endless repetition of the same conflict in the transference. The patient elaboration of the conflict that presents itself with these repetitive characteristics also implies the function of 'holding' originally described by Winnicott (1965) and later emphasized by Kohut. It consists of the analyst's capacity to withstand the onslaught of primitive transferences without retaliation, abandonment of the patient, or a self-devaluing giving up, and the maintenance of a working relationship (or 'therapeutic alliance') that addresses itself consistently to the healthy part of the patient, even when the latter is under the control of his most conflicting behaviors. Bion's (1967) concept of 'containing' is complementary to 'holding,' in the sense that holding deals mostly with the affective disposition of the analyst, and containing with his cognitive capacity to maintain a concerned objectivity and focus on the 'selected fact,' permitting the integration in the analyst's mind what the patient can only express in violently dispersed or split off behavior patterns. Kohut believed that an essential part

of the holding process is the analyst's empathic capacity and the communication of this understanding directly at the appropriate time to the patient. He also believed that this type of communication serves an essential therapeutic function as well.

'Dream analysis' developed in the context of the method of free association, and constituted, in Freud's (1953b) view, a 'royal road to the unconscious.' Freud's discovery of primary process thinking derived from his method of dream analysis. By now, psychoanalytic thinking has evolved into the view that there are many 'royal roads' to the unconscious. The analysis of character defenses, for example, or of particular transference complications, may be equally important avenues of entry into the patient's unconscious mind.

The technique of dream analysis consists, in essence, in asking the patient to free associate to elements of the 'manifest content' of the dream, in order to arrive at its 'latent' content, the unconscious wish defended against and distorted by the unconscious defensive mechanisms that constitute the 'dream work,' and have transformed the latent content into the manifest dream. The latent content is revealed with the help of the simultaneous analysis of the way in which the dream is being communicated to the analyst, the 'day residuals' that may have triggered the dream, the unconscious conflicts revealed in it, and the dominant transference dispositions in the context of which the dream evolved. Dreams especially to Gustave Jung, a contemporary of Freud, also provide some residual, universal symbolic meanings that may facilitate the total understanding of the latent content.

'The analysis of character' may be the single most important element of the psychoanalytic method in bringing about fundamental characterological change. Character analysis is facilitated by the patient's use of reaction formations, that is, his defensively motivated character traits, as transference resistances. Thus, the activation of defensive behaviors in the transference, reflecting the patient's characterological patterns in all interpersonal interactions, facilitates both the analysis of the underlying unconscious conflicts, and in the process, the resolution of pathological character patterns. The result is an increase in the patient's autonomy, flexibility, and capacity for adaptation. As modern research into temperament, that is, the genetically inherited components of personality, evolves, there will certainly be an influence on how psychoanalysts understand and address character during treatment.

The overall objective of psychoanalytic treatment is not only the resolution of symptoms and pathological behavior patterns or characteristics, but fundamental, structural change, that is, the expansion and enrichment of ego functions as the consequence of resolution of unconscious conflict and the integration of previously repressed and dynamically active id and superego pressures into ego potentialities. Such change is reflected in the increasing capacity for both adaptation to, and autonomy from, psychosocial demands and expectations, and an increased capacity for gratifying and successful functioning in love and work.

Treatment Results: Research on Outcome

The Menninger Psychotherapy Research Project, a naturalistic study comparing psychoanalysis, psychoanalytic psychotherapy, and supportive psychotherapy, showed psychoanalysis to be the most effective of these approaches with patients presenting relatively good ego strength, while patients with severe ego weakness – what nowadays would be described as presenting severe personality disorders or borderline personality organization – improved most with psychoanalytic psychotherapy (Kernberg et al., 1972). This research also showed how important supportive elements were throughout all modalities of treatment (Wallerstein, 1986). A comprehensive review of outcome studies on psychoanalytic psychotherapy and psychoanalysis by Bachrach and collaborators (Bachrach et al., 1985) concluded that the improvement rates are in the 60–90% range, but it also pointed to limitations and problems in the methodology utilized.

Recently, studies regarding the treatment process and outcome of psychoanalysis and psychoanalytic psychotherapy have become more precise in defining the specific treatment variables of psychotherapeutic and psychoanalytic treatments, and several systematic studies on psychoanalytic psychotherapies and psychoanalysis are in progress (Fonagy, 1998). A recent study by the Stockholm Outcome of Psychoanalysis and Psychotherapy Project (STOPPP) has found, on the basis of a relatively large patient population, that psychoanalytic treatment, in comparison with psychoanalytic psychotherapy, obtained a significantly higher degree of long range symptomatic improvement (Sandell et al., 1997b).

Modern research into various forms of psychotherapy has been slow to develop, but has expanded dramatically in recent years. While the individual case study was considered the gold standard of psychoanalytic research for much of the twentieth century, which actually impeded use of modern research techniques in psychoanalysis, there is an increasing recognition that larger sample size studies using modern psychotherapy research approaches are necessary to validate the effectiveness of psychoanalysis itself as a treatment. The emerging body of research in both psychoanalytic psychotherapies and cognitive behavioral psychotherapies, treatments which emerged from psychoanalysis itself, does suggest, however, that treatments based on the principles elucidated in the discussion in this article are at least as effective as other forms of psychotherapy, and, in certain psychiatric conditions, such as depression, are at least effective, if not more so, than medications. Finally, all psychiatric treatments appear to be more effective when informed by concepts regarding the nature of and importance of the therapeutic alliance which originated with psychoanalysis.

Psychoanalytic theories continue to evolve, as do the treatment indications, and the techniques utilized. There is a vast-literature which provides more depth of review than is possible here. An example of another well referenced brief review was done by Haggarty and colleagues (Haggard, 2008).

Acknowledgment

This article is a revised version of the article by Otto Kernberg which appeared in the previous edition.

See also: Ethical Guidelines and Codes of Conduct in Social and Behavioral Research; Psychoanalysis in Sociology; Psychoanalysis, History of; Psychodynamic Therapy: Short-Term; Psychology of the Real Self: Psychoanalytic

Perspectives; Psychotherapy Integration; Psychotherapy and Hermeneutics; Research Ethics: Research.

Bibliography

Abraham, K., 1919/1979. A particular form of neurotic resistance against the psychoanalytic method. In: Selected Papers on Psycho-Analysis. Brunner/Mazel, New York, pp. 303–311.

Bachrach, H.M., Weber, J.J., Murray, S., 1985. Factors associated with the outcome of psychoanalysis. Report of the Columbia Psychoanalytic Research Center (IV). International Review of Psychoanalysis 12, 379–389.

Bion, W.R., 1967. Second Thoughts. Selected Papers on Psychoanalysis. Basic Books, New York.

Fonagy, P., 1998. An Open Door Review of Outcome Studies in Psychoanalysis. A report by the Research Committee of the International Psychoanalytical Association, 12 December 1998.

Freud, S., 1953a. In: Three Essays on the Theory of Sexuality, standard ed., vol. 7. Hogarth Press, London, pp. 125–245.

Freud, S., 1953b. In: Fragment of an Analysis of a Case of Hysteria, standard ed., vol. 7. Hogarth Press, London, pp. 3–122.

Freud, S., 1953c. In: The Interpretation of Dreams, standard ed., vol. 4. Hogarth Press, London, pp. 1–338. 5: 339–625.

Freud, S., 1955. Beyond the Pleasure Principles, standard ed. Hogarth Press, London.

Freud, S., 1957a. In: Instincts and their Vicissitudes, standard ed., vol. 14. Hogarth Press, London, pp. 109–140.

Freud, S., 1957b. In: Mourning and Melancholia, standard ed., vol. 14. Hogarth Press, London, pp. 237–258.

Freud, S., 1957c. In: On Narcissism, standard ed., vol. 14. Hogarth Press, London, pp. 69–102.

Freud, S., 1961. In: The Ego and the Id, standard ed., vol. 19. Hogarth Press, London, pp. 3–66.

Gill, M., 1954. Psychoanalysis and exploratory psychotherapy. Journal of the American Psychoanalytic Association 2, 771–797.

Grunberger, B., 1979. Narcissism: Psychoanalytic Essays. International Universities Press, New York.

Haggard, P., et al., 2008. Psychoanalytic Theories. In: Tasman, A., Kay, J., et al. (Eds.), Psychiatry, vol. 1, third ed. John Wiley & Sons Ltd., West Sussex, England, pp. 464–485.

Jacobson, E., 1964. The Self and the Object World. International Universities Press, New York.

Kernberg, O.F., 1975. Borderline Conditions and Pathological Narcissism. Jason Aronson, Inc., New York.

Kernberg, O.F., 1984. Severe Personality Disorders. Psychotherapeutic Strategies. Yale University Press, New Haven, CT.

Kernberg, O.F., Burnstein, E., Coyne, L., Appelbaum, A., Horwitz, L., Voth, H., 1972. Psychotherapy and psychoanalysis: final report of the menninger foundation's psychotherapy research project. Bulletin Menninger Clinic 36, 1–275.

Klein, M., 1952. Notes on some schizoid mechanisms. In: Riviere, J. (Ed.), Developments in Psycho-Analysis. Hogarth Press, London, pp. 292–320.

Klein, M., 1957. Envy and Gratitude. Basic Books, New York.

Kohut, H., 1971. The Analysis of the Self. International Universities Press, New York.

Racker, H., 1957. The meaning and uses of countertransference. Psychoanalytic Quarterly 26, 303–357.

Rosenfeld, H., 1964. On the psychopathology of narcissism: a clinical approach. International Journal of Psycho-Analysis 45, 332–337.

Rosenfeld, H., 1987. Impasse and Interpretation. Tavistock, New York.

Sandell, R., Blomberg, J., Lazar, A., 1997a. When reality doesn't fit the blueprint: doing research on psychoanalysis and long-term psychotherapy in a public health service program. Psychotherapy Research 7, 333–344.

Sandell, R., Schubert, J., Blomberg, J., Carlsson, J., Lazar, A., Bromberg, J., 1997b. The Influence of Therapist Factors on Outcomes of Psychotherapy and Psycho-Analysis in the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (STOPP). Annual Meeting of the Society for Psychotherapy Research, Geilo, Norway.

Wallerstein, R., 1986. Forty-Two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy. Guilford Press, New York.

Winnicot, D., 1965. The Maturational Processes and the Facilitating Environment. International Universities Press, New York.

Psychodynamic Therapy: Short-Term

Jacques P Barber, Adelphi University, Garden City, NY, USA
Richard F Summers, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.P. Barber, R. Stratt, volume 18, pp. 12470–12473, © 2001, Elsevier Ltd.

Abstract

Short-term dynamic therapy (STDP) comprises a group of interventions based on psychoanalytic principles that have been adapted for use in treatments of limited duration. The prototypic definition of STPD includes the following aspects: the theory of psychopathology explaining the disorder and treatment techniques employed (e.g., clarification, interpretation, and confrontation) that are psychoanalytic in origin; the selection of patients on the basis of their ability to withstand intensive psychotherapy; treatment that is time-limited and focused on a specific issue or conflict; and treatment aimed to provide symptom relief or limited personality or character change. STPD was developed by a subgroup of psychoanalysts who were trying to increase the efficiency of psychodynamic therapy, to bring its power to short-term settings (seeing patients once or twice a week, sitting up), and therefore to make the treatment available to a potentially broader group of patients. Subsequent changes in health care financing and reimbursement and increased awareness of data supporting its efficacy have driven increased interest in STDP.

Modern STDP proponents have developed treatment manuals detailing their specific therapeutic techniques and these manuals have not only aided in the training of the next generation of clinicians, but they have also been used in efficacy studies of STPD to ensure that therapists follow treatment guidelines. Meta-analytic reviews of STPD suggest that it is as effective as other therapies for a range of disorders and more effective than control conditions.

Short-term psychodynamic therapy (STPD) denotes a 'family' of psychotherapeutic interventions with a common theory and technique based on psychoanalytic ideas and approaches and adapted for use in short-term settings. These treatments have also been referred to as time-limited dynamic psychotherapy, brief dynamic psychotherapy, or psychodynamic psychotherapy (Crits-Christoph and Barber, 1991; Messer and Warren, 1995). The number of psychotherapies classified as 'short-term dynamic' has increased remarkably in recent decades. As these numbers continue to grow, along with new theory and attempts at integration across psychotherapies, the definition of what constitutes short-term psychodynamic treatment has become blurred.

We employ the following prototypical definition of STPD based on criteria from various authors, including Crits-Christoph et al. (1991):

1. The theory of psychopathology explaining the cause and maintenance of the psychological disorder has a psychoanalytic basis. This may include Freudian, neo-Freudian, interpersonal object relations, self-psychology, mentalization-based, attachment-based, and relational theories. Central constructs include unconscious motivation, internal conflict, a developmental perspective including the formative role of early years on character, the importance of early relationships, the tendency to repeat conflicts, the pervasiveness of transference, and the use of characteristic defenses.
2. The main treatment techniques employed by the STPD therapist include those described in the psychoanalytic literature, such as open-ended exploration; attention to the therapeutic relationship; making clarifications, interpretations, and confrontations; and noting transference and countertransference. The therapist's theoretical convictions will shape his or her approach to each of these basic treatment techniques, for example, an STPD practitioner with a relational background would be more likely to focus on the present moment and the therapist's inner experience even in the short-term setting. Although the therapist maintains a psychoanalytic stance in the sense of neutrality and abstinence, the short-term model promotes a more active exchange between therapist and patient than would be typical in traditional long-term psychodynamic therapy or psychoanalysis. They would also tend to not emphasize free association.
3. The length of treatment is planned, often predetermined, and always time-limited. It usually lasts from 12 to 40 sessions depending on the patient's problem.
4. Patients are typically screened for STPD to try to select those more likely to respond. In the literature, the following characteristics have been mentioned as predictive of a good outcome: psychological mindedness, recognition that present problems are at least partly of a psychological nature, willingness to change, the existence in the present or past of at least one good quality interpersonal relationship, the ability to circumscribe present complaints to one symptom, the capacity to interact flexibly with the interviewer, neurotic and not psychotic defenses, adequate reality testing, and the ability to tolerate frustration and ambivalence.
5. A focus of treatment is maintained and issues not related to the focal point or central conflict are regarded as lower priority. In general, the focus of treatment is developed collaboratively before therapy starts or immediately following the initiation of therapy. The therapist actively works to maintain the treatment focus on the agreed upon core problem.
6. STPD is intended to alleviate specific symptoms and provide limited personality or character change. This is in

contrast to psychoanalysis and long-term psychodynamic therapy where the goal is typically pervasive character change.

7. In general, the use of advice and homework, often practiced in behavioral and cognitive psychotherapies, is not recommended.

Because most forms of STPD do not fulfill all of the aforementioned requirements, these criteria cannot be considered necessary. As components of a prototypic definition, the more criteria a therapy meets, the more consistent it is with STPD.

Historical Background

Modern psychotherapy can be traced to Sigmund Freud's work beginning in the late nineteenth century. Freud developed a method to investigate the mind that ultimately evolved into the therapeutic intervention known as psychoanalysis. In traditional psychoanalysis, the analyst attempts to bring repressed unconscious conflicts to the patient's awareness. By gaining understanding ('insight') into the unconscious aspects of their problems, patients have the opportunity to work them through and subsequently master their difficulties. Although Freud originally worked with patients for a small number of sessions, the duration of psychoanalytic treatment increased into a multiyear process. Gradually, psychodynamic psychotherapy, also referred to as psychoanalytic psychotherapy, emerged as a modified application of psychoanalytic principles and therapeutic techniques.

Long-term psychodynamic therapists tend to be relatively quiet, wait for patients to free associate, and observe how the patient–therapist relationship (or transference) develops. This approach, quite consistent with psychoanalysis, still differs from traditional psychoanalysis in significant ways. While psychoanalytic patients lie on the couch 4–5 times a week, long-term psychodynamic patients sit face-to-face with the therapist once or twice a week for an extended period of time.

Short-term dynamic psychotherapists have modified the traditional principles even further. Work in this area originated with Ferenczi (1950), Rank (1936), and Alexander and French (1946), who recommended the maintenance of an active stance in therapy to hasten the exploration of unconscious material. In spite of their description of a seemingly useful treatment, many psychodynamic clinicians continued to view STPD as inferior to psychoanalysis or long-term psychodynamic therapy. It was not until Malan (1976), Mann (1973), Sifneos (1979), and Davanloo (1980) that STPD was deemed a valuable treatment option. The works of these authors are regarded as the four 'traditional' approaches to STDP (Crits-Christoph et al., 1991). They established a clear theoretical framework and sets of techniques to make focused and brief psychodynamic treatment possible. Malan's emphasis on the importance of careful patient selection through the screening of inappropriate referrals and trial interpretations attuned therapists to the fact that STPD can be effective for a subset of the patient population. Malan and Sifneos were also among the first to stress the significance of defining and maintaining a therapeutic focus as a central aspect of STPD technique.

Malan (1976) expressed what he saw as the essential psychodynamic theory in his description of two triangles. The first triangle is the 'triangle of conflict.' Its apices correspond to 'defense,' 'anxiety,' and 'underlying feeling or impulse.' The second triangle is the 'triangle of person' and refers to 'relationships with current figures,' 'relationship with the therapist (representing transference),' and 'relationships with important figures from the past' (e.g., parents). According to Malan, the therapist's task is to expose the underlying feelings and impulses that the patient has been protecting via defense mechanisms and to elucidate the role of the defenses in reducing the anxiety that the feelings create. He posited that the patient's hidden feelings were originally experienced in important early relationships and since then have frequently recurred with other significant figures in the patient's life, including the therapist. During therapy, the patient must understand the hidden impulses underlying each of the relationships described in the triangle of person. Typically, insight into a current relationship (either with a significant other or the therapist) is achieved first and is then related back to the parental figures.

Davanloo's intensive short-term dynamic psychotherapy (ISTDP) went further and supposedly broadened the range of patients who could be helped with STDP. He emphasized accessing the unconscious emotional content that is at the core of a number of psychoneurotic or borderline disorders by following a series of well-defined phases of clinical processes and interventions (Abbass and Town, 2013). The ISTDP therapist encourages the patient to engage with the emotional content of his or her past experiences (i.e., *pressure*), which in turn activates unconscious anxiety and related defenses. Under *pressure*, the patient experiences *complex transference feelings* and attempts to resist them. As the therapist actively and tenaciously clarifies the resistance and challenges the mobilized defenses, the patient eventually experiences *transference feelings* through somatic pathways, which in turn alleviates unconscious anxiety and eliminates resistance. Thus, the unconscious becomes *unlocked*, and the patient and therapist are able to collaboratively review unresolved, unconscious emotions, especially those related to attachment trauma. As needed, the process is repeated to unlock more unconscious material and work through it collaboratively (Abbass and Town, 2013).

Toward the end of twentieth century, a new generation of short-term psychodynamic therapists appeared on the clinical scene. Among the most important are Luborsky (1984), Horowitz et al. (1984), Strupp and Binder (1984), and Weiss et al. (1986). This new group can be distinguished from previous therapists by their interest in the empirical status of their treatments. Perhaps resulting from their interest in research, this new generation of therapists has written detailed descriptions of their clinical approaches that are very helpful for training and monitoring clinicians in these techniques. In fact, many of their books are considered treatment manuals and have been used in empirical research addressing the efficacy of these therapeutic methods. For example, Barber et al. (2012) recently tested the efficacy of

supportive expressive therapy (SET), (Luborsky, 1984) for depression and Leichsenring et al. (2013) tested its efficacy for social anxiety disorder. SET is an STDP model that has a relatively large amount of research regarding both its central concepts and its efficacy.

Recent developments include protocolized psychodynamic treatments for specific diagnoses. For example, Milrod et al. (1997) developed a treatment manual for the psychodynamic treatment of panic in 24 sessions. Weissman et al. (2007) developed interpersonal therapy for the treatment of depression. While there is some debate about the extent to which interpersonal therapy relies on psychodynamic principles and techniques, it is consistent with many of the criteria mentioned above and was developed to represent a time-limited researchable form of dynamic therapy. The theory behind this treatment holds that depression is caused by a loss or role transition, which the patient must understand using a relationship-focused model; and emphasis is placed on developing a current adaptation that will meet the present emotional needs of the patient and also provide a better solution to the conflicts stirred up by the loss. Emotional exploration of the loss in a time-limited treatment setting is a key element of the technique.

More recently, there has been an increased interest in the concept of mentalization. According to Bateman and Fonagy (2006), mentalization is the process by which we implicitly and explicitly interpret the actions of ourselves and others as meaningful on the basis of intentional mental states. Problems in mentalization are thought to be the basis for a variety of psychiatric problems, chiefly borderline personality disorder. Mentalization-based therapy (MBT) was developed for this patient population to increase mentalization capacity and thereby improve affect regulation and interpersonal relationships. This treatment, which is just beginning to be investigated outside of Borderline Personality Disorder, is delivered in a variety of formats, some of which are consistent with our prototypic definition of STPD. MBT is now being studied for the treatment of depression (Lemma et al., 2011).

The short-term psychodynamic treatments most recently developed, such as panic-focused psychodynamic therapy and MBT, were individually designed to focus on a particular diagnostic problem. Summers and Barber (2009), in their attempt to synthesize contemporary psychodynamic therapy theory and technique, describe an umbrella of treatments, all of which are psychodynamic and share similar underlying principles. They survey the evidence base for psychodynamic therapy efficacy and propose a model emphasizing the conceptual and technical building blocks of the therapy. These building blocks are: therapeutic alliance development, diagnosis of a core psychodynamic problem, psychodynamic formulation, goal setting, new narrative strategies for facilitating change, and the role of positive emotion. An important element of their work is the careful delineation of six core psychodynamic problems – depression, obsessionality, fear of abandonment, low self-esteem, panic, anxiety and trauma – which provide the clinician a roadmap for anticipating and working with patients with these problems; and the relevant modifications of psychodynamic technique are regarded as most helpful. The model is used in both shorter term and longer term settings.

Another noteworthy recent development in STDP is an alliance-focused, manualized treatment model, referred to as Brief relational therapy (BRT; Safran and Muran, 2000), which was derived from research on ruptures in the therapeutic alliance, and incorporates principles from contemporary psychoanalysis (e.g., multiple selves, intersubjective negotiation) and from research on affect regulation and interpersonal process. BRT involves an intensive 'here-and-now' exploration of patient–therapist interactions and intersubjectivity, following the technical principle of meta-communication (i.e., communication about the communication process) with the ultimate aim of expanding 'awareness in relation' – as opposed to resolving core conflicts or correcting pathogenic beliefs. It has been applied as a 30-session protocol to personality-disordered patients with promising empirical support (Muran et al., 2005).

Efficacy of STPD

During the 1980s, researchers began to explore the efficacy of STPD relative to other psychotherapies. Representing the trend toward methodological rigor that prevailed throughout the 1980s and 1990s, manualized STPDs have slowly been incorporated into research studies to provide the standardization of treatment that is needed for the studies to be credible (*see* Psychological Treatment, Effectiveness of). Leichsenring et al. (2004), published one of the most important meta-analysis of STPD in a prestigious journal and showed that when one looked at high-quality studies of STPD ($n = 14$), STPD is significantly more effective than control conditions and not different from alternate treatment. Because the meta-analysis published by Leichsenring et al. did not differentiate among disorders and needed to be updated, several meta-analyses have been conducted since. Most recently, Barber et al. (2013) examined the efficacy of STPD for specific disorders focusing on depressive, anxiety, and personality disorders. They conducted separate meta-analyses involving the comparisons of STPD with control conditions and STPD with individual treatments for each family of disorders. They reported that STPD was significantly more effective than control conditions for depression (5 studies, medium effect size 0.57) and that STPD did not differ from alternative active treatments (12 studies, nonsignificant finding $g = -0.09$); those results were maintained at short-term follow up (7 studies, nonsignificant difference, $g = 0.03$) and at long-term follow up (6 studies, nonsignificant difference, $g = -0.12$). Barber et al. (2013) reported that STPD for patients with anxiety disorders was significantly more effective than control conditions (3 studies, large effect size 0.78) and that STPD did not differ from alternative active treatments (8 studies, nonsignificant finding $g = 0.08$). These results were maintained at short-term follow up (6 studies, nonsignificant difference, $g = -0.16$) and at long-term follow up (6 studies, nonsignificant difference, $g = -0.15$). Patients with personality disorders were more effectively treated with STPD compared to control conditions (7 studies, medium effect size, $g = 0.60$) and STPD did not differ from alternative active treatments (7 studies, nonsignificant finding $g = -0.16$); these results were maintained at short-term follow-up (5 studies, nonsignificant difference, $g = -0.06$).

In a recent development, meta-analyses have been conducted for specific forms of STDP. Abbass et al. (2012) concluded that ISTDP produced large effect sizes as compared to control conditions, and that ISTDP patients maintained gains in general psychopathology and interpersonal functioning at follow up. Most importantly, not only is ISTDP highly effective, but it has also been shown to be a highly cost-effective treatment, resulting in significant reduction in a variety of health care costs as patients maintain their treatment gains and become more functional (Town et al., 2013).

In summary, STDP is more effective than control conditions for a variety of disorders and does not differ from alternative treatments (including cognitive behavior therapy) for a range of disorders. There is some evidence that therapy adding STDP to medication is more effective than medication alone ($g = 0.47$) (Barber et al., 2013).

Process Research

Most of the research associated with STPD involves the study of the therapeutic process underlying therapy in general and STPD in particular. There is a rich literature and it is impossible to give it justice in a brief essay. In their review of the literature on the therapeutic alliance and the other mechanisms of change, Barber et al. (2013) reported that the alliance was associated with outcome (studies, $r = 0.2$). Strong findings in the literature on process research indicate that both nonspecific therapeutic factors (such as a positive therapeutic alliance) and specific therapeutic factors (such as accurate interpretations of the patient's core interpersonal conflicts, change in the maturity of psychological defense mechanisms, and increased self-understanding) are associated with good therapeutic outcome (Barber et al., 2013).

The Future of STPD

Short-term dynamic therapy has flourished in the last 30 years because of cultural change in beliefs about mental health treatment, increased financial pressure to shorten treatment, increased patient interest in effective and quicker treatment, and the social demand for greater transparency. There are burgeoning new forms of STPD as well as increased utilization. Some would say there are more therapies than can be adequately studied. There is little reason to believe that the multiple causes influencing the trend toward brief treatments, including STPD, will reverse any time in the coming decades.

We anticipate at least four lines of change in the field of psychotherapy relevant to STDP in the coming decades: an increased trend toward integration, more interest in defining those subpopulations best helped by different treatment packages and interventions, more emphasis on monitoring outcome, and more need for specific training in diagnostic-specific treatment.

The Trend toward Integration

We are witnessing the increasing use of techniques that are developed for one treatment and are rapidly adapted for other treatments. For example, accelerated experiential dynamic psychotherapy borrows from experiential therapy and targets the deepening of affective experience as the mechanism of therapeutic change (Fosha, 2000; Fosha and Slowiaczek, 1997), while maintaining a relational and psychodynamic therapeutic frame. Most dynamic therapists today pay attention to faulty thinking and dysfunctional attribution, importing traditional cognitive techniques. Similarly, many cognitive behavioral therapists speak about working on the therapeutic relationship and dealing actively with transferential issues. As STPD is becoming increasingly more integrative and eclectic, examining the efficacy of the actual treatments used in clinical settings and defining a given treatment delivered by a specific clinician as STPD is becoming increasingly difficult. Modern psychotherapy including STPD, must constantly evolve to fit with the fast-paced nature and current trends of the modern world. Thus, it is likely that the future practice of STPD will include more components from other psychotherapies.

A Trend toward Individualized Psychotherapies

One conclusion from the research literature on medicine (and especially on psychotherapy outcomes) is that there is no one therapy that works for everyone with a particular diagnosis. More and more, clinicians and researchers agree that a range of treatments are needed to optimize outcome for patients with a specific diagnosis. For example, Barber et al. (2012) recently showed that STPD was not differentially effective from medication or pill placebo for all patients; but it was more effective than the latter for minority male patients.

A Greater Emphasis on Monitoring Patient Outcomes

With the increasing requirement for accountability and the monitoring of patient outcomes, there is likely to be more pressure on clinicians to use evidence-based interventions and for those interventions to lead to a good outcome for a specific patient. There is interesting evidence suggesting that when patients' outcomes are communicated to therapists, these outcomes tend to improve over time.

A Great Need for Specific Training in Diagnostic-Specific Treatment

As the evidence for STDP strengthens, it will be necessary to train more clinicians. This is especially important for the treatment of specific DSM patient populations.

See also: Psychoanalysis: Current Status; Psychological Treatment, Effectiveness of; Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts; Values, Psychology of.

Bibliography

Abbass, A.A., Town, J.M., 2013. Key clinical processes in intensive short-term dynamic psychotherapy. Psychotherapy 50 (3), 433–437.

Abbass, A.A., Town, J.M., Driessen, E., 2012. Intensive short-term dynamic psychotherapy: a systematic review and meta-analysis of outcome research. Harvard Review of Psychiatry 20 (2), 97–108.

Alexander, F., French, T.M., 1946. Psychoanalytic Therapy. Ronald Press, New York.

Barber, J.P., Barrett, M.S., Gallop, R., Rynn, M., Rickels, K., 2012. Short-term dynamic therapy vs. pharmacotherapy for major depressive disorder. Journal of Clinical Psychiatry 73 (1), 66–73. http://dx.doi.org/10.4088/JCP.11m06831.

Barber, J.P., Muran, J.C., McCarthy, K.S., Keefe, R.J., 2013. Research on psychodynamic therapies. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. John Wiley & Sons, Inc, New York, pp. 443–494.

Bateman, A., Fonagy, P., 2006. Mentalization-based Treatment for Borderline Personality Disorder: A Practical Guide. Oxford University Press, Oxford.

Crits-Christoph, P., Barber, J.P. (Eds.), 1991. Handbook of Short-term Dynamic Psychotherapies. Basic Books, New York.

Crits-Christoph, P., Barber, J.P., Kurcias, J.S., 1991. Historical background of short-term dynamic therapy. In: Crits-Christoph, P., Barber, J.P. (Eds.), Handbook of Short-term Dynamic Psychotherapy. Basic Books, New York, pp. 1–16.

Davanloo, H. (Ed.), 1980. Short-term Dynamic Psychotherapy. Jason Aronson, New York.

Ferenczi, S., 1950. The further development of an active therapy in psychoanalysis, (J.I. Suttie, Trans.). In: Ferenczi, S. (Ed.), Further Contributions to the Theory and Technique of Psychoanalysis, second ed. Hogarth Press, London, pp. 198–217 (Original work published 1926).

Fosha, D., 2000. The Transforming Power of Affect: A Model for Accelerated Change. Basic Books, New York.

Fosha, D., Slowiaczek, M.L., 1997. Techniques for accelerating dynamic psychotherapy. American Journal of Psychotherapy 51, 229–251.

Gerber, A.J., Kocsis, J.H., Milrod, B., Roose, S.P., Barber, J.P., Thase, M.E., Perkins, P., Leon, A.C., 2011. A quality-based review of randomized controlled trials of psychodynamic psychotherapy. American Journal of Psychiatry 168, 19–28. http://dx.doi.org/10.1176/appi.ajp.2010.08060843.

Horowitz, M., Marmar, C., Krupnick, J., Kaltreider, N., Wallerstein, R., Wilner, N., 1984. Personality Styles and Brief Psychotherapy. Basic Books, New York.

Leichsenring, F., Rabung, S., Leibing, E., 2004. The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: A meta-analysis. Archives of General Psychiatry 61, 1208–1216.

Leichsenring, F., Salzer, S., Beutel, M.E., Herpertz, S., Hiller, W., Hoyer, J., Huesing, J., Joraschky, P., Nolting, B., Poehlmann, K., et al., 2013. Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: a multicenter randomized controlled trial. The American Journal of Psychiatry 170, 759–767. Epub ahead of print May 17, 2013.

Lemma, A., Target, M., Fonagy, P., 2011. Brief Dynamic Interpersonal Therapy: A Clinician Guide. Oxford University Press, Oxford, 978-0-19-960245-2.

Luborsky, L., 1984. Principles of Psychoanalytic Psychotherapy: A Manual for Supportive–Expressive (SE) Treatment. Basic Books, New York.

Malan, D.H., 1976. The Frontier of Brief Psychotherapy. Plenum, New York.

Mann, J., 1973. Time-limited Psychotherapy. Harvard University Press, Cambridge, MA.

Messer, S.B., Warren, C.S., 1995. Models of Brief Psychodynamic Therapy: A Comparative Approach. Guilford Press, New York.

Milrod, B., Busch, F.N., Cooper, A., Shapiro, T., 1997. Manual of Panic–Focused Psychodynamic Psychotherapy. American Psychiatric Press, Washington, DC.

Muran, J.C., Safran, J.D., Samstag, L.W., Winston, A., 2005. Evaluating an alliance-focused treatment for personality disorders. Psychotherapy 42, 532–545.

Rank, O., 1936. Will Therapy. Knopf, New York.

Safran, J.D., Muran, J.C., 2000. Negotiating the Therapeutic Alliance: A Relational Treatment Guide. Guilford Press, New York.

Sifneos, P., 1979. Short-term Dynamic Psychotherapy: Evaluation and Technique. Plenum Medical Book Co, New York.

Strupp, H.S., Binder, J.L., 1984. Psychotherapy in a New Key: A Guide to Time-Limited Dynamic Psychotherapy. Basic Books, New York.

Summers, R., Barber, J.P., 2009. Psychodynamic Therapy: A Guide to Evidence Based Practice. Guilford Press, New York.

Town, J.M., Abbass, A., Bernier, D., 2013. Effectiveness and cost effectiveness of Davanloo's intensive short-term dynamic psychotherapy: does unlocking the unconscious make a difference? American Journal of Psychotherapy 67 (1), 89–108.

Weiss, J., Sampson, H., Mount, Zion, 1986. The Psychoanalytic Process: Theory, Clinical Observation and Empirical Research. Guilford Press, New York.

Weissman, M.M., Markowitz, J.C., Klerman, G.L., 2007. Clinician's Quick Guide to Interpersonal Psychotherapy. Oxford University Press, New York.

Psychohistory

Rudolph Binion, Brandeis University, Waltham, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12339–12343, © 2001, Elsevier Ltd.

Abstract

Psychohistory developed primarily in America beginning in the 1950s. Two main tenets distinguish its approach to the human past. One is that history is strictly what people did, with its supposedly impersonal developments reducing to so much human performance. The other is that the determinants of human conduct whether individual or collective are largely unconscious. In this, as in its abiding concern with childhood and pathology, psychohistory shows its derivation from psychoanalysis. But it has outgrown its initial dependence on clinical theory to develop an agenda of its own stressing behavior in and by groups. It has also originated concepts later adopted in clinical theory, such as traumatic reliving. This key concept also instances a general trend in the field: mental mechanisms are replacing unconscious motivations as explanatory devices. The methodological problems of sounding groups and verifying findings are daunting. With its rules of evidence tightening under outside criticism, the scholarly field is depopulating. At the same time, the need for psychohistorical insight is being felt increasingly in other walks of history, as for instance in historical demography to explain how populations regulate their total numbers. This growing felt need assures psychohistory of renewal as a discipline in its own right.

Its Meaning

The term *psychohistory* (originally written *psycho-history*) emerged in America in the 1950s to designate historical writings utilizing psychology, for the most part depth psychology, as their primary frame of reference or mode of explanation. Two of its tenets in particular distinguish psychohistory sharply from other systematic approaches to the human past. The first of the two, which is rarely spelled out even in methodological writings, is that the facts of human history consist entirely of what people did whether spontaneously or under pressure of circumstances. To the psychohistorian, an intellectual revolution thus signifies thinkers changing their minds; a decline in fertility translates as women bearing fewer children; urbanization conveys people living increasingly in cities. Other historians may agree that even such seemingly impersonal historic developments all come down to human doings, yet whenever other historians narrate or explain developments of the sort they tend in the first instance to see belief systems or birth rates or urban:rural ratios changing rather than people thinking or acting differently. The other highly distinctive tenet of psychohistory is, by contrast, fully and even emphatically explicit in the psychohistorical literature: that the determinants of human conduct, whether individual or collective, are largely unconscious. Historians of other schools will often freely acknowledge deeply subjective causation of historic human behavior only to shy away from it in their own research as too dauntingly difficult to reconstruct, let alone demonstrate. To the psychohistorian, however, whatever the difficulties involved, there is simply no other way for history to be understood (*see* History, Theories and Methods).

Its Development

Psychohistory originated as applied psychoanalysis, with Freud and his disciples drawing on their clinical findings now and again to illuminate historical facts. At first, those clinical findings bore heavily on the patterns and pitfalls of individual emotional development, with a strong emphasis on pregenital sexuality as it emerges in childhood bit by bit and then undergoes repression even while continuing to inform emotional life. Accordingly, the earliest applications of psychoanalysis to history – including those by Freud himself, who long set the standard for his followers – were biographic and focused on the infantile concerns latent in the words and deeds of eminent or paradigmatic historic figures (*see* Biography: Historical). A landmark work in this category was psychoanalyst Erik H. Erikson's *Young Man Luther* of 1958, which argued that Luther's bold quest for spiritual autonomy was that of his whole age writ large (*see* Erikson, Erik Homburger (1902–94)). Unlike biography in relation to historiography as a whole, psychobiography remains to this day the basic as well as the leading form of psychohistory. At the same time, it was as quick as psychoanalysis itself to move beyond its initial Freudian orthodoxy (*see* Freud, Sigmund (1856–1939)). Post-Freudian concepts utilized in psychobiography when this is practiced as applied psychoanalysis have notably included C. G. Jung's 'individuation,' Alfred Adler's 'inferiority complex,' Erik H. Erikson's 'identity crisis,' Heinz Kohut's 'narcissistic rage,' and Jacques Lacan's 'language of the id.'

Psychohistory long continued to show its derivation from psychoanalysis even where it broadened its scope beyond psychobiography. As late as the 1970s psychohistorians who explored social groups, whole cultures, or long-range collective developments, did so as a rule only in terms of the individual psychodynamics thought to have determined them, tracing their salient characteristics to formative experiences widely shared by the individuals involved, and most often to pathogenic childrearing practices on the order of abrupt weaning or intrusive parenting. Correspondingly, the preferred subjects under the head of group behavior were instances of social pathology such as witchcraft, or of political pathology such as Nazism. The most systematic psychohistorical approach to group behavior to emerge on these individualistic premises is prosopography duly slanted: the profiling of the likes of saints

or witches or Nazis in terms of variables of psychoanalytical significance in their individual family histories – ages of parents and siblings, for instance, or again early parent or sibling losses (*see* Prosopography (Collective Biography)).

A prominent early alternative to the individual personality as the basic unit even of large-scale, broadgaged psychohistorical accounts or analyses was the study of a mass leader whose following interacted with him like a single composite being. Freud himself pioneered this departure from the classic individualistic model that underlay his own science of psychoanalysis. In his *Group Psychology and the Analysis of the Ego* (*Massenpsychologie und Ichanalyse*) of 1922 he treated mass phenomena for the most part as simultaneous, parallel regression by individuals, but in a couple of highly suggestive passages he speculated on human consciousness as originally shared by a Darwinian primal horde from which a leader eventually emerged with a mind and will of his own; such shared group consciousness, he considered, might revive at critical historic junctures under the spell of a charismatic leader who threw back to that aboriginal tribal chief. Freud (1939) worked this speculative construction into a highly imaginative reading of the Moses story, with Moses reincarnating the tribal father supposedly slain and later exalted by his sons acting in concert. Given the very nature of this concept of group consciousness as a reversion to a tribal leader and following, subsequent psychohistorical treatments of leader and led have mostly dealt with political collectivities such as Lincoln's or Nixon's America, or again William II's and especially Hitler's Germany.

On the Freudian model of group process, then, group identity constitutes a regressive throwback to an original shared consciousness in the face of a strong leader. But a rival conception of shared mental life, one that makes it out as permanent and leaderless, came into psychohistory independently. Its *locus classicus* lies outside of psychohistory: in the 'collective consciousness' (*conscience collective*) that Emile Durkheim in his *Suicide* of 1897 inferred from the patterned regularity of suicide rates over and beyond the particular subjective causation of each individual suicide. An early specimen of such a leaderless group model in psychohistory was Caroline E. Playne's *The Neuroses of the Nations* of 1925, which owed nothing to psychoanalysis and indeed wholly ignored it. When William L. Langer, in a resounding presidential address to the American Historical Association in 1957, called the mastery and use of depth psychology the historian's 'next assignment,' he expressly meant its application to leaderless group phenomena, which he exemplified by the epidemic of religious and cultural pathology that followed in the wake of the Black Death of 1348. (Oddly, Langer made no reference to a pioneering work in this mold just then published by his Harvard University colleague Walter Abell (1957) about how the waxing and waning insecurities of medieval European life were projected into Europe's public art of the time.) By the 1970s the dynamics of 'group process,' as it came to be called, topped the psychohistorical agenda, with collective fantasy as its privileged medium of analysis. An exemplary early product of this approach to group process was Paul Monaco's (1976) work, which teased the deep-lying national anxieties of France and Germany in the 1920s out of the recurrent motifs and techniques of each country's most popular movies of those years. Another choice area, still under exploited, for ferreting out the unconscious underpinnings of group activity is historical demography (*see* Historical Demography), especially the adjustments of aggregate birth rates to the felt needs of a population with regard to its overall numbers.

As the term 'group process' itself implies, explanation in psychohistory gradually has shifted from unconscious motives and meanings to unconscious mechanisms governing thought and conduct, while conscious purposes have been even more marginalized in the process. This shift reflects the continuing self-distancing of psychohistory from its parent discipline of psychoanalysis, focused as the latter is on latent impulses originating in childhood. Nowhere is this shift more evident than in the concept of traumatic reliving, which matured within psychohistory well before it was taken up in psychoanalysis. Psychoanalysis had first confronted adult psychic trauma during World War One in the form of battle shock for which, as for other emotional ailments, it sought a childhood referent that in this case was not forthcoming. Comparative psychobiography established two patterned, alternative reactions to emotional trauma: either a continual recall accompanied by the full, tormenting original affect, or a repression of the traumatic affect, and sometimes even of the traumatic memory itself, followed by a contrived, controlled, disguised reexperience or reliving of the trauma to the apparent end, never attained, of mastering the attendant affect that could not be mastered at the time of the experience itself. These same mechanisms have subsequently been shown to operate as well on the collective level in instances of group trauma, and to operate independently of the widely varying reactions of the individuals composing the traumatized group. Precursively, William L. Langer in handing out his 'next assignment' conceived of the after effects of the Black Death, notably its morbid reenactments such as the Dance of Death, as just such a case of group traumatic reliving.

Its Problematic

Psychohistorical research is uniquely demanding of its practitioners because of its concern with subjective phenomena that are rarely if ever directly recorded. To understand a piece of the human past psychohistorically, researchers must first master the relevant facts at least as thoroughly as the practitioners of any other mode of inquiry. But much more, they must become immersed in those facts, make them vicariously their own, and feel out their inner determinants by dint of this imaginative identification with their original experience. The method is strikingly like the one promoted by Henri Bergson under the name of 'intuition': to study a subject intellectually from the outside until one enters into it subjectively through a core insight around which everything known about it will then fall into place. Such intimate, personal self-projection into a historic subject is trebly problematic. For one thing, it presupposes a basic equivalence of mental and emotional constitution as between researchers and their human subjects, however remote those subjects may be from them historically, and even if they are groups rather than individuals like

themselves. For another, researchers' identification with their subject must be thorough-going, yet unless they can also sustain some critical awareness outside of that identification the insights gained through it will be lost. And for a third, pieces of the researchers' own mental underworld may slip unnoticed into their reading of a subject. This danger, often called 'countertransference' by loose analogy with psychoanalysts projecting into their patients, is the risk most cautioned against by psychohistorians themselves. Actually, an affinity in researchers for their subject's inner experience can, on the positive side, sensitize or alert them to connections and meanings they might otherwise miss. Besides, a purely projective finding will not stand up in any case inasmuch as it will fail to fit all the evidence, let alone satisfy anyone apart from the researchers themselves.

Researchers will feel the rightness of their inner grasp of a subject when it not only pulls together all the relevant facts at their disposal, but points the way to unsuspected additional corroborative evidence. Yet the discovery of such new evidence, however suggestive, is in itself no proof that their inner grasp of their subject is universally valid. Because of the very nature of psychohistorical explanation, to validate a psychohistorical thesis takes more than just a wealth of corroborative material, even if the thesis itself is what turned up key pieces of that material. To pass muster, a psychohistorical thesis must also carry subjective conviction, beyond the researchers themselves, with others willing and able to put themselves in the subject's place in their turn and see whether the proposed unconscious source of the thought or action at issue checks out subjectively with them too.

This second, subjective criterion of verification means in brief that a psychohistorical thesis must be emotionally as well as intellectually compelling. Consequently it can rely on no pregiven theory for support. Yet however direct its appeal, its need for subjective on top of objective verification demands too much effort, not to add goodwill, ofcritics accustomed instead to merely material and logical criteria of validation. As a substitute test of validity I therefore propose that, in psychohistory as in any science, a thesis is confirmed if, first of all, the known evidence all runs its way, if moreover it could potentially be refuted by new evidence, and if finally any piece of what it purports to explain cannot very well be explained otherwise. To take a simple example of this last stipulation from my own thesis that Hitler was traumatized by his mother's iodoform poisoning during her terminal cancer treatment: why else did he put it to his generals in late 1944 that his Ardennes offensive had caught the Allies off guard because they believed he was dead already 'or at all odds am down with cancer somewhere and can't live any more or drink any more' except that deadly iodoform poisoning induces a burning thirst together with an inability to drink? (Binion, 1976, p. 143).

Its Crisis Years

As a historical school in its own right, psychohistory flourished above all in the US during the 1960s and 1970s. Even then, like all new departures, it was highly controversial, especially within the historical professions, because of its constitutive tendency to undercut all other modes of historical explanation. It has since suffered considerably from the decline of psychoanalysis, with which it is rightly associated in that the two still share at least the premise of unconscious motives and meanings. Experimental graduate programs in the field have all folded, while course offerings on college campuses have dwindled to the vanishing point. The ranks of researchers in the field are fast thinning even as the stringent standards of verification invoked for it by critics and adepts alike are discouraging prospective practitioners. Fewer psychohistorical books and articles are appearing year by year. Of the two major specialized journals, *The Psychohistory Review* ceased publication in 1998 while *The Journal of Psychohistory* has lost most of its initial circulation. Perhaps worst of all for the discipline, the consensus of outsiders is that its scholarly output is falling off qualitatively as well.

On the other hand, psychohistory has left its impress on most other kinds of historical inquiry, especially those that emerged or developed later than psychohistory itself. Biographies and even general histories now routinely incorporate psychohistorical materials and interpretations. The study of mentalities often verges on psychohistory. So does that of historic memory, especially in the concern with memory gaps or distortions that has dominated the field. Gender history and gay history, as well as the history of emotions, lean heavily on psychohistorical concepts and findings. Many a term launched in psychohistory, such as Erikson's 'identity crisis,' has become common coinage elsewhere, while the mechanism of traumatic reliving by individuals and groups alike has found its way into a broad variety of scholarly contexts, including even a quantitative history of testation in late medieval Tuscany (Cohn, 1992) (*see* Identity in Childhood and Adolescence).

The fact that psychohistory has lost much of its vitality of the 1960s and 1970s, or again that it is being put to subordinate uses by historians of other orientations, does not affect its constitutive claim to provide a privileged access to causality in history. Given this inherently persuasive claim, neither the recent loss of momentum by psychohistory as a method in its own right nor its present auxiliary status in historical research can very well be regarded as permanent.

See also: Emotions, History of; Historical Thought and Historiography: Current Trends; Psychoanalysis, History of.

Bibliography

Abell, W., 1957. The Collective Dream in Art: A Psycho-historical Theory of Culture Based on Relations between the Arts, Psychology, and the Social Sciences. Harvard University Press, Cambridge, MA.

Aldrich, R., 1993. The Seduction of the Mediterranean: Writing, Art, and Homosexual Fantasy. Routledge, London.

Binion, R., 1976. Hitler among the Germans. Elsevier, New York.

Binion, R., 1994. Fiction as social fantasy: Europe's domestic crisis of 1879–1914. Journal of Social History 27, 679–699.

Brink, A., 1989. Bertrand Russell, A Psychobiography of a Moralist. Humanities Press International, Atlantic Highlands, NJ.

Cohn Jr., S.K., 1992. The Cult of Remembrance and the Black Death. Six Renaissance Cities in Central Italy. Johns Hopkins University Press, Baltimore, MD, pp. 116–117, 288.

Demos, J.P., 1970. A Little Commonwealth: Family Life in Plymouth Colony. Oxford University Press, New York.
Erikson, E.H., 1959. Identity and the Life Cycle. International University Press, New York.
Freud, S., 1939. Der Mann Moses unddie Monotheistische Religion. Verlag Albert de Lange, Amsterdam.
Graziano, F., 1992. Divine Violence: Spectacle, Psychosexuality. Radical Christianity in the Argentine 'Dirty War.' Westview Press, Boulder, CO.
Kohut, T., 1985. Mirror image of the nation: An investigation of Kaiser Wilhelm II's leadership of the Germans. In: Strozier, C.B., Ofer, D. (Eds.), The Leader: Psychohistorical Essays. Plenum, New York.
Langer, W.L., 1958. The next assignment. American Historical Review 63, 283–304.
Mack, J.E., 1977. A Prince of our Disorder: The Life of T. E. Lawrence. Weidenfeld and Nicolson, London.
Markowitz, J., 1969. The Psychodynamic Evolution of Groups, first ed. Vantage Press, New York.
Monaco, P., 1976. Cinema and Society: France and Germany during the Twenties. Elsevier, New York.
Roper, L.M., 1994. Oedipus and the Devil: Witchcraft, Sexuality, and Religion in Early Modern Europe. Routledge, New York.
Saslow, J.M., 1986. Ganymede in the Renaissance: Homosexuality in Art and Society. Yale University Press, New Haven, CT.

Psycholinguistics: Overview

Brian MacWhinney, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psycholinguistics combines methods and theories from psychology and linguistics. It attempts to evaluate the psychological reality and underpinnings of linguistic rules and processes. It also seeks to link word and sentence processing to the deeper expressive processes of message construction and interpretation. Modern psycholinguistic theories emphasize the ways in which distinct areas of the brain interact in a dynamic incremental way to process language.

Introduction

Psycholinguistics is a field that combines methods and theories from psychology and linguistics to derive a fuller understanding of human language. From psychology, it inherits experimental methodology and a body of knowledge about processes in perception, memory, attention, learning, and problem solving. From linguistics, it derives detailed descriptions of specific languages, rigorous accounts of the shape of grammar, and ideas about the nature of human language.

The initial issue that motivated the establishment of psycholinguistics as a separate field of study was the problem of the *psychological reality* of linguistic concepts (Sapir, 1933/1968). In the 1960s, researchers wanted to know whether the various rules of Chomsky's new Transformational Generative Grammar (Chomsky, 1957) are actually used by humans. Although the results for this particular issue were largely negative, the core interest in psychological reality has continued unabated. To take a simple example of the issue, consider the claim that speakers of English can form the plurals of nouns by adding the suffix – *s*. This process allows us to form the plural *chandeliers* from *chandelier*. But do we actually use a productive rule to produce this plural or do we simply retrieve the plural *chandeliers* from our long-term memory as a unit? Psycholinguistic research shows that, in fact, both rote and rule are operative at various times in language production (Pinker, 1999).

This issue of psychological reality applies to all levels of language structure and usage, including articulatory phonetics, auditory phonetics, phonology, morphology, lexicon, syntax, semantics, and pragmatics. It applies to all of the constructs of linguistics, including rules, grammars, paradigms, trees, segments, words, and morphemes. When we hear words, do we break them up into their component phonemes or 'letters' or do we recognize them as wholes? When we listen to sentences, do we actually construct treelike representations of the types proposed by linguists or do we somehow extract meaning without building up formal structures? If we do use formal grammars to listen and speak, what is the exact shape of the grammars that we use?

Exploration of the psychological reality of linguistic structures immediately leads us to two related fields of study. The first is developmental psycholinguistics or the study of child language acquisition (Ambridge and Lieven, 2011). If we believe that adults form the plurals of nouns by adding – *s*, we need to consider how young children apply this rule to produce *cats*, but not *foots* or *tooths*. If we argue that children simply learn each form by rote, how can we account for the fact that they make errors like *feets* and the fact that they can produce the plural for a new word like *wug*, even without having been given that form explicitly?

The issue of psychological reality also leads us to consider the even more difficult issue of how rules of grammar might be processed in the brain. Addressing this question has led psycholinguists to explore issues in neurolinguistics (Stemmer and Whitaker, 2008) and cognitive neuroscience (Gazzaniga, 2000). When the field of psycholinguistics first developed in the 1950s, psychologists knew little about the detailed functioning of the human brain and were forced to treat it as a 'black box.' However, as our understanding of the functioning of the human brain grew during the 1980s and 1990s, it became clear that a precise understanding of the functioning of human language would have to make reference to neural mechanisms. The interactive approach to cognition (Rumelhart and McClelland, 1986) used artificial neural networks to model the processing of human language. Directly opposed to connectionism was Fodor's (1983) modular approach to cognition that emphasized the independence of separate cognitive modules for each level of linguistic structure. Researchers have attempted to test the contrasting predictions of the interactive and modular approaches using standard experimental methodology. However, this work has indicated that neither of the strong positions can be maintained (Simpson, 1994).

This article will examine research in six core areas of psycholinguistics: spoken word recognition, sentence comprehension, sentence production, message construction, crosslinguistic comparisons, and neurolinguistics. In addition to these core areas, psycholinguists are involved in the study of reading, conversational interaction, figurative language, text comprehension, aphasia, child language disorders, gesture, prosody, animal communication, and language evolution. However, our analysis here will focus on these six core areas.

Word Recognition

When we listen to speech, we perceive words as following each other in clear temporal succession, like the beads on a string. However, this underestimates the extent to which words are

actually being blurred together by coarticulation and assimilation. Extreme examples of this blurring occur in forms such as *supchu* for 'What is up with you?' However, even nonphrasal sequences such as 'I owe you a yo–yo' show similar blurring. Even without blurring, the job of segmenting sentences into words would be a tough one. Consider a phrase such as 'my catalog value for Mark's piece.' Within this phrase, there are fragments that match a variety of other possible words, such as *mike*, *eye*, *I*, *cat*, *at*, *cattle*, *log*, *you*, *Val*, *ark*, *arks*, *are*, and *form*. The reason why we tend not to hear these alternative forms is fairly simple. If we decided to commit ourselves to having heard the word *mike*, we would end up with the nonsensical segmentation of 'mike at a log value for Mark's piece.' At each point during sentence perception, many of these alternative words are partially active in a short list of competitors. Competitors that recognize larger segments such as 'catalog' are preferred over those that recognize smaller strings or which break up larger strings into pieces, such as 'cat a log.' If a competitor leaves an unrecognized fragment, then its own recognition is weakened. These constraints work together to guarantee a maximally satisfactory segmentation. Although segmentation is primarily driven by the competition of words for matches, it is also facilitated by stress patterns, pauses, and other prosodic patterns (McClelland et al., 2006).

Segmentation relies on word recognition, but this process is also highly dynamic. Words do not have clear and invariant forms. Consonants leave no clear and invariant imprint on the auditory stream, as they are heavily blended into vowels. Because of differences in their vocal tracts, men and women produce vowels in radically different ways. Dialect differences and variations in speech level add further variability to word forms. Again, the solution to this problem involves lexical competition. For example, the sound corresponding to word like *deep* will activate a neighborhood of similar words, such as *deal*, *dean*, *peep*, or *keep*. Within this domain of rhyming forms, competition is particularly keen between words that share the same initial segments, such as *deep*, *deal*, *dean*, and *deed*. The ability of a word to dominate in this competition is a function of the strength of its match to the features of the input (Dahan et al., 2001). Apart from influences coming from a word's sound, there can also be influences from meaning. For example, hearing the word *doctor* will prime the recognition in both listening and reading of the word *nurse* (Balota, 1994).

These competitive models can be given a specific neuronal instantiation in terms of a neural network model based on self-organizing feature maps (Li et al., 2007). In these models, an unorganized, but interconnected, flat sheet of simulated neural tissue is trained to recognize a set of input words. As learning progresses, words that share features move to adjacent areas in the feature map. In our example, the words *deep*, *deal*, *dean*, and *deed* would occupy a small neighborhood on the larger map. The network operates in a winner-take-all fashion so that, when deep starts to receive the highest level of activation, it will inhibit the activation of the competitors through a series of lateral connections. Like actual cortical tissue, these maps use spatial position to represent featural structure, and they rely on lateral inhibition to sharpen the outcome of a competition.

Sentence Comprehension

In the 1960s, psycholinguists thought that the process of word recognition completely preceded the process of sentence comprehension. The idea was that we must first identify all the words in a sentence before we could feed these words to a comprehension mechanism that decided what it all meant. Much of what we have learned about language processing over the last 50 years involves the construction of a detailed rejection of this early vision. As we study how auditory processing, word recognition, and sentence comprehension operate in real time, we have come to see that language processing is both *interactive* and *incremental*.

Incrementalism

In the late 1970s, Marslen-Wilson (1975) and others showed processing keeps up with word recognition in a fully incremental or online fashion. Both word recognition and sentence comprehension attempt to keep up with speech exactly as it comes in, although sometimes there is a slight lag, particularly for sentence comprehension. You might notice this lag if you find that you are trying to read a newspaper while someone is talking to you. You hear their words, but you may not be processing exactly what they say. If someone then asks you whether you have been listening, you may be able to repeat the last seven or eight words they said, but in fact you were not really understanding the message. Rather, you were squirreling away the words in a superficial form in your short-term memory (Botvinick and Plaut, 2006), hoping not to have to actually do the work of fully comprehending the message.

Apart from some such marginal cases of short-term buffering, we process incoming material incrementally, both lexically and conceptually. Take as an example the sentence, 'The boy chased the baboon into the bedroom.' As soon as we hear *the* and *boy*, we immediately begin to relate them to each other. We then relate this unit to the following verb *chased*. Milliseconds after hearing *chased*, we begin to construct an interpretation of the activity in which there is a boy doing some chasing. By the time we hear *baboon*, we can begin to sketch out the figure that the boy is chasing. We do not need to wait until we have heard all the words in the sentence to begin to extract these meanings. In this sense, sentence processing is both interactive and incremental – we tend to make decisions about material as soon as it comes in, hoping that the decisions that we make will not be reversed. Although processing is incremental, it only builds as much structure as it needs in order to keep words related. The real job of comprehension is delayed until more of the message is heard.

Garden Pathing

There are times when the initial decisions that we have made take us down the garden path. A classic example of garden-path processing occurs with sentences such as 'The communist farmers hated died.' It often takes the listener awhile to realize that it was the *communist* that died and that it was the *farmers* who hated the *communist*. Inclusion of

a relativizer to produce the form, 'The communist that farmers hated died' might have helped the listener sort this out. A somewhat different example is the sentence, 'The horse raced past the barn fell.' Here, we need to understand 'raced past the barn' as a reduced relative clause with the meaning 'The horse who was raced past the barn.' If we do this, the appearance of the final verb *fell* after *barn* no longer comes as a surprise.

Garden paths arise when a word or suffix has two meanings, one of which is very common and one of which is comparatively rare (MacDonald et al., 1994). In a sentence like 'The horse raced past the barn fell' the use of the verb 'raced' as a standard transitive verb is much more common than its use as the past participle in a reduced passive. In such cases, the strong meaning quickly dominates over the weak meaning. By the time we realize our mistake, the weak meaning is fully suppressed by the strong meaning and we have to try to comprehend the sentence from scratch. A classic garden-path example from Lashley (1951) is the sentence 'Rapid righting with his uninjured left hand saved from destruction the contents of the capsized canoe.' When this sentence is read aloud, listeners find it extremely difficult to understand the second word as *righting* rather than *writing*.

Lexical Effects

Current models of sentence processing emphasize the extent to which lexical and syntactic ambiguities of individual words trigger competing syntactic structures (MacWhinney, 1987; Trueswell and Gleitman, 2004). For example, there are two readings of the sentence 'Flying planes can be dangerous.' The planes can be either dangerous to their pilots and passengers or dangerous to onlookers down on the tarmac. Both interpretations of the participle *flying* are fairly strong. Because the two readings are of similar strength, they can compete with each other and no garden pathing arises. Another example of this type is 'He bought her pancakes,' in which *her* can be either the possessor of the pancakes or the recipient of the pancakes. Both meanings are strong and can compete with each other during sentence processing, yielding a clear ambiguity.

Sometimes lexically based expectations can be fairly complex. Consider these sentences in which the verbs *criticize* and *apologize* set up contrasting expectations:

John criticized Mary, because she hadn't delivered the paper on time.
John apologized to Mary, because he hadn't delivered the paper on time.
John criticized Mary, because he hadn't delivered the paper on time.
John apologized to Mary, because she hadn't delivered the paper on time.

Processing of the first two sentences is quick and easy, because the gender of the pronoun matches that of the expected agent. However, the processing of the second pair is more problematic because the gender of the pronoun forces the selection of an unexpected causal agent of the criticism or the apology (McDonald and MacWhinney, 1995).

Modularity

Although processing on individual linguistic levels is highly incremental, the interaction between levels is not immediate. During the first 300 ms after hearing a word, we attend primarily to its auditory shape, rather than the degree to which it fits into some grammatical context. Take as an example the sentence 'The sailors took the port at night.' Here the word *port* could refer to either the wine or the harbor. We can ask subjects to listen to sentences like this while watching a computer screen. Directly after subjects hear the word *port*, we can present one of these three words on the computer screen: *wine*, *harbor* and some control word such as *shirt*. If we do this, we will find that the recognition of both *wine* and *harbor* is facilitated in comparison to the control word *shirt*. If we change the sentence to something like 'The sailors drank the port at night,' we might expect that the context would bias the subject to respond more quickly to *wine* than to *harbor*, because one is not likely to drink a harbor. However, there is evidence that both *wine* and *harbor* are facilitated in comparison to the control word *shirt*, even when the context tends to bias the *wine* reading of *port*. This facilitation is fairly short lived and the contextually appropriate reading soon becomes dominant.

This type of result indicates that, in the first fraction of a second after hearing a word, we rely most strongly on auditory cues to guide our processing. This is not to say that context is not present or not being used as a cue. However, during the first fraction of a second, we need to focus on the actual auditory form in order to avoid any 'hallucinatory' effects of paying too much attention to context too soon in processing.

Sentence Production

There are many similarities between sentence comprehension and sentence production. In both activities, we rely heavily on the words in our lexicon to control syntactic structures. Both activities make use of the same patterns for determining grammatical structures. The most important difference is that, during sentence production, we are in full control of the meanings we wish to express. In comprehension, on the other hand, we are not in control and have to follow the ideas of others.

The production of sentences involves at least four processes (Levelt, 1989). The first process is message construction. This process takes our goals and intentions and builds up a thread of ideas to be articulated. The process of lexical access then converts these ideas in worded forms. The third process uses positional patterning to order words into phrases and clauses. The fourth process activates a series of verbal gestures through articulatory planning. As in the case of comprehension, these four stages are conducted not in serial order, but in parallel, interactively and incrementally. Even before we have finished the complete construction of the message underlying a sentence, we begin the process of articulating the utterance. Sometimes we find out in the middle of an utterance that we have either forgotten what we want to say or do not know how to say it. It is this interleaved, incremental, online quality of speech production that gives rise to the various speech errors, pauses, and

disfluencies that we often detect in our own speech and that of others.

Speech errors come in many different forms. Some involve simple slurring of a sound or retracing of a group of words. Others provide more dramatic evidence of the nature of the language planning process. Some of the most entertaining speech errors are spoonerisms, which owe their name to an English clergyman by the name of William Spooner. Instead of 'dear old queen,' Spooner produced 'queer old dean.' Instead of 'ye noble sons of toil,' he produced 'ye noble tons of soil.' Instead of 'I saw you light a fire in the back quad, in fact you wasted the whole term,' he said 'I saw you fight a liar in the back quad, in fact you tasted the whole worm.' These errors typically involve the transpositions of letters between words. Crucially, the resulting sound forms are themselves real words, such as *liar*, *queer*, and *worm*. The tendency of these errors to produce real words is known as the 'lexical bias' in speech errors and indicates the extent to which the lexicon itself acts as a filter or checker on the articulatory process.

Another illuminating group of errors is named after a character named Mrs. Malaprop in the play *The Rivals* by Sheridan. Some examples of malapropisms include *Judas Asparagus* for *Judas Iscariot*, *epitaphs* for *epithets*, or *croutons* for *coupons*. Some malapropisms arise when uneducated speakers attempt to produce unfamiliar words, but many of these errors are true slips of the tongue. In a malapropism, the two words have a similar sound, but a very different meaning. The fact that speakers end up producing words with quite the wrong meaning suggests that, at a certain point during speech planning, the output processor handles words more in terms of their phonological form than the meaning they express. It is at this point that malapropisms can occur.

When two words end up directly competing for a single slot in the output articulatory plan, the result is a lexical blend. For example, a speaker may want to describe both the flavor and the taste of some food and end up talking about its *flaste*. Or we might want to talk about someone's *spatial performance* and end up talking instead about their *perfacial performance*. These errors show the extent to which words are competing for placement into particular slots. When two words are targeting the same slot, one will usually win, but if there is no clear winner there can be a blend.

Another remarkable property of speech errors is the way in which grammatical markers seem to operate independently of the nouns and verbs to which they attach. Consider an exchange error such as 'The floods were roaded' for 'the roads were flooded.' In this error, the plural marker – *s* and the past tense marker – *ed* go to their correct positions in the output, but it is the noun and the verb that are transposed. This independent behavior of stems and their suffixes indicates that words are being composed from their grammatical pieces during sentence production and that grammatical markers contain clear specifications for their syntactic positions.

Sigmund Freud believed that slips of the tongue could provide evidence for intrusions from underlying anxieties, hostilities, and worries. From this theory, arose the notion of a 'Freudian slip.' We now know that the majority of speech errors are not of this type, although there are some that can be viewed as intrusions from the context. For example, if a speaker is discussing some activities surrounding a local barber, he might say 'He made hairlines' instead of 'He made headlines.' Even rare examples of this type of contextual influence seldom point to underlying hostilities or neuroses.

Message Construction and Perspective

The basic goal of both sentence comprehension and sentence production is the linking of spoken forms to underlying conceptual meanings. In the 1970s, psycholinguists thought of these conceptual structures in terms of links between simple sentences or propositions. These propositions represented relations that had no clear relation to the actual physical perspective of the speaker. More recently, work in philosophy, robotics, psychology, cognitive linguistics, and cognitive neuroscience has recast this view of meaning structures in terms of dynamic activities that are linked to the embodied perspective of the speaker.

Consider what occurs when a speaker relates a fable, such as the story of the Tortoise and the Hare. In this story, there are basically three perspectives, one for the tortoise, one for the hare, and one for the narrator. The narrator says, 'the hare decided to rest.' At this point, the narrator is taking the perspective of the hare and describing the hare's actions from that perspective. Next, the narrator shifts to the perspective of the tortoise and describes how he passes the sleeping hare. These various shifts and blends of perspective are important aspects of all narrative structures. However, we can see their impact everywhere in language. For example, if we want to point out that there is a dent in the side of the refrigerator, we may use the impersonal construction, 'someone knocked a dent in the refrigerator' or we may use the pseudopassive, 'the refrigerator got a dent knocked into it.' In the impersonal we are forced to point a finger at an unspecified perpetrator of the deed. In the pseudopassive, we background or deemphasize the doer of the activity, focusing instead on the result. These choices all reflect variations in perspective.

Another choice we make is between alternative perspectives on a single action. For example, we could describe a picture by saying 'the girl gave a flower to the boy.' In this case, we are taking the viewpoint or perspective of the girl and describing the activity of giving from her perspective. Alternatively, we could say, 'the boy got a flower from the girl.' In this alternative form, we view the action from the perspective of the recipient. Depending on whether we choose to view the action from the viewpoint of the giver or the receiver, we will use either the verb *give* or the verb *get*. Language is full of choices of this type and sentence production can be viewed as involving a continual competition between choices for perspectives. The selection of pronouns is determined by the need to link up referents to earlier perspectives. Shifts from the present tense to the past also reflect movements in perspective to various points in time.

Crosslinguistic Comparisons

All languages have sound segments, syllables, words, phrases, and sentences. In all languages, processing is online, incremental, and interactive. However, languages differ markedly in the particular grammatical contrasts they choose to mark and

the devices that they use to mark them. In languages like Warlpiri or Hungarian, word order is remarkably flexible. In languages like Navajo or Eskimo, a single complex word may express what English does in a seven-word sentence.

Psycholinguistic comparisons of processing in different languages (MacWhinney and Bates, 1989) use a simple common task to examine the effects of radically different structures. In one type of experiment, subjects listen to a sentence such as 'The eraser is pushed by the cat' and have to point to a picture that shows who is the actor. Consider a comparison of these sentences from English and Spanish:

The lion kisses the cow.
El león besa la vaca. (The lion kisses the cow).

A major difference between these two languages revolves around the use of variable word orders. In Spanish, it is possible to invert the word order and produce 'la vaca besa el león' while still meaning that the lion is kissing the cow. This inversion is even clearer if the particle *a* is added to marker the direct object as in these variant orderings in which the two nouns are moved into different places around the verb *besa*.

El león besa a la vaca.
A la vaca besa el león.
Besa el león a la vaca.
Besa a la vaca el león.

These differences between English and Spanish can be traced to the relative cue validities of word order and object marking in the two languages. In English, it is virtually always the case that the noun that appears before the verb is the subject of the sentence. If the verb is an active verb, the preverbal noun is almost always the actor. This means that the cue of preverbal positioning is an extremely reliable guide in English to assignment of the actor role. In Spanish, there is no such simple rule. Instead, the best cue to assignment of the actor role is the presence of the object marker particle *a*. If one of the nouns in a two-noun sentence is marked by *a*, then we know that the other noun is the agent or subject.

Other languages have still other configurations of the basic cues to sentence interpretation. For example, Hungarian makes reliable use of a single case-marking suffix on the direct object. German uses the definite article to mark case and number. The Australian language 'Walpiri' marks the subject with a variety of complex case markings. Navajo places nouns in front of the verb in terms of their relative animacy in the Navajo 'Great Chain of Being' and then uses verbal prefixes to pick out the subject. These languages and others also often rely on number agreement between the verb and the noun as a cue to the subject. English also requires subject–verb agreement, but this cue is often missing or neutralized. In languages like Italian or Arabic, subject–verb agreement marking is extremely clear and reliable.

When a monolingual speaker comes to learn a second language, the learner needs to fundamentally retune his or her sentence-processing mechanism. First, the learners need to acquire a new set of grammatical devices and markings. Second, they need to associate these new devices with the correct cue validities. Third, they need to reorganize their expectations for particular sequences of cues and forms. Initially, the learner simply transfers the cues, cue validities, and habits from the first language to processing of the second language. Over a period of time, the cue validities change smoothly, until they eventually match that of the native monolingual (MacWhinney, 2012).

Neurolinguistics

Recent advances in Cognitive Neuroscience have had a major impact on psycholinguistic research. Increasingly, researchers are asking not just when and how linguistic processes interact, but also where and how processing occurs in the brain. The bulk of this work has used the neuroimaging methods of functional magnetic resonance imaging (fMRI) and event-related potential (ERP). Other methods include diffusion tensor imaging (DTI), structural MRI, lesion studies, near-infrared spectroscopy (NIRS), mismatch negativity (MMN), and single-cell recordings during surgery. Using these various methods, researchers have begun to piece together a new vision of language processing in the brain. This new account provides important roles for many cortical and subcortical regions. Frontal areas such as the dorsal lateral prefrontal cortex (DLPFC) and the anterior cingulate cortex (ACC) are seen as playing a role in the type of attentional switching needed for perspective shifting and code switching in bilinguals (Roelofs, 2011). Midbrain structures such as the striatum and basal ganglia are implicated in the development of language fluency (Ullman, 2004). A linked system of temporal and frontal areas is involved in both auditory and articulatory processing of sounds (Hickok, 2009). Studies show that the superior temporal sulcus (STS) controls many aspects of lexical processing, although other temporal areas also have some involvement. Increasing detail from DTI and other methods (Bookheimer, 2007) has shown that inferior frontal cortex or Broca's area has multiple areas for semantic, syntactic, and phonological processing, each with detailed connections through white matter tracts to other areas. The study of articulatory control in stutterers (Howell, 2011) points to a role for between-area connectivity in producing timing errors in phonological and lexical access.

The work attempting to identify areas involving aspects of language processing has shown how difficult it is to peg a function to a single area. Instead, it is becoming clear that language processing works in terms of dynamic interactions across multifunctional areas organized into flexible neural circuits (Anderson, 2010; MacWhinney, 2009). As we begin to understand the dynamic nature of these interactions, we will be able to construct a much more accurate understanding of how language is processed in the brain.

Summary

In this article, we have examined six core issues being addressed by psycholinguists. These areas are central to the study of psycholinguistics, because they allow us to evaluate the psychological reality of linguistic formalisms. The interaction between psycholinguistics and linguistic theory has been intense, reciprocal, and dynamic. We have learned that language processing is interactive and incremental in regard to

word recognition, sentence comprehension, sentence production, and message construction. We know that crosslinguistic differences in processing are closely related to differing levels of reliability for the cues involved and the ways in which structures impose demands on verbal memory capacity. Some versions of linguistic theory have moved to adapt to some of these findings from psycholinguistics (Pollard and Sag, 1994). However, a satisfactory resolution of the core issue of the psychological reality of linguistic structures has not yet been obtained.

See also: Functional Brain Imaging of Language Processes; Second Language Acquisition; Sentence Comprehension, Psychology of; Speech Errors, Psychology of; Speech Perception; Speech Production, Psychology of; Word Order; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Ambridge, B., Lieven, E., 2011. Child Language Acquisition: Contrasting Theoretical Approaches. Cambridge University Press, New York.

Anderson, M., 2010. Neural reuse: a fundamental organizational principle of the brain. Behavioral and Brain Sciences 33, 245–313.

Balota, D.A., 1994. Visual word recognition: the journey from features to meaning. In: Gernsbacher, M. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA.

Bookheimer, S., 2007. Pre-surgical language mapping with functional magnetic resonance imaging. Neuropsychological Review 17, 145–155.

Botvinick, M., Plaut, D.C., 2006. Short-term memory for serial order: a recurrent neural network model. Psychological Review 113, 201–233.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.

Dahan, D., Magnuson, J., Tanenhaus, M., Hogan, E., 2001. Subcategorical mismatches and the time course of lexical access: evidence of lexical competition. Language and Cognitive Processes 16, 507–534.

Fodor, J., 1983. The Modularity of Mind: An Essay on Faculty Psychology. MIT Press, Cambridge, MA.

Gazzaniga, M. (Ed.), 2000. The New Cognitive Neurosciences. MIT Press, Cambridge, MA.

Hickok, G., 2009. The functional neuroanatomy of language. Physics of Life Reviews 6, 121–143.

Howell, P., 2011. Language–speech processing in developmental fluency disorders. In: Guendozi, J., Loncke, F., Williams, M. (Eds.), The Handbook of Psycholinguistic and Cognitive Processes: Perspectives in Communication Disorders. Psychology Press, New York, pp. 435–462.

Lashley, K., 1951. The problem of serial order in behavior. In: Jeffress, L.A. (Ed.), Cerebral Mechanisms in Behavior. Wiley, New York.

Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Li, P., Zhao, X., MacWhinney, B., 2007. Dynamic self-organization and early lexical development in children. Cognitive Science 31, 581–612.

MacDonald, M.C., Pearlmutter, N.J., Seidenberg, M.S., 1994. Lexical nature of syntactic ambiguity resolution. Psychological Review 101 (4), 676–703.

MacWhinney, B., 1987. The competition model. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Lawrence Erlbaum, Hillsdale, NJ, pp. 249–308.

MacWhinney, B., 2009. The emergence of linguistic complexity. In: Givon, T., Shibatani, M. (Eds.), Linguistic Complexity. Benjamins, New York, pp. 405–432.

MacWhinney, B., 2012. The logic of the unified model. In: Gass, S., Mackey, A. (Eds.), The Routledge Handbook of Second Language Acquisition. Routledge, New York, pp. 211–227.

MacWhinney, B., Bates, E. (Eds.), 1989. The Crosslinguistic Study of Sentence Processing. Cambridge University Press, New York.

Marslen-Wilson, W., 1975. Sentence perception as an interactive parallel process. Science 189, 226–227.

McClelland, A., Mirman, D., Holt, L., 2006. Are there interactive processes in speech perception? Trends in Cognitive Sciences 10, 363–369.

McDonald, J.L., MacWhinney, B.J., 1995. The time course of anaphor resolution: effects of implicit verb causality and gender. Journal of Memory and Language 34, 543–566.

Pinker, S., 1999. Words and Rules: The Ingredients of Language. Basic Books, New York.

Pollard, C., Sag, I., 1994. Head-Driven Phrase Structure Grammar. Chicago University Press, Chicago, IL.

Roelofs, A., 2011. Modeling the attentional control of vocal utterances: from Wernicke to WEAVER+. In: Guendozi, J., Loncke, F., Williams, M. (Eds.), The Handbook of Psycholinguistic and Cognitive Processes: Perspectives in Communication Disorders. Psychology Press, New York, pp. 189–208.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.

Sapir, E., 1933/1968. The psychological reality of phonemes. In: Mandelbaum, D. (Ed.), Selected Writing of Edward Sapir. University of California Press, Berkeley, CA, pp. 46–60.

Simpson, G.B., 1994. Context and the processing of ambiguous words. In: Gernsbacher, M. (Ed.), Handbook of Psycholinguistics. Academic Press, San Diego, CA.

Stemmer, B., Whitaker, H. (Eds.), 2008. Handbook of the Neuroscience of Language. Academic Press, New York.

Trueswell, J., Gleitman, L., 2004. Children's eye movements during listening: evidence for a constraint-based theory of parsing and word learning. In: Henderson, J., Ferreira, F. (Eds.), Interface of Language, Vision, and Action: Eye Movements and the Visual World. Psychology Press, New York.

Ullman, M., 2004. Contributions of memory circuits to language: the declarative/procedural model. Cognition 92, 231–270.

Psychological Anthropology

Claudia Strauss, Pitzer College, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological anthropology is the anthropological study of people's beliefs, feelings, and motivations in cultural contexts. In the nineteenth and early twentieth centuries, questions about psychic unity and diversity were a central part of anthropology, not a separate subfield. Later, Freudian theory and the growth of the cognitive sciences inspired distinct schools within anthropology: one devoted to psychoanalytic studies of culture and personality and the other to formal models of cultural knowledge. Current research draws on variety of theories, such as relational psychology, activity theory, schema theory, neurobiology, and critical theory, creating possibilities for studies that integrate thought with feeling, mind with body, and psyche with society.

Introduction

Psychological anthropology is a subfield of anthropology devoted to the way cognition, emotion, and motivation are shaped in sociocultural settings and to the psychological factors that are important in culture learning and expression. Psychological anthropologists study topics such as narrative, identity, experience, emotion, memory, discourse, belief, conscious goals and unconscious desires, conceptual structures, gender, sexuality, trauma, mental illness, stigmatization, and psychological development in diverse social and cultural contexts.

Unlike the norm in related subfields of psychology (e.g., cultural psychology, social psychology, or cross-cultural psychology), methods in psychological anthropology are usually drawn from the anthropological tool kit, including participant observation and qualitative interviews. Formal methods or experiments may be used in addition. Psychological anthropologists are often critical of the approach sometimes seen in psychological research in which culture is considered a single variable.

Unlike other subfields of anthropology that are also concerned with cultural values, beliefs, and identities, psychological anthropologists typically employ 'person-centered' methods (LeVine, 1982), that is, ones that attend to individuals' thoughts, feelings, and motivations (Hollan, 2001).

Some major subfields of psychological anthropology include cognitive anthropology (*see* Cognitive Anthropology), psychoanalytic anthropology, psychiatric anthropology, cross-cultural studies of child development, ethnopsychology, and biocultural studies.

Intellectual Background

Investigations of personality and cognition in culturally diverse societies have been conducted by social and cultural anthropologists from at least the nineteenth century on, not as a distinctive subfield but as an outgrowth of anthropologists' central interest in human differences and universals. Early in the twentieth century, Franz Boas declared in a paper titled "Psychological problems in anthropology" that his concern was "'the influence of the society' on 'the habits of action and thought of the individual'" (quoted in Stocking, 1992: pp. 318–319). Boas described the two main branches of anthropology as 'psychological anthropology' and 'biological anthropology.' For Boas, 'psychological anthropology' included all aspects of the study of 'the mind of man' (Stocking, 1992: p. 312). Debates about human psychic unity were central to anthropological theories in the late nineteenth and early twentieth centuries regarding the origins of magical, scientific, and religious beliefs. For example, Lucien Lévy-Bruhl postulated that mystical 'pre-logical' (or a-logical) beliefs were characteristic of 'primitive' mentality (Lévy-Bruhl, 1910). By contrast, E.E. Evans-Pritchard stressed the logical reasoning that upheld Zande divination practices (Evans-Pritchard, 1976 [1937]). Such studies were considered an integral part of the investigation of culture, defined by E.B. Tylor as "that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society" (Tylor, 1958[1871]: p. 1).

Freudian theory provided the impetus for a separate subfield of anthropology devoted to psyche. Freud's concerns with the shaping of personality by methods of child rearing, his posited universal libidinal and aggressive drives, and his emphasis on subconscious psychic conflict that may be indirectly reflected in individual and collective symbolic expressions had a considerable impact throughout the discipline (e.g., on Claude Lévi-Strauss, Victor Turner, and Meyer Fortes) and inspired a particular approach called 'culture and personality' studies from the 1930s to the 1960s. Culture and personality studies focused on the personality traits that are predominant in a society due to its patterns of child rearing. The prominence of a few well-known culture and personality studies obscures diversity in their methods and theories. Some practitioners emphasized cultural differences in child rearing and consequent differences in typical personality. An example is Margaret Mead's early study of adolescence in Samoa, which she described as less psychologically fraught than Western adolescence because of the freer expression of libidinal instincts (Mead, 1949[1928]). Mead also described extreme variability in sex roles in three New Guinea societies, concluding that biological sex does not determine personality traits and that "human nature is almost unbelievably

malleable" (Mead, 1963: p. 280). Other culture and personality researchers began with assumptions about universals of human personality, such as the Oedipus complex. Some, such as Ruth Benedict in *Patterns of Culture* (1934) and *The Chrysanthemum and the Sword* (1946), characterized the typical personality structure of a society on the basis of their rituals, folklore, and other aspects of public culture; others administered projective tests and reported the individual variation they found (DuBois, 1961; Wallace, 1952), or made detailed observations of children's and caretakers' behaviors (Whiting and Whiting, 1975).

Starting in the 1950s, culture and personality studies came under criticism. Questions were raised about whether child rearing has lasting personality effects or, if it does, whether it has the particular effects postulated in Freudian theory. Criticisms were made of practitioners' emphasis on shared personality types at the expense of intracultural diversity, some researchers' circular method of inferring typical personalities from the social institutions they are supposed to cause, the possible lack of cross-cultural validity of Western projective tests (or their possible lack of validity in any society), ethnocentric and sometimes prejudiced descriptions of personality types, and lack of clarity about the meaning of 'culture' and 'personality' and their relation to each other (Bock, 1988; LeVine, 1982; Wallace, 1970[1961]). In the 1960s, the term 'psychological anthropology' came to the fore instead, particularly among American anthropologists. Unlike Boas's earlier inclusive meaning, 'psychological anthropology' now referred to a subfield of anthropology open to a variety of anthropological studies of psyches.

Beginning in the mid-1950s, a different set of anthropologists began using linguistic and cognitive theories and methods to investigate cultural systems of knowledge. As noted earlier, it has been common for anthropologists to describe cultural values and beliefs. What set apart the new endeavor was its sustained focus on what people think, its use of systematic person-centered methods instead of methods that infer beliefs from public culture, and the inspiration it drew from theories of cognition in the interdisciplinary cognitive sciences. This was the beginning of the major subdivision of psychological anthropology now termed 'cognitive anthropology.' Whether or not cognitive anthropologists agreed with Ward Goodenough's definition of culture as "whatever it is one has to know or believe in order to operate in a manner acceptable to its members" (Goodenough, 1957: p. 167), which reduces culture to shared knowledge (see criticism of that definition in D'Andrade, 1995), they considered such beliefs a central part of culture that could be described using some of the methods that were being developed in other fields.

Initial proposals were inspired by Prague School linguists' descriptions of the sets of distinctive features that differentiated phonemes (meaningful sounds in a language) and linguists' extension of that approach to word meanings. To give an anthropologically relevant example, kin terms in a language can be thought of as a contrast set in which items are differentiated by the presence or absence of a small number of features, such as male/female, consanguineal (blood)/collateral relation to the speaker, generation in relation to speaker, elder than/younger than the speaker (in some languages), patrilateral/matrilateral kin (in some languages), and possibly other features. Other cultural domains, such as terms for plants, animals, and foods, were modeled as taxonomies (*a*, *b*, and *c* are types of *d*; *d*, *e*, and *f* are types of *g*, and so on). Cognitive structures of this sort reveal something about the beliefs of a group at a particular time. Studies of the meanings of terms and relations between them in a cultural domain were sometimes called ethnoscience or ethnosemantics during the 1960s and the 1970s.

While logically satisfying, there were some problems with this approach to the study of people's cultural knowledge. It limited cultural knowledge to word meanings, and limited word meanings to tidy models that might not replicate how speakers actually think about meanings or use the words (Quinn, 2011). Contemporary cognitive anthropologists have a variety of approaches to the study of cultural knowledge that will be described in the next section.

Current Knowledge

At present, psychological anthropology includes a rich variety of methods, theories, and topics in the study of subjective experience in cultural contexts. Some psychological anthropologists obtain deep accounts of one person's life story; others observe child rearing; still others measure cortisol levels. Theoretical frameworks come from a variety of sources, including artificial intelligence, narrative psychology, relational psychoanalysis, feminist and critical theories, phenomenology, evolutionary theory, and the neurosciences. The following overview presents some major approaches but cannot describe all.

An important trend in the last 20 years is the emergence of theoretical frameworks that can integrate studies of cognition with studies of emotions, motivations, and personality dynamics. For heuristic purposes, cognitive and psychoanalytic personality studies will be discussed in separate sections, followed by theoretical efforts to bring them together.

Cognitive Anthropology

Cognitive anthropologists study shared and varying beliefs among members of a social group, information processing in particular social contexts, and other aspects of cognition such as memory or interpretive devices such as narrative structures. Some major schools of cognitive anthropology include cultural models, cultural consensus analysis, cognition in practice/distributed cognition, and cultural epidemiology.

Cultural Models

As the limitations of ethnosemantics were becoming more apparent, some anthropologists became interested in new ways of modeling word meanings and other types of cultural knowledge. They turned to the cognitive sciences where, from the mid-1970s on, linguists, psychologists, and others began formulating a way of modeling knowledge now called schemas (or schemata; other names are frames and scripts). The basic insight of schema theory is that explicit beliefs rest on a wealth of sometimes unspoken understandings. These understandings are not of disconnected propositions or features, but are

holistic models that provide context and interrelations. For example, instead of defining grandmother in terms of the features that distinguish this kin term from others (female second-generation ascending consanguineal), a schema analysis of the meaning of grandmother for many contemporary Americans might include a more psychologically realistic description of the kin relation (grandmother is the mother of one's mother or father), as well as expectations about how grandmothers typically behave (spoil their grandchildren), what they are supposed to look like (gray hair and unfashionable clothes), and so on. The idea of spoiling in turn rests on schemas about human nature as well as values about child rearing. All these implicit or explicit understandings are part of or connected to a grandmother schema. Each person has their own schemas; the ones that are learned and widely shared in a social group are cultural models or cultural schemas (Casson, 1981; Holland and Quinn, 1987; Quinn, 2011). Stereotypes are a type of schema because schemas represent simplified, generic knowledge, rather than knowledge of specific cases that may depart from the generic schema. Schemas are shaped not only by recurring patterns of one's first-hand experience but also by the repeated images and ideas propagated through media. Psychologists have found that schemas influenced memory and other aspects of information processing because cuing part of a schema tends to bring the rest of it to mind; thus, people may reconstruct memories to include the elements they expect on the basis of the schemas they hold.

There is no mechanical procedure for uncovering the elements of a cultural schema. Semistructured interviews about the topic in question are generally employed, with interviewees' comments then analyzed for repeated patterns in their explicit statements as well as the unstated assumptions that are revealed in the metaphors they use, their reasoning, and their trains of association (see essays in Holland and Quinn, 1987; Quinn, 2005). Cultural models research has been used to study conflicts between contemporary Chinese parents' explicitly stated values and their more complex implicit cultural models of appropriate behavior (Fong, 2007), Americans' ideas about marriage (Quinn, 1996), and Americans' conflicting schemas about social policy issues such as immigration and social welfare programs (Strauss, 2012).

Some cultural models researchers have proposed that connectionist models of cognition, inspired by the way neurons interact, can explain the ways in which schemas are built up through experience but allow creative responses to new circumstances. In older models of cognition, information is assumed to reside in the form of sentences in the people's minds. Adaptation to novel circumstances can be explained only as new sentences or as highly complex representations (if conditions *p*, *q*, *r*, *s*, and *t* apply, do this) that it is hard to imagine anyone learning. Suppose instead that learning consists of strengthening connections among different networks of neuronlike units that come to be more closely connected when they are activated together. Knowledge on this view is best thought of the weights that make it likely that a certain set of sensory inputs will activate other units and those still others, eventually leading to a particular response. Those weights can be modified by experience, and all the features of a situation will activate a particular set of units specific to that experience, creating the possibility of a new response. Connectionist models are particularly useful for modeling implicit associations and procedural knowledge, i.e., knowledge that is embodied in skills, habits, and associations that may be at odds with explicit beliefs (Strauss and Quinn, 1997).

Cultural Consensus/Cultural Domain Analysis

Consensus analysis was developed by Romney et al. (1986) as a replicable method for studying the extent of agreement in beliefs. The basic assumption is that in many cultural domains there are some agreed-upon propositions among members of a social group, although some people may have greater cultural expertise in that domain (football rules, disease symptoms, or celestial navigation) than others. If the researcher observes a response pattern in which there is convergence among some participants on a set of answers, and no competing agreed-upon answers, the consensual answers are considered cultural knowledge and those who gave the culturally correct responses are considered to have greater competence in that domain (assuming certain conditions hold, such as the respondents are not discussing their answers with each other while taking the test). Consensus analysis is generally conducted with true–false or multiple-choice questions that may be derived from qualitative interviews. This method can also be used to uncover competing understandings of the correct answer, as well as individual differences in knowledge. Like all standardized measures, it limits participants' opportunity to reveal their thinking beyond the questions asked, but the methods usefully reveal cultural and individual differences.

Cultural consensus analysis is one of many standardized approaches (e.g., free listing, pile sorts, and triad tests) used in cultural domain analysis, which is a continuation of earlier ethnosemantic studies of the interrelation of terms in a particular cultural domain (Weller and Romney, 1988). In free listing, participants are asked to simply list a number of terms in a domain (e.g., kinds of illnesses). The next step might be sorting those terms into piles of like items, which indicates taxonomic categories, or judging which out of three terms is the most different, which can reveal the features that are psychologically significant in making judgments of similarity and difference. For example, Moore et al. (1999) found shared bases for differentiating emotion terms among speakers of English, Chinese, and Japanese using these techniques. (See also the much-cited research on color terms of Berlin and Kay, 1991[1969].)

Cognition in Practice/Distributed Cognition

The theories of the Russian developmental psychologist Lev Vygotsky have led other anthropologists to stress the way thinking emerges from activities in particular settings. On this view, cognition is situated in particular social and material settings, with information processing distributed among different objects and social roles in a 'community of practice' (Lave and Wenger, 1991). From this perspective, culture may be defined not as what people know, but what people do. Thus, the primary research method employed is observation of ongoing activities (arithmetic while shopping, Lave, 1988; or navigation on board a ship, Hutchins, 1995), perhaps asking the person being observed to think out loud while they are in the midst of their activity. Nardi (1996) examines human–technology interaction, distinguishing the insights of

activity theory, situated action models, and distributed cognition. While some followers of this approach reject the significance of mental representations such as schemas, others see mental schemas as symbolic creations just like other cultural artifacts that mediate the effects of situational stimuli on human thoughts, feelings, and actions (Hazlehurst, 2011; Holland and Cole, 1995).

Cultural Epidemiology

Cultural epidemiologists are concerned with why some ideas and behaviors are widespread in a population and others are not. Some focus on features that make information easy to remember and in evolved human psychological processing biases, such as attributions of intentionality to others, taxonomic categorization of natural kinds, and the tendency to believe and emulate prestigious others (Boyd and Richerson, 1985; Henrich and Boyd, 1998; Sperber, 1996). Such studies may also take into account environmental factors (e.g., geographical barriers to the spread of knowledge), social ties (social networks, expert networks, and types of intergroup contact), and institutional structures and practices (e.g., authority structures, schooling, and mass media). Cultural epidemiologists have studied why religious beliefs are widespread (Atran, 2002), and the social distribution of ecological beliefs among different Mayan and Ladino groups in Guatemala (Atran et al., 2005), among other topics. The methods of cultural epidemiology draw upon a mix of cultural consensus analysis and distributed cognition approaches. There is overlap between this field and evolutionary anthropology (Barkow et al., 1992), which proposes ways in which minds have been shaped by evolutionary adaptation. These proposals are controversial among psychological anthropologists.

Psychodynamic Approaches and Studies of Childhood

Contemporary psychoanalytic anthropology retains basic Freudian concerns with psychic conflict, tension between conscious and unconscious beliefs and feelings, and the expression of the unconscious in individual fantasies and public culture. Contemporary psychoanalytic anthropologists usually draw upon neo-Freudian theories such as ego psychology, self-psychology, object relations theory, and relational psychoanalysis (Westen, 2001), or poststructuralist psychoanalytic theories such as those of Jacques Lacan. Psychoanalytic anthropologists have also been influenced by ethnographic observations outside of Europe and North America. The result is a greater attention than in classical Freudian theory to the impact of social and cultural contexts, formative psychological development not only in early childhood but later as well, mother–child relationships or interactions with multiple caretakers rather than Oedipal father–son relationships, and creative and adaptive psychological ego defenses more than destructive drives (LeVine, 2010; Paul, 1989). Some psychoanalytic anthropologists reject the assumption of universal libidinal and aggressive drives; others find support for them in newer evolutionary psychological theories of reproductive fitness. Unlike earlier culture and personality studies, the primary goal of contemporary psychoanalytic anthropologists is not to characterize typical personality structures in a society, although researchers may discuss recurrent types of psychic conflicts and ways of dealing with them. At present, psychoanalytic anthropologists typically rely on in-depth research that produces a wealth of interesting data.

Some psychoanalytic anthropologists analyze collective forms of expression, such as myths and rituals, seeing in them socially instituted expressions of and defenses against recurrent psychological concerns, e.g., about gender identity (Herdt, 1994 [1987]). Other psychoanalytic anthropologists are more interested in the personal uses of collective symbols (Spiro, 1987). Such researchers will focus on the lives of particular people to see how they turn collective symbols into personally meaningful ones (Obeyesekere, 1981). Several psychoanalytic anthropologists have studied dreams as revealing a nexus of psychic conflicts and cultural symbolism. For example, Waud Kracke (1981) explores a series of dreams reported by a Kagwahiv Indian man in Brazil that may be disguised reactions to the recent deaths of his two young children, losses that Kagwahiv are supposed to avoid dwelling upon. (See also Hollan, 2003; Mageo, 2011.) Others have suggested that possession trance may be a culturally sanctioned channel for indirect expression and gratification of wishes for those who lack power in other ways, such as women in many societies (Bourguignon, 2004).

Another influence of Freudian theory is continuing interest in child rearing. For example, Jean Briggs's *Inuit Morality Play* (1998) focuses on the social dilemmas that Inuit adults create for young children. Briggs proposes that their teasing questions (e.g., "Why don't you kill your baby brother?" "Your mother's going to die – look, she's cut her finger – do you want to come live with me?") teach children to be highly tuned to hidden social meanings. They also seem to deliberately cultivate ambivalent attachment feelings, which is an interesting contrast with the ideal of secure attachment in Western attachment theory. Studies of child rearing in psychological anthropology do not need to rest on psychodynamic theories, however, to investigate universals and cross-cultural differences in the ways children are raised from infancy onward (LeVine, 2007; *see* Childhood: Anthropological Aspects).

Current and Emerging Directions in Psychological Anthropology

Cognitive and psychoanalytic approaches may appear to be opposing, following the distinction between thought and feeling in Western folk categories, and for a long time they have been treated that way. One of the more interesting developments in recent decades has been efforts toward a more integrated approach to the study of culture and self.

For example, D'Andrade and other cognitive anthropologists have proposed that the cognitive representations of cultural models can be motivating if they represent widely shared goals (D'Andrade, 1992). Other cultural models theorists focus on the way personal experiences may or may not lend emotional and motivational force to cultural schemas for particular persons (Strauss, 1992; Westen, 2001). Connectionist models of schemas can readily incorporate feelings and desires that are associated with features of

objects or events (Strauss and Quinn, 1997). Freud's theory of dream interpretation rested on a similarly associationist model of mind that combined symbols, meanings, and emotions.

While connectionist models are neurally inspired but are models only, other anthropologists have begun to study the interactions of culture and human biology (biocultural studies), including neurobiology (neuroanthropology, Lende and Downey, 2012). For example, Worthman examines the way maternal stress affects in utero environments and their children's neuroendocrine flow, which regulates emotional reactions to traumatic events (Worthman, 2009). Seligman and Kirmayer (2008) discuss the neurobiological and physiological basis of dissociative states, while showing that the way such states are experienced depends on whether they are culturally normative. Quinn draws on neurobiological accounts of fragmentation and integration of self, expanding the notion of self beyond conscious self-representations (2006). Much of this research moves away from explicit beliefs to stress the way bodily habits, implicit mental associations, and emotions are developed through people's experiences in specific sociocultural settings. The emerging work in this area is not neurally reductive; instead, the goal is to think about the interrelationships of body, mind, and context.

From the other side, anthropologists whose primary interest is in emotions, mental illness, or child rearing have become interested in the shared schemas that shape the way emotions and mental illnesses are conceptualized or the folk theories of human nature or child development that affect practices of child rearing. Psychiatric anthropology is an interesting example of convergence among different theoretical approaches. Both mainstream Western psychiatry and psychiatric anthropology have moved away from Freudian approaches to mental illness, but they have gone in different directions. Contemporary psychiatrists usually stress biomedical definitions of mental illness, while contemporary psychiatric anthropologists try to broaden discussion of mental illness to encompass subjective personal experience and larger social and cultural contexts, including folk theories of psychology (ethnopsychology).

Interest in narrative is another area of convergence between psychoanalytic and cognitive anthropologists (Garro, 2001). Personal narratives rely upon shared cognitive schemas of typical story structures and life course sequences. Telling a story rehearses and fixes some memories; not telling elides other memories. At the same time, clinical practice is increasingly focused not so much on revealing true memories but on helping patients to construct more adaptive narratives relating themselves to others (Paul, 1989). Whether starting with more cognitive or more psychoanalytic concerns, anthropologists are often as interested in how people tell their life story as in the historical record of the facts relayed, unlike life story research in sociology or history. Ordering experience in emotion-laden narratives and other cultural schemas is part of a larger 'effort after meaning' that the mid-twentieth-century psychologist Frederick Bartlett said was part of all human cognition (Garro, 2001: p. 110).

Another important development in psychological anthropology is to place subjective experience in a larger social and political context. Psychological anthropologists are concerned with the subjective experience of unequal power relations, violence, stigma, suffering, and with the development of political consciousness and various kinds of group identities (see the essays in Casey and Edgerton, 2005; Mageo, 2002; Quinn and Strauss, 2006).

Methodological and Theoretical Issues

Tribal differences among psychological anthropologists are often based on preferred methods. For example, even though cultural models and cultural consensus researchers are both concerned with the content of cultural beliefs, their methods differ. Cultural models research relies on qualitative analysis of lengthy semistructured interviews; cultural consensus analysis draws upon quantitative analysis of responses to standardized questions. Cultural models researchers believe their approach to be more revelatory of implicit beliefs; cultural consensus researchers stress the replicability of their methods. In practice, however, cultural consensus analyses usually begin with the more open-ended approaches that are typical of cultural models research, and members of both camps would agree on the importance of person-centered empirical research as opposed to, say, methods that draw inferences about people's beliefs from collective expressions such as rituals or movies.

Psychological anthropologists have differed in the past over the extent to which there are universals of human psychology. Some declare that mainstream academic psychology is only a Western folk theory. While most would not endorse such a sweeping dismissal, psychological anthropologists over the years have questioned the cross-cultural applicability of particular psychological theories, from Freud's Oedipus complex and Piaget's presumed universals of cognitive development to Bowlby's theories of attachment (see LeVine and Norman, 2001; Quinn and Mageo, 2013) at present. Criticism of particular theories does not imply rejection of all psychological theories. If people are acquiring different folk theories of psychology, then culture learning is occurring, which can only be explained by reference to general psychological theories. Cross-cultural comparative research has produced important findings about the interacting effects of universal tendencies and variable experiences in realms ranging from the effects of father's absence on boys' sex-typed behavior to emotion concepts (Munroe and Munroe, 2001).

Although there is no unanimity about what psychological universals there may be, universalism versus particularism is not a central debate at present. Instead of either–or debates, current research draws eclectically on a variety of approaches that reveal the complexity of the ways brains and bodies are shaped in particular settings to create people's interpretations, experiences, and practices.

Electronic Resources

The Society for Psychological Anthropology (SPA), a section of the American Anthropological Association, maintains a Web site with links to syllabi, graduate programs, prizes, funding, and conferences sponsored by the society. The SPA listserv creates opportunities for distant researchers to connect (http://www.aaanet.org/sections/SPA/).

The Neuroanthropology blog is frequently updated with articles and links of broad interest to anthropologists interested in the mind (http://blogs.plos.org/neuroanthropology/).

The Foundation for Applied Psychiatric Anthropology provides links to information about films, books, articles, conferences, organizations, and other resources (http://www.psychanthro.org/FAPA_Home.html).

See also: Affect and Emotion, Anthropology of; Belief, Anthropology of; Body: Anthropological Aspects; Childhood: Anthropological Aspects; Cognition, Evolution of; Cognitive Anthropology; Culture, Cognition and Embodiment; Education: Anthropological Aspects; Human Cognition, Evolution of; National Character; Neo-Vygotskian Developmental Theory; Personhood, Anthropology of; Religion, Anthropology of; Senses, Anthropology of The; Social Cognition; Values, Anthropology of; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development; Worldview.

Bibliography

Atran, S., 2002. In Gods We Trust: The Evolutionary Landscape of Religion. University Press, New York, Oxford.

Atran, S., Medin, D.L., Ross, N.O., 2005. The cultural mind: environmental decision making and cultural modeling within and across populations. Psychological Review 112 (4), 744–776.

Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), 1992. The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford, UK.

Berlin, B., Kay, P., 1991[1969]. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA.

Bock, P., 1988. Rethinking Psychological Anthropology: Continuity and Change in the Study of Human Action. W.H. Freeman, New York.

Bourguignon, E., 2004. Suffering and healing, subordination and power: women and possession trance. Ethos 32 (4), 557–574.

Boyd, R., Richerson, P., 1985. Culture and Evolutionary Process. University of Chicago Press, Chicago.

Briggs, J., 1998. Inuit Morality Play: The Emotional Education of a Three-Year-Old. Yale University Press, New Haven.

Casey, C., Edgerton, R.B. (Eds.), 2005. A Companion to Psychological Anthropology. Blackwell, Malden, MA.

Casson, R., 1981. Language, culture, and cognition. In: Casson, R. (Ed.), Language, Culture, and Cognition. Macmillan, New York, pp. 11–22.

D'Andrade, R.G., 1992. Schemas and motivation. In: D'Andrade, R.G., Strauss, C. (Eds.), Human Motives and Cultural Models. Cambridge University Press, Cambridge, UK, pp. 23–44.

D'Andrade, R.G., 1995. The Development of Cognitive Anthropology. Cambridge University Press, Cambridge, UK.

DuBois, C., 1961[1944]. The People of Alor, vols. 2. Harper and Row, New York.

Evans-Pritchard, E.E., 1976[1937]. Witchcraft, Oracles and Magic Among the Azande. Oxford University Press, Oxford.

Fong, V., 2007. Parent-child communication problems and the perceived inadequacies of Chinese only children. Ethos 35, 85–127.

Garro, L., 2001. The remembered past in a culturally meaningful life: remembering as cultural, social, and cognitive process. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, Cambridge, UK, pp. 105–147.

Goodenough, W., 1957. Cultural anthropology and linguistics. In: Garvin, P. (Ed.), Report of the Seventh Annual Round Table Meeting in Linguistics and Language Study, Monograph Series on Language and Linguistics, No. 9. Georgetown University, Washington, DC, pp. 167–173.

Hazlehurt, B., 2011. The distributed cognition model of mind. In: Kronenfeld, D.B., Bennardo, G., de Munck, V.C., Fischer, M.D. (Eds.), A Companion to Cognitive Anthropology. Wiley-Blackwell, Malden, MA, pp. 471–488.

Henrich, J., Boyd, R., 1998. The evolution of conformist transmission and the emergence of between-group differences. Evolution and Human Behavior 19, 215–241.

Herdt, G., 1994[1987]. Guardians of the Flutes: Idioms of Masculinity. Edition with New Preface. University of Chicago Press, Chicago.

Hollan, D., 2001. Developments in person-centered ethnography. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, Cambridge, UK, pp. 48–67.

Hollan, D., 2003. Selfscape dreams. In: Mageo, J. (Ed.), Dreaming and the Self: New Perspectives on Subjectivity, Identity, and Emotion. State University of New York Press, Albany, NY, pp. 61–74.

Holland, D., Cole, M., 1995. Between discourse and schema: reformulating a cultural-historical approach to culture and mind. Anthropology and Education Quarterly 26 (4), 475–490.

Holland, D., Quinn, N. (Eds.), 1987. Cultural Models in Language and Thought. Cambridge University Press, Cambridge, UK.

Hutchins, E., 1995. Cognition in the Wild. MIT Press, Cambridge, MA.

Kracke, W., 1981. Kagwahiv mourning: dreams of a bereaved father. Ethos 9 (4), 258–275.

Lave, J., 1988. Cognition in Practice: Mind, Mathematics and Culture in Everyday Life. Cambridge University Press, Cambridge, UK.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, UK.

Lende, D.H., Downey, G. (Eds.), 2012. The Encultured Brain: An Introduction to Neuroanthropology. MIT Press, Cambridge, MA.

LeVine, R.A., 1982. Culture, Behavior, and Personality, second ed. Aldine, Chicago.

LeVine, R.A., 2007. Ethnographic studies of childhood: a historical overview. American Anthropologist 109 (2), 247–260.

LeVine, R.A., 2010. Introduction—after Freud: dramas of the psyche in cultural context. In: LeVine, R.A. (Ed.), Psychological Anthropology: A Reader on Self in Culture. Wiley-Blackwell, Malden, MA, pp. 119–123.

LeVine, R.A., Norman, K., 2001. The infant's acquisition of culture: early attachment reexamined in anthropological perspective. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, Cambridge, UK, pp. 83–104.

Lévy-Bruhl, L., 1926[1910]. How Natives Think. Knopf, New York.

Mageo, J. (Ed.), 2002. Power and the Self. Cambridge University Press, Cambridge, UK.

Mageo, J., 2011. Dreaming Culture: Meanings, Models, and Power in U.S. American Dreams. Palgrave Macmillan, New York.

Mead, M., 1949[1928]. Coming of Age in Samoa. Mentor, New York.

Mead, M., 1963[1935]. Sex and Temperament in Three Primitive Societies. Apollo, New York.

Moore, C.C., Romney, A.K., Hsia, T., Rusch, C.D., 1999. The universality of the semantic structure of emotion terms: methods for the study of inter- and intra-cultural variability. American Anthropologist 101, 529–546.

Munroe, R.L., Munroe, R.M., 2001. Comparative approaches to psychological anthropology. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, Cambridge, UK, pp. 223–237.

Nardi, B.A., 1996. Studying context: a comparison of activity theory, situated action models, and distributed cognition. In: Nardi, B. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction. MIT Press, Cambridge, MA., pp. 69–102.

Obeyesekere, G., 1981. Medusa's Hair: An Essay on Personal Symbols and Religious Experience. University of Chicago Press, Chicago.

Paul, 1989. Psychoanalytic anthropology. Annual Review of Anthropology 18, 177–202.

Quinn, N., 1996. Culture and contradiction: the case of Americans reasoning about marriage. Ethos 24, 391–425.

Quinn, N. (Ed.), 2005. Finding Culture in Talk: A Collection of Methods. Palgrave Macmillan, New York.

Quinn, N., 2006. The self. Anthropological Theory 6 (3), 362–384.

Quinn, N., 2011. The history of the cultural models school reconsidered: a paradigm shift in cognitive anthropology. In: Kronenfeld, D.B., Bennardo, G., de Munck, V.C., Fischer, M.D. (Eds.), A Companion to Cognitive Anthropology. Wiley-Blackwell, Malden, MA, pp. 30–46.

Quinn, N., Mageo, J. (Eds.), 2013. Reconsidering Attachment: Cultural Perspectives on a Western Theory. Palgrave Macmillan, New York.

Quinn, N., Strauss, C. (Eds.), 2006. The missing psychology in cultural anthropology's key words. Anthropological Theory 6 (3), 267–279 (special issue).

Romney, A.K., Weller, S., Batchelder, W., 1986. Culture as consensus: a theory of culture and informant accuracy. American Anthropologist 88, 313–338.

Seligman, R., Kirmayer, L.J., 2008. Dissociative experience and cultural neuroscience: narrative, metaphor and mechanism. Culture, Medicine and Psychiatry 32, 31–64.

Sperber, 1996. Explaining Culture: A Naturalistic Approach. Blackwell, Oxford.

Spiro, M.E., 1987. Collective representations and mental representations in religious symbol systems. In: Kilborne, B., Langness, L.L. (Eds.), Culture and Human Nature: Theoretical Papers of Melford E. Spiro. University of Chicago Press, Chicago, pp. 161–184.

Stocking, G., 1992. Polarity and plurality: Franz Boas as psychological anthropologist. In: Schwartz, T., White, G.M., Lutz, C.A. (Eds.), New Directions in Psychological Anthropology. Cambridge University Press, Cambridge, UK, pp. 311–323.

Strauss, C., 1992. What makes Tony run? Schemas as motives reconsidered. In: D'Andrade, R., Strauss, C. (Eds.), Human Motives and Cultural Models. Cambridge University Press, Cambridge, UK, pp. 197–224.

Strauss, C., 2012. Making Sense of Public Opinion: American Discourses About Immigration and Social Programs. Cambridge University Press, New York.

Strauss, C., Quinn, N., 1997. A Cognitive Theory of Cultural Meaning. Cambridge University Press, Cambridge, UK.

Tylor, E.B., 1958[1871]. Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art and Custom. J. Murray, London.

Wallace, A.F.C., 1952. The Modal Personality Structure of the Tuscarora Indians, As Revealed by the Rorschach Test. Bureau of American Ethnology, Bulletin 150, Washington, DC.

Wallace, A.F.C., 1970[1961]. Culture and Personality, second ed. Random House, New York.

Weller, S., Romney, A.K., 1988. Systematic Data Collection. Sage, Newbury Park, CA.

Westen, D., 2001. Beyond the binary opposition in psychological anthropology: integrating contemporary psychoanalysis and cognitive science. In: Moore, C.C., Mathews, H.F. (Eds.), The Psychology of Cultural Experience. Cambridge University Press, Cambridge, UK, pp. 21–47.

Whiting, B.B., Whiting, J.W.M., 1975. Children of Six Cultures—A Psychocultural Analysis. Harvard University Press, Cambridge, MA.

Worthman, C., 2009. Habits of the heart: life history and the developmental neuroendocrinology of emotion. American Journal of Human Biology 21 (6), 772–778.

Psychological Contracts

Simon LD Restubog, Research School of Management, College of Business and Economics, The Australian National University, Canberra, ACT, Australia
Kohyar Kiazad, Faculty of Business and Economics, Monash University, Caulfield, VIC, Australia
Christian Kiewitz, School of Business Administration, University of Dayton, Dayton, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research in psychological contracts at work continues to expand and advance. In this article, we summarize and integrate this scholarship, addressing definitional issues, theoretical perspectives, plus antecedents and consequences of psychological contract breach (PCB). We identify future research opportunities that can enrich our understanding of PCB and its consequences for employees' work attitudes and behaviors.

Over the last two decades, the concept of psychological contract (Rousseau, 1990) has received much theoretical and empirical attention in organizational behavior and industrial/organizational psychology (e.g., Zhao et al., 2007). Spawning much of this interest has been the changing nature of the employee–organization relationship (Rousseau, 1995). Early writings on the psychological contract trace back to Argyris (1960) and Levinson et al. (1962) who loosely referred to psychological contract to describe the subjective aspects of employment relationships, referring to the unspoken agreement between supervisors and employees or the unconscious psychological needs (i.e., affection, dependence, aggression, growth, and development) that affect and define expectations forming within the employee–organization exchange relationship. Although the majority of scholarship emphasized needs and expectations as the primary concepts characterizing psychological contracts, Rousseau (1989) challenged these long-standing assumptions by redefining the psychological contract, thus reinvigorating scholarly interest in the construct.

Accordingly, psychological contracts capture "individual beliefs, shaped by the organization, regarding terms of an exchange agreement between individuals and their organization" (Rousseau, 1990: p. 9). Departing from conventional wisdom emphasizing the role of psychological needs, Rousseau (1990) argued that beliefs regarding the terms of the psychological contract are grounded in the explicit (e.g., verbal and written agreements) or implicit (e.g., repeated patterns of exchange) promises employees perceive their organization to have made (e.g., job security, bonuses) in exchange for their contributions (e.g., commitment and performance). Hence and importantly, employees' psychological contracts consist only of those obligations that are grounded in perceived promises. Besides its promissory nature – which distinguishes the psychological contract from more generic employment expectations (e.g., Robinson, 1996) – Rousseau (1989, 1990) described three additional characteristics of psychological contracts. First, they are perceptual and idiosyncratic, and hence held and understood by the employee alone: that is, mutual understanding of or agreement between employee and employer regarding the contract's content is not a necessary condition (Rousseau, 1995). In this way, psychological contracts differ from formal or implied contracts, which consist of commonly understood or shared beliefs (Morrison and Robinson, 1997).

Second, Rousseau (1990, 2000) identified a typology (transactional, relational, and balanced) of psychological contracts along dimensions of time frame (short-term versus open-ended), degree of specificity (highly versus loosely specified), resources exchanged (tangible versus intangible), and performance–reward contingency (highly contingent versus low or noncontingent). Accordingly, transactional contracts have a short-term and narrow focus, where tangible resources are exchanged for specific times and tasks (e.g., an employee is promised commission-based pay contingent upon meeting performance objectives); relational contracts involve highly subjective, open-ended exchanges that are based on trust between employee and organization (e.g., employee is promised developmental opportunities in exchange for extrarole services); while balanced contracts blend features of transactional and relational arrangements and allow for greater flexibility and changing contract requirements (e.g., employee is promised advancement opportunities in exchange for escalating performance requirements). Psychological contract scholarship has largely emphasized the transactional-relational distinction (Zhao et al., 2007).

Third, Rousseau (1989) speculated that psychological contract violation – or the "failure of organizations or other parties to respond to an employee's contribution in ways the individual believes they are obligated to do so" (p. 128) – constitutes the main process through which psychological contracts impact employees' attitudes and behaviors. This perspective differed from earlier views, which tended to focus on the discrepancy between employee and employer expectations (e.g., Schein, 1980), or thwarted needs (Levinson et al., 1962), regardless of what was promised. Rousseau's (1990) reformulation, has dominated much of the scholarly writing on the effects of psychological contracts on employees' work-related attitudes and behaviors (Zhao et al., 2007). Indeed, there is a wealth of empirical evidence to confirm that psychological contracts matter because failing its (even implicit!) terms (now commonly referred to as psychological contract breach (PCB); Morrison and Robinson, 1997) can result in significant negative consequences for employees and organizations (Zhao et al., 2007).

Overall, Rousseau's (1989, 1990) reformulation advanced greater precision for understanding conceptual and theoretical underpinnings of the psychological contract. More importantly, by introducing and highlighting process mechanisms such as violation (or PCB), her work triggered an explosion of empirical studies focused on the antecedents and consequences of PCB. Building on these insights, scholars have also explored consequences of psychological contract fulfillment (Turnley et al., 2003), including overfulfillment (i.e., delivering more than what was promised; Lambert et al., 2003). Generally speaking, fulfillment of psychological contract obligations elicits positive employee responses, which, however, begin to diminish as delivered inducements exceed promised inducements (overfulfillment). While contract fulfillment is an interesting area and our literature review aims to be broad, we pay particular attention to the state of literature addressing PCB because this is and has been the focus of the theoretical and empirical scholarship (e.g., Zhao et al., 2007).

Theoretical Perspectives

Several theoretical perspectives have been used to explain why PCB results in negative consequences. The underlying theoretical frameworks in psychological contract research include equity theory, social exchange theory, and the group value model.

Equity Theory

Equity theory proposes that individuals estimate the ratio of what has been contributed (i.e., inputs) to what has been received (i.e., outcomes) for both themselves and a chosen referent other (Adams, 1965). Inputs are the employee's contributions (e.g., education, performance, work experience), while outcomes constitute rewards that employees receive in exchange for their contributions (e.g., compensation, promotions, career development opportunities). If an imbalance between inputs and outcomes occurs, individuals are likely to reduce the discrepancy by adjusting their attitudes, contributions, or outcomes to restore the balance between the two ratios.

Social Exchange Theory

Scholars have utilized this theory to conceptualize social exchange in the workplace as a series of mutual, obligation-based exchanges between interdependent parties (Blau, 1964). Exchanges at work are often governed by reciprocity norms (Gouldner, 1960) such that when individuals receive something from another individual, group, or organization, they feel compelled to reciprocate, typically in kind (i.e., positively or negatively). Moreover, the theory differentiates between two types of exchanges (Blau, 1964), with social exchanges involving unspecified obligations and economic exchanges encompassing specified obligations. Notably, social exchanges require trust between parties whereas economic exchanges do not. In the context of psychological contracts, when employees believe that the organization has not adequately fulfilled its obligations as promised, they may become motivated to reciprocate this unfavorable treatment. Such negative reciprocity, may take the form of reduced loyalty or withdrawal of extrarole services, or in some cases, intentional efforts to harm the organization and its members (e.g., Restubog et al., in press; Zhao et al., 2007).

Group Value Model

Inspired by social identity theory (Tajfel and Turner, 1979), the group value model posits that the manner in which employees are treated by organizational authorities communicates consequential identity-relevant information. Because authorities act as prototypical representatives of groups, employees perceive them to express group norms and being salient indicators of what the broad group membership thinks. Fair and respectful treatment by authorities suggests that the employee is a valued group member, whereas unfair treatment signals occupying a marginal position within the group/organization. Importantly, the fairness of decision-making procedures also sends symbolic messages about whether members can take pride in their group or organizational membership(s). If authorities publicly display fair procedures, it hence helps foster pride and identification among members within that group. These processes are significant because individuals become attached to their organizations by integrating perceived attributes of the organization into their own self-concept (Ashforth and Mael, 1989). In contrast, authorities behaving unfairly signals that group members are unworthy of respect or pride. The group value model is particularly useful in explaining responses to PCB when symbolic issues are involved. Responses to PCB are likely to be more pronounced when the nature of the PCB is relational rather than transactional, as the former emphasizes long-term commitment and socioemotional elements in the employment relationship.

Theoretical Distinctions between PCB and Related Constructs

PCB captures "the employee's perception regarding the extent to which the organization has failed to fulfill its promises or obligations" (Zhao et al., 2007: p. 649). Paralleling the idiosyncratic nature of psychological contracts, employees might perceive PCB, regardless of "whether a promise was actually broken, or even whether an obligation was ever established" (Robinson and Morrison, 2000: p. 526). In other words, PCB need not be real to be perceived, such as when employees misinterpret information received during on-the-job socialization (e.g., "people tend to get promoted rapidly here"; Robinson and Morrison, 2000: p. 526). Earlier, confusion existed regarding the distinction between PCB and unmet job expectations, with the latter referring to employees' job experiences in comparison to their prehire expectations (Wanous et al., 1992). Indeed, some scholars even speculated that PCB would not contribute anything beyond the effects of unmet expectations (Guest, 1998). However, the experience of PCB has been shown to go beyond unmet expectations, because PCB denotes that promises have been broken, and thus norms governing reciprocal exchange (i.e., good faith, fair dealing, and trust) have been dishonored (Morrison and Robinson, 1997; Robinson, 1996). In a longitudinal study of MBA graduates, for example, Robinson (1996) found that PCB explained variance

beyond that accounted for by unmet expectations in predicting organizational citizenship behavior (i.e., employee behaviors that are not formally sanctioned but greatly contribute to organizational functioning; Organ, 1988). Similarly, Zhao et al. (2007) concluded that, relative to unmet expectations, PCB elicits stronger effects on a range of outcomes, including job satisfaction, turnover intentions, and job performance.

PCB also differs conceptually and empirically from the experience of psychological contract violation. In the past, researchers used the terms breach and violation interchangeably until Morrison and Robinson (1997) introduced the now commonly accepted distinction. Accordingly, PCB represents the cognitive awareness that the organization has broken promises, whereas violation represents the affective state (e.g., anger, frustration, bitterness) that may or may not accompany this awareness.

Antecedents and Consequences of PCB

To date, psychological contract scholars have overwhelmingly studied the consequences of PCB relative to its antecedents. Yet the few existent exceptions have identified a range of antecedents (mostly exchange-based), including supervisors' experience of PCB (Bordia et al., 2008), leader–member exchange (LMX; Henderson et al., 2008), supervisor–subordinates (dis) similarity (Suazo et al., 2005), lack of trust (Robinson, 1996), informal socialization (Robinson and Morrison, 2000), history of PCB (Lo and Aryee, 2003), organizational and employee performance (Robinson and Morrison, 2000), and personality (Raja et al., 2004). In contrast, there is a solid body of work examining employees' responses to PCB. This is particularly relevant as it illustrates how social exchange theory, equity theory, and the group value model can be used to explain employee outcomes. To integrate and succinctly present the large volume of literature on the consequences of PCB, we organize our review around the following outcome categories: attitudes, affective reactions, and work behaviors.

In terms of employee attitudes, PCB is negatively related to job satisfaction (Bal et al., 2013), trust (Restubog et al., 2008; Robinson, 1996), commitment (Restubog et al., 2006), and identification, while being positively associated with turnover intentions (Dulac et al., 2008), revenge cognitions (Restubog et al., in press), and cynicism (Johnson and O'Leary-Kelly, 2003). With respect to affective reactions, PCB is positively related to feelings of violation (Restubog et al., in press), negative affect (Lo and Aryee, 2003), and emotional exhaustion (Johnson and O'Leary-Kelly, 2003) but negatively to enthusiasm (Conway and Briner, 2002). Finally, in terms of behavioral reactions, PCB generally leads to more workplace deviance (Restubog et al., in press), reduced in-role (Zhao et al., 2007) and extrarole performance, including citizenship behavior (Zhao et al., 2007), proactive behavior (Bal et al., 2011), and innovation (Ng et al., 2010).

Moderators of PCB–Outcomes Relationships

Researchers have also made considerable advances identifying boundary conditions that impact employees' perceptions of and reactions to psychological contract-related issues, with the majority of studies having investigated moderators that either amplify or ameliorate the effects of PCB and that fall into one of two categories: individual differences or situational influences.

Person Factors as Moderators

Early psychological contract research investigated the potential of attitudes – such as trust – and personality characteristics – such as equity sensitivity, attribution style, and agreeableness – to moderate relationships between PCB and work-related outcomes. For instance, a longitudinal study by Robinson (1996) found that initial trust in one's employer at the time of hiring moderated the PCB-subsequent trust relationship such that low initial trust led to a stronger decline in subsequent trust after perceived PCB compared to high initial trust. Another moderator is equity sensitivity that captures individuals' levels of sensitivity toward unfair situations, with entitled individuals emphasizing personal outcomes and exhibiting a low tolerance for being underrewarded in contrast to benevolents' higher tolerance, which stems from their concern for the employer–employee relationship. Unsurprisingly, research has found entitled individuals to generally respond more negatively to PCB than their benevolent counterparts. In the context of specific PCB facets, for example, those with a more entitled orientation reacted more adversely to breaches impacting tangible extrinsic outcomes such as pay or benefits and to breaches impacting growth and development opportunities; yet benevolent employees exhibited more negative responses when the employer broke promises about autonomy and control (Kickul and Lester, 2001). Concerning global effects of PCB, entitled as opposed to benevolent individuals engaged in less organizational citizenship behaviors and in more interpersonal workplace deviance following PCB (Restubog et al., 2007). Notably, Coyle-Shapiro and Neuman (2004) have attributed null findings involving equity sensitivity to the construct being overly broad, instead advocating the usage of exchange and creditor ideologies because they capture responses to social exchange issues more proximally. Accordingly, these ideologies not only impact individuals' orientation toward exchange relationships with employers in general but also perceived employer obligations and the fulfillment of those obligations in particular. In support, their study showed that individuals subscribing to an exchange ideology closely monitor what their employer is obligated to provide and make their contributions contingent upon the employer's potential inducements, whereas individuals endorsing a creditor ideology over match employer-provided inducements because they believe in repaying more than they receive.

More recent psychological contract research has extended the above list of variables by examining the moderating role of individuals' age and future time perspective (FTP). In a longitudinal study, Bal et al. (2013) reported weaker associations of PCB with job performance – partially mediated by job satisfaction – for older workers, irrespective of PCB type (i.e., economic, socioemotional, or developmental). A related, interesting moderator is FTP, which reflects individuals' belief about how much time they have left in the future (Bal et al., 2010). Accordingly, high FTP individuals have more

active coping styles and focus on the future, including future rewards, learning, and long-term exchanges, while low FTPers exhibit more passive coping styles and focus more on personal than organizational relationships. It follows that organizational inducements in the form of economic and social emotional fulfillment should have a stronger potential for high FTP employees to increase motivation and induce stronger felt obligations. In support, a study by Bal et al. (2010) showed that high FTP moderated positively the relationships between economic and socioemotional psychological contract fulfillment and employee obligations, whereas low FTP yielded nonsignificant or even negative effects in a sample of postretirement employees (i.e., those who continue working past retirement age).

In deliberating the wider implications of the above study results, it is instructive to consider works that have employed emotion regulation as well as the Big Five personality dimensions, as moderators of PCB. Framing PCB as an affective event, Bal et al. (2011) found emotion regulatory strategies to moderate the negative impact of PCB, with cognitive change buffering detrimental PCB effects on feelings of psychological contract violation and on the proactive employee behaviors of sharing knowledge and taking charge at work; yet attentional deployment forms of emotion regulation accentuated the negative effects of both social and generative forms of PCB on those proactive behaviors. Regarding conscientiousness, Orvis et al.'s (2008) longitudinal study showed, as expected, that low-conscientious employees answer to PCB with decreased organizational loyalty and job satisfaction plus with increased turnover intentions; yet unexpectedly, they also found that it is those high in conscientiousness who subsequently reduce their levels of task performance. Finally, Ho et al. (2004) found that agreeableness interacted with PCB severity and type such that its association with trust in the other party weakened only for severe cases of social breach, but not for economic ones.

Situational Factors as Moderators

Psychological contract scholars acknowledge employees' interpretations of the situational context in their studies (e.g., Morrison and Robinson, 1997). Examining the role of organizational justice, for example, Kickul (2001) found a three-way interaction in predicting workplace deviance such that PCB resulted in more deviant behavior when both interactional and procedural justice surrounding the PCB were low. Similarly, Turnley and Feldman (1999) found that PCB was more likely to trigger job search behaviors when employees perceived low procedural justice, external job opportunities, and the PCB as not sufficiently justified. These proximal factors are augmented by research that has investigated more distal contextual/cultural influences. For example, a study with Filipino nationals demonstrated that PCB elicits more negative reactions among those high in workplace familism (i.e., those who hold strong expectations of being cared for and nurtured by organizational agents, especially their supervisors; Restubog and Bordia, 2006). In a similar vein, the notion developed above – that variables cannot be assumed to function with PCB in a uniform fashion simply based on their inherent qualities – also applies to leader–member relationships. For instance, Restubog et al. (2010) contrasted the traditional social support perspective of LMX, which suggests that LMX minimizes PCB's adverse impact on work performance, with a betrayal perspective positing high LMX to aggravate negative effects, and found support for the latter.

Future Research Opportunities

We propose two research directions to broaden and enrich psychological contract scholarship. The first opportunity relates to theoretical framing of psychological contract development and the effects of PCB by integrating Conservation of Resources (COR) theory (Hobfoll, 2001). COR theory submits that individuals (1) are motivated to create, maintain, and increase their stock of valued resources and (2) experience stress when resources are lost or threatened. Given that psychological contracts are fundamentally based on exchange of valued resources (Bal et al., 2010), subsequent inquiries might adopt a resource-based perspective on psychological contracts to identify a broader range of resources exchanged beyond those captured by transactional and relational dimensions. Regarding theoretical framing of the effects of PCB, most previous inquiries have invoked the negative reciprocity principle (Gouldner, 1960) of social exchange theory, which implies that the only possible employee reaction to PCB is a withdrawal of effort or contribution. Yet employees can sometimes react proactively and constructively to negative states within the workplace (Zhou and George, 2001), which is not easily explained by the reciprocity principle. However, COR theory – with its focus on motivation to obtain, retain, foster, and protect valued resources – can explain proactive employee efforts to recover from PCB. Specifically, Hobfoll (2001) delineated several propositions related to resource loss, plus the conditions under which individuals are likely to pursue proactive coping in response to resource loss. In general, the theory posits that individuals well-endowed with resources (e.g., social support, self-efficacy, time) are more likely to respond to resource loss with proactive efforts to acquire additional resources to offset or compensate the loss. In the context of psychological contracts, it is possible to conceptualize PCB as an experience of resource loss (e.g., Restubog et al., 2013), as the core definition of PCB implies that valued resources (e.g., training, development, pay) were not provided. By offering a theory-based explanation for why and under what circumstances employees might respond to PCB in a proactive and constructive way, COR theory can hence broaden our understanding of PCB effects, along with potential moderators that condition PCB effects.

A second future research opportunity relates to advancing normative psychological contracts. Most research in this area has treated psychological contracts as individually held beliefs, even though there are theoretical and empirical grounds to suggest that individuals working in teams develop shared beliefs toward their job and working environment (Rousseau, 1978). Employment experiences such as poor management and unsatisfactory working conditions are likely to be shared by team members, and these experiences have been documented to impact work attitudes (Rousseau, 1978). Indeed, the psychological contract literature suggests that individuals belonging to the same team tend to exhibit normative behavior (Coach and

French, 1948), affect (George, 1990), and job attitudes (Pfeffer, 1980). Following this line of reasoning, it can be argued that the commonality of the work environment experienced by team members should contribute to the formation of shared psychological contracts. Social information processing theory (Salancik and Pfeffer, 1978) suggests that social information cues facilitate shaping interpretations about job and work environments, including the employing organization. For example, Rousseau and Greller (1994) noted that employees develop an understanding of their psychological contract from a wide variety of sources, which may include organizational agents (e.g., senior managers), or organizational structures and systems (e.g., company policies) that create obligations on behalf of the employer. Along similar lines, team members are potent sources of information because they typically provide their own subjective evaluation of the organization, which they likely discuss with coworkers. In support, Dabos and Rousseau (2004) found that network interaction among employees contribute to the formation of psychological contract beliefs. Hence, exposure to similar communication networks increases the propensity for team members to develop similar beliefs and assumptions about their job and organization. We can refer to this group level construct as 'normative psychological contracts,' defined as *employees' shared mental models concerning what the organization is obligated to provide to the work team*. Future inquiries might validate this proposed construct and explicate its antecedents, processes, and outcomes.

See also: Leader–Member Exchange Theory; Organizational Citizenship Behavior; Organizational Justice; Work Motivation; Workplace Deviance.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. Academic Press, New York, pp. 267–299.

Argyris, C., 1960. Understanding Organizational Behavior. Dorsey Press, Homewood, IL.

Ashforth, B.E., Mael, F., 1989. Social identity theory and the organization. Academy of Management Review 14, 20–39.

Bal, P.M., Chiaburu, D.S., Diaz, I., 2011. Does psychological contract breach decrease proactive behaviors? The moderating effect of emotion regulation. Group and Organization Management 36, 722–758.

Bal, P.M., de Lange, A.H., Jansen, P.G.W., van der Velde, M.E.G., 2013. A longitudinal study of age-related differences in reactions to psychological contract breach. Applied Psychology: An International Review 62, 157–181.

Bal, P.M., Jansen, P.G.W., van der Velde, M.E.G., de Lange, A.H., Rousseau, D.M., 2010. The role of future time perspective in psychological contracts: a study among older workers. Journal of Vocational Behavior 76, 474–486.

Blau, P.M., 1964. Exchange and Power in Social Life. Wiley, New York.

Bordia, P., Restubog, S.L.D., Bordia, S., Tang, R.L., 2008. Breach begets breach: trickle-down effects of psychological contract breach on customer service. Journal of Management 36, 1578–1607.

Coach, L., French, J., 1948. Overcoming resistance to change. Human Relations 1, 512–532.

Conway, N., Briner, R.B., 2002. A daily diary study of affective responses to psychological contract breach and exceeded promises. Journal of Organizational Behavior 23, 287–302.

Coyle-Shapiro, J.A.-M., Neuman, J.H., 2004. The psychological contract and individual differences: the role of exchange and creditor ideologies. Journal of Vocational Behavior 64, 150–164.

Dabos, G., Rousseau, D., 2004. Social Interaction Patterns Shaping Employee Psychological Contracts: Network-wide and Local Effects. Best Papers Proceedings, 64th Annual Meeting of the Academy of Management, New Orleans, Louisiana, USA.

Dulac, T., Coyle-Shapiro, J.A.-M., Henderson, D., Wayne, S., 2008. Not all responses to breach are the same: a longitudinal study examining the interconnection of social exchange and psychological contract processes in organizations. Academy of Management Journal 51, 1079–1098.

George, J.M., 1990. Personality, affect, and behavior in groups. Journal of Applied Psychology 75, 107–116.

Gouldner, A.W., 1960. The norm of reciprocity: a preliminary statement. American Sociological Review 25, 161–178.

Guest, D.E., 1998. Is the psychological contract worth taking seriously? Journal of Organizational Behavior 19, 649–664.

Henderson, D.J., Wayne, S.J., Shore, L.M., Bommer, W.H., Tetrick, L.E., 2008. Leader-member exchange, differentiation, and psychological contract fulfillment: a multilevel examination. Journal of Applied Psychology 93, 1208–1219.

Ho, V.T., Weingart, L.R., Rousseau, D.M., 2004. Responses to broken promises: does personality matter? Journal of Vocational Behavior 65, 276–293.

Hobfoll, S.E., 2001. The influence of culture, community, and the nested self in the stress process: advancing conservation of resources theory. Applied Psychology: An International Review 50, 337–370.

Johnson, J.L., O'Leary-Kelly, A., 2003. The effects of psychological contract breach and organizational cynicism: not all social exchange violations are created equal. Journal of Organizational Behavior 24, 627–647.

Kickul, J., 2001. When organizations break their promises: employee reactions to unfair processes and treatment. Journal of Business Ethics 29, 289–307.

Kickul, J., Lester, S.W., 2001. Broken promises: equity sensitivity as a moderator between psychological contract breach and employee attitudes and behavior. Journal of Business and Psychology 16, 191–217.

Lambert, L.S., Edwards, J.R., Cable, D.M., 2003. Breach and fulfillment of the psychological contract: a comparison of traditional and expanded views. Personnel Psychology 56, 895–934.

Levinson, H., Price, C.R., Munden, K.J., Mandl, H.J., Solley, C.M., 1962. Men, Management, and Mental Health. Harvard University Press, Cambridge, MA.

Lo, S., Aryee, S., 2003. Psychological contract breach in a Chinese context: an integrative approach. Journal of Management Studies 40, 1005–1020.

Morrison, E.W., Robinson, S.L., 1997. When employees feel betrayed: a model of how psychological contract violation develops. Academy of Management Review 22, 226–256.

Ng, T.W., Feldman, D.C., Lam, S.S.K., 2010. Psychological contract breaches, organizational commitment, and innovation-related behaviors: a latent growth modeling approach. Journal of Applied Psychology 95, 744–751.

Organ, D.W., 1988. Organizational Citizenship Behavior: The Good Soldier Syndrome. Lexington Books, Lexington, MA.

Orvis, K.A., Dudley, N.M., Cortina, J.M., 2008. Conscientiousness and reactions to psychological contract breach. Journal of Applied Psychology 93, 1183–1193.

Pfeffer, J., 1980. A partial test of the social information processing model of job attitudes. Human Relations 9, 712–721.

Raja, U., Johns, G., Ntalianis, F., 2004. The impact of personality on psychological contracts. Academy of Management Journal 47, 350–367.

Restubog, S.L.D., Bordia, P., 2006. Workplace familism and psychological contract breach in the Philippines. Applied Psychology: An International Review 55, 563–585.

Restubog, S.L.D., Bordia, P., Tang, R.L., 2006. Effects of psychological contract breach on performance of IT employees: the mediating role of affective commitment. Journal of Occupational and Organizational Psychology 79, 299–306.

Restubog, S.L.D., Bordia, P., Tang, R.L., 2007. Behavioural outcomes of psychological contract breach in a non-western culture: the moderating role of equity sensitivity. British Journal of Management 18, 376–386.

Restubog, S.L.D., Bordia, P., Tang, R.L., Krebs, S.A., 2010. Investigating the moderating effects of leader-member exchange in the psychological contract breach-employee performance relationship: a test of two competing perspectives. British Journal of Management 21, 422–437.

Restubog, S.L.D., Hornsey, M.J., Bordia, P., Esposo, S.R., 2008. Effects of psychological contract breach on organizational citizenship behaviour: insights from the group value model. Journal of Management Studies 45, 1377–1400.

Restubog, S.L.D., Zagenczyk, T.J., Bordia, P., Tang, R.L., 2013. When employees behave badly: the roles of contract importance and workplace familism in predicting negative reactions to psychological contract breach. Journal of Applied Social Psychology 43, 673–686.

Restubog, S.L.D., Zagenczyk, T.J., Bordia, P., Bordia, S., Chapman, G.J. If you wrong us, shall we not revenge? Moderating roles of self-control and perceived aggressive work culture in predicting responses to psychological contract breach. Journal of Management. doi:10.1177/0149206312443557, in press.

Robinson, S.L., 1996. Trust and breach of the psychological contract. Administrative Science Quarterly 41, 574–599.
Robinson, S.L., Morrison, E.W., 2000. Development of psychological contract breach and violation: a longitudinal study. Journal of Organizational Behavior 21, 525–546.
Rousseau, D.M., 1995. Psychological Contracts in Organizations: Understanding Written and Unwritten Agreements. Sage, Thousand Oaks, CA.
Rousseau, D.M., 1978. Characteristics of departments, positions, and individuals: contexts for attitudes and behavior. Administrative Science Quarterly 23, 521–540.
Rousseau, D.M., 1989. Psychological and implied contracts in organizations. Employee Responsibilities and Rights Journal 2, 121–139.
Rousseau, D.M., 1990. New hire perceptions of their own and their employer's obligations: a study of psychological contracts. Journal of Organizational Behavior 11, 389–400.
Rousseau, D.M., 2000. Psychological Contract Inventory: Technical Report. Tech. Rep. No. 2000-2. Carnegie Mellon University, Pittsburgh, PA.
Rousseau, D.M., Greller, M.M., 1994. Human resource practices: administrative contract makers. Human Resource Management 33, 385–401.
Salancik, G.R., Pfeffer, J., 1978. A social information processing approach to job attitudes and task design. Administrative Science Quarterly 23, 224–253.
Schein, E.H., 1980. Organizational Psychology, third ed. Prentice-Hall, Englewood Cliffs, NJ.
Suazo, M.M., Turnley, W.H., Mai-Dalton, R.R., 2005. The role of perceived violation in determining employees' reactions to psychological contract breach. Journal of Leadership and Organizational Studies 12, 24–36.
Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W., Worchel, S. (Eds.), The Social Psychology of Intergroup Relation. Brooks/Cole, Monterey, CA, pp. 33–47.
Turnley, W.H., Feldman, D.C., 1999. The impact of psychological contract violations on exit, voice, loyalty, and neglect. Human Relations 52, 895–922.
Turnley, W.H., Bolino, M.C., Lester, S.W., Bloodgood, J.M., 2003. The impact of psychological contract fulfillment on the performance of in-role and organizational citizenship behaviors. Journal of Management 29, 187–206.
Wanous, J.P., Poland, T.D., Premack, S.L., Davis, K.S., 1992. The effects of met expectations on newcomer attitudes and behaviors: a review and meta-analysis. Journal of Applied Psychology 77, 288–297.
Zhao, H., Wayne, S.J., Glibkowski, B.C., Bravo, J., 2007. The impact of psychological contract breach on work-related outcomes: a meta-analysis. Personnel Psychology 60, 647–680.
Zhou, J., George, J.M., 2001. When job dissatisfaction leads to creativity: encouraging the expression of voice. Academy of Management Journal 44, 682–696.

Psychological Treatment, Effectiveness of

Michael J Lambert, Brigham Young University, Provo, UT, USA; and Queensland University, Brisbane, QLD, Australia
Kara L Harmon, Grand Junction VAMC, Grand Junction, CO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.J. Lambert, E.J. Hawkins, volume 18, pp. 12372–12377, © 2001, Elsevier Ltd.

Abstract

Empirical evidence supports the efficacy and effectiveness of psychological treatments for a wide variety of psychological disorders. It appears that about two-thirds of patients secure positive outcome even though the problems they are struggling with have been long-standing. The gains that are made can be achieved very rapidly in many cases (first five sessions) with more disturbed patients taking longer to return to a state of normal functioning. Positive outcomes are maintained over follow-up periods of 2–5 years. Changes appear to be due to factors common across theory-based treatments.

Introduction

Empirical evidence supports the effectiveness of psychotherapy. Researchers in the field have confirmed empirically that psychotherapy produces beneficial and relatively lasting effects. Contributions from the dose-effect literature suggest that patients improve as the number of sessions increases with diminishing returns at higher dosage levels. Factors important for patient improvement are generally considered features common across rather than unique to particular theories of psychotherapy. Despite intentions to facilitate improvement, a small percentage of patients seem to deteriorate while participating in psychotherapy treatment.

The Definition of Psychotherapy

In general, psychotherapy is a collaborative enterprise (between a trained professional clinician and patient) developed to provide relief from problematic symptoms and help prevent future relapses, support desired personality change, enhance adaptive role functioning in relationships and at work/school, increase the ability to make healthy life choices, and promote overall quality of life through helping patients to negotiate challenges and make changes in their behaviors, emotions, cognitions, and/or other characteristics. Approaches to psychotherapy are numerous and diverse including orientations such as cognitive, behavioral, psychodynamic, humanistic, and eclectic combinations of these theory-based orientations.

The Effectiveness of Psychotherapy

Over the past 40 years, a myriad of research designs, methodologies, and meta-analytic investigations have combined to create a sizable body of scientific evidence that attests to the effectiveness of psychotherapy across diverse settings (hospitals, counseling centers, clinics, military, public health, primary care, etc.), conditions (diagnoses, addictive, health, psychological, and relational problems), and populations (adults, children, adolescents, and older adults; Burlingame et al., 2013; Hofmann and Smits, 2008; Lambert, 2013; Sexton et al., 2013; Weisz et al., 2013). Overall, research has demonstrated that psychotherapy interventions are efficient, safe, demonstrate efficacy across most diagnostic conditions, and have far-reaching effects. Assessments (via self-report, therapist/expert evaluations, evaluations of significant others, and/or societal records) of the benefits of psychotherapy have focused on such domains as reduction in symptoms, increase in problem-solving skills, enhanced health, and improvements in interpersonal/social functioning. Reviews of this literature have confirmed the beneficial effects of therapy concluding that treated patients consistently fare better than untreated patients; in fact, research reveals that 65% of treated patients can expect to have a positive outcome as compared to only 35% of patients who receive no treatment during the same period of time (Lambert, 2013).

Furthermore, when compared with interventions like psychopharmacology, the results of psychotherapy tend to last for longer periods of time with lower relapse rates in many cases. Moreover, given that patients often acquire new knowledge and a variety of useful skills during the course of treatment that they continue to utilize after the treatment has ended, it is common for patients to experience further improvement following treatment termination. Moreover, psychotherapy has proved to be very cost-effective as it often pays for itself in terms of medical cost offset (as patients engaged in psychotherapy are less likely to use inpatient medical services, tend to have significantly shorter hospital stays, etc.), life satisfaction, and increased productivity in work and home roles (e.g., Chiles et al., 1999; Crane, 2008).

Treatment Duration and Lasting Effects

In contrast to the length of treatment associated with psychoanalytic psychotherapy (which often lasted years with multiple visits per week), the extant literature suggests that substantial benefits from treatment occur in relatively short durations (e.g., 14 weekly sessions), and equally important, these gains are maintained over substantial periods of time. Examinations of changes over weeks of treatment have confirmed a positive response to treatment as the dose (i.e., number of sessions) increases, with diminishing returns at higher dosage levels. Studies from the dose-effect literature suggest that, for many

clinical disorders, a range of 10–20 sessions is sufficient for 50% of patients to return to normal functioning (Harnett et al., 2010). Moreover, promising results occur as the dosage level is increased to the extent that 75% of patients can be expected to recover substantially after 26–50 sessions (Hansen et al., 2002).

Although a majority of patients experience significant benefits from psychotherapy, there is also evidence that the more disturbed a person is at the outset of therapy, the longer (more sessions) it will take for them to return to a normal state (Lambert et al., 2001). Observed trends in patient improvement have led some researchers to propose a phase model of recovery in which personal welfare, symptom reduction, and interpersonal/social functioning improve in sequence (Kopta et al., 1994).

In addition to results supporting relatively short recovery periods for a majority of patients, there is evidence that beneficial gains from psychotherapy are enduring. Although adjustment from psychotherapy treatment does not prevent recurrences of psychological disturbance, it appears that many patients maintain healthy functioning for prolonged periods of time subsequent to treatment (Dugas et al., 2010). However, exceptions to the notion of lasting treatment gains are noted in work with patients who are obese, have substance abuse, eating disorders, or who are suffering from symptoms of recurrent depression or personality disorders. Conversely, research has also detected a large minority of patients, perhaps 30%, who are able to attain benefit from therapy after a surprisingly small number of sessions (three to five); these clients are said to have 'sudden gains' or 'early dramatic treatment response' and approximately 80% have been found to maintain gains and remain improved 1 year later (Haas et al., 2002).

Factors Important to Psychotherapy Outcome

In addition to the research focused on examining the effectiveness of psychotherapy, considerable emphasis has been placed on the factors that contribute to positive patient outcome. Findings from this research fall into three general categories: patient variables, common variables, and specific intervention variables.

Patient Variables

The patient is most likely to determine successful outcome. The literature focusing on the study of patient variables is too voluminous to repeat, and thus only the most important findings are reported here. Each patient entering treatment brings a diverse array of factors when presenting to the clinician. These include different psychological disorders, historical backgrounds, stresses, personality structures, and the quality of their social support networks, to name just a few. Patient variables found to have relationships with outcome are severity of disturbance, motivation and expectancy for treatment, capacity to relate, integration, psychological mindedness, and ability to recognize relevant problems (Bohart and Wade, 2013). Therefore, patients suffering from challenging symptoms as well as relative deficits in numerous patient variable categories are likely to experience negligible improvement during treatment. For example, the borderline, alcoholic patient with suicidal tendencies who is forced into treatment, believing that most marital problems are a result of an insensitive spouse, is likely to benefit less from treatment than the depressed patient who voluntarily begins treatment and is determined to make personal changes that will lead to marital harmony (Sexton et al., 2013).

Although patient variables have been conveniently categorized, the impact each has on the outcome of therapy is not equal. In general, the severity of disturbance is the patient variable that is most related to the outcome of the treatment (Garfield, 1994). Patients with serious mental disorders, such as the schizophrenic, schizoaffective, and bipolar disorders are typically treated primarily with psychopharmacological management, suggesting the challenge these disturbances provide for psychotherapy alone (Forand et al., 2013). Personality disturbances also appear resistant to change in the short-term use of psychotherapy (i.e., 12 sessions).

Common Factors

Because outcome researchers have consistently failed to find sufficient evidence indicating superior improvement of patients as a consequence of receiving a particular kind of psychotherapy, some have concluded that psychotherapy equivalence is a result of factors common to all psychotherapy (Cuijpers et al., 2012; Grawe, 2002). Common factors propose the possibility that there are components utilized in therapy that are curative and universal – regardless of specific school or theory of change – which include support, learning, and action elements (Lambert, 2013). The developmental nature of the support–learning–action sequence presumes that the supportive elements precede modifications in attitudes and beliefs which, consequently, precede the clinician's attempts to encourage client action. It is presupposed that all therapies provide for a collaborative working enterprise wherein the client's sense of safety, trust, and security (along with resultant decreases in anxiety, threat, and tension) lead to changes in conceptualizing problems and, ultimately, in behaving differently by taking informed risks to talk about a sensitive concern, experiment with new coping skills, problem-solve through purposely exposing oneself to anxiety-producing situations, and/or engage in other growth-promoting behaviors that are commonly challenging and distressing.

More specifically, it has been proposed that the therapist–client relationship is a common factor crucial to positive treatment outcomes that has multidimensional and substantial research base (Wampold, 2001). While techniques are influential, the extent to which unique methods can be used to effect change is largely determined by the nature of the relational context in which they are employed. In clinical trial studies comparing multiple treatments for particular psychological disorders, patients' positive assessment of the therapeutic alliance is one of the best predictors of psychotherapy outcome (e.g., Krupnick et al., 1996). If factors like acceptance, trust, kindness, understanding, warmth, and human consideration are not present, it is unlikely that the interventions utilized will result in desired transformation. Furthermore, a systematic review of psychotherapy research by Norcross and Lambert (2011) on the role of relationship factors in therapy revealed that empathy, alliance, and collecting client feedback were all 'demonstrably effective' in

facilitating positive therapy outcomes while positive regard, goal consensus, and collaboration were found to be 'probably effective' components of therapy (*see* Therapist–Patient Relationship).

However, although a significant amount of evidence points to common factors as mediators of patient change, the notion of beneficial techniques unique to specific therapies cannot be ruled out, as occasionally such interventions can be shown to contribute to successful outcomes in patients independent of common factors (*see* Psychological Treatments: Randomized Controlled Trials).

Specific Interventions

The American Psychological Association and other professional associations in other countries (e.g., Germany and England) are constantly in the process of developing clinical treatment guidelines (e.g., National Institute of Health and Care Excellence (NICE) guidelines in the United Kingdom) that can help identify particular treatments and procedures that research has demonstrated to be most effective with specific problems and/or patients. The evidence for treatments is usually judged to be strongly supported (SS) or moderately supported (MS) based on the kind and amount of evidence that has accumulated (see Chambless and Hollon, 1998; for an example of one rating system). One should keep in mind, however, when reviewing the following sections that denote specific therapeutic approaches that the empirically supported treatments that are summarized are not necessarily uniquely effective nor are the outcomes in most cases so superior to other forms of therapy to justify their sole promotion or use. In fact, as stated prior, differences in outcome when comparing various forms of therapy are most often not as distinct as expected, hoped for, or insinuated.

Moreover, other significant obstacles to this type of specificity and approach arise including the sheer number of therapeutic approaches/theories that exist (which is in the hundreds if not the thousands), the constant evolution of therapies subsequent to research and in its absence, and the emergence of new methodologies that appear promising. Some would argue that it is not reasonable to expect a single therapist to sufficiently master the number of therapies it would require to serve a general clientele given that the number of discrete disorders surpasses 300. The expense, impracticality, and lack of research evidence on so many therapies with so many disorders may prove insurmountable. Furthermore, most often clients present with a myriad of complex symptoms and circumstances of life that do not easily fall into definite categories to be matched up with a single clear-cut treatment.

Lastly, as mentioned before, while technical proficiency undoubtedly makes a unique contribution to therapy, any efforts to propagate a best practice or an empirically supported psychotherapy while dismissing the impact and importance of the therapeutic relationship is potentially misleading and incomplete (Norcross and Wampold, 2011) as it is this positive interpersonal affective environment that often stimulates patient improvement in the therapeutic encounter. Nevertheless, at least empirically supported treatments have been rigorously tested and undergone the peer review necessary for publication and there is some information about their comparative effects. We support the clinical use of empirically supported treatments over treatments with no empirical evidence.

Anxiety

Specific disorders within the anxiety spectrum for which particular techniques have been researched include generalized anxiety disorder (GAD), panic disorder, and posttraumatic stress disorder (PTSD).

Generalized Anxiety Disorder

The basic premise of cognitive behavioral therapy (CBT) for GAD (SS) is that thoughts, behaviors, and feelings are interconnected and so using an intervention to change one aspect (e.g., behavior) can lessen difficulties in the others (e.g., thoughts and feelings). CBT typically refers to a treatment wherein combinations of cognitive and behavioral techniques are utilized in the same course of treatment. In CBT for GAD, these interventions might include modifying negative thinking (e.g., that worrying is beneficial in some way), taking part in relaxation training, planning pleasurable activities, and/or intentional exposure to avoided situations that are objectively safe to learn that anticipated, feared outcomes do not occur. CBT for GAD typically encompasses 16–20 sessions over a 4-month period. Acceptance and commitment therapy (ACT) for GAD (MS) is a behaviorally based intervention that focuses on acceptance, rather than avoidance, of internal experience (which in GAD typically includes the unwanted emotion of anxiety, anxiety-inducing thoughts, and/or bodily sensations like racing heart, sweating, rapid breathing, etc.). Through using defusion and mindfulness techniques, patients are able to reduce avoidance and commit to behaving in ways that are consistent with chosen values (listed with support for chronic pain, depression, anxiety, obsessive–compulsive disorder (OCD), and psychosis).

Panic Disorder

CBT (SS) has strong research support for treating panic disorder. Most frequently, the therapist utilizes a combination of psychoeducation (about panic/fear and the body's physiological response), cognitive techniques (helping the patient identify, evaluate, and modify problematic thoughts related to panic such as perceived catastrophic consequences of specific places, people, bodily sensations, feelings, etc.), and exposures that incorporate *in vivo* (e.g., intentional approach of real-world, trauma-related situations that are objectively safe yet have been avoided previously due to trauma-related fear) and/or interoceptive (e.g., inducing bodily sensations like a racing heart, trouble breathing, dizziness, and sweating) aspects. CBT for panic is typically 12–16 sessions and can be delivered in either an individual or group format. The focus of applied relaxation for panic disorder (MS) is helping the patient to identify situations in which panic is likely and to raise awareness of panic cues. Patients are then trained in the skill of progressive muscle relaxation where they are taught to relax quickly while in increasingly stressful situations.

Posttraumatic Stress Disorder

Prolonged exposure for PTSD (SS) is a specific cognitive behavioral approach for PTSD that includes education regarding common reactions to trauma and PTSD, breathing

retraining (for relaxation), repeated recounting and recording of trauma memory(ies) during therapy sessions, and homework assignments that include listening to the recorded trauma account daily and engaging in *in vivo* exposure (confronting trauma-related, real-world situations that are objectively safe yet have been avoided previously due to trauma-related fear). This treatment typically takes between 10 and 16 sessions. Cognitive processing therapy (CPT) for PTSD (SS) involves psychoeducation, written homework assignments, and a Socratic style of exploration during therapy that helps patients to examine their beliefs about self, others, and the world – especially in the areas of safety, trust, power/control, esteem, and intimacy. Together, the client and therapist identify assimilation, overaccommodation, and accommodation in the client's thoughts and the end goal is to develop beliefs that are balanced and realistic regarding the trauma which, consequently, reduces symptomatology. CPT can be administered in individual or group format and is typically 12 sessions. Present-centered therapy for PTSD (SS) is a supportive intervention that focuses on current life problems as manifestations of PTSD. Psychoeducation about PTSD is provided and the therapist then aids the client in identifying stressors in their everyday life; these general difficulties are then reviewed and discussed in a nondirective, supportive fashion. Stress inoculation therapy for PTSD (MS) teaches patients a variety of coping skills (e.g., breathing exercises, thought stopping, assertiveness skills, deep muscle relaxation, role playing, cognitive restructuring, and/or guided self-dialogue) to help them more effectively manage feelings of anxiety and stress. This approach is often used concurrently with other therapies (e.g., CBT).

Depression

This section will review the individual/group as well as specific couple approaches that have been researched for treating depression.

Individual and/or Group Approaches

CBT for depression (SS) includes interventions such as identifying and modifying negative thinking (e.g., all/nothing or catastrophic thinking and jumping to conclusion) and simultaneously incorporating behavioral interventions like exercise, planning pleasurable activities, etc. It should be noted that behavioral therapy (BT) and cognitive therapy (CT) have been utilized individually and independently (as BT and CT rather than in combination as in CBT), each with strong support for efficacy; hence, a brief review of each will follow. BT/behavioral activation (BA) (SS) is a short-term approach (approximately 12 sessions) that utilizes strategies that focus on increasing the quality and frequency of pleasant activities so that patients can increase their sense of mastery, decrease aversive consequences (e.g., of withdrawal, escape behaviors, and disrupted routines), and improve mood. BA uses graded exercises to systematically increase contact with sources of reward (e.g., pleasant activities), identify barriers to activation, teach problem-solving skills, and improve quality of life. It is believed that no modification of thoughts or insight is necessary for change to occur. Likewise, CT and mindfulness-based cognitive therapy (MBCT) (SS) focus almost solely on awareness and modification of thoughts for improving depression. CT is a time-limited, problem-focused, structured therapy that emphasizes the identifying, monitoring, recording, and modification of maladaptive thoughts and beliefs (e.g., catastrophic thinking, all/nothing thinking, emotional reasoning, jumping to conclusions, and personalization); this challenging and altering of unhelpful cognitions and beliefs is often called cognitive restructuring. Problem solving and development of adaptive coping skills is also a frequent focus. CT for depression can be delivered in both group and individual formats and is typically 14–16 sessions in length. MBCT combines CT techniques with mindfulness meditation. In mindfulness meditation, patients learn to observe, recognize, and change their relationship with negative thought processes. Mindfulness promotes acceptance, nonjudgmentalness, and balanced awareness of self that reduces relapse risk in depression and also often leads to various other medical and psychological benefits. In interpersonal psychotherapy (IPT) for depression (SS), it is proposed that, as psychiatric difficulties most often occur within an interpersonal/social context, teaching patients to identify and understand the relationship between the onset and fluctuation of symptoms and interpersonal problems is essential. Specific interpersonal domains include grief, role transitions, interpersonal role disputes, and interpersonal deficits. The therapist (using techniques that often include clarification, encouragement of affect, communication analysis, and supportive listening) helps to motivate the patient to work toward change in the area of interpersonal difficulties, which, consequently, alleviates symptoms. This treatment is typically 12–16 sessions and can be delivered in both individual and group formats. Problem-solving therapy (PST) (SS) teaches patients to more successfully generate solutions for problems (such as the pursuit of goals or resolution of interpersonal conflicts). PST is a short-term therapy that typically runs between 4 and 16 sessions that can be provided in individual or group formats. Self-management/self-control (SS) is a 10-session program based on Kanfer's (1970) model of self-control that can be delivered in group or individual therapy. Homework assignments, didactic presentations, and instructional exercises are utilized to teach skills and concepts to help patients reduce self-punishment, unrealistic self-evaluative standards, and inaccurate overresponsibility for negative events and to increase self-reinforcement and awareness. Cognitive behavioral analysis system of psychotherapy (SS) is an integrative approach for chronically depressed adults that combines aspects of interpersonal, cognitive, behavioral, and psychodynamic therapies to help patients become more empathetic, alter problematic interpersonal dynamics, and heal from trauma using three primary techniques (situational analysis – raising patient awareness of interpersonal impact on others and altering problematic behaviors; interpersonal discrimination exercises – comparing/contrasting traumatic and healthy relationships to differentiate and allow for more informed decisions; and behavioral skills training/rehearsal – e.g., assertiveness training). ACT for depression (MS) is a behaviorally based intervention that focuses on acceptance, rather than avoidance, of internal experience related to depression (typically unwanted emotions, depression-oriented thoughts, unwelcome memories, and/or bodily sensations) by using defusion and mindfulness techniques so that the patients can increase their activity and ability to engage in behaviors that are meaningful and allow them to move forward in valued directions. ACT also has empirical support for being effective with other difficulties that can contribute to depression such as chronic pain, OCD, and

psychosis. Emotion-focused therapy (EFT) (MS) is built on a process-experiential approach designed to facilitate patient identification, utilization, and processing of emotion in the environment of a safe, secure, therapeutic relationship. Three specific phases (emotion awareness, emotion regulation, and emotion utilization/transformation) assist patients to increase awareness of emotions, deepen emotional experiences, recognize/regulate unhealthy emotional responses, and utilize healthy emotions to guide behavior. This approach is typically delivered in an individual format and is 16–20 sessions in length. Rational emotive behavioral therapy (MS) is a type of short-term (i.e., 14–20 sessions), present-focused CBT that focuses almost exclusively on efforts to diminish demanding beliefs (e.g., 'musts,' 'shoulds,' and 'oughts'), reduce secondary problems (e.g., depressive feelings about depression), and promote unconditional self-acceptance. The A-B-C model (action, belief, and consequence) and other variants are also often utilized to identify problematic beliefs to focus on in therapy. It can be delivered in either individual or group formats. Short-term psychodynamic therapy (MS) is a brief dynamic therapy aimed at increasing the patient's insight and awareness of core conflictual relational themes and problematic patterns that contribute to depression. Other common foci might include expressions of emotion and identification of how current functioning is impacted by past experiences. This approach typically takes place individually in 16–20 sessions.

Couple Therapy Approaches for Depression

Behavioral couples therapy (BCT) (MS) is designed for couples wherein one of the partners is depressed. BCT focuses on problem solving, communication, and the exchanges that occur between partners to improve relationship functioning, support, and to decrease conflict, which, consequently, decrease depressive symptoms in the affected partner. This treatment typically takes between 12 and 20 couple sessions. Emotionally focused couples therapy (EFCT) (MS) focuses on the here-and-now construction of the emotional experience and interpersonal patterns of interactions of a distressed couple. Psychoeducation regarding adult love and bonding processes is provided and the EFCT therapist aids the couple in reshaping interactions so that a more secure attachment can be formed. Through reorganizing interactions, reprocessing about experience, and increasing safety, trust, and contact, the couple is able to create a more secure bond that benefits both. Treatment typically comprises 8–20 couple sessions.

Eating Disorders and Obesity

Empirically supported interventions for anorexia nervosa, bulimia nervosa, binge eating disorder, and obesity have been proffered, several of which include family components.

Anorexia Nervosa

Family-based treatment (FBT) (SS) for anorexia nervosa is a three-phase treatment that may follow a brief stay in an inpatient unit (if medical concerns are present) or can begin on an outpatient basis. In the first phase, the parents collaborate with the therapist to assume responsibility for weight restoration and nutritional rehabilitation (while the adolescent maintains autonomy in other areas of life as developmentally appropriate). In the second phase, (after acute starvation is reversed), control over eating is returned to the adolescent. The third phase includes addressing issues of family structure, normal adolescent development, and termination. This approach typically takes place in 20 sessions over the course of 12 months. CBT for anorexia nervosa (MS) is a post-hospitalization relapse intervention designed to prevent a return of symptoms following the completion of inpatient treatment and subsequent weight gain. It is typically conducted in individual sessions that occur over 1 year wherein behavioral strategies (regular pattern of eating, systematic exposure to 'forbidden' foods) and attending to cognitive features of the disorder (disturbances in perception of weight/shape, motivation for changing) are utilized in order to accomplish the explicit goals of achieving and maintaining a healthy weight.

Bulimia Nervosa

CBT for bulimia nervosa (SS) is a present-centered treatment that targets the current symptom cycle rather than attempting to determine etiology. It includes three phases of treatment over approximately 20 weekly sessions. In the first phase, psychoeducational information regarding adverse physiological consequences of extreme dieting, bingeing, purging, and weight is shared and the aim is to establish a regular pattern of eating as well as developing an appropriate schedule for monitoring weight. Phase two is designed to reducing diet behaviors as well as weight/shape concerns and identification of relapse warning signs is also a focus. In the third phase, the focal point is maintenance of progress and relapse prevention. CBT is utilized throughout the phases to disrupt variables (e.g., cognitive distortions) that perpetuate the binge–purge cycle and then to achieve and maintain abstinence from bingeing and purging behaviors. IPT has also been modified for treating bulimia nervosa (SS). While the connection between bulimia nervosa and interpersonal difficulties is identified at the beginning of therapy, the symptoms of bulimia nervosa are never an explicit focus. Rather, an interpersonal problem domain is identified (grief, role transitions, interpersonal role disputes, and interpersonal deficits) and interpersonal precipitants are identified. The focus then shifts to facilitating change in the interpersonal realm with the third phase focused on progress made and relapse prevention. Treatment typically takes 20 weekly sessions and can be delivered in individual or group format. While strong research support for IPT for bulimia nervosa exists, it has been noted that these gains tend to be made more slowly than in CBT. FBT for bulimia nervosa (MS) was adapted from FBT for anorexia nervosa and, consequently, also has three phases of treatment. This treatment typically encompasses 20 sessions over 6 months.

Binge Eating Disorder

CBT (SS) for binge eating disorder is a three-phase treatment that primarily focuses on the maladaptive eating patterns and feelings of distress and loss of control that precipitate binge eating behaviors. In the first phase, the approach is dependent upon whether the individual is obese or not; if obesity is an issue the aim is to persuade the patient to prioritize termination of binge eating over immediate weight loss so that an appropriate weight monitoring and eating schedule can be established (if the person is not obese this step is not necessary). Psychoeducational information regarding nutrition is supplied and a regular

exercise program is encouraged. Phase two encompasses identification of problematic beliefs ('food addiction,' self-esteem based solely on appearance, etc.) and a goal of reducing weight/shape concerns and increasing the definition of self-worth ensues. The last phase is devoted to maintenance and relapse prevention planning. CBT techniques are used throughout and aimed at the disruption of and cessation from factors (thoughts, emotions, behaviors, etc.) that perpetuate binge eating. This treatment typically takes place over approximately 20 weekly sessions in either individual or group formats. IPT has also been adapted for binge eating disorder (SS). Typically, IPT for binge eating disorder takes place over 20 weekly sessions in a group or individual format.

Obesity and Pediatric Overweight

Behavioral weight loss (BWL) treatment (SS) is a short-term approach aimed at establishing new behavioral patterns to achieve acute weight loss and maintain weight reductions attained. The LEARN (lifestyle, exercise, attitudes, relationships, and nutrition) Program is a 12-session holistic approach for adults that encourages small lifestyle changes that, along with reasonable weight goals and moderation in food choices, yields weight loss. Adaptations of BWL for children have also been developed.

The Negative Effects of Psychotherapy

While research overwhelmingly suggests that psychotherapy facilitates reductions in symptomatology, improves everyday functioning, and results in appreciable gains and clinically meaningful change, contrary to the intent of psychotherapy, research also shows that there are some clients who actually worsen during the course of treatment. Similar to client improvement, research shows the phenomenon of deterioration to be equivalent across theoretical orientations (Lambert et al., 1977). Likewise, negative outcomes have been observed across all client populations, treatment interventions, and family and group therapies. In fact, research shows a relatively consistent proportion of adults (5–10%) and an alarmingly high portion of children (14–24%) deteriorate when participating in treatment (Lambert, 2013; Warren et al., 2010). While therapy factors (inadequate conceptualization, applying treatments that do not cover the full range of patient problems, unsuitable treatment) may be liable, this finding does not mean that all instances of worsening are attributable to therapy itself as a myriad of other factors may be present.

Instances that are independent from treatment can occur such as an inevitable negative life event (e.g., loss of loved one and medical condition) or a progressive decline that is inescapable (such as dementia). Unfortunately, as mentioned prior, severity of mental illness alone is predictive of unsuccessful outcome. Higher levels of interpersonal problems and a higher degree of seriousness in presenting problems also predict poor results. In addition, certain diagnoses and symptoms (e.g., borderline personality disorder, OCD, psychotic, schizophrenic, and bipolar) seem less amenable to psychotherapeutic intervention. Client, setting, and therapist factors related directly to treatment might also help account for treatment failure (e.g., the inability of therapists to repair ruptures in the therapeutic alliance; Muran and Barber, 2006).

Just as some client characteristics help maximize gains, other client factors such as poor motivation, sporadic attendance, hostility, resistance to therapist suggestions and prescriptions, complicated problems, negative expectations of treatment, dysfunctional relationship dynamics or environment, and/or unreported problems/issues predict poor outcomes (Bohart and Wade, 2013). The setting in which the client is seen also seems to play an important role in whether clients improve or deteriorate (Warren et al., 2010). In general, about two-thirds of adults who participate in therapy in a randomized controlled trial (where client comorbidity is minimized, therapists are experts in the approach, manuals standardize treatment, etc.) have a positive outcome in approximately 14 sessions and about 30% of patients show no benefit or worsen; in routine practice settings (e.g., community health centers and employee assistance programs) research shows that clients do not fare as well as only about one-third demonstrated improvement or recovery, more than 50% did not respond, and an average of 8% deteriorated (Hansen et al., 2002).

Lastly, specific therapist behaviors have also been found to lead to treatment failure. Clinician competence and mastery vary and, not surprisingly, some therapists respond better to challenging patient presentations than others. Studies focused on deterioration have found lack of empathy, underestimation of the severity of client concerns, and/or overestimation of client progress to be detrimental; multiple studies also underscore the destructiveness of therapist negative reactions toward the client that reflect irritation, hostility, and disappointment (Mohr, 1995; Henry et al., 1986).

Fortunately, innovative resources have been developed to help reduce rates of deterioration in psychotherapy. Routine monitoring of outcome measures taken at each contact with the client can provide real-time clinically relevant feedback to the therapist that provides a benchmark for the client's progress in therapy (Boswell et al., 2013). One such resource has created software that employs score-specific algorithms that track progress, detect 'off-track cases,' sends alarm signals to the therapist for those predicted to have a negative outcome (e.g., deterioration and premature termination) so that ameliorative action can be taken; the result of such an intervention resulted in reducing deterioration rates from 20 to 13%. When clinical support tools were incorporated that helped to identify possible causes of negative response and provide possible solutions/ideas for corrective interventions, the deterioration rate dropped to approximately 8% while positive outcomes doubled (Lambert, 2010; Shimokawa et al., 2010).

Conclusion

Decades of research compellingly demonstrate that when aligned with a wise, skillful, sound provider, clients entering therapy can expect to reduce symptoms, accelerate natural healing processes, improve life functioning, and learn additional methods and coping strategies to present return of their symptomatology. Those who seek therapy consistently have better outcomes than those who do not receive treatment (or are on a wait list) and a broad range of therapies have be shown to assist patients make changes that are both statistically and

clinically meaningful in relatively short periods of time. For many clinical disorders, generally one-half of all patients return to normal functioning in 10–20 sessions. An additional 25% of patients experience substantial improvement when dosage levels are increased to 30–50 sessions. With the exception of disorders that tend to recur (e.g., addiction disorders and personality disorders), the benefits from treatment appear to be durable as the results of follow-up studies suggest that the effects of psychotherapy endure for at least 1–2 years subsequent to treatment. These results seem at least partially attributable to the emphasis and efforts that psychotherapists make to help clients make enduring (as opposed to temporary) changes through consolidating gains, strengthening coping skills, and safety planning toward the end of therapy.

Unfortunately, not all patients experience beneficial effects of treatment and investigation of factors that contribute to both positive and negative change are ongoing. One confirmed way to monitor the efficacy of treatment and improve client success is to implement ongoing monitoring of client mental health functioning with alarm signals and problem-solving tools for off-track clients (Shimokawa et al., 2010).

See also: Evidence-Based Psychotherapy; Imaging Methods in Psychotherapy; Manual Based Treatments; Program Evaluation; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy and Pharmacotherapy; Psychotherapy Integration; Psychotherapy Process Research; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research; Therapist–Patient Relationship.

Bibliography

Bohart, A.C., Wade, A.G., 2013. The client in psychotherapy. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 219–257.

Boswell, J.F., Kraus, D., Miller, S., Lambert, M.J., 2013. Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychotherapy Research. http://dx.doi.org/10.1080/10503307.2013.817696 (Epub ahead of print).

Burlingame, G.M., Strauss, B., Joyce, A., 2013. Small group treatment: evidence for effectiveness and mechanisms of change. In: Lambert, M.J. (Ed.), Bergin & Garfield's Handbook of Psychotherapy & Behavior Change, sixth ed. Wiley, New York, pp. 588–640.

Chambless, D.L., Hollon, S.D., 1998. Defining empirically supported therapies. Journal of Consulting and Clinical Psychology 66, 7–18.

Chiles, J.A., Lambert, M.J., Hatch, A.L., 1999. The impact of psychological interventions on medical coast offset: a meta-analytic review. Clinical Psychology, Science and Practice 6, 204–220.

Crane, D., 2008. The cost-effectiveness of family therapy: a summary and progress report. Journal of Family Therapy 30 (4), 399–410.

Cuijpers, P., Driessen, E., Hollon, S.D., van Oppen, P., Barth, J., Andersson, G., 2012. The efficacy of non-directive supportive psychotherapy for adult depression: a meta-analysis. Clinical Psychology Review 32, 280–291.

Dugas, M.J., Brillon, P., Savard, P., Turcotte, J., Gaudet, A., Ladouceur, R., Gervais, N.J., 2010. A randomized clinical trial of cognitive behavioral therapy and applied relaxation for adults with generalized anxiety disorder. Behavior Therapy 41, 46–58.

Forand, N., DeRubeis, R., Amsterdam, J., 2013. Combining medication and psychotherapy in the treatment of major mental disorders. In: Lambert, M.J. (Ed.), Bergin & Garfield's Handbook of Psychotherapy & Behavior Change, sixth ed. Wiley, New York, pp. 691–735.

Garfield, S.L., 1994. Research on client variables in psychotherapy. In: Bergin, A.E., Garfield, S.L. (Eds.), Handbook of Psychotherapy and Behavior Change, fourth ed. Wiley, New York, pp. 190–228.

Grawe, K., 2002. Psychological Therapy, second ed. Hogrefe & Huber, Gottingen.

Haas, E., Hill, R., Lambert, M.J., Morrell, B., 2002. Do early responders to psychotherapy maintain treatment gains? Journal of Clinical Psychology 58, 1157–1172.

Harnett, P., O'Donovan, A., Lambert, M.J., 2010. The dose response relationship in psychotherapy: implications for social policy. Clinical Psychologist 14, 39–44.

Hansen, N., Lambert, M.J., Forman, E.M., 2002. The psychotherapy dose-response effect and its simplification for treatment delivery services. Clinical Psychology: Science and Practice 9, 329–343.

Henry, W.P., Schacht, T.E., Strupp, H.H., 1986. Structural analysis of social behavior: application to a study of interpersonal process in differential psychotherapeutic outcome. Journal of Consulting and Clinical Psychology 54, 27–31.

Hofmann, S., Smits, J., 2008. Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. Journal Clinical Psychiatry 69, 621–632.

Kanfer, F.H., 1970. Self-regulation: research, issues and speculations. In: Neuringer, C., Michael, L. (Eds.), Behavior Modification in Clinical Psychology. Appleton-Century-Crofts, New York.

Kopta, S.M., Howard, K.I., Lowry, J.L., Beutler, L.E., 1994. Patterns of symptomatic recovery in psychotherapy. Journal of Consulting and Clinical Psychology 62, 1009–1016.

Krupnick, J.L., Sotsky, S.M., Simmens, S., Moyer, J., Elkin, I., Watkins, J., Pilkonis, P.A., 1996. The role of the therapeutic alliance in psychotherapy and pharmacotherapy outcome: findings in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Journal of Consulting and Clinical Psychology 64, 532–539.

Lambert, M.L., 2010. Prevention of Treatment Failure: The Use of Measuring, Monitoring, & Feedback in Clinical Practice. American Psychological Association Press, Washington, DC.

Lambert, M.J., 2013. The efficacy and effectiveness of psychotherapy. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 139–193.

Lambert, M.J., Bergin, A.E., Collins, J.L., 1977. Therapist-induced deterioration in psychotherapy. In: Gurman, A.S., Razin, A.M. (Eds.), Effective Psychotherapy: A Handbook of Research. Pergamon, New York, pp. 452–481.

Lambert, M.J., Hansen, N.B., Finch, A.E., 2001. Patient-focused research: using patient outcome data to enhance treatment effects. Journal of Consulting and Clinical Psychology 69, 159–172.

Mohr, D.C., 1995. Negative outcome in psychotherapy: a critical review. Clinical Psychology: Science and Practice 2, 1–27.

Muran, S., Barber, J., 2006. The Therapeutic Alliance: An Evidence-Based Guide to Practice. Guilford Press, New York, NY.

Norcross, J.C., Lambert, M.J., 2011. Evidence-based therapy relationships. In: Norcross, J.C. (Ed.), Psychotherapy Relationships that Work, second ed. Oxford University Press, New York, NY, pp. 3–21.

Norcross, J.C., Wampold, B.E., 2011. Evidenced-based therapy relationships: research conclusions and clinical practices. In: Norcross, J.C. (Ed.), Psychotherapy Relationships that Work, second ed. Oxford University Press, New York, NY, pp. 423–430.

Sexton, T.L., Datchi, C., Evan, L., LaFollette, J., 2013. Levels of evidence for the models and mechanisms of therapeutic change in family and couple therapy. In: Lambert, M.J. (Ed.), Bergin & Garfield's Handbook of Psychotherapy & Behavior Change, sixth ed. Wiley, New York, pp. 542–587.

Shimokawa, K., Lambert, M.J., Smart, D.W., 2010. Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of Consulting and Clinical Psychology 78, 298–311.

Wampold, B.E., 2001. The Great Psychotherapy Debate: Models, Methods, and Findings. Erlbaum, Mahwah, NJ.

Warren, J.S., Nelson, P.L., Mondragon, S.A., Baldwin, S.A., Burlingame, G.M., 2010. Youth psychotherapy change trajectories and outcomes in usual care: community mental health versus managed care settings. Journal of Consulting and Clinical Psychology 78 (2), 144–155. http://dx.doi.org/10.1037/a0018544.

Weisz, J., Ng, M.Y., Rutt, C., Lau, N., Masland, S., 2013. Psychotherapy for children and adolescents. In: Lambert, M.J. (Ed.), Bergin & Garfield's Handbook of Psychotherapy & Behavior Change, sixth ed. Wiley, New York, pp. 691–735.

Psychological Treatments Delivered Nontraditionally

Marion K Jacobs, University of California, Los Angeles, CA, USA
Jim Rue, Laguna Beach, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological treatment traditionally implied an ongoing, face-to-face relationship between a patient and a trained mental health professional. However, over the last 35 years, cultural shifts, new technologies, and a reshaped health care delivery system have resulted in the emergence of new types and formats of psychological assistance. Nontraditional treatments and formats now include: self-help/mutual support groups; computerized programs; radio and TV talk shows; Internet offerings via blogs, chat rooms, instant messaging, bulletin boards, e-mail, and social media outlets; and self-help books, CDs, and DVDs. Not only mental health professionals but paraprofessionals and nonprofessionals (widely varying in their qualifications to offer such help) create these materials and market these services.

As delivery formats have extended to new forms, so has the nature of the contact between user and provider. While some treatments are still face to face, newer alternatives are direct but electronic communications (not necessarily in real time) through telephone, TV, radio, web sites, or e-mail. The connection between the helper and the person receiving help is further attenuated through indirect interactions, for example, using interactive computerized programs. Finally, when printed materials or materials offered on CDs and DVDs are used on a self-help basis, there is no two-way connection between those offering and those receiving psychological treatments.

The varieties and uses of nontraditionally delivered psychological assistance, both as adjuncts to traditional treatment and as the sole form of intervention, have burgeoned in the last 10 years, spurred by advances in the field of Information Technology (IT). A notable example is the 'virtual environment' technology that provides users with the opportunity to learn and practice new behaviors under conditions mimicking the real-life experience. Another is e-mobile applications ('apps') downloaded to mobile phones, tablets, and e-readers. Using apps for guidance, help seekers can practice skill development at their convenience, as often as they like. If a treatment provider is directly delivering services to that user, treatment-oriented apps provide an additional opportunity for the provider to monitor the user's guided behavioral practice between sessions. Given the pace of technological innovation, this trend seems likely to accelerate. However, priority needs to be given to researching the safety and effectiveness of these new mental health interventions.

An Evolving Picture

The term psychotherapy traditionally connoted an interactive, face-to-face process in which a trained mental health professional attempts to help a client, or group of clients, effect personal change. Authority and responsibility for facilitating such change reside with the professional, whose presence and expertise are deemed necessary to guide and control the meetings in accordance with sound clinical and ethical practices.

Since the 1970s, an array of interventions for delivering psychological assistance has come into being that deviates significantly from the traditional psychotherapy format. Much of this changing landscape is based on developments in IT, meaning the use of computers and telecommunication equipment, including TV, telephones, and tablets, to store, retrieve, transmit, and manipulate data. In general, these new formats enjoy widespread popularity among the general public and their use is less directly controlled by a mental health professional. Such alternative psychological interventions include radio and TV shows dispensing psychological advice, (sometimes seemingly more for purposes of entertainment); interactive software for addressing a wide variety of personal problems and providing practice exercises; self-help books, CDs, and DVDs; and mutual support groups that arose after TV talk shows began focusing on problems such as sexual abuse, alcoholism, and other topics previously not openly discussed. More recently, the spate of cyberspace offerings includes self-help guides, single session e-mail and bulletin board advice, ongoing e-mail counseling, real-time chat groups, web telephony, and videoconferencing.

An emerging technological development is the use of 'virtual environments' to deliver psychological treatment. These are computer-controlled, interactive experiences, some using avatars (online proxies) that propel people into simulated three-dimensional worlds, which appear and feel emotionally close to the real-life experience of being in such a setting. Exposure to problematic situations in these virtual environments allows users to experience emotions in a controlled way and practice constructive emotional and behavioral responses in a realistic yet safe setting, thus developing the confidence to deal with the problem effectively in the real world. Virtual environments may be under the guidance of a therapist. For an example, a patient might use realistic flight simulator software in the therapist's office to help overcome a fear of flying. Therapeutic programs can also be utilized solely on one's own, for example, employing avatars at the popular Second Life (www.secondlife.com). Early research findings on the efficacy of virtual approaches to treatment appear promising (DeAngelis, 2012). Virtual reality technology offers important opportunities to bolster the power of psychological treatment approaches. This newer capability of simulating a total reality is a powerful advance in creating the necessary conditions for personal change.

Just as virtual reality represents a nontraditional approach to psychological treatment, so too does the widespread use of applications ('apps') on mobile devices. Many helpful therapeutic programs and other resources are now available on smartphones, tablets, and e-readers, allowing unlimited access, at the convenience of the user. For example, Psychological First Aid (PFA) Mobile is a mobile phone application for mental health providers and other response workers who provide PFA shortly after disasters and emergencies. It represents a collaborative effort among four US government agencies. For research on an app for therapy via cell phone, see Morris et al. (2010). Indeed, as it relates to psychological interventions, easy, low cost, around-the-clock availability of mental health apps, whether in conjunction with other treatments (Morris et al., 2010), or strictly on their own, is an important breakthrough, part of what some call a 'mobility revolution.'

There are promising early research results about the efficacy of Internet-delivered treatments (e.g., Cuijpers et al., 2010; Hester et al., 2011; Andrews et al., 2010). More research is needed to evaluate their effectiveness and safety. That said, there is no denying that convenient, lower cost formats are appealing. They also make psychological aid available to people who would not, or could not, utilize more traditional psychotherapies. Included here are the homebound people living in remote areas, those who shun psychotherapy as stigmatizing and people who insist on total privacy and independence when dealing with their psychological issues.

Social Climate and the Rise of New Forms of Psychological Treatment

Social climate is another important factor giving rise to these new models of psychological assistance. As compared to traditional forms of psychotherapy, nontraditional approaches tend to involve less direct professional contact and control, even when the treatment is designed or delivered by a mental health professional. This shifts away from strictly professional delivery of mental health services, at least as it has evolved in the United States of America, reflects an intersection of several cultural developments.

At a broad societal level, in the United States of America, considerable value has been placed on empowering the individual, as opposed to almost total reliance in the past on institutions and professionals for information, support, and assistance. This has a lot to do with the ease of using search engines, but it also stems from a post-Watergate, post-Vietnam eroding of American public's confidence in the corporate world, government, the medical establishment, and other institutions. Activist movements (civil rights and minorities, women, consumers, gays and lesbians, and others), further reinforce values promoting personal entitlement and empowerment as well as greater acceptance of cultural diversity.

Today, people have more information and demand greater involvement and control over health care decisions than they did in the past. Advances in communications support that demand. For example, before the 1980s, personal issues such as alcohol and drug abuse, child molestation, family violence, sexual orientation, bulimia and anorexia, cancer, sexually transmitted diseases, Alzheimer's disease, fetal alcohol syndrome, mental illness, and posttraumatic stress disorder among Vietnam veterans, were topics rarely spoken of in public. Talk shows of the early 1980s played an important role in bringing such closeted issues into open discussion as people became increasingly sophisticated about such problems. Learning about the mind–body connection and the potentially damaging effects of stress on mental and physical health helped set the stage for receptivity to new delivery formats that have made psychological aid less costly and more widely accessible.

New models meet a clear-cut public need. Scientific surveys of the prevalence and incidence of mental health problems make it clear that we do not have, and probably never will have, a sufficient number of trained mental health professionals to deliver traditional services to all those needing them. What is more, the need for psychological services continues to grow, fed by a continuing population growth and density. That growth is further fed by advances in medical science that increase the survival rates of people with serious chronic medical conditions, putting added stresses on those patients and their families (Kazdin and Rabbitt, 2013). At this writing, the stress caused by the recent and protracted economic recession also creates an increased need for mental health services (Burgard, 2012).

There is increasing awareness at governmental and scientific levels that widespread mental health problems exact huge social and economic costs to a nation. Further, a growing body of research from the field of Psychoneuroimmunology shows that psychological stressors negatively impact the immune system, general health, and overall functioning. The same body of research also shows that social support and the opportunity to constructively discuss problems can positively affect health, enhance the quality of life, and, by strengthening the immune system, in some instances even promote longevity by reducing the incidence of life-threatening illnesses (Kiecolt-Glaser, 2009). Although Western medicine claims one of its missions is promoting health and preventing the occurrence or the worsening of disease, it often fails to employ what is known about the mind–body connection and its impact on health and well-being.

Further, under managed care, doctors and nurses, traditionally an important source of informal psychological assistance, now have less time to spend with their patients than in the past. The paradoxical result is a public increasingly aware of the connection between stress and illness and desirous of help, while affordable psychological services are less accessible than in the past. Psychologists Alan Kazdin and Sarah Rabbitt (2013) see this lack of services as very detrimental to individuals, families, and even nations. They argue for the necessity of invoking new models of service delivery that are capable of reaching wider swaths of people at lower costs. The models they propose for mental health care delivery derive from practices commonly used today in business, public health, and medicine.

To summarize, four factors stand out as particularly influential in promoting nontraditional psychological treatments: (1) the growing public awareness and desire for help while the availability of affordable professional psychotherapy is decreasing, (2) the value placed on personal entitlement and empowerment in today's society, (3) an ongoing revolution in

IT and its electronic forms of communication, including often via mobile devices, and (4) the rapid dissemination of information by today's highly competitive mass media businesses. Given these forces, it is not surprising that enterprising people continue to seek alternative, accessible, and less costly (although sometimes financially profitable) ways of delivering psychological treatments.

One example of a break from traditional psychotherapy began in the late 1970s, as the number and type of self-help/mutual support groups multiplied rapidly (Goodman and Jacobs, 1994). Some groups had existed for decades, the best known being Alcoholics Anonymous and Recovery, Inc. Suddenly groups sprang up for almost every conceivable personal problem from substance abuse, to parenting concerns; to domestic violence; to transitions such as divorce, retirement, and bereavement; to lifestyle issues like sexual orientation, transvestitism, and open marriage. Groups also came into being for almost every disease listed by the World Health Organization. Millions of people joined them. Although some groups utilize professional leaders or professionally developed ideas, many have little or no professional input. Sociologist Frank Reissman (Gartner and Reissman, 1985) terms the heart of the self-help group movement the 'helper-therapy' principle – the idea that when people suffering from a common problem come together for mutual support, they not only receive help from others, particularly from those more experienced in coping with the problem, but also give help, and in so doing are themselves strengthened and empowered. In the 1980s, many health professionals held negative views about self-help groups, asserting that lay people operating on their own were likely to harm each other by dispensing bad advice. While this remains a concern, the less openly discussed issue was their concerns over loss of professional control and income. In time, professionals have become more supportive of self-help groups as an adjunct to traditional treatment and as a useful, member-governed intervention in its own right. Today, many professionals are directly involved with groups in a variety of roles such as consultants, researchers, leaders, coleaders with group members, and even as participants who share the group's common concern.

The relatively unregulated nature of the various forms of nontraditional psychological treatment offered to the lay public continues to raise concerns about their safety, quality, and effectiveness. Research into the use of paraprofessionals vs professionals, self-help books vs therapist-delivered services, and computer-assisted vs therapist-delivered treatment generally shows these alternatives similar in their effectiveness to professionally delivered psychotherapy (Christensen and Jacobson, 1994; Patel et al., 2011). But, some formats have never been studied. Even among those that have, many important questions remain. For example, are some nontraditional formats more appropriate for certain types of people, or certain types or levels of severity of problem? Further, the relatively unregulated nature of the Internet leaves the door open to poor quality offerings, untested claims, and even outright deception from those designing and offering psychological assistance. In the interest of assuring quality services for the public, priority needs to be given to educating Internet users and service providers, and researching appropriate standards.

See also: Multimodal Therapy in Clinical Psychology; Psychotherapy Integration; Virtual Reality in Psychotherapy.

Bibliography

Andrews, G., Cuijpers, P., Craske, M., McEvoy, P., Titov, N., 2010. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS One 5, e13196.

Burgard, S., 2012. Health, Mental Health and the Great Recession. Stanford Center on Poverty, Stanford, CA. http://www.stanford.edu/group/recessiontrends-dev/cgi-bin/web/sites/all/themes/barron/pdf/Health_fact_sheet.pdf.

Christensen, A., Jacobson, N.S., 1994. Who (or what) can do psychotherapy: the status and challenge of nonprofessional therapies. Psychological Science 5, 8–14.

Cuijpers, P., Donker, T., van Straten, A., Li, J., Andersson, G., 2010. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. Psychological Medicine 40, 1943–1957.

DeAngelis, T., 2012. Second life for practice? Monitor on Psychology 43, 48.

Gartner, A., Reissman, F., 1985. Self-Help Revolution. Human Services Press, New York.

Goodman, G., Jacobs, M.K., 1994. The self-help, mutual support group. In: Fuhriman, A., Burlingame, G.M. (Eds.), Handbook of Group Psychotherapy. Wiley, New York.

Hester, R., Delaney, D., Campbell, W., 2011. ModerateDrinking.com and moderation management: outcomes of a randomized clinical trial with non-dependent problem drinkers. Journal of Consulting and Clinical Psychology 79, 215–224.

Kazdin, A., Rabbitt, S., 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. Clinical Psychological Science 1, 170–191.

Kiecolt-Glaser, J., 2009. Psychoneuroimmunology: psychology's gateway to the biomedical future. Perspectives on Psychological Science 4, 367–369.

Morris, M., Kathawala, Q., Leen, T., Todd, K., Gorenstein, E., Guilak, F., Labhard, M., Deleeuw, W., 2010. Mobile therapy: case study evaluations of a cell phone application for emotional self-awareness. Journal of Medical Internet Research 12, 33–50.

Patel, V., Chowdhary, N., Rahman, A., Verdeli, H., 2011. Improving access to psychological treatments: lessons learned from developing countries. Behaviour Research and Therapy 49, 523–528.

Psychological Treatments, Empirically Supported

William C Sanderson, Hofstra University, Hempstead, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The last decade has witnessed efforts to bring findings from clinical research on psychosocial interventions into everyday clinical practice. These efforts have sought to identify treatments for specific diagnoses that have shown efficacy through empirical study. The identified therapies have been termed empirically supported treatments. In this article, the forces that are driving the move toward empirically supported treatments will be discussed. In addition, the specific criteria used to determine empirically supported treatments will be presented and links to an up-to-date list of these treatments will be provided. Finally, several prominent controversies and criticisms of the move toward empirically supported treatments will be reviewed.

The scientific study of psychological treatments has a long history in clinical psychology. The interest in using scientific evidence to inform clinical practice is best characterized by Paul's (1967) quote: "What treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances?" (p. 111). On a practical level, identifying empirically supported treatments (ESTs) for specific conditions allows service providers to utilize the most effective treatment and as a result, improve outcomes for clients.

Recently, there has been a renewed interest – with urgency – in identifying empirically supported psychological treatments. This is particularly a function of changes in the US health care system, set in motion during the mid-1990s and continuing to evolve, which has increased the need for accountability in the treatments that are provided. Two primary forces are driving these changes: (1) the increasing penetration of managed care and other types of cost-conscious medical insurance plans, and (2) the development and proliferation of clinical-practice guidelines. Essentially, managed care and practice guidelines ask the same question: "Does a treatment work, and if so, is it cost-effective?" Although these forces are influencing every area of health care, the current article will discuss their specific impact on the practice of psychotherapy.

Managed Care

More than any previous event, managed care organizations (or any other system monitoring the utilization and cost of services such as health maintenance organizations, capitated contracts with providers, etc.) are reshaping the practice of psychotherapy (*see* Managed Care and Psychotherapy in the United States). In the traditional fee-for-service model, decisions about the cost and length of treatment were primarily a function of choices made by the doctor and patient. The allocation of resources (i.e., cost of psychotherapy) was of little concern to the clinician. In fact, the fee-for-service model encourages the provision of service, as more service creates more income. However, in response to the increased costs of psychotherapy, and in particular to the perceived 'endless' nature of psychotherapy, managed care organizations are pressuring clinicians to allocate decreasing (less than usual) amounts of service. While to date, the focus of managed care organizations' cost cutting has been almost entirely devoted to limiting the number of sessions a patient receives, ultimately, in order to compete, managed care organizations will have to focus on the quality and efficacy of the psychotherapeutic interventions as well. Managed care organizations are clearly motivated to cut costs. However, they must also satisfy the consumer (i.e., patient), and payer (e.g., employer providing health benefits) and therefore they will have to balance cost-cutting with patient outcome. Simply reducing the length of care will not accomplish this goal, because many, and perhaps the majority, of psychotherapists are trained in 'long-term' therapeutic interventions, which by nature require more sessions than managed care organizations are willing to allocate. Simply reducing the length of therapies devised to take place over a longer period of time is likely to result in ineffective outcomes, leading both to consumer dissatisfaction and increased costs down the road as the severity of the disorder may increase, becoming less responsive to treatment. Thus, concern for the effectiveness of interventions will eventually temper the managed care organizations' focus on economy. Ultimately, managed care organizations will be interested in clinicians providing the "optimal intervention…the least extensive, intensive, intrusive, and costly intervention capable of successfully addressing the presenting problem" (Bennett, 1992: p. 203).

Practice Guidelines

A second force that is having a significant impact on the practice of psychotherapy is the establishment of clinical practice guidelines and treatment consensus statements. As noted by Smith and Hamilton (1994), "Guidelines are now being developed because there is a perception that inappropriate medical care is sometimes provided and that such inappropriate care has both health and economic consequences" (p. 42). It is well known that in all medical fields there is significant variability in treatments delivered to patients with various illnesses. Patients are not necessarily receiving the most effective treatment. Consequently, in order

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21079-9

to ensure that patients uniformly receive the optimal intervention – whether it be a type of medication, surgical procedure, or psychotherapeutic intervention – clinical practice guidelines are being developed. For the most part, the development of practice guidelines relies on systematically evaluating scientific and clinical inputs.

One organization developing practice guidelines is the federal Agency for Healthcare Policy and Research (AHCPR). In order to create impeccable guidelines, only treatments with documented efficacy from randomized, controlled trials are considered. Therefore, for an intervention to be included as a first-line treatment, research evidence attesting its efficacy must be available. Without this evidence, one cannot conclude if a treatment is effective or ineffective, or make judgments about its relative efficacy compared with other available treatments.

The recommendations of the clinical practice guidelines are quite clear. Consider the wording from the AHCPR guideline for depression (Depression Guideline Panel, 1993). When psychotherapy is going to be selected as the sole treatment, the guideline states "Psychotherapy should generally be time-limited, focused on current problems, and aimed at symptom resolution rather than personality change as the initial target. Since it has not been established that all forms of psychotherapy are equally effective in major depressive disorders, if one is chosen as the sole treatment, it should have been studied in randomized, controlled trials" (Depression Guideline Panel, 1993: p. 84). In addition to endorsing specific treatments for depression that has sufficient empirical evidence (e.g., cognitive, behavioral, interpersonal), the report goes on to state "Long-term therapies are not currently indicated as first-line acute phase treatments" (p. 84).

Along the same lines as treatment guidelines, consensus statements are likely to carry significant weight in determining the efficacy of treatments. For example, in 1991 a consensus development conference on the 'Treatment of Panic Disorder' was held, sponsored by the National Institute of Mental Health. As was true in the process of developing the AHCPR treatment guidelines, the merits of various treatments was determined by the available scientific evidence. The format of the conference was to let 'each camp put its best data forward' and the panelists would evaluate and weigh the evidence to formulate a consensus statement (Wolfe and Maser, 1994). As in the case of treatment guidelines, specific treatments were judged to be effective for panic disorder including several pharmacological compounds and cognitive behavioral psychotherapy (cf Panic Consensus Statement published in Wolfe and Maser, 1994). Once again, as is true of treatment guidelines, the wording of the Panic Consensus Statement on the use of treatments not supported by empirical evidence, whether they be psychological or pharmacological, is quite clear, "One risk of maintaining individuals in nonvalidated treatments of panic disorder is that misplaced confidence in the therapy's potential effectiveness may preclude application of more effective treatment." The statement spells out a specific concern about the use of psychotherapies without demonstrated effectiveness for panic disorder, "The nature of the therapeutic relationship makes it difficult for the patient to seek additional or alternate treatment."

As clinical guidelines and treatment consensus statements continue to emerge for a wide array of emotional disorders, they will have further impact on the way clinicians practice psychotherapy. In effect, these documents set standards of care, which if ignored, leave the clinician both ethically and legally vulnerable. Health insurance companies and managed care plans are now providing their practitioners with copies of these guidelines and asking that they be followed. The implications are clear: failure to provide an EST from these guidelines, where one exists, may constitute malpractice in the eyes of the payer. For example, a number of states have passed, or are in the process of passing, legislation that gives these guidelines the 'force of law' by protecting clinicians who follow the guidelines from malpractice litigation (Barlow and Barlow, 1995).

Identifying Empirically Supported Treatments

Recognizing the lack of utilization of ESTs and considering the trends noted above, the Society of Clinical Psychology (Division 12 of the American Psychological Association) established a task force to identify, promote, and assist in the dissemination of ESTs. The original Task Force on Promotion and Dissemination of Psychological Procedures (presently named the Committee on Science and Practice; CSP) was created by David H. Barlow, then President of the Society. The CSP is now a standing committee in the Society of Clinical Psychology, and the work will continue into the future (all reports from this committee are available at http://www.div12.org/psychological-treatments/).

The primary mission of the original task force was to define criteria to determine whether treatments qualified as ESTs. The following criteria were established (Chambless et al., 1998). For treatments to be considered by the task force, research must have been conducted with treatment manuals. This requirement was set because treatment manuals make for better research design, and the use of a treatment manual allows readers to know what specific procedures have been supported. For example, although many interventions fall under the rubric of behavior therapy (e.g., exposure, relaxation training, cognitive restructuring), only exposure and response prevention have been demonstrated as efficacious for obsessive-compulsive disorder. Stating that behavior therapy is effective for obsessive-compulsive disorder does not allow the reader to know what specific strategies were tested and how they were applied. Use of the treatment manual solves this dilemma.

As for whether or not the treatment qualifies as an EST, the task force distinguishes between two levels of EST – 'well established' and 'probably efficacious.' EST status is determined by the results from group design and single-case design studies. With regard to group design studies, to qualify as a well established treatment, at least two studies conducted by different investigators must be presented as supporting studies. The following criteria must be satisfied for qualification as a supporting study: (1a) treatment is superior to pill placebo or psychological placebo, or (1b) treatment is equivalent to an already established treatment; (2) a treatment manual was used in the study; (3) characteristics of the client sample must be

clearly specified. A treatment may not meet the criteria for being 'well established' but may still satisfy the less conservative criteria for a 'probably efficacious' treatment if (1) two studies show that the treatment is more effective than a waiting list control group, or (2) only one supporting study exists, or if more than one exists, all were conducted by the same investigator. (For details concerning the definition of EST, including use of single subject designs and resolution of conflicting data, the reader is directed to the task force report itself located at http://www.div12.org/psychological-treatments/.)

Beginning in 2007 an online version of the list of ESTs was developed, and since it is updated regularly, it represents the latest list ESTs: http://www.div12.org/PsychologicalTreatments/index.html. (It is important to note that it is not assumed that treatments without empirical evidence supporting them are ineffective, but instead, that in the absence of scientific data, no conclusion has been drawn about their efficacy or lack thereof.)

The movement toward ESTs has not been welcomed by everyone within our profession, and indeed, several authors have raised important concerns and criticisms (e.g., Garfield, 1998; Kovacs, 1995; Nathan, 1998) regarding the identification of 'validated' therapies. Although space does not permit a discussion of the specific issues here, in fact, the CSP agrees that there are difficulties and risks in establishing so-called psychotherapy guidelines – or any treatment guidelines for that matter (cf Pincus 1994 concerning the development of psychiatry guidelines) – yet the CSP and other professional organizations believe that the benefits clearly outweigh the risks (Pincus, 1994; Sanderson, 1995), and thus, will continue the work outlined above (cf Weisz et al., 2000 for a discussion of the future directions of the CSP).

Why Is There Controversy over ESTs?

In principle, most psychologists are likely to agree on the necessity of providing empirical support for the treatment interventions they use. It is also reasonable to assume that consumers expect to receive a mental health treatment whose effectiveness has been proven, when such a treatment is available. Therefore, one would expect clinicians to embrace ESTs. However, psychologists are split on the evidence-based practice (EBP) movement and it has been a hotly debated issue in the literature as well as among practicing psychologists today (e.g., Garfield, 1998; Kovacs, 1995; Nathan, 1998). The controversies over the use of ESTs are briefly summarized below:

First, many clinicians claim that EBPs essentially favor cognitive-behavioral therapies because these represent standardized treatments that can be tested more readily in controlled clinical trials. As a result, they believe that viable alternative treatments (e.g., psychodynamic therapy), which cannot be as easily standardized, will fail to measure up to the EBP criteria, even though they may be effective in clinical practice. Secondly, it has been argued that most studies supportive of EBPs fail to capture the complexity and uniqueness of patients seen in 'actual' practice. Consequently, these clinicians question the applicability of these manualized treatments to patients seen in the community. Along the same lines, thirdly, while ESTs clearly work in the controlled research environment, data are lacking on how well they generalize to real-world settings such as hospitals, community mental health clinics, private practice offices, etc. Since these settings usually differ substantially from clinical research centers (with regard to therapist skill and patient characteristics), many argue that one cannot assume ESTs generalize to real-world clinical settings. Finally, many clinicians have a negative reaction to being 'told what to do' by researchers who generate these data. This is part of an ongoing tension between scientists and practitioners as to who 'knows best' about treatment decisions – those who generate the research findings or those who provide clinical treatment in actual practice.

Of course, many of these questions can be answered by empirical research, which must ultimately provide the answers in order to allow clinicians in the field to fully embrace ESTs. Although space does not permit a discussion of the specific issues here, in fact, there are difficulties and risks in establishing so-called 'psychotherapy guidelines' – or any treatment guidelines for that matter (cf Pincus, 1994 concerning the development of psychiatry guidelines). However, many believe that the benefits clearly outweigh the risks (e.g., Pincus, 1994; Sanderson, 1995).

See also: Evidence-Based Psychotherapy; Manual Based Treatments; Motivation in Clinical Interventions; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Trials.

Bibliography

American Psychiatric Association, 1993. Practice guidelines for major depressive disorder in adults. American Journal of Psychiatry 150 (4), 1–26 (Suppl. 5).

Barlow, D.H., 1994. Psychological interventions in the era of managed care. Clinical Psychology: Science and Practice 1, 109–122.

Barlow, D.H., Barlow, D.G., 1995. Guidelines and empirically validated treatments: ships passing in the night? Behavioral Healthcare Tomorrow 4, 25–29.

Bennett, M.J., 1992. The managed care setting as a framework for clinical practice. In: Feldman, J.L., Fitzpatrick, R.J. (Eds.), Managed Mental Healthcare: Administrative and Clinical Issues. American Psychiatric Press, Washington, DC, pp. 203–218.

Chambless, D.L., Sanderson, W.C., Shoham, V., Johnson, S.B., Pope, K., Crits-Christoph, P., Baker, M., Johnson, B., Woody, S.R., Sue, S., Beutler, L., Williams, D., McCurry, S., 1996. Empirically validated therapies: a project of the division of clinical psychology, American Psychological Association, task force on psychological interventions. The Clinical Psychologist 49, 5–15.

Chambless, D.L., Baker, M.J., Baucom, D., Beutler, L.E., Calhoun, K.S., Crits-Christoph, P., Daiuto, A., DeRubeis, R., Detweiler, J., Haaga, D.A.F., Johnson, S.B., McCurry, S., Mueser, K.T., Pope, K.S., Sanderson, W.C., Shoham, V., Stickle, T., Williams, D.A., Woody, S.R., 1998. Update on empirically validated therapies: II. The Clinical Psychologist 51, 3–16.

Depression Guideline Panel, 1993. Depression in Primary Care: Volume 2. Treatment of Major Depression. Clinical Practice Guideline, Number 5. Washington, DC. US Government Printing Office, Rockville, MD (Department of Health and Human Services, Public Health Service, Agency for Healthcare Policy and Research. AHCPR Publication No. 93-0551).

Garfield, S., 1998. Some comments on empirically supported treatments. Journal of Consulting and Clinical Psychology 66, 121–125.

Kovacs, A., 1995. We have met the enemy and he is us! The Independent Practitioner 15 (3), 135–137.

Nathan, P.E., 1998. Practice guidelines: not yet ideal. American Psychologist 53, 290–299.

Paul, G.L., 1967. Strategy of outcome research in psychotherapy. Journal of Consulting Psychology 31, 109–118.

Pincus, H.A., 1994. Treatment guidelines: risks are outweighed by the benefits. Behavioral Healthcare Tomorrow 4, 40–45.

Sanderson, W.C., 1995. Can psychological interventions meet the new demands of health care? American Journal of Managed Care 1, 93–98.

Sanderson, W.C., Woody, S., 1995. Manuals for empirically validated treatments. A project of the division of clinical psychology, American Psychological Association, task force on psychological interventions. The Clinical Psychologist 48, 7–11.

Smith, G.R., Hamilton, G.E., 1994. Treatment guidelines: provider involvement is critical. Behavioral Healthcare Tomorrow 4, 40–45.

Task force on promotion and dissemination of psychological procedures, 1995. Training in and dissemination of empirically-validated psychological treatments. The Clinical Psychologist 48, 3–23.

Weisz, J.R., Hawley, K.M., Pilkonis, P.A., Woody, S.R., Follette, W.C., 2000. Stressing the other three Rs in search of empirically-supported treatments: review procedures, research quality, relevance to practice and public interest. Clinical Psychology: Science and Practice 7, 243–258.

Wolfe, B.E., Maser, J.D. (Eds.), 1994. Treatment of Panic Disorder: A Consensus Statement. American Psychiatric Association Press, Washington, DC.

Psychological Treatments: Randomized Controlled Trials

William S Agras, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The evolution of controlled clinical trials for psychological treatments began in the 1930s and became more frequent with the development of behavior therapy following Eysenck's critique of the effectiveness of psychological therapies in 1966. Now there are a number of different experimental designs suitable for the investigation of different questions posed by investigators. Multisite clinical trials are useful because they can recruit larger number of participants with a particular problem in a shorter time than can be accomplished at one site and are more likely to lead to policy change. Issues such as mediators and moderators of outcome, cost-effectiveness, and future developments are also highlighted.

The aim in developing a psychological therapy is to produce an effective treatment that can be used by therapists in many different clinical settings (*see* Evidence-Based Psychotherapy). The development of a psychological treatment begins with observations flowing from the treatment of patients based on a theoretical model of behavior change. The process continues with a series of cases in which the treatment is refined and its initial effectiveness confirmed. At this point, a therapist manual, a detailed account of the therapeutic procedures, is developed enabling training of therapists and allowing others to replicate the treatment. The treatment is then tested in randomized controlled clinical trials. If these trials continue to demonstrate effectiveness with replication by other researchers, a new evidence-based psychotherapy will have been developed. All too often the process ends at this stage with the result that few practitioners actually use the treatment.

The next stage in the development of a psychotherapy is to disseminate the results through publications and workshops. Unfortunately, this rarely results in more widespread use of the treatment, because research has shown that such dissemination does not increase the use of a novel psychotherapy. Hence, it is important to implement the therapy in settings in which it will be used (*see* Implementation Science). Before implementation certain questions need to be considered. First, in what setting or settings are the treatment needed? Second, what is the nature of those settings? Third, are the therapists in those settings interested in using a new evidence-based psychotherapy? Hence, the researcher needs to understand the nature of these settings in which psychotherapy is to be implemented and the needs of therapists in those settings. Moreover, it is important to engage psychotherapists in those settings because they may provide valuable input into changes in the process of therapy that would need to be incorporated into the adapted treatment to fit the needs of their practice. The next step is to adapt the newly developed treatment to fit those settings. For example, if the new psychotherapy was developed as a 20-session treatment but most clinical settings in which it will be used do not deliver more than 8 sessions of psychotherapy to their clients, the original treatment will need to be shortened, adapted, and tested in further randomized controlled trials in these clinical settings to determine if it is efficacious. Unfortunately, this 'ideal process' takes several years, and steps are often skipped or assumed to be unnecessary resulting in the situation commented upon by the Director of the National Institute of Mental Health (Insel, 2009) "We have powerful, empirically-supported psychosocial interventions, but they are not widely available … A serious deficit exists in training for empirically supported psychosocial interventions … Translational research will need to focus not only on 'bench to bedside' but also on 'bedside to practice.'"

The Development of Controlled Clinical Trials in Psychology

Parallel developments in two clinical disciplines, medicine and psychology, beginning in the 1930s, edged both fields toward the use of controlled trials. In an early study spanning both, medicine and psychology (Chappell and Stephenson, 1936), the authors demonstrated that an educational group program resulted in greater improvement and less relapse in a group of peptic ulcer patients than the control group, both groups having received the same medical care. Another study, the Cambridge–Somerville Youth Project, begun in the late 1930s and continuing with a long-term follow-up into the 1950s, was one of the earliest trials of a psychological treatment (Teuber and Powers, 1953). This large-scale, well-designed study of 650 boys thought to be at high risk of developing delinquency was aimed at evaluating a prevention program. The outcome, as these individuals were followed over time, demonstrated that the treatment group fared less well than the control group for a range of delinquent behaviors. Such negative results, while disappointing, are important because they allow ineffective approaches to treatment to be abandoned.

These early controlled trials were the exception rather than the rule. Most studies of the effectiveness of psychological therapies consisted of treating a series of patients without a control group. Hence, it was impossible to be sure whether the results were due to spontaneous recovery or due to the treatment. Two developments catalyzed the use of the controlled trial. Hans Eysenck, Professor of Psychology at London University, reviewed the psychotherapy treatment literature in 1952 and again in 1966 (see Eysenck, 1952, 1966). On both occasions, Eysenck concluded that there was little evidence that psychotherapy was effective, that is, the results did not exceed the expected rate of spontaneous recovery.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21080-5

While this opinion resulted in a vigorous debate and defense of psychotherapy, including the addition of a few treatment trials that demonstrated some success for various forms of psychotherapy that Eysenck had not considered, Eysenck's conclusions were widely accepted, reaffirming the need for well-conducted controlled trials. Some, however, argued that the effects of psychological therapies, because they were aimed at 'inner' essentially unmeasurable changes, could never be adequately assessed in a controlled trial. This problem was largely overcome by the next development.

The second, and the most important, catalyst for the use of controlled trials of psychological treatments was the emergence of behavior therapy (*see* Behavior Analysis: Applied; Psychology, History of (Twentieth Century)). Eysenck (1966: p. 39) had concluded, "With the single exception of the psychotherapeutic methods based on learning theory, results of published research … suggest that the therapeutic effects of psychotherapy are small or non-existent …." Because behavior therapy was based on learning theory, the focus of treatment was directly on the problem behaviors comprising the clinical syndrome. This allowed for the development of objective measures of treatment outcome (*see* Behavioral Assessment for a complete discussion of objective evaluation methodology). In addition, because treatment was scientifically based, its practitioners were more likely to use scientific methods, including the randomized controlled trial, to assess the effectiveness of these new treatments.

Types of Controlled Trials

The most commonly used clinical trial design involves randomization of participants to two or more groups receiving different treatments. For randomization to be effective it is important that the treatment conditions are reasonably equivalent (*see* Placebo-Controlled Trials, Ethics of). For example, randomizing individuals to either a treatment or no treatment is not equivalent probably leading participants randomized to the no-treatment condition to drop out of the study more frequently than those in the active treatment, thus biasing the trial results. This situation can be addressed by delaying treatment for the no-treatment group. This design is referred to as a waiting-list control. However, this design only tests whether the new treatment is better than no treatment, a comparison that is of little scientific or practical use. Hence, a comparison with another form of treatment is preferable. Usually the number and duration of the treatment sessions will be the same in both groups controlling for many 'shared elements' among treatments and allowing differences in outcome to be assigned to operational differences between treatments. Controlled treatment trials should have a primary outcome to assess the differences between treatments, an outcome that measures as objectively as possible the primary behavior or symptom that the treatment is aimed at. For example, for panic disorder, the number and severity of panic attacks would comprise the main outcome; for depression, a measure of the severity of depression; for bulimia nervosa, measures of binge eating and purging. Because these measures test the primary hypothesis, that is, the null effect of the treatment in question as compared with a control group, the study should stand or fall depending on whether the primary outcome is positive or not. A clinical outcome is also often reported. For example, for panic disorder the number of participants in each group who no longer experience panic attacks would be such a measure. All controlled clinical trials, even or perhaps especially when outcomes are equivalent, should perform a moderator analysis that may throw light on subgroups of patients with better recovery in one or other of the treatments (see Section Predictors and Moderators of Outcomes). When the primary outcome is positive then secondary measures – usually associated psychopathology – are also examined for the effect of treatment. Because some symptoms such as mild anxiety or depression are normative in a community, the extent to which the treated population now approximates community norms may also be reported.

Complex Clinical Trial Designs

Because most treatments improve only a fraction of those with the disorder, a stepped-care design may be used to test the effects of adding a second treatment for those who do not respond to the primary treatment. There are three approaches to this design. First, the original treatment may be compared to an additive treatment (original treatment plus secondary treatment), thus allowing an assessment of the effects of adding a second treatment to the original treatment. Second, in what is known as an adaptive design (Nahum-Shani et al., 2012), the study may begin by treating everyone with the standard treatment and then randomizing treatment failures to either a second treatment or continuation with the standard treatment. This design requires a standardized measure of success or failure at the end of treatment and has the advantage of more closely following clinical decision making. The third option, again an adaptive design, depends on finding an indicator that predicts success or failure reasonably, accurately, and early in the treatment process. Many treatments are associated with a rapid response to treatment in a proportion of individuals often measured as a percentage decrease in the primary outcome early in treatment. For example, for bulimia nervosa, a 70% reduction of purging rates by week 6 of treatment has been shown to predict the treatment success or failure. Probable failures can then be randomly allocated to either the standard treatment or the new treatment. A variant of this design is to randomize participants into two groups and use the early marker in only one group to treat probable failures with a second treatment. Because only a fraction of patients fail treatment, these designs require a large sample size and may require a multisite study. However, such trials may lead to the development of treatment algorithms useful to clinicians.

Multisite Studies

Multisite studies provide more definitive estimates of effectiveness by increasing sample size and implementing the treatment in different sites that may have unique characteristics in terms of patient population, therapist characteristics, and so on. Hence, generalizability of the results of treatment to the clinical world is greater than in a single-site study. Another reason to conduct a multisite study is when treatment for

a relatively rare disorder is being studied. If the study were to be done at only one site it might take many years before sufficient participants were entered to the study. A multisite study provides more rapid results. Another reason for considering a multisite study is when the question of interest shifts to what treatment or treatments might benefit those who do not respond to an initial treatment as discussed earlier. Because only treatment failures will be entered into such a study, a large number of individuals must be treated in order to generate a sufficiently large sample for the study.

Multisite studies are necessarily more complex to carry out than a study at a single site. They are also more expensive than single-site studies and for that reason should be considered only when a reasonable number of single-site studies have suggested that a treatment is effective. The investigators involved in a multisite trial must agree to a common protocol. The treatment(s) being studied must be well documented in a treatment manual and therapists at each of the sites be trained in the delivery of the treatments. Moreover, the therapists should be carefully supervised throughout the treatment phase of the study to ensure that they are delivering the treatment(s) in accord with the protocol. A sample of tape recordings of treatment sessions should also be audited to ensure fidelity to the treatment. In addition, the assessments used must be standardized and performed comparably between sites. This may mean that assessors will also need to be trained and supervised across sites in the particular assessment methods. Because several sites are involved, data are usually collected by an independent data center responsible for collection and analysis of the data. The data center may also be responsible for the trial coordination.

Noninferiority Designs

Rather than hypothesizing that a novel treatment will be more effective than a standard treatment, an investigator may pose the question whether a novel treatment is similarly effective to a standard treatment. For example, there may be reason to think that a shorter, less expensive treatment may be similar in effectiveness compared to a longer, more expensive standard treatment or that a new medication will be similar to the standard effective medication. In this case, a noninferiority design may be called for. Sample size will vary according to the margin of difference that is set between the two treatments. If the new treatment's effect is within a small margin then it is very similar in effectiveness to the standard treatment. However, setting a small margin will necessitate a larger sample size. If the new treatment has not been tested against a placebo then a 3-group design including a placebo group may be required. With this design similarity to both the standard treatment and placebo can be determined.

Therapy Process

Although assessing outcome is the major aim of a controlled clinical trial, consideration of the processes underlying the outcome is also important to ensure that the results of the trial are valid (*see* Psychotherapy Process Research). Among the factors to be considered are following: factors common to all treatments such as the participant's expectancy of benefit, the therapeutic relationship, the fidelity of treatment delivery, therapist competence, and the adherence of participants to the key elements of treatment. Differences between treatments in any of these factors will threaten the conclusion that treatment differences are due to the specific elements that make up the treatment package.

Common Factors

Treatments can be regarded as a package containing both specific procedures aimed at a particular behavior or disorder, together with elements common to all therapies such as the amount of therapist attention, and therapeutic instructions that include the rationale for therapy together with engendering an expectancy that treatment is likely to be successful. Research has demonstrated that nonspecific elements lead to significant behavior change. Hence, it is important to know that to what extent procedures specific to the therapy contribute to a successful outcome over and above nonspecific elements. This raises a particular problem for psychological therapies because there is no placebo pill. This problem can be solved in several ways. The first solution is to use a nonspecific therapy, that is, one without the specific therapeutic procedures being examined and containing all the nonspecific procedures, but which also has face validity as a therapy with which to compare the active treatment. It is necessary to assess the participant's expectancy of a successful outcome to ensure equality between treatments, an assessment that should be made in all comparative treatment studies. Another design is to compare the new therapy with a standard therapy for the condition that has already been demonstrated more effective than a nonspecific therapy. This is the preferred design when a standard therapy exists.

Treatment Manuals

The treatment manual, a detailed account of the therapeutic procedures, was developed to facilitate the training of therapists and to allow others to replicate the treatment in further controlled trials and eventually in clinical practice. Such manuals are now regarded as an essential element of a controlled trial of psychological therapies. Therapists are trained to utilize the therapeutic procedures often in a didactic workshop in which the therapy is discussed in detail followed by therapists practicing the treatment on several patients before the controlled trial begins. During this period, therapists are supervised by an expert in the treatment and audio or videotapes of their sessions are audited and feedback is provided to the therapist. Often, a second, more advanced workshop is held when therapists have treated at least one patient. When therapists attain suitable fidelity to the treatment procedures, they are certified to begin treating participants within the clinical trial. Supervision of therapists continues throughout the treatment phase of the trial. There is some controversy regarding the use of treatment manuals. Some argue that manuals lock the therapist into a rigid method of providing treatment, thus eliminating the opportunity to individualize the treatment. Others suggest that manuals provide essential guidelines for therapists and that the experienced therapist can individualize treatment to the extent

necessary. There are relatively little experimental data on the effectiveness of the use of manuals in the clinic; however, existing data suggest that manuals are likely to enhance therapy probably by ensuring that the therapist stays on track during the treatment process. *See* Manual Based Treatments, for a complete discussion of this important issue.

Treatment Fidelity and Therapist Competence

Once the treatment phase of a clinical trial begins, therapy sessions are audio- or videotaped and a random sample of such tapes is audited using a checklist to ensure that all the elements of treatment are being provided, and that no foreign elements have crept in. Because the relation between fidelity and treatment outcome is at best weak, therapist competence that better relates to outcomes is often assessed. Hence, the auditor will determine if a specific element of therapy is present and also how competently that element was delivered.

Outcomes of Controlled Trials

There are several potential outcomes from a randomized clinical trial. First and foremost is whether the treatment is as effective as the comparison group; second, whether there are subgroups of patients who do best with the treatment under study; and third, whether variables that mediate the outcome of treatment can be identified. Such mediators may throw light on how a given therapy works.

Predictors and Moderators of Outcomes

It is usual in a controlled trial to assess whether baseline variables affect the outcome of treatment. Examples of potential predictors include the following: the duration and severity of the problem, the age and social class of the patient, and associated psychopathology. Unfortunately, such predictors are sufficient for a general description of the patient most likely to improve, but not specific enough to identify such patients before treatment begins. A more useful approach is to identify moderators of treatment outcome, an analysis that should be performed in all randomized clinical trials with two treatment groups. A moderator is a baseline variable that interacts with outcome, thus identifying groups of individuals with particular characteristics with better or worse outcomes. Because such individuals can usually be identified by a cut point on a questionnaire or a particular characteristic such as age, the identification of moderators is a clinically more useful predictor of outcome affording improved guidance to clinicians as to what therapy may work best for which patient.

Treatment Mediators

Mediators are either partially or fully responsible for an effective treatment outcome. They consist of a therapeutic procedure that occurs prior to the benefits of treatment and that significantly mediates outcome. Rather than measuring such variables before treatment begins, as is the case with moderators, potential mediators are measured early in the course of treatment usually on multiple occasions. Potential mediators are based on hypotheses about how the treatment works. The identification of mediators can throw important light on the mechanisms underlying the effectiveness of a given psychological treatment and may provide insight as to how to improve the treatment. See also Psychotherapy Process Research in which interactions between therapist and patient, and their effect on outcome, form the question of interest.

Cost-Effectiveness

The costs of treatment may inform which of two similarly effective treatments to use as a first-level treatment. There are two forms of cost: direct and indirect costs of the psychotherapy itself together with costs of ancillary treatments required during the treatment phase of the trial, e.g., hospitalization, medications, etc. These costs are usually fairly easy to acquire and are, therefore, most often reported. In addition, social costs such as employment of the patient and other family members, school attendance, home nursing, and visits to physicians or emergency rooms not related to the condition being studied may be monitored. These costs are more difficult to acquire because they need to be frequently monitored and may require a family member to report them in addition to the patient. Treatment costs are usually reported as cost per patient treated or cost per recovered patient.

Are Controlled Psychological Treatment Trials Useful?

There is still some controversy as to whether clinical trials for psychological treatments can reliably inform clinical work and social policy (Crits-Christoph et al., 2005; Weston et al., 2004, 2005). Criticism of evidence-based treatments includes arguments that such treatments have only been tested in settings that do not represent the 'real world.' To some extent this critique is true and more research needs to be performed in such settings. A further criticism is that clinical trials do not address the types of patients seen in community practice. This critique reveals a dilemma facing the clinical researcher. To reduce variability, narrowly selected clinical populations are usually entered into clinical trials. However, such selection methods exclude large numbers of treatment-seeking individuals. The converse approach would be to study treatments for a wide range of individuals with a particular disorder, hence excluding fewer individuals and better testing the limits of a treatment. Both approaches seem reasonable and the latter approach should be recognized as an important area for future research. Despite such controversy, the controlled trial has become the accepted method of evaluating the effectiveness of psychotherapy. Most individuals would prefer that their medical treatments were informed by controlled trials; similarly, consumers of psychotherapy should demand that these procedures are well tested.

Future Developments

The use of randomized controlled trials to evaluate the effectiveness of psychological treatments will undoubtedly

continue. A relatively large number of treatment procedures have been identified as effective for specific psychological problems (*see* Psychological Treatment, Effectiveness of and Psychological Treatments, Empirically Supported for a detailed review of such treatments). Many, but not all, such therapies fall under the rubric of behavior therapy. These research efforts have now generated a growing number of multisite trials in which the treatment is put to a more definitive test, compared with another effective psychotherapy, or with medication either separately or in combination with psychotherapy. As is the case in medical research, it would be expected that multisite trials will be increasingly used in psychotherapy research to obtain more definitive information concerning treatment effectiveness and beginning to answer the question: 'What psychotherapy is best for which person?'

In addition, clinical trials will be increasingly used to answer other questions important to psychotherapy research. For example, because moderator analyses are essentially exploratory in nature, the next step would be to test the effectiveness of using moderators as predictors of outcome in controlled trials to ascertain whether the information provided by a moderator enhances outcome. Another arena for controlled trials relevant to psychotherapy research is to compare different methods of training psychotherapists to accurately and effectively use particular psychotherapies. Such work will also lead to the development and testing of different methods to implement evidence-based psychotherapies in community settings. Different methods of delivering psychotherapeutic treatments are also being developed and such methods will need to be compared to standard treatments in randomized controlled trials. For example, treatments can be delivered to areas in which there are few therapists by means of videoconferencing. In addition, the Internet is now being used with increasing frequency to deliver interactive treatment packages. Internet trials where participants are not seen face-to-face pose a problem in terms of assessments that go beyond the use of questionnaires such as interviews. It could be argued that having in-person assessments would alter the participant pool. Potential participants who did not want to interact with anyone about their problem would be less likely to enroll in such studies, hence not representing the potential Internet pool of participants accurately. Nonetheless, optimal assessment methods must eventually be used to evaluate Internet-based treatments.

Finally, clinical trials will be used with increasing frequency to elucidate biologic mechanisms that are altered by psychotherapy and are associated with outcomes.

It may also be that the development of psychotherapies will ultimately be changed to better meet the needs of community therapists. At present, psychotherapies are designed and tested in somewhat rarified venues and community therapists are then expected to adopt them whether or not they meet the needs or practice of such therapists, particularly in clinics treating large numbers of patients. The development of adaptions of existing psychotherapies that better fit community practices should probably become a priority in the future development of psychotherapies. This will entail working more closely with therapists practicing in the community and developing and testing in controlled trials variants of psychotherapies adapted to particular settings (*see* Evidence-Based Psychotherapy).

See also: Cluster Randomized Trials, Ethical Implications of; Evidence-Based Psychotherapy; Implementation Science; Manual Based Treatments; Placebo-Controlled Trials, Ethics of; Psychological Treatment, Effectiveness of; Psychological Treatments Delivered Nontraditionally; Psychological Treatments, Empirically Supported; Psychology, History of (Twentieth Century); Psychotherapy Process Research.

Bibliography

Agras, W.S., et al., 2000. Outcome predictors for the cognitive behavior treatment of bulimia nervosa: data from a multisite study. American Journal of Psychiatry 157, 1302–1308.

Chappell, M.N., Eysenck, H.J., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting and Clinical Psychology 16, 319–324.

Chappell, M.N., Stephenson, I.T., 1936. Group psychological training in some organic conditions. Mental Hygiene 20, 588–597.

Crits-Christoph, P., Wilson, G.T., Hollon, S.D., 2005. Empirically Supported Psychotherapies: Comment on Westen, Novotny, and Thompson-Brenner (2004).

Eysenck, H.J., 1952. The effects of psychotherapy: an evaluation. Journal of Consulting and Clinical Psychology 16, 319–324.

Eysenck, H.J., 1966. The Effects of Psychotherapy. International Science Press, New York.

Goldfried, M.R., Wolfe, B.E., 1998. Toward a more clinically valid approach to therapy research. Journal of Consulting and Clinical Psychology 66, 143–150.

Insel, T.R., 2009. Translating scientific opportunity into public health impact. A strategic plan for research on mental illness. Archives of General Psychiatry 66, 128–133.

Kazdin, A.E., 2011. Evidence-based treatment research: advances, limitations, and next steps. American Psychologist 66, 685–698.

Kraemer, H.C., et al., 2002. Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry 59, 877–883.

Langer, D.A., et al., 2011. Do treatment manuals undermine youth-therapist alliance in community clinical practice? Journal of Consulting and Clinical Psychology 79, 427–432.

McHugh, R.K., Barlow, D.H., 2010. The dissemination and implementation of evidence-based psychological treatments. A review of current efforts. American Psychologist 65, 73–84.

Nahum-Shani, I., et al., 2012. Experimental design and primary data analysis methods for comparing adaptive interventions. Psychological Methods 17, 457–477.

Persons, J.B., Silberschatz, G., 1998. Are results of randomized clinical trials useful to psychotherapists? Journal of Consulting and Clinical Psychology 66, 126–135.

Russell, R.L., Orlinsky, D.E., 1996. Psychotherapy research in historical perspective: implications for mental health care policy. Archives of General Psychiatry 53, 708–715.

Safer, D.L., Hugo, E.M., 2006. Designing a control for behavioral group therapy. Behavior Therapy 37, 120–130.

Schoenwald, S.K., Garland, A.F., 2013. A review of treatment adherence methods. Psychological Assessment 25, 146–156.

Stephenson, I.T., 1936. Group psychological training in some organic conditions. Mental Hygiene 20, 588–597.

Strunk, D.R., et al., 2010. Therapist competence in cognitive therapy for depression: predicting subsequent symptom change. Journal of Consulting and Clinical Psychology 78, 429–437.

Teuber, N.L., Powers, E., 1953. Evaluating therapy in a delinquency prevention program. Proceedings of the Association for Research in Nervous & Mental Disease 31, 138–147.

Weston, D., Novotny, C.M., Thompson-Brenner, H., 2004. The empirical status of empirically supported psychotherapies: assumptions, findings, and reporting in controlled clinical trials. Psychological Bulletin 130, 631–663.

Weston, D., Novotny, C.M., Thompson-Brenner, H., 2005. EBP and EST: reply to Crits-Christoph et al. (2005) and Weisz et al. (2005). Psychological Bulletin 131, 427–433.

Wilson, G.T., 1998. The clinical utility of randomized controlled trials. International Journal of Eating Disorders 24, 13–29.

Psychologism, History of

Martin Kusch, University of Vienna, Wien, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of psychologism is the history of an accusation and of a threat. To label authors as advocates of psychologism is to claim that they use psychological concepts and ideas in the wrong place. The accusation first became popular in the German-speaking world with the rise of experimental psychology in the second half of the nineteenth century. The most famous episode in the history of the accusation is the 'psychologism controversy' in the German-speaking world at the beginning of the twentieth century. Frege and Husserl were among the leading critics of psychologism in logic, mathematics, and epistemology. The German debate over psychologism influenced subsequent philosophical thinking in most Western countries.

The history of psychologism is the history of an accusation and of a threat. To label an author an advocate of psychologism is to claim that he or she uses psychological concepts and ideas in the wrong place. The accusation first became popular in the German-speaking world with the rise of experimental psychology in the second half of the nineteenth century. The most famous episode in the history of the accusation is the German 'psychologism controversy' at the beginning of the twentieth century. It will be the central topic of this article. It is widely accepted that this controversy played an important role in shaping the character of Western philosophy in general and analytic philosophy in particular. It also influenced subsequent debates over 'sociologism.'

The Psychologism Controversy in Early-Twentieth-Century Germany

From the 1880s onward, a number of German logicians and epistemologists started to attack some of their philosophical colleagues for presenting and endorsing the following kinds of 'psychologistic' arguments:

Logic studies specific 'laws of thought.' Psychology is the study of all kinds of laws of thought. Therefore, logic is a part of psychology.

Normative disciplines must be based upon descriptive-explanatory disciplines. Logic is a normative discipline with respect to human thinking. Psychology is the descriptive-explanatory discipline with respect to human thinking. Therefore, logic must be based upon psychology.

Logic is concerned with judgments, concepts, and inferences. Judgments, concepts, and inferences are human mental entities. Therefore, logic is a theory about human mental entities.

The touchstone of logical truth is self-evidence. Self-evidence is a human mental experience. Therefore, logic is about a human mental experience (and thus a part of psychology).

We cannot conceive of a radically different logic. The limits of conceivability are our mental limits. Therefore, logic is relative to the human species, and thus a subfield of human psychology.

In some cases, there was considerable dispute concerning the question who actually held these views. The first argument can be found in Theodor Lipps and Gerardus Heymans, the second in Wilhelm Wundt, the third in Wilhelm Jerusalem and Christoph Sigwart, the fourth in Theodor Elsenhans, and the fifth in Benno Erdmann. All the accused had published widely on psychological issues, and some of them openly advocated treating psychology as the foundation of philosophy.

Arguments against these and other allegedly psychologistic authors were provided by many writers, among them the logician and mathematician Gottlob Frege, the phenomenologist Edmund Husserl, and the Neo-Kantians Hermann Cohen, Paul Natorp, Heinrich Rickert, and Wilhelm Windelband. Of these and the many other critics, Husserl was undoubtedly the most successful. Husserl's arguments appeared in the Prolegomena to his *Logische Untersuchungen*. The following arguments figured centrally in the debate:

If logical rules were based upon psychological laws, then all logical rules would have to be as vague as the underlying psychological laws. But, not all logical rules are vague. Therefore, not all logical rules are based upon psychological laws.

If laws of logic were psychological laws, then they could not be known a priori. They would be more or less probable rather than certain, and justified only by reference to experience. But, laws of logic are known a priori; they are justified by apodictic self-evidence, and certain rather than probable. And therefore, laws of logic are not psychological.

If logical laws were psychological laws, they would refer to psychological entities. But, logical laws do not refer to psychological entities. And therefore, logical laws are not psychological laws.

All forms of psychologism imply relativism. Psychologistic thinkers take logic to be true only for the human species. But this contradicts the meaning of truth. Truths are unchanging and independent of the human mind.

A good deal of the psychologism controversy concerned the validity and originality of Husserl's arguments. Several commentators maintained that Husserl's first three arguments all begged the question. Others challenged Husserl's claim that the laws of logic did not imply matters of fact. Husserl's critics also denied that truths were eternal or independent of human thought.

The debate over psychologism was vigorous. This can be seen, for instance, from the fact that more than 130 philosophers were accused of advocating psychologism and that there were as many symptoms and definitions of psychologism as there were philosophical schools (Kusch, 1995). It was not merely the relationship between logic and psychology that was at issue, but the relationship between psychology and philosophy as a whole. This was a pressing issue not least because the practitioners of the then new (experimental) psychology worked in philosophy departments, and because between 1873 and 1913 the number of full professorships held by these 'psychologists' increased from 1 to 10. Of the 39 full professorships in philosophy in 1892, practitioners of experimental psychology held 3; of the 42 full professorships in 1900, they occupied 6; while of the 44 full professorships in 1913, they had already gained 10. Practitioners and advocates of the new psychology sometimes claimed that psychology provided the key to real progress in philosophy or that it was the heart of philosophy (Ash, 1995; Kusch, 1995, 1999).

This last-mentioned contention was opposed by a quickly growing group of 'pure' philosophers – led by the Neo-Kantians, and by Rudolf Eucken, Wilhelm Dilthey, and Husserl. They deemed it harmful to philosophy and its progress that experimental psychology should be regarded as part and parcel of philosophy. To put it another way, these philosophers believed that the role of the philosopher, as suggested by the academic success of experimental psychology, needed purification.

The prolonged debate between the advocates and opponents of psychology as a philosophical discipline was eventually brought to an end by World War I. The war brought about an atmosphere in which attacks on one's colleagues were regarded as utterly inappropriate. Moreover, it also led to a clear division of labor between pure philosophers and psychologists: While pure philosophy concentrated on the ideological task of celebrating the German 'genius of war,' experimental psychology focused on the training and testing of soldiers. After the war, both academic pure philosophy and experimental psychology had to cope with, and accommodate to, an intellectual environment that was hostile to science, rationality, and systematic knowledge. The project of a scientifically minded philosophy with experimental psychology as its central pillar quickly lost support. Advocates and practitioners of experimental psychology therefore had to find new ways of justifying their work. Many of them continued their wartime involvement with applied psychology (Kusch, 1995).

The Influence of the Psychologism Controversy

Psychologism and Sociologism

The way the psychologism controversy was fought by the defenders of philosophical purity became an important model for subsequent disputes between advocates and opponents of 'naturalism' (the view that philosophical questions can be answered using the results of the natural and social sciences). Many later arguments against 'sociologism' or 'biologism' closely mirror Frege's and Husserl's ideas.

For instance, in the German-speaking world the psychologism controversy of the first two decades of the twentieth century was followed by the dispute over the nature and philosophical aspirations of the sociology of knowledge. Accusations of 'sociologism' surfaced quickly and 'sociologism' was seen by many as the 'successor' to psychologism. Again many philosophers felt that a special science – this time sociology – was trying to occupy philosophical terrain. Critics of the sociology of knowledge accused it of both psychologism and sociologism, and in so doing, they drew on Husserl's work in particular. Defenders of the sociology of knowledge tried a similar feat in presenting their work as a bulwark against psychologism (Meja and Stehr, 1990). This idea has become a general feature of the field to this day (Bloor, 1983). The close link between the two 'isms' has been transmitted to subsequent Anglophone thinking mainly through the work of Karl Popper.

Psychologism and Philosophy

German-speaking philosophy at the beginning of the twentieth century exerted a powerful and lasting influence on philosophy in many other countries. Frege's (1893) and Husserl's (1970) antipsychologism thus became the dominant view in Western philosophy for much of the twentieth century. This is not to deny that other countries also had their own debates over the relation between psychology and philosophy. In the UK, for instance, George Frederick Stout and James Ward emphasized the philosophical significance of psychology; British idealists like Francis Herbert Bradley and Brand Blanshard were sympathetic to psychology, even though they denied its philosophical credentials; and realists like John Cook Wilson, Bertrand Russell, Herbert Joseph, and Herbert Prichard insisted on a separation. For Prichard, psychology was not even a 'proper science' (Passmore, 1994).

The institutional separation of psychology from philosophy eventually removed one important source of conflict. Once institutional separation had been achieved, 'pure' philosophers and 'impure' philosophers (i.e., psychologists) were no longer in direct competition over academic chairs. On the plane of ideas, however, the psychological challenge has remained. Throughout the twentieth century, psychology exerted a powerful pull on philosophers who wished to put philosophy on the sure path of a natural science. Willard Van Orman Quine explicitly called for a return to psychologism in the 1960s. Such calls have kept the psychologism accusation alive. Among the accused, one can find the names of Rudolf Carnap, Michael Dummett, Peter Geach, Nelson Goodman, Thomas Kuhn, John McDowell, Karl Popper, Wilfrid Sellars, and Ludwig Wittgenstein.

It is now common among historians of philosophy and psychology to lament the outcome of the psychologism controversy. The separation of philosophy from psychology delayed the development of research programs like 'cognitive science' that combine psychological experimentation with philosophical aspiration.

See also: Determinism: Social and Economic; Experimentation in Psychology, History of; Husserl, Edmund (1859–1938); Knowledge, Sociology of; Logic and Linguistics; Naturalism; Psychology of Inferences; Relativism: Philosophical Aspects.

Bibliography

Ash, M.G., 1995. Gestalt Psychology in German Culture 1890–1967: Holism and the Quest for Objectivity. Cambridge University Press, Cambridge, UK.

Bloor, D., 1983. Wittgenstein: A Social Theory of Knowledge. Columbia University Press, New York.

Frisby, D., 1983. The Alienated Mind: The Sociology of Knowledge in Germany 1918–33. Heinemann, London.

Frege, G., 1893. Grundgesetze der Arithmetik: Begriffsschriftlich Abgeleitet, vol. 1. Jena, Pohle.

Haaparanta, L. (Ed.), 1994. Mind, Meaning and Mathematics: Essays on the Philosophical Views of Husserl and Frege. Kluwer, Dordrecht, The Netherlands.

Husserl, E., 1970 (J.N. Findlay, Trans.). Logical Investigations, vol. I. Routledge, London.

Kusch, M., 1995. Psychologism: A Case Study in the Sociology of Philosophical Knowledge. Routledge, London.

Kusch, M., 1999. Psychological Knowledge: A Social History and Philosophy. Routledge, London.

Meja, V., Stehr, N., 1990. Knowledge and Politics: The Sociology of Knowledge Dispute. Routledge, London.

Notturno, M.A., 1985. Objectivity, Rationality and the Third Realm: Justification and the Grounds of Psychologism: A Study of Frege and Popper. Nijhoff, Dordrecht, The Netherlands.

Passmore, J., 1994. A Hundred Years of Philosophy, second ed. Penguin, London.

Popper, K., 1945. The open society and its enemies. In: The High Tide of Prophecy: Hegel, Marx, and the Aftermath, vol. II. Routledge, London.

Quine, W.V.O., 1969. Ontological Relativity and Other Essays. Columbia University Press, New York.

Psychology of Inferences

Janice M Keenan, University of Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The goal of comprehension is to understand what the speaker intended the text to mean. Inferences are driven by a desire to make the interpretation both more coherent and more elaborate than the text itself. The goal of research on inferencing is to specify how the computational processes involved in making inferences vary with the comprehender's knowledge, the conditions that promote inferencing, the various types of inferences and methodological problems involved in assessing them, and, most recently, the neural bases of inferencing.

From a legal perspective, inferences are easily defined by contrasting them with direct assertions. Assertions have physical representations, e.g., they exist on paper, whereas inferences do not; asserting a falsehood can lead to a criminal conviction, whereas implying a falsehood is typically not a crime. One of the important findings from research on the psychology of inferences has been the demonstration that this black-and-white distinction between inferences and assertions, which is so evident and straightforward at the level of the text, is not that consequential at the level of understanding the text. Because a comprehender's goal is to understand the author's message, not to represent the text, the information in the text is integrated with other information that the comprehender knows in order to construct an interpretation. What is important for constructing this interpretation is not where the information came from (the text or the comprehender's knowledge), but rather what it means and how it coheres. Inferencing is driven by a desire to make the interpretation both more coherent and more elaborate than the text. The goal of research on inferencing is to specify the conditions that promote inferencing, the various types of inferences that are made, the computational processes involved in making them, and, most recently, the neural bases of inferencing. Like any research domain, inference research has had its share of controversy and issues with methodological problems; nonetheless, the research has provided a solid foundation for understanding individual differences in inferencing. It is now clear that inferencing is not a single cognitive skill that individuals may be more or less good at. Inferencing is a set of processes that vary in their computational nature depending on both the comprehender's cognitive skills and the comprehender's specific knowledge relevant to the text.

Is Explicit Always Easier?

Although intuition might suggest that explicit information should always be easier to process than implicit information, one of the great insights of inference research has been to show that this notion of processing ease rests on a faulty assumption about the primary goal of language comprehension. It was initially thought that the goal was to construct a representation of the meaning of the text. In this view, explicit information is easy because it is direct, whereas implicit information requires an effortful process of determining what information is missing and then supplying it. Early studies of inferencing reinforced this view that inferring implicit information is more difficult than processing explicit information. For example, Haviland and Clark (1974) found it took longer to read a target sentence when the preceding sentence only implicitly mentioned the referent of the target sentence rather than explicitly mentioned it, as in:

Implicit: *Herb took the picnic supplies out of the car.*
Explicit: *Herb took the beer out of the car.*
Target: *The beer was warm.*

At the same time when reading time studies were reporting that implicit statements were more difficult to process than explicit ones, there were many early studies showing that, as far as memory was concerned, implicit information was functionally no different than explicit. For example, Bransford et al. (1972) found that people could not even recognize whether statements had been implicit or explicit when the statements could be integrated to form a story. Perhaps more remarkable was the finding that implicit information could be even more accessible in memory and easier to verify than information that was explicitly presented. Specifically, when statements refer to a linear ordering, e.g., *A is bigger than B*, *B is bigger than C*, explicit statements that refer to adjacent terms are more difficult to verify than implicit statements (e.g., *A is bigger than C*) referring to referents more distantly spaced in the ordering. These memory studies were soon supported by reading time studies, such as reported by Sanford and Garrod in their 1981 book, *Understanding Written Language*, showing that it took no more time to read *The clothes were made of pink wool* when it was preceded by the implicit *Mary dressed the baby* than by the explicit *Mary put the clothes on the baby*.

In order to accommodate findings such as these, the field underwent a change in its metaassumption about the goal of comprehension. The goal for the comprehender was no longer representing the text but rather representing what the speaker intended the text to be about. This was referred to as the mental model view of discourse. In this view, the difference between what is given and what must be filled in with an inference is defined relative to the comprehender's knowledge structures, not the text. But since knowledge structures cannot be observed and because they vary across individuals, the distinction is not

 http://dx.doi.org/10.1016/B978-0-08-097086-8.57011-1

as straightforward as when it is based on the text. The advantage, however, is that this perspective better explains the data. It allows explicit information to not always be easier than implicit. It also accommodates the fact that making information explicit is not even necessary for effective communication. For example, when two people know each other well, just a few words can function to activate the appropriate knowledge. It also accommodates the fact that being explicit can actually be objectionable. Try making all the thoughts behind even a simple sentence explicit. The result is typically the kind of incomprehensible verbiage found in legal documents; incomprehensible because the words serve to detour one from the intended knowledge structures.

The mental model view not only better accommodated the data, it also put the focus on specifying the computational processes involved in inferencing. How does the comprehender's knowledge get activated, and once activated, how does it get integrated with the text to build a model? Answering these questions involved first determining which types of inferences are drawn.

Types of Inferences

The question of which types of inferences are typically drawn during comprehension is important to answer not only for developing computational models of inferencing, but also for addressing the needs of other enterprises. One is the design of artificial intelligence language understanding systems; designers need to know which types of inferences people typically draw in comprehension so that they can program their systems to make the same inferences and thus function like typical natural language users. Another is the study of individual differences in language and cognition. The evaluations that are done in this field to determine whether someone has deficits in inferencing can only be done if one has a baseline understanding of which inferences are typically made by non-impaired individuals.

The most commonly drawn inferences are those that bridge a coherence gap between two textual elements, such as two phrases. Early research identified these inferences as *bridging inferences*; the example text given earlier involving *beer* and *picnic supplies* requires the comprehender to bridge the two sentences by inferring that the picnic supplies must have included beer. These are also called *backward inferences* because when a gap in coherence is detected, the comprehender works backward to build a bridge to the preceding text.

What constitutes a coherence gap is a function of linguistic features of the text as well as the comprehender's knowledge. The field is far from a complete theory of how a text maps onto that knowledge, but important steps have been taken in that direction. One is the identification of the dimensions of coherence. Several categories of cohesion signals and cohesive relations that exist in narrative and informational text have been specified and supported by research showing that inferences are made in order to maintain referential, causal, and temporal coherence. Furthermore, an automated system using these cohesion cues for assessing the coherence of texts, CohMetrix, has been developed (McNamara et al., 2010). Finally, the field has come to recognize how there exist individual differences in 'standards of coherence' – how much coherence the comprehender seeks (van den Broek et al., 1995). Those with high standards of coherence engage in more inferencing so as to achieve greater connectedness of the representation they construct, whereas those willing to tolerate less coherence will inference less. Individual differences in standards of coherence arise from factors such as motivation, relevant background knowledge, and purposes for reading. Standards of coherence vary not only across readers but also within the same reader across different reading situations.

Whereas backward inferences are made to establish coherence and are assumed to occur at coherence breaks in the text, *forward inferences*, also called *elaborative* or *predictive* inferences, merely elaborate the developing mental representation and can occur at any point. Compelling reasons have been proposed for why readers and listeners ought to make forward inferences, and it is often said that the more skilled the comprehender, the more forward inferences they make. But there are also costs associated with making such inferences. For example, if *spoon* is inferred from *stirred the soup*, reading a subsequent sentence about *the spoon* will certainly be easier if *spoon* is already in an activated state. However, if the text proceeds with some other construct, e.g., *stick*, mentioned as the instrument of the action, then the reader who inferred spoon will need to correct their understanding at a cost to their processing resources. The mix of potential costs and benefits associated with forward inferences has made them the focus of much research, primarily assessing the extent to which they are actually drawn. Research has focused on types of forward inferences thought to be most likely to occur: instrument inferences (e.g., inferring *spoon* from *stir the soup*), inferences about the likely consequences of events (e.g., inferring *cut* from *stepped on jagged glass*), and instantiations of general terms (e.g., inferring *cow* from *milk the animal*).

In a review of this research, McKoon and Ratcliff (1992) concluded that fewer of these forward inferences are drawn than what might be thought. They characterized the comprehender as operating according to what they called the *minimalist hypothesis*. It states that the only inferences that are reliably drawn are those required for local coherence and those involving activations of readily available preexisting associations. This started quite a controversy. To many researchers minimalism represented a return to the earlier text-centered view of comprehension. These researchers argued for a more knowledge-based, *constructivist* view (Graesser et al., 1994), and produced several studies arguing against minimalism by showing that even when texts are locally coherent, inferences related to global coherence are still typically drawn. Some of this controversy also revolved around methodological issues.

Methodological Issues

The main paradigms used to assess inferencing can be divided into two classes: *activation measures* and *memory measures*. Activation measures detect inferences by determining whether the inference concept is in an activated state relative to some control condition where the inference would not be expected to be drawn. Activation measures include reading times for a sentence where the inference might be drawn, reaction times

to make a lexical decision on the inference word or a word related to it, and time to name a word referring to the inference. Memory measures require subjects to access their mental representations of the text and determine if the inference is part of that representation. They include cued recall, sentence verification, question answering, and recognition tests.

Much has been learned about the methodological problems associated with each of the various measures (Keenan et al., 1990). The main problem with memory measures is that they do not allow one to determine whether the inference was made during comprehension or if it was made as a result of the test. Because memory tasks require a person to go back to their representation of the text to determine whether the information is present, they provide the person with another opportunity to draw the inference. This is a real problem if one seeks to know whether inferences are being drawn spontaneously at the time that the text is being understood or if they are only being drawn as the result of questions.

Activation measures have potential problems as well. First, not all activated concepts will get integrated in the comprehender's mental model as an inference. Some inference concepts may be activated spuriously by virtue of their association to other words in the text, a phenomenon referred to as word-based priming. For example, if the concept *dead* is activated by the sentence *the fashion model fell from the roof during the photo shoot*, it may either indicate that a forward inference was drawn about the model being dead from her fall or it may be that *dead* is activated because it is related to *shoot*. Thus, activation measures of inference require assessing activation relative to a sentence that controls for word-based priming. Second, the point in time when activation of an inference concept is assessed may affect whether or not it appears to have been drawn. Activation assessed immediately following the target text segment may suggest an inference was drawn; but if the inference is not incorporated into the long-term representation of the text, then a test occurring at the end of the passage may suggest that the inference was not drawn because the activation will have dissipated from not being integrated into the representation. Good methodology would suggest using multiple time points for assessment.

Because the different types of tasks used to assess inferencing can yield different results, it is important that research aimed at assessing whether inferences are made during comprehension use more than one measure of inferencing, with multiple control conditions, and assessments at various points in time. Converging operations such as these can help dispel the distorted picture of inferencing that can emerge when classes of inferences are confounded with the methods used to assess them.

Computational Processes in Inferencing

The computational processes of inferencing are not the same as computational models of these processes. The computational processes of inferencing are basically theories derived from experimental data of how text elements activate and generate knowledge structures during comprehension; these theories generate predictions that are tested in behavioral experiments. Computational models of inferencing, on the other hand, are particular instantiations of a theoretical framework on a computer. Examples are Kintsch's Construction–Integration model and van den Broek's Landscape model. Computational models provide an additional way in which to test theories because they provide simulation data that can be compared to data obtained from people comprehending the same texts; note, however, that when a model deviates from human behavior, it is sometimes difficult to know if it is because of its assumptions about inferencing or because of assumptions required to make the specific implementation run. Computational modeling has contributed significantly to unraveling the complexity of text comprehension processes by demonstrating for example how inference processes vary with dynamical changes to the reader's working memory, the specifics of retrieval mechanisms, and decision rules for integration.

It is clear from both empirical work and modeling that inferencing refers not to a single cognitive skill, but to a broad range of computational processes, varying from simply activating knowledge in the comprehender's memory to generating new knowledge structures. The outputs can similarly range in complexity, varying from a single concept (e.g., *spoon*) to a proposition (e.g., *that one event caused another*) to a set of propositions or schemas (e.g., *one's schema for romantic relationships*).

Kintsch's (1998) influential Construction–Integration model describes two computational phases of comprehension: first possible relevant knowledge structures are activated (Construction); second, there is an attempt to integrate these with the evolving representation of the text which has the consequence of selecting the interpretation (Integration). The computational processes of inferencing are similar. What varies across inferences is the complexity of activating information that allows the current text to be integrated and made coherent with the representation of what has been understood so far.

At the simplest computational level, inferencing involves merely activating concepts that a person already knows through a process called *automatic passive resonance*, where concepts that are associatively related to those explicitly occurring in the text get activated and integrated – e.g., activating *spoon* when comprehending *stirred the soup*. Concepts that get activated by passive resonance, however, may not always be sufficient to bridge coherence gaps. To illustrate with the earlier example about *picnic supplies*, *beer* is unlikely to get activated in most comprehender's minds by the phrase *picnic supplies*. In that case, when the comprehender encounters the phrase, *the beer*, a search of relevant information in memory is initiated to try to *retrieve* the fact that *beer can be a picnic supply*, and thus permit a bridging inference. This retrieval process requires more effort than producing inferences by passive resonance. If retrieval fails, then an even more computationally complex process of knowledge generation must take place. Using knowledge that the two sentences must somehow cohere and that definite reference is only used when a concept has previously been mentioned, the reader determines that *beer* must be considered part of *picnic supplies*. The reader thus makes an inference by generating this new knowledge.

It should be evident from this discussion that the ease of processing an inference is not just a function of information in the text. The types of computations required, and the cognitive effort required to construct an inference, can vary considerably

as a function of the comprehender's knowledge. The more knowledge one has, the less effortful inferencing is.

An interesting finding is the complex relation that exists between effort in drawing an inference and subsequent retention of the information. This is illustrated by studies of causal inferencing (Keenan et al., 1984), using two-sentence texts like the following in which various versions of the first sentence described a cause that varied in its relatedness to the event described by the second target sentence.

Highly Related: *Joey's big brother punched him again and again.*
More Moderately Related: *Racing down the hill, Joey fell off his bike.*
Less Moderately Related: *Joey's crazy mother became furiously angry with him.*
Unrelated: *Joey went to a neighbor's house to play.*
TARGET: *The next day his body was covered with bruises.*

Reading times increased as the relatedness of the sentences decreased, reflecting the likelihood that passive resonance from the first sentence produced the event described by the second sentence. However, as we noted above, when passive resonance fails, then more effortful memory retrieval or knowledge generation processes come to the fore. What is interesting is that these more effortful processes lead to better memory than passive resonance. When memory was tested for the causes by giving people the second sentence as a cue, moderately related causes were remembered best. The increased effort required to integrate moderately related causes improved retention, but only when the causes could be integrated. When they could not, as with unrelated sentence pairs, the increased effort did not aid memory. Thus, an inverted-U shape function was obtained between relatedness and memory.

Neural Bases of Inferencing

Early views of language and the brain proposed that higher-level language processes, such as inferencing, took place in the right hemisphere. This view was compatible with ideas about hemispheric differences derived from visual spatial research. Because left hemisphere processes were considered analytic, whereas the right hemisphere was integrative and synthetic, it fit that the integrative process of text inferencing would be right hemisphere. Two lines of evidence were also offered in support of this view. One came from behavioral studies of typical readers showing more activation of inference concepts when they were presented to the left visual field (right hemisphere) than to the right visual field. The other line of evidence came from patients with lesions to the right hemisphere; these patients showed great difficulty in drawing inferences to integrate sentences and maintain coherence.

More recent studies of the neural basis of inferencing in language comprehension using brain imaging (positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)) have shown that while the right hemisphere makes a considerable contribution, it is not the exclusive domain of complex language processing. In a review of studies of the neural basis of inferencing during discourse processing Ferstl (2007) reported considerable evidence for left-lateralized or left-dominant activation. In addition, more recent behavioral studies of patients show that difficulties in inferencing are not just associated with right hemisphere lesions; they also occur when cognition is compromised for other reasons, such as developmental disabilities like Fragile X. Thus, inferencing does not appear to be confined to the right hemisphere.

The picture that emerges from imaging studies of the neural basis of inferencing is rather complex, as might be expected from the previous discussion of the varieties of computational processes that can underlie inferencing. Like the behavioral studies showing that there is not a single cognitive process underlying inferencing, imaging studies find that there is not a single underlying neural network involved in inferencing. The field has moved away from the idea of a one-to-one mapping of cortical activation on to specific cognitive processes – e.g., this area does inferencing – to thinking in terms of organizations of cortical regions into functional networks.

A variety of cortical networks appear to be involved in inferencing. The specifics of which regions participate in a particular inference depend on the difficulty of the inference being computed and the knowledge domain in which it is being made. Although imaging work on discourse processing and inferencing is still in its infancy, the networks tentatively identified as associated with component processes of inferencing according to Mason and Just (2011) include a coarse semantic processing network (right middle and right superior temporal areas), a network associated with the difficulty of the inference (bilateral dorsolateral prefrontal cortex), a network associated with memory for the inference (left inferior frontal and left anterior temporal), and different networks associated with different types of information being inferred, such as spatial inferences (intraparietal and middle occipital) and emotional inferences (right temporoparietal and medial frontal).

One of the first studies to show how the computational difficulty of the inference affects cortical activation was Mason and Just's (2004) imaging study of causal inferencing, using texts which varied in their causal relatedness like the Joey passage described above. They found bilateral cortical activation in dorsal lateral prefrontal cortex (DLPFC) that linearly increased as levels of causal relatedness decreased. This pattern was similar to what is typically observed for reading times on the second sentence; it reflects the differential effort required as the computational process of inferencing shifts from easily activating inference knowledge by passive resonance when the cause is highly related to more effortful processing when it is less related. Right hemisphere homologues of left hemisphere language areas appeared to be specifically involved in inference integration because they showed the same inverted-U previously reported for memory for the causes.

How the type of knowledge involved in the inference can also affect the specific cortical network activated was first demonstrated using PET imaging (Fletcher et al., 1995) of people comprehending passages on various topics. Differential cortical activation was observed when the topic required the comprehender to invoke a theory of mind to keep track of character's goals and feelings versus when it required understanding spatial configurations described in the text. This was an important finding because it showed

early on that there is not a single neural network associated with inferencing, and that the specific network activated depends on the passage content and the reader's knowledge. Imaging studies, which now use fMRI instead of PET, have recently replicated this finding of differential recruitment of cortical areas as a function of the type of knowledge required to make the inference (Mason and Just, 2011).

How the cortical networks associated with inferences differ depending on the type of knowledge involved is also apparent from imaging studies comparing individuals with and without autism. Cortical activation associated with reading passages that either required inferences based on intentions and emotional states or on spatial relations and physical causality was compared; deficits associated with autism were found for social inference networks but not those underlying other inferences (Happe, 1994). We can expect to learn more about the specifics of how inference processes map onto cortical networks as imaging researchers begin to do more research like this that focuses on individual differences and populations with cognitive and language deficits. Another avenue for future imaging research on inferences will likely involve understanding the brain activity associated with event segmentation and episode structure because inferences are often required to connect these segments. Together with enhanced imaging and enhanced data analytic techniques, studies manipulating passage content and the skills of the comprehender should shed light on the component processes of inferencing.

In sum, the neural basis of inferencing revealed through brain imaging studies supports the richly complex view of inferencing that has emerged from behavioral studies. Together these two sources should someday provide the basis for understanding why there are individual differences in inferencing skill.

Individual Differences in Inferencing

Studies of individual differences in inferencing often refer to it as a unitary skill; one that young readers need to develop, one that poor readers have a deficit in, and one that can be selectively impaired with brain damage. However, it should be evident from our discussion so far that such a view is an oversimplification. Inferencing is not a unitary skill with a defined locus in the brain. It represents a range of computational processes that vary not only with the skills of the comprehender but also with the comprehender's knowledge. Thus, individual differences in inferencing can stem from differences in knowledge as well as differences in cognitive skills. Young readers may therefore show less inference skill than mature readers in part because they have less knowledge of the world and its situations. Similarly, poor readers may appear to have deficiencies in inferencing that can be largely due to their having less knowledge (because they read less) as well as fewer resources to allocate to inferencing because of their struggles with word identification. Furthermore, knowledge differences are reflected in different neural networks, allowing inference deficits to be confined to a specific knowledge domain, as in high-functioning individuals with autism who show inference deficits only in the social–emotional domain.

Given the variety of skills underlying inferences and the fact that inference skill can vary even within an individual depending on the knowledge domain (Noordman and Vonk, 1992), the field is moving away from simply categorizing individuals as *generally* good at inferencing or generally poor at it. The focus of individual differences research now is to develop a deeper understanding of the cognitive and neural factors contributing to the range of differences across inferences of various computational types in various knowledge domains and to determine the extent to which these differences are due to variations in genes and environment.

It is important to note that the fundamental reliance of inferences on specific knowledge structures does not mean that knowledge differences are the only reason for individual differences in inferencing. Differences can also arise from individual differences in skills that inferencing processes draw on, such as working memory capacity, comprehension monitoring, or standards of coherence. For example, greater working memory capacity can facilitate inference processing because it permits more widely separated segments of text to be connected, and those who are better monitors and have higher standards of coherence will make more inferences in order to have a text be more coherent. Individual differences in component skills of inferencing tend to be positively correlated; so one goal of research on individual differences has been to determine whether performance on inference tests can be fully explained by individual differences in these component skills. If not, researchers make the case that separate inferencing specific skills are required to explain variance above and beyond that explained by component skills. A good example of such work is to be found in Oakhill and Cain's (2012) longitudinal study of children's comprehension processes.

It is common to think of differences in cognitive mechanisms as arising from biological differences coded by genes, and differences in knowledge arising mainly from the environments provided by family and schools. But recent studies using twins to assess the etiology of individual differences suggest a more complicated picture. Twin studies have compared the similarities of identical twins, who shared all their genes as well as family and school environments, to the similarities of fraternal twins, who shared their environments but only half their genes on average, in order to distinguish effects due to genes and to environment. Although there have been no twin studies of inferencing, it has been shown that individual differences in comprehension are largely explained by genetic influences, although environmental influences are also apparent (Keenan et al., 2006). At first glance, this may seem difficult to reconcile with what we know about comprehension and inferencing being so dependent on knowledge. Because knowledge is acquired from the environment, it might seem that influences on individual differences in comprehension might be more environmental than genetic. What the twin studies make evident, however, is that there are genetic influences on the cognitive and neural mechanisms that process that knowledge and determine the ease with which acquisition and use of knowledge occurs. These findings illustrate how important it is to examine genetics and understand their influence and interplay with

environmental variables like exposure to knowledge. We can expect that future research on comprehension and inferencing will represent a synthesis of cognitive, neural, and genetic approaches.

See also: Cortical Areas Engaged in Movement: Neuroimaging Methods; Implicit Memory; Implicit Social Cognition; Inferences in Discourse, Psychology of; Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Neural Basis of Reading; Speech Processing, The Cortical Organization of.

Bibliography

Bransford, J.D., Barclay, J.R., Franks, J.J., 1972. Sentence memory: a constructive vs. interpretative approach. Cognitive Psychology 3, 193–209.

Ferstl, E.C., 2007. The functional neuroanatomy of text comprehension: what's the story so far? In: Schmalhofer, F., Perfetti, C. (Eds.), Higher Level Language Processes in the Brain. Lawrence Erlbaum, Mahwah, MJ, pp. 53–102.

Fletcher, P.C., et al., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57, 109–128.

Graesser, A.C., Singer, M., Trabasso, T., 1994. Constructing inferences during narrative text comprehension. Psychological Review 101, 371–395.

Happe, F., 1994. Communicative competence and theory of mind in autism: a test of relevance theory. Cognition 48, 101–119.

Haviland, S.E., Clark, H.H., 1974. What's new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior 13, 512–521.

Keenan, J.M., Baillet, S.D., Brown, P., 1984. The effects of causal cohesion on comprehension and memory. Journal of Verbal Learning and Verbal Behavior 23, 115–126.

Keenan, J.M., Betjemann, R.S., Wadsworth, S.J., DeFries, J.C., Olson, R.K., 2006. Genetic and environmental influences on reading and listening comprehension. Journal of Research in Reading 29, 79–91.

Keenan, J.M., Potts, G.R., Golding, J.M., Jennings, T.M., 1990. Which elaborative inferences are drawn during reading? A question of methodologies. In: Balota, D., Flores d'Arcais, G., Rayner, K. (Eds.), Comprehension Processes in Reading. Erlbaum, Hillsdale, NJ, pp. 377–402.

Kintsch, W., 1998. Comprehension. Cambridge University Press, Cambridge.

Mason, R.A., Just, M., 2004. How the brain processes causal inferences in text. Psychological Science 15, 1–7.

Mason, R.A., Just, M., 2011. Differential cortical networks for inferences concerning people's intentions versus physical causality. Human Brain Mapping 32, 313–329.

McKoon, G., Ratcliff, R., 1992. Inference during reading. Psychological Review 99, 440–466.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C., 2010. Coh-Metrix: capturing linguistic features of cohesion. Discourse Processes 47, 292–330.

Noordman, L.G.M., Vonk, W., 1992. Reader's knowledge and the control of inferences in reading. Language and Cognitive Processes 7, 373–391.

Oakhill, J., Cain, K., 2012. The precursors of reading ability in young readers: evidence from a four-year longitudinal study. Scientific Studies of Reading 16, 91–121.

van den Broek, P., Risden, K., Husebye-Hartmann, E., 1995. The role of readers' standards for coherence in the generation of inferences during reading. In: Lorch, R., O'Brien, E.J. (Eds.), Sources of Coherence in Reading. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 353–373.

Psychology of the Real Self: Psychoanalytic Perspectives

Roger MK Ng, Kowloon Hospital, Kowloon, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.F. Masterson, volume 18, pp. 12405–12409, © 2001, Elsevier Ltd.

Abstract

This article explores the concept of selfhood from a variety of psychological perspectives. The psychoanalytic perspective offers a hypothesis that a stable core selfhood arises from a secure infant–mother relationship and grows in response to further interpersonal relationships in adult lives. The cognitive accounts offer additional models on how the self-concept is being maintained, clarified, evaluated, and developed with time and across different situations. The Buddhist perspective on selfhood nevertheless offers a radically different point of view about the value of selfhood in the maintenance of positive mental health.

Introduction

In the West, the self has all along been considered as an enduring entity, as established, for example, by Descartes, 'I think therefore I am' or by Freud, 'where Id was there ego shall be.' From the point of view of Western psychology, the self is unique in that it is both subject and object so that the self can be simultaneously experienced and observed. This duality of self allows people to think and talk about ourselves as both 'I' and 'me,' and thereby enables the two aspects of the self to converse with one another. Understanding how self and identity are constructed and maintained in healthy individuals enhances the understanding of the maintenance of negative self-views in psychological disorders. The experience of a continuous and coherent self is the result of complex interplay between knowledge, memory, experience, and various self-regulatory processes. Furthermore, the definition of self is also a dynamic interplay between self as an entity, the interface of self with others, and the interface of self with the outside world. The concept of self has been discussed and elaborated in different schools of psychology. The current article attempts to give an overview of the development of the concept of self, with due references to the psychoanalytic, cognitive behavioral, and existential schools of thought about development of self-concepts.

Self as an Entity: Psychoanalytic Perspectives

Influenced by his scientific education and by the ideas generally accepted in the cultural era that he worked, Freud first explained human motivation by reference to two innate instincts, the drives of sex and aggression. At the early stage of the evolution of his psychoanalytic theory of the self, Freud adhered to his academic background of basic science and proposed his first theory of anxiety as sexual instinct being damned up in the form of sexual tension that intruded into the consciousness (Freud, 1959). This idea tends to picture the psyche as a mechanism, an impersonal arrangement for securing detensioning, and a homeostatic organization. Acknowledging the shortfall of this theory, he revised this physical theory for a more sophisticated theory that anxiety was an ego defense reaction to danger that implied a signal in one's self-experience of some imminent threat. This is an important shift in Freud's theories that recognized the ego as the personal self. Freud further expanded this concept of selfhood by developing a more object-relation psychology (see below) of the influence people have on each other's lives, especially parents on children in the form of Oedipus complex. The shift of emphasis toward parent–child relationship led to the development of the superego, a mental structure that internalized the moral judgments, standards, and rules set by the parents and society at large. This led to the revision of the 'hydraulic' model to the 'structural' model of the self, viewing the ego as a control-system agency being tossed about among the instinctual drives, of the outside world, and of a forbidding conscience ('superego') (Freud, 1949). However, the incorporation of the superego into the above mentioned structural model can still be considered a mechanistic model, which explains behaviors as ego defenses against encroaching instinctual drives. Sullivan attempted to synthesize the mechanical model with the object-relations model by proposing that the biological substrates of personality as responsible for taking care of the appetites as organic needs, and the brain and the nervous system as the machinery of perception, thinking, and bodily functions. These substrates were then owned and operated within the psychic self or the ego (Sullivan, 1968). This model allows the overall state of the ego or personal self to determine the thoughts, emotions, and behaviors responsible for satisfying the organic needs. He also emphasized the importance of the subjectivity of experiencing as the true concern of psychodynamic studies. Sullivan also explored the growth of the individual ego in its ever-widening social milieu through interpersonal relationships. This marks a clear separation of interpersonal model of selfhood from a biological one as advocated by Freud.

On the other hand, Melanie Klein, Fairbairn, and Winnicott explored the internal psychic drama of the growing ego through tracing its development to its earliest beginnings in the context of infant–mother relationships. Klein regarded an infant as an arena for an internal struggle between life and death instincts. The immature ego of the infant is therefore exposed to the anxiety stirred up by the inborn polarity of instincts. The inner drama is projected onto the outer world, as

the infant's neurological systems develop the capacity to discern external objects. In other words, the infant only experiences the external world and objects through the lens of his innate makeup rather than through their actual behaviors to him. The death instinct is projected into the mother's breast to ward off his anxieties, which is then introjected, so that his experience of the outer world serves to increase and confirm his anxieties toward the internal dangers and inner conflicts (Segal, 1964). The external objects are therefore indispensable, as the infant needs them to project its death instinct onto them, beginning with the mother's breast. This first hate-object is necessary in order to project the death instinct that threatens the infant with psychic destruction. The first hate-object now becomes a real external threat as the infant has no means of dealing with it physically. In order to cope with this threat, the bad breast is now introjected as an object and not as an instinct. The external objects are therefore valued as receptacles for projection of instincts. The internal psychic life therefore consists of a hostile tension between two contradictory forces, and a pattern world is created into which the child's experience of the external world is fitted. What Melanie Klein did is to display the internal psychic life not as a battleground of instincts but a highly personal inner world of ego-object relationships. This inner world may come to conscious expression in play and dreams, and be disguised in symptoms and in disturbed behavior-relations to real people in daily encounters.

While Melanie Klein still developed her object-relation theories within the framework of life and death instincts, it is Fairbairn who considered the self as a unitary concept (Fairbairn, 1954). He argued that impulses could not be considered apart from object or ego-structures. The life of ego-structures is considered to be consisting of activities arising from impulses. He believed that there was a unitary dynamic ego that was primitive and undeveloped during infancy, which then further grew and developed with age. Mental health with preserved selfhood would then be regarded as a preservation of the ego wholeness in passing through the maturational stages on the way to adult maturity (Fairbairn, 1954). The entire process of growth, disturbance, and restoration of wholeness as an ego or personal self then depends on the ego's relations with objects, primarily in infancy, and thereafter on the object relations in real life. Whole ego development depends on good object relations in real life whereas split ego development occurs due to bad object relations in real life. This is where Fairbairn radically differs from Melanie Klein's views about split ego development arising from the constant battles between life and death instincts from the start. To Fairbairn, the infant is whole and will remain so if protected long enough by good object relationships in his encounters with the real world, starting from early relationships with his mother. Good object experience simply leads to good ego development. Extending the importance of good object relations in early life as suggested by Fairbairn, Winnicott proposed the concept of 'average expectable environment' as essential to the growth of healthy personality in the child, and the important element in it is the 'good enough mother' (Winnicott, 1965a). According to Winnicott, a continuously helping and facilitating environment that accepts the infant's immature dependence while supporting his tentative adventures into individuality and independence is essential for healthy personality development. With such a secure maternal environment, the child is then able to have a capacity to be alone (Winnicott, 1965b). The infant does not feel he has lost his mother when he cannot see her and he does not feel isolated when he is physically alone. This built-in experience is considered by Winnicott to last for a lifetime as a foundation on which all the child's later growth can take place. Winnicott believed that, when an adult develops this mature capacity to be alone without feeling isolated, he will then be able to enjoy solitude, delve into creative work, and commit himself to his true values and life goals (Winnicott, 1965b). The above psychoanalytic accounts of the development of self-concept suggest that the current theories have moved from the Freudian concept of control of instincts to the development of a stable core of selfhood through a good mother–infant relationship at the start of life.

Self as an Entity: A Cognitive Perspective

While psychoanalytic writers have provided different models in explaining the development and splitting of selfhood, the writers from cognitive science provided different models in explaining similar phenomena. Markus (1977) introduced the concept of self-schemas, which are cognitive structures that help to organize, summarize, and explain behavior in a particular sphere of life. Self-schemas serve as templates that represent connected beliefs about the self. These schemas help in the processing of information about the self through selective attention to certain information consistent with the current self-schemas and filtering out of information that are inconsistent or threatening to the self-schemas. As the current self-view depends on the combination of self-representations at any one time, it is better considered as a working self-concept (Markus and Nurius, 1986). If viewed from the framework of self-schemas, one can consider that a particular working concept is active when a particular set of self-schemas are being activated (Beck et al., 1979). If the self-concept can take on different working self-concepts at different times, it is possible that the self-concept can take on the form of 'possible selves' (Markus and Nurius, 1986). These possible selves are cognitive structures that represent concepts about what a person may become, what he or she would like to become, and what he or she is afraid of becoming. People with positive selves tend to experience more positive affect and persist longer on tasks than those with negative selves (Fluery et al., 2002). At times, possible selves may be represented by their own images being involved in certain situations or performing certain tasks. These images may have powerful effects on emotions being experienced and on increasing the likelihood to act in response to these images (Ng et al., 2013).

Possible selves may also be viewed from another perspective. Higgins (1987) suggests that people have an 'actual' self or self-concept, an 'ought' self with beliefs about obligations and personal standards, and an 'ideal' self with beliefs about personal aspirations. Studies have found that discrepancies between different selves may trigger different feelings (Strauman and Higgins, 1987). People with depression are more likely to experience actual–ideal discrepancies while those with anxiety are more likely to experience ought–ideal discrepancies. It is conceivable that these are self-concepts in

imagination. Yet studies have found that these concepts may be crystallized in the form of future-oriented (prospective) images about oneself. For example, patients with depression were found to have negative prospective images (Morina et al., 2011) while those with hypomania had positive prospective images (Gregory et al., 2010). Another related concept is self-concept clarity, the stability of self-concepts in terms of clarity of content, consistency across different situations and at different times. People with low self-concept clarity are highly susceptible to external environmental influence, so that their feelings of self-worth will be highly dependent on the outcome of social events. Indeed, studies have found that people with low self-concept clarity are more likely to be anxious, depressed, and have low self-esteem (Butzer and Kuiper, 2006). Conversely, people with frequent emotional swings are also more likely to have low self-concept clarity, as the individual may experience very different working self-concepts when excited, depressed, or anxious compared to normal mood states.

The above discussion mainly focuses on the content of the self-concept. However, the manner in which self-concepts are being organized may also have a profound impact on mood states. Showers (1992) proposed two types of self-organization: compartmentalized and integrative. In compartmentalized self-organization, positive self-beliefs and roles are segregated clearly from negative ones in the form of separate knowledge structures. There is no overlap between the attributes that are used to describe these two different aspects of the self. For example, a person with compartmentalized self-organization may describe himself or herself as being incompetent and weak as an employee but as being supportive and caring as a friend. However, a person with integrative self-organization will describe himself or herself with both positive and negative attributes. When asked about their performance as an employee, he or she might still coin negative terms like 'incompetent' or 'lazy' but also add positive terms like 'reliable' or 'loyal.' In other words, their views about themselves in different roles are more likely to be less extreme or dichotomous, with positive information being available to balance against negative information. Indeed, people with compartmentalized self-organization were found to be more likely to experience extreme mood swing in response to social stress than people with integrative self-organization (Showers, 1992). As self-images are likely to contain a collection of absolutely positive or negative attributes about oneself, people with compartmentalized self-organization may be more likely to report having intrusive self-images (Luke and Stopa, 2009). Indeed, negative self-images have been reported by people with a wide range of psychological disorders (Hackmann and Holmes, 2004). It is worth noting that this concept of self-organization has close similarities with Fairbairn's concept of split ego formations as a result of poor mothering.

Buddhist Concept of 'No Self'

As shown in the last section on psychoanalytic perspective of self as an entity, psychoanalysts strongly believe the importance of a strong ego as a key to success in interpersonal relationships and pursuit of life goals. A stable and clear self-concept is also considered in the cognitive science field as conducive to positive mental health. As knowledge arises in the context of cultural influence, it is understandable that a heavy emphasis on a stable core of selfhood as being conducive to adult maturity may be a product of Western culture (Williams, 1997). For many Buddhists, the existence of a conventional self-concept is only accepted on a relative level of reality. The self is affirmed as a nonenduring entity arising from countless mental and physical aggregates. This false self-representation is considered as being constructed through avoidance of non-pleasurable feelings and as being nourished by pleasant sensations. Such attachment to incessant cravings is considered as the cause of sufferings in life. The self is perceived as a false belief, an illusion leading to the creation of a sense of false continuity that resists the impermanence that comes with life (see the section on stability of self over time for a Western view). If the self is a construct that leads to suffering, this faulty perception is to be deconstructed and replaced with the notion of 'no self' (Michalon, 2001). Through this cultivation of a nonenduring self, an individual will perceive the interconnectedness of himself with the rest of the world and the reason for his existential suffering (Engler, 1986). The means of achieving this awareness of 'no self' is through meditation that leads to deconstruction and reorganization of one's inner ego world (Engler, 1986). This perspective is in direct contrast with the psychoanalytic and cognitive accounts that emphasize on the importance of establishing a stable, clear, and positive self-concept.

Stability of Self Over Time

The above discussion on self-concept clarity has highlighted the importance of temporal stability of self-concepts in the maintenance of personal well-being. Yet even people with high self-concept clarity will have a shift of self-concepts with time, usually in the form of a sense of growth and development of the self over time (Sedikides and Gregg, 2008). To have a sense of continuity of self across time, one will need to have an ability to access past memories. According to Conway and Pleydell-Pearce (2000), autobiographical memories are events that are recalled from the past and are stored in a self-memory system. Such a self-memory system contains these autobiographical memories and current self-relevant goals, and regulates the environmental cues required to access this knowledge base. People are motivated to enhance the self by recalling the past and predicting the future in a manner that make themselves more favorable (Ross and Wilson, 2000). People usually recall their past selves as more negative or mixed than their current self, and they view their future selves as more positive (Ross and Buehler, 2004), thereby contributing to a sense of growth and development.

A breakdown in this system is likely to lead to a failure to link the past with the present self-concepts. For example, a traumatic event may lead to an experience of vulnerability and precede the onset of post-traumatic stress disorder. The memories of the traumatic events may be used as evidence for the existence of the new self-concept as being a weak and vulnerable person. Failure to retrieve past autobiographical memories that contradict this new self-concept would add to a further sense that the old self has died (Stopa, 2009).

Furthermore, lack of access to such autobiographical memories would also reduce the ability to solve current stressors based on past learning experiences, thereby leading to the maintenance of the psychological disorder (Williams et al., 2007) and the new emerging negative self-concept. In addition, generic memories without specific details may provide general evidence that supports global and dysfunctional self-views, whereas specific, episodic memories with great details about the circumstances may help to contextualize a negative outcome as caused by a set of specific circumstances. Such autobiographical memories will then serve as convincing evidence that the negative outcome is not caused by certain global, internal, and stable self-attributes that are less amenable to changes.

While the above cognitive accounts focus on how the past information would affect the current self-concepts, one can readily discern that both psychoanalytic and cognitive science writers both placed a heavy emphasis on the importance of past experiences in shaping current self-concepts. On a different note, existential writers provide an alternative account on how future information might affect current self-concepts. A threat to the meanings in one's life in accordance with one's culture worldview can lead to various cultural worldview defenses, including harming those people that threaten these meanings and rewarding those who support these meanings, so as to protect one's self-worth. When death is becoming apparent or imminent to a person, a person will approach individuals who offer protection and comfort their life meanings and avoid those who threaten these meanings (Arndt et al., 2002). Reminding individuals with obsessive-compulsive disorder about the salience of death and enhancing their awareness of their death anxieties can exacerbate the frequency and intensity of compulsive rituals (Strachan et al., 2007). Furthermore, people with high self-concepts were found to have lower death anxiety and fewer cultural worldview defenses than people with low self-concepts (Strachan et al., 2007).

Self-Regulation and Self-Monitoring

The individual has both a self that can be observed, monitored, and evaluated (the object) and a self that does the judging (the subject). Self-regulation and self-evaluation are important processes that facilitate people to reflect upon and to monitor who they are, what they aspire to become, and what they would like to avoid becoming. There are a number of contrasting motives for self-evaluation. The first one is self-enhancement motive, which describes the wish to elicit positive information and to discard negative information about the self. The second one is the self-verification motive, which describes the search for consistency within the self. The third is the self-assessment motive, which describes the search for accurate information about the self. The fourth is the self-improvement motive, which describes the search for information that will lead to self-improvement. People tend to search for positive, consistent, accurate, and improving information from other sources (Taylor and Brown, 1988). However, when faced with different types of information, people tend to search for information that enhances their self-esteem (Sedikides, 1993). Once the motive of self-enhancement has been met, people are then more likely to be receptive to negative feedback that may help one to improve in the future (i.e., fulfilling the self-improvement motive) (Trope and Pomerantz, 1998). What is intriguing is that people with psychological disorders tend to put the self-verification motive before the self-enhancement motive when seeking out personal information. Therefore, they are more likely to seek information consistent with their negative self-views and filter out information that disconfirms their negative self-views (Giesler and Swann, 1999). The relevance of other motives of self-regulation to psychological disorders is not systematically studied so far.

Self-Affirmation and Self-Compassion

Self-affirmation is about increasing the positivity of the self, especially after a threat to self-esteem, while self-compassion is about accepting the self as it is. Affirming an important value is the most common method of self-affirmation, which is activated when the self-concept is under threat (Steele, 1988). This is consistent with the self-enhancement motive to seek out positive information to maintain high self-esteem. Such self-affirmation strategies have been found to be helpful in reducing depressive ruminations and increasing positive mood (Koole et al., 1999). On the other hand, self-compassion is concerned about accepting oneself for one's faults and failures and viewing such experiences as normal human experiences necessary for growth and maturation. It is considered as having three components: self-kindness, mindfulness, and common humanity (Neff, 2003). Rather than focusing on positive aspects of the self, self-compassion focuses on kind and compassionate acceptance of self as a human being with flaws and limitations. People with self-compassion are less likely to be self-critical, anxious, or depressed, and are more likely to be associated with greater life satisfaction (Neff et al., 2007).

Self-Awareness

Another aspect of the self being the subject of observation of the object of self is the concept of self-awareness. It refers to the concept that one is aware of the independence of self from others and of having independent thoughts, feelings, and behaviors from others in this world. Objective self-awareness is defined as the tendency to focus on the self and compare the self against a standard (Duval and Wicklund, 1972). The standard is a representation of what constitutes the right attitude or behavior. This concept is similar to the actual-ought selves discrepancy (Higgins, 1987) and, to a certain extent, the structural model of anxiety as proposed by Freud (Freud, 1957), except that there is an emphasis on triggering a restorative mechanism of consistency through changing attitudes, behaviors, or standards. If the inconsistency cannot be resolved, there is a vicious cycle of increasing focusing on internal states, decreasing attention from external cues for problem solving or seeking disconfirmatory evidence, and increases avoidant behaviors (Wells and Mathews, 1994). In support of the dysfunctional nature of excessive self-focus, heightened self-awareness has been found in people with depression (Schwartz, 2001) and anxiety disorders (Bogels and Mansell, 2004). People with negative self-evaluations

(and negative self-beliefs) would have exacerbations in negative affect when they increased their self-awareness (Figurski, 1992), as enhanced self-awareness would lead to increased attention toward body states, negative emotions, and avoidant behaviors. The role of heightened self-awareness in maintenance of psychological disorders possibly explains why mindfulness and self-compassion (in which self-compassion includes mindfulness and commonality of humanity) are beneficial to mental health.

The Self through Interaction with Others

The self does not develop in isolation. People define themselves through interactions with other people at different stages of life. During childhood, the self is defined through interactions with family members, classmates, and teachers at school. In the above section on 'self as an entity,' the psychoanalytic writers of object-relations school are the scholars who wrote extensively about the importance of early object relations in the development of personality and selfhood. John Bowlby, another child analyst, suggests that early attachment experiences (e.g., parental care) influence the development of internal working models, which are important variables in subsequent social behavior throughout the life span (Bowlby, 1980). These internal working models are complex emotional-cognitive schemas that contain feelings of self-worth and beliefs about the availability and responsiveness of important attachment figures, which guide subsequent behaviors. If a child has received consistent and sensitive care from parents, they are likely to develop a positive internal working model of high self-worth and of perceiving significant others as reliable and trustworthy. However, if a child has received inconsistent care, neglect or even abuse, he or she will view himself or herself as unworthy and significant others as unreliable or trustworthy (i.e., a negative internal working model). Hazan and Shaver (1987) has hypothesized that internal working models have direct impact on subsequent adult romantic relationships. Bartholomew and Horowitz (1991) have delineated four different attachment styles: secure, preoccupied, fearful, and dismissive. The latter three styles were also considered as insecure styles. People with a preoccupied attachment style have a view of unworthy self but a positive view of others. People with a fearful attachment style have a view of unworthy self and a negative view of others. Finally, people with a dismissive attachment style have a positive view of worthy self but a negative view of others. An insecure attachment style has been associated with anxiety disorder, bipolar disorder, various personality disorders, and schizophrenia (Dozier et al., 2009). The internal working model is in fact quite similar to the self-schemas with associated emotions, beliefs, attitudes, and behaviors as mentioned in the previous section.

Apart from this interpersonal model of interactions of self with others, a more direct model of social interaction with others in the formation of self-concept is provided by social comparison theory (Festinger, 1954). Festinger believed that people would compare themselves with others when there was inadequate information to understand about their own personal abilities. Depending on the intimacy of the targets of social comparison and the importance of tasks in question, different affects could be elicited through social comparisons. For example, people feel worse about themselves when they are outperformed by someone close to them on a task or an attribute that is considered as important to them (Beach and Tesser, 2000). Individuals with low self-esteem tend to engage in social comparisons when comparisons are unfavorable, thereby maintaining their negative self-concepts about themselves (Santor and Yazbek, 2006). People with depression are more likely to compare with people who outperform themselves (i.e., upward comparison) (Butzer and Kuiper, 2006). Furthermore, the domains of performance selected for social comparison depend on the personality of the individuals and the perceived importance of the domains in attracting the interest of others (Thwaites and Dagnan, 2004). For example, dysphoric individuals with autonomous personality are likely to compare their achievements with others (Wood and Lockwood, 1999). In addition, people with anxiety disorders tend to make upward rather than downward social comparisons (Antony et al., 2005).

Conclusions

The psychology of self is unique in that the self can act as both the subject and the object. The self as a subject can observe, monitor, and evaluate the self as an object. Both the content of self-concepts and the processes of organization are key areas that require further research and understanding. Current understanding of psychological disorders appears to focus on the contents of cognitions and cognitive processes leading to negative affect and maladaptive behaviors. Yet the current models of the entity of self and the interactions of self with others provide testable hypotheses for understanding of the contribution of these models to the onset and maintenance of psychological disorders.

See also: Borderline (Patient) Personality; Hysteria: History and Critiques; Normality and Mental Health; Personality and Self; Post-Traumatic Stress Disorder; Psychiatry and Religion; Psychoanalysis: Current Status; Self and Brain; Self-Concepts: Educational Aspects; Social Work in Eastern Europe; Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts.

Bibliography

Antony, M.M., Rowa, K., Liss, A., Swallow, S.R., Swinson, R.P., 2005. Social comparison processes in social phobia. Behavior Therapy 36, 65–75.

Arndt, J., Greenberg, J., Schimel, J., Pyszczynski, T., Solomon, S., 2002. To belong or not to belong, that is the question: terror management and identification with gender and ethnicity. Journal of Personality and Social Psychology 83, 26–43.

Bartholomew, K., Horowitz, L.M., 1991. Attachment styles among young adults: a test of four-category model. Journal of Personality and Social Psychology 61, 226–244.

Beach, S.R., Tesser, A., 2000. Self-evaluation maintenance and evolution: some speculative notes. In: Suls, J., Wheeler, L. (Eds.), Handbook of Social Comparison: Theory and Research. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 123–140.

Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G., 1979. Cognitive Therapy of Depression. Penguin Books, New York, USA.

Bogels, S.M., Mansell, W., 2004. Attention process in the maintenance and treatment of social phobia: hypervigilance, avoidance and self-focused attention. Clinical Psychology Review 24, 827–856.

Bowlby, J., 1980. Attachment and Loss. In: Loss, Sadness and Depression, vol. 3. Hogarth Press, London.

Butzer, B., Kuiper, N.A., 2006. Relationships between the frequency of social comparisons and self-concept clarity, intolerance of uncertainty, anxiety, and depression. Personality and Individual Differences 41, 167–176.

Conway, M.A., Pleydell-Pearce, C.W., 2000. The construction of autobiographical memories in the self-memory system. Psychological Review 107, 261–288.

Dozier, M., Stovall, K.C., Albus, K.E., 2009. Attachment and psychopathology in adulthood. In: Cassidy, J., Shaver, P.R. (Eds.), Handbook of Attachment: Theory, Research and Clinical Applications. Guildford Press, New York, pp. 497–519.

Duval, S., Wicklund, R.A., 1972. A Theory of Objective Self-awareness. Academic Press, Oxford, UK.

Engler, J., 1986. Therapeutic aims in psychotherapy and meditation. In: Wilber, K., Engler, J. (Eds.), Transformation of Consciousness. Shambala, Boston, pp. 384–390.

Fairbairn, W.R.D., 1954. An Object-Relations Theory of Personality. Basic Books, New York.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

Figurski, T.J., 1992. Everyday self-awareness: implications for self-esteem, depression, and resistance to therapy. In: deVries, M.W. (Ed.), The Experience of Psychopathology: Investigating Mental Disorders in Natural Settings. Cambridge University Press, New York, pp. 304–313.

Fluery, J., Sedikides, C., Donovan, K.D., 2002. Possible health selves of older African Americans: towards increasing the effectiveness of health promotion efforts. Topics in Geriatric Rehabilitation 18, 52–58.

Freud, S., 1949. In: Strachey, J. (Ed.), The Outline of Psychoanalysis. The Hogarth Press, London.

Freud, S., 1957. Beyond the Pleasure Principle. International Universities Press, New York.

Freud, S., 1959. On the sexual theories of children. In: Freud, S. (Ed.), The Complete Psychological Works (Vol IX, 1906-1908): Jensen's 'Gradiva' and Other Works, pp. 205–226.

Giesler, R.B., Swann, W.B.J., 1999. Striving for confirmation: the role of self-verification in depression. In: Joiner, T., Coyne, J. (Eds.), The Interactional Nature of Depression: Advances in Interpersonal Approaches. American Psychological Association, Washington, DC, pp. 189–217.

Gregory, J.D., Brewin, C., Mansell, W., Donaldson, C., 2010. Intrusive memories and images in bipolar disorder. Behaviour Research and Therapy 48, 698–703.

Hackmann, A., Holmes, E., 2004. Reflecting on imagery: a clinical perspective and overview of the special issue of memory on mental imagery and memory in psychopathology. Memory 12, 389–402.

Hazan, C., Shaver, P., 1987. Romantic love conceptualized as an attachment process. Journal of Personality and Social Psychology 52, 511–524.

Higgins, E.T., 1987. Self-discrepancy: a theory relating self and affect. Psychological Review 94, 319–340.

Koole, S.L., Smeets, K., van Knippenberg, A., Dijksterhuis, A., 1999. The cessation of rumination through self-affirmation. Journal of Personality and Social Psychology 77, 111–125.

Luke, M., Stopa, L., 2009. Psychological theories of the self. In: Stopa, L. (Ed.), Imagery and the Threatened Self. Perspectives on Mental Imagery and the Self in Cognitive Therapy. Routledge, Hove, East Sussex, pp. 15–64.

Markus, H., 1977. Self-schemata and processing of information about the self. Journal of Personality and Social Psychology 35, 63–78.

Markus, H., Nurius, P., 1986. Possible selves. American Psychologist 41, 954–956.

Michalon, M., 2001. "Selflessness" in the service of the ego. American Journal of Psychotherapy 55, 202–219.

Morina, N., Deeprose, C., Pusowski, C., Schmid, M., Holmes, E., 2011. Prospective mental imagery in patients with major depressive disorder or anxiety disorders. Journal of Anxiety Disorders 25, 1032–1037.

Neff, K.D., 2003. Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. Self & Identity 2, 85–101.

Neff, K.D., Rude, S.S., Kirkpatrick, K.L., 2007. Self-compassion and its link to adaptive psychological functioning. Journal of Research in Personality 41, 139–154.

Ng, R.M.K., Krans, J., Holmes, E., 2013. Mental imagery and psychopathology: examples of post-traumatic stress disorder and bipolar disorder. In: Lacey, S., Lawson, R. (Eds.), Multi-sensory Imagery. Springer, New York, pp. 365–384.

Ross, M., Buehler, R., 2004. Identity through time: constructing personal pasts and futures. In: Brewer, M.B., Hewstone, M. (Eds.), Self and Social Identity. Blackwell Publishing, Malden, MA, pp. 25–51.

Ross, M., Wilson, A.E., 2000. Constructing and appraising past selves. In: Schacter, D.L., Scarry, E. (Eds.), Memory, Brain, and Belief. Harvard University Press, Cambridge, MA, pp. 231–258.

Santor, D.A., Yazbek, A.A., 2006. Soliciting unfavorable social comparison: effects of self-criticism. Personality and Individual Differences 40, 545–556.

Schwartz, R.C., 2001. Self-awareness in schizophrenia: its relationship to depressive symptomatology and broad psychiatric impairments. Journal of Nervous and Mental Disease 189, 401–463.

Sedikides, C., 1993. Assessment, enhancement, and verification determinants of the self-evaluation process. Journal of Personality and Social Psychology 65, 317–338.

Sedikides, C., Gregg, A., 2008. Self-enhancement: food for thought. Perspectives on Psychological Science 3, 102–116.

Segal, H., 1964. Introduction to the Work of Melanie Klein. Heinemann Medical Books, London.

Showers, C.J., 1992. Compartmentalization of positive and negative self-knowledge: keeping bad apples out of the bunch. Journal of Personality and Social Psychology 62, 1036–1049.

Steele, C.M., 1988. The psychology of self-affirmation: sustaining the integrity of the self. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, Social Psychological Studies of the Self: Perspectives and Programs, vol. 21. Academic Press, San Diego, pp. 261–302.

Stopa, L., 2009. Re-conceptualizing the self. Cognitive and Behavioral Practice 16 (2), 142–148.

Strachan, E., Schimel, J., Arndt, J., Williams, T., Solomon, S., Pyszczynski, T., et al., 2007. Terror management: evidence that mortality salience exacerbates phobic and compulsive behaviors. Personality and Social Psychology Bulletin 33, 1137–1151.

Strauman, T.J., Higgins, E.T., 1987. Automatic activation of self-discrepancies and emotional syndromes: when cognitive structures influence affect. Journal of Personality and Social Psychology 53, 1004–1014.

Sullivan, H.S., 1968. The Interpersonal Theory of Psychiatry. The Hogarth Press, London.

Taylor, S.E., Brown, J., 1988. Illusion and well-being: a social psychological perspective on mental health. Psychological Bulletin 103, 193–210.

Thwaites, R., Dagnan, D., 2004. Moderating variables in the relationship between social comparison and depression, an evolutionary perspective. Psychology and Psychotherapy: Theory, Research and Practice 77 (3), 309–323.

Trope, Y., Pomerantz, E.M., 1998. Resolving conflicts among self-evaluative motives: positive experiences as a resource for overcoming defensiveness. Motivation & Emotion 22, 53–72.

Wells, A., Mathews, G., 1994. Attention and Emotion: A Clinical Perspective. Psychology Press, Hove, UK.

Williams, P., 1997. Nina, Anatta, and no-self. British Journal of Psychotherapy 14, 527–530.

Williams, J.M., Barnhofer, T., Crane, C., Herman, D., Raes, F., Watkins, E., et al., 2007. Autobiographical memory specificity and emotional disorder. Psychological Bulletin 133, 122–148.

Winnicott, D.W., 1965a. The Family and Individual Development. Basic Books, New York.

Winnicott, D.W., 1965b. The Capacity to Be Alone, the Maturational Process and the Facilitating Environment. The International Psychoanalytical Library, International Universities Press, New York.

Wood, J.V., Lockwood, P., 1999. Social comparisons in dysphoric and low self-esteem people. In: Kowlaski, R., Leary, M.R. (Eds.), The Social Psychology of Emotional and Behavioral Problems: Interface of Social and Clinical Psychology. American Psychological Association, Washington, DC, pp. 97–135.

Psychology, History of (Early Period)

Tracy B Henley, Texas A & M University – Commerce, Commerce, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.G. Ash, volume 18, pp. 12399–12405, © 2001, Elsevier Ltd.

Abstract

From the Pre-Socratic Greeks to Wundt's day, the history of psychology has followed shifts in epistemology. Are the senses to be trusted? Is rational analysis more reliable than experience? Such questions have shaped the conduct of science, and relate to psychology in that way, but are also intrinsically psychological. This article traces the evolution of epistemology from Ancient Greece and Rome into Medieval Europe, then the Renaissance, and on to Descartes. From there, both Empiricism and Rationalism are detailed, and their connection to the advent of academic psychology is discussed.

The history of psychology is at least as old as the history of humankind. Nicholas Humphrey (1983) notes that earliest humans not only sought to understand, predict, and control nature, but also each other. Indeed, the survival of early humans was tied to their ability to observe and then anticipate what their comrades were feeling and intending: What in modern parlance we call Theory of Mind. This though, is by definition, prehistorical. History begins when we have enough artifacts to extract a theory or practice. Psychology then, like most other academic disciplines, typically starts its history with the Pre-Socratic Greeks.

Pre-Socratic Philosophy

The relevance of the Pre-Socratic Greeks, indeed of much that will be covered herein is found in the evolution of epistemology. Are the senses to be trusted? Is rational analysis more reliable than experience? Such questions will ultimately shape the conduct of science, and relate to psychology in that way. These questions also invited speculation about perception and reasoning itself, and as such are intrinsically psychological.

Thales (*c*.625–547 BC) has been called the 'first philosopher,' though he would have pointed to mathematicians and physicians in Egypt and Babylon as providing him a foundation. For Thales (and us today), the world was made up of a finite number of elements. This begged the question of which element might be the most primary or fundamental. Using both natural observation and reason, Thales concluded it was water – as all life depends on water. Although Thales is rightly heralded for giving a data-based and rational, as opposed to a mystical account, it was not long before his literal analysis of the matter came to be challenged.

Heraclitus (*c*.540–480 BC) argued that it was neither water, nor air as others had asserted, but fire that was the most fundamental element. Unlike Thales, who had pointed objectively to the water in things, Heraclitus reasoned that all things change, and that fire as the essence of change itself, was thus the most primary of the elements. Since all things change, even those that exhibited no observable fire still shared in its essential nature. This foreshadows much that will follow in epistemology, as it calls into question the value of empirical observation (from our limited senses) as opposed to rational analysis.

Pythagoras (*c*.580–500 BC), like Heraclitus, favored reason and trusted its revelations about reality over what the senses observed. Pythagoras held that the *logos* (as in logical structure) for everything in the cosmos was found in mathematics, and in numerical *ratios* (the root of the Pythagorean's rationality). Practical demonstrations using geometry and music showed that true relationships could be determined through mathematical analysis far more accurately than from empirical observation. Although it is a bold statement, it can be argued that real roots of modern scientific psychology began here with Pythagoras. To the degree that we have confidence (at least at $p < .01$) in what our statistical findings and models can tell us about the general principles of human nature, we are following a tradition that begins with Pythagoras and moves forward across the span of Western civilization. That is, scientists trust the results of their statistics and mathematical models even when they extend beyond the limits of our direct sensory capabilities.

The Golden Age of Greece

Pythagorean ideas were adopted by the later Greeks. The entrance to Plato's Academy stated "Let no one without an understanding of mathematics enter here," and by that he was endorsing the Pythagorean approach. Alcmaeon, usually seen as the first to offer a 'scientifically' grounded approach to medicine, was also a Pythagorean. And building upon the tradition of Alcmaeon was the greatest of Greek physicians – Hippocrates (*c*.460–377 BC). Indeed, Hippocrates was such a great figure that for more than thousand years works written by others would be credited to him so they would be accepted as authoritative. This leads historians to speak more often about Hippocratic writings than the writings of Hippocrates.

Hippocratic writings held that human nature was controlled by chemicals, and that health (both mental and physical) was a matter of keeping these chemicals in proper ratio. The modern treatment of depression and anxiety by medications that increase or block various neurotransmitters is the exact same idea. Even aspects of the ethical code endorsed by contemporary clinical psychologists can first be seen in the Hippocratic oath: "I will not abuse my position to indulge in sexual contacts with the bodies of women or of men… Whatever I see or hear,

professionally or privately, which ought not to be divulged, I will keep secret and tell no one."

Socrates (*c*.470–399 BC) is seen as a major transitional figure in the history of Western philosophy. During Socrates' day Athenians reveled in 'philosophical disputation' of every kind – from human nature to natural science to politics – these Greeks loved an academic argument. In turn, this love of debate afforded an entire class of what is perhaps best understood as private tutors called the Sophists.

In addition to speculative philosophy (most often a form of Skepticism – or the belief that objective truth is often unknowable), the Sophists taught logic, politic, and rhetoric to those early advocates of democracy, serving very much the role of a modern political (or advertising) campaign manager. Indeed, it was their willingness to sell their knowledge and expertise, at times to the highest bidder in a political debate (instead of backing the side that was 'right') that, although consistent with their Skepticism, has earned them derision.

Socrates differed from most of the other Athenian Sophists in that he sought to find the universals or essences that formed the basis of our concepts: What do all beautiful objects have in common? Even such a fuzzy and subjective concept as beauty could be defined then using inductive reasoning. Once so defined, philosophers could also communicate with one another objectively about the concept, which challenged one of the central assumptions of Skepticism.

Like the Sophists, Socrates understood that knowledge meant power. But, Socrates advocated self-exploration as a variant of the day's popular skeptical philosophy. He taught that the very purpose of life was to gain knowledge and from that, personal wisdom. Socrates most famous student, and erstwhile chronicler, was Plato (*c*.427–347 BC).

Like the Pythagoreans, Plato understood that although mathematical relationships were abstractions that could not be experienced through the senses, they were nevertheless real and provided the 'schema' (logical form) for everything in the empirical world. This, in a nutshell, is his Theory of Forms. The thousands of birds that one encounters with the bodily senses are but empirical instances of an ideal bird form that exists in the abstract. For Socrates, essences were merely defining features, but Plato's forms had an actual existence separate from their manifold worldly occurrences.

In his Allegory of the Cave, Plato describes prisoners who have lived their entire lives in the depths of a cave as a metaphor to show the relationship between our knowledge and the forms. Above and behind the prisoners is an unseen path, and as the masters pass by with various objects shadows are cast onto the wall in view of the prisoners. These shadows are all the prisoners know of the world, and constitute the reality given by their senses. Imagine now that one of the prisoners was to escape, and see then the path, people, and objects for what they 'really' were instead of mere shadowy reflections.

Holding the allegory in abeyance, exactly how does one come to know the forms if not through sensory experience? The answer must be that such knowledge is innate, and as such Plato was a nativist as well as a protorationalist. Plato's ideas as sketched here were revitalized in the heyday of cognitive psychology through Eleanor Rosch's work on conceptual categories and Noam Chomsky's idea of an inborn deep grammatical structure. Indeed, when more than 2000 years later, Alfred North Whitehead quipped that the history of Western philosophy is but "a series of footnotes to Plato" he was affirming that Plato had advanced the ideas of the Pythagoreans that would come to shape Western civilization's continued fascination with scientific understanding.

Aristotle's (*c*.384–322 BC) father was the physician of the Macedonian king, and in time Aristotle himself would tutor the Macedonian prince that would grow up to be Alexander the Great. Aristotle studied for many years with Plato in Athens, but also traveled extensively. Around age 50, Aristotle returned to Athens and soon took over one of its famed schools, The Lyceum. With a large staff, library, museums, and 'laboratories' it certainly resembled a modern university. Perhaps more accurately, it could be said that when 'modern' universities were being developed in Europe, they followed an Aristotelian model.

The most amazing thing about Aristotle is the scope of his work. His writings on logic, metaphysics, epistemology, ontology, ethics, and aesthetics are still taught at face value, and in some cases continue to represent philosophy's best understanding of these matters. His thoughts on language, rhetoric, and poetics essentially shape how Western civilization composes its written works (from dramatic narrative to technical argument). Given his substantial and enduring contributions to natural science, he is rightfully viewed as the patriarch of such diverse modern studies as optics, physics, biology, botany, taxonomy, and geology (among others). Arguably, his impact on medicine, theology, and political science has been greater still. And yet, perhaps the contemporary field that he most clearly anticipated was psychology. Aristotle wrote about the physiology of our sense organs, sensation and perception, consciousness, sleep and dreams, learning, memory, language, reasoning, executive functioning, emotion and motivation, development, morality, pedagogy, pathology, and social behavior. Indeed, as one flips through the chapters of an introductory textbook there is little therein that Aristotle did not address.

Of all these topics, his postulation of laws of association is clearly the most important. It was Aristotle who first understood that recall occurred in accord with principles of contiguity (when we think of a thing, we also think of things that were experienced with it) and similarity (when we think of a thing, we also think of things that are like it). Indirectly then, Aristotle explains learning as a function of the frequency of these associations. From the British Empiricists (below) to the first academic psychologists (e.g., Alexander Bain, William James, Wilhelm Wundt), on to American Behaviorism and even modern day neural networks, these insights would explain learning and memory evermore.

It is tempting to describe Plato as the 'grandfather' of rationalism, and Aristotle as the 'grandfather' of empiricism. Aristotle did believe in empirical information as a source of data, but also that such sensory information was to be used as the basis for rational analysis. As Immanuel Kant might have understood him, for Aristotle sensory experience was necessary, but not sufficient, for generating knowledge. For Aristotle (like Thales long before him) knowledge came from the application of reason to empirical data.

Considerations of Ancient Greece typically conclude with a look at the sorts of 'folk psychology,' or everyday philosophical

beliefs that were widely endorsed. Many different –ism's existed, but Epicureanism (from Epicurus, *c*.341–270 BC) can serve as example. This was a materialistic theory emergent from the Greek's understanding that everything in the cosmos was composed of atomic particles. As such, there was no immaterial soul that existed after death. Since there was no afterlife, the goal was to make the most of this one, and that led to a clever version of hedonism.

Epicurus wrote "I am thrilled with pleasure in the body when I live on bread and water, and I spit on luxurious pleasures, not for their own sake, but because of the inconveniences that follow them." Hedonism is the pursuit of pleasure, but the Epicurean insight was that absence of pain was the greatest of pleasures. Stated more bluntly, *overall* there is more pleasure to be had from avoiding the hangover, the bloating, and genital itching than there is from such 'pleasures' as drunkenness, gluttony, and risky sex – followed then by the hangover, indigestion, and all that scratching.

Rome

Popular culture tends to treat all parts of the Ancient World the same. In truth, the sociology (or a day in the life) of Golden Age Greece could not have been more different from Rome after Caesar. The Greeks flourished in the anarchy of democracy, but the Romans prized law and order. Whereas the Greeks valued speculative philosophy, the Romans found it useless. Rome was so fiercely pragmatic that by the height of the Empire even Plato would have been all but unknown. What the Roman's valued they excelled at – their roads, aqueducts, sewers, and architecture are truly marvelous even when set against modern engineering efforts. The battle plans and field medicine that drove the Legions are still in use today, after some 1500 years of warfare.

Roman education and philosophy was no less practical. The goal was productive citizens, efficient cogs in the machine that was Rome. The best example of this comes from Marcus Aurelius' *Meditations* (written just before AD 180). This masterwork of Stoic philosophy came to define a good Roman citizen. Good Romans were expected to accept their lot in life and be dutiful without question: "To live each day as though one's last, never flustered, never apathetic, never attitudinizing – here is the perfection of character." Pleasure came in realizing that one had done their best, and was part of a grander plan, even if that plan was unknown or inconvenient to the individual. Pain comes from reaching beyond one's station, from worrying about things beyond one's control.

In 312, just before the Battle of the Milvian Bridge, the emperor Constantine is said to have had a vision that changed history: the Christian cross in the sky, accompanied by the message "By this sign you shall conquer." Christianity soon became a tolerated, then favored, and within 75 years, the official religion in Rome. From Rome to Europe, from Europe to the World, Christian ideals have certainly influenced how Western civilization thinks about virtually everything, including psychology.

To the extent that there is an actual Christian psychology, it is to be found initially in the New Testament writings of Paul (*c*.10–64, and himself a Roman citizen). Soon however, church figures such as Augustine (*c*.354–430) would become the most significant thinkers for the next 1000 years.

The Middle Age

There is no one date that historians agree to be the start of the Middle Age. From the sack of Rome by the Visigoths in 410 up through events in the early 500s, several dates are offered. Historian generally do agree that the Middle Age was the confluence of three things: (1) the vast infrastructure of roads and cities built by the Roman Empire; (2) the ideas and ideals of Christianity; and (3) the largely Germanic peoples that now controlled what would become Europe. Indeed, the Middle Age is roughly the next 1000 years, during which Europe is formed.

The beginning of this medieval period is also referred to as the Dark Ages, as following the fall of Rome the 'light of knowledge' grew dim in what would become Europe. Matters begin to improve, albeit slowly, following the creation of Charlemagne's Palatine School in the late 700s. Nevertheless, education remained limited to priests, monks, and the nobility for generations. Indeed, the Middle Ages are often skipped over within academic histories, viewed as an anti-intellectual time utterly dominated by the Church. This is an unfortunate oversimplification, as surely the legacy of Roman pragmatism contributed as much to the absence of speculative thought as did the Church itself.

Just after AD 1000, a new era in which philosophical reason and theological faith are united begins. The name most associated with the start of this movement is Anselm (*c*.1033–1109), who felt that secular logic could be used to support religious beliefs. Such ideas are generally known as *scholasticism*, and one of the important early scholastics was Peter Abelard (*c*.1079–1142). Within philosophy Abelard is best associated with the 'realism versus nominalism' debate, a consideration akin to that which we introduced with Socrates and Plato concerning the nature and universality of the essences that define our concepts. Concurrent with scholasticism was the development of schools and even major universities (Abelard lectured at the cathedral school in Paris); and also, the Crusades.

Whereas Greek thought had barely survived Rome and the Dark Ages, its preservation and study had continued with far less disruption in the Middle East – lands now under sway of Islam. Via the Crusades, works by Aristotle and a variety of other Greeks made their way into those fledgling European universities. Additionally, there came new innovations in science and mathematics penned by Islamic scholars, as well as detailed analyses of Aristotle's vast psychology by thinkers such as Al-Kindi, Avicenna, and Averroes.

Scholasticism would come to be associated with the renewed study of Aristotle, and following the model of Islamic scholars, the integration of Aristotle's philosophy (and psychology) with theology. Such was the life's work of Albert Magnus (*c*.1200–1280), and his student Thomas Aquinas (*c*.1225–1274), the most heralded of the scholastics.

The latter part of the Middle Age is filled with colorful figures who would anticipate much of interest to modern psychology. From William of Occam (*c*.1285–1349) we get his razor, or admonition that explanations should always be kept as simple as possible and devoid of extraneous

assumptions. This 'law of parsimony' is now seen as a cornerstone of science, and was a guiding principle of early animal researchers such as Conwy Lloyd Morgan and the ethologists. And as a final example, Ramon Llull (*c*.1232–1315) sought to mechanize Aristotelian logic. As such, he is surely the father of computational approaches to generating knowledge, and perhaps even of computer science (see Bonner, 2007).

The Renaissance

Renaissance means 'rebirth,' and was a time (from the mid-1400s to the early 1600s) when philosophical thought in Europe gradually switched from being God-oriented to being human-oriented. It was concurrently a rebirth of objective inquiry and a rejection of authority. It was fueled by many convergent factors – such as the explorations of Asia and the New World, the advent of movable type printing in the West (it had long been known in China), the continued 'rediscovery' of Western civilization's past (Greece and Rome), and the rise of a new movement commonly known as 'humanism.'

Humanism is characterized by its focus on the individual as opposed to an establishment (such as the Church, the Guild, the State, even Aristotle). Key figures in the intellectual history of the Renaissance include Leonardo da Vinci (1452–1519), Niccolo Machiavelli (1469–1527), Desiderius Erasmus (1466–1536), Martin Luther (1483–1546), Juan Luis Vives (1492–1540), and William Shakespeare (1564–1616), all of whom would have an impact (direct or indirect) on psychology.

In addition to philosophy, the Renaissance afforded a resurgence in science. It was the age of Nicolaus Copernicus (1473–1543), Galileo Galilei (1564–1642), Johannes Kepler (1571–1630), and Isaac Newton (1642–1727). Although a general foundation for scientific inquiry was provided by these giants, the figure most aligned with psychology was Francis Bacon (1561–1626). A witty and clever critic of Aristotle, he shared the man's belief that nature could best be understood by studying it directly. Whereas Galileo (in the Pythagorean tradition) advocated the discovery of general principles that could be formulated mathematically and from which deductions about nature could be made, for Bacon science should be based only on the facts of observation. Bacon feared that theory-laden observations would be biased, and this led to his very psychological consideration of our idols – or those aspects of our nature and experience that color our perceptions and cognitions.

The idols of the cave refer to how our knowledge influences our perceptions. When a child and a divinity student are told a simple Biblical parable, do they really hear the same thing? The idols of the tribe are those biases inherent in human nature, what the modern philosopher Ludwig Wittgenstein would call our 'form of life,' and why, as he claimed, "If a lion could speak, we could not understand him." The idols of the marketplace are the confounds to reasoning introduced by communication itself. Elizabeth Loftus and John Palmer (1974) famously reported that subjects estimated the speed at which two cars struck very differently if prompted by "how fast were they going when they smashed" versus "how fast were they going when they bumped." And, the idols of the theater concern the influence of our allegiances. Albert Hastorf and Hadley Cantril (1954) had students from Princeton and Dartmouth watch clips from a football game between those teams, then answer questions about controversial plays. Results correlated with which school the subjects attended.

Bacon believed in a 'better life through science' and may have been the first to write that "knowledge is power." His approach is seen as an important forerunner to positivism (considered below), and was consciously accepted by B.F. Skinner (1956) and his behavioral followers.

Descartes

Rene Descartes (1596–1650) is typically seen as the transitional figure between the Renaissance and what is called the 'Modern' era in intellectual history. Modern does not exactly mean contemporary, but more that the era's questions and their answers are still seen as relevant to our time.

Famed as a mathematician (e.g., the Cartesian coordinate system) and as a psychological philosopher (e.g., *Cogito ergo sum*), Descartes was also a military engineer and amateur biologist. Although much in his philosophy is still considered both valuable and 'modern,' for psychology there has perhaps never been a more important question than the mind–body problem.

The story often begins with Descartes' interest in the mechanical automatons at St. Germain. In the gardens there, the statues of animals and mythological creatures were animated as onlookers unknowingly triggered elaborate hydraulic and pneumatic systems by their own movements. The amazing result was sounds and actions that seemed both lifelike and situationally appropriate.

Understanding that the automatons shared much with human anatomy – an articulate skeleton, a sort of musculature, and an extensive network of tubings to convey liquids, Descartes pondered what the difference really was? Were we not just biological machines of the same sort? And like the automata, might not our behaviors just be programmed reflexive responses resulting from environmental stimulation?

Descartes came to believe that all animal behavior (and most of our own behavior and cognitive activity) could be explained mechanically. He was compelled however to view at least part of the mind (and his Catholic soul) as immaterial, and as such not part of the physical body. That then is *dualism*, or the belief that the mind and body are separate entities. But given that, what is the relationship between this incorporeal mind and the material form? That alone would get us a mind–body problem, but for Descartes it was worse still, as he also imagined that the immaterial mind could causally influence the corporeal body. How then could that happen? How could a thing without substance act on the physical world?

Descartes offers some thoughts on resolving the problem (including that the pineal gland might somehow afford this metaphysical interaction), but not a satisfactory resolution. He is famous then for giving us the problem, and what a fine problem it is, as no universally accepted to solution (that still supports dualism) has yet been discovered.

In 1649, Descartes assumed a position as tutor and court philosopher to Queen Christina of Sweden. Within 6 months

he fell ill and died. Sixteen years later it was decided that his remains should be returned to France. Somewhat ironically (given his mind–body problem), the coffin was too short, so the skull was removed to be shipped separately. Although the body was interred in Paris, the head was stolen and spent over 150 years in various private collections before reaching the Academie des Sciences (Boakes, 1984).

Empiricism

Although it is anachronistic to say that the epistemological debate between Empiricism and Rationalism goes back to the Greeks, a tension between the place of knowledge gained by experience and knowledge given innately or gained by reason certainly does. One side of the debate would come to be best expressed by a series of philosophers known as the British Empiricists, the British Associationists, as well as related French thinkers. This tradition will eventually lead to Bain, (aspects of) Wundt, and Edward B. Titchener, as well as American Behaviorism.

Bacon (above) has been suggested as the first British Empiricist, and Thomas Hobbes (1588–1679) analysis of human nature is also often considered a good starting point for empirical 'psychology' given his interest in such topics as instinct, appetites, attention, and thought. Nevertheless, it is John Locke (1632–1704) who provides us with the most complete view of this position.

Locke was a physician by training, but likely thought of himself as a political scientist. For Locke, to understand the optimal political system meant to understand man, most especially the origins of what he claimed to know. Locke strongly disagreed with Descartes that certain ideas were innate, arguing that if true, then all men should reason the same. For Locke, the alternative was that we were born 'as white paper' and that experience writes on us. For Locke, ideas came to us passively, from sensation. As such, our minds automatically take in (as ideas) the sensations that experience provides. Locke also believed that the mind actively reflected on those ideas, which explained reasoning, memory, etc. That is, although Locke did not believe in any innate ideas (only those acquired passively from sensation), he did accept intrinsic and active mental faculties.

Simple ideas from sensation could through the reflective processes become more complex ideas. Likewise, Locke believed that ideas could be associated with the experience of pleasure and pain, and thus emotions such as fear could be explained (sensing fire is associated with pain, hence I come to be afraid of fire). Locke (following Aristotle) also understood that some ideas are naturally associated based on contiguity and similarity, and was even aware that ideas could come to be spuriously associated. In fact, he clearly anticipates several points from modern behaviorism, including conditioned aversions and superstitious learning, even outlining a 'behavioral therapy' for removing such.

George Berkeley (1685–1753) and David Hume (1711–76) are usually portrayed as extending Locke's work on sensation and association, as well as bolstering it philosophically. Berkeley's most psychological contribution came in showing how complex perceptions arise from basic sensations. Hume starts with the same assumption as Locke – that understanding human mind is a prerequisite for the study of politics (and science itself), but sought to enrich Locke's system by adding a consideration of causality and expanding on the role of association.

Hume wrote "There is no relation which produces a stronger connexion in the fancy, and makes one idea more readily recall another, than the relation of cause and effect betwixt their objects." It was certainly not Hume's aim to deny that cause and effect exist in nature, but it was his goal to question how we *know* these relationships. We see one billiard ball roll across the felt and strike another, which now begins to move. In fact, we may have seen this many times. Naturally we infer cause from what we have seen and that we have seen it repeatedly. But upon reflection, we surely did not *see* a cause (transfer of force), merely a conjunction of events. It is only by belief and custom that we accept what we have seen as causal.

The major British associationists typically include David Hartley (1705–57) as well as James (1773–1836) and his son, John Stuart, Mill (1806–73). The associationists (as the name suggests) sought to further analyze the associative processes of the mind – culminating in what can be called a 'mental chemistry,' and advancing empirical philosophy ever nearer to the advent of an autonomous academic psychology.

In 1749, Hartley published *Observations on Man, His Frame, His Duty, and His Expectations* which was a combination of physiology and theological philosophy. Although today we would consider almost all of it factually false, it is nevertheless a milestone in the history of psychology. Specifically, Hartley's juxtaposition of the questions of philosophy with the methods of physiology would mark an important step toward psychology having its own niche among the academic disciplines. That niche would be to bridge the philosophical questions about mind and moral behavior with the data of physiology. In Great Britain, this goal soon comes to be realized with the work of Alexander Bain (1818–1903).

French philosophy of the era was often an extension of Locke, or a response in some way to Descartes. Etienne Bonnet de Condillac (1714–80) provided an enduring (and wonderful) example of empirical thought through his 'sentient statue,' whereas Francois-Pierre Maine de Biran's (1766–1824), like the associationists, sought to further understand associative learning. Pierre Jean Georges Cabinas (1757–1808) was a physician during the French Revolution whose interest in the mind–body problem even led him to study those executed by the guillotine. And, Julien Offray de La Mettrie (1709–51) believed that Descartes knew the truth – that man is a machine – but had been unwilling to speak it. Rejecting Dualism for Materialism (there is only matter) La Mettrie had no such fear, and gleefully articulated the view that mind *is* brain – a position that is still endorsed by most psychologists today. La Mettrie also saw a continuity between animal and man, and provides a logic that would serve subsequent scientists who believed that the study of animals could advance (even human) psychology.

Also associated with the wider aims of empiricism is Auguste Comte's (1798–1857) positivism (as in, to posit, or assert as foundational). What was being posited as foundational was an understanding based on science over religion. Even in the 1800s the Church was still taken as the authority

on many matters, and it was Comte's goal to see that authority displaced by science (the rational analysis of empirical data). Subsequent positivists, such as Ernst Mach (1838–1916), would in turn be instrumental in establishing guidelines for the proper conduct of valid science. In fact, science and positivism came to be so widely acclaimed in the industrial revolution of the mid-1800s that Romanticism emerged as a response. Fearing that a place for human emotional experience was being lost, and building on such thinkers as Johann Wolfgang von Goethe (1749–1832), intellectual attention shifted to understanding phenomenological experience. Although psychology was to become a science, it continues to feature some aspects of phenomenology, and echoes of the tension between positivism and Romanticism remain.

Rationalism

Where the empiricist emphasizes sensory information and postulates a relatively passive mind that functions according to mechanistic laws, the rationalist postulates an active mind with preexisting content (innate structures and/or concepts) that interprets data from the senses. The earliest rationalists include Baruch Spinoza (1632–77) and Gottfried Wilhelm von Leibniz (1646–1716), both of whom were influential and eclectic thinkers. For example, Leibniz (independent of Newton) developed calculus, devised early computing machines, and his unique monadology. Importantly for our story, he also disagreed with Locke, and Leibniz' alternative views would form the basis of German Rationalism.

Immanuel Kant (1724–1804) was a disciple of Leibniz, but it is said that reading Hume's analysis of causation awakened him from his 'dogmatic slumbers.' Kant would argue that a priori (innate) knowledge of matters such as space and time must exist for us to make sense of our experience. Even Hume's analysis of causation as perceived correlation relied on an innate understanding of time, specifically the notion of before and after. Kant did not postulate specific ideas as innate, but 'categories' that were used in organizing sensory experience.

Kant would also seek to redress the ethical view of Utilitarianism that had emerged as a part of the empirical tradition (e.g., Jeremy Bentham, 1748–1832). This was a sort of hedonic calculus in which the good or pleasurable was weighed against the bad or painful – the right action being the one that produced the greatest good. In contrast, Kant proposed a categorical imperative, according to which, "I should never act except in such a way that I can also will that my maxim should become a universal law."

Like Plato before him, Georg Wilhelm Friedrich Hegel (1770–1831) believed the goal of philosophy was to understand the nature of the cosmos, what Hegel called the Absolute. Our sensory experience of some small part of it was of little value; what was needed was to find the underlying rational structure of reality. For Hegel, this structure was the dialectical process in which a thesis interacts with its antithesis, and the result produces a synthesis, which is also a new thesis, and so the dialectic continues. In ways both overt and subtle, Hegel's thought can be seen in almost every academic discipline in the late 1800s, psychology included.

Existentialism, as typified by Soren Kierkegaard (1813–55) and Friedrich Nietzsche (1844–1900), emerged in part as a response to Hegel's idealism. With its focus on subjective experience as opposed to rational analysis, existentialism would come to have a profound influence on psychology, especially early 'clinical' psychology. To summarize existentialism with a short cliché: "Never let other people do your thinking for you." That is, to explain our psychology mechanistically or nativistically is to give away the importance of the individual, and the free choices of will.

Beyond Kant and Hegel there were other rationalists with a direct connection to psychology, such as Arthur Schopenhauer (1788–1860), and the Scottish School founded by Thomas Reid (1710–96). Adherents of the Scottish School advocated a consideration of the mind as a collection of faculties united by a common sense, or executive function. This school was especially popular as a moral philosophy, as a model of mind endorsed by the physiologists of the day, and in America's first universities.

Just as empiricism culminates with Bain's psychology in Britain, rationalism will produce a German psychology. Johann Friedrich Herbart's (1776–1841) works include his *Textbook in Psychology* (1816) and *Psychology as a Science Based on Experience, Metaphysics, and Mathematics* (1824–25). The *Textbook* was certainly not the first work on psychology per se, but may be the oldest volume wherein psychology is being treated as an actual (independent) academic discipline. Similarly, Rudolph Hermann Lotze (1817–81), like Hartley before him, would advocate for an independent academic psychology that would juxtapose physiology (and medicine) with philosophy. Lotze would live to see Wundt, 1879, and all that, and his own Chair at Göttingen would next be filled by the famed early *psychologist*, G.E. Müller.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Cognitive Psychology: History; Empiricism, History of; Functionalism, History of; Habit: History of the Concept; Positivism, History of; Psychoanalysis, History of; Psychology, History of (Twentieth Century); Scientific Knowledge, Sociology of.

Bibliography

Boakes, R., 1984. From Darwin to Behaviourism: Psychology and the Minds of Animals. Cambridge University Press, New York.

Bonner, A., 2007. The Art and Logic of Raymond Llull. Brill, Leiden, The Netherlands.

Hergenhahn, B.R., Henley, T.B., 2014. An Introduction to the History of Psychology, seventh ed. Wadsworth, Belmont, CA.

Hastorf, A., Cantril, H., 1954. They saw a game; a case study. Journal of Abnormal and Social Psychology 49, 129–134.

Humphrey, N., 1983. Consciousness Regained. Oxford University Press, Oxford, England.

Loftus, E.F., Palmer, J.C., 1974. Reconstruction of automobile destruction: an example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior 13, 585–589.

Skinner, B.F., 1956. A case study in scientific method. American Psychologist 11, 221–233.

Psychology, History of (Twentieth Century)

Spencer A McWilliams, California State University San Marcos, San Marcos, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.G. Ash, volume 18, pp. 12399–12405, © 2001, Elsevier Ltd.

Abstract

Twentieth-century academic psychology emerged formally in 1879, originally defined as the science of mental life, with the goal of uncovering the facts of consciousness. Evolutionary theory expanded its emphasis on how mental processes and behavior function to help organisms adapt to environments. The study of learning dominated the early half of the twentieth century, with the goal of predicting and influencing behavior. Various subfields and a positivist emphasis on experimental methods led to consolidation of the discipline by midcentury. Technological developments stimulated resurgence of the study of cognition and neuropsychology and increased disciplinary fragmentation later in the century.

Nineteenth-Century Foundations

Modern academic psychology emerged during the latter part of the nineteenth century as a way of addressing age-old questions regarding human nature and built upon work in biology, physiology, medicine, and philosophy – particularly philosophical debates over explanations for how humans develop knowledge, how humans represent environmental objects in conscious awareness, and how best to account for human behavior (Hergenhahn and Henley, 2014; also see article on the Psychology, History of (Early Period)). Twentieth-century psychology built upon these and other topics and branched out into a wide variety of issues, theories, and methodologies. This article addresses psychology as an academic undertaking, primarily in the form of an experiment science, thus (unfortunately) omitting discussion of the many applied areas of psychology that are covered in the encyclopedia article on the history of clinical psychology, including the important influences of psychoanalytic theory, other theories of personality and psychotherapy, and psychological testing and measurement, which also had a large impact on the work of academic psychologists.

Hermann von Helmholtz's (1821–94) mechanistic approach to physiology and his contributions to understanding the physiology of visual and auditory perception, along with such findings as Paul Broca's (1824–80) and Carl Wernicke's (1848–1905) identification of localized brain functions, stimulated research into the relationship among sensations, brain processes, and conscious experience, and in particular, the measurement of these relations. Ernst Weber (1795–1878) pioneered the study of psychophysics – the relationship between physical characteristics of objects and the psychological perception of them. He asked people to discriminate the weight differences of two objects by comparing a target weight with a standard and determined the relationship between the physical weight of the object and the participant's perception of whether the object was heavier than, lighter than, or equal to the standard. From these data, he derived the just noticeable difference (jnd) – the smallest weight that the participant could discriminate – and found that the jnd did not reflect an absolute weight difference, but rather a ratio that is relative to the weight of the standard. This finding, in 1834, Weber's law, is regarded as the first 'law' in psychology.

Further research by Gustav Fechner (1801–87) confirmed and refined Weber's law, generalized it to auditory and visual stimuli, and, in 1860, stated it mathematically that sensations increase arithmetically as physical stimuli increase geometrically (Weber's law), indicating that humans perceive sensations relative to the background context in which they occur. Building on Fechner's work, Hermann Ebbinghaus (1850–1909) studied the processes of learning and memory as they occurred and his findings, still regarded as valid today, demonstrated that forgetting occurs rapidly during the first few hours following learning and then declines more slowly. He also found that learning distributed over a period of time led to better retention than massed repetition at a single time and that overlearning already memorized material reduced the rate of forgetting.

Charles Darwin's (1809–82) theory of evolution stimulated an interest in individual differences among people in terms of intelligence and personal characteristics, based on the assumption that such differences might account for success in adaptation and survival. Darwin's English cousin Francis Galton (1822–1911) began attempts to measure intelligence, following the British empiricist tradition, by emphasizing innate sensory acuity as the primary way humans come to know and adapt to the world. His work led to statistical concepts such as the mean and correlation, which became central to the twentieth-century psychology emphasis on quantitative analysis and the search for general principles. Charles Spearman (1863–1945), also in England, studied the relation between sensory acuity and children's success in school and suggested that an overall inherited factor, called general intelligence, accounted for most of intelligence, with specific factors reflecting learning and experience. In France, Alfred Binet (1857–1911) pursued the measurement of intelligence from the Continental rationalist perspective by measuring a variety of complex mental operations that reflected his view of intelligence as active and complex, rather than a single ability, and that could be nurtured rather than being fixed by inheritance. While a revised version of Binet's intelligence test became the most popular method for assessing children's intelligence in the USA, most psychologists adopted Spearman's general factor understanding of intelligence.

Beginnings of Experimental Psychology

Voluntarism, Structuralism, and Phenomenology

Scholars typically date Wilhelm Wundt's (1832–1920) founding of a laboratory for psychological research at the University of Leipzig in 1879 as the formal beginning of modern psychology. Wundt's voluntarism, which focused on the active processes involved in selective attention, incorporated the work of Helmholtz, Weber, and Fechner into an organized research program that employed scientifically controlled introspection to determine the underlying processes involved in human conscious awareness of environmental objects. He viewed the goal of psychology as identifying general principles of perception and learning that applied to all normal, adult humans by identifying how higher-order concepts derived from raw sensory elements. Wundt built on the European Continental rationalist tradition's view of a structured, active mind that transforms sensations and particularly focused his research on selective perception, the question of what determines how, among the entire perceptual field, people attended to specific sensations and objects.

Following his graduation from Oxford, Edward Titchener (1867–1927) studied with Wundt but followed his native British empiricist tradition, which emphasized a blank mind that passively builds up perceptions through association of sensations. He came to the USA in 1892 and developed a successful doctoral program in psychology at Cornell University. His structuralist approach pursued the search for mental elements further by attempting to identify the immediate experience of raw sensations and determine the laws through which they combine to make perceptions and more complex mental phenomena, attempting to create a chart of psychological elements analogous to chemistry's period chart of elements. Although Titchener trained many doctoral graduates who became highly productive and well-known psychologists, his structuralist approach declined, primarily because even highly trained introspectionists could not agree on how to describe a raw sensation, and the retrospective nature of their descriptions hampered the attempt to study sensory events as they occurred.

European phenomenologists, such as Franz Brentano (1838–1917) and Carl Stumpf (1848–1936), opposed the emphasis on raw sensory elements and championed a phenomenological approach that attempted to understand consciousness in terms of whole, meaningful, intact perceptions, mental processes rather than contents, and the functional nature of mental acts. Edmund Husserl (1859–1938) further proposed that scientific psychology should examine the pure phenomenology of personal subjective experience, and he searched for mental essences and meaning. These phenomenological approaches influenced Gestalt psychology and existential approaches to psychology during the twentieth century.

Functionalism

The publication of William James' (1842–1910) Principles of Psychology in 1890 incorporated and elaborated on each of these traditions and expanded the field to embrace any useful topic or methodology that would shed light on the human condition. Influenced by Darwin's theory of evolution, James' functionalist psychology emphasized how mental processes serve to aid adaptation to the environment and included the study of individual differences as well as general principles, the application of psychology to improving the human situation, and the study of behavior as well as consciousness. James likened consciousness to an unbroken stream rather than a chain of independent elements, emphasizing its personal, continuous, and impermanent qualities and its function in enhancing adaptation to the environment. His views on habits, the self, emotion, volition, and cognition, along with the importance of useful practical applied psychology, elaborated by other functionalist colleagues such as John Dewey (1859–1952), James Angell (1869–1949), and Harvey Carr (1873–1954) established a foundation for the eclecticism of twentieth-century psychology. The functionalists also embraced a pragmatist philosophical perspective that rejected the positivist search for independent truths and instead proposed that the validity of ideas and theories lay in their usefulness in solving human-defined problems and anticipating future events.

Behaviorism

The desire for a more objective approach to psychology, which would avoid the subjectivity of studying conscious experience and using introspection, led Russian physiologists Ivan Sechenov (1829–1905), Vladimir Bekhterev (1857–1927), and Ivan Pavlov (1849–1936) to propose that psychology studies the relationship between environmental stimuli and behavior, both physiological and overt responses. The American psychologist John Watson (1878–1958) adopted Pavlov's respondent conditioning approach and elaborated on this view with his Behaviorist Manifesto, in 1913, which proposed that the goal of psychology should be the prediction and control of behavior. These approaches renewed interest in the empiricist tradition that emphasized learning experiences as the foundation of knowledge, rather than rationalist view that emphasized the importance of innate characteristics. Although initially resisted by most contemporaries, the behaviorist view gradually came to dominate American psychology for half of the twentieth century, changing the definition of the discipline from "the science of mental life" (James, 1890: p. 1) to "the science of human and animal behavior."

Within this behaviorist emphasis researchers differed in the type of behavior they investigated. Some pursued Watson's Pavlovian emphasis on associative, reflexive, respondent (Classical) conditioning. However, the emphasis eventually shifted to the study of purposive, goal-directed behaviors, which organisms voluntarily emit and which lead to consequences that influence future behavior. Edward Thorndike (1874–1949), a pioneer in the study of goal-directed behavior, used the term 'law of effect' in 1898 to emphasize the role of the consequences of behavior to explain its purposive aspects. In 1912, William McDougall (1871–1938) emphasized a behavioral perspective in studying purposive behavior, which he defined as spontaneous, rather than a response to a known stimulus, and which persists until the attainment of a goal. This emphasis on purposive, goal-directed behavior came to characterize the era of the learning theories in the USA, from the 1930s through the 1960s, which used nonhuman animals to

study learning, the primary mechanism by which organisms adapt to environmental change.

Three major approaches vied for dominance during this period, erecting theories of learning following the logical positivist approach to theoretical concepts (see below). Edwin Guthrie (1886–1959) remained close to Watson's respondent learning approach by attempting to explain purposive behavior in terms of the laws of associative learning; he depicted purposive learning as a response to the most recent stimuli associated with that response. Clark Hull (1884–1952) applied the model of Newton's physics to psychology, developing a psychological theory explaining purposive behavior in mechanistic terms. He emphasized the importance of biological drive as motivating behavior, with reduction of the drive serving to reinforce, or increase the probability of, the behavior that led to it. Edward Tolman (1886–1959) attempted to explain purposive behavior in behavioral terms, but he operationally defined concepts like purpose and cognition as intervening variables in goal-directed behavior. His theory sustained a cognitive emphasis that supported the resurgence of cognitive psychology in the 1960s.

The behaviorist perspective reached its zenith in the 1960s with B.F. Skinner's (1904–90) operant conditioning approach, which rejected the need for theories of learning and use of cognitive, or mentalistic, explanations of behavior. Instead, he analyzed the functional relationships among environmental situations, overt behaviors, and consequences, and how these three factors influenced the future probability of behavior. Using evolutionary theory as a metaphor, he described how, during an organism's lifetime, the environment selected its behaviors through the process of conditioning. Reinforcing environmental consequences strengthened (increased the probability of) behaviors, while other consequences decreased behavioral probability through punishment or extinction. Skinner's radical behaviorism became widely embraced in experimental psychology and also led to application in many fields of human services.

Influence of Logical Positivism

Throughout the behaviorist heyday, many psychologists in the USA, Europe, and elsewhere continued to pursue research on psychology's original topics such as sensation, perception, memory, problem-solving, and other cognitive topics. Although many investigators on these topics did not self-identify as behaviorists, most came to accept observable behaviors as the proper valid evidence for psychological processes, even if these behaviors consisted of responses to self-report questionnaires or scales. Based on the success of Newtonian physics, the methods of the natural sciences, grounded in the positivist belief in an absolute independent external reality accessible to human knowing through empirical research, emerged as psychology's dominant methodological approach. Twentieth-century psychologists reaffirmed natural sciences methodology and embraced the logical positivist argument of the Vienna Circle in the 1920s that legitimated the use of theoretical concepts as long as they were operationally defined. Eventually, logical positivism failed as a philosophy of science due to its self-referential incoherence (logical positivism states that meaningful statements require empirical verifiability, at least in principle; therefore, logical positivism itself cannot be meaningful, by its own criterion). However, despite strong criticism from within the field (Bakan, 1973), its adoption by psychologists, in conjunction with the positivist methodology of Newtonian physics and psychology's self-conscious focus on gaining scientific respectability, led to the primacy of research methodology, more than content, as the defining common identity of academic psychologists.

Subfields and Subspecialties

Gestalt Psychology

Although empirically based behaviorism dominated psychology research in America for the first half of the twentieth century, the Continental rationalist tradition remained active in Europe, particularly in the work of the German Gestalt psychologists, led by Max Wertheimer (1880–1943). Gestalt psychology continued to elaborate on the rationalist view of perception as an active process in which mental or brain processes structure and transform sensory data. Along with colleagues Wolfgang Köhler (1887–1967) and Kurt Koffka (1886–1941), Wertheimer provided persuasive support for the perspective that the tendency to perceive forms, structure, and patterns dominated the perception process more than particular sensory inputs, that the whole (perception) is greater than the sum of the parts (sensations) that generate it. Many Gestalt psychologists found a welcome home in America immediately before and after World War II, increasing interest in Gestalt principles, perception, and the active role of mental processes in America.

Social Psychology

Gestalt psychology also had an impact on the emerging subfield of social psychology. Several of the Gestalt psychologists who immigrated to the USA focused their work on applying the perspective to the influence of social factors, defined as the real or imagined presence of other people, on psychological processes. Kurt Lewin's (1890–1947) field theory pioneered the Gestalt approach to the study of personality and social behaviors and its application to motivation, conflict, and group dynamics. Lewin's and Koffka's student Tamara Dembo (1902–93) contributed to the study of frustration and aspiration, and Lewin's student Leon Festinger (1919–89) became well known for his work on social comparison and cognitive dissonance. Solomon Asch (1907–96) studied Gestalt psychology and taught at Swarthmore College with Köhler. He gained recognition for his work on how humans form impressions of others and for his research in conformity that demonstrated that groups can influence even obvious perceptions. Social psychology evolved into an active field that employed creative methodologies to study a variety of contemporary social topics that interested the psychologists of the time (Gergen, 1973).

Developmental Psychology

Darwin's theory also influenced the emerging field of developmental psychology by suggesting that functional psychological processes evolve in individuals across the life span, particularly during infancy, childhood, and adolescence. G. Stanley Hall (1844–1924), founder of the American

Psychological Association (APA), proposed that human development proceeds in a manner that mirrors the increasing complexity of organisms on the phylogenic scale, that 'ontology recapitulates phylogeny.' He built a strong developmental psychology program at Clark University that included Heinz Werner's (1890–1964) pivotal role in bringing the European, Kantian, intellectual traditions to America. Werner elaborated the orderly directionality of developmental changes in terms of a succession of structures that succeed each other (Glick, 1992). The Continental rationalist tradition also guided the work of the Swiss genetic epistemologist, Jean Piaget (1896–1980), whose constructivist cognitive approach greatly influenced developmental psychology and also sustained interest in cognitive processes during the behaviorist era. Piaget showed how a child's experience of the environment becomes more complex and adapted as the child articulates a cognitive schema in the process of biological maturation. These pioneering developmental themes of Werner and Piaget appear in a variety of subsequent developmental psychology models and approaches (Kegan, 1982).

Neuropsychology

Early-twentieth-century advances in studying the relationship between psychological and neurological processes continued to contribute to the evolution of psychological research. Karl Lashley (1890–1958) sought to identify the neurological bases of conditioned reflexes as studied by the Russian physiologists and championed by Watson's behaviorism. He reasoned that innate, instinctive reflexive responses to certain stimuli require an innate neurological structure, and thus conditioned associated responses to these stimuli might also come to have a specific neurological location. The term 'engram' referred to this hypothetical location of learning and memory, and Lashley devoted his career to attempting to find it. Ultimately, he failed to find a single location in the brain for specific learning and proposed that the brain operates as a unified whole rather than with specific physical links, a principle he called mass action.

Donald Hebb (1904–85) saw the newborn's brain as unstructured and undifferentiated, with organized connections among neurons occurring through experience. He surmised that sensory organs respond to environmental objects by firing a complex of neurons, which he called cell assemblies, as a result of reverberating neural activity that enabled separate neurons to establish connections while stimulated. He further suggested that the assemblies of neurons representing an environmental object become connected in a series of activities that he called phase sequences. These sequences come to represent complex trains of memories and thoughts. Hebb articulated the neurological level of connection by positing what came to be known as 'Hebb's rule' that neurons that fire together become wired together. These principles of mass action and cell assemblies influenced developments in cognitive science and neuroscience in the latter part of the twentieth century.

Psychology in Midcentury

Altman (1987) described the period of 1900–60 as a major era of consolidation of the discipline of psychology, within the various perspectives described above establishing the field's topics and defining its content. Methodologically, the emphasis on experimentation, quantification, and laboratory-based research unified the prominence of experimental methodology as the discipline's standard. By reaffirming the logical positivist perspective, the methodology supported operational definitions of theoretical concepts, the quest for causal relationships, and the 'discovery' of universal laws and principles of human and animal cognition and behavior.

Graduate students in psychology during this period typically received broad exposure to psychological theory, research, and content that led to their identification with the discipline as a whole. In sum, "psychology during the 1900–60 period displayed a general pattern of centripetality in research and theory, philosophy of science, graduate education, and organizational infrastructure. This centripetal pattern resulted in consolidation and unification of the discipline" (Altman, 1987: p. 1061). This sense of unification existed even though psychology had not generated an agreed-upon core lexicon, or shared paradigm (Kuhn, 1962), as was the case in other established scientific disciplines.

Late-Twentieth-Century Developments

In the latter half of the twentieth century, two interrelated themes influenced research into the questions and topics that characterized this unified discipline of psychology. The waning popularity of behaviorism and its emphasis on learned behavior influenced a resurgence of interest in cognition and invigorated the importance of nativist, biological explanations of psychological phenomena. Additionally, rapid technological improvements, which routinely stimulate advances in many scientific fields (Pickering, 1984, 1995), had an increasing impact on academic research on psychology's core questions and topics.

Resurgence of Cognitive Psychology and Cognitive Science

A symposium at MIT on information theory in 1956 sparked a revival of research in cognition within the context of the development of sophisticated computer technology, cybernetics, and information processing models. George A. Miller (1920–2012) demonstrated that humans can discriminate seven aspects of events and retain seven units of experience, and published his landmark text Psychology: The Science of Mental Life in 1962, reclaiming James' original definition of the discipline. Jerome Bruner (b. 1915) studied abstract thinking, concept learning, and the use of active cognitive strategies. Ulric Neisser (1928–2012) defined cognition as any processes that transform, reduce, elaborate, store, recover, and use sensory input, further clarifying this renewed approach to the topics that occupied early experimental psychology.

Subsequent developments in cognitive psychology focused on computer programs as a metaphor for the processes of the human mind, leading to the field of artificial intelligence, a branch of computer science. Englishman Alan Turing (1912–54) founded this field by asking whether a machine can actually think, creating his famous Turing test in which

participants attempt to consistently differentiate a human responder from that of a computer. The question of whether computers actually think like a human or only follow rules of program received further discussion by John Searle (b. 1932), who argued against the likelihood of valid artificial intelligence. By the mid-1970s, Neisser criticized the computer and information models of cognition as too sterile and argued that cognitive psychologists should study cognition as it naturally occurs in people's ordinary lives and how it functions in their purposeful activities. However, computer modeling remained the major methodology.

Toward the end of the twentieth century, an interdisciplinary field of cognitive science emerged, combining the work of psychologists with anthropologists, linguists, computer scientists, neuroscientists, and philosophers. An approach called connectionism surfaced as an active model in this field, attempting to provide a computer-based virtual model of how the cognitive and neurological system learns by studying the relationship between input and output in terms of Hebb's rule and Lashley's principle of mass action.

Neuropsychology, Neuroscience, and Evolutionary Psychology

Late-twentieth-century development of sophisticated neuroimaging measurement techniques such as functional magnetic resonance imaging and computed axial tomography scanning, along with advances in human genetics and electrophysiology, enhanced the progression of mapping human emotions and cognitions to specific neural areas, rapidly extending the earlier work of Lashley and Hebb. These developments led to emergence of the interdisciplinary field of neuroscience, including disciplines such as biology, linguistics, mathematics, computer science, and philosophy, as well as psychology, which studies neural circuitry and its relationship to psychological functions.

As we have seen, Darwin's theory of evolution influenced many, if not most, areas of psychology. Toward the latter part of the twentieth century, however, we see the development of a specific approach to psychology called evolutionary psychology. This perspective evolved from sociobiology, founded by Edward O. Wilson (b. 1929), which proposed that evolution has selected for humans with innate tendencies toward social behavior, language, creating cultures, social conventions, phobias, etc. The question of the relative strength of these innate tendencies versus the influence of culture and learning experiences generated a lively debate, with Wilson suggesting that nature gives us a 'long leash,' enabling flexible modes of adapting to changing environments, rather than a 'short leash' that strictly determined specific behaviors.

Evolutionary psychology regards human psychological processes as a collection of functionally specialized regulatory structures or circuits that organize how humans interpret experiences and create concepts and frameworks of meaning. This evolutionary perspective recognizes the universality of 'normal' human behavior through the study of natural competencies that all humans share. Leda Cosmides and John Tooby (1997) thus define psychology as a branch of biology that studies how brains develop programs to process information, emphasizing the brain as a physical system that functions as a computer and generates behavior appropriate to the environment. They regard the brain's neural circuitry as a complex system of specialized pathways that evolved through natural selection to assist in solving problems that the humans encountered over the course of evolutionary history. Both sociobiology and evolutionary psychology base their work on the assumption that we can understand psychological processes as selected for by their benefit in perpetuation of copies of genes into subsequent generations.

Psychology at the End of the Twentieth Century

Irwin Altman (1987) in the USA, Theo Herrmann (2009) in Germany, and Tommy MacKay (2001) in the United Kingdom, each in their own way, characterized the latter half of psychology's twentieth century, in contrast to the first half, as a time of centrifugal trends and discontinuity. They portray the field as a proliferation of specialty areas, loss of a shared common history (Fuchs and Viney, 2002; Samelson, 1999) and core of mutual beliefs, and a weaker identification of its members with the field as a whole. Graduate programs came to focus more on specialized areas, with each field developing its own curriculum and cohort of students, organizational structures, and separated physical quarters. Applied and professional psychology programs became distanced from the basic science of the field, and professional schools and psychotherapy training programs proliferated independently of academic research psychology.

Much of what had become the traditional content of psychology has been absorbed into the newly emerging interdisciplinary specialties of cognitive science and neuroscience (Gardner, 2009; Gentner, 2010), decreasing psychology's absolute claim to this territory (MacKay, 2001). The author's own graduate education exemplifies this trend. I completed my PhD in the Department of Psychology at the University of Rochester in 1971, with a major in clinical psychology and a minor in perception. In 1995, the Department of Psychology ceased to exist, as it split into two departments: the Department of Clinical and Social Sciences in Psychology (locus of my major field) and the Department of Brain and Cognitive Sciences (locus of my minor field).

We can also see this fragmentation reflected in the organization structure of the discipline in the United States. The APA originally primarily represented those involved in academic and basic research and only slowly admitted professional psychologists over its early developing years. By the 1980s, however, professional psychologists came to outnumber academics in APA, and it currently functions as an organization serving the interests of professional practitioners. This trend led to the establishment of an alternative organization in 1988, The Association for Psychological Science (APS), which focuses on research and the scientific aspects of the field. APS has embraced an international perspective and the inclusion of the interdisciplinary fields described above, sponsoring an international convention of psychological science in the Netherlands in 2015.

Although we might feel compelled to do so, we need not regard fragmentation, discontinuity, and centrifugal trends as harmful to the discipline. Rather, we can view them as reflecting an energetic period of change and reformulation that can imbue the discipline with new directions in methods and

theories. By regarding psychology as a relatively young discipline, even characterizing it as in a medieval period (Cromwell, 2010), we can see the state of the discipline in the beginning of the twenty-first century as providing intriguing opportunities for new ideas and formulations.

Challenges to Twentieth-Century Conventions

Although most mainstream academic and research psychologists maintain a commitment to the scientific goal of establishing general laws of human nature through empirical research and logical positivist theorizing, alternative approaches began to receive growing attention by the end of the twentieth century. Sometimes viewed as part of a broader postmodern movement and reflecting challenges to scientific absolutism spawned by Kuhn's (1962) critique of the correspondence theory of scientific truth, many psychologists have begun to question the prevailing view that scientific methods uncover objective, permanent truths. Instead, they suggest that humans play a central role in the construction of knowledge. A variety of approaches to psychology, including personal constructivism (Raskin, 2002), social constructionism (Gergen, 2009), contextualism (Hayes et al., 2001), hermeneutics (Martin and Sugarman, 1999; Slife and Christensen, 2013), Buddhism (McWilliams, 2014), and pragmatism (Fishman, 1999), emphasize the active responsibility of human psychological and social processes in generating ever-changing knowledge and the temporal embeddedness of these processes in historical, cultural, and language contexts. Albert Bandura's (2006) social cognitive approach also came to strongly embrace the active role of human agency in coevolution and creating unanticipated futures in a world of flux and change. Although not without criticism from those committed to conventional psychological models (Held, 2007), these perspectives suggest alternative models that may have promise for a new view of the field.

We might anticipate that as many of the topics that originally fell within psychology's twentieth-century domain migrate to interdisciplinary, biologically based fields, psychology may come to focus more on the processes by which people, personally and collectively, actively create knowledge and meaning, including the approaches to knowledge and meaning that we call science. Perhaps then, rather than viewing psychology as a branch of the natural sciences, we may come to see the natural sciences, the humanities, the arts, religion, philosophy, and other human undertakings as the subject matter of psychology.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Cognitive Psychology: History; Industrial–Organizational Psychology: Science and Practice; Psychology, History of (Early Period); Social Psychological Theory, History of; Social Psychology.

Bibliography

Altman, I., 1987. Centripetal and centrifugal trends in psychology. American Psychologist 42 (12), 1058–1069.
Bakan, D., 1973. On Method: Toward a Reconstruction of Psychological Method. Jossey-Bass, San Francisco, CA.
Bandura, A., 2006. Toward a psychology of human agency. Perspectives on Psychological Science 1, 164–180.
Cosmides, L., Tooby, J., 1997. Evolutionary Psychology: A Primer. http://www.cep.ucsb.edu/primer.html (accessed 17.03.14.).
Cromwell, R.L., 2010. Being Human: Human Being. Manifesto for a New Psychology. iUniverse, New York.
Fishman, D.B., 1999. The Case for Pragmatic Psychology. New York University Press, New York.
Fuchs, A.J., Viney, W., 2002. The course in the history of psychology: present status and future concerns. History of Psychology 5 (1), 3–15.
Gardner, H., 2009. Scientific psychology: should we bury it or praise it? Humana.Mente 11, 11–23.
Gentner, D., 2010. Psychology in cognitive science: 1978–2038. Topics in Cognitive Science 2, 328–344.
Gergen, K.J., 1973. Social psychology as history. Journal of Personality and Social Psychology 26 (2), 309–320.
Gergen, K.J., 2009. An Invitation to Social Constructionism, second ed. Sage, London.
Glick, J.A., 1992. Werner's relevance for contemporary developmental psychology. Developmental Psychology 28 (4), 558–565.
Hayes, S.C., Barnes-Holmes, Roche, B. (Eds.), 2001. Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition. Kluwer Academic/Plenum, New York.
Held, B.S., 2007. Psychology's Interpretive Turn: The Search for Truth and Agency in Theoretical and Philosophical Psychology. American Psychological Association, Washington, DC.
Hergenhahn, B.R., Henley, T.B., 2014. An Introduction to the History of Psychology, seventh ed. Wadsworth, Belmont, CA.
Herrmann, T., 2009. What's next?: continuity and discontinuity in psychology. Zeitschrift für Psychologie 217 (2), 95–102.
James, W., 1890. The Principles of Psychology, vols. I & II. Holt, New York.
Kegan, R., 1982. The Evolving Self: Problem and Process in Human Development. Harvard University Press, Cambridge, MA.
Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
MacKay, T., 2001. The future belongs to psychology. The Psychologist 14 (9), 466–469.
Martin, J., Sugarman, J., 1999. The Psychology of Human Possibility and Constraint. SUNY Press, Albany, NY.
McWilliams, S.A., 2014. Foundations of mindfulness and contemplation: traditional and contemporary perspectives. International Journal of Mental Health and Addiction. http://dx.doi.org/10.1007/s11469-014-9486-1.
Pickering, A., 1984. Constructing Quarks: A Sociological History of Particle Physics. University of Chicago Press, Chicago.
Pickering, A., 1995. The Mangle of Practice: Time, Agency, and Science. University of Chicago Press, Chicago.
Raskin, J.D., 2002. Constructivism in psychology: personal construct psychology, radical constructivism, and social constructionism. American Communication Journal 5 (3), 1–24.
Samelson, F., 1999. Assessing research in the history of psychology: past, present, and future. Journal of the History of the Behavioral Sciences 35 (3), 247–255.
Slife, B.D., Christensen, T.R., 2013. Hermeneutic realism: toward a truly meaningful psychology. Review of General Psychology 17 (2), 230–236.

Psychometrics

Denny Borsboom and Dylan Molenaar, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.O. Ramsay, volume 18, pp. 12416–12422, © 2001, Elsevier Ltd.

Abstract

Psychometrics is a scientific discipline concerned with the construction of measurement models for psychological data. In these models, a theoretical construct (e.g., intelligence) is systematically coordinated with observables (e.g., IQ scores). This is often done through latent variable models, which represent the construct of interest as a latent variable that acts as the common determinant of a set of test scores. Important psychometric questions include (1) how much information about the latent variable is contained in the data (measurement precision), (2) whether the test scores indeed measure the intended construct (validity), and (3) to what extent the test scores function in the same way in different groups (measurement invariance). Recent developments have focused on extending the basic latent variable model for more complex research designs and on implementing psychometric models in freely available software.

Definition

Psychometrics is a scientific discipline concerned with the question of how psychological constructs (e.g., intelligence, neuroticism, or depression) can be optimally related to observables (e.g., outcomes of psychological tests, genetic profiles, neuroscientific information). This problem is most often approached through the construction of measurement models, in which the construct of interest is represented as a latent variable that acts as a common determinant of a set of observable variables (Bollen, 2002; Sijtsma, 2011). Psychometrics is a highly interdisciplinary field, with connections to statistics, data theory, econometrics, biometrics, measurement theory, and mathematical psychology. Psychometricians may be involved in the design of psychological tests, the formalization of psychological theory, and the construction of data-analytic models. The latter activity has, in the past century, been the main focus of the discipline, especially through the development of test theory and associated latent variable modeling techniques. The main professional organization of psychometricians is the *Psychometric Society*, which publishes the journal *Psychometrika* and organizes the annual *International Meeting of the Psychometric Society*.

History

The birth of psychometrics is usually situated at the end of the nineteenth century, when Francis Galton set up the Antropometric Laboratory with the intention to determine psychological attributes experimentally. Among the first constructs of interest to be subjected to measurement were keenness of sight, color sense, and judgment of eye (Galton, 1884). Galton attempted to measure such attributes through a variety of tasks, recording performance accuracy as well as reaction times. In the early twentieth century, the interest in measuring human qualities intensified greatly when the United States implemented a program to select prospective soldiers using tests (the Army Alpha and Beta) that purported to measure a range of abilities deemed relevant for military performance. Such tests produced a great deal of data, which led to questions that inspired the birth of psychometric theory as we currently know it: how should we analyze psychological test data? Which properties determine the quality of a psychological test? How may we find out whether a test is suited for its purpose?

Two important properties of tests were almost immediately identified (Kelley, 1927). The first referred to the question of whether a test produced *consistent* scores when applied in the same circumstances. This question introduced the notion of *reliability*. In general, a measurement instrument is reliable to the extent that it yields the same outcomes when applied to persons with the same standing of the measured attribute under the same circumstances (e.g., to what extent does an IQ test produce the same score if administered to people with the same level of intelligence?). The second important question is whether the test measures what it purports to measure. This question defines the concept of *validity*. Although different views of validity exist (see later text), the textbook definition is that a measurement instrument is valid if and only if it measures what it should measure (e.g., an IQ test is valid if and only if it actually measures intelligence). The concepts of reliability and validity are still among the most important ones in the evaluation of psychological tests, although they have a very different status in psychometric theory. The definition of reliability is widely accepted, but the definition of validity is widely contested.

The development of reliability theory in the first half of the twentieth century culminated in the work of Lord and Novick (1968), who presented the currently accepted definition of reliability as a signal-to-noise ratio (the ratio of true score variance to observed score variance). The concept may be somewhat differently conceptualized in different theoretical frameworks (e.g., as measurement precision in latent variable theory (Mellenbergh, 1996) and as generalizability in generalizability theory (Cronbach et al., 1972)). However, in such cases Lord and Novick's definition typically follows as a special case, which indicates the consistency of the general psychometric framework. Reliability can, under appropriate assumptions, be estimated in various ways; for instance, from the correlation between two test halves, from the average correlation between test items, and from the correlation

between two administrations of the same test at different times. Strictly speaking, such reliability estimates signify properties of test scores rather than of tests themselves (e.g., administering the test to different populations will ordinarily lead to different values for reliability). Although some discussion may exist about how to optimally estimate reliability, or about which coefficient should be preferred in a given context (Sijtsma, 2009), there is little discussion about how to define reliability theoretically or on how to determine it empirically.

The opposite holds for validity. There are no widely accepted methods to determine whether a test is valid (or to estimate its degree of validity, for theories that conceptualize validity as a matter of degree). In fact, there is no agreement on what validity refers to (Borsboom et al., 2004; Lissitz, 2009; Sijtsma, 2011; Newton, 2012). These disagreements result from the fact that the question of whether a test measures what it purports to measure raises fundamental questions about the nature of psychological constructs themselves. For instance, can psychological constructs literally be measured, in the same way that length and mass are (Michell, 1999)? If not, what notion of measurement should be invoked? Should psychological constructs be given a realist interpretation, or are they merely summaries of data (Borsboom, 2005; Cronbach and Meehl, 1955; Trout, 1999)? Does it make sense to talk about the validity of tests at all, or should we rather talk about the validity of interpretations of test scores or of uses of test scores (Newton, 2012)? Should the societal consequences of test use be counted in the evaluation of validity (Messick, 1989)? Perhaps the only conclusion about validity that is widely accepted is that it is the most problematic among the psychometric concepts.

It is an interesting fact of scientific history that psychometric theory and practice developed largely in the absence of definitive answers to these fundamental questions. Perhaps as a result of the inability to come to grips with the slippery constructs of psychological science, psychometrics instead took its lead from statistics and from general ideas on how measurement should be conceptualized. In doing so, it became a largely technical discipline, which identified as its main task the construction of measurement models for psychological data. Usually, such models contain a 'stand-in' for the psychological construct to be measured (e.g., the expectation of the observed scores – the 'true score' – or a latent variable that is assumed to 'underlie' the responses to test items); they do not, however, contain psychological theory.

Although the detachment of modeling techniques and substance matter is occasionally lamented in psychometric circles, it is undeniable that the statistical models developed along statistical, data-analytic lines define some of the most important contributions of psychometric theory to science in general: classical test theory (Lord and Novick, 1968), latent class analysis (Lazarsfeld and Henry, 1968), the congeneric model (Jöreskog, 1971), the modern test theory models of Rasch (1960) and Birnbaum (1968), and the nonparametric item response models (Mokken, 1971). After these models had been constructed, the development of software to fit and estimate them became one of the main topics of psychometric research, and psychometricians played a leading role in the development of estimation algorithms (e.g., Bock and Aitkin, 1981), model fit tests (Andersen, 1973), software for test analysis (Zimowski et al., 1996; Thissen, 1983; Wilson et al., 1991; Bowler et al., 2007), and general latent variable modeling (e.g., Jöreskog and Sörbom, 1974; Muthén and Muthén, 1998; Bentler, 2000; Arbuckle, 2010; Vermunt and Magidson, 2000). This development took place during the last three decades of the twentieth century, at the end of which most of the basic psychometric models could be estimated and fitted.

Main Concepts of Standard Psychometric Theory

Psychometric measurement models relate a latent structure to a set of observed variables by mapping positions in the latent structure to distributions or densities of the observed variables. This is usually done by specifying the conditional distribution function of the observables, given the latent structure. Thus, the general framework can be thought of in terms of a simultaneous regression of the observed variables on a latent variable or a set of latent variables. Many different models can be derived from this idea through variations on (1) the form of the latent structure (e.g., a continuous line or a set of latent classes), (2) the form of the regression function (e.g., a step function or a logistic function), and (3) the distribution or density appropriate to the observations (e.g., a binomial distribution or a normal density). For instance, if the latent structure is a unidimensional continuum, the regression function is linear, and the observables follow a normal density, the resulting model is the linear common factor model (Jöreskog, 1971); if the latent structure is a unidimensional continuum, the regression function is logistic, and the observables follow a binomial distribution, we get the two-parameter logistic model (Birnbaum, 1968); and if the latent structure is categorical and the observed variables are binary, we obtain the latent class model (Lazarsfeld and Henry, 1968). Mellenbergh (1994b) provides a systematic overview of the connections between these and other models.

The latent structure is generally viewed as a representation of the construct to be measured (e.g., intelligence), while the observed scores typically represent concrete behavioral responses (e.g., answers to items in an IQ test). As a whole, the psychometric model thus coordinates the correspondence between observational and theoretical terms, and in that sense is a measurement model. It is important to note that the notion of measurement intended here is broad and covers the entire spectrum from categorical unordered models to quantitative continuous models. Thus the meaning of the term 'measurement' is not limited to the classical interpretation of the term, which implies quantitative continuous structure (Michell, 1986), although models that are consistent with stricter interpretations may be derived from the general model as a special case (e.g., Rasch, 1960). Thus, the psychometric model is a measurement model in the sense that it coordinates theory with observation, but not in the sense that it assumes that human behavior can be successfully analyzed in terms of quantitative laws.

Measurement precision. Reliability relates to the psychometric model via the concept of measurement precision. The measurement precision of test scores is inversely related to the

variance of the observed scores conditional on a given position in the latent structure (Mellenbergh, 1996). Thus, the higher the variance of the conditional distribution of the observables, the lower the measurement precision of the observables. Note that measurement precision may, but need not, be identical for different positions of the latent structure. For instance, in the linear factor model, which has homoscedastic residuals, measurement precision is identical for all values of the latent variable; in contrast, in the Rasch (1960) model, it is highest for the latent position where the logistic regression of the observable has its inflection point and lower for positions father away from that point; note also that, at the point where this curve has its lowest measurement precision, it yields maximal information. The reliability of test scores, as proposed in classical test theory (Lord and Novick, 1968), is an unconditional index of measurement precision that can be straightforwardly derived from the conditional definition (Mellenbergh, 1996).

Item Response Theory. If the observed variables are responses to test items (e.g., items in an IQ test), then the measurement model falls under the scope of item response theory (IRT), a subfield of psychometrics that has come to play an important role in the analysis of educational tests. In IRT, the function that specifies the regression of the observed variables on the latent structure is known as an item characteristic curve (ICC). Usually, IRT models assume a unidimensional and continuous latent structure, which means the ICC is a smooth curve as in Figure 1. Because items in educational testing are typically scored dichotomously (1: correct, 0: incorrect), the ICC is bounded from above and from below, and hence is often modeled using a nonlinear function. The slope of the ICC at a given point on the latent scale is proportional to the ability of the item to discriminate between positions above and below that point, and thus determines the amount of information that the item yields at that point. Plotting the item information against the latent variable then gives the item information function (IIF).

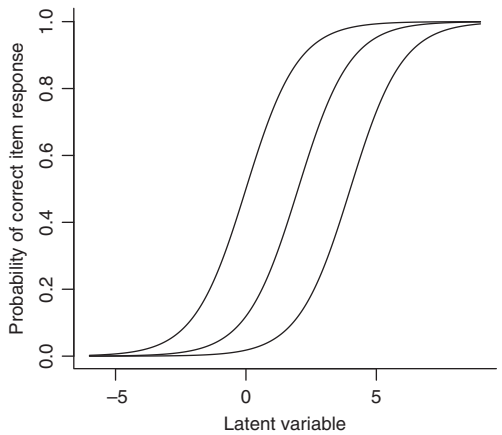

Figure 1 Item response theory relates the probability of a correct response to an item to a continuous latent variable through item characteristic curves (ICCs). When ICCs are parallel, as in this figure, items located further to the right have a lower probability of a correct item response for a given level of the latent variable and are typically interpreted as more difficult.

Adaptive testing. The IIF plays an important role in psychometrics because it can be used to regulate the selection of items. This idea forms the basis for adaptive testing (Van der Linden and Glas, 2000), a strategy of adaptively assembling tests that has become increasingly important with the advent of computerized test administration. In adaptive testing, items are administered sequentially and selected for administration adaptively, i.e., on the basis of the previous item responses of the respondent. This typically works as follows: at each point in the item administration process, examinee ability (the position in the latent structure) is estimated on the basis of the item responses given so far. The next item to be administrated is then chosen on the basis of the slope of the IIF at the estimated examinee ability, so that the administered items yield optimal information. In this way, tests can be shortened without affecting reliability.

Measurement invariance. Test scores are often used to select individuals (e.g., in job selection, student placement). Typically, these selection processes operate on populations that are heterogeneous with respect to background variables like sex and ethnicity. In these cases, considerations of fairness suggest that the test should function in the same way across different subpopulations, in the sense that it should not produce systematic bias in the test scores against a certain group. Such bias may, for instance, arise when an IQ test contains questions that are easier for test takers with a specific background regardless of their intelligence level. This may, for instance, occur if the test contains general knowledge questions that draw on a specific cultural background, so that they are more difficult for ethnic minorities for reasons independent of intelligence level. The concept of measurement invariance (Mellenbergh, 1989; Meredith, 1993) formalizes this idea and allows for testing it against empirical data.

Alternative psychometric models. Although the latent variable model is the primary race horse of current psychometrics, it is not the only available model to connect theoretical terms to observations. An important alternative in the psychometric literature concerns multidimensional scaling (MDS). MDS is a psychometric tool to infer the number of underlying dimensions in proximity data, i.e., data consisting of measures of similarity among pairs of stimuli (e.g., the degree in which different facial expressions are judged to be similar). Depending on whether the measurement scales of the similarity measures are either continuous or ordinal, one speaks of metric MDS (Torgerson, 1952) or nonmetric MDS (Shepard, 1962), respectively. In MDS, individual differences are taken into account by weighting the underlying dimensions differently across subjects (Carroll and Chang, 1970). By doing so, each subject receives a different weight for each dimension indicating which dimensions are more important for a subject in deciding which stimuli are similar. These weights enable identification of different subtypes of subjects. An important instance of MDS with individual differences is unfolding analyses, which is suitable for preference data. In unfolding analyses (Coombs, 1964), each subject is assumed to have an ideal point on the dimensions underlying the preference data. When a given stimulus is close to the subjects ideal point, that stimulus is preferred more.

In addition to MDS, other possibilities to make the connection between theory and data are by (1) representing the

construct as a common effect of the observed variables (Bollen and Lennox, 1991), (2) taking the construct to be a universe of behaviors, of which the observables are assumed to be a sample (Cronbach et al., 1972; McDonald, 1999), and (3) interpreting the construct as a causal system in which the observables influence each other (Cramer et al., 2010).

Recent Developments

The past couple of years, advances in psychometrics have mainly focused on extensions of the traditional measurement models to deal with more complex situations. Important extensions include the incorporation and development of multilevel and random effects structures in IRT models (Fox and Glas, 2001), factor models (Rabe-Hesketh et al., 2004), and latent class models (Lenk and DeSarbo, 2000). In these models, item parameters – considered fixed effects in the traditional measurement models – may themselves become random variables. This enables psychometric analyses of nested data, which are common in large-scale assessments. These extended models have been formulated within integrated frameworks that include the different models as special cases (Moustaki and Knott, 2000; Skrondal and Rabe-Hesketh, 2004).

Other extensions were motivated by the increasing popularity of computerized test administration. This causes response times to become available to the researcher in addition to the ordinary responses. Various suitable models have been proposed including models in which response times are modeled as collateral information (Van der Linden, 2007, 2009; Van Breukelen, 2005) or as manifestation of the underlying decision process (Tuerlinckx and De Boeck, 2005; Van der Maas et al., 2011).

Furthermore, recent advances in computer technology enabled more refined estimation techniques and model fit assessment, particularly for multidimensional IRT models (Béguin and Glas, 2001; Reckase, 2009), factor analyses of non-normal and categorical data (Satorra and Bentler, 2001; Molenaar et al., 2012), cognitive diagnosis models (Junker and Sijtsma, 2001; De la Torre and Douglas, 2004), unfolding analyses (Busing et al., 2005), and nonlinear factor models (Klein and Moosbrugger, 2000; Lee and Zhu, 2002).

On the conceptual side, three discussions have dominated the psychometric literature of the past decade. One concerns the status of psychometric measurement models and the relation between psychometrics and psychology (Michell, 1999; Borsboom, 2006). A second important discussion concerns the usefulness of Cronbach's alpha as a measure of reliability, which has been forcefully contested (Zinbarg et al., 2005; Sijtsma, 2009). Third, the proper definition of validity remains a contentious topic that continues to generate debate (Lissitz, 2009; Newton, 2012).

With respect to psychometric software, recent developments have mainly been inspired by the rise of the open source statistical software program R (R Development Core Team, 2012). This program enables psychometricians to develop their own model estimation packages and share these with other researchers. This trend is illustrated by two recent special issues of the *Journal of Statistical Software* (2007, 2012) that focused exclusively on psychometrics in R. Among the most popular R packages for psychometric analyses are *ltm* (Rizopoulos, 2006) and *eRm* (Mair and Hatzinger, 2007), which can fit various IRT models to data; *mokken* (Van der Ark, 2012), which may be used to fit nonparameteric IRT models to data; *sem* (Fox, 2006) and *lavaan* (Rosseel, 2012), which enables structural equation modeling; *smacof* which involves MDS (De Leeuw and Mair, 2009); and *qgraph* (Epskamp et al., 2012), which produces network visualizations of psychometric data and models. In addition, the *OpenMx* package (Boker et al., 2010) is a more general R package that can be used to fit various (extensions of) parametric models, including advanced possibilities like mixtures and joint modeling of discrete and continuous data.

See also: Classical (Psychometric) Test Theory; Reliability: Measurement; Selection Bias, Statistics of.

Bibliography

Andersen, E.B., 1973. A goodness of fit test for the Rasch model. Psychometrika 38 (1), 123–140.

Arbuckle, J.L., 2010. IBM SPSS Amos 19 User's Guide. SPSS, Chicago, IL.

Béguin, A.A., Glas, C.A.W., 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66 (4), 541–561.

Bentler, P.M., 2000. EQS 6 Structural Equations Program Manual. Multivariate Software Inc., Encino, CA.

Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 397–479.

Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 46 (4), 443–459.

Bollen, K.A., 2002. Latent variables in psychology and the social sciences. Annual Review of Psychology 53 (1), 605–634.

Bollen, K.A., Lennox, R., 1991. Conventional wisdom on measurement: a structural equation perspective. Psychological Bulletin 110 (2), 305–314.

Boker, S., Neale, M.C., Maes, H.H., Wilde, M., Spiegel, M., Brick, T., et al., 2010. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76 (2), 306–317.

Borsboom, D., 2005. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, Cambridge.

Borsboom, D., 2006. Attack of the psychometricians. Psychometrika 71 (3), 425–440.

Borsboom, D., Mellenbergh, G.J., Van Heerden, J., 2004. The concept of validity. Psychological Review 111 (4), 1061–1071.

Bowler, D.R., Miyazaki, T., Gillan, M.J., Ohno, T., Brázdová, V., et al., 2007. CONQUEST β. 1: User manual.

Busing, F.M.T.A., Groenen, P.J.F., Heiser, W., 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika 70, 71–98.

Carroll, J.D., Chang, J.J., 1970. Individual differences and multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35, 282–319.

Coombs, C., 1964. A Theory of Data. Wiley, New York.

Cramer, A.O.J., Waldorp, L.J., van der Maas, H., Borsboom, D., 2010. Comorbidity: a network perspective. Behavioral and Brain Sciences 33 (2–3), 137–193.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley, New York.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52 (4), 281–302.

De la Torre, J., Douglas, J.A., 2004. Higher order latent trait models for cognitive diagnosis. Psychometrika 69 (3), 333–353.

De Leeuw, J., Mair, P., 2009. Multidimensional scaling using majorization: SMACOF in R. Journal of Statistical Software 31 (3), 1–30.

Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D., Borsboom, D., 2012. qgraph: network visualizations of relationships in psychometric data. Journal of Statistical Software 48 (4), 1–18.

Fox, J., 2006. Teacher's corner: structural equation modeling with the sem package in R. Structural Equation Modeling 14, 465–486.

Fox, J.-P., Glas, C.A.W., 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66 (2), 271–288.

Galton, F., 1884. Measurement of character. Fortnightly Review 36, 179–185.

Jöreskog, K.G., 1971. Statistical analysis of sets of congeneric tests. Psychometrika 36 (2), 109–133.

Jöreskog, K.G., Sörbom, D., 1974. LISREL III. Scientific Software International, Inc., Chicago, IL.

Junker, B.W., Sijtsma, K., 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement 25 (3), 258–272.

Kelley, T.L., 1927. Interpretation of Educational Measurements. Macmillan, New York.

Klein, A., Moosbrugger, H., 2000. Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika 65 (4), 457–474.

Lazarsfeld, P.F., Henry, N.W., 1968. Latent Structure Analysis. Houghton Mifflin, Boston, MA.

Lee, S.-Y., Zhu, H.-T., 2002. Maximum likelihood estimation of nonlinear structural equation models. Psychometrika 67 (2), 189–210.

Lenk, P., DeSarbo, W., 2000. Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika 65 (1), 93–119.

Lissitz, R.W., 2009. The Concept of Validity: Revisions, New Directions, and Applications. Information Age, Charlotte, NC.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Mair, P., Hatzinger, R., 2007. Extended Rasch modeling: the eRm package for the application of IRT models in R. Journal of Statistical Software 20 (9), 1–20.

McDonald, R.P., 1999. Test Theory: A Unified Treatment. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Mellenbergh, G.J., 1989. Item bias and item response theory. International Journal of Educational Research 13 (2), 127–143.

Mellenbergh, G.J., 1994a. A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research 29 (3), 223–236.

Mellenbergh, G.J., 1994b. Generalized linear item response theory. Psychological Bulletin 115 (2), 300–307.

Mellenbergh, G.J., 1996. Measurement precision in test score and item response models. Psychological Methods 1 (3), 293–299.

Meredith, W., 1993. Measurement invariance, factor analysis, and factorial invariance. Psychometrika 58 (4), 525–543.

Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement. American Council on Education and National Council on Measurement in Education, Washington, DC, pp. 13–103.

Michell, J., 1986. Measurement scales and statistics: a clash of paradigms. Psychological Bulletin 100 (3), 398–407.

Michell, J., 1999. Measurement in Psychology: A Critical History of a Methodological Concept. Cambridge University Press, Cambridge.

Mokken, R.J., 1971. A Theory and Procedure of Scale Analysis. Mouton, The Hague.

Molenaar, D., Dolan, C.V., de Boeck, P., 2012. The heteroscedastic graded response model with a skewed latent trait: testing statistical and substantive hypotheses related to skewed item category functions. Psychometrika 77, 455–478.

Moustaki, I., Knott, M., 2000. Generalized latent trait models. Psychometrika 65 (3), 391–411.

Muthén, L.K., Muthén, B.O., 1998. Mplus User's Guide. Authors, Los Angeles.

Newton, P.E., 2012. Clarifying the consensus definition of validity. Measurement: Interdisciplinary Research & Perspective 10 (1–2), 1–29.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.Rproject.org.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modelling. Psychometrika 69 (2), 167–190.

Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Paedagogiske Institut, Copenhagen.

Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer, London.

Rizopoulos, D., 2006. ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software 17 (5), 1–25.

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48 (2), 1–36.

Satorra, A., Bentler, P.M., 2001. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66 (4), 507–514.

Sijtsma, K., 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74 (1), 107–120.

Sijtsma, K., 2011. Introduction to the measurement of psychological attributes. Measurement. Journal of the International Measurement Confederation 44 (7), 1209–1219.

Shepard, R.N., 1962. Analysis of proximities: multidimensional scaling with an unknown distance function, I. Psychometrika 27, 125–140.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.

Thissen, D., 1983. MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory [Computer Software]. Scientific Software International, Chicago, IL.

Tuerlinckx, F., De Boeck, P., 2005. Two interpretations of the discrimination parameter. Psychometrika 70 (4), 629–650.

Trout, J.D., 1999. Measurement. In: Newton-Smith, W.H. (Ed.), A Companion to the Philosophy of Science. Blackwell, Oxford, pp. 265–276.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–409.

Van der Ark, L.A., 2012. New developments in mokken scale analysis in R. Journal of Statistical Software 48 (5), 1–27.

Van Breukelen, G.J.P., 2005. Psychometric modeling of response speed and accuracy with mixed and conditional regression. Psychometrika 70 (2), 359–376.

Van der Linden, W.J., 2007. A hierarchical framework for modeling speed and accuracy on test items. Psychometrika 72 (3), 287–308.

Van der Linden, W.J., 2009. Conceptual issues in response-time modeling. Journal of Educational Measurement 46 (3), 247–272.

Van der Linden, W.J., Glas, C.A.W. (Eds.), 2000. Computerized Adaptive Testing: Theory and Practice. Kluwer, Norwell, MA.

Van der Maas, H.L.J., Molenaar, D., Maris, G., Kievit, R.A., Borsboom, D., 2011. Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. Psychological Review 118 (2), 339–356.

Vermunt, J.K., Magidson, J., 2000. Latent Gold: User's Manual. Statistical Innovations Inc., Boston.

Wilson, D.T., Wood, R., Gibbons, R., 1991. TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis [Computer Software]. Scientific Software International, Chicago.

Zimowski, M.F., Muraki, E., Mislevy, R.J., Bock, R.D., 1996. BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items [Computer Software]. Scientific Software International, Chicago.

Zinbarg, R.E., Revelle, W., Yovel, I., Li, W., 2005. Cronbach's α, Revelle's β, and McDonald's ωH: their relationships with each other and two alternative conceptualizations of reliability. Psychometrika 70 (1), 1–11.

Psychometrics: Classical Test Theory

James Algina, University of Florida, Gainesville, FL, USA
Hariharan Swaminathan, University of Connecticut, Storrs, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The principal developments in classical test theory are presented, including the classical true score model, the observed score random variable, the true score, the error score random variable, the reliability coefficient, the standard error of measurement, the effect of test length on statistical properties of test scores, single-administration coefficients of reliability, and models for degrees of equivalence among test scores.

Classical Test Theory

Assessment, evaluation, and research in social and behavioral sciences are typically based on the responses to stimuli presented to research participants or on observations of the participants in laboratory or natural settings. The research participants are more often than not people, but may be animals or some other entities such as schools. For simplicity we will refer to people in the following. The responses or observations are usually used to create scores on one or more variables. We will refer to the procedures for collecting the data as tests and the resulting scores as test scores. Following is a list of tasks, adapted from Gulliksen (1950), involved in constructing, scoring, and evaluating tests:

1. Writing and selecting the test items or specifying the behaviors to be observed and the procedure for making the observation.
2. Assigning scores to participant. Scaling and linking procedures are sometimes used to conduct this task.
3. Assessing the degree of accuracy of the test scores. Methods for assessing reliability and measurement error variance are used to address this task.
4. Assessing the degree to which interpretations of test scores and/or decisions based on the test scores are warranted. Methods for investigating validity are used to address this task.

Test theory comprises the statistical models used in accomplishing these tasks, the procedures for estimating the models, and the experimental designs used for collecting data to estimate the models. The body of literature on each is quite large, even when limited to classical test theory (CTT). In this article we focus primarily on statistical models for reliability and measurement error variance. One implication of this focus is that estimation of reliability and measurement error variance is, for the most part, not addressed. Developments that are not addressed in this article can be traced in the four editions of *Educational Measurement* (Lindquist, 1951; Thorndike, 1971; Linn, 1989; Brennan, 2006). Changes in the conception of validity can also be traced in the *Standards for Educational and Psychological Testing* published in 1954, 1966, 1974, 1985, and 1999. Algina and Penfield (2009) provided a brief account of these changes. See Borsboom (2005; 2006) for critiques of CTT and current conceptions of validity.

Classical Test Theory

CTT is a branch of test theory that is identified in contrast to modern approaches to test theory. According to Traub (1997) (1) the origins of test theory are in three intellectual achievements in the eighteenth and nineteenth centuries: "A recognition of the presence of errors in measurements, a conception of that error as a random variable, and a conception of correlation and how to index it" (p. 8) and (2) the first publication in CTT was Spearman (1904). Lord (1952), in a monograph on item response theory (IRT), referred to 'the classical theory of error' and distinguishes conclusions implied by that theory from conclusions implied by IRT, a prominent modern test theory. Du Mas (1956) referred to the 'newer theory of latent classes' in contrast to CTT. Loevinger (1957) wrote "The central concepts of classical test theory are reliability and validity" (p. 636) and commented on revisions to the concept of validity: "The rather sudden appearance of the term 'construct validity' indicates that the second concept of classical test theory is undergoing criticism and revision" (p. 636). In 1963, Cronbach, Rajatanam, and Gleser referred to the 'classical theory of reliability' to distinguish it from generalizability theory, a more general approach to assessing measurement reliability.

Although new developments in CTT are still being made, it is fair to say that seminal work took place between 1904 and the mid-1970s. Brown and Thompson (1921) provided comprehensive presentation of early work in the field. Thurstone (1932) was also influential. Gulliksen (1950) provided summary of results in CTT developed in the first half of the twentieth century. Lord and Novick (1968) summarized results in CTT as well as modern approaches to test theory. Koutsopoulos (1962; 1964) provided an axiomatic approach to the CTT as did Novick (1966). The books by McDonald (1999), Osterlind (2010), and Raykov and Marcoulides (2010) provide reviews of elements of both classical and modern test theory.

Classical True Score Model

A central idea of CTT is the classical true score model (CTSM), which expresses the observed score on a measurement (X) as the sum of a true score (T) and an error score (E):

$$X = T + E$$

There are different conceptions of the true score. For example, Lord and Novick (1968: p. 28), present three conceptions. In the conception they favor, when a particular person is measured at a particular point in time by using a particular test, there is a distribution of possible outcomes for the person's measurement. This distribution, which Lord and Novick call a propensity distribution, arises due to the random processes that operate in measuring a person's fluctuations of memory and attention in cognitive tests. The true score for a person is the mean of the propensity distribution. Consequently, the observed score for a person is an unbiased estimate of the person's true score. Note that the calculation of the true score is not possible unless the propensity distribution is known, which is typically not true. Therefore the true score variable is called a latent (i.e., hidden) variable. In passing we note that the term 'true score' does not imply that it would necessarily be a useful score for the assessment, evaluation, or research purposes for which the data were collected. For example, suppose a university develops a test to select applicants for admission. If true scores on the test do not accurately predict grade point average scores, then the true scores may not be useful for admission decisions. The error score is the difference between the observed score and the true score: $E = X - T$. Because E is the difference between a variable (X) and a constant (T), the propensity distribution for the error score is the same as the propensity distribution for the observed scores except that the mean of the former distribution is zero rather than T. The error score is also a latent variable. The variance of the propensity distribution is the person-specific error variance which we denote as $\sigma^2_{E_j}$ to emphasize that it applies to person j.

Principal Statistical Results Used in CTT

Although T and E are defined in terms of the propensity distribution, in CTT the means, variances, covariances, and correlation coefficients of X, T, and E over a population of persons are usually of interest and the principal results of CTT are expressed in terms of these parameters. Derivation of the principal results of CTT depends on four results from multivariate statistics. Although these results are presented below in terms of observed variables, they apply as well to true and error variables.

1. The correlation coefficient for two variables is the covariance between the variables divided by the product of the standard deviation of each of the variables. For example, if X and Y are two variables then $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, where ρ denotes the correlation, σ with two subscripts (i.e., σ_{XY}) is a covariance, and σ with a single subscript (e.g., σ_X) is a standard deviation. Rearranging the formula for the correlation coefficient results in $\sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$. The correlation of a variable with itself is equal to one and therefore $\sigma_{XX} = 1\sigma_X\sigma_X = \sigma^2_X$ the covariance of a variable with itself is the variance of the variable.
2. If X_1 to X_K are K variables, the mean of the linear composite $\tilde{X} = X_1 + \cdots + X_K$ is the sum of the means of the variables: $\mu_{\tilde{X}} = \mu_{X_1} + \cdots + \mu_{X_K}$. The ellipsis between the plus signs denotes all of the intervening terms. The variance of a linear composite is the sum of the variances of the variables plus twice the sum of the covariances between the variables: $\sigma^2_{\tilde{X}} = \sigma^2_{X_1} + \cdots + \sigma^2_{X_K} + 2(\sigma_{X_1X_2} + \cdots + \sigma_{X_{K-1}X_K})$.
3. The covariance between one linear composite and a second is the sum of the covariances of each variable in the first composite with each variable in the second. If $\tilde{Y} = Y_1 + \cdots + Y_J$, then

$$\sigma_{\tilde{X}\tilde{Y}} = \sigma_{X_1Y_1} + \sigma_{X_1Y_2} + \cdots + \sigma_{X_1Y_J} + \sigma_{X_2Y_1} + \sigma_{X_2Y_2} + \cdots + \sigma_{X_2Y_J} + \cdots + \sigma_{X_KY_1} + \sigma_{X_KY_2} + \cdots + \sigma_{X_KY_J}$$

In the special case of the covariance between one variable and a sum of variables, the covariance is the sum of the covariances of the variable with each of the variables in the sum. For example,

$$\sigma_{\tilde{X}Y_1} = \sigma_{X_1Y_1} + \cdots + \sigma_{X_KY_1}$$

4. The regression relationship between two variables is $Y = \mu_Y + (\sigma_Y/\sigma_X)\rho_{XY}(X - \mu_X) + \varepsilon$. The predicted value of the Y variable is $\hat{Y} = \mu_Y + (\sigma_Y/\sigma_X)\rho_{XY}(X - \mu_X)$.

Principal Implications of the CTSM

Mean of the Observed Score Variable

From $X = T + E$, we find $\mu_X = \mu_T + \mu_E$. Because the average error score is zero for each person, the average error score is zero over a population of persons and, therefore $\mu_X = \mu_T$. Because a sample mean is an unbiased estimator of a population mean, the sample mean of an observed score is an unbiased estimator of the population true score mean. That is, the average observed score provides correct information about the average true score.

Variance of the Error Score Variable

As noted earlier, the error variance $\sigma^2_{E_j}$ is the error variance for person j. The variance of E over a population of persons is the average value of $\sigma^2_{E_j}$ for that population and is denoted by σ^2_E.

Variance of the Observed Score Variable

An important feature of the distribution of X, T, and E is the relationship between T and E. Figure 1 is a plot of error scores against true scores that conform to the CTSM. The plot, which is for a sample size of 1000, indicates that the correlation of the error and true scores is near zero. In fact the sample correlation is 0.005. It can be shown that the CTSM implies that the population correlation of true and error scores, denoted by ρ_{TE}, is exactly zero. From $X = T + E$, $\sigma^2_X = \sigma^2_T + \sigma^2_E + 2\sigma_T\sigma_E\rho_{TE}$. Because $\rho_{TE} = 0$, $\sigma^2_X = \sigma^2_T + \sigma^2_E$. That is, under the CTSM, observed score variance is equal to true score variance plus error score variance.

Reliability Coefficient

Consider a hypothetical situation A in which $\sigma^2_T = 0$; then all persons have the same true score, $\sigma^2_X = \sigma^2_E$, and all differences

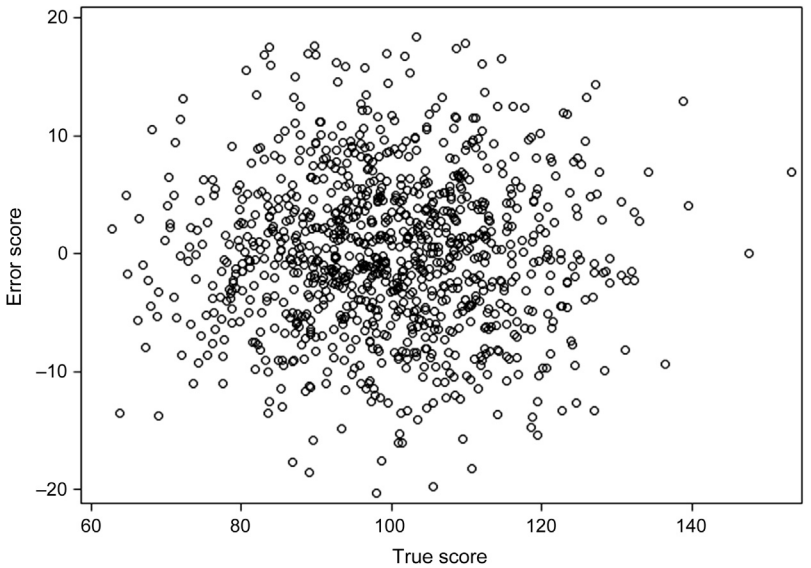

Figure 1 Plot of Error Scores vs True Scores.

among observed scores are due to errors of measurement. Any use of the observed scores to compare different people is fatally flawed because the differences are merely due to random errors of measurement and should be ignored. Now consider a hypothetical situation B in which $\sigma_E^2 = 0$; then $\sigma_X^2 = \sigma_T^2$, all differences among observed scores are due to true score differences: any comparison of observed scores for different people is equivalent to comparison of true scores and is not influenced by measurement error. Considering these observations, a useful parameter in CTT is the ratio of true score variance to observed score variance: σ_T^2/σ_X^2. This ratio is called the reliability coefficient. In situation A, the reliability coefficient is zero, which indicates that any comparison of observed scores will be problematic. In situation B, the reliability coefficient is one indicating that measurement error does not influence the comparison of scores. As these examples suggest, reliability coefficients near one are desirable whenever test scores are used to compare individuals. It can be shown that the correlation between the observed and true score variables is $\rho_{XT} = \sigma_T/\sigma_X$, which is called the index of reliability. Squaring the index gives $\rho_{XT}^2 = \sigma_T^2/\sigma_X^2$: in the CTSM the squared correlation between the observed and true score variables is equal to the reliability coefficient. Further, using $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, we find that $\rho_{XT}^2 = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$.

Effect of Heterogeneity on Reliability

Suppose an achievement testing company has developed a reading comprehension test that can be used with children in the following grade range: beginning of second grade to the end of third grade. One researcher administers the test to children who are beginning second grade and another administers the test to children who are either beginning second grade or finishing third grade. We would expect the true score variance to be larger for the second group because the mixed-grade group is likely to have more variation in reading comprehension than does the single-grade group. For which group of children will the reliability coefficient likely be larger? Suppose it is known that the error variance is unaffected by the grade of the children, then $\rho_{XT}^2 = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$ indicates that reliability will be larger for the children in the mixed-grade group because true score variance is likely to be larger for that group. As this example illustrates, reliability is a function of the heterogeneity of the individuals being measured and therefore is not solely a property of the test itself. For this reason, we should refer to reliability of test scores, rather than reliability of the test.

Standard Error of Measurement

The standard error of measurement is the standard deviation of the error scores: σ_E. To a rough approximation 95% of the error scores fall in the interval $[-2\sigma_E, 2\sigma_E]$. Thus, the standard error of measurement provides information about how large the error scores are likely to be. Suppose a psychologist administers a self-esteem instrument comprising ten 7-point items that are scored 1–7. If the score on the test is the sum of the item scores, the potential range of scores is then 10–70. If $\sigma_E = 1$, the psychologist knows that error scores tend to be quite small, particularly in comparison to the possible range of observed scores. Using $\rho_{XT}^2 = \sigma_T^2/\sigma_X^2$ and $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, it can be shown that $\sigma_E = \sigma_X\sqrt{1 - \rho_{XT}^2}$. Thus if we know the reliability coefficient, we can calculate the standard error of measurement. CTT methods have been developed to estimate the standard error of measurement for a particular person and for all persons with a particular observed score. See Feldt and Qualls (1996) for a method and literature review.

Correction for Attenuation

Spearman (1904) was motivated by a concern that measurement error attenuates the correlation between scores on two variables. In learning goal theory, a mastery goal refers to a goal of learning material for the sake of learning it. Educational

psychologists have been interested in the relationship between mastery goals and achievement test scores. Let X denote mastery goals and Y denote achievement and assume that measurements of each conform to the CTSM: $X = T_X + E_X$ and $Y = T_Y + E_Y$. In addition assume that the error score variables are not correlated. Applying the definition of the correlation coefficient to the true score variables and using the definition of the reliability coefficient, it can be shown that $\rho_{XY} = \rho_{T_X T_Y}\rho_{XT_X}\rho_{YT_Y}$. Because an index of reliability must be between 0 and 1, the equation shows that if at least one of the indices is smaller than 1, the correlation between the observed score variables must be attenuated. That is, the correlation between the observed score variables must be closer to zero than is the correlation between the true score variables.

Estimated True Score

The estimated true score is given by $\widehat{T} = \mu_X + \rho^2_{XT}(X - \mu_X)$, where $\widehat{T}$ is the estimated true score. This equation is derived by considering the linear regression relationship between T and X. The equation shows that if the reliability coefficient is zero, the estimated true score for each person in the population is equal to the observed score mean. This is appropriate because when the reliability coefficient is zero, all differences among observed scores are due to errors of measurement and the differences should be ignored. The equation also shows that if the reliability coefficient is 1, $\widehat{T}$ is equal to the observed score. Again this is appropriate because when the reliability coefficient is 1, the observed score is equal to the true score. The estimated true score has two sources of error as an estimate of T. First, when the reliability coefficient is between 0 and 1, the estimated true score is a biased estimator of the true score. By contrast the observed score is an unbiased estimator. Thus bias favors the observed score over the estimated true score. Second, the variance of $\widehat{T} - T$, the error in using $\widehat{T}$ as an estimate of T, is not larger than the variance of $E = X - T$, the error in using X as an estimate of T. Error variance favors the estimated true score. These two sources of error can be combined in a parameter known as the mean square error (MSE). The MSE for $\widehat{T}$ cannot be larger than the MSE for X, but can be smaller. Using the MSE as a criterion for picking an estimator, $\widehat{T}$ should be used. In practice $\widehat{T}$ is used infrequently if at all. There are two reasons. First, $\widehat{T}$ and X are perfectly correlated, so for many purposes it is irrelevant which is used. Second, if X is larger than its mean then $\widehat{T} < X$ and if X is smaller than its mean then $\widehat{T} > X$. Using the estimated true score requires explaining why more capable examinees are penalized and less capable examinees are advantaged when estimated true scores are used.

The estimated true score given above may be expressed as $\widehat{T} = \rho^2_{XT}X + (1 - \rho^2_{XT})\mu_X$. This form was first given by Kelly (1942). In recent years, this formula for the estimate of true score has been recognized as the prototype of the Bayes (or empirical Bayes) procedure for estimating a random variable (Novick and Jackson, 1974).

Parallel Measurements Model

Much of the early development in CTT was based on the parallel measurements model (PMM). The PMM concerns two or more measurements that conform to the CTSM. The PMM assumes that (1) each participant has the same true score on the measurements:

$$X_k = T + E_k$$

where the subscript indicates the kth of K measurements, (2) the error score variance is equal for all tests, and (3) error score variables are uncorrelated for all pairs of tests. Based on these assumptions and using the statistical results presented above, it can be shown that the following are equal for all K measurements (1) true score means, (2) observed score means, (3) true score variances, (4) error score variances, (5) observed score variances, (6) reliability coefficients, and (7) the correlation coefficient of each of the parallel measurements with any other variable, a variable we denote as Y. These equalities imply that parallel measurements are interchangeable. The equalities are summarized in Table 1 and compared to equalities that can be derived under alternative measurement models. Derivation of (7) assumes that Y conforms to the CTSM and that the error score variable for Y is uncorrelated with the error score variables for each of the parallel measurements.

Correlation Between Parallel Measurements

An important implication of the PMM is that $\rho_{X_k X_{k'}} = \sigma^2_T/(\sigma^2_T + \sigma^2_E)$: the correlation between two parallel measurements is equal to the reliability for each measurement. Assuming parallel measurements can be identified, this provides a basis for using observed score variables to estimate the reliability coefficient and the standard error of measurement. If the measurement error variables for X_k and $X_{k'}$ are not uncorrelated then $\rho_{X_k X_{k'}} = (\sigma^2_T + \sigma_{E_k E_{k'}})/(\sigma^2_T + \sigma^2_E)$ and may be larger or smaller than the reliability coefficient.

Effect of Test Length

Let $\tilde{X} = X_1 + \cdots + X_K$. The variable $\tilde{X}$ can be thought of as the result of using a longer test than any one of the K tests which result in the variables $X_1,\ldots,X_K$, which are assumed to be parallel. Let parameters for the longer test be denoted by using $\tilde{X}$,$\tilde{T}$, and $\tilde{E}$ subscripts and parameters for any one of the parallel measurements be denoted by using X, T, and E subscripts. The following are the effects of lengthening the test (1) true score mean: $\mu_{\tilde{T}} = K\mu_T$, (2) observed score mean: $\mu_{\tilde{X}} = K\mu_X$, (3) true score variance: $\sigma^2_{\tilde{T}} = K^2\sigma^2_T$, (4) error score variance: $\sigma^2_{\tilde{E}} = K\sigma^2_E$,

Table 1 Summary of equality features of five measurement models

Equality	*PMM*	*EPMM*	*TEMM*	*ETEMM*	*CMM*
Observed score means	Yes	No	Yes	No	No
True score means	Yes	No	Yes	No	No
Observed score variance	Yes	Yes	No	No	No
True score variance	Yes	Yes	Yes	Yes	No
Error score variance	Yes	Yes	No	No	No
Reliability coefficient	Yes	Yes	No	No	No
Correlation with criterion variable	Yes	Yes	No	No	No

PMM, parallel measurements model; EPMM, essentially parallel measurements model; TEMM, tau-equivalent measurements model; ETEMM, essentially tau-equivalent measurements model; CMM, congeneric measurements model.

(5) observed score variance: $\sigma^2_{\tilde{X}} = K\sigma^2_X[1 + (K-1)\rho^2_{XT}]$, (6) reliability: $\rho^2_{\tilde{X}\tilde{T}} = K\rho^2_{XT}/[1 + (K-1)\rho^2_{XT}]$, and (7) correlation with a criterion variable, denoted by Y: $\rho_{\tilde{X}Y} = \sqrt{K\rho^2_{XY}/(1 + (K-1)\rho^2_{XT})}$.

Generalized Spearman–Brown Formula

A comparison of the formulas in (3) and (4) in the preceding paragraph shows that, as a test is lengthened, the true score variances increase more rapidly than the error score variances. As a result the reliability of the scores on the longer test is larger than the reliability for the scores on the shorter tests. The effect of test length on reliability is summarized by the formula in (6), which is known as the generalized Spearman–Brown formula. A few calculations with example values of ρ^2_{XT} and K will show that, provided that $\rho^2_{X_kT}$ is between 0 and 1, (1) the reliability of scores on $\tilde{X}$ is larger than the reliability of scores on any one of the parallel measurements and (2) the effect of lengthening the test is larger when the $\rho^2_{X_kT}$ is smaller. However, these conclusions may not be correct if the error score variables are correlated for any of the parallel measurements.

Effect of Test Length on Criterion-Related Evidence of Validity

The correlation coefficient of a test with another variable is often called criterion-related evidence of validity for the test. For example if a company wants to use a test to select salespeople, the correlation of scores on the test with sales figure (i.e., the criterion) is criterion-related evidence of the validity. The formula in (7) shows the correlation of the longer test as a function of the number of parallel measurements, the correlation of each of these measurements with the criterion variable, and the reliability coefficient for each of the parallel measurements. Lengthening the test increases the correlation with the criterion. The impact of lengthening the test is larger when the reliability of the parallel measurements and the correlation of the parallel measurements with the criterion are smaller. The formula can be misleading if the error score variable for Y is correlated with the error score variables for one or more of X_1–X_K.

Reliability from a Single Administration of a Test

Suppose a test can be split into two half-tests that can be scored independently. For example, a test comprising 20 items could be split into two 10-item half-tests. Denote scores on the test as $\tilde{X}$ and on the half-tests as X_1 and X_2. Then $\tilde{X} = X_1 + X_2$. Assume the half-tests yield parallel measurements. Then the correlation coefficient for X_1 and X_2 is the reliability coefficient for each and we can use the generalized Spearman–Brown formula with $K = 2$ to express the reliability coefficient for $\tilde{X}$: $\rho^2_{\tilde{X}\tilde{T}} = 2\rho^2_{XT}/[1 + \rho^2_{XT}]$. This simplified formula is the Spearman–Brown formula and provides a method for computing the reliability coefficient for a test without requiring administration of more than one test. It is an example of a single-administration coefficient for test reliability. An alternative to the generalized Spearman–Brown formula is coefficient alpha: $\alpha = [K/(K-1)][1 - (\sigma^2_{X_1} + \cdots + \sigma^2_{X_K})/\sigma^2_X]$. Cronbach (1951) provided the name coefficient alpha, but noted that it has a long history in test theory dating back to Kuder and Richardson (1937), who derived it for the special case of dichotomously scored items. It can be shown that under the PMM, the generalized Spearman–Brown formula and the formula for coefficient alpha are equivalent. If $K = 2$ then $\alpha = 2[1 - (\sigma^2_{X_1} + \sigma^2_{X_2})/\sigma^2_X]$. Formulas that are equivalent to coefficient alpha with $K = 2$ have been attributed to Flanagan (Kelly, 1942), Guttman (1945), Mosier (1941), and Rulon (1939). Both the Spearman–Brown formula and coefficient alpha can be smaller than or larger than the reliability coefficient, if the error score variables are correlated for some or all of X_1 to X_K. (See, for example, Lucke, 2005; Green and Hershberger, 2000; or Raykov, 2001.) Sijtsma (2009) presented a general critique of coefficient alpha.

Equivalence Models

Beginning in the 1960s and continuing through the early 1990s, a series of models collectively referred to as equivalence models were introduced. These models are generalizations of the PMM obtained by relaxing assumptions made in the PMM. Haertel (2006) provides a valuable account of these developments. In this section, We describe these models and present important differences between the implications of these models and the implications of the PMM.

The Essentially Parallel Measurements Model

The name essentially parallel measurements model (EPMM) was coined by Millsap and Everson (1991). The model can be found as early as Kristof (1969). The EPMM is a modification of the PMM in which true score variables for two observed score variables X_k and $X_{k'}$ are related by an additive constant:

$$T_{k'} = c_{k'k} + T_k$$

As shown in Table 1, all of the equalities based on the PMM hold for the EPMM, except the result that the measurements have equal true score means and equal observed score means. The fact that under the EPMM, measurements do not have equal means implies that the measurements are not interchangeable. For example, if several final examinations are available in a course and yield essentially parallel measurements, students would prefer to take the exam that results in the largest mean. Using $\mu_{X_k} = \mu_{T_k}$ and the CTSM, it can be shown that $c_{kk'} = \mu_{X_k} - \mu_{X_{k'}}$, which indicates that scores for measurement k' can be transformed to the scale for measurement k by using $X_k = X_{k'} - \mu_{X_{k'}} + \mu_{X_k}$. This transformation is an example of test linking. Holland and Dorans (2006) describe two types of test linking: scale alignment and equating. Linking is called equating when the transformed scores are interchangeable because the test measures the same construct, the transformed scores are on comparable scales, and the scores are equally reliable. (See Holland and Dorans, 2006, for other criteria required for linking to be equating.) If the measurements conform to the EPMM model and the data are transformed by using $X_{k'} - \mu_{X_{k'}} + \mu_{X_k}$, the scales have been equated because X_k and $X_{k'}$ measure the

same variable, have equal reliability, and the transformed scores have been placed on the same scale.

The Tau-Equivalent Measurements Model and Essentially Tau-Equivalent Measurements Model

The tau-equivalent measurements model (TEMM) is a modification of the PMM in which the assumption of equal error score variances is dropped. Similarly, the essentially tau-equivalent measurements model (ETEMM) is a modification of the EPMM in which the assumption of equal error score variances is dropped. The TEMM and ETEMM were introduced by Novick and Lewis (1967). **Table 1** shows the equality results that do not hold for these two models. A particularly important implication of these models is that different measurements do not have the same reliability coefficient. We will trace the implication of this result for the ETEMM; the same implication applies to the TEMM. Let $X_1 = T_1 + E_1$ and $X_2 = T_2 + E_2$ denote two essentially tau-equivalent measurements. The reliability coefficients for the two measurements are $\rho^2_{X_kT_k} = \sigma^2_T/\sigma^2_{X_k}, \quad k = 1, 2$. The true score variance is the same for each of the two measurements because the true score variables differ by an additive constant. The observed score variances are not equal because the error score variances are not equal. The correlation between the two measurements is $\rho_{X_1X_2} = \sigma^2_T/(\sigma_{X_1}\sigma_{X_2})$ and is not equal to either of the reliability coefficients. In fact $\sigma^2_T/(\sigma_{X_1}\sigma_{X_2})$ is equal to the geometric mean of the reliability coefficients. Therefore the correlation coefficient between two tests that conform to the ETEMM must be larger than the smaller reliability coefficient and smaller than the larger reliability coefficient. It can be shown, for example, that $\rho^2_{X_1T_1} = (\sigma_{X_2}/\sigma_{X_1})\rho_{X_1X_2}$ which provides a basis for computing the reliability coefficient for X_1. The reliability coefficient for X_2 is $\rho^2_{X_2T_2} = (\sigma_{X_1}/\sigma_{X_2})\rho_{X_1X_2}$. Under the ETEMM, the variable with the larger observed score variance has the larger error score variance and, as the expressions for the reliability coefficients imply, has the smaller reliability coefficient. If a test can be split into two halves that conform to the ETEMM, coefficient alpha can be applied to scores on the two half-tests and equals the reliability coefficient for the test scores on $X = X_1 + X_2$ (Novick and Lewis, 1967). If the test can be split into K parts that yield ETEMM measurements then the general form of coefficient alpha is equal to the reliability of the test. The Spearman–Brown formulas do not apply to these models because the reliability coefficients are not the equal for all measurements. Finally measurements that conform to the TEMM or ETEMM cannot be equated because the measurements do not have the same error score variance. The measurement scales can be aligned, but even after the alignment the scores are not interchangeable because of the differences in error variance. In general, among measurements that conform to the TEMM or ETEMM, the measurement with the smaller error score variance is preferred.

The Congeneric Measurements Model

The Congeneric Measurements Model (CMM), introduced by Joreskog (1971), is the most general of the equivalence models. In the CMM, $X_k = T_k + E_K$, that is, the true score variable is not the same for all measurements. However, $T_k = \lambda_k(T_{k'} - c_{k'})/\lambda_k + c_{k'}$. That is, the true score variable for measurement k is linearly related to the true score variable for measurement k'. Scores on the various measurements can be linked but are not interchangeable because they have different measurement error variances (see **Table 1**). Measurements that conform to the CMM have different reliability coefficients. Consequently, the correlation between two congeneric measurements is not equal to the reliability coefficient of either and the Spearman–Brown formula should not be used. It can be shown that coefficient alpha for a sum of congeneric measurements is smaller than the reliability coefficient for the sum. Contributions to reliability theory under the CMM have been provided by Joreskog (1971), Kristof (1971; 1974), and Feldt (1975). Among others, Feldt (1975) provides contributions for a special case of the CMM in which $\sigma^2_{E_k} = \lambda_k\sigma^2_{\tilde{E}}$ and $\sigma^2_{\tilde{E}}$ is the variance of $\tilde{E} = E_1 + \cdots + E_K$.

Lower Bounds to the Reliability of a Composite Variable

Recall that a composite variable is $\tilde{X} = X_1 + \cdots + X_K$, a sum of observed score variables. The observed variables can be separately scored parts of a test that sum to the score on the test. Therefore, the lower bounds provide a method for approximating the reliability of the test from a single administration of the test. Novick and Lewis (1967) showed that a necessary and sufficient condition for coefficient alpha to be equal to the reliability of the composite $\tilde{X}$ is that the measurements conform to the ETEMM. This implies that for the models that are special cases of the ETEMM, that is the PMM, EPMM, and TEMM, coefficient alpha is equal to the reliability of the composite $\tilde{X}$. For more general models, alpha is smaller than the reliability coefficient for the composite $\tilde{X}$. These claims are based on the assumption that the error variables for the components of $\tilde{X}$ are uncorrelated.

Guttman (1945) introduced six lower bounds to reliability. One of these, λ_3, is equal to α and one of the lower bounds, $\lambda_2 = \{\sigma^2_{\tilde{X}} - (\sigma^2_{X_1} + \cdots + \sigma^2_{X_K}) + \sqrt{[K/K-1][\sigma^2_{\tilde{X}} - (\sigma^2_{X_1} + \cdots + \sigma^2_{X_K})]}\}/\sigma^2_{\tilde{X}}$ is always at least as large as α. Ten Berge and Zegers (1978) introduced an infinite series of lower bounds in which coefficients α and λ_2 are the first two terms and each term in the series is at least as large as the preceding term. According to Ten Berge and Zegers the terms beyond α and λ_2, typically, do not get substantially larger than the first two terms. Bentler (1972) introduced the dimension-free lower bound to reliability. Jackson and Agunwamba (1977) introduced the greatest lower bound (glb) to reliability, a value such that the reliability of $\tilde{X}$ must be at least as large as that value. For example, a glb equal to 0.90 implies the reliability of the test scores is at least 0.90. Unfortunately, when an estimate of glb is calculated from sample data, the estimate is positively biased unless the sample size is very large. Consequently, an estimated glb that is equal 0.90 may be a spuriously optimistic lower bound for reliability and reliability of the composite may be much smaller. This bias has hampered the use of the glb. However, Li and Bentler (2011)

have published corrections to the bias, which may make the glb more useful.

The common factor model for the observed variables $X_1,\ldots,X_K$ is

$$X_k = \mu_k + \lambda_{k1}F_1 + \cdots + \lambda_{kJ}F_J + S_k + E_k$$

where F_1 to F_J ($J < K$) are common factors, S_k is a specific factor, E_k is a measurement error variable, $\lambda_{k1},\ldots,\lambda_{kP}$ are factor loadings, and $\mu_1,\ldots,\mu_K$ are the means of the observed variables. The factor loadings are measures of the relationship between the observed variables and the common factors. A common factor, F_j, is a latent variable and has a relationship with at least two observed variables. The specific factor, S_k, is a latent variable that is only related to X_k. The common factors can be mutually correlated, but are not correlated with either the specific factors or the error variables. The specific factors and error variables are also uncorrelated. Defining $C_k = \mu_k + \lambda_{k1}F_1 + \cdots + \lambda_{kJ}F_J$ as the common part of X_k, we can write $X_k = C_k + S_k + E_k$. Then $\tilde{X} = \tilde{C} + \tilde{S} + \tilde{E}$, where $\tilde{C}$, $\tilde{S}$, and $\tilde{E}$ are the sums of C_k, S_k, and E_k variables, respectively. The composite common part (i.e., $\tilde{C}$) is uncorrelated with the composite specific factor (i.e., $\tilde{S}$) and the error of measurement for the composite (i.e., $\tilde{E}$). Because $\tilde{X} = \tilde{T} + \tilde{E}$, $\tilde{T} = \tilde{C} + \tilde{S}$. That, is the composite specific factor, $\tilde{S}$, is a component of $\tilde{T}$. Because $\tilde{S}$ is uncorrelated with $\tilde{C}$, $\sigma^2_{\tilde{T}} = \sigma^2_{\tilde{C}} + \sigma^2_{\tilde{S}}$ and the reliability coefficient is $\rho^2_{\tilde{X}\tilde{T}} = (\sigma^2_{\tilde{C}} + \sigma^2_{\tilde{S}})/\sigma^2_{\tilde{X}}$. To estimate $\sigma^2_{\tilde{C}} + \sigma^2_{\tilde{S}}$ and the reliability coefficient, it is necessary to estimate the factor loadings, variances of the common factors, covariances of the common factors, and variances of the specific factors. Factor analysis provides a method for estimating these parameters, though in most cases the variances for the specific factors cannot be estimated. When $\sigma^2_{\tilde{S}}$ cannot be estimated, the results of the factor analysis can be used to estimate a lower bound to reliability due to McDonald: $\omega = \sigma^2_{\tilde{C}}/\sigma^2_{\tilde{X}}$ (see McDonald, 1999). Coefficient ω is a lower bound to reliability because $\sigma^2_{\tilde{C}} \leq \sigma^2_{\tilde{T}}$. Revelle and Zinbarg (2009) present examples in which an estimate of ω exceeds the estimate of the glb, a demonstration that raises the issue of whether estimates of ω are positively biased in small samples. Ten Berge and Socan (2007) argue that the glb is preferable to ω because the glb is formulated in terms of true and error score variance whereas ω will only be equal to the reliability coefficient, if $\sigma^2_{\tilde{S}} = 0$.

The equivalence models presented earlier are special cases of the common factor model. All of the equivalence models assume that measures of the same construct have a nonzero loading on only one factor and that there are no specific factors. Thus, for the equivalence models, the common factor model specializes to

$$X_k = \mu_k + \lambda_{kr}F_r + E_k$$

and implies that if all observed variables measure the same true score variable there will be just one factor. If there are several true score variables and each observed score variable measures just one of these, there will be more than one factor in the model, but measures of the same true score variable will have a nonzero loading on only one factor. **Table 2** shows the equations for the equivalence models. In **Table 2**, a k subscript on a mean or on a factor loading means that the parameter is not restricted to be constant across measures of the same construct. The models are presented from the least restrictive model (CMM) to the most restrictive model (PMM). For the CMM, the true score variable is equal to $\mu_k + \lambda_{kr}F_r$. For the other models the true score variable is equal to the restricted form of $\mu_k + \lambda_{kr}F_r$. See Raykov and Marcoulides (2010) for additional information about using factor analysis to estimate reliability coefficients.

Table 2 Equations and restrictions on error variances for five test theory models

Model	*Equation*	*Error score variance*
Congeneric measurements model	$X_k = \mu_k + \lambda_{kr}F_r + E_k$	Not equal
Essentially equivalent measurements model	$X_k = \mu_k + \lambda_r F_r + E_k$	Not equal
Tau-equivalent measurements model	$X_k = \mu + \lambda_r F + E_k$	Not equal
Essentially parallel measurements model	$X_k = \mu_k + \lambda_r F + E_k$	Equal
Parallel measurements model	$X_k = \mu + \lambda_r F + E_k$	Equal

Summary

CTT has a long history with a major focus on the definition of the reliability coefficients; the relationship between reliability and various parameters such as coefficient alpha, λ_2, and ω; the correlation between parallel or nonparallel measurements; the effects of lengthening tests by adding parallel measurements; and the effect of measurement error on correlation coefficients. Despite this long history, two difficult problems in applying CTT are determining the best method for estimating reliability for a given data set and assessing the degree to which applications of CTT produce misleading results when the assumptions of the theory are not met. Despite these problems, CTT remains as the foundation for more complex measurement theories.

See also: Content Validity; Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Generalizability Theory; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Negotiations: Statistical Aspects; Reliability and Validity Assessment: New Approaches; Reliability: Measurement; Scaling and Classification in Social Measurement; Test Theory: Applied Probabilistic Measurement Structures.

Bibliography

Algina, J., Penfield, R.D., 2009. Classical test theory. In: Millsap, R.E., Maydeu-Olivares, A. (Eds.), Handbook of Quantitative Methods in Psychology. Sage Publications, Thousand Oaks, CA, pp. 93–122.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, 1954. Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin 51 (Suppl.).

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, 1966. Standard for Educational and Psychological Tests. American Psychological Association, Washington, DC.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, 1974. Standard for Educational and Psychological Tests. American Psychological Association, Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurements in Education, 1985. Standard for Educational and Psychological Testing. American Psychological Association, Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999. Standard for Educational and Psychological Testing. American Psychological Association, Washington, DC.

Bentler, P.M., 1972. A lower-bound method for the dimension-free measurement of internal consistency. Social Science Research 1, 343–357.

Boorsboom, D., 2005a. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, Cambridge.

Boorsboom, D., 2006. The attack of the psychometricians. Psychometrika 71, 425–440.

Brennan, R.L. (Ed.), 2006. Educational Measurement, fourth ed. Praeger Publishers, Westport, CT.

Brown, W., Thompson, G.H., 1921. The Essentials of Mental Measurement. Cambridge University Press, New York.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Cronbach, L.J., Rajaratnam, N., Gleser, G.C., 1963. Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology 16, 137–163.

Du Mas, F.M., 1956. Behavioral scaling of personality tests. Psychological Reports 2, 195–204.

Feldt, L., 1975. Estimation of the reliability of a test divided into parts of unequal length. Psychometrika 40, 557–561.

Feldt, L., Qualls, A. l, 1996. Estimation of measurement error variance at specific score levels. Journal of Educational Measurement 33, 141–156.

Green, S.B., Hershberger, S.L., 2000. Correlated errors in true score models and their effect on coefficient alpha. Structural Equation Modeling 7, 251–270.

Gulliksen, H., 1950. Theory of Mental Tests. Wiley, New York.

Guttman, L., 1945. A basis for analyzing test-retest reliability. Psychometrika 10, 255–282.

Haertel, E.H., 2006. Reliability. In: Brennan, R.L. (Ed.), Educational Measurement, fourth ed. Praeger Publishers, Westport, CT, pp. 65–110.

Holland, P.H., Dorans, N.J., 2006. Linking and equating. In: Brennan, R.L. (Ed.), Educational Measurement, fourth ed. Praeger Publishers, Westport, CT, pp. 187–220.

Jackson, P., Agunwamba, C., 1977. Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I: algebraic lower bounds. Psychometrika 42, 567–578.

Joreskog, K.G., 1971. Statistical analysis of a set of congeneric tests. Psychometrika 36, 109–133.

Kelly, T.L., 1942. The reliability coefficient. Psychometrika 7, 75–84.

Koutsopoulos, C.J., 1962. The Mathematical Foundations of Classical Test Theory: An Axiomatic Approach I. Educational Testing Service Research Bulletin 62-17. Educational Testing Service, Princeton, NJ.

Koutsopoulos, C.J., 1964. The Mathematical Foundations of Classical Test Theory: An Axiomatic Approach It. Educational Testing Service Research Memorandum 64-3. Educational Testing Service, Princeton, NJ.

Kristof, W., 1969. Estimation of true score and error variance for tests under various equivalence assumptions. Psychometrika 34, 489–507.

Kristof, W., 1971. On the theory of tests which differ only in length. Psychometrika 36, 207–225.

Kristof, W., 1974. Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika 39, 491–499.

Kuder, G.F., Richardson, M.W., 1937. The theory of estimation of test reliability. Psychometrika 2, 151–160.

Li, L., Bentler, P.M., 2011. The greatest lower bound to reliability: corrected and resampling estimators. Modeling and Analysis of Data 1, 87–104.

Lindquist, E.F. (Ed.), 1951. Educational Measurement. American Council on Education, Washington, DC.

Linn, R.L. (Ed.), 2006. Educational Measurement, third ed. Macmillan, New York.

Loevinger, J., 1957. Objective tests as instruments of psychological theory. Psychological Reports 3, 635–694.

Lord, F.M., 1952. A theory of test scores. Psychometric Monographs 7.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Lucke, J.F., 2005. "Rassling the hog": the influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. Applied Psychological Measurement 29, 106–125.

McDonald, R.P., 1999. Test Theory: A Unified Treatment. Lawrence Elbaum, Malwah, NJ.

Millsap, R.E., Everson, H., 1991. Confirmatory measurement model comparisons using latent means. Multivariate Behavioral Research 26, 479–497.

Mosier, C.I., 1941. A short cut in the estimation of split-halves coefficients. Educational and Psychological Measurement 1, 407–408.

Novick, M.R., 1966. The axioms and principal results of classical test theory. Journal of Mathematical Psychology 3, 1–19.

Novick, M.R., Jackson, P.H., 1974. Statistical Methods or Educational and Psychological Research. McGraw Hill, New York.

Novick, M.R., Lewis, C., 1967. Coefficient alpha and the reliability of composite measurements. Psychometrika 32, 1–13.

Osterlind, S.J., 2010. Modern Measurement: Theory, Principles, and Applications of Mental Appraisal, second ed. Prentice-Hall/Merrill, Upper Saddle River, NJ.

Raykov, T., 2001. Bias of coefficient α for fixed congeneric measures with correlated errors. Applied Psychological Measurement 25, 69–76.

Raykov, T., Marcoulides, G.A., 2010. Introduction to Psychometric Theory. Taylor & Francis Group, New York.

Revelle, W., Zinbarg, R.E., 2009. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. Psychometrika 74, 145–154.

Rulon, P.J., 1939. Simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review 9, 99–103.

Sijtsma, K., 2009. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 94, 107–120.

Spearman, C., 1904. The proof and measurement of association between two things. American Journal of Psychology 15, 72–101.

Ten Berge, J.M.F., Socan, G., 2007. The set of feasible solutions for reliability and factor analysis. In: Lee, S.-Y. (Ed.), Handbook of Latent Variables and Related Models. North Holland, Amsterdam, pp. 303–320.

Ten Berge, J.M.F., Zegers, F.E., 1978. A series of lower bounds to the reliability of a test. Psychometrika 43, 575–579.

Thorndike, R.L. (Ed.), 1971. Educational Measurement, second ed. American Council on Education, Washington, DC.

Thurstone, L.L., 1932. The Reliability and Validity of Tests. Ann Arbor, MI.

Traub, R.E., 1997. Classical test theory in historical perspective. Educational Measurement: Issues and Practice 16, 8–14.

Psychometrics: Combinatorial Data Analysis

Michael Brusco, Florida State University, Tallahassee, FL, USA
Douglas Steinley, University of Missouri-Columbia, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by L.J. Hubert, P. Arabie, volume 4, pp. 2263–2267, © 2001, Elsevier Ltd.

Abstract

Combinatorial data analysis (CDA) consists of a collection of methods for confirmatory and exploratory data analyses. The common component of these methods is the use of combinatorial operations, such as permutations or combinations, to either generate reference distributions for a statistic (in the confirmatory case) or optimize some objective function (in the exploratory case). This article especially highlights recent advancements for exploratory CDA applications such as seriation, unidimensional scaling, multidimensional scaling, clustering, and variable selection.

Introduction

Combinatorial data analysis (CDA) is a term coined by Arabie and Hubert (1992) that spans several areas of scientific study. Hubert and Arabie (2001) organized their entry for CDA in the first edition of this encyclopedia around two broad subareas: (1) confirmatory combinatorial data analysis (CCDA) and (2) exploratory combinatorial data analysis (ECDA). The CCDA applications are characterized by 'combinatorial computation,' whereby combinatorial methods are used to generate a reference distribution to obtain a *p*-value for an observed statistic. Accordingly, most nonparametric statistical methods, along with most randomization and permutation tests, fall under the CCDA umbrella. The ECDA methods focus on using combinatorial optimization methods to uncover structure and include seriation, scaling, and clustering. Our coverage of CDA tips heavily in favor of ECDA, where considerable methodological advancements have emerged in the past 10–15 years. In addition to coverage of the ECDA methods discussed by Hubert and Arabie (2001), we broaden the scope of CDA to include another important class of problems, namely, variable selection in multivariate statistics.

The structure of our entry is organized into four major sections. Within each of these sections, there are two main subsections: (1) Models and Applications and (2) Methods. Section Confirmatory Combinatorial Data Analysis provides a coverage of CCDA. Section Seriation and Scaling reviews recent methods for problems in seriation and scaling. Section Clustering focuses on combinatorial problems in clustering, Section Variable Selection discusses the role of combinatorial optimization in the development of variable selection for multivariate statistical applications. The article concludes with a brief summary.

Confirmatory Combinatorial Data Analysis

The notion of generating reference distributions to measure the significance of an observed statistic dates back to the 1930s. More recent and extensive treatments of CCDA topics are provided by Edgington (1995), Hubert (1987), and Mielke and Berry (2001). Hubert and Arabie (2001) organized their coverage of CCDA along the topics of linear assignment and quadratic assignment. One of the inherent advantages of this organization is that linear and quadratic assignment together contain many interesting applications, such as rater agreement, regression, and ANOVA. We also use the linear assignment and quadratic assignment distinction, but also include a section on multinomial models that are widely used in the analysis of contingency tables.

Models and Applications

Linear assignment

The linear assignment model is, for all intents and purposes, a matching model. Given an $n \times n$ matrix, the observed statistic is typically an index measure computed as the sum of the main diagonal elements, and the reference distribution is obtained by recomputing the index for permutations of the columns of the matrix. Hubert and Arabie (2001) note that Pearson's product–moment correlation and Cohen's kappa are among the applications of linear assignment.

Quadratic assignment

The quadratic assignment model assumes two $n \times n$ matrices as input. The observed statistic is typically an index measure computed as the sum of elementwise products of the two matrices. The reference distribution is obtained by recomputing the index for permutations of the rows and, simultaneously, the columns of one of the two matrices while holding the other matrix fixed. Hubert and Arabie (2001) observe quadratic assignment applications such as one-way ANOVA and spatial autocorrelation. We note that quadratic assignment methods have been increasingly popularized in the past 10–15 years because of their role in hypothesis testing within the context of social network structures.

Multinomial (fixed-marginal) models

An exact permutation test for a variety of contingency table statistics can be developed using the principles of fixed-marginal tests (see Hubert, 1987 for a discussion). An 'observed' contingency table is constructed and established and the appropriate statistic is computed. The marginal frequencies of the table are the row and column sums. All contingency tables yielding the same fixed marginal frequencies as the observed table are generated, and their probabilities are computed by

the multinomial distribution. Summing the probabilities for tables that yield a statistic greater than or equal to the observed statistic produces an exact p-value.

Methods

Exact p*-values*

Complete enumeration of the entire reference distribution enables an exact p-value to be established for the observed statistic. This approach becomes impractical as the number of objects increases. For example, evaluation of all $n!$ permutations might be computationally feasible for $n = 10$, but probably not for $n = 15$. It might also be possible to explicitly evaluate all contingency tables with the same fixed marginals, but computational feasibility of this approach depends on the number of objects and the dimensions of the contingency table. Some authors have shown that partial enumeration strategies can often enable exact p-values to be established without explicit generation of the entire reference distribution.

Resampling

A more common approach to obtaining p-values is to randomly sample from the set of all possible configurations to obtain the reference distribution. Hubert (1987), for example, often uses a sample size of 10 000 to obtain the reference distribution, which is more than sufficient to perform a test at, say, $\alpha = .05$ or $\alpha = .01$. The key takeaway here is that even when computation of the exact p-value is possible, resampling might be perfectly sufficient for the analyst's hypothesis testing needs.

Distributional tests

Another approach for estimation of p-values is to use asymptotic distribution theory to estimate means and variances for the various indices. Although this approach can be effective when sample sizes are large, its utility is diminished for small sample sizes. Moreover, the derivation of the variance for some index measures is a nontrivial task.

Seriation and Scaling

There are a variety of problems in the social sciences that require the identification of a permutation or sequence of the objects to uncover underlying structure based on proximity information among pairs of objects. These are known as *seriation* problems. Unidimensional scaling is a closely related problem that incorporates the additional requirement of the specification of distances between pairs of objects in the permutation in an attempt to closely recover dissimilarity information among the objects. This process can also be generalized to multidimensional scaling in the city-block metric (Hubert et al., 2006).

Models and Applications

Paired comparison ranking

Paired comparison ranking is an important type of seriation problem that requires the sequencing of objects based on information in an asymmetric proximity matrix. The objective of the problem, which also has application to triangularization of input–output matrices in economics, is to find a reordering of the rows and columns that maximizes the sum of the matrix elements above the main diagonal in the reordered matrix.

Gradient-based seriation

In some applications, it is desirable to obtain a permutation of the rows and columns of a matrix such that the elements in the reordered matrix exhibit a desirable pattern. These patterns typically involve either nonincreasing or nondecreasing elements when moving away from the main diagonal within rows or columns. An excellent discussion of gradient-based seriation is provided by Hubert et al. (2001, Chapter 4), who provide a number of possible indices that can be used for this procedure.

Unidimensional scaling

The goal of unidimensional scaling is to place a set of n objects along a continuum such that distances between the pair of objects approximate, as closely as possible, the dissimilarities between pairs of objects in a set of observed proximity data. Typically, the quality of the fit is measured by a least squares loss function (Brusco and Stahl, 2005; Hubert et al., 2001; Chapter 4); however, a least absolute deviation loss function can also be used.

Multidimensional city-block scaling

The principles of unidimensional scaling can be generalized to multidimensional scaling in the city-block metric (Hubert et al., 2006). The only difference is that permutations of the objects are established for two or more dimensions, and the computed distances between pairs of objects are the sum of the distances across multiple dimensions.

Methods

In our classification of methods for problems of seriation and scaling, we distinguish between *exact* methods and *approximation* (or heuristic) methods. Exact methods are guaranteed to provide a globally optimal solution for the underlying discrete optimization problem. For small problems, this can be accomplished using exhaustive enumeration of all possible solutions, whereas partial enumeration strategies such as dynamic programming and branch-and-bound are often feasible for large problems. Approximation procedures do not afford the guarantee of globally optimal solutions, yet they often provide globally optimal solutions for small problems and remain scalable for larger ones.

Exact methods

Dynamic approaches for seriation and unidimensional scaling are based on pioneering work in electrical engineering. These approaches have been adapted for seriation of asymmetric matrices within the context of analysis of paired comparison data, as well as for unidimensional scaling (see Hubert et al., 2001, Chapter 4). Dynamic programming is a fast and reliable alternative for seriation of problems with up to about 30 objects. One of its most desirable properties is its lack of sensitivity to the degree of 'structuredness' of the input proximity data. Nevertheless, beyond 30 objects, the

computer storage required for dynamic programming makes it computationally infeasible for most hardware platforms.

Branch-and-bound algorithms have been developed for seriation of asymmetric proximity data within the context of paired comparison ranking, as well as unidimensional scaling. A review of these algorithms, along with other applications, is provided by Brusco and Stahl (2005, Chapters 7–11). Linear and integer programming approaches for seriation of asymmetric matrices have proved successful for many years. In fact, quite sizable problems (100 or more objects) can often be solved exactly using current commercial software programs. The computational evidence for using integer programming to solve gradient-based seriation problems and least-absolute deviation of unidimensional scaling problems is less encouraging.

Approximation methods

Although exact methods for seriation, unidimensional scaling, and multidimensional scaling are feasible for some problems of practical size, heuristics are required for larger problem instance. De Soete et al. (1988) proposed a general heuristic solution strategy for seriation and unidimensional scaling based on iterative quadratic assignment. This approach has also been adapted for multidimensional city-block scaling. The computational evidence suggests that iterative quadratic assignment works well for modestly sized problems, particularly those where there is some inherent structure in the data.

Simulated annealing has been one of the most popular heuristic approaches for seriation, unidimensional scaling, and multidimensional scaling for the past dozen years. A review of recent simulated annealing approaches for unidimensional scaling, paired-comparison ranking seriation, and other application is provided by Brusco et al. (2008), who also integrated simulated annealing in conjunction with a heuristic implementation of dynamic programming to obtain solutions for large-scale seriation of asymmetric matrices, gradient-based seriation, and unidimensional scaling.

Clustering

The literature on clustering methods is vast and, therefore, it is impossible to provide a comprehensive review within the confines of this treatment of CDA. We will not attempt to cover the literature on mixture-model clustering, which tends to use smooth optimization procedures rather than combinatorial ones. Moreover, we will only address several areas of combinatorial clustering literature that we believe are apt to be of continued or emerging importance in the social sciences literature. In the following Section Models and Applications, we focus on four classes of models and applications of combinatorial clustering in the social sciences: (1) distance-based clustering of metric data, which is most common in the literature; (2) clustering of nonmetric data; (3) clustering of social network data; and (4) multiobjective clustering.

Models and Applications

Distance-based clustering of metric data

In many applications, researchers have measurements for N objects (customers, students, patients, etc.) on each of P metric variables. The most common clustering approach for such data is K-means clustering (see Steinley, 2006, for a review), which seeks to find a partition of the objects that minimizes the sum of the squared Euclidean distances of the objects to their cluster centroids. In this context, the cluster centers are virtual because they are variable means that do not necessarily correspond to any given object. A closely related option is to cluster the cases to minimize the sum of the distances of each object from an 'exemplar' object in its cluster (Kaufman and Rousseeuw, 2005). When using the exemplar clustering model, the cluster centers are actual objects. Moreover, exemplar-based clustering methods tend to be more general in nature and can be used for a broad range of similarity and dissimilarity data.

Clustering nonmetric data

Combinatorial methods for clustering nonmetric data also have an extensive history, with their roots dating back to literature on majority rule and social choice. Clique partitioning is a popular approach that is based on the establishment of binary equivalence relations among the categorical measures. A partition is sought that minimizes deviation in a least squares sense. The formation of dissimilarity data based on any number of relations also enables the implementation of minimum-diameter partitioning, which seeks to establish compact partitions by minimizing the maximum dissimilarity between any two objects in the seam cluster. This approach is analogous to the nonhierarchical clustering analog of complete-line hierarchical clustering and is closely related to the graph-coloring problem.

Blockmodeling of social network data

The substantial interest in social network data during the past 10 years creates the need for a special treatment of this topic. Although network data can be analyzed by establishing dissimilarity measures among the nodes of the network and using standard hierarchical and nonhierarchical methods, Doreian et al. (2005) have proposed a more elegant approach known as 'generalized blockmodeling.' For one-mode networks, blockmodeling can be as simple as establishing a bipartition of the nodes of the network, such that one cluster is a highly interconnected core and the other is a weakly interconnected periphery. However, blockmodeling can also accommodate much more formal multiple cluster approaches, as well as different relations based on regular, structural, and other types of equivalence. The blockmodeling of signed networks and two-mode networks has also received extensive attention.

Methods

Exact methods

Considerable progress has been made in the development of exact solution procedures for K-means clustering problem. Although dynamic programming methods (Hubert et al., 2001; Chapter 3) are limited to problems with roughly 30 objects, branch-and-bound methods have successfully solved K-means clustering problems with up to 150 objects (Brusco and Stahl, 2005; Chapter 5). New mathematical programming formulation and solution techniques for

clustering problems have provided even more impressive results for both *K*-means (Aloise et al., 2012), solving problems with thousands of objects. When used in conjunction with heuristics, Lagrangian relaxation, and/or column generation, branch-and-bound algorithms can also successfully solve large instances (several hundred objects or more) of the *K*-median problem. Mathematical programming approaches can also be gainfully employed for other clustering problems, such as clique partitioning.

Approximation methods

There is an assortment of heuristic procedures that are interchangeably referred to as *K*-means clustering methods (Steinley, 2006). These procedures typically iterate between phases of cluster centroid computation and object reassignment until convergence. These heuristics can be gainfully augmented with exchange algorithms. In fact, there is evidence to suggest that using thousands of restarts of the *K*-means heuristic from different seed points can yield performance with much more sophisticated metaheuristics. Similar types of exchange algorithms are also effective for *K*-medoids clustering.

Metaheuristics for clustering problems have abounded during the past couple of decades. For *K*-means clustering, metaheuristics include, but are not limited to, simulated annealing, tabu search, genetic algorithms, and variable neighborhood search. Brusco and Steinley (2007) provide a review and comparison of these heuristics. The same type of heuristics have also been gainfully employed for *K*-medoids clustering and clique partitioning.

Variable Selection

In many multivariate statistical applications, researchers are faced with a large set of potential variables. Moreover, it has long been recognized that, in many applications, at least some of the potential variables will not contribute to the data analysis and might even be detrimental. The problem of variable selection is an inherently combinatorial one because of the exponential growth in the number of subsets as a function of the number of candidate variables. When there are P candidate variables, there are 2^P possible subsets of variables.

Models and Applications

Regression

The selection of an appropriate subset of variables in regression is important regardless of whether the principal focus of the application is description or prediction. In studies where the emphasis is on describing which variables contribute most to explanation of the dependent variable, the presence of multicollinearity can make it difficult to disentangle relative predictor importance. In the case of predictive applications, noisy variables can lead to overfitting and poor prediction of new cases. Although most of the variable selection literature pertains to standard ordinary least squares (OLSs) regression, adaptations have also been designed for logistic regression and polynomial regression.

Discriminant analysis

As in the case of regression, there is often a distinction made between descriptive and predictive applications. Although most of the variable selection literature in discriminant analysis centers on the predictive side, the measurement of predictor contribution may also be relevant in some applications.

Factor analysis and principal component analysis

There is a tendency to conclude that variable selection is not particularly relevant for factor analysis or principal component analysis, which are dimension reduction approaches in their own right. However, it is important to recognize that the number of factors extracted, the loadings, and the factor scores are all contingent on the variables entered into the model. Recent research has shown that considerable improvement in interpretability can often be realized by applying variable selection methods prior to running a principal component or factor analysis.

Cluster analysis

It is long been recognized that even a few irrelevant variables can substantially distort the recovery of structure in a cluster analysis. Not surprisingly, this has led to a substantial effort devoted to the development of variable selection methods for clustering.

Methods

Exact methods

Exact solution procedures for variable selection fall into two major categories: (1) all-possible-subsets methods and (2) branch-and-bound methods. All-possible-subsets methods are precisely what the name implies. Solutions for all 2^P variable subsets are analyzed and the best subset (or, possibly, the top 5 or 10 subsets) based on the appropriate objective criterion is stored further analysis. Although this option was once computationally daunting, it is now often feasible for multivariate statistical applications where P is in the low 20s or less. Moreover, in some applications, such as regression, there are computational aspects of the problem that can be exploited for additional efficiency.

In some applications, there is a monotonicity property for the objective function that enables branch-and-bound strategies to be gainfully employed. The fact that the R^2 measure in regression cannot decrease when a predictor is added is a simple example. One of the most important contributions in this respect was Furnival and Wilson's (1974) branch-and-bound method for multiple regression. This algorithm has been the underpinning for more recent adaptations for regression, as well as extensions for polynomial regression, discriminant analysis, and principal component analysis.

Approximation Methods

For applications where the monotonicity property of the objective function does not hold, or P is too large to enable branch-and-bound methods, approximation procedures are necessary. Among the earliest heuristic procedures are forwards selection, backward elimination, and stepwise methods. Forward selection begins with an empty set of selected variables, and variables are added one at a time until a stopping

criterion is reached. Backward elimination begins with all variables selected and eliminates variables one at a time until a stopping criterion is reached. Stepwise methods typically combine elements of forward selection and backward elimination in some fashion. Although once the workhorses of variable selection (and still present in some commercial software packages), forward selection, backward elimination, and stepwise methods are now generally considered outmoded approaches.

Transfer and exchange algorithms for variable selection typically evaluate three types of neighborhood search operations: (1) adding, i.e., moving a variable from the unselected to selected subset of variables; (2) dropping, i.e., moving a variable from the selected to unselected subset of variables; and (3) interchanging, i.e., swapping a variable in the selected subset with one from the unselected subset. Although effective in their own right, these algorithms often form the engine of more sophistical search procedures such as simulated annealing, tabu search, genetic algorithms, and variable neighborhood search.

Metaheuristic approaches have been gainfully employed for variable selection in several areas of multivariate statistics. For example, the effectiveness of genetic algorithms, tabu search, and variable neighborhood search have been evaluated within the context of discriminant analysis. Simulated annealing, tabu search, and variable neighborhood search have been used for regression applications. A number of metaheuristic approaches have also been evaluated for principal component analysis.

Summary

In our preparation of this article for CDA, we have sacrificed a considerable amount of depth for any given topic area in order to achieve greater breadth in the topic areas covered. In closing, we offer one final treatment of an emergent method in ECDA, namely, multiobjective CDA, which has become increasingly common in the past 10–15 years. Multiobjective CDA applications abound in seriation and scaling, as well as clustering (see, for example, Brusco and Stahl, 2005, Chapters 6 and 11). Moreover, multiobjective programming can sometimes offer a linchpin between seriation and clustering. It seems likely that continued advancements in computational hardware and software, along with the need to obtain solutions that accomplish multiple objectives, will result in multiobjective methods playing a greater role in future applications.

See also: Best Worst Scaling: Theory and Practice; Clustering and Cohesion in Networks: Concepts and Measures; Psychometrics; Scaling and Classification in Social Measurement; Self-Organizing Dynamical Systems.

Bibliography

Aloise, Daniel, Hansen, Pierre, Liberti, Leo, 2012. An improved column generation algorithm for minimum sum-of-squares clustering. Mathematical Programming A 131, 195–220.

Arabie, Phipps, Hubert, Lawrence, 1992. Combinatorial data analysis. Annual Review of Psychology 43, 169–203.

Brusco, Michael J., Köhn, Hans-Friedrich, Stahl, Stephanie, 2008. Heuristic implementation of dynamic programming for matrix permutation problems in combinatorial data analysis. Psychometrika 73, 503–522.

Brusco, Michael J., Stahl, Stephanie, 2005. Branch-and-Bound Applications in Combinatorial Data Analysis. Springer, New York.

Brusco, Michael J., Steinley, Douglas, 2007. A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. Psychometrika 72, 583–600.

De Soete, Geert, Hubert, Lawrence, Arabie, Phipps, 1988. The comparative performance of simulated annealing on two problems of combinatorial data analysis. In: Diday, E. (Ed.), Data Analysis and Informatics. Amsterdam, North-Holland, pp. 489–496.

Doreian, Patrick, Batagelj, Vladimir, Ferligoj, Añuska, 2005. Generalized Blockmodeling. Cambridge University Press, Cambridge, UK.

Edgington, E.S., 1995. Randomization Tests, third ed. Marcel Dekker, New York.

Furnival, G.M., Wilson, R.W., 1974. Regression by leaps and bounds. Technometrics 16, 499–512.

Hubert, Lawrence J., 1987. Assignment Methods in Combinatorial Data Analysis. Marcel Dekker, New York.

Hubert, Lawrence, Arabie, Phipps, 2001. Combinatorial data analysis. In: International Encyclopedia of Social & Behavioral Sciences, pp. 2263–2267.

Hubert, Lawrence, Arabie, Phipps, Meulman, Jacqueline, 2001. Combinatorial Data Analysis: Optimization by Dynamic Programming. Society for Industrial and Applied Mathematics, Philadelphia.

Hubert, Lawrence, Arabie, Phipps, Meulman, Jacqueline, 2006. The Structural Representation of Proximity Matrices with MATLAB. Society for Industrial and Applied Mathematics, Philadelphia.

Kaufman, Leonard, Rousseeuw, Peter J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis, second ed. Wiley, New York.

Mielke, Paul W., Berry, Kenneth J., 2001. Permutation Methods: A Distance Function Approach. Springer-Verlag, New York.

Steinley, Douglas, 2006. *K*-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology 59, 1–34.

Psychometrics: Item Response Theory

David Thissen, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Item response theory (IRT) comprises a set of generalized latent variable models designed specifically for data arising from educational and psychological tests and questionnaires. IRT is used for item analysis and scale development and provides the statistical underpinnings for computerized adaptive testing. There are IRT models for items with either dichotomous or polytomous response formats on scales that measure one or more dimensions of individual differences.

In contemporary statistical parlance, item response theory (IRT) comprises a set of generalized latent variable models (Skrondal and Rabe-Hesketh, 2004) designed specifically for data arising from educational and psychological tests and questionnaires (Bock and Moustaki, 2007). While there are many meanings of the term *latent variable* (see Skrondal and Rabe-Hesketh, 2004: Chapter 1), in the context of IRT *latent variable* refers primarily to a hypothetical construct that explains the observed covariation among the item responses, and secondarily to *latent responses* that are continuous variables underlying observed categorical responses. The models are *generalized* in the sense that they use nonlinear link functions to relate (usually linear) functions of the underlying continuous latent variables to the observed responses. Closely related models with discrete latent variables are referred to as *latent class models*.

Current uses of IRT include item analysis, scale development, detecting group differences in item responses, estimating scores for computerized adaptive testing (CAT), accounting for violations of local dependence with the use of testlets, as well as developing an understanding of the psychological response process underlying people's answers to academic, social, and personality questions.

Intellectual History

Latent variable models have only relatively recently become respectable subjects of study in the field of statistics; however they have a long history of development in psychology and the social sciences. Indeed, the use of the terms *manifest data* for observed responses, and *latent continuum* for the underlying unobserved variable, may have originated with Lazarsfeld (1950: p. 367), who wrote that "We shall now call a *pure test* of a continuum x an aggregate of items which has the following properties: *All interrelationships between the items should be accounted for by the way in which each item alone is related to the latent continuum*." (Emphasis in the original.) We now say that a 'pure test,' in Lazarsfeld's language, is a test in which the item responses fit well with a model with the properties of *unidimensionality* and *local independence*.

The intellectual antecedents of IRT within quantitative psychology predate even the distinction between latent variables and manifest data. The *normal ogive* model for dichotomous data (Lord, 1952), based on theory summarized by Lord and Novick (1968: pp. 370–371), has its roots in Thurstone's (1925, 1927) ideas about scaling tests and psychological stimuli. The intellectual history of IRT models has been summarized by Bock (1997) and Thissen and Steinberg (2009). Embretson and Reise (2000) and DeMars (2010) have offered introductory texts on IRT.

Models for Dichotomous Responses

A prototypical IRT model is the two-parameter logistic (2PL) model for dichotomous responses. Common notation for a logistic *trace line* that 'traces' (Lazarsfeld, 1950: p. 364) the probability of endorsement of item i is

$$T(u_i = 1) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]}, \qquad [1]$$

in which u_i is the item response (1 for correct or endorsement, 0 otherwise), θ is the latent variable being measured, and a_i and b_i are two item parameters – hence, the 2PL. The a parameter is referred to as the *discrimination* or *slope* parameter. The slope parameter is usually positive; its magnitude indicates the strength of the relation between the item response and θ. Equation [1] is written in *slope-threshold* form, in which the threshold parameter b_i is the value of the latent variable θ at which a person has a 50-50 chance of endorsing the item.

The upper panel of **Figure 1** shows trace lines for four dichotomous items in a format that is modeled loosely after Guilford's (1936: p. 427) **Figure 1**, in the sense that it uses the axis labels 'scale of ability' and 'probability of success.' The modern notation θ is added to the abscissa label, which is marked in standard units. Three of the curves are 2PL trace lines: The curve marked A is for a relatively easy ($b = -1.75$), highly discriminating ($a = 3$) item; that marked B is for a slightly easier ($b = -1.25$), less discriminating ($a = 1$) item; and the one marked D is for a more difficult ($b = 1$), similarly less discriminating ($a = 1$) item. The curve marked C is an example of a three-parameter logistic (3PL) trace line. The 3PL model adds a parameter that represents the probability that individuals at low levels of θ may respond correctly or positively, as they do by guessing on multiple choice items. For the item marked C in **Figure 1**, the guessing level $g = 0.2$, the asymptotic level of the curve on the left; setting aside guessing, the item is of moderate difficulty ($b = 0$), and it is relatively discriminating ($a = 2$).

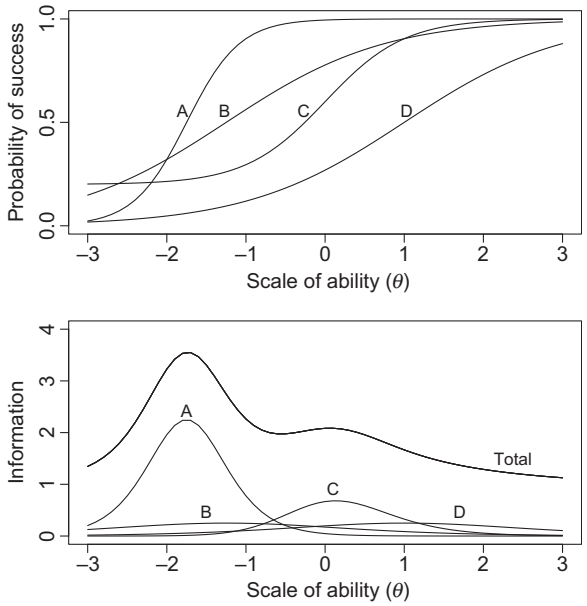

Figure 1 Upper panel: Trace lines for the positive response for four dichotomous items: A, B, and D are two-parameter logistic trace lines, and C is a three-parameter logistic trace line. Lower panel: The information functions for the four items with trace lines in the upper panel, and the total information for the four items.

From a statistical point of view, the crucial aspect of IRT models is that the unobserved variable accounts for the observed interrelationships among the item responses. That means that, given (or *conditional on*) any particular value of the latent variable θ, the item responses are independent; this is usually called the assumption of *conditional independence* or *local independence*. Then the probability of a vector of item responses $\mathbf{u} = [u_1, u_2, \ldots u_I]$ from a particular person for a collection of items indexed by i is

$$P(\mathbf{u}) = \int \left[\prod_i T(u_i) \right] \phi(\theta) \mathrm{d}\theta. \qquad [2]$$

The term $\phi(\theta)$ represents the population distribution of θ, and is usually assumed to be Gaussian. Maximum likelihood (ML) estimation of the item parameters involves choosing their values for all items such that the joint probability (likelihood) of all of the observed item responses from all persons is maximized. The item parameters are embedded in $T(u_i)$ in eqn [2].

IRT provides a detailed quantitative description of each item in the form of the item parameters; these descriptions are useful in test and questionnaire assembly, and in test scoring (Thissen and Wainer, 2001). A very useful derivative function of the item parameters is the item information curve (Birnbaum, 1968); the lower panel of **Figure 1** shows the information functions for the four items with trace lines in the upper panel. The ordinate for the curves in the lower panel is the amount of information (in the likelihood sense) provided by each item at each level of θ. For the 2PL or 3PL model, each information curve has a peak at the value of b; their heights are proportional to the squared value of the a parameter. The information provided by a test is the sum of the information curves for its items, so information curves can be used in test assembly to construct tests with a desired level of information, or measurement precision, at any value, or for any range, of the latent variable.

Models for Polytomous Responses

Samejima's model (1969) for graded item responses is often applied to the analysis of items that are accompanied by Likert-type response scales. The model for an item with K ordered response alternatives $k = 0, 1, 2, \ldots K - 1$ is

$$T(u_i = k) = T^*(u_i = k) - T^*(u_i = k + 1), \qquad [3]$$

in which $T^*(u_i = 0) = 1$ and $T^*(u_i = K) = 0$. $T^*(u = k)$ is the trace line describing the probability that a response is in category k or higher, for each value of the underlying construct. The probability that a response is in a particular category (k) is the probability of observing category k or higher minus the probability that the response is in a higher category. The logistic version of the model is

$$T(u_i = k) = \frac{1}{1 + \exp[-a_i(\theta - b_{ik})]} - \frac{1}{1 + \exp[-a_i(\theta - b_{ik+1})]}. \qquad [4]$$

The slope or discrimination parameter (a) has the same meaning as it does for the 2PL model, and the b_k parameter is the threshold for a category or higher. The value of b_k is the point on the construct axis at which the probability that the response is in category k or higher passes 50%. **Figure 2** shows trace lines from Samejima's (1969) graded model for five Likert-scale responses to an attitude item; for this hypothetical item, $a = 2$ and the four b values are -1.5, -0.5, 0.0, and 1.75.

The statistical model for a collection of polytomous responses remains the same as for dichotomous items; the polytomous trace lines represented by eqn [4] are substituted in eqn [2] to yield a likelihood function that may be used in parameter estimation. Indeed, dichotomous and polytomous items may be mixed in a single test, and analyzed and scored together. There are other models for graded responses, and even for nominal responses with no order specified a priori; chapters summarizing many such models are in volumes on

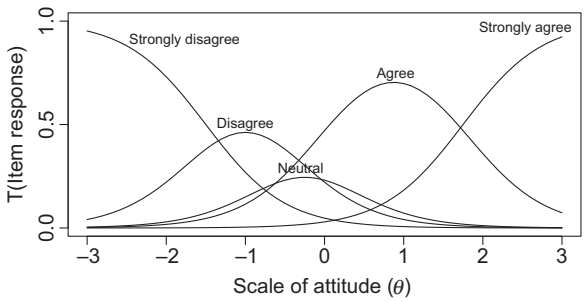

Figure 2 Trace lines from Samejima's (1969) graded model for five Likert-scale responses to an attitude item.

the subject edited by van der Linden and Hambleton (1997) and Nering and Ostini (2010).

Compensatory Multidimensional Item Response Theory

Until relatively recently, applications of IRT were limited to measurement of a single construct, using models in which θ is unidimensional. This was largely a computational limitation; the conceptual generalization of IRT to *multidimensional item response theory* (MIRT) has been the subject of active research for several decades. Reckase (2009) has provided a book-length summary of research on the topic. Advances in computing and algorithms for statistical parameter estimation have made MIRT applications practical.

IRT becomes MIRT when more than one latent variable is required to account for the observed pattern of covariation among item responses. It is conventional to refer to the collection of several latent variables as $\boldsymbol{\theta}$; the bold typeface indicates that $\boldsymbol{\theta}$ represents a vector with as many values as there are latent dimensions. The most commonly used MIRT models are *compensatory*; they assume that the probability of an item response depends on a linear combination of the component values of $\boldsymbol{\theta}$. It is convenient to represent that linear combination using vector multiplication, in which $\mathbf{a}'\boldsymbol{\theta} = a_1\theta_1 + a_2\theta_2 + \cdots + a_p\theta_p$ for a p-dimensional latent variable. Using that notation, the trace surface for the multidimensional generalization of the 2PL model is

$$T(u_i = 1) = \frac{1}{1 + \exp[-(\mathbf{a}_i'\boldsymbol{\theta} + c_i)]}. \qquad [5]$$

Here T is the surface over the $\boldsymbol{\theta}$ (hyper)plane that traces the probability of a correct or positive response ($u_i = 1$) to item i. For two-dimensional models, $\boldsymbol{\theta}$ is a two-dimensional plane with values of θ_1 from low to high in one direction, and values of θ_2 from low to high in the other direction; T is a surface that rises from zero (for low values of θ_1 and θ_2) to one (for high values of θ_1 or θ_2 or both). Models with more than two latent dimensions are difficult to visualize.

Comparing eqn [5] with eqn [1], the only difference is that the notation $\mathbf{a}'$ represents a vector of as many slope or discrimination parameters as there are dimensions, and $\boldsymbol{\theta}$ is a vector of latent variable values of the same dimensionality. This MIRT model adds one more parameter (one more a value) for each additional dimension. The a parameters each measure the degree of relationship of the item response with the corresponding dimension of $\boldsymbol{\theta}$ (that is, with each construct). The scalar-valued parameter c is an intercept parameter that reflects the overall probability of endorsement of the item. In contrast to the unidimensional 2PL model, an alternative parameterization with a threshold parameter is more difficult to interpret, because for MIRT models the location of $T = 0.5$ is a line for a two-dimensional model, or a (hyper)plane for higher dimensional models – there is no scalar-valued threshold.

Compensatory MIRT models are factor analysis models for categorical item response data, or *item factor analysis* models (Bock et al., 1988; Bartholomew and Knott, 1999; Bock and Moustaki, 2007; Wirth and Edwards, 2007). The differences between MIRT and factor analysis involve the use of nonlinear models for the probability of a categorical item response in MIRT vs linear models for continuous scores in factor analysis, and the fact that the regression parameters measuring the association between the observed variables and the latent variables are estimated as a parameters in MIRT and *factor loadings*, λ, in factor analysis. The former is unavoidable given the nature of the observed data; the latter is a matter of convention.

Noncompensatory Multidimensional Models

In the compensatory models described in the preceding section, the combination rule for a positive response is that the respondent may be high on one *or* another of the latent variables. An alternative model formulation is based on the idea that an item response may depend upon two or more latent variables, and both (or all) must be high to observe a positive response. This formulation yields models in which probabilities of success on the component tasks are multiplied to obtain the probability of overall success; such multiplicative models have most often been called *noncompensatory* models, although Reckase (2009: p. 79) uses the term *partially compensatory*.

As an example of such models for cognitive tests, Embretson (1997) summarizes research on the *General Component Latent Trait Model*, which was designed for cognitive items with *components* or *subtasks*, all of which must be completed for a correct response. An example from the social measurement domain is the *two decision model* (Thissen-Roe and Thissen, 2013) for Likert-type responses in contexts in which individual differences in a tendency to select extreme alternatives may be one of the factors underlying Likert-type scale responses.

Noncompensatory (or partially compensatory) models, like the two mentioned above, have generally been developed for specific item types and testing contexts. While such specialized models may often be useful in their context, they are not intended for general purpose item analysis.

Parameter Estimation

Bock and Aitkin (1981) provided an elegant and practical procedure for item parameter estimation; their algorithm is currently the most commonly implemented in computer software for item calibration, because it is efficient for large numbers of persons and items. Skrondal and Rabe-Hesketh (2004) described a general framework integrating IRT with multilevel models, structural equation models, and longitudinal models, and procedures that may be used to compute ML estimates for complex models that are combinations of all of those. This more general conception is theoretically elegant but computationally intensive, so it is limited to problems with relatively small numbers of variables. Skrondal and Rabe-Hesketh's (2004): Chapter 6 includes a detailed review of virtually all extant ML estimation procedures from a statistical perspective. Bock and Moustaki (2007) provide a similarly general framework for IRT and MIRT models with covariates, and the use of the Bock–Aitkin

EM (Expectation-Maximization) algorithm to estimate the parameters of more elaborate models.

An increasingly widely used alternative to ML estimation is Bayesian estimation using Markov chain Monte Carlo (MCMC) algorithms. Albert (1992) provided the first MCMC algorithm for estimation of the parameters of the normal ogive model, using *data augmentation*, which is one common approach to MCMC estimation for latent variable problems. Patz and Junker (1999a,b) described MCMC algorithms based on *Metropolis–Hastings within Gibbs* sampling, the other common approach to complex MCMC estimation, for the 2PL, 3PL, and other models. An advantage of Bayesian MCMC estimation is the relative ease with which the parameters of more complex models may be estimated. This advantage has been used in recent developments of *testlet response theory* (Wainer et al., 2007) to provide estimates for item response models with an additional parameter that absorbs and corrects for local dependence that may appear on tests that include clustered items.

Scoring Tests Using IRT

In the context of IRT, computing scores is taken to be a statistical estimation problem, in which the goal is to provide a point estimate of an individual respondent's location on the θ dimension(s), with an associated statement of (im)precision, or standard error. A commonly used way to do that is to compute a *scale score* as the expected *a posteriori* estimate of θ, or $\mathrm{EAP}[\theta|\mathbf{u}]$, which is notation for the average of the posterior density for the response pattern $\mathbf{u}$. The posterior density is the integrand of eqn [2], in which the trace lines $T(u_i|\theta)$ may be those of any item response model or a mixture of those. The standard deviations ($\mathrm{SD}[\theta|\mathbf{u}]$) are often reported as the standard error of the scale score.

The process of computing the scale score is independent of the particular items presented to the respondent, because it is a function of the product of trace lines associated with the item responses. If one has item parameters for a large collection of items assembled into an *item bank*, in which all of the items are calibrated with respect to the same latent variable θ, subsets of those items may be administered and scale scores computed that are comparable even for persons who responded to mutually exclusive sets of items. Scoring IRT linked alternate test forms and CAT, which can be viewed as the administration of a large number of custom-made alternate test forms, make use of these properties of scale scores.

Chapters 3 and 4 of the volume by Thissen and Wainer (2001) describe in detail the computations underlying EAP $[\theta|\mathbf{u}]$ and other, alternative, estimators; they also describe an algorithm for computing the expected *a posteriori* estimate given the summed score x ($\mathrm{EAP}[\theta|x]$). IRT scale scores for summed scores are sometimes used in testing programs that want to use the properties of IRT and at the same time base score reporting on summed scores, not response patterns.

See also: Bayesian Statistics; Binary Response Models and Logistic Regression; Computation: Expectation-Maximization Algorithm; Factor Analysis and Latent Structure Analysis: Overview; Generalized Linear Mixed Models; Log-Linear and Marginal Models; Psychometrics: Classical Test Theory.

Bibliography

Albert, J.H., 1992. Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational Statistics 17, 251–269.
Bartholomew, D.J., Knott, M., 1999. Latent variable models and factor analysis. In: Kendall's Library of Statistics, second ed., vol. 7. Arnold, London.
Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 392–479.
Bock, R.D., 1997. A brief history of item response theory. Educational Measurement: Issues and Practice 16, 21–33.
Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. Psychometrika 46, 443–459.
Bock, R.D., Gibbons, R., Muraki, E., 1988. Full-information item factor analysis. Applied Psychological Measurement 12, 261–280.
Bock, R.D., Moustaki, I., 2007. Item response theory in a general framework. In: Rao, C.R., Sinharay, S. (Eds.), Handbook of Statistics, Psychometrics, vol. 26. North Holland, Amsterdam, pp. 469–513.
DeMars, C., 2010. Item Response Theory. Oxford University Press, New York, NY.
Embretson, S.E., 1997. Multicomponent latent trait models. In: van der Linden, W., Hambleton, R. (Eds.), Handbook of Modern Item Response Theory. Springer-Verlag, New York, pp. 305–322.
Embretson, S.E., Reise, S.P., 2000. Item Response Theory for Psychologists. Erlbaum, Mahwah, NJ.
Guilford, J.P., 1936. Psychometric Methods. McGraw-Hill, New York, NY.
Lazarsfeld, P.F., 1950. The logical and mathematical foundation of latent structure analysis. In: Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A., Clausen, J.A. (Eds.), Measurement and Prediction. Wiley, New York, pp. 362–412.
van der Linden, W.J., Hambleton, R.K. (Eds.), 1997. Handbook of Modern Item Response Theory. Springer-Verlag, New York, NY.
Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.
Lord, F.M., 1952. A theory of test scores. Psychometric Monographs 7.
Nering, M.L., Ostini, R. (Eds.), 2010. Handbook of Polytomous Item Response Theory Models. Routledge, New York.
Patz, R.J., Junker, B.W., 1999a. A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics 24, 146–178.
Patz, R.J., Junker, B.W., 1999b. Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics 24, 342–366.
Reckase, M.D., 2009. Multidimentional Item Response Theory. Springer, New York, NY.
Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph 17, 34 (Pt 2).
Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.
Thissen, D., Steinberg, L., 2009. Item response theory. In: Millsap, R., Maydeu-Olivares, A. (Eds.), The Sage Handbook of Quantitative Methods in Psychology. Sage Publications, London, pp. 148–177.
Thissen, D., Wainer, H. (Eds.), 2001. Test Scoring. Lawrence Erlbaum Associates, Mahwah, NJ.
Thissen-Roe, A., Thissen, D., 2013. A two-decision model for responses to Likert-type items. Journal of Educational and Behavioral Statistics 38, 522–547.
Thurstone, L.L., 1925. A method of scaling psychological and educational tests. Journal of Educational Psychology 16, 433–449.
Thurstone, L.L., 1927. The law of comparative judgment. Psychological Review 34, 278–286.
Wainer, H., Bradlow, E.T., Wang, X., 2007. Testlet Response Theory and Its Applications. Cambridge University Press, New York.
Wirth, R.J., Edwards, M.C., 2007. Item factor analysis: current approaches and future directions. Psychological Methods 12, 58–79.

Relevant Websites

http://echo.edres.org:8080/irt/ – The Basics of Item Response Theory by Frank Baker.
http://www.metheval.uni-jena.de/irt/VisualIRT.pdf – A visual guide to item response theory.

Psychometrics: Preference Models with Latent Variables

Ulf Böckenholt, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Current psychometric models of choice behavior are strongly influenced by Thurstone's (1927) experimental and statistical work on measuring and scaling preferences. Aided by advances in computational techniques, preference models can now accommodate a wide range of data types and sources of preference variability among respondents induced by diverse factors such as person-specific choice sets or different functional forms for the underlying utility representations. At the same time, these models are increasingly challenged by behavioral work demonstrating the prevalence of choice behavior that is not consistent with the underlying assumptions of these models.

Introduction

Much of human behavior may be viewed as decisions among options or alternatives. Familiar instances are a citizen deciding whether to vote and if so for whom, an investor considering various stock options, or a shopper contemplating different brands in a product category. In addition to these observational settings, choice data may also be collected in experimental studies asking persons to state their preferences for a single option or multiple sets of options. For example, in a market research study, consumers may be shown a set of products with the instruction to choose from them or to rank them.

Choices collected in observational and experimental settings, which are referred to as revealed and stated preference data (Samuelson, 1938), may complement each other. For instance, in an election, votes for political candidates represent revealed choice data. Rankings of the same candidates in a survey shortly before the election are an example of stated choice data. Stated preference data may prove useful both for predicting the election outcome and for providing more information about the preference differences among the candidates than would be available from the election results alone. The collection of stated preference data is also useful in situations when it is difficult or impossible to collect revealed preference data because choices are made infrequently or new choice options offered in the studies are yet unavailable on the market. For example, in taste-testing studies, consumers may be asked to choose among existing as well as newly developed products.

A major challenge in analyzing choice data is that the observed choice outcomes cannot always be interpreted in an unambiguous way. Decisions may be determined not only by the attractiveness of the options, but also by seemingly irrelevant features of the choice environment. For example, contextual features – such as the order and frames in which information about the choice options is presented – have been shown to influence choice outcomes in systematic ways. In addition, characteristics of the decision maker – such as constraints in attention, knowledge, or time and money budgets – can play a significant role in the way information about the choice option is processed. It is not always possible to separate the influence of the different factors related to the choice options, choice situation, and decision maker on the underlying decision processes based on the observed choices alone (Manski, 2007). However, progress can be made in identifying the factors' separate effects by contrasting revealed and stated preference data.

In psychometrics, most preference models for the analysis of choice data have been developed for stated response settings to serve the dual purpose of summarizing choice outcomes and facilitating the forecasting of choices made by decision makers facing possibly new or different variants of the choice options. These models assume both that certain, possibly unobserved, characteristics of the choice options determine which option is selected, and that decision makers differ in their evaluation of these option characteristics. For example, in an election, voters may base their decision on a set of attributes such as integrity and leadership, but differ in the weights they assign to the attribute values for the different candidates. The application of preference models to voting data may reveal these attributes and provide insights about how voters vary in their attribute-specific assessments of the political candidates. This information can also be useful in forecasting how voters may behave if the attributes of the political candidates are changed in some systematic ways.

The strength of the current arsenal of preference models lies in providing parsimonious descriptions of how decision makers perceive and evaluate choice options. This approach is particularly effective when the different sources of individual variability are understood and readily identifiable. The next section reviews these sources of variability and their analysis with regard to preference models. This review emphasizes Thurstone's (1927) choice paradigm, which has been highly influential in the development of preference models, especially in the psychometric, econometric, and transportation literature. Because emphasis is on conceptual issues, technical topics such as sampling, estimation, and inference are not discussed. Related reviews by Böckenholt (2006), Böckenholt and Tsai (2006), Maydeu-Olivares and Böckenholt (2009), and Train (2009) cover the literature on these subjects in more detail.

Valuation of Choice Options

Most preference models are based on the notion of decision utility. This approach infers the value or utility of an option from observed choices and in turn uses this value to explain choices. For example, if a decision maker chooses option j and rejects option k, this behavior is interpreted as indicating that option j

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43082-5

has a higher utility than option k for this person. A number of consistency conditions have been proposed in the literature to characterize choices based on decision utilities (Tversky, 1969). These consistency conditions are important because they allow nontrivial predictions across new choice settings. Consider pairwise choices among the three options j, k, and l. If option j is preferred to option k, and option k is preferred to option l, we expect that option j is preferred to option l when decision makers are consistent in their ranking of preferences. In contrast, the selection of option l in the last paired comparison indicates an intransitive cycle, which suggests a change in valuation or choice strategy.

To determine whether the observed choice outcome is caused by systematic or random factors, Thurstone (1927) promoted an extended version of the decision–utility concept that allowed for stochastic variability in choices. By invoking Fechner's (1860) psychophysical concept of a sensory continuum, Thurstone proposed that choice options can be represented along a continuum by random variables that correspond to the options' effects on a person's sensory processing. The choice process can be described as the selection of an option with the greatest momentary utility realization. As a result, variability in choices can then be used to recover the features of the underlying utility distribution. Moreover, when the means of the utility realizations are stable across time and situations, consistency conditions such as contraction and expansion consistency across different choice settings can be defined for stochastic preferences (Block and Marschak, 1960). These conditions are supported in many applications, demonstrating the usefulness of this approach in predicting choice behavior for different sets of choice options. However, systematic violations in experimental studies have also been reported, as discussed in the final section.

It is noteworthy that Thurstone (1927) used randomness in choice as a device to represent factors that determine the formation of preferences but are unknown to the observer. This specification was driven by pragmatic considerations to avoid the question of whether a choice process is inherently random or determined by a multitude of different factors that may need to be measured. By not specifying the process by which a person "identifies, distinguishes, discriminates, or reacts to stimuli" (1927, p. 369) and defining it to be exogenous to all the factors that can affect a choice process, Thurstone broadened the applicability of this approach to the modeling of choice data in many different settings. Important to note is that he also utilized these ideas for modeling individual differences in choice behavior by arguing that heterogeneity in individual assessments of the same option can also be described as realizations of an underlying distribution. This approach proved instrumental for the development of econometric choice models (McFadden, 2001). When introducing Thurstone's work to economics, Marschak (1960) coined the term random utility maximization when referring to choice probabilities that result from the maximization of utilities with random elements.

Predicting Choices

Typically, stated choice data are collected in the form of incomplete and/or partial rankings. Incomplete ranking data are obtained when a decision maker considers only a subset of the choice options. For example, the method of paired comparison presents two choice options at a time, and the decision maker is asked to select the more preferred one. In contrast, in a partial ranking task, a decision maker is confronted with all choice options and asked to provide a ranking for a subset of the options. For instance, in the best–worst method, a decision maker is instructed to select the best and worst options from the offered set of choice options. Both partial and incomplete approaches can be combined by offering multiple distinct subsets of the choice options and obtaining partial or complete rankings for each of them. Pick-any, constant-sum, and ordinal versions of paired comparison and rankings are alternative methods for collecting stated preference data (Bock and Jones, 1968; David, 1988; Louviere et al., Hensher and Swait, 2000; Marshall and Bradlow, 2002).

We define a population P of decision makers ($i \in P$) and a finite set C of choice alternatives ($j \in C$). For each decision maker and choice alternative, a vector $\mathbf{x}_{ij}$ of observed variables is available that partially describes the pair (i,j). The underlying momentary utility of option j to person i is denoted by υ_{ij} and is a function of the observed and unobserved attribute values. In this section, distributional assumptions are made to capture the variation of the unobserved attribute values and preference differences among the decision makers. In combination with a decision rule for selecting an option, the utilities determine the options' choice probabilities. As a result, the choice probabilities can be represented by an a priori specified member of a restricted family of functions, which is partially determined by the distributional assumptions about the unobserved attribute values and utility variations.

To be specific, under Thurstone's (1927) approach, the choice of an option j is determined by the unobserved utility assessment, υ_{ij}, which can be decomposed into a systematic and a random part:

$$\upsilon_{ij} = \mu_{ij} + \epsilon_{ij} \qquad [1]$$

where μ_{ij} represents the mean evaluation of option j by person i. In contrast to the option mean, which is assumed to stay the same in repeated evaluations, the random contribution ϵ_{ij} varies from evaluation to evaluation according to some distribution.

In general, the choices made by person i from a single choice set of J options can be summarized by an ordering vector $\mathbf{r}_i$. For instance, $\mathbf{r}_i = (h, j, \ldots, l, k)$ indicates that choice option h is judged superior to option j, which in turn is judged superior to the remaining options, with the least preferred option being k. The probability of observing this ordering vector for person i can be written as

$$\Pr(\mathbf{r}_i = (h, j, \ldots, l, k)) = \Pr[(\upsilon_{ih} - \upsilon_{ij} > 0) \cap \ldots \cap (\upsilon_{il} - \upsilon_{ik} > 0)]$$

More compactly, a $(J-1)\times J$ contrast matrix, $\mathbf{C}_i$, can be defined to indicate the sign of the differences among the ranked options. For example, for the ordering vector $\mathbf{r}_i = (j, l, k)$ of $J = 3$, the corresponding contrast matrix takes on the form

$$\mathbf{C}_i = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

where the three columns of the contrast matrix correspond to the options j, k, and l, respectively.

The distributional assumptions involving ϵ_{ij} determine the probability for each choice pattern. A multivariate probit model is obtained when assuming that the latent judgments $\boldsymbol{\upsilon}_i = (\upsilon_{i1}, \ldots, \upsilon_{iJ})$ of the decision maker i are multivariate normal with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_\nu$. In this case, the probability of observing the ordering vector $\mathbf{r}_i$ is obtained by evaluating a $(J-1)$-variate normal distribution,

$$\Pr(\mathbf{r}_i|\boldsymbol{\mu}_i, \Sigma_\nu) = \frac{|\boldsymbol{\Gamma}_i|^{-1/2}}{(2\pi)^{(J-1)/2}} \int_0^\infty \cdots \int_0^\infty \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\delta}_i)'\boldsymbol{\Gamma}_i^{-1}(\mathbf{x}-\boldsymbol{\delta}_i)\right\} d\mathbf{x}$$

where $\boldsymbol{\delta}_i = \mathbf{C}_i\boldsymbol{\mu}_i$ and $\boldsymbol{\Gamma}_i = \mathbf{C}_i\boldsymbol{\Sigma}_\nu\mathbf{C}_i'$ (Hausman and Wise, 1978).

Typically, in experimental settings, choice options are presented to decision makers in multiple blocks, especially when the set of choice options is large. This multiple-judgment approach is used to simplify the choice task since only a few options need to be considered at a time. It also facilitates studying whether decision makers are consistent in their evaluations of the choice options.

Consider a paired comparison task in which a person is asked to select the preferred option for a subset or all of the pairwise choice sets. Extending [1], the utilities of decision maker i that underlie the pairwise comparison of options j and k, respectively, can be written as

$$\begin{aligned} \upsilon_{ij(k)} &= \mu_{ij} + \epsilon_{ij(k)} \\ \upsilon_{ik(j)} &= \mu_{ik} + \epsilon_{ik(j)} \end{aligned}$$

The terms $\epsilon_{ij(k)}$ and $\epsilon_{ik(j)}$ capture random variations in the choice process, where $\epsilon_{ij(k)}$ denotes the random variation of option j when compared to option k. Thus, the choice between j and k is determined by the sign of the pairwise comparison outcome y_{ijk} computed as the difference between the corresponding utilities

$$y_{ijk} = \upsilon_{ij(k)} - \upsilon_{ik(j)} = \mu_{ij} - \mu_{ik} + \epsilon_{ij(k)} - \epsilon_{ik(j)} \qquad [2]$$

It can be assumed that choices across choice sets are independent or dependent at the individual level. Under independence, one obtains

$$y_{ijk} = \mu_{ij} - \mu_{ik} + \epsilon_{ijk}$$

where $\epsilon_{ijk} = \epsilon_{ij(k)} - \epsilon_{ik(j)}$ and the variability in the choices comes solely from the ϵ's, which are independently distributed given the decision makers' parameters. For multiple comparisons of J options, the latent pairwise judgments of decision maker i can then be written conveniently as a linear model, which simplifies the estimation of this model.

Instead of assuming the random terms, such as $\epsilon_{ij(k)} - \epsilon_{ik(j)}$ and $\epsilon_{ij(l)} - \epsilon_{il(j)}$ in [2], to be uncorrelated, it is also meaningful to consider conditions that may lead to dependencies between them. For example, when an option is compared with two similar options, the utilities in these two comparisons are presumably more correlated than those obtained in a comparison with two less similar options. In this case, the difference in $\epsilon_{ij(k)}$ and $\epsilon_{ik(j)}$ cannot be reduced to the pair-specific term ϵ_{ijk}. This more general approach of allowing for choices among choice sets to be dependent proved useful in understanding intransitive choice behavior in temporal discounting and multiattribute choice applications (Tsai and Böckenholt, 2006; 2008).

Identifiability and Interpretation

Since choosing the most preferred option is assumed to be equivalent to selecting the option with the greatest utility, an important feature of the postulated choice process is its comparative nature. A selected option may be the best one from a set of available options, but it may be rejected in favor of other options that subsequently are added to the set. As choices are inherently comparative, the origin of the utility scale cannot be identified on the basis of the choices alone. One option may be preferred to another, but this result does not allow any conclusions about whether either of the options is inherently attractive or unattractive.

The lack of an origin has important consequences for identifying the underlying utility distributions. In particular, it is not possible to recover all elements of the options' covariance matrix of the unobserved utilities $\boldsymbol{\Sigma}_\nu$ on the basis of choice data. The two covariance matrices $\boldsymbol{\Sigma}_\nu$ and $\Sigma_\nu + \mathbf{d}1' + 1\mathbf{d}'$ are empirically indistinguishable provided the vector $\mathbf{d}$ is chosen so that the resulting covariance matrix is positive-definite (Tsai, 2003). Thus, a row- and column-specific constant can be added to each row and column of the covariance matrix without affecting the choice probabilities. As a result, when estimating $\boldsymbol{\Sigma}_\nu$, constraints need to be imposed to ensure identifiability.

The equivalence relation suggests defining the identified parameters based on the utility differences instead of the utilities of the options. For any contrast matrix $\mathbf{C}$ of rank $J-1$, the corresponding identified parameters for the model with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_\nu$ are $\mathbf{DC}\boldsymbol{\mu}$ and $d^{-1}\mathbf{C}\boldsymbol{\Sigma}_\nu\mathbf{C}'$, where $\mathbf{D} = \mathrm{Diag}(\mathbf{S}\boldsymbol{\Sigma}_\nu\mathbf{S}')^{-1/2}$, and d is the (1, 1) element of $\mathbf{C}\boldsymbol{\Sigma}_\nu\mathbf{C}'$. However, other identification constraints can also be chosen that do not involve a contrast matrix. Regardless of the constraints that are being selected, it is important to take into account that the original parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\nu$ cannot be recovered from preference data but only a lower dimensional function of them.

The existence of equivalence models for preference data complicates the interpretation of the results since preference models that have the same identified parameters may have substantively different interpretations. Consider the following two ranking models, A and B, for three normally distributed utilities with the respective mean vectors $\boldsymbol{\mu}_A = (2 \quad 5 \quad 0)'$ and $\boldsymbol{\mu}_B = (\sqrt{.8} \quad \sqrt{5} \quad 0)'$ and the covariance matrices $\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ and $\boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & .7 & .6 \\ .6 & 1 & .5 \\ .6 & .5 & 1 \end{bmatrix}$. Because these two models yield the same ranking probabilities of (0.038, 0.002, 0.802, 0.148, 0.002, 0.008) for the rank ordering patterns {(1, 2, 3), (1, 3,2), (2,1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)}, respectively, they are empirically indistinguishable. However, the interpretation of these two covariance matrices is rather different. Model A has different option variances and zero option correlations. In contrast, model B has equal option variances and nonzero correlations between options. Because of the models' empirical indistinguishability, an interpretation in favor of either model A or model B needs to depend on other factors, such as prior

knowledge or some theory about the underlying choice mechanism. This numerical example also illustrates a frequently overlooked implication in the literature: investigating the model fit of restrictions imposed on either $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}_{\nu}$ requires taking into account that any evidence in support of this restriction speaks for an entire class of equivalence models and not just for the model form under consideration.

It is possible to identify the origin of the utility scale by extending the choice task – for example, by including comparisons not only between single options but also between pairs of options. Böckenholt (2004) discusses different variations of the choice paradigm that can be used for this purpose. In general, methods for identifying a utility scale origin not only are instrumental for avoiding difficulties in the interpretation of the estimated parameters of a choice model, but also may provide useful insights about the underlying judgmental process.

Modeling Effects of Individual Differences, Time, and Social Dependencies

Over the years, the Thurstonian choice paradigm has been extended in different directions. Much work is underway to enhance the tool set for modeling individual differences both at a particular point in time and over time. There is also strong interest in going beyond modeling the relationship between choice options and decision makers and considering a wider range of factors, including social context and social influences.

Individual Differences

The analysis of individual difference effects is one of the most fruitful aspects of preference models. Past work considered nested and crossed random effects, person-specific regression weights, and factor structures to capture dependencies among options under flexible distributional assumptions that include mixtures of the extreme value distribution, multivariate normality, scale mixtures of multivariate normal distributions, and Dirchlet processes. For example, the following decomposition of an individual utility vector allows for the effects of person-specific covariates $\mathbf{x}_i$ as well as observed and unobserved attributes of the options, given by $\mathbf{M}$ and $\boldsymbol{\Lambda}$, respectively:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu} + \mathrm{K}\mathbf{x}_i + \mathrm{M}\boldsymbol{\beta}_i + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\nu}_i \qquad [3]$$

where $\mathbf{K}$, $\boldsymbol{\beta}_i$, and $\mathbf{f}_i$ contain regression coefficients. Effects that are not accounted for by the covariates are represented by $\boldsymbol{\nu}_i$. Under normality assumptions, $\boldsymbol{\nu}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\nu})$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\beta})$, and $\mathbf{f}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_f)$. A detailed discussion and application of this modeling framework is presented in Maydeu-Olivares and Böckenholt (2009), who first introduced the observed covariates describing the choice options and decision makers and then allow for unobserved option attributes to account for any remaining systematic variations in political preferences.

Although the interpretation of individual taste differences is much simplified under the assumption that decision makers use the same set of attributes in assessing the choice options, some care must be taken in relating these attributes to the option utilities. For example, the relationship between attributes and utilities may not be linear, as postulated by the factor-analytic model in [3], but nonmonotonic, as in the ideal-point model if individuals choose the option that is closest to their 'ideal' or most preferred option (Coombs, 1964).

Temporal Preferences

When preferences for only a single choice set are elicited, the effects of between- and within-decision-maker variability are confounded. For example, if decision makers are asked to rank J options only once, it is not possible to identify separately the effects of ϵ_{ij} and μ_{ij} in [1]. Fortunately, with multiple choice sets or when choices are observed over time, it becomes possible to distinguish explicitly these sources of variability. In the context of time-dependent choices, parameter-driven dynamic approaches have proved useful to capture dependencies on the within- and between-person levels. For example, in the context of modeling voting data collected before and after the 1992 presidential election, Böckenholt (2002) applied the following vector-autoregressive model with

$$\boldsymbol{\mu}_{it} = \mathbf{Z}\boldsymbol{\mu}_{i,t-1} + \boldsymbol{\zeta}_{it} \qquad [4]$$

where $\boldsymbol{\mu}_{it}$ are the options' mean evaluations of person i at time t, and the $(J \times J)$ coefficient matrix $\mathbf{Z}$ contains the multivariate regression effects between the evaluations of the options at two adjacent time points. The diagonal elements of $\mathbf{Z}$ contain each option's autoregressive term and the off-diagonal elements contain the lagged effects of the $J-1$ remaining random utilities. The J-dimensional random effect $\boldsymbol{\zeta}_{it}$ captures the part of $\boldsymbol{\mu}_{it}$ that cannot be predicted on the basis of past evaluations of the options.

Social Dependencies

Aside from temporal dependencies, there is also strong interest in modeling proximity-based dependencies that are introduced by known or unknown (latent) relationships among individuals. This work goes beyond the standard assumption that individuals make choices in isolation by allowing explicitly for the possibility that choices of individuals are correlated or influenced by each other. Although it has long been recognized (Veblen, 1899) that choices of individuals may depend on their social network, empirical research on this topic has been slow, mainly because of methodological problems (Brock and Durlauf, 2001). Modeling issues that need to be addressed in empirical work on social interactions are (1) the identification of the reference group for which social-interaction effects are sought to be established, (2) self-selection processes of peer or group members, (3) controls for correlated effects that affect all group members in a similar way, and (4) controls for contextual effects such as exogenous social background characteristics of group members (Manski, 2007). Because these issues are difficult to tackle in any observational study, progress is most likely to be made by combining observational with quasiexperimental or laboratory studies.

Outlook

Parallel to this expansion of preference models, a growing literature has developed over the years that questions their

descriptive value. This work showed that the consideration of unsystematic error in [1] is insufficient to account for inconsistent choice behavior. May (1954) presented initial evidence that strategies for comparing alternatives can lead to systematic violations of such consistency criteria as weak stochastic transitivity, which is defined for the three options j, k, and l as

$$\Pr(j > k) \geq .50 \quad \Pr(k > l) \geq .50 \Rightarrow \Pr(j > l) \geq .50$$

More generally, in the same way as the meaning of a word is clarified by the context of the sentence in which it appears, the valuation of a choice option frequently seems to depend on the context of the choice set. The 'asymmetric dominance' and the 'compromise' effects are two well-known examples of systematic violations of expansion consistency (Simonson & Tversky, 1992). In the asymmetric dominance case, the dominance relation between two options affects their evaluations. This effect is not present when each of the options is presented in combination with an option that is neither dominating nor dominated. The compromise effect occurs when a two-attribute option is presented as falling between two extreme options that are strong on one attribute but weak on the other attribute. In this case, the 'compromise' option seems to receive a boost for having medium values on both attributes. Such relational effects cannot be captured by models that consider context-independent evaluations of choice options.

Many factors have been shown to affect decision processes. This list includes contextual effects (e.g., relational features such as dominance among choice options and temporal features of the choice situation), choice processes (e.g., the evaluability of options, decision strategies, and information search strategies), presentation formats, frames, cultural and social norms, as well as characteristics of the decision maker (e.g., emotional state, general intelligence, and numeracy). In view of this list, it is perhaps surprising how well the basic decomposition in [1] can work in providing satisfactory accounts of choice data.

Although experimental work has convincingly revealed the limitations of preference models based on Thurstone's (1927) framework, few would dispute their usefulness in rendering parsimonious descriptions of how individuals perceive and evaluate choice options. Many applications have shown that preference models can provide both a parsimonious quantitative description of choice outcomes and a flexible framework for modeling individual and contextual differences in choice behavior at a particular point in time or over time. Concurrently, there has been renewed interest in having foundational work on preference modeling go beyond the restrictive assumptions that utilities are predetermined, decomposable, and stable both across time and situations. Better measures and indicators of utility are being investigated that may prove useful to capture the hedonic experience associated with a choice outcome. For example, to infer utility from choices alone without taking into account anticipated or experienced hedonic reactions no longer seems sufficient (Kahneman, 2011). Measures of brain activity and brain electrochemistry in combination with experimental treatments provide much needed insights on the links between choice and sensations of pleasure and pain (Glimcher, 2011). Ways to relax the notion of utility maximization are also under investigation to take into account that the accuracy of perceptions is constrained by the information-processing capacity of one's nervous system.

See also: Best Worst Scaling: Theory and Practice; Decision and Choice: Random Utility Models of Choice and Response Time; Dynamic Decision Making.

Bibliography

Block, H., Marschak, J., 1960. Random orderings and stochastic theories of response. In: Olkin, et al. (Eds.), Contributions to Probability and Statistics. Stanford University Press, Stanford, pp. 97–132.

Bock, R.D., Jones, L.V., 1968. The Measurement and Prediction of Judgment and Choice. Holden-Day, San Francisco.

Böckenholt, U., 2002. A Thurstonian analysis of preference change. Journal of Mathematical Psychology 46, 300–314.

Böckenholt, U., 2004. Comparative judgments as an alternative to ratings: Identifying the scale origin. Psychological Methods 9, 453–465.

Böckenholt, U., 2006. Thurstonian–based analyses: Past, present and future utilities. Psychometrika 71, 615–629.

Böckenholt, U., Tsai, R., 2006. Random-effects models for preference data. In: Rao, C.R., Sinharay, S. (Eds.), Handbook of Statistics, vol. 26. Elsevier Science, Amsterdam, pp. 447–468.

Brock, W.A., Durlauf, S.N., 2001. Interactions-based models. In: Handbook of Econometrics, vol. 5. North Holland, Amsterdam, pp. 3297–3380.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

David, H.A., 1988. The Method of Paired Comparisons. Griffin, London.

Fechner, G.T., 1860. Elemente der Psychophysik. Breitkopf und Hartel, Leipzig.

Glimcher, P.W., 2011. Foundations of Neuroeconomic Analysis. Oxford University Press, New York.

Hausman, J., Wise, D., 1978. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. Econometrica vol. 48, 403–429.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Louviere, J.J., Hensher, D.A., Swait, J.D., 2000. Stated Choice Methods. Cambridge University Press, New York.

Manski, C.F., 2007. Identification for Prediction and Decision. Harvard University Press, Cambridge.

Marschak, J., 1960. Binary choice constraints on random utility indictor. In: Arrow, K.I., Karlin, S., Suppes, P. (Eds.), Stanford Symposium on Mathematical Methods in the Social Sciences. Stanford University Press, Stanford, pp. 312–329.

Marshall, P., Bradlow, E.T., 2002. A unified approach to conjoint analysis models. Journal of the American Statistical Association 97, 674–682.

May, K.O., 1954. Intransitivity, utility, and the aggregation of preference patterns. Econometrica 22, 1–13.

Maydeu–Olivares, A., Böckenholt, U., 2009. Preference models. In: Millsap, R.E., Maydeu–Olivares, A. (Eds.), Handbook of Quantitative Methods in Psychology. Sage, Thousand Oaks.

McFadden, D., 2001. Economic choices. American Economic Review 91, 351–378.

Samuelson, P., 1938. A note on the pure theory of consumers' behaviour. Economica 5, 61–71.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.

Train, K., 2009. Discrete Choice Methods with Simulations. MIT Press, Cambridge.

Tsai, R., 2003. Remarks on the identifiability of Thurstonian paired comparison models under multiple judgment. Psychometrika 68, 361–372.

Tsai, R., Böckenholt, U., 2006. On the importance of distinguishing between within- and between-subject differences in intransitive choice data. Journal of Mathematical Psychology 52, 10–20.

Tsai, R., Böckenholt, U., 2008. On the importance of distinguishing between within– and between–subject effects in intransitive choice data. Journal of Mathematical Psychology 52, 10–20.

Tversky, A., 1969. Intransitivity of preference. Psychological Review 76, 31–48.

Veblen, T., 1899. The Theory of the Leisure Class. MacMillan, New York.

Psychoneuroimmunology

Adam J Massey, Faculty of Medicine and Health, School of Clinical Sciences, Queen's Medical Centre, University of Nottingham (UK), Nottingham, UK
Kavita Vedhara, Faculty of Medicine and Health, School of Medicine, University of Nottingham (UK), Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the field of Psychoneuroimmunology (PNI) which is a multidisciplinary field comprising the fields of psychology, immunology, and the neurosciences. The aim of PNI is to understand the bidirectional interrelationships between the mind, body, and disease. Therefore, PNI seeks to understand how stress and other psychological processes may impact upon nervous, endocrine, and immune system functioning. This article discusses the historical background of PNI, including key research findings as well as providing an overview of key psychological interventions used within this area.

Introduction

The field of psychoneuroimmunology (PNI) is the discipline that seeks to understand the complex interplay between the mind and body and how these connections may help explain the role of psychological processes in relation to health (Vedhara and Irwin, 2005). As such, PNI is inherently multidisciplinary in nature drawing predominantly upon the fields of psychology, immunology, and the neurosciences.

History

Two key figures regarded as the pioneers of PNI are Aaron Fredrick Rasmussen and George Solomon. In 1957, Aaron Fredrick Rasmussen and colleagues conducted a landmark study in which they demonstrated that mice that were put under stress were rendered more susceptible to the herpes simplex virus (Rasmussen et al., 1957). This, and a series of other studies, provided preliminary evidence that the brain may be capable of affecting the immune system. At the time this idea was considered with skepticism by the scientific community who largely believed that the central nervous system and immune systems were separate. Rasmussen's findings gained further credence from the work of George Solomon who also showed that psychological or environmental stressors may influence immunity (Solomon, 1969). In the ensuing years, Robert Ader and Nicholas Cohen were credited for advancing the field of PNI based upon a number of studies using animals, which further developed our understanding of how the brain may relate to the immune system. In one of their seminal experiments, Ader and Cohen (1975) gave rats varying volumes of a saccharin-laced water-based solution to drink while at the same time the rats were given Cytoxan, which is known to suppress the immune system and to cause unpleasant gastrointestinal symptoms. Ader and Cohen observed that the rats quickly began to avoid consuming the solution presumably because they had learned or had been conditioned to associate the solution with the unpleasant side effects. However, when they stopped giving the rats the drug, but continued to give them the saccharin-laced water alone, not only did the rats avoid drinking the solution, but some of the rats exhibited the same suppressed immune response shown previously. This finding was particularly interesting because the saccharin-laced water alone could not have been responsible for the impact the solution still appeared to have on the rats' immunity. Ader and Cohen reasoned that the rats had undergone a form of psychological conditioning and now maybe the brain was the mechanism by which the rats' immune systems were being suppressed.

These observations are regarded as playing a fundamental role in helping transform the way that we now think about the relationship between external stimuli, such as stressors, the brain, and the effect these factors can have on immunity. Robert Ader was later credited for coining the term 'psychoneuroimmunology.' A new era of science based upon understanding the role of the mind and body was born.

PNI and Stress

A key focus of PNI is the role that stress and other psychological processes may have on health through interactions between the nervous, endocrine, and immune systems. While many definitions exist, stress within this context is generally considered to be a cognitive process in which an individual subjectively appraises the demands of a situation, the resources they have to deal with the situation, and the resulting affective, behavioral and/or biological responses to this process (Cohen et al., 1995). Distress is thought to arise when the demands of a situation are believed to outweigh an individuals' subjective appraisal of their own ability to cope with the source of the stress i.e., the 'stressor,' which is typically characterized as chronic or acute. It is thought that the two forms of stress may impact upon the immune system differently. Chronic stressors are enduring major life stressors such as bereavement whereas acute stressors are shorter lived and include the everyday hassles that most of us are likely to encounter such performing a public-speaking task. It is believed that our default response to stress may serve an evolutionary role in that energy resources are mobilized when we perceive danger or threat.

The two main systems, which become activated in response to this process are the sympathetic–adrenal–medullary (SAM) system and the hypothalamus–pituitary–adrenal (HPA) axis. Activation of these systems results in increased secretion of stress hormones. The SAM system is activated by the autonomic

nervous system (ANS) and is characterized by the 'fight or flight' response of the sympathetic nervous system. The sympathetic branch of the ANS stimulates the release of norepinephrine, which results in elevated heart rate, blood pressure, and blood flow to essential organs. Several minutes later the HPA axis becomes activated resulting in the release of the stress hormone cortisol which, as well as other functions, plays a role in the regulation of immune responses. The release of stress hormones provides immediate physiological benefits when confronted with danger, such as fueling immediate energy release, yet it is rare that the daily demands that we encounter in the modern world always require this level of physiological output. The demands on an individual who is running away from a lion are quite different to the demands of bereavement, for example. Thus, while the stress response process is thought to serve important evolutionary functions, problems may arise when the SAM and HPA systems are chronically activated in response to relatively benign everyday situations. This is because chronic secretion of stress hormones are thought to cause changes to the distribution and activity of immune cells, which may well lead to poorer immune functioning (Vedhara and Irwin, 2005). The current understanding in PNI is that acute stress may enhance immune function whereas chronic stress may be detrimental to it.

A plethora of PNI studies have now shown that chronic or long-term emotional states such as long-lasting stress, anxiety, and depression are associated with alterations in immune functioning and worsened health outcomes for many conditions. Indeed, while the scope of PNI has broadened over the last 30 years, four main areas remain useful paradigms for exploring the role of emotional states in relation to health outcomes. The four main areas include wound healing, common cold studies, infections and vaccinations, and heart disease. These contexts allow psychoneuroimmunologists to examine the role of psychological processes and gain clinically relevant information on stress-related immune function and health. An overview of research and current understanding in each of these areas will now be considered.

Wound Healing

Wound healing provides a good model for investigating how psychological processes may impact upon immunity because the immune system is heavily involved at all stages of the wound repair process. Evidence for the role of stress on wound healing comes from both experimental and clinical studies. In a classic experimental study, Kiecolt-Glaser et al. (1995) demonstrated that wound healing took 24% longer in those individuals experiencing a form of everyday chronic stress. In this study, the researchers conducted a biopsy on the arms of 13 women who provided long-term care for an ailing relative and compared their healing rates to 13 aged matched controls who also underwent the biopsy procedure. The authors found that the carers took 9 days longer to heal compared to the age matched controls. These findings suggest that caregiving, a chronic stressor, may impair the early inflammatory response resulting in delayed wound healing. In a further study, Marucha et al. (1998) performed two biopsies on the hard palate within the mouths of 11 dental students. The first biopsy was conducted during vacation time and the second biopsy was performed 3 days prior to the students' examination period. Marucha et al. observed that the students took 40% longer to heal during their examination period, and none of students healed as rapidly during their examination period as they did during their vacation. In a similar study, Ebrecht et al. (2004) performed a 4 mm dermal biopsy on 24 healthy young men and monitored the men's healing rates 7, 14, and 21 days after the biopsy was performed using ultrasound scanning. Ebrecht et al. found that the slower healers exhibited significantly higher stress levels compared to those healing quicker.

Clinical studies also provide valuable insight and a number of studies have now been conducted, which suggest that stress, and other psychological factors, may impact upon the wound-healing process. Broadbent et al. (2003) assessed the effects of preoperative stress on postsurgical wound repair by assessing levels of stress before patients underwent a hernia operation. The authors found that higher stress levels predicted lower immune responses and greater worry about the operation predicted a more painful and slower recovery. Further, Doering et al. (2005) examined 72 patients undergoing coronary bypass surgery and found that higher stress-related symptoms at discharge predicted infection, impaired wound healing, and poor emotional and physical recovery. In fact, the authors found that the patients were 3.7 times more likely to experience wound infections and healing problems some 6 weeks after discharge.

Walburn et al. (2008) conducted a systematic review of published research in this area and found that stress was associated with healing rate across a range of wound contexts including acute and chronic forms of wounds within clinical and experimental studies. Taken together, evidence from both experimental and clinical studies combined suggest that stress and other psychological factors impact upon the wound-healing process. Other research in this field also highlight the importance of coping styles, positive affect, and the role of environmental factors such as social support in the healing process (Broadbent et al., 2012).

Common Cold Studies

We are all continually exposed to infectious agents, such as common cold viruses, yet only a proportion of us develop an infectious illness. Research conducted within this area suggest that stress may increase susceptibility to infectious illness. In a key study, Cohen et al. (1991) assessed the self-reported stress of 394 participants who rated their stress levels over the previous 12 months. Upon rating their stress levels, participants were then given one of five respiratory viruses, which were administered by the researchers using nasal drops. Cohen et al. found that the degree of stress the participants reported was related to the occurrence of infection in a dose-dependent manner. That is, the more life stress experienced over the previous 12 months the more likely the person was to become infected with the virus or suffer a cold. These findings were supported in a further study that found lower resistance to the common cold virus in participants reporting severe chronic stress for more than 1 month (Cohen et al., 1998).

Research examining naturally occurring viral infections also suggest that stress may be associated with an increased natural incidence of upper respiratory illness (Cohen et al., 2002).

Smolderen et al. (2007) investigated the stress levels, personality characteristics, and self-reported influenza-like symptoms in 5404 respondents from the general population and found that low mood and perceived stress were associated with increased reporting of influenza-like symptoms over a 4-week period. Further, Pederson et al. (2010) conducted a systematic review and meta-analysis of 27 published studies exploring the association between psychological stress and the risk of developing upper respiratory infection (URI). Overall, the findings of the review were that increased stress was associated with increased susceptibility to URI. Further research also suggest that these findings are independent of the age-related declines in immunity (Cohen et al., 2013).

Infections and Vaccinations

Vaccinations are commonly administered to prevent diseases such as influenza, hepatitis A and B, with the widespread use of vaccinations offering a useful paradigm for PNI research. Studies within PNI suggest that psychological factors may be important in determining how well an individual responds to a variety of vaccinations (Burns et al., 2003, 2002; Glaser et al., 2000; Jabaaij et al., 1993; Marsland et al., 2001; Phillips et al., 2005). Medical vaccinations, in particular hepatitis B and influenza, have received most of the attention to date. For example, psychological stress has been found to be inversely associated with the immune response to hepatitis B (Burns et al., 2002; Glaser et al., 1992; Jabaaij et al., 1993; Marsland et al., 2001). Results from other studies using the influenza vaccination have also found psychological stress to have negative consequences for immune responses in both younger (Burns et al., 2003; Phillips et al., 2005; Pressman et al., 2005) and older samples (Moynihan et al., 2004; Phillips et al., 2006). Responses to vaccinations have been found to be poorer in elderly individuals who provide long-term care for an unwell spouse compared with elderly non-caregivers (Kiecolt-Glaser et al., 1996; Vedhara et al., 1999; Wong et al., 2013). This is because providing care for an ailing spouse is regarded as a chronic form of life stress, which has both emotional and physical demands. This form of chronic stress may be particularly detrimental to the health of elderly individuals because of the underlying immunity impairments associated with aging. For example, no difference in vaccination response was observed between younger caregivers of multiple sclerosis patients compared to controls (Vedhara et al., 2002). This raises the issue of whether older caregivers, as a consequence of an aging immune system, are more susceptible to the effects of chronic stress. Alternatively, the findings could also reflect differences in caregiving experiences between caregiver populations. The findings of studies using younger populations have been mixed. The results suggest that general life stress may be associated with immune responses in younger populations (Burns et al., 2003; Phillips et al., 2005), whereas in older adults bereavement and nongeneral life events tended to be associated with impaired immunity (Phillips et al., 2006). Taken together, these studies suggest that psychological stress impacts upon immune responses, but that these factors may differ in salience across the life span.

A wide range of psychological factors such as an individual's perception of their social support and feelings of loneliness has also been found to affect vaccination responses. For example, feelings of loneliness and smaller social network size have been associated with lower antibody responses to influenza vaccination (Pressman et al., 2005). Students reporting lower social support produced poorer antibody response to pneumococcal vaccination whereas students reporting greater numbers of stressful life events also produced poorer responses to hepatitis A vaccination (Gallagher et al., 2008). Research also suggest that those individuals exhibiting a general feeling of elevated mood demonstrate more robust immune responses to vaccinations (Marsland et al., 2006) while those individuals exhibiting greater neuroticism have been found to produce weaker immune responses (Phillips et al., 2005).

Heart Disease

Several longitudinal studies have been conducted which suggest that stress may lead to a higher incidence of cardiac disease. For example, a longitudinal study over 14 years found that workers reporting higher chronic stress were more likely to develop risk factors for cardiovascular disease (Chandola et al., 2006). Further, a study of over 6000 men followed on average for almost 12 years showed that men reporting higher stress at the beginning of the study were more likely to develop coronary artery disease (Rosengren et al., 1991). Other studies in this area have compared cases of heart attack with control groups and demonstrated that higher stress (e.g., work, family, and financial) is associated with increased risk of myocardial infarction (Rosengren et al., 2004). Indeed, epidemiological studies suggest that stress and other psychological factors may contribute to the development of coronary artery disease while it is also recognized that factors such as personality characteristics and social isolation may also play an important mediator roles (Rozanski et al., 1999). More recently, Carroll et al. (2012) found that blood pressure reactions to a laboratory stressor were related to cardiovascular disease mortality 16 years later in a sample of 431 patients. These findings suggest that exaggerated cardiovascular responses in relation to acute stressful events may, if experienced frequently, impact upon long-term cardiovascular mortality. In a meta-analytic review, Steptoe and Kivimaki (2013) report that early life stressors, such as childhood abuse and early socioeconomic adversity, are linked to increased cardiovascular morbidity in adulthood. Further, the authors found an increased risk of coronary heart disease among adults experiencing job-related work stress and social isolation while stress, anger, and depressed mood may act as triggers of major cardiac events.

How Can Psychoneuroimmunology Be Used to Improve Health Outcomes?

While researchers have made remarkable progress in understanding the role of stress and other psychological factors on health, an important next step is to consider what interventions may be used to alleviate the adverse effects of

psychological functioning on immunity. One way that psychological interventions such as stress management, relaxation training, and emotional disclosure may confer benefits is by decreasing negative mood states and, therefore, reducing the associated nervous system, hormonal activation, and immune system functioning. Here we discuss some of the main interventions that have been used in PNI.

Emotional Disclosure

Emotional disclosure refers to acknowledging and discussing personal issues with attention placed on labeling the source of the problem and discussing its causes and consequences (Pennebaker, 1999). Research suggest that writing about important emotional issues for three to five consecutive days for 15–30 min each day may have an impact upon number of physician visits, immune responses, autonomic activity, as well as encourage improvements in overall mood (Pennebaker, 1999). An advantage of this technique is that a therapist does not need to be present as it is thought that the process of writing about one's feelings drives the reappraisal process. Indeed, writing about an issue may cause short-term distress but it is thought the long-term gains outweigh this effect (Booth, 2005). Petrie et al. (1995) showed that healthy medical students who wrote continuously for 20 min about emotional life events over a period of 4 days showed an enhanced antibody response to vaccination 4 and 6 months after the vaccination compared to a control group who wrote about unemotional life events only. Furthermore, research also suggest that emotional disclosure may impact upon wound healing. For example, Weinman et al. (2008) conducted a study in which 36 healthy participants were randomized to either an expressive writing group, which wrote about emotional events or a time management writing group. The two groups were given a punch biopsy 2 weeks later. Findings were that the wounds healed quicker in the emotional disclosure group. Several meta-analyses have also been published, which suggest that emotional disclosure, through the use of expressive writing may well be a cost-effective intervention which may be used to combat the influence of negative mood on health (Smyth, 1998; Frattaroli, 2006).

Cognitive Behavior Therapy and Psychotherapy

Cognitive behavioral interventions have been used in PNI in an attempt to alleviate stress and prevent the progression of disease by enhancing emotional well-being. A study by Antoni et al. (2000) found that a cognitive behavioral stress management (CBSM) intervention reduced anxiety, anger, and stress, and was associated with better immune responses in HIV-infected men compared with a control group. Cruess et al. (2000) found that a 10-week CBSM intervention provided positive immune benefits to HIV patients. Further, a study investigating 43 elderly caregivers demonstrated that a CBSM intervention, which included a weekly 1-hour counseling session delivered over 8 weeks was associated with enhanced antibody responses to influenza vaccination following the intervention (Vedhara et al., 2003). In other research, Antoni et al. (2009) conducted a study in which women breast cancer patients were randomized to receive either a 10-week CBSM intervention or a 1-day educational program 4–8 weeks after undergoing surgery. Those patients undergoing the CBSM intervention exhibited lower anxiety and better immune responses compared to controls.

A number of studies in this area have been conducted on the effects of psychological interventions on cancer survival. Spiegel et al. (1989) found that of 86 women diagnosed with breast cancer, those who participated in weekly expressive-supportive group therapy for 1 year survived 18 months longer over a 10-year period. However, the authors replicated this study some years later and found no statistically significant effect on survival time (Spiegel et al., 2007).Other research conducted in this area suggest that psychotherapy may well play a role in immune functioning and cancer survival. For example, Fawzy et al. (1993) evaluated the recurrence and survival of 69 patients diagnosed with malignant melanoma who had participated in a 6-week intervention, which involved education, coping, and stress management techniques. Fawzy et al. found that patients exhibited increased immune activity and survived for longer compared to those not receiving the intervention. However, evidence for the role of psychotherapy on cancer survival is currently mixed (Coyne et al., 2007).

Relaxation Techniques

The role that relaxation interventions may have on immunity responses has been investigated in a number of studies using nonpatient populations. For example, elderly individuals who engaged in muscle relaxation training for 1 month exhibited an improved immunity response to herpes simplex virus (Kiecolt-Glaser et al., 1985). Further, Irwin et al. (2003) observed improved immune responses in elderly individuals who underwent a 15-week tai chi program. Meditation, which is another form of mental relaxation has also been found to show positive effects on a number of immune and health outcomes. For example, a study of 48 healthy individuals found that those who attended an 8-week meditation course produced better immune responses in response to influenza vaccination and showed increased brain activity normally associated with positive emotion (Davidson et al., 2003). Nevertheless, while research using nonpatient populations are valuable, the use of clinical populations provides the most insightful and realistic information regarding the efficacy of relaxation techniques. Broadbent et al. (2012) examined the effects of relaxation on wound-healing processes in patients undergoing elective surgery. Participants were randomized to either receive standard care or standard care plus a relaxation intervention, which included guided imagery and take-home relaxation CDs that the patients listened to before and after undergoing surgery. Perceived stress before and after surgery was assessed using questionnaires. The authors found that those patients randomized to receive the relaxation intervention healed significantly quicker than the patients only receiving standard care. These findings provide preliminary evidence that a brief and relatively low-cost

relaxation intervention prior to surgery can reduce stress and improve the wound-healing process in a clinical population.

Conclusion

Since the early observations of Rasmussen in 1957, and the subsequent birth of psychoneuroimmunology some years later, PNI has provided almost unprecedented insight into the role of the mind in determining health outcomes. The multidisciplinary nature of PNI can be complex, and indeed present challenges; yet contributors to the field have been instrumental in breaking down multidisciplinary boundaries, which once restricted our understanding of the mind–body link. It is clear that PNI researchers have come a long way in demonstrating how stress and other psychological factors may impact upon health through interactions with the nervous, endocrine, and immune systems. In doing so, this has not only provided a sound evidence base but also improved our understanding of the psychological interventions and techniques, which may be used to alleviate the deleterious role of stress.

In future years, advances in the related fields of immunology and the neurosciences will likely prove promising as new methods of assessment and new ways of understanding immunity and disease are developed. Such advances will enable the field of PNI to further our understanding of how we as human beings function and how cognitions, emotions, and behaviors relate to health and well-being.

See also: Cancer: Psychosocial Aspects; Caregiver Burden; Coronary Heart Disease: Psychosocial Aspects; Depression, Pessimism, and Health; Stress, Coping and Health.

Bibliography

Ader, R., Cohen, N., 1975. Behaviourally conditioned immunosuppression. Psychosomatic Medicine 37, 333–340.

Antoni, M., Cruess, D.G., Cruess, S., Lutgendorf, S., Kumour, M., Ironson, G., 2000. Cognitive behavioural stress management intervention effects on anxiety, 24 hour urinary norepinephrine and T cytotoxic suppressor cells over time among symptomatic HIV-infected gay men. Journal of Consulting and Clinical Psychology 68, 31–45.

Antoni, M.H., Lechner, S., Diaz, A., Vargas, S., Hollay, H., Phillips, K., McGregor, B., Carver, S.C., Blomberg, B., 2009. Cognitive behavioural stress management effects on psychosocial and physiological adaptation in women undergoing treatment for breast cancer. Health, Psychology and Immunity 23 (5), 59.

Booth, R., 2005. Emotional disclosure and psychoneuroimmunology. In: Vedhara, K., Irwin, M.R. (Eds.), Human Psychoneuroimmunology. Oxford University Press, Oxford, pp. 319–341.

Broadbent, E., Kahoekehr, A., Booth, R.J., Thomas, J., Windsor, J.A., Buchanan, C.M., Wheeler, B.R., Sammour, T., Hill, A.G., 2012. A brief relaxation intervention reduces stress and improves surgical wound healing response: a randomised trial. Brain, Behavior, and Immunity 26, 212–217.

Broadbent, E., Petrie, K.J., Alley, P.G., Booth, R.J., 2003. Psychological stress impairs early wound repair following surgery. Psychosomatic Medicine 65, 865–869.

Burns, V.E., Carroll, D., Ring, C., Harrison, L.K., Drayson, M., 2002. Stress, coping, and hepatitis B antibody status. Psychosomatic Medicine 64, 287–293.

Burns, V.E., Carroll, D., Drayson, M., Whitham, M., Ring, C., 2003. Life events, perceived stress and anti-body response to influenza vaccination in young healthy adults. Journal of Psychosomatic Research 55, 569–572.

Carroll, D., Ginty, A., Der, G., Hunt, K., Benzeval, M., Phillips, A.C., 2012. Increased blood pressure reactions to acute mental stress are associated with 16 year cardiovascular disease mortality. Psychophysiology 49 (10), 1444–1448.

Chandola, T., Brunner, E., Marmot, M., 2006. Chronic stress at work and the metabolic syndrome: prospective study. British Medical Journal 332, 521–525.

Cohen, S., Tyrrell, D.A.J., Smith, A.P., 1991. Psychological stress and susceptibility to the common cold. New England Journal of Medicine 325, 606–612.

Cohen, S., Kessler, R.C., Gordon, L.U., 1995. Measuring Stress – A Guide for Health and Social Scientists. Oxford University Press, New York.

Cohen, S., Frank, E., Doyle, W.J., Skoner, D.P., Rabin, B.S., Gwaltney, J.M., 1998. Types of stressors that increase susceptibility to the common cold in healthy adults. Health and Psychology 17, 214–223.

Cohen, S., Hamrick, N., Rodriguez, M.S., Feldman, P.J., Rabin, B.S., Manuck, S.B., 2002. Reactivity and the vulnerability to stress associated risk for upper respiratory illness. Psychosomatic Medicine 64, 302–310.

Cohen, S., Janicki – Deverts, D., Turner, R.B., Casselbrant, M.L., Li Korotky, S., Epel, E.S., Doyle, W.J., 2013. Association between telomere length and experimentally induced upper respiratory viral infection in healthy adults. Journal of the American Medical Association 309 (7), 699–705.

Coyne, J.C., Stefanek, M., Palmer, S.C., 2007. Psychotherapy and survival in cancer: the conflict between hope and evidence. Psychological Bulletin 133 (3), 367–394.

Cruess, S., Antoni, M., Cruess, D., Fletcher, M.A., Ironson, G., Kumar, M., 2000. Reductions in herpes simplex virus type 2 antibody titers after cognitive behavioural stress management and relationships with neuroendocrine function, relaxation skills and social support in HIV-positive men. Psychosomatic Medicine 62, 828–837.

Davidson, R.J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, M., Santotorelli, S.F., 2003. Alterations in brain and immune function produced by mindfulness mediation. Psychosomatic Medicine 65, 564–570.

Doering, L.V., Moser, D.K., Lemonkiewicz, W., Luper, C., Khan, S., 2005. Depression healing and recovery from coronary artery bypass surgery. American Journal of Critical Care 14, 316–324.

Ebrecht, M., Hextall, J., Kirtley, L.G., Taylor, A., Dyson, M., Weinman, J., 2004. Perceived stress and cortisol levels predict speed of wound healing in healthy male adults. Psychoneuroendocrinology 29 (6), 798–809.

Fawzy, F.F., Fawzy, N.W., Hyun, C.S., Elashoff, R., Gothrie, D., Fahey, J.L., 1993. Malignant Melanoma: effects of an early structures psychiatric intervention, coping, and affective state on recurrence and survival 6 years later. Archives of General Psychiatry 50, 681–689.

Frattaroli, J., 2006. Experimental disclosure and its moderators: a meta-analysis. Psychological Bulletin 132, 823–865.

Gallagher, S., Phillips, A.C., Ferraro, A.J., Drayson, M.T., Carroll, D., 2008. Psychosocial factors are associated with antibody response to both thymus-dependant and thymus independent vaccines. Brain, Behaviour and Immunity 22, 456–460.

Glaser, R., Kielcolt-Glaser, J.K., Bonneau, R.H., Malarkey, W., Kennedy, S., Hughes, J., 1992. Stress induced modulation of the immune response to recombinant hepatitis B vaccine. Psychosomatic Medicine 54, 22–29.

Glaser, R., Sheriden, J., Malarkey, W.B., MacCallum, R.C., Kielcolt-Glaser, J.K., 2000. Chronic stress modulates the immune response to pneumococcal pneumonia vaccine. Psychosomatic Medicine 62 (6), 804–807.

Hemingway, H., Malik, M., Marmot, M., 2001. Social and psychosocial influences on sudden cardiac death, ventricular arrhythmia and cardiac autonomic function. European Health Journal 22, 1082–1101.

Irwin, M.R., Pike, J.L., Cole, J.C., Oxman, M.N., 2003. Effects of a behavioural intervention, tai chi chih, on varicella-zoster virus specific immunity and health functioning in older adults. Psychosomatic Medicine 65, 824–830.

Jabaaij, L., Grosheide, P.M., Heijtink, R.A., Duivenvoorden, H.J., Ballieux, R.E., Vingerhoets, A.J., 1993. Influence of perceived psychological stress and distress on antibody response to low dose rDNA hepatitis B vaccine. Journal of Psychosomatic Research 37, 361–369.

Kiecolt-Glaser, J.K., Glaser, R., Williger, D., Stout, J., Messick, G., Sheppard, S., 1985. Psychosocial enhancement of immunocompetence in a geriatric population. Health Psychology 4, 25–41.

Kiecolt-Glaser, J.K., Marucha, P.T., Malarkey, W.B., Mercado, A.M., Glaser, R., 1995. Slowing of wound healing by psychological stress. Lancet 346, 1194–1196.

Kiecolt-Glaser, J.K., Glaser, R., Gravenstein, S., Malarkey, W.B., Sheridan, J., 1996. Chronic stress alters the immune response to influenza virus vaccine in older adults. Proceedings of the National Academy of Sciences 93, 3043–3047.

Marsland, A.L., Cohen, S., Rabin, B.S., Manuck, S.B., 2001. Associations between stress, trait negative affect, acute immune reactivity, and antibody response to hepatitis B injection in healthy young adults. Health Psychology 20, 1–8.

Marsland, A.L., Cohen, S., Robin, B.S., Manuck, S.B., 2006. Trait positive affect and antibody response to hepatitis B vaccination. Brain, Behaviour and Immunity 20, 261–269.

Marucha, P.T., Kiecolt-Glaser, J.K., Favagehi, M., 1998. Mucosal wound healing is impaired by examination stress. Psychosomatic Medicine 60, 362–365.

Moynihan, J.A., Larson, M.R., Treanor, J., Duberstein, P.R., Power, A., Shore, B., Ader, R., 2004. Psychosocial factors and the response to influenza vaccination in older adults. Psychosomatic Medicine 66, 950–953.

Pederson, A.F., Zachariae, R., Boubjerg, D., 2010. The influence of psychological stress on upper respiratory infection: meta-Analysis of prospective studies. Psychosomatic Medicine, 823–832.

Pennebaker, J.W., 1999. Writing about emotional experiences as a therapeutic process. Psychological Science 8, 162–166.

Petrie, K.J., Booth, R.J., Pennebaker, J.W., Davison, K.P., Thomas, M.G., 1995. Disclosure of trauma and immune response to a hepatitis B vaccination program. Journal of Consulting and Clinical Psychology 63, 787–792.

Phillips, A.C., Carroll, D., Burns, V.E., Drayson, M., 2005. Neuroticism, cortisol reactions to acute stress, and antibody response to influenza vaccination. Psychophysiology 42, 232–238.

Phillips, A.C., Carroll, D., Burns, V.E., et al., 2006. Bereavement and marriage are associated with antibody response to influenza vaccination in the elderly. Brain, Behaviour and Immunity 20, 279–289.

Pressman, S.D., Cohen, S., Miller, G.E., Barkin, A., Rabin, B.S., Treanor, J.J., 2005. Loneliness, social network size and immune responses to influenza vaccination in college freshman. Health Psychology 24, 297–306.

Rasmussen Jr., A.F., Marsh, J.T., Brill, N.Q., 1957. Increased susceptibility to herpes simplex in mice subjected to avoidance-learning stress or restraint. Proceedings of the Society for Experimental Biology and Medicine 96, 183.

Rosengren, A., Tibblin, G., Wilhelmsen, L., 1991. Self-perceived psychological stress and incidence of coronary artery disease in middle aged men. American Journal of Cardiology 68, 1171–1175.

Rosengren, A., Hawken, S., Gunpuu, S., Sliwa, K., Zubaid, M., Almahmeed, M., 2004. Association of psychosocial risk factors with risk of acute myocardial infarction in 11,119 cases and 13,648 controls from 52 countries (The INTER-HEART study): case control study. Lancet 364, 953–962.

Rozanski, A., Blumenthal, J.A., Kaplan, J., 1999. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation 99, 2192–2217.

Smolderen, K., Vingerhoets, A., Croon, M., Denollet, J., 2007. Personality, psychological stress and self-reported influenza symptoms. BMC 7, 339.

Smyth, J., 1998. Written emotional expression: effect sizes. Outcome types and moderating variables. Journal of Clinical and Consulting Psychology 66, 174–184.

Solomon, G.F., 1969. Stress and antibody response in rats. International Archives of Allergy 35, 97.

Spiegel, D., Bloom, J.R., Kraemer, H.C., Gotteril, E., 1989. Effect of psychosocial treatment on survival of patients with metastatic breast cancer. Lancet 2, 88–891.

Spielgel, D., Butler, L.B., Glese-Davis, J., Koopman, C., Miller, E., Dimiceli, J., 2007. Supportive-expressive group therapy and survival in patients with metastatic breast cancer: a randomised clinical trial. Cancer 110, 1130–1138.

Steptoe, A., Kivimaki, M., 2013. Stress and cardiovascular disease: an update on current knowledge. Annual Review of Public Health 34, 337–354.

Vedhara, K., Cox, N.K.M., Wilcock, G.K., Perks, P., Hunt, M., Anderson, S., 1999. Chronic stress in elderly carers of dementia patients and antibody responses to influenza. Lancet 353, 627–631.

Vedhara, K., McDermott, M.P., Evans, T.G., Treanor, J.J., Plummer, S., Tallon, D., Cruttenden, K.A., Schifitto, G., 2002. Chronic stress in nonelderly caregivers: psychological, endocrine and immune implications. Journal of Psychosomatic Research 53, 1153–1161.

Vedhara, K., Bennett, P.D., Clark, S., Lightman, S.L., Shaw, S., Perks, P., 2003. Enhancement of antibody responses to influenza vaccination in the elderly following a cognitive behavioural stress management intervention. Psychotherapy and Psychosomatics 72, 245–252.

Vedhara, K., Irwin, M., 2005. Human Psychoneuroimmunology. Oxford University Press, Oxford.

Walburn, J., Vedhara, K., Hankins, M., Rixon, L., Weinman, J., 2008. Psychological stress and wound healing in humans: a systematic review and meta-analysis. Journal of Psychosomatic Medicine 67, 253–271.

Weinman, J., Ebrecht, M., Scott, S., Walburn, J., Dyson, M., 2008. Enhanced wound healing after emotional disclosure intervention. British Journal of Health Psychology 13 (95), 102.

Wong, S.Y.S., Chan, F.W., Wong, C.K., Chan, P.K.S., Ngai, K., Mercer, S., Woo, J., 2013. Chronic psychosocial stress: does it modulate immunity to the influenza vaccine in Hong Kong Chinese elderly caregivers. Age 35 (4), 1479–1493.

Relevant Website

www.pnirs.org – Psychoneuroimmunology Research Society.

Psychopathy

Stuart F White and James R Blair, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Abstract

Psychopathy is a developmental disorder associated with increased risk of instrumental and reactive aggression and with chronic and violent antisocial behavior in both youth and adults. Individuals with psychopathy show specific impairments in stimulus-reinforcement learning, emotional processing, and moral decision making. These deficits have been tied to dysfunctional amygdala response and to problems appropriately representing reinforcement expectancies in ventromedial prefrontal cortex. While treatment recommendations are currently sparse, recent work has shown that previous assessments of treatment amenability in this population may have been overly pessimistic.

Introduction

Psychopathy is a disorder that characterizes an individual showing pronounced emotional deficits and an increased risk for displaying antisocial behavior (Frick, 1995; Hare, 2003). The affective component of psychopathy, often called callous-unemotional (CU) traits, is the distinguishing feature of the disorder relative to other behavioral profiles associated with an increased likelihood of antisocial behavior (Blair, 2007). CU traits refer to a lack of guilt, remorse, and empathy (Frick et al., 2005). While it is past antisocial behavior that is particularly important in predicting future criminal activity (Walters, 2003), it is CU traits that are at the core of developmental trajectory associated with psychopathy (Frick and White, 2008). The disorder is developmental. It has been shown that CU traits in particular and the psychopathy more generally are relatively stable from childhood into adulthood (Lynam et al., 2007; Munoz and Frick, 2007). In addition, the functional impairments seen in adults with psychopathy (e.g., in responding to emotional expressions, aversive conditioning, passive avoidance learning, reversal learning, extinction) are also seen in adolescents with psychopathic tendencies (see later).

Several formal assessment tools for psychopathy exist, including the Antisocial Process Screening Device (Frick and Hare, 2001) and Psychopathy Checklist – Youth Version (Forth et al., 2007) for adolescents and the Psychopathy Checklist-Revised (Hare, 2003) for adults. Most research has identified three dimensions of behavior that compose psychopathy (Cooke et al., 2006; Frick et al., 2000; Neumann et al., 2006), though other models have been proposed (Cooke et al., 2006). These three dimensions are as follows: an affective factor that focuses on CU traits; an arrogant, deceitful, and narcissistic interpersonal style; and an impulsive, irresponsible, and antisocial behavior (Cooke et al., 2006; Frick et al., 2000; Neumann et al., 2006).

Psychopathy is not equivalent to the psychiatric conditions of conduct disorder (CD) or antisocial personality disorder (ASPD) as defined by DSM-5 or their ICD-10 counterparts. The diagnostic criteria for these disorders focus on antisocial behaviors rather than on etiological factors such as the emotion dysfunction seen in psychopathy (Blair et al., 2005). As such these psychiatric conditions describe individuals with difficulties in executive dysfunction (Moffitt, 1993), as well as individuals with symptoms stemming from CU traits. Consequently, individuals with psychopathy are a more homogenous group than those individuals meeting the criteria for CD and ASPD (Karnik et al., 2006). It should be noted, however, that DSM-5 includes the specifier for CD 'with limited pro-social emotions,' which stem directly from research on youth with CD and CU traits (Pardini et al., 2010; Pardini and Fite, 2010). Furthermore, the diagnosis of ASPD now includes components of psychopathy (APA, 2013). While the disorder of psychopathy will still not be equivalent to the DSM-5 diagnoses of CD and ASPD, there will be greater overlap in diagnostic conceptualization.

Psychopathy is characterized by an increased risk for antisocial behavior (Frick and Dickens, 2006; Hare, 2003). While several psychiatric disorders and neurological conditions, including CD and ASPD (APA, 2013), confer an increased risk of reactive aggression (Anderson et al., 1999; Leibenluft et al., 2003), psychopathy is unique in that it conveys increased risk for instrumental aggression (Frick et al., 2003). Furthermore, psychopathy is strongly predictive of institutional maladjustment and recidivism (Edens and Campbell, 2007; Walters, 2003), particularly violent recidivism (Edens et al., 2007; Hemphill et al., 1998).

Causes of Psychopathy

Genetic

Behavioral genetics work in children and adolescents (for a review, see Viding and McCrory, 2012) suggests that CU traits and psychopathy are highly heritable. Various twin studies across different groups have found the genetic influences account for between 40 and 78% of the variance in CU traits in the population (Viding and McCrory, 2012). Critically, in a large twin sample, antisocial behavior in youth high on CU traits was highly heritable ($h^2 = 0.81$), while antisocial behavior was much less heritable ($h^2 = 0.30$) in youth low on CU traits (Viding et al., 2005).

Despite the evidence suggesting that genetics plays an important etiological role in psychopathy, relatively little is known about the molecular genetics of the disorder. Studies in youth suggest that the long allele of the serotonin transporter

gene may be a risk factor for psychopathy (Sadeh et al., 2010, 2013). Additionally, Hirata et al. (2013) report an association between the single-nucleotide polymorphism rs6269 of catechol-*O*-methyltransferase and CU traits in a modest sample of antisocial youth. Though these early findings are promising, they are preliminary and further work is needed. Furthermore, it is quite possible that different genotypes are associated with the various deficits associated with psychopathy (see later).

Environmental

Relatedly, antisocial behavior in psychopaths and youth with CD and CU traits appears to be less influenced by environmental factors than antisocial behavior in youth with low levels of CU traits. Poor parenting is less associated with antisocial behavior in youth with high levels of CU traits (Hipwell et al., 2007; Oxford et al., 2003; Wootton et al., 1997; Yeh et al., 2011). Moreover, youth with CD and high levels of CU traits are less likely to have a history of trauma than youth with CD and low levels of CU traits (Kahn et al., 2013; Kimonis et al., 2012). Critically, despite the increased genetic influence and reduced environmental influence on the development of psychopathy and CD with high levels of CU traits, this does not mean that psychopathy is genetically determined (Viding and McCrory, 2012). In fact, there have been data to suggest that socioeconomic status plays a role in the expression of genetic risk for psychopathy (Sadeh et al., 2010), though the mechanisms of this gene by environment interaction are not understood.

Pathophysiology of Psychopathy

With respect to cognitive processing (including emotional processing) in psychopathy, the disorder is particularly interesting given the selectivity in the impairments seen. For example, executive functioning (Morgan and Lilienfield, 2000), theory of mind (Blair et al., 1996), and episodic memory (as long as it does not rely on augmentation by emotional content; Christianson et al., 1996) are intact in individuals with this disorder. There are, however, specific cognitive deficits in the processing of certain emotional expressions and in stimulus-reinforcement learning associated with psychopathy.

Empathy Impairment

Individuals with psychopathy show reduced autonomic responses to the pain and distress of others (Aniskiewicz, 1979; Blair et al., 1997; House and Milligan, 1976) and reduced recognition of emotional expressions (see Dawel et al., 2012; Marsh and Blair, 2008). Interestingly, this impairment in expression recognition is relatively selective. Recognition of fearful, sad, and happy expressions is reduced while the recognition of disgusted and angry expressions remains intact (Dawel et al., 2012; Marsh and Blair, 2008).

Decision-Making Impairment

Furthermore, psychopaths show impairment in stimulus-reinforcement learning. Psychopaths show reduced aversive conditioning; they are less likely to show autonomic activity to stimuli associated with shock (Rothemund et al., 2012). These individuals also show pronounced difficulties with reinforcement-based decision-making modalities including extinction (Newman et al., 1987), reversal learning (Budhani et al., 2006), economic decision making (Koenigs et al., 2010), and moral judgments (Aharoni et al., 2011; Blair, 1995; Blair et al., 1995; Koenigs et al., 2011; Young et al., 2012).

Moral Judgment Impairment

Individuals with psychopathy show specific deficits in certain kinds of moral judgments. It has been argued that different forms of norms are acquired and processed by at least partially dissociable systems. An early version of this position relates to the 'moral/conventional distinction'; the distinction between norms whose violation results in unjust/unfair treatment or in harmful consequences to others (harm/welfare-based; e.g., one person hitting another) as opposed to those whose violation challenges social conventions/norms (conventional; e.g., one person talking to another during class; Nucci and Nucci, 1982a,b; Smetana, 1995; Turiel, 1983). Psychopaths show impairment in distinguishing between harm/welfare-based and conventional norms (Blair, 1995). Individuals with psychopathy show reduced endorsement of harm/welfare-based norms (Aharoni et al., 2011; Glenn et al., 2009). They also show an increased likelihood to allow actions that indirectly harm another (Koenigs et al., 2011).

Neural Impairment

Structural Neuroimaging

There have been three relatively consistent structural MRI findings with respect to psychopathy. Psychopathy is associated with reduced amygdala volumes (Ermer et al., 2012; Fairchild et al., 2013, 2011; Huebner et al., 2008; Sterzer et al., 2007; Yang et al., 2010). Reduced ventromedial prefrontal cortex (vmPFC) has been seen in adults, though not consistently in youth (Ermer et al., 2012; Ly et al., 2012; Yang et al., 2009). Similarly, diffusion tensor-imaging studies have found reduced structural integrity of the uncinate fasciculus in adults, but not youth, with psychopathy relative to comparison individuals (Craig et al., 2009; Motzkin et al., 2011; Sundram et al., 2012). The uncinate fasciculus is a white matter tract providing extensive connection between the orbitofrontal cortex and the amygdala. Furthermore and again, only in adults, four studies have reported reductions in temporal pole (Ermer et al., 2012; Gregory et al., 2012; Ly et al., 2012; Yang et al., 2009), two in STS (de Oliveira-Souza et al., 2008; Yang et al., 2009), and three in orbitofrontal cortex (de Oliveira-Souza et al., 2008; Ermer et al., 2012; Yang et al., 2009).

Functional Neuroimaging

Findings in functional MRI studies (where IQ has been appropriately considered) have been relatively consistent for the amygdala. Individuals with psychopathy have been

reported to show amygdala activity during moral judgment (Glenn et al., 2008; Harenski et al., 2010). Additionally, a reduction in amygdala activity was seen in an emotional memory paradigm (Kiehl et al., 2001). In youth with psychopathic traits, reduced amygdala response to fear/distress in others has been found consistently (Carre et al., 2012; Jones et al., 2009; Marsh et al., 2008; Viding et al., 2012; White et al., 2012). In adults, this finding has been less consistent. Violent schizophrenic patients with psychopathy show reduced amygdala responses to fearful expressions (Dolan and Fullam, 2009) – though a relationship between psychopathic traits in aggressive individuals and amygdala responsiveness to fear/distress expressions was not seen in another study (Pardini and Phillips, 2010).

Reduced vmPFC and caudate activation to reward has been seen in youth with psychopathic traits (Finger et al., 2011), though findings in adults have been inconsistent (Bjork et al., 2012; Buckholtz et al., 2010; Pujara et al., 2013). Importantly, recent neuroimaging work suggests that the reduced response to reward in vmPFC and caudate is associated with antisocial behavior generally and is not specific to psychopathy (Crowley et al., 2010; Rubia et al., 2009; White et al., 2013b). It is important to emphasize that while vmPFC and caudate dysfunction may not be specific to psychopathy, deficits in these regions have been observed in psychopathy and particularly vmPFC dysfunction in psychopathy is well documented (Blair, 2007).

Development of Psychopathy

As noted earlier, psychopathy is associated with deficits in emotional responses (particularly to fear/distress in others), impaired decision making, and problems making moral judgments. These deficits stem directly from neural impairments observed in both the structural and functional neuroimaging work. The amygdala is a key structure in stimulus-reinforcement learning (Everitt et al., 2003). In other words, the amygdala is necessary to learn the 'goodness' and 'badness,' or reinforcement value, of actions, objects, or events. In humans and other primates, fear/distress in conspecifics serves as a negative reinforcer, which is learned, via the amygdala, to be avoided (Hooker et al., 2006; Mineka and Zinbarg, 2006). Individuals with psychopathy fail to represent the fear/distress of others as aversive due to amygdala dysfunction (Sylvers et al., 2011; White et al., 2012). Furthermore, the amygdala projects forward into vmPFC, supplying important information regarding the valence of reinforcement expectations for use in decision making (Blair et al., 2006). As the amygdala is dysfunctional in psychopathy, inappropriate reinforcement expectation information is likely passed forward for use in decision making (Blair, 2007).

These deficits come together to predispose individuals with psychopathy to antisocial behavior (Blair, 2007). Deficits in harm/welfare-based norm processing make it more difficult for psychopaths to distinguish between prosocial and antisocial behavior. Deficits in representing the fear/distress of others as aversive make it much easier for psychopaths to engage in instrumental aggression. Finally, the decision-making deficits seen in psychopathy will likely increase the exposure to frustration. As humans exposed to frustration are more likely to reactively aggress (Berkowitz, 1993), psychopaths are also at increased risk of reactive aggression (Blair, 2010).

Treatment of Psychopathy

Initial examinations of treatment amenability in psychopathy were relatively negative. A number of investigations found that both adults (Ogloff et al., 1990) and youth with psychopathy (Falkenbach et al., 2003; Gretton et al. 2001; Hawes et al., 2013; O'Neill et al., 2003; Spain et al., 2004) showed worse response to treatment than other antisocial individuals. More recent work, however, has begun to examine factors underpinning differential treatment response in psychopathy (Hawes and Dadds, 2005) and suggests that previous pessimism regarding psychopathy and treatment response may have been overstated (Caldwell et al., 2006; Hawes and Dadds, 2005; Kolko and Pardini, 2010; White et al., 2013).

Summary

Psychopathy is a developmental disorder associated with increased risk of both instrumental and reactive aggression. The disorder is associated with chronic and violent antisocial behavior in both youth and adults. Individuals with psychopathy show specific impairments in stimulus-reinforcement learning, emotional processing, and moral decision making. These deficits have been tied to dysfunctional amygdala response and to problems appropriately representing reinforcement expectancies in vmPFC. While treatment recommendations are currently sparse, recent work has shown that previous assessments of treatment amenability in this population may have been overly pessimistic.

Acknowledgments

This work was supported by the Intramural Research Program of the National Institute of Mental Health, National Institutes of Health under grant number 1-ZIA-MH002860-08.

See also: Forensic Psychology.

Bibliography

Aharoni, E., Antonenko, O., Kiehl, K.A., 2011. Disparities in the moral intuitions of criminal offenders: the role of psychopathy. Journal of Research in Personality 45 (3), 322–327. http://dx.doi.org/10.1016/j.jrp.2011.02.005.

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., Damasio, A.R., 1999. Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nature Neuroscience 2 (11), 1032–1037. http://dx.doi.org/10.1038/14833.

Aniskiewicz, A.S., 1979. Autonomic components of vicarious conditioning and psychopathy. Journal of Clinical Psychology 35, 60–67.

APA, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Berkowitz, L., 1993. Aggression: Its Causes, Consequences, and Control. Mcgraw-Hill Book Company, New York, England.

Bjork, J.M., Chen, G., Hommer, D.W., 2012. Psychopathic tendencies and mesolimbic recruitment by cues for instrumental and passively obtained rewards. Biological Psychology 89 (2), 408–415. http://dx.doi.org/10.1016/j.biopsycho.2011.12.003 pii:S0301-0511(11)00301-2.

Blair, K., Marsh, A.A., Morton, J., Vythilingam, M., Jones, M., Mondillo, K., Blair, J.R., 2006. Choosing the lesser of two evils, the better of two goods: specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice. Journal of Neuroscience 26 (44), 11379–11386. http://dx.doi.org/10.1523/JNEUROSCI.1640-06.2006 pii:26/44/11379.

Blair, R.J., 1995. A cognitive developmental approach to mortality: investigating the psychopath. Cognition 57 (1), 1–29 pii:001002779500676P.

Blair, R.J.R., 2007. The amygdala and ventromedial prefrontal cortex in morality and psychopathy. Trends in Cognitive Sciences 11 (9), 387–392. http://dx.doi.org/10.1016/j.tics.2007.07.003.

Blair, R.J.R., 2010. Psychopathy, frustration, and reactive aggression: the role of ventromedial prefrontal cortex. British Journal of Psychology 101 (3), 383–399. http://dx.doi.org/10.1348/000712609x418480.

Blair, R.J.R., Jones, L., Clark, F., Smith, M., 1995. Is the psychopath "morally insane"? Personality and Individual Differences 19, 741–752.

Blair, R.J.R., Jones, L., Clark, F., Smith, M., 1997. The psychopathic individual: a lack of responsiveness to distress cues? Psychophysiology 34, 192–198.

Blair, R.J.R., Mitchell, D.G.V., Blair, K.S., 2005. The Psychopath: Emotion and the Brain. Blackwell, Oxford.

Blair, R.J.R., Sellars, C., Strickland, I., Clark, F., Williams, A., Smith, M., Jones, L., 1996. Theory of mind in the psychopath. Journal of Forensic Psychiatry 7, 15–25.

Buckholtz, J.W., Treadway, M.T., Cowan, R.L., Woodward, N.D., Benning, S.D., Li, R., Zald, D.H., 2010. Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. Nature Neuroscience 13 (4), 419–421. http://dx.doi.org/10.1038/nn.2510.

Budhani, S., Richell, R.A., Blair, R.J., 2006. Impaired reversal but intact acquisition: probabilistic response reversal deficits in adult individuals with psychopathy. Journal of Abnormal Psychology 115 (3), 552–558. http://dx.doi.org/10.1037/0021-843X.115.3.552 pii:2006-09167-016.

Caldwell, M., Skeem, J., Salekin, R., Van Rybroek, G., 2006. Treatment response of adolescent offenders with psychopathy features: a 2-year follow-up. Criminal Justice and Behavior 33 (5), 571–596. http://dx.doi.org/10.1177/0093854806288176.

Carre, J.M., Hyde, L.W., Neumann, C.S., Viding, E., Hariri, A.R., 2012. The neural signatures of distinct psychopathic traits. Society for Neuroscience. http://dx.doi.org/10.1080/17470919.2012.703623.

Christianson, S.A., Forth, A.E., Hare, R.D., Strachan, C., Lidberg, L., Thorell, L.H., 1996. Remembering details of emotional events: a comparison between psychopathic and nonpsychopathic offenders. Personality and Individual Differences 20, 437–443.

Cooke, D.J., Michie, C., Hart, S., 2006. Facets of clinical psychopathy: toward clearer measurement. In: Patrick, C.J. (Ed.), The Handbook of Psychopathy. Guilford Press, New York, pp. 91–106.

Craig, M.C., Catani, M., Deeley, Q., Latham, R., Daly, E., Kanaan, R., Murphy, D.G., 2009. Altered connections on the road to psychopathy. Molecular Psychiatry 14 (10), 946–953, 907 http://dx.doi.org/10.1038/mp.2009.40, pii:mp200940.

Crowley, T.J., Dalwani, M.S., Mikulich-Gilbertson, S.K., Du, Y.P., Lejuez, C.W., Raymond, K.M., Banich, M.T., 2010. Risky decisions and their consequences: neural processing by boys with antisocial substance disorder. PLoS ONE 5 (9). http://dx.doi.org/10.1371/journal.pone.0012835.

Dawel, A., O'Kearney, R., McKone, E., Palermo, R., 2012. Not just fear and sadness: meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy. Neuroscience and Biobehavioral Reviews 36 (10), 2288–2304.

Dolan, M.C., Fullam, R.S., 2009. Psychopathy and functional magnetic resonance imaging blood oxygenation level-dependent responses to emotional faces in violence patients with schizophrenia. Biological Psychiatry 66, 570–577.

Edens, J., Campbell, J., 2007. Identifying youths at risk for institutional misconduct: a meta-analytic investigation of the psychopathy checklist measures. Psychological Services 4 (1), 13–27. http://dx.doi.org/10.1037/1541-1559.4.1.13.

Edens, J.F., Campbell, J.S., Weir, J.M., 2007. Youth psychopathy and criminal recidivism: a meta-analysis of the psychopathy checklist measures. Law and Human Behavior 31 (1), 53–75. http://dx.doi.org/10.1007/s10979-006-9019-y.

Ermer, E., Cope, L.M., Nyalakanti, P.K., Calhoun, V.D., Kiehl, K.A., 2012. Aberrant paralimbic gray matter in criminal psychopathy. Journal of Abnormal Psychology 121 (3), 649–658. http://dx.doi.org/10.1037/a0026371.

Everitt, B.J., Cardinal, R.N., Parkinson, J.A., Robbins, T.W., 2003. Appetitive behavior: impact of amygdala-dependent mechanisms of emotional learning. Annals of the New York Academy of Sciences 985, 233–250.

Fairchild, G., Hagan, C.C., Walsh, N.D., Passamonti, L., Calder, A.J., Goodyer, I.M., 2013. Brain structure abnormalities in adolescent girls with conduct disorder. Journal of Child Psychology and Psychiatry, and Allied Disciplines 54 (1), 86–95. http://dx.doi.org/10.1111/j.1469-7610.2012.02617.x.

Fairchild, G., Passamonti, L., Hurford, G., Hagan, C.C., von dem Hagen, E.A., van Goozen, S.H., Calder, A.J., 2011. Brain structure abnormalities in early-onset and adolescent-onset conduct disorder. American Journal of Psychiatry 168 (6), 624–633. http://dx.doi.org/10.1176/appi.ajp.2010.10081184.

Falkenbach, D.M., Poythress, N.G., Heide, K.M., 2003. Psychopathic features in a juvenile diversion population: reliability and predictive validity of two self-report measures. Behavioral Sciences and the Law 21 (6), 787–805. http://dx.doi.org/10.1002/bsl.562.

Finger, E.C., Marsh, A.A., Blair, K.S., Reid, M.E., Sims, C., Ng, P., Blair, R.J., 2011. Disrupted reinforcement signaling in the orbitofrontal cortex and caudate in youths with conduct disorder or oppositional defiant disorder and a high level of psychopathic traits. American Journal of Psychiatry 168 (2), 152–162. http://dx.doi.org/10.1176/appi.ajp.2010.10010129 pii:appi.ajp.2010.10010129.

Forth, A.E., Kosson, D.S., Hare, R.D., 2007. The Psychopathy Checklist: Youth Version. Multi-Health Systems, Toronto, Ontario, Canada.

Frick, D., 2006. Current perspectives on conduct disorder. Current Psychiatry Reports 8 (1), 59–72.

Frick, S., Dandreaux, Farrell, K., 2005. Callous-unemotional traits in predicting the severity and stability of conduct problems and delinquency. Journal of Abnormal Child Psychology 33 (4), 471–487.

Frick, P.J., 1995. Callous-unemotional traits and conduct problems: a two-factor model of psychopathy in children. Issues in Criminological and Legal Psychology 24, 47–51.

Frick, P.J., Bodin, S.D., Barry, C.T., 2000. Psychopathic traits and conduct problems in community and clinic-referred samples of children: further development of the psychopathy screening device. Psychological Assessment 12 (4), 382–393.

Frick, P.J., Cornell, A.H., Barry, C.T., Bodin, S.D., Dane, H.E., 2003. Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency. Journal of Abnormal Child Psychology 31 (4), 457–470.

Frick, P.J., Hare, R.D., 2001. The Antisocial Process Screening Device. Multi-Health Systems, Toronto.

Frick, P.J., White, S.F., 2008. Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. Journal of Child Psychology and Psychiatry 49 (4), 359–375. http://dx.doi.org/10.1111/j.1469-7610.2007.01862.x.

Glenn, A.L., Iyer, R., Graham, J., Koleva, S., Haidt, J., 2009. Are all types of morality compromised in psychopathy. Journal of Personality Disorders 23, 384–398.

Glenn, A.L., Raine, A., Schug, R.A., 2008. The neural correlates of moral decision-making in psychopathy. Molecular Psychiatry 14, 5–6.

Gregory, S., ffytche, D., Simmons, A., Kumari, V., Howard, M., Hodgins, S., Blackwood, N., 2012. The antisocial brain: psychopathy matters. Archives of General Psychiatry 69 (9), 962–972. http://dx.doi.org/10.1001/archgenpsychiatry.2012.222.

Gretton, H.M., McBride, Michelle, Hare, Robert D., O'Shaughnessy, Roy, Kumka, Gary, 2001. Psychopathy and recidivism in adolescent sex offenders. Criminal Justice and Behavior 28 (4), 427–449. http://dx.doi.org/10.1177/009385480102800403.

Hare, R.D., 2003. Hare Psychopathy Checklist-Revised. PCL-R, second ed. Multi Health Systems, Toronto.

Harenski, C.L., Harenski, K.A., Shane, M.S., Kiehl, K.A., 2010. Aberrant neural processing of moral violations in criminal psychopaths. Journal of Abnormal Psychology 119 (4), 863–874. http://dx.doi.org/10.1037/a0020979 pii:2010-23724-009.

Hawes, D.J., Dadds, M.R., Brennan, J., Rhodes, T., Cauchi, A., 2013. Revisiting the treatment of conduct problems in children with callous-unemotional traits. The Australian and New Zealand Journal of Psychiatry 47 (7), 646–653. http://dx.doi.org/10.1177/0004867413484092.

Hawes, D.J., Dadds, M.R., 2005. The treatment of conduct problems in children with callous-unemotional traits. Journal of Consulting and Clinical Psychology 73 (4), 737–741. http://dx.doi.org/10.1037/0022-006x.73.4.737.

Hemphill, J.F., Hare, R.D., Wong, S., 1998. Psychopathy and recidivism: a review. Legal and Criminological Psychology 3, 139–170.

Hipwell, A.E., Pardini, D.A., Loeber, R., Sembower, M., Keenan, K., Stouthamer-Loeber, M., 2007. Callous-unemotional behaviors in young girls: shared and unique effects relative to conduct problems. Journal of Clinical Child and Adolescent Psychology 36 (3), 293–304. http://dx.doi.org/10.1080/15374410701444165.

Hirata, Y., Zai, C.C., Nowrouzi, B., Beitchman, J.H., Kennedy, J.L., 2013. Study of the catechol-*O*-methyltransferase (COMT) gene with high aggression in children. Aggressive Behavior 39 (1), 45–51. http://dx.doi.org/10.1002/ab.21448.

Hooker, C.I., Germine, L.T., Knight, R.T., D'Esposito, M., 2006. Amygdala response to facial expressions reflects emotional learning. Journal of Neuroscience 26 (35), 8915–8922. http://dx.doi.org/10.1523/jneurosci.3048-05.2006.

House, T.H., Milligan, W.L., 1976. Autonomic responses to modeled distress in prison psychopaths. Journal of Personality and Social Psychology 34, 556–560.

Huebner, T., Vloet, T.D., Marx, I., Konrad, K., Fink, G.R., Herpertz, S.C., Herpertz-Dahlmann, B., 2008. Morphometric brain abnormalities in boys with conduct disorder. Journal of the American Academy of Child and Adolescent Psychiatry 47 (5), 540–547. http://dx.doi.org/10.1097/CHI.0b013e3181676545.

Jones, A.P., Laurens, K.R., Herba, C.M., Barker, G.J., Viding, E., 2009. Amygdala hypoactivity to fearful faces in boys with conduct problems and callous-unemotional traits. American Journal of Psychiatry 166 (1), 95–102. http://dx.doi.org/10.1176/appi.ajp.2008.07071050 pii:appi.ajp.2008.07071050.

Kahn, R.E., Frick, P.J., Youngstrom, E.A., Kogos Youngstrom, J., Feeny, N.C., Findling, R.L., 2013. Distinguishing primary and secondary variants of callous-unemotional traits among adolescents in a clinic-referred sample. Psychological Assessment 25 (3), 966–978. http://dx.doi.org/10.1037/a0032880.

Karnik, N.S., McMullin, M.A., Steiner, H., 2006. Disruptive behaviors: conduct and oppositional disorders in adolescents. Adolescent Medicine Clinics 17 (1), 97–114.

Kiehl, K.A., Smith, A.M., Hare, R.D., Mendrek, A., Forster, B.B., Brink, J., Liddle, P.F., 2001. Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. Biological Psychiatry 50, 677–684.

Kimonis, E.R., Frick, P.J., Cauffman, E., Goldweber, A., Skeem, J., 2012. Primary and secondary variants of juvenile psychopathy differ in emotional processing. Development and Psychopathology 24 (3), 1091–1103. http://dx.doi.org/10.1017/s0954579412000557.

Koenigs, M., Kruepke, M., Newman, J.P., 2010. Economic decision-making in psychopathy: a comparison with ventromedial prefrontal lesion patients. Neuropsychologia 48 (7), 2198–2204. http://dx.doi.org/10.1016/j.neuropsychologia.2010.04.012 pii:S0028-3932(10)00145-4.

Koenigs, M., Kruepke, M., Zeier, J., Newman, J.P., 2011. Utilitarian moral judgment in psychopathy. Social Cognitive and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nsr048.

Kolko, David J., Pardini, Dustin A., 2010. ODD dimensions, ADHD, and callous-unemotional traits as predictors of treatment response in children with disruptive behavior disorders. Journal of Abnormal Psychology 119 (4), 713–725. http://dx.doi.org/10.1037/a0020910.

Leibenluft, E., Blair, R.J., Charney, D.S., Pine, D.S., 2003. Irritability in pediatric mania and other childhood psychopathology. Annals of the New York Academy of Sciences 1008, 201–218.

Ly, M., Motzkin, J.C., Philippi, C.L., Kirk, G.R., Newman, J.P., Kiehl, K.A., Koenigs, M., 2012. Cortical thinning in psychopathy. American Journal of Psychiatry 169 (7), 743–749. http://dx.doi.org/10.1176/appi.ajp.2012.11111627.

Lynam, D.R., Caspi, A., Moffitt, T.E., Loeber, R., Stouthamer-Loeber, M., 2007. Longitudinal evidence that psychopathy scores in early adolescence predict adult psychopathy. Journal of Abnormal Psychology 116 (1), 155–165. http://dx.doi.org/10.1037/0021-843X.116.1.155 pii:2007-01891-015.

Marsh, A.A., Blair, R.J., 2008. Deficits in facial affect recognition among antisocial populations: a meta-analysis. Neuroscience and Biobehavioral Reviews 32 (3), 454–465.

Marsh, A.A., Finger, E.C., Mitchell, D.G., Reid, M.E., Sims, C., Kosson, D.S., Blair, R.J., 2008. Reduced amygdala response to fearful expressions in children and adolescents with callous-unemotional traits and disruptive behavior disorders. American Journal of Psychiatry 165 (6), 712–720. http://dx.doi.org/10.1176/appi.ajp.2007.07071145 pii:appi.ajp.2007.07071145.

Mineka, S., Zinbarg, R., 2006. A contemporary learning theory perspective on the etiology of anxiety disorders: it's not what you thought it was. The American Psychologist 61 (1), 10–26. http://dx.doi.org/10.1037/0003-066x.61.1.10.

Moffitt, T.E., 1993. The neuropsychology of conduct disorder. Development and Psychopathology 5, 135–152.

Morgan, A.B., Lilienfield, S.O., 2000. A meta-analytic review of the relation between antisocial behavior and neuropsychological measures of executive function. Clinical Psychology Review 20, 113–136.

Motzkin, J.C., Newman, J.P., Kiehl, K.A., Koenigs, M., 2011. Reduced prefrontal connectivity in psychopathy. Journal of Neuroscience 31 (48), 17348–17357. http://dx.doi.org/10.1523/jneurosci.4215-11.2011.

Munoz, L.C., Frick, P.J., 2007. The reliability, stability, and predictive utility of the self-report version of the Antisocial Process Screening Device. Scandinavian Journal of Psychology 48, 299–312.

Neumann, C.S., Kosson, D.S., Forth, A.E., Hare, R.D., 2006. Factor structure of the hare psychopathy checklist: youth version (PCL: YV) in incarcerated adolescents. Psychological Assessment 18, 142–154.

Newman, J.P., Patterson, C.M., Kosson, D.S., 1987. Response perseveration in psychopaths. Journal of Abnormal Psychology 96, 145–148.

Nucci, L.P., Nucci, M., 1982a. Children's responses to moral and conventional transgressions in free-play settings. Child Development 52, 1337–1342.

Nucci, L.P., Nucci, M., 1982b. Children's social interactions in the context of moral and conventional transgressions. Child Development 53, 403–412.

O'Neill, M.L., Lidz, V., Heilbrun, K., 2003. Adolescents with psychopathic characteristics in a substance abusing cohort: treatment process and outcomes. Law and Human Behavior 27 (3), 299–313. http://dx.doi.org/10.1023/a:1023435924569.

Ogloff, J.R., Wong, S., Greenwood, A., 1990. Treating criminal psychopaths in a therapeutic community program. Behavioral Sciences and the Law 8 (2), 181–190. http://dx.doi.org/10.1002/bsl.2370080210.

Oxford, M., Cavell, T.A., Hughes, J.N., 2003. Callous/unemotional traits moderate the relation between ineffective parenting and child externalizing problems: a partial replication and extension. Journal of Clinical Child and Adolescent Psychology 32 (4), 577–585. http://dx.doi.org/10.1207/s15374424jccp3204_10.

de Oliveira-Souza, R., Hare, R.D., Bramati, I.E., Garrido, G.J., Ignacio, F.A., Tovar-Moll, F., Moll, J., 2008. Psychopathy as a disorder of the moral brain: fronto-temporo-limbic grey matter reductions demonstrated by voxel-based morphometry. Neuroimage 40, 1202–1213.

Pardini, D.A., Frick, P.J., Moffitt, T.E., 2010. Building an evidence base for DSM-5 conceptualizations of oppositional defiant disorder and conduct disorder: introduction to the special section. Journal of Abnormal Psychology 119 (4), 683–688. http://dx.doi.org/10.1037/a0021441 pii:2010-23724-002.

Pardini, D.A., Phillips, M., 2010. Neural responses to emotional and neutral facial expressions in chronically violent men. Journal of Psychiatry and Neuroscience 35 (6), 390–398. http://dx.doi.org/10.1503/jpn.100037.

Pardini, D.A., Fite, P.J., 2010. Symptoms of conduct disorder, oppositional defiant disorder, attention-deficit/hyperactivity disorder, and callous-unemotional traits as unique predictors of psychosocial maladjustment in boys: advancing an evidence base for DSM-V. Journal of the American Academy of Child and Adolescent Psychiatry 49 (11), 1134–1144.

Pujara, M., Motzkin, J.C., Newman, J.P., Kiehl, K.A., Koenigs, M., 2013. Neural correlates of reward and loss sensitivity in psychopathy. Social Cognitive and Affective Neuroscience. http://dx.doi.org/10.1093/scan/nst054.

Rothemund, Y., Ziegler, S., Hermann, C., Gruesser, S.M., Foell, J., Patrick, C.J., Flor, H., 2012. Fear conditioning in psychopaths: event-related potentials and peripheral measures. Biological Psychology 90 (1), 50–59. http://dx.doi.org/10.1016/j.biopsycho.2012.02.011.

Rubia, K., Smith, A.B., Halari, R., Matsukura, F., Mohammad, M., Taylor, E., Brammer, M.J., 2009. Disorder-specific dissociation of orbitofrontal dysfunction in boys with pure conduct disorder during reward and ventrolateral prefrontal dysfunction in boys with pure ADHD during sustained attention. The American Journal of Psychiatry 166 (1), 83–94. http://dx.doi.org/10.1176/appi.ajp.2008.08020212.

Sadeh, N., Javdani, S., Jackson, J.J., Reynolds, E.K., Potenza, M.N., Gelernter, J., Verona, E., 2010. Serotonin transporter gene associations with psychopathic traits in youth vary as a function of socioeconomic resources. Journal of Abnormal Psychology 119 (3), 604–609. http://dx.doi.org/10.1037/a0019709.

Sadeh, N., Javdani, S., Verona, E., 2013. Analysis of monoaminergic genes, childhood abuse, and dimensions of psychopathy. Journal of Abnormal Psychology 122 (1), 167–179. http://dx.doi.org/10.1037/a0029866.

Smetana, Judith G., 1995/1994. Morality in Context: Abstractions, Ambiguities and Applications. In: Annals of Child Development: A Research Annual, vol. 10. Jessica Kingsley Publishers, London, England, 83–130.

Spain, S.E., Douglas, K.S., Poythress, N.G., Epstein, M., 2004. The relationship between psychopathic features, violence and treatment outcome: the comparison of three youth measures of psychopathic features. Behavioral Sciences & the Law 22 (1), 85–102. http://dx.doi.org/10.1002/bsl.576.

Sterzer, P., Stadler, C., Poustka, F., Kleinschmidt, A., 2007. A structural neural deficit in adolescents with conduct disorder and its association with lack of empathy. Neuroimage 37 (1), 335–342. http://dx.doi.org/10.1016/j.neuroimage.2007.04.043.

Sundram, F., Deeley, Q., Sarkar, S., Daly, E., Latham, R., Craig, M., Murphy, D.G., 2012. White matter microstructural abnormalities in the frontal lobe of adults with antisocial personality disorder. Cortex 48 (2), 216–229. http://dx.doi.org/10.1016/j.cortex.2011.06.005.

Sylvers, P.D., Brennan, P.A., Lilienfeld, S.O., 2011. Psychopathic traits and preattentive threat processing in children: a novel test of the fearlessness hypothesis. Psychological Science 22 (10), 1280–1287. http://dx.doi.org/10.1177/0956797611420730.

Turiel, E., 1983. The Development of Social Knowledge: Morality and Social Convention. Cambridge University Press, Cambridge, England.

Viding, E., Blair, R.J., Moffitt, T.E., Plomin, R., 2005. Evidence for substantial genetic risk for psychopathy in 7-year-olds. Journal of Child Psychology and Psychiatry, and Allied Disciplines 46 (6), 592–597. http://dx.doi.org/10.1111/j.1469-7610.2004.00393.x pii:JCPP393.

Viding, E., McCrory, E.J., 2012. Genetic and neurocognitive contributions to the development of psychopathy. Development and Psychopathology 24 (3), 969–983. http://dx.doi.org/10.1017/s095457941200048x.

Viding, E., Sebastian, C.L., Dadds, M.R., Lockwood, P.L., Cecil, C.A., De Brito, S.A., McCrory, E.J., 2012. Amygdala response to preattentive masked fear in children with conduct problems: the role of callous-unemotional traits. American Journal of Psychiatry 169 (10), 1109–1116. http://dx.doi.org/10.1176/appi.ajp.2012.12020191.

Walters, G.D., 2003. Predicting institutional adjustment and recidivism with the psychopathy checklist factor scores: a metaanalysis. Law and Human Behavior 27, 541–558.

White, S.F., Frick, P.J., Lawing, K., Bauer, D., 2013a. Callous-unemotional traits and response to functional family therapy in adolescent offenders. Behavioral Sciences & the Law 31 (2), 271–285. http://dx.doi.org/10.1002/bsl.2041.

White, S.F., Marsh, A.A., Fowler, K.A., Schechter, J.C., Adalio, C., Pope, K., Blair, R.J., 2012. Reduced amygdala response in youths with disruptive behavior disorders and psychopathic traits: Decreased emotional response versus increased top–down attention to nonemotional features. American Journal of Psychiatry. http://dx.doi.org/10.1176/appi.ajp.2012.11081270.

White, S.F., Pope, K., Sinclair, S., Fowler, K.A., Brislin, S.J., Williams, W.C., Blair, R.J., 2013b. Disrupted expected value and prediction error signaling in youths with disruptive behavior disorders during a passive avoidance task. American Journal of Psychiatry 170 (3), 315–323. http://dx.doi.org/10.1176/appi.ajp.2012.12060840.

Wootton, J.M., Frick, P.J., Shelton, K.K., Silverthorn, P., 1997. Ineffective parenting and childhood conduct problems: the moderating role of callous-unemotional traits. Journal of Consulting and Clinical Psychology 65 (2), 301–308.

Yang, Y., Raine, A., Colletti, P., Toga, A.W., Narr, K.L., 2009. Abnormal temporal and prefrontal cortical gray matter thinning in psychopaths. Molecular Psychiatry 14, 561–562.

Yang, Y., Raine, A., Colletti, P., Toga, A.W., Narr, K.L., 2010. Morphological alterations in the prefrontal cortex and the amygdala in unsuccessful psychopaths. Journal of Abnormal Psychology 119 (3), 546–554. http://dx.doi.org/10.1037/a0019611.

Yeh, M.T., Chen, P., Raine, A., Baker, L.A., Jacobson, K.C., 2011. Child psychopathic traits moderate relationships between parental affect and child aggression. Journal of the American Academy of Child and Adolescent Psychiatry 50 (10), 1054–1064. http://dx.doi.org/10.1016/j.jaac.2011.06.013.

Young, L., Koenigs, M., Kruepke, M., Newman, J.P., 2012. Psychopathy increases perceived moral permissibility of accidents. Journal of Abnormal Psychology 121 (3), 659–667. http://dx.doi.org/10.1037/a0027489.

Psychopharmacotherapy, Costs of

J Fritze, Association of Private Health Insurers, Cologne, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12428–12434, © 2001, Elsevier Ltd.

Abstract

Treatment with drugs is an important aspect of the therapeutic interaction and management in all medical conditions including psychiatric disorders. Many psychiatric disorders take on a chronic, long-term course which produce significant economic impact on the individual as well as the society. It is important for clinicians to be aware of the economic aspects of drug prescription. With increasing longevity there is every likelihood of co-morbidity which will lead to increased burden. In various health care systems thus there is also likely to be a competition for increased resources. In order to avoid arbitrariness clinical guidelines can enable and support the clinicians but it is also important that resource allocation is made on a scientific basis which has been established by pharmacoeconomics. Development of atypical neuroleptics, selective antidepressants and acetylcholinesterase inhibitors has produced major additions to psychopharmacotherapy. It is important that clinicians and policymakers alike are aware of economic costs and aspects of medications being prescribed and the cost–benefit ratio.

Introduction

Psychopharmacotherapy is one of the mainstays of the treatment of psychiatric disorders. These are highly prevalent and frequently take a chronic or chronic remittent course, thus having a high economic impact. Whereas the demand for medical services is increasing, the growth of the resources allocated to health care is increasingly restricted. Thus, diseases and the patients suffering from them compete for the resources. In order to avoid arbitrariness, resource allocation needs a scientific basis which has been established by pharmacoeconomics. Major breakthroughs have been achieved in psychopharmacotherapy particularly by the introduction of atypical neuroleptics, selective antidepressants, and acetylcholinesterase inhibitors. Although there is a need for more research, the pharmacoeconomic analyses available suggest that the new drugs may at least be cost neutral or even cost saving despite their higher acquisition costs, essentially by reducing institutionalization. Nevertheless, the access of patients to the new therapeutic chances is restricted on economic grounds although this varies internationally.

Psychopharmacotherapy is one of the mainstays of the treatment of psychiatric disorders besides social therapy and psychotherapy. Psychiatric disorders essentially comprise the affective disorders (depression, bipolar, i.e., manic-depressive disorder), the schizophrenias, the anxiety disorders (panic disorder, generalized anxiety, post-traumatic stress disorder, obsessive–compulsive disorder, social phobia), the eating disorders (bulimia and anorexia nervosa), the substance abuse and dependence disorders including alcoholism, the personality disorders, and the dementias (with Alzheimer-type dementia being the cause in about 70% of cases). They are highly prevalent (Table 1) and frequently take a chronic or chronic remittent course. Therefore, it is easily understandable that the total costs of psychiatric disorders compare very well with the costs of the leading somatic illnesses (Table 2). The prevalences of depression and of the eating disorders are increasing. The WHO burden of illness study (Murray and Lopez, 1997) suggests that depression will be the leading

Table 1 Annual rate of behavioral disorders in the USA (adapted from NIMH)

Disorder	*Percent afflicted*
Anxiety disorders (all)	13
Depressive disorders (all)	10
Alcohol abuse/dependence	7
Drug abuse/dependence	3
Severe cognitive impairment	3
Antisocial personality	2
Schizophrenia	1
Somatization	0.2

Table 2 Estimated annual total costs of some psychiatric and somatic diseases (American Diabetes Association, 1996, American Heart Association, 1998, Davidson et al., 1999; Greenberg et al., 1993, 1999; Whetten-Goldstein et al., 1998)

Disease	*Estimated total cost ($US billion)*	*Year considered*
Schizophrenia	33	1991
Depression	43.7	1990
Bipolar (manic-depressive) disorder	25.6	1991
Alcoholism	98.6	1991
Anxiety disorders	42.3	1990
Alzheimer's disease	87.6	1991
Coronary heart disease	43	1990
Coronary heart disease	99.8	1999
Cancer	104	1991
Acquired immunodeficiency syndrome (AIDS)	66	1991
Multiple sclerosis	100.8	1994
Diabetes mellitus	91.8	1992
Hypertensive disease	33.3	1999
Congestive heart failure	21	1999
Stroke	45.3	1999

cause of disability in the developed countries in 2020. Psychiatric disorders (except the dementias) manifest in early adulthood, thus limiting the abilities of patients for social achievement.

Psychopharmacotherapy essentially comprises seven classes of drugs:

1. neuroleptics for the treatment of psychotic states, especially the schizophrenias;
2. antidepressants for the treatment of depression of whatever origin as well as for the long-term treatment of anxiety disorders;
3. mood stabilizers (lithium, some anticonvulsants) for the short-term treatment of mania and for the long-term relapse and recurrence prevention of affective disorders;
4. minor tranquilizers (essentially benzodiazepines) for the short-term treatment of anxiety and agitation of whatever origin and of alcohol withdrawal syndromes;
5. hypnotics (essentially benzodiazepines and benzodiazepine-like drugs) for short-term treatment of insomnia;
6. anticraving compounds (bupropion for smoking cessation, acamprosate for alcohol abstinence, opiate antagonists for alcoholism and opiate dependence) which, however, are not yet approved in all countries; and
7. cognition enhancers (nootropics such as piracetam, ginkgo biloba) for the nonspecific treatment of dementias and acetylcholinesterase inhibitors specifically for Alzheimer-type dementia.

Concerning the eating disorders, only one drug, the antidepressant fluoxetine, is approved for supportive treatment of bulimia.

Psychopharmacotherapy competes with psychotherapy in the treatment of mild to moderate depression and the anxiety disorders. The relative merits of both alternatives are subject to debate which is fueled by a lack of scientific data, ideology, and economic interests of the health care providers involved. As a rule of thumb, pharmacotherapy and psychotherapy may be thought not to be alternatives but supplements whenever a psychiatric disorder is moderate to severe, although a synergistic action remains to be definitely proved.

Life expectancy increased dramatically in the twentieth century, and this was accompanied by an increase in the prevalence of age-associated diseases. In the field of psychiatry these are the dementias. Thus, although the number of healthy life years is generally increasing, the prevalence of chronic diseases is increasing in parallel. Consequently, the demand for health care services is increasing. Costs for drug therapy are increasing overproportionately because drug development has become increasingly complicated and costly. The mean investment necessary for developing a single new chemical entity to approval amounts to about US$500 million.

Worldwide, health care is one of the most rapidly growing markets with positive impacts on employment. Nevertheless, governments internationally tend to restrict the growth because at least part of it is publicly funded. The fundamental but unanswered question is what proportion of the gross domestic product society is willing to invest in health care. Obviously, however, most people value health and thus health care very highly. Thus, disorders, i.e., the patients suffering from them, compete for limited resources. Economic criteria have become a third dimension besides efficacy and effectiveness as well as tolerability and safety when selecting between treatment options. It is easily conceivable that psychiatric disorders, faced with persistent stigma, have a hard standing in this competition.

Pharmacoeconomics, a young scientific discipline, has developed a sound scientific basis for the evaluation of the economic dimension (Gold et al., 1996). This is indispensable in the interests of an ethically justifiable allocation of resources. The question is whether a new, more costly drug allows one to achieve better outcomes so that it is worth its higher costs. In principle, a new drug might even spare money in that the better outcome might help avoid costly, higher-level health care such as inpatient treatment and reduce indirect costs, e.g., by less absenteeism.

Basic Terms of Pharmacoeconomics

Pharmacoeconomics is the science of measuring *costs* and *outcomes* associated with the use of pharmaceuticals in health-care delivery. Its objective is to improve public health through rational decision making when selecting among alternative therapies, e.g., for formularies and their impact on costs and outcomes. Several methodological guidelines have been published (e.g., CCOHTA, 1994).

Outcomes are the possible end result of a particular treatment. Outcomes depend on the *effectiveness* of the treatment under the influence of all the confounding variables in real life. Therefore, costs have to be measured under real-life conditions. In contrast, the experimental *efficacy* trials are designed to eliminate as completely as possible all confounding influences. The outcome parameter chosen should have high everyday relevance, e.g., the rate of full remissions of depression.

Costs are the resources consumed by the illness and by its treatment. *Direct costs* are those directly related to treatment such as prescribed medication, self-medication, consulting the physician, blood tests, transportation, psychotherapy, occupational therapy, informal care by relatives, etc., as well as costs for hospital and nursing-home treatment. *Direct costs* include those for diagnosis and treatment of side effects and complications of the initial therapy. There is methodological controversy as to whether the costs of social support (welfare, unemployment benefit, disability pension, housing benefit, etc.) should be included in the direct costs. A complete evaluation of direct costs should consider the opportunity costs, i.e., the lost potential of the resources invested to provide more efficient benefits in other fields.

Indirect or *future costs* are those related to lost productivity due to illness, i.e., due to absenteeism, premature pensioning because of disability, and premature death (e.g., by suicide) as well as those related to health care consumption during the years of life gained owing to the intervention. Concerning psychiatric disorders, costs of violence and crime also have to be considered. *Intangible costs* comprise those attributable to impairments of the quality of life, e.g., of the patients themselves and their relatives and acquaintances involved in the care.

The economic impact of treatment extends into the future. If costs and benefits accrue during different periods, costs and

savings have to be discounted for the remaining life span of the patient sample in question. Discount rates are supposed to reflect the rates of return on private-sector investments and are generally set at 3–5%.

If the costs are low and the outcomes excellent the decision in favor of the treatment is obvious as is its rejection if the costs for poor outcomes are high. Therefore, pharmacoeconomic analyses make sense only when both the costs and the quality of treatment are high or low. The economic *efficiency* then is the ratio of costs divided by the outcomes. When allocating resources the drug yielding the most favorable outcomes at the lowest cost will be favored.

Types of Pharmacoeconomic Analysis

The basic requirement of a pharmacoeconomic analysis is the definition of the perspective, e.g., that of the patient, the health care provider, the health insurance as the payer, or the societal perspective of government. The categories of costs to be considered depend on this perspective. For example, from a provider's perspective, the value of an antidepressant drug under capitation may be expressed in terms of cost per reduction in some depression scale score. From the patient's perspective, the value may be in cost per full treatment response with least side effects. From the societal perspective, the value may be in cost per patient returning to continuous employment at or above the level of previous education. Simply to compare the daily costs of medication does not fulfill the criteria of pharmacoeconomic analysis.

The principle types of analysis include cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis and cost–benefit analysis. *Cost-minimization* analysis assumes that the interventions in question yield identical outcomes; this is rarely valid. *Cost-effectiveness* analysis compares the costs of alternative treatments in terms of a specific health outcome such as life-years gained, e.g., by reducing mortality from suicide, or responder rates. *Cost-utility* analysis (CUA) adjusts the life-years gained by some measure for quality of life. Thus, the results of CUA are typically expressed as costs per quality adjusted life years (QALYs). Quality of life is assessed either by some rating scale for health-related quality of life or, more typically, by determining the preferences of patients suffering from specific diseases for one or the other disease state as opposed to perfect health where perfect health is set as '1' and death as '0.' In *cost–benefit* analysis, the QALYs gained are transformed into monetary terms using the human capital approach or willingness-to-pay approach.

As there is no one-to-one relationship between therapeutic intervention and outcome and its financial sequelae, the impact of these uncertainties has to be estimated. This is done by *sensitivity analysis*, where key variables are systematically modified and the economic consequences calculated yielding best- versus worst-case scenarios.

Pharmacoeconomic Study Designs

There are two principle approaches to pharmacoeconomic analysis: the clinical trial approach (experimental or observational), and the modeling or systems approach. The former means gathering cost data (retrospectively or, preferably, prospectively) in clinical trials. The latter '*in vitro*' approach applies decision-tree analysis to data available from published literature and databases.

Current guidelines favor the clinical trial approach as the 'gold standard.' Modeling, however, can overcome limitations implicit in clinical trials. These include the selection bias concerning the patients recruited in randomized, strictly controlled, and therefore artificial efficacy trials, thus limiting generalizability. Furthermore, their observation period rarely exceeds a few months whereas economic consequences may extend far into the future. Efficacy trials necessarily are underpowered for economic analysis because sample size estimation is based on the critical difference and variability of the efficacy parameter. Obviously, the critical difference and variability of economic data will fundamentally differ. Clinical trials using the mirror image design, i.e., comparing costs and benefits in a defined period before and after administering a specific drug, are of limited value because at least in psychiatric disorders there is a spontaneous decrease of the costs per outcome during the course of illness irrespective of the mode of treatment.

The putatively best experimental design is a prospective, randomized, 'naturalistic,' i.e., non-blind, adequately powered clinical trial. Both principle approaches, the clinical trial approach and modeling, have their weaknesses and strengths so that they are not really exclusive alternatives but complementary.

Friedberg et al. (1999) found that studies sponsored by drug companies were less likely to report unfavorable results than studies funded by nonprofit organizations. This finding could be due to various sources of bias which, however, if present, still have to be identified.

The organization of health-care supply, the health-care utilization patterns, and the prices differ considerably between and even within cultures. Therefore, the results of pharmacoeconomic analyses cannot simply be transferred from one health care market to another.

Pharmacoeconomic Findings with Neuroleptics

Since the 1990s new neuroleptics have been introduced (amisulpride, clozapine, olanzapine, quetiapine, risperidone, sertindole, zotepine), although not yet approved in all countries. These compounds are called 'atypical' because they have a lower incidence of extrapyramidal side effects (e.g., Parkinsonism) than conventional neuroleptics (e.g., haloperidol, fluphenazine). Moreover, they possibly have superior efficacy concerning cognitive dysfunction and so-called negative symptoms as well as quality of life in schizophrenia. These advantages might improve compliance, which is most important in long-term relapse and recurrence prevention. The costs of daily treatment with atypical neuroleptics exceed those of conventional neuroleptics by a factor of 2–8 depending on the compound and the country considered. On the other hand, neuroleptic medication costs generally make only about 4–8% of direct costs.

The pharmacoeconomic analyses actually available for neuroleptics are limited to amisulpride, clozapine, olanzapine,

risperidone, and sertindole in schizophrenia. The majority of studies address clozapine, olanzapine, and risperidone. The costs are mostly considered from the payer's perspective, i.e., addressing direct costs. Most are cost-effectiveness analyses, some cost-utility analyses. The majority follow the modeling approach. The economic analyses based on clinical efficacy trials have the limitations discussed above, especially those restricting the analysis to study completers. This is far from the naturalistic, real-world approach necessary. There is only one randomized, naturalistic, adequately powered trial comparing clozapine with conventional neuroleptics, where clozapine significantly reduced the frequency and duration of inpatient treatment; the cost data are to be reported. All in all, the current evidence suggests that clozapine is cost-effective for neuroleptic-refractory schizophrenia and that olanzapine and risperidone therapy may be cost-neutral despite higher medication costs (Revicki, 1999). Nevertheless, the access of patients to the new therapeutic chances is restricted on economic grounds, although this varies internationally.

Pharmacoeconomic Findings with Antidepressants

Since the 1980s, a number of new antidepressants have been introduced (selective serotonin reuptake inhibitors (SSRI): citalopram, fluoxetine, fluvoxamine, paroxetine, sertraline; the selective norepinephrine reuptake inhibitor reboxetine; the serotonin–norepinephrine reuptake inhibitor venlafaxine; the norepinephrine auto- and heteroreceptor blocker mirtazapine; the reversible and selective inhibitor of monoamine oxidase moclobemide), although not yet approved in all countries. Their efficacy is comparable to that of traditional tricyclic antidepressants (TCAs such as amitriptyline, imipramine, desipramine, nortriptyline), but owing to selectivity their tolerability and safety are superior. Safety is of specific concern in view of the suicide risk, although antidepressants are involved only in the minority (<10%) of suicides. These advantages help reduce drop-out rates although to a limited degree (about 25% versus 29%). Nevertheless, this might be particularly relevant in long-term recurrence prevention which is necessary in about 60% of patients in whom antidepressants are indicated. There is a dramatic underrecognition and undertreatment worldwide (Davidson et al., 1999; Tylee et al., 1999). Undertreatment causes a doubling of direct costs due to increased utilization of other medical services (Manning and Wells, 1995) and increases indirect costs due to absenteeism. The advantages of innovative antidepressants might increase acceptability and thus utilization of antidepressants. The costs of daily treatment with selective antidepressants, however, exceed those of TCAs by a factor of 2–20 depending on the compound and the country considered. On the other hand, antidepressant medication costs generally constitute only about 3–6% of direct costs.

Pharmacoeconomic analyses published cover all innovative antidepressants except reboxetine. All but one (Gould et al., 1995) are restricted to depression. The costs are mostly considered from the payer's perspective. Most are cost-effectiveness analyses, only a few cost-utility and cost–benefit analyses. There are no randomized, open-label, naturalistic, adequately powered trials but some studies analyzed large, complete samples of HMO participants. By far the majority follow the modeling approach. There are some outstanding analyses in which classes of compounds are compared (e.g., CCOHTA, 1997; Casciano et al., 2000). Overall, the current evidence suggests that extended-release venlafaxine might be more cost-effective than SSRIs and that the cost-effectiveness of SSRIs might equal or even exceed that of TCAs, despite higher medication costs (Conner et al., 1999; Casciano et al., 2000). However, more research is needed before this can be taken for granted. Accordingly, there is debate as to whether the access of patients to innovative antidepressants should be restricted or whether they should be recommended as first-choice drugs (CCOHTA, 1997).

The limited evidence available suggests that psychotherapy is more likely to be cost-effective the more severely afflicted the patients are (Gabbard et al., 1997). The cost of psychotherapy is about 3–6 times higher than that of antidepressant drug therapy. One modeling study (Antonuccio et al., 1997) suggests, however, that cognitive behavioral therapy (CBT) might be more cost-efficient than fluoxetine treatment of depression owing to lower recurrence rates after psychotherapy. In a CUA based on a randomized controlled trial (Lave et al., 1998), nortriptyline was slightly superior to interpersonal psychotherapy. One meta-analysis (Gould et al., 1995) suggests that CBT is superior to imipramine and cost-effective in panic disorder. Economic analyses addressing the relative merits of pharmacotherapy and psychotherapy are completely lacking for prominent diseases such as obsessive–compulsive disorder, generalized anxiety, and social phobia.

Pharmacoeconomic Findings with Mood Stabilizers

Mood stabilizers by definition reduce the risk of relapses and recurrences and thus putatively hospitalization. Therefore, they may be expected to reduce direct costs (about 60% for hospitalization). Lithium is the only drug reducing mortality essentially by reducing the suicide rate (Ahrens, 1997). Only a few pharmacoeconomic analyses have been published. Lithium was more cost-effective than carbamazepine from the payer's perspective in a modeling approach based on retrospectively gathered data from a randomized controlled clinical trial (Dardennes et al., 1999). Valproate was more cost-effective than lithium over 1 year from the payer's perspective in a modeling approach (Keck et al., 1996); the prophylactic efficacy of valproate, however, remains to be established.

Pharmacoeconomic Findings with Tranquilizers and Hypnotics

There are almost no economic analyses published on tranquilizers or hypnotics. Yuen et al. (1997) found that hospital stay was longer and costs were higher in elderly people if sedative–hypnotic medications were prescribed, even if prescription adhered to established guidelines. Research is needed especially to clarify whether the higher costs of the newer hypnotics (zaleplon, zolpidem, zopiclone) compared with benzodiazepines are economically justified.

Pharmacoeconomic Findings with Anticraving Drugs

Long-term acamprosate adjunctive to conventional therapy was found to be cost-effective in the treatment of alcoholism in the German setting in a single, retrospective modeling study (Schädlich and Brecht, 1998). Opiate antagonists (Howard et al., 1996), opiate agonist (methadone) substitution, and bupropion have not been evaluated economically.

Pharmacoeconomic Findings with Cognitive Enhancers

There are no economic analyses published on nootropics. The acetylcholinesterase inhibitors donepezil (Foster and Plosker, 1999) and rivastigmine (Fenn and Gray, 1999) seem to be at least cost-neutral, especially by delaying the need for institutionalization. The current evidence is restricted, however, to a few observational studies and modeling analyses using the cost-effectiveness approach from the payer's perspective. Economic analyses on galanthamine, an acetylcholinesterase inhibitor and nicotinic agonist expected to be approved, remain to be published.

Conclusions and Some Precautions

The innovative neuroleptics and antidepressants not only have the potential for a balance of costs despite higher costs of medication, but might even save direct costs essentially by reducing the risk of institutionalization owing to lower drop-out rates in comparison with traditional compounds. This conclusion, however, is based essentially on decision analyses, i.e., modeling. More research, particularly large-scale, randomized trials under naturalistic conditions, is needed. From a systemic perspective, the potential cost savings can only be realized if the institutions concerned, e.g., hospitals, reduce their capacities accordingly. About 80% of the costs of hospitals are common costs, mostly wages and salaries. Whereas in a free health-care market the savings might be realized by market forces, in systems bound in economic welfare a political will is needed to enforce the inevitable dismissal of staff.

See also: Antidepressant Drugs; Health Economics; Placebo Studies (Double-Blind Studies); Prophylaxis in Psychiatry; Psychiatric Care, Economics of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research.

Bibliography

Ahrens, B., 1997. Mortality studies and the effectiveness of drugs in long-term treatment. Pharmacopsychiatry 30 (Suppl. 1), 57–61.

Antonuccio, D.O., Thomas, M., Danton, W.G., 1997. A cost-effectiveness analysis of cognitive behavior therapy and fluoxetine (Prozac) in the treatment of depression. Behavior Therapy 28, 187–210.

Bacaltchuk, J., Trefiglio, R.P., Oliveira, I.R., Hay, P., Lima, M.S., Mari, J.J., 2000. Combination of antidepressants and psychological treatments for bulimia nervosa: a systematic review. Acta Psychiatrica Scandinavica 101, 256–264.

Casciano, R., Arikian, S.R., Tarride, J.-E., Casciano, J., Doyle, J.J., 2000. Antidepressant selection for major depressive disorder: the budgetary impact on managed care. Drug Benefit Trends 12, 6–16.

CCOHTA., 1994. Canadian Coordinating Office for Health Technology Assessment 1994. Guidelines for the Economic Evaluation of Pharmaceuticals. Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Canada. Ottawa, ON.

CCOHTA., 1997. Canadian Coordinating Office for Health Technology Assessment 1997. Selective Serotonin Reuptake Inhibitors (SSRIs) for Major Depression. Part II. The Cost-effectiveness of SSRIs in Treatment of Depression. Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Canada. Ottawa, ON.

Conner, T.M., Crismon, M.L., Still, D.J., 1999. A critical review of selected pharmacoeconomic analyses of antidepressant therapy. Annals of Pharmacotherapy 33, 364–372.

Dardennes, R., Lafuma, A., Watkins, S., 1999. Prophylactic treatment of mood disorders: a cost-effectiveness analysis comparing lithium and carbamazepine. Encephale 25, 391–400.

Davidson, J.R.T., Meltzer-Brody, S.E., Goldberg, D., Wittchen, H.-U., Magruder, K.M., Lecrubier, Y., Ballenger, J.C., Tylee, A., Schulberg, H.C., Nutt, D.J., 1999. The underrecognition and undertreatment of depression: what is the breadth and depth of the problem? Discussion. Journal of Clinical Psychiatry 60 (Suppl. 7), 10–11.

Fenn, P., Gray, A., 1999. Estimating long term cost savings from treatment of Alzheimer's disease: a modelling approach. PharmacoEconomics 16, 165–174.

Foster, R.H., Plosker, G.L., 1999. Donepezil: pharmacoeconomic implications of therapy. PharmacoEconomics 16, 99–114.

Friedberg, M., Saffran, B., Stinson, T.J., Nelson, W., Bennett, C.L., 1999. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. Journal of the American Medical Association 282, 1453–1457.

Gabbard, G.O., Lazar, S.G., Hornberger, J., Spiegel, D., 1997. The economic impact of psychotherapy: a review. American Journal of Psychiatry 154, 147–155.

Gold, M.R., Russell, L.B., Siegel, J.E., Weinstein, M.C. (Eds.), 1996. Cost-effectiveness in Health and Medicine. Oxford University Press, Oxford, UK.

Gould, R.A., Otto, M.W., Pollack, M.H., 1995. A meta-analysis of treatment outcome for panic disorder. Clinical Psychological Review 15, 819–844.

Greenberg, P.E., Stiglin, L.E., Finkelstein, S.N., Berndt, E.R., 1993. The economic burden of depression in 1990. Journal of Clinical Psychiatry 54, 405–418.

Greenberg, P.E., Sisitsky, T., Kessler, R.C., Finkelstein, S.N., Berndt, E.R., Davidson, J.R.T., Ballenger, J.C., Fyer, A.J., 1999. The economic burden of anxiety disorders in the 1990s. Journal of Clinical Psychiatry 60, 427–435.

Howard, M.O., McGuffin, R.W., Saxon, A.J., Sloan, K.L., Walker, R.D., 1996. Clinical issues related to the costs of alcoholism. PharmacoEconomics 9, 134–145.

Keck Jr, P.E., Nabulsi, A.A., Taylor, J.L., Henke, C.J., Chmiel, J.J., Stanton, S.P., Bennett, J.A., 1996. A pharmacoeconomic model of divalproex vs. lithium in the acute and prophylactic treatment of bipolar I disorder. Journal of Clinical Psychiatry 57, 213–222.

Lave, J.R., Frank, R.G., Schulberg, H.C., Kamlet, M.S., 1998. Cost-effectiveness of treatments for major depression in primary care practice. Archives of General Psychiatry 55, 645–651.

Manning, W., Wells, K.B., 1995. How can care for depression become more cost-effective? Journal of the American Medical Association 274, 1931–1934.

Murray, C.J.L., Lopez, A.D., 1997. Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. Lancet 349, 1498–1504.

Revicki, D.A., 1999. Pharmacoeconomic studies of atypical antipsychotic drugs for the treatment of schizophrenia. Schizophrenia Research 35 (Suppl.), S101–S109.

Schädlich, P.K., Brecht, J.G., 1998. The cost effectiveness of acamprosate in the treatment of alcoholism in Germany: economic evaluation of the prevention of relapse with acamprosate in the management of alcoholism (PRAMA) study. PharmacoEconomics 13, 719–730.

Tylee, A., Gastpar, M., Lepine, J.-P., Mendlewicz, J., 1999. DEPRES II (Depression Research in European Society II): a patient survey of the symptoms, disability and current management of depression in the community. International Clinical Psychopharmacology 14, 139–151.

Whetten-Goldstein, K., Sloan, F.A., Goldstein, L.B., Kulas, E.D., 1998. A comprehensive assessment of the cost of multiple sclerosis in the United States. Multiple Sclerosis 4, 419–425.

Yuen, E.J., Zisselman, M.H., Louis, D.Z., Rovner, B.W., 1997. Sedative–hypnotic use by the elderly: effects on hospital length of stay and costs. Journal of Mental Health Administration 24, 90–97.

Psychopharmacotherapy: Side Effects

Nitasha Shetty and Handan Gunduz-Bruce, Yale University School of Medicine, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N.R. Schooler, H. Gunduz, volume 18, pp. 12434–12439, © 2001, Elsevier Ltd.

Abstract

In parallel to advances in neuroscience, more treatment options for psychiatric disorders have become available. Since the original preparation of this article, for example, atypical antipsychotics have been in wide use for the treatment of psychotic, even for mood symptoms. Significant challenges remain though, as atypical antipsychotics are associated with side effects that continue to negatively impact the medical health of majority of patients. Special patient populations such as children and the elderly can be more vulnerable to side effects. In this article, we have included subsections for these populations with further references so that more detailed information can be gathered by the interested reader. This article is written with an intention to educate all clinicians, including those in training, involved in the care of psychiatric patients. Because comprehensive care includes several disciplines addressing the different needs of patients consistent with a psychosocial model, it is imperative for not only patients, but also support staff to be knowledgeable of these basic facts. As detailed in the article, side effects can often mimic underlying pathology and easily be confused the actual symptoms of the illness. Therefore, a sound knowledge base, careful assessment and vigilance are necessary for the assessment of side effects of psychoactive medications. For the psychiatrist in clinical practice, frequent monitoring of side effects is critical. In addition, development of psychoactive medications with favorable side-effect profile needs to be priority for the industry.

Psychopharmacology studies drug-induced alterations in the cognitive function and behavior of humans and aims to develop treatment strategies for neurological and psychiatric disorders. Like all drugs, psychoactive drugs are associated with side effects (adverse effects) which may be generated both centrally and peripherally. The spectrum of drug effects originates largely from their effects on multiple neurotransmitter systems, and the interactions of these systems with each other and other sites of action in the central nervous system. One of the strongest predictors of treatment noncompliance is early onset and persistence of side effects. The strongest determinant of the interindividual variation in the experience of side effects is believed to be in the genetic make up. The study of pharmacogenetics targets drug development that will be both efficacious and well tolerated for each individual. Management of side effects includes a baseline assessment, careful choice of medication, patient and family education, minimization of the number of drugs (to decrease potential side effects and to avoid drug interactions), single bedtime dosing, the availability of clinicians for frequent monitoring, adjustment in dosage, counteractive drug treatment, and continuous assessment. In this article, major behavioral side effects of psychoactive drugs such as sedative, cognitive, and motor effects are examined. These factors play a critical role in the quality of life of patients.

Psychopharmacotherapy refers to the treatment implications of the science of psychopharmacology. Psychopharmacology is the study of drug-induced alterations in the cognitive function and behavior of humans that aims to develop treatment strategies for neurological and psychiatric disorders. Rapid technological progress has led to an abundance of scientific techniques in neurochemical and animal studies on the interaction between drug action and behavior (Feldman et al., 1997).

Most psychoactive drugs exert their effects first by interfering with the communication between nerve cells (neurons). At the molecular level this occurs by their interaction with receptors of neurotransmitters located on the surface or inside the neuron. This step initiates a cascade of events that eventually leads to alterations in the neurotransmitter and/or hormonal systems. Nonspecific drug action at any level in this process results in side effects. The term side effect implies an undesirable (adverse) drug effect; a particular side effect of a drug for a given condition may be a desirable one for another condition. For instance, benzodiazepines may be prescribed for anxiety disorders, but as a side effect they cause sedation. On the other hand, they may also be prescribed for the treatment of insomnia as a primary indication. Indeed, the psychoactive potential of several drugs was discovered as a result of unexpected but beneficial side effects in psychiatric patients who were taking these drugs for other therapies (see below). The spectrum of drug effects originates largely from their effects on multiple neurotransmitter systems, and the interactions of these systems with each other and other sites of action in the central nervous system.

At times it is a challenge in the practice of psychiatry to be able to determine if a given behavior is a symptom of an illness or a side effect of drug treatment, as the latter may mimic the former. For instance, untreated patients with schizophrenia frequently display decreased spontaneity, lack of motivation, and social isolation. These symptoms resemble some of the extrapyramidal system (EPS) side effects induced by conventional antipsychotics. A practical management of this potential difficulty involves careful assessments of the patient's functions e.g., sleep patterns, daily living routines, and sexual function prior to treatment.

Historical Perspective

Treatment with substances that alter mood, thinking, and cognition date back to prehistoric times (Spiegel, 1996).

Various psychoactive substances have been used in religious ceremonies, rituals, and for healing of pain.

Alcohol is the oldest drug used for its calming and sedative effects. The disinhibition caused by alcohol is described in the first book of the Old Testament. Apart from consumption in social settings, alcohol was introduced for medical use in the nineteenth century as a general anesthetic. The narrow therapeutic index (the ratio of toxic over desired dose) of alcohol limited its usefulness for this purpose.

Opium (*papaver somniferum*), *hellebore*, and *rauwolfia* roots are among the oldest compounds that were extracted from plants in prehistoric and ancient times. Morphine was isolated from opium in the beginning of the nineteenth century and has been available as an analgesic with considerable addiction potential.

In ancient times, different forms of hellebore root were used as emetics, laxatives, and for mental conditions such as mania and melancholy. Risk of seizures limited its use. *Rauwolfia serpentina* roots were used by Indians to treat snake bites and also in some mental conditions. Reserpine, a rauvolfia alkaloid was noted to have a calming effect in patients with schizophrenia. Because reserpine induces severe depression as a side effect, its use as an antipsychotic has been abandoned.

Cocaine is an alkaloid derived from the shrub *Erythroxylon coca*; it originated in South America where it was used for its stimulating effects. It was isolated in the nineteenth century and is still used as a local anesthetic. Paranoid delusions, hallucinations, seizures, and dependency are some of the adverse effects of systemic cocaine use.

The discovery of psychoactive medications has resulted from the side-effect profiles of drugs used in medicine. For instance, the antipsychotic chlorpromazine was initially used to help patients with presurgical preparation and as an adjuvant to anesthetics. Its sedative and calming effects easily found its way to treatment in psychiatric disorders as a primary indication soon after its antipsychotic properties were discovered. Similarly, iproniazid (an antidepressant of the monoamine oxidase inhibitor – MAOI type) was originally developed as a drug for the treatment of tuberculosis. In clinical drug trials, patients treated with this drug were observed to become stimulated and cheerful. This observation led to the investigation of the antidepressant effect of iproniazid.

Genetic Factors and Side-Effect Vulnerability

Side effects are not uniformly observed in everyone. While some patients never develop a certain side effect, others are affected from day one of treatment. The strongest determinant of this interindividual variation is believed to be in the genetic complexion. Variations in the sequence of the DNA called polymorphisms are believed to play a role in expression of a wide spectrum of reactions to drugs: from full response to nonresponse; minimal to severe side effects. The completion of the Human Genome Project may aid in the identification of patient subgroups according to their drug response patterns. For instance, certain variations in the gene that codes for the enzyme CYP2D6 lead to slow metabolism of some drugs such as tricyclic antidepressants, therefore these patients run higher levels of drugs and are more vulnerable to side effects (Schatzberg and Nemeroff, 1998). The application of pharmacogenetics will be the identification of drugs that will be both efficacious and well tolerated for each individual according to his or her genetic makeup.

Assessment and Management of Side Effects

As compared with the rating scales used to assess the symptoms of diseases, there are relatively few rating scales for the assessment of side effects. The most comprehensive of these is the Systematic Assessment for Treatment Emergent Events which has been designed for use in clinical drug trials. Udvalg for Kliniske Undersogelser Side-effect Rating Scale is also comprehensive and was developed for use in clinical trials and in routine practice. The Rating Scale for Extrapyramidal Side Effects (Simpson–Angus Extrapyramidal System Scale), Barnes Akathisia Rating Scale, and Abnormal Involuntary Movement Scale were developed to measure the EPS side effects, akathisia, and dyskinetic movements, respectively, that are associated with antipsychotic drug use (American Psychiatric Association, 2000). Even though each side effect ultimately requires individual management, a comprehensive approach is vital for general management. Such an approach involves a baseline assessment, careful choice of medication, patient and family education, minimization of the number of drugs (to decrease potential side effects and to avoid drug interactions), single bedtime dosing, the availability of clinicians to discuss the side effects, adjustment in dosage, counteractive drug treatment, and continuous assessment.

Noncompliance

Among factors that play a role in treatment noncompliance, such as denial of the illness and lack of supervision, one of the strongest predictors is early onset and persistence of side effects. Noncompliance by itself is a predictor of relapse in patients with chronic mental disorders who are recommended to receive lifelong maintenance treatment. One of the studies in this area found that half of the schizophrenic patients took less medication than recommended, and the reluctance was associated with the side effects of antipsychotic drugs (Van Putten, 1974). On the positive side, many studies indicate improvement in treatment compliance with patient education. Particular approaches in patient education have been proposed that maximize the information conveyed while avoiding suggestibility (Balon, 1999).

Classification of Side Effects

One of the common characteristics of psychoactive drugs is their capability to cross the blood–brain barrier and reach the central nervous system. Consequently, side effects of psychoactive drugs may be generated both centrally and peripherally. Here, the distinction between drug action and drug effect is also relevant: drug action refers to alterations produced by the drug at the molecular level, whereas drug effect (as well as side effect) refers to the observed reactions of the organism. Below, we will examine the behavioral side effects of psychoactive

medications; a review of physiological systems is beyond the scope of this article and can be found elsewhere (Schatzberg and Nemeroff, 1998).

Autonomic

The autonomic nervous system controls homeostasis in the different functional systems in the body mostly through involuntarily mechanisms. Major neurotransmitter systems involve those of acetylcholine and norepinephrine (β or α receptors).

Autonomic side effects of psychoactive drugs are most commonly due to *anticholinergic effects*. Because acetylcholine is widely distributed in the central and peripheral nervous systems, adverse anticholinergic effects can be observed both behaviorally and systemically. Most antiparkinsonian (benztropine, trihexiphenidyl) and the majority of the antidepressant and antipsychotic drugs produce anticholinergic effects. Among the tricyclic antidepressants, tertiary amine tricyclics (amitriptyline, clomipramine, doxepin, imipramine, trimipramine) produce more anticholinergic side effects as compared with secondary amine tricyclics (desipramine, nortriptyline, protriptyline) and tetracyclic antidepressants (amoxapine, maprotiline). As compared with tricyclic antidepressants, anticholinergic side effects with selective serotonin reuptake inhibitors (SSRIs) are minimal. Conventional low-potency antipsychotics (drugs in this category, such as chlorpromazine, thioridazine, and mezoridazine require higher doses to produce equal efficacy than the high-potency antipsychotics) have significant anticholinergic properties as compared with high-potency antipsychotics (fluphenazine, haloperidol, perphenazine, pimozide, tiotixene or (thiothixene), trifluoperazine). Among the new generation antipsychotics, clozapine has the most potent anticholinergic effects, as well as having properties of acetylcholine release. Olanzapine also possesses anticholinergic effects, but less than clozapine. Risperidone, quetiapine, and ziprasidone have minimal anticholinergic activity. Peripheral anticholinergic effects include warm skin, flushing, dryness, dilated pupils (and blurred vision), increased intraocular pressure, dry mouth (if unrecognized this can cause dental problems, such as caries), tachycardia (increased heart rate), reduced acid secretion in the stomach, reduced pulmonary secretions, diminished bowel motility (constipation), urinary retention, and delayed or retrograde ejaculation.

Central anticholinergic effects of psychoactive drugs involve features such as memory impairment, confusion, disorientation, delirium, visual hallucinations, and worsening of preexisting psychotic symptoms. Combination of psychoactive drugs with anticholinergic effects, such as a tricyclic antidepressant, and a low-potency antipsychotic will highly increase the likelihood of an adverse reaction. Geriatric patients are particularly sensitive to anticholinergic side effects and deserve special consideration in the choice of psychoactive drugs and evaluation of a deteriorating course of their clinical situation (*see* Antidepressant Drugs).

Dependence

Dependence is an important component of the side-effect spectrum of some of the psychoactive drugs (barbiturates, benzodiazepines). The *Diagnostic and Statistical Manual of Mental Disorders, fourth edition* (soon to transition to the fifth edition) defines substance dependence as a maladaptive pattern of substance use, leading to clinically significant impairment or distress, as manifested by three or more of the following criteria: tolerance, withdrawal, consumption of large amounts or over a longer than intended time period, persistent desire to cut down, significant effort to obtain the substance, social, occupational sacrifice, and continued usage despite known hazards. It is of note that in this definition a biopsychosocial approach is undertaken. The treatment models for dependency are conducted in a multidisciplinary fashion that address physical, psychological, and environmental factors. Chronic drug use involves tolerance, physical dependence, and psychological dependence. Even though related, tolerance and physical dependence are different concepts:

Tolerance refers to the need of increased amount of a drug with repeated usage in order to achieve the same magnitude of effect. Interestingly, not all of the behavioral effects of drugs show similar tolerance patterns; in the case of benzodiazepines, for instance, tolerance develops for the sedative effects, but not for the antianxiety effects. Tolerance is a reversible condition; abstinence from the drug over a period of time will reduce the amount of drug needed to exert the particular effects. Cross-tolerance is a phenomenon where the tolerance for a particular drug effect operates for another drug effect as well. It is demonstrated by the diminished anticonvulsant effects of the barbiturates in people with chronic alcohol use. Both changes in cellular physiology and behavioral conditioning systems have been postulated as underlying mechanisms for the development of tolerance.

Physical dependence refers to the condition where withdrawal symptoms emerge upon discontinuation of the drug. Withdrawal symptoms are specific to different drugs and quite unpleasant, so that continued use is sought to avoid them. Most withdrawal symptoms are separate from the ones in the intoxication state and can be identified without difficulty.

Endocrine, Metabolic and Cardiovascular

Most antipsychotics elevate serum prolactin secondary to the blockade of dopamine. Increased prolactin may cause menstrual irregularities in females, gynecomastia (abnormal enlargement of the breast) in males, and galactorrhea (spontaneous milk flow) in both females and males. Amenorrhea (complete cessation of menses) occurs frequently in women taking antipsychotics. These changes persist with treatment and resolve with discontinuation. Galactorrhea and amenorrhea may also be observed with tricyclic antidepressants.

Hypothyroidism (decreased function of the thyroid gland) may be associated with long-term lithium therapy. The elderly and females are at greater risk for lithium-induced hypothyroidism. Weight gain is a common side effect of antipsychotics, particularly of new generations, cyclic antidepressants, MAOIs, and SSRIs. Studies suggest that increased weight gain confers a greater risk of insulin resistance, dyslipidemia, and hyperglycemia (Newcomer, 2005). Among the atypical antipsychotic agents, those associated with the greatest risk of weight gain are clozapine and olanzapine. Risperidone, quetiapine, amisulpride, and zotepine are associated with intermediate risk of

clinically significant weight gain. Ziprasidone and aripiprazole have the lowest risk of weight gain, and no known increased risk of diabetes or dyslipidemia. Among typical antipsychotic medications, phenothiazine derivatives (i.e., chlorpromazine, fluphenazine, perphenazine) are associated with dyslipidemia, whereas butyrophenone derivatives (i.e., haloperidol) are not (Newcomer, 2005). Weight changes are also seen with antidepressant medications. The antidepressants most likely to cause weight gain in both the acute and maintenance phases of treatment are amitriptyline and mirtazapine. Nortriptyline is associated with weight gain during the acute phase, but not maintenance treatment. Paroxetine is associated with weight loss in the acute phase, but overall weight gain is observed after long term use. Several other antidepressants are associated with weight loss in the acute phase, but only bupropion has demonstrated long term weight loss (Serretti and Mandelli, 2010). Stimulants (amphetamine, dextroamphetamine, methylphenidate), are associated with appetite suppression and weight reduction. Stimulants are also associated with adverse growth effects (height and weight suppression); however, these effects are minimal (i.e., 2.0–2.7 cm less growth in school age children after 3 years of medication) and can be compensated with drug holidays (Swanson et al., 2007). Cardiovascular disease results in sequelae such as heart attack, angina, and heart failure. Cardiovascular disease risk factors include hypertension, hyperlipidemia, diabetes, and obesity. The risk of having cardiovascular disease has been found to be elevated among those with schizophrenia as as compared with healthy controls (9.4 vs 7.0% in men and 6.3 vs 4.2% in women) (Goff et al., 2005). Metabolic side effects of medications, as well as lifestyle factors such as smoking, lack of exercise, and poor diet lead to a two to three times increased risk of cardiovascular disease mortality among those with schizophrenia than the general population (Newcomer, 2005). Even prior to initiating antipsychotic medications, people with schizophrenia demonstrate increased visceral adiposity and abnormal glucose regulation (Newcomer, 2005). This indicates that medication side effects may not be the sole contributor to increased metabolic risk in this population.

Cardiovascular disease is the most common cause of mortality among patients with schizophrenia, and patients with affective disorders also have higher rates of cardiovascular and cerebrovascular mortality (Newcomer, 2005). In addition to the burden of cardiovascular morbidity and mortality, metabolic side effects attributed to atypical antipsychotics also lead to increased health care costs and disability (Colombo et al., 2008; Chwastiak et al., 2009). Early intervention to modify risk factors and treat hypertension, diabetes, and dyslipidemia in this population is crucial. Current practice guidelines recommend baseline measurements of blood pressure, weight, BMI, waist circumference, lipoprotein profile, and glucose level upon initiation of an antipsychotic agent. Weight should be monitored at 4, 8, and 12 weeks following initiation or change in antipsychotic medication, and every 3 months thereafter. Glucose level, lipoprotein profile, and blood pressure should be followed up 3 months after antipsychotic initiation, and annually thereafter. If the lipoprotein profile is within normal limits, it can be monitored every 5 years (Consensus Development Conference Report, 2004). These parameters should be monitored more frequently for those at higher cardiovascular risk (Casey et al., 2004). Several studies have suggested that treatment with metformin is helpful in reducing weight gain and insulin resistance associated with atypical antipsychotic use. However, further long term studies are needed before implementing this strategy in clinical practice (Björkhem-Bergman et al., 2011; Wang et al., 2012).

Insomnia

Insomnia is the most common transient side effect of stimulants and MAOIs. It is also important to consider that underlying conditions may be linked with insomnia. In most cases, individuals develop tolerance to this side effect after the initial couple of weeks of treatment.

Cardiac

Antipsychotic medication use may increase an individual's risk of developing prolongation of the QT interval, and as a result, can in rare cases lead to ventricular tachyarrythmia and sudden cardiac death. Incidence rates of sudden cardiac death, while rare, are approximately twice as high among current users of both typical and atypical antipsychotic medications as among nonusers (18 events per 10 000 person-years, or less than 0.2%) (Ray et al., 2009). Risk increases with increased antipsychotic dose, older age, and male gender. Incidence-rate ratios increase from approximately 1.45 to 2.64 between those taking low doses of antipsychotic medication as compared with high doses (Ray et al., 2009). Individuals taking high doses of thioridazine have the greatest increased risk (incidence-rate ratio 5.05). Among those taking antipsychotics, the rate of sudden cardiac death in those 70–74 years of age is 10-fold higher than in those 30–34 years of age (4.7 vs 47.6 deaths per 10 000 person-years). Incidence rates are twice as high for men as for women (27.1 vs 12.9 deaths per 10 000 person-years).

Neurological/Behavioral

Cognitive Effects

Even though the motor and cognitive functions of the brain do not operate independently from each other, specific neuropsychological tests that capture primarily motor or cognitive functions have been applied to measure the effects of psychoactive drugs in these domains. While several studies comparing neurocognitive improvement with antipsychotic treatment suggested advantages with atypical agents over typical agents (Keefe et al., 2004; Velligan et al., 2003) more recent evidence indicates no difference (Keefe et al., 2007). When perphenazine was compared with equivalent doses of atypical antipsychotics (olanzapine, risperidone, and quetiapine), all groups demonstrated a small improvement in processing speed, reasoning, working memory, verbal memory, and attention. Most of the improvement occurred by 2 months of treatment. Prior evidence suggesting neurocognitive worsening with typical antipsychotics may have been due to use of higher doses of high potency medications (i.e., haloperidol), resulting in extrapyramidal side effects and need for co-administration of anticholinergic medications. Double-blind studies comparing the new generation antipsychotic clozapine and the conventional antipsychotic haloperidol revealed a significant

interaction on measures of executive/attention, verbal fluency, and visuospatial functions (Lee, 1999). Further analysis revealed that some of these effects resulted from worsening performance of the haloperidol-treated patients. Patients treated with clozapine showed an insignificant improvement (Buchanan et al., 1994). Anticholinergic doses have been associated with poor verbal learning, verbal fluency, and motor speed. The cognitive effects of antipsychotic medications are especially salient since schizophrenia is associated with cognitive deterioration over time. Differences in academic achievement between those who later develop schizophrenia as compared with their healthy peers are evident as early as the first grade, and the gap continues to widen by twelfth grade. There is also a substantial decline in cognitive ability associated with the initial psychotic episode (Bilder et al., 2006). Lithium may also interfere with memory and motor speed functions. Valproate, carbamazepine, and topiramate may reduce the speed of cognitive processing (Loring et al., 2011; Hessen et al., 2006). Benzodiazepines may cause slowed cognitive function, impaired memory, and behavioral disinhibition.

Motor Effects

Conventional antipsychotic drugs (haloperidol, fluphenazine), and to a lesser extent, new generation antipsychotic drugs (risperidone, olanzapine), some of the selective serotonin reuptake inhibitors (sertraline, fluoxetine), and cyclic antidepressants (amoxapine) affect motor functions of the EPS. Acute conventional antipsychotic treatment interferes with simple repetitive motor tasks such as finger tapping. These effects are related to the drugs' dopamine blocking properties in the basal ganglia. Lithium and valproic acid may cause hand tremors. Stimulant use may be associated with development or worsening of tics.

Acute motor effects of the conventional antipsychotics include parkinsonism, dystonic reactions, and akathisia. The major chronic effect of antipsychotic use is tardive dyskinesia. *Parkinsonism* is a word derived from Parkinson's Disease, which is characterized by tremor, rigidity, and bradykinesia (slowed movement). All three features are observed in susceptible patients in the initial phases of treatment. Tremor with a regular rate may be observed in the hands and feet, and is present at rest. Tremor around the mouth also occurs. Akinesia (absence of movement) of the facial muscles sometimes leads to lack of expression which is known as masked face. Patients walk slowly, sometimes with shuffling, short steps, and the normal arm swing is diminished. Rigidity of muscles can be felt on examination; cogwheeling (increased tone in all directions of movement and superimposed ratchet-like resistance) may be elicited. Handwriting is characteristically small and tremulous.

One of the most disturbing side effects of antipsychotic medications (mostly of the high-potency conventional ones and some of the new generation antipsychotics) is akathisia, a subjective experience of motor restlessness. Patients complain of not being able to sit still, they pace in hallways, constantly move their legs to help the discomfort. They describe it as "… I want to run a hundred miles … ." Significant akathisia may lead to insomnia because of having trouble lying in bed. It may be accompanied by a feeling of anxiety and agitation. Restlessness and agitation are frequently observed as part of psychosis in the absence of any drug treatment. Drug-induced akathisia therefore is very important to diagnose differentially for appropriate treatment. If it is mistaken as psychosis-related agitation, increased antipsychotic drug dose will only make it worse. Akathisia has been identified as a predictor of antipsychotic noncompliance. Several treatment options are available.

Acute dystonic reaction is another side effect of the conventional antipsychotics that requires immediate treatment. Dystonic reactions consist of abruptly developing muscle contractions which frequently involve the muscles of the face, tongue (protrusions, twisting), neck (spasmodic torticollis), jaw (trismus), eyes (oculogyric crisis, blepharospasm), and limbs. It is most common in young men in the initial phases of treatment. It is treatable with anticholinergic and antihistamine drugs. Prophylactic treatment with these drugs may increase the chances of compliance.

Tardive dyskinesia is a long-term side effect of conventional antipsychotics characterized by involuntary choreiform (random, brisk) or athetoid (slow, regular continuing twisting) movements of the head, face, tongue, limbs, and the trunk. These movements may range from mild and unrecognized to severe and incapacitating. Movements around the mouth are the most common. Dyskinesia may be exacerbated with stress and disappears with sleep. About 10–20% of patients treated with conventional antipsychotics for more than a year develop tardive dyskinesia. Children, women, the elderly, patients with brain damage or mood disorders constitute high-risk groups. While tardive dyskinesia is associated with antipsychotic use, 1–5% of patients with schizophrenia who lived before the antipsychotic era have been reported to exhibit similar dyskinetic movements. Thus, tardive dyskinesia may be related to the pathophysiology of schizophrenia. The new generation antipsychotics (clozapine, olanzapine, risperidone, quetiapine, ziprasidone) were previously believed to have significantly lower incidence of EPS side effects and tardive dyskinesia as as compared with the conventional antipsychotics (Correll et al., 2004). However, more recent studies find minimal difference between first and second generation antipsychotics in the incidence of EPS side effects or tardive dyskinesia over a period of 18 months at equivalent doses (Woods et al., 2010; Miller et al., 2008). It is suggested that earlier discrepancy was based on use of higher doses of first generation agents. Numerous animal studies have demonstrated that clozapine fails to induce EPS-related movements, and it is the only antipsychotic that might have a potential to prevent tardive dyskinesia from deteriorating (*see* Schizophrenia, Treatment of).

Sedation

Sedation is a common side effect of most antipsychotics, mood stabilizers, benzodiazepines, barbiturates, and antidepressants (tertiary amine cyclic antidepressants, MAOIs, trazodone, sometimes with other SSRIs, venlafaxine, and nefazodone). In particular clinical situations, such as severe agitation, sedation may be desirable, whereas in others such as a negative symptoms of schizophrenia (affective flattening, lack of spontaneous speech, avolition), or depression with hypersomnia, it would not be. The choice of drug therefore largely depends on the clinical necessity. In undesirable conditions (night time dosing, reduction in dose) switching to a less sedating medication is considered.

Seizure

Antipsychotic drugs, particularly the ones with low potency reduce the seizure threshold. In a patient with no history of seizure disorder, incidence is low and no prophylactic treatment is warranted. Patients with a history of seizure disorder deserve special consideration. Clozapine increases risk of seizures in a dose-related fashion. The frequency of seizures in those taking clozapine is 1.3% overall, with greater risk observed during the initial titration phase than during the maintenance phase (Pacia and Devinsky, 1994). Individuals receiving doses equal to or greater than 600 mg daily have a 1.9% risk of seizure at 6 months (Pacia and Devinsky, 1994). Anticonvulsants such as valproic acid can benefit the patient to continue treatment with clozapine.

Among the antidepressants, clomipramine and amoxapine may lower the seizure threshold while incidence of seizures with SSRIs and Serotonin and Norepinephrine Reuptake Inhibitors (SNRIs) is low, but present. Estimated seizure risk for these agents are 0.4–1.0% for tricyclic antidepressants, 0.26% for venlafaxine, 0.2% for fluoxetine, 0.15% for bupropion, 0.1% for paroxetine, and less than 0.1% for sertraline (Dunner et al., 1998). Tricyclic antidepressants, MAOIs, anticonvulsants, lithium, and antipsychotics may also induce myoclonic (irregular rapid muscle contraction) movements.

Hypomania/Mania

Even though it remains controversial, one of the important side effects of antidepressant treatment is the induction of hypomania/mania. In a patient with no history of mania, undergoing antidepressant treatment, emergence of manic features merits consideration of the possibility of drug-induced mania. In patients with bipolar disorder, on the other hand, the risk of mania induction with sole antidepressant (in the absence of a mood stabilizer) treatment is established (Truman et al., 2007).

Delirium

Delirium associated with psychotropic medication is most likely to occur in the elderly population and those with dementia. The psychotropic medication classes most likely to cause delirium are anticholinergics (such as benztropine and trihexylphenidyl), benzodiazepines, levodopa, and tricyclic antidepressants. Infrequently, worsening of psychosis, confusion, and agitation may occur with antipsychotic treatment. Most cases are, however, associated with those antipsychotics with greater anticholinergic properties such as clozapine and chlorpromazine, with polypharmacy, or in the elderly (Moore and O'Keefe, 1999). While several psychiatric medications have anticholinergic properties, many other commonly prescribed medications do as well, such as antiemetic, anticoagulant, cardiac, genitourinary, allergy, and muscle relaxant medications (Cancelli et al., 2009). The additive anticholinergic properties of multiple medications are more likely to induce delirium.

Neuroleptic Malignant Syndrome

This is a potentially life-threatening complication of the antipsychotic treatment characterized by muscle rigidity, hyperthermia, mental state changes (stupor, agitation, confusion), and autonomic dysfunction (blood pressure instability, increased pulse, respiration, sweating). The incidence is about 0.1% of patients exposed to conventional antipsychotics, and more likely to occur in men (approximately 70% of cases) (Troller et al., 2012). High-potency antipsychotics may have a higher risk than low-potency antipsychotics. Rhabdomyolysis (muscle necrosis) may occur that leads to myoglobinuria (passing a muscle-related protein – myoglobin – in urine) and subsequent renal failure. Immediate intervention is necessary. As opposed to conventional antipsychotics, NMS with atypical antipsychotics have been shown to manifest in different ways. Patients tend to be younger (age 47 vs 62), are less likely to present with muscle rigidity (especially in the case of clozapine), and the mortality rate is lower (3 vs 16.3% with conventional agents) (Troller et al., 2012).

Placebo Effects

A placebo is a pharmacologically inactive agent administered in a group of patients in a double-blind clinical trial to improve the true estimate of the active agent's efficacy and side effects. Placebo effects may be relevant to the concept of side effects from the perspective of experiences of drug effects in the absence of drug action. Substantial evidence suggests emergence of physiological changes associated with placebo administration. While placebo drug efficacy approaches significant percentage of patients in some clinical trials, incidences of side effects are also variably present. Placebo response rates vary widely among medication trials for depression, anxiety, and schizophrenia (21–70%) depending on the outcomes measured (Laporte and Figueras, 1994).

Sexual

In males impotence, decreased libido, changes in quality of orgasm, priapism (painful erection), and erectile and ejaculatory disturbances have been reported with antipsychotic use. In women orgasmic dysfunction and decreased libido are frequently observed with antipsychotic treatment. Sexual dysfunction associated with SSRI use impairs quality of life, and is a reason for noncompliance. Bupropion and nefazodone may have a more favorable profile as compared with TCAs, MAOIs, and venlafaxine. In contrast, some antidepressants are prescribed for the treatment of premature ejaculation.

Children and Adolescents

Children and adolescents experience higher rates of certain antipsychotic-induced side effects such as sedation, extrapyramidal side effects (except for akathesia), withdrawal dyskinesia, prolactin elevation, and weight gain. Young people are less likely to develop tardive dyskinesia or diabetes than adults treated with antipsychotics (Correll et al., 2011). Children may have more prominent weight gain during treatment with antipsychotic medications than adults (Newcomer, 2005). The risks of obesity, as well as cardiovascular, cerebrovascular or hypertensive adverse events are increased with the concomitant use of other psychotropic agents (Maayan and Correll, 2011).

Psychostimulants are often prescribed for the treatment of ADHD in children. These agents have the potential to cause

delayed onset of sleep, appetite suppression, headache, abdominal pain, growth delay, tics, and increases in pulse and blood pressure (Correll et al., 2011).

SSRI and SNRI medications have been found to be effective in the treatment of depression, anxiety, and obsessive-compulsive disorders in children and adolescents. Young people most frequently experience transient side effects such as nausea or headache. Rarely, they may also experience behavioral activation. Earlier studies led to concern for increased suicidal thinking or behavior among antidepressant-treated youth (Brent et al., 2009; Hammad et al., 2006). Subsequent studies indicate that depressed youth receiving antidepressants are at lower risk for completed suicide than those who are untreated, especially when cognitive behavioral therapy is added to medication management (Correll et al., 2011).

Elderly

Older individuals are more sensitive to side effects of psychotropic medications. Most notably, the elderly may experience more dyskinesia, anticholinergic side effects, cardiovascular effects, and sedation. Studies have indicated that elderly patients with schizophrenia may benefit from atypical antipsychotic agents over typical agents due to lower rates of dyskinesia and improved quality of life (Ritchie et al., 2003). Elderly patients may not exhibit significant weight gain with atypical antipsychotic use, but nonetheless are at increased risk of adverse health outcomes due to insulin resistance, dyslipidemia, and glucose intolerance with these agents (Newcomer, 2005).

Quality of Life and Side Effects

The World Health Organization has defined Quality of Life as:

> An individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept incorporating in a complex way the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of the environment.
>
> WHOQOL, 1999

This definition emphasizes a comprehensive approach in the assessment of quality of life. Disciplines of medicine, particularly cancer research have concentrated on global measures of quality of life, as symptom improvement by itself does not account for it. In psychiatry, the slow pace of research in this area may be due to difficulties of valid and reliable quality of life assessments from psychiatric patients. Pioneered by Lehman (Lehman et al., 1982), renewed interest has led to research in debilitating psychiatric disorders such as schizophrenia and demonstrated the significance of a multidisciplinary treatment model. While the strongest predictor of symptom improvement is identified as compliance with recommended drugs, its sole impact on quality of life in the long term is yet to be elucidated. The issue gets complicated due to the side-effect profile of psychoactive drugs, which may mimic the symptoms of diseases, and also due to the direct impact of side effects on quality of life. For patients with schizophrenia, symptoms, their severity, and subjective distress caused by side effects accounted for half the variance in quality of life (Awad et al., 1997).

In one study where 53 schizophrenic outpatients treated with depot neuroleptics were studied, 70% complained about side effects when 94% were found to experience them. Among the patients who denied having side effects, 87% had some form of it (Larsen and Gerlach, 1996).

The recognition of the negative impact of tardive dyskinesia on, for instance, patients' self-esteem, incapacity, and perception by society has been essential toward targeting more favorable treatments. The introduction of new generation psychoactive drugs, particularly antidepressants and antipsychotics is therefore encouraging. Such drugs have not only led to superior efficacy, but also have more favorable side-effect profiles. However, there remains significant room for further improvement in side-effect profiles of psychoactive drugs.

See also: Antidepressant Drugs; Endocrinology and Psychiatry; Geriatric Psychiatry; Medicalization: Sociological and Anthropological Perspectives; Opiate Use and Abuse, History of; Patient Adherence to Health Care Regimens; Prophylaxis in Psychiatry; Psychopharmacotherapy, Costs of; Psychotherapy and Pharmacotherapy; Schizophrenia: Metabolic Consequences.

Bibliography

American Psychiatric Association, 2000. Handbook of Psychiatric Measures. APA, Washington, DC.

Awad, A.G., Voruganti, L.N., Heslegrave, R.J., 1997. A conceptual model of quality of life in schizophrenia: description and preliminary clinical validation. Quality of Life Research 6 (1), 21–26.

Balon, R., 1999. Practical Management of the Side Effects of Psychotropic Drugs. Dekker, New York.

Bernstein, J.G., 1995. Handbook of Drug Therapy in Psychiatry. Mosby-Year Book, St Louis, MO.

Bilder, Robert E., et al., 2006. Cognitive development in schizophrenia: follow-back from the first episode. Journal of Clinical and Experimental Neuropsychology 28, 270–282.

Björkhem-Bergman, L., Asplund, A.B., Lindh, J.D., March 2011. Metformin for weight reduction in nondiabetic patients on antipsychotic drugs: a systematic review and metaanalysis. Journal of Psychopharmacology 25 (3), 299–305.

Bloom, F.E., Kupher, D.J. (Eds.), 1994. Psychopharmacology the Fourth Generation of Progress. Raven, New York.

Brent, D.A., et al., 2009. The Treatment of Adolescent Suicide Attempters Study (TASA): predictors of suicidal events in an open treatment trial. Journal of the American Academy of Child and Adolescent Psychiatry 48 (10), 987–996.

Buchanan, R.W., Holstein, C., Breier, A., 1994. The comparative efficacy and long-term effect of clozapine treatment on neuropsychological test performance. Biological Psychiatry 36 (11), 717–725.

Chwastiak, L.A., Rosenheck, R.A., McEvoy, J.P., Stroup, T.S., Swartz, M.S., Davis, S.M., Lieberman, J.A., January–February, 2009. The impact of obesity on health care costs among persons with schizophrenia. General Hospital Psychiatry 31 (1), 1–7.

Cancelli, I., et al., 2009. Drugs with anticholinergic properties: cognitive and neuropsychiatric side-effects in elderly patients. Neurological Sciences 30 (2), 87–92.

Casey, D.E., et al., 2004. Antipsychotic induced weight gain and metabolic abnormalities: implications for increased mortality in patients with schizophrenia. Journal of Clinical Psychiatry 65 (Suppl. 7), 4–18.

Colombo, G.L., Caruggi, M., Di Matteo, S., Rossi, A., October 2008. An economic evaluation of aripiprazole vs olanzapine adapted to the Italian setting using outcomes of metabolic syndrome and risk for diabetes in patients with schizophrenia. Neuropsychiatric Disease and Treatment 4 (5), 967–976.

Consensus Development Conference on Antipsychotic Drugs and Obesity and Diabetes, 2004. Diabetes Care 27 (2), 596–601.
Correll, C.U., et al., 2011. Developments in pediatric psychopharmacology: focus on stimulants, antidepressants, and antipsychotics. Journal of Clinical Psychiatry 72 (5), 655–667.
Correll, C.U., Leucht, S., Kane, J.M., March 2004. Lower risk for tardive dyskinesia associated with second generation antipsychotics: a systematic review of 1 year studies. American Journal of Psychiatry 161 (3), 414–425.
Dunner, D.L., et al., 1998. A prospective surveillance study for bupropion sustained-release in the treatment of depression. Journal of Clinical Psychiatry 59, 366–373.
Feldman, R.S., Meyer, J.S., Quenzer, L.F., 1997. Principles of Neuro-psychopharmacology. Sinauer Associates, Sunderland, MA.
Goff, D.C., et al., 2005. A comparison of 10 year cardiac risk estimates in schizophrenia patients from the CATIE study and matched controls. Schizophrenia Research 80, 45–53.
Goldberg, T.E., Weinberger, D.R., 1996. Effects of neuroleptic medications on the cognition of patients with schizophrenia: a review of recent studies. Journal of Clinical Psychiatry 57 (Suppl. 9), 62–65.
Hammad, T.A., et al., 2006. Suicidality in pediatric patients treated with antidepressant drugs. Archives of General Psychiatry 63 (3), 332–339.
Hessen, E., et al., 2006. Influence of major antiepileptic drugs on attention, reaction time, and speed of information processing: results from a randomized, double-blind, placebo-controlled withdrawal study of seizure-free epilepsy patients receiving monotherapy. Epilepsia 47 (12), 2038–2045.
Kaplan, H.I., Sadock, B.J., 1999. Comprehensive Textbook of Psychiatry. Williams and Wilkins, Baltimore, MD.
Keefe, R.S.E., et al., 2004. Comparative effect of atypical and conventional antipsychotic drugs on neurocognition in first-episode psychosis: a randomized, double-blind trial of olanzapine versus low doses of haloperidol. American Journal of Psychiatry 161, 985–995.
Keefe, R.S.E., et al., 2007. Neurocognitive effects of antipsychotic medications in patients with chronic schizophrenia in the CATIE trial. Archives of General Psychiatry 64, 633–647.
Laporte, J.-R., Figueras, A., 1994. Placebo effects in psychiatry. Lancet 344, 1206–1209.
Larsen, E.B., Gerlach, J., 1996. Subjective experience of treatment, side-effects, mental state and quality of life in chronic schizophrenic out-patients treated with depot neuroleptics. Acta Psychiatrica Scandinavica 93, 381–388.
Lee, M.A., 1999. A comparison of the effect of clozapine with typical neuroleptics on cognitive function in neuroleptic-responsive schizophrenia. Schizophrenia Research 37, 1–11.
Lehman, A.F., Ward, N.C., Linn, L.S., 1982. Chronic mental patients: the quality of life issue. American Journal of Psychiatry 139 (10), 1271–1276.
Loring, D.W., et al., 2011. Topiramate dose effects on cognition: a randomized double-blind study. Neurology 76 (2), 133–137.
Maayan, L., Correll, C., 2011. Weight gain and metabolic risks associated with antipsychotic medications in children and adolescents. Journal of Child and Adolescent Psychopharmacology 21 (6), 517–535.
Miller, D.D., Caroff, S.N., Davis, S.M., Rosenheck, R.A., McEvoy, J.P., Saltz, B.L., Riggio, S., Chakos, M.H., Swartz, M.S., Keefe, R.S., Stroup, T.S., Lieberman, J.A., 2008. Extrapyramidal side effects of antipsychotics in a randomized trial. British Journal of Psychiatry 193, 279–288.
Moore, A.R., O'Keefe, S.T., 1999. Drug-induced cognitive impairment in the elderly. Drugs & Aging 15 (1), 15–28.
Newcomer, J.W., 2005. Second generation (atypical) antipsychotics and metabolic effects: a comprehensive literature review. CNS Drugs 19 (Suppl. 1), 1–86.
Pacia, S., Devinsky, O., 1994. Clozapine-related seizure: experience with 5.629 patients. Neurology 44, 2247–2249.
Van Putten, T., 1974. Why do schizophrenic patients refuse to take their drugs? Archives of General Psychiatry 31 (1), 67–72.
Ray, W.A., et al., 2009. Atypical antipsychotic drugs and the risk of sudden cardiac death. The New England Journal of Medicine 360 (3), 225–235.
Ritchie, C.W., Chiu, E., Harrigan, S., Hall, K., Hassett, A., Macfarlane, S., Mastwyk, M., O'Connor, D.W., Opie, J., Ames, D., May 2003. The impact upon extra-pyramidal side effects, clinical symptoms and quality of life of a switch from conventional to atypical antipsychotics (risperidone or olanzapine) in elderly patients with schizophrenia. International Journal of Geriatric Psychiatry 18 (5), 432–440.
Schatzberg, A.F., Nemeroff, C.B., 1998. The American Psychiatric Press Textbook of Psychopharmacology. American Psychiatric Press, Washington, DC.
Serretti, A., Mandelli, L., 2010. Antidepressants and body weight: a comprehensive review and metaanalysis. Journal of Clinical Psychiatry 71 (10), 1259–1272.
Spiegel, R., 1996. Psychopharmacology: An Introduction, third ed. Wiley, Chichester, UK.
Swanson, J.M., et al., 2007. Effects of stimulant medication on growth rates across 3 years in the MTA follow-up. Journal of the American Academy of Child and Adolescent Psychiatry 46 (8), 1015–1027.
Troller, J.N., et al., 2012. Comparison of neuroleptic malignant syndrome induced by first generation and second generation antipsychotics. British Journal of Psychiatry 201, 52–56.
Truman, C.J., et al., 2007. Self-reported history of manic/hypomanic switch associated with antidepressant use: data from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). Journal of Clinical Psychiatry 68 (10), 1472–1479.
Velligan, et al., 2003. The effectiveness of quetiapine versus conventional antipsychotics in improving cognitive and functional outcomes in standard treatment settings. Journal of Clinical Psychiatry 64, 524–531.
Wang, M., et al., 2012. Metformin for treatment of antipsychotic-induced weight gain: a randomized, placebo-controlled study. Schizophrenia Research 138, 54–57.
Woods, S., et al., 2010. Incidence of tardive dyskinesia with atypical versus conventional antipsychotic medications: a prospective cohort study. Journal of Clinical Psychiatry 71 (4), 463–474.
World Health Organization, 1999. WHOQOL Annotated Bibliography. WHO, Geneva, Switzerland.

Psychophysical Theory and Laws, History of

Stephen W Link, University of California, San Diego, La Jolla, CA, USA; and McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychophysics is a scientific approach to the measurement of mental processes. In *Elemente der Psychophysik* in 1860, Gustav Fechner defined the field, quantified Weber's law, created a theory of judgment errors in sensory comparisons, and proposed a logarithmic law of sensation strength. In 1927, theoretical advances by Thurstone expanded the applications of psychophysical methods to stimuli not measurable on a physical scale, thereby permitting measurements of attitudes, preferences, and feelings. Stevens (1956) argued against these ideas and proposed, instead, a power law of sensation strength, an idea previously proposed by Plateau (1872). New experimental data required both camps to accommodate the influence on performance of 'cognitive factors,' which caused deviations from the power law of sensation. In classical psychophysical experiments, the effects of cognitive factors led to the theory of ideals, the theory of signal detection, receiver operating characteristics, and a sequential theory of psychological discrimination. At the beginning of the twenty-first century psychophysics is poised for major theoretical advances toward more refined and more extensive measurements of mental processes.

Introduction

Psychophysics is a scientific approach to the measurement of mental processes. The empirical laws of psychophysics are based on measures such as the magnitudes of stimuli just noticeably different from each other, the proportion of errors made when comparing two stimuli, the time required to complete such judgments, and self-reports of the confidence in comparative judgments or the strength of a stimulus. Some laws define relations between physical events and their mental representations such as sensations and feelings. The theories of psychophysics provide explanations of these empirical relations and suggest neural mechanisms that give rise to experimental results.

Nineteenth-Century Psychophysics

The first law defined the 'just noticeable difference' (jnd) in sensation. A jnd occurred, phenomenally, when an increment in the size of a physical stimulus evoked a sensation of a difference in magnitude. The German physiologist Ernst Heinrich Weber (1834/1978) discovered, for many sense modalities, the increment in a physical stimulus needed to evoke a jnd. Leipzig physicist Gustav Fechner (1860/1966) quantified Weber's observations and created the First Law of Psychophysics, Weber's law,

$$\Delta s/S = c \qquad [1]$$

where S is the magnitude of the physical stimulus, Δs is the just noticeable increment, and c is called Weber's constant. For discriminating lifted weights $c = .043$. Thus, an initial weight of 1000 g will require an increment of 43 g (to $S + \Delta s = 1043$ g) before the new weight will feel just noticeably different from 1000 g.

The subjective nature of these self-reports led Fechner to wonder whether the person experiencing a jnd could identify which of the two stimuli, S or $S + \Delta s$, was larger. Fechner discovered that two stimuli may feel just noticeably different, but may be identified correctly in only about 75% of the comparisons made. Fechner described these experiments and officially established the field of psychophysics in the two-volume *Elemente der Psychophysik* (1860/1966).

In *Elemente* Vol. 1, Fechner created a theory of how humans comparing two sensory stimuli, such as a pair of lights for brightness, two weights for heaviness, or two lines for length, make errors in their judgments. The theory assumed that the sensory system provided the mind with measurements of the external world. However, the process by which the sensory/nervous system created an internal sensory representation of an external physical stimulus introduced error into the measurement. A fixed weight might at one time feel heavier, or lighter, depending upon the amount of error added to, or subtracted from, the true sensory value for the external stimulus. Extending the theory of error created by Gauss (1809) for physical measurement devices, Fechner invented a theory of judgment about invisible mental events that allowed him to measure the size of the error in the nervous system.

Figure 1 illustrates the essential features of Fechner's theory as it applies to the comparison of two just noticeably different weights, W_a and W_b. Along the abscissa are true sensory values, μ_a and μ_b, for the two weights undergoing comparison by lifting each once. Two Gaussian (normal) probability distributions describe the error of sensory measurement and surround the true sensory values of μ_a and μ_b. Fechner postulated that a (decision) threshold, t, for deciding which weight was the larger existed exactly midway between μ_a and μ_b, that is, at

$$t = \mu_a + (\mu_a - \mu_b)/2 \qquad [2]$$

A lifted weight sensed to be greater than this threshold was thought to be the heavier weight.

The shaded area of Figure 1 to the right of t represents the probability of making an error of judgment in deciding that the smaller of two stimuli was in fact the larger. The symmetric nature of Fechner's theory predicts that the probability of judging the larger of two stimuli to be the smaller equals the same probability, and is shown by the shaded area to the left of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43084-9

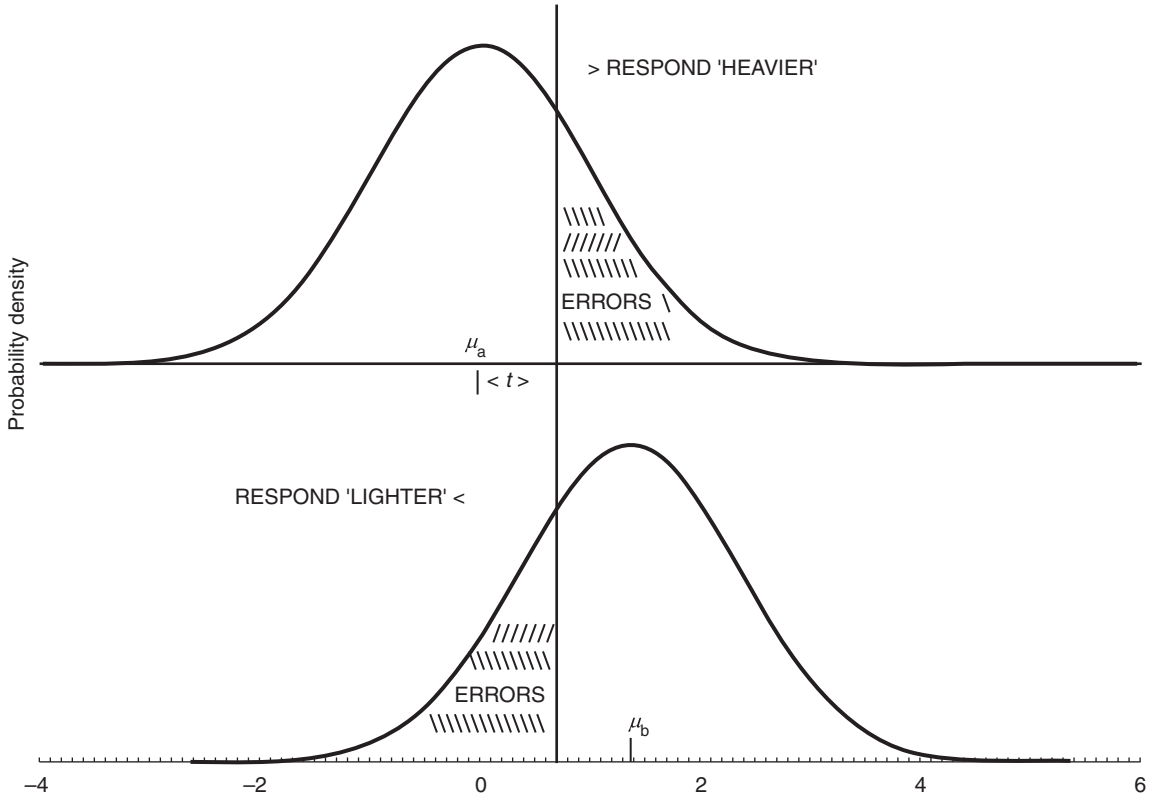

Figure 1 Fechner's theory of mental judgment proposed that Gaussian variability perturbed nervous system transmission of peripheral sensory stimuli. Weight W_a was incorrectly judged to be heavier than weight W_b whenever the sensory value for W_a exceeded the value t. Similarly, W_b was incorrectly judged to be lighter than W_a whenever the sensory value for W_a was less than t. The probabilities of an error are represented by the shaded areas to the right and left of t.

the threshold, t. The two error probabilities increase as the variability in the sensory system increases. Such an increase in variability causes a greater spread in each Gaussian probability distribution.

Fechner's goal was to measure the invisible variability of the sensory system in units of the physical stimuli. As a measure of variability, Fechner chose the value $\sqrt{2}\sigma$, Gauss's measure of the spread of the Gaussian distribution. The value of σ, a parameter of the Gaussian probability distribution, creates a scale of measurement units along the abscissa of Figure 1. The unknown value for t is located at unknown number of σ units above the true sensory value μ_a. However, once the probability of an error of judgment is known, one can determine from tables (many created by Fechner) the corresponding distance in σ units of the (decision) threshold t above the true sensory value μ_a. For example, if the probability of an error in judging which weight is larger equals .25, then the (decision) threshold, t, is $.674\sigma$ units greater than the true sensory value μ_a.

To determine the value of σ, Fechner replaced the unknown values of μ_a and μ_b by the physical values that gave rise to them, W_a and W_b. The distance from μ_a to t is defined as $t = .675\sigma$. That is,

$$t = .675\sigma = (W_b - W_a)/2, \text{or} \quad \sigma = (W_b - W_a)/(2^*.675) = (W_b - W_a)/1.35 \qquad [3]$$

Using the example above, the smaller of the two just noticeably different weights equaled 1000 g, and the larger, 1043 g (and the probability of an error equaled .25). Thus the measure of variability is

$$\sigma = (1043\,\text{g} - 1000\,\text{g})/1.35 \approx 32\,\text{g} \qquad [4]$$

Fechner summarized these ideas in a famous formula that is the Second Law of Psychophysics:

$$t = hD/2 \qquad [5]$$

Here $h = 1/\sqrt{2\sigma}$, and D equals the difference between two physical stimuli, $W_b - W_a$. The term h measures the sensitivity of the system. When σ is large, indicating great sensory variability and a small value of h, many errors occur and the system is said to be insensitive. A small value for σ leads to few errors, a larger value for h, and the system is said to be more sensitive.

This first method of measuring the invisible, internal variability of the sensory system established psychophysics as a truly scientific discipline. Measuring the unknown variability of the sensory/nervous system in units of physical stimuli marks a revolution in the scientific investigation of the human mind. This powerful theory became the basis for the fields of experimental psychology and statistical hypothesis testing.

Fechner also theorized, in *Elemente* Vol. 2, about a measure of how much sensation a stimulus generated. Placing each jnd in a sequence is a way of adding together phenomenally equal

units of sensation. Using Weber's law as a base, Fechner argued that on a continuous scale of sensation Weber's constant,

$$c = \Delta s/S \quad [6]$$

is a differential equation. Integrating the differential equation is equivalent to adding together the jnds of sensation increments to obtain a total amount of sensation. The result of the integration is called Fechner's Law of Sensation: the total amount of sensation generated by a stimulus is given by the equation

$$\xi = k\log_e(S) \quad [7]$$

where k is a constant of integration, S is the magnitude of the stimulus, and ξ is the amount of sensation.

Fechner's law is perhaps the most widely known of any psychophysical law. Yet criticisms of Fechner's idea never abated (Luce and Edwards, 1958; Stevens, 1975; Krueger, 1989; Lockhead, 1992; Murray, 1993). The notion of adding together jnds to create a yardstick of sensory measurement is appealing, but the 'law' provides for no independent experimental means for testing its prediction. How is internal sensation to be measured, or as Laming (1997) wonders, 'Can sensation be measured?' The answers to this question fall in the domain of representational measurement – how numbers are attached to qualitative objects (Narens, 2007).

The father of American philosophical pragmatism, C. S. Peirce, put Fechner's ideas to another test. Peirce argued that the sense of confidence in a judgment was a sensation directly related to the probability of making a correct judgment. Peirce's law states that if C is the average confidence in a judgment, p the probability of making a correct judgment, and k a constant, then

$$C = k\log_e[p/(1-p)] \quad [8]$$

Like Fechner's law there is a logarithmic relation between two variables. However, in this case each variable can be measured through direct experimentation. Peirce and Jastrow (1884, see Link, 1992; Vickers, 1979), and later experiments, confirmed this prediction (Link, 2003).

The profound impact of Fechner's Psychophysics led to Wilhelm Wundt's 1879 creation of the first psychological laboratory at Leipzig University. A student, James McKeen Cattell, proposed a new measure of the amount of sensation generated by perfectly discriminable stimuli, as opposed to jnd stimuli. Reaction time, the time taken to respond to the presence of a stimulus, could measure stimulus strength. Cattell's experiments showed remarkable decreases in response time as stimulus magnitude increased. Later (Piéron 1914, see Welford, 1980; Luce, 1986) proposed the general principle Piéron's law:

$$\mathrm{RT} = a + kS^{-\beta} \quad [9]$$

where a is a constant due to nonsensory effects, k and β are constants, and S is physical stimulus intensity.

Twentieth-Century Psychophysics

By 1900, Fechner's experimental methods and a measure of the response threshold, t, in units of the physical stimulus were well established as part of the new psychology's experimental methods. To create a uniform basis for comparison of experimental results, the stimulus generating a value of t corresponding to 75% correct responses was chosen as a standard measure of sensitivity. When a large stimulus was required for 75% correct responses the responder was less sensitive than when a smaller stimulus generated 75% correct responses. Thus, thresholds became a tool useful in comparing performances across subjects and stimuli. The choice of this threshold value derives in part from the often-used nineteenth-century measure of variability the 'probably error.' Uses of the 75% threshold are still found in the current literature.

Determining the stimulus that generated 75% correct response was formalized by Urban (1907). His method of constant stimuli required that a fixed set of comparison stimuli be compared at random against a fixed standard stimulus. For example, a set of weights varying from 88 to 112 g, compared at random against a fixed standard of 100 g, generated a series of 'heavier' responses varying from near zero for the 88 g comparison to nearly 100% for the 110 g comparison. Urban defined this function as a 'psychometric function.' An estimate of the comparison weight generating 75% correct heavier was then estimated by fitting a smooth curve to the empirical psychometric function. In vision studies the psychometric function is called the 'frequency of seeing curve.'

Urban suggested that another response is available beyond the usual two choices when using the method of constant stimuli. This new response alternative allowed for the judgment of equal when no difference between the standard and comparison was perceived. Studies of equality judgments themselves became stable features of psychophysical experiments for many decades and led to questions about sameness versus difference.

Measuring the magnitudes of perfectly discriminable stimuli remained elusive. Thurstone (1927, 1955), Bock and Jones (1968) proposed that every stimulus projects a value on an internal psychological continuum. For example, stimuli such as crimes may vary on a continuum of the seriousness of crime. When asked to judge which of two crimes is more serious, a subject will compare the two crime's seriousness values and choose the crime with the greater seriousness value. This is the famous Law of Comparative Judgment. For each subject, crimes vary in their degree of seriousness but for a fixed subject each crime's seriousness is fixed. For a fixed crime, individual differences across subjects created variability across the dimension of seriousness of crime. Thurstone represented this variability by a Gaussian distribution of seriousness values. For a group of subjects the theory is formally identical to Fechner's, but the emphasis on measuring stimulus values, rather than variability, created a powerful new direction in psychophysical research.

Thurstone proved that by comparing all stimuli against each other, each stimulus could be measured on the underlying psychological continuum. Now, mental phenomena as elusive as attitudes could be measured. In an application to classical psychophysics, Guilford (1928, see 1954) used the method to determine heaviness values for difficult-to-discriminate lifted weights.

World War II placed new demands on psychologists. As an example, SONAR (SOund Navigation And Ranging) allowed for the detection of enemy submarines shadowing convoys or

warships in wait for an opportunity to attack. A SONAR operator wearing earphones listened to amplified underwater sounds in an effort to detect the presence of an enemy submarine. In the colorful language of signal detection, the operator made a hit when detecting a submarine and a false alarm when claiming a submarine was present when it was not. The human operator suffered stress due to fatigue and required a period of learning before providing a sufficiently low rate of false alarms to be of military effectiveness.

The obvious theoretical treatment of this problem of recognizing the sound of a submarine imbedded in the noisy sounds of the underwater environment is the traditional theory of discrimination due to Fechner. The sound of the noisy underwater environment is one stimulus and a second stimulus is created by embedding the sound of a nearby submarine into the noise. The SONAR operator must be continually alert so as not to miss the presence of an enemy. As was soon discovered, both fatigue and prior information affected the ability to detect a signal embedded in noise.

Experiments from many different laboratories presented conflicting pictures of human signal detection. To account for many experimental results, Peterson et al. (1954) developed the Theory of Signal Detectability. The theory is an extension and deeper development of the theory for deciding between two stimuli first described by Fechner (1860). In Figure 1 let the upper probability distribution represent the distribution of psychological values generated by just the noise alone as heard by the SONAR operator. The lower probability distribution represents the distribution of psychological values when the submarine (the signal) is also present. Let t, called a criterion in signal detectability theory (SDT), be the smallest value needed to respond that a submarine is present. The shaded area to the right of t in the upper distribution represents the probability of a false alarm, the probability of claiming a submarine to be present when only noise is actually present. The unshaded area to the right of t in the lower probability distribution represents the probability of a hit, of detecting a submarine when present.

To measure performance, hit and false alarm probabilities may be plotted against each other in a receiver operating characteristic (ROC) space as shown in Figure 2. In Figure 1 the false alarm probability equals 1 minus the hit probability. This probability lies somewhere on the antidiagonal of Figure 2 depending on the difference between the means of the two probability distributions. If $\mu_a = \mu_b$, then the distributions are identical, and the probability of a hit equals .50 as does the probability of a false alarm. This outcome is represented in Figure 2 at the point where the antidiagonal and the main diagonal, showing guessing, intersect. As the mean of the signal plus noise distribution increases, this point will ascend the antidiagonal, showing an increase in the probability of a hit and a decrease in the probability of a false alarm, due to better discriminability of the two stimuli. This is in accord with Fechner's theory of discrimination.

However, factors were soon discovered that drove the hit and false alarm probabilities to deviate from the antidiagonal. The curved line in Figure 2 represents hit and false alarm probabilities that occur if the value of t in Figure 1 is allowed to vary away from its position midway between the means of the

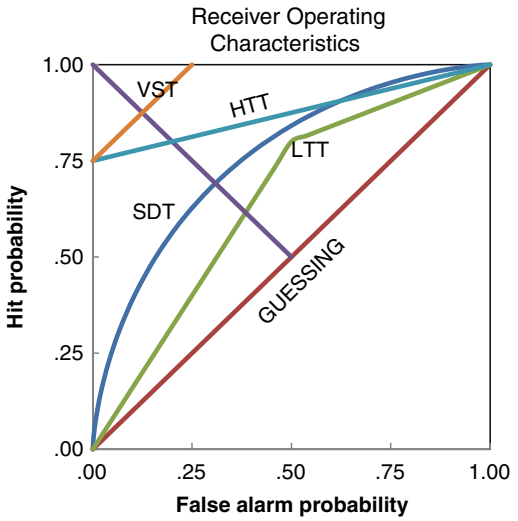

Figure 2 Illustrations of receiver operating characteristics predicted by different psychophysical theories of discrimination for judgments between two stimuli. The abscissa is the probability of an incorrect response to one of the two stimuli (false alarm), and the ordinate is the probability of a correct response (hit) to the other stimulus. The main diagonal represents guessing. Different ROCs correspond to theories described in the text: High threshold theory (HTT), signal detectability theory (SDT), low threshold theory (LTT), and variable sensitivity theory (VST). All these ROCs are given a unitary basis by Link (1992).

two distributions. At the most extreme, with t to the very left side of Figure 1, hits and false alarms occur with probability 1, the areas under the probability distributions to the right of t. This point is at the upper right hand corner of the ROC curve for SDT. As the value of t increases, the hit and false alarm probabilities sweep out the smooth curve labeled SDT in Figure 2. Many experiments supported this view of how changes in performance were caused by changes in the value of t. Changes in the value of t away from its midway position between the two means represented 'response bias.'

Many sensory experiments produced the form of the ROC predicted by SDT (Swets et al., 1961; Swets, 1964, 1996). Experimental subjects were able to bias responding in the light of knowledge about the relative frequencies of the two stimuli, monetary payoffs for correct or error responses, and verbal suggestions. Additional research revealed other forms of the ROC inconsistent with the predictions of SDT. Blackwell (1953) suggested a high threshold theory (HTT), with guessing occurring when the stimulus produced a sensory value less than a threshold. The straight line ROC in Figure 2 illustrates the predicted HTT relation between hits and false alarms replicated in many experiments. Luce (1963) suggested a low threshold theory (LTT) that generated the 'dog-leg' ROC illustrated in Figure 2. Atkinson (1963) also suggested the variable sensitivity theory (VST) that generated the HTT ROC and the uppermost ROC shown at the upper left corner of Figure 2. All these various theories received solid support from experimental results.

A consolidation of these various ROCs resulted from a theory suggesting a different account of their common occurrence. Link (1975, 1992) and Link and Heath (1975) derived these various ROCs from a distribution-free

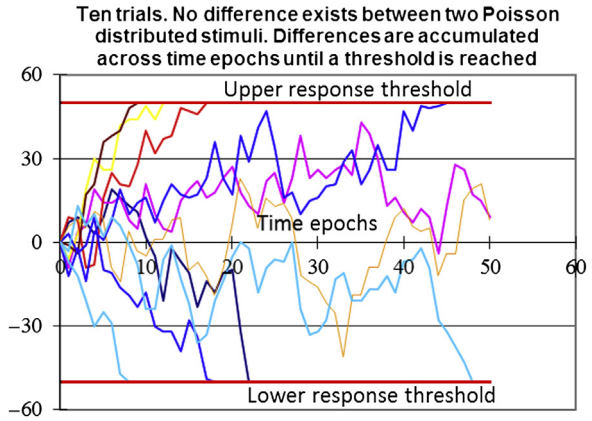

Figure 3 Ten accumulations of comparative differences across time epochs when two stimuli are actually identical. Each stochastic path represents the result of a single experimental trial. The upper and lower response thresholds determine which of two possible responses can occur. The distance between thresholds can be manipulated by the subject as more or less total comparative difference is required for making judgments. Response bias is determined by the starting position of the accumulation process. Link (1978, 1992) shows how response time deadlines imposed on a subject's performance cause the subject to move the thresholds, thereby reducing response time and response accuracy.

sequential theory of psychological judgment. In this relative judgment theory (RJT) comparative differences accumulate during the process of decision making until either an upper or a lower response threshold is exceeded, then a response occurs. As shown in the examples in Figure 3, even when there is no difference between two stimuli, the accumulation of comparative differences can generate very fast response acquisition for some experimental trials and much slower responses on other trials, although nothing changes in either the stimuli or the subject making decisions.

As these response thresholds are manipulated by subjects to meet the demands of the experimental setting all the different ROCs result. By increasing the distance between thresholds subjects require more total comparative difference before responding, leading to longer response times and greater accuracy of discrimination. For a fixed difference between the upper and lower thresholds moving the starting position of the process from the upper threshold toward the lower threshold will sweep out performances like SDT. The HTT ROC occurs when subjects hold the lower threshold constant but vary the upper threshold in response to performance demands. For LTT first one threshold is held constant and the other threshold is allowed to vary. Then the previously variable threshold is held constant and the other threshold varies. The VST ROC results from using two different sets of response thresholds to meet performance demands (Link and Heath, 1975; Link, 1992). Especially important to RJT are predicted relations between response probabilities and response times, between the psychometric function and the corresponding chronometric function (Link, 1975, 1992). The theory spawned a new era of random walk theories applied to psychological research.

However, perplexing irregularities in the predicted ordering of probabilities in the Thurstonian full paired comparison matrix were the basis for Clyde H. Coombs to propose an important extension of the Law of Comparative Judgment, the Theory of Ideal Comparisons. Rather than comparing two stimuli against each other, Coombs (1964) proposed that each stimulus be compared against an ideal defined on a psychological continuum. Choosing between two stimuli equally distant from the ideal led to indifference with respect to them and the peculiar response proportions often observed in some paired comparison matrices. For example, a continuum of shades of gray may contain an ideal gray. However, two shades of gray equally distant on either sides of this ideal would be equally preferred. The subject faced with choosing between either of these stimuli as to which was the better gray would be indifferent to them and chose either with probability .50.

In spite of its immense value, the Law of Comparative Judgment said nothing about how response time or confidence should vary when stimuli are compared against each other. (Johnson 1939, see Link, 1992) combined choice, confidence, and response times (RT) in an experiment on judgments of differences between line lengths and discovered Johnson's law:

$$\mathrm{RT} = a(10)^{b(1-C)} \qquad [10]$$

where $0 \leq C \leq 1$ is a measure of confidence and a and b are constants. Johnson's law, together with Peirce's law, suggests a relationship between confidence and response probability. Link (2003) confirmed predicted relations between response time and the discriminability measures of the sequential theory of decision making first proposed in 1975.

Stevens (1975) dismissed the theoretical reasoning leading to Fechner's law and the Law of Comparative Judgment, and proposed an alternative formulation for determining the amount of sensation generated by a single stimulus. According to Stevens, "equal stimulus ratios produce equal sensation ratios." This idea is the basis for the famous power law of sensation,

$$\xi = \lambda S^{\gamma} \qquad [11]$$

where ξ is the strength of response to a stimulus of intensity S, λ is a nonsensory scaling constant, and γ measures the effects of peripheral sensory transducers that are unique to a sensory modality (Plateau, 1872; Laming and Laming, 1996). Stevens (1975) determined power law exponents for many sensory quantities and gave convincing evidence of the satisfactory representation of stimulus strength provided by the power law.

Although the power law contradicts Fechner's law, the power law must be consistent with Weber's law. Link (1992) derived the prediction that Stevens' exponents and Weber's constants are reciprocals of each other. Using values reported by Teghtsoonian (1971), the prediction holds for sensory quantities investigated by Stevens such as shock, heaviness, loudness, finger span, vibration, saturation, taste, and brightness. Ekman's law (Ekman, 1956) provides yet another relation between jnds and the power law by asserting that increments in sensation on a psychological continuum are a linear function of sensation magnitude, as if Weber's law applied to both the psychological continuum and the physical stimulus continuum. Gescheider (1997) shows how Ekman's law leads to Stevens' power law.

Cognitive factors also affect the magnitude of response. For example, Parducci (1965) develops a theory of how the range and frequency of stimuli affects judgments. Birnbaum (1974) examines context effects on comparative judgments and Birnbaum and Mellers (1978) examine ratio and difference judgments of distances in mental maps. Baird (1996) suggests that beyond the sensation effects caused by the peripheral nervous system are cognitive factors that demand a reinterpretation of the power law. Proposing a complementarity theory of psychophysics, Baird accounts for a wide variety of experimental results unaccounted for by the power law. In a different approach, Norwich (1993) developed a general theory of the sensory language of the brain. Using the fundamental equation

$$F = kH \qquad [12]$$

where F is a perceptual variable, k is a positive constant, and H is a measure of entropy, Norwich derives both the power law and Fechner's law from the same basis. Poulton (1989) reveals how 'biases may conceal the linear relation between magnitude judgments and the stimulus magnitudes that appear to the observer to be equally spaced.' Warren's (1958) physical correlate theory holds that sensory magnitude estimates are based upon the manner in which sensory excitation is correlated with the quantity of a physical attribute of the stimulus.

For choices between many stimuli, Luce (1959, see Falmagne, 1985) created the choice axiom. Let $S_1, \ldots, S_n$ be a set of commensurate stimuli. Luce's choice axiom states that the probability of choosing stimulus S_i is

$$W(S_i)/\left(\sum_n W(S_n)\right) \qquad [13]$$

where W_i is the psychophysical value of stimulus S_i. For example, given five breakfast cereals, $B_1, \ldots, B_5$, the probability of choosing cereal B_i is

$$p_i = W(B_i)/\left(\sum_h W(B_n)\right) \qquad [14]$$

Luce's choice axiom receives support from numerous experiments in psychology and economics.

The majority of psychophysical theories and experiments focus on the ability to discriminate between stimuli rather than on the similarity between them. Shepard (1957, 1987, see Borg and Lingoes, 1987) proposed that similarity between two stimuli, (S_i, S_k), is a continuously changing negative exponential function of the distance, d_{ik}, between them. The probability of correctly identifying a particular stimulus from a set of n stimuli is often assumed to be

$$p_{ik} = \exp(-d_{ik})\Big/\sum_n \exp(-d_{in}) \qquad [15]$$

a form similar to Luce's choice axiom.

Following Thurstone's (1933) suggestion that there are many underlying psychological continua, Anderson (1981) proposed an information integration theory of how sensory values on different dimensions are combined to create perceptual experience. Anderson's cognitive algebra creates perceptual magnitudes from a linear model of sensory values defined on the dimensions on which a stimulus projects its values. Subjectively, a fine dinner is a sum of the psychological values of its many components.

The question as to how these complex stimuli are represented in a psychological space has moved beyond the earlier considerations of single dimensions to questions of whether multiple dimensions define a space in which stimuli are represented. Ashby and Townsend (1986) suggest ways to determine separability, independence, and orthogonality, while Suppes (2002) and Colonius and Dzhafarov (2006) provide deeper analyses of the necessary conditions for representations and issues surrounding multiple and infinite dimensional representation spaces. As psychophysics enters the twenty-first century, its applications in many areas of psychology and related disciplines are described in *Psychophysics Beyond Sensation* (Kaernbach et al., 2004).

Conclusion

During nearly 185 years, the field of psychophysics introduced new measurements of, and theories about, mental processes. Following Weber's discovery of the jnd, Fechner created Weber's law, the field of psychophysics, and a theory of error in sensory comparisons. Fechner's law, that the amount of sensation equaled a logarithmic transformation of the physical stimulus intensity, provoked continual controversy. Piéron showed that the time required to respond to the presence of a stimulus obeyed a power law of stimulus intensity. Stevens proposed that the amount of sensation obeyed a power law with an exponent unique to a sensory modality. Urban created the psychometric function as a means to determine accurately the 75% threshold. Thurstone introduced the idea of psychological continua on to which stimuli project psychological values. Comparisons between two psychological values led to the Law of Comparative Judgment and measurements of stimuli not defined on a physical scale. Anderson proposed a cognitive algebra for adding or multiplying psychological values combined across several such psychological continua.

However, cognitive factors were also shown to affect judgments of stimuli. Coombs proposed that psychological values were compared against ideals rather than against each other, and suggested a method of locating an ideal. The signal detection theory provided for a decision criterion that could vary in response to cognitive factors. The various ROC results previously requiring separate theories for their explanation were shown to follow from a sequential theory of psychological discrimination. Theoretical ideas suggested new experiments. New empirical discoveries challenged old theories. Relations between response probabilities, response times, and response confidence provided new ways to test theories. Now, at the beginning of the twenty-first century, psychophysics is poised for major theoretical advances toward more refined and more extensive measurements of mental processes as evidenced in *Measurement with Persons* (Berglund et al., 2012).

See also: Sensation and Perception: Direct Scaling; Signal Detection Theory, History of.

Bibliography

Anderson, N.H., 1981. Foundations of Information Integration Theory. Academic Press, New York.

Ashby, F.G., Townsend, J.T., 1986. Varieties of perceptual independence. Psychological Review 93, 154–179.

Atkinson, R.C., 1963. A variable sensitivity theory of signal detection. Psychological Review 70, 91–106.

Baird, J.C., 1996. Sensation and Judgment: Complementarity Theory of Psychophysics. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Berglund, B., Rossi, G.B., Townsend, J.T., Pendrill, L.R. (Eds.), 2012. Measurement with Persons: Theory, Methods and Implementation Areas. Taylor and Francis, New York.

Birnbaum, M.H., 1974. Using contextual effects to derive psychophysical scales. Perception and Psychophysics 15 (1), 89–96.

Birnbaum, M.H., Mellers, B.A., 1978. Measurement and the mental map. Perception and Psychophysics 23 (5), 403–408.

Blackwell, H.R., 1953. Psychophysical Thresholds: Experimental Studies of Methods of Measurement. University of Michigan Engineering Research Bulletin, No. 36, University of Michigan Press, Ann Arbor.

Bock, R.D., Jones, L.V., 1968. The Measurement and Prediction of Judgment and Choice. Holden-Day, San Francisco.

Borg, I., Lingoes, J., 1987. Multidimensional Similarity Structure Analysis. Springer-Verlag, New York.

Colonius, H., Dzhafarov, E.N. (Eds.), 2006. Measurement and Representation of Sensations. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

Ekman, G., 1956. Discriminal sensitivity on the subjective continuum. Acta Psychologica 12, 233–243.

Falmagne, J.C., 1985. Elements of Psychophysical Theory. Oxford University Press, New York.

Fechner, G.T., 1860/1966. In: Boring, E.G., Howes, D.H. (Eds.), Adler, H.E. (Trans.), Elements of Psychophysics. Holt, Rinehart & Winston, New York.

Gauss, C.F., 1809. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Perthes and Besser, Hamburg.

Gescheider, G.A., 1997. Psychophysics: The Fundamentals. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Guilford, J.P., 1928. The method of paired comparisons as a psychometric method. Psychological Review 35, 494–506.

Guilford, J.P., 1954. Psychometric Methods, second ed. McGraw-Hill, New York.

Johnson, D.M., 1939. Confidence and speed in two-category judgment. Archives of Psychology 241, 1–52.

Kaernbach, C., Schröger, E., Müller, H. (Eds.), 2004. Psychophysics Beyond Sensation: Laws and Invariants of Human Cognition. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Krueger, L.E., 1989. Reconciling Fechner and Stevens: toward a unified psychophysical law. Behavioral and Brain Sciences 12, 251–267.

Laming, D.R.J., 1997. The Measurement of Sensation. Oxford University Press, New York.

Laming, J., Laming, D., 1996. J. Plateau: on the measurement of physical sensations and on the law which links the intensity of these sensations to the intensity of the source. Psychological Research 59, 134–144.

Link, S., 1975. The relative judgment theory of two-choice reaction time. Journal of Mathematical Psychology 12, 112–135.

Link, S.W., 1978. Relative judgment theory analysis of response time deadline experiments. In: Restle, F., Castellan Jr., N.J., Birnbaum, M., Link, S.W. (Eds.), Cognitive Theory III. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

Link, S.W., 1992. The Wave Theory of Difference and Similarity. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

Link, S.W., 2003. C. S. Peirce, Confidence and Random Walk Theory. In: Fechner Day 2003. International Society for Psychophysics, Stockholm.

Link, S.W., Heath, R.A., 1975. A sequential theory of psychological discrimination. Psychometrika 40, 77–105.

Lockhead, G.R., 1992. Psychophysical scaling: judgments of attributes or objects? Behavioral and Brain Sciences 15, 543–558.

Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.

Luce, R.D., 1963. A threshold theory for simple detection experiments. Psychological Review 70, 61–79.

Luce, R.D., 1986. Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York.

Luce, R.D., Edwards, W., 1958. The derivation of subjective scales from just noticeable differences. Psychological Review 97, 66–77.

Murray, D.J., 1993. A perspective for viewing the history of psychophysics. Behavioral and Brain Sciences 16, 115–137.

Narens, L., 2007. Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Norwich, K.H., 1993. Information, Sensation and Perception. Academic Press, San Diego, CA.

Parducci, A., 1965. Category judgment: A range-frequency model. Psychological Review 72 (6), 407–418.

Peirce, C.S., Jastrow, J., 1884. On small differences in sensation. Memoirs of the National Academy of Sciences 3, 73–83.

Peterson, W.W., Birdsall, T.G., Fox, W.C., 1954. The theory of signal detectability. Transactions of the Institute of Radio Engineers: 1954 Symposium on Information Theory 4, 171–212.

Piéron, H., 1914. Researches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices. L'Année Psychologique 20, 17–96.

Plateau, J., 1872. Sur la mesure defs sensations physiques, et sur la loi qui lie l'intensité de ces sensations à la l'intensité de la cause excitante. Bulletins de L'Academie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique. 2me Sér 34, 250–262.

Poulton, E.C., 1989. Bias in Quantifying Judgments. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Shepard, R.N., 1957. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika 22 (4), 325–345.

Shepard, R.N., 1987. Towards a universal law of generalization for psychological sciences. Science 237, 1317–1323.

Stevens, S.S., 1975. In: Stevens, G. (Ed.), Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects. Wiley, New York.

Suppes, P., 2002. Representation and Invariance of Scientific Structures. CSLI Publications, Stanford.

Swets, J.A. (Ed.), 1964. Signal Detection and Recognition by Human Observers: Contemporary Readings. Wiley, New York.

Swets, J.A., 1996. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. Collected Papers. Lawrence Erlbaum Associates, Publishers, Mahwah, NJ.

Swets, J.A., Tanner, W.F., Birdsall, T.G., 1961. Decision processes in perception. Psychological Review 68, 301–340.

Teghtsoonian, R., 1971. On the exponents in Stevens' law and the constants of Ekman's law. Psychological Review 78, 71–80.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.

Thurstone, L.L., 1933. Theory of Multiple Factors. Edwards Bros, Ann Arbor, Mich.

Thurstone, L.L., 1955. The Measurement of Values. University of Chicago Press, Chicago.

Urban, F., 1907. On the method of just perceptible differences. Psychological Review 14, 244–253.

Vickers, D., 1979. Decision Processes in Visual Perception. Academic Press, New York.

Warren, R.M., 1958. A basis for judgments of sensory intensity. American Journal of Psychology 71, 675–687.

Weber, E.H., (1978/1834). Der tastsinn. In: Ross, H.E., Murray, D.J. (Eds.), Ross, H.E. (Trans.), E. H. Weber: The Sense of Touch. Academic Press, London (Original published 1834).

Welford, A.T. (Ed.), 1980. Reaction Times. Academic Press, New York.

Psychosocial Work

Timothy Sim, The Hong Kong Polytechnic University, The Hong Kong Special Administrative Region, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins by defining the variously used term 'psychosocial' across disciplines, followed by an examination of its ambiguous position in social work. The dual-foci debate in social work provides the context for understanding the evolution of psychosocial work through time and applications. Eschewing the narrow conceptualizations related to psychopathology and trauma is a major shift in international psychosocial work, particularly in emergencies, to address issues of resilience, community, and cultural resources, wider social influences, and human rights. After presenting the emerging trends in the 'both–and' perspective in psychosocial work and reviewing research, ethics, and other critical issues, this article charts the future direction of psychosocial work.

Introduction

Psychosocial practice is one of the oldest, if not the oldest, professional practice framework in social work, and it has continued to evolve over time (Mattaini and Lowery, 2007). With a strong psychodynamic tradition to begin with, the approach gradually recognized the place of developmental processes over the life course, including environmental forces. Despite its legacy and evolution, psychosocial practice is not the most established or widely applied framework in social work. This is not unconnected to its vague conceptualization and varied definitions.

Definitions

Psychosocial work is an ambiguous concept in social work, though the term 'psychosocial' is commonly, and sometimes variously used to refer to a plethora of concepts and phenomena, such as psychosocial factors, psychosocial status, psychosocial support, psychosocial adjustments, psychosocial interventions, psychosocial contexts, psychosocial well-being, psychosocial effects, and psychosocial risks, with little consensus about what 'psycho' and 'social' entail. This is not unrelated to social work drawing knowledge heavily from the fields of psychology and sociology, and the new wave of psychosocial theory currently under construction in sociology, social policy, social research, and psychology. In sociology, the 'psycho' component of the term psychosocial often refers to psychoanalytically informed theory with its intellectual inheritance of critical theory. Though the 'social' component in social work traditionally referred to the context of interfamilial relations, it has more recently been extended to gender, culture, youth, and the reproduction and reinforcement of social stratifications from structural perspectives. (Cooper, 2000; Frost, 2008). This shift means greater focus on the social dimensions of structural issues, such as poverty, inequality, oppression, power in relationships (including relationships between the professionals and service users), moving away from the psychodynamic emphasis of the internal world and interpersonal processes.

In human development studies, which social work draws upon heavily, psychosocial development conventionally refers to change and stability in emotions, personality, and social relationships that can affect cognitive and physical functioning (Papalia et al., 2009). Deriving from the psychoanalytic tradition, Erik Erikson's (1950) seminal psychosocial development theory utilizes 'crisis' or conflicting tendencies in each of the eight stages as a major psychosocial theme throughout life. But critical psychologists are moving from psychosocial theory in a different direction beyond the psychodynamic tradition, and referring to social change to treat and prevent psychological problems (Frost, 2008; Kovel, 2011).

Currently, one of the most commonly quoted psychosocial definitions was developed by the World Health Organization (WHO). Since WHO (2003) defined health as a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity, it acknowledged that psychosocial interventions imply that social interventions have secondary psychological effects, and that psychological interventions have secondary social effects. Given this comprehensive definition, a wide range of international psychosocial programs are included. Some of the programs focus primarily on promoting and strengthening human rights and social justice issues, while others are more orientated toward community development; still others gravitate toward implementing therapeutic programs. Along this line, the Psychosocial Working Group (2003) (established in Oxford in the United Kingdom) reiterated that:

> The term 'psychosocial' is used to emphasize the close connection between psychological aspects of our experience (our thoughts, emotions and behaviors) and our wider social experience (our relationships, traditions, and cultures). These two aspects are so closely inter-twined in the context of complex emergencies that the concept of 'psychosocial well-being' is probably more useful for humanitarian agencies than narrower concepts such as 'mental health'. Interventions focusing narrowly on mental health concepts such as psychological trauma run the risk of ignoring aspects of social context that are vital to well-being. The psychosocial emphasis on social as well as psychological aspects of well-being also ensures that the family and community are fully brought into the picture in assessing needs. (p. 3)

The term psychosocial, regardless of its evolving definition, has offered an alternative paradigm that requires due consideration of both the 'psycho' and 'social' components, and does

not treat the two as an 'either/or' parallel. This calls for a shift to a more dynamic, interactive, and multilayered definition.

Psychosocial Work History, Theoretical Underpinnings and Leading Figures

Though Hamilton (1940) was among the first to use the term 'psychosocial,' Florence Hollis is widely recognized as the most authoritative in developing the psychosocial framework (Hollis, 1964; Woods and Hollis, 1990; Goldstein, 2008) and psychosocial therapy (Turner, 2009) in social work. Hollis' psychosocial framework emphasized the interface between the client and his or her environment. Traditionally, the framework integrated ecological and systems concepts with contemporary psychodynamic theories, including Freudian psychoanalysis, ego psychology, attachment and object relations theory, to understand the in-depth personalities and the coping capacities of individuals, couples, families, and groups in treatment (Woods and Hollis, 1990). Over time, it continues to evolve by incorporating psychological and social theories, as well as new perspectives related to gender, culture, and diversity. Grounded on this development, psychosocial casework is most widely associated with clinical social workers today (Goldstein, 2008). This development has generally continued to enjoy an important position in current social work and has flourished by emphasizing access and sensitivity to minority groups. But radical social work in the United Kingdom (Cooper, 2000) and structural social work that first emerged in Canada (Hick and Murray, 2013) have fiercely challenged psychosocial and clinical social work since the 1970s. They argued for mobilizing people as agents of social change and were strongly against encouraging conformity and adaptation to unequal power relations through casework (Cooper, 2000; Hick and Murray, 2013).

This long-standing argument on the focus of change in social work can be traced to the early roots of social work, and it is apparent that the development of the psychosocial framework can be better appreciated by reviewing the earliest origins of social casework (Kam, 2012). One of the earliest origins of social work is related to the legacies of the Charity Organization Society (COS) established in March 1870 in London, which aimed to provide principles to guide charitable giving and coordinate charitable organizations (Payne, 2005). COS members investigated and assessed whether applicants for assistance were deserving of help (Dominelli, 2004; Payne, 2005). This establishment rapidly spread to the United States especially in view of the immense social needs in the post–Civil War era. Mary Richmond could be considered one of the foremost pioneers in the United States, who advocated social change through improvement in the lives of individuals and families. She recognized that "mass betterment and individual betterment are inter-dependent; however, social reform and social case work of necessity progressing together" (Richmond, 1965: p. 25). Many proponents of clinical social work and psychosocial therapy have continued to defend their practice from a social justice perspective (e.g., McLaughlin, 2002; Parker, 2003) and have argued that a lopsided emphasis on the changing of environment may well end up in neglecting the psychosocial adjustments and well-being of poor and marginalized people. From the social justice perspective, Wakefield (1988) highlighted that psychological and emotional good health should not be restricted to the wealthy.

The other origin could be associated to settlements, which were founded as a movement for working-class education and the maintenance of moral Christian social behavior in poor neighborhoods. Along with relief of poverty, debate and campaigning on social policy was a part of the education activities in new cities (Payne, 2005). The first settlement was Toynbee Hall, founded by Canon Samuel Barnett in 1884 in the East End of London. As with the COS movement, the settlement idea was rapidly imported into the United States, and the first settlement was founded in New York in 1886. The famous Hull House in Chicago was formed in 1889, and one of its leaders, Jane Addams not only worked with the poor but also engaged in political action aimed at establishing new laws to protect them, and pioneered many social changes and reforms. She has been an important influence on the nature of social work (Payne, 2005). These crucial origins of social work has helped shape the contentious dual foci of change in social work today, which is unlikely to be resolved given its purpose to enhance adaptations among individuals, families, communities, or other collective entities within the ecological fields which they are embedded, consistent with social justice (Mattaini and Lowery, 2007).

In the global definition of social work jointly adopted by the International Federation of Social Workers (IFSW) and the International Association of Schools of Social Work (IASSW) in 2001 (each organization had worked on it for several years previously and adopted it separately in 2000), the dual foci of the interaction between the two separate and interactive entities of the environment and the individual was further augmented, in that "Social work intervenes at the points where people interact with the their environments" (IFSW, n.d.). Moreover, the definition's emphasis on the principles of human rights and social justice, articulated clearly the meaning and purpose of social work. Therefore, it is not the particular strategy of changing the individual and/or social reforms, but ultimately attaining the fundamental goal of promoting human right and justice that makes the profession unique, with due respect to varying social, cultural, and political contexts. Highlighting social justice in social work practice further enriches the dual-foci discussion as the profession has always involved both 'cause' (the organized effort to advocate for oppressed populations) and 'function' (a set of direct-practice activities carried out with the sanction of the larger society) (Lee, 1929). The distinction and balance between the two dimensions have ebbed and flowed and has been an ongoing tension within the profession since its beginning (Mattaini and Lowery, 2007). As Payne (2005) poignantly pointed out:

> Social workers will always be dealing with change…. The balance at any time and in any situation between social order and caring, between the state and the individual and between male administrative power and female caring power will be played out in their practice. (p. 247)

Between meeting individual needs and engaging in social changes, some social workers believe that the "middle ground, if one ever existed, is quickly receding. Social workers must decide on which side they stand" (e.g., Abramovitz, 1998: pp. 524–525). But this article suggests that the profession needs to duly consider the dual foci of 'psycho' and 'social' within the social justice framework to best serve the clients that include individuals, families, groups, and communities. The following sections review the application of psychosocial work in different settings in the United Kingdom, the United States, and international emergency situations, with focus on the distinction and interaction between the 'psycho' and 'social' dimensions.

Current Psychosocial Work Applications

Walker (2004), moving away from the arbitrary division of social work orientations, presented a modern psychosocial model of social work, particularly in working with children in the United Kingdom. His flexible model was community-oriented, service-user influenced, and took into account "the relationship between the external social world and the inner psychological experience of the individual that caused some citizens pain and suffering" (p. 163). He called for the consideration of both the inner and outer world of service users, especially those who had to contend with poverty, marginalization, social exclusion, and racism in established minority communities and those traumatized by war, ethnic persecution, and disasters. Intervention at one level included work carried out with individuals, families, and local networks to tackle problems that directly affected them. At another level, interventions also included advocacy and social action with community groups and other professionals and agencies to tackle problems affecting a range of people. Moreover, this modern psychosocial model offered social work a framework to engage both the professionals in the community as well as to collaborate with the service users. This holistic model envisaged that individual therapeutic or micro-interventions that focused on individual or family problems can also contribute to wider social transformation in neighborhoods, and conversely, in community practice that can also enable personal change and growth in individuals through social action and the fostering of cooperative capacity. However, community-based psychosocial and mental health care is not without its constraints and setbacks. Rose (1996) identified three broad defining features of community care that set it apart from earlier phases of mental health care. Sawyer (2008) referenced these features to highlight the constraints of community-based mental health services in Australia: (1) greater involvement with other agencies, which means that a higher level of communication and coordination is demanded among more stakeholders – this could be rare; (2) neoliberal norms of personal responsibility, choice, and empowerment, which implies that service users are encouraged to 'take responsibility' for themselves through learning about their illnesses and discouraging them from becoming too dependent on the given service; (3) and a culture of accountability and blame, in which someone must be held accountable. These three features have heightened mental health professionals' consciousness of risk and evoked concern to protect their professional integrity.

In the USA, psychosocial (or psychiatric) rehabilitation (United States of America Psychiatric Rehabilitation Association (USPRA), n.d.) promotes recovery, full community integration, and improved quality of life for persons who have been diagnosed with any mental health condition that seriously impairs their ability to lead meaningful lives. While its services are collaborative, person-directed, and individualized, the focus is on helping individuals develop skills and access the resources needed to increase their capacity to be successful and satisfied in the living, working, learning, and social environments of their choice. Taintor (2008) considered social workers as especially well qualified to work on the psychosocial multidisciplinary treatment teams because they have a mixture of skills and support for persons severely impaired by mental illness. Specifically, social workers are well trained to work with community organizations and to help service users obtain social support, and are trained to assist them through counseling and skills development. Moreover, social workers are used to working with their strengths, and advocate for their rights to be socially included and be treated as another individual in the society. Such an orientation has also been found to be relevant in promoting the strengths and rights of individuals affected by disasters and emergency settings recently, and the next section will tracks its recent developments with different client groups across cultures (Arntson and Knudsen, 2004; Kos and Zemljak, 2007; Ehrenreich, 2001; Lith, 2007; North Atlantic Treaty Organization (NATO), 2008).

Shifts in International Psychosocial Work: Promoting Resilience and Community Resources

Posttraumatic stress disorder (PTSD) has commonly been used as the predominant lens to implement interventions with individuals and populations affected by disasters. But there are several setbacks. First, PTSD is only one of the many psychological and social problems after a disaster. Secondly, only a small proportion of people may suffer from PTSD and acute distress, both immediately and in the long term, while the majority may have a range of other needs and situations, including those who may not need clinical interventions, such as physical and mental exhaustion, acute stress, and anxiety from being separated from family members (Arntson and Knudsen, 2004; Bourassa, 2009; Inter-Agency Standing Committee, 2007). Thirdly, psychological distress and trauma have a social and cultural dimension. The ensuing interventions recommended for individuals diagnosed with PTSD may unfortunately not be culturally appropriate. For instance, immediately after the Wenchuan earthquake in Sichuan of China in May 2008, there was a widespread dictum: Run away from dogs (that may be been infectious after the earthquake due to poor hygiene), run away from fire (due to the close proximity of makeshift housing), and run away from psychotherapists (who may do harm). This dictum was a culmination of many earthquake survivors being subjected to a 'second disaster' when international aid agencies, local psychologists and psychotherapists, academics

and well-meaning volunteers went around conducting culturally irrelevant psychotherapeutic assessments and interventions (Sim, 2009). Last but not least, privileging the pathological lens inevitably implies that social and political contexts and processes that occur and endure over extensive periods of time may be unduly neglected. Pupavac (2001) raised two specific concerns when psychosocial interventions in emergency settings focus on individuals' deficits and pathology. First, individual rights are being reconceptualized in psychological terms. This may lead to individuals being protected by official bodies, rather than protection from official bodies in some disaster or conflict situations. For example, understandably anxious, tired, or depressed survivors in post-disaster situations are being pathologized as unable to function by public health professionals. As a consequence of the pathologization of their condition, psychosocial intervention implicitly denies the capacity of populations to negotiate with the authorities about their well-being and even participate in the post-disaster reconstruction process. Second, psychosocial initiatives may potentially erode community and family cohesion. External and professional intervention may unwittingly undermine the sense of intimacy necessary for communal and family bonds. Moreover, such interventions encourage identification with and dependence on professional bodies thereby corroding local ties and institutions. This shift in focus away from pathology is also evident in psychosocial work with children internationally. Instead of focusing on remedial psychosocial interventions that put individual problems center stage, developmental approaches that capitalize on community resources are recommended. For example, the United Nation Children's Fund (UNICEF, 2007) argued that an important aim of all psychosocial support activities should be the strengthening of the capacity of families to nurture and support their children at the household level, as well as fostering of the social networks that would normally sustain mothers and fathers in emergencies. The United Nations High Commissioner of Refugees (UNHCR, 2004) in its guidelines on protection and care for refugee children indicated that the best way to promote the psychosocial well-being of children is to support their families and communities, and highlighted the role of schools in providing structure and predictability. International humanitarian and development agencies such as Save the Children (Arntson and Knudsen, 2004) and the International Rescue Committee (IRC, 2003) published similar recommendations, while taking into consideration the rights of children. Several arguments are apparent for a developmental psychosocial approach in work with children: only a small portion of young persons sustain serious psychological problems requiring specialized care in disaster and emergency situations; Western mental health therapies based on trauma and related mental pathological classifications are not suitable in culturally different settings; and the well-being of children often depends on secure family relationships and a stable environment with strong social support and social ties (Arntson and Knudsen, 2004; Duncan and Arntson, 2004).

Garcia del Soto (2008) highlighted the use of the term 'community approaches' in psychosocial work to contrast to individual approaches to mental health (p. 228). Such a shift in perspective highlights the need to promote the individual's trust about living in his or her community and society, apart from the challenge to promote individual resilience to maintain control of life in emergency situations. There is, therefore, a need to design and implement programs in ways that make sense to local populations by respecting their contexts. This requires not only consideration of socio-economic, political, generational, gender, spiritual, ethnic, and cultural practices, but also reinforcement of existing resilient resources within the community in the face of adversity (Duncan and Arntson, 2004; Lith, 2007; NATO, 2008; Tol et al., 2011). As Ehrenreich (2001) aptly reasoned, the individual and community approaches are not really in opposition to one another. The healing and rebuilding of the community is an essential underpinning for the healing of individuals and families, and the healing of individuals and families is necessary for the successful reconstruction of the community. In each case, the underlying principle is to encourage healing processes, in individuals, families, and communities. Thankfully, the perennial struggle between the two foci of change in social work has provided a fluid imagination for the profession, and nurtured an inclusive and flexible culture that allows the profession to include a rich and diverse set of skills, knowledge, and forms of practice (Yan et al., 2012).

The Inter-Agency Standing Committee (IASC) worked out a set of guidelines on mental health and psychosocial support in emergency settings (IASC, 2007) to facilitate an integrated approach to address the most urgent mental health and psychosocial issues in emergency situations. It addresses the social risk factors that affect mental health and psychosocial well-being that most mental health professionals seldom work in, including disaster management, human rights protection, general health, education, water and sanitation, food security and nutrition, shelter, camp management, community development, and mass communication. The set of guidelines recognizes that mental health and psychosocial problems in emergencies encompass far more than the experience of PTSD and not everyone has or develops significant psychological problems during an emergency. On the contrary, the guidelines capitalize on the fact that many people show resilience, that is, the ability to cope relatively well in situations of adversity. More importantly, attention is given to numerous interacting social, psychological, and biological factors that affect people in developing psychological problems or exhibiting resilience in the face of disasters. The IASC (2007) reiterated that a 'common error' in work on mental health and psychosocial well-being is to ignore the resources of different age, gender, socio-cultural context, and the emergency environment, and to focus solely on deficits – the weaknesses, suffering, and pathology – of the affected group (p. 4). Examples of potentially supportive social resources include families, local government workers, teachers, women's groups, youth clubs, and community planning groups, among many others. Affected communities may have economic resources such as savings, land, crops and animals; educational resources such as schools and teachers; and health resources such as health posts and staff; significant religious and spiritual resources including religious leaders, local healers, practices of prayer and worship, and cultural practices, such as burial rites. Increasingly, more social workers have used the individual-environment

perspective, due to its flexible and open nature, to develop antioppressive and culturally sensitive practices in international psychosocial work, particularly in emergency settings and humanitarian work (Bourassa, 2009; Ehrenreich, 2001; Frost, 2008). These shifts in perspectives and strategies correspond well with the dual-foci debate in social work, which allow social workers to be at the forefront of emergency and humanitarian aid at local and international levels. But the question of whether general psychosocial guidelines are in line with social work values, ethics and best practices remain to be researched in depth. Moreover, when applying or exporting them to different communities, both social work and general psychosocial principles may contradict the unique local characteristics. The need to engage with local groups and local professionals so as to be relevant and respectful cannot be overemphasized.

Emerging Trends and Directions

There is a recent call to end the debate between the two foci of 'psycho' and 'social' and to accept a two fold approach tailored to the specific needs and strengths of individuals in the respective contexts (Loughry and Eyber, 2003). While developmental responses are regarded as the most appropriate as generic types of interventions, curative methods are said to be useful to a smaller group of people, such as children who are extraordinarily affected, or those who have not benefited from generic interventions. Specifically, 'a stepped model' whereby the needs of the individuals are assessed according to levels may be useful. For example, Save the Children adopted the approach of disaggregating children in emergencies to three groups according to needs: the severely affected group, the 'at-risk' group, and the generally affected group (Arntson and Knudsen, 2004). The severely affected group is often expected to be a small percentage of children whose psychosocial functioning may be severely compromised. They would benefit from individualized treatment. The at-risk group is expected to be larger than the severely affected one and consists of those who have experienced severe losses and disruptions. They may be significantly distressed, despairing, and hopeless, but their social and psychological ability to function has not yet been overwhelmed. Without timely community and social support, their psychosocial functioning may be at risk of deteriorating. Short-term crisis interventions and support groups may be useful interventions for this group. The generally affected group refers to the broadest segment of the population who may not have been directly affected though they may be disturbed or stressed. Over time, this group will become better even without direct support. Prevention, structured activities, and normalization may suffice for this group of people. This stepped model is useful in meeting the varying needs of affected individuals, and at the same time helps to engage different professionals or provide different types of services according to the needs of affected groups. In the aftermath of a population's exposure to severe stressors, the WHO advocated a focus on medium- and long-term development of community-based and primary mental health care services and social interventions as they are much more effectively implemented over a protracted time during subsequent years. Unfortunately, impetus and funding for mental health programs is highest during or immediately after acute emergencies and not for medium- and long-term programs (WHO, 2003).

When working with children, young people, women, families, and special groups in varying contexts, the focus of change should not be that of 'either/or,' but 'both–and' according to the context of needs, as feminist social work has promulgated (Brandwein, 1986; White, 2006). Dominelli (2002) has updated 12 feminist principles as relevant to practice and evident in feminist social work. The following are a few principles that echo the discussion on the dual focus of change in psychosocial work:

- recognizing the diversity of women;
- valuing women's strengths;
- locating individual women in their social situations and acknowledging the interconnections between the individual and collective entities relevant to them;
- acknowledging that the principle, 'the personal is political' is relevant at macro-, meso-, and micro-level of practice;
- redefinining private woes as public issues;
- recognizing that women's individual problems have social causes and addressing both levels in each intervention; and
- looking for collective solutions to individual problems (Dominelli, 2002: pp. 162–163).

These feminist social work principles help to reconcile the tensions between the individual and social by suggesting that the personal, individual, and private experiences are a legitimate focus for, and site of, political action (Orme, 2013). While this duality of focus and the mission to pursue human rights and social justice set social work apart from other professions, 'psychosocial work' can be considered a term that encapsulates the dual strategies of the profession. It also reflects the need for social workers to continually consider the different foci of change when working different groups.

Critical Issues in Psychosocial Work

Psychosocial work has rather limited quantities of and in quality research. Tol et al.'s (2011) comprehensive review of mental health and psychosocial support in humanitarian setting concludes that there is little link between practice and research. Specifically, research is infrequently implemented. Existing outcome assessments have focused predominantly on PTSD, though there is still disagreement about the public health value of this disorder. Psychoeducation, structured social activities, and counseling had been frequently practiced but were included in few evaluation studies that often presented mixed results. However, limited and mixed results do not necessarily indicate that these approaches are not effective. Rather, these point to the need for more rigorous examination. In the quest for more evidence, psychosocial workers and researchers ought to be mindful of the impact of evaluation and research on the people affected by disasters and crises that they endeavor to serve and understand. Some evaluation methods may not be appropriate for some socio-cultural contexts (Ehrenreich, 2001; Kos and Zemljak, 2007; Lith, 2007; WHO, 2003). For example, in India, Pakistan,

and Jordan, local workers trained in North America or Europe could not translate the term 'psychosocial' into Hindi, Urdu, or Arabic for monolingual populations because theories from other fields such as religion or political science prevail regarding the individual's relationship to society (Aggarwal, 2011). And, the evaluation of psychosocial work, within a short frame of time given the complexity of contexts and issues involved, may be arduous.

Conclusion

Psychosocial work embodies the long-standing dual-foci debate in the social work profession that has instrumentally allowed the highest flexibility for the profession to negotiate its roles and tasks with different stakeholders, domains, and contexts. Due consideration of the balance offered by the dual-foci approach has opened up possibilities for the profession to include a diverse set of values, knowledge, and skills of practice. Careful deliberation of the interaction between the 'psycho' and 'social' dimensions highlight the need to consider salient perspectives in social work, including but not limited to 'both–and' and 'either/or,' resilience and pathology, past and present, the inner and outer worlds of service users, foreign and local contexts, the individual and community resources of different service users, and the settings they are in. When carrying out psychosocial work, social workers who promote human rights and social justice would reveal a marked difference from other professional groups.

See also: Critical Social Work Practice; Mental Health; Radical Social Work.

Bibliography

Abramovitz, M., 1998. Social work and social reform: an arena of struggle. Social Work 43 (6), 512–526.

Aggarwal, N.K., 2011. Defining mental health and psychosocial in IASC Guidelines. Intervention 9 (1), 21–25.

Arntson, L., Knudsen, C., 2004. Psychosocial Care and Protection of Children in Emergencies. Save the Children, UK.

Bourassa, J., 2009. Psychosocial interventions and mass populations: a social work perspective. International Social Work 52 (6), 743–755.

Brandwein, R., 1986. A feminist approach to social policy. In: Van Den Bergh, N., Cooper, L.B. (Eds.), Feminist Visions for Social Work. NASW, Maryland, pp. 250–261.

Cooper, A., 2000. Psychosocial perspectives. In: Davies, M. (Ed.), The Blackwell Encyclopedia of Social Work. Blackwell Publisher, Oxford, pp. 276–277.

Dominelli, L., 2002. Feminist Social Work Theory and Practice. Palgrave, New York.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity, Cambridge, UK.

Duncan, J., Arntson, L., 2004. Children in Crisis: Good Practices in Evaluating Psychosocial Programming. Save the Children, UK.

Ehrenreich, J.H., 2001. Coping with Disasters: A Guidebook to Psychosocial Intervention, revised ed. http://www.keithwilliamslawgroup.com/library/CopingWithDisaster.pdf (accessed 03.07.13.).

Erickson, E.H., 1950. Childhood and Society. Norton, New York.

Frost, L., 2008. Why teach social work students psychosocial studies? Social Work Education: The International Journal 27 (3), 243–261.

Garcia del Soto, G., 2008. Psychosocial community approaches and practices in Latin America. Intervention 6 (3/4), 228–231.

Goldstein, E.G., 2008. Psychosocial framework. In: Mizrahi, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed. Oxford University Press, New York, pp. 462–467.

Hamilton, G., 1940. Theory and Practice of Social Casework. Columbia University Press, New York.

Hick, S.F., Murray, K., 2013. Structural social work. In: Gray, M., Webb, S.A. (Eds.), Social Work: Theories and Methods, second ed. SAGE, California, pp. 110–125.

Hollis, F., 1964. Casework: A Psychosocial Therapy. Random House, New York.

Inter-Agency Standing Committee (IASC), 2007. IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings. IASC, Geneva.

International Federation of Social Work (IFSW), n.d. Global Standards. Retrieved from: http://ifsw.org/policies/global-standards/ (accessed 19.06.13.).

International Rescue Committee (IRC), 2003. Guiding Principles for Aiding and Protecting War Affected Children and Youth. IRC, New York.

Kam, P.K., 2012. Back to the 'Social' of Social Work: Reviving the Social Work. International Social Work. Epub ahead of print 5 October 2012. http://dx.doi.org/10.1177/0020872812447118. The on-line version of this article can be found on http://isw.sagepub.com/content/early/2012/10/05/0020872812447118.full.pdf+html.

Kos, A.M., Zemljak, B., 2007. Psychosocial support for children, families and teachers in Iraq. Intervention 5 (2), 150–158.

Kovel, J., 2011. Erik Erikson's psychohistory. In: Parker, I. (Ed.), Critical Psychology. Routledge, London, pp. 248–257.

Lee, P.R., 1929. Social work: cause and function. In: Proceedings of the National Conference of Social Work. Columbia University Press, New York, pp. 3–20. The on-line version of this article can be found on http://quod.lib.umich.edu/n/ncosw/.

Lith, B.K.V., 2007. Psychosocial interventions for children in war-affected areas: the state of the art. Intervention 5 (1), 3–17.

Loughry, M., Eyber, C., 2003. Psychosocial Concepts in Humanitarian Work with Children: A Review of the Concepts and Related Literature. National Academies Press, Washington, DC.

Mattaini, M.A., Lowery, C.T., 2007. Foundations of social work practice. In: Mattaini, M.A., Lowery, C.T. (Eds.), Foundations of Social Work Practice, fourth ed. NASW Press, Washington, DC, pp. 3–30.

McLaughlin, A.M., 2002. Social work's legacy: Irreconcilable differences? Clinical Social Work 30 (2), 187–198.

North Atlantic Treaty Organization (NATO), 2008. Psychosocial Care for People Affected by Disasters and Major Incidents. NATO, Belgium. The on-line version of this article can be found on http://www.coe.int/t/dg4/majorhazards/ressources/virtuallibrary/materials/Others/NATO_Guidance_Psychosocial_Care_for_People_Affected_by_Disasters_and_Major_Incidents.pdf.

Orme, J., 2013. Feminist social work. In: Gray, M., Webb, S.A. (Eds.), Social Work: Theories and Methods, second ed. SAGE, California, pp. 87–98.

Papalia, D.E., Olds, S.W., Feldman, R.D., 2009. Human Development. McGraw-Hill Higher Education, New York.

Parker, L., 2003. A social justice model for clinical social work practice. Affilia 18 (3), 272–288.

Payne, M., 2005. The Origins of Social Work: Continuity and Change. Palgrave Macmillan, New York.

Psychosocial Working Group, 2003. Psychosocial Interventions in Complex Emergencies: A Framework for Practice. Retrieved from: http://www.forcedmigration.org/psychosocial/papers/A%20Framework%20for%20Practice.pdf (accessed 19.06.13.).

Pupavac, V., 2001. Therapeutic governance: psycho-social intervention and trauma risk management. Disasters 25 (4), 358–372.

Richmond, M.E., 1965. Social Diagnosis. First Free Press, Toronto, ON.

Rose, N., 1996. Psychiatry as a political science: advanced liberalism and the administration of risk. History of the Human Sciences 9 (2), 1–23.

Sawyer, A.M., 2008. Risk and new exclusions in community mental health practice. Australian Social Work 61 (4), 327–341.

Sim, T., 2009. Crossing the river stone by stone: developing an expanded school mental health network in post-quake Sichuan. China Journal of Social Work 2 (3), 165–177.

Taintor, Z.C., 2008. Psychosocial and psychiatric rehabilitation. In: Mizrahi, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed. Oxford University Press, New York, pp. 458–462.

Tol, W.A., Barbui, C., Galappatti, A., Silove, D., Betancourt, T.S., Souza, T., Golaz, A., van Ommeren, M., 2011. Mental health and psychosocial support in humanitarian settings: linking practice and research. The Lancet 378 (9802), 1581–1591.

Turner, F.J., 2009. Psychosocial therapy. In: Roberts, A.R. (Ed.), Social Workers' Desk Reference, second ed. Oxford University Press, New York, pp. 248–253.

United Nations High Commissioner for Refugees (UNHCR), 2004. Refugee Children: Guidelines on Protection and Care. UNHCR, Geneva.

United Nations Children's Fund (UNICEF), 2007. Psychosocial Support of Children in Emergencies (field version). UNICEF, Geneva.

United States of America Psychiatric Rehabilitation Association, n.d. About the U.S. Psychiatric Rehabilitation Association. https://netforum.avectra.com/eWeb/StartPage.aspx?Site=USPRA (accessed 19.06.13.).

Walker, S., 2004. Community work and psychosocial practice – chalk and cheese or birds of a feather? Journal of Social Work Practice: Psychotherapeutic Approaches in Health, Welfare and the Community 18 (2), 161–175.
Wakefield, J.C., September 1988. Psychotherapy, distributive justice, and social work part 2: psychotherapy and the pursuit of justice. Social Service Review 62 (3), 353–382.
White, V., 2006. The State of Feminist Social Work. Routledge, Oxon.
Woods, M.E., Hollis, F., 1990. Casework: A Psychosocial Therapy, fourth ed. McGraw-Hill, New York.
World Health Organization (WHO), 2003. The World Health Report 2003: Shaping the Future. WHO, Geneva.
Yan, M.C., Tsui, M.S., Chu, W.C.K., Pak, C.M., 2012. A profession with dual foci: is social work losing the balance? China Journal of Social Work 5 (2), 163–172.

Relevant Websites

http://www.who.int/mental_health/emergencies/guidelines_iasc_mental_health_psychosocial_june_2007.pdf – Inter-Agency Standing Committee (IASC), 2007. IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings.
http://www.iassw-aiets.org/ – International Association of Schools of Social Work (IASSW).
http://ifsw.org/ – International Federation of Social Work (IFSW).
http://www.rescue.org/about – International Rescue Committee (IRC).
http://www.interventionjournal.com/index1.html – Intervention: International Journal of Mental Health, Psychosocial Work and Counselling in Areas of Armed Conflict.
http://www.forcedmigration.org/psychosocial/ – Psychosocial Working Group.
http://www.unicef.org/protection/57929_57998.html – United Nations Children's Fund (UNICEF) Child Protection from Violence, Exploitation and Abuse – Psychosocial Support and Well-being.
http://www.unhcr.org.hk/unhcr/en/our_work.html – United Nations High Commissioner for Refugees (UNHCR).
http://www.who.int/whr/2003/en/ – World Health Organization – The World Health Report 2003: Shaping the Future.

Psychosomatic Medicine

Santosh K Chaturvedi, National Institute of Mental Health and Neurosciences, Bangalore, India
Soumya Parameshwaran, Kasturba Medical College, Mangalore, India

© 2015 Elsevier Ltd. All rights reserved.
This is article is a revision of the previous edition article by D. Oken, volume 18, pp. 12452–12457, © 2001, Elsevier Ltd.

Abstract

Psychosomatic medicine focuses on holistic care, involving the biopsycho-social aspects and mind–body relationships. It involves the role of psychosocial factors in etiopathogenesis and maintenance of medical conditions, and the importance of holistic interventions. The term psychosomatic medicine has undergone transformation following the growing rejection of psychodynamic and psychoanalytic schools of thought and replaced by contemporary concepts of consultation liaison psychiatry, behavioral medicine, and general hospital psychiatry. The role of stress and neurohumoral activity still remain of importance, and behavioral and alternative medicine gain significance. Diagnostic criteria for psychosomatic research have been proposed and methodologically stringent studies are needed to validate the concepts of psychosomatic conditions.

Concepts

The term 'Psychosomatic' has been used in several ways and various definitions have been put forth. Not only has this term been in dispute but the field of psychosomatic medicine itself has often been rife with ambiguity and controversy, due to various reasons.

Psychosomatic medicine provides a comprehensive framework for a holistic (biopsycho-social) consideration of patient care encompassing: (1) the role of psychosocial factors in affecting individual vulnerability to all types of diseases; (2) the interaction between psychosocial and biological factors in the course and outcome of disease; and (3) the application of psychological therapies to the prevention, treatment, and rehabilitation of physical illness (Fava and Sonino, 2000).

Modern psychosomatic medicine originated in the developed Western world, but it is equally relevant to all nations. In the areas of clinical practice and the training of health workers, psychosomatic postulates regarding comprehensive care of patients and the behavioral aspects of prevention of disease are just as applicable in developing as in developed countries.

Psychosomatic medicine arose from the psychodynamic theory and practice in the twentieth century, and referred to the concept that psychological factors are important in the etiology and course of certain physical/medical disorders. It became influential in the 1930s in the United States and in the German-speaking countries. It continued to flourish in parts of Europe, as a medical specialty, which was different from both internal medicine and psychiatry. In other parts of the world, it did not become popular after the 1960s. Outside German-speaking countries, psychosomatic theory is not popular and not seen as useful. The emphasis on the interaction of physical and psychological process is beneficial, but a simplistic assumption about possible psychological factors affecting physical disorders is a bit difficult to accept and understand. Psychosomatic medicine has had a lasting legacy within medicine in promoting greater understanding of psychological aspects of physical symptoms and disorders (Mayou, 2008).

The psychosomatic approach has explored linkages between psychological and neurophysiological processes influencing health and disease. The largest share of current psychosomatic research focuses on mediation by the brain from relationships between behavioral traits and disease state. The advent of the cognitive neuroscience methods has ushered in a new era in which the brain basis of mental states relevant to psychosomatic medicine can begin to be identified. Investigators in psychosomatic medicine should support their theoretical models and empirical investigations by including further consideration of the brain and mind–body interactions (Lane et al., 2009).

Alternative Terminology

Psychosomatic medicine has now been replaced by terms like consultation liaison (C-L) psychiatry, behavioral medicine, health psychology, and general hospital psychiatry.

Behavioral medicine is an interdisciplinary field concerned with the development and integration of behavioral and biomedical science knowledge and techniques relevant to health and illness and the application of this knowledge and these techniques to prevention, diagnosis, and rehabilitation.

C-L psychiatry is the study, practice, and teaching of the relation between medical and psychiatric disorders. The psychiatrists as consultants play many roles: interviewer, therapist, teacher, physician as part of the team, diagnosis, and treatment. Principal contribution to medical treatment involves a comprehensive analysis of a patient's response to illness, psychological and social resources, coping style, and psychiatric illness.

Consultation psychiatry refers to individuals trained as psychiatrists providing consultations to colleagues from other medical specialties upon receiving a specific referral. Occurring both in inpatient and outpatient settings, the consultation is

limited to evaluating the referred patient for psychiatric symptoms and suggesting a plan of management. It involves limited interaction with the referring physician. The psychiatrist is usually not involved in the day-to-day care of the patient.

Liaison psychiatry is a much closer integration of psychiatric services with the medical services. Typically occurring in the inpatient setting, it involves a physician or psychiatrist (the C-L psychiatrist) being a part of the medical team and seeing all patients under their care irrespective of presence of psychiatric symptoms or referral.

C-L psychiatry's growth continued as a specialization in cancer units, ICUs, burns, dialysis, obstetrics and gynecology, pediatrics, cardiology, neurology units, reaching nonreferred populations by screening and triage methods. Liaison psychiatry encompasses clinical, teaching, and research activities of psychiatrists in the nonpsychiatric divisions of general hospitals. The Liaison psychiatrist seeks to enhance the psychological status of all the medical patients, aims to participate in case detection rather than waiting for referral and provides educational programs (Chaturvedi and Tipirneni, 2005; Chaturvedi and Jaideep, 2007).

There has an upsurge of interest in another related discipline, subsumed under the rubric of mind–body interrelationships, sharing the same holistic, biopsycho-social connotation with psychosomatic medicine. However, the concept of mind–body medicine is inextricably linked to that of alternative therapy (Pert et al., 1998). Psychosomatic medicine continues to manage patients in medical settings that cut across a continuum from primary care to complex medical and surgical centers. The clinical problems containing a medical–psychiatric interface offer a clinical challenge and an area for new knowledge and better interventions (Wise, 2008).

Criticisms of the Term 'Psychosomatic'

The term 'psychosomatic disorder' was strongly criticized by several psychosomatic researchers. Lipowski (1986) criticized the obsolete notion of psychogenesis, since it was incompatible with the doctrine of multicausality, which constitutes a core postulate of current psychosomatic medicine. Engel (1967) wrote that the term 'psychosomatic disorder' was misleading, since it implied a special class of disorders of psychogenic etiology and, by inference, the absence of a psychosomatic interface in other diseases. On the other hand, he viewed reductionism, which neglected the impact of nonbiological circumstances upon biological processes, as a major cause of mistreatment (Fava, 1992). Tinetti and Fried (2004) suggested abandoning disease as the primary focus of medical care. When disease became the focus of medicine in the past two centuries, the average life expectancy was 47 years and most clinical encounters were for acute illness. Currently, the life expectancy in many developed countries is much higher and most of clinical activities are concentrated on chronic disease or nondisease-specific complaints. As time went by, psychosomatics gained a rather broader connotation and refers to the study and treatment of those conditions, which are related to the interaction of psychic and somatic phenomenon. The various categories of clinical phenomenon included in psychosomatic medicine thus defined, are enumerated below:

1. Psychological reactions to physical illness, surgical procedures, radiation therapy, etc., like emotional and behavioral responses to physical illness.
2. Disturbances resulting from disorders affecting brain function, acute or chronic organic brain syndromes.
3. Medical complications of maladaptive behavior, like alcohol and drug abuse, obesity, eating disorders like anorexia and bulimia.
4. Emotional disorders manifested by somatic symptoms with no organic basis or functional somatic symptoms or medically unexplained symptoms, e.g., somatoform disorders, neurotic disorders, masked depression, conversion disorder, hypochondriasis, chronic pain, chronic fatigue, etc.
5. Physiologic concomitants of emotional states.
6. Psychosomatic diseases, the classical psychosomatic disorders like hypertension, bronchial asthma, rheumatoid arthritis, ulcerative colitis, neurodermatitis, duodenal ulcer.
7. Physical illnesses presenting as emotional or psychological problem, e.g., myxedema, thyrotoxicosis.
8. Psychological adverse drug reactions to drugs used for medical illnesses, e.g., steroids, antihypertensives.

Both psychophysiological and psychosomatic disorders are associated with prolonged psychological stress. Some of the important psychosomatic disorders, psychophysiological symptoms and functional somatic symptoms areas follows:

Physiological somatic system	*Psychosomatic syndrome*	*Psychophysiologic symptom*	*Functional symptom*
1. Cardiovascular	Coronory heart disease Essential hypertension	Migraine headache Vasovagal syndrome Tachycardia Palpitations	Noncardiac chest pain
2. Gastrointestinal	Peptic ulcer Ulcerative colitis Mucous colitis	Irritable colon Psychogenic vomiting Aerophagy Abdominal cramps	Irritable bowel syndrome
3. Genitourinary pain	Dysmenorrhoea Premenstrual syndrome Female 'Dhat' syndrome	Psychogenic impotence Dyspareunia Anorgasmia	Chronic pelvic
4. Musculoskeletal	Rheumatoid arthritis Low-back pain	Psychogenic torticollis Muscle tension, pain, fatigue, tremor	Fibromyalgia Repetitive strain injury
5. Respiratory	Bronchial asthma	Hyperventilation Bronchospasm Hiccup (psychogenic), yawning, air hunger	Hyperventilation syndrome

Continued

Physiological somatic system	*Psychosomatic syndrome*	*Psychophysiologic symptom*	*Functional symptom*
6. Skin	Urticaria Dermatitis Eczema	Pruritus Hyperhidrosis	
7. Miscellaneous	● Chronic pain syndromes ● Obesity ● Anorexia ● Bulimia		

Other functional somatic symptoms, by specialty, are as follows:

Dentistry – atypical facial pain, temporomandibular joint dysfunction, burning mouth syndrome.
Ear nose and throat – Globus syndrome.
Neurology – tension headache.
Nonallied FSS – Gulf war syndrome, chronic whiplash, sick-building syndrome, candidiasis hypersensitivity, organophosphate poisoning, and multiple chemical hypersensitivity (Page and Wesseley, 2008).

The diagnosis of psychophysiological and psychosomatic disorders is done by the following:

1. The presence of symptoms and signs characteristic of a particular syndrome. Appropriate interview technique should elicit psychological symptoms in first contact.
2. The exclusion of organic disease through appropriate physical examination and laboratory procedures.
3. The presence of chronic stress, tension, episodes of anxiety, and/or stressful life situations, like exit life events, separation or death.

Diagnosing Psychosomatic Disorders

Psychosomatic disorders must be diagnosed precisely and must be distinguished from physical illnesses, somatoform disorders, hypochondriasis, conversion disorders, and other psychiatric illnesses with somatic presentations.

Certain characteristics of psychosomatic patients are as follows:

1. They present to physician, invariably.
2. May have inherited or early acquired (through illness or conditioning) organ or system vulnerability.
3. Evidence of chronic and sustained emotional or autonomic arousal on response to stress.
4. Precipitating life events could be identified.
5. Certain personality traits, like type-A personality for coronary heart diseases.
6. Difficulty in expressing feelings and emotions, have affective inhibition and alexithymia.
7. Poor psychological mindedness or sophistication.
8. Show abnormal illness behavior in the form of disease conviction, somatic preoccupation, somatic attribution, somatic focusing, nonresponse to reassurance and a need to deny psychological component.

Classification of Psychosomatic Disorders

The third edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-III) introduced somatoform disorders as a speculative diagnostic category. It remained in the successive versions of DSM-III-R and DSM-IV. Somatic syndromes that were neither explained by a general medical condition nor clearly associated with depressive or anxiety diagnoses were combined to create a new category of somatoform disorders. In the subsequent revisions of DSM-III, minor changes to the definitions of these disorders were made. There was also one major change, which was the introduction of undifferentiated somatoform disorder. The addition of this new diagnosis was necessary to categorize the large number of patients who, although clearly ill, did not fall within the existing somatoform categories. As a result, the category of somatoform disorders changed between DSM-III and DSM-IV from being a small grouping of relatively uncommon conditions to a general category covering a wide range of illnesses. The tenth revision of International Classification of Diseases (ICD-10) was developed in parallel with DSM-IV and also includes a similar, but not identical, category of somatoform disorders embedded within the broader "Neurotic, stress-related, and somatoform disorders 'category'" (Mayou et al., 2005).

Proposals for DSM-V/ICD-11

The DSM-IV-TR category of somatoform disorder introduced ambiguity into the classification of 'Specialty' syndromes such as irritable bowel syndrome (IBS), and fibromyalgia (FM). Such syndromes are characterized by medically unexplained physical symptoms and are found in many different medical specialties, such as IBS in gastroenterology, and FM in rheumatology.

The unifying principle for the category, the feature that the somatic symptoms are 'medically unexplained' is highly unsuitable as the unifying principle for a category of mental disorder. Hence, elimination of any reference to the physical symptoms of being 'medically unexplained' has been proposed in the forthcoming classifications (Sykes, 2012). Further, if a positive psychological criterion is provided, this will overcome the limitations of a negative diagnosis based on the feature of the physical symptoms being medically unexplained.

The diagnostic category of somatoform disorders is currently under revision for DSM-5 and ICD-11. In the current diagnostic draft of DSM-5, somatic symptom disorder includes the former diagnoses of somatization disorder, undifferentiated somatoform disorder, pain disorder, and partly hypochondriasis. It introduces two major changes in comparison with DSM-IV: First, the criterion of somatic symptoms is no longer limited to medically unexplained symptoms, thus eliminating the contentious and unreliable distinction between medically explained and medically unexplained symptoms as well as an implicit mind–body dualism. Second, positive psychological diagnostic criteria are added to the diagnosis,

which covers excessive thoughts, feelings, and behaviors related to these somatic symptoms or associated health concerns. Disproportionate and persistent thoughts about the seriousness of one's symptoms, a persistently high level of anxiety about health or symptoms, and excessive time and energy devoted to these symptoms or health concerns are other descriptions, which should persist for more than 6 months (Voigt et al., 2012).

Historical Aspects

The word 'psychosomatic' was first used in medical literature by Heinroth in 1818, though it may not have meant what it came to be considered a century later. Later, in 1922, Felix Deutsch introduced the term psychosomatic medicine, and in 1935, Dunbar described psychosomatic interrelationships in medical practice. This resulted from the confluence of two concepts having an ancient tradition in Western thought and medicine: those of psychogenesis of disease and of holism (Lipowski, 1986). The concept was formalized in 1939 by the editors of the journal *Psychosomatic Medicine*, a journal which is still a popular one dealing with 'biopsycho-social' and more aspects of health.

The idea of psychogenesis characterized the first phase of the development of psychosomatic medicine (1930–60) and resulted in the concept of psychosomatic disease (a physical illness believed to be caused by psychological factors, such as peptic ulcer). Diseases regarded previously as obscure of origin, had found an explanation and the prospects of treating them by psychotherapy appeared exciting. However, these expectations did not survive the test of scientific evidence and led to profound crisis (Wittkower, 1974). The identification of helicobacter pylori and its association with gastric and duodenal ulcers dumped the psychosomatic and psychogenic etiology of peptic ulcers!

Engel developed a multifactorial 'biopsycho-social' model of illness. It allows illness to be viewed as a result of interacting systems at the cellular, tissue, and organismic, interpersonal, and environmental levels. As a result, the study of every disease must include the individual, his body and his surrounding environment as essential components of the total system (Engel, 1967). Engel formulated a sharp criticism of the concept of disease as what can be understood and recognized by the physician, and proposed a unified concept of health and disease that was later elaborated upon within the biopsychosocial approach. Lipowski (1977) gave an invaluable contribution in setting the scope, mission, and methods of psychosomatic medicine. He identified three interrelated facets:

1. It is a scientific discipline concerned with the study of the relationships of biological, psychological, and social determinants of health and disease.
2. It embodies a holistic approach to the practice of medicine.
3. It encompasses C-L psychiatry.

Kissen (1963) provided a better specification of the term psychosomatic. He clarified that the relative weight of psychosocial factors may vary considerably from one individual to another within the same illness and underscored the basic conceptual flaw of considering diseases as homogeneous entities. Instead of asking, 'Which psychological factors give rise to which illnesses?' Kissen suggested asking, "who are the patients within a given illness population for whom psychosocial variables are of primary significance?"(Fava and Sonino, 2000).

Psychosomatic Disorders

Psychosomatic disorders or conditions that involve a major interaction of the mind–body interrelationships, occurring in different physiological systems can be described as follows.

Cardiovascular System

Psychocardiology is the study and practice of psychiatric aspects of cardiac disorders. Hypertension and coronary heart diseases are the common psychosomatic disorders seen in medical clinics and are frequently referred to psychiatrists for management of the psychological distress and psychiatric reactions. Hypertension may result from a variety of factors including psychological and emotional stresses. Though in the persistent and the more pathological varieties, the evidence of psychological factors is not quite so obvious and has given rise to considerable controversy.

Coronory heart disease (CHD, Ischemic heart disease (IHD)) like Angina Pectoris and Myocardial infarction are the two commonly encountered coronary heart disorders, where emotional and psychological factors have been implicated. Many studies have isolated a type-A behavior pattern, which has an overt style and has features of hard driving, competitiveness, impatience, sense of time urgency, and intensely committed to vocational goals. Other risk factors that have been identified in major collaborative studies are: high serum cholesterol, elevated systolic blood pressure, and cigarette smoking. Common reactions after a myocardial infarction include anxiety, depression, worry about future capabilities and sexual functioning, and hypochondriacal features.

Advancing technologies of cardiac catheterization, bypass surgeries, angioplasty, artificial heart valve surgery, atherectomy, cardiomyoplasty, stent procedures, transmyocardial revascularizations are likely to be associated with anxiety, health worry, and psychological reactions in the patients and their caregivers, which would need to be addressed by the C-L psychiatrists.

Gastrointestinal System

The most common focus for psychophysiologic and psychosomatic disorders was believed to be the gastrointestinal tract since it was thought to be associated with the symbolic expression of emotion through out life, starting from nursing in infancy and proceeding to toilet training. Medically unexplained symptoms related to the gut have been the most challenging. Peptic ulcer and dyspepsia have multifactorial etiology and the role of psychosocial factors is quite important. In Ulcerative Colitis, stressful psychological conditions had been observed as potential etiologic agents in most cases especially as separation, loss of love object, painful life situations, and life dissatisfaction. The principal etiologic factors are not known, however, psychological factors are thought to be important.

Respiratory System

The common cold, allergic rhinitis, and many cases of pharyngitis are thought to have a psychosomatic etiology. Bronchial asthma is the most serious psychosomatic disorder of this system. Bronchial Asthma is one of the commonest and most challenging of the psychosomatic disorders and has attracted considerable attention to its psychological and physical components. The specific role of psychological factors and psychotherapeutic interventions is debatable. The hyperventilation syndrome is found both in children and adults and may be initiated as a voluntary act, though in many instances it is a result of anxiety. It may not be recognized as such, for most do not present with respiratory distress, but with the physical sequalae of the hyperventilation. Other respiratory conditions considered as psychosomatic include psychogenic dyspnoea, allergic rhinitis, and the common cold.

Oncology

Psychooncology is the study of psychological, social, and environmental factors in relation to cancer, which have been implicated in the development, course, terminal phases, and consequences of cancer. This has been perhaps the fastest growing specialty on almost all aspects of cancer biology, disease, and treatment. The journal Psycho Oncology run by the International Psycho Oncology Society is a rich source of developments in this field.

Nephrology

Psychonephrology deals with the study of psychological and emotional aspects of renal diseases. Patients who develop end stage renal disease are not as a group predisposed to psychopathology and have low incidence of severe psychiatric disorder. They may develop maladaptive styles of coping with illness and thus present difficult problems for the staff such as noncompliance.

Dermatology

Psychodermatology is the study of psychosocial factors related to skin conditions like dermatitis, eczema, and urticaria, which are considered to have psychological and emotional factors involved in their causation and outcome.

Obstetrics and Gynecology

Psycho gynecology involves the study of numerous conditions, which have important mind–body interrelationships like premenstrual syndrome, postpartum disorders, pelvic pain, and perimenopausal disorders.

Somatizing Patient in Medical Clinics

All patients presenting with somatic complaints or chronic pain to physicians do not have physical disease. In fact, a sizable proportion of the medical practices of most physicians consist of patients for whom no objective evidence of a pathophysiologic process can be established. The prevalence of somatizing behavior is highest when the type of medical practice studied includes patients who complain of vague, poorly differentiated, or nonspecific symptoms. Medically unexplained/functional somatic symptoms are considered to have an unidentified interplay of body–mind mechanisms, with more role of the mind on the body, than vice versa!

Management of Psychomatic Disorders

Assessment

A considerable body of evidence has accumulated in psychosomatic medicine related to concepts such as quality of life, stressful life events, somatization, illness behavior, and personality, but it has not resulted in methods whereby psychosocial aspects of medical disease could be routinely assessed. The Diagnostic Criteria for Psychosomatic Research (DCPR) has been proposed to deal with the lack of specificity of earlier descriptions. The DCPR attempts to translate psychosocial characteristics observed in various medical diseases. The DCPR have now undergone extensive validation and have been found to be more suitable than DSM criteria in describing psychological distress in a variety of medical settings (Porcelli and Sonino, 2007).

The DCPR are a set of 12 psychosomatic syndromes whose prognostic role in the development, course and outcome of physical diseases, regardless of their 'organic' or 'functional' nature, was documented by a large body of literature (Fava et al., 1995). Eight of them refer to the concept of abnormal illness behavior: persistent somatization, functional somatic symptoms secondary to a psychiatric disorder, conversion symptoms, anniversary reaction, disease phobia, thanatophobia, health anxiety, and illness denial. The four syndromes of alexithymia, type A behavior, demoralization, and irritable mood are mostly related to the field of psychological factors affecting medical conditions.

Sirri and Fava (2013) discuss how the DCPR may contribute to both macroanalysis (the evaluation of functional relationships between different problems) and microanalysis (a detailed analysis of specific symptoms). As to macroanalysis, the DCPR expands the understanding of the relationships between psychosomatic and psychiatric diagnoses, such as illness denial and agoraphobia, and between different psychosomatic syndromes, such as irritable mood and type A behavior. At the level of microanalysis, the DCPR provides a better characterization of psychological distress, such as the identification of impairments in specific dimensions of well-being in demoralization. The inclusion of the DCPR in the clinical process may expand the spectrum of information about patients' functioning, leading to an increased accuracy in the decision-making phase.

Treatments

The fundamentals of treatment for patients with psychosomatic conditions were quite well described by Schwab (1979), which

1. provide the best medical care available with the help of laboratory investigative procedures wherever necessary,

2. prescribe medications according to a rational plan both for the physical as well as psychiatric components of the disorders,
3. work psychotherapeutically with the patient and preferably one doctor should have the basic responsibility for the patient, and
4. marshal social support systems for the patients.

Pharmacological Treatment

This is usually two pronged. Pharmacological agents appropriate for the physical or medical problem are necessary. Psychotropic agents like antidepressants are advocated as and when necessary. Usually, the severity of psychological or emotional problems determines the use of psychopharmacological agents.

Psychotherapies

Individual psychotherapy has an important role in alleviating psychological or emotional stressors. Many psychosomatic patients do not express their feelings verbally and many are probably alexithymic. Psychotherapy in such patients may not have much role, but in most other patients supportive psychotherapeutic efforts may prove to be beneficial.

Behavior Therapy

Behavior therapy appears to be a useful approach to certain types of problems. The emphasis is on how the patient acts and reacts in everyday life, and modifying behaviors associated with psychosomatic symptoms. Biofeedback has been demonstrated to be useful in certain psychosomatic disorders as migraine, hypertension, certain cardiac arrhythmias, and tension headache.

Usually, the approach to treatment has to be multidisciplinary, with the involvement for different professionals for complete management of psychosomatic disorders. Principles of yoga can be incorporated and tried systematically in the management of such disorders. Tinetti and Fried (2004) suggested that the aim of treatment should be the attainment of individual goals and the identification of all modifiable biological and nonbiological factors according to the biopsychosocial model constructed by Engel (vide supra).

The traditional medical specialties appear to be inadequate in dealing with functional symptoms and problems, which cut across organ systems. A holistic approach is needed with interdisciplinary involvement. There is a need to include functioning in daily life, productivity, performance of social roles, intellectual capacity, emotional stability and well-being, as an important part of clinical investigation and patient care. These aspects have become particularly important in chronic diseases, where cure cannot take place, and also extend over family caregivers of chronically ill patients and health providers. Patients have become increasingly aware of these issues. Psychosomatic interventions may respond to these emerging needs within the established medical system and may play an important role in supporting the healing process (Fava et al., 2012).

Conclusions

Psychosomatic medicine may be considered as a comprehensive, interdisciplinary framework for the assessment of psychosocial factors affecting individual vulnerability, course and outcome of medical diseases; holistic consideration of patient care; integration of psychological therapies in the prevention, treatment, and rehabilitation of medical conditions. Its primary goal is its incorporation into clinical practice. The interdisciplinary dimension that includes the psychosocial domain characterizes many rehabilitation units and pain clinics and is an operational translation of psychosomatic models (Fava and Sonino, 2009). The present challenge of psychosomatic research is to demonstrate its usefulness with stringent research methodologies. However, irrespective of the results of such research, and whatever name is given to the phenomenon, the care of medical disorders will stay to be a combination of pharmacological and psychosocial. In due course of time, spiritual factors would also gain more importance than at present, into the holistic care theme, and lead to research and care of bio-psycho-socio-spiritual medicine.

See also: Hysteria: History and Critiques; Somatoform Disorders; Somatotherapy, History of; Women and Psychiatry.

Bibliography

Chaturvedi, S.K., Tipirneni, S.K., 2005. General hospital and liaison psychiatry. In: Bhugra, D., Ranjith, G., Patel, V. (Eds.), Handbook of Psychiatry – A South Asian Perspective. Byvord Viva Publishers, New Delhi, pp. 405–421.

Chaturvedi, S.K., Jaideep, T., 2007. Working in liaison psychiatry. In: Bhui, Kamal, Bhugra, Dinesh (Eds.), Textbook of Cultural Psychiatry. Cambridge University Press, pp. 526–534.

Engel, G.L., 1967. The concept of psychosomatic disorder. Journal of Psychosomatic Research 11, 3–9.

Fava, G.A., 1992. The concept of psychosomatic disorder. Psychotherapy and Psychosomatics 58, 1–12.

Fava, G.A., Sonino, N., 2000. Psychosomatics: emerging trends and perspectives. Psychotherapy and Psychosomatics 69, 184–197.

Fava, G.A., Sonino, N., 2009. Psychosomatic assessment. Psychotherapy and Psychosomatics 78, 333–341.

Fava, G.A., Sonino, N., Wise, T.N., 2012. Principles of psychosomatic assessment. Advances in Psychosomatic Medicine. Basel, Karger 32, 1–18.

Fava, G.A., Freyberger, H.J., Bech, P., Christodoulou, G., Sensky, T., Theorell, T., Wise, T.N., 1995. Diagnostic criteria for use in psychosomatic research. Psychotherapy and Psychosomatics 63, 1–8.

Kissen, D.M., 1963. The significance of syndrome shift and late syndrome association in psychosomatic medicine. Journal of Nervous and Mental Disease 136, 34–42.

Lane, R.D., Waldstein, S.R., Chesney, M.A., Jennings, J.R., Lovallo, W.R., Kozel, P.J., Rose, R.M., Drossman, D.A., Schneiderman, N., Thayer, J.F., Cameron, O.G., 2009. The rebirth of neuroscience in psychosomatic medicine part I: historical context, methods, and relevant basic science. Psychosomatic Medicine 71, 117–134.

Lipowski, Z.J., 1977. Psychosomatic medicine in the seventies: an overview. American Journal of Psychiatry 134, 233–244.

Lipowski, Z.J., 1984. What does the term 'psychosomatic' really mean? A historical and semantic inquiry. Psychosomatic Medicine 46, 153–172.

Lipowski, Z.J., 1986. Psychosomatic medicine: past and present. Canadian Journal of Psychiatry 31, 2–21.

Mayou, R., 2008. The development of general hospital psychiatry. In: Lloyd, G., Guthrie, E. (Eds.), Handbook of Liaison Psychiatry. Cambridge University Press, New Delhi, pp. 3–23.

Mayou, R., Kirmayer, L.J., Simon, G., Kroenke, K., Sharpe, M., 2005. Somatoform disorders: time for a new approach in DSM-V. American Journal of Psychiatry 162, 847–855.

Page, L., Wesseley, S., 2008. Functional somatic syndromes. In: Lloyd, G., Guthrie, E. (Eds.), Handbook of Liaison Psychiatry. Cambridge University Press, New Delhi, pp. 125–148.

Pert, C.B., Drehert, H.E., Ruff, M.R., 1998. The psychosomatic network: foundations of mind-body medicine. Alternative Therapies in Health and Medicine 4, 30–41.

Porcelli, P., Sonino, N. (Eds.), 2007. Psychological Factors Affecting Medical Conditions. A New Classification for DSM-V. Karger, Basel.

Schwab, J.J., 1979. Psychosomatic disturbances. In: Usdin, G., Lewis, J.M. (Eds.), Psychiatry in General Medical Practice. McGraw Hill, New York.

Sirri, L., Fava, G., 2013. Diagnostic criteria for psychosomatic research and somatic symptom disorders. International Review of Psychiatry 25 (1), 19–30.

Sykes, R., 2012. Somatoform disorder and the DSM-V workgroup's interim proposals: two central issues. Psychosomatics 53, 334–338.

Tinetti, M.E., Fried, T., 2004. The end of the disease era. American Journal of Medicine 116, 179–185.

Voigt, K., Wollburg, E., Weinmann, N., Herzog, A., Meyer, B., Langs, G., Löwe, B., 2012. Predictive validity and clinical utility of DSM-5 somatic symptom disorder. Comparison with DSM-IV somatoform disorders and additional criteria for consideration. Journal of Psychosomatic Research 73, 345–350.

Wise, T.N., 2008. Update on consultation-liaison psychiatry (psychosomatic medicine). Current Opinion in Psychiatry 21, 196–200.

Wittkower, E.D., 1974. Historical perspective of contemporary psychosomatic medicine. International Journal of Psychiatry in Medicine 5, 309–319.

Psychotherapy and Hermeneutics

Gabriele Chiari, School of Specialization in Constructivist Psychotherapy, CESIPc, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The postmodern approaches in psychology (psychological constructivism, social constructionism, symbolic interactionism, narrative psychology) share the attempt at overcoming the Cartesian subject–object dualism and affirming the view of people as meaning-generating beings. For this reason, they take advantage of a dialogue with the philosophical tradition of hermeneutic phenomenology, concerned with the problem of interpretation and understanding of human experience. This article briefly reviews the similarities between hermeneutic phenomenology and the postmodern movement in psychology, before critically outlining the variegated field of hermeneutic approaches to psychotherapy.

The field of human sciences in recent years is showing an unusual kind of dialogue. Some biologists, cyberneticians, computer scientists, cognitive scientists, psychologists, and psychotherapists converse with each other, and – more surprisingly – they all converse with philosophers, in particular with phenomenologists and hermeneutic phenomenologists. What makes such a dialogue possible is an increasingly widespread dissatisfaction with the dominant view of psychology as a natural science, and the search for an alternative answer to the question 'what it is to know.'

A strong impulse to this trend has come from the spreading of broadly speaking 'postmodern' approaches in psychology, particularly in the form of social constructionism and psychological constructivism.

The movement of social constructionism (Gergen, 1985) is characterized by the recognition of the role played by language in the discourse about the world viewed as an artifact of communal interchange – following Berger and Luckmann's (1963) seminal volume *The Social Construction of Reality* – and acknowledges its roots in various phenomenologists who adopted an 'endogenic perspective' about the origins of knowledge.

Psychological constructivism developed from Ernst von Glasersfeld's (1984) interpretation of Piaget's view of knowledge in terms of a radical constructivist epistemology, according to which knowledge has to be meant as a construction of possible realities rather than as a representation of a world out there. Psychological constructivism too, has many commonalities with phenomenology and hermeneutics, as well as with the biological theory of knowledge developed by Maturana and Varela (1987), in its turn associated with Heideggerian hermeneutics (Winograd and Flores, 1986).

Close relationship with the above perspectives also have symbolic interactionism derived from American pragmatism, particularly from the work of George H. Mead (1934) on the emergence of mind and self from the social process of communication, which constitutes human experience, and the narrative approaches in psychology and psychotherapy fostered by Jerome Bruner's (1990) argument for a narrative organization of personal experience.

Though differing from each other with respect to the different emphasis given to the biological, cultural, and social factors implied in the personal construction of experience, all the above approaches share the attempt at overcoming the Cartesian subject–object dualism as well as a view of people as meaning-generating beings. For such features, they take advantage of a dialogue with the philosophical tradition of hermeneutic phenomenology (Chiari and Nuzzo, 2010), concerned with the problem of interpretation and understanding of human experience.

Postmodernism, Phenomenology, and Hermeneutic Phenomenology

The similarities between phenomenology and the postmodern movement in psychology appear so close that some expressions of the latter can be regarded as the present attempt to recover Husserl's (1970/1936) project of a refoundation of science – more specifically psychology – without foundations, that is, on the basis of people's lived, personal experience of phenomena. The interest in the human subject as an independently acting and responsible being also connects phenomenology with existentialism, whose main exponents are considered Friedrich Nietzsche, Søren Kierkegaard, and Jean-Paul Sartre.

Following Wilhelm Dilthey's (1988/1923) criticism to the application of the naturalistic method to the 'sciences of the spirit' supported by a differentiation between causal explanation and understanding as ways of acquiring knowledge, Franz Brentano (1995/1874) identified the peculiar characteristic of psychic phenomena in their intentionality (aboutness), that is, in their being directed on an object.

One of Brentano's students, Edmund Husserl (1983/1913), is often referred to as the father of phenomenology. He proposed a new way of looking at objects by examining how people, in their being intentionally directed on them, actually constitute them. Husserl's focus was on trying to understand human experience as it is lived prereflectively, that is, before resorting to any categorization or conceptualization. The notion of objects as real is not excluded, but 'bracketed' by means of the method of phenomenological reduction, or *epoché*, consisting in questioning what appears to be obvious and taken for granted in order to uncover new meanings.

Like phenomenology, Martin Heidegger's (1962/1927) hermeneutic phenomenology is concerned with the human experience as it is lived. However, Heidegger extended Husserl's proposal by describing the primary mode of human experience as engaged activity rather than detached knowing, and emphasizing the dimensions of historicality and sociality. Being is always a 'being-in-the-world' (*Dasein*), and finds itself thrown into the world and its possibilities via a history and a tradition. Unlike Husserl, Heidegger held that preunderstanding – the meanings of a culture that are already present before we understand and become part of our historicality of background – is not something a person can put aside. In other words, all understanding is connected to a given set of forestructures that cannot be eliminated, but that can be accounted for in the interpretive process. Ultimately, there is an indissoluble unity between the person and the world, and understanding is a basic form of human existence in that it is not a way we know the world, but rather the way we are (Polkinghorne, 1983).

Based on the hermeneutic phenomenology initiated by Heidegger, Hans-Georg Gadamer (1989/1960) argued that interpretation of a text arises from the dialectical interaction between the horizon supplied by the text and the horizon given to it by the reader, the 'horizon' being the range of vision that includes everything that can be seen from a particular vantage point. In the same way, each person in understanding the world, himself/herself and other people, is continuously involved in a process of interpretation. This, in turn, is bound and embedded in history because understanding deploys the knower's 'effective history,' that is, his or her personal experience and cultural traditions, to assimilate new experiences. Effective history constitutes the 'prejudices' or 'preconceptions' brought to bear in understanding, and implies the historicality of any interpretation. Both Heidegger and Gadamer share the notion that language is the foundation of being.

The role of the body – meant as 'phenomenal body' rather than 'objective body' – was brought into the phenomenological inquiry by the French philosopher Maurice Merleau-Ponty (1962/1945). The phenomenal body, or body-subject, represents a permanent condition of experience in the form of embodied subjectivity. Another French hermeneutic philosopher, Paul Ricoeur (1984–1988/1983–1985), argued that the self is essentially embodied, and that personal identity, as a whole, is a narrative identity in that we make sense of our own identities by telling ourselves a story about our own life experiences. Therefore, identities are not fixed structures or substances; instead, they are fluid and, until the story is finished, subject to revision.

The assumed susceptibility to change of personal narrative structures or meaning systems through an open-ended dialectical process of interpretation justifies the application of hermeneutic-phenomenological speculations to the field of psychotherapy.

Hermeneutics and Psychotherapeutic Approaches

Even though hermeneutic phenomenology and psychotherapy are distinctly different disciplines with different origins and goals, they show striking commonalities. To begin with, both disciplines are based primarily on language. Besides, hermeneutic phenomenology aims to understand and interpret meanings as expressed in written texts or in lived experience, as well as psychotherapy is concerned with the understanding and interpretation of the life experiences of people as expressed in their narratives. Such a community of interests should point at a close connection between the two; but this is not exactly the case, at least until few decades ago. The reason for the missed marriage resides in the development of both psychology and psychotherapy as naturalistic disciplines within a scientific discourse dominated by positivism and rationalism, and the resulting spreading of behavioral, cognitive-behavioral, and cognitive therapies. The psychoanalytic approaches escaped long an epistemological affiliation due to their distinctive subject – the unconscious – but for this same reason they were deemed untestable and unscientific by mainstream psychology.

Before the appearance of the postmodern trend, the phenomenological and hermeneutic-phenomenological criticism to the Cartesian, naturalistic approach had an important even though minority place in the development of psychotherapy.

The variegated field of existential-phenomenological approaches to psychotherapy sees the psychiatrist and philosopher Karl Jaspers as the acknowledged founder. His *General Psychopathology* (1997/1913), mainly based upon the empathic understanding of the personal worldviews of the patients and the subsequent search for general principles transcending the individual case, represents an alternative to traditional psychiatry and its categorical diagnosis.

Whereas Jaspers' approach to mental disorders can be regarded as a method of investigation capable of leading to more accurate generalization and classification, Ludwig Binswanger's (1963) and Medard Boss's (1963) *daseinsanalysis*, Viktor E. Frankl's (1959/1946) *logotherapy*, and Eugène Minkowski's (1970) *structural-phenomenological method* are more specifically designed to produce therapeutic effects. Significantly, Minkowski is quoted on the first page of *The Divided Self*, the most famous book of Ronald Laing (1955), which gave fresh contributions to the existential-phenomenological approach in psychiatry. Existential psychotherapy too (May et al., 1958) is based on the direct application of Heidegger's philosophy to the understanding of psychopathology, and aims at favoring the acceptance of the normal existential anxieties that come from confronting life.

The passage from early existential-phenomenological approaches to more recent hermeneutic-phenomenological ones entails a more vague delimitation of the field. The application of the basic tenets of hermeneutic phenomenology to the psychotherapeutic practice shows two main characteristics: (1) It is not identifiable as a specific school with adequate inner cohesiveness, but as the integration of some hermeneutic principles into several, already formalized psychotherapies, in ways depending on the basic assumptions of each; (2) Partly as a consequence, if interpretation, understanding, meaning, intersubjectivity, and narrative are the fundamental tenets of contemporary hermeneutic speculation, their closeness to the original thought of hermeneutic phenomenologists can significantly vary.

It is the case of many psychoanalytic perspectives, particularly Schafer's and Spence's proposals, which are held to have

opened the hermeneutic and narrative turn in psychoanalysis. Roy Schafer (1980) and Donald Spence (1982), in the same period but separately, offered interpretive perspectives in psychoanalysis by coupling the narrative tradition with the Freudian model. People tell stories about themselves and others, and these lend them to various understandings, each of which can legitimately claim to be true. In making interpretation, the analyst retells over and over these stories, accentuating certain features and placing others in parentheses. The end product is a radically new, jointly authored work. However, both the approaches have been criticized for not having altogether relinquished a correspondence theory of truth (Moore, 1999; Sass and Woolfolk, 1988).

Another break with the Cartesian 'myth of the isolated mind' espoused by Freud comes from George E. Atwood and Robert D. Stolorow's (1984) proposal of an *intersubjective psychoanalysis*, to some extent connected with Heidegger's hermeneutics and Husserl's approach to empathy and intersubjectivity. The basic unities of analysis are structures of experience open to a dialogical exploration with the contribution of the therapist's experiential world, within a contextual psychology that recognizes the constitutive role of relatedness in the making of all experience. Even their approach, however, has been said not satisfactorily in-line with the tradition of thinking to which it purports to belong.

The perspective called *dialectical constructivism* suggested by Irwin Z. Hoffman (1998) attributes to both the patient and the analyst the responsibility for the coconstruction of reality in the analytic process in order to promote new experience. Hoffman considers the terms 'constructivist' and 'hermeneutic' as interchangeable, but neither accepts von Glasersfeld's radical constructivism regarded as a kind of solipsistic relativism nor seems to overcome the traditional subject–object dichotomy as hermeneutic phenomenology does.

Much more absorbed by philosophical hermeneutics is the approach developed by Donnel B. Stern (1997). His redefinition of unconscious in terms of *unformulated experience* draws heavily on Gadamer's view of interpretation as a dialogical process, but also on Merleau-Ponty's notion of creative language, and Sartre's understanding of self-deception as *mauvaise foi* (bad faith).

Apart from the psychoanalytic approaches, Gestalt therapy (Perls et al., 1951) is an established existential-experiential form of psychotherapy that emphasizes phenomenological method, dialogic process, personal choice, and responsibility, recognizing its principal influences in European phenomenology and existentialism.

Since the end of the 1980s the application of the hermeneutic theory began to spread to other psychotherapeutic approaches (Messer et al., 1988).

The blend with hermeneutic phenomenology proves itself particularly well matched and fertile with the postmodern perspectives applied to psychotherapy – namely, the psychological constructivist and the social constructionist ones – due to the similarities above described.

Particularly with regard to the theory of disease, both the hermeneutic and the constructivist/constructionist approaches share an antimedical stance. The symptoms of mental disturbance are not regarded as expressions of a physical disease, but, in generic terms, as expressions of the person's attempt at conserving an adaptation with the social environment: they reveal primarily ontological – that is, 'being'-related – issues. They are part of personal narratives that reflect ways of life, manners of living, even though they can imply suffering. In addition to their objection to physical methods of treatment, both phenomenology and constructivism/constructionism reject the rationalistic psychotherapies aimed at correcting false perceptions or beliefs. Actually, the view of knowledge that both share implies that there is no way to have access to an ultimate reality, but that there are only personal interpretations open to alternatives in meaning. Such interpretations cannot be 'correct' or 'wrong,' since they are not based upon external, objective laws or truths that have been universally ascertained. This results in the therapist's attitude of acceptance of the clients' views of the world and themselves, as well as in the effort of empathically understanding them, as a common basis for favoring a therapeutic process and relationship in ways that can differ according to the different theoretical approaches.

A Gadamerian hermeneutic learning is clearly and explicitly present in the social constructionist paradigm as applied to family therapy. According to Anderson and Goolishan (1992), human beings are language-generating and meaning-generating systems, and the therapeutic system itself is linguistic in nature. Starting from a 'not-knowing' position – an attitude expressing a need to know more about what has been said rather than convey preconceived opinions and expectations about the client, the problem, or what must be changed – the therapist becomes part of a linguistic system and a therapeutic conversation wherein the therapist and the family together explore the family's problem through dialogue, aimed at its dissolution and at increasing the potential for the narrative development of new agency and personal freedom.

The constructivist approach to psychotherapy derived by Kelly's (1955) personal construct theory, the forerunner of contemporary constructivist psychotherapies, is having in recent years a narrative hermeneutic development (Chiari and Nuzzo, 2010) through a comparison with Gadamer's hermeneutic conversation and Ricoeur's hermeneutics of narrative self and recognition. Even though Kelly explicitly rejected a comparison with phenomenology, his theory shows many commonalities with its basic tenets which recent analyses are revealing and emphasizing.

As the postmodernist-hermeneutic phenomenological approaches in psychotherapy began to counter the supremacy of the naturalistic-positivist ones, the research methods also showed a significant shift (Polkinghorne, 1983). Qualitative research, once confined to anthropological and sociological research – or, in psychotherapy, to the case studies of rare pioneers as Freud, Maslow, and Rogers – claimed its justification within a scientific inquiry dominated by quantitative research methods and statistical correlations, which had produced results with limited relevance to the analysis of the therapeutic process and relationship. In this case too, though qualitative research aimed at the interpretation of textual material, spoken narratives, therapeutic conversations or change processes are professed to be hermeneutic or interpretive, seldom obeys adequately the rules of philosophical hermeneutics. Anyway, it represents at least a third road of inquiry between the natural scientistic and the genuine

hermeneutic approach in the complex field of psychotherapy research.

See also: Constructivism and Constructionism: Methodology; Constructivist Theory and Psychotherapy; Hermeneutics, History of; Hermeneutics; Phenomenology: Philosophical Aspects; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Science: Constructivist Perspectives, History of; Social Constructivism.

Bibliography

Anderson, H., Goolishian, H., 1992. The client is the expert: a not-knowing approach to therapy. In: McNamee, S., Gergen, K.J. (Eds.), Therapy as Social Construction. Sage, London, pp. 25–39.

Atwood, G.E., Stolorow, R.D., 1984. Structures of Subjectivity: Explorations in Psychoanalytic Phenomenology. The Analytic Press, Hillsdale, NJ.

Berger, P.L., Luckmann, T., 1963. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, Garden City, NY.

Binswanger, L., 1963. Being-in-the World: Selected Papers of Ludwig Binswanger. Basic Books, New York.

Boss, M., 1963. Psychoanalysis and Daseinsanalysis (L.E. Lefebre, Trans.). Basic Books, New York.

Brentano, F., 1995. Psychology from an Empirical Standpoint (L.L. McAlister, Trans.). Routledge, London (Original work published 1874).

Bruner, J.S., 1990. Acts of Meaning. Harvard University Press, Cambridge, MA.

Chiari, G., Nuzzo, M.L., 2010. Constructivist Psychotherapy: A Narrative Hermeneutic Approach. Routledge, London.

Dilthey, W., 1988. Introduction to the Human Sciences: An Attempt to Lay a Foundation for the Study of Society and History (R.J. Betanzos, Trans.). Wayne State University Press, Detroit, MI (Original work published 1923).

Frankl, V.E., 1959. Man's Search for Meaning: An Introduction to Logotherapy. Beacon Press, Boston (Original work published 1946).

Gadamer, H.G., 1989. Truth and Method (J. Weinsheimer and D.G. Marshall, Trans.), second ed. Continuum, London (Original work published 1960).

Gergen, K.J., 1985. The social constructionist movement in modern psychology. American Psychologist 40, 266–275.

von Glasersfeld, E., 1984. An introduction to radical constructivism. In: Watzlawick, P. (Ed.), The Invented Reality: How Do We Know What We Believe We Know? Contributions to Constructivism. Norton, New York, pp. 17–40.

Heidegger, M., 1962. Being and Time (J. Macquarrie and E. Robinson, Trans.). Harper & Row, New York (Original work published 1927).

Hoffman, I.Z., 1998. Ritual and Spontaneity in the Psychoanalytic Process: A Dialectical-Constructivist View. The Analytic Press, Hillsdale, NJ.

Husserl, E., 1983. Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: General Introduction to a Pure Phenomenology (F. Kersten, Trans.). Martinus Nijhoff, The Hague (Original work published 1913).

Husserl, E., 1970. The Crisis of European Sciences and Transcendental Philosophy (D. Carr, Trans.). Northwestern University Press, Evanston, IL (Original work published 1936).

Jaspers, K., 1997. General Psychopathology (J. Hoenig and M.W. Hamilton, Trans.). Johns Hopkins University Press, Baltimore, MD (Original work published 1913).

Kelly, G.A., 1955. The Psychology of Personal Constructs, vols. 2. Norton, New York (Reprinted ed. Routledge, London, 1991).

Laing, R.D., 1955. The Divided Self: A Study of Sanity and Madness. Tavistock, London.

Maturana, H.R., Varela, F.J., 1987. The Tree of Knowledge: The Biological Roots of Human Understanding. Shambhala, Boston.

May, R., Angel, E., Ellenberger, H.F., 1958. Existence. Basic Books, New York.

Mead, G.H., 1934. Mind, Self and Society. University of Chicago Press, Chicago.

Merleau-Ponty, M., 1962. Phenomenology of Perception. Routledge, London (Original work published 1945).

Messer, S.B., Sass, L.A., Woolfolk, R.L. (Eds.), 1988. Hermeneutics and Psychological Theory: Interpretive Perspectives on Personality, Psychotherapy, and Psychopathology. Rutgers University Press, New Brunswick, NJ.

Minkowski, E., 1970. Lived Time: Phenomenological and Psychopathological Studies (N. Metzel, Trans.). Northwestern University Press, Evanston.

Moore, R., 1999. The Creation of Reality in Psychoanalysis: A View of the Contributions of Donald Spence, Roy Schafer, Robert Stolorow, Irwin Z. Hoffman and Beyond. The Analytic Press, Hillsdale, NJ.

Perls, F.S., Hefferline, R.F., Goodman, P., 1951. Gestalt Therapy: Excitement and Growth in the Human Personality. Julian Press, New York.

Polkinghorne, D.E., 1983. Methodology for the Human Sciences: Systems of Inquiry. State University of New York Press, Albany.

Ricoeur, P., 1984, 1985, 1988. Time and Narrative (K. McLaughlin and D. Pellauer, Trans.). University of Chicago Press, Chicago (Original works published 1983, 1984, 1985).

Sass, L.A., Woolfolk, R.L., 1988. Psychoanalysis and the hermeneutic turn: a critique of narrative truth and historical truth. Journal of the American Psychoanalytic Association 36, 429–454.

Schafer, R., 1980. Narrative Actions in Psychoanalysis. Clark University Press, Worcester, MA.

Spence, D.P., 1982. Narrative Truth and Historical Truth: Meaning and Interpretation in Psychoanalysis. Norton, New York.

Stern, D.B., 1997. Unformulated Experience: From Dissociation to Imagination in Psychoanalysis. The Analytic Press, Hillsdale, NJ.

Winograd, T., Flores, F., 1986. Understanding Computers and Cognition: A New Foundation for Design. Ablex, Norwood, NJ.

Relevant Websites

http://plato.stanford.edu/entries/hermeneutics/ – Hermeneutics in Stanford Encyclopedia of Philosophy.

http://web.lemoyne.edu/~hevern/narpsych.html – Narrative Psychology.

http://plato.stanford.edu/entries/phenomenology/ – Phenomenology in Stanford Encyclopedia of Philosophy.

http://www.postmodernpsychology.com – Postmodernism and Psychology.

http://www.univie.ac.at/constructivism/ – Radical Constructivism.

http://www.taosinstitute.net – Social Constructionism.

Psychotherapy and Pharmacotherapy

Richard Balon, Wayne State University School of Medicine, Detroit, MI, USA
Michelle B Riba, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the last 30 years, we have seen a rise and now prevalence of the combination of pharmacotherapy and psychotherapy for patients with a wide variety of psychiatric conditions. The treatment modality that is often used is a collaborative one, in which there is a therapist who provides the psychotherapy and a physician or other clinician, such as nurse practitioner, who provides the medication. There may be more than two clinicians involved, such as a family physician who oversees the care, a psychiatrist, and a psychologist or social worker who provides the psychotherapy. There are positive aspects and negative aspects for this split type of care. Certain group of patients might be at higher risk for difficulties with split treatment, such as those with cluster B types of personality disorders. This is an important area for research and education and improved methods of communication between the providers and the patients.

Combined psychopharmacology and psychotherapy is a common treatment form in modern psychiatric practice. Numerous research studies confirm that the combination of psychopharmacology and psychotherapy treatments are more effective than each alone. Such clinical care is often provided in collaborative treatment arrangements wherein a psychiatrist or other physician provides the medication and manages the medical issues while a nonphysician therapist provides the psychotherapy. In the last several years, the term 'integrated' care is sometimes used in place of the term 'collaborative' while in some health systems, the terms are not necessarily synonymous.

It should be noted that treatments with psychopharmacology and psychotherapy have become common for all psychiatric patients, reflecting the importance of biological psychiatry. Since the early 1990s, there has also been an expansion in specific types of psychotherapies with personality disorders, such as dialectical behavior therapy (Linehan et al., 1993), mentalization-based therapy based on dynamic therapy (Bateman and Fonagy, 1999, 2001), fueling the need for collaborative treatment.

There are positive and negative aspects of providing collaborative treatment and various associated ethical and legal dilemmas. Some of the positive aspects include better utilization of the special talents of both therapist and psychiatrist; cost-effective utilization of resources; and enhanced opportunity for patients to select clinicians with similar ethnic and/or gender background. Negative aspects include splitting of clinicians; potential problems in prescribing decisions due to inadequate knowledge of the content of psychotherapy; and clouded legal and clinical responsibilities. It is important that the clinicians and patients involved in such treatment understand and define their own roles and responsibilities, treatment objectives, and optimal ways of communicating patient care issues.

Background and Introduction

Combined psychopharmacology and psychotherapy is a treatment form quite accepted and ubiquitous in modern psychiatric practice (Beitman, 1983). It may have its origins in the writings of Fromm-Reichman wherein the role and responsibilities of the analytic treating therapist were differentiated from the managerial therapist, a feature of long-term psychotherapy in the 1940s (Fromm-Reichman, 1947). Further explicated by Firman, the rationale of this division of labor was based on analytic principles: the 'psychotherapist' should ideally not be in any managerial relationship with respect to the patient so that the patient's fantasies about the therapist could be viewed and analyzed as material emanating from the patient's own self (Firman, 1982). According to Firman, the usefulness of this relationship is twofold: the 'outsider' (managerial) therapist can serve as a negative split object upon whom the patient can focus her own negative features. This serves to allow the tenuous whole-object relationship with the treatment therapist to proceed without jeopardy. Firman viewed the splitting that would occur in such collaborative relationships to be in a therapeutic setting and thus under better and safer control for the patient (Balon, 1999; Firman, 1982).

Since then, the techniques used in analytic practice and with patients in long-term psychotherapy have been transferred to almost all types of psychotherapy. Fueling this have been newer and simpler types of psychotropic medication which physicians, other than psychiatrists, feel comfortable prescribing (Horgan, 1985; Regier et al., 1978; Schurman et al., 1985) and systems of care, which have advocated for lower costs for patient care. Additionally, numerous research studies confirm that the combination of psychopharmacology and psychotherapy treatments are more effective than each alone (Goldsmith et al., 1999; Karasu, 1992; Koenigsberg, 1994; Manning et al., 1992) adding incentives for therapists to seek consultation for their patients from physicians regarding medication.

In modern psychiatry, collaborative treatment wherein a psychiatrist or other physician provides the medication and manages medical issues while a nonphysician therapist provides the psychotherapy is a common and growing treatment modality. While the field has not definitively defined this paradigm, the American Psychiatric Association published guidelines describing the responsibilities in

a collaborative relationship, noting that the psychiatrist and therapist share responsibility for the care of the patient in accordance with the qualifications and limitations of each person's discipline and abilities (American Psychiatric Association, 1980). Various terms have been used in the literature and in clinical care to describe this practice, which include medication backup, split treatment, combined treatment, medical management, cotreatment, triangular treatment, collaborative treatment, medical backup, and parallel treatment. For the purposes of this section, we will call this practice 'collaborative treatment.'

What has not been definitively researched, however, are the best methods for providing collaborative treatment. Recognizing the cognitive and emotional problems for all parties, Goldsmith and coworkers and others have sought to better explicate the potential difficulties and sources of conflict (Berg, 1986; Bradley, 1990; Chiles et al., 1984; Goldsmith et al., 1999; Kingsbury, 1987). Others have described the positive aspects of this form of treatment (Balon, 1999; Michels, 1995).

The remainder of this article will focus on some of the highlights of collaborative treatment: the positive and negative aspects, legal and ethical issues, and suggestions for maximizing the collaborative process.

Positive and Negative Aspects of Collaborative Treatment

The combination of psychopharmacology and psychotherapy, or collaborative treatment, has some positive and negative aspects. As noted by Beitman (1983), collaborative treatment is based on a triangular relationship between psychiatrists, patients, and nonmedical therapists. In an ideal world, such a triangle should have just positive aspects, e.g., utilization of special talents. Triangular relationships, however, could also be a breeding ground for difficulties or negative aspects, such as splitting. The negative aspects of collaborative treatment do not arise only from patients' pathology. As Brill (1977) pointed out, "More and more nonmedical mental health professionals began to resist and resent working under the direction of a psychiatrist. They believe that their own professions and therapeutic skills have advanced to the point that independence and self-determination were justified." He emphasized the emergence of territorial conflicts arising in collaborative 'team' treatment. The territorial or turf conflicts and concerns are a bilateral issue – psychiatrists, as much as nonmedical therapists, frequently view the 'other side' as a competitor rather than a collaborator.

Collaborative treatment could benefit the patient only if it is truly collaborative, when positive aspects are maximized and negative aspects minimized. Awareness of the possible positive and negative aspects of collaborative treatment is the first step in maximizing the collaborative relationship (see below).

Positive Aspects of Collaborative Treatment

Utilization of Special Talents of Both Therapist and Psychiatrists

As Balon (1999) pointed out, both professionals may possess expertise in special areas, e.g., mood disorders, eating disorders. Firman (1982) also pointed out that a psychiatrist who is a skillful psychotherapist could also serve as a second therapist, or a negative split object, on whom the patient focuses many of his/her own disowned negative features.

Cost-Effective Utilization of All Resources

This is an issue especially important to patients. Due to various economic pressures, most patients cannot afford psychotherapy by psychiatrists unless they have good mental health care benefits. In many cases, insurance companies either do not cover psychotherapy by psychiatrists or pay an amount that is not acceptable by psychiatrists. Psychotherapy provided by nonmedical psychotherapists is considered cheaper (a contentious issue; see Goldman et al., 1998) and frequently is covered by insurance companies. Most patients thus can still benefit from the combination of pharmacotherapy and psychotherapy.

It is important to note there are managed care companies, which are reluctant to approve treatment sessions with seriously ill patients who are not receiving medication; this tenet is another factor in the use of the collaborative care model.

Better Opportunity for Patient to Select Therapist with Similar Ethnic and/or Gender Background

Some authors (e.g., Foulks and Pena, 1995) have advocated that matching a patient with a therapist or physician of a similar ethnic or cultural background may facilitate the process of psychotherapy or pharmacotherapy. Others (Yamamoto et al., 1993) have warned that this matching could have its pitfalls and could be discriminatory at times. Nadelson and Zimmerman (1993) have recommended that a patient's request for a therapist or psychiatrist of a specific gender, ethnicity, or race, or a patient's uneasiness with a particular therapist's or psychiatrist's gender, ethnicity, or race should be an issue explored and addressed in therapy.

Increased Time and Resources Available for Patients

The combination of a 20-min medication review and 50-min therapy session represent more resources than just one of these modalities and is an unquestionable benefit. The combination of resources is also advantageous during the temporary absence of a psychiatrist or therapist.

Increased Amount of Clinical Information

There may be information that the patient feels more comfortable sharing with their therapist and vice versa. Thus, psychiatrist and therapist should set time aside to share information, impressions, and concerns and to support each other in their efforts to help patients (Pilette, 1988).

Enhancement of Adherence

As Paykel (1995) pointed out, psychotherapy may assist or possibly hinder medication compliance. Similarly, collaborative treatment may also increase compliance with psychotherapy. To date, the difference in adherence between collaborative and split treatment has not been delineated, although collaborative care seems to improve adherence. Levine Baruch (2012) has studied the patient's sense of satisfaction and subsequent adherence in split treatment is

complex and multifactorial depending on the matching of the treatment with the patients' treatment preferences, prioritized needs, and relational style, the patient may exhibit varying levels of satisfaction and adherence behaviors (Levine Baruch, 2012). Much of the studies to date have focused on adherence to medication, not to psychotherapy, in collaborative versus split treatment.

Professional and Emotional Support for Both Therapist and Psychiatrist

As Pilette (1988) pointed out, such support is especially important when the patient is difficult to treat and also during times of crisis. When there is collaboration with a primary care physician, the mental health professional can confer with someone who may have a long-standing relationship with and understanding of the patient, which may enhance adherence.

Negative Aspects of Collaborative Treatment

Inappropriate Prescribing Decisions without Knowledge of the Content of Therapy

The prescribing psychiatrist may not always be aware of all the inner conflicts or external pressures felt by the patient. For instance, the psychiatrist may prescribe medication for anxiety that arose after an important intrapsychic conflict was addressed in therapy with the nonmedical therapist.

Potentially Discrepant Information Given by the Patient to Each Clinician

Frequently, the discrepant information provided by the patient to both treating professionals could lead to decisions with serious consequences. This issue is directly related to the next one – splitting.

Splitting of Clinicians

This is probably the most frequently discussed negative aspect of collaborative treatment. Treatment provided by two persons provides a fertile ground for a patient to develop negative transference, to introduce problematic countertransference, and to split the prescribing psychiatrist and therapist. Splitting may lead to noncompliance with any of the treatments.

Patients with personality disorders, in particular Cluster B traits with emotional liability, depressivity, separation insecurity and hostility, may 'split' even without a 'split' treatment relationship so this problem must be acknowledged when a patient is treated in a collaborative care model (Schlesinger and Silk, 2009). Patients may idealize or devalue one or the other clinician making the roles of 'good' and 'bad' treater shift. The treaters may unconsciously identify with the projected characteristics promoted by the patient and may respond accordingly (Schlesinger and Silk, 2009; Gabbard, 1989; Gabbard and Wilkinson, 1994).

Unclear Confidentiality

As Pilette (1988) pointed out, is the collaborative treatment an occasion for free communication between a therapist and psychiatrist without the patient's consent? What if the patient asks that some important information not be related to the other treating professional? These issues are complicated and need to be discussed and resolved.

Clouded Legal and Clinical Responsibility

It is not always clear who is legally responsible for various aspects of collaborative treatment. It is rather safe to assume, however, that the psychiatrist will most likely be found legally responsible for all aspects of clinical care including inappropriate or negligent supervision (Balon, 1999).

Lack of Reimbursement for Collaboration

The fact that third party payers do not reimburse for collaboration somewhat hinders the collaborative treatment.

Various Misconceptions of Patient, Therapist, and Psychiatrist

These include issues such as therapists and psychiatrists ignoring the psychological meaning of medication, psychiatrists' belief that therapists should always agree with and support physicians' decisions, and patients' common perception that psychiatrists are only interested in prescribing medication (Goldsmith et al., 1999).

This is only a brief summary of positive and negative aspects of collaborative treatment. More thorough discussion of both positive and negative aspects is provided in the two introductory chapters of *Psychopharmacology and Psychotherapy. A Collaborative Approach* (Goldsmith et al., 1999; Riba and Balon, 1999).

Legal and Ethical Issues

Legal

The major worry in collaborative treatment is that with two clinicians delivering the care, there will be a loss of information or understanding of what may be occurring for the patient. There are currently no standards in place for how frequently or for how long patients should be seen by physicians if collaborative treatment is provided. There is, therefore, a broad range of frequency of visits and communication patterns between the patient, therapist, and physician, which range from ideal to inadequate. Because medications are involved in collaborative treatment and historically have been the source of significant legal exposure, the issue of medical practice is an important topic for discussion. Negligence underlying malpractice liability is of two types: (1) carelessness in gathering necessary data about a patient and (2) negligence in assessing that information in determining appropriate treatment modalities, interventions, providing precautions, etc. (MacBeth, 1999). Psychiatric medications, like all medications, are associated with side effects. Physicians who prescribe medications become liable, if there is a deviation in the accepted uses of the medications and there results harm to the patient. It is critical for physicians to obtain informed consent from patients when prescribing medication and also to ascertain liability for patient suicide.

If patients are going to be seen infrequently by the prescribing physician, there must be a way for the physician to set up a communication pattern where side effects and sociality are made known, to the extent possible, to the physician, on a regular basis. For example, if a patient were being seen by the physician every 6 months but the therapist is seeing the patient weekly, and then the therapist and physician must discuss how issues and developments about the patient can be communicated.

In collaborative treatment, it must be determined how the treatment plan is shared between the clinicians; how emergencies including hospitalization will be managed; and who is directing the psychotherapy. The setting may play a large role in the extent to which the physician, therapist, and patient communicate. In many public settings, for example, the psychiatrist only sees the patient infrequently. The patient, however, has a case manager whose job it is to communicate to the various clinicians about the patient. In managed care settings, psychiatrists may find themselves in relationships with therapists whom they do not know and who are unaware of their skills, talents, experience, etc. There may be disincentives for the physician to see the patient often, even when more psychosocial problems accumulate for the patient. In collaborative treatment in certain settings, patients sometimes do not even know who the doctor is and often the physician has not had the chance or taken the opportunity to meet with the patient's family, since others (therapists, case managers, nurses, and a previous physician) might have already done so.

There is, therefore, a range of potential legal problems for physicians who engage in collaborative treatment. Physicians are usually viewed as 'the deep pocket' when there is a lawsuit regarding a patient's treatment. The physician usually carries more insurance than the nonphysician therapist and is viewed by juries as 'being in charge' by virtue of education and experience. Even if the plaintiff's attorney does not believe that the psychiatrist was central to the matter, it is often likely that the psychiatrist will be part of the suit to extend the limits of insurance and to try to reach 'the deep pocket.' There is also the problem that occurs when there is more than one defendant. One defendant can try to make their case stronger by blaming the other defendant, thereby turning against one another. Another problem is that even if medications or the medical management provided by the physician is of minor importance to the reason for the suit, regardless of the realities, the physician is often viewed as being in charge while the therapist is viewed as subordinate to the physician.

Finally, physicians are increasingly signing contracts to provide services for patients in managed care. These contracts are usually written by the companies themselves and the fine print is not read carefully by the physician, causing them to be implicitly involved in relationships with therapists that are not collaborative in nature but more supervisory. This practice, therefore, often puts the physician at further risk.

There are some suggestions for clinicians to heed in order to decrease legal risk in collaborative treatment:

1. Clinicians should understand the settings they are practicing in, including its operations and routines.
2. Care should not be compromised because of the setting or because the patient is being treated in a collaborative model.
3. Use good communication techniques with professionals with whom you are sharing responsibility for the patient. Document these discussions.
4. If supervision is part of the collaborative relationship, then set regular supervision times and amend depending on the patient's condition and needs.
5. Read the fine print in all managed care and other contracts that are signed.
6. For a more thorough discussion of these issues, please see the work of MacBeth (1999).

Ethical

When patients receive collaborative treatment because that is what is deemed best for their care, then there is not an ethical problem. An ethical and clinical dilemma, however, may exist if such care is provided because of cost considerations (Lazarus, 1999). While some patients clearly benefit from team approaches to care, others do not. Research is starting to help us discriminate between such patients and determine mental health outcomes in different practice and fee arrangements (Rogers et al., 1993; Wells et al., 1992). Goldman et al. (1998), for example, has shown that for many patients, psychotherapy and psychopharmacology provided by a psychiatrist alone is as effective as and less costly than collaborative treatment.

There are four bioethical principles that need to be considered when one is providing collaborative treatment (Beauchamp and Childress, 1994):

1. Beneficence: the covenant between professional and patient to do what is in the patient's best interest;
2. Nonmaleficence: the principle of doing no harm;
3. Autonomy: respect for the individual rights of patients, especially in the area of informed consent and treatment refusal; and
4. Justice: the fair treatment of patients.

As noted by Lazarus, when determining ethical dilemmas related to collaborative treatment, any two or more of the above principles may be compromised or in conflict (Lazarus, 1999). In this era of cost containment and lack of standards regarding collaborative treatment, all of these principles are being tested. The American Psychiatric Association (APA, 1993) and the American Medical Association (AMA, 1994) have addressed ethical principles and have recognized the importance of collaborative treatment. Section Conclusion of the AMA principles states: "A physician shall continue to study, apply and advance scientific knowledge, make relevant information available to patients, colleagues, and the public, obtain consultation, and use the talents of other health professionals when indicated" (AMA, 1994).

It is a complex area where psychiatrists and other mental health professionals engaged in collaborative treatment can find themselves in ethical dilemmas. Some of this has to do with the settings and the rules within those settings; physician–therapist relationships that should be modified because of the patient's condition but may not be; need for differential types of supervision and monitoring between therapist and physician but which are not taken into consideration; fee schedules that cause disincentives for providers to see patients more often, even when needed; lack of understanding by the patient of the nature of the collaborative relationship between the providers, etc.

In order for health professionals to provide psychotherapy and psychopharmacology in an ethical manner, Lazarus (1999) suggests:

1. The physician and therapist keep themselves adequately and appropriately informed of the nature of the treatment and the progress of those patients cared for in this type of treatment.
2. The physician and therapist assure themselves that the treatment is being carried out competently and adequately.
3. The physician and therapist provide enough supervision, consultation, or collaboration to ensure that their ethical, medical, and legal responsibilities toward the patient are satisfied.
4. The collaborative treatment is consistent with any local, state, or national guidelines outlining psychiatrists' responsibilities to patients and in conformity with licensure requirements and the scope of practice of all the health care professionals involved.

Maximizing the Collaborative Relationship

Providing combined psychopharmacology and psychotherapy is a complex enterprise, much more difficult than a single therapist providing either the psychotherapy, psychotherapy and pharmacotherapy, or pharmacotherapy. There are, therefore, some general suggestions offered that might be useful.

It is incumbent upon the clinicians and patients involved in such treatment to be aware of the complicated nature of this form of care. While combining psychopharmacology and psychotherapy offers distinct and unique advantages, there are problems, as noted in the above section. Time and attention must, therefore, be given to the process of treatment as well as the content. Written contracts between clinicians have even been suggested as a way to delineate the specific duties and responsibilities of providers (Appelbaum, 1991). Other authors have suggested that the patient be part of the three way contract and so be asked to sign such agreements for care (Chiles et al., 1984; Woodward et al., 1993). Specific guidelines for combined care based on the phases of treatment (initial, mid, termination) have been described by Rand (1999). Kahn (1990) has delineated the ideological differences that the psychotherapist and psychopharmacologist often have, and the identified roles they must play in order to provide such treatment.

It is probably prudent for clinicians, both physicians and nonphysician therapists, to review their case mix from time to time. Goals would be to limit the numbers of such collaborative treatments based on the amount of time it takes to speak with other therapists, review and develop joint treatment plans, read other clinicians' notes, determine the background and treatment orientation of the clinicians who are collaborating on patient care, work with patients' families and significant others, develop vacation and coverage schedules with collaborating clinicians, discuss such arrangements with patients, etc. Such thoughtful processes may lead to a decrease in malpractice exposure and ultimately more efficient and useful clinical care.

Attention must be paid carefully to the issue of confidentiality and communication between clinicians and with the patient. Consent to discuss treatment plans as well as content of sessions must be clarified. Since it is impossible to initially predict what will be important or relevant, the patient must be cognizant of these communication patterns and provide consent to clinicians at the beginning of treatment. Any barriers to such communication should be discussed with the patient as well as any relevant family members. It should be determined how the information will be transmitted, that is by telephone, electronic communication, mail, in person, etc., and how often. Communication must be regular, not just when there is a crisis (Vasile and Gutheil, 1979). It is critical that the patient understands which type of problems should be discussed with which clinician. Further, patients should not be placed in the position of needing to relay information between clinician and the other.

Even with that said, it will sometimes be difficult to sort out. As an example, often what seems like uncomfortable medication side effects could be resistance to taking medications. What seems like a dynamic issue in therapy could in fact be a medical problem. Clinicians need to be supportive of the treatment plan; listen for resistances to either psychotherapy or pharmacotherapy; try to delineate the boundary of the treatment being provided by the individual clinician; and talk regularly with one another about the patient's progress. Because of financial remuneration patterns and modern contractual agreements between clinicians and payers, it is easier than ever for a clinician to find themselves responsible for large numbers of patients, many of whom are not being seen frequently or regularly. This can lead to problems with the inherent doctor–patient relationship that is the hallmark of providing care, whether it is pharmacotherapy, psychotherapy, or both. The clinician is well advised to review the ethical boundaries of the work situation and the nature of the tasks required.

Certain patient populations may be at higher risk in such combined therapies. For example, patients with personality disorders are notoriously difficult to treat given their problems with interpersonal relationships. As summarized by Silk (1999) and others, patients with Cluster B diagnoses are prone to splitting (Kernberg, 1975); often do not tell the entire story equally to all clinicians (Main, 1957), and tend to externalize problems (Silk et al., 1995). When there is concomitant substance abuse, emotional liability, and self-mutilating behaviors, combined treatment with more than one clinician can be enormously difficult and complex.

Finally, psychiatry residency training programs and graduate programs in social work and psychology must develop education and training in the specific area of combined treatment. Teaching psychopharmacology alone or psychotherapy without integrated pharmacotherapy is no longer sufficient. There needs to be curricula developed and teachers who are trained themselves in how to deliver such care with emphasis on learning to communicate clearly with other professionals; thinking clearly about diagnosis and treatment strategies; and developing leadership abilities (Spitz et al., 1999).

Conclusion

Providing combined pharmacotherapy and psychotherapy is a growing and important part of psychiatric care. It is a complex dynamic treatment, fraught with potential pitfalls but offering many advantages. Importantly, it offers a unique opportunity for a bio-psychosocial and fiscal model of care and interesting areas for future research and clinical development.

See also: Antidepressant Drugs; Behavior Therapy: Psychiatric Aspects; Cognitive Behavioral Therapy; Psychopharmacotherapy, Costs of; Therapeutic Community.

Bibliography

American Medical Association Council on Ethical and Judicial Affairs, 1994. Code of Medical Ethics: Current Opinions with Annotations. American Medical Association, Chicago, IL.

American Psychiatric Association, 1980. Guidelines for psychiatrists in consultative, supervisory, or collaborative relationships with nonmedical therapists. American Journal of Psychiatry 137, 1489–1491.

American Psychiatric Association, 1993. The Principles of Medical Ethics with Annotations Especially Applicable to Psychiatry. American Psychiatric Association, Washington, DC.

Appelbaum, P.S., 1991. General guidelines for psychiatrists who prescribe medication for patients treated by nonmedical psychotherapists. Hospital and Community Psychiatry 42, 281–282.

Balon, R., 1999. Positive aspects of collaborative treatment. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 1–31.

Bateman, A., Fonagy, P., 1999. Effectiveness of partial hospitalization in the treatment of borderline personality disorder: a randomized controlled trial. American Journal of Psychiatry 156, 1563–1569.

Bateman, A., Fonagy, P., 2001. Treatment of borderline personality disorder with psychoanalytically oriented partial hospitalization: an 18 month follow-up. American Journal of Psychiatry 158, 35–42.

Beauchamp, T.L., Childress, J.F., 1994. Principles of Biomedical Ethics. Oxford University Press, New York.

Beck, A.T., Freeman, A., 1990. Cognitive Therapy of Personality Disorders. Guilford, New York.

Beitman, B.D., 1983. The demographics of American psychotherapists: a pilot study. American Journal of Psychotherapy 37, 37–48.

Berg, M., 1986. Toward a diagnostic alliance between psychiatrist and psychologist. American Psychology 41, 52–59.

Bradley, S.S., 1990. Nonphysician psychotherapist-physician pharmacotherapist: a new model for concurrent treatment. Psychiatric Clinics of North America 13, 307–322.

Brill, N.Q., 1977. Delineating the role of the psychiatrist on the psychiatric team. Hospital and Community Psychiatry 28, 542–544.

Bhugra, Dinesh, 2014. In: Wright, James (Ed.), Psychiatry Encyclopedia of Social and Behavioral Sciences. Elsevier.

Chiles, J.A., Carlin, A.S., Beitman, B.D., 1984. A physician, a nonmedical psychotherapist, and a patient: the pharmacotherapy-psychotherapy triangle. In: Beitman, B.D., Klerman, G.L. (Eds.), Combining Psychotherapy and Drug Therapy in Clinical Practice. SP Medical and Scientific Books, New York, pp. 89–101.

Davidson, K., Norrie, J., Tyrer, P., et al., 2003. The effectiveness of cognitive behavior therapy for borderline personality disorder: results from the borderline personality disorder study of cognitive therapy (BOSCOT) trial. Journal of Personality Disorders 20 (5), 450–465.

Firman, G.J., 1982. The psychiatrist-nonmedical psychotherapy team: opportunities for a therapeutic synergy. Journal of Operational Psychiatry 13, 32–36.

Foulks, E.F., Pena, J.M., 1995. Ethnicity and pharmacotherapy: a component in the treatment of cocaine addiction in African Americans. Psychiatric Clinics of North America 18, 607–620.

Fromm-Reichman, F., 1947. Problems of therapeutic management in a psychoanalytic hospital. Psychoanalysis Quarterly 16, 325–356.

Goldman, W., McCulloch, J., Cuffel, B., et al., 1998. Outpatient utilization patterns of integrated and split psychotherapy and pharmacotherapy for depression. Psychiatric Services 49, 477–482.

Goldsmith, R.J., Paris, M., Riba, M.B., 1999. Negative aspects of collaborative treatment. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 33–63.

Gabbard, G.O., 1989. Splitting in hospital treatment. American Journal of Psychiatry 146, 444–451.

Gabbard, G.O., Wilkinson, S.M., 1994. Management of Countertransference with Borderline Patients. American Psychiatric Press, Washington, DC.

Horgan, C.M., 1985. Specialty and general ambulatory mental health services: comparison of utilization and expenditures. Archives of General Psychiatry 42, 565–572.

Kahn, D.A., 1990. The dichotomy of drugs and psychotherapy. Psychiatric Clinics of North America 13, 197–208.

Karasu, T.B., 1992. The worst of times, the best of times: psychotherapy in the 1990s. Journal of Psychotherapy Practical Research 1, 2–15.

Kernberg, O.F., 1975. Borderline Conditions and Pathological Narcissism. Jason Aronson, New York.

Kingsbury, S.J., 1987. Cognitive differences between clinical psychologists and psychiatrists. American Psychology 42, 152–156.

Koenigsberg, H.W., 1994. The combination of psychotherapy and pharmacotherapy in the treatment of borderline patients. Journal of Psychotherapy Practical Research 3, 93–107.

Lazarus, J., 1999. Ethical issues in collaborative or divided treatment. Risk. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 159–177.

Levine Baruch, R., 2012. Understanding the Experiences of Patients with Depression in Split Treatment. Department of Psychology, Fordham University, New York.

Linehan, M.M., Heard, H.L., Armstrong, H.E., 1993. Naturalistic followup of a behavioral treatment for chronically parasuicidal borderline patients. Archives of General Psychiatry 50, 971–974.

MacBeth, J.E., 1999. Divided treatment: legal implications and risk. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 111–158.

Main, T.F., 1957. The ailment. British Journal of Medical Psychology 30, 129–145.

Manning, D.W., Markowitz, J.C., Frances, A.J., 1992. A review of combined psychotherapy and pharmacotherapy in the treatment of depression. Journal of Psychotherapy Practical Research 1, 103–116.

Michels, R., 1995. Commentary: medication use during psychoanalysis: a survey. Journal of Clinical Psychiatry 56, 179.

Nadelson, C.C., Zimmerman, V., 1993. Culture and psychiatric care of women. In: Gow, A.C. (Ed.), Culture, Ethnicity and Mental Illness. American Psychiatric Press, Washington, DC, pp. 501–515.

Paykel, E.L., 1995. Psychotherapy, medication combinations, and compliance. Journal of Clinical Psychiatry 56 (Suppl. 1), S29–S30.

Pilette, W.L., 1988. The rise of trepany treatment relationships. Psychotherapy 25, 420–423.

Rand, E.H., 1999. Guidelines to maximize the process of collaborative treatment. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 353–380.

Regier, D.A., Goldberg, I.D., Taube, C.A., 1978. The de facto US mental health services system: a public health perspective. Archives of General Psychiatry 35, 685–693.

Riba, M.B., Balon, R. (Eds.), 1999. Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC.

Rogers, W.H., Wells, K.B., Meredith, L.S., et al., 1993. Outcomes for adult outpatients with depression under prepaid or fee-for-service financing. Archives of General Psychiatry 50, 517–525.

Schlesinger, A., Silk, K.R., 2009. Collaborative treatment. In: Oldham, J.M., Skodol, A.S., Bender, D.S. (Eds.), Essentials of Personality Disorders. American Psychiatric Publishing, Washington, DC, pp. 321–341.

Schurman, R.A., Kramer, P.D., Mitchell, J.B., 1985. The hidden mental health network: treatment of mental illness by nonpsychiatrist physicians. Archives of General Psychiatry 42, 89–94.

Silk, K.R., 1999. Collaborative treatment: personality disorder patients. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 221–277.

Silk, K.R., Lee, S., Hill, E.M., et al., 1995. Borderline personality-disorder symptoms and severity of sexual abuse. American Journal of Psychiatry 152, 1059–1064.

Spitz, D., Hansen-Grant, S., Riba, M.B., 1999. Residency training issues in collaborative treatment. In: Riba, M.B., Balon, R. (Eds.), Psychopharmacology and Psychotherapy: A Collaborative Approach. American Psychiatric Press, Washington, DC, pp. 279–306.

Vasile, R.G., Gutheil, T.G., 1979. The psychiatrist as medical backup: ambiguity in the delegation of clinical responsibility. American Journal of Psychiatry 136, 1292–1296.

Wells, K.B., Burnam, M.A., Rogers, W.H., et al., 1992. The course of depression in adult outpatients: results from the medical outcomes study. Archives of General Psychiatry 49, 788–794.

Woodward, B., Duckworth, K.S., Gutheil, T.G., 1993. The pharmacotherapist-psychotherapist collaboration. In: Oldham, J.M., Riba, M.B., Tasman, A. (Eds.), American Psychiatric Press Review of Psychiatry, vol. 12. American Psychiatric Press, Washington, DC, pp. 631–639.

Yamamoto, J., Silva, J.A., Justice, L.R., et al., 1993. Cross-cultural psychotherapy. In: Gaw, A.C. (Ed.), Culture, Ethnicity and Mental Illness. American Psychiatric Press, Washington, DC, pp. 101–124.

Psychotherapy, History of: Psychiatric Aspects

Sidney Bloch and Edwin Harari, St Vincent's Hospital, University of Melbourne, Melbourne, VIC, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12484–12491, © 2001, Elsevier Ltd.

Abstract

Psychotherapy has its origins in two healing traditions – the magico-religious and the medico-scientific. It was with the publication of *Studies in Hysteria*, by Josef Breuer and Sigmund Freud, and their concept of 'free-association' – the patient was instructed to disclose whatever came into their mind while the therapist adopted an attitude of 'evenly-suspended attention' in listening to and interpreting the material – that the latter was firmly launched. Freud revised his theories at numerous points and many of his colleagues extended the boundaries of psychoanalytic thinking. But differences of opinion soon surfaced, which led notable figures to evolve their own 'schools'. Jung and Adler were two foremost dissidents in Europe, Horney, Fromm, and Sullivan in the US, and Klein in Britain. Humanistically oriented schools became popular in the 1960s and 1970s, mainly in the US. Among their chief features were challenging patients with the way they avoid emotionally significant matters in the 'here-and-now'. Feminist-based psychotherapies evolved in the 1980s, reflecting a range of opinion from condemning Freud as a misogynist to appreciating his linkage of life experience, innate drives, and defenses in influencing vulnerability. The second half of the twentieth century saw the evolution of treatments conducted with more than one person: groups of strangers and families.

Although the therapy of stranger groups was experimented on early in the twentieth century, it was not until World War II that major developments ensued. In the US, the move to the study of group process was spearheaded by social psychologists. Paralleling the human potential movement in the 1960s, the group movement transferred its focus from group to personal dynamics, and spawned the encounter movement. Formal group therapy under professional leadership then assumed a more prominent role. Yalom pioneered the study of therapeutic factors specific to the group rather than transposing psychoanalytic or other theories from the individual setting. Alongside this interest in group therapy came the idea of working with the family. Many competing theories of how families become dysfunctional resulted, but principles of Systems Theory dominated. Therapists also began to consider that families might be stymied by their interpretation of past experience. With the narrative a family conveys about their lives regarded as a construction which organizes experience in particular ways, other narratives are excluded. Systematic research in the psychotherapies was a low priority for decades. An impetus was Eysenck's critique of treatment as ineffective. Since then research has yielded much knowledge about whether psychotherapy works and how it works.

Several reviews, including sophisticated methods such as meta-analysis, point to psychotherapy benefiting suitably selected patients. Research on comparative effectiveness has shown a consistent pattern – that 'everyone has won and must have prizes,' a finding probably attributable to factors common to all approaches. Attention has also been devoted to the therapeutic process, which has been shown to be as salient as that on outcome since only with the study of what occurs in treatment can therapists appreciate factors accounting for improvement.

Psychotherapy – the systematic application of psychological principles to accomplish symptomatic or more substantial personality change – has its origins in two healing traditions – the magico-religious and the medico-scientific (Bromberg, 1959). The idea that human experience is influenced by supernatural forces and that certain people have the power to intercede with these forces dates back to antiquity. Shamans and sorcerers have resorted to amulets, magical potions, and incantations for centuries. They have also conducted exorcism and induced altered states of consciousness.

With the Age of Reason, interest was aroused in the newly discovered phenomena of electricity and magnetism. Anton Mesmer (1734–1815), a brash physician, drew on the latter when propounding his theory of animal magnetism. Since the body contained 'magnetic fluid,' and this could become disturbed, by applying magnets to various parts a patient could be relieved of pain and other ailments. While Mesmer's work was ultimately denounced by a Royal Commission and he was forgotten, the observation that his methods appeared to help patients continued to attract attention. James Braid (1795–1860), a Manchester doctor, noted that he could induce a similar trance-like state by getting people to fix their gaze on a luminous object. Labeling the phenomenon hypnotism, he proceeded to demonstrate its effectiveness in treating a range of conditions.

The doyen of French neurology, Jean Martin Charcot (1835–93), considered hypnosis a neurophysiological process, and set about its study in states like somnambulism and hysteria. His pupil Pierre Janet (1859–1947) suggested that the latter was brought about by weakening of a higher brain function, resulting in a constriction of consciousness. In this 'dissociated' state thoughts could not be integrated, and symptoms were beyond the reach of consciousness.

In an alternative explanation, Ambroise Liebeault (1823–1904), and Hippolyte Bernheim (1840–1919) in France posited a narrowing of attention in hypnosis which rendered the patient vulnerable to the therapist's influence by suggestion. Far from invalidating hypnosis, the power of suggestion could be investigated scientifically and accepted as a bona fide therapy.

Freud (1856–1939) had studied briefly with Charcot in 1885–86 and became enthusiastic about new possible treatments for hysteria. Freud's mentor, a respected Viennese physician, Josef Breuer (1842–1925), had described the treatment of a young woman who suffered from an array of marked hysterical symptoms. Instead of suggesting disappearance of

the symptoms, Breuer had encouraged Anna O (the pseudonym given to her) to talk freely about her life under hypnosis. She did so and over time began to share memories about her father's illness. Breuer noted that after each session, his patient became less distressed and her symptoms improved. Anna herself referred to this approach as her 'talking cure' or 'chimney sweeping'; Breuer termed it 'catharsis'. Whereas Charcot had described physically traumatic events as possibly causal, in his hysterical patients, Breuer and Freud's interest focused on psychological trauma such as humiliation and loss. Freud also became less keen on applying hypnosis.

Their volume, *Studies in Hysteria*, which included the case of Anna O., and four others treated by Freud himself, also contained ideas on defense mechanisms and an account of various techniques including suggestion, hypnosis, and catharsis. Most importantly, however, was the new concept of 'free-association': the patient was instructed to disclose whatever came into their mind, without any censoring, while the analyst adopted an attitude of 'evenly-suspended attention' in listening to and interpreting the material.

On the basis of his subsequent self-analysis and further clinical experience, Freud elaborated his theory of infantile sexuality. Sexual fantasies in young children centered around the triangular relationships of love and rivalry with their parents (the Oedipal complex). Promoting free association of this theme, as well as of dreams, slips of the tongue, and other unconsciously-based thoughts, feelings, and fantasies, Freud emphasized the centrality of the complex in the neuroses. While catharsis and insight (making the unconscious conscious) seemed pivotal to therapy, Freud soon realized that other factors also operated, particularly the role of transference, i.e., how feelings, thoughts, and fantasies stemming from childhood experiences and revived in current relationships are 'transferred' onto the analyst.

Interpretations were directed at this threefold nature of the analytic experience. The therapist's response to transference (countertransference) led to analysts having their own personal therapy, the aim being to minimize it, and thus to be thoroughly objective in attending to the patient's disclosures (Ellenberger, 1970).

In summary then, psychoanalytic psychotherapy elaborated the following processes: the patient's unbridled disclosure of whatever enters his or her mind – free association; the transference of infantile and childlike feelings and attitudes to the therapist which were previously directed to key figures in the patient's earlier life; interpretation of the transference as well as of defenses the patient applies to protect him/herself and the resistance he/she manifests to self-exploration; and finally, the repeated working through of the discoveries made in treatment. The ultimate aim is insight with translation into corresponding change in behavior and personality.

Variations on a Theme

Freud revised his theories at numerous points in his long professional life. In addition, many colleagues extended the boundaries of psychoanalytic thinking. But differences of opinion soon surfaced, some radical. The nature of psychoanalysis then, especially Freud's intolerance of dissent, led notable figures to leave the fold and to evolve their own models on which corresponding schools of therapy were established. Jung and Adler were two foremost European dissidents; Karen Horney, Erich Fromm, and H.S. Sullivan were pioneering neo-Freudians in the US; and W.D. Winnicott and Melanie Klein were prominent in Britain (Mitchell and Black, 1995).

Carl Gustav Jung is perhaps the most critical dissenter since he was clearly being groomed to take over leadership of the psychoanalytic movement. But he was also celebrated because of his own contribution. Specifically with regard to psychotherapy, he advanced the notion of 'individuation' – the aim was to discover all parts of oneself and one's creative potential. Jung was less concerned than Freud with the biological roots of behavior, especially infantile sexual development, but rather emphasized social and cultural factors. The role of transference was replaced by a more adult type of collaboration.

Furthermore, the unconscious for Jung is not merely the repository of the individual's history but also a wider social history, a phenomenon Jung labeled the collective unconscious. He arrived at this notion through a study of myths, legends, and symbols in different cultures, in different epochs. That these are shared by a variety of cultures is not fortuitous but reflects cosmic mythical themes or archetypes, a salient feature of man's collective history.

Alfred Adler (1870–1937), a Viennese physician, was like Jung a prominent dissenter. Adler joined Freud's circle in 1902 but broke away to form his own school of 'individual psychology' after 9 years. An important tenet concerns individual development. We begin life in a state of inferiority, weak, and defenseless, for which we compensate by striving for power and by evolving a lifestyle to make our lives purposeful. The pattern that emerges varies and may include such goals as acquisition of money, procreation, high ambition, or creativity. The drive for power and choice of lifestyle may, however, go awry in which case a neurosis results. A path is followed which leads to ineffective efforts to cope with the feeling of inferiority and assumption of a facade or false self.

Adler regarded therapy as a re-educative process in which the therapist, who serves as model and source of encouragement, engages in a warm relationship with the patient and enables him or her to understand the lifestyle he/she has assumed. Unconscious determinants of behavior are less crucial than conscious ones and the term 'unconscious' is used only descriptively to refer to aspects of the person that are not understood by him.

Freud's insistence on innate drives and infantile sexuality not only led to the schisms with Jung and Adler but also spurred a new generation of analytically oriented analysts to concentrate on interpersonal aspects of psychological experience. In the US, Sullivan pioneered the school of 'interpersonal psychiatry,' in which the therapist adopted the role of participant observer in treatment, the transference providing one opportunity to explore communication and its breakdown in the patient's interpersonal world.

We can consider the chief features in the approach of the neo-Freudians by looking briefly at Karen Horney (1885–1952). She became disenchanted with Freud's rigid focus on instinctual biological factors, arguing that cultural factors were more salient, as reflected in the differences in psychological development and behavior among different

sociocultural groups. Indeed, behavior regarded as normal in one culture could be viewed as neurotic in another.

In line with her emphasis on culture, Horney advanced the role of parental love in the life of a young child. Children typically suffer anxiety, a consequence of feeling helpless in a threatening world. The child reared in a loving atmosphere succeeds in overcoming basic anxiety. By contrast, the deprived child comes to view the world as cruel and hazardous. An inevitable result is low self-esteem.

The task of therapy is to examine the patient's defective patterns in relating to others. In part this is achieved through study of what transpires between patient and therapist. But there is no emphasis on transference as occurs in Freudian analysis. Therapy aims to enable the patient to move *with* others, by engaging in relationships which are reciprocal and mutual. Another goal is greater self-realization, with freedom from determined modes of thought and action.

While the neo-Freudians were establishing a new pathway for psychoanalysis, a group was equally innovative in the UK and France. In the UK, Melanie Klein (1882–1960), a physician hailing from Berlin, concluded from her study of children's play that the Oedipal complex, as described by Freud, was a late expression of primitive unconscious anxieties which children experienced in the first 18 months of life. Given that these states could be reactivated in adult life, Klein labeled them 'positions' rather than stages of psychological development. She also supported the move away from Freud's formulation of instincts and psychosexual energy toward one of object relations, namely the representations of the perception and experience of significant others. Similarly, transference was to be understood not only as a repetition of past or current relationships, but as evidence of relational patterns in the patient's current *internal* world.

In tandem with Klein, other analysts in Britain proposed ways in which significant early relationships, particularly between infant or young child and its mother, exerted a formative influence on the developing psyche. John Bowlby (1907–90), in particular, observed that children who had suffered a prolonged separation from their mother underwent a grief experience which predisposed them to a range of psychopathology in later life. Donald Winnicott (1896–1971), a pediatrician and analyst, proposed that a mother's attunement to her child, enabling him to feel psychologically 'held' and understood, was an essential ingredient for a stable sense of self. The American psychoanalyst Heinz Kohut advanced a similar concept, but highlighted the place of empathy whereby the mother functions as a 'self object' for the child.

Wilfred Bion, a student of Klein, suggested that the analyst functions as a 'container' into whom the patient projects their 'unprocessed' feelings and experiences; the analyst returns them later in a form accessible to him to work on. This experience is deemed to be crucial to the development of a person's ability to examine his or her own state of mind and that of others.

Michael Balint, an analyst at the Tavistock Clinic, worked along similar lines to Winnicott, focusing on the opportunity for the patient to 'regress' in the safety of the therapeutic relationship to a state of mind in which differentiating between patient and therapist may blur. The features of the encounter – its regularity, predictability, and the therapist's empathic attitude – creates circumstances for such regression which can then be used to examine and alter defense mechanisms.

These models have implications for practice in that the therapist expresses moment-to-moment empathic understanding of the patient's inner experiences and as this is repeated time and again the process promotes the evolution of a coherent identity.

Jacques Lacan (1901–81), the maverick French analyst, claimed a 'return to Freud,' his early work in particular, but through the prisms of semiotics and linguistics. These disciplines, he contended, do not see language as a value-free means to express ideas and convey meaning but as a system of signs that communicate symbolically tacit rules and power arrangements. Language imposes on the child awareness of a separateness from its mother who represents a fantasized internal object (the imaginary mother), in whose eyes the child mirrors itself in an attempt to achieve a state of psychological unity. Lacan's controversial proposals regarding variable duration of treatment and trainees themselves determining their competence to qualify as psychoanalysts led to his expulsion from the International Psychoanalytic Association, albeit with a dedicated following.

Humanistic experientially oriented schools of therapy became popular in the 1960s and 1970s, mainly in the US. Their chief features were: challenging patients with the way they avoid emotionally significant matters in the 'here-and-now'; increasing awareness of nonverbal aspects of communication; facilitating emotional arousal; and providing a forum in which patients were encouraged to experiment with new behaviors.

Some of these schools have incorporated aspects of psychoanalytic theory. Transactional analysis, for instance, makes playful use of the concept of ego states, represented as parent, adult, and child; primal scream therapy is predicated on heightened arousal of the purported emotional trauma of birth; and psychodrama encourages people to enact their conflicts by personifying various roles in their psychological lives.

The most influential humanistic school has been the client-centered therapy of Carl Rogers, although it has been absorbed into psychotherapy generally and not always with appropriate acknowledgment. Its premise is that if the therapist creates an atmosphere that is nonjudgmental, empathic, and warm, people can realize their potential for self-expression and self-fulfillment.

Feminist-based psychotherapies, still evolving since the 1980s, reflect a range of opinion from those which condemn Freud as a misogynist who deliberately recanted his discovery of the actuality of childhood sexual abuse, to those who appreciate his linkage of life experience, innate drives, and corresponding defense mechanisms in influencing vulnerability to psychopathology (Elliot, 1991; Appignanesi and Forrester, 1992). The latter point of view, echoing a broader debate within psychoanalysis of the respective roles of internal and external reality, has led some feminist therapists to stress social reality, and others to stress physical reality.

The Development of Other Modes of Psychotherapy

Our focus, hitherto, has been on psychotherapy of the individual patient. The second half of the twentieth century saw the evolution of treatments conducted with more than one

person: groups of strangers and families. In this section, we provide a brief account of the development of these two modes.

Although therapy of stranger groups was experimented on early in the century, it was not until World War II that major developments ensued. The exigencies of war were the spur in that the group format proved highly economical for dealing with huge numbers of soldier-patients. The Northfield Military Hospital in the UK was a center of remarkable innovation, led by psychoanalytically oriented therapists like Wilfred Bion and S.N. Foulkes. Bion's (1961) influence pervaded subsequent work at the Tavistock Clinic, while Foulkes (1965) was the founding father of the Institute of Group Analysis (both in London).

In the US, the move to the study of group process was spearheaded by social psychologists, particularly Kurt Lewin. A national center was established to train people in human relations and group dynamics. Participants from diverse backgrounds studied group functioning in order to act more effectively in their own work settings. Paralleling the human potential movement in the 1960s, the group movement transferred its focus from group to personal dynamics, and before long the encounter group movement had evolved with its thrust of promoting greater self-awareness and personal growth. Encounter groups became widespread during the 1960s and 1970s but after an initial fervor, declined both in terms of membership and appeal.

Formal group therapy under professional leadership then assumed a more prominent role. The most popular model was fathered by Irvin Yalom who drew upon an interpersonalist approach, one originally molded by Harry Stack Sullivan who proposed that personality is chiefly the product of a person's interaction with other significant people. Yalom pioneered the study of therapeutic factors specific to the group rather than following the pattern of transposing psychoanalytic or other theories from the individual setting. His *Theory and Practice of Group Psychotherapy* (Yalom, 1995) became exceedingly influential, attested to by the appearance of four editions.

Constrained perhaps by Western medicine's focus on the individual patient, psychiatry was slow to develop an interest in the family (Gurman and Kniskern, 1991). Scattered through Freud's writings are interesting comments about family relationships and their possible roles in both individual development and psychopathology. His description of processes like introjection, projection, and identification explained how individual experiences could be transmitted across the generations in a family. Influenced by the work in the UK of Anna Freud, Melanie Klein, and Donald Winnicott, the child guidance movement developed a model of one therapist working with a disturbed child and another with the parents, most often the mother on her own. The two clinicians collaborated in order to recognize how the mother's anxieties distorted her perception and handling of her child, which were added to the child's own anxieties.

Things took a different turn in the US. There, Ackerman (1958) introduced the idea of working with the family of a disturbed child using psychoanalytic methods in the 1950s. An interest in working with the family, including two or more generations, arose concurrently. Thus, Murray Bowen (1978) found that the capacity of psychotic children to differentiate themselves emotionally from their families was impaired by the consequences of unresolved losses, trauma, and other upheavals in the lives of parental and grand-parental generations.

Boszormenyi-Nagy and Spark (1984) also addressed this transgenerational theme by describing how family relationships were organized around a ledger of entitlements and obligations which conferred upon each participant a sense of justice about their position. This, in turn, reflected the experience in childhood of neglect or sacrifices for which redress was sought in adult life.

Bowen also introduced the principles of Systems Theory into his work but it was Salvador Minuchin, working with delinquent youth in New York, who highlighted the relevance of systems thinking to their interventions. The youngsters often came from emotionally deprived families, headed by a demoralized single parent (most often the mother) who alternated between excessive discipline and helpless delegation of family responsibilities to a child. Minuchin's Structural Family Therapy deploys a series of action-oriented techniques and powerful verbal metaphors which enable the therapist to join the family, and to reestablish an appropriate hierarchy and generational boundaries between the family subsystems (marital, parent/child, siblings).

Another major development took place in Palo Alto, California, where a group of clinicians gathered around the anthropologist Gregory Bateson (1972) in the 1950s. They noted that implicit in communication were *meta-communications*, which defined the relationship between the participants. Any contradiction or incongruence carried great persuasive, moral, or coercive force and formed part of what they labeled a 'double-bind'; they proposed this as a basis for schizophrenic thinking.

All these system-oriented views assume that the family is a system observed by the therapist. However, therapists are not value-neutral. They may take an active role in orchestrating change in accordance with a model of family functioning. Yet these models ignore therapists' biases as well as the relevance of their relationships with families. This probably reflected the determination of some American family therapists to distance themselves from psychoanalytic theory, and also led them to neglect the family's history, how it changed throughout the lifecycle, and the significance of past traumatic and other notable life events.

In response to these criticisms there was a move away from the problem-focused approach which had characterized most behavioral and communication views of psychopathology. The Milan school (1980) whose founders were psychoanalysts developed a new method of interviewing families in conjunction with observers behind a one-way screen formulating and then presenting to the family and therapist hypotheses about 'their' system.

Family therapists also began to consider that families might be constrained from experimenting with new solutions because of the way they had interpreted their past experiences or internalized explanatory narratives of their family, the expert's, or society at large. This led to a shift from considering the family as a social to a linguistic system.

The narrative a family conveys about their lives is a construction which organizes past experience in particular ways; other narratives are excluded. When a family with an ill

member talks to a professional, conversations are inevitably about problems (a problem-saturated description). The participants ignore times when problems were absent or minimal, or when they successfully confined problems to manageable proportions. A different story might be told if they were to examine the context that might have led, or could still lead, to better outcomes.

A number of narrative approaches have applied these concepts. Philosophically, they align themselves with post-modernism, which challenges the notion of a basic truth known only by an expert.

Many criticisms of systems approaches have been leveled including:

1. Disregard of the subjective and intersubjective experiences of family members.
2. Neglect of their history.
3. Denial of unconscious motives which affect relationships.
4. Although family members are reciprocally connected, the power they exert on one another is not equal; this is highlighted in the violence against women and in child abuse.
5. Inequality and other forms of injustice based on societal attitudes toward differences in gender, ethnicity, class, and the like, are uncritically accepted as 'givens'.
6. Minimizing the role of the therapeutic relationship, including attitudes members develop toward the therapist and her feelings toward them.

This critique has led to growing interest in integrating systems-oriented and psychoanalytic ideas, particularly those derived from object-relations theory. One variant is John Byng-Hall's (1995) synthesis of attachment theory, systems-thinking, and a narrative approach.

A further criticism of systems-oriented approaches is that they minimize the impact of external reality such as physical disability or biological factors, in the etiology of mental illness, and sociopolitical phenomena like unemployment, racism, and poverty. One result is the 'psychoeducational' approach, which has evolved in the context of the burden schizophrenia places on the family and the potential for its members to influence its course. This has led to a series of interventions including educating the family about the nature, causes, course, and treatment of schizophrenia; providing them with opportunities to discuss their difficulties in caring for the patient, and to devise appropriate strategies; and helping them to resolve conflict related to the illness, which may be aggravated by the demands of caring for a chronically ill person.

The Scientific Era

Systematic research in the psychotherapies was a low priority for many decades; instead, practitioners' interests focused on theory and technique. Investigation of the subject only took off in earnest in the early 1950s. A notable impetus was the critique by H.J. Eysenck (1952) in which he argued that the treatment of neurotic patients was no more effective than no treatment at all. Two-thirds of both groups showed improvement with time. In later reviews, Eysenck was even more damning: the effects of psychoanalysis in particular were minuscule. The attack led to much rancor and an extended battle between the psychoanalytic and behavioral camps (Eysenck featured prominently in the latter). Fortunately, a positive repercussion was the sense of challenge experienced by the analytic group. They had been riding high for several years, particularly in the US, and barely questioned whether their concepts and practice required scientific appraisal. Thus, although Eysenck's interpretation of the research literature was flawed and biased, he had stirred a hornet's nest.

Since the 1950s research has burgeoned and yielded much knowledge about whether psychotherapy works (outcome research) and how it works (process research). These developments were not without incident. Many therapists challenged the relevance of conventional research methodology to psychotherapy. A key argument was that the encounter between therapist and patient is unique, involving two people in a complex collaborative venture, and cannot be subject to the same form of scrutiny as occurs in the natural sciences. Moreover, the latter approach is necessarily mechanistic and reductionistic.

There is merit to the argument but actual research practice reveals that methodology can accommodate, at least in part, respect for the highly personal nature of psychological treatment. Sophisticated statistical procedures can also contribute to maintaining an appreciation that many characteristics in both patient and therapist are relevant, and they cannot be viewed as homogeneous groups. An illustration is the multidimensional measurement of outcome. Instead of restricting this to one or two variables only, several can be examined concurrently which together encompass the patients' internal and interpersonal world, e.g., quality of life, self-awareness, authenticity, self-esteem, target problems, and social adjustment.

Another criticism of outcome research is examining effectiveness without adequate treatment being given. Duration of research-bound therapy in fact is often much briefer than in customary practice because of funding constraints. A similar objection relates to the follow-up period usually set. A robust test of effectiveness entails examining outcome one or more years following the end of therapy in order to judge clinical projects independent of the therapist.

The Effectiveness of Psychotherapy

Several reviews have been conducted since the 1970s. A sophisticated method – meta-analysis – which relies on a rigorous statistical approach, points to psychotherapy overall generating benefits for patients who have been suitably selected. Encouragingly, better designed studies show positive results more commonly than inadequate ones. In one trail-blazing meta-analysis, levels of effectiveness were computed from 375 studies in which some 25 000 experimental and 25 000 controlled patients were tested (Smith et al., 1980). The average patient receiving psychological treatment of one kind of another was better off than 75% of those who were untreated – a clear demonstration of the beneficial effects of treatment in general.

With tens of 'schools' available, many of them claiming a distinctive approach, the obvious question arises as to

whether some are superior in producing benefit. This question of comparative effectiveness is complicated in that not only may 'schools' be compared but also specific procedures such as setting a time limit or not or conducting therapy individually or within a group. Notwithstanding these difficulties, considering research on comparative effectiveness is worthwhile. Several reviews show a consistent pattern – that 'everyone has won and must have prizes'. This was of course the judgment of a race handed down by the dodo bird in *Alice in Wonderland*. In the psychotherapy stakes, it appears that everyone has won too, a finding probably attributable to factors common to all therapeutic approaches. These factors, set out originally by Jerome Frank (1973), include a confiding relationship with a helping person, a rationale which provides an account of the patient's problems and of the methods for remedying them, instillation of hope for change, opportunity for 'success experiences' during treatment, facilitating emotion in the patient, and providing new knowledge, so promoting self-awareness. These 'nonspecific' factors comprise a significant theme in psychotherapy; they probably serve as a platform for benefits from all treatments.

Any consideration of effectiveness brings up the issue of harm. In other words, treatment may be for better or for worse. After a long gestation, the concept of a 'negative effect' attracted widespread attention from the 1970s (Hadley and Strupp, 1976). A growing sense of confidence perhaps permitted therapists to be more open to a potential harmful impact.

To distinguish between a patient becoming worse because of their intrinsic condition or following an adverse life event, a negative effect has been defined as deterioration directly attributable to treatment itself. Such a causal link is difficult to prove but genuine negative effects certainly do occur. The definitional difficulty, however, leads to estimates of prevalence ranging from rare to common. The type of therapy appears to influence the rate. Evidence points to worsening in about 5–10% of cases in psychotherapy generally. The reasons for deterioration are not well established although a common view is that patient factors probably contribute, particularly selecting people for treatments for which they are unsuited. In this situation, refining assessment, thus enhancing clinicians' ability to predict response to specific treatments, is deemed as helpful to reduce the casualty rate.

Another obvious facet is the therapist and/or technique. Here, inadequate training has emerged as salient, with poor judgment leading to inappropriate interventions. It has also been recognized that a proportion of therapists perform poorly because of their personality traits, no matter what their level of training or experience. In other words, the role of therapist does not suit all those who wish to practice. In a noteworthy study of different forms of encounter group, a substantial percentage of 'casualties' were produced by four leaders, who typically pummeled members to disclose personal information and express intense emotion.

During the half century of systematic research, attention has also been devoted to the processes that take place in the therapeutic encounter.

Process and outcome research are inevitably linked. If we can identify factors that promote or hinder effectiveness, we may be able to modify these and then note the result. The advent of recording techniques such as audio and video has facilitated studies of process and a wide range of observations of both verbal and nonverbal behavior, in both therapist and patient and their interaction, have been made. One illustration is the premise that a necessary condition for effective group therapy is group cohesiveness. It follows that anything enhancing cohesiveness could be advantageous. Compatibility between members has been proposed as pertinent and indeed been shown to relate to cohesiveness. Group therapy therefore could conceivably be more effective if, in selecting members, compatibility were taken into account. For example, a patient incompatible with all his peers would presumably not be placed in that group.

Research on process has been shown to be as salient as that on outcome since only with the study of what occurs in treatment can therapists appreciate its inherent nature and the factors for optimal improvement of participants. Diligent process research has provided solid foundations for establishing hypotheses about outcome.

We have dealt in general terms with the theme of psychotherapy research. The second half of the twentieth century has been a fertile period and seen much achieved; to do it justice is beyond our remit. The interested reader is recommended to consult the four editions of the *Handbook of Psychotherapy and Behavior Change* (Bergin and Garfield, 1994) which has served the field in distilling work done and providing a critique of its quality.

See also: Bowlby, John (1907–90); Freud, Sigmund (1856–1939); Janet, Pierre (1859–1947); Jung, Carl Gustav (1875–1961); Klein, Melanie (1882–1960); Psychoanalysis, History of; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported.

Bibliography

Ackerman, N.W., 1958. The Psychodynamics of Family Life. Basic Books, New York.
Appignanesi, L., Forrester, J., 1992. Freud's Women. Basic Books, New York.
Bateson, G., 1972. Steps to an Ecology of Mind. Chandler Publishing Co., San Francisco.
Bergin, A.E., Garfield, S.L. (Eds.), 1994. Handbook of Psychotherapy and Behaviour Change, fourth ed. Wiley, New York.
Bion, W., 1961. Experiences in Groups. Tavistock, London.
Bromberg, W., 1959. The Mind of Man: A History of Psychotherapy and Psychoanalysis. Harper, New York.
Boszormenyi-Nagy, I., Spark, G.M., 1984. Invisible Loyalities: Reciprocity in Intergenerational Family Therapy. Brunner-Mazel, New York.
Bowen, M., 1978. Family Therapy in Clinical Practice. Jason Aronson, New York.
Byng-Hall, J., 1995. Rewriting Family Scripts. Improvisation and Systems Change. Guilford, New York.
Ellenberger, H., 1970. The Discovery of the Unconscious. Basic Books, New York.
Elliot, P., 1991. From Mastery to Analysis, Theories of Gender in Psychoanalytic Feminism. Cornell University Press, Ithaca, NY.
Eysenck, H.J., 1952. The effects of psychotherapy. Journal of Consulting Psychology 16, 319–324.
Foulkes, S.H., Anthony, E.J., 1965. Group Psychotherapy: The Psychoanalytic Approach. Penguin, Harmondsworth, UK.
Frank, J.D., 1973. Persuasion and Healing: A Comparative Study of Psychotherapy. Johns Hopkins University Press, Baltimore, MD.
Gurman, A.S., Kniskern, D.P. (Eds.), 1991. Handbook of Family Therapy, vol. II. Brunner/Mazel, New York.

Hadley, S.W., Strupp, H.H., 1976. Contemporary views of negative effects in psychotherapy. Archives of General Psychiatry 33, 1291–1302.
Mitchell, S.A., Black, M.J., 1995. Freud and beyond: A History of Modern Psychoanalytic Thought. Basic Books, New York.
Selvini-Palzzoli, M., Boscolo, L., Cecchin, G., Prata, G., 1980. Hypothesizing-circularity-neutrality: three guidelines for the conductor of the session. Family Process 19, 3–12.
Smith, M.L., Glass, G.V., Miller, T.I., 1980. The Benefits of Psychotherapy. Johns Hopkins University Press, Baltimore, MD.
Yalom, I.D., 1995. The Theory and Practice of Group Psychotherapy, fourth ed. Basic Books, New York.

Psychotherapy Integration

John C Norcross, University of Scranton, Scranton, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychotherapy integration is characterized by a desire to look across and beyond theories to see how patients can benefit from multiple ways of conducting psychotherapy. Formal ideas on synthesizing therapies appeared in the 1930s, but it is only since the 1980s that integration developed formally. A number of mutually reinforcing motives that have fostered the recent growth of integration are reviewed. The most frequent integrative pathways – technical eclecticism, theoretical integration, common factors, and assimilative integration – are considered and compared. Although integration faces several obstacles, its research support and clinical popularity suggest that it will probably constitute the future of the discipline.

Psychotherapy integration is characterized by dissatisfaction with single-school approaches and a concomitant desire to look across school boundaries to see how patients can benefit from other ways of conducting psychotherapy. Although various labels are applied to this international movement – including integration, eclecticism, rapprochement, convergence, responsiveness, and prescriptionism – the goals are similar indeed. The ultimate goal is to enhance the efficacy and applicability of psychotherapy.

Applying identical psychosocial treatments to all patients in all circumstances is now recognized as inappropriate and perhaps unethical. Different folks require different strokes. The efficacy of psychotherapy will be enhanced by tailoring it to the unique needs of the client, not by imposing procrustean methods on unwitting consumers of psychological services. The integrative mandate is embodied in Gordon Paul's (1967) iconic question: *What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem and under *which* set of circumstances?

Any number of indicators attests to the popularity of psychotherapy integration. *Eclecticism*, or the increasingly favored term *integration*, is the first or second most frequent theoretical orientation of English-speaking psychotherapists (e.g., Norcross and Karpiak, 2012). It is less popular but increasing in other countries. Leading psychotherapy textbooks routinely identify their theoretical persuasion as integrative, and an integrative chapter is regularly included in compendia of treatment approaches. The publication of books that synthesize various therapeutic concepts and methods continues unabated; they now number in the hundreds. Handbooks on psychotherapy integration have now been published in at least a dozen countries (e.g., Norcross and Goldfried, 2005; Stricker and Gold, 2006).

The prevalence of psychotherapy integration can be ascertained directly by assessing endorsement of the integrative orientation (as noted in this article) or gleaned indirectly by determining endorsement of multiple orientations. For example, in a study of British counselors, 87% did *not* take a pure-form approach to psychotherapy (Hollanders and McLeod, 1999). In a study of clinical psychologists in the United States, for another example, fully 90% embraced several orientations (Norcross and Karpiak, 2012). Very few contemporary therapists adhere exclusively to a single therapeutic tradition.

The Historical Context

Integration as a point of view has probably existed as long as philosophy and psychotherapy. In philosophy, the third-century biographer Diogenes Laertius referred to an eclectic school that flourished in Alexandria in the second century (Lunde, 1974). In psychotherapy, Freud consciously struggled with the selection and integration of diverse methods. As early as 1919, he introduced psychoanalytic psychotherapy as an alternative to classical psychoanalysis in recognition that the more rarified approach lacked universal applicability (Liff, 1992).

However, the discipline of psychotherapy historically has been organized around singular and rival theories, such as psychoanalysis, existentialism, behaviorism, and family systems. In the infancy of the field, therapy systems, like battling siblings, competed for attention and affection in a 'dogma eat dogma' environment. Clinicians traditionally operated from within their own particular theoretical frameworks, often to the point of being blind to alternative conceptualizations and potentially superior interventions.

Formal ideas on synthesizing the psychotherapies appeared in the literature in the 1930s (Goldfried et al., 2005). Thomas French (1933) stood before the 1932 meeting of the American Psychiatric Association and drew parallels between certain concepts of Freud and of Pavlov. In 1936, Sol Rosenzweig published an article that highlighted commonalities among various systems of psychotherapy. These and other early attempts at integration, however, were largely theory driven and empirically untested.

If not conspiratorially ignored altogether, these precursors to integration appeared only as a latent theme in a field organized around discrete theoretical orientations. Although psychotherapists secretly recognized that their orientations did not adequately assist them in all they encountered in practice, a host of political, social, and economic forces – such as professional organizations, training institutes, and referral networks – kept them penned within their own theoretical schoolyards and typically led them to avoid clinical contributions from alternative orientations.

Systematic integration was probably inaugurated in the modern era by Frederick Thorne (1957), who is credited with being the grandfather of eclecticism in psychotherapy. Persuasively arguing that any skilled professional should

come prepared with more than one tool, Thorne emphasized the need for clinicians to fill their toolboxes with methods drawn from many theoretical orientations. He likened psychotherapy at that time to a plumber who used only a screwdriver. Like such a plumber, inveterate psychotherapists applied the same treatment to all people, regardless of individual differences, and expected the patient to adapt to the therapist rather than vice versa.

Since the mid-1980s, psychotherapy has witnessed a decline in ideological struggle and the stirrings of rapprochement. The theoretical substrate of each system is undergoing intensive reappraisal, as psychotherapists acknowledge the inadequacies of any one system and the potential value of others. Distinctive of the present era is the tolerance for and assimilation of formulations that were once viewed as deviant, or even heretical.

Integration Defined

As the field of psychotherapy has matured, integration has emerged as a dominating force. The dramatic interest in psychotherapy integration has crystallized into a formal movement, or, more dramatically, a 'metamorphosis' in mental health (London, 1988). At the very least, integration represents a revolutionary departure from decades of school-ism.

Psychotherapy integration is characterized by dissatisfaction with single-school approaches and a concomitant desire to look across and beyond school boundaries to see what can be learned – and how patients can benefit – from other ways of thinking about psychotherapy and behavior change. The ultimate outcome of doing so, which is not yet fully realized, is to enhance the effectiveness, efficiency, and applicability of psychotherapy.

As with many new movements, integration has been plagued by confusing vocabulary and multiple variants. Various labels are applied to the movement, but *psychotherapy integration* has emerged as the overarching consensual term. Similarly, a plethora of integrations have been advanced: combining different therapy formats (such as individual, family, and group therapies), blending self-help and psychotherapy, integrating pharmacotherapy and psychotherapy, synthesizing research and practice, integrating Western and Eastern perspectives, and so on. This article confines itself to the traditional and primary meaning of integration as the blending of diverse theoretical orientations or schools of psychotherapy.

Reasons for Recent Growth

A number of interacting, mutually reinforcing motives have fostered the recent growth of integration (Prochaska and Norcross, 2014):

- *Proliferation of therapies*. Which of 500-plus therapies should be studied, taught, or bought? The hyperinflation of brand-name therapies has produced narcissistic fatigue: "With so many brand names around that no one can recognize, let alone remember, and so many competitors doing psychotherapy, it is becoming too arduous to launch still another new brand" (London, 1988: pp. 5–6). This might also be called the 'exhaustion theory' of integration: Peace among warring schools is the last resort.
- *Inadequacy of any single theory for all patients and problems*. No therapy or therapist is immune to failure. It is at such times that experienced practitioners often wonder if the clinical methods from orientations other than their own might have been more effective. Underlying the ecumenical spirit is the stark realization that narrow conceptual positions and limited clinical methods cannot begin to address the diverse needs and disorders of our patients. Clinical realities have come to demand a more flexible, if not integrative, perspective.
- *External socioeconomic contingencies*. Psychotherapy has experienced mounting pressures from not easily disregarded sources such as policy makers, informed consumers, and insurance companies. Everyone is demanding hard research evidence of effective treatments. Unless they demonstrate accountability, psychotherapists stand to lose prestige, customers, and money. Attacks from outside the mental health professions have started to propel them together. There is something to be said for having the different therapies 'hang together' rather than 'hang separately.'
- *Ascendancy of short-term, problem-focused treatments*. The brief problem focus has brought divergent therapies closer together and has created variations of different treatments that are more compatible with each other. Integration, particularly in the form of eclecticism, responds to the pragmatic time-limited injunction of "whatever therapy works better – and quicker – for this patient with this problem."
- *Opportunities for therapists to observe and experiment with various treatments*. The establishment of specialized clinics for the treatment of specific disorders – personality disorders, trauma, and eating disorders, to name a few – has afforded exposure to other therapies, and stimulated some therapists to consider the other therapies more seriously. Treatment manuals have also induced an informal version of 'theoretical exposure': Previously feared and unknown therapies were approached gradually, anxiety was dissipated, and the previously feared therapies were integrated into the clinical repertoire.
- *Recognition that therapeutic commonalities heavily contribute to treatment outcome*. Only about 8–12% of outcome is generally accounted for by the specific treatment method. Therapeutic success can be best predicted by common elements of psychotherapy, such as the contributions of the patient, the therapy relationship, and the facilitative qualities of the therapist.
- *Development of professional societies for integration*. The creation of professional networks has been both a consequence and cause of interest in psychotherapy integration. Several organizations, principally the Society for the Exploration of Psychotherapy Integration, have brought together integration enthusiasts through conferences, networking, and journals. Integrationists now have a professional home.

These factors speak to the entire field of psychotherapy moving toward integration. But what motivates individual

therapists to embrace integration? Robertson (1979) identified six personal factors that facilitate the acquisition of an eclectic or integrative stance. The first is a lack of pressure in training to adopt a doctrinaire position and the absence of a charismatic figure to emulate. The second factor, which has been substantiated by research (Norcross, 2005), is length of clinical experience: As therapists encounter complex clients and problems over time, they are more likely to reject a single theory as too simple. A third factor is the extent to which doing psychotherapy is making a living or making a philosophy of life; Robertson asserts that integration is more likely to follow the former.

The remaining three factors are personality variables: an obsessive–compulsive desire to give order to all the interventions of the therapeutic universe, a maverick temperament to move beyond a theoretical camp, and a skeptical attitude toward the status quo. These personality traits are frequently found in the personal histories of prominent integrative therapists (e.g., see, Goldfried, 2001).

Pathways to Integration

There are numerous pathways toward the integration of the psychotherapies. The four most popular routes appear to be technical eclecticism, theoretical integration, common factors, and assimilative integration. Research (Norcross et al., 2005) reveals that each is embraced by a considerable number of self-identified eclectics and integrationists (19–28% each). All four attempt to look beyond the confines of single theories, but they do so in rather different ways and at different levels.

Technical eclecticism seeks to improve our ability to select the best treatment methods for the person and the problem. This search is guided primarily by research on what specific methods have worked best in the past with similar problems and patient characteristics. Eclecticism focuses on predicting for whom interventions will work; its foundation is actuarial rather than theoretical. Exemplars of eclecticism include Lazarus's (1967, 1989) influential multimodal therapy and Beutler et al.'s (2000) systematic treatment selection.

Technical eclectics use procedures drawn from different therapeutic systems without necessarily subscribing to the theories that spawned them, whereas theoretical integrationists draw their concepts and techniques from diverse systems that may be epistemologically or ontologically incompatible. For technical eclectics, no necessary connection exists between conceptual foundations and techniques. "To attempt a theoretical rapprochement is as futile as trying to picture the edge of the universe. But to read through the vast amount of literature on psychotherapy, *in search of techniques*, can be clinically enriching and therapeutically rewarding" (Lazarus, 1967: p. 416).

In *theoretical integration*, two or more therapies are united with the hope that the hybrid result will be better than the constituent therapies alone. As the name implies, there is an emphasis on integrating, or at least bridging, the underlying theories of psychotherapy along with the techniques from each. Wachtel's (1977) integration of psychoanalysis and behaviorism, as well as Prochaska and DiClemente's transtheoretical (or stages of change) model (Prochaska et al., 1994), illustrate this path. Some would also add here the infusion of acceptance and mindfulness of the 'third-wave' therapies to cognitive-behavior therapy.

Theoretical integration involves a commitment to a conceptual or theoretical creation beyond a technical blend of methods. The goal is to create a conceptual framework that synthesizes the best elements of two or more therapies. Integration aspires to be more than a simple combination; it seeks an emergent theory that is more than the sum of its parts.

The *common factors* approach seeks to identify core ingredients shared by different therapies with the eventual goal of creating more parsimonious and efficacious treatments based on those commonalities. This search is predicated on the belief that commonalities are more important in accounting for therapy success than the unique factors that differentiate among them. The research-informed models of Jerome Frank (1973) and Bruce Wampold (2001) exemplify the common factors route.

Literature reviews have discovered that the most consensual commonalties across theories are the development of a therapeutic alliance, opportunities for catharsis, acquisition and practice of new behaviors, and clients' positive expectancies (Grencavage and Norcross, 1990).

Marvin Goldfried (1980), a leader of the integration movement, argued,

> [to] the extent that clinicians of varying orientations are able to arrive at a common set of strategies, it is likely that what emerges will consist of robust phenomena, as they have managed to survive the distortions imposed by the therapists' varying theoretical biases. (p. 996)

In specifying what is common across orientations, we may also be selecting what works best among them.

The fourth and most contentious pathway is *assimilative integration*, which entails not only a firm grounding in one system of psychotherapy but also a willingness to selectively incorporate (assimilate) practices and views from other systems (Messer, 2001). In doing so, assimilative integration combines the advantages of a single, coherent theoretical system with the flexibility of a broader range of technical interventions from multiple systems. A cognitive therapist, for instance, might use the gestalt two-chair dialogue or a systemic paradoxical directive in an otherwise cognitive course of treatment.

To its proponents, assimilative integration is a realistic way station on the path to a sophisticated integration; to its detractors, it is a waste station of people unwilling to commit themselves to a full evidence-based eclecticism. Both camps agree that assimilation is a tentative step toward full integration: Most therapists gradually incorporate parts and methods of other approaches once they discover the limitations of their original approach. Inevitably, therapists gradually integrate new methods into their home theory.

These four integrative pathways are not mutually exclusive, of course. No technical eclectic can disregard theory, and no theoretical integrationist can ignore technique. Without some commonalities among different schools of psychotherapy, theoretical integration would prove impossible. And even the most ardent proponents of common factors cannot practice

'nonspecifically' or 'commonly' on their own; specific methods must be applied.

While the field has not settled on a single route to the integrative summit, it has firmly rejected *syncretism* – uncritical and unsystematic combinations. This haphazard approach is primarily an outgrowth of pet techniques and inadequate training. It is an arbitrary blend of methods without systematic rationale or empirical verification (Eysenck, 1970). Psychotherapy integration, by contrast, is the product of years of painstaking training, research, and experience. It is integration by design, not default; that is, clinicians competent in several therapeutic systems who systematically select treatment methods and therapeutic relationships on the basis of outcome research and patient need.

Research Support

The research support for the effectiveness of psychotherapy integration comes in several guises. First and most generally, the entire body of empirical research on psychotherapy informs treatment selection. A genuine advantage of being an eclectic is the vast amount of research attesting to the efficacy of psychotherapy and pointing to its differential effectiveness with certain types of disorders and patients. Of course, this research translates into improved outcomes only if the psychotherapist is aware of, and adheres to, the research conclusions. In this respect, a natural affinity has arisen between integration and evidence-based practice.

Controlled outcome research has been conducted on integrative therapies. A comprehensive review of the adult literature (Grawe et al., 1998) located 22 controlled studies covering 1743 patients treated with diverse therapies described as eclectic: either explicitly no school connection or multiple combinations of methods. In 9 of the 13 comparisons, the eclectic therapy outperformed the control treatment in terms of symptom relief, and in four of the six comparisons in terms of subjective well-being. A comprehensive meta-analysis of the child literature (Weisz et al., 1995) located 20 controlled studies on mixed treatments, which produced a respectable effect size of 0.63 compared to control conditions, a moderate to large effect. The interpretative problem, however, is that we possess little understanding of what these multifarious treatments represent.

A corrective strategy would be to assess whether more complex and integrative treatments outperform simpler and single-theory treatments. One clever meta-analysis did just that (Stevens et al., 2007). In 14 studies, more complex therapies were compared with less complex treatments. The effect size for the complex therapies was slightly higher (0.15), which translates into an increase in improvement rates from 66 to 81%.

More specifically, programmatic research supports the effectiveness of matching different treatment methods and relationship stances to different people. All the patient qualities that predict differential effectiveness have not yet been determined, but meta-analytic research demonstrates that tailoring therapy to patients' transdiagnostic characteristics demonstrably improves the outcome of psychosocial treatments (Norcross, 2011).

For example, matching therapist directiveness to the client's degree of resistance improves therapy outcome. In 80% of the studies, clients presenting with high resistance benefit more from self-control methods and minimal therapist directiveness, such as in motivational interviewing. By contrast, clients with low resistance benefit more from therapist directiveness and explicit guidance. This strong, consistent finding can be expressed as a large effect size averaging 0.76 (Beutler et al., 2011). The clinical implication is to match the level of the therapist's directiveness to the patient's level of resistance.

For another example, 30 years of research in behavior change and psychotherapy converge in showing that different processes of change are differentially effective in certain stages of change (precontemplation, contemplation, preparation, action, and maintenance). A meta-analysis (Rosen, 2000) of 47 cross-sectional studies examining the relationships among the stages and the processes of change showed large effect sizes ($d = 0.70$–0.80). Methods associated with psychoanalytic and insight-oriented psychotherapies are most useful during the earlier precontemplation and contemplation stages. Existential, cognitive, and interpersonal therapies are particularly well suited to the preparation and action stages. Behavioral, exposure, and solution-focused therapies are most useful during action and maintenance. Each therapy system has a place, a differential place, in the big picture of behavior change (Norcross et al., 2011).

These and other research-supported prescriptive matches improve the effectiveness of psychotherapy, integrative and otherwise. How directive should a therapist be? Which treatment approach – insight oriented or symptom focused – works better? In the integrative tradition, it depends upon the particular client.

Future Directions

Transtheoretical dialogue and cross-fertilization fostered by the integrative spirit have produced new ways of thinking about psychotherapy and researching behavior change. This closing section outlines probable directions of the integration movement.

First and foremost, psychotherapy integration, in some guise, will likely emerge as the Zeitgeist. A dispassionate panel of psychotherapy experts portended its escalating popularity in the coming decade (Norcross et al., 2013). The integrative fervor will apparently persist well into the twenty-first century.

Integrationists will increasingly attend to the potential obstacles and tradeoffs of the process of integration. These will include establishing ambitious training programs and processes that ensure student competence in multiple theories and interventions, an unprecedented task in the history of psychotherapy; acknowledging the underlying epistemological and ontological differences embodied by the disparate theories of psychotherapies, which precludes facile integration; and addressing the language problem that confounds understanding of disparate jargons and, in some cases, leads to active avoidance of theoretical constructs.

At the same time, psychotherapy integration will evolve past its initial step of widening therapeutic horizons and embracing

multiple treatments. The next step is to specifically know when and where to use these multiple techniques and theories in order to enhance outcomes. While the emphasis has been traditionally on matching Treatment X to Disorder Y, the integration movement is more and more committed to adapting Therapeutic Relationship X to Person Y. Paralleling the notion of treatment of choice in terms of techniques, clinicians will determine the relationship of choice in terms of interpersonal stances for individual clients and their singular situations. Such tailoring or adapting will occur on the basis of diagnostic and especially cross-diagnostic considerations, such as the patient's preferences, culture, coping style, personality traits, and stage of change (Norcross, 2011).

Conclusion

Integration can take several different paths – theoretical integration, technical eclecticism, common factors, and assimilative integration – but it consistently searches for new ways of conceptualizing and conducting psychotherapy that go beyond the confines of single schools. Integration encourages practitioners and researchers to examine what other therapies have to offer, particularly when confronted with difficult cases and therapeutic failures. Rival therapy systems are increasingly viewed not as adversaries, but as welcome partners (Landsman, 1974); not as contradictory, but as complementary.

Integrative psychotherapies are intellectually vibrant, clinically popular, and, when evaluated, demonstrably effective. Integration converges with the evidence-based movement in emphasizing that different people require different solutions and that these solutions increasingly can be selected on the basis of empirical research. It offers the evidence, flexibility, and responsiveness to meet the multifarious needs of individual patients and their unique contexts. For these reasons, integration will assuredly be a therapeutic mainstay of the twenty-first century. In that ideal future, psychotherapy will be defined not by its theoretical orientations or brand names but by its effectiveness and applicability (Norcross and Beutler, 2013).

See also: Classical Conditioning Methods in Psychotherapy; Concepts and Methods of Cognitive Therapies; Constructivist Theory and Psychotherapy; Experiential Psychotherapy; Exposure Therapies and Stress Inoculation: A Brief Overview; Feminist Psychotherapy; Group Psychotherapy, Clinical Psychology of; Hypnosis; Imaging Methods in Psychotherapy; Interpersonal Psychotherapy; Manual Based Treatments; Motivation in Clinical Interventions; Multimodal Therapy in Clinical Psychology; Narrative Exposure Therapy; Psychodynamic Therapy: Short-term; Psychological Treatment, Effectiveness of; Psychological Treatments Delivered Nontraditionally; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Process Research; Psychotherapy and Hermeneutics; Virtual Reality in Psychotherapy.

Bibliography

Beutler, L.E., Clarkin, J., Bongar, B., 2000. Guidelines for the Systematic Treatment of the Depressed Patient. Oxford University Press, New York.

Beutler, L.E., Harwood, T.M., Michelson, A., Song, X., Holman, J., 2011. Reactance/resistance. In: Norcross, J.C. (Ed.), Psychotherapy Relationships that Work, second ed. Oxford University Press, New York, pp. 261–278.

Eysenck, H.J., 1970. A mish-mash of theories. International Journal of Psychiatry 9, 140–146.

Frank, J.D., 1973. Persuasion and Healing, second ed. Johns Hopkins University, Baltimore, MD.

French, T.M., 1933. Interrelations between psychoanalysis and the experimental work of Pavlov. American Journal of Psychiatry 89, 1165–1203.

Goldfried, M.R., 1980. Toward the delineation of therapeutic change principles. American Psychologist 35, 991–999.

Goldfried, M.R. (Ed.), 2001. How Therapists Change: Personal and Professional Reflections. American Psychological Association, Washington, DC.

Goldfried, M.R., Pachankis, J.E., Bell, A.C., 2005. History of psychotherapy integration. In: Norcross, J.C., Goldfried, M.R. (Eds.), Handbook of Psychotherapy Integration, second ed. Oxford University Press, New York, pp. 24–60.

Grawe, K., Donati, R., Bernauer, F., 1998. Psychotherapy in Transition. Hogrefe & Huber, Seattle, WA.

Grencavage, L.M., Norcross, J.C., 1990. Where are the commonalities among the therapeutic common factors? Professional Psychology: Research and Practice 21, 372–378.

Hollanders, H., McLeod, J., 1999. Theoretical orientation and reported practice: a survey of eclecticism among counsellors in Britain. British Journal of Guidance and Counselling 27, 405–414.

Landsman, J.T., August 1974. Not an Adversity but a Welcome Diversity. Paper presented at the meeting of the American Psychological Association, New Orleans, LA.

Lazarus, A.A., 1967. In support of technical eclecticism. Psychological Reports 21, 415–416.

Lazarus, A.A., 1989. The Practice of Multimodal Therapy. Johns Hopkins University, Baltimore, MD.

Liff, Z.A., 1992. Psychoanalysis and dynamic techniques. In: Freedheim, D.K. (Ed.), History of Psychotherapy. American Psychological Association, Washington, DC, pp. 571–586.

London, P., 1988. Metamorphosis in psychotherapy: slouching toward integration. Journal of Integrative and Eclectic Psychotherapy 7, 3–12.

Lunde, D.T., 1974. Eclectic and integrated theory: Gordon Allport and others. In: Burton, A. (Ed.), Operational Theories of Personality. Brunner/Mazel, New York, pp. 381–404.

Messer, S.B., 2001. Introduction to the special issue on assimilative integration. Journal of Psychotherapy Integration 11, 1–4.

Norcross, J.C., 2005. The psychotherapist's own psychotherapy: educating and developing psychologists. American Psychologist 60, 840–850.

Norcross, J.C. (Ed.), 2011. Psychotherapy Relationships that Work, second ed. Oxford University Press, New York.

Norcross, J.C., Beutler, L.E., 2013. Evidence-based relationships and responsiveness for depression with substance abuse. In: Barlow, D.H. (Ed.), Clinical Handbook of Psychological Disorders, fifth ed. Guilford, New York.

Norcross, J.C., Goldfried, M.R. (Eds.), 2005. Handbook of Psychotherapy Integration, second ed. Oxford University Press, New York.

Norcross, J.C., Karpiak, C.P., 2012. Clinical psychologists in the 2010s: fifty years of the APA division of clinical psychology. Clinical Psychology: Science and Practice 19, 1–12.

Norcross, J.C., Karpiak, C.P., Lister, K.M., 2005. What's an integrationist? A study of self-identified integrative and (occasionally) eclectic psychologists. Journal of Clinical Psychology 61, 1587–1594.

Norcross, J.C., Krebs, P.M., Prochaska, J.O., 2011. Stages of change. In: Norcross, J.C. (Ed.), Psychotherapy Relationships that Work, second ed. Oxford University Press, New York, pp. 279–300.

Norcross, J.C., Pfund, R.A., Prochaska, J.O., 2013. Psychotherapy in 2022: A Delphi poll on its future. Professional Psychology: Research & Practice 44, 363–370.

Paul, G.L., 1967. Strategy of outcome research in psychotherapy. Journal of Consulting Psychology 31, 109–118.

Prochaska, J.O., Norcross, J.C., 2014. Systems of Psychotherapy: A Transtheoretical Analysis, eighth ed. Brooks/Cole, Pacific Grove, CA.

Prochaska, J.O., Norcross, J.C., DiClemente, C.C., 1994. Changing for Good. William Morrow, New York.

Robertson, M., 1979. Some observations from an eclectic therapist. Psychotherapy 16, 18–21.
Rosen, C.S., 2000. Is the sequencing of change processes by stage consistent across health problems? A meta-analysis. Health Psychology 19, 593–604.
Rosenzweig, S., 1936. Some implicit common factors in diverse methods in psychotherapy. American Journal of Orthopsychiatry 6, 412–415.
Stevens, S.E., Hynan, M.T., Allen, M., et al., 2007. Are complex psychotherapies more effective than biofeedback, progressive muscle relaxation, or both? A meta-analysis. Psychological Reports 100, 303–324.
Stricker, G., Gold, J. (Eds.), 2006. Casebook of Psychotherapy Integration. American Psychological Association, Washington, DC.
Thorne, F.C., 1957. Critique of recent developments in personality counseling theory. Journal of Clinical Psychology 13, 234–244.
Wachtel, P.L., 1977. Psychoanalysis and Behavior Therapy: Toward an Integration. Basic Books, New York.
Wampold, B.E., 2001. The Great Psychotherapy Debate: Models, Methods, and Findings. Lawrence Erlbaum, Mahwah, NJ.
Weisz, J.R., Weiss, B., Han, S.S., Granger, D.A., Morton, T., 1995. Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. Psychological Bulletin 117, 450–468.

Relevant Websites

sepiweb.org/ – The homepage of the Society for the Exploration of Psychotherapy Integration (SEPI), the foremost integration organization.

www.uri.edu/research/cprc/ and www.ChangeologyBook.com – These sites provide resources, references, and self-assessments on the stages of change and transtheoretical model.

www.innerlife.com/ – This website provides an online treatment-matching system, rooted in Systematic Treatment Selection, for use by patients and professionals.

Psychotherapy Process Research

David Orlinsky, University of Chicago, Chicago, IL, USA
Erkki Heinonen, National Institute for Health and Welfare, Helsinki, Finland
Armin Hartmann, Universitätsklinikum Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.E. Orlinsky, volume 18, pp. 12499–12504, © 2001, Elsevier Ltd.

Abstract

Psychotherapy process research examines the interactions and experiences of patients and therapists in psychosocial mental health treatments. Early psychotherapy researchers differentiated between 'outcome' studies that focus on the impact of treatment on patients, and 'process' studies that examine the events of therapy. Process studies typically focus on verbal and nonverbal communications between patients and therapists in sessions, treatment methods, qualities of the patient–therapist relationship, and experiences of patients and therapists during and between therapy sessions. This article discusses basic concepts and methods of process research, the main findings linking process to outcome, and new areas of research.

Psychotherapy process research is the scientific study of what happens in psychotherapy. It involves the precise, systematic, and controlled observation and analysis of the events that constitute psychotherapy. The basic elements of psychotherapy are simple – a patient, the distress or illness experienced by the patient, a therapist, and the treatment that the therapist provides – but psychotherapy occurs in a wide variety of forms and settings, involves a wide diversity of practitioners, and is conducted according to widely divergent theories about the causes and cures of psychological illness and distress. Since psychotherapy process research takes all these variations as its subject, a brief overview of the field it studies may serve as an introduction.

Psychotherapy

The forms of psychotherapy most commonly include 'individual therapy' (one therapist with one patient), 'couple therapy' (one therapist with two patients who are married or partners in a relationship), 'family therapy' (one or more therapists with members of a multigenerational family), and 'group therapy' (one or two therapists with several patients who were previously strangers).

The settings in which psychotherapy is practiced include community mental health centers; private professional practices; university or college counseling centers, research clinics, and psychotherapy training centers; inpatient or outpatient psychiatric facilities; social work agencies; corporate employee assistance programs; and church-sponsored counseling programs.

The practitioners of psychotherapy are similarly diverse in background and training. Psychotherapists are not members of a single profession, but instead have typically been trained in one of the following: clinical or counseling psychology, clinical social work, psychiatry, psychosomatics, general medicine, psychiatric nursing, marital counseling or other counseling specialties, or pastoral care. The level of involvement of different professions in psychotherapeutic practice, and their relative standing vis-à-vis one another, varies from country to country.

Still, the most striking difference among psychotherapeutic practitioners is the diversity of theoretical models used in guiding their clinical work. Broadly, these models include variations on the following: analytic and psychodynamic theories; behavioral, cognitive, and cognitive-behavioral procedures; existential, experiential, and humanistic approaches; hypnotherapy; and interpersonal, relational, and family systems orientations. Initially, these theories formed the bases for rival 'schools' of psychotherapy. More recently, many therapists have adopted an eclectic or integrative mode of practice, seeking to broaden their effectiveness by combining the strengths of different models or using those that seem to suit the needs of particular clients (e.g., Goldfried et al., 2011). The increasingly multicultural nature of modern countries has also created pressure on practitioners to adapt their traditional treatment models to fit the needs and circumstances of particular socioeconomic, cultural, ethnic, and linguistic groups (Zane et al., 2004).

For simplicity in discussion, all the mental health practices based on interpersonal influence through the application of psychological methods to improve the mental health and emotional well-being of distressed or troubled persons will be termed psychotherapy – whether their practitioners call it psychotherapy, therapeutic counseling, psychoanalysis, behavior modification, growth facilitation, or some other term. Similarly, all the mental health workers who engage in these practices professionally will be referred to as psychotherapists, and those whom they treat will be referred to as patients.

Psychotherapy Research

The scientific study of psychotherapy emerged as a field of academic research in the mid-twentieth century, following World War II, through the pioneering work of teams of investigators such as those led by Carl Rogers at Ohio State University and the University of Chicago, Jerome Frank at Johns Hopkins University, and researchers at the Menninger Clinic. Readers may consult Castonguay et al. (2010) for valuable biographical accounts of these pioneers and other later significant contributors.

During the decades since the 1950s and 1960s, the field of psychotherapy research continued to grow at an increasing pace

and volume of studies (Orlinsky and Russell, 1994; Strupp and Howard, 1992) to the present (cf Norcross et al., 2011: chapters 10–12). Within the field, a distinction soon grew up between two kinds of research, depending on the principal focus: 'psychotherapy outcome research,' which studied the impact of therapy on patients after treatment concludes (called efficacy or effectiveness, respectively, based on whether treatment was conducted in laboratory or naturalistic settings); and 'psychotherapy process research,' which focused instead on the interactions and experiences of patients and therapists during therapy.

Even from the outset, however, some studies included measures both of process and outcome, and over time 'psychotherapy process–outcome research' has become an increasingly important focus of inquiry because such studies link specific aspects and kinds of events in therapy to their immediate and long-term impact. After an extensive review of three decades of findings from process–outcome research, Orlinsky and Howard (1986) presented an integrative conceptualization broadly based on a twofold distinction: first, between therapeutic processes (the observable and experienced aspects and events of therapy) and their functional contexts (the psychological, social, and cultural circumstances in which therapy takes place); and second, between the states of these contexts as they exist prior to therapy or to specific moments of process within therapy (e.g., before a particular intervention) – the 'inputs' or determinants of process – and those contexts as they exist subsequent to specific moments of therapy process (e.g., a patient's experience of insight) or to the course of therapy as a whole – the 'outputs' or consequences of therapy process.

The inputs or determinants of process include, most immediately, the psychological characteristics and personal problems that patients bring to therapy, as well as the theoretical treatment model that therapists apply in therapy, their professional skills in doing so, and their personal manner and presence with patients. Less directly, but of potential importance, the inputs or determinants of process would also include the social characteristics of the treatment setting; the nature and state of the mental health service delivery system of which it is a part; the current economic, political, and societal pressures and stressors upon them; and the prevailing cultural standards regarding selfhood, well-being, and personal maturity (e.g., Kirmayer, 2007).

The outputs or consequences of therapy processes include, most immediately, the changes that occur (usually positive but possibly negative) in the patient's life and personality (clinical 'outcome'); they also include, viewed in this broad framework, the impact of therapeutic events on the professional and personal life of the therapist (e.g., growth in experience and expertise vs emotional depletion and burnout), and, less directly, the aggregate influence of patients' and therapists' experiences on social institutions and cultural values.

This conceptualization of therapy as involving contextual influences, therapeutic processes, and contextual consequences has served for nearly two decades to integrate the findings of many research studies into a coherent body of knowledge (Orlinsky et al., 1994; Orlinsky and Howard, 1986; Orlinsky et al., 2004). It was called the 'generic model of psychotherapy' to reflect its use as a framework that can encompass and can be used to compare the varied types or species of psychotherapy (e.g., Orlinsky and Howard, 1995). Between the contextual inputs or determinants and the contextual outputs or consequences lies the complex set of psychological interactions and transformations that comprise the psychotherapeutic process. The generic model includes these as specific variables grouped in the following process categories: (1) the therapeutic contract or organizational aspect of process (understanding between participants with respect to treatment goals, setting, schedule, etc.); (2) therapeutic operations, the technical–instrumental aspect of process (participants' engagement in specific treatment interventions, in their roles as therapist and patient); (3) the therapeutic bond or interpersonal aspect of therapy (the quality of rapport and feelings of participants toward one another as persons); (4) psychological self-relatedness, the reflexive-intrapersonal aspect of process (the self-awareness, self-management, and self-feelings of the participants as therapy proceeds); (5) in-session impacts, the clinical aspect of process (e.g., enhanced insight and morale for patients and sense of efficacy and work satisfaction for therapists); and, as these unfold over time, (6) the sequential process patterns, the temporal aspect of therapy (including interaction sequences within sessions, between sessions, and over the course of treatment).

Before embarking on a study, researchers who would observe and analyze therapeutic process must make a number of decisions about framing and focusing their studies. Framing a psychotherapy process study entails decisions about the location of the phenomenon (where to look), the temporal unit of observation (how long to look), and the observational perspectives to be employed (from which angle to look). Focusing the study involves making decisions about the facets and phases of therapeutic process to assess (which 'variables' to measure).

Process Research Frames

Location

While therapeutic process may be defined abstractly in terms of psychological interactions and transformations, process researchers have differed over where and how they can best be observed. Initially, most researchers viewed therapeutic process as consisting of the externally observable interactions and communications between the patient and therapist during their therapy sessions. Subsequently, many also included post-session reports by patients and therapists about their experiences during the sessions; and, more recently, researchers have come to recognize that therapy also occurs actively between therapy sessions, whether by design in the form of 'homework assignments' (Kazantzis and L'Abate, 2007) or spontaneously in the reminiscences, reflections, and imaginings of patients and therapists about their therapy and one another (e.g., Hartmann et al., 2011).

A further difference in view is whether therapeutic process includes all the events that occur during and between therapy sessions or just those that are specifically directed toward patient improvement. The former view, which is largely descriptive and discovery-oriented, appeals primarily to researchers who assume that all the factors that produce positive or negative patient outcomes are not yet known. The latter

view focuses specifically on incident of in-session impact ('change events') and tends to appeal to investigators who usually hold, and seek to test, some particular theoretical conception of therapy (e.g., Barber et al., 2013; Elliott et al., 2013; Emmelkamp, 2013; Hollon and Beck, 2013). To distinguish clearly between these contrasting but somewhat overlapping views, the former will be called therapy session process and the latter therapeutic change process.

Temporal Units

The concept of process refers most generally to activities and events that proceed or evolve over time – but some activities and events proceed or evolve so rapidly, or so slowly, that they normally cannot be noticed. Examples of the former are the interactions that physicists observe among subatomic particles or, more familiarly, the rapid projection of still photographs that one does not notice separately but experience sequentially as 'moving' pictures. Examples of the latter include the geological raising of mountain ranges by the gradual collision and subduction of tectonic plates, and their imperceptible erosion by millennia of wind and rain. The point is commonplace but basic and relevant. To be observed at all, therapeutic process must be framed within the temporal span of human consciousness, but there is still a range of noticeable rates at which process events occur. The researcher may choose to examine very brief events, like changes in facial expression or vocal intonation; somewhat longer events, like the statement by one person of a coherent line of thought involving several sentences, or dialogue between patient and therapist on a particular topic that includes successive statements by each; still longer events, like the quality of whole therapy sessions; even longer events, like the opening, middle, and closing phases of an extended course of treatment; or the clinical history of successive psychotherapies over a patient's life. The temporal units or frames of reference that researchers select for study inevitably determine the kind of therapeutic events they can observe, and the descriptions of therapeutic process they give.

Observational Perspectives

Any complex event can be viewed from several angles, and each observational perspective has the potential to reveal a different aspect of the event. In psychotherapy, there are two distinct perspectives, that of patients and that of therapists (e.g., Orlinsky and Howard, 1975), who as participant observers experience the events of therapy directly but from separate vantage points. To these, psychotherapy research has added the perspectives of nonparticipant observers by relying on clinical raters who are not part of the events being observed and whose view of therapeutic process is generally mediated by audio or video recordings of therapy sessions or transcriptions of the verbal content of sessions. Participant observers have not only a 'first-hand' and 'inside' but also fleeting view of the events they experience. They have direct awareness of their own thoughts, feelings, and intentions, and immediate impressions of their own and others' behaviors, but these are typically organized around their goals in action. By contrast, nonparticipant observers gain only 'second-hand' knowledge of therapeutic process, and can only infer the covert internal states of the participants. However, they have the advantage of being able to carefully and repeatedly study the recorded events without the situational pressures to respond that participants experience.

Researchers may use, and have used, one, two, or all these observational perspectives in their studies; depending on their interests and the questions they investigate. It should not seem strange to researchers, or an indication of methodological weakness, if observers who view therapeutic phenomena from different perspectives often produce discrepant observations, since those discrepancies may validly reflect the inherent complexity of social phenomena.

Process Research Focus

When undertaking a study of therapeutic process, researchers understand that they are dealing with a set of events that is too complex to be observed in its entirety. Consequently, they must simplify the focus of the study on a few specific facets of therapeutic process, which are the 'variables' to be studied. For the simplification to be meaningful, these variables must be defined in terms of a theoretical model that will be understood and accepted by other researchers.

Research Variables

Traditionally, the variables examined in psychotherapy process studies have been differentially defined according to whether they focus on the task-instrumental or technique aspect of therapist and patient behavior (referred to generically as 'therapeutic operations'), or focus instead mainly on the social–emotional or relational aspect of therapist and patient behavior (generically encompassing both the 'therapeutic bond' and consensus with respect to 'therapeutic contract').

Task-instrumental behaviors include the interventions that clinical theories suggest are helpful to patients: for example, interpretation of the patient's free associations, dreams, or resistances and transference in psychodynamic therapies; empathic reflection of feeling in person-centered therapy; or relaxation training to counteract phobic anxiety in cognitive-behavioral therapies. Research variables often focus on the content, timing, skillfulness, and appropriateness with which therapists use these techniques. Other variables may focus on the patient's cooperative or resistant response to therapist interventions and on the specific effects of those interventions ('in-session impacts'). Research procedures focused on instrumental aspects of therapeutic process include those described in Greenberg and Pinsof (1986: chapters 2–8, 14).

Researchers have also very extensively studied varied aspects and effects of the relationship between patient and therapist (Horvath et al., 2011; Norcross, 2011). Relationship variables generally focus on the interpersonal manner (as distinct from the task-instrumental content) of therapist's and patient's interactions during sessions; their attitudes and feelings toward one another; and their respective contributions to the formation and development of a therapeutic alliance. In this area too, researchers espousing different theoretical frameworks have focused on the specific relationship qualities suggested by their

respective orientations. Thus, researchers influenced by psychoanalytic theory have focused on issues related to the concept of transference (Levy and Scala, 2012), such as the 'core-conflictual themes' expressed in patients' personal relationships (Luborsky and Crits-Christoph, 1997), and the therapist's countertransference (Gelso and Hayes, 2007). Others, influenced by client-centered theory, have focused on therapist-offered relationship conditions such as genuineness, empathy, and positive regard (e.g., Elliott et al., 2011). Social psychological theories of interpersonal behavior have provided another conceptual approach to research on relationships, emphasizing the reciprocal influences between patient and therapist, focusing on dimensions of behavior such as acceptance vs rejection and control vs autonomy (e.g., Brehm and Smith, 1986). However, given the advent of technological developments such as computer-administered therapeutic programs and text message interventions, which do not involve direct face-to-face interaction with a clinician, the concept of therapy as an essentially interpersonal process may need to be extended. Researchers have also given special attention to the importance of repairing 'alliance ruptures,' which may occur when therapists misunderstand or lack empathy for what is troubling their patients (Safran et al., 2011). Useful procedures for studying therapeutic relationships are described by their proponents in Greenberg and Pinsof (1986: chapters 9–12, 15).

Research Methods

The events of psychotherapy occur in a series of meetings between patient and therapist called therapy sessions. The most directly observable aspects of those events consist of conversation. In some types of therapy, conversation may be supplemented by having patients engage in activities or exercises (e.g., role-play) designed to provide significant learning or corrective emotional experiences. Given the prominence of verbal and nonverbal communications, researchers have relied on audio- and videotapes of therapy sessions to record the events of therapy. Recording and replay technology have not only provided investigators with voluminous records for detailed examination from an 'external' or nonparticipant observational perspective, but also constrained them to concentrate on temporal units that have generally been briefer than whole sessions, and sometimes no longer than one complete statement or interchange between a patient and therapist. But, conversation is only the outward aspect of therapeutic events. The inner side of process is comprises patients' and therapists' experiences of themselves and one another during their sessions, and during the imagined conversations or fantasy interactions that they may have in the times between sessions. This aspect of therapeutic process has been studied mainly through the use of post-session questionnaires. These retrospective ratings utilize patients and therapists as participant observers of therapy to provide research data, and generally take the whole session as the temporal unit of observation. Greenberg and Pinsof (1986) present methods that rely on nonparticipant observation (chapters 2–8) as well as methods that draw upon participant observation (chapters 12–16), and a general overview of methodological issues is provided by Hill and Lambert (2004).

Main Process–Outcome Findings

Established Findings

A half century of accumulated research on psychotherapy has convincingly demonstrated that qualities of the therapeutic relationship, bond, or alliance robustly predict treatment outcomes – whether the relationship or alliance is observed from varied observational perspectives, and whether outcome is assessed by patients, therapists, clinical judges, and psychometric measures. Moreover, the issue of whether these relationship qualities produce outcome or merely reflect its ongoing attainment has apparently been resolved in favor of the former (Crits-Christoph et al., 2013). Global ratings of relationship quality are demonstrably linked to outcome from all observational perspectives. More specifically, the patient's personal investment, active collaboration, expressiveness, emotional responsiveness, and positive feelings toward the therapist seem particularly important (Bohart and Wade, 2013), as is the therapist's empathy and credibility (Norcross, 2011). This has led to a widespread view of the therapeutic bond or alliance as the most salient aspect of therapeutic process, and the 'common factor' that links therapists with patients to enable different types of psychotherapy to be more or less equivalent in effectiveness (e.g., Wampold, 2010).

With regard to therapeutic operations, some types of technical interventions (e.g., interpretation, experiential confrontation, and paradoxical intention), when skillfully focused on patients' life problems, have also been shown persuasively to be associated with positive outcomes (Orlinsky et al., 2004). In terms of participant self-relatedness, assessments of patient openness (vs defensiveness) have been very consistently associated with outcome. Likewise, patients' positive in-session impacts have been consistently found to predict positive treatment outcomes. However, in their recent overview of this area, Crits-Christoph et al. (2013) call for added methodological rigor, increasing controls of potentially confounding variables in order to establish which findings may reflect causal influences on outcome.

New Research Areas

Several new areas of study have emerged in recent years that expand the bounds of process research. Process studies traditionally focus on the interactions and experiences of patients and therapists within therapy sessions, as distinct from outcome studies that focus on patients' mental health status and emotional well-being in their lives outside of therapy. One new area of research explores the experiences that patients and therapists have regarding each other and their therapy outside of therapy in the time between sessions. Study of these inter-session experiences (IEs) extends the concept of therapeutic process into the traditional realm of outcome insofar as they shed light on how patients use what they learn in sessions to improve their outcome status. IEs are conscious, reportable events that presumably reflect a psychological process of internalization, which occurs largely outside the participants' awareness between sessions and enables episodic, relatively infrequent therapy sessions (typically 1 h per week) to become a continuous, cumulative experience for the participants (Orlinsky and Geller, 1993). Measurable indicators of that

psychological process are provided by the thoughts, feelings, memories, dreams, and fantasies that patients and therapists regularly report having between sessions about their therapy and each other. Patient IE reports have been related to the amount of personality disturbance, the therapeutic alliance and type of therapy, treatment outcome, and the internalization of therapists and therapy long after the termination of therapy. Therapist IE reports show how practitioners cope with difficulties encountered in their work with patients, and prepare for upcoming sessions. The value of research on IEs is twofold: showing how the psychological attributes and experiences of patients and therapists, in their lives outside of therapy, influence their actions and experiences during therapy (e.g., in the form 'transference' and 'countertransference' experiences); and showing how the events and experiences in therapy sessions in turn influence the subsequent psychological attributes and experiences of patients and therapists (e.g., how patients apply the new insights and skills acquired in sessions to problematic 'real-life' situations outside therapy).

A reciprocal extension of outcome evaluation into the traditional domain of process has occurred through research on the effects of feedback to therapists about patients' psychological progress or deterioration, and the utilization of this information by therapists to maintain or modify their practice (Lambert, 2010). Programs that monitor the patient's psychological state on a regular basis during the course of therapy were developed through advances in statistical analysis, which made it possible to project a patient's expected treatment response, combined with the use of computer technology to make real-time reports to therapists about their patients' progress or deterioration. This has proven especially helpful with patients who fail to meet expected benchmarks for improvement, or seem to be getting worse despite treatment, particularly if the therapists are provided with and use prompts about how to enhance the process (e.g., working to repair an alliance rupture).

Finally, a third new area of study uses recent technological developments to explore neurophysiologic concomitants and foundations of psychological change during the process of therapy (e.g., Carrig et al., 2009). These newly observable aspects of therapeutic process have shown early promise (e.g., Buchheim et al., 2012) and may gain more prominence in the future. They have already inspired several theoretical accounts of how behavioral and psychological changes may be understood in terms of neurobiological processes (e.g., Schore, 2003), but use of functional neuroimaging and electrophysiology is still largely uncharted territory and its value for studying therapy process remains to be demonstrated.

Process Research and Practice

Psychotherapy process research thus far has proven itself a scientifically successful field of inquiry. Initially, it had to overcome the skepticism and even hostility of some clinicians to having their work studied, as typified in a oft-quoted statement by Freud (1926/1959) that "the 'analytic situation' allows the presence of no third party …" and that "An unauthorized listener who hit upon a chance one of them would as a rule form no useful impression; he would be in danger of not understanding what was passing between the analyst and the patient …" (p. 185). Thus, in the earliest days, investigators were mainly concerned to demonstrate the feasibility of systematic and methodologically objective research on therapeutic process. Once that had been established (e.g., Rogers and Dymond, 1954), another 15 years or so were devoted primarily to devising reliable measures and rigorous methods of process research, during which time some therapists who had taken an interest in the field were disappointed by the methodological preoccupations of researchers and the paucity of clinically applicable results. However, the rate of research on process and process–outcome patterns increased steadily from about 1970 onward and, as the 1990s approached, a considerable body of well-replicated findings had been amassed, from which implications for therapeutic practice could begin to be drawn (e.g., Talley et al., 1994). Current pressures for accountability and quality assurance have forced therapists of all schools to become interested in outcome research, in order to demonstrate the effectiveness of their treatments (e.g., Goodheart et al., 2006). As a result of this pressure, clinical awareness of the relevance and applicability of process research has increased through a number of research-based contributions addressed to practitioners (e.g., Castonguay and Beutler, 2005; Duncan et al., 2010; Greenberg et al., 1993; Hubble et al., 1999; Lebow, 2006; Muran and Barber, 2010; Summers and Barber, 2010). Engaging practitioners has also been done through the development of 'practice research networks' (Parry et al., 2010), in which investigators and therapists collaborate to define research questions relevant to their clinical work, and collect and analyze data both for academic studies and for use to enhance practice. Further development of scientifically meaningful and clinically valuable process research may be anticipated through innovations like these.

See also: Counseling and Psychotherapy: Ethnic and Cultural Differences; Evidence-Based Psychotherapy; Manual Based Treatments; Mental Health and Social Work; Mental Health; Psychiatric Social Work in the United Kingdom; Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychotherapy Integration; Psychotherapy, History of: Psychiatric Aspects; Therapist–Patient Relationship.

Bibliography

Barber, J.P., Muran, J.C., McCarthy, K.S., Keefe, J.R., 2013. Research on dynamic therapies. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ.

Bohart, A.C., Wade, A.G., 2013. The client in psychotherapy. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 219–257.

Brehm, S.S., Smith, T.W., 1986. Social psychological approaches to psychotherapy and behavior change. In: Garfield, S.L., Bergin, A.E. (Eds.), Handbook of Psychotherapy and Behavior Change, third ed. Wiley, New York, pp. 69–116.

Buchheim, A., Viviani, R., Kessler, H., Kächele, H., Cierpka, M., Roth, G., Carol, G., Kernberg, O.F., Bruns, G., Taubner, S., 2012. Changes in prefrontal-limbic function in major depression after 15 months of long-term psychotherapy. PLoS One 7 (3), e33745.

Carrig, M.M., Kolden, G.G., Strauman, J.T., 2009. Using functional magnetic resonance imaging in psychotherapy research: a brief introduction to concepts, methods, and task selection. Psychotherapy Research 19, 409–417.

Castonguay, L.G., Beutler, L.E., 2005. Principles of Therapeutic Change that Work. Oxford University Press, New York.

Castonguay, L., Muran, J.C., Angus, L., Hayes, J.A., Ladany, N., Anderson, T. (Eds.), 2010. Bringing Psychotherapy Research to Life: Understanding Change through the Work of Leading Clinical Researchers. American Psychological Association, Washington, DC.

Crits-Christoph, P., Connolly Gibbons, M.B., Mukherjee, D., 2013. Psychotherapy process-outcome research. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 298–340.

Duncan, B.L., Miller, S.D., Wampold, B.E., Hubble, M.A. (Eds.), 2010. The Heart and Soul of Change: Delivering What Works in Therapy, second ed. American Psychological Association, Washington, DC.

Elliott, R., Bohart, A.C., Watson, J., Greenberg, L.S., 2011. Empathy. In: Norcross, J.C. (Ed.), Psychotherapy Relationships that Work: Evidence Based Responsiveness, second ed. Oxford University Press, New York, pp. 132–152.

Elliott, R., Greenberg, L.S., Watson, J., Timulak, L., Freire, E., 2013. Research on humanistic-experiential psychotherapies. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 495–538.

Emmelkamp, P.M.G., 2013. Behavior therapy with adults. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 343–392.

Freud, S., 1926/1959. The question of lay analysis. In: Strachey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. XX. Hogarth Press and the Institute of Psychoanalysis, London.

Gelso, C.J., Hayes, J.A., 2007. Countertransference and the Therapist's Inner Experience: Perils and Possibilities. Earlbaum, Mahwah, NJ.

Goldfried, M.R., Glass, C.R., Arnkoff, D.B., 2011. Integrative approaches to psychotherapy. In: Norcross, J.C., Vandenbos, G.R., Freedheim, D.K. (Eds.), History of Psychotherapy, second ed. American Psychological Association, Washington, DC, pp. 269–296.

Goodheart, C.D., Kazdin, A.E., Sternberg, R.J., 2006. Evidence-Based Psychotherapy: Where Practice and Research Meet. American Psychological Association, Washington, DC.

Greenberg, L.S., Pinsof, W.M. (Eds.), 1986. The Psychotherapeutic Process: A Research Handbook. Guilford, New York.

Greenberg, L.S., Rice, L.N., Elliott, R., 1993. Facilitating Emotional Change: The Moment-by-Moment Process. Guilford, New York.

Hartmann, A., Orlinsky, D., Zeeck, A., 2011. The structure of intersession experience in psychotherapy and its relation to the therapeutic alliance. Journal of Clinical Psychology 67 (10), 1–20.

Hill, C.E., Lambert, M.J., 2004. Methodological issues in studying psychotherapy processes and outcomes. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, fifth ed. Wiley, New York, pp. 84–135.

Hollon, S.D., Beck, A.T., 2013. Cognitive and cognitive-behavioral therapies. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. Wiley, Hoboken, NJ, pp. 393–442.

Horvath, A.O., Del Re, A.C., Flückiger, C., Symonds, D., 2011. Alliance in individual psychotherapy. Psychotherapy 48 (1), 9–16.

Hubble, M.A., Duncan, B.L., Miller, S.D. (Eds.), 1999. The Heart and Soul of Change: What Works in Therapy. American Psychological Association, Washington, DC.

Kazantzis, N., L'Abate, L. (Eds.), 2007. Handbook of Homework Assignments in Psychotherapy: Research, Practice and Prevention. Springer, New York.

Kirmayer, L., 2007. Psychotherapy and the cultural concept of the person. Transcultural Psychiatry 44 (2), 232–257.

Lambert, M.J., 2010. Yes, it is time for clinicians to routinely monitor treatment outcome. In: Duncan, B.L., Miller, S.D., Wampold, B.E., Hubble, M.A. (Eds.), The Heart and Soul of Change: Delivering What Works in Therapy, second ed. American Psychological Association, Washington, DC, pp. 239–266.

Lebow, J., 2006. Research for the Psychotherapist: From Science to Practice. Routledge, New York.

Levy, K.N., Scala, J.W., 2012. Transference, transference interpretations, and transference-focused psychotherapies. Psychotherapy 49, 391–403.

Luborsky, L., Crits-Christoph, P., 1997. Understanding Transference: The Core Conflictual Relationship Theme Method, second ed. American Psychological Association, Washington, DC.

Muran, J.C., Barber, J.P. (Eds.), 2010. The Therapeutic Alliance: An Evidence-Based Guide to Practice. Guilford, New York.

Norcross, J.C. (Ed.), 2011. Psychotherapy Relationships that Work: Evidence Based Responsiveness, second ed. Oxford University Press, New York.

Norcross, J.C., Vandenbos, G.R., Freedheim, D.K., 2011. History of Psychotherapy, second ed. American Psychological Association, Washington, DC.

Orlinsky, D.E., Geller, J.D., 1993. Patients' representations of their therapists and therapy: New measures. In: Miller, N.E., Luborsky, L., Barber, J.P., Docherty, J.P. (Eds.), Psychodynamic Treatment Research. Basic Books, New York, pp. 423–466.

Orlinsky, D.E., Grawe, K., Parks, B.K., 1994. Process and outcome in psychotherapy – noch einmal. In: Bergin, A.E., Garfield, S.L. (Eds.), Handbook of Psychotherapy and Behavior Change, fourth ed. Wiley, New York.

Orlinsky, D.E., Howard, K.I., 1975. Varieties of Psychotherapeutic Experience: Multivariate Analyses of Patients' and Therapists' Reports. Teachers College Press, New York.

Orlinsky, D.E., Howard, K.I., 1986. Process and outcome in psychotherapy. In: Garfield, S.L., Bergin, A.E. (Eds.), Handbook of Psychotherapy and Behavior Change, third ed. Wiley, New York.

Orlinsky, D.E., Howard, K.I., 1995. Unity and diversity among psychotherapies: a comparative perspective. In: Bongar, B., Beutler, L.E. (Eds.), Comprehensive Textbook of Psychotherapy: Theory and Practice. Oxford University Press, New York, pp. 3–23.

Orlinsky, D.E., Rønnestad, M.H., Willutzki, U., 2004. Fifty years of psychotherapy process-outcome research: continuity and change. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, fifth ed. Wiley, New York.

Orlinsky, D.E., Russell, R.L., 1994. Tradition and change in psychotherapy research: notes on the fourth generation. In: Russell, R.L. (Ed.), Reassessing Psychotherapy Research. Guilford, New York.

Parry, G., Castonguay, L.G., Borkovic, T.D., Wolfe, A.W., 2010. Practice research networks and psychological services research in the UK and USA. In: Barkham, M., Hardy, G.E., Mellor-Clark, J. (Eds.), Developing and Delivering Practice-Based Evidence: A Guide for the Psychological Therapies. Wiley, Hoboken, NJ.

Rogers, C.R., Dymond, R.F. (Eds.), 1954. Psychotherapy and Personality Change. University of Chicago Press, Chicago, IL.

Safran, J.D., Muran, J.C., Eubanks-Carter, C., 2011. Repairing alliance ruptures. Psychotherapy 48 (1), 80–87.

Schore, A.N., 2003. Affect Regulation and the Repair of the Self. Norton, New York.

Strupp, H.H., Howard, K.I., 1992. A brief history of psychotherapy research. In: Freedheim, D. (Ed.), History of Psychotherapy: A Century of Change. American Psychological Association, Washington, DC, pp. 309–334.

Summers, R.F., Barber, J.P., 2010. Psychodynamic Therapy: A Guide to Evidence-Based Practice. Guilford, New York.

Talley, P.F., Strupp, H.H., Butler, S.F. (Eds.), 1994. Psychotherapy Research and Practice: Bridging the Gap. Basic Books, New York.

Wampold, B.E., 2010. The research evidence for common factors models: a historically situated perspective. In: Duncan, B.L., Miller, S.D., Wampold, B.E., Hubble, M.A. (Eds.), The Heart and Soul of Change: Delivering What Works in Therapy, second ed. American Psychological Association, Washington, DC, pp. 49–82.

Zane, N., Nagayama Hall, G.C., Sue, S., Young, K., Nunez, J., 2004. Research on psychotherapy with culturally diverse populations. In: Lambert, M.J. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, fourth ed. Wiley, New York, pp. 767–804.

Public Administration: Organizational Aspects

Judy Johnston, Centre for Management and Organization Studies, University of Technology, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers 'Organizational Aspects' of Public Administration from its origins in the United Kingdom and United States of America prior to the 1850s, for the period from the 1970s to the present. It also considers the future of public administration.

Introduction

Public administration is the term traditionally used to define the formal arrangements under which public organizations serve a government, ostensibly in the public interest. The development of the public administration model dating from the mid-1800s to the early 1900s was influenced primarily by Weber's theory of bureaucracy, Northcote and Trevelyan's recommendations relating to the establishment of a professional civil service in Britain, and Woodrow Wilson's ideas in the United States for the separation of policy from administration (Hughes, 1994). Other managerial theories and concepts over the years have been relevant including Taylor's scientific management and Simon's rational decision-making. Since the 1980s, the traditional model of public administration has been largely rejected by governments in favor of a more focused managerialist model based on private-sector practice, within the context of a market-based economic model of public organization. In more recent years, there have been attempts to balance the economic focus with a renewed emphasis on public organization creating public value. This contribution will first examine the traditional model of public administration including the conceptual and theoretical bases and how this affected organizational aspects. Second, it will explore how and why there has been a paradigm shift from public administration to public management. Third, it will consider future trends.

Public administration is the term used to define the formal procedural and organizational arrangements under which public employees serve a government, by implementing and advising on policy, and managing resources. Organizational aspects refer to both the overall structures as well as the relationships that occur within public administrations. This could include: The organizations that make up a civil service, sometimes referred to as the machinery of government; internal organizational arrangements; and/or organizational behavior. Thus, organizational aspects can be studied in a broad sense or within several defined fields. This contribution covers organizational aspects, widely interpreted.

Beyond public administration as a discrete body of knowledge, organizational aspects can be examined through other theories and practices relating to, for example, political science, public policy, sociology, economics, and management. In this sense, each area of study has its own theories and concepts. Furthermore, each state has its own history, organizational form, and approach, although there are many universal, common elements, which have developed through the international transfer of ideas. New governments, formal review processes, focused research, and events have often stimulated notable change. Therefore, the area of public administration is a difficult area to research, and over the years studies have been largely descriptive rather than empirical.

For the purpose of this article, organizational aspects of public administration are examined as four phases of development and evolution, representing important paradigm shifts or prospective shifts. These are the premodern, modern, postmodern, and virtual/supranational phases. Critical issues relating to the theories, models, trends, and practices of organizational aspects of public administration during these times are explored.

Premodern Public Administration, Organizational Aspects (Pre-1850s)

The organizational aspects of public administration have been apparent from ancient times, through many civilizations, such as Greek, Roman, and Chinese dynasties, and have often been the subject of criticism and analysis. Indeed, the ancient Greek satirist and playwright Aristophanes was probably one of the earliest in the literate world to highlight the lack of accountability and fairness in relation to public office. The works of ancient philosophers, such as Plato and Aristotle, address pertinent issues of interest in this regard.

Military histories over many years from the ancient Greeks and Romans to Machiavelli, to Frederick the Great, and to Napoleon provide further insights into the evolution of organizational aspects of public administration. For example, in the 1700s, in the context of militarism and empire building, the Prussian leader, Frederick the and his successor, Frederic William III, structured an exemplary public administration. Furthermore, France's public administration, which had been evolving for decades through constitutional reform following the French Revolution of 1789, was organized on the basis of similarly functioning districts, or *départements*. In the late-1700s to early 1800s, Napoleon oversaw the central appointment of officials charged with the administrative responsibility for the *départements*. He also developed bodies, which undertook administrative review, and involved the separation of juridical and public administrative powers. The civil service was made up of a number of distinct corps, or bodies, supported by training schools.

In the context of royalty, an early patronage system of public administration developed in England, primarily, as a support function for the royal Anglo-Saxon courts. The later structuring of a distinct, centralized public administration was related to a concept of a parliamentary cabinet. Some senior government ministers had portfolio responsibilities and a small supporting staff. Separate departments eventually existed to provide administrative functions to the ministers in charge of the Treasury, the Exchequer, the Chancellery, and the State. Accountabilities of the ministers were divided, between the Parliament and the monarch. While ministers undertook many of the administrative duties because of a general lack of suitably qualified men [*sic*], a small, elite corps of public administrators was established with a resultant continuity of tenure from government to government. To a small extent, these public administrations served a legitimate purpose, but for the most part they existed to meet the needs and ends of the administrators themselves.

Some of the public administration ideas from Europe were also adopted in New World states. In the United States, in the context of a revolutionary federal constitution, which came after the War of Independence, their early patronage system was based on political connections, rather than royalty or class. Not surprisingly, these arrangements meant that the administrative organizations of government became highly corrupted, as Woodrow Wilson observed (1887). The dissatisfaction with this type of system and a promoted interest in reform processes led to the eventual development of professional public administrations in many states.

Modern Public Administration, Organizational Aspects (*1850s–1970s*)

Organizational aspects of public administration continued to evolve because of constitutional and parliamentary reform processes and, in some cases, shifts from royal courts to representative governments. The industrial revolution, the rise of capitalism, and the new bourgeoisie all contributed to pressures for professional civil services. Thus, in modern times, organizational aspects of public administration were based largely on the need for professionalism and often developed, further, because of administrative review.

In Britain, for example, the Northcote–Trevelyan Report (1853) recommended major changes in civil service recruitment and other processes. Fundamental to the changes was the establishment of a central Board of Public Administration, which eventually became a Civil Service Commission in 1870. The Board was responsible for the competitive recruitment, based on merit, of professional civil servants, and their appointment to departments; it was largely independent of government. From an organizational aspect, the idea of a central and independent agency of government responsible for recruitment and other personnel matters was instituted from that time. The best educated men [*sic*] with knowledge of history, classics, language, and mathematics were appointed to the service. Three classes of civil servant also applied: the elite administrative classes; clerical support classes; and the technical or professional classes.

In the United States, European and British theories and practices continued to influence reform of public administration, especially through the Civil Service (Pendleton) Act of 1883. However, there was a determination, as Woodrow Wilson (1887) notes, to 'Americanize' the system. The state was regarded as an instrument of social cooperation, and a more bureaucratic public administration structure was needed to manage the state. There was, however, always a fundamental rejection that the senior administration could perform in a nonpartisan way. This meant that there were dual streams of career civil servants and political appointees at the high levels of the civil service. Therefore, it was recognized that there was a blurred delineation between the political and public administration environments. High-level entry requirements were embedded into the career civil service, and a political education was favored as the main qualification. Taylor's (1912) theory of scientific management and related developments in the field were promoted as appropriate models of organization and work distribution.

Another major contributor to the evolution of ideas about public administration and organizational aspects was the German economist and sociologist, Max Weber. Weberian principles of bureaucracy have had a universal impact on the organizational aspects of public administration. These principles included the matching of fixed job responsibilities to defined, permanent positions; a hierarchy of official authority; the requirement for appointed, suitably qualified, and anonymous staff with continuity of tenure until retirement; recognition that high-level appointments bring social status; and entitlement for a salary and a retirement pension (Weber, 1946).

As an example of Weber's force, his ideas were evident as far away as Japan. Public administration and organizational aspects were based largely on German bureaucracy and Weberian principles, with other local attributes. This included the preferential acceptance of law graduates into the administrative elite stream, which has meant that the critical decision processes have been rule- and precedent bound. However, the developmental role of the state led to the dominance of certain elite governmental organizations, such as the Ministry of International Trade and Industry.

Furthermore, major events and trends, such as world wars and the growth of global trade, have impacted upon organizational aspects of modern public administration. Consequently, public sectors have increased in certain periods to address major reconstruction initiatives. Cycles of nationalization of certain industries have also impacted upon organizational aspects. Political history is important, too, and organizational change at a microlevel will usually occur when new governments are elected. Thus, there is a strong link between significant history and highly-dynamic organizational aspects of public administration.

Public administration has evolved as a distinct area of research and study in modern times. This is supported organizationally where, in concept, there is meant to be a clear separation of powers and responsibilities between political executives (governments); public administrations (civil or public services); and the juridical arena (the law and courts). There are obvious accountabilities between governments and public organizations, as well, given that there is usually a correlation between the portfolio responsibilities of members of the government and the way public administrative arrangements are determined. For example, members of a government are

allocated portfolio responsibilities pertinent to particular public policy fields, such as health, defense, or education. Typically, there is a defined public administration department dedicated to the implementation of government policy and/or service provision in the corresponding policy field.

Moreover, within public administrations, organization can take a number of forms. While categorization may vary from government to government, in general terms, five types of organization may be apparent. First, there are central agencies dedicated to coordinating and supporting government effort, responsible for areas such as financial and human resource management and management improvement. Second, there are departments charged with the responsibility for direct service provision to the public, across a broad range of policy fields. Third, there are government business enterprises (GBEs), which operate at a distance from government, in a more business-like way, usually providing some essential or commercial service. Fourth, there are the review and regulatory agencies, such as auditors-general, ombudsmen [*sic*], and anticorruption and whistle-blowing protection agencies. Fifth, there are the more peripheral boards and agencies, often semiautonomous entities that conduct other, sometimes more obscure, aspects of government business.

In general, these organizational aspects still apply, but the public administration bureaucratic paradigm, which endured for over 100 years were to a significant extent rejected by governments in postmodern times. By the 1960s, especially in Anglo-American polities, it was apparent that the modern paradigm of public administration, as it had evolved, was no longer serving governments, as intended. Public administrations were seen by critics and reviewers to be too big, too inefficient, and too powerful in comparison to governments. These criticisms were also linked to the perceived failure of Keynesian economic principles, which had underpinned governments' strategic policy decisions since the 1930s.

Postmodern Public Administration, Organizational Aspects (*1979–Present*)

It took until the late-1970s and early 1980s for fundamental changes to be implemented in a major paradigm shift toward a market-based model of public administration. Conservative political leaders Prime Minister Thatcher in Britain and President Reagan in the United States were the first politicians demonstrably to sustain such an approach, which led to early adoption in other Anglo-Saxon polities, and later around the globe. Similar earlier attempted changes had been evident in Latin America, where free-market concepts of the Chicago School were influential for a relatively short time. The market-based model, in varying forms, has been adopted in over 100 countries, worldwide, with some convergence of ideological thought between traditionally conflicting political parties. The market-based model supplanted communism in Eastern Europe in the late-1980s as a guiding ideology and led to revolutionary change in organizational form in other largely totalitarian states, such as the Peoples Republic of China.

Supranational governance structures, involving organizations such as the International Monetary Fund (IMF), the World Bank, the Organization for Economic Cooperation and Development (OECD), the Group of 7–8 (or more) leading trading nations, the World Trade Organization (WTO), and the European Union (EU), have all supported and promoted the market-based model. This may account for the active transfer and adoption processes in place from polity to polity. The EU, as the most comprehensive organizational form for supranational governance, operates with its own parliament and public administration system.

Several common themes based partly on public choice, and other less well-defined neoclassical economic theories of rational choice are characteristic of the market-based model. They all have implications for public administration and organizational aspects. First, for example, there is a move away from a mixed economy to a firm belief in markets as the primary and more efficient economic mechanism. To this end, deregulation relating to trade, finance, and monetary policies has taken place.

Second, if markets are to prevail, then the size of governments and their public sectors needs to be reduced. Included in the radical downsizing of public sectors is commercialization, and/or the privatization of GBEs where it is deemed inappropriate for governments to continue to operate in these areas. Major utilities and commercial businesses traditionally involving governments, in areas of activity such as telecommunications, power generation, distribution and transmission, water, banking, insurance, postal services, and airlines have been targeted.

Third, and most importantly, for the public sector, the strategic focus is on organizational efficiency, the creation of internal market-style competitive conditions, and the more purposive application of private-sector business techniques to public management. These changes have involved a shift from a focus on processes of administration and policy, however flawed, to organizational management. A business, or corporation, metaphor may be used to define organizational form. This change has been given various names in different polities; the most commonly adopted phrase is the New Public Management.

As a subset of public choice theory, principal–agent theory focuses on the relationships between governments and their top bureaucrats. In this interpretation, public bureaucrats are perceived to be budget-maximizing, self-interested agents, not trusted by governments. Therefore, governments need to develop executive structures where principal–agent theory can be applied directly so that governments can compel their administrative elite to act in the government's interests. The establishment of formal senior executive services, with formal performance agreements attached to incentives and fixed-term contracts may result.

Organizationally, the acceptance of the market-based model by governments has seen other major changes. These include, first, substantial machinery of government changes including the separation of policy, regulatory, commercial, and operational services and programs into discrete organizations. Second, the purchaser/provider split, where governments are the contract managers and service provision is contracted out. Third, more openly political appointments at the top of the public service are apparent in previously nonpartisan services. Frequently, this has involved the centralization of power away

from independent personnel bodies, which have been abolished. These senior executive managers are expected to support, publicly, governments' electability. In such circumstances, the possibility of ministers accepting responsibility for departmental managerial failure, within the construct of ministerial responsibility, considerably diminishes.

The rational choice logic, which underpins the market-based model, is open to question, given the presupposition of perfect markets and exchange processes. Such a conceptual framework is unrealistic for the political environment of public administration. Rewards for certain actions and behaviors of agents, required by the principals, can be achieved through both formal and informal power structures. This is an important organizational aspect of the market-based model, where parallel systems of exchange, reward, and relationships operate as a matter of course. While the formal systems may promote a constancy of action and process, the informal system can support implicitly institutionalized power–political relationships.

Whether the paradigm shift has resulted in better public administration and organizational aspects is still to be determined as the key economic indicators of performance, predicated within this new paradigm, suggest a mixed result. There has certainly been a shift from idiosyncratic or confined interpretations of public administration, relevant to particular nations, regions, or systems, to a more universal construct. Nevertheless, there are many variations on the market-based theme depending on governments in power and individual leaderships at any given time.

From an organizational aspect, one of the weaknesses of the market-based model is that it focuses strongly on economic issues and individualism, rather than on the collective responsibility of the state to its citizens. In this sense, social and related policy fields may be diminished in their importance. Some governments have attempted to redress this imbalance with more attention to social, environmental, and other imperatives, relating to the development of human and social capital. United Kingdom Prime Minister Tony Blair (1997–2007) coined the concept of New Labour, which was inspired by the advice and work of Anthony Giddens. The idea of New Labour was an attempt to balance market considerations with greater social concerns. To this end, in many polities there has been a shift toward an underpinning belief that governments and their public organizations are elected to create public value. Policies should, therefore, create a better balance between economic and social outcomes. Additionally, the continuing growth of supranational structures in terms of greater organizational form, and advances in information technology, now provide different challenges for governments.

Virtual and Supranational Public Administration, Public Administration (*Now and into the Future*)

These two developing areas of interest will undoubtedly continue to have major organizational impacts on public administration. With continuing extraordinary developments in information technology, governments are adding more virtual organizational forms to public administrations. For example, especially with customer service initiatives, knowledge management becomes increasingly important. Free information lines, supported by computer technology, provide coordinated details of government services. Similarly, virtual locations allow citizens to complete a range of government requirements, such as license renewals and the payment of bills. Many processes that involved members of communities attending government organizations in person now can be achieved more efficiently online. Comprehensive information can be obtained from government Internet Web sites, and civil servants can be contacted, more informally and directly, by e-mail. Many governments are now making extensive use of social media platforms such as Twitter and Facebook to inform and communicate with constituents. E-commerce, relating to procurement, contracting, and the lodging of documents may now also apply. Within governments, more virtual communication also takes place involving typed, aural, and visual media. Undoubtedly, this will continue to impact upon organizational aspects of public administration in the future, with a likely increase in virtual government. Organizational aspects of public administration will continue to expand through Web sites.

Second, the continuing development of supranational governance structures, in the context of globalization, has led to both cogent and conspiratorial debate in recent years about future organizational aspects of public administration. The Parliament of the EU as a structure challenging the individual sovereignty of national member governments is an example of how future organizational aspects may evolve as other trading coalitions consolidate their form. Governments in rapidly developing economies, such as the People's Republic of China, continue to make many changes to the organizational aspects of public administration as the shift toward a market economy continues to evolve. Responses to globalization, especially negative ones, however, prompt support for localism, which emphasizes the role of individual communities. The Global Financial Crisis of 2008 also has shown the vulnerability of the economically globalized world and how governments need to address crises through urgent administrative restructuring at such times.

With all these complex local and global pressures, it is likely that organizational aspects of public administration will continue to experience quite dramatic change over the decades of the twenty-first century. Whether the more traditional structures of agencies, departments, trading enterprises, and review bodies, with the complex relationships between key actors, will remain as the primary organizational aspects is difficult to predict. Even if they do remain, the more virtual aspects of public administration are expected to endure and grow to the point where there will be far more virtual public administration than there is in the second decade of the twenty-first century.

See also: Administration in Organizations; Bureaucracy and Bureaucratization; Civil Service; Organization: Overview; Organizations and the Law; Organizations, Sociology of; Organizations: Authority and Power; Parliamentary Government; Public Bureaucracies; Public Sector Organizations; Weber, Max (1864–1920).

Bibliography

Abonyi, G., Van Slyke, D.M., December 2010. Governing on the edges: globalization of production and the challenge to public administration in the twenty-first century. Public Administration Review 70 (Suppl. 1), s33–s45 (see this issue for other articles on the future of public administration).

Alford, J., O'Flynn, J., 2009. Making sense of public value: concepts critiques and emergent meanings. International Journal of Public Administration 32 (3), 171–191.

Barratt Brown, M., 1995. Models in Political Economy. Penguin Books, London.

Bozeman, B., 2007. Public Values and Public Interest: Counterbalancing Economic Individualism. Georgetown University Press, Washington D.C.

Clegg, S., 1989. Frameworks of Power. Sage, London.

Galbraith, J.K., 1994. The World Economy since the Wars. Mandarin, London.

Giddens, A., 2002. Where Now for New Labour? Polity Press, Cambridge.

Halligan, J., Power, J., 1991. Political Management in the 1990s. Oxford University Press, Melbourne, Australia.

Hood, A., March 1998. Public officials, government and the public trust. Australian Journal of Public Administration 57 (1), 98–114.

Hughes, O., 1994. Public Management and Administration. St. Martin's Press, New York.

Hutton, W., 1995. The State We're in. Jonathan Cape, London.

Institut Internationale d'Administration Publique, 1996. An Introduction to French Administration. Institut Internationale d'Administration Publique, Paris.

Kingdom, J., 1991. Government and Politics in Britain. Polity Press, Cambridge, UK.

Laffin, M., Painter, M. (Eds.), 1995. Reform and Reversal: Lessons from the Coalition Government in New South Wales 1988–1995. Macmillan, South Melbourne, Australia.

Lynn Jr., L.E., 1996. Public Management as Art, Science and Profession. Chatham House, Chatham, NJ.

Northcote–Trevelyan Report, P, 1853. Report on the Organisation of the Permanent Civil Service. In: Appendix B in Fulton, Lord (Committee under the Chairmanship of) 1968 the Civil Service, vol. 1. HMSO, London.

Peters, B.G., 1996. The Future of Governing: Four Emerging Models. University Press of Kansas, St Lawrence.

Pollitt, C., 1993. Managerialism and the Public Services, second ed. Blackwell, Oxford, UK.

Raadschelders, J., Lee, K.-H., January/February 2011. Trends in the study of public administration: empirical and qualitative observations from public administration review, 2000–2009. Public Administration Review 71 (1), 19–33.

Rainey, H., 1991. Understanding Public Organizations. Jossey-Bass, San Francisco, CA.

Savoie, D., 1994. Thatcher, Reagan, Mulroney. University of Toronto Press, Toronto, Canada.

Self, P., 1993. Government by the Market? Macmillan, London.

Shafritz, J., Hyde, A., 1992. Classics of Public Administration. Brooks/Cole, Pacific Grove, CA.

Spann, R.N., 1979. Government Administration in Australia. Allen & Unwin, Sydney, Australia.

Sturgess, G., September 1993. Can parliament cope? Towards a new era in public accountability. Canberra Bulletin of Public Administration 7, 45–52.

Taylor, F., 1912. Scientific Management (Reprinted in Shafritz, J., Hyde, A., 1992. Classics of Public Administration). Brooks/Cole, Pacific Grove, CA.

Wanna, J., O'Faircheallaigh, C., Weller, P., 1992. Public Sector Management in Australia. Macmillan, South Melbourne, Australia.

Weber, M., 1946. Bureaucracy. Reprinted in Shafritz, J., Hyde, A., 1992. Classics of Public Administration. Brooks/Cole, Pacific Grove, CA.

Wilenski, P., 1986. Public Power and Public Administration. Hale and Ironmonger, Sydney, Australia.

Wilson, W., 1887. The Study of Administration. Reprinted in Shafritz, J., Hyde, A., 1992., Classics of Public Administration. Brooks/Cole, Pacific Grove, CA.

Public Broadcasting

Michael Tracey, University of Colorado at Boulder, Boulder, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For much of the twentieth century, the dominant form of broadcasting globally was public service broadcasting. The essential premise of this form of broadcasting was that its role was to serve the public interest. To achieve this end, it was deemed crucial that the broadcaster should be free from commercial and political pressure, should seek to inform and educate the audience always with an eye to the quality of the programming on offer. This article offers a description of the principles of public service broadcasting, particularly as represented by the most famous and influential broadcasters, such as the BBC (British Broadcasting Corporation) .

The history of broadcasting, beginning in the early 1920s, has seen three main types or models of broadcasting organizations. One is the commercial model, in which advertising is used to provide revenues, with programs seen as vehicles for providing an audience to which products can be sold. A second model is one in which the broadcasting organization is funded by the state, usually so that the state can project its political message and thus influence the public. The third model is known in the United States as public broadcasting but in most of the rest of the world as public service broadcasting. The most famous and influential example of this model is the BBC. It is this model with which this article is concerned, particularly by defining it in ways that will allow the reader to see the fundamental difference between this and the other two models. The article examines the principles that were deemed to define the purpose and editorial policies of such organizations as the BBC. The article argues that the fundamental case that was, and is, made for public broadcasting is that absence of overt state and commercial pressure is the only way in which one can guarantee programming of quality and diversity, objectivity, and balance.

The General Concept and Background of Public Broadcasting

In a public system, television and radio producers acquire money to make programs. In a commercial system they make programs to acquire money. However simple, this little epithet articulates the divergence of basic principles, the different philosophical assumptions, on which broadcasting is built. History and experience fashioned inside public broadcasting a definable canon, a set of principles and practices which constitute the core theses around which the institution has been formed and which have guided its performance.

The particular institutional structures and forms of funding may vary from place to place, but public broadcasting is above all else an ambition, a belief that the sheer presence of broadcasting within all our lives can and must be used to nurture society, to provide the opportunity that society and its inhabitants can be better served than by systems which primarily seek consumers for advertisers or subjects for the state.

Some of the most powerful visions of the purpose of broadcasting emerged within unusual and trying circumstances. Consider, for instance, the cultural histories of the occupations of Germany and Japan in the late 1940s and the formulation of Allied policy for broadcasting in the rebuilding of those societies. Here was testament to the idea of broadcasting as primarily a social rather than an economic process, as something with moral, cultural, intellectual, and creative purposes and not just as a source of mild comment and moderate pleasure. The charters that established broadcasting in Japan and in Germany, by the Allied occupying powers, were replete with the public service ideal. If broadcasting was to comment, it should do so with a flourish. If it was to amuse, it should do so with elan. If it was to educate, it should do so with real professionalism. It was believed by the Allied leadership, including that of the United States, that the life of the mind of a society was far too precious and important to be left to what they took to be the vagaries of a commercial system or the oppressive force of state control.

It could be argued that such policies were creatures of the moment, as massive destruction demanded enormous reconstruction, of which communications would inevitably be part. But what was required was the restoration not just of highways, buildings, plants, but also of the shattered imaginative lives of whole populations. The architects of postwar Germany and Japan sensed correctly that healthy, diverse cultural institutions were a prerequisite to a functioning liberal democracy. Broadcasting was thus to be used as a key part of the cultural and social regeneration of those societies. In that lies the real clue to the nature and purpose of great public broadcasting: that it makes best sense when it represents a national and moral optimism within a society, when it suggests through the diversity and quality of its programs that societies be better than they are, better served, better amused, better informed, and, thus, better societies.

A period which is widely regarded within the advanced industrial societies as a high water mark of public service broadcasting is that of the BBC in the early 1960s. A key figure from those years was Sir Arthur, possibly the greatest of the chairmen of the BBC. In a pamphlet that he privately published in 1963 titled "What Is Broadcasting About?" he quoted from the report of the Pilkington Committee which had been set up by the British government to look into the future of broadcasting and published its report in 1962: "By its nature broadcasting must be in a constant and sensitive relationship with the moral condition of society." Sir Arthur felt that the moral establishment had failed modern society and that

broadcasting was a way in which that failure could be rectified. He added, quoting again from the Report, that it "is of cardinal importance that everyone in a position of responsibility should be ready to set himself or herself the duty of assuring, to those creative members of staff … that measure of freedom, independence and élan without which the arts do not flourish" (quoted in Tracey, 1983, p. 235).

That idea of providing a protective layer within which the imaginative spirit might create, lay at the heart of the BBC version of public service broadcasting which flourished in the postwar years and which constituted a model that other nations sought to emulate. Ian Jacob, Director-General of the BBC from 1952 to 1959, refined the notion. In 1958, in an internal document called 'Basic Propositions,' he described public service broadcasting as:

> … a compound of a system of control, an attitude of mind, and an aim, which if successfully achieved results in a service which cannot be given by any other means. The system of control is full independence, or the maximum degree of independence that Parliament will accord. The attitude of mind is an intelligent one capable of attracting to the service the highest quality of character and intellect. The aim is to give the best and the most comprehensive service of broadcasting to the public that is possible. The motive that underlies the whole operation is a vital factor; it must not be vitiated by political or commercial consideration.

This is one of the best attempts to capture in words a concept and view of broadcasting which continues to circulate in the world of cultural politics. Yet, even here the vision, the articulation, is limited. Jacob's words imply that we understand the nature of public service broadcasting not by defining it, but by recognizing its results, rather as one plots the presence of a hidden planet or a subatomic particle not by 'seeing' it, but by measuring the effects of its presence. The Pilkington Committee, on the future of British broadcasting, said as much in 1962 when it observed: "though its standards exist and are recognizable, broadcasting is more nearly an art than an exact science. It deals in tastes and values and is not precisely definable" (Home Office, 1962, p .13, para. 34).

One obvious characteristic of efforts to define public service broadcasting is the elegant vagueness that is studied and deliberate. In effect what is suggested is that one thing that has to be understood about defining public service broadcasting is that in any abstract sense one cannot; that whatever the definitional uncertainties, public service broadcasting can be experienced and recognized but never properly be captured by language. Inevitably, however, someone has to decide on what is 'good' and 'bad,' and that there is a guiding hand in making decisions as to what will and will not be produced, what has been referred to as 'the custodians, the caretakers' of culture. As traditionally understood by public broadcasters the whole idea of custodianship was totally unproblematic, as was understanding just who the custodians were, the characteristics they would possess, and the locations in which they would be found. Such presuppositions appeared only natural throughout the history of broadcasting because that history was embedded within a social order in which hierarchy was also assumed: hierarchies of social status and cultural judgment. And in a curious kind of way the point of hierarchy is to reproduce itself since the fundamental belief of the hierarchical system is, and has to be, that such arrangements have worth and merit. The plausibility of such a thesis becomes very much dependent on delivering evidence that attests to both worthiness and merit in such a way as to drown out the noise of any emergent countervailing thesis, most notably that this is an odiously elitist viewpoint.

A key justification for the custodial role in most societies where public service broadcasting was established was that since the radio frequency used for transmissions was a limited natural resource, someone had to ensure that its use served the public good, the whole community. The cultural geology of this decision had, however, a deeper level to it, based on nineteenth century assumptions about the ways in which the arts and humanities could elevate the human condition. In an essay on the formation of modern literary analysis, Steiner wrote that behind the study of literature in the nineteenth century lay:

> a kind of rational and moral optimism … a large hope, a great positivism … The study of literature was assumed to carry an almost necessary implication of moral force. It was thought self-evident that the teaching and reading of the great poets and prose writers would enrich not only taste or style but moral feeling: that it would cultivate human judgment and act against barbarism.
>
> Steiner, 1985, pp. 77–78

Reading this account one could quite properly substitute the word 'broadcasting' for 'literature' and have a powerful explanation of what the creation of the model of public service broadcasting was all about: a relocation of a nineteenth century humanistic dream of human betterment. The fear that drove that dream was of 'the mob,' the pervasive belief among cultural, religious, and political elites that there was indeed a dark side to the human soul that was, when let loose, dangerous and devastating to the flesh as well as the spirit. Whatever the apparent elitism who is to say that they were wrong nestling, as they did, between the first great war and a looming second? There remained, however, a residual faith, tied to the whole condition of the Enlightenment, humanism, and belief in progress, that popular culture need not be debauched but could in fact transcend itself. Consider this key passage from the Pilkington Report:

> Television does not, and cannot, merely reflect the moral standards of society. It must affect them, either by changing or by reinforcing them…Because the range of experience is not finite but constantly growing, and because the growing points are usually most significant, it is on these that challenges to existing assumptions and beliefs are made, where the claims to new knowledge and new awareness are stated. If our society is to respond to the challenges and judge the claims, they must be put before it. All broadcasting, and television especially, must be ready and anxious to experiment, to show the new and unusual, to give a hearing to dissent. Here, broadcasting must be most willing to make mistakes; for if it does not, it will make no discoveries.
>
> Home Office, 1962, pp. 19–20, paras. 50–53

The suggestion here is not that public broadcasters are all hoping and dreaming that their programs will transform

people from cultural and intellectual slobs into something of which one can more readily approve, but rather that objectively some such argument must be the public broadcaster's last line of defense. The language is of standards, quality, excellence, and range. The logic is of social enrichment, that in however indefinable a manner this society is in, it is 'better' for having programs produced from within the framework of those social arguments termed public service broadcasting, compared to those programs produced within an environment in which commerce or politics prevail.

One cannot, however, escape the charge that those sentiments rest on a set of ill-explored assumptions about the sociological organization of modern culture. How are we better? What are the mechanics? And, vitally, where is the evidence? Indeed, what would such evidence even look like?

The Public Service Idea: Eight Principles

What then are the definable principles of public service broadcasting? Eight principles were identified by the London based Broadcasting Research Unit (of which the author was the director) and are summarized in the following sections.

Universality of Availability

Public broadcasting has historically sought to ensure that its signals are available to all, that no one should be disenfranchised by distance or by accident of geography. The imperative that guides this principle is not that of maximizing customers in a market but of serving citizens in a democracy, that if one defines one's audience as the citizens of a country, then logically one has to reach them all. To a remarkable extent in country after country this principle has been made real.

Universality of Appeal

Public broadcasting seeks to provide programs that cater to the many different tastes and interests that constitute a society's life. The public broadcasting community understands that each of us, at different moments, is part of a majority and a minority. In seeking to provide programs for a wide range of tastes and interests, public broadcasting does so with an eye cocked to the need to ensure that whether the program is pitched at the many or the few it is done so with real quality. Public broadcasting does not expect that it can please all the people all of the time – indeed it sees in that approach precisely the kind of populism that nurtures cultural mediocrity, as quality is sacrificed on the altar of maximizing the audience size.

It is an important element of this principle that public broadcasting serves not only tastes and interests that are readily apparent, but also those that are dormant and latent – that may be part of the potential we all possess but that circumstance may not have allowed us to develop.

Provision for Minorities, Especially Those Disadvantaged by Physical or Social Circumstances

It is commonplace to characterize the medium of television as essentially serving 'the mass.' Certainly, public broadcasting understands the vast capability of one medium to reach enormous numbers of people. It sets its face, however, against the logic of commercial systems to see people as no more than statistics in skins, with a definable value captured in the most desirable rates, demographic buys, and cost per thousand. As suggested in the second principle, public broadcasting views the public as a rich tapestry of tastes and interests, each of which, insofar as possible, should be served.

In particular, it recognizes that there are whole subcultures of minority social experiences crying out for attention. People of different color, language groups, and religious preferences all have vital needs for expression in the political and social discourse of the nation. Public broadcasting is dedicated to a dual role here – on the one hand to give access to such groups, to provide them with the opportunities to speak to one another, and to voice the issues as they see them, and on the other to provide coverage of their histories, interests, and concerns for public at large.

Serving the Public Sphere

Some television programs are successful because they get a fair-sized audience, make some money, and sometimes even exemplify the craft of popular television. Other programs are successful because they reach out and touch a small, particular, but powerful audience. Some programs are successful because the craft of the program maker is used to speak to us all. They touch us, move us, make us laugh and cry, and cheer. They speak to us because they speak for us. Like all great art, they help us make sense out of life, to see and understand things with a fresh eye, and give us a burning sense of the collective, of belonging to the nation-as-community. It is an increasingly vital principle of the work of public broadcasting that it recognizes its special relationship to a sense of national identity and broad community. Any nation is a patchwork of localities and regions, but it also is a nation, heterogeneous and homogeneous to a remarkable degree at one and the same time. Public broadcasting's very nature is then to nurture the public sphere as a means of serving the public good. It understands that while within civil society individuals pursue their own private self-interests, it is within the public sphere that they function as citizens. It is a fundamental principle then that public broadcasting must motivate the viewers as citizens possessing duties as well as rights, rather than as individual consumers possessing wallets and credit cards.

A Commitment to the Education of the Public

The most outstanding example of public broadcasting's commitment to the audience-as-citizen is the long-time provision in almost all systems of educational programming at every level. Public broadcasting knows that political and social literacy, as well as of course literal literacy, is an essential prerequisite to the healthy working of a democratic order. Above all else, the commitment to this principle requires that it treats its audience as mature, rational beings capable of learning and growing in many ways. Thus, much of public broadcasting has retained its commitment to institutional services. Daytime school broadcasting and formal learning services of all kinds continue to play a role in most national services.

Independence from Vested Interests

It is a simple but key principle of public broadcasting that its programs can best serve the public with excellence and diversity when they are produced from within a structure of independence. Programs funded by advertising necessarily have their character influenced in some shape or form by the demand to maximize the garnering of consumers. Programs directly funded by the government, and with no intervening structural heat shield, inevitably tend to utter the tones of their master's voice.

The whole history of public broadcasting has been dominated by the idea that it can best serve the nation when it remains distanced from any particular commitment to any particular power structure inside the nation. Of particular importance to this principle is the ability of public broadcasting to support a cadre of independent-minded program makers, who are thus well able to speak with authentic tones and to offer that singularity of vision allied to creativity and passion that has traditionally produced some of public television's finest moments. It follows that the political and economic architecture of this principle is such as to support the making of programs that are good in their own terms, whatever their intended audience. In the making of programs for public broadcasting, there should be no ulterior purpose or motives. It is axiomatic to this principle that the funding of public broadcasting should be such, in total amount and in the absence of any strings attached, as to encourage rather than negate the independence enjoyed.

Encouraging Competition in Good Programming Rather than Competition for Numbers

This principle is central to public service broadcasting and essentially involves a commitment to making programs which, whatever their intended audience, are of high quality. The overwhelming brunt of the evidence leads to the conclusion that the most important aspect of such structuring relates to the forms of finance. Where commercial sources of revenue are dominant, or even present, or where there is direct subvention from government, the program maker's eye is almost inevitably diverted away from what should be the main focus, the inherent quality of the program he or she is making.

The Rules of Broadcasting Should Liberate Rather than Restrict the Program Maker

While all broadcasting will inevitably be governed by certain prescriptions – 'educate,' 'inform,' 'entertain,' 'balance,' and 'fairness' – and certain broadly drawn restrictions – obscenity and national security – the essence of the legislative foundation by which it is empowered should sustain a liberal function for the program maker. The legislation should create secure living space, arena for action, for broadcasters with all kinds of interests in possible programs and possible varieties of audience, rather than leaving the field to those who are interested chiefly in delivering maximum audiences most of the time. The legislation should also ensure that the higher echelons of broadcasting contain individuals who understand its potential and who themselves care for the importance of the creative work of their staff. Part of that understanding would be the need for experiment and innovation in broadcasting, the need to provide a focus for a society's quarrel with itself and the recognition that mistakes will be made but as such may signify the health of the system.

Perhaps above all else, such leadership should be helped to understand that experiment, innovation, quarrel, and mistake are likely to come from the younger program maker, without whom the system is in danger of institutional arteriosclerosis.

Public Service Broadcasting in the Twenty-First Century

In recent years public broadcasting everywhere has faced increasing challenges from a number of quarters. The digital revolution and the rise of the multichannel world of cable and satellite television has posed serious competitive pressure on national public broadcasters. This is because not only is there greater competition for the audience, but also because public broadcasters have felt the need to offer their programming not only in the traditional way, but also to develop new platforms and new programming product to meet the digital challenge. There has, however, in recent years been another, more subtle, but none the less important, change in what is expected of them. In almost every country governments have developed economic policies which encourage the play of market forces in almost every facet of life, and culture and broadcasting are no exception. The increase in competition has already been referenced, but public broadcasters are increasingly being asked to justify their value to the society not in some loosely defined manner, but in what can only be called econometric ways. A key concept which has emerged and, indeed, which has been incorporated into the official policy of such broadcasters as the BBC, is that of 'Public Value.' The term has been defined by the man most closely associated with it, Mark Moore, in this way: "public value is that which the public values." This implies that increasingly the basic model that determines the ways in which public broadcasters will function has moved from the classic public service/culture model, to one that more closely resembles the market model. Some observers have interpreted this as suggesting not just a different role for public broadcasting, but also that there is now a very serious question mark over the whole future of public broadcasting in the very different world of the twenty-first century.

See also: Broadcasting: Regulation; Mass Media, Political Economy of; Mass Media, Representations in; Public Sphere and the Media.

Bibliography

Avery, R.L. (Ed.), 1993. Public Service Broadcasting in a Multichannel Environment: The History and Survival of an Ideal. Longman, New York.

Briggs, A., 1985. The BBC: The First Fifty Years. Oxford University Press, Oxford, UK.

Broadcasting Research Unit, 1985. The Public Service Idea in British Broadcasting – Main Principles. BFI Publishing, London.

Bullert, B.J., 1997. Public Television: Politics and the Battle over Documentary Film. Rutgers University Press, New Brunswick, NJ.

Home Office, 1962. Report of the Committee on Broadcasting (Pilkington Report). HMSO, London.
Jarvik, L.A., 1997. PBS, behind the Screen. Forum, Rocklin, CA.
Ledbettor, J., 1997. Made Possible by: The Death of Public Broadcasting in the United States. Verso, New York.
Lowe, G.F., 2010. The Public in Public Service Media. Nordicom, Sweden.
Skene, W., 1993. Fade to Black: A Requiem for the CBC. Douglas and McIntyre, Vancouver, BC.
Smith, A., Paterson, R. (Eds.), 1998. Television: An International History. Oxford University Press, New York.
Steiner, G., 1985. Language and Silence: Essays 1958 – 1966. Faber, London.
Stewart, D., 1999. The PBS Companion—a History of Public Television. TV Books, New York.
Stuart, C. (Ed.), 1975. The Reith Diaries. Collins, London.
Tracey, M., 1983. A Variety of Lives: A Biography of Sir Hugh Greene. Bodley Head, London.
Tracey, M., 1998. The Decline and Fall of Public Service Broadcasting. Oxford University Press, Oxford, UK.
Tracey, M., 2012. Public Service Broadcasting: A Bibliographical Essay. Oxford Bibliographies On-line, Oxford University Press, New York.

Public Bureaucracies

Francis E Rourke, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12519–12521, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Throughout history public bureaucracies have served as a source of continuity and stability in the life of the State. This role has sometimes led to their becoming a significant obstacle to innovation in government policies. More recently, however, these bureaucracies have also become useful agents of change in many societies. The power that bureaucrats exercise comes mainly from the expertise they command both as individuals and when linked together in organizations. This power has come under increasing challenge in the industrialized democracies. Other institutions – governmental and nongovernmental – have begun to acquire their own expertise, and public bureaucracies are less able to cloak their operations in secrecy. Moreover, many public agencies that formerly performed regulatory or service delivery functions have been abolished, and private or quasi-public organizations have begun to perform tasks that were once within the exclusive preserve of government agencies. By the end of the twentieth century it had become clear that public bureaucracy was not the growth industry it had appeared to be in the aftermath of World War II, 50 years earlier.

Throughout history public bureaucracies have served as a source of continuity and stability in the life of the state. This role has sometimes led to their becoming a significant obstacle to innovation in government policies. More recently, however, these bureaucracies have also become useful agents of change in many societies.

The power that bureaucrats exercise comes mainly from the expertise they command both as individuals and when linked together in organizations. This power has come under increasing challenge in the industrialized democracies. Other institutions – governmental and nongovernmental – have begun to acquire their own expertise, and public bureaucracies are less able to cloak their operations in secrecy. Moreover, many public agencies that formerly performed regulatory or service delivery functions have been abolished, and private or quasi-public organizations have begun to perform tasks that were once within the exclusive preserve of government agencies. By the end of the twentieth century, it had become clear that public bureaucracy was not the growth industry it had appeared to be in the aftermath of World War II, 50 years earlier.

Bureaucracy: The Permanent Government

Although perhaps the least venerable, public bureaucracies rank among the oldest of all existing government institutions. Such bureaucracies played a central and continuing role in the operation of two of the largest and the most effective of the empires that emerged in the early stages of human history, the Chinese and the Egyptian, as they did in the imperial systems of Rome and Byzantium. Each of these was a political system that endured for centuries.

As history evolved and as nation-states emerged as the dominant form of political community in Western society, public bureaucracies were still able to demonstrate as formidable a capacity for holding on to power as their ancient forbears. Even when other political institutions, such as elected executives and legislatures, began to emerge as new centers of government authority, these public bureaucracies still managed to maintain a leading position in crafting and carrying out state programs and decisions.

This continuity in the exercise of power enabled public bureaucracies to be a source of stability in the life of a nation even when fundamental changes were taking place in the nature of the regime they served. Alexis De Tocqueville has given us a memorable description of bureaucratic organizations serving in this way amidst the turbulence of French politics at the end of the eighteenth century: "For though in each successive revolution the administration was, so to speak, decapitated, its body survived intact and active. The same duties were performed by the same civil servants…These men administered the country or rendered justice in the name of the King, then in that of the Republic, thereafter in the Emperor's" (De Tocqueville, 1955).

Bureaucracy and Change

Since De Tocqueville wrote, the role that public bureaucracies can play as a stabilizing force in societies beset by a time of troubles has sometimes been seen in a less favorable light. It is argued, for example, that their durability as government institutions has enabled bureaucracies to function as an obstacle to change in the social order, long after it has become clear that such change is urgently needed. During the twentieth century, some bureaucracies, especially in the military sector, have played this kind of obstructive role in the emerging nations of Africa and Asia. In its most extreme form, resistance to change on the part of a public bureaucracy may even lead a government agency to sabotage programs placed under its control that it opposes.

Bureaucracies as Change Agents

In the modern world, however, it is also possible to find public bureaucracies that serve not as barriers to innovation but as

agents of change in their own societies. When Social Democratic political parties came to power in Europe after World War II, it was widely anticipated that the reforms they intended would be undercut by the conservative bureaucracies in their own country that would be responsible for putting these programs into effect. This kind of sabotage did not in fact occur. Such bureaucracies carried out the changes proposed by the Labour Party when it took control over the British government after the parliamentary elections of 1945 with the same diligence they had shown when administering the programs of the Conservative governments under which they had previously served.

Moreover, one of the enduring legacies of that great period of social and economic reform in the USA in the 1930s, which came to be known as the New Deal, was the creation of new bureaucratic organizations such as the Rural Electrification Administration (REA) and the Tennessee Valley Authority (TVA) that were designed to serve neglected sectors of the American population. In the 1960s a similar reform movement gave birth to agencies at all levels of government in the USA that were authorized to protect women and minority groups from discrimination in employment and other spheres of everyday life. Of course, even when a bureaucracy is created to serve as an agent of change, it can still find itself under attack from its own supporters on the grounds that it is insufficiently zealous in pursuing this objective. As a result, the stereotype of bureaucracy as an obstacle to change tends to endure, even in settings where an organization is breaking new ground.

The effort to use public bureaucracies as instruments of social reform is the product of a general movement toward democracy that has taken place throughout the Western world. As this has occurred, victory in regularly scheduled elections has become the chief source of empowerment, and the political parties competing in such electoral contests have every reason to see public bureaucracies as vehicles for promoting policy changes or carrying on programs that will ultimately provide the political support necessary to win or maintain control of the government. This convergence of interest between party and bureaucracy has energized the effort in modern democratic societies to use bureaucratic institutions as instruments for breaking with the past rather than preserving it.

Bureaucratic Expertise

Throughout history the fundamental source from which public bureaucracies derive their power is the possession of a body of knowledge or skill indispensable to the successful performance of the government in which they operate. In many cases, especially during the early stages in their historical development, these organizations held a monopoly of such expertise. In addition to the influence it gave them within the councils of government, it enabled them to win a certain measure of deference within the community at large. To be sure, this deference was usually coupled with hostility, since ordinary citizens commonly resent the power exercised by bureaucrats in democratic as well as authoritarian societies. Such antagonism toward bureaucracies is perhaps the chief way in which these disparate systems of government continue to resemble one another in the modern world.

Over the course of modern history societies have differed in their willingness to allow expertise to become the signature characteristic of their public bureaucracies. The UK, modeling itself after the example of ancient China, pioneered in this respect, and various segments of the British Empire followed this example even after they won independence. In the more egalitarian culture of the USA, proposals to establish such a Mandarin-style bureaucracy met with strong initial resistance in the nineteenth century as betrayal of the country's democratic traditions. A similar outcry can still be triggered today in US politics by proposals that the country create an institutional counterpart to France's '*Ecole Nationale d'Administration*,' a school for the training of the French administrative elite.

Competition among Experts

One development that has greatly diminished the advantage public bureaucracies derive from their command over bodies of expertise in specialized areas of government policy has been the gradual migration of such expertise to other institutions of government. In the USA, for example, Presidents now have a large number of experts of their own clustered in various units within the Executive Office of the President. While Prime Ministers and Cabinet members in the parliamentary systems that prevail elsewhere in democratic societies do not have similar access to such organized cadres of experts, they are often able to turn elsewhere to satisfy their need for expert assistance.

The Rise of Think-Tanks

Another important source of expertise widely available to elected officials today lies within the private sector in the form of private research organizations known as 'think-tanks' that have emerged and begun to focus on current issues of government policy in the Western democracies. The research findings of these organizations sometimes attract almost as much, if not more, public attention than the reports emanating from the public bureaucracies whose work they monitor. In earlier days these think-tanks tended to provide intellectual support for programs and policies that led to an expansion in the size and power of public bureaucracies, especially in the areas of the social services and economic regulation. More recently, however, think-tanks with a conservative orientation have begun to emerge, and they have provided intellectual ammunition for groups critical of public bureaucracies and the programs and policies these organizations pursue.

Bureaucratic Secrecy and Open Government

The emergence and expansion of new forms of mass communication has tended to make public bureaucracies increasingly vulnerable to scrutiny and criticism from institutions and groups outside the government. Pressure from such media organizations has been chiefly responsible for the enactment in a number of modern societies of 'freedom of information' statutes which require public bureaucracies to provide information in their files to private citizens or groups upon request, except for certain specified categories of data. Max Weber long ago saw the secrecy in which the public bureaucracies of his day

normally operated as a major source of the power they exercised. As a study edited by Itzhak Galnoor has shown, the scope of such secrecy has been greatly narrowed in both Europe and the USA today.

Alternatives to Public Bureaucracies

Critics of the role of public bureaucracies in government, or of the programs they administer, have long argued that they are in fact immortal institutions, which, once established, are impossible to terminate. Certainly it is true that such bureaucracies are tenacious organizations, sinking deep roots into the fabric of the societies they serve.

Still, recent developments in modern societies make it abundantly clear that public bureaucracies are far from immortal. During the past two decades a wide variety of regulatory agencies have been put to death in both the USA and Europe, casualties of a bi-partisan consensus that they mainly serve not the public but the interests of the industries they regulate. Public bureaucracies have also been victims of the growing belief that private ownership is more efficient than state operation. A striking example of this trend occurred in Canada in 1996, when Parliament transferred the vital task of carrying on the country's air traffic control system from the government's transport department to a nonprofit firm created by the legislature.

Along with these developments, there has been a continuing expansion in both Europe and the USA of organizations in the nonprofit sector that administer programs that were once thought to lie within the exclusive domain of government. These nongovernmental agencies, or 'semi-public' bureaucracies as they might be called, supplement rather than supplant the work of public bureaucracies in such areas as health, welfare, and education, and much of their revenue actually comes from the government itself. One such organization in the international sphere – Medecins Sans Frontieres – received the Nobel prize for Peace in 1999 for its work in providing medical care and moral support for population groups in distress around the world.

Editor's Note

Recent works of interest on public bureaucracies include: Piore, 2011; Stazyk and Goerdel, 2011; Denhardt and Catlaw, 2015; Milakovich and Gordon, 2013.

See also: Administration in Organizations; Administrative Law: The United States and Beyond; Bureaucracy and Bureaucratization; Bureaucratization and Bureaucracy, History of; Civil Service; Public Policy Schools; Public Sector Organizations; Think Tanks.

Bibliography

De Tocqueville, A., 1955. The Old Regime and the French Revolution, first ed. Doubleday, Garden City, New York.

Denhardt, Robert B., Catlaw, Thomas J., 2015. Theories of Public Organization, seventh ed. Wadsworth, Cengage, Stamford, CT.

Galnoor, I. (Ed.), 1977. Government Secrecy in Democracies. Harper and Row, New York.

Gidron, B., Kramer, R.M., Salamon, L.M. (Eds.), 1992. Government and the Third Sector. Jossey-Bass, Inc, San Francisco.

Huddleston, M.W., Boyer, W.W., 1996. The Higher Civil Service in the United States: Quest for Reform. University of Pittsburgh Press, Pittsburgh, PA.

Kaufman, H., 1976. Are Government Organizations Immortal? Brookings Institution, Washington, DC.

Milakovich, Michael E., Gordon, George J., 2013. Public Administration in America. Wadsworth, Cengage, Boston, MA.

Piore, Michael J., 2011. Beyond markets: sociology, street-level bureaucracy, and the management of the public sector. Regulation & Governance 5 (1), 145–164.

Rourke, F.E., 1984. Bureaucracy, Politics, and Public Policy. Little, Brown, Boston.

Stazyk, Edmund C., Goerdel, Holly T., 2011. The benefits of bureaucracy: public managers' perceptions of political support, goal ambiguity, and organizational effectiveness. Journal of Public Administration Research and Theory 21 (4), 645–672.

Public Choice Economics

Michael C Munger, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public choice is the application of economic methods and behavioral assumptions to nonmarket collective choice institutions. There are six major questions that public choice has focused on: collective action, controlling Leviathan, delegation, democratic coherence, information problems, and rent-seeking. Public choice has important implications for institutional design, particularly under the assumption that political actors may be motivated by interests other than the public interest. Public choice has given rise to important areas of study in empirical behavioral economics, including experimental economics, and to the study of common pool resource management institutions.

Introduction

The school of social science called 'public choice' is not sharply identified as an approach or a subject matter, but includes a variety of types of research. Nonetheless, as a general matter, public choice can be defined as the application of economic methods and behavioral assumptions to collective nonmarket institutions.

The foundational assumptions used are often methodological individualism, constrained utility maximization for all human agents, high transactions costs, and incomplete and/or asymmetric information. The focus of study is on institutions, formal or informal, that aggregate information or preferences, and that foster (or thwart) cooperative action intended to capture collective benefits. James Buchanan, 1986 Nobel Laureate and one of the founders of the public choice movement, has argued in several publications (Buchanan and Tullock, 1962; Buchanan, 1975) that public choice is a theory of exchange-based government by self-interested individuals. More briefly still, Buchanan (1999) simply said that public choice is 'politics without romance,' but he was careful to define 'politics' as a means of exchange, or achieving cooperative outcomes and mutual gains.

Six Key Problems

There are six key problems that public choice theory (PCT) has focused on. These six are listed below, along with the scholar or scholars who made the most important contributions in describing the problem:

- Collective action (Olson)
- Controlling Leviathan (Buchanan)
- Delegation (Niskanen and McNollGast)
- Democratic coherence (Arrow, Black, and McKelvey)
- Information problems (Coase, Downs, Hayek, and Tullock)
- Rent-seeking (Krueger and Tullock)

Each of these fundamental problems has important connections to other work in economics, law, or political science, but the particular contributions of PCT are unique and identifiable. (For an extended commentary on this work, and other aspects of the development and impact of PCT, see Mueller, 2003; Rowley and Schneider, 2003.) That is, research on these six problems is not necessarily a contribution to PCT, but PCT is largely composed on work on these questions.

Collective Action

The classical theories of interest groups (Bentley, 1908; Truman, 1951) conceived of group preference intensity as an anodyne counterweight to simple majorities. But there was an underlying assumption that intensity of preference measured total costs and benefits, defined somehow in interpersonally comparable utility units or in money. This meant that outcomes of interest group contests were more likely to track the optimal outcomes for the society. The reason is that interest groups were conceived as vectors of force (intensity multiplied by group size) allowing a comparison of policy choices, weighted by utilities.

This optimistic view of government action was forever shattered by the publication of Mancur Olson's *Logic of Collective Action* (1962). Olson's argument had two steps. First, 'groups' are different from collections of citizens. To form a group, citizens need some form of organization, and means of acting collectively, to focus their political power. Second, not all collections of citizens are equally likely to form an effective group. There are three factors that substantially determine the success of latent collections of citizens in engaging in effective collective action. These are (1) group size; (2) individual benefits and costs of participation; and (3) the capacity to provide selective incentives for group participation.

Olson pointed out that small groups, because they face lower costs of communicating and negotiating, and are more likely to have personal connections, are more likely to form effective political units. Large groups may fail to form effective units, precisely because their large size and anonymity allow individuals to 'free ride,' and fail to contribute.

The free rider problem is exacerbated if benefits are dispersed. It is easy to imagine that a group stands to benefit $1000 if it can raise $200 to organize. If the group contains four members, each of whom expects $250 in benefits, it is likely one member would pay the entire $200 cost, if necessary, because there is still a net benefit of $50 to that individual. But if there are 10 members of the group, each of whom expects a benefit of $100, no one will pay $200. If the group fails to organize, there is no chance of acquiring the benefit.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71071-3

Olson's conclusion was not that groups fail to organize. Rather, he concluded that successful groups managed to communicate and solve the free rider problem either through social capital or through the provision of direct benefits to participation that might be quite orthogonal to the group's public purpose. He called these orthogonal benefits 'selective incentives.' One might not be interested in a group's political agenda, but might be willing to contribute if the group can provide access to a magazine with pictures he or she enjoys (guns for the National Rifle Association, cute owls and seals for the Sierra Club). Or, one might contribute to public radio because he or she really wants the mug and the tote bag.

Controlling Leviathan

The symbol of the journal, *Constitutional Political Economy*, cofounded by James Buchanan, is a struggling Odysseus bound to the mast, with the Sirens in the background. The key problem Buchanan has been concerned about is reaching and enforcing agreements among groups of people who recognize that all have substantial incentives to cheat on the agreement once it is reached. If the enforcement powers of the state that the group constitutes are too weak, then cheating will be immanent. But if the enforcement powers are too strong, or cannot be constrained to the domain and procedures outlined in the agreement, then the Leviathan itself will cheat. Hobbes (1651/1990) despaired when confronted with this problem, and concluded a bad autocrat is better than anarchy. Buchanan argues that there is another option: constitutional contractarianism.

PCT (as Buchanan and Tullock, 1962, argued) conceives of government, for the most part, as an enforcement mechanism for capturing gains from exchange, based on voluntary agreements among groups of citizens. Government's function is to specialize in the production of public goods, controlling for externalities, and adjudicating disputes over property rights. Consequently, there are two stages of interaction. The first is the selection of rules – including rules for changing the rules – for which unanimity or something close to it is desirable. After rules are selected, participants choose actual strategies and actions, and outcomes are the product of the interactions of those strategies, under the rules. No one can choose outcomes, alone.

The problem is that constitutions may not be able to restrict the use of coercion to the domain everyone agreed on (Buchanan, 1975). Further, even a properly constrained, well-functioning government will sometimes engage in actions that look very much like coercion, because majorities choose policies and people who violate the legislated rules are punished. But the 'coercion,' in this view, is voluntary, because participants consented to the rules. As Brennan and Buchanan (1985) note, the problem of enforcement may be less severe than what a pure self-interest model might predict, because if the rules are consistent, and fairly enforced, citizens may internalize the rules as being 'the right thing to do,' creating moral sanctions against rule violation.

This suggests two important roles for limiting Leviathan. The first is that a set of choices and actions is set aside for individual autonomy, beyond the power of majorities to dictate or interfere. In the United States, this takes the form of a 'Bill of Rights,' limiting the power of government. The second is that rule of law and equal protection are absolute requirements for the application of legitimate coercion. A group of people might decide, for example, to impose taxes on themselves to finance the provision of a public good they all desire (North and Weingast, 1989). But the tax rate could not discriminate, and all citizens would have to be treated equally, and have recourse to due process for appealing decisions of the taxing authority.

Delegation

To 'delegate' is to entrust obligations to others, at the same time giving those others the powers to fulfill the obligations. In government systems, the policy/administration dichotomy – or any 'separation of powers' structure – requires an act of delegation. There are two large literatures that come together in the study of delegation in PCT. The older literature is on bureaucracy; the second, newer literature addresses the principal-agent problem.

Max Weber wrote that the bureaucracy represented an improvement over previous systems of organization. Ludwig von Mises (1944) and later William Niskanen (1971) were less optimistic about delegation of powers to bureaucracy, because bureaucrats are people, too. A 'public interest' theory of the institution of government is plausible enough, because government is established to secure and enforce the terms of the collective contract. But it is not plausible to assume that, just because bureaucracies are established to advance the public interest, the actual men and women who work in bureaucracies have the public interest as their personal goal. Consequently, they are no less likely than any other kind of citizen to try to manipulate pay, work conditions, and information to their advantage. The problem is that bureaucracy is uniquely insulated from normal economic incentives for efficiency. This is true partly because of the structure of bureaucracy, focusing on rules and procedures rather than outcomes or production.

But this is not a flaw of bureaucracy itself. Rather, the reliance on government provision in a territorially extensive state requires the use of bureaucracy. There is actually no alternative form of organization. The question, posed pointedly by Niskanen, is whether this means that the selection of government provision is correct in the first place. Niskanen noted that bureaucrats can be expected to try to expand their budgets, motivated either by simple self-interest or a public-spirited honest belief that their agency needs to increase its level of service provision and transfer payments. Given the ability of top agency officials to disguise their true costs and manipulate budgets, this means that elected officials charged with oversight are unlikely to be able to have enough information to control agency actions effectively.

Of course, Niskanen goes on to point out that it may not matter much anyway. The official who selects a position on the oversight committee for an agency is likely to have a personal or electoral interest in that agency's activity and budget. Consequently, if oversight officials are themselves relatively high demanders of the agency's services, the 'oversight' process is, in Niskanen's phrase, a 'stylized farce,' with oversight committees' main job being to check to see if the agency has correctly estimated the maximum budget that can pass the entire legislature on final passage.

The second major approach in PCT, the principal-agent literature, considers some of the same questions, but looks at the problem through the lens of contract design and – in the case of bureaucracy at least – the promulgation of administrative procedures and law as a means of eliciting the desired performance in bureaucratic agencies. One of the first articles to take this problem seriously was Romer and Rosenthal (1978), who argued that the problem of control in delegation rests on three institutional variables: agenda control, the possession of the *de jure* or *de facto* veto powers, and the reversion point, or outside option in negotiations. Barro (1973) and Ferejohn (1986) made important contributions to the idea of delegation of authority from voters to elected officials.

Some of the most important work in principal-agent delegation in PCT is by the composite author whimsically referred to as 'McNollGast,' or McCubbins, Noll, and Weingast. The precursors to this work include Weingast and Moran (1983), Weingast (1984), and McCubbins and Schwartz (1984). This led to McCubbins et al. (1987, 1989), which, along with Miller (1992) are the major summary of the administrative and management procedures problem in principal-agent PCT.

With Miller (1992), the literature had connected back to the origins, in Niskanen (1971), of a general model of organizational structure. Though Niskanen had written about bureaucracy and 'representative government,' his original inspirations to think about the bureaucratic form had come while he was employed as chief economist at Ford Motor Company, and witnessed there also the difference between 'markets and hierarchy,' as Miller put it. The fact is the market processes and bureaucratic processes are substitutes only at the margin. In most instances, it is clear what organizational form is required, if the activity is to be carried out efficiently.

Democratic Coherence

Groups of people who all agree, unanimously, have no problem deciding things. Literally any responsive (i.e., based on voter preference) decision rule would yield the same, correct outcome. Democracy matters most when there is disagreement. The difficulty is that when there is disagreement there is a chance of political incoherence, because democratic procedures are indeterminate.

This possibility was originally recognized long ago (McLean and London, 1990, 1992). But the problem was rediscovered, for majority rule, by Black and Newing (1951) and most importantly Black (1958), one of the founding documents of the public choice movement. It turns out that if there are three choices, three choosers, and complex disagreement, then majority rule may be radically indeterminate. Black (1958), to his credit, investigated whether this problem had been investigated before, and uncovered contributions by Condorcet and Borda whose implications had not been recognized – not even by Condorcet and Borda themselves!

A parallel development (Arrow, (1951/1963)) connected to this result by showing that the indeterminacy was not simply a property of majority rule, but of any aggregation mechanism. Arrow showed that, making only apparently innocuous assumptions about the features of the preference aggregation rule, or (in the case of welfare economics) the social welfare function, there is no mechanism other than dictatorship that always produces a complete and transitive aggregate ordering over outcomes. To the extent that dictatorship is ruled out by normative concerns for citizen participation and autonomy, then, there is no consistently ethically defensible solution to the aggregation problem.

And yet groups have to choose. The public choice conclusion is that what is required is humility; the outcomes selected by voting processes, ranging from elections to simple democratic choices over policy, are the results of a decision process, and nothing more. A different decision process would almost certainly have yielded a different outcome, even if exactly the same preferences were being 'represented.' Thus, since there is no unique mapping from preferences to outcomes, institutions of choice are central problems for a polity. As William Riker (1982) put it:

> We should never take the results of any method always to be a fair and true amalgamation of voters' judgments. Doubtless the results often are fair and true; but unfortunately, we almost never know whether they are or not. Consequently, we should not generally assume that the methods produce true and fair amalgamations. We should think of the methods, I believe, simply as convenient ways of doing business, useful but flawed. This gives them all a place in the world, but it makes none of them sacrosanct. (p. 113)

This work has led to an accumulation in the understanding of the nature of political power in a democracy. It is now known that political power lies at the 'center' of the distribution of preferences effectively enfranchised by the institutions of the society (Downs, 1957; Black, 1958). Majority rule has substantial power, from an ethical perspective, as a normatively defensible decision rule (May, 1952). But it turns out the sufficient conditions for the existence of equilibrium in majority rule decision (Plott, 1967) still leave the question of the necessary conditions for equilibrium dependent on a variety of assumptions about information, institutions, agenda control, and voter sophistication (McKelvey, 1976; Hinich, 1977; Enelow and Hinich, 1983; McKelvey and Schofield, 1987).

There are two separate problems of democratic coherence. One is the problem of agenda control and manipulation. The power to choose institutions can be tantamount to dictatorship, or at best give the agenda controller disproportionate power inconsistent with a democratic 'one person, one vote' ethic (McKelvey, 1976; Miller, 1995). Of course, disputes over outcomes can metastasize into disputes over institutions of choice, as Riker (1980) famously pointed out. The second problem is the arbitrary or 'imposed' nature of institutions. Having one institution for choosing, rather than another, can have a striking impact on outcomes, just for reasons of historical accident or path dependence (Shepsle, 1979, 1991; North, 1981; Baron and Ferejohn, 1989). Since institutions are costly to change, and since the translation of preferences into outcomes is so uncertain, collective choices may be locked into suboptimal arrangements for long periods.

Information Problems

There are at least three contexts in which the problem of costly, limited, or asymmetric information is central to PCT. The first is the claim, derived from the Austrian position in the 'socialist

calculation debate' (Hayek, 1945) of the 1930s and 1940s, that governments could not possibly be able to gather enough information to make accurate judgments about resource flows and production decisions. Information is simply too dispersed, too specific, and too dynamic to allow aggregation.

The second information problem is the 'rational ignorance' of voters. Downs (1957) pointed out that individual voters have at best limited incentives to acquire accurate information about candidates, without which the accuracy of representational choices in a republic is suspect. As a result, institutions such as parties (Aldrich, 1995) or ideologies (Hinich and Munger, 1994) are important components of large, complex political systems. In fact, if Downs is right, democracy would be effectively impossible with parties and ideologies.

Third, there is the micro-level concern about specific local knowledge needed to resolve disputes and impose judicial remedies. As discussed above, the macro problem of organizing the economy is an obstacle to valuing commodities in the aggregate, and in rationing valuable resources in the face of competing uses. But the problem of transaction costs in local disputes over externalities is the micro-level analog, and represents a substantial difficulty for regulatory decisions. As Coase (1960) pointed out that, the very transaction costs that make private negotiations over nuisances and externalities problematic are likely to hamper government efforts to solve the same problem. Again, this problem is related to Austrian economists' (Kirzner, 1985) concerns over the knowledge required to assign the 'correct' value to resources and nuisances. But the problem is even more severe, because of the subjective nature of value, and the requirement that government officials must make interpersonal utility comparisons.

These information problems come together in the so-called revelation problem, long of interest to PCT scholars. After all, there exists a determinate solution to the problem of optimal tax shares for provision of a public good (Lindahl, 1919), but this solution relies on the availability of accurate – and honest – valuations by citizens. There was interest in PCT on the question of demand revelation (Tideman and Tullock, 1976; Tullock, 2006), but the complexity of the tax schemes involved turned out to be prohibitive.

Rent-Seeking

Economists have long believed that competition is good, and more competition is better. For this reason, many economists support antitrust policies that limit the market power of existing firms, because of the deadweight losses associated with pricing above marginal cost, or 'rents.' Harberger (1954) famously disputed whether monopolies really cause that much harm, and offered empirical measures that suggested that the static deadweight losses, at least, were not enormous. There was general agreement that the rectangle of monopoly profits, or rents, was simply a transfer and therefore economically innocuous.

But Tullock (1967) recognized a mechanism by which not just the triangle, but also the rectangle, was actually social costs. The reason is that a 'rent seeker,' or economic agent pursuing the rents of the rectangle of monopoly profits, would be willing to devote resources up to the value of the rectangle to get, and then retain, access to those rents. The resources spent on rent-seeking were consequently not a transfer, but might represent pure deadweight losses. If the activities involved lobbying, bribery, or the creation of asset-specific capital to capture the rent, those resources would be taken away from other productive activities.

Worse, if government has been complicit, creating artificial benefits such as trade barriers against foreign competition, or licensing or other requirements to erect entry barriers domestically, firms competing for these benefits will dissipate much of the benefit. If government can change the rules to create an unfair benefit for one competitor, in terms of access to contracts or patents, competition over that rule change will dissipate a value approaching the expected rent from the benefit.

A business firm might invest in new plant and equipment, achieving lower prices and higher quality in a way that benefits consumers but harms the rest of the industry. The harm to the rest of the industry takes place within the existing rules of competition, and is generally more than offset by the far larger benefits to consumers. But a business firm might also invest in lobbying the government to create new rules, such as protections from competition that raise prices and destroy incentives for innovation. The benefit to the industry results from a political 'investment,' and the harm to consumers is far larger than the benefit to firms and workers.

The literature on rent-seeking has become quite large, and has expanded beyond PCT. The actual term 'rent-seeking' was first used by Krueger (1974). Later work on measuring and refining rent-seeking has included estimates of the proportion of the 'rent-seeking sector' of economies, the size of growth loss to rent-seeking, and other topics (see Rowley et al., 1988; Tollison, 2007).

Human Motivation and the Problem of Institutional Design

PCT is sometimes criticized as being a cynical or oversimplified view of human action. But PCT is, for the most part, self-consciously modest about what it can explain and where it should be applied. The first step in a PCT explanation is deceptively narrow: what would happen if people were self-interested? But it is important to keep track of just what the theory is trying to explain. The goal is to ask what institutions would look like if a very simple model of human behavior were assumed and its implications explored. This is quite different from saying that certain conclusions will be derived, but that these conclusions depend for their validity on the literal correctness of the self-interest assumption.

A clear example is Mayhew (1974). He asks, what would the institutions of the US Congress look like if members of Congress were single-minded seekers of reelection? Then he compares the answer to that 'as if' question against US Congressional institutions one actually observes.

What did he find? There are no important differences. The institutions of Congress one actually observes are indistinguishable from those implied by a consciously oversimplified behavioral assumption of narrow self-interest.

The organization of Congress meets remarkably well the electoral needs of its member. To put it another way, if a group

of planners sat down and tried to design a pair of American national assemblies with the goal of serving members' electoral needs year in and year out, they would be hard pressed to improve on what exists (pp. 81–82).

None of this implies members of Congress are exclusively, or even primarily, selfish. But actual institutions can be explained, at least in part, by a model that allows selfishness. An earlier example, using assumptions much like what later became PCT, is James Madison's description of the problem of institutional design.

Ambition must be made to counteract ambition. The interest of the man must be connected with the constitutional rights of the place. It may be a reflection on human nature, that such devices should be necessary to control the abuses of government. But what is government itself, but the greatest of all reflections on human nature? If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government that is to be administered by men over men, the great difficulty lies in this: one must first enable the government to control the governed; and in the next place oblige it to control itself. A dependence on the people is, no doubt, the primary control on the government; but experience has taught mankind the necessity of auxiliary precautions (Madison, Federalist #10, Rossiter, 1961).

The problem, then, is not whether, or how often, the PCT assumption of self-interest is correct. Rather, the problem is to design institutions that will function even if political actors were motivated by ambition rather than the public interest.

The fact is that PCT has long recognized that the assumption of self-interest is at most one view among many. Often, all that is required is the notion that political actors are purposive, or goal-oriented, and that those purposes or goals motivate the choices and strategies actors select. As Buchanan and Tullock (1963: pp. 16–17, emphasis added) note, the idea of genuine egoism, pure self-interest, as a motivation has always been more interesting to critics – as a straw man – than it has been to proponents.

The first point to be noted is that economic theory does not depend for its validity or its applicability on the presence of the purely economic man. *This man of fiction, who is motivated solely by individual self-interest in all aspects of his behavior, has always represented a caricature designed by those who have sought to criticize rather than to appreciate the genuine contribution that economic analysis can make, and has made, toward a better understanding of organized human activity….*

It is also necessary to emphasize that [PCT] *does not try to explain all human behavior, even all of that which might be called 'economic' in some normally accepted sense of this term. At best, the theory explains only one important part of human activity in this sphere. It examines one relationship among individuals in isolation….*

Empirical Implications and Connections

When James Buchanan won the Nobel Prize in 1986, Alistair Cooke (in his 'Letter from America' program on BBC 4) noted that PCT was committed to the "homely but important truth" that politicians are really no different from the rest of us (quoted in Brennan and Lomasky, 1993: p. 2). The idea that human behavior is of a piece, affected most by differences in incentives and institutions, rather than having entirely different goals and motivations in different contexts, has had an impact beyond the boundaries of traditional PCT.

Some of the most interesting and important work in the last two decades has been in experimental and behavioral political economy, as well as the study of comparative institutions. In addition to James Buchanan, two other past Presidents of the Public Choice Society have been awarded Nobel Prizes: Vernon Smith (2002) and Elinor Ostrom (2009). Smith played a central role in the creation and expansion of the Economic Science Association, which is the flagship organization for presenting work in behavioral economics experiments. And Ostrom played a central role in the creation and expansion of the International Association for the Study of Common Property, which is concerned with institutions for solving common pool resource problems and other problems of capturing gains from cooperation and exchange in nonmarket settings.

As a body of work, PCT is in many ways a counterpoint to the tradition of welfare economics in the 1940s and 1950s, which looked at tax, expenditure, and regulatory policies as solutions to problems of 'market failure'. The counterclaim of the public choice movement is not that markets are perfect. Rather, what PCT does is point out that the logic of 'markets fail, therefore government can do better' is at best incomplete. Each of the six problems described here involve a difficulty that government action has to overcome. Taken together, these problems represent sources of 'government failure', which would have to be balanced against market failure in deciding how a society might best organize itself.

PCT is a direct answer to the assumption of a benevolent government, both in terms of collective action and agency problems. Political groups and officials cannot simply be assumed to act in the public interest. The 'will of the people' may be hard to gauge, and the coercive powers of the state may be very difficult to constrain. Taken together, these observations raise important questions about the 'market failure' justification for state action. The claim is not that market failure is not real; rather, PCT is skeptical that government has the incentives (as noted above) or the information necessary to do better.

Interestingly, these problems had been recognized much earlier, by the very scholar usually credited with originating the idea of 'corrective' taxes and subsidies as a means of correcting market failures. A.C. Pigou (1920) aptly summarized the humility that PCT would urge on planners and government officials, and recognized the inherent limitations of state action:

> It is not sufficient to contrast the imperfect adjustments of unfettered enterprise with the best adjustment that economists in their studies can imagine. For one cannot expect that any State authority will attain, or even wholeheartedly seek, that ideal. Such authorities are liable alike to ignorance, to sectional pressure, and to personal corruption by private interest. A loud-voiced part of their constituents, if organized for votes, may easily outweigh the whole. (p. 296)

See also: Antitrust Policy: Lessons from the US; Behavioral Economics; Global Public Policy; Governance; Public Administration: Organizational Aspects; Public Sector Organizations; Regulation, Political Economy of.

Bibliography

Aldrich, John, 1995. Why Parties? University of Chicago Press, Chicago.

Arrow, K.J., 1951/1963. Social Choice and Individual Values, first ed. Yale University Press, New Haven, CT. published by Wiley, 1952.

Baron, D.P., Ferejohn, J.A., 1989. Bargaining in legislatures. The American Political Science Review, 1181–1206.

Barro, Robert, 1973. The control of politicians: an economic model. Public Choice 14, 19–42.

Bentley, Arthur, 1908. The Process of Government. University of Chicago Press, Chicago.

Black, Duncan, 1958. The Theory of Committees and Elections. Cambridge University Press, New York. Reprinted from: Kluwer Academic, Netherlands (1987).

Black, Duncan, Newing, R.A., 1951. Committee Decisions with Complementary Valuation. Lowe and Brydon, London.

Brennan, Geoffrey, Buchanan, J.M., 1985. The Reason of Rules: Constitutional Political Economy. Cambridge University Press, Cambridge.

Brennan, Geoffrey, Lomasky, Loren, 1993. Democracy and Decision: The Pure Theory of Electoral Politics. Cambridge University Press, New York.

Buchanan, James, Tullock, Gordon, 1962. The Calculus of Consent. University of Michigan Press, Ann Arbor.

Buchanan, James, 1975. The Limits of Liberty: Between Anarchy and Leviathan. University of Chicago Press, Chicago.

Buchanan, James, 1999. "Public Choice: Politics without Romance." The Logical Foundations of Constitutional Liberty. Liberty Fund.

Coase, R.H., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Downs, Anthony, 1957. An Economic Theory of Democracy. Harper and Row, New York.

Enelow, James, Hinich, Melvin, 1983. On Plott's pairwise symmetry condition for majority rule equilibrium. Public Choice 40, 317–321.

Ferejohn, John, 1986. Incumbent performance and electoral control. Public Choice 50, 5–25.

Harberger, A.C., 1954. Monopoly and resource allocation. The American Economic Review 44 (2), 77–87.

Hayek, F.A., 1945. The use of knowledge in society. The American Economic Review 35 (4), 519–530.

Hayek, F.A., 1978. Law, Legislation and Liberty. The Mirage of Social Justice, vol. 2. University of Chicago Press, Chicago, IL.

Hayek, F.A., 1988. The Fatal Conceit: The Errors of Socialism. University of Chicago Press, Chicago.

Hinich, Melvin, 1977. The median voter result is an artifact. Journal of Economic Theory 16, 208–219.

Hinich, Melvin, Munger, Michael, 1994. Ideology and the Theory of Political Choice. University of Michigan Press, Ann Arbor.

Hobbes, Thomas, 1651/1990. Leviathan. Hackett Publishers.

Kirzner, I.M., 1985. The perils of regulation: a market-process approach. Discovery and the Capitalist Process. University of Chicago Press, Chicago.

Krueger, Anne, 1974. The political economy of the rent seeking society. The American Economic Review 64, 291–303.

Lindahl, Erik, 1919. Just taxation: a positive solution (E. Henderson, Trans.). In: Musgrave, R.A., Peacock, A.T. (Eds.), Classics in the Theory of Public Finance. Macmillan, London, pp. 168–176 (1958).

May, K.O., 1952. A set of independent, necessary, and sufficient conditions for simple majority decision. Econometrica 20, 680–684.

Mayhew, David, 1974. Congress: The Electoral Connection. Yale University Press, New Haven.

McCubbins, M.D., Schwartz, T., 1984. Congressional oversight overlooked: police patrols versus fire alarms. American Journal of Political Science 28, 165–179.

McCubbins, M.D., Noll, R.G., Weingast, B.R. (McNollGast), 1987. Administrative procedures as instruments of political control. Journal of Law, Economics, and Organization 3, 243–277.

McCubbins, M.D., Noll, R.G., Weingast, B.R. (McNollGast), 1989. Structure and process, politics and policy: administrative arrangements and the political control of agencies. Virginia Law Review 75, 431–482.

McKelvey, Richard, 1976. Intransitivities in multidimensional voting bodies and some implications for agenda control. Journal of Economic Theory 30, 283–314.

McKelvey, Richard, Schofield, Norman, July 1987. Generalized symmetry conditions at a core point. Econometrica 55, 923–933.

McLean, Iain, London, John, 1990. The borda and condorcet principles: three medieval applications. Social Choice and Welfare 7, 99–108.

McLean, Iain, London, John, 1992. Ramon Llull and the theory of voting. Studia Lulliana 32 (1), 21–37.

Miller, Gary, 1992. Managerial Dilemmas: The Political Economy of Hierarchy. Cambridge University Press, New York.

Miller, Nicholas, 1995. Committees, Agendas, and Voting. Chur, Switzerland.

Mises, Ludwig (von), 1944. Bureaucracy. Yale University Press, New Haven.

Mitchell, W.C., 1988. Virginia, Rochester, and Bloomington: twenty-five years of public choice and political science. Public Choice 56 (2), 101–119.

Montesquieu, Baron de (Charles de Secondat), 1750/1989. Spirit of the Laws. (English edition). Cambridge University Press, New York.

Mueller, Dennis, 2003. Public Choice III. Cambridge University Press, New York.

Niskanen, William, 1971. Bureaucracy and Representative Government. Aldine-Atherton, Chicago.

North, Douglass, 1981. Structure and Change in Economic History. Norton, New York.

North, D.C., Weingast, B.R., 1989. Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. The Journal of Economic History 49 (04), 803–832.

Nozick, Robert, 1974. Anarchy, State, and Utopia. Basic Books, New York.

Olson, Mancur, 1962. The Logic of Collective Action. Harvard University Press, Cambridge.

Ostrom, Elinor, 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Pigou, A.C., 1920. The Economics of Welfare. MacMillan and Company, London.

Plott, Charles, 1967. A notion of equilibrium and its possibility under majority rule. The American Economic Review 57, 787–806.

Rawls, John, 1971. The Theory of Justice. Harvard University Press, Cambridge, MA.

Riker, W.H., 1980. Implications from the disequilibrium of majority rule for the study of institutions. American Political Science Review 74, 432–446.

Riker, W.H., 1982. Liberalism against Populism: A Confrontation between the Theory of Democracy and the Theory of Social Choice. W.H. Freeman and Company, San Francisco, CA.

Romer, Thomas, Rosenthal, Howard, 1978. Bureaucrats vs. voters: on the political economy of resource allocation by direct democracy. Public Choice 33, 27–44.

Rossiter, Clinton, 1961. The Federalist Papers. New American Library, New York, p. 558.

Rousseau, Jean-Jacques, 1762/2003. The Social Contract (G.D.H. Cole, Trans.). Dover Thrift Editions, London.

Rowley, C.K., Schneider, F. (Eds.), 2003. The Encyclopedia of Public Choice. Springer.

Rowley, C.K., Tollison, R.D., Tullock, G.C. (Eds.), 1988. The Political Economy of Rent Seeking. Springer.

Shepsle, Kenneth, 1979. Institutional arrangements and equilibrium in multidimensional voting models. American Journal of Political Science 23, 27–60.

Shepsle, Kenneth, 1991. Models of Multiparty Electoral Competition. Harwood Academic, Chur, Switzerland.

Tideman, Nicolaus, Tullock, Gordon, 1976. A new and superior process for making social choices. Journal of Political Economy 84, 1145–1159.

Tollison, R.D., 2007. Rent seeking: a survey. Kyklos 35 (4), 575–602.

Truman, David, 1951. The Governmental Process. Alfred A. Knopf, New York.

Tullock, Gordon, 1967. The welfare costs of tariffs, monopolies, and theft. Western Economic Journal 5, 224–232.

Tullock, Gordon, 2006. Demand revealing, transfers, and rent seeking. In: Rowley, C.K. (Ed.), Selected Works of Gordon Tullock. Liberty Fund, Indianapolis, pp. 142–149.

Weingast, B.R., 1984. The congressional-bureaucratic system: a principal-agent perspective (with applications to the SEC). Public Choice 44, 147–191.

Weingast, Barry, Moran, Mark, 1983. Bureaucratic discretion or congressional control? Regulatory policy-making by the federal trade commission. Journal of Political Economy 91, 775–800.

Public Goods: International

Adrienne Héritier, Robert Schuman Centre for Advanced Studies, European University Institute, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

International public goods are goods that are accessible to all at the international level, the consumption of which by one individual does not reduce the possibility of consumption by other individuals. However, there are few public goods of this pure type. Instead, nearly all are characterized by mostly unlimited accessibility and simultaneous rival consumption, i.e., the use of the resource by one individual reduces the possibilities for others to consume this good. These types of goods are called common-pool resources. A number of institutional possibilities exist to secure the provision of public goods and common-pool resources. In what follows, public goods will first be defined, and then different paths to the development of institutional answers to collective action problems will be presented, before different institutional modes of provision of public goods and common-pool resources as well as their possible implications will be discussed.

International public goods are goods that are accessible to all at the international level, the consumption of which by one individual does not reduce the possibility of consumption by other individuals. However, there are few public goods of this pure type. Instead, nearly all are characterized by mostly unlimited accessibility and simultaneous rival consumption, i.e., the use of the resource by one individual reduces the possibilities for others to consume this good. These types of goods are called common-pool resources. Where there is free accessibility and rival consumption, a collective action problem or social dilemma arises: how can the individual user be prevented from using the limited resource and jeopardizing the long-term utility of all as a result of maximizing their own short-term individual interest?

A number of institutional possibilities exist to secure the provision of public goods and common-pool resources. However, the possibilities for collective action within a specific institutional context available at national level – such as legal prescriptions imposed by majority decision in order to force individual actors to comply with a scheme of collective action – are not all applicable at the international level because of nation-state sovereignty. In what follows, public goods will first be defined, and then different paths to the development of institutional answers to collective action problems will be presented, before different institutional modes of provision of public goods and common-pool resources as well as their possible implications will be discussed.

Definition of Public Goods

Public goods and common-pool resources – as stated above – may be defined by the analytic economic criteria of universal accessibility and (non)rival consumption. These attributes may be properties that are inherent to the goods themselves, or they may be properties that are attributed to goods through societal consensus and/or political decision. In the older economic literature, the central attributes of public goods – nonexcludability and nonrivalry (Samuelson 1954; Musgrave and Musgrave 1976) – have been regarded as sources of market failure, from which the need for state intervention has been derived. The problem with common-pool resources and public goods lies in their limited excludability. As a result, prices cannot be charged and hence private provision through the market cannot be relied upon.

More recently, economic theory (Cornes and Sandler, 1996) has argued in a more differentiated way that there are exceptions to this rule when a particular actor has an incentive to provide a public good irrespective of the free-riding behavior of other beneficiaries. The same holds true when, in the case of a common-pool resource, the stakes of provision are high enough to induce cooperation among actors (Cornes and Sandler, 1996). Only in cases where the private provision of common goods fails are governments called upon to remedy the market failure (Bator, 1958). Another perspective from which public goods can be viewed is starting from the phenomenon that some market activities generate external effects. Positive, as well as negative, external effects are not part of the actor's economic calculation and, as a result, they are under- or overproduced from an aggregate view. In the traditional perspective, externalities are seen as by-products of economic activities and internalization of externalities traditionally has been viewed as a government task (Pigou, 1920). However, in the 1960s, a new interpretation started to dominate the economic literature. According to this view, Pareto-relevant externalities are internalized through market forces if property rights are well defined and transaction costs are sufficiently low (Coase, 1960; Buchanan and Stubblebine, 1962).

While economic theory views the properties of a public good as inherent attributes of such a good from which a specific institutional form of provision is deduced, political science goes beyond this notion to insist additionally on a process view of common goods. It argues that what constitutes such a good and how it is provided for institutionally depends on specific situational and technological conditions, and, moreover, is determined in a political and social process. First, it is not clear in many cases what constitutes a public good or a common-pool resource in so far as there are no absolute and objective criteria for defining them in a general sense. Instead, what constitutes a public good depends in part on the specific situational context of provision and consumption of the good, and its technical properties.

Thus, the importance of a specific situational context for the definition of a good as a public good in terms of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75045-8

nonexcludability and nonrival consumption is illustrated by the example of the protection offered by the ozone layer vis-à-vis radiation for sunbathers. This is a public good in the classical sense. One sunbather enjoying the sun does not impinge upon the protection offered to other sunbathers. In this sense, the ozone layer is a public good, accessible to all with nonrival consumption. In another sense, however, it may be perceived as a common-pool resource. Seen in terms of the relationship between producers and sunbathers, and given that scientific research has identified a problem of scarcity in the ozone layer, certain productive activities, such as the production of refrigerators with chlorofluorocarbons, diminish the protective quality of the ozone layer and jeopardize the protection offered to all. This is rival consumption. In this specific context, the ozone layer constitutes a common-pool resource that is accessible to all and is depletable (Snidal, 1984).

Technological innovation may force a reconsideration of the definition of a public good in cases where a good, which had hitherto been regarded as universally accessible and nondiminishable, has been rendered potentially excludable. A case in point is broadcasting. This good used to be accessible to all and was of nonexcludable nature, but as a result of technological change, access to the good can now be limited to those paying for the service via a telematic coding system. Put in different terms the transaction costs for establishing property rights were reduced drastically by technological progress so that the gains from excluding free riders and the precise allocation of user costs now exceed the costs of establishing property rights. Second, the definition of what constitutes a public good and what does not may not necessarily be determined by 'objective' characteristics, but instead by political and social definition. There are huge cultural variations across polities and time as to what is considered a public good or common-pool resource and what is a private good.

Thus, universal and free health care may be formulated as a political goal and consequently be considered as a common-pool resource in one polity. But in another, the same service may be regarded as a private good that each individual has to buy on the market. In the same way, policy makers in one society may decide that its provision should be left entirely to the buying power of the individual through the market. Hence, what constitutes a public good or common-pool resource can be answered either in terms of analytic economic criteria or in terms of a process of social and political definition. Once a good has been identified as a public good or as a common-pool resource, the institutional mode of provision of the good has to be determined.

The Provision of Institutional Answers to a Collective Action Problem

Different paths may lead to an institutional answer for a stable solution for providing public goods at the international level. The solutions may be designed intentionally, as is the case of international agreements. They may develop without being planned, as in a case of spontaneous coordination, for example, in the case of the standardization of products. Either way, once in place, they develop an institutional dynamism of their own and over a period of time they may transform the processes in the institution and, in turn, transform the preferences of the actors involved. The intentional formulation of an institutional solution to a public good problem may presuppose an acute awareness of an existing social dilemma and a need for collective action where there is cross-boundary problem interdependence. For example, in the case of an imminent cross-border environmental problem, an international conference will be convened that seeks to develop a format of cooperation.

After a lengthy process of bargaining, institutional answers for solving the problem will be found and agreed. All those involved in the negotiations will seek to maximize their own interests, for example, by minimizing the implicated costs. In other words, institutional solutions to collective action problems are interpreted in terms of their distributional impacts on the actors involved (Knight, 1995). In the great majority of cases, the negotiators do not all enjoy the same degree of power, and in the event of an agreement not being reached, the stronger actors may take up a fallback position. While a solution may be desirable for them, they may not consider it vital. In other words, they have more resources for negotiating an institutional solution. This route to an institutional solution, i.e., interest-oriented, power-based negotiation with a view to the distributional results, generally applies where the interests at stake are clear cut and substantial.

A case in point is the negotiation of CO_2 emission limits at international climate conferences. All the states engaged in the bargaining process, to find an institutional solution for reducing emissions, view the process in terms of its distributional impacts on the market position of domestic industries, and so forth. However, there are situations where collective action problems need to be solved and the actors involved do not really appreciate the nature of the problem at hand, much less which interests are at stake. Consequently, the preferences of the actors involved are uncertain. In such situations deliberation, exchange of arguments, and persuasion will prevail as the actors strive to reach, at best, a common perspective on the problem at hand together with possible solutions; or as a minimum, to establish clarity as to the different stakes involved (Knight and Johnson, 1994). Only then does interest-oriented bargaining follow and asymmetries of interests and power come into play. Cases in point are complex environmental problems or the problem of providing more transparency and certainty in financial markets.

The second avenue for developing an institutional solution to a collective action problem is that of the spontaneous emergence, rather than the intentional development, of a solution in a planned social interaction at international level. Here, the strategic uncertainty as to the preferences of other actors promotes behavior likely to resolve a social dilemma. It makes sense to shape other actors' perceptions of one's preferences and to give signals of willingness to coordinate and be trustworthy. These signals enable like-minded actors to locate others who are equally so, and who can in turn build up a reputation for responding cooperatively to trust (Bates, 1988). Because actors prefer coordination as one of the outcomes, as opposed to noncoordination, they will use whatever salient information (focal point) they have in order to achieve that coordination (Schelling, 1960). In time, other actors will follow suit, and a social convention is established,

which provides information about the future actions of others (Knight, 1995). The focal point may be relatively contingent, but henceforth serves as a point of reference for future coordinating actions (Calvert, 1986; Schelling, 1960).

Third, irrespective of the path along which an institutional solution develops, be it design or spontaneous emergence, once established, it gains a dynamism of its own. The institutional process then begins to develop its own rules beyond the formal and informal rules originally instituted, or interprets the original rules in the light of the day-to-day need to make decisions, thereby transforming and developing those rules incrementally (March and Olsen, 1989). This process may change the preferences of the actors involved. Not only the paths of development of an institutional solution to secure the provision of a public good or a common-pool resource but also the specific context in which it is shaped and the institutional arrangements in which it is embedded have been discussed in the literature.

The Context and Institutional Mode of Provision

What is the general context and what are the more specific institutional and instrumental conditions in which the provision of a public good and a common-pool resource occurs? The overall question is: under which conditions do particular institutional arrangements apply particular instruments to produce favorable results for providing public goods at the international level? With respect to the overall context, it is argued that the greater the number of participants involved in an attempt to resolve a social dilemma and the greater the heterogeneity of the actors involved, the greater is the difficulty in arriving at a solution. In this sense, the prospects of an efficient solution to collective action problems for agreements at the international level, with its heterogeneous actors and large numbers of participants, would appear to be bleak.

However, as has been pointed out (Ostrom, 1990; Martin, 1994; Snidal, 1984), this relationship is not as simple as it first appears. Heterogeneity may constitute a precondition for actors to take the lead in organizing collective action because they have a specifically strong incentive for this action to be successful. Thus, the United States has repeatedly acted as a hegemonic actor in cooperating with Western public and private actors when it comes to solving collective action problems such as sharing banking information to enable countries to uncover money laundering (Schmidt, 1995). Furthermore, the problem of coordinating an unwieldy number of actors is often alleviated by organizing the participants into groups according to commonly shared criteria. Each group is represented by a speaker, as in the case of climate negotiations, thereby reducing the number of actors sitting at the bargaining table. Moreover, the extent to which homogeneity facilitates negotiations in collective action problems depends upon the nature of the homogeneity: if all actors are intent on having the same resource, homogeneity may render agreement difficult. If, by contrast, the preferences of actors differ over time and in intensity, they may be made complementary through issue linkage (Martin, 1994).

Beyond the general context conditions, various institutional arrangements and instruments are applied in the provision of public goods and common-pool resources at the international level that determine the degree of success in solving the collective action problem. The institutional arrangement denotes the structure and the nature of the relationship among the actors involved; the instruments that are applied define how the desired behavior of the target actor is to be elicited. The basic types of institutional settings for providing public goods are hierarchy, negotiation, societal self-regulation, and markets (Mayntz, 1997; Scharpf, 1999). Hierarchy is an arrangement in which a central actor (government) has an electoral mandate that permits it to legislate specific provisions, such as the statutory requirement to use a fully regulated catalytic converter, with noncompliance being sanctioned in order to protect the quality of ambient air. Thus, legal obligation is backed up by a bureaucratic apparatus to ensure implementation.

In an international context, hierarchy as an institutional arrangement for providing public goods and protecting common-pool resources can only be applied to a very limited extent. However, even in international negotiations among sovereign states seeking to solve a collective action problem, a certain amount of initiative needs to be taken by one or a group of actors if public goods are to be provided jointly. As a result, there is power being yielded by central actors in varying degrees in order to facilitate cooperation. Thus, a central body may consist of member actors who only make framework decisions that subsequently are made more specific by actors at the decentral level, or the central body may delegate power to members to implement the decisions. However, it is also conceivable that the outcome of autonomous actors cooperating voluntarily is strict implementation by command-and-control once a decision is agreed. Depending on whether the decision is vaguely, as opposed to precisely, formulated, the ensuing latitude for such actors is either great or little. Power to sanction noncompliance may again be vested in a central body or, alternatively, in decentral bodies (Martin, 1994).

Another important institutional setting in which solutions to collective action problems at the international level are found is negotiations in bipartite, tripartite, or multipartite actor constellations. However, it can often be linked to elements of various forms of hierarchy. Thus, if negotiations are not successful in bringing about a solution to a problem, actors may move out of the arena and turn to a court or a tribunal for a ruling in order to overcome the conflict. In anticipating this possibility, actors may be more willing to compromise. Thus, in the European dispute over social policy, if social partners do not negotiate an agreement successfully, the Commission may propose legislation for submission to the Council of Ministers. The 'shadow of hierarchy' (Scharpf, 1997) speeds up negotiations. Some actors are particularly apt at brokering agreements in negotiations at international level in order to bring about a common solution. In the European context, the Commission has played a preeminent role in hammering out such agreements by acting as a policy entrepreneur and broker in different arenas, bargaining with various actors simultaneously in order to develop solutions acceptable to all, for example, by proposing package deals and compensation payments. (Grande, 1995).

Yet another institutional arrangement is the self-organization and self-regulation of actors. Self-regulation in providing a public good or common-pool resource implies that

the same actors who make joint decisions about how to provide the common good also implement these decisions. The actors concerned build on the motivation, expertise, and resources of the actors immediately affected in order to provide public goods (Mayntz and Scharpf, 1995). These forms of self-governance are also linked frequently to elements of hierarchy, such as the possibility of appeal to a third-party tribunal (Ostrom, 1990). Finally, markets may serve as an institutional context for the provision of public goods and common-pool resources under specific conditions, for example, when a group of actors have such high stakes that they engage in the provision. It has been increasingly claimed that quasimarkets should be created to provide public goods and common-pool resources (Le Grand, 1991).

Thus, many Western countries have promoted the liberalization of state monopolies in the provision of network utilities. However, a pure form of market as an institutional arrangement for providing public goods and common-pool resources can rarely be found either. Rather – as, for example, in the case of the provision of network utilities – markets are regulated at two levels: in terms of market creation to guarantee competition on the one hand, and at the level of market correction as regards 'service public' obligations to guarantee the accessibility and affordability of these services on the other hand. Here – as with other institutional arrangements – two or three different types of institutional arrangements for providing public goods or common-pool resources are combined rather than existing separately as pure forms.

Moreover, institutional solutions for the usage of common-pool resources may be developed if different types of property rights are distinguished. Schlager and Ostrom (1992) distinguish the 'right of access' to a defined physical area and its nonrival use from the 'right of withdrawal' of resource units; the 'right of management' of a resource (i.e., regulate its use); the 'right of exclusion,' i.e., to determine who has access; and the 'right of alienation,' i.e., the right to sell or lease management and exclusion rights. Thus, in the case of groundwater basins, individual actors may have a withdrawal right of a specific quantity of water, but the groundwater basin may be managed by government (Ostrom, 2002: p. 49).

A further, more contextual institutional mechanism for solving collective action problems is the development of 'social capital' and trust among the actors concerned (Taylor, 1996; Hardin, 1993). The underlying notion is that one can trust the others involved to cooperate in the provision of a public good. Trusting others is reasonable because it is in their interest to be trustworthy in a specific situation in which a social dilemma is to be overcome. One's own trust depends on the interest of the trusted (Hardin, 1999). The instruments of command-and-control, bargaining, setting incentives, persuasion, and information are often closely linked to the institutional settings in which they are applied. Thus, command-and-control is often applied in a hierarchical context where a government authority requires the regulatees to behave in a particular way and where sanctions are applied in cases of noncompliance. However, a majority decision can also be made in favor of applying incentives.

Hence, hierarchy may be linked to instruments other than command-and-control. In the institutional setting of negotiation in bi-, tri-, and multipartist settings, bargaining, together with issue linkage, compensation payments, and compromises are the main instruments used to reach agreement, while deliberation and argument also play an important part. The negotiation process may, however, also result in the application of a strict prescription of behavior linked with the levying of fines. Incentives are set in the institutional environment of the market. Trust would be linked to information and persuasion.

Implementation Agreement

Implementation agreement on a joint institutional solution to provide a public good does not automatically guarantee its implementation. The path from decision making to the final realization of the measure is a long one, and many obstacles may prevent it from being a smooth process (Pressman and Wildavsky, 1973; Mayntz, 1983; Windhoff-Heritier, 1980).

At the international level, typically we are faced with the situation that the definition of the goods and institutional solutions for providing the public good or common-pool resource occur at international level, while implementation takes place at the national level. This may raise particular problems of compliance. Thus, for example, if an international environmental agreement stipulates the extent and time span for a reduction in CO emissions, this does not mean that the signatory states, in cooperation with industry, will necessarily honor their obligations. Research on implementation concludes that solutions for providing public goods that are redistributive in nature have little chance of being realized, whereas solutions with distributive outcomes, where all stand to win, meet with fewer difficulties. This applies as long as the redistributive solutions do not rely on hierarchy command-and-control that prevents actors who are adversely affected by the decision from trying to change the contents of the decision in the course of implementation, and as long as those who are to benefit from the potential solution do not mobilize in order to exert pressure for a speedy implementation process (Windhoff-Heritier, 1980).

However, since hierarchical guidance and command-and-control can only be used to a limited extent in international governance, effective implementation depends much more on setting incentives which induce actors, over whom there is no hierarchical control, to comply with the agreed solution. Alternatively, it depends on societal mobilization where the actors profiting from the agreement press those who resist implementation into compliance. The same effect may be achieved if an agreed provision is deeply embedded in existing societal value systems. Furthermore, by using reputational mechanisms parties which do not comply are 'named and shamed' into compliance. The noncompliance of a state would result in the loss of reputation.

Outcomes and Impact of the Institutional Provision of Public Goods

A final question relates to the outcomes and impacts of specific institutional provision of a public good at the international level. What are the policy outcomes of such a solution and how

does it affect existing governance structures? The first question addresses the problem-solving capacity of an institutional solution, i.e., the performance of the institutional setting that has been created and the instruments applied in solving the problem. Thus, the question may be raised: does an international environmental agreement help protect endangered natural resources? Since institutional solutions are largely negotiated with a view to their potential distributional outcomes, their de facto distributive effects must be identified. Who gains and who loses on account of a specific institutional solution, and what does this mean in terms of the stability of the chosen solution? The second aspect relates to the fact that specific institutional solutions to public good problems across boundaries affect existing governmental and governance structures. Thus, the formal decision-making powers and the de facto power of national actors are changed by specific institutional solutions to problems.

A case in point is the attempts of the European Union to solve collective action problems in environmental policy that have challenged the competencies of subnational actors in the member states. Similarly, the de facto involvement of Euro associations in policy shaping has implied a loss of political power of national associations. Finally, the question must be raised as to how the overall structure and functions of the traditional state change in an environment where public goods are provided by forms of international governance. Does it imply the demise of the traditional state (Majone, 1996; Grande and Eberlein, 1999) in direction of the regulatory state?

See also: Collective Action; Global Governance.

Bibliography

Bates, R.H., 1988. Contra contractarianism: some reflections on the new institutionalism. Politics and Society 16, 387–401.

Bator, F.M., August 1958. The anatomy of market failure. Quarterly Journal of Economics 72, 351–379.

Buchanan, J.M., Stubblebine, W.C., 1962. Externality. Economica 29, 371–384.

Calvert, R.L., 1986. Models of Imperfect Information in Politics. Harwood Academic, Chur, Switzerland.

Coase, R.H., 1960. The problem of social cost. Journal of Law and Economics 3.

Cornes, R., Sandler, T., 1996. The Theory of Externalities, Public Goods, and Club Goods, second ed. Cambridge University Press, Cambridge, UK.

Eberlein, B., 1999. Der Aufstieg des Regulierungsstaates. Zur Transformation der politischen Okonomie der Bundesrepublik Deutschland [The Rise of the Regulatory State. The Transformation of the Political Economy of the Federal Republic of Germany]. Presentation March 1999 at the Max-Planck-Projektgruppe Recht der Gemeinschaftsgu ter, Bonn, Germany.

Grande, E., 1995. Regieren im verflochtenen Verhandlungssystemen. In: Mayntz, R., Scharpf, F.W. (Eds.), Gesellschaftliche Selbstregelung und politische Steuerung. Campus, Frankfurt, Germany Grande E.

Hardin, R., 1993. The street-level epistemology of trust. Politics and Society 21 (4), 505–529.

Hardin, R., 1999. Liberalism, Constitutionalism, and Democracy. Oxford University Press, Oxford, UK.

Knight, J., 1995. Models, interpretations and theories: constructing explanations of institutional emergence and change. In: Knight, J., Sened, I. (Eds.), Explaining Social Institutions. University of Michigan Press, Ann Arbor, MI, pp. 95–119.

Knight, J., Johnson, J., 1994. Aggregation and deliberation: on the possibility of democratic legitimacy. Political Theory 22, 277–296.

Le Grand, J., 1991. Equity and Choice: An Essay in Economics and Applied Philosophy. Harper Collins Academic, London.

Majone, G., 1996. Regulating Europe. Routledge, London.

March, J.G., Olsen, J.P., 1989. Rediscovering Institutions: The Organizational Basis of Politics. Free Press, New York.

Martin, L.L., 1994. Heterogeneity, linkage and commons problems. Journal of Theoretical Politics 6 (4), 473–493.

Mayntz, R., 1983. Zur Einleitung: Probleme der Theoriebildung in der Implementationsforschung. In: Mayntz, R. (Ed.), Implementation politischer Programme II: Ansatze zur Theoriebildung [The Implementation of Political Programmes II: First Attempts to Develop a Theory]. Westdeutscher Verlag, Opladen, Germany.

Mayntz, R., 1997. Soziologie der offentlichen Verwaltung [Sociology in Public Administration], fourth ed. Mueller Verlag, Heidelberg, Germany.

Mayntz, R., Scharpf, F.W. (Eds.), 1995. Gesellschaftliche Selbstregelung und politische Steuerung [Societal Self-regulation and Political Steering]. Campus, Frankfurt, Germany.

Musgrave, R.A., Musgrave, P.B., 1976. Public Finance in Theory and Practice, second ed. McGraw-Hill, New York.

Ostrom, E., 1990. Governing the Commons. The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge, UK.

Ostrom, E., 2002. Property rights regimes and common goods: a complex Link. In: Héritier, A. (Ed.), Common Goods. Reinventing European and International Governance. Rowman and Littlefield, pp. 29–58.

Pigou, A.C., 1920. The Economics of Welfare. Macmillan, London.

Pressman, J.L., Wildavsky, A.B., 1973. Implementation: How Great Expectations in Washington Are Dashed in Oakland. University of California Press, Berkeley, CA.

Samuelson, P.A., 1954. The pure theory of public expenditure. Review of Economics and Statistics 36, 387–389.

Scharpf, F.W., 1997. Games Real Actors Play: Actor-Centered Institutionalism in Policy Research. Westview Press, Boulder, CO.

Scharpf, F.W., 1999. Regieren in Europa: Effekti und demokratisch [Governing Europe: Effective and Democratic? Oxford University Press, Oxford, UK]. Campus Verlag, Frankfurt, Germany.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Schlager, E., Ostrom, E., 1992. Property rights regimes and natural resources: a conceptual analysis. Land Economics 68 (3), 239–262.

Schmidt, V.A., 1995. The New World Order, Incorporated: The Rise of Business and the Decline of the Nation-State. European University Institute. Working Paper 5.

Snidal, D.J., 1984. Interdependence, Regimes and International Cooperation. University Microfilms International, Anne Arbor, MI.

Taylor, M., 1996. Good government: on hierarchy, social capital and the limitations of rational choice theory. The Journal of Political Philosophy 4 (1), 1–28.

Windhoff-Heritier, A., 1980. Politikimplementation: Ziel und Wirklichkeit politischer Entscheidungen [Policy Implementation: Aims and Reality in Political Decision-Making]. Hain, Ko nigsteinTaunus, Germany.

Public Health as Social Science

Brian Oldenburg and Fiona Cocker, Melbourne School of Population and Global Health, University of Melbourne, VIC, Australia
Benjamin Schüz, School of Psychology, University of Tasmania, TAS, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by B. Oldenburg, volume 18, pp. 12540–12546, © 2001, Elsevier Ltd.

Abstract

Behavioral and social science research and theories are relevant to promoting health and preventing disease in the population. This article discusses the major health challenges confronting developed and developing countries, the contribution of social and behavioral factors to current global health challenges, and current approaches to addressing health problems. Finally, it considers the importance of using social and behavioral theories and models to providing a comprehensive understanding of all the determinants and influences relevant to a particular health issue, and informing the development of efficacious, practitioner-delivered lifestyle–change interventions to address these issues at a public health and societal level.

This article focuses on the extent to which behavioral and social science research and theories are relevant to the challenge of promoting health and preventing disease in the whole population. An overview is provided of the major health challenges confronting developed and developing countries, including a discussion of the contribution made by social and behavioral factors to these global health challenges. It is important to note some key differences between the so-called individual or high-risk approach to addressing health problems, and the whole-of-community or population approach. The 'high-risk' approach, used to develop efficacious, intensive, practitioner-delivered lifestyle–change interventions for individuals, needs to be integrated within a broader public health approach. The latter approach involves a consideration of the whole spectrum of sociobehavioral and environmental factors that are relevant determinants of health to promote health and prevent disease. The importance of the wider dissemination and uptake of interventions at an institutional or policy level is also discussed. Social and behavioral theories and models are very important in public health; they not only help provide a more complete understanding of all of the determinants and influences relevant to a particular health issue, but are also very important for informing how to address the health issue at a public health and societal level.

What Is Public Health?

Public health has been classically defined as

> "… the combination of science, skills and beliefs that are directed to the maintenance and improvement of the health of all people through collective or social actions. The programs, services, and institutions involved emphasize the prevention of disease and the health needs of the population as a whole. Public health activities change with changing technology and values, but the goals remain the same: to reduce the amount of disease, premature death and disability in the population."
>
> Last, 1995

Health was defined by the World Health Organization (WHO) more than 60 years ago as "a state of complete mental, physical, and social well-being, and not merely the absence of disease or infirmity." By this definition, health is more than just a physical state and is inextricably linked to personal well-being and potential, and incorporates many other dimensions of human experience, such as quality of life, psychological health, and social well-being. *Public health* is different from other fields of health and medicine as it strongly emphasizes the promotion of health and well-being and the prevention of disease, illness, and disability at the population level. *Populations* are groups of individuals with some similar characteristic(s) that unify them in some way, such as people living in rural areas, older people, women, an indigenous group, or the population of a specific region or country.

Contemporary public health is particularly concerned with those sections of the population who are under- or poorly served by traditional health services and, consequently, are most disadvantaged and vulnerable in terms of their health. Because such disadvantage or vulnerability usually has its origins in the very early years of childhood and/or is related to an individual or a population subgroup's social and economic circumstances, public health is therefore very concerned with the more upstream or distal (i.e., distant from the person), social, physical, economic, and environmental determinants of a person's health, in addition to the more proximal determinants (i.e., closer to the person), such as lifestyle risk factors, health behaviors, and biological factors.

Preventing any disease is commonly grouped into three levels, named for the stages of disease that they target; primary (pre-disease), secondary (latent disease), and tertiary prevention (symptomatic disease). *Primary prevention* aims to reduce both the incidence and prevalence of a disease before it occurs. For example, encouraging people to protect themselves from ultraviolet radiation or tobacco prevention and health promotion programs which target young people are examples of this level of prevention. *Secondary prevention* (or early detection) refers to the situation after a disease has started to occur, but before the person has noticed anything is wrong. The goal of secondary prevention is to find and treat a disease early – such as cardiovascular disease or diabetes – and in many cases, the disease can then either be cured or the risk of progression can be significantly reduced. *Tertiary prevention* targets people who are already experiencing symptoms of

a disease or diseases and thus aims to reduce the complications associated with poor health and minimizing disability and suffering, such as cardiac rehabilitation programs for patients after a heart attack. There is also a fourth level of prevention, *primordial prevention*, which involves actions and measures to prevent the emergence of risk factors in the form of environmental, economic, social, and cultural patterns of living. For example, many adult health problems, such as obesity and hypertension, have their origins in childhood, because this is the time when the determinants of lifestyle are developing. Subsequently, improving the social and economic circumstances of families during these early years can have a profound influence on later health. Sometimes the boundaries of prevention strategies are indistinct, and it can be common for a given intervention to address a combination of the primary, secondary, and tertiary levels.

The World Health Organization recommends the combined use of population (or mass) and high-risk strategies for prevention. The 'high-risk' strategy aims to provide preventive care to individuals at special risk; hence it requires the detection of individuals with high risk. The *population strategy* is directed at the whole population, or subgroups in the population, irrespective of their individual risk levels. For example, research has demonstrated even a small reduction in the average blood pressure or the serum cholesterol of a population can produce a large reduction in the incidence of cardiovascular disease. Health promotion strategies are often very important for prevention and the WHO definition for health promotion is "the process of enabling people to increase control over the determinants of health and thereby improve their health" (World Health Organisation, 2005). Successful prevention depends upon the following: a knowledge of causation and the determinants; an understanding of the dynamics of transmission; identification of risk factors and risk groups; availability of evidence-based preventive or early detection and treatment measures; an organization and systems for delivering their measures to appropriate persons or groups; and continuous monitoring and evaluation and further development of the strategies that have been implemented.

The application of behavioral and social science theories and methods to prevention can occur at multiple levels and in many different settings in the community. Many public health initiatives are aimed at the individual or interpersonal level, and are delivered through the primary care setting, workplaces, and schools. Such programs typically have a low reach, in that only a limited number of people are impacted on by, and/or have access to the programs. However, these programs can have high levels of exposure for the individuals and can be quite intensive. Other programs can be organized and implemented through whole organizations or systems. Finally, community-wide or societal-level interventions may involve mass or social media, policy, or legislative action at a national or governmental level, and these can have a very large reach. However, program exposure for any given individual is typically quite low.

The practice of public health is therefore very broad and can include the delivery of many different types of programs and services throughout the community. It can include efforts to reduce health disparities, increase access to health services, improve childhood rates of immunization, control the spread of HIV/AIDS, reduce uptake of smoking by young people, and implement legislation and policies to improve the quality of water and food supplies, reduce the occurrence of motor vehicle accidents, and reduce pollution. In fact, the overall health and well-being of the population is more directly related to the effective and widespread implementation of these kinds of efforts, than it is by the society's direct investment in direct health-care delivery.

This article discusses the extent to which theories and models from the behavioral and social sciences can better underpin the understanding and practice of public health. The following sections of this article provide a brief overview of the major public health problems and challenges confronting both developed and developing countries; the contribution made by social and behavioral factors to many of these health problems; and how social and behavioral theories and models can be usefully applied to the development, implementation, and evaluation of public health programs.

Trends in the Global Burden of Disease

The past 100 years have seen substantial change in health and the patterns of disease burden around the world. The findings of the Global Burden of Disease Study 2010 resulted from the collaborative efforts of 486 scientists from 302 institutions in 50 countries and it has made a very important contribution to our understanding of present and future health priorities for specific countries and the global community (Horton, 2012). Expanded in scope since the first study of the global burden of disease completed in 1990, in which 107 diseases and injuries and 10 risk factors were assessed, the GBD 2010 included 235 causes of death and 67 risk factors. This has facilitated the comparison of diseases which cause early death with those that cause prolonged suffering or disability in any country or region of the world.

Disease burden refers to the impact of diseases, injuries, and risk factors as measured by financial cost, mortality, morbidity, and/or other indicators. It is often quantified in terms of quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs), both of which measure the number of years lost due to disease. One DALY can be thought of as 1 year of healthy life lost, and the overall disease burden can be thought of as a measure of the gap between current health status and the ideal health status. Such measures allow for a comparison of difference causes of disease burden, and they can also be used to evaluate the possible impacts of health intervention for similar or different causes. The WHO has now also provided a set of detailed guidelines for measuring disease burden.

The findings GBD 2010 (Horton, 2012) also allowed a comparison of health life expectancy, a summary measure which captures mortality and nonfatal outcomes in a single measure, between countries from 1990 to 2010. Contributions of changes in child mortality, adult mortality, and disability to overall change in population health between 1990 and 2010 were also considered. Findings revealed that in 2010, global healthy life expectancy at birth was 59 years for males and approximately 63 years for females, and across countries in 2010, life expectancy at birth ranged from almost 28 years in Haiti to 71 years in Japan for males, and 37 and 75 years for

females in Haiti and Japan, respectively (Salomon et al., 2012). In the past 20 years healthy life expectancy has increased by 5 years or more for males in 48 countries compared with 43 countries for females. However, healthy life expectancy decreased for males in 22 countries compared with 11 countries for females. Worthy of note is that increases in life expectancy have corresponded with an increase in the number of healthy years lost to disability (YLDs) in most countries. This finding has implications for health planning and health-care expenditure. That is, although substantial progress has been made in terms of the reduction of mortality over the past two decades, comparatively little improvement has been made in reducing the overall effect of nonfatal disease and injury on population health. Of course, this provides a very strong argument for the importance of prevention.

GBD 2010 findings also reveal that although almost 53 million deaths occurred in 2010, up from approximately 46 million deaths in 1990, some very important progress has been made in global health and the health of the majority of countries. For example, life expectancies are increasing for men and women and a greater proportion of deaths are taking place among people aged 70 and older (Horton, 2012). The burdens of HIV and malaria are falling, and fewer children aged 5 years or less are dying (Horton, 2012). The number of deaths attributable to communicable, maternal, neonatal, and nutritional causes were also down from those attributed to these causes in 1990, due largely to decreases in mortality from diarrheal disease (2.5–1.4 million), lower respiratory infections (3.4–2.8 million), neonatal disorders (3.1–2.2 million), measles (0.6–0.13 million), and tetanus (0.3–0.06 million) (Horton, 2012). Furthermore, infectious diseases control is improving in most countries.

Despite this progress, old threats remain and new threats have emerged. In developing countries, major public health problems continue to include infectious diseases, undernutrition, and complications of childbirth, as well as the increasing prevalence of chronic diseases such as cardiovascular disease, cancer, and diabetes. In Sub-Saharan Africa communicable, maternal, neonatal, and nutritional causes still account for 76% of premature mortality in 2010, and newborn, and child mortality, and various vaccine-preventable and other communicable diseases remain urgent concerns (Horton, 2012). Furthermore, the considerable impact of HIV/AIDS on mortality and disability in most developing regions in 2010 is still of concern, as it is now the fifth leading cause of DALYs worldwide (up from 33rd in 1990; a 351% increase) (Murray et al., 2012).

The increasing prevalence of chronic disease is a significant new threat with blood pressure identified as the biggest global risk factor for disease, closely followed by tobacco, alcohol, and poor diet. Deaths attributable to noncommunicable diseases rose from approximately 8 million between 1990 and 2010, to account for two of every three deaths worldwide by 2010. Specifically, 8 million people died from cancer in 2010, 38% more than two decades ago, ischemic heart disease and stroke collectively accounted for one in four deaths worldwide, compared with one in five in 1990, and deaths due to diabetes (1.3 million) were double that in 1990. Deaths due to injuries (5.1 million) were marginally higher in 2010 compared with 1990, a figure driven by a 46% rise in deaths worldwide due to road traffic accidents. In short, ischemic heart disease, stroke, chronic obstructive pulmonary disease, lower respiratory infections, lung cancer, and HIV/AIDS were the leading causes of death in 2010.

Other emerging areas of concern are the increasingly widespread and rising prevalence of diabetes in China and other low- or middle-income countries (Rawal et al., 2012), such as India, Russia, Brazil, Pakistan, Indonesia, and Bangladesh, and the very large unrecognized burden of mental illness in developed and developing countries (Murray et al., 2012). A recent systematic review revealed an increase in diabetes mellitus prevalence from 2.6 to 9.7% in China over the period of 2000–2010 (Li et al., 2012). This finding, combined with the increasing prevalence of undiagnosed, untreated, and suboptimally managed diabetes and prediabetes in China and a population of 1.3 billion people, is cause for concern. In regards to mental illness, results of the GBD 2010 revealed that mental and substance use disorders accounted for approximately 7% of DALYs globally, 8.6 million of all years of life lost (YLL), and are the leading cause of years lived with disability (YLD)s worldwide. The burden of mental and substance use disorders increased by approximately 40% between 1990 and 2010, a figure largely driven by population growth and aging (Whiteford et al., 2013). These findings suggest that mental and substance use disorders pose a significant challenge for health systems in developed and developing regions, and the prevention and treatment of mental and substance use disorders is vital if improvements in population health are to be achieved.

With more people spending more years of their lives with more illnesses, disability from disease and injury is also becoming an increasingly important public health issue. Women are impacted on especially hard by disability with those aged 15–65 years losing more healthy life to disability than men. Although global DALYs remained stable from 1990 to 2010, an important shift occurred in DALY composition with the contribution of deaths and disability among children (aged 5 years and younger) declining from 41% of global DALYs to 25%. YLLs typically account for about half of disease burden in more developed regions (high-income Asia Pacific, Western Europe, high-income North America, and Australasia), rising to over 80% of DALYs in sub-Saharan Africa. Changes in DALYs over the past 20 years were observed for communicable, maternal, neonatal, and nutritional disorders, as well as noncommunicable diseases, and injuries. Furthermore, ischemic heart disease became the leading cause of DALYs worldwide in 2010, up from fourth in 1990, followed by lower respiratory infections (first in 1990), stroke (fifth in 1990), diarrheal diseases (second in 1990), and HIV/AIDS (33rd in 1990). Other significant contributors to global DALYs were major depressive disorder (up 37%) and road injuries (up 34%). In summary, the global disease burden has continued to shift away from communicable to noncommunicable diseases and from premature death to years lived with disability (YLDs). Furthermore, population aging has dramatically changed the global burden of illness worldwide, hastened by improvements in health related to increases in income and reductions in family sizes.

Publications which continually update the GBD estimates are an important source of information for framing health

policies and objectives not just globally but also within specific countries (Murray et al., 2013). For example, the findings of GBD 2010 suggest rising burden chronic conditions and their comorbidities, and the shift away from communicable to noncommunicable diseases and from premature death to years lived with disability (YLDs), particularly in developing or low- and middle-income countries, are significant problems global public health needs to address. In short, disability from disease and injury has become an increasingly important consideration for all health systems as more people will be spending more years of their lives with more illnesses, and suggests a need to focus on the growing burden of noncommunicable diseases and to make a shift from preventing premature death to reducing years lived with chronic illnesses and debilitating conditions (Murray et al., 2012). However, the small evidence base from developing countries is underdeveloped which serves to highlight the need for appropriate cultural adaptation/translation of programs based on social and behavioral models and theories, from more developed settings.

Theories and Models

Understanding the alterable causes of the burden of disease can help inform responses to these challenges nationally, regionally, and internationally, as well as within population subgroups. There are numerous factors which explain the changing patterns of morbidity, mortality, and the spread of disease, globally and between regions and countries, relating to population growth, aging and changes to family, social structures, education and reproductive patterns, and changes in ecological and living environments. An evidence base for these multilevel determinants of health has been extensively developed over the past decade (Roux, 2008), in particular, particularly the important relationships between behavior and health. Fisher et al. (2011) have recently reviewed the central role that behavior has in health and the importance of behavioral interventions to prevent disease, improve management of existing disease, increase quality of life, and reduce health-care costs.

Mokdad et al. (2004) demonstrated the relative contribution of alterable health behaviors and social factors to mortality and morbidity in the United States. They concluded that the primary causes of death in 2000 were tobacco and alcohol use, poor diet, and physical inactivity (World Health Organisation, 2005); however, the conditions of the society were also very important as more fundamental causes of health. More recently, the United States Centers for Disease Control and Prevention released a key report estimating that 14 of the top 15 causes of death in the United States are attributable to modifiable behaviors, including physical inactivity, diet, smoking, alcohol and other drug use, injury control, sun protective behaviors, appropriate use of medicines, immunization, sexual and reproductive health, oral hygiene, and mental health (Heron et al., 2009). An increased understanding of the social, economic, and environmental influences on behaviors leading to health (and ill health) from an individual to a neighborhood/community level and to a global level is consistent with a socioecological model of health that can also be used to understand the rapid increase in chronic NCDs reflected by tobacco, poor diet, physical inactivity, and sedentariness. These findings from a considerable amount of social and behavioral epidemiologic research conducted over the past 40 years demonstrate that a substantial part of the global disease burden is attributable to these multiple levels of influences from the society where we live. This underlines the importance of gaining more of an understanding of the social, environmental, and other determinants of human behavior and how these can be influenced and changed. Theories and models are very important for this improved understanding.

Theories and models from the social and behavioral sciences not only help to explain health-related behaviors and their downstream and more upstream determinants, they can also help guide the development of interventions to influence and change both the determinants of and health-related behaviors themselves and, thereby, improve health of individuals and populations. Theories can also help understand adherence to a health regimen, identify the kinds of information and other details needed to develop effective intervention programs, provide guidelines on effective communication, and identify barriers to be overcome in implementing an intervention.

This article adopts a widespread definition of theory (Glanz et al., 2008), by which a theory is seen as a set of related concepts, definitions, and assumptions about relations between concepts that allow understanding events or situations in a systematic way. The most important notions perhaps are that theories can be applied to multiple problems and they should also be testable, which means that we can generate and examine data that tells us something about how well a theory is suited to explain a specific public health problem. This means that we should select theories (or elements of theories) on the basis of the evidence of such tests of how well they explain and predict a specific problem rather than their face validity or what might appear plausible. Theories can be useful for understanding the determinants of behavior and how to improve health at multiple levels, that is, at an individual, interpersonal, institutional/organizational, community, and/or societal level.

Theories can also be used as basis for public health measures to specific populations to increase effectiveness. For example, target group-specific theories such as the Gelberg–Andersen behavioral model for vulnerable populations (Gelberg et al., 2000) explicitly include frequent behavioral problems in vulnerable populations as predisposing variables that impact on health-care utilization. On the other hand, demographic characteristics (e.g., age, sex, cultural background, behavioral characteristics) of the target group can be taken into account in identifying the most promising theory-based methods.

Furthermore, an important distinction can be made between theories that explain behavior (e.g., 'The Theory of Planned Behavior,' Ajzen, 1991) and theories that explain change in individuals and organizations (e.g., 'Diffusion of Innovations,' theory Oldenburg and Glanz, 2008). The former are useful in outlining factors that are for example responsible for engaging in a specific behavior, the latter are useful in identifying factors that promote change.

On the level of individual behavior, a range of theories have been examined in the public health context. These include theories such as the Health Belief Model, the Theory of Planned

Behavior, interpersonal theories such as Social Cognitive Theory, stage theories such as the Transtheoretical Model, or hybrid theories such as the Health Action Process Approach. These theories outline variables on the level of the individual that explain individual health-related behavior. For example, a very parsimonious approach, the Health Belief Model, proposes the more susceptible someone feels to a health problem, the more severe they see this health problem, and the more advantages as well as the less costs of a protective behavior they perceive, the more likely they are to engage in this protective behavior. Theories such as the Health Action Process Approach (Schwarzer and Luszczynska, 2008) are more sophisticated, include several stages of behavioral preparedness with different psychosocial influence factors, and assume different phases of behavior initiation, maintenance, and recovery. Practical approaches go beyond the application of one specific model and move toward the identification of effective behavior change techniques.

However, in public health, the scope of intervention usually goes beyond the individual and therefore needs to be based on theories that go beyond the individual. For example, interventions that target health at the workplace need to account for workplace factors that interact with individual health such as an imbalance between rewards and efforts (Siegrist, 1996) or workplace demands and control (Karasek, 1979). More integrative theories such as the Reserve Capacity Model (Gallo, 2009) can help to understand how socioeconomic factors interact with individual factors in determining health. There have also been recent efforts to integrate environmental factors with individual predictors of health behaviors, for example, the interactions between environmental factors such as district-level resources and attitudes, or self-efficacy and the built environment (Carlson et al., 2012) in predicting health behaviors such as physical activity.

Theories such as Social Cognitive Theory (for an overview, see McAlister et al., 2008) assume that health-related behavior is both shaped by the individual and the social environment within which individuals live. By observing others' behavior and the results of others' behaviors, we shape our own expectations, make learning experiences, and determine our capability to change health-related behavior (social learning). Theories that operate on this level of influence on health can help to understand and (potentially) modify the effects of interpersonal networks such as families, friends, or peers on behavior.

Finally, on the level of organizations and society more broadly, theories about the functioning of social systems consider factors that impact on change of institutions, organizations, or communities. For example, the Diffusions of Innovations Theory (Oldenburg and Glanz, 2008) proposes that the degree to which innovations (such as health initiatives) permeate a social group is determined by the communication channels available, the time since the innovation and properties of the social system. Understanding these parameters can help to identify access points for interventions to achieve wide-reaching distribution.

From a practical viewpoint, however, the interaction between environmental, social, and individual influences on behavior can be very complex, so that more integrative models or frameworks that operationalize health promotion and prevention as a process through multiple levels, from individual to policy level, are often employed as tools to develop, implement, and evaluate such programs.

One such framework, the PRECEDE–PROCEED model (Green and Kreuter, 2005) outlines eight steps that allow mapping a series of measures to promote health on different levels – from the individual to policy makers. The model includes a detailed needs assessment of the health problem on a social, an epidemiological, behavioral and environmental, an educational and ecological, and an administrative and policy level. More importantly, this model also includes a framework for evaluating the impact of interventions on these different levels and thus provides an integrative framework that allows including theories for individual and organizational change on different levels. Furthermore, Intervention Mapping (Bartholomew et al., 2011) has also been developed as a very comprehensive model that identifies theory as well as theory- and evidence-based methods to promote behavior change that can be integrated into the PRECEDE–PROCEED framework. It involves the specification of a series of matrices outlining the change objectives and determinants of change on various levels (e.g., determinants of individual increases of physical activity in a specific target population need to be linked with the determinants of an organization adopting a workplace health promotion intervention based on these individual determinants of physical activity). Such integrative and multilevel models are increasingly used as comprehensive frameworks to understand the determinants of a particular health issue or behavior and to develop effective public health interventions.

Implementation and Wider Uptake of the Social and Behavioral Science Evidence Base to Improve Public Health

With the development, implementation, and evaluation of a series of major community-based cardiovascular prevention trials during the 1970s and 1980s, widespread use was made of multilevel intervention or change strategies underpinned by behavioral and social science theories and research for the first time. Importantly, these strategies did not focus exclusively on change in the individual. Instead, preventive strategies using multiple approaches were directed at the media, legislation, and the use of restrictive policies, and involved specific settings such as schools, the workplace, health-care settings, and other key groups in the community, with the aim of reducing population rates of risk behaviors, morbidity, and mortality.

The North Karelia Project in Finland (Puska et al., 1985) was initiated in response to concern that the community, at that time, had the highest risk of heart attack worldwide. Strategies included tobacco taxation and restriction, televised instruction in skills for nonsmoking and vegetable growing, and extensive organization and networking to build an education and advocacy organization. After 10 years, results indicated significant reductions in smoking, blood pressure, and cholesterol, and a 24% reduction in coronary heart disease mortality among middle-aged males. The Stanford Three Community Study (Maccoby et al., 1977) demonstrated the feasibility and effectiveness of mass media-based educational campaigns and achieved significant reductions in cholesterol

and fat intake. The Minnesota Heart Health Project (Mittelmark et al., 1986) used interventions aimed at raising public awareness of risk factors for coronary heart disease, and changing risk behaviors through education of health professionals and environmental change programs, such as grocery store and restaurant labeling.

Worksite intervention trials, such as the Working Well Trial (Sorensen et al., 1996), also utilized interventions aimed at favorably altering social and psychological factors known to influence health-related behaviors, including increased awareness of benefits, increased self-efficacy, increased social support, and reduced perception of barriers. Strategies in this setting can include promotion of nonsmoking policies, increasing the availability of healthy foods in canteens and vending machines, group education sessions, and access to exercise facilities. School-based interventions provide an established setting for reaching children, adolescents, and their families to address age-pertinent health issues such as smoking, teenage pregnancy, and substance use. Strategies have targeted social reinforcement, school food service, health-related knowledge and attitudes, local ordinances, peer leaders, and social norms.

In evaluating the effectiveness of such trials, Rose (1992) states that it is important to acknowledge the differences between clinical and more population-based approaches. Clinical approaches, while more likely to result in greater mean changes in individuals, are less significant at a population level because fewer individuals are involved. In contrast, the small mean changes achieved by population-based approaches, if applied to a large number of people, have the potential to impact much more significantly on population rates of disease and health. Furthermore, one of the difficulties in evaluating such population-based approaches is the length of time required to demonstrate changes in population rates of health outcomes. While researchers can readily measure changes in a behavioral outcome – such as the rate of smoking cessation – following a widespread and well-conducted program, the time required for such behavioral changes to impact upon long-term morbidity and mortality – such as lung cancer – is much more problematic. Nevertheless, many of the above-identified behavior-based trials demonstrate the value of social and behavioral theories and the feasibility of activating communities in the pursuit of health.

The ultimate value of this research must be determined by the extent to which the knowledge generated is widely disseminated, adopted, implemented, and maintained by 'users' and, ultimately, by its impact on systems and policy at a regional, state, and/or national level (i.e., institutionalization). While considerable effort and resources are devoted to developing effective interventions, relatively little attention has been given to developing and researching effective methods for the *diffusion and implementation* of their use. It is important to acknowledge that the availability of relevant research findings does not in itself guarantee good practice. Such knowledge transfer involves the development of formal research policies, formalized organizational structural support, appropriate and targeted funding, formal monitoring of activity, and ongoing training.

In essence, achieving the effective diffusion of innovations, both within the general community and in organizational settings, involves change, and the change principles which underpin the diffusion process are not so different to those previously identified for understanding change at the individual, organizational, or community levels. At the level of the individual, family, or small group, uptake of a health promotion innovation typically involves changes in behaviors or lifestyle practices which will either reduce risk factors or promote health. At an organizational level, such as the workplace, school, or the health-care setting, successful uptake of an innovation may require the introduction of particular programs or services, changes in policies or regulations, or changes in the roles and functions of particular personnel. At a broader community-wide or even societal level, the change process can involve the use of the media and changes in governmental policies and legislation, as well as coordination of a variety of other initiatives at the individual and the settings level.

However, in considering the principles that underpin the diffusion of innovations at a population level, further complexity arises from the need to consider change occurring at multiple levels, across many different settings, and resulting from the use of many different change strategies. This requires the application of multiple models and theories to develop frameworks with sufficient explanatory power.

Looking to the Future

It is difficult to be certain about the public health challenges that lie ahead in the twenty-first century. It is clear that in all countries, there is still a need to reduce preventable disease and premature death, reduce health inequalities, and develop sustainable public health programs. Much of the social and behavioral science research conducted with individuals is still relevant in the field of public health, as are many of the models and theories developed and tested over the past 40 years. To move to a population-based focus, we need innovative means of implementing such knowledge, and methodologies so as to increase the reach of such programs to more people, particularly the most disadvantaged individuals and sections of society.

Adopting preventive approaches to improve health, acknowledging and acting on the importance of social and environmental factors as major determinants of population health, and reducing health inequalities provides a substantial challenge to public health now and in the future. Furthermore, limited resources place an emphasis on evaluating the affordability and sustainability of public health interventions. Public health should be at the forefront of developing and researching effective intervention methods and strategies for current and future health challenges, with global sharing of information in terms of monitoring health outcomes and development of the evidence base. However, this research should not focus solely on developing strategies and testing their efficacy and effectiveness. There is an important need for research that examines larger scale implementation and diffusion of these strategies to whole communities and populations, through the innovative use of a variety of psychological, social, and environmental intervention methods, including evidence from the social and behavioral sciences.

See also: Globalization and Health; Health Education and Health Promotion; Health Interventions: Community-based; Health Literacy; Health Promotion in Schools; Life Course Epidemiology; Psychosocial Theories of Health Behavior; Social Epidemiology.

Bibliography

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Bartholomew, L.K., et al., 2011. Planning Health Promotion: An Intervention Mapping Approach. Jossey-Bass, San Francisco, CA.

Carlson, J.A., et al., 2012. Interactions between psychosocial and built environment factors in explaining older adults' physical activity. Preventive Medicine 54, 68–73.

Fisher, E.B., et al., 2011. Behavior matters. American Journal of Preventive Medicine 40, e15–e30.

Gallo, L.C., 2009. The reserve capacity model as a framework for understanding psychosocial factors in health disparities. Applied Psychology: Health and Well-being 1, 62–72.

Gelberg, L., et al., 2000. The behavioral model for vulnerable populations: application to medical care use and outcomes for homeless people. Health Services Research 34, 1273–1302.

Glanz, K., et al., 2008. Theory, research, and practice in health behavior and health education. In: Glanz, K., et al. (Eds.), Health Behavior and Health Education. Jossey-Bass, San Francisco.

Green, L.W., Kreuter, M.W., 2005. Health Program Planning: An Educational and Ecological Approach. McGraw-Hill, New York.

Heron, M., et al., 2009. Deaths: final data for 2006. National Vital Statistics Reports 57 (14). Centers for Disease Control and Prevention, Hyattsville, MD.

Horton, R., 2012. GBD 2010: understanding disease, injury, and risk. The Lancet 380, 2053–2054.

Karasek Jr., R.A., 1979. Job demands, job decision latitude, and mental strain: implications for job redesign. Administrative Science Quarterly 24, 285–308.

Last, J., 1995. Dictionary of Epidemiology. Oxford University Press, New York.

Li, H., et al., 2012. Diabetes prevalence and determinants in adults in China mainland from 2000 to 2010: a systematic review. Diabetes Research and Clinical Practice 98, 226–235.

Maccoby, N., et al., 1977. Reducing the risk of cardiovascular disease. Journal of Community Health 3, 100–114.

McAlister, A.L., et al., 2008. How individuals, environments, and health behaviors interact: social cognitive theory. In: Glanz, K., et al. (Eds.), Health Behavior and Health Education. Jossey-Bass, San Francisco, pp. 169–188.

Mittelmark, M.B., et al., 1986. Community-wide prevention of cardiovascular disease: education strategies of the Minnesota Heart Health Program. Preventive Medicine 15, 1–17.

Mokdad, A.H., et al., 2004. Actual causes of death in the united states, 2000. Journal of the American Medical Association 291, 1238–1245.

Murray, C.J., et al., 2012. GBD 2010: a multi-investigator collaboration for global comparative descriptive epidemiology. The Lancet 380, 2055–2058.

Murray, C.J.L., et al., 2013. UK health performance: findings of the Global Burden of Disease Study 2010. Lancet 381, 997–1020.

Oldenburg, B., Glanz, K., 2008. Diffusion of innovations. In: Glanz, K., et al. (Eds.), Theory, Research, and Practice in Health Behavior and Health Education. Jossey-Bass, San Francisco.

Puska, P., et al., 1985. The community-based strategy to prevent coronary heart disease: conclusions from the ten years of the North Karelia project. Annual Review of Public Health 6, 147–193.

Rawal, L.B., et al., 2012. Prevention of type 2 diabetes and its complications in developing countries: a review. International Journal of Behavioural Medicine 19, 121–133.

Rose, G., 1992. The Strategy of Preventive Medicine. Oxford University Press, Oxford, UK.

Roux, A.V.D., 2008. Next steps in understanding the multilevel determinants of health. Journal of Epidemiology and Community Health 62, 957–959.

Salomon, J.A., et al., 2012. Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the Global Burden Disease Study 2010. The Lancet 380, 2144–2162.

Schwarzer, R., Luszczynska, A., 2008. How to overcome health-compromising behaviors. European Psychologist 13, 141–151.

Siegrist, J., 1996. Adverse health effects of high-effort/low-reward conditions. Journal of Occupational Health Psychology 1, 27–41.

Sorensen, G., et al., 1996. Work site-based cancer prevention: primary results from the Working Well Trial. American Journal of Public Health 86, 939–947.

Whiteford, H.A., et al., 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet.

World Health Organisation. The Bangkok Charter for Health Promotion in a Globalized World. 6th global Conference on Health Promotion, 2005, Bangkok, Thailand.

Public Interest

Mark Blitz, Claremont McKenna College, Claremont, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The question of public interest must be considered under four headings. First is the formal or conceptual question. This includes problems such as the meaning of public as opposed to private, self, and common interest, and the meaning of interests as distinguished from other goods and qualities. Second is the question of what substantively the public interest is. Third is the issue of how the public interest can be used as a guide for specific actions and concretely implemented. Fourth is the issue of the degree to which voters, politicians, and bureaucrats actually seek in their actions to serve public rather than private concerns.

Scholars debate the question of the public interest somewhat less than they once did, but it remains an important issue, even if "much maligned and dismissed" (King et al., 2010). As is true of other central concepts, the passage of time does not always lead to an increase in acuity, although it does lead to changes in emphasis (Friedrich, 1962; Lewin, 1991). Much that might be considered under the rubric of the public interest, moreover, continues to be discussed in terms of associated ideas such as justice, responsibility, and community (Blitz, 2005; King et al., 2010). This ambiguity complicates a topic that is itself elusive.

What Is the Public Interest, Conceptually?

The issue of the public interest is best organized under four headings. First is the formal or conceptual question. This includes problems such as the meaning of public as opposed to private, self, or common interest, and the meaning of interests as distinguished from other goods. 'Public' sometimes is taken as anything that is not private, so that regard for the public interest means any regard for interests or groups that are not strictly or exclusively one's own (Lewin, 1991). Less broadly, a public often is said to be those affected by private groups, in the sense that people who are neither labor nor management are affected by strikes. Such publics, moreover, sometimes are taken to refer to those potentially and not just actually affected as, say, any potential passenger might be inconvenienced by a strike of air traffic controllers.

Someone can, therefore, have both a private and a public interest in the same matter, and 'self' interest need not be identical to private interest. In this sense we are all members of many publics (Dewey, 1927). Often, however, the public interest is taken to be what is in the interests of 'all' of us, presumably as citizens within some boundary, although the questions of scope and authority are often curiously undiscussed. There are not many publics in this view but, finally, one public (Friedrich, 1962; Barth, 1992). The public is the group that makes law. Of course, its interests may be the sum of those of the many affected publics into which it can be divided.

Many discussions of the 'public' in these terms have in mind economic matters, or the point of view of academic economics (Friedrich, 1962). A public is seen as an affected or lawmaking group of individuals that seeks satisfaction in economically calculable terms. Publics want things that market failures cause them to receive inadequately. It is serving such publics that defines the responsibilities of professionals and of businesses 'affected with the public interest.' One strength of this conception is that since the 1950s it has brought to bear an increasingly sophisticated understanding of economically rational action, both strategic and straightforward.

There are, however, limits to this perspective. They are evident in the narrowness of the economic view of the public when it is compared to other notions of what is potentially common in common goods. For, what is common is not limited to what each independently and equally can use or enjoy, as we use markets. The common also can involve things of which we are only parts, sometimes equal or identical and sometimes not, as a face is common to its different parts or a team or family common to its members. The ties that characterize what is common, moreover, are not limited to what covers or blankets us universally, but include attaching, placing, fitting, ordering, attracting, and other links integral to images of politics from Plato through Hegel. Practical and economically based notions of the public interest will, of course, often take these distinctions for granted, but deeper exploration of the concept of the public interest cannot.

Just as 'public' is often narrower than common, so too are 'interests' conceptually different from goods more broadly conceived. Discussions of the public interest often refer to Adam Smith and, therefore, to a view of what is good as what satisfies desire in the somewhat flat, calculable, and exchangeable manner of post-Hobbesian or Lockean bourgeois life. Interests are linked to what is merely interesting or desired, not to goods as objects of passion, longing, pride, or nobility or to the character and institutions that support and direct these. Public interests are what several of us happen to desire or means to satisfaction that we generally want. Part of the conceptual task in understanding the public interest is to consider the different ways of experiencing goods so that we can grasp how interests are goods conceived in a limited manner that is linked to the premises of liberal individualism.

What Is the Public Interest, Substantively?

A second question about the public interest is what it is substantively. This issue involves both scope and content. Some scholars seek to limit questions of the public interest

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93112-X

only to certain kinds of public goods. Others use it to stand for national goals, purposes, or goods more comprehensively (Long, 1991). In practice, too, speaking of the public interest generally differs from considering the public's interest in assuring the provision to all of some specific good such as telephone service.

When the scope of the public interest is limited to economic concerns, so too is the content of what is said to comprise it; for some the public interest does not even include protecting rights. To limit the concept in this way risks confusion when 'public interest' retains its broad use in practice and even in theory; to use it broadly, however, seems to involve every question of human purpose and to require some other concept to take the place of the idea used narrowly.

Just as our understanding of the public interest conceived narrowly has benefited from developments in economic understanding, so too has substantive discussion about the public interest conceived broadly undergone notable developments since the 1950s. The triumph of liberal democracies in the Cold War has directed the concerns of the left away from socialism simply to expanding the place of egalitarianism and community within liberal democracy itself. A wish among those who reject the revivified concept of natural rights to nonetheless ground the place of rights and defend them from simple utilitarian calculations has become widespread. Concern with the place of virtue and nobility in liberal democracies and, indeed, concern with these phenomena simply has been renewed (Rawls, 1971; Strauss, 1952). All these trends have enriched and re-energized discussions of the public interest.

Can the Public Interest Be Implemented?

The third question is how the public interest might be implemented concretely and used as a guide for action. Some thinkers argue that the 'public interest' is too vague to be used other than rhetorically or that it is merely a name that sums up a country's political and economic life at some moment frozen in analytical time (Friedrich, 1962). More charitably, one might concede that it means a certain level of economic satisfaction and political freedom, but that these are best reached through private and political competition, that is, when no one aims at them directly.

When we conceive the public interest broadly, however, it is not always so difficult to discern how it might guide action. Constitutions, for example, cannot be formed well without explicit attention to more than private concerns: Charles Beard's worm's eye view of the framing of the American Constitution has been debunked by more thoughtfully elevated judgments (Kesler, 2000; Mansfield, 1991). Once constitutionally secured, individual rights and liberties must be exercised individually, but steps to protect these rights publicly must be attended to explicitly by jurists, legislators, and executives. National defense is a public interest that guides specific legislative and executive action.

To say that the public good often receives explicit attention is not to say that it goes without debate, but to call it debatable is not to call it meaningless. The deliberations and result of the US constitutional convention prove the contrary. Similarly, if we take our private interest to include moral and intellectual excellence and not only economic opportunity and satisfaction we would not say that the disputability of these ends makes them pointless. On the contrary, their subtlety may make our attention to them all the more intense.

If we conceive the public interest more narrowly as assuring some equitable or safe provision and distribution of certain goods then it is again clear that the public interest can be a concrete guide. Are expanding access to communications and increasing public health served better by explicit government regulation with a defined public interest in mind of such and such a level of drug safety, health care, and access to the internet, or are they better served by discovery and application fueled largely by private desires for wealth, knowledge, and mastery, with little regulatory adjustment of the private application of resources? Is improved education secured better by unfunded parental choice, by funded choice, or by public provision or standards? Explicit attention to health, education, and responsible choice as public interests seems necessary to help secure their sufficient provision and fair distribution, if only to ward off excessive regulation and allow private efforts to flourish.

Is the Public Interest Followed?

The fourth question is whether anyone in practice pays attention to the public interest or whether private interest in fact dominates even in public matters. It is argued that the actions of voters, representatives, and bureaucrats all can be explained as attempts to enhance private goods; this analysis can be expanded to professionals, journalists, and intellectuals who do public talking, if not public business, for a living. Even if the public interest could guide us, in this view it does not.

Some research in the US and Western Europe indicates that voters do not simply or usually vote their immediately selfish economic concerns, that the behavior of representatives cannot always or primarily be explained by their wish for re-election, that bureaucrats do not always want to increase their agency's size, and that groups and individuals sometimes do seek benefits for publics from which they and their members do not receive special advantages (Lewin, 1991; Wilson, 1989). Of course, a link between private and public interest is to be expected and largely desired in liberal democratic regimes, for the responsibility we encourage in citizens and officials is meant to benefit not just others but oneself as well (Kesler, 2000).

Obviously, whether political action actually attempts to serve public interests requires continued study. Only the excessively naive would believe that unbridled devotion to a clearly conceived common good directs each and every political choice. But only the false or foolish sophisticate would believe the public interest to be as fleeting and airy as the summer snow.

See also: Interest: History of the Concept; Political Representation; Political Sociology; Public Interest Law: The United States and Beyond; Public Reason.

Bibliography

Antitrust and the economics of the public interest 1997. Symposium: Economics of Antitrust Enforcement Antitrust Bulletin 42 (Spring).

Barth, T.J., 1992. The public interest and administrative discretion. American Review of Public Administration 22, 289–300.

Blitz, M., 2005. Duty Bound: Responsibility and American Public Life. Rowman and Littlefield, Lanham, MD.

Box, R.C., 1992. The administrator as trustee of the public interest: normative ideals and daily practice. Administration & Society 24, 323–345.

Dewey, J., 1927. The Public and Its Problems. Holt, New York.

Friedrich, C.J. (Ed.), 1962. The Public Interest. Atherton, New York.

Goodin, R.E., 1996. Institutionalizing the public interest: the defense of deadlock and beyond. American Political Science Review 90, 331–343.

Kesler, C. (Ed.), 2000. Alexander Hamilton, John Jay and James Madison, the Federalist Papers. Random House, New York.

King, S.M., Chilton, B.S., Roberts, G.E., 2010. Reflections on defining the public interest. Administration and Society 41, 954–978.

Lewin, L., 1991. Self Interest and Public Interest in Western Politics (D. Lavery, Trans.). Oxford University Press, Oxford, UK.

Long, N.E., December 1991. Politics, political science and the public interest. Political Science and Politics 24, 670–675.

Mansfield Jr., H.C., 1991. America's Constitutional Soul. Johns Hopkins University Press, Baltimore, MD.

Rawls, J., 1971. A Theory of Justice. Belknap Press of Harvard University Press, Cambridge, MA.

Rohr, J.A., 2002. Civil Servants and Their Constitutions. University Press of Kansas, Lawrence.

Strauss, L., 1952. Natural Right and History. University of Chicago Press, Chicago.

Wilson, J.Q., 1989. Bureaucracy. Basic Books, New York.

Public Interest Law: The United States and Beyond

Scott L Cummings, UCLA School of Law, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Halpern, volume 18, pp. 12549–12553, © 2001, Elsevier Ltd.

Abstract

Public interest law is a branch of legal practice committed to representing traditionally underrepresented groups and interests. It has historically been associated with test case litigation designed to not merely affect the outcome of a particular dispute, but to influence public policy. Yet it has also been – and perhaps is now increasingly – focused on a broad range of legal advocacy: some before government administrative agencies and legislative bodies, and others in connection with grassroots organizing, economic development initiatives, and social movement campaigns. The defining feature of public interest law is that, unlike individual legal representation, it focuses on reshaping public policy and affects the interests of a substantial population.

Introduction

Public interest law is a branch of legal practice committed to representing traditionally unrepresented groups and interests. It went through a period of rapid growth in the United States in the 1970s and in the early decades of the twenty-first century is an important, though small component of the legal system in the US and many other countries. Public interest law has traditionally been associated with test case litigation designed to not merely affect the outcome of a particular dispute, but to influence public policy. Yet it has also been – and perhaps is now increasingly – focused on a broad range of legal advocacy: some before government administrative agencies and legislative bodies, and others in connection with grassroots organizing, economic development initiatives, and social movement campaigns.

The defining feature of public interest law is that, unlike individual legal representation, it focuses on reshaping public policy and affects the interests of a substantial population. The class action device has at times been used in such cases, although similar goals have been advanced by lawyers representing an organization or, in appropriate cases, even by representation of a single individual on a matter where substantial issues of legal principle are involved. In the classic public interest model, a lawsuit in which a doctrinal innovation is made is followed by strategic efforts to elaborate, or expand, the new principles. In addition, public interest law has sought to use legal advocacy to build policy change through legislative advocacy and to change public attitudes about social problems, as the recent campaign for same-sex marriage in the United States has illustrated (Cummings and NeJaime, 2010).

During the last third of the twentieth century, there were significant shifts in public policy in the United States in critical areas such as race relations, environmental protection, and women's rights. In each of these areas, public interest lawyers played an important role in tandem with citizen organizing and political action. When legislatures and other agencies of government were closed to citizen advocates, courts were often used to advance causes. Court decisions in public interest cases often precipitated action in administrative agencies, legislatures, and the executive branch, which would otherwise not have come about.

Antecedents

Although the name public interest law was not coined until 1969, its antecedents existed as a set of practices for much of the twentieth century. Historically, public interest law in the United States is associated with several movements, some of which were concerned with broad societal issues, others of which were more concerned with the individual's access to basic legal representation.

One important precursor of public interest law was the decades-long campaign to redress racial discrimination and racist institutional structures. In this arena, the National Association for the Advancement of Colored People's Legal Defense and Education Fund (LDEF) was the paradigm. The LDEF was established in the mid-1930s, and a brilliant group of civil rights lawyers led by Thurgood Marshall (later a Supreme Court justice) devised a strategy to bring a succession of test cases through the federal courts to the Supreme Court that would gradually undermine the legal foundations for racial segregation. The sequence of cases led to the landmark decision in 1954 of *Brown v. Board of Education*, which ended formal, legally mandated segregation in American public schools. In subsequent decades, a series of lawsuits followed to apply the principles of *Brown* in specific situations and to carry them forward into other areas where racial discrimination was practiced such as employment and housing.

Another important pillar upon which public interest law rests is the movement for freedom of expression advanced by the American Civil Liberties Union, which grew out of the National Civil Liberties Bureau's defense of antiwar advocates during the years of World War I. Also helping to build public interest law was the consumer movement, embodied in various organizations founded by Ralph Nader, and tracing its lineage back to Supreme Court Justice Louis Brandeis and progressive-era antitrust legislation. The start of the twentieth century saw the emergence of the idea that the American government had

a right, and often a duty, to regulate the marketplace and ameliorate some of the injustices of a *laissez-faire* economic structure. Brandeis developed the idea that a 'private attorney general' could go to court to enforce the law if government failed to act. This idea has remained a foundational element of contemporary public interest practice. The legal aid movement, which began in the late nineteenth century in New York as a way of providing charity-based legal representation for people too poor to pay private attorneys, also influenced public interest law. Legal aid expanded in the post–World War I period in the wake of Reginald Heber Smith's 1919 book *Justice and the Poor*. In the 1960s, as part of President Lyndon Johnson's War on Poverty, a major federally funded legal services program dramatically expanded and strengthened legal aid. With the expansion of federal antipoverty programs in the 1960s, many lawyers within legal services developed public interest law strategies and highly specialized knowledge, litigating issues such as welfare policy, health care for the poor, and the availability of subsidized housing. They used class actions and other litigation strategies to achieve important policy shifts.

Since the 1960s, public interest law has broadened its scope to include such areas as environmental law, in which public interest advocacy groups have built a framework of effective judicial precedents, federal statutes, and regulations. In the 1970s, public interest groups litigated suits that delayed the construction of the Trans-Alaska oil pipeline and forced its redesign; challenged, and eventually halted, the domestic use of the pesticide DDT; and won stricter oversight of nuclear power plants. In the following decades, public interest law organizations specializing in environmental issues have developed into large membership organizations with increasingly sophisticated technical expertise, and a range of coordinated strategies involving litigation, administrative advocacy, grassroots organizing, political advocacy, and public relations. Their efforts have been joined by other groups addressing a broad range of issues, such as gender-based and sexual orientation-based discrimination, and serving a wide range of constituencies, including immigrants, people with disabilities, children, prisoners, and many others.

The Public Interest Law Firm

Sarat and Scheingold (1998) describe public interest law as using 'legal skills to further a vision of the good society'. Most legal work is, in theory, value-neutral, and the attorney representing a fee-paying client without regard to the merits of his claim. Public interest law is, at its core, a systematic effort to bring new voices into policy formulation, and to use the law to attain certain public policy objectives, such as the abolition of legally authorized segregation or the implementation of better environmental protection. Lawyers accept lower levels of remuneration in exchange for having the freedom to pursue social goals and to represent only clients whose causes they believe have genuine merit.

Many public interest lawyers in the United States work in nonprofit public interest law firms, granted tax-exempt status by the Internal Revenue Service on the condition that they are paid a salary by the firm instead of fees by their clients, and that they take on cases broadly defined as having implications for the public good. The firms' efforts are supplemented by the work of lawyers in private firms who are willing to take on a limited number of *pro bono* cases, and by law students who receive academic credit for their work in law school clinics.

The Internal Revenue Service's recognition of the charitable status of public interest law firms has allowed foundations to channel money to these groups. Philanthropic organizations such as the Ford Foundation, the Rockefeller Brothers Fund, and the Edna McConnell Clark Foundation donated millions of dollars to public interest law organizations. Foundations were particularly important in providing seed money to the public interest movement in the 1970s. Subsequently, many of them became reluctant to fund public interest law activities, especially litigation, because of the controversy such activities engender. Nonetheless, many progressive foundations have continued to support regional organizations to provide funds for local and regional public interest law activities.

The first of the new generation of public interest law firms, and one representative of the field, is the Center for Law and Social Policy (CLASP), established in 1969. It was founded by four attorneys (including one of the authors) experienced in the distinctive legal culture of Washington, DC. They had backgrounds in corporate law practice and the executive and legislative branches of the federal government. They believed that effective legal advocacy on behalf of corporations often subverted the mechanisms of government; that the process was impoverished by the absence of effective advocacy on behalf of consumers, environmentalists, the poor, the victims of discrimination, and others' interests. They organized a bipartisan group of eminent bar members to serve as a board of directors for CLASP. Within a year, CLASP had won landmark victories in two major environmental lawsuits. Early CLASP clients included Ralph Nader, the Wilderness Society, and the American Public Health Association.

Historically, public interest law organizations have focused on issues that can be broadly defined as progressive or liberal, acting as a counterbalancing force to corporate power and serving the needs of the poor and powerless. Public interest lawyers frequently provided a counterweight to corporate advocacy. As such, it was essential that they render top-flight legal services and that they be backed up by the specialized technical expertise that was vital in many complicated cases. Public interest law thrives on both publicity and broader citizen participation. Often litigation launched on behalf of individuals or groups by public interest lawyers is one element in broad campaigns involving political lobbying, sophisticated media campaigns, and sometimes direct action.

The Evolution of Public Interest Law

Growth

Public interest law occupies a radically different place within the profession 40 years after its inception (Chen and Cummings, 2013). No longer in 'fragile alliance' with the organized bar (Scheingold and Sarat, 2004: p. 44), public interest law in the United States now enjoys the status of a stable and strongly supported occupational category within the legal profession. From an organizational perspective, the field has shown impressive expansion. Although direct

historical comparisons are not possible, the available data point to significant growth in both the number and size of public interest law groups over the movement's life span. Using survey data from 1975, Handler et al. (1978) identified 576 lawyer positions in 86 public interest law organizations nationwide (excluding legal aid organizations), yielding an average of approximately seven attorneys per group. In addition, Johnson (1974) reported that in 1972 there were 2660 lawyers in roughly 850 federally funded legal services offices (for an average of three attorneys per office). Taken together, these public interest lawyers (3236 in all) were approximately 0.7% of the total bar at the time. According to Nielsen and Albiston's 2004 survey data, there were slightly more than 1000 public interest law organizations (including legal aid organizations) with an average of 13 lawyers per group, for an estimated total of 13 715 attorneys in the field – approximately 1.3% of the total bar (Nielsen and Albiston, 2006). Although the 1975 and 2004 data are not directly comparable – and it is clear that public interest lawyers remain a tiny fraction of the overall bar – it does seem likely that significant growth has occurred, with these figures suggesting roughly a doubling of the public interest sector relative to the total bar. And this figure does not include public interest lawyers working in other sites, such as government, small firms, or clinical programs.

As the field has expanded, there appears to be more law school support for public interest students, which may promote supply – and increase competitive pressures within the field. Granfield's study (1992) of Harvard and Northeastern law students in the 1980s documented the socialization pressures that led to public interest 'drift.' Since that time, there have been new investments of resources in law schools devoted to 'nurturing' public interest law (Abel, 2002). According to The Equal Justice Works Guide to Law Schools in 2014, 21 US law schools (out of 198) reported offering public interest or public service specializations, while 35 provided public interest scholarships. Granfield's research suggests that institutional commitments to public interest law may nonetheless be overwhelmed by the dominant pedagogical style of law school and the availability of elite firm jobs for graduates. Public interest specializations are designed to counteract these influences by selecting students with public interest experience and fostering a subcommunity of ideological resistance.

In a similar sign of US law school commitment to public service, Rhode (2005) reports that one-fifth of law schools mandate pro bono or public service requirements as a condition of graduation, while half have formally administered voluntary programs. In the race to attract the best students and provide secure opportunities for summer employment, it has also become commonplace for top schools to offer summer funding for public interest jobs through some combination of law school financial support and student fundraising. Hiring processes for public interest jobs, too, have become more systematic and 'firm-like.'

A major issue affecting entry into public interest practice is quality of life. New public interest lawyers are relatively happy with their job settings, the substance of their work, and their engagement in social issues, but report sharply lower satisfaction levels when it comes to salary and opportunities for professional mobility (Dinovitzer et al., 2004). At $38 500 in 2004, average entry-level salaries of public interest lawyers were the lowest of any practice setting. Moreover, the disparity between public interest and law firm salaries has grown significantly. In the early 1970s, the ratio of private firm to public interest salaries was 1.5:1 (Komesar and Weisbrod, 1978). In 2004, the ratio of private firm (over 20 lawyers) to public interest salaries was roughly 3:1 and the ratio of big firm (over 250) to public interest salaries was 3.6:1. Law school debt has also increased – climbing sharply over the past several years – to $100 000 for private school graduates and over $60 000 for public school graduates as of 2009 (Schrag and Pruitt, 2011). The combination of stagnating salaries and rising debt would be expected to affect the supply as well as the potential quality of public interest lawyers, as those with opportunities to take high-paying jobs may do so despite a commitment to public interest practice. The development and expansion of law school Loan Repayment Assistance Programs (LRAP; and the 2007–08 enactment of federal loan forgiveness for public interest lawyers) to subsidize loan repayment for law school graduates who take low-paying public interest jobs was intended to alleviate the debt burden (and thus close the salary gap). The number of LRAPs at law schools doubled between 2000 and 2006, from 47 to 100 (Schrag and Pruitt, 2011). Fellowship programs have also played a significant role in producing public interest lawyers – at least at the elite end of the public interest bar. The Reggie program funded roughly 2000 fellows between 1967 and 1985 to enter legal services practice. The contemporary counterparts are the Skadden Fellowship and the EJW Fellowship programs. Both have been responsible for providing opportunities for elite law school graduates with demonstrated commitment to enter public interest practice by providing full salary and loan assistance.

Ideology

At its inception, the public interest law movement was explicitly a progressive one. Its 'public' was composed of those whose voices were deemed underrepresented in American politics: Blacks, women, the poor, consumers, environmentalists, and other marginalized groups. And its 'law' was mainly that of the federal judiciary, to which liberal lawyers turned to protect minority rights and promote (their version of) the collective good.

Political conservatism has reshaped the field of public interest law in three related ways: politically, organizationally, and ideologically. Politically, conservatism has transformed the 'social context' of public interest law. From an advocacy perspective, the major change has been the declining role of the federal government as the guarantor of legal rights associated with political liberalism. The most striking change has come in the judicial arena, where the struggle over the ideological composition of the federal bench has moved the weight of the judiciary toward a constitutional vision skeptical of economic regulation and minority rights. Republican control of the presidency for 7 of the 10 terms before the 2008 election of Barak Obama reversed majorities of democratically appointed judges at both the circuit and district court levels, and solidified the conservative majority on the US Supreme Court. The federal court, deemed the final arbiter of liberal claims to social justice during the civil rights period, has thus been transformed

into, at best, an unreliable ally and, at worst, a hostile enemy to be avoided.

Research suggests that liberal public interest lawyers are adapting their strategies to respond to the political realities – a trend that appears supported by the limited empirical evidence available. Although public interest groups continue to focus on filing lawsuits, Rhode (2008) reports that "it is with a more realistic vision of how they will serve long-term goals." While the mean percentage of legal work at public interest groups has remained constant since 1975, more groups are doing little or no legal work and spending more time on research, education, and outreach (Nielsen and Albiston, 2006). At the elite end of the public interest bar, there is also evidence to suggest that groups are engaged in more legislative work than in the early public interest law phase.

Organizationally, conservatism has both challenged public interest groups on the left and strengthened groups on the right. There have been well-documented attacks on liberal public interest lawyers, most notably in the 1996 the Legal Services Corporation (LSC) restrictions prohibiting lawyers in LSC-funded organizations from taking on class actions or collecting attorney's fees – thus limiting the type of cases undertaken by legal aid lawyers to routine individual service. In another example, Albiston and Nielsen found that the US Supreme Court's 2001 decision in *Buckhannon v. West Virginia*, which repealed the 'catalyst theory' for attorney's fee awards, has discouraged public interest groups from undertaking civil rights enforcement actions, particularly when the remedies are limited to injunctive relief. On the right, as Southworth (2008) has demonstrated, conservative groups have also contested public interest law from 'inside' the movement by establishing a countervailing set of legal organizations dedicated to advancing a right-wing political agenda under the rubric of serving the public interest. As a result, on the most contentious sociopolitical issues of the day, liberal public interest groups – standing for economic regulation, redistributive social welfare, the separation of church and state, the rights of criminal defendants, and protections for minority groups – have found themselves pitted against their conservative counterparts advocating a mirror-image agenda – free markets, small government, a prominent role for religion in public life, law and order, and an end to affirmative action.

Funding

Public interest law organizations were catalyzed in the 1970s by an infusion of funding from the Ford Foundation and other philanthropic institutions. In the early phase of public interest law, it was foundation funding that constituted the most significant source at over 40% of total income as of 1975. Recent research suggests that foundations play a less significant – though still important – funding role. In their survey of public interest law organizations, Nielsen and Albiston (2006) reported that foundation funding had decreased to 21% of total income by 2004, while state and local funding had increased to 28%. However, part of this reported decrease may be due to the fact that the study included legal aid organizations relatively more dependent on governmental funds. Rhode, in contrast, found that for the prominent public interest groups that she surveyed, foundation funding constituted an average of 37% of organizational budgets, while government funds represented only 1%. In addition, Rhode (2008) found that individual donations represented an average of 28% of total income, with corporate donations constituting another 14%. This suggests that elite public interest groups may be as dependent – or even more dependent – on private individuals and corporations for financial support (42% of total income) than foundations. What does this mean for organizational governance and programmatic content? One interpretation might be that such private donations reflect significant law firm financial involvement in these organizations. And one would expect such involvement to be accompanied by law firm representation on the groups' boards of directors. How far such financial relationships extend and how much they influence organizational policy are the questions that have troubled observers of public interest law since Bell (1976) charged the Legal Defense Fund with 'serving two masters' in its pursuit of school desegregation. If it is the case that law firms and other private sector actors are exercising more financial control over public interest law organizations, business considerations may compete with ideology in shaping their dockets.

Perhaps no public interest institution has been as politicized as the legal services program, which began in 1965 with federal funding as its primary source. The political backlash to its liberal agenda resulted in efforts to privatize the program and brought significant federal budget cuts that forced legal services offices to rely more heavily on foundations, law firms, individuals, and state and local government funding. In 1981, LSC mandated that its grantees make a 'substantial amount' of funds available for Private Attorney Involvement (PAI). Although the PAI program resulted in some direct payments to private practitioners, its major effect was to stimulate the expansion of programs designed to recruit, train, and connect pro bono volunteers with low-income clients. Spurred by the PAI mandate, the number of pro bono programs rose from about 50 in 1980 to over 500 in 1985. By the early 1990s, there were approximately 900 pro bono programs, and in 2004 there were roughly 1000. In terms of funding, LSC financial support declined just under 50% (in real terms) between 1980 and 2009. Overall legal services funding has held relatively steady, but has been diversified. As of 2005, LSC funds constituted only about one-third of the total legal services budget in the United States, with state and local government funds contributing about one-third, Interest on Lawyers Trust Accounts (IOLTA) about 10%, foundations about 7%, and private lawyer contributions roughly 4% (Houseman, 2005).

The constraints are much clearer in the context of big firm pro bono, which has grown in institutional scope and constitutes an important vehicle for delivering legal services to the poor. Pro bono's institutional ascendance occurred against the backdrop of several trends (Cummings, 2004): declining federal support for legal aid, which prompted the organized bar to promote pro bono as a complement to state-sponsored legal aid; the growth of big firms, which permitted the large-scale mobilization of pro bono resources through centralized planning; the advent of law firm rankings, which focused attention on pro bono as a recruitment strategy; and pressure from attorneys to give back in the face of law firm financial success. As pro bono has become institutionalized, it has also emerged as a significant component of US civil legal assistance.

Sandefur (2004) reports that, in 1997, lawyer pro bono work constituted "more than one-quarter of the full-time equivalent lawyer staff providing civil legal assistance." In addition, pro bono cocounseling is commonplace in public interest impact litigation, with Rhode (2008) finding that "almost all of the large national organizations relied heavily on pro bono counsel for impact litigation, and involved them in at least half of their major cases."

From a quality perspective, the rise of pro bono means that there are more attorneys working on public interest cases episodically – as part-time relief from their billable client work. This has implications for commitment and expertise. Some public interest attorneys complain about the need to closely monitor the quality of the work of pro bono volunteers, who at times are not closely supervised and are pulled away from pro bono work by their billable commitments. In addition, nonexpert pro bono lawyers may miss opportunities for connecting cases to broader law reform or political organizing efforts, and may be less likely to take risks to advance larger-scale social change agendas. From the perspective of institutional design, increased reliance on pro bono for the delivery of legal services to the poor may mean that certain categories of cases are less likely to receive attention. Positional conflicts and other business constraints make big firms reluctant to take on cases that strike at the heart of their business clients' interests. Thus, labor and employment law cases are less likely to receive significant support from firms that regularly defend business clients (Cummings and Rhode, 2010). Legal services groups may, in turn, organize their in-house programs around their appeal to private firm volunteers who will ultimately help staff the cases. Similar business constraints affect the ability of public interest groups to find pro bono counsel in impact cases against corporate defendants.

When public interest groups are unable to find big firm cocounsel in impact cases, they frequently turn to small plaintiff-side firms, for which the motivation is not 'giving back' through pro bono service, but rather the opportunity to collect attorney's fees. These 'private' public interest firms have been held out as an alternative site for 'doing well' and 'doing good,' allowing lawyers to take on large-scale social change litigation that nonprofit groups cannot because of resource limits – and big firm pro bono programs will not because of business conflicts – while also addressing other deficits associated with NGO practice, such as low salaries, lack of training, and high turnover. There are no systematic data on these firms, but the available evidence suggests that they have grown in number. Handler and his colleagues estimated that there were 66 'mixed' private public interest law firms in the mid-1970s. Cummings' 2012 compilation of private public interest law firms, drawn from lists in which private lawyers are identified as practicing in the public interest, put the number at over 600 (Cummings, 2012). Although this recent list includes firms engaged in activities such as securities and mass tort litigation, the majority are firms that undertake civil rights, human rights, and environmental litigation that have analogues in the NGO sector. Some private public interest firms only work on cause-oriented cases, while others supplement their public interest docket with cases taken on solely to generate fees.

The law school clinic constitutes another important (and underexplored) site of public interest practice – embedded within legal academia – that is characterized by distinct lawyering trade-offs. The clinical movement, which began in tandem with public interest law in the 1970s when Ford granted over $10 million to the Council on Legal Education for Professional Responsibility to promote clinical education, occupies a paradoxical space within law schools: radically institutionalized and yet often marginalized. Its institutional presence is powerful: In 1999, the Association of American Law Schools (AALS) Section on Clinical Legal Education's database showed 183 law schools with clinics staffed by over 1700 clinicians – 80% of whom reported teaching in live-client clinics. Despite these numbers, lower institutional status is a pervasive complaint among clinicians, who regularly do not enjoy the security and prestige of academic tenure and its attendant benefits.

In line with its original mission of client service, law school clinics now constitute a significant force in public interest law, yet we know very little about how clinics select cases, what trade-offs they face, and how their work relates to that of the broader NGO community. There are some obvious and well-documented limitations. Clinicians must select cases based in part on pedagogical value and – as high-profile cases of political backlash against environmental clinics at Tulane and Maryland for challenging industry pollution – the representation of progressive clients or causes may place clinical programs in political (and financial) risk.

Globalization

Public interest law, like its private sector counterpart, has also experienced the impact of global forces. This has been evident in two directions: from the outside-in, with the influence of globalization on US-based public interest practice, and from the inside-out, with the interaction between US public interest law models and legal activism around the world.

Within the United States, global interdependence has produced a greater degree of internationalization within the domestic public interest law field (Cummings, 2008). There have been three key transnational processes – immigration, market integration, and human rights – that have influenced US practice. With respect to immigration, since the 1980s there has been a quantitative increase in migration to the United States combined with a qualitative change in both its pattern (more geographically dispersed) and composition (more undocumented entrants). These changes have redefined immigrant advocacy, transforming it from an ancillary part of public interest practice into a distinctive field. Significantly, the influx of undocumented immigrants seeking employment in the exploitative low-wage sector has focused public interest resources on enforcing legal protections in the workplace, leading to the development of new organizational investments promoting 'immigrant worker rights'.

As immigration has brought market integration home, the outflow of US corporations in search of investment opportunities and low-cost production locales has extended open market policies to developing countries abroad. Within these new arenas of US economic activity, the main focus of public interest law has become upgrading systems of legal governance and regulatory enforcement outside of US borders. This has involved advocating new theories of US jurisdiction under the Alien Tort Statute; entering new venues of global economic

governance, such as the NAFTA system and the WTO; and building alliances with new partners, both transnational activist networks challenging market integration from below, and governmental and philanthropic institutions promoting the rule of law from above.

Domestic practice has also been influenced by the movement to 'bring human rights home'. Whereas the international human rights system promoted in the Cold War era was, in part, a way to export American-style public interest law to activists in foreign countries resisting authoritarian regimes, the US human rights movement represents an effort by public interest lawyers to import the very norms and methods built through international struggle to contest what they view as the erosion of domestic legal standards resulting from new American policy imperatives: market integration, conservatism, and the War on Terror. Accordingly, there have been significant efforts by mainstream public interest law groups to deploy human rights strategies to advance domestic agendas in the areas of criminal justice, immigrant rights, women rights, worker rights, and the environment.

Globalization has therefore influenced what public interest lawyers do inside the US legal system (which clients they represent and which causes they pursue) as well as what types of activities they undertake outside of American borders (which international venues they enter and which networks they support). From a tactical perspective, lawyering within the international arena embraces a broad range of nontraditional techniques such as lobbying, reporting, and organizing. It also operates across institutional contexts, moving from US courts to institutional venues – like the UN and Inter-American Commission on Human Rights – where lawyers think that they can amplify their claims and have political impact.

These global trends highlight the reversals of fortune, strategic adaptations, and deep tensions that characterize the American public interest movement in the contemporary era. The international turn is itself a product of domestic political realignment: Inside the United States, the liberal public interest law movement, built upon a symbiotic relationship with the federal government, has recently found itself in opposition to the main levers of federal power. It has, therefore, looked outside US borders – not just for legal resources but also for connections with international struggles – to infuse it with a renewed sense of movement energy and political mission. And it is there that US lawyers have found new political allies, as well as opportunities to engage in large-scale reforms that seem only a dim possibility at home. Particularly on issues of labor rights and environmental justice, US lawyers have found global partners eager to assert social standards within the regime of free trade. US lawyers have similarly invested in rule-of-law reforms in developing countries, not out of an impulse to remake the world in the American image, but rather drawn by the lure of enormous possibilities for profound legal and political change. Back at home, lawyers have also tapped into international movements to promote domestic reforms, taking up the banner of immigrant rights and enlisting the legal and rhetorical power of human rights in the service of domestic causes. In contrast to the self-confident insularity of public interest law during its early phase, these movements suggest that US lawyers now perceive that the rest of the world has political lessons to teach and legal models to emulate.

When one looks outside of the United States, the questions for public interest law are different. Key issues concern why public interest law is developing abroad and the degree to which foreign public interest law resembles – or resists – US models. With respect to the spread of public interest law, three developments are important: (1) the rise of the 'new' law and development movement beginning in the 1980s, which has promoted legal aid, clinical education, and pro bono as part of a broader set of market-friendly rule-of-law reforms supported by international financial institutions like the World Bank and IMF, as well as foundations like Ford and the Open Society Institute; (2) the emergence of constitutional democracies in places like Central and Eastern Europe, which provide new opportunities to use domestic courts to advance political ends, thus allowing activists to shift away from prior reliance on the human rights system; and (3) the development of organic movements to resist neoliberal globalization through the deployment of domestic and international law in coordination with indigenous organizing groups. Scholars are beginning to focus on the shape and meaning of this advocacy and its relation to the US model of public interest law. One question concerns the degree to which legal organizations abroad receive external funding and technical assistance that may connect them to public interest networks in the global North. Such foreign influence shapes the types of cases brought and strategies deployed. On the other hand, foreign public interest law systems must adapt to local legal, political, and professional environments, which create distinct advocacy opportunities and constraints. Because these hybrid systems combine elements of the global and local, there are inevitable struggles over authorship and power, with resistance to the notion of outside intervention, even while it leaves a distinctive imprint. As the interaction between foreign transmission and local adaptation plays out, global public interest lawyers will play leading roles in their own countries' movements for democratic reform in the global era – providing new perspectives on the evolving role of law in social change across a diverse range of political and economic systems.

Conclusion

Public interest law is an assertion of a commitment within the justice system to fairness, open process, and diversity of viewpoints. In response to the challenges the movement has faced over the past 25 years, particularly more conservative courts, many public interest lawyers have developed a range of alternate strategies: working more directly in the legislative and administrative processes; filing suits in state courts; and relying more heavily on citizen actions.

Despite these challenges, public interest law in the United States – and beyond – has continued to thrive, affecting policy decisions and providing a measure of balance in the legal system. There are generally more law school graduates seeking public interest law work than there are jobs; and there are many situations in which effective public interest advocacy, were it available, would lead to better policy outcomes. Additional efforts to build the field are thus still necessary. The development of public interest law in other countries holds significant

promise for fair processes and broader public participation in critical policy decisions ranging from environmental protection to the protection of racial minorities. Globalization creates new needs and opportunities for public interest lawyers to function in transnational settings.

See also: Discrimination and the Law; Equality and Inequality: Legal Aspects; Justice and Law; Justice, Access To; Law and Democracy; Law and Society: Development of the Field; Law as an Instrument of Social Change; Political Lawyering; Public Interest; Race and the Law.

Bibliography

Abel, R., 2002. Choosing, nurturing, training and placing public interest law students. Fordham Law Review 70, 1563.

Aron, N., 1989. Liberty and Justice for All: Public Interest Law in the 1980s and Beyond. Westview Press, Boulder, CO.

Bell, D., 1976. Serving two masters: integration ideals and client interests in school desegregation litigation. Yale Law Journal 85, 470.

Chen, A., Cummings, S., 2013. Public Interest Lawyering: A Contemporary Perspective. Wolters Kluwer, New York.

Council for Public Interest Law, 1976. Balancing the Scales of Justice: Financing Public Interest Law in America. Council for Public Interest Law, Washington, DC.

Cummings, S., 2008. The internationalization of public interest law. Duke Law Journal 57, 891.

Cummings, S., 2012. The pursuit of legal rights—and Beyond. UCLA Law Review 59, 506.

Cummings, S., NeJaime, D., 2010. Lawyering for marriage equality. UCLA Law Review 57, 1235.

Cummings, S., Rhode, D., 2010. Managing pro bono: doing well by doing better. Fordham Law Review 78, 2357.

Dinovitzer, R., et al., 2004. After the JD: First Results of a National Study of Legal Careers. The NALP Foundation for Law Career Research and Education (NALP) and the American Bar Foundation (ABF), Overland Park, KS.

Granfield, R., 1992. Making Elite Lawyers: Visions of Law at Harvard and Beyond. Routledge, New York.

Handler, J., et al., 1978. The public interest law industry. In: Weisbrod, B., et al. (Eds.), Public Interest Law: An Economic and Institutional Analysis. University of California Press, Berkeley.

Houseman, A., 2005. Civil Legal Aid in the United States: An Overview of the Program in 2005.

Johnson Jr., E., 1974. Justice and Reform: The Formative Years of the OEO Legal Services Program. Transaction Publishers, New Brunswick, NJ.

Komesar, N., Weisbrod, B., 1978. The public interest law firm: a behavioral analysis. In: Weisbrod, et al. (Eds.), Public Interest Law: An Economic and Institutional Analysis. University of California Press, Berkeley.

McClymont, M., Golub, S. (Eds.), 2000. Many Roads to Justice: The Law Related Work of Ford Foundation Grantees around the World. The Ford Foundation, New York.

Nielsen, L.B., Albiston, C., 2006. The organization of public interest practice: 1975–2004. North Carolina Law Review 84, 1591.

Rhode, D., 2005. Pro Bono in Principle and in Practice. Oxford University Press, New York.

Rhode, D., 2008. Public interest law: the movement at midlife. Stanford Law Review 60, 2027.

Sandefur, R., 2004. Lawyers' pro bono service and American-style civil legal assistance. Law and Society Review 41, 79.

Sarat, A., Scheingold, S. (Eds.), 1998. Cause Lawyering: Political Commitments and Professional Responsibilities. Oxford University Press, New York.

Scheingold, S., Sarat, A., 2004. Something to Believe in: Politics, Professionalism, and Cause Lawyering. Stanford University Press, Stanford.

Schrag, P., Pruitt, C., 2011. Coordinating loan repayment assistance programs with new federal legislation. Journal of Legal Education 60, 583.

Southworth, A., 2008. Lawyers of the Right: Professionalizing the Conservative Coalition. University of Chicago Press, Chicago.

Weisbrod, B.A., et al., 1978. Public Interest Law: An Economic and Institutional Analysis. University of California Press, Berkeley, CA.

Public Opinion: Social Attitudes

Edward P Freeland, Survey Research Center, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public opinion and social attitudes are broadly defined as the expression of beliefs, ideology, and sentiment within a citizenry with regard to the conduct of human affairs. The concept of public opinion arose as a consequence of the rise of democracy, premised on the notion that government should rule with the consent of the governed. Methods for collecting and analyzing data on public opinion based on survey samples were first developed in the US and Europe and are now used in most countries throughout the world. Within the field of survey research, the scientific study of public opinion has emerged as a specialized area of research that is done by sociologists, psychologists, political scientists, and market researchers. In the late 1960s, political scientists in the US were surprised to find that very few people base their political opinions and vote choices on a coherent system of beliefs. They also found that large percentages of respondents seemed to switch their opinions when asked the same question at different points in time. This instability and apparent lack of ideological consistency poses serious challenges for the theory of government by democracy. Subsequent research has shown that most people's opinions are influenced by family socialization and education and are often based on perceptions of various social groups. Researchers also study the limitations of the methods and techniques used to measure public opinion in order to minimize biases that cause researchers to misinterpret their findings. Unlike the patterns of instability observed in individual opinions, public opinion at the aggregate level appears much more stable and consistent over time. Research on long-term trends in aggregate opinion has also shown that government does indeed respond to citizen preferences, but there is evidence to suggest that it has become more responsive (at least in the US) to the preferences of affluent citizens. Due to the globalization of the survey research industry, social scientists now have access to public opinion data from the vast majority of countries throughout the world.

Public opinion and social attitudes are broadly defined as the expression of beliefs, ideology, and sentiments within a citizenry with regard to the conduct of human affairs. Opinions are usually understood as an expression of underlying attitudes and sentiments, which tend to remain stable over time while opinions can change more rapidly depending on public events or exposure to new information. Social attitudes are more closely tied to widely shared human values, namely our most basic and important beliefs about human life and social order. Opinions are assessments of particular people, events, or policies as they are understood through the lenses of our basic social attitudes or ideologies. While most people generally agree on and share basic core values in a society, opinions about specific areas of life can vary widely from person to person and within individuals over the course of the life span.

Opinions are generally understood to be a result of life experiences, exposure to information, peer pressure, and other forms of social influence. Opinions vary widely in terms of substantive focus and in terms of their importance to each person. Thus, in one dimension, opinions differ in their *valence* (e.g., favor or oppose laws restricting gun ownership) and, in a separate dimension, in their *intensity* (e.g., strongly care or do not care at all about this issue). This means that while many people may share the same opinion about an issue, they may, at the same time, disagree about the importance of that issue (Petty and Krosnick, 2014). Collective action or social movements occur when individuals sharing the same opinion on an issue and a sense of urgency about the importance of that issue come together to form a group intent on changing laws, policies, or social norms. In more extreme cases, collective action driven by shifts in public opinion can result in rebellions or social revolutions (Olson and Olson, 2009).

Opinions are distinct from perceptions, which are based on observations of factual reality and can be judged as either correct or incorrect. By contrast, opinions cannot be judged as correct or incorrect: they are the result of moral or esthetic judgments. They pertain to matters of right and wrong or personal taste that are typically thought of as subjective. Opinions are part of the sphere of cognition in which individuals are free to think and judge as they wish, although the public expression of opinions may be more or less tightly controlled by various social institutions.

Historically speaking, the concept of public opinion arose as a consequence of the rise of democracy. Inherent in the principle that government should rule with the consent of the governed is the notion that government must understand and respond to the needs and aspirations of its citizens. Throughout most of human history, the exercise of power in human affairs paid little or no heed to the opinions of ordinary people. However, with the rise of the modern state and the concept of citizenship, rulers of both democratic and nondemocratic states have understood that citizens are capable of expressing their collective will in a variety of ways that may support or threaten the operations of government. This possibility is implied in the classic definition of public opinion offered by the political scientist V.O. Key (1961: 14) as the "opinions held by private persons which governments find it prudent to heed."

One of the earliest and most important studies of public opinion was first published in 1922 by journalist Walter Lippmann. In his book, Public Opinion, Lippmann argued that the complexities of modern life make it impossible for human beings to pay attention to more than just a fraction of what goes on in their environments. To cope with these complex environments and the unseen realities outside of their local

communities, people create simplified mental images of the world based on stereotypes. In forming opinions, people draw on these simplistic stereotypes to make sense of the world and events happening beyond their immediate grasp. These stereotypes are resistant to contradictory information, but at the same time vulnerable to manipulation by skilled politicians and newspaper editors. As a reporter during World War I, Lippmann saw firsthand the myriad ways that newspaper accounts were manipulated by military and political leaders to suit their own purposes. He also understood that the business of selling newspapers was based on maximizing circulation and advertising revenues, and not necessarily on delivering relevant facts in an objective and comprehensive way. By virtue of being profit-driven enterprises, newspapers are more likely to cater to their readers' stereotypes rather than challenging them. Lippmann was prescient in understanding the way most people think about politics and world affairs and the way news media play a crucial role in shaping public opinion.

> Thus the environment with which our public opinions deal is refracted in many ways, by censorship and privacy at the source, by physical and social barriers at the other end, by scanty attention, by the poverty of language, by distraction, by unconscious constellations of feeling, by wear and tear, violence, monotony. These limitations upon our access to that environment combine with the obscurity and complexity of the facts themselves to thwart clearness and justice of perception, to substitute misleading fictions for workable ideas, and to deprive us of adequate checks upon those who consciously strive to mislead.
>
> (1932: 76)

The Scientific Study of Public Opinion

The measurement of public opinion is accomplished through opinion polls or social surveys. A poll or a survey is a systematic effort to collect information about opinions, perceptions, behaviors, or intentions among a random sample of people that is more or less representative of a larger population. While polls and surveys are a seemingly ubiquitous feature of modern life, the techniques for measuring public opinion took many years to develop. The methods used by researchers to conduct a survey or poll evolved over a 100 year period that began with the founding of professional statistical societies in Great Britain and the US in the 1830s and culminated in the first use of modern sampling methods in nationwide surveys of employment and unemployment during the Great Depression of the 1930s. During this time the techniques for creating representative samples were developed by a group of statisticians and social scientists from many different countries, including Anders Kiaer, a statistician who conducted numerous experiments with Norwegian census data in the 1890s, and Jerzy Neyman, a Polish statistician who provided the mathematical foundation for modern sampling techniques based on classical theories of probability (Converse, 1987: 41–45). Opinion polling and market research has grown into a sizable worldwide industry that employs thousands of survey researchers, interviewers, and statisticians.

Surveys and polls are now used throughout the modern world as efficient and cost-effective methods for gathering data on public opinion, but also for collecting many kinds of information needed for the normal operations of government and industry. Within the field of survey research, the scientific study of social attitudes and public opinion has emerged as a specialized area of study that includes research done by sociologists, psychologists, political scientists, and market researchers. Their data come from interviews conducted with large numbers of survey respondents in many countries around the world. These interviews are done in person, over the telephone, by mail using printed questionnaires, or over the Internet using online questionnaires. In addition to surveys and polls, researchers who study public opinion also use observations and data gathered from controlled laboratory experiments, focus groups, observational studies, and the analysis of media content such as newspaper articles and radio and television broadcasts. These researchers study the factors that influence public opinion as well as processes of social change that are driven by shifts in public opinion.

One of the first systematic, large scale efforts to study public opinion in the United States was carried out by Phillip Converse (1964), a political scientist at the University of Michigan. Converse analyzed national survey data collected from election studies in 1956, 1958, and 1960. One important feature of Converse's research was that while each of the samples were nationally representative, each one also included sample members who were interviewed at all three points in time using the same set of survey questions. This innovative design feature enabled Converse to measure change in *aggregate* opinion over the 4 year period of the study, as well as change within *individual* opinions during the same period. At the time of the study, most social scientists like Converse assumed that people generally based their opinions on a basic set of core values or social attitudes in a logically coherent way. They further assumed that most individuals maintain relatively stable opinions and attitudes over time. However, for the most part, both assumptions turned out to be wrong.

Using responses from a wide range of policy-related survey questions as well as detailed explanations for their opinions and vote choices, Converse divided the mass electorate into five different categories, configured as a pyramid, with the most highly educated and attentive voters at the top of the pyramid, moderately interested and lower educated voters near the middle, and the least interested and least educated citizens at the bottom. Converse designated the elite group at the top as ideologues, namely people who rely on clear and consistent principles (e.g., liberal vs conservative, prolabor vs probusiness, etc.) in deciding how to vote or respond to questions about their opinions on policy issues. Just below this level was a group Converse called near-ideologues – people who mention any kind of evaluative dimension in their orientation to politics, but rely on that dimension less often than the ideologues. In the third level, Converse placed respondents who do not rely on any apparent evaluative dimension, but instead explain their thinking about candidates and issues in terms of the impact of politics on various social groups. People at this level tend to see politics as the interplay between competing interest groups, so that their political opinions derive largely from their own sense of group membership and their perceptions of which groups have more power and which deserve more help. Below this group are people who talk about some policy considerations in explaining their vote and policy

preferences, but apply those considerations in a haphazard way. The last group comprises people whose views and vote choices reflect neither any kind of structured ideas or principles nor any reference to competing social interests.

When Converse sorted his survey respondents into these five categories, he found that fewer than 3% of the total sample fell into the first category (ideologues) and only 9% fell into the second category (near-ideologues). Slightly more than 40% fell into the interest group category (level 3) and the rest (almost half the sample!) were found to have little or no coherent ideological underpinnings associated with their opinions or vote choices. Thus, it was not surprising to find that many sample members also gave entirely different answers in 1958 and again in 1960 to the same set of policy questions they had been asked two years earlier. Converse noted with great concern that while opinions at the aggregate level appear to be relatively stable over time, that stability masks a great deal of instability that seems to fluctuate in an almost quasirandom manner in the minds of many individual voters. The results of his study indicated that "large portions of the electorate do not have meaningful beliefs, even on issues that have formed the basis for intense political controversy among elites for substantial periods of time" (1964: 245). Beyond giving credence to Lippmann's pessimistic view of public opinion, Converse's findings cast a long shadow over prevailing assumptions about the need for an active, well-informed and thoughtful electorate for the success of government by democracy. In fact, much of the research that has been done in the field of political science for the past 50 years can be understood as attempts to address the implications of Converse's research.

It is important to consider whether some of the incoherence and long-term instability that Converse found to be characteristic of most people's belief systems is due to limitations in the methods used to collect data on opinions. Incoherent or unstable responses are sometimes the result of poorly designed questionnaires, biases in the way questions are worded, or measurement errors related to the interview process itself. Researchers have documented a 'social desirability' bias that occurs when respondents seek to avoid presenting themselves in a negative light, and withhold from expressing opinions that might be perceived as negative (Tourangeau and Yan, 2007). This even extends to a desire to avoid appearing ignorant on issues about which the respondent has limited knowledge or no knowledge at all. An experiment by Bishop et al. (1986) showed that when asked questions about a policy proposal that was entirely fictitious, significant numbers of respondents offered favorable or unfavorable opinions when the question did not include the option of answering with a 'do not know' response. Similar kinds of experiments have demonstrated that responses to questions can be significantly altered by the use of open-ended responses vs lists of closed-ended responses; by differences in the way scale values are labeled (e.g., 0 to 10 vs −5 to +5); by differences in the end points of ordinal response categories; and by changes in the order in which questions are presented (Schwarz, 1999). Response biases have also been observed as a result of using different modes of interviewing (e.g., telephone vs in person) (Holbrook et al., 2003).

Measurement error may also result from the use of simplistic questions to gauge opinion on complex issues. Some researchers claim that Converse's methods do not adequately capture the complex process by which core values contribute to the expression of specific policy preferences. Achen (1975) argues that much of the response instability that Converse claims to detect can be attributed to vagueness in the survey questions themselves. His reanalysis of Converse's data shows that response variability over time is uncorrelated with education, income, and occupation – a strong indication that even sophisticated respondents had trouble interpreting the survey questions that showed the most response variability over time. In this way, measurement error obscures the process by which core beliefs inform opinion in a meaningful way. To better understand this process, Zaller (1992) proposes a model of survey response in which core beliefs and predispositions act as filters or gatekeepers for information flowing through our political environments. Core values affect our receptiveness to ideas and information and constrain the process through which we evaluate information and decide what to store in memory. When asked to express an opinion in response to a question, respondents call to mind a range of relevant facts, ideas, and experiences drawn from memory, and 'average' across the most salient considerations in deciding how to answer a question. Seen from this perspective, response instability does not indicate the absence of coherent core beliefs, but instead indicates the limitations inherent in the typical question and answer format used in most opinion polls.

Factors That Shape Individual Opinions

If, for most people, attitudes and opinions are not grounded in any logical, coherent way on underlying belief systems or ideologies, what then determines their attitudes and opinions? Most research on the factors that shape an individual's political opinions point to the importance of family socialization, group identity and prejudice, and the effects of education. There is even some evidence to support the notion that genetics may play a role in determining one's political outlook. Other research has examined the extent to which people base their opinions on social class or self-interest.

In reaction to the rise of Nazi Germany and the events of World War II, many psychologists and social scientists were concerned about the potential for social institutions such as the family to raise children with strong prejudices against people different from themselves. Those prejudices were part of a set of traits that became known as the 'authoritarian personality' (Adorno et al., 1950). Children reared in family environments that cultivated this type of personality were seen as strictly obedient to authority and intolerant of social groups other than their own. This was seen as an important factor in understanding anti-Semitism and other forms of prejudice. Later, during the civil rights movements of the 1960s, many researchers who studied social attitudes about race and racial equality focused less on personality traits and more on social learning processes and the role of parents in transmitting attitudes of racial prejudice to their children (Allport, 1954; Levy et al., 2010). In terms of political attitudes more generally, recent research has shown that under certain circumstances, parents exert a significant influence on the political views of their children. Using longitudinal data from three generations

of parents and children, Jennings et al. (2009) find that children are most likely to adopt parents' political orientations and opinions when parents have a greater interest in politics, provide consistent cues to their children over time and share the same views as their spouses. The effect of parent–child transmission is particularly strong when it comes to affect laden issues such as party identification, vote choice, religiosity, opinion on school prayer, and racial attitudes. This effect of parental influence holds up, though its impact is somewhat diminished, even when controlling for contextual factors such as parents' education and income, school environment, and local economic conditions. Additional evidence suggests that the timing of parental influences is unclear, with some researchers arguing that parental impacts are strongest in early childhood and others claiming that parental influence is equally or more important during adolescence (Sears and Funk, 1999).

More recent research indicates that the link between parent–child political views may occur as a result of biological or genetic influences. Using survey data collected from thousands of twins born in the US and in Australia, Alford et al. (2005) compare differences in attitudinal measures on a wide range of political issues among both identical (monozygotic) and fraternal (dizygotic) twins. They find that correlations on attitude measures are significantly stronger for identical twins when compared with correlations for fraternal twins, and this remains true even when accounting for whether or not both types of twins shared the same family environment during childhood. Although this work has been criticized for inadequately testing its assumptions about the equivalence of the environments in which twins are being raised (for example, see Beckwith and Morris, 2008), the strengths of the findings, coupled with weak but suggestive evidence on the mechanisms through which familial and environmental influences actually operate, mean that genetic backgrounds, or at least the interaction of genetics and environment, may be a significant factor in understanding the role of parental influence on attitudes and opinions.

Converse's observation that people are much more likely to think about politics in terms of interest groups than in terms of a coherent belief system helped lay the groundwork for research on the effects of ethnocentrism and its role in the formation of political identities. This view can be seen as a more generalized application of the basic ideas behind research on the authoritarian personality – beyond simply creating negative stereotypes of certain groups (such as Jews or Blacks), reference group theory, as applied in political analysis, assumes that most people see the world as composed of many different kinds of groups, some of which they view negatively and some of which they view positively (Merton and Kitt, 1950; Fiske, 1998). In addition, individuals may see themselves as members of several different groups at the same time (Catholic, working class, white, immigrant, etc.) (Tajfel, 2010). This leads to the question of whether and how often attitudes and opinions can be understood as manifestations of underlying group identities and the stereotypes people hold about groups other than their own. In reviewing past research on this idea, Kinder (2003: 18) argues that "[g]roup sentiment is not the only thing driving opinion in… various policy disputes, but it is always present, and of all the diverse ingredients that make up opinion, it is often the most powerful."

Ironically, some research has shown that higher levels of education do not always serve as an antidote to the prejudices that inform many people's thinking about politics. Stouffer (1955) was one of the first American researchers to develop measures of political tolerance and to provide evidence that education was positively associated with higher levels of tolerance for nonconformist groups. Bobo and Licari (1989) use data from the 1984 General Social Survey to show that education has a positive, independent effect on people's willingness to tolerate freedom of expression for extremist political groups. This was true even among respondents who had a negative view of those groups. Gibson (2013) finds a similar independent effect of education on political tolerance using more recent data from five national surveys conducted between 2007 and 2011. Thus, while the evidence suggests that education makes people more tolerant of various social and political groups, there is not much evidence to support Converse's original assumption that more highly educated people are less likely to think about politics in terms of racial or ethnic stereotypes (Kinder, 2003: 38).

Marxist theories of public opinion view it as an accommodation to or outgrowth of the system of control over a society's material and productive resources. People in capitalist societies may have many opinions about what policies or laws are in their best interest, but Marxist theorists assert that people's real interest lies in understanding their true relationship to the means of production – all other interests are subordinate. Thus, the only opinions that matter are those concerned with understanding and opposing the oppressive power structures that characterize life in capitalist societies. According to Marxists, people often misperceive their 'true' interests and develop 'false' beliefs (such as religion) that mask the true source of their oppression. Marxists refer to such beliefs as a form of 'false consciousness' (Rosen, 1996). This concept is premised on the idea that what is in one's best interest can be determined objectively and is not simply a matter of subjective preference (Mansbridge, 1983).

Given the capacity of democratic governments to respond to working class demands and the absence of revolutions in modern capitalist societies, Marxist theories of public opinion have not enjoyed much popularity. Critics have also pointed to the difficulty of identifying and prioritizing objective interests in modern, affluent societies (Bartels, 1996). Nevertheless, some scholars have studied instances in which people vote and express opinions that seem directly contrary to their own economic interests. Thomas Frank (2004), for example, looks at the evolution of the conservative movement throughout rural Mid-Western states to try to explain how economic issues were supplanted by cultural and symbolic issues in the minds of many voters. In his book *What's the Matter with Kansas?* Frank argues that economic conservatives in America have exploited 'culture war' issues such as abortion, school prayer, and gay rights to lure white working class voters away from their long-standing tradition of supporting labor unions, progressive taxation, and protectionist trade policies – issues that directly impact their income and employment. In another study, Bartels (2009) examines the campaign to eliminate the federal estate tax in the US. Although the tax itself affected only a very small

percentage of the nation's wealthiest families, many low- and middle-income voters expressed strong opposition to the tax even though its elimination would cause a shortfall in government revenues that would inevitably impact all taxpayers. These studies show that material interests are just one of many factors that influence public opinion.

Factors That Shape Aggregate-Level Opinion

In looking at patterns of change and stability in public opinion over time, Page and Shapiro (1992) claim that patterns of response instability that Converse and other researchers observe in individuals has limited relevance to patterns observed for the population as a whole. In contrast to individual opinion, public opinion at the aggregate level has been, for most issues, relatively stable over time. Page and Shapiro argue that even if individual response variability has a significant random component, the net result of aggregating data across many individuals is to cancel out any bias that might result from random variation in individual responses. Thus, the authors argue, while individuals may not always appear rational in their policy choices, the American public as a whole has, for the most part, been stable and consistent in its policy preferences. Page and Shapiro use data compiled from thousands of national opinion surveys conducted in the US between 1935 and 1990 to demonstrate that significant, large-scale shifts in the policy preferences of the public as a whole are relatively rare. Although major changes have occurred on issues such as civil rights, women's rights, civil liberties, the Vietnam War, and the use of nuclear power, opinions on most other issues show only moderate or negligible long-term change.

In a similar analysis of polling data, Stimson (2004) also finds long-term steady growth in the acceptance of equal rights for racial minorities, for women, and, more recently, for homosexuals. In each of these issue domains, laws and government policy changed as a result of clear shifts in public preferences. The long-term trend on nearly all other issues (including gun control, welfare spending, taxes, the environment, crime, and health care) is a pattern that fluctuates over time. Stimson argues that these fluctuations in public preference build up gradually in the years before major elections and presage subsequent shifts in political party dominance, in particular, dominance of the White House. Stimson also speculates that these fluctuations are consistent with movements of public opinion that signal changes in the public's demand for more or less spending in certain program areas. Earlier research by Wlezien (1995) finds evidence that changes in appropriations for national defense are preceded by changes in public opinion on whether the government should be spending more or less on defense. Wlezien argues public opinion in this case functions much like a thermostat, with elected officials responding to demands for increased spending and then opinion moderating as those demands are met.

Public Opinion as a Social Force

These findings raise the question of how much elected officials are aware of public opinion and how much it influences their actions. Page and Shapiro (1983) argue that aggregate-level trends in public opinion data provide clear evidence that government policy changes in response to major shifts in citizen preferences. But aggregate-level data are of limited value in understanding how government officials translate signals from voters into concrete political agenda. In a study of the ways that legislative staffers, journalists, and interest group leaders interpret and react to public opinion, Herbst (1998) finds that elected officials want to know not just what percent of citizens favor a certain policy, but how *important* that policy is to them. Opinion polls typically fail to convey how much an issue really matters to voters because they usually lack measures of attitude strength. Polls give equal weight to voters who care passionately about an issue and those who do not. For politicians, *intensity* of opinion matters as much as its *valence*. Herbst finds that most politicians use an operative definition of public opinion that takes account of more than just numbers from polls and surveys. Elected officials depend on interest group leaders, lobbyists, reporters, newspaper editorials, constituent communications, and networks of local party officials for clues about the issues that concern voters and how much they are concerned. As such, data from polls and surveys reveal only one dimension of what elected officials regard as a multifaceted phenomenon.

This leads to the question of whether elected officials treat some expressions of public opinion differently from others. Certainly all officials find it 'prudent to heed' the opinions of their constituents, but do some constituent preferences get more attention than others? Gilens (2014) analyzes trends in US public opinion and changes in federal government policy over a four-decade period to examine this question. In general, he finds that government responsiveness to public opinion has shifted markedly in favor of the preferences of more affluent citizens, in particular, the wealthiest 10% of Americans. There are some indications that government is more responsive to low and middle income Americans during presidential election seasons and in cases of policy debates in which powerful interests groups are engaged on their behalf. But these are rare exceptions in a growing tendency for government policy to be increasingly driven by the opinions of the nation's most affluent citizens. Gilens claims that "under most circumstances, the preferences of the vast majority of Americans appear to have essentially no impact on which policies the government does or doesn't adopt" (2014: 2).

The Globalization of Opinion Research

Data on public opinion are now collected in almost every country throughout the world. The World Values Survey (WVS) began as the European Values Survey in 1981. Representative samples of citizens were interviewed in 10 western European countries with a common core of questions about moral and social values. For its second wave of interviews, 12 non-European countries were added. The sixth and most recent wave of data collection covers more than 50 countries, enabling researchers to track changes in moral values in a wide range of countries over a period of more than 40 years. One of the most important studies to emerge from the WVS is based on data

analyzed by Inglehart and Baker (2000) showing that traditional cultures have long-lasting impacts on moral values in countries undergoing modernization. This occurs in spite of the fact that modernization "is associated with shifts away from absolute norms and values toward values that are increasingly rational, trusting, tolerant and participatory."

The Global Barometer surveys grew out of the Eurobarometer Survey that was initiated by the European Commission in 1973. The surveys address a wide range of public issues on topics related to European citizenship, international relations, health, culture, and the environment. Similar multinational surveys are conducted for the Latinobarometer, which gathers data from more than 20 000 interviews in 18 countries in Central and South America, the Asiabarometer, which covers 18 countries in Southeast Asia, and the Afrobarometer, which began in 1999 and has completed five waves of data collection in more than 35 African nations.

The Gallup World Poll originated from an international consortium of independent polling companies. The Gallup International Research Institute held annual meetings from 1937 to 1939 at which representatives from 50 different countries shared their findings from polls conducted on a wide range of topics. Today, the Gallup World Poll conducts survey and opinion polls in 160 different countries. The poll is known especially for its comparative measures of personal well-being, which are based in part on a question that asks respondents to describe their life satisfaction by placing themselves on an imaginary ladder with steps numbered 0–10. Results from the poll show that people living in Scandinavian countries are most satisfied with their lives while people in the poorest countries of Eastern Europe, Southeast Asia, and Africa are among the least satisfied. Scholars have used data from the Gallup World Poll to study the relationship between personal well-being, income, aging, and health (Deaton, 2007; Stevenson and Wolfers, 2008).

See also: Citizenship, Sociological Aspects of; Collective Behavior, Sociology of; Collective Memory, Psychology of; Movements, Social; Prejudice in Society: Sociological Perspectives; Voting, Sociology of.

Bibliography

Achen, C.H., 1975. Mass political attitudes and the survey response. The American Political Science Review 69 (4), 1218–1231.

Adorno, T.W., Levinson, D.J., Sanford, N., 1950. The Authoritarian Personality. Harper, New York.

Alford, J.R., Funk, C.L., Hibbing, J.R., 2005. Are political orientations genetically transmitted? American Political Science Review 99 (02), 153–167.

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.

Bartels, L.M., 1996. Public opinion and political interests. In: Annual Meeting of the Midwest Political Science Association, Chicago. Revised August 1996.

Bartels, L.M., 2009. Unequal Democracy: The Political Economy of the New Gilded Age. Princeton University Press, Princeton.

Beckwith, J., Morris, C.A., 2008. Twin studies of political behavior: untenable assumptions? Perspectives on Politics 6 (04), 785–791.

Bishop, G.F., Tuchfarber, A.J., Oldendick, R.W., 1986. Opinions on fictitious issues: the pressure to answer survey questions. Public Opinion Quarterly 50 (2), 240–250.

Bobo, L., Licari, F.C., 1989. Education and political tolerance testing the effects of cognitive sophistication and target group affect. Public Opinion Quarterly 53 (3), 285–308.

Converse, J.M., 1987. Survey Research in the United States: Roots and Emergence, 1890–1960. University of California Press, Berkeley, CA.

Converse, P.E., 1964. The nature of belief systems in mass publics. In: Apter, D. (Ed.), Ideology and Discontent. John Wiley, New York, pp. 202–261.

Deaton, A., 2007. Income, Aging, Health and Wellbeing around the World: Evidence from the Gallup World Poll (No. W13317). National Bureau of Economic Research.

Fiske, S.T., 1998. Stereotyping, prejudice and discrimination. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed. McGraw-Hill, Boston.

Frank, T., 2004. What's the Matter with Kansas? How Conservatives Won the Heart of America. Holt, New York.

Gibson, J.L., 2013. Measuring political tolerance and general support for pro–civil liberties policies notes, evidence, and cautions. Public Opinion Quarterly 77 (1), 45–68.

Gilens, M., 2014. Affluence and Influence. Princeton University Press, Princeton.

Herbst, S., 1998. Reading Public Opinion: How Political Actors View the Democratic Process. University of Chicago Press, Chicago.

Holbrook, A.L., Green, M.C., Krosnick, J.A., 2003. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. Public Opinion Quarterly 67 (1), 79–125.

Inglehart, R., Baker, W.E., 2000. Modernization, cultural change, and the persistence of traditional values. American Sociological Review 65 (1), 19–51.

Jennings, M.K., Stoker, L., Bowers, J., 2009. Politics across generations: family transmission reexamined. The Journal of Politics 71 (03), 782–799.

Key, V.O., 1961. Public Opinion and American Democracy. Knopf, New York.

Kinder, D.R., 2003. Belief systems after Converse. In: MacKuen, M.B., Rabinowitz, G. (Eds.), Electoral Democracy. University of Michigan Press, Ann Arbor.

Levy, S.R., Rosenthal, L., Herrera-Alcazar, A., 2010. Racial and ethnic prejudice among children. In: Chin, J.L. (Ed.), The Psychology of Prejudice and Discrimination. Greenwood, Santa Barbara, CA, pp. 37–50.

Lippmann, W., 1932. Public Opinion. The Macmillan Co., New York.

Mansbridge, J., 1983. Beyond Adversary Democracy. University of Chicago Press, Chicago.

Merton, R.K., Kitt, A.S., 1950. Contributions to the theory of reference group behavior. In: Merton, R.K., Lazarsfeld, P.F. (Eds.), Continuities in Social Research: Studies in the Scope and Method of "The American Soldier". Free Press, Glencoe, IL, pp. 40–105.

Olson, M., Olson, M., 2009. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge.

Page, B.I., Shapiro, R.Y., 1992. The Rational Public: Fifty Years of Trends in Americans' Policy Preferences. University of Chicago Press, Chicago.

Page, B.I., Shapiro, R.Y., 1983. Effects of public opinion on policy. The American Political Science Review 77 (1), 175–190.

Petty, R.E., Krosnick, J.A. (Eds.), 2014. Attitude Strength: Antecedents and Consequences. Psychology Press, New York.

Rosen, M., 1996. On Voluntary Servitude: False Consciousness and the Theory of Ideology. Polity Publishers, Cambridge.

Schwarz, N., 1999. Self-reports: how the questions shape the answers. American Psychologist 54 (2), 93.

Sears, D.O., Funk, C.L., 1999. Evidence of the long-term persistence of adults' political predispositions. The Journal of Politics 61, 1–28.

Stevenson, B., Wolfers, J., 2008. Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox (No. W14282). National Bureau of Economic Research.

Stimson, J.A., 2004. Tides of Consent: How Public Opinion Shapes American Politics. Cambridge University Press, Cambridge.

Stouffer, Samuel C., 1955. Communism, Conformity, and Civil Liberties. Doubleday, New York.

Tajfel, H. (Ed.), 2010. Social Identity and Intergroup Relations, vol. 7. Cambridge University Press, New York.

Tourangeau, R., Yan, T., 2007. Sensitive questions in surveys. Psychological Bulletin 133 (5), 859.

Wlezien, C., 1995. The public as thermostat: dynamics of preferences for spending. American Journal of Political Science 39, 981–1000.

Zaller, J.R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, Cambridge.

Relevant Websites

http://www.aapor.org/ – American Association of Public Opinion Research.
http://www.europeansurveyresearch.org/ – European Survey Research Association.
http://www.gallup.com/strategicconsulting/en-us/worldpoll.aspx – The Gallup World Poll.
http://www.globalbarometer.net/ – Global Barometer Surveys.
http://wapor.org/ – World Association for Public Opinion Research.
http://www.worldpublicopinion.org/ – World Public Opinion.
http://www.worldvaluessurvey.org/ – World Values Survey.

Public Policy

Anke Hassel, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public policy is a set of decisions by governments and other political actors to influence, change, or frame a problem or issue that has been recognized as in the political realm by policy makers and/or the wider public. Scientific approaches toward public policies have proliferated over the postwar period as the size and scope of government interventions have continuously expanded. The study of public policy includes policy analysis or policy science, which identifies effective policy measures, policy instruments, which a government can employ, and the policy process, which analyses how a government comes to take a decision.

Introduction

Public policy is a set of decisions made by governments and other political actors to influence, change, or frame a problem or issue that has been recognized as being in the political realm by policy makers and/or the wider public. Public policy can also take the form of a nondecision (Dye, 1972) or a deliberate neglect (Klein and Marmor, 2006). Strategic approaches toward public policies have proliferated over the postwar period, as the size and scope of government interventions have continuously expanded. Today, there are still high expectations about the problem-solving capacities of public policies, even though, in the last couple of decades, many policy problems could not be solved. Trust in government is also declining and problem solving has been increasingly taken up by private sector initiatives. These can be found in the form of social enterprises, charities, or business activities.

The study of public policy includes policy analysis or policy science, which identifies effective policy measures, policy instruments, which a government can employ, and the policy process, which analyses how a government comes to make a decision. The fundamental question in the study of public policy is still to reconcile the fundamentally different constraints in the field of politics and the field of policies. While many policies are now available for solving problems and policy knowledge is vastly improved, many policies remain politically contested.

Origins

Until the full democratization of modern societies set in, politics was the process of establishing order and power. Public policies were one part of politics as they were employed to satisfy certain political demands. They were usually part of the political process and not necessarily aimed at particular political goals in their own right or to solve problems. The German chancellor, Otto Bismarck, introduced social insurance to pacify the increasingly restless working class in Germany and to pre-empt sympathy for the fledgling social democrats. The underlying problem of social risks was hardly on policy makers' minds until well into the twentieth century. Democratization, however, gave rise to higher expectations from the public with regard to governmental responsibility. The consequences of the Great Depression led to an increased study of policy tools as a way to stabilize national economies in an international context. The failure to find and implement an appropriate response to the underlying problems of the Great Depression, and the subsequent rise of fascist governments in Europe and Japan, followed by World War II, increased awareness about the need for more clearly designed and planned public policy for both political and economic stability in industrialized countries.

The postwar period, therefore, saw the development of more professional policy planning across the industrialized world and the rise of public policy as a function of governments. In the socialist Eastern bloc all economic and social activities by governments were based on complex planning procedures in state bureaucracies. A detailed command and control structure was applied to major spheres of life. In the Western world, education, health, social, and pension policies became important fields for public policies. But also industrial and economic policy, the labor market and defense were key activities of central governments. Today, public policies have now been developed for almost all issues and questions in modern societies, ranging from birth control and consumer protection to assisted suicide. There is hardly any aspect of modern living which is not touched upon by public policy. Governments have approaches and positions on virtually everything, with maybe the exception of fashion and cultural tastes. But even here major mass media outlets are in the public sector and shape the consumption of popular culture.

The process of expanding public policy to all parts of life was accompanied by a continuous increase of public spending as part of GDP in the industrialized world between WWII and the mid-1980s, as well as a massive expansion of legislative activities and regulation. This has not been reversed by trends toward privatization and deregulation since the 1980s as both generally require new rules and public supervision. Public spending is stagnating at a relatively high level and today's economies cannot survive without active involvement by governments.

The initial enthusiasm for public policy as a problem-solving device has waned. Trust in governments to solve pressing problems is at an all-time low. The first two decades after WWII were full of hope that an orderly approach toward policies could eradicate the most serious problems in modern

societies. Poverty, crime, drug abuse, low skills, and unemployment were seen as social problems which could be addressed and solved with public policies, if enough knowledge and research were applied to find the best solution. These hopes were dashed for several reasons: first, finding good and appropriate policies to address problems was more complex than initially thought. There is still no one decisive answer as to how to eliminate poverty or crime in open societies. Second, potential policy tools such as redistribution, public spending, social spending, and public work schemes are strongly politically contested in all modern democracies, even where they have been proven effective. The policies of the Nordic countries, which generally have higher levels of education and health and lower levels of crime and unemployment, have rarely exported these policies to other areas as they rely on historical political compromise as well as political institutions which were not in place elsewhere. Third, problem definitions were also subject to political contestation: do governments have to be concerned with consumption patterns of their citizens if it affects their health or wellbeing? Fourth, the size of government as well as its role remains a topic of debate. Fifth, policy effectiveness is only one criterion among many for policy makers and often not seen as the most important one. Some policies are not in line with general values or political preferences. Many studies have shown, for instance, that incarceration rarely prevents reoffending. Nevertheless, the value of punishment for offenders remains a strong reason for imprisonment. Finally, policies very often produce winners and losers and impact the political process (Lowi, 1971). Many insights in public health on healthy eating and exercise have not led to policies as they potentially have a negative impact on producers. For instance, the effects of meat and high levels of sugar and alcohol are well known but have not led to effective policies to curtail them. Banning sugar or even increasing information about sugar in food, meets fierce resistance from major manufacturers, and yet tobacco is under heavy attack by policy makers. In other words, the politics of policy making is still the main factor influencing public policies, not their problem-solving capacity.

Policy Analysis, Policy Science and Evidence-Based Policy Making

The academic field of public policy originated in the United States as a turn toward a scientific analysis of policies (rather than politics, as in the much older field of political science). The growth of government and the increasing expectation that governments were to solve specific public problems, such as unemployment, poverty or housing, led to the emergence of policy science as a new approach (Lasswell, 1951). Howard Lasswell suggested that better policy and better government could be achieved through the intelligent use of the social sciences. Just as no energy grid could work without the knowledge of engineers and monetary policy is based on the advice of economists, all parts of government should be organized in a way that independent scientific knowledge would lead the way toward better policy making. For maximizing the effects of policy science, research should focus on three aspects: to be multidisciplinary, problem solving, and normative (Deleon, 2006; Howlett et al., 2009).

In particular the practical application of policy solutions for real-world problems was the driving force for setting up schools and research institutes dealing with public policy. Government bureaucracies themselves integrated policy sciences into their administrative structures during the 1960s. In the United States by the 1980s virtually every federal office had a policy branch (Deleon, 2006). They aimed to identify objective and scientific solutions for clearly defined policy problems. The task of the policy analyst was to apply the best available knowledge from all disciplines to a given problem in order to find the best available solution and to mediate between science and politics. Dunn defined policy analysis, for example, as an applied social science discipline that uses multiple research methods in a context of argumentation, public debate in order to create, critically evaluate and communicate policy-relevant knowledge (Dunn, 1994). Policy science increasingly built on existing knowledge to propose a policy solution. Eugene Bardach, at the Goldman School of Public Policy at the University of California Berkeley, published *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. The book identifies in detail the steps toward successful policy analysis as follows: (1) define the problem; (2) assemble some evidence; (3) construct the alternatives; (4) select the criteria; (5) project the outcomes; (6) confront the trade-offs; (7) decide; and (8) tell your story. Steps (1)–(6) would be based on scientific research dealing with a policy problem. Only the last two steps would put the policy issue in a wider political context. The persuasion, argumentation and reasoning were an important part of policy science in addition to policy analysis which determined the best approach.

After an initial phase of enthusiasm for large-scale policy planning, such as the War on Poverty in the United States, or initiatives for economic planning in Europe, the 1970s and 1980s saw an increasing disillusionment with the power of scientifically based policies. Many policy suggestions were not taken up by policy makers and the implementation of policies was found to be harder than the policy designers had anticipated. Also many policy fields are complex and particular measures have a knock-on effect. Attempts to fix one problem create several more (Moran et al., 2006). Primarily in the field of economic policy, the recession during the 1970s triggered by oil-shocks demonstrated the limits of economic management and the fact that governments were not in control of the business cycle. Unemployment, poverty, and low educational attainment returned to the policy agenda. In many countries a new debate about the limits of the state and state responsibility emerged. Whereas previously governments aimed to alleviate social problems, they now seem to increasingly accept the nature of societies as unequal and imperfect.

Today, specialist research institutes and universities deal with specific policy concerns in great detail. Research in economics has advanced the understanding of different parts of the economy and influences public policy a great deal. Sociology, demography, public health, engineering, and environmental studies have also provided many policy suggestions to policy makers. Government bureaucracies issue policy research either by assessing potential outcomes of particular policies or evaluating existing policies. Microeconomic research models incite structures and frequently propose public policies to improve regulation, taxation, and spending.

Policy science today is based on a model of demand (by governments) and supply (by researchers) (Deleon, 2006). Rather than integrating research into government bureaucracies and establishing expanded research and planning units, as was foreseen in the 1960s, many governments rely on a host of university and institute-based research for individual policy problems. Independent research output competes for attention in the political arena and is occasionally used in either parliamentary hearings or government programs to improve policy design. The separation of governments and research institutions also allows for a greater plurality in methods and approaches, as well as political preferences. The hope by Laswell that each policy problem would have one best solution, which could be identified by applying the best knowledge, has been replaced by a market place for solutions that offers different pros and cons as well as different distributional outcomes. Expectations toward social engineering as well as economic management today are much more modest than in the early postwar period. Policy science as a science in itself has been in disarray as it could not establish and institutionalize one dominant approach toward problem solving (Klein and Marmor, 2006). At the same time, the role of scientific advice for policy makers has never been stronger.

Evidence-Based Policy Making

Recently a new form of policy science has emerged under the rubric of evidence-based policy making (Radaelli, 1995). Pioneered in medicine, where evidence of effectiveness is now a major criterion for developing guidance, the British government under Tony Blair engaged in further developing a science-based policy prescription. In the UK, evidence-based policy making was supported by grants from the Economic and Social Research Council (ESRC) to the Evidence Network in 1999. The Evidence Network is a center for evidence-based policy and practice. At the core of evidence-based policy making is the systematic testing of the policy's effectiveness. It includes the counterfactual and asks what would have happened without the policy change. Effects of a policy are measured both directly as well as indirectly. The decision in favor or against a policy is then taken in a cost–benefit perspective which aims to assess the net benefit of a policy change.

Evidence-based policy is used in development policy and has been advanced by the RAPID Outcome Mapping Approach (ROMA) through the British Overseas Development Institute (ODI). It is based on six key assumptions (Young and Mendizabal, 2009). They include first, that policy processes are complex and rarely linear or logical. Simply presenting information to policy makers and expecting them to act upon it is very unlikely to work. They are not purely linear as they have various stages that each take varying lengths of time to complete and may be conducted simultaneously. Strategies must be fluid. Second, many policy processes are only weakly informed by research-based evidence. Third, research-based evidence can contribute to policies that have a dramatic impact on lives. Fourth, policy entrepreneurs need a holistic understanding of the context in which they are working. While there are an infinite number of factors that affect how one does or does not influence policy, it is relatively easy to obtain

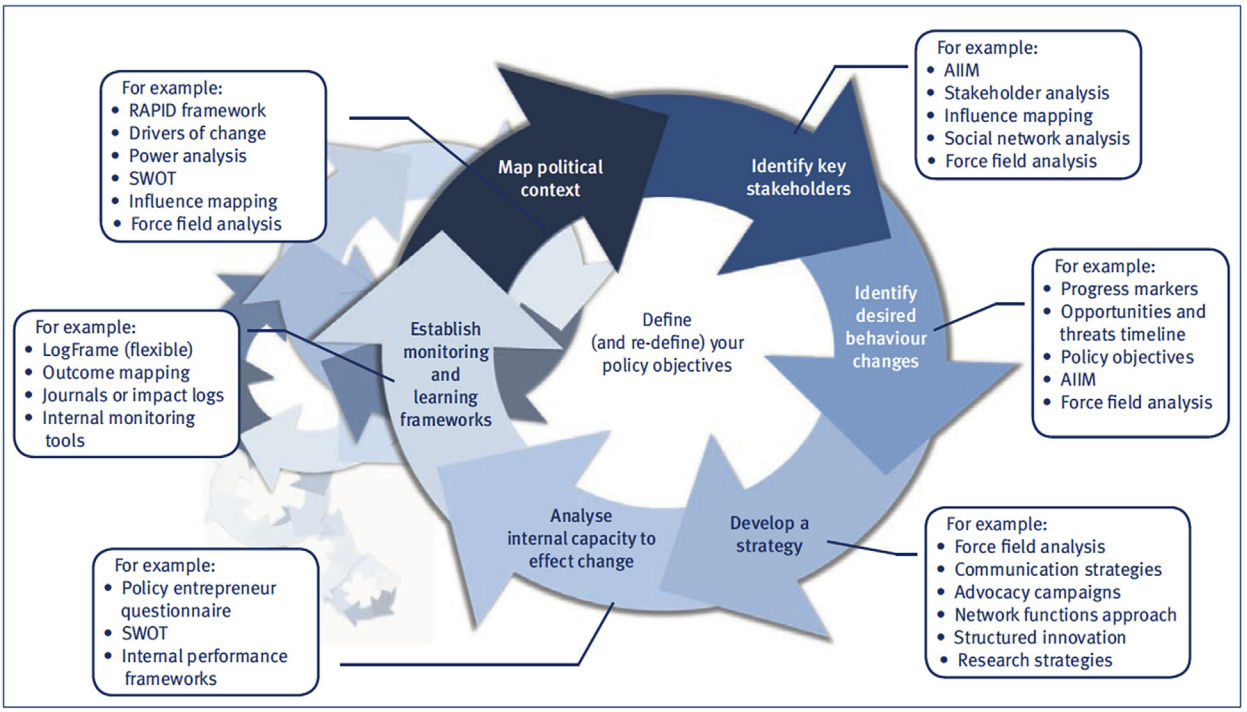

Figure 1 The RAPID outcome mapping approach (ROMA). Reproduced from Young, J., Mendizabal, E., 2009. Helping Researchers Become Policy Entrepreneurs: How to Develop Engagement Strategies for Evidence-Based Policy-Making. Overseas Development Institute.

enough information to make informed decisions on how to maximize the impact of research on policy and practice. Fifth, policy entrepreneurs need additional skills to influence policy. They need to be political fixers, able to understand the politics and identify the key players. Finally, policy entrepreneurs need clear intent – they need to really want to do it. Turning a researcher into a policy entrepreneur, or a research institute or department into a policy-focused think tank involves a fundamental reorientation toward policy engagement rather than academic achievement; engaging much more with the policy community; developing a research agenda focusing on policy issues rather than academic interests; acquiring new skills or building multidisciplinary teams; establishing new internal systems and incentives; spending much more on communications; producing a different range of outputs; and working more in partnerships and networks.

As the key lessons from evidence-based policy making show, understanding the relationship between research, policy and implementation is as complex as the older and more traditional policy science. To develop policy advice for policy makers is not easy or straightforward (**Figure 1**).

Tools of Government

Public policies come in many different forms and sizes as governments have a whole array of instruments at their disposal. Policy problems can be defined and solved in many different ways as the following example can illustrate. The pollution of a public space can be stopped by either convincing citizens not to drop their litter; by introducing a fine for littering; by supplying more bins or by giving firms subsidies for removing and recycling the litter. Depending on the approach chosen, very different groups of the public will be made responsible for solving the problem: the citizens through persuasion or coercion; the city council through the provision of bins or private business.

There are many categorizations and typologies of different tools (Dahl and Lindblom, 1953; Lowi, 1971; Salamon, 1981). More recently the discussion of policy instruments is often centered on an individual's behavioral change to solve policy issues. Through 'nudging' individuals are tempted to live healthier lifestyles and thereby reduce health risks which would lead to higher spending on healthcare (Thaler and Sunstein, 2008). Behavioral economics is a relatively new but very active field for policy analysis.

Nevertheless, the most intuitive and still powerful categorization was developed by Christopher Hood (1986). He distinguished between four different types of resources that governments could employ for achieving policy goals: Nodality, Authority, Treasure, and Organization (NATO) (**Table 1**).

Information-based policy instruments often take the form of public information campaigns which address citizens directly. They can either inform them about new government policies, entitlements, approaches or target behavioral change. Governments can also use information campaigns to change the perception of a problem or to influence the mood about a particular topic. This can facilitate or endorse other public policy changes. For instance, commissions and inquiries can be used to put a policy problem on the agenda and raise awareness for policy change. At first glance, information seems to be a weak policy instrument since it cannot enforce a particular behavior. However, many behavioral changes are based on an awareness of a situation. The perception of a particular problem is often an essential precondition for compliance with other policies.

Authority-based policy instruments are regulatory government tools that cover almost all policy fields. Regulation includes rules, norms, standards, directives and legislation. The violation of regulation is frequently punished by law. It is therefore a strong policy instrument in theory as long as the public authority is in a position to monitor compliance. If compliance cannot be observed, regulation is often ineffective. Compliance costs can also be high and outweigh the benefits. Self-regulation or delegated self-regulation can help to improve compliance when those actors who are the recipients of regulation can influence and participate in the setting of the regulation.

Financial instruments generally are very effective if used appropriately. Subsidies, transfers, grants and taxes have a direct effect on citizens' behavior. Taxation on alcohol and tobacco reduces consumption and improves public health. Subsidies to farmers protect them from the effects of falling prices. Grants to service providers establish particular services. Financial instruments are, however, problematic as they can become politically entrenched. Once introduced, it is very hard to abolish transfers, tax-breaks or subsidies as people become dependent on them.

Organization-based instruments are provisions through the government directly or through agencies under governmental control. Schools, tax authorities, prisons hospitals, and armies

Table 1 Policy instruments

Nodality	*Authority*	*Treasure*	*Organization*
Information collection and release	Command-and-control regulation	Grants and loans	Direct provision of goods and services and public enterprises
Advice and exhortation	Self-regulation	User charges	Use of family, community and voluntary organizations
Advertising	Standard-setting and delegated regulation	Taxes and tax expenditures	Market creation
Commissions and inquiries	Advisory committees and consultation	Interest group creation and funding	Government reorganization

Howlett, Ramesh, Perl, 116. Based on Hood, C., 1986. Tools of Government. Books on Demand.

are mostly under governmental control. For childcare and other social facilities, voluntary organizations can be supported through government agencies. Market creation through voucher schemes is another way of developing services. While public enterprises have largely been privatized throughout the last two decades, public–private partnerships have become more important. Government reorganization occurs in the context of privatization and the creation of new government agencies for supervising privatized utility providers or transport providers.

Very often governments employ a mixture of all instruments in order to implement public policies. For instance, vocational training in Germany is an important policy instrument that simultaneously keeps the rate of youth unemployment low and improves the skills of school leavers. In order to ensure a high rate of voluntary participation by firms to provide training places, the government runs information campaigns, gives tax breaks for hiring disadvantaged youths, enables and fosters firms' self-regulation with regard to the content of training and runs public training facilities for those who do not find an apprenticeship. These different approaches are not chosen indiscriminately but with the expectation that they will reinforce each other. Often a range of policy tools is employed because the effectiveness of each individual tool is not known.

Which policy instrument will be chosen depends on a number of factors. First and foremost, policy choice depends on the broader institutional context of a country. There are different national approaches toward risk regulation, for instance, in the United States and Europe (Vogel, 2012). There are also differences in legal systems which make the use of regulation more or less appropriate. In some countries public ownership of firms or agencies is more accepted (as in France) than in others (for example in the UK). The use of tax-breaks is heavily used in the United States, where public spending is seen more critically, compared with Europe where both taxes and public spending is higher. Large scale changes often develop as an accumulation of small scale changes and major policy changes occur only very rarely in the context of political upheaval.

Politics and Institutional Constraints

Public policies are often path dependent and policy changes occur incrementally (Pierson, 2000). Once a particular problem is addressed by policies, the chances are high that future policies will be similar. As Aaron Wildavsky pointed out: the main predictor for a federal budget is last year's budget (Pressman and Wildavsky, 1984). Once a particular public policy response to an existing problem has been developed, it is very likely that future responses will be at least influenced if not determined by it. A pension reform in the context of a Bismarckian pension system will not easily give up on pension insurance and move to a Beveridge public pension scheme. Rather, policy makers open up an avenue for another noninsurance pension pillar to lift the burden off the insurance-based pension.

This has several reasons: first, citizens' expectations are shaped by previous policies. Second, policy makers themselves have a cognitive bias toward established responses. Third, political actors might have vested interests based on previous policies and fourthly, the costs for changing an established path continually rise even if the costs of an existing and unsuccessful policy might rise too. Peter Hall (1993) has distinguished between three different kinds of policy learning and has acknowledged the high barriers for fundamental policy change. Third order policy change, which involves a shift of paradigm as to how to interpret and solve a problem, occurs only after much experimentation and several crises, after which existing policies have been proven to have failed.

Besides path dependence and barriers for policy learning there are clear institutional effects on policy making. Not all public policies are equally likely in all institutional settings. Rather, different kind of political systems, political institutions and electoral rules have different effects on policy making. These effects are not well understood yet and research on institutional factors determining policies is still in its infancy (Immergut, 2006).

One influential study that looked at the effects of political institutions on public policies is Lijphart (Lijphart, 1984, 1999). The study distinguishes between consensus and majoritarian democracies. Political institutions in majoritarian democracies provide governments with a strong majority and the capacity to implement far-reaching policy changes which cannot be vetoed by other actors. In contrast, consensus democracies heavily constrain governments in their decision-making power. Variables that are used to classify political systems are the number of political parties, coalition governments, federalism and also the role of the central bank and constitutional courts. Lijphart correlates political systems with policy outcomes such as economic growth, inflation, social spending and citizen satisfaction. He comes to the conclusion that consensus democracies, which are mainly to be found in continental Europe, tend to be the kinder, gentler form of democracy (Lijphart, 1999).

A similar approach is pursued by Iversen and Soskice (2006) who use electoral rules as a major factor determining policy outcomes. They also divide modern democracies into two types: those with proportional representation and those with majoritarian electoral rules. In democracies with majoritarian electoral rules, we tend to find two-party systems. This is because losers of elections are heavily punished and small parties cannot survive. In a two-party system, the middle class voter has the choice to vote for a left-wing party which might tax the middle classes to spend on the poor or to vote for a right-wing party which constrains taxation. On balance, middle class voters will vote for the right, leading to lower levels of social spending in countries with majoritarian institutions. The evidence shows that majoritarian democracies have experienced far more years of conservative rule than of left-wing rule, as well as lower levels of social spending. Similarly but with different arguments, Persson and Tabellini, have argued that in proportional electoral systems politicians maximize votes (and not districts as in majoritarian systems) and tend to go for redistributive policies such as social spending or public pensions (Persson and Tabellini, 2002).

Other studies have looked at the capacity of other actors to constrain governments. Ellen Immergut has, for instance, identified veto points in the policy process, which can be used by opposing actors to stop a particular policy. Depending on

the availability of veto points, some policies, such as comprehensive healthcare, might be prevented by special interest groups (Immergut, 1990). Similarly George Tsebelis has identified veto players in the policy process. He argues that policy change will be more difficult when the number of veto players increases (Tsebelis, 2002).

Theories as to the effects of electoral rules and political institutions have been powerful for explaining redistribution and social spending and, therefore, the size of the government and the welfare state (Huber et al., 1993). They have been less influential for other policy fields and policy change. It is, however, currently undisputed that in advanced democracies there is a major difference between countries with majoritarian and proportional representation for social distribution.

Outlook

Our knowledge about the effectiveness of policies and the role of public policy has vastly improved over the last 50 years. The scientific approach toward policies as well as a more systematic evaluation of existing public policies has created a wealth of knowledge of what works. The move toward evidence-based policy making and behavioral economics will improve our knowledge base even more. In a couple of decades, policy makers in advanced democracies will know exactly what the best approach for solving pressing social and political problems will be.

However, the most effective policies are still not the most likely to be adopted. In public health, a healthier lifestyle through exercise and nutrition is frustrated by producers and consumers of unhealthy products. Introducing taxes or banning sugary drinks have been very controversial and politically contested. The fight against the tobacco industry has been ultimately successful in the developed world while more smokers exist today due to an increase in the developing world. In environmental policy, the predominance of individual transport by car undermines global targets of lower carbon emissions. In global environmental policy, free-riding and collective action problems undermine common solutions. In education policy early childhood education still receives the least funding even though experts agree that early investment for children is highly effective. There are many more examples that prove that policy makers do not choose the best policy to address public concerns. Rather, political interests, power constellations and political institutions heavily influence the choice of public policies. The flaws in the democratic policy process which gives other interests, including a general skepticism toward government intervention in society and markets, an avenue for lobbying and intervention, has so far prevented a more systematic science-based policy approach.

Recently, public policies have also been increasingly influenced by private actors. Business and civil society actors are far more active in policy debates than they were before. Iron triangles of policy makers, where bureaucrats and experts jointly dominate a policy field have been replaced by a more fluid policy process. Governments control policies less now than they did in the past. They also use many more market-based policy instruments than before. This is not necessarily based on evidence but more often on general political motivation to include the private sector in the provision of services. Privatization and deregulation have created private interests in previously government controlled policy fields. These have created problems of their own which are still ongoing.

On the whole, the field of public policy is likely to further expand and professionalize. Modern societies are facing many problems which governments are expected to solve. Finance and trade, climate change, global migration, and health threats have evolved into global policy fields as crises can be contagious and are not confined to national borders. The application of existing knowledge to policy problems will remain a major issue for the years to come even though the exact form of policy research and application is likely to change.

See also: Policy Analysis; Policy Cycle; Policy Networks; Policy Process: Nonprofit Institutions in the United States.

Bibliography

Dahl, R.A., Lindblom, C.E., 1953. Politics, Economics and Welfare. Transaction Publishers.

Deleon, P., 2006. The historical roots of the field. In: Moran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, United States.

Dunn, W.N., 1994. Public Policy Analysis: An Introduction. Prentice Hall.

Dye, T.R., 1972. Understanding Public Policy. Prentice Hall.

Hall, P., 1993. Policy paradigms, social learning and the state: the case of economic policy making in Britain. Comparative Politics 25, 275–296.

Hood, C., 1986. Tools of Government. Books on Demand.

Howlett, M., Ramesh, M., Perl, A., 2009. Studying Public Policy: Policy Cycles and Policy Subsystems. Oxford University Press, Canada.

Huber, E., Ragin, C., et al., 1993. Social democracy, Christian democracy, constitutional structure, and the welfare state. American Journal of Sociology 99, 711–749.

Immergut, E.M., 1990. Institutions, veto points, and policy results: a comparative analysis of health care. Journal of Public Policy 10, 391–416.

Immergut, E.M., 2006. Institutional constraints on policy. In: Moran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, United States.

Iversen, T., Soskice, D., 2006. Electoral institutions and the politics of coalitions: why some democracies redistribute more than others. American Political Science Review 100, 165–181.

Klein, R., Marmor, T.R., 2006. Reflections on policy analysis: putting it together again. In: Moran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, United States.

Lasswell, H.D., 1951. The policy orientation. In: Lerner, D., Lasswell, H.D. (Eds.), The Policy Sciences: Recent Developments in Scope and Method. Stanford University Press, Palo Alto, California.

Lijphart, A., 1984. Democracies: Patterns of Majoritarian and Consensus Government in Twenty One Countries. Yale University Press, New Haven.

Lijphart, A., 1999. Patterns of Democracy. Government Forms and Performance in Thirty-Six Countries. Yale University Press, New Haven; London.

Lowi, T.J., 1971. The Politics of Disorder. Basic Books Inc., New York; London.

Moran, M., Rein, M., Goodin, R.E., 2006. The Oxford Handbook of Public Policy. Oxford University Press, United States.

Persson, T., Tabellini, G.E., 2002. Political Economics: Explaining Economic Policy. MIT Press.

Pierson, P., 2000. Increasing returns, path dependence, and the study of politics. American Political Science Review 94, 251–267.

Pressman, J.L., Wildavsky, A., 1984. Implementation: How Great Expectations in Washington Are Dashed in Oakland: Or, Why It's Amazing that Federal Programs Work at All, This Being a Saga of the Economic Development Administration as Told by Two Sympathetic Observers Who Seek to Build Morals on a Foundation of Ruined Hopes. Colegio Nacional de Ciencias Políticas y Administración Pública: Fondo de cultura económica.

Radaelli, C.M., 1995. The role of knowledge in the policy process. Journal of European Public Policy 2, 159–183.
Salamon, L.M., 1981. Rethinking public management: third-party government and the changing forms of government action. Public Policy 29, 255–275.
Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions about Health, Wealth and Happiness. Yale University Press, New Haven.
Tsebelis, G., 2002. Veto Players. How Political Institutions Work. Russell Sage Foundation, New York.
Vogel, David, 2012. Politics of Precaution: Regulating Health, Safety and Environmental Risks in Europe and the United States. Princeton University Press, NJ.
Young, J., Mendizabal, E., 2009. Helping Researchers Become Policy Entrepreneurs: How to Develop Engagement Strategies for Evidence-Based Policy-Making. Overseas Development Institute.

Public Policy Schools

John W Ellwood and Eugene Smolensky, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12563–12567, © 2001, Elsevier Ltd.

Abstract

Since the 1970s graduate education in public policy has come to dominate elite graduate education for public service in the United States. It has done so through an emphasis on the imparting of a variety of rigorous analytic techniques – centered on microeconomics, descriptive and inferential statistics, and political and organizational analysis. Today, however, public policy education and the field of policy analysis faces a number of challenges – a decline in the belief that social problems can be solved through governmental action, the continuing weakness of the political and management parts of the policy curriculum, the potential negative effects of trying to serve an ever more heterogeneous set of student interests and expectations, and the failure of the field to foster the equivalent theoretical breakthroughs that finance theory provided for professional business administration programs.

What Is Policy Analysis?

Before describing public policy schools, it is first necessary to describe the field of policy analysis. Although neither academics nor practitioners of policy analysis have been able to agree on a common formal definition, almost all would agree that policy analysis involves the use of formal reasoning to formulate advice for the solution of problems faced by governmental and nonprofit sectors. Of course, political leaders, decision-makers, managers, and administrators have relied on advisors over recorded time. Scholars and practitioners in many fields, particularly in the professions such as law, medicine, public health, social welfare, and urban planning, claim to incorporate policy analysis as part of their task. How do public policy schools and their graduates differentiate themselves from the mass of advisors then? They do so by using formal analytic techniques to make informed guesses about the outcomes of alternative choices, as they would be implemented on the ground in the real world.

The modern policy analysis movement had its genesis in the United States and has spread very slowly, if at all, to other countries. For that reason, the history of public policy as it evolved into an academic curriculum is a history of the application of policy analysis in the United States.

The staffing effort of the World War II and the subsequent Cold War with the Soviet Union brought the policy analysis movement into existence. Although many historians of the policy analysis movement point to Daniel Lerner and Harold Lasswell's call for a policy sciences field that would apply reasoned advice to all stages of the policy process (deLeon and Steelman, 1999; Lerner and Lasswell, 1951), in practice the field developed with the application of what might be called the optimization sciences – particularly micro-economic analysis and the techniques of operations research – to policy choices. These techniques were first used in this way during the World War II when the need for war mobilization led to a planned economy significantly replacing the market economy. After the war the new techniques found a home at a newly formed home for defense analysis (the RAND Corporation). Ultimately these techniques, and many defense analysts, found a home in offices or ministers of defense and then diffused throughout government. The analytic offices that were set up became an established source of policy advice using the techniques of benefit-cost analysis, cost-effectiveness analysis, operations research, and policy evaluation. These techniques became the heart of the new policy analysis curriculum.

Since the 1970s in the United States, public policy schools have become major centers for graduate education for service in the public and nonprofit sectors in the United States. At many prestigious US universities they have replaced schools and programs of public administration. Their number has grown from under 10 in 1970 to over 60 in 2000. Many existing public administration programs, moreover, have incorporated parts of the public policy curricula, and in some cases have adopted public policy concentrations.

The growth in public policy schools and programs has been associated with the creation of an academic field complete with an association – the Association for Public Policy Analysis and Management – and a journal – *Journal of Policy Analysis and Management*. Of equal, if not greater, importance has been the spread of the intellectual approach of policy analysis as graduates of policy schools have replaced traditional public administrators in high-level staff positions at the US federal, state, and local governments.

As graduates of policy programs have replaced students of accounting and public administration in various governmental staff agencies – particularly budget offices – the policy analysis movement and approach have been institutionalized within the governmental service. This institutionalization was reinforced by the slow economic growth and resulting budgetary constraints of the 1970s and 1980s which forced decision-makers to continually face tighter constraints and trade-offs. New institutions, such as the Congressional Budget Office of the US Congress, were largely staffed by those with policy analysis training obtained either in schools of public policy or PhD programs in economics.

These new staff organizations reflected the norms and values of their dominant staff professions – the graduates of

economics departments and policy analysis schools and programs. Chief among these was a belief that, all things being equal, markets are the best way to allocate society's goods and services. That a decision to use other allocation mechanisms was justified only on grounds of equity or rights (the distribution of benefits and costs and/or the existence of some activities that society believed should not be freely exchanged), or when one could show that markets had failed and that governmental intervention would (not could) lead to a more economically efficient allocation than was the case under the failed market. This reliance on markets separated the new policy analysis movement from the traditional planning approach and built within the policy analysis framework a conservative bias when it came to justification of public sector activity.

The centrality of the norms of economics to the new policy analysis movement led to a belief that analysis of policy requires the laying out of the trade-offs inevitable in the creation and implementation of policy. These trade-offs are reflected in the centrality of two ideas of economics – opportunity costs and the margin. The policy analyst relies on opportunity cost to inform the decision-maker that no option is free, that one can always use the resources for other activities. The concept of the margin allows the economist or policy analyst to tell the decision-maker what he/she can obtain from adding or subtracting one unit of resources. The use of these concepts separates the economist and policy analyst from the political analyst, who often seeks to suppress trade-offs so as to increase the membership of political coalitions.

A third effect was a conscious attempt to present decision-makers with analyses of policy options that make choices concrete by clearly defining problems, setting out possible solutions, and specifying the criteria to be used to judge those solutions so that those decision-makers can evaluate the solutions in a comparative, predictable manner. Although students of policy analysis often wrote about qualitative as well as quantitative indicators, in practice practitioners of the new field found that their comparative advantage vs other advisors was in the use and presentation of quantitative measures – particularly those in a benefit/cost format.

The reliance on analytic techniques and on policy options separated economists and policy analysts from traditional American public administrators. Since the progressive era and the early academic writings of Woodrow Wilson on public administration (Wilson, 1887), students of public administration had operated under the belief that it was possible to separate the creation of policy from its efficient administration. Although by the 1960s most public administrators knew that this dichotomy between politics and policy on the one hand and administration on the other was a myth, much, if not most, of the public administration approach still centered on the efficient running of bureaus and departments. Consequently, to a great extent, the educational core of masters in public administration programs reflected this approach with required courses on the functional skills needed by a line manager – budgeting, organizational behavior, personnel management, and the role of an unelected civil servant in representative government.

What Are Public Policy Schools and How Do They Differ from Other Programs?

The first public policy schools and programs were created in the US in the late 1960s. With considerable philanthropic support, 1971 had founded nine schools or programs in the United States. In a pattern that has continued to this day, several existing programs quickly adopted the new approach and curricula of the policy school movement.

The advocates and sponsors of these new schools and programs sought to apply the new optimizing techniques of policy analysis to the problems facing nations in the late 1960s – particularly the problems facing urban areas. At this time it was still widely believed that government could solve social problems if the right policies were found, created, adopted, and implemented. Thus, the initial educational effort sought to support positive public sector activity to solve problems, a view that has much less support today. The creators of the new programs also sought to do for education toward public service what the changes in business administration education had done for business administration in the mid-1950s – replace a curriculum centered on institutional description and anecdotes with one grounded on analytics.

Similarities Across the Schools

Although one can point to many differences among the original nine policy schools and programs – one had a heavy emphasis on undergraduate education, a second only offered a PhD – viewed from afar their similarities far outweighed their differences. With the exception of the one that only offered a PhD, all focused their graduate education efforts on a Masters degree that was viewed as a terminal professional degree rather than a step toward a doctorate. In most cases they labeled this primary degree the masters in public policy (MPP).

Concentration on Analytics and Professionalism

The basic analytic public policy core set of courses has remained quite similar since the founding of the nine original schools. It centers on a combination of three yearlong sequences of analytic skill offerings – in economics, quantitative methods, and political and organizational analysis – one or more workshops in which students undertake policy analyses for real, as against hypothetical, clients, and a required summer internship. Some programs also required courses in individual and/or policy ethics. Others required courses in legal analysis. Initially the economics sequence centered on microeconomics, although several of the programs required a microeconomics–macroeconomics sequence. Although centered on courses on descriptive and inferential statistics, many of the initial required curricula also included material drawn from decision analysis, game theory, and operations research.

The political and organizational analysis sequences were included so that students trained in technical skills would also graduate from the programs with the political and managerial skills needed to enact, implement, and administer policies. Of all the required sequences in the policy analysis curriculum, political and organizational analysis has remained the most

problematical. From their inception public policy programs rejected the public administration separation of politics from administration. Instead it was assumed that to be an effective analyst one had to be an effective political actor and manager. But unlike economics with its instrumental approach, the academic discipline of political science was largely descriptive and as such was less useful for a professional curriculum. This posed and still poses a great challenge to the public policy faculty charged with the political and managerial core of the policy curriculum. Attempts have been made at a number of institutions to take the findings of political science and public administration research and, through the use of decision forcing case studies, to build a prescriptive and strategic approach to management. Students were challenged to go from answering the descriptive question, "What are the political forces or organizational tendencies which affect the kind of policy we can formulate?" to the more instrumental question, "What do I, as analyst or manager, need to do, as I shape my policy prescriptions or implement the chosen policy, to ensure that the political forces and organizational tendencies do not succeed in frustrating or changing desirable policy in undesirable ways?" (Fleishman, 1990: pp. 743–744).

Many observers have felt that this attempt has been only partially successful, however. To overcome this deficiency some scholars have focused on identifying examples of 'best practice' that might form the basis of improved political and organizational behavior (Bardach, 1994, 2000). Others, criticizing the lack of scientific rigor in this approach, have urged their colleagues to create a body of knowledge in which statements of appropriate political and organizational actions are subject to formal hypothesis testing (Lynn, 1994).

The professional nature of public policy programs is reflected in the fact that, unlike the traditional pattern in which professional masters students simply take graduate courses offered by the economics and political science departments, the required policy analysis courses were and are especially designed for a professional program. The goal in most cases is to recast traditional material in a policy centered, problem solving framework and to present the material in an instrumental rather than descriptive manner. As indicated, political science approaches are largely descriptive, and the new programs, therefore, had to take this material and turn it on its head by presenting it in a fashion that would allow future practitioners to use it as a lever to achieve policy and managerial goals – what might be called political and managerial engineering.

One constant across nearly all the original courses was the focus on domestic policy. Over time international material slowly entered the curricula, but this occurred largely through applications of the analytic skills of the curricula (such as benefit/cost analysis) to international situations, such as whether a water project should be built in another country. In recent years several policy schools have moved into the area of economic development, but almost all have avoided the traditional subjects of international relations and country analysis.

Differences between Public Policy and Public Administration Programs

The new public policy programs sought to replace the existing public administration educational model of public administration. In addition to their focus on analytic skills, the new programs differed from their public administration counterparts in their large required set of core courses, the absence of required concentrations, and the greater number of economists and statisticians on the faculty.

Existing public administration programs had a small set of core required programs (four to five courses) and lasted a year to a year and a half. The new public policy programs were two-year programs with a large number of required core courses (at least a full year's worth at a minimum and in some cases up to two-thirds of a standard two-year course load). In most cases students were required to take these courses in a set sequence. The average public administration program (MPA) offered many specialties or concentrations (with a mean of five in 1984) (Ellwood, 1984). Initially at least, MPP programs almost totally focused on the analytic techniques taught in their large cores. In 1971, in the United States, the average MPA program faculty member's doctorate was either in political science or public administration (21 and 20% in 1971). Economists represented only 9%. In the new public policy schools and programs economists represented a plurality (31% in 1971) and in some cases a majority of the faculty with political scientists being the second most represented discipline (23%) (Ellwood, 1984).

Changes and Challenges to Public Policy Schools

Masters programs in public policy currently dominate graduate education for public and nonprofit service in the United States. Graduates from the programs are now found in key decision-making positions throughout government and the nonprofit sector. However, recent developments provide a challenge to the future of the approach and the field.

The Decline in the Belief That the Public Sector Can Solve Problems

The public policy school initiative was fostered by the belief that better public policies could lead to the solution of public problems – particularly those of the poor and of urban areas. Over the past three decades this belief has attenuated if not disappeared in the United States and elsewhere. To paraphrase Ronald Reagan, in the United States government is now seen as the problem, not the solution. Public service itself has become less attractive. The private sector increasingly has come to be seen as the creator and locus of heroes. Private sector remuneration has grown at a far faster pace than public or nonprofit sector pay.

The effect of all these changes has been to draw graduates of public policy programs away from careers in the public and nonprofit sectors toward jobs in the private sector. One recent study found that between 1973 and 1993, in the United States, the percentage of MPP graduates taking jobs in the public sector declined from 76 to 49% (Light, 1999). At the largest US MPP program this percentage dropped to 28% for 1 year in the late 1990s before rebounding to the mid-40% range. This trend poses this question, will policy programs continue to attract the 'best and the brightest' as they have over the past three decades?

The Difficulty of Management and Leadership Education

As indicated above, from the beginning of the policy school movement it was understood that high technical skill was but a necessary condition for an effective policy analyst. The sufficient condition was the acquisition of political, leadership, and managerial skills. The inclusion of a required sequence of political analysis and management in most programs was meant to impart these skills. But this effort has always been the most problematic endeavor of public policy education. Academics simply know less and have a more difficult time imparting what they know to practitioners in these areas than they do when it comes to technical skills. Moreover, from the beginning, management skills were grafted on to a curriculum focusing on the skills needed by staff analysts. In contrast to masters of business administration programs in business schools, where the entire curriculum (or at least the first year of required courses) was seen as management education, policy schools sought to condense courses in accounting, finance, marketing, and operations into one or two required offerings. The result has never been totally satisfying.

The Trade-Off between Breadth and Depth

The original public policy curricula focused on imparting the skills that would be needed by policy analysts who, acting as advisors to decision-makers, would create policies that would hopefully solve society's problems. Over time policy schools have moved away from this narrow focus on the training of staff to a broader focus of training future managers and leaders. Moreover, the initial narrow focus on domestic policy has also broadened. Today the policy schools seek to serve the needs of students with a wide set of career goals. Some want to become elective officials, some political activists, some social activists, some career civil servants, some managers of nonprofit organizations, and some private sector consultants. The percentage seeking careers in technical policy staff positions has become smaller and smaller. As the programs have tried to meet the needs of this increasingly diverse set of students, they have to some extent lost the coherence and depth of the required set of skills that made up the core of their curricula.

As they have tried to serve future elected officials, political activists, public and nonprofit managers, as well as policy analysts, some policy schools have dropped such topics as decision analysis and operations research from their core set of requirements. Others have moved to the creation of tracks or concentrations, with some providing more quantitative, analytic rigor. Some have approved these changes as responding to the needs of a new set of students while others fear that public policy by trying to be relevant to all will lose its distinctive purpose.

The Intellectual Challenge

Professional education always faces the challenge that it is simply an applied or less rigorous version of the material offered by the traditional academic professions. Frequently this charge is misplaced. The need to apply theory to actual situations often leads to new breakthroughs – sometimes theoretical breakthroughs. Such was the case with the creation of the modern theory of finance in business schools.

It was hoped that implementation theory or principal-agent theory would be the public policy equivalents of the business schools' finance theory. But this has not occurred. Implementation theory has never got beyond its inductive beginnings and principle-agent theory is limited when applied to the multiple-goal, multiple-agent world of the public sector (Moe, 1984). This has meant that public policy education has remained derivative and applied. For the long-term health of the public policy movement within the academy, overcoming this weakness is perhaps the greatest challenge to the field.

See also: Policy Knowledge: Universities.

Bibliography

Bardach, E., 1994. Comment: the problem of best practice research. Journal of Policy Analysis and Management 13 (2), 260–268.

Bardach, E., 2000. A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving, Part III. Chatham House Publishers, New York.

deLeon, P., Steelman, T.A., 1999. The once and future public policy program. Policy Currents: Newsletter of the Public Policy Section of the American Political Science Association 9 (2), 1–9.

Ellwood, J.W., 1984. A Morphology of Graduate Education for Public Service in the United States. Report for the Study of Trends and Innovations in Graduate Education for Public Service. National Association of Schools of Public Affairs and Administration.

Fleishman, J.L., 1990. A new framework for integration: policy analysis and public management. American Behavioral Scientist 33 (6), 733–754.

Lerner, D., Lasswell, H.D., 1951. The Policy Sciences. Stanford University Press, Palo Alto, CA.

Light, P.C., 1999. The New Public Service. The Brookings Institution Press, Washington, DC.

Lynn, L.E., 1994. Public management research: the triumph of art over science. Journal of Policy Analysis and Management 13 (2), 231–259.

Moe, T.M., 1984. The new economics of organization. American Journal of Political Science 28 (4), 739–777.

Wilson, W., 1887. The study of administration. Political Science Quarterly 2, 197–222.

Public Policy: Ethics

Andrzej Klimczuk, Warsaw School of Economics, Warszawa, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There are many ethical dimensions of public policy. Public policy as actions to solve the collective problems includes directly or indirectly making ethical judgments. Public policy takes into account reconciliation of conflicting interests of individuals, groups, and organizations which is based on the values agreeing which influences on the objectives, principles, and styles of policy implementation. Ethical judgments about selecting more and less important as well as more positive problem solutions are present on all stages of policy cycle.

Public policy refers to structuring the government and society actions, which is based on objectified knowledge and is taken to solve the key collective problems. These activities are carried out by using tools such as regulations, public programs, and strategies, stimulating cooperation, arguing, debating, studying, developing positive incentives to adopt expected attitudes, indicators of action effectiveness, evaluation, expertise, as well as by institutions including public offices and networks of commercial and nongovernmental organizations (Howlett and Ramesh, 2003). Analysis of public policy includes both actions that are implemented, those that are not implemented, and those that represent alternative solutions to problems and may be taken.

Analyze and implementation of public policies is a cycle which usually includes the successive steps of agenda setting, policy formulation, decision making, policy implementation, and policy evaluation. There are different styles of action at different stages of policy cycle, which are subject of ethical choices and decisions of individuals, groups, and organizations (policy entities, actors). Public policy can also be understood as functional and technical aspect of governance, whose essence is the control over public resources that need to be redistributed. The choice of mechanisms and logic of redistribution requires an ethical judgment. There are many types of public policies. Among the main are sectoral policies – covering one thematic issue (e.g., health policy, education policy, environmental policy, innovation policy, economic policy); horizontal – crossing various topics (e.g., regional policy, family policy, social cohesion policy); and strategic, redistributive – related to financial transfers from one social group to another.

Ethical Decision Making and Public Policy

Ethics concerns the assessment of activities that we intent to take or do not intend to take. Includes consciously or subconsciously assessment in terms of good and evil, positive and negative actions, justice and injustice, as well as explaining how people define these differences. Choice of public policy is related directly to such ethical judgments during selecting what matters most and what matter not at all in area affected by that policy. Ethics can be understood as a theory of morality or customs and habits. Morality as a social phenomenon is intended to regulate relationship between individuals and social groups. Morality involves views and beliefs actually functioning in society, manifested in people's attitudes toward others, relations between them, and the capacity to cooperate. Morality is a form of social control including evaluation, feedback, norms, sanctions, penalties, awards, role models, and ideals. Ethics is therefore a set of norms and ethical judgments which characterize a society or a certain ethical system (e.g., professional ethics, social ethics, honorary ethics). Ethics is divided into descriptive ethics – descriptive explanations of the morality; normative ethics – evaluative analysis and assessment of morality; applied ethics – applying norms or standards to actual situations, issues; and meta-ethics – philosophical analysis of moral standards and assessments. Ethics includes the theory of values, that is setting ontological status of values, their hierarchy, and their possibility to achievement by human. Based on research ethics as a science formulates sets of rules enabling the implementation of the selected values. Values are all valuable, worthy of desire and choice of things, purposes, and meanings of human actions.

Applied ethics is mostly related to public policy as a moral reasoning and behavior of policy makers, policy analyst, and other stakeholders. Ethical judgments are expressed here by people behavioral choices with social consequences, allocating resources, entitlements, and authorities. In Western tradition, ethics was always related to politics (Capurro, 2005; Wolff, 2011). Basic relationships between public policy and morality can be described as four approaches (Dębowski and Jedynak, 1999). These are the primacy of politics over morality, the primacy of morality over politics, recognition policy and morality as two autonomous areas, as well as the unity of politics and morality. The first approach assumes that political interest is more valuable than the existing moral principles and if whenever breaking the rules and norms is needed this can be done – because 'the end justifies the means.' This position represents, among others, Machiavellian doctrine. Second approach – the primacy of morality over politics – assumes that loyalty to moral integrity is the most important and states that this position shall legitimate any political action. Among representatives of this position were Socrates and Mahatma Gandhi. The third approach is based on the assumption that politics and morality are two autonomous areas. This view assumes that political actions are taken by individuals and groups as far as they moral standards allow

to it. At the same time separation of morality and politics is assumed here as equivalent to no engagement into politics of those persons who are not politicians. This view was formulated by Julien Benda. The fourth position is assuming unity of politics and morality. It is considered here that all policies are evaluated with regard to their moral significance. At the same time the abstract moral standards should be subject to political assessments of effectiveness.

The three most important contemporary theories of ethical decision making in public policy are consequentialism, deontological ethics, and virtue ethics (Sullivan and Segers, 2007; Boston et al., 2010). Consequentialism is related to evaluating human actions by good results, best outcome, the greatest possible increase of pleasure over pain, and the possible satisfaction of preferences in welfare economics. Critics show that this approach countenances sacrificing the interests of a few for the sake of the many, that there is no possibility to estimate all the possible or even probable consequences of a particular action or policy, and that the calculation of consequences is always subjective selection and interpretation of beneficiaries. Deontological ethics points out moral obligations or duties that people should fulfill apart from consideration of consequences. Rights and norms are priority. People should be treated as ends rather than as means to purposes outside of themselves. Deontological approach promotes using published codes of ethics, constructed by professional associations and public agencies. Virtue ethics refers to the sources of people morality in their inner life and character. This approach is focused on certain ideals toward which people should strive to development of their humanity such as honesty, courage, compassion, generosity, fidelity, integrity, fairness, and prudence.

In practice none of these approaches independently provides a complete guide to action under all circumstances and on all stages of policy cycle. Policy makers and analyst may employ different ethical principles and change them to improve policy outcomes.

Ethical Concepts in Public Policy

Number of ethical concepts are used in public policy studies and practice. Particularly important are the concepts of interest, values, principles, and policy intervention styles.

Interest

The ability to create public policies adequate to solve the problems in the country can be understood as the ability to preserve its attributes of sovereignty. In other words, well-prepared mechanisms of public policies give a chance to adapt the country and its society to functioning in modern conditions, that is to compete for resources that will allow a decent quality of life. In this sense, public policy is a set of skills on management available public resources to achieve the highest possible added value for society (Zybała, 2012).

Ethical decision making is a highly complex process in contemporary public life. The difficulty lies here in the fact that public policy is usually practiced in a situation of limited resources. Therefore it requires from government the ethical decisions on selecting activities that should engage these resources for best results (Bellinger, 2007). In addition, public policies are implemented in a situation where different social groups have different or conflicting interests. These groups have different resources, strength, and ability to pursue their interests. Implementation of policies in an efficient manner usually requires the establishment of a consensus among them or even a framework of consensus, which is, for example, public interest, social interest, or national interest. Excessive social conflict may in fact result that the economically weakest groups will hamper (by, e.g., protests, strikes, and social campaigns) important public actions and reforms. Therefore, governments typically run processes such as social dialogue and consultations in various forms.

Interests, as benefits of individuals, groups, or organizations usually includes the values and based on them rules of conduct (Supińska, 2008a). Interests primarily are features that enable and increase the chance of their satisfaction, they tend to give free access to the means of satisfying them, and they cause actual high level of satisfaction. Diagnosis of interests requires a comparison of benefits and losses that may bring a concerned action in a specific situation. Thus, the interests are actions based on reflexivity, in contrast to actions based on impulses, emotions, or feelings. Conflicts of interest arise from the adoption by different individuals, groups, and organizations of different values, principles, needs, and goals, that although may be beneficial for some, can also be bad and harmful to others (Shue, 2006). For example, the decision to build the highway can improve transport of goods, but also degrade the quality of life of communities living close to it, increase the number of fatal traffic accidents, and lead to the destruction of valuable natural habitats; funding of research on human genetic material may raise questions about using their results for medical and military aims; policy on illegal migrants can protect the jobs of citizens in current country, but it can be implemented by rigorous methods that leads to violations of human rights and damage the country's positive image abroad.

Stakeholder analysis in public policy is useful for mapping of the power and interests of different groups (stakeholders) and resulting in a diagram that helps to plan their participation in policy implementation. This includes then, for example, informing and empowering stakeholders (Hermans and Cunningham, 2013). Diagnosis of interests and actors may also serve as a basis or justification for a program of action that will achieve that interest. For example, the program of education development at the same time may aim to increase the intergenerational relations and to protection of nature after analyzing the expectations and opportunities for the involvement of older and younger people. Democratic procedures in the political system allows the articulation of interests. To gain political supporters' interests can be expressed in terms of protecting the interests of vulnerable groups. For example, the interests of our company is the interest of the country and the interests of many small shareholders, who work hard and save money; assistance to the poor is not to summarily grant money (short-term and material interest), but on arousing and sustaining motivation to work (long-term and intangible interest). Public policy therefore requires ethical actions in the interests of the

powerful and the weak, dispersed, and concentrated interests, public and particularistic interests.

Values of Public Policy

Values are what the individual, group, or organization considers as desirable, worth achievement, and action. In public policies primarily evaluated are social relations, the state of meeting the needs, institutions, and relationship between man and nature. The values of social relations includes, for example, justice, freedom, solidarity, equality, and cooperation. The values associated with the meeting the needs are, for example, social security, welfare, health, dignity, and self-realization. The values appearing in evaluation of institutions include self-governance, agency, partnership, common good, and efficiency. The relationship between man and nature determines among other values such as work and ecological balance. Indicators of choices are individual and group behaviors including verbal (e.g., demands, evaluation) and nonverbal (e.g., efforts to gain access to desired goods, emotional states, such as satisfaction or dissatisfaction).

For the public policy crucial are relationships between the values of individuals and groups (Supińska, 2008a). These values can be fully compatible with each other (internalization of the group characteristics by the individuals). They can also be partially compliant when individual accepts some part of the group values and recognizing them as a tool for achieving individual values. For example, some individuals accept freedom in social life since this can guarantee their own freedom. The third type of relationship is inconsistency of group and individual values. This occurs, for example, when individual matured in different sociocultural context and not adapted to the new environment or when is rebelling against the group values. Moreover public policy is examined whether the values adopted by the individuals are in accordance with the values advocated by various entities that are creating and implementing public policy. Thus analyzed whether the values underlying the selected public programs are known for members of the society and whether they agree with the values of individuals and groups. Values in public policy are used for programming, organizing objectives, motivate to action, and justification of adopted legal norms. Knowledge of the society values give possibility to anticipate the arrival of new needs, increasing or decreasing of already occurring needs, or the efforts to meet the needs in a new way, by other means.

Valuing by public policy actors involves the use of assessment criteria related to the broader social, economic, cultural, political, and environmental context. For example, those are criteria on obligatory features (responding to the questions: what is consistent with the values? what corresponds to the norms and standards? what is good for society and for the development of the human person?) or instrumental features (whether achievement of group values will lead to the individual values? whether the implementation of the individual values is not detrimental to the common good?).

There are many sources of values in public policy. These include, among others, ideologies, social and economic doctrines, customs, and attitudes of significant moral and scientific authorities. In modern societies particular importance to ethical debates in public policy is gained by public opinion and mass media (radio, television, Internet), which need guarantee of speech and freedom of the press (Capurro, 2005). Choices of values in public policy and therefore choices of aims, principles, and their implementation programs take place under conditions of conflict of interest and building compromises. Public policies are not, therefore, implementing a rather uniform and consistent value systems.

Values also defines the objectives and principles of public policy. Public policy objectives are determined by groups and organizations during policy formulation. These objectives may be general such as comprehensive development, economic stability and competitiveness, social cohesion, high quality of life, access to nuclear power, adaptation to climate change, and ensuring help for refugees. Under such general goals are specific objectives such as within social cohesion: improving the position of vulnerable economically and socially groups, inequality and poverty alleviation, minimizing the social risk. Fundamentally ethical questions may concern whether these and other goals are actually important and if so how and why exactly they should be achieved (Supińska, 2008a).

Principles of Public Policy

From values are derived the principles of public policy, which can be understood as general guidelines and standards of activities, which should be guided by public policy actors in achieving their goals (Supińska, 2008a). These rules are different depending on the value systems defined in various social and economic doctrines as well as depending on the values adopted and implemented as significant (not always consciously) by public policy actors (which is described as policy style). The principles are an expression of values and directives designed to protect them.

Certain public policy principles are the same as values. For example solidarity as a value describes the desirable attributes of social relations and institutions, and at the same time is the principle used as part of the public policy instruments such as social security systems. For example, self-governance is both a desirable feature of the institutions and the principle of their organization in form of the territorial public administration.

Among the principles of public policy can be distinguished: comprehensiveness, self-limitation, common good, providence, forethought, self-help, solidarity, subsidiarity, participation, sustainable development, self-governance, and multisectorality (Luks, 2005; Bryner, 2006; Supińska, 2008a). These principles can be described as follows:

- The principle of comprehensiveness means that public policy entities should consider all the interests and opportunities to achieve goals and analyze all internal and external factors that may affect their usage.
- The principle of self-limitation means that the government does not have to intervene in all possible areas and sectors of the economy and society, but should choose areas for action.
- The principle of the common good (cohesion) is the realization of individual interests so as to minimize conflicts between them (e.g., environmental conflicts, conflicts between the individual and social interests).

- The principle of providence means that under conditions of limited resources is essential to reduce costs but also to limit intervention, which when is used in excess can inhibit the private and social initiative.
- The forethought principle means that the safety of the individual can not only be the result of benefits from the public, but should also result from individual responsibility for its own future. Forethought is the ability to give up part of the ongoing benefits to meet future needs, and the pursuit of such activities, which minimizes the risk of threats (e.g., saving, healthy lifestyle, organization of insurance).
- The principle of self-help refers to the existence and development of mutual aid of people struggling with similar problems in life as well as support by stronger people for the weak. This assistance usually takes place within a small, informal groups.
- The principle of solidarity is most often understood as a transfer of social risk consequences from individuals to society as well as the superiority of the common interests of society members over the interests of individuals, groups, and social classes.
- The principle of subsidiarity means the adoption of a specific order in which the various institutions undertake interventions such as providing support to the poor. This principle is based on the support in the first instance derived from closest entities (e.g., family, local community), and later help from the state institutions.
- The principle of participation describes an organization of social life, which allows individuals to the full realization of their social roles, and allows groups having equal position in society. This rule specifies the possibility of being a member of different social groups and communities and participate in their actions.
- The principle of sustainable development refers to the integration of political, economic, and social action on preserving the natural balance to ensure the possibility of satisfying the basic needs of both the present and future generations.
- The principle of self-governance is the implementation of such values as freedom and human agency. It is achieved through the organization of social life which guarantees individuals and groups active participation in existing institutions and creating new institutions (informal entities and no-governmental organizations) to better meet the needs and pursue interests.
- The principle of multisectorality refers to simultaneous functioning of different entities of public policy. This include public entities (public administration), commercial entities (private, market), and nongovernmental organizations (civil society, voluntary) that provide resources and services aimed at meeting the needs of society.

These principles do not cover all which are present in the public policy making and implementation. Different types of public policies such as economic and social policies are characterized by different rules. For example, social policies in liberal regimes of welfare state are less focused on solidarity than on forethought and individual savings. While conservative regimes are highlighting subsidiarity and solidarity.

Policy Style

Styles of intervention are derived from public policy values. Policy style concept refers to the assumption that human behavior is important factor in ethical choices and decisions of the policy actors which influence on selecting and using more or less persuasive instruments to achieve policy objectives (Supińska, 2008b). Behaviors are to a large extent rational decisions of people based on the use of their freedom and to pursuit of their life aspirations. However, the behavior of individual groups, for example, employees, students, parents, consumers, criminals, sick, teachers, police officers, and doctors, may differ from the socially expected norms. For example, teachers may focus on work with good students and not promote equality of opportunity for all, consumers can choose cheaper but less healthy food, police officers may be willing to accept bribes. Public policy interventions primarily aim to prevent further dissemination of the behaviors with negative social consequences as well as intervene to maintain and strengthen the already existing positive behaviors, and seek ways to mobilize people to achieve objectives recognized publicly as the correct.

Policy style (style of policy implementation) refers to a set of public policy tools that are adapted to the characteristics of program recipients, which are the causes of undesired behaviors. The choice of intervention style should be preceded by research and public debate which will allow to understand the reasons of undesirable behaviors, overcoming stereotypes, to recognize the social, economic, administrative, political, and environmental conditions of that behaviors as well as to choose relevant objectives and principles for implementation of public policy.

Four basic policy styles may be described (Supińska, 2008b). These are liberal style, caring style, stimulating style, and rigorous style. Liberal style proves to be effective if the only problem with the policy recipients, who have appropriate aspirations and competency, is that they cannot realize their ambitions due to barriers that appeared in their life (such as low education, living in periphery region, lack of institutional capacity). Activities in this style gives freedom to individuals and groups, expressing confidence to their competence. Caring style can be implemented when the main reason for lack of desired behavior is the lack of skills and abilities among policy recipients. Recipients of programs based on this style are treated as persons that need permanent or transitional support until they achieve the desired behavior, skills, and competencies (e.g., help for people with disabilities). Stimulating style assumes that individuals, groups, and organizations can and may keep from negative behaviors or take the socially desired behaviors, but they do not want to (e.g., people choosing to drive they own car than public transport). This style assumes that the desired behaviors can be achieved by offering people some valuable benefits in return. For example, unwillingness to change the organization of business in accordance to the principles of safety and low energy usage can be reduced by proposing a public financial contribution to these investments; the reluctance to change the location of plants can be reduced by lowering taxes. Rigorous style refers to using sets of policy tools that may force the desired behavior. It is assumed here that recipients of the program (e.g., alcohol abusers or

criminals) are able and may but do not want to abstain from condemned behaviors. So that the desired behaviors can be activated upon them only under the pressure of fear from the onerous sanctions (e.g., they do not want to live honestly, but even more do not want to go to jail or pay the penalty).

Policy styles are also varied in different countries due to the various relationships between government and the public (consensus or imposition) as well as due to dominant approach to problem solving (anticipatory or reactive) (Richardson et al., 1982). During formulation of any public programs is reasonable to determine the style which will be fundamental in pursuing of specific objectives. It should also be noted that, in practice, public policies are combinations of the various styles. The different stages of public policy may cover different possible styles. For example on the agenda-setting stage there may be used styles of outside initiation, inside initiation, consolidation, or mobilization (Howlett and Ramesh, 2003). While on the stage of policy implementation may be chosen institutionalized voluntarism, directed subsidization, representative legalism, or directed provision.

There are also different sequences of using policy styles. For example, first may be assumed the citizens' freedom, then the punishment of those who abuse this freedom, then may take place reducing undesirable behaviors of the punished or encouraging to change aspirations. Each style allows not only to achieve short-term objectives of the program but also to contribute to the long-term transformation of personality and attitudes of citizens. This effect is particularly evident in the societies in which there has been long tradition of using rigorous and caring styles by state that also includes interventions into private life (e.g., it affected on low rates of social trust and citizen activities in postsocialist countries like Poland and Hungary).

Ethics in the Policy Cycle

Ethical decision making occurs at all stages of policy cycle (Howlett and Ramesh, 2003; Donahue, 2008). Each stage includes different disputes over values and ethical dilemmas:

1. Agenda setting is a process by which problems come to the attention of governments. This stage includes problem recognition by public, private, and nongovernmental organizations. Important issues are selected by governments into agenda as those that can and should be resolved. Ethical decisions are made during selecting issues because alternative values of policy makers may lead to different choices and objectives. Ethical judgments are also made during considering resource that may be used and stakeholders that interests should be satisfied.
2. Policy formulation refers to how policy options are formulated within government. It starts from policy analysis by which possible solutions to problem are identified and proposed. Crucial part is cost and benefits analysis of material and nonmaterial resources. Ethical judgments are related here to choosing analysis methods, allocating resources, making assumptions on findings, and interpreting them.
3. Decision making refers to the process by which governments adopt a particular course of action or nonaction. On this stage choices between alternative solutions are made by positive or negative decisions or nondecisions. This process allows the preparation of the program as well as clarify the objectives, activities, policy style, and balancing between conflicting objectives. Policy is discussed to be acceptable, legitimated by avoiding bad and doing good. Then adoption of policy is done by law, public programs and strategies, executive orders, rules, and regulations.
4. Policy implementation refers to how governments put policies into effect. Implementation may be organized according to program or may be general and freely structured for policy improvements and flexible actions. Implementation usually includes a series of activities done by variety of policy actors and stakeholders. Here public administrators and servants are confronted with ethical dilemmas about responding to authorities and stakeholders, action according to the professional codes. For example dilemmas may occur when policy goals are too general, when stakeholders are trying to halt or alter implementation, and when different authorities have conflicting expectations.
5. Evaluation refers to the processes by which the results of policies are monitored by state, private, and societal actors, the outcome of which may be reconceptualization of policy problems and solutions. Implemented solutions are evaluated in terms of whether or not they brought the expected results and whether there are any unexpected effects. Policies may not achieve all goals due to ambiguous goals, choosing inadequate action, not taking into account specific conditions, incorrect selection of indicators. Evaluation helps to find reasons why the programs did not respond to the needs of their recipients and explain processes of intervention and outcomes. Ethical dilemmas refer here to choosing appropriate measures, objectively collecting data and interpreting findings. Ethical questions are also posed in responding to stakeholders and deciding on continuing policy, changing it or discontinuing.

Described examples show only some of the possible problems, dilemmas, and ethical decisions which occur in practice of public policy and that may be the subject of further analysis.

Conclusions

Public policy includes directly or indirectly making ethical judgments. Ethical decisions are taken mainly by policy makers by selection and reconciliation of interests represented by individuals, groups, and organizations. Public policy is based on the balancing individual and social values. From those value arise its objectives, principles, and styles of policy implementation and intervention. All choices and decisions in public policy on each stages of policy cycle are ethical judgments because they presuppose that some things are more important than other, that some actions will have positive and other will have negative impact on society. Policy debates can be more productive by using ethical approaches. Ethics enables a systematic analysis of the rules and standards as well as the selection of rational public decisions.

See also: Attitudes, Political and Civic Culture; Civil Service; Corruption: Political and Public Aspects; Decision Making, Psychology of; Economics and Ethics; Leadership; Planning Ethics; Quality of Government; Responsibility: Philosophical Aspects.

Bibliography

Bellinger, W.K., 2007. The Economic Analysis of Public Policy. Routledge, London, New York.

Boston, J., Bradstock, A., Eng, D., 2010. Ethics and public policy. In: Boston, J., Bradstock, A., Eng, D.L. (Eds.), Public Policy: Why Ethics Matters. ANU E Press, Canberra, pp. 1–17.

Bryner, G., 2006. Ethics and public policy. In: Peters, B.G., Pierre, J. (Eds.), Handbook of Public Policy. Sage Publications, London, Thousand Oaks, CA, pp. 433–442.

Capurro, R., 2005. Ethics and public policy within a digital environment. In: General Report on the Activities of the European Group on Ethics in Science and New Technologies to the European Commission, 2000–2005. Office for Official Publications of the European Communities, Luxembourg, pp. 19–25.

Dębowski, J., Jedynak, S., 1999. Mały słownik etyczny (Glossary of Ethics). Branta, Bydgoszcz, Poland.

Donahue, A.K., 2008. Ethics and public policy. In: Berman, E.M., Rabin, J. (Eds.), Encyclopedia of Public Administration and Public Policy. Taylor & Francis, New York, London, pp. 696–700.

Hermans, L., Cunningham, S., 2013. Actor models for policy analysis. In: Thissen, Wil A.H., Walker, W.E. (Eds.), Public Policy Analysis. Springer, New York, pp. 185–213.

Howlett, M., Ramesh, M., 2003. Studying Public Policy: Policy Cycles and Policy Subsystems. Oxford University Press, Toronto, New York.

Luks, K., 2005. Polityka gospodarcza: Wybrane zagadnienia (Economic Policy: Selected Issues). Zakład Wydawnictw Naukowych Instytutu Morskiego, Gdańsk, Poland.

Peterson, S.A., 2008. Values and policy analysis. In: Berman, E.M., Rabin, J. (Eds.), Encyclopedia of Public Administration and Public Policy. Taylor & Francis, New York, London, pp. 2026–2027.

Richardson, J., Gustafsson, G., Jordan, G., 1982. The concept of policy style. In: Richardson, J.J. (Ed.), Policy Styles in Western Europe. Allen & Unwin, London, pp. 1–16.

Shue, H., 2006. Ethical dimensions of public policy. In: Moran, M., Rein, M., Goodin, R.E. (Eds.), The Oxford Handbook of Public Policy. Oxford University Press, Oxford, New York, pp. 709–728.

Sullivan, E., Segers, M., 2007. Ethical issues and public policy. In: Fischer, F., Miller, G., Sidney, M.S. (Eds.), Handbook of Public Policy Analysis: Theory, Politics, and Methods. CRC/Taylor & Francis, Boca Raton, pp. 309–328.

Supińska, J., 2008a. Wartości i zasady polityki społecznej (Values and principles of social policy). In: Firlit-Fesnak, G., Szylko-Skoczny, M. (Eds.), Polityka społeczna (Social Policy). PWN, Warszawa, Poland, pp. 71–80.

Supińska, J., 2008b. Style i instrumenty polityki społecznej (Style and instruments of social policy). In: Firlit-Fesnak, G., Szylko-Skoczny, M. (Eds.), Polityka społeczna (Social Policy). PWN, Warszawa, Poland, pp. 81–91.

Weimer, D.L., Vining, A.R., 1999. Policy Analysis: Concepts and Practice. Prentice Hall, Upper Saddle River, NJ.

Wolff, J., 2011. Ethics and Public Policy: A Philosophical Inquiry. Routledge, Milton Park Abingdon Oxon, New York.

Zybała, A., 2012. Polityki publiczne (Public Policy). Krajowa Szkoła Administracji Publicznej, Warszawa, Poland.

Public Psychiatry and Public Health Mental Systems

Sokratis Dinos, School of Health, BPP University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.A. Talbott, volume 18, pp. 12567–12572, © 2001, Elsevier Ltd.

Abstract

The development of public psychiatry and public mental health systems historically have revolved around three main periods: the rise of the asylums, the decline of asylums, and the reform of public mental health systems through new treatments and a new care philosophy. This article presents the most important shifts that shaped public psychiatry during each historical period. The major reforms of the past two decades will also be discussed, with particular focus on some countries as examples of the latest developments in public psychiatry and public mental health systems.

Introduction

Mental health problems account for over 30% of work absences and according to the World Health Organization (WHO, 2004) at least one in four people meet diagnostic criteria for mental disorder at some point in their lives. This trend has been evident in most surveys around the world. Furthermore, the prevalence of mental health problems is more than twice that of physical health problems including cancer and cardiovascular disease. The importance of public mental health to the social, physical, economic, and overall well-being of a society is tremendous. The role of public psychiatry and public mental health systems in society are therefore of crucial importance. The path of evolution of public psychiatry through history and until today has not been an easy one. The field has had to change direction and reconceptualize itself a few times during its development. As will be discussed, the field of public psychiatry is multilayered and is often difficult to define as its involvement with society and public mental health systems encompasses a number of different facets of engagement including public, political, academic, social, and clinical engagement. The evolution of public psychiatry and public mental health systems will be discussed with an emphasis on particular trends and movements that shifted the field along the way and until today.

The Rise of the Asylums

Psychiatry like any medical specialism did not exist before the end of the eighteenth century. However, mental disorder had always been part of humanity and different societies employed different ways of dealing with it. It was most often the case though that mental illness was a family affair as no other systems were in place. In the absence of family support, local communities were in charge of their own facilities and often people with mental illness were sent to jailhouses or any other facilities designed for criminals and people considered disorderly in general.

The Middle Ages

In England, one well-known institution founded as a priory in 1247 was at the Bethlem Hospital (or 'Bedlam' as it became known) in London, which started providing short-term treatment for mental illness a century after it was founded. Thomas Guy within Guy's Hospital in London established one of the first known psychiatric hospitals in 1726 (Murphy, 1991). In the United States, it was not until 1752 that the first hospital for mental health problems opened in Virginia but it was private and it was not until the 1840s that state or public mental health hospitals appeared. During the 1870s, a census of the mentally ill took place in the French-speaking Swiss canton of Fribourg and revealed that a significant number of people with mental illness were kept restrained in unheated rooms and stables. The lack of provision gave rise to the asylum (Murphy, 1991). However, there are some accounts suggesting the existence of psychiatric hospitals in the middle years in many parts of the world including in Islamic countries during the eighth century, in India during the tenth century, and in England during the twelfth century (Howells, 1975). Although such public provision was evident, this was based on archaic and primitive treatments, which are unacceptable in today's standards.

The Renaissance

Organized systematic public mental health systems started appearing during the nineteenth century and asylums started springing up more frequently. In particular, toward the end of the eighteenth century, in many parts of the world and in particular in the United States and Europe there was a notable shift from unregulated provision to the development of mental health systems. Such systems were more centralized and better managed (Rogers and Pilgrim, 2001). During this time and alongside the asylum movement and the subsequent growth of asylums, a new therapeutic movement sprang up in France and Britain, which expanded the focus from care to include cure from mental illness. The movement was based on the doctrine of what was called the 'moral treatment' and was led by a French physician Philippe Pinel (Murphy, 1991). Moral treatment was based on the belief that mental illness is not a biological disease, which was the accepted belief during this time, but that it is caused by psychological factors. Therefore, a cure was possible if patients were cared for utilizing a number of elements such as physical nourishment and care as well as introducing a conversational element to the care by presenting patients with logical arguments. Pinel's doctrine gained a lot of

popularity for the care of people with mental illness in the asylums.

Asylums provided residential care, which could last for a lifetime and hosted a number of illnesses, which during this time would be classified as mental health problems such as dementia, syphilis, learning disability, epilepsy, and even conditions with moral correlates such as single mothers (Porter, 2002). As a consequence, their growth was immense. For example, the United States (during the 1820s and 1830s) and England (i.e., Lunacy Act and County Asylums Act 1845) saw the first pieces of legislation, where it became compulsory for the state government in the United States and local authorities in England to provide public asylums for people with mental health problems. Similar movements were happening at the same time in most developed countries including France, Italy, and the former Soviet Union. Asylums were the main provision of public mental health systems and public psychiatry well into the twentieth century and until after World War II (WWII). It was not until the 1950s when a number of developments started taking place (e.g., antipsychotic medication, care in the community philosophy), which made it possible for a major change, which gave way to the second era in public psychiatry – the decline of the asylum.

The Decline of the Asylum

There were instances during the nineteenth century where thinkers or movements would try and shift the focus of care provision inside the asylums to care provision in the community. These views gained a lot more popularity during WWII. The psychoanalytic school of thought was also gaining popularity in Europe, and Freud's talking therapy was perceived as a good treatment that could replace the asylum system with a system based on community care. During this time, outpatient treatment in the community was a private practice and was not commonplace in public psychiatry. The polemic of two American sociologists, Belknap and Goffman, on the asylums made the criticism to grow larger. In particular, Goffman (1961) described how the asylum environment promoted adverse effects on the treatment of mental illness by removing individual agency, encouraging apathy, withdrawal, and loss of social and personal care skills amongst patients. In 1970, Ervin Goffman became one of the founders of the American Association for the Abolition of the Involuntary Mental Hospitalization, which provided legal aid to psychiatric patients. The growing skepticism toward asylums was helped by the new drug treatments, the rise of psychotherapy, and the deinstitutionalization movement, all of which played a significant role in the decline of the asylums in the 1960s and also gave rise to antipsychiatry. Antipsychiatry was the belief that psychiatric treatments and practices create more damage than good (Burns, 2006). This view gained further popularity through the writings of psychiatrists such as RD Laing, Theodore Lidz, and Silvano Arieti as well as earlier existential philosophers such as Michel Foucault and later psychoanalytic theorists such as Jacques Lacan and Erich Fromm. A common criticism was that more than any other discipline it involves an unequal power between doctor and patient, with the former being able to exercise power in a number of situations including hospitalization, the law and the courts, and norms of normality. Paradoxically, during the decline of the asylums and for the next 30 years the antipsychiatry views were prominent, while the development of public psychiatry and public mental health systems were also prominent.

The Rise of Psychotherapy

Public psychiatry came to the forefront around the time of WWII. Although the idea of the merits of psychoanalysis started appearing toward the last decade of the nineteenth century, it was before WWII that the psychoanalytic school of thought through Sigmund Freud's ideas began to be adopted as a treatment for mental illness (Shorter, 1998). Psychoanalysis also helped further the idea of moving away from the asylums and treating patients in the clinic. The rise of Nazism and other fascist movements around Europe during WWII brought psychoanalysis to a halt and many practitioners moved to America making the treatment popular. The American Psychiatric Association worked in tandem with the American Psychoanalytic Association and soon the majority of members from the former association became members of the latter association too (Shorter, 1998). It needs to be noted though that psychoanalysis as a treatment in psychiatry was very much part of private psychiatry and a very lucrative practice. Nevertheless, the view that the cause of mental illness was not biological, and therefore could be treated with psychoanalysis, paved the way for the decline of the asylums and the new public psychiatry, which was shifting toward providing care in the community.

Advances in Psychiatric Treatments and the Rise of Public Psychiatry

A number of treatments for mental disorders have been used historically. It was mainly during the past 60 years that psychiatric treatments have seen major developments. A controversial treatment called lobotomy was introduced in 1936. The treatment was used to alter mood and regulate disturbed behavior or disturbed thinking by severing connections within the brain. Thousands of these operations took place in Britain and the United States but due to lack of regulations they started declining dramatically toward the end of the 1950s (Bewley, 2008). Electroconvulsive therapy (ECT) was also a common treatment, which was used around the same time and in particular in the beginning of the 1940s. ECT replaced an earlier treatment called convulsive therapy, which was introduced in Hungary in 1934 by a neuropsychiatrist called Ladislas Meduna. Convulsive therapy was based on seizure induction using a circulatory stimulant drug called Metrazol. It was not until 1937 that the first human experiment of seizure induction through electric shocks took place. ECT involved the use of electric currents to induce seizures in order to shift the brain chemistry responsible for mood and/or thought disorders. The use of ECT declined dramatically during the 1960s, but it is still used today to treat severe forms of mental illness (and in particular severe depression), which do not respond to any other treatment. During the early 1950s and after lobotomy and ECT treatments started declining, new treatments became of physiological nature and targeted

glucose levels and the adrenal system through injecting insulin to induce hypoglycemic comas. Insulin shock therapy targeted the adrenal system, which was seen as responsible for mood regulation and in particular as the cause of psychotic disorders (Shives, 2008). The major shift in psychiatric treatments took place in the 1950 and 1960s as the new psychotropic drugs made their appearance. In particular, in 1950 the first anti-psychotic and mood regulatory drugs appeared such as chlorpromazine and lithium, respectively. During the 1960s the rise of talking therapies (e.g., psychotherapy and psychoanalysis) diversified public psychiatry. However, it needs to be noted that the new scientific evidence during the 1970s onwards was a major driver for the rise in public psychiatry, which had seen its popularity dwindling. The new scientific evidence pointed toward the biological basis of mental illness and in particular the 'chemical imbalance' hypothesis for mental illness such as the 'monoamine hypothesis' for depression and the 'dopamine hypothesis' for schizophrenia. The technological advancements of the 1970s allowed for brain imaging techniques to be developed and make these findings possible and paved the way for treatments to have both a physiological/biological basis and a psychological basis.

The Deinstitutionalization Movement

During the 1960s, advances in public psychiatric and other sciences (e.g., social, health and human sciences), drug treatments as well as the emergence of human rights played a significant role in the provision and treatment of mental illness. This brought about the deinstitutionalization movement, which was prominent in many developed countries (Wright, 1997). In England, a number of asylums started closing and in Italy asylums were abolished altogether. It needs to be noted that Italy was the first country in the world to introduce care in the community and it is also a good example of a successful shift from asylum-based to community-based care. Deinstitutionalization also paved the way for new advances in mental healthcare and in particular for the development of community psychiatry, which sought to provide care and support in the community as this was seen a better predictor of recovery from mental illness. Recovery from mental illness and life adjustments were reconceptualized to integrate a number of public mental health systems including medical and psychological treatment and social support. This brought a shift from prior systems, which were focused on short-term hospital-based care (e.g., asylums and psychiatric hospitals) to long-term, ongoing, and sustained support. The deinstitutionalization movement and the subsequent community-based mental health care and services took place in a number of countries of varying economic development. Following Italy, United States, England, Australia, New Zealand, and the Scandinavian countries were amongst the first countries to implement community-based mental health care.

During and Post-Deinstitutionalization

The deinstitutionalization movement, which was particularly active during the 1970s onwards, reformed public mental health care systems. Such systems, which developed around the world, revolved mainly around providing supporting housing, very often with supervision of care workers on the premises as well as community mental health teams. Such teams comprised a group of specialists including psychiatrists, psychologists, occupational/vocational workers, substance abuse specialists, social workers, and psychiatric nurses. The aim of these teams was to enhance independence, recovery, and rehabilitation and therefore, prevent the negative impact of severe mental illness such as homelessness, substance abuse, hospitalization, and any other negative consequences that arise from mental illness. Examples of community teams around the world included the assertive community treatment (ACT) teams in the United States and the early intervention teams in the United Kingdom. In particular, during the 1970s, ACT became the predominant model adopted by public mental health systems. It was first used in a state psychiatric hospital in Madison, Wisconsin and since then it was replicated and was applied widely within the United States as well as in a number of countries around the world including Australia, Canada, and the United Kingdom.

In the 30 years and until the late 1990s, psychiatric hospitals closed rapidly. However, the closure of hospitals was not accompanied by a surge of provision in community care. This resulted in a lack of suitable provision. The lack of suitable provision with an additional lack of suitable governance and coordination failed to provide patients with the desired social integration. As a consequence, public mental health systems failed to deliver integration of psychiatric patients in the community and provide care for a complexity of needs. Patients were very often discharged too early without suitable post-hospital provisions, were very often marginalized in sheltered environments, and a great deal of patients ended up homeless or in prisons (Fakhourya and Priebe, 2007). A number of new policies were introduced around the world throughout the 1970s and 1980s in order to shape public mental health systems further and provide appropriate social integration to patients with mental health problems. However, the main issue of how to provide social integration was still evident because community services alone could not care for a multiple of needs. For public mental health systems to work effectively, a psychiatric and/or discharged patient needed to have a strong social network (e.g., friends, family), which very often was not the case.

Public Psychiatry and Psychiatric Rehabilitation

The failure of public mental health systems during the deinstitutionalization movement to provide effective care in the community gave rise during the 1980s to the reconceptualization and refinement of service systems. This resulted in the development of the concept of psychiatric rehabilitation. The emphasis was on the treatment and management of the consequences of mental illness and not on the illness. Public mental health systems were reconceptualized to include a number of elements such as treatment, crisis intervention, case management, rehabilitation, enrichment of activities, rights protection, basic support, and self-help. The World Health Organization developed a model of psychiatric rehabilitation by conceptualizing how severe mental illness affects people's lives (see **Box 1**).

Box 1 The psychiatric rehabilitation model

The stages of the impact of severe mental illness:

- Impairment (e.g., hallucinations)
- Dysfunction (e.g., lack of vocational and social skills)
- Disability (e.g., unemployment)
- Disadvantage (e.g., poverty, discrimination, stigmatization)

By conceptualizing the different stages that severe mental illness impacts on, the overarching aim was to provide patients with help to achieve illness management as well as social and vocational skills. Interestingly, psychiatrists and public psychiatry generally were largely pushed away from shaping up and providing psychiatric rehabilitation. This was partly because of the rising of antipsychiatry feelings amongst service users and part of this was a result from the failed promises of the deinstitutionalization movement. Psychiatrists also found themselves as part of multidisciplinary teams with a role that was often unclear and undefined. For example, public mental health authorities were often utilizing other specialists in providing medical treatments, and psychiatrists were seen as unsuitable in providing patients with the necessary skills for social and vocational integration.

The Reform of Public Psychiatry and Public Mental Health Systems in the Past 20 Years

The deinstitutionalization movement brought a number of policy changes from the 1960s onwards and shaped public psychiatry further. More recently, international and European policy was also introduced such as the United Nation's "Principles for the Protection of Persons With Mental Illness and the Improvement of Mental health Care" in 1991 and "The Charter of Fundamental Rights of the European Union" in 2000, both of which emphasized the need for community care for people with mental health problems (Knapp et al., 2011). The problems in the implementation of care in the community of the past 50 years pushed the policy agenda further and introduced the idea that people with mental illness need more than symptom management and relief. This brought about the latest developments in public psychiatry and public mental health systems, which revolved around the concept of recovery and the so-called 'third wave' of psychological treatments (e.g., cognitive behavioral therapy).

Recovery Movement

In the early 1990s and as asylums continued closing, the recovery movement started gaining popularity, especially amongst service users. The recovery movement became popular both in the United States and Europe and underlined the need to shift the focus from symptom management and service use to indicators of well-being and service user empowerment. Public mental health systems acknowledged that people with mental illness have multiple needs such as social (including removing the stigma attached to mental illness), vocational, residential, educational, etc., which cannot be met through hospitalization or by just treating illness symptoms (Anthony, 1993). Although the identification of these needs (e.g., vocational, residential, social rehabilitation) was not new, the recovery movement added a new dimension to the conceptualization of these needs. The movement realigned the relationship between public psychiatry and service users and emphasized the role of the latter in being responsible both for having control over their treatment plan and for their own recovery process. The concept of recovery within public mental health systems is not fully researched and therefore its application is still under investigation. Following the care in the community and the recovery movement, it is a common process in many countries for service users along with their families to influence their treatment and recovery process by being involved in the decision-making process.

Latest Initiatives in Public Psychiatry and Public Mental Health Systems

The past decade has also seen a number of further initiatives in public psychiatry and public mental health systems across the world. WHO (2007) outlined the need to further develop public or community-based mental health systems based around a number of key components (see **Box 2**).

A review of public mental health systems around the world through data collected by international organizations and published country profiles was developed by WHO (2009). According to the report, most countries and in particular middle- and low-income countries do not offer adequate mental health care. For example, approximately 28% of countries have a budget for public mental health care, many countries have notable shortages of staff in mental health-related services and many systems still rely on psychiatric hospitals for care provision. Nevertheless, the last decade has seen the emergence of a number of major initiatives in the area of public psychiatry to reform public mental health systems further. Looking at some country examples, it is evident that public psychiatry today involves a number of activities and roles extending from providing a wide range of treatments for mental health problems to educating the public, designing stigma reduction campaigns and programs, working closely

Box 2 Key components of public mental health systems (WHO, 2007)

Key components of public mental health systems:

- Enable the provision treatment for mental illness in primary care
- Enable wide access to psychotropic medication
- Enable the provision of care in the community
- Educate the public (e.g., remove the stigma of mental illness)
- Incorporate communities, families, and service users in decision-making
- Develop mental health-related policies, programs, and legislation
- Develop human resources and create links with other sectors
- Support research- and evidence-based practice and monitor public mental health

with the government and the charity sector to influence policies and legislation, as well as working closely with patients and carers. Examples of such initiatives can be seen through the latest public health system reforms in the United States and the United Kingdom.

Recent Examples of Public Mental Health System Reforms: the United States and England

In recent years in the United States some leading professional bodies in the field of psychiatry such as the American Association of Community Psychiatrists (AACP), the National Council on Community Behavioral Healthcare, and the Community Behavioral Health Organization, to name a few, have driven the development of public psychiatry and public mental health systems. For example, members of the AACP developed the Psychiatric Leadership Development Program in 2007, which aims at providing further training for psychiatrists. Training involves enhancing knowledge of public mental health systems and skills in psychiatric-related administration as well as coaching psychiatrists in organizational leadership. Similarly, in 2008 AACP developed a new set of guidelines for developing advanced postgraduate training fellowships in public and community psychiatry. The elements of the fellowships were divided into core elements, which included academic, research, teaching, clinical work, and staff supervision elements and elective ones, which included systems management, community advocacy, recovery-oriented services, and cultural competence. Since AACP developed these guidelines there have been a growing number of US universities providing training in public psychiatry.

Similar activity has been taking place in England. In particular, the National Service Framework for Mental Health (NSF-MH) was published by the Department of Health in 1999 setting a 10-year plan for reform in public mental health systems. Since then the NSF-MH has been refined and extended to include additional policy documents and guidelines from the National Institute for Health and Care Excellence (NICE). The aims of the report were structured around seven standards, which public mental health systems needed to incorporate. These standards also cover funding, service provision, and staffing. The specifications covered the whole spectrum of public mental health provision including both hospital- and community-based care as well as services of both primary and specialist care. The guidance also aimed at (1) program development to reduce stigma and discrimination associated with mental health and (2) improving access to mental health services. As part of the NICE guidance, the Improving Access to Psychological Therapies (IAPT) program was launched in 2005 and aimed to improve access to treatment for depression and anxiety. The aim was to increase well-being amongst the working population and reduce the cost of providing incapacity benefits. When IAPT was first launched, it only provided cognitive behavioral therapy to the working age population but since then it has increased its accessibility to all age groups as well as its availability of therapies. Additional therapies include counseling, couple counseling, dynamic interpersonal therapy and interpersonal psychotherapy. This brought further reforms to public psychiatry as IAPT services require and provide training in psychological therapies, which traditionally were not part of the psychiatric training.

The Future

The latest initiatives in public psychiatry and the subsequent reforms in public mental health systems suggest that the field has been going through significant reconceptualization and has become a lot more multifaceted encompassing public, political, and social engagement. As a consequence, there is some lack of clarity with regards to how public psychiatry can promote the interaction of all these different facets and/or how these facets fit into the role of a public psychiatrist. It seems plausible to suggest that in the next decade public psychiatry will be adding clarity and defining its boundaries and place in the field of mental health further. Furthermore, the progress of the past decade in public psychiatry suggests that there is a variety of new training opportunities (e.g., psychotherapies, governance, management, public engagement, etc.), which expand beyond the psychiatric training received during specialization. This will help public psychiatry to fulfill the expectation of a multifaceted field, which has a significant role to play in society and which extends beyond the clinic and/or the academia.

See also: Antidepressant Drugs; Behavior Therapy: Psychiatric Aspects; Cognitive Behavioral Therapy; Deinstitutionalisation in Low and Middle Income Countries; Deinstitutionalization, Psychiatry of; Mental Disorders, Epidemiology of.

Bibliography

Anthony, W.A., 1993. Recovery from mental illness: the guiding vision of the mental health service system in the 1990s. Rehabilitation Journal 16, 11–23.
Bewley, T., 2008. Madness to Mental Illness: A History of the Royal College of Psychiatrists. RCPsych Publications, London.
Burns, T., 2006. Psychiatry: A Very Short Introduction. Oxford University Press, Oxford.
Fakhourya, W., Priebe, S., 2007. Deinstitutionalization and reinstitutionalization: major changes in the provision of mental healthcare. Psychiatry 6, 313–316.
Goffman, E., 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Penguin, England.
Howells, J., 1975. A World History of Psychiatry. Bailliere Tindall, New York.
Knapp, M., Beecham, J., McDaid, D., Matosevic, T., Smith, M., 2011. The economic consequences of deinstitutionalization of mental health services: lessons from a systematic review of European experience. Health and Social Care in the Community 19, 113–125.
Murphy, E., 1991. After the Asylums: Community Care for People with Mental Illness. Faber and Faber, London.
Porter, R., 2002. Madness: A Brief History. Oxford University press, Oxford.
Rogers, A., Pilgrim, D., 2001. Mental Health Policy in Britain, second ed. Palgrave Macmillan, New York.
Shorter, E., 1998. A History of Psychiatry: From the Era of the Asylum to the Age of Prozac. Jossey Bass, New York.
Shives, L.R., 2008. Basic Concepts of Psychiatric Mental Health Nursing. Lippincott Williams & Wilkins, Philadelphia.
World Health Organization, 2007. Monitoring and Evaluation of Mental Health Policy and Plans: Mental Health Policy and Service Guidance Package. http://www.who.int/mental_health/policy/services/14-monitoring%20evaluation_HKprinter.pdf (accessed 15.08.14.).
World Health Organisation, 2009. Improving Health Systems and Services for Mental Health. http://whqlibdoc.who.int/publications/2009/9789241598774_eng.pdf (accessed 15.08.14.).
Wright, D., 1997. Getting out of the asylum: understanding the confinement of the insane in the nineteenth century. Social History of Medicine 10, 137–155.
WHO, 2004. Prevention of Mental Disorders: Effective Interventions and Policy Options. WHO.

Public Reason

Gerald Gaus and Chad Van Schoelandt, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.F. Gaus, volume 18, pp. 12572–12578, © 2001, Elsevier Ltd.

Abstract

This article discusses various theories of, and problems for, publicly shared reasons and reasoning. The Enlightenment ideal that a common basis of reasoning, such as scientific method, faces a number of modern challenges that are discussed. These challenges include those of cultural diversity, the view in philosophy of science that different epistemic values must be weighed as the data always underdetermines theory, and the reasonable moral pluralism that arises from what Rawls calls the 'burdens of judgment.' Attempts discussed include appeal to common core of epistemic norms, the prerequisites of mutual understanding, shared liberal-democratic commitments, and convergence on social norms with diverse justifications.

Theories of public reason are responses to a distinctively modern worry: that the free use of human reason, rather then producing consensus in belief and moral conviction, leads to disagreement and conflict. The worry stems from a variety of sources: social science, the philosophy of science, and the moral pluralism of modern societies. This article begins by examining the lines of inquiry that have led many to question whether all fully informed and fully competent inquirers would come to the same conclusions about science and morals. According to some philosophers and social scientists, the very concept of reason has broken apart, leaving different individuals and groups with different conceptions of rationality, and so of reasons for belief and action. The article then considers the different ideals of public reason that have been advanced to overcome worries about the 'fragmentation of reason.'

The Diversity of Reason

The Enlightenment

According to the conception of reasoning that dominated the European Enlightenment of the seventeenth and eighteenth centuries, reason is inherently universal, and so shared and public. On this Enlightenment conception, reason is a shared capacity of all human beings, and norms of good reasoning are universal. Any premise p that is true for one person is necessarily true for all others; if the inferential rule '$(p\&(p\rightarrow q))\rightarrow q$' is valid for one person, it is necessarily valid for all. The true and valid results of one person's reasoning are necessarily true and valid for all. According to this standard Enlightenment view, the use of common human reason produces consensus not only on scientific beliefs, but also on moral and political opinions. To be sure, people might disagree on matters of science or ethics, but that would be due to mistaken beliefs or irrationality: some have arrived at the wrong answer. The process of Enlightenment was the only remedy for this – the increasingly better use of reason to uncover truths about the natural and the social world. The ideal or model was that of Newtonian physics: just as our common reason had uncovered the laws of matter and motion, so too could it be expected to uncover the laws of human nature, society, morals, and politics. Each field was awaiting its Newton.

Relativism and Social Science

It is a familiar story how the systematic study of cultures with radically different norms from those of Europe gave rise to cultural relativism. When confronted with the very different mores of the Zuñi and Kwakiutl Indians and Dobu Islanders of Papua New Guinea, Ruth Benedict (1934) endorsed a relativistic view – their norms were justified for them, just as Western norms were for Europeans. As anthropologists began to study non-European cultures in more depth, and became increasingly more sophisticated and self-reflective in their techniques, disputes arose about the proper method for anthropologists to interpret the belief systems of other cultures. For example, was a belief in magic or a deity to be understood as simply a false belief on which the 'natives' relied? John Stuart Mill (1974[1843]: pp. 766–767) believed that an investigator could best understand the magical beliefs of other cultures by attributing to them erroneous beliefs and invalid inferential rules. Others, adopting a principle of charity, have insisted that the best interpretation of a culture minimizes the number of false beliefs attributed to its members (Davidson, 1973). Thus the best interpretation – which makes sense of the 'native's' metaphysical theories, religious convictions, and their beliefs about nature – might seek to show that a belief in spirits is, after all, rational given their world view. Thus, the second step in the relativist attack on the Enlightenment's ideal of reason was to endorse relativism concerning what beliefs are rational in different cultures. The last step in the relativistic project was to apply tolerance to the idea of reason itself. Can Western anthropologists properly interpret other cultures if they apply their Western conception of reason in their interpretation? Is the very idea of reason culturally relative?

Pluralism and Indeterminancy

Social science thus provided one source of the 'fragmentation of reason.' Of course, there was, and still is, lively debate within social scientific and philosophical circles about whether any such relativistic view of reasoning is justified, and if so to what extent individuals employ different, but equally good, norms of reasoning. Nevertheless, some form of cognitive relativism is widely accepted. Concurrently, developments in the philosophy of science were challenging the idea of common, human,

reasoning in the very citadel of the Enlightenment – science itself. Whatever doubts may be entertained about the application of Western reason to other cultures, or departures from expert norms by ordinary reasoners, surely science is still the model of rational discourse tending to convergence of opinion leading to the truth. To many, Thomas Kuhn's analysis of scientific practice demonstrated that, rather than a shared project of pursuit of the truth, judged by rules of reason held by all, scientific inquiry takes place within 'paradigms' that determine not only what constitutes research problems but what constitutes success in a research program. Within a group of scholars sharing a paradigm, the Enlightenment's ideal of common norms of reason and consensus on the truth will be approximated; and insofar as 'normal science' is dominated by a single paradigm in each field, there often is a common mode of reasoning. But in times of scientific crisis, when paradigms compete, some have insisted that there is no rational, impartial way to adjudicate among the competing theories.

The crux of Kuhn's analysis is that because data underdetermines theories, scientists draw on values in adjudicating among theories. "To a greater extent than other sorts of components of the disciplinary matrix," writes Kuhn, "values may be shared by men who differ in their application. Judgments of accuracy are relatively, though not entirely, stable from one time to another and from one member to another in a particular group. But judgments of simplicity, consistency, plausibility, and so on often vary greatly from individual to individual … Even more important, in those situations where values must be applied, different values, taken alone, would often dictate different choices. One theory may be more accurate but less consistent or plausible than another" (Kuhn, 1970: p. 185). The crucial claim is that there is no uniquely rational way to order these various desiderata – simplicity, consistency, plausibility – and different orderings endorse different scientific theories upholding competing truth claims. Equally well-informed scientists employing their reasoning in perfectly legitimate ways can arrive at different judgments about what is true. Fred D'Agostino (1996, 2003) generalized the analysis. In any field of inquiry, when there exists both a plurality of criteria and no impartial way to order the criteria, the criteria are apt to be indeterminate in their application. Different orderings produce different outcomes – reasons for belief. According to D'Agostino, the criteria only yield a determinate result when an ordering is provided, but impartial reason cannot give us that. Much of postmodernism is based on this key idea. Reason is inherently perspectival because of the inherent plurality of relevant considerations for which there is no rational, impartial ranking.

Reasonable Pluralism

Kuhn's analysis of scientific practice and D'Agostino's generalization rely on the idea of indeterminancy: reason underdetermines choice between various theories and perspectives. A more modest view maintains that our powers of reasoning are inconclusive on many complex matters of science, morality, and politics. Claim C is characterized by reasonable pluralism if some perfectly reasonable agents do, while others do not, have good reasons for accepting C. For a controversy to be characterized by reasonable pluralism, it is not simply the case that people actually disagree about the merits of C, but that it is reasonable for one person to assert C and another not-C. This is a far less radical claim than was examined in the last section: it is not argued that the question 'C or not-C?' is inherently indeterminate but only that present beliefs about C are inconclusively justified. According to John Rawls, our disputes about value seem subject to reasonable disagreement because our understanding of what is good and valuable is especially subject to what he has called the 'burdens of judgment.' According to Rawls, reasonable judgments so often are at odds because

1. the evidence is often conflicting and difficult to evaluate;
2. (as in Kuhn's example) even when we agree on the relevant considerations, we often weigh them differently;
3. because our concepts are vague, we must rely on interpretations that are often controversial;
4. the manner in which we evaluate evidence and rank considerations seems to some extent the function of our total life experiences, which of course differ;
5. because different sides of an issue rely on different types of normative considerations, it is often hard to assess their relative merits;
6. in conflicts between values, there often seems to be no uniquely correct answer (Rawls, 1993: p. 57).

Because these matters are so complex and uncertain, different people will reach different, competing, credible, or reasonable conclusions. In the face of complexity, our powers of reasoning do not produce convergence.

Conceptions of Public Reason

Public Rationality and Public Reasons

Important movements in social science, the philosophy of science, and moral and political philosophies thus have combined to cast grave doubt on the Enlightenment's confidence that the free exercise of human reason leads us to agree. In the face of these challenges to the ideal of common human reason, theories of distinctly public reason have been advanced to show how people can share reasons (Korsgaard, 1996: Chap. 10). However, this idea is ambiguous. It may constitute a response to relativistic social science (see Section Relativism and Social Science), which challenges the very idea of a shared conception of human reason. Understood thus, the ideal of public reason is an ideal of public rationality, or a shared conception of rationality. The challenge posed by reasonable pluralism (see Section Reasonable Pluralism) is more modest. Advocates of reasonable pluralism maintain that even if we apply shared standards of reason, our deliberations will not yield shared reasons for belief. In a weaker sense, then, the ideal of public reason is an ideal of public reasons, or how we can arrive at shared reasons to believe and act. A theory of public reason may advance a theory of public rationality and/or public reasons.

Four main conceptions of public reason stand out: (1) epistemic, (2) shared practice, (3) political, and (4) convergence.

Epistemic Theories of Public Reason

In philosophy, 'epistemology' refers to the theory of knowledge or, more broadly, the theory of justified belief. Epistemic

theories of public rationality, then, maintain that at least some common norms of reasoning can be justified to everyone. The Enlightenment view was that all valid norms of reasoning can be justified to everyone – reason was inherently public. Epistemic theories of public rationality need not, however, advance such a sweeping claim. Importantly, an epistemic theory of public rationality may adopt a modest version of the 'relativism of reasons' (see Section Relativism and Social Science), according to which in some cases person Alpha may have a fully justified norm of reasoning N while Beta accepts a fully justified norm of reasoning M, where it is the case that N and M are inconsistent. In cases where only N and M are relevant to deliberation, a relativism of rationality will manifest itself; Alpha may have reasons based on N that Beta will deny on the basis of M, and both will be justified. However, to establish a thoroughgoing relativism of reasoning that undermines the possibility of public reason, it is not enough to establish that this can sometimes occur; it must be shown that no common norms of rationality can be justified to all – that it is always (or at least usually) the case that for any justified norm of reasoning held by Alpha, some person Beta will be justified in accepting a competing norm. That is the sweeping relativistic claim rejected by epistemic theories of public rationality (Gaus, 1996).

In reply to the anthropological argument (see Section Relativism and Social Science), epistemic theorists of public reason can embrace Martin Hollis' view that an anthropologist must build up an interpretation of an alien language by locating a 'bridgehead.' The assumption underlying the bridgehead strategy is that others generally perceive what we perceive and tend to say about it the sorts of things we would say. Relying on these assumptions, the anthropologist begins by translating everyday, basic perceptual sentences such as 'Yes, this is a brown cow' and 'No, it is not raining right now.' Now, Hollis insists, this bridgehead not only includes translations of such basic beliefs, but basic logical rules, since "what a sentence means depends on how the beliefs which they express are connected, and that to justify a claim to have identified a belief one must show the belief is connected to others." The point, then, is that the context that sets the stage for the possibility of relativism – that we confront others, different from us but of whom we can make some sense – presupposes widespread shared norms of inference as well as beliefs. Thus, the very possibility of mutual intelligibility sets a limit on the extent to which we can understand others as employing cognitive processes different from our own.

Shared Practice Theories of Public Reason

Epistemic accounts of public reason uphold our ability to reason together because they insist that we do, indeed must, share basic rules of inference. In contrast, shared practice accounts of public reason conceive it as arising out of a shared social life and/or shared discourse (*see* Shared Belief: Philosophical Aspects). A number of otherwise diverse philosophers such as Ludwig Wittgenstein, Jürgen Habermas, Richard Rorty, Susan Hurley, and Philip Pettit all point to the idea that reasoning is a social phenomenon. In order to be a deliberator, one must be a member of a community in which one's beliefs and rules are intelligible to others. One comes to understand what is a justified or unjustified belief, and what is a good or a bad way of reasoning, by sharing the norms and standards of communities, including communities of inquirers. Thus, some insist that "ultimately, there is only one criterion by which beliefs can be judged valid, and that is that they are based on agreement reached by argumentation" – we converge on them (this is Habermas's (1990: p. 14) gloss on Rorty's 'particular version of discourse theory'). Habermas draws an intimate connection between mutual understanding (*Verständingung*) and agreement (*Einverständnis*). It would seem that to understand others one must in some way arrive at intersubjective agreement about norms of rationality and good reasons. Consequently, the phrase 'public reason' becomes almost redundant: something can only be identified as reason because it is the object of public, interpersonal, agreement.

Wittgenstein and his contemporary followers such as Hurley (1989) and Pettit (1993) suggest a broadly similar view. Reasoning itself presupposes a 'community' – common rules of thought. Wittgenstein (1958) argued that one could not have a private rule that one only followed once – for in such a case we could not know whether the rule was actually followed. More generally, it has been argued that one cannot have a purely private rule, for one needs to be able to distinguish when one has correctly followed the rule and when one has gotten it wrong. But, assert Wittgensteinians, the difference between correct and incorrect applications of a rule cannot be located in individual belief systems, but only by the individual comparing one's application of the rule to those of one's fellows: each must look to others to check and correct his or her rule-based performances. Since thinking is necessarily rule driven, the analysis applies to reason itself. Thus, it seems that the very idea of rule following – and, so, of reasoning – supposes agreement with others. If so, the search for reason is a search for agreement.

In one respect, this view, somewhat surprisingly perhaps, returns to the basic idea of the Enlightenment – all reason is public reason. The Enlightenment was confident in asserting this because of the widespread faith in a universal, common, human faculty of reason that could accurately represent reality, whereas shared practice theorists see reason as arising out of a shared public life. To put it crudely, rather than reason producing agreement, agreement generates reason. Thus, the publicness of such reason, unlike that of the Enlightenment's, is consistent with a cultural relativistic conception of rationality. The public, intersubjective agreement, reached by one culture – which is constitutive of their very conception of rationality – may be very different from that upheld by another.

Because shared practice accounts of public reason stress the constitutive nature of public agreement to the very idea of reason, and because public agreement may be promoted by political institutions, more radical shared practice theorists such as D'Agostino give political institutions a constitutive role in determining what reason is (1996: Chap. 9). This is a radical view indeed – however, it is not entirely novel, perhaps going back as far as Hobbes. For Hobbes (1948[1651]), the conflict in the state of nature arises from conflicting private judgments; people's private reasoning yields conflicting judgments of right and wrong, as well as matters of fact, and

this leads to the less intellectual conflict that characterizes the state of nature. Hobbes's solution is to appoint an 'arbitrator.' This 'judge,' says Hobbes, provides 'public reason' to which private reason 'must submit.' The arbitrator proclaims what each has reason to do, and so defines a single, coherent conception of reason. Because, in Hobbes' view, we have authorized the judge (i.e., sovereign) to define public reason for us, the sovereign's pronouncements constitute a shared public reason on which there is agreement.

Public Reason qua Political Reason

D'Agostino's radical view allows the political to shape the rational. A more modest proposal with which this radical view is easily confused maintains that the notion of public reason is a purely political idea, in the sense that it concerns only how we reason together about politics. This understanding of public reasons lies at the heart of what is commonly known as 'political liberalism.'

Rawls's political liberalism is a response to the problem posed by reasonable pluralism (see Section Reasonable Pluralism). For Rawls, the crucial problem is that citizens in democratic societies entertain a plurality of what he calls reasonable 'comprehensive doctrines' – overall philosophies of life centered on religious, philosophical, or moral beliefs. This pluralism seems a permanent feature of modern societies. "Political liberalism assumes that, for political purposes, a plurality of reasonable yet incompatible comprehensive doctrines is the normal result of the exercise of human reason within the framework of free institutions of a constitutional regime" (Rawls, 1993: p. xvi). The reasonable pluralism of comprehensive views renders such views unacceptable as bases for the justification of political power. Rawls endorses the 'liberal principle of legitimacy,' that is, that "our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may be reasonably expected to endorse in the light of principles and ideals acceptable to their common human reason" (p. 137). Thus, because there exists a reasonable plurality of comprehensive doctrines, basing the justification of political power on any of them violates the liberal principle of legitimacy. This leads Rawls to seek a political conception that 'all affirm' and that is 'shared by everyone.' Such a conception would be supported by the diverse reasonable comprehensive doctrines that characterize our democratic societies. It would generate public reasons that would fit into our many reasonable, but irreconcilable, comprehensive views. Some version of the principle of political legitimacy motivates the search for public reasons in contemporary liberal political philosophy: if the exercise of coercive political power relies on nonpublic reasoning, it is oppressive and illegitimate in relation to those who do not share those reasons.

Rawls's core argument – and so his entire political liberalism – apparently depends on the contrast between comprehensive views and the shared, political conception. Comprehensive conceptions produce diverse reasons that conflict, whereas the political point of view provides public reasons – reasons that all reasonable comprehensive views can endorse.

Habermas (1995) suggests that Rawls posits an 'a priori' distinction between the political and nonpolitical spheres. At least at times, Rawls seems to suggest that the political point of view can be conceptually distinguished from moral, religious, and philosophical matters; the former identifies a common point of view that can be affirmed by all, whereas when human reason is applied to moral, religious, and philosophical issues, it leads to reasonable disagreement. It is difficult, however, to accept the claim that some single conception of the political is shared by all reasonable persons in modern democratic societies. Political theorists have often pointed to the 'political' as an 'essentially contested' concept – one that is open to divergent, reasonable interpretations (Connolly, 1983: Chap. 1). If citizens of modern democratic societies disagree about the concept of the political, it cannot serve the function of identifying public reasons that overcome reasonable pluralism. Like the moral and philosophical, the nature of the political itself seems a matter of reasonable disagreement.

An alternative interpretation of Rawls' view is to take the political conception as constructed out of that which is shared. On this reading, the nonpolitical is, by definition, those matters on which the free use of reason by citizens in democratic societies leads to different, reasonable conclusions. It is, by its very nature, the realm of reasonable pluralism. In contrast, one might define the political as those matters on which reason converges in democratic societies, and so necessarily generates constitutional principles that satisfy the principle of liberal legitimacy. The political is thus characterized as the overlapping consensus of those diverse comprehensive views characterizing modern societies. The problem for this interpretation is that Rawls himself indicates that the free use of human reason leads us to reasonable disagreement about conceptions of justice and constitutional essentials. The 'political' qua 'shared perspective' is limited to the abstract concept of a 'liberal political order' (Rawls, 1993: p. 241). "By this [Rawls says] I mean three things: first, it specifies certain basic rights, liberties, and opportunities (of the kind familiar from constitutional democratic regimes); second, it assigns a special priority to these rights, liberties and opportunities, especially with respect to claims of the general good and of perfectionist values; and third, it affirms measures assuring all citizens' adequate all-purpose means to make effective use of their basic liberties and opportunities" (p. 223). Thus stated, the content of public reason seems sparse indeed. Unfortunately, if a more detailed articulation of this idea is provided, Rawls indicates that the use of human reason again leads to disagreement. Apparently, modern democratic societies agree on only the most abstract of public – political – reasons.

Convergence Accounts

We have thus far supposed that public reasons are shared or common reasons. Following D'Agostino (1996: pp. 30–31), Vallier (2011) contrasts this 'shared' or 'consensus' view to a 'convergence' conception. According to public reason qua convergence, Alpha and Beta may endorse common conclusions for different reasons; the conclusion is vindicated by public reason because each person has some sufficient reason for it from the perspective of his or her distinctive

commitments. For example, the separation of church and state is supported by some secularists because they do not want the state promoting religion, and also by the religious with doctrines of free faith or who fear the corruption of religion by politics. The key is that the reasoning of different members of the public may converge on some principles, rules, laws, or policies even if they lack consensus regarding why those principles, etc., are appropriate.

To advocates such as Vallier, convergence holds out the possibility of commonly endorsed rules and laws in a social world in which we deeply disagree about the sources of normative justification. For the convergence theorist, our differences are not simply obstacles to, but resources for, achieving a public perspective. However, many reject the idea of convergent public reason.

Jonathan Quong argues that it valorizes a type of insincerity (cf. Larmore, 1996: pp. 134–135; Macedo, 1995; Schwartzman, 2011). Convergence allows Alpha to offer consideration *C* to Beta to justify law *L* because Beta accepts *C*, although Alpha himself thinks *C* quite misguided. But offering as 'reasons,' considerations one rejects may seem like mere rhetoric or manipulation. Quong insists that only shared reasons can ground each seeing the other as having good reason, and thus as truly justified in endorsing *L* (2011: pp. 265ff). Quong's argument here depends upon holding that Alpha can see Beta as justified in holding views that he or she rejects. If, however, Alpha accepts the sources of the diversity of reason and reasoning we canvassed above (see Section The Diversity of Reason), he or she is apt to reject the deep assumption of Quong's critique: that what are good reasons for Alpha must be what are good reasons for Beta (Gaus, 1990: Part 1; Gaus, 1996: Part 1; Gaus, 2010: Sect. 13).

Convergence theorists not only defend convergence reasoning as compatible with respectful engagement, but also seek to turn the tables on their critics, arguing that the demand for consensus is what fails to respect diverse individuals. Jeffrey Stout argues that an alternative to the retreat to a common ground is a Socratic dialogue, in which I show you my reasons and you show me yours, or, more fully, employing the mixed "strategy of expressing one's own (perhaps idiosyncratic) reasons for a political policy while also directing fair-minded, nonmanipulative, sincere immanent criticism against one's opponent's reasons" (Stout, 2004: p. 85). Furthermore, respecting people in light of reasonable pluralism may require taking their diverse comprehensive views seriously instead of prohibiting appeal to them (see Vallier, 2011: pp. 264, 273; Stout, 2004: p. 73).

It must be noted that convergence reasoning dissolves the Enlightenment target of the best principles of justice and public policies. As A. Sen argues, members of the public with diverse perspectives rarely converge upon judgments about the best principles and policies, but will often agree to a broader set of policies worth endorsing (2002: Chap. 4). Consider Alpha who believes there are special obligations to the poor because of his or her religious belief and Beta who believes that preventing the desperation of severe poverty reduces crime. Alpha and Beta likely find many poverty-relief polices worth supporting and better than the injustice of having no such policy, although they disagree about which is best. They each can compromise to accept a policy that promotes, without maximizing, his or her values. Convergence provides a set of eligible policies within which public reason is inconclusive. The legitimacy of a particular policy within this eligible set requires use of a political decision procedure (Gaus, 1996: Part III) or appeal to the existing, partially contingent, status quo (Gaus, 2010: Sect. 21). Convergence reasoning thus significantly reinterprets, or moves beyond, the Enlightenment ideal.

Membership in the Public

Members of the public are those whose perspectives must be accounted for in public reason. Since public reason accounts arise from recognition of reasonable pluralism, membership to the public cannot be restricted to believers in 'the Truth.' If we had an independently defined truth, we may want to investigate what sort of people would understand and believe it, or what reasoning capacities facilitate discovering that truth. Public reason, however, treats the public as an independent variable while searching for the conclusions such a public could endorse.

The members must be picked out by criteria at least some of which are independent of the conclusions they support. Epistemic theories (see Section Epistemic Theories of Public Reason), for instance, provide basic features of rational thought that anyone must follow to count as a member of the public. This requirement will rule out some conclusions, such as contradictions, but does so without narrowing the public to a single perspective. Membership may also require moral capacities, such as Rawls's requirement that members not only be rational, but also have an interest in forming, and ability to comply with, fair terms of cooperation. Moral restrictions may be very basic, such as simply excluding psychopaths and others lacking developed moral emotions (Gaus, 1990: Sect. 17). Some theorists, on the other hand, argue that particular substantive commitments are necessary, excluding perspectives that do not give sufficient priority to democratic equality or other values. Requiring these value commitments may seem necessary to ensure that our political principles are not 'held hostage' by the demand of immoral "or otherwise troublesome people in the world as we know it …" (Quong, 2011: p. 37). The justification of even basic liberal rights could be at stake if the public includes those with highly illiberal values. It is hard to see, however, why the conclusions of public reason with a value-restricted membership should be authoritative over people with dissenting values. From the dissenting perspective, the 'public' conclusions will appear to be sectarian and merely authoritarian impositions. An important alternative to value-based exclusions is to find shared presuppositions, such as the conception of ourselves as agent, from which public reason can craft liberal principles (Gaus, 2010: Chap. VI). People not sharing such a conception are outside the public not for being immoral or troublesome, but instead because they would seem to not be persons at all and their perspective is too distant to form a public with us.

Conclusion: Reasoning in a Plural Society

Puzzles about the possibility of shared, public rationality and reasons derive from a variety of sources: social science, the

philosophy of science, and the moral pluralism of modern societies. Although developments in all these fields serve to cast doubt on the Enlightenment's ideal of a common universal reason, they pose different challenges. The ideal of public reason is as diverse as the sources of our doubts about the possibility of shared ways of reasoning in contemporary societies. In all its manifestations, however, the ideal of public reason seeks to maintain the heritage of the Enlightenment – the free use of human reason will lead us to agreement in our public life.

See also: Contractarianism; Democracy: Normative Theory; Intercultural Legal Discourse: Philosophical Aspects; Pluralism and Tolerance; Relativism: Philosophical Aspects; Value Pluralism.

Bibliography

Benedict, R., 1934. Patterns of Culture. Houghton Mifflin, Boston, MA.
Connolly, W.E., 1983. The Terms of Political Discourse, second ed. Princeton University Press, Princeton, NJ.
D'Agostino, F., 1996. Free Public Reason. Oxford University Press, New York.
D'Agostino, F., 2003. Incommensurability and Commensuration: The Common Denominator. Ashgate Publishing Ltd, Burlington, VT.
D'Agostino, F., Gaus, G.F., 1998. Public Reason. Ashgate, Aldershot, UK.
Davidson, D., 1973. Radical interpretation. Dialictica 27, 314–328.
Gaus, G.F., 1990. Value and Justification: The Foundations of Liberal Theory. Cambridge University Press, Cambridge, UK.
Gaus, G.F., 1996. Justificatory Liberalism: An Essay on Epistemology and Political Theory. Oxford University Press, New York.
Gaus, G.F., 2010. The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World. Cambridge University Press, Cambridge, UK.
Habermas, J., 1990. Moral Consciousness and Communicative Action. MIT Press, Cambridge, MA.
Habermas, J., 1995. Reconciliation through public reason: remarks on John Rawls's political liberalism. Journal of Philosophy 92, 109–131.
Hobbes, T., 1948[1651]. Leviathan. Basil Blackwell, Oxford, UK.
Hollis, M., 1970a. The limits of irrationality. In: Wilson, B.R. (Ed.), Rationality. Blackwell, Oxford, UK, pp. 214–220.
Hollis, M., 1970b. Reason and ritual. In: Wilson, B.R. (Ed.), Rationality. Blackwell, Oxford, UK, pp. 221–239.
Hurley, S., 1989. Natural Reasons: Personality and Polity. Oxford University Press, New York.
Korsgaard, C.M., 1996. Creating the Kingdom of Ends. Cambridge University Press, Cambridge, UK.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago, IL.
Larmore, C., 1996. The Morals of Modernity. Cambridge University Press, Cambridge, UK.
Macedo, S., 1995. Multiculturalism for the religious right? Defending liberal civic education. Journal of Philosophy of Education 29, 223–238.
Mill, J.S., 1974[1843]. A system of logic. In: Robson, J.M. (Ed.), The Collected Works of John Stuart Mill, Vols. 7 and 8. University of Toronto Press, Toronto, ON.
Pettit, P., 1993. The Common Mind: An Essay on Psychology, Society and Politics. Oxford University Press, New York.
Qoung, J., 2011. Liberalism without Perfection. Oxford University Press, New York.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Rorty, R., 1991. Objectivity, Relativism and Truth. Cambridge University Press, Cambridge, UK.
Schwartzman, M., 2011. The sincerity of public reason. Journal of Political Philosophy 19, 375–398.
Sen, A., 2002. Rationality and Freedom. Harvard University Press, Cambridge, MA.
Stout, J., 2004. Democracy and Tradition. Princeton University Press, Princeton, NJ.
Vallier, K., 2011. Convergence and consensus in public reason. Public Affairs Quarterly 25 (4), 261–279.
Wittgenstein, L., 1958. Philosophical Investigations, third ed. Macmillan, London.

Public Sector Organizations

Renate E Meyer and Stephan Leixnering, WU Vienna University of Economics and Business, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Callender, volume 18, pp. 12581–12585, © 2001, Elsevier Ltd.

Abstract

New public management-inspired reforms created numerous autonomous units with many different faces and labels. This variety of organizations and organizational forms precludes a straightforward definition of what constitutes a public sector organization and blurs the boundaries between the public and private sectors as well as the boundaries of single organizations. In addition, the complexity of the interlocking arrangements and relationships in this public organizational landscape has resulted in considerable governance problems with serious implications for coordination and policy coherence.

Modern society is a society of organizations and this is also true for the public sector where organizations have become the central building blocks of states: Public services are delivered by organizations; organizations monitor compliance with legal requirements, regulate liberalized public services, or manage the distribution of public funding. This ubiquity of organizations has been observed by many scholars in the field: Terms like 'para-government organizations' (Hood and Schuppert, 1990), 'agencification' (Pollitt et al., 2001), 'quangocratization' (Hood, 1986), or 'corporatization' (Ducan and Bollard, 1992) emphasize the relevance and heterogeneity of organizations in the public sector. Moreover, observations that citizens are "now separated from their representatives by an asteroid belt of organizations" (Barley, 2007: p. 202) and that the current "arrangements raise questions of considerable constitutional and political significance, including the possibility of state sovereignty being diffused" (Thynne, 2003: p. 318) give evidence of the concerns connected with their proliferation. Brunsson and Sahlin-Anderson (2000) have argued that the public sector reforms of the 1980s and 1990s can indeed be summarized as turning public services into organizations. These far-reaching transformations of the architecture of states seem to have converted public sector reforms into organizational change processes.

Despite their ubiquity, the question of what defines a public sector organization (PSO) is far from easy to answer. What is the distinctive nature of a PSO besides the very general observation that it is an entity located in the diffuse "sphere of delegated governance" (Flinders and Skelcher, 2012: 330) of the state? Is it their ownership structure, the task they perform, their legal form? What is the appropriate level of attention: The individual organizations or the whole network of organizations? Today, not least due to the developments of the past three decades, there is no typical organizational form that encompasses all PSOs and no single classification that captures their diversity. To use Hood's (1986) metaphor: PSOs are not a particular animal or even species, they make up a whole administrative zoo. Apart from this heterogeneity in form, PSOs exist in ambiguous contexts. While, on the one hand, individual PSOs are often very stable, inflexible, and in some cases even immortal (Kaufman, 1976); on the other hand, they exist in a field which is highly dynamic and subject to ongoing change and intense reform activities.

We will, in the following, first outline the developments that have contributed to the increase in numbers, variety, and relevance of PSOs. We then discuss shifts in the various types of activities these organizations perform. In order to give an overview of the heterogeneity of organizational forms in the public sector and the current discussion, we briefly review different criteria used for classification and point to the blurring of sector and organizational boundaries that has been a result of the organizational pluralism. Finally, we address the consequences for governing this heterogeneous organizational landscape and close with a short conclusion.

New Public Management Reforms and Public Sector Organizations

Reform per se is, as Mayntz (1997) remarked, a permanent reflex of administrative praxis. The past three decades have certainly been no exception from this general observation. During this time, the public sectors in most Western democracies have undergone a fundamental change, starting in the mid-1980s in Anglo-Saxon countries (e.g., United Kingdom, New Zealand) and the Scandinavian countries, and have been followed, finally, by the Continental European Countries with a *Rechtsstaat* tradition (e.g., Germany or Austria) about 10 years later. Hereby, new public management (NPM) served as an umbrella term for diverse reform initiatives to such an extent that the term has come to denote the reform paradigm of the whole era.

Being an umbrella concept, NPM was broad enough to be a reference model for reforms from various ideological backgrounds (e.g., 'Thatcherism' and 'New Labor' in the United Kingdom; the Reagan and the Clinton/Gore administration in the United States) and with quite diverse scopes. In the different countries, reforms connected with NPM ranged from modest modernizing of the administration to fundamentally rethinking the state's role and its administrative architecture (Hood, 1991; Pollitt and Bouckaert, 2011). NPM's principles were mainly underpinned by new institutional economics (e.g., minimization of transaction costs by reconsidering agency-dilemmas and the distribution of property rights) and explicitly directed against Weberian-style bureaucracy. Despite this divergence in scope and content of actual reforms, NPM

entailed some core elements, which, independently from the administrative tradition or political party in power, almost all governments referred to. The creation of (semi-)independent, single-purpose PSOs and the reliance on private-sector management instruments are two of them.

Although (semi-)independent public entities have already existed in pre-NPM times, their quantity as well as their quality has been significantly influenced by NPM-inspired reforms. A common denominator was the assumption that organizational disaggregation, together with the coupling of responsibility for resources and outputs within one single organizational entity, would lead to more efficient and effective service delivery. As a consequence, alongside the traditional centralized and decentralized government bodies (ministries, departments, etc.), public tasks and services were increasingly to be carried out by such (semi-)independent entities. The fact that they were predominantly created ad hoc following a case-by-case approach (Van Thiel et al., 2012) is one of the reasons for the generally low level of formal standardization of PSOs even within single countries.

Moreover, within the traditional administrational units (e.g., governmental departments), measures like the introduction of decentralized lump-sum budgets, contract management with goals and targets, performance measurement, and incentivation, gave rise to the formation of quasiorganizations. Inspired by managerialism – "let the managers manage" (Kettl, 1997: p. 447) – the management of public sector units was widely seen as the main field of professionalization. Business companies served as a role model, which also became apparent in the linguistic usage of terms typical for the business world. Hence, a multiplicity of concepts and instruments was transferred from the private into the public sector (for an overview, see Rainey, 2009), such as the product logic and the idea of a production process of public services, or transfer pricing systems between administrative units. From a steering at arm's length perspective (Hood, 1991), the predominant question of how to "make the managers manage" (Kettl, 1997: p. 454) established a highly performance-oriented view. As a consequence, the decentralization of managerial decision making, management by objectives, and performance-related pay were seen as a proper means to encourage management's commitment with service delivery on the one hand and to foster cost awareness among public managers on the other hand. They also increased management's identification with their single units rather than with the whole government body this unit was part of.

The focus on disaggregation and decentralization led not only to a significant increase in the number of individual organizations, but also to the enlargement of types of organizations – organizational forms – to be found in the public sector. This development was reinforced by a discussion of new roles and changing tasks of the state.

New Types of Activities

The variety of organizational forms is matched by the range of different types of activities PSOs have to perform: Organizational forms and types of activities have evolved separately, but definitely not independently from each other.

Public tasks change over time due to historical, demographic, or technological factors: Some services may become obsolete (such as care for war invalids in Western Europe), some increase or alter (e.g., the need for elderly care), and some new services arise (e.g., e-government services such as online tax declarations). Tasks and activities also change as the underlying discourses and narratives of the state and its public administration shift (Fleischer and Jann, 2011). The movement away from the older welfare state toward the notions of an enabling state, guaranteeing state, or activating state has gone hand in hand with an increase in responsibilities such as regulating and monitoring of liberalized domains and decentralized units that had been created to deliver the devolved tasks. Today, PSOs perform functions as different as service delivery (which is the classical function), regulation (e.g., in the field of infrastructure) and supervision (e.g., of the capital market) of policy fields, or the transfer of various types of subsidies. These different types of activities cover the entire spectrum from acting as a public authority to taking the role of a private market actor.

As a result, the relationship between the state and its PSOs is multifaceted, too: The state – the role of the state may be assumed by central (e.g., federal), regional, or local governments or other legal public bodies such as municipalities, provinces, regions, etc. – acts in a variety of roles, such as owner, funder, regulator, or purchaser of services provided by a PSO. It is equally possible that different units of the state embody different roles in relation to the same PSO and therefore have differing or conflicting interests. State-owned enterprises (SOEs) play a central role not only in Continental Europe, but also in the rapidly developing economies in Asia (e.g., China or Indonesia): They have always been a good example for such internal conflicts of interests, as they are often market-based and for-profit and may compete with private companies. The state as an owner may be interested in returns (or, at least, a minimization of losses), while as a customer, it may be interested in reasonably calculated products of high quality (e.g., postal services). Additionally, the public/electorate (and especially the SOEs' employees) may expect the state to promote fair wages, which alludes to its role as a (direct or indirect) employer. State-owned energy providers are a good and complex example, as their purpose is twofold: On the one hand, they have to secure the state's stake in a field of high strategic importance (organization as instrument); on the other hand, they are expected to deliver returns that go into the overall budget (organization as investment). In cases like rail services, liberalization approaches and, thus, the emergence of competing service providers have led to single entities where infrastructure provider and one core competitor are under the same organizational roof. Problems arising from these models have mostly resulted in the further decomposition into smaller single-purpose units. Market share and power, governance issues (e.g., public corporate governance), regulation under conditions where the regulator owns the biggest or only competitor, or accountability concerns in the absence of a democratic electorate are but a few of the questions that arise from the multifaceted trends related to PSOs.

Heterogeneous Organizational Forms and Blurring Boundaries

PSOs comprise not one prototypical organizational form, but many. While some are very similar to private sector companies and some share many characteristics with NPOs, others like courts, federal banks, or military organizations are fundamentally different. There is no easy or clear answer to questions such as: What defines a PSO? What is the common denominator of this diverse collection of organizations? What features does an organization have to exhibit in order to qualify as PSO apart from the fact that it is somehow located in the public sphere? In fact, questions like these challenge and make blurry or even fluid both the boundaries between the public and the private sector and – from an organizational perspective – the boundaries of a single organization.

As a consequence of this heterogeneity, one can find a considerable number of categories and labels for the different decentralized public entities, such as agencies, nondepartmental public bodies, corporatized units, SOEs, quagos, quangos, etc. Their definitions vary and so do the presumed relationships among them. To give just one example: There is no consensus as to whether 'agency' is an encompassing category that does not denote any particular organizational form and thus, also embraces government administration and departmental units (Australian Public Service Act of 1999 cit. in Wettenhall, 2003), or a specific type describing specific PSOs – those that are "considered suitable for activities that can be shifted out from core government departments, but not very far" (Pollitt et al., 2001: p. 276). While in the first understanding, agency is a synonym for PSO, in the latter agencies are a 'marginal case' of PSOs only, even if states suffered from an 'agency fever' (Pollitt et al., 2001). This multiplicity of forms and labels in the Anglo-Saxon administrative tradition is mirrored in other traditions and languages. It is obvious that this is a considerable challenge for comparative descriptions and research (Verhoest et al., 2012).

What Defines a PSO?

In the literature, several attempts to define and map PSOs have been made (for overviews see Roness, 2003; Thynne, 2003; Wettenhall, 2003). The labeling as 'public sector organizations' points to public sector as opposed to private sector as primary delineating factor. Yet, the decentralization and autonomization trend together with the shift in the predominant conception of the state have led to a situation where public sector activities are increasingly performed in an overlapping area between the three sectors (public, private, and not-for-profit). Pollitt et al. (2001: p. 274) note that the public–private borderline "is increasingly a rather large patchwork zone rather than a line." The adequate blend of sectors in the provision of public services – including contracting out, outsourcing, and public–private partnerships (PPPs) – is even one of the objectives of an enabling state. This complex array of hybrid or intermediate types of organizations – Wettenhall (2003: p. 229) ironically calls it the 'go-quago-quango-nongo spectrum' – has, to a certain extent, dissolved, or at least blurred, the sector boundaries and precludes a straightforward delineation along this feature. This difficulty to clearly differentiate between the private and the public sector is, according to Wettenhall (2003), the 'biggest hurdle' in the way of developing a Linnéan classification of PSOs.

Classifications of PSOs are based on a variety of criteria such as task, ownership structure, legal status, degree of autonomy, financing and budget structure, distribution of financial surpluses, or 'publicness.' Building on the tasks devolved (see also the section above), PSOs are generally categorized as to whether they perform classical functions based on government authority, welfare state activities or economic market-based activities (Bouckaert et al., 2000). For organizations with market-based activities, the term SOE is frequently used (OECD, 2005). Based on the primary type of expenditure, Dunleavy (1989), for example, distinguished between delivery, regulatory, transfer, contract, and control agencies.

Ownership as ordering principle of a classification scheme mostly distinguishes between units that are fully owned by one governmental body, PPPs, and public–public partnerships, i.e., multiple governmental bodies (e.g., in federal systems) together own an organization. However, ownership as a criterion may also turn out to be misleading (e.g., with supranational organizations like the UN), rather vague (as, for example, with trusts or funds) or complex (e.g., when multiple other PSOs are the owners). Drawing on an analogy to private sector group consolidation rather than ownership, control has become useful in defining boundaries (see e.g., the definition of in-house procurement by the EU court, or International Public Sector Accounting Standard 6 on Consolidated and Separate Financial Statements, issued by the International Federation of Accountants). However, the question of control equally requires a case-by-case consideration as it may be based on organizational (e.g., ownership of the majority of shares, contractual agreements to have the right to appoint the majority of board directors), financial (e.g., subsidies that account for the majority of the organization's income), or economic terms (e.g., the state being the only purchaser of the organization's products).

Using the legal status of a unit as characteristic feature differentiates, firstly, entities that do not have a separate legal status (e.g., departmental units) from those that do. Secondly, depending on the legal instruments by which entities have been established, the latter category is further split into public law or private law entities. The practical decision between private or public law is often informed by tax, liability, transfer, or public procurement considerations. Moreover, in the public debate, 'privatization' is often also used for the setting up of an entity according to private law, irrespective of the fact that it remains fully government owned. Thus, the choice of legal status may also be driven by the expected degree of approval by the electorate for privatizations rather than by characteristics of the respective unit. Hence, legal status alone is a weak indicator as to whether the unit should be regarded as a public entity or not.

Also with regard to the level of autonomy, PSOs come in a broad variety, ranging from administrative units given slightly more freedom (e.g., by the implementation of lump-sum budgets or service contracts) to entities with considerable degrees of legal, economic, and managerial autonomy. Although one of the core principles of NPM demanded the separation of policy formulation from execution, some PSOs

were given policy autonomy (in addition to their legal and managerial autonomy), which means that they play a key role not only in the execution of political objectives but also in the know-how based formulation of those goals and targets (Van Thiel et al., 2012). A recent suggestion for a Linnéan System of categorization that builds on different levels of autonomy has been developed, for example, by Van Thiel (2012).

A common distinction between private and nonprofit sector hinges on the distribution of financial surpluses among owners. Although public sector activities are prototypically seen as nonprofit, there are many PSOs that gain profits and are even designed to do so (see especially market-based SOEs). Thus, this distinction is useful for PSOs only to a very limited extent. In some cases, for-profit and nonprofit orientation (and often also public and private law) is mixed, when, for example, PSOs, such as universities, that are generically designed as nonprofit, create for-profit spin-offs (e.g., executive education centers, university-owned consulting companies, university as incubator for innovations, etc.).

Some scholars recommend giving up the categorical distinction in favor of a degree of publicness. Antonsen and Jorgensen (1997: p. 340), for instance, studied publicness as the "degree to which public sector organizations adhere to public sector values" across different functional categories and policy sectors. Another classification of PSOs that tries to account for several of the categories discussed above has been suggested by Pollitt et al. (2001). In their model, different types of PSOs are clustered along two dimensions, ownership (public institutions, public bodies, PPPs, and civil society) and the extent to which the PSO fulfills commercial purposes.

Redefining Organizational Boundaries

Apart from the definition of 'public' in PSOs and the related problem of blurring sector boundaries, the questions of what constitutes an organization and what is the appropriate unit of attention (Roness, 2003) are equally relevant as difficult to answer.

NPM's core idea was to focus on smaller, decentralized single-purpose units; the role model for creating these units was, in many respects, the basic organizational form of the private sector: The corporation. While for management and incentive purposes, this focus on the individual unit may be appropriate, from the perspective of the state and its public sector or even from the perspective of a single policy field, it falls short in many ways. In fact, this imagery of an isolated organization engaging in transactions is no longer regarded as adequate even in the private sector. In the public sector, it is fundamentally at odds with the complexity of the interlocking arrangements and relationships between the state and its various PSOs. Including into the picture arrangements such as outsourcing and contracting out makes the question of organizational boundaries even more challenging with the obvious danger being to miss the forest for the trees.

In the private sector that originally served as the model for NPM-inspired reforms, similar concerns with the focus on single organizational units have been voiced and much effort has recently been devoted to the study of embedded groups of organizations and the network relations between them. In the public sector, an analogous perspective encourages making the whole formation of PSOs and its governance the center of attention. (The analogy of the public organizational landscape and business groups is especially popular among German scholars: see e.g., Reichard (2006) on 'municipal holdings.') The heterogeneity of organizations within the network and the diverse and intertwined power relationships between them certainly provide enough material for in-depth analyses that can generate interesting insights into the working and dynamics of the post-NPM PSO landscapes. Being complex systems, these network formations have characteristic features (Goergen et al., 2010) such as the emergence of typical patterns of behavior, shared norms and values, and a distinct identity. Another common feature of complex systems is entropy, which means that within such systems, strong centrifugal forces are at work. Moreover, their specific formation and the practices, norms, values, and identities they develop are shaped by the cultural, historical, and political context in which they are formed. Thus, although the decisions to construct the individual organizations in whatever form may be ad hoc and nonstandardized, the overall formation is far from being random, but reflects its contextual embeddedness. Consequently, measures to strengthen PSOs' structural and cultural ties within the public organizational landscape are currently at the center of many reform initiatives under the banner of a (new) public governance.

Consequences for Governing

More than two decades of reforms inspired by the principles of NPM have had considerable consequences for the public organizational landscape. Van Thiel et al. (2012: p. 419) mirror very well the disillusionment with these reforms when they state that autonomy alone, even if understood as multidimensional quality combining legal, managerial, financial, and policy aspects, "has proven to be an insufficient condition to increase agency management and performance." In addition to autonomy, means of trust and control, the establishment of a stimulating organizational culture and appropriate governance structures are required. Thus, reforming PSOs in the post-NPM era frequently aims at establishing a shared culture and a general, widely understood, and encompassing governance framework for all organizations in the landscape.

Reform initiatives attend to a range of concerns and introduce related measures that are frequently subsumed under the label of (new) public governance (Osborne, 2010; Pollitt and Bouckaert, 2011) – arguably the new kid on the block of umbrella terms and reform paradigms. A first bundle of concerns addresses the management of networks and multi-actor systems (Rhodes, 1997) as well as legitimacy issues ('good governance': European Commission, 2001). On the one hand, the enabling state and the integration of multiple external actors (such as NPOs and private sector companies) into the processes of public service provision and the creation of hybrid organizational forms demand management systems and steering mechanisms outside the classical hierarchical authority relations. On the other hand, transparency, accountability, or participation are equally demanded from

units operating at arm's length. However, it is much more difficult to secure compliance in fragmented systems.

It has been argued that the main motives for delegated governance and the creation of autonomous organizations lie beyond the increase of effectiveness and efficiency: Instead, depoliticization and impartial execution of public tasks (Van Thiel et al., 2012) was the promise. However, as the authors argue, from a quality of democracy perspective, the increased distance between PSOs and the parliaments and the resulting weakening of democratically legitimized institutions are highly problematic. Concerns not only with increasing expenditure, but with a dilution of political responsibility and accountability, while at the same time political patronage is rising, were based on the insight that PSOs "occupy a niche whereby ministers can remove themselves from direct responsibility for policy implementation" (Flinders and Skelcher, 2012: p. 327). PSOs, initially designed to withstand political interference, show to be likely of falling victim to politicization no less than centralized governmental units, only more hidden.

According to empirical evidence from different countries (Verhoest et al., 2012), the administrations' ability to stay in command of the organizational landscape seems to be rather low, irrespective of different administrative traditions: What is more, forms and instruments of steering, controlling, and directing PSOs turn out to be insufficient to a great extent. As a consequence of the disaggregation, formerly closely tied policy fields seem only loosely coupled and fragmented today. Apart from coordination and coherence difficulties, this leads to a considerable lack of transparency not only for the external public, but also for within the administration itself. Many administrations no longer have a comprehensive overview of their independent bodies. Flinders and Skelcher (2012: p. 329), for example, report that in the United Kingdom, following the new government's attempt to reduce the number of quangos, first attempts to quote correct figures and map the organizational landscape showed that "*nobody* actually possesses a blueprint of the structure of British governance or understands how the various tentacles of the bureaucracy are expected to work together." Undoubtedly, this is no exception.

Although it is certainly true that coordination is "an old administrative doctrine" (Hood, 2005: p. 19), its lack is at the center of reform initiatives attending to the fragmentation and control deficits that have resulted from the unbundling and splitting up of the state into 'single-purpose organizations,' sometimes effected by means of reintegration, in-sourcing and deagencification. Counteracting the focus on measurement and indicators, post-NPM approaches rely on shared norms and values, ethics and the fostering of a distinct public service motivation. The aim is to regain cohesion and coordination ('whole-of-government' or 'joined-up-government' initiatives: among many others see Christensen and Laegreid, 2007; Hood, 2005) across and between levels of government and these highly fragmented and complex bundles of organizations by increasing structural and cultural embeddedness without reestablishing the status quo ante by merely reaggregating them. The challenge is once again to keep a balance of embedded autonomy: This time between the advantages of PSOs' autonomy and the reestablishment of their ties to the public sector and the governmental body they are part of.

Conclusion

NPM-inspired reforms created numerous autonomous units with many different faces and labels. This variety of organizations and organizational forms precludes a straightforward definition of what constitutes a PSO and blurs the boundaries between the public and private sector as well as the boundaries of single organizations. In addition, the complexity of the interlocking arrangements and relationships in this public organizational landscape has resulted in considerable governance problems with serious implications for coordination and policy coherence.

See also: Boundaries and New Organization Forms; Bureaucracy and Bureaucratization; Corporate Governance; Development: Organizational; Governance; Old and New Institutionalism in Economics; Public Administration: Organizational Aspects; Quality of Government; Regulatory Agencies.

Bibliography

Antonsen, M., Jorgensen, R.B., 1997. The 'Publicness' of public organizations. Public Administration 75 (2), 337–357.

Barley, S.R., 2007. Corporations, democracy, and the public good. Journal of Management Inquiry 16 (3), 201–215.

Bouckaert, G., Ormond, D., Peters, B.G., 2000. Restructuring Central Government. Stage I. The International Dimension, Helsinki.

Brunsson, N., Sahlin-Anderson, K., 2000. Constructing organizations: the example of public sector reform. Organization Studies 21 (4), 721–746.

Christensen, T., Laegreid, P., 2007. The whole-of-government approach to public sector reform. Public Administration Review 67 (6), 1059–1066.

Ducan, I., Bollard, A., 1992. Corporatization and Privatization: Lessons from New Zealand. Oxford University Press, Auckland/New York.

Dunleavy, P., 1989. The architecture of the British central state, Part II: empirical findings. Public Administration 69 (4), 391–417.

European Commission, 2001. European Governance - A White Paper.

Fleischer, J., Jann, W., 2011. Shifting discourses, steady learning and sedimentation: the German reform trajectory in the long run. In: Eymeri-Douzans, J.-M., Pierre, Jon (Eds.), Administrative Reforms and Democratic Governance. Routledge Chapman & Hall, New York, pp. 68–79.

Flinders, M., Skelcher, C., 2012. Shrinking the quango state: five challenges in reforming quangos. Public Money & Management 32 (5), 327–334.

Goergen, M., Mallin, C., Mitleton-Kelly, E., Al-Hawamdeh, A., Hse-Yu Chiu, I., 2010. Corporate Governance and Complexity Theory. Edward Elgar Publishing Limited, Cheltenham/Northampton.

Hood, C., 1986. The hidden public sector: the 'Quangocratization' of the world. In: Kaufman, F.-X., Majone, G., Ostrom, V. (Eds.), Guidance, Control and Evaluation in the Public Sector. de Gruyter, New York, pp. 183–207.

Hood, C., 1991. A public management for all Seasons? Public Administration 69 (1), 3–19.

Hood, C., 2005. The idea of joined-up government: a historical perspective. In: Bogdanor, V. (Ed.), Joined-Up Government. Oxford University Press, Oxford/New York, pp. 19–42.

Hood, C., Schuppert, G.F., 1990. Para-government organization in the provision of public services: three explanations. In: Anheier, H.K., Seibel, W. (Eds.), The Third Sector: Comparative Studies of Nonprofit Organizations. de Gruyter, Berlin/New York, pp. 93–106.

Kaufman, H., 1976. Are Government Organizations Immortal? Brookings Institution, Washington, DC.

Kettl, D.F., 1997. The global revolution in public management: driving themes, missing links. Journal of Policy Analysis and Management 16 (3), 446–462.
Mayntz, R., 1997. Verwaltungsreform und gesellschaftlicher Wandel. In: Grande, E., Prätorius, R. (Eds.), Modernisierung des Staates? Nomos, Baden-Baden, pp. 65–74.
OECD, 2005. Guidelines on Corporate Governance of State-Owned Enterprises, Paris.
Osborne, S. (Ed.), 2010. The New Public Governance. Emerging Perspectives on the Theory and Practice of Public Governance. Routledge, London/New York.
Pollitt, C., Bathgate, K., Caulfield, J., Smullen, A., Talbot, C., 2001. Agency fever? Analysis of an international policy fashion. Journal of Comparative Policy Analysis 3 (3), 271–290.
Pollitt, C., Bouckaert, G., 2011. Public Management Reform: A Comparative Analysis. NPM, New Public Governance and the Neo-Weberian State, third ed. Oxford University Press, Oxford, UK.
Rainey, H.G., 2009. Understanding and Managing Public Organizations, fourth ed. Jossey-Bass, San Francisco, CA.
Reichard, C., 2006. Strengthening competitiveness of local public service providers in Germany. International Review of Administrative Sciences 72 (4), 473–492.
Rhodes, R.A.W., 1997. Understanding Governance: Policy Networks, Governance, Reflexivity and Accountability. Open University Press, Maidenhead/Philadelphia, PA.
Roness, P.G., 2003. Categorizing State Organizations in Theory and Practice, in General and in Norway. Seminar on Organizational Forms, Autonomy and Control in the Public Sector, Bergen.
Thynne, I., 2003. Making sense of organizations in public management: a back-to-basics approach. Public Organization Review 3 (3), 317–332.
Van Thiel, S., 2012. Comparing agencies across countries. In: Verhoest, K., Van Thiel, S., Bouckaert, G., Laegreid, P. (Eds.), Government Agencies – Practices and Lessons Learned from 30 Countries. Palgrave Macmillan, Hampshire/New York, pp. 18–26.
Van Thiel, S., Verhoest, K., Bouckaert, G., Laegreid, P., 2012. Lessons and recommendations for the practice of agencification. In: Verhoest, K., Van Thiel, S., Bouckaert, G., Laegreid, P. (Eds.), Government Agencies – Practices and Lessons Learned from 30 Countries. Palgrave Macmillan, Hampshire/New York, pp. 413–439.
Verhoest, K., Van Thiel, S., Bouckaert, G., Laegreid, P. (Eds.), 2012. Government Agencies – Practices and Lessons Learned from 30 Countries. Palgrave Macmillan, Hampshire/New York.
Wettenhall, R., 2003. Exploring types of public sector organizations: past exercises and current issues. Public Organization Review 7 (3), 219–245.

Relevant Website

www.ifac.org – International Federation of Accountants.

Public Sphere and the Media

Slavko Splichal, University of Ljubljana, Ljubljana, Slovenia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Garnham, volume 18, pp. 12585–12590, © 2001, Elsevier Ltd.

Abstract

Public sphere as a specific sphere, domain, an imagined space of communicatively mediated social life, or a discursively constructed social space between the state and civil (initially bourgeois) society represents the infrastructure for social integration through public discourse. It is taken for granted that no specific social norms (e.g., privacy, ownership) and regulatory rules (e.g., formal membership) limit citizens' access to the public sphere, and no specific knowledge and competences are needed to make them able to participate in it. A public sphere could either be (1) the sphere of publicness, i.e., communication spaces created and maintained by the media, which involve relations of power and dominance ('public space') or (2) the sphere of the public/s consisting of free and equal citizens participating in public reasoning. In either case, the principle of publicness represents the very heart of the concept of the public sphere. In empirical terms, the public sphere is often conceived of as a kind of 'compromise' of these two 'ideal types.' The concept of the public sphere largely refers to the role of the media and public opinion in modern representative democracies; the normative conceptualization of the media comes to the forefront in the idea of public service media.

To analyze the media from the perspective of the public sphere is to focus upon the media's relationship with democracy. The concept of the public sphere has been used to describe the historical development of modern, post-Enlightenment forms of representative democracy and to conceptualize the role of the media and public opinion within them, and has also served as a normative concept with which to pass judgment on the current performance of the media from the perspective of the creation and maintenance of democracy.

Principle of Publicness and the Public Sphere

Central to the concept of the public sphere is the normative principle of publicness. Modern democracy is usually thought of as a product of the enlightenment, when the idea of publicness was raised to a fundamental moral principle; I. Kant (1784) outlined the 'public use of reason' as the fundamental and universal principle of human public agency, and J. Bentham (1791) defined 'publicity' as the foundation of public opinion and people's sovereignty – a safeguard for public confidence in the rulers and an assurance that they would perform their duties. Since the principle of publicness only makes sense in reference to forms of communication beyond transitory interpersonal interaction, the nature of 'being public' – the publicness – in the strict sense could not have existed before the ample expansion of the temporal and spatial horizons of communication brought about by the development of the press. In contemporary (Western) democracies, the idea of publicness primarily refers to the public sphere, where the public use of reason or public discussion of free and equal citizens should result in the formation and expression of public opinion.

In recent years, the public sphere perspective has become dominant in thinking about the media and their political role, by providing, in a return to the central questions of enlightenment political thought, a bridge between Marxist theories of ideology and the liberal 'free press' tradition. It provides a critique of Marxist theories of ideology by revalidating the specificity of the political, by giving due weight to the emancipators potential of liberal bourgeois concepts of free assembly and debate, and by shifting attention from worker to citizen. On the other hand, it provides a critique of an abstract, property-based 'free press' model by stressing both the systemic social barriers to entry and the manipulative distortions to which its discourse is subject. Thus the idea of the public sphere follows Marx's critique of the bourgeois idea that the press might attain its freedom more easily and fully by adopting the laws of free economic operations that rested on the right to private property, which paved the way to the subordination of the press to freedom of ownership as a peril no less dreadful to a genuine freedom of the press than ideological censorship.

The concept of the public sphere was introduced into media analysis with the English translation of J. Habermas's book *Strukturwandel der Öffentlichkeit* (1962, Trans. 1989). In Habermas's original formulation, the term was designed to describe a historical process whereby in the shift from premodern feudal, absolutist rule to forms of modern representative democracy public opinion based upon publicly available information and debate, developed as both a check on, and as a source for the legitimacy of, government. The words 'public sphere' originated in Habermas's *The Structural Transformation of the Public Sphere* 1989, but the idea dates back to Marx's (1843a) early conceptualization of the press as 'the third element' mediating between decision-making political authorities (the 'first element') and civil society (the 'second element'), in which the rulers and the ruled alike can criticize their principles and demands on equal terms, emancipated from their authoritative officiality and private interests as citizens of the state; no longer as individuals, but as intellectual forces, as exponents of reason. Ever since Bentham, the press was considered constitutive of the public and public opinion, but K. Marx was the first to realize that the press is more than just an 'organ of public opinion.' The distinctive feature of the press is, according to Marx (1843b: p. 155), that it can transform a particular interest to a common interest by confronting

the rulers and the ruled; it 'harmoniously combines all the true elements of the popular spirit.' It took more than a hundred years that Marx's idea of 'the third element' has been rediscovered as 'the public sphere' by Habermas.

This return has focused on concepts of the public sphere, civil society, citizenship, and identity. It has in particular pitted those who see the emancipatory project of the enlightenment as unfinished business and wish, in a movement of immanent critique, to hold liberalism to its emancipatory ideals, against a range of communitarians and postmoderns. The latter, in the name of various versions of identity politics, see liberal, rights-based political theories as inherently repressive and their emancipatory and universalizing tendencies as bogus (Gray, 1995).

The public sphere approach has provided, against the background of the turn away from Marxism associated with the collapse of 'actual existing socialism,' an alternative to theories of dominant ideology or hegemony as an explanation for the coincidence, within what used to be called 'bourgeois democracies,' of growing social inequality and political apathy on the one hand with the relative stability and increasingly consensual, nonideological nature of representative party politics on the other. Its emphasis on discursive practices and on communicative action as central to democratic practice and legitimization fitted well with the wider 'linguistic turn' in the human and social sciences. At the same time, its use of the spatial metaphor of sphere, and its stress on the necessary institutional foundations for the realization of those citizen rights to free expression and debate central to democratic theory, addressed the questions arising from both the perceived empirical reality in the mature democracies of an increasingly media-saturated, image-driven, public relations–oriented politics and the task of constructing democracy in both state and civil society in newly democratized countries with little or no historical traditions of democracy to call upon.

It was Marx who first conceptualized that the press not as a simple instrument to create and reflect public opinion but as *a mediator* between the state and civil society. He saw the press as *the site* in which the political state and civil society could meet on equal terms, where *citizens of the state* – "no longer as *individuals*, but as *intellectual forces*, as exponents of reason" – are emancipated from their authoritative officiality and private interests. The free press, Marx argues, can mediate between civil society and the state because it is "*political* without being official, hence not based on bureaucratic premises," and it is "of a *civil* nature without being bound up with private interests and their pressing need" (Marx, 1843a: pp. 189–190). It is based both on reason and emotion; its discourse, both rational and passionate, differs from that of official reports. The distinctive feature of the free press is that it can transform a particular interest to a common interest by confronting the rulers and the ruled, provided that a genuinely public sphere is established in which confrontation between opposing viewpoints is possible, because the difference between rulers and ruled vanishes. It is not only an organ through which the public can express its 'public opinion'; it is also the condition for public opinion and its 'producer' – the public – to come into existence and stay alive.

During the last century, Marx's 'third element' mediating between the state and civil society has developed into a complex 'public sphere' – a kind of communication network and infrastructure enabling the formation of public opinion. The public sphere is normatively understood as a specific sphere, domain, or imagined space of social life existing between, and constituted by, the power structure of the state and civil society. It is the context of public opinion rather than its generator, which always is a public. There is no public without a public sphere, and no public sphere without a public. The existence of the public sphere is vital for a public to become visible through public opinion, and an acting public is a necessary condition for a public sphere to really exist.

The public sphere represents the locus where private citizens engage in the 'public use of reason' (Kant); an infrastructure of social integration through (mass mediated) public discourse, in contrast to political power in the sphere of politics and market mechanisms in the sphere of economy, both of which represent specific forms of social integration through competition. It is activated in the communicative interdependence and rational–critical discourse among citizens potentially affected by transactions in which they do (or did) not participate, in which the legislated laws of the state and the market laws of the economy are (ought to be) suspended.

Only as long as economic and social conditions giving everyone an equal chance to meet the criteria for admission to the public sphere are effectively met, the existence of the public sphere is protected. A clear interest in the (re)production of a genuine public sphere derives from less numerous and less powerful actors, which may be classified into three main categories: (1) citizen demand for media uses and interest in receiving information and opinions; (2) various civil society groups' interest in having access to the media to publish their opinions; (3) a general (ethical) interest of citizens in maintaining the citizen rights and in the media performing their public service functions. However, in practical terms, citizens *qua* citizens do not seem to be among key actors in the public sphere anymore but rather, as in the old Lippmann's theorization, 'deaf spectators in the back row' observing public actors (Lippmann, 1925: p. 13).

The public sphere and the public are inseparably connected to each other – like a system and its component, which must produce results congruent with the defined goal – but the difference between the two is essential in their relation to the organized power of the state. The public sphere is *inseparably connected* with the state, which is one of the actors in the public sphere and, at the same time, its legal guarantor, whereas a public is formed more or less *in opposition* to the organized power of the state.

We may think of the public sphere as a kind of communication framework of the 'body public,' a sphere of channels of opinion circulation binding and protecting its constituent publics. The public sphere is the arena in which civil society informs itself and exchanges ideas and opinions with other social actors representing the two remaining realms, those of the state and the economy. It would be delusive, however, to equate 'the public sphere' with 'the sphere of publics.' Beside public/s as a communicatively acting social category consisting of formally equal citizens appearing and acting in relation to some important and controversial social–political issues, whose participation is constitutive of the public sphere, two additional major classes of actors appear in the public sphere *in front of*

publics: (1) state authorities, political parties, interest groups, politicians, media owners, commercial corporations, business elite, and similar actors as coauthors and addressees of public opinions who use the already constituted public sphere, and (2) media gatekeepers – journalists, editors, and publicists – who are the agenda-setting gatekeepers of public discourse. The former potentially imperil publicness of the public sphere in the measure that they bring manipulative publicity in the public sphere by subsidizing information to frame and set the agenda of public discourse. This 'new publicity' represents a constant danger of confounding public actions of elites with a public sphere.

Mass Media, the Public, and the Public Sphere

The public sphere as the infrastructure for social integration through public discourse is largely based on media operation. Mass media are not only constitutive of the public as an 'imagined communion,' which can only come into existence through the media, but also vitally important for the creation of an institutionalized (infra)structure of the public sphere. Mass media are, as Tönnies (1887: p. 221) suggested for the press, "comparable and, in some respects, superior to the material power which the states possess through their armies, their treasuries, and their bureaucratic civil service." Ideally, they channel the flows of discourses from opinion formation in the networks of the public sphere to the political will-formation in the political system and vice versa.

It is not just any media on which the public sphere can be build up. As 'organs' of (opinion of) the public, the media have to perform specific normative functions along the principles of publicness set forth already by Kant and Bentham. Personal freedom of public use of reason and the right of the public to control political authorities are hardly compatible with the media owned by private capital or the state, which adopt media-centered strategies to maximize commercial profit or political power. The initially beneficial consequences of incorporated media as private businesses faded away with the industrialization, commercialization, and monopolization of the press and other media. Marx (1842: p. 68) admitted that "it is no transgression, when a German perceives the unknown goddess of freedom of the press as one of the goddesses he already knows, [as] freedom of entrepreneurship or freedom of ownership." Even if 'enacted' as freedom of property rather than personal freedom to communicate, freedom of the press had temporarily beneficial consequences for the autonomy of civil society and the public sphere in the early phase of liberal capitalism. In the long run, however, the subsumption of freedom of the press in the 'general freedom' of entrepreneurship as the fundamental principle is a fatal degradation of the press.

The subordination of newspapers to powerful commercial and party-political interests in the late nineteenth century brought about the idea of public service newspapers even before the emergence of radio. Many press reformers pleaded for 'security valves in public life' based on charitable funds, which ought to limit the dependence of newspapers on owners, political parties, and advertisers, to prevent sensationalism, and empower people's voices to find their expression in the newspaper to serve the common good together with journalists and editors (Tönnies, 1922). Similarly, C. Bücher (1926: p. 53) saw the main task of the press in enabling the circulation of results of intellectual work in politics, science, arts, and technics, which is why it should be made available to citizens for free. The idea of public service clearly failed in the field of the press even though newspapers were constitutive of the bourgeois public sphere emerging in the seventeenth century. The reason seems to be rather simple. The idea of the 'public service press' emerged rather late; it was actually behind time since newspapers already became powerful economic players, and no social force or political actor existed influential enough to limit their power at that time. Newspapers' owners would ideologically legitimize the rejection of such ideas by denoting them as attacks on 'freedom of the press.'

Essentially, we can find those ideas condensed in subsequent definitions of *public service (radio) broadcasting*. This model prevailed in Britain until the 1960s and influenced the organization of broadcasting in many other countries, particularly (but not exclusively) in Europe. Later on, it was complemented by private commercial broadcasters in the so-called 'dual broadcasting systems.' A rather clear echo of early public service ideas one can come across in a modernized version of the definition of public service broadcasting adopted by UNESCO, although it fails to recognize the importance of the 'direct expression of people's voices' in the media: "Public Service Broadcasting (PSB) is broadcasting made for the public and financed and controlled by the public. It is neither commercial nor state-owned. It is free from political interference and pressure from commercial forces. Through PSB, citizens are informed, educated, and also entertained. When guaranteed with pluralism, programming diversity, editorial independence, appropriate funding, accountability, and transparency, public service broadcasting can serve as a cornerstone of democracy."

In an economic perspective, the idea of public service is often associated primarily with the natural scarcity of goods (e.g., the scarcity of radio frequencies in the predigital era) or complexity of services exceeding individual investment capacities (e.g., military defense), which would justify a strict state control or even monopoly. Since there is no such scarcity in the digital era any more, it should follow – according to this view – that there is also no need for public service broadcasting any more; thus consumer preferences and demands could 'freely' determine the program supply on the media market.

The scarcity argument against media regulation is questioned for three main reasons. First, a true competition of the mass media is not likely to materialize since the entry costs to the media market are prohibitively high. Media concentration in the hands of a small number of corporations taking place both nationally and globally may severely restrict the possibility of nonmainstream 'consumers' to influence the program supply.

Second, legal regulation has been introduced by modern states in virtually all areas of production of goods and services – from food and pharmaceutical industry to aircraft industry and transportation – to protect consumers and society at large.

Third, the mere plurality of even perfectly competitive content providers would not guarantee diversity of the supply. Extremely competitive media markets may tend toward

homogeneity even more likely than monopolistic or oligopolistic media systems because media markets tend to reflect only mainstream preferences and demands of audiences.

In a broader perspective, the concept of public service is rooted in historical processes of human emancipation and thus closely linked with the virtues of human freedom and equality. It implies a broad societal consensus that certain services should be made available to all members of society on a non-profitable basis, regardless of the differences between them in terms of income, social status, cultural habitus, or ethnic affiliation. Such a consensus is based on the recognition that those services are fundamental to human life and thus have to be universally guaranteed and provided. The universal provision of such services associated with fundamental human rights and liberties is an ethical imperative. Even if some public services are not publicly financed (e.g., education provided by private schools), they are subject to a more strict legal regulation and public control than pure commercial sectors.

Mass media can be considered a case in point; since any kind of publication through the media has a public nature and public significance, it calls for specific legal regulation to implement the principle of publicness and privacy protection, which are both related to the personal right, and human generic ability and need, to communicate. It is often taken for granted that freedom of thought and opinion does not legitimately invite public regulation because it is individual rather than intrinsically collective exercise of freedom. Such a belief is essentially delusive as the personal right to use one's reason *in public* can only be realized in community with others.

The normative view that mass media have specific responsibility to serve (civil) society in which they operate is for some an axiom that hardly needs explanation since the media are an integral part of modern social system with significant political and cultural functions for individuals and society as a whole. The opposite view holds that a true freedom of expression implies not only an absence of any outside interference (market forces are considered 'insiders' in the process), but also a lack of obligation to meet any outside claims, especially those backed by force or authority.

Early radical critiques of the commercialization and politicization of the press in the late nineteenth century considered socialization of the press a fundamental condition of its liberation from the subordination to the state, political parties and, particularly, commodity production. They related the ideas of press reforms that would establish the press as public service to the 'extension' of personal freedom of expression and democratization of society, which would make the press a matter of genuine common concern and decision making. But the early critics also realized that reforms could not be limited to the questions of political and economic (in)dependence of the press because the genuine socialization of communication activity stands and falls together with the general democratization of economic and political relations, which substantially affect all forms of human communication.

The general idea of public service points out the principle of equal access to consumption of services or goods in question without any reference to the principle of *publicness*, whereas the public service model of the media is not defensible as just *any* kind of service to the public. This is where the specific idea of public service in the sphere of (mass) communication substantially diverges from the idea of public service in general. Since the media may be considered the guardians of the principle of publicness in modern societies, this principle is central to the concept of public service media and its unique feature.

The specificity of public service media can be defined in reference to two related issues: (1) the nature of contents produced and published (broadcast) by the media and (2) the constitution of the public, which ought to use public service media as its 'organs.' In normative terms, public service media as organs of the public (or publics) provide a service *of* the public, *by* the public, and *for* the public. It is a service *of* the public because (if) it is financed by the public and controlled by it. A service *by* the public – not only financed and controlled, but also produced by it. A service intended *for* the public – but also for the government and other powers acting in the public sphere.

Table 1 presents schematically the main differences between paternal, commercial, and public service media systems in terms of their obligations to those with legitimate power to influence and/or control the media. The media owe responsibility to seven major groups of legitimate actors: (1) their owners, private or state/governmental, (2) clients from the world of the corporate capital, such as advertisers, (3) politicians, (4) interest

Table 1 Modes of media responsibility in three media systems

			Normative power of actors to influence the media		
Media responsibility to	*Source of legitimacy of responsibility claims*	*Prevailing type of regulation of obligations*	*State/Paternal*	*Private/Commercial*	*Public service*
Private media owners	Property right	Legal	None	High	Moderate
State media owners	Property right	Legal/political	High	None	Moderate
Corporate economic clients	Profit maximization	Contractual (market)	Low	High	Low
State – policy makers and regulatory bodies	Political	Legal/political	High	Low	Moderate
Interest groups	Political	Political	Moderate	High	Low
Content creators/Journalists	Professional norms	Ethical	Low	Moderate	High
Audiences	Consumers' rights (e.g., right to choose)	Cost-effective (market)	Low	High	Moderate
Citizens and civil society	Communication rights	Ethical/political/legal	None	None	High

Based on Splichal, S., 2011. Transnationalization of the Public Sphere and the Fate of the Public, Hampton, New York.

groups, (5) professional organizations of occupational groups employed in the media, e.g., associations of journalists, (6) media audiences, and (7) civil society. The relationships between the media and seven groups of actors differ in terms of the source of legitimacy of responsibility claims by these actors and the ways the claims are regulated but also depend on the amount of power specific actors have. The dominant actors influence the media not only to maximize profits or power in the political sphere but also to maintain influence over and against other actors in the public sphere.

For example, privately owned commercial media are more likely to respect responsibility claims by their owners, major commercial partners, and audiences. If the media are kept under state (governmental) surveillance, they are not able to expose political elites to public surveillance. Inasmuch as political and corporate economic elites overlap, corporate control of the media may also disable them from performing their watchdog function.

In contrast, public service media are expected to have stronger professional (ethical) commitments and, in particular, obligations to civil society. As a result, public service media may enjoy greater autonomy from the political and economic power actors due to more dispersed responsibilities whereas in commercial media, profit-seeking interests may limit substantially authentic media autonomy. Moreover, participation of civil society associations in operation of (public) media not only enables civil society to act more effectively, but also contributes to democratization of society by and large.

Without an effective power of civil society to influence the media in terms of its values, interests, and preferences, the idea of public service media would be deprived of its constitutive principles of publicness and personal communication freedom. The power of other actors influencing public service media would excessively increase, and the difference between private and public service media would eventually disappear – to the detriment of civil society and democratic social order.

Public service media may be seen as representing a shrinking minority voice in today's global media landscape. Nevertheless, in the period of newspaper and journalism crisis, the idea of public service becomes redemptive not only in terms of (normative) theory, but also in terms of practical solutions. The recent global newspaper crisis indicates that the production and consumption of media contents could not be left to the market without serious 'collateral' damages, as Bücher and Tönnies have argued long ago. Following the model of public service broadcasting or public university by providing economic stability to the press and promoting journalistic autonomy necessary to serve the public good is believed to be an important step to protect the opinion-forming press, which is fundamental to the existence of the public sphere. It is primarily the responsibility of public service media to be instrumental in the formation of an inclusive public sphere because they are best suited – as indicated in Table 1 – to make it accessible to all citizens regardless of their social status.

Critical Issues

Research and debate around the public sphere and the media now focus on five issues. The first is the question of democratic achievements and failures of the Internet in democratizing public communication. The popularity of the public sphere concept was given a strong impetus by digital(ization of) media in the 1990s. The ubiquity of the Internet in social relations also brought about new puzzles about the public sphere concept. The Internet is much more than just new information and communication technology; its development embodies, reproduces, and sustains forms of social organization spread through the entire social structure. Although the public sphere is not formed around a single medium but is built up as an infrastructure with a variety of communication media, institutional and cultural frameworks as its pillars for organizing public deliberation, the rise of the Internet enabled several civil society actors, particularly in the realm of 'alternative politics' (e.g., women, religious actors, environmentalists, labor, peace activists, gay and lesbian activists, humanitarians) to act on a larger scale and to become more visible both nationally and transnationally. How effective their actions are because of that in terms of impact on public opinion and decision makers is another question. In some cases, the Internet may have contributed to the transgression of fragmented cultural and political interests, but it also may have deepened the fragmentation. It is less efficacious in generating public opinion than the 'old' media. The question is whether the online media and networks across the world can lead toward a 'transnational public sphere' (and transnational publics) or rather toward the fragmentation of mass audiences – who are (were) large, dispersed, and predominantly national, but nevertheless focused on political issues defined by the media – into many dissociated 'issue publics.'

The second question is the institutional question of the relationship between the political public sphere and the nation-state. Both the actual institutional development of the modern mass media, the public sphere, and representative democracy, and thinking about them, have taken place within a nation-state framework. It is argued that current globalizing trends, both within society at large and within the media, are undermining this nation-state foundation of both democratic politics and its associated political public sphere. The question is then whether this is a process of cultural and discursive emancipation and the beginnings of the construction of transnational public spheres or, on the contrary, is a further tightening of the screw of refeudalization, which places the key centers of power outside the structures of democratic accountability and rational critical debate.

On the one hand, the trends toward globalization within the media can be exaggerated. For reasons of audience taste and interest, most media output, and the organizations that provide it, remain obstinately linked to local, often national, markets, cultures, languages, and political structures. There are few examples of successful transnational newspapers or broadcasting news services. Even in the field of entertainment their spread is limited. This is itself, as the example of the European Union shows, a major barrier to the construction of a transnational public sphere. On the other hand, it can be argued that current developments on the World Wide Web, particularly its use for mobilization, debate, and advocacy by NGOs, are a welcome start in the construction of a global

public sphere. What certainly is the case is that the development of global media has begun to create a situation where nation-states, both within their own borders and in their inter-state relations, are having increasingly to take account of global public opinion (e.g., the cases of Tiananmen Square protests of 1989, the Balkan wars in 1990s, the 2003 invasion and subsequent occupation of Iraq, the Arab spring in 2011).

Third is the discursive question of how we can both analyze and normatively judge from the perspective of democracy the wide range of rhetorical forms used and identities appealed to within the range of modern media and the changing nature of publicness. The type of reflexive publicness developed with newspapers and reading publics cannot be compared with publicness as visibility that emerged with television. Television brought about a new type (or at least dimension) of publicness – direct visibility – and consequently changed the nature of the public and public discourse. The availability of broadcasting to almost every citizen (at least in the most developed countries of the time) changed the character of publicness not only because of the (re)introduction of the visual dimension, but also because it widened the notion of 'public interest' by the inclusion of ordinary people into the public domain. This process continues with the development of many-to-many mass communication enabled by the Internet. Here the debate around the public sphere reprises an older debate about mass culture.

The question at issue is whether current developments in media forms, genres, and their associated audiences/publics is a liberatory process of demassification and cultural pluralism on a global scale or, on the contrary, whether it is a process of 'dumbing down,' which in its appeals to irrationality and social fragmentation, undermines the rational/universalistic norms upon which citizenship, and thus democracy, rest. In the press, attention is focused on tabloidization, i.e., the stress on the scandalous and the sensational and the coverage of politics according to the norms of show business; in broadcasting on the shift in the balance of the programing mix in favor of entertainment at the expense of news, current affairs, and political coverage and debate, a shift associated, at least in Europe, with the retreat of public service broadcasting in the face of increasingly intense commercial competition for audiences and revenues.

Analysis and debate have focused in particular on the spread of audience participation shows of the Jerry Springer type, of 'docu-soaps' and of a general sound-bite coverage of politics and public issues. In this debate, there are no easy answers. First, empirically the trends are by no means as clear cut as many critics claim. It is as easy to argue for a rise in standards. Second, it is all too easy to fall into familiar elitist or populist stances, neither of which take due account of the complexity, both actual and historical, of the construction of the identity of citizens or of the rhetorical problems of communicating with the generality of such citizens. There is also often a danger of stretching beyond breaking point the proper and necessary boundaries of politics and the public sphere.

Fourth, there is the question of the role of the mediators. Public sphere theory finds it difficult to break free from a participatory model of direct democracy. Representation always introduces a potentially distorting element into the communication process between the private citizen and political whole. The public sphere is conceived of as an infrastructure for private citizens to publicly express their opinions and participate in public debates to which all have equal access. The ideal model is the Athenian agora. The scale, social complexity, and specialization of the modern world make such a model unrealistic. It has clearly never fitted the actual forms and practices of the (public service) media. It therefore leads to unrealistic demands being made upon them and to an excessively dismissive critique when these demands are not met. It has, in particular, led to exaggerated claims being made for the democratic potential of the Internet on the grounds precisely that cyberspace is a new agora without mediators or mediation. The challenge, therefore, is to think of a legitimate role for mediators and gatekeepers within the public sphere – for journalists, TV producers, bulletin board moderators, etc. – which is neither one of pure transparent 'objective' mediation nor one of the mere pursuit of instrumental professional or economic self-interest. This, in its turn, links to the wider question of the social and political role of intellectuals and experts within modern democratic polities.

The fifth question addresses the controversial relationship between publicness (free public expression and access) and secrecy (censorship) as its opposite, which became a burning issue with recent WikiLeaks' and E. Snowden's actions toward disclosure of 'classified, censored, or otherwise opaque documents' to the public record, by following the Benthamite and Kantian belief that "the public scrutiny of otherwise unaccountable and secretive institutions forces them to consider the ethical implications of their actions" (https://wikileaks.org/). Yet the 'ideology of publicness' is charged by others not to be able to acknowledge that, because of the persistent presence of legitimate secrecy, the public in the strict sense cannot exist. Rule- and decision-making bodies often try to avoid publicness, which they believe would slow down the process or make it less effective because of the inclusion of all potentially affected in public debate. Secrecy is believed to be well (legally) grounded for certain activities of (particularly global) governance such as small-group decision making, political negotiations, diplomacy, gathering of commercially, military, or other sensitive information. Others would argue that this is the conflation or even the swap of privacy with secrecy (secreted private interests, which I may not declare publicly), which for Kant was an exemplary instance (e.g., secret treaties) of an unjust maxim representing a threat to everyone and to universal peace. Although secrecy is essentially the negation of publicness, it is also codeterminative of publicness: by defining its limits, secrecy also constitutes publicness, whereas public reports by exposing certain issues to public attention at the same time hide from it issues considered unworthy of publication (Gouldner, 1976). Consequently, it is exactly actions directed against the conditions (re) producing secrecy, which can bring us ever closer to the counterfactual ideal of 'the public' rather than prove that it cannot exist.

See also: Attitudes, Political and Public Opinion; Civil Society and the Public Sphere: History of the Concept; Critical Theory: Contemporary; Critical Theory; Journalism; Legal Issues: Public Opinion; Mass Communication: Normative Frameworks; Mass Media, Political Economy of; Mass Media: Introduction and Schools of Thought; Mass Society: History of the Concept; Political Communication; Television: History.

Bibliography

Bentham, J., [1791]1994. Of publicity. Public Culture 6 (3), 581–595.

Bohman, J., 2004. Expanding dialogue: the Internet, the public sphere and prospects for transnational democracy. The Sociological Review 52 (Suppl. s1), 131–155.

Bücher, C., 1926. Gesammelte Aufsätze zur Zeitungskunde. Laupp'sche Buchhandlung, Tübingen.

Calhoun, C. (Ed.), 1992. Habermas and the Public Sphere. MIT Press, Cambridge, MA.

Castells, M., 2008. The new public sphere: global civil society, communication networks, and global governance. The Annals of the American Academy of Political and Social Science 616 (1), 78–93.

Dahlgren, P., 2005. The Internet, public spheres, and political communication: dispersion and deliberation. Political Communication 22 (2), 147–162.

Gray, J., 1995. Enlightenment's Wake: Politics and Culture at the Close of the Modern Age. Routledge, London.

Gouldner, A.W., 1976. The Dialectic of Ideology and Technology: The Origins, Grammar, and Future of Ideology. Seabury Press, New York.

Gripsrud, J., et al. (Eds.), 2010. The Idea of the Public Sphere: A Reader. Lexington Books, Lanham, MD.

Habermas, J., 1989. The Structural Transformation of the Public Sphere. Polity, Cambridge, UK.

Kant, I., [1784]1983. An answer to the question: what is enlightenment? In: Perpetual Peace, and Other Essays on Politics, History, and Morals (T. Humphrey, Trans.). Hackett, Indianapolis, IN.

Kant, I., [1795]1983. Perpetual Peace, and Other Essays on Politics, History, and Morals (T. Humphrey, Trans.). Hackett, Indianapolis, IN.

Lippmann, W., 1925. The Phantom Public. Harcourt, Brace and Co., New York.

Marx, K., [1842]1974. Die Verhandlungen des 6. rheinischen Landtags. Debatten über Pressfreiheit und Publikation der Landständischen Verhandlungen. In: Marx–Engels Werke, vol. 1. Dietz Verlag, Berlin, pp. 28–77.

Marx, K., [1843a]1974. Rechtfertigung des ++-- Korrsepondenten von der Mosel. In: Marx–Engels Werke, vol. 1. Dietz Verlag, Berlin, pp. 172–199.

Marx, K., [1843b]1974. Das Verbot der Leipziger Allgemeine Zeitung. In: Marx–Engels Werke, vol. 1. Dietz Verlag, Berlin, pp. 152–171.

Peters, J., 1993. Distrust of representation: Habermas on the public sphere. Media, Culture and Society 15 (4), 541–572.

Schudson, M., 1998. The Good Citizen: A History of American Civic Life. Free Press, New York.

Splichal, S., 2011. Transnationalization of the Public Sphere and the Fate of the Public. Hampton, New York.

Tönnies, F., [1887]1957. Community and Society (C.P. Loomis, Trans.). Michigan State University Press, East Lansing.

Tönnies, F., 1922. Kritik der öffentlichen Meinung. Julius Springer, Berlin.

UNESCO, n.d. Public Service Broadcasting. http://portal.unesco.org/ci/en/ev.php-URL_ID=1525&URL_DO=DO_TOPIC&URL_SECTION=201.html.

Public Sphere: Eighteenth-Century History

R Chartier, Ecole des Hautes Etudes en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12590–12641, Elsevier Ltd, with revisions made by the Editor.

Abstract

This article examines the emergence of a public sphere in eighteenth-century Europe. Its first part is devoted to the successive or competing historical and historiographical definitions of what the 'public' considered either as located within the new forms of intellectual and polite sociability (salons, clubs, masonic lodges), or as opposed to the private use of the reason and made possible by the circulation of written and printed culture, or as conceived in contrast with the popular opinions. The works by Jürgen Habermas, Immanuel Kant, and the French philosophers of the Enlightenment illustrates these different perspectives. The second part of the article identifies the metaphors used for designating public opinion in the eighteenth century. It is conceived either as a unanimous voice to be heard or as a tribunal of which the judgments are endowed with a supreme authority. Defined as a conceptual entity, public opinion was incarnated by different printed genres the memories published by the lawyers, which transformed particular causes into political or social issues, or the pamphlet and libels, which condemned in the name of the opinion the despotism and depravation of the monarchy. The article finally stresses the limits implied by such a conception of the public sphere: on the one hand, the popular milieux are excluded from the public use of the reason reserved to literate and learned individuals, and, on the other hand, the women are deprived since the 1770s of the role they played before in the absolutist or enlightened salon society.

Any reflection trying to understand how the notion of public sphere was constructed in the eighteenth century is necessarily referred to the classical book published by Jürgen Habermas in 1962 (Habermas, 1962). According to Habermas, the 'political public sphere' was construed in the eighteenth century (in some places sooner, in others later) as a space for discussion and exchange removed from the control of the state. The new 'public sphere in the political realm,' founded on the exercise of public criticism was thus opposed to the 'sphere of public authority' identified with the exercise of state power. For Habermas, this new intellectual and political 'space' was distinct from the court, which belonged within the domain of state power, and from the common opinions of the people who had no access to enlightened critical debates. This is why it is possible for him to qualify it as 'bourgeois.'

Definitions of the Public

There were different ways of conceiving the construction of the new political public sphere of the eighteenth century. Was it issued from the literary controversies of the seventeenth century and rendered possible by the diffusion of new forms of intellectual sociability? Or was it the extension of the modalities of print communication proper to the Republic of Letters to a greater number of individuals, which made possible the transformation of dispersed private persons into a unified public community?

Habermas's Political Public Sphere

For Habermas such a 'public' supposed, on the one hand, that the various participants were not acting as agents or subjects of the ruler and, on the other hand, that they were considered as equal. The political public sphere thus ignored both the obedience required by the exercise of state power and the distinctions of 'orders' and 'estates' imposed by the traditional social hierarchies.

No limit could subject the critical exercise of reason, and no domain was to be forbidden to it. Overcoming the division instituted by the cartesian methodical doubt between opinions that could legitimately be submitted to examination and obligatory credences and obediences, the members of the new public sphere considered that no domain of thought, including the political and religious ones, was to be removed from critical and reasonable judgment.

This judgment was exercised within the different forms of sociability, which had been previously erected as a tribunal of aesthetic criticism: the salons, the cafés, and the clubs. In the seventeenth century, this new public challenged the traditional monopoly of the court, the official academies, or the connoisseurs in such matters. It was enlarged during the eighteenth century, thanks to the multiplication and diffusion of the periodicals, and the success of new forms of association: the book clubs, the reading societies, the masonic lodges. But it supposed also the exclusion of all those who lacked the competence which made possible the participation into a critical community welcoming only the persons who had sufficient wealth and culture.

Kant's Public Use of Reason

Habermas's perspective was based on one of the most famous text of the eighteenth century, Kant's *What Is Enlightenment?* published as an article in the periodical *Berlinische Monatsschrift* in 1784. But it also distorted it. In his text Kant proposed a new way of conceptualizing the relation of the public to the private. He equated the public exercise of reason with judgment produced and communicated by private individuals acting 'as scholars before the reading public' and he defined the public as the sphere of the universal in opposition to the private considered as the domain of particular and domestic interests (which may even be those of an army, a church, or a state).

He inverted, thus, the accepted meanings of the two terms by associating the private use of reason with the exercise of an office

or with the duties toward the state and by locating the public use of reason within a community without limitations or boundaries. If the private use of reason can be restrained legitimately in the name of public ends without prejudice to the progress of enlightenment, the liberty of the public use of the reason was absolute and has to be respected by the prince himself.

In this same text, Kant shifted the way in which the legitimate limits put on critical judgment should be conceived. Such limits were no longer defined by domains forbidden to methodical doubt; they laid in the position of the subject who exercises his reason. He could be constrained legitimately when he was executing the duties of his charge or of his status, but he was necessarily free when he acted as a member of the 'society of world citizens.'

That universal society was unified by the circulation of written works that authorized the communication and discussion of thoughts without the necessity of physical presence of the different interlocutors. Kant systematically associated the public use of reason with the production or reading of written matter by private individuals. The public was not construed on the basis of new forms of intellectual sociability, such as clubs, cafés, literary societies, or lodges, because those associations were modalities of particular, circumscribed, and domestic entities. For him, only written communication was admissible as a figure for the universal.

Kant's conception of the public use of reason was drawn for the notion of the Republic of Letters as it was constructed through correspondence and the circulation of manuscript or printed texts from the late seventeenth century onwards (see Goldgar, 1995). Founded on the free engagement of its members, on equality among its interlocutors, and on reciprocity and disinterestedness, the Republic of Letters which united the learned and the scholars provided a model and a support for thinking of the new public sphere. At the same time, it showed the distance separating the universality of the 'society of world citizens' and the restricted social composition of the 'reading public.' In Kant's time, this reading public was not the whole of society by any means, and those capable of writing were even less numerous. It is only when everyone will be able to exercise critically one's reason that humanity will experience 'an enlightened age.'

The French Philosophers: Public Opinion versus Popular Opinions

Kant held the distinction between the public and the people to be temporary, transitory, and characteristic of a century which was an 'age of enlightenment' but not yet an 'enlightened age.' But for many thinkers of the eighteenth century, particularly within the framework of the French Enlightenment, the two notions constituted an irreconcilable dichotomy (see Ozouf, 1989: pp. 21–35). Public opinion was defined precisely in contrast to the opinions of the multitude. Public opinion, set up as a sovereign authority and a final arbiter, was necessarily stable, unified, and founded on reason. The universality of its judgments and the constraining self-evidence of its decrees derived from that unvarying and dispassionate constancy. On the other hand, popular opinions were multiple, versatile, and inhabited by prejudice and passion. Such a qualification reveals a strong persistence of older representations of the people considered as subject to extremes, inconstant, and foolish.

When the concept of public opinion did emerge in France around 1750 as the superior authority to which all particular opinions must bow, the distinction between the public and the popular became essential. A new political culture thus took shape, which transferred the seat of authority from the will of the king alone, who decided without appeal and in secret, to the judgment of an abstract entity embodied in no institution, which debated publicly and was more sovereign than the sovereign who was obliged to seek to win its approval and support.

This gave acuity and urgency to a fundamental question: how could the people, eliminated from the exercise of critical judgment by their lack of competence, be 'represented' in the new political space? A first answer to this question assigned this function to one or another of the theories of political representation confronted after the mideighteenth century. Following Keith Baker, three of the theories were fundamental: the absolutist theory, which made the person of the king the only possible representative of a kingdom divided into orders, estates, and bodies; the judiciary theory which instituted the Parliaments as interpreters of the consent or remonstrances of the nation; and the administrative theory, which attributed the rational representation of social interests to municipal or provincial assemblies founded not on privilege but on property (see Baker, 1990).

An alternative model of representation that removed the notion institutional setting – monarchical, parliamentary, or administrative – was offered by the category of public opinion. Substituting the self-evidence of unanimity to the uncertainties of particular and popular opinions, detached from any form of exercise of governmental and state authority, the concept delegated the representation of the whole nation to those who were able to express its reasonable decrees: the men of letters.

Public Opinion: Voice or Tribunal?

Although everyone recognized the existence of public opinion and postulated its unity, there was no unanimous consent on two issues: first, who were its true spokesmen and, second, how was one to evaluate the self-evidence of its judgments? In eighteenth-century France, the answers to such questions were expressed through the metaphors used to designate this new entity: public opinion as a voice to be heard; public opinion as a tribunal, which had to be persuaded.

From Theater to Tribunal

In 1775, in his *discours de réception* before the *Académie francçaise*, Malesherbes forcefully expresses the idea, by then commonly accepted, that public opinion was to be considered a court of justice more imperious than any other. It is a tribunal 'independent of all powers and that all powers respect,' composed by the men of letters who are 'amid the public dispersed, what the orators of Rome and Athens were in the middle of the public assembled.'

There are several arguments contained in this comparison. First, it invested the new 'judges' with an authority that ordinary judges did not have. Their competence knew no bounds and their jurisdiction no limits; their freedom of judgment was

guaranteed because they were in no way dependent upon the power of the ruler; their decrees had the force of self-evident propositions. With such a definition of public opinion, the men of letters were invested with the judiciary legitimacy of all traditional powers and with veritable public office.

Reference to this dispersed judiciary function had another meaning, however. It aimed at establishing a connection between the universality of judgments and the dispersal of persons, and at constructing a uniform opinion, which, unlike that of the ancients, had no physical location in which it could construe and experience its unity. As for Kant later, it was the circulation of written matter and particularly the printed one, which made it possible to envisage the constitution of a unified public in a nation in which people were necessarily separated from each other and formed their ideas individually. By associating the public nature of the written word, vastly increased by the invention of the printing press, with the supreme authority of the judgments pronounced in the new political sphere, Malesherbes (as many others) converted the plurality of reflections and criticisms that emerged from solitary reading into a collective and anonymous conceptual entity.

From the seventeenth century to the eighteenth century there had been a radical shift in the manner of conceiving the public. In the age of baroque politics the traits that defined the public were the same as those which characterized theatrical audiences: heterogeneous, hierarchized, and formed into a public only by the spectacle in which the most powerful effects were produced by hidden machines and secret maneuvers.

When the concept of public opinion did emerge, it effected a dual rupture. It opposed the art of dissimulation and secrecy by appealing to transparency. Before the tribunal of opinion all causes were to be argued without duplicity. But all citizens were not (or not yet) adept at joining together to form enlightened opinion. If the audience which mingled in the theaters potentially was composed of men and women from all social levels, the public which served as the tribunal to judge literary merits, religious matters, or political issues was more selected and homogeneous. When opinion was thought of more as actor than as acted upon, it became public, and by that token it excluded many people who lacked the competence to establish the decrees, which it proclaimed.

From Audience to Public

Constituting the public as an entity the decrees of which had more force than those of the established authorities supposed several operations. They can be exemplified by the memoirs published in great number by both French lawyers and litigants from 1770 onward. They took the judicial comparison literally and exposed before opinion affairs, which were examined by a court of justice. The cases subjected to the secret procedures of justice were thus transformed into public debates, shifting the place in which judgment had to be pronounced (see Maza, 1993).

The most fundamental strategy required by such an operation consisted in endowing particular cause with general and exemplary value. The debt that a court noble refused to pay to his bourgeois creditors became an ideal occasion for denouncing unjust privilege, just as the arbitrary imprisonment of a young gentleman was an opportunity to criticize the *lettre de cachet*. To achieve such goals supposed, first, that the secrecy of judicial procedure had to be broken by the circulation of printed memoirs on the largest possible scale, and, second, that a dramatic style or a first-person narrative had to replace the customary legal prose. Universalizing the particular, making public what had been secret, and 'fictionalizing' discourse were the techniques that lawyers used to appeal to opinion and, in doing so, to proclaim themselves the authorized and legitimate interpreters of that opinion.

The traditional, discrete, and exclusive relationship that bound individuals to the king, the guarantor and guardian of domestic secrets, gave way to a totally different situation in the public exposition of private conflicts. From that point of view, judicial memoirs are the exact inverse of the *lettres de cachet* accorded by the sovereign in response to requests from families interested in concealing familial disorders, which sullied their honor. The memoirs displayed what the *lettres* concealed; they expected from the judgment of opinion what the *lettres* hoped to gain from the omnipotence of the monarch; they converted into public issues the scandals that the *lettres* were charged with burying.

It is in this sense that the new public space was built on conflicts proper to the private sphere, assigning political significance to simple familial or conjugal strife. In a parallel manner, the clandestine pamphlets stigmatized the aristocracy, the courtiers, the queen, and finally the king by revealing their corrupt mores. People became accustomed to seeing individual affairs transformed into general causes. This procedure was not restricted to pamphlet literature but also lay behind the lawsuits instituted by rural communities against theirs landlords or by journeymen workers against their masters. It was the process of 'privatization' in which individuals conquered autonomy and freedom removed from state authority, which permitted the very existence of a new public sphere.

Even if, or because, it was defined as a conceptual entity, and not in sociological terms, the notion of public opinion which invaded the discourse of all segments of society – political, administrative, judicial – in the second half of the eighteenth century operated as a powerful instrument for social division and legitimization. The mobilization of the category founded the authority of all who, by affirming that they recognized its decrees alone, set themselves up as mandated to pronounce its judgments. Universal in its philosophical essence, the unified, enlightened, and sovereign public was far from universal in its social composition. The public sphere, emancipated from the domain in which the ruler held sway, thus had nothing in common with the shifting opinions and blind emotions of the multitude. Between the people and the public there was a fundamental difference.

Prohibited Books, Subversive Words

According to Darnton, this new public was detached from the traditional authorities by a wider and wider dissemination of philosophical ideas. In this process the role played by the great classics of the Enlightenment was perhaps less important than the texts that the eighteenth-century book trade called 'philosophical books': critical pamphlets, scandal mongering

chronicles, anticlerical satires, pornographic works. The circulation of this clandestine and seditious literature, which depicted the despotic corruption of the monarchy or the depraved mores of the court, engendered a change in thought and in collective representations, henceforth detached the subjects from their old loyalties: "Sedition was hatching. It was instilled in people's minds […] We know for certain that it was communicated by a formidable instrument: the book" (see Darnton, 1991, 1995).

But other historians like Farge have challenged this interpretation that sees the increase in seditious writings in the last two or three decades of the *ancien régime* as the matrix of a desacralization of the monarchy. She affirms: "Popular opinion did not emerge from the cumulative reading of pamphlets and placards; it was not unilinear and did not base arguments on the sum of what is read" (see Farge, 1994: p. 35).

Statements of disapproval or hatred of the sovereign, attacks on his acts, and expressions of a desire to kill him did not arise from the flourishing clandestine literature. Nor did they start with Damiens's attempt to assassinate Louis XV, which, rather than unleashing a proliferation of regicide discourse, persuaded the authorities of the reality of Jansenist and Jesuit plots and set them in pursuit of such literature. For Farge, "the failed murder of Louis XV had a ready-made public opinion; it tells us more about the monarchy's reactions than about any new and original turn of popular thought" (p. 174).

The symbolic and affective disinvestment, which transformed relations with authority when it was deprived of all transcendence was manifested but not caused by the wide diffusion of 'philosophical books.' The erosion of founding myths of the monarchy, the desacralization of royal symbols, and the distancing of the person of the king formed a body of representations which was 'already there,' 'ready made' when the public sphere arose conceptually and sociologically.

Women and the Public Sphere

What was the role played by women in this new public sphere? Joan Landes' thesis opposes radically the "symbolic politics of the emerging bourgeois public sphere (which) was framed from the outset by masculinist interests and assumptions" with the forms of sociability dominated by women in the *ancien régime*, for example, the salon society conceived as "an alternative sphere of cultural production inside absolutism" (Landes, 1988: p. 40). This is the reason why the exclusion of women from political rights during the French Revolution is inscribed by Landes within the logics of the enlightened condemnation of the role that women played within the *absolutist* public sphere: "the structures of modern republican politics can be construed as part of an elaborate defense against women's power and public presence" (p. 203).

Against such a perspective Goodman argues that women played a central role both in the collective project of the Enlightenment and in the construction of a new political and critical public space. For her, "the French Enlightenment was grounded in a female-centered mixed-gender sociability that gendered French culture, the Enlightenment, and civilization itself as feminine" (Goodman, 1994: p. 6). In this perspective the salon society did not belong to the old regime absolutist public sphere but had a fundamental role in shaping a 'discursive space' opened to public and political criticism. Between 1750 and 1775 women were essential in the government of such society because they kept within the limits of civility and politeness the tensions, which necessarily arose from the free confrontations of opinions and the competitions for intellectual leadership: "Enlightenment salons were places where male egos were brought into harmony through the agency of female selflessness" (p. 104).

According to Goodman (1994: p. 280), it is during the 1770s that women began to be excluded from the new public sphere. On the one hand, the harmony of Republic of Letters was then destroyed by a series of a ferocious discords starting with controversies about Physiocracy. On the other hand, new forms of association (among them, literary societies called *lycées* or *musées* and masonic lodges) either excluded women or gave them a subordinate position. The return to 'masculine self-governance' prepared women's expulsion from the revolutionary political space: "When the literary public sphere was transformed into the political public sphere in 1789, it had already become masculine; the 'democratic' republic of 1792 would reflect the limitations and exclusions of the Republic of Letters of the 1780s."

It is possible to discuss such an interpretation. The Freemasonry did not ignore women's presence and gave them an important role in the 'lodges of adoption.' In these lodges many women were involved in the new values coined by the late Enlightenment: the philanthropic activities, the interest for political issues, the friendship between men and women (Burke and Jacob, 1996: pp. 513–49). Besides that, it is clear that even between 1750 and 1775, the salon society was always very minoritary and limited to a restricted number of participants. They cannot be considered as the center of the Republic of Letters.

The collective construction of the new public sphere cannot be reduced either to a unique process or to a sole place. It arose from the intertwining between conceptual novelties and social practices, the alliance between speech and print, and the common (but dissimilar) participation of men and women.

See also: Absolutism, History of; Civil Society, Concept and History of; Enlightenment; Intellectual History; Kant, Immanuel (1724–1804); Public Sphere: Nineteenth and Twentieth-Century History; Space and Social Theory in Geography.

Bibliography

Baker, K.M., 1990. Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century. Cambridge University Press, Cambridge, UK.

Burke, J.M., Jacob, M.C., 1996. French freemasonry, women, and feminist scholarship. Journal of Modern History 68 (3), 513–549.

Chartier, R., 1991. The Cultural Origins of the French Revolution. Duke University Press, Durham and London.

Crow, T.E., 1985. Painters and Public Life in Eighteenth-century Paris. Yale University Press, New Haven and London.

Dalton, S., 2003. Engendering the Republic of Letters: Reconnecting Public and Private Spheres in Eighteenth-century Europe. McGill-Queen's University Press, Montreal, Que.

Darnton, R., 1991. Edition et sédition. L'univers de la littérature clandestine au XVIIIe sicle. Gallimard, Paris.

Darnton, R., 1995. The Forbidden Best-sellers of Pre-revolutionary France. Norton, New York.

Farge, A., 1994. Subversive Words: Public Opinion in Eighteenth-century France. Polity Press, Cambridge, MA.

Goldgar, A., 1995. Impolite Learning: Conduct and Community in the Republic of Letters 1680–1750. Yale University Press, New Haven.

Goodman, D., 1994. The Republic of Letters: A Cultural History of the French Enlightenment. Cornell University Press, Ithaca.

Habermas, J., 1962. Strukturwandel der ffentlichkeit. Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft. Luchterhand, Neuwied (English Trans., 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. Polity Press, Cambridge, MA).

Kant, E.I., 1975. Foundations of the Metaphysics of Morals and what Is Enlightenment (1784). Bobbs-Merrill, Indianapolis.

Kwass, M., 2000. Privilege and the Politics of Taxation in Eighteenth-century France: Liberté, Égalité, Fiscalité. Cambridge University Press, Cambridge.

Landes, J.B., 1988. Women and the Public Sphere in the Age of the French Revolution. Cornell University Press, Ithaca.

Maza, S., 1993. Private Lives and Public Affairs: The Causes Célbres of Prerevolutionary France. University of California Press, Berkeley, CA.

Melton, J.V.H., 2001. The Rise of the Public in Enlightenment Europe. Cambridge University Press, Cambridge, UK.

Nechtman, T.W., 2010. Nabobs: Empire and Identity in Eighteenth-century Britain. Cambridge University Press, Cambridge.

Ozouf, M., 1989. L'homme régénéré. Essais sur la Révolution française. Gallimard, Paris.

Pollock, A., 2009. Gender and the Fictions of the Public Sphere, 1690–1755. Routledge, New York.

Roche, D., 1988. Les Républicains des Lettres. Gens de culture et Lumires au XVIIIe sicle. Fayard, Paris.

Sauter, M.J., 2009. Visions of the Enlightenment: The Edict on Religion of 1788 and the Politics of the Public Sphere in Eighteenth-century Prussia. Brill, Leiden.

Shimko, R.B., Freeman, S., 2012. Public Theatres and Theatre Publics. Cambridge Scholars Pub, Newcastle upon Tyne.

Swaim, B., 2009. Scottish Men of Letters and the New Public Sphere, 1802–1834. Bucknell University Press, Lewisburg, PA.

Wittman, R., 2007. Architecture, Print Culture, and the Public Sphere in Eighteenth-century France. Routledge, New York.

Wittman, D., 2012. Architecture, Print Culture and the Public Sphere in Eighteenth-century France. Routledge, London.

Public Sphere: Nineteenth and Twentieth-Century History

C Calhoun, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12954–12959, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

The development of institutions and practices of open communication concerning affairs of common interest is basic to democracy. This is analyzed as involving transformations of public space from the Athenian agora to the metaphorical space of electronic communications. These spaces provide for communication among strangers as well as those joined by bonds of community; they are arenas for the expression of individual and collective identities as well as for the making of collective decisions. Analytic attention focuses largely on questions of access to publics, which is often restricted on bases such as gender, ethnicity, or citizenship. It focuses also on institutional supports for public communication, on the relationship between rational-critical communication and other forms of action in public, and on the differentiation of public communication into multiple partial publics.

The closely related concepts of public sphere and public space are rooted in ancient Greek distinctions of public and private and developed in early modern Europe through changing ideas of citizenship, civil society, and republican virtue (*see* Civil Society and the Public Sphere: History of the Concept). They refer first and foremost to settings for political discussion among citizens about the order of their lives together, and secondarily to a much wider range of public culture including art, literature, and popular discourse on a host of subjects.

Publicness as a Modern Concern

New ideas about public discourse were complemented by development of new communications media, especially those dependent on print; rising literacy and education levels; growth of the state; and expansion of popular political participation. In this process, the distinction of public and private took on new importance and complexity. On the one hand, the realm of public interaction expanded; cities were the primary setting for this, especially cosmopolitan trading and capital cities. Public spaces appeared literally with coffee houses, parks, theaters, and other places where people who were not bound by private relations gathered and communicated. They also grew metaphorically with printed sermons, pamphlets, newspapers, books in vernacular languages, journals that reviewed them, and other media of public communication. On the other hand, the state also expanded and with it the range of *res publica*, public things, which included property held in common and matters of concern to the whole polity. Publicness took on a dual sense, referring both to openness of access and interaction and to collective affairs as managed by the state. The public referred both to the collective subject of democracy – the people organized as a discursive and decision-making public – and as its object – the public good.

The two dimensions were linked in the notion that political debate among responsible citizens was a way to arrive at sound understanding of common affairs. This depended on notions of participation in debate that developed in the realms of science and religion (Ezrahi, 1990; Zaret, 1999) and literature (Habermas, 1962; Hohendahl, 1982) as well as politics. Processes of rational-critical debate were held to form educated public opinion as distinct from other forms such as the 'representative publicity' of monarchs appearing before their subjects or the 'mere opinion' of uneducated masses. Interest in such public opinion grew alongside civil society as a self-organizing realm of social relations, and especially with the rise of democracy.

Complementing the growth of the public realm were new senses of the private. The 'privacy' of the family was more sharply affirmed. Women and children were increasingly sequestered in private homes, especially among the bourgeoisie; the modern idea of the individual included the notion of nurturance and maturation in private as preparation for action in public, but this applied in the first instance to men. At the same time, economic activity was increasingly moved out of the household. In one sense, then, going to work meant going out into public, being exposed to the public gaze. In another sense, however, property relations continued to be understood as private in the sense that they were to be managed by individual persons and not the state. The eventual rise of business corporations further complicated the distinction, since these held property as artificial private persons but operated as collective, public actors, especially when shares of ownership were openly available on the market rather than closely held within families.

As the example of corporations and private property suggests, the distinction of public and private was sometimes difficult to sustain. This undermined the classical notion of the public sphere: "The model of the bourgeois public sphere presupposed strict separation of the public from the private realm in such a way that the public sphere, made up of private people gathered together as a public and articulating the needs of society with the state, was itself considered part of the private realm" (Habermas, 1962: pp. 175–176). As corporations became public actors while still claiming the status of private property, states began to intervene ever more into civil society and even intimate life. These trends joined with the rise of mass media, and especially professions and institutions devoted to the manipulation of public opinion through mass media (advertising, public relations) to undermine the conditions for the effective operation of the public sphere as a source for educated public opinion.

History as Decline

For many later analysts, the eighteenth century was a sort of 'golden age' of modern public life (*see* Public Sphere: Eighteenth-Century History). Hannah Arendt (1958) theorized 'public' in terms of creative action, the making of a world shared among citizens, and saw the founding of the USA as a crucial example. Jürgen Habermas (1962) idealized eighteenth-century English parliamentarianism, newspapers, and coffee house conversation. He presented the public sphere as a realm of civil society in which private citizens could communicate openly about matters of public concern, transcending their particular statuses and addressing the state without becoming part of it. Others have focused on Parisian clubs in the Revolution of 1789 (Amann, 1975) or the literary journalism of the German Enlightenment (Hohendahl, 1982).

Idealization of the eighteenth- and sometimes the early nineteenth-century public sphere underwrites explicit or implicit narratives of decline. In Habermas's classic *Structural Transformation of the Public Sphere*, for example, nineteenth- and twentieth-century public discourse is analyzed in terms of the loss of rational-critical capacity attendant on expansion of scale and the rise of public relations management that incorporated the public into the realm of administered society (see also Horkheimer and Adorno, 1944). Such accounts reveal that the notion of public is not only basic to democracy, but also linked to an apparent contradiction in democratic ideals. At least in its most prominent modern, liberal forms, democracy seems to promise both a politics based on reason rather than power, and the inclusion of all adults in the political process. The latter, the classic accounts of the public sphere suggest, undermines the former. The nineteenth- and twentieth-century history of democratic politics is taken to demonstrate this. Public opinion is increasingly 'managed' on the basis of research and manipulation rather than developed in open, rational-critical discourse.

Already in the nineteenth century itself, political theory of many persuasions emphasized the fragility and limitations of the liberal democratic conception of the public. Tocqueville, most famously, argued as early as the 1830s that the democratization of society tended to eliminate the *intermediary* public bodies, which had a traditionally refined opinion and furnished the individual as a collective social identity outside the state. Engaged, politicized publics composed of distinct views and interests could be reshaped over time into mass publics – passive, conformist, and atomized before the state. Tocqueville's fear of the unmediated state would resonate with generations of critics of mass society.

Central to the analyses of historical decline in the public sphere is the idea that as its scope expanded, it was unable to sustain a high level of rational-critical debate. This happened with the rise of various sorts of opinion managers, from pollsters to public relations specialists, and with the institutionalization of political parties as products of negotiations among interest groups rather than arenas for debate about the public good in general. A key phase in this, according to the critics, was the institutionalization of class politics during the late nineteenth and early twentieth centuries. Habermas treated this in terms of negotiation among interests, as distinct from reasoned debate about the public good. Negt and Kluge ([1964]1993), however, famously emphasized the existence of a proletarian public sphere, rooted in distinctive experiences, and productive of reasoned internal discourse. Indeed, the German Social Democratic Party, like successful workers organizations elsewhere, created a parallel public sphere to that dominated by the bourgeoisie (Eley, 1992; see also Katznelson and Zolberg, 1987). This faltered principally because of a nationalization of popular politics, not class politics per se (Mosse, 1975, 1987). Such nationalization was often accompanied by mass spectacle and reliance on mass media. Both did frequently collapse the distinction of public and private, especially in emotive appeals to national unity. Equally important was the rise of two-party electoral systems. Rather than being able to participate activity in public life, late-twentieth-century citizens of democratic societies were reduced to being invited to respond to plebiscites (Lowi, 1985; Ginsburg, 1986).

Historical research on the public sphere has questioned both the accuracy of the late-eighteenth-century ideal and the narrative of subsequent decline. Schudson (1998), for example, has argued that active participation of citizens in public discourse and politics has ebbed and flowed without linear trend. The ideal of the good citizen as an active participant in the public sphere has long been contrasted with the failings of actual citizens. A substantial line of research on 'civic culture' shifted the emphasis away from the problematic of public debate and toward a more general idea of participation and incorporation into civic life, for example, in voluntary associations (see, in particular, Almond and Verba, 1966 and critical analysis in Somers, 1993, 1995). This, as most researchers argued, grew through the nineteenth and early twentieth centuries. More recently, however, critics of late-twentieth- and early twenty-first-century democracies have emphasized a decline in civic participation (e.g., Putnam, 2000).

Participation and Exclusion

Although openness was basic to the ideology and theory of the public sphere, various forms of exclusion were basic to actually existing publics in the nineteenth and twentieth centuries. Women, for example, had been prominent in late-eighteenth-century salon society and elite public spheres, but were widely excluded from nineteenth- and twentieth-century public space; this was mirrored by a commensurate bias in political thought (Landes, 1988; Elshtain, 1993). Feminist struggles made gender a basic political issue in new ways (Ryan, 1990, 1992). In addition, women were prominent in other movements, such as abolitionism in the USA, and may have figured less in subsequent historical accounts than in actual nineteenth-century social movements. This may be partly because of different styles of address in the public sphere, including a greater emphasis on narrative and less on argumentation (Lara, 1999). Similarly, Blacks were largely excluded from the US public sphere, even after abolition. This encouraged a similar duality: struggles for inclusion and development of parallel public discourse marked by distinctive communicative styles (Diawara, 1995).

This has occasioned more general inquiry into 'counter-publics,' which contest the apparently neutral dominance of more mainstream publics (Fraser, 1992; Warner, 2001). This

argument is rooted in Negt and Kluge's ([1964]1993) account of the proletarian public sphere. Members of religious communities, ethnic groups, and the women's movement have been among the many who have constructed counterpublics. During the classic heyday of the late eighteenth- and early nineteenth-century British public sphere, thus, many artisans and workers were denied participation in the public sphere. They were not simply and unambivalently members of a proletarian public sphere, though they did develop their own media and organizations and to some extent constitute a counterpublic (Gilmartin, 1999). They claimed the right to participate in the dominant, unmarked public sphere, and challenged those who introduced restrictive measures to make it a specifically 'bourgeois' (or more generally, propertied) public sphere (Jones, 1983). The same people who excluded those with less wealth from the public sphere nonetheless claimed it in unmarked form as simply *the* British public.

Accounts of the public sphere by both contemporaries and historians have been tacitly linked to the idea of nation, since nation defined the range of citizens entitled to participate in public affairs and the basis for legitimacy of modern, especially democratic, states (Calhoun, 1997). In liberalism's triumphant early phase, experiments in republicanism and constitutionalism placed these issues squarely on the table. They asked how new political structures were to acquire legitimacy and how, within those structures, the plurality of opinion was to be represented. Defenses of free speech and the press, too, signaled an appreciation of the importance of venues in which views could be widely and, in principle, freely circulated. At the same time, the rhetoric of nationalism typically presented the nation not as a creature of public discourse but as a preexisting unity, commonly on ethnic or cultural grounds. This is true even, if ironically, in the USA, the nation quintessentially created through the public action of its citizens. A tension running through American history has pitted the capacity for creation of a new political order exemplified by the founders against claims that the defining national essence is already in place and needs only defense from the loyal.

Notions of a vital republican public sphere have been influential especially in American historical studies, visible in the work of J.G.A. Pocock, Bernard Bailyn, Gordon Wood, and more recently, Michael Warner. Here the focus, in Hannah Arendt's terms, is on the capacity of action in public not only to define individual identity but also to create the world citizens share in common. Republicanism, implied several things: (1) a concerted effort to reach agreement about the public good, as distinct from the attempt to find the net balance of private, individual goods – the essence of utilitarianism; (2) a sense of politics as a work in progress regarding the appropriate structures and limits of the political, as opposed to a hallowed set of procedures for achieving particular ends – such as the cult of the constitution; and (3) a belief that individuality and identity are only fully expressed and enjoyed through engagement in public life, rather than constituted beforehand in the private realm. These writers emphasize the ephemeral nature of republicanism and the rapid emergence in the nineteenth century of a variously liberal, national, and representative (rather than participatory) model of the public sphere. In American Studies, as in Frankfurt School critical theory, this decline is commonly associated with the rise of mass culture – and the blurring of divisions between reasoned political discourse and other social processes of opinion formation.

Cities and Media

Transformations of public life were also shaped by the ways in which communications media and the physical public spaces of cities changed through the nineteenth and twentieth centuries. Richard Sennets (1977) *The Fall of Public Man*, for example, contrasts a vibrant and highly varied eighteenth- and nineteenth-century urban life (especially among the bourgeoisie) to a reduction in both the occasions for interaction across lines of difference and the psychological ability to manage open relations with strangers well. Citizens retreated, Sennett suggests, into both privacy and the conformity of mass culture.

Urban public spaces anchor face-to-face interaction, which was basic to the classical image of the public sphere, conceived of largely in terms of direct interaction. This is important for a sense of serendipitous contact among people of diverse backgrounds. Parks, cafes, and theaters all inform this. Many of Europe's cities, especially older ones, were distinctive in their pedestrian character and their scale. From the mid-nineteenth century, however, this began to shift. Hausmann's post-1848 construction of the Paris boulevards and destruction of some of the local *quartiers* is taken as emblematic (e.g., by Harvey, 1985). Both suburbanization and less pedestrian-friendly, larger-scale designs for newer cities have changed the character of public interaction in Europe, largely in ways familiar earlier in the US Commentators throughout the nineteenth and twentieth centuries have been concerned for the implications of this for democracy. As Mumford (1938: p. 483) wrote, "One of the difficulties in the way of political association is that we have not provided it with the necessary physical organs of existence: we have failed to provide the necessary sites, the necessary buildings, the necessary halls, rooms, meeting places …."

This has been linked to a huge increase in scale. One crucial result is that directly interpersonal relations organize proportionately less of public life; mediations of various kinds have become increasingly important (Calhoun, 1988; Garnham, 1992; Keane, 1991; Thompson, 1995). This is partly a matter of communications media per se; it is also a matter of growing importance for bureaucracies and other formal organizations. The eighteenth- and early nineteenth-century public sphere was closely tied to the spread of print media. The nineteenth and early twentieth centuries were the heyday of great urban newspapers; after this, media that transcended locality became increasingly important. First radio and then television fundamentally altered the public sphere. They contributed to a shift in what was publicly visible as well as in how public discourse was organized (Meyrowitz, 1985; Thompson, 1995). New media shared both information and emotionally powerful images widely. Broadcast media are often charged with furthering the debasement of reason by substituting powerful images for sustained analysis, by appealing to a lowest common denominator in audiences, by blurring the lines between entertainment and critical

discourse, and by centralizing control over messages in the hands of a few corporations. Some caution is in order about generalizations from medium, though, as newspapers produced their own mass forms, with broadly similar if perhaps less extreme effects.

Crucially, the changes in scale and media forms of the twentieth century challenge the classical eighteenth-century model of the public sphere, which was grounded in readership on the one hand and directly interpersonal discourse on the other. Any account of contemporary public life must deal with a dramatic increase in indirect, mediated relationships. Transformations of media continue, of course, and the implications of the Internet and computer-mediated communication are important topics of contemporary investigation (Rasmussen, 2000). Relatedly, the dominant approaches to the public sphere have emphasized domestic politics within nation-states. Especially during the last years of the twentieth century, however, communications media and the rise of transnational organizations have facilitated growth of an international public sphere (Thompson, 1995; Held, 1995; Calhoun, 2002). To what extent transnational public opinion can steer global society directly, and to what extent only through the agency of state structures, remains a crucial question.

Finally, though this article concentrates on the public sphere in Western European and American history, it is important to note that the nineteenth and twentieth century also saw dramatic and transformative growth in public discourses elsewhere. In late Imperial and early Republican China, for example, the category of 'public' became influential and both directly interpersonal and mediated public spheres were created (Wakeman, 1993; Rankin, 1993). This was basic to development of the idea of citizen, as distinct from subject, to constitution of notions of public welfare, and to cultural transformations such as the introduction of new vernacular writing. It also contributed to the growth of both nationalist and class politics in the twentieth century. Further transformations followed, attendant on changes in media, politics, and culture, and creating important linkages within the Chinese diaspora as well as with the West.

An unusually vital print-mediated public sphere developed in India. The early history of this is closely interwoven with the history of British colonialism and opposition to it. The history of Indian public life has been an important setting for examining the role of both international power relations and issues of domestic diversity. Not least, questions like the claims of nonreaders in a public sphere dominated by print media have surfaced and shaped not only Indian but international discussions. Broadcast media have played an important role recently in the constitution of a new Hindu nationalism (Rajogopal, 2001). In many settings, including in the European countries and the USA as well as both India and China, the development of a public sphere has been closely related to the development of national consciousness, identity, and political contestation.

Conclusion

Growth and transformation of public spheres has been a major theme in nineteenth- and twentieth-century history. It is crucial to the history of democratic politics, including not only electoral and parliamentary regimes, but also broader struggles in civil society and beyond national borders. A dominant historical theme is the inclusion of a broader range of participants in public discourse. This has been coupled with analyses of the implications of growing scale, including mismatch between local – especially urban – public spaces and enlarged range of potential participants, dynamics of inclusion and exclusion, and changing roles of media. A prominent line of analysis suggests a concomitant decline in the rational-critical capacity of public discourse. This is challenged by both research on the limits of earlier public spheres and the strengths of later counterpublics. The importance of the public sphere is in any case not limited to argumentation but includes also the formation and spread of culture, as in the development of national identity.

See also: Attitudes, Political and Civic Culture; Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Mass Society: History of the Concept; Political Communication; Politics and Participation; Public Sphere and the Media; Public Sphere: Eighteenth-Century History.

Bibliography

Almond, G., Verba, S., 1966. The Civic Culture.
Amann, P., 1975. Revolution and Mass Democracy: The Paris Club Movement in 1848. Princeton University Press, Princeton, NJ.
Arendt, H., 1958. The Human Condition. University of Chicago Press, Chicago.
Bellah, R., 1985. Habits of the Heart: Individualism and Commitment in American Life. University of California Press, Berkeley, CA.
Boggs, C., 2000. The End of Politics: Corporate Power and the Decline of the Public Sphere. Guilford Press, New York.
Calhoun, C., 1988. Populist politics, communications media, and large scale social integration. Sociological Theory 6 (2), 219–241.
Calhoun, C., 1997. Nationalism and the public sphere. In: Weintraub, J., Kumar, K. (Eds.), Public and Private in Thought and Practice. University of Chicago Press, Chicago, pp. 75–102.
Calhoun, C., 2002. Constitutional patriotism and the public sphere: interests, identity, and solidarity in the integration of Europe. In: DeGreiff, P., Cronin, C. (Eds.), Transnational Politics. MIT Press, Cambridge, MA, pp. 275–312.
Cuddy-Keane, M., 2003. Virginia Woolf, the Intellectual, and the Public Sphere. Cambridge University Press, Cambridge.
Diawara, M. (Ed.), 1995. The Black Public Sphere. University of Chicago Press, Chicago.
Eley, G., 1992. Nations, publics and political cultures: placing Habermas in the nineteenth century. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA, pp. 289–339.
Elshtain, J.B., 1993. Private Man, Public Woman. rev. ed., Princeton University Press, Princeton, NJ.
Ezrahi, Y., 1990. The Descent of Icarus: Science and the Transformation of Contemporary Democracy. Harvard University Press, Cambridge, MA.
Fraser, N., 1992. Rethinking the public sphere: a contribution to the critique of actually existing democracy. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA, pp. 109–142.
Garnham, N., 1992. The media and the public sphere. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, MA.
Gilmartin, K., 1999. Print Politics: The Press and Radical Opposition in Early Nineteenth-Century England. Cambridge University Press, Cambridge, UK.
Ginsburg, B., 1986. The Captive Public: How Mass Opinion Promotes State Power.
Goode, L., 2005. Jürgen Habermas: Democracy and the Public Sphere. Pluto Press, London.

Habermas, J., 1962. The Structural Transformation of the Bourgeois Public Sphere: An Inquiry into a Category of Bourgeois Society (T. Burger, Trans.). MIT Press, Cambridge, MA.
Habermas, J., 1984[1988]. In: Theory of Communicative Action, vols. 1 and 2. Beacon Press, Boston.
Habermas, J., Crossley, N., Roberts, J.M., 2004. After Habermas: New Perspectives on the Public Sphere. Blackwell Pub, Oxford, UK.
Harvey, D., 1985. Consciousness and the Urban Experience. Johns Hopkins University Press, Baltimore, MD.
Held, D., 1995. Democracy and the Global Order. Polity, Cambridge, UK.
Hohendahl, P.U., 1982. The Institution of Criticism. Cornell University Press, Ithaca, NY.
Horkheimer, M., Adorno, T., 1944. Dialectic of Enlightenment. Continuum, New York.
Jones, G.S., 1983. Languages of Class: Studies in English Working Class History, 1832–1982. Cambridge University Press, Cambridge, UK.
Katznelson, I., Zolberg, A. (Eds.), 1987. Working-Class Formation: Nineteenth-Century Patterns in Western Europe and the United States. Princeton University Press, Princeton, NJ.
Keane, J., 1991. The Media and Democracy. Polity, Cambridge, UK.
Landes, J., 1988. Women and the Public Sphere in the Age of the French Revolution. Cornell University Press, Ithaca, NY.
Lara, M.P., 1999. Moral Textures: Feminist Narratives in the Public Sphere. University of California Press, Berkeley, CA.
Lowi, T., 1985. The Personal President. Cornell University Press, Ithaca, NY.
Meyrowitz, J., 1985. No Sense of Place. Oxford University Press, New York.
Mosse, G.L., 1975. The Nationalization of the Masses. Cornell University Press, Ithaca, NY.
Mosse, G.L., 1987. Masses and Man: Nationalist and Fascist Perceptions of Reality. Wayne State University Press, Detroit, MI.
Mumford, L., 1938. The Culture of Cities. Secker and Warburg, London.
Negt, O., Kluge, A., 1964. The Public Sphere and Experience. University of Minnesota Press, Minneapolis, MN, 1993.
Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.
Rajogopal, A., 2001. Politics after Television: Religious Nationalism and the Reshaping of the Indian Public. Cambridge University Press, Cambridge, UK.
Rankin, M., 1993. Some observations on a Chinese public sphere. Modern China 19 (2), 158–182.
Rasmussen, T., 2000. Social Theory and Communication Technology. Ashgate, London.
Ryan, M., 1990. Women in Public: Between Banners and Ballots, 1825–1880. Johns Hopkins University Press, Baltimore, MD.
Ryan, M., 1992. In: Calhoune (Ed.), Habermas and the Public Sphere. MIT Press, Cambridge, pp. 259–288.
Salvatore, A., 2007. The Public Sphere: Liberal Modernity, Catholicism, Islam. Palgrave Macmillan, New York.
Schudson, M., 1998. The Good Citizen. A History of American Civic Life. Free Press, New York.
Sennet, R., 1977. The Fall of Public Man. Knopf, New York.
Somers, M., 1993. Citizenship and the place of the public sphere: law, community, and political culture in the transition to democracy. American Sociological Review 58, 587–620.
Somers, M., 1995. What's political or cultural about political culture and the public sphere? Toward an historical sociology of concept formation. Sociological Theory 13, 113–144.
Thompson, J., 1995. Media and Modernity. Stanford University Press, Stanford, CA.
Tocqueville, A., 1840. Democracy in America. Scribners, New York, 1844.
Wakeman, F., 1993. The civil society and public sphere debate: Western reflections on Chinese political culture. Modern China 19 (2), 102–138.
Warner, M., 2001. Publics and Counterpublics. Zone Books, Cambridge, MA.
Yeo, E.J., 1998. Radical Femininity: Women's Self-Representation in the Public Sphere. Manchester University Press, Manchester, UK.
Zaret, D., 1999. Origins of Democratic Culture. Princeton University Press, Princeton, NJ.

Publishing as Medium

Beth Luey, Arizona State University, Fairhaven, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.J. Miller, volume 18, pp. 12599–12603, © 2001, Elsevier Ltd.

Abstract

Book publishing has been the subject of study by historians, economists, bibliographers, legal scholars, and sociologists. At the junction of culture and commerce, the industry is a focus for researchers interested in issues such as commodification, cultural identity and protectionism, mass versus elite culture, intellectual property, freedom of expression, and business practices. The technology of publishing, from scribal to print to digital cultures, has affected all of these issues. Since World War II, book publishing has become increasingly concentrated, with mergers occurring across economic sectors and national borders, raising questions about organization and cultural protectionism. The rise of digital media has intensified debate about intellectual property law, authorship, and the future of books and reading.

Publishing is the act of making a work of words and/or images available to the public. Book publishing is differentiated from newspaper, magazine, or journal publishing by the nature of the product (volumes in codex form, not published periodically) and its business practices (including royalties to authors rather than single payments, selling mainly through retailers rather than by subscription, and the absence of advertising within the publication). Book publishing encompasses a range of activities through which a manuscript is turned into a printed and bound book or an electronic publication that can be sold or otherwise distributed to an audience. These activities are carried out by organizations and individuals who coordinate creation, production, promotion, and distribution.

History

In the ancient world, texts were reproduced and circulated under the auspices of religious, political, and commercial bodies; in some places, individuals copied and traded books among themselves. During the Middle Ages, publishing in Europe was centered in the monasteries, where manuscripts were reproduced primarily for religious purposes. After the twelfth century, scribes associated with universities took over this task, serving the clergy as well as new readers among officials, the nobility, professionals, and merchants. Until the appearance of the printing press in Europe in 1450, the relevant public was extremely small, limited to those who could read the classic languages and afford hand-copied books. With the advent of print, however, book publishing and bookselling became organized into an industry, its members issuing books in vernacular languages as well as Greek and Latin to sell to a growing and increasingly diverse audience.

Before the seventeenth century in Europe, and the nineteenth century in North America, the book trade remained undifferentiated and unstandardized. A single person usually performed the functions of today's publisher, printer, and bookseller, printing and selling books at the request (and at the expense) of authors or their patrons. Most printers issued newspapers and pamphlets as well as books, and they derived a large portion of their income from job printing of bills, stationery, business forms, and other products. As the market for books grew, a more complex division of labor gradually emerged in some companies, with different segments of the industry responsible for publishing, printing, and selling. (Despite this general trend toward specialization, some large publishing houses continued to do their own printing and maintained retail outlets well into the twentieth century.) Publishers and booksellers also began to specialize in specific types of books, such as school books, religious texts, or medical works.

Within the modern book industry, the publisher is central to the transmission of books from author to reader. Publishing firms select, edit, design, and prepare manuscripts for production; acquire copyrights and sell subsidiary rights such as translation and film rights; create marketing campaigns; sell through wholesale and retail outlets as well as directly to the consumer; and assume the financial risks involved in bringing a book to the public. The publisher interacts and contracts with the author (or the author's representative, the agent). A separate wholesaling sector performs the functions of warehousing and order fulfillment for some publishers. A more differentiated retail sector has also developed, as books are sold through both bookstores and retailers whose primary business is not in books (online and brick-and-mortar); through educational, employment, and voluntary organizations; and through mail-order organizations such as book clubs.

Since the early twentieth century, the publishing industry has been characterized by specialization in response to the markets for particular kinds of books and the channels of distribution. The most visible form of publishing is in trade books, those sold through retailers to a general audience. However, both textbooks (K-12 and college) and titles for professionals (scientific, technical, and medical (STM) publishing as well as books for lawyers, accountants, and educators) are a significant and profitable part of the book market. In many developing countries textbooks make up the single largest segment. Other important segments include the publishers, nonprofit and for-profit, who issue scholarly and religious books. As these divisions between book sectors developed during the twentieth century, different organizational and sales practices appeared within each type of

publishing house. These distinctions have sometimes broken down when mergers have brought together dissimilar publishing houses under a single roof.

The advent of the World Wide Web has reinvigorated the tradition of self-publishing, in which no one performs the gatekeeping or selection process of the traditional publisher. Earlier self-publishers included both authors whose work did not interest mainstream publishers and who availed themselves of the services of vanity presses, and those evading censorship. The latter included most notably the authors of *samizdat* – Russian for 'self-published' – in the Soviet Union. Online publication, which requires no investment in paper or printing, has made it possible for authors to publish their work independently, or with the assistance of an online producer. The willingness of electronic retailers to distribute these books has made the process more attractive and practical.

Publishing as a Field of Study

The world of book publishing has been viewed in various and sometimes contradictory ways: as an elite environment of literary connoisseurs and as a producer of commodified ephemera; as an industrial backwater and as the site of technological transformation; as a promoter of the status quo and as the medium of protest. These contrasting images grow out of the wide range of activity encompassed by book publishing and by the range of disciplines that analyze the industry. Literary scholars focus on trade publishing; scholars interested in education analyze textbook houses; sociologists interested in the transmission of knowledge study university presses and STM publishers. Journalists and people who work in book publishing have joined scholars in studying the ways the industry interacts with the larger society.

Historical Perspectives

Numerous publishers have penned memoirs and histories of individual houses. These, along with the archives of publishing houses, provide primary sources for historians who have sought to provide more analytical accounts of the industry as a whole. 'The history of the book,' which many date to the publication of Febvre and Martin's (1958) *L'Apparition du Livre*, analyzes the production, distribution, and reception of books. Febvre, a prominent member of the Annales school, and his coauthor described early publishing and bookselling in Europe and attempted to account for the book's impact on the spread of various ideas and cultural movements. For some, such as Darnton (1982), the field of research that has evolved since Febvre and Martin's book is equivalent to the social and cultural history of print communication. In this view, the object is to understand the role books have played in society at various times and in various places, and to understand how books have interacted with other social developments and processes. Darnton's proposal is also methodological, suggesting that the book be studied in terms of a communication circuit that runs from author to publisher to printer to shipper to bookseller and finally to reader. The history of the book may also encompass the history of reading: the rise of literacy, the way readers interpret and use books, and the meaning of books in individual lives (for example, Rose, 2001).

Other historians have emphasized the history of printing and the subject of bibliography, that is, the study of books as physical objects. Bibliography is concerned with describing books and identifying their attributes, such as paper, type, and bindings, in order to distinguish different editions or printings. This kind of examination can provide information about particular publishers and manufacturers, about labor practices and productivity, and about authorial intention (see, for example, McKenzie, 1986). The underlying purpose of such studies is to make the book itself the central focus of attention rather than using the book as a means for discovering other social patterns.

Social Science Perspectives

Social science research on book publishing has been intermittent and sparse compared to social scientists' interest in other forms of mass media. Instead, some of the most detailed descriptions and critiques of publishing have been written by people directly associated with the publishing industry. Of course, books and book publishing, central to academic work in the humanities and social sciences, have not escaped academic scrutiny altogether. The state of the publishing world is taken to be a matter of public concern, and scholars and other public intellectuals frequently address a general audience when discussing the significance of developments in publishing for print culture. Throughout the twentieth century, such discussions have often reflected the assumptions of mass culture theory. Commentators question whether publishing practices that cater to a mass market leave room for high-quality books that sell in moderate numbers, and whether increasing commercialism and centralization will homogenize the available literary fare (see, for example, Dutscher, 1954). As this suggests, the curiosity of publishing tends to lie in its dual nature as a commercial enterprise on the one hand, and as a producer of literary objects associated with artistic, educational, and religious values on the other. Indeed, the perceived difficulty of reconciling these two dimensions of publishing – often described as the conflict between culture and commerce – is probably the central problem analysts of the book industry have confronted.

Beginning with work by Escarpit (1958), a relatively small group of sociologists has tried to dissect the workings of the publishing industry. These studies drew on organizational theory (sometimes called the production of culture perspective) which, especially during the 1970s and 1980s, was being applied to various media firms and industries. This perspective begins with the assumption that publishers are similar to other kinds of bureaucratic organizations, facing the same problems of how to reduce uncertainty and risk (Hirsch, 1972). Scholars in this tradition have provided specific accounts of how books are acquired and distributed, and how publishing strategies vary according to size and type of firm. Common to much of this work is a stress on the gatekeeping role that publishers play in determining what new knowledge, theories, and literary forms will enter the marketplace of ideas.

Congruent with trends in mass media and popular culture studies, more recent researchers have reacted against theories of

mass culture and hegemony, which posit audiences as easily manipulated and degraded by mass-market literature. Such scholarship grants the reader a more privileged place in analyses of publishing, both acknowledging the reader's active role in choosing and finding meaning in reading material, and studying how conceptions of the audience shape the work of book professionals. Moreover, this theoretical stance questions the very distinction between popular and quality books. Instead, some suggest that the fear of commercialism has been primarily a cover for the discomfort of cultural elites who see their authority over cultural matters challenged (Long, 1985–1986, Luey, 2010; Radway, 1997). Similarly, it has been argued that publishers play an important role in the production of literary value (Verdaasdonk and van Rees, 1991).

Themes in Publishing Research

Changes within the publishing industry during the last few decades of the twentieth century shaped many of the questions posed in publishing research. Renewed interest in publishing has been concurrent with its integration into multinational entertainment and information conglomerates. Researchers have been especially attentive to the consequences of concentration, the significance of publishing as a global phenomenon, and the convergence between books and other entertainment and information media. Electronic publishing – like the distribution of art and music online – has raised questions of the adequacy of copyright laws and the role of the publisher as gatekeeper.

The Dynamics of Concentration

The American situation best illustrates certain trends in the industry. During the post-World War II era, most major New York publishing houses were transformed from private or family-owned companies into public corporations. Various social changes contributed to this development. The postwar baby boom expanded the market for children's books and textbooks, and many publishers sought to exploit this market by acquiring other houses, creating new juvenile and textbook divisions, or expanding existing divisions. At the same time, the aging heads of publishing houses founded in the first quarter of the century wished to retire. Few had heirs interested in continuing their work. Issuing stock to the public raised capital for expansion and made it possible to transfer a firm's assets in an orderly fashion.

As the assets of publishing houses began to appear in the market in the 1960s, they became attractive not only to other book publishers but also to news and entertainment conglomerates and technology firms seeking synergy. Book publishers would not add much revenue, but entertainment conglomerates hoped that book content could be reproduced across a variety of media, and cross-promotional strategies could be employed. Technology firms believed that the use of television and computers in the classroom would make it profitable to acquire the expertise and content available in textbook publishing. Merger and acquisition activity continued throughout the century, including purchases by foreign-owned companies such as Bertelsmann, News Corporation, and Pearson. The result has been the consolidation of the nation's largest publishers into a handful of companies, which also include American-owned Viacom, AOL Time Warner, and Disney.

Concentration of distribution began somewhat later. Chain booksellers began to grow in the 1960s with the establishment of standardized outlets in shopping malls, but their decisive dominance of the bookstore market did not occur until the 1990s, with the appearance of the superstore. The rise of Amazon.com and its online competitors had drastically altered the retail market by the turn of the twenty-first century. Independent booksellers, already endangered by the chains, virtually disappeared, and the chains themselves were threatened. With the introduction of the Kindle, Nook, and other e-book readers, online booksellers have been able to use their vast market share to bargain with publishers more effectively than other retailers ever could. Amazon.com has also begun to publish original works itself, blurring the lines between publisher and bookseller in a new way.

The effects of concentration have been a constant source of debate. Along with longstanding concerns about declining quality and the homogenization of reading material, these developments raise questions about changes in the organizational culture of the book business. Attempts to rationalize the publishing process, invest heavily in promotion, and exploit new markets have altered the balance of power within the industry, diminishing the autonomy of editors while giving increased clout to accounting, marketing, and subsidiary rights divisions. At the same time, publishers have grown dependent on a few large retail customers who as a result play a more authoritative role in the industry. Complicating the issue of concentration is the increase in numbers of independent publishers and the trend toward self-publishing. Some argue that these trends create diversity at the margins of the industry, and a less-elitist conception of books and publishing at the center. For others, the concentration of control of the book business means that a small number of for-profit companies have undue power over what individuals can read.

Globalization and National Identities

The role that the circulation of books plays in maintaining or undermining specific identities has long been an important theme in publishing research. The concept of globalization is readily applied to the publishing industry: multinational companies with far-flung holdings are now among the major players. But the integration of international markets is the primary way in which publishing can be conceived as a global phenomenon. Since World War II, selling rights to publish or distribute books in foreign territories has become one of the most important sources of revenue for major publishers. The trade in books goes in many directions but disproportionately travels from Europe and especially the United States to the rest of the world.

This raises both theoretical and policy questions about national sovereignty and the maintenance of national identities. In Canada, for instance, the overwhelming presence of first British and then American books has hampered the growth of a national publishing industry. While government policy has helped to preserve Canadian-controlled firms,

foreign-controlled publishers in Canada remain more profitable. Also, many Canadian publishers can survive only because they have the exclusive right to sell American books, which dominate best-seller lists in the English-language market. Even in Québec, where French publishers play a strong role, translations of American books are plentiful. Many argue that this situation has helped to further Americanize Canadian culture, diffusing American tastes, styles, and values through the medium of popular books.

Economic agreements between nations to eliminate trade barriers have, in the past, often excluded cultural products, with books in particular singled out for exception. The 2005 UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expression reaffirmed the right of nations to control the entry of cultural products from other nations. There are significant obstacles to maintaining a viable publishing industry in less-developed countries, and as the Canadian situation demonstrates, even highly developed nations find it difficult to make the economics of publishing work without protective measures. Internet retailing and online content pose new challenges for cultural control.

Technological Challenges

Trade in books depends on the effectiveness of copyright law, and a significant body of research has examined how copyright and its breach have affected the publishing industry. However, the development of new technologies along with publishing's integration into entertainment conglomerates has raised new concerns about the control of literary property. Content is now easily transportable across media, and entertainment firms attempt to create synergies between book, film, television, music, magazine, and electronic divisions. This introduces a new level of instability into the publishing process. Traditional arrangements for the control of intellectual property are challenged when content is so mobile, with publishers and authors in conflict over who owns the new forms that emerge. The notion that 'information wants to be free,' a rallying cry of the Internet, has intensified pressure on copyright.

The death of the book has been predicted with increasing frequency in the twenty-first century. It is not clear exactly what is moribund. Some see the doom of the codex but assume that the content of books will survive in newer formats. Others fear that the very concept of the book – a reasonably long work of scholarship or the imagination – is threatened, either because creators will not be rewarded sufficiently to motivate them or because new media will so reduce the attention span of readers that there will be no audience. The future of book publishing is thus fraught with uncertainty in technology, law, economics, and culture.

See also: Audiences, Media; Literacy and Illiteracy, History of; Literacy and Orality; Mass Communication: Normative Frameworks; Mass Media, Political Economy of; Mass Media: Introduction and Schools of Thought; Textbooks.

Bibliography

Altbach, P.G., Hoshino, E.S. (Eds.), 1995. International Book Publishing: An Encyclopedia. Garland Publishing, New York.

Coser, L.A., Kadushin, C., Powell, W.W., 1982. Books: The Culture and Commerce of Publishing. Basic Books, New York.

Darnton, R., 1982. What is the history of books? Daedalus 111, 65–83.

Dutscher, A., 1954. The book business in America. Contemporary Issues 5, 38–58.

Escarpit, R., 1958. Sociologie de la Litterature. Presses Universitaires de France, Paris.

Escarpit, R., 1971. Sociology of Literature. Frank Cass, London.

Febvre, L.P.V., Martin, H.-J., 1958. L'Apparition du livre. Editions Albin Michel, Paris.

Febvre, L.P.V., Martin, H.-J., 1990. The Coming of the Book: The Impact of Printing 1450–1800. Verso, London.

Goldstein, P., 2003. Copyright's Highway: From Gutenberg to the Celestial Jukebox, revised ed. Stanford University Press, Palo Alto, CA.

Hirsch, P.M., 1972. Processing fads and fashions: an organization-set analysis of cultural industry systems. American Journal of Sociology 77, 639–659.

Horowitz, I.L., 1991. Communicating Ideas: The Politics of Scholarly Publishing, second ed. Transaction, New Brunswick, NJ.

Long, E., 1985–1986. The cultural meaning of concentration in publishing. Book Research Quarterly 1, 3–27.

Luey, B., 2010. Expanding the American Mind: Books and the Popularization of Knowledge. University of Massachusetts Press, Amherst.

McKenzie, D.F., 1986. Bibliography and the Sociology of Texts. British Library, London.

Radway, J.A., 1997. A Feeling for Books: The Book-of-the-month Club, Literary Taste, and Middle-class Desire. University of North Carolina Press, Chapel Hill.

Rose, J., 2001. The Intellectual Life of the British Working Class. Yale University Press, New Haven, CT.

Verdaasdonk, H., van Rees, C.J., 1991. The dynamics of choice behavior towards books: the role of cultural products in cultural behavior. Poetics 20, 421–437.

Publishing: Academic and University Presses

Peter J Dougherty, Princeton University Press, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the emergence of the modern university press nearly 150 years ago, these academic publishers have played an essential role in the world of scholarship, not only by making knowledge widely available through the books they publish but also by validating the quality of those books through the process of peer review and by actively engaging with authors to generate ideas for future works. The advent of three relatively new trends – corporate consolidation, globalization, and digitization – has now brought scholarly publishing to a crossroads, and how presses respond to these forces will determine their role in the years to come.

Introduction to University Presses

Academic publishing, and, more specifically, university press publishing, has been a central part of scholarly communications ever since the first books published by the universities of Oxford and Cambridge in the fifteenth and sixteenth centuries. Since the *modern* university press emerged nearly 150 years ago, the number and variety of university presses have proliferated and effectively served the main purpose of illuminating academic issues for researchers and students far and wide, while also playing the functional role of validating the scholarly quality of books through peer review, which is needed by academic authors trying for tenure and promotion. While the title of this article suggests a sprawling subject that reaches into many forms of scholarly communication produced by myriad kinds of publishers, nonprofit as well as commercial, this article will mainly tell the story of the university press itself, particularly of the American university press. Though this may seem like a parochial choice, it is the opposite for two reasons.

First, American university presses sit at the center of a vast international network of scholarly publishers (some university-based, most not), united mainly through translation activities. That is, global scholarly book publishing is heavily defined by the buying and selling of foreign-language rights between American university presses and their fellow scholarly publishers around the world.

Second, as the prevailing model of scholarly publishers, the American university press serves as a sociological and economic model for the debates surrounding scholarly publishers everywhere. As Mark Mazower recently wrote of David Runciman's 2013 history of democracy in crisis, *The Confidence Trap*, "What Runciman's focus on American democracy helps to do is to remind us that there is an international dimension to this subject that is closely connected to American self-perceptions" (Mazower, 2013). This applies to university press publishing as well.

From the early decades of the twentieth century, the period defining the scope of this article, through the 1980s, scholarly publishing at university presses was a relatively stable business. University presses and a handful of commercial scholarly presses produced and published mainly specialized academic books and journals funded primarily by sales to college and university libraries and secondarily by sales to students due to course adoptions. Many presses – notably at state universities such as the University of Texas or the University of California – also supplemented their publishing activity with books of regional or local interest. A certain subset of presses helped finance their operations through the publication of scholarly journals, though books have been the main product of university presses (and so sit at the center of this article).

The distinguishing, even defining, economic characteristic of the modern system of scholarly publishing is that university presses are built to be largely financially self-sustaining. In this regard, they are markedly different than most of their corresponding university-based units such as academic departments, museums, laboratories, or libraries. Aside from certain subsidies and endowment funds, presses have to recover roughly 80% of their costs in order to remain financially viable. Their growth – and their ability to maintain and even expand their annual title output – has been conditioned largely by the budgetary fortunes of higher education in general. As the funding trends have changed, so has the ability of university presses to sustain their title output.

This economic equation changed after the early 1980s with the advent of three powerfully disruptive trends: first, the corporate consolidation of the publishing business at large; second, globalization; and third, digitization and the revolution in information technology. The three-way pressure exerted by these forces on university presses has placed most presses, and indeed the entire enterprise of scholarly publishing, at a crossroads – some would say a crisis point – that threatens the very institution of the university press. How presses respond strategically to these very challenging forces, and how their universities support them in this effort, will determine their future trajectory.

Following some background on their history (where university presses came from) and function (what university presses do), this article will take up these forces and consider the cumulative effect they are having on the university press and its commercial corroborates.

A Thumbnail History of Modern University Presses

The historical impact of university presses can best be understood in three ways: first, through a chronological accounting

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41035-4

of presses themselves; second, through their publications; and third, through their people.

The Evolution of University Presses

The ancient history of the modern university press originates with Oxford and Cambridge vying for primacy in the fifteenth century, joined briefly by Harvard in the seventeenth. The story begins in 1478 when Oxford published a commentary on the *Apostle's Creed*, followed in 1521 by the establishment of a university printer in Cambridge. In 1640, on the other side of the Atlantic, Harvard University, then only 4 years old, issued the *Bay Psalm Book*.

But according to Peter Givler, former executive director of the Association of American University Presses, the modern character and personality of university presses only developed in the latter part of the nineteenth century with the establishment of The Johns Hopkins University Press, just 2 years after the university was founded in 1878. Unbeknownst to him at the time, Hopkins founder Daniel Coit Gilman gave university presses their motto when he wrote, "It is one of the noblest duties of the university to advance knowledge, and to diffuse it not only among those who can attend the daily lectures, but far and wide."

Gilman and the Hopkins leaders also gave university presses their structure, describing the press as a publishing house operating under the nonprofit charter of the university and charged with publishing the results of postdoctoral research.

The Johns Hopkins University Press was followed in 1891 by the establishment of the University of Chicago Press, 2 years after that university was founded. Two years after that, both Columbia University and the University of California began their presses.

The business of scholarly book publishing grew over the next several decades with the appearance of the University of Toronto Press in 1901, Princeton in 1905, Fordham in 1907, Yale in 1908, Washington in 1909, and Harvard in 1913 (following the closing of its original press in 1692). The second decade of the twentieth century closed out with the establishment of NYU in 1917, Stanford in 1918, and Illinois in 1919.

As university presses grew over the ensuing decades, they eventually came together in the establishment of the Association of American University Presses (AAUP) in 1946. The founding of the AAUP preceded the great boom in twentieth-century university press publishing, marked by the beginning of the race for space triggered by the launching of Sputnik in 1957. America's government, determined to meet the Soviet challenge, began investing massively in science and in research in other areas of higher education. Scholarly book budgets grew, and funds flowed to their publishers, especially university presses.

As much as the mid-century growth of university presses was marked, ironically enough, by outer space with the 1957 launching of Sputnik, it ended there too, with the Apollo 1969 moon walk of US astronaut Neil Armstrong. University press growth began to level off with the ebbing of the Cold War as government funding for publisher-produced research declined. This reduction in funding has continued gradually but steadily since then, beggaring the library market for scholarly monographs, the main stock in trade of university presses. The response on the part of presses, adapting to the realities of a withering market for their books, would stamp their history until today. (The foregoing discussion is based on Givler, 2002.)

As Cold War library funding began to dry up, university presses began to diversify. Fortunately for presses, the opportunity to do so coincided with consolidation in commercial publishing, the first factor to give the modern university press its shape. This consolidation would reconfigure the map of scholarly publishing largely in favor of university presses, and it has played into the hands of university presses at two levels: text and trade. Some university presses, in an effort to generate funds from greater sales, began to publish more trade books – that is, books intended for a broader cross-disciplinary, and even public, audience – and textbooks. Many university presses stepped into the advanced, first-year PhD level courses largely abandoned by commercial college publishers seeking greater fortunes by publishing for large introductory courses. This move to diversify the portfolio of university press publishing output would show up impressively in some of the outstanding titles produced during this era.

Important Contributions of University Presses

A brief (and informal) sampling, dating from the days around the establishment of the AAUP in 1946, suggests the extraordinarily, and too often forgotten, scholarly and literary legacy of university press publishing.

Some of the greatest university press books published graced the social sciences and include the following:

- Friedrich Hayek, *The Road to Serfdom* (Chicago)
- Milton Friedman, *Freedom and Capitalism* (Chicago)
- Daniel Patrick Moynihan and Nathan Glazer, *Beyond the Melting Pot* (MIT)
- David Reisman, *The Lonely Crowd* (Yale)
- Robert Shiller, *Irrational Exuberance* (Princeton)
- Robert Bellah et al., *Habits of the Heart* (California)
- Carol Gilligan, *In Another Voice* (Harvard)
- Thomas Piketty, *Capital in the Twenty-First Century* (Harvard)
- Albert O. Hirschman, *The Passions and the Interests* (Princeton)
- Alfred Chandler, *The Visible Hand* (Harvard)
- Richard White, *The Middle Ground* (Cambridge)
- Paul Samuelson, *Foundations of Modern Economic Analysis* (Harvard)

The humanities sit at the heart of the university press publishing portfolio, and it is there that university presses have made some of their greatest impact. Among the most outstanding humanities titles one thinks of the following:

- Mark Girouard, *Life in the English Country House* (Yale)
- E. H. Gombrich, *Art and Illusion* (Princeton)
- Jacques Derrida, *Of Grammatology* (Johns Hopkins)
- Northrop Frye, *Anatomy of Influence* (Princeton)
- Harold Bloom, *Anxiety of Influence* (Yale)
- Camille Paglia, *Sexual Personae* (Yale)
- Paul Fussell, *Great War and Modern Memory* (Oxford)
- T. Harry Williams, *Lincoln and His Generals* (LSU)
- Peter Brown, *Augustine of Hippo* (California)

- Mary Beard, *The Parthenon* (Harvard)
- Terry Eagleton, *Literary Theory* (Minnesota)
- John Rawls, *A Theory of Justice* (Harvard)
- Richard Rorty, *Philosophy and the Mirror of Nature* (Princeton)
- Charles Taylor, *A Secular Age* (Harvard)
- Harry Frankfurt, *On Bullshit* (Princeton)

Science is not thought of as a core university press publishing area, yet it has produced some of the most influential and enduring university press books:

- Albert Einstein, *The Meaning of Relativity* (Princeton)
- Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago)
- Richard Dawkins, *The Selfish Gene* (Oxford)
- Claude Shannon, *The Mathematical Theory of Communication* (Illinois)
- Linus Pauling, *The Nature of the Chemical Bond* (Cornell)
- Stephen Jay Gould, *The Structure of Evolutionary Science* (Harvard)
- E. O. Wilson, *Sociobiology* (Harvard)

University presses are hardly thought of as publishers of original literary and cultural works, but here again their output has been impressive. Highlights include the following:

- Eugene O'Neill, *Long Day's Journey into Night* (Yale)
- John Kennedy Toole, *Confederacy of Dunces* (LSU)
- Bohumil Hrabal, *Closely Watched Trains* (Northwestern)
- Ted Kooser, *Poetry Home Repair Manual* (Nebraska)
- John Neihardt, *Black Elk Speaks* (Nebraska)
- Norman McLean, *A River Runs Through It* (Chicago)
- Carlos Casteneda, *The Teachings of Don Juan* (California)
- *The I Ching* (Princeton)

Last but not least, in reference, university presses sport perhaps the most iconic book in all of publishing, *The Chicago Manual of Style*.

University presses tend to be talked about in terms of the quantity of the books they publish. The aforementioned list makes an impressive case for their quality.

What Do University Presses Do?

The story of scholarship in the twentieth century and beyond, especially in the humanities and the social and behavioral sciences, can be told largely through its books. University presses have figured powerfully in creating this story. From the early days of the university presses, they have brought the seminal work of authors such as Albert Einstein, Woodrow Wilson, Thomas Kuhn, David Reisman, William McNeil, Martha Nussbaum, Henry Louis Gates, Robert Shiller, and Thomas Piketty to market, and they continue to do so. And yet presses have done much of their work quietly and with little understanding on the part of observers of what actually happens within their walls. A brief accounting will help clarify their function within the culture of scholarly communications.

A few basics are in order: First, all university presses are nonprofit publishing entities either owned by or closely affiliated with their host universities. Second, university presses publish only peer-reviewed books, evaluated and accepted by panels of professors within their host universities, which are called editorial boards, faculty committees, or some variation thereof. Third, university presses are structured largely along the lines of their counterpart commercial publishers. That is, they are led by a director (historically an acquisitions editor promoted from within the ranks), and they employ acquisitions editors who acquire and build their lists of books, production teams (including production coordinators and designers) who create these books, marketing departments (including publicists, sales staffs, and direct mail and Web marketers) who sell the books, contracts departments, foreign rights departments, finance and administration departments, and in many cases entire warehousing and distribution departments that house and distribute the books.

The engine that drives the university press and that shapes its scholarly output is editorial acquisitions. Editorial acquisitions departments, typically headed by an editor-in-chief (or, at smaller presses, the director), are organized to generate a strong and steady supply of books in each area in which the press publishes, typically at most presses some combination of history, a cluster of humanities disciplines, and various social and behavioral sciences. Not surprisingly, university presses have traditionally organized their acquisitions efforts around those fields where books (as opposed to journal articles) form the foundation of academic knowledge while also serving as the basis of tenure and promotion decisions for their authors. Most great university press lists consist of a portfolio of books that include not only the works of mid-career and senior scholars, but also critically of junior scholars seeking to build their literary output and secure their career trajectories.

The Editorial Imagination

Acquisitions Editors

Within universities most discussion of university presses revolves around their instrumental functions: title output, revenues and costs, production capacity, and the practical value of the press in verifying the quality and kind of monographs as a basis for tenure decisions. The press is seen in this formulation as a relatively passive agent in the scholarly process. Yet the impulse that defines and drives the university press, and gives it life and personality, is what might be called the editorial imagination.

By applying their command of ideas and enthusiasm for cultivating ideas, acquisitions editors actively play a vitally important role in the propagation of scholarly knowledge. Acquisitions editors, often self-educated in the academic cultures of the fields in which they publish, reach out to scholars through university travel and conference attendance to create the author networks that attract the books to their lists. Further, these editors actively suggest and work with prospective authors to generate ideas for books. Examples abound, but a recent and significant example is the widely admired *Princeton Companion to Mathematics*, initially inspired by Princeton University Press editor David Ireland. Another is *The Classical Tradition*, initially suggested by Harry Haskell at Yale University Press. Series, such as The Princeton Economic History of the Western World, also came about because of the initiative of

acquisitions editors. This activity, centrally important to the development of outstanding publishing programs, is the engine that drives the university press.

The application of the editorial imagination contributes to the distinctive and diverse character of university press publishing. Each list is different and reflects the taste and intellectual curiosity of its editor. Some of the outstanding recent lists are Lindsay Waters' humanities list at Harvard University Press, Douglas Mitchell's sociology list at Chicago, Patricia Fidler's art and art history list at Yale, Roger Conover's cultural studies list at The MIT Press, and the economics list that several of us have built at Princeton. Other editors who have distinguished themselves in past years are the late Sheldon Myers, history editor at Oxford University Press in New York, Susan Abrams, who built the history of science list at Chicago, Marian Ash for her political science list at Yale, and Terry Moore, philosophy editor at Cambridge.

While acquisitions editors serve as the university press's main conduit to the outside world of authors, they also lead the system that connects these authors and their books to the internal staff of presses. These members then communicate with external corresponding communities vital to the propagation of scholarly ideas. Acquisitions editors 'publish their books within the press' to the production editors, designers, and copywriters who shape and structure books. They are also responsible for coordinating book releases with publicists and foreign rights directors, who communicate the virtues of books to the scholarly community. Publicists pitch to the vital networks of book reviewers, journalists, and periodical editors and producers responsible for reviewing and discussing books in their publications; rights directors pitch the books to scholarly and commercial foreign-language copublishers around the world for translation. This publishing circuitry, structured around the imagination and initiative of acquisitions editors, sustains the scholarly book publishing enterprise.

Marketing and Sales

Publishing the book within the press also requires acquisitions editors to coordinate with marketing and sales departments. This means equipping these colleagues with the necessary information for marketing to create the Web pages, direct mail pieces, and advertising necessary to engage prospective buyers, e.g., researchers, librarians, and faculty adopting books for courses. Meanwhile, supplying book information to sales directors and their staff makes it possible for these colleagues to actively pitch distributors and booksellers – local, chain, and online – on the press's forthcoming publications. Increasingly, university press sales departments have developed relationships with online aggregators such as Books at JSTOR, Project MUSE, and Walter de Gruyter, which supply research libraries with digitized collections of scholarly books from many publishers. This activity is finding new expression in e-book lending to public libraries through firms such as Overdrive and subscription services such as Oyster.

Further, most university presses maintain relationships with certain overseas sales representatives and agents (often serving small groups and press consortia, such as the University Press Group in Europe or Zhao Wei in China), who sell to local distributors and library wholesalers on information provided by university press sales directors and supported by reviews and other promotion generated by press publicists.

Completing the circle, actively engaged acquisitions editors cull feedback from these various sources to help inform themselves – and by extension, their authors – of the changing landscape of global scholarly book readership; what works, what does not work, and why.

Publishing Culture

Beyond the publishing process sparked by the engaged imagination of acquisitions editors, university presses operate at the intersection of several networks that help define their distinctive role in the scholarly and publishing cultures. Formal relationships include membership in the AAUP, the Association of American Publishers, and the International Publishers Association. These organizations convene to address issues bearing on the viability and health of their members and to address communally important issues such as copyright and piracy, while also providing educational services for member publishers.

Informal relationships include ongoing correspondence and meetings with foreign publishers with whom university presses meet regularly at the Frankfurt, London, and Beijing book fairs. Included in this category are organizational relationships American university presses have developed with mirror organizations in China, South America, Europe, and Japan. Finally, this constellation of corresponding organizations also includes related associations such as the Association of American Research Libraries, the Association of American Universities, and the American Library Association. It is within this institutional network that university presses advance the agenda for their ongoing contribution to knowledge.

Three Forces Reshaping the University Press

Neil Armstrong's walk on the moon may have marked the end of steady, predictable, and perhaps even complacent growth on the part of university presses. But three forces closer to Earth, all mentioned at the beginning of this article – corporate consolidation in publishing, globalization, and digitization – have shaped the recent history of university press publishing in equally profound ways, for better and for worse.

Corporate Consolidation in Publishing

What could the first of these forces, corporate consolidation in commercial book publishing, have done to the university press? Plenty. With the market for scholarly monographs slowly and steadily in decline by the early 1980s, university presses began to look around for ways of supplementing their revenues while remaining within the scholarly idiom. Changes in commercial publishing would provide the means. At the time, commercial publishing began a historic consolidation that would streamline what once had been a large and diverse business into a handful of megacorporations. The economies achieved in this transformation occasioned changes in two parts of commercial publishing in particular that would

provide opportunities for university presses: crossover nonfiction and college publishing, and to a lesser degree specialty publishing (including regional books).

Crossover nonfiction publishing is that part of the business most associated with houses such as Basic Books, The Free Press, Pantheon, Athenaeum, and a few others. From the World War II onward these houses thrived on the strength of publishing outstanding intellectuals for broadly cross-disciplinary and even general readerships (hence the term 'crossover'). This style of publishing is most prominently associated with Basic Books, which in the 1970s reached its apogee in the works of Robert Nozick, Michael Walzer, James Q. Wilson, and others. After the 1980s, these crossover houses and others came under pressure from the corporations that owned them to produce greater growth and higher profits through a steady supply of best sellers than typically associated with their traditional publishing. As a result, many authors, and even some editors, who would have published with or worked for one or more of the crossover houses migrated to the university presses. The presses that most benefited from this migration were the big Ivy League presses and the University of Chicago Press, each of which successfully built up its crossover publishing by signing major authors who might otherwise have continued to publish with commercial crossover presses. The most recent example is Thomas Piketty's extraordinarily influential 2014 book, *Capital in the Twenty-First Century*, published by Harvard. And while riskier and more expensive, this style of publishing provided the university presses capable of doing it with a new revenue stream.

A similar shift took place with college publishing. The concentration that occurred in college publishing was dramatic, resulting in the reduction in the number of college publishers from around 30 in the early 1970s to only a handful – five or six major higher education houses – by the late 1980s. Corporate pressure in college publishing came in the form of a powerful concentration on introductory and intermediate-level textbooks, central to big enrollments and potential big revenues. This concentration opened up the smaller upper-level undergraduate and graduate course markets for publishers at the business's margins. Certain university presses were only too happy to step in, first and most notably MIT Press, which pioneered the art of the excellent advanced textbook in technical areas such as economics and computer sciences. Other university presses followed suit, integrating the advanced textbook into the broadening portfolio of scholarly books.

For many state university presses, this exercise in portfolio diversification arose not from textbooks, but from another source: regional books. These books included kinds of titles once associated mainly with the ever-concentrating business of commercial publishing, such as books on cuisine, architecture, design, natural history, and sport. These genres, embraced by university presses from Rutgers in New Jersey to the University of Hawaii Press, have supplemented presses' more mainstream scholarly publishing while enriching the cultural lore of their home cities and states.

Globalization

The opportunity for diversified publishing occasioned by the commercial consolidation of publishing in the 1980s, and the concomitant search for new sources of growth on the part of university presses, helped to preserve some very important strains of publishing while strengthening university presses and animating their lists. Globalization, the second of the forces under discussion, would have similarly salutary results for those university presses prepared and poised to take advantage of it.

Declining library sales – historically the mother's milk of scholarly monograph publishing – combined with a drop-off in independent bookstore business to advance the steady desiccation of the domestic market for scholarly books in North America. The domestic bookstore market staged a rally in the 1990s with the rise of chains such as Borders and Barnes & Noble, but this trend was short-lived, owing to the increasing dominance of Amazon and online book shopping. The net effect was a gradual but dramatic reduction in the amount of shelf space available for scholarly books – and hence a further shrinkage contributing to the now generation-old decline of university presses. Presses would shore themselves up not only through diversification of their publishing portfolios but also by looking for new readers in new places: this meant overseas, an opportunity accompanied by the globalization of higher education.

Setting the stage for the globalization of the university press are several forces: the growth of the developing world, the spread of English as the lingua franca of scholarship, the rise of cross-national scholarship, the investment in higher education among governments in developing countries, and the growth of literacy rates around the world. Concomitant with the expansion of global markets has been the corresponding growth and integration of the media, giving publishers access to book reviews, editorial pages, and eventually blogs and social media as conduits of exposure for news about university press books.

Leading the way in accessing these global markets have been those presses with lists that 'travel,' meaning clusters of books in areas such as science, economics, philosophy, and world history that appeal as greatly across other continents as they do in North America. Among the most prominent global presses are Princeton, spurred by its economics list since the 1990s; Yale, strategically well placed to grow internationally through its London office, established in the 1970s; The MIT Press, well situated in globally relevant science fields; and smaller presses such as Georgetown and Fordham with their strong lists in languages and international culture. Sales of English-language editions of university presses for globally minded presses have been abetted by expanded translation business, especially in growing countries such as China, Korea, and Brazil.

Globalization continues to provide a source of residual growth for university presses, especially those deploying strategies around the development of editorial programs geared toward international readerships.

Digitization

The final force in reconfiguring the university press, digitization, has had both positive and destabilizing effects on presses. While university presses tend to be regarded by university administrators and librarians as technologically old-fashioned institutions, presses actually were early movers in

implementing digital technologies, blazing a trail in book production through the use of digital files in the copyediting and typesetting of their books. By the early 1980s, digital book production had become the norm across university press publishing, while the use of computer technology had begun to spread throughout the other functions across the business. As the quality and sophistication of digital technology grew, the digital production of university press books evolved.

The next turn in this evolution came with the rise of the Internet, which enabled presses to apply digital technology as a means of communication. The Internet made it possible for university presses to reach out to book review editors, editorial page editors, critics, journalists, and public intellectuals writing for publications all over the world. This global publicity landscape was later enriched by blogs, and eventually by social media. Presses that held world rights to their books could, and did, constructively pursue global publicity and review coverage for their books, occasionally resulting in worldwide sales success.

Global university press publishing took its next step on the sales front with the spread of online book merchant sites such as Amazon and The Book Depository as well as national sites such as Flipkart in India. Readers in countries serviced by these sites were able to discover and purchase university press books with greater ease than ever before. Meanwhile, the rise of library aggregators made possible the online purchase of university press books by libraries around the world, including university library consortia in developing countries, which were able to acquire complete collections of dozens of university press offerings at minimal prices and make them freely available to researchers in their networks.

While corporate consolidation and globalization have been unambiguously positive influences on the university press, digitization is a much more complicated story. On the one hand, it has enabled presses to be far more effective in reaching markets and making their books discoverable than ever in their history; on the other, it has created pressure from parent institutions to push presses into making their publications freely accessible on the Web without accounting for the operating costs that make the publication of books possible in the first place. This is the crossroads noted at the beginning of this article, and its resolution will define the fate of scholarly books in the next generation.

The underside of digitization for university presses and other book publishers is captured in Joseph Schumpeter's (1942) notion of creative destruction – that is, technological innovation itself contains the seeds of institutional subversion. This Schumpeterian subversion has come to scholarly book publishers in two forms: via outside attack by digital piracy and via institutional pressure through universities' calls for open access to books.

Online book piracy, though seemingly a recent plague for publishers, is as old as online publishing itself. Dating back to the early 2000s, online pirate sites such as Pirate Bay have captured book files from university presses and commercial publishers alike, making these files freely available to prospective readers all over the world. Individual publishers as well as national associations of publishers have endeavored to immobilize pirate sites by legal threats and even by legislation against Web providers that host pirates, but efforts to this day have been largely ineffectual. Estimates place the losses from piracy to publishers at around a tenth of net revenues.

While piracy amounts to a direct assault on university presses and their authors through copyright violation, digitization presents another threat to university presses and the cost-recovery economic model on which they operate by way of 'open access.' Open access is the name of a movement, championed by some university leaders, librarians, academics, and foundation executives, for presses to make their books freely available on the Web. Although there have been ideas proposed for university funding of open access university press books, nobody has yet devised a way to effectively underwrite university presses while offering their books free of charge.

An intermediate disruption caused by digitization appeared a few years ago in the form of a challenge to the fair use of published materials in the classroom. The crux of this challenge arose in a dispute between universities, which embraced a broad interpretation of fair use in order to make copious amounts of published materials available to instructors for use as classroom materials, and publishers, who hewed to a more limiting interpretation of fair use. This issue is still being adjudicated in the courts, just as the influence of digitization on university presses continues to evolve.

Challenges to the Future of the University Press

As noted at the outset of this article, university presses are poised at a crossroads. How they frame the value of what they do, how this value is perceived by university leaders, and how the presses enhance their value proposition in the future will determine their fate. There are basically two scenarios surrounding this value proposition – a low road and a high road. The low road charts a diminishing role for university presses. It defines presses in minimal terms, casting them as little more than third-party credentialing agents for tenure candidates at colleges and universities near and far. This function can and will – and, some would argue, should – be performed by educational institutions themselves, largely on the Internet. The university-based library publishing movement now taking shape is a first step in this direction.

The high road for university presses traces a different future. The high road casts university presses less as passive credentialing agents than as active publishers, purposefully engaged in shaping the scholarly conversation and in reshaping publishing itself. The high-road university press is an editorially driven publishing enterprise in which editors are actively working with authors to develop ideas for new books of various kinds – monographs, textbooks, trade books, and reference works – while the leadership of such a press seeks new ways to connect its books to the scholarly world at large and especially internationally, through traditional print as well as technologically advanced media.

The high-road university press eagerly engages other parties in the global republic of letters – book reviews, editorial pages, syndication services, bloggers, and public intellectuals – in framing and airing the scholarly conversation.

The high-road university press actively builds the narrative of scholarship, across disciplines and around the world, extending fully into the developing world. Ultimately, it will find a way not only to engage, but also to cultivate the next generation of scholars, both from the industrialized and the developing world.

See also: Awards, Public Recognition, and Evaluation; Computers: Impact on the Social Sciences; Cultural Mediators and Gatekeepers; Education, History of; Educational Media: Potentials for Learning; Encyclopedias, Handbooks, and Dictionaries; Globalization in Education; Information Society; Journals, Scholarly; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Libraries; Peer Review and Quality Control; Peer Review: Organized Skepticism; Publishing as Medium; Schumpeter, Joseph A. (1883–1950); Strategic Adaptation.

Bibliography

Givler, Peter, 2002. University press publishing in the United States. In: Abel, Richard E., Newman, Lyman W. (Eds.), Scholarly Publishing: Books, Journals, Publishers and Libraries in the Twentieth Century. Wiley, New York, pp. 107–120.

Mazower, Mark, 2013. The Future of Democracy. Financial Times. http://www.ft.com/intl/cms/s/2/016c6aea-3c08-11e3-b85f-00144feab7de.html (accessed 12.08.14.).

Schumpeter, Joseph, 1942. Capitalism, Socialism, and Democracy. Harper & Row, New York.

Relevant Websites

http://www.ala.org – American Library Association.
http://www.publishers.org – Association of American Publishers.
http://www.arl.org – Association of American Research Libraries.
http://www.aau.edu – Association of American Universities.
http://www.aaupnet.org – Association of American University Presses.
http://about.jstor.org/content-on-jstor-books – Books for Jstor.
http://www.internationalpublishers.org – International Publishers Association.
http://muse.jhu.edu – Project MUSE.
http://www.degruyter.com – Walter de Gruyter.

Punishment

Claire EM Hamilton, Queen's University Belfast, Belfast, Northern Ireland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Given the focus elsewhere in this collection on the comparative politics and functions/history of punishment, respectively, this article focuses on the meaning of the concept itself. The legal prerequisites for formal punishment will be discussed before proceeding to examine its main components, namely, censure and sanction. The final two sections critically discuss the traditional view that punishment comprises a sentence imposed by a criminal court following a conviction.

Legal Requirements and Justification

Despite its popular usage as a term for anything, which is painful or demanding, formal punishment in the sense that interests lawyers and criminologists denotes penalties authorized by the state in response to a criminal offense. This has been further elaborated by Hudson (2003: 1–2) as a set of six criteria, adapting work done by philosophers, Flew (1954) and Benn and Peters (1959):

- It must involve an evil or the infliction of some unpleasantness.
- It must be for an offense, actual, or supposed.
- It must be of an offender, actual, or supposed.
- It must be the work of personal agencies.
- It must be imposed by an authority conferred through or by the institutions against the rules of which the offense has been committed.
- The pain that is inflicted should be intentional, not accidental or coincidental.

Clearly, the authority averred to in the fifth criteria is a judicial authority in the form of a court and the institutional rules are the rules of the criminal law and penal system. Zedner (2004) additionally draws attention to the foundational legal principles that protect citizens from arbitrary prosecution and sentencing, namely, "there can be no crime without law" and 'no punishment without law.' While, as will be seen below, this narrow view of punishment as pain inflicted following conviction by a criminal court should not be accepted uncritically, the above remain the prerequisites of formal or legal punishment in many contemporary Western societies. The systematic study of punishment following crime, together with the study and evaluation of penal institutions, can also be said to form the basis for the field of 'penology,' though this is arguably a term less in common usage now than previously (Hudson, 2003; Sparks, 2013).

Broadly following Zedner's (2004) discussion of punishment in the context of criminal justice, discussion below will focus on 'censure' and 'sanction' as component parts of punishment, as well as revisiting traditional notions of punishment as pain delivered by the state at the end of criminal process. Given their treatment elsewhere in this collection (*see* Punishment: Comparative Politics of; Punishment: Social and Legal Aspects), the social roles of punishment and the various philosophies of punishment will not be comprehensively dealt with here. Suffice it to say at this juncture that the former has assumed a growing importance within penology in recent years, replacing the positivist 'punishment' with the more constructivist 'penalty' viz. the complex of ideas, institutions and relationships that deploy, construct and legitimize the power to punish (Garland and Young, 1983). The latter are generally divided into two groups, with the major point of distinction being their prospective or retrospective orientation. Theories such as retribution and denunciation differ from the rest of the rationales for punishment in that they are deontological and backward looking in their outlook (Von Hirsch, 1993). Punishment is gauged in relation to the harm caused by the offense and culpability of the offender. Instrumentalist or utilitarian theories such as deterrence, rehabilitation, and incapacitation, on the other hand, are forward looking, i.e., they hope to achieve something through punishment. They are also often referred to as consequentialist because they seek to justify punishment in relation to its anticipated future consequences (Braithwaite and Pettit, 1990) or reductivist because they aim to reduce crime (Cavadino and Dignan, 2012). In considering, and indeed seeking to reconcile, the various rationales, it is useful also to bear in mind another distinction, namely, that between the general justifications of punishments and the principles of distribution (Hart, 1968). Many people would accept that the overall aim of the system of punishment is preventative or deterrent, but this does not answer key questions about how much punishment there should be for certain types of offences/offenders and what types of punishment there should be – arguably, concepts such as desert or proportionality best supply the answer to these questions. The tensions between the various schools of thought may be easier to resolve, if this hybrid approach is adopted as a starting point.

In discussing the philosophical basis of punishment, consensus should not of course be assumed. The work of the abolitionist movement presents an important challenge to mainstream penology given their arguments that, far from reducing crime and protecting the public, the criminal justice system engenders further criminality, reproducing social divisions. While sometimes dismissed as 'sociological dinosaurs' (Sim, 1994), abolitionist criticisms have been highly influential

in moderate penal policies in Scandinavia and the Netherlands and indeed, can be viewed as the conceptual forerunners of what would become the restorative justice paradigm (Snacken et al., 2013).

Censure and Sanction

Moving to the essential elements or characteristics of punishment, these have been identified by Zedner (2004) as censure and sanction. Censure or condemnation of an individual's actions forms an essential element of punishment, giving further, more forceful expression to the censure already conveyed by a criminal conviction. This idea is closely allied to the notion of communicative or expressive punishment or indeed its 'symbolic significance' (Feinberg, 1994: 74). Thus, for example, a fine imposed as punishment carries a censure or condemnation that a civil penalty such as a tax does not, even though both are payable to the state. Much work has been done on censure by retributivists such as Andrew von Hirsch given their emphasis on punishment as the expression of blame or a moral judgment on an individual (Duff, 1986; Von Hirsch, 1993; Narayan, 1993). In *Censure and Sanctions*, von Hirsch argues that censure is addressed not only to the offender but also to the victim and others in society:

- Offender: it sends the message that society disapproves of what she/he has done and that a moral response is expected;
- Victim: it acknowledges the victim's hurt and loss and that this occurred at the hands of the perpetrator;
- Society: it provides reasons for desistance from wrongdoing, not because they fear the consequences but because it is morally wrong.

Assuming censure was sufficient to these tasks, questions may legitimately be raised about the justification for inflicting further pain in the form of a sanction on the offender. Punishment involves deliberately harming individuals, and therefore stands in very strong need of justification (Lacey, 1988). Von Hirsch (1993) and Narayan (1993) answer this question by arguing that the sanction acts as 'prudential disincentive' or reason for the offender to resist temptation and pursue the proper course of action. This supplements rather than replaces, however, appeals to the individual's moral sense of the wrongfulness of the act and it should be noted that the sanction must comport with the seriousness of the crime so as to reflect the appropriate degree of censure. At the other end of the spectrum, Braithwaite and Pettit's (1990) republican theory of criminal justice seeks to decouple censure from sentencing. On this view, censure may be achieved more effectively by reintegrative or nonstigmatizing shaming or other forms of social intervention. This is not predicated on abstract notions of individual responsibility, however, and the idea of proportionality is also rejected.

Turning now to the 'sanction' aspect of punishment, questions arise about the defining element of this pain-causing activity. Does the importance of the pain or unpleasantness imposed reside in the intention of the person imposing the sentence or the experience of the person receiving it? Some theorists prefer the former interpretation, with Hudson (2003) citing Walker (1991: 3) that "it is the belief or intention of the person who orders something to be done and not the belief or intention of the person to whom it is done that settles the question whether it is punishment." Walgrave (2004: 48–49) similarly distinguishes between, on the one hand, fines and incarceration as punishments or 'intended deprivations' and, on the other, probation or 'reparative measures' as nonpunishments or measures that are 'intended to be constructive.' Yet, to exclusively privilege the intentions of the sentencer over the sentenced in the punishment process also presents difficulties. The story reported in Cork, Ireland of two homeless brothers who threw a stone through the window of a bank because they could not cope with sleeping rough demonstrates well that for certain types of offender, punishment may be experienced as a benefit (Kilcommins et al., 2004). Kolber (2009: 214) also makes a compelling argument for recognizing subjective experience in calibrating punishments, arguing that "it is this very concern about inflicting distress that leads us to seek a justification for punishment in the first place."

The centrality of pain to the notion of sanction raises further questions about whether punishment loses its punitive quality where it imposed for the purpose of doing good. Arguments such as those referred to above by Walgrave have been challenged by writers such as Davis (1992), Crawford and Newburn (2003), and Levrant et al. (1999), in the context of a discussion of whether outcomes from restorative justice process are types of punishment. On this view, punishment includes anything which is a burden or imposition on the offender; the intentions of the sentencer are not relevant. While clearly claims about 'doing good' should be treated cautiously, research with sentencers often reveals rehabilitative intentions behind the ostensibly 'harsh' decision to sentence an individual to a term of imprisonment. McAra (2011: 100), for example, discusses research with Scottish sentencers suggesting that "prison (no matter its brutalizing reality) has been successfully marketed as a therapeutic institution capable of delivering reintegrative programmes." Imprisonment was thus in this context not always a punitive choice. The objective and subjective meaning of a related, but distinct, term, 'punitiveness' (see Pratt et al., 2005 among others on the 'new punitiveness') is an issue, which has recently received some attention from criminologists such as David Nelken (2005). He writes "are we mainly interested in the causes and conditions of variations in levels of punishment, regardless of the intentions of those involved? Or are we trying to explain why members of a given society think they are being punitive or non-punitive?" (2005: 221). Continuing in this vein, Nelken (2009) seeks to unpack assumptions commonly made about 'punitiveness' and 'leniency' in Italian-English youth justice. These countries are often put at opposite ends of the punitiveness spectrum in terms of their youth justice systems and are often characterized as 'lenient' and 'punitive,' respectively. Yet, researchers often fail to engage with what policymakers and other actors think they are doing and why it makes sense *to them*. Nelken notes that legal actors in the English and Welsh youth justice systems simply claim to be doing their best for young people before they get into further trouble and that "different working assumptions lie behind the levels of tolerance or leniency in dealing with young people in different places. In England and Wales it is thought that children will not easily grow out of delinquency without official intervention. In Italy, on the other hand, there is a legally enshrined presumption that in the majority of cases they will do

so" (Nelken, 2009: 307–308). His argument, therefore, is that in addressing questions of punishment, the researcher needs to keep the objective and subjective aspects distinct: it is important to differentiate between the deliberate infliction of pain/leniency and that which is motivated by other (worthy) concerns or the product of neglect. In an increasingly complex penal environment, 'punitive' outcomes can occur for 'good' reasons and vice versa (Daems, 2007).

A related point may also be made about community sanctions, which are often perceived by the public in relatively lenient, nonpunitive terms. The misperception (termed by Von Hirsch (1993) as the 'anything but imprisonment fallacy'), however, is revealed in research by Hudson (2006) and others (e.g., Dumortier, 2003) on the highly 'punitive' effects of interventions such as treatment programs and restorative justice. For some offenders, treatment programs in a noncustodial environment may be experienced in a very punitive manner given their demands on their privacy and the 'secrets of the self.' Recent work by Durnescu (2011) has very usefully attempted to identify more precisely the 'pains of probation' ranging from deprivation of autonomy to stigmatization effects through to the financial costs of attending meetings with their probation officer. The punitive ethos of many of these orders is also revealed in deliberate attempts by policymakers in some jurisdictions to introduce tougher 'intermediate' community sanctions with the explicit aim of enhancing the legitimacy of alternatives to imprisonment (Tonry, 1996). In the UK, this has led to the rebranding of community service orders, for example, as 'community punishment orders.'

Punishment as Sentencing?

The above analysis treats as unproblematic the fact that punishment follows conviction and sentence and that the 'object' of punishment must be convicted in a criminal court. Of course, in practice such neat divisions are rarely adhered to as suggested by research on the variety of ways in which police can informally punish suspects (for example, through denial of 'station' bail) (see, for example, McConville and Hodgson, 1993). While technically the purpose of the earlier stages is not to punish but to determine guilt or innocence, the reality for the offender is much different. As first recognized by Malcolm Feeley in his book *The Process is the Punishment* (1979), the distinction between the pretrial and sentencing stages (attendant on formal adjudication) is largely illusory for the offender. From his/her vantage point, punishment in the form of arrest, detention, prolonged pretrial custody, etc., begins prior to formal adjudication and this is becoming increasingly significant as the difference between 'accused' person and 'offender' is incrementally blurred (Brown, 2008).

The objection to such a position is that these informal, and sometimes illegitimate, practices by police may be analogous to punishment but to describe them as such would be to endow them with a legitimacy, which they do not merit. This is, however, to privilege a formal or 'legal' view of punishment over the offender's perspective; to prefer the 'law in the books' to 'the law in action.' A proponent of the broader view is Lynch (1995: 27) who argues that sanctions can be administered at many points along the criminal process:

> Police officers can punish by taking you into custody or giving you a warning or a citation. The prosecutor's decision to prosecute or not can be construed as a form of punishment. Judges can punish at sentencing. Prison and parole officials can punish through classification, decisions regarding good time or early release, decisions regarding institutional discipline, and revocation decisions.

The point is particularly important when it comes to cross-national comparison. The paucity of data on decisions made earlier in the criminal justice process often leads to a concentration or focus on the sentencing decision and beyond. This results in cross-national studies of punishment policy in fact becoming comparisons of prison use (Young and Brown, 1993). In her work on punitiveness in Ireland, Scotland, and New Zealand, Hamilton (2014) developed and applied a multidimensional test for punitiveness, which incorporated aspects of the criminal process beyond the narrow focus on the sentencing stage. The surprising homogeneity suggested by the results suggests the utility of a broader test in order to capture the variability and complexity of punishment in contemporary penal systems. Similarly, expansive approaches to the cross-national measurement of punitive practices have also been suggested by Pease (1994), Whitman (2003), and Kutateladze (2009).

More formal legal mechanisms also provide for the imposition of 'penalties' on individuals who have not been convicted of a criminal offense but who have been brought before the courts for civil wrongs. These hybrid civil orders generally place liberty restrictions on their subjects that may be as burdensome as punishment but are not subject to the same sorts of protections of criminal procedure, for example, the criminal standard of proof. A well-known British example is the anti-social behavior order ('ASBO'), which can be made when a court is satisfied that the individual has acted in a manner 'likely to cause harassment, alarm, or distress to another' and which attracts a penalty of up to 5 years imprisonment if breached. The issue of whether ASBO proceedings are criminal, 'in reality and in substance' has been considered and rejected by the UK House of Lordsin *R (on the application of McCann) v. Manchester Crown Court* ((2003) 1 AC 787). Perhaps most significantly, however, for the purposes of the present discussion, the Law Lords held that ASBOs, unlike ordinary criminal penalties, are designed to *prevent* antisocial behavior rather than to *punish* the offender. Ashworth (1998, 2004) has questioned the approach of the English courts to the issue and has drawn attention to the robust case law developed by the European Court of Human Rights in this area. (See *Engel v Netherlands* (1976) 1 EHRR 647; *Welch v. UK* (1995) 20 EHRR 247.) The Court has held on several occasions that, in order to prevent subversion of the procedural safeguards guaranteed in the fair trial provisions of Article 6 of the Convention, the term 'charged with a criminal offense' should be given an autonomous meaning, which looks to the substance rather than the form. What is known as the 'antisubversion doctrine' exists to prevent states simply reclassifying criminal proceedings as civil in order to avoid the protections attaching to defendants in criminal proceedings. As Ashworth (2004) argues, this approach is preferable in that it allows consideration to be given to the (potentially far-reaching) impact and

consequences of the order for the 'offender,' as well as the purpose of the order as imputed by the state. This is an area, which is likely to be the focus of much debate into the future, given the rise of coercive measures pursued through civil or administrative channels. As will be further discussed, these are particularly prevalent in the counterterrorism sphere where threatened states have introduced measures incorporating forms of highly restrictive house arrest (e.g., control orders in the UK and Australia and Security Certificates in Canada).

Beyond the Penal State?

Thus far punishment has been discussed in the context of pain or hard treatment imposed by authorized agents of the criminal justice system. Extending conceptions of punishment beyond the criminal justice system may appear to blur the boundaries between punishment and broader mechanisms of social control, sometimes treated as the independent variables in any analysis. Low levels of imprisonment in Japan, for example, are often explained with reference to the high levels of informal social control and the risk of ostracism for rule-breaking (Nelken, 2010). Convincing arguments may be made, however, for the inclusion of broader regulatory mechanisms in examinations of punishment practices and in some cases, it is arguable that a satisfactory understanding of the harshness of states or communities at given points in their history cannot be achieved without such an expanded definition. A few examples serve to illustrate that it may be difficult to justify *not* taking account of these factors. In the United States, there were 15 times fewer black people in prison prior to the abolition of slavery than afterward (Melossi, 2002 cited in Nelken, 2006), yet it is difficult to claim that American society was really more lenient then. Further, in an Irish context, O'Sullivan and O'Donnell (2007, 2012) have convincingly argued that the reversal of previous levels of noncriminal incarceration such as asylums and Magdalene homes casts contemporary increases in imprisonment rates in a very different light, particularly for women and children. Using the alternative analytic framework of 'coercive confinement,' they chart not only the numbers detained in prisons and borstals but also the thousands of men, women, and children who were incarcerated in psychiatric hospitals, mother and baby homes, Magdalen homes, reformatory and industrial schools during the first 50 years of Irish independence in the middle years of the twentieth century. In demonstrating that incarceration rates were eight times higher in 1951 than 2002, they both locate the prison as part of a wider network of institutions or 'carceral archipelago,' which was integral to the early Irish state and serve to complicate current dystopian analyses of contemporary punishment in Ireland and elsewhere.

Narrow conceptions of punishment as sanction imposed for breaches of the criminal law are also highly challenged in light of coercive practices deployed by states in the recent 'war on terror.' Counterterrorism measures adopted since the events of 9/11 by countries such as the United States and UK rather evoke a 'coercive spectrum' with criminal law at one pole and a series of extralegal measures such as extraordinary rendition, torture, and military detention at the other. In the middle are a range of legal measures such as control orders (in the UK and Australia), civil forfeiture laws and immigration detention, all of which operate on the margins, or outside of, the framework of the criminal law. In an important analysis of recent sociolegal changes, Lazarus et al. (2013: 463) argue that "various domestic and international pressures have led to a significant shift away from the punishment model in the exercise of state coercion." For them, the emerging use of these preventative techniques raises serious questions about the extent to which we are witnessing a "displacement of legal punishment as the primary mode of delivery of state coercion" (2013: 484). The focus on 'security' and 'prepunishment' rather signifies a temporal shift within the field of coercion frequently justified in the name of exceptionalism than by more traditional notions of retributivism, deterrence, or rehabilitation (see also Zedner, 2007).

While the trend toward extralegal measures is perhaps most obvious in the United States (Baker, 2013), it has been argued by some scholars that European policies such as interception of communications, data retention, data profiling, and the introduction of the European Arrest Warrant represent more punitive measures, as measured by interference with the human rights of those involved (Hosein, 2005). Similarly, Baker (2010) speaks about the extent to which the European Union has exchanged 'governing through crime' for 'governing through security,' given its degree of investment in initiatives to strengthen security. Certainly, there has been a definite shift within the European Union in recent years toward the criminalization of immigration (Melossi, 2013). At a rhetorical level at least, Dumortier et al. (2012) observe a remarkable blurring of the boundaries between the immigration and criminal justice systems, as evidenced in European Union initiatives such as the Framework Decisions of 13 June 2012 concerning terrorism.

Analytic attention may also be brought to bear on the question of whether the 'authorized agents of the state' mentioned above refer strictly to 'state' or 'public' actors or whether this may also include certain nonstate bodies to which the state has delegated its power for certain purposes. While punishment may be regarded as an inherently public function, justification of the privatization or marketization of prisons, which has occurred to a significant degree in countries such as the United States, UK, and Australia has rested on the distinction between the *allocation* and *delivery* of punishment (Logan, 1990). Sparks (1994) takes issue with the sharp distinction drawn between these two functions, arguing that this arrangement "tends to sandpaper away the contested and 'dismaying' nature of state punishment" (Garland, 1990: 1). In other words, the managerialist tendency to treat punishment as a mere 'technical' matter capable of delivery by any agency to which the state delegates this function serves to divorce the debate from punishment's core characteristics, namely, the deliberate imposition of pain requiring the highest possible justification. It is also, as Zedner (2004) remarks, a highly artificial position, one which ignores the role of prison authorities in determining (through prison disciplinary proceedings and other discretionary mechanisms) the amount of punishment, as well as its delivery. Whether the delivery of pain by private agents should be described as 'punishment' poses similar questions to those raised above concerning punitive practices engaged in by the police, namely, whether

this serves in some way to legitimate their actions. As also discussed above, however, and given the extent of private involvement in contemporary punishment, an authentic approach to this issue should embrace a more expansive definition of punishment in order to reflect the reality of pain delivery for the offender.

See also: Crime Policy in the United States; Death Penalty; Deterrence Theory: Crime; Deterrence; Deterrent Effect of Police and Prisons; Juvenile Justice: International Law and Children's Rights; Prisons and Imprisonment; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Sanctions in Political Science.

Bibliography

Ashworth, A., 1998. Neighbouring on the oppressive: the government's "anti-social behaviour order" proposals. Criminal Justice 16, 7–14.

Ashworth, A., 2004. Social control and "anti-social behaviour": the subversion of human rights? Law Quarterly Review 120, 263–291.

Baker, E., 2010. Governing through crime? The case of the European Union. European Journal of Criminology 7 (3), 187–213.

Baker, E., 2013. The emerging role of the EU as a penal actor in European penology? In: Daems, T., et al. (Eds.), European Penology? Hart Publishing, Oxford.

Benn, S.I., Peters, R.S., 1959. Social Principles and the Democratic State. Allen and Unwin, London.

Braithwaite, J., Pettit, R., 1990. Not Just Deserts: A Republican Theory of Criminal Justice. Clarendon Press, Oxford.

Brown, M., 2008. Risk, punishment and liberty. In: Cuneen, C., Anthony, T. (Eds.), The Critical Criminology Companion. Hawkins Press, Sydney.

Cavadino, M., Dignan, J., 2012. The Penal System: An Introduction. Sage, London.

Crawford, A., Newburn, T., 2003. Youth Offending and Restorative Justice. Willan, Cullompton.

Daems, T., 2007. Deconstructing punitiveness: review: Natasha A. Frost (2006): the punitive state: crime, punishment and imprisonment across the United States. New York, LFB Scholarly Publishing. Crime, Law and Social Change 47 (2), 129–133.

Davis, G., 1992. Making Amends: Mediation and Reparation in Criminal Justice. Routledge, London.

Duff, R.A., 1986. Punishment, Communication and Community. Oxford University Press, Oxford.

Dumortier, E., 2003. Legal rules and safeguards within Belgian mediation practices for juveniles. In: Weitekamp, E., Kerner, H.J. (Eds.), Restorative Justice in Context: International Practice and Directions. Willan, Cullompton.

Dumortier, E., Gutwirth, S., Snacken, S., De Hert, P., 2012. The rise of the penal state: what can human rights do about it? In: Snacken, S., Dumortier, E. (Eds.), Resisting Punitiveness in Europe. Routledge, Abingdon, Oxon.

Durnescu, I., 2011. Pains of probation: effective practice and human rights. International Journal of Offender Therapy and Comparative Criminal Justice 55 (4), 530–545.

Feeley, M., 1979. The Process Is the Punishment: Handling Cases in a Lower Criminal Court. Russell Sage Foundation, New York.

Feinberg, J., 1994. The expressive function of punishment. In: Duff, A., Garland, D. (Eds.), A Reader on Punishment. Oxford UniversityPress, Oxford.

Flew, A., 1954. The justification of punishment. Philosophy 29, 291–307.

Garland, D., 1990. Punishment and Modern Society. Oxford University Press, Oxford.

Garland, D., Young, P. (Eds.), 1983. The Power to Punish. Heinmann, London.

Hamilton, C., 2014. Reconceptualising Penality: A Comparative Perspective on Punitiveness in Ireland, Scotland and New Zealand. Ashgate, Farnham, UK.

Hart, H.L.A., 1968. Punishment and Responsibility: Essays in the Philosophy of Law. Clarendon Press, Oxford.

Hosein, G., 2005. Threatening the Open Society: Comparing Anti-terror Policies in the US and Europe. Privacy International, London.

Hudson, B., 2006. Secrets of the self: punishment and the right to privacy. In: Claes, E., Duff, A., Gurwirth, S. (Eds.), Privacy and the Criminal Law. Intersentia, Oxford and Antwerp.

Hudson, B., 2003. Understanding Justice: An Introduction to Ideas, Perspectives and Controversies in Modern Penal Theory, second ed. Open University Press, Buckingham.

Kilcommins, S., O'Donnell, I., O'Sullivan, E., Vaughan, B., 2004. Crime, Punishment and the Search for Order in Ireland. Institute of Public Administration, Dublin.

Kolber, A.,J., 2009. The subjective experience of punishment. Columbia Law Review 109, 182–236.

Kutateladze, B., 2009. Is America Really So Punitive?: Exploring a Continuum of US State Criminal Justice Policies. LFB Scholarly Publishing, New York.

Lacey, N., 1988. State Punishment: Political Principles and Community Values. Routledge, London.

Lazarus, L., Goold, B., Goss, C., 2013. Control without punishment: understanding coercion. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London.

Levrant, S., Cullen, F., Fulton, B., Wozniak, J., 1999. Reconsidering restorative justice? The corruption of benevolence revisited? Crime and Delinquency 45 (1), 3–27.

Logan, C., 1990. Private Prisons: Cons and Pros. Oxford University Press, Oxford.

Lynch, J., 1995. Crime in international perspective. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime. Institute for Comparative Studies Press, San Francisco.

McAra, L., 2011. Review symposium on Nicola Lacey, the prisoners' dilemma: political economy and punishment in contemporary democracies: global politics and local culture a response to Nicola Lacey. Punishment & Society 13 (1), 96–104.

McConville, M., Hodgson, J., 1993. Custodial Legal Advice and the Right to Silence. HMSO, London.

Melossi, D., 2002. Stato, ControlloSociale, Devianza. Bruno Mondadori, Milan.

Melossi, D., 2013. Punishment and migration between Europe and the USA: a transnational "less eligibility. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London.

Narayan, U., 1993. Appropriate responses and preventative benefits: justifying censure and hard treatment in legal punishment. Oxford Journal of Legal Studies 13, 166–182.

Nelken, D., 2005. When is a society non-punitive?: the Italian case. In: Pratt, J., et al. (Eds.), The New Punitiveness: Trends, Theories, Perspectives. Devon: Willan, Cullompton.

Nelken, D., 2006. Patterns of punitiveness. Modern Law Review 69 (2), 262–277.

Nelken, D., 2009. Comparative criminal justice: beyond ethnocentrism and relativism. European Journal of Criminology 6 (4), 291–311.

Nelken, D., 2010. Comparative Criminal Justice. Sage, London.

O'Sullivan, E., O'Donnell, I., 2007. Coercive confinement in the Republic of Ireland: the waning of a culture of control. Punishment & Society 9 (1), 27–48.

O'Sullivan, E., O'Donnell, I., 2012. Coercive Confinement in Ireland: Patients, Prisoners and Penitents. Manchester University Press, Manchester.

Pease, K., 1994. Cross-national imprisonment rates: limitations of method and possible conclusions. British Journal of Criminology 34 (5), 116–130.

Pratt, J., Brown, D., Brown, M., Hallsworth, S., Morrison, M., 2005. In: The New Punitivneess: Trends, Theories and Perspectives. Willan, Culllompton.

Sim, J., 1994. The abolitionist approach: a british perspective. In: Duff, A., Marshall, S., Dobash, R.,E., Dobash, R.P. (Eds.), Penal Theory and Practice: Tradition and Innovation in Criminal Justice. Manchester University Press/Fulbright Commission, Manchester.

Snacken, S., Bauwens, A., van ZylSmit, D., Tournel, H., Machiels, R., 2013. Prisons and punishment in Europe. In: Body-Gendrot, S., Hough, M., Kerezsi, K., Levy, R., Snacken, S. (Eds.), The Routledge Handbook of European Criminology. Routledge, Abingdon, Oxon.

Sparks, R., 1994. Can prisons be legitimate? penal politics, privatization and the timelessness of an old idea. British Journal of Criminology 34 (5), 14–28.

Sparks, R., 2013. Penology. In: McLaughlin, E., Muncie, J. (Eds.), Sage Dictionary of Criminology, third ed. Sage, London.

Tonry, M., 1996. Sentencing Matters. Oxford University Press, Oxford.

Von Hirsch, A., 1993. Censure and Sanction. Oxford University Press, Oxford.

Walgrave, L., 2004. Has restorative justice appropriately responded to retribution theory and impulses? In: Zehr, H., Toews, B. (Eds.), Critical Issues in Restorative Justice. Criminal Justice Press, New York.

Walker, N., 1991. Why Punish? Oxford University Press, Oxford.

Whitman, J., 2003. Harsh Justice: Criminal Punishment and the Widening Divide between America and Europe. Oxford University Press, New York.

Young, W., Brown, M., 1993. Cross-national comparisons of imprisonment. In: Tonry, M. (Ed.), Crime and Justice: A Review of Research, vol. 17. University of Chicago Press, Chicago.

Zedner, L., 2004. Criminal Justice. Oxford University Press, Oxford.

Zedner, L., 2007. Pre-crime and post-criminology? Theoretical Criminology 11 (2), 271–275.

Punishment: Comparative Politics of

Robert C Fried, University of California, Los Angeles, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The most striking development in twentieth-century penal politics was the emergence in the advanced democracies of the state that punishes as few and as little as it can. Convergence on policies of minimal punishment, as measured by the ratio of prisoners to robberies, occurred in spite of high unemployment, growing ethnic diversity and conflict, a drugs pandemic, and an average of sevenfold increase in serious crime. Convergence on leniency was paradoxical – given democratic competition, constant crime coverage in the media, and the persisting, even growing, crime-induced popularity of punishment among voters. The paradox is perhaps explained by such factors as Elias' civilizing process, penal sentimentalism, secularization, the welfare state, professionalism, and persisting elite control.

The Minimally Punishing Democratic State: Can a Democracy Punish?

A not-much-noted yet striking development in twentieth-century penal politics was the emergence in the advanced democracies of the minimally punishing state. Convergence on policies of minimal punishment, as measured by the ratio of prisoners to robberies, occurred in spite of high unemployment, growing racial and ethnic diversity and conflict, a drugs pandemic, and an average sevenfold increase in serious crime. Convergence on leniency was paradoxical – given democratic competition and the persisting, even growing, crime-induced popularity of punishment among voters. And indeed, the minimally punishing democracy emerged during one of the largest crime waves in recorded history – a Great Crime Wave impressively absorbed and ignored by all democratic political systems for one or more decades (Figure 1).

Modern democracies were strong enough and morally neutralized enough to dismiss the demand for a state response proportionate to the crime wave – a wave reflected in the annually mounting numbers of thousands, even millions of crime reports. Most democratic governments displayed surprising capacities not to punish in spite of the victim's entitlement to punishment implied by their penal codes. (No penal code contains a warning label to citizens that the punishments it threatens will be applied at most to 3% of those arrested and 1% of those who violate its terms.) Attrition rates for criminal cases became extremely high in all democracies, from the apparently most punitive to the least. (It would be comforting if this severe attrition reduced punishment to only and all the worst offenders. But the criminal law is easily manipulated and its administration heavily affected by fortune and Fortuna.)

A conclusion seems to have been reached in most democracies that the state can get away with never needing to punish (or at least imprison) much more than 100 per 100 000 people at any time, no matter which or how many or against whom crimes are committed (Figures 2–6).

These historical graphs incidentally show no impact of the gradual democratization that took place in those countries between 1850 and 1950.

The Convergence on Minimal Punishment

Nonenforcement has been the usual case in all democracies, of whatever repute for punitiveness. The facts of convergence on minimal punishment levels are clear. In 1940, most of the currently advanced democracies had capital punishment; by 2000, only two of them still inflicted the death penalty and only on a much reduced scale. In 1960, in the advanced democracies on average 11 prisoners (for all offenses) per reported robbery; by 2010, the average democracy had one prisoner per robbery. Between 1960 and 2010, the democratic state largely relinquished (despite public opinion) the threat of death punishment and reduced to lottery-odds the likelihood of prison punishment (Figure 7).

What Caused the Historical Rarefaction of Punishment?

Theories of overpunishment abound, given the hostility of most academic theorists to putting lower class criminals in jail or prison, not to mention executing them for murder. Harder to find are theories of underpunishment, unless one deals with crimes of the rich and powerful, where "the rich get away with murder" is axiomatic. (In spite of the axiom, the rich and powerful also remain the succulent targets of the media and ambitious prosecutors.)

The State Monopoly of More or Less Legitimate Punishment

One of the integral elements in the formation of states was the state monopolization of legitimate drastic punishment. For centuries, feudal lords, towns, and counties proudly displayed gallows as the mark of their sovereignty. Once local control of gallows was surrendered, monopoly allowed the state to reduce and then eliminate capital punishment.

State Elites' Conversion to Leniency

The decision to ration punishment came to rest on what had become, by the 1960s, a gap between the lenient mentalities of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45036-1

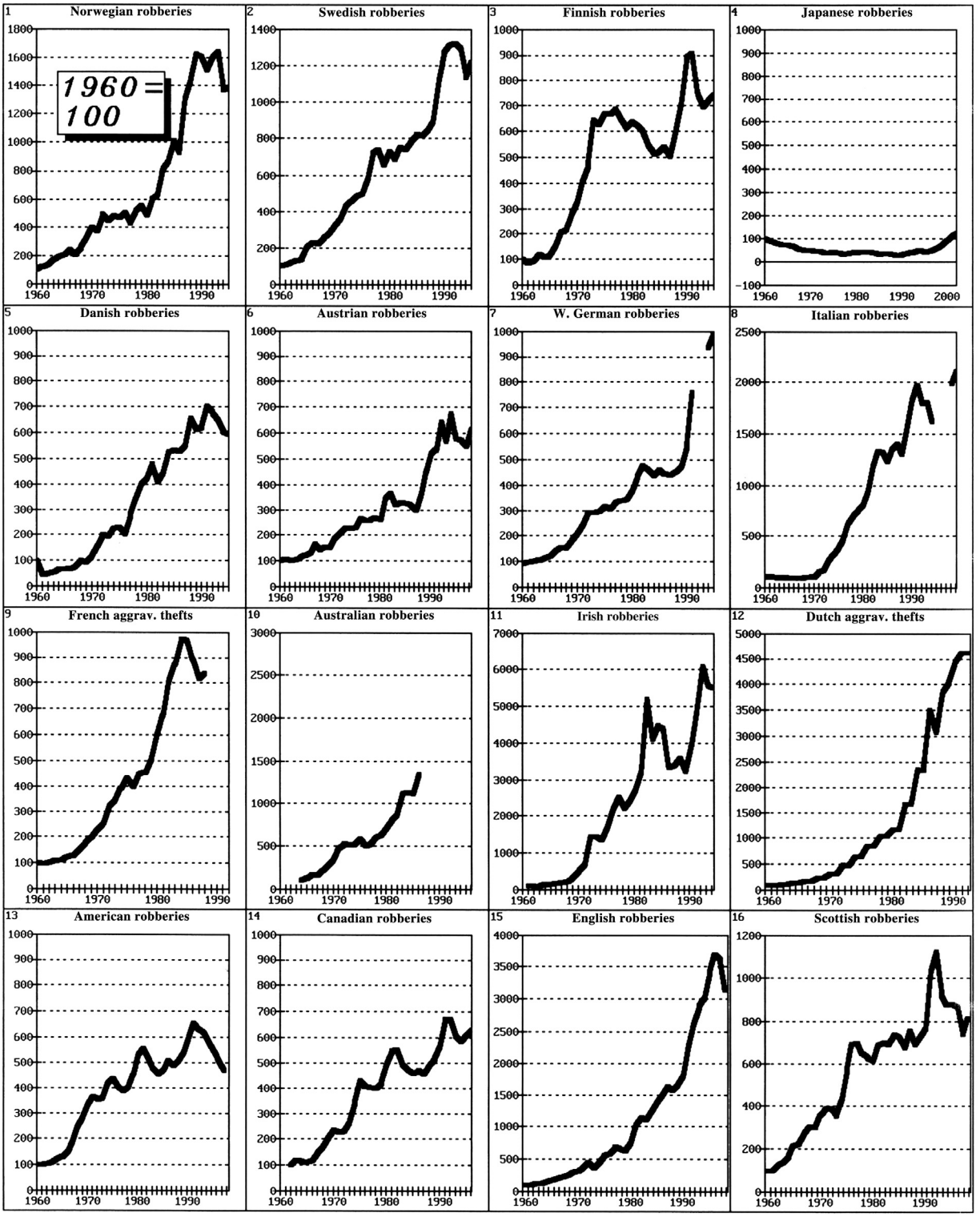

Figure 1 Rise in robberies.

political elites and the more traditionally punitive, if sometimes also ambivalent, mentalities of the mass public. This gap was best explained by Norbert Elias, who traced what he called "the civilizing process" – the diffusion of antiviolence mentalities among and by the middle and upper classes since the Middle Ages (Elias, 2000). Elites in the Western democracies became remarkably comfortable with lower class violence especially since the targets of that crime were usually other lower class

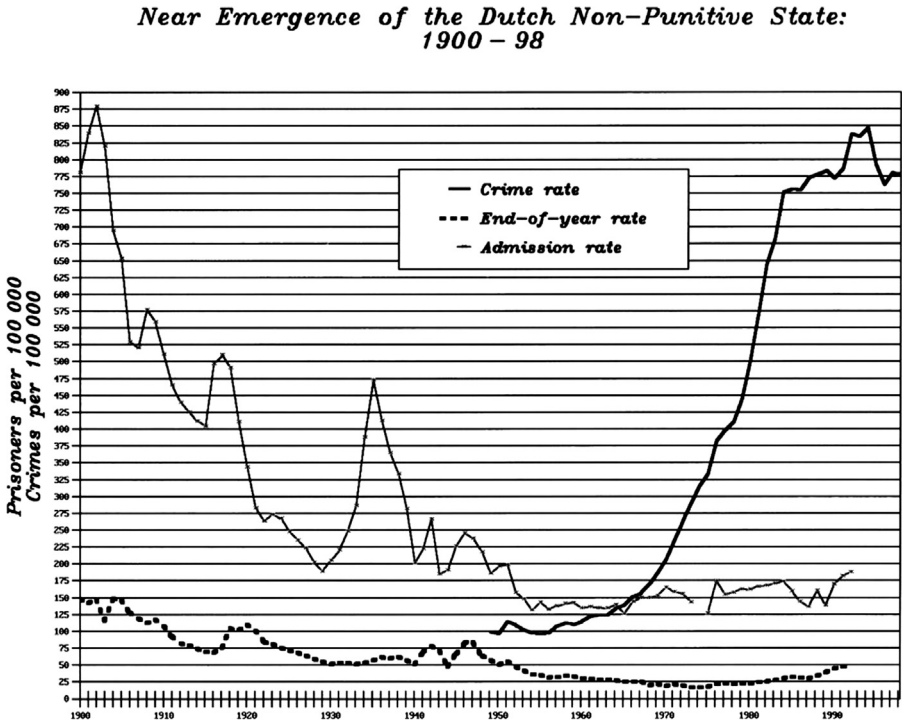

Figure 2 Decline of punishment in the Netherlands.

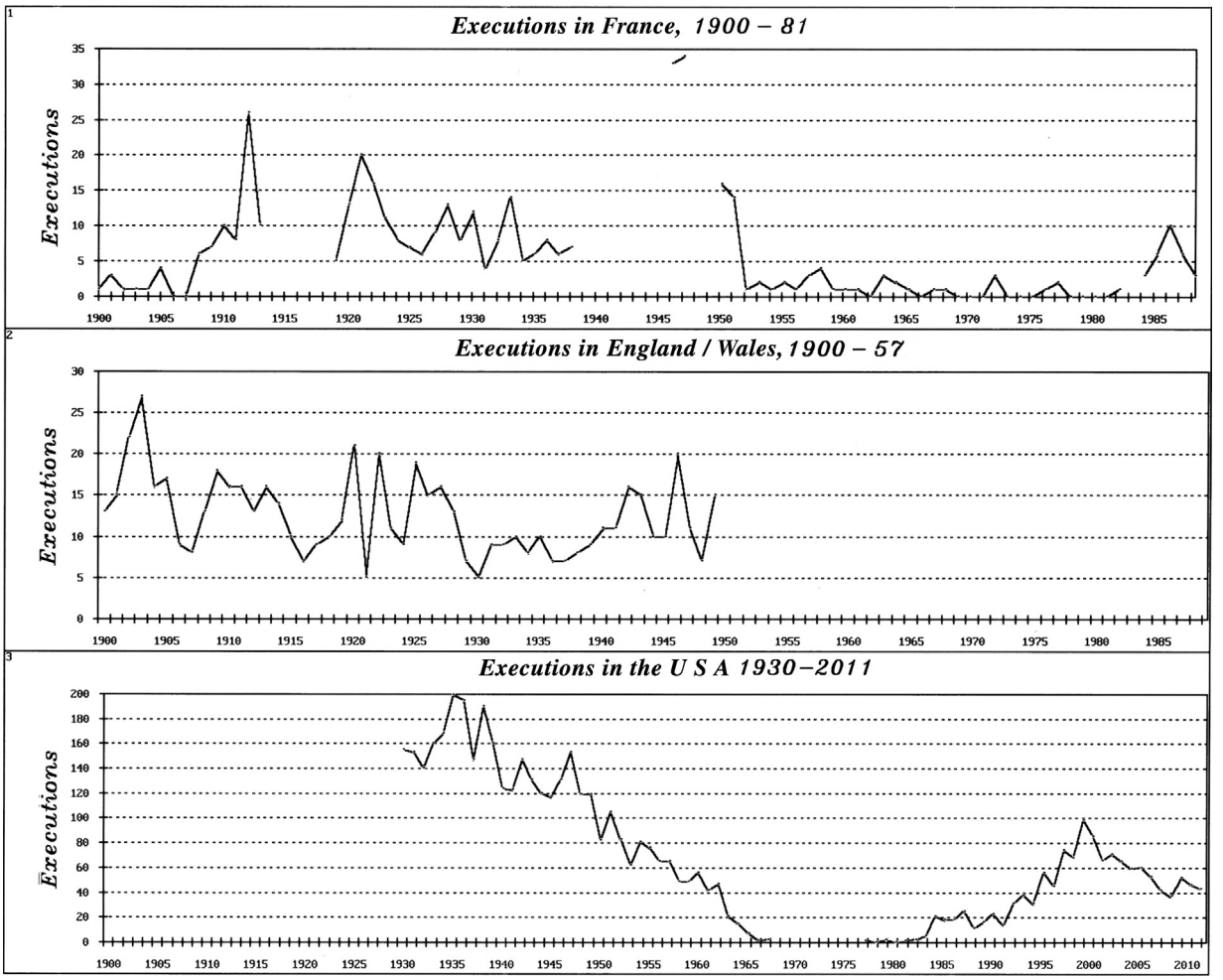

Figure 3 Decline of executions in France, England, and the United States.

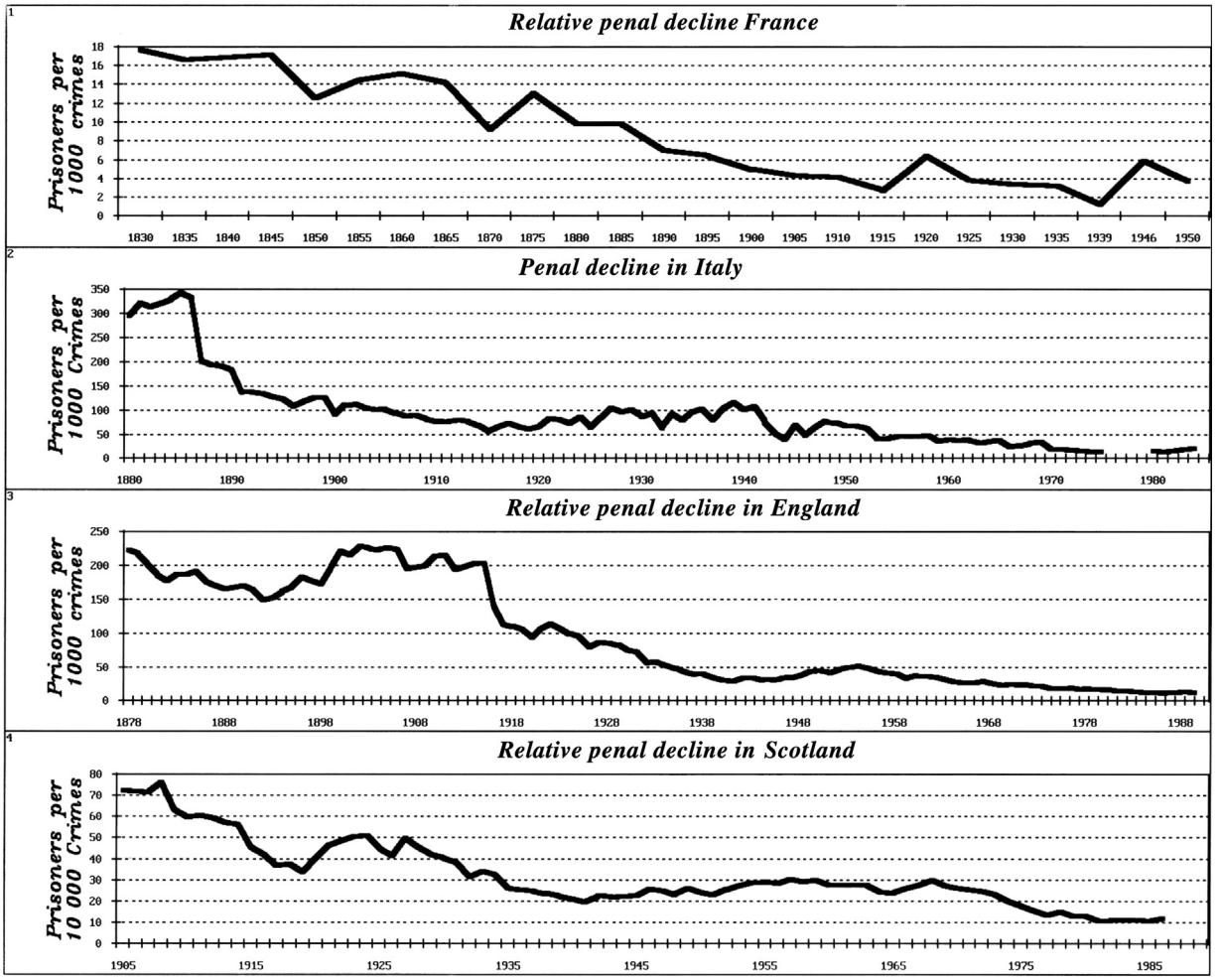

Figure 4 Decline of the prison/crime ratio in France, Italy, England, and Scotland.

victims. The police were invented in part to keep lower class crime targeted on other lower class people.

Penal Sentimentalism Rationalizes Elite Indifference to Lower Class Crime

The medieval elite – mass gap in penal sentimentalism widened with the further top-down spread of the ideas of the eighteenth-century Enlightenment and of the nineteenth-century Romantic, liberal, and socialist movements – humanitarianism, human perfectibility, and society's responsibility for crime. There arose an implicit understanding that civilized people do not believe in punishment. Criminals indirectly acquired great political influence: for many in the elite, they became objects to protect, not punish.

There was also a darker side to elite tolerance. Many people, especially intellectuals, were excited by the idea of violence, especially political violence or class warfare violence. The powerful appeals of Marxism (and Nazism) for Western intellectuals often rested on the legitimacy of violence for noble purposes. Sympathy for criminals as victims of capitalism rested (and rests) on elite contempt for capitalism.

Secularization and the Christianization of Penal (and Social) Policy

Surprisingly, Michelangelo's *Last Judgment*, in the Sistine Chapel, has not been painted over – not this time, of course, to cover nudity but to combat any possible notion of Christian, divine, or Church endorsement of hellish punishment. The *Last Judgment* survives even though it is judgmental. Since the eighteenth century, and in spite of the twentieth century, pessimism about human nature and the concept of original sin (although not in its Freudian reincarnation as universal, genetically innate aggression) have become outdated, reactionary, intolerable ideas.

Late in the nineteenth century, a new democracy emerged, the Nietzschean democracy where God is dead, Hell is only a bad dream, and paradise on earth is achievable. Crime is no longer immoral, but merely deviant behavior.

The private sector has taken up some of the slack created by the defaulting state: private policing and private punishment thrive. Much crime consists of acts of private, drive-by, Union Corse, and Godfather justice – never displaced by public justice.

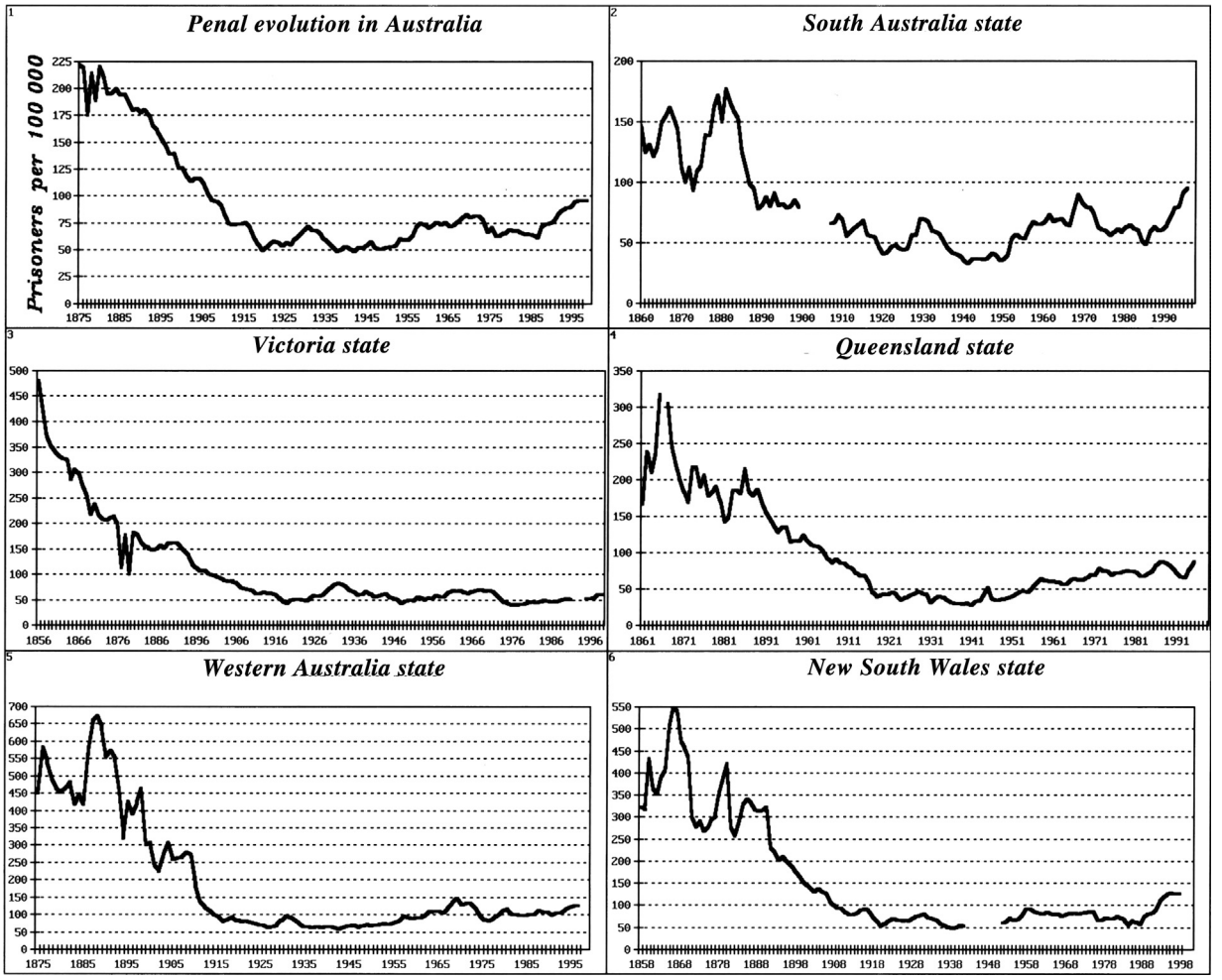

Figure 5 Decline of punishment in the Australian states.

Juvenilization of Justice: Age Discrimination in Penal Codes

One of the great challenges any society faces is to domesticate ever oncoming cohorts of young males. Of all the groups in democratic societies, the most challenging to peace and order, life and limb, are not the proletariat or the poor, but emerging cohorts of young males. These are in any society the readiest by hormonal inheritance to seize opportunities to commit violence, riot, and massacre (Banfield, 1974). They are also the readiest to expose themselves to violence to defend society against the young men of other societies – the readiest to be heroes in war, as well as the readiest to be villains in peace.

In spite of biology, penal sentimentalism has found ready application to young males ('adolescents'), the young offenders (usually male) who have not been convinced by the age of 5 or 10 or 15 or 20 that hurting other people is wrong and not a legitimate source of pleasure. To be sure, growing up in rougher communities, it is safer to be known as violent.

The Minoritization of Crime

Given millennia of racism as a set of normal attitudes, it is surprising that in the twentieth century, the domination of much crime in the Western democracies by racial and alien minorities has had relatively little impact on punishment rates. Many places have kept low punishment rates by simply replacing domestic offenders with minority offenders. One of the reasons behind the Great Crime Wave is that in the 1960s for the first time, the violent crime in minority areas ('China-towns') began to be reported to and recorded by the police (Figures 8 and 9).

One might have thought, given 'tribal' prejudices against outsiders and feared minorities, that punishment rates would have soared. But, thanks to Hitler and his enormities, overt racism, rather than normal became abnormal, even detestable. Capital punishment in the United States almost disappeared and reappeared only on a minor scale – even though half the murders involve blacks (only 12% of the population).

Welfare State Clientelism

The welfare state discredits the very notion of punishment because the welfare state was founded to reduce, not to increase suffering. Even today, there is a seldom-questioned belief that criminal impulses will be eliminated by a big enough welfare state. Thus, the welfare state as built in Sweden (having

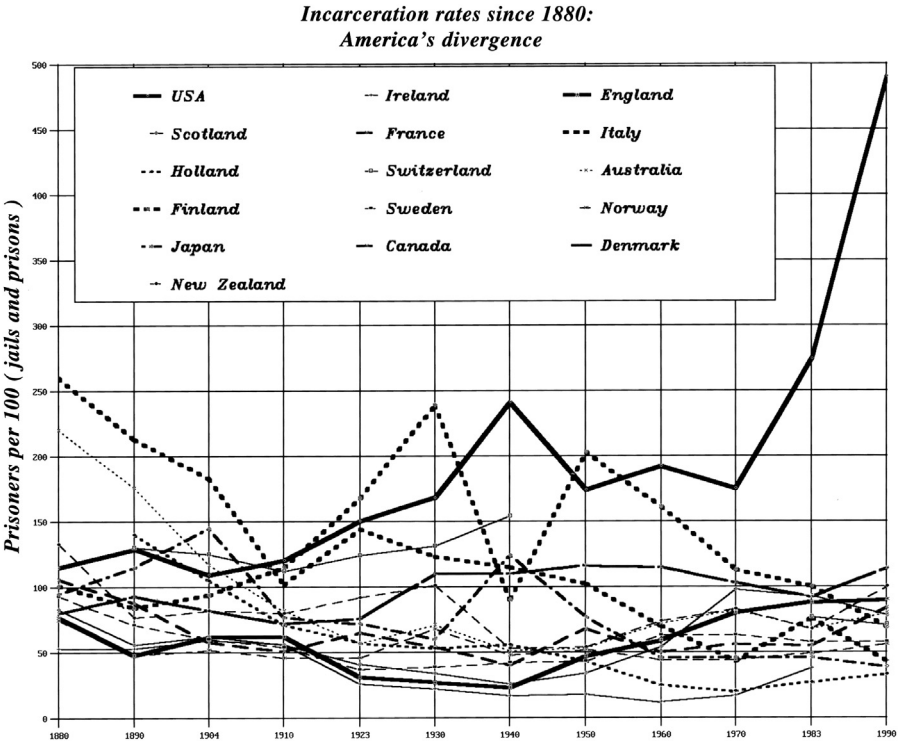

Figure 6 Decline of punishment since 1880.

eliminated need and greed) should have had little or no crime. Figure 10 shows that crime has more than thrived in the beacon welfare state. *Au contraire*, the welfare state has sharply reduced the negative consequences of doing crime.

The welfare state is designed to eliminate the unpleasant consequences of any kind of personal 'misconduct': crime, heavy drinking, casually impregnating women, spreading a social disease, dropping out from school, beating one's wife and kids, mugging someone on the street, and taking drugs. No one is to be 'marginalized' in the secular version of a Christian community. Thus, in German politics, it is still common and not silly for politicians to quote an aphorism of the progressive penal theorist von Liszt (1851–1919): Sozialpolitik ist die beste Kriminalpolitik. "Social policy is the best anti-crime policy."

The proper welfare state will also produce full employment and this, in the minds of many people, will prevent crime. It is received wisdom (no matter how often challenged by negative correlations between crime and unemployment) that unemployment means crime. For example, unemployment in the United States rose from 4% in mid-2008 to 10% in 2009 and 2010, dropping to 8% in 2012. The violent crime rate dropped from 458 in 2008 to 404 in 2010 and the property crime rate dropped from 3214 to 2941 (Figure 11).

Capitalism

Persisting neo-Marxists, seeing class economics always as *primum mobile* (the 'root' cause, cause of causes) have returned to the idea that capitalism (now in the guise of neoliberalism) fosters (needs?) increased imprisonment (Nelken, 2009). They argued that at least some types of capitalist systems (neoliberal ones) are exclusivist, with prison just one form of social exclusion. Imprisonment, rather than crime-related, results from neoliberal propensity to exclude undesirables. Other capitalist systems (such as in Europe) prefer to include (keep?) criminals around in society. This is perhaps why the leading neoliberal system (the United States) has come to have many lower crime rates than the inclusive (social democratic) capitalist systems. Strangely, neoliberal systems like England, the United States, and Switzerland have radically different per-capita prison populations: in 2010, 79 (Switzerland), 153 (England), and 756 (the United States).

Prisons and Budgets: Prisons versus Payouts

In democracies, government programs compete for shares in the budget. In the West, when budgets are made by government leaders alone, prisons often fare badly. Justice ministers are rarely powerful members of the cabinet. When prison-building bonds are put to the public for a vote, they fare better. Generally, however, welfare programs have much greater budget appeal than prisons: they involve tangible, spendable individual benefits and produce powerful interest groups to defend and expand their benefits. There are also greater payoffs (metaphorically speaking) for spending on policing than on prisons. Policing is an at-hand benefit compared to prisons. Moreover, governments have found that solving a few scandalous crimes and demonstrating government control of the streets may obviate the need for large numbers of routine arrests and prosecutions.

The drive to separate punishment policy from crime policy is often mirrored in the creation of separate ministries for crime (Interior) and punishment (Justice) – headed by ministers

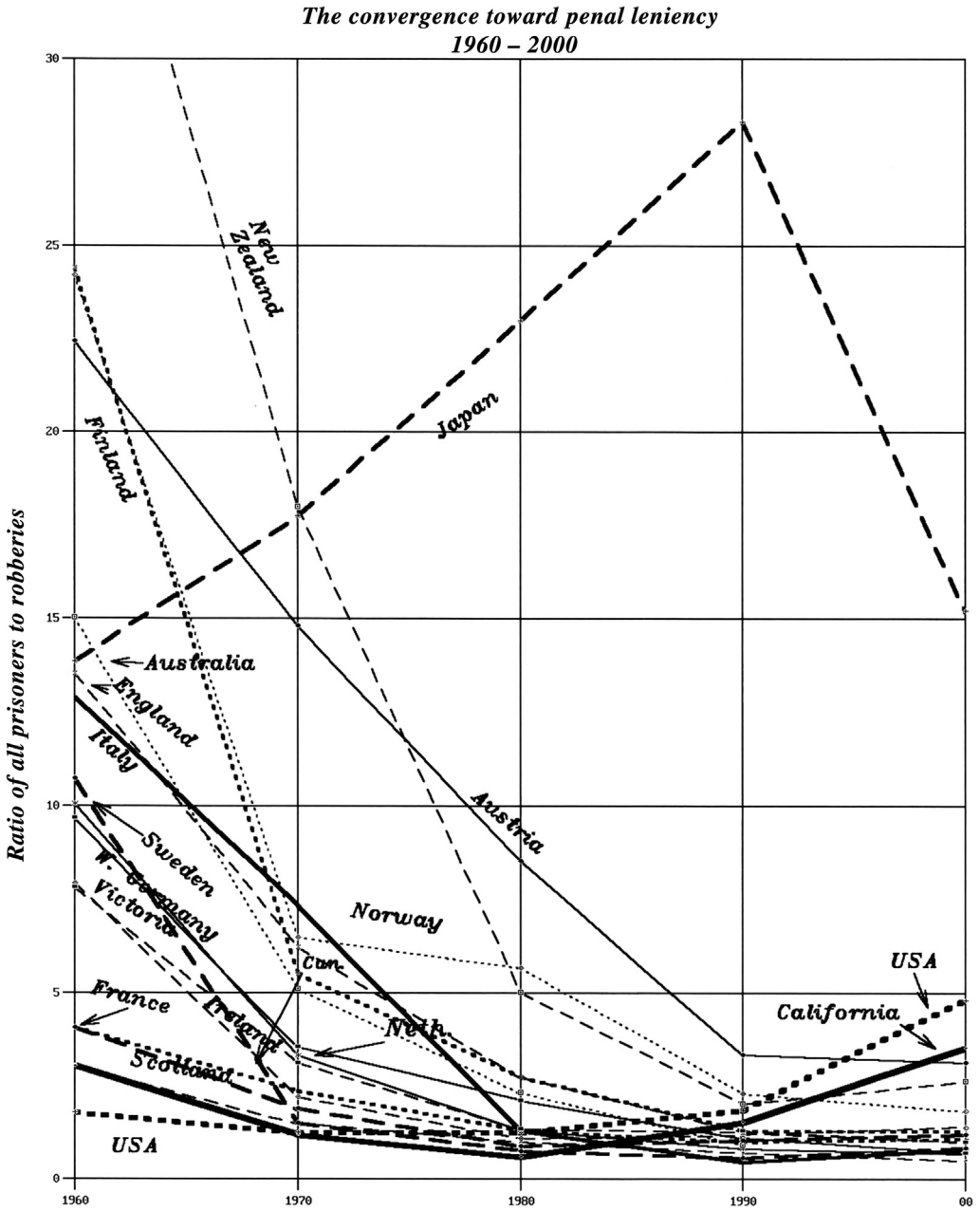

Figure 7 Convergence on minimal punishment.

known for being 'tough on crime' (Interior) and *laxiste* 'soft on crime' (Justice). To separate punishment from crime is the triumph of the antipunishment movement. Whence the regular complaint: "Interior catches, Justice releases."

Utilitarian Considerations: Prisons Fail versus Prison Works

Since at least the eighteenth century, prison reformers have tried to convince themselves and government that prisons are not (or should not be) punishment institutions but treatment or 'correctional' institutions. But as the belief in correctional benefits of prison has gradually faded, some social reformers have lost enthusiasm and joined the antiprison crusade. Belief in rehabilitative prisons, however, still survives in the twenty-first century. Tory Minister of Justice Clarke announced in 2011 that prisons were not (as he had somehow after all these years thought) rehabilitative and therefore should be emptied. Clarke's decarceration plan was opposed even by Labour leaders and most of the Tory Party.

Antiprison Professionals and Experts: Professionalism as Power

Foucault was all too right about the importance of professionals in modern life. Professionals want punishment policy to be a private preserve free from 'punitive populism' or populism of any kind; 'elite antipunitiveness' is to be like most government policies, almost totally invisible. Professionals

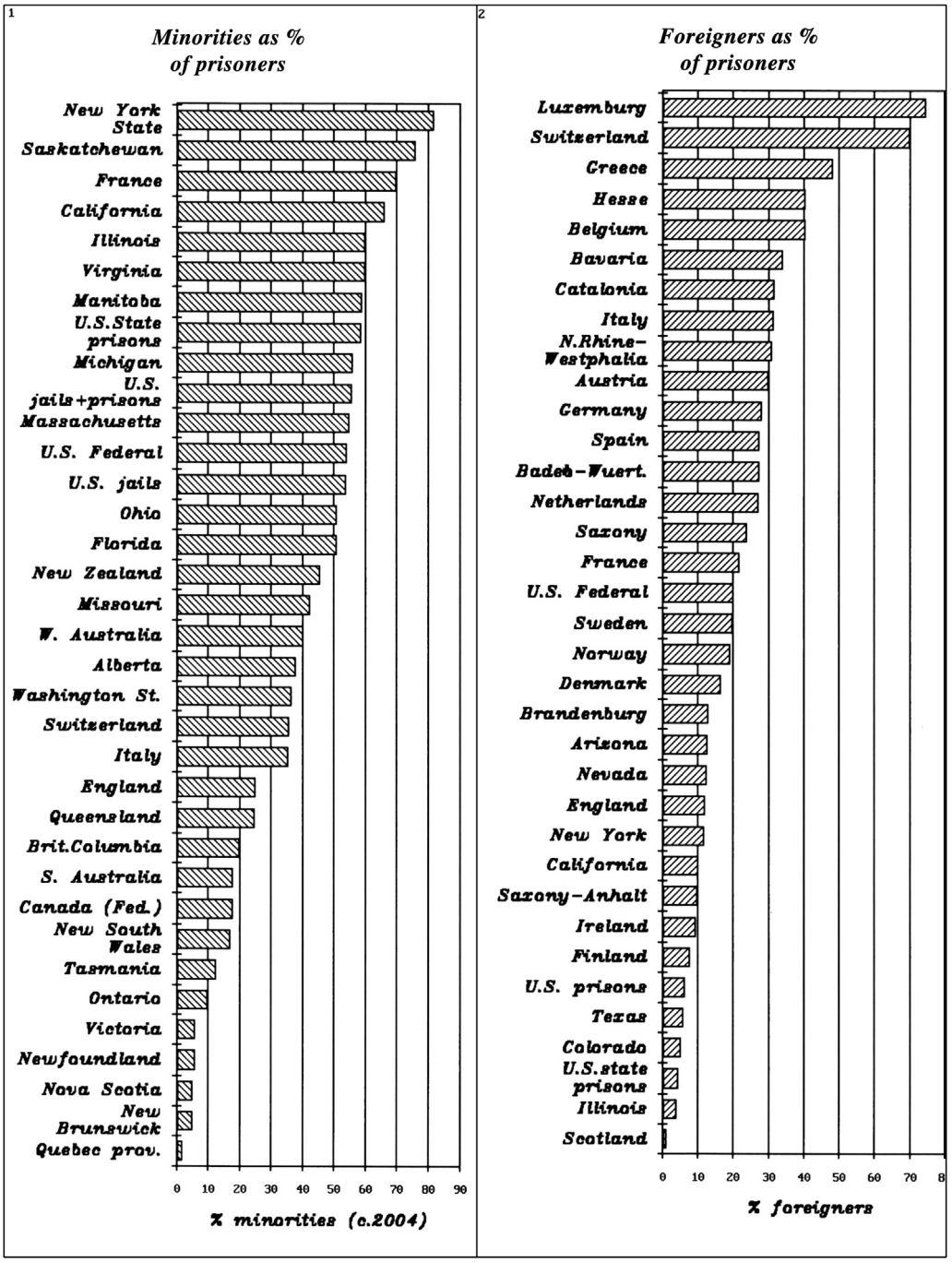

Figure 8 Minorities in prison *c.* 2000.

have been successful in Europe in separating punishment from crime because of the strong European acceptance of what the Swedes call (and routinely accept) *styrning uppifrån* (governing from above).

Alfred Blumstein, one of America's most eminent penalists, has even written that "one's normal conception of punitiveness [as measured for example by prison rate] [is] inconsistent with the principles of democracy" (Blumstein, 2007). Professionals have invented a set of insults to apply to any negative public response to crime and criminals: 'moral panic,' for example, suggests that crime fear (rather than based on burglaries and muggings in the neighborhood) is only fear of the bogeyman.

Opposition to prison (except as alternative to the death penalty) became in the nineteenth and twentieth centuries the mark of an enlightened, educated person and a prerequisite to standing in the helping professions. Expert opinion matters politically because governments often seek to legitimize action as being based on the best advice of the best professional experts. Given the domination of antipunishment attitudes

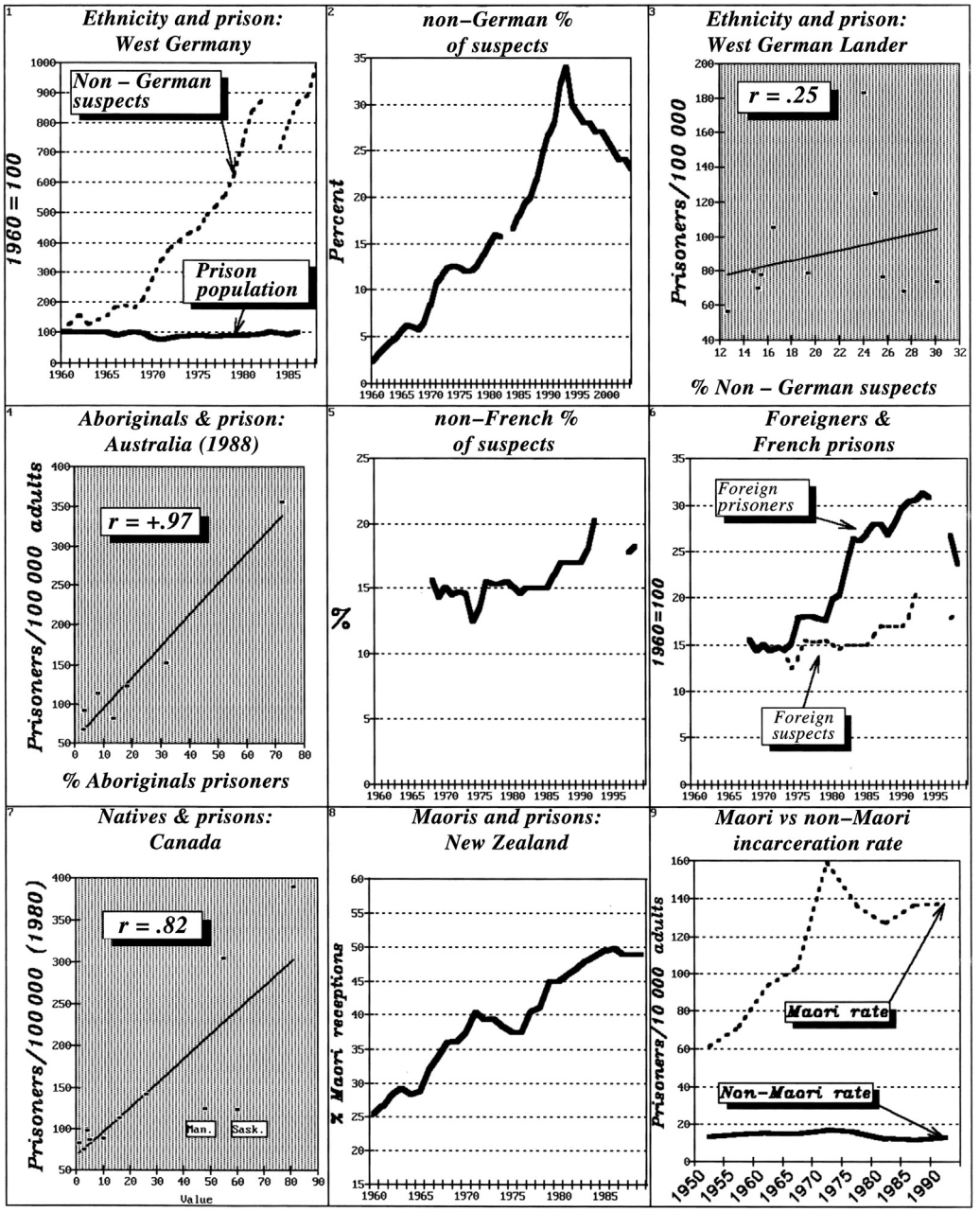

Figure 9 Minorities in prison.

among the experts, whether lawyers, sociologists, penologists, correctional officials, justice ministry bureaucrats, psychiatrists, journalists, clerics, or social workers, the separation of prison and crime and indulgent advice emerges routinely (e.g., Downes, 1992).

Finland is much feted for reducing its prison population by more than two-thirds since 1950 (despite the rise in the number of robberies from 210 in 1950 to 2600 in 2000); not incidentally does its leading penal expert describe the penal policy process in that country as "exceptionally expert-orientated. Reforms have been prepared and conducted by a relatively small group of experts…The impact of these professionals was further reinforced by close personal and professional contacts between some politicians, state officials and academic researchers. Consequently…crime control has never been a central political issue in election campaigns" (Lappi-Seppälä, 2002).

Typifying the bias of almost all penal experts has been the standard use of per-capita denominators in measuring prison rates and 'punitiveness.' Overpunishment is almost universally defined as anything above 100 prisoners per 100 000. The crime rate is bizarrely considered an irrelevant denominator by most experts in determining desirable punishment levels. The prisoners-per-capita rate, moreover, is not only easy to calculate but has great utility in the fight against imprisonment: comparisons are so easy and they all show how prison rates can be held low in such respectable countries as Sweden and

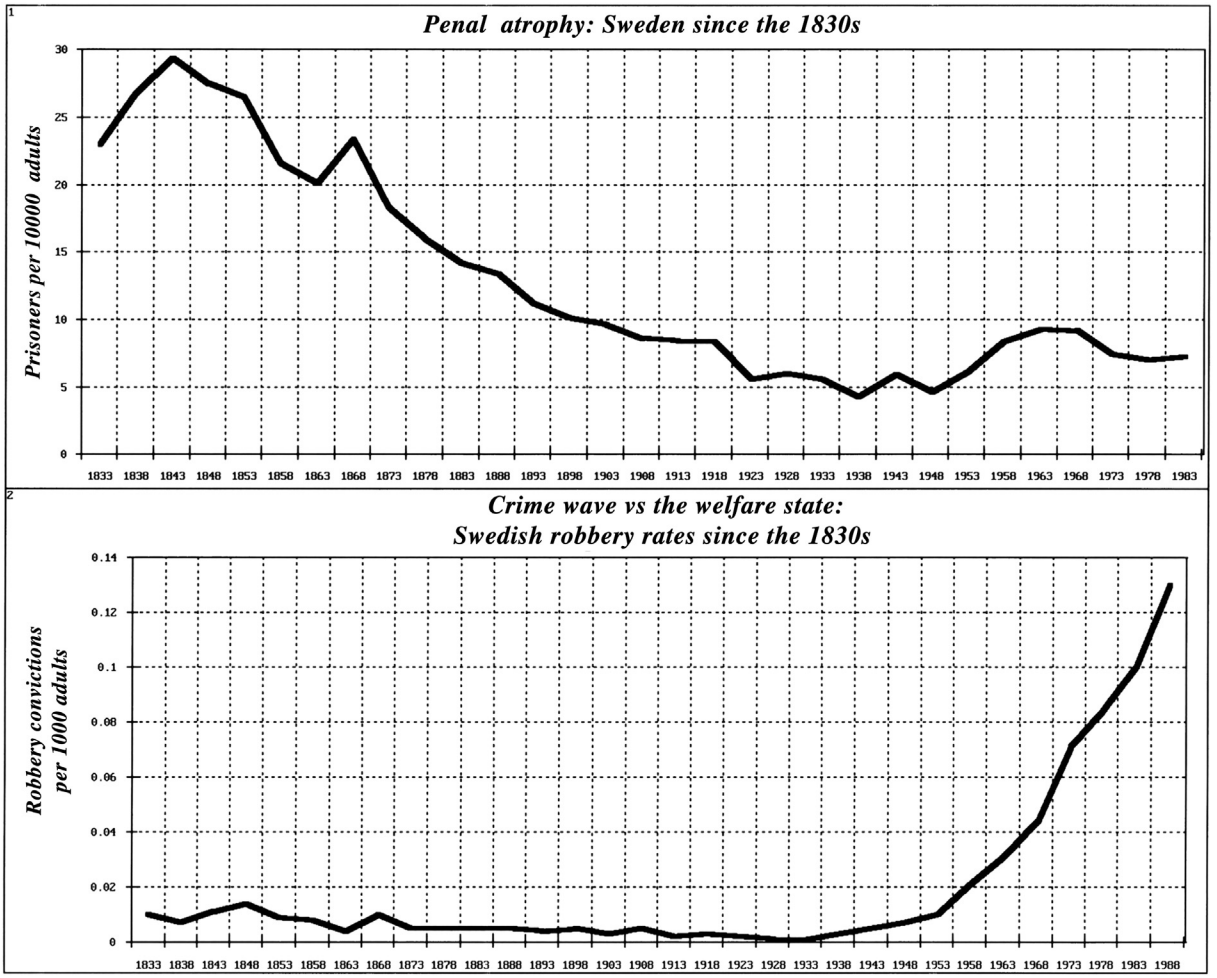

Figure 10 Crime prevention in the Swedish welfare state.

Switzerland – only the will to take expert advice and make imprisonment rare is necessary (Figures 12 and 13).

The author uses instead a nonconformist ratio: prisoners per robbery. Robbery has the advantage of being a serious crime, not that rare, that has for decades been recorded by most jurisdictions. Using the prisoner-per-robbery ratio produces radically different rankings in punitiveness and has the logic of relating punishment levels to serious crime rather than, say, the price of cabbage.

Eminent experts also argue that if crime has gone down in country A, which has raised imprisonment rates (say the United States), but also gone down in country B that did not raise imprisonment rates (say Canada), the raising of imprisonment rates cannot have been the cause of the drop in US crime. For some undisclosed scientific reason, every result must have the same cause everywhere.

Major comparative works have been published by professionals in the 2000s. An impressive number were sponsored and written by Professor Michael Tonry. Tonry's rich and detailed work represents the antipunishment movement at its most sophisticated. It is, of course, unashamedly antidemocratic, castigating 'politicization' and 'polemicization' when delicate questions of penal policy are involved. Election campaigns involving justice issues are debased by 'emotive appeals' resulting in 'coarse' policies (Tonry, 2007).

Hitler-Tainted (but Not Stalin-Tainted) Penal Policy

Hitler's glorification of law and order discredited the very notion of law and order. Punishment as a normal sequel to crime became tainted as fascistic. In the democracies, the antifascist parties after World War II (after inflicting sometimes indiscriminate punishments on collaborators) found punishment legitimate only as 'treatment.' One lesson of the Hitler experience was not learned: penal laxity breeds political extremism. The extreme right-wing parties to emerge in the later twentieth-century Europe often derived from the need for some party in France, the Netherlands, Italy, Norway, and Sweden to express popular feelings about the Great Crime Wave.

The Defendant-Rights State

Procedurally, the ideal democracy remains the Blackstonian state, which proudly releases 10, 20, even 100 guilty if necessary to acquit one innocent (Volokh, 1997). The Blackstonian state found its apotheosis in the US Supreme Court Miranda-

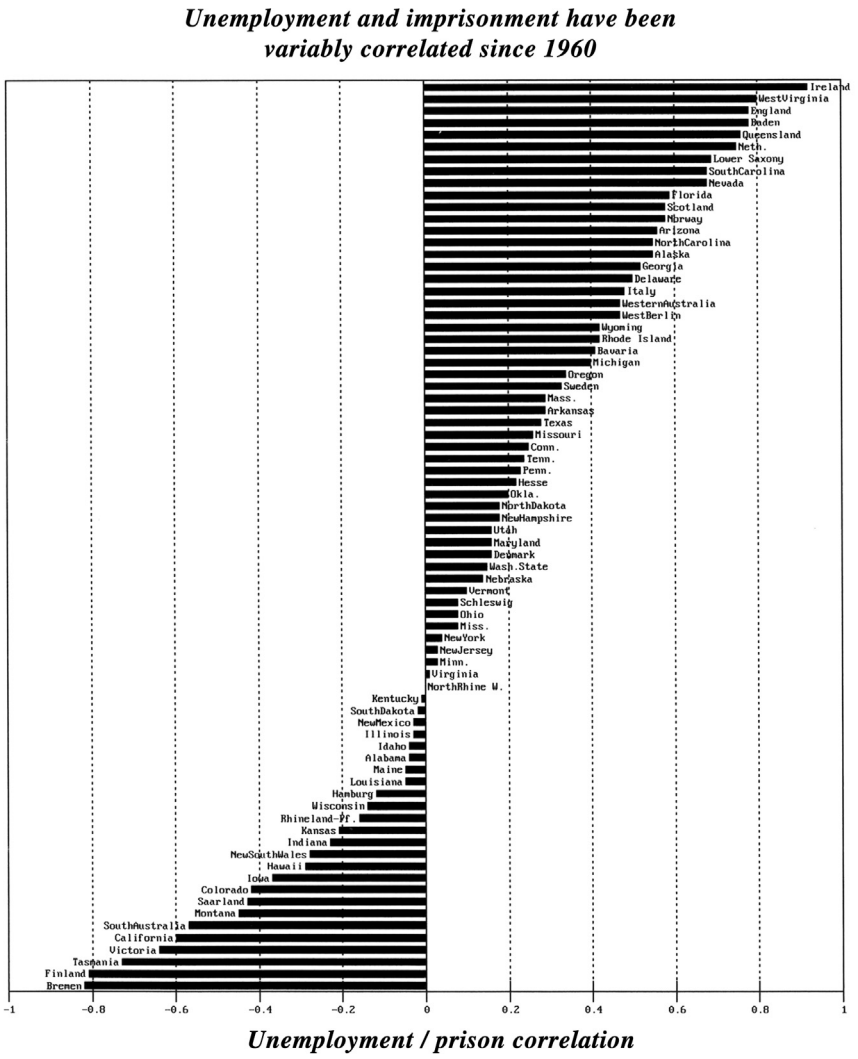

Figure 11 Unemployment rates vs prison rates.

type decisions of the 1960s and 1970s designed to make conviction and punishment as difficult as possible – this during the worst crime wave in American history. In the 1960s, the accusing state was presumed guilty.

Paradoxically, the more a country's justice system resembles the 'crime control' model rather than the 'due process' model (Packer, 1964), the more lenient its penal practices may be. Dutch criminal process is perhaps the least due process system among the advanced democracies; but criminals are totally willing to confess their crimes, knowing that the penalty will be slight. Criminals need no Blackstone when penalties are penalties-lite.

Cultural Consistency

Criminal justice in modern democracies is not unique. The decline of punishment is not restricted to the handling of criminals; it occurs in all spheres of society: the family, schools, the military, the university, jails and prisons, unionized business firms. A nonspanking family corresponds to a non-punishing state. To be legitimate, punishments have to be called something else and to be mild, painless, judicially supervised, easily obstructed by lawyers, and drowned in documentation.

The Great Crime Wave: The Role of Criminals (Deviant Individuals) in Reducing Punishment Rates

The prisoner-per-robbery ratio sank low after World War II because of the surge in robberies and a slump in prison sentencing. Figure 1 shows the dramatic success (numbers and productivity) of robbers in the democracies since 1960 (at least those reported). The democracies for some decades remained almost penalty-free paradises for robbers. Once-safe democracies after the 1950s were betrayed by their governments into allowing criminal activity to become normal. For example, violent crime increased by a factor of seven in Denmark and a factor of nine in the burgeoning welfare state, Sweden (Balvig, 2009). Reported robberies in Sweden rose from 190 in 1950 to 9570 in 2009. By the Nietzschean morality of the experts, tolerance for the Crime Wave made the Netherlands a beacon state 'beyond good and evil.' The average sentence for rape in the Netherlands was and is 11–12 months.

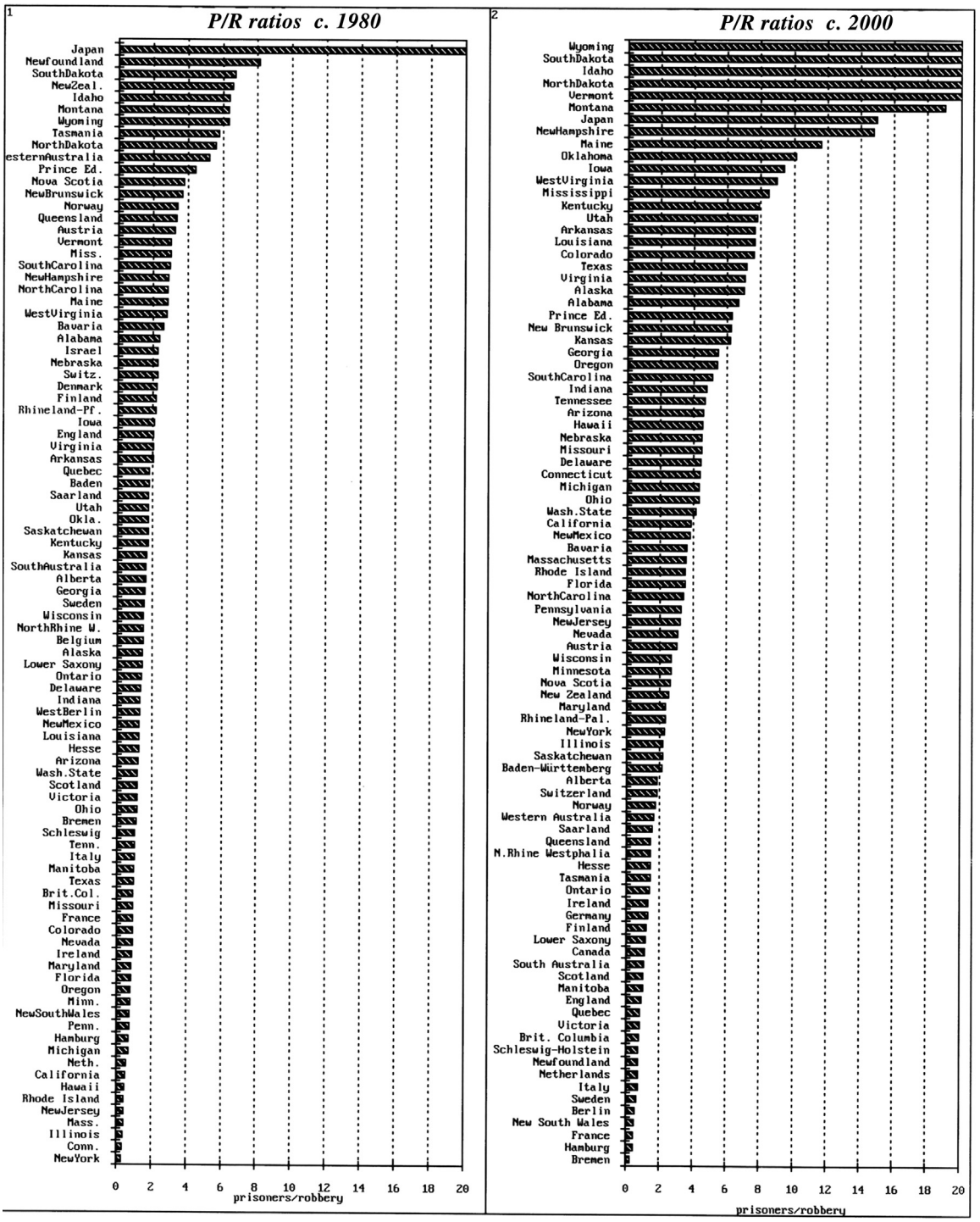

Figure 12 Per robbery prison rates *c.* 2000.

The Successful Challenge of Criminals to Mediocre Criminal Justice Systems

The 1950s were an idyllically safe period for most people – including those in minority neighborhoods. In the 1960s, however, thousands, indeed millions, of mostly young men rightly sensed a weakness in the authorities and stole, robbed, raped, and murdered – with impunity. And eminent jurists seemed oblivious in their rulings to any possibility that they might be encouraging the surge in rapes, aggravated assaults, and murders.

Overwhelmed by the mass of new candidates for punishment, the disjointed 'chain' of public authorities paid to deal

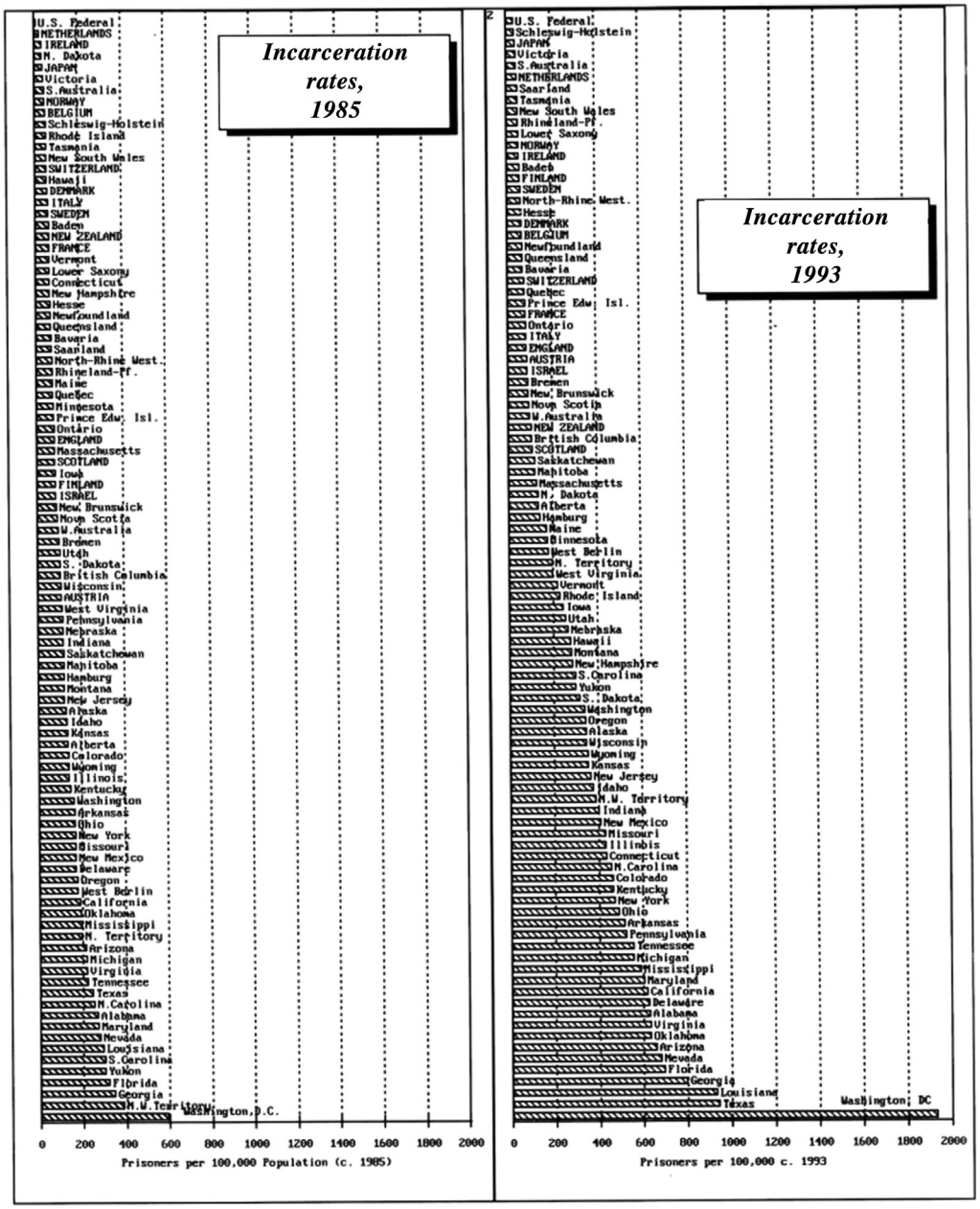

Figure 13 Per-capita prison rates *c.* 2000.

with criminals was able (with little criticism) to solve few crimes, prosecute (with little criticism) few of the few delivered into their hands, negotiate (with little criticism) convictions for only a few – on the basis of reduced charges. But the workload of judges was kept manageable. Prison sentences became rare but seldom served, and prison rates were kept respectably low (Figures 14–16).

What counted for the institutional survival and maintenance of these underpowered justice systems was the ability of the police, prosecutors, and judges to achieve occasional success in dealing with media cases. (Ironically, the institutional apparatus did play an important part in making the public aware of the seriousness of the crime problem – simply by repeating annual demands for more funds, personnel, and power to cope with 'an increasing crime problem.' This reinforced the conclusions from daily experience.)

The Netherlands, seemingly proud to tolerate crimes against its citizens, had no rivals when it came to refusing to acknowledge the Great Crime Wave. In 1980, the Netherlands, despite having a soaring crime rate, had the same small number of prison cells it had in 1840 when the country had one-fifth of the population.

Internationalization of Punishment Rules and Norms: Using the Treaty Power against the Voters

Prison reform, as well as the crusade against the death penalty and torture, have been international since Beccaria's *Dei Delitti*

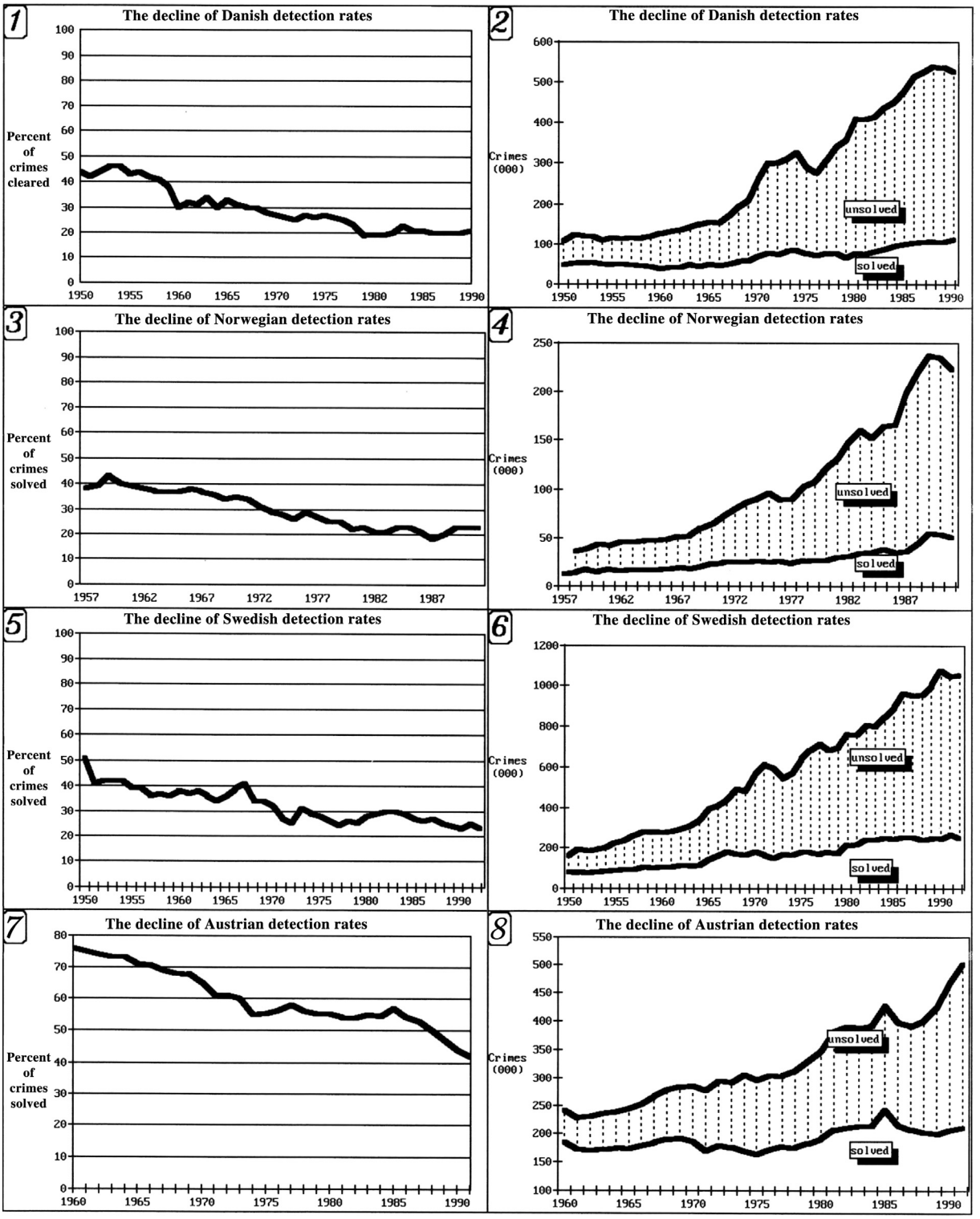

Figure 14 Getting away with crime.

e dei Peni (1764). In the subsequent Age of Democracy, internationalization became a major way elites could escape democratic control over crime and punishment policy. Penal experts much prefer international definitions of rights, crimes, and punishments – definitions made by fellow experts (seldom by elected representatives) in far-off places, selected by national executive elites. Thus, adoption into domestic law of the European Human Rights Convention by Britain allows judges (picked far from any public's scrutiny) to overrule any English or Scottish judicial decision seen as violating a 'human right'

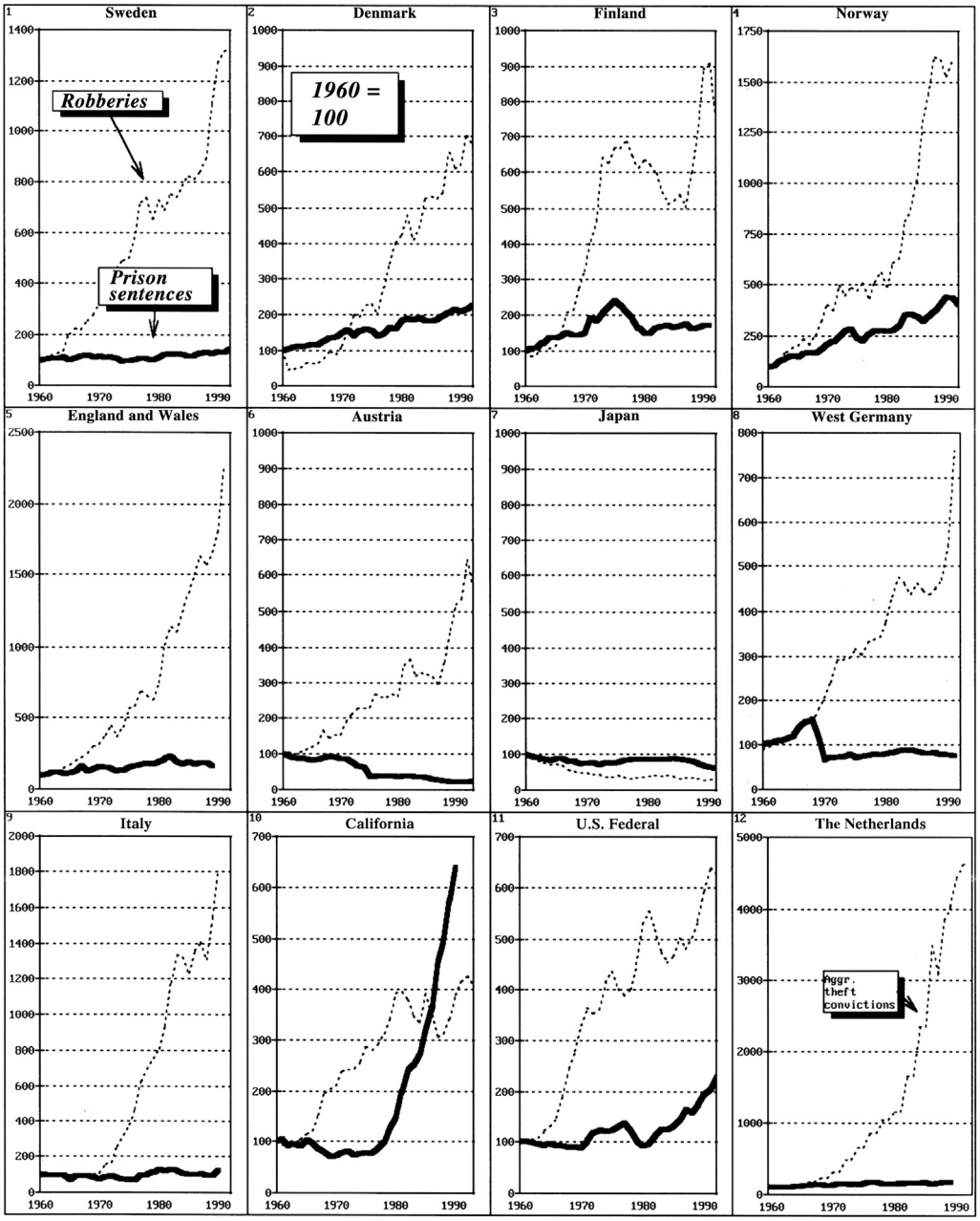

Figure 15 The gap between serious crimes and sentencing.

in the eyes of, say, an Armenian, Croatian, Montenegrin, or Russian justice.

Swiss elites, normally extravagantly concerned with national autonomy (they still refuse to join the European Union and joined the United Nations (UN) only in 2002), were happy to commit to the International Criminal Court, sign the European Convention on Human Rights, and accept the jurisdiction of the European Court of Human Rights – all means to remove justice questions from referendum voting. Adoption of that charter by the Swiss Parliament allowed European judges to overrule the usually sacred *vox populi* – a strong national referendum in 2004 in favor of life imprisonment for "sexual and violent criminals judged very dangerous and not criminal."

The mythology of internationalism is powerful – that international officials, bodies, courts, etc., are by some cloudy process made more objective, humanistic, and rights, committed than mere elected national authorities. In reality, the UN Human Rights Commission has routinely included representatives from states without human rights. Perhaps as important as formal international pressures against punishment are informal and regional pressures, such as those exerted on Finland by other Nordic nations. Finland maintained typically high Eastern European prison rates until the demise of

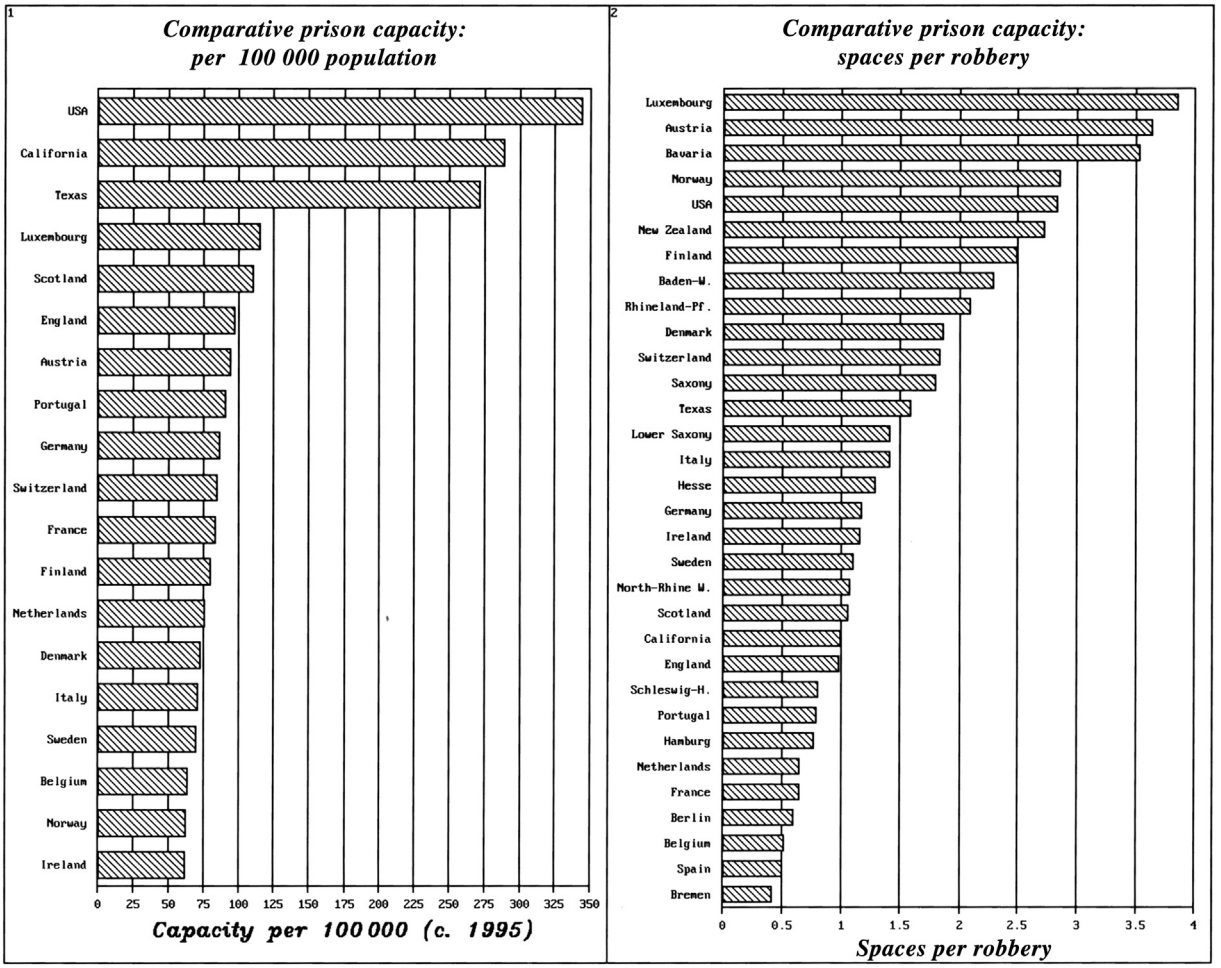

Figure 16 Comparative prison capacity.

its unofficial Big Brother, the Union of Soviet Socialist Republics. Then, in a dramatic effort to demonstrate its (west) European character, Finnish experts were able with no interference from Russia, Finnish party leaders, or Finnish voters to reduce its prison population to the lowest rate in northern Europe. Finland has thus made itself a 'beacon state' – a state that punishes criminals the least.

European countries like Finland want not only to be a beacon state but to be echt European (positive tropism), and not like the United States (negative tropism). Negative tropism toward the United States is one of the driving forces in European and Canadian government and politics, including penology.

Pretrial Detentions as Punishment

In a large number of democracies – national and local – pretrial detention serves informally as the principal 'dose' of punishment. Jails primarily house those who will never go to trial or if convicted, will be sentenced to 'time served.' Pretrial detention (deserved or not) is punitive but legally not punishment, at least in name, although conditions may be much worse than in convict prison. This is a major way in which police and prosecutors try to impose some punishment (Figure 17). (The population in pretrial status is, however, included in per-capita and per-robbery prison populations.)

Invention of Alternatives to Death and Prison: 'Net Widening'

Reformers have been fertile in creating ways for judges to avoid prison and death sentences via 'alternative punishments.' The death penalty has been replaced by imprisonment for life, life without parole, but most often, a dozen years or so in prison. Now also common in the United States is a 15-year delay between death sentence and execution. Generally speaking, taking a human life has been devalued, from being worth "a life for a life" to a dozen years in prison with color television and internet access. Incarceration as traditionally threatened by penal codes has been replaced by cautions, pretrial detention, deportation, probation, local imprisonment, part-of-the-week and part-of-the-day imprisonment, home arrest, restitution, community service, and shaking hands with the victim. But one must not forget queuing up for prison spaces: "we'll call when there is an opening for you." The Netherlands went in for prison waiting lists in a big way but eventually

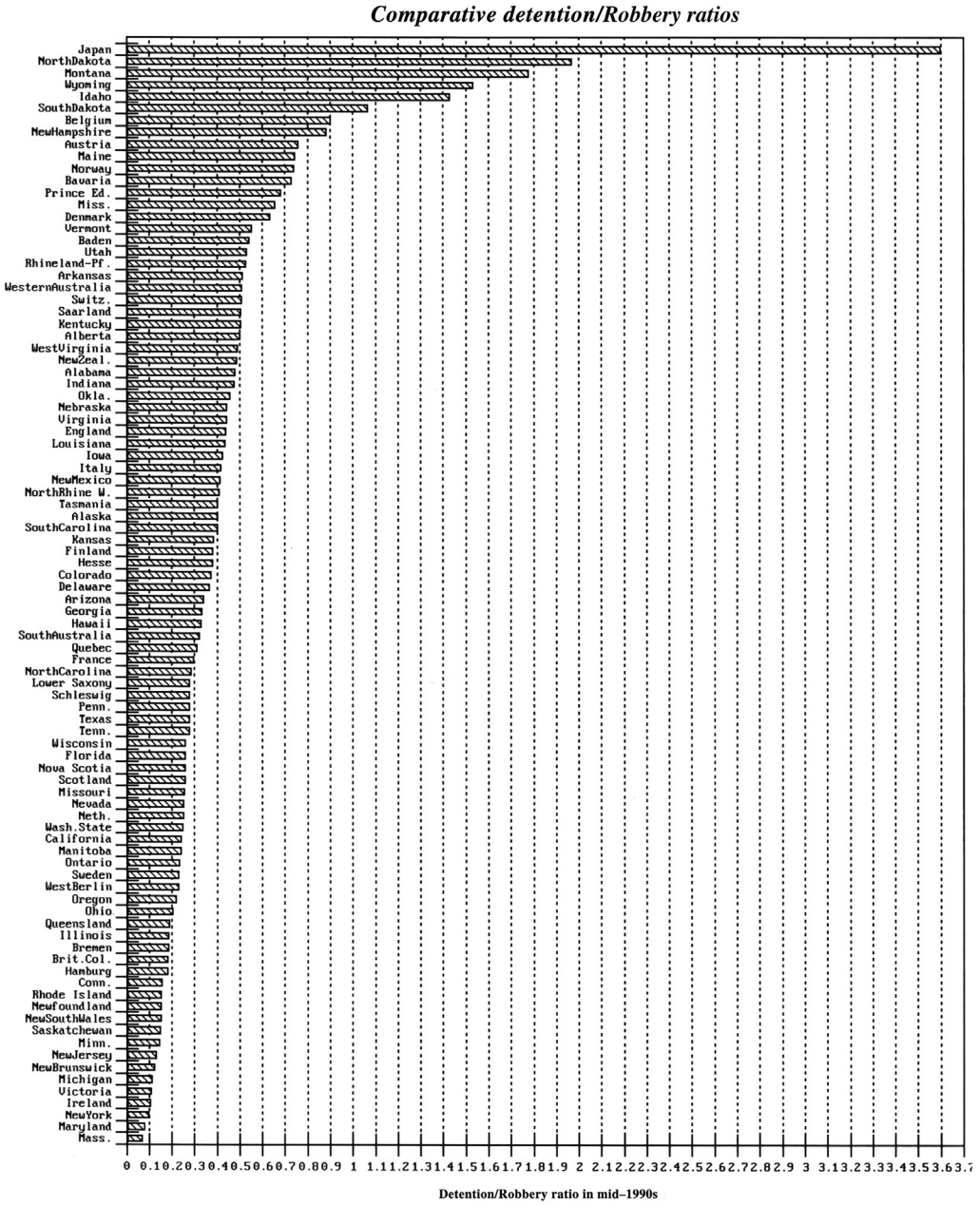

Figure 17 Comparative detention rates.

doubled their minuscule prison capacity after a man accused of rape was put on a waiting list and then made the mistake of raping a journalist.

Ironically, the concept of 'net widening' was developed by professionals just when the net of justice caught few and punished fewer. Normally pro-state when welfare and business regulation was involved, professionals were antistate when crime or mental illness was involved. Prisons reached their lowest populations in the United States in the midst of the Great Crime Wave and, not coincidentally, at the same time that 'deinstitutionalization' began delivering thousands of formerly hospitalized mental patients onto the streets. Mental patients have come to form a growing percentage of prison populations.

Process Can Be Made Costlier: Fiscalism

Capital punishment has been made prohibitively expensive in most places in the United States. In order to boost the cost of capital trials and appeals to millions of dollars, lawyers and judges have been allowed to require the employment of expensive lawyers and experts in any capital trial before it can be considered 'due process.' With the rulings of two sets of appellate judges, state and federal, prisons in the United

States can be made costly to operate, difficult to manage, racially explosive, and unsafe. American judges routinely decide whether any particular set of prison conditions constitutes 'cruel and unusual' punishment. (Ironically, California was criticized for spending more money on prisons than universities, as if prisons might not make greater immediate improvements of life quality for more people than universities.) European prison systems, for their part, are bombarded with criticism from domestic and international bodies, sometimes because prisons truly look like hotels, more often because they can easily seem to be what American judges would consider 'cruel and unusual' punishment.

Prisons Showed How They Could Be Made Ungovernable

In 1971, riots and violent state suppression made 'supermax' prison Attica a symbol of state sadism or prisoner sadism, perhaps both. In the rebellious atmosphere of the 1960s and 1970s, supposedly isolated prisoners demonstrated their ability to become sympathetic actors on the international stage, challenging the right of governments to govern them, much less punish them. Prison ungovernability came from the weakening of elite confidence in its right to govern and punish. It also came from the extension of outside revolts against authority and racial/ethnic tensions into prison. Guards with uncertain authority in Europe and America are now supposed to govern prisoners of diverse language, nationality, race, religion, urban versus rural origin, status in prison and society, not to mention reputation for brutality and gang connection.

The Defeat of Democracy: Elite Control of Crime and Punishment Policy by 'Platonic Guardians' (Loader, 2006)

Democracy is a very inefficient system of punishing politicians. The great mass of people, untrained in graduate schools and not having a candidate for felony punishment in the family, might be expected to punish lax governments, judges, and prosecutors when they can. But there are powerful elite barriers to taking penal codes seriously or to responding to 'punitive populism' (i.e., kindergarten notions that you will be punished if you hurt other people). In most elections, there are too many issues besides crime at stake. A party that is tender toward criminals is likely also to be generous in welfare state benefits. Decades of social democratic leniency toward criminals were decades in which half or more of the voting population became dependent on government jobs, contracts, medical care, housing, and money.

Then, too, law and order rhetoric by parties and media and interest groups often serves to mask lenient prosecution and sentencing. The Thatcher period was 'full of sound and fury,' unrelated to continuing lenient imprisonment policies (Farrall and Hay, 2010). Even when politics seems to be dominated by law and order rhetoric, this often has little impact on police, prosecutor, court, and executive-release preferences and routines.

Savelsberg describes Germany's version of 'not too much democracy' as compared to America's version of 'too much democracy' as follows:

> [German politicians are] much less likely to be responsive [than American politicians] … Their re-election as legislators depends little on their responsiveness to public sentiments…And judges and prosecutors never face elections but bureaucratically managed
>
> promotion systems. … New interest coalitions are hard to forge, new ideas are slow to spread, and decision-makers are slow to respond … We should not be surprised by the great stability of the macro-outcomes of decision-making in Germany, or by the great stability in incarceration rates (Savelsberg, 1999).

Democratic societies, if criminal justice is any indicator, survive by performing at 5–10% performance levels. Democracies seem to handle the crime and punishment problem mostly by relying on class structure and social geography: quarantining violence, focusing police attention and jail sentences on ghettoes and *banlieues*, and policing ghetto borders.

The Occasional Defeats of the Platonic Guardians of Minimal Punishment

Voters, even in west Europe, do not always supinely accept the nonpunishment of serious criminals – often revealed only in media reports on the past judicial handling of those arrested or convicted of serious crimes. There are occasional revolts against elite high-mindedness and apathy.

The Netherlands

Even the much-lauded deferential Dutch voters eventually learned that they could disobey their subcultural masters and vote to express displeasure with rampant criminality. Their Catholic and Liberal masters, joined in the 1980s even by their Socialist masters, grudgingly began to acknowledge the horrendous crime problem and for the first time in a century to build some prisons. Holland in the 1950s had been one of the safest places in the world; it has since become, with the help of antipunishment prosecutors and judges, one of the safest places in the world to do crime.

It was and is very difficult for Dutch voters to affect national policy: they have no individual district representatives in parliament; their mayors and governors are appointed by The Hague; the party system is so fragmented that governing coalitions emerge only months after elections, often in no relation to the vote; and judges are all civil servants. But Dutch voters did, as noted, get some prisons built after two or three decades of suffering rising violent crime. In 2012, the government even began shutting some of them down in view of declining crime. But at the acme of its apparent punitiveness, the Netherlands, as **Figure 13** shows, still ranked in the bottom 1–2% of jurisdictions in the ratio of prisoners to robberies.

The Blair Revolution

After four consecutive defeats at the polls (1979, 1983, 1987, and 1992), the British Labour party chose Tony Blair as Shadow Home Secretary in 1992, and he shockingly declared that crime

was a socialist issue. Ironically, the Tories under Margaret Thatcher, the supposed Iron Lady, had appointed a long series of lenient Home Secretaries (Farrall and Hay, 2010). To satisfy Labour's traditionalists, Blair's slogan, which helped win the 1997 elections, was: "Tough on crime, but also tough on the causes of crime." The number of prison sentences did increase after Labour took over in 1997. (The Tories, in response to Blair's radically new stand, appointed their first tough Home Secretary, Michael Howard, in 1993.)

The US Penal Revolutions

When democracy is not artfully stifled, there is likely to be more punishment for more crime. Some American states carried out great reversals of penal policy. The lenient policies of the 1960s and early 1970s – including abolitions of capital punishment and reductions in prison populations – were rejected in state after state, beginning in the later 1970s. Ordinary Americans, unlike ordinary Europeans, can sometimes shape penal policy. California voters by direct democracy brought back the death penalty a year after its supreme court abolished it. A decade later, in retention elections, they removed supreme court justices who had found a reason to reject every one of the hundreds of death penalty case submitted to them. In other states, lenient prosecutors and judges were removed from office. Judges were encouraged to take crime seriously and incarcerations increased at the state and federal levels. But, as Figure 12 shows, punishment rates still vary greatly among the American states: states with big cities remain the least punitive.

The White March in Belgium in 1996

In 1996, hundreds of thousands of Belgians, although divided and subdivided by culture, language, religion, region, gender, and class, gathered together to protest the years of mishandling of pedophile cases by what passes for a state in Belgium with a march in Brussels. Although not a riot, the march achieved a reorganization of Belgian justice and the sentencing of the multiple child rapist/murderer to a 'life' sentence – 8 years after the first rape-murder.

Conclusion

Most often, however, democracy allows criminals (the few), allied with lenient elites ('Platonic guardians'), to outweigh their victims (the many). In perhaps most instances, punishment decisions in democracies (even in the direct democracies like the Swiss cantons and California) are made by focused and interested elites. (The nonpunishment of white collar crime, financial crime, political corruption, fraud, tax evasion, and multiple others may seem less outrageous than nonpunishment of violent crimes. But some of these crimes cause lots of direct hurt to people. Comparative study of democracies and their nonpunishment of bourgeois crime, however, is another, and probably equally dismaying study.)

See also: Age and Crime; Class and Crime; Communities and Crime; Community Supervision and Diversion in the United States; Comparative Studies: Method and Design; Control Theories and Crime; Crime and Gender; Crime, Desistance, and Persistence; Crime, Geography of; Crime: Knowledge about and Prevalence; Crime, Race, and Ethnicity; Crime Trends and Debates; Data Bases and Statistical Systems: Crime Measurement; Data Bases and Statistical Systems: Crime; Delinquency, Sociology of; Deterrence Theory: Crime; Deterrence; Immigration and Crime in the United States; Imprisonment: Sociological Aspects; International and Comparative Criminology; Mass Imprisonment and Contemporary Collateral Consequences; Media, Crime and Crime Waves; Police, Sociology of; Policing; Prisons and Imprisonment; Quantitative Cross-national Research Methods; Race: Prespectives and Crime; Rehabilitation of the offender; Sentencing, Trends and Disparities: Crime; Sociology of Crime; Transnational Organized Crime.

Bibliography

The sources of the statistics in the charts come from mostly printed but scattered materials consulted in the libraries of UCLA, Yale, LSE, Institute of Criminology (Cambridge), Swiss Federal Statistical Office, British Home Office, Dutch Ministry of the Interior, Max Planck Institute (Freiburg), and more. Of primary importance were the sometimes long runs of prison statistics contained in the statistical yearbooks of nearly twenty countries and about 80 sub-national states in those countries. Current statistics are available also in the Council of Europe, Annual Penal Statistics. Invaluable is the European Sourcebook of Crime and Criminal Justice Statistics. Several volumes of the Criminal Justice annual published by the University of Chicago Press have been devoted to comparative penal politics.

Balvig, Flemming, 2009. Violence. In: Looking at Violence in the Nordic Countries. http://www.norden.org/en/publications/publikationer/2009-542.

Banfield, Edward, 1974. The Unheavenly City. Little-Brown, Boston (Chapter 9).

Blumstein, Alfred, 2007. The roots of punitiveness in a democracy. Journal of Scandinavian Studies in Criminology and Crime Prevention 8 (Suppl. 1), S2–S16.

Downes, David, 1992. Contrasts in Tolerance: Post-war Penal Policy in the Netherlands and England and Wales. Oxford University Press, New York.

Elias, Norbert, 2000. The Civilizing Process. Sociogenetic and Psychogenetic Investigations. Blackwell, Oxford.

Farrall, Stephen, Hay, Colin, 2010. Not so tough on crime? British Journal of Criminology 50, 550–569.

Lappi-Seppälä, Tapio, 2002. Penal policy and incarceration rates. Corrections Today 64, 30–33.

Loader, Ian, 2006. Fall of the 'platonic guardians': liberalism, criminology and political responses to crime in England and Wales. British Journal of Criminology 16, 561–586.

Nelken, David, 2009. Comparative criminal justice: beyond ethnocentrism and relativism. European Journal of Criminology 6, 291–311.

Packer, Herbert, 1964. Two models of the criminal process. University of Pennsylvania Law Review 113, 1–68.

Savelsberg, Joachim, 1999. Knowledge, domination, and criminal punishment revisited. Punishment and Society 1, 45–70.

Tonry, Michael, 2007. Determinants of penal policy. In: Tonry, Michael (Ed.), Crime, Punishment, and Politics in Comparative Perspective. The University of Chicago Press, Chicago, pp. 1–48.

Volokh, Alexander, 1997. n guilty men. University of Pennsylvania Law Review 146, 173–216.

Punishment: Social and Legal Aspects

Jonathan Simon, School of Law, UC Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While punitive sentiments come as close as any human motivations to being universally observed, the major finding from the history of punishment is one of change and variation. The scholarly study of punishment as social and humanistic inquiry has sought to make sense of the major shifts in punishment and society through historical and comparative research drawing on insights and inquiries from five theoretical 'toolkits.'

Punishment, Law, and Society

> the concept of 'punishment' presents, at a very late stage of culture … not just one meaning but a whole synthesis of 'meanings.'
>
> Nietzsche, 1994: p. 57

Research from fields as diverse as psychology (Vidmar and Miller, 1980), evolutionary biology (Gachter et al., 2008), and archeology (Gager, 2006) suggests that the perceived/internal imperative to punish, the sense of satisfaction at the thought of a wrongdoer receiving some compensating sanction, is a cross-cultural universal found in virtually every society of which we have records. It is arguably rooted in the evolutionary advantages achieved by primitive human groups who practiced it. But while the imperative to punish may be a general human trait, the historical record suggests that the methods, the meanings, the motivations, the justifications, the organization, and the myriad effects of punishment vary enormously and change over time according to a wide variety of social, economic, political, legal, and cultural factors (Durkheim, 1969; Rusche and Kirchheimer, 1939, 2003; Garland, 1990; O'Malley, 1999; Hudson, 1996).

It is the whole 'penal field' (Page, 2013), defined by those means, meanings, and consequences, and the dynamics of change within it, that forms our subject. A generation ago, it would have been subsumed in the field of penology, the scientific study of criminal deviance and methods, for its diagnosis and reform. The history of previous penal logics or methods would be little more than a prolegomenon to the progressive march of experimentation. Today, for reasons suggested later, the historical change in this field has become and remains central. We can use the term 'penality' (Garland and Young, 1983; Garland, 1990) to describe this complex of techniques, institutions, discourses, and their historical genealogy. This article introduces the major historical transition points in Western penality since the early modern period and to the most significant 'toolkits' that have been deployed by punishment, law, and society scholars since the revival of the field in the early 1970s.

Part I. The Historical Context of Western Penal Development: Capitalism and the Nation-State

The contemporary penal landscape in the world today is largely the result of five transformational ruptures in Western penality since the early modern period.

Rupture One: The Festival of Cruelty

The rise of the nation-state beginning in Europe in the sixteenth century co-opted the traditional role of families and private local landholders in administering punishments, making public punishments and important theater for the display and affirmation of political authority (Spierenburg, 1984; Foucault, 1977). This seems to have gone along with an overall increase in the frequency and severity of punishment, especially the use of tortuous capital punishment as well as a gradual legalization of punishment (increasing levels of due process and the involvement of professional lawyers) beginning in the mid-eighteenth century (Langbein, 1994).

Capital punishment played a quantitatively small role in early modern penality, but its rituals played a central role in the system of state trial and punishment. In England, for example, black-hooded judges pronounced sentence in capital trials, followed by public performance of a vividly violent execution to eventual display of the corpse (Gattrell, 1994). In a typical European ritual, the performance of various forms of minor humiliation or painful treatment would precede the execution of capital punishment as the final event (Spierenburg, 2004). As the culmination of a 'festival of cruelty,' the execution emphasized physical suffering and its display (Foucault, 1977; Garland, 1990: p. 64).

Rupture Two: The Birth of the Prison

Beginning in the late eighteenth century, the states with the most advanced central administrative capacity began to reduce their reliance on capital punishment, especially torture-laden public execution spectacles, in favor of a range of less visible and sanguinary practices, including transportation, public labor, and carceral confinement. By the middle of the nineteenth century, the scaffold as the main event in public capital punishment largely disappeared within the walls of a form of

punishment that had become the most successful alternative to replace it, the prison.

Rupture Three: Surveillance Societies

Beginning in the late nineteenth century, and through the mid-twentieth century the most advanced states professionalized and extended their penal system from prisons to community surveillance. Specialized carceral regimes for juveniles and other classes of offenders emerged (Garland, 1985; Rothman, 1980). The focus of prison regimes on discipline and hard labor also shifted in favor of education and training. A new set of penal practices, including parole, probation, and juvenile justice, saw modern state penal control extend from the prison into the community. Historians have treated this as a considerable expansion of state power over the poor and working classes. Using the metaphor of a net, these new penal forms are said to have widened the extent of the net (extending social control to more citizens) and thinning its mesh (closing the gaps through which lower-level offenders previously escaped sanction; see Platt, 1972; Rothman, 1980; Cohen, 1985; Rubin, 2012).

A new penological discourse combined the promise of rehabilitation through psychological and educational interventions, with eugenic prevention through incapacitation. It promised to reform the nearly century-old penitentiary project and create new forms of community supervision. The 'new penology' as it was sometimes called, promoted the ideal of individualized treatment, normalization (that is, preparation for return to the community), and progressivity but only for those deemed biologically suitable for success in free society. The strong influence of Darwinian population eugenics and scientific racism informed penal policy even more profoundly. The result was a broad extension of a penal field that remained heavily oriented toward imprisonment in non-therapeutic penitentiaries designed to incapacitate large numbers along with the most basic and externally unpromising forms of labor (Rothman, 1980; McLennan, 2008). World War II and global revulsion at the crimes of the Axis regimes helped to discredit the eugenic aspects of this approach, while its rehabilitative and correctional features became inscribed in central human rights charters. They were also embraced by the penal laws in most parts of the world, albeit with widely varying degrees of actual compliance, by the middle of the twentieth century. Following the war, penal systems in the wealthy democracies slowly moved toward a greater realization of the therapeutic model of reform penology, one bolstered in part by the rising wave of human rights thinking (International Covenant of Civil and Political Rights, Art. 10.3).

Rupture Four: Mass Incarceration

In the late twentieth century, the United States experienced a rapid and dramatic rise in numbers of people incarcerated in jails or prisons, and under community supervision, along with a qualitative shift toward harsher and more degrading punishments (Garland, 2001; Gottschalk, 2006; Western, 2006; Wacquant, 2009; Lynch, 2010; Whitman, 2003). This has been experienced by minority populations in particular, who saw already high mid-twentieth century levels of imprisonment reach supersaturation rates (Davis, 2003; Western and Petit, 2004; Alexander, 2013). A similar, but less dramatic growth in the scale of penal populations has occurred in other parts of the Anglozone, Europe, as well as parts of Latin America and Asia, mostly with a lag of a decade or more (Lacey, 2013; Pratt, 2007; Nelken, 2011) although nowhere near the levels of the US. As with racial minorities in the US, immigrants, especially those from outside Europe with marginal legal status have been a major focus of increased imprisonment in these nations (De Giorgi, 2006).

In the United States, the increase in prisoners has been fueled by a populist politics of punishment and the abandonment of modern correctional objectives and restraints (Scheingold, 1991; Garland, 2001; Simon, 2007). The modernist correctional apparatus associated with twentieth-century penology has been transformed from a normalizing force of social control, on both society and the penal system, into a managerial expertise with little knowledge of crime and little commitment to humane treatment (Feeley and Simon, 1992; Bottom, 1995; Garland, 2001). Rehabilitation and selective incapacitation of the incorrigible as modernist penal rationales became discredited from a variety of scientific and political perspectives, and a more generalized logic of incapacitation – anchored in a perception of prisoners as representing a high, unspecified, and unchanging criminal risk (Zimring and Hawkins, 1995; Simon, 2014) – took over.

Rupture Five: Dignity

There is evidence of a fifth rupture in process today in the emergence of globalization and institutions of international legality, and framed around the problem of human rights. Since the adoption of the UN Universal Declaration of Human Rights in 1948, considerations of human dignity and protection from torture, cruelty, and degradation have gradually become more central to legal challenges to penal practices and policies. In some countries and whole regions, human rights norms are reshaping the penal field through legal constraints on severe punishments as well as normative pressure on penal institutions to shape routines that conserve the dignity of those being punished and those carrying out the punishments (Whitman, 2003; Van Zyl Smit and Snacken, 2009).

This represents a major shift away from the idea of punishment as a form of state policy aimed at managing and securing the population. Whether in the form of the interventionist treatment-oriented penal welfare state of the mid-twentieth century, or the punitive and incapacitation-oriented penal state of the late twentieth century and early twenty-first century, modern punishment has presumed a strong state utilizing punishment to defend society (or critiqued for using it as a political tool). Human rights norms were largely relegated to the periphery, in the gradual abandonment of certain severe penalties as inhumane (torture, mutilation, slavery, solitary confinement). In the robust version now operating in Europe, the entire course of imprisonment is subject to three key demands necessary to preserving the human dignity of prisoners: (1) individualization, that is, decisions about them must be made with consideration of their individual circumstances and how that may change; (2) normalization, as much as is feasible, prison life should harmonize with free life and prepare

prisoners for return to freedom no matter what their actual prospects for release; and (3) progressivity: the possibility of change and improvement in the prospects of prisoners as their sentence is served (e.g., presuming good behavior) no matter what their actual terms of release. Citing the dignity principle, a number of globally influential human rights courts found unlawful the issuance of life sentences (known as 'life sentence without parole' or 'LWOP' in the United States) (Vintner et al. vs the United Kingdom, 2013).

Part II. Punishment and Social Theory: A Toolkit for Exploring Penal Variation and Change

The question of how these ruptures in penality relate to particular changes in law and society has produced an extensive and growing body of both social theory and empirical study. This literature is far too extensive for a handbook (Simon and Sparks, 2013) let alone an encyclopedia article to begin to summarize, but following the productive model of Garland's (1990) *Punishment and Modern Society*, I describe five major 'toolkits' drawn from both classical and contemporary social and cultural theory, being used in sum or in part by many scholars, to describe and explain changes in punishment, law, and society.

Toolkit One: Punishment and Solidarity

For Durkheim (1964, 1969) crimes are any action that violate or offend the fundamental moral values, or collective conscience, of society. When members of society come together to punish the crime, their moral outrage reassures the people of their collective commitment to these shared values, thereby reinforcing society's solidarity. Punishment thus helps to hold society together and provides its normative structure.

To Durkheim, these visible changes in penal practice were an empirical reflection of underlying changes in the social solidarity of modern societies. Traditional societies, characterized by little division of labor or variation in social forms, and unified by powerful, demanding, and often sacred norms in what Durkheim called 'mechanical solidarity' (a strong form of mutual identification through likeness), produced harsh punishments that reflected the large role played by the penal law in reproducing such solidarity. Modern societies, characterized by an extensive division of labor, diverse social forms, and unified by shallower but more extended norms associated with rational and secular values, produced more lenient punishments that reflected the reduced role of penal law in reproducing solidarity (the greater burden of which was now carried by what Durkheim called 'restitutive law,' including the law of contracts).

In a later elaboration (1903) Durkheim largely credited the reduction in traditional religious values for reduction in both the qualitative demands of punishment (whether it demands permanent mutilation or death on one hand or 'merely' a term of penal service on the other) and the overall severity of its output (whether the number and severity of sentences goes up). In addition to these long developmental processes, Durkheim recognized the significance of shorter term political developments, such as the potential for authoritarian governments to use punishment to inspire popular fear and awe among the masses.

Durkheim was reflecting on ruptures one and two. Since the rebirth of interest in punishment and society in the 1970s, scholars have drawn on aspects of his account to explain the emergence of penal welfarism in the twentieth century (rupture three; Rothman, 1980; Garland, 1985). If human rights norms like dignity succeed in restraining mass incarceration and realigning global practices toward restraint and leniency, it will no doubt fuel Durkheimian theories of the emerging global society.

Toolkit II: Punishment and Social Structure

Marxist scholars have related penal change to the class structures of society, and particularly the changes associated with the rise of capitalism. Georg Rusche and Otto Kirchheimer (1939, 2003) explored ruptures one, two, and the beginnings of three as a function of the political economic conditions and the structure of class power underlying punishment and society. The harsh punishments of physical mutilation and death frequently exacted by courts of the early modern period reflected the lack of economic value in human labor created by conditions of surplus labor supply, as feudalism broke down but before capitalism had produced a demand for unskilled labor. The trend toward leniency, which Durkheim had also observed, reflected the value of labor as commercial society began to develop and the population surplus declined through colonialism and the relocation of many laborers to the New World. New punishments, like convict transportation and houses of correction, emerged to use the labor of those undergoing punishment. These punishments, and the criminalization of vagrancy (the state of being unemployed), also helped to incentivize prospective workers to stay in the labor market instead of waiting for better conditions or higher wages. For Rusche and Kirchheimer (who wrote before World War II), the promise of penal welfarism to reform prisons along treatment lines which found reflection in the legal code of Weimar Germany and other European societies during the 1920s depended largely on the ability of global capitalism to escape crisis, and the success of the countermove toward penal harshness associated with the Nazi government in Germany.

Punishment may also serve a sophisticated political tool of class rule available to leaders attempting to mediate the conflicts inherent in class-divided societies like England during the rise of capitalism (Thompson, 1975; Hay et al., 1975; Linebaugh, 2006).

More recently, sociologists and geographers have suggested that mass incarceration in the late twentieth century reflects a strategy of control associated with globalization and neoliberalism (Garland, 2001; Western, 2006; Gilmore, 2007; Wacquant, 2009).

Toolkit Three: Punishment and Political Technologies of the Body

After Durkheim (1969) and Rusche and Kirchheimer (originally published 1939), there is a gap of some three decades until a reawakening of punishment and society scholarship.

During the high tide of scientific penology, most social scientific reflection on the social role of punishment and the causes of penal change was concentrated on the technical goal of testing and advancing methods of treatment presumed and required by the success of penal welfarism in the decades following World War II.

It is the crisis of this model, marked by prison disorders and outside political protests against the prison system, that sparked renewed interest in the Durkheimian and Marxist approaches. It also led to the development of a third approach by two independent scholarly projects, both inspired in large part by the political protests on behalf of prisoners and against the modern correctional establishments in France (Foucault, 1977) and Italy (Melossi and Pavarini, 1982). Both projects took the work of Rusche and Kirchheimer as a starting point, but taking political economy as a horizon, both independently examined the prison in its relationship to the disciplinary technologies of power, whose rise both in state agencies and in private business establishments was crucial to the successful organizing and protection of capitalism.

Prisons emerged from a now forgotten period of experimentation in punishment at the end of the eighteenth century, because they fulfill the judicial function of delivering painful (or at least distasteful) treatment on a mass scale without unduly disrupting the coordination demanded by capitalism and democracy.

Responding to Foucault (1977) and Melossi and Pavarini (1982), punishment and society scholars in the 1980s began a vigorous debate over whether rupture three, the development of twentieth-century practices, like parole, probation, and juvenile justice, should be seen primarily as an extension of the disciplinary logic of the prison (which Foucault suggested at the end of his book) or a radical transformation (Bottoms, 1983; Mathiessen, 1983; Shearing and Stenning, 2003; Cohen, 1985; Feeley and Simon, 1992). An alternative is to see rupture three as reflecting a fundamentally new technology of power, one associated with social and medical diagnosis, specialized treatment institutions, as well as social insurance and welfarism and what some sociologists of governance have called 'the social' (Donzelot, 1979; Rose, 1999).

The rise of mass incarceration at the end of the twentieth century, in the US and in some other countries, may reflect the influence of a new political technology, that of containment or waste management. These new techniques, coming from fields like private security, operations research, and cybernetics, emphasize incapacitation and control (Deleuze, 1992; Feeley and Simon, 1992; Garland, 2001; Young, 1999; Shalev, 2009).

Toolkit Four: Punishment and Civility

In his sweeping history of manners and state formation from the late Middle Ages through the eighteenth century, Norbert Elias (1978) laid the groundwork for a powerful account of increasing penal leniency (ruptures two and three). Observing what he called the 'arc of civility,' the spread of refined manners and diminishing violence that spreads from the aristocracy to the industrial working classes in a period of about 400 years, Elias theorized that growing centralized state control in the external world correlates to increasing internal psychological controls (and corresponding tensions and anxieties) within the 'civilized' human subject (Mennell, 1992). Building on Elias, historians and sociologists have generated substantial studies of the long-term development of punishment in Europe (Spierenburg, 1984, 1991, 2004; Garland, 1990; Pratt, 2007).

Drawing on Elias, Spierenburg (1984, 1991) has challenged the claims implicit in toolkit one (Rusche and Kirchheimer, 2003) and explicit in toolkit two (Foucault, 1977) – that penal change happens in a series of ruptures or sharp breaks with the past. Rather, according to this research, the decline of scaffold punishment and the rise of imprisonment were gradual changes reflecting slowly shifting social sentiments of citizens and especially elite actors rather than catastrophic failures of top–down control strategies (see also Masur, 1989; Gattrell, 1994).

John Pratt (2007) has applied Elias' theory to explicating the rise of mass incarceration, rupture four, as a 'de-civilizing' process. Pratt cogently argues that the civilizing process is not a one-way ratchet, and that the linkage between greater levels of state centralization of life and subjective pacification can unwind.

Although America has adopted more extreme punishments, it has not necessarily broken with the trends suggested by the civilizing process. David Garland's (2010) study of capital punishment in the United States during a period when, both in the scale of imprisonment and the persistence of capital punishment, the country seemed to diverge significantly from its peers suggests that the constraints of civility remain strongly at work. Rather than being an exception, the contemporary American death penalty, with its often excruciating search for certainty as to the degree of suffering experienced by prisoners during the execution, may exemplify the demands that the processes of state formation ultimately place on the exercise of state power.

Punishment and Legality

Punishment as a problem for modernizing societies has been a distinctly legal problem (Hart, 2012). The principle of legality can be traced through the history of Western penal practices since the end of the Middle Ages, and the progressive legalization of the power to punish has been a central finding of historical and sociological studies (Garland, 1990: p. 184; Weber, 1978: p. 37; Jacobs, 1977).

But punishment is a particularly fragile institution when it comes to legitimacy. The ways in which it can fail are multiple and its satisfactory completion often leaves many unsatisfied. This makes punishment among the most volatile and change-prone institutions of society. Indeed, the significant penal reforms in America's Early Republic period were preceded by mandates in states' new constitutions (Rothman, 1972; McLennan, 2008: pp. 21–22), while prison labor was authorized as the sole exception to the Thirteenth Amendment's prohibition of slavery (Lichtenstein, 1996; McLennan, 2008). More recently, conditions in super-maximum-security (supermax) prisons and secure housing units have been redesigned to hypercomply with Supreme Court ruling (Reiter, 2012).

For much of ruptures two, three, and four, national legislation and rights derived from national constitutions have been the crucial legal influence on penal practices. To these have been added in recent decades a body of human rights law

anchored in treaties as well as in a growing body of customary law. The prohibition on torture is considered a matter of customary international law obligatory on all state actors whether those under treaties, like the United Nations Convention Against Torture (1984). The Universal Declaration of Human Rights (1948) specifically provides that "No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment." The International Covenant of Civil and Political Rights provides that: "The penitentiary system shall comprise treatment of prisoners the essential aim of which shall be their reformation and social rehabilitation" and which limits the death penalty (in those countries that have not abolished it altogether) to the most serious crimes. Some regional human rights treaties, like the European Convention on Human Rights, offer a particularly strong version of the interaction human rights law and modern penal practices (Van Zyl Smit and Snacken, 2009).

These norms, which are anchored in the broad concept of 'human dignity' (Whitman, 2003), have helped shape an emerging 'pan European penal sensibility,' which treats imprisonment as a last resort, requires that prisoners be kept in conditions as closely corresponding to community life as possible, and requires that prisoners be prepared for eventual reintegration into the community (Van Zyl Smit and Snacken, 2009). One of the most significant research questions of our time is whether greater court recognition of the human dignity of prisoners will counterbalance the pressures on penal bureaucracy coming from punitive populism in many contemporary societies.

Toolkit Five: Punishment and Culture

Wielding Durkheim, Marx, or even Foucault to explicate a particular penal policy or event can be less useful than finer tools of cultural and organizational analysis. Cultural sociologist Philip Smith (2013) argued that the theories mentioned earlier are most useful in making sense of what might be thought of as the 'penal climate,' that is, stable long-term patterns in the practices and discourses that define punishment in society, but less useful in making sense of the short- and medium-term developments. He provocatively analogizes these more temporary developments to the 'weather' while relating large-scale shifts to the 'climate' ('climate is what you expect, weather is what happens'). These include tools of narrative analysis that allow the analyst to unpack the meanings coded in the innumerable penal stories and statements that punishment gives rise to (Smith, 2008; Brown, 2009; Garland, 2010). It also includes sociological analysis of distinctions within social fields (Bourdieu, 1980; Page, 2013). Another important line of punishment and society scholarship connects cultural sociology to issues of social structure and class hegemony (Hall et al., 1978).

Part III. New Problems and Horizons of Punishment and Society

Punishment and Migration

In almost all the advanced economies of the world today, non-citizens are the fastest growing segment of the prison population (De Giorgi, 2006; Calavita, 2005; Melossi, 2013). Frequently, this incarceration is not considered 'punishment' as a legal matter because the detention is ordered on the basis of a 'civil' judgment that the person is in the country without proper permission and for the purposes of determining his legal status and deporting him; it is not a criminal penalty for a conviction. As a practical matter, immigration detention facilities are equivalent to prisons in their restrictions on liberty and interference with normal family and work life, while often lacking the programs and legal protections accorded to prisoners.

Mass Supervision/New Banishments

The rise of imprisonment rates in some advanced societies, and in particular the US (rupture four), seemed particularly surprising against the trends of the 1960s and 1970s, when institutional populations, including prisons, were declining in many advanced economies including the US and the UK (Cohen, 1985). In its place, forms of social supervision that emerged at the turn of the twentieth century (rupture three) were expected to become a more substantial part of the overall penal effort and with growing significance for the lives of people under legal sentence. During mass incarceration, parole and probation populations grew commensurately with prison populations, and increasingly became contributors to the growing prison population rather than an alternative to it (Simon, 1993; Petersilia, 2003).

Today it appears that many advanced societies that have avoided radical increases of imprisonment have greatly expanded the scale and reach of supervision, now often enhanced with electronic monitoring (Robinson et al., 2013; Nellis, 2010). Countries like the US that have experienced mass incarceration and its rising costs now seem poised to increase the role and power of supervision sanctions; a shift that could mark a return to twentieth-century norms, or a new form of 'mass supervision.' Researchers in the US have also noted the rise of spatial sanctions aimed at managing the access of individuals perceived as risky to social spaces (Beckett and Herbert, 2010).

Punishment, Risk, and Prevention

At least since the birth of the prison, rupture two, penal practices have been forward looking, acting not just to expiate past bad acts but to correct future behavior (Foucault, 1977). The late twentieth-century surge in imprisonment reflected a highly pessimistic assessment of the future dangerousness of prisoners (and pessimism about rehabilitative interventions), leading to a 'new penology' based on managing permanently high-risk populations through imprisonment (Hannah-Moffat, 2013; Feeley and Simon, 1992; Garland, 2001; Hudson, 2003; but see Lynch, 1998; Hannah-Moffat, 2004).

Conclusion

Research on the social and legal aspects of punishment has grown from a series of provocative social theories about Western state formation to an interdisciplinary field of

punishment and society research (Simon and Sparks, 2013). The field was initially concerned with the political claims of prisoners and the implications of the penitentiary project for the claims of welfare liberalism in the 1960s and 1970s. The rise of mass incarceration has been the primary focus of researchers since the late 1990s. As prison populations come down in the US, the field has recruited a large number of well-trained interdisciplinary students and is primed for a comparative and historical look beyond the ruptures in Western penality that have been its primary focus.

See also: Death Penalty; Durkheimians, The; Human Rights, History of; Law, Sociology of; Mass Imprisonment and Its Consequences; Prison Life, Sociology of; Punishment: Comparative Politics of; Torture: Legal Perspectives.

Bibliography

Beccaria, C., 1996. Of Crime and Punishment (J. Grigson, Trans.). Marsilio Publishers, New York.

Beckett, K., Herbert, S., 2010. Banished: The New Social Control in Urban America. Oxford University press, New York.

Bourdieu, P., 1980. The Logic of Practice. Stanford University Press, Palo Alto, CA.

Bottoms, A., 1983. Neglected features of contemporary penal systems. In: Garland, D., Young, P. (Eds.), The Power to Punish: Contemporary Penality and Social Analysis. Humanities Press, Atlantic Highlands, NJ, pp. 166–202.

Bottom, A., 1995. The philosophy and politics of punishment and sentencing. In: Clarkson, C., Morgan, R. (Eds.), The Politics of Sentencing Reform. Clarendon Press, Oxford, UK.

Brown, M., 2009. The Culture of Punishment. NYU Press, New York.

Calavita, K., 2005. Immigrants at the Margins: Law, Race and Exclusion in Southern Europe. Cambridge University Press, New York.

Cohen, S., 1985. Visions of Social Control. Polity Press, Cambridge, UK.

De Giorgi, A., 2006. Re-thinking the Political Economy of Punishment: Perspectives on Post-Fordism and Penal Politics. Ashgate, Aldershot.

Deleuze, D., October 1992. Postscript on the Societies of Control 59, 3–7.

Donzelot, J., 1979. Policing the Family (R. Hurley, Trans.). Pantheon, New York.

Durkheim, E., 1964. The Division of Labor in Society (G. Simpson, Trans.). Free Press of Glencoe, New York.

Durkheim, E., 1969. Two laws of penal evolution, Deux lois l'evolution penale (Jeffrey, Jr. 1969, Trans.). Originally published in Anne Sociologique 4, 65–95. University of Cincinnati Law Review 38, 32–60.

Elias, N., 1978. The Civilizing Process. Blackwells, London.

Feeley, M., Simon, J., 1992. The new penology: notes on the emerging strategy of corrections and its implications. Criminology 30, 449–474.

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison (A. Sheridan, Trans.). Pantheon, New York.

Gachter, S., Renner, E., Sefton, M., December 5, 2008. The long run benefit of punishment. Science Magazine 322 (5907), 1510.

Gager, J., 2006. Curse tablets and binding spells in the Greco-Roman world. In: Wygant, Amy (Ed.), The Meaning of Magic from the Bible to Buffalo Bill. Bergan Books.

Garland, D., 1985. Punishment and Welfare. Gower, Brookfield, VT.

Garland, D., 1990. Punishment and Modern Society: A Study in Social Theory. University of Chicago Press, Chicago.

Garland, D., 2001. The Culture of Control: Crime and Social Order in Contemporary Society. University of Chicago Press, Chicago.

Garland, D., 2010. Peculiar Institution: America's Death Penalty in an Age of Abolition. Harvard University Press, Cambridge.

Garland, D., Young, P., 1983. Toward a social analysis of penality. In: Garland, D., Young, P. (Eds.), The Power to Punish: Contemporary Penality and Social Analysis. Heinemann, London.

Gattrell, V.A.C., 1994. The Hanging Tree: Execution and the English People 1770–1868. Oxford University Press, Oxford.

Gilmore, R., 2007. Golden Gulags: Prisons, Surplus, Crisis, and Opposition in Globalizing California. University of California Press, Berkeley.

Gottschalk, M., 2006. The Prison and the Gallows: The Politics of Mass Incarceration in America. Cambridge University Press.

Hall, S., Critcher, C., Jefferson, T., Clarke, J., Roberts, B., 1978. Policing the Crisis: Mugging, the State, and Law and Order. MacMillan, London.

Hannah-Moffat, K., 2004. Criminogenic need and the transformative risk subject: hybridizations of risk/need in penality. Punishment and Society 7, 29–51.

Hannah-Moffat, K., 2013. Punishment and risk. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 129–151.

Hart, H., 2012. The Concept of Law, third ed. Clarendon Press, Oxford.

Hay, D., Linebaugh, P., Rule, J.G., Thompson, E.P., Winslow, C., 1975. Albion's Fatal Tree: Crime and Society in Eighteenth-Century England. Pantheon, New York.

Hudson, B.A., 1996. Understanding Justice: An Introduction to Ideas, Perspectives and Controversies in Modern Penal Theory. Open University Press, Philadelphia, PA.

Hudson, B., 2003. Justice in the Risk Society: Challenging and Re-affirming 'Justice' in Late Modernity. Sage, London.

Jacobs, J., 1977. Stateville: The Penitentiary in Mass Society. University of Chicago Press, Chicago.

Kant, I., 1974. The Philosophy of Law (W. Hastie, Trans. 1877). Clifton, NY. Reprinted 1974, A.M. Kelley.

Lacey, N., 2013. Punishment, neoliberalism, and social democracy. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 260–280.

Langbein, J., 1994. The historical origins of the privilege against self incrimination at common law. Michigan Law Review 92, 1047–1085.

Lichtenstein, A., 1996. Twice the Work of Free Labor: The Political Economy of Free Labor in the New South. Verso, London.

Linebaugh, P., 2006. The London Hanged: Crime and Society in Eighteenth Century England, second ed. Verso, London.

Lynch, M., 1998. Waste managers? The new penology, crime fighting and parole agent identity. Law & Society Review 32 (4), 839–870.

Lynch, M., 2010. Sunbelt Justice: Arizona and the Transformation of American Punishment. Stanford University Press.

Masur, L., 1989. Rites of Execution. Oxford University Press, New York.

Mathiessen, T., 1983. The future of control systems – the case of Norway. In: Garland, D., Young, P. (Eds.), The Power to Punish: Contemporary Penality and Social Analysis. Humanities Press, Atlantic Highlands, NJ, pp. 130–145.

Melossi, D., 2013. Punishment and migration between Europe and the USA. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 416–431.

Melossi, D., Pavarini, M., 1982. The Prison and the Factory: Origins of the Penitentiary System (G. Cousin, Trans.). The Macmillan Press, Ltd, London.

Mennell, S., 1992. Norbert Elias: An Introduction. Blackwell, Oxford.

McLennan, R., 2008. The Crisis of Imprisonment: Protest, Politics, and the Making of the American Penal State, 1776–1941. Cambridge University Press.

Nelken, D., 2011. Comparative Criminal Justice and Globalization. Ashgate.

Nellis, M., 2010. Electronic monitoring: towards integration into offender management. In: McNeill, F., Raynor, P., Trotter, C. (Eds.), Offender Supervision: New Directions in Theory, Research and Practice. Willan, Cullumpton, pp. 509–533.

Nietzsche, F.W., 1994. On the Genealogy of Morals (C. Diethe, Trans.). Cambridge University Press, Cambridge, UK.

O'Malley, P., 1999. Volatile punishments: contemporary penality and neo-liberal government. Theoretical Criminology 3, 175–196.

Page, J., 2013. Punishment and the penal field. In: Simon, Jonathan, Sparks, Richard (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 152–166.

Page, J., 2011. The Toughest Beat: Politics, Punishment, and the Prison Officers Union in California. Oxford University Press, New York.

Petersilia, J., 2003. When Prisoners Come Home: Parole and Prisoner Reentry. Oxford University Press, New York.

Platt, A., 1972. The Child Savers, Revisited. University of Chicago Press, Chicago.

Pratt, J., 2007. Penal Populism. Routledge, London.

Reiter, K, 2012. The Most Restrictive Alternative: The Origins, Functions, Control, and Ethical Implications of the Supermax Prison, 1976–2010. Dissertation, UC Berkeley. Jurisprudence and Social Policy Program.

Robinson, G., McNeill, F., Maruna, S., 2013. Punishment in society: the improbable persistence of probation and other community sanctions and measures. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 321–340.

Rose, N., 1999. The Powers of Freedom: Reframing Political Thought. Cambridge University Press, Cambridge.

Rothman, D.J., 1980. Conscience and Convenience: The Asylum and Its Alternatives in Progressive America. Little Brown, Boston.
Rothman, D.J., 1972. The Discovery of the Asylum: Social Order and Disorder in the New Republic, rev. ed. Little Brown, Boston.
Rotman, E., 1995. The failure of reform: United States, 1865–1965. In: Marris, N., Rothman, D. (Eds.), The Oxford History of the Prison: The Practice of Punishment in Western Society. Oxford University Press, New York, pp. 169–197.
Rubin, A., 2012. The unintended consequences of penal reform: a case study of penal transportation in eighteenth-century London. Law & Society Review 46 (4), 815–851.
Rusche, G., Kirchheimer, O., 1939, 2003. Punishment and Social Structure, rev. ed. Transaction Books.
Scheingold, S., 1984. The Politics of Law and Order: Street Crime and Public Policy. Longman, New York.
Scheingold, S., 1991. The Politics of Street Crime: Criminal Process and Cultural Obsession. Temple University Press, Philadelphia, PA.
Shalev, S., 2009. Supermax: Controlling Risk through Solitary Confinement. Willan, London.
Shearing, C., Stenning, P., 2003. From the panopticon to disney world: the development of discipline. In: McLaughlin, E., Muncie, J., Hughes, G. (Eds.), Criminological Perspectives: Essential Readings. Sage, London, pp. 335–349.
Simon, J., 1993. Poor Discipline: Parole and the Social Control of the Urban Underclass, 1890–1990. University of Chicago Press, Chicago.
Simon, J., 2007. Governing through Crime: How the War on Crime Transformed American Democracy and Created a Culture of Fear. Oxford University Press, New York.
Simon, J., Sparks, R., 2013. Punishment and society: the emergence of an academic field. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 1–20.
Smith, P., 2008. Punishment and Culture. University of Chicago Press, Chicago.
Smith, P., 2013. Punishment and meaning: the cultural sociological approach. In: Simon, J., Sparks, R. (Eds.), The Sage Handbook of Punishment and Society. Sage, London, pp. 114–128.
Simon, J., 2014. Mass Incarceration on Trial: A Remarkable Court Decision and the Future of Imprisonment. New Press, New York.
Spierenburg, P., 1984. The Spectacle of Suffering: Executions and the Evolution of Repression from a Pre-Industrial Metropolis to the European Experience. Cambridge University Press, Cambridge, UK.
Spierenburg, P., 1991. The Prison Experience: Disciplinary Institutions and Their Inmates in Early Modern Europe. Rutgers University Press, New Brunswick, NJ.
Spierenburg, P., 2004. Written in Blood: Fatal Attraction in Enlightenment Amsterdam. Ohio State University Press, Columbus, OH.
Swidler, Anne, 1986. Culture in action: symbols and strategies. American Sociological Review 51 (2), 273–286.
Thompson, E.P., 1975. Whigs and Hunters: The Origin of the Black Act. Pantheon, New York.
Van Zyl Smit, D., Snacken, S., 2009. Principles of European Prison Law and Policy. Oxford University Press, Oxford.
Vidmar, Neil, Miller, Dale, Spring, 1980. Social psychological processes underlying attitudes toward legal punishment. Law & Society Review 14 (3), 565–602. Contemporary Issues in Law and Social Science.
Wacquant, L., 2009. Punishment the Poor: The Neoliberal Government of Social Insecurity. Duke University Press, Durham.
Weber, M., 1978. In: Roth, G., Wittich, C. (Eds.), Economy and Society. University of California Press, Berkeley.
Western, B., 2006. Punishment and Inequality in America. Russell Sage, New York.
Western, B., Petit, B., 2004. Mass imprisonment and the life course: race and class inequality in U.S. incarceration. American Sociological Review 69 (2), 151–169.
Whitman, J., 2003. Harsh Justice: Criminal Punishment and the Widening Divide between America and Europe. Oxford University Press, New York.
Young, J., 1999. The Exclusive Society: Social Exclusion, Crime and Difference in Late Modernity. Sage, London.
Zimring, F., Hawkins, G., 1995. Incapacitation: Penal Confinement and the Restraint of Crime. Oxford, New York.

Cases, Statutes, and Treaties

http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx – International Covenant of Civil and Political Rights, 16 December 1966. United Nations General Assembly, Resolution 2200A. (site last visited on 30.12.13.).
http://hudoc.echr.coe.int/sites/fra/pages/search.aspx?i=001-122664#{"itemid":["001-122664"]} – Vinter and others vs The United Kingdom, Grand Chamber, European Court of Human Rights (Applications nos. 66069/09, 130/10 and 3896/10). (last accessed on 30.12.13.).

Putnam, Hilary (1926–)

Maria Baghramian, School of Philosophy, University College Dublin, Dublin, Ireland

Published by Elsevier Ltd.

Abstract

American philosopher, Hilary Whitehall Putnam (1926–) is one of the most influential analytic philosophers of our time. He is particularly well known for his contributions to philosophy of science, philosophy of mind, and philosophy of language. His work is also marked by a critical engagement with his own earlier views and several changes of key philosophical positions. This article distinguishes between early, middle, and late periods in Putnam's writing and discusses the most seminal ideas of each period. Particular attention is paid to Putnam's treatment of the question of realism over the three periods.

Life

Hilary Whitehall Putnam is one the most influential and prolific analytic philosophers of our time. He has written extensively on a wide array of philosophical topics but is particularly well known for his contributions to philosophy of science, philosophy of mind, and philosophy of language. Putnam was born in Chicago in 1926, but the family soon moved to Paris where his father, the eminent translator and literary editor Samuel Putnam, was to translate the complete works of Rabelais and edit the *Paris Review*. In 1933, the family returned to the United States and eventually settled in Philadelphia where Putnam lived until his graduation from the University of Pennsylvania in 1948, with a degree in Philosophy, Linguistics, and German. After a year in Harvard, Putnam moved to UCLA in 1949 to do a PhD under the supervision of the eminent logical positivist Hans Reichenbach. Reichenbach's profound influence led to a lifelong preoccupation with philosophy of physics and mathematics, and increasingly critical engagements with logical positivism.

Putnam's early academic career was at Northwestern (1951–52) and Princeton University (1953–61) followed by 4 years in MIT (1961–65), where he was Professor of Philosophy of Science and instrumental in setting up a new department of Philosophy. From 1965 until his retirement, Putnam has been a member of the Department of Philosophy in Harvard, first as Walter Beverly Pearson Professor of Modern Mathematics and Mathematical Logic and later as the Cogan University Professor. He is a Fellow of the American Academy of Arts and Sciences, a Corresponding Fellow of the British Academy, and was a president of the American Philosophical Association (Eastern Division). His numerous awards include the Rolf Schock Prize in Logic and Philosophy from the Royal Swedish Academy and The Ulysses Medal from University College Dublin, Ireland. He is married to the philosopher Ruth Anna Putnam.

Work

The sheer breadth and variety of Putnam's writing, accompanied by a radical practice of rejecting or reformulating his own earlier positions, make the task of summarizing and evaluating the full span of his work difficult, if not impossible. This article distinguishes between the early, middle, and late periods in Putnam's writing – what I call Putnam 1, 2, and 3 – and discusses the most seminal ideas of each period. Despite frequent changes of views and philosophical orientations, however, a number of prominent themes are evident in Putnam's writing, among them are the questions of realism and its discontents, a preoccupation with questions about the status of science, a continuous concern with questions of norms and values, and a general aversion to dichotomized thinking.

Early Putnam

Putnam's early publications have the imprint of his encounter with Rudolf Carnap, the most important of the logical positivists, whom he came to know during his Princeton years. Two topics were central. The first was on questions of synonymy and analyticity, the topic of W.V.O. Quine's seminal "Two Dogmas of Empiricism" (Quine, 1951), where Quine had argued, against Carnap, that the distinction between analytic and synthetic statements, or sentences true by virtue of their meaning only versus those true by virtue of facts, is untenable. Putnam's *The Analytic and the Synthetic* (1962) is where he establishes his own distinct philosophical voice. The article is also the first clear example of Putnam's aversion to philosophical dichotomies and beginnings of a lifelong critique of logical positivism. Putnam agrees with Carnap that there is an obvious and "gross distinction between 'All bachelors are unmarried' and 'There is a book on this table'" (1975b: 36). However, he is also in agreement with Quine that "'overworking the analytic–synthetic distinction' distorts our views of its significance for science or Philosophy" (1975b: 33). The distinction, he argues, could have a role provided that we do not treat it as a sharp dichotomy but as a continuum. There is a small group of concepts, he calls them 'one-criterion concepts,' which have a single criterion for application. For instance, the only criterion for the application of 'bachelor' is that of being an unmarried man and we cannot abandon this criterion without depriving 'bachelor' of its use, hence the analyticity of 'all bachelors are unmarried men.' A majority of concepts, on the other hand, are 'law cluster' concepts where "any one law can be abandoned without destroying the identity of the … concept involved" (Putnam, 1975b: 52). He thus

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61263-1

shows a way of thinking about the Quine–Carnap debate that retains the initial insight of the defenders of analyticity while avoiding its philosophical pitfalls.

The second area of focus in this early period was mathematics and its philosophy. Putnam's copublication of the proof of "Hilbert's Tenth Problem (1961)," in collaboration with Martin Davis, established his reputation as a mathematician. A collection of articles on philosophy of mathematics, coedited with Paul Benacerraf in 1964, was his first book-length publication. Putnam, in that early period, was a realist about both mathematics and science but his realism about mathematics piggybacks on realism about science. In what has become known as the "Quine–Putnam Indispensability Argument," Putnam, in similar lines of thought put forward by Quine (e.g., Quine, 1960), argues that much of science is either expressed in mathematical terms or uses mathematics as an indispensable component. And to the extent that scientific hypotheses are true, in the sense of corresponding with a mind-independent reality, then so are the mathematical theorems that underpin it. He says, "quantification over mathematical entities is indispensable for science, both formal and physical: therefore we should accept such quantification; but this commits us to accepting the existence of mathematical entities in question" (Putnam, 1972: 57). A realist interpretation of science, thus, supports realism about mathematics.

Science

Putnam's key idea on realism about science is known as the 'no miracle' or Putnam–Boyd argument to the best explanation. Using an abductive format and taking a similar approach to the indispensability argument for mathematics, Putnam and Boyd argued that the best explanation for the predictive and explanatory success of a scientific theory is its truth or approximation to truth. Their argument, in a nutshell, states that

1. Mature, that is, well-supported, and long-established scientific theories, result in successful predictions.
2. The best explanation for these predictive successes is that such theories are true or approximately true, i.e., they reflect what is going on in the natural world.
3. Nonrealist arguments, such as instrumentalism, cannot explain the success of science.

Therefore, mature scientific theories are (approximately) true.

Scientific realism is the claim that mature scientific theories are true.

Therefore, scientific realism is true.

Putnam explains:

> The positive argument for realism is that it is the only philosophy that does not make the success of science a miracle. That terms in mature scientific theories typically refer (this formulation is due to Richard Boyd), that the theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when they occurs in different theories – these statements are viewed not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate description of science and its relations to its object
>
> Putnam, 1975b: 73.

Mind

Putnam's other major contribution in this period was to philosophy of mind. The two dominant theories of the mind in the second half of twentieth century were logical behaviorism, which reduced the mind to dispositions to behavior and patterns of stimulus and response, and mind–brain identity theory claimed that mental states are identical with states of the central nervous system. Putnam, starting in the 1963 article "Brain and Behavior," republished in Putnam 1975a, argued that the view that a mental state, such as experiencing pain, can be reduced to certain types of behavioral states is untenable because it ignores the fact that pain is a private sensation. Identity theory is also defective because it disallows the possibility of ascribing mental states, such as pain, to creatures with a physical makeup different from that of humans. If pain is identical with certain neuronal activities in the human brain, then creatures with no such neurons would not be capable of experiencing pain.

Functionalism was Putnam's positive contribution to philosophy of mind. His version of the theory is known as Machine Functionalism because of the parallels he draws between mental states and the functional states of universal Turing machines or computers. A mental state is characterized by its causal relations with external stimuli, or input from the senses, behavioral responses, and relations with other mental states, e.g., beliefs and desires. Mental states are the causes of our actions within the background of our beliefs and desires about the world, but are not reducible to behavior. Even sensations, such as pain, could be explained functionally. Things that feel pain do so not because of the particular brain state they are in, but because they are in a certain functional state. One strength of Functionalism is that it ensures the autonomy of the mind and shows that "[t]he question of the autonomy of our mental life does not hinge on and has nothing to do with that all too popular… question about matter or soul-stuff. We could be made of Swiss cheese and it wouldn't matter" (Putnam, 1975a: 302). Just as different machines need not share the same hardware to carry out the same computation, mental states, including sensations and emotional states, could be functionally alike but physically different. What matters, Putnam argued, is functional organization, which is realizable in very different hardware, be it by silicon-based robots or by Martians with a different evolutionary history and biological makeup. The 'multiple realizability' of mental states became orthodoxy in cognitive science.

Decades later, Putnam would repudiate functionalism for reasons that are quite similar to his rejection of behaviorism and identity theory (Putnam, 1987b, 1992a). Earlier Putnam had rejected a reductionist account of the relationship between the mind and brain, he now warns against the reduction of mental states to machine states or brain software. In the same way that mental states cannot readily be identified or corelated with behavioral or brain states, mental and psychological states, such as hoping, desiring, fearing, etc., also cannot be corelated to software in a straightforward singular manner. In arguing against mind–brain identity theory, Putnam had maintained that mental states are multiply realizable because they are 'compositionally plastic.' He now believes that mental states are also 'computationally plastic' because a propositional attitude, such as a belief, can be realized by many different

computational states (Putnam, 1994: 47). The morale is that intentional mental states cannot be reduced to either physical or computational states.

Language

Putnam's contribution to philosophy of language has been momentous. For much of the twentieth century, philosophy of language was dominated by what is commonly known as 'descriptivism' – the approach to meaning adopted by the founders of analytic philosophy, Gottlob Frege and Bertrand Russell. According to descriptivism, names refer to what they stand for via what Frege calls 'sense' or conceptual content and what Russell calls 'definite descriptions.' For Russell, for instance, what defines 'gold' is a conjunction of properties such as its color, texture, weight, etc.; 'gold,' thus, is a shorthand term standing for a longer list of the properties we associate with that substance. Meaning, thus, is distinct from reference and is mediated by descriptions of that reference. Putnam and Saul Kripke, in his *Naming and Necessity* (1980), turned this position on its head.

In "The Meaning of 'Meaning,'" Putnam sets out to refute what he sees as a 'grotesquely mistaken' view of language (1975a: 21) arising from the tendency to ignore the role of the natural and social environment in establishing the meaning of words in general and names in particular. Two incompatible assumptions are at fault, he says. First, that to know the meaning of a term is to *be* in an appropriate psychological state and second, that meaning determines reference. Traditional theories of meaning are unsatisfactory because they are individualistic rather than social and neglect the contribution of external reality to meaning. The crucial point missed by descriptivism was that both the reference and meaning of natural kind terms, 'water,' 'gold,' 'brain,' for instance, are determined through interactions between these physical entities and referring minds, i.e., language users.

Putnam's argument proceeds through several thought experiments, the best known of which is the Twin Earth thought experiment, where he asks us to imagine a distant planet identical in most respects to our planet. Twin Earth has the same geographical features as earth – seas, lakes, mountains, etc., it has a climate similar to earth, and its inhabitants seem to speak the same languages that the earth people do. The one major difference with earth is that the stuff in their lakes and oceans, used for drinking, bathing, and swimming, does not have the chemical constitution H_2O but is XYZ. Except for having a different underlying chemical composition, XYZ resembles earth water in all respects, Moreover, the word used by Twin Earthians to refer to XYZ is 'water.' However, according to Putnam, when the inhabitant of the Twin Earth utter a sentence such as 'bring me a glass of water,' they are referring to XYZ, while when an earthling utters the same request she is talking about H_2O. So, the two seemingly identical sentences are about two very different things, and hence they have different meanings. The Twin Earth person is in effect talking about Twin Earth water or twater and not water.

According to Putnam, natural kind words have an "unnoticed indexical component: water is stuff that bears a certain similarity relation to the water around here" (1975a: 234). The reference of the term 'water,' has been fixed by pointing out to a stuff and saying something like "this is water" (Putnam, 1975a: 231). This initial baptism and the continuity of usage have establish causal connections between the speakers and what they are talking about. People on earth and Twin Earth may have the same kind of mental images running through their head when they think, respectively, of water and twater. Yet 'water' in English has a different meaning from 'water' in Twin Earth English. Thus, meaning is independent of the 'inner' mental states that people possess; rather, it depends on external, natural, and social factors. As he famously puts it "cut the pie any way you like, 'meanings' just ain't in the *head!*" (Putnam, 1975a: 227).

Not all language users would have knowledge of the underlying composition of natural kinds or always be able to always identify such kinds accurately. Putnam introduces the idea of division of linguistic labor to explain such cases (1975a: 227). Words, he argues, have producers and consumers, the experts and the ordinary users of language. The users often know only the stereotypical descriptions associated with a natural kind term and sometimes they get even those wrong. One role of stereotypes is to facilitate communication between speakers who ordinarily have very little knowledge of the underlying structure of natural kinds. To find out whether a particular object falls under a natural kind term users have to defer to the knowledge of the experts. For instance, ordinary speakers know gold through the stereotypical properties (shiny, yellowish, precious metal), but in order to authenticate whether a piece of metal is gold they defer to the superior knowledge of the relevant experts.

The Twin Earth thought experiment has become a paradigm of philosophical methodology in the analytic tradition and its use has been extended beyond the semantics of natural kind terms to ethics, philosophy of mind, and theories of content, despite many changes of position over the years, Putnam has remained constant in defending the view.

Middle Putnam

A preoccupation with what Putnam calls the "great question of realism" (Putnam, 1994: 295) or the question of how language connects with the world, is a unifying thread in Putnam's work, but the issue takes center stage in Putnam 2. The early Putnam was a realist not only about the existence of material objects but also about mathematical and purely theoretical entities such as numbers, fields, and physical magnitudes. The most striking feature of Putnam's work in the middle period is the rejection of some of the core assumptions of realism that he had previously embraced.

The emergence of Putnam 2 can be dated back to the presidential address he gave at the Eastern Division of the American Philosophical Association in December 1976, subsequently published as "Realism and Reason" (1977). Putnam used the occasion, rather dramatically, to repudiate his earlier views, a position he came to call 'metaphysical realism.' Metaphysical realism, he has argued, is committed to the following:

a. The world consists of a fixed totality of all objects.
b. The world has a fixed totality of properties.
c. There is a fixed set of 'language-independent' objects (some of which are abstract and others are concrete) and a fixed 'relation' between terms and their extensions.

d. There is a sharp distinction between properties we discover, the factual, and the properties we project onto the world (i.e., the evaluative).
e. There is a fixed relation of correspondence in terms of which truth can be defined (Putnam, 1981: 43 and 1999: 183).

Putnam offers several arguments to show the incoherence of metaphysical realism, the earliest and most controversial of which is the model theoretic argument. The argument was presented in three different formulations and, at least initially, relied on the *Löwenheim–Skolem* theorem in mathematics (Putnam, 1977). The most influential version is the permutation argument (Putnam, 1981) where it is claimed that it is possible to systematically permute the way singular terms and predicates refer without changing the truth values of the sentences containing these terms. The argument particularly targets the metaphysical realist's assumption that there is a fixed 'relation' between terms and their extensions (c above) and, that there is a fixed relation of correspondence between propositions and states of affairs (e above). Take any epistemically ideal theory, one that meets all operational and theoretical constraints possible, for instance it is consistent, complete, well formulated. According to the metaphysical realists, for such a theory to be true then it also should satisfy a unique word–world relationship, i.e., its terms should refer determinately to objects and entities (or classes of objects) in the world. However, Putnam contends that there could be no such thing as determinate reference as required by the realists. Any assignment of truth, by envisaging a relationship of reference, presupposes an interpretation of the referring language; the model theoretic argument shows that for any given language there could be many such interpretations and consequently many different ways of assigning (correct) reference. Consider Putnam's example of 'the cat is on the mat.' The sentence comes out true in an interpretation where 'cat' refers to the animal cat and 'mat' to an object we put on the floor, but it also comes out true if we use an alternative interpretation where 'cat' refers to leaves and 'mat' refers to trees. Thus metaphysical realism is shown to be based on an incoherent demand. It may be argued, using Putnam's own externalist theory of meaning, that 'cat' refers to cats not to leaves because it stands in the appropriate causal relation to cats and not to leaves. Putnam responds "How 'causes' can uniquely refer is as much of a puzzle as how 'cat' can on the metaphysical realist picture" (1978: 126). To think that there is metaphysical glue binding 'cat' to the animal cat is to subscribe to "a magical theory of reference" (1981: 47).

Putnam's second major argument against metaphysical realism is the famous brain in the vats argument (Putnam, 1981). The structure of the argument is quite simple. To believe that the world is mind-independent, as the metaphysical realist does, is to allow for the possibility of being radically mistaken in our beliefs about that world. Relying on a technologically updated retelling of Descartes' evil demon argument, Putnam proposes a matrix-type scenario, where the skeptic challenges us to prove that we are not brains kept alive by mad scientists dreaming of what we call 'reality.' He draws on arguments from semantic externalism to prove that such a scenario is incoherent and therefore so is metaphysical realism. As we saw, according to semantic externalism, the reference of words is at least partially determined by relations between language users and the natural and social world they inhabit. The word 'tree,' in a speaker's mouth, refers to the object tree because of a variety of social and physical connections with trees in the world. The same is true of the words 'brain,' 'vats,' etc. But if we were brains in a vat, we could not possibly be using the word 'brain' the way unenvatted people use for there are no actual connections between the brains in vats and their brains or vats. The words 'brain,' 'vat,' etc., as uttered or thought by brains in vats, would be connected only to neural impulses or images of brains and not to real brains in real-life vats. The skeptical hypothesis therefore is either false or cannot be coherently stated. But if the skeptical hypothesis is incoherent, then so is metaphysical realism.

Internal Realism

'Internal realism' is Putnam 2's proposed alternative to metaphysical realism, where, under the influence of Michael Dummett, in place of his earlier correspondence theory, he advocates an epistemic view of truth. Internal realism is committed to the following three core doctrines:

1. There is more than one true theory or description of the world.
2. There is no 'ready-made world.'
3. "truth is an *idealization* of rational acceptability – some sort of ideal coherence of our beliefs with each other and with our experiences *as those experiences are themselves represented in our belief system* – and not correspondence with mind-independent or discourse-independent 'state of affairs'" (Putnam, 1981: 49–50).

A crucial feature of internal realism was its commitment to an epistemic view of truth. Truth, Putnam 2 argued, is not independent from knowledge, but at the same time, it should not be identified with contingent and changeable epistemic achievements of a single time and a place, but with what we can know and justify under ideal epistemic conditions. The assumption is that if there were such things as epistemically ideal conditions, then statements that were justified under such conditions would be true. As Putnam puts it, truth is independent of justification here and now, but not independent of all justification (Putnam, 1981). This view aimed to deal with the problem of access by minded beings, such as us, to a mind-independent world. We cannot step outside our language, our mind-set or perspective and compare our beliefs about the world to the world as it is in itself. To give any substance to such a view we need to postulate a God's eye view of reality, a view not accessible to us mere mortals. The epistemic view of truth bypasses the need for a God's eye perspective while allowing for the possibility of objectivity in truth and knowledge.

The second important component of internal realism is conceptual relativity, the view that there could be alternative, noncompatible descriptions of objects and their properties. The key motivation for conceptual relativity is the conviction that the very *idea* of an unconceptualized world is incoherent. All our experiences of the world are shaped by our concepts, here too we cannot have a God's eye point of view, but only "the various points of view of actual persons reflecting various

interests and purposes that their descriptions and theories subserve" (Putnam, 1981: 50).

To show the force of the argument Putnam proposes a scenario where one and the same situation can be described as involving very different numbers and kinds of objects (Putnam, 1987a, 1990b). Let us take a world containing objects a, b, c. Our usual way of counting tells us that there are three objects in that world. However, according to 'mereology,' a theory of part/whole relationship first formalized by the Polish logician Stanislaw Leśniewski in the early twentieth century, for every two particulars there is an object that is their sum. Given this basic mereological starting point, the world just described will contain not three but seven objects. Putnam asks: Faced with a world with three individuals, does the question how many objects are there in this world have a determinate reply? The answer is no, because any reply would depend on how we interpret the word 'object.' From an atomist perspective, the perspective most of us take for granted, there would be three independent, unrelated objects a, b, c in this world. While, according to the mereological conceptual scheme the world will consist of seven objects, a, b, c, a + b, a + c, b + c, a + b + c. The example makes the internal realist's point that we cannot coherently talk of a world consisting of language-independent objects or properties for the very idea of object is mediated by the concepts we use.

In the late 1980s, Putnam reached the conclusion that internal realism, and its epistemic view of truth, in particular, faced problems similar to those besetting metaphysical realism and its concomitant correspondence theory of truth. The metaphysical realist needs to explain how access to the external world is possible, how the human mind could reach out, so to speak, and touch or represent a mind-independent world. Putnam came to believe that there was a similar problem in explaining our "referential access to 'sufficiently good epistemic conditions'"(Putnam, 1994: 462). The growing concern with this question led to the emergence of a third phase in Putnam's philosophical journey.

The Later Putnam

The emergence of Putnam 3 can be dated to his 1990 Gifford lectures in St. Andrews University. Ongoing change of position, however, particularly in comparison with the more radical break between Putnam 1 and 2, has been more nuanced and gradual. Putnam 3 defends what he variously calls 'common sense,' 'natural,' or 'pragmatic realism' and rejects both his early metaphysical realism and the later epistemic view of truth. While internal realism, at least in its inception, had a Kantian flavor; Putnam's work in the 1990s was influenced by Wittgenstein and by Pragmatism. The most crucial feature of this latest philosophical project is "a recovery of the natural realism of the common man" (Putnam, 1999: 24).

In earlier work, Putnam had treated sensations as the interface between the human mind and the external world. By early 1990s, however, he came to find this option unworkable – a realization that set the scene for the gradual emergence of his third reply to the big question of realism. Natural or Pragmatic realism emphasizes the role of our direct, unencumbered, contact with the world, unhampered by sense data or any other intermediaries. We can be natural realists because we perceive the world itself, and not the sense data caused by that world. Natural realism, Putnam claims, reflects our common sense "that 'external' things, cabbages and kings, can be *experienced*." (Putnam, 1999: 20).

One important feature of his thinking in this phase, and a carryover from Putnam 2, concerns the role and status of values in a world of facts. Putnam first questioned the intelligibility of a sharp division between factual statements and normative judgments, so popular among the logical positivists, in 1986. Some of his more recent publications (Putnam, 2002, 2004, 2012) deal in greater detail with the same theme. Putnam's aim is to dispel the wide-spread beliefs that while science deals with objective facts, moral values are purely subjective. Here, as in much of his writing in this phase, Putnam credits his views to classical Pragmatists such as William James and John Dewey.

Since David Hume's argument concerning the logical gap between 'is' and 'ought,' the fact/value division has been a cornerstone of naturalist or scientific worldview. Putnam argues that this orthodoxy is the wrong starting point for understanding both science and ethics. It is often claimed that scientific statements have a representational content; they represent states of affairs, while value judgments express our attitudes and emotions and give prescriptions and commands. For this reason, factual judgments are objectively true or false while ethical positions are projections and expressions of our subjective preferences. Putnam's key argument against this popular position is that the distinction is unsustainable because factual statements, as well as the practices of science, which enables us to decide what counts as a fact, are always value laden, in the sense that they presuppose some epistemic values (Putnam, 1981: 128). In other words, the very act of theorizing about the factual world is a value-laden activity. Here, once again, we see Putnam's desire "to break the grip that a certain picture has on our thinking; the picture of a dualism, a dichotomous division of our thought" (Putnam, 1987a: 71).

Conclusion

Philosophy for Hilary Putnam is an ongoing project involving a critical but generous engagement with numerous philosophical ideas and perspectives, including those of his older self. It is a never-ending attempt at renewing the subject matter and acknowledging our fallibility in the face of the formidable task of addressing some of the most fundamental questions facing us. The best summary of Putnam's approach to philosophy is by Paul Franks, one of his many students to have become prominent philosophers in their own right. Franks says, "Putnam is not best understood in terms of his 'position' or 'positions' but in terms of the dialectical tensions and complexity that expresses themselves through his positions" (quoted in Putnam, 2012: 2). This article has been a summary of both his various influential positions and the dialectical relations between them.

See also: Dewey, John (1859–1952); Ethics and Values; Functional Explanation: Philosophical Aspects; Hume, David (1711–76); James, William (1842–1910); Logical Positivism and Logical Empiricism; Pragmatism: Philosophical Aspects; Reference and Representation: Philosophical Aspects.

Bibliography

Baghramian, M. (Ed.), 2012. Reading Putnam. Routledge, London/New York.

Baghramian, M. Hilary Putnam: Mind, Language and the World. Polity Press, London, in press.

Baghramian, M., Jorgensen, A., 2013. Quine, Kripke, Putnam: meaning, necessity and intuitions. In: Beaney, M. (Ed.), Oxford Handbook of the History of Analytic Philosophy. Oxford University Press, Oxford.

Ben-Menahem, Y., 2005. Hilary Putnam. Cambridge University Press, Cambridge.

Boolos, G. (Ed.), 1990. Meaning and Method: Essays in Honour of Hilary Putnam. Cambridge University Press, Cambridge.

Clark, P., Hale, B. (Eds.), 1994. Reading Putnam. Blackwell, Oxford.

Davis, M., Putnam, H., Robinson, J., 1961. The decision problem for exponential diophantine equations. Annals of Mathematics 74 (3), 425–436.

De Gaynesford, M., 2004. Hilary Putnam. Acumen Publishing, Chesham, UK.

Kripke, S., 1980. Naming and Necessity. Harvard University Press, Cambridge, MA.

Norris, C., 2002. Hilary Putnam: Realism, Reason and the Uses of Uncertainty. Manchester University Press, Manchester.

Pessin, A., Goldberg, S., 1996. The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'". M.E. Sharpe, New York.

Putnam, H., 1962. The analytic and the synthetic. In: Herbert Feigl, Grover Maxwell (Eds.), Scientific Explanation, Space, and Time. Minnesota Studies in the Philosophy of Science, vol. 3. University of Minnesota Press, Minneapolis, pp. 358–397.

Putnam, H., 1963. Brains and behavior. In: Butler, R.J. (Ed.), Analytical Philosophy, Second Series. Basil Blackwell, Oxford, pp. 211–235.

Putnam, H., 1972. Philosophy of Logic. Harper and Row, 1971, New York. George Allen and Unwin, London.

Putnam, H., 1975a. Mind, Language and Reality: Philosophical Papers. Cambridge University Press, pp. 33–69.

Putnam, H., 1975b. Mathematics, Matter and Method. Harvard University Press, Cambridge.

Putnam, H., 1977. Realism and reason. Proceedings and Addresses of the American Philosophical Association 50 (6), 483–498.

Putnam, H., 1978. Meaning and the Moral Sciences. Routledge & Kegan Paul, Boston.

Putnam, H., 1981. Reason, Truth, and History. Cambridge University Press.

Putnam, H., 1983. Realism and Reason. Cambridge University Press, Cambridge.

Putnam, H., 1987a. The Many Faces of Realism. Open Court, La Salle, IL.

Putnam, H., 1987b. Representation and Reality. MIT Press, Cambridge, MA.

Putnam, H., 1951(1990a). The Meaning of the Concept of Probability in Application to Finite Sequences (Ph.D. dissertation). University of California, Los Angeles (1951). Garland, New York (1990).

Putnam, H., 1990b. Realism with a Human Face. Harvard University Press, Cambridge.

Putnam, H., 1992a. Renewing Philosophy. Harvard University Press, Cambridge, MA.

Putnam, H., 1992b. Pragmatism. An Open Question. Blackwell, Oxford.

Putnam, H., 1994. Words and Life. Harvard University Press, Cambridge.

Putnam, H., 1999. The Threefold Cord. Columbia University Press, New York.

Putnam, H., 2002. The Collapse of the Fact/Value Dichotomy and Other Essays. Harvard University Press, Cambridge, MA.

Putnam, H., 2004. Ethics without Ontology. Harvard University Press, Cambridge, MA.

Putnam, Hilary, 2012. In: De Caro, Mario, Macarthur, David (Eds.), Philosophy in an Age of Science: Physics, Mathematics and Skepticism. Harvard University Press, Cambridge, MA.

Putnam, H., Benacerraf, P. (Eds.), 1964. Philosophy of Mathematics: Selected Readings, second ed. Prentice-Hall, Englewood Cliffs, NJ. Cambridge University Press, Cambridge (1983).

Putnam, H., Walsh, V. (Eds.), 2012. The End of Value-Free Economics. Routledge, London, New York.

Quine, W.V., 1951. Two dogmas of empiricism. The Philosophical Review 60, 20–43.

Quine, W.V., 1960. Word and Object. M.I.T. Press, Cambridge.

Zeglen, U., Conant, J. (Eds.), 2002. Hilary Putnam: Pragmatism and Realism. Routledge, London & New York.

Q

Qualitative Methodology in Developmental Psychology

Carolin Demuth, Aalborg University, Aalborg, Denmark
Günter Mey, University of Applied Sciences, Hochschule Magdeburg-Stendal, Hansestadt Stendal, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Qualitative methodology presently is gaining increasing recognition in developmental psychology. Although the founders of developmental psychology to a large extent already used qualitative procedures, the field was long dominated by a (post) positivistic quantitative paradigm. The increasing recognition of the sociocultural embeddedness of human development, and of the importance to study individuals' subjective experience, however, calls for adequate methodological procedures that allow for the study of processes of transformation across the life span. The wide range of established procedures in qualitative research offers a promising avenue to advance the field in this direction.

Qualitative research is composed of many different procedures to collect and analyze data. In addition to some general common premises, these procedures are based on distinct theoretical approaches (e.g., social constructionism, ethnomethodolgy, hermeneutics, symbolic interactionism, phenomenology, psychoanalysis) to give access to the social, cultural, and psychological realities of the individuals or groups of individuals (e.g., for an overview, see Flick et al., 2004). A number of handbooks of qualitative research exist in the field of psychology (Camic et al., 2003; Forrester, 2010; Lyons and Coyle, 2007; Smith, 2008; Willig, 2008; Willig and Stainton-Rogers, 2007) and several chapters are dedicated to qualitative research in the recent American Psychological Association handbook of research methods in psychology (Cooper, 2012).

This article outlines how qualitative methods can be applied fruitfully – in particular in the field of developmental psychology – by adopting a life-span perspective and considering different contexts of development and different fields of research. To do this, we first give an overview of the history of qualitative research in developmental psychology and introduce some general shared premises of qualitative methodology; we next present an overview of established procedures suitable for collecting and analyzing qualitative data in developmental research.

The New Interest in Qualitative Developmental Research

Qualitative methods are presently gaining increasing recognition in developmental psychology. This development can be related to the following reasons: for decades, the field – like psychology in general – was dominated by a positivistic and postpositivistic paradigm and accordingly by quantitative methods, such as experiments, surveys, and objective testing. The focus was primarily on age-related cognitive competencies and their effects on some aspects of social and nonsocial behavior with the aim of identifying general laws of development. Underlying this focus was an understanding of development as a series of universal and natural ('biologically based') processes within the child that are largely independent from historical and cultural contexts. These ontological and epistemological assumptions have been challenged within recent years by a paradigm shift toward an interactional and sociocultural understanding of human psychological functioning and in line with a social constructivist epistemology. Human development now is acknowledged to be inseparably and dialogically intertwined with individuals' participation in mundane everyday social interaction. Through repeated participation in social interaction individuals learn to make sense of their experience and gain an understanding of the world. Hence, increasingly it is acknowledged that human development cannot be understood adequately without taking its cultural embeddedness seriously. Social interactions are always culturally prestructured and the interpretations offered always are constructed culturally versions of the world. This line of reasoning makes developmental psychology inseparable from cultural psychology that is interested in meaning-making processes that occur in social interactions and everyday contexts. Psychological subjects, such as the mind and the self, are conceived of not as entities but as processes that are constituted, or made up, within specific social and cultural practice. Culture is seen as the human ability to draw on interpretative procedures to make sense of one's experience in social interaction, which in turn is embedded in a specific structure of social organization of the society in which someone lives.

Awareness is increasing that mainstream developmental psychology has failed to adequately describe and understand children's ordinary lives and their active participation in their social worlds. Rather than treating children as 'objects' of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23156-5

research, a stronger focus on the study of their subjective experience is required (Greene and Hogan, 2005; Burman, 2008; Mey, 2010; *see* Street Children: Cultural Concerns). This critique arose with the increasing recognition that individuals are actively acting and potentially self-reflexive subjects and constructors of their own life. Thirty years ago, Urie Bronfenbrenner (1977: p. 513) got at the heart of this critique with his well-known statement: "it can be said that much of developmental psychology, as it now exists, is the science of strange behavior of children in strange situations with strange adults for the briefest possible periods of time."

Both a dialogical understanding of human development and culture and a recognition of the importance of individual's subjective experience require rethinking the adequacy of prevailing methodological standards. In fact, the very object of developmental psychology – processes of change with age in the psychological functioning of individuals – requires a discrete research program based on a methodology and methods that allow for the study of these developmental processes. Much of the twenty-first century's developmental psychology's standardized methods, however, still are borrowed from the realm of nondevelopmental psychology and hence largely inadequate to capture developmental processes (Valsiner, 2000).

Accordingly, methodological procedures within a qualitative paradigm that aim at reconstructing developmental conditions and contexts, life stories, and individual experiences from the subjective perspective of the participants, as well as sociocultural coconstruction processes of meaning-making, increasingly are called for (Mey, 2010; Valsiner, 2000; Greene and Hogan, 2005). Although neighbor disciplines like sociology, educational science, linguistics, and anthropology have well-established research procedures to study individuals' perspectives and everyday lives across diverse age-groups, mainstream developmental psychology, especially in North America where most developmental research takes place, has long been unaware of these challenges and methodological consequences. Recently, however, the field has seen an emerging period of critical self-reflection and change and a new reorientation toward qualitative methods.

Pioneers of Qualitative Developmental Psychology

Developmental psychology, in fact, often has benefited from qualitative research: Several of the founding fathers and mothers of the discipline applied distinct interpretative procedures in their work, albeit not in such a systematic way as would be the standard for qualitative methods. They provided many arguments for the establishment and the development of a qualitative developmental psychology, although they never used the label qualitative themselves (for a further discussion of the historical development, see Mey, 2010).

These early works were in line with the main stances of qualitative methodology: They followed a knowledge-producing and theory-generating logic (rather than testing existing theories and ex ante formulated hypotheses). They developed theoretical stances strictly from the perspective of the individuals; they tried to avoid artificial settings and to do research in the everyday context of the individuals; and they reflected the process of production and construction of the data, mostly detailed and in a way of thick description using only a small number of cases.

Taking this into account, it can be stated that some of the most influential developmental theories are derived from qualitative research (e.g., Erikson's theory of psychosocial development, Piaget's theory of cognitive development, Kohlberg's stages of moral development). Although the work is quoted widely in the field, little is mentioned about the methodological approaches they used to come to their conclusions.

Other examples of early ethnography of the children's concrete everyday lives with the aim of setting a counterbalance to the prevailing mainstream in the discipline at the time are the diary studies of their own children by Darwin and Preyer at the end of the nineteenth, and by Clara and William Stern at thebeginning of the twentieth century. They kept a separate record for each of their three children that allowed for individual case studies. Only in a second step was the data discussed in relation to a broader theoretical context. In William Stern's approach – the theory of personalism – a qualitative research style was embedded from the very beginning, a fact that often remains unmentioned, while at the same time, Stern often is acknowledged as the one who coined the term 'intelligence quotient'.

Similarly, the ethnographical work on life space of urban children by Martha Muchow, a research associate at Stern's Hamburg Institute is an example of the groundbreaking work in the first decades of the last century. Muchow used different open-ended methods like participant observation, conversations, methods based on essays next to the analyses of documents, and standardized methods. She acknowledged that to understand the particular nature of children requires one to understand the difference between adults (also researchers) and children, a position that also is crucial for contemporary research in the interdisciplinary field of children and childhood studies and children's rights (e.g., see Burman, 2008; Greene and Hogan, 2005). Kurt Lewin's renowned film of a young child learning to sit on a stone that demonstrates the barriers and field forces at play is another example of a radical departure from orthodox research methodology that arose from a dissatisfaction with the apparent triviality of much experimental psychology at the time.

Jean Piaget – probably the most prominent developmental psychologist of the twentieth century – postulated a clear qualitative perspective in his theory of genetic epistemology and his innovations of methods. His research can be described as a version of ethnography: For instance, in his clinical method, Piaget expanded the boundaries of the experimental approach to include context-specific tasks that can be applied to children in everyday life settings. He also considered standardized observations to be unsuitable for the study of child development. According to Piaget, traditional test procedures – such as intelligence test items – fail to address relevant questions, such as inquiry into children's spontaneous interests and their emerging entirely original understandings.

Similarly, Lev Vygotsky's reflections on (developmental) psychological hermeneutics – laid down in his work in 1930s – show how developmental psychology can be built only on qualitative grounds. His demand was that a psychology that wants to investigate 'whole objects' in their complexity requires analysis into units (rather than into elements) while at the same time preserving the whole objects in their complexity. Two case studies presented by Vygotsky's partner Alexandr

Luria – in the context of his 'romantic science' – in the second half of the twentieth century are brilliant examples of longitudinal studies that are informative for a qualitative developmental psychology. One of these case studies, described in his book *The Man with a Shattered World*, is about a man who suffered a left brain injury from shell splinters during World War II. The study that is based on more than 3000 pages of diaries and annotations by Luria captures the development of a man over a time span of 25 years. Erikson's biographical reconstructions – beside his identity theory based on his work as a psychoanalyst – of the lives of prominent persons like Luther or Gandhi, can be seen as movement in this direction.

To define and found a qualitative developmental psychology, including rethinking the theoretical and methodical assumptions, it is essential to remember one's own research tradition. From a twenty-first-century perspective, this early works reveals an understanding that already points to central features of qualitative research, such as wholeness and holistic methods, historicity, and reference to single-case studies, and it is linked closely to some more general premises of a qualitative approach outlined in the next section. Moreover, although early works of developmental psychology focused on early and middle childhood, the field has expanded to include human development over the entire life span.

General Methodological Premises of a Qualitative Developmental Psychology

Since the early days of developmental psychology, a number of systematic procedures have been established based on general premises of a genuinely qualitative paradigm that allow for empirically sound studies. Although these principles make clear that qualitative methods need to be understood on their own epistemological terms, they also can be fruitfully combined with quantitative methods. Within a qualitative paradigm, systematic integration of different procedures is understood as taking different perspectives on the investigated phenomenon to provide a holistic description of the phenomenon. This differs from triangulation approaches that are concerned with the mutual validation of findings or with cross-validation, as found in many mixed-methods studies that are oriented more toward a quantitative paradigm. A genuinely qualitative research style aims to overcome an abstract notion of development, as defined by the increase or decrease of specific abilities or by the transition from one hypostatized development stage to another. Such a style pursues a holistic view of abilities and functions within a developing person and between a developing and actively constructing person in a developing environment. To adequately access individuals' perception, description, and interpretation of reality, it is committed to the following three general and central principles of qualitative research (see also Flick et al., 2004).

The Principle of Openness: Individuals' Perspectives

Qualitative research generally follows an inductive-deductive and theory-generating logic. The guiding principle of openness aims to appropriately reconstruct (subjective) reality from the subjects' perspectives. The aim is to use this as a basis to (further) generate theories, to attain thick descriptions of developmental–psychological phenomena, and to understand the possible different development paths (that run parallel). Theoretical preconceptions by the researchers in qualitative research are understood to be 'sensitizing concepts' that can be used fruitfully in the analysis. This is in stark contrast to the traditional conception of development as more or less linear finality (that is defined ex ante from the perspective of the researcher).

The Postulate of Foreigness – Diversification and Development

The postulate of foreigness is linked closely to the principle of openness. This approach prohibits the researcher from (self-evidently) equating his or her own concepts (of development) with those of the culture, group, or person studied. This is an essential principle in cultural developmental psychology and childhood studies in which he difference between children and adults becomes evident when the 'culture of the child' meets the 'culture of the adult' (see Burman, 2008). A qualitative approach helps to prevent researchers from confusing their experiences and concepts with those of the children researched. At the same time, children and childhood studies and related areas of research pose an exceptional challenge for qualitative research as, until now, it has been much focused on language. Forms of expression typical for children can be of limited verbal nature and therefore it often is necessary to interpret preverbal articulation, gestural symbolizations, playful expressions, and the like. Research from the 'perspective of the children' or the 'aging people' is not possible without additional reflection on the researcher as a person, the way he or she is conducting the research, and consequently the conditions under which scientific knowledge is generated.

The Principle of Communication – Development as a Process

The relationship between the researcher and the researched also is reflected in the principle of communication. This principle is based on the assumption that all data are jointly constructed by those involved in the situation. Every research situation entails an intervention and hence a modification of the object of research. The principle of communication plays a pivotal role within process analysis and the development of process models, one main task of a qualitative developmental psychology, as it enables one to adequately conceptualize data as having a temporal dimension: that is, data in time that can be analyzed in a sequential and developmental–psychological manner. Any research situation hence needs to be understood not as a unit but as a sequence of events, taking place between those (researchers and researched) involved, that can be analyzed.

Qualitative Methods in Contemporary Developmental Psychology

As becomes clear from the overview of the historical roots of the field, a number of qualitative methods can be applied fruitfully to developmental psychology research in terms of studying processes and transformations across the life span. Moreover, the

number of established procedures in qualitative research is expanding (see Flick et al., 2004 for an overview). Qualitative research is a heterogenous field; however, besides the general premises they share, these methods need always to be understood within the respective theoretical and epistemological backgrounds in which they were developed. This implies that some methodologies in qualitative research need to be considered as approaches or research styles rather than as specific methodological procedures, which makes it difficult to allocate them exclusively as procedures for data assessment or data analysis. Overall, however, it is crucial that procedures of data assessment fit the procedures of data analysis with regard to their epistemological stance.

Procedures for Data Assessment

Interviews

Interviews are the most commonly used methods in qualitative research (for an overview, see Flick et al., 2004) and also are prominent in developmental psychology. Some early ambassadors of developmental psychology, like Mary Main or James Marcia, a scholar of Erik Erikson, already developed and popularized specific interview techniques in their line of research (on attachment and identity, respectively). Although many developmental researchers use semistructured interviews that often are closer to questionnaires than to a genuinely qualitative approach, a number of qualitative interview variants available have proved fruitful for developmental–psychological research. They range from special interview techniques that support the dialogue between interviewer and interviewees by using elements from the Rogerian psychotherapy approach to more one-sided and open narrative interviews. Specifics of conducting interviews with children have been discussed, for example, by Greene and Hogan (2005).

Written Data

Existing written documents are another form of data that can be used to gain insight into developmental processes. Collecting and analyzing diaries was, for instance, a prominent way to study the emotional life and thoughts of adolescents in early developmental psychology. Likewise, diaries can be used to study processes of transition within the family – for example, through the birth of a child and becoming a mother or father. Contrasting the perspectives of men and women can provide interesting insights. Studying diaries also can serve to investigate processes of coping with critical life events over time, like severe illness or mourning a beloved person. A modern version of this line of research is the study of Internet-based presentations (blogs, social networks, personal websites, or newsgroups – e.g., parentship or in regards to other developmental issues) also referred to as qualitative online research or 'netnography' (Kozinets, 2006). A common procedure in youth and childhood studies is the study of essays produced in the context of school exercises. An interesting aspect in this regard is how the constraints given in this context will contribute to the construction of the data. A final example of written data material used in qualitative developmental research is the completing sentences test – a procedure in which the participants are asked to complete a partial sentence without being given further restrictions.

Group Discussions and Focus Groups

The terms group discussion and focus group often are used interchangeably. Some scholars, however, relate the term group discussion to a more genuine understanding of qualitative research, whereas the term focus group then is used for structured (standardized) discussions with a group of participants (Bohnsack, 2004). Group discussions are of particular use when it comes to analyzing how opinions, attitudes, and orientations emerge, constitute, influence, and modify each other in an exchange of views – in a phrase, when it comes to the genesis of (developmental) processes. As such, they are of particular value for developmental psychology. Possible areas of inquiry are the study of patterns of argumentation with respect to specific ways of reasoning, such as moral dilemmas or social rules.

Furthermore, the focus on discursive, argumentative, and communicative contexts opens up additional possibilities to study negotiation processes in age-homogenous and -heterogeneous groups. Likewise, the study of the meaning of (self) socialization in peer cultures is a particularly prominent example of how group discussions do provide an extraordinary access to developmental processes. Group discussions in general provide an opportunity to investigate the genesis of knowledge. Piaget already had convincingly demonstrated how children's thinking developed after a quarrel took place between children and how during a quarrel or discussion functional moments emerge that initiate the development of self-reflection.

Qualitative Experiments

Qualitative experiments often remain unnoticed in the methodological literature. Disregarding qualitative experiments may be due to the fact that for many the combination of the experiment (believed to be the epitome of nomothetic science) and qualitative research (and its implications and aims) seems impossible.

In developmental psychology it was Piaget who recognized the value of experiments within a qualitative design. With his clinical method, which he understood to be an experimental method in the broader sense, Piaget hoped to counterbalance some of the shortcomings of rigid experimental designs. In a similar vein, Bronfenbrenner (1977) in his harsh criticism of experimental methods with regard to ecological validity pleaded for not using experiments as a verification tool but rather as a heuristic strategy.

Visual Data

The value of nonverbal and visual data and particularly of children's drawings, films, photos, or other artifacts created during play and other mundane activities can be considered as a valuable source for inquiry. Participants, for example, might be invited to photograph their social and physical environments (like kindergarten, school environments, houses for elderly, etc.) and to explore them with techniques interesting for them. In addition to analyzing the product (e.g., photographs, drawings), analysis of the process of producing (e.g., photographing, drawing) allows one to gain insight into the way individuals construct person–environment relations.

Contrary to photographs, videography and observation studies have gained far-reaching acceptance within developmental psychology (starting with the early work by Gesell or

Lewin). To a large extent, these studies draw on standardized designs and statistical analysis. From a qualitative research perspective, such an approach risks ignoring that records produced in such a way capture only reduced fragments of reality that furthermore are restricted to the static perspective of the camera and, in many respects, are inferior to the human capability for sensual and Gestalt perception. This perspective lends to the illusion that video recordings 'capture' reality – an impression that overlooks that they do not relieve from the duty of interpreting what has been 'captured'. A growing number of developmental psychologists, however, use videography and subscribe to the principles of qualitative research. Most of these studies can be found within cultural and discursive developmental psychology as well as in the field of language socialization (e.g., see the Center on the Everyday Lives of Families [CELF and ICELF]).

Participant Observation and Ethnography

Like qualitative research in general, ethnography does not constitute a single set of methods and procedures, but rather it is marked by systematic but flexible deployment of a variety of such methods as participant observation; videography; field notes; questionnaires; and open, semistructured, and focus group interviews (Jessor et al., 1996; see also Camic et al., 2003). An important common characteristic of ethnography is the sustained contact and involvement with a community that allows the researcher to become deeply engaged in the lives and mundane everyday practices of their participants (one difference to the early studies by people like Preyer's or the Stern's is that in this case, the researcher's involvement in the observed situation is reflected). The aim is to understand meaning from the perspective of the local participants within the context of mundane everyday life. It is this aspect that makes ethnography of vital importance for developmental psychology, as it permits a methodically comprehensive approach to studying human development where it actually takes place. Staying in the field, close to the participants (and their interactions), allows for detailed descriptions of different situations over longer periods of time and is an important strategy for collecting longitudinal data and studying processes of transition and change.

Finally, collecting and systematizing extensive field knowledge helps the researcher to generate adequate interpretations and 'thick descriptions' of a particular phenomenon – a term coined programmatically by Clifford Geertz (1973). Particularly with respect to the study of culturally distinct developmental pathways, ethnographic knowledge that only can be obtained through sustained fieldwork and engagement with the participants is indispensable. This might include, for example, knowledge about caregiver's ideas of good child care, their socialization goals, and an understanding of these goals in light of the prevailing sociocultural and sociohistorical living conditions.

Autoethnography

A procedure that has gained increasing attention in recent years with the raise of postmodern thinking in psychology is 'autoethnography' (Ellis, 2004). This approach to research and writing seeks to describe and systematically analyze personal experience in order to understand cultural experience. Methodologically, it combines characteristics of autobiography and ethnography and is an approach to research rather than a specific research method. Through systematic analysis of one's personal experience and by comparing and contrasting this experience against existing research, the researcher aims to identify how personal experience is made possible by being part of a culture or by possessing a particular cultural identity. For developmental psychologists, it offers insights into individual psychological transformation processes across the life span and does so in light of the person's sociocultural embeddedness – for example, the process of becoming a mother, the loss of a beloved person, or coping with illness.

Participatory Action Research

Participatory action research, like ethnography, constitutes a research approach rather than a specific research method and is based on an understanding that research should be conducted not only for but primarily with the people under study (Greene and Hogan, 2005). It commonly is related to the work of Kurt Lewin and has a long tradition within community psychology and social action projects. The outstanding characteristic of action research is its commitment to active, democratic participation, and empowerment of individuals under study as foundation for social change. Procedures range from naturalistic observation to interviews, case studies, surveys, and even experiments. Human psychological functioning is understood to be embedded in and systematically linked to various micro and macro levels of society as postulated in the ecological perspective of Uri Bronfenbrenner (e.g., 1977). This approach has a strong potential for applied developmental psychology, especially for developing intervention programs in many areas and across the life span – for example, to improve the quality of day care centers, youth delinquency prevention programs, preparation programs for parents, family-centered services, and residential homes for elderly. Action research in childhood studies explicitly promotes the participation of children and young people as active researchers to give them a voice and take them seriously as social actors in their own right. Hence, this approach raises the potential for children's self-development and political agency.

Procedures for Data Analyses

Reconstructing subjective experiences and coconstructive meaning-making processes requires adequate and open procedures not only for data assessment but also for data analysis. The material to be analyzed may vary from transcripts derived from interviews or group discussion, essays, observation protocols, diaries, photos, videos, and blogs, to media reports and other collected material. Depending on the specific research question and theoretical background, a number of established procedures are outlined in the following (see also Flick et al., 2004). They can broadly be categorized as coding procedures or and sequential procedures.

Qualitative Content Analysis

Qualitative content analysis is a distinct procedure from quantitative content analysis. This form of analysis is useful for analyzing large amounts of verbal data collected through interviews or focus groups and also offers possibilities for quantification of categories (Schreier, 2012). Its restriction to

a coding manual makes it less appropriate for reconstructing meaning-making processes and for highly open-ended research questions. Thus, it is less suitable for developmental–psychological research that is concerned with the temporal nature of phenomena that need to be reconstructed.

Grounded Theory Methodology

Grounded theory methodology (GTM) is considered to be one of the most popular methods of qualitative research. First formulated within the tradition of the Chicago School of Sociology by Glaser and Strauss and further developed by Strauss and Corbin, it has been developed in a number of ways, leading to a variety of existing versions of GTM (for an overview, see Bryant and Charmaz, 2010). This methodology allows for rigorous and systematic theory development that is grounded in the data. Although GTM envisages an explicit process perspective in its original sociological approach, a procedure elaborated from a genuine developmental–psychological perspective still needs to be developed. Similarly, the paradigm model proposed by Strauss and Corbin as a dynamic conditions-strategies-consequences matrix still needs to be transferred to developmental psychology realms of inquiry.

Interpretative Phenomenological Analysis

Psychological phenomenology recognizes that a whole set of factors leads to differences in people's perception of reality and is committed to the exploration of individual lived experience. Interpretative phenomenological analysis is composed of a series of sets of factors that aim to interpret the accounts of people of their personal lived experience and explicate an underlying structure in these accounts (see Smith and Dunworth, 2003). Its potential for developmental psychology lies in the focus on the subjective experience of individuals and the possibility to study developmental processes in detail and is suitable for longitudinal designs (e.g., identity change during the transition to motherhood).

Discursive Psychology and Discourse Analysis

In line with the recent sociocultural and interactional turn in the field, discourse analysis as developed within discursive psychology (e.g., see Camic et al., 2003; Willig, 2008) constitutes a particularly promising avenue to developmental psychology research. This procedure is oriented to examining social interaction as the site in which psychological phenomena that traditionally have been conceived of as developing inside the mind (attitudes, traits, knowledge, intentions, agency, emotions, identity) emerge in a process of intersubjective agreement. It treats language not as referential, that is, as a means to refer to these phenomena as hidden 'entities' in the isolated mind, but rather as constitutive of social reality and of the mind. Methodologically, it draws on procedures from conversational analysis and applies them to psychological inquiry. It increasingly is used in combination with videography, which allows for paralinguistic features to be included in the analysis. For developmental psychology, it offers possibilities to study early socialization processes in childhood (e.g., in mundane family or peer interactions) as well as locally situated identity construction across the life span (e.g., Bamberg, 2012; for a conversation analytical approach to developmental psychology, see Gardner and Forrester, 2010). Promising fields of discourse in analytical developmental research include the study of Internet-based communication like chats and blogs as they constitute a central arena for social interaction especially in young people's everyday life.

Other forms of discourse analysis also aim at somewhat-different research questions – for example, Foucauldian discourse analysis, critical discourse analysis, or genre analysis (e.g., see Flick et al., 2004; Willig, 2008 for an overview). Their focus is on institutional and social structures that frame specific discursive events and the ways in which ideological and political domination is reproduced in text and talk. Their relevance for developmental psychological research lies in studies concerned with cultural and historical ideologies of various aspects of human development, like childhood, parenting, or aging.

Metaphor Analysis

A more linguistic approach that could be applied fruitfully in future developmental–psychological research is metaphor analysis (Schmitt, 2000). The use of metaphor analysis is obvious in that the term 'development' can be understood as a metaphor itself (although mostly restricted to the path-metaphor; e.g., 'to forge ahead,' 'to stay behind') or can further enter concepts of education (e.g., 'to look after'/'to protect,' 'keep an eye on'). A metaphor analysis offers an exceptional potential for the reconstruction of the description of development from the perspective of the subject, for instance, as a deficit-oriented imagery in such terms as 'age degeneration' or 'loss of functioning' and their important role for understanding aging.

Narrative Psychology and Narrative Analysis

Narrative analysis is concerned with the structure, content, and function of stories in written and oral communication (Bamberg, 2012; Crossley, 2000; see also, Camic et al., 2003; Lyons and Coyle, 2007; Flick et al., 2004). It emerged from the narrative turn in psychology and the seminal work of Jerome Bruner more than 20 years ago. Another early forerunner of narrative analysis can be seen in Nelson's classic case study of a 2-year-old. Analyses of autobiographical life stories provide the opportunity to explicitly address temporality and the individual's reconstruction of developmental processes and are linked closely with the study of narrative identity. Social–constructivist approaches, however, stress the local and situated nature of identity construction and are more concerned with the analysis of 'small stories' in mundane everyday interaction (Bamberg, 2012).

Video Analysis

Video recordings allow one to capture the high complexity of mundane social interaction and accordingly offer a number of possibilities for qualitative analysis, depending on the specific research question and epistemological stance of the researcher (Heath et al., 2010). Videographed data, for instance, can be used partly in interview studies and group discussions as an additional source of information, or as part of larger ethnographic research (e.g., for the study of social interaction in kindergarten, at school, or within the family). Although technological advancements allow for increasingly sophisticated

analysis of video material, transcription of selected portions that will be analyzed in more detail usually is required after a substantive review of the data corpus. The most prominent procedures for qualitative video analysis in developmental psychology are discourse and conversation analysis. Embedded in a longitudinal and ethnographically informed design, it allows microanalytical analysis of change and transformation of interactional patterns over time and, hence, the study of developmental processes.

Performative Social Science

Performative social science refers to the deployment of different forms of artistic performance in the execution of a scientific project (Gergen and Gergen, 2010). Its aim is to offer a social constructionist alternative to traditional empiricist forms of communication in social science, such as theater, dance, fiction, poetry, or multimedia performances. It challenges the assumption that traditional forms of social science writing are the only suitable way to communicate scientific knowledge. Research findings, for example, might be presented in a poetic form or with visual and auditory expositions of data as a new way to disseminate and appropriate scientific research. This procedure strives for an interdisciplinary exchange between science and arts by creating new ways to make findings available and understandable to the participants under study.

Future Directions

The history of qualitative developmental psychology since its beginnings 120 years ago provides us with a variety of examples of how qualitative methods can advance our understanding of human developmental processes. Besides classic procedures like interviews and observations, recent technical advancements have lead to several new forms of collecting and analyzing data. A number of computer programs have been developed specifically for qualitative research (Aquad, Atlas.ti, Dedoose, HyperRESEARCH, MAXQDA, NVivo, QDA Miner, XSight, Qiqqa, Transana). Computer-Assisted Qualitative Data AnalysiS (CAQDAS) allows one to systematically retrieve, organize, and efficiently manage larger data sets and hence to expand the size of the samples. New approaches have emerged from various paradigmatic shifts within the past decades like the narrative turn, discursive turn, interactional and sociocultural turn, and more recently the performative turn embedded within the postmodern turn. Moreover, an increasing interdisciplinary exchange allows for adapting procedures originally developed in different disciplines for the specific demands of the field. In the twenty-first century, qualitative research in developmental psychology is a flourishing field that can be applied to studies of emotional, cognitive, social, and language development across all age-groups, developmental contexts, and cultural milieus.

Recent textbooks of developmental psychology not only acknowledge qualitative methods but also dedicate whole chapters to the topic (e.g., Smith and Dunworth, 2003). But have these methods really arrived in developmental psychology? Indeed, in a 1999 roundtable discussion, nearly two-thirds of editors of English-language mental health and child development journals argued in favor of publishing more empirical qualitative studies to, as they emphasized, initiate a necessary shift in this area of research (see Mey, 2010). Conversely, some of the most important mainstream journals in the field persist in opposing publication of genuinely qualitative research. Although interest has been increasing in mixed-method designs in the past years, the qualitative part of such designs often consists of open procedures rather than genuinely qualitative methods that are used within a quantitative paradigm. A systematic integration of genuinely qualitative and quantitative research designs seems to be difficult in light of the distinct epistemological underpinnings. If, however, we take one of the central criteria for the any empirical study seriously – the adequacy of the method for the phenomena under study – then the decision for or against specific methods needs to be derived from the research questions and the object of investigation, respectively. Newer methodological developments provide a number of opportunities to help the field study what it aims to study – that is, processes of change with age in the psychological functioning of individuals – in a systematic and more effective way. An adequately reflected methodological pluralism in light of the nature of human development as embedded in mundane everyday interaction and meaning-making processes within culturally prestructured patterns of society constitutes a promising avenue to advance the field.

See also: Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions; Mother-Child Communication: Cultural Differences.

Bibliography

Bamberg, Michael, Cooper, Harris, 2012. Narrative analysis. In: APA Handbook of Research Methods in Psychology, 3 Vols. APA Press, Washington, DC.

Bohnsack, Ralf, 2004. Group discussion and focus groups. In: Flick, Uwe, von Kardorff, Ernst, Steinke, Ines (Eds.), A Companion to Qualitative Research. Sage, London, pp. 214–221.

Bronfenbrenner, Urie, 1977. Toward an experimental ecology of human development. American Psychologist 32, 513–531.

Burman, Erica, 2008. Developmental psychology. In: Willig, Carla, Stainton-Rogers, Wendy (Eds.), The Sage Handbook of Qualitative Research in Psychology. Sage Publications, London, pp. 407–430.

Bryant, Antony, Charmaz, Kathy (Eds.), 2010. The Sage Handbook of Grounded Theory. Sage, London.

Camic, Paul M., Rhodes, Jean E., Yardley, Lucy (Eds.), 2003. Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design. American Psychological Association, Washington, DC.

Cooper, Harris (Ed.), 2012. APA Handbook of Research Methods in Psychology (3 Volumes). APA Press, Washington, DC.

Crossley, Michele L., 2000. Introducing Narrative Psychology: Self, Trauma and the Construction of Meaning. Open University Press, Buckingham.

Ellis, Carolyn, 2004. The Ethnographic I: A Methodological Novel about Autoethnography. AltaMira Press, Walnut Creek, CA.

Flick, Uwe, K., Ernst von, Steinke, I. (Eds.), 2004. A Companion to Qualitative Research. Sage, London.

Forrester, Michael A. (Ed.), 2010. Doing Qualitative Research in Psychology. A Practical Guide. Sage, London.

Gardner, Hilary, Forrester, Michael A., 2010. Analysing Interactions in Childhood: Insights from Conversation Analysis. Wiley-Blackwell, Chichester.

Geertz, Clifford, 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.
Gergen, M., Gergen, K., 2010. Performative social science and psychology. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 12 (1). Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs1101119 (accessed on 04.04.14.).
Greene, Sheila, Hogan, Diane (Eds.), 2005. Researching Children's Experience – Approaches and Methods. Sage, London.
Heath, Christian, Hindmarsh, Jon, Luff, Paul, 2010. Video in Qualitative Research. Analysing Social Interaction in Everyday Life. Sage, London.
Jessor, Richard, Colby, Anne, Shweder, Richard A. (Eds.), 1996. Ethnography and Human Development. Context and Meaning in Social Inquiry. The University of Chicago Press, London/Chicago.
Kozinets, Robert V., 2006. Netnography 2.0. In: Belk, Russell W. (Ed.), Handbook of Qualitative Research Methods in Marketing. Edward Elgar Publishing, Northampton, pp. 129–142.
Lyons, Evanthia, Coyle, Adrian (Eds.), 2007. Analyzing Qualitative Data in Psychology. Sage, London.
Mey, Günter, 2010. Qualitative developmental psychology. In: Toomela, Aaro, Valsiner, Jaan (Eds.), Methodological Thinking in Psychology: 60 Years Gone Astray? Information Age Publishers, Charlotte, NC, pp. 209–230.
Nelson, Katherine (Ed.), 1989. Narratives from the Crib. Harvard University Press, Harvard.
Schmitt, Rudolf, 2000. Notes towards the analysis of metaphor (10 paragraphs). Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 1(1), art. 20, http://nbn-resolving.de/urn:nbn:de:0114-fqs0001206; http://nbn-resolving.de/urn:nbn:de:0114-fqs0001206 (accessed on 04.04.14.).
Schreier, Margrit, 2012. Qualitative Content Analysis in Practice. Sage, London.
Smith, Jonathan A. (Ed.), 2008. Qualitive Psychology. A Pratical Guide to Research Methods, second ed. Sage, Los Angeles; London; New Delhi; Singapore.
Smith, Jonathan, Dunworth, Fraser, 2003. Qualitative methodology. In: Valsiner, Jaan, Connolly, Kevin (Eds.), Handbook of Developmental Psychology. Sage, London, pp. 603–621.
Valsiner, Jaan, 2000. Data as representations: contextualizing qualitative and quantitative research strategies. Social Science Information 39 (1), 99–113.
Willig, Carla, 2008. Introducing Qualitative Research in Psychology: Adventures in Theory and Method, second ed. Open University Press, Berkshire, UK.
Willig, Carla, Stainton-Rogers, Wendy (Eds.), 2007. The Sage Handbook of Qualitative Research in Psychology, Sage Publications.

Relevant Websites

www.qualitative-research.net – Forum: Qualitative Social Research.
http://www.qualitative-forschung.de/en/index.html – Qualitative reserch at Forschung.
http://www.listserv.uga.edu/archives/qualrs-l.html – The University of Georgia - Qualitative Research for the Human Sciences.
http://www.clarku.edu/∼mbamberg/ – Dr. Michael Bamberg - Clark University.
http://www.lboro.ac.uk/departments/socialsciences/research/groups/darg/ – Loughborough University - Discourse and Rhetoric Group.
http://www.celf.ucla.edu/pages/about.php (access date on 03.04.13.) – Center on Everyday Lives of Families.
http://icelf.weebly.com/english-version.html – ICELF - Italian Center on Everyday Lives of Families.
http://childrens-research-centre.open.ac.uk – The Open University - Children's Research Centre.
http://www.aaanet.org/sections/spa/?page_id=190 – Society for Psychological Anthropology.
https://email.rutgers.edu/mailman/listinfo/exploring_childhood_studies – Exploring Childhood studies - Rutgers University.
http://www.celf.ucla.edu/pages/about.php (accessed on 03.04.13.) – Center on the Everyday Lives of Families [CELF and ICELF].

Qualitative Methods, History of

John R Hall, University of California, Davis, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of qualitative methodological practices concerns development of research practices, objects of inquiry, and analytic strategies in relation to archival, observational, and interview data. There are three unevenly autonomous and connected streams: (1) historical inquiry and historical social science (especially historical sociology), (2) ethnographic research in anthropology, and (3) social science methodologies of interviewing and field research. Despite the diverse domains, qualitative methods have often converged in shared analytic issues, logics of comparison, and epistemological issues. Recently, qualitative methods have been pushed in two contradictory directions: toward increasingly powerful logical formalization and toward 'postmodern' emphasis on historicity and reflexivity.

This brief history traces the development of sociohistorical methodologies employed in relation to data that are either archival or based on direct social observation and interaction, insofar as these methods do not depend on quantitative analysis. The history of qualitative methodologies involves three unevenly autonomous and connected streams: (1) historical inquiry and historical social science (especially historical sociology), (2) ethnographic research in anthropology, and (3) social science methodologies of interviewing and field research. Despite the diverse domains and their differentiation, qualitative methods have often converged in shared analytic issues, for example, the relation between qualitative data and social theories and concepts, logics of comparison, epistemological issues concerning objectivity versus relativism, and questions of analytic methods of deduction versus induction. These developments can be explored in three respects: (1) institutionalization, (2) how objects of inquiry have been construed, and (3) logics and analytic strategies. Recently, debates in the wake of the 'postmodern' turn have created new possibilities for understanding qualitative research, and research more broadly, as a terrain of interdiscursive communication with properties of 'integrated disparity.'

Institutional Emergences of Qualitative Methods

Sociohistorical inquiry emerged long before the rise of the modern academic disciplines. In China, statecraft early on was informed by typologies of social circumstances and transitional possibilities of action specified by the *I Ching*. In the fifth century before the modern era, the ancient Greek Herodotus wrote as much as an observer of the contemporary as a historian. And during the fourteenth century, in the Arab world, Ibn Khaldun wrote in a way that anticipated comparative and historical sociology. These and other relatively sophisticated practices of inquiry coexisted with much more basic efforts of social groups to document their existences through the development of archives, king lists, annals, and chronicles. Thus, the arrangements through which sociohistorical knowledge has been produced are long-standing and diverse.

However, contemporary qualitative inquiry is especially a product of modernity, and of critical and postmodern reactions to the research practices spawned of modernity. Initially, qualitative research was inextricably linked with the gradual emergence of modern national societies and their colonial expansion, most concertedly during the nineteenth and twentieth centuries (Vermeulan and Roldán, 1995). Intellectually, the rise of social and humanistic disciplines can be traced to the European Enlightenment of the eighteenth century. Eighteenth-century thinkers such as Giambattista Vico and Johann Gottfried Herder increasingly began to characterize social and historical processes as subject to understanding in their own terms rather than as products of the will of God alone, and Edward Gibbon developed analytic history by use of concrete social explanations (i.e., in his *Decline and Fall of the Roman Empire*).

In the nineteenth century, history developed under two aligned tendencies: (1) efforts to differentiate the professional discipline from popular historical writing and (2) attempts to harness sociohistorical inquiry to the 'liberal' project of constituting the State, citizens, and the economy as a viable assemblage for social development. Thus, Leopold von Ranke sought to establish accounts of 'what really happened' on the basis of 'scientific' validation of facts drawn from archival sources, while simultaneously positing the history of political and religious elites as the subject matter that would reveal the meaning of history as a whole, that is, as a manifestation of the will of God. Historians in the United States for the most part embraced a modernist quest for 'objectivity' inspired by Ranke (Novick, 1988), while leaving the ideological and theological underpinnings of Ranke's scientific historicism largely uninspected. By contrast, the exemplary efforts of Max Weber to offer an alternative notwithstanding, historical sociology developed through the middle of the twentieth century mostly in relation to grand evolutionary social theories, at points aligned with anthropological ethnology, which was focused on constructing racialized world evolutionary sequences of social group development. Only after World War II, and especially in the 1960s, did historical sociology develop a methodologically sophisticated array of comparative and theoretical strategies for the analysis of history (Barnes, 1948; Smith, 1991).

As for field research, it emerged in relation to modern colonial expansion and capitalist industrial development. Among Europeans, especially from the sixteenth century

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44044-4

onward, knowledge gained from travel became a basis of colonial power sustained by exploratory expeditions, practical sciences, missionary work, and territorial penetration of colonial civil services. Sometimes authors combined traveling for a trading company or state with writing for a popular audience that wished to visit other lands vicariously. But epistemological principles of social knowledge were also at stake. Samuel Johnson's later eighteenth-century trip to western Scotland implicitly demonstrated that cultural differences would require forms of knowledge different from the then emerging rationalized social quantification (Poovey, 1998).

'Exploration' journals became increasingly systematic not only in describing the flora and fauna of the 'discovered' world but also in inventorying the languages and practices of non-Western peoples. Organizations such as the Smithsonian Institute (founded in 1846) coordinated efforts to gather data and circulated inventory questionnaires that could be filled out by nonprofessional travelers. Yet travelers' accounts – often saturated with racist and ethnocentric assumptions – largely served as grist for evolutionary ethnology and philosophical ('armchair') anthropology, for example, in Emile Durkheim's work on religion and James Frazer's *The Golden Bough*.

By the late nineteenth century, ethnology and philosophical anthropology still dominated the disciplinary space where modern ethnography was beginning to emerge. Lewis Henry Morgan's field study of the Iroquois, published in 1851, was largely a springboard to his ethnology. Only with the work of Franz Boas and Bronislaw Malinowski did inquiry begin to abandon the ethnological search for transgroup anthropological principles of development and emphasize analysis of social groups in their contemporaneous existences (a shift much bemoaned by the theorists). Boas was a crucial transitional figure. He initially intended his late-nineteenth-century observation and interviewing of indigenous peoples such as the Kwakiutl to advance his studies of perception and speech, to the point that he sometimes became impatient with ceremonial potlatches that wasted research time! Nevertheless, Boas pioneered efforts to create valid information based on direct, systematic field research, and he trained students such as Edward Sapir and Margaret Mead.

During and after World War I, Malinowski even more strongly embraced sustained participant observation oriented toward characterizing a group's present way of life. His Trobriand Islands fieldwork, published as *Argonauts of the Western Pacific*, proposed a methodology borrowed from sociology, placing empirical observations within a functionalist framework, yet it also consolidated the modern ethnographic role as that of the expert offering an authoritative account aimed at an emic understanding of the 'other's' point of view. This tension between an objective theoretical framework and emic analysis became a hallmark of modern anthropology, for example, in the French Durkheimian tradition, carried forward by Claude Lévi-Strauss (and in Britain by Alfred R. Radcliffe-Brown), and even in the way Clifford Geertz sometimes combined 'thick description' with social systems rhetoric.

Ethnographic study of the exotic 'other' was hardly the only impetus to institutionalization of field research. Famously, Alexis de Tocqueville voyaged to the United States in the 1830s to learn more about the American prison system. During the late nineteenth and early twentieth centuries, sociology was influenced by similar practical filed observational quests for knowledge about social conditions, for example, in studies of social issues sponsored by the *Verein für Socialpolitik* in Germany, by socialists in England, and among social workers such as those at Chicago's Hull House in the United States.

Building on the Hull House tradition, during the early twentieth century, sociology's strongest institutional development of qualitative methodology came at the 'Chicago school' founded by Robert Park, who took inspiration from Boas and Malinowski to encourage field research (DeVault, 2007). The city of Chicago became an urban laboratory for the direct observational study of social phenomena. Chicago faculty and students produced a stream of monographs such as *The Hobo* (1923) and *The Gold Coast and the Slum* (1929), and their approach found common cause with observational and in-depth interviewing field studies of communities, such as *Middletown* (1929) and the *Yankee City* studies (1941–59), work influenced by anthropological ethnographies (and sometimes carried out by anthropologists).

Beginning in the 1930s, American sociology underwent a broad disciplinary shift toward positivism (Steinmetz, 2007), the increasingly hegemonic position of which shaped the context in which qualitative field researchers operated. Only toward the end of the twentieth century did major positivist and quantitatively oriented departments such as the University of Wisconsin begin to reaffirm the importance of qualitative methods. However, that development occurred precisely because qualitative researchers maintained and developed diverse enterprises during the later half of the twentieth century. In the 1950s, Herbert Hyman formalized interviewing as a methodology, and Chicago students such as Howard S. Becker, Jr, took up the pragmatic and symbolic interactionist Chicago tradition that had been initiated by George Herbert Mead and developed by Everett Hughes. During the 1960s, phenomenological sociology, inspired by Alfred Schutz, and ethnomethodology, promoted by Harold Garfinkel, emerged as alternative approaches that problematized objectivity, realism, and the social activities through which qualitative knowledge is created. Added to this mix during the same decade were the constructivism of Peter Berger and Thomas Luckmann, the dramaturgical approach of Erving Goffman (himself trained at Chicago), and a nascent bridge between field research and history via the emerging interest in oral history. These developments, together with the 1960s retooling of sociohistorical inquiry at the hands of scholars such as Reinhard Bendix, E.P. Thompson, and Barrington Moore, set the stage for an efflorescence of qualitative research in the last decades of the twentieth century.

Yet, just when qualitative research flourished, it also underwent differentiation, along with sometimes fruitful methodological crisis. Some symbolic interactionists, ethnomethodologists, and comparative and historical sociologists sought to affirm the scientific legitimacy of their enterprises and gain acceptance among wider audiences by formalizing research methods and strategies, while other researchers took seriously the rejection of scientism consolidated from the 1960s onward by critical theorists such as Jürgen Habermas, hermeneuticists such as Paul Ricoeur and Charles Taylor, philosophers of history (most notably Hayden White), feminist epistemologists such as Dorothy Smith and Sandra

Harding, anthropologists like Marshall Sahlins and James Clifford, and the wave of poststructuralist and postmodern 'turns' heralded by Michel Foucault and Jacques Derrida. Some ethnographers became reflexive to the point of envisioning research itself as performative or eschewing study of any 'other' in favor of autoethnography. Ethical considerations about privacy, asymmetry of field research, and exploitation of subjects – raised especially by feminists and within anthropology – cast long shadows on the high modernist quest for objectivity and validity in qualitative analysis (for discussions of a variety of such issues, see Denzin and Giardina, 2007). During the past two decades of the century, these developments, along with the rise of world-systems and globalization theories, led to the formulation of new hybrid qualitative methodologies by such researchers as Michael Burawoy and Philip McMichael that linked the local and the global, ethnographic and historical analysis. The old dream of the ethnologists, of theorizing case studies in relation to general social processes, was reborn, but in a dispensation that rejected ethnological ethnocentrism. On a different front, historical and comparative sociologists have continued to reflect on their own recent history and how they might connect more effectively both with substantive topoi of sociology and with social theory (for a set of pivotal essays, see Adams et al., 2005). In general, then, research using qualitative methodologies has become strongly institutionalized within the social sciences, yet methodological practices and relationships with disciplines remain very much in flux.

The Object of Inquiry, a Brief History

One enduring problem of qualitative research concerns how to construe the object of inquiry, that is, what is studied. At the end of the eighteenth century, Immanuel Kant raised the philosophical problem of selection, namely, which of myriad events revealed the 'guiding thread' of history. In the nineteenth century, under the influence of teleologies such as those of Hegel and Ranke, historians were generally content to regard their studies as centered on 'natural' objects that had their own coherences in the world itself. However, as Hayden White (1973) has shown, the nineteenth-century European historians established rhetorical coherence through one or another genre of emplotment that offered a metanarratological structure into which to fit events. The alternative possibility, which developed outside of history (and increasingly within it), was to constitute the object of analysis via theory. Thus, Karl Marx and Frederick Engels theorized about capitalism as a dynamic totality changing through its own contradictions over time, and they used their perspective to thematize social events and processes in places as diverse as India, the United States, and Manchester, England. Within academia, the Marxian approach found its counterpart in the later nineteenth-century program of German historian Karl Lamprecht, who sought to displace Ranke's elites as the object of historical analysis with the study of diverse social groups and their interaction.

The Kantian problem of the relation between object and selection was more directly addressed during the German *Methodenstreit* toward the end of the nineteenth century, especially through the effort of Wilhelm Rickert to posit the objective value of truth in a way that would yield a shared way of looking at the world, and by extension, a shared object. Among mainstream historians during the twentieth century, a sort of ontological historicism emerged, in which shifts in styles (e.g., the emergence of social history) depended on shifting the object of historical analysis, rather than its methods. The major alternative to historicism, which Max Weber spelled out at the beginning of the twentieth century, depended on two central assertions:

1. that because values cannot be warranted scientifically, the objects of sociohistorical inquiry cannot be scientifically determined, even though, once objects are determined, scientific methods can be applied to their study; and
2. that the objects of sociohistorical inquiry are not natural kinds, but rather, infinitely varying sociohistorical phenomena that become objects of inquiry through cultural interest in them.

Alternatives to Weber's approach can be traced in history to the Annales School and Marxism, and in the social sciences, methodologically to the objectivism of Emile Durkheim, and analytically to Georg Simmel. Durkheim had provided the general warrant for objectivist approaches in his 1895 *Rules of Sociological Method*. And, in a way that is not widely recognized, Durkheim's structuralism had its equivalent in Simmel's early-twentieth-century identification of social 'forms' (such as 'sociability'), a central inspiration for a number of qualitative symbolic interactionists (Hall, 1999: Chapter 4). Within structural Marxism, the perspective of the proletariat was assumed to offer an objective standpoint for the analysis of capitalism, although during the first half of the twentieth century, this position was rendered problematic by the scholarship of Georg Lukács and critical theorists Max Horkheimer and Theodor Adorno. If the object of inquiry for structuralist Marxism was a totality, for the Annales School, in the formulation of Fernand Braudel, the totality was the grid of objective temporality, onto which multiple levels of change – ecological, structural, *histoire évènmentielle* – could all be mapped (Hall, 1980; Burke, 1990).

Ethnographic field researchers during the first half of the twentieth century generally treated their cases as naturally occurring 'primitive' societies subject to 'salvage anthropology.' Yet, asserting the supposed isolation of the natural case became increasingly untenable with the post-World War II emergence of first modernization and then world-systems-theory-approaches, both of which drew into question the autonomy of indigenous social groups. A parallel tension developed within sociological domains of field research. At the moment of high modernism, symbolic interactionists like Herbert Blumer would posit the empirical world 'out there' as the object of 'naturalistic research.' Yet, interactionists shared with phenomenologically oriented researchers and ethnomethodologists the sense that the empirical object of research was 'constructed,' that is, that the meaningful world was constituted by its participants' practices, which logically would include researchers themselves.

In the early twenty-first century, the objectivist–relativist divide in constituting objects of qualitative inquiry persisted, but the cultural and Foucauldian turns in history, the linguistic

turn in anthropology, increased understanding of how ethical and personal safety choices shape data from field research, and the challenges of analyzing social phenomena in relation to the internet all created conditions in which answers to the questions of how research is to be framed and 'what is a case?' had become considerably more nuanced and sophisticated (on the latter issue, see, e.g., essays in Ragin and Becker, 1992). In particular, by 1978, Arthur Stinchcombe argued against comparing cases as total entities one against another, and in favor of searching out the deeply analogous social processes that might be at work in cases that, on the face of it, might not even be considered as members of the same set. In a different vein, George Marcus has tracked the shift in anthropology from the deep study of a delimited ethnographic site to more integrated analyses of developments across disparate venues.

Overall, in today's postpositivist climate, reflexivity rules. Even objectivists and realists increasingly recognize that theoretical orientations and concepts have consequences for how cases are construed, while dialogic relativists, some of whom doubt the existence of any straightforwardly accessible empirical world to be studied, have had to come to grips with the problem of how adequately to 'represent' the (constructed) realities of their encounters with the people about whom they write.

Developments in Logics and Methodologies of Qualitative Analysis

Philosophers have debated the logics of human reasoning for millennia. But the most compelling specification directly relevant to modern sociohistorical inquiry came in John Stuart Mill's 1843 philosophy of scientific method, which specified two methods of nonexperimental research. The method of agreement entailed isolating crucial similarities in a set of diverse cases, yielding the claim that if two aspects co-occur under otherwise diverse circumstances, they are likely to be somehow connected with each other. On the other hand, the indirect method of difference extended the logic of control group experiments to nonexperimental settings, by searching for contingent corelated variations. In effect, two (or more) findings derived from parallel uses of the method of agreement could be placed side by side. If a particular propensity were found only in the presence of another particular propensity, and never when the second propensity was absent, Mill reasoned that one was "the effect, or the cause, or an indispensable part of the cause" of the other.

The application of Mill's logic to sociohistorical inquiry met with opposing visions of qualitative analysis that became crystallized by issues raised during the late-nineteenth-century *Methodenstreit*. Whereas the logic advanced by Mill yielded a nomothetic search for general principles, an alternative, idiographic view suggested that the study of human affairs necessarily dictates emphasis on the individually distinctive, indeed unique, aspects of social phenomena. Whatever the resolution to this question, a second issue concerned whether the special character of sociohistorical phenomena made them necessarily the subjects of a different kind of knowledge than the explanatory search for universal causal laws, namely, a cultural science based on understanding social meanings and actions. Three broad logics of qualitative methodology – interpretive strategies, ideal-type analysis, and formal analysis – can be traced in relation to these contestations.

First, early in the twentieth century, Wilhelm Dilthey proposed a historicist hermeneutics that would seek to understand history as embedded in the life meanings of the people who participate in it. This basic approach has its parallels in both ethnography and sociological field research. Elaborations of analytic logics were especially concentrated in sociology. In the 1940s, Alfred R. Lindesmith developed the method of 'analytic induction' as a procedure for successively refining hypotheses in relation to cycles of field research, an approach that Barney Glaser and Anselm Strauss subsequently reinterpreted as 'grounded theory.' And in the 1950s, a now-classic essay by Howard S. Becker reflected on 'problems of inference and proof' in the analysis of the substantial amounts of field notes and other documents. By the 1960s, 'participant observation' (Platt, 1983) and in-depth interviewing became increasingly institutionalized as legitimate approaches to inquiry, with their own epistemological rationales, challenges (notably how to construe and know subjectivity of the other person), research procedures, and handbooks, such as those of Buford Junker, Severyn Bruyn, and John Lofland and Lyn Lofland. Ultimately, epistemologically, as Isaac A. Reed (2011) demonstrated, the broadly interpretivist logic at the heart of much qualitative research has become a fruitful basis by which to probe the deep cultural structures of both positivist and realist approaches to inquiry.

Second, working at the same time as Dilthey, Max Weber consolidated a *Verstehende* approach that was both comparative and interpretive. He did not deny the possibility that sociohistorical inquiry might discover general regularities of the sort Mill sought to identify, but he regarded individualizing knowledge as the central value interest of sociohistorical inquiry. Therefore, he combined explanation and understanding, by using generalized ideal types that would be 'adequate on the level of meaning' and thus serve as meaningfully coherent analogs to the patterned character of empirical cases, yet, nevertheless, allow for comparison beyond individual particularities. Weber's methodology became central to the renaissance of historical sociology consolidated in the 1960s. In the 1990s, Geoffrey Hawthorn and David William Cohen, among others, developed Weber's logic of the 'mental experiment' into 'counterfactual analysis.'

Third, Mill's methodological effort became the touchstone for subsequent formalizations of qualitative logic. The most important elaboration has been the Qualitative Comparative Analysis pioneered by Charles Ragin, which builds on Mill's principles by using set theory and Boolean algebra to identify multiple configurational patterns within an array of systematically described cases.

In different lines of development, with the late twentieth century turn toward narrative and the study of discourse (partly inspired both by ethnomethodology and poststructuralism), scholars such as Larry Griffin theorized ways in which narrative accounts might become subjected to formal analysis that would yield descriptions of discrete social processes. And researchers more generally have begun to use one or another of various software programs for organizing and analyzing qualitative data – for example, Atlas.ti, Ethno, and NVivo. These

and related developments of formalization – as well as techniques of 'triangulation' that compared findings gained from diverse methodologies and the use of mixed and multiple methods more generally – increasingly blurred the lines dividing qualitative from quantitative analysis. With logical development, powerful new analytic techniques, and mixed methods came increased acceptance of qualitative research, but also critique. Three major recent developments stand out.

First, Stanley Lieberson (in Ragin and Becker, 1992) questioned whether formalization alone could resolve logical difficulties of comparative studies with small numbers of cases. For qualitative researchers, however, his critique may not be so devastating as would seem at first blush, because nested and intersecting comparisons and rich uses of data can be used to safeguard against the narrow errors of logic that Lieberson so effectively documented.

Second, a more convoluted set of issues arose a decade later in a review symposium in the *American Journal of Sociology*, where Loïc Wacquant (2002) critiqued the methodologies and logics of three widely read urban ethnographies – by Mitch Duneier, Elijah Anderson, and Katherine Newman. Responding in the symposium, the authors discussed by Wacquant characterized his readings of their work as tendentious and ideologically driven. The manifold (and often riveting) challenges and disagreements in this exchange are less important for general issues of qualitative methodology than the often shared bases of critique. Notably, all participants in the symposium treated morals, values, and ideology as both insidious in their capacity to underwrite reductionist metanarratives and important to understand – both in the subjects and reflexively – in the practices of research. The most serious methodological divides among symposium participants concern: (1) what credence should be extended to subjects' accounts versus explanation in relation to wider social contexts – revisiting a long-standing anthropological debate over emic versus etic ethnography; (2) whether to embrace theoretically driven or more grounded research (the latter position especially strongly defended by Anderson); and (3) whether, in order to enhance validity, the subjects of research ought to be guaranteed anonymity or, with their permission, have their identities revealed – and relatedly – offered the opportunity to read a researcher's accounts about them (practices championed by Duneier in both his book *Sidewalk* and his response to Wacquant). The symposium's *frisson* shows that enduring issues in qualitative methodology remain unresolved, and subject to alternative conventional resolutions. Thus, in relation to Dunieir's use of real names, although historians would not think twice about doing so in their studies, a sociologist criticized such a practice in a review of the author's 1987 sociological history of Jim Jones, *Peoples Temple*, and the 1978 mass suicide at Jonestown. On a different front, the divide between grounded and theoretically driven approaches is firmly ensconced in enduring alternative styles and logics of social inquiry.

Third, in *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables* and in a symposium on 'Varieties of evidence and method in cultural sociology' (Berezin, 2014), Richard Biernacki (2012) strongly criticizes research practices that code texts in relation to observer's categories. A number of approaches to content analysis used in qualitative research, he argues, efface meaning and yield analytically incoherent results. Biernacki's critique does not resolve the status of measurement in relation to meaning, but it sounds an alarm about an issue that qualitative researchers need to give serious attention.

The developments in logics and methodologies of qualitative analysis over the long term have not settled fundamental issues. However, they have brought increased sophistication to the practice of qualitative research. Moreover, they have increasingly raised issues of wider importance. Specifically, the Millsian logic of comparison, the relation between values and knowledge production, meaning and the problem of measurement, how to construe the character of theoretical concepts in relation to variables, cases, and social phenomena are issues that *quantitative* methodologists took very seriously some decades ago. But more recently, their focus has been more on improvement of statistical techniques. Because *qualitative* methodologists in recent decades have sought to confront these issues, their efforts now bear a salience for methodology more generally.

Postmodern Reflexivity in the Shared Domain of Qualitative Methods

By the end of the twentieth century, there had been substantial differentiation and development of qualitative methods, as well as experimentation in both research role relationships in field research and the form and rhetoric of texts reporting qualitative research more generally. Yet, these methodological shifts typically transpired within relatively autonomous schools where one or another approach was alternatively embraced, reinvented under a new flag, or syncretically revised in relation to alternatives. The differences between inductive and deductive historical sociologists, like those between mainstream symbolic interactionist field researchers and those who had taken the postmodern turn, often loomed as important as the disjuncture between quantitative and qualitative research more generally.

Under these conditions, a fruitful development has been the specification of typologies that describe alternative practices of inquiry in terms of the shared or contrasting cultural logics of their research methods, rather than assuming an equivalence between particular qualitative methodologies and various schools or genres (e.g., symbolic interactionism, cultural history). Theda Skocpol and Margaret Somers, and Charles Tilly developed typologies of historical and comparative sociology, and John R. Hall (1999) used discursive analysis to elaborate a general typology describing eight core cultural logics of research practice. In these terms, the three methodologies most favored in field research and historical sociology are historicism, contrast-oriented comparison, and analytic generalization. However, beyond these methodologies is a wider array of possibilities, including practices such as 'configurational history' that are often employed without benefit of formalization, and other practices, such as 'universal history,' that persist despite postmodern condemnation of 'metanarratives.' Typologies of qualitative analysis might seem to harden the boundaries between alternatives. However, the outcome of discursive analysis is the opposite, for it reveals

a condition of 'integrated disparity' among research practices that share cultural logics in ways not previously understood, and it clarifies the conditions and potential for translation of research knowledge produced through radically alternative research practices (Hall, 1999: Chapter 9). Thus, although qualitative research is shot through with contemporary tensions between positivism, realism, relativism, and constructivism, qualitative researchers may find that general recognition of sociohistorical inquiry as a reflexive enterprise will create ripe conditions for a new era of sophistication about qualitative research, with implications for research practices more widely. But to consider those possibilities moves beyond history into the realm of as-yet contingent possibility.

See also: Ethnomethodology, General; Field Observational Research in Anthropology and Sociology; Grounded Theory: Methodology and Theory Construction; Health Research, Qualitative; Interactionism, Symbolic; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Qualitative Methods in Geography; Science and Technology Studies, Ethnomethodology of; Symbolic Interaction: Methodology.

Bibliography

Adams, Julia, Clemens, Elisabeth S., Orloff, Ann S. (Eds.), 2005. Remaking Modernity: Politics, History, and Sociology. Duke University Press, Durham, NS.

Barnes, Harry E., 1948. Historical Sociology. Philosophical Library, New York.

Berezin, Mabel, symposium editor, 2014. Varieties of evidence and method in cultural sociology. Qualitative Sociology 37, forthcoming.

Biernacki, Richard, 2012. Reinventing Evidence in Social Inquiry: Decoding Facts and Variables. Palgrave Macmillan, New York.

Burke, Peter, 1990. The French Historical Revolution: The Annales School, 1929–89. Polity, Cambridge, UK.

Denzin, Norman K., Giardina, Michael D. (Eds.), 2007. Ethical Futures in Qualitative Research. Left Coast Press, Walnut Creek, CA.

DeVault, Marjorie L., 2007. Knowledge from the field. In: Calhoun, Craig (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 155–182.

Hall, John R., 1980. The time of history and the history of times. History and Theory 19, 113–131.

Hall, John R., 1999. Cultures of Inquiry: From Epistemology to Discourse in Sociohistorical Research. Cambridge University Press, Cambridge, UK.

Novick, Peter, 1988. That Noble Dream: The 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge, UK.

Platt, Jennifer, 1983. The development of the 'participant observation' method in sociology: origin myth and history. Journal of the History of the Behavioral Sciences 19, 379–393.

Poovey, Mary, 1998. A History of the Modern Fact. University of Chicago Press, Chicago, IL.

Ragin, Charles C., Becker, Howard S. (Eds.), 1992. What Is a Case?. Cambridge University Press, Cambridge, UK.

Reed, Isaac A., 2011. Interpretation and Social Knowledge: On the Use of Theory in the Human Sciences. University of Chicago Press, Chicago, IL.

Smith, Dennis, 1991. The Rise of Historical Sociology. Temple University Press, Philadelphia, PA.

Steinmetz, George, 2007. American sociology before and after World War II: the (temporary) settling of a disciplinary field. In: Calhoun, Craig (Ed.), Sociology in America: A History. University of Chicago Press, Chicago, pp. 314–366.

Vermeulan, Han F., Roldán, Arturo A. (Eds.), 1995. Fieldwork and Footnotes: Studies in the History of European Anthropology. Routledge, London.

Wacquant, Loïc, 2002. Scrutinizing the street: poverty, morality, and the pitfalls of urban ethnography. American Journal of Sociology 107, 1468–1532.

White, Hayden, 1973. Metahistory: The Historical Imagination in Nineteenth-Century Europe. Johns Hopkins University Press, Baltimore, MD.

Qualitative Methods in Geography

Isabel Dyck, Queen Mary University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Qualitative methodology is employed in human geography in a variety of ways, reflecting different topics, theoretical questions and philosophical approaches. This article traces the use of qualitative methods over time in the discipline, identifying key intellectual turns and ethical concerns that have influenced how and why particular methods are employed. Continued methodological innovation is observed as geographers explore the complex interconnectedness of human action, space and place.

Contemporary human geography is characterized by wide-ranging topic areas and theoretical discussion, accompanied by methodological plurality. A central concern with spatiality is constant, but a diversity of geographical imaginations (Gregory, 1994) is reflected in a continuing exploration of methodological approaches that may be used to investigate how 'geography matters' in the unfolding of social, economic, and political life. Among these, qualitative methodology provides an alternative paradigm of knowledge creation to that of quantitative techniques. With its admission of subjective experience and focus on meaning, qualitative methodology provides a variety of ways through which to analyze and interpret landscape, region, location, space, place, and the complex intersections of social and spatial relations. It has been critical to the ongoing reimagining of the discipline as shifting philosophical and theoretical debates are integrated into empirical inquiry.

A Paradigm of Inquiry and Collection of Methods

Rather than accepting the possibility of a unitary, objective explanation of observed phenomena, qualitative research approaches acknowledge and seek the multiple and partial versions of reality that are constituted in the course of social life. Embracing a variety of approaches and field methods, a common interest in qualitative research is to enhance understanding through description and concepts grounded in the perspective of those studied. The meanings brought to bear by individuals and groups in ordering and making sense of their everyday social life and its spatiality are privileged in analysis. Methods of inquiry include various ethnographic approaches, including participation in and observation of the worlds of those studied and nonstandardized forms of interviewing with individuals or groups; text and visual analysis; and a growing repertoire of other nonnumerical ways of observing, describing, and interpreting the world.

The shifting understandings of relationships between people and places and society and space are important in circumscribing how specific qualitative research strategies are employed. Interest in epistemological issues, particularly through the critical arm of the discipline, has fostered debate about the production of geographical knowledge and further shaped the choice of methods. There are close associations among theoretical developments, philosophies underlying inquiry, societal change, and the now well-established use of qualitative methods in geography. Such methods are used in various ways according to the questions and purpose of research seeking to understand complex sociospatial relations from different subdisciplinary and theoretical perspectives. They can provide descriptive data in mixed strategy designs to enhance findings from quantitative data, but more commonly are associated with postpositivist analyses that center the interpretation of meanings in explanation of narrative and visual data. Inductive analyses of narrative data, informed by theory, focus on generating concepts grounded in categories emerging from the data. Interpretation of various types of visual data, such as paintings, films, photographs, maps, and historical documents, similarly are sensitized by the theoretical and philosophical lenses of the 'reader.'

Qualitative Methodology in a Context of Philosophical and Theoretical Turns

Following the dominance of quantitative methods in the 1960s, intellectual turns have brought new questions and new topics to the agenda of geographical inquiry, together with a rethinking of appropriate methods. Geographers from a number of subdisciplines have explored qualitative approaches in their research as the respective influences of humanism, postmodernism, feminist scholarship, cultural studies, a new material turn, and an interest in emotional geographies have problematized conventional methods of inquiry and categories of analysis.

Humanistic geography, with its philosophical inspirations from idealism, phenomenology, and existentialism, first challenged the strong quantitative turn of the 1960s (Ley and Samuels, 1978). Interest in human agency, the inclusion of people's intentionality, a focus on lived experience, and the quest of understanding how meanings were produced in the context of interacting social and geographical worlds marked the value of qualitative research strategies. Participant observation studies produced ethnographic accounts that permitted a linking of the social with the geographical in ways that opened up inquiry to numerous topic areas that privileged the perspective of the human subject. Interpretation of landscapes included concerns with sense of place, while

case study approaches informed by a theoretical framework of social constructionism explored the embeddedness of human action and perceptions in place and provided models of inquiry for the 1980s. The qualitative methods text of Eyles and Smith (1988) signaled the establishment of qualitative research as an appropriate and useful paradigm of knowledge production within the discipline of geography.

Paralleling these developments, the influence of feminist scholarship resulted in further embracing of qualitative strategies in research. Noting the absence of women in geographical representations of the world, feminist geographers set out to include the neglected 'half' of the population in the mappings and narratives of the discipline. In making the lives and geographies of women visible, qualitative methods, as in other disciplines, were privileged in their ability to uncover the everyday lives of women and the spatialities of their domestic and paid work (McDowell, 1992). Bringing subjective realities of women into geographical explanation exposed the masculinist partiality of geographical inquiry and permitted the construction of concepts reflecting the geographies of women's lives. Linking principles of research with an emancipatory aim, qualitative approaches remain important in feminist methodology as a vehicle, together with the contribution of quantitative methods, in exploring the spatiality of gendered social relations (see Jones et al., 1997 for an example).

Postmodernism, the ongoing development of feminist theorizing and the later influence of cultural studies brought a 'cultural turn' to human geography (Massey et al., 1999). The importance of discourse in constructing human subjectivity and representing difference, whether reflected in the organization of space, social life, or in visual representations, was an important aspect of a complex literature on culture that influenced methodological approaches. Throughout human geography, qualitative methods enabled inquiry to reveal the lifeworlds of marginalized 'others,' to engage critically with various media as cultural representations, to expose the multiple, partial meanings of landscapes and texts, and to disrupt many taken-for-granted categories of analysis. Immigrants, the ill, the disabled, the colonized, children, the aged, and other groups on the margins of earlier inquiry emerge as 'others' whose subjectivities, emotions, and material geographies are included in a broadened geographical knowledge. Ethnographies of workplaces and interviews with elites, analysis of contemporary visual representations of social diversity, and the inclusion of historical texts that include women's experiential accounts are examples that further reflect a concern with competing knowledge and representations of the world. These journeys into the lives of others have led to an expansion and innovation in field techniques, including the use of life stories, focus groups, autobiographical writing, photography, film, and the adaptation of interviewing techniques and formats to meet the needs of particular research projects.

Constructing Knowledge through Qualitative Research: Methodological Issues

In tandem with the use of qualitative research strategies, geographers have engaged with a methodological literature in the social sciences discussing the difficulties that emerge when conducting research with the 'other' of studies. Reflection centers on the complexity of constructing knowledge through reliance on the intersubjective relationship between researcher and researched and the interpretation of data through the self as a socially positioned researcher who is part of the world under study. Issues of rigor in research design and practice have been addressed with attempts made to translate the concepts of quantitative research into those more amenable to qualitative methodology, but most discussion has focused on the close association between practical and theoretical dilemmas in research, and their location in epistemological concerns.

The particular attentiveness of feminist scholars to the micropolitics of research encounters, together with the acknowledgment of the situatedness and partiality of all experience and knowledge, have become the concerns of many qualitative researchers in the discipline. Gaining access to previously subjugated knowledge is shown to be a process located within a layering of power-laden social relations that situate the face-to-face research encounter within the wider politics and processes of 'the field,' theoretical and philosophical tensions, and the institutional values and processes of universities and funding bodies. Specificity of place is shown to mediate how this intricate layering plays out in the microdynamics of particular fieldwork situations.

Particular issues vary in different types of inquiry, but the exercise of reflexivity by researchers has problematized the whole enterprise of knowledge production. Whether in the context of reading landscape as text, rewriting colonial geographies, the demedicalization of medical geography, the shifting concerns of feminist inquiry, interviewing elites in economic geography, or exposing processes of exclusion of marginalized groups in society, researchers consider the different ways their own identities and contextuality influence their analyses.

That the findings of research are theory laden and interpreted through the researcher's subjectivity have brought attention to the presentation and writing up of research. The representation of the multiple voices of the researched is seen as a politically charged and theoretically tricky enterprise. Both in relation to narrative and visual data interpretation, the text produced and its claims to knowledge are problematized. The authority of the author is challenged, in that a particular representation is one of several possibilities. Following from this, the reading and consumption of research writing are viewed as crucial dimensions of the research process and the development of knowledge, rather than separate activities.

Attempts to disrupt hierarchical research relations, including issues concerning the politics and ethics of qualitative research, have remained at center stage in methodological debate. For whom research is conducted, how findings are used and related to theory production, and concern over the possible exploitative nature of the relationship with research participants are important considerations in research conduct. Exploration of participatory and emancipatory models of research, with different forms of collaboration between 'the researched' and 'researcher' reflect a thinking through how academia and activism may be effectively connected (Cahill, 2007; Sharp, 2005). The ethical issues involved in such attempts to 'empower' are a necessary and important focus of discussion. Visual methodology is attracting considerable

attention as the potential of visual methods in data collection are examined, whether in the form of photoelicitation interviewing, participatory video, or film (see, for example, Garrett, 2011; Oldrup and Carstensen, 2012; Ortega-Alcázar and Dyck, 2012; Rose, 2012). The connection made between visual methodology and social action, and the potential of different types of visual methods to level out power relations in research, make innovative use of the visual in research particularly appealing to those working with children, youth, and more vulnerable groups in society. In addition, research output in the form of film or photo-exhibition creates an accessible platform for discussion of research findings across a variety of audiences outside the conventional output of written scholarship.

Continuing Contributions

Expansion of the use of qualitative methods in geography is reflected in special issues of geographical journals and new texts on research methods that focus on qualitative methods or include qualitative approaches as important, rather than subordinate research strategies (Davies and Dwyer, 2007, 2008; DeLyser et al., 2010; Dwyer and Davies, 2010; Hay, 2012). Their value is recognized as theoretical developments have signaled the need to find new ways to discover and analyze data appropriate to current topics and questions. The various methods of qualitative research have opened up different ways of 'reading' and 'seeing' landscape, space, place, and the relationships between people and environments, so continually reshaping the nature and scope of geographical inquiry. The ongoing theorization of space, which has emphasized its relational, social, and recursive dimensions, has been accompanied by qualitative studies that are inclusive of human experience, perceptions, and meaning. Such studies have brought into view the diversity of local, everyday geographies, while at the same time locating individual and group experience within complexly linked local, national, and global relations.

The use of qualitative and quantitative methods is sometimes presented in terms of competing paradigms, sitting uncomfortably with each other. Theoretical turns and epistemological concerns may suggest that one be privileged over the other in the quest for valid knowledge in geographical inquiry, but different questions require different approaches. Both quantitative and qualitative works can and do contribute to furthering the understanding of spaces, places, locations, and how variously positioned individuals and groups live within and experience them. Qualitative geographical inquiry, however, has exposed the complexity of gaining access to and representing the situated knowledge of research participants in ways that preclude the development of rigid protocols or 'ways of doing' qualitative research. Different ways of telling, listening, and knowing require flexibility in research design and approaches. Ways of doing qualitative research need to be situated within and be sensitive to the many situations that comprise the academic geographer's field. It may be anticipated that further innovation of methods and flexible models will remain the hallmark of research concerned with the interconnections of people, place, and space.

See also: Phenomenology: Philosophical Aspects; Qualitative Methods, History of.

Bibliography

Cahill, C., 2007. The personal is political: developing new subjectivities through participatory action research. Gender, Place and Culture 14 (3), 267–292.
Davies, G., Dwyer, C., 2007. Qualitative methods: are you enchanted or alienated? Progress in Human Geography 31, 257–266.
Davies, G., Dwyer, C., 2008. Qualitative methods II: minding the gap. Progress in Human Geography 32, 399–406.
DeLyser, D., Herbert, S., Aitken, S., Crang, M., McDowell, L. (Eds.), 2010. The Sage Handbook of Qualitative Geography. Sage Publications, London.
Dwyer, C., Davies, G., 2010. Qualitative methods III: animating archives, artful interventions and online environments. Progress in Human Geography 34 (1), 88–97.
Eyles, J., Smith, D.M., 1988. Qualitative Methods in Human Geography. Blackwell, Oxford, UK.
Garrett, B.L., 2011. Videographic geographies: using digital video for geographic research. Progress in Human Geography 35 (4), 521–541.
Gregory, D., 1994. Geographical Imaginations. Blackwell, Cambridge, Oxford, UK.
Hay, I. (Ed.), 2012. Qualitative Research Methods in Geography. Oxford University Press, Oxford.
Jones III, J.P., Nast, H.J., Roberts, S.M. (Eds.), 1997. Thresholds in Feminist Geography: Difference, Methodology, Representation. Rowman and Littlefield, Lanham, MD.
Ley, D., Samuels, M.S., 1978. Humanistic Geography. Maaroufa Press, Chicago, IL.
Massey, D., Allen, J., Sarre, P. (Eds.), 1999. Human Geography Today. Polity Press, Cambridge, UK.
McDowell, L., 1992. Doing gender: feminism, feminists and research methods in human geography. Transactions of the Institute of British Geographers 17, 399–416.
Oldrup, H.H., Carstensen, T.A., 2012. Producing geographical knowledge through visual methods. Geografiska Annaler: Series B, Human Geography 94 (3), 223–237.
Ortega-Alcázar, I., Dyck, I., 2012. Migrant narratives of health and wellbeing: challenging 'othering' processes through photo-elicitation interviews. Critical Social Policy, themed issue 'Images and inequalities: implications for policy and research'. Critical Social Policy 32 (1), 106–125.
Rose, G., 2012. Visual Methodologies: An Introduction to the Interpretation of Visual Materials, third ed. Sage, London, UK and Thousand Oaks, CA.
Sharp, J., 2005. Geography and gender: feminist methodologies in collaboration and the field. Progress in Human Geography 29 (3), 304–309.

Qualitative Research

Joseph A Maxwell and L Earle Reybold, George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Qualitative research, characterized by substantially inductive and open-ended methods, is an important and accepted approach in the social sciences. Four specific features of qualitative research – understanding research participants' meanings, investigating the influence of the specific contexts in which the individuals and activities studied are situated, elucidating the processes by which these meanings and contexts lead to particular features or outcomes, and explicitly incorporating the subjectivity of the researcher – are central to qualitative research. These features shape the distinctive methods, analysis strategies, and approaches to validity and generalizability that give qualitative research its unique value for social research.

Qualitative researchers seek to understand lived experiences in context and the meanings associated with these experiences, usually from the perspective of participants. Qualitative data collection typically involves interviews, observations, and documents. Interview types include various options that range from personal, in-depth, active interviews to open-ended surveys or focus group interviews. Gubrium and Holstein (2002), in their edited volume, *Handbook of Interview Research: Context and Method*, offer a number of interview types and uses, but still critique the so-called 'interview society' that is pervasive in public life, not just in qualitative research. Likewise, observations can be exploratory, useful for documenting a setting and identifying contextual clues for future inquiry; or they can be used to compare participant perspectives with actual behavior. Finally, document collection, a secondary method in many disciplines, is of primary importance in historical studies. Documents include a variety of items: diaries, journals, photographs, census records, recordings, organizational memos, to name a few. The methods for qualitative data collection, as noted above, have a common goal of obtaining a detailed, chronological record of the events witnessed or the interviewee's statements. In actual use, these types of data overlap. For example, the use of video recording has provided an important additional way of recording interviews (individual and group), settings, and activities. Talk-aloud protocols blend interview with observation. Mobile interviewing connects the interview process to participant memories in real time. Finally, the Internet has provided many new ways of collecting data, from online interviews to accessing online discussions and social media.

What is most distinctive about qualitative research, however, is not data collection, but data analysis. In general, most qualitative data are recorded as transcripts and field notes or video, and analyzed using inductive, holistic, and narrative strategies rather than through quantification and statistical methods. Although numbers may be used in qualitative research (for example, to document how many individuals expressed a particular view or meaning, or the number of times a particular event occurred), qualitative analysis does not begin with converting the data to numerical form. There are three main types of analysis that are commonly used in qualitative research. First, in listening to and transcribing an interview, or in expanding rough field notes on an observation, researchers typically write about their reactions, interpretations, and ideas for further analysis. These writings, often called 'analytic memos' or simply 'memos,' are a form of analysis, recording the researcher's thinking about how to make sense of the data.

Second, the most prevalent form of qualitative analysis is coding, but the purpose of this is very different from quantitative coding. In the latter, the goal is to sort the data into mutually exclusive categories (typically established in advance) and counting the number of individuals, events, or responses in each category, in preparation for further statistical analysis. In qualitative research, in contrast, the main task is *developing* the analytical categories, and using them to generate themes, narratives, or even a 'grounded theory' of the phenomena. The latter was the major focus of the 'grounded theory' approach developed by Barney Glaser and Anselm Strauss – 'grounded' in the sense that the theory and the hypotheses being tested were not drawn primarily from existing theories, but were inductively generated from the data, and then tested and modified through additional data collection and analysis in a process of constant comparative analysis. More recent work in grounded theory has explored constructivist and postmodern applications of theorizing from data.

Finally, there is a less formalized and theorized form of analysis in qualitative research, one that is often implicit rather than explicit. This type of analysis, rather than fragmenting the data into categories across data sources, approaches the data more holistically, looking for connections *within* an interview, observation, or document. This is involved in many types of narrative analysis, discourse analysis, life history, and case studies, and is central to the goal of analyzing process, rather than simply providing a static description. Most qualitative studies involve all three of these forms of analysis, although the first and third may be less explicit.

The last 30 years have seen the development of a large number of computer programs (the common acronym for these is CAQDAS, for Computer-Assisted Qualitative Data Analysis Software) to facilitate the process of qualitative data analysis. Unlike statistical analysis programs, this software, although it has vastly increased in sophistication from the early forms, does not actually do the analysis; it simply facilitates the process of creating coding categories, assigning data segments to categories, and compiling all of the excerpts in a particular category for examination. The categories themselves, and the

interpretations of the categorized data, must still be created by the researcher. Although the more advanced programs can link memos to categories or data segments, and create links between categories (often called 'theory building'), their main advantage is to make the coding process easier and faster. (This can be a disadvantage if the research questions call for a primarily holistic, connecting analysis.)

A number of qualitative scholars have defined qualitative research in relation to quantitative methods. However, a precise definition of qualitative research and a clear distinction between qualitative and quantitative research remain elusive. The simplest definition, and one that is stated in some textbooks, is that qualitative research categorizes data using words (nominal- or categorical-scale measurement) and quantitative research uses numbers (ordinal- or interval-scale measurement). However, although the distinction between different scales of measurement is well established, this definition is obviously inadequate as a characterization of qualitative research, or for distinguishing qualitative from quantitative research. Qualitative researchers often use numbers in presenting their data and results, and a fixed-response survey in which the response categories are words would be seen by virtually all qualitative researchers as just as 'quantitative' as one in which the responses are numbers on a scale.

At the opposite and most complex end of the definitional scale, many qualitative researchers believe that the two approaches involve fundamentally different and irreconcilable 'paradigms,' with different philosophical assumptions about reality and how we can understand this, different goals and methods, and even different ethical premises. For example, some qualitative researchers claim that genuinely qualitative research requires a radical constructivist ontology and epistemology, one that denies the existence of any 'objective' reality; they assert that 'reality' is socially constructed by different individuals and societies, rather than existing independently of human constructions. The goal of qualitative research is seen as limited to understanding these different constructions and working to support those groups that are oppressed by the constructions of dominant groups.

This definition, which emerged from the 'paradigm wars' between qualitative and quantitative researchers during the 1970s and 1980s, also fails to stand up to scrutiny. Qualitative researchers differ widely in their philosophical assumptions; these include not only the constructivism described above, but also realism, pragmatism, postmodernism, critical approaches such as feminism and critical race theory, and many others. In addition, many qualitative researchers, particularly in applied fields, simply ignore the philosophical debates and focus on understanding their subject matter.

Because qualitative researchers usually are socialized to the method through a particular discipline, many adopt design 'types' more associated with their fields of study. Certain design types have gained popularity in specific disciplines because of the perceived alignment with foundational questions: What is the nature of reality? How does the nature of reality influence researcher access? How does this access impose a set of acceptable set of methods for exploring that reality? Some authors emphasize discrete design types that include case study, ethnograpy, phenomenology, grounded theory, and historical inquiry; the rationale is that these designs offer distinct 'recipes' for the conduct of qualitative research associated with certain approaches to inquiry. On the other hand, qualitative researchers often blend these 'recipes' to meet the needs of their specific research goals. Thus, researchers are expected to detail their decision-making process and actual methods used; in other words, qualitative researchers document the process of inquiry through an audit trail.

Given that qualitative research is not method dependent, it is tempting to argue that different forms of, and approaches to, qualitative research have what Wittgenstein called a 'family resemblance,' one that cannot be reduced to an explicit definition or distinction from quantitative research. However, there are a number of important conceptual features of qualitative approaches that are widely shared. Four of these features, mentioned above, are particularly salient for understanding qualitative research, and while not all studies will exhibit all of these, it is difficult to imagine a qualitative study that possesses none of them.

First, qualitative researchers typically focus on *meaning*, rather than simply behavior or 'attitudes.' By this term, qualitative researchers mean the beliefs, values, and perspectives held by the people studied, and as the people themselves understand these, rather than as conceptualized by the researcher. In the classic phrasing of the anthropologist Bronislaw Malinowski, a qualitative researcher seeks to understand things "from the natives' point of view," rather than by imposing prior theories or external frameworks. Doing this requires understanding what people say and do from their own perspective, and constructing what linguists and anthropologists call an 'emic' account, one grounded in the meanings that the actors themselves place on these actions and statements. This is what qualitative researchers typically mean by saying that qualitative research is interpretive – that it seeks to interpret the participants' own understandings. This does not preclude using prior theories and analytic frameworks, but it does require that the researchers not rely exclusively on these, and that they examine these theories critically, noting where they fail to capture the participants' understandings.

Second, qualitative researchers pay close attention to the *contexts* in which the individuals and events that they study occur, rather than simply attempting to generalize across contexts or individuals. Each person and situation is unique, and a qualitative study will address these differences and the effect that they have on the people and social groups studied. This is particularly true of the type of qualitative research known as case study research, in which the researcher seeks to understand a particular individual, group, or activity in its unique context, as a complex phenomenon that is intrinsically shaped by that context.

Third, qualitative researchers seek to understand *process* – not simply showing *that* the people, groups, or settings studied have certain features, but also *how* these features operate in the lives of these people and communities. As Robert Weiss described this, rather than seeking a correlation between two variables – for example, between the level of a wife's education and the presence of a companionable marriage – "[i]n qualitative studies we would look for a process through which the wife's

education or factors associated with her education express themselves in marital interaction."

Finally, qualitative researchers honor the role of *subjectivity*, sometimes defined in contrast to objectivity, across the research process. However, subjectivity is a distinct concept that acknowledges the place of personal history and experience on the part of the researcher in making decisions across the research process. These aspects of self – the researcher as instrument – encourage critical reading of qualitative findings and are recognized as a strength of certain qualitative designs. Related to this notion of subjectivity are the concepts of positionality and standpoint epistemology; these terms situate and critique qualitative inquiry through the lens of race, class, sex, gender, and other identity markers. Subjectivity is not bias; in fact, a deep awareness of one's own assumptions and focus strengthens the qualitative researcher's claim of quality. (This will be discussed in more detail in the section on validity and quality.)

These four features are shaped by a fundamental difference between quantitative and qualitative research. Quantitative researchers necessarily think about the things they study in terms of *variables* – properties of persons, settings, or events that can vary, and can thus be measured or counted, compared, and correlated. Qualitative researchers, in contrast, do not typically think in terms of variables, but in terms of entities, meanings, contexts, and processes. Their research questions, instead of focusing on correlation and comparison (*whether* there is a relationship, and its magnitude), typically address some or all of the first three of these issues – *how* and *why* things happen, how they are shaped by, and shape, their context, and what meaning these have for participants.

These features are responsible for many of the more obvious differences between qualitative and quantitative research. For example, because most of the specific, local meanings, contexts, and processes that qualitative researchers study are not well understood in advance, qualitative designs are typically inductive and flexible, allowing the researcher to modify the data collection and analysis methods during the research to better fit the particular situation and the insights that emerge from this. This flexibility is less problematic than in quantitative research, because the primary goal is not standardization for the purpose of comparison and generalization, but particular understanding.

Similarly, qualitative researchers typically study naturally existing settings, rather than contrived, experimental environments. This is because they are less interested in manipulating particular features of the environment to determine their effect on participants (i.e., treating these as variables) than in understanding these 'real-life' contexts holistically. Although some important qualitative studies have been done in experimental settings, these researchers have treated these settings *as* 'natural,' seeking to understand the meanings, actions, and processes that are found there.

Finally, a qualitative approach brings into question the goal of 'objectivity' in the sense of detachment from, and lack of influence on, the things studied. In conducting observations and interviews, a qualitative researcher is inevitably part of the situation studied, and cannot help influencing and being influenced by this, a phenomenon often termed 'reflexivity.' Rather than attempting to 'control for' this (i.e., to prevent it from being an 'extraneous variable' that may influence the relationships among other variables), qualitative researchers seek to *understand* this influence, both to prevent it from having unwanted, negative consequences, and to use it productively to investigate the phenomenon studied. This is most obvious in interviewing; the interviewer is intrinsically influencing the interviewee simply by asking questions, and the goal is not to 'standardize' the interview in order to prevent variation in the interviewer's behavior from affecting the variables studied, but to adapt to the particular interviewee and setting to best understand that person's views and perspective.

Qualitative research has a long and diverse history, arising from a variety of sources, including ethnographic research writing in anthropology since the seventeenth century (particularly that of Bronislaw Malinowski and Clifford Geertz), the work of Charles Booth, Max Weber, and the Chicago School in sociology, and the ideas of Wilhelm Wundt in psychology. Denzin and Lincoln divided the more recent history of qualitative research into seven 'moments,' emphasizing controversies and new approaches since 1950. However, most of their different 'moments' are clearly less 'stages' in the development of qualitative research than they are ongoing, diverse approaches to qualitative research; much qualitative research is still done using methods and assumptions developed before the current proliferation of approaches began.

This proliferation is perhaps the most significant feature of qualitative research at present – qualitative research includes a wide and expanding range of conceptions of what qualitative research is and can be. These include approaches that draw as much from the humanities as from the social sciences, including narrative inquiry, phenomenology, autoethnography, performance ethnography, queer theory, bricolage, and testimonio, as well as more traditional approaches such as grounded theory, case study, and participatory action research. It is beyond the scope of this article to attempt to characterize these varied strategies.

There are, however, some generalizations that can be made about qualitative methods, applying to many, but not necessarily all, approaches. First, qualitative researchers do not usually employ random or other probabilistic sampling methods, which are necessary for making statistical generalizations from a sample to the population sampled. Because their main goal is particular understanding rather than generalization, they typically use what is called 'purposeful' or 'theoretical' sampling, selecting individuals, cases, and settings that are best able to answer the researcher's questions or that are most relevant to the theory being developed. (See the later discussion of generalization in qualitative research.) In fact, many qualitative researchers avoid the use of the term 'sampling' because it connotes statistical representation of a broader population, not usually a goal of qualitative research. Instead, they use the term *selection* to remind the reader that the scope and boundaries of any qualitative study are determined by researcher choices.

The presentation of qualitative studies has traditionally been less formal and structured than in quantitative articles, more narrative in form and more open to personal details. However, this openness and diversity has increased substantially in recent work, along with the growing incorporation of humanistic and artistic forms of presentation, including performance as an alternative to written presentation of

a study's results. One such development has been the use of what has been termed 'poetic rendition'; this involves presenting the data (sometimes with some reorganization and editing) in poetic form, bringing out images and connections that might not be apparent in a prose account. Approaches such as autoethnography require researchers to include themselves as a source of data and an object of analysis. A number of analytic techniques are seeing a resurgence in recent scholarship, but incorporated through constructivist and critical theory lenses rather than the original structural perspectives, including dramaturgical analysis and portraiture. Newer forms of qualitative analysis such as meta-synthesis are gaining attention across disciplines, especially those in which pragmatic, somewhat generalizable findings are important (e.g., nursing and community health).

This proliferation of alternative types of qualitative research, and the prior polarization between qualitative and quantitative approaches that resulted from the 'paradigm wars,' led many qualitative researchers to question the traditional concepts of 'validity' and 'generalizability' as defined by quantitative researchers. Many qualitative researchers reject the concept of validity entirely, substituting terms such as 'trustworthiness' or 'authenticity' and substantially broadening the criteria used to assess these, including ethical issues such as whether their results are empowering for the people studied. Despite this, some qualitative researchers have continued to use the term 'validity,' seeking to define this in ways more appropriate for qualitative research while retaining the basic idea that validity is fundamentally about the relationship between one's theories or findings and a 'reality' to which these refer.

Generalizability has been a similarly contested concept for qualitative research. The traditional, quantitative definition (often termed 'external validity') refers to the extent to which a study's results apply to a wider range of people and settings than those actually studied. Since qualitative research typically involves much smaller samples, and rarely uses probability sampling that allows statistical generalization to the population sampled, this conception is clearly problematic for qualitative research. As a result, qualitative researchers have often employed the concept of 'transferability'; this concept does not imply that results can uncritically be generalized, but that they *may* apply more broadly, depending on differences in the nature and context of the situation to which they are transferred. This puts the burden of determining transferability on the person making the transfer, rather than on the original researcher.

Howard Becker, discussing the applicability of the results of a study of men's prisons to women's prisons, stated that such generalizations "are not about how all prisons are just the same, but about a process, the same wherever it occurs, in which variations in conditions create variations in outcomes." Robert Yin has termed this strategy 'analytic generalization' or 'generalizing to theory,' contrasting this with statistical generalization. In Becker's example, the theory involved is about the process by which a prison culture develops; this process may result in different outcomes in different contexts.

This issue of generalizability has led scholars to once again debate the value of qualitative research, thus politicizing the method in some policy venues. Still, qualitative inquiry has endured this persistent suspicion from methodological and epistemological skeptics concerned about the lack of generalizability (in the traditional quantitative sense of the term). Surviving the most strident of these debates, qualitative inquiry has proved to be an innovative genre of research for those interested in unfolding process and participant meaning, and especially in those situations where local cases do not fit the more generalized understanding.

The core issue in this methodological debate is one of quality and who gets to define it. Researchers new to qualitative inquiry may be surprised by a lack of essential standards for quality. The subjectivity of quality is particularly pertinent in qualitative inquiry where there is no pretense of universal standards for goodness, nor is there an expectation for traditional generalizability. This is not to say that qualitative inquiry is not valuable to broader audiences; rather, the issue is one of definition and conceptualization of goodness or application.

Quality does count. Every qualitative researcher, whether cognizant of it or not, applies some mental model of research goodness to his or her inquiry process. Every research choice reflects the underlying assumptions about quality in research. These mental models incorporate two aspects of research quality – technical and epistemological – that guide the reasoning process in research. Technical quality attends to methodology, specific design, and methods of data collection and analysis. Emphasizing qualitative inquiry as a creative process, Patton (2001) reminds researchers of the need for technically sound analysis, but not at the expense of the creative aspects of qualitative research.

Epistemological quality, on the other hand, introduces elements of perspective, subjectivity, and meaning inherent to qualitative inquiry. Good research is more than requiring a set of techniques to guarantee some level of validity and reliability during design and implementation. At the epistemological level, quality research questions the very representation of the study, both within and beyond the design. Why are certain participants selected? How are interview questions crafted? What observation notes are left unwritten? Who am I in this decision-making process? Epistemological quality accounts for multiple interpretations and readings of the research text; this approach assumes that many 'others' will determine the goodness of the research, including participants, journal editors and reviewers, and the reading audience.

Despite the differences between qualitative and quantitative approaches and methods, the two are not inherently incompatible; they are increasingly combined in what is typically called 'mixed method' or 'multimethod' research. (Many such studies are considered to be 'qualitatively driven'; the qualitative component of the study is primary, and its assumptions and goals largely shape the methods of the research, with the quantitative component having a supplementary role.) It is in fact the differences between the two approaches that create the complementarity that is the main strength of mixed methods research, enabling both triangulation (checking one method against another) and a deeper and wider understanding of the phenomena studied. The distinctive strengths of qualitative research – in particular, its ability to address meaning, context, and process – provide both a valuable alternative to strictly quantitative studies and complementary strategies that can be integrated with quantitative assumptions, goals, and methods.

There are numerous journals and handbooks specifically devoted to qualitative research, and qualitative studies are regularly published in many other journals. See the Web site citations for journals dedicated to qualitative research methods generally; there are many other journals devoted to qualitative research in specific fields.

See also: Case Studies; Constructivism and Constructionism: Methodology; Ethnography in Applied Social Research; Generalization: Conceptions in the Social Sciences; Grounded Theory: Methodology and Theory Construction; Interpretation in Anthropology; Interpretive Methods: Micromethods; Mixed Methods; Oral History; Participatory Action Research in Social Research; Qualitative Methods, History of; Realism, in Applied Research.

Bibliography

Bazeley, P., 2013. Qualitative Data Analysis: Practical Strategies. Sage Publications Ltd, London, UK.

Becker, H.S., 1991. Generalizing from case studies. In: Eisner, E., Peshkin, A. (Eds.), Qualitative Inquiry in Education: The Continuing Debate. Teachers College Press, New York, pp. 233–242.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage, Newbury Park.

Denzin, N.K., Lincoln, Y.S. (Eds.), 2011. The SAGE Handbook of Qualitative Research, fourth ed. Sage Publications, Thousand Oaks, CA.

Emerson, R.M., Fretz, R.I., Shaw, L.L., 2011. Writing Ethnographic Fieldnotes, second ed. University of Chicago Press, Chicago.

Flick, U. (Ed.), 2014. The SAGE Handbook of Qualitative Data Analysis. Sage Publications Ltd, London, UK.

Gallagher, K. (Ed.), 2008. Methodological Dilemma: Creative, Critical, and Collaborative Approaches to Qualitative Research. Routledge, London.

Given, L. (Ed.), 2008. The SAGE Encyclopedia of Qualitative Research. Sage Publications, Thousand Oaks, CA.

Glesne, C., 2011. Becoming Qualitative Researchers: An Introduction, fourth ed. Pearson, Boston, MA.

Gubrium, J.F., Holstein, J.A. (Eds.), 2002. Handbook of Interview Research: Context and Method. Sage, Thousand Oaks, CA.

Hesse-Biber, S.N., Leavy, P. (Eds.), 2006. Emergent Methods in Social Research. Sage, Thousand Oaks, CA.

Luttrell, W. (Ed.), 2010. Qualitative Educational Research: Readings in Reflexive Methodology and Transformative Practice. Routledge, New York.

Maxwell, J.A., 2012. Qualitative Research Design: An Interactive Approach. Sage Publications, Thousand Oaks, CA.

Patton, M.Q., 2001. Qualitative Research and Evaluation Methods, third ed. Sage Publications, Thousand Oaks, CA.

Rose, G., 2007. Visual Methodologies: An Introduction to the Interpretation of Visual Materials, second ed. Sage, Thousand Oaks, CA.

Weiss, R., 1994. Learning from Strangers: The Art and Method of Qualitative Interviewing. The Free Press, New York.

Willis, J.W., 2007. Foundations of Qualitative Research: Interpretive and Critical Approaches. Sage, Thousand Oaks, CA.

Relevant Websites

Journals

http://fmx.sagepub.com/ – Field Methods.

http://www.qualitative-research.net/index.php/fqs – Forum: Qualitative Social Research.

http://www.tandfonline.com/toc/tqse20/current#.U6CuRPldWSo – International Journal of Qualitative Research in Education.

http://www.jeqr.org/ – Journal of Ethnographic & Qualitative Research.

http://www.brill.com/journal-phenomenological-psychology#INDABS_1 – Journal of Phenomenological Psychology.

https://benjamins.com/#catalog/journals/ni/main – Narrative Inquiry.

http://www.sagepub.com/journalsProdDesc.nav?prodId=Journal200926 – Qualitative Health Research.

http://qix.sagepub.com – Qualitative Inquiry.

http://www.emeraldgrouppublishing.com/products/journals/journals.htm?PHPSESSID=3678quhhf4j63fqlpru81rget1&id=qmr – Qualitative Market Research.

http://qrj.sagepub.com – Qualitative Research.

http://www.sagepub.com/journalsProdDesc.nav?prodId=Journal201566 – Qualitative Social Work.

http://link.springer.com/journal/11133 – Qualitative Sociology.

http://www.qualitativesociologyreview.org/ENG/index_eng.php – Qualitative Sociology Review.

Quality of Government

Bo Rothstein, The Quality of Government Institute, University of Gothenburg, Gothenburg, Sweden
Jan Teorell, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The authors argue that the concept of quality of government should be best understood as that of having impartial government institutions. This definition avoids functionalism, ignores the contents of specific policies in favor of the procedures for how they are implemented, and pertain to the output side of the political system. They discuss this concept in relation to several proposed alternative definitions.

Introduction

The concept of quality of government – or good governance – has multiple meanings, but most generally refers to a standard or model for how states or other political entities should govern and be governed. Generally, this includes a long list of normative principles to which these entities should adhere, such as transparency, accountability, inclusiveness, effectiveness, and impartiality. Quite commonly, good governance is defined in terms of its antonyms, that is, by referring to phenomena that indicate its absence, such as corruption, nepotism, favoritism, particularism, or patrimonialism. Within policy circles of the developmental aid community, the concept has been used as an agenda for reform of developing countries, such as civil service reforms, securing property rights, or installing judicial independence. Within the academic community, the concept has been systematically linked to several highly desirable outcomes, most notably economic growth and long-term development (Rothstein, 2011).

But what, more exactly, does this new concept entail? It has been argued that there is a serious lack of conceptual precision in the use of these concepts (Fukuyama, 2011: p. 469; Andrews, 2010). In this article, the authors argue that the quality of government should be defined as having *impartial* government institutions. They will frame their justification for this conceptualization around a short intellectual background to the current issues at stake, followed by a criticism of previous efforts to define the concept. After having outlaid their own definition, the authors conclude by discussing some of the most challenging questions on the future research agenda of this field.

The Intellectual Background to the Debate

One of the major sources for the rise of the good governance and quality of government agenda has been the 'institutional turn' in the social sciences. Around 1990, three major works were published that have had a profound impact on the analysis of the importance of institutions, namely, James B. March and Johan P. Olsen's (1989) *Rediscovering Institutions*; Douglass C. North's (1990) *Institutions, Institutional Change and Economic Performance*; and Elinor Ostrom's (1990) *Governing the Commons*. Although coming from different intellectual traditions, they had one thing in common, namely, they challenged the then dominating societal views in studies of social and economic outcomes. These paradigms in the social sciences (for example, pluralism, elitism, and Marxism) all argued that variables like economic power configurations, systems of social stratifications, or the structure of class divisions were central in explaining political and thereby social and economic outcomes. Contrary to this, the institutionally orientated scholars argued that political institutions, broadly understood, were central in explaining social and economic outcomes. In political science, this became known as *Bringing the State Back In* (Evans et al., 1985; Steinmo et al., 1985).

In short, instead of focusing on how economic and sociological variables determined politics and outcomes of the political systems, the institutional approach turned the causal logic around by arguing that the character of a society's political institutions to a large extent determined its economic and social development. In common language, the institutional turn in the social sciences showed why 'the rules of the game' should have a more central role in social science research. This led to a number of interesting questions for research, such as, why societies had different institutions, what was the relation between institutions and social/economic outcomes, and if some type of institutions were better in producing valued social outcomes than others.

Different Conceptions of Quality of Government and Good Governance

To apprehend the scope of meanings attached to the concept, a series of distinctions are in order. First, good governance may be defined in terms of what it is (as a 'good in itself') or in terms of what it is good for. A prominent example of the latter approach is the definition of quality of government as what is 'good for economic development' (La Porta et al., 1999: 223). The former be called an *internal* and the latter an *external* (or functionalist) definition of good governance. In the internal case, some properties of the governance system of a country must be singled out in advance, together with a notion of why these properties make the system 'better' or of 'higher quality.' This could, for example, be the idea that good governance is to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75048-3

be equated with the rule of law. Countries abiding by this principle are then considered to have 'good governance,' regardless of whether the rule of law leads to other desirable consequences. In the external case, it is left for empirical scrutiny what properties of the system are considered to be 'good,' namely, those that are found to exert an influence on some preferable, external, outcome. In this case too, the rule of law might be considered to be part of good governance, but only to the extent that the rule of law can be shown to impact, e.g., economic growth or other desirable phenomena.

Second, good governance may be defined in terms of *procedures* or policy *content*. Procedural definitions refer to regulatory principles, codes of conduct, norms, or other value-laden criteria that constrain the forms in which politics may be conducted. Definitions demanding certain contents, by contrast, require certain policies to be enacted by the political system, such as market deregulation or state privatization.

Third, there is a fundamental difference between procedural definitions referring to the input and the output sides of the political system. The former, where *access* to public authority is regulated, is the realm of decision-making bodies, such as parliaments, cabinets, juntas, party executive committees, or royal councils. The output side concerns instead the *exercise* of public authority: how reforms are implemented, rules are enforced, court cases are decided, and so on. This is the realm of the public administration, the courts, and the law enforcement bodies of the state apparatus.

One of the most frequently used definitions of good governance has been launched by the World Bank Research Institute and reads as follows:

> The traditions and institutions by which authority in a country is exercised. This includes (1) the process by which governments are selected, monitored and replaced, (2) the capacity of the government to effectively formulate and implement sound policies, and (3) the respect of citizens and the state for the institutions that govern economic and social interactions among them.
>
> Kaufmann et al., 1999: p. 1

This definition forms the basis of the World Bank's widely used Worldwide Governance Indicators that have measures for 'voice and accountability,' 'political instability and violence,' 'government effectiveness,' 'regulatory quality,' 'rule of law,' and 'control of corruption.' This is a very broad definition and can be criticized for including both policy content ('sound policies') and procedures ('rule of law'), and for containing both the institutions for access to political power ('the process by which governments are selected') and those that exercise and implement laws and policies. In the words of Keefer (2004: p. 5), "if the study of governance extends to all questions related to how groups of people govern themselves …, then there are few subjects in all of political science and political economy that do not fall within the governance domain." Or put differently, if quality of government is everything, then maybe it is nothing.

Yet, clearly some political institutions or aspects of 'politics' must matter more than others for what should count as quality of government. The authors thus agree with the critique launched by Grindle (2007) that the World Bank's good governance agenda, not least for many poor countries, can be overwhelming and, in particular, with her argument that it fails to distinguish between various institutional particularities and more basic principles.

As an alternative, the authors propose an *internal* definition, which is purely *procedural* and based only on the *exercise* of public authority (Rothstein and Teorell, 2008). By way of defending this position, they discuss the limitations of a series of other approaches to conceptualizing quality of government.

Good Governance as the Absence of Corruption

One way out of the definitional problem would be to define quality of government and/or good governance simply as the absence of corruption. This turns out to be problematic for several reasons. First, corruption is in itself difficult to define. The standard definition is that corruption is 'the abuse of public power for private gain.' The problem with this definition is that it is relativistic since what counts as 'abuse' (or 'misuse') would vary in different parts of the world (Kurer, 2005). Needless to say, this would dramatically increase problems of operationalization and measurement but it would also carry all the difficulties connected to relativistic definitions that are known from discussions about human rights and democracy. Without a universally accepted normative standard about what forms of behavior are acceptable and appropriate, there is no way to know (and measure) what should count as 'abuse' when various systems of governance are compared in order to see if they would qualify for the epithet 'good' or not.

The second reason why quality of government cannot be equated with the absence of corruption is that there may exist many problems when governing societies are not confined to what is usually understood as corruption. A high degree of corruption is certainly an antithesis to good governance, but so are many other practices that are usually not seen as corruption, such as clientelism, lack of respect for the rule of law and property rights, nepotism, cronyism, patronage, systemic discrimination, and cases where administrative agencies are 'captured' by the interest groups that they are set out to regulate and control.

Good Governance as the Rule of Law

Perhaps as central as corruption, establishing the rule of law is usually key in any discussion on good governance and placed high on the agenda for reforming developing and transitional countries (Carothers, 1998). Still, although unequivocally embraced as a virtue of any system of good governance, the concept is rarely defined. One reason for this may of course be that the concept is inherently ambiguous and legal scholars argue over its exact meaning (Rose, 2004). To begin with, they dispute whether or not the rule of law should be given a purely procedural interpretation, bearing no implications for the actual substance of promulgated laws. Those that defend a procedural notion claim that the rule of law must be distinguished from the rule of 'good' law. Critics argue that this would allow morally detested regimes, such as Nazi Germany, to be classified as abiding by the rule of law. Against the procedural view, these critics seek to inscribe into the rule of

law various substantive moral values of liberal democracy (cf Bratton and Chang, 2006: pp. 1077–1078). Yet, even among proceduralists, who adhere to a narrower conception, ambiguities remain. Usually, more attention is paid to the internal qualities of the laws themselves – such as the need for the law to be clear, understandable, general, internally consistent, prospective, and stable – rather than to define the core principles that a political system must abide by in order to be in accordance with the rule of law.

Searching for these core principles, one may instead turn to conceptions developed within political science. Weingast (1997: p. 245) defines the rule of law as "a set of stable political rules and rights applied impartially to all citizens." Similarly, O'Donnell (2004: p. 33) states a minimal definition of the rule of law as "that whatever law exists is written down and publicly promulgated by an appropriate authority before the events meant to be regulated by it, and is fairly applied by relevant state institutions including the judiciary." He then specifies his normative term:

> By "fairly applied" I mean that the administrative application or judicial adjudication of legal rules is consistent across equivalent cases; is made without taking into consideration the class, status, or relative amounts of power held by the parties in such cases; and applies procedures that are preestablished, knowable, and allow a fair chance for the views and interests at stake in each case to be properly voiced.

The rule of law thus embodies the principle 'equality before the law.' It entails "a crucial principle of fairness – that like cases be treated alike" (O'Donnell, 2004: pp. 33–34). Although similar in spirit to the definition the authors suggest below, one problem with the rule of law approach is that good governance also applies to spheres of state action other than those directly governed by law. When public policy is to be enacted in the so-called 'human processing' areas, such as education, health care, welfare benefits, and active labor market programs, widely discretionary powers usually need to be transferred to lower level government officials and professional corps responsible for implementing policy. The reason is that they have to adapt actions to the specific circumstances in each case and it has turned out to be administratively impossible to enact precise 'rule-of-law-type' laws and regulations that can guide this. In many areas, governance is carried out by professional corps who are for the most part guided by professional standards issued by their organizations, which are not connected to 'rule of law' principles. The conclusion is that while the 'rule of law' principles in most approaches serve as a central ingredient in good government, they do not cover the full spectrum of the concept.

Good Governance and Democracy

Establishing representative democracy has often been championed as an effective antidote to everything from corruption to poverty. This is because it is linked to accountability, which helps to reduce the discretionary powers of public officials (Deininger and Mpuga, 2005: p. 171). This would indicate that democracy, quality of government, and good governance could possibly conceptually overlap, as is the case in the World Bank's definition mentioned above. This raises the question why concepts like good governance, state capacity, and quality of government are needed since we could just talk about 'democracy' as the overall standard for evaluating the polity.

The problem, however, is that empirically, there is no straightforward relationship between establishing electoral representative democracy and many features of good government. On the contrary, the relationship between measures of degree of democracy and degree of corruption shows that the curve is U-shaped or J-shaped (Montinola and Jackman, 2002; Sung, 2004; Bäck and Hadenius, 2008). Empirical research indicates that corruption is worst in countries that have newly been democratized. For example, some of the worst cases of corruption have appeared in newly democratized countries, such as Peru under its former president Fujimori (McMillan and Zoido, 2004) and Jamaica since the mid-1970s (Collier, 2006). At the more country-specific levels, one can note that studies of Italy show that politicians who stand accused of or are under investigation for corruption do not stand a lesser chance of being elected than 'clean' politicians (Chang et al., 2010). One should also keep in mind that the two states that have made the greatest progress in curbing corruption over the past few decades – Singapore and Hong Kong – have not been and still are *not* democracies (Uslaner, 2008). From this, and from the empirical research (referred to above) showing that measures of various aspects of quality of government have a much greater impact on human well-being (and perceptions of political legitimacy) than measures of democracy, it may be concluded that quality of government is different from, and should not conceptually be equated with, democracy.

Quality of Government as Government Efficiency

It would certainly be strange to argue that a government that is very inefficient or ineffective could be of high quality or produce good governance (Fukuyama, 2004). Would it then be possible to define quality of government in terms of government efficiency or effectiveness? The reasons why this is problematic are two. First, the notions of 'good' or 'high quality' usually imply other things than just economic efficiency. It is easy to think of things that a government can carry out in an efficient way that normatively would be just the opposite of 'good.' Secondly, defining concepts like quality of government in terms of administrative and regulative efficiency would border on establishing a tautology. One should bear in mind that the good governance agenda largely came about in studies trying to understand why many developing countries were unable to increase growth. Defining good governance in terms of efficiency (or efficient policies) would be tantamount to saying that efficiency causes efficiency. Not much would be gained by saying that societies with efficient (good, high-quality) governance systems produce efficiency. If not a tautology, one could say that such a definition would make the distance between independent and dependent variables minimal. Instead, what needs to be known is if societies that are socially and economically efficient, that is, are able to solve the problem of producing the amount and type of public goods they need, have institutions that are qualitatively different in their operative principles than the opposite type of societies.

In the long discussion of how representative democracy should be defined, the distinction between procedural and substantive definitions is a central theme (cf Dowding et al., 2004). Since we are striving for a universal and procedural definition that could be acceptable to groups in a democracy with, to quote John Rawls (2005: p. xvi), "a pluralism of incompatible yet reasonable," comprehensive religious, philosophical, and moral doctrines, including substantial policies in the definition is a very risky business and not likely to achieve broad-based legitimacy. This is also why the authors prefer the term 'quality of government' to the World Bank term 'good governance' since the latter has too broad connotations.

Quality of Government as Impartiality

To sum up, neither the absence of corruption, nor democracy, the rule of law, or administrative effectiveness capture what should be counted as quality of government. Moreover, including 'sound policies' in the definition raises the quite problematic question whether international (mostly economic) experts really can be expected to be in possession of reliable answers to the question of what is 'sound policies.' For example, should pensions, health care, or education be privately or publicly funded (or a mix of these)? To what extent and how should financial institutions be regulated?

Searching for a definition, it is notable that the conceptual discussion has largely been detached from normative political theories about social justice and the state. It should be obvious that when terms like 'good' or 'quality' are placed in political concepts, it is impossible to refrain from entering the normative issues that are raised in political philosophy. Obviously, a definition of quality of government need not relate to a specific set of institutional arrangements. Instead, one has to look for some *basic norm* that characterizes their institutional systems as a whole. The question is what would be the equivalent to Dahl's (1989) political equality for the 'output' side of the political system. Based on the type of rights-based liberal political theory launched by philosophers such as Brian Barry and John Rawls, the authors have suggested such a basic norm, namely, *impartiality* in the exercise of public power. This is defined in the following way: "When implementing laws and policies, government officials shall not take anything about the citizen or case into consideration that is not beforehand stipulated in the policy or the law" (Rothstein and Teorell, 2008: p. 170; cf Strömberg, 2000: p. 66). This definition is fairly precise and can be applied universally. It makes clear what basic norm is being 'abused' when corruption, clientelism, favoritism, discrimination, patronage, nepotism, or undue support to special interest groups occurs. It excludes the content of policies since it is strictly procedural.

A related suggestion on how to define quality of government launched by Mungiu-Pippidi (2006) is *universalism* (as different from *particularism*) in public policy formation and implementation. By this she means that public policies should not be directed to favoritism of individuals because they belong to particular groups of elites or have had resources to bribe officials. Yet another suggestion comes from North et al. (2009) and is *impersonal* enforcement of rules as opposed to *personalistic* forms of wielding public power. A closer inspection reveals that the differences between the notion of impartiality of these authors and the authors of this article are mostly terminological. What is important is that they all suggest a similar universal standard for what should count as quality of government and its opposite.

What does it mean to be impartial in the exercise of public power? Cupit writes: "To act impartially is to be unmoved by certain sorts of considerations – such as special relationships and personal preferences. It is to treat people alike irrespective of personal relationships and personal likes and dislikes" (Cupit, 2000; cf Barry, 1995: p. 11). The connection to 'good' or 'quality' is motivated by the fact that impartiality is the driving notion behind John Rawls' liberal right-based theory of justice. As Goodin argues: "Certainly, the antithesis of justice is favouritism" (2004: p. 100). In this context, impartiality is not a demand on actors on the input side of the political system, but first and foremost an attribute of the actions taken by civil servants, professional corpses in public service, law enforcement personnel, and the like.

Equally important, however, are the things that the norm of impartiality does *not* rule out. Since quality of government as impartiality is a procedural norm confined to the exercise of public power, one important field that is not affected by this conception is the substance of the content of policies. This builds on the idea that noncorruption implies that "a state ought to treat equally those who deserve equally" (Kurer, 2005: p. 223). This is in line with the argument that the content of public policies should not be included in the definition. Instead, it is impartiality in the exercise of power (the 'ought to treat equally' principle) that is the central component of quality of government. To treat equal does of course not imply that everyone should get the same. Only people who are in need of kidney transplantation should get one. Instead, this follows the idea of 'equal concern and respect' launched by Ronald Dworkin (1977).

It should be underlined that the argument is not that impartiality is equivalent to 'objectivity.' Still it could be said that, as a concept, objectivity has an absolute and perfectionist ring to it that implies that humans can have full knowledge of a case and weigh all things equal and come down with a decision as if the outcome was decided by some natural law process. The authors would argue that impartiality implies somewhat more human and realistic demands. First, it is about a 'matter of factness,' implying that things that according to the policy/law should not have an impact on the decision are to be left out. Secondly, it requires that the public official should not be a party to the case, neither directly nor indirectly. Moreover, the idea of quality of government as impartiality stands in sharp contrast to the public choice idea of public officials maximizing their self-interest. For example, an impartial civil servant should not be susceptible to bribery, should not decide in cases where his or her friends and relatives are involved, and should not favor any special (ethnic, economic, or any other type of organized) interest when applying laws and rules.

Conclusions

In this article, the authors have argued that impartiality provides a coherent, encompassing, and universally applicable

conception of quality of government. Elsewhere, they have also shown that it is a measurable conception (Rothstein and Teorell, 2012). This leaves two very broad sets of questions on the table. The first concerns the consequences of impartiality. After all, the new research agenda on good governance and quality of government has mostly been driven by a desire to understand what government institutions lead to preferable outcomes, such as peace, prosperity, and general human well-being. Can quality of government as impartiality live up to these expectations? In other words, can it be shown empirically that this novel theory and conception of quality of government really is what produces these outcomes?

The other central question is of course the origins of impartiality. Why have some countries in the world been more successful than others in turning their nepotistic, clientelistic, patrimonial, and corrupt government institutions into impartial state structures? Is there a common cure that could be transferred from previously successful cases to new political orders, or is each country trajectory unique, forever bound up in its own historically predetermined path dependence? One should not underestimate the problems in making transitions from low to high quality of government.

One particular problem is that, precisely because high-quality government institutions are impartial, they have no obvious interest group that is their natural supporter. The natural thing for any interest group, whether based on economic, ideological, or ethnical orientation, is to strive for political institutions and regulations that serve their particular interests and such institutions are by nature not impartial. Instead, impartial government institutions must be seen as 'public goods' type of institutions and are thereby suffering from all the well-known problems of collective action in creating goods such as various forms of 'free-riding' and others sorts of opportunistic behavior. An example of the magnitude of this problem is the Nobel Prize in Economics in 2009 that went to the political scientist Elinor Ostrom for her studies in how impartial government institutions can be established for preserving natural 'common pool' resources. While the authors think this prize was very well deserved, it should be noted that Ostrom's cases are relatively small local groups where the agents have known each other for a very long time and, as she states, have been able to develop norms about reciprocity, trust, and social capital (Ostrom, 1990: p. 35). The authors' project is to understand how such institutions can come about in much larger settings where one cannot assume the existence of such norms of reciprocity, social capital, or interpersonal trust among the agents.

See also: Executive Government; Governance; Governments.

Bibliography

Andrews, M., 2010. Good government means different things in different countries. Governance 23 (1), 7–35.

Barry, Brian, 1995. Justice as Impartiality. Oxford University Press, Oxford.

Bratton, M., Chang, E.C., 2006. State building in sub-Saharan Africa – forwards, backwards, or together. Comparative Political Studies 39, 1059–1083.

Bäck, H., Hadenius, A., 2008. Democracy and state capacity: exploring a J-shaped relationship. Governance 21 (1), 1–24.

Carothers, T., 1998. The rule of law revival. Foreign Affairs 77, 95–106.

Chang, E.C.C., Golden, M.A., Seth, H.J., 2010. Legislative malfeasance and political accountability. World Politics 62 (2), 177–197.

Collier, M.W., 2006. Political Corruption in the Caribbean Basin. Routledge, London.

Cupit, G., 2000. When Does Justice Require Impartiality? Presented at the Political Studies Association-UK 50th Annual Conference, London.

Dahl, R.A., 1989. Democracy and its Critics. Yale University Press, New Haven, CT.

Deininger, K., Mpuga, P., 2005. Does greater accountability improve the quality of public service delivery? Evidence from Uganda. World Development 33 (1), 171–191.

Dowding, K., Goodin, R., Pateman, C. (Eds.), 2004. Justice and Democracy. Cambridge University Press, Cambridge.

Dworkin, Ronald, 1977. Taking Rights Seriously. Duckworth, London.

Evans, P., Rueschemeyer, D., Skocpol, T. (Eds.), 1985. Bringing the State Back In. Cambridge University Press, Cambridge.

Fukuyama, F., 2004. State-building: Governance and World Order in the 21st Century. Cornell University Press, Ithaca.

Fukuyama, F., 2011. The Origins of Political Order: From Prehuman Times to the French Revolution. Farrar, Straus and Giroux, New York.

Grindle, M.S., 2007. Good enough governance revisited. Development Policy Review 25 (5), 553–574.

Kaufmann, D., Kraay, A., Zoido-Lobatón, P., 1999. Governance Matters. Policy Research Working Paper 2196. World Bank Institute, Washington, DC.

Keefer, P., 2004. A Review of the Political Economy of Governance: From Property Rights to Voice. World Bank Policy Research Working Paper No. 3315.

Kurer, O., 2005. Corruption: an alternative approach to its definition and measurement. Political Studies 53 (1), 222–239.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1999. The quality of government. Journal of Law, Economics and Organization 15 (1), 222–279.

March, J.B., Olsen, J.P., 1989. Rediscovering Institutions: The Organizational Basis of Politics. Basic Books, New York.

McMillan, J., Zoido, P., 2004. How to subvert democracy: montesinos in Peru. Journal of Economic Perspectives 18 (4), 69–92.

Montinola, G.R., Jackman, R.W., 2002. Sources of corruption: a cross-country study. British Journal of Political Science 32, 147–170.

Mungiu-Pippidi, A., 2006. Corruption: diagnosis and treatment. Journal of Democracy 17 (3), 86–99.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge, MA.

North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History. Cambridge University Press, Cambridge, MA.

O'Donnell, G., 2004. Why the rule of law matters. Journal of Democracy 15 (4), 32–46.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Rawls, J., 2005. Political Liberalism. Columbia University Press, New York.

Rose, J., 2004. The rule of law in the western world: an overview. Journal of Social Philosophy 35, 457–470.

Rothstein, B., 2011. The Quality of Government: Corruption, Social Trust and Inequality in a Comparative Perspective. University of Chicago Press, Chicago.

Rothstein, B., Teorell, J., 2008. What is quality of government: a theory of impartial political institutions. Governance 21 (2), 165–190.

Rothstein, Bo, Teorell, J., 2012. Defining and measuring quality of government. In: Holmberg, S., Rothstein, B. (Eds.), Good Government: The Relevance of Political Science. Edward Elgar, Cheltenham.

Strömberg, H., 2000. Allmän Förvaltningsrätt. Liber, Malmö.

Steinmo, S., Thelen, K., Longstreth, F., 1985. Structuring Politics: Historical Institutionalism in Comparative Perspective. Cambridge University Press, Cambridge.

Sung, H.-E., 2004. Democracy and political corruption: a cross-national comparison. Crime. Law & Social Change 41, 179–194.

Teorell, J., 2010. Determinants of Democratization: Explaining Regime Change in the World, 1972–2006. Cambridge University Press, Cambridge.

Uslaner, E.M., 2008. Corruption, Inequality, and the Rule of Law: The Bulging Pocket Makes the Easy Life. Cambridge University Press, Cambridge.

Weingast, B.R., 1997. The political foundations of democracy and the rule of law. American Political Science Review 91 (2), 245–263.

Quantification in History

Konrad H Jarausch and Peter A Coclanis, University of North Carolina at Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of quantification is defined and the evolution of quantification as a methodological tool in history is traced. Originating in the early modern period and gaining greater prominence in the nineteenth century, quantification in history came of age in the 1960s. For a time, in the 1970s and early 1980s, quantitative approaches seemed poised to revolutionize history as a field. Beginning in the mid-1980s, however, other approaches and methods, often drawn from anthropology and literature and inspired by postmodernism, began to surpass quantification in popularity, and since the late 1980s, quantitative work in history has been increasingly marginalized. But over the past decade, postmodernist approaches have lost momentum in history, and there have been some small victories by quantifiers, particularly in the field of economic history. Yet despite the fact that the approach remains indispensable for answering certain kinds of historical questions and despite the effective employment of quantitative approaches by many historically inclined scholars in other social sciences, the future of quantification in the discipline of history itself remains uncertain.

'Quantification' refers to a specific method of historical research that departs from the prevalent practice of textual analysis by actually seeking to measure changes in the past. Due to the mass of data to be handled and the complexity of the calculations involved, the application of quantitative methods usually has to be computer assisted. In contrast to a mere illustration of certain tendencies by tables or graphs, quantification tends to use analytical statistics to probe relationships between and among certain variables. Such statistical analysis also requires a high degree of formalization in the research design that formulates hypotheses to be tested and tries to combine them into models. The final aim of the quantitative approach is not just to explain a particular set of developments but to understand broader processes of change and make a theoretical contribution to some facet of historical understanding.

The Rise of Quantification

Within historical research, the development of quantitative methods was a product of a particular constellation during the 1960s. Already during the Enlightenment, some scholars had sought to describe socioeconomic developments by compiling time series, and by the middle of the nineteenth century national governments took on the task of determining population trends through a series of censuses. Nonetheless, it took the shift to the 'new social history' a century later to get historians to employ quantitative data in systematic ways and to direct historical interests away from affairs of state to mass politics, societal changes, demographic transformations, or economic growth. At the same time, the arrival of mainframe computers, followed eventually by powerful PCs, made it possible to handle large amounts of data and to use complicated statistical procedures in customized form via statistical packages such as SPSS or SAS. In order to explore these exciting new subjects with novel techniques, historians became interested in the theoretical constructs of the social sciences, borrowing eclectically from political science, sociology, or demography and somewhat less so from economics. The result of this interaction was the rapid development of a new hybrid, combining elements of historical research with social scientific analysis.

During the 1970s, this 'social science history' or 'historical social science' was successfully institutionalized. In the United States, the multitude of initiatives in history and the neighboring disciplines combined in founding the Social Science History Association in 1975, and made its annual meeting a showcase of the most innovative work. In France, many of the long-term community studies, produced by the Annales School, turned quantitative, compiling detailed time series of population, prices, or wages and probing their interrelationship. In Germany, the quantifiers linked with the theory-oriented project of a Historische Sozialwissenschaft to create a vigorous organization, called QUANTUM (Kocka, 1984). Within the decade, a spate of new journals emerged to provide publication outlets for the new work, such as the *Journal of Social History*, *Journal of Interdisciplinary History*, *Historical Methods*, *Social Science History*, *Histoire et Mésure*, and *Historical Social Research*. New textbooks (Dollar and Jensen, 1971; Floud, 1979; Jarausch and Hardy, 1991) as well as influential anthologies (Aydelotte et al., 1972; Lorwin and Price, 1972) provided a powerful platform for the propagation of quantitative research.

Research Contributions

The first field transformed by quantitatively oriented work was the new political history, which differed from its predecessor by its attention to collective phenomena (Bogue, 1983). For instance, Michael Kater investigated the rise of the Nazi movement in Germany by analyzing its membership files, indicating precisely which social stratum, religious group, gender, region, etc. joined the bandwagon at what time. For democratic societies like the United States, Lee Benson's or Alan Bogue's analysis of electoral behavior proved quite productive, since comparing returns over time made it possible to establish which polls were the critical elections when party realignments

were transformed by ethno-cultural or other factors. For parliaments without party discipline, Thomas B. Alexander's roll-call analysis also offered a powerful instrument for establishing who actually voted with whom, indicating what networks dominated legislative behavior. This retro-projection of political science methods gradually provided a firmer empirical basis for an analysis of the transformation of mass politics in the last two centuries.

The second area that was deeply affected by quantitative methods and conceptual borrowing from sociology was the new social history (Stearns, 1985). As 'history from the bottom up,' this perspective examined previously untouched serial records such as police files, employment registers, or draft books in order to infer from the behavior of ordinary people what they might have felt and thought. The appeal of this effort was the diffusion of its topics, since the labor history of the Tillys included subjects as different as proletarian living conditions and workers' collective efforts at protesting through strikes. Another question, treated, for example, by Hartmut Kaelble, concerned the pattern of social mobility, comparing the relative openness of societies to upward or downward movement over time or across borders. A related issue was the societal impact of education that required an analysis of the enrollment and composition of schools or universities in order to get at changes in educational opportunities (Ringer, 1979). These quantitative probes produced much new and sometimes conflicting information about the lower orders, the make-up of institutions, and basic trends in societal transformations.

A third area propelled by computer-assisted source collection and statistical analysis was the field of historical demography (Hareven and Vinovskis, 1978). For time periods before national censuses, the demographers faced the challenge of reaggregating vital statistics from the ground up by converting parish registers into reliable measures of natality, nuptiality, or mortality. Under the direction of Anthony Wrigley, the Cambridge population project established the outlines of a population history for England, while the demography center in Umea undertook a national project of quantifying church records so as to explain demographic changes in Sweden. American demographers like John Demos and Tamara Hareven instead combined individual life-event data to scrutinize the development of family structures from extended via stem to nuclear families and the like. Other scholars such as Roderick Floud and Richard Steckel studied changes in the height and stature of military recruits or slaves in order to develop more accurate measures of biological well-being and social deprivation. This historical demography amassed a source base for the prestatistical era and produced interesting interpretative concepts for explaining the population transition.

The final area in which quantification transformed research priorities and methods was the new economic history (Davis and Engerman, 1987). Redirecting attention away from entrepreneurship, 'cliometrics' as it was also called inspired efforts to establish indicators of national production and income for the prestatistical age as well as to transform information from manuscript censuses into published form. In the United States, a heated debate revolved around the profitability of slavery, with Robert Fogel's and Stanley Engerman's claims producing much controversy, due to their combination of economic and moral argumentation. Much attention also focused on the key factors of economic development, such as railroads, with some scholars actually arguing that their impact had been exaggerated. In Europe, discussions revolved around the pattern of industrialization instead, with some specialists like Sidney Pollard proposing a more complex regional diffusion model. On the whole, the quantitative impulse made economic history more rigorous and theoretical, but removed it at the same time further from the mainstream of historical debate.

By the 1980s, these signal contributions had managed to overcome traditionalist skepticism and to establish quantification as a new and growing subfield within the historical discipline. Though they were sometimes difficult to read for the layperson, high-quality monographs convinced reluctant granting agencies to fund major collective initiatives such as the Philadelphia Social History Project, directed by Theodore Hershberg. The number of articles with graphs and tables in respected historical journals began to reach almost one-quarter (Kousser, 1989). Opinion surveys of the profession suggested that most scholars expected quantification to be here to stay. Even if some critics still bemoaned the spottiness of graduate training in this area, the rapid rise of quantitative methods in historical research had been impressive and their future prospects seemed secure.

The Culturalist Challenge

During the decade of the 1980s, however, the seemingly unstoppable progress of the quantitative history 'project' began to stall (Johnson, 1989). By the middle of the decade, the number of quantitative journal articles and the frequency of graduate courses in quantitative techniques suddenly began to drop precipitously (Reynolds, 1998; Ruggles, 2012). This trend has continued ever since, and over the past 20 years (plus or minus five!) quantitative history has become increasingly marginalized, as a variety of other methods rose to prominence, making so-called quants seem rather quaint. Ironically, the demise of quantitative history has occurred simultaneously with the development of enhanced technology, more complete datasets, and increasingly sophisticated models and methods, which, taken together, have rendered quantitative history potentially more useful than ever before.

The factors responsible for this loss of attention are still being debated vigorously. Some scholars attribute the problems afflicting quantitative history to the initial overselling of the method by insufferable practitioners who promised more insights than their research was able to deliver. These critics, comprising largely traditionalists, have always been suspicious of the proposition that history had anything to learn from the social sciences and have long bristled at heady sentiments such as that expressed by Emmanuel Le Roy Ladurie, who predicted in 1967 that in the future "the historian will be a programmer or he will be nothing" (Rabb, 1983). The most famous (or infamous) of such critiques is almost certainly that of Carl Bridenbaugh, who in his American Historical Association presidential address in 1962 stated that "[t]he finest historians will not be those who succumb to the dehumanizing methods of the social sciences," nor those who "worship at the shrine of that Bitch-goddess, QUANTIFICATION" (Bridenbaugh, 1963). In later years, other skeptics such as Bernard Bailyn and

Lawrence Stone were more subtle, calling not for the explicit abandonment of quantification, but instead for a revival of narrative history.

Another reason was the cultural turn of many younger historians who rejected both the materialist premises and the positivist vision of most practitioners of quantitative history. To these critics, quantitative representations of the past were at once too cold, too remote, too impersonal, and too abstract. Large datasets were said to dehumanize, and 'soulless' quantifiers were assumed to be ready, like Marx's bourgeoisie, to drown all human sentiments "in the icy water of egotistical calculation" (Coclanis, 1992). Moreover, in making fetishes of rigor and precision, quantitative historians purportedly attempted to exact more from the vast, vague, largely unknowable past than historians had any right to claim. Instead, the past was reconfigured as a foreign country, to invoke the title of historian David Lowenthal's influential 1985 book (Lowenthal, 1985).

The changing demographics of the historical profession, particularly in the United States, also served to reinforce the trend toward qualitative approaches. As history opened up to women and minorities – groups, generally speaking, less inclined to pursue quantitative approaches to the past – quantitative history could be expected to decline in a relative sense. Just as women and minorities have been underrepresented in the hard sciences over the years, they have proven less interested in the more social scientific sides of history. Though both groups have embraced social history as a vehicle for their identity politics, relatively few of their members can be found in fields such as economic history, historical demography, or quantitative political history.

During the 1980s, many historians turned instead to cultural anthropology for methodological inspiration. Rejecting the rigorous, formal, and remote data-driven 'truths' sought by quantifiers, scholars interested in experience and meaning preferred in-depth microstudies, whether of single villages, small groups, or even well-documented individuals. Gradually, 'thick description' came to trump ecological regression, crude death rates, and multiple R-squares, no matter how robust! The philosophical basis of this shift to anthropology was the rise of postmodernism, which deprecated the rationalism, universalism, and empiricism of historical social science. By the late 1990s, the ascendancy of cultural studies thus left quantifiers off in a corner silently counting all by themselves.

Recent Developments

In a general sense, the spiral has continued downward for quantitative history over the past 10 years, but, when examined more closely, the overall pattern – and prognosis – for the approach becomes far more complicated. On the one hand, there is abundant evidence that quantitative approaches continue to languish among historians. With such trends as the 'history of emotions,' today few articles and books emanating from scholars in history departments employ quantitative methods; history courses in quantitative methods – and, indeed, those that once taught them – have largely been mothballed, and few departments display interest in, much less privilege, the quantitative skills of applicants to their graduate programs. Some associations and journals have withered and died – the British-based Association for History and Computing and its journal *History and Computing*, most notably – and there are virtually no job listings for quantitative historians (Anderson, 2008).

On the other hand, excellent historical work employing quantitative methods continues to appear, albeit generally by social scientists working with historical data. Economic historians, the vast majority trained as economists and working out of departments of economics, have produced an impressive body of work over the past decade, and an exceptional new journal – *Economics and Human Biology* – has become the principal venue for quantitative work in historical anthropometrics (Coclanis, 2010). Some of the interest among economic historians was spurred by the so-called Great Recession (2007–09), which led many economists to look anew at earlier periods of severe economic distress (particularly the Great Depression of the 1930s, but also the depressions of the 1870s and 1890s).

Quantitative history has flourished as well over the last decade in disciplines other than economics. Indeed, what (quantitative) historian Steven Ruggles has labeled QUASSH (quantitative social science history) is booming in allied social sciences, with top journals in demography, sociology, political science, and geography frequently featuring historical work with a quantitative bent (Ruggles, 2012). Nor does this trend begin and end with the traditional social sciences. For example, the emergence of new fields such as digital humanities and historical information science that place heavy emphasis on computing and on electronic-data storage and retrieval has proven a boon to quantitative history. Moreover, many historically minded linguists and specialists in literature have embraced quantitative methods in their work, with Stanford's Franco Moretti probably the most notable – and certainly controversial – of the literary scholars going the quantitative route (Moretti, 2005). Thus, if all is not well in history department corridors for quantitative approaches to the past, such approaches are more than holding their own in other preserves of the academic realm.

The Continued Importance of Quantification

The current disregard of quantitative methods in the discipline is unfortunate, since quantification continues to be an essential technique for historical research. Some questions can best be answered through recourse to what Sir William Petty referred to in the seventeenth century as 'Number, Weight, or Measure,' and it is more than 'physics envy' that has drawn scholars to quantitative analyses of the past. Even after conceding that many early practitioners were a bit overly zealous in their claims, it seems fair to suggest that quantitative history can provide the methodological discipline, intellectual rigor, and moral purpose that subjectivist forms of history often lack. With all due respect to such influential postmodernist thinkers such as Baudrillard and Lyotard, not every aspect of the past is positional, free-floating, and contextual. Moreover, historians are better trained to handle and preserve data from the past than social scientists who often take figures at their face value. The challenge therefore is not to abandon quantification to the

social scientists, but to create a constructive dialogue between quantitative and qualitative methods (Grossbart, 1992; Haskins and Jeffrey, 1990; Hudson, 2000).

The utility, indeed, the necessity of quantitative history seems more compelling in light of important new developments in the field. Technological constraints have eased considerably, for example, as miniaturization and digitalization in the electronics industry have advanced. Today's personal computers, programmable calculators, and other platforms offer tremendous power and, thus, computational possibilities to quantitative historians. Similarly, new developments in statistical packages, spreadsheets, GIS (geographic information systems) programs, text analysis and retrieval software, file-sharing packages, and the like have vastly expanded the array of data-gathering and analytical tools that can be used in quantitative research on historical questions.

The increasing availability of new and/or enhanced sources and datasets also presents unprecedented research opportunities. Materials readily retrievable from established institutions such as the ICPSR (Interuniversity Consortium for Political and Social Research) in Ann Arbor, Michigan, and data available online from the World Bank, the International Money Fund (IMF), the Eurostat (EU), and other governmental, quasigovernmental, and nongovernmental organizations and agencies are transforming the nature (and sites) of scholarly production. The online version of the millennial edition of the *Historical Statistics of the United States*, datasets assembled by Yale's Cambodia Genocide Program, the University of Minnesota Population Center's IPUMS (Integrated Public Use Micro Sample) Project, drawn from US census records, and the East German central cadre files – to cite but four examples – are of similar scholarly importance, and have tremendous potential as teaching tools. Various research projects undertaken under the auspices of, or promoted by the now defunct Association for History and Computing, founded in England in 1987, have proven extremely useful as well (Anderson, 2008).

At the same time, quantifiers have in recent years developed more sophisticated methods, allowing them to analyze and interpret historical questions with greater sensitivity and precision. Economic historians, for example, have used game-theoretical approaches, complex experimental approaches in laboratory settings, and advanced econometric techniques to excellent effect, and anthropometric historians have employed important new procedures from auxology and biostatistics to facilitate their reconstruction of our biological past (Komlos, 1992; Steckel, 1995; Steckel and Rose, 2002). Economic and social historians, along with historical sociologists, have also employed regression analysis in more subtle and refined ways, using temporally recursive regression procedures, when appropriate, instead of standard time-series regression techniques. And psephologists – political scientists specializing in election analysis – are using increasingly powerful analytical tools in their analyses of historical voting data.

Other quantitative historians have begun to 'push the envelope' in different ways, in so doing, often trespassing on the territory of scholars in other social sciences or even in the humanities. Thus, as procedures such as social network analysis (SNA), event-structure analysis (ESA), and quantitative language analysis are successfully employed in answering historical questions – quantitative analyses of major 'texts' in order to produce so-called semantic grammars, for example – non- and anti-quantifiers ignore the results at their peril (Anderson, 2008). In addition, quantitative historians' recent incorporation of spatial relationships into historical analysis through spatial-effects regression models and the like has produced similarly impressive results (Griffin and van der Linden, 1998).

In the second decade of the twenty-first century, the future prospects of quantification in history are therefore indistinct. Clearly, individual quantitative historians, among the profession's most avid proponents of scientific rationalism, universalism, and empiricism, are continuing to expand the frontiers of their research and are still making important contributions. But whether or not the interest of mainstream historians will return to quantification and the subfield as such will ever again recapture the excitement and energy it possessed in the 1970s remains an open question. The relative popularity of various approaches does not just depend on the quality of their scholarly work but on the larger tides of intellectual fashion. However, one unexpected development offers some cause for optimism: Though quantitative historians are hanging on for their professional lives, historically inclined political scientists, economists, demographers, and sociologists are winning kudos in their respective disciplines by bringing quantitative methods to bear on the past.

See also: Anthropology and History; Data Bases and Statistical Systems: Archives and Historical Databases; Economic History; Event History Analysis: Applications; Historical Thought and Historiography: Current Trends; History and the Social Sciences; Hypothesis Testing: Methodology and Limitations; Linguistic Turn and Discourse Analysis in History; Measurement Theory: History and Philosophy; Network Analysis; Postmodernism: Methodology; Slaves and Slavery, History of; Social Mobility, History of.

Bibliography

Anderson, I., 2008. History and Computing. Making History. Institute of Historical Research University of London. http://www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html.

Aydelotte, W.O., Bogue, A.G., Fogel, R.W. (Eds.), 1972. The Dimensions of Quantitative Research in History. Princeton University Press, Princeton, NJ.

Bogue, A.G., 1983. Clio and the Bitch Goddess. Sage Publications, Beverly Hills, CA.

Bridenbaugh, C., 1963. The great mutation. American Historical Review 68, 315–331.

Coclanis, P.A., 1992. History by the numbers: why counting matters. OAH Magazine of History 7, 5–8.

Coclanis, P.A., 2010. The audacity of hope: economic history today. AHA Perspectives in History 48, 21–25.

Davis, L.E., Engerman, S., 1987. Cliometrics: the state of the science. Historical Methods 20, 97–106.

Dollar, C.M., Jensen, R.J., 1971. Historian's Guide to Statistics. Holt, Rinehart and Winston, New York.

Floud, R., 1979. An Introduction to Quantitative Methods for Historians, second ed. Methuen, London.

Griffin, L.J., van der Linden, M. (Eds.), 1998, New methods for social history. International Review of Social History 43 (Suppl. 6), 3–8.

Grossbart, S.R., 1992. Quantitative and social science methods for historians. Historical Methods 25, 100–120.

Hareven, T.K., Vinovskis, M.A. (Eds.), 1978. Family and Population in Nineteenth Century America. Princeton University Press, Princeton, NJ.

Haskins, L., Jeffrey, K., 1990. Understanding Quantitative History. MIT Press, Cambridge, MA.

Hudson, P., 2000. History by Numbers: An Introduction to Quantitative Approaches. Oxford University Press, New York.
Jarausch, K.H., Hardy, K., 1991. Quantitative Methods for Historians. University of North Carolina Press, Chapel Hill, NC.
Johnson, E., 1989. Reflections on an old 'new history': quantitative social science history. Central European History 22, 408–426.
Kocka, J., 1984. Theories and quantification in history. Social Science History 8, 169–178.
Komlos, J., 1992. Anthropometric history: what is it? OAH Magazine of History 6 (4), 3–5.
Kousser, J.M., 1989. The state of social science history in the late 1980s. Historical Methods 22, 13–20.
Lorwin, V.R., Price, J.M. (Eds.), 1972. The Dimensions of the Past. Yale University Press, New Haven, CT.
Lowenthal, D., 1985. The Past is a Foreign Country. Cambridge University Press, New York and Cambridge.
Moretti, F., 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. Verso, London and New York.
Rabb, T.K., 1983. The development of quantification in historical research. Journal of Interdisciplinary History 13, 591–601.
Reynolds, J.F., 1998. Do historians count anymore? Historical Methods 31, 141–148.
Ringer, F.K., 1979. Education and Society in Modern Europe. Indiana University Press, Bloomington.
Ruggles, S., 2012. The decline of quantitative history. Unpublished paper. Annual Meeting of the American Historical Association.
Stearns, P.N., 1985. Social history and history: a progress report. Journal of Social History 19, 319–334.
Steckel, R.H., 1995. Stature and the standard of living. Journal of Economic Literature 33, 1903–1940.
Steckel, R.H., Rose, J.C. (Eds.), 2002. The Backbone of History: Health and Nutrition in the Western Hemisphere. Cambridge University Press, New York.

Quantification in the History of the Social Sciences

Theodore M Porter, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The wide role of quantification in the modern world is not to be understood as the importation of objective methods from outside of history, but the cultivation of tools and concepts in close alliance with their contexts of use. While differentiated by field, quantitative methods have been shaped also by interactions across barriers of discipline, and even between scientific and political or administrative applications.

Quantification means, very broadly, the use of numbers to comprehend objects and events. While quantification is often identified with science, it has been equally significant for engineering, commerce, labor, government, professions, and human use of the natural environment. Quantitative methods had a crucial role in shaping the social and behavioral sciences, and remain central to them. The purpose of this article is not to supply a practical guide to quantitative tools in social science, but to examine their backgrounds and assess their broader significance from the standpoint of science studies and the history of science. In modern times, quantitative methods and the numbers they produce have assumed a powerful role in most societies as intersecting discourses of science, administration, and the public sphere. This article emphasizes the more public uses of quantification.

The broad category of quantification refers to several allied but distinct human activities. *Mathematization* means the characterization of objects and events in terms of mathematical relationships. While it is not the main topic of this article, it is of course relevant. The expression of scientific theories in mathematical terms has provided one important basis for relying on quantitative or numerical data. There have been other reasons of a technical, administrative, or moral kind. Among the more empirical forms of quantification, *measurement* grew out of such practical activities as surveying lands or buying and selling. *Counting*, the basic tool for assigning numbers to sets of discrete objects, was deployed for such tasks as inventorying one's possessions or tallying populations. *Calculation* refers loosely to methods for combining and manipulating numerical information into more readily useable forms. *Statistics*, a wide-ranging body of methods for analyzing quantitative information, is discussed historically in other articles in these volumes, and receives also some attention here.

Theories and Measures

As early as the 1660s, and more systematically from about 1830, social and economic investigators in the European world held up quantification as the proper method of science. The growth of university-based social disciplines from about 1880, and especially after 1930, was attended by a widespread faith that through numbers, social research could be made genuinely scientific. A generation or more of positivistically inclined researchers looked to systematic measurement as the proper foundation for solid, mathematized theory. Some compared its progress with the development of physics in the seventeenth century, and anticipated a breakthrough in the near future.

The most influential historical and philosophical writers on science in the postwar period took almost the opposite view. Alexandre Koyré (1968), whose philosophical studies of seventeenth-century science inspired the first generation of professional historians of science in the Anglophone world, challenged the supposition that indiscriminate data collection would ever lead to serious science. He doubted, for example, that Galileo had learned much about the laws of motion from rolling balls down inclined planes. The new kinematics applied not to real objects, he argued, but to abstract bodies in geometrical space. Galileo derived his law of falling bodies mainly by working out the logical consequences of what he already knew. This mathematical and metaphysical work was, for Koyré, the most important achievement of the new science. Indeed, only through theory had the measures attained scientific meaning. Thomas Kuhn (1961, 1976) reaffirmed Koyré's broad argument and provided a historical perspective that reached forward to the nineteenth century. The most successful sciences of the seventeenth-century 'Scientific Revolution,' including mechanics, astronomy, and geometrical optics, had already been mathematical in ancient times. The more Baconian experimental studies of the early modern period were comparatively ineffective until a 'Second Scientific Revolution,' beginning around 1800, when new scientific objects such as electricity and heat began to be comprehended mathematically.

Scholars in science studies since about 1980 have inclined to dethrone theory in favor of a more respectful attention to the practices and material culture of science. Accordingly, they take more seriously the quantification of phenomena that had not been grasped by exact laws. Just as the life of experiment has never been strictly subordinated to theory, so numbers have often functioned independently of higher mathematics. They are linked not just to theoretical science, but also, and perhaps even more consequentially, to economic and administrative changes. The history of ideas alone cannot grasp their full significance.

Quantities and Standards

Quantity is not first of all the language of science, imported as something alien into everyday affairs, but an indispensable part

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03227-X

of the fabric of life. Yet the aspiration to science has gone a long way toward reshaping the role of numbers and measures in the world. In this – as Max Weber suggested – we find much in common between bureaucratic and scientific rationality. Mathematics has often, and plausibly, been held up as an exemplar of universalized knowledge. The rules of arithmetic, or constants like π, are the same everywhere, at least in principle. Yet the universalism of measures and numbers, particularly when these have empirical content, has been achieved only within confined human spaces, and only by overcoming great obstacles. Even now, even in the industrialized West, quantities and prices are managed in everyday activities such as shopping by techniques quite different from those taught in schools (Lave, 1986). Before the rise of modern science, in the preindustrial world, quantification had very different meanings. Numbers were pervasive, but for most purposes their use and meanings were thoroughly local. The process through which they gained general validity is one of the crucial transformations of modern life.

A Plethora of Units

As Witold Kula (1986) has shown, measures remained immensely variable in Europe as late as the eighteenth century. This was partly by design, since small cities and lesser states often claimed the right to define their own weights and measures as a mark of autonomy and an obstacle to rule by central authorities. A diversity of measures, however, is the natural state of things wherever the pressures of commerce and administration have been insufficient to bring about unification. At a local level, authorities took some pains to reduce the ambiguity of measurement. A bushel measure, valid for the region, would often be fixed to the town hall, and could be used to settle disputes. In the next town, however, its volume would be slightly different, and over greater distances they varied considerably. The subtle geography of measures, moreover, was not defined exclusively by distance. Milk, oil, and wine came in different units; measures of cloth depended on the width of looms, and hence on the costliness of fabric and weave. The practices of measurement, too, were complexly nuanced. A bushel measure, even if the volume was fixed, could be tall and narrow or short and wide; the wheat might or might not be compacted by being poured in from a considerable height; and a complex set of conditions governed whether to heap the bushel, and by how much. Parties to a transaction often negotiated about measures rather than about prices. Add to this the consideration that almost nothing was decimalized, not even money, and it is evident that arithmetical calculation provided no fixed routine for managing affairs or settling disputes (Kula, 1986; Frängsmyr et al., 1990). The conversions required by trade provided an important source of employment for mathematicians.

The simplification and standardization of measures was encouraged by the expansion of long-distance trade, but was achieved primarily by governments in alliance with scientists. Among Europeans, it was the British who came nearest to creating uniformity of measures in the eighteenth century. A more systematic, almost Utopian, scheme, the metric system, was initiated in the 1790s under the French Revolution. In its most ambitious form this included a new calendar using months of 30 days divided into weeks of 10 days, and even a few decimalized clocks, beating out a hundred seconds per minute. The meter was set so that 10 million of them would reach from pole to equator, in the vain hope of integrating geodetic and astronomical measures with those on a human scale (Gillispie, 2004). To this end, bold expeditions were sent out into a disordered countryside to survey and measure a meridian. Although the meter was justified in part as a response to organized petitions (*cahiers de doléance*) calling for fixed measures as protection against seigniorial discretion, the highly rationalized metric system of decimalized conversions and strange Greek names – of milliliters, kilometers, and centigrams – went well beyond this. Not until the July Monarchy, after 1830, did the metric system begin to succeed even in France (Alder, 1995, 2002). In time, however, it has come to provide a common set of measures for most of the world, with the United States a notable, though partial, holdout.

A precise and maximally rigorous definition of units is a basic element in the infrastructure of modern life. The role of standardized measures in systems of manufacturing based on interchangeable parts is widely recognized. They are indispensable also for large-scale networks such as electric power distribution, telephony, and, by now, computers and the Internet. A collaboration of scientists and bureaucrats was required to create them. In some cases, as with the electrical units defined in the later nineteenth century, eminent scientists such as James Clerk Maxwell, William Thomson (Lord Kelvin), and Hermann von Helmholtz were deeply involved in the definition of standards.

Other units, far from the domain of scientific theory, play an indispensable role in the regulation of modern life. Environmental agencies, for example, deploy measures of effluent concentrations, to regulate pollutants. They depend on meticulous specification not only of instruments, but also of sampling methods, training of technicians, custody of materials, and so on. Since penalties and taxes often depend on them, they are always vulnerable to challenge. In this domain, as in many others, an immense apparatus of instrumentation and control lies behind what scientists and citizens have the luxury (usually) of regarding as 'raw,' or merely 'descriptive,' data.

Producing Uniformity

Units and instruments are not by themselves sufficient to quantify effectively. A quantifiable world contains reasonably homogeneous objects, objects whose measures can be compared, or that can be grouped together and counted. Nature and unregulated human labor produce a few of these, though real uniformity occurs mainly on a scale – for example, the molecular one – that is not readily accessible to the human senses. Physics was quantified partly in terms of abstractly 'mathematical' entities, such as forces, and partly through increasingly rigorous experimental controls, involving an expanding apparatus of instruments.

In the social world, where quantification is also a tool of administration, these same regulatory processes have helped to produce new quantifiable entities. The investigation and

compensation of joblessness, for example, solidified the category of 'the unemployed.' In most countries, the official definition of an unemployed person has been linked to a state obligation to provide benefits. To collect unemployment insurance requires that a person meet certain standards, which are more or less closely monitored, and to which persons who lack work are encouraged to conform. They also must enroll themselves, thereby adding their unit datum to the numbers. The point here is not that government offices cause people to be without work, though mechanisms of compensation will certainly affect the rates. It is that official categories sharpen definitions, and so help to create countable entities. Those entities are also vulnerable to manipulation: a looser definition will be popular among those collecting benefits while a tighter one will not only reduce costs, but in the same stroke will lower the official unemployment measure. Most human and social types are variable, or at least fuzzy at the margins (Hacking, 1995). Measurement, for example, of crime, ethnicity, or mental illness, is allied to processes that make them more uniform, and hence countable. Economic categories such as profits, investment, and trade are dependent on the processes through which they are recorded and regulated. Financial and other measures can be *performative* in the sense that a theoretical articulation may cause values in the world to conform more closely to them (MacKenzie, 2006). But there is another side to this dialectic, since the thinness of quantitative description can invite manipulation, leading sometimes to breakdown and even catastrophic failure of an order based on measurement.

The Authority of Calculation

Numbers and measures are not merely tools of description and analysis. They work also as guides to action, especially public action. An explicit and formal analysis, often largely quantitative in form, is by now required for a variety of public decisions involving investment or regulation. Business organizations depend increasingly on measures and projections in order to make decisions. These analyses are often expressed in money terms, and their modern expansion attests to the triumph of capitalism. But public reliance on economic quantification is by no means a mere imitation of its commercial uses. The tools of quantitative analysis were developed as much for government as for business purposes, and have become increasingly dependent on academic researchers. In the background is an evolving political and administrative culture, reflecting changes in the status and composition of elites.

Commensuration

It is characteristically modern to suppose that a conscientious decision should involve explicit consideration of the available data. Decision by calculation goes beyond this, requiring a course of action to be dictated 'objectively' by facts. One prototype of this ideal of public reason is benefit–cost analysis (BCA). In practice, BCA involves predictions in many domains, including weather patterns, health, hydrology, ecology, and commerce, but its most distinctive aspect is the attempt to convert a great diversity of considerations into a common financial metric. The obstacles to commensuration are most compelling in regard to such considerations as human lives lost or saved, the disappearance of biological populations, or the degradation of a beautiful landscape. In public debate, such exchanges may seem profoundly immoral. Yet commensuration flourishes particularly as a strategy of simplifying, and perhaps of ranking, for lay people trying to make choices regarding complex problems for which they lack real access to specialized tools of knowledge (Espeland and Stevens, 1998).

Economics provides a rationale for much commensuration, but BCA did not develop out of theoretical economics. It was imposed by the U.S. Congress and by the exigencies of legal and administrative challenges during a massive expansion of federal water projects beginning in the 1930s. Engineers, not economists, initially faced the immense difficulties of incorporating life, health, recreation, and scenic beauty along with economic development as they proffered measures of the costs and benefits of water projects. Despite political opposition to some of these measures, a political logic of rationalized choice made them necessary nonetheless. The task was to devise methods that would hold up to administrative and legal challenge. In this context, what was not quantified tended to disappear, and the experts found it necessary to put aside what seemed the implausibility or even immorality of expressing everything, including lives saved or lost and traditional ties to the land, as a sum of money. Beginning in the 1950s, economists and social scientists were recruited into agencies like the U.S. Army Corps of Engineers and the Bureau of Reclamation to provide more consistent or more acceptable solutions to these thorny problems. Finally, during the 1960s, BCA emerged as an applied field within welfare economics (Porter, 1995; Espeland, 1998).

Trust and Impersonality

Academic problems of social measurement and commensuration are thus embedded in social and political conditions that put a premium on scientific objectivity in the specific sense of freedom from personal or self-interested bias. Since David Hume, at least, facts have often been sharply distinguished from values, and for some purposes the credibility of science is enhanced by its separation from their 'subjective' domain. The ambition to provide a scientific, possibly quantitative, basis for public decisions remained quixotic or utopian so long as an aristocratic political class retained power. Few would assent now to Frédéric Le Play's assessment that numbers became necessary as a second-best option under democratic governments, in the wake of revolution, when leaders no longer possessed the inherited wisdom of birth and experience (Porter, 2011). Yet the aspiration to replace judgment with objective, quantitative rules grew up in situations where elite authority had become suspect.

The move by American state engineers to an increasingly rigorous benefit-cost methodology was a response to systematic challenge, and exemplifies this point. So also do other exemplary tools of 'mechanical objectivity.' Social and medical researchers have been criticized for preferring the impersonal rigor of the statistically significant result over proper measures of quantities that matter for health and

wealth (Ziliak and McCloskey, 2008). Accountants did not willingly abandon what many saw as their professional responsibility to offer an expert assessment of the financial condition of firms. These choices may be understood as a sacrifice of meaning for the sake of quantified objectivity, whose ascent is part of a broad cultural and political history of the twentieth century. Some basic features of that history would include the growth of government, bureaucratic conflict, resistance to administrative secrecy, and public distrust of 'bureaucracy.' The unprecedented quantitative enthusiasm of postwar social science involved, along with its optimism about the 'methods' of science, an ethic of renunciation that more and more was to be anchored in rules of method rather than high moral character (Porter, 2004).

A Historical Perspective on Social Science Quantification

Reflective accounts about the social and behavioral sciences have often supposed that these fields have short histories, and that the pursuit of quantification means following the path of natural science triumphant. But the rise of a quantitative ethos in natural science can be dated no earlier than the early eighteenth century. Systematic measurement until then was almost confined to astronomy, and with it astrology, which extended over every inhabited continent. For a long time, astronomical measurement was geometrical and not numerical. Its purposes were practical as well as scientific: to survey land, navigate ships, calculate calendars, and horoscopes. Barometers and thermometers helped to extend the domain of measurement into meteorology during the seventeenth century. Measurement in early modern Europe was associated as closely with natural history as with experimental physics. By 1750, savants bearing instruments wandered the Earth measuring mountains and other natural wonders, conjoining precision to their appreciation of nature. The development of an exact science of electricity in the later eighteenth century marked the advent of systematic experimental quantification (Heilbron, 1979). Measurement had also become routine in technological endeavors, such as mineral assaying, decades before Lavoisier and his contemporaries began to insist on its centrality to chemistry.

From this perspective, it would be difficult to argue that political and economic studies lagged behind natural science in their reliance on numbers. Among the early members of the Royal Society of London, founded in 1660, were William Petty and John Graunt. Petty announced a new science of 'political arithmetic,' which in his hands involved speculative estimates of population and of wealth. Graunt is noted for his more strictly empirical compilations from the London bills of mortality. Just a few decades later, Jan de Wit, the Dutch Stadtholder (magistrate), used rudimentary probability theory to assign rates for the sale of annuities, and the astronomer Edmond Halley calculated the first mortality table. Political arithmetic, which entailed the compilation as well as the analysis of demographic and medical records, flourished in the eighteenth century, and public health was among the earliest and most influential of the quantitative studies. At the same time, probability theory was pursued as a mathematical theory of belief and credibility (Hacking, 1975; Daston, 1988; Rusnock, 2002). The *philosophe* and mathematician Condorcet, a leader of the French Revolution who then was hunted down by it, hoped to develop probability into the theoretical basis for elections and judicial decisions (Baker, 1975).

The Nineteenth Century and the Rise of Statistics

By the early nineteenth century, as Kuhn observed, the natural historical investigation of heat, electricity, and the physics of light had given way to sciences that were both mathematical and quantitative. Possibly the analysis of mortality records for insurance can be regarded as comparable in certain respects. But the explosive growth of quantification beginning in the 1820s, what Ian Hacking (1990) has called an 'avalanche of printed numbers,' was not concentrated in the experimental sciences. Rather, it involved systematic surveys of stellar positions, terrestrial magnetism, weather phenomena, tides, plant and animal distributions, and social statistics.

The collection and investigation of social numbers had expanded gradually over the course of the eighteenth century. Few eighteenth-century states had the bureaucratic means to conduct a full census, and most were disinclined to release such sensitive information to their enemies or even to their subjects. In many places, too, the people would have viewed census officials as harbingers of new taxes or military conscription, and in Britain the census was blocked by a gentry class that refused to be counted. In contrast to piecemeal efforts at home, there were some systematic surveys abroad, the machinery of centralized state information gathering having been turned loose first of all in colonial situations. Among European states, Sweden pioneered the regular census, in 1749; its results remained secret until 1764. The French used sampling, of a sort, and enlisted mathematicians such as Pierre Simon Laplace to estimate the population. The U.S. census of 1790 heralded a new era of regular, public censuses, which were encouraged in Europe by French military conquest and by the pressure of wartime mobilization. The British instituted their decennial census in 1801. Finally in the 1830s, bureaucratic professionals began to take over the collection of social and economic numbers in the most advanced European states.

At the same time, an increasing pace of quantitative study by voluntary organizations made 'statistics' a plausible 'science of society.' One model for it was developed by the Belgian Adolphe Quetelet, whose quantitative pursuits ranged from observational astronomy and meteorology to the bureaucratic organization of census work. He argued for an alliance of data gathering with mathematical probability, in the quest to uncover 'social laws.' Apart from life insurance, however, most statistical collection was under the control of reformers and administrators with little mathematical knowledge. Their enthusiasm for counting was perhaps indiscriminate, yet it was nurtured by specific anxieties about poverty, sanitation, epidemic disease, economic dislocation, and urban unrest. Social investigation, still more than natural science, was an object of contest, and there were active efforts to organize a social science to speak for working people (Yeo, 1996). Karl Marx's immense statistical labors for his *Capital*, drawing mainly from official British inquiries whose

results he praised as honest and reliable, testify impressively to the credibility of public numbers. Middle-class organizations had more staying power than working-class ones, and better standing to claim the mantle of 'social science.' In Britain, the Statistical Society of London (founded 1834) joined bureaucrats with professionals and even aristocratic political leaders. In Germany, statistics became a university field, which, from about 1860 to 1900, combined empirical quantification with a historicized 'national economics,' in a quest for state action on behalf of workers and peasants. Some of the founders of social science in America were brought up in this tradition, as was Max Weber. Emile Durkheim, too, engaged with it, as his *Suicide* of 1897 plainly attests (Porter, 1986).

The nineteenth-century science of statistics embraced much of the subject matter of the modern social sciences. It overlapped, and sometimes competed, with political economy, political science, geography, anthropology, ethnology, sociology, and demography. These stood for genres, more or less distinct, though not for anything so definite as modern professional disciplines. Their quantitative methods were not identical, but neither were they sharply differentiated from each other or from measurement in biology. There was much uncertainty and debate from 1830 through the end of the century as to whether statistics was a science, with its own subject matter, or only a method. It was resolutely empirical, to the point that some considered sampling too speculative. Far from being allied to any mathematical social science, statistics was more often at war with deduction and abstract theory. Some, like the economist William Stanley Jevons, would claim that political economy should be mathematical because it was inherently quantitative, but he could not attach the mathematical formulations he introduced to empirical numbers.

Quantification in the Modern Social and Behavioral Sciences

Toward the end of the nineteenth century, a more mathematical field of statistics began to develop. The most influential form of the new statistics was British biometric version created by Francis Galton and Karl Pearson (Stigler, 1986). At almost the same time, however, distinct social and behavioral sciences began to crystallize, in part around somewhat distinct quantitative methods. All were conceived as ways to enhance the scientific standing of their fields, and each was engaged also with issues of politics, policy, and the management of economies and populations. The typology to follow is intended to be suggestive, and certainly cannot be exhaustive.

Physical Anthropology and Eugenics

The measurement of humans was carried out for bureaucratic as well as scientific purposes throughout the nineteenth century. The increasing concern with race in anthropology was linked to a preoccupation with skulls, whose measures were used to distinguish distinct 'races' of Europeans. What Francis Galton called 'correlation' was developed to measure the interconnections among physical human characteristics, their perpetuation from one generation to another, and their relation to measures of merit or achievement.

Education and Psychology

Correlation provided also a solution to a problem of contemporary psychophysics, related to education. Gustav Theodor Fechner began applying the methods of error theory to the measurement of sensory discrimination in the mid-nineteenth century. Systematic use of experimental randomization was developed for studies of this kind (Dehue, 1997). They provided also a wide field of application for correlational methods. Did childhood study of dead languages really train the mind for speedier or more thorough learning of other subjects, such as science, history, or literature? The growth of mental testing in the early twentieth century posed a more general question of correlations. To what extent did students who proved themselves adept in some particular study excel also in others? Charles Spearman's *g*, or general intelligence, and the 'intelligence quotient' or IQ, presupposed a basic unity of the mental faculties. Psychometricians and educational psychologists developed a body of statistical methods to measure this intelligence, or to decompose it, and to assess the relations of its components. The measurement of intelligence assumed an important role in sorting institutionalized populations of students, soldiers, and the like (Zenderland, 1998; Carson, 2007).

From the 1930s, as psychology turned resolutely experimental, quantification became mandatory. Once again, educational psychology as well as parapsychology took the lead in the assimilation of novel methods of experimental design and analysis. The new paradigm of a proper experiment involved a randomized design of control and experimental populations, yielding results that would, in most cases, be subjected to analysis of variance. A worthwhile result was defined in part by a test of significance: the 'null hypothesis' that the treatment had no effect should be rejected at some specified level, commonly 0.05. This was the methodology of the English statistician R. A. Fisher, and for half a century or more beginning in the 1920s his name was practically synonymous with modern statistics (Gigerenzer et al., 1989).

Econometrics and Economics

Economic numbers were preeminent among the concerns of statistics from the eighteenth century. From the late nineteenth century, new tools grew up with new quantified objects such as unemployment and the cost of living (Stapleford, 2009). A new economic mathematics of marginal utility appeared in the 1870s and was self-consciously reconciled to the classical tradition in the 1890s by Alfred Marshall. His work included abundant graphical illustrations, but these presented curves that could not readily be measured. The sources of econometrics are largely distinct. From origins in the 1870s to the establishment of a society and a journal in the 1930s, the field focused on the study of business cycles (Morgan, 1990). Econometricians took their numbers where they could get them, often from public sources. They were highly dependent on national accounts, prepared by governments, and indeed participated in structuring these accounts. The paradigmatic statistical tool of econometrics, adapted from biometrics mainly after 1950, was regression analysis.

Surveys

From the founding of statistical societies in the 1830s, social science has relied on surveys to understand social problems and learn about how the lives and opinions of ordinary people. Charles Booth's monumental studies of East London beginning in the 1880s inspired a movement of surveys of towns and regions in Britain and America. For most of the nineteenth century, statisticians typically aimed to count exhaustively, but then began relying more and more on 'representative' sampling, which was valued for its speed and adaptability. The origins of survey sampling are partly conceptual, but were tied to specific practical problems, as in late imperial Russian efforts by officials, sometimes allied to mathematicians, to manage budgets for communities of recently freed serfs. Arthur Bowley and then Jerzy Neyman worked to develop rigorous sampling methods that would permit generalization using probability theory. The techniques of survey sampling have linked sociology and political science to marketing and political campaigns (Desrosières, 1998; Mespoulet, 2001; Igo, 2007).

Medicine

The statistics of public health, including the treatment of mental illness, was long focused mainly on gathering information and presenting it in tables. Pearson's biometric statistics began the formalization of experimental and quasi-experimental methods, which were reshaped according to a new theoretical logic by Fisher. Health officials institutionalized the randomized, controlled trial in medicine from the 1940s. It soon came to define the 'gold standard' in experimental medicine. More broadly, statistical data and analyses, including 'metastatistical' methods for combining the results of different studies, have become exemplary not just of 'evidence-based medicine,' but of evidence itself in its most public, bureaucratically acceptable form (Marks, 1997; Jorland et al., 2005).

Conclusion

Quantification is as central to social as to natural science. In an era of vast interlinked computational systems and almost infinite data banks, its public prominence and its secret manipulations are more impressive than ever. The expansion of quantification over the last three centuries reveals some key characteristics of the social and behavioral disciplines, among them, the drive for the status of science, the role of applied studies in forming academic ones, the shaping of knowledge by data technologies, and the often problematic relationship of theoretical to empirical work. The quantification of human science draws from centuries of history, and is by no means a mere product of the modern effort to attain scientific standing. Indeed, the advance of measurement and statistical analysis in social science is not distinctively academic, but has been stimulated by administrative and political demands. Yet quantification in social science has, for two centuries, been enveloped in a certain mystique, regarded as an indication of rigor or at least of scientific maturity. We should understand it not as the key to all mysteries, but as a powerful set of tools and concepts, to be integrated as much as possible with theoretical understanding and with other ways of comprehending social phenomena.

See also: Applied Social Research, History of; Computers: Impact on the Social Sciences; Demography, Early History of; Empirical Social Research, History of; Experimentation in Psychology, History of; Intelligence: History of the Concept; Political Science, History of; Positivism, History of; Psychology, History of (Twentieth Century); Quantification in History; Sample Survey Methodology, History of; Science and Technology Studies, History of.

Bibliography

Alder, K., 1995. A revolution to measure: the political economy of the metric system in France. In: Wise, N. (Ed.), The Values of Precision. Princeton University Press, Princeton, pp. 39–71.
Alder, K., 2002. The Measure of All Things. The Seven-Year Odyssey and Hidden Error that Transformed the World. Free Press, New York.
Baker, K., 1975. Condorcet: From Natural Philosophy to Social Mathematics. University of Chicago Press, Chicago.
Carson, J., 2007. The Measure of Merit: Talents, Intelligence and Inequality in the French and American Republics. Princeton University Press, Princeton.
Daston, L., 1988. Classical Probability in the Enlightenment. Princeton University Press, Princeton.
Dehue, T., 1997. Deception, efficiency, and random groups: psychology and the gradual origination of the random group design. Isis 88, 653–673.
Desrosières, A., 1998. The Politics of Large Numbers. Harvard University Press, Cambridge, MA.
Espeland, W., 1998. The Struggle for Water. University of Chicago Press, Chicago.
Espeland, W., Stevens, M., 1998. Commensuration as a social process. Annual Review of Sociology 24, 313–343.
Frängsmyr, T., Heilbron, J., Rider, R. (Eds.), 1990. The Quantifying Spirit in the Eighteenth Century. University of California Press, Berkeley.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L., 1989. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press, Cambridge.
Gillispie, C., 2004. Science and Polity in France: The Revolutionary and Napoleonic Years. Princeton University Press, Princeton.
Hacking, I., 1975. The Emergence of Probability. Cambridge University Press, Cambridge.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge.
Hacking, I., 1995. Rewriting the Soul. Princeton University Press, Princeton.
Heilbron, J., 1979. Electricity in the 17th and 18th Centuries: A Study in Early Modern Physics. University of California Press, Berkeley.
Igo, S., 2007. The Averaged American: Surveys Citizens, and the Making of a Mass Public. Harvard University Press, Cambridge.
Jorland, J., Opinel, A., Weisz, G. (Eds.), 2005. Body Counts. Medical Quantification in Historical and Sociological Perspectives. McGill-Queens University Press, Montreal and Kingston.
Koyré, A., 1968. Metaphysics and Measurement. Harvard University Press, Cambridge, Mass (Original papers 1943–1960).
Kuhn, T., 1961. The function of measurement in modern physical science. In: Woolf, H. (Ed.), Measurement. Bobbs-Merrill, Indianapolis, pp. 31–63.
Kuhn, T., 1976. Mathematical versus experimental traditions in the development of physical science. Journal of Interdisciplinary History 7, 1–31. Reprinted in: Kuhn, 1977. The Essential Tension. University of Chicago Press, Chicago, pp. 31–65.
Kula, W., 1986. Measures and Men (R. Szreter, Trans.). Princeton University Press, Princeton.
Lave, J., 1986. The values of quantification. In: Law, J. (Ed.), Power, Action, and Belief: A New Sociology of Knowledge? Routledge, London, pp. 88–111.
MacKenzie, D., 2006. An Engine Not a Camera: How Financial Models Shape Markets. MIT Press, Cambridge.
Marks, H., 1997. The Progress of Experiment: Science and Therapeutic Reform in the United States 1900–1990. Cambridge University Press, Cambridge.
Mespoulet, M., 2001. Statistique et revolution en Russie: Un compromis impossible (1880–1930). Presses Universitaires de Rennes, Rennes.
Morgan, M., 1990. The History of Econometric Ideas. Cambridge University Press, Cambridge.

Porter, T., 1986. The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton.
Porter, T., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton.
Porter, T., 2004. Karl Pearson: The Scientific Life in a Statistical Age. Princeton University Press, Princeton.
Porter, T., 2011. Reforming vision: the engineer Le Play learns to observe society sagely. In: Daston, L., Lunbeck, E. (Eds.), Histories of Scientific Observation. University of Chicago Press, Chicago, pp. 281–302.
Rusnock, A., 2002. Vital Accounts: Quantifying Health and Population in Eighteenth-Century England and France. Cambridge University Press, Cambridge.
Stapleford, T., 2009. The Cost of Living in America: A Political History of Economic Statistics, 1880–2000. Cambridge University Press, Cambridge.
Stigler, S., 1986. The History of Statistics: The Measurement of Uncertainty to 1900. Harvard University Press, Cambridge.
Yeo, E., 1996. The Contest for Social Science. Rivers Oram Press, London.
Zenderland, L., 1998. Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing. Cambridge University Press, Cambridge.
Ziliak, S., McCloskey, D., 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor.

Quantifiers

David Gil, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

Published by Elsevier Ltd.

Abstract

Quantifiers are freestanding expressions whose meanings involve the notion of quantity. The basic semantic structure of quantification is bipartite, consisting of the quantifier itself plus the expression that it quantifies. Quantification has traditionally been of great interest to semanticists, logicians, and philosophers of language, due at least in part to the perceived 'logical' or 'mathematical' nature of the meanings involved. However, in different languages, such meanings are often expressed by means of very different morphosyntactic strategies. Quantifiers may be classified in terms of their inherent meanings, as count or mass; as existential, midrange, or universal; as strong or weak; and as increasing, decreasing, or neither. They may also be classified in terms of their role within the sentence, as nominal or verbal, as continuous or discontinuous, and as simple or scopal. In general, quantifiers do not constitute a coherent grammatical category, either within individual languages or crosslinguistically. Although prototypical instances of quantifiers are easily identifiable, there are many cases in which a given expression straddles the boundary of what counts as a quantifier, either because it is not clearly freestanding or because its meaning shades off into other domains not generally characterized as quantificational.

Introduction

Quantifiers are freestanding expressions whose meanings involve the notion of quantity, such as, in English, *three*, *several*, *numerous*, *most*, *every*, *one hundred and twenty three*, *all but seventeen*, and so forth.

The basic semantic structure of quantification is bipartite, consisting of the quantifier itself plus the expression that it quantifies. For example, in a sentence such as *John saw three cats*, *three* is the quantifier, and *cats* the quantified expression.

Quantification has traditionally been of great interest to semanticists, logicians, and philosophers of language, due at least in part to the perceived 'logical' or 'mathematical' nature of the meanings involved. As such, it is striking to observe how such basic and seemingly immutable meanings may, in different languages, be expressed with very different morphosyntactic strategies, exhibiting a great degree of crosslinguistic variation. Some seminal recent studies of quantification include Keenan and Westerståhl (2011), Westerståhl (2011), and Keenan and Paperno (2012).

Quantifiers may be classified internally, with respect to their morphosyntactic structure and basic semantic properties. In addition, they may be classified externally, with regard to the syntactic and semantic relationships between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

Internal Classifications

Internal classifications of quantification make reference to a variety of fundamental morphosyntactic and semantic properties.

Mass and Count Quantifiers

One of the most basic distinctions is between mass and count quantifiers (Pelletier, 1979; Allan, 1980). Mass quantifiers constitute expressions which denote an undifferentiated homogeneous mass; for example, English *much* is a mass quantifier because it forms expressions such as much milk, much cake. In contrast, count quantifiers constitute expressions which refer to one or more countable units of characteristic size and shape; for example, English *many* is a count quantifier because it forms expressions such as *many cats*, *many cakes*. Whereas many quantifiers belong to just one of the two types, mass or count, some quantifiers are undifferentiated with respect to the mass/count distinction and may appear in expressions of either kind.

An interesting case is that of the so-called 'classifier languages,' many of which are located in the East Asian linguistic region (Craig, 1986; see example (4) below). In such languages, many or all of the count quantifiers cannot occur in immediate construction with the noun that they quantify; instead, the quantifier must occur in construction with a classifier, and the quantifier-plus-classifier constituent may then occur in construction with the noun. Typically, the classifier makes reference to various characteristics – size, shape, function, and so forth – of the quantified noun. Accordingly, different nouns often require different classifiers. However, in some cases, different classifiers may occur with the same noun, sometimes resulting in subtle differences in meaning (*see* Classifiers, Linguistics of).

Existential and Universal Quantifiers

Of particular interest to logicians are the existential and universal quantifiers (Vendler, 1967; Heim, 1982; Gil, 1995). Existential quantifiers, such as English *some* and *a*(*n*), form expressions denoting at least a minimal, nonzero amount or number of the quantified expression, for example, *some water*, *a boy*. Universal quantifiers, such as English *all* and *every*, form expressions referring to an exhaustive amount or number of the quantified expression, for example, *all the water*, *every boy*. Existential and universal quantifiers are semantically related to each other through negation by means of the classical 'square of opposition': thus, in English, *Not a boy came* is equivalent

to *All the boys did not come*, and *Not all the boys came* is equivalent to *A boy did not come*.

Existential quantification, in many languages, is inextricably intertwined with singular number and/or indefiniteness: in some cases they are expressed identically, while in other cases they are expressed by means of the same marker occurring in different constructions, or by means of similar though not identical markers. Universal quantification, too, is expressed in a variety of different ways across languages. Some languages, such as English, do have 'dedicated' universal quantifiers, that is to say, words such as *all* and *every*, whose primary or exclusive function is the expression of universal quantification. However, many other languages lack such words, instead making use of more complex circumlocutions if and when the expressive need arises.

Given the importance of existential and universal quantifiers in mathematical logic, it is worthy of note how relatively unimportant these quantifiers are within the grammatical systems of languages. It is indeed not all that common to find a single form, in any language, whose interpretation corresponds precisely and unproblematically to either the existential or the universal quantifier. Moreover, there is probably no language in which expressions of existential and universal quantification constitute a natural grammatical class to the exclusion of other expressions. The diversity of morphosyntactic expression of existential and universal quantifiers across the world's languages suggests that, although greatly valued by logicians, they may not be endowed with privileged status vis-à-vis the structure of language.

Midrange Quantifiers

'In between' the existential and the universal quantifiers lie a variety of midrange quantifiers, such as, in English, *few*, *several*, *many*, *most*, and so forth. Among the midrange quantifiers, a particularly privileged category is that of cardinal numerals – quantifiers which refer to natural numbers, for example, *one*, *two*, *thirteen*, *one hundred and twenty seven* (Menninger, 1969; Hurford, 1975). Complex phrasal numerals often make use of syntactic patterns available elsewhere in the language; however, in addition, they sometimes exhibit unique structural features not found in other grammatical domains. For example, in Georgian, phrasal numerals are exceptional in that they are the only multiword expressions that may undergo reduplication (thereby forming distributive-share numerals of the kind discussed in the Section Scope and Scopal Quantifiers below).

Other Internal Classifications

Quantifiers are sometimes classified as either *strong* (alternatively referred to as definite) or *weak* (indefinite) (Milsark, 1977). Strong quantifiers, such as English *all* and *most*, are inherently definite in their meaning, whereas weak quantifiers, such as English *some* and *three*, are not. Strong and weak quantifiers often exhibit different syntactic behavior; for example, strong quantifiers are typically awkward or ungrammatical in existential sentences, e.g., *There is all the water on the floor*.

Some further semantic distinctions are captured by the categories of *increasing* (alternatively referred to as upward entailing) and *decreasing* (or downward entailing) (Keenan and Stavi, 1986). A quantifier is increasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is larger than that explicitly stipulated by the quantifier, for example, *at least three*. Conversely, a quantifier is decreasing if a sentence containing it would be logically true in a situation in which the amount or number of quantified expressions involved is smaller than that explicitly stipulated by the quantifier, for example, *at most three*. Other quantifiers are neither increasing nor decreasing, for example, *exactly three*.

An interesting problem is posed by simple monomorphemic quantifiers such as *three*. At first blush, such quantifiers appear to be neither increasing nor decreasing: if somebody says *There are three cats on the bed*, the mental picture that is conjured up is, of course, one of three cats, not two, or four, or any other number. However, a moment's reflection will reveal an important asymmetry. If in fact there are exactly two cats on the bed, then the above sentence is clearly false. In contrast, if there are exactly four cats on the bed, then, in most contexts, the above sentence would obviously be misleading and inappropriate – but it would not, strictly speaking, be false. Thus, while the pragmatic appropriateness conditions of *three* resemble those of *exactly three*, the actual truth conditions of *three* are identical to those of *at least three*. Hence, *three* is, in fact, increasing.

External Classifications

External classifications of quantification make reference to the morphosyntactic and semantic relationships which obtain between the quantifiers and the quantified expressions and/or the remainder of the sentences in which they occur.

Nominal and Verbal Quantifiers

One basic semantic classification relates to the category of the quantified expression, specifically whether it is of a nominal or verbal nature (Gil, 1993). For example, in the English sentence *John saw three cats*, the numeral *three* quantifies a noun; accordingly this sentence illustrates nominal quantification. In contrast, in the sentence *John sang three times*, the numeral phrase *three times* quantifies a verb; hence this sentence instantiates verbal quantification.

Up to a point, nominal and verbal quantification exhibit parallel grammatical properties. Obvious, but still worth noting, is the fact that in both of the preceding sentences, the same numeral, *three*, is used. However, nominal and verbal quantification also differ in many ways; typically, verbal quantification is more highly marked than its nominal counterpart.

Continuous and Discontinuous Quantifiers

A distinct albeit related syntactic classification pertains to the position of the quantifier with respect to the quantified expression, specifically, whether or not the two form a continuous constituent (Gil, 1993; Bach et al., 1995). Sentences (1)–(4) below all instantiate nominal quantification, in which a numeral quantifies an expression referring to an animal;

however, each of the examples exhibits a different syntactic configuration.

(1) HEBREW

raiti	[*šaloš*	arnavot]
see-PAST:1:SG	three-F	rabbit-PL:F

'I saw three rabbits'

(2) TAGALOG

Tatlo	[*ang*	*nakita*	
three	TOP	PAT.TOP:PFV-see	
kong	*baboy*]		
DIR:1:SG-LIG	pig		

'I saw three pigs'
('The pigs seen by me were three')

(3) KUTENAI (Matthew Dryer, p.c.)

hu	[*qă są*	*wu:kati*]	*cupqa*
1:SG	three-PRV	see-IND	deer

'I saw three deer' ('I three-saw deer')

(4) JAPANESE

Watasi	*wa*	*inu*	o
1:SG	TOP	dog	ACC
sanbiki	*mita*		
three-CLF	see-PFV		

'I saw three dogs' ('I saw dogs threely')

In modern Hebrew, (1), as in its English gloss, the quantifier groups with the quantified expression, *šaloš arnavot*, to the exclusion of the remainder of the sentence; in Tagalog, (2), the quantifier *tatlo* stands alone as a sister to the remainder of the sentence, which forms a constituent; in Kutenai, (3), the quantifier occurs in construction with the verb, *qăsă wu:kati*, to the exclusion of the remainder of the sentence; while in Japanese, (4), the quantifier *sanbiki* is a sentential adverb not forming an obvious constituent with any part of the remainder of the sentence.

Constructions (1) and (2) exemplify continuous quantification, in which the quantifier and the quantified expression form a constituent. In contrast, constructions (3) and (4) exemplify discontinuous quantification, in which the quantifier occurs apart from the quantified expression. Examples (1)–(4) above illustrate the distinction between continuous and discontinuous quantification with respect to nominal quantification. In principle, a similar distinction may also be drawn with respect to verbal quantification, though examples of discontinuous verbal quantification are more hard to come by.

Scope and Scopal Quantifiers

When two quantified expressions are present in the same construction, a variety of semantic relations may obtain between the two expressions (Jackendoff, 1971; Gil, 1982b). The following English sentence, containing two numerically quantified NPs, may, at least potentially, be interpreted in any or all of the following four ways:

(5) Three boys saw two girls

(a) 'Each of three boys saw each of two girls' not scope differentiated: strong symmetric.

(b) 'Three boys saw two girls between them' not scope differentiated: weak symmetric.

(c) 'Three boys saw two girls each' scope differentiated: wide scope for *three* boys distributive key: *three boys*; distributive share: *two girls*.

(d) 'Two girls were seen by three boys each' scope differentiated: wide scope for *two girls* distributive key: *two girls*; distributive share: *three boys*.

Interpretations (5a) and (5b) both involve a single set of three boys and a single set of two girls: each of the two NPs has independent reference. These interpretations are accordingly characterized as non-scope-differentiated or symmetric. Interpretations (5c) and (5d) present a more complex picture. In (5c) there is a single set of three boys; however, each of the three boys is associated with a different set of two girls. Thus, while the subject NP has independent reference, the direct-object NP is referentially dependent on the subject NP. In this interpretation, then, the subject NP has scope over the direct-object NP. More precisely, the two NPs are in a relationship of distributivity, with the subject NP as *distributive key*, and the direct-object NP as *distributive share*. In (5d), a mirror-image situation obtains. Here, there is a single set of two girls; however, each of the two girls is associated with a different set of three boys. Accordingly, while the direct-object NP has independent reference, the subject NP is referentially dependent on the direct-object NP; the direct-object NP has scope over the subject NP. Here too then, a relation of distributivity obtains; this time, though, it is the direct-object NP which is the distributive key and the subject NP which is the distributive share.

The above four interpretations are not equally readily available for speakers of English. Specifically, the symmetric interpretations (5a) and (5b) are more easily available than the asymmetric interpretations (5c) and (5d). Moreover, between the latter two interpretations, (5c) with wide scope for the subject NP is more readily available than (5d) with wide scope for the direct-object NP. These preferences are not accidental; rather, they are particular consequences of more general principles governing the assignment of quantifier scope across languages.

Given the general preference for symmetric interpretations, languages typically have, at their disposal, various lexical and morphosyntactic devices whose function is to induce the more marked asymmetric interpretations, when these are required. Consider the following English sentences:

(6) (a) *All* the men carried three suitcases.

(b) *Every* man carried three suitcases.

Sentences (6a) and (6b) differ in the lexical choice of quantifier within the subject NP. Whereas in (6a), with *all*, a variety of scope relationships may obtain between the two NPs, and

the number of suitcases may vary between three in total and three per man, in (6b), with *every*, the subject NP has scope over the direct-object NP, and the suitcases number three per man. Whereas *all*, like *three*, is a simple quantifier, *every* is a *scopal* quantifier, that is to say, a quantifier endowed with an additional denotational component, one that forces a particular scope relationship to obtain between the NP in which it occurs, and some other expression. In particular, it is a distributive-key quantifier, inducing a relationship of distributivity, in which the NP containing the quantifier in question is the distributive key, or element with wide scope (Gil, 1995).

A different kind of scopal quantifier is exemplified by the following pair of Maricopa sentences:

(7) Maricopa

(a) *ʔipač* ʔii *xmokm*
man-PL:NOM stick 3-three-SG-DS
upaavk
3-carry-PL-REAL
'The men carried three sticks'

(b) *ʔipač* *ʔii* *xmokxperm*
man-PL:NOM stick 3-three-SG-DIST.SHARE-DS
upaavk
3-carry-PL-REAL
(i) 'The men carried three sticks each'
(ii) 'The men carried the sticks in threes'

Sentences (7a) and (7b) differ only in the form of the numeral within the direct-object expression. Sentence (7a) contains a simple numeral, *xmok-* 'three,' and is ambiguous in roughly the same ways as its English gloss. In contrast, sentence (7b) contains the same numeral plus the suffix *-xper*, whose effect is to mark the numeral *xmok-* and the expression containing it as distributive share. The ambiguity exhibited by sentence (7b) results from different expressions being selected as the distributive key counterpart. Under interpretation (7b(i)), the subject NP *ʔipač* 'men' is chosen as the distributive key: the sticks accordingly number three per man. Under interpretation (7b(ii)), the verb *upaavk* 'carried' is chosen as the distributive key: here the sticks accordingly number three per carrying. Crosslinguistically, distributive-share numerals are formed by a variety of formal processes, including affixation (as in the above example), periphrasis, reduplication, and suppletion (Gil, 1982a).

Quantifier as a Grammatical Category

A quantifier is primarily a semantic category: an expression is a quantifier only if it bears certain meaning properties. One of the most striking characteristics of quantifiers is the absence of any distinctive shared formal morphosyntactic features, which might justify the positing of a corresponding grammatical category of quantifier. Within languages, different quantifiers often display different grammatical properties. And across languages, the 'same' quantifier may exhibit different morphosyntactic behavior as well. Thus, in terms of their formal properties, quantifiers constitute a heterogeneous collection, exhibiting a great amount of morphological and syntactic diversity.

Looking within individual languages, there is probably no language within which there is a formal grammatical category consisting exactly of all quantifiers but no other expressions. Rather, in most or all languages, different quantifiers exhibit different arrays of grammatical properties, some grouping together with expressions belonging to one category, others patterning together with expressions belonging to some other category, and yet others exhibiting idiosyncratic morphological or syntactic behavior.

Looking across languages, quantifiers with the same meaning often exhibit quite different morphological and syntactic behavior. Examples (1)–(4) above showed how the numeral 'three,' as a nominal quantifier, may occur, in different languages, in a variety of different constructions, exhibiting continuous and discontinuous quantification of various kinds. Even if attention is restricted to the most well-known type, that of continuous NP-internal quantification, as in (l), a great amount of crosslinguistic variation may be observed. In particular, in different languages, nominal quantifiers in attributive NP internal position may resemble, to various degrees, nouns, adjectives, and/or verbs.

Delimitation of Quantification

The discussion in the preceding sections was based on a bipartite definition of quantifiers as 'freestanding expressions' whose meanings 'involve the notion of quantity.' Like many similar definitions, this one is useful, but not unproblematical. Specifically, in addition to prototypical quantifiers such as those discussed above, there are numerous expressions and construction types for which it is difficult to adjudicate whether or not they may appropriately be characterized as quantificational.

The first part of the definition is formal: in order for an expression to be considered as a quantifier, it must be freestanding – a word or a phrase, rather than an affix or some other bound morphological unit. Accordingly, various kinds of clitics, particles, and other intermediate units smaller than words but larger than affixes may provide borderline cases straddling the boundary between what is a quantifier and what is not. The distinction between free and bound forms is what underlies the distinction between quantification and the semantically related grammatical category of *number*. Specifically, whereas quantifiers are freestanding, number markers are bound forms. However, in some instances it may be difficult to adjudicate whether a form associated with a quantificational meaning is free or bound, and hence whether it constitutes an instance of quantification or of number. In order to determine whether forms such as these are quantifiers or number markers, it may be necessary to make use of additional diagnostic criteria distinguishing quantifiers from number markers.

The second part of the definition of quantifier is the obvious semantic condition: its meaning must involve the notion of quantity. While in many instances this is quite straightforward, in other cases, quantificational meanings may shade off gradually into other kinds of meanings not generally considered to be quantificational. One class of expressions related in meaning to

nominal quantifiers are those denoting size and other scalar properties. An important class of expressions related to verbal quantifiers are those marking aspectual categories, such as perfect and progressive. More generally, different quantifiers in different languages may be related to an extremely broad range of nonquantificational concepts, synchronically and/or diachronically.

See also: Classifiers, Linguistics of; Lexical Semantics; Logic and Linguistics; Parts of Speech; Semantics.

Bibliography

Allan, K., 1980. Nouns and countability. Language 56, 541–567.
Bach, E., Jeline, K.E., Kratzer, A., Partee, B.H. (Eds.), 1995. Cross-linguistic Quantification. Kluwer Academic Publishers, Dordrecht, The Netherlands.
Craig, C. (Ed.), 1986. Typological Studies in Language: 7. Noun Classes and Categorization.
Gil, D., 1982a. Distributive Numerals. Ph.D. dissertation. University of California, Los Angeles, CA.
Gil, D., 1982b. Quantifier scope, linguistic variation, and natural language semantics. Linguistics and Philosophy 5, 421–472.
Gil, D., 1993. Nominal and verbal quantification. Sprachtypologie und Universalienforschung 46, 275–317.
Gil, D., 1995. Universal quantifiers and distributivity. In: Bach, E., Jeline, K.E., Kratzer, A., Partee, B.H. (Eds.), Quantification in Natural Languages. Kluwer, Dordrecht, The Netherlands, pp. 321–362.
Heim, I., 1982. The Semantics of Definite and Indefinite Noun Phrases. Ph.D. dissertation. University of Massachusetts, Amherst, MA.
Hurford, J.R., 1975. The Linguistic Study of Numerals. Cambridge University Press, Cambridge, UK.
Jackendoff, R.S., 1971. Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA.
Keenan, E.L., Stavi, J., 1986. A semantic characterization of natural language determiners. Linguistics and Philosophy 9, 253–326.
Keenan, E.L., Westerståhl, D., 2011. Generalized quantifiers in linguistics and logic. In: van Benthem, J., ter Meulen, A. (Eds.), Handbook of Logic and Language. Elsevier, Amsterdam, pp. 837–893.
Keenan, E.L., Paperno, D. (Eds.), 2012. Handbook of Quantifiers in Natural Language. Springer, Berlin.
Menninger, K., 1969. Number Words and Number Symbols (P. Broneer, Trans.). MIT Press, Cambridge, MA.
Milsark, G., 1977. Toward an explanation of certain peculiarities in the existential construction in English. Linguistic Analysis 3, 1–30.
Pelletier, F.J. (Ed.), 1979. Mass Terms: Some Philosophical Problems. Synthese Language Library, Reidel, Dordrecht, The Netherlands.
Vendler, Z., 1967. Linguistics and Philosophy. Cornell University Press, Ithaca, NY.
Westerståhl, D., 2011. Generalized quantifiers. In: Zalta, E.N. (Ed.), Stanford Encyclopedia of Philosophy.

Quantile Regression

Roger Koenker, University of Illinois, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Classical least squares regression may be viewed as a natural way of extending the idea of estimating an unconditional mean parameter to the problem of estimating conditional mean *functions*; the crucial link is the formulation of an optimization problem that encompasses both problems. Likewise, quantile regression offers an extension of univariate quantile estimation to estimation of conditional quantile functions via an optimization of a piecewise linear objective function in the residuals. Median regression is a special case and minimizes the sum of absolute residuals, an idea introduced by Boscovich in the eighteenth century.

The asymptotic theory of quantile regression closely parallels the theory of the univariate sample quantiles. Computation of quantile regression estimators may be formulated as a linear programing problem and efficiently solved by simplex or barrier methods. A close link to rank-based inference has been forged from the theory of the dual regression quantile process, or regression rankscore process.

Quantile regression is a statistical technique intended to estimate, and conduct inference about, conditional quantile functions. Just as classical, linear regression methods based on minimizing sums of squared residuals enable one to estimate models for conditional mean functions, quantile regression methods offer a mechanism for estimating models for the conditional median function, and the full range of other conditional quantile functions. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables.

Quantile regression has been used in a broad range of application settings. Reference growth curves for childrens' height and weight have a long history in pediatric medicine; quantile regression methods may be used to estimate upper and lower quantile reference curves as a function of age, sex, and other covariates without imposing stringent parametric assumptions on the relationships among these curves. See, e.g., Wei et al. (2005). Quantile regression methods have been widely used in economics to study determinants of wages, discrimination effects, and trends in income inequality. Several recent studies have modeled the performance of public school students on standardized exams as a function of socioeconomic characteristics like their parents' income and educational attainment, and policy variables like class size, school expenditures, and teacher qualifications. It seems rather implausible that such covariate effects should all act so as to shift the entire distribution of test results by a fixed amount. It is of obvious interest to know whether policy interventions alter performance of the strongest students in the same way that weaker students are affected. Such questions are naturally investigated within the quantile regression framework.

In ecology, theory often suggests how observable covariates affect limiting sustainable population sizes, and quantile regression has been used to directly estimate models for upper quantiles of the conditional distribution rather than inferring such relationships from models based on conditional central tendency. In survival analysis, and event history analysis more generally, there is often also a desire to focus attention on particular segments of the conditional distribution, for example, survival prospects of the oldest-old, without the imposition of global distributional assumptions.

Quantiles, Ranks, and Optimization

We say that a student scores at the τth quantile of a standardized exam, if he performs *better* than the proportion τ, and *worse* than the proportion $(1-\tau)$, of the reference group of students. Thus, half of the students perform better than the median student, and half perform worse. Similarly, the quartiles divide the population into four segments with equal proportions of the population in each segment. The quintiles divide the population into five equal segments; the deciles into 10 equal parts. The quantile, or percentile, refers to the general case.

More formally, any real valued random variable, Y, may be characterized by its distribution function,

$$F(y) = \text{Prob}(Y \le y)$$

while for any $0 < \tau < 1$,

$$Q(\tau) = \inf\{y : F(y) \ge \tau\}$$

is called the τth quantile of X. The median, $Q(1/2)$, plays the central role. Like the distribution function, the quantile function provides a complete characterization of the random variable, Y.

The quantiles may be formulated as the solution to a simple optimization problem. For any $0 < \tau < 1$, define the piecewise linear 'check function,' $\rho_\tau(u) = u(\tau - I(u < 0))$ illustrated in Figure 1. Minimizing the expectation of $\rho_\tau(Y - \xi)$ with respect to ξ yields solutions, $\xi(\tau)$, the smallest of which is $Q(\tau)$ defined above.

The sample analogue of $Q(\tau)$, based on a random sample, $\{y_1, \ldots, y_n\}$, of Y's, is called the τth *sample* quantile, and may be found by solving,

$$\min_{\xi \in R} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

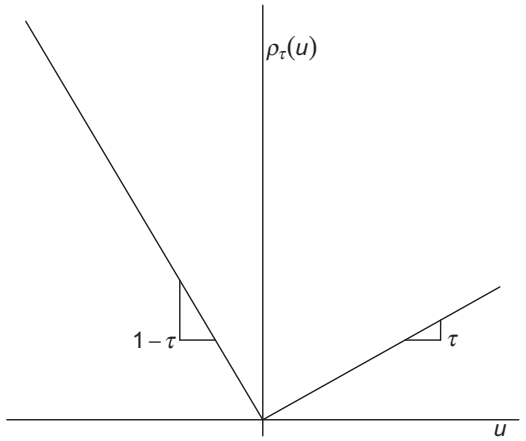

Figure 1 Quantile regression ρ function.

While it is more common to define the sample quantiles in terms of the order statistics, $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, constituting a sorted rearrangement of the original sample, their formulation as a minimization problem has the advantage that it yields a natural generalization of the quantiles to the regression context.

Just as the idea of estimating the unconditional mean, viewed as the minimizer,

$$\widehat{\mu} = \text{argmin}_{\mu \in \mathbb{R}} \sum (y_i - \mu)^2$$

can be extended to estimation of the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving,

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum \left(y_i - x_i'\beta\right)^2$$

the linear conditional quantile function, $Q_Y(\tau|X = x) = x_i'\beta(\tau)$, can be estimated by solving,

$$\widehat{\beta}(\tau) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum \rho_\tau(y_i - x'\beta)$$

The median case, $\tau = 1/2$, which is equivalent to minimizing the sum of *absolute* values of the residuals has a long history. In the mid-eighteenth century, Boscovich proposed estimating a bivariate linear model for the ellipticity of the earth by minimizing the sum of absolute values of residuals subject to the condition that the mean residual took the value zero. Subsequent work by Laplace characterized Boscovich's estimate of the slope parameter as a weighted median and derived its asymptotic distribution. F.Y. Edgeworth seems to have been the first to suggest a general formulation of median regression involving a multivariate vector of explanatory variables, a technique he called the 'plural median.' The extension to quantiles other than the median was introduced in Koenker and Bassett (1978).

An Example

To illustrate the approach, we may consider an analysis of a simple first-order autoregressive model for maximum daily temperature in Melbourne, Australia. The data are taken from Hyndman et al. (1996). In **Figure 2**, we provide a scatter plot of

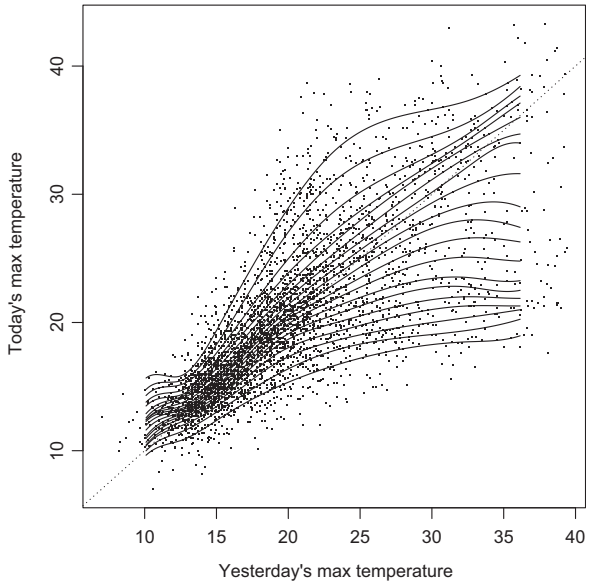

Figure 2 Melbourne maximum daily temperature: the plot illustrates 10 years of daily maximum temperature data (in degrees centigrade) for Melbourne, Australia as an AR(1) scatterplot. The data are scattered around the (dashed) 45° line suggesting that today is roughly similar to yesterday. Superimposed on the scatterplot are estimated conditional quantile functions for the quantiles $\tau \in \{0.05, 0.10, \ldots, 0.95\}$. Note that when yesterday's temperature is high the spacing between adjacent quantile curves is narrower around the 45° line and at about 20 °C than it is in the intermediate region. This suggests bimodality of the conditional density in the summer.

10 years of daily temperature data: today's maximum daily temperature is plotted against yesterday's maximum. Our first observation from the plot is that there is a strong tendency for data to cluster along the (dashed) 45° line implying that with high probability today's maximum is near yesterday's maximum. But closer examination of the plot reveals that this impression is based primarily on the left side of the plot, where the central tendency of the scatter follows the 45° line very closely. On the right side, however, corresponding to summer conditions, the pattern is more complicated. There, it appears that *either* there is another hot day, falling again along the 45° line, *or* there is a dramatic cooling off. But a mild cooling off appears to be more rare. In the language of conditional densities, if today is hot, tomorrow's temperature appears to be bimodal with one mode roughly centered at today's maximum, and the other mode centered at about 20°.

Several estimated quantile regression curves have been superimposed on the scatterplot. Each curve is specified as a linear B-spline. Under winter conditions these curves are bunched around the 45° line, however, in the summer it appears that the upper quantile curves are bunched around the 45° line and around 20°. In the intermediate temperatures, the spacing of the quantile curves is somewhat greater indicating lower probability of this temperature range. This impression is strengthened by considering a sequence of density plots based on the quantile regression estimates. Given a family of reasonably densely spaced estimated conditional quantile functions, it is straightforward to estimate the conditional density of the response at

various values of the conditioning covariate. In Figure 3, we illustrate this approach with several of density estimates based on the Melbourne data. Conditioning on a low previous day temperature we see a nice unimodal conditional density for the following day's maximum temperature, but as the previous day's temperature increases we see a tendency for the lower tail to lengthen and eventually for very warm summer days we see a clearly bimodal density. In this example, the classical regression assumption that the covariates affect only the location of the response distribution, but not its scale or shape, is clearly violated.

Interpretation of Quantile Regression

Least squares estimation of mean regression models asks the question, 'How does the conditional mean of Y depend on the covariates X?' Quantile regression asks this question at each quantile of the conditional distribution enabling one to obtain a more complete description of how the conditional distribution of Y given $X = x$ depends on x. Rather than assuming that covariates shift only the location or scale of the conditional distribution, quantile regression methods enable one to explore potential effects on the shape of the distribution as well. Thus, for example, the effect of a job training program on the length of participants' current unemployment spell might be to lengthen the shortest spells while dramatically reducing the probability of very long spells. The mean treatment effect in such circumstances might be small, but the treatment effect on the shape of the distribution of unemployment durations could, nevertheless, be quite significant.

Quantile Treatment Effects

The simplest formulation of quantile regression is the two-sample treatment–control model. In place of the classical Fisherian experimental design model in which the treatment induces a simple location shift of the response distribution, Lehmann (1974) proposed the following general model of treatment response:

> *Suppose* the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be x. Then the distribution G of the treatment responses is that of the random variable $X + \Delta(X)$ where X is distributed according to F.

Special cases obviously include the location shift model, $\Delta(X) = \Delta_0$, and the scale shift model, $\Delta(X) = \Delta_0 X$, but the general case is natural within the quantile regression paradigm.

Doksum (1974) shows that if $\Delta(x)$ is defined as the 'horizontal distance' between F and G at x, so

$$F(x) = G(x + \Delta(x))$$

then $\Delta(x)$ is uniquely defined and can be expressed as

$$\Delta(x) = G^{-1}(F(x)) - x$$

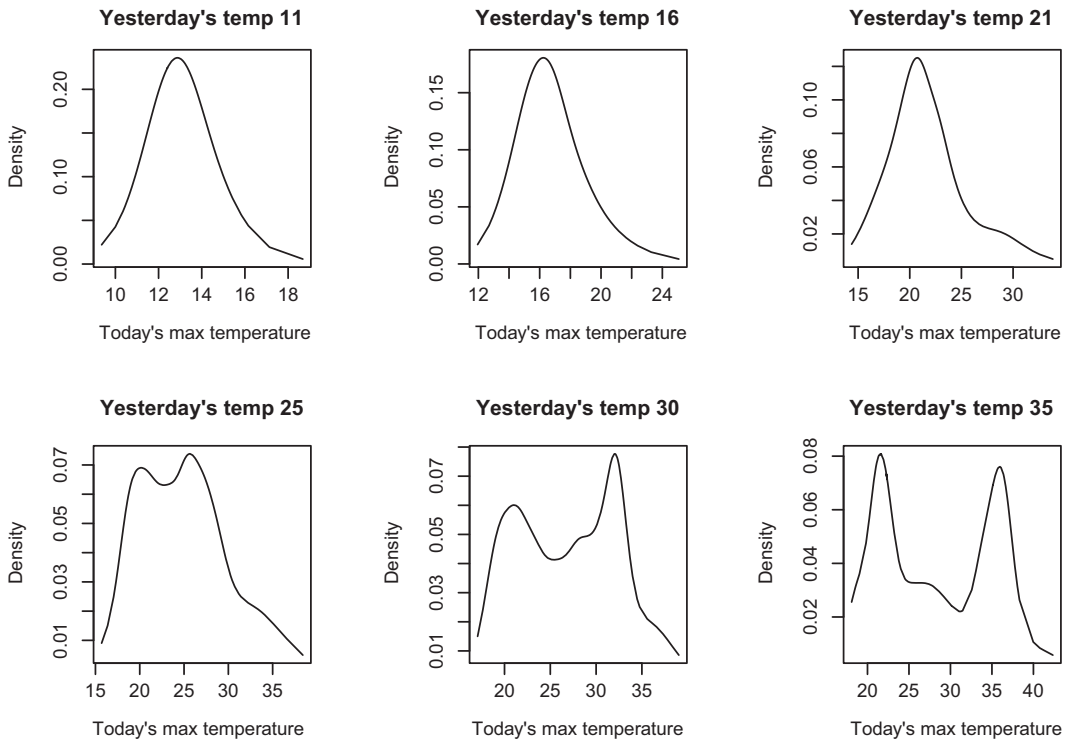

Figure 3 Conditional density estimates of today's maximum temperature for several values of yesterday's maximum temperature based on the Melbourne data: these density estimates are based on a kernel smoothing of the conditional quantile estimates as illustrated in the previous figure using 99 distinct quantiles. Note that temperature is bimodal when yesterday was hot.

Changing variables so $\tau = F(x)$ one may define the *quantile treatment effect*,

$$\delta(\tau) = \Delta\left(F^{-1}(\tau)\right) = G^{-1}(\tau) - F^{-1}(\tau)$$

In the two-sample setting, this quantity is naturally estimable by

$$\widehat{\delta}(\tau) = \widehat{G}_n^{-1}(\tau) - \widehat{F}_m^{-1}(\tau)$$

where G_n and F_m denote the empirical distribution functions of the treatment and control observations, based on n and m observations, respectively.

Formulating the quantile regression model for the binary treatment problem as,

$$Q_{Yi}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i$$

where D_i denotes the treatment indicator, with $D_i = 1$ indicating treatment, $D_i = 0$, control, then the quantile treatment effect can be estimated by solving,

$$\left(\widehat{\alpha}(\tau), \widehat{\delta}(\tau)\right)' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \delta D_i)$$

The solution $(\widehat{\alpha}(\tau), \widehat{\delta}(\tau))'$ yields $\widehat{\alpha}(\tau) = \widehat{F}_n^{-1}(\tau)$, corresponding to the control sample, and

$$\widehat{\delta}(\tau) = \widehat{G}_n^{-1}(\tau) - \widehat{F}_n^{-1}(\tau)$$

Doksum suggests that one may interpret control subjects in terms of a latent characteristic: for example, in survival analysis applications, a control subject may be called frail if he is prone to die at an early age, and robust if he is prone to die at an advanced age. This latent characteristic is thus implicitly indexed by τ, the quantile of the survival distribution at which the subject would appear if untreated, i.e., $(Y_i|D_i = 0) = \alpha(\tau)$. And the treatment, under the Lehmann model, is assumed to alter the subjects control response, $\alpha(\tau)$, making it $\alpha(\tau) + \delta(\tau)$ under the treatment. If the latent characteristic, say, the propensity for longevity, were observable *ex ante*, then one could view the treatment effect $\delta(\tau)$ as an explicit interaction with this observable variable. In the absence of such an observable variable, however, the quantile treatment effect may be regarded as a natural measure of the treatment response.

It may be noted that the quantile treatment effect is intimately tied to the two-sample QQ-plot, which has a long history as a graphical diagnostic device. The function $\widehat{\Delta}(x) = G_n^{-1}(F_m(x)) - x$ is exactly what is plotted in the traditional two-sample QQ-plot. If F and G are identical then the function $G_n^{-1}(F_m(x))$ will lie along the 45° line; if they differ only by a location scale shift, then $G_n^{-1}(F_m(x))$ will lie along another line with intercept and slope determined by the location and scale shift, respectively. Quantile regression may be seen as a means of extending the two-sample QQ-plot and related methods to general regression settings with continuous covariates.

When the treatment variable takes more than two values, the Lehmann–Doksum quantile treatment effect requires only minor reinterpretation. If the treatment variable is continuous as, for example, in dose–response studies, then it is natural to consider the assumption that its effect is linear, and write,

$$Q_{Yi}(\tau|x_i) = \alpha(\tau) + \beta(\tau)x_i$$

We assume thereby that the treatment effect, $\beta(\tau)$, of changing x from x_0 to $x_0 + 1$ is the same as the treatment effect of an alteration of x from x_1 to $x_1 + 1$. Note that this notion of the quantile treatment effect measures, for each τ, the change in the response required to stay on the τth conditional quantile function. As in the Melbourne example, nonlinear conditional quantile functions can be estimated by simply expanding covariates in spline basis functions.

Transformation Equivariance of Quantile Regression

An important property of the quantile regression model is that, for any monotone function, $h(\cdot)$,

$$Q_{h(T)}(\tau|x) = h(Q_T(\tau|x))$$

This follows immediately from observing that

$$\text{prob}(T < t|x) = \text{prob}(h(T) < h(t)|x)$$

This equivariance to monotone transformations of the conditional quantile function is a crucial feature, allowing one to decouple the potentially conflicting objectives of transformations of the response variable. This equivariance property is in direct contrast to the inherent conflicts in estimating transformation models for conditional mean relationships. Since, in general, $E(h(T)|x) \neq h(E(T|x))$ the transformation alters in a fundamental way what is being estimated in ordinary least squares regression.

A particularly important application of this equivariance result, and one that has proven extremely influential in the econometric application of quantile regression, involves censoring of the observed response variable. The simplest model of censoring may be formulated as follows. Let y_i^* denote a latent (unobservable) response assumed to be generated from the linear model

$$y_i^* = x_i'\beta + u_i \quad i = 1, \ldots, n$$

with $\{u_i\}$ independent and identically distributed (iid) from distribution function F. Due to censoring, the y_i^*'s are not observed directly, but instead one observe

$$y_i = \max\{0, y_i^*\}$$

Powell (1986) noted that the equivariance of the quantiles to monotone transformations implied that in this model the conditional quantile functions of the response depended only on the censoring point, but were independent of F. Formally, the τth conditional quantile function of the observed response, y_i, in this model may be expressed as

$$Q_i(\tau|x_i) = \max\{0, x_i'\beta + F_u^{-1}(\tau)\}$$

The parameters of the conditional quantile functions may now be estimated by solving

$$\min_b \sum_{i=1}^{n} \rho_\tau\left(y_i - \max\{0, x_i'b\}\right)$$

where it is assumed that the design vectors x_i, contain an intercept to absorb the additive effect of $F_u^{-1}(\tau)$. This model is computationally somewhat more demanding than conventional linear quantile regression because it is nonlinear in parameters.

Robustness

Robustness to distributional assumptions is an important consideration throughout statistics, so it is important to emphasize that quantile regression inherits certain robustness properties of the ordinary sample quantiles. The estimates and the associated inference apparatus have an inherent distribution-free character since quantile estimation is influenced only by the local behavior of the conditional distribution of the response near the specified quantile. Given a solution $\widehat{\beta}(\tau)$, based on observations, $\{y, X\}$, as long as one does not alter the sign of the residuals, *any* of the y observations may be arbitrary altered without altering the initial solution. Given the x's only the signs of the residuals matter in determining the quantile regression estimates, and thus outlying responses influence the fit in so far as they are either above or below the fitted hyperplane, but how far above or below is irrelevant. While quantile regression estimates are inherently robust to contamination of the response observations, they can be quite sensitive to contamination of the design observations, $\{x_i\}$. Several proposals have been made to ameliorate this effect.

Computational Aspects of Quantile Regression

Although it was recognized by a number of early authors, including Gauss, that solutions to the median regression problem were characterized by an exact fit through p sample observations when p linear parameters are estimated, no effective algorithm arose until the development of linear programing in the 1940s. It was then quickly recognized that the median regression problem could be formulated as a linear program, and the simplex method employed to solve it. The algorithm of Barrodale and Roberts (1974) provided the first efficient implementation specifically designed for median regression and is still widely used in statistical software. It can be concisely described as follows. At each step, we have a trial set of p 'basic observations' whose exact fit may constitute a solution. We compute the directional derivative of the objective function in each of the $2p$ directions that correspond to removing one of the current basic observations, and taking either a positive or negative step. If none of these directional derivatives are negative the solution has been found, otherwise one chooses the most negative, the direction of steepest descent, and goes in that direction until the objective function ceases to decrease. This one dimensional search can be formulated as a problem of finding the solution to a scalar weighted quantile problem. Having chosen the step length, we have in effect determined a new observation to enter the basic set, a simplex pivot occurs to update the current solution, and the iteration continues.

This modified simplex strategy is highly effective on problems with a modest number of observations, achieving speeds comparable to the corresponding least squares solutions. But for larger problems with, say $n > 100\,000$ observations, the simplex approach eventually becomes considerably slower than least squares. For large problems recent development of interior point methods for linear programing problems are highly effective. Portnoy and Koenker (1997) describe an approach that combines some statistical preprocessing with interior point methods and achieves comparable performance to least squares solutions even in very large problems.

An important feature of the linear programing formulation of quantile regression is that the entire range of solutions for $\tau \in (0, 1)$ can be efficiently computed by parametric programing.

At any solution $\widehat{\beta}(\tau_0)$ there is an interval of τ's over which this solution remains optimal, it is straightforward to compute the endpoints of this interval, and thus one can solve iteratively for the entire sample path $\widehat{\beta}(\tau)$ by making one simplex pivot at each of the endpoints of these intervals.

Statistical Inference for Quantile Regression

The asymptotic behavior of the quantile regression process $\{\widehat{\beta}(\tau) : \tau \in (0, 1)\}$ closely parallels the theory of ordinary sample quantiles in the one sample problem. Koenker and Bassett (1978) show that in the classical linear model,

$$y_i = x_i\beta + u_i$$

with u_i iid from *df* F, with density $f(u) > 0$ on its support $\{u|0 < F(u) < 1\}$, the joint distribution of $\sqrt{n}(\widehat{\beta}_n(\tau_i) - \beta(\tau))_{i-1}^m$ is asymptotically normal with mean 0 and covariance matrix $\Omega \otimes D^{-1}$. Here $\beta(\tau) = \beta + F_u^{-1}(\tau)e_1$, $e_1 = (1, 0, \ldots, 0)'$, $x_{1i} \equiv 1, n^{-1}\sum x_i x_i' \to D$, a positive definite matrix, and

$$\Omega = \left(w_{ij} = \left(\min\{\tau_i, \tau_j\} - \tau_i\tau_j\right)\right) / \left(f\left(F^{-1}(\tau_i)\right) f\left(F^{-1}(\tau_j)\right)\right)$$

When the response is conditionally independent over i, but not identically distributed, the asymptotic covariance matrix of $\xi(\tau) = \sqrt{n}(\widehat{\beta}(\tau) - \beta(\tau))$ is somewhat more complicated. Let

$$\xi_i(\tau) = x_i\beta(\tau)$$

denote the conditional quantile function of y given x_i, and $f_i(\cdot)$ the corresponding conditional density, and define,

$$J_n(\tau_1, \tau_2) = (\min\{\tau_1, \tau_2\} - \tau_1\tau_2)n^{-1}\sum_{i=1}^{n} x_i x_i'$$

and

$$H_n(\tau) = n^{-1}\sum x_i x_i' f_i(\xi_i(\tau))$$

Under mild regularity conditions on the $\{f_i\}$'s and $\{x_i\}$'s, we have joint asymptotic normality for vectors $(\xi(\tau_i), \ldots, \xi(\tau_m))$ with mean zero and covariance matrix

$$V_n = \left(H_n(\tau_i)^{-1} J_n(\tau_i, \tau_j) H_n(\tau_j)^{-1}\right)_{i=1}^m$$

An important link to the classical theory of rank tests was made by Gutenbrunner and Jurečková (1992), who showed that the rankscore functions of Hájek and Šidák (1967) could be viewed as a special case of a more general formulation for the linear quantile regression model. The formal dual of the quantile regression linear programing problem may be expressed as,

$$\max\{y'a | X'a = (1 - \tau)X'1, \ a \in [0, 1]^n\}$$

The dual solution $\widehat{a}(\tau)$ reduces to the Hájek and Šidák rankscore process when the design matrix, X, takes the simple form of an n vector of ones. The regression rankscore process $\widehat{a}(\tau)$ behaves asymptotically much like the classical univariate

rankscore process, and thus offers a way to extend many rank-based inference procedures to the more general regression context.

Extensions and Future Developments

There is considerable scope for further development of quantile regression methods. Applications to survival analysis and time-series modeling seem particularly attractive, where censoring and recursive estimation pose, respectively, interesting challenges. Portnoy (2003) and Peng and Huang (2014) provide analogues of the Kaplan–Meier and Nelson–Aalan estimators for accelerated failure time type models. This work generalizes Powell's approach to fixed censoring to handle conventional random censoring and their promising extensions to competing risk and recurrent events models.

Model selection and applications to high dimensional data have also attracted considerable interest. This is natural given the close connection the lasso (ℓ_1, total variation) penalty methods that have proven to be highly influential throughout statistics, see, e.g., Belloni and Chernozhukov (2011) and Zou and Yuan (2008).

Nonparametric estimation of conditional quantile functions has developed very rapidly: Chaudhuri (1991) introduced local polynomial methods, Koenker et al. (1994) suggested total variation penalty methods, Koenker (2011) provides an overview of this literature. There is an emerging literature on instrumental variable methods for quantile regression for models with endogenous covariate effects. Building on early work of Chernozhukov and Hansen (2005) and Chesher (2003), this work also has implications for errors in variables and missing data methods for quantile regression, Wei and Carroll (2009).

There is a growing literature on time-series applications of quantile regression; Koenker and Xiao (2004) describe some results on quantile autoregression models, and there is considerable work on stochastic volatility models and density forecasting primarily for financial applications. Frequency domain methods have also been considered in work by Li (2012) and Hagemann (2011). Longitudinal data analysis has also received considerable recent attention, see, e.g., Koenker (2004) and Galvao, Jr. (2011).

Quantile regression for discrete response models remains a serious challenge. Machado and Silva (2005) have proposed convenient methods for count data. For binary response, Kordas (2014) has extended the Manski (1975) maximum score estimator, but computational issues remain an impediment.

There is no general agreement about how to generalize univariate quantiles to multivariate settings, but this has perhaps encouraged a variety of new proposals, see, e.g., Wei (2008), Kong and Mizera (2012), and Hallin et al. (2010) among others.

Conclusion

Classical least squares regression may be viewed as a natural way of extending the idea of estimating an unconditional mean parameter to the problem of estimating conditional mean *functions*; the crucial step is the formulation of an optimization problem that encompasses both problems. Likewise, quantile regression offers an extension of univariate quantile estimation to estimation of conditional quantile functions via an optimization of a piecewise linear objective function in the residuals. Median regression minimizes the sum of absolute residuals, an idea introduced by Boscovich in the eighteenth century.

The asymptotic theory of quantile regression closely parallels the theory of the univariate sample quantiles; computation of quantile regression estimators may be formulated as a linear programing problem and efficiently solved by simplex or barrier methods. A close link to rank-based inference has been forged from the theory of the dual regression quantile process, or regression rankscore process.

Nontechnical introductions to quantile regression are provided by Cade and Noon (2003), Hao and Naiman (2007), and Koenker and Hallock (2001) from the perspective of ecology, sociology, and economics, respectively. A more complete introduction is provided by the monograph of Koenker (2005). Most of the major statistical computing languages now include some capabilities for quantile regression estimation and inference. The quantile regression package *quantreg* for the R language is available from the CRAN archive at http://cran.r-project.org. Stata provides some limited quantile regression estimation and inference functions. SAS also offers some facilities for quantile regression.

See also: Linear Hypothesis: Regression (Basics).

Bibliography

Barrodale, I., Roberts, F., 1974. Solution of an overdetermined system of equations in the ℓ_1 norm. Communications of the ACM 17, 319–320.

Belloni, A., Chernozhukov, V., 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics 39, 82–130.

Cade, B., Noon, B., 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1, 412–420.

Chaudhuri, P., 1991. Global nonparametric estimation of conditional quantile functions and their derivatives. Journal of Multivariate Analysis 39, 246–269.

Chernozhukov, V., Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica 73 (1), 245–261.

Chesher, A., 2003. Identification in nonseparable models. Econometrica 71, 1405–1441.

Doksum, K., 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. Annals of Statistics 2, 267–277.

Galvao Jr., A.F., 2011. Quantile regression for dynamic panel data with fixed effects. Journal of Econometrics 164, 142–157.

Gutenbrunner, C., Jurečková, J., 1992. Regression quantile and regression rank score process in the linear model and derived statistics. Annals of Statistics 20, 305–330.

Hagemann, A., 2011. Robust Spectral Analysis arXiv preprint arXiv:1111.1965.

Hájek, J., Šidák, Z., 1967. Theory of Rank Tests. Academia, Prague.

Hallin, M., Paindaveine, D., Šiman, M., 2010. Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth. The Annals of Statistics 38, 635–669.

Hao, L., Naiman, D.Q., 2007. Quantile Regression. Sage, Thousand Oaks.

Hyndman, R.J., Bashtannyk, D.M., Grunwald, G.K., 1996. Estimating and visualizing conditional densities. Journal of Computational and Graphical Statistics 5, 315–336.

Koenker, R., 2004. Quantile regression for longitudinal data. Journal of Multivariate Analysis 91, 74–89.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, London.
Koenker, R., 2011. Additive models for quantile regression: model selection and confidence bandaids. Brazilian Journal of Probability and Statistics 25, 239–262.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Koenker, R., Hallock, K., 2001. Quantile regression. Journal of Economic Perspectives 15, 143–156.
Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.
Koenker, R., Xiao, Z., 2004. Unit root quantile autoregression inference. Journal of the American Statistical Association 94, 775–787.
Kong, L., Mizera, I., 2012. Quantile tomography: using quantiles with multivariate data. Statistica Sinica 22, 1589–1610.
Kordas, G., 2014. Smoothed binary regression quantiles. Journal of Applied Econometrics 21, 387–407.
Lehmann, E., 1974. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco.
Li, T.-H., 2012. Quantile periodograms. Journal of the American Statistical Association 107, 765–776.
Machado, J.A.F., Santos-Silva, J.S., 2005. Quantiles for counts. Journal of the American Statistical Association 100, 1226–1237.
Manski, C., 1975. Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics 3, 205–228.
Peng, L., Huang, Y., 2014. Survival analysis with quantile regression models. Journal of the American Statistical Association 103, 637–649.
Portnoy, S., 2003. Censored quantile regression. Journal of the American Statistical Association 98, 1001–1012.
Portnoy, S., Koenker, R., 1997. The Gaussian hare and the Laplacian tortoise: computability of squared error versus absolute-error estimators, with discussion. Statistical Science 12, 279–300.
Powell, J., 1986. Censored regression quantiles. Journal of Econometric 32, 143–155.
Wei, Y., 2008. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. Journal of the American Statistical Association 103, 397–409.
Wei, Y., Carroll, R.J., 2009. Quantile regression with measurement error. Journal of the American Statistical Association 104, 1129–1143.
Wei, Y., Pere, A., Koenker, R., He, X., 2005. Quantile regression for reference growth charts. Statistics in Medicine 25, 1369–1382.
Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108–1126.

Quantitative Cross-national Research Methods

Gosta Esping-Andersen, Universitat Pompeu Fabra, Barcelona, Spain
Adam Przeworski, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12649–12655, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Quantitative cross-national comparisons usually are based on smaller *N*'s. This implies that theory needs to be stronger and that counterfactuals need to be made explicit. Bayesian estimation is, in this situation, an attractive possibility. Because dependent variables are often categorical or limited, it is often preferable to use nonlinear models, like logit and probit. Selection bias is potentially acute due to nonrandom sampling. This is a problem of identification that usually cannot be overcome by quasi-experimental designs. It calls, again, for strong counterfactual thinking. Cross-national comparisons often cannot assume independence between observations on a variable (nations form 'families'), and this provokes biased coefficients and likely heteroskadisticity. Finally, the article examines the endogeneity problem (*X* is influenced by *Y*, or both jointly by an unobserved *Z*) which is potentially serious in cross-national comparisons because the meaning of a variable is embedded in the nation's entire history. The bias can derive from variable omission in which case correction calls for added controls. The article also examines the tradeoffs between large-*N* and smaller-*N* comparisons in terms of generalization and context, and stresses the need for stronger theory the smaller the number of cases being sampled.

The Main Challenges of Quantitative Nation Comparisons

Quantitative cross-national comparisons have become an important feature of all the social sciences. Their popularity has been spurred, in part, by the increasing availability of good data for a growing number of countries and, in part, by the diffusion of statistical skills among researchers. There are many constraints and potential pitfalls involved in cross-national statistical analyses, and these will be the main theme of this article. But this research strategy has, regardless, proven itself of great scientific value. It has, above all, helped highlight how fundamental many international variations are, thus, providing a powerful check against parochialism and unwarranted generalizations. Quantitative macrocomparisons have also proven themselves eminently useful in testing the validity of core theoretical assumptions, such as the convergence thesis that has long dominated development and modernization theory in political science and sociology especially.

Quantitative nation comparisons will, however, inevitably face a number of tradeoffs; some more serious than others. One is that the capacity for broader generalization that quantification permits is bought at the price of less attention to the contextual reality of individual nations. The uniqueness of a nation's culture, historical heritage, and endemic logic is lost easily when all we actually observe are a few indicators. Moreover, these indicators may mean somewhat different things in the unique context of specific countries (Ragin, 1987). Scholars are frequently quite sensitive to this issue and so attempt to embed their statistical analyses within the histories and social uniqueness of the countries being studied. Also, as Ragin (1994) argues and empirically demonstrates, Boolean analysis can help overcome this dilemma by enabling us to build conjunctural models with very few cases and to analyze nonevents. But the approach is somewhat circumscribed. It needs to be guided by strong theory and background knowledge, its applicability is limited to relatively few cases, and it may be biased in favor of corroborating nonadditive, conjunctural models (Ragin et al., 2003).

A second constraint has to do with the often limited number of observations available. This is especially true of studies of advanced (OECD) democracies where the *N* rarely exceeds 25. In broader worldwide comparisons, however, the *N* approaches 200. Many researchers attempt to supplement the paucity of nations by collecting longitudinal data on each nation, as with pooled cross-sectional and time series analyses. A 20-year time series for 20 countries yields 400 observations; a 20-year time series for 200 nations yields 4000. As ever, longer data series for individual countries become available, the small *N* problem will gradually diminish.

Small samples do not necessarily pose problems of statistical inference, especially when the substantive theory is strong. They do, however, limit what statistical tools can be brought to bear. Fearon (1991) argues that the smaller the sample size, the greater is the need to make counterfactuals explicit – in other words, to tighten the theoretical formulation about the precise conditions that will (or will not) produce an outcome. Vague theory implies uncertainty about the number of contending explanations (which can grow very large) and about their mutual relationship within a causal order. The consequences of this include causal misspecification, multicollinearity and large error terms. The last two of these can only be managed by having more observations. Without them, estimations are likely to yield results that are not robust. The less ambiguous the theory, the less the need for large *N*'s. Small *N*'s are, therefore, an especially acute problem in disciplines without a firm theoretical architecture, like sociology or political science. Western (1998a) argues that Bayesian estimation, by allowing for greater uncertainty, produces more robust results when *N*'s are few, theory vague, and explanations are many.

Small *N*'s can even hold advantages, because a limited nation sample implies that the researcher can gain more easily maximum advantage from diagnostic scrutiny of the residual plots. A set of systematic outliers, for example, can be easily

identified in terms of their nationhood, and any good comparativist should be able to pin down what variable(s) drives their deviance from the fitted regression line. Diagnostics on small samples can approach the advantages of in-depth case studies.

The constraints imposed by small samples are generally well known to researchers. There are additional problems inherent in quantitative national comparisons that are both more generic and potentially more serious, but also less recognized. Chief among these are issues related to qualitative and limited dependent variables, the potential for selection bias, possible lack of independence between observations and, finally, the question of endogeneity of variables.

Qualitative and Limited Dependent Variables

Many dependent variables in cross-national research are either qualitative (assuming discrete values) or limited (they assume continuous values within some range). The choice whether to measure a variable in a qualitative or continuous way is often controversial. Bollen and Jackman (1989), for example, argue that difficulties in classifying some political regimes speak in favor of using continuous scales because 'dichotomizing democracy blurs distinctions between borderline cases.' In contrast, Przeworski et al. (2000) prefer to treat regimes dichotomously or multinomially.

When the dependent variable is either qualitative or limited, we need to use nonlinear models. Whether we give political regimes the values of 0–1 (as do Przeworski et al., 2000) or 1–100 (as does Bollen, 1980), it remains that the value on the dependent variable cannot exceed its maximum when the independent variable(s) tend to infinity, and it cannot fall below its minimum when these variables tend to infinity. Linear models, at best, can provide an approximation within some range of the independent variables.

The standard model when the dependent variable is multinomial is

$$\Pr(Y = j, X = x) = F(x) \qquad [1]$$

where $j = 0, 1, \ldots, J - 1$ and F is the cumulative distribution function. Since such models are covered by any standard textbook we need not present them here. Such models can be applied to panel data unless the number of repeated observations is large.

'Event history analysis' is one particular class of nonlinear models applied in cross-national research. In such models the dependent variable is an event, such as a revolution, regime transition, or a policy adoption. The general model is

$$\Pr(Y(t + \mathrm{d}t) = j) = f(y(t), x(t)), \mathrm{d}t0 \qquad [2]$$

Most often, such models can be conveniently estimated as

$$\log S(t) = \log(1 - F(t)) \qquad [3]$$

where $S(t)$ is the survival function, or the probability that an event lasts beyond time t, and $F(t)$ is the cdf.

For dichotomous dependent variables, logit and probit give very similar results. For multinomial variables, it is often assumed that the errors are independent across the values of the dependent variable, which leads to a logit specification. But this implies a strong assumption, namely the irrelevance of independent alternatives that rarely holds in practice. Multinomial probit, in turn, requires computing multiple integrals, which was until recently computationally expensive. Alternatives would be semi- and nonparametric methods.

The distributions that are commonly used in estimating survival models include the exponential, Weibul, logistic, and Poisson distributions. But, except for the Weibul and exponential, it is very difficult to distinguish statistically between such distributions.

Methods for studying qualitative dependent variables are now standard textbook fare, but what warrants emphasis is that the traditional distinction between 'qualitative' and 'quantitative' research is becoming increasingly obsolete. Phenomena such as social revolutions may be rare, but this just means they occur with a low probability. They can, nevertheless, be studied systematically using maximum likelihood methods.

The Problem of Selection Bias

In comparative research, we are often interested in the effect of some systemic feature, institution, or policy on some outcome. Examples include the effect of labor market institutions on unemployment, the effect of political regimes on economic growth, or the effect of electoral rules on the number of political parties.

In such cases, the question is how some feature of X affects some outcomes Y in the presence of conditions Z, or PR (Y, X, Z), where the hypothesis is that $y = f(x)$. The generic problem is how to isolate the effects of X and Z on Y.

If sampling is exogenous (i.e., the probability of observing any y is independent of z), the $pr(Y = yz) = pr(y)$, and standard statistical methods can be used. But if $pr(Y = yz)$ is different under different conditions Z, sampling is endogenous and the assumptions of the standard statistical model are violated.

To exemplify the problem: Suppose we want to know the impact of political regimes on economic growth. We observe Chile in 1985 (Z), which was a dictatorship (X), in which case per capita income declined at the rate of $Y = -2.26$. To assess the effect of political regimes, we need to know what would have been the rate of growth (Y) in Chile had it been a democracy. This cannot be answered with the available observations because Chile in 1985 was indeed a dictatorship. The comparativist in such a case would proceed quasi-experimentally, and look for a case that matches Chile in all aspects except authoritarianism. In this way one would compare an authoritarian with a democratic 'Chile.'

But what if Chile's status as a dictatorship in 1985 was due to some factors that also affected its economic performance? This could be because growth itself influences the survival of regimes; a country's level of development may be associated both with regime selection and growth; some unobservable factors (say 'enlightened leadership' or, perhaps, measurement error) may be common to both variables. Whatever the underlying selection mechanism, the basic consequence is that there will be cases without a match, and this implies that our inferences will be biased.

The problem raised by nonrandom selection is how to make inferences from what we observe to what we do not. It is

a problem of identification. For what do we know when we observe Chile as a dictatorship in 1985? We observe the fact that, given conditions Z, the Chilean regime (X) is a dictatorship. We also know its rate of economic growth, given that it is a dictatorship under conditions Z. We do not observe, however, its rate of growth as a democracy under conditions Z. The issue is one of counterfactual observations: the rate of growth that an observation with $Z = zi$, observed under dictatorship, would have had under democracy. We do not know what this value is and, hence, we face underidentification.

What we need to know are two distributions, $P(Y = y, x = j, Z)$ and $P(Y = y, x = k, Z)$, and their expected values, where we now think of the possible values of X more generally as j, $k = 0, 1, 2, \ldots, N$. The first is the distribution of Y in the entire population if all cases were observed as dictatorships; the second is the distribution of Y in the entire population if all cases were observed as democracies. Clearly, if we want to know the impact on y of being in states x, we need to determine the difference between these two distributions, conditional on z.

It may seem strange to refer to the 'population' as including unobserved cases. But in order to compare the effects of states x on the performance of an individual case characterized by a zi, we must allow the case to be potentially observable under the full range of X. We observed this case as $X = j$, but must also imagine the possibility that it would have been Xj. We must think in terms of a 'superpopulation' consisting of a continuum of potential cases. The actual sample is then regarded as having been drawn by 'nature' from this superpopulation.

If the actual population is an exogenous sample of the potential one, any multivariate combination (xi, zi) can be drawn with a positive probability. Such exogenous sampling allows us to isolate the effect of x on y given z, that is, to 'control' for z the effects of x on y. If the sample is exogenous, we can 'match' pairs ($xi = j$, zi) and ($xi = k$, zi). But if sampling is endogenous, there will be some cases without a 'match': some of the xi's will not have support. Where we have nonrandom selection, quasi-experimental comparisons fail regardless of the number of observations.

The methods to correct for selection bias consist of constructing counterfactuals, that is, of filling the unobserved supports of the distribution Y for all X. Following Heckman (1988), when sampling is not random, the regressions $y = f(x, z)$ suffer from omitted variable bias. When such regressions are estimated on the basis of the observed sample, the variable that is omitted is the expected value of the error in the equation that specifies how the observations are selected into the sample.

$$\Pr(x = j) = F(Z) + u \qquad [4]$$

If

$$E(uj\, x = j)\ 0 \qquad [5]$$

then

$$E(uj\, x) = juE(u\, x) \qquad [6]$$

where ju is a regression coefficient of uj on u.

Finally, $E(ux = j) = j$, where the j's are the inverse Mill ratios, or the hazard rates. Since we know that Yj is observed when $x = j$, we can write the expected values in the observed sample as

$$E(yj\, x = j, z) = zj + ju\, j \qquad [7]$$

Note that if this equation is estimated on the basis of the observed sample, the variable j is omitted from the specification.

We can now see why controlling for the variables that enter both into selection and outcome equations may indeed worsen the selection bias. Following Heckman (1988), we distinguish first between selection on observables and on unobservables. Selection on observables occurs when the expected covariance $E(uju\, z)\ 0$, but once the observed variables Z are controlled for it vanishes, so that $E(uju\, z) = 0$. Selection is on unobservables when $E(uju)\ 0$ and $E(uju\, z)\ 0$, which means that controlling for the factors observed by the investigator does not remove the covariance between the errors in the outcome and the selection equations. Now note that the regression coefficient $ju = \mathrm{cov}(uju)/\mathrm{var}(uj)$. If selection is on unobservables, controlling for some variable x in the outcome equation may reduce the error variance uj without equally reducing the covariance uju. Hence, the coefficient on the omitted variable will be larger and the bias will be exacerbated.

In short, the expected values of the observed cases will be biased because they covary with the variable which determines which cases are observed. If selection is exclusively on observables, this bias can be corrected by traditional controlling techniques. But if it is on unobservables, such controls only worsen the bias.

Correcting for selection bias is not uncontroversial. It has been found that corrections for selection are not robust, but are highly sensitive to relatively minor changes in assumptions about distributions (Goldberger, 1983; Robins et al., 2000). Others have found that some estimation methods fail to correct for this bias and may even exacerbate it. As Heckman (1988) argues, the quandary we face is that different methods of correcting for selection bias are robust if there is no bias to begin with; if there is, there is no guarantee that the methods are robust.

The logic of the problem is similar whether we study large or small N's. When de Toqueville concluded that revolutions do not bring about social change, the reason might be that they occur only in countries where it is difficult to change society. Even when $N = 1$, the issue of selection bias does not disappear: The French revolution in 1789 may have been caused by the same conditions that made social change so difficult. It is possible that a revolution in a country where social relations are easier to change would have provoked change. But then a revolution would not have been necessary.

Comparativists who conduct case studies cannot benefit from statistical distributions to generate the counterfactuals. Yet, as in all cases of comparison the problem remains and it should, therefore, be standard practice to ask counterfactual questions of one's case or cases.

The Problem of Independence

Statistical inference must assume that the observations on a variable are independent one of the other. Is country A's

performance truly independent of what happens in country B? Is what happens at $t + 1$ independent of events in t? Usually not, and this implies the need for corrective procedures. In most cases, however, rigorous correction will entail that the *de facto N* (nations or years) diminishes; in some instances, statistical dependency cannot be resolved at all.

Cross-sectional analysis almost invariably assumes that nations and their properties (say budgets or institutions) are independent one of the other. This should not be assumed. We know that the Scandinavian countries have a shared history, deliberately learning from each other through centuries, thus creating similar institutions and path dependencies. The same goes for Austria and Germany, for Belgium and the Netherlands and, arguably, for all the Anglo-Saxon nations. World samples have a similar problem: Japan's long hegemony in East Asia will have influenced Korean society; Confucianism has had a pervasive influence throughout the region. Similar stories are easily told for Latin America and Africa. If the world is a set of nation clusters, the real *N* is not 20-odd OECD countries or 150-odd world nations.

When nations form families, but are treated as if they were all unique and independent, we are likely to get biased coefficients and, very probably, unequal error variance (heteroskadicity). Sweden alone will drive the regression line in just about any welfare state analysis, and when also Denmark and Norway are treated as discrete observations, the bias is multiplied in so far as all three in reality form part of the same political family ('Scandinavia'). Diffusion effects that operate between members of a nation cluster can also result in heteroskadistic disturbance in the cross-section. Such can be corrected by, for example, adding a variable that captures the common underlying property that drives the disturbance (say, a dummy for being 'Scandinavia') but, again, this correction absorbs precious degrees of freedom in a small *N* study and, substantively, amounts to reducing the three nations to one observation.

Lack of independence in a time series is normally taken for granted, since this year's budget or election outcome is almost inevitably related to last year's budget or the previous election. The standard assumption is a first-order (AR1) serial correlation. In comparative research virtually all time series applications are pooled with cross-sections. But, where *N*'s are very small, one may as well simply compare across individual time series estimations, as do Esping-Andersen and Sonnberger (1991). Time series are meant to capture historical process. Yet as Isaac and Griffin (1989) argue, they easily end up being a historical.

Pooling cross-sectional with time series data (panel regressions) has become very widespread, especially in studies of the limited group of advanced (OECD) societies. In many cases, the panel design is chiefly cross-sectional (more nations than years); others are temporally dominated (for a discussion, see Stimson, 1985). Panel models are especially problematic because they can contain simultaneous diachronic and spatial interdependence and, worse, the two may interact. The standard method for correcting contemporaneous error correlation (GLS) applies only where the *t*'s well exceed nations (which is rare). The consequence is that *t*-statistics are overestimated, errors underestimated, and the results may therefore not be robust (Beck and Katz, 1995). The Beck and Katz (1995) procedure can correct for temporal and cross-sectional dependency one at a time, but if the two interact, no solution exists. See also Beck et al. (1998) for an application to maximum likelihood estimation.

Panel models can be based on two types of theoretical justification. One is that events or shocks occur over time that affect the cross-sectional variance. There is, for example, a huge recent literature on the impact of labor market 'rigidities' on unemployment: regulations vary across nations but also across time because of deregulatory legislation (see, for example, Nickell, 1997). Deregulation in a country should produce a break in its time series, and the autocorrelation element will be split into the years preceding and following the break. The second justification, not often exploited, is to interpret autocorrelation as an expression of institutional or policy path dependency. In this instant, the rho must be treated as a variable. The problem, of course, is that the rho is likely to combine theoretically relevant information as well as unknown residual autocorrelation. The researcher can accordingly not avoid including a variable that explicitly measures path dependency.

If we insist on faithful adherence to the real world, panel regressions may require so much correction against dependency that the hard-won additional degrees of freedom that come with a time series are easily eaten up. And how many can truthfully claim that time and country dependencies do not interact? Indeed, most sensible comparativists would assume they do: if nations form part of families it should also be the case that the timing of their shocks, events, or policies is interdependent. Such intractable problems are certainly much more severe in small-*N* comparisons. Here, a marginal difference in measurement, the inclusion or exclusion of one country, the addition or subtraction of a year here or there, or the substitution of one variable for another, can change the entire model.

One alternative is to construct multilevel models which explicitly take into account the possibility that nations may 'cluster' (for an overview, see Gelman and Hill, 2006). Nieuwbeerta and Ultee (1999) have, for example, estimated a three level (nation, time, and individual) model of the impact of class on party choice within the context of nations' social mobility structure. A second alternative, in particular when the dependent variable is categorical, is to exploit the advantages of event history analysis. But, here the time series needs to be quite long considering that theoretically interesting events, such as revolutions, democratization, or even welfare reforms, are far between. In the event history context, analytical priority usually is given to temporal change, which brings it much closer to traditional time series analysis. But rather than having to manipulate autocorrelation, time sequencing (states and events) is actively modeled and thus gains analytic status. For an application to nation comparisons, see, for example, Western (1998b), which also can stand as an exemplar of how to minimize the interdependency problem.

There are two particular cases where the lack of independence among observations simply prohibits adequate estimation. The first, noted above, occurs when time and nation dependencies interact. The second is when 'globalization' penetrates all nations and when many nations (such as the European Union) become subsumed under identical constraints. In this instance, existing cross-national correlations

will strengthen and we may, indeed be moving toward an $N = 1$. Of course, global shocks or European Union membership do not necessarily produce similar effects on the dependent variable across nations or time. If nations' institutional filters differ, so will most likely the impact of a global shock on, say, national unemployment rates. Here we would specify interaction effects, but that would be impossible in a pure cross-section, and extremely difficult in a time series, unless we already know how the lag structure will differ according to institutional variation.

The Endogeneity Problem

All probabilistic statistics require conditional independence, namely that the values of the predictor variables are assigned independently of the dependent variable. The basic problem of endogeneity occurs when the explanans (X) may be influenced by the explanandum (Y) or both may be jointly influenced by an unmeasured third. The endogeneity problem is one aspect of the broader question of selection bias discussed earlier.

The endogeneity issue has been debated intensely within the economic growth literature in terms of the causal relationship between technology and growth. But it applies equally to many fields. For example, comparativists often argue that left power explains welfare state development. But are we certain that left power, itself, is not a function of strong welfare states? Or, equally likely, are both large welfare states and left power just two faces of the same coin, different manifestations of one underlying, yet undefined, phenomenon? Would Sweden have had the same welfare state even without its legendary social democratic tradition? Perhaps, if Sweden's cultural past overdetermines its unique kind of social democracy and social policy. If this kind of endogeneity exists, the true X for Sweden is not left power but a full list of all that is 'Sweden.' The vector of the X's becomes a list of all that is nationally unique.

The endogeneity problem becomes easily intractable in quantitative cross-national research because we observe variables (Y's and X's) that represent part of the reality of the nations we sample. Our variables are in effect a partial reflection of the society under study, and the meaning of a variable score for one nation may not be metrically equivalent to that of another – a weighted left cabinet score of 35 for Denmark and 40 for Sweden probably misrepresents the Danish–Swedish difference if, that is, the two social democracies are two different beasts.

A related, and equally problematic, issue arises in the interpretation of coefficient estimations. In the simple cross-sectional regression, the coefficient for left power, economic development, or for coordinated bargaining can only be interpreted in one way: as a constant, across-the-board effect irrespective of national context. If Denmark had five points more left power, its welfare state should match the Swedish (all else held constant). In fixed-effects panel regressions, the X's are assumed to have an identical impact on Y, irrespective of country. Often we cannot accept such assumptions.

The reason that we cannot is that it is difficult to assume that variables, say a bargaining institution or party power balance, will produce homogenous, monotonically identical, effects across nations for the simple reason that they are embedded in a more complex reality (called nation) which has also given rise to its version of the dependent variable. In this case, the bias remains equally problematic whether we study few or many N's; it will not disappear if we add more nations. Quantitative national comparisons rarely address such endogeneity problems. Studies of labor market or economic performance routinely presume that labor market regulations or bargaining centralization are truly exogenous variables, whose effects on employment is conditionally identical whether it is Germany, Norway, or the United States. The same goes for welfare state comparisons with their belief that demography and left power are fully exogenous and conditionally unitary.

There exist several, not necessarily efficient, ways of correcting for the endogeneity bias. If we have some inkling that the bias comes from variable omission, the obvious correction entails the inclusion of additional controls. An example of this kind was Jackman's (1986) argument that Norway's North Sea Oil was overdetermining the results in the Lange and Jarrett (1985) study. Small N studies with strong endogeneity have little capacity to extend the number of potentially necessary controls. A second approach is to limit endogeneity in X by reconceptualizing and, most likely, narrowing Y. The welfare state literature provides a prototypical example: aggregate social expenditure increasingly was replaced by measures of specific welfare state traits.

Controls, no matter how many, will however not resolve the problem under conditions of strong sampling bias. As discussed above, the best solution in such a situation is to concentrate more on the theoretical elaboration of causal relations between variables. If we can assume that our estimations are biased because Y affects the values on X, or because both are jointly attributable to a third underlying force, thinking in counterfactual terms ('would Sweden's welfare state be the same without Sweden's social democracy') will force the researcher to identify more precisely the direct or derived causal connections (Fearon, 1991; Morgan and Winship, 2007). If bias can be assumed to come from the assumption of monotonically homogeneous effects of the X across all nations, the researcher's attention should concentrate on identifying more precisely the conditional mechanisms that are involved in the causal passage from an X to a Y (why and how will a 5-point rise in left power make the Danish welfare state converge with the Swedish?).

Conclusion

The difficulties and tradeoffs that we have synthesized above are intrinsic to all quantitative research, but are often more difficult to deal with in national comparisons because of the inevitability of small samples of countries and of the ways in which nations have developed together in interaction and will continue to do so even more in the future, given globalization. However, the difficulties should not be seen as paralyzing research but rather as stimulating attempts to deal with them. Indeed, in most practical situations today, each of these problems can be dealt with, albeit not perfectly. So, today's partial vices may be the impetus to tomorrow's virtues. Although we may never be able to completely resolve all problems, ongoing trends suggest that our capacity to conduct good quality

research in this field is improving rapidly. For one, the constraints of small *N*'s are diminishing because of rapid progress in data quality and quantity, covering both more years and nations. Second, the field has made great strides in terms of researchers' substantial comparative knowledge of the nations and cultures under study. Hence, our ability to interpret and contextualize statistical findings has greatly strengthened. And this, in turn, is a major precondition for what we believe is the single most effective tool with which to manage the challenges of quantitative comparison: namely, to systematically think in terms of counterfactuals.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Control Variables in Research; Cross-Cultural Research Methods in Sociology; Endogeneity: Methodology; Event–History Analysis in Discrete Time; Experimental Design: Bayesian Designs; Experimental Design: Overview; External Validity; Internal Validity; International Social Survey Programme; Latent Structure and Causal Variables; Longitudinal Causal Inference; Mechanism-Based Causal Analysis; Moderator Variable: Methodology; Nonequivalent Group Designs; Panel Surveys: Uses and Applications; Regression Discontinuity Design; Selection Bias, Statistics of.

Bibliography

Amemyia, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.
Beck, N., Katz, J.N., 1995. What to do (and not to do) with times-series, cross-section data. American Political Science Review 89, 634–647.
Beck, N., Katz, J.N., Tucker, R., 1998. Taking time seriously: time series-cross-section analysis with a binary dependent variable. American Journal of Political Science 42, 1260–1288.
Blossfeld, H.P., Roehwer, G., 1995. Techniques of Event History Modelling: New Approaches to Causal Analysis. L. Erlbaum, Mahwah, NJ.
Bollen, K.A., 1980. Issues in the comparative measurement of political democracy. American Sociological Review 45, 370–390.
Bollen, K.A., Jackman, R.W., 1989. Democracy, stability, and dichotomies. American Sociological Review 54, 612–621.
Esping-Andersen, G., Sonnberger, H., 1991. The demographics of age in labor market management. In: Myles, J., Quadagno, J. (Eds.), States, Labor Markets, and the Future of Old-Age Policy. Temple University Press, Philadelphia, PA, pp. 227–249.
Fearon, J.D., 1991. Counterfactuals and hypothesis testing in political science. World Politics 43, 169–195.
Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.
Goldberger, A., 1983. Abnormal selection bias. In: Karlin, S., Amemyia, T., Goodman, L.A. (Eds.), Studies in Econometrics, Time Series, and Multivariate Statistics. Academic Press, New York, pp. 67–84.
Goldstein, H., 1987. Multilevel Models in Educational and Social Research. Oxford University Press, New York.
Heckman, J.J., 1988. The microeconomic evaluation of social programs and economic institutions. In: Chung-Hua Series of Lectures by Invited Eminent Economists No. 14. The Institute of Economics, Academia Sinica, Teipei, China.
Hsiao, C., 1986. Analysis of Panel Data. Cambridge University Press, Cambridge, UK.
Isaac, L.W., Griffin, L.J., 1989. A historicism in time-series analyses of historical process. American Sociological Review 54, 873–890.
Jackman, R., 1986. The politics of economic growth in industrial democracies, 1974–1980: leftist strength or North Sea oil? Journal of Politics 49, 242–256.
Janoski, T., Hicks, A. (Eds.), 1994. The Comparative Political Economy of the Welfare State. Cambridge University Press, Cambridge, UK.
Lange, P., Jarrett, G., 1985. The politics of growth. Journal of Politics 47, 792–827.
Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge, UK.
Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, New York.
Nickell, S., 1997. Unemployment and labor market rigidities: Europe versus North America. Journal of Economic Perspectives 11 (3), 55–74.
Nieuwbeerta, P., Ultee, W., 1999. Class voting in Western industrialized countries, 1945–1990: systematizing and testing explanations. European Journal of Political Research 35, 123–160.
Przeworski, A., Alvarez, M.R., Cheibub, J.A., Limongi, F., 2000. Democracy and Development: Political Regimes and Material Welfare in the World, 1950–1990. Cambridge University Press, New York.
Ragin, C., 1987. The Comparative Method. University of California Press, Berkeley, CA.
Ragin, C., 1994. A qualitative comparative analysis of pension systems. In: Janoski, T., Hicks, A.M. (Eds.), The Comparative Political Economy of the Welfare State. Cambridge University Press, Cambridge, UK, pp. 320–343.
Ragin, C.C., Shulman, D., Weinberg, A., Gran, B., 2003. Complexity, generality, and qualitative comparative analysis. Field Methods 15, 323–340.
Robins, J.M., Rotnitzky, A., Scharfstein, D.O., 2000. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer, New York, pp. 1–94.
Stimson, J.A., 1985. Regression in time and space: a statistical essay. American Journal of Political Science 29, 914–947.
Western, B., 1998a. Causal heterogeneity in comparative research: a Bayesian hierarchical modeling approach. American Journal of Political Science 42, 1233–1259.
Western, B., 1998b. Between Class and Market: Postwar Unionization in the Capitalist Democracies. Princeton University Press, Princeton, NJ.

Quantitative Linguistics

Stefan Th Gries, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article surveys a selected variety of statistical methods that are currently used in experimental and observational studies in linguistics. It covers goodness-of-fit tests, monofactorial and multifactorial hypothesis testing methods, and hypothesis-generating techniques. In addition, for the two major sections of significance testing and exploratory methods, the article also discusses a wide range of statistical desiderata, i.e., perspectives and methods whose more widespread recognition or adoption would benefit linguistics as a discipline.

Introduction

Over the last 20 or so years, linguistics has taken a decidedly quantitative turn. While subdisciplines such as phonetics, sociolinguistics, and experimental psycholinguistics have employed statistical methods for a long time, work in most other central subdisciplines – morphology, syntax, semantics, to name but a few – has only done so since about the 1990s. This shift has no doubt been facilitated by a variety of developments.

For example, the predominance of generative linguistics, with its stance that experimental or observational evidence for one's judgments is not really required, has waned and functional, cognitive, and exemplar-usage-based theories have become more widely adopted; since these theories have relied much on empirical data, an increase in the use of statistical tools was a natural by-product of these theoretical developments.

In addition to any major theoretical shifts in linguistics, there are also growing overall tendencies to (1) study linguistic phenomena from quantitative perspectives including, but not limited to, probability theory, information theory, etc. and (2) cross-disciplinary boundaries and, thus, get involved with neighboring disciplines in which statistical modeling has been much more widespread such as psychology, sociology, and communication.

As yet another example, technological progresses – computers with faster processors, more RAM, and larger hard drives as well as the invention of the WWW – have made it much easier to compile and process large(r) corpora and other kinds of 'big data.' Large(r) corpora in turn yield higher frequency data, which cannot be studied by mere eyeballing; so this, too, has led to a burgeoning use of statistics in all areas of linguistics in which corpora are studied.

As a result of these converging trends, quantitative methods in linguistics are now established and constitute a vibrant methodological domain. However, given the ever-evolving nature of the field of statistics and the recency of the more widespread adoption of quantitative methods in linguistics, the latter are also still in a field of flux as best practices are still being developed/established. This article provides an overview of how statistical tools are used in linguistics – given the vastness and diversity of linguistics as well as all the ways in which statistical methods can be used; however, this can only be a highly selective bird's-eye overview. The section 'Hypothesis-Testing Methods' covers quantitative methods that are hypothesis testing in nature, i.e., that return, among other things, *p*-values from significance tests. The section 'Hypothesis-Generating Methods', then, deals with hypothesis-generating approaches, i.e., methods that aim at detecting or hypothesizing structures in data without necessarily testing these for significance. Each of these sections will first provide an overview of what may be considered the state of the art before turning to a variety of desiderata, i.e., methods or developments. The section 'Concluding Remarks' then concludes with a few general developments that linguistics as a discipline would benefit from.

Hypothesis-Testing Methods

Statistical tools that belong to the domain of null-hypothesis significance testing are the most widespread in linguistics. Within this set of tools, one needs to distinguish between *goodness-of-fit tests* and *tests for independence/differences*. The former are concerned with testing whether a characteristics of a particular data set – a mean, a standard deviation, the overall distribution – is different from that of some other data set (e.g., one from a previous study) or a known distribution (e.g., the bell-shaped normal distribution). The latter can be divided up into *monofactorial* and *multifactorial* tests: both involve one *dependent variable* (or response or effect), but the monofactorial designs contain only one *independent variable* (or predictor or cause) whereas multifactorial designs contain more than one independent variable. While not a standard statistical terminology, it has occasionally been didactically useful to distinguish two kinds of multifactoriality; Adopting this distinction, multifactorial$_1$ refers to designs in which multiple independent variables are involved but without interactions, whereas multifactorial$_2$ then refers to designs in which multiple independent variables are involved such that they may interact with each other.

The following sections will survey monofactorial and multifactorial approaches. However, given the fact that no linguistic phenomenon is truly monofactorial in nature and that,

correspondingly, the field has been moving in the direction of multifactorial testing, the emphasis will be on this latter type of approach.

Goodness-of-Fit Tests in Common Use

Nearly all statistical tests can be said to involve one of the following five statistics:

- Distributions
- Frequencies
- Averages (such as means or medians)
- Dispersions (such as standard deviations or interquartile ranges)
- Correlations (such as Pearson's r or Kendall's τ).

Thus, for each of these statistics, goodness-of-fit tests are conceivable. The probably most frequent goodness-of-fit test involves testing distributions, more specifically whether data are normally distributed, i.e., whether a histogram or density plot would result in a symmetric bell-shaped form. This scenario is frequent because many statistical tests require that the data tested exhibit this shape lest their significance tests produce unreliable results. Figure 1 exemplifies this scenario: a researcher may have collected the data represented by the histogram and the dashed density curve and now needs to determine whether the data are sufficiently similar to the heavy normal curve to assert that his data are normally distributed.

However, goodness-of-fit tests are available for the other four statistics, too. An example for frequencies involves testing whether the frequencies with which foreign language learners choose three modal verbs in a gap-filling experiment differ significantly from native speakers' choices in some prior study. An example for averages involves testing whether foreign language learners' average acceptability judgment for n sentences is significantly different from native speakers' average acceptability judgment in previously studied data. And an example for correlations involves testing whether the correlation between the age of a child and the child's mean length of utterance in morphemes in corpus data is the same as that found for other children from a comparable corpus.

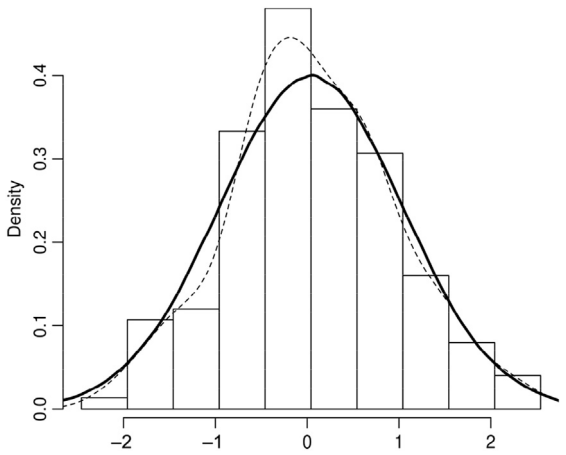

Figure 1 Histogram and density curve (dashed line) of fictitious data compared to a normal distribution with the same mean and standard deviation (heavy line).

Tests for Independence/Differences in Common Use

The much more frequent kind of statistical scenario involves both dependent and independent variables, and the statistic of the dependent variable is typically one of the five listed above.

Monofactorial Designs

In monofactorial hypothesis tests, one is usually interested in whether the value of the statistic in question of the dependent variable is dependent on the value of the independent variable; put differently, the question is, 'do the values of the independent variable make a difference for those of the dependent variable?' An example for a test for independence/differences of distributions would be the question of whether the similarities of the source words of blends (e.g., *breakfast* and *lunch* for *brunch*) are differently distributed from the similarities of the source words of complex clippings (e.g., *system* and *administrator* for *sysadmin*). If one was *not* interested in whether the average source word similarities are different, but just in the overall distributions of source word similarities – are the more larger values for blends? – then one could run a Kolmogorov–Smirnov test. If one *was* interested in the difference of average source word similarities of blends and complex clippings, then one could perhaps run a t-test for independent samples. If the source word similarities are not normally distributed and have different variances, one could perhaps run a U-test.

An example for a test for independence/differences of dispersions would be the question of whether two groups of subjects – say, native speakers and foreign language learners – exhibit differently variable sentence lengths (because, perhaps, the native speakers' higher competence allows them to exploit longer sentence lengths more for stylistic/expressive means); such a question could, if certain conditions are met, be tested with an F-test for variance homogeneity.

Finally, testing frequencies for independence/differences are often applied to frequency tables such as Table 1. If 100 subjects were asked "at what time does your shop close?" and an additional 100 subjects were asked "what time does your shop close?" one noted the frequencies with which one received the answers "at five o'clock" and "five o'clock," then one could test the resulting Table 1 with a chi-squared test, which would here return a significant result indicating that subjects prefer to use *at* when they were asked with *at*.

The above tests all have in common that the data evaluated are all independent; for example, no subject or no word provides more than one data point. In cases where this is not

Table 1 Fictitious data from a priming experiment

	Answer with at	*Answer without at*	*Totals*
Question with *at*	70	30	100
Question without *at*	45	55	100
Totals	115	85	200

the case, as in when subjects take a test before and after a treatment, alternatives to many of the above tests are available.

Multifactorial Designs

As mentioned above, statistical designs in many subdisciplines of linguistics now do more justice to the fact that linguistic choices are bound to always be affected by multiple causes. As a result, multifactorial approaches are now more common than ever. Most such studies involve different types of regression modeling, which can be considered state of the art at this point:

- *linear modeling*, a type of approach where the predictors can be categorical or numeric and where the dependent variable is numeric. In traditional references, such models may be treated under the headings of ANOVA (analysis of variance) or ANCOVA (analysis of covariance);
- *generalized linear modeling*, a type of approach where the predictors can be categorical or numeric and where the dependent variable can be of different forms: binary, ordinal, categorical/multinomial, or frequencies, which give rise to binary logistic regression, ordinal logistic regression, multinomial/polytomous regression, and Poisson regression respectively. (Strictly speaking, multinomial regression is not a type of generalized linear model, but for simplicity's sake and because it is closely related to a sequence of binary logistic regressions, it is included here in this bullet point.) This process still involves linear modeling after the application of a so-called link function to the dependent variable (which makes sure that the model generates only sensible predictions, e.g., no negative values are predicted for frequencies).

Regression analyses are usually conceptualized as shown in (1) and (2) using an example where a constructional choice by a speaker for an *of*- or an *s*-genitive (will the speaker say *the nut of the squirrel* or *the squirrel's nut*?) may be modeled as a function of the animacy of the possessor (*squirrel*), the animacy of the possessed (*nut*) and the difference of the lengths of the possessor and the possessed (say, in characters). The two regression equations begin with the dependent variable (Genitive), followed by a tilde meaning 'as a function of,' followed by all included predictors connected with pluses.

(1) Genitive (*of* vs *s*) ~ PossessorAnim + PossessedAnim + LenDiff
(2) Genitive (*of* vs *s*) ~ PossessorAnim + PossessedAnim + LenDiff + PossessorAnim:PossessedAnim + PossessorAnim:LenDiff + PossessedAnim:LenDiff + PossessorAnim:PossessedAnim: LengthDiff

The equation in (1) is multifactorial$_1$: multiple independent variables are studied at the same time. The equation in (2) is multifactorial$_2$: multiple independent variables and their interactions (indicated by the colons) are studied at the same time, which means that the regression in (2), but not the one in (1), can determine whether PossessorAnim has the same effect on the choice of genitive regardless of the level of PossessedAnim, whether PossessorAnim has the same effect on the choice of genitive regardless of what the length differences are, etc. The second type of modeling is statistically much more challenging, but it is often also the more revealing since many studies have shown that independent variables can interact in quite complicated ways. Often, researchers begin with a larger model such as that in (2) and then use various statistical operationalizations of Occam's razor (such as significance tests or information criteria) to trim the larger model down in a process called model selection to a so-called minimal adequate model; the model that, with some simplification, contains all, and only all, predictors that are significantly correlated with the dependent variable.

These regressions are immensely powerful and versatile tools and are now applied to many experimental and observational data sets (but see below). For instance, dependent variables that are reaction times or acceptability judgments would be analyzed with linear regression modeling; binary-dependent variables such as above in (1) and (2) and many other cases of syntactic alternations (such as the dative alternation, particle placement, preposition stranding, analytic vs synthetic comparisons, *will* vs *going to*, etc.) have been studied with binary logistic regressions; and dependent variables with more than two levels have been studied with multinomial regressions in, for instance, the semantic analysis of which of several near synonyms is the most likely choice given a particular context and why.

The above two kinds of models assume that all data points are independent of each other. However, especially in multifactorial settings, this is hardly ever the case: speakers provide more than one response in an experiment, verbs are reacted to by more than one subject, an author provides more than one constructional choice in a corpus file, etc. At the same time, these three variables – Speaker, Verb, and Author – are typically also different from variables such as TimeOfTest (before vs after), SexOfSpeaker (male vs female), or Animacy (animate vs inanimate): The levels of the latter variables (often called *fixed effects*) cover all the possibilities one would expect to exist in the population whereas the levels of the former variables (often called *random effects*) cover only the (ideally random) sample of speakers, verbs, or authors that was included in the study although a researcher would probably want to generalize beyond the set of speakers, verbs, or authors tested.

Dealing with this kind of scenario requires adjustments to the kind of (generalized) linear modeling where independence of data points is required. For quite some time, experimental psycholinguistic studies were heavily influenced by Clark (1973) and Forster and Dickinson (1976), two of the most widely cited early discussion of this problem in psycholinguistics. The current state-of-the-art approach in linguistics, however, involves what is called (generalized linear) *mixed-effects modeling*, or *multilevel modeling*. These models address the dependence by taking into account the variability that is associated with random effects. Consider a scenario in which one dependent variable is modeled as a function of one independent variable on the basis of 10 data points from three speakers each. **Figure 2** represents a regular linear model of some data in the left panel (the numbers 1, 2, and 3 represent the speakers' data points) and it is obvious that a linear model cannot account well for the data. In the right panel, a linear mixed-effects model is computed on the same data where every speaker gets his own intercept but they all share the same slope. In that panel, the dashed line reflects the overall trend and the three straight lines are the separate regression lines for each

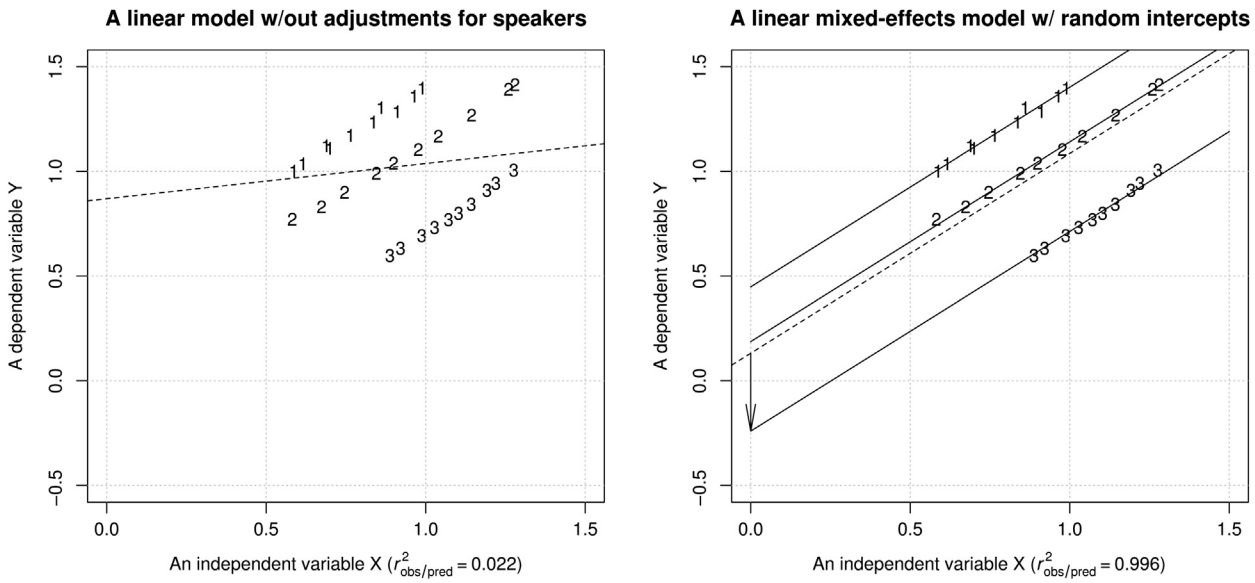

Figure 2 A linear model and a linear mixed-effects model (random intercepts) analysis of a fictitious data set.

speaker; the difference between the overall trend and the regression line for speaker 3 is represented by the arrow. Again, it is obvious that this model, which takes the relation between data points of a speaker into consideration, accounts very well for the data.

More complex versions of such mixed-effects model are available, e.g., models where speakers do not just get their own intercepts, but also their own slopes, and these adjustments can be included for crossed and nested variables that have made this a very powerful and popular approach in a very wide variety of experimental and observational studies. While some details are currently still being worked out – how to compute *p*-values for predictors in some models, how to do model selection with mixed effects, how many different random effects to include, etc. – mixed-effects modeling is bound to become one of the most important tools of the trade.

While regression models are probably the most widespread tool for multifactorial statistical analysis, another increasingly popular method is that of *classification and regression trees* or related, but more advanced, methods such as *random forests*. Such trees are based on the idea of successively splitting the data in a binary fashion such that each split groups the data based on one independent variable into two groups that predict the behavior of the dependent variable best. The output of such an analysis is typically a prediction/classification accuracy and a decision tree-like structure and the analyst interprets the tree in a top-down fashion to see which variable splits explain the dependent variable. This method is sometimes used an exploratory precursor to regression modeling, but can also be illuminating in its own right.

Desiderata

The field of linguistics has evolved enormously with regard to the number and sophistication of hypothesis-testing techniques that are now used regularly. However, given the recency of this development, many of the techniques mentioned above need to become more fleshed out in how they are applied to linguistic data and more entrenched among a larger number of practitioners. In addition to these developments, there is also a need for linguists to recognize the multitude of techniques that are already regularly applied in fields that struggle with data that pose the exact same distributional challenges that linguists face every day. Typically, psychology is mentioned as one of the closest neighboring fields whose statistical tools linguists might inspire – however, the field of ecology is maybe an even better contender (cf Zuur et al., 2007, 2009 for examples). This section will briefly outline a few methods that more linguists ought to pay more attention to, given how they can help tackle data and questions that are sometimes hard to address with the currently established tools.

One very basic but still quite useful improvement would involve linguists realizing more the power of regression approaches. For example, if a researcher has done a corpus study and represented his data in a 2×2 frequency table and then wanted to compare these results to another 2×2 frequency table from someone else's previous study, then he might resort to a heterogeneity chi-squared test to see if both data sets may represent the same population. However, not only is this somewhat cumbersome, but the heterogeneity chi-squared test also does not generalize to $r > 2 \times c > 2$ frequency tables. Instead, one natural solution may be to just analyze the data with a regression and test whether the two variables tested for in the chi-squared test also interact with a third variable that indicates which study the data are from. If there is no such significant three-way interaction, then the two data sets can be argued to represent the same population, otherwise, they cannot. Thus, while it may feel like an exaggeration to run a binary logistic regression where a chi-squared test would be entirely appropriate, adopting the regression perspective allows one to pursue a wider range

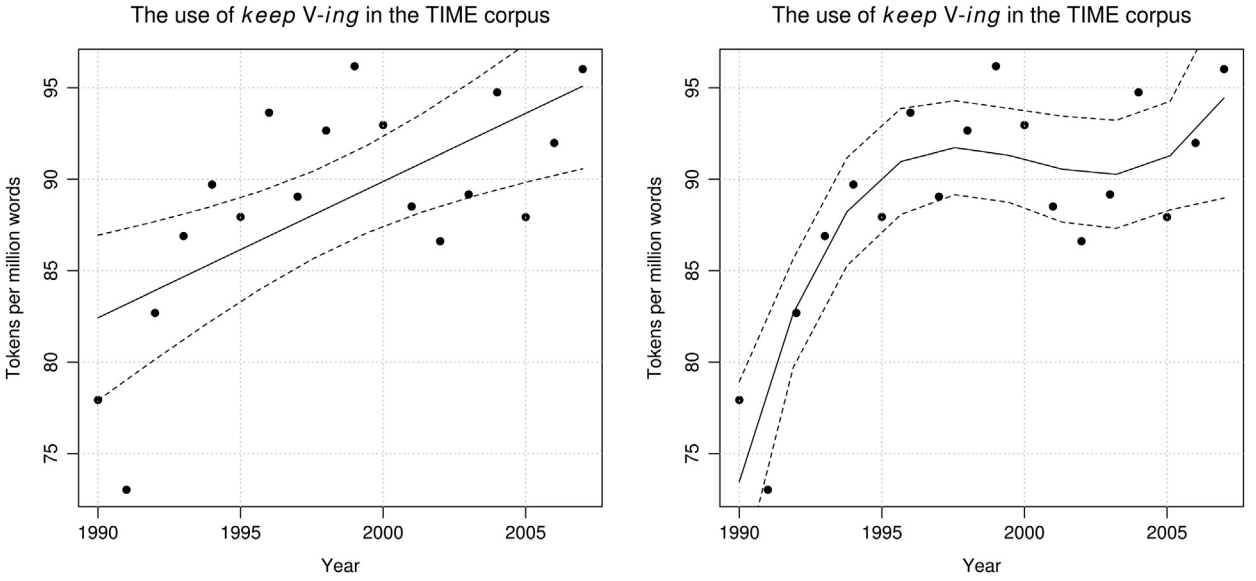

Figure 3 A linear regression model without polynomial predictors (left panel) and a linear regression model with a predictor as a third-degree polynomial (right panel).

of questions with a single unified modeling approach. (See below for another similar example.)

While conceptualizing some data sets in terms of regressions is already useful, there are of course also extensions and additional methods. Some of these extensions/methods are very closely related to regression approaches. One kind of particularly useful extension allows researchers to handle curvature in data or nonlinear effects. One example would be to include polynomial terms in regression analyses, another more general one is *generalized additive modeling*. Consider **Figure 3** for data on the frequency of the *keep* V-*ing* construction in the TIME corpus (from Hilpert, 2011). While the left panel shows that it is possible to fit a simple linear regression to these data (and even obtain a significant result), the right panel shows that a polynomial to the third degree provides a much better fit with the data and suggests that, after 1995, the growth trend seems to be leveling off.

Another technique linguists may consider using much more is *general linear hypothesis test*. This is a test that can be used as follow-up analysis of a regression model that involves at least one categorical independent variables with three or more levels. Even if such a categorical variable makes a significant contribution to a dependent variable, this does not mean that all its levels need to be maintained. For example, a researcher may distinguish five levels of animacy in his coding of corpus data – e.g., human and deities, animate but not human, plants, concrete objects, abstract entities – and the corresponding variable Animacy might be significant. Then, one can use general linear hypothesis tests to determine whether all five levels of Animacy differ from each other or whether Occam's razor would in fact require to conflate the first two and the next two, which would result in a new version of Animacy with only three levels.

Interestingly enough and as mentioned above, this logic does not only apply to, and would benefit, complex multifactorial regression models – on the contrary: Instead of submitting a, say, 4×2 frequency table to a chi-squared test, one could analyze these data with a Poisson regression and follow up with general linear hypothesis tests to determine whether all four levels of the independent variable are indeed necessary/justified.

Then, there are many statistical tools with huge potential for linguistics. One is *naïve discriminative learning* (NDL) as discussed by Baayen (2010). The overall goal is very similar to that of (generalized) linear mixed-effects modeling but, unlike regression modeling, this algorithm is based on a general model of human probabilistic learning and equations rooted in research on associative learning. As Baayen shows, NDL, while cognitively better grounded than traditional regression models, achieves classification accuracies comparable to other models and can even handle random effects in insightful ways.

Another set of relevant tools includes methods that are designed to address more explicitly the question of causal relations between predictors and responses. Strictly speaking, regression models of the above type only speak to whether there are significant correlations between the predictors on the one hand and the response on the other. However, techniques such as *structural equation modeling* or *Bayesian networks* can help study causal relations better by forcing the researcher to explicitly formulate expectations about how predictors are intercorrelated with each other and correlated with the dependent variable. For example, it is well known that many syntactic/constructional alternations such as the dative alternation or particle placement are correlated with phonological predictors (e.g., stress and syllabic length), syntactic predictors (e.g., definiteness and complexity), and discourse–functional predictors (e.g., givenness/inferrability of referents). However, these predictors are all correlated with each other, which not only poses problems to most regression models but also makes it harder to disentangle cause–effect relations. Applying

the above kinds of tools would require the researcher to hypothesize, for example, that discourse-functional characteristics are causes for phonological and syntactic effects, which in turn are causes for syntactic/constructional choices – rather than treating all predictors as mere causes and only the syntactic/constructional choice as an effect. Since usually causal relations are what we are most interested in, these techniques hold high promise.

Finally, a more general recommendation is that linguists may want to become more familiar with the field of robust statistics (cf Wilcox, 2012). Since linguistic data are rarely well-behaved in the sense of being normally distributed (instead, many phenomena are Zipfian-distributed), not suffering from outliers (many linguistic variables such as lengths and frequencies regularly have outliers), and since grouped data often differ in their dispersion, etc., robust statistics offer ways to analyze data with all these properties that undermine the use of the currently more widespread regular statistical tools.

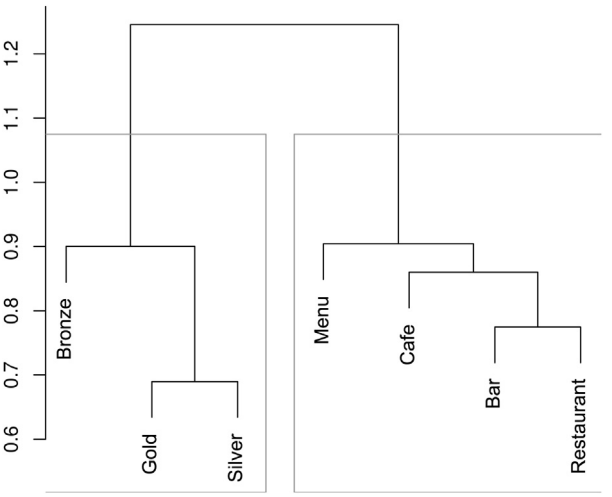

Figure 4 Results of an HCA of seven nouns on the basis of the frequencies of words they co-occur with.

Hypothesis-Generating Methods

The techniques discussed so far are all usually employed to subject hypotheses to significance testing and decide on which hypothesis to adopt based on a *p*-value. A different set of techniques is concerned with generating hypotheses in the first place. That is, these techniques serve exploratory functions and are often applied to data sets whose size and complexity defies an analyst's eyeballing and pattern-matching skills. Again, this section will begin with a discussion of what are arguably the currently most frequently used methods before turning to some refinements and desiderata that would benefit the field; as before, given that the number of techniques is vast, the overview has to be quite selective.

Exploratory Methods in Common Use

The probably most widespread exploratory technique is *hierarchical cluster analysis* (HCA). As most exploratory methods, this one, too, is typically applied with an eye to determining how n entities – sentences, NPs, words, subjects, etc. – can be grouped into $m < n$ groups/clusters that exhibit high within-group similarity and low similarity to other groups.

HCAs involve several steps: First, one computes how similar each of the n items is to each other item based on a user-defined similarity metric. The results of this step is usually represented in a (dis)similarity matrix. Crucially and unlike other exploratory tools, the items can be compared even if they involve both categorical and numeric characteristics (although numeric characteristics are most frequent). Second, in an iterative process, the pairs of items with the highest similarity values are identified and merged into groups. Once an item enters into a group, it cannot be considered in isolation anymore – what is considered for the next merger is then the newly formed group. In the most frequent type of HCA, this process is repeated until all items have been merged into one group. Third, this grouping process is visualized with a so-called dendrogram and the researcher has to decide how many clusters the dendrogram reflects most revealingly. An example is represented in Figure 4, which is the result of clustering seven nouns on the basis of the frequencies of words they co-occur with in corpus data. Both semantic considerations – the left cluster contains precious-metal nouns, the right one gastronomical terms (although note the ambiguity of *bar*) – and follow-up diagnostics to be mentioned below suggest that assuming the two clusters highlighted with gray boxes is most useful here.

HCA is a very flexible tool, suitable for very many different kinds of both experimental and observational data and often relatively straightforward to interpret, and they are becoming increasingly frequent in linguistic research.

Another frequent technique is *principal component analysis* (PCA, and its 'sibling' principal factor analysis). The overall goal of a PCA is quite similar to that of an HCA, but the underlying mathematics are quite different and PCA is only applied to all-numeric data. The input data is the same as for an HCA and a PCA proceeds by computing all possible intercorrelations between the n items/columns being studied. Then, the PCA merges/compresses the n items/columns into $m < n$ principal components, such that the items/columns within a principal components are highly intercorrelated but all principal components are usually mutually orthogonal, or independent of each other. This can mean, for example, that a PCA would convert a data set in which NPs are classified according to $n = 20$ characteristics into a new data set where, say, only four principal components retain more than 90% of the information of the originally 20 columns. This kind of result is why PCA and other similar methods – for example, *multidimensional scaling* or *correspondence analysis* – are also referred to as dimension-reduction methods. A researcher can then 'just' analyze the nature of the principal components – what are the linguistic characteristics that make up the principal components and why? – and often this interpretive task is facilitated by a two-dimensional visual representation of the results, as exemplified in Figure 5, where the seven nouns are grouped in a way that resembles the above results of the HCA.

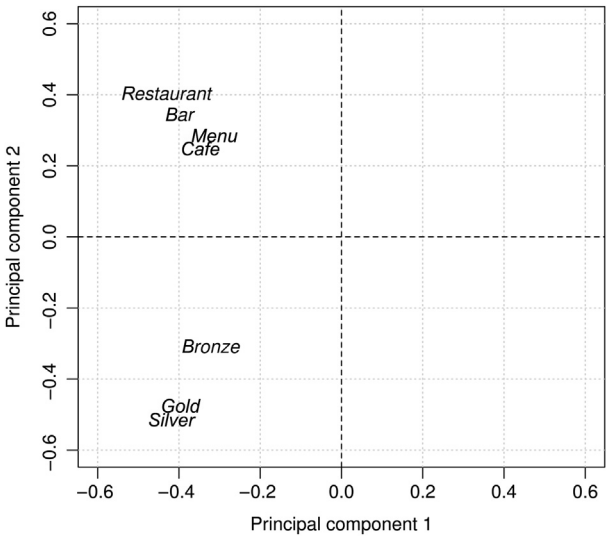

Figure 5 Results of a PCA of seven nouns on the basis of the frequencies of words they co-occur with.

Alternatively, the compressed data set can also be used as input to hypothesis-testing methods, which are now more likely to return useful results, given that instead of 20 intercorrelated variables, now only 4 uncorrelated variables, the principal components, are studied. In corpus linguistics, factor analyses have been extremely influential in, for example, Biber's (1995) work on register/genre variation, but many studies in computational semantics employ mathematically very similar methods.

Desiderata

Just as above, the desiderata regarding exploratory methods come in two kinds: First, as extensions of, or follow-ups to, the methods discussed above; second, methods in addition to the ones discussed above.

As for the former, while cluster analyses enjoy quite some use now, the results they provide are often not yet analyzed to their fullest extent. For instance, not all dendrograms are easy to interpret in terms of how many clusters a researcher should assume but very few studies discuss principled ways of arriving at such decisions. One approach to this question involves a statistic called average silhouette widths, which quantifies how similar elements are to the clusters which they are in relative to how similar they are to other clusters; in the above case of **Figure 4**, this statistic 'recommends' to assume the highlighted two clusters. Another approach toward the same problem involves resampling approaches that, ultimately, assign a *p*-value to each possible cluster so that researchers can pick the most strongly supported and most robust clusters for interpretation. In spite of the availability of such methods, they are still too rare, which can render HCA results less insightful than they would need to be. Finally, one rarely sees decisions for a particular number of clusters validated by some other statistical technique. For instance, if a cluster analysis really reveals two clusters, then a binary logistic regression (or some other non-clustering technique) should also be able to recognize these two clusters in the data – attempts at validation like these are still hard to find.

A different set of extensions involves the interpretation of the clusters: what in the data is most responsible for a particular cluster; what gives rise to the structures; which variables are reflected most in the clustering? There are diagnostic statistics available to answer such questions, but they have not yet made their way into the relevant research literature.

As for the latter kind of desiderata, there are several relatively new methods that should be highly useful to linguistic research. One of these is actually another cluster algorithm, but one with underutilized properties. This approach, so-called *fuzzy clustering*, allows for items to be members of more than one cluster and to different degrees, which is often a much more realistic assumption than the seemingly clear-cut divisions of items into clusters suggested by traditional dendrograms. Another technique that might be very useful is that of *association rules*, a data mining procedure suitable to uncover co-occurrence and *if … then* relations. The most attractive features of association rules are their conceptual simplicity and that this method can handle extremely large data sets (think Amazon shopping-basket data) consisting of categorical variables, which are not amenable to multiple cross-tabulation anymore. Thus, especially large-scale corpus data might benefit from this method.

Concluding Remarks

It is useful to conclude by mentioning a few, more general desiderata which, if they were more widely implemented or adhered to, would also tremendously help the discipline of linguistics evolve. The first of these is concerned with the fact that we not only need to broaden and deepen our knowledge of statistical methods per se, but we also need more awareness of the distributional assumptions that many methods come with. Many published studies are quite cavalier when it comes to these, but, as discussed at length in the above-mentioned Wilcox (2012) study, failure to deal with violations of such assumptions can lead to hugely anticonservative results.

In addition, the field also needs more and firmer guidelines on what is important in statistical analysis, what is important to report (e.g., checking distributional assumptions), and how methods and results should be reported. Other fields have had long and intense discussions about these issues whereas linguistics, for the most part, has not. We should be prepared to be inspired by how other disciplines with similar kinds of questions and data have come to grips with these challenges and recommendations such as Wilkinson and the Task Force on Statistical Inference (1999) provide many essential tips (e.g., to always include effect sizes to distinguish significance from effect size and make analyses comparable).

Finally, two areas in statistics that linguists should strive to learn about more: (1) developments in *resampling approaches* such as bootstrapping, which can be quite useful to enhance the precision and generalizability of our results; (2) *Bayesian statistics*, which adopts a perspective on statistical analysis that is very different from the predominant null-hypothesis significance testing approach, but which is ultimately more

in line with what we as researchers are interested in: the confirmation of reasonable hypotheses and the quantification of effects on the basis of accumulated knowledge. Given the diversity and complexity of linguistic data, any kind of tool can only be a welcome addition to our toolbox.

See also: Bayesian Statistics; Hypothesis Testing in Statistics; Invariance in Statistics; Likelihood in Statistics; Mixture Models in Statistics; Nonparametric Statistics: Advanced Computational Methods; Order Statistics.

Bibliography

Baayen, R. Harald, 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge.

Baayen, R. Harald, 2010. Corpus linguistics and naïve discriminative learning. Brazilian Journal of Applied Linguistics 11 (2), 295–328.

Biber, Douglas, 1995. Dimensions of Register Variation. Cambridge University Press, Cambridge.

Clark, Herbert H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12 (4), 335–359.

Forster, Kenneth I., Dickinson, R.G., 1976. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and min F′. Journal of Verbal Learning and Verbal Behavior 15 (2), 135–142.

Gries, Stefan Th, 2013. In: Statistics for Linguistics Using R: A Practical Introduction, second rev. & ext. ed. De Gruyter Mouton, Berlin & New York.

Hilpert, Martin, 2011. Diachronic collostructional analysis: how to use it and how to deal with confounding factors. In: Allan, Kathryn, Robinson, Justyna (Eds.), Current Methods in Historical Semantics. Mouton de Gruyter, Berlin & New York, pp. 133–160.

Johnson, Keith, 2008. Quantitative Methods in Linguistics. Blackwell, Malden, MA & Oxford.

Larson-Hall, Jenifer, 2012. A Guide to Doing Statistics in Second Language Research Using R. Routledge, New York, London. http://cw.routledge.com/textbooks/9780805861853/guide-to-r.asp.

Oakes, Michael P., 1998. Statistics for Corpus Linguistics. Edinburgh University Press, Edinburgh.

Rietveld, Toni, Hout, Roeland van, 2005. Statistics in Language Research: Analysis of Variance. Mouton de Gruyter, Berlin & New York.

Theijssen, Daphne, Louis ten, Bosch, Lou, Boves, Bert, Cranen, Hans van, Halteren. to appear. Choosing Alternatives: Using Bayesian Networks and Memory-based Learning to Study the Dative Alternation. Corpus Linguistics and Linguistic Theory 9 (2), 227–262.

Wilcox, Rand, 2012. Introduction to Robust Estimation and Hypothesis Testing, third ed. Elsevier, Amsterdam.

Wilkinson, Leland, The Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: guidelines and expectations. American Psychologist 54 (8), 594–604.

Woods, Anthony, Fletcher, Paul, Hughes, Arthur, 1986. Statistics in Language Studies. Cambridge University Press, Cambridge.

Zuur, Alain F., Ieno, Elena N., Smith, Graham M., 2007. Analysing Ecological Data. Springer, Berlin & New York.

Zuur, Alain F., Ieno, Elena N., Walker, Neil, Saveliev, Anatoly A., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin & New York.

Quantitative Methods and Crime

David McDowall, School of Criminal Justice, University at Albany–SUNY, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Crime emerged as a topic of quantitative study in the eighteenth century and scholars today analyze it using methods drawn from all of the social sciences. Crimes are distinctive for their rarity and for their social and geographic concentration. One response to these features is to adjust for them when performing standard statistical analyses. An alternative is to take the features as essential aspects of the phenomenon under study, and give them an important role within models for longitudinal and spatial data.

Crime has been a subject of quantitative investigation for more than 200 years, and sophisticated analyses became common well before criminology emerged as a self-conscious area of inquiry. Crime and related phenomena are serious concerns in large-scale urban societies, and they generate a relative wealth of formal records. Due to their availability and inherent interest, these records are then natural objects of study.

History of Quantitative Crime Analysis

Criminal trial verdicts were a topic of much discussion in eighteenth- and nineteenth-century France, and important statistical ideas resulted from analyses of jury decision-making behavior (Hacking, 1990). Early work on conviction probabilities by Condorcet and Laplace was entirely of a mathematical character and did not consider the results of actual court cases. Research on trials adopted an increasingly empirical focus, however, as the collection of criminal justice statistics became routine. Among the prominent contributors to this literature was Poisson, who developed his probability distribution for rare events and his law of large numbers in studying trial data (Stigler, 1986).

Expansion in the number and variety of state-recorded social statistics also encouraged research on the incidence and correlates of crime. Both Quetelet and Guerry used data on crime rates to undertake tests of causal hypotheses, and crime came to join suicide as a staple topic of statistical analysis (Hacking, 1990). In the succeeding years, studies of crime data have been at least a minor interest in all of the social sciences, as well as in public health and related fields. As analytical procedures developed in these areas, scholars applied them to questions involving criminal offending and victimization.

Today most major analytical methods see use in criminology, and researchers are quick to adopt new approaches when they appear. Criminology is an interdisciplinary area of study, and it maintains few fixed ideas about proper research designs or analysis techniques. As a result, the field is receptive to innovative strategies of all types. The articles in Piquero and Weisburd's (2010) edited volume provide examples of the wide scope of quantitative criminology during the early twenty-first century.

Sources of Quantitative Crime Data

Most quantitative studies of crime and criminal justice issues rely on the analysis of observational data. Experiments and quasi-experiments also appear with some frequency in the empirical literature, and criminologists hold the rigor of these research designs in high esteem. Practical and ethical considerations nevertheless limit the situations in which experimental control is feasible, and researchers must usually instead take their data as they naturally occur.

Quantitative information on crime is available in diverse forms, varying in purposes and methods of collection. Although the accuracy of these numbers is always open to some degree of doubt, they are often publicly available or otherwise easy to obtain. Most analyses of crime use data from one of three general sources.

First, police agencies collect reports of the crime incidents that come to their attention. In the United States, the Federal Bureau of Investigation (FBI) has since the 1930s compiled counts of crimes and arrests through the Uniform Crime Reporting (UCR) program (Barnett-Ryan, 2007). The UCR data cover a range of serious offenses at the city, state, and national levels, and they support aggregate time series and cross-sectional studies. Local police agencies often provide additional information about criminal incidents, including their locations and characteristics. The FBI's National Incident-Based Reporting System offers a similar level of detail, but police agencies have been slow to embrace it. Other nations also maintain publicly available crime statistics, although differences in legal definitions and questions about accuracy limit their uses in international comparisons (Newman, 1999).

Surveys of victims are a second major data source. A victimization survey draws a sample, determines if the respondents were crime victims, and gathers information about both the victims and crimes. The National Crime Victimization Survey (NCVS) in the United States is an example of this approach (see Rennison and Rand, 2007, for a description of NCVS methods). Surveys like the NCVS allow study of crime outcomes for individuals or large aggregates. They do not have as fine a level of geographic resolution as do police-recorded crime counts, but they give more detailed data on individual incidents. Many nations sponsor their own victimization surveys, and the International Crime Victimization Survey

collects data from multiple areas of the world using a standard format (Van Dijk et al., 2008).

The third source of data comes from surveys of offenders, most often from adolescent populations. The surveys ask respondents to self-report their participation in criminal activities, permitting comparisons of the offenders with their law-abiding peers. Unlike the other two data sources, no government agency sponsors a long-term self-report survey. Numerous private surveys exist, however, and data from these studies are often freely accessible to the broader research community.

Besides data about crime, extensive amounts of quantitative information also exist on associated phenomena. These include data on prisons and jails, law enforcement agencies, and criminal courts. The statistics about the criminal justice system are helpful in understanding crime, and they involve significant areas of study themselves. Within the United States, the federal Bureau of Justice Statistics compiles and distributes several large-scale data series on these crime-related topics (Groves and Cork, 2009).

Analysis of Quantitative Crime Data

Quantitative criminology has yet to develop a distinctive set of analytical techniques, and researchers instead use methods that originally appeared in other fields. As in most areas of the social sciences, studies in criminology rely heavily on the linear regression model and its many variations.

Two features of crime data nevertheless have implications for empirical analysis. First, crime is a rare event, and most members of a population manage to avoid it as either offenders or victims. In the United States as a whole, for example, only one person in 20 000 was a murder victim during 2010 (UCR data). Considering more common offenses, less than 1 US household in 40 suffered a burglary during that year, and less than 1 in 200 reported a motor vehicle theft (NCVS data).

Second, the frequency of crime varies greatly across social units and physical locations. A relative handful of offenders commits a disproportionately large share of serious criminal acts, and the burden of crime falls heavily on a similarly small group of victims. Offending is also concentrated within group settings, so that delinquents more often commit crimes when they are in the company of others (Thornberry et al., 2003). Besides these social concentrations, criminal incidents cluster in space, with many crimes occurring in a small set of geographical places.

Among the large body of research documenting the unequal distribution of crime across individuals, a study by Wolfgang et al. (1972) was particularly influential for later work. Wolfgang and colleagues followed a birth cohort over time, and they found that about 6% of the cohort's members accounted for 52% of its total arrests. The spatial concentration of crime also has a high level of empirical support. A prominent study by Sherman et al. (1989), for example, found that 3% of Minneapolis street addresses generated 50% of all calls for police services in the city.

The rarity and concentration of crime in turn imply that the frequency of criminal incidents will have a highly skewed distribution, especially in small samples. Most of the cases will be at the distribution's low end, and a long right tail will extend off toward the higher counts. Due to the social and spatial clustering, criminal incidents will also be correlated with each other over time and space.

Skewed and correlated data pose problems for conventional regression-based analysis methods, which usually assume independent Gaussian observations. Application of transformations to induce normality can often reduce or eliminate the long tails in distributions of offenses and victimizations. Appropriately weighting the observations or substituting distributions like the negative binomial for the Gaussian can similarly allow for both skews and lack of independence (see, for example, Osgood, 2000). In practice, these adaptations give researchers the ability to conduct analyses using extensions of the conventional models for regression and other popular methods.

Longitudinal and Spatial Analysis of Quantitative Crime Data

Although quantitative studies of crime can in most cases successfully apply modifications of standard techniques, the unique features of the data make some types of methods especially attractive. The clustering of criminal incidents creates problems for conventional practice, but it also opens analytical possibilities. In particular, methods for studying within-unit changes and spatial dependencies exploit the information that clustering provides, and crime researchers often apply them.

Longitudinal analyses take advantage of the concentration of offending within units to examine how criminal activity varies over time. Longitudinal studies observe a sample of units in two or more time periods and use the repeated measurements to compute individual levels of change. One class of methods accomplishes this goal by using fixed or random effect analytical models. Longitudinal studies often focus on juveniles, and the interest is in how involvement in crime evolves over the life course. By linking the past and future, these models also allow study of dynamic temporal patterns of change (see, for example, Horney et al., 1995).

Extensions of longitudinal methods consider the shape of the time path connecting earlier and later periods. The simplest extensions use polynomials in time to model nonlinear developmental sequences. More elaborate variations include growth curves that differ across individuals depending on the values of other variables. The growth curve models generally involve the use of methods for the analysis of multilevel data (Singer and Willett, 2003; Bryk and Raudenbush, 2002).

A technically different but conceptually related approach addresses the possibility of distinctive growth trajectories among population subgroups (Nagin and Land, 1993). Although several classes of trajectory models are available (Kreuter and Muthén, 2008; Nagin, 2005), all have the goal of identifying latent classes of cases that follow a common time path. Trajectory models consider if a set of clearly defined developmental patterns exists within a population, and they allow study of how these patterns differ depending on the values of other variables.

Collectively, longitudinal methods attempt to take advantage of the information that within-unit clustering provides.

Spatial methods have a similar goal, but here the clustering is cross-sectional, over geographical units at a single time point. Also, while the units in longitudinal studies are often individuals, spatial analysis units are almost always collective aggregates, such as neighborhoods or other large areas.

In elementary applications, spatial analyses plot the locations of criminal incidents on maps and use these visual displays to make inferences about the distribution of offending. This type of analysis has blossomed with the availability of capable computing equipment and the development of geographical information system (GIS) software.

Police departments increasingly attach street addresses or global positioning system data to crime incident reports, allowing the creation of detailed and question-specific displays (Parker and Asencio, 2008). The New York City Police Department was an early leader in applying GIS technology to identify crime concentrations, or 'hot spots.' By some accounts, increasing police activity in hot spot areas helped produce the dramatic crime reductions that the city began to experience in the 1990s (Batton, 1998).

Beyond mapping, statistical analyses of spatial crime data take several forms. One approach is to estimate a regression model that adjusts the results for spatial dependence in the error term. This type of analysis gives no explicit attention to the geographical characteristics of the data, and instead regards them as nuisance features that complicate an analysis. An alternative is to allow the spatial structure of the data to affect crime outcomes directly through mechanisms of contagion and diffusion. Identifying clusters of criminal offenses and their boundaries is also a common application of spatial methods, and is a natural accompaniment to mapping (Parker and Asencio, 2008). In these cases, the geographic distribution of crimes is a central element of the analysis.

Future Directions

Quantitative study of crime is now almost entirely a matter of data analysis, and most researchers devote their efforts to estimating population characteristics and relationships between variables. Existing work correspondingly gives little attention to developing a mathematical understanding of the processes that underlie crime and criminal justice outcomes. Despite its current neglect, formal mathematical modeling is a potentially important avenue for future research in the area. Such an effort could contribute to an explanation of crime clustering, for example, rather than simply taking the concentrations as a given. Adopting a more theoretical focus would coincidentally also move quantitative investigation of crime back toward some of the intellectual approaches from which it began.

See also: Crime: Knowledge about and Prevalence; Data Bases and Statistical Systems: Crime; Growth Curve Analysis; Hierarchical Models: Random and Fixed Effects; Linear Hypothesis: Regression (Basics); Longitudinal Data: Event–History Analysis in Discrete Time; Spatial Dynamics and Crime; Spatial Statistical Methods; Statistical Analysis: Multilevel Methods; Statistical Clustering.

Bibliography

Batton, W., 1998. Turnaround: How America's Top Cop Reversed the Crime Epidemic. Random House, New York.

Barnett-Ryan, C., 2007. Introduction to the uniform crime reporting program. In: Lynch, J.P., Addington, L.A. (Eds.), Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR. Cambridge University Press, New York, pp. 55–89.

Bryk, A.S., Raudenbush, S.W., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, second ed. Sage, Thousand Oaks, CA.

Groves, R.M., Cork, D.L. (Eds.), 2009. Ensuring the Quality, Credibility, and Relevance of U.S. Justice Statistics. National Academies Press, Washington, DC.

Hacking, I., 1990. The Taming of Chance. Cambridge University Press, New York.

Horney, J.D., Osgood, D.W., Marshall, I.H., 1995. Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. American Sociological Review 60, 655–673.

Kreuter, F., Muthén, B., 2008. Analyzing criminal trajectory profiles: bridging multilevel and group-based approaches using growth mixture modeling. Journal of Quantitative Criminology 24, 1–31.

Nagin, D.S., 2005. Group-Based Modeling of Development. Harvard University Press, Cambridge, MA.

Nagin, D.S., Land, K.C., 1993. Age, criminal careers, and population heterogeneity: specification and estimation of a nonparametric, mixed poisson model. Criminology 31, 327–362.

Newman, G. (Ed.), 1999. Global Report on Crime and Justice. Oxford University Press, New York.

Osgood, D.W., 2000. Poisson-based regression analysis of aggregate crime rates. Journal of Quantitative Criminology 16, 21–43.

Parker, R.N., Asencio, E.M., 2008. GIS and Spatial Analysis for the Social Sciences. Routledge, New York.

Piquero, A.R., Weisburd, D. (Eds.), 2010. Handbook of Quantitative Criminology. Springer Science+Business Media, New York.

Rennison, C.M., Rand, M., 2007. Introduction to the national crime victimization survey. In: Lynch, J.P., Addington, L.A. (Eds.), Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR. Cambridge University Press, New York, pp. 17–54.

Sherman, L.W., Gartin, P.R., Buerger, M.E., 1989. Hot spots of predatory crime: routine activities and the criminology of place. Criminology 27, 27–56.

Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press, New York.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, MA.

Thornberry, T.P., Krohn, M.D., Lizotte, A.J., Smith, C.A., Tobin, K., 2003. Crime and Delinquency in Developmental Perspective. Cambridge University Press, New York.

Van Dijk, J.J.M., van Kesteren, J.N., Smit, P., 2008. Criminal Victimization in International Perspective: Key Findings from the 2004–2005 ICVS and EU ICS. Boom Legal Publishers, The Hague.

Wolfgang, M.E., Figlio, R.M., Sellin, T., 1972. Delinquency in a Birth Cohort. University of Chicago Press, Chicago.

Quasi- and Field Experiment

Alex Díaz, María Jiménez-Buedo, and David Teira, UNED, Madrid, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Field trials and quasi-experiments are comparative tests in which we assess the effects of one intervention (or a set thereof) on a group of subjects as compared to another intervention on another group of similar characteristics. The main difference between field trials and quasi-experiments is in the way the interventions are assigned to the groups: in the former the allocation is randomized whereas in the latter it is not. We are going to see first the different roles played by randomization in medical experiments. Then we will discuss how controlled field trials, originating in psychology, spread to the social sciences throughout the twentieth century. Finally, we will show how the idea of a quasi-experiment appeared around a debate on what constitutes a valid test and what sort of controls guarantee it.

Field Experiments and Quasi-Experiments

Field experiments and quasi-experiments are comparative tests in which we assess the effects of an intervention (or a set thereof) on a group of subjects as compared to another intervention on another group of similar characteristics. The main difference between field trials and quasi-experiments is in the way the interventions are assigned to the groups: in the former the allocation is randomized whereas in the latter it is not. For more than a century now, there is an ongoing debate about what sort of methodological controls in the comparison warrant causal inferences in such experiments. In particular, it is under discussion in what sense randomized allocations provide superior grounds for causal inferences. We are going to see, in the first place, the roles played by randomization in medical experiments. Then we will discuss how controlled field experiments, originating in psychology, spread in the social sciences throughout the twentieth century. Finally, we will show how the idea of a quasi-experiment appeared around a debate on what constitutes a valid test and what sort of controls guarantee it.

Randomized Experiments

Randomized experiments originate mainly in medicine, where the fair comparison of therapies is well documented centuries ago. As early as in 1662, the Flemish physician J.B. Van Helmont suggested settling a dispute about the efficacy of bloodletting and purging with a randomized allocation of treatments: "Let us take out of the hospitals […] 200 or 500 poor people, that have fevers, pleurisies. Let us divide them into halves, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation; but you do, as ye know […]. We shall see how many funerals both of us shall have." A 1747 test of six remedies for scurvy is considered the first actually controlled trial: conducted by James Lind, a naval surgeon, it showed the superior efficacy of (vitamin-C-containing) oranges and lemons as compared to the other alternatives (against the opinion of both the Royal College of Physicians and the Admiralty). As Iain Chalmers (2006) has extensively argued, randomization originated in comparative experiments as a method for the correction of biases in the allocation of the tested treatments. If randomized, the preferences of the experimenter allocating them would not contaminate the comparison (e.g., assigning her favorite treatment to one particular group of patients). Randomization was just one possible *control* for the selection bias; an alternative one was alternating treatments between patients. Medical experimenters identified many other biases in their tests, developing *controls* for them (e.g., blinding as early as 1784). As it often happens with error-correction procedures in most scientific fields, these debiasing methods were grounded in the experimenters' experience, without any substantial theory to account for their efficacy. In the first decades of the twentieth century, medical trials implementing all sorts of controls were common, but the standardization of the randomized clinical trial (RCT) began only with the 1947 test of Streptomycin, an antibiotic drug, commissioned to Austin Bradford Hill, a medical statistician, by the British Medical Research Council (Teira, 2013a).

Hill relied on the foundations for experimental design developed by the great statistician Ronald Fisher for agricultural research on fertilizers and soils. Fisher used randomization (now explicitly defined on the probability of receiving a given treatment) to ground significance testing: how likely it is to observe an experimental outcome equally or more extreme than the one actually obtained in an infinite series of repetitions of the trial? Randomization secured that the actual allocation of treatments was just one random draw in that series so we could calculate the *p*-value for the actual outcome obtained on the basis of the distribution of all outcomes for every possible allocation. Furthermore, Fisher defended the role of randomization in causal inference: The interference of potential confounding factors could be disconnected from the treatment outcome in a series of replications of the experiment if the allocation was randomized.

Bradford Hill merged these three roles of randomization (debiasing, significance testing, causal inference) in the trial of Streptomycin, establishing the first template for the RCT. The pharmaceutical revolution of the 1950s also brought about the expansion of drug testing, and the template was gradually refined with further debiasing controls (e.g., double blinding) and statistical tools (e.g., power calculations, showing the

adequate sample size for a statistically sound test). In 1962, RCTs for safety and efficacy became compulsory for the regulatory approval of new drugs by the US Food and Drug Administration. This created a booming industry of pharmaceutical trials still alive today, although under serious discussion (Goldacre, 2012).

Field Experiments in the Social Sciences

The history of field experiments in the social sciences is still to be written. The few existing accounts have not so far been inclusive of all the social scientific disciplines, and this leaves us with a fragmented portrayal of experimentation in the social science. In the case of economics, for instance, and according to Steven Levitt and John List (2009), the use of field experiments originated in the early twentieth century, growing gradually to an upsurge in the last decade (the 2000s). Here we will motivate the importance of two questions that arise in the light of this yet fragmentary history: How did social scientists reach a consensus on the experimental design of field trials? Why did they use this design so scarcely until recently? The methodological design and interpretation of a field trial is in itself far from obvious: In principle, we are trying to grasp the causal effect of an intervention comparing it with an alternative one on two groups of subjects, both entirely alike except for the effect of the intervention. We find explicit methodological discussions on such comparisons in the nineteenth century (e.g., Mill's method of difference), but experimenters need to agree, on the one hand, on the controls they impose on the two groups in order to make sure that no confounding factors make a difference between them. And, on the other hand, if there is a quantitative difference between the outcomes, experimenters need to agree also on a statistical procedure to assess the effects.

Historians of statistics, following Theodore Porter (1995), have shown how the use of statistical techniques in public policy was presented as an alternative to expert judgment. RCTs in medicine came to replace the opinion of individual doctors about the efficacy of treatments (sometimes based on laboratory evidence, often just on their own clinical experience). Of course, many physicians considered their judgment objective enough to ground the assessment. And, furthermore, a comparative approach was often considered unethical since it would deprive a group of patients of a treatment that a physician would have recommended otherwise. This is why the agreement on a comparison constrained by a series of controls (often on the physician administering the treatment) was only reached after several decades in the 1970s, when the first meta-analyses started ranking trials for the aggregation of evidence (for an instance of such lists of controls, see Bero and Rennie, 1996).

As to the assessment of the comparison, right from the beginning there were serious divergences on what were the proper statistical grounds to judge the difference between interventions (Teira, 2011). The controversy between Fisher and Jerzy Neyman in the 1930s is well documented; for instance, both parties shared a frequentist conception of probability, but disagreed on how to apply it to an experimental test of a given hypothesis. Fisher wanted to rely exclusively on *p*-values: Assuming the truth of the hypothesis that there is no difference between treatments, if we observe a substantive divergence between the two groups, either an exceptionally rare chance has occurred or the hypothesis is not true. But the data alone cannot establish whether the former or the latter is the case (or whether both are). Neyman, together with Egon Pearson, developed a different rationale for the testing of hypotheses: Instead of assessing the plausibility of a single (null) hypothesis, we should have a criterion for choosing between alternative exclusive hypotheses, with a known probability of making the wrong choice in the long run. Fisher considered this latter approach suitable only for industrial quality controls, but epistemically ungrounded. Neyman opened the way instead to a decision-theoretic appraisal of hypothesis testing. Eventually, it led to a full-fledged Bayesian alternative, already emerging in the 1960s and thriving today, in which RCTs can be designed and interpreted without neither randomization nor *p*-values.

Hence, the history of field trials cannot be written, as Leavitt and List suggest, as if once Fisher statistically grounded the design (randomization plus significance testing), it just spread to the medical and socials sciences without controversy. For instance, Stephen Ziliak (2014) shows that even randomization was controversial in proto-economic trials: Already in the 1910s, the pioneer of statistics, *Student* (the pen name of William Gosset) defended a balanced design of agricultural plots as an alternative to randomization, since the confounding factors are not independent and identically distributed.

Historically, field trials emerged among many other approaches to social observation (see, e.g., Maas and Morgan, 2012 regarding economics) and the different controls in their experimental design were tried one by one in a variety of settings. According to Stephen Stigler (1992) and Ian Hacking (1988), the philosopher Charles S. Peirce, together with his student Joseph Jastrow, conducted the first sound randomized experiment in psychology between 1883 and 1884. It tested Fechner's hypothesis about the existence of a perception threshold below which one cannot discern small differences. The experimenter presented different weights to the subject, ordering them according to the cards drawn from a standard deck. Peirce used this randomization scheme as a debiasing procedure (in order to control the subject's guesses about weights), without any explicit statistical implication. Yet the procedure was not adopted in subsequent experiments on the topic, although it was actively used in all sorts of experiments on telepathy for several decades.

According to Trudy Dehue (1997), by the end of the nineteenth century, out of Fechner's psychophysics research program emerged a causal approach to the study of learning transfer (subjects were asked to perform an ability task with either the left or the right hand to measure gradual performance improvement and then they were tested to see whether the trained skill had transferred to their other hand). This causal approach, of which randomization of treatments was already an integral part, was gradually adopted by psychologists of education, led by the Columbia psychologist Edward L. Thorndike, and eventually came to be increasingly implemented at a major scale in American educational research. Here, still according to Dehue, the Stanford psychologist John Edgard Coover presented the control group in 1907 as a methodological device to correct

a confounder in training experiments: Anyone who takes the same test twice is likely to do better the second time – on top of the multiple factors that may affect the subject in between. In the 1910s, control groups were used in experimental educational research conducted in schools on the most effective means to teach. In the 1920s, we found a variety of comparative studies with different controls (Oakley, 2000, pp. 164–168) and different degrees of quantitative analysis. For instance, after graduating with Thorndike in Columbia, William McCall (1923) started testing the validity of classroom practices with control groups, using randomization as an allocation procedure.

The 1920s were the age of quantitative research in American social science (Porter and Ross, 2003; Oakley, 2000, pp. 163–197). Although educational research was gradually left behind, psychology became a science of controlled experiments, with controls intending to warrant bias-free causal analysis, favoring the laboratory to the field. This is how it features in the methodological handbook of the Social Sciences Research Council (SSRC, 1932), a major funding body of American social science at that point. Statistics, and in particular statistical economics, dealt instead with experimentally uncontrolled factors. Nonetheless, other social scientists tried to do controlled field research: still in Chicago and under the influence of Charles Merriam, the political scientist Harold Gosnell (1927) tried to find a third way studying the factors influencing voter turnout on 6000 Chicago citizens. Gosnell divided them into an experimental and a control group organized by place of residence; the former received notices urging them to register and vote, and their turnout was followed up in elections in 1924 and 1925.

The statistical techniques implemented in these early field experiments (often Pearson's correlation theory) did not contribute much to differentiate between true and spurious causes (Turner, 2007). The weight of causal inference still lied on the controls, where randomization was not always the most prominent. For example, another Columbia graduate, the sociologist Francis Stuart Chapin (1947), introduced the term *ex post facto experiment* to name studies where the conditions are not manipulated but selected by the investigator and the comparison is controlled by matching groups – considered a licit alternative to randomization.

Fisher's statistical foundations were explicitly incorporated into the design of these trials already in the 1940s (e.g., Lindquist, 1940). Yet, in fields such as educational research this approach apparently did not yield significant results stimulating further research (Oakley, 2000). However, field trials in the social sciences did not take off until the 1960s (Dehue, 2001). Just as it happened with RCTs in medicine, its success owes much to American regulatory pressure: Several Acts enacted by Congress in the early 1960s mandated evaluation of the programs implemented, just as the 1962 Food and Drug Administration required 'adequate and well-controlled clinical studies' for proof of efficacy and safety of new treatments. However, whereas in the three following decades thousands of RCTs would be conducted, in the social sciences there were only a few hundreds. A paradigmatic field trial of this "golden age of evaluation" (Oakley, 2000) is the New Jersey Income Maintenance Experiment (NJIME), conducted from 1968 to 1972, with nearly 1500 participant households at a cost of $7.8 million (1971). Based on the doctoral dissertation of Heather Ross (1970), the trial tested a welfare policy suggested by the economist Milton Friedman: a negative income tax, by which people earning below a certain amount would receive supplemental pay from the government instead of paying taxes to the government. The general intuition is that, unlike indirect assistance schemes, this negative tax would give people incentives to work and sustain themselves. The NJIME tested whether different combinations of income guarantees plus a tax rate on other income had an effect on the work effort of the participants or prompted any other lifestyle changes. This was all followed by questionnaires every 3 months. The initial assessment of the outcome concluded that work effort did not decline in the treatment group (measured by the number of hours of employment), although further analyses of the data pointed out that there was indeed a reduction. The negative income tax worked better in this respect than some welfare schemes though, but it did not outperform others. The NJIME had a control group and treatment allocation was randomized, but a number of potential confounders and biases were soon detected (Greenberg and Shroder, 2004, pp. 194–195): for example, misreporting of earnings or misrepresentation of the sources of income of families. Many other trials dealt with a variety of policies and social settings, with different degrees of success – for a review see Oakley (2000), pp. 198–230; Greenberg and Shroder (2004).

During this golden age of trials in the 1960s and 1970s, a new discipline, *evaluation research*, gradually emerged. In close collaboration with administration researchers, truly interdisciplinary groups of social scientists coming from different disciplines, such as economists, sociologists, educational researchers, and psychologists, participated in evaluating the different aspects of implemented policies and programs, making widespread use of both experimental and nonexperimental methods (Alkin, 2012). Central to this team of new policy evaluators was Donald T. Campbell, a Berkeley graduate in psychology, with broad interests in the social sciences, philosophy, and the methodology of causal inference. The publication of Campbell and Stanley's book in 1963 (republished in 1966 under a more general heading) was a major breakthrough in methodological research for the social sciences (experimental and else), and it still is among the most widely cited social scientific methodological works. As we will see, Campbell and his collaborators helped to create a unified language for policy evaluators in their analysis of both experimental and nonexperimental data and their work was instrumental in the listing and systematization of the possible biases that can threaten experiments, some of which are specific to experimentation in the social sciences.

From a statistical standpoint, field trials reached maturity, integrating full-fledge hypothesis testing and power calculations – although the controversy on its proper interpretation has not completely vanished: it is still open to discussion under which circumstances statistical significance provides good enough grounds to interpret the test of hypotheses (e.g., Ziliak and McCloskey, 2008).

The experience accumulated in this golden age showed a number of biases that might interfere in the outcome of the experiment. Levitt and List (2009) list, for instance, the following: *randomization bias*, by which the allocation process

would exclude risk averse participants; *attrition bias*, participants drop out of the experiment more in one of the groups, biasing the comparison; the *Hawthorne effect*, participants change their behavior because they know they are being observed in an experiment; *substitution bias*, by which the participants in the control group violate the protocol seeking alternatives to the treatment they receive. The problem with these biases is that there might not be an obviously effective debiasing procedure (e.g., Teira, 2013b). In other words, the list of controls that would bring about a consensus on the outcome of trial is still under discussion.

In recent years, a third generation of field trialists has emerged with great strength in economics. Led by the Abdul Latif Jameel Poverty Action Lab (J-PAL) at the MIT, a cohort of development economists are testing social interventions in third world countries, where the costs of implementing such programs are significantly lower. The randomization of treatment often takes place in naturally occurring settings, and the participants often remain unaware of the experiment conducted (correcting thus the randomization bias and the Hawthorne effect). Trials are often designed for theoretical purposes (theory testing or parameter measurement), but their current popularity owes much to their promise of delivering evidence-based policies to fight poverty. In this respect, trialists such as Esther Duflo (2010) advocate the necessity of an evidence hierarchy for policy making: such as physicians use evidence-based rankings to decide on prescriptions according to the best available data (with meta-analyses and RCTs on top of the pyramid), social scientists should ground their recommendations according to the same principles. Are field trials as reliable regarding policies as RCTs with drugs? The future of the field will probably depend on the answer social trialists deliver in the forthcoming years (Teira and Reiss, 2013).

Quasi-Experiments

Although an enthusiast and public advocate of social experimentation, Campbell was well aware of the fact that randomization was often not feasible in the social realm (due to ethical, administrative, or pragmatic obstacles). With this in mind, Campbell developed the idea of quasi-experiments and developed a system of different designs that fell under the category (Campbell and Stanley, 1963). Although the term quasi-experiment had previously been used by other authors, Campbell fixed and systematized its use to the current one. Quasi-experiments share with all other experiments a similar purpose (to test descriptive causal hypotheses about manipulable causes) as well as many structural details, such as control groups and pretest measures. But, by definition, quasi-experiments lack random assignment of groups to treatments. Assignment to the different interventions can occur either by means of participants self-selection or, alternatively, by means of administrator selection, by which teachers, bureaucrats, or researchers decide who gets what treatment. In the latter case, quasi-experimenters may still have a considerable control over how the assignment is executed.

What gives quasi-experiments their close-to-experimental nature is that the cause is manipulable and occurs before the effect is measured. However, since the allocation to conditions is not performed at random, the inferences that can be made from results are typically limited, because we cannot rule out the existence of systematic differences between groups originating in a confounding factor. These differences will often constitute alternative explanations for the observed effect, in direct competition with the tested treatment. Researchers, though, may explore these alternative explanations and rule some or even all of them through other alternative tests either *ex ante* or *ex post*. However, in cases in which this potential list of alternative explanations is very long, quasi-experimental results may not be conclusive. Quasi-experimental designs need to be very carefully planned in order to minimize the potential threats that stem from the lack of randomization. For this reason, quasi-experiments tend to be more complex than true experiments both in their design and in their analysis (Shadish et al., 2001).

One of the most important outputs of Campbell's methodology is his approach to the question of experimental validity, an approach he transposed from previous concepts of measurement validity in psychology into social scientific experimentation (Heukelom, 2011). Campbell first introduced the terms in his 1957 article "Factors Relevant to the Validity of Experiments in Social Settings." In this first definitional attempt, internal validity was seen as the basic minimum requirement of an experiment, expressed by reference to the following question: "did in fact the experimental stimulus make some significant difference in this specific instance?"; and external validity, in turn, referred to the question of representativeness, or generalizability: "to what populations, settings, and variables can this effect be generalized?"(1957, p. 297). Though certain aspects of Campbell's validity typology have been controversial since the beginning, the distinction between internal and external validity remains a central tenet of the contemporary conceptual assessment of experimental validity (Jiménez-Buedo and Miller, 2010).

After this initial definition, Campbell, together with a number of collaborators gradually reformulated the distinction between internal and external validity and gradually added more validity types into what is now a stable fourfold validity typology: *statistical conclusion validity*, *internal validity*, *construct validity*, and *external validity*. Statistical conclusion validity has been defined as the validity of inferences about the correlation between treatment and outcome. Construct validity refers to the inferences about the higher order constructs that represent sampling particulars (Shadish et al., 2001). These latter two concepts, though common currency in social psychology and related fields in the social sciences, are much less used in more recent experimental fields, like behavioral economics. Often, the first two concepts (statistical conclusion and internal validity) are usually grouped together, whereas construct and external validity are also often conflated, suggesting, in accordance to Campbell himself (Shadish et al., 2001), that all of the categories can be subsumed under a broader, more encompassing, internal/external pair.

Campbell developed this methodological discussion as a reaction against an unreflective use of randomization in social psychology, as if it were the only necessary control to secure the validity of an experiment. Through the years, the validity typology helped to identify and analyze a series of

common sources of biases, or threats to validity. Examples of threats to internal validity are phenomena like *selection* (of units to treatments), *history* (i.e., the existence of concomitant events to the treatment that may be actually responsible for the observed effect), *differential attrition* (or mortality) of subjects in the different experimental groups, or *maturation*, where naturally occurring changes over time could be confused with a treatment effect. All of these are confounding factors that can be mistaken for effects of the implemented treatment. In turn, threats to external validity usually originate in confounders that may lead us to infer erroneously that a causal relation will hold over contexts other than the one under study. Examples of such threats are the interaction of selection and treatment (where the relation between x and y will not hold beyond the type of subject who took part in the experiment), or interaction of setting and treatment (where the relationship between x and y will only hold in the experimental setting). Knowledge about these 'threats' was, for Campbell and his collaborators, a matter of practice: detecting ways in which experimentalists were arriving at wrong inferences from looking at the results of their experiments was thus for Campbell a collective, inductive process and very importantly, relative to the details of a particular study. Therefore, the lists of validity threats never intended to be exhaustive, and many additions (and subtractions, and reclassifications) were made over the years.

The list of threats is still to date routinely used for those purposes by field social experimentalists that contrast it, when choosing experimental designs, with feasibility or budgetary constraints. Campbell though was always careful to remind us that performing these checks did not safeguard any design from a potentially infinite list of threats. In this sense, Campbell showed some concern and went as far as considering the lists potentially 'dangerous if they led readers' to overestimate their capacity as shields from error (1986, p. 154). This overestimation, for its practical implications, seemed a bigger concern to Campbell than the fact, of which he was also aware, that his distinction between internal and external validity had been often misunderstood in ways that portrayed it as having to do with the distinction between field experiments and pure laboratory research (Jiménez-Buedo, 2011).

As regards the choice between lab versus field experiments, there has been, historically, a vivid debate concerning the advantages of both in social psychology. Among the early supporters of field experiments, we find Kurt Lewin (e.g., 1939), who highlighted the need to focus on the balance of forces influencing behavior in social situations, developing methods to perform experiments in everyday life scenarios. During the twentieth century, psychology mostly focused on controlled laboratory experiments, but in the 1960s, authors like McGuire (1969) vindicated again field experiments in social psychology from a methodological standpoint.

The enthusiasm for the field and immersive experimental settings in the 1960s also yielded Stanley Milgram's and Philip Zimbardo's famous experiments on obedience and interpersonal dynamics. Milgram (1963) reproduced a clinical scenario to observe obedience to authority: Participants were instructed to administer electrical shocks to other participants and witnessed the (pretended) consequences of pain and disgust. Zimbardo and colleagues reproduced the structure of a prison in the basement of a Stanford building and instructed participants to behave like wards and prisoners, which resulted in a hostile atmosphere as participants followed the instructed roles (see, e.g., Haney et al., 1973). Such extreme social interactions somehow dissuaded other psychologists from following this experimental path and, as of today, social psychology still hinges more on the laboratory than on the field (Salovey and Williams-Piehota, 2004).

See also: Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Experimental Economics; Experimentation in Psychology, History of; Experimenter and Subject Artifacts: Methodology; External Validity; Field Experiments; Hypothesis Testing: Methodology and Limitations; Internal Validity; Nonequivalent Group Designs; Optimal Experimental Design; Quasi-Experimental Designs; Random Assignment: Implementation in Complex Field Settings; Reliability and Validity Assessment: New Approaches; Selection Bias, Statistics of.

Bibliography

Alkin, M.C., 2012. Evaluation Roots: A Wider Perspective of Theorists' Views and Influences. SAGE Publications, Los Angeles, CA.
Bero, L.A., Rennie, D., 1996. Influences on the quality of published drug studies. International Journal of Technology Assessment in Health Care 12, 209–237.
Campbell, D.T., 1957. Factors relevant to the validity of experiments in social settings. Psychological Bulletin 54, 297–312.
Campbell, D.T., Stanley, J.C., 1963. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L. (Ed.), Handbook of Research on Teaching. Houghton Mifflin, Boston, MA.
Campbell, D.T., Stanley, J.C., 1966. Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin, Boston, MA.
Chalmers, I., 2006. Why fair tests are needed: a brief history. Evidence Based Medicine 11, 67–68.
Chapin, F.S., 1947. Experimental Designs in Sociological Research. Harper, New York.
Dehue, T., 1997. Deception, efficiency, and random groups: psychology and the gradual origination of the random group design. Isis 88, 653–673.
Dehue, T., 2001. Establishing the experimenting society: the historical origin of social experimentation according to the randomized controlled design. American Journal of Psychology 114, 283–302.
Duflo, E., 2010. La politique de l'autonomie. Seuil, Paris.
Goldacre, B., 2012. Bad Pharma. Fourth State, London.
Gosnell, H.F., 1927. Getting Out the Vote; an Experiment in the Stimulation of Voting. The University of Chicago Press, Chicago, IL.
Greenberg, D.H., Shroder, M., 2004. The Digest of Social Experiments. Urban Institute Press, Washington, DC.
Hacking, I., 1988. Telepathy: origins of randomization in experimental design. Isis 79, 427–451.
Haney, C., Banks, W.C., Zimbardo, P.G., 1973. Interpersonal dynamics in a simulated prison. International Journal of Criminology and Penology 1, 69–97.
Heukelom, F., 2011. How validity travelled to economic experimenting. Journal of Economic Methodology 18, 13–28.
Jiménez-Buedo, M., 2011. Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. Journal of Economic Methodology 18, 271–282.
Jiménez-Buedo, M., Miller, L.M., 2010. Why a trade-off? The relationship between the external and internal validity of experiments. Theoria 25, 301–321.
Levitt, S.D., List, J.A., 2009. Field experiments in economics: the past, the present, and the future. European Economic Review 53, 1–18.
Lewin, K., 1939. Field theory and experiment in social psychology: concepts and methods. American Journal of Sociology 44, 868–896.

Lindquist, E.F., 1940. Statistical Analysis in Educational Research. Houghton Mifflin, Boston, MA.
Maas, H., Morgan, M.S., 2012. The observation and observing in economics. History of Political Economy 44, 1–24.
McCall, W.A., 1923. How to Experiment in Education. The Macmillan Company, New York.
McGuire, W.J., 1969. Theory-oriented research in natural settings: the best of both worlds for social psychology. In: Sherif, M., Sherif, C.W. (Eds.), Interdisciplinary Relationships in the Social Sciences. Transaction Publishers, New Jersey, NJ, pp. 21–51.
Milgram, S., 1963. Behavioral study of obedience. The Journal of Abnormal and Social Psychology 67, 371–378.
Oakley, A., 2000. Experiments in Knowing: Gender and Method in the Social Sciences. New Press, New York.
Porter, T.M., Ross, D., 2003. The Cambridge History of Science. The Modern Sciences, vol. 7. Cambridge University Press, Cambridge, MA.
Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.
Ross, D., 1991. The Origins of American Social Science. Cambridge University Press, Cambridge, MA.
Ross, H.L., 1970. An Experimental Study of the Negative Income Tax. Doctoral Dissertation. MIT.
Salovey, P., Williams-Piehota, P., 2004. Field experiments in social psychology. American Behavioral Scientist 47, 488–505.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA.
Social Science Research Council. Committee on Scientific Method in the Social Sciences, Rice, S.A., Secrist, H., Maciver, R.M., Lasswell, H.D., 1932. Methods in Social Science, a Case Book Compiled under the Direction of the Committee on Scientific Method in the Social Sciences of the Social Science Research Council. The University of Chicago press, Chicago, IL.
Stigler, S.M., 1992. A historical view of statistical concepts in psychology and educational research. American Journal of Education 101, 60–70.
Teira, D., 2011. Frequentist versus Bayesian clinical trials. In: Gifford, F. (Ed.), Philosophy of Medicine. Elsevier, Amsterdam, pp. 255–297.
Teira, D., 2013a. On the impartiality of early British clinical trials. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44, 412–418.
Teira, D., 2013b. Blinding and the non-interference assumption in medical and social trials. Philosophy of the Social Sciences 43, 358–372.
Teira, D., Reiss, J., 2013. Causality, impartiality and evidence-based policy. In: Chao, H.-K., Chen, S.-T., Millstein, R. (Eds.), Mechanism and Causality in Biology and Economics, vol. 3. Springer, London, pp. 207–224.
Turner, S.P., 2007. Defining a discipline: sociology and its philosophical problems, from its classics to 1945. In: Turner, S.P., Risjord, M.W. (Eds.), Philosophy of Anthropology and Sociology, first ed. Elsevier, Amsterdam, pp. 3–69.
Ziliak, S., 2014. Balanced versus randomized field experiments in economics: why W.S. Gosset aka "Student" matters. Review of Behavioral Economics 1, 167–208.
Ziliak, S.T., McCloskey, D.N., 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor, MI.

Quasi-Experimental Designs

Kelly Hallberg and Jared Eno, American Institutes for Research, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.R. Shadish, volume 18, pp. 12655–12659, © 2001, Elsevier Ltd.

Abstract

Quasi-experimental designs (QEDs) are a subset of research designs intended to test causal hypotheses about the effect of manipulable causes, generally policies, programs, and treatments, when randomized controlled trials are not feasible. This article provides a summary of efforts to assess the validity of QEDs to provide unbiased causal inference as well as an overview of the most prevalent types of QEDs used in practice, including regression discontinuity designs, comparative interrupted time series, and nonequivalent comparison group designs.

Quasi-experimental designs (QEDs) are a subset of research designs intended to test causal hypotheses about the effect of manipulable causes, generally policies, programs, and treatments. QEDs are primarily defined by what they are not: experiments, also known as randomized controlled trials (RCTs). In the simplest case, RCTs randomly assign units to either receive a treatment (the treatment group) or not receive the treatment (the control group). Random assignment ensures that the treatment and control groups will be roughly equal as long as the groups are of sufficient size. As a result, the impact of the treatment may be measured by simply taking the difference of the average outcomes of the treatment and control groups, although more complex statistical methods are sometimes used to increase the precision with which the treatment effect is measured.

QEDs differ from RCTs in that units (e.g., individuals, schools, communities) are not randomly assigned to treatment conditions. Rather, units self-select into treatment or are assigned to treatment by program administrators through some nonrandom process. As such, unlike in the case of random assignment where the treatment and control groups are equivalent in expectation, units that receive the treatment and those that do not (often called comparison units in the QED literature) may not be comparable before the treatment begins. Simple comparisons of the outcomes of treated and untreated cases will confound any initial differences with any actual treatment effects. Therefore, researchers using nonexperimental data must carefully identify plausible alternative explanations of the apparent effect and rule them out using logic, design, and measurement (Shadish et al., 2002). However, QEDs may be feasible when RCTs are too costly, unethical, or politically unpopular.

Approaches to Assessing the Validity of QEDs

Quasi-experiments actually predate the use of RCTs in the social sciences, but it was not until the 1950s that Donald T. Campbell and his colleagues began to formalize the notion of quasi-experimentation (Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish et al., 2002). The cornerstone of this formalization is a validity typology and the systematic categorization of the associated threats to validity. Validity is understood here as, "the approximate truth of an inference" (Shadish et al., 2002, p. 34). Campbell initially ascribed two types of validity to QEDs. A quasi-experiment is internally valid if it can demonstrate whether the relationship between a treatment and an outcome is causal. A quasi-experiment is externally valid if the causal relationship is true across units, treatments, outcomes, settings, and times (Campbell, 1957; Campbell and Stanley, 1963). The validity framework was later expanded to include construct validity (the validity with which the operations within a particular study appropriately represents that, which it was intended to represent) and statistical conclusion validity (the validity of the correlation between the treatment and the outcome). Of these four types of validity, Campbell saw internal validity as most important, arguing that without internal validity, the other validity types were of little consequence (Cook and Campbell, 1979; Shadish et al., 2002).

Campbell went on to articulate potential threats to each type of validity in practice. These threats can be thought of as alternative explanations or plausible alternative hypotheses for an apparent treatment effect, which could render causal conclusions completely or partially incorrect. The initial list of threats, articulated by Campbell in 1957, included eight threats to internal validity and four threats to external validity. By 2002, the list had grown to include 37 threats (Shadish et al., 2002). The goal of strong quasi-experimentation is to retrospectively, and more importantly proactively, analyze and prevent plausible threats to validity through the inclusion of design features in a quasi-experimental (QE) study. In doing so, the list of threats to validity developed by Campbell and his colleagues serve as a heuristic tool to ensure that possible threats are considered and addressed.

Another important approach for understanding causal inference that is useful for assessing the validity of QEDs is the Rubin causal model, or the potential outcomes model, developed by Donald Rubin and colleagues. This model characterizes causal effects as unit-specific differences between the outcomes that would be achieved under different treatment conditions and provides a clear theoretical rationale for the conditions under which observational studies can produce unbiased estimates of causal effects (Rubin, 1978; Holland, 1986; Neyman, 1935; Heckman, 1979). Under this framework, each unit is characterized by a set of variables ($Y_i(1)$, $Y_i(0)$, X_i, T_i), where $Y_i(1)$ is the potential outcome for the ith unit after having received the treatment, $Y_i(0)$ is the potential

outcome for that same unit had it not received the treatment, X_i is a vector of baseline characteristics, and T_i is an indicator of treatment receipt such that $T_i = 1$ if the ith unit received the treatment and $T_i = 0$ if the unit received the control condition. While values of $(T_i, X_i,)$ are observable for every member of the population, only one of the two potential outcomes $(Y_i(1), Y_i(0))$ can be observed for any single individual. The observed outcome for each unit is formally given as: $Y_i = (1 - T_i) Y_i(0) + T_i Y_i(1)$. However, the *unobserved, counterfactual* potential outcome represents what this treatment recipient's outcome *would have been* if she/he had experienced the alternative treatment. Every unit in the population is missing one of these two potential outcomes and this means that we can never directly measure a treatment–control contrast at the unit level. It should be noted that this formulation of the problem implicitly assumes that each units' potential outcomes depend only on whether or not that individual receives the treatment, not on the treatment assignment of other units. This critical assumption, commonly known as the Stable Unit Treatment Value Assumption, is what makes the desired causal quantity, $Y_i(1) - Y_i(0)$, tractable as a simple mean difference.

While the missing potential outcomes make it difficult to estimate individual treatment effects, it is possible to estimate average group differences. In an RCT, randomization assures potential outcomes are missing at random, so the average treatment effect (ATE), $E[Y_i(1) - Y_i(0)]$, can be estimated using the difference in mean outcomes in the treatment and control groups. Put another way, randomization ensures the mean of those treated cases that we observe, $\overline{Y_i(1)}$, is an unbiased estimate of the potential outcome of $E[Y_i(1)]$ for the population. In the case of QEDs, however, the researcher has no such assurance. Instead, he or she must adopt additional assumptions to estimate average causal effects. We will explore the assumptions needed in several common QED designs below using Rubin's potential outcomes framework.

Both Campbell and Rubin provide a theoretical framework for assessing the validity of QEDs, but one may ask whether these approaches produce trustworthy causal inference in practice. Within study comparisons (WSCs) have been increasingly used to address these questions. WSCs compare results from a QED and RCT that shared the same treatment group. The first WSCs drew on data from RCTs designed to study the effects of job training programs. QE estimates of the same treatment effects were calculated by drawing on extant data sets like the Current Population Survey or the Panel Study of Income Dynamics, and statistical adjustment procedures were then used to determine whether various QE methods would replicate the RCT's causal estimate. The early conclusions were that none of these data analytic adjustment techniques were successful (LaLonde, 1986; Fraker and Maynard, 1987; Friedlander and Robins, 1995; summarized in Glazerman et al., 2003). However, these early WSCs had RCT and QE units that differed in many ways other than whether the comparison group was formed at random or not – the key independent variable of a WSC. In some cases, the comparison units were drawn from distant locales and measured at different times on different scales than in the RCT benchmark. Later studies took more pains to draw comparisons from the same target population as the RCT, that shared the same geographic location, and that were measured at the same time and on the same scale. They also moved beyond asking whether QED methods could replicate RCT results and came to focus instead on the conditions under which QED and RCT methods generated similar causal results. There is now an increasing number of WSCs that have been conducted across domains and have begun to provide empirical evidence on the validity of various QEDs in practice. Key findings from the WSC literature are included in the discussions below.

Types of QEDs

Although there are many types of QEDs, the most commonly used designs include regression discontinuity designs (RDDs), interrupted time series (ITS) and comparative ITS (CITS) designs, and nonequivalent comparison group designs.

Regression Discontinuity Designs

RDDs are used to estimate the effect of a treatment that is assigned based on a threshold for a continuous variable. For instance, students may be assigned to a remedial educational program if they score below a certain cutoff score on an assessment. The use of a threshold for treatment assignment produces a discontinuity in the probability of treatment, which, in the presence of a treatment effect, leads to a discontinuity in outcomes. RDDs assume that units just above or below the threshold are roughly equivalent, and estimate a treatment effect by comparing the outcomes of these two groups. RDD was used infrequently after the original conceptualization in 1960 by Thistlewaite and Campbell, but recently has received increased attention (Cook, 2008; Lee and Lemieux, 2010). The design is now designated as the preferred QED by the Institute of Education Sciences when an RCT is not possible (Schochet et al., 2010), reflecting a widespread belief that it is among the most rigorous of the QEDs (Cook and Wong, 2008).

RDD estimates the counterfactual based on units close to but not meeting the threshold for treatment. The simplest case is a 'sharp' RDD, in which the probability of treatment is completely determined by the threshold of an ordinal or continuous variable, referred to as the 'forcing' variable: $T_i = 0$ if $Z_i < z_c$ and $T_i = 1$ if $Z_i > z_c$ (or the reverse). Suppose that the outcome of unit i depends on the treatment (This approach follows Hallberg et al. (2013); see also Imbens and Wolldridge (2008) and Lee and Lemieux (2010).):

$$Y_i = \begin{cases} Y_i(1) + \varepsilon_i, & T_i = 0 \\ Y_i(0) + \vartheta_i, & T_i = 1 \end{cases}$$

As above, $Y_i(0)$ is the outcome in the absence of treatment, $Y_i(1)$ is the outcome in the presence of treatment, and ε_i and ϑ_i are random error terms. The treatment effect to be estimated is the difference in potential outcomes at the threshold value $\tau = E(Y_i(1) - Y_i(0)|Z_i = z_c) = E(Y_i(1)|Z_i = z_c) - E(Y_i(0)|Z_i = z_c)$. Only one of these potential outcomes is observed at the cutoff. Therefore, outcomes for units close to this cutoff are used instead, and the treatment effect is $\tau = \lim_{z \downarrow z_c} E(Y_i(1)| Z_i = z) - \lim_{z \uparrow z_c} E(Y_i(0)|Z_i = z)$.

'Fuzzy' RDD occurs when the threshold of the forcing variable influences but does not completely determine treatment assignment. In this case, the treatment effect is the ratio of the discontinuity in the outcome to the discontinuity in the treatment $\tau = [\lim_{z \downarrow z_c} E(Y_i(1)|Z_i = z) - \lim_{z \uparrow z_c} E(Y_i(0)|Z_i = z)]/[\lim_{z \downarrow z_c} E(T_i|\ Z_i = z) - \lim_{z \uparrow z_c} E(T_i|Z_i = z)]$. This setup is similar to an instrumental variables approach in which the forcing variable is an instrument for treatment assignment.

For the RDD treatment effect estimates to be internally valid, there can be no discontinuity of potential outcomes at the cutoff (Hahn et al., 2001). If potential outcomes – that is, f_0 or f_1 – were not smooth at the cutoff, this would suggest that other factors also influenced the outcome based on whether the forcing variable met the threshold value. RDD estimates would then confound the effects of the treatment with the effects of these other factors. For instance, if a threshold determined eligibility for two programs, RDDs would not be able to distinguish between their respective effects. Actors manipulating the forcing variable in order to control selection into or out of the treatment may also create discontinuities in potential outcomes that could affect the internal validity of RDD estimates. Tests for discontinuities in the forcing variable or other covariates can provide evidence on whether the continuity restriction holds, although they cannot provide definitive proof (McCrary, 2008).

RDD estimates may also be biased by incorrect assumptions about the relationship of the outcome and the forcing variable at the cutoff. Nonparametric estimation is possible but requires many data points near to the cutoff, which may not be available in practice (Wing and Cook, 2013; Schochet, 2008). Assuming that the relationship takes a specific functional form can support estimation in these cases, but incorrect assumptions will bias estimates. Therefore, weaker or more flexible functional forms are generally favored, such as polynomial functions or local linear regression (Imbens and Lemieux, 2008; Hallberg et al., 2013). Another consideration is how close data points must be to the cut point to be included in the analysis, referred to as the 'bandwidth.' Some approaches to choosing bandwidth have been proposed (Ludwig and Miller, 2005; Imbens and Lemieux, 2007; Imbens and Kalyanaram, 2008), with cross-validation as an important robustness check (Hallberg et al., 2013; Imbens and Wolldridge, 2009). (A further consideration for local regression methods is the choice of the kernel. However, Imbens and Lemieux (2008) argue that because results not robust to the choice of kernel would be unconvincing, there is little benefit to using kernels more complex than a uniform distribution.)

The external validity of RDD estimates is limited because the design focuses on the outcomes of units near the threshold. RDDs estimate a local average treatment effect (LATE) that applies to units assigned to treatment because of the threshold, rather than a general ATE across all units. For sharp RDD, this includes all units close to the threshold. For fuzzy RDD, this includes units close to the threshold who were induced to either take up or abstain from the treatment as a result of their score on the assignment variable, referred to as 'compliers.' The LATE and ATE will differ if treatment effects vary based on the value of the forcing variable.

WSCs have generally found that RDDs can reproduce experimental estimates of LATE estimates. These studies generally restricted experimental samples to units close to the cutoff to estimate the LATE to find general, though not universal, agreement between the two methodologies (Aiken et al., 1998; Buddelmeyer and Skoufias, 2003; Black et al., 2005; see Cook and Wong, 2008 for a review). An additional WSC compared an RDD estimate of the LATE to an RCT estimate of the ATE for units below the cutoff, also finding similar results (Berk et al., 2010). Shadish et al. (2011) randomized units into two different treatment assignment mechanisms, random or cutoff-based. They compared RCT estimates of the former with RDD estimates of the latter, finding general agreement between the two. Wing and Cook (2013) propose adding a pretest measure of the outcome or a comparison group to the RDD. Using a WSC, they find that this can improve the precision, internal, and external validity of RDD estimates.

Comparative ITS Designs

ITS designs estimate a program's effect by comparing an outcome for a treatment group before and after the beginning of a program. Change in the outcomes at the time of the intervention is attributed to the program. CITS approaches add a comparison group to address the possibility that the change in outcomes was due to factors other than the treatment, assuming that these other factors affected the treatment and comparison groups equally.

ITS models estimate the counterfactual by modeling past trends in outcomes and projecting them into the future. Suppose that the outcome of unit i depends on both time t and treatment T_i, such that (This approach follows Wong et al. (2013).):

$$Y_{it} = \begin{cases} f_0(t) + \varepsilon_{it}, & T_i = 0 \\ f_1(t) + \vartheta_{it}, & T_i = 1 \end{cases}$$

Here, f_0 is the course of the outcome in the absence of treatment, f_1 is the course of the outcome in the presence of treatment, and ε_{it} and ϑ_{it} are random error terms. The treatment effect to be estimated is the difference in potential outcomes, $\tau(t) = f_1(t) - f_0(t)$. ITS capitalizes on the fact that parts of both functions are observed before or after the time of treatment, t_c. Before treatment ($t < t_c$), f_0 is observed; after treatment ($t > t_c$), f_1 is observed. The change at the time of treatment is used to estimate the treatment effect; in this way, ITS is similar to an RDD in which time is the forcing variable. As with RDD, a key assumption is the functional form of the outcome around the time of treatment. Bloom (2003) provides a summary of different modeling approaches.

For the ITS treatment estimate to be causal, no factors other than the treatment contribute to the change in outcomes at the time treatment occurs. ITS estimates will be biased in the presence of such simultaneous effects, which may be due to history, selection effects, or instrumentation (Shadish et al., 2002). A comparison group may be added in an attempt to rule out such confounding effects. CITS designs benchmark the change in the treatment group against the change in the comparison group. Here, the *change* in outcomes for the comparison group at the time of treatment serves as the counterfactual for the change in outcomes for the treatment group. Still, any factors

other than the treatment that lead to different changes in outcomes at the time of treatment for the two groups will still cause bias in CITS estimates. The selection of comparison cases therefore becomes an important consideration for CITS.

A modest but growing body of work seeks to examine the validity of ITS analyses through WSCs comparing their estimates to those from experimental designs. Multiple studies have found that ITS or CITS analyses can produce estimates similar to those from RCTs (Schneeweiss et al., 2004; Shadish et al., n.d.; Fretheim et al., 2013; St. Clair et al., 2014) or RDDs (Somers et al., 2012). However, evidence suggests that the validity of ITS analyses depends on the context of the intervention. For instance, St. Clair et al. (2014) find that misspecification of functional form can produce bias in the presence of linear trends in outcomes over time.

Nonequivalent Comparison Group or Matching Designs

When random assignment is not feasible, researchers often seek to identify comparison units that are similar to those receiving the treatment to approximate a randomly assigned control group. While randomization to ensure that the observed treated and control case are similar in both observed and unobserved ways, matching approaches rely solely on comparing treatment and comparison cases that are equivalent on all *observed* characteristics. As such, matching approaches rely on a critical assumption known alternatively as the strong ignorability assumption, the nonconfoundedness assumption, or the conditional independence assumption. This assumption states that

$$Y_i(1), Y_i(0) \perp T_i | \boldsymbol{x_i}$$

That is, matching approaches can produce unbiased causal inference if the potential outcomes are independent of treatment receipt controlling for observed characteristics. If this assumption holds, Rosenbaum and Rubin (1983) show that, at each value of $\boldsymbol{x_i}$, the difference between the treatment and control means is an unbiased estimate of the ATE at that value. If there are both treated and untreated cases $(0 < \mathrm{P}(T = 1|\boldsymbol{x}) < 1)$ for every value of $\boldsymbol{x}$, then we are able to use matching methods to estimate the ATE for the population. If this 'common support' or 'overlap' assumption fails, causal inference is limited to the values of $\boldsymbol{x}$ in which there are both treatment and comparison cases, the region of common support.

In the absence of complete knowledge of treatment assignment process, there are often a large number of pretreatment covariates that could ostensibly influence selection into treatment and the outcomes of interest. In these cases, it is often difficult to find exact matches on each of a wide array of covariates. To address this 'curse of dimensionality,' Rosenbaum and Rubin (1983) introduced propensity score approaches. The propensity score is defined as the conditional probability of treatment receipt given a vector of observable characteristics:

$$e(x) = P(T_i = 1|\boldsymbol{x}_i)$$

The propensity score is often estimated using a logit or probit model to calculate each case's expected probability of treatment given the observable covariates. Rosenbaum and Rubin show that adjustment on the basis of the estimated propensity score is sufficient for removing bias due to all observed covariates. As such, if strong ignorability is met, propensity score adjustment, whether through case matching or stratification, can provide unbiased estimates of causal effects.

Critically, it is impossible to test the strong ignorability assumption empirically. Rather, researchers must rely on substantive theory to determine the plausibility of whether all of the potential confounding variables (those correlated both with selection into treatment and the outcome) have been observed (Rosenbaum and Rubin, 1983).

Qualitative syntheses of WSCs to date have suggested at least three promising cases in which observational approaches seemed to perform well: (1) when researchers condition on pretreatment measures of the outcome of interest (Glazerman et al., 2003; Bloom et al., 2005; Smith and Todd, 2005; Bifulco, 2012); (2) when a rich and reliable set of pretreatment covariates are available (Cook et al., 2008; Shadish et al., 2008; Steiner et al., 2011), and (3) when matches are geographically local (Friedlander and Robins, 1995; Bell et al., 1995; Heckman et al., 1998; Bloom et al., 2005; Bifulco, 2012).

Conclusion

QEDs use logic, design, and measurement to estimate causal effects of programs and policies when RCTs are not possible or feasible. A rich body of literature provides theoretical guidance on the conditions under which sound causal inference is possible in QEDs. Further, a growing number of WSCs have been conducted and are beginning to provide practical guidance to applied researchers implementing these research designs. Findings from these WSCs suggest that RDD, CITS, and some kinds of matching designs *can* replicate experimental findings in practice. More studies are needed to examine how replicable these findings are across contexts and fields. Additional WSC studies that consistently probe the conditions under which QEDs can provide trustworthy results are needed, especially in the area of nonequivalent comparison group or matching designs. Much of this work is currently underway across a variety of venues and we expect that in a few years, even more practical guidance will be available to applied researchers.

See also: Experimental Design: Randomization and Social Experiments; Field Experiments.

Bibliography

Aiken, L.S., West, S.G., Schwalm, D.E., Carroll, J., Hsuing, S., 1998. Comparison of a randomized experiment and two quasi-experimental designs in a single outcome evaluation efficacy of a university-level remedial writing program. Evalaution Review 22, 207–244.

Bell, S.H., Orr, L.L., Blomquist, J.D., Cain, G.G., 1995. Program Applicants as a Comparison Group in Evaluating Training Programs. W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Berk, R., Barnes, G., Ahlman, L., Kurtz, E., 2010. When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. Journal of Experimental Criminology 6, 191–208.

Bifulco, R., 2012. Can nonrandomized estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. Journal of Policy Analysis and Management 31 (3), 729–751.

Black, D., Galdo, J., Smith, J.C., 2005. Evaluating the Regression Discontinuity Design Using Experimental Data. Working paper from: http://www.personal.ceu.hu/departs/personal/Gabor_Kezdi/Program-Evaluation/Black-Galdo-Smith-2005-RegressionDiscontinuity.pdf.

Bloom, H., 2003. Using "short" interrupted time-series analysis to measure the impacts of whole schools reforms: with applications to a study of accelerated schools. Evaluation Review 27, 3–49.

Bloom, H.S., Michalopoulos, C., Hill, C.J., 2005. Using experiments to assess nonexperimental comparison-group methods for measuring program effects. In: Bloom, H.S. (Ed.), Learning More from Social Experiments. Russell Sage Foundation, New York, pp. 173–235.

Buddelmeyer, H., Skoufias, E., 2003. An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA. IZA, Bonn, Germany.

Campbell, D.T., 1957. Factors relevant to the validity of experiments in social setting. Psychological Bulletin 54 (4), 297–312.

Campbell, D.T., Stanley, J., 1963. Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin Company, Boston, MA.

Cook, T.D., 2008. Waiting for life to arrive: a history of regression discontinuity in psychology, statistics, and economics. Journal of Econometrics 142 (2), 636–654.

Cook, T.D., Wong, V.C., 2008. Better quasi-experimental practice. In: Alasuutari, P., Brannen, J., Bickman, L. (Eds.), The Handbook of Social Research. Sage, London.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Designs and Analysis for Field Settings. Houghton Mifflin Company, Boston, MA.

Cook, T.D., Shadish, W.J., Wong, V.C., 2008. Three conditions under which observational studies produce the same results as experiments. Journal of Policy Analysis and Management 27 (4), 724–750.

Fraker, T., Maynard, R., 1987. The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources 22 (2), 194–227.

Fretheim, A., Soumerai, S.B., Zhang, F., Oxman, A.D., Ross-Degnan, D., 2013. Interrupted time-series analysis yielded an effect estimate concordant with the cluster randomized controlled-trial result. Journal of Clinical Epidemiology 66, 883–887.

Friedlander, D., Robins, P.K., 1995. Evaluating program evaluations: new evidence on commonly used nonexperimental methods. American Economics Review 85 (4), 923–937.

Glazerman, S., Levy, D., Myers, D., 2003. Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy 589, 63–91.

Hahn, J., Todd, P., van de Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica 69 (1), 201–209.

Hallberg, K., Wing, C., Wong, V., Cook, T.D., 2013. Experimental designs for causal inference: clinical trials and regression discontinuity designs. In: Little, T.D. (Ed.), Oxford Handbook on Quantitative Methods. Oxford University Press, Oxford.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrika 47, 153–161.

Heckman, J.J., Ichimura, H., Smith, J.A., Todd, P.E., 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.

Holland, P., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142 (2), 615–635.

Imbens, G.W., Kalyanaraman, K., 2009. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. (Unpublished working paper).

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142 (2), 615–635.

Imbens, G.M., Wolldridge, J.M., 2008. Recent Developments in the Economterics of Program Evaluation. National Bureau of Economic Research Working Paper Series, Cambridge, MA.

LaLonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. Annual Economic Review 76, 604–620.

Lee, D.S., Lemieux, T., 2010. Regression Discontinuity Designs in Economics, Journal of Economic Literature, American Economic Association 48 (2), 281–355.

Ludwig, J., Miller, D., 2005. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design (NBER Working Paper 11702). National Bureau of Economic Research, Cambridge, MA.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. Journal of Econometrics 142, 2.

Neyman, J., 1935. Statistical problems in agricultural experimentation (with discussion). Supplement to the Journal of the Royal Statistical Society 2, 107–108.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Annals of Statistics 6, 34–58.

Schochet, P.Z., 2008. Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations (NCEE 2008-4026). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J., Porter, J., et al., 2010. Standards for Regression Discontinuity. Retrieved 15.10.10, from What Works Clearinghouse: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Schneeweiss, S., Maclure, M., Carleton, M., Glynn, R.J., Aron, J., 2004. Clinical and economic consequences of a reimbursement restriction of nebulized respiratory therapy in adults: direct comparison of randomized and observational evaluations: direct comparison of randomized and observational evaluations. British Medical Journal 328 (7439), 560.

Shadish, W.R., Clark, M.H., Steiner, P.M., 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. Journal of the American Statistical Association 103, 1334–1343.

Shadish, W.R., Cook, T., Campbell, D., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston, MA.

Shadish, W.R., Rindskopf, D.M., Boyajian, J.G. n.d. A comparison of results from single-case designs to a randomized experiment. Manuscript under review.

Smith, J., Todd, P., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 305–353.

Somers, M., Zhu, P., Jacob, R., Bloom, H., 2012. The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation (Working Paper). MDRC, New York.

St. Clair, T., Cook, T.D., Hallberg, K., 2014. Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. American Journal of Evaluation 35 (3).

Steiner, P.M., 2001. Propensity Score Methods for Causal Inference: On the Importance of Covariate Selection, Reliable Measurement, and Choice of Propensity Score Technique. Almalaurea Working Papers, no. 9.

Steiner, P.M., Cook, T.D., Shadish, W.R., 2011. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. Journal of Educational and Behavioral Statistics 36 (2), 213–236.

Thistlewaite, D., Campbell, D., 1960. Regression discontinuity analysis: an alternative to the ex-post-factor experiment. Journal of Educational Psychology 309–317.

Wing, C., Cook, T.D., 2013. Strengthening the regression discontinuity design using additional design elements: a within study comparison. Journal of Policy Analysis and Management 32 (4), 853–877.

Wong, V.C., Wing, C., Steiner, P.M., Wong, M., Cook, T.D., 2013. Research designs for program evaluation. In: Velicer, W., Schinka, J. (Eds.), Handbook of psychology: Research methods in psychology, second ed., Wiley, Hoboken, NJ, pp. 316–341.

Queer Anarchism

Jamie Heckert, Anarchist Studies Network, UK
Deric Shannon, Emory University, Oxford, GA, USA
Abbey Willis, University of Connecticut, Storrs, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article briefly draws out some historical and contemporary relationships among queer and anarchist theories and practices, highlighting the rise of a distinctive form of queer anarchism.

This queer anarchism has roots in the earliest anarchist theorists who, like contemporary queer theorists, saw clear links between hierarchies of political economy and rigid constructions of the individual (e.g., vis a vis gender and sexuality). Drawing strength from roots in the overlaps and differences in queer and anarchist traditions, queer anarchism offers cultural, analytical, and practical tools for addressing contemporary challenges to individual and collective freedom, equality, and vitality.

Queer anarchism is a term arising in recent years to signify the overlaps, affinities, and mutual influences between queer and anarchist traditions of theory and practice. Queer theory emerged in the early 1990s inspired by the work of key figures including Judith Butler, Gloria Anzaldua, Eve Sedgwick, Michel Foucault, and Audre Lorde. Noted for its critique of fixed identities, a decentralized understanding of power and an intersectional analysis of oppression, queer theory deconstructs conventional ideas of politics in order to make space for new forms. Soon after, anarchism reemerged on the global stage with the 1999 Battle of Seattle where a loose coalition of anarchists, feminists, labor organizations, environmentalists, and others shut down the World Trade Organization Ministerial Conference. Anarchist(ic) methods of scholarship, horizontal organizing, direct action, and direct democracy played a key role in the success of the protests and continue to do so in the ongoing global social justice movements (e.g., Occupy, the Arab Spring, Idle No More, Zapatismo, and many more) (Bray, 2013; Day, 2005; Gordon, 2008; Kinna, 2012). Both of these seemingly disparate traditions have shared roots in the student/worker uprisings of May 1968, in feminist and queer movements including those which parallel anarchist critiques of hierarchy (e.g., AIDS Coalition to Unleash Power (ACT UP) and Greenham Common Women's Peace Camp), and in a variety of theoretical developments as noted above (Heckert and Cleminson, 2011a). Recognition of these shared roots and the resultant affinities between the critique of normativity in queer theory and activism and anarchism's critique of hierarchy, as well as actually existing queer/anarchist movements (Brown, 2007) and cultural productions (Keenaghan, 2008; Nyong'o, 2009) have inspired the ascendance of a distinctly queer anarchism (Daring et al., 2012; Heckert, 2010; Heckert and Cleminson, 2011b).

A Queer History of Anarchism

Queer theory and anarchism might seem like an odd match to some, given their different intellectual trajectories. Anarchism is often associated with the Enlightenment project of progressive modernist politics; as a class struggle movement to revolutionize society, overthrow capitalism and the state, and institute new social relations based on the common ownership of society in a stateless form of democratic and participatory socialism (Schmidt and Van der Walt, 2009). If one accepts that way of narrating anarchism, then queer theory might, indeed, seem at odds with anarchism. Queer theory, after all, is typically associated with the *post*modern condition and post-structuralist theory. Post-structuralism is often described as *critiques* Enlightenment politics, progressive views of history, and what were, according to many of post-structuralism's advocates, the essentialist politics and grand narratives mobilized by the radical theories of the modern era. Yet many contemporary anarchists are drawing productively on post-structuralist theory, some even sketching out what has come to be called *postanarchism* (Rousselle and Evren, 2011). In many ways, anarchism *anticipated* some of the arguments within post-structuralism broadly and queer theory specifically, providing the roots of what came to be called 'queer anarchism.'

Foucault's (1977a, 1977b, 1980, 1991) critique of a particular Marxist reduction of power to economics and his consequent analysis of decentralized micropolitical forms of disciplinary power, including specifically in the realm of sexuality, was essential to the development of queer theory. In this, Foucault was long preceded by anarchists in the late nineteenth and early twentieth centuries who were also explicitly attentive to institutions like the state and other forms of oppression. The anarcho-syndicalist, Rudolf Rocker, held "that the war against capitalism must be at the same time a war against all institutions of political power," because "exploitation has always gone hand in hand with political and social oppression" (Rocker, 2004: p. 11). Likewise, the Italian anarchist, Errico Malatesta (1920) suggested a *reciprocal* relationship between people and institutions (rather than seeing the economic structure as being 'the real foundation' of social existence): "Between man [sic] and his social environment there is a reciprocal action. Men make society what it is and society makes men what they are, and the result is therefore a kind of vicious circle. To transform society men must be changed, and to transform men, society must be changed."

Similarly, early anarchists often saw their *technique* differently, as a way of acting on/in the world that tended to stress relationality rather than simply a struggle between the workers and the owners of society. This relationality is perhaps best embodied in a quote by the German anarchist, Gustav Landauer (Kuhn, 2010: p. 214), when he argued that "the state is a social relationship; a certain way of people relating to one another. It can be destroyed … by people relating to one another differently." Landauer (Kuhn, 2010, emphasis in original) argued passionately that "(t)he entire system would vanish without a trace if the people began *to constitute themselves as a people apart from the state*." Landauer wrote these words in 1910, yet they have a striking resemblance to contemporary radical theories influenced by post-structuralism that advocate a strategy of exodus out of capitalism. For a queer anarchist approach, the queer critical emphasis on forces of normalization and normativity derived from Foucault and others is the flip side of the anarchist emphasis on hierarchy. At the same time for many anarchists (e.g., Emma Goldman), normalization, uniformity, and sameness were recognized as fundamental to the constitution of authoritarianism. While Foucault refused to align himself with any political identity or tradition, he did acknowledge that his efforts to understand power and the state constituted a form of anarchaeology, taking no power as necessarily to be accepted (Day, 2005).

Transformations of social relations and the challenging of normativity and various forms of power have long included those *within* anarchist movements. Gender norms and relations offer a prime example. Some anarchist men saw "the issue of women's subordination as, at best, secondary to the emancipation of workers, a problem that would be resolved 'on the morrow of the revolution'" (Ackelsberg, 2005: p. 38). Ackelsberg (2005: p. 21) painstakingly outlines the efforts of the Mujeres Libres, an anarchist women's group that "mobilized over 20 000 women and developed an extensive network of activities designed to empower individual women while building a sense of community" during the Spanish Civil War of the early twentieth Century. Ackelsberg (2005: pp. 21, 37) notes that the Mujeres Libres were "dedicated to the liberation of women from their 'triple enslavement to ignorance, as women, and as producers'" largely adopting a theory of anarchism that "insisted that power has its own logic and will not be abolished through attention to economic relations alone." It was quite common for early 'classical' anarchists including Peter Kropotkin, Emma Goldman, and Voltairine de Cleyre to write about (and organize around) issues such as prostitution and sex trafficking, forced sterilizations, and marriage, widening the anarchist critique of hierarchy to give critical concern to women's issues in their own right.

Perhaps more to the point for *queer* anarchism, this was certainly true of anarchist theorizing and organizing around sexuality specifically (though both 'queer' and 'queer anarchism' frequently engage in intersectional analyses and practices not limited to gender and sexuality). Early anarchists had a noted role in a radical politics of same-sex relationships, stressing human freedom and emancipation from repressive sexual norms that saw same-sex love and sexuality as unnatural, sinful, and subordinate (Kissack, 2008). Nearly a century before queer theory's declarations of the limits of identity, Emma Goldman noted that sexual identities cannot (honestly) bound the huge variety of individual experiences of gender and sexuality, that "[i]t is a tragedy, I feel, that people of a different sexual type are caught in a world which shows so little understanding for homosexuals, is so crassly indifferent to the various gradations and variations of gender and their great significance in life" (in Katz, 1976: pp. 378–379). Queer anarchism is committed to a recognition of difference, both through honoring self-determination of gender and other identities (Stanley and Smith, 2011) and by encouraging sabotage of, playfulness with, or nonattachment to, identity (Heckert, 2013; Sallydarity, 2012). Similarly, many early anarchists advocated free love, same-sex love, and support for sex workers (Kissack, 2008). Like queer theory, the politics of the historical anarchist tradition was often relational, creative, and not centered solely on class and those politics expanded beyond humanity's interactions and relationships with each other. Bookchin (1977), for example, notes how anarchists during the Spanish Civil War often experimented with nudism, vegetarianism, and relating differently with the land underneath their feet. Early anarchists also saw a need for developing new relationships with the Earth, nonhuman animals, and even their notion of *selves*. We might also note that like today's queer theorists, anarchist figures such as Emma Goldman, among many others less famous, were also engaged in discursive forms of activism through public talks and writing, which included what today would be called cultural studies (Ferguson, 2011).

Similarly, early anarchists frequently put forward ideas about enacting anarchy that did not follow the progressive politics of radical modernist perspectives. While some radicals argued for communism as an end of history, following discrete stages of development in society (particularly Marx), many early anarchists did not see a clear beginning nor end to their project. Malatesta (1920: p. 75) famously declared that the point was "not whether we accomplish Anarchism today, tomorrow, or within 10 centuries, but that we walk toward Anarchism today, tomorrow, and always." Emma Goldman (1972[1908]), likewise, stated that what she believed in was "a process rather than a finality. Finalities are for gods and governments, not for the human intellect."

One might look, then, for the historical beginnings of queer anarchism within the early anarchist movement around the turn of the nineteenth Century. Anarchism anticipated many of the arguments developed by contemporary theorists of post-structuralism and queer theory. Early anarchists were often quite critical of theories of society that reduced complex human interactions to simple economic functions. They held radical views of gender and sexuality, noting a wide range for gender and sexual practices. They saw a need for relational politics, often in addition to theoretical models of class struggle, which stressed a need to develop new relationships with each other, our selves, and the entire nonhuman world. And rather than narrating history and the future as a series of stages of development with a discrete end to political struggle, many early anarchists stressed anarchism as a process – as an experience – with no definite end. These connections allow us to stress very early relationships between *queer* and anarchism that can, perhaps, problematize the notion that queer anarchism is a particularly *recent* phenomenon.

From those early roots to today, the anarchist tradition is rich in resources for radical queer theory and politics. Anarchist

polymath Paul Goodman, for example, was one of the first intellectuals to publicly reclaim the word queer. Daniel Guerin linked the politics of a hetero/homo divide to wider questions of political economy. Gay and Women's Liberation Movements from the 1970s used anarchic methods in attempts to organize as free equals. Anarcha-feminists around the world including Peggy Kornegger, Takamure Itsue, Mujeres Creando, and many more engage(d) with the intersections of gender, sexuality, and other hierarchies. Anarchist novelists including Ursula LeGuin, Marge Piercy, and Starhawk queer gender, sexuality, and authority. ACT UP, who inspired later queer theorists, developed nonhierarchical methods later taken up by alter-globalization activists. Greenham Common was a site of queer anarchist activism effectively challenging militarism, patriarchy, and compulsory heterosexuality. In each of these cases, critics note that nonhierarchy was often not as well implemented as it might have been. Oppressive practices can still occur within spaces declared anarchist or queer. Anarchist ideals are high and living up to them is an incredible challenge. One of the key challenges anarchism attempts to put into practice, and therefore brings to queer theory, is the task of combining individuality *and* community, or, in other words, making space for difference while being able to live well together. Indeed, contemporary queer theorists including Judith Butler, Lauren Berlant, and Judith Jack Halberstam are increasingly influenced by explicitly anarchist and autonomist movements including Occupy (transnational), Anarchists Against the Wall and Black Laundry (Israel/Palestine), Pink Panthers (Canada), Queeruption (transnational), Church Ladies for Choice (USA), Proyectil Fetal (Argentina), Mujeres Creando (Bolivia), and other queer anarchists groups.

Queering Identity/Politics

Queer theory and politics arose, in part, as diverse responses to sexual identity politics that were based on a centralized understanding of sexual orientation and gender identity functioning as a truth of the self. This understanding lends itself to a state-centered politics of formal equality under the law through the liberal discourse of human or civil rights (Conrad, 2012; Liesegang, 2012; Lydon, 2012). In this situation, a feedback loop can develop whereby attempts to become comprehensible to the gaze of the state encourages a statelike interpretation of gender and sexuality: both are assumed to be static, fixed, and conforming to clear borders (Carr, 1999; Heckert, 2011a; Stanley and Smith, 2011). The strategy is highly successful for some Lesbian, Gay, Bisexual and Transgender (LGBT) people and in certain terms. However, a queer approach, whether in the university, the streets or across/outside any such division, questions the meaning and cost of institutionalized 'success' and offers alternatives (Against Equality, 2013; Brown, 2011; Heckert, 2011b; Jeppesen and Nazar, 2012; Roseneil, 2000).

Insider/Outsider

The first question of a queered identity/politics might be, who ends up outside the borders of identity politics? When politics is defined in terms of state citizenship, there will always be outsiders. Bisexual people, for example, are often labeled in statist terms (e.g., tourist, traitor, criminal, and spy) for not conforming to a hetero/homo border. Transgender, too, can trouble tidy borders of male/female, hetero/homo placing it lower down the sexual minority hierarchy. In addition, the politics of race, class, nationality, and other social divisions are often defined as 'outside' a politics of gender and sexuality. In statist thought, these are all separate categories, separate issues. In lived life, as feminism by women of color has brought to queer theory and to anarchism, oppressions intersect. When sexual politics is imagined to be separate from race or class, it simply means that it is associated with racial and class privilege. In statelike politics, it is those placed on the outside of the border or at the bottom of the hierarchy, which are simultaneously constructed as *having* the problem and as *being* the problem. In a queer anarchist politics, it is the border/hierarchy itself, which is in question. Where identity politics focuses on issues, such as marriage, the military, and other examples of equality within institutions, queer anarchism asks what exclusions, what violence, results from hierarchical institutions? Anarchist groups are more likely to focus campaigns on gentrification, state violence, homo nationalism, queer youth homelessness, sex work, health care, and other issues disproportionately affecting working class and/or queers of color.

Equality

An intertwined query concerns the nature of equality. In state-centered discourse, this refers to formal equality under the law. Queer and anarchist commentators may ask what inequalities are *produced* by law. For example, the legal construction of private property, including that claimed by the legal fictions known as corporations, is fundamental to maintaining local and global inequalities in access to the Earth's common resources including food, fuel, medicine, clean water, land, information, and more. When such a situation is discursively constructed as normal, any politics, which questions its legitimacy is clearly queer. Moreover, anarchism queries the inequality of a small group of people creating laws for entire populations. In anarchism, equality is not a formal matter, a practical one. In practice, how do we learn to relate freely as equals? 'With practice,' anarchist and other traditions show. This includes various forms of direct action and collective decision making, rather than representative democracy, whereby members of collectivities learn to collectively make decisions themselves in order to live, work, and play as free equals. Significantly, for queer anarchism, relating as equals also applies to romantic, sexual, familial, pedagogical, friendly, and other relations deemed to be outside the liberal political sphere.

Queering Representation

This critique of the state as institution is embedded within a broader and deeper questioning of the emphasis placed on representation. Where Deleuze (Foucault, 1977c) spoke of 'the indignity of speaking for others,' queer anarchism might emphasize the dignity of listening. Where many mainstream LGBT equality organizations lobby governments and other bodies by claiming to speak for a minority population, queer

anarchist approaches emphasize a decentralized bridging methodology. Groups form in order to directly address the concerns of their members and network horizontally with other groups also facing oppression. Through cooperation and listening, mutual transformations occur, personal connections are formed, collective lessons are learned, and practical grassroots work addressing social–ecological needs is undertaken. Where institutional success is frequently measured in terms of organizational longevity and measurable outcomes, a queer anarchist approach recognizes the value of experimentation, temporary structures, evolving networks, and cultural shifts imperceptible to the distant biopolitical gaze of the state.

The State of Sexuality

Not only does queer anarchism query speaking for LGBT people, it also asks how the very idea of 'LGBT people' acts as a limiting form of representation. For example, one queer anarchist formulation (Heckert, 2011a) reads sexual orientation as a prime example of Deleuze and Guattari's 'state form.' For Deleuze and Guattari, like Landauer mentioned above, the state is not merely a distant institution but a social pattern replicated at micro- as well as macrolevels. As an institution, it depends upon the reproduction of statelike relationships throughout society. These relationships are characterized by overcoding. If coding is a linguistic necessity, labeling objects in order to communicate about them, overcoding is the reification of those codes, categories, labels over, and above the diversity and complexity of lived experience. In other words, it prioritizes representation over direct relationship, fixity over fluidity, and judgment over compassion. By looking at the statist nature of gender and sexual normativity, queer anarchism acts *to queer* anarchism. By linking normativity to hierarchy, queer anarchism highlights the anarchic impulse within queer theorizing and organizing.

Compassionate Self-Inqueery

Queries regarding equality and representation might also be referred reflexively, compassionately applied by queer anarchist groups to themselves. There is always the potential of queer anarchism to transform from the process of questioning, experimenting, and experiencing into the reified fantasy of itself as a superior identity position (a form of representation). This may be defined as smarter than more mainstream LGBT movements, more socially engaged than queer theorists or more sexy than other anarchists. While disdain or resentment toward movements that hurt, invisibilize or exclude can be effective mechanisms for avoiding pain, disconnecting from the experience of being alive (including pain) can make the practice of liberty, equality, and vitality impossible. Thus, an effective queer anarchism looks not only at external relations but also at one's own relationships with beliefs, desires, emotions, and embodiment. Recognizing these relations as the effects of social conditioning, they need to be seen to represent the judgmental truth of oneself as is encouraged in a society of control. Rather, some might be perceived as patterns helpful for surviving a hierarchical society and helpfully released for building an anarchist one. Queer and/or anarchist is not a state of being but a continual process of becoming. In this sense, then, queer anarchism can help create a form of self-reflexivity that does not rely on blame (toward self or others), but instead looks to understand the external social basis for all of our beliefs, desires, and feelings. Instead of fault, the focus, then, looks toward transformation (creative and destructive).

Queering Anarchism

Queering functions to keep topics of gender and sexuality, emotion and embodiment, and love and desire on the anarchist map. It does so by highlighting the relevance of diverse issues related to sexual categorization and beyond (including Bondage & Discipline; Dominance & Submission; Sadism & Masochism (BDSM), sex work, non/monogamy, pregnancy, gender diversity, asexuality, sex education, communication skills, bodily practices of freedom, and more) to the continuous production of borders and hierarchies, as well as to practices of collective decision making, community building, and libertarian political economy. Queer anarchist methods include creative direct action, cultural productions, radical pedagogy, social research, cultural studies, subtle interventions into dominant patterns, cultivating self awareness, carving out queer autonomous spaces, and various forms of playful subversion. Queering also functions to keep anarchism from becoming that which it aims to replace. In addition to playing with diverse subjects and methodologies, queering anarchism involves experimenting with the meaning of anarchism itself. Anarchism, like any ideology, has the potential to become rigid, sectarian, and abstracted from lived experience. Queering, as an ongoing process of loosening borders, playing with identities, and accepting uncertainties, can help keep anarchism true to its ideals of equality, vitality, and diversity.

See also: Anarchism; Citizenship: Political; Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Queer Geography; Sexual Minorities and Sexual Citizenship; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Ackelsberg, Martha A., 2005. The Free Women of Spain: Anarchism and the Struggle for the Emancipation of Women. AK Press, Oakland.

Against Equality: Queer Challenges to the Politics of Inclusion. Website. http://www.againstequality.org/ (accessed 08.01.13.).

Bookchin, Murray, 1977. The Spanish anarchists: the heroic years, 1868–1936. Free Life Editions, New York.

Bray, Mark, 2013. Translating Anarchy: The Anarchism of the Occupy Wall Street Movement. Zero Books, Ropley, UK.

Brown, Gavin, 2007. Mutinous eruptions: autonomous spaces of radical queer activism. Environment and Planning A 39 (11), 2685–2698.

Brown, Gavin, 2011. Anarchism and amateurism in the creation of autonomous queer spaces. In: Heckert, J., Cleminson, R. (Eds.), Anarchism & Sexuality: Ethics, Relationships and Power. Routledge, Abingdon.

Carr, C. Lynn, 1999. Cognitive scripting and sexual identification: essentialism, anarchism and constructionism. Symbolic Interaction 22 (1), 1–24.
Cleminson, Richard, 2004. Making sense of the body: anarchism, nudism and subjective experience. Bulletin of Spanish Studies 81 (6), 697–716.
Conrad, Ryan, 2012. Gay marriage and queer love. In: Daring, et al. (Eds.), Queering Anarchism: Addressing and Undressing Power and Desire. AK Press, Oakland.
Daring, C.B., Rogue, J., Shannon, Deric, Volcano, Abbey, 2012. Queering Anarchism: Addressing and Undressing Power and Desire. AK Press, Oakland.
Day, Richard J.F., 2005. Gramsci is Dead: Anarchist Currents in the Newest Social Movements. Between the Lines, Toronto.
Ferguson, Kathy E., 2011. Emma Goldman: political thinking in the streets. Rowman & Littlefield, Lanham.
Foucault, Michel, 1977a. Discipline and Punish: The Birth of the Prison (A. Sheridan, Trans.). Pantheon, New York.
Foucault, Michel, 1977b. The History of Sexuality, vol. 1: An Introduction (R. Hurley, Trans.). Pantheon, New York.
Foucault, Michel, 1977c. In: Bouchard, Donald F. (Ed.), Language, Counter-Memory, Practice: Selected Essays and Interviews by Michel Foucault. Cornell University Press, Ithaca, NY. Available online: http://libcom.org/library/intellectuals-power-a-conversation-between-michel-foucault-and-gilles-deleuze.
Foucault, Michel, 1980. Power/Knowledge: Selected Interviews and Other Writings 1972–1977 (C. Gordon et al., Trans.). Pantheon, New York.
Foucault, Michel, 1991. In: James Goldstein, R., Cascaito, James (Eds.), Remarks on Marx: Conversations with Duccio Trombadori. Semiotext(e), New York.
Goldman, Emma, 1972[1908]. What I believe. In: Shulman, A.K. (Ed.), Red Emma Speaks: An Emma Goldman Reader. Humanity Books, Amherst.
Gordon, Uri, 2008. Anarchy Alive: Anti-authoritarian Politics from Practice to Theory. Pluto, Ann Arbor.
Heckert, Jamie, 2004. Sexuality/identity/politics. In: Purkis, Jonathan, Bowen, James (Eds.), Changing Anarchism: Anarchist Theory and Practice in a Global Age. Manchester University Press, Manchester, pp. 101–116.
Heckert, Jamie, 2010. "Anarchism and sexuality", a special issue of sexualities. 13 (4).
Heckert, Jamie, 2011a. Sexuality as state form. In: Rousselle, D., Evren, S. (Eds.), Postanarchism: A Reader. Pluto, Ann Arbor.
Heckert, Jamie, 2011b. On anarchism: an interview with Judith Butler. In: Heckert, J., Cleminson, R. (Eds.), Anarchism & Sexuality: Ethics, Relationships and Power. Routledge, Abingdon.
Heckert, Jamie, 2013. Another state of mind is possible: anarchism and psychology. Social and Personality Psychology Compass 7 (8), 513–525.
Heckert, Jamie, Cleminson, Richard, 2011a. Ethics, Relationships and Power: An Introduction. Anarchism & Sexuality: Ethics, Relationships and Power. Routledge, Abingdon.
Heckert, Jamie, Cleminson, Richard, 2011b. Anarchism & Sexuality: Ethics, Relationships and Power. Routledge, Abingdon.
Jeppesen, Sandra, Nazar, Holly, 2012. Genders and sexualities in anarchist movements. In: Kinna, R. (Ed.), The Continuum Companion to Anarchism. Continuum, London.
Katz, Jonathan, 1976. Gay American History: Lesbians and Gay Men in the U.S.A. Thomas Y. Crowell Company, New York.
Keenaghan, Eric, 2008. Life, war, and love: the queer anarchism of Robert Duncan's poetic action during the Vietnam war. Contemporary Literature 49 (4), 634–659.
Kinna, Ruth (Ed.), 2012. The Continuum Companion to Anarchism. Continuum, London.
Kissack, Terence, 2008. Free Comrades: Anarchism and Homosexuality in the United States 1895–1917. AK Press, Oakland.
Kuhn, Gabriel (Ed.), 2010. Revolution and Other Writings: A Political Reader. PM Press, Oakland.
Liesegang, Jerimarie, 2012. Tyranny of the state and trans liberation. In: Daring, et al. (Eds.), Queering Anarchism: Addressing and Undressing Power and Desire. AK Press, Oakland.
Lydon, Jason, 2012. Tearing down the walls: queerness, anarchism and the prison industrial complex. In: Daring, et al. (Eds.), Queering Anarchism: Addressing and Undressing Power and Desire. AK Press, Oakland.
Malatesta, Errico, 1974[1920]. Towards anarchism. In: Graham, M. (Ed.), Man! Cienfuegos Press, London, pp. 73–78.
Nyong'o, Tavia, 2009. Unfinished situations. Women & Performance: A Journal of Feminist Theory 19 (2), 187–203.
Rocker, Rudolf, 2004. Anarcho-syndicalism: Theory and Practice. AK Press, Oakland.
Roseneil, Sasha, 2000. Common Women, Uncommon Practices: The Queer Feminisms of Greenham. Cassell, London.
Rousselle, Duane, Evren, Süreyyya (Eds.), 2011. Post-anarchism: A Reader. Pluto, London.
Sallydarity, 2012. Gender sabotage. In: Daring, et al. (Eds.), Queering Anarchism: Addressing and Undressing Power and Desire. AK Press, Oakland.
Schmidt, Michael, van der Walt, Lucien, 2009. Black Flame: The revolutionary class politics of anarchism and syndicalism. AK Press, Oakland.
Stanley, Eric, Smith, Nat (Eds.), 2011. Captive Genders: Trans Embodiment and the Prison Industrial Complex. AK Press, Oakland.

Queer Anthropology

Cymene Howe, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Queer anthropology is defined as the analysis of social systems, practices, and symbolic resources in contexts of sexual difference from the (presumed) norms of heterosexuality and gender ascription. This article argues that while the study of sexuality has enjoyed a place in anthropology since its inception, homoerotic or homosexual practices and identities have only relatively recently secured legitimacy as objects of study; this overview of the subfield focuses on queer anthropology's contributions to the discipline of anthropology and to transdisciplinary gender and sexuality studies more broadly.

One might argue that the discipline of anthropology has always been a little bit 'queer.' Historically, it has taken human marginality as its object of study and it has consistently demonstrated that social norms are contingent. It was Malinowski himself after all who insisted that sexuality and sexual practices were not simply vehicles for legitimating (heterosexual) marriage but, rather, a more polysemic set of social behaviors. However, one could also argue that the discipline of anthropology has been quite un-queer. Although the study of sexuality has enjoyed a place in the discipline since its inception, homoerotic or homosexual practices and identities have only recently found legitimacy as objects of study in their own right. As recently as the early 1990s, lesbian and gay studies in anthropology were still being described as an 'emergent field' within the discipline (Weston, 1993). I begin with this paradoxical argument in order to index the ways in which queer anthropology is an uneasy category. As a subfield, or perhaps better put – using the terminology of queer communities – as a 'bent' within the discipline, queer anthropology has been a particularly adaptive genre of inquiry. Queer, same-sex, or lesbian, gay, bisexual, transgender sexuality has never enjoyed the hallowed status of anthropology's hallmark objects of study such as indigenous peoples, kinship dynamics, or political systems. Yet, the subfield has engaged canonical anthropological topics, often reinterpreting their assumptions and providing a queered vantage on classical questions. Provocative new readings on language and kinship that have sought to look beyond heteronormative hegemony, for example, have enlivened debates not only in queer anthropology but in the discipline more broadly.

Queer anthropology can be broadly defined as the effort to examine social systems, practices, and symbolic resources that emerge in contexts of sexual difference from the (presumed) norms of heterosexuality and gender ascription. While queer anthropology has, by definition, contranormative impulses, it is also true that the subfield has struggled to be recognized and legitimated; queer anthropology has had to balance its political, theoretical, and substantive penchant for subversions against conceits that would consign it to triviality, lewd preoccupation, or a 'minority' concern. Like feminist anthropological work that was so foundational to its evolution, queer anthropology is an experimental mode of ethnographic practice. It prides itself on innovative, transdisciplinary engagements and, at times, rather radical departures from the status quo. While 'queer anthropology' is a capacious, contemporary way to describe anthropological work that examines sexual diversity, both ethnographically and theoretically, queer anthropology is also a designation that has its limits and controversies. In order to uncover these contingencies, queer anthropology must be situated in historical perspective and grounded in earlier anthropological studies of sexuality. However, this does not mean defining a linear trajectory from A to Q; instead, the queer anthropological works I overview here will be juxtaposed thematically rather than chronologically, highlighting focal areas of research and key questions that have emerged within the field. With a better understanding of the issues and insights that have animated queer anthropological analysis we can then turn to potential anthropological futures for cultural erotics and sexual imaginaries.

Histories of Sexuality and Queer Paradoxes

Anthropology has never been particularly prudish in its treatment of sexuality. But historically, it has been inclined to engage 'human sexuality' broadly, privileging heterosexual relationships and practices over nonheterosexual ones. In the early twentieth century, Bronislaw Malinowski endeavored to understand sexual practices and nonmarital 'amorous adventures' among the Trobriand Islanders. Margaret Mead was also rather famously fascinated by gender and sexual roles in the South Pacific, including premarital erotic contact among adolescents and Samoans' treatment of *moetotlos* ('sleep crawlers'). Ruth Benedict, one of Mead's mentors and one of anthropology's founding foremothers, suggested that concepts of 'deviance' were relative and contingent upon social circumstances. Her psychologically oriented research offered a challenge to the presumed psychoanalytic 'truth' of sexual deviance that was dominant at the time. It is also a rather open secret that both Benedict and Mead were far from heteronormative in their practices. E.E. Evans-Pritchard's collection of pre-European contact narratives gathered between the 1920s and the early 1960s in southern Sudan among the Azande, was another early surfacing of queer anthropology. Pritchard's discussion of 'male wives' (Evans-Pritchard, 1970) demonstrated how transgenerational and ranked homosexual relationships were founded and maintained in the hypermasculinized social

domains of Azande warriors. He also indicated that female same-sex sexual relationships were (although prohibited) prevalent in polygamous households. However, Benedict's critiques of social deviance models and Evans-Pritchard's 'queer' moment were but two instances within a more general tendency in the discipline to view sexuality through a decidedly heterosexual lens. The Human Relations Area Files (HRAF), for instance, a catalog of ethnographic data from nearly everywhere on the planet, indexed levels of 'sexual promiscuity' as well as levels of 'tolerance for homosexual behavior' within civilizational groups. The HRAF model was symptomatic of a more generalized western fascination with sexual taboos and anxieties about how sexuality operated to stabilize or threaten heterosexual marriage and reproduction. Given the history of western worries about sexuality in general – as a polluting practice, as a dangerous endeavor, and as a sinful preoccupation (Douglas, 1966; Vance, 1984) – anthropology's ability to establish the study of sexuality as a scientifically legitimate project is notable and even laudable. These were the beginnings of analytic perspectives that had the potential to illustrate the diversity of the human condition from the vantage of sexual behavior and sexual sensibilities.

Defining what constitutes 'homosexuality' was, and continues to be, an unsettled question. A relatively early debate in queer anthropology is a case in point. Working with 'the Sambia' (a pseudonym) in Papua New Guinea, Gilbert Herdt (1984) described a series of institutionalized encounters where Sambian boys were expected to ritually fellate their older male mentor. Boys were required to physically ingest the masculine power that resided in a particular fluid: semen. This was, according to Herdt's analysis, how boys became fully socialized men. Herdt's study, unsurprisingly, provoked considerable reaction, in part because it involved sexual partners of different ages. For many western readers, this was a culturally institutionalized form of child sexual abuse. It was a test of anthropological relativism to be sure. But beyond a kind of moral provocation, Herdt's study animated other questions about what specifically constitutes 'sexual' practice and 'homosexual' behavior in particular. Herdt described Melanesian semen ingestion practices as 'ritualized homosexuality,' suggesting that they challenged heterosexual sureties and disputed the cross-cultural reproductive logics of heterosexuality. Deborah Elliston (1995), however, argued against Herdt's interpretation. She claimed that to call these acts 'ritualized homosexuality' was to export a Western model of sexuality to non-Western societies. For Elliston, Melanesian semen practices ought to be understood not through the western category of homosexuality but as a comment upon the value of fluids in Melanesian societies. The debate between Herdt and Elliston, though it may have been limited to a particular region and historical moment, was paradigmatic because it animated a series of questions about the proper 'object' of study for queer anthropology.

Are two same-sex persons who regularly engage in same-sex genital contact 'homosexual' by definition (and, by whose definition)? Or to pose the question in a more philosophical mode, are these practices necessarily 'sexual' or are they instead dependent upon the meaning given to eroticism, sexuality, and sexual subjectivity in any given case and context? Should a queer anthropology center attention on the lives of those who explicitly define themselves as 'lesbian,' 'gay,' 'queer,' or some other sexually inflected term of identity? Or should the subfield 'queerly' analyze the many social dynamics with which human beings engage? These kinds of theoretical propositions are ongoing debates within the field, and, I would argue, are not necessarily meant to be resolved. Rather these are opportunities to test and reset the boundaries of queerness, homosexuality, and sexual and gendered subjectivity. This is also precisely what queer anthropology has been adept at doing. Rather than providing comforting answers, queer anthropology has been far more interested in probing assumptions. In good deconstructionist form, it has been, and continues to be, an exercise in principled interrogations of sexuality as it is manifested in social relationships, identifications, affective practices, and political propositions.

Sexuality in the Field: Reflexivities and Methodologies

The designation 'queer anthropology' has been a point of debate for good reason. Many of the early works that would now fall under the rubric of 'queer anthropology' were not originally conceptualized as queer at all. Instead, importance was placed on their specific attention to 'lesbian' and 'gay' issues. Kath Weston (1993) notes that it was not until the 1960s that the classic studies in gay and lesbian anthropology were published. Research on homosexuality or transgendering continued to be conditioned by 'fear of professional repercussions' and 'reticence on the part of ethnographers' (1993: p. 340) for the next several decades. Queer ethnography tended to follow an 'ethnocartographic' logic "looking for evidence of same-sex sexuality and gendered ambiguity in 'other' societies" (Weston, 1993: p. 341). *The Many Faces of Homosexuality* (Blackwood, 1986), for example, featured essays on same-sex practices and identifications in a variety of national contexts. However, the volume marked an important advance since, unlike many academic disciplines analyzing sexuality, anthropologists were willing to cross borders and depart from the 'west.'

Institutionally, queer anthropology was also becoming established. In the 1970s, a handful of pioneering anthropologists founded ARGOH (The Anthropology Research Group on Homosexuality) within the American Anthropological Association in order to encourage research on gay and lesbian communities and to condemn (what was called at the time) 'homoerotophobia.' Membership expanded in the late 1980s and the organization changed its name to SOLGA (The Society of Lesbian and Gay Anthropologists). However, the nominal emphasis in the Society's title centered on the identity of the anthropologist him or herself, rather than on lesbian and gay subjects. Many, but not all, anthropologists researching sexual diversity identified themselves as lesbian, gay, or queer. While some forms of nonheteronormative sexuality that anthropologists were encountering across the world could be categorized as 'lesbian' or 'gay,' many others were not identity-based, or combined autochthonous forms of same-sex erotics with globally disseminated terms (Blackwood, 2010; Boellstorff, 2005; Howe, 2013; Sinnott, 2004). For the emerging field, it was critically important to designate an intellectual space for

(specifically) lesbian and gay anthropology. At the same time, it was clear from the outset that the categories of gay and lesbian could not fully describe transnational sexual identities, erotic experiences, or naming practices.

The 'queer turn' of the early 1990s in the humanities and social sciences was an auspicious intellectual shift for queer anthropology. Inspired in great part by Foucault's *History of Sexuality* and soon after, among others, Judith Butler's *Gender Trouble*, Michael Warner's *Fear of a Queer Planet*, Eve Sedgwick's *Epistemology of the Closet*, Gloria Anzaldúa's work on race, nation, and sexuality and Gayle Rubin's (Rubin, 2012) analyses of sex/gender systems, the development of queer theory was a rather epochal event. As a capacious umbrella category of belonging, queerness became a designation for many 'alternative' configurations of sexuality, gender, and desire. Yet, 'queer' was not without its problems. The term was originally derogatory, an insult that held a certain amount of venom in later iterations as it was appropriated as a term of empowerment and political progressiveness through organizations such as Queer Nation. Second, the concept was patently northern in origin (as an English language word). As it began to be applied in non-western settings, 'queerness,' whether in its subjective form or its theoretical life, still carried 'baggage' as a concept that was invented and reworked largely in the global north. Anthropologists, unsurprisingly given their methodological and theoretical orientation toward relativism, were cautious of projects that would erase cultural and social particularities, which 'queer,' nominally, seemed to risk.

The transition from lesbian and gay anthropology to queer anthropology was not without contention as scholars weighed the political, analytical, and categorical values of refashioning the field as a specifically queer anthropology (Boellstorff, 2007). In 2010, SOLGA was officially retitled as AQA (the Association for Queer Anthropology), thus claiming a queer designation while also removing the emphasis on the anthropologist's identity. These reflexive deliberations within queer anthropology also found themselves replayed in important methodological engagements. In *Out in the Field: Reflections of Lesbian and Gay Anthropologists* (Lewin and Leap, 1996), for example, ethnographers describe the contingencies of negotiating one's sexuality in the context of field research. In some cases, being marked as sexually 'other' is a liability and a struggle, and in other cases it is an advantage. Discussions of how sexual identity helps or hinders the research process raised broader concerns about how sexuality has (undoubtedly, always) conditioned the subjectivity of field researchers as well as how they are interpreted by their interlocutors. Lewin and Leap's collection posed the larger question of how sexuality and sexual identifications are managed in professional settings. The methodological paradigms of queer anthropology have simultaneously asked, in reflective and reflexive ways, how sexuality marks the field and how anthropologists themselves are marked by their sexuality.

Kinship, Sociality, and Queering Community

While anthropology may have always been a little bit queer, it is also true that queer anthropology has often tried to be concertedly 'anthropological.' Kinship, for example, one of the classic anthropological objects of inquiry, has been an important initial, and continuing, engagement for queer anthropology. In *Families We Choose* (1991), Kath Weston delineated the ways in which lesbian and gay families in the San Francisco Bay Area established family networks that were not predicated on biological or 'blood' relations, but rather on networks, connectivities, and shared values. Weston's interpretation of queer kinship practices offered a counterpoint to the dominant, prevailing thesis that American kinship was (necessarily) bound by rules of consanguine relationality. Ellen Lewin's work on lesbian and gay commitment ceremonies (1998) and lesbian motherhood (1993) illustrated both divergences and congruencies between queer experiences and normative practices of reproduction, kinship, and sexual identity in the United States. Comparing lesbian mothers and divorced heterosexual mothers in the United States, beginning in the late 1970s, Lewin found parallels between lesbian women's coming out experiences and straight women's divorces: each established greater autonomy. Being a mother, in both cases, emerged as a central identity and basis for relationships, illustrating how kinship profoundly determines social value, even, or perhaps especially, in the context of queer communities.

Queer kinship studies were at the forefront of a broader conversation regarding rather dramatic changes in family forms brought about by technology, globalization, and new opportunities for parenting and partnership. Echoes of these changes continue to resound as a new generation of queer anthropologists embark on the legal and cultural dimensions of same-sex marriage in the United States, especially as same-sex partnerships are, increasingly, becoming legally codified and receiving increased public support. Queer parenting practices also continue to be an important area of research, exploring, for example, how gay men conceive of their roles as co-fathers (Lewin, 2009). The next challenge for queer kinship studies will be to move discussion beyond the United States and Europe where it has largely focused and to extend analyses of queer parenting and partnerships to other regions of the world that have been conditioned by different histories and where we might expect very different futures. Discussions about queer kinship, family formation, and generational commitments are elements of a broader context that includes queer sociality and community formation, each of which has been a mainstay for queer anthropology.

Sexual identity has often served as a basis for solidarity and a shared sense of identification. However, because homoerotic desire and sexual orientation do not always, or in all places, translate into a collective logic, queer anthropology has sought to distinguish how sexuality operates in conjunction with other elements of social belonging. Mary L. Gray's *Out in the Country* (2009), for instance, demonstrated how queer youth in rural Kentucky have appropriated public spaces (such as high schools), neoliberal private institutions (such as Walmart), and social media spaces on line, to establish rural queer visibility. Gray's focus on rural queer populations shifted the optic of queer visibility from its usual locale in the urban metropole and illustrated important differences about queer lives beyond the 'meteronormative' focus of many queer studies. In the *Spaces between Us: Queer Settler Colonialism and*

Indigenous Decolonization (Morgenson, 2011), Scott-Morgenson illustrates how Native American sovereignty articulates with queer sociality, showing how white settler colonialism has been a primary condition for the development of modern queer politics in the United States. In *The Empire of Love* (2006), Elizabeth Povinelli contrasted indigenous communities in Northern Australia and a progressive, queer, radical fairy contingent to focus attention on what she calls 'social intimacy,' rather than identity. Her analysis, among others, signals a move within queer anthropology (and in queer theory more broadly) to examine postidentity paradigms. Moving the 'post' yet further is anthropological work that extends the category of queerness to include non-normative sexual practices that may or may not be same-sex or same-gendered. In *Techniques of Pleasure: BDSM and the Circuits of Sexuality*, Margot Weiss (2011) demonstrates how consumerism, notions of self-mastery, and strict regulatory regimes of behavior all operate to constitute a collective ethos among bondage, domination and submission (BDSM) communities. In this work and others, queer anthropology is not solely defined by same-sex sexuality, but rather more expansively. Here, queer anthropology echoes queer and postcolonial theories which suggest that dominant discourses are always embedded in, putatively counterhegemonic communities. Radically different forms of sociality are possible, but normative values are not easily sloughed off. Ultimately, subjectivity is not singularly aroused through sexual orientation, erotic behaviors, or affective desires, but instead defined by multiple, intersecting factors including gender, race, socioeconomic status, (post)-colonial conditions, and geolocation.

Performance and Performativity: Language and Gender

The relationship between genders and sexual behavior, as well as the values associated with both, has been a pivotal debate in the field of queer anthropology. Early on, Gayle Rubin articulated partial economic and feminist concerns, establishing a key theoretical device for sexuality studies including queer theory itself. In 'Thinking Sex,' Rubin described how some sexual acts are negatively marked as unnatural while others remain in what she called the 'charmed circle' of sexual behavior; that is, those that are at least nominally qualified by reproductive, marital, or romantic purposes. Sex/gender systems, as well as their origins and effects, would prove to be a critical domain for queer anthropological analysis. *Two-Spirit People: Native American Gender Identity, Sexuality, and Spirituality*, for example (1997), challenged the ways in which non-normative gendered behaviors have been, at times, mistakenly attributed as gay or lesbian identification. Through the voices of Native American two-spirit people themselves, the book uncovers the intimacies of these misapprehensions (including on the part of anthropologists) and in turn, further complicates questions surrounding the definitional boundaries of gendered and sexual being. Anthropological work in South Asia has demonstrated similar negotiations. Ethnographies centered on the lives of *hijras*, whom Serena Nanda (1998) called 'neither man nor woman,' have provided regional and historical lineages to practices of gender and sexual diversity. As Gayatri Reddy (2005) found, the third gendered lives of *hijras* are also shaped by spiritual commitments that provide an important point of definition for *hijra* subjectivity.

Queer anthropology has also grown through its engagements with the performative elements of gender and sexuality, as Rosalind Morris detailed in her review essay, 'All Made Up: Performance Theory and the New Anthropology of Sex and Gender' (1995). Esther Newton's *Mother Camp* (1979) was an early example of how sexual subcultures of camp and drag reproduced forms of inclusion through performances, ritualized over time. Later, Newton would revisit drag in her study of Cherry Grove (1993) a long time lesbian and gay enclave. Employing a historical perspective on the region's development, Newton also underscored how dimensions of class and gender functioned to establish status differentials between men and women. The fascination with drag, for Newton as for many queer theorists, has centered on how it re-performs gender bending and in turn, works as a deconstructive exercise of gender norms. Drag is a relatively small gesture that reveals worlds about the values surrounding gender systems. The iterative qualities of gender performance have been important to queer anthropological texts seeking to illuminate how individuals and collectives make, and make sense of, their social location.

Many gendered and sexed practices cannot be easily categorized as feminine/masculine, female/male, or homosexual/heterosexual. Whether under the mantle of 'trans' (-gender, -sexual, -vestism), or 'third' (or fourth) gender, these embodied and socially situated experiences have been an important area of research for queer anthropology. Don Kulick's (1998) vivid ethnography details the lives of Brazilian *travestis* – biological men who dress in feminine clothing, some of whom have taken radical steps to make their bodies more *feminino*. Injecting industrial silicon or motor oil to create the appearance of hips, breasts, and larger buttocks (or *bundas*), *travestis* who make their living through sex work often see these steps as a way to garner better pay. But more important is how these exercises in embodiment signal a greater desire to be, as Kulick puts it, 'not men.' David Valentine's *Imagining Transgender: An Ethnography of a Category* (2007) provides a close interpretation of the remarkably rapid rise of the category of 'transgender' in the United States, fomented through a series of academic engagements and political projects. The burgeoning field of transgender studies has an early precedent in queer anthropology. There are multiple examples of gendering practices in distinct regions of the world, such as fa'afafines in the South Pacific or women's masculinities in Asia, Africa, and Latin America (Wieringa et al., 2007; Blackwood and Wieringa, 1999). Queer ethnographic work on transgendering, has posed a series of theoretical challenges and generative discussions about how sexuality and gender become articulated and in turn, how anthropologists and others can do representative justice to them.

The ways in which gender and sexuality cohabitate, both intellectually and experientially, have been foundational to queer anthropology. In a parallel way, gender performance and the performativity of language have been central to debates in sociolinguistic queer anthropology (Kulick, 2000). Queer

theory drew very dramatic attention to the concept of linguistic and gender performativity. Following the linguist J.L. Austin, Judith Butler argued for example, that illocutionary acts of 'hailing' a sexual subject performs the act it is said to represent. In other words, to call someone 'gay' is not a descriptive exercise but a constitutional/determinative one that, repeated over time, comes to be accepted as a putative truth. How naming acts are performed and understood and how sexual categories are applied, appropriated, or retooled in queer worlds continues to be a dynamic domain for critical queer inquiry. While language use within queer communities had been analyzed in a folkloric mode cataloging lexical items, in *Word's Out: Gay Men's English* William Leap (1996) argued for a more sociodynamic approach that highlighted how gay men negotiated risk, cooperation, and coming out through specific linguistic forms and performances. Using life stories and interviews as well as gay magazines and books, Leap's ethnography foreshadowed later interventions in queer anthropology that would focus on gay pornography as cultural texts. The edited collection, *Speaking in Queer Tongues: Globalization and Gay Language* (Leap and Boellstorff, 2003), further transnationalized the analysis of queer idioms and speech acts, illustrating how global discursive flows and travel have influenced gay and lesbian lexicons and transformed the content of queer language systems.

Transnationalities and New Sexual Subjectivities

While queer anthropology can be traced to historical moments in the anthropological canon, queer anthropology has unmistakably burgeoned in the more recent era of globalization. In what they called a more generalized 'transnational turn' in lesbian, gay, and queer studies, the anthropologist Elizabeth Povinelli and the historian George Chauncey (1999), signaled how neoliberal globalization was being played out within the intimate domains of sexual subjectivity. For queer anthropology and the discipline of anthropology more generally, processes of globalization – including the flow of goods, ideologies, financial instruments, and human beings themselves – were not entirely new terrain. However, with the acceleration of these flows and their interrealtionships with putatively private domains, sexuality was becoming a more explicit, expanded, and vitally important area of queer anthropological research. Earlier works, such as *De Los Otros* (Carrier, 1995), *Latin American Homosexualities* (Murray, 1995), and *Islamic Homosexualities* (Murray and Roscoe, 1997), for example, outlined what can be glossed as 'indigenous' 'traditional' 'autochthonous,' or 'historical' modes of same-sex practices. Using the idiom of sexual models, these projects differentiated gender role-based behaviors (transgenderal), age-distinguished sexual practices (transgenerational), and reciprocal sexual behavior (egalitarian). Emphasis was also placed on how social stigma was apportioned according to one's 'role' in sexual acts, or how sexual shame was distributed. Roger Lancaster (1992) argued, for instance, that gender and sexual values in Nicaragua, historically, had allowed for a degree of low-status social recognition for *cochones* (loosely, 'fags'). A masculine man could, in this pattern of transgenderal sexuality, have sex with another man, as long as he maintained the 'active,' 'insertive,' and 'manly' role in the encounter. Similar sexual status hierarchies were, and continue to be, familiar in Mediterranean and Middle Eastern settings, among others. However, 'role-based' approaches have found significant critique in queer anthropology, and many have argued that sexual desire and erotic practices are far more diverse and malleable than role-based models would suggest. In the age of globalization, this criticism has found even more traction. It is not that sexual models or role-based practices have disappeared entirely, but sexual practices and the social values underlying them are increasingly intertwined with globalized forms of intimacy and epistemologies of identity.

The transnational turn in sexuality studies, and its continued relevance in queer anthropology, has demonstrated how erotiscapes around the world are changing as populations and queer communities come into greater and more regular contact with one another, either virtually or *in vivo*. As with all things regarding globalization, it is impossible to disregard the multiplicity of forces at work, including increases in commerce, trade, and tourism, as well as the emergence of new sexual subjectivities. In *Beneath the Equator* (1999), Richard Parker demonstrated how 'cruising' and sex work coincided in Brazilian urban space to engender gay men's communities. Parker's ethnographic material also raised the issue of HIV/AIDS and the dynamics of transitional health management (and failures), an area of research that has continued to be an important emphasis within queer anthropology. Analyzing the dynamics of sexual commerce and tourism in the Caribbean (2007), Mark Padilla has shown how gender and sexual behaviors have been transformed as nominally 'straight' men engage in sex with foreign (male) travelers while continuing to maintain heterosexual relationships with their female partners. In addition to transforming labor, gender roles, and sexual behavior, Padilla notes how the transnational flow of tourism and sexual commerce has created new challenges for HIV prevention and public health management.

As the insistence of globalization has increased, sexual subjects everywhere are being conditioned by a greater diversity of discourses and possibilities for imagining the sexual self. The dissemination of global media, internet communications, travel, and capital flows have also produced worries about how sexual difference may be subsumed under a monological 'global gay' sexual citizen. Because of its longstanding commitment to investigating how ideals and practices of sexuality travel, queer anthropology is well equipped for these moments of transnational erotic transformations. Even as sexuality – as a concept, a category, a practice – is in motion, this does not necessarily portend the wholesale erasure of non-western (or non-northern) forms of sexual identification or practice. Martin Manalansan's (2003) work in New York City, for example, describes how queer Filipino immigrants have created new modes of sociality in a diasporic context. They have retained certain cultural forms from their homeland, responded to new challenges in their diasporic home, and found ways to mobilize and articulate sexual sensibilities that eschew a strict distinction between (putatively) 'traditional' or 'modern' forms of sexual practice. Indeed, disrupting this developmentalist logic of sexuality, one that begins in 'indigenous' forms of same-sex sexuality and ends in

so-called egalitarian, 'modern,' or global models of sexual subjectivity, has been a central argument in transnational queer anthropology. Following the quotidian experiences of gay Indonesian men in the world's largest Muslim country – and examining how their sexual subjectivity has been informed by mass media, nationalism, and globalization – Tom Boellstorff (2005) has suggested that his interlocutors' experiences reveal 'archipelagic' contrasts and philosophical questions about what constitutes sameness and difference. Engaging with ongoing anthropological debates about nation states, Boellstorff argues that the cohering logic of the Indonesian nation state allowed for the formation of particular kinds of queer subjectivities. In addition to charting how global discourses and transnational diasporas shape and define sexual subjectivities, these sorts of studies challenge assumptions about assimilation more broadly. They demonstrate how immigrant practices and globally disseminated models of identity are being taken up and refashioned in global contexts.

Understanding how queer communities and individuals reconceptualize and rearticulate seemingly hegemonic global gay identities has been a critical emphasis in transnational queer anthropology. Evelyn Blackwood's (2010) work in Indonesia among the Minangkabau, for instance, has shown how same-sex loving women have adopted and adapted local, regional, and transnational forms of queer subjectivity. Megan Sinnott's (2004) fieldwork in Thailand also illustrates how women's same-sex sexuality has been negotiated over time and in particular gendered contexts. These studies, among others, have provided an important counterbalance to the gender disparity that has been the legacy of much of queer anthropology. Note that in essays that reviewed the state of the field of lesbian and gay anthropology in 1993 (Newton, 1993) and again in 2007 (Boellstorff, 2007), authors drew attention to the paucity of anthropological work on women's same-sex sexuality and affective relationships. Unfortunately, this is still true as men's sexuality continues to dominate the field. However, it is also true that women (as lesbians, or by some other moniker) have become a more robust subject in queer ethnographic studies, from social movements to shifting subjectivities (Dave, 2012; Howe, 2013; Wekker, 2006). This is a promising sign suggesting that the subfield's androcentric bias is being recalibrated. Scholarship on women's masculinities, transgender subjectivity, and the articulations between sexuality and feminism should continue to be a generative arena for future queer anthropological work.

Futures and Post-Queering Anthropology

Maybe it is unwise to prognosticate the future of any academic field but we do have precedent and patterns to discern our way forward, and queer anthropology has many vibrant prospects on the horizon. Continuing in the mode of transnational inquiry, the subfield will undoubtedly generate more accounts of queer lives around the world, bringing them into conversation with other elements of the socioscape including changing economic conditions and epistemologies of rights and relationalities between, among, and beyond queer communities. Many queer ethnographic works have been situated in the global north (especially the United States); other areas of the world, though to a lesser degree, have also been represented in queer anthropology (Latin America, the Caribbean, South Asia, Southeast Asia, East Asia, and Melanesia). However, more work is clearly needed in the Middle East and North Africa, as well as East and West Africa. While some queer anthropological work has been carried out in these regions (Gaudio, 2009), several are in the ethnocartographic mode (Murray and Roscoe, 1997) or only focused on queer sexuality in an ancillary way (Borneman, 2007), meaning that a significant gap remains.

The dynamics of communication, particularly in the era of global media, social networks, and online life, will undoubtedly continue to be of interest to queer anthropology. Anthropological work on new media forms has demonstrated how queer subjectivity can be created and contextualized in digital forums (Boellstorff et al., 2003; Gray, 2009; Herdt and Howe, 2007). Queer subjectivities that may be impossible elsewhere have found social space and new possibilities through the use of digital platforms and new media outlets. Future research will continue to complement social scientific analyses of media more broadly, including how media is produced, consumed, interpreted, and engaged. In the spirit of analyzing performance, performativity, and the impact of the Internet, queer anthropological studies of pornography also promise to generate provocative readings and interpretations of how queer bodies are rendered through cinematic tropes and circuits of consumption.

The next generation of studies in queer anthropology will continue to analyze the confluences of globality, (post)-identities, and shifting subjectivities and allegiances to nation, state, race, gender, and sexuality. Our current global episteme is one that circulates massive amounts of knowledge and information implicating political, social, and sexual selves. New queer anthropological work on transnational advocacy and ethnographies of activism (Howe, 2013) uncover how queer political agendas become formed in part through social movements and the ideologies that have preceded them as well as how they are conditioned by neoliberal or neosocialist contexts. These studies should continue to illustrate how queer activism also informs other movements for social justice and equity. Following earlier studies in the anthropology of emotion, the affective turn in the humanities and social sciences will likely continue to inform queer anthropology, demonstrating, for example, how emotive, affective practices condition political formations (Dave, 2012). The meaning of queer personhood and what constitutes equality are intellectual questions that are being answered, in part, through analyses of how human rights models become articulated with autochthonous juridical forms and social values. Sexual asylum claims, queer migration policies, transnational same-sex marriage, and hate crime legislation will continue to be critical areas of engagement going forward (Murray, 2014). In the domain of the anthropology of experts and expertise, queer anthropology will also likely continue to find productive sites of analysis that center on innovations in queer law and scientific treatments of queer biopolitics.

Michael Warner, one of queer theory's founders, has suggested that we are now entering a post-queer phase of scholarship, one that has shifted queer theory's abiding concerns to

new rubrics of inquiry including the deep inequalities of globalization, and a fascination with temporality, affect, and normative repudiations. For queer theory, as for its anthropological kindred, these turns in the queer imaginary are not divested from understanding how sexuality is maintained and transformed, but they are open to new contingencies and mediations. Whether post-queer-ness is upon us or not, we know that queer anthropology has proven itself agile in comprehending new theoretical and ethnographic coordinates. It has been an anthropological 'bent' with a heterogeneous and generous intellectual ontology that bodes well for future innovation.

See also: Discourse and Indentity; Expertise, Anthropological Approaches To; Feminism in Anthropology; Fieldwork in Social and Cultural Anthropology; Gender: Anthropological Aspects; Identity in Anthropology; Kinship in Anthropology; Language and Gender; Performance: Anthropological Aspects; Reflexivity in Anthropology; Sociolinguistics.

Bibliography

Blackwood, Evelyn, 2010. Falling into the Lesbi World: Desire and Difference in Indonesia. University of Hawai'i Press, Honolulu.
Blackwood, Evelyn (Ed.), 1986. The Many Faces of Homosexuality: Anthropological Approaches to Homosexual Behavior. Routledge, New York.
Blackwood, Evelyn, Wieringa, S. (Eds.), 1999. Female Desires: Same-Sex Relations and Transgender Practices across Cultures. Columbia University Press, New York.
Boellstorff, Tom, 2005. The Gay Archipelago: Sexuality and Nation in Indonesia. Princeton University Press, Princeton.
Boellstorff, Tom, 2007. Queer studies in the house of anthropology. Annual Review of Anthropology 36, 17–35.
Boellstorff, Tom, Berry, Chris, Martin, Fran, Yue, Audrey (Eds.), 2003. Mobile Cultures: New Media in Queer Asia. Duke University Press, Durham.
Borneman, John, 2007. Syrian Episodes: Sons, Fathers, and an Anthropologist in Aleppo. Princeton University Press, Princeton.
Carrier, Joseph, 1995. De Los Otros: Intimacy and Homosexuality Among Mexican Men. Columbia University Press, New York.
Dave, Naisargi, 2012. Queer Activism in India: A Story in the Anthropology of Ethics. Duke University Press, Durham.
Douglas, Mary, 1966. Purity and Danger: An Analysis of the Concepts of Pollution and Taboo. Routledge, London and New York.
Elliston, Deborah, 1995. Erotic anthropology: 'ritualized homosexuality' in Melanesia and beyond. American Ethnologist 22 (4), 848–867.
Evans-Pritchard, E.E., 1970. Sexual inversion among the Azande. American Anthropologist 72 (6), 1428–1434.
Gaudio, Rodolfo, 2009. Allah Made Us: Sexual Outlaws in an Islamic African City. Wiley-Blackwell, New York.
Gray, Mary, 2009. Out in the Country: Youth, Media, and Queer Visibility in Rural America. New York University Press, New York.
Herdt, Gilbert H. (Ed.), 1984. Ritualized Homosexuality in Melanesia. University of California Press, Berkeley.
Herdt, Gilbert, Howe, Cymene (Eds.), 2007. 21st Century Sexualities: Contemporary Issues in Health, Education and Rights. Routledge, New York.
Howe, Cymene, 2013. Intimate Activism: Sexual Rights in Postrevolutionary Nicaragua. Duke University Press, Durham.
Jacobs, Sue-Ellen, Thomas, Wesley, Lang, Sabine (Eds.), 1997. Two-Spirit People: Native American Gender Identity, Sexuality, and Spirituality. University of Illinois Press, Urbana.
Kulick, Don, 1998. Travesti: Sex, Gender, and Culture Among Brazilian Transgendered Prostitutes. University of Chicago Press, Chicago.
Kulick, Don, 2000. Gay and lesbian language. Annual Review of Anthropology 29, 243–285.
Lancaster, Roger N., 1992. Life is Hard: Machismo, Danger, and the Intimacy of Power in Nicaragua. University of California Press, Berkeley.
Leap, William, 1996. Word's Out: Gay Men's English. University of Minnesota Press, Minneapolis.
Leap, William, Boellstorff, Tom (Eds.), 2003. Speaking in Queer Tongues: Globalization and Gay Language. University of Illinois Press, Urbana.
Lewin, Ellen, 1993. Lesbian Mothers: Accounts of Gender in America. Cornell University Press, Ithaca.
Lewin, Ellen, 1998. Recognizing Ourselves: Ceremonies of Lesbian and Gay Commitment. Columbia University Press, New York.
Lewin, Ellen, 2009. Gay Fatherhood: Narratives of Family and Citizenship in America. University of Chicago Press, Chicago.
Lewin, Ellen, Leap, William (Eds.), 1996. Out in Theory: the Emergence of Lesbian and Gay Anthropology. University of Illinois Press, Urbana.
Manalansan IV, Martin F., 2003. Global Divas: Filipino Men in the Diaspora. Duke University Press, Durham.
Morgenson, Scott L., 2011. Spaces between Us: Queer Settler Colonialism and Indigenous Decolonization. University of Minnesota Press, Minneapolis.
Morris, Rosalind, 1995. All Made up: performance theory and the new anthropology of sex and gender. Annual Review of Anthropology 24, 567–592.
Murray, Stephen O., 1995. Latin American Male Homosexualities. University of New Mexico Press, Albuquerque.
Murray, Stephen O., Roscoe, Will (Eds.), 1997. Islamic Homosexualities: Culture, History, and Literature. New York University Press, New York.
Murray, David (Ed.), January/February 2014. Queering Borders: Language, Sexuality and Migration. Journal of Language and Sexuality 3 (1).
Nanda, Serena, 1998. Neither Man Nor Woman: The Hijras of India. Cengage Learning, New York.
Newton, Ester, 1979. Mother Camp: Female Impersonators in America. University of Chicago Press, Chicago.
Newton, Ester, 1993. Cherry Grove, Fire Island: Sixty Years in America's First Gay and Lesbian Town. Beacon, Boston.
Padilla, Mark, 2007. Caribbean Pleasure Industry: Tourism, Sexuality, and AIDS in the Dominican Republic. University of Chicago Press, Chicago.
Parker, Richard, 1999. Beneath the Equator: Cultures of Desire, Male Homosexuality, and Emerging Gay Communities in Brazil. Routledge, New York.
Povinelli, Elizabeth A., 2006. The Empire of Love: Toward a Theory of Intimacy, Genealogy, and Carnality. Duke University Press, Durham.
Povinelli, Elizabeth A., Chauncey, George, 1999. Thinking sexuality transnationally: an introduction. GLQ: A Journal of Lesbian and Gay Studies 5 (4), 439–450.
Reddy, Gayatri, 2005. With Respect to Sex: Negotiating Hijra Identity in South India. University of Chicago Press, Chicago.
Rubin, Gayle, 2012. Deviations: A Gayle Rubin Reader. Duke University Press, Durham.
Sinnott, Megan, 2004. Toms and Dees: Transgender Identity and Female Same-Sex Relationships in Thailand. University of Hawai'i Press, Honolulu.
Valentine, David, 2007. Imagining Transgender: An Ethnography of a Category. Duke University Press, Durham.
Vance, Carole (Ed.), 1984. Pleasure and Danger: Exploring Female Sexuality. Routledge, New York.
Wekker, Gloria, 2006. The Politics of Passion: Women's Sexual Culture in the Afro-Surinamese Diaspora. Columbia University Press, New York.
Weiss, Margot, 2011. Techniques of Pleasure: BDSM and the Circuits of Sexuality. Duke University Press, Durham.
Weston, Kath, 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.
Weston, Kath, 1993. Lesbian/Gay Studies in the House of Anthropology. Annual Review of Anthropology 22, 339–367.
Wieringa, Saskia, Blackwood, Evelyn, Bhaiya, Abha (Eds.), 2007. Women's Sexualities and Masculinities in a Globalizing Asia. Palgrave Macmillan, New York.

Queer Geography

Andrew Gorman-Murray and Scott McKinnon, University of Western Sydney, Penrith, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Queer geographies overlap with the geographies of sexualities, but the two fields are not entirely interchangeable. Queer geographies offer an approach that combines queer theory with spatial analytics to describe and analyze the social coconstitution of sexuality and space. This article discusses the contributions of queer geographies to scholarship, focusing on three geographical lenses: scale, mobility, and place-making. The article first charts inter-relations of sexuality and space at the scales of the body, home, city, and nation-state, and then traces queer processes of emplacement and displacement ranging across the local, national, global, urban, and rural.

Introduction

The disciplinary label 'Queer Geography' maps out a wide field, the boundaries of which are somewhat indistinct and remain in dispute. In this article, we allow for these uncertain borders by offering, not so much a definitive cartography, but a journey of exploration. We outline here the emergence of sexuality as a concern in the discipline of geography, including the expanding influence of queer theory; suggest ways in which sexuality and space both constitute and are constituted by one another at a range of scales; note an increasing disciplinary interest in mobilities within, across, and through these scales; and explore the concept of place-making as it intersects with mobilities and queer sexual identities. We allow 'queer' two – admittedly contradictory – meanings here. Firstly, queer operates as a catch-all term for a range of 'alternative' sexual and gender minorities, including those who identify as lesbian, gay, bisexual, queer, and/or transsexual/transgender, as well as those who, while homosexually active, do not adopt an identity based on their sexual desires. Secondly, queer can be defined as a process which highlights fluidity and complexity and which specifically disrupts sexual and gender identity labels. By incorporating these two definitions, we are able to explore 'queer geography' as, firstly, an examination of intersections between sexuality and space as studied within the field of geographies of sexualities and, secondly, as its own distinct field which initially emerged from geographies of sexualities but which has since expanded to encompass a range of other concerns (Knopp, 2007).

It was in the 1970s and 1980s that homosexual identities emerged as a disciplinary concern for some geographers (Bell and Valentine, 1995). The primary interests of this early research were in mapping the location and development of the social, commercial, and residential urban spaces of gay men and, to a lesser extent, lesbians. Gradually, feminist geographers and Marxist economic geographers began to push lesbian and gay geography away from its initial positivist epistemologies and toward more radical explorations of power relations. Larry Knopp, for example, explored intersections between sexuality and the spatial dynamics of capitalism (Knopp, 1992). Into the 1990s, the influence of queer theory – drawing particularly on the work of Judith Butler – located performance and representation as critical factors in studies of sexuality and space. The 1995 publication of the edited volume *Mapping Desire* (Bell and Valentine, 1995) marked a significant moment for poststructuralist geographies of sexualities that sought to queer understandings of how sex and sexuality operate in and through space.

What such research achieved was a new visibility for the role of sexuality in how space is constituted, imagined, and experienced, and just as importantly, for the role of space in the development and construction of sexual identities and behaviors. Concentrating largely (but not exclusively) on the experiences of queer populations, the research considered how heterosexuality is often spatially constituted as normative, thus rendering nonheterosexual people out of place or as needing to queer space to both contest heterosexual norms and to create a sense of belonging. Space and sex were seen as intertwined and constitutive at a range of scales including the body, the urban and the nation; as intersecting with race, gender, and class; and as critical to a range of geographical disciplinary areas, most notably social and cultural geographies, but also economic, political, and health geographies.

Even though the field has been heavily influenced by queer theory since the 1990s, research within geographies of sexualities has more recently been critiqued by some queer geographers (or at least by geographers using queer theory) as overly concerned with identity politics and with the twin binaries of male/female and heterosexual/homosexual. Queer theory itself largely defies definition, seeking to maintain its radical potential by refusing to adhere to any fixed meaning. At its core, however, queer theory rejects sexual and gender binaries in favor of fluidity and complexity, finding sexual identity and its role in space not as fixed and constituted but iterative and mutable. To queer theorists, 'queer' is not an identity but a process; not a noun but a verb. In this light, queer geography can at times be seen to include geographies of sexualities, but the two should not be understood as interchangeable labels for the same field (Knopp, 2007). Queer geography, thus defined, seeks not to examine homosexual (or queer) space as the nonnormative 'other' to the normative of heterosexual space, but instead urges a queer approach to space which rejects any such binary (Oswin, 2008). Unlocked from a binary view it is possible to see, for example, how materialized performances of

homosexuality may in fact enforce normativity, and/or to locate in space nonnormative performances of heterosexuality.

Scale

Both geographies of sexualities and queer geographies frequently examine the spatial operation or constitution of sexuality using scale as a conceptual and methodological framework. For those not familiar with the concept, scale is a quintessential focus of geographical thinking that refers to the size and scope of the 'space' under examination, as well as the relationships between scales (e.g., hierarchical, networked, etc.). In disciplinary debate, the notion of scale is increasingly complex. If once presented as a fixed set of analytical categories, often arranged in a fairly rigid hierarchy from smaller to larger (e.g., body, urban, regional, national, global), these geographical scales have themselves come to be understood as socially constructed, negotiated, and reproduced. Sexuality often plays an important role in this construction across various scales from, for example, the body to the home, the community, the city, the rural, the nation, and the globe. Below, we take four of these scales as examples of how geographers have investigated intersections between sex, sexuality, and space.

The Body

According to Johnston and Longhurst, the body is the "primary space in which and around which we construct our subjectivities" (Johnston and Longhurst, 2010: p. 39). Our bodies operate both *as* space and *in* space and determine ways in which we are included or excluded; experience regulatory approval or disapproval; are materialized as sexualized and gendered beings; and physically or metaphorically move within and through space. The body is a site in and at which we construct our sexual identities and through which we perform those identities. It is also a heavily monitored space which is subject to the same socially constructed discursive regulation as wider scales, and thus has become a significant site of interest for geographers seeking to understand the spatial aspects of sex, gender, and sexuality.

One indicative element of sexuality at this scale is the metaphor of the closet (Brown, 2000). Homosexual people are frequently described as being *in* or *out* of the closet according to whether they are attempting to hide or determined to reveal their (homo)sexual identity. This is an explicitly spatial metaphor in which the body is materialized as a site of identity performance. Along with this spatial metaphor, Michael Brown has shown how gay men at times describe the closet as a 'real' physical or material space and as a "way their physical geographies are demarcated" (Brown, 2000: p. 48). Migration from a town, city, or nation considered homophobic (and in which one's sexuality must be kept private) to a location in which homosexuality is more widely accepted (and in which a gay identity can be publicly performed) can therefore be seen as both a migration across wider scales and also as a move from *in* to *out* of the closet at the scale of the body (Gorman-Murray, 2007).

Geographers have also examined how the sexual and gendered body is operated upon by regulatory forces. This research draws on the work of Foucault, who has argued for the powerful, historically, and geographically constituted role of discourse in shaping and controlling sexuality from the individual body to the nation. Displays of physical affection are frequently cited as examples of how, through social regulation, the heterosexual body and the homosexual body experience space differently. While a public kiss between a heterosexual couple may go unnoticed or uncommented on, a kiss between a homosexual couple may well elicit comment, abuse, or even violence. Such regulation delineates sexual identity performance according to a public/private binary, denying or troubling access to, or safety in public space for nonheterosexual people (Johnston and Longhurst, 2010).

Similarly, the gendered body is regulated according to a male/female binary, leaving intersex, trans, or ambiguous bodies out of place and subject to the possibility of censure or attack (Doan, 2007). The normative body is delineated as clearly and publicly either male or female. By conforming to this binary through the space of the body – for example, by wearing clothing deemed 'appropriate' to our gender and comporting ourselves according to masculine or feminine expectations – we may ease the process of our participation in public space. Bodies which refuse or are unable to conform in this way may well find their access to space limited or uncertain, and potentially dangerous.

The Home

The public/private binary can also be queried, queered, and deconstructed at the scale of 'the home,' which is often synonymous with 'private space' or the 'private sphere.' The space of the home is often thought of as a refuge or sanctuary from the public sphere. We go home to relax and be ourselves. In terms of sexuality, this may mean that the home is a safe space for the development and performance of sexual or gender identities, away from the prying eyes of regulatory forces. The research of a number of geographers has found, however, that the home is a complex space imbued with a variety of (at times conflicting) meanings and in which intersecting categories of sexuality, gender, race, ethnicity, age, and class continue to hold influence.

The public/private binary, which would at first appear to exclude the outside world from the sanctity of the home, has been found to be unstable and unclear (Gorman-Murray, 2012). The monitoring and regulation that structures performances of sexuality and gender in public may equally find its way into the home through several possible avenues (while some ostensibly public spaces can also become 'private' at certain times or through particular uses – e.g., public toilet stalls).

For instance, home is often thought of as a space of family, generally imagined and structured as a heterosexual, monogamous couple raising children (Johnston and Longhurst, 2010). Geographers of sexualities have complicated and expanded this imagining by investigating the ways in which nonheterosexual individuals, couples, and families make or experience home. How, for example, do the experiences of lesbian or gay family members living with heterosexual parents complicate both the notion of home as a place for heterosexual families and/or as a place of safe gay or lesbian identity performance? (Valentine et al., 2003). For queer teenagers or young adults, the home

may in fact be a form of closet, in which they must hide or mask their sexual identities for fear of negative repercussions from heterosexual family. Under such circumstances, public spaces such as bars or nightclubs may at times offer a safer venue for identity development and performance than the home, and indeed, can become home*like* (Gorman-Murray, 2006), revealing that public/private permeability flows both ways.

Andrew Gorman-Murray has found that home can also act as a space for a queer politics of identity (Gorman-Murray, 2012). In configuring home as a space from which a homophobic world is excluded, gay men may carefully monitor public intrusions and attempt to enforce a division between the public and the private. Equally, however, some gay men were found by Gorman-Murray to exploit the ambiguity of the public/private binary in order to enact a political identity. The flying of a rainbow flag (a symbol of the gay pride movement), for example, is a public performance of sexual identity enacted in or at the scale of the home. This is a specifically political act, defying attempts to exclude gay men from the public sphere by placing the sexual make-up of a private home on public view.

Further complicating the idea of home as a space of safety and security is the fact that, particularly for women and children, home may be a site of abuse and danger (Johnston and Longhurst, 2010). While public space is often presented as the dangerous sexual other from which home can provide sanctuary, such an imagining elides experiences of family violence and sexual abuse in the home. By making such experiences visible, queer geographies reveal how fixed notions of sexualized space, as well as set binaries of public/private or safe/unsafe, mask the ways in which spaces like the home remain iterative, unstable, and uncertain.

The City

In terms of geographies of sexualities, 'the city' is perhaps the scale at which the most substantial amount of work has been conducted over the longest time period (Brown and Knopp, 2003). As noted above, it was the urban spaces adopted by visible gay and lesbian communities that first prompted a geographical interest in sexuality in the 1970s and 1980s. Since that time, the field has expanded to investigate the role of urban space in the development of a more diverse range of sexual and gender identities (including heterosexuality), as well as to consider intersections between sexuality and race, age, ethnicity, ability, gender, and class.

Across that time period, gay men and lesbians in many of the large cities of the Western world have experienced radical changes in public acceptance, legislative reform, and access to urban space. While these changes arguably reflect a growing level of social equality for nonheterosexual individuals and a growing ability to create or produce queer urban spaces, queer geographers have examined ways in which sex and gender continue to be regulated and in which the production of queer urban space may in fact produce inequalities and enforce normativity.

Petra Doan, for example, has examined the gendered nature of queer urban spaces, highlighting a potential lack of tolerance for transgender or gender variant people (Doan, 2007). While noting that transgendered individuals do report feeling safer in visibly queer areas than in cities without a queer space, Doan also finds that lesbian and gay residential or commercial areas "replicate a strongly heteronormative gender dichotomy" (Doan, 2007: p. 70). Thus, sexuality can in fact intersect with gender to produce normativities that would at first seem out of place in a queer space.

Intersections between sexuality, race, and class may similarly trouble the place of queer individuals in some queer urban spaces. Berlant and Warner, for example, note attempts by local property owners to exclude 'queer pilgrims' – frequently young, African-American people – from making use of gay neighborhoods in New York City (Berlant and Warner, 1998: p. 563). To Berlant and Warner, access to such spaces should not be predicated on the ability to buy into an expensive property market. Visitors play an important role in the production of queer urban space while intersections of class, race, and sexuality may produce normativities which exclude those visitors.

The Nation-State

Geographers have also examined complex constitutive interactions between sexuality and the nation-state. Perhaps the clearest example of how the nation constitutes the experience of sexuality for its queer citizens is the legal regulation of homosexual sex acts (Binnie, 2004). While, for example, sex between consenting male adults was decriminalized in Great Britain in 1967, nation-states including Indonesia, Iran, and Malaysia maintain strict bans (ILGA, 2013). The ability to perform (both in public and private) and to develop a (homo)sexual identity is thus heavily constituted by the state. Even in those nations which have decriminalized homosexual acts, the state continues to play a role in deciding which kinds of sexual activity are 'moral' or acceptable (e.g., monogamous relationships) and which are not (e.g., public sex, sadomasochistic (S&M) sex). Sexual subjectivities are thereby specifically constituted at the scale of the nation-state, while sexual identities and behaviors demarcate and limit the boundaries of citizenship.

It is also true, however, that sexuality plays a significant role in how a nation will imagine itself into being. In part, this may be enacted through attempts to exclude nonheterosexual identities from the national imagination. Recent legislation in Russia, for example, has banned the holding of gay pride parades and the distribution of lesbian, gay, bisexual, and transsexual or transgendered (LGBT) information to children (Associated Press, 2013). Gordon Waitt has shown that it is possible for sexuality to intersect with and disrupt national mythologies in a way that (even if only temporarily) queers those mythologies (Waitt, 2005). Waitt explores the centrality of sport to Australian nationalist imaginary and notes that the Gay Games sporting festival in Sydney in 2002 provided, for some participants, a sense of queering the presumptive heterosexuality of the Australian sportsperson's body.

In many ways, the aims of lesbian and gay activism can be seen as efforts to gain inclusion for nonheterosexual people as full national citizens. For many queer individuals, there remains a significant tension, however, between the desire for inclusion within the nation-state (and acquisition of the benefits of full citizenship) and a feeling that such inclusion is predicated on participation in a consumerist system which demands conformity and elides difference. As Waitt notes,

some in Sydney felt that participation in the Gay Games signified less a queering of the sporting body as it did a conformist acceptance of nationalist normativities. The predominance of military service and same-sex marriage as aims for lesbian and gay activism in the United States (and other countries) also signify this tension (Binnie, 2004). While some argue that participation in the military and access to marriage signify a queering of these institutions and of the nation-state by which they are regulated, others see instead the dequeering of communities now conforming to the dominant (hetero)normative values of the state.

These examples of scale are by no means conclusive, but are instead offered as examples of how geographers (and others concerned with sex, sexuality, and space) have theorized and examined the spatial aspects of sexuality. If geographers have taken a specific interest in connections between sex, sexuality, and space at various scales, it is also true that an increasing focus has been applied to movement within, between, and through scales, and the relations between movement and place-making. Below, we explore some examples of research into how sexuality impacts on entwined processes of mobility and place-making across scale and space.

Mobility and Place-Making

Mobilities

Sex and sexuality can be seen as critically important inspirations for or limitations to mobility and migration. Queer geographies can offer valuable insights into mobility, for example, by investigating the role of sexuality in encouraging migration or in constituting the experience of movement across and within space. As argued by Nicola Mai and Russell King, "love, and sometimes alongside it, sex can play a decisive role in the imagination and enactment of the choice to migrate" (Mai and King, 2009: p. 296). Mobility itself can be enacted at and through different scales, with various terms applied to these movements. At a local, everyday level, we might think of such mobilities as commuting (to and from work) or 'cruising' (a form of everyday mobility perhaps quite resonant with gay male subcultures). At broader scales, both temporal and geographical, we can refer to ideas of migration, which might be intraregional, intracity, intranational, or international. We consider some of these scales of mobility in relation to sexuality and space below.

Taking the broadest scale and range of mobility, in terms of international migration, Mai and King note a "sexual turn" (Mai and King, 2009: p. 296) in migration studies which, heavily influenced by queer theory, has placed migrants (including migrant workers, refugees, and others) as sexual beings. Such a framing allows for emotional and sexual desires to be seen as central issues in why people migrate or how they experience the spaces from, to, and through which they move in the process of migration. This research has included, for example, study into the impacts of migration on the 'traditional' or established sexual and gender norms understood or practiced by migrant groups. Nader Ahmadi, for example, argues that among Muslim Iranian migrants in Sweden, "traditional authoritarian patriarchal sexual relationships are giving way to more egalitarian relationships" (Ahmadi, 2003: p. 702).

Mobilities and sexuality also significantly intersect in what Larry Knopp has called "queer quests for identity" (Knopp, 2004), which are often perceived and experienced at the intranational scale, although at various levels of displacement within the nation-state. These are the journeys taken, at a range of scales (within cities, between cities, between regions, etc.), by queer people seeking a space in which to develop and perform nonnormative sexual or gender identities. This may include, for example, gay men and lesbians who travel from their homophobic home-of-origin (be it a rural area, city, or nation) to a more accepting location in search of a safer space in which to come out as gay or lesbian. At a simplistic level, such a journey may mark out an almost cinematic narrative, in which our hero(ine) escapes from the dark closet of the rural, travels through a series of adventures and is left, in the last scene, finally free and forever out in the city. Away from the cinema screen, mobility for the queer pilgrim is more complex and may involve a series of journeys, returns, and an ongoing, iterative process of making and unmaking both place and identity (Gorman-Murray, 2007).

It is important to note that mobility does not only signify migration across wider scales such as rural to urban, city to city, or nation to nation. If we accept that sexuality is spatially constituted, then it is possible to see how queer people may adapt, (re)make, reveal, and hide their sexual identities in different ways when moving across smaller scales. A businessman who is out as gay at home but closeted when traveling on business, or a married man who only explores his bisexual identity online or in specific material spaces, suggest examples of how hybrid sexual subjectivities may be materialized differently as queer people move through space (Knopp and Brown, 2003). 'Cruising' is another classic example, in which gay men, bisexual men, or men-who-have-sex-with-men seek connections and/or sexual encounters with like others as they move through public urban spaces day-to-day, including streets, parks, train stations, and pubs. Cruising palpably links same-sex activities and subjectivities – both deliberate and furtive – to everyday urban mobilities such as walking, commuting, and driving. Queer mobilities – and the queerness of mobility – thus encompass banal or everyday processes as much as national or global displacement.

Place-Making

At first glance, the concept of place-making may seem to sit in contrast to – or perhaps be seen as the conclusion of – mobility. At the end of their travels, for example, migrants may appear to come to a rest and make a place of their own. Geographers have, however, examined place-making as intertwined and in constant interaction with mobility (Knopp, 2004; Gorman-Murray, 2007). As we have argued above, mobility indicates everyday movements as much as it does global migrations. Place-making is a process throughout our lives and, for that matter, throughout each day, as we move at the scale of the body, the nation, and the globe. Geographers of sexualities have examined the significance of sexual identity in the need and the desire both to move through space and to make place. Making a place for sex, for community, or for the construction and performance of an identity takes place at the scale of the body, the home, the neighborhood, the city, and beyond.

Below, we look at some examples of how place-making is performed by queer populations.

The gay neighborhood (or gay territory, gay ghetto, or 'gayborhood') is a frequently cited example of queer place-making, in which nonheterosexual urban populations have established – through commercial enterprises, community organizations, and residential spaces – a visible presence in a particular suburb or area of a city (Brown, in press). If once seen in negative terms as a kind of containment of homosexual populations as enforced by the mainstream, these spaces would eventually be understood in terms of a politics of identity, in which gay populations appropriated urban space as a means of gaining economic and political power as well as social protection. The place-making process here is in some ways depicted as an end-point of queer migrations, in which gay men are best able to be gay (or to adopt a particular form of gay identity) once they have made a place for themselves in the gay neighborhood of a large city.

Geographies of sexualities have often depicted this process of place-making as one in which a gay community reacts against the heteronormativity of urban spaces by queering a space of its own. This is never a fixed and final ownership of space, however, but is an ongoing and contested process which – while offering queer populations a level of safety and a feeling of belonging – may also enhance the dangers of being out and visible as queer. In some instances, therefore, gay men may prefer to make place in a way that masks or hides their sexual identities (Brown, 2000). This form of place-making, rather than queering space, may actually contribute to spatializing and reinforcing heterosexual normativities.

Although place-making in the gayborhood has often framed homosexual identities in terms of the urban, geographers have taken an increasing interest in how queer populations perform processes of place-making in rural areas (Gorman-Murray et al., 2013). Whether as transient moments or long-term migrations, investigations into queer rural place-making have the potential to both highlight the role of sexuality in the constitution of rural spaces and to examine how sexuality is constituted differently between the rural and the urban, while also drawing connections between these material and imaginative spaces.

In terms of transient or temporary rural place-making, geographers have, for example, looked at the meanings and impacts of annual womyn's (Browne, 2011) and gay, lesbian, bisexual, transsexual or transgendered, and queer (GLBTQ) (Gorman-Murray, 2009) festivals held in rural areas of the United States and Australia, respectively. Kath Browne notes ways in which womyn festival–goers seek out or attempt to create a rural utopia which specifically excludes men (and, more controversially, trans women). For the time of the festival, this creates a space for sexual exploration and frames the rural in terms of an escape from the heteronormative regulation of the city. Gorman-Murray contrasts the different meanings of a rural GLBTQ festival for rural residents and urban visitors, with the first viewing the event as a political assertion of their visibility and presence in rural spaces and the latter seeking a rural idyll which places rurality as an "aesthetic commodity for urban consumption" (Gorman-Murray, 2009: p. 81).

It is equally possible to contrast the experiences of lesbian and gay rural residents who were born in a rural area with those who have decided to migrate from the urban to the rural. Geographers have successfully disrupted fixed narratives which frame lesbian and gay migrations as escape from the oppressive rural and toward urban freedom by examining the place-making processes of this latter group. Lesbian separatist communities, for example, as well as gay male groups such as the radical faeries, have undertaken a form of rural place-making in which they seek freedom *from* the city, not freedom *in* the city. Rather than finding an urban space in which to come out as lesbian or gay, some instead seek out an escape from heteronormative (or homonormative) urban experiences and create a rural space for themselves in which to construct and perform a differently constituted form of homosexual identity.

Conclusion

The bourgeoning range of scholarship within queer geography demonstrates that the diverse facets of sex and sexuality – comprising social norms, institutions, and regulations, alongside bodily desires and pleasures – cannot be understood without an interrogation of the spaces, places, scales, and mobilities through which sexual lives, identities, behaviors, and relationships are constituted and practiced. Viewing the world through a queer geographical lens, it is possible to challenge the idea that the spaces in which we live and the sexual and gender identities we adopt are natural and fixed. Instead, we can see that socially constructed ideas of space and of sexuality (as well as gender, race, and class) are constantly in dialog and forever in a process of creation, regulation, and melting away. Sexuality constitutes and is constituted by space such that our experiences of, participation in, or movement through one is intrinsically tied to the other.

See also: Gender and Feminist Studies in Geography; Heterosexism and Homophobia; Lesbian and Gay Activism in Latin America; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations; Place in Geography; Scale in Geography; Sexual Minorities and Sexual Citizenship; Sexual Minorities in China; Sexual Minorities in Modern Russia; Sexual Minorities in South Africa; Sexual Minorities in the Maghreb; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Sexuality and Space; Social Geography; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Ahmadi, N., 2003. Migration challenges views on sexuality. Ethnic and Racial Studies 26 (4), 684–706.

Associated Press, 1 July 2013. Russia passes anti-gay law. The Guardian. http://www.guardian.co.uk/world/2013/jun/30/russia-passes-anti-gay-law.

Bell, D., Valentine, G. (Eds.), 1995. Mapping Desire: Geographies of Sexualities. Routledge, London and New York.

Berlant, L., Warner, M., 1998. Sex in public. Critical Inquiry 24 (2), 547–566.

Binnie, J., 2004. The Globalization of Sexuality. Sage Publications, London.
Brown, M., 2000. Closet Space: Geographies of Metaphor from the Body to the Globe. Routledge, London and New York.
Brown, M. Gender and sexuality II: there goes the gayborhood? Progress in Human Geography. http://dx.doi.org/10.1177/0309132513484215, in press
Brown, M., Knopp, L., 2003. Queer cultural geographies – we're here! We're queer! We're over there, too! In: Anderson, K., Domosh, M., Pile, S., Thrift, N. (Eds.), Handbook of Cultural Geography. Sage Publications, London, pp. 313–324.
Browne, K., 2011. Beyond rural idylls: imperfect lesbian utopias at Michigan Womyn's music festival. Journal of Rural Studies 27 (1), 13–23.
Doan, P.L., 2007. Queers in the American city: transgendered perceptions of urban space. Gender, Place and Culture 14 (1), 57–74.
Gorman-Murray, A., 2006. Homeboys: uses of home by gay Australian men. Social and Cultural Geography 7 (1), 53–69.
Gorman-Murray, A., 2007. Rethinking queer migration through the body. Social and Cultural Geography 8 (1), 105–121.
Gorman-Murray, A., 2009. What's the meaning of chillout? Rural/urban difference and the cultural significance of Australia's largest rural GLBTQ festival. Rural Society 19 (1), 71–86.
Gorman-Murray, A., 2012. Queer politics at home: gay men's management of the public/private boundary. New Zealand Geographer 68 (2), 111–120.
Gorman-Murray, A., Pini, B., Bryant, L. (Eds.), 2013. Sexuality, Rurality and Geography. Lexington Books, Lanham, MD.
ILGA, 2013. International Lesbian, Gay, Bisexual, Trans and Intersex Association. Website. Retrieved 25 June 2013.
Johnston, L., Longhurst, R., 2010. Space, Place and Sex: Geographies of Sexualities. Rowman & Littlefield Publishers, Plymouth, UK.
Knopp, L., 1992. Sexuality and the spatial dynamics of capitalism. Environment and Planning D: Society and Space 10 (6), 651–669.
Knopp, L., 2004. Ontologies of place, placelessness, and movement: queer quests for identity and their impacts on contemporary geographic thought. Gender, Place and Culture 11 (1), 121–134.
Knopp, L., 2007. From lesbian and gay to queer geographies: pasts, prospects and possibilities. In: Browne, K., Lim, J., Brown, G. (Eds.), Geographies of Sexualities: Theory Practices and Politics. Ashgate Publishing Group, Brookfield, VT, pp. 21–27.
Knopp, L., Brown, M., 2003. Queer diffusions. Environment and Planning D: Society and Space 21 (4), 409–424.
Mai, N., King, R., 2009. Love, sexuality and migration: mapping the issue(s). Mobilities 4 (3), 295–307.
Oswin, N., 2008. Critical geographies and the uses of sexuality: deconstructing queer space. Progress in Human Geography 32 (1), 89–103.
Valentine, G., Skelton, T., Butler, R., 2003. Coming out and outcomes: negotiating lesbian and gay identities with, and in, the family. Environment and Planning D: Society and Space 21 (4), 479–499.
Waitt, G., 2005. The Sydney 2002 gay games and querying Australian national space. Environment and Planning D: Society and Space 23 (3), 435–452.

Queer Theory and Intersectionality

Notisha Massaquoi, Women's Health in Women's Hands Community Health Centre, Faculty of Social Work, Ryerson University, Toronto, ON, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article uses theoretical insights to explore the future of queer theory as it engages with intersectionality in a manner that will lead to the development of a framework that articulates the complexities of the lives produced by intersections of gender, race, culture, identity, and sexuality. This direction involves exploring the nonnormative alignments of sexuality, resisting the tendency to essentialize identity or conflate it with mainstream definitions of gender and sexual expressions.

Introduction

Contemporary progressive political movements have often faced the problem of proving legitimacy and authority due to lack of representation or diversity. Activists have long taken issue with the politics of representation and the problem of who speaks for and about others. Much of the experience-based literature articulates a clear divide between a sexuality-based lesbian, gay, transgender, or queer cultural identity and resulting theories and ones that are based on race or ethnicity with the two paradigms rarely intersecting to form a comprehensive legitimate discourse. Intersectionality alone according to Butler and Scott (1992) does not produce good theory. Such a theoretical practice would not end confusion or ambiguity but would allow for attention to and critical analysis of inequalities. Intersectionality points to the fact that the work of intersectionality is more complex and contrary to anything that is theoretically imagined.

Academic interest in how intersectional discourse shapes racialized, gendered, sexualized, and oppositional subjectivities is often tied to academic attempts to shape intellectual commitment, which I would say is very much the case with my following exploration of queer theory and intersectionality. In the following discussion, I am using Shoats' (2002) line of analysis in terms of looking at relational maps of knowledge that highlight the negotiation of sexuality as it is presented and understood in diverse situations and context. I will also explore the future imaginings of a queer theory that engage intersectionality in a manner, which will lead to the development of a framework that enables one to articulate the complexities of the lives produced by interactions of gender, race, culture, identity, and sexuality.

Development of the Concepts

With the increased emergence of queer subjects within mainstream discourse and the elevated demands for political voice and representation has come a stronger strategic resistance to the mainstream tendency of cultural homogenization. Queer subjects or those whose sexual identity, orientation, or behavior is seen as oppositional to the heterosexual norm are counteracting dominant discourses and requiring from social work practitioners a significant change in methodological, political, and culturally based theoretical approaches. This shift would allow for the articulation of identity and the claiming of space in history, while creating a political platform in terms of rather than in spite of the dissention that marks queer subjects as different from the mainstream. Histories of queer cultures and communities are scarce, not only in a Western context, but also in a global one (Manalansan, 2003). Very little is known about same-sex practices or how same-sex desire is theorized. Even more unsettling is the limited attention paid to same-sex desire, sociopolitical location, and the conveniently constructed silences around the relations between and around sexual orientation, gender, race, identity, and subjectivity (Massad, 2007).

To fully understand and support a queer theoretical framework, readers need to move beyond the gaze of queer subjects as the history of a marginalized minority population and look at the organization of social locations, the resulting intersections, and the impact on queer subjects in the widest possible sense. This group ought to be considered as part of and as central to the historical analysis, which validates the Foucauldian concept of the production of sexualities (Foucault, 1990).

The current academic study of sexual orientation that has been connected to the emergence of queer politics of identity and queer theoretical formations has historically focused on Western examples, with primary emphasis placed on white, middle-class gay life (Berube, 2001). A review of the writings of racialized lesbians, gay men, and queers does not produce a critical mass or a coherent body of published works or authors. This is due in part to historically limited institutional access but more so to differing or competing priorities as well as divergent definitions and ways of addressing sexuality and queerness. Queer communities globally may very well express solidarity with each other, but in the day-to-day working out of collective liberation struggles what must be acknowledged is a twofold reality: a clear and active recognition of the extent and nature of the varying contexts and a rich and deeply reflective understanding of the contrasting priorities that these contexts create. Being a queerly marked subject entails the negotiation of identity, politics, and desire as one moves between cultures, regions, and nations. It involves answering to a plethora of terms depending on language and location. It further involves acting on emotions, feeling desire, and engaging in the sexual practices that have diverse meanings and

priorities depending on locality and situation. While other marginalized groups have been defined depending on location and historical and situational context, the current discussion strategically focuses on queer identity.

In the following discussion, I highlight that the failure to use an intersectional approach in the discussion of queer theory can lead to the dislocation of particular marginal identities and the dismantling of progressive coalition building (Coloma, 2006). A queer theoretical view that centralizes intersectionality, I believe, will inform and signify the dialogical relationship between categories of social location, historically specific categories of sexualized identities, and the assumptions of race and location in the lives of queer subjects.

The Unpacking of Queer Identity

According to Epstein (1994), sexual meanings are not universal absolutes but ambiguous and problematic categories, with this becoming more evident as sexuality and sexual meanings are explored globally. For example, I would argue that most forms of oppression are based on what a person is or what they are perceived to be, irrespective of what they do or how they live. Individuals are discriminated based on who they are, for example, being a woman, a Muslim, or a racialized subject. This oppression occurs not because of what that person does but because the individual does it in a marginalized body. In the case of queerness, I believe that it is less a matter of who or what they are that is considered offensive but rather what that person is perceived of sexually doing or not doing.

Queer subjects and their nonconforming sexual behavior have been historically persecuted by state-sanctioned legislation and resulting discrimination. For example, Canada's unreformed same-sex laws, like those of most former European colonies, were modeled after European sodomy statutes, which were developed to control same-sex practices deemed as endangering the nation-state. Religious antihomosexual ordinances were also a part of this construction of same-sex laws (Kinsman, 2001). These laws were in keeping with the imperial and colonial agenda of the day, which aimed to wipe out and exterminate any sexual practices in violation of heterosexual normativity and those who engaged in them (Epprecht, 1998). Despite the historical creation of global state-sanctioned homophobia, most Western countries have long since reformed these laws, but the damage is irreversible and far-reaching. These statutes have not only aided in the erasure of queer history, the elimination of queer subjects, and the silencing of narratives through fear and state-sanctioned violence, but also left societies with the mechanisms that enable the continued persecution of queer subjects, since many former colonies still practice British, French, and Roman Dutch law and since the moral doctrines of Christianity have replaced traditional indigenous values. This persecution is also extended to countries where Islam is dominant (Tamale, 2011).

The most recent example of how colonial law still applies to modern-day democracies can be seen in the decision of India's Supreme Court in December 2013, which took the world's largest democracy a step back when it decided to overturn the 2009 lower court ruling to decriminalize homosexual acts between consenting adults (Vasudevan, 2013). The original Indian Penal Code mirrors the language that can be found in the laws of all former European colonies: "punishment for activity against the order of nature…" This application of colonial law is also seen across the African continent. The number of constitutional articles regulating homosexuality varies, but the wording of the offense is virtually identical. For example, in Uganda, "any person who has carnal knowledge of any person against the order of nature is guilty of an offense and is liable to life imprisonment" and the Nigerian Penal Code states that "any person who has carnal knowledge of any person against the order of nature … is liable to imprisonment for 14 years" (Madi, 2005).

The concern is as previously stated not with who queers are but with what they do. In current political articulations of queer sexualities in the West, social and cultural norms have been reconceptualized and are standing on their own merit, even though definitions remain in my estimation, unstably defined and coded. In the Western context, the phrase 'lesbian and gay' has become the standard way of referring to what only a few years ago was simply gay and prior to that homosexual (de Lauretis, 1991). Queer politics, on the other hand, has come to replace older modes of lesbian and gay politics by coexisting alongside them while at the same time opening up new possibilities of sexual expression. Queer subjectivity creates a new space in the framework of sexuality. The concept of identity and identity politics becomes crucial in the understanding of the functioning of queer theory.

Queer is said to transcend difference, to be gay or lesbian is merely to be assigned an identity, and to be queer requires the visioning and production of abstract genderless free agents who are definitionally indeterminate (Warner, 1991). Queer theory, then, attempts to unpack the identities 'lesbian' and 'gay,' and subsequently demonstrate how these categories are overdetermined by factors such as heteronormativity, race, gender, and ethnicity (Martinez, 2003). Queer theory thoroughly investigates sexuality by exploring themes of history, marginalization, exclusion, normalcy, social location, and political agency, all of which are crucial in the investigation of lives that stand outside any dominant culture.

Primarily concerned with documenting past and current manifestations and implications of same-sex attractions, queer theory operates from the perspective that heterosexuality or normative sexuality could not exist without queer or anti-normative sexualities. What, then, are the discursive constructions and constructed silences around the relations of race to identity and subjectivity in the practice of same-sex desire within the development of intersectional queer subjectivity? How do sexualized identities shaped by silence coexist with other sexualities that have not experienced historical erasure? How do these subjects articulate the discourse that so forcefully shapes their lives?

The Tale of Two Theories

The introduction of the term 'queer theory' by Theresa de Lauretis in 1991 aimed to encompass new histories of gays and lesbians (Goldman, 1996). In 1989, Crenshaw had introduced the term 'intersectionality,' which set out to amplify the systematic ways in which differences in social location such as,

but not limited to, race, gender, sexuality, and socioeconomic status were interrelated as distinct categories rather than conflated into a single experience of oppression. Difference was seen by both queer and intersectionality theories as central to the understanding of how marginalization was experienced and lived. Queer theory was seen as having the ability to problematize what it meant to be and live as a lesbian or gay person. Queer theory was also seen as the vehicle of desirous utopian imagination, as in the style of Warner (1991), fighting for equality while claiming difference, and demanding political representation, while insisting on its material and historical specificity. Queer celebrated the diversity of subjects who experienced heterosexist oppression without essentializing identity, while at the same time acting as an umbrella term for all those who had been marginalized because they transgressed the heterosexual norm (Khayatt, 2002). Intersectionality's project was to celebrate diversity without conflating identities or creating hierarchies of oppressions, while at the same time acting as an umbrella term for the expressed experience of multiple oppressions. The multiple identities of oppression formed an axis that amplified the magnitude of social inequalities one experiences through categories of oppression that are shaped by one another rather than as a single, discrete, and unrelated event (Crenshaw, 1989).

A queer theoretical practice that operates through an intersectional lens becomes a useful tool for making visible how some desirous subjects pass as normal and how others are rendered marginal. This plays out in countries where the rights of queer citizens are protected, while racialized queers are still experiencing discrimination within these communities. It allows people to see how both social processes and social representation are crucial in the understanding of oppression. Without a doubt, queer can be seen as a term that negatively erases difference, but I am holding on to the utopian use of the term, which is theoretically structured around the concept of intersecting identities as they relate to sexuality.

Readers can also understand queer theory by looking at how it functions. It is my view that queer theory performs as a constant narrative with the power to challenge all norms relating to sexual identity. Intersectionality accounts for the multiple levels across which oppression operates, and queer theory pushes us to deconstruct categorical identities and question normativity. Intersectionality articulates the site of becoming marginalized or oppressed, while queer theory allows us to question norms that perpetuate marginalization or oppression. Together, these theories force us to ask different questions and ask questions differently (Coloma, 2006), and their theoretical systems can be combined not only for academic exercises and purposes, but also as political tools in the creation of epistemologies, social justice coalition building, and practices of resistance. Queer theory, then, can be seen as innovative because of the degree to which it facilitates self-consciousness about how society mediates self-understanding of sexual identities and for its promotion of queer culture building. By this I mean the radical project of recognizing identities, desire, and sex that can exist when the heterosexual subject is not the referent point or the normal discussant of sexual culture.

An intersectional queer theory is a transformative political project that can be used as a critical lens in scholarly practices to decenter the practice of naming mainstream queer subjects as having fixed definitions within collective queer identities or what I also term queer communities of necessity or survival, where diversity is overlooked in favor of gaining perceived monolithic political power. I want to develop a queer academic tradition that focuses on the lived experience of structural inequalities in order to reconceptualize the queer subject as embedded in difference and cultural diversity. It is important to refocus the emphasis on the everyday lives and intimate experiences of those who practice a queer tradition and to take into account the other identities that intersect to place them queerly in the world.

Queer Repositioning

I know that if queer theory is to be truly seen as innovative and to challenge what is 'normal,' it must provide a framework within which to challenge racist, misogynistic, and other oppressive discourses, as well as those that are heterosexist and homophobic, through an intersectional lens. It cannot simply challenge heteronormativity but must instead question and resist the very systems that sustain heteronormativity. There needs to be a clear understanding of the ways in which systems of oppression are linked and create discourses that seek to undermine and expose the entire system through frameworks that address both local and global sexual identities as well as offer ways of moving toward nonoppressive practices.

I have been focusing on the ability of this theory to articulate both a sexual culture and a movement of political resistance, as well as demand a change in sexual politics. Queer theory mirrors the discourse of white gay and lesbian historiography and sociology, both of which have haphazardly added on marginalized groups such as racialized communities as an afterthought with little or no understanding of sociosexual specificity. According to Escoffier (1990), queer theory has placed emphasis on the intellectualized interpretations of texts rather than on the social history or the sociology of gay life. He also notes that this queer theory at this point had failed to incorporate women and people of color in any significant way and postmodern theoretical concepts make interpretation difficult and obtuse for anyone outside of the academy.

I have set up the concept of queer theory to deliver so much, despite the fact that it is a theoretical framework that is thoroughly embedded in white Western academic culture. Those pioneers who produced queer theory have made abstractions of queers of color in particular (Berube, 2001). This reality forces us to investigate and explore how queer theory can be circulated and, at the same time, moved away from its origins in order to be used critically and for the development of oppositional practices. The thought being that in order to increase the political usefulness of queer theory, there needs to be mindful consideration that the theory as a framework has been disciplined by academia.

Just as heterosexuality limits the discourse on other sexualities, the structuring of queer theory around whiteness, for example, limits racial or ethnic discussions of queerness. The continued constructed silence around race and non-Western queer perspectives, or those identities that have not been constructed by Western definitions of sexuality and sexual identity

essentially forces us to essentialize our identities as ethnic and racialized queer communities. This flags how racialized groups are yet again being left out of another theoretical framework. This makes it a priority to investigate exclusion within frameworks that address sexuality and the importance of queer approaches to highlighting the marginalization of certain categories and experiences within its own movement.

At the same time I call into question the use of 'queer theory' as a tool to articulate the lives of racialized queers and also critique the concept of queer resistance or queer liberation. Valdes (2002) wonders whether queer liberation is possible without reference to gender, race, or class. If queer liberation only stands for sexual liberation, then what kind of foundation is it standing on if it is supporting and perpetuating queer male privilege, class privilege, and/or white privilege? There can no longer be a universal category of 'queer' that includes all queers with gender, race, and ethnicity, for instance, being subsumed under some neutral umbrella and used merely as a descriptor. In the critical project that I am proposing, gender, race, and ethnicity are central to the content of a queer theoretical framework.

This debate has existed since the emergence of queer theorizing. Hammonds (1994) suggests that the use of the term queer without the qualifier 'black' first of all highlights the lack of meaningful theoretical engagement with the intersectionality of race and sexuality. For her, the term lesbian, for example, without the racial qualifier can be simply read as white lesbian. What methodologies, then, need to be created to read and understand how sexualities shaped by silence, erasure, and invisibility coexist with other sexualities and address the impact that the visible has on the invisible? Hammonds (1994) articulates this thought: "When I am asked if I am queer I usually say yes even though the ways in which I am queer have never been articulated in the body of work that is now called queer theory" (p. 1).

In queer theory, questions of race and ethnicity tend to be overlooked in order to analyze sexuality in the same manner that race-based scholarship tends to minimize the importance of sexuality and erase homosexuality from its analysis (Somerville, 2000). The notion of race must be central in the historicity and representational formation of a sexuality shaped by Black femaleness and compulsory heterosexuality. Such an analysis shifts queer theories' gaze from the resistance of mainstream heterosexuality to the elaboration of one's identity being located in constantly shifting intersections of oppression that are not only resisting heterosexuality, but other queers as well. As Hammond says:

> Torn between the homophobia of the black community and the racism of the white lesbian community, I need, as a black lesbian, to speak for myself and in my own voice, which is not the voice of the white world." While she does not doubt the good intentions of her colleagues, she does doubt their ability to understand her lesbianism which cannot be taken out of her context as a black woman: "I do not want my black experience filtered through your white academic language, the rage and passion edited out, explained away" (p. 88).

I use this example to propose the implementation of a logic that requires a critical reassessment of the ways in which queer individuals understand and pursue their history, politics, and sexualities and to articulate them in terms of the very differences that characterize queer communities and membership in them. This logic sheds light on the dynamics of heterosexism, homophobia, coming out, and existence in families as well as communities in specific gendered, racial, and ethnic contexts. It will help all individuals move away from primary resistance to negative images of nonwhite or non-Western queer identity toward challenging queer theory that lacks gendered, racial, and ethnic dimensions.

Intersectional Queerness

It might be useful to think of the articulation of an intersectional queer framework as a project of elaboration since the politics of marginal sexualities has not been so much neglected as it has been shielded from public view. Many members of society think that sexuality is a very 'straight'forward political matter and that queer people should not be discriminated against and that is as far as their political interest as a group should go. The question still remains as to whether or in what context do queers have political interests that connect them to broader demands for justice and freedom. Warner (1991) asks the simple but poignant question, 'What do queers want (p. 3)?' Any person who comes to a queer understanding of themselves understands that his/her oppression is connected but not limited to desire, family, fear, shame, health, personal freedom, gender, reproductive rights, parenting, silence, invisibility, public display, terror, violence, intimacy, depression, love, and marriage (GLAD, 2004). Being queer means to address these issues all the time, and deal with the ensuing consequences. As queer activists from non-Western contexts become more and more involved in setting political agendas and, as the rights discourse of internationalism is increasingly extended to more cultural contexts, mainstream queer theorists will have to become more conscious of the limiting localizing tendencies of their theoretical language.

An intersectional queer theory not only allows for the articulation of the lives of complex racialized subjects, but also challenges the mainstream frame of reference. It does not allow for the recycling sociosexual and socioeconomic hierarchies of oppression, which are fought against. The use of the term queer is based on a particular historical experience and many marginalized communities have limited if any historical connection with the term (Walcott, 2006). How, then, do marginalized communities insert themselves into this discussion and how do they lay claim to the discourse? I would say that there is need to interrogate the complex system of sexual norms within all cultures, where the normal racialized subject is not queer and the normal queer is not racialized. To firmly locate queerness in a historical, political, and intellectual context, and to identify the culturally based problematic that queer theory intends to remedy, theorists must determine what Penn (1995) would define as the conceptual category of intersectional queerness.

A poststructural approach to sexuality understands that there is neither true queer history but many histories nor is there a true queer identity but many identities that are fluid and dynamic. One needs to question whether it is a personal, political, or linguistic one? For you to call yourself queer possibly means that you are a certain type of person, that you

engage in certain sexual behaviors, that you will have a certain kind of political orientation, and that you will potentially dress in a particular type of way and look a certain way. Until a word gains wide usage in the culture, the category, identity, or type of person it represents does not exist (Khayatt, 2002). Queerness is socially constructed and cannot be understood without recognizing that an identity and the resulting behavior are derived from social norms. Being queer entails the realignment of identity, politics, and desire as one moves between cultures, regions, and nations (Massaquoi, 2013). For example, within the vastness of languages in the African continent, the term 'queer' could be theoretically interchanged with the terms buyazi, mashoga, tousso bakari, woubie, supi, or kuchu, to name a few. What does it mean to be queerly marked and to answer to those terms in Africa? What happens when that queerly marked African body becomes transnational in a Western context? It involves answering to those terms, acting on emotions, feeling desire, and engaging in the sexual practices that have different meanings, depending on the locality and the circumstances.

With the emergence of the queer refugees in the West, theoretical changes to political and culturally based approaches are required to allow for the articulation of identity, the claiming of space in queer history, and the creation of a political platform in terms of rather than in spite of the dissention that marks them differently from the mainstream queer communities. With this emergence, the complex system of sexual norms in both the refugee's country of origin and country of adoption requires interrogation. The queer refugee is traveling to Western destinations in which queer subjects are fighting for political rights and freedoms for the mainstream white queer communities while often engaging in discriminatory exclusionary patterns of behavior that disregard and invalidate the issues and experiences of queer transnationals (Patton and Sanchez-Eppler, 2000). It is often the case that queer refugees are coming from nations in which there are widespread beliefs supported by religious institutions, cultural norms, and political leadership that homosexuality does not exist, and that it is a Western construction or a Western disease (Hoad, 2007).

Thus, one of the biggest challenges for queer subjects globally is forced invisibility. By using an intersectional lens in the development of queer theoretical frameworks, invisibility could become the starting point for launching a discussion of same-sex desire in which queer refugees could be central figures. This is something that traditional queer theory is unable to accommodate. The queer refugee is invisible in both the country of origin and the country of adoption. Addressing this invisibility becomes the point from which the work of identifying the ways in which queer refugee subjects occupy a perverse subject position begins. Reimagining and reconstituting their relationships within multiple identities and interrogating how nonnormative erotic and gender configurations erupt within sites of extreme heteronormativity will lead to change. A queer theory that is not conscious of the privileging of Western definitions of same-sex sexual practices and experiences risks duplicating an imperial gaze in relation to non-Western sexual practices and identities. An intersectional queer theory would then be required to function not as an autonomous narrative but as one that is enmeshed and immersed within multiple discourses.

Identity based on same-sex desire may be potentially productive of new forms of self, community, and social relations. Bannerji (1993) comments on her disconnection from the white queer androgynous esthetic as she adorns herself in traditional South Asian attire, which has multiple readings for her as a South Asian Canadian lesbian but read as a 'femme' or hyperfeminine identity in a white lesbian context. She challenges the roles and behaviors attributed to that presentation, which complicates her attempt to negotiate ethnicity and sexual identity in a Western context. Queer theory speaks to the need for a particular strategy of reading sexuality outside dominant configurations of visibility, same-sex desire, and identity within cultures and against other cultures. It speaks to Foucault's 'paramount moment of possibility,' the positioning of bodies and pleasures that are not already enmeshed in the discourse of queer theory (Foucault, 1990).

One of the biggest challenges facing any project that attempts to dismantle or interrogate a system in which one group dominates another is to provide for a new system that does not reproduce the structure of the old one (Fanon, 1963). In proposing a need for an intersectional reading of queer theory to emerge, there must be a strategic effort to avoid simply assimilating dominant Western culture and essentializing notions of queer identity. Theorists must become fluent in each other's histories if the goal is to undo the ingrained mythologies that constrain us (Alexander, 1997).

Queer theory aims to identify some of the changes in political and cultural approaches that are necessary for realizing such potential. These changes, I believe, will facilitate positive modes of thinking and behavior that exceed the current exclusionary logic. Heterosexism, for example, can be overcome only by actively imagining a necessarily and desirably queer world (Warner, 2002). This project highlights that any movement of queer liberation cannot be representative without proactively addressing community building, providing the tools for communities to do so, and reaching beyond the intellectual landscape of academic theories to impact the day-to-day lives and practices of queers. It requires looking at the rearticulation and reconfiguration of queer lives that do not necessarily resolve or dissolve their conflicts and tensions but instead refashion them into proactive patterns of thought and behavior.

Conclusion

By expanding the ways of thinking about queerness, queer theories will be opened up to a wider audience (Valdes, 2002). Queer theory does its work by contributing to diverse and sometimes conflicting projects, across interdisciplinary academic fields. Under contemporary social and political conditions, queer theory cannot be understood as ideal in its resistance to heteronormativity, but can ask what the possibilities are for the theory's expansion in a given political moment. What are the possibilities, conditions, and future imaginations of a queer theory that can thoroughly embrace intersectionality? There needs to be an engagement in intersectional bound self-discovery, which involves the understanding of what it means to be an individual whose multiple identities have been imposed upon or premised on exclusion

and invisibility to create a subordinate other who can also recreate the self by acts of refusal. The conceptualization of intersectional queer subjects displaces the neutral (read white) heterosexual norm and launches a critique of queer theory from the locational politic where it is articulated. Collective queer identities acquire specific content from the narratives of both belonging and exclusion and become strategically essentialized as a necessary component of agency, resistance, and survival. It is an identity that is attained through an ongoing process of interpretation of social position and multiple intersecting social locations.

See also: Critical Social Work Practice; Feminist Theory: Radical Lesbian; Heterosexism and Homophobia; Intersectionality and Crime; Intersectionality; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Queer Anthropology; Queer Geography; Sexual Orientation and the Law; Sexual Orientation: Genetic Aspects; Sexual Orientation: Historical and Social Construction; Sexuality and Social Work; Social Work Theory.

Bibliography

Alexander, J., 1997. Erotic autonomy as a politics of decolonization: an anatomy of feminist and state practice in the Bahamas tourist economy. In: Alexander, J., Mohanty, C. (Eds.), Feminist Genealogies, Colonial Legacies, Democratic Futures. Routledge, New York, pp. 63–100.

Bannerji, K., 1993. No apologies. In: Ratti, R. (Ed.), A Lotus of Another Color: An Unfolding of the South Asian Gay and Lesbian Experience. Toronto South Asian Review Press, Toronto, pp. 59–61.

Berube, A., 2001. How gay stays white and what kind of white it stays. In: Rasmussen, B.B. (Ed.), The Making and Unmaking of Whiteness. Duke University Press, Durham, pp. 234–265.

Butler, J., Scott, J., 1992. Feminist Theorize the Political. Routledge, New York.

Coloma, R., 2006. Putting queer to work: examining empire and education. International Journal of Qualitative Studies in Education 19 (5), 639–657.

Crenshaw, K., 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, vol. 140. University of Chicago Legal Forum, 139–167.

Epprecht, M., 1998. Good god almighty, what's this! Homosexual "crime" in early colonial Zimbabwe. In: Murray, S., Roscoe, W. (Eds.), Boy-Wives and Female Husbands: Studies of African Homosexualities. St. Martin's Press, New York, pp. 197–221.

Epstein, S., 1994. A queer encounter: sociology and the study of sexuality. Sociology Theory 12 (2), 188–202.

Escoffier, J., 1990. Inside the ivory closet: the challenges facing lesbian and gay studies. Out/Look 10, 40–48.

Fanon, F., 1963. The Wretched of the Earth. Grove Weidenfeld, New York.

Foucault, M., 1990. The History of Sexuality: Volume One: An Introduction. Vintage Books, New York.

Gays and Lesbians of African Descent (GLAD), December 4, 2004. Focus group discussion, Toronto.

Giffney, N., 2004. Denormatizing queer theory: more than (simply) lesbian and gay studies. Feminist Theory 5 (1), 73–78.

Goldman, R., 1996. Who is that queer queer? Exploring norms around sexuality, race, and class in queer theory. In: Beemyn, B., Eliason, M. (Eds.), Queer Studies: A Lesbian, Gay, Bisexual and Transgender Anthology. New York University Press, New York, pp. 169–182.

Hammonds, E., 1994. Black (w)holes and the geometry of black female sexuality. Differences 6 (2–3), 1–10.

Hoad, N., 2007. African Intimacies: Race Homosexuality and Globalization. University of Minnesota Press, Minneapolis.

Khayatt, D., 2002. Toward a queer identity. Sexualities 5 (4), 487–501.

Kinsman, G., 2001. Challenging queer and Canadian nationalisms. In: Goldie, T. (Ed.), In a Queer Country, Gay and Lesbian Studies in the Canadian Context. Arsenal Pulp Press, Vancouver, pp. 135–209.

de Lauretis, T., 1991. Queer theory: lesbian and gay sexualities. Differences 3 (2), iii–xviii.

Madi, Thuli, September 14, 2005. "Country Profiles" behind the Mask. Web 32 April, 2008.

Manalansan, M., 2003. Global Divas: Filipino Gay Men in the Diaspora. Duke University Press, Durham.

Martinez, J.M., 2003. Racisms, heterosexisms, and identities: a semiotic phenomenology of self-understanding. Journal of Homosexuality 45 (2/3/4), 109–127.

Massad, J., 2007. Desiring Arabs. University of Chicago Press, Chicago.

Massaquoi, N., 2013. No place like home: African refugees and the emergence of a new queer frame of reference. In: Epprecht, M., Nyeck, S. (Eds.), Sexual Diversity in Africa: Politics, Theory and Citizenship. McGill-Queen's University Press, Montreal, pp. 37–53.

Nyeck, S., 2013. Mobilizing against the invisible: erotic nationalism, mass media, and the "paranoid style" in Cameroon. In: Epprecht, M., Nyeck, S. (Eds.), Sexual Diversity in Africa: Politics, Theory and Citizenship. McGill-Queen's University Press, Montreal, pp. 151–169.

Patton, C., Sanchez-Eppler, B., 2000. Queer Diasporas. Duke University Press, Durham.

Penn, D., 1995. Queer: theorizing politics and history. Radical History Review 62, 24–42.

Shoat, E., 2002. Area studies, gender studies, and the cartographies of knowledge. Society Text 20 (3), 67–78.

Somerville, S., 2000. Queering the Color Line: Race and the Invention of Homosexuality in American Culture. Duke University Press, Durham.

Tamale, S., 2011. Introduction. In: Tamale, S. (Ed.), African Sexualities: A Reader. Pambazuka Press, Cape Town, pp. 1–7.

Valdes, F., 2002. Mapping the patterns of particularities: queering the geographies of identities. Antipode 34 (5), 974–987.

Vasudevan, Aniruddan, 23 December, 2013. "India's Gay Sex Ruling Met with Global Outrage" Aljazeera America. Web 29 January, 2014.

Warner, M., 1991. Introduction: fear of a queer planet. Society Text 29, 3–17.

Warner, M., 2002. Publics and Counterpublics. Zone Books, New York.

Walcott, R., 2006. The new African American studies: Blackening queer studies and sexing black studies. A Journal of Lesbian and Gay Studies 12 (3), 510–513.

Questionnaires: Cognitive Approaches

Norbert Schwarz, Department of Psychology, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Respondents' answers to standardized questions presented in self-administered questionnaires or face-to-face and telephone interviews are a major source of social science data. These answers, however, are profoundly influenced by characteristics of the question asked, including the wording of the question, the format of the response alternatives, and the context in which it is presented. The cognitive and communicative processes underlying these influences are increasingly well understood, and this article reviews their implications for question wording and questionnaire construction.

Respondents' Tasks

Answering a question requires that respondents (1) interpret the question to understand what is meant and (2) retrieve relevant information from memory to form an answer. In most cases, they cannot provide their answer in their own words but (3) need to map it onto a set of response alternatives provided by the researcher. Finally, (4) respondents may wish to 'edit' their answer before they communicate it to the interviewer for reasons of social desirability and self-presentation. Respondents' performance at each of these steps is context dependent and profoundly influenced by characteristics of the questionnaire (Sudman et al., 1996, Chapter 3), including its administration mode (self-administered or a face-to-face or telephone interview).

Understanding the Question

The key issue at the question comprehension stage is whether respondents' understanding of the question matches the meaning the researcher had in mind. Not surprisingly, methodology textbooks urge researchers to write simple questions and to avoid unfamiliar or ambiguous terms (for good practical advice, see Bradburn et al., 2004). However, understanding the *literal meaning* of a question is not sufficient to provide an informative answer. When asked, "What have you done today?" respondents are likely to understand the words – yet, they still need to determine what kind of activities the researcher is interested in. Should they report, for example, that they took a shower, or not? Providing an informative answer requires inferences about the *questioner's intention* to determine the *pragmatic meaning* of the question.

To infer the pragmatic meaning, respondents draw on a broad range of contextual information, including the content of adjacent questions, the nature of the response alternatives, and the graphical layout of questions. Their use of this information is licensed by the tacit assumptions that govern the conduct of conversation in daily life, which respondents apply to the research interview (Clark and Schober, 1992; Schwarz, 1996). These pragmatic inferences have important implications for questionnaire construction.

Question Context

Respondents interpret questions in the context of the overall interview and a term like 'drugs' acquires different meanings when presented in a survey pertaining to respondents' medical history rather than to crime in the neighborhood. Moreover, they assume that adjacent questions are meaningfully related to one another, unless otherwise indicated. They may infer, for example, that an ambiguous 'educational contribution' refers either to fellowships that students receive or to tuition they have to pay, depending on the content of adjacent questions (Schwarz, 1996, Chapters 3–4). Finally, they hesitate to reiterate information they have already provided, consistent with conversational norms that discourage redundancy. For example, questions pertaining either to happiness with one's life or to life satisfaction typically elicit highly similar answers unless both questions are presented together. In the latter case, respondents assume the researcher wouldn't ask for the same thing twice and differentiate between both questions, resulting in less similar answers (Schwarz, 1996, Chapter 6). In self-administered questionnaires, question interpretation is influenced by preceding as well as following questions, whereas only preceding questions can exert an influence in face-to-face and telephone interviews.

Response Alternatives

Most questions are presented with a fixed set of response alternatives, which facilitates administration and data analysis. Far from being 'neutral' measurement devices, however, response alternatives influence question interpretation (Schwarz, 1996, Chapter 5). First, consider the differences between *open* and *closed* response formats. When asked in an open response format, "What have you done today?" respondents are likely to omit activities that the researcher is obviously aware of (e.g., "I gave a survey interview") or may take for granted anyway (e.g., "I took a shower"). Yet, if these activities were included in a closed list of response alternatives, most respondents would endorse them. At the same time, such a list would reduce reports of activities that are not represented on the list, even if an 'other' option were provided, which respondents rarely use. Both of these question form effects reflect that response alternatives can clarify the intended meaning of a question by specifying what the researcher is interested in. When using a closed response format, it is therefore important to ensure that the list of response alternatives covers the full range of behaviors or opinions.

Similarly, suppose respondents are asked how frequently they felt 'really irritated' recently. To infer if 'really irritated'

refers to major or to minor annoyances, respondents may draw on an apparently formal feature of the question, namely the numeric values of the *frequency scale*. When the scale presents low- (high-) frequency alternatives, respondents infer that the researcher is interested in rare (frequent) events. Hence, they report on major annoyances (which are relatively rare) when given a low-frequency scale, but on minor annoyances when given a high-frequency scale. As a result, the same question elicits reports of substantively different experiences, depending on the frequency scale used (Schwarz, 1996, Chapter 5).

In sum, apparently formal differences in the response alternatives can profoundly influence respondents' interpretation of the question. Far from reflecting superficial responding, such effects indicate that respondents do their best to make sense of the question asked by drawing on the full range of information available to them – including information the researcher never intended to provide. To safeguard against unintended question interpretations, researchers have developed a number of different pretesting procedures, reviewed in Section Pretesting.

Age-Related Differences

Respondents' extensive use of contextual information requires that they hold the question in mind and relate it to other aspects of the questionnaire to determine its intended meaning. This entails considerable demands on respondents' cognitive resources. Given that these resources decline with increasing age, we may expect that older respondents are less likely to use, or less successful in using, contextual information at the question comprehension stage. Indeed, question order effects decrease with age, and this has a greater impact the poorer respondents' memory is (for a review, see Schwarz and Knäuper, 2000). Children and adolescents, whose cognitive capabilities are not yet fully developed, appear to show a similar deficit in incorporating relevant contextual information into survey responding (Fuchs, 2005). Age-related differences in the use of contextual information are more likely in face-to-face and telephone interviews, where respondents cannot look back to earlier questions, than in self-administered questionnaires.

Recalling or Computing a Judgment

Once respondents have determined what the researcher is interested in, they need to recall relevant information from memory. In some cases, respondents may have direct access to a previously formed relevant judgment that they can offer as an answer. In most cases, however, they will not find an appropriate answer readily stored in memory and will need to compute a judgment on the spot. The different judgmental processes pertaining to behavioral questions and opinion questions are discussed in Section Asking Questions about Behavior.

Formatting the Response

Unless the question is asked in an open response format, respondents need to format their answer to fit the response alternatives provided by the researcher. Respondents observe these *question constraints* and avoid answers that are not explicitly offered, as noted in Section Understanding the Question. For example, many more respondents answer, 'Don't know,' or choose a middle alternative, when these options are explicitly offered than when they are not (see Sudman et al., 1996, Chapters 3–4). Under which conditions the addition or omission of these response options results in different substantive conclusions is the topic of some debate.

Response-formatting processes are particularly relevant when *rating scales* are used (Sudman et al., 1996, Chapter 4). Specifically, respondents use the most extreme stimuli to anchor the endpoints of the scale. Hence, a given stimulus will be rated as less extreme if presented in the context of a more extreme one than if it is presented in the context of a less extreme one. In addition, if the number of stimuli to be rated is large, respondents attempt to use all categories of the rating scale about equally often. Hence, two similar stimuli may receive different ratings when only a few stimuli are presented, but may receive the same rating when a large number of stimuli has to be located along the scale. Accordingly, ratings of the same object cannot be directly compared when they were collected in different contexts, rendering comparisons over time or between studies difficult. The psychometric properties of rating scales have been the topic of extensive investigation (see Tourangeau et al., 2000, Chapter 8). Seven-point scales seem to be best in terms of reliability, percentage of undecided respondents, and respondents' ability to discriminate between scale points. Moreover, scales that provide verbal labels for each scale value seem more reliable than scales with verbal endpoints only.

Finally, respondents' choice of a given response alternative is often influenced by the order in which the alternatives are presented. Under the auditory presentation conditions of telephone interviews, a given alternative is more likely to be endorsed when it is presented last rather than first (referred to as a recency effect). Under the visual presentation conditions of a self-administered questionnaire, a given alternative is more likely to be endorsed when it is presented first rather than last (referred to as a primacy effect; see Sudman et al., 1996, Chapter 6, for a discussion of the underlying processes). Because limited cognitive resources make it more difficult to consider multiple response alternatives, response order effects increase with age and lower educational attainment, which can undermine group comparisons (see Schwarz and Knäuper, 2000).

Editing the Response

Finally, respondents may want to edit their response before they communicate it, reflecting considerations of social desirability and self-presentation (for a review, see DeMaio, 1984). Not surprisingly, the impact of self-presentation concerns is more pronounced in face-to-face interviews than in self-administered questionnaires. It is important to emphasize, however, that influences of social desirability are limited to potentially threatening questions and are typically modest in size. These influences can be reduced through techniques that ensure the confidentiality of the answer (for detailed advice, see Bradburn et al., 2004).

Asking Questions about Behavior

Most questions about respondents' behavior are frequency questions, pertaining, for example, to how often the

respondent has bought something or has missed a day at work during some specified period of time. Researchers would ideally like respondents to identify the behavior of interest, to scan the reference period, to retrieve all instances that match the target behavior, and to count these instances to determine the overall frequency of the behavior. This, however, is the route that respondents are least likely to take.

Except for rare and very important behaviors, respondents are unlikely to have detailed representations of numerous individual instances of a behavior stored in memory. Instead, the details of various instances of closely related behaviors blend into one global representation, rendering it difficult to distinguish and retrieve individual episodes. Accordingly, a 'recall-and-count' model does not capture how respondents answer questions about frequent behaviors or experiences. Rather, their answers are likely to be based on some fragmented recall and the application of inference rules to compute a frequency estimate. Accordingly, the best we can hope for under most conditions is a reasonable estimate, unless the behavior is rare and of considerable importance to respondents.

Conway (1990) provides an introduction to psychological research into autobiographical memory (i.e., individuals' memory for their experiences and behaviors), and Schwarz and Oyserman (2001), Sudman et al. (1996, Chapters 7–9), and Tourangeau et al. (2000, Chapters 3–5) review the implications for questionnaire construction. The following sections summarize what researchers can do to facilitate recall and review what has been learned about the influence of questionnaire characteristics on respondents' estimation strategies.

Facilitating Recall

Recall Cues and Reference Periods

If researchers are interested in obtaining reports that are based on recalled episodes, they may simplify respondents' task by providing appropriate *recall cues* and by restricting the recall task to a short and recent reference period. There are, however, important drawbacks to both of these strategies. Although the quality of recall will generally improve as the retrieval cues become more specific, respondents are likely to restrict their memory search to the particular cues presented to them, reflecting that the cues constrain the meaning of the question. As a result, respondents will omit instances that do not match the specific cues, resulting in underreports if the list is not exhaustive. Moreover, using a short reference period may result in many 'zero' answers from respondents who rarely engage in the behavior, thus limiting later analyses to respondents with a high behavioral frequency. Finally, reference periods of different lengths may result in different question interpretations: long reference periods (How often have you been angry last year?) suggest that the researcher is interested in relatively rare events, whereas short reference periods (How often have you been angry last week?) indicate an interest in frequent events. Hence, respondents may report on substantively different behaviors, for example rare and intense versus frequent but minor episodes of anger.

Not surprisingly, different recall cues are differentially effective. The date of an event is the poorest cue, whereas cues pertaining to what happened, where it happened, and who was involved have been found to be very effective. In addition, recall will improve when respondents are given sufficient time to search memory. Recalling specific events may take up to several seconds, and repeated attempts to recall may result in the retrieval of additional material, even after a considerable number of previous trials. Unfortunately, respondents are unlikely to have sufficient time to engage in repeated retrieval attempts in most research situations, and may often not be motivated to do so even if they had the time. This is particularly crucial in the context of survey research, where the available time per question is usually less than 1 min.

Moreover, the direction in which respondents search memory may influence the quality of recall. Specifically, better recall is achieved when respondents begin with the most recent occurrence of a behavior and work backward in time than when they begin at the beginning of the reference period. This presumably occurs because memory for recent occurrences is richer and the recalled instances may serve as cues for recalling previous ones. Given free choice, however, respondents tend to prefer the less efficient strategy of forward recall.

Yet, even under optimal conditions, respondents will frequently be unable to recall an event or some of its critical details, even if they believed they would 'certainly' remember it at the time it occurred. In general, the available evidence suggests that respondents are likely to underreport behaviors and events, which has led many researchers to assume that higher reports of mundane behaviors are likely to be more valid. Accordingly, a 'the-more-the-better' rule is frequently substituted for external validity checks.

Dating Recalled Instances

After recalling or reconstructing a specific instance of the behavior under study, respondents have to determine if this instance occurred during the reference period. This requires that they understand the extension of the reference period and that they can accurately date the instance with regard to that period. Reference periods that are defined in terms of several weeks or months are highly susceptible to misinterpretations. For example, the term "during the last 12 months" can be construed as a reference to the last calendar year, as including or excluding the current month, and so on. Similarly, anchoring the reference period with a specific date (Since March 1, how often …?) is not very helpful because respondents will usually not be able to relate an abstract date to meaningful memories.

The most efficient way to anchor a reference period is the use of salient personal or public events, often referred to as 'landmarks.' In addition to improving respondents' understanding of the reference period, the use of landmarks facilitates the dating of recalled instances. Given that the calendar date of an event will usually not be among its encoded features, respondents were found to relate recalled events to other, more outstanding events in order to reconstruct the exact time and day. Accordingly, using public events, important personal memories, or outstanding dates (such as New Year's Eve) as landmarks was found to reduce dating biases (for reviews, see Sudman et al., 1996; Tourangeau et al., 2000).

Without a chance to relate a recalled event to a well-dated landmark, time dating is likely to reflect both forward and backward telescoping. That is, distant events are assumed to have happened more recently than they did, whereas recent

events are assumed to be more distant than they are (Sudman et al., 1996, Chapter 8).

Calendar Methods

The currently most successful approach to improving autobiographical recall is offered by calendar methods that combine key characteristics of the above strategies in a single data collection format (see the contributions in Belli et al., 2009). To help respondents recall their alcohol consumption during the last week, for example, respondents may be given a calendar grid that provides a column for each day of the week, cross-cut by rows that pertain to relevant contexts. Thus, they may be asked to enter for each day what they did, who they were with, if they ate out, and so on. Reconstructing the last week in this way provides a rich set of contextual cues, with entries in one row often prompting memories relevant to a different row. Based on this rich network of associations, individual episodes are more likely to be retrieved than under other methods, and each recalled episode may prompt additional memories. The method also encourages respondents to correct earlier answers as additional information comes to mind. This approach is costly in terms of interview time, but excessive cost can be avoided by implementing it in self-administered or online formats.

Estimation Strategies

Even with extensive encouragement and opportunity to search memory, respondents will often need to rely on estimation strategies to arrive at an answer (Sudman et al., 1996, Chapter 9) – and under the time pressure of most survey interviews, estimation is more likely than detailed recall. Which estimation strategy they use is often a function of the research instrument and hence to some extent is under researchers' control.

Decomposition Strategies

Many recall problems become easier when the recall task is decomposed into several subtasks. To estimate how often she has been eating out during the last three months, for example, a respondent may determine that she eats out about every weekend and had dinner at a restaurant this Wednesday, but apparently not the week before. Based on such partial recall, she may arrive at an estimate of "18 times during the last three months." Such estimates are likely to be accurate if the respondent's inference rule is adequate, and if exceptions to the usual behavior are rare.

In the absence of these fortunate conditions, however, decomposition strategies are likely to result in incorrect estimates because people usually overestimate the occurrence of low-frequency events and underestimate the occurrence of high-frequency events. Hence, asking for estimates of a global, and hence frequent, category (e.g., 'eating out') is likely to elicit an underestimate, whereas asking for estimates of a narrow, and hence rare, category (e.g., 'eating at a Mexican restaurant') is likely to elicit an overestimate. The observation that decomposition usually results in higher estimates does not therefore necessarily reflect better recall (see Schwarz and Oyserman, 2001).

Subjective Theories

A particularly important inference strategy is based on subjective theories of stability and change (Ross, 1989). In answering retrospective questions, respondents often use their current behavior or opinion as a benchmark and invoke an implicit theory of self to assess whether their past behavior or opinion was similar to, or different from, their present behavior or opinion. The resulting estimates are correct to the extent that the subjective theory is correct. This, however, is often not the case.

In many domains, individuals assume a rather high degree of stability, resulting in retrospective reports that are too close to the current state of affairs, thus underestimating the degree of change that has occurred over time. For example, retrospective estimates of income or of tobacco, marijuana, and alcohol consumption were found to be heavily influenced by respondents' income or consumption habits at the time of the interview. On the other hand, when respondents have reason to believe in change, they will detect change, even though none has occurred. Assuming, for example, that therapy is helpful, respondents may infer that their problem behavior prior to therapy was more frequent than was actually the case. Relying on such theory-driven retrospective reports, researchers may infer therapeutic success where none has occurred (see Ross, 1989, for example).

Response Alternatives

In many studies, respondents are asked to report their behavioral frequencies by checking the appropriate alternative on a frequency scale of the type shown in Table 1. Assuming that the researcher constructed a meaningful scale, respondents commonly believe that the values in the middle range of the scale reflect the 'average' or 'usual' behavioral frequency, whereas the extremes of the scale correspond to the extremes of the distribution. Given this assumption, respondents can use the range of the response alternatives as a frame of reference in estimating their own behavioral frequency. This strategy results in higher estimates along scales that present high- rather than low-frequency response alternatives, as shown in Table 1. As may be expected, the impact of frequency scales is most pronounced for frequent and mundane behaviors, which are poorly represented in memory (for a review, see Schwarz and Oyserman, 2001).

Table 1 Reported daily TV consumption as a function of response alternatives

Reported daily TV consumption (in hours)			
Low-frequency alternatives		*High-frequency alternatives*	
Up to $^1/_2$	7.4%	Up to $2^1/_2$	62.5%
$^1/_2$ to 1	17.7%	$2^1/_2$ to 3	23.4%
1 to $1^1/_2$	26.5%	3 to $3^1/_2$	7.8%
$1^1/_2$ to 2	14.7%	$3^1/_2$ to 4	4.7%
2 to $2^1/_2$	17.7%	4 to $4^1/_2$	1.6%
More than $2^1/_2$	16.2%	More than $4^1/_2$	0.0%

Note: $N = 132$.
Adapted from Schwarz, N., Hippler, H.J., Deutsch, B., Strack, F., 1985. Response scales: effects of category range on reported behavior and comparative judgments. Public Opinion Quarterly 49, 388–395. Reprinted by permission.

In addition to influencing respondents' behavioral reports, response alternatives may also affect subsequent *comparative judgments*. For example, a frequency of '$2^1/_2$ h per day' constitutes a high response on the low-frequency scale, but a low response on the high-frequency scale shown in Table 1. Checking this alternative, a respondent may infer that her own TV consumption is above average in the former case, but below average in the latter. Hence, respondents in this study reported lower satisfaction with the variety of things they do in their leisure time when the low-frequency scale suggested that they watch more TV than most other people.

To avoid these systematic influences of response alternatives, it is advisable to ask frequency questions in an open response format, such as "How many hours a day do you watch TV? ___ hours per day." Note that such an open format needs to specify the relevant units of measurement, such as 'hours per day,' to avoid answers like 'a few.'

As another alternative, researchers are often tempted to use 'vague quantifiers,' such as 'sometimes,' 'frequently,' and so on. This, however, is the worst possible choice. Most importantly, the same expression denotes different frequencies in different content domains; for example, 'frequently' suffering from headaches reflects higher absolute frequencies than 'frequently' suffering from heart attacks. Moreover, different respondents use the same term to denote different objective frequencies of the same behavior. For example, suffering from headaches 'occasionally' denotes a higher frequency for respondents with a medical history of megrim than for respondents without that megrim history. Accordingly, the use of vague quantifiers reflects the objective frequency relative to respondents' subjective standard, rendering vague quantifiers inadequate for the assessment of objective frequencies, despite the popularity of their use (Pepper, 1981).

Pretesting

Given the complexity of the question-answering process, extensive pretesting is of paramount importance. Traditionally, questions were often assumed to 'work' if respondents could provide an answer. Recent psychological research, however, facilitated the development of more sensitive pretest procedures (for comprehensive reviews, see the contributions in Schwarz and Sudman, 1996; for good practical advice, see Willis, 2005). These procedures include the extensive use of probes and think-aloud protocols (summarily referred to as *cognitive interviewing*), detailed coding of interview transcripts, and the use of expert systems that alert researchers to likely problems. Drawing on these procedures, researchers can increase the odds that respondents understand the question as intended and use cognitive strategies that are likely to produce a meaningful answer.

See also: Attitude Measurement.

Bibliography

Belli, Robert, Alwin, Duane, Stafford, Frank (Eds.), 2009. Using Calendar and Diary Methods in Life Events Research. Sage, Newbury Park, CA.

Bradburn, Norman, Sudman, Seymour, Wansink, Brian, 2004. Asking Questions, second ed. Jossey Bass, San Francisco, CA.

Clark, Herbert H., Schober, Michael F., 1992. Asking questions and influencing answers. In: Tanur, Judith M. (Ed.), Questions about Questions. Russel Sage, New York, pp. 15–48.

Conway, Michael A., 1990. Autobiographical Memory: An Introduction. Open University Press, Buckingham, UK.

DeMaio, Teresa J., 1984. Social desirability and survey measurement: a review. In: Turner, Clyde F., Martin, Elizabeth (Eds.), Surveying Subjective Phenomena, vol. 2. Russell Sage, New York, pp. 257–281.

Fuchs, Marek, 2005. Children and adolescents as respondents: experiments on question order, response order, scale effects and the effect of numeric values associated to response options. Journal of Official Statistics 21, 701–725.

Pepper, S.C., 1981. Problems in the quantification of frequency expressions. In: Fiske, D.W. (Ed.), Problems with Language Imprecision, New Directions for Methodology of Social and Behavioral Science, vol. 9. Jossey-Bass, San Francisco, CA, pp. 25–41.

Ross, Michael, 1989. The relation of implicit theories to the construction of personal histories. Psychological Review 96, 341–357.

Schwarz, Norbert, 1996. Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation. Erlbaum, Hillsdale, NJ.

Schwarz, Norbert, Knäuper, Bärbel, 2000. Cognition, aging, and self-reports. In: Park, Denise, Schwarz, Norbert (Eds.), Cognitive Aging. A Primer. Psychology Press, Philadelphia, PA, pp. 233–252.

Schwarz, Norbert, Oyserman, Daphna, 2001. Asking questions about behavior: cognition, communication and questionnaire construction. American Journal of Evaluation 22, 127–160.

Schwarz, Norbert, Sudman, Seymour (Eds.), 1996. Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. Jossey-Bass, San Francisco.

Sudman, Seymour, Bradburn, Norman M., Schwarz, Norbert, 1996. Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco, CA.

Tourangeau, Roger, Rips, Lance J., Rasinski, K., 2000. The Psychology of Survey Responses. Cambridge University Press, Cambridge, UK.

Willis, Gordon B., 2005. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Sage, Thousand Oaks, CA.

Questionnaires: Collection of Subjective Expectations Data

Charles F Manski, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Questionnaires are used to elicit from individuals a broad array of objective information as well as subjective information about their thoughts and perceptions. This article focuses on a single topic about questionnaires – how to elicit respondents' subjective expectations of future events. Subjective expectations are being measured with increasing frequency, yet the social and behavioral sciences have not formed a common view on how expectations should be conceptualized and measured. A particularly intriguing controversy, described in this article, concerns the relative merits of verbal and probabilistic forms of questioning.

Questionnaires are used to elicit from individuals a broad array of objective information about their experiences and environments, as well as subjective information about their thoughts and perceptions. An important objective of social and behavioral science research is to make questionnaires more informative. In practical terms, how should the questionnaire be administered and what should be its content? Aspects of questionnaire administration include when and where the data are collected, who does the data collection, and the mode of administration (e.g., face-to-face, telephone, or the Internet). Aspects of content begin with the basic matter of determining the purposes of the questionnaire and end with specific issues of item wording, sequencing of questions, and formatting of responses.

This article focuses on a single topic about questionnaires – how to elicit respondents' subjective expectations of future events. For example, demographers ask women to predict their future fertility, political scientists ask citizens to predict their voting behavior, and economists ask persons to predict their future incomes. Researchers use the responses to expectations questions to predict these and other future events. They also use the responses to assess how individuals' expectations are related to the decisions they make today.

Subjective expectations of future events are being measured with increasing frequency. However, the social and behavioral sciences have not formed a common view on how expectations should be conceptualized and measured. A particularly intriguing controversy, described in this article, concerns the relative merits of verbal and probabilistic forms of questioning.

Subjective Concepts and Measures

A US National Research Council (NRC) panel study on *Surveying Subjective Phenomena* distinguished subjective from objective phenomena as follows (Turner and Martin, 1984, p. 8):

> Subjective phenomena are those that, in principle, can be directly known, if at all, only by persons themselves, although a person's intimate associates or a skilled observer may be able to surmise from indirect evidence what is going on 'inside.' Objective phenomena are those that can be known by evidence that is, in principle, directly accessible to an external observer.

The NRC panel acknowledged that the sharp dichotomy just drawn is not always tenable, but the essential idea clearly has content: subjective phenomena are not amenable to external verification.

The social and behavioral sciences have a common interest in understanding subjective phenomena, but the various disciplines have not been able to agree on concepts or measures. The conceptual problem is to characterize what is going on 'inside.' A reader of social and behavioral science research encounters an enormous variety of subjective concepts in use, with concepts that are central to one discipline often playing no role in others. A sense of this variety can be obtained by scanning the introductory chapter of Turner and Martin (1984), which mentions affect, aspirations, attitudes, disappointments, dispositions, emotion, frustrations, fulfillment, happiness, hopes, intentions, motives, opinions, plans, preferences, quality of life, satisfaction, and values. The measurement problem is to obtain empirical evidence on what is going on 'inside.'

A reader of empirical research in sociology, social psychology, or political science finds a long history of analysis of subjective data elicited from survey respondents through questionnaires. A reader of economics finds relatively little use of questionnaire responses until the early 1990s but a growing body of research since then. Economists long sought to infer subjective phenomena purely from observations of behavior and experiences but have increasingly become comfortable with measurement of expectations and some other subjective concepts. A reader of anthropology finds little reliance on questionnaire responses. Anthropologists, who tend to recoil from the tightly scripted structure of questionnaires, have preferred to elicit information through conversations.

To illuminate the diversity of perspectives on subjective phenomena, this article examines how different disciplines have used questionnaires to study subjective expectations about potential future events. How persons perceive the future is a shared concern of all social and behavioral science disciplines, yet there is a wide spectrum of views on how expectations should be conceptualized and measured.

Section Inference on Expectations in Economics examines the historical practice in economics of inferring expectations from realizations data, and recalls early efforts to collect data on economic expectations through questionnaires. Section Qualitative Attitudinal Research describes qualitative attitudinal research as practiced by sociologists and social

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44046-8

psychologists. Section Elicitation of Probabilistic Expectations describes the growing body of research in economics and cognitive psychology eliciting subjective probabilities of future events. Section Qualitative versus Probabilistic Expectations considers the relative merits of qualitative and probabilistic modes of questioning. Section Frontiers of Research discusses the frontiers of research on subjective expectations. See Manski (2004) for an extensive exposition of the subjects discussed below.

Inference on Expectations in Economics

Early in their careers, economics students are commonly taught that a good economist believes only what people do, not what they say. Skeptical of subjective data, economists have sought to infer expectations of future events by combining data on past experiences with assumptions about processes of expectations formation. This is a daunting task. To infer expectations from data on past experiences, economists must speculate about what information persons possess and how they use the available information to form their expectations.

Consider, for example, the matter of income expectations. Economists often assume that persons use their own past incomes to forecast their future incomes. Perhaps so, but how do persons form expectations of future income conditional on past income? Economists often assume that persons have *rational expectations*; that is, they know the actual stochastic process generating their income streams. Perhaps so, but what is this stochastic process? The practice has been to specify the income process up to some parameters and use available data on income realizations to estimate the parameters. Authors differ, however, in their specifications. Therefore, the literature displays competing models of expectations formation, with no clear way of distinguishing among them.

If inference on expectations from data on experiences is so challenging, why would economists not collect and analyze data on expectations? The answer lies at least partly in the history of a scientific controversy that began in the 1940s and persisted until the 1960s. In the 1940s, the US Federal Reserve Board began to fund an annual Survey of Consumer Finances, conducted by the University of Michigan Survey Research Center (SRC), that elicited qualitative assessments of expected household finances. A typical question took this form:

> How about a year from now – do you think you people will be making more money or less money than you are now, or what do you expect?

The usefulness of responses to such qualitative questions was controversial and the Federal Reserve Board appointed a committee to assess their value. The Federal Reserve Consultant Committee on Consumer Survey Statistics (1955) issued findings that questioned the predictive power of the SRC expectations data. The negative findings of the committee were challenged by SRC researchers, notably George Katona, a leading proponent of research on consumer attitudes and expectations. A contentious conference on expectations data followed (National Bureau of Economic Research, 1960). Then Juster (1964) reported an intensive study, drawing largely negative conclusions, of the usefulness of qualitative expectations data in predicting individual behavior.

The Federal Reserve – SRC controversy only concerned the usefulness of a certain type of qualitative expectations data in predicting consumer behavior. Academic economists, however, appear to have drawn the broader conclusion that all expectations data are suspect. Economists also failed to take note when Juster (1966) proposed elicitation of probabilistic expectations of consumer durable purchases and reported favorable empirical evidence. Exploration of this idea had to wait a quarter century, as described in Section Elicitation of Probabilistic Expectations.

Qualitative Attitudinal Research

The Federal Reserve – SRC controversy may have soured most economists on the usefulness of qualitative expectations data, but the collection and analysis of such data have long been important activities in other behavioral and social science disciplines. Many questions in common use have a structure similar to that of this General Social Survey (GSS) question eliciting expectations of job loss:

> Thinking about the next twelve months, how likely do you think it is that you will lose your job or be laid off – very likely, fairly likely, not too likely, or not at all likely?

This GSS question illustrates well a basic persistent problem that researchers face in interpreting qualitative expectations data – assessment of the interpersonal and intrapersonal comparability of responses. Respondents are often asked to report whether a specified event is 'expected,' 'likely,' or 'probable' to occur. Sometimes they are asked to report their strength of belief in the event on a numerical scale with undefined units. Do different respondents interpret these verbal phrases and numerical scales in the same way? Does a given respondent interpret them in the same way when asked about different events? Cognitive research does not give reason to think that responses should be or are comparable. Indeed, the available empirical evidence indicates that interpretations of qualitative expectations questions vary substantially between persons (Section Qualitative versus Probabilistic Expectations).

Concern about interpersonal comparability comes into even sharper focus when considering the events about which respondents are asked. Attitudinal researchers often ask respondents about vaguely defined future outcomes, as in this SRC question:

> Now looking ahead – do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?

How should a respondent interpret the phrase "better off financially?" Do different respondents interpret it the same way? The data provide no way of knowing.

A second persistent problem in qualitative attitudinal research is that the coarseness of the response options limits the information contained in the responses. Consider, for example, the fertility question posed to female respondents in the annual June Supplement to the Current Population Survey of the US Bureau of the Census:

> Looking ahead, do you expect to have any (more) children?

> Yes No Uncertain

The three response options can express little of the richness of the uncertainty that women may perceive about their future childbearing. Savage (1971, p. 795) called attention to the problem when he observed that "Yes, No, or Maybe is Not Enough." Juster (1966) hypothesized that a person facing a Yes/No expectations question responds as would a statistician asked to make a best point prediction of a future random event. Manski (1990) formalized this hypothesis and derived an upper bound on the information about expectations a researcher can extract from responses to Yes/No questions.

Elicitation of Probabilistic Expectations

A growing body of research in economics and cognitive psychology elicits expectations in the form of subjective probabilities. For example, the Survey of Economic Expectations used this question to elicit expectations of job loss:

> What do you think is the percent chance that you will lose your job during the next 12 months?

The National Longitudinal Study of Youth: 1997 Cohort of the US Bureau of Labor Statistics elicited the fertility expectations of teenage (girls/boys) as follows:

> What is the percent chance that you will become the [mom/dad] of a baby sometime between now and when you turn 20?

Probabilistic questions of similar structure have been posed in other national surveys, such as the Health and Retirement Study, The Bank of Italy Survey on Household Income and Wealth, and the American Life Panel.

Expectations have been elicited for macroeconomic events (e.g., stock market returns), for risks that a person faces (e.g., job loss, crime victimization, and mortality), for future income (e.g., the earnings returns to schooling and Social Security benefits), and for choices that persons make (e.g., durable purchases and voting choices). The review articles of Manski (2004) and Hurd (2009) describe the emergence of this field of empirical research and summarize a range of applications.

The idea that probabilistic elicitation of expectations might improve on the traditional qualitative approaches of attitudinal research developed independently in economics and in cognitive psychology. In economics, the idea appears to have originated with Juster (1966), who recommended elicitation of purchase probabilities for consumer durable goods. Market researchers were immediately attracted to Juster's proposal, and elicitation of purchase probabilities has since become a common practice in market research. However, the idea that expectations might be elicited probabilistically from survey respondents did not draw serious attention within economics until the early 1990s, when economists began to reassess the discipline's self-imposed prohibition on the collection and analysis of subjective data.

Elicitation of subjective probabilities has a more continuous history of research in cognitive psychology. Psychologists have long conducted calibration studies assessing the accuracy of elicited expectations reports when the outcome is known. Lichtenstein et al. (1982) review findings from calibration studies dating back to 1906. The older studies mostly examined the accuracy of experts (e.g., weather forecasters' reported probabilities of precipitation), whereas more recent research typically analyzes the expectations of nonexperts, especially students situated in a cognitive laboratory.

Qualitative versus Probabilistic Expectations

Economists have embraced probabilistic measurement of expectations, whereas attitudinal researchers continue to pose questions that seek verbal responses. Psychologists use both practices.

There is reasonably widespread agreement that elicitation of probabilistic expectations has several desirable features. Perhaps the most basic attraction is that probability provides a well-defined absolute numerical scale for responses. Hence there is some reason to think that responses may be interpersonally and intrapersonally comparable.

A second attraction is that some empirical assessment of the internal consistency and external accuracy of respondents' expectations is possible. A researcher can use the algebra of probability (Bayes theorem, the law of total probability, etc.) to examine the internal consistency of a respondent's elicited expectations about different events. In those cases where probability has a frequentist interpretation, a researcher can compare elicited subjective probabilities with known event frequencies, as is done in calibration studies, and reach conclusions about the correspondence between subjective beliefs and frequentist realities.

A third consideration is the usefulness of elicited expectations in predicting prospective outcomes. Suppose that respondents have reasonably accurate expectations about the likelihood of future events. Then, as argued by Juster (1966), Savage (1971), and Manski (1990), numerical responses to probability questions should have more predictive power than do categorical responses to qualitative expectations questions.

These a priori desirable features of probabilistic expectations data must be weighed against a set of issues. Perhaps the most basic concern is that there is no direct way to assess how well elicited subjective probabilities reflect respondents'

thinking. Of course this concern is not specific to elicitation of probabilities. Interpretation of responses to subjective questions of any type runs up against the generic problem that a researcher cannot directly observe a respondent's thinking.

A specific concern in the elicitation of probabilities is that respondents may not think probabilistically about uncertain events. A large and diverse literature suggests that persons may think about uncertain events using less than the full structure of modern probability theory. If so, then questions eliciting precise probabilities force respondents to give point responses when only interval responses may be meaningful. In a study of rounding of probabilistic responses, Manski and Molinari (2010) first asked respondents to the American Life Panel to provide precise probabilities of future events and then gave them the opportunity to provide interval responses. They found that some respondents continued to provide precise probabilities but others reported intervals when permitted to do so.

Within cognitive psychology, there has long been controversy about the way in which humans internally represent their beliefs, and their ability and willingness to express their internal beliefs as numerical probabilities. Wallsten et al. (1993) reported that a majority of respondents prefer to communicate their own beliefs verbally and to receive the beliefs of others in the form of numerical probabilities. This asymmetry is intriguing but only marginally relevant to the design of expectations questions. The relevant question is not what communication mode respondents prefer to use, but rather what modes they are willing and able to use. Wallsten et al. (1993) reported that virtually all their respondents were willing to communicate their beliefs numerically, should the situation warrant it.

A related theme of cognitive research has been to learn how individuals associate numerical probabilities and verbal phrases such as 'very likely,' 'probable,' or 'rarely' (e.g., Wallsten et al., 1986). A central finding is that individuals vary considerably in the numerical probabilities that they associate with given verbal phrases. There is, however, controversy about the proper conclusion to be drawn. Should one conclude that individuals differ in their interpretation of verbal phrases, in their interpretation of numerical probabilities, or in their interpretation of both? The empirical findings are compatible with all these conclusions.

Frontiers of Research

Empirical research measuring expectations has made substantial progress and has informed social science on many important issues. Two frontiers of research warrant particular attention.

First, future research should do more to understand how persons form their expectations. Attitudinal researchers have given little systematic thought to this basic matter. Researchers who conceptualize expectations in probabilistic terms have not been able to agree on how persons update their subjective probabilities in light of new evidence. Economists and statistical decision analysts have commonly assumed that persons apply the Bayes rule to update their expectations. Psychologists have reported empirical evidence to the contrary, but the practical import of this evidence remains unsettled. Many social scientists hypothesize that persons form expectations by observation of the behavior and experiences of others in their environments. However, we know precious little about the nature of social learning processes.

Second, research on subjective expectations should aim at understanding how a person's perception of the future affects behavior. Economists suppose that youth make schooling choices based on their expectations of the returns to schooling relative to labor market activity. They also hypothesize that uncertainty about future income induces individuals to increase savings, thereby decreasing consumption. Criminologists suggest that perceived risks of crime victimization may cause individuals to take precautions in their daily activities. Public health researchers speculate on how perceptions of health risks affect smoking, drug use, and unprotected sexual behavior. To shed light on these and similar questions of practical import requires intensive collection and analysis of expectations and behavioral data.

The research of Delavande (2008) demonstrates what may be accomplished by using expectations data when studying behavior. Delavande studied women's choice of contraception methods. She assumed that women base their decisions on their beliefs regarding method-related outcomes, including pregnancy and contraction of a sexually transmitted disease (STD). To learn about these beliefs, she surveyed women and asked them to state their beliefs about the chance of becoming pregnant or contracting an STD if they were to use alternative contraceptive methods. She then combined the measured expectations with data on contraceptive behavior to estimate an econometric model of choice of contraceptive method. The availability of expectations data made it possible for her to identify preferences more persuasively than if she had made unfounded assumptions about expectations.

Bibliography

Delavande, A., 2008. Pill, patch, or shot? Subjective expectations and birth control choice. International Economic Review 49, 999–1042.

Federal Reserve Consultant Committee on Consumer Survey Statistics, 1955. Smithies committee report. In: Reports of the Federal Reserve Consultant Committees on Economic Statistics. Hearings of the Subcommittee on Economic Statistics of the Joint Committee on the Economic Report, 84th U.S. Congress.

Hurd, M., 2009. Subjective probabilities in household surveys. Annual Review of Economics 1, 543–564.

Juster, F.T., 1964. Anticipations and Purchases. Princeton University Press, Princeton, NJ.

Juster, F.T., 1966. Consumer buying intentions and purchase probability: an experiment in survey design. Journal of the American Statistical Association 61, 658–696.

Lichtenstein, S., Fischoff, B., Phillips, L., 1982. Calibration of probabilities: the state of the art to 1980. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Manski, C., 1990. The use of intentions data to predict behavior: a best case analysis. Journal of the American Statistical Association 85, 934–940.

Manski, C., 2004. Measuring expectations. Econometrica 72, 1329–1376.

Manski, C., Molinari, F., 2010. Rounding probabilistic expectations in surveys. Journal of Business and Economic Statistics 28, 219–231.

National Bureau of Economic Research, 1960. The Quality and Economic Significance of Anticipations Data. Princeton University Press, Princeton, NJ.

Savage, L., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783–801.

Turner, C., Martin, E. (Eds.), 1984. Surveying Subjective Phenomena. Russell Sage Foundation, New York.

Wallsten, T., Budescu, D., Rapoport, A., Zwick, R., Forsyth, B., 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General 115, 348–365.

Wallsten, T., Budescu, D., Zwick, R., Kemp, S., 1993. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. Bulletin of the Psychonomic Society 31, 135–138.

Quetelet, Adolphe (1796–1874)

Ron Lesthaeghe, Free University, Brussels, Belgium

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 18, pp. 12673–12676, © 2001, Elsevier Ltd.

Abstract

This article gives a short biography of the Belgian mathematician, astronomer, and statistician Lambert-Adolphe-Jacques Quetelet and highlights his contributions to both theoretical and empirical foundations of a variety of social sciences (particularly sociology, demography, social statistics, and psychology). A bibliography is attached of selected works by Quetelet and of critical evaluations of his contribution. The article should also be of interest to historians of science.

Life and Career

Born in Ghent (Belgium) to a French father and Flemish mother, the young Quetelet started his career as a teacher, and besides his interest in mathematics, he also had artistic talents.

Quetelet was the first student to obtain a doctorate in 1819 at the newly founded State University of Ghent. In his dissertation he presented a new geometric function called the focal curve. In 1819 he was appointed mathematics professor at the Atheneum in Brussels and in 1820 he was elected to the *Académie royale des sciences et des belles-lettres.* The academy became the central place from which Quetelet directed all his other activities. In 1823 his scientific orientation took a decisive turn: he proposed the foundation of a new observatory to the Dutch government, and in the same year he was sent to Paris for further preparation of this plan. At the Paris observatory, contacts were established with Laplace, Poisson, and Fourier, and Quetelet's interests in probability theory and astronomy stem from this short period in France. The construction of the Brussels observatory was completed only in 1832, but in the meantime Quetelet had also become involved in the planning of the 1830 population census.

Unfortunately this census was never completed as a result of the Belgium's secession from The Netherlands in 1830. But, elected as secretary for life at the Academy in 1834, Quetelet remained in control not only of the activities in meteorology, astronomy, and physical geography located at the observatory, but also of the activities in demography and statistics at the *Bureau de Statistique* and, from 1841 on, at the *Commission centrale de Statistique.* The procedures to be used in the new census were tried out first in Brussels in 1842, and the nationwide census took place in October 1846. Both the method of census taking and the subsequent analyses with detailed breakdowns by age and marital status set new standards for similar undertakings in most other European countries. In the meantime Quetelet had gained an international reputation, first by the publication of his *Physique Sociale* in 1835, in which he brought together his earlier statistical work on demographic subjects, anthropometry, and criminology, and then by his relentless efforts to establish statistics as a new internationally oriented and empirically grounded scientific discipline. The latter was accomplished mainly through organizing a steady succession of international statistics conferences and by inspiring the foundation of statistical societies in other countries.

In 1855, however, Quetelet suffered a major stroke, and from then on his own scientific contribution was seriously reduced. But, until his death in 1874, he continued to inspire statistical applications in a wide variety of fields, and above all, to promote international comparability of statistical information. In Desrosieres' words (1993, p. 95), 'Quetelet was the orchestra conductor of 19th Century statistics.'

Quetelet's Contribution to Statistics and Demography

Quetelet's statistical work was profoundly influenced by early probability theory – particularly the binomial distribution of events with equal odds – and by the use that Laplace had made of the Gaussian bell curve in astronomy. From Laplace he had gained the insight that the binomial distribution could also represent measurement error in astronomical observations, and Quetelet was convinced that this would also hold for measurements in the physical, social, and 'moral' domains. In the 1820s, measurements and frequency distributions pertaining to these domains and based on large numbers were still very rare. Quetelet had to use the frequency distribution of the chest circumferences of over 5000 Scottish recruits to verify his hypothesis that these would conform to the binomial distribution (Hacking, 1990, pp. 108–111). In a series of heights of French recruits, however, he detected a deviation from the expected distribution and therefore suspected the presence of systematic errors associated with attempts to avoid conscription. This illustrates how Quetelet used theoretical distributions to assess measurement reliability.

His use of the bell curve led him also into far deeper water: he proposed the notion of *l'homme moyen* or the average man, of which all members of a given population would be imperfect copies. For Quetelet, and for many after him, the emergence of a bell curve is a strong indication that there must be a link between the individuals producing this particular frequency distribution. Individuals would be drawn as balls from the same urn, and populations would correspond to different urns. The average men, each reconstructed from a different population, would therefore facilitate comparisons between populations, which themselves would be sufficiently homogeneous. Quetelet's 'average man' has often been misunderstood and misused. First, Quetelet was never preoccupied exclusively by averages. Whenever possible he presented

complete distributions, and one of his contributions to demography is precisely his systematic presentation of *age-specific distributions* of vital events or of other occurrences (e.g., crime). This is the beginning of the statistical study of the life cycle. Second, Quetelet would be quite shocked to see his average man being associated with national prejudice, e.g., the stingy Scot, the thrifty Frenchman,…, for the simple reason that these would all be *unmeasured* attributes. In fact, he perceived his average man as an antidote against statements based on prejudice, anecdote, and impression. It is for this reason that Quetelet can be considered as one of the founding fathers of modern empirical sociology: adequate statistical measurement presupposes operationalization with satisfactory reliability and validity.

The most fundamental critique of the average man came from A. Lexis who discovered that the variances of observed distributions are generally too large to uphold the thesis of balls drawn from the same urn. The hypothesis of a homogeneous population ceased to be tenable, and this would have major consequences for the advancement of statistics and for theories in the biological and social sciences (cf. Desrosiéres, 1993: 111ff). In present day multivariate analyses we now routinely calculate 'little average men' in terms of subgroup means, odds ratios for different subpopulations at risk, and for different combinations of categories of co-variates. Furthermore, these numerous 'little average men' are compared and the results are tested for the presence or absence of significant differences.

The use of the bell curve was taken one step further by Quetelet's younger colleague and professor of mathematics at his alma mater in Ghent. PierreFrançois Verhulst (1804–49) imagined that the growth rate r of a population would evolve according to a normal distribution. The size of a population, or of any other stock, would then follow a logistic curve, i.e. the elongated S-curve, with 50 % of the total growth corresponding to the maximum of r, and then leveling off toward a saturation plateau. The logistic curve proved to have many applications: later demographic transitions produced population evolutions that very closely resemble the logistic curve, and diffusion processes modeled along the principle of contagion (of rumors, knowledge, disease, technology, etc.) or of competition commonly led to growth curves in accordance with Verhulst's logistic.

It is rather ironic that Verhulst's view of this particular form of population change did not feed back into Quetelet's own demographic work. Quetelet continued to think within the framework of a stationary population. Like several observers before him (e.g., Vauban, Sussmilch), Quetelet had been impressed by the observation that both numbers and distributions of vital events (deaths, births, marriages, ages at marriage, and age differences between spouses) showed a remarkable stability over time. His early work with Smits on the demography of the low countries had convinced him even more. Only major disturbances, such as a revolution, were capable of producing a temporary distortion (cf. his *causes constantes* et *causes accidentelles* in the *Physique sociale*). Quetelet kept thinking in terms of a homeostatic model, in the same way as Malthus had before him.

In his construction of the Belgian life table of 1841–50, centered around the census of 1846 and in his 'population tables' (population by age, sex, and marital status simultaneously) of 1850, Quetelet explicitly discusses the properties of a stationary population and shows that the actual age composition ought to be the same as the l_x or ${}_nL_x$ functions of the life table (l_x = number of survivors at each exact age x; ${}_nL_x$ = number of person-years lived in the age interval x to $x + n$). Furthermore, Quetelet goes on to show that the hypothesis of constant mortality could be relaxed (1835, p. 310; 1850, p. 16): 'the necessary condition for deducing a population table (i.e., an age structure) from a life table is that the deaths by age annually preserve the same ratios [in French: *rapports*] between them' (present author's translation and clarifications).

Quetelet was clearly on the way to show that there exists a neutral mortality decline which does not affect the shape of the age distribution (for the proof see A. J. Coale, 1972, pp. 33–36). What Quetelet never developed – probably because there was not yet a need for it in pre-transition Belgium – was the model of a stable population with a growth rate different from zero. Yet, his demographic work shows that he can be considered as an early founding father of the branch of contemporary demography called 'indirect estimation.' In this branch extensive use is made of population models (stationary, stable, quasistable), of model life tables, and of the mathematical properties of particular model distributions for detecting and correcting errors in recorded data, and especially for estimating basic demographic parameters of mortality, fertility, and nuptiality on the basis of fragmentary information. Modern historical demography and the demography of developing countries progressed rapidly thanks to these techniques and ways of thinking.

Quetelet's Contribution to Sociology

In contrast to Auguste Comte, Quetelet never developed a general plan for a new social science, and by 1900 many of his ideas had already become outdated (cf. J. Lottin, 1912, pp. 391–515). He is, however, highly typical of the beginnings of sociology: there is this new entity, called 'society,' that can be studied and analyzed from the outside with objective methods. This entity follows its own 'laws' (i.e., patterning, regularities) and these can be detected via statistical methods. At first, Quetelet's sociology had a deterministic ring, and his distinction between *causes constantes* and *causes accidentelles* led straight into the debate pitting collective determinism against individual freedom of action. Quetelet is clearly more impressed by the *causes constantes*, but in various places in his work he also proposed a solution to the conundrum. In an early letter to Villerme of 1832, for instance, Quetelet writes the following: "As a member of society [in French: *corps social*], the individual experiences at each moment the social requirements [la *necessite des causes*] and pays them regular tribute; but as an individual he masters these influences [*causes*] to some extent, modifies their effect and can seek to reach a superior state" (quoted in Desrosieres, 1993, pp. 103–104; present author's translation). In other words, through socialization the actor becomes an integral part of society, but individual and institutional agency (for Quetelet: the advancement of the sciences) play decisive roles in social change.

For many of his contemporaries the deterministic ring of the *Physique Sociale* of 1835 continued to be a source of steady criticism, particularly when Quetelet went on with his assertion that the *causes accidentelles* merely compensate each other. Later on, however, Emile Durkheim viewed the social influence on individuals and the feedback stemming from individual and collective action in very much the same way as Quetelet. Durkheim's equally holistic view of society, as in his *Formes élémentaires de la vie religieuse,* was directly influenced by Quetelet's work.

In actual practice, however, Quetelet remains essentially a master of 'comparative statics' rather than of 'social dynamics.' He paid little attention to the effects of early industrialization (definitely a *cause constante*) or to the impact of the economic, social, and health crises of the 1840s. This is very surprising for a chief statistician and demographer. The social reform movements, the rise in real wages, and the decline in adult mortality in Belgium since the 1860s also remain without comment. His successors in sociology were much more alert to analyzing the dynamics of the major societal transformations they were witnessing.

Contributions to Psychology

Quetelet's 1835 work on the development of human faculties has also had a major impact on the field of developmental psychology (Baltes, 1983; Lindenberger and Baltes, 1999). His work is the first effort to consider, based on empirical evidence, the nature of intellectual development from childhood into old age.

Treating human-ontogenetic development as a lifelong process is a fairly recent advent in psychology. Researchers in this field have been drawn to Quetelet's early efforts in this regard. They have also emphasized that Quetelet in his discussion of methodological issues was able to prefigure some of the issues that are of high significance for modern researchers. Among these are the role of period effects and selective survival.

See also: Censuses: History and Methods; Demography, Early History of; Distributions, Statistical: Approximations; Distributions, Statistical: Special and Continuous; Distributions, Statistical: Special and Discrete; Elicitation of Probabilities and Probability Distributions; Individuals in Society: History of the Concept; Life Table; Measurement Theory: History and Philosophy; Population Size Estimation and Capture–Recapture Methods; Simultaneous Equation Estimates (Exact and Approximate), Distribution of; Statistics, History of.

Bibliography

Académie Royale de Belgique, 1997. Actualité et universalité de la pensée scientifique d'Adolphe Quetelet. Classe des sciences, actes du colloque, Brussels, Belgium. 24–25.10.96.

Baltes, P.B., 1983. Life-span developmental psychology: Observations on history and theory revisited. In: Lerner, R.M. (Ed.), Developmental Psychology: Historical and Philosophical Perspectives. Erlbaum, Hillsdale, NJ, pp. 79–111.

Coale, A.J., 1972. The Growth and Structure of Human Populations—A Mathematical Investigation. Princeton University Press, Princeton, NJ.

Desrosiéres, A., 1993. La politique des grands nombres—Histoire de la raison statistique. Editions la Découverte, Paris.

Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.

Lindenberger, U., Baltes, P.B., 1999. Die Entwicklungspsychologie der Lebensspanne (Lifespan-Psychologie): Johann Nicolaus Tetens (1736–1807) zu Ehren. Zeitschrift für Psychologie 207, 299–323.

Lottin, J., 1912. Quetelet, statisticien et sociologue. Institut Supérieur de Philosophie, Louvain, France.

Mailly, E., 1875. Essai sur la vie et les ouvrages de L-A-J. Quetelet. Annuaire de l'Académie Royale de Belgique 41, 5–191.

Porter, T.M., 1986. The Rise of Statistical Thinking. Princeton University Press, Princeton, NJ.

Quetelet, A., 1832. Sur la possibilité de mesurer l'influence des causes qui modifient les éléments sociaux—Lettre à M. De Villermé. Correspondance Mathématique et Publique 7, 321–346.

Quetelet, A., 1833. Recherches sur le penchant au crimes aux différents âges. M. Hayez, Brussels, Belgium.

Quetelet, A., 1835. Sur l'homme et le développement de ses facultés, ou essai de physique sociale. Editions Bachelier, Paris (2nd edn., 1869, eds C. Muquart, Brussels).

Quetelet, A., 1847. De l'influence du libre arbitre de l'homme sur les faits sociaux, et particuliérement sur le nombre des marriages. Bulletin de la Commission Centrale de Statistique Brussels 2, 135–155.

Quetelet, A., 1846. Lettres à S.A.R. le Duc Regnant de Saxe Cobourg et Gotha, sur la theorié des probabilités, appliquée aux sciences morales et politiques. M. Hayez, Academie Royale, Brussels, Belgium.

Quetelet, A., 1848. Du système social et des lois qui le régissent. Guillaumin et Cie, Paris.

Quetelet, A., 1849. Nouvelles tables de mortalité pour la Belgique. Bulletin de la Commission Centrale de Statistique, Brussels 4, 1–22.

Quetelet, A., 1850. Nouvelles tables de population pour la Belgique. Bulletin de la Commission Centrale de Statistique, Brussels 4, 1–24.

Stigler, S.M., 1986. The History of Statistics—The Measurement of Uncertainty Before 1900. Belknap Press of Harvard University Press, Cambridge, MA.

Queues

Carmen Armero and Susie Bayarri, Universitat de València, Spain

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 18, pp. 12676–12680, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Queueing systems are simplified mathematical models to explain congestion. Broadly speaking, a queueing system occurs any time 'customers' demand 'service' from some facility; usually both the arrival of the customers and the service times are assumed to be random. If all of the 'servers' are busy when new customers arrive, these will generally wait in line for the next available server. Simple queueing systems are defined by specifying the following (a) the arrival pattern, (b) the service mechanism, and (c) queue discipline. From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. However, in all but the simplest queues, determination of the state probabilities is extremely difficult. Often, however, it is possible to determine their large time limit, the so-called equilibrium or steady-state distribution. This distribution does not depend on the initial conditions of the system and it is stationary. The ergodic conditions give the restrictions on the parameters under which the system will eventually reach the equilibrium. For the most part, queueing theory deals with computations of the steady-state probabilities and their use in computing other (steady-state) measures of performance of the queue. When only the expected values are required, an extremely useful formula for systems in equilibrium is Little's law. Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties; that is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. Inference in queueing systems is not easy: development of the necessary sampling distributions can be very involved and often the analysis is restricted to asymptotic results. The statistical analysis is simpler if approached from the Bayesian perspective.

Since Bayesian analyses are insensitive to (noninformative) stopping rules, all that it is required from the data is a likelihood function, which combined with the prior distribution on the parameters produces the posterior distribution from which inferences are derived. This is an important simplification in the analysis of queues where there are a variety of possible ways of observing the system, many providing proportional likelihood functions but very different sampling distributions. The prior distribution quantifies whatever is known about the system before the data is collected. Usually, there is plenty of information a priori about the queue, especially if it is assumed to be in equilibrium. However, it is also possible to keep the parallelism with a likelihood analysis and to avoid the incorporation of further subjective inputs, by carrying out a Bayesian analysis usually called 'objective' because the prior distribution used is of the 'noninformative' or 'objective' type. From the posterior distribution, computation of estimates and standard errors is immediate. Also, probabilities of direct interest (like the probability that the ergodic condition holds) can be computed. Most importantly, restrictions in the parameter space imposed by the assumption of equilibrium are readily incorporated into the analysis. Prediction of measures of congestion of the system (number of customers waiting, time spent queuing, number of busy servers, and so on) is carried out from the corresponding predictive distributions which are also very useful for design and intervention in the system.

Queueing systems are simplified mathematical models to explain congestion. Broadly speaking, a queueing system occurs any time 'customers' demand 'service' from some facility; usually both the arrival of the customers and the service times are assumed to be random. If all of the 'servers' are busy when new customers arrive, these will generally wait in line for the next available server. Simple queueing systems are defined by specifying the following (1) the arrival pattern, (2) the service mechanism, and (3) queue discipline. From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. However, in all but the simplest queues, determination of the state probabilities is extremely difficult. Often, however, it is possible to determine their large time limit, the so-called equilibrium or steady-state distribution. This distribution does not depend on the initial conditions of the system and it is stationary. The ergodic conditions give the restrictions on the parameters under which the system will eventually reach the equilibrium. For the most part, queueing theory deals with computations of the steady-state probabilities and their use in computing other (steady-state) measures of performance of the queue. When only the expected values are required, an extremely useful formula for systems in equilibrium is Little's law. Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties; that is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. Inference in queueing systems is not easy: development of the necessary sampling distributions can be very involved and often the analysis is restricted to asymptotic results. The statistical analysis is simpler if approached from the Bayesian perspective. Since Bayesian analyses are insensitive to (noninformative) stopping rules, all that it is required from the data is a likelihood function, which combined with the prior distribution on the parameters produces the posterior distribution from which inferences are derived. This is an important simplification in

the analysis of queues where there are a variety of possible ways of observing the system, many providing proportional likelihood functions but very different sampling distributions. The prior distribution quantifies whatever is known about the system before the data is collected. Usually, there is plenty of information a priori about the queue, especially if it is assumed to be in equilibrium. However, it is also possible to keep the parallelism with a likelihood analysis and to avoid the incorporation of further subjective inputs, by carrying out a Bayesian analysis usually called 'objective' because the prior distribution used is of the 'non-informative' or 'objective' type. From the posterior distribution, computation of estimates and standard errors is immediate. Also, probabilities of direct interest (like the probability that the ergodic condition holds) can be computed. Most importantly, restrictions in the parameter space imposed by the assumption of equilibrium are readily incorporated into the analysis. Prediction of measures of congestion of the system (number of customers waiting, time spent queuing, number of busy servers, and to on) is carried out from the corresponding predictive distributions which are also very useful for design and intervention in the system.

The intuitive notion that everybody has of the so familiar 'queues' or waiting lines closely parallels its mathematical meaning. Queueing systems are simplified mathematical models to explain congestion. This is a field of impressive activity and growth, which has mainly been concerned with the building of the models and the study of its probability properties. Relatively little effort has been put into their statistical analysis. In this article, the main concepts underlying simple queuing systems are discussed; some Bayesian inferential tools are also given.

Queueing Systems

Broadly speaking, a queueing system occurs any time 'customers' demand 'service' from some facility; usually both the arrival of the customers and the service times are assumed to be random. If all of the 'servers' are busy when new customers arrive, these will generally wait in line for the next available server. In some examples, the terms 'customer,' 'service' and 'server' have the usual, everyday meaning. For instance, queueing systems occur when customers check out in a grocery store, or require the service of a teller in a bank; here, the queueing term 'customer' refers to the actual customers at the store or the bank, the required service is to be checked out or to resolve the bank inquiry they have, and if the server (the check-out person or the teller) is busy when a customer require his/her service, then the customer has to wait in line, that is, queue up for service. In general, the terms 'customers,' 'service,' and 'servers' are very broad and can refer to almost anything; for instance, telephone calls ('customers') arriving at a switchboard (the service time being the time each call needs to be completed), machine failures (the 'customer' is the machine, and the server the repair person), demands of CPU time (server is the CPU, and the service the completion of the task), etc. Sometimes, a 'waiting line' (real or conceptual) never occurs. For instance, buses following their route can be considered a queueing system in which the customers are the buses, and the service is the completion of the route; here, customers are immediately 'served' and no waiting is necessary (technically, infinite servers, or as many servers as the number of buses, are assumed). In many interesting applications, customers require more than one service from different servers, and upon completion of service in one node, they have to proceed to another node. If the new server is busy, the customer has to queue up again for service. This is a typical situation modeled by the so called networks of queues, which are not treated in this article. An example occurs when patients enter an emergency department in a hospital: they might have to 'queue up' for primary attention, X-rays, lab service, surgery, and so on.

Queues (and queueing networks) have become very useful tools for evaluating (at least approximately) the performance of complex interacting stochastic systems in which congestion occurs. Obvious applications occur in airport terminals, health-care centers, and check-out lines at different counters. They have also been traditionally applied to biology (migration, population models), electrical engineering (flow models) and chemistry (polymerization, clustering) among others. The most prevalent applications of queueing systems and networks occur in the fields of manufacturing, computer networking, and telecommunications and broadcasting. But there are also many applications in the social sciences:

1. Several types of airport queues: aircrafts waiting to take off or to land, queues of passengers at terminals (they have to go through ticket counters, check-in counters, security and passport controls, baggage carousels, customs, etc.).
2. Police, fire, and emergency medical systems can all be regarded as queueing systems. Here the main concern is with responding to the calls for assistance as quickly as possible, important questions being to decide the number of units and which unit or units to dispatch to a specific call.
3. Health-care systems provide many examples of analyses well suited for queueing systems, as modeling the movement of coronary patients in a hospital, determining the staffing of a department, modeling waiting queues for transplants, etc.
4. In justice scenarios, cases queue up for trial and the length of the trial depends on the actions of both the 'customer' (that is, the lawyers), and the judge.
5. Analyzing and evaluating the delays in the mail, studying queues at voting polls to optimize the number (and type) of voting machines, and so on.
6. Latent phase of diseases (such as Aids) in which the times of infection are arrivals to the queuing system, the incubation is the 'service,' and the onset (or diagnosis) of the disease is the departure from the queuing system. Questions of interest are the number of people in the system and the time that 'customers' stay in the system (that is, from infection to onset or diagnosis).

Since the economical and social impact of congestion can be considerable, the importance of analyzing queueing systems has been clear ever since Erlang's pioneering work on telecommunications at the beginning of the twentieth century. For mathematicians, queues offer an added bonus, namely that the mathematics involved can be complex and intriguing; they also have an enormous potential for varying the conditions and the assumptions of the system. It does not come as a surprise that queues developed quickly into the huge area of study that it is

today. Thousands of papers, and many journals, meetings, and books are entirely devoted to the subject. It is impossible to list here even the key references, and only a small selection will be given, from which other references can be obtained. Among the numerous books, it is worth mentioning the general books of Cox and Smith (1961), Takács (1962), Cooper (1981), Gross and Harris (1985), Medhi (1991), and Nelson (1995); Kleinrock (1975, 1976) and Newell (1982) are good sources of applications; Cooper (1990) is a nice survey with an extensive bibliography. Very useful reviews of many different aspects of queueing theory (including history, statistical analysis, simulation methods, applications, etc.) can be found in Dshalalow (1995, 1997).

Simple queueing systems are defined by specifying the following:

1. *The arrival pattern*, which among other things can be deterministic or stochastic, individual or in batches (like travelers arriving at an airport passport control), etc. Also customers can abandon the line before being served, or can interact in some other ways with the rest of the customers.
2. *The service mechanism*, which describes the number of servers, how many customers at the time the server can handle (usually one, but bulk servers such as buses, elevators, etc. can serve more than one), and the distribution of service times (which can be independent among servers, time-dependent, dependent on queue length, etc.).
3. *Queue discipline*, usually FIFO (first-in-first-out), but others are possible, such as last-in-first-out (LIFO) (as the subroutines in a computer program), priority customers (as express check-out lines, giving priority to shorter service times), customers switching lines, etc. Besides these, other aspects can also be relevant, such as the capacity of the waiting room, or the size of the population of customers.

In the early 1950s, Kendall introduced a useful shorthand notation for simple queueing systems. It consists on a series of symbols separated by slashes, as $A/B/c$. 'A' refers to the distribution of the inter-arrival times between customers; thus Er stands for Erlang, G for General, M for Exponential (Markovian), etc. 'B' characterizes the service distribution for each server, with the same symbols as before, and 'c' refers to the number of servers, who are usually assumed to be equally efficient. More symbols are added when needed, to impose restrictions on the system capacity, and/or on the size of the population that feeds the system, and/or to specify the queue discipline if other than FIFO. One of the simplest and most studied queues is the $M/M/c$ queue, especially the $M/M/1$ and $M/M/\infty$; the latter are particularly useful for modeling self-service facilities or as approximations to systems with a very large number of servers (such as emergency services).

Equilibrium: Measures of Performance

From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. Let $N(t)$ denote the number of customers in the system at time $t \geq 0$. The *state probabilities* are defined as $P_j(t|\theta) = Pr(N(t) = j|\theta)$, for $j = 0, 1, 2, \ldots$, where θ denotes all parameters governing the arrival and service processes. The derivation of $P_j(t|\theta)$ is crucial in the probabilistic analysis of queueing systems. For many simple systems, the stochastic process $\{N(t),\ t \geq 0\}$ is a birth–death process (a 'birth' occurring with every arrival to the system, and a 'death' with every departure), greatly simplifying their study. (A brief introduction to birth–death, and other, stochastic processes with their connections to queues can be found, for example, in Ross (1993), as well as in the general references for queues given in Section Queueing Systems). More complex systems sometimes resort to studying other *embedded* stochastic (hopefully Markovian) processes by contemplating the system as it would be seen by an arriving customer or by a departing customer.

In all but the simplest queues, determination of these probabilities is, unfortunately, extremely difficult. Often, however, it is possible to determine their large time limit, the so-called *equilibrium* or *steady-state* distribution

$$P_j^{(\theta)} = \lim_{t \to \infty} P_j(t|\theta)$$

without actually having to compute $P_j(t|\theta)$. The equilibrium distribution does not depend on the initial conditions of the system and it is stationary (does not depend on t once the statistical equilibrium is reached). The ergodic conditions give the restrictions on θ under which the system will eventually reach the equilibrium (with probability one). Thus, for instance, in an $M/M/1$ queue, the ergodic condition establishes that the arrival rate (λ) should be smaller than the service rate (μ) (that is, the average number of customers arriving in a unit time is smaller than the average number that the server can handle, a most intuitive condition).

For the most part, queueing theory deals with computations of the steady-state probabilities $P_j(\theta)$, and their use in computing other (steady-state) measures of performance of the queue, such as the expected number of customers in the system and/or in the waiting line, the distribution of the time spent queueing or in the system or their expected values, or the mean (or distribution) of the time the servers are busy (or idle). These steady-state distributions are also often very difficult to compute, and many require solving systems of differential equations, inversion of transforms (generating probability functions, Laplace transforms,…), etc. When only the expected values are required, an extremely useful formula for systems in equilibrium is *Little's law* which establishes that the expected number of customers in the system (L) is equal to the average time that a customer spends in the system (W) times the rate (λ) at which customers enter the system.

Statistical Analysis: A Bayesian Approach

Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties; that is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. A good review is Bhat et al. (1996). It could be argued that the general results of the (well developed) area of inference for stochastic processes (see for instance,

Basawa and Prakasa Rao, 1980; Lehoczky, 1990) could be applied to queues. These general results do not take advantage of the special structure of queueing systems, however, nor do they address its specific questions.

Inference in queueing systems is not easy: From a frequentist perspective, development of the needed sampling distributions can be very involved and often the analysis is restricted to asymptotic results. Also, estimates can be different, for example, if the observational process differ, if the initial queue length is or is not included in the analysis. The statistical analysis is simpler if approached from the Bayesian perspective, and we take this approach below (*see* Bayesian Statistics). More arguments in favor of Bayesian analyses of queueing systems are given in Armero and Bayarri (1999), where a comprehensive list of references with Bayesian analyses of queues is also given. As an illustration of the main issues, the analysis of the much studied *M/M/1* queue is outlined here.

According to the assumptions of the *M/M/1* queue, the interarrival times, Y, are i.i.d. exponential with arrival rate λ, that is

$$f(y/\lambda) = \lambda e^{-\lambda y}, \quad y > 0. \qquad [1]$$

Similarly, the service times X are also i.i.d., $X \sim \mathrm{Ex}(\mu)$, independently of Y, where μ is the service rate. That is

$$f(x/\mu) = \mu e^{-\lambda x}, \quad x > 0. \qquad [2]$$

Steady-state distributions exist if and only if the ergodic condition

$$\rho = \lambda/\mu < 1 \qquad [3]$$

holds, where ρ is the called traffic intensity and plays an important role in the study of $M/M/1$ queues.

A queue can be observed in a variety of ways and different quantities can be recorded. The most traditional experiments (providing complete information) in queues observe the system over a continuous period of time $(0, T]$. T can either be a value fixed in advance, or be determined by a suitable stopping rule. For instance, the system can be observed until the busy time reaches some preassigned fixed value, until a fixed number of customers have departed from the system, until a fixed number of transitions have been recorded, and so on (references can be found in Armero and Bayarri, 1997). Different quantities can be recorded (instants of arrivals, of departures, of commencing of service, number of arrivals and/or departures, interarrival times, service times, number of transitions, busy times, etc). If the queue is assumed to be in equilibrium, the observation of the number of customers at time 0 can also be added, providing more informative but rather cumbersome likelihood functions.

Suppose that an experiment providing n_a interarrival times, $Y_1,\ldots,Y_{na}$ and n_s service completions, $X_1, \ldots, X_{ns}$ has been performed. Then, from [1] and [2], a likelihood function for (λ, μ) based on the observed $\mathbf{x} = (x_1, x_2,\ldots, x_{na})$ and $\mathbf{y} = (y_1, y_2,\ldots, y_{na})$ is

$$\mathrm{L}(\lambda,\mu)\alpha f(\mathrm{x},\mathrm{y}|\lambda,\mu)\alpha\lambda^n a\mu^n s e^{-\mu t_s} \qquad [4]$$

where $t_a = \sum_{i=1}^{n_a} y_i, \quad t_s = \sum_{j=1}^{n_s} x_j$ are the totals of observed interarrival times and service times, respectively. Notice that this likelihood is valid whether n_a and/or n_s are fixed or random (as long as the likelihood function is proportional to [4]). Also, the observation of the arrival and the service processes does not need to be simultaneous and the X's and Y's does not need to correspond to successive customers. (Of course, many of the observational procedures mentioned above result in likelihood functions proportional to [4].) Notice also that [4] applies whether or not the system is in equilibrium and whether or not equilibrium can eventually be reached.

Maximum likelihood estimators are immediate from [4], and given by.

$$\widehat{\lambda} = \frac{n_a}{t_a}, \widehat{\mu} = \frac{n_s}{t_s}, \widehat{p} = \frac{\widehat{\lambda}}{\widehat{\mu}} \qquad [5]$$

However, standard errors of confidence intervals do require careful consideration of the specific stopping rule used, and different experimental designs will in general produce different sampling distributions for these estimators, and hence different standard errors and confidence intervals.

Bayesian analyses are insensitive to (noninformative) stopping rules, and all that it is required from the data is the likelihood function [4]. A prior distribution has also to be assessed quantifying whatever is known about the system before the data is collected. Usually, there is plenty of information a priori about the queue, especially if it is assumed to be in equilibrium (which implies, at the very least, that the system has been working for an extended period of time). A prior distribution incorporating all this information should then be assessed. However, to keep the parallelism with the likelihood analysis and also to avoid the extra work of quantifying subjective beliefs, we use a Bayesian analysis usually called 'objective' because the prior distribution used is of the 'non-informative' or 'objective' type. Since both λ and μ are scale parameters, the invariant measures for each will be used, and they will be assumed independent a priori, that is.

$$\pi(\lambda,\mu)\alpha\frac{1}{\lambda}\ \frac{1}{\mu} \qquad [6]$$

Note that no assumptions about the stationarity of the queue are made in [6]. In fact, they are not needed to produce posterior distributions, although they could be incorporated if desired (see [9] below). If it is wished to incorporate prior information about the values of λ and μ, a very simple prior distribution to use is the conjugate $\pi(\lambda, \mu)\alpha Ga\ (\lambda|a, b)\ Ga\ (\mu|c, d)$, a product of two independent gamma distributions, with the hyperparameters a, b, c, d chosen to reflect the prior information. (Again, this prior makes no assumptions about stationarity). In fact, the prior [6] can be seen as a lmiting case of this conjugate prior when all hyperparameters tend to 0 ('vague' prior information).

Combining [4] with [6] with Bayes theorem, we get the joint posterior distribution

$$\pi(\lambda,\mu/x,y)\alpha\pi(\lambda,\mu)f(\mathrm{x},\mathrm{y}/\lambda,\mu)\alpha Ga(\lambda/n_a,t_a)Ga(\mu/n_s,t_s) \qquad [7]$$

so that λ and μ are independent *a posteriori* with scaled chi-squared distributions $2t_\alpha\lambda \sim \chi^2_{(2n_a)}, \quad 2t_s\mu \sim \chi^2_{(2n_s)}$, and since they are independent, $(t_a/t_s)\rho \sim F_{2na,2n}$, an F distribution. These distributions will be *exactly* valid for all experiments

producing the likelihood function [4], irrespective of the stopping rule. Usual estimates of λ, μ and ρ are the means (or the modes) of their (marginal) posterior distributions, and 'standard errors' associated to these estimates are most naturally given by the standard deviations of those distributions. Intervals with a given probability content can similarly be given. All these will numerically reproduce their non-Bayesian counterparts when the design has n_a and n_s fixed in advance.

An important issue when studying queues is how likely it is that steady-state can be reached. From the marginal distribution of ρ, this and other questions of direct interest can be answered. In particular, the probability that the ergodic condition holds is computed from the F distribution above as

$$\pi_e = \Pr(\rho < 1/\mathrm{x}, \mathrm{y}) = \int_0^1 \pi(\rho|\mathbf{x}, \mathbf{y})\mathrm{dp} = \Psi_{(2n_a, 2n_s)}(t_a/t_s) \quad [8]$$

where $\Psi_{(2n_a,2n_s)}$ is the distribution function of an F_{2na}, $2n_s$ distribution.

Most often, queues are assumed to be in equilibrium. Under this assumption, the restriction $\rho < 1$ imposed by the ergodic conditions should be incorporated into the analysis. This is easily done, and in fact amounts to truncating the unrestricted posterior distribution for ρ giving

$$\pi(\rho|\mathbf{x}, \mathbf{y},\ \rho < 1) = \frac{\pi(\rho|\mathbf{x}, \mathbf{y})}{\pi_e} \quad \text{for } \rho < 1 \quad [9]$$

and $\pi(\rho|\mathbf{x},\ \mathbf{y},\ \rho < 1) = 0$ for $\rho \geq 1$. From [9], estimates, standard errors, credible intervals, etc. that take into account the restriction can be produced. (In this simple case, the probability π_e in [8], the proportionality constant in [9] and the moments of the truncated posterior can be given close form expressions in terms of Gauss' hypergeometric functions.) In Armero and Bayarri (1999), an explicit expression for a factor that can be used for adjusting the usual MLE ρ is given, as well as (adjusted) 'standard errors.' It can there be seen that incorporating the restriction $\rho < 1$ has the effect of 'shrinking' the usual MLE, the shrinkage being larger the larger ρ and the smaller the sample sizes, a most intuitive result.

When statistically analyzing queueing systems, inference about the parameters is rarely (if ever) of direct interest. Usually the main goal is prediction of observable quantities that describe the utilization and degree of congestion of the system. From the Bayesian perspective, it is also very easy to produce predictive distributions incorporating the uncertainty in the estimation of the unknown parameters. The predictive distribution of the steady-state number, N, of customers in the system is computed as

$$\rho(N = n/\mathbf{x}, \mathbf{y},\ \rho < 1) = \int_0^1 \Pr(N = n|\rho)\pi(\rho|\mathbf{x}, \mathbf{y},\ \rho < 1)\mathrm{dp} \quad [10]$$

where $Pr(N = n|\rho)$ is given by the Geometric $Ge(1 - \rho)$ distribution and $\pi(\rho|\mathbf{x},\ \mathbf{y},\ \rho < 1)$ is given in [9]. In this case, again, [10] can be given a close form expression (see Armero and Bayarri, 1994; for details).

Similar arguments apply for prediction of the number of the (steady-state) number of customers in the waiting line, time spent by a customer in the system (or waiting in line), length of busy (idle) periods, etc. All these predictive distributions are a most natural tool for deciding future interventions in the queue (see Armero and Bayarri, 1997).

Conclusion

Queues systems are useful to stochastically model systems in which subjects interact and congestion occurs. They can describe a wide variety of such situations, in virtually all areas of application. In probabilistic terms, queues have been extensively studied under the assumption that the governing parameters are known. Inferences about these parameters can be very challenging when approached from a frequentist perspective, and the common approach of simply replacing the unknown parameters by some estimates when predicting measures of performance of the queues can produce deceiving results. Bayesian methods appear to be ideally suited to handle inferential aims in queuing systems, as incorporating restrictions in the parameter space (imposed by the ergodic conditions), predicting measures of performance, designing the system, etc. Moreover, its independence of the stopping rule makes the inferential task much simpler.

Bibliography

Armero, C., Bayarri, M.J., 1994. Bayesian prediction in M/M/1 queues. Queueing Systems 15, 401–417.

Armero, C., Bayarri, M.J., 1997. Bayesian questions and answers in queues. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, UK, pp. 3–23.

Armero, C., Bayarri, M.J., 1999. Dealing with uncertainties in queues and networks of queues: a Bayesian approach. In: Ghosh, S. (Ed.), Multivariate Analysis, Design of Experiments and Survey Sampling. Marcel Dekker, New York, pp. 579–608.

Asmussen, Søren, 2003. Applied Probability and Queues. Springer, New York.

Basawa, I.V., Prakasa Rao, B.L.S., 1980. Statistical Inference for Stochastic Processes. Academic Press, New York.

Bhat, U.N., Miller, G.K., Rao, S.S., 1996. Statistical analysis of queueing systems. In: Dshalalow, J.H. (Ed.), Frontiers in Queueing: Models and Applications in Science and Engineering. CRC Press, Boca Raton, FL, pp. 351–394.

Cooper, R.B., 1981. Introduction to Queueing Theory, second ed. North-Holland, New York.

Cooper, R.B., 1990. Queueing theory. In: Heyman, D.P., Sobel, M.J. (Eds.), Handbooks in OR and MS, vol. 2. Elsevier, Amsterdam, pp. 469–518.

Cox, D.R., Smith, W.L., 1961. Queues. Wiley, London.

Dshalalow, J.H. (Ed.), 1995. Advances in Queueing: Theory, Methods, and Open Problems. CRC Press, Boca Raton, FL.

Dshalalow, J.H. (Ed.), 1997. Frontiers in Queuing. Models and Applications in Science and Engineering. CRC Press, Boca Raton, FL.

Gross, D., Harris, C.M., 1985. Fundamentals of Queueing Theory, second ed. Wiley, New York.

Kleinrock, L., 1975. Queueing Systems. In: Theory, vol. 1. Wiley, New York.

Kleinrock, L., 1976. Queueing Systems. In: Computer Applications, vol. 2. Wiley, New York.

Khinchin, Aleksandr Yakovlevich, 2013. Mathematical Methods in the Theory of Queuing. Courier Dover Publications, Mineola, New York.

Lehoczky, J., 1990. Statistical methods. In: Heyman, D.P., Sobel, M.J. (Eds.), Handbooks in OR and MS, vol. 2. Elsevier, Amsterdam, pp. 469–518.

Medhi, J., 1991. Stochastic Models in Queueing Theory. Academic Press, Boston.
Nelson, R., 1995. Probability, Stochastic Processes and Queueing Theory. Springer Verlag, New York.
Newell, G.F., 1982. Applications of Queueing Theory, second ed. Chapman and Hall, London.
Ross, S.M., 1993. Introduction to Probability Models, fifth ed. Academic Press, New York.
Takács, L., 1962. Introduction to the Theory of Queues. Oxford University Press, New York.
Whitt, Ward, 2002. Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues. Springer, New York.

R

Race and Academic Motivation

Sandra Graham, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews contemporary research on achievement motivation in racial/ethnic minority youth. The review is organized around two main constructs in motivation research: expectancy for success, or beliefs about ability (Can I do it?) and how much that success is valued or desired (Do I want it?). The topics covered under expectancy for success are racial stereotypes about intelligence, students' implicit theories about intelligence, and coping with discrimination. The topics covered under achievement values are oppositional identity, the valued characteristics of peers, self-affirmation, and the immigrant paradox. Implications for interventions to improve motivation in ethnic minority youth are discussed.

Motivation is the study of *why* people behave as they do: what energizes them to start behaving in a particular way, what directs their behavior and keeps them going, and what leads them to eventually terminate the behavior. In the achievement domain, think about a typical behavior such as a student studying for an exam, and view it as a temporal sequence that is initiated, directed, sustained, and then terminated. Motivation researchers examine what the student is doing, or behavioral *choice*; how long it takes before the student begins to study, or the *latency* of behavior; how hard the student actually studies, or behavioral *intensity*; how long the student is willing to study, or the *persistence* of behavior; and what the student is thinking and feeling during or after the study session, or the *cognitions and emotions* that accompany or follow the behavior.

There are a few dominant contemporary approaches to motivation that address one or more of these temporally ordered constructs (see review in Graham and Weiner, 2011). The most popular of these approaches fit broadly within an expectancy-value framework. Motivation (e.g., choice, latency, persistence) is determined by both the likelihood of goal attainment, which is the expectancy component, and how much that goal is desired, which is the value component. The study of expectancy addresses beliefs about ability (Can I do it?), whereas the study of values addresses preferences and desires (Do I want it?). The notions of *can* and *want* closely correspond to the distinctions between aptitude and effort expenditure, which are two of the most important determinants of the motivation to achieve.

This article on race and motivation is organized around overarching questions about *can* and *want*. In so doing, it departs from the general motivation literature in two ways. First, the starting point is the larger sociocultural context and factors that are unique to the everyday lives of people of color. Some of those factors are historical and structural in nature. Many racial and ethnic minority groups are positioned at the bottom of a status hierarchy wherein barriers to opportunity often override personal strivings for achievement. Second, race and motivation are examined from a social contextual perspective. Central to that perspective is the belief that students' social lives and their academic lives are interrelated and that one cannot fully understand an individual student's motivation to achieve without also understanding their relationships with other people. The social nature of motivation is particularly important during adolescence when peer approval takes on heightened significance.

With this framework in mind, the first part of the article addresses beliefs about ability (Can I do it?) and how these beliefs are shaped by racial stereotypes about intelligence, students' implicit theories about intelligence, and coping with discrimination. The second part of the article turns to achievement values (Do I want it?) in which the focus is on oppositional identity, the valued characteristics of peers, self-affirmation, and the immigrant paradox. Addressing these topics highlights the unique motivational challenges of racial and ethnic minority youth as they strive to master their environment.

Beliefs about Ability: Can I Do It?

Racial Stereotypes about Ability

Stereotypes are culturally shared beliefs, both positive and negative, about the characteristics and behaviors of particular groups. Even though beliefs about African-Americans have become more positive over the last 50 years, studies of cultural stereotypes continue to show that respondents associate being black (and male) with hostility, aggressiveness, violence, and danger (e.g., Correll et al., 2002). The recent acquittal in Florida of George Zimmerman for the murder of Trayvon Martin, an unarmed black youth, is a stark reminder of that fact. The much smaller stereotype literature on other ethnic groups in the United States also portrays the more marginalized groups in a negative light. For example, cultural stereotypes of Latino

youth depict them as unintelligent and with little personal ambition (e.g., Kao, 2000).

African-Americans and Stereotype Threat

Research on stereotype threat (Steele, 2010) has provided important insights into the negative consequences of racial stereotypes about intelligence. Stereotype threat is the awareness that individuals have about negative stereotypes associated with their group. Although considered to be a general psychological state applicable to any negative group stereotype, the construct originated in the achievement domain and it has been applied to African-American students' awareness of the cultural stereotype associating their race with intellectual inferiority. That awareness can be quite debilitating, especially for those African-American students who are invested in doing well in school. For example, in a series of studies with black and white students attending Stanford University, Steele and colleagues found that black students performed more poorly than whites on standardized test items when they were told that the test was diagnostic of their abilities. When told that the test was a problem-solving activity unrelated to ability, there was no difference in the performance of the two racial groups. In ability-related contexts, therefore, what became threatening for black students was the fear that they might confirm the stereotype or be judged by others based on that stereotype. Stereotype-threatened students often are dividing their attention between the task itself (e.g., taking a SAT exam) and ruminating about the meaning of their performance (e.g., what does this say about me or about members of my racial group?). Such negative ruminations have been shown to both arouse anxiety and deplete the executive resources needed to do well on academic tasks. Furthermore, it is not necessary that a student endorse the stereotype; mere awareness of its existence is sufficient to activate threat.

Stereotype threat researchers have documented two motivational consequences of the anxiety associated with thinking about race and intelligence in highly evaluative achievement contexts. Some students may choose to work especially hard as a way of disconfirming the stereotype; but high effort in the face of increasing academic challenge may be difficult to sustain and may even lead one to question their abilities. Stereotype threat can also have the opposite effect, causing students to minimize effort and downplay the importance of doing well in school. Steele coined the term academic *disidentification* to describe students who no longer view academic achievement as a domain that is either important to them or their self-definition. Disidentification has been associated with declining achievement from middle-school to high-school, particularly among African-American boys. A similar process, labeled academic *disengagement*, occurs when students begin to discount the feedback they receive about their performance or to devalue achievement altogether. While disidentification and disengagement may be self-protecting mechanisms for coping with negative racial stereotypes, in the long run their detrimental effects on motivation will outweigh any short-term self-enhancing effects.

In the last two decades, hundreds of studies on stereotype threat have been conducted and the Steele and Aronson (1995) inaugural article is considered to be a modern classic in social psychology (Devine and Brodish, 2003). However, it is also important to point out that the phenomenon is not without its critics (e.g., Sackett et al., 2004). Most of the studies are laboratory-experimental in which threat is activated by some experimental manipulation, leading to questions about ecological validity. It is also not clear that undoing stereotype threat actually reduces the racial achievement gap, which some supporters of the theory have proposed. Moreover, few studies thus far have been conducted with children in elementary and secondary schools, which raise issues about the age at which stereotype threat begins to affect motivation and performance of racial minority groups. The absence of a developmental literature is probably one of the major limitations of this literature.

Asian-Americans and the Model Minority Stereotype

There is also a literature on the motivational effects of racial stereotypes about intelligence at the other end of the ability spectrum. Unlike African-Americans, the cultural stereotype about Asian-Americans is that they are hardworking and intellectually gifted high achievers who are especially competent in math and science. The term *model minority* was coined in the 1960s to capture those characteristics and to account for the seemingly unprecedented successful entry of East-Asian immigrants into mainstream American society. Many studies have now documented that Asians and non-Asians alike are aware of the culturally shared association between high academic achievement strivings and being an Asian-American (e.g., Kao, 2000).

While it may be more tolerable to know that one's racial group is viewed as smart and hardworking rather than dumb and lazy, that stereotype also has its own set of challenges. Ethnographic, survey, and experimental research all point to psychological and emotional costs associated with living up to the model minority stereotype. Ethnographic studies, for example, detail the anxiety that many Asian-American students feel when forced to cope with the perception of their group as academic superstars. Many report feeling frustrated and pressured to attain or maintain high academic achievement because of the expectations placed upon them by their parents and teachers. Thus, while some African-American students worry that their performance might confirm a negative stereotype, some Asian-American students worry that their performance might not confirm a positive stereotype. These worries may be exacerbated among groups from Southeast Asia (e.g., Cambodians, Laos) whose immigrant experiences have been quite different from those of their East Asian and South Asian counterparts (Ngo and Lee, 2007). Southeast Asian students may be derogated by teachers and authority figures because of their lower-academic achievement yet not enjoy any of the social benefits of being perceived as a model minority.

Entity versus Incremental Theories about Intelligence

Whether individuals are vulnerable to racial stereotypes or other challenges to 'I can' is in part determined by their beliefs about the meaning of intelligence. For example, does it make a difference if students view their abilities as stable and

essentially beyond their control ("either you have it or you don't") or as changeable and largely determined by how much effort they exert? To address this question, Dweck proposed a distinction between individuals who hold entity versus incremental theories of intelligence (Dweck, 2006). Some individuals are entity theorists: they believe that intelligence is basically fixed, each person has their allotted share, and there is little they can do to increase that allotment. In contrast, other individuals appear to be incremental theorists: they believe that intelligence is modifiable through increased effort. These implicit theories have been related to a variety of achievement-related cognitions and behaviors. For example, students who subscribe to an incremental theory of intelligence prefer challenging tasks so that they can both increase their ability and achieve greater mastery. Entity theorists, on the other hand, often avoid challenge because their primary concern is with the adequacy of their presumably fixed ability and what failure might convey about that ability.

It should be evident that the negative motivational consequences of an entity theory of intelligence resemble the vulnerability to stereotype threat discussed above. Recognizing this overlap, a number of social psychologists have designed interventions for stereotype-threatened students that focus on changing beliefs about intelligence (see review in Yeager and Walton, 2011). In these interventions, ethnic minority college students and middle-school students were trained to view intelligence as malleable rather than fixed. For example, in one representative study with middle-school students, African-American and Latino seventh graders who were struggling in math were randomly assigned to an incremental theory experimental condition or to a control condition (Blackwell et al., 2007). The intervention group received eight weekly lessons on how the brain works, emphasizing its plasticity and malleability (e.g., one lesson was entitled "You Can Grow Your Intelligence"). Compared to control-group classmates, those in the treatment group showed significant improvement in math performance over a 1-year period. The greatest improvements were found for those who changed their theories of intelligence from entity to incremental beliefs, suggesting that the intervention was effective in changing a very specific kind of maladaptive motivational belief.

Coping with Discrimination

Like people who hold entity theories of intelligence, individuals of color who believe that they are treated unfairly because of their race or ethnicity can lose confidence in themselves and their willingness to persist in the face of academic challenge. Among the most prevalent kinds of unfair treatment reported by ethnic minority youth are receiving a lower grade than deserved from teachers, being the recipient of unusually harsh discipline from authority figures, being accused of behaving suspiciously in public places, and being the target of verbal, psychological, or physical abuse from peers at school (Fisher et al., 2000).

Racial discrimination can take its toll on the mental, physical, and academic well-being of youth. Such experiences have been linked to decreased perceptions of mastery among early adolescents of color and increased negative attitudes about school. Perceived discrimination can lead to mistrust of teachers and to the general belief that the school rules and policies are unfair. In an effort to offer a more nuanced approach to the experience of discrimination, recent research has also documented that different sources of unfair treatment might be related to different outcomes (Benner and Graham, 2013). For example, does it make a difference if the perpetrators are peers versus teachers? It appears that race-based unfair treatment by peers (e.g., physical bullying, ostracism) is uniquely linked to mental health difficulties, whereas perceived discrimination by teachers (e.g., discipline or assignment of grades) is more directly predictive of academic disengagement. Because Asian students are more likely to report maltreatment from peers and African-American students are more likely to perceive teacher discrimination (Rosenbloom and Way, 2004), it is important to identify the particular negative consequences of each discrimination source.

Not all ethnic minority youth succumb to the motivational challenges of perceived discrimination. There may be some circumstances under which youth transform discriminatory experiences so that they become an impetus for increased effort in school. Eccles et al. (2006) asked middle-school youth whether they believed that being African-American would make it difficult for them to be successful in their future educational and occupational pursuits. Analyses showed that youth who anticipated future race-based discrimination displayed *higher* levels of academic motivation than those who did not. Rather than focusing on day-to-day discrimination experiences, some youth may conceptualize unfair treatment as a structural barrier that can be overcome through academic excellence.

Perhaps the best antidote to the negative effects of perceived discrimination is a combination of strong racial/ethnic identity and parental socialization about race. For example, when African-American youth report that their race is central to their self-concept and that parental messages communicated both racial pride and vigilance (preparation) against racism, then the relations between experiences with discrimination and academic disengagement were weakened (e.g., Neblett et al., 2008). It appears that a moderate amount of preparation for bias and discrimination is most effective. Too little parental socialization may leave ethnic minority children adolescents vulnerable to the harmful impact of discrimination whereas too much preparation could heighten awareness of racial barriers to a level that results in the erosion of children's self-efficacy (Harris-Britt et al., 2007).

To summarize, this section of the article focused on how beliefs about ability can be challenged for ethnic minority youth. Such youth must cope with stereotypes linking race to low or high intelligence and with the widely shared perception that intelligence is a fixed entity that cannot be changed. These youth also confront discrimination, both overt and subtle, that can undermine achievement strivings. Seeing race or ethnicity as central to one's identity and receiving a moderate level of parental messages that prepare one for racism can buffer some of the negative consequences of discrimination.

Beliefs about Effort: I Can, but Do I *Want* to?

This next section focuses on the second main construct in contemporary motivation research – achievement values. Unlike expectancy, which is largely about ability beliefs, values

have to do with desires and preferences and are more concerned with the perceived importance, attractiveness, or usefulness of achievement activities. Like expectancies, values have motivational significance because they guide thoughts, feelings, and behavior. For example, what we judge to be important, attractive, or useful can influence the activities we choose, how we evaluate other people and events, and our worldviews. Values also are rooted in the moral constructs of 'ought' and 'should' as illustrated by the belief that one should try hard in school regardless of their perceived abilities. Particularly during adolescence, trying hard in school is sometimes compromised because of the tension between the need to achieve and the need to belong, both of which are fundamental human motives. The sociologist James Coleman astutely observed more than 50 years ago that adolescents care more about being popular than being perceived as smart and that they even downplay effort in order to be liked by popular peers (Coleman, 1961).

Oppositional Identity

The tension between having friends and being a high achiever can play out in particular ways for youth of color, African-American youth in particular, in a phenomenon emerging from anthropological analyses known as oppositional identity (Fordham and Ogbu, 1986). African-Americans and some Latino groups have been labeled as an involuntary minority – that is, a group which has become part of the American fabric not by choice, but as a result of slavery, conquest, or colonization. One consequence of this history is that acceptance of mainstream values about working hard and school success may be perceived as threatening to one's social identity. Particularly during adolescence, African-American and Latino youth may adopt oppositional identities whereby they show relative indifference, or even disdain toward achievement behaviors that are valued by the larger society. Fordham and Ogbu (1986) coined the term *acting white* to describe African-American high-school students' perceptions of their same race peers who work hard to do well in school. Antiachievement peer group pressure from involuntary minorities can take many forms, including being labeled an 'oreo' or 'sellout,' social ostracism, and even physical assault (Kao, 2000).

The discourse surrounding oppositional identity during adolescence has remained active, in part because it provides a motivational explanation for the achievement gap between black and white students (Downey, 2008). One would be hard-pressed to find an article on academic motivation in African-American adolescents in the last 20 years that does not explicitly or implicitly make reference to oppositional identity. That construct also has been linked to other motivational phenomena discussed earlier in this article, such as stereotype threat and disidentification, as a way to fully capture the motivational challenges that African-American students face.

Other than the rich qualitative studies that followed Fordham and Ogbu's inaugural research, there is not a lot of hard evidence that black students or members of other involuntary minority groups endorse the attitudes assumed by an oppositional identity perspective. For example, survey methods were used to test hypotheses about oppositional identity with data from the National Education Longitudinal Study, a nationally representative panel study of 25 000 ethnically diverse students who were assessed in 8th, 10th, and 12th grade (see review in Downey, 2008). Contrary to predictions from oppositional identity, black students in these analyses were more likely than white students to endorse positive school attitudes and just as likely to hold high educational aspirations (although they were doing more poorly in school). What is true – and what may be the lasting contribution of the oppositional identity literature – is that some high-achieving students of color experience a particular kind of conflict between achievement strivings and their desire to be accepted by members of their racial/ethnic group and this conflict may at times interfere with their motivation to do well in school.

Characteristics of Valued Peers

Survey methods that directly ask students of color whether they value effort in school sometimes yield inconclusive findings, in part because of the social desirability built into such questions (almost everyone agrees that effort is important). Recent motivation research has offered less reactive ways to measure the degree to which students of color report that they value working hard and doing well in school. For example, Taylor and Graham (2007) used peer nomination procedures in which they asked elementary and middle-school African-American and Latino youth to list the names of classmates whom they admire, respect, and want to be like. The rationale for these questions was that if researchers can identify the characteristics of classmates whom a student admires, respects, and wants to be like, this tells us something about the characteristics that the student values. A conceptually similar approach (the 'admiration ladder') was used in early motivation training work to enhance the achievement motive. Taylor and Graham (2007) were particularly interested in how nominations as someone that the nominator admired, respected, and wanted to be liked varied as a function of the achievement level of the person nominated. The research documented that girls of all grade levels overwhelmingly nominated high achievers as classmates they most valued, whereas boys were less likely to nominate high-achieving peers as they approached middle-school. By the time African-American and Latino boys reached adolescence, they appeared to look more admiringly upon less academically inclined peers.

Such findings highlight the particular motivational challenges of ethnic minority males, a gender pattern that has not adequately been acknowledged. In mainstream gender research on motivation, a dominant theme is the heightened vulnerability of girls to motivational deficits (Meece and Askew, 2011). The argument is that gender role socialization and cultural stereotypes about women and achievement lead many girls to question their academic competence more than boys, particularly in math; to display more maladaptive reactions to failure, and to perceive more barriers to success. However, these arguments may need to be reframed because it is evident that ethnic minority males (i.e., African-American and Latino) are faring more poorly than females (e.g., Way and Chu, 2004). The ethnicity by gender differences increase across the school years and are particularly apparent when the measures are so-called markers of adolescent success (i.e., high-school graduation) and young adult mobility (i.e., enrollment in and

completion of college). Harsh disciplinary practices like school suspension also fall disproportionately on African-American males (American Psychological Association, 2008). These findings indicate that ethnic minority boys, more than girls, must cope in unique ways with the dual stressors of academic challenge and stereotypes about their group and these stressors can undermine the endorsement of achievement values.

Self-Affirmation

Self-affirmation theorists have examined how an opportunity to express one's core personal values can have positive motivational consequences. In one novel intervention, Cohen et al. (2009) had African-American and white middle-school students choose from a set of values (including those related to achievement) the two or three that were most important to them and to write a paragraph about why they selected those values. Students in the control condition wrote about the values that were least important to them. The rationale for the intervention was that reflecting on personally important values can be self-affirming and can therefore buffer some of the challenges associated with ethnic minority status. This simple intervention yielded powerful effects. African-American students in the values affirmation condition attained higher grades over the course of the school year and were actually able to reduce the achievement gap between black and white students.

Achievement Values and the Immigrant Paradox

Driven largely by immigration during the past 40 years, the racial and ethnic landscape in the United States has been dramatically redrawn. Although whites are still the majority group in the nation as a whole, Asians and Latinos are now the fastest growing ethnic groups. In some states, like California and Texas for example, that growth has been so dramatic that it is no longer meaningful to talk about majority and minority groups inasmuch as no single ethnic group holds the numerical balance of power. The increased presence of immigrant children of color in the schools has led to an interest in the impact of acculturation on academic motivation (see Urdan, 2011). As schools become more multicultural, immigrant students cope simultaneously with increased cross-ethnic contact and pressures to adjust to the dominant culture. These acculturation pressures have largely been studied in the context of children's valuing of effort and doing well in school.

One of the most provocative findings in the literature on motivation among immigrant groups is the so-called immigrant paradox. Defying 'melting pot' theories that outcomes improve across longer residence in the host culture, recent research has documented that many immigrant groups do more poorly across successive generations of residence in the United States (Garcia-Coll and Marks, 2011). For example, in some studies first- and second-generation adolescents of Latino, Asian, or black (Caribbean) descent did better in school and maintained more positive attitudes about achievement than did same-ethnicity youth whose families had resided in this country for three or more generations. This phenomenon has been labeled a paradox because immigrant youth tend to have less well-educated parents and be of lower socioeconomic status than native-born peers.

Communicated parental values about hard work, the importance of a good education, and family obligation appear to be among the most important factors accounting for higher achievement strivings among immigrants and children of immigrants (second generation) compared to their counterparts of third generation and beyond. Family obligation refers to how much family members feel a sense of duty to help one another and to take into account family needs when making personal decisions (Fuligni, 2001). It has been shown that Latino and Asian immigrant youth are more likely than their American-born counterparts to report a belief in family duty. Family obligation, in turn, is correlated with achievement values inasmuch as many of the youth feel that doing well in school is something that they owe their parents.

Conclusions

This discussion of motivation in racial/ethnic minority youth has focused on the challenges to expectancy and value, in part due to structural barriers that often accompany minority status in the United States. But it is also important to acknowledge resilience in youth of color and successful intervention efforts to overcome those challenges. There are promising motivation interventions, such as those related to beliefs about intelligence and self-affirmation, that foster achievement strivings among youth at risk for school failure. These interventions have generated a good deal of excitement because of their potential to deliver short but powerful treatments that not only boost immediate achievement but also reduce the racial achievement gap (see review in Yeager and Walton, 2011). However, given the structural barriers that influence both *can* and *want*, it may be unrealistic to think that changing one set of beliefs – be it worries about confirming racial stereotypes, implicit theories about intelligence, or the importance of affirming personal values – will have lasting effects on motivation and achievement no matter how powerfully they are delivered. Multifaceted approaches that address multiple belief systems are probably needed.

Furthermore, the burden for motivational change to enhance achievement should not remain solely with the individual student. In light of challenges to perceptions of 'I can,' teachers need to be especially vigilant about not employing instructional practices that can indirectly and unintentionally communicate low-ability messages or activate stereotype threat. For example, it has been documented that undifferentiated praise for success at easy tasks, unsolicited offers of help, and too much sympathy following failure can lead students to attribute their academic setbacks to low ability (Graham, 1991; Rattan et al., 2012). Furthermore, altering pedagogical practices to be more effort- rather than ability-oriented can have immediate impact on students' motivation, even among those who are highly identified with the achievement domain. Cohen et al. (1999) found that African-American college students displayed more subsequent task motivation when feedback about poor performance was accompanied by criticism and communicated high expectations than when the same criticism

was accompanied by general praise as a buffer. Such wise feedback can shift the attribution for failure away from low ability and toward those factors such as lack of effort that are under volitional control. And school administrators should become more sensitized to how policies related to discipline, such as suspension and expulsion, can promote perception of barriers to opportunity and discrimination among ethnic minority adolescents (APA Task Force, 2008). In other words, there are both individual-level variables and structural or systemic factors that need to be targets of motivational interventions for ethnic minority students. Such a multilevel approach, albeit challenging, will be the best way for motivation researchers to bring their expertise to bear on ways to enhance the achievement striving of ethnic minority youth and make it last.

See also: Academic Motivation and Performance: Task Value Interventions; Control Behavior: Psychological Perspectives; Expectancy-Value-Cost Model of Motivation; Gender and Academic Motivation; Grit; Internal/External Frame of Reference Model; Motivation and Actions, Psychology of; Motivation in Australian Aboriginal Populations; Motivation, Familial Influences on; Motivation, Learning, and Instruction; Personal Projects; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Teacher Motivation.

Bibliography

American Psychological Association Zero Tolerance Task Force, 2008. Are zero tolerance policies effective in the schools? Evidentiary review and recommendations. American Psychologist 63, 852–862.

Benner, A., Graham, S., 2013. The antecedents and consequences of racial/ethnic discrimination during adolescence: does the source of discrimination matter? Developmental Psychology 49, 1602–1613.

Blackwell, L.S., Trzesniewski, K.H., Dweck, C.S., 2007. Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. Child Development 78, 246–263.

Cohen, G., Garcia, J., Purdie-Vaughns, V., Apfel, N., Brzustoski, P., 2009. Recursive processes in self-affirmation: intervening to close the achievement gap. Science 324, 400–403.

Cohen, G., Steele, C., Ross, L., 1999. The mentor's dilemma: providing critical feedback across the racial divide. Personality and Social Psychology Bulletin 25, 1302–1318.

Coleman, J.S., 1961. The Adolescent Society. Greenwood Press, Westport, CT.

Correll, J., Park, B., Judd, C.M., Wittenbrink, B., 2002. The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. Journal of Personality and Social Psychology 83, 1314–1329.

Devine, P.G., Brodish, A.B., 2003. Modern classics in social psychology. Psychological Inquiry 14, 196–202.

Downey, D.B., 2008. Black/white differences in school performance: the oppositional culture explanation. Annual Review of Sociology 34, 107–126.

Dweck, C.S., 2006. Mindset: The New Psychology of Success. Ballantine Books, New York.

Eccles, J.S., Wong, C.A., Peck, S.C., 2006. Ethnicity as a social context for the development of African-American adolescents. Journal of School Psychology 44, 407–426.

Fisher, C., Wallace, S., Fenton, R., 2000. Discrimination distress during adolescence. Journal of Youth and Adolescence 29, 679–694.

Fuligni, A.J., 2001. Family obligation and the academic motivation of adolescents from Latin American, Asian, and European American backgrounds. New Directions for Child and Adolescent Development 94, 61–76.

Fordham, S., Ogbu, J., 1986. Black students' school success: coping with the "burden of 'acting white'". Urban Review 18, 176–206.

Garcia-Coll, C., Marks, A.K. (Eds.), 2011. The Immigrant Paradox in Children and Adolescents: Is Becoming American a Developmental Risk? American Psychological Association, Washington, DC.

Good, C., Aronson, J., Inzlicht, M., 2003. Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. Journal of Applied Developmental Psychology 24, 645–662.

Graham, S., 1991. Communicating low ability in the classroom: bad things good teachers sometimes do. In: Graham, S., Folkes, V. (Eds.), Attribution Theory: Applications to Achievement, Mental Health, and Interpersonal Conflict. Erlbaum, Hillsdale, NJ, pp. 17–36.

Graham, S., Weiner, B., 2011. Motivation: past, present, future. In: Harris, K., Graham, S., Urdan, T. (Eds.), APA Educational Psychology Handbook, vol. 1. American Psychological Association, Washington, DC, pp. 367–397.

Harris-Britt, A., Valrie, C., Kurtz-Costes, B., Rowley, S.J., 2007. Perceived racial discrimination and self-esteem in African American youth: racial socialization as a protective factor. Journal of Research on Adolescence 17 (4), 669–682.

Kao, G., 2000. Group images and possible selves among adolescents: linking stereotypes to expectations by race and ethnicity. Sociological Forum 15, 407–431.

Meece, J.L., Askew, K.J.D., 2011. Gender, motivation, and educational attainment. In: Harris, K., Graham, S., Urdan, T. (Eds.), Educational Psychology Handbook, vol. 2. American Psychological Association, Washington, DC, pp. 139–162.

Ngo, B., Lee, S.J., 2007. Complicating the image of model minority success: a review of Southeast Asian America Education. Review of Educational Research 77, 415–453.

Neblett Jr., E.W., White, R.L., Ford, K.R., Philip, C.L., Nguyan, H.X., Sellers, R.M., 2008. Patterns of racial socialization and psychological adjustment: can parental communications about race reduce the impact of racial discrimination? Journal of Research on Adolescence 18 (3), 477–515.

Rattan, A., Good, V., Dweck, C.S., 2012. "It's OK – not everyone can be good at math": instructors with an entity theory comfort (and demotivate) students. Journal of Experimental Social Psychology 48, 732–737.

Rosenbloom, S., Way, N., 2004. Experiences of discrimination among African American, Asian American, and Latino adolescents in an urban high school. Youth and Society 35, 420–451.

Sackett, P.R., Hardison, C.M., Cullen, M.J., 2004. On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. American Psychologist 59, 7–13.

Steele, C.M., 2010. Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us. Norton & Simon, New York.

Steele, C.,M., Aronson, J., 1995. Stereotype threat and the intellectual performance of African Americans. Journal of Personality and Social Psychology 69, 797–811.

Taylor, A.Z., Graham, S., 2007. The development of relations between achievement values and perceptions of barriers in African American and Latino youth. Journal of Educational Psychology 99, 52–64.

Urdan, T., 2011. Factors affecting the motivation and achievement of immigrant students. In: Harris, K., Graham, S., Urdan, T. (Eds.), Educational Psychology Handbook, vol. 2. American Psychological Association, Washington, DC, pp. 293–313.

Way, N., Chu, J. (Eds.), 2004. Adolescent Boys: Exploring Diverse Cultures of Boyhood. New York University Press, New York.

Yeager, D.S., Walton, G.M., 2011. Social-psychological interventions in education: they're not magic. Review of Educational Research 81, 267–301.

Race and Racism in the Twenty-First Century

Yolanda T Moses, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Race and racism in the contemporary globalized world is a product of historical forces that still resonate today. With a focus on the United States and its relation to race–racism research both in the United States and abroad. By reexamining the Atlantic slave trade, social scientists and historians tie its establishment to the development of the modern world economy. This article then examines the specific ways that these linkages are still operative in the United States and globally. The author also suggests new research strategies for understanding the specific ways to expose the depth of these structural racial inequalities, and how to begin to eradicate them.

Racial Formation in the Twentieth Century by Omi and Winant (1984, 1994, 2002) provided a framework through which sociologists, anthropologists, and others could locate the role of race in structuring broader social formations. It was quite useful to use to look at the political and social terrain of the 1970s and 1990s in the United States, especially. It arose at the same time politically that the conservatives and neoliberals were forging a 'new world order' that introduced the concepts of 'colorblindness' and 'postracial America' that one still hears echoed among conservatives during the first Obama administration (Goodman et al., 2012). Also, in the 1990s, anthropologists, sociologists, and biologists reaffirmed the fact that race as a biological construct was not real, but was an illusion, a social construct (Mukhopadhyay and Moses, 1997; Goodman et al., 2012; Smedley, 1993). Omi and Winant challenged that benign view by indicating that race was not an illusion, but was very much a foundational axis of American society (Omi and Winant, 2012: p. 14). Mullings, Harrison, and others also comment that anthropologists, by denying the biological basis of race, actually left a void and a vacuum to talk about what race really was, and how it continued to operate as a powerful social construction (Mullings, 2005; Harrison, 1995). It was Mukhopadhyay and Moses in 1997 who challenged the field of American anthropology to reestablish race in its anthropological discourse (Mukhopadhyay and Moses, 1997). Since then, anthropologists have looked at the social construction of race and racism from many perspectives, including intersectionality of issues like gender, sexual orientation, class, and identity (Mullings, 1997; Bolles, 2001; McClaurin, 2001); power and hegemony (Sanjek, 1998; Smith, 2012); globalization and capitalism (Wade, 2003; Brodkin, 2000); genetics and genomics (Goodman et al., 2012; Bliss, 2012); immigration, migration, and transnationalism (Kearney, 1995); racial critique of the discipline of anthropology itself (Baker, 1998; Mullings, 2005; Marks, 2005; Blakey, 1997; Harrison, 2008; Susser and Patterson, 2001); antiracism research (Mullings, 2005); and the future of race–racism in the United States and the world (Winant, 2001; Goodman et al., 2012).

Pushing the Racial Formation Paradigm

Since the 1990s, there have been developments in the research on race in the United States and abroad that both critique and build on the foundation of Omi and Winant's work. The edited book *Racial Formation in the 21st Century* (2012) is a place to begin this critique and new vision of racial formation and its processes. In the introduction, while paying homage to Omi and Winant, the editors, HoSang, LaBennett, and Pulido point out at least three ways that *Racial Formation in the 20th Century* fell short. They are specifically that Omi and Winant focus too much on the United States as the global model for race, there was a conflation of the concept of race and ethnicity, and class is treated in a dismissive way (2012: p. 6). This book goes on to take a holistic view of new racial theory formation since the original Omi and Winant publication, as well as what they call dialogues and conversations with other racial formation theories that were also in circulation during the 1990s. Two of these theories are critical race theory (Crenshaw, 1991; Crenshaw et al., 1995) and intersectionality theory (Crenshaw et al., 1995; Twine, 1998). Racial formation theory of Omi and Winant did tackle the issue of ethnicity in the United States by explaining that ethnicity is a process by which individuals and groups (come) to understand, or understand themselves as separate or different from others based on social practices such as language, religion, rituals, and other patterns of behavior. In this analysis, they note that the use of the term ethnicity instead of race offers the opportunity for conservatives and neoliberal policies to question or negate the legal and policy gains of people of color in this country, when new policies and programs are focused not on race, but on the creation of a 'colorblind society,' which ignore the history and the intractability of structural racism (Goodman et al., 2012).

Other racial theorists outside the fields of sociology and anthropology, such as Thomas Holt (2002) in history and Paul Gilroy (1993, 2002) in cultural studies and American studies have used the concept of Omi and Winant's racial formation to build theoretical frameworks to better understand and explain the history of the formation of black expressive culture, black diasporic identity, and hybridity. Winant (2000, 2001) argues that there was no such thing as race in classical antiquity. Isaac in his book *The Invention of Racism in Classical Antiquity* did find what he calls significant examples of 'protoracism' in Greek and Roman literature, "despite his observance that racism did not exist in its current form of biological determinism, nor was there any systematic persecution of any ethnic group or another." Other historians and anthropologists also found it useful to make a distinction

between prototypical forms and the systematic racial classification that took center stage over the past two centuries in our modern world (Smedley, 1993). Indeed, Mullings contends that a reexamination of history and race, especially the slave trade puts the concept of race in a different light (2005). Historians like Holt (2002), and anthropologists like Brodkin (2000) along with Winant (2001) all tie the centrality of the African enslaved labor trade to the development of the modern capitalist world economy. Stuart Hall and Paul duGay (1996) and his concept of new ethnicities fall in line with how literary studies have conceptualized race and ethnicity. This field has highlighted how over recent years transnational literacy influences multilingual traditions that traverse national boundaries, and that new multinational public spheres are produced by the translation of diverse texts, ideas, and people. HoSang, Bennett, and Pulido also note that by focusing on the circuits of ethnicity formation in conjunction with racial formation and exploring these relations in and across nations, such studies began to diverge from sociological models of racial formation in which the state plays a central role (HoSang et al., 2012: p. 9). In other words, the sheer movement of vastly diverse groups of people is producing especially for the late twentieth and early twenty-first centuries, new ways of thinking about and enacting racial formation on a global scale. Paul Gilroy in *Black Atlantic* (1993) emphasizes the common global routes that historically link black expressive culture. Foregrounding common experiences of oppression rather than defining blackness in relation to Africa as an 'actual or mythical' homeland, Gilroy was able to conceptualize black communities as constructed in defiance of racial essentialism and as connected through popular cultural productions (Gilroy, 1993).

Racial Formation in the Twenty-First Century

Racial Formation in the Twenty-First Century takes us to old and new intersections, beyond the borders of the United States, back to biology and back to the unfinished job of dismantling structural racism here in the United States and beyond. Research in the area of intersectionality in the twenty-first century examines how sociopolitical categories and identities based in race, gender, and class interact, sometimes systematically structuring social inequality and injustice, and sometimes profoundly challenging such social structures. This notion of challenging structures builds on Kimberle Crenshaw's concept of intersectionality (1991) that sought to explain how the experiences of women of color are "simultaneously the product of intersecting patterns of racism and sexism, and how these experiences tend not to be represented within the discourses of either feminism (read white women) or anti-racism (read men of color)" (Crenshaw, 1991: p. 299). Priya Kandaswamy (2012) points in this new direction as she builds on legal scholar Crenshaw's model and suggests that racial formation is fundamentally a gendered and sexualized process and further argues that viewing race as unstable, historically produced, and changing in the ways that Omi and Winant demonstrates complicates the meaning of intersectionality (Kandaswamy, 2012: p. 24). For her, it is precisely the intentional interrogation of the historical embeddedness of race, gender, sexuality, and class with each other that will produce new theories of intersectionality. In a similar vein, Ferguson (2012) challenges the historiographical framework proposed by Omi and Winant in *Racial Formation* that helped to theorize US national liberation movements as more than instances and outcomes of social agitation, showing that they were historiographical provocations that pushed people in the United States to think differently about racial difference and identity. He is fine with the theorizing, but he feels that they leave out a significant group of people who were not officially a part of the national civil rights or national liberation movements of the time. He says "those [smaller] movements were initiated by women of color and queers of color within the United States. Indeed, reperiodizing anti-racist politics according to the emergence of those movements constituted by people of color who were marginalized by sexuality and gender allows us not only to observe the specificities of civil rights and national liberation movements but also to see how race and anti-racist politics were rearticulated in those very movements…" (Ferguson, 2012: p. 45).

Andrea Smith (2012) moves us in still another direction that links indigeneity, settler colonialism, and the uninterrogated terrain of white supremacy. In her article, she challenges the very racial theory framework that is used in Ethnic Studies and Native Studies programs. She agrees that "native peoples are treated as ethnic groups suffering from discrimination rather than as nations undergoing colonization, and that these two groups should get together to discuss that erroneous framework … at the same time because of this political and intellectual divide, there is little dialogue among them that would help us to understand how white supremacy and settler colonialism intersect, especially in the United States" (2012: p. 66). In this chapter, she argues that scholars' lack of attention to settler colonialism hinders the analysis of race and white supremacy developed by scholars who focus on race. In the end, she focuses on ways that native scholars and practitioners can rearticulate their relationship with the land, and with the white power structure (2012: p. 88).

Where does one go from here? Future research should focus among other things on relationships among antiracist, women's, anticolonial/anti-imperial, and labor-based/antipoverty movements. There is also a powerful 'intersectionality' that exists on the political right (especially in the United States), where antifeminist, antilabor, 'blame the victim'-oriented, homophobic, and ambiguously antiracist orientations operate and need to be interrogated. There is a need for more research that investigates the instability in practice of intersectional situations and social structures. For example, how do individuals and movements navigate among the raced, gendered, queer, class, and caste-based dimensions of social, cultural, and political life; what is the relationship between intersectionality and indignity; what are the elusive linkages among axes of identity and difference, and those between alliance and antagonism? There is also a continuous need for historical views of intersectionality both in the United States and transnationally. Even more fundamentally, what are the emerging contested views and definitions of intersectionality itself? Is intersectionality in both local and global contexts different? If so, how? How are intersectionality

and postcolonialism connected, and what are the newly emerging theories around intersectionality and the politics of identity? Finally, theoretical frameworks and methodologies are needed that highlight micro–macro linkage problems and social movements in particular comparative and historical contexts.

Race and Racism beyond the Borders of the United States

For the past two decades, scholars in anthropology, sociology, and other fields have looked at issues of immigration, transnationalism, race and ethnicity, diasporic studies (Ong, 1999), and intersectionality through cross-cultural, transnational, and global lenses. The idea of race and the hierarchical institutionalization of racial difference emerged dialectically in relation to sixteenth-century economic transformations that ultimately created what is now known as 'the modern West' (Holt, 2002). While notions of difference operated prior to this period, the expulsion of Muslims from Europe, the initial European voyages of exploration and discovery, and the development of mercantile trade generated a novel situation whereby, for the first time, racialized labor became crucial to the mobilization of productive forces on a world scale (Holt, 2002: p. 32; Thomas and Clark, 2006). At the same time that the associations between nation building and imperialism and between racial slavery and the development of export-oriented mass agricultural production became more tightly integrated, new philosophical and scientific ideologies began to circulate in Europe about the nature of man. Within religious, philosophical, scientific, and political discourses, hierarchies of human value over time were mapped onto gendered, racial, and civilizational differences (Thomas and Clark, 2006; Goodman et al., 2012).

In this way, early state formation (in the old and new worlds) and mercantile capitalism inaugurated material and ideological processes that indelibly linked the 'New World' and the 'Old' in a common project of defining people in a commodified social hierarchy that was expressed in gendered racial and class terms. In other words, the development of a highly structured social and economic system in the New World is the direct result of Imperial conquest (by Europeans), plantation slavery, and structural racial domination. These new racialized processes also had an impact back home on metropolitan Europe as well. According to Gilroy (2002) and Thomas and Clark (2006), this formulation of the modern world conceptualizes the Atlantic Ocean as an integrated geohistorical unit in which the structural transformations associated with early European expansion westward created what ultimately became a triangular web of political, economic, and sociocultural relations joining individuals, communities, and classes on three continents in a single sphere of interaction (Thomas and Clark, 2006; Mullings, 2005; Mukhopadhyay et al., 2007).

Three Hundred Years and Counting

Imperialism, slavery, and colonial exploitation in the fifteenth, sixteenth, and seventeenth centuries created enduring global linkages that were sustained through the postemancipation period of the mid-nineteenth century and into the twentieth and twenty-first centuries. At the same time when the age of exploration was taking place in the fifteenth and sixteenth centuries, both science and religion colluded (sometimes consciously and sometimes not) as the two dominant discourses of colonialization and the institutionalization of a permanent racial hierarchy in the Americas, Africa, and South Asia. Advocates for slavery and colonial expansion helped to institutionalize the new science of anthropology, in part to counter Abolitionists' claims based on morality and biblical tenets. As the discipline developed during the late nineteenth century, human difference was parsed along a color-coded hierarchy from savage to civilized – from black through brown, yellow, and red to white (Baker, 1998; Thomas and Clark, 2006; Goodman et al., 2012). The fact that some researchers documented customs and behaviors while others measured brains and bodies did not change this hierarchy because human diversity and cultural differences were blurred and racially mapped in a way that privileged biology as the basis for human difference, with those in power being male Europeans.

However, other changes were afoot in the early twentieth century as altered relationships between the means of production and local and global consumption that also had a hand in transforming racial meanings once again. The rise of resistance movements such as Garveyism and Pan-Africanism in the early twentieth century showed that racialized peoples, both laborers and intellectuals, were also helping to shape notions of their own identity as well as notions of where they belonged, both physically and metaphorically. While socioeconomic and political arrangements in the United States required that racialized labor forces remain fixed within the particular (material and ideological) places to which they were transported, by the middle of the twentieth century models of production and consumption instead relied upon a massive movement of these same labor forces out of their place, "from South to North in the United States, from colony to metropole in the British, French, Portuguese and Dutch West Indies, from country to city in southern and western Africa" (Holt, 2002: p. 70). This movement that was facilitated, in part, by the liberalization of US immigration laws in 1965, helped to generate a transnational wave of immigrants from homelands especially from the Caribbean, Latin America, and Asia, to new lands such as the United States, Europe, and Canada.

New Engagements of Race

By the mid-twentieth century, these changes in demographics helped to push the transformation of locally particular notions of difference into discourses of new meanings of race, ethnicity, and national descent. This reordering of human subjectivity in increasingly ethnic heritage terms follows structurally related shifts in the language of contemporary American racial organization particular to the post-1965 period. During this period, the intensification of civil rights discourses against discrimination generated new ideologies about racial/ethnic identities and racial difference. These

changes also set the stage for the formation of other postcivil rights movements that focused on honoring new cultural traditions and heritages (women's rights, gay rights, Native American rights, and Latino rights, just to name a few). In a commodified world, these new heritages, rather than becoming homogenized in an 'American melting pot' advertising discourse that stressed sameness, new marketing strategies were developed to exploit new lucrative racial and ethnic markets. The cultural formation of a new commercial politics of linkage between people in the Americas and those in related 'homelands' set the stage for the establishment of a heritage and identity category through which linkages to origins were used to supplement national identity and citizenship in the adopted country. In the social sphere, this emphasis on diasporic interconnection redefined prevalent notions of biological race to emphasize what came to be called by some researchers cultural race, shaped by new conceptions of ethnicity, or ancestral heritage (Mukhopadhyay et al., 2007; Thomas and Clark, 2006; Harrison, 2008). This notion of cultural race is also deeply rooted, not in biology but in a historical context.

During the 1980s through the 1990s, the rights discourse of the 1960s all but disappeared from the national political discourse in the United States. In other parts of the world, discourses of equality were also on the wane because states were less powerful in some locations, and in others, less willing to hold 'equality' as a value for all citizens. At the same time, the postcolonial context was one in which migration, movement, and the rise of a global media have created a situation in which as Winant has indicated "the nation-based dimensions of racial solidarity have atrophied" (Winant, 2000: p. 180). Today, no one can deny that race and racism currently exist in a world in which larger and larger percentages of national populations are illiterate, where the avenues for self-advancement have become increasingly limited, where remittances constitute ever greater percentages of developing countries' gross domestic products, where sexual tourism and sex trafficking is on the rise (in the Inland Empire of Southern California as much as it is globally), where the availability of critical social services is on the decline in the United States and elsewhere, where a commitment to social equality becomes framed as anticapitalist by the conservatives and far right politics in the United States, and where an assertive call for global peace is seen as suspiciously unpatriotic. Global economic restructuring has thus resuscitated racialized and naturalized social stratification systems that are currently at work in such industries as banking, and the mortgage foreclosure industry here in the United States that has specifically targeted poor racialized populations to prey on.

New Globalization Research and Race

It has become commonplace to speak of the contemporary intensification of processes of globalization and the ways in which they are continually reconfiguring the structures of everyday life. While scholarly analyses of globalization have proliferated, and while there have been recent attempts within the social sciences to consider the articulations among ethnicity, gender, and sexuality within a global frame of analysis, race and processes of racialization are not usually considered central to academic discussions of global economic and political transformations. Yet, globalization is facilitated by the transmission and reproduction of deeply embedded social prejudices rooted in a past characterized by territorial concepts of identity that both generated and were generated by structural racial inequalities. As such, the contemporary twenty-first century redistribution of wealth has exacerbated historically entrenched racial hierarchies. In other words, racial formations dynamically reflect and shape global processes and are not merely effects of them. As a result, the complexity of contemporary global processes can never be fully grasped without a deep understanding of the historically specific and dynamic ways that race has both constituted and been constituted by global transformations. What is clearly critical, then, to a more complete understanding of contemporary global processes is an integrated analysis of the historical precedents of current circulations, of how imperialism and racial ordering have shaped global movements, and of the ways conceptualizations of identity, membership, and citizenship have been both constructed and institutionalized in racial terms (Mullings, 2005; Thomas and Clark, 2006).

Dismantling Structural Racism: The Unfinished Job

In some ways, this short article has only touched the surface of this topic of 'race and racism' as one looks at the unfinished business of exposing anew, and then eradicating structural racism. It is not enough for researchers to just reveal the breadth and depth of the social and cultural underpinnings of structural racism, and the power dynamics that operate to maintain this highly racialized and stratified social system in US society and elsewhere. One must do more. Sociologist Howard Winant reminds us that as we talk about the issue of 'postracialism' in the United States to remember that some of the incorporative concepts of racial difference tend to actually validate the very identities they purport to question. Rather than serving as catalysts for truly new racial formations they actually "reinforce the racialized social structures such as comprehensive racial stratification, discrimination, xenophobia, etc. that they so loudly repudiate" (Winant, 2001: p. xix).

The author suggests that much more critical research and praxis work still needs to be done in key areas such as looking for new intersections in addition to mining more deeply the intersectionality of race, class, and gender; researching race, migration, and immigration both locally and globally; continued research on the validity of the claim of reverse discrimination by Euro-Americans (especially in the United States); researching the contradictions of claims of 'colorblindness' in the ubiquitous presence of conversations about race; conducting a deeper and more critical analysis of the continuing educational, housing, and health disparities along axis of race, ethnic, and class lines; researching the impact of national, regional, and state demographic changes and race; continued research and study of whiteness and power from the inside out; more research on the links between race and globalization studies; and more research on the relationship of race, gender, and human rights, just to name a few.

Conclusions

Race is about power relations, and continues to be an ideology to legitimize the dominance of certain groups. Race and its consequence racism are fundamentally part of a global system of structural racial stratification and inequality. People indoctrinated in US individualism often have difficulty understanding abstract notions of a social system, social stratification, and how one's social position impacts individual experience and opportunities. For example, fluctuations in racial categorization and definitions of 'whiteness' reflect demands for population growth set against a backdrop of powerful elites seeking to maintain a system of social stratification and cultural dominance. The US race-based system of social classification became a way of maintaining the dominance of elite 'white' groups, politically and socially, while simultaneously recruiting new populations. Race became a central basis for organizing labor and maintaining an economically stratified system, first in the agricultural sector and now in the growing industrial and postindustrial sectors. Racializing the labor force helped to mask the pervasive class stratification and structural inequality that has always characterized American life. The North American system of racial classification is most accurately viewed as a historically and culturally specific 'legitimizing ideology,' a complex and unique way for explaining, justifying, and perpetuating a system of social, economic, and political inequality. But these racial meanings were also contested in the past and will continue to be contested (Mukhopadhyay et al., 2007; Omi and Winant, 2012).

See also: Ethnocentrism; Ghetto; Multiculturalism; Xenophobia and Nativism.

Bibliography

Baker, Lee, 1998. From Savage to Negro: Anthropology and the Construction of Race 1896 to 1954. University of California Press, Berkeley.

Bolles, A. Lynn, 2001. Seeking the Ancestors: Forging a Black Feminist Tradition in Anthropology. See McClaurin, 2001, pp. 24–48.

Brodkin, Karen, 2000. Global capitalism: What's race got to do with it? American Ethnologist 27 (2), 237–256.

Blakey, Michael L., 1997. Past is present: comments on "in the realm of politics: prospects for public participation in African-American plantation archaeology." History of Archaeology 31 (3), 140–145.

Bliss, Catherine, 2012. Race Decoded: The Genomic Fight for Social Justice. Stanford University Press, Palo Alto.

Crenshaw, Kimberle, 1991. Mapping the margins: identity politics, intersectionality and violence against women. Stanford Law Review 43, 1241–1299.

Crenshaw, Kimberle, et al., 1995. Critical Race Theory: Key Writings that Formed the Movement. The New Press, New York.

Ferguson, Roderick A., 2012. On the specifics of racial formation: gender and sexuality in historiographies of race. In: Racial Formation in the 21st Century. University of California Press, Berkeley, pp. 44–56.

Gilroy, Paul, 1993. The Black Atlantic. Verso, New York.

Gilroy, Paul, 2002. Against Race: Imagining Political Culture beyond the Color Line. Harvard University Press, Cambridge.

Goodman, Alan, Moses, Yolanda, Jones, Joseph, 2012. Race: Are We So Different? Wiley-Blackwell, New York.

Hall, Stuart, duGay, Paul, 1996. Questions of Cultural Identity. Sage Publications, London/Thousand Oaks, CA/New Delhi.

Harrison, Faye V., 1995. The persistent power of "race" in the cultural and political economy of racism. Annual Review of Anthropology 24, 47–74.

Harrison, Faye V., 2008. Outsider within: Reworking Anthropology in the Global Age. University of Illinois Press, Champaign – Urbana.

Holt, Thomas C., 2002. The Problem of Race in the 21st Century. Harvard University Press, Cambridge, MA/London.

HoSang, Daniel Martinez, La Bennett, Oneka, Pulido, Laura, 2012. Racial Formation in the 21st Century. University of California Press, Berkeley.

Kandaswamy, Priya, 2012. Gendering racial formation. In: HoSang, LaBennett, Pulido (Eds.), Racial Formation in the 21st Century. University of California Press, Berkeley, pp. 23–43.

Kearney, Michael, 1995. The local and the global: the anthropology of globalization and transnationalism. Annual Review of Anthropology 24 (5), 47–65.

Marks, Jonathan, 2005. Anthropology and the bell curve. In: Besteman, C., Gusterson, H. (Eds.), Why America's Top Pundits Are Wrong: Anthropologists Talk Back. University of California Press, Berkeley/Los Angeles/London, pp. 206–227.

McClaurin, Irma (Ed.), 2001. Black Feminist Anthropology: Theory, Politics, Praxis and Poetics. Rutgers University Press, New Brunswick, NJ.

Mukhopadhyay, Carol, Henze, Rosemary, Moses, Yolanda, 2007. How Real Is Race? A Sourcebook on Race, Culture and Biology. Roman and Littlefield, Latham, Maryland, Toronto and Plymouth, UK.

Mukhopadhyay, Carol, Moses, Yolanda, 1997. Re-establishing race in anthropological discourse. American Anthropologist 99 (3), 517–533.

Mullings, Leith, 1997. On Our Own Terms: Race, Class, and Gender in the Lives of African-American Women. Routledge, New York/London.

Mullings, Leith, 2005. Interrogating racism: toward an anti-racist anthropology. Annual Review of Anthropology 34, 667–693.

Omi, Michael, Winant, Howard, 1984, 1994, 2002. Racial Formation in the 20th Century.

Omi, Michael, Winant, Howard, 2012. Conclusion: racial formation rules, continuity, instability and change. In: HoSang, La Bennett, Pulido (Eds.), Racial Formation in the 21st Century. University of California Press, Berkeley, pp. 302–332.

Ong, Aiwa, 1999. Flexible Citizenship: The Cultural Logics of Transnationality. Duke University Press, Durham, N.C.

Sanjek, Roger, 1998. The Future of Us All: Race and Neighborhood Politics in New York City. Cornell University Press, Ithaca/London.

Smedley, Audrey, 1993. Race in North America: Origin and Evolution of a World View. Westview, Boulder/San Francisco/Oxford.

Smith, Andrea, 2012. Indigeneity, settler colonialism, white supremacy. In: Racial Formation in the 21st Century. University of California Press, Berkeley, p. 66.

Susser, Ida, Patterson, Thomas C. (Eds.), 2001. Cultural Diversity in the United States: A Critical Reader. Blackwell Publishing, Oxford, UK.

Introduction. In: Thomas, Debra, Clark, Kamari (Eds.), Globalization and Race: Transformations in the Cultural Production of Blackness. Duke University Press, Durham, NC.

Twine, France W., 1998. Racism in a Racial Democracy: The Maintenance of White Supremacy in Brazil. Rutgers University Press, New Brunswick, NJ/London.

Wade, Peter, 2003. Race and nation in latin america: an anthropological view. In: Appelbaum, N.P., Macpherson, A.S., Rosemblatt, K.A. (Eds.), Race & Nation in Modern Latin America. University of North Carolina Press, Chapel Hill/London, pp. 263–281.

Winant, Howard, 2000. Race and race theory. Annual Review of Sociology 26, 169–185.

Winant, Howard, 2001. The World Is a Ghetto: Race and Democracy since World War II. Basic Books, New York.

Race and Racism, Geography of

James A Tyner, Kent State University, Kent, OH, USA
Joshua Inwood, University of Tennessee at Knoxville, Knoxville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we highlight the changing geographies of race and racism, from the era of European colonialism to the present. We emphasize the co-constitution of race and space and highlight how epistemological understandings of 'race' greatly inform racist practice. Our objective in this article is to consider the legacy and malleability of racial discourse and how the construction of 'race' has material consequences. Special emphasis is placed on eugenics, racial formation theory, and critical race theory.

Introduction

To begin any conversation on race one must proceed from the standpoint that the very nature of the term is problematic (Kobayashi and Peake, 1994, 2000; Gilmore, 2002). Despite considerable efforts to 'prove' that race exists, it is now a widely accepted fact that racial differences are not rooted in any kind of biological or genetic difference, but are the result of social constructions that meet particular, economic, political, and cultural needs. Furthermore, the understanding of race in society is more often than not connected to overarching power relations. In addition, race is a geographical contextual in a process known as 'differential racialization' (Pulido, 2006). Consequently, by studying and understanding racial formation we are able to understand the sociospatial construction of space and place (cf Alderman, 1996; Holloway, 2000; Delaney, 2002; McKittrick, 2006; McKittrick and Woods, 2007; Heynen, 2009; Winders and Schein, 2013). In other words, to understand that race is a social construction is to understand not only the development of identity, but also the development of many other taken-for-granted concepts that connect to power relations (Sundstrom, 2003; Saldanha, 2006; Wilson, 2006; Inwood and Bonds, 2013).

The Birth of Modern Racial Discourse

While the identification of distinct human groups dates back to some of the earliest human writings "it was only after European explorers reached the Western Hemisphere" did any kind of "distinction and categorization fundamental to a racialized social structure" begin to appear and be used to understand human difference (Omi and Winant, 1994: 61). Thus, as European nations began to establish empires throughout Africa, Asia, and North and South America these states needed to come to terms with their conquests and to justify the exploitation of newly colonized peoples. Subsequently, there emerged an assemblage of scientific discourse that purportedly documented the natural division of humanity into a hierarchically ordered set of discrete races (Goldberg, 2002).

To place an object (such as a 'human body') in one class rather than another establishes its central characteristics and creates assumptions about matters that are not seen (Edelman, 1977: 123). This holds particularly true for the 'scientific' construction and promotion of racial categories. Out of the scientific discourses emerged (supposedly) objective criteria that purported to explain the spatial distribution and diffusion of racialized bodies. Classifications were initially constructed according to observed, yet highly selective, physical traits; these, in turn, were used to explain 'cultural' traits and behaviors, including poverty and promiscuity. In actuality, these presumed correlations worked both ways. The presumptive inferiority of 'blacks' for example (but, in general, all 'nonwhites') was taken as given and subsequently prompted the scientific search for associated morphological and anatomical signs between the 'races' (Tucker, 1994: 22; see also Miles, 1989).

While somatic differences were the most common indicator of racial difference, other measurements compared facial angles, cranial capacity, and hair texture. Other more innovative physical markers were used. Etienne Serres, for example, compared the distance between a man's navel and penis to establish differences between Europeans and non-Europeans. Even more ambitious was the work of F.G. Crookshank (1924), who attempted to establish the relationship between 'Orientals' (Mongols), the 'mentally retarded' (Mongoloids), and orangutans. As evidence, he documented differences between palm markings and hand gestures, body and limb postures, eyes, noses, teeth, arms, feet, mouths, tongues, varicose veins, brains, stomachs, anuses, and scrotums.

The 'medicalization' of 'race' afforded a significant degree of legitimacy and authority to existent racial beliefs, thus infusing them with power. Racial classifications were ordered hierarchically. Northern and western European 'races' were held as normative yardsticks against which all other 'races' were measured. The most common standard included a tripartite division of the 'white race': Nordics, Alpines, and Mediterraneans. The tall and fair-skinned Nordics (e.g., people from northern and western Europe) were thought to represent the highest echelons of humanity; lower down the ladder of civilization were the shorter and somewhat darker Alpines (e.g., people from central Europe), followed by the supposedly much shorter and darker Mediterranean peoples (e.g., people from southern and eastern Europe). Trailing behind the 'white races' were the various 'yellow,' 'brown,' 'red,' and 'black' populations.

The effort to construct (and justify) racial differences intensified as European and later North American colonies and

nations turned to the enslavement of Africans to reap incredible profits out of the lives and work of chattel slaves (Du Bois, 1935). It is important to note that chattel slavery as practiced in North America differed from the way slavery was historically understood and used in the ancient world. As a consequence early colonists and European governments went about codifying slavery into colonial societies in ways that increasingly subordinated slaves thus translating racial hierarchies into law racial that placed white men above all others. According to Du Bois (1935) the role that slavery plays in North America is central to the development of racial hierarchies not only within North America, but also within the broader Atlantic World. He argues

> Black labor [in the form of slavery] became the foundation stone not only of the Southern [US] social structure, but on Northern manufacture and commerce, of the English factory system, of European commerce, of buying and selling on a world-wide scale, new cities were built on the results of black labor.
>
> 1935: 5

As slavery began to play an increasingly important role in the spread and development of capitalism throughout the late eighteenth and nineteenth centuries, slavery is at the center of this economic development.

As national economies began to rest more fully on racially organized exploitation (e.g., slavery) and as these colonies and nations expanded it became ever more difficult to justify and rectify the needs of the nation with the exploitation of people of color (Omi and Winant, 1994: 63). Thus, "the invocation of scientific criteria to demonstrate the 'natural' basis of racial hierarchy was both a logical consequence of the rise of this form of knowledge, and an attempt to provide a more subtle and nuanced account of human complexity" (63). It is from these roots that many of our present day misconceptions about race and the continuing significance of racism emerge.

Critical for geography is the way constructed racial differences have been translated in and through space and place. For example, in North America as slavery became an ever more viable alternative to other forms of labor, laws were passed that restricted the movement of people of African descent (Delaney, 1998). Laws were also passed to prevent the so-called miscegenation of 'races' in an attempt to maintain a 'pure' racially defined society (Tyner and Houston, 2000). The legacy of these racist practices, as detailed below, continues to inform contemporary American society.

Race and Eugenics

Up through the late nineteenth century the physiological mechanisms of heredity were obscure (Paul, 1995). It was generally assumed that characteristics acquired by organisms during their lifetimes were transmissible to their progeny; such as belief was forward by the writings of Jean Baptise de Lamarck. In what became popularly known as *Lamarckism*, it was understood that animals, for example, would alter their behaviors in response to changing environmental conditions. One consequence of such changes was that certain, and selected, organs would come into or out of use (Paul, 1995: 40). When applied to humans, this 'principle of inheritance of acquired characteristics' would have direct policy relevance. Simply put, the environment in which children were born and raised would have a major bearing on subsequent generations.

By the late nineteenth century the scientific study of both heredity and (racialized) populations underwent a paradigmatic change. The publication of Charles Darwin's *On the Origin of Species* (1859), most notably, began to undermine ideas of fixed biological categories and the principle of inheritance of acquired characteristics. Darwin's work posited that humans evolved through natural laws of selection and competition. As Noel Castree (2005) summarizes, Darwin – from his expeditionary observations and his studies of pigeon-breeding – derived the metaphor of 'natural selection'; from his reading of Thomas Malthus's work on population and resources, Darwin derived the idea of the 'survival of the fittest'; and from his understanding of the work of geologists James Hutton and Charles Lyell, Darwin articulated the idea of 'deep time' – the idea that change occurs slowly over the *longue durée*. Combined, these elements suggested to Darwin that species interacted with each other and that those species best able to adapt to their conditions of existence were most likely to survive and to produce offspring similarly well adapted.

Although Darwin's theory of natural selection is in many respects counter to that of Lamarck, Darwin's theory of heredity (which he termed 'pangenesis') premised that acquired characteristics were heritable. According to pangenesis, minute particles thrown off by various cells circulate through the body and ultimately concentrate in the 'germ cells'; through this, changes in parents' bodies could be manifested in their children (Paul, 1995: 41). In opposition to the concept of the inheritance of acquired characteristics was Francis Galton (1822–1911), a statistician, naturalist, policy advocate, and cousin of Darwin. Galton premised that hereditary material was transmitted from one generation to the next. Both specific talents and traits of character and personality were inherited; so too was tuberculosis, longevity, madness, gambling, sexual passion, criminal inclinations, and alcoholism thought to be heritable (Kevles, 1995; Paul, 1995).

The mechanisms for inheritance were most clearly articulated by a German biologist, August Weismann. Throughout his writings, Weismann forwarded his theory of the continuity of the 'germ plasm.' He argued that there existed two types of cells that constituted the human body: germ cells – present in the gonads and which give rise to sperm and egg – and somatic cells, which are present in other tissues in the body. For Weismann, the germ cells (or 'germ plasm') were isolated from somatic cells; and while the somatic cells could be affected by the environment, the hereditary units of the germ plasm could not. In proposing that the germ plasm was completely independent from the rest of the cells, Weismann argued that it was the germ plasm which was inherited continuously by one generation from another without alteration from outside influences. Of significance, and in contrast to Lamarckism, Weismann's theory of germ plasm posited that environmental conditions were largely inconsequential; that no matter how much parents worked to

improve their bodies or minds, their heredity would be unchanged and their children none the stronger or smarter (Paul, 1995; see also Stepan, 1991).

By the turn of the twentieth century, scientists and social reformers had already begun to advance the idea that society should recognize the power of heredity in its social laws, in such a way as to favor reproduction of the physically and morally fit (Stepan, 1991: 26). Now, with the sustained assault against the belief in the inheritance of acquired characteristics, there appeared to be clear understanding of the task at hand. Specifically, the emerging science of genetics – buttressed by statistical and biometric studies – indicated that environmental factors played at best a minimal role in affecting the human condition.

Throughout the late nineteenth and early twentieth centuries, concern grew among the academic community that contemporary society – through charity and welfare programs – was 'tampering' with the laws of nature; that 'unfit' people were being allowed to survive and reproduce; that 'lives not worth living' constituted an unbearable financial burden on national and local budgets. These political valuations of life identified a grievous and imminent social problem and demanded an immediate solution. Fueled by Darwinian concepts of sexual selection and evolution, competing concepts of genetics, Malthusian perspectives on overpopulation, and environmental determinism, widespread fears of racial and societal degeneration permeated both popular and scientific venues. These fears were augmented by a growing – and scientifically documented – understanding of demographic changes. In the United States, for example, long-established patterns of immigration were drastically changing. Prior to the 1880s immigration was dominated by northern and western European nations; in the decades surrounding 1900, however, the main source areas shifted to southern and eastern Europe and, increasingly, Asia. Citing these changes, the labor historian and university professor John Commons (1907: 70) warned that "the change is one that should challenge the attention of every citizen."

Threats to the body-politic of the United States were not simply external. Following the abolition of slavery a moral panic ensued over the moral presence and economic position of newly freed blacks in American society. Thus, in the wake of the American Civil War, Southern politicians implemented a series of laws to retain blacks as docile, landless laborers. Known as the *Black Codes*, these laws were designed to affect white social control over black occupational, social, and geographic mobility. And while varying by state and by more local jurisdictions, the Black Codes generally denied freed slaves the right to marry whites, bear arms, or to assemble after sunset (Hummel, 1996). Unfair labor arrangements were also instituted and, consequently, African-Americans often became debt-bound to their former slave-owning masters.

Throughout the late nineteenth century – but especially during the early twentieth century, the problem of the African-American found its most forceful crystallization in the eugenic-inspired fear of sexuality. In part, (white) scientists reasoned at the time that slavery had been beneficial – from an evolutionary standpoint – to the black. Slavery, in fact, was re-presented as a form of genetic welfare, a viewpoint often reflecting a misappropriation of Darwinian thought and the process of natural selection. George T. Winston (1901: 108), President of the College of Agriculture and Mechanic Arts in North Carolina, for example, argued that slavery "transformed the Negro so quickly from a savage to a civilized man." H.E. Belin (1908: 517), writing in the *American Journal of Sociology*, concurred: "Slavery, so far from degrading the Negro, has actually elevated him industrially, mentally, and morally, the term of his involuntary tutelage [slavery] to the white race raising him to a vastly higher level than that ever occupied by his kinsmen in Africa." According to these experts, however, the abolition of slavery threatened to curtail this evolutionary improvement. The black was, on this account, retrogressing to his or her 'natural' and 'uncivilized' condition. Belin (1908: 519) described this transformation as 'frenzied,' noting that blacks who "but one brief hour before, were laughing, chattering, peaceable members of the community, are subject at any moment to be converted by some trivial occurrence into fierce, howling, blood-thirsty savages."

In short, academics and other so-called experts maintained that African-Americans, now freed from the 'beneficence' of slavery, posed direct threats to the well-being of American society. These threats, moreover, were often sexual in nature and revealed a stark-gendered component. African-American *men* posed a threat to the purity of the white race, through their supposed and unquenchable desire for virginal white women. African-American *women*, on the other hand, posed a threat also by their supposedly hypersexuality and hyperfertility. Combined, these discourses contributed to an overall image of 'uncivilized' and 'uncontrolled' blacks who retained no positive *productive* capability but considerable negative *reproductive* potential (Mencke, 1976; Hodes, 1997; Tyner and Houston, 2000).

Ongoing demographic and family studies revealed that immigrants and African-Americans exhibited higher fertility rates. If left unchecked, these reproductive trends allegedly would lead to an overall decline in the proportion of the 'original' Nordic stock of the United States. Numerous – and widely circulated – texts of the period attest to these fears: Madison Grant's *The Passing of the Great Race* (1916) and *The Conquest of a Continent* (1933), Lothrop Stoddard's *The Rising Tide of Color Against White World-Supremacy* (1920), and James Herbert Curle's *Our Testing Time: Will the White Race Win Through?* (1926).

For many 'social Darwinists' concern was raised that welfare and charity programs were tampering with the 'bloody hand' of evolution; hence, rather than succumbing to nature's law of the 'survival of the fittest,' misguided philanthropy – including minimum wages, set working hours, free public education, and public health reforms – was enabling inferior races to live longer and reproduce. Progress in the natural and social sciences – coupled with a proper political will – was required to understand both the implications of human interference on natural selection and, if possible, to provide viable methods to promote a more healthy, vital population. What was required, in practical terms, was a biopolitical control over the three fundamental processes of population change: migration, fertility, and mortality. What emerged was the eugenics movement.

As Ekland-Olson (2012: 21) writes, the eugenic argument was attractively simple: Would it not be better to have a society enriched by those who are productive, healthy, emotionally stable, and smart, than one stifled by degenerate, feeble-minded, disabled, and criminal citizens? Framed as a concern over differential worth, the eugenics movement effectively combined sexist, racist, classist, and nationalist sentiments. Indeed, as Saul Dubow (1995: 101) writes, eugenics coincided with the rising intensity of imperialist feelings, which also helped augment nationalist fervor and provide a convenient rational for the colonial subjugation of non-European peoples. Thus, as a *globalizing* paradigm, eugenics sought not only to identify differences within the human population, but also to evaluate and calculate the potential contribution of racially defined groups toward the betterment of society. As explained by the eugenicist L.H.M. Baker (1912: n.p.), "When we have ascertained … the qualities we want to preserve and the characteristics we desire to eliminate we must be courageous in the application of our remedy." From the United States to Germany, from Japan to Mexico, an array of segregation policies, immigration legislation, relocation schemes, antimiscegenation laws, sterilization, and euthanasia programs emerged as tangible – and viable – instruments in the practice of statecraft (Burleigh, 1994; Kühl, 1994). These racist practices would find their apogee in the Holocaust.

Late Twentieth- and Twenty-First-Century Race and Racism

After the atrocities of the Nazis the biological classification of race has become discredited; earlier theories of race and racial difference have been demonstrated to be themselves racist. In the years after World War II it became increasingly problematic to justify racial exploitation through pseudoscientific means. This was both a result of the atrocities of the Nazis, which justified the genocide of millions of Jewish and other European minorities, through scientifically grounded understandings of race and racism, and also because of the increasingly globalized struggle to overthrow colonial projects. The practices of colonialism came under increased pressure as global antiracist and anticolonial movements gained visibility and exposed contradictions in the old racial order (Inwood, 2013: 2122). Instead, we have now come to understand that race is not a "biologically given, but rather a socially constructed way of differentiating human beings" (Omi and Winant, 1994: 65) that nonetheless has real and meaningful consequences for society. The question then becomes, how do we come to understand race and why do we continue to live with and through the consequences of racism? The simple answer is because the discourse of race (and racism) is not profitable, it continues to reinscribe hierarchies that reinforce and retrench historic practices of exploitation. How, one might ask? While the answer is complex the social sciences have increasingly turned to questions of 'systemic racism' to address the continuing significance of race.

For the past quarter century understandings of race and racism, particularly in a North American context, have been dominated by the work of Omi and Winant (1994) and their seminal work on racial formation. Omi and Winant argue that to understand race one must focus on the link between "structure and representation" the way "racial projects do the ideological work of making links" between these two positions (1994: 56). They argue we should view race and racial formation as a sociohistorical process in which race and concomitant racial categories are created, transformed, or destroyed depending on contemporary political and economic necessity. Critical for understanding their argument is the way racial ideology is dialectically linked to social structure and the ways these relations come "to shape the nature of racism" (Omi and Winant, 1994: 74).

Recently this approach to racial formation has been challenged by critical race scholars who argue that we need to place racism more centrally into analysis of race (Chari, 2008). This argument has been made most forcefully, and some would argue persuasively, by the work of Joe Feagin and Sean Elias, among others (cf Feagin, 2006; Feagin and Elias, 2013), who argued that we need to focus more intently on 'systemic racism approaches.' Systemic racism critiques argue

> the past and present racial hierarchy and perpetuation of unequal socio-economic power relations among different groups are viewed as endemic to a race-based social system involving much more than conceptual meanings, ideologies and biased actions. In the case of US society, systemic racism is foundational to and engineered into its major institutions and organizations.
>
> Feagin and Elias, 2013: 936

Systemic racism approaches begin from the premise that racism and racial stratification have operated for centuries and, while sociospatially contingent (Pulido, 2006; Chari, 2008), have consistently reinscribed historic racial hierarchies. Systemic racism approaches argue that it is only through the consideration of racial oppression that we can begin to understand the true nature of race (Golash-Boza, 2013). In so doing a 'diverse assortment of racist practices' becomes readily evident including the unjustly gained economic and political power of whites; continuing resource inequalities; the rationalizing white racist frame; and the major institutions created to preserve white power and advantage (Feagin, 2006).

What is central to this approach is that the focus of critical race scholars needs to be on the ways the "recurring and unequal relationships between groups and individuals" (Pulido, 2006: 13) are continuously made and remade through time and space. In other words, and what makes systemic racism approaches relational to critical geography, is the way racism is understood as being dialectically linked to other social, economic, and political systems in the US (and elsewhere) and the way those systems perpetuate white racial rule (Peake and Schein, 2000). In taking such an approach we can more fully come to terms with the reality that white supremacy is a permanent fixture of life in many parts of the world. The continuance of white supremacist and neo-Nazi groups in France, Germany, Italy, and the United Kingdom is testimony to the perniciousness of an inherited legacy of virulent forms of racism and racial violence (cf Blee, 2005).

Racism and racial violence however need not be so apparent. Over the last quarter of a century there has been

a transformation of racism to subtler and more covert manifestations of white supremacy (Coates, 2008; Finney, 2013). Such practices have in large part been aligned with the penetration of neoliberalism. As Berg (2012: 511) argues, neoliberalism "produces an epistemological space in which humans are all the same (the so-called 'level playing field') and – ironically – where to even raise the specter of 'racial difference' is itself read as an act of racism." Berg (2012: 511) continues that the attempt to erase 'race' has a long history; what is distinctive to neoliberalism is that "certain forms of difference are highlighted, but others are effaced." In the maelstrom of neoliberalism, consequently, has emerged the discourse of 'postracialism' and the concomitant claim that North American and European societies are becoming 'race-blind' or 'color-blind' (Haney-López, 2010). Within the United States, the recently enacted Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act of 2009, signed into law by Barack Obama, America's first African-American president, seems to confirm what a number of political pundits – both liberal and conservative – have been arguing: that American society has entered into a postracial era. The term 'postracial,' in this context, encompasses the idea that if racial attitudes today are the same as they were in the last few decades, then it would have been impossible, for example, that the Hate Crimes Prevention Act would have been signed into law; or that Obama would have been elected. According to this view, therefore, the United States has finally transformed from a place where racial slavery and segregation were legally sanctioned and commonplace practices, to a place in which the American people could put aside their racial prejudices and unite to elect a president of African origin (Young, 2012: 501).

Despite such claims, or perhaps because of these claims, the malignant effects of racism have not abated. If anything, racism has intensified – albeit disguised in part by coded language that is not necessarily amenable to either empirical study or criminal prosecution (cf Coates, 2008). Perhaps as a result over the last three decades critical race theory (CRT) has become a dominant paradigm in the study of race and racism (Price, 2010).

Critical race studies is related to the development of CRT that emerged during the 1970s from the work of legal scholars Derrick Bell and Alan Freeman who were frustrated over the 'slow pace of racial reform' and wanted to understand the way race was reproduced in the US legal system (Delgado and Stefancic, 2000: xvi). Since that time, CRT has evolved to encompass a variety of academic disciplines and incorporates a wider critique of the processes involved in sustaining systems of racial oppression in the United States (Lynn et al., 2002). CRT leads to a shift in our understanding of racism from "willful acts of aggression" to "discussions of [the ways] race and racism are deeply embedded within the framework of American society" (Parker and Lynn, 2002: 8). CRT posits that "racism is normal, not aberrant in American society" and argues that current efforts at addressing racial injustices only "remedy the more extreme and shocking forms of injustice," noting that more subtle forms of racism are almost impossible to address using current legal and economic solutions (Delgado and Stefancic, 2000: xvi). The overarching goal of CRT is to "expose the historical, ideological, psychological, and social contexts in which racism has been virtually declared to be eradicated" and instead to show how racism is still a fundamental organizing principle in society (Parker and Lynn, 2002: 10). Those who adopt a CRT approach frequently (1) recognize that race is a social construction and argue for the eradication of racism; (2) draw important relationships between race and other axes of domination, thus illuminating the way racism links with other forms of injustice; and (3) present narratives and other qualitative approaches as valid approaches through which to examine race and racism (Parker and Lynn, 2002: 10).

Critical race studies integrate concepts from CRT through an integrative, qualitative methodology. Critical race methodology (1) foregrounds race and racism in all aspects of the research process; (2) challenges traditional research paradigms used to explain the experiences of persons of color; (3) offers transformative solutions to racial, gender, and class subordination; and (4) focuses on the *experiences* of persons of color (Solorzano and Yosso, 2002: 24). Critical race methodology seeks also to integrate these goals by emphasizing counter-narratives (Milner, 2008; Solorzano and Yosso, 2002; Delgado, 2000). Montecinos (1995) for example argues that the ideology of racism works through the use of master narratives which justify and maintain systems of oppression. Nash (2003: 639) similarly argues that research informed by a critical race studies perspective offers the field of geography a number of useful research tools to address the "multi-scaled processes that shape the lives of racialized groups" and invites the "bridging of perspectives across economic, cultural, urban, social and political geography" (639).

See also: Critical Race Theory; Cultural Geography; Gender and Feminist Studies in Geography; Population Geography; Race: Ethnicity and Health; Social Geography.

Bibliography

Alderman, D., 1996. Creating a new geography of memory in the south: (re)naming of streets in honor of Martin Luther King Jr. Southeastern Geographer 36 (1), 51–69.
Baker, L.H.M., 1912. Race Improvement or Eugenics. New York, n.p.
Belin, H.E., 1908. A southern view of slavery. American Journal of Sociology 13, 513–522.
Berg, Lawrence D., 2012. Geographies of identity I: geography—(neo)liberalism-white supremacy. Progress in Human Geography 36, 508–517.
Blee, Kathleen M., 2005. Racial violence in the United States. Ethnic and Racial Studies 28, 599–619.
Bonds, Anne, 2013. Racing economic geography: the place of race in economic geography. Geography Compass 7 (6), 398–411.
Burleigh, Michael, 1994. Death and Deliverance: 'Euthanasia' in Germany, 1900–1945. Cambridge University Press, Cambridge.
Castree, Noel, 2005. Nature. Routledge, New York.
Chari, Sharad, 2008. Critical geographies of racial and spatial control. Geography Compass 2 (6), 1907–1921.
Coates, Rodney D., 2008. Covert racism in the USA and globally. Sociology Compass 2 (1), 208–231.
Commons, John R., 1907. Races and Immigrants in America. Macmillan, New York.
Crookshank, F.G., 1924. The Mongol in Our Midst: A Study of Man and His Three Faces. E.P. Dutton, New York.
Curle, James Herbert, 1926. Our Testing Time: Will the White Race Win through? George H. Doran, New York.
Delaney, David, 1998. Race, Place, and the Law, 1836–1948. University of Texas Press, Austin.

Delaney, David, 2002. The space that race makes. The Professional Geographer 54, 6–14.
Delgado, R., 2000. Story telling for oppositionists and others: a plea for narrative. In: Delgado, R., Stefancic, J. (Eds.), Critical Race Theory: The Cutting Edge, second ed. Temple University Press, Philadelphia.
Delgado, R., Stefancic, J., 2000. Introduction. In: Delgado, R., Stefancic, J. (Eds.), Critical Race Theory: The Cutting Edge, second ed. Temple University Press, Philadelphia.
Du Bois, W.E.B., 1935/1994. The Souls of Black Folk. Dover, New York.
Dubow, Saul, 1995. Scientific Racism in Modern South Africa. Cambridge University Press, Cambridge.
Edelman, Murray, 1977. Political Language: Words That Succeed and Policies That Fail. Academic Press, New York.
Ekland-Olson, Sheldon, 2012. Who Lives, Who Dies, Who Decides? Abortion, Neonatal Care, Assisted Dying, and Capital Punishment. Routledge, New York.
Finney, Carolyn, 2013. Brave new world? Ruminations on race in the twenty-first century. Antipode. http://dx.doi.org/10.1111/anti.12021.
Feagin, Joe, 2006. Systemic Racism: A Theory of Oppression. Routledge, New York.
Feagin, Joe, Elias, Sean, 2013. Rethinking racial formation theory: a systemic racism critique. Ethnic and Racial Studies 36 (6), 931–960.
Gilmore, Ruth Wilson, 2002. Fatal couplings of power and difference: notes on racism and geography. The Professional Geographer 54, 15–24.
Golash-Boza, Tanya, 2013. Does racial formation theory lack the conceptual tools to understand racism? Ethnic and Racial Studies 36, 994–999.
Goldberg, David, 2002. The Racial State. Blackwell, Malden, MA.
Grant, Madison, 1916. The Passing of the Great Race, or the Racial Basis of European History. Charles Scribner's, New York.
Grant, Madison, 1933. The Conquest of a Continent, or the Expansion of Races in America. Charles Scribner's, New York.
López Ian F, Haney, 2010. Post-racial racism: racial stratification and mass incarceration in the age of Obama. California Law Review 98, 1023–1074.
Hodes, Martha, 1997. White Women, Black Men: Illicit Sex in the Nineteenth-Century South. Yale University Press, New Haven, CT.
Hummel, Jeffrey R., 1996. Emancipating Slaves, Enslaving Free Men: A History of the American Civil War. Open Court, Chicago.
Heynen, Nik, 2009. Bending the bars of empire from every Ghetto for survival: the black panther party's radical antihunger politics of social reproduction and scale. Annals of the Association of American Geographers 99 (2), 406–422.
Holloway, Steve, 2000. Identity, contingency and the urban geography of race. Social and Cultural Geography 1 (2), 197–208.
Inwood, Joshua, 2013. Contextualizing the state mode of production in the United States: race, space, and civil rights. Environment and Planning A 45, 2120–2134.
Inwood, Joshua, Bonds, Anne, 2013. On racial difference and revolution. Antipode 45, 517–520.
Kevles, Daniel J., 1995. In the Name of Eugenics: Genetics and the Uses of Human Heredity. Harvard University Press, Cambridge, MA.
Kobayashi, Audrey, Peake, Linda, 1994. Unnatural discourse: 'race' and gender in geography. Gender, Place and Culture 1, 225–243.
Kobayashi, Audrey, Peake, Linda, 2000. Racism out of place: thoughts on whiteness and an antiracist geography in the new millennium. Annals of the Association of American Geographers 90, 392–403.
Kühl, Stefan, 1994. The Nazi Connection: Eugenics, American Racism, and German National Socialism. Oxford University Press, Oxford.
Lynn, Marvin, Tara J. Yosso, Daniel G. Solórzano, Laurence Parker, 2002. Critical race theory and education: qualitative research in the new millennium. Qualitative Inquiry 8, 3–6.
McKittrick, Katherine, 2006. Demonic Grounds: Black Women and the Cartographies of Struggle. University of Minnesota Press, Minneapolis.
McKittrick, Katherine, Woods, Clyde (Eds.), 2007. Black Geographies and the Politics of Place. South End Press, Cambridge, MA.
Mencke, J.G., 1976. Mulattoes and Race Mixture: American Attitudes and Images, 1865–1918. UMI Research Press, Ann Arbor, MI.
Miles, Robert, 1989. Racism. Routledge, London.
Milner, H., 2008. Disrupting deficit notions of difference: counter-narratives of teachers and community in urban education. Teaching and Teacher Education 24, 1573–1598.
Montecinos, Carmen, 1995. Multicultural teacher education for a culturally diverse teaching force. In: Martin, R. (Ed.), Practicing What We Teach: Confronting Diversity in Teacher Education. State University of New York, Albany.
Nash, C., 2003. Cultural geography: anti-racist geographies. Progress in Human Geography 27 (5), 637–648.
Omi, Michael, Winant, Howard, 1994. Racial Formation in the United States: From the 1960s to the 1990s. Routledge, New York.
Parker, L., Lynn, M., 2002. What's race got to do with it? Critical race theory's conflicts with and connections to qualitative methodology and epistemology. Qualitative Inquiry 8 (1), 7–22.
Paul, Diane B., 1995. Controlling Human Heredity: 1865 to the Present. Macmillan, New York.
Peake, Linda, Schein, Richard, 2000. Racing geography into the new millennium: studies of 'race' and North American geographies. Social and Cultural Geography 1, 133–142.
Price, Patricia L., 2010. At the crossroads: critical race theory and critical geographies of race. Progress in Human Geography 34, 147–174.
Pulido, Laura, 2006. Black, Brown, Yellow and Left: Radical Activism in Los Angeles. Univeristy of California Press, Berekely.
Saldanha, Arun, 2006. Reontologising race: the machinic geography of phenotype. Environment and Planning D: Society and Space 24, 9–24.
Slocum, Rachel, 2009. The embodied politics of pain in US anti-racism. ACME: An International E-Journal for Critical Geographies 8, 18–45.
Solorzano, D., Yosso, T., 2002. Critical race methodology: counter-storytelling as an analytical framework for education research. Qualitative Inquiry 8 (1), 23–44.
Stepan, Nancy L., 1991. 'The Hour of Eugenics': Race, Gender, and Nation in Latin America. Cornell University Press, Ithaca, NY.
Stoddard, Lothrop, 1920. The Rising Tide of Color against White World-Supremacy. Charles Scribner's, New York.
Sundstrom, Ronald R., 2003. Race and place: social space in the production of human kinds. Philosophy & Geography 6, 83–95.
Tucker, William, 1994. The Science and Politics of Racial Research. University of Illinois Press, Urbana.
Tyner, James A., Houston, Donna, 2000. Controlling bodies: the punishment of multiracialized sexual relations. Antipode 32, 387–409.
Wilson, David, 2006. Metaphors, growth coalition discourses and black poverty neighborhoods in a U.S. city. Antipode 28 (1), 72–96.
Winders, Jamie, Schein, Richard, 2013. Race and diversity, what have we learned? Professional Geographer published online.
Winston, George T., 1901. The relation of the whites to negros. Annals of the American Academy of Political and Social Science 18, 103–118.
Young, Donna A., 2012. Post race posthaste: towards an analytical convergence of critical race theory and marxism. Columbia Journal of Race and Law 1, 499–510.

Race and the Law

Richard Ford, Stanford Law School, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.T. Ford, volume 19, pp. 12684–12689, © 2001, Elsevier Ltd.

Abstract

This article examines the use and treatment of race in the legal systems of several countries, including the use of racial categories in case law, statutes, and constitutional texts and the influence of racial politics on the development of law. It addresses the historical use of racial criteria for immigration and naturalization, racial apartheid, and 'Jim Crow' laws. It also discusses the statutory and constitutional prohibitions against racial discrimination, legal and policy initiatives to promote racial integration, and contemporary debates surrounding affirmative action and racial categories. It compares and contrasts the laws of the various nations whenever relevant comparison is useful and will attempt within the limitations imposed by length to provide a snapshot of the development of the legal treatment of race worldwide.

Definition of the Field

The study of race and the law is primarily, but far from exclusively, occupied by scholars trained in law and working in law schools. In Canada and Europe, the field is more evenly divided among lawyers and social scientists trained in anthropology, sociology, and cultural studies.

Mainstream versus Critical Scholarship

Scholarship addressing race and the law is split between mainstream approaches and critical approaches. Mainstream approaches generally conceive of racial discrimination as a minor and isolated exception to the otherwise liberal and enlightened societies of the industrialized West. These scholars understand racial injustice as largely unconnected to other institutions and are therefore optimistic that racism can be eliminated 'surgically,' without requiring significant or widespread changes in society. This is not to say that mainstream scholarship is always sanguine about the prospect for successful antiracist policy, but it tends to look at such policy as an isolated policy intervention that need not affect the basic institutions of the market economy or the majoritarian political process. Critical scholarship, by contrast, understands racism as deeply implicated in the institutions and traditions of societies and is therefore accordingly both more radical in its prescriptions and more circumspect in its predictions of success (Bell, 1987; Crenshaw, 1995; Williams, 1991).

Because the definition of race is malleable, fluctuating, and contested, the study of race and the law encompasses not only 'race' as it is colloquially understood, but also issues of ethnicity, national origin, caste, and religious and cultural differences. The field could therefore be understood to include the study of ethnic and nationalist divisions in Asian countries, caste hierarchy in Hindu societies, tribal conflict in central Africa, and religious conflict worldwide. It implicates the study of nationalism and pannationalist movements, communities of diaspora, and indigenous peoples. Moreover, the field is deeply concerned with the ways that race intersects with issues of gender, sexual orientation, and class (Crenshaw, 1995; Harris, 1990; Thomas, 1995; Williams, 1991; Carbado, 1999; Banks, 2010).

Generally speaking, the study of race and the law encompasses two major projects. The first is a critique that demonstrates how legal institutions and formal laws produce and perpetuate racism and racial hierarchy. The second is the analysis of legal reform that might eliminate or minimize the effects of racial hierarchy. This article will address each in turn.

Law as a Cause of Racism

Legal scholarship addresses the various ways in which law has produced and continues to produce racial hierarchy. Law contributes to racism and racial hierarchy in at least three distinct ways. First, it codifies racial categories, which are then employed by both public and private actors to erect and maintain racial hierarchies. Second, law institutionalizes racial hierarchy by creating formal rules requiring or encouraging racial hierarchy: by segregating the races; by excluding the members of certain races from participation in social, economic, and political activities and institutions; or by legally codifying the subordinate status of one or more races. These first two types of legal action are commonly known as *de jure* racial discrimination because they explicitly employ or create racial categories by force of law. Third, law allows, encourages, and requires the maintenance of institutional practices and social customs that perpetuate racial subordination by entrenching the racial hierarchies of the past, even when no formal legal rule requires or promotes racial subordination on its face. This latter type of legal action is known as *de facto* racial discrimination because it in fact produces and enforces racial subordination, although it does not do so explicitly.

Racial Categorization

Many scholars have studied the development and use of racial classificatory schemes (Davis, 1991; Haney-Lopez, 1994; Omi and Wynant, 1994). Virtually all scholars agree that, historically, schemes of racial classification were not based on sound scientific theories and therefore do not reflect any 'real' biological differences between the races. Given the dubious scientific basis for any racial classification of human beings, it is not an exaggeration to assert that law not only enforced racial hierarchy, but it also (along with other social institutions and practices) in fact produced a specific type of racial difference.

The Ontological Status of Racial Categories

There is disagreement about the ontological status of racial categories and the centrality of law in producing racial categories. Some scholars believe that racial categories have a substance that predates or transcends legally enforced and encoded racial categorization. While few scholars would argue that races are well understood in terms of biological or genetic difference, many argue for or assume that cultural differences distinguish the races (Gotanda, 1995; Peller, 1995). The belief that racial categories describe an essence or positive (in the sense of 'positive science,' not necessarily in the sense of 'good') attribute of human beings – whether biological or cultural – is known as racial essentialism.

Other scholars argue that social practices, including legal codification, produced racial differences where there were no such differences before. These scholars contend race as not an attribute of individuals or groups but instead the effect of a social or legal discourse that ascribes difference to individuals and groups (Appiah, 1996; Haney-Lopez, 1994). They insist that without the discourse of race, there would be no races. The belief that race is not an attribute or essence of individuals or groups but is instead the effect of the social discourses that describe it and the social practices that make it relevant is known as antiessentialism or social constructivism.

The Legal Codification of Racial Categories

Legal historians have examined the development and transformation of racial categories and definitions. A vast literature details the variety of racial classificatory schemes employed in the USA, Europe, and South Africa. The contemporary literature is overwhelmingly critical of most racial categorization, both because of its opportunistic use of scientific data (or its use of questionable science to justify racial categories) and the oppressive use to which racial categories have been put.

There is, however, some dissension on the question of remedial or the so-called benign racial classification. A widespread understanding of constitutional law and a pervasive popular attitude hold that the legal codification of *any* racial classification is socially deleterious and morally indefensible (Thernstrom and Thernstrom, 1999). This view asserts that the law should be *race blind*. Many scholars who may not insist on strict color blindness nevertheless have serious misgivings about the use of racial classifications, even for remedial purposes (Carter, 1991). Others believe that racial classification is not per se undesirable and argue that remedial racial classification is the necessary antidote to centuries of oppressive racial classification. They assert that the law should be affirmatively *race conscious* (Crenshaw, 1995; Gotanda, 1995; Matsuda and Lawrence, 1997; Peller, 1995).

The Legal Enforcement of Racial Hierarchy

Along with the study of racial categorization, a core concern of the study of race and the law is the legal enforcement of racial hierarchy. Of course the two are closely related: racial categories developed hand in glove with policies of racial exclusion, segregation, disenfranchisement, and involuntary servitude. The study of racial hierarchy has traditionally distinguished between formally enforced racial distinctions, hierarchies, and exclusions – *de jure* racial practices – and the formal requirement and enforcement of practices that have the effect of establishing or perpetuating racial hierarchy – *de facto* racial practices.

De Jure *Racial Subordination*

Legal historians have detailed the legal mechanisms that historically have enforced and perpetuated racial hierarchy. Historians of the USA study the institution of racial hierarchy through chattel slavery (Tushnet, 1981) and the post-emancipation development of 'Jim Crow' laws, which required racial segregation in public institutions and accommodations such as school restrooms, drinking fountains, eating facilities, railroad coaches, and theater or stadium seating (Woodward, 1989). Similar if less comprehensive policies targeted Asian-Americans (Kang, 2008; Carbado and Moran, 2014) and Hispanics (Carbado and Moran, 2014). Scholars of South Africa study the emergence, transformation, and eventual dismantling of racial *apartheid*, a policy that codified a belief in black inferiority as explicit policy and required segregation of the races in virtually every aspect of political and social life (Goldberg, 1994). Comparative scholars analyze the difference and similarities of South African apartheid and American Jim Crow laws (Ford, 1999; Goldberg, 1994).

Much of the historical study of European racial policy takes place as part of the study of European colonialism (Appiah, 1993; Bhabha, 1994; Gilroy, 1991; Ruskola, 2013). Explicit colonial racial policy and the treatment of non-whites in Europe are understood to implicate the ideology and political economy of European colonial rule (Ruskola, 2013). Scholarship concerning contemporary issues addresses the treatment of migrants, religious minorities, and non-whites who trace their ancestry to former colonies. Comparative study looks at differences and similarities between the racial policies of different nation states; for instance, studying how the American antiracist policy has been shaped by the experience of slavery and Jim Crow while European antiracist policy has been influenced by the atrocities of World War II (Suk, 2007). The question of religious intolerance, especially with respect to Muslims, is a topic of significant concern; many commentators suggest that Muslims are 'racialized' in much of Europe and the United States (Scott, 2010).

De Facto *Racial Subordination*

Formal racial exclusion, segregation, and hierarchical classification are rare in the postindustrial mass democracies of today. They have been rejected by mass opinion, eliminated by legislation, and struck down by courts applying constitutional law and international law. Nevertheless, racial hierarchy persists as a significant feature of every society with a history of racial differentiation. Accordingly, the study of race and the law focuses a significant amount of attention on informal or implicit practices that effectively perpetuate racial hierarchy.

The focus on *de facto* racial subordination began as a response to racial practices in the postcivil war USA. Federal constitutional guarantees of emancipation from slavery, equal protection, and the right to vote prohibited certain explicit racial barriers (*see* Civil Rights; Civil Liberties and Human Rights; Rights: Legal Aspects). But many states, especially

those of the former confederacy, found mechanisms to effectively reestablish these racial barriers without explicit reference to race. For instance, although blacks had a formal right to vote, many states effectively eliminated the black franchise through poll taxes and literacy requirements that were informally waived or ignored for whites but strictly enforced against blacks. And despite the formal emancipation of black slaves, many blacks were effectively forced into perpetual indentured servitude through the notorious sharecropping system of Southern agriculture (Goluboff, 2010; Klarman, 2004).

Informal and *de facto* forms of racial subordination emerged again as the response to more comprehensive civil rights reforms of the Civil Rights Act of 1964. Significantly, the Civil Rights Act prohibited racial discrimination in employment and accommodation in private institutions such as hotels, restaurants, and theaters, and in the sale and rental of housing (*see* Civil Rights; Civil Liberties and Human Rights; Rights: Legal Aspects). Many of the private institutions covered by the Act sought effectively to reproduce racial exclusion or discrimination through less explicit means. Lawyers, activists, and scholars, armed with a knowledge of postcivil war history, were able to identify and critique these practices (Dudziak, 2000; Mack, 2012). The attack on *de facto* racial subordination and discrimination became a staple of civil rights litigation as scholars and lawyers worked to expose the variety of practices that effectively perpetuated racial subordination. Antiracist scholarship and legal practice have attacked a host of formally color-blind practices as effectively racially discriminatory, including electoral gerrymandering (Guinier, 1994), standardized tests (Matsuda and Lawrence, 1997), neighborhood schools in a context of racially segregated neighborhoods (Ford, 1995; Denton and Massey, 1993), restrictive land use regulations (Ford, 1995; C. Harris, 1995), informal racial steering in real estate sales, municipal boundary formation (Ford, 1995, 1999; Anderson, 2010), and informal racial discrimination in hiring and promotions (Wilson, 2010; Ford, 2008).

Antiracist Legal Reform

Scholarship that examines, analyzes, and advocates the development of antiracist legal reform encompasses a vast range of proposals and is based on a number of methodologies and ideological commitments. Many approaches to antiracist reform build on civil rights reforms of the 1960s in the USA, advocating the expansion of such reforms and or the adaptation of such reforms in other national contexts (*see* Civil Rights; Indigenous Rights; Sexual Orientation and the Law; Law and People with Disabilities for information about how these paradigms have been applied in other settings).

A significant split has emerged between mainstream and critical approaches. Critical approaches question the efficacy and the underlying ideologies of mainstream civil rights and liberal antiracism (*see* Critical Race Theory). Beginning from an analysis that sees the problems of racism as deeply imbedded in the structure of social practices and institutions – including law itself – critical approaches are skeptical of reform proposals that attempt 'surgically' to correct or ameliorate racial bias while leaving most institutions and social practices more or less unchanged (Bell, 1987; Crenshaw, 1995). Nevertheless, many critical approaches often build on mainstream reforms, arguing for their radical expansion (Matsuda and Lawrence, 1997; Williams, 1991). Others suggest that mainstream reforms do more harm than good and should be abandoned altogether in favor of more comprehensive changes (Tushnet, 1984). Still others argue that while mainstream civil rights can play an important part in reducing racial bias, they are both insufficient and subject to abuse (Kelman and Lester, 1997; Ford, 2008, 2011).

Integrationism; Separatism

Crudely put, antiracist legal reform is characterized by a split between two strategies and or normative commitments, one which favors the integration of the races and another which favors the development of distinct, autonomous, and separate communities (Peller, 1995).

Historically, racial hierarchy has been legally enforced through spatial quarantine: American Jim Crow laws enforced the physical separation of the races in public accommodations and state institutions; racial zoning and restrictive covenants have divided the metropolitan landscape into racial neighborhoods; South African apartheid enforced racial segregation on a comprehensive regional scale and imposed racially specific curfews in white-only jurisdictions. Spatial segregation not only had potent symbolic consequences but also facilitated the inequitable distribution of resources: white-only institutions were consistently better situated, better funded, and socially and politically privileged compared to those assigned to racial minorities.

Given this history, a significant approach to remediation has been to encourage the integration of the races in public and private institutions and residential environments. Many American civil rights reformers believed that antiracist legal reform should embody a positive preference for integration *qua* integration (Wilson, 2010). Others believed that integration was desirable only because it would undo or stymie the inequitable distribution of resources in a society. These two rationales for integration produced a powerful consensus that has dominated American civil rights discourse since the 1960s and which has been influential in many other national contexts (Sabbagh, 2007; Bleich, 2003).

At the same time, however, a competing view has held that the development of distinctive and empowered minority communities is essential to racial justice and equality (Peller, 1995). This view focuses on the need for political solidarity, social support, and cultural autonomy. Racial 'separatism' has grown in popularity as many racial minorities have become disillusioned with the limitations of mainstream civil rights and the difficulties and compromises entailed in integration.

Antidiscrimination

A dominant conceptual and practical approach to antiracist legal reform has been that of antidiscrimination. The antidiscrimination concept presupposes that racial hierarchy is the result of irrational prejudice or bigotry based on race. Therefore, legal reform should prohibit intentional racial

discrimination and practices and some policies that disparately impact members of racial minority groups. The anti-discrimination approach focuses on specific wrongful actions by identifiable decision makers: if no wrongful act can be identified, there is no legally cognizable or remediable harm (see Freeman, 1995; Equality and Inequality: Legal Aspects).

Antisubordination

An alternative approach to antiracist legal reform is that of antisubordination. The antisubordination approach views racism as a set of practices that have the effect of systematically reproducing racial hierarchy, even if no identifiable decision maker can be said to be racially biased. Antisubordination approaches therefore insist that practices that adversely affect racial minorities should come under legal scrutiny, even in the absence of evidence of intent or bias. More importantly, antisubordination approaches look for patterns and systems in which several distinct practices combine to produce regular, predictable, and systematic racial subordination even when each individual practice, viewed in isolation, might seem unproblematic (Crenshaw, 1995).

Implicit Bias

Social psychologists have argued that individuals may harbor and act on biases of which they are unaware (Banaji and Greenwald, 2013). Such 'implicit' or unconscious biases are prevalent in many societies and can lead to discrimination in employment, law enforcement, and other social interactions. Some legal scholars have argued that implicit bias has important implications for antiracist law reform (Krieger, 1995; Kang, 2012). Others question the social science underlying claims of implicit bias (Arks and Tetlock, 2004) and/or its importance for civil rights law (Banks and Ford, 2009).

Cultural Discrimination

One significant strand of antiracist politics seeks to advance a project of racial recognition or multiculturalism, emphasizing the distinctiveness of racial minority communities in terms of culture. This strand seeks legal reform prohibiting discrimination against the cultural practices of racial minority groups as well discrimination on the basis of race *qua* status (Gotanda, 1995; Peller, 1995; Kymlicka, 2001; Carbado and Gulati, 2013).

This approach can be understood as a logical extension of the antisubordination approach to antiracist legal reform: discrimination against cultural practices effectively harms members of racial minority groups and can be seen as a component of systematic racial subordination.

One necessary consequence of racial multiculturalism is that it requires a definition or widely shared understanding of the content of racial identity or a racial culture. Such understandings or definitions are often controversial, both in society at large and within the given racial group itself. Moreover, the idea that racial identity or culture exists in a static or stable form is itself controversial. Some scholars would reject the very idea of a static or widely shared understanding of racial identity or culture as mistakenly essentialist (see Section The Ontological Status of Racial Categories).

Criminal Justice

Especially in the United States, racial disparities in the administration of criminal justice has long been a topic of scholarly and political concern. Legal scholarship has focused on racial disparities in the application of capital punishment, length of sentencing, and frequency of conviction (Kennedy, 1998). American prisons have been described as a new form of Jim Crow segregation (Alexander, 2010) and advocates have called for significant reforms ranging from the repeal of drug laws that disfavor minority groups to protest movements urging juries to refuse to convict minority defendants (Butler, 2009).

Antiessentialist and Cosmopolitan Conceptions of Racial Identity

A strand of antiracist scholarship first articulated within postcolonial African, Caribbean, and Indian communities that approach racial identity as fluid and supple, yet also the source of rich political, social, and cultural affiliations. These approaches define racial identity not in terms of a biological or cultural essence, but instead in terms of identification (Appiah, 1996; Bhabha, 1994; Gilroy, 1991; Ford, 2005). According to one strand of this scholarship, members of racial minority groups share common 'passions' based on shared or analogous experiences of racial subordination (Gilroy, 1991). This approach dispenses altogether with any remnant of biological or inheritance notions of racial identity: members of what might traditionally have been several different 'races' can and do share a cosmopolitan racial identity that can transcend national boundaries, cultural traditions, skin color, and phenotype. This scholarship adopts an explicitly antisubordination approach to anti-racism, but rejects the multiculturalist focus on cultural traditions within distinctive racial or ethnic communities as traditionally defined (Ford, 2005). Recent scholarship looks to a future of more complex race relations as racial mixing and new forms of 'racialization' change the nature of racial identity, create new racial categories, and perhaps even alter the conventional understanding of the concept of race (Hochchild et al., 2012).

See also: Affirmative Action, Sociology of; Affirmative Action: Comparative Policies and Controversies; Affirmative Action: Empirical Work on Its Effectiveness; Civil Liberties and Human Rights; Critical Race Theory; Discrimination and the Law; Race: Genetic Aspects; Racism, History of; Racism, Sociology of; Segregation Indices.

Bibliography

Alexander, M., 2010. The New Jim Crow: Mass Incarceration in the Age of Colorblindness. The New Press, New York.

Anderson, M., 2010. Mapped out of local democracy. Stanford Law Review 62, 931.
Appiah, K.A., 1993. In My Father's House, Africa in the Philosophy of Culture. Oxford University Press, New York.
Appiah, K.A., 1996. Race, culture, identity. In: Appiah, K.A., Gutman, A. (Eds.), Color Conscious. Princeton University Press, Princeton, NJ, pp. 30–74.
Arks, H., Tetlock, P., 2004. Attributions of implicit prejudice or "Would Jesse Jackson 'fail' the implicit association test?". Philosophical Inquiry 15, 257–278.
Banks, R., 2010. Is Marriage for White People? Dutton, New York.
Banks, R., Ford, R., 2009. (How) does unconscious bias matter? Emory Law Journal 58, 1053.
Bell, D., 1987. And We Are Not Saved: The Illusive Quest for Racial Justice. Basic Books, New York.
Bell, D., 1992. Race, Racism and American Law. Little Brown, Boston.
Banaji, M., Greenwald, A., 2013. A Blindspot: Hidden Biases of Good People. Delacorte Press, New York.
Bleich, E., 2003. Race Politics in Britain and France: Ideas and Policymaking since the 1960s. Cambridge, Cambridge, UK.
Bhabha, H.K., 1994. The Location of Culture. Routledge, New York.
Butler, P., 2009. Let's Get Free: A Hip-Hope Theory of Justice. New Press.
Carbado, D., Gulati, M., 2013. Acting White: Rethinking Race in a Post-racial America. Oxford, New York.
Carbado, D., 1999. Black Men on Race, Gender and Sexuality: A Critical Reader. NYU Press.
Carbado, D., Moran, R., 2014. Race law cases in the American Story. In: Sarat, A. (Ed.), Civil Rights in American History and Politics. Cambridge.
Carter, S.L., 1991. Reflections of an Affirmative Action Baby. Basic Books, New York.
Crenshaw, K., 1995. Race, reform and retrenchment: transformation and legitimation in anti-discrimination law. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. New Press, New York, pp. 103–126.
Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), 1995. Critical Race Theory: The Key Writings that Formed the Movement. New Press, New York.
Davis, F.J., 1991. Who Is Black? One Nation's Definition. Penn St. Press, Philadelphia, PA.
Delgado, R., Stephanic, J. (Eds.), 1999. Critical Race Theory: The Cutting Edge, second ed. Temple University Press, Philadelphia, PA.
Denton, N.A., Massey, D.S., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.
Dudziak, M., 2000. Cold War Civil Rights: Race and the Image of American Democracy. Princeton University Press.
Ford, R.T., 1995. The boundaries of race. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. New Press, New York, pp. 449–464.
Ford, R.T., 1999. Law's territory (a history of jurisdiction). Michigan Law Review 97, 843–930.
Ford, R.T., 2005. Racial Culture: A Critique. Princeton University Press, Princeton, NJ.
Ford, R.T., 2008. The Race Card: How Bluffing about Bias Makes Race Relations Worse. Farrar, Straus and Giroux, New York.
Ford, R.T., 2011. Rights Gone Wrong: How Law Corrupts the Struggle for Equality. Farrar, Straus and Giroux, New York.
Freeman, A.D., 1995. Legitimizing racial discrimination through anti-discrimination law: a critical review of supreme court doctrine. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. New Press, New York, pp. 29–45.
Gates Jr, H.L. (Ed.), 1986. Race, Writing and Difference. University of Chicago Press, Chicago.
Gilroy, P., 1991. There Ain't No Black in the Union Jack: The Cultural Politics of Race and Nation. University of Chicago Press, Chicago.
Goldberg, D., 1994. Racist Culture. Blackwell, London.
Goluboff, R., 2010. The Lost Promise of Civil Rights. Harvard, Cambridge.
Gotanda, N., 1995. Our constitution is colorblind. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. New Press, New York, pp. 257–275.
Guinier, L., 1994. The Tyranny of the Majority: Fundamental Fairness in Representative Democracy. Free Press, New York.
Haney-Lopez, I.F., 1994. White by Law. New York University Press, New York.
Harris, A.P., 1990. Race and essentialism in feminist legal theory. Stanford Law Review 42, 581–616.
Harris, C., 1995. Whiteness as property. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. The New Press, New York, pp. 276–291.
Hochchild, J., Weaver, V., Burch, T., 2012. Creating a New Racial Order: How Immigration, Multiracialism, Genomics and the Young Can Remake Race in America. Princeton University Press, Princeton, NJ.
Kang, J., 2008. Dodging responsibility: the story of Hirabayashi v. United States. In: Carbado, D., Moran, R. (Eds.), Race Law Stories. Foundation, New York.
Kang, J., 2012. Implicit bias in the courtroom. UCLA Law Review 59, 1124.
Kelman, M., Lester, G., 1997. Jumping the Queue: An Inquiry into the Legal Treatment of Students with Learning Disabilities. Harvard, Cambridge.
Kennedy, R., 1998. Race, Crime and the Law. Vintage Books, New York.
Klarman, M., 2004. From Jim Crow to Civil Rights: The Supreme Court and the Struggle for Racial Equality. Oxford, New York.
Krieger, L., 1995. The content of our categories: a cognitive bias approach to discrimination and equal employment opportunity. Stanford Law Review 47, 1161.
Kymlicka, W., 2001. Politics in the Vernacular: Nationalism, Multiculturalism, and Citizenship. Oxford University Press, Oxford, UK.
Mack, K., 2012. Representing the Race: The Creation of the Civil Rights Lawyer. Harvard, Cambridge.
Matsuda, M.J., Lawrence III, C.R., 1997. We Won't Go Back: Making the Case for Affirmative Action. Houghton Mifflin, Boston.
Omi, M., Wynant, H., 1994. Racial Formation in the United States: From the 1960s to the 1990s, second ed. Routledge, New York.
Peller, G., 1995. Race consciousness. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. The New Press, New York, pp. 127–158.
Ruskola, T., 2013. Legal Orientalism: China, the United States and Modern Law. Harvard.
Sabbagh, D., 2007. Equality and Transparency: A Strategic Perspective on Affirmative Action in American Law. Palgave MacMillian, New York.
Scott, J., 2010. The Politics of the Veil. University Press, Princeton, NJ.
Suk, J., 2007. Equal by comparison: unsettling assumptions of antidsicrimination law. American Journal of Comparative Law 55, 1.
Taylor, C., 1994. The politics of recognition. In: Gutman, A., Taylor, C. (Eds.), Multiculturalism: Examining the Politics of Re-cognition. Princeton University Press, Princeton, NJ.
Thernstrom, S., Thernstrom, A., 1999. America in Black and White: One Nation, Indivisible. Touchstone Books, New York.
Thomas, K., 1995. Rouge et Noir reread. In: Crenshaw, K., Gotanda, N., Peller, G., Thomas, K. (Eds.), Critical Race Theory: The Key Writings that Formed the Movement. The New Press, New York, pp. 465–494.
Tushnet, M.V., 1981. The American Law of Slavery, 1810–1860: Considerations of Humanity and Interest. Princeton University Press, Princeton, NJ.
Tushnet, M., 1984. An essay on rights. Texas Law Review 62, 1363–1403.
Williams, P.J., 1991. The Alchemy of Race and Rights. Harvard University Press, Cambridge, MA.
Wilson, W.J., 2010. More than Just Race: Being Black and Poor in the Inner City. Norton, New York.
Woodward, C.V., 1989. The Strange Career of Jim Crow, third ed. Oxford University Press, New York.

Race: Ethnicity and Health

Nicelma J King, University of California, Davis, Davis, CA, USA
Carolyn B Murray, University of California, Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.R. Williams, volume 7, pp. 4831–4838, © 2001, Elsevier Ltd.

Abstract

This article reviews the relationships among ethnicity, race, and health, based largely on populations located in the United States. Key indicators of health, including life expectancy, infant mortality, and specific disease incidence, vary by race and ethnicity. While improved medical care efforts have contributed to dramatic increases in life expectancy, the health status gap between majority and some minority racial and ethnic groups continues. Racism and economic deprivation underlie many of the observed racial variations in health. Understanding the interrelationships of race, ethnicity, and health requires identifying health-enhancing and health-damaging characteristics linked to socioeconomic status, race, and ethnicity.

This article provides an overview of the relationships among ethnicity, race, and health, based largely on populations located in the United States. The available data indicate that key indicators of health, including life expectancy, infant mortality, and specific disease incidence, vary by race and ethnicity. While people with higher income levels and more education live longer and experience lower levels of infant mortality, racial differences remain even within these higher income levels. Numerous research studies lead to the inevitable conclusion that racial and ethnic statuses are conspicuously related to the prevalence of disease and death (Smedley et al., 2003; Murray and Woods, 2009). Although improved medical care and prevention efforts have contributed to dramatic increases in life expectancy in the United States, the health status gap between majority and some minority racial and ethnic groups remains unchanged (NCHS, 2007).

In racially stratified societies, institutional discrimination linked to both historical and present racism and economic deprivation has resulted in large racial differences in socioeconomic status (SES). In turn, these variations in SES underlie many of the observed racial variations in health. In addition, racism can directly affect health through racial bias in medical and other service agencies and institutions and lead to the development of stress associated with experiences of discrimination and the acceptance of negative societal stereotypes. Advancing our understanding of the role of race and ethnicity in health thus requires better data that can identify the ways in which health-enhancing and health-damaging characteristics linked to SES, race, and ethnicity relate to each other and combine to affect health.

What is Race and Ethnic Status?

Historically, race has been viewed as capturing biological variation within human populations. However, the concept has never rested on a firm, scientific foundation. The delineation of racial taxonomies predates our understanding of genetics (Keita et al., 2004; AAPA, 1996). Thus, race emerged as a sociopolitical construct, useful to justify the exploitation of groups defined as inferior as well as to classify phenotypic variations in the population (Perry, 2011). Historically, the conceptualization of race has been influenced by cultural and political considerations, and race has served important ideological functions in society (AAPA, 1996). For example, medical research in the nineteenth century attempted to document that blacks were biologically inferior to whites and therefore more susceptible to a broad range of diseases. Importantly, such analyses provided a scientific justification for slavery and economic exploitation (Harawa and Ford, 2009).

Understanding and addressing the role of race and ethnicity in health requires a clear understanding of what race and ethnicity are, as well as the identification of the specific risk factors and resources linked to racial and ethnic status that may be responsible for variations in health. A growing body of evidence indicates that the traditional conceptualization of race is flawed. Scientists at the National Institutes of Health, after writing a draft of the entire sequence of the human genome, unanimously declared, that there is only one race – the human race (Angier, 2000; NCHPEG, 2014). Our current racial categories are more alike than different in terms of biological characteristics and genetics, and there are no scientific criteria to classify the human population unambiguously into discrete biological categories with rigid boundaries. Thus, racial taxonomies are arbitrary, and race is more of a social construct than a biological one.

Although ethnic groups can share a range of phenotypic characteristics due to their shared ancestry, the term 'ethnicity' is typically used to highlight cultural and social characteristics rather than biological characteristics. Ethnic groups share some combination of common geographic origins, family patterns, language, common values, cultural norms, religious traditions, literature, music, dietary preferences, and employment patterns. Thus, the current 'racial' categories often capture the essence of ethnic variation that is reflective of common ancestry, history, and/or culture.

Some writers have suggested that the term 'ethnicity' should replace 'race' in health research, and the term race should be abandoned (Afshari and Bhopal, 2002). However, although not useful as a biological category, race is an important social category. It is what sociologists call a master status – a central determinant of social identity and obligations, as well as of access to societal rewards and resources. Given that racial categories have reflected racism and its consequences (oppression, exploitation, and inequality) historically, collecting the appropriate data on race can facilitate ongoing monitoring of the magnitude of differentials and enhance understanding of

their causes and the development of effective interventions to address them.

Relatedly, even though variations in health are evident for many ethnic groups, much of the marked variation is evident for minority racial and ethnic groups. The term 'minority' is not used to represent the numerical size of a population, but rather the historical castelike status and lack of access to power and resources of that group. The current 'minority' groups in the United States, for example, reflect the intersection of racial and ethnic status combined with economic disadvantage. The groups that experienced discrimination historically and have been unsuccessful in assimilating are often termed 'minority' groups.

The Challenge of Analyzing Health Data

Data from the United States also suggest that there is considerable inconsistency in reporting racial, and especially ethnic, status over time. Given that many persons are multiracial or multiethnic in ancestry, self-identity appears to be importantly linked to the particular context of measurement. Categories of race and ethnicity are often assessed differently across societies, over time within particular societies, and by different agencies and institutions even within the same society. In the United States, for example, the standard birth and death certificates were revised in 1989 to include identifiers for Hispanic status. But the wording of the question assessing Hispanic status varies from state to state, and the data on Hispanic ethnicity is not completely reported in all states. Similarly, many health surveys and federal data collection systems in the United States use one question to assess race and a separate one to assess Hispanic origin, while other agencies use one combined question to assess both racial and ethnic status. Importantly, differing wording of racial and ethnic questions will provide varying estimates of the size of racial/ethnic populations.

Concomitantly, there is considerable debate over the optimal terminology to be used for particular racial groups, the number of racial groups for which data should be collected, the boundaries between racial and ethnic status, and whether new categories should be developed for persons who are of mixed ancestry. For a growing number of persons in society, racial and ethnic status is not singular and static, but multiple and dynamic. For example, current US census data uses the term 'Hispanic' to describe persons who identify themselves as 'Latino,' 'Mexican,' 'Mexican-American,' 'Chicano,' 'Cuban,' 'Puerto Rican,' and South or Central American, as well as persons who consider themselves to be of another Hispanic, Latino, or Spanish origin. Hispanics can be of any race, and prior to the 1970 census in the United States, were all categorized either as 'white' or as any other racial designation they chose. In addition to these dynamic classifications, the fastest growing racial designation in the United States is 'mixed race' (Associated Press, 2009).

Racial and Ethnic Variations in Health

Across a broad range of societies, racial and ethnic statuses predict variation in health. The United States has routinely reported health statistics by race, and the role of race and ethnic status is readily evident in US data. National mortality data reveal that blacks have higher death rates than whites for 4 of the 15 leading causes of death in the United States (NCHS, 2010). Cardiovascular disease is the number one cause of death for all groups except Asians, for whom cancer is the most frequent cause of death. Blacks have higher rates of death than whites from diabetes, kidney disease, infection, and human immunodeficiency virus, and the highest rate of death by homicide. Hispanics and American Indians have lower mortality rates than whites for cardiovascular disease and cancer, but higher rates for some causes of death, such as diabetes and liver disease. In contrast, the Asian/Pacific Islander population has higher mortality rates from cancer and cerebrovascular disease than other groups, but slightly lower mortality rates than whites or blacks from cardiovascular disease (NCHS, 2010, 2012).

Internationally, there are marked differences in health status between developing countries and economically developed countries. In part, it appears that these differences represent disparities in access to health care, as well as access to adequate and healthy food and water. The research associated with the Preston curve, which demonstrates a strong relationship between life expectancy and the level of economic development, reveals that even advances in medical technology do not eliminate the differences between rich and poor countries (Bloom and Canning, 2007). **Table 1** illustrates the differences in life expectancy and infant mortality rates in several countries from the developed and developing worlds. On the whole, developing countries have shorter average life expectancies and higher infant mortality rates than economically developed countries.

Very few countries collect health statistics on their constituent ethnic groups. For example, until recently, the United Kingdom did not focus on racial/ethnic variations in health. However, analyses of death rates by country of birth suggest that there may be important racial/ethnic variations in health status in England (Balarajan and Raleigh, 1995). These patterns are complex and vary by both the specific ethnic group and health outcome under consideration. Another example is that the death rate from coronary heart disease and stroke is higher

Table 1 Life expectancy and infant mortality rates per 1000 in selected countries in the developed and developing world

Country	*2012 Male/female life expectancy*	*2012 Infant mortality rate*
United States[a]	76.0/81.0	5.9
France	78.3/84.7	3.3
Japan	80.6/87.4	2.2
United Kingdom	78.5/82.4	4.5
Canada[a]	78.8/84.2	4.8
Mexico	72.9/77.6	16.8
India[a]	66.8/68.3	46.1
Botswana	56.9/54.5	10.5
Brazil[a]	69.2/76.5	20.5
Egypt	70.3/75.7	24.3

[a]Indicates a country with a substantially (25% or more) racial or ethnically diverse population.
Source: Indexmundi.com (accessed 28.05.12.).

than the national average for persons from the Indian subcontinent and the African commonwealth. The experience of the United States and England is not unique. In a comprehensive review of racial and ethnic differences in disease internationally, Polednak (1989) reveals that racial and ethnic status predicts variations in health across a broad range of societies.

Socioeconomic Status: A Central Determinant of Health

Although racial and ethnic status does not tell us much about underlying biological risks for disease, they often reflect SES. SES has traditionally been defined as a composite of education and income, which are in turn closely correlated. Low levels of education are correlated to lower income levels, and persistently low income levels are most often found among ethnic and racial minorities. In many, though not all cases, low-SES populations experience poorer health outcomes, including lower life expectancy, higher infant mortality, lower rates of health insurance or access to health care, and higher rates of death from particular illnesses. Thus, racial and ethnic differences in SES are central to understanding health status variations.

Disadvantaged racial/ethnic groups, such as the Maori in New Zealand, blacks in South Africa, and aboriginal populations in Australia, have higher rates of disease than the more advantaged groups within their societies. For example, in South Africa, despite the dismantling of the apartheid system, the infant mortality rate was three times higher for blacks and 2.3 times higher for Coloureds than for white infants (Burgard and Treiman, 2006). There are a few exceptions; Muslim babies in India have a lower infant mortality rate than the general (mostly Hindu) population, even though the Muslim population has considerably lower incomes and educational levels (Dutta and Chandrashekhar, 2012).

Racial Differences in Socioeconomic Status

There are important differences in power and SES across major racial and ethnic groups in the United States and in other countries. In the United States, the relationship between income, race, and education is clear, as shown in **Table 2**. As the table indicates, groups with fewer bachelor's degrees have lower incomes, and Asians, who have the highest percentage of persons holding at least a bachelor's degree, have the highest income levels. However, within the United States, some Asian subgroups such as the Hmong, Laotians, and Cambodians, have lower levels of income and higher rates of poverty than the black or the American Indian populations. This highlights the heterogeneity that exists within many of the major racial and ethnic categories, and the importance of attending to the variation within the subgroups of which large racial and ethnic categories are composed. For many racial/ethnic groups, an overall statistic masks the variation in health status and outcomes that exists for subgroups within that population. In addition, racial differences in income understate the true magnitude of racial differences in economic circumstances.

Table 2 Median family income and percent with at least a bachelor's degree of racial and ethnic groups in the United States in 2009

Race/ethnicity	*Income*	*% With at least a bachelor's degree*
All races	60 088	31
White	104%	35
Black/African-American	64%	21
Hispanic/Latino	66%	15
American Indian	63%	13
Asian	124%	51

Source: United States Census, 2012. Money Income of Families – Median Income by Race and Hispanic Origin, Detailed Years of School Completed by Persons 25 and Older by Sex, Race and Hispanic Origin. American Fact Finder.

In US census data, income captures the flow of economic resources into the household. In contrast, wealth reflects the accumulated economic reserves that can cushion a decline in income and facilitate long-term planning. Racial/ethnic differences in wealth are considerably larger than those for income. For example, in 2009 in the United States, the median net worth of white families ($113 149) was more than 20 times larger than that of black families ($5677), and almost 18 times larger than that of Latino families ($6325) (Kochhar et al., 2011). Moreover, measures of wealth make an incremental contribution in predicting variations in health over and above measures of education and income (Williams and Collins, 1995).

Race, SES, and Health

Research reveals that SES differences between the races account for many of the racial/ethnic differences in health. For example, adjusting black/white disparities in health in the United States for SES substantially reduces these differences. A 2012 study (Geruso, 2012) found that income alone explained 52% of the difference between the life expectancy of black males and white males, and 59% of the difference between black females and white females. Education alone, which is closely correlated to income, explained 26% of the difference between the life expectancy of black males and white males, and 29% of the difference between black females and white females. In combination, the two factors explained 62% and 63%, respectively. Several important trends are noteworthy. For both racial groups, income is strongly linked to life expectancy. For most racial and ethnic groups, persons with lower levels of income have lower life expectancy than their more economically favored peers. In addition, the difference in health status by SES for both black and white men is larger than the overall racial difference. African-American men in the highest income group live 7.4 years longer than those in the lowest income group; the comparable number for whites is 6.6 years. The overall racial difference for men was 4.9 years. Although SES predicts variation in health for both black and white women, for women, the overall racial difference is somewhat larger than the differences between SES groups. The life expectancy data also reveal an independent effect of race, even when blacks and whites are comparable in income. At every level of income for both men

and women, African-Americans have lower levels of life expectancy than their white peers. This pattern has been observed across multiple health outcomes, and for some indicators of health, such as infant mortality, the racial gap becomes larger as SES increases.

There are exceptions; for instance, life expectancy for both Asian-Americans (87.3 years) and Latino Americans (83.5 years) is higher than that of white Americans (78.7 years; American Human Development Project, 2010/2011). While for Asian-Americans their longevity can be explained partially by their higher economic and educational status, as well as their diet and lifestyle, the longevity of Latino Americans is not so easily explained. Latinos economically have a similar profile when compared to African-Americans (74.3 years), and experience disparate medical treatment (Smedley et al., 2003), yet their group on average lives longer than even the more advantaged white group. In contrast, for whites and African-Americans, wealth predicts longevity, with the wealthier members of each group projected to live half a decade longer. Known as the 'paradox,' Latinos' SES has little bearing on their longevity. Researchers have offered several explanations for this phenomenon. In investigating the finding that Latinos have birth outcomes (i.e., low infant mortality, see **Table 3**) that are better than would be expected, some researchers propose that this has something to do with the 'healthy immigrant phenomenon,' which points out that people who immigrate are young and active and tend to be healthier than those who do not. Furthermore, newcomers are believed to have an ethic of good health habits. Another plausible explanation is that because Latino communities are composed of strong social support networks, including more extended family supports, this acts as an insulation that protects their health. However, the longer they stay within the United States and acculturate to the American lifestyle, their well-being, ambition, and health tend to deteriorate (Lamare, 2012). For instance, current rates of alcohol and substance abuse among Latinos are comparable to or surpass those of other US ethnic groups (Guerrero et al., 2013).

While the life expectancy for Asian-Americans is the highest among the four major ethnic/racial groups, the statistics are not based upon data disaggregated by ethnic subgroups and there is a scarcity of evidence-based interventions specifically adapted for and tested among Asian-American subgroups. Very little is known about disparities among the Asian-American subgroups (South Asian, Chinese, Filipino, Japanese, Korean, and Vietnamese). The lack of disaggregated data has led to overlooking the fact that the poverty rates for some Asian groups (e.g., Hmong and Cambodian Americans) approach 30%. A concomitant statistic is that some subgroups are unhealthier than others and have a much shorter life expectancy. For instance, it is estimated that Asian-Americans constitute close to 50% of the 1.4 million Americans chronically infected with hepatitis B, even though they constitute only 4% of the nation's population (CDC, 2008). Asian-Americans also have the highest race-specific rate of tuberculosis incidence (CDC, 2001) and underutilize preventive services such as colorectal, breast, and cervical cancer screenings when compared with non-Hispanic whites (Ward et al., 2004). Furthermore, many well-to-do Asian-Americans return to their native countries to receive nonurgent medical care. Thus, the present picture of Asian longevity is incomplete.

Table 3 Infant mortality rates in the United States per 1000 by race and Hispanic origin of the mother, 2005–07

Race/ethnic identity of mother	*Rate of infant mortality, 2005–07*
All races	6.8
White	5.7
Black/African-American	13.4
Hispanic/Latino	5.5
American Indian/Alaska native	8.5
Asian/Pacific Islander	4.7

Source: National Center for Health Statistics, 2012. Health, United States 2011. With Special Feature on Socioeconomic Status and Health. Hyattsville, MD, pp. 101–102, Table 18.

Disparate Health Services and Treatment

Improved medical care and prevention efforts have contributed to dramatic increases in life expectancy in the United States, particularly over the past century. At the same time, however, the health status of some racial and ethnic minorities lags far behind that of the majority population (NCHS, 2007). This situation is especially true for African-Americans and Native Americans who both, for a variety of reasons, including poverty and overt discrimination, experience significant health disparities in comparison to white Americans. Given the cost and long-term consequences of health disparities (e.g., decreased productivity, diminished quality of life), identifying the causes can be of utmost importance. Racial disparities in care tend to remain, even after controlling for patient differences in SES, insurance status, and access to a regular primary care provider (Johnson et al., 2004).

The findings from numerous studies that ethnic minorities receive inferior medical treatment are compelling. For instance, the National Cancer Institute Black/White Cancer Survival Study examined whether cancer treatment varied for white American and African-American women. They found that African-American women received less aggressive and less appropriate treatment than white American women (Breen et al., 1999). Although white American women have a higher incidence of breast cancer, they do not have the highest mortality rate; African-American women are more likely to die from breast cancer than white American women (Lantz et al., 2006). Income and education did not account for racial differences in the stage of breast cancer diagnosis, and racial disparities in early-stage diagnosis remained significant even after controlling for income and education. Even when controlling for disease stage, the risk of death is 50% higher for African-American women with breast carcinoma than for white American women (Bradley et al., 2001). While overt racism undoubtedly contributes to this disparate outcome, a report by the Institute of Medicine (IOM) on racial disparities in health care suggests that various aspects of the physician–patient relationship may be

a major cause of health disparities in the United States (Smedley et al., 2003).

Physician–Patient Issues

Physicians diagnose disease and recommend treatment alternatives, and even when patients desire to participate in their medical decisions, they typically depend upon their physicians for guidance, options, and appropriate care. This is especially true for patients who have lower levels of education, health literacy, and SES (Lillie-Blanton et al., 1996). Furthermore, research suggests that some physicians remain unenthusiastic about treating non-white patients, particularly African-Americans (Murray, 1998). For instance, research demonstrates that African-American patients receive less positive communication, less information, fewer treatment options, and overall less instrumentally and interpersonally effective care when compared with white American patients (Cooper-Patrick et al., 1999; Smedley et al., 2003). Furthermore, less positive effect, is central to patient recall of information, treatment adherence, and satisfaction with health care process and outcomes (Johnson et al., 2004).

There is also evidence that all four groups (African-Americans, Asian-Americans, Latinos, and Native Americans), in contrast to white Americans, may have difficulties developing appropriate patient–physician relationship and trust (Vamos et al., 2009). Recent research has addressed the effects of physician–patient concordance on the process and outcomes of medical care; concordance refers to the similarity or shared identity between physician and the patient in the delivery of health care. Most concordance research has examined similarities and differences in the demographic characteristics of race, gender, and age, and this research provides some evidence that medical visits in which physician and patient are similar may be somewhat different in process and outcome than those in which physician and patient are very different. For example, race-concordant medical visits (compared to discordant visits in which patient and physician are of different races) tend to be longer, with a more participatory partnership-oriented style (DiMatteo et al., 2009). Race-concordant care also results in higher patient satisfaction, greater patient trust in and access to the physician, greater trust in the health care system, and greater respect and listening in the process of communication (Cooper-Patrick et al., 1999; Cooper et al., 2003). Also, patients in racially discordant interactions received significantly less information and were significantly less active participants when compared with patients in racially concordant interactions. Controlling for patients' participation reduced the effect of concordance on information giving, suggesting that racial discordance might increase patient passivity and reduce information exchange. It is possible that minority patients in discordant interactions do less to prompt their physicians for information, resulting in physicians offering less information to these patients.

For minorities, the importance of the negative influence of physician communication or lack of communication in health outcomes abound. For example, prostate cancer, which is the leading cancer diagnosed among men in the United States, occurs about 60% more often in African-American men than in white American men, and when it is diagnosed, the cancer is more likely to be advanced (Prostate Cancer Health Center, 2013). A study (Woods et al., 2006) that investigated prostate cancer in black men reported that the men believed that if there were a potential problem, their health care provider would communicate the seriousness; however, few providers did so. Furthermore, they reported that traditional barriers generally found in the literature regarding black men and late health care access because of a lack of health insurance, not going to physician on a regular basis, lack of transportation, and not interested in their health were *not* supported (Woods et al., 2004). They concluded that the findings were consistent with the IOM report on health care access, i.e., those blacks and other minorities receive inferior health care compared with whites across a range of health conditions and procedures, even when insurance status, income, age, and severity of condition are comparable (Woods et al., 2004).

Psychosocial Causes of Health Disparities

Although numerous studies have shown that even when health status and coexisting illnesses are accounted for, minorities are still *less* likely (1) to be referred for evaluation, (2) to receive the 'gold standard' (i.e., most effective treatment available), (3) to be rated as appropriate candidates for organ transplantation, or (4) to receive a transplant, regardless of their age and income (Vamos et al., 2009). One documented explanation is direct doctor/institutional discrimination (Smedley et al., 2003). Some researchers, however, place the burden on minority patients, who are blamed for mismanaging their health by not adhering to the medical regimens, lacking the literacy to understand the protocols, and avoiding professional medical attention. For instance, Latinos are less likely to be insured or more likely to have suboptimal insurance, which therefore limits their ability to receive many health care services. African-American men are more likely to be unemployed and thus, if they seek health care, it is more likely to be for an emergency (Disparities Report, 2003). Even Asian-Americans, who have the highest life expectancy than any of the four major ethnic groups within the United States, in contrast to whites, contend with numerous factors, which may threaten their health. Some negative factors are infrequent medical visits due to the fear of deportation, language/cultural barriers, and the lack of health insurance (OMH, 2013).

Cultural Mistrust: An Added Burden for Nondominant Populations

'Trust' in their health care provider seems to be an issue for many minorities. Cultural mistrust is a construct that was initially conceptualized to describe the distrust among African-Americans of white Americans and mainstream American institutions including the legal system, political system, government agencies, educational system, health care system, employment settings, and other entities that are governed or staffed by white Americans (Terrell and Terrell, 1981). Recently, cultural mistrust of health care providers has been investigated with other minority populations. One study found that Filipino Americans who expressed higher levels of cultural

mistrust were less likely to seek professional psychological help. Furthermore, cultural mistrust predicted variance in mental health help-seeking attitudes that are not accounted for by income, generational status, loss of face, and adherence to Asian cultural values (David, 2010).

Another study that collected data using a stratified, random-digit dial telephone questionnaire of noninstitutionalized households in New York, Maryland, and Puerto Rico, reported that compared with whites, Latinos were nearly two times more likely to report that fear of being a 'guinea pig' and lacking trust in medical people made them unwilling to participate in cancer screening (Davis et al., 2012).

Also, Native Americans, in contrast to whites, who were newly diagnosed cancer patients, expressed significantly higher levels of mistrust and lower levels of satisfaction with their health care. In multivariable analyses, race was the only factor found to be significantly predictive of higher mistrust and lower satisfaction scores. Numerous studies have reported similar findings whether it is beliefs concerning being an organ donor (Breitkopf, 2009), adherence to recommended medical regiments (Shelton et al., 2011), or receiving genetic testing (Suther and Kiros, 2009). In each case, people of color, compared to whites, were significantly more mistrusting of health care providers, and thus, more likely to avoid interacting with health providers and less likely to adhere to recommended health regimes.

A recent study investigating the connection between health and stress (Hicken et al., 2013) found that African-Americans, but not Latinos, in contrast to whites, were more vigilant. This vigilance was positively associated with sleep difficulties in all three racial groups; however, it was attenuated by income and education, and completely attenuated after adjustment for racism-related vigilance. Given the importance of sleep quality to health, anticipation of and perseveration about racial discrimination appears to be an important determinant of racial disparities in health. Interestingly, given that Latinos were less vigilant when it came to racism and discrimination, this may be one explanation of why they live longer than their African-American counterparts.

Changes in Racial Inequalities in Health

Data from the United States suggest that patterns of racial inequality change over time in response to larger socioeconomic changes. For example, the Civil Rights Movement improved the economic circumstances of blacks and narrowed the black/white gap in income during the 1960s and the early 1970s. Correspondingly, mortality rates for blacks declined, both on a percentage and an absolute basis more rapidly than those of whites between 1968 and 1978 (Williams and Collins, 1995). Similarly, during this time period, the life expectancy increases for black males and females were larger on both a relative and an absolute basis than those for whites. However, the economic status of blacks relative to whites stagnated in the mid-1970s and worsened during the 1980s. Similarly, the black/white gap in multiple health indicators worsened during this decade. The black/white gap in life expectancy, infant mortality, and excess deaths increased between 1980 and 1991 (Williams and Collins, 1995). For example, for five consecutive years after 1984, the life expectancy at birth for blacks progressively declined, while that of whites continued to increase progressively.

The economic downturn following 2008 has also been associated with worse health outcomes in the United States, especially among economically disadvantaged minority groups. African-American, Hispanic, and Asian children have been reported as having a higher incidence of asthma in the years immediately following the 2008 downturn (Burgard, 2012). All groups reported increased psychological distress. These data are consistent with research, which shows that the health of populations is affected by a broad range of economic factors, including growth, stability, and inequality (Brenner, 1995). For example, mortality from all causes tends to increase about 2 to 3 years after the low point of a recession. These data highlight that both race and SES have powerful influences on health because they shape the experience of and exposure to almost all psychosocial and environmental risk factors for health.

Resources and Acculturation

Health-enhancing resources exist even in the most disadvantaged racial and ethnic populations. Much research on the health of disadvantaged groups has focused historically on pathology and vulnerability. However, the available health data clearly suggest that health-enhancing factors also exist within contexts of economic deprivation. For example, despite high rates of poverty and low rates of health insurance coverage, Mexican immigrants still have a relatively good health profile, with relatively low rates of infant mortality and longer life expectancy than many other groups. Similarly, although black Americans have lower levels of income and lower life expectancy, their rates of suicide are markedly lower than whites, and blacks tend to have equivalent or lower rates of psychiatric illness compared to whites. These data suggest the need to identify cultural strengths and health-enhancing resources that exist within disadvantaged communities; this may shield these populations from at least some of the negative exposures in their social context. For instance, research has found that quality of family and extended kin relationships, as well as religious participation are key health-enhancing resources that help to reduce some of the negative effects of stress on the health of disadvantaged populations. More research is needed on the mechanisms and processes by which these factors operate, and the ways in which they combine with existing risk factors to affect health.

In many societies, minority racial and ethnic groups are largely immigrant populations. The health profiles of these groups may reflect their immigrant status, and may change with increasing length of stay in the host society. In the United States, immigrants of all the major racial/ethnic groups have lower rates of infant and adult mortality than their native-born counterparts (Green, 2012; Williams, 1999). Importantly, these patterns persist even when immigrant groups, such as Mexicans in the United States, are lower in SES and have limited access to medical care. At the same time, other data suggest that, across a broad range of health status indicators such as infant mortality, low birth rate, cancer, high blood pressure, adolescent pregnancy, and psychiatric disorders, the health status of

immigrants worsens with length of stay in the United States (Williams and Collins, 1995). The mechanisms and processes responsible for these negative effects of acculturation are not fully understood. Several behaviors that adversely affect health status appear to increase with length of stay. It has also been suggested that there may be health-protective aspects of indigenous cultures and belief systems that are lost with increasing time in a new society.

Culture is a construct that is widely invoked in discussions of racial and ethnic variations in health. However, it is often treated as a residual explanation for unexplained differences in health. There is a critical need for the theoretical identification and the empirical assessment of the specific cultural beliefs and behaviors that are presumed to affect health. Moreover, researchers must avoid labeling as cultural what are more structural differences in the social position of different groups. Culture is dynamic and its development and persistence must be understood within the context of powerful reciprocal influences that are operative as social groups interact with the constraints and opportunities of their specific environment.

Concluding Recommendations

Future research on racial/ethnic variations in health should include an assessment of the considerable ethnic heterogeneity that exists within all the widely used racial categories. Researchers in this area must recognize that there is no generic health model that applies equally to all racial or ethnic minority populations, but that particular indicators of disease prevalence or health can be used comparatively if data has been carefully and consistently collected. Research and health policy regarding race/ethnicity should be attentive to the distinctiveness of each group, including using the terms that are broadly recognized and preferred by a substantial proportion of respondents of the groups being studied. The fluidity of racial and ethnic categories requires periodic monitoring and updating measures of racial and ethnic status. Importantly, whenever racial and ethnic data are reported, more attention must be given to the interpretation of these data. For example, it is always necessary to indicate why race or ethnicity is being utilized, the limitations of racial and ethnic data, and how the findings should be interpreted. Given the strong association between racial/ethnic status and SES, the presentation of data on racial differences should routinely stratify them for SES within racial groups. Failure to attend to these issues of presentation and interpretation may oversimplify complex health risks and even perpetuate negative racial stereotypes. Finally, future research must conceptualize and operationalize the specific factors and practices linked to race that might affect health.

See also: Cancer: Psychosocial Aspects; Coronary Heart Disease: Psychosocial Aspects; Culture as a Determinant of Mental Health; Diabetes: Psychosocial Aspects; Doctor–Patient Interaction in the West: Psychosocial Aspects; Explanatory Style and Health; Gender and Health Care; Health Education and Health Promotion; Health Insurance: Economic and Risk Aspects; Health Interventions, Community-based; Health Literacy; Migration and Health; Religion and Health: Clinical Considerations and Applications; Social Exclusion, Social Deprivation and Health; Socioeconomic Status and Health Care; Socioeconomic Status and Health.

Bibliography

AAPA, 1996. AAPA statement on biological aspects of race. American Journal of Physical Anthropology 101, 569–570.

Afshari, R., Bhopal, R.S., 2002. Changing pattern of use of 'ethnicity' and 'race' in scientific literature. International Journal of Epidemiology 31 (5), 1074.

American Human Development Project, 2010/2011. Mapping Risks and Resilience. Project of the Social Science Research Council. http://www.measureofamerica.org/wp-content/uploads/2010/11/AHDP-HEALTH-FACT-SHEET-11.08.10.pdf.

Angier, N., 22 August 2000. Do Races Differ? Not Really, Genes Show. The New York Times Science. http://www.nytimes.com/2000/08/22/science/do-races-differ-not-really-genes-show.html?pagewanted=all&src=pm.

Associated Press, updated 28.05.09. Multiracial America is Fastest Growing Group: That Demographic Now Makes up 5 Percent of the Minority Population. Race & Ethnicity. NBCNews. http://www.nbcnews.com/id/30986649/ns/us_news-life/t/multiracial-america-fastest-growing-group/#.UiUCeOAqwzk.

Balarajan, R., Raleigh, V.S., 1995/2009. Ethnicity and Health in England. Department of Health, London.

Bloom, D.E., Canning, D., 2007. Commentary: the Preston curve 30 years on: still sparking fires. International Journal of Epidemiology 36 (3), 498–499.

Bradley, C., Given, C., Roberts, C., 2001. Disparities in cancer diagnosis and survival. Cancer 91, 178–188.

Breen, N., Wesley, M., Merrill, R., Johnson, K., 1999. The relationship of socioeconomic; status and access to minimum expected therapy among female breast cancer patients in the; National Cancer Institute Black-White Cancer Survival Study. Ethnicity & Disease 9, 111–125.

Breitkopf, C.R., 2009. Attitudes, beliefs and behaviors surrounding organ donation among Hispanic women. Current Opinion in Organ Transplantation 14 (2), 191–195.

Brenner, M.H., 1995. Political economy and health. In: Amick III, S., Levine, B.C., Tarlov, A.R., Walsh, D. (Eds.), Society and Health. Oxford University Press, New York, pp. 211–246.

Burgard, S., 2012. Health, Mental Health and the Great Recession. The Russell Sage Foundation and the Stanford Center on Poverty and Inequality. http://www.stanford.edu/group/recessiontrends/cgi (accessed 07.03.13.).

Burgard, S.A., Treiman, D.J., 2006. Trends and racial differences in infant mortality; in South Africa. Social Science & Medicine 62, 1126–1137.

CDC (Center for Disease Control and Prevention), 2008. Asian Americans and Hepatitis B. http://www.cdc.gov/Features/AAPIHepatitisB/.

Cooper, R.S., Kaufman, J.S., Ward, R., 2003. Race and genomics. New England Journal of Medicine 348, 1166–1170.

Cooper-Patrick, L., Gallo, J.J., Gonzales, J.J., Vu, H.T., Powe, N.R., Nelson, C., Ford, D.E., 1999. Race, gender, and partnership in the patient-physician relationship. Journal of the American Medical Association 282 (6), 583–589.

CDC (Center for Disease Control and Prevention), 2001. Health Disparities Affecting Minorities: Asian Americans. http://www.cdc.gov/minorityhealth/brochures/AA.pdf.

David, E.J.R., 2010. Cultural mistrust and mental health help-seeking attitudes among Filipino Americans. Asian American Journal of Psychology 1, 57–66.

Davis, J.L., Bynum, S.A., Katz, R.V., Buchanan, K., Green, B.L., 2012. Sociodemographic differences in fears and mistrust contributing to unwillingness to participate in cancer screenings. Journal of Health Care for the Poor and Underserved 23 (Suppl. 4), 67–76.

DiMatteo, M.R., Murray, C.B., Williams, S.L., 2009. Gender disparities in physician-patient; communication among African American patients in primary care. Journal of Black; Psychology 35, 204–227.

Dutta, B., Chandrashekhar, 2012. Infant and Child Mortality: A Study of Selected South and South East Asian Countries. Chapter-07-Infant-and-Child-Mortality.pdf. http://hetv.org/india/nfhs/nfhs3/NFHS-3.

Geruso, M., 2012. Black-White disparities in income: how much can the standard SES variables explain? Demography 49, 553–574.

Green, T., 2012. Black and Immigrant: Exploring the Effects of Ethnicity and Foreign-Born Status on Infant Health. Migration Policy Institute. http://www.migrationpolicy.org/pubs/CBI-Green.pdf.

Guerrero, E.G., Marsh, J.C., Khachikian, T., Amaro, H., Vega, W.A., 2013. Disparities in Latino substance use, service use, and treatment: implications for culturally and evidence-based interventions under health care reform. Drug and Alcohol Dependence pii: S0376–8716(13).

Harawa, N.T., Ford, C.L., 2009. The foundation of modern racial categories and implications for research on black/white disparities in health. Ethnicity & Disease 19, 209–217.

HHS (U.S. Department of Health and Human Services), 2008–2010. Office of Minority Health.

Hicken, M.T., Lee, H., Ailshire, J., Burgard, S.A., Williams, D.R., 2013. "Every shut eye, ain't sleep": the role of racism-related vigilance in racial/ethnic disparities in sleep difficulty. Race and Social Problems 5, 100–112.

Johnson, R., Roter, D., Powe, N.R., Cooper, L.A., 2004. Patient race/ethnicity and quality; of patient-physician communication during medical visits. American Journal of Public; Health 94, 2084–2090.

Keita, S., Kittles, R.A., Royal, C., Bonney, G., Furbert-Harris, P., Dunston, G., Rotimi, C., 2004. Conceptualizing human variation. Nature Genetics 36, S17–S20.

Kochhar, R., Fry, R., Taylor, P., 2011. Wealth Gaps Rise to Record Highs between Whites, Blacks, Hispanics. Pew Research Social and Demographic Trends. Pew Foundation. http://www.pewsocialtrends.org/2011/07/26/wealth-gaps-rise-to-record-highs-between-whites-blacks-hispanics/ (accessed 03.07.13.).

Krieger, N., 1999. Embodying inequality: a review of concepts, measures, and methods for studying health consequences of discrimination. International Journal of Health Services 29, 295–352.

Lamare, J., 2012. Hispanic Link News Service. News Report. Jim. Posted: 02 December 2012.

Lantz, P., Mujahid, M., Schwartz, K., Janz, N., Fagerlin, A., Salem, B., et al., 2006. The influence of race, ethnicity, and individual socioeconomic factors on breast cancer stage at; diagnosis. American Journal of Public Health 96, 2173–2178.

Lillie-Blanton, M., Bowie, J., Ro, M., Falik, M., Collins, K., 1996. African American women: social factors and the use of preventive health services. In: Falik, M.M., Collins, K.S. (Eds.), Women's Health: The Commonwealth Fund Survey. Johns Hopkins University Press, Baltimore, pp. 99–122.

Murray, C.B., 1998. Racism and mental health. In: Friedman, H.S. (Ed.), Encyclopedia of Mental Health, vol. 3. Academic Press, San Diego, CA, pp. 345–357.

Murray, C.B., Woods, V.D., 2009. Psychology of health disparities among African; American populations: an overview. Journal of Black Psychology 35, 142–145. NCHPEG (National Coalition for Health Professional Education in Genetics). Race and Genetics FAQ. http://www.nchpeg.org/index.php?option=com_content&view=article&id=142&Itemid=64.

National Healthcare Disparities Report, 2003. Agency for Healthcare Research and Quality. Rockville, MD. http://wwwahrqgov/qual/nhdr03/nhdr03htm.

NCHPEG, 2014. FAQ, National Coalition for Health Professional Education in Genetics. http://www.nchpeg.org/index.php?option=com_content&view=article&id=142&Itemid=64 (accessed 06.03.2014.).

NCHS (National Center for Health Statistics), 2007. Health, United States, 2007 with Chartbook on Trends in the Health of Americans. Author, Hyattsville, MD.

NCHS (National Center for Health Statistics), 2010. Health, United States 2010. With Special Feature on Death and Dying Hyattsville, MD.

NCHS (National Center for Health Statistics), 2012. Health, United States 2011. With Special Feature on Socioeconomic Status and Health. Hyattsville, MD.

OMH (Office of Minority Health), 2013. Asian American/Pacific Islander Profile. U.S. Department of Health and Human Services. http://minorityhealth.hhs.gov/templates/browse.aspx?lvl=2&lvlID=53.

Perry, I., 2011. More Beautiful and More Terrible: The Embrace and Transcendence of Racial Inequality in the United States. New York, University Press, New York, 23.

Polednak, A.P., 1989. Racial and Ethnic Differences in Disease. Oxford University Press, New York.

Prostate Cancer Health Center, 2013. Prostate Cancer Risk Factors. WebMD. http://www.webmd.com/prostate-cancer/guide/prostate-cancer-risk-factors.

Shelton, R.C., Goldman, R.E., Emmons, K.M., Sorensen, G., Allen, J.D., 2011. An investigation into the social context of low-income, urban Black and Latina women: implications for adherence to recommended health behaviors. Health Education & Behavior 38, 471–481.

Smedley, B., Stith, A., Nelson, A. (Eds.), 2003. Unequal Treatment: Confronting Racial; and Ethnic Disparities in Health Care. Institute of Medicine, Washington, DC. National; Academies Press.

Suther, S., Kiros, G.E., 2009. Barriers to the use of genetic testing: a study of racial and ethnic disparities. Genetics in Medicine 11, 655–662.

Terrell, F., Terrell, S., 1981. An inventory to measure cultural mistrust among Blacks. Western Journal of Black Studies 5, 180–185.

United States Census, 2012. Money Income of Families – Median Income by Race and Hispanic Origin, Detailed Years of School Completed by Persons 25 and Older by Sex, Race, and Hispanic Origin. American Fact Finder.

Vamos, E.P., Novak, M., Mucsi, I., 2009. Non-medical factors influencing access to renal transplantation. International Urology and Nephrology 41, 607–616.

Ward, E., Jemal, A., Cokkinides, V., Singh, G.K., Cardinez, C., Ghafoor, A., Thun, M., 2004. Cancer disparities by race/ethnicity and socioeconomic status. CA: A Cancer Journal for Clinicians 54, 78–93.

Williams, D.R., 1999. Race, SES, and health: the added effects of racism and discrimination. Annals of the New York Academy of Sciences 896, 173–188.

Williams, D.R., Collins, C., 1995. US socioeconomic and racial differences in health. Annual Review of Sociology 21, 349–386.

Williams, D.R., Williams-Morris, R., 2000. Racism and mental health: the African American experience. Ethnicity and Health 5, 243–268.

Woods, V.D., Montgomery, S., Belliard, J.C., Ramirez-Johnson, J., Wilson, C., 2004. Culture, Black men, and prostate cancer: what is reality? Cancer control. Journal of the Moffitt Cancer Center 11, 388–396.

Woods, V.D., Montgomery, S.B., Herring, P., Gardner, R.W., Stokols, D., 2006. Social ecological predictors of prostate-specific antigen blood test and digital rectal examination in Black men. Journal of the National Medical Association 98, 492–504.

Race: Ethnicity and Urban Planning

June M Thomas, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses how urban planning has affected different racial and ethnic groups, and in turn how race and ethnicity affect the work of urban planning. The article summarizes some of the concerns with previous planning efforts in minority communities, but it argues that social justice is important for both individuals and institutions and remains an important consideration for urban planning as a field. Short discussions of the subfields of land use regulation and of housing, only two of many possible urban planning specializations, offer examples of the interaction of race, ethnicity, and urban planning.

This article discusses those aspects of urban planning that affect or are affected by varying conditions facing different racial or ethnic groups. 'Urban planning' – also known as urban and regional planning, or as town and country planning, depending on the country of practice – concerns the location, arrangement, and quality of housing, transportation, environmental protection, community services, land use, or commercial facilities. These activities can have many positive effects on residents of a city or region, but in certain contexts they may also exacerbate social inequality.

The word 'race' typically refers to one of several broad categories of humanity, such as Caucasian or 'white,' Asian, or dark African heritage or 'black,' but such categories are not well-defined biologically. Furthermore, different countries define race differently; some countries treat immigrants from the same nation or region as one race, but these groups may have marked internal differences. In some parts of the world, ethnic or religious differences are much more significant than racial differences (Yiftachel, 2006). Therefore, both race and ethnicity are referred to while discussing planning. The term 'ethnic' is used to mean groupings of people bound by historic ties such as language, nationality, or customs, a definition that will vary depending on context. The focus here is on how inequality by race or ethnicity relates to urban planning in a particular place of practice, such as a metropolitan area or a nation.

Background of the Issue

The linkage between urban planning and race arose in modern multiracial societies because decision makers sometimes used urban planning tools as strategies for racial oppression or exclusion of relatively powerless racial or ethnic groups. In the United States, during the early part of the twentieth century, some of the first tools that professional urban planners used to regulate land use enforced the residential segregation of African-American households from white households (Silver, 1997). In societies characterized by white colonialists living in subject territories, European powers used urban planning tools to make sure that the spatial arrangement of urban settlements mirrored official roles. Grand approaches to colonial government buildings sent a visual message that the European colonial governments and their citizens were dominant. Residential areas for Europeans had much higher standards of services and housing than sections set aside for 'natives.' Planning was therefore used as an expression of power and subjugation in the guise of order and sanitation (Boyer, 1983; Hall, 1996; Njoh, 2009).

As multiracial urban societies evolved during the twentieth century, many governments abandoned the purposeful isolation of supposedly subordinate racial groups but developed other mechanisms of oppression. Some US localities used urban redevelopment programs to isolate racial or ethnic minorities, either through building racially segregated public housing or clearing minority-race or minority-ethnic neighborhoods with the help of federal 'urban renewal' funds. The rationale was often the low quality of targeted-area housing stock, due to residential secession and poverty, and many planners in fact saw themselves as serving both the public interest and the poor by carrying out such policies (Thomas, 1997). Yet local planners did not always recognize the right of targeted-area residents to live in low-income but viable neighborhoods or to participate in decision making about clearance of their own communities, and federal policy did not offer adequate financial support for relocation to other areas. Furthermore, such redevelopment took place during an era of racial segregation in housing, leading to few options for ousted residents who were black. Federal 'urban renewal' thus led to spirited protests in the 1950s and 1960s (Wilson, 1966). The outcry over this program, nicknamed 'Negro removal' because so many targeted neighborhoods were black, led to significant reforms in US federal policies and local planning activities, including the rise of a movement of planners devoted to the cause of social justice (Thomas, 2012). Yet racial segregation was so entrenched in other ways that it proved difficult to lessen simply by changing federal and local policies.

In other parts of the world, as colonialism withered, indigenous residents of former colonies arose to plan their own cities but nonetheless faced significant social divisions because of caste, class, or ethnicity. It was not atypical for native-race government workers to move into the same districts once inhabited by white colonial settlers, and to enjoy better benefits in municipal services and housing than poor or working-class residents. In some locales, certain ethnic groups gained higher status over others, again mirrored in living conditions. Although planners often worked to improve living conditions for the less privileged, they also sometimes protected urban structures reflecting the fissures of socioeconomic status.

Although the purposeful use of urban planning tools to enforce racial subordination, isolation, and segregation waned by the early twenty-first century, the problem has not gone

away. Subtle exclusion still characterizes seemingly innocuous urban planning decisions. Yet the nature of the issues has changed markedly as broader societal questions have changed. In urban South Africa, for example, the end of apartheid has ended legally enforced spatial segregation by race, but not uneven settlement or development (Beauregard, 1998; Watson, 1998). In modern Europe, immigrants of color have sometimes taken on the role of 'other' and faced prejudices and exclusions. In the United States, decision makers use planning tools to help maintain autonomy and exclusion by income status if not by race or ethnicity. Thus, the conversations continue even though the particulars may differ.

Areas of Scholarly Inquiry

Bauman has suggested four main categories of scholarships related to race in planning history (Bauman, 2002). Bauman's first category includes scholars who see minority groups as victims of planning decisions. A second category views race, ethnicity, or class as issues necessary to understand the inherent 'structure or contours' of urban society. The third category would treat minority groups as purposeful agents involved in shaping the urban environment. A final category would view such groups as simply part of a larger picture composed of various actors. The author focuses here on Bauman's first, second, and third categories, victimization, impact on urban structure, and activism.

Another way to look at scholarship in this area, other than Bauman's continuum for planning history, would be to consider contemporary planning-related specialty areas such as planning theory, land use, housing, transportation, environment, urban design, economic development, or others, examining each topic for racial or ethnic implications. To do this, however, would be a daunting and lengthy task for this summative narrative. Instead, the next section will briefly discuss social justice in planning theory as well as additional implications of Bauman's continuum; these ideas would apply to any planning specialization. Then a following section will briefly summarize specific concerns in two of the many possible specialty areas in planning: land use and housing.

Social Justice, Urban Structure, and Agency

Historically, urban planning tools and programs have sometimes oppressed racial or ethnic minorities, as in Bauman's first category, victimization. Much of this literature is based in the United States (Thomas and Ritzdorf, 1997; Sandercock, 1998; Gillette, 2006). Such scholarship once focused on planning inequity between blacks and whites, but recently, scholars have offered more careful study of groups such as Hispanics and indigenous peoples (Irazabal and Farhat, 2008). Contemporary planning literature also examines circumstances around the globe, particularly in Europe and the Middle East (Thomas, 2000; Bollens, 2006; Maginn, 2004). Such scholarship has often shown that planning actions do not always protect social justice for subject, oppressed, or disadvantaged populations.

Justice and social justice are complicated topics, but some authors have helped to link these ideas to urban planning. Notable among these are David Harvey, whose book *Social Justice in the City* was foundational (Harvey, 1973), and Iris Marion Young, who helped define justice in the context of distinctly identifiable groups. Young warned against judging people solely by their group membership, as individuals can differ, act on free will, or belong to several overlapping groups; yet, at the same time, she noted, certain definable social groups deserve the ability to give voice to their protests against oppressive treatment. Suggesting that clustering was not always bad, she stressed that certain groups should be able to cultivate their own identity without exclusion or forced immobility (Young, 2000; Fainstein, 2007). Furthermore, she noted that social and political systems need to correct social wrongs even though individuals also have justice-related obligations. Other urban scholars who wrote about social justice often referenced her work (Harvey, 2002; Fainstein, 2000, 2010; Soja, 2010). They echoed the need for planning-related decisions to avoid oppression of various kinds, involve all affected parties in decision making, and offer equitable access to resources.

The responsibility for social justice lies largely with social and political systems, Young argued, but individuals still have responsibilities to help right social wrongs. Urban planners may not always be in a position to counteract such social wrongs easily; they may work as civil servants or consultants, and they may answer to elected or appointed officials as their employers; these employers could very well push for objectionable decisions. Yet planners have advisory power and implement some programs, and they need to have a sense of planning professionalism that eschews social oppression and promotes social justice.

Thus, codes of professional ethics enter this discussion. One way to deal with the historical and contemporary dilemmas caused by a racially fragmented society is to create a professional code of ethics that specifically exhorts support for social justice. Both general principles of ethical behavior promoted by the American Planning Association and the professional code of ethics enjoined by the American Institute of Certified Planners (both in the United States) encourage planners to support opportunities for those who have few, to give voice in planning decisions to those who may be marginalized, and to recruit diverse peoples into the planning profession. Certified planners in the United States, furthermore, should explicitly promote racial integration, according to their code. In South Africa, the code is weaker on this dimension: planners need only promote equality and avoid personal discrimination, while the UK code prohibits discrimination of several types (see Web pages at end of article). Although these official ethical provisions can be quite limited, they do offer at least some support for nondiscrimination.

Bauman suggested a second category of study: the effects of race on urban form and structure. Urban studies in the United States analyze residential segregation as a way of studying race and ethnicity in the context in which planners must work. The topic of segregation has also become increasingly important in many other countries experiencing the clustering of various ethnic groups and nationalities, since such urban patterns affect access to jobs, housing, high-quality schools, and other public services. Journals such as *Urban Studies* have published studies from around the world about similar issues. More connected with the disciplinary areas of the social sciences than with urban planning, these works are nevertheless important to

urban planners. Documentation of segregation by race, ethnicity, and socioeconomic status helps explain the nature of urban form as well as identify possible circumstances needing action. Other works focus not so much on residential segregation as on changes in urban structure, urban policies, or development plans that race or ethnic divisions affect, such as US literature on post-Hurricane Katrina New Orleans (Lowe and Shaw, 2009), on economic opportunity fettered by racial exclusion (Harper-Anderson, 2008), or on metropolitan divisions by race that affect the ability of a region to prosper (Jacobs, 2009).

Activism by minority-races or ethnicities themselves, the third of Bauman's suggested categories, is also an important area of study. One way to frame this activism is simply as community development, which means the organized efforts of groups to develop and support their own communities. Soja (2010) describes such activism in Los Angeles, but a number of other books and articles do as well, and the concept of community development itself is by no means limited to groups defined by race or ethnicity. One line of studies has shown that activism in matters related to land use is a particularly fertile area for indigenous peoples or ethnic groups in multiracial societies (Jojola, 1998; Lane and Hibbard, 2005), including a growing-related literature concerning North American Latinos (Morales, 2010; Carpio et al., 2011). The environmental justice movement also highly values activism carried out by oppressed minorities confronted with disproportionate exposure to poor or toxic environmental conditions (Bullard, 1993; Agyeman, 2005). At this point, however, much more literature exists on the victimization aspect of the relationship between race/ethnicity and planning than on racial or ethnic minorities as activists.

Land Use and Housing

Land use and housing are generally two different planning specialties but they are closely related when discussing race and ethnicity. Zoning refers to one specific aspect of municipalities' legal regulation of land use or buildings and is a commonly used planning tool. Land use decision making may routinely exclude people as possible residents on the basis of socioeconomic status, ethnicity, or race. An example would be zoning ordinances that exclude low-income or multifamily housing, or housing on relatively small lots. The general term for such discriminatory action is 'exclusionary zoning.' In the early twentieth century, some locales set up zoning categories based on racial exclusion, but these are now in most parts of the world illegal; South Africa, under apartheid, was one of the last holdouts. Yet in metropolitan areas in which low-income racial minorities are isolated and segregated in some municipalities, zoning that permits only (or mostly) low-density single-family housing may have much the same effect as would zoning exclusions based on racial characteristics (Massey and Rothwell, 2009; Dyble, 2010; Rothwell, 2011).

In the United States, although overt racial discrimination in housing rental or purchase violates civil rights laws that date back to the 1960s, the federal judicial system will not substantially challenge socioeconomic exclusion through zoning rules, which localities use to regulate or limit various types of residential housing. Yet local residents in privileged or middle-class municipalities commonly oppose subsidized, low-income, or multifamily housing, a situation that may have enormous racial or ethnic implications. The failure to allow low-income plaintiffs who seek affordable housing in suburban communities to gain 'standing' has effectively shut off an important route to access to suburban housing for many inner-city minorities (Federal Courts – Standing, 2007). Although 2010 census figures showed that minority-race suburbanization is increasing, this integration has spread unevenly across metropolitan space; blacks, for example, are moving to some suburbs but not others, and they remain far more isolated than any other minority groups.

Planners in diverse societies are reshaping land use regulations to recognize special needs of ethnic minorities. For example, in Canada many regulations for buildings and land use concerning religious facilities assume that any religious building would be a free-standing church. An immigrant-rich context in which Muslims of different Middle Eastern nationalities are increasing in numbers has required creative flexibility in many ways. A common sequence for such immigrants is to begin to meet congregational needs by holding smaller then larger meetings in residential houses, a situation that might counter traditional zoning regulations. When growth in size or finances allow, these groups may then move to rented stores or other buildings, and then construct mosques with unique architectural or sound-system features not allowed by regulations designed only for Christian churches (Qadeer, 1997). Similar issues have arisen in Great Britain and other countries (Gale, 2008).

Aside from the issue of exclusion of certain kinds of residential land use, the problem of making sure that members of various minority groups have adequate housing faces planners in many different countries. In developing countries, the problem may very well be a lack of societal resources to make sure that the population is well housed, although even there it is not difficult to imagine different access to housing resources based on differing ethnic status or race. In more developed countries, the problem of unequal access to housing is connected with the issue of urban structure; certain areas of cities or metropolitan areas are very likely to have better housing resources than other areas, and when conditions of inequality exist, housing resources likewise differ.

In such situations, the social responsibility of the urban planner who focuses on housing would be threefold. First, as mentioned above, where certified planners have a clearly stated professional obligation to help promote integration, as in the United States, seeking ways to promote this should be the planner's goal. A few experiments in residential integration for racial minorities have led to suburban resistance, such as with the United States' Moving to Opportunity program (Briggs et al., 2010), but other opportunities exist to make less drastic changes in racial patterns. Second, in those places with antidiscrimination laws, the planner would be obligated to follow those laws and to discourage efforts to block racial or ethnic minorities from available housing. But the third strategy is perhaps the most common for housing planners. That is to work within existing programs to find the financial and political resources needed either to protect or upgrade existing housing or to build new residential units. Training for planners interested in promoting housing would include full exposure to

various private and public financing programs for housing rehabilitation or construction. The planner would then need to make sure that such resources were made available with no exclusion based on race and ethnicity and perhaps with special targeting to minority populations.

In some places, governments have been involved in building low-income housing for some time; depending on the country and program, the label for such housing could be public housing, estate housing, social housing, or subsidized housing. Although strategies to purposefully isolate racial or ethnic populations in such low-income housing are increasingly rare in the 2000s, the practical result of building such housing in self-contained complexes may indeed be social isolation. One possible task for planners would be to strive for income or racial diversity in such housing, or to work with ethnic and racial groups to assure that the government-subsidized housing meets their needs. In only a few places, such as Singapore, has the government actually set up quotas for various ethnic groups in different public housing developments (Sin, 2002), a common practice in the United States in the 1940s. In Singapore, the goal apparently is to promote balance rather than exclusion.

Responses to Inequality and Injustice

Urban planners have reacted in several ways to the continued problems related to the connections between racial injustice and urban planning in diverse societies. Scott Bollens has offered a typology of responses that is useful as a framework for discussion (Bollens, 1988).

One category that has already been discussed includes those planners who try to reduce inequalities between groups. Many urban planners in the United States, for example, consider the advocacy planning and equity planning movements to have laid the groundwork for important responses to the racial crisis in their country, and progressive planning has continued as a label for those concerned about such ends. The concept of equity planning is a practical response to the crises of local policy caused by racial segregation and poverty in American cities. This particular approach suggests that urban planners can and do rise above the limitations of government bureaucracy and pursue other local policies to channel direct benefits to people hindered by race, ethnicity, or socioeconomic status (Krumholz and Forester, 1990). Several key planners and scholars have explicitly addressed the important role of race in helping to create the unjust conditions to which planners must respond, and their concepts have influenced the US professional code of ethics. Specific groups of planners have worked to directly address the needs of various racial and ethnic minorities and the poor through their work and to raise attention to their cause in several venues, including the planners' professional and academic organizations. (See Web page links below.)

A second possible response by an urban planner is to side with the status quo in the name of objectivity, even in the face of blatant inequities. The urban planner or policy is supposedly 'neutral,' but in fact may reinforce the position of the dominant group. Such a so-called neutral planner may refuse to acknowledge or address issues of segregation or polarization. The shield, under such circumstances, is that urban planning – or at least their version of it – is a professional exercise not heavily involved with issues of social inequality. This particular attitude may actually be hurtful when exemplified by planners in circumstances of major inequality or racial or ethnic conflict. The claim that the planner is 'color blind' may be only a poor excuse for collaboration with systematic injustice.

A third response might be called 'partisan,' where the planner rejects the concerns of powerless groups and sides with the powerful. The city's government and civic leaders view neighborhoods and policies through a lens colored by race or ethnicity, and the interests of the dominant group sets the norm for professional urban planners. As the political leaders set the policies, the urban planners carry them out, even in the face of blatant favoritism or discrimination. Bollens offered several examples from South Africa, but also noted the same syndrome in Jerusalem, where the systematic and purposeful isolation of Arab neighborhoods had gained the force of righteousness. Similar trends existed in Belfast, Ireland, where religion was the primary divider. As Bollens noted, somehow local urban planners were able to function without any noticeable discomfort about the obvious partisanship of their actions.

Finally, Bollens referred to 'a resolver strategy,' which attempted to address the root causes of power imbalances, competition among racial and ethnic groups, and disempowerment. This approach then aimed to confront the status quo call and profoundly transform the system. Although this approach is in some ways simply a more impassioned form of equity planning, it also has broader implications in countries such as South Africa. There, for example, resolvers have begun to carry out development planning that differs markedly from traditional urban planning activities, which favored white areas and white concepts. Development planning, in contrast, more fully integrates spatial planning with social and economic planning. It restructures policy directions by cutting across traditional departments and governments, and it uses a participatory process to help empower marginalized people. It would involve more members of the victim class as professional planners than was previously the case.

It is this last point, the need to involve the subordinate social group in professional planning, that has led to much discussion about the challenge of diversifying the urban planning profession as well as the planning academy. Although strides have been made in doing so, diversity is still lacking both in the profession and in planning education, and a number of challenges remain in terms of rectifying this shortcoming (Wubneh, 2011; Sweet and Etienne, 2011).

That the urban planning profession maintains an ongoing dialog about issues related to resolving social injustice is a healthy sign. Although many societies remain severely fragmented and alienated by race, the situation is open to proactive effort to improve existing conditions. As larger societies become more comfortable with multiculturalism and multiracialism, urban planning as a field can and will adjust.

See also: Advocacy and Equity Planning; Citizen Participation; Just City; Neighborhood Revitalization and Community

Development; Planning Ethics; Race and Racism, Geography of; Spatial Equity; Urban Geography; Urban Planning (Western), History of; Urban Poverty in Neighborhoods.

Bibliography

Agyeman, Julian, 2005. Sustainable Communities and the Challenge of Environmental Justice. New York University Press, New York.

Bauman, John, 2002. Race, class, gender and sexuality in planning history: a look at trends in the literature. Journal of Planning History 1, 225–229.

Beauregard, Robert, 1998. Subversive histories: texts from South Africa. In: Sandercock, Leonie (Ed.), Making the Invisible Visible: A Multicultural Planning History. University of California Press, Berkeley, pp. 184–197.

Bollens, Scott, 1988. Urban planning amidst ethnic conflict: Jerusalem and Johannesburg. Urban Studies 35, 729–750.

Bollens, Scott, 2006. Urban planning and peace building. Progress in Planning 66, 67–139.

Boyer, Christine, 1983. Dreaming the Rational City: The Myth of American City Planning. MIT Press, Cambridge.

Briggs, Xavier de Souza, Popkin, Susan, Goering, John, 2010. Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty. Oxford University Press, New York.

Bullard, Robert (Ed.), 1993. Confronting Environmental Racism: Voice from the Grassroots. South End Press, Boston.

Carpio, Genevieve, Irazabal, Clara, Pulido, Laura, 2011. Right to the suburb? Rethinking Lefebvre and immigrant activism. Journal of Urban Affairs 33, 185–208.

Dyble, Louise, 2010. Review essay: the continuing saga of zoning in America. Journal of Planning History 9, 140–146.

Fainstein, Susan, 2000. New directions in planning theory. Urban Affairs Review 35, 451–478.

Fainstein, Susan, 2010. The Just City. Cornell University Press, Ithaca.

Fainstein, Susan, 2007. Obituary – Iris Marion Young (1949–2006): a tribute. Antipode 39, 382–387.

Federal courts – Standing, 2007. Third Circuit denies standing to bring claim of racial discrimination in zoning. Harvard Law Review 120, 1105–1111.

Gale, Richard, 2008. Locating religion in urban planning: beyond race and ethnicity? Planning, Practice & Research 23, 19–39.

Gillette, Howard, 2006. Between Justice and Beauty: Race, Planning, and the Failure of Urban Planning in Washington, D.C. University of Pennsylvania Press, Philadelphia.

Hall, Peter, 1996. Cities of Tomorrow: An Intellectual History of Urban Planning and Design in Twentieth Century, second ed. Blackwell, Oxford.

Harper-Anderson, Elsie, 2008. Benchmarks and barriers: African American experiences in the Corporate Bay Area's new economy sector of the 1990s. Journal of Planning Education and Research 27, 483–498.

Harvey, David, 1973. Social Justice in the City. Johns Hopkins University Press, Baltimore.

Harvey, David, 2002. Social justice, postmodernism, and the city. In: Fainstein, Susan, Campbell, Scott (Eds.), Readings in Urban Theory, second ed. Blackwell, Malden, MA, pp. 386–402.

Irazabal, Clara, Farhat, Ramzi, 2008. Latino communities in the United States: place-making in the pre-World War II, post-World War, and contemporary city. Journal of Planning Literature 22, 207–228.

Jacobs, Andrew J., 2009. Embedded contrasts in race, municipal fragmentation, and planning. Journal of Urban Affairs 31, 147–172.

Jojola, Ted, 1998. Indigenous planning: clans, intertribal confederations, and the history of the all Indian Pueblo Council. In: Sandercock, Leonie (Ed.), Making the Invisible Visible: A Multicultural Planning History. University of California Press, Berkeley, pp. 100–119.

Krumholz, Norman, Forester, John, 1990. Making Equity Planning Work: Leadership in the Public Sector. Temple University Press, Philadelphia.

Lane, Marcus, Hibbard, Michael, 2005. Doing it for themselves: transformative planning by indigenous peoples. Journal of Planning Education and Research 25, 172–184.

Lowe, Jeffrey, Shaw, Todd, 2009. After Katrina: racial regimes and human development barriers in the Gulf Coast region. American Quarterly 61, 803–827.

Maginn, Paul, 2004. Urban Regeneration, Community Power, and the (In)Significance of Race. Ashgate Publishing, Aldershot.

Massey, Douglas, Rothwell, Jonathan, 2009. The effect of density zoning on racial segregation in US urban areas. Urban Affairs Review 44, 779–806.

Morales, Alfonso, 2010. Planning and the self-organization of marketplaces. Journal of Planning Education and Research. 30, 182–197.

Njoh, Ambe, 2009. Urban planning as a tool of power and social control in colonial Africa. Planning Perspectives 24, 301–317.

Qadeer, M.A., 1997. Pluralistic planning for multicultural cities. Journal of the American Planning Association 63, 481–494.

Rothwell, Jonathan, 2011. Racial enclaves and density zoning: the institutionalized segregation of racial minorities in the United States. American Law and Economics Review 13, 290–358.

Sandercock, Leonie (Ed.), 1998. Making the Invisible Visible: A Multicultural Planning History. University of California Press, Berkeley.

Silver, Chris, 1997. The racial origins of zoning in American cities. In: Thomas, June M., Ritzdorf, Marsha (Eds.), Urban Planning and the African American Community: In the Shadows. Sage, Thousand Oaks, CA, pp. 23–42.

Sin, Chih, 2002. The quest for a balanced ethnic mix: Singapore's ethnic quota policy examined. Urban Studies 39, 1347–1374.

Soja, Edward, 2010. Seeking Spatial Justice. University of Minnesota Press, Minneapolis.

Sweet, Elizabeth, Etienne, Harley, 2011. Commentary: diversity in urban planning education and practice. Journal of Planning Education and Research 31, 332–339.

Thomas, Huw, 2000. Race and Planning: The UK Experience. Routledge, London.

Thomas, June Manning, 1997. Redevelopment and Race: Planning a Finer City in Postwar Detroit. Johns Hopkins University Press, Baltimore.

Thomas, June Manning, 2012. Social justice as responsible practice: race, ethnicity, and the influence of the civil rights era. In: Sanyal, B., Vale, L., Rosan, C. (Eds.), Planning Ideas that Matter. MIT Press, Cambridge, pp. 359–385.

Thomas, June Manning, Ritzdorf, Marsha (Eds.), 1997. Urban Planning and the African American Community: In the Shadows. Sage, Thousand Oaks, CA.

Watson, Vanessa, 1998. Planning under political transition—lessons from Cape Town's metropolitan planning forum. International Planning Studies 3, 335–350.

Wilson, James Q., 1966. Urban Renewal: The Record and the Controversy. MIT Press, Cambridge.

Wubneh, Mulatu, 2011. Commentary: diversity and minority faculty perception of institutional climate of planning schools – results from the climate survey. Journal of Planning Education and Research 31, 340–350.

Yiftachel, Oren, 2006. Essay: re-engaging planning theory? Towards 'South-Eastern' perspectives. Planning Theory 5, 211–222.

Young, Iris M., 2000. Inclusion and Democracy. Oxford University Press, New York.

Relevant Websites

http://www.planning.org/ethics/ethicscode.htm – American Institute of Certified Planners. "Code of Ethics."

http://www.planning.org/diversity/ – American Planning Association. "Diversity Resources."

http://www.plannersnetwork.org/ – Planner's Network.

http://www.acsp.org/aboutpocig – Planners of Color Interest Group. Association of Collegiate Schools of Planning.

http://www.sapi.org.za/about-sapi/our-code-of-conduct.html – South African Planning Institute. "Code of Conduct."

http://www.rtpi.org.uk/media/831098/code_of_professional_conduct_2012.pdf – The Royal Town Planning Institute. "Code of Professional Conduct."

Race: Genetic Aspects

Guido Barbujani, University of Ferrara, Ferrara, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Genetic research has shown that all individuals and populations are different, but no agreement has ever been reached on the number and definition of human races, with proposed races numbering from none to 200. Human genetic differences are patterned in geographical space, but each population harbors a large proportion of the species' diversity and shares with other populations most of its genetic variants. Race remains an important component of our social and psychological world, but envisaging humans as subdivided in genetically differentiated races leads to poor evolutionary inference and to errors in clinical practice.

Introducing geographic variation in his classical book *Systematics and the Origin of Species*, Ernst Mayr (1947) makes a distinction between species in which biological changes from population to population are continuous, and species in which populations with different character combinations are separated by borders. In the latter, the entities separated by borders are called subspecies or geographic races. Similarly, in human genetics (see, e.g., Vogel and Motulsky, 1986) and anthropological (Brues, 1977; Molnar, 1998) textbooks, races are often envisaged as large populations of individuals who have a significant fraction of their genes in common and can be distinguished from other races by their common gene pool. Under both definitions, increased levels of genetic similarity among members of a race is the result of shared ancestry.

The idea that races are a natural feature of human diversity has long set the standards for anthropological research. For centuries, people sharing the same 'blood,' especially in noble families, were said to form a race (Furetiere, 1690, cited in Cohen, 1991). In the eighteenth and nineteenth centuries, the word acquired a more general meaning, coming to designate constant human types, and it was used both to describe and to explain human diversity. Scientists investigating human variation classified people in different races and explained their differences as a consequence of their belonging to different races. That view is based on an empirical observation and on an assumption. The observation is that members of the same community look similar, whereas people from abroad look different; the assumption is that these similarities and differences reflect our species' natural subdivision in clusters of individuals sharing common ancestors. Testing that assumption at the genetic level became possible only in the second half of the twentieth century.

In the meantime, however, the race concept has become intrinsically ambiguous. For creationists, including Linnaeus, species were fixed entities, and so there was no conceptual difficulty in assigning individuals not only to species, but also to subdivisions of lower taxonomical rank. However, Lamarck's and Darwin's proposal that different species arise with modification from common ancestors implies that borders between them are fluid and evolve with time. As a consequence, species are now regarded as groups of organisms capable of interbreeding and producing fertile offspring, reproductively isolated from other such groups (de Queiroz, 2005). Accordingly, a race or subspecies is a genetically differentiated group of organisms on its way to developing reproductive isolation from other groups of the same species, but not there yet; this way, its borders are, by definition, uncertain.

In practice, some subspecies may be rather obvious. In several plant and animal species (e.g., the Hawaiian amber snail; Holland and Cowie, 2007), the greatest fraction of the overall genetic diversity occurs between populations or groups of populations. When that is the case, most individuals of unknown origin can be assigned with a good degree of statistical confidence to one group on the basis of their morphological or genetic characteristics, and these groups are likely to have a common ancestry. However, in other organisms genetic differences between populations are small and each population harbors extensive variation (one example is the Bluefin tuna, Riccioni et al., 2010).

Tuna are highly mobile animals, snails are not, and in general sharp genetic differences between populations are observed in species in which isolation prevails over migration. One question, then, is whether the patterns of human genetic diversity have been shaped by factors isolating or connecting populations, with isolating factors promoting the establishment of a racial structure. Two related questions are whether genetic differences between humans are large enough for a racial classification to be (1) possible and (2) useful for purposes such as understanding human differences, reconstructing past demography, diagnosing disease, or providing effective medical treatment.

Listing Human Races

Science has left creationism behind, but the idea that a scientific study of humans starts from their racial classification was long unchallenged (Cohen, 1991) and still remains attractive to some. People differ in many aspects; we all recognize variations in facial traits, height, body structure, skin, and hair color; although social status affects the way people are classified in races (Weeks and Lupfer, 2004), often this variation allows one to guess with some accuracy a person's geographical origin. The classical, or typological, approach consists then in identifying some basic human types, defined on the basis of such traits, and then attributing each individual to one of those types, or

races. Ideally, most individuals and populations should fall unambiguously into the appropriate group, except perhaps for a minority of them, which would be regarded as derived from interracial crosses. However, even in the early years it was evident that single morphological traits do not allow a stable classification. As a consequence, races were generally defined by combinations of traits, occasionally including cultural variables such as language, house-building and tool-making techniques, and mating systems (Cohen, 1991).

An incomplete selection of the many racial catalogs proposed using a typological approach is in Table 1; for a broader treatment of this issue, see Cohen (1991) and Molnar (1998). The first scientific attempt can be traced back to Linnaeus' nineth edition of *Systema naturae*, in which six races were proposed. Two of them consisted of individuals with congenital pathologies (*Homo sapiens monstruosus*), and hybrids between humans and other species whose existence has not been confirmed by subsequent studies (*H. sapiens ferus*). Each of the other four races (*europaeus*, *luridus*, *afer*, *americanus*) corresponded to one element of classical Greek philosophy (air, water, earth, fire) and was defined by both physical and psychological traits. Some of the most prominent naturalists of the nineteenth century, Buffon, Blumenbach, and Cuvier among them, modified Linnaeus's taxonomy, splitting or collapsing some of the previously proposed races.

Table 1 Some lists of proposed human races

Author	*No. of races*	*Races proposed*
Linnaeus (1735)	6	Europaeus, Luridus, Afer, Americanus, Ferus, Monstruosus
Buffon (1749)	6	Laplander, Tartar, South Asian, European, Ethiopian, American
Blumenbach (1795)	5	Caucasian, Mongolian, Ethiopian, American, Malay
Cuvier (1828)	3	Caucasoid, Negroid, Mongoloid
Deniker (1900)	29	
Weinert (1935)	17	
Von Eickstedt (1937)	38	
Coon et al. (1950)	30	
Biasutti (1959)	53	
Coon (1962)	5	Congoid, Capoid, Caucasoid, Mongoloid, Australoid
Garn (1961)	9	African, European, Asian, Indian, Amerind, Melanesian, Polynesian, Micronesian, Australian
U.S. Census Bureau (2000)	6	White, Black or African–American, American Indian and Alaska Native, Asian, Native Hawaiian and other Pacific Islander, Hispanic or Latino
U.S. Census Bureau (2010)	15	White, Black or African–American, American Indian and Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Native Hawaiian, Guamanian, Samoan, Other Pacific Islander, Hispanic or Latino
Risch et al. (2002) Figure 1	5	African, Caucasian, Pacific islanders, East Asian, Native American
Risch et al. (2002) Table 3	5	African–Americans, Caucasians, Hispanic Americans, East Asians, Native Americans

Compiled from Madrigal, L., Barbujani, G. 2006. Partitioning of genetic variation in human populations and the concept of race. In: Crawford, M H., (Ed.) Anthropological Genetics. Theory, Methods and Applications. Cambridge University Press, Cambridge, pp. 19–37; Coon, C S., Garn, S M., Birdsell, J B., 1950. Races. C. Thomas, Springfield, Illinois; and Garn, S M., 1961. Human Races. C. Thomas, Springfield, Illinois, where the original bibliographical references can be found.

In Table 1, a problem and a trend are evident. The problem is that it is fairly simple to list some typical features of a population, but each human group, however defined, includes people who do not much resemble the typical individual. This leads to a number of ambiguities, reflecting the classifier's standpoint; for example, the number of possible skin colors recognized by different scientists has fluctuated between 4 and 34 (Cohen, 1991). The trend is partly a consequence of the problem. When the European and North American experts came in contact with new populations, it generally proved difficult to fit them into the existing lists of races, which were repeatedly expanded, reaching figures around 200 in the twentieth century (Molnar, 1998). However, increasing the number of races created additional problems because their borders became more and more elusive. As a result, contradictions emerged between catalogs proposed by different authors, and even by the same author (see Risch et al., 2002 for an example). Things were by no means clearer when it came to practical applications of the race concept. Since 1890, the U.S. Census Bureau has been asking residents to choose the race with which they most closely identify; the number of possible answers changed almost every decade, from 8 (1890) to 11 (1920), 9 (1930), 12 (1960), 6 (2000), and 15 (2010).

In time, this led to a critical reconsideration of the validity of the race concept for humans. In an article entitled "On the Nonexistence of Human Races," Frank Livingstone (1962) observed that different genetic loci give discordant information on the individuals' relationships, that human genetic variation is distributed in continuous gradients over the geographical space, and that boundaries between groups are anything but obvious. Other scientists disagreed; Theodosius Dobzhansky (1973) maintained that human biological diversity is properly described in terms of racial differences; he also admitted that no consensus had yet been reached on the number and definition of human races, but predicted that future genetic studies would settle the issue. The debate is still open. More recently, inconsistencies in racial classification were recognized, but not taken to invalidate the race concept, by authors such as Levin (2002) and Harpending (2007). By contrast, Brown and Armelagos (2001) and Relethford (2003) asserted that a classification system based on ill-defined categories has no scientific value. Dealing with the philosophical aspects of this question would be out of place here, but to test genetically the validity of the race concept, some degree of agreement is needed on the number and definition of human races.

Apportioning Human Diversity

The future that Dobzhansky was looking forward to has come; it is now possible to quantify levels of polymorphisms, that is, levels of genetic diversity, between individuals and

populations. To put these data in a more general context, one has to remember that the (haploid) human genome size is about 3.2 thousand million nucleotides; that some 36 million variable nucleotide sites are known (see http://browser.1000genomes.org/index.html); that the average nucleotide difference between humans and chimps is about 40/1000 (Varki and Altheide, 2005), and between random members of our species 1/1000. The percentages of the global species variance reported below are fractions of the last-mentioned figure.

Various statistical methods are used to apportion diversity within and between human groups, but the conceptual framework is rather constant. Groups are defined on the basis of independent evidence. Each individual is compared to all other individuals studied, and their genetic differences are counted. Three statistics are then estimated from those counts, namely, the average genetic variance: (1) between individuals of the same population; (2) between populations of the same group; and (3) between groups (Table 2). This approach of breaking down the overall genetic diversity in three hierarchical components was pioneered by Lewontin (1972), who in the pre-DNA era analyzed protein variation at 17 genetic loci. In that and successive studies, seven-race (Lewontin, 1972; Latter, 1980) or three-race classification systems (Ryman et al., 1983) were tested; in all cases, individual differences between members of the same population accounted for some 85% of the global human variance (Table 2), with values for individual loci ranging from 63.6 to 99.7%. Belonging to different populations added between 3 and 8% to that value, and belonging to different groups or races between 6 and 11%. At the protein level, therefore, variation among putative races seems to represent less than one-tenth of our species' overall genetic diversity.

At first, these proportions appeared counterintuitive, but later DNA analyses confirmed them with remarkable precision (Table 2). In practice, each human population seems to harbor a large share of the genetic diversity of the whole species. If we arbitrarily set to 100, the maximum genetic difference between two humans, the expected difference between members of the same community (excluding close relatives) is about 85, that is, much more than previously illustrated definitions of race appear to imply. This 85% figure was estimated from isolated populations of anthropological interests, rather than from urban populations, which would be a more typical sample of humanity. A UN document released in 2008 reports that, in fact, half of all humans live in urban areas. This means that the within-population component of variance is substantially underestimated in the available genetic studies and that, if subjects were sampled totally at random over the planet, differences between populations and groups thereof would account for an even smaller fraction of the species' variance.

The Genome Revolution

An impressive amount of data on human genomes, including at present nearly 1000 complete genome sequences, has accumulated in the last decade (see Clarke et al., 2012). Analysis of genomic data is showing why it has proved impossible so far to find an agreement on the main genetic groups of humankind. Xing et al. (2010) analyzed nearly 250 000 Single-Nucleotide Polymorphisms (SNPs), at none of which a fixed difference would distinguish any pair of continental populations. More than 80% of human SNPs are cosmopolitan: That is, the two alleles occur, at different frequencies, in all continents (Jakobsson et al., 2008). Combinations of alleles on the same chromosome, or haplotypes, have a clearer geographical distribution, but, with the only exception of Africa, continent-specific haplotypes are but a small proportion of the total haplotypes of the species (Jakobsson et al., 2008). A similar pattern was found in the analysis of long DNA stretches, or haplotype blocks, in which two main classes of haplotypes were described, one exclusively African, including 26% of the blocks, and the other worldwide distributed, including more than half of the blocks (Gabriel et al., 2002). We can summarize these findings by saying that, with limited exceptions,

Table 2 Estimated fractions of the global human diversity at three hierarchical levels of population subdivision

Polymorphism, number of loci	*Reference*	*Within population*	*Between populations, within race or continent*	*Between races or continents*
Protein, 17	Lewontin (1972)	85.4	8.3	6.3
Protein, 18	Latter (1980)	85.5	5.5	9.0
Protein, 25	Ryman et al. (1983)	86.0	2.8	11.2
Autosomal DNA, 109	Barbujani et al. (1997)	84.4	4.7	10.8
Autosomal DNA, 90	Jorde et al. (2000)	84.8	1.6	13.6
Autosomal DNA, 21	Romualdi et al. (2002)	82.9	8.2	8.9
Autosomal DNA, 377	Rosenberg et al. (2002)[b]	93.2	2.5	4.3
Autosomal DNA, 377	Rosenberg et al. (2002)[c]	94.1	2.4	3.6
Alu insertions, 100	Watkins et al. (2003)	88.6	1.9	9.6
X chromosome, 17	Ramachandran et al. (2004)	90.4	4.6	4.9
Autosomal Insertion/deletions, 40	Bastos-Rodrigues et al. (2006)	85.7	2.3	12.1
Median, all loci		85.5	2.8	9.0
Median, normalized[a]		87.9	2.9	9.2

[a]These values were obtained by considering all studies equally informative. The medians of their results at the three levels of population subdivision were normalized by dividing them by the total (97.3) and multiplying by 100.
[b]Considering five groups: Africa, Eurasia, East Asia, Oceania, and Americas.
[c]Considering the seven groups obtained by further subdividing Eurasia into Europe, Middle East, Central/South Asia.

variable regions in human DNA belong to two main types: either specifically African or generically human. In addition, and in agreement with Livingstone's (1962) intuition, different sets of polymorphisms, and especially Copy Number Variants (CNVs) versus SNPs, show different geographic distributions (Redon et al., 2006), and hence would suggest contradictory subdivisions of individuals and populations (Jakobsson et al., 2008).

In most genomic studies, the global variance is estimated at two levels only, namely, between members of the same population and between populations. The latter component, essentially the sum of the last two columns of Table 2, is summarized by Wright's F_{ST}-statistic. Studies differ in terms of the number of populations and individuals considered, the number of polymorphisms and the class of polymorphisms, whether SNPs or CNVs (Table 3). In principle, values of F_{ST} range from 0 (indicating that all populations are identical) to 1 (indicating maximum genetic differences between populations). When estimated from human genomic data, they are generally close to the normalized median of Table 3, that is, around 0.12 or 0.13, with a minimum of 0.05 (Auton et al., 2009). This confirms that they account for roughly 13% of the species' genetic variance, with the remaining 87% representing, on average, the difference between individuals of the same population. Once again, different loci vary greatly in their F_{ST}s, even when they map on the same chromosome (Coop et al., 2009).

Analysis of interpopulation differences shows that both allele frequencies and DNA sequences vary nonrandomly in space, often forming gradients over entire continents (Ramachandran et al., 2005; Liu et al., 2006). We also know that zones of sharp genetic change, or genetic boundaries, exist and tend to overlap with various kinds of migrational barriers, both physical (e.g., mountain ranges or large bodies of water) and cultural (especially language boundaries) (Sokal et al., 1990). However, very seldom do genetic boundaries form complete rings around single populations or groups of populations. Virtually all populations studied are somewhat different from their neighbors, but continuous variation is the rule over much of the planet, with the genetic characteristics of the various regions changing smoothly in the geographical space.

Table 3 Genomic estimates of F_{ST} for the global human population

Number of markers	*Samples*	F_{ST}
599,356 SNPs	209 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.13
1,034,741 SNPs	71 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.10
1,007,329 SNPs	269 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.12
443,434 SNPs	3845 worldwide-distributed individuals	0.052
2,841,354 SNPs	210 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.11
243,855 SNPs	554 individuals from 27 worldwide populations	0.123
100 *Alu* insertions	710 individuals from 23 worldwide populations	0.095
1,447 CNVs covering 12% of the genome	270 individuals from 4 populations: Caucasian, Chinese, Japanese, Yoruba	0.11

From Barbujani, G., Colonna, V., 2010. Human genome diversity: frequently asked questions. Trends in Genetics 26, 285–295; where the original references can be found

The genetic differences between populations, around 13% of the global species variance, are small, but not zero. Based on these differences, individuals can often be assigned to populations. By selecting polymorphisms among the most variable of the genome, Witherspoon et al. (2007) could correctly classify most people from different geographical origins. However, the error rate remained substantial even in comparisons between continents, such as sub-Saharan Africa and Europe. The accuracy of the assignment depends on the degree of differentiation between populations, on the number of populations being compared (the more, the higher the error), and on the number of polymorphisms considered (Romualdi et al., 2002; Witherspoon et al., 2007).

On the other hand, the genetic differences between populations are not zero, but they are small. Once the analysis of highly variable polymorphisms has shown that a certain individual has a high probability of belonging to a certain population, predicting that individual's genotype at other loci is problematic because different polymorphisms show different patterns of variation. Despite that difficulty, and although few statisticians would try to discriminate groups when the variance within them is 8–9 times as large as that between them, identifying even moderately differentiated genetic groups could be of potential relevance for clinical practice (see Section Races in Medicine). To be useful, however, this exercise should produce a stable classification, that is, a consistent list of biological groups, independent of the markers studied. So far, that has not been the case.

A popular approach to this question is based on an algorithm, STRUCTURE, assigning individual genotypes to an arbitrary number of groups, k. Various k values are tested, and their likelihood is eventually estimated. In the first, and most influential, worldwide analysis based on STRUCTURE, Rosenberg et al. (2002) typed 377 Short Tandem Repeats (STRs) in 1056 individuals from 52 populations. They recognized six clusters, five of them corresponding to continents or subcontinents, and the sixth to a genetic isolate in Pakistan, the Kalash. In general, individuals of the same population fell consistently in the same cluster, or shared similar membership coefficients in two clusters. The authors concluded that self-reported ancestry contains information on DNA diversity, and hence that an objective clustering of genotypes is possible, if large amounts of data are considered.

In fact, clustering is possible, but it is not consistent across studies. In a subsequent paper based on a larger assemblage of data, the same authors failed to confirm the Kalash as a separate unit; instead, the Native American populations were this time split into two clusters (Rosenberg et al., 2005). The Kalash resurfaced as a distinct group when 15 Indian populations were added to the analysis, leading to the identification of seven clusters, with most populations of Eurasia now showing multiple memberships. In these studies, all African genotypes

formed a single group, a result that may point to a weakness in the methodology, since there is broadly replicated evidence of a deep population subdivision within Africa (Hunley et al., 2009). Indeed, when the same dataset was analyzed by a method that highlights zones of sharp genetic change or genetic boundaries, Africa appeared to be subdivided into four groups, and each American population formed an independent group, giving a total of 11 (Barbujani and Belle, 2006). In Jakobsson et al.'s (2008) study of more than half a million SNPs, STRUCTURE indicated different clusterings if the SNPs were individually analyzed or if they were combined to form haplotypes. In turn, both inferred clusterings were inconsistent with those inferred from CNVs in the same individuals.

In practice, no matter whether inferred from protein variation or from recent genomic analyses, the proportion of human genetic variance due to differences between populations is modest. Differences exist and have significant evolutionary implications, but it is an error to consider them typical of populations, and not of individuals, because genetically different individuals coexist in the same population. When the number of markers is large, many dissimilarities are detected, and a fraction of these are likely to achieve statistical significance. However, minor differences in the markers considered, in the sample distribution, or in the method of analysis, lead to different clusters.

An illustration of the consequences of this aspect of human diversity comes from comparisons of complete individual genomes. Among the first people whose genome was entirely sequenced are James Watson and Craig Venter, both U.S. geneticists of European ancestry. Their genomes share 1 715 851 SNPs, which is less than each of them has in common with the genome of the Korean scientist Seong-Jin Kim (1 824 482 and 1 736 340, respectively) (Ahn et al., 2009). The fact that a person of Asian ancestry (Kim) is genetically intermediate between two persons of European ancestry (Watson and Venter) does not mean that on average Europeans are genetically closer to random Koreans than to each other, but highlights the limitations of such coarse categorizations. Populations are structured in the geographical space, Asians and Europeans appear well separated in analyses of population structure, but when it comes to predicting individual DNA features, labels such as 'European' and 'Asian' are misleading, because members of the same group, Watson and Venter in this case, may be very different. If what matters for future research is whether by racial labeling we can approximate what is in a person's genome, it is already clear that the answer is no (Xing et al., 2010). As a matter of fact, the best way to know what is in a person's DNA is to study that person's DNA. In the past, this task was technically challenging, but now it is both feasible and cheap.

Comparisons with Other Species

We are not alone on Earth; examining patterns of population diversity in our own and other species may help us put human diversity in broader evolutionary perspective. Among primates, humans have the lowest levels of DNA diversity, in two different ways. First, comparisons of X-chromosome sequences show that individual variation in humans is half that observed in gorillas, and less than one-third that observed in orangutans and chimpanzee (Kaessmann et al., 2001). Second, F_{ST} values estimated between gorilla ($F_{ST} = 0.38$) and chimpanzee ($F_{ST} = 0.32$) populations are roughly three times as high as most human estimates ($F_{ST} \leq 0.13$, Table 3). Thus, not only are individual differences in humans much smaller than between individuals of other primate species, but also these small differences are mainly accounted for by individual differences within populations. In other words, we are subdivided into more closely related populations than any of our evolutionary relatives (with the possible exception of bonobos, *Pan paniscus*), despite the fact that humans occupy a much broader geographic area.

Templeton (1998) extended this comparison to 12 large-bodied, highly mobile mammals. Two such species (East African Buffalo and Kenyan Waterbuck) show smaller between-population diversity than humans. However, these species are not cosmopolitan like ours, but dwell in restricted geographical zones. All other species compared, including elephant, gazelle, and gray wolf, show higher, or much higher, differentiation among populations and groups, with the values of those variance components reaching 80% of the total. The limited degree of differentiation among human populations does not suggest a history of long-term isolation and differentiation, which was likely the case for other mammals.

Dogs and horses are sometimes cited as examples of species subdivided in clear races or breeds. In comparing them with humans, one should keep in mind that breeds evolved in domesticated animal as a result of human-driven artificial selection only in the last 10 000 years or so, whereas human populations might have diverged for many tens of thousands years. Despite that, F_{ST} among dog breeds is nearly three times as large as in humans, around 0.35 (Von Holdt et al., 2010). F_{ST} has not been estimated yet for the whole horse species, but for both dogs and horses estimates of linkage disequilibrium are available—that is, a measure of the nonrandom association of alleles along the chromosomes. Population-genetic theory demonstrates that linkage disequilibrium increases after a sharp reduction of population size, such as that resulting from the selection of a few individuals to start a new breed (under artificial selection) or whenever a population evolves in isolation (under natural conditions). Levels of linkage disequilibrium are highest in dogs, intermediate in horses, and least in humans. These figures also depend on the distance along the chromosome between DNA regions; considering, for example, regions separated by 0.2 million bases, linkage disequilibrium in dogs is six times as large as in humans, and in horses four times as large as in humans (Wade et al., 2009).

In short, data on genome diversity in mammal species, either evolving under natural conditions (such as elephant, chimpanzee, and orangutan) or under artificial selection (such as dog and horse) show significant differences between these species and humans. Humans show the highest degree of allele sharing between populations, even distant ones; several features of human DNA diversity seem to indicate a comparatively recent origin from a small number of founders, who differentiated while dispersing to colonize the whole planet, and whose populations often exchanged migrants. This history has produced a pattern of genetic variation that is inconsistent with the existence of independently evolving groups (Hunley et al., 2009).

Races in Medicine

The repeated observation that population affiliation is not a reliable predictor of an individual's specific genotype (Xing et al., 2010), has led to a reconsideration of the utility of terms such as 'race,' 'racial differences,' and 'ethnic background' as proxies in medical decisions (e.g., for decisions on the kind of treatment, dosage, etc., but also for diagnosis). Schwartz (2001) wrote that there is no plausible biological justification for racial distinctions. Risch et al. (2002) responded by claiming that they could demonstrate the validity of racial self-categorization 'from an objective perspective.' In fact, they proposed verbal arguments supporting the view that each continent corresponds to a distinct genetic cluster of populations. In addition, they defended the status of race for Hispanics, a group composed of people of different origins who have two features in common, being immigrants in the United States and speaking Spanish, neither of which has a genetic basis. Despite all these problems inherent in using race as shorthand for a cluster of people supposedly sharing the same biological characteristics, racial categories are routinely employed by U.S. medical doctors, scientists, and community health workers. On 10 July 2012, a PubMed search of scientific literature using the term 'human races' returned 133 129 items, and about 1000 new entries are being added every month.

As mentioned above, one reason why medical doctors hesitate to abandon racial categorizations may be the fact that different diseases have different prevalence in different populations. Actually, much like any other allele, alleles associated with genetic pathologies or predisposition to diseases do vary in frequency. Because these frequencies are low, these alleles tend to show a more restricted geographical distribution, and many of them can be regarded as population-specific; examples include many hemoglobin variants, the *HEXA* mutation causing Tay-Sachs disease in Ashkenazi Jews, and the Mendelian alleles associated with the so-called Finnish disease complex. However, in all these cases, the pathologic alleles are carried by a small fraction of the chromosomes in the population, typically less than 1.5%. Defining Ashkenazi Jews as the genetic group carrying the Tay-Sachs disease allele would mean to exclude from the group the 98% of them who do not carry the *HEXA* mutation.

While recognizing that distinguishing groups of humans from each other on genetic grounds is problematic, Harpending (2007) argued that the race concept is indispensable to understand the causes of human disease. He then proposed to regard as races human groups among which genetic differences exist. Since some degree of genetic differentiation has been found between any two sets of humans compared so far, be they arbitrary or defined by geographical, cultural, or epidemiological criteria, Harpending's definition implies that any human population is a race. The meaning of words is conventional, and so one could be tempted to settle the issue by deciding to henceforth use race and population as synonymous. However, that does not seem a good idea. In science, and not only in science, communication is possible when all communicators accept some common criteria. 'Thank you' denotes a higher level of gratitude than 'Get lost.' One may decide to express his gratitude by saying 'Get lost,' but shifting to that convention will harm his social life. Giving the status of race to all populations does not seem to make much more sense.

It is possible that with time the interest in racial categorization will decline also clinical practice as well. Ever since the development of a scientific approach to medicine, the purpose of therapeutic strategies has been to cure the 'average patient,' but this average patient is a statistical abstraction. Real patients have different disease risks and respond differently to pharmaceutical treatment. In a sense, then, the so-called racial medicine is an attempt to come to terms with these differences and somehow take them into account. But it is a scientifically superficial attempt, based on the erroneous assumption that people classified in the same racial category are genetically homogeneous. Contrary to the claim that racial stereotypes capture some meaningful aspects of biological variation, the available data suggest that a good way to predict whether certain individuals will have certain health risks or will benefit from pharmaceutical treatment is to study their genes. Individuals vary in their drug response as a consequence of variation at several genes coding for the enzymes responsible for drug metabolism. In a classical study of CYP2D6 (Ingelman-Sundberg et al., 2007), the best-known drug-metabolizing enzyme, Chinese and Swedes were shown to differ in their average rates of metabolism, with an excess of slow-metabolizing individuals in the Swedish sample. However, both populations included the whole range of possible phenotypes, from poor to ultra-rapid metabolizers (Johansson et al., 1991). Therefore, a racial-medicine approach consisting in looking for a dosage of the drug that will make everybody happy in Sweden or China is a waste of time. What is conceivable and, in the future, hopefully feasible, is to identify the alleles responsible for individual variation, hence giving to each person, no matter whether Chinese or Swedish, the drug dosage that fits the individual therapeutic needs. This is part of a recent, serious attempt to replace the use of proxies such as gender, age, and race in medical decision making by a multitude of integrated data characterizing individuals at the molecular level (Committee on a Framework for Developing a New Taxonomy of Disease, 2011).

Racial Variation in Intelligence?

Very little can currently be said about the inheritable bases of variation in cognitive abilities. Biological determinism, that is, the idea that what we are and can possibly be is written in our genes, reached its most radical modern expression with the best seller *The Bell Curve*. In that book, Herrnstein and Murray (1994), respectively, a psychologist and a political scientist, claimed that intelligence is accurately measured by the Intelligence Quotient (IQ); that the IQ is genetically determined, constant through life, correlated with race, and basically independent of any social or economical factors; and that a 'cognitive elite' of biologically gifted people is endangered by the U.S. social policies, transferring resources from the rich (and therefore intelligent) to the poor, thus favoring a higher fertility of biologically inferior people. Whatever one might think of the political aspects of Herrnstein and Murray's proposal, its biological bases are phenomenally flawed.

In a sense, biological determinism reflects a deep faith in the ability of biology to explain everything. Unfortunately, that faith

is not justified. To stay within the realm of genetics, we understand well how single genes work, and we can predict the occurrence of many diseases caused by single-gene defects. However, most diseases depend on the interaction of many genes and environmental factors. We also have a good understanding of many gene-to-gene and protein-to-gene interactions, but as the number of interacting genes and factors in the environment exceeds a few units, the available models become unsuitable. That is why we cannot predict the onset of Type 2 diabetes, to which at least 36 genes are known to contribute (Herder and Roden, 2011). The onset of most cancers, cardiovascular, and neurodegenerative diseases depends on greater numbers of genes than diabetes. Understanding variation for common traits such as height, weight, or physical performance is clearly more complicated than predicting disease onset (see Weedon et al., 2008), and by far an even more complex task would be to understand the genetic determinants of human cognitive abilities. More than 1500 genes are expressed in the frontal brain cortex and change their level of expression during fetal or infant development (Colantuoni et al., 2011). This means that, even if what we call intelligence depended only on the frontal cortex (which we can safely rule out), to account for its genetic bases we ought to consider thousands of factors – at present, a totally unmanageable number.

The effects of this enormous collection of genes on individual cognitive abilities are just beginning to be investigated. With regard to genetic differences in intelligence between human groups, we just do not know how to start addressing the problem in scientifically respectable terms. Based on what we know, nobody can seriously claim that there is any evidence that social and economic differences between human groups, defined on the basis of race, class or sex, stem from inherited, inborn differences, and hence that social stratification is a fair or inevitable reflection of biological hierarchies.

Tentative Conclusions and Future Prospects

Population-genetics theory and empirical data show that genetically differentiated groups tend to evolve when genetic drift affects populations connected by little (or no) migratory exchanges. In that case, the populations' internal diversity will decrease as alleles are lost by chance and are seldom reintroduced by gene flow; because in each population different alleles are lost and different mutations occur, genetic boundaries may arise. Conversely, boundaries are unlikely to develop if the effects of gene flow exceed those of drift.

All humans are genetically different, but our species shows lower degrees of interpopulation differentiation than most other mammal species. Geographical patterns in human genome diversity exist, but each human population harbors a large share of the species' diversity; that diversity is largely continuous in space, rather than interrupted by boundaries. This probably means that subdivision and isolation were more the exception than the rule in the human past. Taken together with the fact that no agreement has ever been found on the number and definition of human races, this finding casts radical doubts on the race concept as a useful tool for describing human biodiversity.

Now that hundreds of whole genome sequences have been studied, there is no reason to suspect that the comparatively weak geographical structure of humans might be an artifact, due to our limited ability to detect the alleles that make us, apparently, so different. Thus, if we are to understand our demographic history, a good starting point is to ask what combination of evolutionary factors has led to the coexistence in the same populations of a broad spectrum of genetically different individuals. This way, we might be able to reconstruct aspects of our migratory history that still escape us. The fossil record documents an origin of anatomically modern humans in Africa, more than 100 000 years ago. We do not know yet if they expanded through Palestine in a single wave, some 60 000 years ago (Ramachandran et al., 2005; Liu et al., 2006), or if there was an earlier exit through a Southern route, that is, through the horn of Africa into Southern Asia and Australia (Lahr and Foley, 1994). A better understanding of the times and tempos of dispersal from Africa is crucial for testing the hypothesis of a possible (if limited) hybridization between anatomically archaic and anatomically modern humans, recently supported by the publication of a draft sequence of the Neandertal genome (Green et al., 2010). Finally, different hypotheses exist on the details of the peopling of several areas of the world, and by carefully investigating current and past genomic diversity we might be able to reject some of these hypotheses.

Envisaging our species as genetically subdivided in differentiated races is of no use for addressing questions of this kind, and leads to errors in clinical practice. Still, race remains an important component of our social and psychological world. This component affects human interactions and social policies, and will not vanish just because some scientists say it should. In a sense, then, races do exist, and they should be kept into consideration. However, on the basis of what is known, they exist only in the sense that the labels we stick on ourselves and on the others have practical consequences, even if they do not correspond to empirically identifiable biological realities (Glasgow, 2009). Should new biological evidence emerge, this conclusion will have to be reconsidered. However, at present, the burden of the proof is definitely on those supporting the existence of a racial structure in the human species.

See also: Genomics, Ethical Issues In; Human Evolutionary Genetics; Human Genome Diversity Project: History; Human Genome Project: History and Assessment.

Bibliography

Ahn, S.M., et al., 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Research 19, 1622–1629.

Auton, A., et al., 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Research 19, 795–803.

Barbujani, G., Belle, E.M., 2006. Genomic boundaries between human populations. Human Heredity 61, 15–21.

Barbujani, G., Colonna, V., 2010. Human genome diversity: frequently asked questions. Trends in Genetics 26, 285–295.

Barbujani, G., Magagni, A., Minch, E., Cavalli-Sforza, L.L., 1997. An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences USA 94, 4516–4519.

Bastos-Rodrigues, L., Pimenta, J.R., Pena, S.D.J., 2006. The genetic structure of human populations studied through short insertion-deletion polymorphisms. Annals of Human Genetics 70, 658–665.

Brown, R.A., Armelagos, G.J., 2001. Apportionment of racial diversity: a review. Evolutionary Anthropology 10, 34–40.

Brues, A.M., 1977. People and Races. Macmillan, New York.

Clarke, L., et al., 2012. The 1000 genomes project: data management and community access. Nature Methods 9, 459–462.

Cohen, C., 1991. Les races humaines en histoire des sciences. In: Hublin, J.J., Tillier, A.M. (Eds.), Aux Origines D'Homo Sapiens. Presses Universitaires de France, Paris, pp. 9–47.

Colantuoni, C., et al., 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature 478, 519–523.

Committee on A Framework for Developing a New Taxonomy of Disease, 2011. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. The National Academies Press, Washington.

Coon, C., 1962. The Origin of Races. Knopf, New York.

Coon, C.S., Garn, S.M., Birdsell, J.B., 1950. Races. C. Thomas, Springfield, Illinois.

Coop, G., et al., 2009. The role of geography in human adaptation. PLoS Genetics 5, e1000500.

de Queiroz, K., 2005. Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences USA 102 (Suppl. 1), 6600–6607.

Dobzhansky, T., 1973. Genetic Diversity and Human Equality. Basic Books, New York.

Gabriel, S.B., et al., 2002. The structure of haplotype blocks in the human genome. Science 296, 2225–2229.

Garn, S.M., 1961. Human Races. C. Thomas, Springfield, Illinois.

Glasgow, J., 2009. A Theory of Race. Routledge, New York.

Green, R.E., et al., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722.

Harpending, H.C., 2007. Anthropological genetics: present and future. In: Crawford, M. (Ed.), Anthropological Genetics. Cambridge University Press, Cambridge, pp. 456–466.

Herder, C., Roden, M., 2011. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. European Journal of Clinical Investigation 41, 679–692.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.

Holland, B.S., Cowie, R.H., 2007. A geographic mosaic of passive dispersal: population structure in the endemic Hawaiian amber snail *Succinea caduca* (Mighels, 1845). Molecular Ecology 16, 2422–2435.

Hunley, K.L., Healy, M.E., Long, J.C., 2009. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. American Journal of Physical Anthropology 139, 35–46.

Ingelman-Sundberg, M., Sim, S.C., Alvin, A., Rodriguez-Antona, C., 2007. Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacology & Therapeutics 116, 496–526.

Jakobsson, M., et al., 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998–1003.

Johansson, I., Yue, Q.Y., Dahl, M.L., Heim, M., Säwe, J., Bertilsson, L., Meyer, U.A., Sjöqvist, F., Ingelman-Sundberg, M., 1991. Genetic analysis of the interethnic difference between Chinese and Caucasians in the polymorphic metabolism of debrisoquine and codeine. European Journal of Clinical Pharmacology 40, 553–556.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T., Batzer, M.A., 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. American Journal of Human Genetics 66, 979–988.

Kaessmann, H., Wiebe, V., Weiss, G., Pääbo S, 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature Genetics 27, 155–156.

Lahr, M.M., Foley, R.A., 1994. Multiple dispersals and modern human origins. Evolutionary Anthropology 3, 48–60.

Latter, B.D.H., 1980. Genetic differences within and between populations of the major human subgroups. American Naturalist 116, 220–237.

Levin, M., 2002. Races unreal? Mankind Quarterly 42, 413–417.

Lewontin, C., 1972. The apportionment of human diversity. Evolutionary Biology 6, 381–398.

Liu, H., Prugnolle, F., Manica, A., Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history. American Journal of Human Genetics 79, 230–237.

Livingstone, F.B., 1962. On the nonexistence of human races. Current Anthropology 3, 279–281.

Madrigal, L., Barbujani, G., 2006. Partitioning of genetic variation in human populations and the concept of race. In: Crawford, M.H. (Ed.), Anthropological Genetics. Theory, Methods and Applications. Cambridge University Press, Cambridge, pp. 19–37.

Mayr, E., 1947. Systematics and the Origin of Species, third ed. Columbia University Press, New York.

Molnar, S., 1998. Human Variation. Races, Types and Ethnic Groups. Prentice Hall, Upple Saddle River, NJ.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., Cavalli-Sforza, L.L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences USA 102, 15942–15947.

Ramachandran, S., Rosenberg, N.A., Zhivotovsky, L.A., Feldman, M.W., 2004. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Human Genomics 1, 82–97.

Redon, R., et al., 2006. Global variation in copy number in the human genome. Nature 444, 444–454.

Relethford, J.H., 2003. The Human Species. An Introduction to Biological Anthropology, fifth ed. McGraw –Hill, New York.

Riccioni, G., et al., 2010. Spatio-temporal population structuring and genetic diversity retention in depleted Atlantic Bluefin tuna of the Mediterranean Sea. Proceedings of the National Academy of Sciences USA 107, 2102–2107.

Risch, N., Burchard, E., Ziv, E., Tang, H., 2002. Categorization of humans in biomedical research: genes, race and disease. 1. Genome Biology 3 (7) comment2007.

Romualdi, C., Balding, D., Nasidze, I.S., Risch, G., Robichaux, M., Sherry, S.T., Stoneking, M., Batzer, M.A., Barbujani, G., 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Research 12, 602–612.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., Feldman, M.W., 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genetics 1, e70.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. Genetic structure of human populations. Science 298, 2381–2385.

Ryman, N., Chakraborty, R., Nei, M., 1983. Differences in the relative distribution of human gene diversity between electrophoretic and red and white cell antigen loci. Human Heredity 33, 93–102.

Schwartz, R.S., 2001. Racial profiling in medical research. New England Journal of Medicine 344, 1392–1393.

Sokal, R.R., et al., 1990. Genetics and language in European populations. American Naturalist 135, 157–175.

Templeton, A.R., 1998. Human races: a genetic and evolutionary perspective. American Anthropologist 100, 632–650.

Varki, A., Altheide, T.K., 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Research 15, 1746–1758.

Vogel, F., Motulsky, A.G., 1986. Human Genetics: Problems and Approaches, second ed. Springer-Verlag, Berlin, p. 534.

Von Holdt, B.M., et al., 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464, 898–903.

Wade, C.M., et al., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867.

Watkins, W.S., et al., 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Research 13, 1607–1618.

Weedon, M.N., et al., 2008. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40, 575–583.

Weeks, M., Lupfer, M.B., 2004. Complicating race: the relationship between prejudice, race, and social class categorizations. Personality & Social Psychology Bulletin 30, 972–984.

Witherspoon, D.J., Wooding, S., Rogers, R.A., Marchani, E.E., Watkins, W.S., Batzer, M.A., Jorde, L.B., 2007. Genetic similarities within and between human populations. Genetics 176, 351–359.

Xing, J., Watkins, W.S., Witherspoon, D.J., Zhang, Y., Guthery, S.L., Thara, R., Mowry, B.J., Bulayeva, K., Weiss, R.B., Jorde, L.B., 2010. Fine-scaled human genetic structure revealed by SNP microarrays. Genome Research 19, 815–825.

Race Identity

Clifford L Broman, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by T. Duster, volume 19, pp. 12703–12706, © 2001, Elsevier Ltd.

Abstract

This article discusses the research on race identity. We examine some of the more significant behavioral research traditions that have addressed this issue. Issues such as the development of racial identity, white identity, and current and future trends in the broader study of identity are also addressed.

Politics and Identity Politics

In every society, racial identity is generated by massive historical, economic, and political forces that far transcend the level of individual choice (Banton, 1977; Miles, 1989). Racial identity is socially constructed in a society as a consequence of these forces. By social construction, we refer to the fact that a society defines racial groups and defines who is in them. In attempting to explain the differences between people, Europeans as far back as the sixteenth century attempted to classify people according to skin color (Jordan, 1968). Though the concept of race as we know it today was slow to develop, by the nineteenth century, relatively clearly defined racial groups existed throughout much of the world. In the worldwide quest for white supremacy, those with the power to define themselves and others thrust a racial label on themselves, 'white,' and a contrasting label on those they considered 'not white,' such as blacks, Indians, and others. This was the result of politics and differential societal power, not biological or medical science.

Indeed, there is overwhelming evidence that race is arbitrary, shifting, and often biologically and sociologically inconsistent and contradictory (Corcos, 2014). A consortium of leading scientists across the disciplines from biology to physical anthropology issued a 'Revised UNESCO Statement on Race' in 1995 – a definitive declaration that summarizes 11 central issues and concludes that in terms of 'scientific' discourse, there is no such thing as a 'race' that has any scientific utility: "… Many scholars now acknowledge the fact that the biological concept of race as a means of characterizing human populations is so flawed that it is no longer a scientifically valid concept. In fact, the statement makes clear that the biological concept of race as applied to humans has no legitimate place in biological science" (Corcos, 2014; Katz, 1995: pp. 4, 5).

However, the power of social construction leads to a social definition of racial groups and who is in those groups (Omi and Winant, 1986). This leads people to come to identify with a group. Again, a group label is thrust upon people, as a consequence of marking people as different from one another and as a vehicle for achieving group solidarity (Kelley, 1997; Fredrickson, 1995; Crenshaw et al., 1995). There is an important distinction between being labeled as a member of a particular group and psychologically identifying with that group. For example, while one may be a member of the Jones family, one may not feel a psychological connection with the Jones family. Thus, racial labeling is not necessarily the same as racial identification.

Racial Identity and Individual Identity

Assimilation into a racial identity involves two components: racial group identification and racial group consciousness. Racial identity is defined as one's sense of self that is related to racial group membership (Belgrave et al., 2000). Racial identity is a multidimensional construct that includes the strength of one's identification with one's racial group, a sense of attachment to other group members, an evaluation of group membership (e.g., how much the individual likes or dislikes being white, for example) and may include group-relevant attitudes and behaviors (Allen et al., 1989; Phinney, 1996; Sellers et al., 1997). Foundational studies of racial group consciousness argued that group consciousness also implied an action orientation and a set of political beliefs attached to the racial group (Gurin et al., 1980). In other words, some argue there is an ideological meaning attached to a racial identity (Gurin et al., 1980).

There is a large body of research going back to the Clark and Clark (1947) doll studies that has investigated a set of constructs that purport to measure racial identity. They are sometimes labeled differently, such as self-consciousness, racial group identification, racial group perceptions, and acculturation (Baldwin and Bell, 1985; Broman et al., 1988; Landrine and Klonoff, 1996a,b; Parham and Helms, 1981; Utsey, 1998). A key commonality for many of these constructs is the sense of strong psychological and emotional connection to being a member of the racial group. For many scholars, this sense of strong psychological and emotional connection defines racial identity (for example, see Broman et al., 1988; Cross et al., 1998; Landrine and Klonoff, 1996a; Sellers et al., 1997). Thus, an individual who does not feel this strong sense of emotional and psychological connection is commonly regarded of as not identifying with the racial or ethnic group.

The Development of Racial Identity

Two of the most important factors in the development of racial identity are racial discrimination and interracial contact. Several studies have focused on racial discrimination and its

impact on African-Americans and other minorities (Broman et al., 2000; Chou and Feagin, 2014; Essed, 1991; Feagin, 1991; Kessler et al., 1999). Several studies have shown that racial discrimination is relatively pervasive (Broman, 1996; Broman et al., 2000; Essed, 1991; Feagin, 1991; Kessler et al., 1999; Swim et al., 2003). Racial discrimination is thought to reinforce the racial label applied to an individual and to groups, and the boundaries between groups, and to also increase the tendency to racially identify with the group (Branscombe et al., 1999).

Interracial contact is also an important factor in racial identity. Studies in the tradition of the 'contact hypothesis' have long suggested that interracial contact affected both majority and minority groups (Allport, 1954; Deutsch and Collins, 1951; Wilner et al., 1955). Early studies found support for the hypothesis, but more sophisticated studies found more limited effects (Ellison and Powers, 1994). Scholars have also noted the reverse: the lack of interracial contact corresponds to racial insulation. Studies examining this phenomenon have found that the lack of interracial contact, or racial insulation, affects both Black self-esteem and racial identity (Broman et al., 1989; Rosenberg, 1979). The insulation hypothesis argues that because of the segregation in US society, many minorities are insulated from the negative esteem, racism, and discriminatory behaviors of Whites (Rosenberg, 1979; Simmons, 1978). The significant, day-to-day interactions that many minorities have occur in primarily minority contexts. It is only in dissonant contexts, where minorities are in primarily White contexts, that they are subject to the direct expression of racial hostility, experiencing themselves as 'marginal,' and more likely to compare themselves to Whites rather than to other minorities. Thus, interracial contact – the amount and the context – is argued to have direct impact on racial identity.

Building on the idea that people develop identities based on how other people see them and treat them, early researchers focused on low self-esteem and self-hatred as key components of black identity (Horowitz and Murphy, 1938; Clark and Clark, 1947). Given that African Americans commonly experienced exclusion and discrimination by whites and were generally stigmatized in American society, blacks were thought to internalize this societal rejection and devaluation leading to self-hatred and low self-esteem. This view of black identity came under intense criticism in the wake of the black power movements of the 1960s, and during the 1970s researchers shifted to an emphasis on resilience and formulated new theoretical models that emphasized the development of strong, healthy black identities in opposition to white racial oppression. Cross (1971, 1991), for example, proposed a model of Nigrescence by which African Americans proceeded in five stages to shift from a non-Afrocentric to an Afrocentric identity. The five stages were preencounter, encounter, immersion-emersion, internalization, and internalization-commitment and Hall et al. (1972) developed a 55-item questionnaire to measure them.

Parham and Helms (1981) built on Cross's work, theorizing that the critical period for identity formation was late adolescence and young adulthood and developing the 30-item Racial Identities Attitudes Scale to measure the preencounter, encounter, immersion-emersion, and internalization stages of Nigrescence. They argued that progress through the successive stages was not inevitable and that identity formation could stagnate or regress back to earlier stages of development, a process they called recycling. Milliones (1980) saw black identity developing in four stages – preconscious, confrontation, internalization, and integration – and developed the 65-item Developmental Inventory of Black Consciousness (DIBC) to assess them. Research done using the DIBC found that high preconscious scores (indicating low black consciousness) were associated with low self-esteem and dissatisfaction with appearance (Smith et al., 1991) and that high confrontation scores (indicating high black consciousness) were associated with unwillingness to share information with other African Americans (Taylor and Rogers, 1993).

Baldwin (1985) developed his own theory about the structure and content of black personality and created the 42-item African Self Consciousness Scale (ASC) to assess individual awareness of African identity and culture, the degree to which African survival is recognized as a paramount goal, the extent of appreciation for African institutions and artifacts, and awareness of people and things that are African and non-African. Not surprisingly, students attending black colleges and universities that tend to have higher ASC scores than those attending predominantly white institutions (Baldwin et al., 1987). Jackson et al. (1988) included a variety of items on the National Survey of Black Americans to assess feelings of closeness to a variety of different kinds of African Americans and used them to create a scale of black identity. Older, less-educated, urban African Americans tend to score high on this scale, as do blacks who came of age in schools, neighborhoods, and workplaces characterized by high proportions of African Americans (Broman et al., 1989).

Demo and Hughes (1990) factor analyzed items from the National Survey of Black Americans and argued that African American identity had three components: closeness, black separatism, and black group evaluation. Closeness indicates the degree to which respondents feel that their feelings and thoughts are similar to those of other African Americans; black separatism indicates the extent of commitment to black culture and inclination to restrict social relationships to other African Americans; and black group evaluation is the degree to which respondents believe that African Americans display positive characteristics. They found that integration into mainstream American society is associated with less in-group solidarity but that mainstream integration and higher economic status enhance the positive evaluation of African Americans.

In their analysis, Allen et al. (1989) identify five dimensions of black racial identity: closeness to elite African Americans, closeness to common African Americans, positive stereotypes, negative stereotypes, and black autonomy, with the latter being close to the black separatism component of Demo and Hughes (1990). According to Allen et al. (1992), lower-status African Americans hold more negative stereotypes about their group and feel closer to elite blacks than those of higher status, but status had no effect on black autonomy or belief in positive stereotypes. In contrast, higher-status African Americans were less supportive of black autonomy and felt distant from lower class members of their group.

Sanders-Thompson (1991, 1995) also proposes a multidimensional model of black racial identity, though not one that is developmental in the manner of Cross (1971, 1991). Her model consists of four components. The physical dimension assesses the degree of acceptance of African American physical attributes; the cultural dimension measures awareness of black cultural and social practices; the sociopolitical dimension indicates awareness of black social and political issues; and the psychological dimension reflects pride and belongingness to African Americans as a social group. Her Multidimensional Racial Identification Scale presently includes 30 items (Myers and Sanders-Thompson, 1994). She found that physical identity was associated with experiences of racism and greater racial socialization within the family. Psychological racism was associated with experiences of discrimination and greater involvement in the black community; and sociopolitical and cultural identity both reflect greater community involvement and family socialization (Sanders-Thompson, 1991, 1994, 1995).

Like Sanders-Thompson, Sellers et al. (1997) propose a multidimensional model of black racial identity that incorporates four components: racial centrality (whether race is a core part of individual identity over time), racial salience (whether race is an important part of an individual identity at a point in time), racial ideology (philosophical beliefs about how African Americans should relate to American society), and racial regard (evaluations about an individual's own race). The ideology dimension has four subscales: nationalist (emphasizing the uniqueness of African origins), oppressed (emphasizing commonality with other disadvantaged groups), assimilationist (emphasizing similarities between blacks and other Americans), and humanist (emphasizing the similarity of all human beings). They found that respondents who reported a black best friend displayed high values on the centrality and nationalist scales and lower values on the assimilationist and oppressed scales.

In their evaluation of the various black racial identity scales, conclude that "there is no consensus as to the best way to conceptualize or measure racial identity in African Americans." For members of racial groups other than African-Americans, there has been little research on how they might racially identify, though Phinney (1992) has proposed a general model of ethnic identity. There is more research upcoming on white racial identity, and the literature on acculturation is, at least peripherally concerned with Latino identification. In the past, it is plausible that there was simply no justification for the pursuit of such a research question for whites and other racial or ethnic groups.

White Identity

In recent years, the concept of white racial identity has come into increasing focus as a basis for a research agenda. Blumer's (1958) classic article reframed race as a structural relationship between competing group interests. This position was adopted and developed by an early entry into this field, David Wellman's 'Portraits of White Racism' (1993). Roediger (1991, 1994), Frankenberg (1993), Allen (1994), Jacobson (1998), Lipsitz (1998), and Perry (1998) have each examined the ways in which white privilege is embedded in a faceless and taken-for-granted normalness of white identity. Twine and Warren (2000) have assembled an anthology in which they bring to greater consciousness the ways in which researchers in racially stratified societies ignore at their peril the subject world of race as an integral feature of methodology, largely through the racial identity of subjects and researchers. However, as yet, the field of white racial identity lacks a coherent theoretical focus and a body of empirical research. This is something we look forward to in the near future.

The Future of Racial Identity: How Does 'Mixed' Race Figure In?

A recent development in this field of study concerns mixed race. Heretofore there was no such thing in the United States, as people were labeled according to a strict protocol. This meant that people were of one race or another; there was no such thing as a mixed race person. The infamous one-drop rule is an example of this. Pele with a drop of black blood were considered to be black. However, recently, a number of scholars of race have noted the increasing use of the category of 'mixed race' as the self-identity of a younger generation, mainly of relatively privileged and middle-class persons. Maria Root (1996) has culled a collection of essays that deal with identity construction along ethnic and racial lines and further explores these issues well beyond the black and white focus that has riveted much of this literature. The difference between 'individual choice' of a racial identity as opposed to having a racial identity 'imposed' by the state and/or the obdurate empirical reality of a group consensus (based primarily upon phenotype-stereotype) is profound, yet variable over time and place. This focus on mixed race is currently in vogue and is an example of the fact that there are concrete economic, political, social, and sometimes psychological benefits to choosing certain identities. That is why scores of thousands of light-skinned 'blacks' under slavery in Brazil, Jamaica, and the United States chose to 'pass' as white (Small, 1994). For parallel reasons, thousands of persons in the United States have attempted to reclassify themselves as Native American or Latino or African American to benefit from affirmative action laws. It remains to be seen if this interest in mixed race will continue into the future.

See also: Civil Rights Movements; Integration, Social; Racial Relations; Social Identity in Social Psychology.

Bibliography

Allen, R.L., Dawson, M.C., Brown, R.E., 1989. A schema-based approach to modeling an African-American racial belief system. The American Political Science Review 83 (2), 421–441.

Allen, R.L., Thornton, M.C., Watkins, S.C., 1992. An African American racial belief system and social structural relationships: a test of invariance. National Journal of Sociology 6 (2), 157–186.

Allen, T., 1994. The Invention of the White Race. Verso, New York.

Allport, G., 1954. The Nature of Prejudice. Beacon Press, Boston.

Baldwin, J.A., 1985. African (Black) Personality: From an Afrocentric Framework. Third World Press, Chicago.

Baldwin, J.A., Bell, Y.R., 1985. The African self-consciousness scale: an Africentric personality questionnaire. Western Journal of Black Studies 9, 61–68.

Baldwin, J.A., Duncan, J.A., Bell, Y.R., 1987. Assessment of African self-consciousness among black students from two college environments. Journal of Black Sociology 9 (2), 27–41.

Banton, M., 1977. The Idea of Race. Tavistock, London.

Belgrave, F.Z., Brome, D.R., Hampton, C., 2000. The contribution of Africentric values and racial identity to the prediction of drug knowledge, attitudes, and use among African American youth. Journal of Black Psychology 26 (4), 386–401.

Blumer, H., 1958. Race prejudice as a sense of group position. Pacific Sociological Review 1 (Spring), 3–6.

Branscombe, N.R., Schmitt, M.T., Harvey, R.D., 1999. Perceiving pervasive discrimination among African Americans: implications for group identification and well-being. Journal of Personality and Social Psychology 77, 135–149.

Broman, C.L., 1996. The health consequences of racial discrimination: a study of African Americans. Ethnicity and Disease 6, 148–153.

Broman, C.L., Neighbors, H.W., Jackson, J.S., 1988. Racial group identification among black adults. Social Forces 67, 146–158.

Broman, C.L., Jackson, J.S., Neighbors, H.W., 1989. Sociocultural and racial group identification among black adults. Revue Internationale de Psychologie Sociale 2, 367–378.

Broman, C.L., Mavaddat, R., Hsu, S., 2000. The experience and consequences of racial discrimination. Journal of Black Psychology 26, 165–180.

Broman, C.L. (Ed.), 2014. Understanding Race and Ethnicity in Contemporary Society. Cognella Academic Publishing, California.

Clark, K.B., Clark, M.P., 1947. Racial identification and preference in negro children. In: Newcomb, T.M., Hartley, E. (Eds.), Readings in Social Psychology. Henry Holt, New York, pp. 602–611.

Corcos, A.F., 2014. Race is a slippery word. In: Broman, C.L. (Ed.), Understanding Race and Ethnicity in Contemporary Society. Cognella Academic Publishing, California, pp. 5–8.

Chou, R.S., Feagin, J.R., 2014. Everyday racism: anti-Asian discrimination in public places. In: Broman, C.L. (Ed.), Understanding Race and Ethnicity in Contemporary Society. Cognella Academic Publishing, California, pp. 5–8.

Crenshaw, K., Gotanda, N., Peller, G., Thomas, K., 1995. Critical Race Theory. The Free Press, New York.

Cross, W., 1971. The Negro-to-Black conversion experience. Black World 13–27.

Cross, W., 1991. Shades of Black: Diversity in African-American Identity. Temple University Press, Philadelphia, PA.

Cross, W.E., Parham, T.A., Helms, J.E., 1998. Nigrescence revisited: theory and research. In: Jones, R.L. (Ed.), African American Identity Development: Theory, Research, and Intervention. Cobb & Henry, Hampton, VA.

Demo, D.H., Hughes, M., 1990. Socialization and racial identity among black Americans. Social Psychology Quarterly 53 (4), 364–374.

Deutsch, M., Collins, M., 1951. Interracial Housing. University of Minnesota Press, Minneapolis.

Ellison, C.G., Powers, D.A., 1994. The contact hypothesis and racial attitudes among black Americans. Social Science Quarterly 75, 385–400.

Essed, P., 1991. Understanding Everyday Racism: An Interdisciplinary Theory. Sage Publications, Newbury Park, CA.

Feagin, J.R., 1991. The continuing significance of race: anti-black discrimination in public places. American Sociological Review 56, 101–116.

Frankenberg, R., 1993. White Women, Race Matters: The Social Construction of Whiteness. University of Minnesota Press, Minneapolis, MN.

Fredrickson, G.M., 1995. Black Liberation: A Comparative History of Black Ideologies in the United States and South Africa. Oxford University Press, New York.

Gurin, P., Miller, A.H., Gurin, G., 1980. Stratum identification and consciousness. Social Psychology Quarterly 43, 30–47.

Hall, W.S., Freedle, R., Cross, W.E., 1972. Stages in the Development of Black Identity. ACT Research Report No. 50. The American Testing Program, Research and Development Division, Iowa City.

Horowitz, R., Murphy, L.B., 1938. Projective methods in the psychological study of children. Journal of Experimental Education 7 (1), 133–140.

Jackson, J.S., McCullough, W.R., Gurin, G., 1988. Family, socialization environment, and identity development in black Americans. In: Black Families, second ed. Sage Publications, Newbury Park, CA, pp. 242–256.

Jacobson, M.F., 1998. Whiteness of a Different Color: European Immigrants and the Alchemy of Race. Harvard University Press, Cambridge, MA.

Jordan, W.D., 1968. White over Black, White over Black: American Attitudes toward the Negro, 1550–1912. University of North Carolina Press, Chapel Hill, NC.

Katz, S.H., February 1995. Is Race a Legitimate Concept for Science? The AAPA Revised Statement on Race: A Brief Analysis and Commentary. University of Pennsylvania.

Kelley, R.D.G., 1997. Identity politics and class struggle. New Politics VI (2), 84–97.

Kessler, R.C., Mickelson, K.D., Williams, D.R., 1999. The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. Journal of Health and Social Behavior 40, 208–230.

Landrine, H., Klonoff, E.A., 1996a. The schedule of racist events: a measure of racial discrimination and a study of its negative physical and mental consequences. Journal of Black Psychology 22, 144–168.

Landrine, H., Klonoff, E.A., 1996b. African American Acculturation. Sage, Thousand Oaks, CA.

Lipsitz, G., 1998. The Possessive Investment in Whiteness: How White People Profit from Identity Politics. Temple University Press, Philadelphia, PA.

Miles, R., 1989. Racism. Routledge, London.

Milliones, J., 1980. Construction of a black consciousness measure: psychotherapeutic implications. Psychotherapy: Theory, Research, & Practice 17 (2), 175–182.

Myers, M.A., Thompson, V.L.S., 1994. Afrocentricity: an analysis of two culture specific instruments. Western Journal of Black Studies 18 (4), 179–184.

Omi, M., Winant, H., 1986. Racial Formation in the United States: From the 1960s to the 1980s. Routledge & Kegan Paul, New York.

Parham, T.A., Helms, J.E., 1981. Relation of racial identity attitudes to self-actualization and affective states of black students. Journal of Counseling Psychology 32, 431–440.

Perry, P., 1998. Beginning to See the White (Unpublished Ph.D. thesis). University of California, Berkeley.

Phinney, J.S., 1992. The multi-group ethnic identity measure: a new scale for use with diverse groups. Journal of Adolescent Research 7 (2), 156–176.

Phinney, J.S., 1996. Understanding ethnic diversity: the role of ethnic identity. American Behavioral Scientist 40 (2), 143–152.

Roediger, D.R., 1991. The Wages of Whiteness, Race and the Making of the American Working Class. Verso, New York.

Roediger, D.R., 1994. Towards the Abolition of Whiteness: Essays on Race, Politics, and Working Class. Verso, New York.

Root, M.P.P., 1996. The Multiracial Experience: Racial Borders as the New Frontier. Sage, Thousand Oaks, CA.

Rosenberg, Morris, 1979. Conceiving the Self. Basic Books, New York.

Sanders-Thompson, V.L., 1991. African American race consciousness and racial identification. The Western Journal of Black Studies 15 (3), 154–158.

Sanders-Thompson, V.L., 1994. Socialization to race and its relationship to racial identification among African Americans. Journal of Black Psychology 20 (2), 175–188.

Sanders-Thompson, V.L., 1995. Sociocultural influences on African American racial identification. Journal of Applied Social Psychology 25 (16), 1411–1429.

Sellers, R.M., Rowley, S.A., Chavous, T.M., Shelton, J.N., Smith, M.A., 1997. Multidimensional inventory of black identity: a preliminary investigation of reliability and construct validity. Journal of Personality and Social Psychology 73 (4), 805–815.

Simmons, R.G., 1978. Blacks and high self-esteem: a puzzle. Social Psychology Quarterly 41, 54–57.

Small, S., 1994. Racialised Barriers: The Black Experience in England and the United States in the 1980s. Routledge, New York.

Smedley, A., 1993. Race in North America: Origin and Evolution of a Worldview. Westview Press, Boulder, CO.

Smith, L.R., Burlew, A.K., Lundgren, D.C., 1991. Black consciousness, self-esteem, and satisfaction with physical appearance among African American female college students. Journal of Black Studies 22 (2), 269–283.

Swim, J.K., Hyers, L.L., Cohen, L.L., Fitzgerald, D.C., Bylsma, W.H., 2003. African American college students' experiences with everyday racism: characteristics of and responses to these incidents. Journal of Black Psychology 29, 38–67.

Taylor, J., Rogers, J., 1993. Relationship between cultural identity and exchange disposition. Journal of Black Psychology 19 (3), 248–265.

Twine, F.W., Warren, J. (Eds.), 2000. Racing Research, Researching Race. New York University Press, New York.

Utsey, S.O., 1998. Assessing the stressful effects of racism: a review of instrumentation. Journal of Black Psychology 24, 269–288.

Wellman, D.T., 1993. Portraits of White Racism, second ed. Cambridge University Press, New York.

Wilner, D.M., Walkley, R.P., Cook, S.W., 1955. Human Relations in Interracial Housing: A Study of the Contact Hypothesis. University of Minnesota Press, Minneapolis.

Race: Prespectives and Crime

Michael J Leiber, Department of Criminology, University of South Florida, Tampa, FL, USA
Jennifer H Peck, School of Criminology and Criminal Justice, Florida Atlantic University, Boca Raton, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The empirical and theoretical literature surrounding the relationship between race and crime has found that in general, minorities (especially Blacks) are overrepresented as offenders compared to their representation in the general US population. Two potential explanations for this racial disparity are differential offending and selection bias. This article describes various theoretical perspectives of the relationship between race and crime, with a primary focus of why minorities may commit more crime/more serious crime and are overrepresented as offenders through both the juvenile and adult justice system.

Introduction

Race/ethnicity has generated an enormous amount of research and debate in criminology (e.g., Leiber, 2003; Wang and Mears, 2010). One factor that underlies much of this activity is the overrepresentation of minority youth and in particular, Blacks, in both the criminal and juvenile justice systems. This article provides a discussion of explanations for the social control of minorities throughout both justice systems. The next section of the article provides official statistics concerning the relationship between race and crime. The next section discusses the theoretical perspectives for understanding why minorities commit more crime or more serious crime. This explanation is known as the differential offending perspective. The final section focuses on the differential selection or selection bias argument. Theoretical perspectives of race and crime that center on potential racism or bias that results in minority representation throughout juvenile justice and criminal justice systems are discussed to conclude the article.

Official Statistics on Race and Crime

According to information provided from the 2010 US Census, the overall Hispanic population surpassed the Black population as the largest minority group in the United States for the first time since the US Census began its data collection. Hispanics comprised 16% of the US population, while 84% of individuals identified as non-Hispanic, but other racial categories. Specifically, Whites represented 64% of the general US population while 12% of all individuals identified as Black. Asians/Pacific Islanders (5%), American Indians/Alaskan Natives (1%), other racial groups (1%), and individuals of two or more races (1%) were also represented (U.S. Census Bureau, 2000). When the juvenile population is disaggregated from the overall US population, White youth represented 76% of all youth up to 17 years old, while Black youth comprise 16% in 2012. American Indians and Asians represented 2 and 6% of all juveniles in 2012, respectively (Puzzanchera et al., 2013).

For the purpose of this article on race perspectives and crime, it is important to understand the representation of different racial/ethnic groups in the general US population in order to compare them to the proportion of racial/ethnic groups involved in crime. For example, the FBI's Uniform Crime Reports (UCR) is one form of crime measurement that collects data on crimes reported to the police across the United States. **Table 1** presents the UCR's count and percentage of violent crime and property crime arrests in 2012 for Part 1 index crimes by race. Part A includes arrests for individuals 18 years and under, while Part B includes arrests for ages 18 years and older.

As shown in Part A of **Table 1**, data from the 2012 UCR shows that White youth are underrepresented in arrests across both violent and property crime arrests compared to their representation in the general population. In particular, White youth represent 47% of all violent crime arrests, and 62% of all property crime arrests. Black youth, however, are overrepresented in arrests across both offense types. Black youth represent 51% of all violent crime arrests and 36% of property arrests. For the most part, American Indian/Alaskan Native youth are arrested at a similar proportion to their representation in the general youth population, while Asian/Pacific Islander youth are underrepresented in UCR violent crime and property crime arrests.

To some degree, Part B of **Table 1** presents similar results as Part A for individuals who are arrested over the age of 18 years. Across both violent and property arrests, White adults are arrested similar to their representation in the general population (60–69%). Black adults, however, are consistently

Table 1 Arrests in 2012 for part 1 index crimes by race

	White	*Black*	*AI or AN*[a]	*Asian or PI*[a]
Part A: ages 18 years and under				
Crime type				
Violent crime	21 682 (47)[b]	23 982 (51)	393 (1)	487 (1)
Property crime	139 749 (62)	80 467 (36)	3002 (1)	3755 (1)
Part B: ages 18 years and over				
Crime type				
Violent crime	214 712 (60)[b]	131 106 (37)	4805 (1)	5303 (2)
Property crime	727 053 (69)	293 496 (28)	14 802 (2)	12 991 (1)

[a]AI: American Indian; AN: Alaskan Native; PI: Pacific Islander.
[b]Numbers in parentheses are percentages.
Data source: Federal Bureau of Investigation, 2012. Crime in the United States.

Table 2 Cases handled by the juvenile court in 2010 by race

	White	*Black*	*Native American*	*Other*[a]
Referred[b]	876 411 (64)[c]	451 063 (33)	21 142 (2)	19 536 (1)
Detained	163 766 (57)	113 925 (40)	5054 (2)	4162 (1)
Formally charged	441 890 (60)	267 543 (37)	12 655 (2)	11 084 (1)
Adjudicated delinquent	266 182 (62)	146 839 (35)	8752 (2)	6383 (1)
Probation	166 921 (64)	83 642 (32)	5464 (2)	4322 (2)
Placed out of home	65 196 (58)	43 623 (39)	2349 (2)	1391 (1)
Waived to adult Court	3122 (52)	2654 (45)	140 (2)	48 (1)

[a]Other includes Asian, Hawaiian, and Pacific Islander.
[b]Referral offense includes person, property, drugs, and public order.
[c]Numbers in parentheses are percentages.
Data source: National Center for Juvenile Justice, 2013. National Juvenile Court Data Archive: Juvenile Court Case Records 1985–2010 (machine-readable data files). NCJJ (producer), Pittsburgh, PA.

overrepresented in across both arrest categories. In fact, the proportion of all arrests for Blacks is at least two times larger for property crime, and three times larger for violent crime than in the general population (12%). American Indian/Alaskan Native adults are slightly underrepresented in violent and property crime arrests, while Asian/Pacific Islander adults are arrested at comparable proportions to the general population across crime types.

According to FBI arrest statistics, over 1.6 million youth under the age of 18 years were arrested for either violent, property, or other (e.g., vandalism, disorderly conduct) types of crimes in 2010 (Puzzanchera et al., 2012). Out of the 1.6 million youth who were arrested, over 1.3 million of those youth were subsequently referred to the juvenile court (Puzzanchera et al., 2012). **Table 2** presents the racial characteristics of cases handled by the juvenile court in 2010.

Of person, property, drugs, and public order cases referred to the juvenile court, White youth represent 64% of all referrals, while Black youth are overrepresented with 33% of all referrals. Out of all of the cases that were detained, 40% were Black youth, and 57% were Whites. Black youth also represented 37% of all cases that were formally charged, compared to 60% of Whites. Throughout these three stages, Whites are consistently underrepresented within the juvenile justice system, while Blacks are overrepresented. Native American and other youth are either equally distributed or underrepresented compared to their representation in the general population. With regard to the adjudication stage, White youth are slightly more represented at adjudication compared to earlier proceedings (i.e., detention and formal charging), as they are the largest racial group to be adjudicated delinquent (62%). Recall, although, that White youth comprise 76% of the overall juvenile population, therefore they are underrepresented throughout these stages. Once again, Black youth are overrepresented at the adjudicatory stage, as they make up 35% of all youth who are adjudicated.

In addition, data are provided on a yearly basis from the Office of Juvenile Justice and Delinquency Prevention (OJJDP) for three potential outcomes at judicial disposition: probation (community supervision), out of home placement (residential placement), and other (not included). Across racial groups, Black youth more often received the most severe outcome of residential placement (39%) compared to probation (32%). Both Native American and Other youth were either under or similarly represented at both dispositional outcomes, and were equally likely to receive either community sanctions or residential placement.

It is also important to note that there are a large number of youth each year who are removed from the juvenile court and transferred/waived to the adult court. While the percentage of all cases waived is relatively small (0.4%) approximately 5900 juvenile cases were waived to adult court in 2010. Of those waived cases, disparities across racial groups were once again found, as minorities in general represented 48% percent of all waived cases, while Blacks alone represented 45% of all youth waived. Overall, there is considerable evidence that race disparities occur throughout all stages of juvenile justice proceedings, as Black youth are consistently overrepresented throughout the juvenile justice system across all other racial groups.

While **Table 2** described the representation across racial groups through various stages of juvenile court proceedings, **Table 3** presents counts and percentages of federal defendants handled by the US District Court in 2010 by race and ethnicity. However, a limitation of **Table 3** compared to earlier tables is that all racial groups exclude individuals who are also of Hispanic or Latino origin (because they constitute their own

Table 3 Federal defendants handled by the US District Court in 2010 by race

	White[a]	*Black*[a]	*Hispanic*	*AI or AN*[a,b]	*Asian or NH or OPI*[a,b]	*Other*[a]
Charged	17 021 (21)[c]	14 201 (18)	45 333 (57)	1217 (2)	1138 (1)	255 (1)
Sentenced in custody	51 091 (27)	68 180 (37)	60 268 (32)	3331 (2)	2280 (2)	
Released in 2008	14 227 (24)	14 004 (24)	29 390 (50)	817 (1)	910 (1)	
Returned in 2010	1843 (22)	2608 (31)	3644 (43)	316 (3)	58 (1)	
Supervised release	35 643 (35)	38 698 (38)	22 405 (22)	2037 (2)	3056 (3)	
Probation	11 116 (49)	5898 (26)	4083 (18)	681 (3)	907 (4)	
Parole	548 (27)	1238 (61)	183 (9)	41 (2)	20 (1)	

[a]Excludes individuals of Hispanic or Latino origin.
[b]AI: American Indian; AN: Alaskan Native; NH: Native Hawaiian; OPI: Other Pacific Islander.
[c]Numbers in parentheses are percentages.
Data source: Bureau of Justice Statistics, based on data from Administrator Office of the US Courts, Probation, and Pretrial Automated Case Tracing System (PACTS), fiscal year 2010.

category in Table 3). Therefore, the counts and percentages included in Table 3 may be underestimated compared to earlier tables.

Across the six racial/ethnic categories, Hispanics were the largest group that was charged by the US District Court in 2010. Blacks and Hispanics are overrepresented as federally sentenced offenders in the custody of Federal Bureau of Prisons, comprising 37 and 32%, respectively. The next two categories focus on the count and percentage of offenders released from federal prison, then subsequently return to prison within their 3 years of release. Therefore, while Whites comprise 24% of offenders released in 2008, they represent 22% of offenders returning into federal custody. The proportion of Blacks released in 2008 (24%) was smaller than those who returned to custody in 2010 (31%). Hispanics comprised 50% of all offenders released in 2008 from federal prisons, and represented 43% of those offenders who returned 3 years later. The other three racial/ethnic categories are either similarly or underrepresented as offenders within each category discussed until this point.

Table 3 also includes the racial/ethnic characteristics of offenders within the US District Court under post-conviction federal supervision. Of the total number of offenders who were granted supervised release, two potential options are available: probation and parole. Whites represent 35% of all offenders under supervised release, while Blacks comprise 38%, and Hispanics, 22% of this category. When supervised release is disaggregated between probation and parole, racial/ethnic disparities are even more evident, as Whites and Hispanics are more likely to receive the less severe outcome of probation instead of parole, while Black offenders are almost 2½ times more likely to receive parole as a supervised release option than probation.

Theoretical Perspectives of Differential Offending

Most criminological theories have generally failed to pay specific attention to racial and ethnic differences in the extent and severity of offending. For some, criminological theories are believed to universally apply to all racial and ethnic groups (Hirschi, 1969). More specific, race/ethnic differences in delinquency and criminal offending are seen as the result of individual differences in offenders, family processes, or structural and community explanations (Gottfredson and Hirschi, 1990; Sampson and Wilson, 1995). Still, there are a few race-specific theories and perspectives for understanding why minorities are overrepresented as offenders. The following section will describe how some general and race-specific theories have explained racial differences in offending.

Social and Self-Control Theories

Hirschi's (1969) version of social control theory contends that individuals conform because they have strong affective attachments to parents, stakes in conformity, involvement in conventional activities, and belief in social norms. Conversely, those youth who have weak attachments, low stakes in conformity, little involvement in conventional activities, and/or poor attitudes regarding societal norms are more likely to participate in delinquent behavior. Hirschi (1969) originally argued that the attachment between parent and child is paramount and that the strength of this relationship is the most important factor in deterring delinquent behavior. In other words, it is the quality, and not the quantity, of bonds that determines delinquency. Implicit within Hirschi's (1969) version of social control theory is the applicability of the perspective to all youth, irrespective of race or ethnicity. In fact, Hirschi (1969) stated that he found nothing to contradict the "assumption that the causes of delinquency are the same among Negroes as among whites," and for this reason he stated that "we need not study Negro boys to determine the causes of delinquency" (p. 80).

Gottfredson and Hirschi's (1990) version of the general theory of crime contends that differences in offending between minorities (more specifically Blacks) and Whites are due to Blacks having lower levels of self-control. The lower level of self-control is believed to be by-product of faulty and erratic socialization and supervision provided by parents. Minority families seen as less able to monitor, recognize, and punish deviant behavior in children, and are ineffective at childrearing and in turn minorities commit more frequent and serious crime. Gottfredson and Hirschi (1990) further argue that minority families in general are more likely to include parental criminality, larger family sizes, more single-parent families, and mothers working outside of the home, which are additional examples of why minority youth have lower levels of self-control and an increased probability of involvement in crime.

General Strain Theory

Strain theory and in particular, general strain theory (GST) is another individual-level perspective that has been applied to examine race and ethnic differences in criminal offending. Some have argued that GST can explain racial differences in offending by implying that Blacks and minorities in general experience a disproportionate and qualitative different types of strain that lead to increased levels of involvement in crime. It is argued that Blacks experience different types of strains more than other groups including economic strain (poverty, unemployment), family strain (poor parenting practices), educational strain (problems with students and teachers, poor grades), criminal victimization, racial discrimination, and community strain (living in structurally disadvantaged neighborhoods). Prior research has supported this argument where Black youth who experience welfare and not being happy in their neighborhood are considered types of strains that contribute to delinquent behavior compared to Whites and Hispanics (Peck, 2013).

Social Disorganization Theories

Social disorganization theory is a perspective that emphasizes structural and cultural factors to understand why minorities offend more than Whites. Shaw and McKay (1942) studied deteriorated neighborhoods to better delineate how neighborhood conditions in urban areas created a crime-prone environment. Communities with low socioeconomic status, ethnic heterogeneity (many varied ethnic groups living

among each other), and residential mobility (where people are constantly moving in and out of the neighborhood) are believed to be indicators of social disorganization, which in turn leads to a subculture or the adoption of values that are contrary to those of larger society – gangs and crime. Community characteristics such as low socioeconomic status and residential mobility lead to weak social ties among residents of the community and its institutions such as school or church. Since these communities are unable to provide basic services such as education, health care, and proper housing, there is a sense of disorganization and a lack of social ties and community. Therefore, minorities offend more because they are likely to reside in such communities. Underlying social disorganization theory is a strong economic component.

Sampson and Wilson (1995) offer a revised interpretation of social disorganization theory with a focus on race. More specific, they integrate social disorganization with Wilson's (1987) perspective of economic inequality and racial discrimination to understand minority poverty and involvement in crime. Living in impoverished and segregated communities that evidence unemployment, lack of quality jobs, and family disruption, minorities face structural barriers, isolation, and adapt subcultural values as a means of coping that include a devaluation of educational aspirations, drug use, and crime.

Subcultural Theories

Subcultural theories are considered race-specific perspectives because some perspectives consist of race-related themes that attempt to explain why minorities differentially offend compared to Whites. The most popular subcultural theory is Wolfgang and Ferracuti's (1967) subculture of violence theory that was based on the relationship between race and homicides, specifically the influence of victim-precipitated homicides. It was argued that there was a subculture of violence which does not consider assaults as wrong or antisocial, where physical aggression is socially approved, and where violence is considered normal but often ends in death. The theory articulated seven propositions that encompass the subculture of violence. Some of these propositions include that the subculture is found within all ages of society; however, violence is most prominent in late adolescent and middle age. Violence is not seen as an illicit occurrence, and there are no feelings of guilt. Those in the subculture are not violent all of the time. Throughout the 1970s and 1980s support was mixed, with some studies finding a subculture of violence among Blacks while others failed to find support who argued that poverty was the most consistent predictor of homicide, not race.

Anderson's (1999) 'Code of the Street' is another perspective that has racial and cultural overtones as well as an emphasis on the structural deficiencies of communities. Residing in disadvantaged communities and facing racial discrimination youth and in particular, minority youth adopt a 'street code' to survive. Research suggests that this perspective is applicable to not only Blacks but also Hispanics. The theory has also been applied to victimization in that residents who adopt a 'street code' are at a heightened risk for victimization, above and beyond living in disorganized and violent community.

Life-Course Perspectives

Recently, a few studies have been conducted that examine race/ethnicity and crime from a life-course perspective. Although there are some differences between developmental or life-course theories, these perspectives embody the principle that there is continuity, persistence, and possible desistance in offending (Moffitt, 1993; Sampson and Laub, 2005). In addition, relative offending rates in one age-period link to relative offending rates in others. Continuity in offending is also the hallmark of the most high-rate persistent offenders that are thought to begin their offending careers earlier, have longer careers, and commit more crimes over comparable periods than other types of offenders. While some offenders persist well past the aggregate peak age in offending, most desist in a normative manner in late adolescence/early adulthood (e.g., Moffitt, 1993; Sampson and Laub, 2005). Although distinct, persistence, and desistance are related through a dynamic process involving the consequences associated with the interrelationships between choice, life circumstances, turning points, and criminality (Sampson and Laub, 2005). Ge et al. (2001), for example, examined racial/ethnic differences in offending over time by using data from the California Youth Authority and discovered that up to age 21 years, there were no statistically significant differences in arrest frequencies between Whites and Blacks or Whites compared to Asians. After age 21 years, Blacks, however, were arrested more frequently. Hispanics were also arrested more frequently than Whites but only after age 25 years.

Moffitt's (1993) developmental theory or taxonomy of antisocial behavior argues that Blacks are more likely to be overrepresented as both life-course persistent (LCP) offenders (offend throughout the life course) as well as adolescent-limited (AL) offenders (offend only during adolescent years). Moffitt's justification for this argument is that Black offenders are likely to live in structurally disadvantaged communities that hinder their life chances at success. Thus, Black LCP offenders have experienced inadequate prenatal care, exposure to hazardous toxins, lack of familial attachment and bonds, and disadvantaged school environments. These developmental situations result in poor educational success and limited employment opportunities. Accordingly, Blacks will be more likely to continue offending as adults because they as a group experience higher levels of risk factors (unemployed, single, drinking, and drug use) than do Whites.

In terms of Blacks as AL offenders, Moffitt argues that this type of offender is stuck within a maturity gap that White youth do not experience for as long of duration of time. More specific, Black youth spend more years within a maturity gap because they do not necessarily shift to an 'adult status' with valued adult roles (e.g., employment) and privileges. This delayed entrance into adulthood and conventional adult roles results in Blacks becoming overrepresented as AL offenders, who mimic offending patterns of delinquent peers. These types of offenses symbolize adult social status, such as theft, drinking, and drug use.

Empirical research has found support for Moffitt's explanation of racial differences in offending. For example, results have found that Blacks compared to Whites were more likely to persist in violent offending due to reduced opportunities in economic well-being and employment when in adolescence, supporting an economic maturity gap perspective. Research has also found that for Blacks, certain LCP characteristics of offenders (developmental risk and adverse family situations) were intensified by community disadvantage. This result was not applicable to Whites (Haynie et al., 2008).

Theoretical Perspectives of Selection Bias

Prior research has found that the differential offending explanation alone has not been able to fully account for the relationship between race and crime, specifically how minorities are overrepresented in the juvenile and criminal justice systems (Leiber and Rodriguez, 2011). In particular, the presence of selection bias has evolved into an additional reason why minorities are treated differently compared to Whites. Both overt and covert racial bias and stereotypes have found to account for minorities receiving disadvantaged outcomes compared to Whites (Bridges and Steen, 1998). From these findings, theoretical perspectives of selection bias surrounding the relationship between race and crime have focused on both structural and individual-level factors that influence decision-making of both Black and White offenders.

There are a number of macro-level perspectives of the relationship between race and crime that focus on how community characteristics may influence social control to the disadvantage of minority groups. Contextual indicators such as urbanism, minority representation, structural disadvantage, and racial inequality are some of the factors noted in the literature as being important measures concerning race perspectives and crime. In the following paragraphs, the macro-level perspectives of urbanism and formal rationality (Weber, 1969), minority/racial group power threat thesis (Blalock, 1967), and the theory of inequality and social control (Sampson and Laub, 1993) are discussed in detail.

Urbanism and Formal Rationality

One macro-level perspective that focuses on the influence of community context on crime and court outcomes is a perspective that is characterized by formal and substantive rationality (Weber, 1969). Weber (1969) argued that criminal justice decision-making is based on the interrelationships urbanization and bureaucratization. Specifically, he believed that the economy and legal system are two institutions that guide all individuals within society. The link between both systems is formal rationality, which translates to universally applied laws, rules, and standards in decision-making. Weber (1969) proposed that courts that are located in urban communities make court decisions within the formal rational legal system, which are grounded in the rules themselves. The law is applied equally across extralegal factors (e.g., race, gender, and age) based on legal criteria (e.g., crime severity and prior record) in the formal rational system. Therefore, the amount of discretion given to decision-makers is limited based on the notion that court outcomes depend on legal factors. Few race differences in court outcomes between Black and White offenders should exist concerning their contact and involvement in the criminal justice system.

Compared to formal rationality, substantive or 'informal' rationality is more likely to occur in rural settings, where decision-making is more often based on external values of beliefs, respect, and informal values. Within this type of legal system, discretion is enhanced among decision-makers, which can result in greater disparities among racial groups with regard to involvement in the criminal justice system. Weber (1969) argued that in more rural courts, issues such as racism, stereotypes, and other discretionary factors influence decision-making and result in more disadvantage outcomes for minorities compared to Whites.

However, it has also been argued that urban courts will subject minorities to greater social control than rural courts, based on the argument that the formal rational legal system may not question criminal justice procedures but accept them as binding and rely on factors that may be biased against minorities (e.g., prior record and family status). Since urban courts have a greater number of cases compared to rural courts, decision-makers may classify offenders into 'routine types' that may be based on racial, gender, and class stereotypes (Gaarder et al., 2004). Therefore, while Weber (1969) proposed that race differences will be less likely to occur in urban courts compared to rural jurisdictions, research has also argued that the influence of racial stereotypes may be more prominent in urban settings and result in more severe court outcomes for minorities (especially Blacks) compared to Whites.

Minority/Racial Group Power Threat Thesis

In contrast to Weber (1969) who focused on differences across urban and rural courts based on the degree of bureaucracy in the legal system, Blalock (1967) proposed a macro-level theoretical perspective of race and crime that focuses on how 'racial threat,' 'minority group threat,' and general 'group threat' result to race differences in social control. Two interrelated factors, the size of the minority (e.g., Black) population, and the economic and political situation of the minority group compared to the majority group (e.g., Whites) are argued by Blalock (1967) to result in increases in punitive court outcomes for minorities. More specific, as the proportion of minority residents increase within a community, this situation translates into an increase in competition for economic resources (e.g., jobs, property, and money) and political power of the minority group compared to the majority group. In turn, Whites feel that there is a perceived challenge to their status quo compared to Blacks.

In other words, increases in economic wealth, prestige, and prosperity by Blacks will make Whites (the majority group) feel more threatened. As a result, prejudicial and discriminatory attitudes and practices will be enforced across racial lines to decrease the amount of power by the minority group and perceived notions of threat by the majority group. Overall, Blalock (1967) has found some degree of support depending on the outcome examined. For example, increases in minority populations and political threat resulted in more punitive sanctioning for minorities in the form of prison and jail

sentences (Wang and Mears, 2010) and additional studies have found that minority economic threat resulted in larger police force sizes.

Theory of Inequality and Social Control

One macro-level perspective of race and crime integrated aspects of structural disadvantage, racial stereotyping based on Tittle and Curran's (1988) symbolic threat hypothesis, and the influence of the War on Drugs to explain why minorities receive disadvantaged court outcomes compared to Whites. Specifically, Sampson and Laub's (1993) macro-level of inequality and social control argued that structural indicators of underclass poverty and racial inequality influence juvenile justice decision-making. It was proposed that counties characterized by underclass poverty and racial inequality were more likely to subject youth, especially Black youth, to harsh juvenile court outcomes compared to youth who reside in more affluent counties.

Paralleling Tittle and Curran's perspective (1988), Sampson and Laub (1993) argued that decision-makers will enforce greater social control on populations that threaten middle and upper-class standards. Minority populations, the poor, and the unemployed represent these threatening groups to mainstream American. Decision-makers will perceive these groups are more threatening and subject them to harsh court outcomes. In other words, the term 'the underclass' corresponds to negative perceptions of threatening populations. Furthermore, the perspective states that threatening populations can be disaggregated into more specific groups. Based on the historical context of the War on Drugs, the relationship between race and crime was focused on negative perceptions and racial stereotypes of minorities, especially Black drug offenders. The combination of both aspects of the symbolic threat hypothesis (Tittle and Curran, 1988) and the War on Drugs resulted in Sampson and Laub's (1993) perspective to propose that young, Black male drug offenders pose a specific threat to the general population and will result in race differences in social control.

Overall, the perspective argued that youth who are referred to the juvenile court are more likely to receive severe court outcomes when they reside in counties characterized by underclass poverty and racial inequality. The effect was stated to be larger for Blacks compared to Whites, and drug offenders compared to other types of offenders. For the most part, support for the macro-level theory of inequality and social control is mixed. Research testing Sampson and Laub's (1993) theory has found that race differences in court outcomes emerged regardless of the level of community disadvantage (Leiber, 2003), and both harsh and lenient outcomes were found for minorities compared to Whites once community characteristics were taken into consideration.

While there are a number of macro-level theories that focus on the relationship between race and crime within the selection bias argument, there are also various micro-level perspectives that emphasize the role that individual-level racial stereotypes play in the treatment of minority youth. These perspectives argue that racial biases and stereotypes influence the decision-making of court officers which results in greater social control for minorities. In particular, the symbolic threat hypothesis (Tittle and Curran, 1988), liberation hypothesis (Kalven and Zeisel, 1966), and attribution theory (Albonetti, 1991) are discussed as a few examples of micro-level theories.

Symbolic Threat Hypothesis

As introduced earlier, the symbolic threat hypothesis (Tittle and Curran, 1988) was integrated into Sampson and Laub's (1993) macro-level theory of inequality and social control. One justification for this integration is based on the notion that the symbolic threat hypothesis has a slight focus on how community characteristics influence the perceptions of court actors. The second justification is that the hypothesis primarily emphasizes certain micro-level contingencies of decision-making based on the characteristics of youth (specifically, minority youth) and the social psychological emotions of juvenile justice decision-makers.

The perspective argues that juvenile court actors may perceive minority youth as aggressive, sexual, not taking responsibility for their offenses, lacking discipline, and other negative qualities. These qualities translate into social psychological emotions of fear, resentment, uneasiness, or jealousy against minorities who are referred to the juvenile court. Tittle and Curran (1988) argue that court actors are unable to identify with youth or their offending behavior. In turn, the negative emotions and lack of identification translate to disadvantaged court outcomes of minority youth because these stereotypes and negative emotions influence the decision-making of court officers. However, Tittle and Curran (1988) argued that court decisions are based on the negative perceived thoughts of minority youth, not the actual offending behavior that resulted in minority youth being brought to the juvenile court in the first place. The threat that minorities pose to juvenile court actors is more symbolic than real. The symbolic threats are connected to structural characteristics (e.g., size of the minority population and size of the youth population), and negative emotions of court officers to result in greater social control of Black youth compared to Whites.

Overall, Tittle and Curran (1988) found support for the symbolic threat hypothesis in that counties that were characterized by a large percentage of minority residents and large percentage of youth resulted in harsh court outcomes by court actors against minority youth who were charged with drug and/or sex offenses. It may be that decision-makers viewed minority youth who engaged in these specific offenses as more threatening than youth who were referred to the court for other types of illegal behaviors. Additional support has also been found in the literature that has studied the relationship between race and crime within the juvenile justice system.

Liberation Hypothesis

One micro-level perspective that identifies certain case conditions that influence the decisions of court actors can be tied to racial stereotyping is the liberation hypothesis (Kalven and Zeisel, 1966). Kalven and Zeisel (1966) originally formulated the liberation hypothesis from examining the conditions that affected jury decision-making in sexual assault cases. The hypothesis identifies certain case conditions that result in jurors' ability to 'liberate' back and forth between legal and extralegal influences when determining case outcomes.

The perspective stated that in cases that involve serious crimes (e.g., homicide, robbery, and rape) with strong evidence, discretion is limited and jurors rely on case characteristics and legal criteria. Court actors are more likely to rely on legal factors such as crime severity and prior record when determining case outcomes (Spohn and Cederblom, 1991), and less on extralegal factors such as victim and offender race, gender, and age. Stated differently, decision-makers are less likely to 'liberate' from the legal factors of the case and not base decisions on their own perceptions, beliefs, and even stereotypes. The social control of offenders for serious crimes is more likely to occur based on legal criteria, and less on personal feelings.

With regard to less serious crimes with weak evidence, decision-makers may be more apt to utilize large amounts of discretion, and let personal typescripts and racial stereotypes influence outcomes. In turn, less serious offenses with limited amounts of evidence can result in jurors 'liberating' from the case characteristics and introduce potential extralegal factors into the decision-making process (Horney and Spohn, 1996). Concerning the relationship between race and social control, the liberation hypothesis predicts that minorities will be subjected to increased social control compared to Whites only in less serious cases. This is because discretion is enhanced by decision-makers and race as an extralegal factor within less serious cases will subsequently influence case outcomes to the disadvantage of minority offenders (Spohn and Cederblom, 1991).

Focal Concerns Perspective

The 'focal concerns' perspective is a theoretical model that corresponds to the relationship between race and crime based on the research by Steffensmeier et al. (1998). The perspective specifically focuses on the amount of discretion awarded to court actors and corresponding social control within adult court processing decisions, yet has also more recently been applied within the juvenile justice system. The underlying premise of the focal concerns model is that decision-makers (i.e., judges) have a limited amount of time and information about criminal defendants, and may rely on three focal concerns or attributions involving race, gender, and class stereotypes. The three focal concerns include: (1) the defendants' culpability and blameworthiness based on the seriousness of the offense; (2) society's concern to protect society and the community from the offender; and (3) organizational considerations involving available correctional resources. Even though the three focal concerns are interrelated, Steffensmeier et al. (1998) argue that judges do not have complete information about each type of concern. In turn, judges develop a 'perceptual shorthand,' where they rely on legal factors (e.g., prior record, crime severity) *and* racial stereotypes to determine the outcomes of defendants. In particular, Steffensmeier et al. (1998) found support for their perspective in that young, Black males received more severe sentences compared to any other age, race, or gender category. These results confirmed Steffensmeier et al. (1998) proposed stereotype that young, Black males are often perceived as more dangerous, threatening, and unsuitable for release.

Attribution Theory and Racial Stereotyping

An additional prospective that focuses on the role of racial stereotypes is attribution theory. The perspective proposes that individuals rely upon a set of rules or inferences when explaining individual's behavior (Albonetti, 1991). Albonetti (1991) integrated aspects of the structural organizational approach to decision-making and causal attribution theory to understand judicial outcomes of criminal justice defendants. The structural organizational approach argues that in order for decision-makers to have the ability to make rational decisions, they must be knowledgeable of all relevant information pertaining to the decision and possible outcome alternatives. Decision-makers, however, rarely have complete information to make an informed decision. To reduce the amount of uncertainty and make a more rational decision, actors develop 'patterned responses' that to varying degrees are based on past stereotypes and experiences in determining outcomes (Ulmer and Kramer, 1998).

At the micro-level, Bridges and Steen (1998) proposed that court actors rely on past experiences, cases, offenders, and racial discrimination to influence future case outcomes. Specifically, it was argued that racial bias can be found in different attributions of delinquents, especially the interrelationships between race and class (see also Rodriguez, 2010). Race differences in social control are based on the perceptions that Black youth encompass more negative *internal* attributions (i.e., bad attitude, lack of remorse, lack of cooperation, and failure to take the offense seriously), while White youth may represent more negative *external* attributions (i.e., hostile living environment, poor school performance, and dysfunctional family). Internal attributions of delinquent offenders compared to external attributions are therefore perceived by court officers as being more dangerous, having a higher risk of recidivism, and less amenable to treatment. This would result in more social control and harsher punishments for Black referred to the juvenile court compared to Whites (Rodriguez, 2010). It has also been proposed that certain attributions have been assigned to other minority groups in addition to Blacks. Rodriguez (2010) argued that cultural differences across racial/ethnic groups may result in negative attributions that perceive a Latino youth as prone to violence, more likely to be involved in a gang, and indifferent to juvenile justice intervention (Demuth, 2003).

Conclusion

Minority (especially Black) overrepresentation in arrests and presence in both the juvenile justice and criminal justice systems represent significant issues throughout the field of criminology. While both the differential offending and the race/ethnic selection bias perspectives have been used to explain these occurrences, these explanations have all too often have been pitted against one another. Research has shown that both can explain the disproportionate social control of minorities. Support for each of these positions have led the charge for the need to link factors associated with differential offending (i.e., inequality, subcultural values as adaptations to structural inequities) with the contexts of race and social control in terms of police contact, processing within the court,

and discrimination (Leiber and Rodriguez, 2011). Thus, theory and policy efforts need to focus on the theoretical conditions that contribute to involvement in crime that fuel the reproduction by justice officials of cultural stereotypes of minority offenders, their families, and the neighborhoods in which they reside.

See also: Communities and Crime; Control Theories and Crime; Crime and Class; Crime, Race and Ethnicity; Crime, Sociology of; Criminology: Development and Current Status of the Discipline; Data Bases and Statistical Systems: Crime Measurement; Intersectionality and Crime; Life Course Criminology; Self-Control Theory and Crime; Strain Theories and Crime.

Bibliography

Albonetti, C., 1991. Integration of theories to explain judicial discretion. Social Problems 38, 247–266.

Anderson, E., 1999. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton & Company.

Blalock, H.M., 1967. Toward a Theory of Minority Group Relations. Wiley, New York.

Bridges, G.S., Steen, S., 1998. Racial disparities in official assessments of juvenile offenders: attributional stereotypes as mediating mechanisms. American Sociological Review 63, 554–570.

Demuth, S., 2003. Racial and ethnic differences in pretrial release decisions and outcomes: a comparison of hispanic, black, and white felony arrestees. Criminology 41 (3), 873–908.

Gaarder, E., Rodriguez, N., Zatz, M., 2004. Criers, liars, and manipulators: probation officers' views of girls. Justice Quarterly 28, 547–578.

Ge, X.J., Donnellan, M.B., Wenk, E., 2001. The development of persistent criminal offending in males. Criminal Justice and Behavior 28, 731–755.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Haynie, D.L., Weiss, H.E., Piquero, A., 2008. Race, the economic maturity gap, and criminal offending in young adulthood. Justice Quarterly 25, 595–622.

Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley.

Horney, J., Spohn, C., 1996. The influence of blame and believability factors on the processing of simple versus aggravated rape cases. Criminology 34, 135–162.

Kalven Jr., H., Zeisel, H., 1966. The American Jury. Little, Brown, Boston.

Leiber, M.J., 2003. The Contexts of Juvenile Justice Decision Making: When Race Matters. State University of New York Press, Albany, NY.

Leiber, M.J., Rodriguez, N., 2011. The implementation of the disproportionate minority confinement/contact (DMC) mandate: a failure or success? Race and Justice 1, 103–124.

Moffitt, T.E., 1993. Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. Psychological Review 100, 674–701.

Peck, J.H., 2013. Examining race and ethnicity in the context of general strain theory, depression, and delinquency. Deviant Behavior 34, 706–726.

Puzzanchera, C., Sladky, A., Kang, W., 2012. Easy Access to FBI Arrest Statistics 1994–2009. Online. Available: http://www.ojjdp.gov/ojstatbb/ezaucr/.

Puzzanchera, C., Sladky, A., Kang, W., 2013. Easy Access to Juvenile Populations: 1990–2012. Online. Available: http://www.ojjdp.gov/ojstatbb/ezapop/.

Rodriguez, N., 2010. The cumulative effect of race and ethnicity in juvenile court outcomes and why preadjudication detention matters. Journal of Research in Crime and Delinquency 47, 391–413.

Sampson, R.J., Laub, J.H., 1993. Structural variations in juvenile court processing: inequality, the underclass, and social control. Law & Society Review 27, 285–311.

Sampson, R.J., Wilson, W.J., 1995. Toward a theory of race, crime, and urban inequality. Race, Crime, and Justice: A Reader 177–190.

Sampson, R.J., Laub, J.H., 2005. A life-course view of the development of crime. The Annals of the American Academy of Political and Social Science 602, 12–45.

Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas. Chicago, Illinois.

Spohn, C., Cederblom, J., 1991. Race and disparities in sentencing: a test of the liberation hypothesis. Justice Quarterly 8, 305–327.

Steffensmeier, D., Ulmer, J., Kramer, J., 1998. The interaction of race, gender, and age in criminal sentencing: the punishment cost of being young, black, and male. Criminology 36, 763–797.

Tittle, C.R., Curran, D.A., 1988. Contingencies for dispositional disparities in juvenile justice. Social Forces 67, 23–58.

Ulmer, J.T., Kramer, J., 1998. The use and transformation of formal decision-making criteria: sentencing guidelines, organizational contexts and case processing strategies. Social Problems 45, 248–267.

United Sates Census Bureau, 2000. Census Summary File 1. Retrieved from http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t.

Wang, X., Mears, D.P., 2010. A multilevel test of minority threat effects on sentencing. Journal of Quantitative Criminology 26, 191–215.

Weber, M., 1969. The nature of the city. In: Senne, R. (Ed.), Classic Essays on the Culture of Cities. Meredith Corporation, New York.

Wilson, W.J., 1987. The Truly Disadvantaged. University of Chicago Press, Chicago.

Wolfgang, M.E., Ferracuti, F., 1967. The Subculture of Violence. Towards an Integrated through in Criminology. Tavistock Publications, London.

Racial Relations

Joe R Feagin, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines racial relations in the United States, France, and South Africa. After examining the ideas of race, racial group, and ethnic group, the article explores social science theories of racial and ethnic relations – order-assimilation theories and power-conflict theories. The article concludes with an examination of challenges to white European domination of non-European peoples around the globe.

This article examines the reality and development of racial relations in the United States, France, and South Africa. After an examination of the ideas of race, racial group, and ethnic group, the article explores two broad categories of social science theories of racial and ethnic relations – order-assimilation theories and power-conflict theories. Then the article briefly examines typical racial relations in three countries – the United States, France, and South Africa – and concludes with an examination of coming challenges to white European domination of non-European peoples around the globe.

The term *racial relations* denotes relationships between important societal groups socially defined as races. In common usage, this term *race* is ambiguous and taken by many popular and some scholarly analysts to mean that racial relations are not primarily matters of societal power and oppression, but rather mostly involve relationships between groups with relatively comparable societal resources. In many cases, however, racial relations are much more accurately termed *racist relations*. Thus, in areas of the globe where there has been significant European colonization and imperialism – eventually most areas of the globe – systems of racist oppression were created by European colonists and their descendants, who typically enriched themselves substantially at the expense of indigenous peoples' land, resources, and labor. This often brutally executed white Western enrichment became part of the economic foundation for Europe's and North America's long-term economic prosperity and dominant wealth, as well as for the impoverishment of indigenous peoples around the globe (Du Bois, 1965 [1946]).

The Idea of Race

Before the eighteenth century, the English word *race* was used for descendants of a common ancestor, emphasizing kinship rather than skin color or other physical characteristics. By the eighteenth century, this word had evolved fully into its modern racist sense of a category of human beings with distinctive and determining physical characteristics transmitted by descent. In the 1770s, the preeminent Western philosopher Immanuel Kant was the first major scholar to use the term 'races of mankind' (in German) in the sense of physically and biologically distinct 'race' categories. Soon thereafter, the German anatomist Johann Blumenbach laid out a ladder of races that became influential (Hannaford, 1996, pp. 205–207). At the top were the 'Caucasians,' a racist term Blumenbach coined for Europeans, whom he considered to be the superior racial group. Groups lower on his racial ladder were, in order, the Mongolians (Asians), the Ethiopians (Africans), the Americans (Native Americans), and the Malays (Polynesians).

In the late-eighteenth and early nineteenth centuries, Kant, Blumenbach, Thomas Jefferson, and other leading European and North American intellectuals increasingly proclaimed the superior virtues and privileges of being white and of European origin. Educated Europeans and European Americans crystallized the distinction between the 'superior white race' and 'inferior races' and disseminated these views to the general public in their countries and also globally. This *social* construction of races developed dramatically in impact and assertion as Europeans and European Americans expanded their colonial excursions and exploitation of more groups overseas. European and European American colonizers developed an array of racist stereotypes, explanations, and rationalizations for their destruction of indigenous societies, enslavement of Africans, and other colonial exploitation. This array congealed in a broad *white racial framing*, a pervasive and increasingly dominant white worldview with racialized ideological, narrative, and other elements that emphasized the superior white 'race' and legitimated this extensive oppression and exploitation around the globe (Feagin, 2010b).

Racial and Ethnic Groups

Taking into account this history of asymmetrical racial relations and racial oppression, most social scientists define a *racial group* as a socially constructed group that people in- or outside that group have decided is important to single out as inferior or superior, typically on the basis of physical characteristics subjectively selected. For marking off racial groups, certain physical characteristics like skin color are often chosen, and other characteristics like eye or hair color usually are rejected, solely for social reasons. Similarly, an *ethnic group* is a group that is set apart by insiders or outsiders primarily on the basis of cultural or national-origin characteristics subjectively selected (Feagin and Feagin, 2011).

Racial and ethnic groups are not fixed groups or essential categories that endure forever, but rather are temporary human inventions that are constructed socially and usually are shaped in sociopolitical struggles at particular historical times and geographical places. Most important, they have significance and meaning only in relationship to group interaction.

Although the two types, racial and ethnic groups, can be distinguished on theoretical grounds, in the everyday operation of societies, racial and ethnic markers, identities, and oppressions usually are intertwined.

Conceptual Approaches to Racial Relations

Most conceptual approaches to racial (and ethnic) group relations can be classified roughly as (1) order-assimilation theories or (2) power-conflict theories. Order-assimilation theories tend to emphasize the gradual incorporation (assimilation) of immigrant groups into the dominant culture of a society and to give more attention to stability over time in intergroup relations. In contrast, power-conflict theories give central attention to the outcomes of racial (and ethnic) group contacts (immigrants and others) that result in oppression, labor exploitation, and genocide.

Assimilation Approaches

In Europe and the United States, much theorizing about racial and ethnic group relations has placed an emphasis on assimilation, on the more or less orderly adaptation of a migrating group, such as Italian or Russian immigrants to the United States in the early 1900s, to the ways of an established host group, such as the dominant white-Anglo-Protestant core culture of the United States then (and now). Related theories stress individual or group competition seen as resolving itself in the direction of eventual accommodation and assimilation (e.g., Banton, 1983). In most order-assimilation theorizing, relatively little attention has been paid to outcomes of intergroup contact taking the form of long-term or foundational racial oppression. One early US theorist, the influential University of Chicago sociologist Robert E. Park (1950), argued that intergroup relations generally move through the stages of a *race relations cycle* – contact, competition, accommodation, and eventual assimilation – that are more or less continuous and largely irreversible. Scholars in this assimilationist tradition typically have assumed that there is a long-term trend toward eventual adaptation of immigrant groups in Western societies to the dominant group and culture and that racial-ethnic conflicts will diminish with the eventual assimilation of incoming groups to the dominant core culture.

Giving primary attention to immigrants to the United States, influential sociologist Milton Gordon (1964) viewed the typical trend of such racial-ethnic group adaptation to be in the direction of immigrants giving up much of their home-country heritage for the preexisting and dominant white-Anglo-Protestant core culture. This group adaptation takes place in seven different and defined arenas of assimilation: cultural assimilation, structural (primary group) assimilation, marital assimilation, attitude-receptional assimilation (breakdown of prejudices), behavior-receptional assimilation (decline of discrimination), identificational assimilation, and civic assimilation. Mainstream assimilation theorists like the pioneering Gordon and subsequent researchers utilizing assimilation theory usually have had in mind as their main examples of racial-ethnic adaptation European immigrant groups migrating voluntarily to the United States or to various European nations. Over recent decades, this assimilation framework has persisted, in more nuanced form, as an influential mainstream perspective. Examples of the contemporary use of Gordon's typology can be seen in the research of sociologist Silvia Pedraza (1985) on Latino immigration to the United States and the research of sociologists Richard Alba and Victor Nee (2003) on white ethnic adaptation. Alba (2009) has drawn on the mainstream conceptualization of assimilation in critically assessing the contemporary loss of strong ethnic identities among white Americans. Richard Alba and Victor Nee (2003) have amplified this theoretical perspective by demonstrating reciprocity in the assimilation process – that is, how the cultures of immigrants have had an impact on the 'core' culture of society. Contemporary assimilation theory has become considerably more developed and nuanced than in earlier decades.

Still, mainstream US assimilation theorists have as a rule given relatively less attention to the adaptation of non-European groups to the US core culture than have power-conflict theorists. When they have, they have viewed non-Europeans as likely assimilating more or less like earlier white European immigrants. One prominent European researcher of US racial relations, Gunnar Myrdal (1964 [1944], vol. 2, p. 929), has argued that as a practical matter it is to the advantage of African Americans "as individuals and as a group to become assimilated into American culture, to acquire the traits held in esteem by the dominant white Americans." Even today, many order-assimilation analysts expect the importance of initial racial-ethnic stratification to decline as powerful universalistic forces, such as industrialization and urbanization, wipe out the vestiges of earlier particularistic value and racial-ethnic (including racist) coding systems in societies such as the United States.

Several social scientists (e.g., Mia Tuan, 1998) have used empirical field studies to examine certain problematical aspects of the assimilation perspective in regard to the contemporary experiences of immigrants of color and their immediate descendants. Their findings show that, even as Asian American immigrants and their children substantially acculturate to the Anglo-American core culture, most retain a strong racialized identity – substantially because they regularly suffer discrimination from whites who constantly impose on them the racially stereotyped identity of undesirable 'alien' and 'foreigner.' Contrary to the perspective of Gordon and some recent mainstream assimilation thinkers, most Asian Americans, in spite of significant cultural and socioeconomic assimilation, remain racialized and thus a substantial distance from being fully accepted and assimilated into white middle-class society.

Power-Conflict Approaches

Generally speaking, mainstream assimilation theorists are most concerned with cultural (such as language) or primary-group network (such as intermarriage) features of the adaptation of migrating groups to a host society, whereas power-conflict theorists are most concerned with racial-ethnic oppression, exploitation, and subordination that affect incoming and other groups in society. An early US advocate of this perspective, Oliver C. Cox (1948), emphasized the role of the capitalist class

in creating racial conflict by exploiting the labor of groups like African Americans and propagating racist ideologies to divide the working class in Western societies such as the United States. Providing much evidence, scholars in this power-conflict tradition have shown that early racial (e.g., white-on-black) stratification largely grew out of the global and other territorial expansion of profit-oriented capitalism and has continued to be reinforced and maintained by still-extant capitalism and associated imperialism.

Some scholars have drawn on theoretical discussions of external colonialism and developed the concept of *internal colonialism* to describe what happens to many non-European groups coming into and integrated into Western societies. An emphasis on great power inequalities between whites and people of color is generally at the heart of this model. Thus, people of color, such as Native Americans and African Americans, were brought into the sphere of expanding capitalistic system of North America by force. Subsequent group adaptation was characterized by extensive dominant-group coercion and oppression of these racially subordinated groups, as well as by significant group resistance by the latter. Racial stratification thus has a firm material base, an underpinning rooted in a specific political–economic histories. Internal-colonialism theorists are not concerned primarily with white immigrant groups but rather with the persistence of systemic oppression and the asymmetrical control processes that maintain white dominance over various groups of color in a society like the United States (Blauner, 1972).

Power-conflict analysts look at the immigration of non-European peoples as involving much more than continuous assimilation. John Rex (1973) has analyzed the immigration of such groups from the former European colonies to the metropolitan cities of numerous colonizing countries. In the latter countries, most of these colonial immigrants of color historically have gotten inferior and marginal jobs and often have suffered higher unemployment – thereby typically becoming structurally distinct from the native-born, white working and middle classes and, thus, a low-ranking social group often viewed as alien or foreign.

Generally, power-conflict analysts emphasize the direct and continuing impact of structural racial barriers on people of color in countries around the globe. In the 1980s and 1990s, innovative sociologists Michael Omi and Howard Winant (1994) developed an important theory of 'racial formation.' Racial formation is the "sociohistorical process by which racial categories are created, inhabited, transformed, and destroyed" in societies like the United States (Omi and Winant, 1994, p. 55). Critically reviewing prior theorizing – especially that of assimilation (ethnicity) and Marxist theorists – that avoided or downplayed racial oppression issues in favor of accents on ethnicity or class, Omi and Winant have emphasized the continuing significance of racial formations in the United States over the generations. From their influential perspective, asymmetrical racial relationships are substantially defined by state action, such as that defining who is 'black' and 'white' (the infamous 'one drop of blood' US rule) or that restricting immigration to the United States from Asian and Latin American countries on the basis of racist reasoning. Numerous social science and humanities scholars have made significant use of this racial formation theory in theorizing about issues, such as struggles among whites and Mexican Americans, racism and popular culture, and interpreting the societal construction of race (see Gomez, 2002; Brock, 2009).

Other power-conflict analysts (see Feagin and Elias, 2012) have demonstrated that the racial formation approach of Omi and Winant does not go far enough in probing empirically and theorizing thoroughly the foundational and systemic reality of racial oppression in white-settler societies like the United States. The concept of *systemic racism* (also termed *institutional racism*) emphasizes the importance of these determining patterns of racial oppression that are threaded throughout a society's everyday operations and institutions – and well beyond the state institutions. From its major development in the seventeenth century to the twenty-first century, white-imposed racism has been about much more than whites'racist attitudes and biases, for it has long encompassed a complex array of racially inegalitarian economic, political, and other social relationships developed and sustained over many generations and embedded in all major institutions in the United States and other Eurocentric societies (Du Bois, 1965 [1946]; Feagin, 2010a).

In the US case, from the 1600s to the present, the dominant white group – especially, the white male elite – has created and maintained the hierarchical patterns of racial oppression. They have created and perpetuated the oppressive practices that unjustly enrich the dominant white group and unjustly impoverish groups of color. The long history of white oppression in the US case has been well documented by an array of social scientists (Frederickson, 2008/1971, 1987; Singh, 2005; Feagin, 2010a). In racist systems, such as that of the United States, most whites become privileged stakeholders in a hierarchical structure of opportunities, resources, wealth, and power linked to a long history of racial exploitation and oppression.

Among others, Joe Feagin (2010a; see also Feagin and Feagin, 2011) has argued that the concept of *systemic racism* encompasses both the creation and sustained maintenance of substantial white privilege and power in society and the construction of a broad *white racial frame* to explain and rationalize that unjustly gained privilege and power. Individuals, whether they are the perpetrators or recipients of racial oppression, are caught in a complex web of alienating individual and group relationships. Systemic racism divides human beings and severely impedes the development of a common consciousness. At the macro level, large-scale institutions routinely perpetuate the subordination of certain racial groups to yet other racial groups. In turn, these racially inegalitarian institutions are maintained by recurring discriminatory practices of dominant-group individuals at the micro level of everyday interaction. Individual acts of discrimination activate, on a recurring basis, whites' major power-and-privilege position, a point underscored by, among others, the pioneering European scholar of institutional racism, Philomena Essed (1991).

In some theorizing about contemporary racial conflict and inequality, there is an emphasis on the growing importance of class, defined in the Weberian sense of socioeconomic status. Some of these analysts argue that systemic racism is of declining importance in Western societies like the United States and increasingly is consigned to societal fringes, whereas socioeconomic (class) divisions are increasing (Wilson, 1978). Yet research evidence suggests that such a conclusion is

premature. Critics (e.g., Omi and Winant, 1994; Feagin, 2010a) of this view of the 'declining significance of race' have shown that, for the United States at least, race cannot be reduced to socioeconomic status and divisions, but rather it remains an autonomous and continuing field of much intergroup conflict, political organization, and ideological meanings.

Some contemporary theorizing about racial and ethnic relations, especially 'segmented assimilation' theory, draws heavily on the mainstream assimilation tradition, yet also makes use of power-conflict ideas. Researchers like Alejandro Portes (1995) have argued that societal outcomes of immigrant adaptation vary greatly, with groups of color like Mexican Americans substantially confined to the lower rungs of US society's economic ladder in later generations, even as other groups such as early twentieth-century European immigrants have experienced over generations much faster socioeconomic mobility. Researchers (Portes, 1995) have found that the second-generation children of relatively recent Mexican immigrants, unlike children of European immigrants, continue to face racial-ethnic discrimination and related difficult economic circumstances, including disproportionate placement in low-paying jobs and inadequate housing. These researchers underscore the significant problems associated with traditional immigrant-assimilation theories when they are applied to immigrants of color and have insisted on spelling out who is being assimilated and under what circumstances.

Continuing Racial Oppression: The US Case

In the twenty-first century, people of African descent appear to be the globe's largest racially oppressed group; they are now resident in a number of countries as a result of the overseas slave trade carried out by Europeans centuries ago. One of the major groups in this large-scale African diaspora includes the residents in the United States, where much research evidence remains of continuing white-on-black discrimination and other racial oppression.

Recent field studies and opinion surveys show that among white Americans, racist attitudes and other racial framing remain widespread. Psychologists recently have given implicit racial attitude tests online to thousands of white Americans. The overwhelming majority quickly associate black faces with negative words and traits, yet they have more difficulty in quickly associating black faces with pleasant words and positive traits than they do for white faces. In these and other research tests, most whites of various educations and occupations have revealed a strong antiblack and a prowhite bias (for a summary, see Feagin, 2010a, pp. 102–103). In addition, recent surveys of African Americans have found that a substantial majority report that they as a group face significant discrimination from whites in local housing, employment, restaurants, and other institutional areas (Feagin, 2010a, p. 120). One focus-group study (Nuru-Jeter et al., 2009) involving black women found that they reported numerous instances of discrimination in schools, housing, shopping, and health care institutions; additionally, they stressed how vigilant they must be in countering the everyday racism faced by themselves and their families. Numerous studies demonstrate that African Americans continue to face significant discrimination from police officers doing racial profiling and from real estate agents and landlords who engage in antiblack housing discrimination (Smith, 2006).

In the United States, as in numerous other white-controlled countries, whites also direct much racist framing, hostility, and discrimination at other groups of color. Numerous research studies have shown that many Asian Americans are moving into the mainstream of the US economy and suburban neighborhoods, yet as they do so, they still encounter much racial discrimination at the hands of whites. Blue-collar and white-collar workers, including administrators and professionals in historically white institutions, report ingrained patterns of anti-Asian discrimination (Cheng, 1997; Feagin, 2010a, pp. 244–245). Similarly, Latino Americans report much continuing discrimination in employment and housing. Political discrimination targets Latino and Asian Americans. Over recent decades, voters in the state of California, especially white voters, have supported ballot propositions designed to abolish government services for Latino and Asian immigrants and eliminate antidiscrimination (e.g., affirmative action) and bilingual programs. White hostility to Spanish-speaking Americans of color persists, such as in the widespread mocking of the Spanish language and other aspects of Latino culture in the United States (Feagin, 2010a, pp. 221, 240–247).

Racism in Europe: The Case of France

The United States is not the only Western nation pervaded by a white racial framing of society and thus of racist relations. The reality of racial oppression can be seen in the growth of neo-Nazi groups and the often-associated outbreak of hate crimes in many European countries since the 1980s, as well as in routine discrimination in housing and employment in those European countries. For example, in France, immigrants from Algeria and other African countries often face racist framing, hostility, and discrimination much like that of the United States. France imposed colonial rule on Algeria in the nineteenth century; later, beginning in the 1960s, large numbers of Algerian immigrants migrated to France. Defining these and other African immigrant groups as undesirable, many French whites have publicly alleged the lack of assimilation by African immigrants into (white) French culture as the major issue. Yet, these whites typically play down or ignore the way in which such societal realities as housing separation and other social segregation result from institutionalized, race-linked discrimination by whites against French people of African descent. In France racist–right political groups like the French National Front – with its strong white racial framing, anti-immigrant, and anti-multiculturalism agendas – have grown in social and political influence in recent years. In the 2012 election, the head of the National Front ran as a presidential candidate and got millions of votes. The national government in France periodically responded to white voters' pressures with its own racialized anti-immigrant framing and repressive attacks on African immigrants and their French descendants (Batur-VanderLippe, 2011; Hollinger, 2012).

For numerous European countries, overseas colonialism in areas like Africa involved the use of organized violence to exploit the resources of the colonies there and to oppress their indigenous peoples – practices often legitimated with a strong white-racist framing and Eurocentric ideologies (Du Bois, 1965 [1946]; Batur-VanderLippe, 2011). More recently, the descendants of the formerly colonized often face racial oppression in the colonizers' home countries. In a number of European countries, the legacy of colonialism can be seen graphically in the often-violent expressions of racialized hatred for the identity and culture of African, Asian, Middle Eastern, or West Indian immigrants and their descendants in Europe.

Still, there are some differences between Europe's racist patterns and those of the United States. One example can be seen in the dramatic increase in white supremacist and other hate-oriented websites on the Internet. The number of these websites is estimated to be at least 5000, with some estimates being much higher. Although European countries do have a serious problem with racist and other hate websites and their mostly white followers, a majority of the major racist and other hate-speech websites are based in the United States – in part because of the way that judges and politicians in the United States interpret the US Constitution's first amendment to protect most hate speech. In contrast, many European countries, including Italy, Austria, England, France, Germany, and the Netherlands, have important laws that significantly restrict white supremacist and other hate speech both on the Internet and in other societal settings.

Racism in Former European Empires: The South African Case

In many ways South Africa is a major symbol of racist relations as they developed between European and non-European peoples in the ever-spreading European empires from the sixteenth to the nineteenth centuries. In the South African case, two colonizing powers, Great Britain and Holland, invaded the lands of the indigenous African peoples, and their white descendants became the economically and politically dominant group in the country. In the first decade of the twentieth century, a new country, called the Union of South Africa, was formed, combining white residents of British and Dutch origin as the ruling group over several black African groups. Black Africans gradually were forced into a system of comprehensive racial segregation (called apartheid) where the relations between the white and black groups, as well as between these groups and middle groups like the Asians and the 'coloreds' (those of mixed racial backgrounds), were determined by an asymmetrical racial hierarchy and an extensive system of racial segregation laws. Over time, the racial apartheid system came to encompass most aspects of the relations between white residents and residents of color, including relationships of great racial oppression in the major institutional areas of politics, employment, education, and housing (Batur-VanderLippe, 2011; Feagin and Feagin, 2011).

As in the case of African Americans, this comprehensive system of racial oppression especially targeted *black* South Africans. As a result, apartheid generated much black resistance in the form of major organizations like the African National Congress (ANC). The ANC, working with other antiapartheid groups, engaged in collective protests and eventually forced an end to white-imposed apartheid, thereby bringing new hope for improved living conditions for the oppressed black majority. Racial relations in numerous spheres dramatically changed. In the mid-1990s, the Republic of South Africa (the country's name changed in the 1960s) moved from white to black political control and began to change its traditional patterns of political, economic, and social apartheid. This societal change has so far marked only a significant beginning, for the important political changes have not yet ended extensive economic domination by whites, and the country still ranks very high among the world's countries in unemployment and socioeconomic inequality along racial lines. Racially discriminatory practices – blatant, subtle, and indirect – still are found in numerous areas, including employment, housing, and education. Educational and other socioeconomic mobility opportunities, although much better than in the past, are limited severely by the unjust impoverishment of most black African families over generations of racial apartheid imposed by whites and its continuing impact on the assets and resources available to black families to the present day. Attempts at remedial and reparatory action by the black-led government have helped to create a black middle class, but they have not yet provided much sustained improvement in socioeconomic conditions for the black working-class majority (Batur-VanderLippe, 2011; Feagin and Feagin, 2011).

These late-twentieth-century events in South Africa have helped to link black African-origin peoples around the globe in a common struggle against the descendants of European colonists, who have long maintained economic or political control in numerous countries where people of color are the numerical majority. Struggles against systemic racism in one country often inspire or inform struggles in other countries. Since the early 1900s, for example, black American efforts for liberation from white racism, including recurring black civil rights movements, have provided inspiration and assistance for struggles of people of color in South Africa and other countries with significant oppressed populations (e.g., Brazil). In a similar fashion, the successes of black Africans in South Africa in bringing down political and other apartheids have encouraged African-origin peoples in the United States and other countries in the Americas in their struggles against racial framing and racial discrimination. Indeed, African Americans have long drawn on the history, spirituality, and symbolism of Africa as part of their individual and group strategies for countering and coping with a white-dominated and racialized US society. These cross-national examples again suggest how racial relations and racist histories in one country have been set within a larger global racial context. This has been true for centuries now (Feagin and Feagin, 2011; Batur-VanderLippe, 2011).

The Future of the Global Racial Order

Over several centuries, the global expansion of capitalism by European entrepreneurs, colonists, and missionaries – European colonialism and imperialism – frequently generated an extensive system of white-racist oppression and

framing, one with a global scope. Eventually, in the early twentieth century, a substantial majority of this planet's people of color were dominated in one way or another by people of European descent (see Batur-VanderLippe, 2011). Yet this oppression eventually generated large-scale resistance and periodically substantial change.

In many countries around the globe, recurring challenges to white racial domination and its discriminatory consequences are arising not only from human-rights organizations' efforts but also from important demographic and sociopolitical changes. Europeans and European Americans are a decreasing fifth of the world's population and a decreasing proportion of the US population. In the second decade of the twenty-first century, US whites are a statistical minority in many large cities and in the states of New Mexico, Hawaii, California, and Texas. Demographic projections indicate that no later than the year 2050, white Americans will be a statistical minority of the US population, for the first time since the country's beginning in the 1790s. According to a 2011 US Census Bureau's official count, the white non-Hispanic population percentage had declined to just 63%, and the white proportion of babies born had dropped for the first time to just under 50%.

As whites become the statistical minority in the United States and in the many national and international organizations they once dominated, traditional patterns of white-dominant racial relations will be under greater pressure to change. The demographic changes have significant political, economic, and other social implications. For the US case, by the 2040s, the majority of the labor force will very likely not be white, and the population and labor force will be older. This likely will give Americans of color yet more political and economic clout. Political–economic tensions may well increase between a mostly white retired population and the majority of the nonwhite working population. In the United States and other Western countries, white politicians intensely opposed to legal immigration by non-Europeans and to major remedial programs for racial discrimination likely will not be elected when much larger proportions of political constituencies become voters of color. As these constituencies change away from white dominance, it seems probable that juries and justice systems, education systems, government policy-making arenas, and other societal institutions also will change in the direction of more influence and power for citizens of color.

Around the globe, the relations of colonialism and systemic racism are slowly but dramatically changing. The deep-lying contradictions of the global racial order were set into motion by the colonial and imperialistic ventures of numerous European nations. This European colonialism generated sociopolitical structures that have long embedded the practices, ideologies, and institutions of racist subordination in numerous countries. In the twenty-first century, international relations, global markets, the global mainstream media, and multinational corporations are all racialized, with white European perspectives and white or white-oriented executives usually in control. Half a millennium of European-dominated national and international institutions is slowly coming to an end. As the twenty-first century develops further, asymmetrical racial relations and racial hierarchies in Europe, the United States, and other parts of the globe are likely to change substantially. Many once racially oppressed groups probably will move farther out from under the racialized dominance of Europeans and European Americans.

Not surprisingly, over the past few decades, strong anti-racism movements have arisen in South Africa, Brazil, the United States, and other countries with colonial histories. The possibility of a global democratic order substantially rid of the oppressive relations of white-imposed racism remains only a dream, but the US civil rights movements and the South African political revolution, among other racialized global struggles, demonstrate that it is still a powerful dream for a great many people around the globe.

See also: Affirmative Action, Sociology of; Colonization and Colonialism, History of; Discrimination: Racial; Environmental Justice in the United States; Ethnic Conflicts; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity, Sociology of; Hate Crimes; Policing; Race Identity; Race and the Law; Race: Ethnicity and Urban Planning; Racism, History of; Racism, Sociology of; Residential Segregation in the United States; Slavery as Social Institution; Slaves and Slavery, History of.

Bibliography

Alba, R., 2009. Blurring the Color Line: The New Chance for a More Integrated America. Harvard University Press, Cambridge.

Alba, R., Nee, V., 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Banton, M.P., 1983. Racial and Ethnic Competition. Cambridge University Press, Cambridge.

Batur-VanderLippe, P., 2011. Colonialism and post-colonialism: the global expansion of racism. In: Feagin, J.R., Feagin, C.B. (Eds.), Racial and Ethnic Relations, ninth ed. Prentice-Hall, Upper Saddle River, NJ.

Blauner, R., 1972. Racial Oppression in America. Harper & Row, New York.

Brock, A., 2009. Life on the wire. Information, Communication & Society 12, 344–363.

Cheng, C., 1997. Are Asian American employees a model minority or just a minority? Journal of Applied Behavioral Science 33, 277–290.

Cox, O.C., 1948. Caste, Class, and Race. Doubleday, Garden City, NY.

Du Bois, W.E.B., 1965 [1946]. The World and Africa. International Publishers, New York.

Essed, P., 1991. Understanding Everyday Racism. Sage, Newbury Park, CA.

Feagin, J.R., 2010a. Racist America: Roots, Current Realities, and Future Reparations. Routledge, New York.

Feagin, J.R., 2010b. The White Racial Frame: Centuries of Racial Framing and Counter-framing. Routledge, New York.

Feagin, J.R., Elias, S., 2012. Rethinking racial formation theory: a systemic racism critique. Ethnic and Racial Studies 36, 1–30.

Feagin, J.R., Feagin, C.B., 2011. Racial and Ethnic Relations, ninth ed. Prentice-Hall, Upper Saddle River, NJ.

Frederickson, G., 2008/1971. The Black Image in the White Mind: The Debate on African-American Character and Destiny, 1817–1914. ACLS Humanities E-Book, New York.

Gomez, L.E., 2002. Race mattered: racial formation and the politics of crime in territorial New Mexico. UCLA Law Review 49, 13–95.

Gordon, M.M., 1964. Assimilation in American Life. Oxford University Press, New York.

Hannaford, I., 1996. Race: The History of an Idea in the West. Johns Hopkins University Press, Baltimore, MD.

Hollinger, P., 2012. Le Pen daughter applauds Cameron. Financial Times, February 9, 2011. http://www.ft.com/intl/ (accessed 17.05.12.).

Myrdal, G., 1964 [1944]. An American Dilemma. McGraw-Hill, New York.

Nuru-Jeter, A., et al., 2009. It's the skin you're in: African-American women talk about their experiences of racism. Maternal and Child Health Journal 13, 33–39.

Omi, M., Winant, H., 1994. Racial Formation in the United States: From the 1960s to the 1990s, second ed. Routledge, New York.

Park, R.E., 1950. Race and Culture. Free Press, Glencoe, IL.

Pedraza, S., 1985. Political and Economic Migrants in America: Cubans and Mexicans. University of Texas Press, Austin.

Portes, A., 1995. Segmented assimilation among new immigrant youth: a conceptual framework. In: Rumbaut, R., Cornelius, W. (Eds.), California's Immigrant Children. Center for Mexican American Studies, University of California, La Jolla, CA.

Rex, J., 1973. Race, Colonialism and the City. Routledge and Kegan Paul, London.

Singh, N.P., 2005. Black Is a Country: Race and the Unfinished Struggle for Democracy. Harvard University Press, Cambridge, MA.

Smith, S.L., February 28, 2006. Testimony before the House Financial Services Committee, Subcommittee on Housing and Community Opportunity. National Fair Housing Alliance.

Tuan, M., 1998. Forever Foreigners or Honorary Whites?: The Asian American Experience Today. Rutgers University Press, New Brunswick, NJ.

Wilson, W.J., 1978. The Declining Significance of Race: Blacks and Changing American Institutions. University of Chicago Press, Chicago, IL.

Racism, History of

GM Fredrickson, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 12716–12720, © 2001, Elsevier Ltd.

Abstract

Initially seen by historians as a theory of difference, racism is now usually seen as an ideology with an integral relationship to domination or exclusion. The most conspicuous forms of modern racism – white supremacy and antisemitism – have been studied separately. A general or comparative history could be based on a definition of racism as an ideology sanctioning the domination or exclusion of one ethnic group by another on the basis of difference believed to be hereditary and unalterable. Racism can, thus, be distinguished from xenophobia and religious intolerance. It emerged in the late medieval and early modern periods in conceptions of Jewishness as ancestry rather than belief and of blackness as a curse condemning Africans to servitude. The scientific revolution and the Enlightenment provided foundations for naturalistic racism. The struggle over slavery and the accelerated expansion of Europe into Africa and Asia brought color-coded racism to full consciousness in the nineteenth century. European antisemitic racism emerged in the context of modernity and the emancipation of the Jews. Racism climaxed during the twentieth century in the segregation of blacks in the American South and in South Africa, and in the genocidal assault on European Jewry.

Initially viewed by historians as a theory of difference, racism is now usually seen as an ideology with an integral relationship to domination or exclusion. The most conspicuous forms of modern racism – white supremacy and antisemitism – have been studied separately. A general or comparative history could be based on a definition of racism as an ideology sanctioning the domination or exclusion of one ethnic group by another on the basis of difference believed to be hereditary and unalterable. Racism can, thus, be distinguished from xenophobia and religious intolerance. It emerges in the late medieval and early modern periods in conceptions of Jewishness as ancestry rather than belief and of blackness as a curse condemning Africans to servitude. The scientific revolution and the Enlightenment provided foundations for naturalistic racism. The struggle over slavery and the accelerated expansion of Europe into Africa and Asia brought color-coded racism to full consciousness in the nineteenth century. European antisemitic racism emerged in the context of modernity and the emancipation of the Jews. Racism climaxed during the twentieth century in the segregation of blacks in the American South and in South Africa and in the genocidal assault on European Jewry.

In common usage the term racism is often employed in a loose and unreflective way to describe the hostile or negative feelings of one ethnic group or historical collectivity toward another and the actions resulting from such attitudes. But historians and social scientists normally require a more precise and limited definition, although they may disagree on what it should be. It is useful, therefore, to review the historiographic discourse on the meaning of racism before attempting a brief survey of what contemporary historians might accept as its principal manifestations.

Historiography of Racism

The term racism was coined in the 1930s mainly in reference to the theories on which the Nazis based their persecution of the Jews. The first historian to focus directly on the subject was Barzun. His book *Race: A Study in Superstition* (originally published in 1937 and reprinted in 1965) set two important precedents for most future historians of racism: it presumed that claims of the innate inferiority of one 'race' to another were false or at least unproven, and its main concern was with the history of ideas rather than with the social and political applications of prejudiced beliefs and attitudes. Barzun's concern with the intellectual origins of Nazi antisemitism was shared by the many scholars who wrote on the subject after World War II in response to what the Holocaust had revealed about the horrifying consequences of racist ideas. A notable example was Poliakov's *Le Myth Aryen*, translated into English as *The Aryan Myth* (1974). But disagreements developed on whether biological racism was continuous with earlier antisemitic attitudes that were, ostensibly at least, based on religion or whether it was a radical new departure, a *sine qua non* for something like the Holocaust.

In the immediate postwar decades, the term racism was also applied with increasing frequency to the relationships between whites or Europeans and people of color, especially Africans or those of African descent, in colonial, postcolonial, and former slave societies. The rising concern with black-white relations in the US, South Africa, and (to a somewhat lesser extent) in Great Britain, beginning in the 1950s and early 1960s, gave rise to an enormous literature on the origins and development of white supremacy as an ideology. Historians of ideas did most of the early work, but gradually the emphasis shifted somewhat from the racist ideas themselves to the patterns or racial discrimination and domination for which they served as a rationale. Many historians of black-white relations used the term casually and without reflecting on its exact meaning. Most imprecisely, it could mean anything that whites did or thought that worked to the disadvantage of blacks.

There have been relatively few attempts to write general histories of racism, which encompass both the antisemitic and color-coded varieties. For the most part, the two historiographies have not addressed each other, and the definitions

employed have sometimes been incompatible. Some historians of antisemitic racism have found the term applicable only if the aim was the elimination of the stigmatized group. Some historians of white supremacy, on the contrary, would limit the term to ideologies associated with patterns of domination or hierarchical differentiation. The first attempt at a comprehensive history of American racism, Gossett's *Race: The History of an Idea in America* (1963) traced consciousness of race to the ancient world. Its treatment of specifically American manifestations cast its net quite wide to include representation and treatment of American Indians and immigrants from southern and eastern Europe, including Jews, as well as blacks. But most subsequent work on the history of American racism has been group-specific and has concentrated most heavily on attitudes toward African Americans. Mosse's general history of European racism, on the other hand, pays most of its attention to the growth of antisemitism between the Enlightenment and the Holocaust (Mosse, 1978). But there appear to be only two significant attempts to cover Western attitudes toward race comprehensively: Hannaford's *Race: The History of an Idea in the West* (1996) and Geiss' *Geschichte des Rassismus* (1988). Hannaford's study, as its title indicates, is strictly an intellectual history and considers race as a concept more than racism as an ideology. It argues strenuously that no clear concept of race existed before the seventeenth century, thus raising the issue of whether anything that existed before the invention of race in the modern sense can be legitimately labeled racism. Geiss, on the contrary, sees racism as anticipated in most respects by the ethnocentrism or xenophobia that developed in the ancient world, as reflected, for example, in the Old Testament.

Toward a Definition of Racism for Historical Purposes

Somewhere between the view that race is a peculiar modern idea without much historical precedent and the notion that it is simply an extension of the ancient phenomena of ethnocentrism and xenophobia may lie a working definition that is neither too broad for historical specificity or too narrow to cover more than the limited span of Western history during which a racism based on scientific theories of human variation was widely accepted. If racism is defined as an ideology rather than as a theory, links can be established between belief and practice that the history of ideas may obscure. But ideologies have content, and it is necessary to distinguish racist ideologies from other belief systems that emphasize human differences and can be used as rationalizations of inequality. The classic sociological distinction between racism and ethnocentrism is helpful, but not perhaps in the usual sense, in which the key variable is whether differences are described in cultural or physical terms. It is actually quite difficult in specific historical cases to say whether appearance or 'culture' is the source of the salient differences, because culture can be reified and essentialized to the point where it has the same deterministic effect as skin color. But it would stretch the concept of racism much too far to make it cover the pride and loyalty that may result from ethnic identity. Such groupcenteredness may engender prejudice and discrimination against those outside the group, but two additional elements would seem to be required before the categorization of racism is justified. One is a belief that the differences between the ethnic groups involved are permanent and ineradicable. If conversion or assimilation is genuinely on offer, we have religious or cultural intolerance but not racism. The second is the social side of the ideology – its linkage to patterns of domination or exclusion. To attempt a short formulation, we might say that racism exists when one ethnic group or historical collectivity dominates, excludes, or seeks to eliminate another on the basis of differences that it believes are hereditary and unalterable.

The Emergence of Racism

Historians who can accept this middle-of-the-road definition can agree with the scholars who have examined conceptions of human diversity in classical antiquity and the early Middle Ages without finding clear evidence of color prejudice or racism. Neither the classical conception that one becomes civilized through being able to engage in politics nor the Christian belief that everyone regardless of color or ancestry is a potential convert with a soul to be saved could readily sustain a racist view of the world. The period between the fourteenth and the seventeenth centuries is much more problematic. The literal demonization of the Jews in the late Middle Ages, the belief that they were in league with the devil and plotting the destruction of Christianity, could lead unsophisticated folk to believe that they were outside the bounds of humanity, even though church authorities persisted in advocating their conversion. Folk antisemitism resulted in massacres and expulsions of Jews in virtually every country in Western Europe. A degree of elite or official sanction for such attitudes came in sixteenth century Spain when Jews who had converted to Christianity and their descendants became the victims of a pattern of discrimination and exclusion. When purity of blood, or *limpieza de sangre*, became a qualification for certain offices and honors it signified that what people could accomplish or achieve no longer depended on what they did but on who they were. Precedents for the notion that blood will tell could be found in the inheritability of royal or noble status but the designation of an entire ethnic group as incapable of a reliable conversion to Christianity constituted, at the very least, a transitional stage between religious bigotry and biological racism.

The period of the Renaissance and Reformation was also the time when Europeans were coming into increasing contact with people of darker pigmentation in Africa, Asia, and the Americas and were making judgments about them as they traded with them, fought with them over territory, or enslaved them. A belief that all these non-Europeans were children of God with souls that missionaries could save did not prevent a brutal expropriation of their land and labor, but it did inhibit the articulation of a racial justification for how they were being treated. When Africans were carried by force to the New World to labor on plantations, slave traders and slave owners sometimes invoked an obscure passage in the book of Genesis to explain the color of their human property. Ham had committed a sin against his father Noah that resulted in a divine curse on his descendents to be 'servants unto servants.' When Africans were enslaved, first by Arabs, and then by Europeans, it became convenient to believe that the

punishment of God had included a blackening of the skin. But the passage had other possible meanings, and no orthodox religious authorities ever endorsed the interpretation of the curse popular among slaveholders, because they feared that it would impede conversion of blacks to Christianity. The standard and official justification for enslaving Africans was that they were heathens and that enslavement made heaven accessible to them.

An implicit racism became evident, however, when converted slaves were kept in bondage, not because they were actual heathens but because they had heathen ancestry. As with the doctrine of purity of blood in Spain, descent rather than performance became the basis for determining the qualifications for membership in a community that was still theoretically based on a shared Christian faith. Beginning in the late seventeenth century laws were passed in English North America forbidding marriage between whites and blacks and discriminating against the mixed offspring of informal liaisons. Without clearly saying so, such laws implied that blacks were unalterably alien and inferior.

The Rise of Modern Racism

A modern naturalistic form of racism could not be articulated fully until the Enlightenment. A secular or scientific theory of race required a new way of looking at human diversity, one that moved the subject away from the Bible, with its insistence on the essential unity and homogeneity of the human race and its collective elevation above the animal kingdom. Eighteenth-century ethnologists began to think of human beings as part of the natural world and subdivided them into three to five races, usually considered as varieties or subspecies. Most eighteenth-century ethnological theorists believed that whiteness was the original color of humanity and that blackness or brownness resulted from a process of degeneration caused by climate and conditions of life. Although the white or Caucasian race was generally considered more beautiful and civilized than the Mongolian and especially the Negro, the belief that human beings constituted a single species with a common origin remained dominant. In the late eighteenth and early nineteenth centuries, however, an increasing number of ethnologists asserted that the races were separately created and constituted distinct species. As debates over the morality of slavery and the slave trade erupted in Western Europe and the US in the wake of democratic revolutions and the rise of humanitarian movements, some of the most militant defenders of black bondage adopted these polygenetic arguments as a justification for enslaving Africans and holding them in permanent servitude. But the fact that polygenesis seemed to conflict with scripture limited the usage of this extreme version of scientific racism. Many defenders of slavery in the American South, where evangelical religion enforced biblical literalism, reverted to the Curse on Ham or made the degeneracy hypothesis of Enlightenment ethnology into an argument for the permanent or irreversible divergence of the races from their common origin.

The nineteenth century was an age of emancipation, nationalism, and imperialism – all of which contributed to the growth and intensification of ideological racism in Europe and in the US. Although the emancipation of blacks from slavery and Jews from the ghettoes, which seemed a logical consequence of the democratic revolutions of the late eighteenth century, received most of its support from religious or secular believers in an essential human equality, the consequences of these reforms was to intensify rather than diminish racism. Race relations became less paternalistic and more competitive. Poor or working class whites in the US, for example, feared competition with emancipated blacks for work, land, and social status. Middle-class (especially lower middle class) Germans viewed the increasing involvement of Jews in finance, commerce, the professions, journalism, and the arts during the late nineteenth century as a potential threat to their own economic security, social status, and traditional values. Of course, it was the legacy of inherited stereotypes about 'the Other,' in addition to the immediate economic and social circumstances, that gave edge and intensity to these feelings of insecurity and hostility. The idiom of Darwinism, with its emphasis on 'the struggle for existence' and concern for 'the survival of the fittest' was conducive to the development of a new and more credible scientific racism in an era that increasingly viewed race relations as an arena for conflict rather than as a stable hierarchy.

The growth of nationalism, especially romantic cultural nationalism, encouraged the growth of a more mystical variant of racist thought. When it could be assumed that one had to be of German ancestry to participate in the German *Volkgeist*, Jews could be discriminated against or excluded just as effectively as if their deficiencies were clearly defined as genetic or biological. For most of the nineteenth century the question of whether German Jews could or should be converted to Christianity was an open one. Beginning in the late 1870s and early 1880s, however, the coiners of the term 'antisemitism' made explicit what some cultural nationalists had previously implied – that to be Jewish was not simply to adhere to a set of religious beliefs or cultural practices but meant belonging to a race that was the antithesis of the race to which true Germans belonged. The latter was variously designated as Teutonic, Aryan, Nordic, or simply Germanic. In Great Britain and the US romantic nationalism took the form of a veneration for Anglo-Saxon ancestors and legacies that could turn racist, as it did for a time in the US, when confronted with large-scale immigration from eastern and southern Europe. In France, an ethnocentric cultural nationalism fed the fury against the Jews that surfaced in the Dreyfus affair at the turn of century. It did not triumph against the universalistic civic nationalism inherited from the revolutionary past, but it did make antisemitic racism a strain of right-wing French thought that would resurface during the 1930s and especially under the German occupation.

The climax of Western imperialism in the late nineteenth century 'scramble for Africa' and parts of Asia and the Pacific represented an assertion of the competitive ethnic nationalism that existed among European nations (and which, as a result of the Spanish-American War came to include the US). A belief that different races, subraces, or racial mixtures inhabited each nation led some, especially the Germans and the English, to suppose that the quality of each nation's racial stock was being tested for its fitness. But imperialism also required the massive subjugation and colonial domination of non-European populations. An ideology proclaiming that whites were superior to

'lesser breeds' and responsible for ruling over them and tutoring them in the rudiments of civilization was a prime rationale for this new burst of European expansionism, even if it was not a cause of it. The relation of racism to imperialism is more problematic than one might suppose, however, because the most consistent and extreme racists were often anti-imperialists. They believed that nothing useful could be done with people whose inferiority was so profound and permanent that they were incapable of being civilized. For them the taming or domestication of such savages was not worth the trouble. Within colonies that attracted substantial numbers of European settlers, such as French Algeria or some of the British colonies of South and East Africa, a more extreme and explicit racism than the one usually professed by government officials or missionaries could be found in the discourse of white colonists about their 'native' servants or farm laborers.

The Climax: Racism in the Twentieth Century

The climax of the history of racism came in the twentieth century in the rise and fall of what might be called overtly racist regimes. In the American South, the passage of racial segregation laws and restrictions on black voting rights reduced African Americans to lower caste status, despite the Constitutional Amendments that had made them equal citizens. Extreme racist propaganda, which represented black males as ravening beasts lusting after white women, served to rationalize the practice of lynching. These extralegal executions increasingly were reserved for blacks accused of offenses against the color line, and they became more brutal and sadistic as time went on; by the early twentieth century victims were likely to be tortured to death rather than simply being killed. A key feature of the racist regime maintained by state law in the South was a fear of sexual contamination through rape or intermarriage, which led to efforts to prevent the conjugal union of whites with those with any known or discernable African ancestry. The effort to guarantee 'race purity' in the American South anticipated aspects of the official Nazi persecution of Jews in the 1930s. The Nuremberg laws of 1935 prohibited intermarriage or sexual relations between Jews and Gentiles, and the propaganda surrounding the legislation emphasized the sexual threat that predatory Jewish males presented to German womanhood and the purity of German blood. Racist ideology was eventually of course carried to a more extreme point in Nazi Germany than in the American South of the Jim Crow era. Individual blacks had been hanged or burned to death by the lynch mobs to serve as examples to insure that the mass of southern African Americans would scrupulously respect the color line. But it took Hitler and the Nazis to attempt the extermination of an entire ethnic group on the basis of a racist ideology.

Hitler, it has been said, gave racism a bad name. The moral revulsion of people throughout the world against what the Nazis did, reinforced by scientific studies undermining racist genetics (or eugenics), served to discredit the scientific racism that had been respectable and influential in the US and Europe before World War II. But explicit racism also came under devastating attack from the new nations resulting from the decolonization of Africa and Asia and their representatives in the United Nations. The Civil Rights movement in the US, which succeeded in outlawing legalized racial segregation and discrimination in the 1960s, was the beneficiary of revulsion against the Holocaust as a logical outcome of racism. But it also drew crucial support from the growing sense that national interests were threatened when blacks in the US were mistreated and abused. In the competition with the Soviet Union for 'the hearts and minds' of independent Africans and Asians, Jim Crow and the ideology that sustained it became a national embarrassment with possible strategic consequences.

The one overtly racist regime that survived World War II and the Cold War was the South African, which did not in fact come to fruition until the advent of Apartheid in 1948. The laws passed banning all marriage and sexual relations between different 'population groups' and requiring separate residential areas for people of mixed race ('Coloreds'), as well as for Africans, signified the same obsession with 'race purity' that characterized the other racist regimes. However, the climate of world opinion in the wake of the Holocaust induced apologists for apartheid to avoid straightforward biological racism and rest their case for 'separate development' mainly on cultural rather than physical differences. The extent to which Afrikaner nationalism was inspired by nineteenth-century European cultural nationalism also contributed to this avoidance of a pseudoscientific rationale. No better example can be found of how a 'cultural essentialism' based on nationality can do the work of a racism based squarely on skin color or other physical characteristics. The South African government also tried to accommodate itself to the age of decolonization. It offered a dubious independence to the overcrowded 'homelands,' from which African migrants went forth to work for limited periods in the mines and factories of the nine-tenths of the country reserved for a white minority that constituted less than a sixth of the total population.

The End of Racism?

The defeat of Nazi Germany, the desegregation of the American South in the 1960s, and the establishment of majority rule in South Africa suggest that regimes based on biological racism or its cultural essentialist equivalent are a thing of the past. But racism does not require the full and explicit support of the state and the law. Nor does it require an ideology centered on the concept of biological inequality. Discrimination by institutions and individuals against those perceived as racially different can long persist and even flourish under the illusion of nonracism, as historians of Brazil have discovered recently. The use of allegedly deep-seated cultural differences as a justification for hostility and discrimination against newcomers from the Third World in several European countries has led to allegations of a new 'cultural racism.' Similarly, those sympathetic to the plight of poor African Americans and Latinos in the US have described as 'racist' the view of some whites that many denizens of the ghettos and barrios can be written off as incurably infected by cultural pathologies. From the historian's perspective such recent examples of cultural determinism are not in fact unprecedented. They rather represent a reversion to the way that the differences between ethnoracial groups could be made to seem indelible and unbridgeable before the articulation of a scientific or naturalistic conception of race in the eighteenth century.

See also: Antisemitism; Discrimination and the Law; Discrimination, Economics of; Discrimination: Racial; Ethnic Identity, Psychology of; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity: Anthropological Aspects; Holocaust, The; Race Identity; Race and the Law; Racial Relations; Racism, Sociology of.

Bibliography

Banton, M., 1998. Racial Theories. Cambridge University Press, Cambridge, UK.

Barzun, J., 1965/1937. Race: A Study in Superstition. Harper and Row, New York.

Bernasconi, R., Anderson, S.C., 2003. Race and Racism in Continental Philosophy. Indiana University Press, Bloomington.

Bhattacharyya, G., Gabriel, J., Small, S., 2002. Race and Power: Global Racism in the Twenty-First Century. Routledge, London.

Dwyer, C., Bressey, C., 2008. New Geographies of Race and Racism. Ashgate, Aldershot, England.

Farber, P.L., Cravens, H., 2009. Race and Science: Scientific Challenges to Racism in Modern America. Oregon State University Press, Corvallis, OR.

Fredrickson, G.M., 1981. White Supremacy: A Comparative Study in American and South African History. Oxford University Press, New York.

Geiss, I., 1988. Geschichte des Rassismus. Suhrkamp, Frankfurt, Germany.

Gossett, T.F., 1963. Race: The History of an Idea in America. Southern Methodist University Press, Dallas, TX.

Gotkowitz, L., 2011. Histories of Race and Racism: The Andes and Mesoamerica from Colonial Times to the Present. Duke University Press, Durham.

Hamilton, C.V., Comparative Human Relations Initiative (Project), 2001. Beyond Racism: Race and Inequality in Brazil, South Africa, and the United States. Lynne Rienner Publishers, Boulder, CO.

Hannaford, I., 1996. Race: The History of an Idea in the West. Johns Hopkins University Press, Baltimore, MD.

Howard, D.J., 2001. Coloring the Nation: Race and Ethnicity in the Dominican Republic. Signal Books, Oxford, UK.

Johnson, M.D., 2007. Race and Racism in the Chinas: Chinese Racial Attitudes toward Africans and African-Americans. AuthorHouse, Bloomington, IN.

Kidd, C., 2006. The Forging of Races: Race and Scripture in the Protestant Atlantic World, 1600–2000. Cambridge University Press, Cambridge, UK.

Kowner, R., Demel, W., 2013. Race and Racism in Modern East Asia: Western and Eastern Constructions. Brill, Leiden.

Lang, B., 2000. Race and Racism in Theory and Practice. Rowman & Littlefield, Lanham, MD.

Langmuir, G.I., 1990. Toward a Definition of Antisemitism. University of California Press, Berkeley, CA.

Lauren, P.L., 1988. Power and Prejudice: The Politics and Diplomacy of Racial Discrimination. Westview Press, Boulder, CO.

Love, E.T.L., 2004. Race over Empire: Racism and U.S. Imperialism, 1865–1900. University of North Carolina Press, Chapel Hill.

Moore, J.H., 2008. Encyclopedia of Race and Racism. Macmillan Reference USA/ Thomson Gale, Detroit.

Mosse, G.L., 1978. Toward the Final Solution: A History of European Racism. H. Fertig, New York.

Perry, R.J., 2007. "Race" and Racism: The Development of Modern Racism in America. Palgrave Macmillan, New York.

Pieterse, J.N., 1992. White on Black: Images of Africa and Blacks in Western Popular Culture. Yale University Press, New Haven, CT.

Pitcher, B., 2009. The Politics of Multiculturalism: Race and Racism in Contemporary Britain. Palgrave Macmillan, Basingstoke.

Poliakov, L., 1974. The Aryan Myth: A History of Racist and Nationalist Ideas in Europe. Chatto for Essex University Press, London.

Sheriff, R.E., 2001. Dreaming Equality: Color, Race, and Racism in Urban Brazil. Rutgers University Press, New Brunswick, NJ.

Smedley, A., 1993. Race in North America: Origin and Evolution of a World View. Westview Press, Boulder, CO.

Taguieff, P., 1997. Le Racisme. Flammarion, Paris.

Racism, Sociology of

Matthew Clair, Harvard University, Cambridge, MA, USA
Jeffrey S Denis, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P.L. Van Den Berghe, volume 19, pp. 12720–12723, © 2001, Elsevier Ltd.

Abstract

The sociology of racism is the study of the relationship between racism, racial discrimination, and racial inequality. While past scholarship emphasized overtly racist attitudes and policies, contemporary sociology considers racism as individual- and group-level processes and structures that are implicated in the reproduction of racial inequality in diffuse and often subtle ways. Although some social scientists decry this conceptual broadening, most agree that a multivalent approach to the study of racism is at once socially important and analytically useful for understanding the persistence of racial inequality in a purportedly 'postracial' society.

At root, racism is "an ideology of racial domination" (Wilson, 1999: p. 14) in which the presumed biological or cultural superiority of one or more racial groups is used to justify or prescribe the inferior treatment or social position(s) of other racial groups. Through the process of racialization (see Section Racism as a Social Process), perceived patterns of physical difference – such as skin color or eye shape – are used to differentiate groups of people, thereby constituting them as 'races'; racialization becomes racism when it involves the hierarchical and socially consequential valuation of racial groups.

Racism is analytically distinct from racial discrimination and racial inequality. Racial discrimination concerns the unequal treatment of races, whereas racial inequality concerns unequal outcomes (in income, education, health, etc.). While racism is often implicated in both processes, contemporary racial inequalities and forms of discrimination are not always the immediate result of contemporary racism (Pager and Shepherd, 2008). The sociology of racism investigates the relationships between these three phenomena, asking when, how, why, and to what extent they reproduce one another. In the post-Civil Rights era, with (overt) racism now widely condemned, one challenge for social scientists is to conceptualize and measure its more subtle and diffuse manifestations and lasting effects.

Definitions

Racism cannot be defined without first defining race. Among social scientists, 'race' is generally understood as a social construct. Although biologically meaningless when applied to humans – physical differences such as skin color have no natural association with group differences in ability or behavior – race nevertheless has tremendous significance in structuring social reality. Indeed, historical variation in the definition and use of the term provides a case in point.

The term race was first used to describe peoples and societies in the way we now understand ethnicity or national identity. Later, in the seventeenth and eighteenth centuries, as Europeans encountered non-European civilizations, Enlightenment scientists and philosophers gave race a biological meaning. They applied the term to plants, animals, and humans as a taxonomic subclassification within a species. As such, race became understood as a biological, or natural, categorization system of the human species. As Western colonialism and slavery expanded, the concept was used to justify and prescribe exploitation, domination, and violence against peoples racialized as nonwhite. Today, race often maintains its 'natural' connotation in folk understandings; yet, the scientific consensus is that race does not exist as a biological category among humans – genetic variation is far greater within than between 'racial' groups, common phenotypic markers exist on a continuum, not as discrete categories, and the use and significance of these markers varies across time, place, and even within the same individual (Fiske, 2010).

For most social scientists, 'race' is distinct from 'ethnicity'. A major distinction is the assumption of a biological basis in the case of race. Races are distinguished by perceived common physical characteristics, which are thought to be fixed, whereas ethnicities are defined by perceived common ancestry, history, and cultural practices, which are seen as more fluid and self-asserted rather than assigned by others (Cornell and Hartmann, 2006). Thus, Asian is usually considered a 'race', whereas Tibetans and Bengalis are considered ethnicities. Although ethnicity and nationality often overlap, a nationality, such as American, can contain many ethnic groups (e.g., Italian-Americans, Arab-Americans). Yet, all three categories – race, ethnicity, and nationality – are socially constructed, and, as such, groups once considered ethnicities have come to be seen as races and vice versa. Moreover, some groups who are now taken for granted as 'white', such as the Irish, Italians, and Jews, were once excluded from this racial category. The definitional boundaries of race and ethnicity are shaped by the tug and pull of state power, group interests, and other social forces.

From a sociological perspective, it is this social construction of race – not its 'natural' existence – that is the primary object of inquiry in the study of racism. Bundled up with eighteenth century classifications of various racial groups were assertions of moral, intellectual, spiritual, and other forms of superiority, which were used to justify the domination of Europeans over racialized others. In the North American context, racist ideology served as justification for land appropriation and colonial violence toward indigenous peoples as well as the enslavement

of Africans starting in the sixteenth century. It was later used to justify the state-sanctioned social, economic, and symbolic violence directed at blacks and other minorities under Jim Crow laws. In the mid-twentieth century, the American Civil Rights Movement, global anticolonial movements, and increasing waves of non-European immigration to the West changed how individuals, groups, and nation-states talked about, viewed, understood, and categorized race. A major task for sociologists has been to assess these changes and their implications for racial discrimination and inequality.

Intellectual History

There are at least two distinct phases in the sociology of racism, demarcated by the changing nature of race and racism as constructed by social actors and social forces after World War II. The first phase – from the late nineteenth to the mid-twentieth century – typically considered racism as a set of overt individual-level attitudes; the second phase – from the mid-twentieth century to the present – considers racism as not simply explicit attitudes but also implicit biases and processes that are constructed, sustained, and enacted at both micro- and macro-levels. While the first phase focused on the direct relationship between racism and racial inequality, the second phase considers diffuse relationships between these concepts and the ways in which historical, unconscious, institutional, and systemic forms of racism interact with other social forces to perpetuate racial inequality.

In the late nineteenth century, as sociology emerged as a social scientific discipline, few scholars studied racism. (One notable exception was W.E.B. Du Bois, who analyzed the political economic roots of racism and its perverse impacts on Western institutions and psyches.) Instead of studying racism as a social problem, many social scientists – truly products of their time – maintained racist attitudes and incorporated racist assumptions into their explanations of racial group differences in social outcomes. Racism pervaded society, including sociology, and was legitimated by dominant scientific discourses such as Social Darwinism, which misapplied the concept of natural selection to the social world to account for why some (racial, class, etc.) groups excel more than others.

Beginning in the 1920s, when the scientific validity of 'race' came under closer scrutiny, some sociologists – primarily associated with the Chicago School – began to view racism as a distinct social problem worthy of study. The issue took on greater urgency during and after World War II when the devastating consequences of racism reached their ugly peak. This period can be characterized as the first phase in the sociological study of racism, in which dominant theories, drawing on psychology's emphasis on individual prejudice, conceived racism as a set of explicit individual-level beliefs and attitudes that were a historical relic (as opposed to a systemic social process) that would inevitably fade with time. This assumption of inevitable attenuation was most evident in theories of immigrant assimilation, which proposed an inexorable straight-line process, whereby 'ethnic' European immigrants originally racialized as 'other' would gradually assimilate into the American mainstream as full-fledged 'white' citizens.

In this first phase, defining racism as prejudicial beliefs and attitudes provided little difficulty for the social scientist, as individuals, organizations, and the state were explicit about how race mattered for the distribution of material and symbolic resources. For example, income inequality between whites and blacks could be readily explained by workplace discrimination and policies excluding blacks from well-paid jobs; differences in educational attainment could be explained by legally segregated schools; etc. The 1950s and 1960s, however, witnessed a shift in how individuals, groups, and nation-states used race in everyday life and social systems. In the West, the confluence of the Civil Rights Movement, increasing immigration, the fall of colonialism abroad, and the economic rise of developing nations coincided with the precipitous decline in overtly racist attitudes, as measured by representative opinion surveys. As racial prejudice declined (unevenly) in the United States and the world (Bobo et al., 1997), theories arose to explain why racism, racial discrimination, and racial inequality persisted, emerged, or changed form in some places more than others.

This moment may be characterized as the start of the second (and contemporary) phase in the sociological study of racism. It has witnessed the (re)emergence of once-ignored critical and structural analyses of racism (à la DuBois) as well as manifold new theories to account for the subtlety of present-day racism. These theories often focus on group-level processes and social structures as opposed to, or in interaction with, the individual. For example, whereas earlier scholars defined racism as primarily an individual problem of overt hostility that could be diminished through interracial interaction (e.g., Allport, 1954), later sociologists viewed racism as fundamentally rooted in political, economic, and/or status resource competition (e.g., Blalock, 1967; Blumer, 1958); under these conditions, intergroup contact could exacerbate the perceived group threat that, in this view, drives racial prejudice and discrimination (cf Nagel, 1995). Building on this latter perspective, other scholars have examined the intersections of racism with colonialism (e.g., Blauner, 1969), class conflict (e.g., Bonacich, 1972), and gender (e.g., Collins, 1990). In the 1980s and 1990s, various theories of 'new racisms' (see Section New Racisms) and implicit biases (see Section Implicit Bias) emerged, suggesting that racism itself has transformed into more covert forms. Sociologists have also elaborated theories of institutional racism (see Section Institutional Racism), exploring how racist ideologies and discriminatory practices have become embedded in taken-for-granted laws, policies, and norms that systematically (dis)advantage certain groups. And since the turn of the century, social scientists have turned attention to the social processes whereby race, racism, and racial inequalities are constructed and challenged at micro-, meso-, and macro-levels (see Sections Racism as a Social Process and Responses to Racism).

Contemporary Definitions of and Approaches to Racism

Contemporary approaches to racism center on explaining the well-documented persistence of racial inequality and racial discrimination in an era of declining overtly racist attitudes.

Without explicit ideologies of racial domination as a direct cause, how can we explain persistent racial inequalities in criminal sentencing, health, and wealth; persistent rejection of policies meant to alleviate racial inequalities; and persistent racial discrimination in hiring, credit markets, and housing (Bobo et al., 1997; Fiske, 2010; Massey and Denton, 1993; Pager and Shepherd, 2008)?

Various theories have arisen to account for this paradox. Some scholars point to the (alleged) cultural deficiencies of people of color – in particular, inner city blacks. These theories often acknowledge the history of racism in shaping contemporary inner city black culture but argue that subordinates' cultural behaviors are at least one immediate cause of continuing racial inequality (e.g., Moynihan et al., 1967; Patterson, 1998). Instead of focusing on the cultural problems of historically oppressed groups, other scholars have attempted to explain persistent inequality by showing how racism endures today – if not so much within individuals, then at least within institutions and organizations, and if not so much as explicit attitudes, then at least as implicit or covert biases. While some social scientists claim this 'broadening' of the concept of racism serves a 'political agenda' (van den Berghe, 2001: p. 12721), most consider it a necessity – offering an empirically and socially useful toolkit of approaches for understanding the durability of racial inequality in the twenty-first century. We now turn to these approaches.

New Racisms

One major line of work in the contemporary sociology of racism examines whether the observed decline in racist attitudes on opinion surveys represents an actual decline in racism or merely a decline in the social acceptability of expressing such attitudes; perhaps some individuals consciously hold racist attitudes but withhold them when surveyed. Since the 1970s, social scientists have developed various techniques – from more subtle questions to new forms of discourse analysis – to alleviate respondents' hesitancy to report socially undesirable attitudes and to draw out the deeper meanings behind ambiguous or contradictory responses.

Using these techniques, sociologists have uncovered new forms of racism that are expressed not in avowed racist attitudes but rather in contextually specific moral and symbolic principles that stereotype subordinated racial groups as undeserving and thereby justify existing racial inequalities. For example, surveys repeatedly show that many whites support racial equality in principle but resist policies to implement it (e.g., affirmative action and reparations). Kinder and Sears (1981) attribute this principle–implementation gap to 'symbolic racism', which merges a genuine belief in the universalistic principles of Western liberal democracy with stereotypes and moral resentments (rooted in childhood socialization) toward 'irresponsible' blacks. Bobo et al.'s (1997) concept of 'laissez-faire racism' also highlights persistent antiblack (and antinative) stereotyping and a tendency to blame blacks (and other minorities) for their social problems despite increased support for racial equality in principle. Unlike symbolic racism, however, they argue that (white) opposition to racial equality policies is rooted in perceived racial group threat (Blumer, 1958), which is "triggered when the dominant group's sense of entitlement to resources and privileges appears threatened by subordinate group gains or aspirations" (Denis, 2012: p. 456). Similarly, 'colorblind racism' refers to a set of frames, styles, and scripts that are used to explain and justify the racial status quo without sounding racist (Bonilla-Silva, 2010). (For additional variations on the new racism theme, see Quillian, 2006.)

Despite this outpouring of research and theorizing, some critics (e.g., Sniderman and Carmines, 1997) argue that the problem is not 'hidden' racist sentiments or cultural stereotypes; rather, many whites reject policies such as affirmative action because of principled opposition to government intervention. Yet, regardless of whether racism or 'political principle' directly motivates opposition to policies intended to advance racial equality, such opposition effectively replicates racial inequality (see Section Institutional Racism). In short, although the observed decline in overt racist attitudes shows that racism is socially unacceptable in much of contemporary society, the extent to which individuals still hold racial stereotypes, prejudices, or ideologies – and the precise form(s) these take – remains contested.

Implicit Bias

Another explanation for persistent racial discrimination and inequality despite the decline in overt racist attitudes can be found in the growing literature on implicit bias. An implicit bias is an unconsciously triggered belief in the inferiority of, or negative attitude toward, a group(s). The assumption, drawing from recent findings in cognitive psychology and its nascent interaction with cultural sociology, is that implicit biases can impact expectations and actions. As such, unconscious negative beliefs and feelings about racial groups may not appear on a survey but may be revealed in everyday interpersonal interactions at work, at school, or on the street.

Relying on psychological experiments, and an Implicit Association Test (IAT), scores of studies have considered the association between race and various attributes and judgments, revealing that, on average, individuals more readily associate positive attributes and stereotypes with whites than with other races, particularly blacks (Banaji and Greenwald, 2013). Furthermore, these biases predict distinct (especially spontaneous, nonverbal) behaviors and emotions that could affect interpersonal interaction (Lane et al., 2007). More recent work has shown that minorities (not just whites) sometimes hold implicit biases against their own group despite articulating explicit beliefs in racial equality (Livingston, 2002). The implicit antiblack biases of many blacks, for example, may be interpreted as a form of internalized racism, in which members of a subordinated racial group accept the negative stereotypes and attitudes toward their group. While inquiry into internalized racism has been criticized as victim blaming and racially essentialist, Pyke (2010) argues that this work is critical to understanding the mechanisms whereby racial inequality is reproduced and can even be empowering, as it demystifies an insidious element of white privilege and systems of power.

Research on implicit biases has its limitations. Some scholars have questioned the reliability and validity of the IAT. Others suggest that laboratory experiments on nonrepresentative participants may not generalize to real-world contexts

where behavioral and attitudinal dynamics may differ. Moreover, studies of implicit biases would benefit from deeper engagement with macrosociological research on the media and other institutions that could shape implicit attitudes, and new institutional theories that view contemporary workplaces as contexts in which ostensibly neutral scripts and procedures could moderate autocognitive processes.

Racism as a Social Process

The upheavals of the post–World War II period made plain the mutability of racial classifications and racism. While research on new racisms and implicit biases documents the changing content, character, and implications of new forms of racism, another line of scholarship seeks to identify and explain the social processes and actors that account for change (and persistence) in racism. Within this broad, process-based approach there are two major lines of research – one that considers racism in historical context, focusing on general mechanisms that account for macrolevel changes over time, and another that considers microsocial processes that operate in interpersonal interaction.

The term 'racialization' tends to designate theories of racism as a macrolevel historical process. Racialization is the process of constructing racial meaning, including the creation of racial categories and the signification of these categories in relation to people, objects, and ideas (Murji and Solomos, 2005; Satzewich, 2011). Racialization's focus on the constructedness of racial meaning shifts analytic attention away from the content of racism and toward the individuals, groups, and social forces constructing and signifying race. The increasing focus on racialization, as a process, is part of the growing sociological focus on uncovering general social processes and mechanisms (Lamont et al., 2014). Indeed, research on racialization has revealed fundamental processes and mechanisms that contribute to the persistence of system-wide racial inequalities – from cognitive classification processes that allow humans to identify and categorize others (Massey, 2007) to nation-state bureaucracies and methods (such as the census or standardized education) that categorize, define, and distribute resources to groups at the macrolevel (Omi and Winant, 1994; Wimmer, 2013). One benefit to the study of racism as a social process is attention to the fluidity of racial categorizations and their embeddedness in power relations (e.g., Saperstein and Penner, 2012). One difficulty of macrolevel approaches is their tendency to present 'grand' narratives that abstract or totalize the way race operates. We may lose explanatory precision on the nuanced operation of race in particular contexts, institutions, and organizations at particular times and places (see Hall, 1980 on 'historically specific racisms').

Another process-oriented approach considers how microlevel social processes maintain and change racial categories, stereotypes, attitudes, and beliefs. Drawing on Goffman and the symbolic interactionist tradition, some scholars explore how racial identities and statuses are embodied and enacted in interpersonal interaction through distinctive habits, styles, scripts, and modes of self-presentation that inadvertently reproduce racial inequalities (e.g., Anderson, 2000). More broadly, Schwalbe et al. (2000) have identified generic processes in face-to-face interaction and meaning-making pursuits – othering, subordinate adaptation, boundary maintenance, and emotion management – that create and reproduce inequalities, including racial inequalities. Similarly, cultural sociologists examine how symbolic boundaries are constructed against racialized others, often translating into social, spatial, and temporal boundaries at the mesolevel (e.g., neighborhood) that perpetuate racial segregation (e.g., Small, 2004). In a small-town ethnography, Denis (2015) illustrates how the mutually reinforcing processes of subtyping, ideology-based homophily, and political avoidance norms enable many white Canadians to maintain laissez-faire racist attitudes, despite frequent positive contact with indigenous peoples. In social psychology, status characteristics theory considers how status characteristics (e.g., race, gender, and age) shape the influence and prestige of individuals in small group settings; it highlights the role of self-fulfilling prophecies, as lower status members are viewed and treated in terms of, and may even conform to, the stereotypical low expectations of their group (Fiske, 2010). While setting aside the question of why 'race' initially became a salient marker, all of these microlevel studies offer insights into how racial inequalities and stereotypes are reproduced in concrete settings. Much of this work also considers the everyday responses of historically marginalized groups to racism, and how these responses perpetuate or challenge the racial status quo (see Section Responses to Racism).

Institutional Racism

One major departure from past scholarship is contemporary sociology's shift from locating racism in individual beliefs and attitudes to considering it as primarily a phenomenon of higher-level entities, such as social processes, social forces, and institutions. Rather than explaining racial inequality via individual-level racism (conscious or otherwise), theories of institutional racism give analytic primacy to the taken-for-granted policies, practices, and norms of organizations, systems, and structures.

First defined by Stokely and Hamilton (1967), institutional racism refers to particular and general instances of racial discrimination, inequality, exploitation, and domination in organizational or institutional contexts, such as the labor market or the nation-state. While institutional racism can be overt (e.g., a firm with a formal policy of excluding applicants of a particular race), it is more often used to explain cases of disparate impact, where organizations or societies distribute more resources to one group than another without overtly racist intent (e.g., a firm with an informal policy of excluding applicants from a low income, minority neighborhood due to its reputation for gangs). The rules, processes, and opportunity structures that enable such disparate impacts are what constitute institutional racism (and variants such as 'structural racism', 'systemic racism', etc.).

The literature on institutional racism is diffuse but coheres on the point that macrostructural processes, as opposed to individual acts, provide more meaningful explanations of contemporary racial inequality. Bonilla-Silva's (1997) structural conception of racialized social systems' emphasizes how political, economic, and social arrangements are structured by racial hierarchy and supported by colorblind ideology.

Similarly, Feagin (2010) analyzes the institutional spaces, societal norms, and 'white racial frames' that can be traced back to sixteenth century slavery and imperialism, concluding that American society is fundamentally racist. Another variant is the (internal) colonialism model, in which one racialized group imposes its ideas, systems, and lifestyles on another via policies (such as Canada's Indian Act) and practices (such as the Indian residential school system) that are justified by the former group's presumed superiority (Frideres and Gadacz, 2008). These conceptions of institutional racism and systemic racial/colonial oppression are similar to macrohistorical process-centered conceptions in their totalizing nature (see Section Racism as a Social Process); they suggest that race and racism operate on various levels and in various corners of the social system, which is 'racialized' as such. They have been criticized for downplaying diversity within both dominant and subordinate groups, whose life chances vary considerably by class, gender, and other crosscutting categories (Satzewich, 2011).

Alternative approaches to institutional racism are less totalizing. Audit studies, for example, seek to measure racial discrimination at the firm or organization level by sending equally qualified resumes or research participants to employers, creditors, landlords, and the like. These studies have found that racial discrimination still exists in hiring, credit, and housing markets (Quillian, 2006; Pager and Shepherd, 2008). While audit studies suffer from uncertain generalizability (case studies do not tell us about all firms) and cannot specify the mechanisms that account for racial discrimination (e.g., is it due to employers' covert or implicit biases or uncontrolled differences between research participants?), they do enable us to identify the precise levels of discrimination in specific firm contexts. Moreover, when coupled with complementary evidence about labor market processes, such as Kirshenman and Neckerman's (1991) analysis of employers' statistical discrimination, we gain insight on important avenues through which racial discrimination and inequality persist.

Still other approaches focus on state and nonstate institutions – such as the criminal justice and education systems and the media – assessing the rules, regulations, norms, laws, discourses, and procedures that account for pernicious and racialized differential treatment and outcomes. While an institutional racism perspective highlights the general systems and processes that maintain racial inequality, one criticism is that it dilutes the meaning of racism (Miles, 1989) and attaches moral condemnation where it is undue (van den Berghe, 2001). Is it fair, or even accurate, to label an institution, organization, or society 'racist' if such intent is lacking? On the other hand, Allport (1954: p. 12) reminds us, "The unpleasant flavor of a word should not mislead us into believing that it stands only for a value-judgment"; the point of these theories is not to blame the innocent but to improve understanding so that the policies, practices, and ideas that perpetuate racial inequality can be identified and dismantled. Conversely, institutional racism's focus on the extraindividual might obscure and absolve the role of individual actors in maintaining racism.

Responses to Racism

While the contemporary sociology of racism focuses mostly on the processes, people, and organizations that perpetuate racism, increasing attention is being paid to those who experience racism. The literature on internalized racism (see Section Implicit Bias) illuminates how the targets of racism may accept dominant racist ideologies and engage in habitual interactions that perpetuate their own subordination. Yet, a growing literature also interrogates the multiple experiences, knowledges, and practices of subordinated racial groups. Increasingly, studies of 'everyday racism' (Essed, 1991) and responses to stigmatization (Lamont and Mizrachi, 2012) are providing insights into the effects of, responses to, and strategies for combating various forms of racism.

In one groundbreaking study, Essed (1991) examined how black women in the Netherlands and California navigate everyday racism, showing how it covertly shapes their daily life routines but also illuminating their diverse and often subtle forms of resistance. She argues that these women's narratives are important knowledge *ipso facto* because their unique social positions and experiences (raced, classed, gendered, etc.) offer unique perspectives on the inner workings of oppression and resistance (cf Collins, 1990); from a sociology of knowledge standpoint, subordinated groups' knowledge serves as a corrective to the elite, Western knowledge systems of dominant social scientists (Feagin, 2010). Building on this work, Lamont and Mizrachi (2012) systematically compare the responses of stigmatized groups in Brazil, Israel, and the United States, arguing that the comparative study of responses illuminates how institutions, ideologies, and varying national contexts constrain and enable individual opportunities. Other scholars have assessed antiracist and multicultural policies and affirmative action and diversity programs at the state and organizational levels (e.g., Henry and Tator, 2010). Still others have examined the conditions under which racial minority groups and their allies mobilize in social movements for racial justice (e.g., McAdam, 1982).

This (increasingly comparative and cross-national) work on responses to racism has blossomed alongside a robust sociology of immigration literature, which highlights the agency of immigrants in shaping their incorporation into host societies (Alba and Nee, 2003). Different immigrant groups employ varying strategies, seize unique opportunities in the form of ethnic niches, and encounter different obstacles. Contemporary assimilation is variegated and the incorporation of new, often racialized, groups has proven anything but straightforward. Moreover, competition and tension between native-born minority groups and recent immigrants has highlighted racial discrimination between subordinate groups (Bobo and Hutchings, 1996; Waters et al., 2014). Indeed, the relative ease with which some immigrants (despite being racialized) have integrated into the mainstream economically and socially only underscores the uniquely pernicious nature of antiblack and anti-indigenous racism in North America (Dixon, 2006). Numerous studies, for instance, have documented the stigmatization and racism faced by upwardly mobile middle-class blacks (e.g., Feagin and Sikes, 1995) and indigenous peoples (e.g., Denis, 2012), as well as the creative strategies they use to cope with and respond to racism.

One criticism of these studies is that they appear to place the burden of overcoming racism on the shoulders of the targets of racism rather than on the actors and institutions that perpetuate it. Yet, until racism is eliminated, it is important to

investigate how the targets interpret and respond to it in order to understand how racism affects them, ensure their voices are heard, and develop more effective strategies to combat racism itself.

Challenges and Future Directions

'Racism' today is a dirty word; yet, significant racial inequalities in socioeconomic, health, and other outcomes persist; racial minorities around the world continue to report frequent experiences with racial discrimination; and racial flashpoints – from the recent backlash to First Nations protests in Canada, to the Trayvon Martin shooting and trial in the US, to ongoing European debates over immigration reform – abound. Our contemporary moment presents theoretical and methodological challenges. What is the most productive way to study race and racism in an era when openly espousing racist views is anathema?

Rather than simply studying overt categorical hostility or explicitly racist laws and their effects, some sociologists have developed more nuanced understandings of what racism means, how it operates, and how it relates to racial inequality. These concepts include the multidimensional 'new racisms' (laissez-faire, colorblind, etc.), systemic and institutional racism, and even unconscious or implicit racism, which are often studied through experiments, audit studies, critical discourse analysis, and other innovative techniques. Other scholars eschew consideration of the proper definition of racism and instead examine the processes that construct racial categories, responses to racism by historically marginalized groups, and the (deliberate) strategies and (taken-for-granted) routines that reproduce (or challenge) racial inequalities.

From a pluralistic and pragmatic standpoint, we urge continued attention on all fronts. Greater interchange between these literatures should also lead to creativity and innovation as scholars seek to bridge levels of analysis and make sense of seemingly contradictory themes and trends. Although some scholars decry the conceptual broadening of 'racism', the sociology of racism is best served by this pluralism. As Western countries incorporate more 'nonwhite' immigrants and racial boundaries collapse, transform, and (re)emerge, scholars must use multivalent approaches to gain a comprehensive understanding of the nature, causes, and consequences of contemporary racism, racial discrimination, and racial inequality.

See also: Crime, Immigration, and Ethnicity; Discrimination: Racial; Ethnicity, Sociology of; Inequality, Social; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Race Identity; Race and Racism in the Twenty-First Century; Racial Relations; Racism, History of; Racism: Social Psychological Perspectives; Residential Segregation in the United States; Underclass.

Bibliography

Alba, R.D., Nee, V., 2003. Remaking the American Mainstream: Assimilation and Contemporary Immigration. Harvard University Press, Cambridge, MA.

Allport, G., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.

Anderson, E., 2000. Code of the Street: Decency, Violence, and the Moral Life of the Inner City. Norton, New York.

Banaji, M.R., Greenwald, A.G., 2013. Blindspot: The Hidden Biases of Good People. Delacorte Press, New York.

van den Berghe, P.L., 2001. Sociology of Racism. International Encyclopedia of Social and Behavioral Sciences. Elsevier.

Blalock, H.M., 1967. Toward a Theory of Minority-Group Relations. Wiley, Somerset, NJ.

Blauner, R., 1969. Internal colonialism and ghetto revolt. Social Problems 16, 393–408.

Blumer, H., 1958. Race prejudice as a sense of group position. The Pacific Sociological Review 1 (1), 3–7.

Bobo, L., Hutchings, V.L., 1996. Perceptions of racial group competition: extending Blumer's theory of group position to a multiracial social context. American Sociological Review 61, 951–972.

Bobo, L., Kluegel, J., Smith, R.A., 1997. Laissez-faire racism: the crystallization of a 'kinder, gentler' anti-black ideology. In: Martin, J.K., Tuch, S.A. (Eds.), Racial Attitudes in the 1990s. Greenwood, Westport, CT (Chapter 2).

Bonacich, E., 1972. A theory of ethnic antagonism: the split labor market. American Sociological Review 37, 547–559.

Bonilla-Silva, E., 1997. Rethinking racism: toward a structural interpretation. American Sociological Review 62 (3), 465–480.

Bonilla-Silva, E., 2010. Racism without Racists: Color-blind Racism and Racial Inequality in Contemporary America, third ed. Rowman & Littlefield, New York.

Collins, P.H., 1990. Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge, New York.

Cornell, S., Hartmann, D., 2006. Ethnicity and Race: Making Identities in a Changing World, second ed. Pine Forge Press, Thousand Oaks, CA.

Denis, J.S., 2012. Transforming meanings and group positions: Anishinaabe-white relations in Northwestern Ontario, Canada. Ethnic and Racial Studies 35 (3), 453–470.

Denis, J.S., 2015. Contact theory in a small-town settler-colonial context: The reproduction of prejudice in indigenous-white Canadian relations, American Sociological Review.

Dixon, J.C., 2006. The ties that bind and those that don't: toward reconciling group threat and contact theories of prejudice. Social Forces 84 (4), 2179–2204.

Essed, P., 1991. Understanding Everyday Racism: An Interdisciplinary Theory. Sage, New York.

Feagin, J.R., 2010. The White Racial Frame: Centuries of Racial Framing and Counter-Framing. Routledge, New York.

Feagin, J.R., Sikes, M.P., 1995. Living with Racism: The Black Middle-Class Experience. Beacon Press, Boston, MA.

Fiske, S.T., 2010. Interpersonal stratification: status, power, and subordination. In: Fiske, S.T., Gilbert, D.T., Lindzey, D. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, New York, pp. 941–982.

Frideres, J., Gadacz, R.R., 2008. Aboriginal Peoples in Canada, eighth ed. Pearson, Don Mills, ON.

Hall, S., 1980. Race, articulation, and societies structured in dominance. In: UNESCO (Ed.), Sociological Theories: Race and Colonialism. UNESCO, Paris, pp. 305–345.

Henry, F., Tator, C., 2010. The Colour of Democracy: Racism in Canadian Society, fourth ed. Thomson-Nelson, Scarborough, ON.

Kinder, D.R., Sears, D.O., 1981. Prejudice and politics: symbolic racism versus racial threats to the good life. Journal of Personality and Social Psychology 40 (3), 414–431.

Kirshenman, J., Neckerman, K.M., 1991. 'We'd love to hire them, but…': the meaning of race for employers. In: Jencks, C., Peterson, P.E. (Eds.), The Urban Underclass. Brookings Institution, Washington, DC, pp. 203–234.

Lamont, M., Beljean, S., Clair, M., 2014. What is missing?: cultural processes and causal pathways to inequality. Socioeconomic Review. http://dx.doi.org/10.1093/ser/mwu011.

Lamont, M., Mizrachi, N. (Eds.), 2012. Responses to Stigmatization in Comparative Perspective. Routledge, New York.

Lane, K.A., Banaji, M.R., Nosek, B.A., Greenwald, A.G., 2007. Understanding and using the implicit association test IV: what we know (so far). In: Wittenbrink, B., Schwarz, N.S. (Eds.), Implicit Measures of Attitudes: Procedures and Controversies. Guilford, New York, pp. 59–102.

Livingston, R.W., 2002. The role of perceived negativity in the moderation of African; Americans' implicit and explicit racial attitudes. Journal of Experimental Social Psychology 38, 405–413.

Massey, D.S., 2007. Categorically Unequal: The American Stratification System. Sage, New York.

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

McAdam, D., 1982. Political Process and the Development of Black Insurgency, 1930–1970. University of Chicago Press, Chicago, IL.

Miles, R., 1989. Racism. Routledge, London.
Moynihan, D.P., Rainwater, L., Yancey, W.L., 1967. The Negro Family: The Case for National Action. MIT Press, Cambridge, MA.
Murji, K., Solomos, J., 2005. Racialization: Studies in Theory and Practice. Oxford University Press, New York.
Nagel, J., 1995. Resource competition theories of ethnicity. American Behavioral Scientist 38, 442–458.
Omi, M., Winant, H., 1994. Racial Formation in the United States: From the 1960s to the 1990s, second ed. Routledge, New York.
Pager, D., Shepherd, H., 2008. The sociology of discrimination: racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology 34, 181–209.
Patterson, O., 1998. Rituals of Blood: Consequences of Slavery in Two American Centuries. Basic Books, New York.
Pyke, K.D., 2010. What is internalized racial oppression and why don't we study it? Sociological Perspectives 53 (4), 551–572.
Quillian, L., 2006. New approaches to understanding racial prejudice and discrimination. Annual Review of Sociology 32, 299–328.
Saperstein, A., Penner, A.M., 2012. Racial fluidity and inequality in the United States. American Journal of Sociology 118 (3), 676–727.
Satzewich, V., 2011. Racism in Canada. Oxford University Press, Don Mills, ON.
Schwalbe, M., Godwin, S., Holden, D., Schrock, D., Thompson, S., Wolkomir, M., 2000. Generic processes in the reproduction of inequality: an interactionist analysis. Social Forces 79 (2), 419–452.
Small, M., 2004. Villa Victoria: The Transformation of Social Capital in a Boston Barrio. University of Chicago Press, Chicago, IL.
Sniderman, P., Carmines, E.G., 1997. Reaching beyond Race. Harvard University Press, Cambridge, MA.
Stokely, C., Hamilton, C.V., 1967. Black Power: The Politics of Liberation in America. Vintage Books, New York, pp. 2–6.
Waters, M.C., Kasinitz, P., Asad, A.L., 2014. Immigrants and African Americans. Annual Review of Sociology 40 (1).
Wilson, W.J., 1999. The Bridge over the Racial Divide: Rising Inequality and Coalition Politics. University of California Press, Berkeley, CA.
Wimmer, A., 2013. Ethnic Boundary Making: Institutions, Power, Networks. Oxford University Press, New York.

Racism: Social Psychological Perspectives

Martha Augoustinos, School of Psychology, University of Adelaide, Adelaide, SA, Australia
Danielle Every, Appleton Institute, CQUniversity, Wayville, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Racism has been a core topic in social psychology since the 1930s. Central to most definitions of racism is the belief in a biological hierarchy between different social groups based on perceived racial differences. As a complex social issue multiple social psychological perspectives have been advanced to understand and theorize beliefs, behavior, and social practices that sustain racial inequality ranging from the individual to the social level of explanation. These include personality theories; social cognition models; realistic group conflict; social identity theory; critical discursive approaches, studies of White privilege; and embodied racism. Although these perspectives are frequently argued to be inconsistent with each other, recent attempts at integration are providing richer accounts of this phenomenon.

Defining Racism

Although many definitions of racism have been proposed, an all-inclusive definition has yet to be agreed upon, particularly as researchers have identified a variety of 'racisms' (Miles and Brown, 2003; Richards, 1997). Central to most definitions of racism is the belief in a biological hierarchy between different social groups based on race, and the associated practices that maintain and reproduce social inequalities between groups based on such beliefs. The belief that different racial groups reflect a natural evolutionary hierarchy, at the top of which are European (White) people was central to scientific racism which was widely promoted as an ideology between 1850 and 1910. During this period, European imperialist expansion and colonial rule over Indigenous peoples created the ideal conditions for the proliferation of such Social Darwinist beliefs (Richards, 1997).

The concept of race is entrenched in both popular usage and scientific discourse as a taken-for-granted, essentialist category that categorizes people into groups based on assumptions that surface phenotypic characteristics such as skin color reflect deeper genotypic features. Despite its ubiquitous taken-for-granted usage, geneticists and biologists discredited the validity of race as a scientific category as long ago as the 1930s (Richards, 1997), and more recently via the mapping of the human genome (McCann-Mortimer et al., 2004). Nonetheless the concept of race continues to be used uncritically both in the scientific community and in everyday discourse as a 'natural' kind variable in ways that reinforce the commonplace view that it is a biological and genetic reality that reflects real differences between groups.

Although the notion of a biological hierarchy between groups is generally eschewed today and indeed is associated with blatant forms of racism, it has been replaced with beliefs in a cultural hierarchy between groups where the dominant group's social values, norms, and practices are represented as superior to those of less dominant groups. As we will discuss further below, this has come to be known as the 'new racism.' Group differences therefore, whether biological or cultural, continue to operate as socially meaningful cues by which to categorize and differentiate people. That is, 'race' has become a reified and objectified social representation through which group differences come to be understood and explained (Moscovici, 1988).

Prejudice versus Racism

There has been a tendency within psychology to use the terms prejudice and racism interchangeably. Jones (1997), among others, argues that racism is distinct from prejudice. Prejudice has commonly been defined as negative attitudes and behavior toward a social group and its members. Prejudice is typically regarded as an individual phenomenon, whereas racism is a broader construct that links such individual beliefs and behavior to broader social and institutional norms and practices that systematically disadvantage particular groups. The second important difference between prejudice and racism relates to the role of power. At an individual level, a person can display prejudice, but this in itself does not necessarily constitute racism. Central to racism is the ability of dominant groups to systematically exercise *power* over outgroups. If we define racism without reference to power differentials between groups, it is clear that anyone can engage in ingroup preference and outgroup bias. 'Everybody is racist' is a claim that is often used to counter accusations of racism (Hage, 1998). Importantly, the power one group has over another transforms prejudice into racism and links individual prejudice with broader social practices (Jones, 1997).

Racism, practiced at a structural and cultural level, maintains and reproduces the power differentials between groups in the social system. Racism practiced at this broad societal level has been referred to as institutional and cultural racism (Jones, 1997). Institutional racism refers to the institutional policies and practices that are put in place to protect and legitimate the advantages and power one group has over another. Institutional racism can be overt or covert, intentional or unintentional, but the consequences are that racist outcomes are achieved and reproduced. Cultural racism occurs when the dominant group defines the norms, values, and standards in a particular culture. These mainstream ideals permeate all aspects of the social system and are often fundamentally antagonistic with those embraced by particular minority

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24082-8

groups. To participate in society, minority groups often have to surrender their own cultural heritage and adopt those of the dominant group (e.g., the White majority).

Although prejudice has been condemned within psychology as negative and pernicious, it has also been criticized for depoliticizing the issue of racial inequality. Because the concept of prejudice is primarily seen as located within the psychology of the individual, it fails to recognize the wider historical, social, and institutional structures that support racial inequality. Because of this narrowness, the concept of prejudice is often challenged as actually part of the problem of racial inequality – by making it an individual pathology rather than a political and social reality. As many social theorists have argued, this has had the net effect of obscuring the political and ideological dimensions of prejudice. Racism can persist in institutional structures and policies in the absence of prejudice at the individual level (Henriques, 1984).

New Racism

Over the past 50 years, social psychologists and social scientists more broadly have argued that contemporary racism has become less about beliefs in a biological hierarchy between groups, and increasingly about beliefs in the cultural superiority of a dominant group's values, norms, and practices (Barker, 1981). Survey studies consistently demonstrate that blunt, hostile, segregationist, and White supremacist beliefs are less openly acceptable to White majority group members in Western liberal democracies. However, racial inequality continues to exist. To explain this, a distinction is therefore commonly made between 'old-fashioned racism' and 'modern' (McConahay, 1986) or 'symbolic racism' (Kinder and Sears, 1981), which in contrast, is subtle, covert, and paradoxically, endorses egalitarianism. Modern racism rejects racial segregation and notions of biological supremacy, and is instead, based on feelings that certain social groups transgress important social values such as the work ethic, individualism, self-reliance, and self-discipline: values that are embodied in the Protestant ethic. Symbolic or modern racism justifies and legitimates social inequities based on moral feelings that certain groups violate such traditional values.

Gaertner and Dovidio (1986) have also proposed models of racism that address the changing and complex nature of contemporary beliefs about race. Their 'ambivalent racism' and 'aversive racism' models both posit that contemporary racial attitudes have become complex, contradictory, and multidimensional. In the ambivalent racism model, pro-Black and anti-Black sentiments are seen to coexist within the person and to reflect different value structures held by the individual. Pro-Black attitudes reflect humanitarian and egalitarian values that emphasize equality and social justice, whereas anti-Black attitudes reflect individualism, the Protestant ethic, hard work, individual achievement, and self-reliance. Similarly, the aversive racism model emphasizes the coexistence of a contradictory complex of attitudes: on the one hand, liberal egalitarian principles of justice and equality; and on the other, a residue set of negative feelings and beliefs about particular groups that are learned early in life, and which are difficult to completely eradicate. Gaertner and Dovidio (1986: p. 63) describe these negative feelings as "discomfort, uneasiness, disgust, and sometimes fear, which tend to motivate avoidance rather than intentionally destructive behaviors." In both of these accounts, individuals strive to maintain a nonprejudiced image, both to themselves and to others, and struggle unconsciously to resolve the internal psychological ambivalence that is produced by maintaining a contradictory set of attitudes and beliefs. By justifying and legitimating social inequalities between groups on the basis of factors other than race, members of dominant groups can avoid attributions of racism and thus maintain and protect a nonprejudiced self-image. Indeed the psychological and social motivation to dodge a prejudiced identity is a common thread in contemporary theorizing on racism.

Contemporary racism, therefore, is seen as more insidious and difficult to identify because of its subtle and covert nature. This has led to the proliferation of implicit measures to identify and measure this more subtle racial bias (Greenwald et al., 2003). For example, the Implicit Association Test is a response latency measure using subliminal primes to test the strength of association between social categories (e.g., 'Black' or 'White') and positive and negative trait characteristics. Slower responses to stereotype inconsistent associations (Black + positive traits and White + negative traits) than to stereotype consistent associations (Black + negative traits and White + positive traits) is treated as evidence for an implicit bias or prejudice toward Blacks. Indeed the distinction between implicit and explicit racial bias is now so ubiquitous in social psychology that it is sometimes (erroneously) assumed that implicit measures reflect people's true or real attitudes whereas explicit measures merely reflect social desirability norms. It has been argued, however, that implicit measures do not tap racial attitudes or beliefs per se but deeply ingrained stereotypes strongly associated with particular groups. Devine's (1989) dissociation model of prejudice is consistent with this view and posits that stereotypes are more primitive cognitive structures learned early in life that can be automatically activated, whereas racial attitudes (prejudice) are learned later in life and can be either inconsistent or consistent with these stereotypes. The fact that negative stereotypes can be unconsciously activated even among people with low levels of explicit prejudice should not be taken as evidence that prejudice is an inevitable and natural cognitive tendency in everyone. As we will discuss below, the inevitability of prejudice perspective is associated with cognitive models of prejudice.

Theories of Racism

A variety of explanations for prejudice and racism have been advanced by social psychologists throughout the twentieth century. The prevalence of particular kinds of explanations has shifted during this time depending on wider historical and social factors and the dominance of specific paradigmatic frameworks within social psychology itself. Here we provide an overview of six current approaches to racism ranging from the individual to the social level of explanation: personality theories; social cognition models; realistic group conflict and social identity theory; critical discursive approaches, studies of White privilege; and embodied racism.

Personality Theories

Freudian psychodynamic accounts of prejudice were prevalent between 1930 and 1960. Prejudice was largely understood as a product of intrapsychic unconscious impulses primarily related to sexual and aggressive desires within the person. To reduce tension, negative emotions such as fear, anger, and disgust generated by these internal psychological conflicts are projected outward onto outgroups.

The most well known of these psychodynamic approaches is *The Authoritarian Personality* by Adorno et al. (1950). Published soon after the end of the Second World War, Adorno et al. were interested in explicating a theory that accounted for the widespread support for fascism as was seen in Nazi Germany. Adorno et al. (1950) argued that parent–child relationships with severe and punitive parental discipline produce children with an authoritarian personality characterized by a rigid adherence to conventional social values and mores, an unquestioning subservience to one's moral and social superiors, and a vigilance for, and hostile rejection of, those who violate conventional social values and mores. The F Scale was developed to measure levels of authoritarianism and was widely used as a personality measure. High levels of authoritarianism were found to be associated with all types of prejudice (racism, sexism, homophobia). Despite the widespread use of the F Scale, by the 1960s, the theory of the authoritarian personality was strongly criticized for its emphasis on internal psychological predispositions at the expense of social and cultural norms that tolerated prejudice and sanctioned institutionalized racism, for example, racial segregation in the United States and apartheid in South Africa (Pettigrew, 1958).

Interest in authoritarianism was revived in 1981 with Altemeyer's theory of right-wing authoritarianism (RWA). RWA is described as a rigid adherence to social conventions, submission to established authorities, and a strong rejection of outgroups who are perceived to be culturally and ethnically different. Unlike Adorno et al. whose work was heavily influenced by psychodynamic theory, Altemeyer theorized RWA as an individual personality characteristic that was predominantly shaped by social learning experiences. RWA has been found to be a good predictor of racial and ethnic prejudice in a variety of different settings, more so than the early authoritarianism scales.

The most recent personality approach to prejudice is that of 'social dominance orientation' or SDO (Sidanius and Pratto, 2001). SDO is purported to be a stable individual difference that refers to a person's level of support for group-based hierarchies in society such as racial/ethnic, gender, and socioeconomic hierarchies. Like RWA, SDO is strongly correlated with prejudice. SDO scores vary with gender (males score higher), personality and temperament, education, religion, and whether one is a member of a dominant or subordinate group. Although the concept of SDO is embedded within a wider social theory, it too has been criticized for reducing prejudice to a psychological trait rather than a social phenomenon that requires a social/structural explanation. Perhaps the most controversial aspect of SDO is the claim that group-based hierarchies and the legitimating beliefs that support them have an evolutionary basis.

Significant limitations have been identified with personality accounts of prejudice. Most notably is the issue of why certain groups rather than others become the targets for prejudice by authoritarians, those high on RWA and/or SDO. In addition, although such theories recognize that economic, historical, and social factors contribute to these predispositions, the potential interplay between individual psychology and social structural factors is rarely dealt with explicitly or integrated thoroughly into these models.

Social Cognitive Theories

Gordon Allport's seminal work, *The Nature of Prejudice* (1954), provides the foundational basis for social cognitive models of prejudice which have become dominant and influential in social psychology since the 1980s. Allport's definition of prejudice as "an antipathy based upon a faulty and inflexible generalization" (1954: p. 9), about a social group and its members emphasizes the prominent role that social categorization and stereotyping as perceptual–cognitive processes are given in social cognition models. According to these models, categorizing people into their respective group memberships (such as race, gender, age) is driven by our cognitive need to simplify the overwhelming amount of stimulus information we receive from our environment. This group-based or category-based perception is seen as distorting reality because people are not viewed as individuals in their own right but rather as prototypical group members. In turn, this leads to stereotyping, which recent social cognition research suggests can occur automatically and outside conscious awareness (Nosek et al., 2011). Stereotyping of course is just one step away from prejudice – literally prejudging someone based solely on their group membership. This inextricable relationship between categorization, stereotyping, and prejudice is central to social cognition models of prejudice and notwithstanding some of the qualifications that recent research has placed on this directional, and by implication, causative link between these three processes (e.g., Devine, 1989, see above), it is nonetheless the case that categorization in and of itself is seen as the cognitive basis for prejudice, driven primarily by our limited processing capacities.

Social cognition models have been criticized for normalizing prejudice and racism as inevitable products of our cognitive hard-wiring. Critics have also argued that by treating racial categories and racial categorization as natural rather than social and ideological constructs, social cognition models themselves reproduce racism in psychology (Hopkins et al., 1997).

Realistic Group Conflict and Social Identity Theory

Realistic group conflict theory and social identity theory are intergroup approaches to racism in social psychology that emphasize the role that relations of power and dominance between different social groups play in determining patterns of intergroup hostility. As the name suggests, realistic group conflict views intergroup hostility as arising from competition between social groups for economic, social, and cultural resources. Unlike personality theories that see racism and

prejudice as outcomes of internal psychological drives or differences in personality, in this approach conflict is viewed as emerging from 'real' group-based interests. The famous boys' camp field studies by Sherif et al. (1961) demonstrated how the creation of two competing groups was able to produce ingroup favoring and outgroup derogating attitudes and behavior between the two groups. When the social conditions were changed, however, and the two groups were required to cooperate to obtain resources or to complete valued tasks, intergroup hostility began to diminish.

Another series of famous studies, the minimal group experiments by Henri Tajfel and his colleagues formed the foundations upon which social Identity theory (SIT) was built (Tajfel and Turner, 1986). Many social psychologists have concluded erroneously from the minimal group experiments that the mere categorization of people into ingroups and outgroups is sufficient to trigger intergroup discrimination and prejudice. Although SIT stresses the psychological importance for groups to differentiate themselves positively from other groups this does not necessarily go hand-in-hand with ingroup enhancement and outgroup derogation though regrettably these are all too frequent occurrences. Groups can maintain a positive social identity without threatening the social identity of others. SIT posits that groups and their members strive to achieve some sort of differentiation from other groups, in ways that are shaped by the nature of the intergroup context and on dimensions of importance to them. Sometimes those dimensions of importance emphasize tolerance, generosity, and beneficence, but again all too often these dimensions emphasize superiority, dominance, and preserving ingroup privilege (Ellemers and Haslam, 2012).

Critical Discursive Research

Critical discursive research views racism as interactive and communicative and as located within the language practices and discourses of a society. This body of work emphasizes the ambivalent and contradictory nature of contemporary racism, but explicitly avoids making claims about the psychology of individual perceivers. Discursive studies analyze how people talk, discuss, and debate matters to do with 'race' and intergroup relations in both formal and informal settings (van den Berg et al., 2003).

It is through everyday language practices, both in formal and informal talk that relations of power, dominance, and exploitation become reproduced and legitimated. The analytic site for discursive research is how discursive resources and rhetorical arguments are put together to construct different social and 'racial' identities, and to provide accounts that legitimate these differences and identities as 'real' and 'natural.' Discursive studies locate these language practices or 'ways of talking' at a societal level, as products of a racist society rather than as individual psychological and/or cognitive products. The analytic site therefore is not the prejudiced or racist individual, but the discursive and linguistic resources that are available within an inequitable society (Wetherell and Potter, 1992). This approach has been able to identify how linguistic resources are combined in flexible and contradictory ways to reproduce and justify racist outcomes in modern liberal democracies. In some instances, existing relations of power, dominance, and privilege are maintained through overt racial ideology, but given the increasing opprobrium against the expression of such views, social inequalities are more commonly legitimated through the flexible and contradictory use of liberal egalitarian arguments that draw on principles such as freedom, individual rights, and equality. Discursive studies in several Western countries including Australia, Britain, Belgium, the Netherlands, New Zealand, and the United States have demonstrated how majority group members express negative and even hostile views of minorities by the use of self-sufficient rhetorical arguments premised on liberal and egalitarian values such as treating everybody equally.

Critical discursive research has demonstrated the varied ways the category of racism is itself highly contested in everyday life: that is, defining what is and what is not racist is far from being a value-free, neutral assessment. Rather, the category of racism itself is constructed flexibly and variably, and can be used to manage the moral accountability and identity of an individual or group. van Dijk (1992) documents the ubiquitous nature of denials of racism through the use of disclaimers such as "I'm not racist but … ." Contemporary race talk therefore, is strategically organized to deny prejudice and racism. By redrawing the boundaries of what may be legitimately defined as 'racist,' the category of racism may be used to position a person or group as 'not racist' by placing their own behavior and views outside of these boundaries.

These language practices are forms of power that are products of particular historical, hierarchical relationships between groups of people, in which some people have unjustly and unfoundedly claimed dominance over others. Understood as power relationships, racism shapes the lives of everyone within these hierarchies, both the oppressed and the oppressors. In this sense, 'race' is a form of categorization that reflects particular forms of power relations between groups of people, rather than reflecting the actual attributes (whether they be physical or behavioral) of any particular group of people.

Embodied Racism

While critical discursive research has been invaluable for elucidating the perpetuation and legitimation of racism, it has also been criticized for its privileging of discourse; for ignoring the materiality of oppression. Discursive work has tended to become heavily involved in identifying the rhetorical aspects of racism and oppression, but has been less concerned with the nexus between discourse, space, and place. Such a research focus has been taken up by Durrheim and Dixon (2005) in South Africa.

Durrheim and Dixon (2005) combine discourse analysis with ethnographic mapping to identify the ways in which discourses are embodied in people's use of public places such as beaches. They demonstrate that, even in post-Apartheid South Africa, there remains significant physical segregation on South Africa's beaches – Whites and Blacks use different areas of the beach, and use the beach at different times. These spatial practices are legitimated through discourses about 'appropriate beach behavior' with White beach-goers explaining that they leave the beach because Black beach-goers are too loud, create mess, and do not respect personal space. Their

research demonstrates the interdependence of discourse and embodied practices and how "'race relations' are constructed both in language and in located bodily practices, emphasising how people describe and account for the racialised features of social life that they participate in" (Durrheim and Dixon, 2005: p. 459). They argue that without such a combined approach, research on racism cannot capture the lived experience of racism, or of anti-racism. Their observations and interviews capture the ongoing nature of racism in everyday life, even where overt official structures maintaining inequality are dismantled.

Racism as White Privilege

While traditional research on racism focuses on attitudes and practices toward minorities, recent research on White privilege turns the gaze from minorities to the majority group. Thus there is a shift in focus – the gaze moves away from those who bear the brunt of racism, and toward the discourses and institutional practices of those who benefit from racism (Aveling, 2004). This body of work focuses on White identity construction (e.g., Carter, 1997). It also examines the ways in which whiteness is produced and reproduced in different social and cultural sites; and the implications of these constructions for intergroup relations and anti-racism (e.g., Hage, 1998). Whiteness studies is primarily concerned with how White people's identities are shaped by broader institutionalized forms of racism and brings to the fore both the benefits that White people accrue because of their privileged position in society and the responsibilities they have for addressing racism (Giroux, 1997).

Clearly the category 'White' or 'White people' is problematic as a way of referencing the dominant majority in Western liberal democracies as it fails to adequately reflect the ethnic and cultural heterogeneity of group members that may identify as 'White.' Nonetheless by marking 'whiteness' (not only as a category of skin color, but of cultural capital (Hage, 1998)), there is an attempt to make visible the unearned power and privilege that accrues to the dominant majority, especially those that have access to the highest cultural and social capital – in terms of appearance, ancestry, religion, socioeconomic status, education, and employment (Hage, 1998).

Broadly speaking, whiteness is "… the production and reproduction of dominance rather than subordination, normativity rather than marginality, and privilege rather than disadvantage." In this definition, whiteness is something that places White people in normative positions and grants White people unfair privileges. These positions and privileges are often invisible to White people, because of this normativity. Indeed, it is this normativity that gives whiteness its power (Frankenberg, 1993: p. 236).

Integrating Multiple Perspectives

This overview of social psychological approaches to racism always raises the difficult question of whether it is possible to integrate these different theories. This question has always been a bone of contention within psychology, as analytic frameworks differ significantly in their epistemological assumptions and orientations. It also raises the question of whether integration is desirable.

Duckitt (1992) argues that these multiple social psychological perspectives are not necessarily competing paradigms, but rather each is a valid response to different aspects of this social phenomenon: unconscious processes, personality, cognitions, social norms and linguistic practices, power, social structure, and intergroup relations.

Duckitt (1992) has proposed an integrative framework that identifies four primary causal processes of prejudice: internal psychological processes; social and intergroup dynamics; social transmission; and individual differences. He argues that each of these causal processes provides a partial but essential contribution to the explanation of prejudice: psychological processes build a human propensity for prejudice; social and intergroup dynamics elaborate this propensity into socially shared patterns of interaction; these patterns are socially transmitted throughout social groups; and individual differences in susceptibility to prejudice modify these social norms. Each theory is limited on its own as it focuses on and seeks to elaborate just one of these causal processes.

While Duckitt's integrative framework has considerable appeal more recent critical approaches that focus on racial discourse, power, whiteness, and embodiment are largely ignored in this model. Bringing all of these together we conclude that social psychology has conceptualized racism to be a normative, often invisible system of social practices, cognitions, emotions, and discourses that are perpetuated through all levels (individual, interpersonal, intergroup, institutional) that privilege one social group and disadvantage and marginalize other social groups. These practices can be overt assertions of biological difference, but in today's social and political climate, are more likely to be covert and implicit.

See also: Authoritarian Personality; Critical Psychology; Desegregation; Embodied Social Cognition; Ethnicity and Migration in Europe; Immigration: Social Psychological Aspects; Intergroup Relations; Levels of Analysis in Social Psychology; Social Categorization; Social Cognition; Social Dominance Orientation; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology; Xenophobia: Social Psychological Aspects.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper and Row, New York.

Allport, G.W., 1954. The Nature of Prejudice. Perseus Books, Cambridge, MA.

Altemeyer, B., 1981. Right-wing Authoritarianism. University of Manitoba Press, Winnipeg.

Aveling, N., 2004. Being the descendant of colonialists: White identity in context. Race, Ethnicity and Education 7, 57–71.

Barker, M., 1981. The New Racism. Junction, London.

van den Berg, H., Wetherell, M., Houtkoop, H., 2003. Analysing Race Talk: Multidisciplinary Perspectives on the Research Interview. Cambridge University Press, New York.

Carter, R., 1997. Is white a race? Expressions of white racial identity. In: Fine, M., Weis, L., Powell, L.C., Mun Wong, L. (Eds.), Off White: Readings on Race, Power, and Society. Routledge, New York, pp. 198–209.

Devine, P., 1989. Stereotypes and prejudice: their automatic and controlled components. Journal of Personality and Social Psychology 56 (1), 5–18.

Duckitt, J., 1992. Psychology and prejudice: an historical analysis and integrative framework. American Psychologist 47 (10), 1182–1193.
Durrheim, K., Dixon, J., 2005. Studying talk and embodied practices: toward a psychology of materiality of 'race' relations. Journal of Community and Applied Social Psychology 15, 446–460.
van Dijk, T.A., 1992. Discourse and the denial of racism. Discourse & Society 3, 87–118.
Ellemers, N., Haslam, S.A., 2012. Social identity theory. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (Eds.), Handbook of Theories of Social Psychology, vol. 2. Sage, London, pp. 379–399.
Frankenberg, R., 1993. White Women, Race Matters: The Social Construction of Whiteness. University of Minnesota Press, Minneapolis.
Gaertner, S.L., Dovidio, J.F., 1986. The aversive form of racism. In: Dovidio, J.F., Gaertner, S.L. (Eds.), Prejudice, Discrimination and Racism: Theory and Research. Academic Press, Orlando, FL, pp. 61–89.
Giroux, H.A., 1997. Racial politics and the pedagogy of whiteness. In: Hill, M. (Ed.), Whiteness: A Critical Reader. New York University Press, New York.
Greenwald, A., Nosek, B., Banaji, M., 2003. Understanding and using the implicit association test: an improved scoring algorithm. Journal of Personality and Social Psychology 85 (2), 197–216.
Hage, G., 1998. White Nation: Fantasies of White Supremacy in a Multicultural Society. Taylor & Francis, UK.
Henriques, J., 1984. Social psychology and the politics of racism. In: Hollway, W., Venn, C., Walkerdine, V., Henriques, J., Urwin, C. (Eds.), Changing the Subject: Psychology, Social Regulation and Subjectivity. Routledge, UK, pp. 60–89.
Hopkins, N., Reicher, S.D., Levine, M., 1997. On the parallels between social cognition and the 'new racism'. British Journal of Social Psychology 36, 305–330.
Jones, J.M., 1997. Prejudice and Racism, second ed. McGraw-Hill, New York.
Kinder, D.R., Sears, D.O., 1981. Prejudice and politics: symbolic racism versus racial threats to the good life. Journal of Personality and Social Psychology 40, 414–431.
McCann-Mortimer, P., Augoustinos, M., LeCouteur, A., 2004. 'Race' and the human genome project: constructions of scientific legitimacy. Discourse & Society 15 (4), 409–432.
McConahay, J.B., 1986. Modern racism, ambivalence, and the modern racism scale. In: Dovidio, J.F., Gaertner, S.L. (Eds.), Prejudice, Discrimination, and Racism. Academic Press, New York, pp. 91–125.
Miles, R., Brown, M., 2003. Racism. Routledge, UK.
Moscovici, S., 1988. Notes towards a description of social representations. European Journal of Social Psychology 18, 211–250.
Nosek, B., Hawkins, C., Frazier, R., 2011. Implicit social cognition: from measures to mechanisms. Trends in Cognition Science 15 (4), 152–159.
Pettigrew, T.F., 1958. Personality and sociocultural factors in intergroup attitudes: a cross-national comparison. Journal of Conflict Resolution 2, 29–42.
Richards, G., 1997. Race, Racism & Psychology: Towards a Reflexive History. Routledge, UK.
Sherif, M., Harvey, O.J., White, B.J., Hood, W.R., Sherif, C., 1961. Intergroup Conflict and Cooperation: The Robber's Cave Experiment. University of Oklahoma Press, Norman.
Sidanius, J., Pratto, F., 2001. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, UK.
Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup relations. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 7–24.
Wetherell, M., Potter, J., 1992. Mapping the Language of Racism: Discourse and the Legitimation of Exploitation. Harvester Wheatsheaf, Hemel Hempstead.

Radical Social Work

Hilde Bloemen, KHLeuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work's legacy of 'progressive politics' or radical traditions shifts according to changes in political, economic, and social environments. Working on social justice becomes an international matter as social problems become more complex, poverty increases, community life worsens. This demands that the nature of social work be clarified. Will this lead to further radicalization?

Radical social work from the 1970s evolved into critical social work, currently termed 'structural social work' and/or antioppressive practice. This article sketches a short history of radical social work practice and theory, focusing on Flanders and the Netherlands. A survey of social work educators and practice-oriented research illustrate where radical social work's core might lie.

Introduction

Increasing unemployment, migration, and bank crises, have characterized welfare states over the past decades. When searching for answers within a neoliberal and neocapitalist context, social work can play a maintenance or critical role (Dominelli, 2010). Given this history, does the practice of social work offer answers for a social justice society? Social work's international bodies suggest that the mission of social work is social change, solving difficult social problems and enhancing peoples' well-being (IFSW and IASSW, 2004). Does social work neglect its mission by identifying itself too much with psychotherapy? Sometimes social workers find their clients or participants difficult and marginalized. On whose side are they? Do they spontaneously choose the side of oppressed, disadvantaged, and vulnerable people? Do they recognize the barriers in peoples' lives or do they see people as impeding their own self-empowerment? Are organizations, workers, and policymakers held accountable? Is working on social justice, equality, and emancipation at the foreground of reclaiming the radical social work, a tradition?

This article develops another definition of social work by considering how radical social work is described in the literature and what its traditions are. An exploration of the existing vision of social work was conducted by asking questions of methodology tutors and social workers. Interviews and interactions with the focus groups were conducted in order to explore the meaning and values that tutors, social workers, service users and managers assigned to social work. These results ground my description of radical social work and how social work is practiced, and help social workers and tutors to specify the heart and identity of social work.

Social Work and Radical Social Work

Social work is anchored in society. The profession of social workers has provided professional answers to requests for help. Berteke Waaldijk (1996) argues that from the beginning of the professionalization of social work, social workers realized that the profession constantly needed to legitimize itself according to the rapidly changing contexts in which it operated. Consequently, social workers have to immerse themselves thoroughly in the broader contexts of the profession and appreciate such influences on their practice. Van Riet (2006) states that, if they fail to do so, practitioners would become victims similar to the people with whom they work when providing care and enhancing well-being. But what is social work about?

In describing social work, I look into its history to determine why it came into existence. "Social work is a fluid entity – it grows, develops and changes" (Thompson, 2005: 16). Besides its history, the objectives or purposes of social work, i.e., why it exists, needs clarification. The question, 'What is social work for?' is highly contested. Several authors have developed different typologies or classification systems for social work theory. This discussion is, however, limited to the difference between 'social work and social order' and 'social work and social change' (Thompson, 2005: 22). Payne (2005) describes the gap between liberal and critical visions. Professional social work is a classical analysis of social work as a discourse among three aspects of practice: social order, therapy, and transformation (Dominelli, 1997; Payne, 2006). The theoretical debate continues to be held from among these three visions of social work, linked to liberal, humanistic and critical visions of welfare and humanity (Howe, 2008). Critical and radical theories of social work practice stem from Marxist social theories. First, these theories are transformational and aim for social change. More specifically, they reject capitalist, neoliberal and rational approaches to managing the economy because of contradictions between those who own/manage the means of production and the workers. Radical and critical theories stress the structural elements of social problems and focus on inequality and oppression. These theories are emancipatory or committed to the liberation of peoples from the barriers imposed by the existing social order. Radical and critical practices try to *raise awareness* about social injustice, political action to enhance social change (Dominelli, 2002b). Raising awareness helps to battle cultural hegemony through which powerful people uphold the existing social order and take advantage of people's social and cultural convictions through the media and education (Payne, 2005).

However, radical social work has never been a dominant tendency in social work. According to Midgley (2001), from the beginning of social work practice, three (sometimes overlapping) approaches could be distinguished: the (traditional)

'remedial social work' which has been the approach most often used historically and has an individual focus; 'developmental social work' which rose as a reaction on remedial work and focuses on processes of economical and social development; and (progressive) 'activist social work,' better known as radical, structural, or antioppressive social work (Midgley, 2001). Radical work is hidden in the history of social work. Initially, the radical approaches criticized the traditional social work for its focus on individual methods of helping and individualized notions of personal problems (Leonard, 1975; Fook, 2002). In Flanders, this criticism appeared when social casework failed to promote individual development and the structural causes and effects of problems gave way to collective interventions through community work (Herman, 1996). Social workers reacted with alternative small social services agencies that integrated social care, education, or therapeutic consciousness and social action with political aspects (De Turck, 1978). This political social work was based on rights and developed into a rights discourse. During recent decades, this discourse has oscillated between the political climate and social work (Driessens and Geldof, 2009). The task of social work is to support oppressed people, groups, and communities by challenging systematic discrimination and inequality (Leonard, 1975; Corrigan and Leonard, 1978; Dominelli, 1988, 1996; Mullaly, 1997, 2001; Thompson, 2005; Verzelen, 2005; Driessens and Geldof, 2009).

Hence, international, and to some extent Flemish/Dutch, literature shows that radical social work emphasized the heart of social work. Its theories can be called radical because they describe how social inequality and oppression and dominant power relations can be changed through raising consciousness and social action. Various authors define radical or structural or critical social work in their own way, based on the radical tradition (Bailey and Brake, 1975; Galper, 1975; Throssell, 1975; Leonard, 1975; Corrigan and Leonard, 1978; Howe, 1994; Dominelli, 1988; Dominelli and McLeod, 1989; Dominelli, 1996, 2002a,b; Fook, 1996, 2002; Mullaly, 2001; Ferguson, 2008). An examination of these traditions makes substantive points about the root of social work. With the focus on inequality and oppression, the political and ethical or normative aspect as the core business of social work is proposed.

Traditions from the 1970s

From the 1970s, radical social work or the 'old' radical social work approach focused mainly on Marxist social theories. It criticized the capitalist system and traditional social work and launched a call for changes that enabled social work to change capitalist society, favor the working class and eliminate its pathologizing approach to individuals needing help. Social work practice, i.e., casework, was criticized because of its individual approach perpetuating the social order, to promote instead, community work as a collective approach (Bailey and Brake, 1975; Dominelli, 1990; Fook, 2002; Ferguson, 2008). The aim was not to eliminate casework, but to raise awareness of its deficit approach to service users (Dominelli, 1988). This awareness and the simultaneous coming together of critical social workers, service users, students, and academics was of vital importance. Within the current or 'new' radical social work, sociocritical theories provide a theoretical frame alongside social and sociological theories. Initially, critical theory was a tradition of thinking ('theorizing') referring to Marx and Hegel, while emancipating itself from this tradition. A core principle of critical theory is analyzing society holistically (relationships and systems) rather than as a collection of individuals. It focuses on structural power relations and models of domination while seeking structural change, and encouraging positive ideals (Laermans, 2009). Within this tradition dominant power relations are analyzed, giving social workers and clients/service users the opportunity to develop critical awareness for undertaking critical social work practice (Mullaly, 2001; Fook, 2002; Ferguson, 2008).

Radical social work since the 1970s has been a strong movement. Rather than disappearing, it finds roots in emancipatory social movements concerning women's rights and the oppression of specific social groups on the basis of 'race,' ethnicity, disabilities, age, and sexual orientation (Dominelli, 2002c). Radical Marxist theories of social work practice were critiqued for not having models for practice (Dominelli and McLeod, 1989). Neoliberalism from the 1990s marked a shift in the political climate and required engagement with professionalization and managerialism. Instead of equality and solidarity, contemporary politics focus on the strength of the individual and the idea of the free market (Hoogerwerf, 2001) and impacts adversely on what social workers can do (Dominelli, 1996, 1997). Also psychosocial discourses from the 1970s alongside Rogers' client-centered practice continue to promote human potential and growth (Payne, 2005).

In the first decade of the twenty-first century, radical or critical social work prevails again. Environmental- and climate movements, as well as anticapitalist, antiglobalist, and antiwar movements predominate, given the global crises (Dominelli, 2009, 2012; Ferguson, 2008). Under the influence of postmodernism and social constructionism of the 1990s, critical ideas have prospered and led to the reformulation of Marxist social work, utilizing critical theories of society, and the practice of empowerment and advocacy (Payne, 2005). A keystone of the radical, structural or critical approach is the focus on social structures and 'not blaming the individual victim' for problems (Fook, 2002). These approaches do not take the social order for granted.

Radical social work, the critical movement of social workers in the 1970s, occurred rather moderately in Flanders. Most social workers focused more on the external and societal causes of the problems their clients faced. Accordingly, social workers engaged in the battle to tackle and change these societal causes, rather than accept that problems would be solved on an individual level (Verzelen, 2005). Societal critiques illustrate the highly unequal distribution of resources and chances to purchase economic, political and cultural goods (income, power, knowledge, status). These critique produced alternative and politicized social work, social action, education (awareness), and social work. Politicized work is primarily structural and environmentally based, and entails a collectivization of the issues instead of their individualization. Characteristic of this perspective is social workers' preference for client choice and their activation through externalizing societal and social problems instead of internalizing them (Snellen, 2002). The dominant attitude of social workers is that they show solidarity

with the client, that they struggle against the internalization of social problems, and that they fight structural causes, not least, within their own institutions. The conflict within their own social work organizations is addressed by paying attention to the role their superiors/managers have in limiting clients in the interest of reproducing their power as owners and/or managers siding with the ruling interests (Verzelen, 2005).

The emancipatory approach takes this societal critique a step further. Society is not just a combination of external problem-causing factors. It is also within the person, a factor of one's identity. This identity is reproduced by the processes of socialization and patterns of human interaction. Discipline and the power to discipline are general societal phenomena which occur both within societal institutions (state, capitalism) and interactions occurring on the microlevel (Snellen, 2002). According to Snellen (2002), emancipatory work (cf the empowerment-approach from the 1990s), aims to help people to control their own lives, discover their own possibilities, have their voices heard, and fight situations of inequality and oppression. When taking on the problem/situation, their focus lies with the client, as an involved 'expert' actor. Now, more than ever, social workers have to immerse themselves in the causes of people's problems, working informally and being with or deeply involved in poor people's situations, and acting to improve situations and make their work matter (Van Riet and Wouter, 2002).

Lost in the 1990s

Critical trends connect social work with the need for change in social and institutional relations in society and illustrate how social work too easily withdraws into the microlevel (De Bie and van Ewijk, 2008). From the roaring 1960s to the 1980s, the critical social work movement continued its lead on the internal critique of the psychologization of social problems and the external one that ignored the politicization of the profession (Herman, 2001). Verzelen (2005) argued that social work discourses traverse two traditional forces, going back and forth between the emphasis on individual responsibility and that of societal and social environmental factors.

Simultaneously, through extensive professionalization and institutionalization, social work made itself rather immobile and impotent. Van Riet and Wouter (2002) uses Procrustes' metaphor (literally the stretcher): the needy are stretched or customized to the supply of aid. With their supply and expertise, social workers preserve their own market and deprive clients of their own capabilities (Achterhuis, 1980). This message brings great uncertainty and reticence within social work and helps legitimize government cuts. The neoliberal climate of the 1980s and 1990s, led to marketization, commercialization, and the withdrawal of government provisions, while postmodernism challenged the prevailing grand narratives and ideologies that resulted in deprofessionalization (Herman, 2001). Neoliberal politics initiated the transformation from a welfare state to an activation welfare state (De Bie and van Ewijk, 2008) whereby individuals were held accountable for meeting their own needs. These initiatives prioritized individual progress over collective progress. The liberal injunction that each individual is responsible for him/herself is based on transference of free market principles to human relations (Hoogerwerf, 2001). The tendency to commercialize services increasingly leads to a certain 'contractualization' of relationships: contracts instead of contacts. Professionalism increasingly stands for rationality and detachment, a new business-like character in which well-organized activities preferably yielding quantifiable outcomes. Notions of attention, being with, equality, seem old-fashioned and unprofessional because outcomes are wanted and only talking is not expected from professional social workers (Jacobs and Van Doorslaer, 2000).

Apart from the neoliberal approach, also called neoconservatism, Reaganomics, or Thatcherism, postmodernism played an important role, rejecting the 'grand narratives' for paying limited attention to equality and emphasizing instead the diversity of people and cultures. The welfare issues of the most vulnerable are deemed safety problems. Hence, the care of vulnerable groups has shifted to the fear of a new 'underclass,' which is a risk for the middle and higher classes (Wacquant, 2012). Risk images are dictated by fear of the unknown, unfamiliar, and unconventional. Negative solidarity and notions such as risk, harm, and mostly 'nuisance' prevail. As a reaction, the demand for safety, security, and control increases. This is where a paradox originates – the tendency to consider danger as manageable risk, makes people see more danger, at which point feelings of fear and insecurity will increase even more. Fears are dramatized, due to the way the media and politics respond to this matter. Consequently, more means of social control arise with the aim of restricting feelings of fear and preventing crime. This phenomenon is not new. Scientific research plays a greater role and empirical research on crime increasingly acts as a scientific base for policies on criminality. Along with new technological possibilities this leads to different interventions.

The government increasingly focuses on 'controlling risks and evil.' A new sort of state is created, a new kind of governance (van Swaaningen, 2001). The government rejects certain tasks and is led by experts (the power of consultants), entrepreneurs (the power of the market), and interest groups (the power of the lobby). Foucault (1991) discusses the 'microphysics of power': power relations are not presented in a pyramid structure with a ruler or monarch giving top-down orders. Power is exercised on different levels and places in society and stems from a network of private and public bodies without absolute power, but with many different powers, based on professional knowledge and the resulting relations of authority (van Swaaningen, 2001).

This 'remote administration,' as Foucault puts it, implies that the government draws policies and that the precision and implementation of these policies is left to third parties in the private sector. The withdrawing government has a different role, a different character, and is rather an enterprising government. Governance is a question of business by making use of tactics rather than laws, and even the laws are used as tactics (Foucault, 1991). Due to the withdrawal of government, a democratic shortage can occur. A radical depoliticization of society takes place: politics becomes an arena parallel to other arenas such as economy, law, and art (Visker, 2005). Politics is more about technical and governmental guidance of processes (resources) and is less based on substantive arguments and fundamental choices or deliberation (goals) (van Swaaningen, 2001).

Reclaiming the Twenty-First Century

Social work has become too much an 'executor' and evaluator of risks, thereby falling into bureaucratic routine and marginalization (De Bie and van Ewijk, 2008). Professionals have been hiding behind office windows and procedures and have increasingly isolated themselves from people's living environments (Van der lans, 2010). However, there is another aspect: 'corrective external influences,' such as the development of user movements and experiential knowledge of poverty reduction, where approaches of living environment and presence, critique bureaucratic welfare work (Verzelen, 2005).

Radical social works connects a structural analysis of individual problems with the ethical imperative to act. During past decades, this combination has been attractive to social workers, who accept that social justice and equality are a central part of social work (Dominelli, 1997; Powell in Ferguson, 2008: 96). Social workers should report more on the impact of (neoliberal capitalist) policy on their client's lives. Social workers can make the difference and play a role in securing equality and justice in society (Bailey and Brake, 1975; Galper, 1975; Throssell, 1975; Dominelli, 1990, 2009; Thompson, 2005; Ferguson, 2008). They can formulate an ethical answer, based on a critical analysis of the political context in which social work is situated. It is important that social workers unite to develop a new vision, paradigm or ideology for practice (Mullaly, 2001).

The current international definition of social work states that principles of human rights and social justice are fundamental to practice and that social work targets social change, problem solving in interpersonal relations and empowerment and freedom of people, considering their well-being (IFSW and IASSW, 2004). This could indicate a positive attitude for a transformational socialist collectivist vision, but with the focus on the social and political position of service users/clients (Payne, 2005). Social work as emancipatory practice is person-centered, empowering, and critical of power structures and systems of resource distribution that undermine the well-being of both the many and the individual (Dominelli, 2009: 2). Working on social change is trying to change the negative and destructive aspects of society, such as discrimination and oppression. Working on social change does not mean that social stability is not important. Social work should question which aspects should be retained. Thompson (2005) argues that a 'realistic' vision of social work focuses on social stability and social change. Providing social stability strengthens those aspects that promote individual and social well-being/prosperity. Social workers address social problems that have their roots in society. The stress on the social dimension of problems would pave the way for social work to promote social change again (Jordan, 2006; Roose, 2006).

Theoretical Evolution

Theory-development in Flanders is reflected in limited publications. Some specific theoretical bases for Flemish social work stand out (Verzelen, 2005). During the period 1960–80, human and social sciences were evolving and adult-learning about change and expertise began to fill the gap of the missing basic science for social work (Verzelen, 2005). This 'new' science feeds welfare work with an individualistic approach alongside sociocultural work with its commitment to changing society. For this approach, Freire's work (1972) leads on adult-learning theories. Freire directly influenced Flemish sociocultural work in general, and literacy work in the 1980s in particular (Verzelen, 2005). Freire connected the living environment of his participants to their thinking about inequalities and injustice in society (Verschelden and De Droogh, 2009). Indirectly, he influenced social work more broadly, because Piet Reckman and Bert De Turck politicized social work's development within welfare and sociocultural work.

The sociocritical approach is also fed by sociological theories, i.e., 'critical theory,' mainly from the Frankfurt School including Marcuse, Fromm, and Habermas. With a model of communicative dialogue, Habermas presents a way for reaching binding (ethical) conclusions (Verbruggen, 2007). Both Freire's (1972) thinking and critical theory helped realize the connection between engaged emancipation perspectives and techniques of support (Verzelen, 2005).

Finally, other theories for mention regard the rise of target group approaches in the 1990s and include theories on street outreach work or Baart's (2006) presence theory. These theories consider welfare work and poverty, i.e., custom work, connective strength, and empowerment (tailoring care) (Van Regenmortel, 2002), and become a binding force in poverty (Driessens and Van Regenmortel, 2006). Simultaneously, Boszormenyi-Nagy and Krasner's (1994) and Boszormenyi-Nagy's (2000) contextual approaches are, together with the systems approach, the most important source for methodological ideas for current welfare workers (Verzelen, 2005).

Even though psychotherapy's influence on social work is more extensive in the United States than in the Netherlands and Flanders, social work must realize that it will neglect its mission by identifying too much with psychotherapy (Van Riet, 2006). Because social workers are the closest to social problems, they also have a public role and they should present themselves socially, given growing poverty, unemployment, social inequality, and extreme exclusion and stigmatization (Driessens and Geldof, 2009). Hence, the tension of individual and structural change remains relevant in the twenty-first century. Also, the international definition of social work links individual and structural work with one another to promote social change. Different perspectives or theories stipulate whether social workers think and act from an individual or institutional responsibility perspective (Van Regenmortel, 2002; Driessens and Geldof, 2009).

Exploring Social Work Practice

Nowadays, the debate on radical or structural social work is also being held in the field. Regarding practice-oriented research by KH Leuven and KU Leuven (Bloemen, 2012), a survey has been conducted to address actual needs expressed by social work organizations willing to work structurally. Participants from various organizations expressed concerns about this topic to the advisory committee for this research project. Both regional and Flemish partners were represented on this committee. Its members were from a local social services

agency in Genk (OCMW), an organisation supporting experts by experience in Leuven (Leren Ondernemen vzw), a center for general welfare in Leuven (CAW), street outreach work in Flanders (Vlastrov), community development in Flanders (Samenlevingsopbouw), and youth work in Flanders (Uit de Marge). During this research the education institution and field closely collaborated alongside the students taking a Masters in Social Work degrees (KU Leuven).

Interviews had been conducted and three focus groups were constituted. The members of these focus groups were specifically chosen for their experience in social work. Participants defined structural work as caring (primary) and political (secondary). It does not exclude individual work, but combines individual and collective work (group-centered, project-based, community-centered). Structural work is a laborious effort about major, long-term changes as well as small, short-term activities and is rooted in people and anchored in society. It centers people in different ways by initiating policy changes from the bottom up. Instead of guiding people, it creates space to develop new structures. A development of transparent organizational structures and communication channels seem crucial to working structurally. Advocacy groups can be mandated to undertake such work, given their independence and organizational structure. Networking is crucial for structural work and making links with movements such as professional associations and civil society. Solidarity is crucial in making contact with people, to stand among them and be on the same side of the barricade.

Strikingly, social workers are hindered to do their work given the absence of space for critique and the lack of an ideological framework. More specifically, the lack of vision and ideology and the lack of sociocritical analysis appear to be a great obstruction. Being able to analyze is a must and a starting point for critical social work. Insight into politics, mechanisms of exclusion, policymaking, law, and their implementation are imperative to work transformationally and emancipatorily. A critical–ideological and collective framework as a base for action is meaningful and needs to be acquired by social workers. Dealing constructively with anger and completely supporting the goal of equality are mentioned as social work drivers.

During their education, social workers have to be enabled to critique society, develop a critical attitude and insights into the backgrounds and underlying paradigms of societal phenomena (Cuyvers, 2009). Bloemen (2010) conducted interviews to explore ideas about radical social work and the profession of social work. Four Flemish polytechnic universities were included in the study. Analysis of the 18 in-depth interviews highlighted the importance of social workers' (critical) reflective capacity and vision development; the challenges of increasing collaboration between education and field; the lack of social work influence on the structural societal level; and the demand for unity among social workers. Most respondents name the capacity of reflection a skill which students acquire during their social work education, alongside communication skills. They think that the development of vision is a task for the School to provide frameworks that promote thinking and gaining insights into society. Respondents generally did not specify these frameworks. Within 'new' radical social work, critical theories are put forward (alongside sociological theories) as theoretical frameworks because they analyze society as a whole, taking power relations into account.

Most respondents answered positively and convincingly the question of whether social work can influence social change. In microlevel or individual contact with people, social workers definitely make a difference, according to the methodology tutors. Some respondents pointed out that influencing the meso- or organizational level can be taught. Social work indicates that by developing activities of individual and collective resilience, social workers can promote change. Thus, social change can be initiated by questioning the organization. The policy or macrolevel was unanimously considered as the level on which too little action took place. Social work can steadily evolve into becoming a codesigner of social politics and social policy, by emphasizing the link between social work and the basic objectives of social justice and equality. This approach considers social work as having the expertise to make critical connections between individual situations and societal developments (Driessens and Geldof, 2009). However, the question arises as to how this is to be put into practice?

Social Work Making Radical Progress?

The information this research project provided shows that social work practice and theory developments influence one another. From its beginning onward, social work has been philanthropic, caring for social stability. It has also focused on social change to stress the structural elements of social problems and inequality and oppression. Social work is emancipatory or, committed to removing the barriers that limit peoples' aspirations within the existing social order. Social inequality, oppression, and dominant power relations can be changed by raising consciousness and social action. Social workers are aware of the influence of the broader context and hegemony of neoliberalism and capitalism. The analyses of dominant power relations gives social workers and clients/service users the opportunity to develop critical awareness as a base for a critical social work practice (Fook, 2002; Ferguson, 2008; Mullaly, 2001).

During recent years, there has been an increase of social movements identifying capitalism as the culprit of poverty and environmental destruction (Dominelli, 2012). According to Mullaly (2001) structural social work can support social movements that promote social transformation. By stressing the connection between social work and basic goals such as social justice and equality, social work will be able to become a codesigner of social politics and social policy. This approach considers social work as having the expertise to link individual situations to societal developments (De Bie and van Ewijk, 2008).

The dominant attitudes among social workers in this project were that of showing solidarity with the client, and battled against the internalization of social problems, and tackled structural causes, not least within their own institutions. Both within and outside of their organizations and systems, social workers can engage in and encourage social change. It is imperative for social workers to have a clear view

on the immediate, varied, and changing situations that people live within from day to day (Saleebey, 2006). A societal viewpoint implies that the engagement of social work is not limited to microsituations, but extended through a broader societal and historical analysis, conducted by critical social workers. Social work radicals cannot simply discover and analyze the causes of oppression or injustice. The next stage is to transform their insights into immediate social action (Reisch and Andrews, 2001). This alternative approach is something that clients and practitioners do together, forming alliances with each other as necessary (Dominelli, 2004).

See also: Critical Reflexivity and Social Work Practice; Critical Social Work Practice; Global Agenda for Social Work and Social Development; International Social Work; Marxist Approaches to Social Work; New Managerialism and New Public Sector Management; Social Action and Social Justice; Social Protest Movements and Social Work Practice; Social Work Education; Social Work Theory.

Bibliography

Adams, R., Dominelli, L., Payne, M., 2002. Critical Practice in Social Work. Palgrave Macmillan, London.

Achterhuis, H., 1980. De markt van welzijn en geluk. Een kritiek van andragogie. Ambo, Baarn.

Bailey, R., Brake, M., 1975. Radical Social Work. Eduard Arnold (Publishers), London.

Baart, A., 2006. Een theorie van de presentie. Lemma, Den Haag.

Bloemen, H., 2010. Radicaal sociaal werk: realiteit of illusie? KULeuven.

Bloemen, H., 2012. Structureel sociaal werk. De structurele component van het sociaal werk methodisch onderbouwen. KHLeuven.

Boszormenyi-Nagy, I., 2000. Grondbeginselen van de contextuele benadering. De Toorts, Haarlem.

Boszormenyi-Nagy, I., Krasner, B., 1994. Tussen geven en nemen. Over contextuele therapie. De Toorts, Haarlem.

Chambon, S., Irving, A., Epstein, L., 1999. Reading Foucault for Social Work. Columbia University Press, New York.

Coben, D., 1998. Radical heroes. Gramsci, Freire en de politics of adult education. Garland Publishing, New York-London.

Corrigan, P., Leonard, P., 1978. Social Work Practice under Capitalism: A Marxist Approach. Macmillan, London.

Cuyvers, G., 2009. Sociaal werk: ontwikkelen in een netwerk van relationele praktijken. In: Brodala, J., Cuyvers, G., Devillé, Van den Eeckhaut, G. (Eds.), Met een dubbele lus. Prikkelend nadenken over sociaal werk. Antwerpen-Apeldoorn: Garant.

De Bie, M., van Ewijk, H., 2008. Sociaal werk in Vlaanderen en in Nederland: een begrippenkader. Wolters Kluwer, Mechelen.

De Turck, B., 1978. Politiserende hulpverlening, een oriëntering. Link, Nijmegen.

Dominelli, L., 1988. Anti-racist Social Work, second ed. Macmillan, London. third ed. 2005.

Dominelli, L., 1990. Women and Community Action, second ed. Venture Press, Birmingham. 2006.

Dominelli, L., April 1996. Deprofessionalising social work: equal opportunities, competence and postmodernism. British Journal of Social Work 26, 153–175.

Dominelli, L., 1997. Anti-racist Social Work, second ed. Palgrave, Basingstoke.

Dominelli, L., 2002a. Anti-oppressive Social Work Theory and Practice. Palgrave-Macmillan, London.

Dominelli, L., 2002b. Feminist Social Work Theory and Practice. Palgrave-Macmillan, London.

Dominelli, L., 2002c. Antioppressive practice in context. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Socical Work: Themes, Issues and Critical Debates, second ed. Macmillan, Basingstoke.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.

Dominelli, L., 2009. Introducing Social Work. Polity Press, Cambridge.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.

Dominelli, L., McLeod, E., 1989. Feminist Social Work. Macmillan, London.

Driessens, K., Geldof, D., 2009. Individu en/of structuur? Of wat wil het sociaal werk aanpakken? http://www.canonsociaalwerk.be/essays.php (accessed 05.06.11.).

Driessens, K., Van Regenmortel, T., 2006. Bind-kracht in armoede. Leefwereld en hulpverlening, LannooCampus, Leuven.

Ferguson, I., 2004. Neoliberalism, the third way and social work: the UK experience. Social Work and Society 2 (1), 1–9.

Ferguson, I., 2008. Reclaiming Social Work. Challenging Neoliberalism and Promoting Social Justice. Sage, London.

Fook, J. (Ed.), 1996. The Reflective Researcher: Social Workers' Theories and Practice Research. Allen & Unwin, Sydney.

Fook, J., 2002. Social Work. Critical Theory and Practice. Sage, London.

Fook, J., Gardner, F., 2007. Practising Critical Reflection. A Resource Handbook. Open university press - McGraw-Hill education, Berkshire.

Foucault, M., 1991. Governmentality. In: Burchell, G., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality. Harvester Wheatstaf, London, pp. 87–104.

Freire, P., 1972. Pedagogie van de onderdrukten. In den toren (Anthosboek), Baarn.

Galper, J., 1975. The Politics of Social Services. Prentice-Hall, Englewood Cliffs.

Herman, S., 1996. Maatschappelijk werk. Tussen gisteren en morgen. Garant, Leuven-Apeldoorn.

Herman, S., 2001. Onvoltooid verleden tijd. Maatschappelijk werk en social casework. Leuven-Apeldoorn: Garant.

Hermans, K., 2009. De complexe relatie tussen sociaal werk en sociaal beleid ontrafeld. In: Brodala, J., Cuyvers, G., Devillé, A., Van den Eeckhaut, G. (Eds.), Met een dubbele lus. Prikkelend nadenken over sociaal werk. Garant, Antwerpen-Apeldoorn.

Hoogerwerf, A., 2001. De onweerstaanbare gelijkheid. Een geschiedenis van het sociaalpolitieke denken. Damon, Budel.

Howe, D., 1994. Modernity, postmodernity and social work. British Journal of Social work 24 (5), 513–532.

Howe, D., 2008. An Introduction to Social Work Theory, revised ed. Ashgate, Hampshire.

International Federation of Social Workers and International association of Schools of Social Work, 2004. Ethics in Social Work, Statement of Principles. http://www.ifsw.org/f38000032.html.

Jacobs, R., Van Doorslaer, J., 2000. Het pomphuis van de 21ste eeuw. Educatie in de actieve welvaartsstaat. Epo, Berchem.

Jordan, B., 2006. Emancipatorisch social work? Mogelijkheid of oxymoron. Paspoort Maatschappelijk Werk (1), 73–91.

Laermans, R., 2009. Kritische maatschappijtheorie. Katholieke Universiteit Leuven, Leuven.

Leonard, P., 1975. Towards a paradigm for radical practice. In: Bailey, R., Brake, M. (Eds.), Radical Social Work. Eduard Arnold (Publishers), London.

Midgley, J., 2001. Issues in international social work. Resolving critical debates in the profession. Journal of social work 1 (1), 21–35.

Mullaly, B., 1997. Structural Social Work. Ideology, Theory and Practice. Oxford University Press, Ontario.

Mullaly, B., 2001. Confronting the politics of despair: toward the reconstruction of progressive social work in a global economy and postmodern age. Social Work Education 20 (3), 303–320.

Payne, M., 2005. Modern Social Work Theory. Palgrave Macmillan, London.

Payne, M., 2006. What is Professional Social Work? Policy Press, Bristol.

Reisch, M., Andrews, J., 2001. The Road Not Taken. A History of Radical Social Work in the United States. Brunner-Routledge, Philadelphia.

Roose, R., 2006. De vraag naar profilering: hulpverlening tussen begrenzing en ontgrenzing. Sociale Interventie (4), 17–25.

Saleebey, D., 2006. De macht van het kleine: een andere kijk op de omgeving. Paspoort Maatschappelijk Werk (1), 7–29.

Snellen, A., 2002. Vijftig jaar methodiek. Maatwerk, vakblad voor maatschappelijk werk (4), 4–11.

van Swaaningen, R., 2001. Criminaliteitsbeleid en risicobeheersing. In: Lissenberg, E., van Rullier, S., van Swaaningen, R. (Eds.), Tegen de regels IV. Een inleiding in de criminologie. Ars Aequi Libri, Nijmegen.

Thompson, N., 2005. Understanding Social Work. Preparing for Practice, second ed. Palgrave Macmillan, Hampshire.

Throssell, H. (Ed.), 1975. Social Work: Radical Essays. University of Queensland Press, St Lucia Queensland.

Van der Lans, J., 2010. Erop-af! De nieuwe start van het sociaal werk. Augustus, Amsterdam.

Van Regenmortel, T., 2002. Empowerment en maatzorg. Acco, Leuven.
Van Riet, N., 2006. Social work. Mensen helpen tot hun recht te komen. Van Gorcum, Assen.
Van Riet, N., Wouter, H., 2002. Helpen is leren. Emanciperende hulpverlening als methode. Van Gorcum, Assen.
Verbruggen, A., 2007. De ziel van de stiel. Maatschappelijk assistenten en hun beroep: bouwstenen voor een gedeeld verhaal. Academia Press, Gent.
Verschelden, G., De Droogh, L., 2009. Van educatie naar leren en terug? Over de verschuivende functie van leren en educatie in onze samenleving. In: Brodala, J., Cuyvers, G., Devillé, A., Van den Eeckhaut, G. (Eds.), Met een dubbele lus. Prikkelend nadenken over sociaal werk. Antwerpen-Apeldoorn: Garant.
Verzelen, W., 2005. Sociaal werk. In- en uitzichten. Antwerpen-Apeldoorn: Garant.
Visker, R., 2005. Vreemd gaan en vreemd blijven, Filosofie van de multiculturaliteit. Sun, Amsterdam.
Waaldijk, M.L., 1996. Het Amerika der vrouw. Sekse en geschiedenis van maatschappelijk werk in Nederland en de Verenigde Staten. Erasmus universiteit, Rotterdam.
Wacquant, L., 2012. Paria's van de stad. Nieuwe marginaliteit in tijden van neoliberalisme. Epo, Berchem.

Radio as a Medium

Joy E Hayes, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Due to its mobility and low cost, radio is the most widely used electronic medium of communication on the globe. Although public monopoly broadcasting dominated until the 1980s, US-based commercial broadcasting was a global force beginning in the 1930s. Radio played a key role in promoting national political and cultural integration under both models. By the late 1980s, most of the world's radio systems faced liberalization of some kind, although the Arab world held out until the 2010s. A global community radio movement emerged in the 1980s and expanded in the 2000s, nurtured by democratization struggles.

Radio, the transmission of sound via electromagnetic waves, is the most widely used electronic means of communication on the globe. It is second only to spoken language and nonverbal interaction as a medium of human communication. Despite the broad reach of television and the growing role of cable and Internet, radio remains an important arena of cultural expression, commerce, politics, and education in advanced capitalist countries. In many developing countries, including most of Africa, much of Asia and the Middle East, and parts of Latin America, Eastern Europe, and Oceania, radio continues to be the primary means of mass communication. Due to its mobility, low cost, and flexibility, radio is one of the fastest growing media of communication worldwide.

Although radio has a variety of scientific and commercial applications as a *technology of radio wave transmission* (e.g., radio astronomy, radiotelegraphy, radiotelephony), as a *social medium* radio has been configured principally as a means of point-to-mass communication or broadcasting. Two broad models of radio broadcasting emerged during the first half of the twentieth century: commercial broadcasting and public monopoly broadcasting. While both models defined the radio wave spectrum as a public resource or good, they differed on whether the spectrum should be exploited by competing private entities for profit or developed by a single public entity as a nonprofit service. During the last quarter of the twentieth century both commercial broadcasters and community-based nonprofit broadcasters challenged the public monopoly model and launched new and hybrid broadcasting forms. The following article examines the cultural, technological, economic, and political factors that shaped radio as a social medium of broadcasting.

Technology and Cultural Form

Radio is comprised of a dynamic sound technology wedded to a flexible yet resilient set of cultural practices. This combination of technological and cultural factors both enables and constrains radio as a medium. Radio production, for example, is shaped by the capacities of the technology (real-time, sound-only) as well as the forms of expression that have been adapted to the medium over time (program genres, formats, and presentation styles). Radio reception is likewise constrained. For example, a portable, personal radio differs from a console radio in both its technological configuration and in the cultural assumptions that shape its use (mobile-individualized listening vs situated-shared listening). This interaction of technology and culture shapes not only the form and content of the radio medium, but radio's relationship to time, space, community, and identity.

In this sense, a defining feature of radio as a technology and cultural form is its conflicted relationship to modernity. On the one hand, radio broadcasting represents a heightened sense of movement toward modernity; that is, movement toward a condition in which human social relations are increasingly abstract, individualized, and future-directed. To the extent that radio is understood to speed up processes of social progress and development (through, for example, the diffusion of innovations) the medium evokes a sense of 'hypermodernity.' On the other hand, radio provides a means of resisting or allaying the impact of increased mobility and rationalization in social life. To the extent that radio is used as a tool for building intimacy and community in an individuated and changeable modern world, broadcasting practices are 'antimodern.'

Radio broadcasting both reflects and constitutes uniquely modern forms of social organization. It works to integrate and coordinate an increasingly mobile, dispersed, and complex society, at the same time that it promotes the privatization of social life. The simultaneous access of vast numbers of isolated listeners to a single radio content creates a virtual common space that tends to become coterminous with the nation (due, in part, to government regulation and commercial market formation). In this way, radio, like print and cinema, acts as a medium for the "community in anonymity which is the hallmark of modern nations" (Anderson, 1991: 36).

At the same time, however, radio voices have a materiality that reproduces the immediate, tactile quality of primary orality in a number of ways. Roland Barthes's concept of the 'grain' of the voice is helpful in characterizing the somatic presence of radio voices. Barthes (1990) describes the grain as "the body in the voice as it sings"; that is, the listener's sense of the physical presence of the performer in the sound of his or her voice (299). Walter Ong (2012) argues that radio talk has the same ability as oral speech to envelope and immerse listeners, uniting the audience "with themselves and with the speaker" (73). Indeed, broadcasters actively work to re-create

the sociability and interactivity of oral settings by developing intimate, chatty forms of radio talk. In this way, radio gives listeners a sense of continuity with older forms of intimacy and community at the same time that it enables distinctly modern forms of social life.

History and Political Economy

Radio has historically been deployed by a range of political and economic institutions largely within the bounds of the nation-state. During the last two decades of the twentieth century radio changed from a predominantly public service medium to an eminently commercial one. To understand the meaning of this global transformation it is necessary to investigate the social forces that guided and shaped radio broadcasting beginning in the 1920s. Among the most important influences on the medium have been national and transnational communication corporations, the state (nationalist and expansionist) and community groups.

Beginnings: The 1920s and 1930s

Radio expanded from a means of point-to-point communication to a medium of point-to-mass communication, or broadcasting, during the 1920s and 1930s. The rise of broadcasting was both shaped by and reflective of larger political and economic changes in the first quarter of the twentieth century, including the rise of transnational communication corporations following World War I (e.g., Philips, RCA, Telefunken), the growth of product marketing, increased trade in cultural goods (e.g., film and recorded music) and the escalation of both nationalism and international expansion. Technological developments including voice broadcasting (replacing Morse code) and high-power radio transmissions were also necessary ingredients for broadcasting.

Approaching the medium as a strategic communication technology, scarce resource and tool for national integration, central governments began asserting control over the radio spectrum in the 1910s and 1920s. In part, this followed the pattern established by state intervention in telegraph and telephone development. As a U.S. Commerce Department publication lamented, "When the social possibilities of broadcasting became apparent many countries determined upon maintaining it as a public utility, particularly those where government ownership of communication services had previously been the rule" (Radio Markets, 1932: 2). Even in the United States, however, the central government claimed radio as a public resource and used regulatory powers to promote a national broadcasting system with citizenship and public service requirements for broadcasters. These regulations, codified in the 1927 Radio Act and again in the Communications Act of 1934, granted concessions or licenses for commercial broadcasting and actively encouraged the development of commercial networks on a national scale.

Radio was shaped by a climate of intense nationalism and international expansion. In industrialized countries radio was viewed as a means of consolidating the nation as both market and polity, strengthening ties with colonies and dependencies and opening foreign markets. For example, the British Broadcasting Corporation (BBC) began its Empire Service in 1932, while the Japanese government established broadcasting stations in occupied territories in both Korea and China. The U.S. media corporations such as the Radio Corporation of America, General Electric, and Westinghouse financed broadcasting stations in foreign markets in order to promote the sale of radio receivers (e.g., Latin America). In less-industrialized countries and colonial territories broadcasting rarely reached beyond a few major cities and was often limited to the interests of colonial governments (e.g., Hong Kong, Kenya).

In postrevolutionary societies such as Mexico and Soviet Russia, as well as fascist countries like Germany and Italy, the state turned to radio as a tool for building political solidarity and cultural identity through official broadcasting projects and the distribution of radio sets for both public and in-home listening. In Nazi Germany inexpensive radio sets known as the 'People's Wireless' or 'Goebbel's Lips' allowed listeners to tune in Adolf Hitler's speeches but blocked the reception of foreign stations (Euromedia, 1997: 79). In almost all independent countries radio developed within a nationalistic framework: broadcasters were required to be citizens transmitting from national soil, the state was given special access to the medium on the grounds of national interest and security, and content was frequently regulated to promote and 'improve' the national culture (e.g., language and music requirements).

Although radio broadcasting took many different forms, two broad types of radio organization – commercial broadcasting and public monopoly broadcasting – emerged during this period. In the commercial model, radio programs were designed to attract large audiences that could be sold to consumer product advertisers through program sponsorship or 'spot' advertisements. This model predominated only in the United States and Latin America, although Canada, Australia, and New Zealand had significant commercial radio sectors and small pockets of commercial radio existed elsewhere. In France and Luxembourg, for example, private broadcasters transmitted commercial programs in direct competition with the BBC (e.g., Radio Luxembourg, Radio Normandie). In the vast majority of countries and occupied territories radio was a public monopoly (or modified monopoly) financed by government subsidy or receiver license fees and subject to varying degrees of government control. These two general models yielded different geographies of broadcasting: whereas public monopolies tended to produce a small number of stations serving the national community (with limited regional or local service), commercial systems tended to generate a large number of stations serving the local community. With the rise of national networks, however, local commercial affiliates developed a strong national orientation.

Only countries in North America and Western Europe approached universal radio access by the end of the 1930s when, for example, about 75% of the British population had radios. One of the most influential broadcasts of the era, Franklin Roosevelt's first Fireside Chat of 18 March 1933, was heard by an estimated audience of 60 million (out of a total population of about 120 million). Urban and industrializing areas of Latin America, Oceania, and Asia, including the Soviet Union, had small but growing numbers of radio listeners. Rural parts of these regions, along with most of Africa and the Middle East, had negligible radio listenership.

Takeoffs and Transitions: The 1940s and 1950s

World War II had a major impact on radio organization, content, and listenership. Timely broadcasts of news and information greatly increased the utility of the medium, promoted frequent listening, and raised demand for radio sets both during and after the war. Radio listening grew even in countries where most people could not afford radio sets (e.g., Egypt) "because merchants or owners of coffee shops would have a set to attract customers" (Dabbous, 1994: 65). The role of the central state in broadcasting also increased significantly on both the domestic and international fronts. At home, the state took an interest in the security and efficiency of the radio system. Official reports and government-sponsored programs proliferated on both public monopoly and commercial radio stations. In some cases (e.g., Vichy France) the state used the wartime emergency to prohibit private broadcasting and institute a public monopoly system.

International broadcasting was exploited by all of the belligerents during the war, largely in response to Italy and Germany who invested heavily in short-wave broadcasting. The BBC, for example, became the medium for Charles de Gaulle's appeals to the French Resistance, beginning with his celebrated broadcast of 18 June 1940. The U.S. government developed international broadcasting projects aimed at Asia, Europe, and Latin America. In the case of Latin America, the U.S. Coordinator of Inter-American Affairs developed a wartime broadcasting offensive that worked to solidify U.S. hegemony in the region while strengthening the cultural influence of U.S. advertisers and a small number of Latin-American broadcasters. In 1948, the Coordinator's radio unit was transferred to the U.S. State Department where it became the Voice of America broadcasting service. By the end of the war, international broadcasting became the provision of the central state.

Radio broadcasting became one of the most important ideological and political tools of the Cold War. Along with international broadcasting, the United States, Soviet Union, and their allies sponsored clandestine radio stations in countries around the world (e.g., Guatemala, Turkey, Yugoslavia). The mobility and low cost of radio, however, also made it a powerful tool for grassroots revolutionary movements in the postwar period. Communist insurgents in China in the late 1940s and revolutionaries in Cuba in the late 1950s used clandestine broadcasting stations to maneuver against their enemies and mobilize popular support. Fidel Castro's Radio Rebelde, for example, became a model of effective clandestine broadcasting for decades to come.

Most of the countries that fought for independence at the end of World War II (e.g., China, India, Indonesia) followed a path similar to earlier post-Revolutionary societies: they established state-directed broadcasting systems that controlled language and content and used the medium for nationalist ends. New governments struggled to improve broadcasting coverage and increase listenership quickly and efficiently. Consequently, broadcasting services were often highly centralized and limited to one or two national services (e.g., the Central People's Broadcasting Station in Beijing, China).

The history of radio in the postwar period was also shaped by the rise of a second broadcasting medium: television. Although television fundamentally transformed radio broadcasting in the United States during this period, television's impact was much more gradual worldwide. Radio remained a public monopoly in most countries and a strong competitor for the national audience. In parts of Europe (e.g., Italy, Norway, Sweden) television failed to reach a majority of the population until the early 1960s and in Latin America the high cost of television guaranteed that radio would continue as the primary mass medium. Indeed, this period can be characterized as a *golden age of radio* in countries with developed broadcasting systems. In Latin America, for example, popular music (e.g., mambos, rancheras, sambas) provided the primary material for broadcasting while the radionovela (drama serial) reached the height of its popularity. In both commercial and public monopoly systems this was an important period of genre formation, with the live production of musical variety programs, comedies, soap operas, novelas, news, and game shows filling the airwaves.

In the United States, radio changed radically beginning in the early 1950s. In order to develop national television networks, broadcasting corporations like the National Broadcasting Company stripped capital, talent, and program genres from radio and ultimately dismantled their national radio networks. Without national network programming, radio stations were forced to reorient their activities to the local and regional level and cultivate new markets (e.g., teenagers, ethnic minorities, and out-of-home listeners). Radio content became increasingly diversified and localized as radio shifted to a repertoire of recorded music supervised by disk jockeys or DJs who advertised products and promoted the radio station. By the late 1950s, stations were catering to narrow market segments with increasingly standardized program formats (e.g., 'beautiful music' and 'top-40').

Overall the early postwar period was marked by two divergent trends in radio broadcasting. While richer Western countries moved away from radio and toward television as the primary broadcast medium, many newly independent, postcolonial countries looked to radio as the primary tool for national communication and unification projects. Although television came to symbolize development, it remained impractical for the vast majority of countries due to high costs and infrastructure limitations. Beginning in the postwar decades, then, a significant divide emerged and grew between countries that saw high levels of television penetration beginning in the second half of the 1950s and those in which radio remained the primary mass medium through the late 1970s and beyond.

Change and Competition: The 1960s and 1970s

Public monopoly broadcasting remained dominant during this period despite technological changes and increasing competition from television. In the countries that emerged from the last major wave of decolonization (e.g., Algeria, Nigeria, Uganda), public monopoly broadcasting with strong government influence was the norm. In their 1975 study, *Broadcasting in the Third World*, Elihu Katz and George Wedell noted the tendency toward government control of radio broadcasting in many developing countries. In the case of

Nigeria, for example, they observed that "The demands of a government needing to assert its authority over a country of 50 million people of diverse ethnic, linguistic, and cultural characteristics, between 70 and 80 percent of whom were illiterate, made closer control of broadcasting inevitable" (83).

At the same time, however, public monopoly broadcasters faced growing criticism for controlling and limiting public information. In India, for example, a 1964 government report noted that "Suspicion of official information has deepened in India because of an incorrect, even improper use of media for personalized publicity and an undue accent on [government] achievements …" (Ninan, 1998: 4). In France, during the political crisis of 1968, listeners turned to commercial radio stations outside of France (the peripheral stations) to hear balanced coverage of events that French radio covered with a pro-government bias. In many countries protests against the public monopoly system were registered in the form of unlicensed, pirate stations operating within the national territory. While some countries managed to stave off reforms, significant liberalization began during this period. For example, Italy legalized private broadcasting in 1976 and Sweden permitted noncommercial 'neighborhood radio' as early as 1978.

During the same period commercial broadcasting (both television and radio) came under attack as a wasteland of low-quality programming that failed to meet basic public needs for information, education, and culture. In the United States, legislators responded with the Public Broadcasting Act of 1967, which created a public corporation that helped to finance educational television as well as National Public Radio, a public service national radio network. In Latin America, where radio became accessible to the majority of the population during this period, most governments either initiated or greatly expanded the use of radio in national education projects (e.g., Brazil, Venezuela). Overall, this period witnessed a sizable growth in educational and nonprofit broadcasting by diverse civic groups, including religious denominations, universities, and other nonprofit organizations.

Although radio listening increased worldwide during this period, the divide between rich and poor countries continued. Radio's growth in the developed world was driven, in part, by innovations such as frequency modulation (FM) broadcasting and transistor radios. While FM radio made high-fidelity broadcasting possible, inexpensive transistor sets increased radio's portability outside of the home and facilitated the expansion of car radio. In the United States, for example, both innovations helped to push the number of radio sets to 1813 per 1000 population by 1975. In the developing world, however, FM expanded much more slowly due to the costs of conversion from amplitude modulation receivers. In addition, although inexpensive transistor sets made radio much more accessible to communities without electricity, the price of batteries remained prohibitively high for many listeners. Thus radio penetration remained relatively low in poorer regions of the world at the end of the decade. Asia and the Middle East, for example, had 62 sets per 1000 population while Africa had 70 sets per 1000 population in 1977 (Katz and Wedell, 1977: 62).

Liberalization, Deregulation, and Consolidation: The 1980s and 1990s

The move to reform public monopoly broadcasting that began in the 1970s grew enormously during the 1980s. Liberalization was a response to a variety of pressures: social pressure to increase political and cultural pluralism over the air, political pressure to reduce public spending (especially during a time of economic crisis) and ideological pressure in favor of free market capitalism. This process was particularly evident in Western Europe, where local noncommercial stations were legalized in the early 1980s, followed by commercial stations in the mid- to late 1980s (e.g., Belgium, Denmark, W. Germany). Although competition was permitted at the local level, in many cases public service radio retained a privileged, monopoly-like position at the national level through the late 1990s (e.g., Britain, Finland).

After the breakup of the Soviet Union and the Eastern Bloc in the late 1980s, most of the world's radio systems faced liberalization of some kind. Much of Eastern Europe became an unregulated market open to competition from North American and European companies (although, in fact, development remained limited due to severe economic dislocation in the region). In Africa most countries opened their radio systems to commercial broadcasting, although commercial development was also highly circumscribed by economic conditions. South Africa, for example, began to liberalize its broadcasting system in the mid-1990s with the privatization of some government-owned stations, licensing of new commercial stations, and the establishment of community broadcasting. Only in the Arab world did monopoly systems remain largely unscathed.

What began as liberalization in the 1980s became a move toward market regulation in the 1990s. As commercial broadcasters grew and faced local competition, they moved toward networking, syndication, and consolidation at a national level. In Sweden, for example, although the Radio Act of 1996 prohibited multiple ownership "all but a few of the more than 90 commercial stations [were] closely integrated into four more or less national networks" (Euromedia, 1997: 218). A similar process took place in France where a political movement in opposition to the public broadcasting monopoly was ultimately exploited by commercial interests. Whereas the 1980s had been dominated by small, activist stations, the 1990s witnessed the "oligopolization of the private radio sector in France" at the hands of professional broadcasting corporations (Kuhn, 1995: 107).

The move toward market regulation also produced oligopoly in the U.S. radio industry. Beginning in the early 1980s, deregulation removed federal content regulations and reduced the limits on cross-ownership of media. With the Telecommunications Act of 1996 all national limits on station ownership were removed, opening the door to a series of mergers and consolidations. By the year 2000 radio broadcasting had moved from monopolistic competition to oligopoly: many major cities had one station owner with 40% of the market and some had two owners with 70% of the market. According to Charles Fairchild (1999) the 1996 deregulation aimed, in part, to make radio competitive with other consolidated media industries, thus preparing the radio system for exploitation by integrated media conglomerates. At

the same time, developments such as Internet radio and digital audio broadcasting provided ideal media platforms for the convergence of recorded music, text, and image through a real-time broadcasting style format.

Radio reached almost every person on the globe by the end of the twentieth century. Radio listening continued to grow in the world's expanding urban areas, primarily in the form of in-car and portable radio listening. Radio listening also grew in rural areas due to electrification projects as well as the development of low-tech, wind-up radio receivers that finally made radio affordable in areas without electricity. Radio's popularity also increased during this period as broadcasters exploited the potential interactivity and intimacy of the medium. Listener-participation talk shows reinvigorated the medium and promoted public discussion of personal topics (especially personal relationships) as well as political issues in countries as different as China, Mexico, and the United States.

In the growing commercial sector, radio stations delivered increasingly standardized program formats to targeted market segments. Format-specific radio chains developed a 'brand value' or 'lifestyle statement' that could be "applied profitably to other aspects of their listeners' lives," including other forms of media consumption (Hendy, 2000: 47). Convergence and conglomeration in radio and related media industries (recorded music, television, Internet), as well as the growth of satellite delivery systems, set the stage for transnational deployment of increasingly 'branded' radio services. Although linguistic diversity and national regulatory policies continued to pose obstacles, regional radio integration was feasible in a few linguistically and economically unified regions (e.g., the Southern Cone of Latin America and the European Economic Community).

In contrast, the last decade of the twentieth century was also a boom period for decentralized community-based radio broadcasting. Community Radio broadcasting, which emphasized community control and participation, was nurtured by democratization movements around the globe. The World Association of Community Radio Broadcasters provided a venue for the organization and promotion of community radio and lobbied for 'social enterprise' broadcasting as a means of encouraging economically viable community stations. Just such a model of community radio was initiated in South Africa in the second half of the 1990s, where nonprofit community broadcasters were allowed to support themselves through advertising. These community stations, some of which served populations of over 50 000, provided local service where none had existed before and covered events such as the local Truth and Reconciliation hearings. By 2000, despite the vocal protests of the commercial broadcasting industry, even the United States Federal Communication Commission had approved noncommercial low-power FM broadcasting at the community level.

Radio in the First Decades of the Twenty-First Century

Radio's relevance as a cheap, mobile, and 'eyes-free' medium jumped during the first decade of the twenty-first century as a number of crises (natural and human-made) ratcheted up the communication needs of ordinary people. After the Terrorist Attacks of 11 September 2001, Americans turned increasingly to radio for timely international news and talk shows that provided conservative (and for a short-time liberal) interpretations of 9/11 and its aftermath (Sherman, 2005). The Indian Ocean Tsunami in 2004, Hurricane Katrina in 2005, and numerous large-scale disasters across the globe in the first and second decades of the millennium demonstrated the importance of radio broadcasting as a means of information sharing and logistical coordination at the local, regional, and national levels.

Such emergency conditions, combined with broader trends toward economic liberalization and social democratization, sped the development of community radio on a global scale. Although community broadcasters still faced political and economic barriers – and at times violent opposition – under certain regimes, the sector grew in the 2000s both in terms of the number of community stations broadcasting and the number of countries that legalized the service (Buckley, 2008). The role of community radio in the protests of 2010–11 that came to be known as the 'Arab Spring' has yet to be thoroughly documented, and has been largely overshadowed by the focus on digital and social media. There is no doubt, however, that community radio has expanded, and will continue to expand, in the Arab world in the 2010s. Research will be needed to explore the interplay of traditional broadcast and new digital media in the context of societies undergoing rapid political and social change.

Despite increasing competition from a range of digital media platforms, U.S. commercial radio grew in the 2000s and promised continued growth in the 2010s. One forecaster attributed this growth to national advertisers' recognition of the continued relevance of radio, particularly in urban markets, and radio stations' aggressive pursuit of online advertising revenues (BIA/Kelsey, 2011). Although the future of traditional commercial radio remains uncertain, listeners continue to tune in to its national, regional, and local content, and its format of mass intimacy and anonymous interactivity.

See also: Broadcasting: Regulation; International Communication: History; Media Effects; Media Imperialism; Media Talk Shows; Television: History.

Bibliography

Anderson, B., 1991. Imagined Communities. Verso, London.

Barthes, R., 1990. The grain of the voice (1977). In: Firth, S., Goodwin, A. (Eds.), On Record: Rock, Pop, and the Written Word. Routlege, New York, pp. 293–300.

BIA/Kelsey, 2011. BIA/Kelsey Reports Radio Industry Revenues Rose 5.4% to $14.1 Billion in 2010, Driven by Political Season and More Activity by National Advertisers. BIA/Kelsey, Chantilly, VA.

Buckley, S., 2008. Community Broadcasting: Good Practice in Policy, Law and Regulation. UNESCO.

Dabbous, S., 1994. In: Kamalipour, Y.R., Mowlana, H. (Eds.), Egypt. Mass Media in the Middle East. Greenwood Press, Westport, CT.

Douglas, S.J., 1999. Listening In: Radio and the American Imagination. Times Books, New York.

Euromedia Research Group (Ed.), 1997. The Media in Western Europe. Sage, London.

Fairchild, C., 1999. Deterritorializing radio: deregulation and the continuing triumph of the corporatist perspective in the USA. Media, Culture & Society 21, 549–561.

Hayes, J.E., 2000a. Radio. In: Balderston, D., González, M. (Eds.), Encyclopedia of Contemporary Latin American and Caribbean Cultures, vol. 1. Routledge, London, pp. 1234–1238.

Hayes, J.E., 2000b. Radio Nation: Communication, Popular Culture and Nationalism in Mexico, 1920–1950. University of Arizona Press, Tuscon.
Hendy, D., 2000. Radio in the Global Age. Polity Press, Malden, MA.
Hilmes, M., 1996. Radio Voices: American Broadcasting, 1922–1952. University of Minnesota Press, Minneapolis, MN.
Hong, L., 1994. Radio revival in China. Beijing Review 15–18.
Kamalipour, Y.R., Mowlana, H. (Eds.), 1994. Mass Media in the Middle East. Greenwood Press, Westport, CT.
Katz, E., Wedell, G., 1977. Broadcasting in the Third World. Harvard University Press, Cambridge, MA.
Kuhn, R., 1995. Radio. The Media in France. Routledge, London, 77–108.
Lent, J.A. (Ed.), 1978. Broadcasting in Asia and the Pacific. Temple University Press, Philadelphia, PA.
McDaniel, D.O., 1994. Broadcasting in the Malay World. New Jersey Ablex, Norwood.
Miller, T., 1997. Radio. In: Cunningham, S., Turner, G. (Eds.), The Media in Australia. Allen & Unwin, St. Leonards, NSW, pp. 47–69.
Ninan, S., 1998. History of indian broadcasting Reform. In: Price, M.E., Verhulst, S.G. (Eds.), Broadcasting Reform in India. Oxford University Press, Delhi, pp. 1–21.
Ong, W., 2012. Orality and Literacy: The Technologizing of the Word. Routledge, London, 73.
Ritchie, S.E., 1999. The Impact of Community Radio Broadcasting on Rural Development in South Africa. Radio Africa '99: Proceeding of the Third Regional Workshop on Radio Communication in Africa. University of Botswana, Gaborone, Botswana.
Scannell, P., Cardiff, D., 1991. A Social History of British Broadcasting. Basil Blackwell, Oxford.
Sherman, S., May 2005. Good, gray National Public Radio. The Nation 23, 34–40.
Strauss, N. (Ed.), 1993. Radiotext(e). Semiotext(e). Columbia University, New York.
United States Department of Commerce, 1932. Radio Markets of the World. Government Printing Office, Washington.

Random Assignment: Implementation in Complex Field Settings

Robert G St. Pierre, Abt Associates Inc., Breckenridge, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 12731–12734, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Random assignment of units to treatment/control groups allows researchers to make the strongest possible causal connection between the treatment and observed outcomes. This article considers the way in which random assignment is implemented in educational and social field settings. Topics include evidence about the widespread acceptance of randomized experiments, the purpose of random assignment, common objections to randomized studies, conditions that make program implementers more likely to accept random assignment, methods of successfully implementing random assignment, and how to preserve randomization once it has been achieved.

Widespread Acceptance of Random Assignment

The last three decades of the twentieth century have seen widespread acceptance of random assignment as an important aspect of high-quality research studies designed to test the effectiveness of interventions. In the 1970s and 1980s, proponents of randomized designs for studying program impacts began to assemble a literature attesting to their feasibility and frequent use (Boruch, 1974; Boruch et al., 1978; Boruch and Wothke, 1985). By the 1990s, randomized experiments in the social sciences had become so numerous that Greenberg and Shroder (1997) prepared a 500-page compendium of recent experiments in various content areas. The US Congress now routinely specifies that random assignment be used in studies of the effectiveness of federally funded programs, and recently called for a random assignment study of Head Start, America's flagship social program for disadvantaged children (Advisory Committee on Head Start Research and Evaluation, 1999).

Randomized experiments have limitations as research vehicles (see Cook and Campbell, 1979 for a discussion of the strengths and weaknesses of randomized experiments), and they are not universally accepted as the best way to learn about program effectiveness. One sector of the evaluation community espouses the virtues of qualitative research in place of randomized experiments and other quantitative approaches (e.g., Guba and Lincoln, 1982), while another sector is more interested in issues of generalizability than internal validity (e.g., Cronbach, 1982; Heckman and Smith, 1995). Cook (1999) has written about the antirandom assignment culture that exists in American schools of education. Still, there is no denying that American society relies heavily on randomized experiments as one source of information about the effectiveness of social programs (St. Pierre and Puma, 2000).

The Purpose of Random Assignment

As long as researchers take the time to explain the strengths and weaknesses of random assignment studies, those who commonly question experiments are easily able to understand the rationale underlying random assignment. Perhaps most helpful is the argument that random assignment is a special purpose tool, and while it is good only for one purpose, for making causal inferences, it is far better for that purpose than any alternative. As a special purpose tool, random assignment has many limitations. For example, it does not assure group comparability over time (attrition from the evaluation sample may result in nonrandom analytic groups), it cannot ensure that subjects remain in the intervention being studied (with or without random assignment, interventions may incur high dropout rates), and it does not help with issues of generalizability (a one-site random assignment study is no more generalizable than a one-site quasiexperiment). Cook and Campbell (1979) stated it clearly: "Though random assignment is germane to the research goal of assessing whether the treatment may have caused any observed effects, it is conceptually irrelevant to all other research goals" (p. 343).

Common Objections to Random Assignment

Most of the common objections to random assignment in studies of program effectiveness do not dispute its utility as a tool for understanding causal connections. Instead, objections take other forms. Some are made in an attempt to avoid participation in a research study. Take the oft-made claim that 'random assignment is unethical' or that 'random assignment means denying services to some subjects.' These are political statements, either made without understanding what random assignment means or made in an attempt to retain control over which subjects are to be served by a given social program.

When there is an excess demand for a social intervention, program operators are in the position of denying services to needy families every day as they make decisions about which families are the neediest or the most deserving. As long as a randomized experimental study of such an in-demand program does not reduce the total number of subjects served, then there is no denial of services – the same aggregate number of families is served with or without an experiment. What an experiment does change is the mechanism for selecting which families receive the services; random assignment is used to decide who participates in the program instead of the judgment

of program staff. Thus, when program operators say that random assignment is unethical or involves denying services, they usually are saying that they do not want to give up the power to determine which families are served.

But what about denying services to families? Is this always unethical? Suppose a funding agency wants to learn about the effectiveness of one of its social programs and in the absence of evidence of effectiveness, the funding agency plans to withdraw its support. Suppose further that there is no waiting list and hence that a randomized experiment would make it impossible to maintain aggregate service levels since some families would be assigned to a control group. In this case, the ethical problem of doing an experiment which reduces the number of families serviced by a social program has to be weighed against the possibly greater ethical problem of not doing an experiment and risking reduced funding for the entire program.

An underlying, and typically unvoiced, objection to random assignment is that program staff or funders assume the effectiveness of social programs instead of worrying about whether scarce resources are being spent in the most appropriate manner. This is understandable, since scientific skepticism seems to counteract the ability of program operators to do their best work. But again, this is a political issue – that it is appropriate to assume the effectiveness of a social program instead of subjecting it to a rigorous test.

Another objection to random assignment is that assignment to a control group 'feels bad' to program staff and to the families they serve. The argument that randomization is like a lottery does not work in many social program settings. Why? Because a lottery usually has only one winner, and the losers can take solace in the fact that they are in the great majority. On the other hand, those assigned to the control group in a randomized study do not view their assignment the same way they view a lottery loss, since the odds of winning the lottery are miniscule, whereas the odds of being randomly assigned to the desired social program are usually 50/50.

Some other arguments are that randomized experiments are expensive (Levitan, 1992) or that they take a long time. The response to this sort of complaint is, "Compared to what?" A randomized experiment is expensive and time consuming when compared with the alternative of doing no research at all. But this is rarely a realistic alternative. Many quasiexperimental alternatives are just as expensive (any design in which data are collected from a treatment group and a nonequivalent comparison group) and time consuming (the common pre–post study) as randomized studies. Some alternative approaches are quick and inexpensive, for example, relying on data collected for other purposes. However, the fact is that none of these alternatives are capable of doing the one thing that randomized studies excel at – providing strong evidence about causal connections. The problem with using alternatives to random assignment to study program effectiveness is that even if they are less expensive, even if they require less time, even if they are easier to implement, they rarely are worth the time and money spent on them. Why? Because the results are not helpful for their intended purpose – for understanding program effectiveness. To paraphrase the common aphorism about education, "If you think randomized studies are expensive, try ignorance."

Conditions Making Random Assignment Easier

There are several circumstances that reduce the difficulty of implementing random assignment studies. Suppose an intervention program is in short supply and there is a list of subjects waiting to participate. This often happens with Head Start (the largest federal preschool program in the United States) projects in large metropolitan areas. If being placed on a waiting list means that a family is likely to miss a substantial portion of the Head Start year, then randomly assigning families from the waiting list to be in the program or in a control group might well be preferable to being on the waiting list with certainty. The absence of a waiting list does not mean that subjects are uninterested in a program. It often means that program staff do not want to spent time and energy on recruiting until they have openings. Hence, it is possible in short-supply situations for researchers to generate a waiting list by helping to recruit subjects and informing them about the planned study.

Suppose a treatment can be delayed with no ill effect on the participants. This is likely the situation in a school that wishes to test the short-term effectiveness of a one-semester school health curriculum. In this case, it would be possible to randomly assign students to take the school health course in the fall semester or in the spring semester.

Suppose a program's proponents are skeptical and are searching for the best solution to a social problem instead of trying to promote a specific intervention. Or, suppose a demand for evidence of program effectiveness cannot be ignored, for example, a congressional mandate. In these cases, political objections to random assignment are eased because program implementers are supportive of or at least are willing to cooperate with the research.

Suppose that a social program has many local projects, for example, the Head Start program has about 1600 local projects and the Even Start program (a family literacy program) has about 800 projects. An experimental evaluation of these programs could be done be selecting a relatively large number of projects and randomly assigning only a few subjects within each project. This approach minimizes the unpleasant effects of random assignment for any given project, making it more palatable to local program staff.

Assigning higher-level units is another way to ease the pain of random assignment. Suppose that there is an interest in testing three different solutions to a given social problem. One way to construct this test is to randomly assign individuals to one of three interventions implementing the three different solutions. A different approach which avoids many of the common objections to random assignment is to randomize groups of individuals. For example, a federal agency could hold a grant competition to test three different approaches to preschool education. School districts would not be allowed to receive a grant unless they indicated their willingness to implement any of the three approaches. School districts that receive grants would then be randomly assigned to implement one of the three preschool approaches.

Working Toward Correct Random Assignment

Random assignment is more likely to be successfully implemented if done by researchers and research organizations with

prior experience in conducting experimental studies. Certain research organizations have built the capacity to and reputation for conducting random assignment studies. Greenberg et al. (1999) report that almost half of the social experiments started in the US since 1983 have been done by three large research organizations.

In the same vein, random assignment is more likely to be successful if it is under the control of researchers rather than program implementers (Conner, 1977). The latter are likely to make exceptions or may misunderstand random assignment rules, even if those rules are carefully prepared by researchers. The research team does not have to be physically present in order to control the random assignment. One approach is to have program staff be responsible for recruiting subjects, explaining the experimental alternatives, and transmitting lists of study subjects to the research team via fax, email, or other means. The research team then does the random assignment and sends back listings of research subjects, sorted into treatment and control groups. A related approach is for the research team to prepare a random assignment computer program for use by program staff who recruit subjects and enter basic data on a microcomputer which does the random assignment on the spot.

The way in which random assignment is implemented depends on the way in which applicants are recruited for an intervention. Sometimes a large pool of applicants are recruited. In these cases, information on the entire pool can be sent to the research team at the same time, simplifying the random assignment. The process is more complicated when applicants enter a program on a rolling basis. In these cases, a random assignment algorithm has to be developed so that applicants are told their treatment/control status soon after they apply. However, simple assignment rules such as assigning the first applicant to the treatment, the second to control, the third to treatment, etc. are easily manipulated by program staff who often have control over the flow of applicants. More complex assignment systems are less subject to local control. For example, random assignment of families to an intervention could be done in blocks of four, where the order of treatment and control assignment is random within each block.

Preserving Randomization

Implementing random assignment is difficult. Once achieved, how can it be preserved throughout the life of an evaluation? One issue has to do with when randomization takes place. This involves tradeoffs between generalizability and internal validity. Families interested in participating in the Even Start family literacy program could be randomized as soon as they are first recruited for the program, after they have had the evaluation explained to them and have indicated their willingness to accept any of the alternatives, or after they have completed a 1-month 'tryout period' during which they can try Even Start and see whether they truly are interested in the program. Early randomization draws on the largest possible pool of families (increased generalizability) but risks losing many families (decreased internal validity) who find out that they are not assigned to the desired group or who leave the program after a tryout period. Late randomization winnows the pool of families (decreased generalizability) to those who are most likely to participate in the study (increased internal validity). Gueron and Pauly (1991) and Rieken et al. (1974) give additional examples.

Whether assigned to an intervention or a control group, subjects may refuse to cooperate with research measurements once an experiment has been started. Such attrition has the potential to destroy the comparability of the intervention and control groups. Clearly, it is best to use all means possible to avoid attrition of subjects. In some circumstances (e.g., studies of welfare programs), subjects can be sanctioned if they do not participate in research; in other cases, incentives for participation can be offered. Attrition is all but unavoidable in most social experiments since families move, are ill or otherwise unavailable for measurement activities, or simply refuse to participate in the research. Methods for determining whether attrition is related to the treatment and ways of dealing with this problem are discussed by Cook and Campbell (1979).

Are there ways to reduce the likelihood that subjects will refuse to go along with the randomization, or refuse to participate in the study? One approach is to plan the study so that subjects who agree to participate in the randomization have a better chance of getting the desired intervention than they would have if the study had not existed. This works best when there is a waiting list and when the intervention has a large number of openings relative to the number of subjects needed for the study. A second approach is to offer a competing intervention to members of the control group. This could work in several different ways. One competing intervention could be a financial or in-kind incentive for participating in the study. Another competing intervention could be a reduced or modified version of the treatment being studied. Yet a third alternative could be to offer control group subjects the opportunity to enroll in the desired intervention at a later point in time.

See also: Causal Inference: Overview; Comparative Studies: Method and Design; Experimental Design: Bayesian Designs; Experimental Design: Compliance; Experimental Design: Large-Scale Social Experimentation; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Latent Structure and Causal Variables; Mechanism-Based Causal Analysis; Nonequivalent Group Designs; Observational Studies: Overview; Optimal Experimental Design; Quasi-Experimental Designs; Regression Discontinuity Design; Social Experiments.

Bibliography

Advisory Committee on Head Start Research and Evaluation, 1999. Evaluating Head Start: A Recommended Framework for Studying the Impact of the Head Start Program. US Department of Health and Human Services, Washington, DC.

Boruch, R., 1974. Bibliography: illustrative randomized field experiments for program planning and evaluation. Evaluation 2, 83–87.

Boruch, R., McSweeny, A.J., Soderstrom, E.J., 1978. Randomized field experiments for program planning, development and evaluation. Evaluation Quarterly 2, 655–695.

Boruch, R., Wothke, W., 1985. Randomization and field experimentation. New Directions for Program Evaluation 28, 39–53.

Conner, R.F., 1977. Selecting a control group: an analysis of the randomization process in twelve social reform programs. Evaluation Quarterly 1, 195–244.

Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs, first ed. Jossey-Bass, San Francisco.

Cook, T.D., 1999. Considering the major arguments against random assignment: an analysis of the intellectual culture surrounding evaluation in American schools of education. Paper Presented at the Harvard Faculty Seminar on Experiments in Evaluation, Cambridge, MA.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design & Analysis Issues for Field Settings. Rand McNally College Pub. Co., Chicago.

Greenberg, D., Shroder, M., 1997. The Digest of Social Experiments, second ed. The Urban Institute Press, Washington, DC.

Greenberg, D., Shroder, M., Onstott, M., 1999. The social experiment market. Journal of Economic Perspectives 13, 157–172.

Guba, E.G., Lincoln, Y.S., 1982. Effective Evaluation, first ed. Jossey-Bass, San Francisco.

Gueron, J.M., Pauly, E., 1991. From Welfare to Work. Russell Sage Foundation, New York.

Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. The Journal of Economic Perspectives 9, 85–110.

Levitan, S.A., 1992. Evaluation of Federal Social Programs: An Uncertain Impact. George Washington University, Center for Social Policy Studies, Washington, DC.

Rieken, H.W., Boruch, R.F., Campbell, D.T., Coplan, W., Glennan, T.K., Pratt, J., Rees, A., Williams, W., 1974. Social Experimentation: A Method for Planning and Evaluating Social Innovations. Academic Press, New York.

St. Pierre, R.G., Puma, M.J., 2000. Toward the dream of the experimenting society. In: Bickman, L. (Ed.), Advances in Social Research Methods: The Heritage of Donald Campbell. Sage Publications, New York.

Rape and Sexual Coercion

Peggy Reeves Sanday, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the biological and cultural approaches to the understanding of rape in human societies and evaluates these approaches by reference to studies demonstrating the incidence and social correlates of sexual coercion in the United States and cross-culturally. The biological, evolutionary view is compared and contrasted with the cultural approach proposed by anthropologists. The wide variation in the response to rape cross-culturally and intraculturally in the United States suggests that whatever the biological basis of male sexual aggression might be, family background and cultural context make a difference.

The Biological View

The biological explanation for rape and sexual coercion is grounded in the argument that sexual aggression is an evolved adaptation in human males. The argument builds on Charles Darwin's theory of natural and sexual selection. Natural selection is the doctrine that in the struggle for existence evolutionary progress is achieved by the inheritance of advantageous characteristics that prosper at the expense of less advantageous ones. Sexual selection is a specific case of natural selection applied to the evolution of sex differences. Darwin claimed that the greater size and strength of human males was due "in chief part to inheritance from his half-human male ancestors." According to Darwin, such characteristics "would have been preserved or even augmented during the long ages of man's savagery, by the success of the strongest and boldest men, both in the general struggle for life, and in their contest for wives." He measured success by the number of children some men left as compared with their "less favored brethren" (Darwin, 1936: pp. 872–873).

In the nineteenth and early twentieth centuries, the Darwinian doctrine of sexual selection was applied to treatises on human sexual behavior produced by the famous sexologists R.F. von Krafft-Ebing and Havelock Ellis. In his magnum opus, *Psychopathia Sexualis*, first published in German in 1886, Krafft-Ebing argued that "gratification of the sexual instinct [is] the primary motive in man as well as in beast" (quoted by Sanday, 1996: p. 125). In the third volume of his *Studies in the Psychology of Sex*, published in Philadelphia between 1897 and 1910, Ellis comes very close to making rape the *sine qua non* of sexual arousal. Ellis conceives of human sexual behavior as a game of combat. Playing the role of hunted animal the female conceals her sexual passion by adopting a demeanor of modesty in order that the male may be more ardent and forceful. Ellis claims that as the hunt becomes more sexually charged, "an element of real violence, of undisguised cruelty" is introduced. Accepting the Darwinian theory of natural selection with its emphasis on competition and brute strength, Ellis claimed that a woman who resisted "the assaults of the male" aided natural selection "by putting to the test man's most important quality, force" (quoted by Sanday, 1996: pp. 127–128).

At the end of the twentieth century, the Darwinian approach was again applied to the subject of the 'biological bases of sexual coercion' by Randy Thornhill and Craig T. Palmer in their book *A Natural History of Rape* (2000). Like Krafft-Ebing and Ellis, they make a distinction between sexual desire in human males and females. Citing the work of anthropologist Donald Symons (1979: pp. 264–267), these authors claim that rape is due to certain adaptations making males more easily aroused, with a greater sex drive, a reduced ability to abstain from sexual activity, in need of greater sexual variety, and less discriminating in choice of sexual partner (Thornhill and Palmer, 2000: p. 62). However, while Symons suggests that rape is a by-product of such adaptations, Thornhill and Palmer suggest that rape is itself an adaptation, which they call "the human rape-adaptation" (Thornhill and Palmer, 2000: p. 62). By this they mean that rape has a direct biological basis rather than being a behavioral consequence of other evolved behaviors.

To support their argument, Thornhill and Palmer (2000: p. 62) say it is necessary to identify "mechanisms involved in rape that were designed by selection in the past specifically for reproduction by means of rape." Because such mechanisms exist in the morphology of certain insects, they seem to suggest, they must also exist in humans. To make their case, the authors draw parallels with scorpion flies, which they point out have a physical adaptation for rape. Male scorpion flies have an organ, "a pair of clamp-like structures, one on either side of the penis," that serves to keep unwilling females in a mating position (2000: p. 63). Admitting that there is no "conspicuous morphology that might be a rape adaptation" in human males, Thornhill and Palmer (2000: pp. 64–65) look "to the male psyche for candidates for rape adaptations." They propose (2000: pp. 65–66) a number of 'psychological mechanisms' as 'adaptations' that can be viewed as 'analogous' to the rape adaptations observed in 'male insects.'

In a review of the argument put forward by Thornhill and Palmer, Frans B.M. de Waal, a prominent primatologist, points out that they have not demonstrated that rape in humans is a product of natural selection. According to de Waal, "for natural selection to favor rape, rapists would have to differ genetically from nonrapists and need to sow their seed more successfully, so to speak, causing more pregnancies than nonrapists, or at least more than they would without raping"

(de Waal, 2000: p. 24). "Not a shred of evidence for these two requirements is present" de Waal concludes (2000: p. 24). de Waal also points out that the psychological mechanisms Thornhill and Palmer propose do not demonstrate natural selection but have to do "with judgment of people and situations, a multi-purpose capacity also present in women" (2000: p. 24). de Waal's emphasis on judgment illuminates the fact of variation in the incidence of rape, which turns attention to the role of culture.

Early in the twenty-first century, Cheryl Brown Travis, Professor of Psychology and Chair of Women's Studies at the University of Tennessee, published responses to the Thornhill and Palmer argument authored by prominent anthropologists, evolutionary biologists, ecologists, philosophers, primatologists, psychologists, sociologists, and women's studies scholars. This was the first and most up-to-date volume providing multidisciplinary responses to essentialist biological claims regarding male rape from both a cultural and biological point of view. In her introductory article, Travis (2003: p. 4) points out that if evolutionary theory is to be used to understand gender differences and sexual aggression in particular, "it is necessary to take a considered look at basic principles of evolution and it is necessary to examine assumptions about gender in general." "Evolutionary theory is not sexist," she says. However, it can become so when applications of evolutionary principles to gender relations are characterized by grandiose overgeneralization and a politically charged gender philosophy. Travis (2003: p. 4) argues instead "that there should be a broad understanding of gender and of sexuality as phenomena that are as much social as biological." This is a point with which the other authors in the book agree. Claiming that Thornhill and Palmer's book represents bad science, bad history, and bad politics, one of the authors, Michael Kimmel (2003: p. 222), concludes that the book "tells us less about 'the biological bases of sexual coercion' than the ideological fantasies of those who justify sexual coercion."

The Cultural Approach

Against a strictly biological explanation of human behavior anthropologist Clifford Geertz (1973: p. 46) argues that without cultural patterns (i.e., 'organized systems of significant symbols') human behavior would be shapeless and without direction, "a mere chaos of pointless acts and exploding emotions." Applied to sexual behavior, the cultural approach, accepted by many anthropologists and sociologists, suggests that sexual behavior is part of a broader cultural system that includes notions of masculinity and femininity, ideas about appropriate sexual relations, cues for sexual arousal, and messages about the consequences of sexual expression. As anthropologist Malinowski said long ago, "[s]ex in its widest meaning … is rather a sociological and cultural force than a mere bodily relation of two individuals" (Malinowski, 1929: p. xxiii).

Peggy Reeves Sanday (1981) puts Malinowski's claim to the test in a study of the sociocultural context of rape cross-culturally. Using a standard sample of band and tribal societies, she reports that 47% of the societies for which there was adequate information showed a low frequency of rape, compared with 18% that were unambiguously 'rape prone.' Commenting on the pattern of correlations revealed by her study, Sanday (1981: p. 5) concludes that rape by males is part of "a cultural configuration," which includes interpersonal violence among men as well as between men and women and male social dominance. Rape-prone societies are more likely to be characterized by male segregation in separate houses, an ideology of male toughness, emphasis on competition, low respect for women as citizens, and the absence of women in the public domain of economic and political affairs. In the more rape-free societies, there is an ethos of cooperation and consensus in human affairs. The social separation of the sexes is less marked and both sexes are more likely to hold exalted positions in public decision making and to be integrated and equal in the affairs of everyday life. These findings support Malinowski's observation, based on his fieldwork among the Trobriand Islanders of Papua New Guinea, that the expression of sexual behavior is consistent with social forms and cultural meanings.

Anthropologist Carole Vance (1984: pp. 7–9) also cites the diversity in human sexual practices to suggest that biology does not "directly or simply determine the configuration or meaning of sexuality." If this were the case, she argues, one would encounter uniformity cross-culturally rather than the startling diversity that actually exists in which "activities condemned in one society are encouraged in another, and ideas about what is attractive or erotic, or sexually satisfying or even sexually possible vary a great deal." Vance concludes that "the body and its actions [must be] understood according to prevailing codes of meaning." She suggests that an important question for research must be "What is the nature of the relationship between the arbitrariness of social constructions and the immediacy of … bodily sensations and functions?" For similar points, see articles in Caplan (1987).

In two books, Sanday expands the understanding of the cultural foundation of acquaintance rape in the United States by examining fraternity gang rape on college campuses (1990) and the legal and historical contexts of rape in the United States (1996). Sanday (1996: p. 26) coins the term sexual culture to describe the system of meanings, standards, and expected behaviors by which male and female sexuality is judged and understood. She argues that in the United States, theories like those of Darwin, Krafft-Ebing, and Ellis functioned as cultural templates for popular sexual stereotypes by means of which males and females judged appropriate sexual behavior and acted accordingly. Because these stereotypes condoned sexual aggression by making it 'natural,' such as seen in the idea that 'boys will be boys,' US rape laws applied primarily to the use of force by a stranger. This changed after feminist activism in the 1970s resulted in more stringent legal codes. The same activism initiated studies of the incidence and social correlates of rape on US college campuses showing that acquaintance rape was much more common than stranger rape (Sanday, 1996: pp. 161–207). In her ethnographic study of "fraternity gang rape," Sanday (1990) demonstrates how alcohol is used to "hit on" young women guests in the "hooking-up" sexual culture prominent on college campuses that snares unsuspecting young women.

Rape Statistics in the United States

US statistics come from numerous studies utilizing national, community, or college samples. In general, these studies define rape as nonconsensual sexual intercourse irrespective of whether the parties know one another. Nonconsent is defined using the modern legal definition of rape. In most states, rape is defined as nonconsensual sex either due to the use of force or taking advantage of a person's inability to consent because of intimidation or because of alcohol or drugs. Although most studies of the past focused on heterosexual rape, most now include same-sex rape as well.

The first study designed to measure the percentage of those claiming they had been raped in a specified sample population was conducted in 1957 by sociologist Eugene Kanin on a college campus. This survey revealed that 55% of the 291 college women interviewed said they had experienced offensive episodes "at some level of erotic intimacy." Twenty-one percent said they were offended "by forceful attempts at intercourse" and 6.2% by "aggressively forceful attempts at sex intercourse in the course of which menacing threats or coercive infliction of physical pain were employed." Of the latter group, 48% told no one and none reported to anyone in authority (Kirkpatrick and Kanin, 1957: pp. 53, 56).

Another well-known path-breaking study was conducted by Diane Russell in 1978 in San Francisco. Russell interviewed 930 women ranging in the age from 18 to 80 years of diverse social classes and racial/ethnic groups. Twenty-four percent of the women interviewed reported at least one completed rape, and 31% reported at least one attempted rape. The majority experienced rape by acquaintances and a much smaller percentage by strangers (Russell, 1984: pp. 34–38).

Two large-scale national studies yielded similar results. In 1985, Mary Koss joined with *Ms*. magazine to survey a national sample of 6159 students on 32 college campuses. This study found that many more women and men were having sex in the 1980s compared to the figures reported by Alfred Kinsey in his landmark studies of male and female sexual behavior published in the 1940s and 1950s. One-quarter of the men interviewed in the Koss–*Ms*. study reported involvement in some form of sexual aggression, ranging from unwanted touching to rape. A high percentage of the males did not name their use of force as rape. Eighty-eight percent said it was definitely not rape. Forty-seven percent said they would do the same thing again (summarized by Sanday, 1996: pp. 191–192; see also Koss, 1988, 1992; Warshaw, 1988).

One in four of the women surveyed by Koss said yes to questions asking about behaviors that in most states would be legally defined as rape or attempted rape (for this definition see above). The power of popular stereotypes for defining acceptable sexual behavior and the adversarial nature of these stereotypes is reflected in the fact that most of the women reporting incidents that would legally qualify as rape did not call it rape. The term 'date rape' evolved from this study because many of the women surveyed said that the behavior often occurred on dates. Although most did not call it rape, Koss reported that "the great majority of rape victims conceptualized their experience in highly negative terms and felt victimized whether or not they realized that legal standards for rape had been met" (Koss, 1992: pp. 122–126).

By the end of the 1990s, a number of scientifically designed research studies revealed similar statistics (for a summary of studies, see Koss, 1993; Sanday, 1996: pp. 184–206). The findings reported in these studies are corroborated by a national study of sex in America conducted in 1992 employing a representative probability sample of 3432 Americans (see Michael et al., 1994). The findings fall within the range of many other studies: "22 percent of women were forced to do something sexually at some time," while "just 2 percent of men were forced." The authors of this study concluded that "as many as one in five women" experience some form of sexual coercion, including rape, in the United States (Michael et al., 1994: pp. 33, 221–222).

During the first decade of the twentieth century, despite the attention given to 'date' and 'acquaintance rape' in the l990s in the sexual assault policies adopted on many campuses, acquaintance rape statistics did not change. Verbal consent and a partner's ability to withdraw consent continue to be more debated than observed. Alcohol use continued to play a role in the sexual culture of acquaintance rape. Rape statistics reported by various colleges and universities resembled those reported in the late twentieth century. One study, funded by the Department of Justice published in 2007, found that approximately one in five (19%) women surveyed experienced attempted or completed sexual assault during their years in college. Close to the same number (15.9%) reported attempted or completed sexual assault before entering college. Many of the college women said that the incident occurred during their first or second year and was tied to alcohol. A large number also reported that they were at a party when the incident happened (Krebs et al., 2007).

Some campuses now administer occasional surveys to monitor the local incidence of acquaintance rape. The University of New Hampshire (UNH), which is especially known for its support, prevention, and monitoring policies, is one of the few institutions to both conduct campus surveys at regular intervals and publish the findings. In its 2006 survey, UNH compared data for this year with survey results for 2000 and 1988 showing a slight downward trend. Unwanted sexual contact was reported by 35% of women surveyed in 1988 as compared with 20% of those surveyed in 2000 and 23% in 2006. Unwanted sexual intercourse was reported by 10% of women surveyed in 1988, 6% in 2000, and 5% in 2006. The perpetrator's use of alcohol was mentioned in the large majority of cases both of unwanted sexual contact and sexual intercourse for each of the three reported years: www.unh.edu/news/docs/2006unwantedsexualexperiences.pdf.

The relationship between alcohol and rape is evident in a study by the Harvard School of Public Health conducted during the years spanning the late twentieth and early twenty-first centuries. This study surveyed female students at 119 schools representing a national cross section of students enrolled at 4-year colleges. During the 3 years of the study (1997, 1999, and 2001), a total of 23 980 female students were questioned. Roughly 1 in 20, or 5% of these students, reported being raped since the beginning of the school year. Seventy-two percent experienced rape while intoxicated. These young women tended to be under 21, white, resided in sorority houses, used illicit drugs, drank heavily in high school, and

attended colleges with high rates of episodic (binge) drinking (Mohler-Kuo et al., 2004: p. 37).

It is now more common to include males in local collegiate surveys. In the 2006 survey conducted by the UNH, mentioned above, 10% of men reported experiencing at least one unwanted sexual contact; 4% unwanted sexual intercourse, and 8% (as compared with 11% of the women respondents) reported that they had sexual intercourse when they were too intoxicated to consent. The 2009 stats for a southeastern college, published internally about the same time but not for distribution, found that 8% of males reported sexual touching against their will, 3% attempted rape, and 2% reported being raped. This tendency to now include males reflects the growing recognition that sexual equity applies to both sexes with respect to issues regarding consent.

The Social Correlates of Rape in US Studies

A number of studies conducted late in the twentieth century indicate that the adversarial view of male–female sexual interaction described by the early sexologists is frequently held by sexually coercive males. In a study of a large southeastern university, Boeringer (1996: pp. 137–139) reports that 56% of the males he interviewed admitted to obtaining sex by verbal harassment (i.e., "threatening to end a relationship unless the victim consents to sex, falsely professing love, or telling the victim lies to render her more sexually receptive"). One-quarter of the males reported using drugs or alcohol to obtain sex and 9% reported at least one use of force or threatened force to obtain sex. Boeringer concludes that such verbally coercive tactics suggest an "adversarial view of sexuality in which one should use deceit and guile to win favors from a woman" (p. 140).

Sexually aggressive men, from convicted rapists to college males answering questions on social surveys, share a remarkably similar set of attitudes. Most believe that sexual aggression is normal, that sexual relationships involve game playing, that men should dominate women, that women are responsible for rape, and that relations between the sexes are adversarial and manipulative on both sides (Koss and Leonard, 1984: pp. 221, 223). Reanalyzing Koss' data to pinpoint attitudes held by the self-admitted sexually aggressive men of her study (i.e., those who admitted to forcing a woman to have sex in Koss' 1985 questionnaire), Sanday (1996: pp. 196–197) found that these men often expressed adversarial-like beliefs. For example, many stated that they believe that a woman's No means Yes and that women say no to intercourse because they do not want to seem loose but really hope the man will force her. Compared with men who do not admit to forcing a woman, more of these men also think that being roughed up by a man is sexually stimulating to women and that women have an unconscious wish to be raped. Few of the women in Koss' sample hold such attitudes.

Prevention

Examples of prevention strategies based on the biological approach are offered by Thornhill and Palmer (2000: pp. 169–88). Believing that rape occurs "in all the environments in which humans societies [sic] have been known to exist" (p. 171), they are less interested in changing the environment by challenging popular stereotypes than with focusing on the "exact nature of the psychological mechanisms that guide male sexual behavior" (p. 172). They offer a number of suggestions including certain types of educational programs. Citing the view of rape proposed by Camille Paglia, Thornhill and Palmer (2000: p. 183) claim that since men "have evolved sexual preferences for young and healthy women and are attracted to women who signal potential availability by means of dress and behavior," women should be informed of the risk factors and use the information to lower their risk of rape by paying attention, for example, to dress and appearance. They also suggest more direct social interventions such as separating vulnerable females from males.

The cultural approach argues for a change in rape laws and sexual stereotypes so that nonconsensual sex is treated unambiguously as a crime. Increasingly, this approach defines rape not just in terms of the use of force or drugs, but more broadly as failure to obtain verbal consent. The operative words of the new American sexual ideology are 'affirmative verbal consent,' 'freely given agreement,' and 'sexual communication.' Increasingly, this sexual ideology is supported publically by male public figures.

A notable example can be found in an address given in 2011 by Vice President Biden addressing an audience of some 600 students at UNH kicking off its annual 'sexual awareness month' now observed on many campuses.

> Look guys—all you guys in the audience—no matter what a girl does, no matter how she's dressed, no matter how much she's had to drink—it's never, never, never, never, never okay to touch her without her consent. This doesn't make you a man—it makes you a coward. A flat-out coward. (US vice president Joe Biden, 4 April 2011)

The tone of this talk and the events surrounding it marked another turning point in the cultural history of acquaintance rape. The US government is now playing a stepped-up public role. More men are participating in the struggle for sexual equity. Other developments suggest optimism for the future. Policies on many campuses are now in place to support students, prevent rape, and engage in oversight by measuring the incidence of acquaintance rape. These policies are supported by changes in the legal arena leading to more prosecutions of rape (see Sanday, 1996; for updated developments see Afterword in the 2011 Random House ebook edition).

The public, legal, and political actions to stem the incidence of acquaintance rape can be likened to Darwin's view of natural selection because they promote sexual health and well-being in an era of sexually transmitted diseases. One point left out of discussions of the biological bases of rape is the fact that as much as Darwin stressed competition, he also stressed cooperation or acting on behalf of what he called "the general good or welfare of the community" (quoted by Sanday, 1996: p. 288). Called 'altruism' by students of human evolution, some suggest that the primary selective pressure in human

evolution was not for violent men but for those who were able to cooperate with women and other males in organized food gathering. Some anthropologists believe that this ability to cooperate was responsible for the evolution of human culture (Tanner, 1981). In light of the ever-increasing incidence of sexually transmitted diseases together with the evolution of nuclear weapons, one can also note that today there is a fine line between aggressive behavior that results in reproductive success as argued by the biological approach and behavior that leads human populations to the brink of extinction.

Conclusion

There is considerable evidence suggesting that whatever the biological basis of male sexual aggression might be, family background, and cultural context make a difference. The incidence of rape is far lower among men raised in stable, peaceful environments be it a family context in a complex society like the United States or the intimate communities of many small-scale traditional societies in which the ethos is one of mutual respect between the sexes. To say that some men will rape in all societies and to use this fact to make generalizations about the sexual predilections of all men may bring about the very sexual culture one is trying to avoid by creating a lore that makes sexual aggression a 'natural' part of masculinity. More important for the education of young males is the conclusion based on studies in the United States and cross-culturally that most men in most societies do not rape, and those who do face social rejection or incarceration.

See also: Domestic Violence: Sociological Perspectives; Empirical Legal Studies; Feminist Legal Theory; Feminist Theory: Radical Lesbian; Gender and Feminist Studies in Geography; Gender and the Law; Heterosexism and Homophobia; Occupational Health; Perversions, Sexual (Paraphilias); Pornography; Prostitution and Sex Work; Repetitive Sex offenders, Treatment of; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Sex offenders; Sexual Harassment: United States and Beyond.

Bibliography

Boeringer, S.B., 1996. Influences of fraternity membership, athletics, and male living arrangements on sexual aggression. Violence Against Women 2 (2), 134–147.

Caplan, P. (Ed.), 1987. The Cultural Construction of Sexuality. Tavistock, London.

Darwin, C., 1936. The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life and the Descent of Man and Selection in Relation to Sex. The Modern Library, New York.

de Waal, F.B.M., 2 April 2000. Survival of the Rapist. NY Times Book Review, 24–25.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Kimmel, M., 2003. An unnatural history of rape. In: Travis, C.B. (Ed.), Evolution, Gender, and Rape. The MIT Press, Mass., pp. 221–334.

Kirkpatrick, C., Kanin, E.J., 1957. Male sex aggression on a university campus. American Sociological Review 22, 52–58.

Koss, M.P., 1988. Hidden rape: sexual aggression and victimization in a national sample of students in higher education. In: Burgess, A.W. (Ed.), Rape and Sexual Assault II. Garland Publishing, New York, pp. 3–25.

Koss, M.P., 1992. Defending date rape. Journal of Interpersonal Violence 7 (1), 122–126.

Koss, M.P., 1993. Rape: scope, impact, interventions, and public policy responses. American Psychologist 48 (10), 1062–1069.

Koss, M.P., Leonard, K.E., 1984. Sexually aggressive men: empirical findings and theoretical implications. In: Malamuth, N.M., Donnerstein, E. (Eds.), Pornography and Sexual Aggression. Academic Press, New York, pp. 213–232.

Krebs, Christopher P., Lindquist, Christine H., Warner, Tara D., Fisher, Bonnie S., Martin, Sandra L., 2007. The Campus Sexual Assault (CSA) Study. U.S. Department of Justice. http://www.ncjrs.gov/pdffiles1/nij/grants/221153.pdf.

Malinowski, B., 1929. The Sexual Life of Savages in North-Western Melanesia: An Ethnographic Account of Courtship, Marriage, and Family Life among the Natives of the Trobriand Islands, British New Guinea. G. Routledge & Sons, London.

Michael, R.T., Gagnon, J.H., Laumann, E.O., Kolata, G., 1994. Sex in America: A Definitive Survey. Little, Brown and Company, Boston.

Mohler-Kuo, Meichun, Dowdall, George W., Koss, Mary, 2004. Correlates of rape while intoxicated in a national sample of college women. Journal of Studies on Alcohol 65, 37–45.

Russell, D.E.H., 1984. Sexual Exploitation: Rape, Child Sexual Abuse, and Workplace Harassment. Sage, Beverly Hills, CA.

Sanday, P.R., 1981. The socio-cultural context of rape: a cross-cultural study. Journal of Social Issues 37 (4), 5–27.

Sanday, P.R., 1990. Fraternity Gang Rape: Sex, Brotherhood and Privilege on Campus. NYU Press, NY.

Sanday, P.R., 1996. A Woman Scorned: Acquaintance Rape on Trial. University of California Press, Berkeley, CA (Ebook edition published with new Afterword in 2011. Knopf Doubleday Publishing Group.).

Symons, D., 1979. The Evolution of Human Sexuality. Oxford University Press, New York.

Tanner, N.M., 1981. On Becoming Human. Cambridge University Press, Cambridge, UK.

Thornhill, R., Palmer, C.T., 2000. A Natural History of Rape: Biological Bases of Sexual Coercion. MIT Press, Cambridge, MA.

Travis, C.B., 2003. Talking evolution and selling difference. In: Travis, C.B. (Ed.), Evolution, Gender, and Rape. The MIT Press, Mass., pp. 3–28.

Vance, C.S., 1984. Pleasure and danger. In: Vance, C. (Ed.), Pleasure and Danger: Exploring Female Sexuality. Routledge and Kegan Paul, Boston, pp. 1–28.

Warshaw, R., 1988. I Never Called It Rape: The Ms Report on Recognizing, Fighting and Surviving Date and Acquaintance Rape. Harper & Row, New York.

Rational Choice and Organization Theory

Mary A Zey, University of Texas, San Antonio, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 12751–12755, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editors.

Abstract

Rational choice theory (RCT) is based on the premise of individual self-interested utility maximization. Organizational theory (OT) is based on the premise of efficient functioning of organizations through means/ends rationality within organizations. Organizational economic theory, which consists of transaction cost economics (TCE) and agency theory (AT), forms an indirect link between RCT and OT. However, TCE and AT differ in their grounding premises. Thus, economic organizational frameworks, which purport to link micro-level RCT to meso-organizational theory, are unable to do so, due to the incompatibility of the premises on which RCT and OT rest. As one set of premises are assumed, the other becomes problematic. No OT has successfully integrated RCT and OT.

Rational choice theory (RCT) is based on the premise of individual self-interested utility maximization. Organizational theory (OT) is based on the premise of efficient functioning of organizations through means/ends rationality within organizations. Organizational economic theory, which consists of transactions cost economics (TCE) and agency theory (AT), forms an indirect link between RCT and OT. However, TCE and AT differ in their grounding premises. Transaction cost theory holds to the premise of OT that organizational rationality dictates that organizations are efficiently designed means/ends mechanisms, in some cases more efficient than markets, and finds that the RCT premise of individual self-interested utility maximization is problematic under conditions of uncertainty and asset specificity. AT holds to the premise of RCT that individuals are self-interested utility maximizers, but finds that the OT premise, that organizations function efficiently, is problematic. Thus, economic organizational frameworks, which purport to link micro-level RCT to meso-organizational theory, are unable to do so, due to the incompatibility of the premises on which RCT and OT rest. As one set of premises is assumed, the other becomes problematic. No OT has successfully integrated RCT and OT (Scott, 2014).

RCT and OT are two different, yet – in the works of some neoclassical economists, rational choice sociologists, and economic organizational theorists – closely related theories. RCT is used by economic organizational theorists as a method of aggregation. Specifically, it is used when it is necessary to link the change in actions of individuals with change in characteristics of organizations. However, organizational analysts generally take a broader view in which organizational context, organizational structure, and individual actions interact and move together to change organizational functioning. Such a perspective does not rest on the assumption that aggregated individual choices explain organizational behavior. In some cases, OT explains how individual actions articulate with structure. In other cases, OT does not deal with individual actions. Rather, it deals with articulating relationships between and among organizational and contextual characteristics.

The purpose of this analysis is to compare RCT with OT. RCT is based on the premise of self-interested, utility-maximizing actions/choices of individuals. OT is based on the premise of organizational rationality, the efficient functioning of organizations through means/ends relationships. Organizational economic theory, which consists of TCE and AT, forms an indirect link between the RCT and OT. First, differences by defining the origin and premises of RCT and premises of OT are elaborated. This is followed by demonstrating the link between the two, as developed in the premises of transaction cost analysis and AT. Although transaction theory holds to the premises of organizational efficiency, it finds the premise of self-interest utility maximization problematic. On the other hand, while AT holds the premise of self-interested, utility-maximizing individual choice of RCT, it finds the guiding premise of OT, the efficient functioning of firms, to be problematic. Thus, even the economic organizational theories, which purport to link RCT and OT, are able to do so only if they accept either the RCT premises or the OT premises. As one set of premises is assumed, the other becomes problematic and becomes the questions to be examined in the analysis. No OT has successfully integrated RCT and OT.

Rational Choice Theory: Origin and Premises

RCT originates in the British utilitarian tradition, via neoclassical economics. It is the core of classical economic theory. Although often qualified beyond recognition to meet conditional constraints of organizational analysts, it has been adopted in sociology (social choice) and in political science (public choice, game theory). In OT and analysis, rational choice is most often used for the development of economic models of organizations. TCE and AT are the dominant economic organizational perspectives.

The premises at the basis of RCT as derived from neoclassical economic theory are: (1) self-maximizing individual utility-driven actors, unconstrained by others and social norms (Coleman, 1990, p. 503); (2) actors possess complete

information and knowledge of their tastes, existing resources, the resources of others, and market characteristics of the products and services they seek; (3) actors calculate and recalculate the changing conditions and act rationally; (4) however, preferences do not change, they are stable, and they are givens often implied from the very behavior they are to predict; (5) actors are not emotional, irrational, impulsive, habitual in their choices, they are rational; and (6) interaction is defined as an exchange between buyer and seller achieved when utility and cost converge in a market transaction of supply and demand equilibrium.

RCT is an axiomatic system that provides a basis for making individual decisions/choices. The axioms of RCT define rational choice on the basis of choices between pairs of alternatives, which meet the axiomatic requirements of: consistency, transitivity, independence, continuity, and monotonicity. Consistency demands that it be possible for all the decision-maker's options to be ranked. The assumption of consistency is also called the assumption of connectedness. Preferences must be either equal or unequal. Unequal preferences can be rank ordered for comparison across the decision-maker's entire set of preferences. Transitivity is the assumption that if A is preferred to B, and B is preferred to C, then consistency requires that A is preferred to C (Green and Shapiro, 1994, pp. 14–15). The distance between preferences or magnitude of preferences does not have to be known to the subject or the analyst. Independence assumes that all preferences are independent of all other preferences. The continuity assumption is that preferences hold across time and space. Rational choice theorists assume monotonicity – all decision rules and preferences are stable across individuals and time. Allowing for assumptions of interpersonal variation produces enormous tractability problems (Storm, 1990, p. 126). Another assumption of RCT is that individuals "maximize the expected value of their payoff measured on some utility scale" and that, usually, "numerical probabilities can be attached to different eventualities" (Luce and Raiffa, 1957, p. 50). Eventuality is calculated as the value of the outcome "weighted by the probability of achieving it" (Elster, 1986; Harsanyi, 1986).

Organizational Theory: Origin and Premises

RCT and OT are brought together by the term rational. However, the term has different meanings (Zey-Ferrell, 1981) for RCT and OT. RCT is based on the idea that an individual seeks to maximize his/her utilities as he/she conducts organizational business. The central tenant of rational choice is that individuals at all levels of organizations act selfishly based on their own preferences to maximize their utility. This is rational individual behavior. OT defines rational as the way in which economic organizations are designed – the most efficient method of organizing. This is rational organizational behavior.

Rationality in OT originates in Weber's ([1922]1968) concept of 'rational bourgeois capitalism,' which connotes the systematic organization of ideas, people, and resources in the interests of instrumental efficiency and effectiveness. He wrote that rationality and its accompanying bureaucratic organizational form, based on formalization of the rule of law and differentiation of functions, pervaded not only the modern economy, but public administration, the legal system, and religions. The rationalization of bureaucratic structures led to increased efficiency. In Weber's model, organizational rationality and rationalization are essentially the same. This is not an (micro-level) individual process as it is in RCT. It is an organizational process or specification of mean/ends relationships based on systematic calculation and recalculation to achieve the most efficient organizational functioning. Subsequent analyses of organizations have assumed that this rationality is the basis of organizations and developed theories around this premise.

Divergences of Rational Choice and Organizational Theories

In addition to definitional divergence of the term rational, the most apparent divergence is the difference in the unit of analysis. RCT attempts to explain individual choice behavior. RCT is based on the utility-maximizing behavior of individual choice, while OT is based on the assumption of efficiency-seeking organizational design. OT explains behavior of organizations. The utility-maximizing behavior of individuals and the assumption of efficiency-seeking organizational design are similar, but are at different levels of analysis. If the unit is the individual, the utility being maximized is individual owner or employee self-interest. If the unit is the organization, the utility being maximized is profitability. Profitability of the firm does not always translate into income or other self-interests of firm members. Likewise, rational efficient organizational functioning does not always maximize the self-interest of organizational members.

Economic organizational theorists define organizations as firms doing business in markets. Neoclassical economists begin with market assumptions and work backward to firms and individuals assuming that actors operating in firms act in ways consistent with the theory of the market. However, assumptions that hold for RCT do not hold for organizational actors (March, 2014). Let us take one assumption, i.e., assuming that all actors have full information about costs. Because each of them knows the cost and adjusts the quantity of their purchase, it is reasonable to assume that the market will settle on an equilibrium. In organizations these assumptions are false. Many types of decisions are not afforded to each actor. The choices available to each actor are not the same. Some actors' choices impose constraints on the choices available to others. Some actors do not know that a particular choice exists, for example, information about pending mergers and acquisitions. During mergers and acquisitions, parts of corporations maintain secrecy, not only from other firms, but from other units of their own firm and individual within their own units (Zey, 1993, 1999). For example, securities firms are required by law to construct walls of secrecy, e.g., 'Chinese Walls' between M&A units and traders to prevent insider trading. Information is also kept secret to prevent oppositional reactions to present and new allocations of organizational resources. Information is not equally distributed among or within organizational units for deliberate power-based reasons. Finally, the norms of organizational rationality, efficiency dictates that various units within an organization

should not receive the same information, but only what is necessary for their organizational entity. For all units to receive all information would be inefficient.

The Link between RCT and OT

Although RCT and OT are based on different premises and units of analysis, one type of theory, organizational economic theory, has attempted to link the two sets of premises and units of analysis. There are two dominant perspectives in organizational economic theory: TCE and AT.

Economic organizational theory, both AT and TCE, begins with the relationships between individuals and organizations that are often conceptualized as events or transactions in which goods and services are exchanged (transaction cost analysis) and as contracts, which are agreements to conduct specific transactions (agency analysis). One is immediately aware that the unit of analysis is aggregated individuals. Transactions and contracts are aggregated either cross-sectionally or over time. In contrast, OT deals with the properties and characteristics of organizations, not individuals.

Coase (1937), from whom Williamson (2010) elaborates his TCE, explains the origin of firms as alternatives to inefficient markets, while organizational theorists observe that people create organizations to generate income and power, to create larger empires with which they are identified, for themselves or groups to which they belong. Mergers to create new organizations often take place to take advantage of new technologies, to take advantage of tax changes, or to reposition the organization in a growing market. These changes may have little to do with rational economic organization or inefficient market functioning.

Convergence and Divergence of Economic Theories of the Firm and Organizational Theory

TCE and AT, generally labeled economic organizational theories, have attempted to link RCT and OT and their different units of analysis and methods of theorizing through a set of assumptions that aggregate utility-maximizing individual behavior so that their behavior promotes organizational efficiency (profitability). However, survey research has demonstrated that not everyone has the same (or even individually consistent) utility. Thus, the aggregation, specifically nonadditive aggregation functions, of individual choices and action into organizational actions is not clear-cut, because the complexity of the aggregation process often violates RCT assumptions.

Furthermore, the two economic organizational theories, AT and TCE, do not operate under the same premises. AT is based on the utility-maximizing individual of RCT (Jensen, 1983, 1993; Jensen and Meckling, 1976), while TCE assumes that humans are boundedly rational at best. AT attempts to explain why organizations are more or less efficient while TCE assumes that organizations are an efficient alternative to inefficient markets (Williamson, 1975, 1985, 1996). The utility-maximizing behavior of organizations is efficiency-seeking organizational design. If the unit of analysis is the organization, and the utility being maximized is profit, then efficiency maximization is implied when income seeking is not compromised. Organizational theorists question these assumptions because organizations, even firms, are never constraint-free and individuals within organizations do not operate under the same constraints.

AT assumes that individuals are utility maximizers, but does not assume that organizations are structured in the way they are because it is the most efficient method of organizing. This is demonstrated, not only in the assumption that managers and workers act as opportunistic agents – that is, not in the interest of the owners – but also by the preoccupations of agency theorists with governance and methods of controlling nonowners to maximize profits (returns to shareholders) to the owners (Jensen and Ruback, 1983). Agency theorists and economic organizational theorists in general do not integrate into their analyses historical changes, which provide opportunities and constraints on individual and organizational actions as organizational theorists do. Jensen's presidential address to the American Finance Association demonstrates that he had not until recently discovered Chandler's work (1962, 1977, 1990) and that organizational forms, which are most efficient, change over time with each historical context in which they develop.

TCE assumes that organizations are structured in the way they are because it is the most efficient method of organizing, but TCE does not assume that individuals within organizations are utility maximizers. For example, in *Markets and Hierarchies* (1975), Williamson assumes that organizations more efficiently eliminate (by management fiat) opportunistic ('self-interested with guile') individual behavior than market mechanisms. TCE theorists analyze uncertainty and asset specificity of transactions. Individual utility maximization is problematic because of the obvious tendency of rational actors to act opportunistically under conditions of uncertain and asset specificity. TCE, like AT, ignores historical changes and processes which constrain and provide opportunities for organizations. Both contextual and organizational changes are ignored. Organizational change, especially transformational change, cannot be explained by economic theories of the firm (see Zey and Swenson, 1998, 1999). This is problematic because of the major transformations of firms at the turn into the twenty-first century.

In both Williamson's TCE (1975) and in Jensen's AT (Jensen and Meckling, 1976), the idea of self-interest in contractual arrangements between and within organizations is problematic. Ultimately, questions of reconciliation of differential self-interest in the absence of trust are impossible. Trust, rules of fairness, norms of reciprocity, and conflict are nonexistent and irrelevant in RCT and economic theories of the firm. They are essential to organizational formation and perpetuation.

Neither TCE nor AT holds both the basic premise of RCT, self-interested, utility-maximizing individuals, and the basic premise of OT, that firms are designed in the most efficient form of organizational design. AT holds the former, while TCE holds the latter. Each finds the other's basic premises problematic and grist for the mill of organizational analysis. Thus, economic OT has not yet integrated RCT and OT. This author does not think it is possible to do so because of conflicting premises and different units of analysis (**Table 1**).

Table 1 Rational choice theory (RCT) and organizational theory (OT): ideal-type contrasts

	Rational choice theory	*Organizational theory*
Intellectual Origins	Neumann and Morgensterns Milton Freideman Oliver Williamson Michael Jensen	Max Weber and Herbert Simon
Unit of analysis	Individual	Organization
Content of analysis	Preferences	Characteristics of organization and relationships to context
Human nature	Self-interested	Self- and other interested
Rationality	Individual, self-interested utility maximizer	Means/ends efficiency of functioning
Utility maximization	Self-interests	Profits in firms
Human preferences	Fixed hierarchy	Variable over time and place
Basis for interaction	Assumed trust	Trust is problematic
Constraints	Economic actions are constrained by tastes and scarcity of resources	Organization actions are not only constrained by scarcity, but also by social structure and historical context
Method	Predictive and explanation not descriptive. Formal, mathematical model building no empirical data (clean hands)	Many methodologies comparative, historical based on data from empirical observations (dirty hands)

See also: Behavioral Theories of Organization; Decision Making: Nonrational Theories; Decision and Choice: Bounded Rationality; Game Theory; Hierarchies and Markets; Market Structure and Performance; Organization: Overview; Transaction Costs and Property Rights; Utilitarian Social Thought, History of; Weber, Max (1864–1920).

Bibliography

Chandler Jr., A.D., 1962. Strategy and Structure: Chapters in the History of the Industrial Enterprise. MIT Press, Cambridge, MA.
Chandler Jr., A.D., 1977. The Visible Hand: The Managerial Revolution in American Business. Belknap Press, Harvard University Press, Cambridge, MA.
Chandler Jr., A.D., 1990. Scale and Scope: The Dynamics of Industrial Capitalism. Belknap Press, Harvard University Press, Cambridge, MA.
Coase, R., 1937. The nature of the firm. Economica 4, 386–405.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press, Harvard University Press, Cambridge, MA.
Elster, J. (Ed.), 1986. Rational Choice. New York University Press, New York.
Green, D.P., Shapiro, I., 1994. Pathologies of Rational Choice Theory: A Critique of Applications in Political Science. Yale University Press, New Haven, CT.
Harsanyi, J.C., 1986. Advances in understanding rational behavior. In: Elster, J. (Ed.), Rational Choice. New York University Press, New York.
Jensen, M., 1983. Organizational theory and methodology. Accounting Review 55 (2), 319–339.
Jensen, M., 1993. Presidential address: the modern industrial revolution, exit, and the failure of internal control systems. Journal of Finance 48 (3), 225–264.
Jensen, M., Meckling, W., 1976. Theory of the firm: managerial behavior, agency costs, and ownership structure. Journal of Financial Economics 3, 305–360.
Jensen, M., Ruback, R.S., 1983. The market for corporate control. Journal of Financial Economics 11, 5–50.
Luce, D.R., Raiffa, H., 1957. Games and Decisions: Introduction and Critical Survey. A Study of the Behavioral Models Project. Bureau of Applied Social Research, Columbia University. Wiley, New York.
March, J.G., 2014. The two projects of microeconomics. Industrial and Corporate Change 23 (2), 609–612.
Scott, W.R., 2014. Institutions and Organizations: Ideas, Interests, and Identities, fourth ed. Sage, London.
Storm, G.S., 1990. The Logic of Lawmaking: A Spatial Theory Approach. Johns Hopkins University Press, Baltimore, MD.
Weber M., [1922]1968. Economy and society. In: Roth G., Wittich C. (Eds.). Bedminster Press, New York.
Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications: A Study in the Economics of Internal Organization. The Free Press, New York.
Williamson, O.E., 1985. The Economic Institutions of Capitalism: Firms, Market, Relational Contracting. Free Press, New York.
Williamson, O.E., 1996. The Mechanisms of Governance. Oxford University Press, New York.
Williamson, O.E., 2010. Transaction cost economics: the origins. Journal of Retailing 86 (3), 227–231.
Zey-Ferrell, M., 1981. Criticisms of the dominant perspectives in organizations. The Sociological Quarterly 22, 181–205.
Zey, M., 1993. Banking on Fraud: Drexel, Junk Bonds, and Buyouts. Aldine de Gruyter, Hawthorne, NY.
Zey, M., 1998. Rational Choice Theory and Organizational Theory: A Critique. Sage, Thousand Oaks, CA.
Zey, M., 1999. The subsidiarization of the securities industry and the organization of securities fraud networks to return profits in the 1980s. Work and Occupations 26, 50–76.
Zey, M., Swenson, T., 1998. Corporate tax laws, corporate restructuring, and subsidiarization of corporate form, 1981–95. The Sociology Quarterly 39, 555–582.
Zey, M., Swenson, T., 1999. The transformation of the dominant corporate form from multidivisional to multisubsidiary: the role of the 1986 Tax Reform Act. The Sociological Quarterly 40, 241–267.

Rational Choice Explanation: Philosophical Aspects

Russell Hardin, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rational choice explanation is the progeny of the Scottish Enlightenment and neoclassical economics. It is essentially reasoning from economic assumptions of what actors value to reach conclusions about both individual actions and aggregate- or collective-level outcomes. It is applied to virtually any significant behavior or social outcome, from political participation, crime, and drug addiction to institutional structures, legislative and interest group politics, law, norms, failures of collective action, and revolution. The most extensively developed parts of such theory are in democratic theory, the analysis of law under the rubric of law and economics, the study of group behavior, and game theory.

There was a time – the era of Hume and Smith – when it was natural to frame economic and political–philosophical social analyses as a singular coherent enterprise, not partitioned into unrelated disciplines. That breadth of understanding has been taken for granted in contemporary rational choice, in which the focus is on explanation, especially on individual motivations.

Rational choice explanation is the progeny of the Scottish Enlightenment and neoclassical economics. It is essentially reasoning from economic assumptions of what actors value to reach conclusions about both individual actions and aggregate- or collective-level outcomes. It is applied to virtually any significant behavior or social outcome, from political participation, crime, and drug addiction to institutional structures, legislative and interest group politics, law, norms, failures of collective action, and revolution. The most extensively developed parts of such theory are in democratic theory, the analysis of law under the rubric of law and economics, the study of group behavior, and game theory. Rational choice explanations gain their power from the attempt to reduce particular motivations to some narrower set of motivations, often simply to interests, which can be moderately well defined and measured across a large variety of contexts and actions. The article surveys game theory and rational choice value theory, and it applies rational choice explanation to collective action, democracy, voting, norms, and social order.

The most common criticism of rational choice explanation is that it is false to the facts of actual motivations and actions of people. This is a by-inspection claim. Because very much of actual behavior surely is motivated by interests, the dispute must actually be whether rational choice explanation is likely to help understand certain activities but not others. This is a question that cannot be answered in the abstract. Sometimes the criticism of rational choice trades on the fact that we commonly resolve many apparent choices by acting from habit or routine. As A.N. Whitehead quipped, civilization advances by the reduction of more and more choices to habit. A similar criticism of utilitarianism is that, if we spend as much time in calculating costs and benefits of our choices as would seem to be necessary to get our choices right, we will have no time left for life. J.S. Mill (1969) responded, rightly, that a sailor need not work out the positions of the stars anew each trip to sea but can simply rely on an astrolabe.

To illustrate the nature and power of rational choice explanation, as well as its plausible limits, I will address the aggregate-level phenomena of collective or group action, majority rule, and democracy. I will then address game theory and the value theory of rational choice. Finally, I will address the application of rational choice explanation to norms and social order.

Fallacies of Composition

A common criticism of rational choice explanation is that it cannot handle aggregate-level phenomena because it is fundamentally grounded in individual-level interests, choices, and actions. On the contrary, the chief successes of such theory have been to relate aggregate-level phenomena to individual-level actions. Several of the most important findings have been the – typically negative – demonstration that various aggregate-level claims are fallacies of composition from individual-level actions, choices, or values. It is a fallacy of composition to suppose that a collection of items has itself the characteristics of those items. This can be true, but it must be shown and not merely assumed. Consider three of these demonstrations.

The logic of collective action suggests that, although individuals may share a general interest in supplying themselves some collective benefit, they individually may have no interest in contributing effort or resources to the project. My contribution will buy a small part of our good, of which an even smaller part will go to me, so that my cost of contribution will be larger than my benefit from my contribution. All members of a group may therefore be individually worse off because they failed to take some action (Olson, 1965; Hardin, 1982). This is a problem that is pervasive in societies, ranging from very small societies with limited resources to very large and complex industrial societies.

The Arrow theorem (Arrow, 1963; Barry and Hardin, 1982) concludes that a collective utility or welfare function cannot be aggregated from individual utility functions if both are ordinal (see discussion of ordinal utility theory below). In essence, the theorem says that there cannot be a collective preference that has the simple features of an individual preference function. Among the implications of this theorem

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63070-2

is that every standard choice procedure is potentially incoherent. Majority rule, for example, can produce cyclic preferences such that candidate A would defeat B by a majority, B would defeat C, and C would defeat A. Hence, simple majority rule is potentially indeterminate. It often works well enough – or seems to work – in practice, however, because there might not be cyclic preferences or because there are devices that block such preferences from being effective and even from being revealed. For example, plurality elections or the use of run-offs between the top two candidates in a multicandidate election can hide any cyclicities because there cannot be cycles over only two candidates.

The economic theory of democracy explains why candidates might not be substantially differentiated in certain common contexts and why voters often have little interest in voting. It also makes sense of the otherwise surprising level of ignorance of typical voters. In this latter case, however, there may be no fallacy of composition, because individual-level ignorance often translates into collective ignorance. Consider the problem of voting, as analyzed by Anthony Downs (1957). The chance that my vote would make a difference in an election decided by a large electorate is very nearly nil. In a 1974 statewide contest for senator in the very small state of New Hampshire, with somewhat more voters than Iceland, the vote was so close that errors in counting made it practicably undecidable who had won. No-one's single vote mattered. Suppose that the trouble to me of voting is such that I would sooner spend several dollars to vote by telephone to avoid going to the polls. The value to me of electing one candidate rather than another would have to be in the many thousands of dollars to offset my costs of voting, because the benefit would have to be discounted by the very low likelihood that my vote would make a difference. Yet many people do vote. Much behavior seems not to fit the rational choice theory and must be explained in other ways.

If it would not be worth my very slight effort to vote, Joseph Schumpeter (1950) concludes, then it would be even less worth the much more substantial effort it would take for me to be well informed about the candidates and issues at stake. Evidence from many national contexts suggests, in fact, that citizens are abysmally ignorant when they vote. Here rational choice explanations seem to fit well. Editorial campaigns against voter ignorance are often misframed as being about the irrationality of citizens, who are acting all too rationally.

Game Theory

Rational choice explanations are commonly based in game theory, neoclassical economics, or commonsense arguments such as those of Smith and Hume. Game theory is perhaps the framework that best characterizes the nature of most of the problems in rational choice theory, although many of the most important insights in such theory predate game theory, which originated roughly during World War II (Von Neumann and Morgenstern, 1953). Although it is called a theory, game theory is little more than a descriptive framework on which many theories about particular or quite general phenomena are founded. That framework takes many forms, the most transparent of which is the matrix form in which each of the players or participants faces a set of possible strategy choices. Outcomes of a play of a game are jointly determined by the strategy choices of all the players.

This characterization of our individual choices – of strategies, not of outcomes – makes game theory virtually antideontological. Judging action types as right or wrong makes no sense in interactive contexts, which are the principal contexts for most of moral theory. An action theory that takes what are essentially individual choices of outcomes as central is almost irrelevant to the choice in social contexts in which, typically, important outcomes necessarily are not under the control of individual deciders (Hardin, 1988: pp. 68–70). An ethic of rules makes nonsense of choices of strategies, which are chosen because of their effect on outcomes, as these are influenced by the strategy choices of others. Game theory can readily be fitted to consequentialism and especially utilitarianism. Utilitarianism essentially generalizes the principle of rational choice to choosing the welfare of all instead of only the welfare of the individual chooser.

To fit utilitarianism to game theory is to revamp its central concern with outcomes as valued by their benefits and costs. In interactive social choice, we typically cannot say that a particular action is right or wrong in producing a good overall outcome. We must, rather, judge strategy choices and we must judge them contingently. A strategy choice is right only to the extent that, in interaction with the choices of others, it produces a good overall outcome. If the others with whom I am interacting are sometimes utilitarian, sometimes self-seeking, and sometimes rule-guided in their choices, I may have difficulty determining my best strategy choice for producing a utilitarian outcome. This is not different from the difficulty that self-interested players have in determining the best strategy choices for themselves. Game theory is indeterminate in many contexts, and therefore both pragmatic and utilitarian choices in a real-world setting are apt to be indeterminate as well. Merely seeing the nature of interactive choice in descriptive game theory changes our understanding of the possibilities of rational choice theory and of utilitarianism.

Value Theory: Interests, Consumptions, and Welfare

Rational choice accounts are commonly grounded in interests. Sometimes, interests are the whole story of someone's motivations in a particular context. But typically, my interest in having more resources, such as money, is itself grounded in my desire to consume various things. Resources, capabilities, and opportunities matter to us only because they enable us to consume or experience various things, and consuming or experiencing these things is what constitutes our welfare. Resources are worthless otherwise. Focusing only on interests is therefore a shorthand for what the resources can enable us to do. Moreover, we could relate our consumptions to the overall welfare that they bring us. It would be a mistake to suppose interests translate smoothly into welfare or consumptions. Indeed, consumptions typically trade off against each other and against interests. If I invest much of my resources in a holiday, I may not have enough left for other things, such as a new refrigerator. Rational choice explanations involve all three of these categories – interests, consumptions, and welfare.

It would also be a mistake to suppose that my welfare is wholly self-interested. One of the things I might wish to consume or experience might be your happiness. One may say this is against my interests because it means using resources that I could use for my own happiness in some other, perhaps more direct, way. But this is a confusion in at least two ways.

First, the only point in having resources is to be able to consume in order to contribute to one's welfare, as is evident to everyone but a Scrooge for the resource of money. True, there is a conflict between using my resources for one consumption and saving them for another, but this is properly seen not as a conflict between consumptions and interests but as a conflict between different consumptions, or between different paths to welfare.

Second, there is no reason to suppose that, say, enjoying an opera, a good meal, or a lovely journey is a more direct or more successful route to enhancing my welfare with some bit of my resources than providing for my son's or lover's, or friend's happiness with that same bit of resources. More generally, my welfare is enhanced best by a mixture of consumptions and experiences, including contributions to the welfare of others, especially certain others. With those certain others, my welfare will depend on our sharing intentions to do various things for and with each other.

If we could suppose that economic interests as measured in resources were the primary issue for people, we could fairly easily explain actions as though people were all alike except for their circumstances. Rational choice explanations typically reduce the focus to interests, rather than other demands, such as demands for moral positions (ranging from deontological to utilitarian). In part they do so for methodological reasons. Analysis is more productive if we keep it clean and focused, and simplification is required for virtually all social analyses. But in part, the focus on interests is often basically right, especially for aggregate or collective issues. Other concerns matter to collective groups and societies, but they matter the way they do in life more generally. They often do not conflict with interests although, of course, occasionally they do and they may even trump interests in many contexts. For example, voters may vote against their personal interests in support of welfare programs and policies favoring ill-represented ethnic groups. But those who voted against the British Raj in India, French control of Algeria, or the US involvement in Vietnam may generally have voted with their interests as well as with moral concern. Still, many social and individual phenomena cannot readily be explained as the product of interest-seeking by individuals.

Value Theory: Cardinal, Ordinal, Ramsey–Neumann Cardinal

Apart from the structure or content of the value theory discussed above, there has been extensive debate over the form of such theory. Jeremy Bentham supposed that pleasures and pains during a day of someone's life could be defined in such a way that we could add the value of pleasures and subtract the value of pains for an overall measure of benefit to the person for that day. Today, we would speak of such a welfare or utility measure as cardinal in the sense that each experience one has could be assigned a utility number that could be added straightforwardly to other utilities.

Bentham's terminology was confused in that he spoke of the utility one gets from consuming some object and also of the utility inherent in the object. The latter notion underlies the labor theory of value that corrupted much of economic thinking throughout the nineteenth century and often still corrupts it today, in part under the doleful influence – on this matter – of Karl Marx. Bentham contributed to two opposing streams of value theory: One is that of the labor theory of value and the other is that of the subjective theory that holds value to be strictly what a person gets out of an experience or consumption. The value to me of consuming something such as, say, beluga caviar might be far below the value to me of consuming other things that take equal expenditures of my resources. The value to you, even with the same overall level of resources, might be radically higher. Hence, you might consume a lot of beluga caviar and I might consume none after an initial taste. The subjective theory obviously has the disadvantage that we cannot always generalize from one person to others, although often we can generalize enough not to need to do in-depth studies of the tastes and preferences of everyone in order to explain the overall behavior.

Bentham supposed further that the utility for one person could be added to that for another. This supposition would allow us to add up the utilities to all under one set of circumstances and to compare the total to that under another set of circumstances. Supposing we could do the measures, we could then say that, for example, one form of government is superior to another.

The so-called ordinal revolution in demand theory in economics overthrew all of the cardinal assumptions and set in their place a utility theory that is entirely subjective and that cannot be given interpersonal comparisons. The remarkable fact for economics is that such a limited theory clears up many prior problems and makes coherent sense of demand and supply relationships and, hence, prices in market exchange. Unfortunately, this revolution came only in the 1930s when moral theorists had ceased also to be economic theorists, as Smith, Hume, Mill, and Henry Sidgwick had been. It has therefore not had as much influence in general social theory as in economics, although it clearly should be brought to bear on the former as much as on the latter. A clear implication of interpersonally noncomparable value theory is that many traditional questions of interaction between people cannot be readily settled, so that moral and social theory must often be open on many issues.

In what might count as a minor sideline, Frank Ramsey (1931) and John Von Neumann (Von Neumann and Morgenstern, 1953) independently proposed a way to cardinalize preferences that are merely ordinal. Suppose you face a list of mutually exclusive outcomes, and suppose you can rank order any two outcomes. Assign a value of 0 to the least preferred, Z, and a value of 100 to the most preferred, A, outcomes in your list. For some outcome, K, between these two in your ranking, you can rank it against any lottery over A and Z. For example, suppose you would be indifferent to getting K and getting the combination of a p% chance of getting A and a $100 - p$% chance of getting Z ($0 < p < 100$). On the scale of 0–100, K has a cardinal value to you of p.

The Ramsey–Von Neumann (RN) cardinalization of utility clearly applies to personal decisions over outcomes. But we cannot use it to construct a Benthamite cardinal utility that is additive across persons, because there is no sense in which your 0 and my 0 are the same or your scale and mine, because each of them is arbitrary. If I assign a value of p and you assign a value of q to outcome K, we cannot meaningfully say that K has a joint value of $p + q$. Indeed, we can perform a linear transformation on your utility scale without affecting the results that follow from it. For example, we could divide the value of each outcome by 10 and add 37 to it, so that your scale would then run from 37 to 47, and this transformed scale would get all the results that your original scale from 0 to 100 would get. In addition, we can assign utility values to outcomes that fall outside the original ranking from A to Z, so that some outcomes might have negative utility and others might have extremely large positive values. This is suggested by the fact that the original normalization was performed over the arbitrary set of outcomes from A to Z. Hence, there is no meaning to any claim that your RN utility can be added to mine to yield an overall joint utility. The RN theory is therefore primarily of interest in individual decision making under risk, and is seldom of interest for more general social theory. It is irrelevant to Benthamite hopes for a cardinally additive utilitarianism.

Norms

There is a growing literature on the rational choice explication of norms. Edna Ullmann-Margalit (1977) gave rational reconstructions of norms for dealing with strategic contexts that are modeled by prisoner's dilemma, coordination, and partiality or unequal coordination games. In this categorization, the descriptive apparatus of game theory is used to establish and characterize types of norms according to the strategic problems they resolve. The first of these norms essentially handles collective action problems because, without the force of the norms, many individuals would free ride, and any effort to provide a collective good would fail. The second handles such coordinations as the driving convention. That convention surely requires no norm to keep drivers in compliance with it, because the threat of grievous harm from violating the convention is sufficient. But many conventions are less regulated by such clear incentives and some of these might be more forceful if backed by norms. Norms in Ullmann-Margalit's third category handle unequal coordinations such as the division of labor by gender roles or social order in any hierarchical society in which obedience to norms might keep even those at the bottom of the hierarchy loyal and orderly. Again, these conventions are grounded in coordination incentives and they might be fully self-enforcing, but norms that moralize the inequalities might make those in the less well off positions less sullen in acquiescing in the arrangements.

More generally, we may say that there are two broad categories of norms: norms of exclusion and universalistic norms (Hardin, 1995: chapters 4 and 5). Norms of exclusion are specific to particular groups, which can enforce the norms by sanctioning violators within the group; the sanctions can go so far as to shun and even exclude members. The sanctioning is grounded in the interests of individual sanctioners, so that the norms are commonly interest based, although group members might claim they are inherently moral or right. A norm of exclusion can work to help enforce cooperativeness in a group or community, or to enforce a coordination that might serve the interests of some but not all. For example, a minority culture might enforce upon its members the use of a minority language or a particular dress code. Using that language or following that dress code might reduce the possibilities for some members of the community to gain employment outside the community. The community might protect its hegemony over deviant members by shunning those who violate the language or dress code.

Universalistic norms can be dyadic, as in promising or truth telling, or large-number in the sense of applying to actions toward the whole society, as in not littering. Dyadic norms are commonly enforced through the sanctions inherent in the iterated prisoner's dilemma. If you lie to me or break your promise to me, I can simply avoid further interaction with you, to your likely cost. If you break a norm that affects everyone, such as a norm against littering or in favor of voting, no particular individual is likely to have an interest in sanctioning you. Hence, such norms are likely to be relatively weak in their force and in our adherence to them. Oddly, therefore, it is the often odious norms of exclusion, such as racist and ethnic norms, that have strongest force, while universalistic norms are relatively weak unless they are played out essentially in dyadic interactions. Those who insist that, morally, norms should commonly trump interests, should be bothered that it is interests that trump any plausible moral content of many norms, and that norms not backed by interests are likely to be weak. Large-number universalistic norms may be backed primarily by moral or psychologically inculcated commitments. If so, then it is of interest whether they are, in fact, very effective in motivating people.

Social Order

There are three grand schools of social theory. One, associated with the French sociologist Émile Durkheim and many others, holds that society is held together by shared values. The second, associated with Karl Marx, and Plato's *Thrasymachus*, holds that conflict is pervasive and that social order must therefore be maintained through the imposition of power to coerce when necessary. The third, associated with Adam Smith, holds that society is organized by and for exchange, which involves a combination of coordination and conflict. Each of these schools has many strong contemporary advocates. A fourth plausible view is that much of social life is simply organized by coordination (Hardin, 1999: chapter 1). We need not share particular values in order to share interests in avoiding conflicts with each other. The ideal type of this last form of social order is coordination on traffic rules, especially on the rule to drive right (or to drive left). We can be hostile to each other in many ways and still find it sensible to avoid needless conflict over many things, so that we would willingly be coordinated. A shared value theory is also a coordination theory, but its advocates ground it in value commitments and are generally hostile to the inclusion of interests among those value commitments. The conflict and exchange schools

are commonly grounded in interests, although the socially destructive conflict that sparked much of early liberal thought was over commitment to divergent religious views.

The preeminent coordination theorists among classical political philosophers were Thomas Hobbes and David Hume. They were both protogame theorists in their grasp of the fundamental nature of various social interactions. While Hobbes is commonly thought to be a major conflict theorist, his actual resolution of the problem of social order is to arrange for coordination on an all-powerful sovereign, whose task is essentially to coordinate us on order so that we may construct our own separate lives and engage with one another in productive pursuits (Hobbes, 1968). Hume's (1978: book 3) articulation of a remarkable array of coordination problems qualifies him to be seen as the greatest coordination theorist, although his complex insights were largely neglected for two centuries, in part because he lacked the technical vocabulary of game theory and did not himself bother to work out the general principles of it. He did, however, work out a general moral and political theory that was substantially grounded in strategic understanding (Hardin, 1988: chapter 2).

In complex modern societies, coordination may be the central, dominant mode of social order. It does not require consensus on values – pluralists can coordinate on an order for the society in which they seek their diverse values. It does perhaps fit especially well with modern individualist liberalism that allows individuals to live by their own lights. Indeed, coordination is a simple version of Mill's principle that we should be free to pursue our own ends without interference so long as our doing so does not interfere with others (Mill, 1977). When there is such interference, of course, other principles may be required to regulate our interactions.

Concluding Remarks

Rational choice theory is a diverse set of approaches to the study of society that are based in assumptions of individual rationality. Indeterminacies in such theory often mirror indeterminacies in social relations and individual understandings of these. Indeed, many rational choice explanations have demonstrated the indeterminacy of social choice (Hardin, 2003). Among the most important contributions of such theorizing is the descriptive clarity it often adds to our understanding of issues that have been poorly characterized before. Rational choice explanation is inherently about states of affairs, although it can include in a state of affairs concerns for anything that motivates people. It is therefore easy to see how it fits with utilitarianism.

Rational choice value theories, when introduced into the study of law, have led to the burgeoning movement of law and economics, which has produced the most systematic analysis of the broad sweep of law and legal rules that legal scholarship and practice have hitherto seen. The quick march of that movement – roughly beginning with the Coase theorem (Coase, 1988) – through virtually all areas of law has provoked counterattacks, but none of these is itself grounded in a value theory systematic enough to permit more than piecemeal objections. The Coase theorem says, with audacious generality and simplicity, that if there are no transaction costs to impede economic cooperation between owners of various resources, then ownership will not affect production, although it presumably will affect how the profits from the production are divided. That theorem, along with the earlier work of Ronald Coase and many others, has led to the massive study of transaction costs.

We have a rich understanding of the problems of rational choice value theories just because those theories have been a major focus in the development and articulation of the entire discipline of economics. The theories are a product of centuries of determined and often brilliant debate with dozens of major contributions. Alternative value theories have had far too few advocates and critics to yield much understanding at all. Indeed, value claims commonly seem to be ad hoc and not systematically generalizable. Such considerations are no argument for the superiority or rightness of any value theory. But they may be a tonic for those who wonder how the only well-articulated class of value theories is so easy to criticize: It has enough content to be subject to extensive and varied criticism.

See also: Economics and Ethics; Economics, Philosophy of; Game Theory: Noncooperative Games; Irrationality: Philosophical Aspects; Practical Reasoning: Philosophical Aspects; Pragmatism: Philosophical Aspects; Prisoner's Dilemma, One Shot and Iterated.

Bibliography

Arrow, K.J., 1963. Social Choice and Individual Values, second ed. Wiley, New York.
Barry, B., Hardin, R. (Eds.), 1982. Rational Man and Irrational Society? Sage, Beverly Hills, CA.
Coase, R.H., 1988. The problem of social cost. In: Coase, R.H. (Ed.), The Firm, the Market and the Law. University of Chicago Press, Chicago, pp. 95–156.
Downs, A., 1957. An Economic Theory of Democracy. Harper, New York.
Hardin, R., 1982. Collective Action. Johns Hopkins University Press, Baltimore, MD.
Hardin, R., 1988. Morality within the Limits of Reason. University of Chicago Press, Chicago.
Hardin, R., 1995. One for All: The Logic of Group Conflict. Princeton University Press, Princeton, NJ.
Hardin, R., 1999. Liberalism, Constitutionalism, and Democracy. Oxford University Press, Oxford, UK.
Hardin, R., 2003. Indeterminacy and Society. Princeton University Press, Princeton, NJ.
Hardin, R., 2009. How Do You Know? Princeton University Press, Princeton, NJ.
Hobbes, T., 1968. Leviathan. Penguin, London.
Hume, D., 1978. A Treatise of Human Nature, second ed. Oxford University Press, Oxford, UK.
Mill, J.S., 1969. Utilitarianism. In: Robson, J.M. (Ed.), Collected Works of John Stuart Mill, vol. 10. University of Toronto Press, Toronto, pp. 203–259.
Mill, J.S., 1977. On liberty. In: Robson, J.M. (Ed.), Collected Works of John Stuart Mill, vol. 18. University of Toronto Press, Toronto, pp. 213–310.
Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Ramsey, F.P., 1931. Truth and probability. In: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Logical Essays by F.P. Ramsey. Harcourt Brace, New York.
Schumpeter, J.A., 1950. Capitalism, Socialism and Democracy, third ed. Harper, New York.
Ullmann-Margalit, E., 1977. The Emergence of Norms. Clarendon, Oxford, UK.
Von Neumann, J., Morgenstern, O., 1953. Theory of Games and Economic Behavior, third ed. Princeton University Press, Princeton, NJ.

Rational Choice Theory in Sociology

Vincent Buskens, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P. Abell, volume 19, pp. 12768–12771, © 2001, Elsevier Ltd.

Abstract

This article provides a short explanation of what rational choice theory is. Recent theoretical as well as empirical developments are discussed using the example of embeddedness effects. The theoretical developments address macro-level and micro-level extensions. The empirical part focuses on the value of using complementary research methods to test hypotheses derived from rational choice models. This article ends with some thoughts on future challenges and pitfalls for rational choice theory.

Introduction

Rational choice theory is a theoretical framework commonly used in various social sciences including economics, political sciences, and sociology. While in economics, rational choice theory has become the dominant paradigm, this has not been the case in sociology. On the contrary, rational choice theory has been widely criticized for its unrealistic assumptions, lack of empirical testing, and limited empirical validity (cf Green and Shapiro, 1994; Goldthorpe, 2000). While some of these criticisms are related to misconceptions of the interpretation of rational choice theory, others have been seriously taken up by researchers in the field. This has led to more research in which rational choice theory construction is combined with a variety of empirical tests. The *Handbook of Rational Choice Social Research* (Wittek et al., 2013) provides an extensive and recent overview of the literature in this area including theoretical advancements as well as a wide variety of empirical applications of rational choice social research. The handbook is not particularly restricted to sociology, but illustrates research and applications in economics, political science, and history as well (see Kroneberg and Kalter, 2012 for another recent review focusing on Europe and Hechter and Kanazawa, 1997, for an earlier overview focusing on family, gender, and religion). In this article, we first discuss the properties of the standard rational choice model, the so-called baseline model. Thereafter, we address extensions of the model in terms of macro-level conditions and extensions that elaborate on the micro-level assumptions related to individual behavior. We conclude with a short overview of properties of recent empirical research testing rational choice models and an outlook to possible future developments.

The Baseline Model

The classical baseline assumptions in rational choice theory comprise that actors are rational and selfish and that actors have one-shot interactions in a completely transparent context with complete information on all details of the setting. In economic terms, this implies that all interactions resemble atomized interactions in perfect markets. Clearly, rational choice theory currently comprises a much broader set of theories than those resembling neoclassical economic models. To structure current directions in rational choice theories, it is useful to distinguish extensions that focus on the macro level and the micro level. At the macro level, these extensions include broadening the assumption that all interactions resemble interactions in perfect markets and allowing more complex interaction structures in which also assumptions on information availability are relaxed. At the micro level, extensions include changes in the assumptions of rationality as well as relaxing the selfishness assumption. This distinction in extensions of rational choice models at the macro level and the micro level closely relates to Coleman's scheme in which macro-level phenomena are explained as a result of behavior of individual actors and in which macro-level conditions affect micro-level conditions for individual behavior. While micro-level behavior based on these micro-level conditions is aggregated back to macro-level outcomes (Coleman, 1990; see Raub et al., 2011 for an extensive explanation on Coleman's scheme and a sketch of its history). These macro-level outcomes are also the main interest for sociological researchers, because most sociological questions address how macro-level conditions lead to macro-level outcomes mediated through the decisions of individuals. Although there is a strong overlap between sociological models based on Coleman's scheme and rational choice models, not all micro–macro models are considered rational choice models.

Extensions at the Macro Level

The first type of extensions includes models in which the 'embeddedness' of actors is modeled in a more complex manner, for example, because actors do not interact only once but repeatedly and actors are organized in networks. For sociological questions, Coleman (1987, 1990) advocated to insist on simple and robust assumptions on rational behavior at the micro level, while indeed extending the assumptions on the social structure of interaction. This was also what Granovetter (1985) already advocated in his classic article on embeddedness in economic action. Of course, the realization that interdependences between actors have to be modeled in a more complex manner than assumed in the neoclassic model started much earlier and is already implied by the introduction of game-theoretic models in the late 1940s (Neuman and Morgenstern, 1944). In game-theoretic models, it is not assumed that the behavior of other actors is completely

determined by the external circumstances, but actors' outcomes resulting from actions of these so-called players in the game directly depend on the choices of other actors. Rational behavior in a game is defined as equilibrium behavior. The most well-known equilibrium concept is the Nash equilibrium (Nash, 1951). Strategies of actors fulfill the conditions for a Nash equilibrium in a game if none of the actors can improve his or her outcome by changing strategies and, thus, all actors play a best-reply given what all the other actors do. Thus actors are rational in equilibrium in the sense that they follow goal-directed behavior by optimizing their outcome given what others do. Note that strategies are complete specifications of behavior in a game under all possible contingencies that can happen in the game. For a more detailed account of game-theoretic terminology and textbook game-theoretic models, we refer to Rasmusen (2007).

We illustrate some key findings derived from more complex macro-level conditions based on a relatively large class of rational choice models, namely, the class of models that studies behavior in social dilemmas embedded in a social context. The key property of a social dilemma is that if in a single encounter actors follow their own self-interest and play equilibrium behavior, they obtain a worse outcome compared to when the players would choose the 'cooperative' option that would make everyone better off (e.g., Rapoport, 1974). A standard example of a social dilemma is the prisoner's dilemma (e.g., Axelrod, 1984) in which two prisoners get a 1 year sentence if they both confess, while they get a 3 year sentence if they both deny. However, if one denies and the other confesses, the one who denies gets no sentence while the other one gets a 5 year sentence. The implication is that whatever the other does, the focal actor can always best deny. If both actors do this, they are 3 years in jail, while if they would both confess they would only be 1 year in jail. Actors in a social dilemma can escape this trap if they do not have these interactions only once but repeatedly. For example, if the same actors play a dilemma sufficiently often (more precisely, if after every interaction the same actors play the same game again with a probability that is high enough), then there exist also Nash equilibria in which all actors always play the cooperative strategy (Friedman, 1971; Axelrod, 1984). Thus, the embeddedness of social dilemmas in series of repeated interactions changes the prediction from no cooperative behavior to at least a rational possibility also for cooperative behavior. If one would assume that a Pareto optimal equilibrium in which everyone is better off is more likely to result than a Pareto suboptimal equilibrium in which everyone is worse off, one would even indeed predict cooperation if the probability for repeating the game is large enough. Thus, these adaptations of the rational choice model are not only crucial because they take better into account macro-level conditions, but they also substantially change macro-level predictions.

Further extensions assume that not only two prisoners play repeatedly a prisoner's dilemma, but a network of prisoners play prisoner's dilemmas in pairs over time while the network takes care of information exchange about the behavior in the past. Raub and Weesie (1990) analyzed this model showing that more network embeddedness indeed increases the possibilities for cooperative behavior. In all these models, it is crucial that the prisoners do not know what the last time will be that they play. The reason is the so-called backward induction argument. In the last interaction, both actors know that they have no reason to be cooperative, and as a consequence they defect. But given that they will not cooperate in the last interaction, they also have no reason to be cooperative in the last but one interaction. This argument unravels until the very first round and, therefore, no cooperation is predicted for so-called finitely repeated games. The models become a bit more complex if we assume that prisoners are not completely informed about the sentences others will receive given the behavior of both prisoners (Kreps et al., 1982) and that there are some prisoners who have no incentive to deny whatever the other prisoner does. In these models, prisoners are also predicted to behave cooperatively if there are enough interactions to be played, and this result even remains valid if it is known what the last game will be. Although prisoners are the textbook example to illustrate the social dilemma, they are not the best example for embedded social dilemmas. Therefore, one should realize that the prisoners mentioned above represent, for example, buyers and sellers in purchase relations, students collaborating on assignments, partners running a household, and so on.

Examples on social dilemma research as provided earlier are still purely theoretical examples. Game theory is a strong tool, because it allows for many variations in modeling the context of social interactions and relaxing assumptions on who interacts with whom and what people do knowing about each other's preferences as well as who is informed about whose interactions. Still, these models are highly simplified at the individual level, which raises two questions: (1) Are the testable predictions derived from game-theoretic models corroborated by empirical research? (2) If predictions are not in line with empirical findings, is this due to elements at the macro level that are still simplified or are the assumptions on individual behavior wrong and in such a way that this affects the macro-level outcomes of behavior. Although the assumptions at the individual level are simple or, maybe better, parsimonious, they still imply complex rational deliberations of the actors and they also imply that the outcomes in the game reflect exactly the utility of the actors. This brings us immediately at the micro-level extensions. After these extensions, we pay more attention to some empirical work.

Extensions at the Micro Level

The second type of extensions focuses on relaxing the standard assumptions at the individual or micro level, namely that actors are rational and selfish. First, it should be realized that these are different assumptions. Rationality focuses on the ability of actors to optimize their well-being or utility given their preferences, while these can be, for example, altruistic or risk-averse preferences. Selfishness focuses on the extent to which actors optimize their own material benefits. Simon (1982) introduced already the idea of bounded rationality implying that actors are not necessarily able to optimize, but try to achieve an acceptable outcome in more adaptive ways. Limits of rationality occur because actors are not able to handle probabilities in a consistent manner in the sense of expected utility theory or

because actors have inconsistent time preferences. Alternative strategies actors can adhere to might be imitation (e.g., of successful others) or heuristics that lead to acceptable but not necessarily optimal outcomes. Extensive reviews of models on bounded rationality are provided by Gigerenzer and Selten (2001) and Gintis (2009). Lindenberg (2013) formulates a more psychological view on the issue interpreting human rationality as self-regulation in an adapted manner to a specific context or 'frame.' Many empirical examples exist for boundedly rational behavior such as the Ellsberg paradox showing that people often handle probabilities in an inconsistent manner or at least not in line with expected utility theory. Still, given that we know that boundedly rational behavior is not uncommon in certain situations, this does not imply that we need to assume bounded rationality to explain macro-level phenomena. This is necessary merely if the deviation from rational behavior not only affects micro-level decisions of some actors, but also macro-level outcomes for the group as a whole. This issue is discussed later.

The second type of extensions at the micro level relaxes the selfishness assumptions. Related theoretical arguments are developed in social psychology as well as in behavioral economics. In the literature on social value orientations (e.g., Van Lange, 2000), different types of actors are distinguished: individualists, cooperators, and competitors. Individualists only care about their own material outcome, while cooperators also assign a positive value to outcomes of others. Competitors assign a negative value to outcomes of others. While the social psychologists pioneered typifying these different types of social preferences and also accumulated considerable experimental evidence that not all people are selfish, they did not make the step to analyzing what this would imply for rational choice explanations on behavior in larger groups. Experimental economists realized that these different types of actors can be represented through different utility functions that specify how money for the focal actors and money for others can be translated to utility for the focal actor.

The most famous models in experimental economics are the inequality aversion model (Fehr and Schmidt, 1999) and the reciprocity model (Bolton and Ockenfels, 2000). The inequality aversion model by Fehr and Schmidt assumes that utility is based on what actors receive themselves and, in addition, that actors have disutility from differences between what they earn themselves and what others receive. This disutility does not need to be the same depending on whether the focal actor receives more or less than other actors in the game. Bolton and Ockenfels (2000) take a related approach assuming that utility depends on own monetary outcomes and the relative outcome of the focal actor compared to other actors. They focus on the threshold where the importance of the relative outcome takes over the importance of the own money, and thus where the argument more money is better does not hold anymore.

Both the Fehr-Schmidt as well as the Bolton-Ockenfels model focus on the outcomes actors receive in the game, but not on the path how actors reach the stages in the game. This is in contrast with more intention-based approaches by, for example, Rabin (1993) and Falk and Fischbacher (2006). In these models, behavior not only depends on the outcomes of the game, but also on the behavior experienced in the game. If one actor has to make the decision on how to divide a cake, this will typically depend on what happened before the division has to be made. The division will be, on average, fairer if the actor who has to divide the cake has been put in this advantageous position thanks to a decision of the other actor in the game than when the focal actor is put in this position, say, by an experiment leader. An additional assumption when making the rational choice model more complex in terms of rationality or selfishness is whether all actors are the same and, if not, what actors actually believe about the characteristics of other actors. Because given that it seems plausible that some actors are, for example, inequality averse, it is also likely that not all actors are inequality averse to the same extent and that actors realize that this is the case. So actors have beliefs about inequality aversion of others and they can update these beliefs based on past experience. As Aksoy and Weesie (2012) illustrate, these issues raise many new questions: do the beliefs of actors depend on their own characteristics? Is updating of beliefs done in a rational manner? To what extent do we need to take these beliefs into account to understand macro-level outcomes? Aksoy and Weesie's results suggest that, yes, beliefs depend on focal actors' characteristics and, yes, they seem crucial to understand macro-level outcomes.

There are various related approaches in which actors are goal-directed in their decisions, but not necessarily optimizing in the strict rational sense and also not necessarily selfish. Although these approaches are at least closely related to rational choice theory, even the authors in these approaches vary depending on whether they think their theories should be considered variants of rational choice theory or not. What they at least do have in common is that they are analytical in their derivation from explicit assumptions to theoretical predictions. One line of research is in analytical sociology as advocated by Hedström (2005; Hedström and Bearman, 2009). Hedström spells out his 'DBO theory' (desires, beliefs, opportunities) through which he elaborates his program on 'mechanism-based' explanations. Although many of the features in DBO theory resemble properties of more standard rational choice explanations, Hedström (2005: p. 41) seemingly does not consider this theoretical approach as a variant, but more as an alternative to rational choice theory and the discussion on this is still lively in the recent literature (Opp, 2013a,b; Manzo, 2013; Ylikoski, 2013). Note that experimental economists (e.g., Gächter, 2013) describe their approaches as based on preferences, beliefs, and constraints, which not by coincidence resembles the terminology in the DBO theory of Hedström.

Also different learning theories and agent-based simulation models (see Macy and Flache, 2009; Epstein, 2006) are variants of theories in which actors behave according to well-defined behavioral rules although these rules are mostly not rational, but more adaptive in nature and based on experiences actors had in the past. Still, they are worth mentioning here, because although the behavioral or micro-level assumptions are quite different, the macro-level consequences of these assumptions occasionally coincide with predictions from strict rational choice models. For example, the evolutionary stable strategies in evolutionary game theory (a variant of game theory in which indeed the actors react on past

experiences rather than that they base behavior on rational expectations about others) are a subset of the set of Nash equilibrium strategies under a broad range of assumptions (Weibull, 1995). While this is an example where indeed macro-level outcomes of other behavioral rules coincide with rational behavior, many of the adaptations of micro-level assumptions indeed lead to different macro-level predictions. For example, simple divisions of cakes (so-called dictator game outcomes) in the laboratory show that virtually all subjects give more than negligible parts of the cake to other actors in completely anonymous settings and without any future interactions. This behavior cannot be explained through selfish rational choice explanations, but can be explained by models of inequality aversion.

Links to Empirical Research

The value of all the extensions discussed earlier can only be judged by the empirical evidence that corroborates the new predictions of these extensions. Rational choice researchers have more and more learned to pay attention to the empirical testability and value of their theories after criticisms as posed by Green and Shapiro (1994). As a reaction on these criticisms that rational choice research was merely a collection of theoretical exercises, the amount of empirical research testing rational choice models has been seriously extended in the last decades. Related to the models on behavior in embedded social dilemmas, for example, Buskens and Raub (2013) provide an extensive overview of empirical research focusing on trust situations. Their overview illustrates the breadth of empirical research on rational choice models varying from laboratory experiments, via vignette experiments to also survey studies. The laboratory experiments on embeddedness effects in trust situations show that different mechanisms simultaneously affect trust. Actors learn from their own past experiences and others' past experiences whom to trust. Hypotheses on this are derived from learning arguments as mentioned earlier. Also without having previous experiences, actors are more trustful if they expect to have more future interactions with a partner. This corroborates the related hypothesis based on repeated game arguments that allow for cooperative behavior if enough future interactions are expected. The same hypotheses are then shown to be also supported in so-called vignette experiments. These are contextualized hypothetical decision situations for which students as well as experienced professionals answer questionnaires. Finally, the same hypotheses are tested and supported in a survey on buyer–seller transactions, which can also be considered trust situations. We claim that such a combined set of tests of the same hypotheses provides a really strong point in case for the related theory, because the theory is supported in settings in which the tests are very close to the specific causal mechanism as well as in actual daily decision making that is not hampered by the external validity problem. We have to indicate that actually one hypothesis is consistently not supported in our overview and that is the hypothesis that trust increases with possibilities for reputation building through a social network. This hypothesis also follows from an extension of the repeated game model on networks. The consistent lack of support of the hypothesis in various empirical contexts is a strong case that the theory needs to be made more complex for understanding these consistent deviations from the simpler model. One of our conjectures is that the issue is related to rationality, because the reputation effect via social networks is the most complex effect to be incorporated by the decision makers.

Certainly related to extension of the rational choice model at the micro level, the extent of empirical research investigating these extensions has been flourishing especially in the laboratory. As has been realized by many researchers, stretching the assumptions on the utility function by, for example, social preferences opens the possibility to explain post-hoc any behavior through a suitable utility argument (see Gächter, 2013 for a recent account of this). The same problem is true for relaxing assumptions on rationality. As a counter force, Gächter (2013; see also Fehr and Gintis, 2007, for an earlier account and Camerer, 2003, for an extensive overview) discusses the increasing possibilities of experimental research (as done in experimental economics, but also in many related disciplines). Laboratory experiments provide an empirical tool to critically and under controlled conditions study preferences of actors or other possible deviations from the micro-level assumptions of the standard rational choice model. Fehr and Schmidt (1999) advocate their inequality aversion model based on the claim that this relatively simple model explains behavior in many different games in the laboratory. Although this claim has been challenged for this particular case (Binmore and Shaked, 2010), the idea that extensions to the micro-level assumptions should be tested in a variety of contexts is exactly what is necessary to avoid theoretical particularism. Related overviews of experiments on variations of games (Fehr and Gintis, 2007; Gächter and Thöni, 2011; Gächter, 2013) show that some behavioral assumptions seem to reappear regularly in different contexts, while different settings also seem to have their own particularities for which not yet more general explanations are available.

Still, empirical research in the laboratory is only one way to test rational choice theory and in itself not a sufficient way. Because there are serious concerns on the external validity of laboratory experiments, which are often done only with university students, the real proof of the pudding is in putting the theories also at a test outside the laboratory (see Gächter and Thöni, 2011, for a more extensive review of advantages and limitations of experimental research including further references). The recent volumes on rational choice theory and related approaches (Hedström and Bearman, 2009; Wittek et al., 2013) illustrate that also in that respect rational choice research is well on the way.

Looking Ahead

Overlooking the developments of rational choice theory in the last decades especially also after the criticism of Green and Shapiro (1994), rational choice researchers have certainly taken up the glove and strengthened the connection between rational choice theory and empirical findings. In particular, the strong extension of experimental studies (Jackson and Cox, 2013) has enabled rational choice

theorists to test prediction under controlled conditions and provided the potential to measure directly preferences and beliefs. Another consequence of this is that the acceptance of 'thick' rational choice theories rather than 'thin' rational choice theories (Hechter and Kanazawa, 1997) has gained an importance, because the thin rational choice theories do not only seem to violate assumptions of behavior at the individual level, but also do not predict outcomes at the macro level accurately.

So although these developments in rational choice theory have advanced the state of the art quite a bit in the last decades, still several pitfalls need to be avoided. Although the laboratory work indeed is ideal to pinpoint behavioral regularities under different circumstances, the laboratory should not become the main empirical playground for sociology. While in earlier times, theorists were blamed for developing new models only for the sake of new models without an eye at where the model is about, experimentalists sometimes have the tendency to develop new experiments for the sake of the experiments. Such development would still keep the real-world applications out of sight. Experimental tests are meant to test specific theoretical predictions, which in the end should predict behavior also outside the laboratory. Thus, external validity issues of experimental tested are an important concern for social scientists. Therefore, not only the same theoretical extensions of the standard rational choice model should find support in many different experiments, but the same extensions should also be supported with data collected outside the laboratory. It would be of great value to future rational choice research if the same hypotheses would be tested with a multitude of empirical set-ups and researchers would be parsimonious with extensions especially at the micro level until empirical evidence accumulates that a specific hypothesis is not confirmed or that there is repeated evidence that specific behavioral assumptions are violated.

A final issue that might come up given the tendencies to make rational choice theories thicker is that by adding complexities, the consistency of the total theoretical set-up should be kept in mind. As the close reader might have noticed in our account of the studies on trust, some hypotheses were derived from learning models and other hypotheses from more classic game-theoretic models. For combining learning elements in standard game-theoretic models, incomplete information models need to be considered, which again pose very strong assumptions on the rationality of actors, for example, in terms of rationally of updating beliefs. Although the example illustrated that deriving all hypotheses within the same theoretical framework is not always self-evident, freely shopping from different potential extensions to 'thicken' rational choice models should be discouraged as well.

See also: Bounded Rationality; Decision and Choice: Bounded Rationality; Experimental Economics; Intentionality and Rationality; Irrationality: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Rational Choice and Organization Theory; Rationality in Society; Social Dilemmas, Psychology of; Trust, Sociology of.

Bibliography

Aksoy, O., Weesie, J., 2012. Beliefs about the social orientations of others: a parametric test of the triangle, false consensus, and cone hypotheses. Journal of Experimental Social Psychology 48, 45–54.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Binmore, K., Shaked, A., 2010. Experimental economics: where next? Journal of Economic Behavior & Organization 73, 87–100.

Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90, 166–193.

Buskens, V., Raub, W., 2013. Rational choice research on social dilemmas: embeddedness effects on trust. In: Wittek, R., Snijders, T.A.B., Nee, V. (Eds.), Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA, pp. 113–150.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Russell Sage, New York.

Coleman, J.S., 1987. Free riders and zealots. In: Cook, K.S. (Ed.), Social Exchange Theory. Sage, Newbury Park, CA, pp. 59–82.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.

Epstein, J., 2006. Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press, Princeton, NJ.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.

Fehr, E., Gintis, H., 2007. Human motivation and social cooperation: experimental and analytic foundations. Annual Review of Sociology 33, 43–64.

Fehr, E., Schmidt, K., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868.

Friedman, J.W., 1971. A non-cooperative equilibrium for supergames. Review of Economic Studies 28, 1–12.

Gächter, S., Thöni, C., 2011. Micromotives, microstructure and macrobehavior: the case of voluntary cooperation. Journal of Mathematical Sociology 35, 26–65.

Gächter, S., 2013. Rationality, social preferences, and strategic decision-making from a behavioral economics perspective. In: Wittek, R., Snijders, T.A.B., Nee, V. (Eds.), Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA, pp. 33–70.

Gigerenzer, G., Selten, R. (Eds.), 2001. Bounded Rationality: The Adaptative Toolbox. MIT Press, Cambridge, MA.

Gintis, H., 2009. The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences. Princeton University Press, Princeton, NJ.

Goldthorpe, J.H., 2000. On Sociology: Numbers, Narratives, and the Integration of Research and Theory. Oxford University Press, Oxford.

Granovetter, M.S., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.

Green, D.P., Shapiro, I., 1994. Pathologies of Rational Choice Theory: A Critique of Applications in Political Science. Yale University Press, New Haven.

Hechter, M., Kanazawa, S., 1997. Sociological rational choice theory. Annual Review of Sociology 23, 191–213.

Hedström, P., 2005. Dissecting the Social: On the Principles of Analytical Sociology. Cambridge University Press, Cambridge.

Hedström, P., Bearman, P. (Eds.), 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.

Jackson, M., Cox, D.R., 2013. The principles of experimental design and their application in sociology. Annual Review of Sociology 39, 27–49.

Kreps, D.M., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoners' dilemma. Journal of Economic theory 27, 245–252.

Kroneberg, C., Kalter, F., 2012. Rational choice theory and empirical research: methodological and theoretical contributions in Europe. Annual Review of Sociology 38, 73–92.

Lindenberg, S., 2013. Social rationality, self-regulation, and well-being. In: Wittek, R., Snijders, T.A.B., Nee, V. (Eds.), Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA, pp. 72–112.

Macy, M.W., Flache, A., 2009. Social dynamics from the bottom up: agent-based models of social interaction. In: Hedström, P., Bearman, P. (Eds.), The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, pp. 245–268.

Manzo, G., 2013. Is rational choice theory still a rational choice of theory? Social Science Information 52, 361–382.

Nash, J., 1951. Non-cooperative games. Annals of Mathematics 54, 286–295.

Opp, K.-D., 2013a. What is analytical sociology? Strengths and weaknesses of a new sociological research program. Social Science Information 52, 329–360.

Opp, K.-D., 2013b. Rational choice theory, the logic of explanation, middlerange theories and analytical sociology: a reply to Gianluca Manzo and Petri Ylikoski. Social Science Information 52, 394–408.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.

Rapoport, A., 1974. Prisoner's dilemma – recollections and observations. In: Rapoport, A. (Ed.), Game Theory as a Theory of Conflict Resolution. Reidel, Dordrecht, pp. 18–34.

Rasmusen, E., 2007. Games and Information: An Introduction to Game Theory, fourth ed. Blackwell, Oxford.

Raub, W., Weesie, J., 1990. Reputation and efficiency in social interactions: an example of network effects. American Journal of Sociology 96, 626–654.

Raub, W., Buskens, V., Van Assen, M., 2011. Micro–macro links and micro-foundations in sociology. Journal of Mathematical Sociology 35, 1–25.

Simon, H., 1982. Models of Bounded Rationality. MIT Press, Cambridge, MA.

Van Lange, P., 2000. Beyond self-interest: a set of propositions relevant to interpersonal orientations. European Review of Social Psychology 11, 297–331.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Weibull, J., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.

Wittek, R., Snijders, T.A.B., Nee, V. (Eds.), 2013. Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA.

Ylikoski, P., 2013. The (hopefully) last stand of the covering-law theory: a reply to Opp. Social Science Information 52, 383–393.

Rationality in Society

Andreas Flache and Jacob Dijkstra, University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.A. Simon, volume 19, pp. 12782–12786, © 2001, Elsevier Ltd.

Abstract

Contemporary theories of rational behavior in human society augment the orthodox model of rationality both by adding various forms of bounded rationality and by relaxing the assumptions of self-interest and materialistic preferences. This article discusses how these extensions of the theory of rational behavior are increasingly applied to model rational behavior in society. First, the baseline model of rational behavior and its extensions are discussed, with particular emphasis on heuristics, learning theory, and theories of social preferences. Subsequently, applications to rational behavior in societal realms are addressed. Three areas of application are discussed in particular: production of collective goods, social networks, and rationality in organizations.

The orthodox model of rationality construes social actors as decision makers who always select the action that optimally serves their goals, given their information and the perceived costs and benefits of the available courses of action. Contemporary theories of rational behavior in human society augment this model in several ways (Wittek et al., 2013; *see* Rational Choice Theory in Sociology). Such modifications are necessary for several reasons. One reason is that empirical evidence shows that people have limited information processing and computational capabilities and find making multidimensional comparisons difficult. This has led to a range of models of 'bounded rationality' (Simon, 1997). Moreover, the orthodox model is silent on the contents of the goals rational people should strive for. Applications of the orthodox model of rational man, rooting in neo-classical economics (e.g., Cournot, 1838) typically fill this gap with two key assumptions on human goals in social behavior: actors are assumed to be selfish and exclusively interested in tangible material outcomes. But many social scientists criticize these assumptions because models of rationality based on them fail to explain many forms of real social behavior observed. In response to this, a range of theories of rational social behavior have been developed that attribute partially altruistic and nonmaterial preferences to human social actors, while retaining the core assumption that individuals strive to optimally attain these goals, albeit with bounded rationality. This article discusses how these extensions of the theory of rational behavior are increasingly applied to model rational behavior in society. We start from a discussion of the baseline model of rational behavior and its extensions. Subsequently, applications to rational behavior in societal realms are addressed. We focus on three areas of application: production of collective goods, social networks, and organizations.

The Theory of Rationality and Its Extensions

Rationality as Utility Maximization

Utility maximization and its extension of expected utility maximization (von Neumann and Morgenstern, 1944) are the best developed formal notions of rationality and form the core of neoclassical economics. Utility maximization as such does not explicitly refer to the social context of action. It postulates a utility function, which measures the degree to which an individual's (aggregate) goals are achieved as a result of their actions. The rational actor chooses the action, from among those given, which maximizes utility given his (rationally formed) expectations. If the actor's goals are food and sleep, then rationality calls for choosing the attainable combination of food and sleep that yields the greatest utility. But the core theory does not specify the content of the utility function (Lindenberg and Frey, 1993; Simon, 1983).

Without a concrete specification of what it is that generates utility, the theory has few refutable consequences since almost any choice can be rationalized by assuming a convenient goal. Thus, to increase the theory's empirical content additional assumptions must be made. Economic applications usually add the vastly oversimplified assumption that utility is measured by total income or wealth.

Without auxiliary assumptions, the theory is also silent about the processes used by the actors to find the optimal action. Neoclassical economics typically employs the assumption of perfect rationality in this respect. Rational actors never fail to find the action that maximizes their utility, even if this requires unlimited capacities to process and memorize all information available and to have unlimited foresight of the consequences of all available courses of actions in a distant future. More recent extensions of the theory of rationality relax both assumptions about actors' goals and assumptions concerning the process of decision making. We discuss both types of extensions below.

Bounded Rationality: Heuristics and Learning

An action may be regarded as rational because of the way it was chosen (procedural rationality) or because of its outcomes (substantive rationality) (*see also* Decision and Choice: Economic Psychology; Heuristics for Decision and Choice). Attributions of rationality may or may not respect the limits of actors' abilities to adapt actions to goals. Utility-maximizing behavior selects an action that is among those objectively available and attains a real-world optimum. As an alternative, bounded rational behavior selects some action from among those that memory, observation, and thought bring to attention. It is aimed at attaining outcomes that satisfice (attain aspirations) for the wants and needs that have

been evoked (Conlisk, 1996). There is a close affinity between optimizing and substantive rationality and an affinity between bounded and procedural rationality. The utility-maximizing optimum is independent of the choice process; the outcome is all that counts. In bounded rationality, the choice depends upon the process leading up to it. The social sciences, other than neoclassical economics, usually explicitly incorporate human cognitive limitations and social influences into their theories, using some version (seldom highly formalized) of bounded rationality. Bounded and procedural rationality require generating courses of action, assessing their possible outcomes under different futures and estimating how much satisfaction they will attain. They postulate procedures that respect human limits on knowledge of the present and future, on abilities to calculate the implications of knowledge, and on abilities to evoke relevant goals (Simon, 1983). Two important classes of models of bounded rationality focus on heuristics people use to make choices and on backward-looking learning processes.

Heuristics

Heuristics are the actual mental rules humans employ when making decisions such as choosing an action from a set of alternatives or deciding when to stop searching for a better option. A simple illustration is 'one-reason decision making' according to which individuals make choices based on the first relevant criterion that discriminates between the alternatives (Gigerenzer and Todd, 1999). Heuristics generally deviate from optimizing behavior in the sense that individuals are not assumed to gather all relevant information about all available alternatives before making a decision, even when such an exhaustive search was possible. Heuristics are adaptive decision-making rules (Gigerenzer and Selten, 2001) because they are in accord not only with human cognitive abilities but also with the information structure of the environment (Simon, 1956). Since they deviate from optimizing behavior, heuristics are necessarily linked to satisficing behavior (March and Simon, 1958). Originally conceptualized as 'deviations from perfect rationality' in the biases and heuristics program of Kahneman et al. (1982), work on heuristic decision making has shown that simple heuristics can actually outperform more advanced decision models. Recent developments in the decision literature focus on a 'dual mode model of decision making' distinguishing two mental systems, one fast, requiring little effort, and driven by heuristics, and one slow, effortful, and driven by conscious deliberation (Kahneman, 2011). Examining the interactions between these two systems and the conditions that determine which of the two systems is activated is a main frontier of this field of research.

Learning

The most simple approach to learning is reinforcement. Thorndike (1898) first proposed the 'Law of Effect.' If an action is associated with a favorable outcome, this makes it more likely that the decision maker repeats that action. Otherwise, alternative choices will be explored more likely. Reinforcement theory relaxes three key behavioral assumptions in models of forward-looking rationality (Macy et al., 2013). First, association replaces causality as the link between choices and payoffs. Decision makers explore the likely consequences of alternative choices and develop a propensity for those associated with better outcomes. But the outcomes that guide the exploration are those that have already been experienced, not those that a perfectly rational actor might anticipate to occur based on logical deduction. Second, reward and punishment replace utility as the motivation for choice. Decisions are guided toward better outcomes by two distinct cognitive mechanisms, approach (driven by reward), and avoidance (driven by punishment). This distinction means that aspiration levels are very important for learning theory. The effect of an outcome depends decisively on whether it is evaluated as gain or loss, satisfactory or unsatisfactory, or pleasant or aversive. Third, similar to heuristic decision making, reinforcement learning implies satisficing (March and Simon, 1958), not maximizing. Choices with satisfactory outcomes are likely to be repeated even if other choices have higher utility. In contrast, unsatisfactory outcomes induce searching for alternative outcomes, including a tendency to revisit alternative choices whose outcomes are even worse, a pattern called 'dissatisficing.' (Macy et al., 2013).

Reinforcement learning can lead actors out of 'social traps' where perfect rationality would entail socially undesirable outcomes. An important example is cooperation in a finitely repeated Prisoner's dilemma game (Macy and Flache, 2002). For perfectly rational actors, mutual defection in such a game is unavoidable. Whether the other player cooperates or not, own defection always leads to an outcome with a higher utility than cooperation does. However, if both players were to resist the temptation to defect, both would be better off. Reinforcement learners can escape from this trap if their aspirations are such that they are satisfied with an outcome in which the other player cooperates and are dissatisfied otherwise. In that case, actors will feel punished as long as the other player defects, resulting in avoidance of whatever action they have chosen themselves. As a consequence, actors keep changing their behavior until an outcome of mutual cooperation occurs. Then, the behavior of mutual cooperation becomes self-reinforcing because for both players it is associated with a satisfactory outcome. Unilateral defection would still yield a higher utility in this situation, but due to satisficing actors will not explore this possibility.

The theory of reinforcement learning is the most elementary theory of bounded rationality based on learning. In recent years, more sophisticated formal models of learning behavior have been developed that all share the principle that decision optimization is based on experience rather than anticipation. Theories differ in the extent to which rational forward-looking decision making is combined with learning. Bayesian belief learning and fictitious play, for example, are models of how decision makers form beliefs about the expected behavior of other actors in their social environment, to then determine the best reply given these beliefs (see Camerer, 2003; Macy et al., 2013; for brief overviews).

Goals

Social science and behavioral economics research has shown that the assumptions of selfish and of purely materialistic goals are empirically untenable. Research on social preferences has

cast doubt on the selfishness assumption, demonstrating that human decision makers do take the outcomes for other individuals into account for nonstrategic reasons. Forsythe et al. (1994) compare behavior in ultimatum and dictator games showing that pure altruism plays a considerable part in human (strategic) decisions. Several theories have been developed to account for findings like these, including and combining pure altruism (Andreoni and Miller, 2002) with motives such as spite (Levine, 1998) and aversion against unequal distributions, even when in favor of oneself (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

Regarding the standard assumption of materialistic preferences research has shown that human decisions are shaped not only by the distribution of outcomes but also by the perceived fairness of the social process through which the outcomes come about (Rabin, 1993). Charness and Rabin (2002) and Falk and Fischbacher (2006) proposed important models formalizing the idea that in social interactions human actors evaluate an outcome based on the perceived (un)kindness of actions taken by the other individuals involved. Moreover, the perceived (un)kindness of these actions is determined not just by their consequences but by the intentions attributed to the others. Intended (un)kindness then triggers a desire to reciprocate.

A main conclusion from this literature is that humans are heterogeneous in terms of social preferences, leading to classifications of actors in terms of having for instance selfish, inequality averse, surplus-maximizing, or competitive preferences. These distinctions between preference types chime with the social value orientation concept from social psychology (cf. Bogaert et al., 2008; *see also* Social Psychology, Sociological) that classifies humans as proselfs and prosocials. An important classification when it comes to intentions and reciprocity is the one between 'strong reciprocators' and selfish individuals (Fehr and Gintis, 2007), where the former are willing to punish and reward unkind and kind behavior, respectively, at a net cost to themselves. Importantly, when humans interact in groups the composition of the group in terms of social preferences and the interaction between these preferences and the institutional environment have a strong impact on behavior and outcomes.

Rationality in Social Context

In the following, it will be discussed how extensions of the orthodox model have informed the analysis of rationality in different social contexts, notably in the production of collective goods in society, in actor's embeddedness in social networks, and in rational behavior in organizations. More in general, the application of the model of rationality to the analysis of societal realms follows the approach of methodological individualism (*see also* Methodological Individualism in Sociology), where the aim is to explain societal outcomes from assumptions about individual behavior in social contexts.

Collective Goods and Rational Behavior

Collective goods benefit all individuals involved, but cooperation is required to bring them about. The problem is that cooperation is costly for those who contribute and cannot easily be enforced. Many authors therefore see the provision of collective goods as an expression of solidarity (Hechter, 1987; *see also* Solidarity, Sociology of). The reason why cooperation in the provision of collective goods is problematic is that collective goods are characterized to varying degree by 'jointness of supply' (i.e., consumption of the good by a given individual does not reduce the amount available to other individuals) and 'nonexcludability' (i.e., individuals who benefit from the good cannot be excluded from the consumption of the good, even if they did not contribute). 'Free rider' problems at the workplace and environmental pollution are two examples where cooperation fails due to the conflict between individual and collective interests. Nevertheless, for many collective goods a degree of cooperation is observed that is hard to explain by self-interest alone.

Orthodox models predict that (large) groups of individuals will fail to produce collective goods at a sufficient level in the absence of selective incentives (Olson, 1965). This is because the direct gains an individual obtains from her own contribution fall short from compensating the costs. Selective incentives (such as cheap insurance for union members) can alleviate this free-rider problem if they only benefit those who contribute (e.g., pay their union membership fees). However, the work of Marwell and Oliver (1993) on collective action theoretically establishes that the relations between rationality, group size, and the production of collective goods is much more complex than suggested by orthodox treatments. Their approach was developed with particular emphasis on explaining the dynamics of social movements (*see also* Movements, Social). They adopt a bounded rationality perspective in the sense that actors will invest if the immediate marginal gains for them exceed their marginal costs. With this, Marwell and Oliver exclude more sophisticated strategic calculations in which actors may withhold contributions in order to let others shoulder the burden, even if they themselves also would gain from investment. Based on this assumption, Marwell and Oliver show that rational actors can make considerable levels of contributions to the collective good if costs and benefits of contributions depend nonlinearly on the amount of investments already made (modeled in the 'production function' of the collective good) and there is heterogeneity amongst potential contributors in terms of their interest in and ability to contribute to the collective good. For example, models imply that a few particularly interested or resourceful individuals would be willing to shoulder high start-up costs in producing a collective good. They then form a 'critical mass' (Marwell and Oliver, 1993) that triggers subsequent contributions of less resourceful or less interested members, because under an 'accelerating' production function contributing becomes sufficiently cost-effective for the less interested or less resourceful once the collective action has successfully started up. A prototypical example for this are the dynamics of demonstrations against oppressive regimes (Opp and Gern, 1993) in which the risks from the participation decrease and the potential for success increases as more people join in.

In addition to considerations concerning the production function of collective goods much experimental evidence falsifies the orthodox prediction even for situations in which production functions are such that contributions are never cost-effective in material terms (Ledyard, 1995). Famously, Fehr and Gächter (2002) showed that allowing group members to

punish each other greatly improves the chances for successful collective good production even when the experimental set up assured that imposing punishment does not pay off for the enforcer. The authors explain this finding with the strong social preferences of enforcers.

The seminal work of Elinor Ostrom (1990) focuses on how the right social and institutional circumstances can overcome free rider problems in local communities. She offers a set of design principles that are conducive to good governance of 'common pool resources' (CPRs), such as drink water basins, forests, and community irrigation systems. For instance, clear lines of demarcation between users and nonusers are necessary; rules for the governance of the CPR should be attuned to local physical and social conditions, local needs, and local knowledge, and these rules should be drafted by community members themselves; regional and national governments should respect these local rules and accord the local communities sufficient autonomy; community members should monitor and sanction each other and an accessible and low-cost institution for conflict resolution should be created. While not explicated in her work, Fehr and Gächter's theory of strong reciprocity suggests an answer to the question why community members could be motivated to engage in such monitoring and sanctioning in the first place. Ostrom then focuses on the institutional features that further sustain monitoring and sanctioning. She also emphasized the importance of norms (*see also* Norms) and trust (*see also* Trust, Sociology of) between community members as social resources of a community. Importantly, in later work Ostrom et al. (1999) emphasize that the creation of local institutions is also a viable strategy for creating global collective goods, such as slowing down the pace of global warming. In these global cases, local institutions create systems of polycentric governance, which offer points of application for higher-level governance.

Rationality in Social Networks

Networks and Cooperation

A key theme in studies of rational behavior in networks is the relationship between social networks (*see also* Social Networks) and the solution of cooperation problems. Networks are of especial significance in two realms of cooperation. In bilateral interactions, such as between buyers and sellers, or in interfirm alliances, experiments with trust games and surveys have shown how the embeddedness of interacting partners in on-going network relations (between each other and with third parties) stabilizes the trust that is needed for cooperation (Buskens and Raub, 2002). Rational choice theories explain this through rational reciprocity based on conditional cooperation (Axelrod, 1984). Network relations serve as conduits for the flow of reputational information about other peoples' behavior (Buskens and Raub, 2002). Seen from this perspective, networks that allow for a fast and reliable spread of information (e.g., centralized networks) provide incentives for actors to cooperate, because their cooperation avoids possible future sanctions or disruption of a mutually beneficial reciprocal exchange. In the case of production of collective goods, research also has demonstrated how networks foster cooperation, for example in political movements (Opp and Gern, 1993) or organizational settings. Explanations from a rationality perspective in this research again invoke reputational effects on conditional cooperation and also refer to network relations as instruments to impose rewards or sanctions in terms of, e.g., social approval. But from a bounded rationality perspective using learning models, networks can also impede cooperation. In a collective good setting, free riders can learn to use their friends' dependence on social approval from the free rider to deter social sanctions (Flache and Macy, 1996). In terms of the learning approach used in this research, boundedly rational learners find it easier to coordinate on mutual social approval in a bilateral relationship than to coordinate on the more complex multilateral exchange of contributions to the collective good for social rewards. Satisficing then implies that mutual approval without provision of the collective good may be an outcome that is good enough for the actors, while mutual approval combined with collective good provision would still be superior. The possibility that trustful social relations hinder social control empirically resonates in the negative effects of high trust on monitoring that have been found in research on teamwork (Langfred, 2004).

Strategic Formation of Social Networks

Studies of the role of networks in collective action typically treat networks as static and exogenously given. An important new frontier in research on rationality in social networks addresses the role of rationality in the formation of social networks. A prominent line of research conceptualizes network evolution as a process driven by the strategic interaction of rational individuals (Dutta and Jackson, 2003). Models in this field typically assume that individuals' payoffs depend on structural properties of their network, also beyond their direct network neighborhood. A prototypical example is the formation of coauthorship networks. For example, decisions of one author to establish a coauthorship relationship with a colleague can affect the outcomes both of her current coauthors (because she may invest less time into other projects) as well as of the colleagues' coauthors (because the colleague may likewise invest less effort). Severance of an existing network link is typically a unilateral decision, but formation of a new link is either a unilateral or a bilateral decision, depending on the type of model. Individuals are assumed to make link formation and severance decisions independently and are assumed to be motivated by the desire to maximize the utility they obtain from the networks they create. In the models in this literature, the degree of rationality assumed is frequently very low. Typically, individuals are assumed to be entirely myopic in the sense of not anticipating on the reactions of other network members on their networking choices. In a seminal paper Jackson and Wolinsky (1996) show that under these assumptions there is generally a tension between efficiency and stability. Thus, in these models network formation exhibits central characteristics of a social dilemma: many networks that are stable (in the sense that no individual or pair of individuals can unilaterally or bilaterally profitably change their network connections) are not efficient (in the sense of yielding the highest overall payoffs). For example, in a coauthorship setting it may be individually rational to maximize the number of coauthorship

relations because every additional project increases scientific visibility, all other things being equal. At the same time, when everyone follows this strategy this may result in a reduction of overall productivity due to excessive costs of coordination of multiple projects.

Rationality in Organizations

In a modern society, formal work organizations – companies, government agencies, schools – provide environments for a large fraction of human waking hours (Simon, 1997). From a rationality perspective, organizations face difficult problems of coordination and cooperation between employees as well as between employers and employees (*see also* Organizations, Sociology of). At the same time, organizational structures can be seen as institutions (*see also* Institutions; New Institutionalism in the Analysis of Complex Organizations) that form an important part of the solution of these problems. Two prominent theoretical approaches that aim to explain features of organizations from a rationality perspective are principal-agent theory (Petersen, 1993) and transaction-cost theory (Williamson, 1981). Principal-agent theory views members of an organization as perfectly rational actors, typically with materialistic and strictly self-interested preferences. From this perspective, employees would only invest effort for organizational goals if the organization assures that 'shirking' provides inferior outcomes for the employee. Principal-agent theory analyses which features of payment and career systems provide this assurance given its model of man. For example, Petersen (1992) argued that this can explain why under group-payment systems higher levels of effort are obtained under target-rate payment schemes as opposed to piece-rate payment schemes in manufacturing industries. Broadly, the argument is that under a target-rate payment scheme every worker's contribution to the collective effort may be crucial for whether the target is met and thus the group bonus is obtained. In a piece-rate scheme, instead, the gains for an individual worker from investing effort are too small to compensate for costs, because the relatively small pay-increment achieved is spread across the entire group. Similar to Marwell and Oliver's (1993) nonlinear production function, target-rate payment thus reduces incentives to free ride in group production. Principal-agent theory also explains features of the career systems of many organizations. Often, organizations require employees to make large investments in terms of effort, training, and education in early stages of their career for comparatively low wages. Only at later stages of the career higher wages can be obtained. Principal agent approaches argue that organizations choose such structures to avoid 'adverse selection' in a situation of asymmetric information where the employer cannot know the true qualities and motivations of the employee. Only employees who are sufficiently motivated and have enough trust in their own abilities will stay with the organization in the early career phase in such a system, and other employees will self-select into jobs that bear lower risk but also lower payment in later career stages. The career system thus solves the cooperation problem employers and employees face. It assures that the 'right' employees are hired by the firm into a long-term career path which – in turn – provides for the firm the incentive to hire these employees in the first place even when information about their characteristics is inaccurate or incomplete.

Transaction cost theory differs from principal-agent approaches in assuming strongly bounded rationality. The theory then analyzes which features of a 'governance structure,' such as employment contracts or contracts between firms, are best suited to efficiently solve the cooperation problems between economically interdependent boundedly rational actors. The inability of boundedly rational actors to oversee all contingencies of an economic transaction makes contracting incomplete and provides room for 'opportunistic' behavior. According to transaction cost theory, this is the most important reason why hierarchical organizations exist. By integrating the different parts of a production process into a hierarchy, opportunism problems are more efficiently solved through means of hierarchical control. This is especially important when transaction-specific investments in human capital or technology create 'asset specificity.' That is, they create dependencies between contract partners that may give room for opportunistic behavior. For example, when employees need to make high investments to obtain skills that can only be used within a specific firm, this may deter an employee from obtaining the necessary training in the first place. The reason can be fear by the employee that the employer may later abuse the employee's dependence on the firm by making employment conditions less favorable for the employee. In such a case, a long-term labor contract with, e.g., guaranteed wages and career prospects can resolve the cooperation problem to the mutual benefit of both parties. While principal-agent theory and the transaction cost approach typically focus on monetary and contractual incentives in organizations, more recent elaborations add the importance of organizational features that appeal to social and altruistic motivations, such as socialization of employees into company cultures, professional identities, or professional norms (*see also* Norms). One early example of a formal elaboration of such 'soft' instruments of governance in a rational choice framework is the work of Kandel and Lazear (1992).

Conclusion

The goals rational actors pursue as well as the process of rational decision making itself are both shaped by their social environment and help to explain outcomes of social action. Recent developments in the theory of (partially) altruistic motivations, as well as heuristic decision making and learning, help to understand how cooperation in collective goods is possible and how social networks arise, change, and affect cooperative behavior. Moreover, theories of rational behavior in organizations help to understand the structure of organizations and the conditions under which employers and employees can successfully cooperate. Models of rational behavior suitable for understanding rationality in society turn out to be more complex and have more degrees of freedom than the orthodox model of man of neoclassical economics. Current work in rational choice social research (Wittek et al., 2013) focuses on identifying which mix of assumptions makes the rationality perspective most fruitful for addressing a particular societal context, as well as on moving toward an

integration of different perspectives into a general model of rationality in society.

See also: Action, Theories of Social; Decision and Choice: Bounded Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Heuristics; Game Theory; Institutions; Methodological Individualism in Sociology; Movements, Social; New Institutionalism in the Analysis of Complex Organizations; Norms; Organizations, Sociology of; Rational Choice Theory in Sociology; Social Networks; Social Psychology, Sociological; Solidarity, Sociology of; Trust, Sociology of.

Bibliography

Andreoni, J., Miller, J., 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70 (2), 737–753.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Bogaert, S., Boone, C., Declerck, C., 2008. Social value orientation and cooperation in social dilemmas: a review and conceptual model. British Journal of Social Psychology 47 (3), 453–480.

Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. American Economic Review 90 (1), 166–193.

Buskens, V., Raub, W., 2002. Embedded trust: control and learning. In: Thye, S.R., Lawler, E.J. (Eds.), Group Cohesion, Trust and Solidarity, Advances in Group Processes, vol. 19. Elsevier Science, Amsterdam, pp. 167–202.

Camerer, C., 2003. Behavioral Game Theory: Experiments on Strategic Interaction. Princeton University Press, Princeton.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 114 (3), 817–869.

Conlisk, J., 1996. Why bounded rationality? Journal of Economic Literature 34, 669–700.

Cournot, A., 1838. Researches into the Mathematical Principles of the Theory of Wealth. Reprinted 1960. Augustus M. Kelley, New York.

Dutta, B., Jackson, M.O. (Eds.), 2003. Networks and Groups: Models of Strategic Formation. Springer, Heidelberg.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293–315.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.

Fehr, E., Gintis, H., 2007. Human motivation and social cooperation: experimental and analytical foundations. Annual Review of Sociology 33, 43–64.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114 (3), 817–868.

Flache, A., Macy, M.W., 1996. The weakness of strong ties: collective action failure in a highly cohesive group. Journal of Mathematical Sociology 21, 3–28.

Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games and Economic Behavior 6, 347–369.

Gigerenzer, G., Selten, R. (Eds.), 2001. Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge, MA.

Gigerenzer, G., Todd, P.M., The ABC Research Group, 1999. Simple Heuristics That Make Us Smart. Oxford University Press, New York.

Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.

Jackson, M.O., Wolinsky, A., 1996. A strategic model of social and economic networks. Journal of Economic Theory 71, 44–74.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.

Kandel, E., Lazear, E.P., 1992. Peer pressure in partnerships. Journal of Political Economy 100, 801–817.

Langfred, C.W., 2004. Too much of a good thing? Negative effects of high trust and individual autonomy in self-managing teams. Academy of Management Journal 47 (3), 385–399.

Ledyard, J., 1995. Public goods experiments. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton, NJ, pp. 111–194.

Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593–622.

Lindenberg, S., Frey, B., 1993. Alternatives, frames, and relative prices: a broader view of rational choice. Acta Sociologica 36, 191–205.

Macy, M.W., Flache, A., 2002. Learning dynamics in social dilemmas. Proceedings of the National Academy of Sciences United States of America 99 (10), 7229–7236.

Macy, M.W., Benard, S., Flache, A., 2013. Learning. In: Edmonds, B., Meyer, R. (Eds.), Simulating Social Complexity. A Handbook. Series: Understanding Complex Systems. Springer, New York, pp. 431–452.

March, J.G., Simon, H.A., 1958. Organizations. Wiley, New York.

Marwell, G., Oliver, P., 1993. The Critical Mass in Collective Action. A Micro-Social Theory. Cambridge University Press, Cambridge.

von Neumann, J., Morgenstern, O., 1944. The Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Opp, K., Gern, C., 1993. Dissident groups, personal networks and spontaneous cooperation: the East German revolution of 1989. American Sociological Review 58, 659–680.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge UK.

Ostrom, E., Burger, J., Field, C.B., Norgaard, R.B., Policansky, D., 1999. Revisiting the commons: local lessons, global challenges. Science 284, 278–282.

Petersen, T., 1992. Individual, collective and systems rationality in work groups: dilemmas and market-type solutions. American Journal of Sociology 98 (3), 469–510.

Petersen, T., 1993. The economics of organization: the principal-agent relationship. Acta Sociologica 36, 277–293.

Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83 (5), 1281–1302.

Simon, H.A., 1956. Rational choice and the structure of environments. Psychological Review 63, 129–138.

Simon, H.A., 1983. Reason in Human Affairs. Stanford University Press, Stanford, CA.

Simon, H.A., 1997. Models of bounded rationality. In: Empirically Grounded Economic Reason, vol. 3. MIT Press, Cambridge, MA.

Thorndike, E.L., 1898. Animal intelligence: an experimental study of the associative processes in animals. The Psychological Review, Monograph Supplements, No. 8. Macmillan, New York.

Tversky, A., Kahnemann, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Williamson, O.E., 1981. The economics of organization: the transaction cost approach. American Journal of Sociology 87, 548–577.

Wittek, R.P.M., Snijders, T., Nee, V., 2013. Introduction: rational choice social research. In: Wittek, R.P.M., Snijders, T., Nee, V. (Eds.), The Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA.

Rawls, John (1921–2002)

Elaine Kellman, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rawls is often acknowledged as one of the most important political philosophers of the twentieth Century. Rawls' seminal work, *A Theory of Justice*, recreates social contract theory as a hypothetical thought experiment, thereby forming the basis for Rawls' liberal conception of justice as 'Justice as Fairness.' In his second book, *Political Liberalism*, Rawls transfers 'Justice as Fairness' into the political realm and considers the stability of 'Justice as Fairness' in a world of diverging religious, moral, and philosophical beliefs. Rawls' third and final book, *The Law of Peoples*, considers the application of 'Justice as Fairness' to global justice. This bio-entry aims to provide a short introduction to Rawls' life and his work, and to introduce readers to Rawls' most significant concepts including, without limitation, 'the original position,' 'veil of ignorance,' 'principles of justice,' 'overlapping consensus' and 'reasonable pluralism.'

Early Life and Education

Family Origins and Early Education

Rawls was born in 1921 in Baltimore, the second of five sons. His father, despite dropping out of law school, worked hard and became a successful and respected lawyer. His mother, heavily involved in the League of Women voters, fought for women's rights. Both his parents were very interested in politics and justice, and this permutated Rawls' early years.

Unfortunately Rawls' otherwise uneventful early life was twice struck by tragedy: In 1928, Rawls' younger brother Bobby caught diphtheria from Rawls, and sadly passed away. The incident shocked Rawls, and, according to Pogge, Rawls' mother thought it may have triggered the stammer that Rawls had for the rest of his life. Unfortunately a year later, tragedy repeated itself. Rawls' younger brother Tommy contracted pneumonia from Rawls and passed away while Rawls was recovering.

After completing his schooling in Baltimore and Connecticut, Rawls joined his older brother at Princeton. Rawls studied philosophy with Norman Malcom (previously Wittgenstein's student). He wrote his senior thesis on the topic of human evil and seriously considered studying for the priesthood. However, after graduation, like most of his classmates, he joined the army.

Wartime Experiences

Rawls was sent as an infantry man to the Pacific, where he served in Japan, New Guinea, and the Philippines. For Rawls his army service was 'singularly undistinguished;' however he was awarded a Bronze Star for work behind enemy lines. It is widely reported that in Japan, Rawls witnessed the devastated city of Hiroshima and almost 50 years after the war, Rawls would refer to the destruction of Hiroshima as a great evil (Rawls 'Fifty Years After Hiroshima' (1995) in *Collected Papers*). It is clear that Rawls' experiences in the army had a profound effect on him and Thomas Nagel wrote that Rawls' outlook was strongly marked by his experiences of "the deadly fanaticism of the Japanese army, the murderous romanticism of the Nazis and the calculated massacre of civilians by the Allies in Tokyo, Hiroshima, Nagasaki and Dresden."

Rawls' experiences in the army also lead to his disillusionment with religion. In 'On my Religion' (1997) in *A Brief Inquiry into the Meaning of Sin and Faith* Rawls details three incidents that influenced this disillusionment: The first occurred following a particularly brutal period of combat. A Lutheran pastor gave a brief sermon to Rawl's unit, during which Rawls recalled him saying "God aimed our bullets at the Japanese while God protected us from theirs." Rawls did not believe this was an appropriate use of Christian doctrine. The second was the death of Rawls's friend, Deacon, which perhaps was particularly profound for Rawls as he so easily could have been in Deacon's place. A sergeant had sought volunteers: one to go with a colonel to look at Japanese positions, the other to give blood. Rawls and his friend Deacon volunteered. As Rawls had the right blood type he gave blood while Deacon accompanied the colonel. Unfortunately, Deacon and the colonel came under fire from the Japanese and were killed by a mortar shell while sheltering.

The third incident was news of the Holocaust which made Rawls question how he could: "…pray and ask God to help me, or my family, or my country or any other cherished thing I cared about when God would not save millions of Jews from Hitler?"

Post-War Academic Life

Rawls was offered an officer position at the end of the war but declined. Having given up any thoughts of joining a seminary, he returned to Princeton in 1946 and after a year's fellowship at Cornell, completed his PhD at Princeton. While completing his thesis Rawls met Margaret Fox, his wife to be. They married in 1949 and, later, had four children, two boys and two girls.

Rawls taught at Princeton until 1952 when he received a Fulbright Fellowship to attend Oxford University. At Oxford he met and was influenced by HLA Hart, Isaiah Berlin, and Stuart Hampshire. Hart's influence in particular perhaps can be seen in Rawls' second published essay 'Two Concepts of Rules' (1955). Influenced by his time at Oxford, Rawls continued to develop his theory of political justice and he published widely with a flurry of articles on justice appearing between 1958 and 1969.

After his time at Oxford, Rawls obtained a faculty position at Cornell University. He was also a visiting professor at Harvard and a professor at MIT before finally becoming a permanent member of the philosophy department faculty at Harvard from 1962.

In the 1960s Rawls, perhaps influenced by his own memories of war, strongly criticized the Vietnam War. In particular, he opposed the 2-S deferment. This deferment allowed college students to avoid the conscription to focus on their studies but, in practice, it was misappropriated to enable children with wealthy backgrounds to avoid conscription. Rawls believed that this was not only unfair, but logically that a society would be less likely to engage in unjust wars if the children of political elite were on the front lines. In 1966, Rawls therefore introduced a controversial resolution to the Harvard faculty condemning the deferment.

Rawls retired in 1991 but continued to teach until he was unfortunately largely incapacitated in 1995 by the first in a series of strokes. Despite this, he continued to work, and with assistance he completed: a new introduction to *Political Liberalism* (1996, hereafter *PL*); the Idea of Public Reason Revisited; and most substantial of all his third book, *The Law of* Peoples (1999, hereafter *LP*). In addition, he participated in the editing of his collected papers and of three collections of his lectures. Not one to be idle in respect of causes he supported, in 1997, Rawls also joined five other philosophers in a brief to the US Supreme Court supporting a 'right to die.' The Court did not support a constitutional right to die, but nor did it prevent states from establishing laws that would establish such a right. Many supporters of physician-assisted suicide considered this a huge victory.

Rawls passed away at his home in Massachusetts in 2002.

Short Description of Major Works

Main Achievements

Although Rawls published articles and papers throughout his life, the focal point of Rawls' work is contained within three books. The first, *A Theory of Justice* (1971), sets out the basic principles of Rawls' theory of 'Justice as Fairness.' In 1999, Rawls issued a revised edition of *A Theory of Justice* (and all references to *TJ* hereafter shall refer to this revised edition) (and he later further summarized his modification of justice as fairness in *Justice as Fairness: A Restatement* (2001)). The second, *PL*, transfers justice as fairness from a metaphysical to political setting. The third, *LP*, applies Rawls' earlier work to global justice.

A Theory of Justice: Pre-Publication

Rawls is best known for his theory of 'Justice as Fairness.' Rawls began formulating his ideas of Justice as Fairness following the completion of his PhD, but it was not until 20 years later that he detailed his theory in *TJ*.

During that 20-year development period Rawls published many articles, and, in *TJ*, Rawls tried to coherently bring together some of those arguments. Accordingly, the first part of *TJ* elaborates on Rawls' earlier essays 'Justice as Fairness' (1958) (which set out the core of his argument of 'Justice as Fairness') and 'Distributive Justice: Some Addenda' (1968). The three chapters of the second part broadly correspond, respectively, to 'Constitutional Liberty' (1963), 'Distributive Justice' (1967) and 'Civil Disobedience' (1966) and the second chapter of last part covers subjects of 'The Sense of Justice' (1963). It is also worth mentioning two other essays that were important precursors to Rawls' Justice as Fairness. Rawls first began to develop his ideas of reflective equilibrium in his essay 'Outline of a Decision Procedure for Ethics' (1951), and in 'Two Concepts of Rules' (1955) he reframed utilitarianism as a theory specifically addressed to problems of political philosophy.

A Theory of Justice: In Context

Rawls developed his ideas for *TJ* during the 1950s, a time when welfare-state institutions were first introduced and liberals were seeking philosophical justification for their belief that social policies should not be driven by profit, but should be to the benefit of all. Concurrently, the concept of individual rights was becoming dominant. Against this background, Rawls theory was therefore unique in supplying both a commitment to individual rights, and an argument for social justice.

However, Rawls' aim in a *TJ* was to work out a conception of justice that provided an alternative to utilitarianism, which was then the predominant theory of moral and political philosophy. For Rawls, an account of the basic rights and liberties of citizens as free and equal persons was vital to any theory of democratic institutions, and utilitarianism was unable to deliver such an account.

Social Justice and Institutions

Utilitarianism is often interpreted as a complete moral philosophy. However, Rawls addresses a narrower conception of utilitarianism, utilitarianism as a theory of 'social justice.' Accordingly, the primary subject of justice for Rawls is the 'basic structure of society,' and he therefore seeks to evaluate whether the institutions that constitute the basic structure of society are just. Institutions for Rawls are the economic, political, and social systems or arrangements that distribute fundamental rights and duties (what Rawls calls 'primary social goods') and determine the division of advantages in society. Taken together institutions therefore define a citizen's rights and duties, thereby influencing their life plan and prospects. The basic structure is the primary subject of Rawls theory of justice because its effects are so profound.

'Original Position' and the 'Veil of Ignorance'

Traditional social contract theories were extremely unpopular at the start of Rawls' career. However, Rawls, a great admirer of Kant, realized that the sometimes hard-to-decipher concepts in Kant and Locke's writings could be revised as a new theory that could potentially challenge utilitarianism. Rawls therefore reincarnated the social contract as the thought experiment he called the 'original position.' The original position establishes a hypothetical scenario in which a group of equal participants are tasked with reaching an agreement (i.e., the social contract) about the basic structure of a society (and in particular the

principles assigning basic rights and duties and the division of social benefits), which they are, once an agreement has been reached, to occupy.

Rawls wanted the principles of justice to be fair to all citizens. To force participants to select only principles that would be fair to everyone, Rawls determined that participants in the original position would need to agree the principles from behind a 'veil of ignorance.' The intention of the veil of ignorance is to shroud the participants from the knowledge of the actual position they will occupy in society and what their particular interests and inclinations will be. By preventing participants from shaping the principles by reference to personal advantage, participants can only proceed on the basis of securing to the greatest possible degree fairness for all, including themselves. The result of the veil of ignorance is that each participant is equal and no one is able to design principles to favor his or her position. Rawls called this the 'perspective of eternity.' Rawls argued that in this perspective, it would be irrational for participants to propose any ideas that that are innately unjust, for example, racism and sexism.

The specific knowledge that Rawls envisages the veil of ignorance stripping from participants in the original position is their: position in society; class or social status; natural assets or abilities (e.g., intelligence and strength); personal conception of the 'good'; personal life plan; and psychological inclination. In addition, participants will not know the economic and political situation of their society, the level of civilization and culture attained by their society, or the generation to which they belong. The only knowledge the participants retain is that they possess certain attributes and capacities. For example, each participant will know that they have the capacity to construe, hold, and revise 'a conception of the good,' or life plan – even if they are not aware what the content of that good is. In addition each participant is aware that they have the ability to develop a sense of justice and will generally comply with it.

Principles of Justice

The majority of *TJ* is devoted to considering which principles of justice would be chosen by participants in the original position. Rawls believes the participants would take a low-risk approach, agreeing to principles that are fundamentally equal, and that would guarantee for all the highest possible minimum levels of freedom, wealth, and opportunity – even if this means lowering the overall or average levels. In this situation, Rawls is therefore adamant the principle of utility would not be chosen. Stressing the inherent selfishness of people, he believes participants would not agree to a principle that may require lesser life prospects for some if it results in a greater sum of advantages for others.

Instead, Rawls formulates two new principles of justice that he believes people in the original position would choose. The first "requires equality in the assignment of basic rights and duties, while the second holds that social and economic inequalities, for example inequalities of wealth and authority are just only if they result in compensating benefits for everyone and in particular for the least advantaged members of society."

The initial formulation of the first principle, sometimes referred to as the Liberty Principle or Equality Principle, establishes an equality in basic liberties for all citizens and it is expressed in *TJ* as:

Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.

For Rawls basic liberties tend toward civil and political rights, as examples he mentions political liberty (the right to vote and to be eligible for public office); freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person along with the right to hold (personal) property (which is defined in terms of moral capacities and self-respect rather than a natural right of ownership); and freedom from arbitrary arrest and seizure as defined by the rule of law.

Rawls' aims for the second principle are to ensure that basic liberties are available to all and to ensure distributive justice. Accordingly, the focus of the second principle is the distribution of wealth and opportunity and the design of institutions. Like the first principle, Rawls reevaluated and revised the second principle throughout his work, but at its most basic, it consists of two subprinciples. The first, which aims to ensure that persons with comparable talents and motivations face similar life chances, is often referred to as 'The Fair Equality of Opportunity Principle,' the sometimes controversial second, the intention of which is to ensure that inequalities in society work to the benefit of the least advantaged, is the 'Difference Principle.' For Rawls the 'least advantaged' are those that are: (1) persons whose family and class origins are more disadvantaged than others; (2) persons whose natural endowments (as realized) permit them to fare less well; and (3) those whose fortune and luck in the course of life turn out to be disadvantaged.

The second principle is formulated in *TJ* as follows:

Social and economic inequalities are to be arranged so that they are both (1) to the greatest expected benefit of the least advantaged and (2) attached to offices and positions open to all under conditions of fair equality of opportunity.

Rawls stressed that the first and second principles are 'lexically' ordered. The first principle, of liberty, has priority over the second, while the second will always be prior to concerns of 'efficiency' or 'maximization of advantage' and the acceptance of inequality.

The choice of these principles presupposes that participants in the original position are rational. The priority of liberty is, for Rawls, the rational choice because it will benefit each participant by allowing him or her to attain their ultimate success, notwithstanding the parameters of their actual situation. In addition, the acceptance of inequality is also a consequence of the participants rational self-interest. For Rawls each participant aims individually for the best for themselves. Although they seek to ensure the best possible potential position for themselves, in light of the cautious and prudent psychology built into humans, they ensure there is a safeguard for those participants that unfortunately are dealt a hand of disadvantage. This demonstrates the operation of the 'maximin' principle (maximizing the position of the least advantaged) that threads throughout Rawls' writings.

Application of the Principles of Justice

Underlying Rawls' discussion of the principles of justice is his intention that these two principles will steer the design of the

institutions that comprise the basic structure of society. Rawls therefore delineates a four-stage sequence for the application of his principles. If followed, Rawls believes the resulting institutions will be just. The four stages in sequence are the following:

1. An enunciation of principles of justice from the original position.
2. A lifting of the veil of ignorance so that participants become aware of the general conditions of society, but not their own individual circumstances, in order that participants can devise a constitutional system that addresses, in accordance with the principles, the powers of government and rights and duties of citizens. Rawls stresses that this process must among other criteria address the existence of different, and possibly divergent, political viewpoints. However, in consideration of the priority of liberty, Rawls' contention is that the outcome will be a constitutional democracy.
3. The third stage is the legislative stage. Legislation should be drawn up in accordance with the principles of justice as well as constitutional procedures. Again legislators are required to act in light of general, and not personal, interests.
4. Finally, the laws and rules that have been legislated will be applied by judges and administrators – and will govern the actions of citizens generally. At this stage the veil of ignorance is removed completely.

Reflective Equilibrium

TJ also introduced Rawls' concept of 'reflective equilibrium,' and an approach to achieving a state of reflective equilibrium is used in *TJ* as a method for the justification (and to some extent construction) of the principles. Wenar describes reflective equilibrium as a state where "all of one's beliefs on all levels of generality agree perfectly with one another" and he further clarifies that "in reflective equilibrium one's abstract beliefs explain one's more general convictions which in turn explain one's specific judgments." In effect, Rawls' method of achieving a state of 'reflective equilibrium' is a technique for the deliberation and consideration of morality. At its center is a procedure of constant reevaluation, testing of principles against considered judgments, and vice versa.

Rawls' method of achieving reflective equilibrium can be summarized as follows, using justice as an example. First, a group of considered judgments about justice (generally based on secure intuitions) are identified. Attempts are then undertaken to explain, and justify, these judgments through the formulation of abstract principles of justice. Finally, any discrepancy between the principles and considered judgments is addressed. This could be either by modifying the principles or revising existing considered judgments (and for Rawls even judgments that are initially regarded as fixed are liable to revision).

Rawls believed that by going back and forth between these steps and constantly testing ideas, eventually abstract principles would be yielded that matched and explained considered judgments. The 'ideal state' achieved following this method was what Rawls termed reflective equilibrium, stating "it is an equilibrium because at last principles and judgments coincide; and it is reflective since it is clear what principles the judgments confirm and the premises of their derivation."

Political Liberalism: From the Metaphysical to the Political

After *TJ*, Rawls continued to work on his theory of justice as fairness, publishing more articles and revising extending and defending his theory. Rawls also began to reconsider some aspects of his theory and these reconsiderations initially appeared in a series of papers, the three most important are *Kantian Constructivism in Moral Theory* (1980), *Social Unity and Primary Goods* (1982), and *Justice as Fairness: Political not Metaphysical* (1985). Eventually, over 20 years after the publication of *TJ*, Rawls published his second book, *PL*.

Although the theory of justice as fairness remains at the heart of *PL*, Rawls refocuses the theory so that it (1) moves from a metaphysical to political sphere and (2) can be used as a means of coexistence by people who have different theories of the good. *PL* therefore converts the doctrine of justice as fairness presented in *TJ* into a political conception of justice. This reinterpretation of the theory is also continued in Rawls later, unfinished work Justice as Fairness: A Restatement (2001).

At the heart of *PL* is the question of stability and, more particularly, whether a society ordered by the two principles of justice as fairness can endure. Although Part III of *TJ* attempted to give an account of stability, Rawls came to view this account as defective and inconsistent with his theory as a whole and he recognized that the political conception offered in *TJ* had to be extended so that it worked even when faced with differing doctrines and conceptions of the good. In *PL*, Rawls therefore considers how it is possible for a just and stable society to endure, despite diverging philosophical, religious, and moral views held by citizens, what Rawls, borrowing from Cohen, calls 'reasonable pluralism.'

For Rawls, 'reasonable pluralism' is the existence of the various moral and metaphysical 'comprehensive doctrines' that characterize a society. These metaphysical comprehensive doctrines are the systems of beliefs that give life its meaning or value, for example, religious or philosophical doctrines. Although these doctrines can coexist, they are generally incompatible, with often insurmountable differences. The 'Reasonable' in reasonable pluralism refers to 'reasonable' comprehensive doctrines. For Rawls, a reasonable doctrine is one which (1) relies on the employment of practical and theoretical reason, (2) is conscious of its fallibility, and (3) is open to possibility of change.

In the context of reasonable pluralism, the legitimacy of any theory of social justice is, according to Rawls, dependent on whether it can serve as the focus of an 'overlapping consensus.' To fall within Rawls' concept of an 'overlapping consensus' the theory must therefore remain not only neutral between reasonable comprehensive doctrines, but also political, that is, nonmoral, freestanding, and independent of any comprehensive doctrine. Simultaneously, the theory must be acceptable to all in the political community and able to constitute a basis of their beliefs. Any legitimate theory therefore must be such that it can be accommodated within any comprehensive doctrine and acceptable to all.

Of course, for Rawls, Justice as Fairness was the single conception of social justice that could fit within all reasonable comprehensive doctrines and could serve as the focus of an overlapping consensus. To achieve this, some alterations to Justice as Fairness are rendered in *PL*. In particular, to ensure that Rawls' political conception of justice is free-standing (that is not derived from, or part of, any comprehensive doctrine), Rawls modifies his account of primary goods and his argument for the lexical ordering of the two principles. Importantly, Rawls also alters the principles of justice themselves. Although the first principle retains priority over the second, Rawls determined that in the new lexical ordering the first half of the second principle would have priority over the second half. The revised principles are therefore set out as follows:

1. Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal basic liberties, and only those liberties, are to be guaranteed their fair value.
2. Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity; second, they are to be to the greatest benefit of the least advantaged members of society.

The modifications are subtle but important. Most notably, Rawls switches the two parts of the second principle, so that the difference principle becomes subordinate to not only the first principle but also the condition of fair equality of opportunity.

The Law of Peoples: Global Justice and International Relations

Before LP, Rawls only briefly mentioned the application of his theory on a global basis: in *TJ*, he suggested the original position could be used to generate principles of global justice, in the same way it was used to generate the two principles of social justice. However, he made this comment in passing as a way of supplying some possible nonreligious grounds for conscientious refusal. Accordingly, there is no substantial indication as to what principles such a procedure would generate.

Following this, two significant attempts were made to apply the reasoning behind justice as fairness to global justice: Beitz's *Political Theory and International Relations* (1979) and Pogge's *Realising Rawls* (1989). Reaching similar conclusions, both argued that people from all states should be included in the original position. The participants in this global original position would then select principles of global justice in accordance with Rawls' theory of Justice as Fairness. Beitz and Pogge both argue that arguments in favor of the application of the difference principle in a state must be equally valid across states. Consequently, they advocate redistribution from wealthy and advantaged societies, to poor and less advantaged societies.

In *LP*, Rawls decided to address the problem of global justice and international relations. Rawls saw human rights as distinct from liberal rights and did not see it to be the role of liberal peoples to impose liberalism upon 'decent' societies. Many were surprised that Rawls did not therefore endorse Beitz and Pogge's conclusions and he did not apply the difference principle globally, nor endorse a global principle of distributive justice.

Instead, Rawls envisaged that there would be two conceptions of the original position. The first original position would take place separately in each state, as set out in *TJ*. Thereafter, Rawls envisaged that well-ordered societies (by which he meant states whose institutions respect the rights and interests of their citizens) would each send a representative to a 'global original position.' Participants in this global original position would still be subject to a veil of ignorance, except they would be aware that they were not representing the interests of all societies or all citizens but only well-ordered societies. Rawls believes this system would result in principles of global justice that commit to state autonomy rather than global redistribution. Although this account was a more accurate reflection of international affairs it divided Rawls' supporters, with some accusing Rawls of making unjustified compromises.

Lasting Impact/Influence

Rawls is often credited with single-handedly reviving the study of political philosophy the popularity of which, before his writing, was in steady decline. However, Rawls' influence is not limited to the spheres of political science or philosophy but crosses disciplinary lines receiving serious attention from legal scholars, economists, theologians and sociologists as well as political scientists and philosophers.

Although not all agree with Rawls' ideas, his contribution to the character and form of political philosophy has been lasting. Rawls inadvertently reframed the boundaries of political philosophy with debates today almost always formulated in Rawls' language. In addition, Rawls' conception of justice has generated a cornucopia of secondary literature, inspiring numerous academic debates, and different approaches. There are too many to mention here, but worthy of note is Nozick's *Anarchy, State and Utopia* (1974), in which Nozick argues against Rawls' theory of justice as fairness from a broadly libertarian point of view, promoting individual rights as an absolute standard, with primacy over any unwanted interference or other moral and political considerations; Dworkin's *Taking Rights Seriously* (1977) and Ackerman's *Social Justice in the Liberal State* (1980) which present an alternative libertarian egalitarian standpoint; Sandel's *Liberalism and the Limits of Justice* (1982), Walzer's *Spheres of Justice: A Defense of Pluralism and Equality* (1983), and MacIntrye's *After Virtue* (2007) all of which are inspiring communitarian-influenced critiques and/or rework Rawls' theories; Pateman's *The Sexual Contract* (1988) and Okin's *Justice Gender & The Family* (1989) which critique Rawls from a feminist's standpoint, and finally the influence of Rawl's seminal ideas can also be gleaned in Sen's *Inequality Re-examined* (1992) and Nussbaum's *Women and Human Development: The Capabilities Approach* (2000).

See also: Equality of Opportunity; Institutions; Justice: Philosophical Aspects; Kant, Immanuel (1724–1804); Kantian Ethics and Politics; Liberalism, Anthropology of; Liberalism:

Impact on Social Science; Political Science, History of; Political Theory, History of; Public Reason; Reflective Equilibrium; Sandel, Michael (1953–); Social Justice: Historical and Theoretical Considerations; Utilitarian Social Thought, History of; Walzer, Michael (1935–).

Bibliography

Ackerman, B., 1980. Social Justice in the Liberal State. Yale University Press, New Haven, CT.

Aybar, S., Harlan, J., Lee, W., 1991. John rawls: for the record. Available at: Harvard Review of Philosophy http://www.hcs.harvard.edu/~hrp/issues/1991/Rawls.pdf.

Barry, B., 1973. The Liberal Theory of Justice. Oxford University Press, Oxford, UK.

Beitz, C., 1979. Political Theory and International Relations. Princeton University Press, Princeton, UK.

Dworkin, R., 1977. Taking Rights Seriously. Harvard University Press, Cambridge, MA.

Freeman, S. (Ed.), 2003. The Cambridge Companion to Rawls. Cambridge University Press, Cambridge, UK.

Freeman, S., 2007. Justice and the Social Contract: Essays on Rawlsian Political Philosophy. Oxford University Press, Oxford, UK.

Kant, I., 1785. In: Gregor, M. (Ed.), The Metaphysics of Morals. Cambridge University Press, Cambridge, UK.

Lovett, F., 2011. Rawls's a Theory of Justice: A Reader's Guide Continuum. International Publishing Group, London, UK.

Mandle, J., 2009. Rawls's a Theory of Justice: An Introduction. Cambridge University Press, Cambridge, MA.

McIntrye, A., 2007. After Virtue: A Study in Moral Theory, third ed. University of Notre Dame Press, Notre Dame, IN.

Nagel, T., 1999. The Rigorous Compassion of John Rawls: Justice, Justice, Shalt Thou Pursue in the New Republic. http://business.highbeam.com/4776/article-1G1-56259807/rigorous-compassion-john-rawls-justice-justice-shalt.

Nozick, R., 1974. Anarchy, State and Utopia. Basic Books, New York.

Nussbaum, M.C., 2000. Women and Human Development: The Capabilities Approach. Cambridge University Press, Cambridge, UK.

Okin, S., 1989. Justice, Gender and the Family. Basic Books, New York.

Pateman, C., 1988. The Sexual Contract. Polity Press, Cambridge, UK.

Pogge, T., 1989. Realising Rawls Cornell. University Press, Ithaca, NY.

Pogge, T., 2007. John Rawls: His Life and Theory of Justice. Oxford University Press, Oxford, UK.

Rawls, J., 1996. Political Liberalism. Colombia University Press, New York.

Rawls, J., 1999a. A Theory of Justice, revised ed. Belknap Press, Cambridge, MA.

Rawls, J., 1999b. The Law of Peoples: With "The Idea of Public Reason Revisited". Harvard University Press, Cambridge, MA.

Rawls, J., 1999c. In: Freeman, S. (Ed.), Collected Papers. Harvard University Press, Cambridge, MA.

Rawls, J., 2000. Lectures on the History of Moral Philosophy. Harvard University Press, Cambridge, MA.

Rawls, J., 2001. In: Kelly, E. (Ed.), Justice as Fairness: A Restatement. Belknap Press, Cambridge, MA.

Rawls, J., 2009. In: Nagel, T. (Ed.), A Brief Inquiry into the Meaning of Sin and Faith: With "On My Religion". Harvard University Press, Cambridge, MA.

Richardson, H. John Rawls the Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/rawls/.

Rousseau, J., 1987. On the Social Contract in the Basic Political Writings. Hackett Publishing, Indianapolis, IN.

Sandel, M., 1998. Liberalism and the Limits of Justice, second ed. Cambridge Univesity Press, Cambridge, UK.

Sen, A., 1992. Inequality Re-examined. Oxford University Press, Oxford, UK.

Walzer, M., 1983. Spheres of Justice: A Defense of Pluralism and Equality. Basic Books, New York.

Walzer, M., 1993. Interpretation and Social Criticism. Harvard University Press, Cambridge, MA.

Wenar, L., 2013. John Rawls the Stanford Encyclopeadia of Philosophy. http://plato.stanford.edu/archives/win2013/entries/rawls/.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 20

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 20

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 20 *xlv*

VOLUME 20

Reading Nonalphabetic Scripts 1
Takeshi Hatta

Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments 5
Heikki Lyytinen, Mikko Aro, Ulla Richardson, Jane Erskine, Hong Li, and Hua Shu

Real Estate Development 12
Richard Peiser

Realism, in Applied Research 20
George Julnes

Realism/Neorealism 26
Arthur A Stein

Recidivism 31
William D Bales, Kaleena J Burkes, Samuel JA Scaggs, and Catie L Clark

Recognition 37
Elaine J Mahoney, Allison E Nickel, and Deborah E Hannula

Recognition, Politics of: Perspectives from Social Work and Sociology 44
Stephen A Webb

Reconstructive Memory, Psychology of 50
Henry L Roediger, III, and Kurt A DeSoto

Recovery of Function: Dependency on Age 56
Bryan Kolb

Re-entry and Reintegration After Incarceration 61
Christy A Visher

Reference and Representation: Philosophical Aspects 67
Frank Jackson

Reference Group, History of 72
Barbara Hoenig

Reflective Equilibrium 77
Stelios Virvidakis

Reflexivity 82
Nick Gould

Reflexivity in Anthropology 88
George E Marcus

Reflexivity in Science and Technology Studies 93
Malcolm Ashmore

Reformation and Confessionalization 98
Ute Lotz-Heumann

Refugees in Anthropology 105
Marita Eastmond

Refugees: Geographical Aspects 111
Prem Kumar Rajaram

Region: Geographical 117
Andrew EG Jonas

Regional Geography 121
John Harrison

Regional Planning 129
Alex Lord and Mark Tewdwr-Jones

Regional Planning: China 134
Cecilia Wong

Regional Planning, History of 141
David A Johnson

Regional Science 146
Andrew Isserman and Roger R Stough

Regression Discontinuity Design 152
Joseph C Cappelleri and William M Trochim

Regulation and Administration 160
Bronwen Morgan and Susan Sterett

Regulation, Economic Theory of 166
David EM Sappington

Regulation, Empirical Analysis of 172
Nancy L Rose

Regulation, Political Economy of 178
Nancy L Rose

Regulatory Agencies 181
Christopher H Foreman

Rehabilitation (Cardiac): Psychosocial Aspects 187
Martyn Jones, Karen M Smith, and Oliver Herber

Rehabilitation of the Offender 193
Friedrich Lösel

Reification: History of the Concept 203
Frederic Vandenberghe

Reinforcement, Principle of 207
Ian M Evans

Reinforcement: Neurochemical Substrates 211
Larry Stein

Relapse Prevention Training 217
Peter H Wilson

Relational Social Work 221
Fabio Folgheraiter

Relativism, Cognitive 227
Diederick Raven

Relativism: Philosophical Aspects 233
Maria Baghramian

Reliability and Validity Assessment: New Approaches 239
Duane F Alwin

Reliability: Measurement 248
David Rindskopf

Religion and Abuse 253
Nancy Nason-Clark

Religion and Economic Life 258
Richard H Roberts

Religion and Health: Clinical Considerations and Applications 263
Harold G Koenig, Faten N Al Zaben, and Saad Al Shohaib

Religion and Health: Sociological Considerations 269
Ellen L Idler

Religion and New Media 273
Heidi A Campbell and Louise Connelly

Religion and Politics in the United States 279
Paul A Djupe and Andrew R Lewis

Religion and Sciences 284
Gary B Ferngren

Religion and Stratification 292
Lisa A Keister and Brad Fulton

Religion and Youth 298
Lisa D Pearce

Religion, Anthropology of 307
Simon Coleman

Religion: Evolution and Development 313
Robert N Bellah

Religion, Family, and Kinship 317
Annette Mahoney

Religion, History of 322
Friedrich W Graf and Astrid Reuter

Religion: Mobilization and Power 328
Darren E Sherkat

Religion: Morality and Social Control 333
John P Hoffmann

Religion: Nationalism and Identity 338
John Myhill

Religion, Phenomenology of 344
John R Hall and Joshua D Hayes

Religion: Psychiatric Aspects 348
Sheila M LoboPrabhu, Kenneth I Pargament, and James W Lomax

Religion, Psychology of 354
Fraser N Watts and Ryan J Williams

Religion, Sociology of 359
David Voas

Religions of East Asia 366
Ian Reader

Religions of India 373
Peter van der Veer

Religiosity 377
Darren E Sherkat

Religiosity and Prejudice against Minorities 381
Peer Scheepers and Rob Eisinga

Religious Belief, Evolution of 384
Armin W Geertz

Religious Fundamentalism: Cultural Concerns 396
Martin E Marty

Religious Nationalism and the Secular State 401
Mark Juergensmeyer

Religious Organizations 406
James A Beckford

Remedies and Damages: Legal Aspects 413
Theodore Eisenberg

Remote Sensing 419
Assaf Anyamba, J Estes, K Kline, and E Collins

Renaissance 425
Bernd Roeck

Repartnering and Stepchildren 430
Frances K Goldscheider

Repetitive Sex Offenders, Treatment of 434
Robert A Prentky and Raina V Lamade

Replication 441
Lowell L Hargens

Representations, Social Psychology of 443
Ivana Marková

Reproduction and Cultural Anthropology 450
Elizabeth Roberts

Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective 457
Sonia Corrêa

Reproductive Rights in Affluent Nations 468
Roxanne Mykitiuk and Robyn Lee

Republicanism: Philosophical Aspects 477
Mortimer NS Sellers

Reputation 483
Kurt Lang

Resampling Methods of Estimation 492
Bradley Efron, David Rogosa, and Robert Tibshirani

Research and Development in Organizations 496
Keith Pavitt and Mei-Chih Hu

Research and Theory Building in Social Work 502
Jane F Gilgun

Research Ethics, Cross-Cultural Dimensions of 508
Leonardo D de Castro, Peter A Sy, and Nur IB Soonaan

Research Ethics: Research 514
Elizabeth Yuko and Celia B Fisher

Research Funding: Ethical Aspects 523
Virginie Pirard

Research Publication: Ethical Aspects 529
Federico Di Trocchio

Residential Child Care 534
Andrew Kendrick

Residential Segregation in the United States 540
Joanna Marie Pinto-Coelho and Camille Z Charles

Residential Segregation: Geographic Aspects 544
Susan J Smith

Residential Segregation: Recent Trends 549
William AV Clark

Resilience 555
Christoph Steinebach

Resilient City 561
Danny MacKinnon

Resource Geography 568
Roger Hayter and Jerry Patchell

Respiratory Disorders: Psychosocial Aspects 576
Andreas von Leupoldt

Responsibility in Social Work 581
Viviene E Cree

Responsibility: Philosophical Aspects 587
Marina AL Oshana

Retirement and Encore Adulthood: The New Later Life Course 592
Phyllis Moen and Jack Lam

Retirement and Health 598
James Banks, Tarani Chandola, and Katey Matthews

Retirement, Economics of 602
Courtney C Coile

Retrieval 608
Asaf Gilboa

Retroactive Interference 613
Michael Craig, Michaela Dewar, and Sergio Della Sala

Revolution 621
Richard Stahler-Sholk

Revolution, Theories of 627
Rod Aya

Revolutions of 1989–90 in Eastern Central Europe 633
Timm Beichelt

Revolutions, History of 637
Jim Leith

Revolutions, Sociology of 642
John Markoff

Rhetorical Analysis 650
Peter Larsen

Rheumatoid Arthritis: Psychosocial Aspects 655
Christina Bode and Erik Taal

Ricardo, David (1772–1823) 660
Paul A Samuelson and Bryan S Turner

Ricoeur, Paul (1913–2005) 665
Jonathan Roberge

Riesman, David (1909–2002) 669
Christian Fleck

Right to the City: A Liberal-Democratic Perspective 673
Shenjing He

Rights: Legal Aspects 680
Michael McCann and Stuart Scheingold

Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice 686
Tim Blackman and Brid Featherstone

Risk Assessment in Social Work 692
Marilyn Callahan

Risk Screening, Testing, and Diagnosis: Ethical Aspects 698
Göran Hermerén

Risk, Sociology and Politics of 706
Steve Yearley

Risk: Social Psychological Perspectives 711
Dame GM Breakwell

Ritual 717
David J Parkin

Ritual and Symbolism, Archaeology of 721
Rosemary A Joyce

Robinson, Joan (1903–83) 727
Geoffrey C Harcourt and PM Kerr

Robustness 731
Stephan Morgenthaler

Rogers, Carl Ransom (1902–87) 737
Howard Kirschenbaum

Rokkan, Stein (1921–79) 744
Peter Flora

Roma People 748
Michael Stewart

Roma People: Discrimination and Social Work Practice 754
Špela Humljan Urh

Rorty, Richard (1931–2007) 761
Maria Baghramian and Fergal McHugh

Rousseau, Jean-Jacques (1712–78) 767
Jürgen Oelkers

Routine Activities, Criminal Opportunities, Crime and Crime Prevention 772
Pamela Wilcox

Rule of Law (and Rechtsstaat) 780
Martin Krygier

Rumors and Urban Legends 788
Pamela Donovan

Rural Geography 795
Jo Little

Rural Planning: General 801
Hemalata C Dandekar

Rural Social Work 807
Margaret Alston

Rural Sociology 814
Daniel T Lichter

Russian and Post-Soviet Studies: Culture 820
Alexander Etkind

Russian and Post-Soviet Studies: Education 825
Mark S Johnson

Russian and Post-Soviet Studies: Environment 831
Charles E Ziegler

Russian and Post-Soviet Studies: Gender 836
Lynne Attwood

Russian and Post-Soviet Studies: Politics 843
Jeffrey Hass

Russian and Post-Soviet Studies: Society 851
Andrew Buck

Russian Revolution, The 856
Manfred Hildermeier

Sacrifice 860
David Hicks

Safety, Economics of 863
John F Burton, Jr

Same-Sex Sexualities in Turkey 870
Cenk Özbay

Sample Survey Methodology, History of 875
Stephen E Fienberg and Judith M Tanur

Sample Surveys: Cognitive Aspects of Survey Design 881
Judith M Tanur

Sample Surveys: Model-Based Approaches 886
Sharon L Lohr

Sanctions in Political Science 890
David A Baldwin

Sandel, Michael (1953–) 893
Andrew Tsz Wan Hung

Sapir, Edward (1884–1939) 899
Victor Golla

Sapir-Whorf Hypothesis 903
John A Lucy

Sarmiento, Domingo Faustino (1811–88) 907
Carlos Altamirano

Satisfaction and Commitment in Long-Term Heterosexual Relationships 912
Aaron Ben-Ze'ev

Sauer, Carl Ortwin (1889–1975) 918
Michael Williams

Saussure, Ferdinand de (1857–1913) 921
Andrea Cossu

Savage, Leonard J (1917–71) 926
Dennis V Lindley

This page intentionally left blank

CONTRIBUTORS TO VOLUME 20

Saad Al Shohaib
King Abdulaziz University, Jeddah, Saudi Arabia

Margaret Alston
Monash University, Melbourne, VIC, Australia

Carlos Altamirano
Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

Duane F. Alwin
Pennsylvania State University, University Park, PA, USA

Faten N. Al Zaben
King Abdulaziz University, Jeddah, Saudi Arabia

Assaf Anyamba
NASA/Goddard Space Flight Center and Universities Space Research Association, Biospheric Sciences Laboratory, Greenbelt, MD, USA

Mikko Aro
University of Jyväskylä, Jyväskylä, Finland

Malcolm Ashmore
Loughborough University, Leicestershire, UK; and Universidad Nacional de Colombia, Bogotá, Colombia

Lynne Attwood
School of Arts, Languages and Cultures, University of Manchester, Manchester, UK; and Derbyshire, UK

Rod Aya
Amsterdam, Netherlands

Maria Baghramian
School of Philosophy, University College, Dublin, Ireland

David A. Baldwin
Columbia University, New York, NY, USA

William D. Bales
College of Criminology and Criminal Justice, The Florida State University, Tallahassee, FL, USA

James Banks
School of Social Sciences, University of Manchester, Manchester, UK

James A. Beckford
University of Warwick, Coventry, UK

Timm Beichelt
Europa-Universität Viadrina, Frankfurt/Oder, Germany

Robert N. Bellah
University of California, Berkeley, CA, USA

Aaron Ben-Ze'ev
University of Haifa, Haifa, Israel

Tim Blackman
The Open University, Milton Keynes, UK

Christina Bode
University of Twente, Enschede, The Netherlands

Dame G.M. Breakwell
University of Bath, Bath, UK

Andrew Buck
University of Southern Indiana, Evansville, IN, USA

Kaleena J. Burkes
College of Criminology and Criminal Justice, The Florida State University, Tallahassee, FL, USA

John F. Burton
Cornell University, Ithaca, NY, USA; and Rutgers University, New Brunswick, NJ, USA

Marilyn Callahan
School of Social Work, University of Victoria, Victoria, BC, Canada

Heidi A. Campbell
Texas A&M University, College Station, Texas, TX, USA

Joseph C. Cappelleri
Pfizer Inc, Groton, CT, USA

Tarani Chandola
Cathie Marsh Centre for Census and Social Research, University of Manchester, Manchester, UK

Camille Z. Charles
University of Pennsylvania, Philadelphia, PA, USA

Catie L. Clark
College of Criminology and Criminal Justice, The Florida State University, Tallahassee, FL, USA

William A.V. Clark
University of California, Los Angeles, CA, USA

Courtney C. Coile
Wellesley College, Wellesley, MA, USA

Simon Coleman
University of Toronto, Toronto, ON, Canada

E. Collins[†]
University of California, Santa Barbara, CA, USA

Louise Connelly
University of Edinburgh, Edinburgh, UK

Sonia Corrêa
The Brazilian Interdisciplinary from HIV/AIDS, Sexuality Policy Watch, Brazil

Andrea Cossu
University of Trento, Italy

Michael Craig
Human Cognitive Neuroscience, Psychology, University of Edinburgh, UK

Viviene E. Cree
The University of Edinburgh, Edinburgh, UK

Hemalata C. Dandekar
California Polytechnic State University, San Luis Obispo, CA, USA

Leonardo D. de Castro
Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Kurt A. DeSoto
Washington University, Saint Louis, MO, USA

Michaela Dewar
Psychology, School of Life Sciences, Heriot-Watt University, Edinburgh, UK; and Human Cognitive Neuroscience, Psychology, University of Edinburgh, UK

Federico Di Trocchio
Università di Roma Sapienza, Roma, Italy

Paul A. Djupe
Denison University, Granville, OH, USA

Pamela Donovan
Bloomsburg University of Pennsylvania, Bloomsburg, PA, USA

Marita Eastmond
University of Gothenburg, Göteborg, Sweden

Bradley Efron
Stanford University, Stanford, CA, USA

Theodore Eisenberg
Cornell University, Ithaca, NY, USA

Rob Eisinga
Radboud University Nijmegen, Nijmegen, The Netherlands

Jane Erskine
Banff, Aberdeenshire, UK

J. Estes
University of California, Santa Barbara, CA, USA

Alexander Etkind
European University Institute, Florence, Italy

Ian M. Evans
Massey University, Wellington, New Zealand

Brid Featherstone
The Open University, Milton Keynes, UK

Gary B. Ferngren
Oregon State University, Corvallis, OR, USA

Stephen E. Fienberg
Carnegie Mellon University, Pittsburgh, PA, USA

Celia B. Fisher
Center for Ethics Education, Fordham University, Bronx, NY, USA

Christian Fleck
University of Graz, Graz, Austria

Peter Flora
University of Mannheim, Mannheim, Germany

Fabio Folgheraiter
Università Cattolica del Sacro Cuore, Milano, Italy

Christopher H. Foreman
Maryland School of Public Policy, University of Maryland, College Park, MD, USA

Brad Fulton
Duke University, Durham, NC, USA

Armin W. Geertz
Aarhus University, Aarhus, Denmark

[†] Deceased.

Asaf Gilboa
Baycrest, Toronto, ON, Canada; and University of Toronto, Toronto, ON, Canada

Jane F. Gilgun
School of Social Work, University of Minnesota, Twin Cities, MN, USA

Frances K. Goldscheider
Brown University, Providence, RI, USA

Victor Golla
Humboldt State University, Arcata, CA, USA

Nick Gould
University of Bath, Bath, UK

Friedrich W. Graf
University of Munich, Munich, Germany

John R. Hall
University of California, Davis, CA, USA

Deborah E. Hannula
University of Wisconsin, Milwaukee, WI, USA

Geoffrey C. Harcourt
University of Cambridge, Cambridge, UK

Lowell L. Hargens
University of Washington, Seattle, WA, USA

John Harrison
Loughborough University, Loughborough, Leicestershire, UK

Jeffrey Hass
University of Richmond, Richmond, VA, USA; and St. Petersburg State University, St. Petersburg, Russia

Takeshi Hatta
Kansai University of Welfare Science and Nagoya University, Takatsuki City, Osaka, Japan

Joshua D. Hayes
University of California, Davis, CA, USA

Roger Hayter
Simon Fraser University, Burnaby, BC, Canada

Shenjing He
Sun Yat-sen University, Guangzhou, China

Oliver Herber
Institute of General Practice, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Göran Hermerén
Lund University, Lund, Sweden

David Hicks
State University of New York, Stony Brook, NY, USA

Manfred Hildermeier
Goettingen University, Goettingen, Germany

Barbara Hoenig
Innsbruck University, Innsbruck, Austria

John P. Hoffmann
Brigham Young University, Provo, UT, USA

Mei-Chih Hu
National Tsing Hua University, Hsinchu, Taiwan

Andrew Tsz Wan Hung
Hong Kong Community College, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Ellen L. Idler
Emory University, Atlanta, GA, USA

Andrew Isserman[†]
University of Illinois, Urbana, IL, USA; and George Mason University, Fairfax, VA, USA

Frank Jackson
Australian National University, Canberra, ACT, Australia

David A. Johnson
University of Tennessee, Knoxville, TN, USA

Mark S. Johnson
University of Wisconsin—Madison, Madison, WI, USA

Andrew E.G. Jonas
Department of Geography, Environment and Earth Sciences, University of Hull, Hull, UK

Martyn Jones
School of Nursing and Midwifery, University of Dundee, Dundee, UK

Rosemary A. Joyce
University of California, Berkeley, CA, USA

Mark Juergensmeyer
University of California, Santa Barbara, Santa Barbara, CA, USA

George Julnes
University of Baltimore, Baltimore, MD, USA

Lisa A. Keister
Duke University, Durham, NC, USA

Andrew Kendrick
School of Social Work and Social Policy, University of Strathclyde, Glasgow, Scotland, UK

P.M. Kerr
Nedlands, WA, Australia

[†] Deceased.

Howard Kirschenbaum
University of Rochester, Rochester, NY, USA

K. Kline
University of California, Santa Barbara, CA, USA

Harold G. Koenig
Duke University Medical Center, Durham, NC, USA; and King Abdulaziz University, Jeddah, Saudi Arabia

Bryan Kolb
Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

Martin Krygier
University of New South Wales, Sydney, NSW, Australia; and Regulatory Institutions Network, Australian National University, Acton, ACT, Australia

Jack Lam
University of Minnesota, Minneapolis, MN, USA

Raina V. Lamade
Fairleigh Dickinson University, Teaneck, NJ, USA

Kurt Lang
University of Washington, Seattle, WA, USA

Peter Larsen
University of Bergen, Bergen, Norway

Robyn Lee
Banting Postdoctoral Fellow, Brock University, St. Catharines, ON, Canada

Jim Leith
Queen's University, Kingston, ON, Canada

Andrew R. Lewis
University of Cincinnati, Cincinnati, OH, USA

Hong Li
Beijing Normal University, Beijing, China

Daniel T. Lichter
Cornell University, Ithaca, NY, USA

Dennis V. Lindley[†]
Minehead, Somerset, UK

Jo Little
College of Life and Environmental Sciences, University of Exeter, Exeter, UK

Sheila M. LoboPrabhu
Michael E. DeBakey V.A. Medical Center, Houston, TX, USA; and Baylor College of Medicine, Houston, TX, USA

Sharon L. Lohr
Arizona State University, Tempe, Arizona, AZ, USA

James W. Lomax
Baylor College of Medicine, Houston, TX, USA

Alex Lord
University of Liverpool, Liverpool, UK

Friedrich Lösel
University of Cambridge, Cambridge, UK; and University of Erlangen-Nuremberg, Erlangen, Germany

Ute Lotz-Heumann
University of Arizona, Tucson, AZ, USA

John A. Lucy
University of Chicago, Chicago, IL, USA

Heikki Lyytinen
University of Jyväskylä, Jyväskylä, Finland

Danny MacKinnon
Newcastle University, Newcastle Upon Tyne, UK

Annette Mahoney
Bowling Green State University, Bowling Green, OH, USA

Elaine J. Mahoney
University of Wisconsin, Milwaukee, WI, USA

George E. Marcus
University of California, Irvine, CA, USA

John Markoff
University of Pittsburgh, Pittsburgh, PA, USA

Ivana Marková
University of Stirling, Stirling, Scotland, UK

Martin E. Marty
University of Chicago, Chicago IL, USA

Katey Matthews
Cathie Marsh Centre for Census and Social Research, University of Manchester, Manchester, UK

Michael McCann
University of Washington, Seattle, WA, USA

Fergal McHugh
School of Philosophy, University College, Dublin, Ireland

Phyllis Moen
University of Minnesota, Minneapolis, MN, USA

Bronwen Morgan
The University of New South Wales (UNSW), Sydney, NSW, Australia

Stephan Morgenthaler
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[†] Deceased.

John Myhill
University of Haifa, Haifa, Israel

Roxanne Mykitiuk
Osgoode Hall Law School, York University, Toronto, ON, Canada

Nancy Nason-Clark
University of New Brunswick, Fredericton, NB, Canada

Allison E. Nickel
University of Wisconsin, Milwaukee, WI, USA

Jürgen Oelkers
University of Zurich, Zurich, Switzerland

Marina A.L. Oshana
University of California, Davis, CA, USA

Cenk Özbay
Bogazici Universitesi, Bebek, Istanbul, Turkey

Kenneth I. Pargament
Institute for Spirituality and Health at the Texas Medical Center, Houston, TX, USA; and Bowling Green State University, Bowling Green, OH, USA

David J. Parkin
All Souls College, Oxford, UK

Jerry Patchell
Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Keith Pavitt[†]
University of Sussex, Brighton, UK

Lisa D. Pearce
University of North Carolina, Chapel Hill, NC, USA

Richard Peiser
Harvard University, Cambridge, MA, USA

Joanna Marie Pinto-Coelho
University of Pennsylvania, Philadelphia, PA, USA

Virginie Pirard
Institut Pasteur, Paris, France

Robert A. Prentky
Fairleigh Dickinson University, Teaneck, NJ, USA

Prem Kumar Rajaram
Central European University, Budapest, Hungary

Diederick Raven
Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, The Netherlands

Ian Reader
Lancaster University, Lancaster, UK

Astrid Reuter
University of Munich, Munich, Germany

Ulla Richardson
University of Jyväskylä, Jyväskylä, Finland

David Rindskopf
CUNY Graduate Center, New York, NY, USA

Jonathan Roberge
Institut National de la Recherche Scientifique, Québec, QC, Canada

Elizabeth Roberts
University of Michigan, Ann Arbor, MI, USA

Richard H. Roberts
University of Stirling, Stirling, UK

Bernd Roeck
University of Zurich, Zurich, Switzerland

Henry L. Roediger
Washington University, Saint Louis, MO, USA

David Rogosa
Stanford University, Stanford, CA, USA

Nancy L. Rose
Massachusetts Institute of Technology, Cambridge, MA, USA

Sergio Della Sala
Human Cognitive Neuroscience, Psychology and Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, UK

Paul A. Samuelson
Massachusetts Institute of Technology, Cambridge, MA, USA

David E.M. Sappington
University of Florida, Gainesville, FL, USA

Samuel J.A. Scaggs
College of Criminology and Criminal Justice, The Florida State University, Tallahassee, FL, USA

Peer Scheepers
Radboud University Nijmegen, Nijmegen, The Netherlands

Stuart Scheingold[†]
University of Washington, Seattle, WA, USA

Mortimer NS Sellers
University System of Maryland, MD, USA; and University of Baltimore Center for International and Comparative Law, MD, USA

[†] Deceased.

Darren E. Sherkat
Southern Illinois University, Carbondale, IL, USA

Hua Shu
Beijing Normal University, Beijing, China

Karen M. Smith
School of Nursing and Midwifery, University of Dundee, Dundee, UK

Susan J. Smith
University of Edinburgh, Edinburgh, UK

Nur I.B. Soonaan
Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Richard Stahler-Sholk
Eastern Michigan University, Ypsilanti, MI, USA

Arthur A. Stein
University of California, Los Angeles, CA, USA

Larry Stein
University of California at Irvine, School of Medicine, Irvine, CA, USA

Christoph Steinebach
School of Applied Psychology, Zürich University of Applied Sciences, Zürich, Switzerland

Susan Sterett
University of Denver, Denver, CO, USA

Michael Stewart
University College London, London, UK

Roger R. Stough
University of Illinois, Urbana, IL, USA; and George Mason University, Fairfax, VA, USA

Peter A. Sy
College of Social Sciences and Philosophy, University of the Philippines, Quezon City, Philippines

Erik Taal
University of Twente, Enschede, The Netherlands

Judith M. Tanur
Stony Brook University, Stony Brook, Montauk, NY, USA

Mark Tewdwr-Jones
Newcastle University, Newcastle-upon-Tyne, Tyne and Wear, UK

Robert Tibshirani
Stanford University, Stanford, CA, USA

William M. Trochim
Cornell University, Ithaca, NY, USA

Bryan S. Turner
The Graduate Center, The City University of New York, New York, NY, USA

Špela Humljan Urh
Germany

Frederic Vandenberghe
Institute for Social and Political Studies, State University of Rio de Janeiro, Brazil

Peter van der Veer
Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

Stelios Virvidakis
University of Athens, Athens, Greece

Christy A. Visher
University of Delaware, Newark, DE, USA

David Voas
University of Essex, Colchester, UK

Andreas von Leupoldt
University of Leuven, Leuven, Belgium

Fraser N. Watts
University of Cambridge, Cambridge, UK

Stephen A. Webb
Glasgow Caledonian University, Glasgow, Scotland, UK

Pamela Wilcox
University of Cincinnati, Cincinnati, OH, USA

Michael Williams
University of Oxford, Oxford, UK

Ryan J. Williams
University of Cambridge, Cambridge, UK

Peter H. Wilson
Southern Cross University, Coffs Harbour, NSW, Australia

Cecilia Wong
School of Environment, Education and Development, the University of Manchester, Manchester, UK

Steve Yearley
University of Edinburgh, Scotland, UK

Elizabeth Yuko
Center for Ethics Education, Fordham University, Bronx, NY, USA

Charles E. Ziegler
University of Louisville, Louisville, KY, USA

Reading Nonalphabetic Scripts

Takeshi Hatta, Kansai University of Welfare Science and Nagoya University, Takatsuki City, Osaka, Japan

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Hatta, volume 19, pp. 12797–12800, © 2001, Elsevier Ltd.

Abstract

The unique characteristics of Japanese language such as Japanese sound system contains few sounds. Japanese vocabulary consists of four word types (Japanese origin, Sino-Japanese, foreign-loan word, and hybrids), and Japanese writing system consists of five kinds of script type (Kanji, Hirakana, Katakana, Roman alphabet, and Arabic numerals). Among five script types, Kanji (logographic symbol) and two types of Kana (phonetic symbol) are most frequently employed in writing Japanese. Cognitive psychological studies, which examined this uniqueness, are reviewed. Especially, cognitive psychological studies which investigated how people read the Japanese text written by five kinds of scripts and how people select one special script type in writing Japanese text are introduced. Findings of neuropsychological studies by brain-damaged people as well as brain-imaging studies by intact people, which examined if Kanji and Kana are processed differentially in reading also are reviewed.

Japanese Language

Linguistically, Japanese does not relate to Chinese, Korean, and English, though it is similar to Korean in syntax, but not in speech sounds and vocabulary. However, Japanese has borrowed many words and characters mainly from Chinese. Almost all indigenous people living in Japan speak Japanese, and ethnic Japanese living outside of Japan also do. There are about 130 million Japanese speakers, and Japanese ranks sixth among the languages of the world after Chinese, English, Russian, Hindi, and Spanish.

The Japanese language employs five vowels and 16 consonants. Compared with other languages such as English, the Japanese sound system contains many sounds. The Japanese sound system uses a unit called 'mora,' which is a short beat, the time to pronounce a short syllable. Mora is the main unit in Japanese language, and also the unit of the Japanese phonetic scripts, hirakana (or hiragana) and katakana. Each Japanese syllabary sign is based upon mora. There are 110 different moras or syllables in Japanese. It is one of the smallest inventories of any major languages. For example, 400 syllables in Chinese, and several thousand in English. The small inventory of Japanese sounds gives rise to many homophones but it is easy to be represented by a simple syllabary.

The Japanese vocabulary consists of four word types, Japanese origin (47.5%), Sino-Japanese (36.7%; long years ago adopted from China), foreign-loan word (9.8%; mostly European adopted in modern times), and hybrids (6.0%; Sino-Japanese or foreign word stems with Japanese endings).

Japanese Writing System

Japanese writing system consists of five kinds of script type, kanji, katakana, Roman alphabet, and Arabic numerals. Among these, the first three script types are employed most frequently in Japanese writings. Japanese text consists these five kinds of scripts, and it is also true that both the traditional vertical (tategaki) from top to bottom style and the yokogaki from left to right style are used depending on the type of written materials. For example, newspaper is printed in takegaki style whereas academic journals are printed in yokogaki style.

Kanji Script and Reading

As the results of waves of Chinese and Korean emigration to Japan during fourth and seventh centuries AD, Chinese-based culture mainly in the form of Confucian classics, medicine, and Buddhism came to Japan. A writing system with Chinese characters was, therefore, the only script until two forms of syllabary, hiragana and katakana, were created out of kanji during the nineth century AD. 'Kan-on' kanji readings were borrowings from Confucian scholarship. Chinese 'Hanzi,' associated with Han Dynasty period, became Japanese 'Kanji,' which remains the backbone of the Japanese writing system. As Japanese Kanji borrowed Hanzi in distinct historical periods, different pronunciations were adopted for these kanji and makes reading complicated in modern writing systems. The feature of having multiple readings for many kanji characters in Japanese is different from those of Chinese and Korean, which have only a single reading.

The Japanese language uses not only Chinese and native readings to the logographic characters, but also several varieties of each type of reading, which makes reading characters complicated. For example, Mandarin Chinese read the simple character for '10' 十 as one sound 'shi' and Korean read it 'sip' (no tone or soundless). Both Languages use only one reading; however Japanese speakers read the character, 'to,' 'toh,' 'so,' 'jyu,' 'jutt,' all without tone (Taylor and Taylor, 1995). The first three sounds are the example of Kun readings (Japanese native readings), while the others are 'On'-readings (Chinese readings). The Kun reading is quite different from the Chinese because the two languages are unrelated. Besides Kun reading, most Kanji have one or more on-readings which are not identical with Chinese sounds. Depending upon when and where the Chinese characters were borrowed, four different

types of on-reading (go-on, kan-on, toh-soh-on, and habitual one) were agreed. Most Japanese kanji are given only one of the four types of on-readings.

The number of Kanji is enormous. Kanji dictionaries contain between 12 000 and 50 000 entries. In 1981, the Japanese Ministry of Education attempted to restrict the number of kanji as 'joyo kanji' to be used in educational system. This contains 1945 Kanji but because almost kanji possess On- and Kun-readings, Japanese people therefore learn a total of 4084 kanji reading during their first 9 years of education. Generally, each kanji consists of the combination of radicals. There are 250 kinds of radicals, each consisting of between 1 and 17 strokes. Most radicals consist of two to nine strokes (Ueda et al., 1993). As the number of Kanji is so huge, this reduces any difficulties in understanding homophones. However, there is a clear asymmetry between reading and writing abilities and it is true that some adult Japanese fail to read even joyo Kanji.

Cognitive psychologists once suggested that the processing route of reading depended upon the type of orthography, that is, kanji has a single route from orthography to semantics, bypassing the contribution of phonological information, while kana travel a phonological route to access semantics. But this is too simple an explanation. Recent reviews of kanji studies suggest that it is reasonable to conclude that, depending upon the task settings, familiarity, frequency, and complexity of kanji, one of the most efficient processing routes of phonology-based or grapheme-based routes (or both routes) is employed access to semantics (Kess and Miyamoto, 1999). However, in the case of high-frequency and high-familiarity kanji, semantic access is achieved only through the grapheme-based route, and it is not specific to Japanese. There are examples in alphabetic system where the cognitive route is also direct from grapheme to semantics.

As shown in Table 1, complex Kanji consists of semantic and phonetic radicals, hen (left hand), tsukuri (right hand), kanmuri (top), ashi (bottom), kamae (enclosure), tare (top-left), and nyoo (left-bottom). Each acts differentially in reading. That is, generally, semantic radicals give clues to meanings whereas phonetic radicals give clues to pronunciation. Research findings also suggest that right-hand radical is more informative than the left.

Many words in Japanese are represented not only by individual kanji but also by a compound that is polysyllabic composite between two and four kanji usually carrying an on-reading. More than 50% words in most Japanese dictionaries are compound Kanji. Research findings also suggest that reading processes of kanji compound words constitute another groupings. In addition to the frequency of the compound word itself, first kanji in a compound is important, because the first kanji defines the range of possible collocations in the cohort of potential kanji compounds.

There is strong evidence that the individual kanji are processed as a recognition unit and kanji compounds themselves are also the recognition units. A case has been that one agraphic patient was able to read the single kanji 手 'hand' and 紙 'paper' but failed to read their coupling as 手紙 'letter,' while another patient had less difficulty in reading a compound word but could not read single kanji which constituted the compound word (Hatta et al., 1998). This double dissociation suggests that single and compound kanji words are stored with different addresses in mental lexicon.

Table 1 Samples of radicals

Radicals	*Examples*
Hen	波 (Wave) 河 (River)
Tsukuri	清 (Clear water) 晴 (Clear weather)
Kanmuri	草 (Grass) 花 (Flower)
Ashi	熊 (Bear) 馬 (Horse)
Kamae	国 (State) 囲 (To enclose)
Tare	庭 (Garden) 店 (Shop)
Nyo	返 (To return) 逃 (To escape)

Kana Scripts and Reading

In addition to kanji, there are two syllabary types of script in Japanese, hiragana and katakana (see Table 2). The kana unit represents a sound unit, which corresponds roughly to a syllable. Both scripts contain 46 basic symbols and 25 additional symbols with diacritic marks. Hiragana and katakana share the same syllabic reference point; therefore the same syllable can be written by either system. Hiragana is cursive in shape as shown in Table 2 and used mainly for some content words, morphological endings, function words, and the rest of the grammatical scaffoldings of sentences. In contrast, katakana is angular in shape and is used for writing foreign-loan words.

Kana gradually developed out of Chinese characters in the following order: kanji, kanji as phonetic signs, simplified kanji shapes, and kana. During this development, mainly female authors used hiragana to write letters, poems, diaries, and stories; this is because hiragana is a graceful cursive form of its original kanji. For a long time, it has been the convention that a cursively simplified hiragana is used for grammatical morphemes, and angular simplified katakana is used for foreign-loan words and onomatopoeia. However, an explanation for the use of both types of scripts because of different role in syntax cannot be the only reason why Japanese have kept three kinds of script usage irrespective of high cost. Japanese people might have found it useful or convenient to express semantic emotional information by employing different types of script.

Conventionally, there is a connection between orthography type and word used. This has a strong effect on Japanese word readings. Japanese people possess a strong semantic emotional image for each type of script such as kanji is hard, difficult, intellectual, vigorous, old, masculine, formal; hiragana is soft, round, tender, simple, childish, feminine, lovely; and katakana is angular, hard, cold, simple, new, sharp, exotic (Iwahara et al., 2003). In reading text, Japanese must activate different mental representations depending on types of scripts. That is, Japanese

Table 2 Hiragana, Katakana, and Kanji description of the word, firefly dream

Hiragana	ほたる	ゆめ
Katakana	ホタル	ユメ
Kanji	蛍	夢

may employ shallow (morphological) cues to generate mental representations from text.

Japanese people sometimes employ this emotional image of a script when writing, irrespective of the conventional usage of scripts and words. For example, reading experiments suggested that when Japanese see kanji word of chair 椅子, an image of an old fashioned strongly built chair is generated; when they see katakana word for chair イス, they imagine a modern and elegant chair; and in the case of hiragana for chair いす, they may imagine a simple wooden chair (Ukita et al., 1996). Writing experiments also indicated Japanese write text with due regard to the compatibility between mental representations and type of scripts (Iwahara et al., 2003).

Research Findings on Reading Studies in Cognitive and Neuropsychological Studies

The results of earlier cognitive psychological experiments suggest that orthographic configurations can be used to access directly phonological form, meaning, or both, depending on the processing task requirement. Although the role and the importance of these factors may differ for each orthography in respect to the processing strategy favored and the processing times required, Japanese seems not to be an exceptional writing system, but possesses idiosyncratic requirements of the world's writing systems.

However, neuropsychological studies have suggested that kanji and kana are processed differentially and stressed the importance of orthography difference in reading (Iwata, 1984; Yamadori, 1975). In alexia with agraphia due to angular gyrus lesions, kana reading is impaired more severely than kanji, whereas left posterior–inferior temporal lesions impair kanji reading (Iwata, 1984; Kawamura et al., 1987). These findings support the hypothesis that there are two pathways of reading in Japanese: one is the ventral pathway from the visual area in the occipital cortex via the posterior part of inferior and middle temporal gyli of the left hemisphere, which is used in semantic reading of kanji; the other is the dorsal route from the visual area via the angular gyrus to Wernicke's area, and is used for the phonological processing of kana reading.

Regarding the visual half field, several studies revealed that the right hemisphere contributes more than the left in kanji recognition. For example, a tendency toward a right hemisphere advantage was reported in the case of single kanji (Hatta, 1977), and in the case of two kanji nonword pairs (Sasanuma et al., 1977). On the other hand, a left hemisphere advantage was reported in the case of two kanji compound words (Hatta, 1978). Regarding to kana recognition, the left hemisphere contributes more than the right both in cases of single and word as well as in case of English (Hatta, 1978). By employing a new research method (MEG), Koyama et al. (1998) revealed that not only the left but also the right hemisphere is activated in reading kanji words and supports the view of the right hemisphere engagement in kanji reading. Same proposal was reported by the functional MRI method (Nakamura et al., 2005). These studies suggest the fact that hemisphere function in kanji reading is not the same as in the case of the reading of English and kana words.

Research Findings on Writing Studies in Cognitive and Neuropsychological Studies

Research findings for kanji and kana writing systems suggest a dissociation of kanji and kana processing by brain-damaged, brain-imaging, and behavioral research methods.

Just above Broca's area and anterior to the primary motor control area (Exner's area) damaged patients make kana word writing errors though they have not prominent difficulty in single kana writing. On the other hand, a left superior parietal lobule damage generates paragrahia, writing order error, and configuration distortion more in kanji than in kana writing, and a left occipito-temporal area B37 damage shows difficulty in kanji writing but no difficulty in kana writing (Otsuki, 2007). Left parietal and angular gyrus damage makes more errors in kana than in kanji writing, and left medial frontal area and thalamus damage invites difficulty in kanji and kana retrieve and paragraphia (Soma et al., 1989).

Brain imaging studies show a prominent activation in left occipito-temporal area B37 when participants retrieve kanji by semantic strategy (e.g., retrieve kanji that represents color) whereas Exner's area activates when they employ phonetic strategy, for example, retrieve kanji corresponding phoneme (Nakamura and Fukumura, 2007).

Behavioral studies showed that Japanese kanji writing errors consisted of largely three factors, phonological, semantic, and morphological, in spontaneous writing situation. Phonologically related errors are the most prominent, followed by orthographically related errors, and semantically related errors, including some overlap of these three types (Hatta et al., 1998).

Summary for Findings in Cognitive and Neuropsychological Studies

Kanji reading and writing relate to the networks including occipital and temporal lobes, whereas kana relate to the networks in frontal and parietal lobes including left inferior frontal gurus and motor association area (Iwata, 1986). Mechanism of Japanese reading and writing seems more complicate than alphabetic languages, which have only one type of scripts. Further studies concerning nonalphabetical scripts are necessary to answer the question that there is a language nonspecific universal cognitive processing in reading and writing.

See also: Functional Brain Imaging of Language Processes; Linguistic Typology; Neural Bases of Reading; Phonetics, Articulatory; Writing Systems; Writing Systems, Psychology of.

Bibliography

Hatta, T., 1977. Recognition of Japanese kanji in the left and right visual fields. Neuropsychol 15, 685–688.

Hatta, T., 1978. Recognition of Japanese kanji and hiragana in the left and right visual fields. Japanese Psychological Research 20, 51–59.

Hatta, T., Kawakami, A., Tamaoka, K., 1998. Writing errors in Japanese kanji. Reading & Writing 10, 457–470.

Iwahara, et al., 2003. Effects of compatibility feelings between type of scripts and words in writing Japanese. Reading and Writing 16, 377–397.

Iwata, M., 1984. Kanji versus kana: neuropsychological correlates of the Japanese writing system. Trends in Neurosciences 7, 290–293.

Iwata, M., 1986. Neural mechanism of reading and writing in the Japanese language. Functional Neurology 1, 43–52.

Kawamura, M., Hirayama, K., Hasegawa, K., Takahashi, N., Yamaura, A., 1987. Alexia with agraphia of kanji (Japanese morphograms). Journal of Neurology, Neurosurgery, and Psychiatry 50, 1125–1129.

Kess, J., Miyamoto, T., 1999. The Japanese Mental Lexicon. John Benjamins, Amsterdam.

Koyama, S., Kakigi, R., Hoshiyama, M., Kitamura, Y., 1998. Reading of Japanese kanji (morphograms) and kana (syllabograms): a magneto-encephalographic study. Neuropsychol 36, 83–98.

Nakamura, et al., 2005. Hemispheric asymmetry emerges at distinct parts of the occipitotemporal cortex for objects, logograms and phonograms. A functional MRI study, Neuroimage 28, 521–528.

Nakamura, K., Fukumura, H., 2007. A functional brain imaging analysis in Japanese writing. In: Iwata, M., Kawamura, M. (Eds.), Neurogramatography. Igaku-shoin, Tokyo, pp. 201–212.

Otsuki, M., 2007. Neural mechanism in writing. In: Iwata, M., Kawamura, M. (Eds.), Neurogramatography. Igaku-shoin, Tokyo, pp. 179–199.

Sasanuma, S., Itoh, K., Mori, K., Kobayashi, Y., 1977. Tachistoscopic recognition of kana and kanji words. Neuropsychol 15, 547–553.

Soma, Y., Sugishita, M., Kitamura, K., et al., 1989. Lexical agraphia in the Japanese language: pure agraphia for kanji due to left posteroinferior temporal lesions. Brain 112, 1549–1561.

Taylor, I., Taylor, M.M., 1995. Writing and Literacy in Chinese, Korean, and Japanese. John Benjamins, Amsterdam.

Ueda, M., Okada, M., Iijima, T., Sakaeda, T., Iida, D., 1993. Shin Daijiten. Kodan-sya, Tokyo.

Ukita, J., Sugishima, M., Minagawa, M., Inoue, M., Kashu, K., 1996. Psychological Research on Orthographic Forms in Japanese. Psychological Monograph No. 25, Japanese Psychological Association.

Yamadori, A., 1975. Ideogram reading in alexia. Brain 98, 231–238.

Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Heikki Lyytinen, Mikko Aro, and Ulla Richardson, University of Jyväskylä, Jyväskylä, Finland
Jane Erskine, Banff, Aberdeenshire, UK
Hong Li and Hua Shu, Beijing Normal University, Beijing, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reading acquisition is under the influence of the structure and function of the language itself, the individual's capacity to learn and the differential methods of instruction. Alphabetic languages vary from the bidirectionally consistent phoneme–grapheme–phoneme units of Finnish, to English as the most inconsistent for small units, but the latter of which increases in consistency as a function of increasing size of the unit (e.g., rime). Nonalphabetic Chinese places a heavy learning load through its inherently meaning-based logographic structure. Computer-assisted literacy training (e.g., GraphoGame) tailored to language features, teaching practices, and learner specific requirements can provide an adaptive and motivational learning experience.

Introduction

The acquisition of reading skill comprises several steps. The first step of reading is for the child to learn to connect the spoken language with the written language in order to sound and spell out the written language. For the majority, this step is mastered relatively well when formal reading instruction is begun. The next step involves the acquisition of sufficient fluency in reading in order to become fully literate and thus achieve the main goal of reading – that of comprehension. For the most part, this step can be achieved through simple exposure and practice.

There are many factors that affect reading acquisition; some internal, resulting from the readiness factors characteristic to the learner and others external, such as the nature of both the writing system and reading instruction. Below, we examine first the external factors and then those that are internal. In closing, we offer advice with regard to helping those children for whom reading acquisition is a challenge.

External Influences on Reading Acquisition

Reading Acquisition Is Affected by the Writing System

Learning to read starts with the acquisition of the central connections between language that is spoken (phonology) and language that is written (orthography). Such connection building can vary on two dimensions: in what is connected to writing – sounds or meaning – and what is the appropriate size of the to-be-learned unit in order to form a sufficiently reliable connection between spoken and written language. Independent of the writing system – either alphabetic (e.g., English or Finnish) or nonalphabetic (e.g., Chinese) – the typical beginnings to reading involve forming the connection between the spoken and written form.

Alphabetic writing systems range from consistent (or 'transparent' or 'shallow') to very inconsistent (or 'opaque' or 'deep') in terms of the relationship between the spoken and written unit. Alphabetic orthographies vary in their consistency at the level of the smallest functional spoken unit (i.e., the phoneme) and its relationship to its smallest counterpart in the written form – the corresponding letter(s) or grapheme. In nonalphabetic Chinese, writing does not reflect the segmental structure of speech and Chinese characters map to a whole morpheme (the smallest unit of meaning) syllable. Chinese children then have to learn thousands of lexical representations (Chinese characters) that include phonological, orthographical, and semantic (meaning) information.

Alphabetic Writing Systems

Transparent alphabetic writing systems can be so consistent that each phoneme is represented in writing by only one grapheme, independent of the context whereby only phonemic information is marked in the script. In more opaque orthographies, often morphological information is also coded – usually at the expense of phonemic regularity. As an example, in English the past tense ending of regular verbs is always written with '-ed', despite three possible pronunciations (/t/ in 'washed', /id/ in 'shouted', /d/ in 'roared'). Orthographic transparency is also related to historical changes in the development of a language. There is a general tendency that more recent orthographies, such as African local languages, are more transparent. Older orthographies – such as English – have often seen a number of spoken language-related changes that can cause written language to diverge from spoken. English is generally considered the most opaque of major alphabetic orthographies. This status is problematic since the majority of research knowledge regarding reading acquisition and acquisition-related problems stems from observations of readers of English.

Nonalphabetic Writing Systems

The Chinese orthography is often described as both logographic (visual symbols that represent words) and morphosyllabic. The basic unit of interface between the written word

and spoken language in Chinese is the morpheme. A character corresponds with one syllable and usually represents one morpheme. The most prominent characteristic of Chinese that differs from alphabetic languages lies in its complicated orthography and the relationships among visual symbol, speech sound, and semantic meaning. The Chinese writing system is also well known for its visual complexity. The units of orthographic structure in Chinese include strokes (丶, 丨, 丿), radicals (or root forms) (犭, 苗), or recurrent patterns of strokes and characters (猫, /mao1/, cat). The basic unit of Chinese, the character, is square shaped as opposed to the linear arrangement of English and other alphabetic scripts. Each character is a salient perceptual unit that differs from thousands of others in terms of individual strokes or the overall spatial configuration. About 72% of the characters that Chinese children learn during the 6 years of elementary school are compound and consist of radicals or components under the orthographic rules (Shu et al., 2003). There is some phonology information available in the semantic–phonetic compound characters (e.g., 烤, /kao3/, 'to burn'), which consist of a phonetic radical (考, /kao3/, exam) that provides cues to character pronunciation and a semantic radical (火, /huo3/, 'fire') that provides cues to character meaning. However, according to Shu et al.'s (2003) analysis, in all of the semantic–phonetic compound characters, about 29% are regular, in which these characters have exactly the same pronunciation as their phonetic radicals (e.g., 清/qing1/, with the phonetic, 青/qing1/) and about 40% are transparent, in which the radical provides reliable information about the meaning (e.g., 氵means 'water' and 清 means 'clear'). Chinese has a relatively simple syllable structure: a syllable consists of an onset and a rime and the lexical tone. The Chinese character directly maps onto morphemes rather than phonemes. Print-sound mapping in Chinese happens at the morpheme–syllable level, not at the grapheme–phoneme level as in English and other alphabetic scripts. There are approximately 7000 morphemes but only 1200 syllables in Mandarin. Thus, more than five morphemes, or characters, share one syllable. Homophones (same sounds but different meaning), e.g., syllable /yi4/ may correspond to 义 '*meaning*', or 易 '*easy*', 亿 '*a 100 million*', 宜 '*appropriate*', 益 '*benefit*', 艺 '*art*', 议 '*discuss*', and so on and homographs (same written form but different meaning), e.g., 草, may represent 'grass' in 草地 '*lawn*' or '*careless*' in 草率 '*cursory*', are very prevalent in Chinese, which weaken the differentiation of phonology. The connection from orthography to meaning is more systematic and efficient than that from orthography to phonology for Chinese character reading, while the situation is the opposite for English word reading (Yang et al., 2006).

Typical Acquisition of Reading and Writing

Alphabetic Writing Systems

In alphabetic writing systems, the majority of children can learn to read using phonics-based instruction. However, the more transparent the orthography, the faster the process of acquisition, as shown in **Figure 1**.

In contrast, the situation is completely different in nonalphabetic writing where the number of written language items (e.g., characters in Chinese) can be much larger than the number of letters in alphabetic writing (typically, not more than 30). To illustrate, the number of characters that require to be mastered by learners of Chinese over a number of years proceeds as shown in **Table 1**. The challenge faced by learners in these two different types of writing environment varies immensely; not least in terms of the amount of time required to acquire the basics of reading – from a very short time (i.e., a few months) for readers of fully transparent writing environments such as Finnish, to more than 6 years for readers of Chinese.

With speech typically considered as the 'primary' language (Mattingly, 1972), the child's awareness of speech sounds is viewed as implicit and consequently is not readily available for

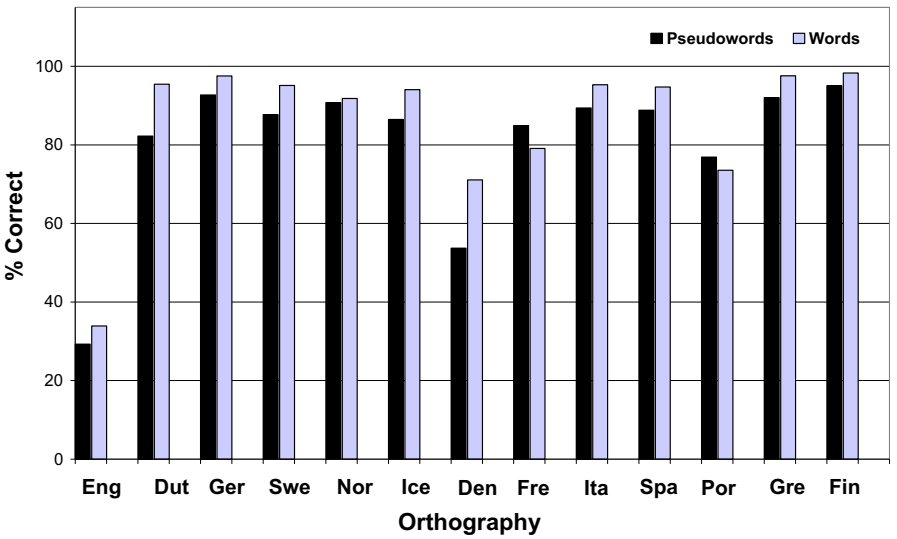

Figure 1 Accuracy of word and pseudoword reading at the end of the first grade in 13 orthographies according to Seymour et al. (2003). Replicated from Handbook of Language Development, Blackwell (Lyytinen, H., Erskine, J., Tolvanen, A., et al., 2006. Trajectories of reading development: a follow-up from birth to school age of children with and without risk for dyslexia. Merrill-Palmer Quarterly 52 (3), 514–546).

Table 1 Number and proportion of characters by grade

Grade	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
Number of characters	436	709	541	358	323	203	2570
Proportion of total	0.17	0.28	0.21	0.14	0.13	0.08	1.00

Table 1 is replicated from Shu, H., Chen, X., Anderson, R.C., et al., January/February 2003. Properties of school Chinese: implications for learning to read. Child Development 74 (1), 31.

intentional manipulation in the learner's mind. The shift to 'secondary' language, in the written form, requires the trigger that makes explicit the awareness of the constituent sounds that make up a given language (Gombert, 1992; Duncan et al., 1997). This transition can present as a challenge because, prior to formal instruction, children use spoken stimuli only to mediate meaningful messages to others and pay little heed to the individual sounds per se, except in language games. However, the step toward explicitly switching attention to the differentiation of small meaningless sound segments from each other is not large, especially when the child is assisted by a consistent alphabetic orthography (e.g., German, Spanish, Italian, Finnish, etc.). Within such contexts, in the reading direction, each letter or grapheme will, at most, cue only the smallest corresponding unit of the spoken language (phoneme). In the opposite direction that is spelling, each phoneme or sound unit will have a direct connection to a single letter when the consistency is high (with the exception of, e.g., German in this direction; see Landerl and Wimmer, 2008).

In the above-described alphabetic writing environments, the greatest challenge posed to the early reader relates to the perceptual differentiation of the sound units that she or he must learn in order to connect these to their corresponding letter(s). In the typical learner, such connections can be instructed effectively by drilling the connections (e.g., presenting a spoken phoneme and asking the child to choose the representative letter). For most children, the acquisition of the names of letters will make them sufficiently aware of phonemes to be able to piece together the constituent sounds in order to produce a pronounceable word sequence and without further reinforcement from intended meaning of the word. This claim is supported by the fact that a substantial number of Finnish children have managed to acquire the basics of grapheme–phoneme connections – the basis of reading – in their fully transparent orthography, prior to any formal instruction beginning at the age seven.

It is important to note that the level of consistent correspondence between spoken and written units of the languages varies along a continuum of 'orthographic depth' from shallow (e.g., Finnish) to deep (e.g., English, where transparency is minimal). The more transparent the orthography on the letter-sound level, the lesser is the learning burden and consequently, the easier it is to learn to sound out the letter sequences.

The development of reading therefore depends on differences in orthographic depth. Cross-language research has shown that learning to read an irregular orthography is harder and takes longer than learning to read a regular orthography (e.g., Frith et al., 1998; Goswami et al., 2001). Seymour et al. (2003) compared reading acquisition in 13 European countries and showed that average reading accuracy in most transparent orthographies reached ceiling at the end of grade 1, but only 34% in English. English-speaking children were estimated to need 2 or more years of experience to achieve the accuracy and fluency levels of the children in shallow orthographies, such as Finnish, on word identification.

Nonalphabetic Writing Systems

Although the mapping in Chinese from spelling to sound is syllable based, numerous studies of Chinese children's reading development have demonstrated that phonological skills, including syllable awareness, onset awareness, and rime awareness, are associated with Chinese character recognition (e.g., Chow et al., 2005; Siok and Fletcher, 2001). Lexical tone is a fundamental feature of Chinese spoken language in which four tones are used to distinguish meanings that are not differentiated by segmental information. Different levels of phonological awareness in Chinese emerge as the result of age or experience (Shu et al., 2008). Recent studies have also suggested Rapid Automatized Naming (RAN) to be a consistent predictor of Chinese reading development, in which linking printed information with a given phonological representation arbitrarily is important. RAN predicts reading development among Chinese children (Ho and Lai, 1999; Lei et al., 2011; Pan et al., 2011).

The acquisition of reading skill in Chinese requires learning to connect not only sounds, as is the beginning phase of Pinyin (Pinyin is a phonological coding system that represents Chinese characters using Roman alphabetic letters and lexical tone transcriptions. It represents single phonemes as in alphabetic scripts, but it is taught in a syllabic way, divided into onset, rime, and four lexical tones, e.g., the character 青 is transcribed as /qing1/ in Pinyin, where /q/ is the onset, /-ing/ is the rime, and /1/ is the lexical tone representing the pitch of speech.), but also meanings to the written symbols. In comparison to alphabetic languages, Chinese presents a huge challenge to learning to read. The Chinese learner must spend more than 5 years in learning more than 3000 simplified Chinese characters before they are equipped to read a newspaper. However, compared to learning to read English, learning to read Chinese is much more of a challenge. Learning to read Chinese depends on memory and integration of information about orthography, phonology, and semantics of thousands of characters in the early stages of reading. Thus, school children learning to read Chinese require a longer period in reading acquisition. Consequently, Chinese developmental dyslexia is usually diagnosed after forth grade when their classmates have learned to read (e.g., Shu et al., 2006).

We have partially unlocked the secrets of reading skill above in that the learner must first acquire the connections between the spoken and written units. The learning load varies from light (in transparent Finnish) to heavy (in nontransparent English), to cumbersome characters (in logographic Chinese). The fewer than 30 connections of transparent orthographies contrasts with the greater burden of learning the hundreds of potential correspondences in English, many of which can thwart pronunciation until all the letters are on view (e.g., 'thought'). In turn, the concept of transparency takes a back seat

in comparison to the learning load that is placed on the shoulders of the young reader of Chinese.

Larger, syllable- or word-level units that prevent a transparent match between the smaller constituent spoken and written units can place a significant load on memory as the current conventional route to their learning is through repetition or rote learning as each must be learned almost individually. Such a writing system could be highly detrimental in the context of agglutinative languages, such as Finnish whose words can have hundreds of different orthographic forms due to inflectional endings. Initial learning, based on whole word recognition, would, in such a situation, be a huge undertaking. Instead, in fully transparent writing systems, reading words or any sequence of letters forming a pronounceable item (i.e., pseudowords), is no different. Any written item can be sounded out accurately, simply by knowing the sounds of each of the letters and phonemic blending.

Although there is systematicity inherent in Chinese character orthography, traditional instruction in the classroom does not take full advantage of this. Teachers rely heavily on rote memorization, generally accomplished for reading by having children memorize characters as a paired associate learning task and for writing, by having children copy characters until they have internalized their production as a motor skill. Teachers typically demonstrate how each character is constructed, stroke by stroke on the blackboard, with students performing a repetitive copying task on worksheets at their desks. Teachers also often demonstrate how the characters are written by tracing the individual strokes with hand movements in the air, with students imitating the teacher's movements in the air at their seats. In addition, in elementary schools in Mainland China, children learn Pinyin in the first 8 weeks of school and then learn characters associated with Pinyin. During the first 2 years, all characters in language arts textbooks are annotated with Pinyin. After second grade, Pinyin only appears above new characters as a tool for decoding unfamiliar characters (Wu et al., 1999).

Internal Influences on Reading Acquisition

Developmental dyslexia is an intrinsic significant difficulty in developing reading skills in spite of adequate learning opportunities, good motivation, a normal or high level of IQ, and normal sensory abilities. Depending on the criteria, it is estimated to affect 6–17% of school age populations regardless of the language/orthography or culture (Fletcher et al., 2007: p. 105). Dyslexia is more common in males, with a gender ratio of 1.5:1 (Pennington, 2009: p. 45). It has been shown that dyslexia is neurobiological in origin, and there is strong evidence of its being heritable. Historically, the core cognitive deficits of dyslexia have been sought in many cognitive domains. However, at present there is a general consensus that dyslexia is caused by problems in linguistic skills, especially phonological processing. Individuals with dyslexia seem to have specific difficulties in processing spoken language. There may be difficulties in the ability to segment and manipulate spoken words, poor verbal short-term memory, and problems with lexical access, i.e., slow to remember/retrieve names of items. Thus, it seems that phonological representations play a part in dyslexia but their quality and/or accessibility is not well understood, even after several decades of research in dyslexia.

Dyslexia in Alphabetic Languages

A research-based definition of dyslexia states that dyslexia is characterized by difficulties in accurate and fluent word recognition, as well as poor spelling and decoding abilities (Lyon et al., 2003). Thus, the core problems of dyslexia are considered to be related to phonological processing and word-level reading, and the often observed problems in reading comprehension are secondary to the basic problem.

The research on dyslexia has a strong anglocentric history. Most of the models and theories of dyslexia are based on studies concerning English-speaking populations. Because recent cross-linguistic research on reading acquisition has shown marked differences in literacy development between alphabetic orthographies, attention has turned its focus to questions with regard to universal vs language-specific aspects of dyslexia. In a recent review of cross-linguistic research on dyslexia, Share (2008) argues that traditional emphasis on English has biased our understanding of both normal literacy development and dyslexia. The studies on neurobiological correlates of dyslexia seem to show that the neurobiological origins of dyslexia are similar across languages (Paulesu et al., 2001). Similarly, it seems that phonological processing skills predict later reading development and dyslexia across the continuum of orthographic transparency. However, the phenotype of dyslexia seems to vary depending on transparency of the orthography. It has been shown that poor accuracy as a core problem in dyslexia is, by large, specific to opaque orthographies, most notably English (Landerl et al., 1997). Instead of poor accuracy, a deficit in reading speed seems to be the most consistently observed problem in dyslexia across languages, also after the initial stages of development (Ziegler et al., 2003). In addition to being more universal, this fluency deficit seems to be also more persistent developmentally (Landerl and Wimmer, 2008). Of practical importance is the fact that among older children, fluency problems do not seem to respond well to training of phonological processing skills (Torgesen, 2005). Developing efficient means to support reading fluency, and to alleviate the reading speed deficit characteristic of dyslexia still remains a challenge for future research.

Research findings indicate that deficits in phonological processing, especially poor phonemic awareness and problems in lexical access (as reflected in problems of fluent and effortless naming), seem to be related to dyslexia in alphabetic orthographies across the continuum of transparency. However, there is research evidence to show that these linguistic skills are related to different aspects of reading development, and consequently, their role as a predictor of reading problems might vary between alphabetic orthographies. Poor phonemic awareness is supposedly more closely related to accuracy problems, and problems in lexical access – naming speed – to rate (see, e.g., Wolf and Bowers, 1999). Longitudinal studies in transparent orthographies support this notion; in German-speaking (Landerl and Wimmer, 2008) and Finnish-speaking

samples (Torppa et al., 2012), an early naming speed deficit predicted reading fluency problems, whereas phonemic awareness was predictive of inaccurate spelling. These studies also support the notion that problems in phonemic awareness and naming speed are partly independent deficits, as suggested by Wolf and Bowers (1999). However, it is worth pointing out that there are as yet no conclusive answers concerning the relation between phonemic awareness and naming speed. Although deficits in both phonemic awareness and naming speed have been shown to affect reading development in alphabetic orthographies in general, their relative relevance to orthographies that vary in transparency is still an open question (see, e.g., Ziegler et al., 2010).

Developmental Factors and Courses That Increase Risk of Compromised Learning due to Biological Reasons

The Jyväskylä Longitudinal study of Dyslexia (JLD; for review, see Lyytinen et al., 2008) examined in detail the developmental factors and courses that increase the risk of encountering difficulties during reading acquisition. Data are based on children with familial background (one parent affected) of dyslexia. Genetic factors explain the starting point (the first candidate gene of dyslexia was detected partly on the basis of the JLD data; Taipale et al., 2003). Half of those at risk who later encountered dyslexic difficulties were identified on the basis of developmental factors (delayed expressive and receptive speech) at age 1–2 years. Later indices include delayed development of phonological sensitivity and, in many cases, very slow naming speed (see Lyytinen et al., 2006; Puolakanaho et al., 2007). However, for practical, preventive purposes, it is sufficient to follow the facility to store letter names. Children who require preferably preventive intervention have difficulty in storing letter names (Lyytinen et al., 2009).

Learning Difficulties in Nonalphabetic Languages

Research has demonstrated that Chinese dyslexic children predominantly experience difficulties with accuracy and speed of word reading and spelling (Shu et al., 2006). Previous studies have demonstrated that phonological processing skills, including phonological awareness and RAN are involved in reading Chinese characters (e.g., Siok and Fletcher, 2001; Perfetti et al., 1992). The conclusion was consistent with that of alphabetic studies. However, numerous studies on Chinese children's reading development and impairment have demonstrated that visual skill and orthographic awareness (e.g., Huang and Hanley, 1995; Ho et al., 2004; Li et al., 2012; Zhang, et al., 2012a), speech perception, especially categorical perception of Mandarin Chinese tones (Zhang, et al., 2012b) and morphological awareness, including sensitivity to homophones/homographs and lexical compounding, and semantic radical awareness (e.g., Chen et al., 2009; Shu et al., 2006), as well as visual–verbal paired associate learning ability (e.g., Li et al., 2009), play significant roles in Chinese reading development.

Dissonance between Learner's Spoken Language and the Language She or He Is Learning to Read

For the most part, children's first spoken language is the same language that is being instructed in written form in school. If this is not the case – as has happened and is still happening in many countries, e.g., in Africa – whereby the learner speaks only the local language(s), but is instructed first to read English or French, the outcomes may not be good. This has been demonstrated, e.g., by the situation in Zambia. Although the formal goal of their school system is to instruct reading of the local language, the results are very poor. Largely, this is due to the fact that reading instruction has followed, from the beginning, the contents of English (Ojanen et al., 2013). A concrete example that dooms acquisition to failure is that the first step of instruction comprises recitation of English letter names when the appropriate letter names (which provide cues to the sounds) in the fully transparent Zambian are very different. A more general reason for compromised learning in multilingual environments can result from a mismatch between the language spoken in the child's natural environment and later attempts to adapt to the written language of school instruction. When a child learns to speak, she or he will become sensitized to perceive the sounds of the language spoken in the community.

Sometimes, a large variation in learning burden is encountered within one country. Spanish has a relatively consistent writing system in comparison to English. Consequently, the potential is there to foster initial instruction in the reading and writing of the native spoken language (Spanish) first, for example in the USA, with focus on the mechanics of learning to corresponding sounds and letters, and as a precursor to beginning formal instruction in reading a more complex second language, English.

The Way Forward: Computer-Assisted Learning for All Writing Environments and Developmental Prospects

In order to provide efficient support for learning to read for different types of learners in different language contexts, the above-described external and internal factors that affect learning skills need to be taken into account. In addition to traditional teacher led learning situations, technology-enhanced learning environments might provide the necessary extra support for learners. Such environments should be based on theories of reading acquisition for typical and atypical learners, relevant research findings, and, above all, provide scientific evidence as to the efficacy of such support. By utilizing technology, the basic processes needed for developing technical reading skills, learners can practice as many times as they need, the connection of good examples of spoken language items (from individual speech sounds to larger units, such as syllables, words, and sentences) to the corresponding written items. Technology has the advantage that it can provide a learning environment where learning is engaging enough (e.g., in a digital game format) and encouraging (by presenting immediate positive feedback when appropriate) for learners to practice the essential skills repeatedly – often one of the crucial factors for gaining good learning results. Also, the advantage of technology is that, if carefully designed and implemented,

individual learners can be provided with just the type of training that is appropriate at the precise stage of their skill level. This is possible, e.g., by basing the adaptation system of the software on mathematical/statistical estimations and linguistic/educational/psychological knowledge on the number and type of learning material (e.g., psycholinguistic grain sizes) that is optimal for each learner. Thus, multidisciplinary approaches to the development of such technologies are needed. One such evidence-based technological support is GraphoGame (grapholearn.info), which has been developed in Finland at the University of Jyväskylä. This technological tool is based on the main findings of 20 years of longitudinal research on dyslexia in the JLD (e.g., Lyytinen et al., 2008). By incorporating appropriate spoken language units to the corresponding written language units in a systematic manner and taking into account the nature of the orthography, such use of technology for training reading skills has been shown to have significant positive effect on learning results (e.g., Lyytinen et al., 2007; Saine et al., 2011). Encouraged by the positive findings, GraphoGame technology is currently being developed and tested in different languages using the above mentioned principles, and the results so far for developing functional reading skills have been very encouraging for alphabetic languages (Kyle et al., 2013; Ojanen et al., 2013). Using neuroimaging techniques, Brem et al. (2010) have demonstrated the neural basis of the mapping between sound and print in the very early stages of reading acquisition. For orthographically complex English, Kyle et al. (2013) showed the superior effects on learning of the more consistent rime-level units in comparison to lesser consistent phonemes.

In the alphabetic language contexts, the fluent and effortless mediation of meaning from written material may be the most challenging skill to develop for those whose learning abilities are affected by the internal factors. The challenge for GraphoGame and other technology-enhanced learning environments is to tackle the central features of nonalphabetic orthographies and establish an efficient way to provide support for building connections between the thousands of written symbols and the corresponding meanings. However, with the joint efforts of expert content providers and software developers, increasingly more affordable technological products may turn out to be a time- and place-independent and economical means to provide appropriate and targeted learning opportunities on a global level, especially to those who face challenges due to either biological (internal) or environmental (external) factors.

Conclusion

Reading is often viewed as the means of progression in modern society. The process of reading acquisition, however, is contingent on the structure and function of the language of instruction. A young reader in one country may have a much lighter workload than their peer in another. This load varies according to the number of units, their size, and the potential for ambiguity in the relationship between what is spoken and what is written. Thus, the child learning a very transparent alphabetic language can quickly latch on to the straightforward relationship between the small sound units and convert these to their consistently matching small units within written words. In contrast, a child's attempt to deploy such a simple strategy in the learning of a more opaque language, is likely to be thwarted in part by inconsistencies within these smaller-size units and thus she or he becomes more reliant on simultaneous use of more consistent larger units. All these, however, tend to pale into insignificance relative to the task of the learner of a logographic language, the acquisition of which entails larger units, in greater numbers and thus requires much longer to achieve. The burden of learning is greater still for those children who encounter difficulty with literacy acquisition – due either to inherent learning disability or externally influenced acquisition of nonnative language.

How then to support today's young reader, irrespective of the structure, function, or difficulty with the language to be learned? Technology-enhanced learning, such as the GraphoGame learning environment can assist in both first and second language learning by fostering the positive triggers that promote learning – a synthesis between the critical linguistic components specific to each language and the individual requirements of the learner – coupled with the enjoyment of learning, motivational sustenance, and positive reinforcement.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Computer Assisted Instruction; Computer Games and Academic Achievement; Diagnostic Processes in Clinical Psychology and Psychotherapy; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Morphology in Linguistics; Neural Bases of Reading; Phonology; Second Language Acquisition; Vocabulary Acquisition.

Bibliography

Brem, S., Bach, S., Kucian, K., et al., 2010. Brain sensitivity to print emerges when children learn letter-speech sound correspondences. Proceedings of the National Academy of Sciences 107 (17), 7939–7944.

Chen, X., Hao, M., Geva, E., et al., 2009. The role of compound awareness in Chinese children's vocabulary acquisition and character reading. Reading and Writing 22, 615–631.

Chow, Y.W., McBride-Chang, C., Burgess, S., 2005. Phonological processing skills and early reading abilities in Hong Kong Chinese kindergarteners learning to read English as a L2. Journal of Educational Psychology 97 (1), 81–87.

Duncan, L.G., Seymour, P.H.K., Hill, S., 1997. How important are rhyme and analogy in beginning reading? Cognition 63, 171–208.

Fletcher, J.M., Lyon, G.R., Fuchs, L.S., et al., 2007. Learning Disabilities: From Identification to Intervention. Guilford, New York.

Frith, U., Wimmer, H., Landerl, K., 1998. Learning to read and phonological recoding in English and German. Scientific Studies of Reading 2 (1), 31–54.

Gombert, J.E., 1992. Metalinguistic Development. Harvester Wheatsheaf, Chicago, IL.

Goswami, U., Ziegler, J., Dalton, L., et al., 2001. Pseudohomophone effects and phonological recoding procedures in reading development in English and German. Journal of Memory and Language 45, 648–664.

Ho, C.S.-H., Chan, D.W., Lee, S.-H., et al., 2004. Cognitive profiling and preliminary subtyping in Chinese developmental dyslexia. Cognition 91, 43–75.

Ho, C.S.-H., Lai, D.N.-C., 1999. Naming-speed deficits and phonological memory deficits in Chinese developmental dyslexia. Learning and Individual Differences 11, 173–186.

Huang, H.S., Hanley, J.R., 1995. Phonological awareness and visual skills in learning to read Chinese and English. Cognition 54, 73–98.

Kyle, F.E., Kujala, J., Richardson, U., et al., 2013. Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the United Kingdom: GG Rime and GG Phoneme. Reading Research Quarterly 48 (1), 61–76.

Landerl, K., Wimmer, H., 2008. Development of word reading fluency and spelling in a consistent orthography: an 8-year follow-up. Journal of Educational Psychology 100, 150–161.

Landerl, K., Wimmer, H., Frith, U., 1997. The impact of orthographic consistency on dyslexia: a German–English comparison. Cognition 63, 315–334.

Lei, L., Pan, J., Liu, H., et al., 2011. Developmental trajectories of reading development and impairment from ages 3 to 8 years in Chinese children. Journal of Child Psychology and Psychiatry 52 (2), 212–220.

Li, H., Shu, H., McBride-Chang, C., et al., 2009. Paired associate learning in Chinese children with dyslexia. Journal of Experimental Child Psychology 103, 135–151.

Li, H., Shu, H., McBride-Chang, C., et al., 2012. Chinese children's character recognition: visuo-orthographic, phonological processing, and morphological skills. Journal of Research in Reading 35 (3), 287–307.

Lyon, G.R., Shaywitz, S.E., Shaywitz, B.A., 2003. A definition of dyslexia. Annals of Dyslexia 53, 1–14.

Lyytinen, H., Erskine, J., Tolvanen, A., et al., 2006. Trajectories of reading development: a follow-up from birth to school age of children with and without risk for dyslexia. Merrill-Palmer Quarterly 52 (3), 514–546.

Lyytinen, H., Ronimus, M., Alanko, A., et al., 2007. Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. Nordic Psychology 59 (2), 109–126.

Lyytinen, H., Erskine, J., Ahonen, T., et al., 2008. Early identification and prevention of dyslexia: results from a prospective follow-up study of children at familial risk for dyslexia. In: Reid, G., Fawcett, A., Manis, F., Siegel, L. (Eds.), The SAGE Handbook of Dyslexia. SAGE Publishers, London, pp. 121–146.

Lyytinen, H., Erskine, J., Kujala, J., et al., 2009. In search of a science-based application: a learning tool for reading acquisition. Scandinavian Journal of Psychology 50, 668–675.

Mattingly, I.G., 1972. Reading, the linguistic process and linguistic awareness. In: Kavanagh, J.F., Mattingly, I.G. (Eds.), Language by ear and by eye: the relationship between speech and reading. MIT Press, Cambridge, MA, pp. 134–147.

Ojanen, E., Kujala, J., Richardson, U., Lyytinen, H., 2013. Technology-Enhanced Literacy Learning in Zambia: Observations from a Multilingual Literacy Environment. Insights on Learning Disabilities: from Prevailing Theories to Validated Practices 10 (2), 103–127.

Pan, J., McBride-Chang, C., Shu, H., et al., 2011. What's in the naming? A 5-year longitudinal study of early rapid naming and phonological sensitivity in relation to subsequent reading skills in both native Chinese and English as a second language. Journal of Educational Psychology 103 (4), 897–908.

Paulesu, E., Démonet, J.F., Fazio, F., et al., 2001. Dyslexia: cultural diversity and biological unity. Science 291, 2165–2167.

Pennington, B.F., 2009. Diagnosing Learning Disorders: A Neuropsychological Framework. Guilford Press, New York.

Perfetti, C.A., Zhang, S., SandBerent, I., 1992. Reading in English and Chinese: evidence for a "universal" phonological principle. In: Frost, R., Katz, L. (Eds.), Orthography, Phonology, Morphology, and Meaning. North-Holland, Amsterdam, pp. 227–248.

Puolakanaho, A., Ahonen, T., Aro, M., et al., 2007. Very early phonological and language skills: estimating individual risk of reading disability. Journal of Child Psychology and Psychiatry 48, 923–931.

Saine, N.L., Lerkkanen, M.-K., Ahonen, T., Tolvanen, A., Lyytinen, H., 2011. Computer-assisted remedial reading intervention for school beginners at-risk for reading disability. Child Development 82, 1013–1028.

Seymour, P.H.K., Aro, M., Erskine, J.M., 2003. Foundation literacy acquisition in European orthographies. British Journal of Psychology 94, 143–174.

Share, D.L., 2008. On the anglocentricities of current reading research and practice: the perils of overreliance on an "outlier" orthography. Psychological Bulletin 134, 584–615.

Shu, H., Chen, X., Anderson, R.C., et al., 2003. Properties of school Chinese: implications for learning to read. Child Development 74, 27–47.

Shu, H., McBride-Chang, C., Wu, S., et al., 2006. Understand Chinese developmental dyslexia: morphological awareness as a core cognitive construct. Journal of Educational Psychology 98, 122–133.

Shu, H., Peng, H., McBride-Chang, C., 2008. Phonological awareness in young Chinese children. Developmental Science 11, 171–181.

Siok, W.T., Fletcher, P., 2001. The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. Developmental Psychology 37, 886–899.

Taipale, M., Kaminen, N., Nopola-Hemmi, J., et al., 2003. A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. Proceedings of the National Academy of Sciences (PNAS) 100, 11553–11558.

Torgesen, J.K., 2005. Recent discoveries from research on remedial interventions for children with dyslexia. In: Snowling, M., Hulme, C. (Eds.), The Science of Reading. Blackwell, Oxford, pp. 521–537.

Torppa, M., Georgiou, G., Salmi, P., et al., 2012. Examining the double-deficit hypothesis in an orthographically consistent language. Scientific Studies of Reading 16, 287–315.

Wolf, M., Bowers, P.G., 1999. The double-deficit hypothesis for the developmental dyslexias. Journal of Educational Psychology 91, 415–438.

Wu, X., Li, W., Anderson, R.C., 1999. Reading instruction in China. Journal of Curriculum Studies 31, 571–586.

Yang, J.F., Zevin, J., Shu, H., et al., 2006. A "triangle model" of Chinese reading. In: The 28th Annual Conference of the Cognitive Science Society Proceeding. Lawrence Erlbaum, Mahwah, NJ, pp. 912–917.

Zhang, Y., Li, J., Tardif, T., et al., 2012a. Association of the DYX1C1 dyslexia susceptibility gene with orthography in the Chinese population. PLoS ONE 7 (10), e43192.

Zhang, Y., Zhang, L., Shu, H., et al., 2012b. Universality of categorical perception deficit in developmental dyslexia: an investigation of Mandarin Chinese tones. Journal of Child Psychology and Psychiatry 53 (8), 874–882.

Ziegler, J.C., Bertrand, D., Tóth, D., et al., 2010. Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. Psychological Science 21, 551–559.

Ziegler, J.C., Perry, C., Ma-Wyatt, A., Ladner, D., et al., 2003. Developmental dyslexia in different languages: language-specific or universal? Journal of Experimental Child Psychology 86, 169–193.

Real Estate Development

Richard Peiser, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Real estate development is the process through which cities are built. It concerns all aspects relating to the creation of new real estate projects and adding value to existing projects – site selection, design, finance, entitlements, construction, leasing, and operations. This article addresses the steps of development, industry structure, property types, design and density, and the stages of development and real estate finance. It deals with the impact of business cycles, institutionalization of the industry, smart growth/urban sprawl, public–private development, technology, and sustainability on real estate development, both public and private. It concludes with the special social responsibilities that developers have for the built environment.

What Is Development?

Real estate development is the process of *adding value* to real estate. While most people think of development in terms of new buildings and subdivisions, development refers to any situation where value is added to a property. Developers take *development risk*, which ranges from leasing-risk associated with buying buildings with high vacancy to buying raw land with no utilities or zoning, and undertaking all the risks associated with obtaining public approvals, financing, construction, leasing, and operations of a new development project.

Developers initiate, plan, and coordinate the entire development process – buying the land, designing the product, arranging financing, obtaining public approvals, constructing the building, leasing, managing, and ultimately selling the completed project. Development is distinguished from owning and managing property because it includes at least some element of adding value. While most developers also own and manage their buildings, many real estate investors and owners buy buildings that are completed and fully leased and that do not involve adding value through some form of development.

What Distinguishes Good Development from Bad?

While there is no absolute standard to distinguish good development from bad, the test is a simple one: Does the development enhance the social welfare of the community?(Peiser, 1990) Providing new housing and office space enhances the community, but if it does so at the expense of environmentally precious resources, then on balance it may reduce the social welfare. Unfortunately, social welfare is difficult to measure – especially the indirect costs and benefits of externalities associated with the development.

Obtaining approvals for development is a political process that is increasingly long and costly. Developers used to be able to obtain their building approvals before they closed on the land, greatly reducing their risk. Today, building approvals usually take longer than the contingent-purchase time that developers can negotiate in their site acquisition contracts. Where they must take down land before they know they can build what they want to greatly increases the risk of development.

One of the biggest challenges facing city planners and council members is to create an approval process that is unbiased and gives the various stakeholders appropriate voice in the outcome. At the same time, they must do their best to streamline the process and prevent unnecessary delays and costs.

Academic and Professional Antecedents

Real estate development relates to a number of academic fields including economics, urban planning, architecture, finance, law, political science, civil engineering, and geography. Urban economics provides the foundation for understanding real estate value and how cities change over time (Mills and Hamilton, 1994; Gordon et al., 1989). Real estate finance, public finance, land use, planning, environmental law and economics, infrastructure, market analysis, demography, architecture, and landscape architecture all have significant bodies of research that contribute to the understanding of urban development and its impact on cities and countryside (see Brueggeman and Fisher (2011); DiPasquale and Wheaton (1996); Case and Shiller (1989); O'Sullivan (2012)).

Developers are generalists. They need to be conversant with many different areas. Because they are responsible for conceiving, planning, and executing the entire process, developers must have strong financial backgrounds and be excellent 'salesmen,' no matter whether they are marketing their projects, raising capital, or speaking to neighborhood groups or planning boards. Because they determine what is built and how the urban landscape appears, they should be intimately familiar with cities – what works and what does not, both from a private development perspective and from a government perspective (see Garvin, 2002).

Successful development requires knowledge of all the real estate disciplines – brokerage, finance, law, political approvals, design, construction, and management. Developers usually have several years of real estate experience before they undertake their own projects, most often in brokerage, finance and investment, or project management. Legal, architectural,

planning, and construction backgrounds also provide excellent training.

Industry Structure

Ownership

Historically, private entrepreneurs who operated in a single market have dominated the development sector. Today, large publicly held real estate investment trusts (REITs) and private equity funds are among the largest owners of property. REITs allow investors to own shares in larger pools of property in the form of a pass-through entity that distributes to the shareholders substantially all of its taxable earnings and capital gains from equity ownership (equity REITs) or mortgages (mortgage REITs) of properties. REITs invest primarily in *income-producing* property such as office buildings and apartments. Private companies still dominate the development sector, although successful development firms are becoming larger, more institutionally financed, and are operating in multiple markets. REITs and private equity real estate funds typically do little ground-up development because they have tax-driven distribution requirements and depend on operating income to create value. However, *opportunistic* and *value-added* funds and REITs often take some development risk such as leasing or rehabilitation risk to generate higher returns (see Shilling and Wurtzebach, 2010).

Altogether, real estate industry comprises almost 20% of the nation's gross domestic product (Urban Land Institute, 1997). The real estate industry is composed of many different players (see Figure 1). Developers are the entrepreneurs who create and produce real estate. They may continue to own and manage the completed buildings or they may sell them to other parties. While developers have traditionally come from the private sector in the United States, redevelopment and economic agencies often function as public developers, either directly or through public–private partnerships. Affordable housing projects are often developed by nonprofit organizations. Many corporations and public agencies including the armed forces also develop property for their own use, which is later recycled into the inventory of space available for anyone.

Each of the players is involved in a different phase of producing or managing real estate.

Property Type

Figure 2 provides a breakdown of the real estate industry by location and property type. Each location represents a different and distinct market. All property types may be found in all locations. Certain property types, however, such as major office buildings, regional malls, and major hotels are more likely to be developed in the central business district and primary suburban nodes – typically located at the intersections of major freeways and mass transit hubs – whereas industrial buildings, apartments, and shopping centers are more ubiquitous.

The real estate industry is divided into six main property types: residential, office, commercial, industrial, hospitality, and land (see Figure 2).

Each property type is virtually a separate industry, serving different purposes, markets, and customers. Although industrial warehouses, for example, may have completely different development requirements from apartments, they share a common institutional framework with respect to land, financing, legal aspects, and even construction. Still, each product is distinct, and value is created from different features. Apartments and industrial both benefit from locations near major highways. Industrial developers, however, prefer to avoid residential areas where homeowners complain about truck traffic. And apartment developers are attracted to areas with good schools, good retail services, recreation, and

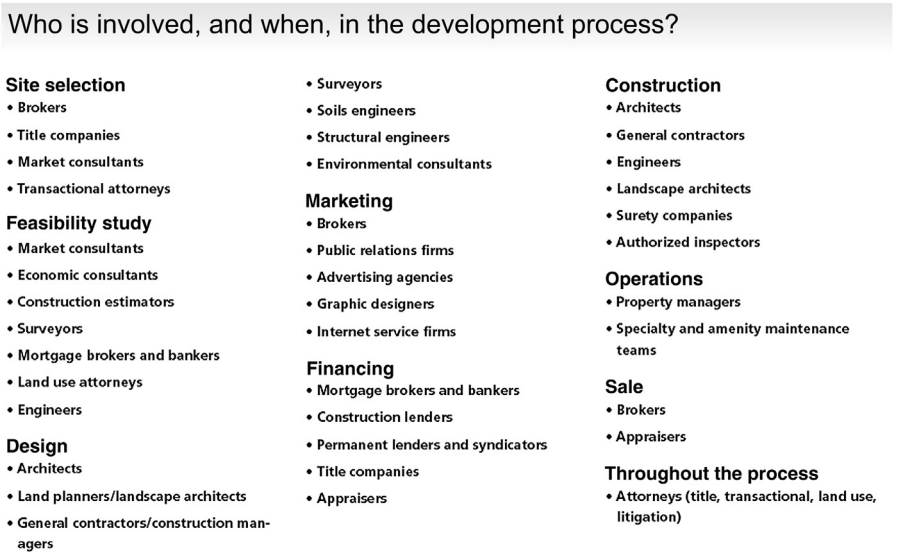

Figure 1 Real Estate Industry Players. Reproduced from Peiser, R.B., Hamilton, D., 2012. Professional Real Estate Development: The ULI Guide to the Business. ULI – The Urban Land Institute, Washington, DC, p. 41.

Location	Property type
CBD	*Residential*
Inner city	*Office*
Suburbs—Wealthy	*Industrial*
Graying	*Retail*
Exurbs	*Hospitality*
Rural	*Land*

Figure 2 Understanding the real estate industry – location and property type.

other amenities that are much less important to industrial tenants.

Developers, lenders, brokers, professional service providers, and most suppliers tend to specialize in one property type. While there is considerable crossover between property types, such as electrical subcontractors who work on both office buildings and apartments, professionals develop a network of relationships over the years that favor those who concentrate on a single property type.

For-Sale versus Income-Producing Properties

A critical distinction between property types is whether they are for sale or are income producing.

For-sale property	*Income-producing property*
Land	Apartments
Single-family homes	Office
Condominiums	Retail
	Industrial
	Hotels

For-sale property is sold to a buyer after it is developed, whereas income-producing property is held for the cash flow it generates. For-sale property is valued based on how much buyers are willing to pay for enjoying the benefits of ownership – such as the consumption and investment value of living in a home in a good school district near shopping and recreation.

Income-producing property includes office, retail, rental apartments, industrial, and hotels. Values for income properties are based on the cash flow (called net operating income or NOI) they generate from leases after operating expenses are deducted. Office buildings, for example, have tenants who typically sign 3- to 10-year leases. The developer uses the lease income to obtain long-term mortgage financing.

Income properties are valued by capitalizing their NOI. Capitalization rates are determined by the market for similar properties sold and represent their NOI divided by their purchase price. For example, if a building produces an NOI of $100 000 per year, and the cap rate for similar properties sold is .08, then its value would be $1.25 million ($100 000 divided by .08). If cap rates were 9%, then it would only be worth $1.11 million ($100 000 divided by 0.09).

When the office building is sold, investors pay a premium for leases signed by large corporations who have high credit ratings. Investors may be willing to purchase a high-quality office building (excellent location and design) occupied by a major corporation such as IBM on a cap rate as low as 5%, whereas they might require an 8 or 9% cap rate for the same building with small noncredit tenants. The rental payments of a major corporation are like a bond – carrying the same credit rating as the corporation itself. In fact, the owner can sue the corporation for lost rent if the tenant defaults on the lease payment.

Design	Density
Hi rise	*80+ Dwelling units per acre*
Low rise/garden	*30–80 Dwelling units per acre*
	18–30 Dwelling units per acre
Single family attached	*8–16 Dwelling units per acre*
Single family detached	*1 or less–12 Dwelling units per acre*

Figure 3 Residential design and density.

Design and Density

The design of a building is much more important than just the way it looks. Good design is a major determinant of real estate value, not only through its appearance and quality of finish but also through the building's attractiveness and efficiency. Well-designed buildings serve the needs and preferences of a specific market in a cost-effective manner.

Building density (measured by dwelling units per acre or floor area ratio – building square footage divided by land area) determines not only how crowded or built-up a neighborhood appears but also how much the land and the building is worth. The relationship, however, is not necessarily straightforward. Even though a developer may get more residential units on a site, the net profit from building the units may be lower than for fewer units. Building costs also go up with density, and buyers may be willing to pay enough more for the lower density units that the net value created goes down as density goes up. Nonetheless, land values usually increase with density.

Densities for residential property range from one unit per two or more acres for single-family detached homes to 80–200 units per acre for a high-rise apartment building. Residential developers tend to focus either on for-sale single-family homes or income-producing apartments. Figure 3 shows the range in densities for different types of residential construction.

Commercial building densities are defined as floor-area ratios (FARs – building area divided by land area). For example, a 20 000-square-foot site where the zoning allows an FAR = 2.0 could have a building with 40 000 square feet. The building could be a two-story building covering all of the site, or a four-story building covering half the site. FARs are usually defined as rentable area and exclude parking area. However, height restrictions may make the building smaller than it otherwise would be when parking is included because it is not cost-effective to put it all underground. Downtown high-rise buildings may have FARs as high as 10–20 or higher.

The Development Process and Stages of Development

Figure 4 shows the time line of development for income-producing property. The two main periods – development and operations – involve dramatically different risks, types of financing, and activities (see Peiser and Hamilton, 2012).

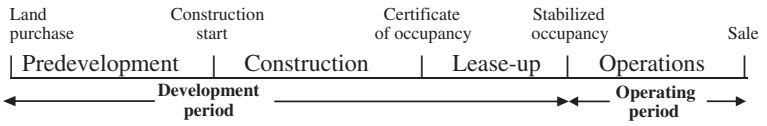

Figure 4 Time line of development.

The *development period* runs from the inception of deal through the time that the building reaches stabilized occupancy (90–95% occupancy). The development period is financed by a construction loan, which covers the construction and lease-up of the building. The building is technically placed in operation when the *certificate of occupancy* is received. *Lease-up* represents a major activity of development as well as risk, and consequently is shown on the above time line as part of the development period. The operating period runs from the point in time when the building reaches stabilized occupancy to the time of sale. Once stabilized occupancy is achieved, the developer is able to obtain a long-term permanent mortgage, which is used to pay off ('take out') the construction loan (see the Section Finance, below).

The development period in Figure 4 includes the stages of predevelopment, construction, and leasing activities. The predevelopment period runs from project initiation up to the start of construction. It includes feasibility studies, site acquisition, design, and finance.

Many of the critical development decisions are made even before a developer acquires a site as part of a series of feasibility studies for the market, site, environmental conditions, and finance. Market feasibility studies help the developer identify the best location and site characteristics for a particular development (see DiPasquale and Wheaton, 1996 or Brett and Schmitz, 2009). Market feasibility differs for each product type. It includes demographic projections, income, employment, retail sales, and growth projections. The market study should conclude with estimates of demand for the product expressed in units or square feet per year in a particular location for a particular product type, as well as estimates of current and projected supply. The market study tells the developer how much space can be built, what rent to charge, and how quickly it can be absorbed by the marketplace. The market study identifies competitive properties, and helps the developer determine the best location in which to develop his or her product. It also tells the developer what the design features and amenities should be in order to attract the greatest number of customers.

While many developers focus on low vacancy rates to justify new development, the best measure of a good market is one in which the months-to-absorb the existing and new supply is within the range of 12–18 months (3 years for high-rise office buildings). This is computed by dividing the total available supply including supply under construction by the absorption per year. For example, if the total supply for competitive office space is 1 million square feet, and the annual absorption of office in that submarket is 750 000 square feet per year, the months-to-absorb the product would be 1.33 years or 16 months – a relatively healthy market.

For site selection, developers will often consider as many as 50–100 different sites in order to find five sites, which they examine in detail. From the five, they may pick one site to purchase. Development is said to be a local business because experienced developers will know information about most of the potential development sites, competition, rents, and costs in their market area.

Site Acquisition

Property acquisition occurs in three main stages: (1) putting the property under contract ('opening escrow'), (2) going hard on the earnest money deposit, and (3) closing. Purchase contracts typically run from 60 to 120 days or longer. When the buyer makes an offer, he or she places an 'earnest money deposit' of 3–5% of the total purchase price. Purchase contracts have a number of contingencies, which the developer wants to satisfy before consummating the purchase. 'Closing' is the transfer of title to the buyer and the payment of all money to the seller.

After a site is placed under contract, the developer will hire consultants to determine the site's suitability for development. The analysis should cover site characteristics, environmental status, topography, drainage, utilities (water, sewer, storm, electricity, phone, gas, and cable), as well as neighboring properties, and the chance of obtaining necessary zoning and building approvals. Once the developer is satisfied that these conditions are suitable for development, he or she will go hard on the contract. 'Going hard' means that the earnest money becomes nonrefundable – when the major contingencies such as good title and finding no environmental problems are satisfied. Since it may take longer than the typical 15- to 30-day contingency period for the developer to perform the full range of studies necessary for development, the developer will often need to ask for an extension. The developer may also do much of the investigation before even placing the property under contract or while the purchase contract is being negotiated. In a hot market, sellers are usually unwilling to extend the contingency period because other buyers are waiting in the wings.

For a large project, the feasibility period can extend several years and cost millions of dollars before the developer receives all the necessary approvals. Such projects almost always require a developer to purchase the property before many of the feasibility studies are done and long before the public approvals are in place. Such projects carry full 'entitlement' risk and represent the most risky (and potentially the highest profitable) form of development.

Design

Design includes land planning, architecture, engineering, and consulting services that go into producing completed working drawings from which the building is built. The developer often

will perform considerable preliminary design work on a property even before he or she 'goes hard' in order to determine how much space can be built, and what it will cost based on preliminary construction bids. Design drawings are used by the developer to obtain financing and political approvals. The architect prepares working drawings after the developer approves the final design. The general contractor and the subcontractors use the working drawings to bid the project and to build it.

Finance

If developers have a distinct expertise, it is their ability to raise equity and debt financing for their development activities. Development financing for income properties has two main components – construction financing to build the project, and permanent financing for operating it. Developers obtain construction financing primarily from commercial banks. Construction loans have short terms – 1–2 years, and higher interest rates – 2–3% above 'prime' – than permanent mortgages. Permanent mortgages are long term – 30 years – and have lower interest rates than construction loans because the completed property, especially if it has high-quality tenants, offers better security to the lender than one under construction. Historically, pension funds, insurance companies, and savings and loan (S&L) associations provided most of the permanent mortgage financing, but today Wall Street is heavily involved in it through securitization, commercial mortgage-backed securities, and conduit lending. Banks and mortgage brokers originate mortgages called 'conduit loans' that are packaged together, securitized, and resold as assets to investors and institutional buyers. As a result, permanent financing is highly competitive in the market place and developers and owners of property have many more sources for permanent debt than they used to.

Equity financing is the most difficult and expensive money that developers raise. It represents the developer's own cash investment in a deal, or the money raised from 'equity investors.' It is much riskier than debt financing because equity investors receive cash flow only after the mortgage payments are made during operations and after the debt is paid off when the property is sold.

For example, if a building costs $10 million altogether, the developer would typically invest 25–30% cash equity and the rest would be covered by debt. Suppose the building at stabilized occupancy (90–95% occupancy) generates $1 000 000 in NOI (rent minus operating expenses). Figure 5 shows the development cost, financing, annual operating income, and mortgage assumptions for the sample building.

In Figure 5, the simple return measures are the *overall capitalization rate* and the *cash on cash return*. The overall capitalization rate equals the NOI/total cost = $1 000 000/$10 000 000 = 10%. The *cash on cash return* equals the cash throw off/equity = $339 612/$2 500 000 = 13.58%.

Sophisticated investors focus on before-tax and after-tax *internal rates of return* (IRRs) to determine whether or not they should invest in a particular real estate deal. IRRs represent annual rates of return on equity that are comparable to returns from investing in stocks or bonds. IRR requirements depend on risk. Investors require IRRs on equity that are typically 15–20% or higher for new properties under development, whereas they may require returns as low as 9–12% for investing in a fully leased office building with high-quality tenants (see Brueggeman and Fisher (2011) for a detailed discussion of IRRs).

Construction

Developers hire general contractors to build most nonresidential buildings, whereas homebuilders and many apartment developers often do their own construction. The general contractor (or homebuilder) hires subcontractors who specialize in constructing various parts of the building such as foundations, framing, glass, granite, electrical, plumbing, heating, ventilating, and air conditioning. Legal arrangements between the developer and the contractor may take several forms depending on who bears the risk for delivering the completed building within budget and on time. Common contract arrangements include fixed price, negotiated bid, design build, fixed price with a savings clause (where savings are split between the general contractor and the developer), and time and materials. Larger contractors are often unionized while small contractors and especially those who work on residential projects and in certain 'right-to-work' states such as

Development cost

Land	$ 2 000 000
Building	$ 8 000 000
Total	$10 000 000
Financing	
Equity	$ 2 500 000
Mortgage	$ 7 500 000
Total	$10 000 000

Annual operating income

Gross rent		$ 1 500 000
Vacancy	5 percent	$ 75 000
Effecting gross income		$ 1 425 000
Less operating expenses		$ 425 000
NOI		$ 1 000 000
Less debt service		$ 660 388
Cash throw-off		$ 339 612

***Mortgage assumptions**

Mortgage amount	$7 500 000
Interest rate	8 percent
Term	30 years
Annual debt service	$660 388

Figure 5 Development cost and operating income.

Texas are typically not unionized and tend to be lower cost. During the construction phase, the developer is most concerned about delivering the building on time and under budget. Since working drawings are often incomplete or have mistakes, a nonconfrontational relationship between the developer and contractor is key to success.

Marketing, Leasing, and Sales

Marketing for a new project begins even before the developer purchases the land. In order to obtain financing for office and retail projects, the developer usually must demonstrate that tenants are interested in the property by obtaining letters of intent or signed leases on 30–50% of the property. Preleasing requirements are higher when credit markets are tight such as in the wake of the global financial crisis of 2008–09.

Developers may have in-house marketing personnel or use outside brokers to lease and sell their property. Exclusive brokerage arrangements are common whereby a brokerage firm is the developer's primary agent. Brokers work on commission, which ranges from 4 to 6% of the lease or sale value for smaller deals down to 1–2% for larger transactions in the millions of dollars. For leasing office or warehouse space, the developer will typically pay the broker whom he or she has hired as the exclusive agent for his property a 3–4% commission for signing a lease if there are no other brokers involved. The tenant often has his own broker, in which case, a higher commission will be split between the two brokers.

Successful leasing involves more than having the lowest rent for the nicest space. The developer typically gives the tenant an allowance for *tenant improvements* (TIs), which covers interior finish. For law offices in which tenant finish can run $75–$100 or more per square foot (compared to standard finish of $20–$40 per square foot), the TI allowance can make the difference in getting the law firm to sign a lease.

The Internet has replaced newspaper advertisements as the primary source of leads for apartments and most other property types. Companies use Internet listing services to place banner advertisements and featured listings, and search engine marketing experts to develop keyword search (pay-per-click) campaigns to boost traffic. Social networking sites such as Facebook are invaluable for enhancing customer service and neighborhood connectivity, although they are less useful for generating leads (Peiser and Hamilton, 2012: p. 227).

Operations

Larger developers usually have their own property management departments to handle the ongoing leasing and management of a building after it is completed. Smaller developers may hire outside firms that specialize in property management. Property management activities include repair and maintenance of the property, leasing, relations with tenants, security, building, and payment of all capital and operating expenses. Good developers give careful attention to property management because it sets the tone of relations with tenants and determines their firm's image in the market place.

Reducing turnover reduces re-leasing costs as well as renovation and cleanup required to prepare the space to be leased again. Good management stresses the importance of communications, customer service, and prompt response to repair requests in order to reduce turnover.

One of the great advantages of owning property is the ability to refinance it and take out the additional financing proceeds (the difference between the new loan and the old loan) tax free. As the NOI increases over time, savvy owners will refinance their properties regularly to take advantage of the tax-free cash they receive by obtaining larger mortgages. The operating period ends with the sale of the property to a third party. Most real estate projects generate a significant part of their financial returns from appreciation in the value of the property relative to its development or purchase cost. IRRs are highly sensitive to the length of the holding period. The shorter the holding period, the higher the IRR is. Owners try to maximize the NOI of properties they are planning to sell in order to generate the highest possible sales prices.

Current Issues and Major Trends

The real estate industry is highly cyclical. Because development can take from 1 to 5 or more years to complete a new project, depending on the difficulty of obtaining public approvals, buildings are often started when the market is hot, and completed when it is not. The real estate industry collapsed in the late 1980s to early 1990s with S&L crisis, and again in the great recession of 2007–09. "Globalization and financial innovation increased both the volume and technical complexity of investment in real estate and resulted in substantial and sustained increase in credit availability to buyers" (Peiser and Hamilton, 2012: p. 372). The great recession was the result not only of cheap and easy credit but also of the failure of regulators and policy makers to understand the increasing connection of property markets via securitization to other areas of the financial system – the so-called systemic risk.

Cycles, however, are endemic to real estate (Wheaton, 1999; Grenadier, 1995). Those who learn how to play them well are the ones who will be most profitable. Many developers and investors who bought property at bargain prices in the early 1990s made fortunes as rents and prices recovered and surpassed prerecession peaks. In the S&L crisis, many institutions went bankrupt and were taken over by the Federal Deposit Insurance Corporation and their assets were sold through the government's Resolution Trust Corporation. Bargains have been slower to materialize in the wake of the global financial crisis of 2008 because banks have not been forced to get rid of nonperforming loans.

Institutionalization of the industry: In the wake of the S&L crisis of the late 1980s, two powerful forces emerged – a vast expansion Wall Street's real estate presence primarily through securitization and commercial mortgage-backed securities, and the reemergence of REITs as a major owner of property (Sanders, 1998). By 2012, REITs had equity value of more than $430 billion, or 3.2% of the total market capitalization – up from 1% 10 years ago. While REITs and other publicly financed companies, especially national homebuilders, play an increasingly important role in real estate development and ownership, entrepreneurial developers will continue to play a significant if less dominant role in the creation of new

product. Private equity real estate funds representing institutional investors are also financing more development projects. They give larger developers who have access to institutional money a competitive advantage in competing for new projects.

Smart growth/urban sprawl: The rapid expansion of suburban growth in the last half of the twentieth century has fueled a backlash against development and *NIMBYism* (not in my backyard). 'Smart growth' addresses the problems associated with urban sprawl – congestion, loss of farmland, lack of functional open space, monotonous and boring development, and inadequate infrastructure (see Peiser, 1989). Smart growth promotes mixed land uses, walkable neighborhoods, less dependency on automobiles, environmental preservation, strong sense of place, and a range of housing opportunities. It is criticized, however, for not having greater impact on development patterns. "Buyers demand low-density development and voters oppose high-density development near them" (Rein, 2009).

Public–private development: As the regulatory process for major development projects has become more time consuming and costly, more developers are joining forces with cities, redevelopment agencies, school districts, armed forces, and other government agencies to develop property. Public–private partnerships help developers obtain lower cost financing and necessary public approvals while giving public agencies a share in the profits and the ability to deliver more public benefits such as parks, plazas, roads, sewer plants, and other infrastructure (Kotin and Peiser, 1997).

Technology: The Internet and other new technologies are having dramatic impacts on all aspects of real estate, from office building construction to shopping center functionality resulting from Internet commerce. The pace of obsolescence of old shopping centers and industrial buildings is speeding up due to changing market conditions, the great recession, and concerns about sustainability. New industrial distribution centers and 'entertainment' shopping centers are replacing them. While home offices and telecommuting have caused alarm for future office space demand and spawned new work patterns such as 'hot desks' (where employees share office space, plugging their laptop into the company computer network as they move from office to office) workers still appear to want the 'social' aspects of coming to work. Suburban corporate campuses with high-tech communications, on-site recreation, sporting facilities, and high security are being built to respond to new workplace needs and employee preferences.

The Internet is changing how developers market their properties, how they reach potential customers, how they build support for public approvals, and how they build a sense of community with tenants and residents. Rapid technological change is impacting building and construction technology, property management, reporting to investors and lenders, and financing including mortgage originations and servicing as well as equity raising.

Sustainability: Principles of sustainability – energy efficiency, lower carbon footprints, water conservation, and environmental protection – have been fully embraced by corporate and commercial real estate leaders. Corporations and real estate fund managers are making investments in projects that demonstrate high levels of sustainability in terms of operations, materials, and impact on the environment. "Energy efficiency and cost savings will remain at the core of real estate sustainability strategy, but a growing number of corporate users see the value of enhanced employee health, well-being, and productivity within workplace environments" (Simpson, 2012). Despite the trend toward ever higher standards of sustainability, developers remain cautious about spending money where they do not see a payback within 5–7 years. Nevertheless, the strong orientation in favor of greater sustainability on the part of corporate users and investors is not only raising property values for sustainable development but also giving those projects an edge in the market place.

Social Responsibility

Real estate development touches on every aspect of people's lives. It determines not only the quality of the built-form environment, but also affects the life style and social well-being of residents and workers in every community. What developers build may last a lifetime or longer. Because of their far-reaching role, developers bear a special responsibility. Most developers care deeply about the quality of their projects, not just making money. High-quality development depends as much on public sector support and services as it does on the private activities and financial resources of the developer. Since real estate involves very long-term investment, the long-term stability and continuing attractiveness of the community to new residents and firms is essential to long-term profitability for the developer. While poor development is often vilified for destroying the environment and hastening neighborhood change, good developers are responsive to neighborhood concerns and sensitive to maintaining and improving environmental conditions. As public funding has been reduced for open space and infrastructure improvements, development is often the instrument through which they are provided. As the role of developers in providing previously public-goods expands, they need to be ever more knowledgeable about the urban context in which they build and about the multitude of forces with which they must contend. The Urban Land Institute's mission statement sums up the developer's charge: "to provide responsible leadership in the use of land and in creating and sustaining thriving communities worldwide" (ULI, 2012).

See also: Cities and the Financial Crisis; City Marketing; Environmental Planning; Family Business; Gentrification; Housing and Cities; Neighborhood Revitalization and Community Development; New Urbanism; Planning Issues and Sustainable Development; Property: Legal Aspects; Property Rights; Suburbanization and Suburbanism; Urban History; Urban Sprawl; Venture Capital.

Bibliography

Brett, D., Schmitz, A., 2009. Real Estate Market Analysis: Methods and Case Studies. The Urban Land Institute, Washington, DC.

Brueggeman, W., Fisher, J., 2011. Real Estate Finance and Investments. Irwin/McGraw-Hill, New York.

Case, K.E., Shiller, R.J., 1989. The efficiency of the market for single-family homes. The American Economic Review 79 (1), 125–137.
DiPasquale, D., Wheaton, W., 1996. Urban Economics and Real Estate Markets. Prentice Hall, Upper Saddle River, NJ.
Garvin, A., 2002. The American City: What Works, What Doesn't. McGraw-Hill, New York.
Gordon, P., Kumar, Richardson, H., 1989. Influence of metropolitan structure on commuting time. Journal of Urban Economics 26 (2), 138–151.
Grenadier, S.R., 1995. The persistence of real estate cycles. Journal of Real Estate Finance and Economics 10 (2), 95–119.
Kotin, A., Peiser, R., 1997. Public-private joint ventures for high volume retailers: who benefits? Urban Studies 34 (12), 1971–1986.
Ling, D., Archer, w., 2012. Real Estate Principles: A Value Approach. Mcgraw-Hill/Irwin Series in Finance, Insurance, and Real Estate.
Mills, E.S., Hamilton, B.W., 1994. Urban Economics. Addison-Wesley, Boston, MA.
O'Sullivan, A., 2012. Urban Economics. McGraw Hill/Irwin, New York.
Peiser, R.B., 1989. Density and urban sprawl. Land Economics 65 (3), 194–204.
Peiser, R.B., 1990. Who plans America: planners or developers. Journal of the American Planning Association 56, 496–503.
Peiser, R.B., Hamilton, D., 2012. Professional Real Estate Development: The ULI Guide to the Business. The Urban Land Institute, Washington, D.C.
Rein, Lisa, November 2, 2009. Study calls Md. smart growth a flop. The Washington Post.
Sanders, A.B., 1998. The historical behavior of REIT returns: a capital markets perspective. In: Garrigan, R., Parsons, J.F.C. (Eds.), Real Estate Investment Trusts, pp. 277–305.
Shilling, J., Wurtzebach, C., 2010. Is Value-Added and Opportunistic Real Estate Investing Beneficial? If So, Why? Real Estate Research Institute. www.reri.org.
Simpson, Larry, January 25, 2012. Sustainability in Real Estate – the Road Ahead in 2012. Environment Leader.
The Urban Land Institute, 1997. America's Real Estate: Natural Resource, National Legacy. The Urban Land Institute, Washington, D.C.
ULI, the Urban Land Institute, 2012. The Mission and Priorities of ULI. The Urban Land Institute, Washington, D.C. http://www.uli.org/about-uli/mission-priorities/.
Wheaton, W., 1999. Real Estate Cycles Some Fundamentals. Real Estate Economics 27 (2), 209–230.

Realism, in Applied Research

George Julnes, University of Baltimore, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Realism is a philosophy of science that, in its many varieties, seeks to provide adequate justification for believing that (1) the commonsense, everyday world we experience is more or less how the world really is and (2) our understanding of this world has improved across the millennia and can be expected to continue to improve, both with regard to the structures in our world (e.g., subatomic particles, as well as social structures) and the processes (e.g., underlying causal mechanisms). While most social and behavioral scientists are realists in this sense, it has proven difficult for philosophers to explicate coherent realist philosophies that can withstand criticism. As a result, there have been dozens of variants of realism proposed over the past two centuries, with the changes over time instructive for the future of realist philosophies and the implications for conducting productive social and behavioral research. One development in recent decades has been greater understanding of how humans make sense of the world, informing what has been called 'naturalized epistemology' with an explosion of cognitive and neuroscience knowledge. This, in turn, has resulted in some realist philosophers revisiting American pragmatism and promoting revised pragmatist theories that offer clearer support for social and behavioral sciences.

Realism Defined

Realism, as a philosophical approach with relevance for the social sciences, is a continuingly developing set of philosophical positions intended to justify taking seriously our commonsense experience of our world. Indeed, realism can be thought of as a position that many, or most, want to hold, with new versions emerging as older ones become seen as vulnerable to criticism. The result is a diversity of realist positions defined as much by a tension with what they oppose as by what they claim.

Common among all recent forms of realism is a critical stance regarding epistemology, acknowledging that our understanding of the real world is partial, fallible, always in revision, and never what Hilary Putnam refers to as a God's Eye View, a single, correct understanding. Another commonality concerns ontology, or the nature of our world, in that realists accept that the world exists to a large degree independent of our experience of it and that there is real causality in this world. That is, there are real underlying dynamics, some refer to mechanisms, that are responsible for what we observe (e.g., John stole the money because he was hungry and realized that no one was watching). However, the outcomes of these dynamics depend on multiple factors (e.g., John was hungry but thought he was being watched), complicating the 'constant conjunction' required by deterministic Humean models and making prediction possible only in special circumstances.

Historically, there are two main ways that realism has been understood and defined, one contrasting realism with *idealism* and the other with *nominalism*.

Reality of 'Things' in the World

When contrasted with idealism, realism is the position that the 'things' in the world exist, in some manner, independently of our perceptions of them. For example, water flowing over a waterfall creates the same erosion of the rock whether or not observed by any humans or other creatures.

Reality of Forms

When contrasted with nominalism, realism is the belief in universals, sometimes presented as Platonic forms or natural kinds. Natural kinds refer to a fundamental similarity among particulars, which "is immediately definable in terms of kind; for, things are similar when they are two of a kind" (Quine, 1991, p. 160).

For example, when we distinguish oak trees from maple trees, we are invoking the notion of natural kinds, as also when distinguishing subatomic protons and electrons. Furthermore, these groupings evolve as our understanding advances – "New groupings, hypothetically adopted at the suggestion of a growing theory, prove favorable to inductions and so become 'entrenched'" (Quine, 1991, p. 165).

When dealing with less clear-cut examples, as when categorizing different kinds of people, the question is whether we are dealing with real categories of things in the world that exist independently of our use of them, or whether we classify things simply as heuristics to simplify our communications about them. In applied research, this debate arises when we try to measure 'constructs,' such as intelligence, depression, job satisfaction, and transformational leadership. Are these constructs real or better understood as convenient labels?

Reality and Scientific Theories

These two senses of realism contribute to a debate with particular relevance to applied social research – how are we to make sense of the relationship between theory and the reality that it seeks to represent? No one currently claims that our theories correspond in some direct and absolute way to a preexisting reality; no one claims that there is no 'reality' that our theories seek to represent. It is a matter of degree. "To assert that something is somehow mind-independent is to move in the realist direction; to deny it is to move in the opposite direction. No sane position is reached at either extreme" (Honderich, 2005, p. 787). To understand the opposition to the apparent commonsense emphasis of realism, we need to appreciate both empiricist and constructivist arguments against realism.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10559-8

Empiricist Antirealism

The empiricist argument against realism is based on objecting to any statements about unobservable phenomena. Imagine two theories that explain observed regularities in terms of different unobservable mechanisms. Since the mechanisms are unobservable, "no scientific evidence can bear on the question of which of these theories provides the correct account of observable phenomena…it follows that scientific evidence can never decide the question between theories of unobservable phenomena and knowledge of unobservable phenomena is thus impossible" (Boyd, 1991, pp. 196–197).

Constructivist Antirealism

The constructivist argument against realism is more subtle. Given that every aspect of the scientific enterprise is theory dependent (theory provides the frame in which we judge the acceptability of methodology, theory, and even the nature of observation), science can work only if what is being studied is also theory dependent.

Varieties of Realism in History

Most scholars of realist philosophy comment on its diversity. For example, Harre titled a book *Varieties of Realism*; Putnam named one *The Many Faces of Realism*. This diversity results in realism being the dominant perspective but also feeling marginalized. In that the history of realism is one of overcoming challenges, we can see an evolution in the balance between direct perception of the world and perception being only indirect, conditioned by what humans bring to the experience. As such, all realist positions described here (see O'Shea, 2008; Reese, 1980; for more background on each of these) are an attempted balance between what is called naïve realism, the view that reality is exactly like our perception of it, and idealism.

Locke's Representational Realism

John Locke's theory of primary and secondary properties was an early attempt to reconcile an understanding that there are things in the world, our senses provide sense data, which are then organized into perceptions. This is a realist theory in that the things of the world exist independently of human experience and understanding, but it avoids the problems of naïve realism by having perception mediated by ideas.

Commonsense Realism of Thomas Reid

Writing in the second half of the eighteenth century and reacting to the theory of knowledge of the British empiricists (no knowledge of unobservables), Reid ascribed more trustworthiness to commonsense descriptions than to the philosophical abstractions of Hume and Berkeley. Similarly, our intuitions of ethics were to be trusted over philosophical analyses.

New Realism

Near the beginning of the twentieth century, a group of philosophers (Holt, Marvin, Montague, Perry, Pitkin, and Spaulding) promoted a position called 'new realism' that sought to counter prevalent interest in neo-Hegelian idealism. Making use of work by G.E. Moore (e.g., "The Refutation of Idealism," 1903), this group published an initial article in 1910 and a book in 1912 titled *The New Realism* (Holt et al.). Rejecting both idealism and materialism, this group argued for a realism that "is opposed to subjectivism or epistemological idealism which denies that things can exist apart from an experience of them, or independently of the cognitive relation" (Holt et al., 1910, p. 396). However, this entailed something of a direct realism. "For the new realists in general, there is an act of cognition and an object or datum directly cognized, without any intermediary ideas or perceptions of the sort proposed by traditional representational or indirect realist (Descartes, Locke)" (O'Shea, 2008, pp. 218–219).

American Critical Realism

In 1920, the book *Essays in Critical Realism* (coauthored by the US philosophers Durant Drake, Arthur O. Lovejoy, James Bissett Pratt, Arthur K. Rogers, George Santayana, and Roy Wood Sellars) presented critical realism as an American movement of a direct realism that also addressed 'the nature and mechanism of knowing' to allow for a critical stance. Critical realism hence was intended, in part, to counter the new realists, who, with their near identification of the things of the world and our sense-datum, were seen as insufficiently 'critical' in framing the strengths and limits of human understanding.

Further developing this critical stance, Roy Wood Sellars (who published a book titled *Critical Realism* in 1916) labeled his variant 'emergent realism' to emphasize the nonreductionist nature of his framing of causality. Specifically, "Sellars, like the British emergentists, made use of the 'layered model' of reality, one according to which entities of the world are ordered in a hierarchy of increasingly complex 'levels,' with humans and higher-organisms, and their emergent characteristics of life and mentality, occupying the higher levels" (Kim, 2003, p. 89).

Sellars was also a proponent of a robust naturalism that appears in all subsequent realist theories. Indeed, his books "outline a conception of the human being as a complex product of organic evolution, resulting in a dynamic and multi-leveled cognitive and causal relationship to its environment" (O'Shea, 2008, p. 220). In line with this naturalist perspective, he also authored the first draft of the 'Humanist Manifesto' of 1933. Hence, we see that the critical realist perspective contained fairly modern notions of human understanding being based on evolved cognitive capacities that were, nevertheless, constrained by what evolution did not provide. While this naturalist perspective was overshadowed by analytic philosophy for much of the middle of the twentieth century, these insights provided the foundation for the scientific realism that was developed in the 1950s and continues to this day.

Current Realist Perspectives

As noted above, the variety of realist accounts is the result of something akin to natural selection – theories overly vulnerable to criticism do not 'survive,' so that realist theories still

being proposed are those that have yet to succumb. The overarching frame for recent approaches is scientific realism, which contributes in large measure by being an alternative to positivism (with logical empiricism being a recent variant) and constructivism (we 'construct' reality and know it only through our constructions). As such, the general scientific realist frame is presented first, followed by variants offered by Hilary Putnam and Roy Bhaskar.

Scientific Realism

In the 1950s Wilfrid Sellars, son of Roy Wood Sellars, promoted scientific realism as more consistent with the world that modern science at that time was presenting. "The nonreductive materialism that was being sketched in original ways by Roy Wood Sellars should certainly be credited… with anticipating many developments later in the century. But it was the tools provided by the various phases of the logical and linguistic turn in the analytic tradition that enabled Wilfrid Sellars to develop more powerful and substantive theories on just how…the perceptual, conceptual, and normative dimensions of the perceptually experienced world or *manifest image* of persons-in-the-world are to be successfully integrated with the comprehensively causal and naturalistic picture of physical nature that is projected in our ongoing scientific image of that same world" (O'Shea, 2008, p. 233).

In its recent form, scientific realism is contrasted with social constructivism as the primary alternatives for a philosophy of science. "According to realists, when a well-confirmed scientific theory *appears* to describe unobservable 'theoretical entities,' it is almost always appropriate to think of its 'theoretical terms' as *really* referring to real unobservable features of the world, *which exist independently of our theorizing about them*" (Boyd, 1991, p. 11). As with other realist theories, it is this insistence on the independent existence of the reality being described that distinguishes realism from social constructivists. This is not to say that social constructivists are trapped in the empty claim that there is no reality, only that everything the realist thinks is 'real' is instead socially constructed. "Socially constructed causal and metaphysical phenomena are, according to the constructivist, real: they are as real as anything scientists can study ever gets" (Boyd, 1991, p. 13).

Elaborating these points, Boyd (1991, p. 195) presents four basic theses of scientific realism; in brief, scientific theories (1) "should be interpreted 'realistically'"; (2) "are confirmable *and in fact often confirmed* as approximately true by ordinary scientific evidence"; (3) progress with "successively more accurate approximations to the truth about both observable and unobservable phenomena"; and (4) describe a reality that "is largely independent of our thoughts or theoretical commitments."

Causal Explanation

Consistent with the philosophy of science emphasis, scientific realism addressed the nature of causal explanation. An early version of this was C.S. Peirce's concept of 'abduction,' which involves the probabilistic inference of observing something surprising, considering possible explanations, and then accepting, at least tentatively, the explanation that best fits what is known. This form of knowing was developed further in scientific realism in the 1950s, with Hanson promoting retroduction as an updated version of abduction. This is currently presented as a natural human tendency to seek an 'inference to the best explanation.'

Naturalized Epistemology

An additional emphasis of scientific realism, continuing to be developed today, is called natural epistemology. One aspect of this is the recognition that many prior philosophical debates on knowing, or epistemology, took place without any interest in or understanding of how humans actually make sense of the world. If a type of inference did not satisfy the demands of deductive logic, then the knowing that resulted was second-rate and suspect. In contrast, naturalized epistemology recognizes that humans have evolved to be reasonably good at making sense of the world, and that these natural capacities should be respected. "The old epistemology aspired to contain, in a sense, natural science; it would construct it somehow from sense data. Epistemology in its new setting, conversely, is contained in natural science, as a chapter of psychology" (Quine, 1994, p. 25). This emphasis on evolved natural capacities supports arguments for trusting our commonsense understanding of the world.

Putnam's Internal Realism

Hilary Putnam has been a prominent American realist for the past half-century. Having studied under the noted scientific realist Hans Reichenbach, Putnam developed a variant (labeled 'internal realism'), understood as a type of commonsense realism.

Commonsense Realism

One commonsense aspect of Putnam's realism concerns the goal regarding our understanding of the world – rather than advancing knowledge claims as being 'justified' in a formal sense, he uses Dewey's term 'warranted assertibility.' From this, his realist argument includes: "In ordinary circumstances, there is usually a fact of the matter as to whether the statements people make are warranted or not…[and yet] Our norms and standards of warranted assertibility are historical products; they evolve in time" (1990, p. 21). This is a realist view that embraces our commonsense world but also is skeptical of any formal, final criteria for justifying acceptance of specific knowledge claims. In addition, the commonsense stance also involves 'antiformalism,' the view that our formal logics and formal accounts of knowing can be useful guides for how we approach the world "but it manifests itself in a sustained critique of the idea that formal models, in particular, systems of symbolic logic, rule books of inductive logic, formalizations of scientific theories, etc. – describe a condition to which rational thought either can or should aspire" (Putnam, 1995, p. 63).

Possibility of Progress

It is possible to 'get it right' in our understanding without 'absolutely getting it right.' So, for Putnam, it is appropriate to conclude "Newton got it right when he said that the tides are caused by the gravitational pull of the moon and the sun" (1990, p. 131) without having to claim that the Newtonian worldview is correct or is a limiting case of the worldview

presented in quantum mechanics. The same can be said for what quantum mechanics gets right – we have every reason to expect our understanding to continue to develop.

As for whether it is appropriate to hold onto the modernist faith in progress, Putnam notes that we cannot prove that faith "but our action is 'fantastic, directed to empty, imaginary ends' if we do not postulate the possibility of progress" (Putnam, 1990, p. 191). On the other hand, this need not suggest that our understanding is converging on some God's Eye View of the world. "There is no more evidence that science converges to one final world-view than there is that literature or morality converge to one final world-view" (p. 130).

Fact–Value Relationship

Putnam rejects the notion of a fact–value dichotomy (though does accept that it can be useful to note distinctions between them that can be illuminating). As such, he also accepts that it is possible to talk about some value judgments being more warranted than others. "As a being who makes value judgments every day, I am *of course* committed to the idea that there are true value judgments; *what must be the case if there are to be true value judgments?* In what kind of world can there be true value judgments?" (Putnam, 1995, p. 43).

Bhaskar's Critical Realism

Another current variant of scientific realism was developed by Roy Bhaskar and his dissertation mentor, Rom Harre. Bhaskar's realism was developed first (1975) as *transcendental realism*, providing a more persuasive realist foundation for the natural sciences. He subsequently elaborated his theory to include *critical naturalism* for the social sciences, emphasizing a reality that is differentiated and stratified. Later commentators, whether aware or not of the American tradition of critical realism, concatenated transcendental realism and critical naturalism into 'critical realism,' and the name stuck.

Transcendental Realism

Bhaskar characterizes his basic position as transcendental realism, in homage and contrast to Kant's transcendental idealism, and states that it "regards the objects of knowledge as the structures and mechanisms that generate phenomena" (1975, p. 25). This is consistent with other variants of realism, as is the claim that "critical realism is realist because it holds, contra postmodernism and social constructionism, that research is about gaining knowledge of a reality that exists independently of our representations of it" (Cruickshank, 2003, p. 3).

One concept central to the ontology of transcendental realism is the notion of 'stratified' layers of our world. At the deepest level (the domain of 'real') are real structures and mechanisms. These underlying mechanisms result in 'events' in the world that are potentially observable in the intermediate level, the domain of 'actual.' At the surface, 'empirical' level, we have 'experiences' of the world and make sense of observed patterns.

Further, the variation, or 'differentiation,' we see in the world, the lack of 'constant conjunction' of cause and effect, arises because "real structures exist independently of and are often out of phase with the actual patterns of events. Indeed it is only because of the latter that we need to perform experiments and only because of the former that we can make sense of our performances of them" (Bhaskar, 1975, p. 13). That is, there are meaningful patterns in reality that justify our efforts to distinguish and explain them, but the underlying mechanisms that we want to understand are not expected to yield the same effects in different contexts.

Critical Naturalism

Elaborating his vision to the social sciences, Bhaskar addressed the complexities of open systems with a theory he called critical naturalism, presented in *The Possibility of Naturalism* (1979). Central to this vision is a dialectic between the sociological structures and psychological agency: "critical realists argue: (1) that structure and agency need to be linked, and that we must avoid putting all the explanatory weight on structures or individuals, as this results in determinism and an inability to explain individuals' social relationships" (Cruickshank, 2003, p. 3).

This approach qualifies as a critical one because of its emphasis on critical dialogue. "Critical realism is critical, as regards methodology, because it hold that the concepts which inform the meta-theory that defines structure and agency can only be developed via a critical dialogue with alternative social ontologies" (Cruickshank, 2003, p. 3). Further, there is also a normative claim, building on Marxist thought that justifies a moral vision of the world. "Critical realism is also critical in a political as well as a methodological sense…to enable the move from facts to values…to explore how existing social, political and economic relations create inequality, and turn on exploitation, in order to develop a normative critique against those relations" (Cruickshank, 2003, p. 3).

Another concept important in transcendental realism but particularly so in critical naturalism is 'emergent order,' the idea that phenomena cannot be reduced to what is understood of lower-order understanding. "It follow from this that the operations of the higher level cannot be accounted for solely by the laws governing the lower level in which we might say the higher-order level is 'rooted' and from which we might say it was 'emergent'" (Bhaskar, 1975, p. 113). This allows us to be informed by lower-order dynamics but not to be constrained by them for our understanding.

Implications of Realism for Social and Behavioral Sciences

The most cited value of realist philosophy for social sciences is the warrant that it provides to avoid choosing sides in some ongoing methodological debates, offering instead a position that, potentially, can support some reconciliation. That is, most social scientists do not embrace standard versions of either positivism or constructivism, instead believing that there is a real world that we want to, and can to some degree, understand. For example, Julnes and Mark (1998, p. 36) ask "Are we to choose, in Bhaskar's words, between either a conceptually impoverished and deconceptualizing empiricism, or a hermeneutics drained of causal import and impervious to empirical controls (1989a, p. 12)?…[instead] we promote an alternative based on the philosophical paradigm known as realism." Similarly, Meehl (1986, p. 322) noted

that "As to realism, I have never met any scientist who, when doing science, held to a phenomenalist or idealist view; and I cannot force myself to take a nonrealist view seriously even when I work at it."

Implications for Ontology and Epistemology

As stated above, the realist ontology begins with the belief that the real world is approximately what we experience, and its epistemology with belief in having fallible, constructed knowledge of a real world. Meehl (1986, p. 322) speaks for this commonsense view in stating, "So I begin with the presupposition that the external world is really there, there is a difference between the world and my view of it, and the business of science is to get my view in harmony with the way the world really is to the extent that is possible."

Recent realist perspectives also acknowledge the difference between closed systems and open systems, wherein the open systems of interest to social and behavioral sciences involve underlying mechanisms operating in complex, changing environments such that (1) variations in contexts can influence expression of those mechanisms and (2) emergent phenomena will result that cannot be reduced directly to the dynamics of less complex phenomena (Julnes and Mark, 1998). These two points underlie Don Campbell's *Evolutionary Epistemology* (1974) and, hence, the position referred to as *Evolutionary Critical Realism* in Cook and Campbell (1979, pp. 28–30). As such, there will be few 'universal laws,' and the perceived regularity of observed patterns will depend on our understanding of the mechanisms underlying them and the ways in which those mechanisms are influenced by contexts.

Implications for Methodology

There are different views of whether there is a 'realist methodology.' For example, Pawson and Tilley (1997) develop a specific methodology for what they call 'realistic evaluation' that is based on identifying patterns of mechanism (M) that operate differentially in different contexts (C) to yield observed outcomes (O). However, in addition to emphasizing a qualitative approach to understanding how mechanisms are triggered in specific contexts (referred to as the CMO framework), they also acknowledge some value for pluralism of methodology. Similarly, others (e.g., Yeung, 1997) criticize the notion that there is a 'realist methodology.'

We can, however, discern a general logic suggested by realist views. The logic is an iterative one that involves the process of abduction that Peirce explicated over a century ago and which has since been elaborated as pattern matching – identifying potential mechanisms responsible for observed regularities, testing out those mechanisms, and then developing ever more refined understandings of mechanisms and testing those.

Pattern Matching

Pattern matching is the general realist process promoted by Campbell for supporting conclusions about causal mechanisms. It is an iterative process, moving from theorizing to testing to theorizing, and so on, but it is useful to highlight the processes of both retroduction and competitive elaboration.

Retroduction: Inferring Underlying Mechanisms

Peirce's 'abduction,' reframed by Hanson and others as retroduction, refers to the natural human capacity for speculating about mechanisms responsible for observed regularities and accepting provisionally the mechanism most consistent with the observed regularities. This has been a consistent feature of realist theories for a century and remains foundational still.

Competitive Elaboration: Testing Presumed Mechanisms

Competitive elaboration is a logic rather than a specific method and "refers to the process by which alternative explanations (mechanisms) – whether alternative program theories or validity threats – are ruled out" (Mark et al., 2000, p. 256).

An important point to make is that this process is used in some form by most, if not all, social scientists. For example, when asked how to improve testing theories in observational studies, "Sir Ronald Fisher replied, 'Make your theories elaborate'…[meaning that] when constructing a causal hypothesis one should envisage as many different consequences of the truth as possible, and plan observational studies to discover whether each of these consequences is found to hold" (Cochran, 1965; cited in Mark et al., 2000, pp. 254–255).

As an example of this approach, consider Milgram's series of studies on the factors, or mechanisms, responsible for the willingness of participants to increase what they believed was the voltage of an electric shock applied to another participant. Through retroduction, one posited mechanism was 'identification with the experimenter' and so, in line with competitive elaboration, "one set of studies varied the exposure to the experimenter, with some subjects having the experimenter present during the experiment and others only meeting the experimenter at the beginning of the experiment" (Julnes, 2005, p. 888).

Future Directions

Realism will continue to be a dominant perspective as a philosophy of science, particularly in the social and behavioral sciences. It will maintain this position in part because people want to believe in realism. In the process, however, it will continue to evolve, both in response to criticism and in response to advances in our understanding.

One form of development involves incorporating new phenomena into the perspective. For example, Bhashkar has continued to elaborate his theory, moving beyond his initial focus on transcendental realism and critical naturalism to emphasize the emancipatory imperative of critical realism (developed around 1986–94 and referred to as 'dialectical critical realism') and later its spiritual aspects ('transcendental dialectical critical realism,' developed around 1994–2000).

Another form of development comes from developing new technology that yields patterns with implications for our ways of knowing and, hence, for the types of realism that seem most consistent with such ways of knowing. Two aspects of this form of development can be seen in the increasingly productive fields of cognitive science and neuroscience. The understanding that is resulting from this work is also increasing interest in American pragmatism, a perspective developed over a century ago but increasingly relevant today.

Cognitive science and Neuroscience

Debates about how to conduct research (e.g., regarding the contrasting strengths and limitations of qualitative and quantitative knowing) will increasingly be framed in terms of our understanding of cognitive science and neuroscience. For example, in the controversy over the appropriate roles for 'rigorous' quantitative designs for causal impact conclusions, research on causal cognition in children showed that they routinely make use of two approaches to understanding causal relations – one, depending on the covariation of events that is the bedrock of quantitative methods and, the other, a sort of a 'folk psychological domain,' that involves the intentionality of actions that makes up the agency studied with qualitative methods. That is, evidence about our evolved capacities for causal understanding suggests two very different evolved capacities that, not surprisingly, have resulted in two sets of evolved methodological tools to support those capacities.

Similarly, neurological research has shown that different parts of the brain are involved in different aspects of valuing. When people are thinking about receiving an immediate reward, the 'emotional' area of the brain (midbrain dopamine system, the nucleus accumbens) is active; when thinking about long-term rewards, the 'rational' prefrontal cortex area is active. We can expect that our efforts to understand value judgments will lead to new methodologies for applied social sciences like program evaluation and policy analysis, some methods supporting our rational, analytical valuing and others our more holistic valuing.

Pragmatism: Theories and Methods as Tools for Assisted Sensemaking

As discussed above, many of the foundations of recent realist thinking, such as Bhaskar's retroduction and evolutionary epistemology, originally were developed by people like Peirce, James, and Dewey who are grouped as early pragmatists (also, in the history of psychology James and Dewey are grouped in the paradigm now known as functionalism, distinguishing it from the earlier structuralist paradigm). To see how this fits with realist theories, "remember the evolutionary roots of pragmatism, wherin our cognitive abilities are viewed as having evolved to solve problems" (Julnes, 2012, p. 112). Once our theories and methods are viewed as tools, we are freed to some degree from dependence on formal philosophic positions to guide the development of the social and behavioral sciences. As Putnam explains, having evolved into a greater appreciation of pragmatism (in part, as he notes, due to the work of his wife, Ruth Anna Putnam) he has come to see that it contains insights that seem 'lastingly right' and which have implications for both realism and the antirealism that it opposes. He expresses appreciation of the "pluralism and thoroughgoing holism which are ubiquitous in Pragmatist writing" (1995, p. xii) and notes that viewing "fact, theory, value and interpretation as interpenetrating undermines a certain sort of metaphysical realism" and also "undermines fashionable versions of antirealism and 'postmodernism'" (1995, p. xii). As this vision of pragmatism develops (and it escapes the discredited one-line description that it is based on the belief that "if it works, it must be 'true'"), new versions of realism will begin to merge with it and will be successful to the extent that they support more productive work in the social and behavioral sciences.

See also: Applied Social Research, History of; Campbell, Donald Thomas (1916–96); Dewey, John (1859–1952); Evolutionary Epistemology; James, William (1842–1910); Logical Positivism and Logical Empiricism; Pragmatism: Philosophical Aspects; Putnam, Hilary (1926–); Social Neuroscience of the Twenty-First Century.

Bibliography

Boyd, R., 1991. Introductory essay. In: Boyd, R., Gasper, P., Trout, J.D. (Eds.), The Philosophy of Science. MIT Press, Cambridge, MA, pp. 3–35.

Bhaskar, R., 1975. A Realist Theory of Science. Verso, London.

Bhaskar, R., 1979. The Possibility of Naturalism. Hemel Hampstead, Harvester Wheatsheaf.

Boyd, R., 1991. On the current status of scientific realism. In: Boyd, R., Gasper, P., Trout, J.D. (Eds.), The Philosophy of Science. MIT Press, Cambridge, MA pp. 195–222.

Campbell, D.T., 1974. Evolutionary epistemology//The philosophy of Carl Popper. Le Selle (Ill.): Open court.

Cochran, W.G., 1965. The planning of observational studies of human populations. Journal of the Royal Statistical Society. Series A (General), 234–266.

Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: design and analysis for field setting. Houghton Mifflin, MA.

Cruickshank, J. (Ed.), 2003. Critical Realism: The Difference It Makes. Routledge, London.

Holt, E.B., Marvin, W.T., Montague, W.P., Perry, R.B., Pitkin, W.B., Spaulding, E.G., 1910. The program and first platform of six realists. The Journal of Philosophy, Psychology and Scientific Methods, 393–401.

Honderich, T., 2005. Oxford Guide to Philosophy. Oxford University Press, Oxford, England.

Julnes, G., 2005. Experiment, overview. In: Kempf-Leonard, K. (Ed.), Encyclopedia of Social Measurement. Academic Press, San Diego, CA, pp. 885–899.

Julnes, G., 2012. Developing policies to support valuing in the public interest. In: Julnes, G. (Ed.), Promoting Valuation in the Public Interest: Informing Policies for Judging Value in Evaluation, New Directions for Evaluation, vol. 133. Jossey-Bass, San Francisco, pp. 109–130.

Julnes, G., Mark, M., 1998. Evaluation as sensemaking: knowledge construction in a realist world. In: Henry, G., Julnes, G., Mark, M. (Eds.), Realist Evaluation: An Emerging Theory in Support of Practice. New Directions for Evaluation, No. 78. Jossey-Bass, San Francisco, pp. 33–52.

Julnes, G., Mark, M., Henry, G., 1998. Promoting realism in evaluation: realistic evaluation and the broader context (conceptual review of Realistic Evaluation by R. Pawson and N. Tilley). Evaluation 4, 483–503.

Kim, J., 2003. The American origins of philosophical naturalism. Journal of Philosophical Research 28, 83–98.

Mark, M., Henry, G.T., Julnes, G., 2000. Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs. Jossey-Bass, San Francisco.

Meehl, P., 1986. What social scientists don't understand. In: Schweder, R.A., Fiske, D.W. (Eds.), Metatheory in Social Science: Pluralisms and Subjectivities. University of Chicago Press, Chicago, pp. 315–338.

O'Shea, J., 2008. American philosophy in the twentieth century. In: Moran, D. (Ed.), The Routledge Companion to Twentieth Century Philosophy. Routledge, London, pp. 204–253.

Pawson, R., Tilley, N., 1997. Realistic Evaluation. Sage, Thousand Oaks, CA.

Putnam, H., 1990. Realism with a Human Face. Harvard University Press, Cambridge, MA.

Putnam, H., 1995. Pragmatism. Blackwell, Cambridge, MA.

Quine, W.V.O., 1991. Natural kinds. In: Boyd, R., Gasper, P., Trout, J.D. (Eds.), The Philosophy of Science. MIT Press, Cambridge, MA, pp. 159–170.

Quine, W.V.O., 1994. Epistemology naturalized. In: Kornblith, H. (Ed.), Naturalizing Epistemology. MIT Press, Cambridge, MA, pp. 15–31.

Reese, W.L., 1980. Dictionary of Philosophy and Religion. Humanities Press, Atlantic Highlands, NJ.

Sellars, R.W., 1916. Critical Realism: A Study of the Nature and Conditions of Knowledge. Rand McNally, New York.

Yeung, H.W.C., 1997. Critical realism and realist research in human geography: a method or a philosophy in search of a method? Progress in Human Geography 21 (1), 51–74.

Realism/Neorealism

Arthur A Stein, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Realism is an approach to international politics that has a long historical tradition and numerous variants. It focuses on the actual state of the world, takes as given that the key actors are self-interested states, and that they interact in an anarchic setting, one in which there is no central authority to enforce order. As a result, states seek power and what emerges is a conflict-ridden world in which the balance of power is the only basis for order. The perspective has problems and limitations but few analyze international politics without focusing on the distribution of power.

Varieties of Realism in International Politics

Realism is a broad approach to conceptualizing international politics. As with any broad approach, it has intellectual antagonists, internal disagreements generating subtle distinctions among types of realism, and experiences recurrent efforts at reformulation.

Realism is a tenet that views the world as it is, and not in terms of an unrealized ideal. It emphasizes the constraints of the real world and the need for pragmatism and prudence. Applied to international politics, realism sees a conflict-ridden world of states concerned preeminently with their security, struggling for advantage, sometimes seeking dominance, and therefore pursuing power as the means to assure these. Realists therefore see the distribution of power as the major determinant of state behavior and international stability. Within this broad rubric, realists have developed alternative formulations, emphases, and distinctions, and others have appropriated aspects of realism and conjoined it with a plethora of other perspectives.

The Realist Vision

The realist vision of international politics holds that states interact in an anarchical international environment in which survival is uncertain. Because power is the key requisite for survival, states define their interests in terms of power and constantly assess their strength relative to that of others. What emerges from their interaction is a balance of power that is the sole basis for order, regularity, and stability in international politics.

Realism emphasizes the competitive and conflictual nature of international politics. States either constantly prepare for war or are actually at war with one another. Fearing adverse shifts in the balance of power, they respond with balancing behavior, either through internal mobilization and, should that prove inadequate, with external alignment. Self-reliance is preferred in such a system of self-help, but allying with others to confront a common threat may be temporarily necessary. Cooperation is thus short-lived, lasting only as long as the common threat, and is a function of common interests and not common values or ideology. Realism can therefore readily explain alliances between wholly antithetical political systems, such as the Nazi–Soviet pact of 1939. The mobilization and alignment that result in response to adverse shifts in the balance of power return the system to a state of balance. Hence, balances of power recur and reemerge in international politics.

Realist Assumptions

Realism is a statist perspective, for it assumes states to be the key actors in the international system. It excludes multinational corporations, nongovernmental organizations, and a variety of transnational and subnational entities that matter only to the extent that they affect states.

The theory assumes that states face recurrent threats and challenges in the international system. Power, especially military power, is the basis of both threats to a state and to its ability to assure its survival. Although some acknowledge a variety of forms of power, the overwhelming focus is on military capabilities, their distribution and those material resources that make them possible.

Realism presumes survival and its requisites to be the highest priority of states, indeed, their very reason for being (*raison d'etat*). States value security considerations above all and respond to the imperatives of the international system. The primacy of security implies that other policies, including foreign economic ones, reflect considerations of international power politics. Other state objectives, such as ideological ones, are inherently secondary. They are only pursued when security is assured, and they are readily abandoned when security is threatened.

Realism Deemphasizes

For realists, international law and international organizations are largely irrelevant to international politics. Both reflect the interests of powerful states, for in international politics might makes right, and international institutions do not constrain great powers but merely provide additional venues for the exercise of power. At most, both play a role only in matters of low politics, in regulating transportation, communication, and economic exchange, for example, and not in matters of high politics, especially military security.

Realists also consider the requisites of survival to dominate any moral considerations. The requisites of survival are more important than any ethical considerations. Indeed, realists transform necessity into a moral imperative.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93119-2

Realism downplays the role of domestic politics. It treats states as unitary actors that are functionally equivalent to one another, and the constraints imposed by the international system generate similar behavior even by dissimilar states. The international system once consisted largely of monarchies and then of authoritarian and military governments, and is now increasingly composed of democratic states. For realists, however, it continues to function in the same way. Differences in the internal composition of states do not matter, except if they affect relative power.

Self-reliance precludes states from allowing the development of any division of labor or any interdependence regarding the requisites of state survival. Small states may have no choice but to rely on others, but great powers do have such choices and thus do not allow the emergence of any meaningful interdependence.

Realism as Equilibrium Theory

The balance of power is an equilibrium theory. The security policies of nations result in a balance of power among states. The equilibrium can be disturbed, but only by exogenous shocks, ones outside the model. The disturbance then sets in motion forces that return the system to a new equilibrium. The implosion of the Soviet Union, for example, disturbed the balance of international power, but was an event outside the model and inexplicable by it. But realism does purport to explain the set of forces that then would come into play to return the system to equilibrium. For some, a balance of power is something states seek; for others, it is an unintended result.

Put this way, balance of power theory functions like Adam Smith's invisible hand in creating an unintended outcome. For Smith, the self-interested behavior of producers in a competitive market generated the best price for consumers. In international politics, the pursuit of national interest and the struggle for power in an anarchical system create a peaceful stable order. It is not the search for peace but the search for security that creates this stability. The preparation for war assures peace.

Realism's Historical Roots

Realism traces its lineage to Thucydides, who argued that his *History of the Peloponnesian War* (1996) provided the one and true explanation of the war: the growth of Athenian power and the fear this struck in Sparta. Thucydides argued that preceding incidents and specific complaints notwithstanding, the true cause was a shift in the balance of power.

As important to current realist formulations is the work of Thomas Hobbes, who translated Thucydides into English and argued (in *Leviathan*) that the state of nature, which he characterized as a war of all against all, is one in which war and preparation for war are a constant. Conflict is limited within societies by the creation of the Leviathan, a coercive state. For modern realists, who distinguish international from domestic politics, international relations remains an anarchic realm; there is no Leviathan and the war of all against all endures.

Realism dominated the study of international relations in the middle of the twentieth century as the world came to grips with the occurrence of the Second World War and the failure of both international law and international organizations to prevent it. Hans Morgenthau's (1948) influential formulation of realism drew upon psychology and rooted realism in a human nature presumed to be inherently aggressive and in a view of individuals as power seeking.

Beginning in the 1950s and increasingly in the 1970s, an alternative realist version emerged. It coincided with the rise of (and often appropriated the terms of) systems thinking in the 1950s and structuralism in the 1970s. This variant rejected basing realism in any theory of human nature or any internalized commitment to a balance of power. This neorealist formulation did not posit any individual or collective propensity toward power maximization per se. Neorealism was a purely systemic theory, one that focused on the characteristics of the international system structure, rather than the nature of individuals or states. Rather, it posited a minimalist concern with survival in an anarchic setting as the only necessary conditions. The anarchic system, the absence of hierarchy assured a struggle for survival and an emphasis on power. All states caught in an anarchic structure have no choice but to arm themselves in preparation for the possibility of war and must assiduously guard their relative power. Neorealism or structural realism firmly rooted the problem of international politics in the international system and not in the nature of individuals or states.

Realism has a core set of assumptions and relaxing any one opens the door to an alternative approach. Neorealists maintain a pristine structural approach that abjures discussion of foreign policy. They provide a theory of international politics and not one of foreign policy. Others have returned to what they describe as a more classical perspective and developed a neoclassical realism that aspires to provide a theory of foreign policy, maintaining the international system as a key independent variable but including domestic constraints and elite perceptions as important intervening variables.

The Realist View of History

For realists, history consists of the rise and fall of great powers and of changes in the distribution of power in the international system. Much like economic structures are characterized by the distribution of firms (monopoly, duopoly, oligopoly, and a competitive market), the international system is characterized by the distribution of power, and more specifically the number of great powers. Multipolarity is a system of three or more powers, bipolarity is a system of two great powers, and unipolarity (or hegemony) is a system with one dominant power. Changes in the distribution of power constitute major break points in history and determine the structure of the international system. The multipolarity of the nineteenth and early twentieth century is contrasted with the bipolarity of the Cold War and contrasted with the unipolarity of the post-Cold War world. Historical periods are characterized by the distribution of power in the system, and the challenge of international politics in any era is that of dealing with changes in the relative power of great powers. The end of the nineteenth and beginning of the twentieth century had to deal with the rise of German power, and the current age is dominated by a focus on the rise of Chinese power.

This characterization of the international politics of the last century contrasts with those that emphasize the role of ideological competition and conflict. Diplomatic histories of the last century invariably focus on ideological conflict. They note the conflict between liberalism, fascism, and communism as the dominant feature of both domestic and international politics. The emergence of a Communist Russia is seen as a central feature shaping international politics from 1917 to 1989. Nazism and fascism are deemed to have posed a challenge to international order in the 1930s and 1940s. The Cold War that followed the end of World War II is not simply a case of bipolarity but to be understood as a conflict between Communism and the liberal west. Finally, the collapse of the Soviet Union is not merely the collapse of one of the two superpowers in a bipolar world but the end of the Communist challenge and the triumph of liberalism. For realists, however, ideological conflict was merely epiphenomenal to changes in the set of great powers and the distribution of power in the international system.

Disagreements and Debates

Despite the consensus in emphasizing states as the central actors, security as their prime concern, and power as their key instrumentality, there is some confusion and disagreement in the core of realist thought. First, there is disagreement about the level of analysis of realist thought and the nature of both its dependent and independent variables. For neorealists especially, realism is a purely systemic argument. It merely explains the recurrence of balances of power in the international system. It is a theory of international politics and not of foreign policy. It cannot explain the foreign policy of any particular state but merely the tendency of the system to arrive at a balance of power. For others, harking back to a classical realism, and dubbing themselves neoclassical or neotraditional realists, the emphases on power and interest imply particular foreign policies in particular circumstances.

A second related source of confusion and disagreement is whether the theory is positive or normative. Realists typically argue that theirs is a positive theory of how the system actually works (an iron law of international politics), yet many (including structural realists) have also used realist precepts to criticize the foreign policies of specific countries. Realists, for example, criticized US military involvement in the Vietnam War in the 1960s and in the Iraq War of 2003. Yet, if realism can be used to criticize extant policy, then it is inadequate as an explanation of that policy. Moreover, if it provides no theory of foreign policy then it should hardly be used as a basis for criticizing a particular foreign policy.

That politicians adopted the argument also led to disagreement. First, the term balance of power has come to be used loosely and applied to virtually every configuration of power. Second, there is a disagreement between scholars as to whether realism operates through intention (states must aim for a balance of power) or irrespective of it (as an unintended consequence).

Disagreement about the nature of the dependent variable (does realism explain foreign policy or just international politics) is matched by disagreement on the independent variable side as well. Although all scholars emphasize the role of power and the distribution of power, there are those (neoclassical realists) who see the impact of power as mediated by domestic politics. Thus they do not argue that dissimilar states will always respond in the same way to a particular distribution of power.

Finally, there is even a little disagreement about the core proposition of balance of power theory, that the workings of the system mean that balances recur and are the basis of international stability. First, there are different balanced configurations of power and there is some disagreement about the relative stability of different configurations of power, about whether bipolar or multipolar international systems are more stable. Most scholars argue that bipolarity is more stable because states readily respond to adverse shifts in the balance of power. In contrast, in multipolar systems, states will pass the buck in the hope that others will deal with adverse changes in the balance. This delays the return to balance and stability. Multipolar systems are thus thought to be more war-prone and unstable. By and large this is the consensus, though there is some disagreement on this point.

Most realists, as balance of power theorists, agree that the international system tends toward balance (though more predictably and more quickly in some settings than others) and that hegemony and unipolarity are therefore not possible or sustainable. Any state aiming for global domination generates countervailing coalitions of power and thus attempts to achieve hegemony have historically failed. Yet, there are those scholars who hold that the system tends toward power preponderance rather than balance, and that hegemony is stable but that a balance of power is not. This has played out in the last two decades in competing predictions of the post-Cold War world. Following the collapse of the Soviet Union and the end of the bipolar order, balance of power realists argued that US hegemony, or global unipolarity, would be short-lived. Coalitions that had been built to contain Soviet power would come apart once the threat that had been the basis for their existence disappeared. Moreover, the US would succumb to imperial temptation, and even if not, others would balance against US power and multipolarity would soon emerge. In contrast, others have argued that unipolarity is stable and sustainable, either because they believe that preponderance is generically stable or because they believe that US hegemony differs from prior attempts to achieve global domination.

A sustained debate has emerged within realism between offensive and defensive realists. The former argue that anarchy and the concern with security, and therefore, power result in states that are power maximizers. All states are revisionists, assertive expansionists seeking to maximize their power relative to others. Like classical realists, they argue that states are power maximizers. But they do not root this in any individual will to power, but in the requisites of survival in an anarchic world. In contrast, defensive realists emphasize that security seeking leads states to respond to adverse shifts in the balance of power but also leads them to forgo seeking advantage over others. Security-seeking states recognize the existence of a security dilemma, in which a state's search for military superiority will result in comparable reactions by others, leaving the state no better off and perhaps worse off. The result is a world in which

status quo states predominate and revisionist states are rare. Defensive realists see a world less conflict-ridden than offensive realists.

Problems

Realism is criticized for leaving too much out. A pristine realism that only says that balances recur cannot explain most foreign policies of most states most of the time. Moreover, the theory is static and cannot explain change over time. Important transformations of recent centuries, such as the rise of nationalism, the industrial revolution, the rise of ideological politics, the spread of democracy, the revolutions in communication, transportation, and warfare, realists judge irrelevant to international politics, which they treat as unchanged as long as anarchy continues to characterize international politics. Realists deem these factors to matter only to the extent that they affect the relative power of states (whose alteration is the only source of change in international relations) and are viewed as exogenous shocks. The only transformation that some realists accept as affecting world politics is the nuclear revolution, which they consider to undercut fundamentally the importance of relative power in a classical military sense.

The theory is also vague and imprecise. It is not clear what factors enter into a calculation of the balance of power, nor how balance is ascertained. It is difficult to test assertions of the theory when there can be wide disagreement over whether a particular distribution of power is balanced or not. Scholars dispute whether stability during the Cold War derived from US military superiority, a balance of power between the US and USSR, or the existence of nuclear weapons. Scholars disagree whether past wars occurred as balance or imbalance was emerging.

More important, the theory is indeterminate. It cannot predict which particular distribution of power will emerge, only that balance will. Nor can it predict the process by which balance is restored. US containment of the Soviet Union during the Cold War can be explained as the working of the balance of power, but so can the US roll back of German and Japanese power during World War II. A US decision simply to contain German and Japanese power in the 1940s would have been as consistent with realism as the actual US decision to wage war and undo the consequences of German and Japanese aggression. The theory cannot explain when war is used to restore the balance of power and when containment restores the balance.

The theory makes precise predictions only when the constraints of the international system are presumed to be tight. Although the theory holds that it is great powers that structure global politics, the international system is most constraining of small powers and least constraining of large ones. Hence, the theory is typically indeterminate regarding the behavior of the great powers, which are presumed to matter most for international politics.

Formal studies have shown that realist premises are consistent with international cooperation, the construction of international organizations, and even the emergence of international norms. These works challenge the deductive logic of realism, arguing that the premises are insufficient to lead to realist conclusions.

Realists have had some difficulty in recent years in explaining significant international developments, as, for example, the disjuncture between military and economic power. Realists argue that economic power is invariably translated into military power relatively quickly. Yet the global disjuncture between the two has been sustained for an extended period of time. During the later stages of the Cold War, the world remained militarily bipolar, dominated by the might of the United States and the Soviet Union. Yet economically the world was multipolar, with Germany (and the EU more broadly) and Japan as significant economic powers that did not translate their economic prowess into military strength. Since the end of the Cold War, the world has been militarily unipolar (US defense spending being greater than the combined spending of the next 15 powers for much of the period) and yet the global economy was distinctly multipolar. Such sustained disjunctures between the distribution of economic and military power is problematic for realism.

As noted earlier, most realists assumed that the unipolarity of the post-Cold War world would be a short-lived phenomenon, that balancing behavior would react to imperious US power. Yet multipolarity has not emerged and most of the other rich and powerful states in the system remain aligned with the US. Some have argued that opposition to US policy by its allies constitutes a new form of soft balancing behavior, thus both conforming to realist expectations and yet constituting a new and distinct phenomenon not adequately conceptualized in realist thought.

Realists also have problems explaining the emergence of a zone of peace in central and western Europe. War is now unthinkable between states that have fought many wars over centuries. Despite remaining sovereign and autonomous, the states of Europe are no longer engaged in a security competition, and their peaceful relationships transcend the mere temporary cooperation envisaged by realism.

Modifications and Varieties of Realism

Realist thought continues to develop, in part to deal with such anomalies. One variant now posits that it is the perception of power rather than the actual distribution of power that matters. Another decouples changes in the distribution of power from the perception of threat, and holds that states respond to threats and not just shifts in power.

Some see realism as characterizing particular eras rather than international politics generally. Eras in which territorial acquisition brings greater rewards than does commercial exchange are ones to which the logic of realism is more applicable. In contrast, eras in which territorial expansion provides little wealth and great cost and in which commercial exchange is the basis for wealth and power are eras in which a realist logic is less applicable. Eras in which the nature of warfare gives the offense an edge are ones in which a realist logic is more likely to prevail, eras in which defensive weapons are superior are more stable and peaceful and not characterized by security competitions.

Realism has a core point with which no international relations scholars will argue: the international system consists of states that are sovereign and independent and there is no power

to enforce norms or law and that therefore power is a central consideration in analyzing international politics. No scholar denies the importance of states, power, and interests. Scholars of all stripes give this central core of realism its due, but then depart from other elements. The resulting desire both to amend realism and to appropriate it for important modifications has resulted in a variety of realisms. Although liberalism (and liberal institutionalism) have been proffered as conceptual alternatives, there are those suggesting a hybrid of democratic realism or liberal realism. Similarly, constructivism, critical theory, and cultural perspectives have been proffered as alternatives to realism, and here too amalgams have been proposed with the labels cultural realism, critical realism, and constructivist realism. Finally, realism is seen as quintessentially immoral, a cynical view of international life that makes war and military necessity an ethical virtue. Yet, here too, some have proposed arguments they have dubbed ethical realism and moral realism.

These varieties of realism (which many realists would deny as realist arguments) attest to the continuing centrality of power and interests in any discussion of foreign policy and international politics. Yet the resulting transmogrifications have raised the question, "is anyone still a realist?" (Legro and Moravcsik, 1999).

Realism has a core strength: its emphasis on power and interest and an unregulated international system. Yet its most pristine formulation is rejected by many, if not most, scholars who nevertheless articulate their positions relative to a core realist argument. Realism is akin to Greenwich Mean Time: more live outside than inside that time zone, but everyone measures themselves by their distance from it.

See also: Balance of Power, History of; Military, War, and Politics; Nation-State and War; National Security Studies and War Potential of Nations; Wars among Nation-States: Patterns and Causes.

Bibliography

Baldwin, D.A. (Ed.), 1993. Neorealism and Neoliberalism: The Contemporary Debate. Columbia University Press, New York.

Bell, D. (Ed.), 2008. Political Thought and International Relations: Variations on a Realist Theme. Oxford University Press, New York.

Booth, K. (Ed.), 2011. Realism and World Politics. Routledge, New York.

Brown, M.E., Lynn-Jones, S.M., Miller, S.E. (Eds.), 1995. The Perils of Anarchy: Contemporary Realism and International Security. MIT Press, Cambridge, MA.

Buzan, B., Jones, C.A., Little, R., 1993. The Logic of Anarchy: Neorealism to Structural Realism. Columbia University Press, New York.

Carr, E.H., 1940. The Twenty Years' Crisis, 1919–1939: An Introduction to the Study of International Relations. Macmillan, London.

Claude, I.L., 1962. Power and International Relations. Random House, New York.

Craig, C., 2003. Glimmer of a New Leviathan: Total War in the Realism of Niebuhr, Morgenthau, and Waltz. Columbia University Press, New York.

Donnelly, J., 2000. Realism and International Relations. Cambridge University Press, New York.

Doyle, M.W., 1997. Ways of War and Peace: Realism, Liberalism, and Socialism. Norton, New York.

Frankel, B. (Ed.), 1996. Realism: Restatements and Renewal. Frank Cass, London.

Gulick, E.V., 1955. Europe's Classical Balance of Power: A Case History of the Theory and Practice of One of the Great Concepts of European Statecraft. Cornell University Press, Ithaca.

Haslam, J., 2002. No Virtue Like Necessity: Realist Thought in International Relations Since Machiavelli. Yale University Press, New Haven.

Jervis, R., 1978. Cooperation under the security dilemma. World Politics 30, 167–214.

Jervis, R., 1997. System Effects: Complexity in Political and Social Life. Princeton University Press, Princeton.

Kaplan, M.A., 1957. System and Process in International Politics. Wiley, New York.

Keohane, R.O. (Ed.), 1986. Neorealism and Its Critics. Columbia University Press, New York.

Legro, J.W., Moravcsik, A., 1999. Is anybody still a realist? International Security 24 (2), 5–55.

Little, R., 2007. The Balance of Power in International Relations: Metaphors, Myths, and Models. Cambridge University Press, New York.

Lobell, S.E., Ripsman, N.M., Taliaferro, J.W. (Eds.), 2009. Neoclassical Realism, the State, and Foreign Policy. Cambridge University Press, New York.

May, E.R., Rosecrance, R., Steiner, Z. (Eds.), 2010. History and Neorealism. Cambridge University Press, New York.

Morgenthau, H.J., 1948. Politics Among Nations: The Struggle for Power and Peace. Alfred A. Knopf, New York.

Stein, A.A., 1990. Why Nations Cooperate: Circumstance and Choice in International Relations. Cornell University Press, Ithaca.

Thucydides, 1996 (R. Crawley, Trans.). In: Strassler, R.B. (Ed.), The Landmark Thucydides: A Comprehensive Guide to the Peloponnesian War. Free Press, New York.

Trachtenberg, M., 2003. The question of realism: a historian's view. Security Studies 13, 156–194.

Vasquez, J.A., 1998. The Power of Power Politics: From Classical Realism to Neotraditionalism. Cambridge University Press, New York.

Waltz, K.N., 1959. Man, the State, and War: A Theoretical Analysis. Columbia University Press, New York.

Waltz, K.N., 1979. Theory of International Politics. Addison-Wesley Publishing Company, Menlo Park.

Waltz, K.N., 2008. Realism and International Politics. Routledge, New York.

Williams, M.C., 2005. The Realist Tradition and the Limits of International Relations. Cambridge University Press, New York.

Recidivism

William D Bales, Kaleena J Burkes, Samuel JA Scaggs, and Catie L Clark, College of Criminology and Criminal Justice, The Florida State University, Tallahassee, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With our correctional facilities suffering from overcrowding, recidivism research has become of extreme importance to policy makers and social scientists. The purpose of this article is to view recidivism from the perspective of unaccounted for populations. Typically, research that studies recidivism looks at certain outcomes (reimprisonment, rearrest, and reconviction) for offenders released from prison. However, the authors suggest that within the population of individuals able to recidivate, failing to include jail and community correction populations leaves out a large portion of the picture. Additionally, the authors discuss risk assessment tools as a means of predicting recidivism among supervised and incarcerated populations.

Sociologists, and particularly criminologists, are often concerned with the behaviors of individuals in society that fall outside of the established social norms. The study of deviance and offending is a vast field, enduring over several centuries and encompassing many disciplines. The nature of crime and offending itself is frequently discussed and hypothesized in the form of a plurality: once individuals begin to engage in crime, it is typically expected that they will continue to engage in criminal acts. The question for many theorists is not why some individuals engage in crime, but why some do not, or, alternatively, why some individuals ultimately desist from offending, while others persist. In this sense, the study of crime is also the study of recidivism.

Recidivism is commonly defined as the continuation of offending following official sanctioning by the criminal justice system (Beck, 2001; Blumstein and Larson, 1971; Maltz, 2001; Webb et al., 1976). Individuals who have been arrested, convicted, incarcerated, and/or placed on a form of community supervision who then continue to engage in illegal behaviors that result in a new arrest, conviction, period of incarceration in jail or prison, term of community supervision, and/or modification of a current term of incarceration or community supervision are said to have recidivated or reoffended. Recidivism is partly a continuation of illegal behaviors and partly the recognition and sanctioning of said illegal behaviors by the criminal justice system. One of the key roles of criminal justice practitioners and policy makers is to reduce recidivism. Actuarial risk assessment tools, which are used at multiple points in the criminal justice system, are used to determine the likelihood that an individual with a particular criminal history and miscellaneous characteristics will commit another criminal act. In a sense, these practitioners and policy makers are in the business of curtailing multiple forms of recidivism at multiple stages in the criminal justice process.

Despite the need to assess risk for reoffending in a variety of ways, the criminological literature has primarily focused on the recidivism of persons released from incarceration and has conceptualized recidivism as rearrest, reconviction, and reimprisonment. By only focusing on a limited population (released inmates) and limited measures of recidivism (rearrest, reconviction, and reimprisonment), only a small part of the recidivism picture is being captured. Ultimately, the prison population is a relatively small part of the criminal justice system compared to individuals incarcerated in local jails or placed on a variety of types of supervision in the community. Discounting the importance that the community corrections and jail populations play in contributing to the overall recidivism rate of offenders in the United States is excluding the majority of the reoffending population. These frequently overlooked and understudied populations, as well as the accuracy of the extant literature in capturing the true nature of recidivism, are focused on in this discussion of recidivism. An understanding of these other recidivism measures permits a better knowledge base about potential contributors and responses to multiple forms of recidivism.

To this end, the authors first describe the types of measures that are generally classified as recidivism as well as community- and jail-based sanctions that are not commonly included. Second, they summarize the correctional context to illustrate the abundance of people being sanctioned to community-based alternatives to incarceration and local jails. Third, they further contend that including nontraditional community-based measures of recidivism would further strengthen the validity of commonly used risk assessment tools.

Probation, Parole, and Jail Populations in the United States

The reality is that there are far more adults serving time under community supervision and in local jails than in state and federal prisons. The Bureau of Justice Statistics in 2010 indicates that 4 887 900 offenders were under some form of community supervision. In the fiscal year 2010, 2074 individuals per 100 000 (or 1 in every in 48 people) in the United States were under some form of community supervision. Based on specific breakdowns, 1721 per 100 000 (or 1 in every 58 people) were on probation in the fiscal year 2010. The overall number of adults on probation has reached 4 055 500, representing nearly 83% of the total supervised population. Alternatively, 357 per 100 000 (or 1 in every 280 people) were on parole in the fiscal year 2010, which represents a total number of 840 700 adults on parole (Glaze and Bonczar, 2011). Jails also house a substantial number of inmates

relative to prisons. According to the Bureau of Justice Statistics' annual survey of jails at midyear 2010, 748 728 inmates were confined in jails, which represents an incarceration rate of 242 per 100 000 persons. This figure is almost equivalent to the average daily population of 748 553 inmates reported by Minton (2011). In contrast, 1 612 395 (500 inmates per 100 000 residents) adults were being in held in state ($N = 1\,402\,624$) and federal ($N = 209\,771$) prisons in 2010 (Guerino et al., 2011). These figures suggest that the virtual absence of recidivism measures related to offenders under community supervision or those released from local jails results in the measurement of a relatively small portion of offenders under correctional control.

Recidivism within a Community Context

How recidivism is measured is largely contingent on the definition used for this critical outcome of sentencing and correctional strategies. While recidivism is frequently operationalized as reconviction, rearrest, and reimprisonment, there are other types of sanctions indicative of recidivism captured within the community context. These sanctions may include technical violations, new arrests, probation and parole revocation due to technical violations or new arrests, and absconding.

Technical violations are generally defined as infractions of noncriminal conditions of supervision (i.e., rule violations), while law violations are the commission of a new crime while on supervision. The process of revocation involves a modification to the individual's prior supervision conditions, which may result in an increase in supervision conditions or an outright elimination of probation or parole as well as a new sentence to jail or prison. Both the commission of a new crime and one or more technical violations may result in the revocation of probation or parole. A probation or parole term may also be revoked for drug-related violations, such as testing positive for drug use, possessing drugs, failing to report for drug testing, or failing to complete alcohol or drug treatment (Gray et al., 2001; Hughes et al., 2001).

Offenders abscond when they leave the jurisdiction in which they are being supervised without receiving prior permission. The reasons for this action are varied, but some assertions are that offenders may be fearful, distrustful of the system, or uncertain about the outcome of their situation. Despite the possibility that offenders on community supervision receive a warning, it is highly unlikely that these informal sanctions will be reliably recorded, if reported at all. Notwithstanding warnings, each type of sanction may be used to represent the act of recidivism within the community.

The Overlooked Local Jail Populations and Recidivists

Over the past several decades, in response to what has been termed 'mass incarceration,' criminological research generally, and recidivism research specifically, has been focused on offender outcomes of state and federal prisoners. Largely absent from the recidivism literature have been studies concerned with the approximately 12.9 million offenders released annually from jails in the United States (Miller and Miller, 2010). Jails are essentially the initial institutional experience that virtually every suspected or convicted offender experiences. Except for individuals suspected of very minor offenses that result in notices to appear in court, civil citations, etc., all arrestees who ultimately are convicted and sentenced to a community-based supervision sanction or to state or federal prison begin with some length of stay in a local jail facility.

The argument can be made that recidivism research devoted to determining what program interventions and correctional policies are most effective in reducing reoffending should be focused more on jail populations than any other form of correctional control for the following reasons. First, the vast majority of offenders sentenced to prison have been through local jails numerous times. Langan and Levin (2002) examined the recidivism rates of the 272 111 inmates released from the prisons systems in 15 states in the United States in 1994 and reported that 93.1% had been arrested prior to the arrest that resulted in their incarceration and averaged 8.8 prior arrests. Second, the opportunity to possibly change the criminal trajectory of individuals arrested, regardless of whether they are ultimately convicted and sentenced to some form of correctional sanction, requires some intervention at the local jail level. The dearth of research examining the correlates of offending subsequent to offenders' initial involvement in the jail system or after repeated stays in jail results in the inability to understand what programs or interventions can be used to reduce the likelihood of further immersion in the correctional system.

The Importance of Research on Jails

As the gatekeeper to the American correctional system, US jails house a significant portion of the total inmate population. For instance, 9% of jails house 48% of all inmates within any state correctional system (Stephen, 2001; Hughes, 2006). Each year, a flow of over 12 million individuals are admitted and released from jails and an average of 766 010 inmates are incarcerated on any given day. Currently, there are 3365 independently and locally administered jail facilities in the country, compared to 50 state prison systems (Beck, 2006). While state and federal prisons incarcerate convicted felons, US jails serve a broader function. First, jails process and house individuals awaiting arraignment, trial, conviction, and sentencing. Additionally, jails house convicted individuals whose sentence is generally less than a year. They may also hold individuals who violate their pretrial release, probation, or parole conditions. In order to alleviate overcrowding in prisons, jails also temporarily hold state and federal prisoners. Prior to their arrival to prisons or when they are transitioning between prisons, state and federal prisoners may temporarily spend time in jail facilities. Aside from the people serving time in jails, individuals are also supervised outside of jail facilities. In the United States in 2010, over 7% of individuals were under jail supervision in the community (excluding persons under probation and parole) (Minton, 2011). These community-based alternatives include weekend programs, electronic monitoring (EM), home detention, day reporting, community service, and treatment programs (Minton, 2011).

Jails present unique challenges for addressing the reentry needs of offenders. Jailed inmates' short incarceration periods along with the low service capacities of jails limits the opportunity for treatment and other services. According to Harlow (2003), only 14% of jail inmates, compared to 52% of state prisoners, participate in education programs. Another challenge is that the size, operation, resources, and populations of the 3365 independently operated jails may not be uniform, which has important implications for the supervision, management, and treatment of inmates. Finally, there is no community-based system to aid in the reentry process (Solomon et al., 2008). Nonetheless, despite the clear need to research jails as well as the problems that exist for jail reentry, there is, as mentioned previously, a surprising lack of empirical research in this area generally (Miller and Miller, 2010) and recidivism specifically.

Only recently have researchers (Miller and Miller, 2010; Osher, 2007; Osher et al., 2002; Solomon et al., 2008) considered the reentry issues, outcomes, and programmatic needs of postrelease jail inmates. Nonetheless, and consistent across various settings and offender populations are demographic factors (i.e., age, race, sex, marital status), socioeconomic characteristics, drug abuse histories, limited education histories, prior offense histories, and poor community support networks, which are well-documented predictors of recidivism (Benedict and Huff-Corzine, 1997; Benedict et al., 1998; Kubrin and Stewart, 2006; Petersilia, 1995; Ulmer, 2001). Another predictor of reentry success is whether jail inmates can find employment and secure housing upon their release. Depending on their offense, offenders may be further marginalized in both areas. As previously discussed, 6 of 10 jail inmates have not yet been convicted and are awaiting arraignment (Sabol et al., 2007). Examining this population of pretrial jail release offenders allows for a more differentiated evaluation of the predictors of reentry outcomes among a specific jail inmate population.

While the empirical literature on jail-based reentry programs remains in its infancy, available studies that have provided insights on this important area of corrections have generally outlined principles for effective intervention (Osher, 2007) and provided profiles of several currently implemented jail-based reentry programs within several states (Soloman et al., 2008). Unfortunately, few studies have evaluated in a methodologically rigorous manner the effect of life skills and substance abuse correctional programming on reducing recidivism. In a review of the literature on evidence-based correctional programs and interventions, MacKenzie (2006) concludes that life skills training programs are ineffective interventions for reducing recidivism because they focus on changing opportunities in the environment rather than the individual. In contrast, MacKenzie (2006) finds that incarceration-based drug treatment programs are effective interventions because they address the individual's criminogenic needs. However, with few exceptions (Turley et al., 2004), research has not yet specifically examined the impact of jail-based life skills and substance abuse treatment programs on both pretrial and sentenced offender outcomes upon release.

Only a few empirical studies have specifically tested for the effect of EM on jail inmates' outcomes during the pretrial stage. In one study, it was found that the use of EM, compared to the absence of EM, is associated with significantly higher rates of technical violations (Cooprider and Kerby, 1990). Another study evaluating pretrial jail release compared offenders on EM with offenders not on EM at the national ($n = 22\ 725$) and the federal district ($n = 7234$) levels (Cadigan, 1991). While there were no statistical controls in the analysis, the author concluded that EM offenders are higher risk defendants than non-EM offenders. Cadigan's outcome measures included failure to appear and felony and misdemeanor rearrest rates. The study's results indicate that failure to appear rates, felony rearrest rates, and misdemeanor rearrest rates for EM offenders are slightly higher than those for non-EM offenders at both the district and national levels. The study did not incorporate relevant control variables or report offense types. An area where EM has not been evaluated in previous studies involves offenders who must serve a period on EM after release from jail.

In summary, the evidence is clear that the dearth of recidivism research on the over 12 million individuals who exit local jails throughout the United States each year is a significant void in the criminological research literature. There have been major advancements over the recent past in the research conducted on offenders exiting state and federal prisons in the United States and our understanding of the correlates of recidivism and the relative effectiveness of various correctional strategies in these types of facilities is improving reentry outcomes. Devoting more effort and resources in the future to understanding the factors that are linked to higher rates of reoffending and reincarceration among local jail releases will advance the recidivism research agenda and provide policy makers and practitioners with invaluable information to reduce future jail admissions and populations.

Risk Assessment Tools

One of the most pragmatic methods for identifying whether someone recidivates is through risk assessment instruments. These tools are used by various correctional officers and staff to predict the likelihood of prison misconduct (Cunningham and Sorensen, 2006), violence, psychopathology (Hare, 2003; Hare et al., 1990), and recidivism upon release from prison. One issue with these instruments is that they only determine whether a range of social, psychological, cognitive, and biological factors predict more common forms of recidivism.

As a result of the ideological backlash in the 1980s and 1990s toward rehabilitative ideals and a reduction in resources allocated to probationary agencies, several correctional and community supervision agencies adopted risk management and avoidance tools to reduce the likelihood of reoffending. This new risk management system is geared toward maximizing the regulation and control of offenders as well as toward enhancing public safety. In what some scholars (Simon and Feeley, 2003) have called 'the new penology,' this approach views crime as a systemic issue that should be managed through actuarial instruments. While the initial goal was to assess the risk of individuals to reduce reoffending, risk management techniques quickly expanded into the realm of treatment services as well. Now risk assessment is one of the most common ways of statistically predicting the likelihood of

recidivating given the characteristics of the offender and the situation (Bonta, 2002).

The normal progression of classifying offenders within the community includes assessing the individual's propensity to engage in violent or criminal behavior, identifying the appropriate interventions given their personal characteristics, and selecting suitable supervision and treatment approaches (Lowenkamp et al., 2006). Although early subjective clinical diagnoses were most frequently implemented to classify offenders, recent empirical research suggests that objective actuarial tools may be the more reliable and efficient options relative to clinical assessments, if administered by trained staff (Bonta et al., 1998; Grove and Meehl, 1996).

The risk–need–responsivity (RNR) model represents the foundation for several instruments that assess and match offenders with corresponding treatment or programmatic needs. The 'risk' principle argues that individual be placed within a category associated with their propensity to engage in violent or criminal behavior. For instance, an individual may be assigned to a low-, medium-, or high-risk classification. According to the 'needs' principle, a criminal justice agent will assess and report the existence and magnitude of an offender's problem behaviors. Due to important considerations pertaining to the offender's amenability to treatment, the 'responsivity' principle examines individual characteristics that may hinder or augment their success from treatment (Van Voorhis et al., 2007). The behavioral philosophy of this model mandates the receipt of rewards and punishments to optimize success. Regardless of whether the focus of the program is supervision oriented or treatment oriented, the most intensive programs should be reserved for those offenders with the highest risk classification (Hanley, 2002). This model also adopts cognitive behavioral and 'social learning' elements and relies on the reinforcement of family and peers to instill positive and prosocial attitudes (Andrews and Bonta, 1998). RNR techniques have garnered some support as an effective approach to reducing recidivism in the community.

Objective instruments predict the statistical likelihood of reoffending given information about the offender. The most effective of these instruments examine both static and dynamic factors. While static factors are those characteristics of the individual that cannot be altered (i.e., age at first offense, prior convictions), dynamic factors or criminogenic needs are variables that can change over time (i.e., drug and alcohol abuse, family and peer relationships, anger management). The latter risk factors are better able to target both positive and negative individual factors that are apt to change over time. Moreover, dynamic risk factors are referred to as criminogenic needs because they represent variables that can be targeted with treatment (Bonta, 2002). A reduction in these needs has been shown to result in lower levels of recidivism (Andrews and Bonta, 1998; Andrews et al., 1990).

As a result of this new risk management approach, a host of risk assessment tools have emerged. One of the most common risk–needs assessment tools is the Level of Service Inventory-Revised, which examines information on criminal history, education, employment, alcohol and drug use, companions, and emotional and personal state (see Andrews and Bonta, 1995). Based on the risk score produced by this 54-item scale through an officer-led interview process, offenders are assessed based on their likelihood of recidivating (Lowenkamp and Bechtel, 2007). Currently, this tool is one of the most theoretically guided assessment instruments used on offender populations (Bonta, 2002) with empirically established predictive validity (Andrews and Bonta, 1995, 1998; Gendreau et al., 2002).

Importantly, these tools are used for assessing individuals exiting prison as well as those offenders who are sanctioned within the community (since the focus in the current article is on community-based measures of recidivism, the authors do not describe those risk assessment tools that measure misconduct (as an indicator of recidivism) within prison). Given that the majority of offenders receive community-based sanctions as an alternative to prison terms, there is a clear need for offenders to be assessed upon receiving a term of supervision. One of the guidelines for implementing any risk assessment tools is achieving optimal predictive validity, which is accomplished through assessing their ability to predict criminal offending (Bonta, 2002). Assuming that these instruments are used to predict offending within the community, what would be considered the optimal measureable outcomes, rearrest, reconviction, or reimprisonment? Should the predictive validity of risk assessment tools not also include other forms of infractions such as technical violations, law violations, and probation and parole revocation? Even if the objective of correctional practitioners is to reduce all forms of infractions, it would be important to examine these other types of recidivism. Additionally, if the objective is to target specific types of infractions, risk assessment tools that measure even arguably minor community-based outcomes could be used to modify the intervention (i.e., degree or conditions of supervision). Overall, the use of recidivism measures within risk assessment tools may ultimately enhance public safety, improve treatment services, and reduce correctional expenditures.

Accurately Assessing and Predicting Recidivism: The Whole Picture

Recidivism is one of the most studied topics in the criminological literature. The failure to provide more focus and attention on local jail and community correction populations results in not accurately and comprehensively assessing the problem of recidivism and how it may be reduced through various correctional strategies. It has been reported that 729 295 inmates are released from prison and return to their communities each year (Bureau of Justice Statistics, 2009, 2010), with approximately 2000 offenders released daily. With policy makers and academics diligently trying to find ways to prevent those individuals from recidivating after their release, their work is only partially addressing the recidivism problem. Even with the proliferation of validated risk assessment tools, only a portion of the story is being told relative to all forms of recidivism.

A primary reason for understanding the causes of recidivism and determining methods of reducing recidivism is to address the crime control problem in communities. Research on recidivism has taken many avenues through the focus on the

differential treatment of minority offenders (Reisig et al., 2007), gender differences in the needs of those who return to communities (Deschenes et al., 2006; Reisig et al., 2006; Holtfreter et al., 2004), gender differences in predicting recidivism (Bonta et al., 1995), offender rehabilitation (Harm and Phillips, 2000), and employment and its effects on recidivism (Wang et al., 2010). There is an importance for these areas to be studied on jail and community corrections populations as well.

Visher and Travis (2003), in an article, titled *Transitions from Prison to Community: Understanding Individual Pathways* discuss the preprison, in-prison, immediate postprison, and postrelease integration experiences (p. 94) of offenders and how each of these experiences play an important role on the decisions made once these individuals are back in the community. The authors stress the importance of how these individual experiences are extremely valuable in determining whether or not someone will or will not recidivate. However, there is only mention of offenders who are released from prison and not those released from local jails or those under community supervision who may also recidivate. Going forward, criminological research must work to incorporate jail and community correction populations in the measurement and assessment of recidivism if the true nature of this phenomenon is to be captured. Simply relying on inmate populations will not serve this discipline in the advancement of understanding offending behaviors as it excludes the largest parts of the picture.

Conclusion

National statistics indicate that the majority of offenders are being supervised in the community, which suggests that seemingly innocuous community-based measures of recidivism are as important if not more important than more conventional forms of reoffending. The observed drop in national crime rates in the past two decades allows researchers the opportunity and a little more freedom to assess alternative forms of recidivism. Eventually, the objective is to develop risk assessment procedures to determine those factors that predict status offenses, such as absconding and receiving a technical violation, and reduce them like commonly measured forms of crime.

In relation to incorporating community-based measures of recidivism, there is also a need to study jail populations when attempting to understand the current recidivism issue. These underexamined populations, once studied, could change how we as researchers and policy makers approach recidivism. There are an extremely large number of offenders (12.9 million released from jail annually and approximately 4.9 million under community supervision) who are left out of policy-making decisions when only those offenders who were released from prison are discussed. In an attempt to create more efficient and effective policies regarding reentry and recidivism in this county, these two underexamined populations must be addressed.

Research in this area should move forward by assessing the similarities as well as differences among offenders who are released from prison, those who are released from jail, and those who are under community supervision. The trends of recidivism for jail populations and community corrections should also be studied. Until recidivism is studied in regard to these two underexamined populations, the true state of recidivism cannot be fully understood.

See also: Data Bases and Statistical Systems: Crime Measurement; Forecasting; Prison Life, Sociology of; Prisons and Imprisonment; Re-entry and Reintegration After Incarceration; Rehabilitation of the Offender.

Bibliography

Andrews, D.A., Bonta, J., 1995. The level of supervision inventory-revised. Toronto: Multi-Health Systems 106, 19–52.

Andrews, D.A., Bonta, J., 1998. The Psychology of Criminal Conduct, second ed. Anderson, Cincinnati, OH.

Andrews, D.A., Zinger, I., Hoge, R.D., Bonta, J., Gendreau, P., Cullen, F.T., 1990. Does correctional treatment work? A clinically-relevant and psychologically informed meta-analysis. Criminology 28, 369–404.

Armstrong, G.S., 2012. Factors to consider for optimal span of control in community supervision evidence-based practice environments. Criminal Justice Policy Review 23 (4), 427–446.

Beck, A.J., 2006. The Importance of Successful Reentry to Jail Population Growth. Presented to the Urban Institute Jail Reentry Roundtable. Urban Institute, Washington, DC.

Beck, A.R., 2001. Recidivism: a fruit salad concept in the criminal justice world. Justice Concepts Incorporated, 1–3.

Benedict, W.R., Huff-Corzine, L., 1997. Return to the scene of the punishment: recidivism of adult male property offenders on felony probation, 1986–1989. Journal of Research in Crime and Delinquency 34 (2), 237–252.

Benedict, W.R., Huff-Corzine, L., Corzine, J., 1998. "Clean up and go straight": effects of drug treatment on recidivism among felony probationers. American Journal of Criminal Justice 22 (2), 169–187.

Blumstein, A., Larson, R.C., 1971. Problems in modeling and measuring recidivism. Journal of Research in Crime and Delinquency 8 (2), 124–132.

Bonta, J., 2002. Offender risk assessment: guidelines for selection and use. Criminal Justice and Behavior 29, 355–379.

Bonta, J., Law, M., Hanson, R.K., 1998. The prediction of criminal and violent recidivism among mentally disordered offenders. Psychological Bulletin 123, 123–142.

Bonta, J., Pang, B., Wallace-Capretta, S., 1995. Predictors of recidivism among incarcerated female offenders. The Prison Journal 75 (3), 277–294.

Bureau of Justice Statistics, 2009. Bulletin: Prisoners in 2009. Bureau of Justice Statistics, Washington, DC (retrieved 28.01.11.).

Cadigan, T.P., 1991. Electronic monitoring in federal pretrial release. Federal Probation 55, 26–30.

Cooprider, K.W., Kerby, J., 1990. Practical application of electronic monitoring at the pretrial stage. Federal Probation 54, 28–35.

Cunningham, M.D., Sorensen, J.R., 2006. Actuarial models for assessing prison violence risk revisions and extensions of the risk assessment scale for prison (RASP). Assessment 13 (3), 253–265.

Davies, G., Dedel, K., 2006. Violence risk Screening in community corrections. Criminology & Public Policy 5 (4), 743–770.

DeMichele, M., Payne, B.K., 2012. Measuring, community corrections' officials perceptions of goals, strategies, and workload from a systems perspective: differences between directors and nondirectors. The Prison Journal 92 (3), 388–410.

Deschenes, E.P., Owen, B., Crow, J., 2006. Recidivism among female prisoners: secondary analysis of the 1994 BJS recidivism data set. Final Report, National Institute of Justice, U.S. Department of Justice.

Gendreau, P., Goggin, C., Smith, P., 2002. Is the PCL-R really the "unparalleled" offender risk assessment measure? Criminal Justice and Behavior 29, 397–426.

Gilsinan, J.F., 1986. Creating a reform environment: a case study in community corrections and coalition building. Criminal Justice Policy Review 1 (3), 329–343.

Glaze, L.E., Bonczar, T.P., 2011. Probation and Parole in the United States, 2010 (NCJ 236019). US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC.

Gray, K.M., Fields, M., Maxwell, S.R., 2001. Examining probation violations: who, what and when. Crime and Delinquency 47 (4), 51–53.

Griffen, M.L., Armstrong, Gaylene S., 2003. The effect of local life circumstances on female probationers' offending. Justice Quarterly 20 (2), 213–239.

Grove, W.M., Meehl, P.E., 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. Psychology, Public Policy, and Law 2, 293–323.

Guerino, P., Harrison, P.M., Sabol, W.J., 2011. Prisoners in 2010. Bureau of Justice Statistics, Washington, DC.

Gunnison, E., Helfgott, J.B., 2011. Factors that hinder offender reentry success: a view from community officers. International Journal of Offender Therapy and Comparative Criminology 55 (2), 287–304.

Hanley, D., 2002. Risk Differentiation and Intensive Supervision: A Meaningful Union? (An unpublished doctoral dissertation) University of Cincinnati, Cincinnati, OH.

Hare, R.D., 2003. The Hare Psychopathy Checklist – Revised, second ed. Multi-Health Systems, Toronto.

Hare, R.D., Harpur, T.J., Hakstian, A.R., Forth, A.E., Hart, S.D., Newman, J.P., 1990. The revised psychopathy checklist: reliability and factor structure. Psychological Assessment 2, 338–341.

Harlow, C.W., 2003. Education and Correctional Populations. In: Bureau of Justice Statistics. U.S. Department of Justice, Washington, DC.

Harm, N.J., Phillips, S.D., 2000. You can't go home again. Journal of Offender Rehabilitation 32 (2), 3–21.

Holtfreter, K., Reisig, M.D., Morash, M., 2004. Poverty, state capital, and recidivism among women offenders. Criminology and Public Policy 3 (2), 185–208.

Hughes, K., 2006. Justice Expenditure and Employment Extracts, 2006. Bureau of Justice Statistics. U.S. Department of Justice, Washington, DC.

Hughes, T.A., Wilson, D.J., Beck, A.J., 2001. Trends in State Parole, 1990–2000. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Hylton, J.H., 1982. Rhetoric and reality: a critical appraisal of community correctional programs. Crime and Delinquency 28 (3), 341–373.

Johnson, B.D., Dipietro, S.M., 2012. The power of diversion: intermediate sanctions and sentencing disparity under presumptive guidelines. Criminology 50 (3), 811–850.

Kubrin, C.E., Stewart, Eric A., 2006. Predicting who reoffends: the neglected role of neighborhood context in recidivism studies. Criminology 44, 165–197.

Langan, P., Levin, D., 2002. Recidivism of Prisoners Released in 1994. In: Bureau of Justice Statistics. U.S. Department of Justice, Washington, DC.

Lawrence, R., 1991. Reexamining community corrections models. Crime & Delinquency 37 (4), 449–464.

Lowenkamp, C.T., Bechtel, K., 2007. The predictive validity of the LSI-R on a sample of offenders drawn from the records of the Iowa department of corrections data management system. Federal Probation 71 (3), 25–29.

Lowenkamp, C.T., Latessa, E.J., Holsinger, A.M., 2006. The risk principle in action: what have we learned from 13,676 offenders and 97 correctional programs? Crime & Delinquency 52, 77–93.

MacKenzie, D.L., 2006. What Works in Corrections: Reducing the Criminal Activities of Offenders and Delinquents. Cambridge University, New York.

Maltz, M.D., 2001. Recidivism. Academic Press, Inc., Orlando, FL.

Miller, H.V., Miller, J.M., 2010. Community in-reach through jail reentry: findings from a quasi experimental design. Justice Quarterly 27 (6), 893–910.

Minton, T.D., 2011. Jail Inmates at Midyear 2010 – Statistical Tables. In: Bureau of Justice Statistics. U.S. Department of Justice, Washington, DC.

Osher, F.C., January/February 2007. Short-term strategies to improve reentry of jail populations. American Jails, 9–18.

Osher, F.C., Steadman, H.J., Barr, H., 2002. A Best Practice Approach to Community Reentry from Jails for Inmates with Co-occurring Disorders: The APIC Model. The National GAINS Center, Delmar, NY.

Petersilia, Joan, 1995. A crime control rationale for reinvesting in community corrections. The Prison Journal 75 (4), 479–496.

Pettus, C.A., Severson, Margaret, 2006. Paving the way for effective reentry practice: the critical role and function of the boundary spanner. The Prison Journal 86 (2), 209–229.

Reisig, M.D., Bales, William D., Hay, Carter, Wang, Xia, 2007. The effect of racial inequality on black male recidivism. Justice Quarterly 24 (3), 408–434.

Reisig, M.D., Holfreter, Kristy, Morash, Merry, 2006. Assessing recidivism risk across female pathways to crime. Justice Quarterly 23 (3), 384–405.

Sabol, W., Minton, T., Harrison, P.M., 2007. Prison and Jail Inmates at Midyear, 2006 (NCJ 217675). U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Simon, J., Feeley, M.M., 2003. The forms and limits of the new penology. In: Blomberg, T.G., Cohen's, S. (Eds.), Punishment and Social Control, second ed. Aldine De Gruyter, New York.

Solomon, A.L., Osborne, J.W.L., LoBuglio, S.F., Mellow, J., Mukamal, D.A., 2008. Life after Lockup: Improving Reentry from Jail to the Community. Urban Institute, Washington, DC.

Stephen, J.J., 2001. Census of Jails, 1999 (NCJ 186633). U.S. Department of Justice, National Institute of Justice, 1997, Washington, DC.

Turley, A., Thornton, T., Johnson, C., Azzolino, S., 2004. Jail drug and alcohol treatment program reduces recidivism in nonviolent offenders: a longitudinal study of Monroe County, New York's, jail treatment drug and alcohol program. International Journal of Offender Therapy and Comparative Criminology 48, 721–728.

Ulmer, J.T., 2001. Intermediate sanctions: a comparative analysis of the probability and severity of recidivism. Sociological Inquiry 71 (2), 164–193.

Van Voorhis, P., Braswell, M., Lester, D., 2007. Correctional Counseling and Rehabilitation, sixth ed. Anderson, Cincinnati.

Visher, Christy, Travis, Jeremy, 2003. Transitions from prison to community: understanding individual pathways. Annual Review of Sociology 29, 89–113.

Wang, X., Mears, D.P., Bales, William D., 2010. Race-specific employment contexts and recidivism. Criminology 48, 1171–1211.

Webb, V.J., Hoffman, D.E., Wakefield, W.O., Snell, J., 1976. Recidivism: in search of a more comprehensive definition. International Journal of Offender Therapy and Comparative Criminology 20 (2), 144–147.

Wodahl, E.J., Garland, B., 2009. The evolution of community corrections: the enduring influence of the prison. The Prison Journal 89 (1), 81–104.

Wodahl, E.J., Ogle, R., Heck, C., 2011. Revocation trends: a threat of legitimacy of community-based corrections. The Prison Journal 91 (2), 207–226.

Wright, K.A., Pratt, T.C., Lowenkamp, C.T., Latessa, E.J., 2012. The importance of ecological context for correctional rehabilitation programs: understanding the micro- and macro-level dimensions of successful offender treatment. Justice Quarterly 29 (6), 775–798.

Recognition

Elaine J Mahoney, Allison E Nickel, and Deborah E Hannula, University of Wisconsin, Milwaukee, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recognition memory permits the distinction of items encountered in the past from those that are novel. How exactly this is accomplished has been a source of considerable theoretical debate and the number of recognition memory investigations that have been reported in the literature is astounding. Here, the authors provide readers with a brief overview of this literature. Methods that permit investigators to provide precise quantitative descriptions of recognition memory performance are described, recognition memory models said to account for empirical findings are summarized, and a subset of factors that affect the accuracy of recognition memory responses are outlined.

Recognition memory is a ubiquitous part of a person's day-to-day experience – it permits differentiation of one's car from others that are parked nearby, identification of a middle school classmate who happens to be standing nearby in the aisle at the grocery store, and selection of correct responses on multiple-choice exams. Recognition memory impairments have been linked to particular types of brain damage and dysfunction that occur as a consequence of injury, disease, or psychiatric impairment, and recognition memory errors can have substantial real-world consequences (e.g., wrongful conviction of a crime). It is not surprising then, that recognition memory has been the subject of extensive investigation in laboratory settings and that the reported results have prompted considerable theoretical debate.

In standard recognition memory experiments, participants attempt to commit information to memory (e.g., words, pictures, and nonsense syllables) so that when a test is administered, they can distinguish studied from novel materials. On these tests, participants may be asked to identify individual items as old or new (old/new recognition), select studied items presented among novel alternatives (forced-choice recognition), identify items that were presented together during a corresponding study phase (associative recognition), or identify the context in which a particular item was initially encountered (source recognition). Participants may also be asked to rate how confident they are in the accuracy of their recognition memory responses, and these confidence ratings can be used to plot receiver operating characteristic (ROC) curves. The sections that follow provide readers with a brief introduction to signal detection theory, which can be used to evaluate recognition memory performance; an overview of competing recognition memory models; and a summary of some key findings from the recognition memory literature. Factors that affect recognition memory performance – i.e., that make recognition memory decisions increasingly accurate or error prone – take top billing in this final section.

Assessing Recognition Memory Performance: Signal Detection Theory

Participants in recognition memory experiments attempt to distinguish studied from novel materials, and signal detection theory (SDT) is frequently used to evaluate performance. SDT permits quantitative assessment of the decision-making process that is used to discriminate signal from noise across a range of domains (e.g., medical diagnosis and industrial inspection) and was originally applied by psychologists to investigations of perceptual processing (e.g., presence vs absence of an auditory stimulus; see Green and Swets, 1966). In the context of recognition memory, SDT permits investigators to determine how effectively participants discriminate studied from novel materials (i.e., sensitivity), and to assess whether or not performance has been affected by any systematic response preferences (i.e., bias) such as a disproportionate tendency to endorse items as old (studied).

According to SDT, recognition decisions are mediated by the strength (or familiarity) of memory representations associated with individual items encountered on a recognition memory test; for each item, the strength of this memory signal is compared with a response criterion set by the observer and a recognition memory decision is made. On average, old items are assumed to have greater memory strength than new items by virtue of having been studied, but the distribution of memory strength across items is not uniform (i.e., some items, studied or not, seem more familiar than others) – this is depicted graphically in Figure 1, which shows that studied and novel items form overlapping Gaussian distributions. 'Old' responses are made when signal strength exceeds the criterion set by the observer and 'new' responses are made otherwise. Where exactly the criterion happens to fall is at the discretion of the participant, but may be influenced by several factors including, but not limited to, instructions administered prior to the recognition memory test (e.g., being encouraged to guess) and the possibility of reward or punishment based on performance.

Corrected Recognition Scores

The resulting model of the decision making provides a framework that can be used to characterize correct and incorrect recognition memory responses. After a recognition test has been administered, individual test trials can be assigned to one of four distinct categories based on the actual status of the test item (i.e., studied or novel) and the corresponding behavioral response made by the observer (i.e., old or new). Studied and novel items that elicit correct responses are referred to as hits and correct rejections, respectively; studied and novel items that elicit incorrect responses are referred to as misses and false alarms (see Figure 1). Something to keep in mind is that

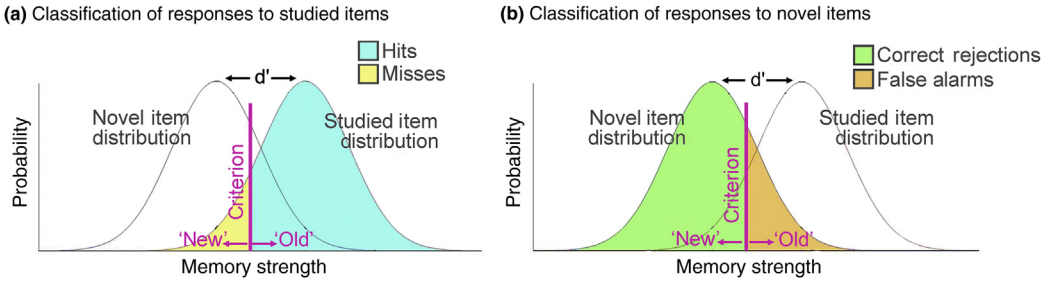

Figure 1 Illustration of signal detection theory. Novel and studied items form overlapping Gaussian distributions. Memory strength, which is greater for the studied item distribution, is plotted on the *x*-axis. (a) Studied items correctly endorsed as 'old' based on the response criterion are shown in blue; studied items that are incorrectly rejected as 'new' are shown in yellow. (b) Novel items correctly rejected as 'new' based on the response criterion are shown in green; novel items incorrectly endorsed as 'old' are shown in orange.

evaluation of the hit rate alone does not provide a complete picture of recognition memory performance because this measure does not take into account how often participants said *old*, in error, when novel items were presented. A participant who responds *old* to most items will have a very high hit rate, but the false alarm rate will also be high, and discriminability will be poor. Therefore, the hit rate and the false alarm rate should be considered together, and investigators typically report a corrected recognition score (e.g., the hit rate minus the false alarm rate).

Discriminability and Response Bias

Although corrected recognition scores are informative, they do not tell the full story. Imagine, for example, that an experimenter has designed a recognition memory experiment with two conditions – participants study neutral or emotionally arousing pictures, and are asked to distinguish studied from novel exemplars on a corresponding recognition memory test. After having calculated corrected recognition scores for each condition, the experimenter finds that performance is better for emotionally arousing than for neutral pictures. While this outcome is notable, it provides no insight into whether the observed difference is due to superior discrimination of emotional (vs neutral) content (i.e., a difference in sensitivity), the influence of a systematic response bias, or some combination of these factors.

To avoid potential misinterpretation of results, SDT can be used to obtain numerical estimates of sensitivity (e.g., d' – pronounced 'd prime') and bias (e.g., c). The sensitivity estimate (d') indicates how far apart the old and new item distributions are from one another (see **Figure 1**) and provides a measure of discriminability. A d' score equal to zero (i.e., no difference in mean signal strength for studied and novel item distributions) indicates that participants cannot reliably distinguish old from new items (i.e., they are essentially guessing), and larger values reflect increasingly improved discriminability. The bias estimate (c) indicates how far the decision criterion is from a neutral point where old and new responses would be made equally often. Graphically, the neutral point is the place where the studied and novel item distributions cross (see **Figure 1**). If criterion is to the right of this point (i.e., c is positive), then participants have adopted a conservative response bias and favor new responses; if criterion is to the left of this point (i.e., c is negative), then participants have adopted a liberal response bias and favor old responses. A more complete explanation of SDT, along with mathematical formulas and descriptions of computer software applications that can be used to obtain quantitative estimates of sensitivity and bias, is provided by Stanislaw and Todorov (1999; for an alternative perspective readers should consult Kroll et al., 2002).

Recognition Memory Models

Several recognition memory models have been proposed in the literature. These models provide conceptual and quantitative descriptions of the processes that support recognition memory performance on a variety of tests (e.g., item recognition and associative recognition), and many of them incorporate aspects of SDT. At the broadest level, these models can be distinguished based on whether recognition memory is said to be supported by a single underlying process (i.e., memory strength/familiarity) or by two processes (i.e., familiarity and recollection). In what follows, the authors provide readers with a brief introduction to a subset of the proposed models. For more complete treatment, including detailed discussion of the points of departure across models, and descriptions of ongoing debate about the relative merits of single- and dual-process accounts of recognition memory, the authors direct readers to several excellent reviews (Clark and Gronlund, 1996; Parks and Yonelinas, 2007; Wixted, 2007; Yonelinas, 2002).

Single-Process Models

Single-process models borrow substantially from SDT and assume that recognition memory decisions are based on the strength (familiarity) of tested items. A number of single-process models, categorized collectively as global matching models (e.g., the *search of associative memory* or SAM model and the *theory of distributed associative memory* or TODAM model) have been described in the literature. These models account for a number of empirical observations – among them, observed reductions in recognition memory performance consequent to an increase in the number of to-be-remembered materials (i.e., list-length effects), incorrect endorsements of novel items as 'old' when they resemble studied materials (i.e., prototype effects), and

influences of context reinstatement on item recognition (i.e., context effects). Context effects have been documented when participants attempt to distinguish studied from novel words. For example, results show that when individual words (e.g., shadow) are studied as pairs (e.g., shadow–ladder and clock–bottle), they are more often recognized when the pairs are reinstated (i.e., shadow–ladder) at test than when they are rearranged (i.e., shadow–bottle). Global memory models predict and account for context effects by proposing that a single, strength-based metric reflects the combined influence of memory for individual items and additional associated information (e.g., the second item in a pair). This global representation of memory characteristics underlies memory strength and guides recognition memory decisions in a process that is well described by standard SDT. The origins of each model, their effectiveness in accounting for empirical findings, and their operating characteristics are evaluated in detail by Clark and Gronlund (1996). As indicated by these authors, a model that stands the test of time must account for reported results and generate new predictions. While the global memory models have been successful on both fronts, there are also empirical findings that the models cannot explain without modification (e.g., associative recognition effects). These difficulties, along with dissociations that have been reported in neuropsychological and neuroimaging experiments, suggest that recognition memory may be best described by more than one process, and that recognition memory decisions may be based on more than just the assessment of memory strength (see Clark and Gronlund, 1996; Yonelinas, 2002 for details).

Dual-Process Models

Dual-process models propose that recognition memory decisions are based on two processes – familiarity and recollection. Familiarity permits us to identify something as having been seen in the past (e.g., I know I have seen this key before), but affords no contextual detail. Recollection permits us to recognize the item and recall additional information about the context in which it was originally encountered (e.g., this is the key that opens the shed behind the house).

Early dual-process models were proposed by Atkinson et al. (Juola et al., 1971) and by Mandler et al. (Mandler et al., 1969). According to the Atkinson model, which has since been dubbed a conditional search model, an effortful memory search (i.e., recollection) is engaged if and only if a fast-acting familiarity process fails to provide conclusive evidence about whether or not an item was studied. As in SDT, studied and novel items are assumed to form overlapping Gaussian distributions, but now participants are said to adopt a low criterion as well as a high criterion. Items with signal strength that exceeds the high criterion are endorsed as old, whereas those with signal strength less than the low criterion are rejected as new. If the memory strength of an item places it in between these two criteria, then familiarity is not diagnostic and an effortful search (i.e., recollection) guides the decision-making process (hence the designation conditional search). Like Atkinson, Mandler proposed that recognition memory is governed by two processes – a fast-acting perceptual matching process (i.e., familiarity) and a slower recall process (i.e., recollection). While the two processes were initially assumed to operate in succession (as above), the model has since been revised and it is now proposed that familiarity and recollection proceed in parallel, but that familiarity is faster and 'wins the race.' According to Mandler, familiarity is the product of a perceptual match between a stimulus and a memory representation. With repeated exposure, the perceptual representation of an item is strengthened and becomes increasingly integrated in memory; this renders the perceptual match more likely during a test phase. At the same time, an increasingly elaborate contextual network is also established. This information is used to guide the memory search process (i.e., recall) and supports recollection (cf Mandler, 1980).

Jacoby et al. (e.g., Jacoby and Dallas, 1981) have also advocated for a two-process model of recognition memory; here, however, the focus has been on factors that give rise to the subjective experiences of familiarity and recollection. According to the model, feelings of familiarity are attributable to perceptual or conceptual processing fluency that has accrued based on past experience with test materials; information that is processed more easily is endorsed as having been studied. Notably, fluency does not have to be a consequence of past experience – a high-contrast word may be processed more fluently than a low-contrast word and mistakenly endorsed as having been seen during study as a result. Familiarity is characterized as fast-acting and automatic (i.e., it does not depend upon conscious deliberation or effortful retrieval attempts), and fluent processing is said to contribute to performance whether memory is tested directly or indirectly (i.e., with or without reference to a previous study experience). Recollection, on the other hand, is characterized as effortful and said to depend upon a conscious, cognitively controlled search process. In this case, participants base recognition memory decisions on the recovery of unique contextual information that was associated with a studied item when it was originally encountered.

Estimates of the contributions of recollection and familiarity to performance can be obtained via the process dissociation procedure (Jacoby, 1991). To obtain these estimates, participants might be instructed to study two separate lists of items (lists A and B). In one test condition (i.e., the inclusion condition), participants are told to endorse studied items from either list as 'old' and to reject novel items. Under these circumstances it is assumed that familiarity and recollection both contribute to accurate recognition. In a second test condition (i.e., the exclusion condition), participants are instructed to endorse items from list A as 'old,' but to reject novel items and items from list B. Here, it is assumed that any incorrect endorsement of items from list B as 'old' is attributable to familiarity because if recollection had been successful, participants would have rejected list B items as instructed. Contributions of recollection to performance are estimated by calculating the proportion of old responses made to list B items in the inclusion and exclusion conditions, and then subtracting the exclusion score (which is assumed to reflect the influence of familiarity alone) from the inclusion score (which is assumed to reflect the influences of recollection and familiarity). When the recollection estimate has been obtained, an equation can be used to estimate the contribution of familiarity to performance. The

equations used to obtain these estimates are subject to several assumptions about the processes that are being measured and the approach is not always endorsed in the literature. For a recent review that advocates for broad applicability of the process dissociation procedure and a discussion of criticisms that have been levied against it, readers should consult Yonelinas and Jacoby (2012).

Perhaps most influential among the two process models is the dual-process signal detection (DPSD) model proposed by Yonelinas (1994). According to the model, which has been applied widely to behavioral, neuroimaging, and neuropsychological investigations (Yonelinas et al., 2010), familiarity is based on the assessment of memory strength as described by SDT and recollection is threshold based. Adopting a memory threshold to characterize the experience of recollection means that this process can fail. Some items may seem familiar, but one simply cannot recall any additional detail about the context in which they were encountered. Under these circumstances, recollection does not provide any information that can be used to distinguish old from new items, but familiarity may still be diagnostic of past experience. A common misconception in the literature has been that the use of a threshold implies that recollection is all or none – i.e., participants will either recover all of the qualitative/contextual information that was associated with a particular item during encoding or nothing. Yonelinas et al. (cf Parks and Yonelinas, 2007) have clarified this point, indicating that while recollection can indeed fail to exceed threshold, above-threshold responses simply imply that some additional information has been recovered. The amount of information that is successfully retrieved may vary from trial to trial or across participants, and strong claims have not been made about the shape taken by the distribution of recollected items.

According to the DPSD model, when recollection occurs, participants will endorse items as old and indicate high confidence in the accuracy of their recognition memory decision. Items that are not recollected will be endorsed as old when memory strength exceeds criterion, and these familiarity-based recognition responses are expected to be endorsed with a range of confidence judgments. When confidence judgments have been obtained, numerical estimates of recollection and familiarity can be calculated by plotting ROC curves and fitting the data to a quantitative model. ROC curves are generated by plotting hit rates (y-axis) against false alarm rates (x-axis) at increasingly liberal levels of recognition memory confidence. The leftmost point on an ROC curve reflects corresponding hit and false alarm rates for the highest confidence responses; subsequent points reflect cumulative hit and false alarm rates at increasingly lower levels of confidence. As described in more detail by Yonelinas et al. (2010), d′ (i.e., the degree of curve in the ROC) is used to estimate contributions of familiarity to performance, and the y-axis intercept provides an estimate of recollection. Much of the current theoretical debate about the validity of proposed recognition memory models centers on the shape of ROC curves and z-transformed ROC curves. In contrast to other dual-process models, the DPSD model can predict the shape of these distributions across a wide range of empirical investigations (see Yonelinas and Parks, 2007). The same effects are also generally well predicted by a recently proposed unequal variance signal detection model in which it is stated that memory strength reflects the combined (rather than independent) influences of familiarity and recollection (Wixted and Stretch, 2004). In contrast to standard SDT, the standard deviation of the studied item distribution is greater than the novel item distribution in this model. Discussions of the relative merits and shortcomings of each approach and points of contention about the ability of each model to fit recognition memory data are provided in companion papers by Wixted (2007) and by Parks and Yonelinas (2007).

Key Findings: Factors that Affect Recognition Memory Performance

Empirical reports indicate that recognition memory performance can be affected by a range of factors (e.g., characteristics of to-be-encoded materials, processing operations engaged during encoding, and context reinstatement). Because the recognition memory literature is substantial, an exhaustive discussion of these factors is beyond the scope of this article. In what follows, the authors endeavor to highlight some of the key empirical observations that have been reported in the recognition memory literature, highlighting a subset of the factors that can make recognition memory more accurate or increasingly error prone.

The Mirror Effect: Characteristics of To-Be-Encoded Materials Affect Recognition Memory Performance

Whether or not recognition memory decisions are correct can be affected by basic properties of the materials that are used in an experiment. Simply put, some materials are more easily recognized than others. They are more likely to be endorsed as old when they have been studied and are more often rejected as new when they are novel. The classic demonstration of this material-specific advantage is based on word recognition, and shows that words with a low normative frequency (e.g., henchman) have higher hit rates and lower false alarm rates than words with high normative frequency (e.g., handyman). This word frequency effect, which has dubbed the mirror effect by Glanzer and Adams (1985), is so robust, and so easily replicated, that it is considered a regularity of recognition memory and is frequently used in classroom demonstrations. Mirror effects have also been documented in studies that manipulate list length and encoding strength. Performance advantages in these investigations are typically documented for items that were encoded in the context of short (vs long) lists, and for items that were presented more (vs less) frequently during encoding (Cary and Reder, 2003).

There is ongoing debate about what exactly drives the reported differences in recognition memory performance for low- versus high-frequency words. According to one view, the low-frequency advantage is a consequence of differences in attentional allocation. Low-frequency words benefit from extra attention during the study phase, which makes them more distinctive and more discriminable from novel exemplars at test (Malmberg and Nelson, 2003). An alternative account,

based on dual-process models of recognition memory, attributes differences in the false alarm rate to baseline levels of memory strength. Because high-frequency words are more familiar than low-frequency words, they are more likely to be endorsed as old in error. The difference in hit rate is accounted for by the observation that low-frequency words are more likely to be recollected. As indicated by Diana and Reder (2006), the above accounts are not necessarily mutually exclusive – findings from their experiments and others indicate that the recollection-based performance advantage for low-frequency words depends upon sufficient allocation of attention to these items during encoding. Future work will undoubtedly further refine theories about the mechanistic underpinnings of the mirror effect.

Facilitative Effects of Deep Encoding on Subsequent Recognition Memory Performance

The processing operations that are engaged during encoding can also affect recognition memory outcomes. According to the levels of processing model (Craik and Lockhart, 1972), materials are better remembered when they have been processed more deeply during encoding. In the context of word learning investigations, deep processing tasks encourage participants to evaluate the meaning of each word on a to-be-remembered list (e.g., indicate whether the word refers to a living or nonliving object), whereas shallow processing tasks encourage participants to evaluate superficial features of these words (e.g., indicate whether each word appears in upper or lower case text). That depth of processing does indeed affect recognition memory performance is not subject to debate and deep processing manipulations are frequently used in recognition memory experiments to facilitate encoding. Critically, evidence suggests that the reported effects are not a consequence of spurious differences in task difficulty or processing time across encoding conditions, and that they do not depend upon the intention to commit studied materials to memory (cf Craik and Tulving, 1975). Robust effects of these encoding manipulations are evident whether participants expect a recognition memory test or not, and performance on incidental recognition memory tests that follow deep encoding is on par with, or even superior to, performance when recognition memory tests are expected (Eagle and Leiter, 1964).

Context Reinstatement Effects

Intuitively, it seems that memory should be affected by whether or not there is a match between an encoding context and the associated test context. Consistent with this intuition, robust context reinstatement effects have been documented when participants attempt to recall encoded content, but the same manipulations have yielded mixed results when recognition memory tests are administered. It is difficult to provide an overarching definition of context, but in general, context refers to factors that are peripheral or subsidiary to intentionally encoded information. The studies described below are representative examples of investigations that have examined the effects of environmental and associative context manipulations on recognition memory performance.

The classic demonstration of environmental context reinstatement on memory performance was based on an experiment conducted by Godden and Baddeley (1975). Participants in this investigation studied a word list on land or underwater, and were either tested in the same environmental context or not. In this initial investigation, memory was tested via recall rather than recognition, and performance was best when the encoding and test contexts were the same (e.g., information encoded underwater was also retrieved underwater). Surprisingly, when the same experiment was conducted using a recognition test, context reinstatement did not affect performance (Godden and Baddeley, 1980). According to the encoding specificity principle (cf Thomson and Tulving, 1970), the effectiveness of a retrieval cue (in this example, environmental context) depends critically upon whether or not that information was processed and stored with to-be-encoded content, and whether or not that information is accessed during retrieval. When these requirements are not met, any anticipated benefit of an experimental manipulation may be absent. Consistent with the encoding specificity principle, it has been proposed that the relative absence of environmental context reinstatement effects may be a consequence of the failure to integrate contextual information into the memory representation that is used to support performance (Murane et al., 1999).

While a match in environmental context may not always impact recognition memory performance, effects of associative context reinstatement have been reported consistently in the literature. In these experiments, items co-occur during encoding (e.g., word pairs, faces, and scenes), but memory for the association is not emphasized and is not necessary for accurate performance during the test phase. One example of an associative context effect was described above – individual words are better remembered when studied pairs are reinstated (vs rearranged) at test despite the fact that memory for the pairs themselves is unnecessary (see Clark and Gronlund, 1996). Similar effects have also been documented in investigations that pair faces with scene contexts. In these experiments, recognition of preexperimentally unfamiliar faces is better when they are presented with the same scenes during encoding and test than when the pairings are rearranged; as in the word recognition investigations, memory for the pairing, and in this case for the scenes themselves, was not an explicitly stated performance requirement. Remarkably, this recognition memory performance advantage is evident even when the scene is a different exemplar from the same basic category (e.g., two different kitchens; Memon and Bruce, 1983; Rainis, 2001). This outcome suggests that the influence of scene context on face recognition is not strictly perceptual, but that semantic (meaning-based) characteristics are abstracted from the picture and that this higher level representation is associated with the face in memory.

These examples illustrate what was stated above – the effects of context reinstatement on recognition memory outcomes have been mixed. While some effects are robust and reproducible, others are not. More information about the effects of context on recognition memory performance, the implications of these findings for models of recognition memory (especially the global memory models), and discussion of when and under what circumstances context effects

might be expected can be found in Smith and Vela (2001) and Murane et al. (1999).

False Recognition: Factors that Make Recognition Memory Unreliable

In this final section, the authors cite work that has identified factors that make recognition memory increasingly error prone. As described above, recognition memory errors can take one of two forms – participants may fail to recognize studied information (i.e., a miss) or they may mistakenly endorse novel information as old (i.e., false alarm). Empirical investigations have focused disproportionately on one type of memory error – false alarms – perhaps because the potential consequences of false recognition can be so grave (e.g., wrongful conviction of a crime; see Wells, 1993).

Reliable demonstrations of false memory have been reported in a number of investigations that use a variety of experimental manipulations. False recognition can be induced by providing participants with misinformation about encoded content prior to test (see Loftus, 2005 for review), by manipulating the ease with which novel materials are processed (Jacoby and Whitehouse, 1989), and by presenting novel words at test that are semantically related to words from a studied list (Roediger and McDermott, 1995). All of these findings have had a substantial impact on the memory literature and are important not only empirically, but practically. For brevity, just one of these effects, the Deese–Roediger–McDermott (or DRM) effect, is described in detail here. The DRM effect is so robust that it is commonly used as a classroom demonstration to illustrate the fallibility of memory. The reliability of the effect and the simplicity of the DRM paradigm have also made it a favorite among investigators, who have used it to identify the neural substrates of false memory and to address questions about memory in special populations (e.g., children, older adults, and brain-damaged patients).

The DRM paradigm is based on a list-learning effect that was originally reported by Deese (1959) and was subsequently revived by Roediger and McDermott (1995). Participants recruited for a DRM experiment are presented with several lists of words that reliably elicit memory for a single target word that was not studied. Six lists of 12 words, each with meaning (semantics) related to one nonstudied word would be presented. A list might contain words like 'tired,' 'bed,' and 'dream,' which are all semantic associates of the nonpresented word 'sleep.' After the word lists were presented, participants would be asked to recognize the words that were studied; related, nonstudied words were presented on the recognition test as foils. Using this approach, Roediger and McDermott found that the hit rate (i.e., correct identification of studied words) was roughly equivalent to the false alarm rate (i.e., mistaken identification of foils as studied). In other words, participants were recognizing nonstudied associates of actual list members at an alarmingly high rate. When participants were asked to indicate whether they remembered the words (i.e., recollected some detail about having studied the word) or whether they simply knew the words had been presented, 72% of the false alarms were endorsed as remembered. This response corresponds to the experience of recollection, so not only were participants endorsing semantically related foils as having been studied, they were also expressing high confidence in the accuracy of their recognition memory responses.

Why exactly participants endorse these nonstudied semantic associates as remembered was puzzling as it is impossible to recover qualitative details of a study experience, when something has not been studied. To determine what was driving the reported outcome, Norman and Schacter (1997) developed a modified version of the DRM paradigm in which they had participants justify their responses when words (studied or not) were identified as old and endorsed as remembered. Participants were asked to explicitly describe details of the corresponding study experience (e.g., to indicate what words preceded/followed the remembered word, to indicate what they were thinking about when that word was presented, and to indicate whether they remembered the word because of its strong semantic relationship to other list items). It was reported that when foils were endorsed with remember responses, the most common explanation participants gave was that they had recollected the word because it was related to others on the list. In contrast, when words that were actually studied had been endorsed as remembered, participants most often reported specific sensory or contextual associations that were related to the experience of having studied those items. The same investigators reported that older adults were more susceptible to false recognition in the context of the DRM paradigm, a result that was attributed to poorer memory monitoring processes among old versus young adults.

Summary and Conclusions

The recognition memory literature has provided a number of important insights about how exactly one distinguishes studied from novel materials, and has pointed to factors that might lead one astray. Quantitative descriptions of recognition memory performance can be obtained by using SDT to analyze recognition memory data, and a number of recognition memory models provide mathematically and conceptually precise characterizations of recognition memory findings. This article was written to provide readers with an overview of the recognition memory literature and highlights a subset of the important methodological and theoretical advances that have been made over the course of the last two or three decades. An exhaustive supplemental reading list cannot be provided here, but excellent suggestions have been made by Professor Ken Malmberg. Those who are interested in learning more about recognition memory are referred to his Web site: http://uweb.cas.usf.edu/~malmberg/Recognition%20Readings.htm.

See also: Declarative Memory, Neural Basis of; Learning and Memory, Neural Basis of; Signal Detection Theory, History of; Signal Detection Theory.

Bibliography

Cary, M., Reder, L.M., 2003. A dual-process account of the list-length and strength-based mirror effects in recognition. Journal of Memory and Language 49, 231–248.

Clark, S.E., Gronlund, S.D., 1996. Global matching models of recognition memory: how the models match the data. Psychonomic Bulletin & Review 3, 37–60.

Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behavior 11, 671–684.
Craik, F.I.M., Tulving, E., 1975. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General 104, 268–294.
Deese, J., 1959. On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology 58, 17–22.
Diana, R.A., Reder, L.M., 2006. The low-frequency encoding disadvantage: word frequency affects processing demands. Journal of Experimental Psychology: Learning, Memory & Cognition 32, 805–815.
Eagle, M., Leiter, E., 1964. Recall and recognition in intentional and incidental learning. Journal of Experimental Psychology 68, 58–63.
Glanzer, M., Adams, J.K., 1985. The mirror effect in recognition memory. Memory & Cognition 13, 8–20.
Godden, D.R., Baddeley, A.D., 1975. Context-dependent memory in two natural environments: on land and underwater. British Journal of Psychology 66, 325–331.
Godden, D., Baddeley, A., 1980. When does context influence recognition memory? British Journal of Psychology 71, 99–104.
Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.
Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.
Jacoby, L.L., Dallas, M., 1981. On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General 110, 306–340.
Jacoby, L.L., Whitehouse, K., 1989. An illusion of memory: false recognition influenced by unconscious perception. Journal of Experimental Psychology: General 118, 126–135.
Juola, J.F., Fischler, I., Wood, C.T., Atkinson, R.C., 1971. Recognition time for information stored in long-term memory. Perception and Psychophysics 10, 8–14.
Kroll, N.E.A., Yonelinas, A.P., Dobbins, I.G., Frederick, C.M., 2002. Separating sensitivity from response bias: implications of comparisons of yes–no and forced-choice for models and measures of recognition memory. Journal of Experimental Psychology: General 131, 241–254.
Loftus, E.F., 2005. Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. Learning & Memory 12, 361–366.
Malmberg, K.J., Nelson, T.O., 2003. The word frequency effect for recognition memory and the elevated-attention hypothesis. Memory & Cognition 31, 35–43.
Mandler, G., Pearlstone, Z., Koopmans, H.J., 1969. Effects of organization and semantic similarity on recall and recognition. Journal of Verbal Learning and Verbal Behavior 8, 410–423.
Mandler, G., 1980. Recognizing: the judgment of previous occurrence. Psychological Review 87, 252–271.
Memon, A., Bruce, V., 1983. The effects of encoding strategy and context change on face recognition. Human Learning: Journal of Practical Research & Applications 2, 313–326.
Murane, K., Phelps, M.P., Malmberg, K., 1999. Context-dependent recognition memory: the ICE theory. Journal of Experimental Psychology: General 128, 403–415.
Norman, K.A., Schacter, D.L., 1997. False recognition in younger and older adults: exploring the characteristics of illusory memories. Memory & Cognition 25, 838–848.
Parks, C.M., Yonelinas, A.P., 2007. Moving beyond pure signal-detection models: comment on Wixted (2007). Psychological Review 114, 188–202.
Rainis, N., 2001. Semantic contexts and face recognition. Applied Cognitive Psychology 15, 173–186.
Roediger, H.L., McDermott, K.B., 1995. Creating false memories: remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 803.
Smith, S.M., Vela, E., 2001. Environmental context-dependent memory: a review of meta-analysis. Psychonomic Bulletin & Review 8, 203–220.
Stanislaw, H., Todorov, N., 1999. Calculation of signal detection theory measures. Behavioral Research Methods, Instruments & Computers 31, 137–149.
Thomson, D.M., Tulving, E., 1970. Associative encoding and retrieval: weak and strong cues. Journal of Experimental Psychology 86, 255–262.
Wells, G.L., 1993. What do we know about eyewitness identification? American Psychologist 48, 553–571.
Wixted, J.T., 2007. Dual-process theory and signal-detection theory of recognition memory. Psychological Review 114, 152–176.
Wixted, J.T., Stretch, V., 2004. In defense of the signal detection interpretation of remember/know judgments. Psychonomic Bulletin & Review 11, 616–641.
Yonelinas, A.P., 1994. Receiver-operating characteristics in recognition memory: evidence for a dual-process model. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 1341–1354.
Yonelinas, A.P., 2002. The nature of recollection and familiarity: a review of 30 years of research. Journal of Memory and Language 46, 441–517.
Yonelinas, A.P., Aly, M., Wang, W.C., Koen, J.D., 2010. Recollection and familiarity: examining controversial assumptions and new directions. Hippocampus 20, 1178–1194.
Yonelinas, A.P., Jacoby, L.L., 2012. The process-dissociation approach two decades later: convergence, boundary conditions, and new directions. Memory & Cognition 40, 663–680.
Yonelinas, A.P., Parks, C.M., 2007. Receiver operating characteristics (ROCs) in recognition memory: a review. Psychological Bulletin 13, 800–832.

Relevant Websites

http://wise.cgu.edu/sdtmod/ – Discussion of Signal Detecton Theory: .
http://psychology.ucdavis.edu/Labs/Yonelinas/PWT/index.cfm?Section=9 – Excel Worksheets for DPSD and PDP.
http://uweb.cas.usf.edu/~malmberg/Recognition%20Readings.htm – Recommended Recognition Memory Readings.

Recognition, Politics of: Perspectives from Social Work and Sociology

Stephen A Webb, Glasgow Caledonian University, Glasgow, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In critically reviewing the literature on the politics of recognition, this article situates concerns for social justice within fundamental problems of acceptance, dignity, and worth. It outlines key contributions to contemporary discussion about recognition theory and the contours of debate within sociological thought. The article demonstrates how social work is an example that illustrates the applicability of this ethical turn to a concrete case. In doing so, it reveals the explanatory power of the redistribution/recognition interaction by connecting socioeconomic and sociocultural dimensions of injustice. The article shows how along with other public good services, reevaluating social work in light of the political applicability of recognition theory can be an integral part of strategies for a just society (Macdonald, C., Merrill, R., 2002. It Shouldn't Have to be a Trade: Recognition and Redistribution in Care Work Advocacy. Hypatia 17 (2), 67–83.).

Introduction

The politics of recognition has been described as a 'turn to ethics' (Garber et al., 2000) whereby the source of social and political conflicts is seen to be rooted in social suffering. The concept of 'social suffering' is used to refer to the lived experience of pain, damage, injury, deprivation, and loss caused by misrecognition. Harm is to be viewed from the standpoint of recognition, and its absence. Recognition theory occupies a privileged place among differing conceptions of ethical worth because it does not start from principles of justice generated by thought experiments, but from the empirical needs of human beings for recognition in the form of legal respect, social care, and distributed worth. These needs can give rise to "struggles for recognition that always take place 'within' the social framework of an inherited cultural horizon" (Honneth, 1995: p. 134).

Our social and political life, with its basic rights and liberties, is thought to be protected by laws of the modern state and yet seemingly with a blink of the eye it can be so quickly transformed into what Agamben (1995) calls naked or bare life, which strips back vital forms of life. Indeed, by decree has become common since World War I in all modern states. With this notion of bare life, the central paradox posed for any progressive politics is not only the erasure of cultural and political distinctions, but also the extreme negative differentiation or alienation produced by such erasure with regard to the worth and value of a people whose form of life has been destroyed. Mobilized by an empire that prides itself on freedom, liberty, and democracy, we have witnessed the apparatus of bare life at work in Guantanamo Bay. Recognition, its political and ethical applications, promises to act as a significant counterpoint to the zones of indistinguishability so starkly exposed by Agamben. For some, recognition offers the potential for a new paradigm of social justice where a difference-friendly world is no longer subject to the assimilation of dominant economic norms or the politics of the majority. Examples include claims for the recognition of distinctive attributes of sexual preference, minority religious practices, and gendered work (Fraser, 1997).

Situating Recognition and Its Political Application

In providing a robust definition, Brandom (2009) sets out parameters for establishing the social and political contours of recognition. Starting from the premise that recognition requires not only that someone recognized by another person or group, but that the person being recognized regards the legitimacy of the other in conferring said recognition. This approach rests on the view that recognition is authorized or conferred on a person. One of the problems in providing an adequate definition of recognition relates to questions about its responsive or generative nature (Markell, 2007). The latter model concentrates on the ways in which recognition reproduces or generates justifications for actions and identity formation. In recognizing, a process of attunement occurs involving a movement of self-adjustment to a disposition or perspective, whereby a person transforms herself in the act. A conscious performance of recognition enables understanding of the processes by which we constitute ourselves and others in a shared horizon of meaning. A good enough parent, for example, is "someone who ought to act in a certain way in virtue of being recognised as, for example, recognising someone as a rational being will generate certain duties and responsibilities for both the person being recognised and those who interact with her" (McQueen, 2011: http://www.iep.utm.edu/recog_sp/). The response model of recognition is bound to the dispositions that someone possesses and normatively articulates in a way that may be subsequently affirmed. Here, to recognize someone is to acknowledge them as they already really are (Appiah, 1994: p. 149). This means that there are justifications for why one ought to confer recognition to someone prior to the act of recognition itself. Thus, for example, we ought to recognize someone's ability to be a good parent because they possess certain dispositions, of bonding and caring.

Recognition theory is thought to be especially well suited to provide for the political analysis of relations of power and associated mechanisms of social and political resistance. "As experiences of misrecognition violate the identity of subjects, the affected are supposed to be particularly motivated to resist, that is, to engage in a 'struggle for recognition'" and as such "promise to illuminate a variety of new social movements – be it the struggles of ethnic or religious minorities, of gays and lesbians or of people with disabilities" (Mattias, 2013: p. 1). Within the sphere of human rights such struggles are supposed to produce ever more just and fulfilling relations of recognition, with the demand for equal recognition of worth as the ultimate goal.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28069-0

The history of the idea of recognition stems back to the foundational work of the German philosopher Georg Hegel. For Hegel, the struggle for recognition found through an intense and enduing tension in the master–slave relation (*Anerkennung*) was famously expressed as "Self-consciousness attains its satisfaction only in another self-consciousness" (1967: p. 173). If self-consciousness exists only when recognized as such, or better still, is satisfied with an external reflection of its own freedom in the other, it is necessarily caught up in the reciprocal movement, or dialectic, in order to be certain of itself. To illustrate the dialectic of recognition, Hegel uses the custom of the greeting as a concrete example of the reciprocity at stake. Hegel has undoubtedly influenced the contemporary understanding of recognition as a political morality more than any other philosopher.

The idea of the demands for recognition features prominently in the work of leading social theorists such as Charles Taylor, Axel Honneth, Nancy Fraser, Iris Young, and Lois McNay. Each is concerned with the centrality of issues of identity in modern society. Thus the politics of recognition agenda regards justice demands as about more than the fair distribution of material goods. Young explains, "People have or lack self-respect because of how they define themselves and how others regard them, because of how they spend their time, because of the amount of autonomy and decision making power they have in their activities, and so on. Some of these factors can be conceptualized in distributive terms, but others cannot. Self-respect is at least as much a function of culture as it is of goods" (1990: p. 27). Crucially, this perspective gained salience in highlighting that concerns for recognition are primarily attempts to compensate for an underlying loss of or a displaced sense of identity. Taylor describes this critical relation between recognition and identity as follows:

> The demand for recognition in these latter cases is given urgency by the supposed links between recognition and identity, where this latter term designates something like a person's understanding of who they are, of their fundamental defining characteristics as a human being. The thesis is that our identity is partly shaped by recognition or its absence, often by the *mis*recognition of others, and so a person or group of people can suffer real damage, real distortion, if the people or society around them mirror back to them a confining or demeaning or contemptible picture of themselves.
>
> 1994: p. 25

With Taylor's Hegelianism, we witness the possibility of the achievement of identity as accorded to each by the other. The fundamental insight that subjectivity is dialogical in nature, that it is created only through interaction with an Other permits social science to recast a model of social justice around the rich life world of everyday existence. While reproduced and manifest at level of the individual, commentators insist on the profoundly collective characteristic of identity. Indeed, it is argued that the significance given to identity in the politics of recognition lies in the strength it has with affinities people have with chosen groups (e.g., organized protest at a natural gas fracking site). Arguably, it is group affiliation and networked status that define our identities and not our individual identities. Since our identity is constituted through relations with others, our being recognized by them, feelings of value, worth, and self regard are most likely to be sustained, if we are positively recognized for who we are. Indeed, in differing ways thinkers such as Taylor, Fraser, and Honneth deploy the idea of recognition to develop accounts of the individual, which are opposed to the individualism of liberal thought and to the psychologization associated with studies of self-realization and subjectivity. The idea of recognition expresses the notion that individuality is an intersubjective phenomenon formed through concrete networked relations with others. Most theories of recognition assume that in order to sustain a concrete identity, persons fundamentally depend on the feedback of other people and of society as a whole. Accordingly, those who experience a deficit of conferred recognition i.e., people or groups who are depicted negatively or perceived as worthless by societal norms and values, will find it much harder to pursue life projects in a positive way. However, distinct forms of recognition are possible only within a certain kind of community. Unlike modern Western society in which beggars are regarded as illegal scroungers in other earlier communities, mendicant orders were honored and highly regarded because they actively refused property ownership. The mendicants were destitute through choice. However, so far there has been very little analysis of the connection between recognition and the constitutive power of networks or groups or historical comparisons on the situated nature of recognition.

Recognition or Redistribution? The Fraser–Honneth Debate

Fowler suggests the redistribution and recognition debate, originally conducted by Nancy Fraser and Axel Honneth (2003), and involving Seyla Benhabib, Jessica Benjamin, and Charles Taylor, promises to be as significant "in the 21st century as the Protestant Ethic debate or the Althusserians/Culturalists debate were in the 20th century" (2009: p. 144). A key issue in the rich exchange between Fraser and Honneth is whether we need to substantively differentiate forms of social suffering: on the one hand, the denial of recognition for certain social groups who are subjected to exploitation and prejudice; and on the other hand, the consequences of vast disparities of stratified wealth and income.

The significance of developing an integrated conception of justice in the social sciences began in 1995 with the publication of articles by Fraser and Honneth. Lovell's *(Mis)recognition, social inequality and social justice* (2007a) and Fraser and Honneth's *Redistribution or Recognition?: A Political-Philosophical Exchange* (2003) trace the recent history and show the significant contribution they have made to a debate running for over a decade. Fowler (2009) claims that while other sociologists such as Pierre Bourdieu, Marion Young, and Charles Taylor have made valuable contributions to this debate, the influence of Fraser and Honneth has been by far the most compelling. At the risk of offering a limited account of the complexities entailed in the redistribution and recognition debate, the concentration here lies with Fraser and Honneth and the way this has been taken up in the valuable collection of essays edited by Lovell (2007a).

The struggle for recognition debate rests on an essential paradox of the human condition. As human beings, we need each

other in order to become human and to sustain our lives as meaningful projects. Our identities are primarily formed by processes of intersubjective recognition, which are then sustained by relations and institutions which consolidate these forms of intersubjectivity. And yet, at the same time we threaten one another's existence. Violence, neglect, abuse, or isolation of others is a permanent feature of human history. In short, we are deeply embedded in struggles for recognition (Thompson, 2006).

Honneth grounded his 'recognition' model in Hegel's early Jenna writings. Here, the idea of a fundamental unity between people, as intersubjectivity, is framed, and one that precedes their understanding of themselves as separate beings. A core preoccupation lies with how subjects related to themselves to enhance their sense of psychological well-being as a result of how significant others reacted to them. For Honneth, there are three types of recognition which lead to ontological well-being for the individual self, namely, recognition of the subject's (1) need for love and care, (2) rights as a human being, and (3) strengths or more precisely, the contribution he or she makes to the community. He regards role taking and internal conversations about expectations of our behavior were formative processes in the recognition of the other (Houston, 2013: p. 80). However, Honneth's reconstruction of social theory consists of a combination of what is most fruitful in Foucault and Habermas. From Foucault, he not only takes the embedded nature of conflictual social life, to be understood at the institutional level, but also the ever presence of power relations at the cultural, symbolic, and microphysical level of bodies and dispositions. While from Habermas, he takes the normative preoccupations with morality and intersubjective aspects of communication to evaluate the prospects for an emancipatory project (Deranty and Renault, 2007: p. 95). Honneth deliberately avoids using the term 'a politics of recognition.' His reluctance to discuss the political and focus on the ethical is justified within his theory. As Deranty and Renault note, "The driving intuition of his model is that social progress is based on the normative expectations of individuals, which must be construed as moral claims, rather than as socio-economic interests" (2007: p. 92). The struggle for recognition thus has at its root an Aristotelian teleology in the form of an 'ethical good life.'

For Fraser, modern society is comprised of two empirically interrelated but analytically distinct orders of stratification: an economic order of distributive relations that generate inequalities of social class, and a cultural order of recognition relations – relating to gender, ethnicity, age, and sexuality – that generate inequalities of status. She argues that although analytically distinct, these orders are deeply intertwined: "Even the most material economic institutions have a constitutive, irreducible cultural dimension they are shot through with significations and norms. Conversely, even the most discursive cultural practices have a constitutive, irreducible political economic dimension, they are underpinned by material supports" (1995: p. 720). As we shall see, against identity politics, Fraser's Weberian turn endorses a 'status model' in which recognition means participation in the dominant culture without discrimination, locating both recognition and misrecognition in structural elements of the wider society, rather than in the activity of the subject itself. A stark example of such forms of misrecognition exists among the 'precariat' black American ghettos and the French *banlieue*, who Wacquant notes are "compelled to form its subjectivity out of its objectification by others" (2007: p. 73).

Honneth's main line of criticism against Fraser is her dualism and the duplication of what he regards as a tired Marxist base–superstructure dichotomy. He argues that while redistribution is essential to remedy material inequalities and injustice, it can be subsumed as a subset of the politics for recognition. In short, he aims to show that the end point of Fraser's felt injustice expressed in communities, is above all, a feature of the injustice of disrespect (Lovell, 2007b: p. 5). For Honneth, Fraser's specifically separate category of economic injustice is surplus to requirements for an adequate theory of social justice. He privileges asymmetrical forms of recognition as providing the foundation of all economic inequalities and system stratifications. In Honneth's account, to be recognized *is* to be treated justly. He also accuses Fraser of being overreliant on those recognition struggles that have been publically mobilized through new social movements. Honneth maintains that the key reference point for injustice is experiential, it is what is experienced as injustice and manifest as 'psychic suffering' that is the most significant marker of exploitation and domination. In response, Fraser argues that Honneth's reductionism is inferior to her own 'perspectival dualism' in which recognition is one crucial but limited component of social justice (2003). She insists in integrating redistribution and recognition "by treating them as irreducible dimensions of a single, overarching idea of justice, which is expressed in the norm 'parity of participation'" (Markell, 2008). This notion of 'parity of participation,' "requires social arrangements that permit all to participate as peers in social life" (Fraser, 2005: p. 73) and is key to Fraser's framework of social justice (Davies et al., 2013). Her model of parity of participation challenges the undue emphasis on identity-focused approaches to participation, whereby the "result is often to impose a single, drastically simplified group identity, which denies the complexity of people's lives, the multiplicity of their identifications, and the cross-pulls of their various affiliations" (2001: p. 24). In doing so, it reveals the ways in which asymmetries of power and cultural misrecognition manifest themselves in deliberative processes.

In summary, a politics of recognition for Fraser can address cultural needs and differences but not the deep cleavages in political economy: class-based divisions, inequalities in living standards, and material disadvantage. Honneth's model is tipped overly toward questions of cultural identity. Indeed, the prevailing preoccupation with difference and celebrations of diversity in social science rests on what Fraser regards as a perverse identity politics that have the effect of displacing redistribution as a principle against economic inequalities and reifying identity. She draws attention to significant problems associated with identity politics.

> Having begun by assuming that identity is dialogical, constructed via interaction with another subject, it ends by valorizing monologism – supposing that misrecognized people can and should construct their identity on their own. It supposes, further, that a group has the right to be understood solely in its own terms – that no one is ever justified in viewing another subject from an external perspective or in dissenting from another's self-interpretation (2005: p. 112).

Thus, for Fraser, identity politics and its concomitant celebration of difference encourages separatism and rigid group enclaves. As we have seen, recognition as a form of identity politics rests on an assumed fundamental relation of self to the Other originally developed by Hegel in his master–slave dialectic. Spanning a long line of philosophers of difference from Hegel, Lacan, Derrida, and Levinas, the feminist philosopher Judith Butler maintains "… there is no self prior to its persecution by the Other. It is that persecution that establishes the Other at the heart of the self, and establishes that 'heart' as an ethical relation of responsibility" (2000: p. 25). 'Othering' is the process that is believed to create a 'them–us' dichotomy or division between people to include those in the 'us' category and exclude those in the 'them' group because they are deemed inferior.

McNay (2007) takes issue with both Fraser and Honneth in arguing that the insights of recognition theorists are undercut by their reliance on an inadequate sociological account of power thus positing a thin political solution to deeply embedded structural problems. She is struck by the way in which debates about identity politics, and identity and recognition, has been gradually overwhelmed by concerns, which in a previous era would have been seen as psychological concerns of the self; things to do with self-discovery and self-realization. The idea of recognition relies on an account of social relations as extrapolations of a primal dyad, or face-to-face dynamic of interaction that overlooks the complex ways in which individuality is connected to abstract social structures in contemporary society. Moreover, recognition is pervasively naturalized and universalized in such a way as to foreclose anything but a narrow understanding of the reproduction of structural inequalities (2007: p. 3). McNay is insistent that liberal-based consensus recognition is only a 'temporary stabilization of power,' which always involves some kind of exclusion. Using Bourdieu's relational sociology, McNay develops an alternative account of individual agency that connects identity to structure. In focusing on issues of gender identity and agency, she opens up fertile pathways to move beyond the oppositions between material and cultural feminisms conveyed in the politics of recognition.

Recognition as a Matter of Concern for Social Work

This section critically reviews the value of politics of recognition in its potential to successfully translate to a substantive field of social intervention, namely, modern social work. Both Garrett and Houston have generously outlined the main contours and contributions of the redistribution/recognition debate, particularly those of Fraser and Honneth, as well as drawing attention to the rich implications these have for social work. Given their significant in-depth treatments any reiteration, other than to outline the broad points of convergence and disagreement, would be superficial. Therefore the coverage here will provide only a schematic overview of the debate in the treatment by Houston (on Honneth) and Garrett (on Fraser and Honneth) (see Webb, 2013).

In a number of important articles, Houston has extolled the virtues of Honneth's cultural theory of recognition, which he regards as providing an ethical framework "for self-realization through symbolic interaction and mutual recognition" (2008a: p. 13). The critical leverage of such a framework is that it has the capacity to move 'individualized' direct work in social work to one 'characterized by symbolic interaction, care, respect, and validation' leading to optimal self-realization (Houston, 2008a). Honneth recruits the theoretical aid of Maslow's hierarchy of needs to cast a progressive version of self. From the standpoint of the potential of human capabilities, society is conceived as an institutional framework that either promotes or hinders individual growth (Honneth et al., 1988). However, the main problem with Maslow's (1962) approach is that individual needs are innate, so that questions of social interaction and culture are seriously downgraded. In his critique of 'individualization' in social work, he helpfully contextualizes the ethics of recognition in the practice dimensions of: (1) social work as 'symbolic interaction,' (2) social work as 'care,' (3) social work as 'respect,' and (4) social work as 'validation' (2008b: p. 8). For Houston these four dimensions are constitutive of an ethical social work that frames the person in a rich interactive social context while embodying a normative set of moral ideals. He effectively pairs the ethical turn in social sciences with a normative moral agenda of social work intervention. Self-realization is inevitably bound up with matters of recognition that are versions of individual life politics rather than universal standards of social justice. Houston considers that Honneth's recognition model can be applied in a range of social work contexts, for example, young care leavers form a group to highlight their relative disadvantage, socially and economically; adults with a sensory impairment seek to realize better practical services from a local authority; or a group of travelers fight to preserve their distinct cultural identity and choice of lifestyle (2013: p. 85). He shows how self-directed groupwork can be captured as an empowering process targeting service users' struggles for recognition centering on their experiences of disrespect, denigration, shame, outrage, and humiliation. Houston argues that social workers must be receptive to service users advocating for public recognition of their particular needs, concerns, identities, and ways of being (2013: p. 85). Davies et al. (2013) have provided empirical support for Fraser's model of parity of participation in a study of mental health and homelessness service users. The inadequacies of rights-based and consumerist approaches to service user participation are shown to reproduce institutionalized forms of cultural value and fail to engage in actual decision-making processes.

For Garrett (2009), the formality of the model of ethical life espoused by both Honneth and Houston is synonymous with a kind of political agnosticism. In Honneth's case, the failure in his conception of subjectivity is problematic, marked as it is by an unbridgeable gulf between an individual subject conceived in psychological terms, and social action conceived in abstract, general terms. Garrett is concerned that Houston reproduces these apolitical tendencies for social work. He notes that in Houston's account "the role that the more powerful structural forces play—internal and external elites, occupying forces—in promoting division and partitioning communities: generating, bolstering, structuring relations of (mis)recognition and social difference"

(p. 13) is strangely absent. Garrett puts this down to perverse blind spot in cultural recognition politics on the hegemonic position of the modern State and the role it plays in reproducing inequalities of recognition. There are two significant contributions the State performs in structuring difference: firstly "it is responsible for the discursive promulgation of categories and labels attached to particular groups" (Garrett, 2009); secondly, across its territorial jurisdictions, it "can be regarded as a literal builder and consolidator of 'difference' which is achieved via spatial and territorial separating practices (under the guise of 'peace' walls or, more routinely, of course, in terms of the control of 'borders'" (p. 14). In part, as a vehicle of the apparatus of the State it would be important to reflect on the way that social work maintains or constructs these inequalities of recognition. It would also be instructive to identify how oppositional anti-State politics of recognition are mobilized by social work and further afield. In this respect, it is important to acknowledge the ambivalent role of social work in relation to the State; not only it is complicitous, as a functionary apparatus, but also decidedly hostile to its machinations (Parton, 2008; Webb, 2009).

Garrett reaches the conclusion that "there is a case for embedding the politics of recognition into social work education and practice: perhaps particularly, the theorization of Nancy Fraser, whose perspective, although mostly neglected in the social work literature, is preferable to the narrower approach of Honneth, which, although achieving dominance within social work's engagement with recognition, is prone to 'psychologization' and psychological determinism" (p. 13). Ironically, McNay – whom Garrett mobilizes, in defense of Fraser, to criticize Honneth's subjectivist politics of recognition – condemns Fraser for the tendency to downplay the significance of experiential and interpretative perspectives. McNay (2008) claims that "Fraser's paradigm is deprived of a perspective upon the submerged experiences of everyday life that precede any politics of recognition. The move away from a focus on subjective experience means that she cannot adequately address pre-political and hidden forms of social suffering" (p. 285). As in the case, for example, of Afghan women who set themselves on fire to escape brutal beatings from husbands and forced marriages; or the thousands of farmers in India who commit suicide because they cannot afford to pay off extortionate bank loan interest. McNay is right to elevate the significance of deeply embedded forms of subjective suffering that never make it into the rights claims of recognition politics. It is relevant to note that social workers intervening in domestic violence, child abuse, families facing poverty, progressive illness in old age, and people with severe mental health problems face this sort of distress on a daily basis. However, Garrett's general point is that in social work we are tending more and more to reinterpret problems that might have been seen to have socioeconomic causation in the past, or problems that in the past would have been interpreted as social in character. The prevailing tendency – dangerously perhaps – is now to frame these material problems as having little more than a psychological dimension to them. In this respect, going full circle back to Hegel, it is worth underlining what Williams notes "For Hegel, the struggle for recognition is primarily about sociopolitical and cultural history, not about personal psychological development" (1997: p. 386).

Conclusion

Houston and Garrett have made a significant contribution to reconceptualizing politics and justice in social work drawing on recognition theory as a source of inspiration. Taken together showing that the task of developing an integrated strategy for social justice in social work is not a job for individual theorists but rather a collective project for researchers, practitioners, and service users engaged in critical deliberation. In extrapolating from both Honneth's and Fraser's contributions, a perspective on social justice must have as its object the correspondence of two homologous structures: the ethics of recognition, concerned with cultural injustice and suffering; and the ethics of redistribution, focused on redressing economic injustice – a field of power relations that is unavoidably an arena of struggle. Fraser (2003) argues that to pose an either/or choice between the politics of redistribution and the politics of recognition is to posit a false antithesis. On the contrary, justice today requires both redistribution and recognition. Indeed, the strategies of social workers engaged in struggles for social justice depends on the position they occupy in the structure of the field of redistribution and/or recognition. These are best evaluated on the basis of the degree to which it redistributes and/or recognizes and in doing so subverts existing dominant rule within the State apparatus and its forms of neoliberal economy. Radical transformation of society toward a more just world, however, ultimately rest on factors other than calculated interests and the dependency on discrete others as its object. It rests on the emotional anticipation of what Ernst Bloch (1954) called the 'principle of hope.'

See also: Critical Social Work Practice; Feminist Research and Social Work; Globalization: Welfare and Social Work; Neoliberalism; Oppression; Radical Social Work; Social Work Ethics; Social Work Theory.

Bibliography

Agamben, G., 1995. Homo Sacer: Sovereign Power and Bare Life. Stanford University Press, Stanford, CA.

Appiah, K.A., 1994. Identity, authenticity, survival: multicultural societies and social reproduction. In: Gutmann, A. (Ed.), Multiculturalism: Examining the Politics of Recognition. Princeton University, Princeton.

Bloch, E., 1954. The Principle of Hope. The MIT Press, Cambridge, MA.

Brandom, R., 2007. The structure of desire and recognition: self-consciousness and self-constitution. Philosophy & Social Criticism 33 (1), 127–150.

Brandom, R., 2009. Reason in Philosophy: Animating Ideas. Harvard University Press, Cambridge, MA.

Butler, J., 2000. Ethical ambivalence. In: Garber, M.B., Hanssen, B., Walkowitz, R.L. (Eds.), The Turn to Ethics. Routledge, London.

Davies, K., Gray, M., Webb, S.A., 2013. Putting the parity into service-user participation: an integrated model of social justice. International Journal of Social Welfare. http://dx.doi.org/10.1111/ijsw.12049. Article first published online: 11 July 2013.

Deranty, J.P., Renault, E., 2007. Politicizing Honneth's ethics of recognition. Thesis Eleven 88, 92–111.

Fraser, N., 1995. From redistribution to recognition? Dilemmas of justice in a 'Post-Socialist' age. New Left Review 212, 68–93.

Fraser, N., 1997. Justice Interruptus: Critical Reflections on the Postsocialist Condition. Routledge, London.

Fraser, N., May–June 2005. Rethinking recognition. New Left Review 3, 107–120.

Fraser, N., Honneth, A., 2003. Introduction: redistribution or recognition? In: Fraser, N., Honneth, A. (Eds.), Redistribution or Recognition?: A Political-Philosophical Exchange. Verso, London.

Fowler, B., 2009. The recognition/redistribution debate and Bourdieu's theory of practice. Theory, Culture and Society 26 (1), 144–156.
Garrett, P.M., 2009. Recognizing the limitations of the political theory of recognition: Axel Honneth, Nancy Fraser and social work. British Journal of Social Work 40 (5), 1517–1533.
Garber, M.B., Beatrice Hanssen, B., Walkowitz, R.L. (Eds.), 2000. The Turn to Ethics. Routledge, London.
Gray, M., Webb, S.A., 2008. The return of the political in social work. International Social Welfare 18 (1), 111–115.
Hegel, G.W.F., 1967. The Phenomenology of Mind. Harper Torch Book, New York.
Honneth, A., Joas, H., Meyer, R., 1988. Social Action and Human Nature. University of Cambridge Press, Cambridge.
Honneth, A., 1995. The Struggle for Recognition: The Moral Grammar of Social Conflicts (J. Anderson, Trans.). Polity Press, Cambridge.
Honneth, A., 2012. The I in We: Studies in the Theory of Recognition. Polity Press, London.
Houston, S., 2008a. Beyond homo economicus: recognition, self-realization and social work. British Journal of Social Work 40 (3), 841–857.
Houston, S., 2008b. Transcending ethno-religious identities in Northern Ireland: social work's role in the struggle for recognition. Australian Social Work 61 (1), 25–41.
Houston, S., 2013. Honneth's Theory of Recognition for Social Work. University of Sussex, Brighton. Keynote.
Houston, S., Dolan, P., 2008. Conceptualising child and family support: the contribution of Honneth's critical theory of Recognition. Children and Society 22 (6), 458–469.
Lovell, T., 2007a. (Mis)recognition, Social Inequality and Social Justice. Routledge, London.
Lovell, T., 2007b. Nancy Fraser's integrated theory of justice: a 'sociologically rich' model for a global capitalist era. Law, Social Justice and Global Development 1. http://www.go.warwick.ac.uk/elj/lgd/2007_1/lovell.
Markell, P., 2003. Bound by Recognition. Princeton University Press, Princeton, NJ.
Maslow, A.H., 1962. Towards a Psychology of Being. Van Nostrand Reinhold, New York.
Markell, P., 2008. Recognition and redistribution. In: Dryzek, J.S., Honig, B., Phillips, A. (Eds.), The Oxford Handbook of Political Theory. Oxford University Press, Oxford.
Markell, P., 2007. The potential and the actual: Mead, Honneth, and the "I". In: Brink, B., Owen, D. (Eds.), Recognition and Power: Axel Honneth and the Tradition of Social Theory. Cambridge University Press, Cambridge.
Mattias, I., 2013. Recognition. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy. Fall 2013 ed. http://plato.stanford.edu/archives/fall2013/entries/recognition/.
Macdonald, C., Merrill, R., 2002. It shouldn't have to be a trade: recognition and redistribution in care work advocacy. Hypatia 17 (2), 67–83.
McQueen, P., 2011. Social and political recognition. Internet Encyclopedia of Philosophy (last accessed on 01.09.14.). http://www.iep.utm.edu/recog_sp/.
McNay, L., 2007. Against Recognition. Polity Press, London.
McNay, L., 2008. The trouble with recognition: subjectivity, suffering, and agency. Sociological Theory 26 (1), 271–296.
Parton, N., 2008. Changes in the form of knowledge in social work: from the 'social' to the 'informational'? British Journal of Social Work 38 (2), 253–269.
Taylor, C., 1994. The politics of recognition. In: Goldberg, D.T. (Ed.), Multiculturalism. A Critical Reader. Blackwell, Oxford.
Thompson, S., 2006. The Political Theory of Recognition: A Critical Introduction. Polity Press, London.
Wacquant, L., 2007. Territorial stigmatization in the age of advanced marginality. Thesis Eleven 91, 66–77.
Webb, S.A., 2009. Against diversity and difference in social work: the case of human rights. International Social Welfare 18 (3), 307–316.
Webb, S.A., 2013. (Re)assembling the left in social work: the politics of redistribution and recognition in social work. British Journal of Social Work 40 (8), 2364–2379.
Williams, R.R., 1997. Hegel's Ethics of Recognition. University of California Press, Berkeley.
Young, Iris Marion, 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Reconstructive Memory, Psychology of

Henry L Roediger, III, and Kurt A DeSoto, Washington University, Saint Louis, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Memories are reconstructions of past events, not perfect recordings. These reconstructions can be accurate, but may also contain errors. Since the early 1900s, researchers have investigated memory processes in the laboratory and have demonstrated that reconstructions occur regardless of whether memories are retrieved after short or long periods. Reconstructive errors in memory retrieval have been shown in memory for basic objects (e.g., words, prose, pictures) and naturalistic materials (e.g., scenes, events). The implication is that understanding how memories are recollected requires considering the complex interaction of individuals' preexisting knowledge, beliefs, experiences, and goals.

Cognitive processes are active. When we perceive and encode events in the world, we construct (rather than copy) the outside world as we comprehend the events. If perceiving is a construction, then remembering the original experience involves a reconstruction. *Reconstructive memory* refers to the idea that remembering the past requires an attempt to reconstruct the events experienced previously. These efforts are based partly on traces of past events, but also on our general knowledge, our expectations, and our assumptions about what must have happened. As such, recollections may be filled with errors when our assumptions and inferences, rather than traces of the original events, determine them. Errors – false memories – constitute the prime evidence for reconstructive processes in remembering. Several different sources of error (inferences during encoding, information we receive about an event after its occurrence, our perspective during retrieval) exist. Contrary to popular belief, memory does not work like a video recorder, faithfully capturing the past to be played back unerringly at a later time. Rather, even when our memories are accurate, we have reconstructed events from the past.

In sum, reconstructive memory refers to the idea that retrieval of memories does not occur in some completely accurate form, as a video might replay a scene, but rather that recollection of memories involves a process of trying to reconstruct past events. Although the reconstruction can be quite accurate, the processes responsible can also introduce errors during retrieval. In fact, systematic errors in memory are the primary evidence for its reconstructive nature. This article provides an overview of the evidence that remembering is reconstructive.

Background

The hypothesis that remembering should be viewed as reconstructive dates to an important book by Sir Frederic Bartlett (1932). Bartlett contrasted veridical, rote forms of memory, such as reproducing a telephone number – which he called *reproductive memory* – with reconstructive memory and argued that the latter was a more typical use of memory outside laboratory and educational circumstances. That is, we are rarely faced with the task of remembering something exactly the way it happened; more frequently, we need only to get the essence of the event right. If a friend asks, "What did Kathleen say last night?" the request is not for a literal rendering of last night's conversation, but rather for the gist of what she said.

Bartlett argued that perceiving and comprehending events do not simply happen automatically, but that every event of comprehension involves the mental construction of one's understanding of the event in the world. In cases of ambiguous or difficult events, we can struggle with an "effort after meaning" in comprehending (as when you overhear a puzzling conversation and have to retrieve some relevant context to understand it). If encoding or perceiving is a construction, then when one wants to recall the events later, the attempt requires the reconstruction of the event.

Bartlett argued that remembering is guided by schemas, or general organizing structures, which aid encoding and guide retrieval. If you see a scene at the beach and are asked to recall it later, you might recall seeing a beach umbrella even if none was present in the actual scene itself, because an umbrella is consistent with the general schema of items that belong in a beach scene. Therefore, although schemas can aid encoding and guide retrieval, they can also lead to errors.

According to the standard textbook account, certain forms of remembering are reconstructive (remembering your time in the sixth grade) whereas others are reproductive (remembering a phone number or your favorite poem). However, a large amount of research is consistent with the idea that nearly all remembering is reconstructive. In fact, reconstructive memory appears to occur even when the reconstruction is quite accurate (Gallo, 2006; Roediger and McDermott, 1995). This article considers various forms of memory as they are experimentally studied and discusses evidence for reconstructive processes at work.

Sensory and Short-Term Memory

The greatest challenge to the claim in the previous paragraph, that all remembering is reconstructive, may come from studies of sensory memory, defined as the borderline between perceiving and remembering. Many theorists have argued that perceiving is itself a constructive activity, with the fascinating phenomena of visual and auditory illusions used as evidence

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51016-2

for this claim (see Hoffman, 1998; among many others). Sensory memory refers to the temporary persistence of information that has struck the senses, which lingers briefly as it is being comprehended. Visual persistence is called iconic memory and auditory persistence is labeled echoic memory. It would seem that iconic memory – essentially a fleeting afterimage of the scene from the outside world – would surely be a form of reproductive memory. Yet even in these situations errors arise, showing that the retrieval processes from this type of memory involve reconstruction. For example, in Sperling's (1967) studies of iconic memory, in which people had to report letters that they had briefly seen on a screen, a common error when people missed a letter was to report another letter that either looked like or sounded like the original letter. This type of error indicates that people may code even such simple items as single letters into visual patterns and associated sounds. When people miss a *b* in these experiments, they may substitute a *v* (which sounds like *b*) or a *p* or *d*, which share both similar appearance (a long line and a curve) and similar sound (the letters rhyme with *b*). In sum, even reports from iconic memory may show reconstructive tendencies.

Short-term and working memories last longer than sensory memories do, but people are still ordinarily accurate in retrieving information from short-term stores if no interference occurs. Does this accuracy reflect a rote, reproductive process? The answer seems to be no, because when the short-term memory system is challenged by having people operate under, for example, fast rates of presentation, errors occur. Errors are often (but not always) phonological in nature. That is, if someone tries to recall letter strings and misses a letter, similar sounding letters are confused (Conrad, 1964). In the case of words, those that share visual and phonemic (sound-based) features are confused (Crowder, 1976; Chapter 4). Therefore, even though short-term memory processes are often considered quite accurate (and they can be), recall in these situations typically occurs under conditions that make for accurate reconstructions (e.g., with short unfilled delays between study and test). Stress these systems by presenting material quickly, or by creating interference, and the characteristic error pattern indicative of a reconstructive process appears.

Long-Term Memory

By far the greatest amount of research on reconstructive memory has occurred in the domain of long-term, episodic memory. Many different kinds of experiments have been conducted and we will consider some of the main types here, which have used a wide variety of materials: word lists, pictures and scenes, sentences, prose passages, and videos. Research with all these materials reveals remembrance to be quite prone to error and filling in the gaps.

Word Lists

Roediger and McDermott (1995) developed a paradigm in which people heard lists of 15 words that were generated as associates to another word, which was not itself presented. For example, subjects heard the list "hot, snow, warm, winter, ice, wet, frigid, chilly, heat, weather, freeze, air, shiver, Arctic, frost." All these words are associated to the word 'cold,' which was the word that generated the list (all the items in it were associates to 'cold'), but 'cold' was not presented. Immediately after hearing the list, subjects were asked to recall the items and were urged to be sure that the words they recalled had been presented in the list. Despite this warning not to guess, Roediger and McDermott found that their subjects recalled missing words like 'cold' at very high levels; in fact, the probability of recalling the critical item often equaled the probability of recall of the list items. After recall of many lists, subjects received a recognition test in which they were instructed to pick out only items that had actually appeared earlier in the experiment. The test included words from the lists (winter, chilly), completely unrelated words (automobile), and the critical words such as 'cold.' Subjects examined each word and, if they judged it to be old (i.e., studied), they were asked to make a remember/know judgment. This judgment (Tulving, 1985) asks people to report if they can remember the moment of the item's occurrence in the study list (a remember judgment) or if they just know it was on the list but cannot remember its moment of occurrence (a know judgment).

The recognition results are shown in Figure 1, where it can be seen that subjects were good at distinguishing studied words from unrelated lures or distracters. However, the critical lures such as 'cold' were recognized with the same likelihood as were the words in the list, about 80% of the time. In other words, people were just as likely to recognize the related but non-studied words as they were items in the list. In addition, subjects claimed to remember the moment of occurrence (as measured by remember judgments) of the critical words (that, of course, were not actually presented) at the same level as for the words that were actually studied. In short, in this paradigm subjects were just as likely to recall, recognize, and claim to remember the occurrence of words that were never presented as words that really were presented, which provides a compelling demonstration of illusory memory. The implication is that subjects reconstruct the list using their general knowledge and insert a specific memory – they remember 'cold' as having occurred – because it fits with the general theme or schema of the list.

The previous example demonstrates how reconstructive memories can be demonstrated with associative lists, but these memories can also be shown for semantic categories, as well. A chapter by Roediger et al. (2012) discusses work by Roediger and colleagues (Meade and Roediger, 2006, 2009; Roediger and DeSoto, 2013), who had subjects study category members belonging to different semantic categories (e.g., vegetables). Subjects studied 15 members from each category that were neither the most typical members of the category (e.g., carrot, lettuce, etc.) nor atypical (e.g., 'artichoke'). They were words like 'corn,' 'spinach,' and others. On a subsequent recognition test, however, subjects were highly likely to recognize highly typical category members that were never presented (e.g., carrot). These false memories were often accompanied by extremely high confidence – subjects were extremely certain that the highly typical category members were presented, even though they never were. Like the Roediger and

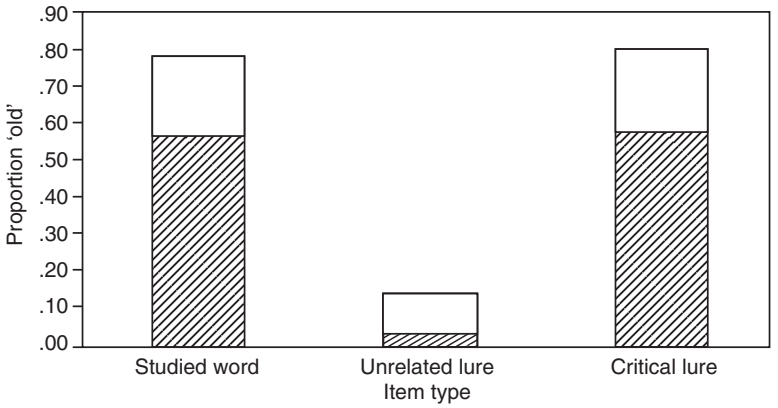

Figure 1 False recognition and false remembering. Mean probability of recognizing (total bar) and remembering (cross-hatched bar) studied words and lures. Unrelated lures were rarely recognized, but critical lures were recognized and remembered at about the same levels as studied words. People remember the occurrence of the critical items at the same level as words that were actually studied. Data reproduced from Roediger, H.L., McDermott, K.B., 1995. Creating false memories: remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 803–814 (Experiment 2).

McDermott (1995) study, this research provides evidence that semantic knowledge and schemas guide reconstructive processes that occur during remembering.

Prose Recall

A large body of research demonstrating reconstructive processes in retrieval has been carried out in prose recall. Brewer and colleagues (Brewer, 1977; Brewer and Sampaio, 2006; Brewer et al., 2005; Sampaio and Brewer, 2009) examined pragmatic implications in retention of sentences. A pragmatic implication is made when the person reading or hearing a sentence infers something that is neither explicitly stated nor logically implied by the sentence. For example, "The karate champion hit the cinder block" implies that the block was broken. However, this is a pragmatic implication; it is perfectly possible that the block was struck but not broken. After people studied sentences of this type mixed with other sentences, they were given the first parts of the sentences as cues to recall the entire sentence. Brewer (1977) showed that in recall of sentences with pragmatic inferences, people were more likely to recall the implied verb than the verb actually used in the sentence. When reconstructing the action, they remembered the champion as breaking the cinder block, not just hitting it. This result reflects a kind of inferential momentum, with people going beyond the literal truth of a statement to infer the events that probably occurred next. Often such inferences may be correct, but they are not necessarily so. When the inference is strongly made, but false, it can generate a false memory.

Sulin and Dooling (1974) also studied people's remembrance of implications as facts. In one case, they presented subjects with short paragraphs about a troubled girl and later tested for recognition both of ideas from the paragraph and of other ideas that had not been presented but that were either consistent or inconsistent with the theme of the passage. The paragraphs studied by two groups of subjects were identical, with the exception of the name above the story. In one case it was Helen Keller and in the other case it was Carol Harris. One item on the recognition test asked subjects if the following sentence had appeared in the narrative: "She was deaf, dumb, and blind." Students who had read the story about Helen Keller frequently made the error of saying that this sentence had appeared in the passage, whereas those reading about Carol Harris almost never made this error. In trying to reconstruct whether the sentence may have been in the story, the knowledge about Helen Keller caused subjects to fill in what "must have been" in the story and led subjects erroneously to conclude that the sentence was there.

Sometimes people remember the sentences that were in a prose passage no better than sentences they have never read before, so long as the new sentences are consistent with the material that was actually presented. Bransford and Franks (1971) gave subjects short sentences such as "The rock rolled down the hill" or "The rock crushed the hut." There were four simple ideas like this that, when put together, would make up the complex sentence: "The rock rolled down the hill and crushed the tiny hut." The experimenters gave their subjects various sentences to build up the complex idea unit. They later tested subjects by asking them to pick out sentences they had actually studied from among other sentences. Two types of distractor sentences were used: some were consistent with the idea unit but had not been studied, whereas other sentences were unrelated to the idea unit. The test sentences themselves could vary in the number of idea units present, from one to four. Subjects rated their belief that the sentences had been previously studied on a 10-point scale, from −5 (sure the sentence was new or nonstudied) to +5 (sure the sentence was old or studied).

The sentences that were not consistent with the idea units were accurately rejected as not having been studied, but the results for the other two types of items (shown in Figure 2) reveal that the more idea units embedded in the test

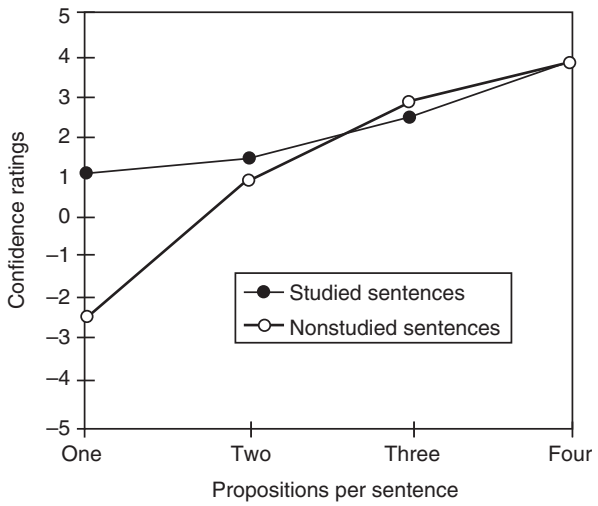

Figure 2 Mean recognition ratings for studied and nonstudied sentences as a function of number of propositions per sentence. Subjects could not distinguish sentences actually presented from those that had not been presented but that were consistent with the ideas. Data reproduced from Bransford, J.D., Franks, J.J., 1971. The abstraction of linguistic ideas. Cognitive Psychology 2, 331–350.

sentence, the higher the confidence rating that the sentence had been studied. Interestingly, this relation held both for sentences that had been studied and for new, nonstudied, sentences that were consistent with the overarching idea unit. Remarkably, except for sentences expressing only one idea, subjects were equally confident when recognizing the sentences that had not been presented in the study phase as when recognizing those that had been presented! Therefore, consistent with Bartlett's ideas, people seemed to retain the meaning of the sentences, but did not retain the literal wording of the sentences.

Bartlett (1932), who began this tradition of work, was interested in repeated recollection of material. He presented subjects with a Native American folktale, *The War of the Ghosts*, which he had them read twice. The story is a bit disjointed, has supernatural elements, and is rather difficult to understand. He found that subjects frequently made errors in recalling the story and that these errors increased over repeated retellings of the story. The stories got shorter, reflecting forgetting, but more importantly for present purposes, the story changed in systematic ways by subjects introducing errors that served to rationalize its bizarre elements. For example, people might have 'remembered' elements that had not been present to explain some of the unusual features in the story. Interestingly, this tendency to rationalize – to reconstruct the story more in line with the subject's own schema – seemed to grow over time and repeated retellings. Although Bartlett's original studies were rather informal, the work has been replicated under more systematic conditions (Bergman and Roediger, 1999).

Pictures and Scenes

Reconstructive memory has also been shown with common pictorial scenes (Miller and Gazzaniga, 1998; Weinstein and Shanks, 2010). Miller and Gazzaniga (1998) developed a procedure that used scenes that had once been covers of the *Saturday Evening Post* magazine. They showed people the scenes with a critical object from the scene removed. To use our earlier example, people might view a beach scene with a beach umbrella either present in the scene or absent from it. Miller and Gazzaniga found that when people were given recognition tests on objects in the scene, they were very likely to recognize the object as having been present when in fact it was not. Again, in reconstructing the scene, the object was inserted and seemed part of the memory.

The Misinformation Paradigm

One of the most popular paradigms for studying reconstructive processes in memory was developed by Elizabeth Loftus and her colleagues (Loftus et al., 1978) and is called the *misinformation paradigm*. It is modeled after the events that an eyewitness to a crime might experience. In the lab, subjects view a set of slides or video depicting an event, then later read a narrative or answer questions about the event. In Loftus et al. (1978), subjects witnessed an automobile accident that occurred at an intersection where a car failed to stop at a stop sign. In the later series of questions, subjects were asked either one of three questions: "Did the car stop at the traffic sign?" (control condition), "Did the car stop at the stop sign?" (consistent information condition), or "Did the car stop at the yield sign?" (inconsistent or misleading information condition). This question served as the experimental manipulation for a question asked at a later point in time: "What kind of sign was at the intersection, stop or yield?"

Results of this experiment are shown in Figure 3, where it can be seen that, relative to the control condition in which the question was posed in a neutral manner, wording of the question that was consistent with the original information improved recognition whereas the conflicting information reduced recognition. This last finding is the misinformation effect – incorrect information received after an event is often incorporated into one's memory for the event. This misinformation effect has been demonstrated in many experiments, although the magnitude of the effect depends on the exact nature of the test.

The general principle to arise from experiments on the misinformation effect is that memory for an event is not encapsulated in time the way the event itself is. Rather, information provided after the event can modify our memories for the event itself. Consistent information improves our later reconstruction, whereas conflicting or misleading information is harmful. When we try to reconstruct events from our past, we draw on all the information and resources at our disposal and the reported memory may include 'facts' that were not really part of the event, but incorporated into it later. Practically, if police or other people ask suggestive questions or embed erroneous information into questions asked of witnesses after the event, then the witnesses' later recollections and testimony might be tainted.

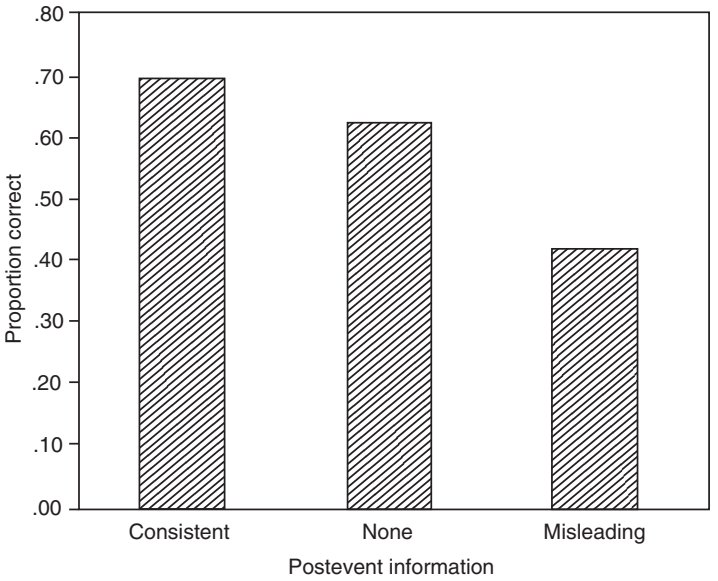

Figure 3 Probability correct (on a two-alternative forced choice recognition test) as a function of whether subjects had received information consistent with what they had previously seen, information inconsistent with what had been seen, or no specific postevent information. Misleading information after the event can reshape memory for the event. Data reproduced from Loftus, E.F., Miller, D.G., Burns, H.J., 1978. Semantic integration of verbal information into a visual memory. Journal of Experimental Psychology: Human Learning and Memory 4, 19–31 (Experiment 2).

Conclusion

This article has reviewed selectively some key evidence that permits us to draw the conclusion that remembering the past should be viewed as reconstructing it. To draw on a metaphor from Hebb (1949), we can think of the process of remembering the past as we conceive of paleontologists' reconstruction of a dinosaur from bone fragments and chips. The archeologist recovers a partial skeleton, but the finished product in a museum is shown as complete, with new bones added, old ones refinished or enhanced, and the entire skeleton reconstructed based on knowledge of what the animal probably looked like. Similarly, Hebb (1949) argued, remembering the past involves recollection of specific facts and details (the bones and bone chips) that are woven together into a complete story of the event (like the skeleton). However, the story about the event might involve considerable constructive activity on the part of the rememberer. A person's present knowledge and goals may shape and determine how he or she remembers the past. A quote from Bartlett captures the essence of this reconstructive approach:

> … the one overwhelming impression produced by this more 'realistic' type of memory experiment is that human remembering is normally exceedingly subject to error. It looks as if what is said to be reproduced is, far more generally than is commonly admitted, really a construction, serving to justify whatever impression may have been left by the original. It is this 'impression,' rarely defined as with much exactitude, which most readily persists. So long as the details which can be built up around it are such that they would give it a 'reasonable' setting, most of us are fairly content, and we are apt to think that what we have built we have literally retained. (Bartlett, 1932: pp. 175–176)

See also: Episodic Memory; Eyewitness Testimony; False Memories; Recognition; Retrieval; Source Monitoring; Visual Memory, Psychology of; Working Memory, Psychology of.

Bibliography

Bartlett, F.C., 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge, UK.
Bergman, E.T., Roediger, H.L., 1999. Can Bartlett's repeated reproduction experiments be replicated? Memory & Cognition 27, 937–947.
Bransford, J.D., Franks, J.J., 1971. The abstraction of linguistic ideas. Cognitive Psychology 2, 331–350.
Brewer, W.F., 1977. Memory for the pragmatic implications of sentences. Memory and Cognition 5, 673–678.
Brewer, W.F., Sampaio, C., 2006. Processes leading to confidence and accuracy in sentence recognition: a metamemory approach. Memory 14, 540–552.
Brewer, W.F., Sampaio, C., Barlow, M.R., 2005. Confidence and accuracy in the recall of deceptive and nondeceptive sentences. Journal of Memory and Language 52, 618–627.
Conrad, R., 1964. Acoustic confusions in immediate memory. British Journal of Psychology 55, 75–84.
Crowder, R.G., 1976. Principles of Learning and Memory. Erlbaum, Hillsdale, NJ.
Gallo, D.A., 2006. Associative Illusions of Memory: False Memory Research in DRM and Related Tasks. Psychology Press, New York.
Hebb, D.O., 1949. The Organization of Behavior: A Neuropsychological Theory. Wiley, New York.
Hoffman, D.D., 1998. Visual Intelligence: How We Create What We See. Norton, New York.
Loftus, E.F., Miller, D.G., Burns, H.J., 1978. Semantic integration of verbal information into a visual memory. Journal of Experimental Psychology: Human Learning and Memory 4, 19–31.
Meade, M.L., Roediger, H.L., 2006. The effect of forced recall on illusory recollection in younger and older adults. The American Journal of Psychology 119, 433–462.
Meade, M.L., Roediger, H.L., 2009. Age differences in collaborative memory: the role of retrieval manipulations. Memory & Cognition 37, 962–975.
Miller, M.B., Gazzaniga, M.S., 1998. Creating false memories for visual scenes. Neuropsychologia 36, 513–520.

Roediger, H.L., McDermott, K.B., 1995. Creating false memories: remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 21, 803–814.

Roediger, H.L., DeSoto, K.A., 2013. Confidence in memory: Assessing positive and negative correlations. Memory.

Roediger, H.L., Wixted, J.H., DeSoto, K.A., 2012. The Curious Complexity between Confidence and Accuracy in Reports from Memory. Memory and Law. Oxford University Press, Oxford, UK.

Sampaio, C., Brewer, W.F., 2009. The role of unconscious memory errors in judgments of confidence for sentence recognition. Memory & Cognition 37, 158–163.

Sperling, G., 1967. Successive approximations to a model for short-term memory. Acta Psychologica 27, 285–292.

Sulin, R.A., Dooling, D.J., 1974. Intrusion of a thematic idea in the retention of prose. Journal of Experimental Psychology 103, 255–262.

Tulving, E., 1985. Memory and consciousness. Canadian Psychology 26, 1–12.

Weinstein, Y., Shanks, D.R., 2010. Rapid induction of false memory for pictures. Memory 18, 533–542.

Recovery of Function: Dependency on Age

Bryan Kolb, Canadian Centre for Behavioural Neuroscience, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

When the brain is injured there is a cascade of cellular events that can both seriously compromise the functioning of the brain and stimulate reparative processes that underlie eventual functional improvement. The likelihood and intensity of these cellular changes vary with the age of the individual. In particular, there are times during brain development, which begins *in utero* and continues well into the third decade of life, when brain injury may lead to strikingly different outcomes. These critical periods are related to the varying maturational events occurring at different stages of brain development.

Brain damage results in lost functions that do not completely return after injury. The brain is capable of compensatory change, however, thus allowing at least partial restitution of function, the extent of recovery varying with the precise developmental state of the brain at the time of injury. Thus, although some compensatory change is possible in the brain at any age the best outcome from cerebral injury occurs after damage during the period of maximal synaptogenesis, which is in the first two years of life in humans. The formation of new synapses provides a mechanism to modify the remaining cerebral circuitry, which in turn allows each neuron to increase its processing capacity. Although it is uncommon, in certain restricted cases the brain also may generate new neurons that can be integrated into the remaining brain and support functional recovery. Treatments that increase synaptogenesis or cell generation will enhance functional recovery and events that reduce synaptic formation will retard recovery.

When the brain is injured in adulthood there is a loss of function followed by a prolonged period during which there is partial recovery of some types of cognitive and motor functions. The extent of functional recovery varies with age and although injury to the infant brain can have less severe consequences than similar injury in adulthood, there are times during development when injury is especially devastating. These differences are related to the details of anatomical changes at different stages of brain development. The difference in functional outcome after injury at different ages is related to the injury-induced changes in neuronal structure, changes that are referred to as plastic changes. These plastic changes are similar to the neural and glial changes that are associated with learning in the normal brain. Various factors, such as experience, hormones, and drugs modulate these plastic changes. Even these factors cannot completely restore lost functions because there are limits to brain plasticity.

The Nature of 'Recovery' after Brain Injury

There are many ways to damage the brain (e.g., stroke, disease, trauma) and it is rare indeed for a family not to contain a member who is affected by one of them. For example, it is estimated that about 400 000 people with new head injuries are admitted to hospitals each year in the United States. Similarly, it is estimated that there are about 400 000 new cases of stroke every year in the United States and most surviving stroke victims experience significant cognitive loss. These figures do not include neurological disorders such as Parkinson's and Alzheimer's diseases, which together affect more than 1% of the population. The estimate in the United States is that about 48 million people have some sort of neurological disorder, excluding mental diseases such as schizophrenia, which are clearly a result of neural dysfunction.

The high incidence of brain injury and disease thus poses perhaps the most perplexing clinical neurological problem, which is the issue of how to repair the damaged brain. It is tempting to assume that if behavior improves after injury, then there is recovery of the original behavior. However, just because a behavior appears to recover does not mean that the lost behavior itself has returned. The problem is to determine when recovery has actually occurred. In fact, true recovery is a rare occurrence after brain injury.

We can identify at least three outcomes that could occur after brain injury. First, there may be a compensation resulting from an adaptation to the lost function. Compensation could reflect a change in strategy or it may represent a substitution of a new behavior for the lost one. For example, a person who has a deficit in moving the eyes can be expected to have difficulty in reading. This difficulty can be solved, however, by moving the head, which allows restitution of function (reading) by substitution of one behavior for another. Similarly, if, following a stroke, a person can walk only with the aid of a cane, the ability to walk has returned but the original walking ability has not returned, just an adaptation to the behavioral difficulty.

A second outcome might be partial restitution of the original behavior. This could reflect recovery from some sort of nonspecific effect of the injury, such as swelling, or it may reflect a genuine return of function. Behaviors that return over a period of months are likely to reflect a partial restitution of function. This is most clearly seen in people with significant language disturbances (i.e., aphasias) in which the severe deficits in making the movements of the tongue and mouth necessary to speak may diminish leaving the person with more subtle language disturbances, such as in finding words (anomia).

Third, there is a possibility that there could be a complete restitution of the original behavior. Although it is at least theoretically possible for lost functions to return completely in a plastic brain, careful behavioral analysis shows that this is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.55048-X

rarely, if ever, the case in either adult laboratory animals or human patients. In principle, however, it ought to be possible to reorganize the brain to allow at least some functions to completely recover. This process is most likely in the developing brain and it has been documented for over 100 years; children with damage to language areas can show an almost complete recovery of lost language functions.

Age-Dependent Recovery

Broca first noticed in the late 1800s that damage to the speech zones of the left hemisphere during infancy did not interfere with the later development of language (Finger and Almli, 1988). Rasmussen and Milner (1977) later showed that the sparing of language functions was due to the shifting of the speech zones either to the opposite hemisphere or, in some circumstances, within the left hemisphere. In these later studies the critical factor was age: lesions of the left hemisphere language zones prior to 5 years of age changed the speech representation so that the affected language zone moved to the right hemisphere.

Similar lesions from the ages of about 6–12 still allowed recovery of language functions but this was accomplished by intrahemispheric reorganization (Krashen, 1973). The studies of the recovery of language functions after infant brain injury led to the idea that damage to the infant brain led to better functional outcomes than damage after similar injury in adulthood. This conclusion was supported by the studies of Margaret Kennard in the late 1930s in which she showed that unilateral motor cortex impairments are less severe after motor cortex injury if the lesion occurs in infancy rather than adulthood (Kennard, 1940). One difficulty with the studies of both Kennard and the language reorganization in children is that there are also many examples in which early brain damage results in rather severe functional loss, such as in cerebral palsy or hemiplegia resulting from unilateral cerebral injury at birth. In fact, Hebb (1949) noted that children with frontal lobe injuries in infancy showed little functional recovery in adulthood and for some behavior, such as social behavior, these children might actually have more severe deficits than adults with similar injuries.

The apparent contradiction between the recovery of language and some motor functions on the one hand, and the poor recovery of some executive, social, and other motor functions on the other hand, has been resolved by laboratory studies in which precise age at injury and the nature of the behavioral deficits have been systematically examined. A key element of these studies is that the age of the brain is measured not in relation to the birthday, but rather it is measured by its stage of anatomical development. For example, compared to humans, newborn rats are approximately equivalent to about 6-month old fetuses when they are born. Thus, if the newborn rat brain is damaged, we would anticipate functional symptoms similar to those of human infants whose brain is injured at the beginning of the third trimester. In order to compare across species we can separate the periods of brain development into five broad stages: neurogenesis, cell migration, cell differentiation, synaptogenesis, and gliogenesis (*see* Prenatal and Infant Development: Overview, Current Trends, Future Directions).

Studies on both laboratory rats and cats have shown that if the cortex is injured bilaterally during neurogenesis, when neurons are formed, there is substantial recovery and, if the injury is not too extensive, there can be virtually complete functional recovery. In contrast, however, if the cortex is injured in the first few days after birth in the rat, which is a time of neural migration and cell differentiation, the effect is functionally devastating as animals show far more severe effects of injury than would be expected even in aging animals. This poor outcome is niether a function of extent of injury, nor localized to particular cortical areas (Kolb, 1995).

Rather, there is something about the rat cortex shortly after birth that makes it especially vulnerable. For example, damage during this time may disturb the subsequent process of synapse formation or may even alter stem cell activity in some way. Once this phase of development is over, however, and brain development moves to the production of connections and glia, the brain shows a remarkable ability to compensate for injury. Rats or cats with cortical injuries during this phase show far better behavioral capacities in adulthood than animals with similar lesions at any other time. In fact, on some behavioral tests these animals show recovery that is virtually complete.

Importantly, this recovery is far more extensive in cognitive tasks, such as the learning of various spatial navigation problems, than it is on tests of species-typical or motor behavior (Bachevalier and Malkova, 2000; Kolb and Whishaw, 1989). In humans, spatial behavior appears to show less recovery than verbal behavior (Stiles, 2000), possibly because spatial behavior is supported by a more diffuse pattern of neural networks (Stiles, 2012). This behavior-specific recovery likely reflects the relative ease of reorganizing cortical circuitry in supporting different types of behavior. The neural substrates for less plastic behavior are presumably more highly specified and constrained during development than the neural substrates for more plastic behavior. It also reflects a difference in the way the brain produces cognitive processes versus basic sensory and motor processing. Thus, the most extensive recovery is associated with lesions of the frontal areas and the least extensive recovery is associated with damage to the primary sensory areas.

Adolescence can be viewed as a behavioral transitional period that is characterized by pruning of synapses throughout the cerebral cortex. Evidence is accumulating from both human clinical studies (Crowe et al., 2012) and laboratory animal studies (Nemati and Kolb, 2010) that the effects of cerebral injury vary with age in a nonlinear manner. For example, both children and laboratory rats with cortical injury in early adolescence have poorer cognitive outcomes than those with earlier or later injuries. In the rodent studies the poor outcomes are correlated with enhanced synaptic loss after injury in early adolescence relative to similar injury in late adolescence. Given that different cortical regions shed synapses at different rates, with the prefrontal cortex being the latest, it might be expected that the critical period for poorer outcome might vary across regions, with the prefrontal cortex being the latest (Nemati and Kolb, 2012). Areal studies are difficult in studies of human children because most research

is conducted on children with traumatic brain injuries, which are normally not well localized.

Mechanisms Underlying Recovery of Function

The fundamental assumption here is that recovery of function can occur only if there is some type of modification of the structure of the remaining brain. This structural change is necessary if the remaining brain cells are to be recruited in some manner to control behaviors that preinjury were controlled by the lost neurons. In principle, there are three ways that the brain could make changes that might support recovery. First, there could be changes in the organization of the remaining, intact, circuits in the brain. The general idea is that the brain could reorganize in some way to do 'more with less.' It is unlikely that a complexly integrated structure like the cerebral cortex could undergo a wholesale reorganization of cortical connectivity, but rather, recovery from cortical injury would be most likely to result from a change in the intrinsic organization of local cortical circuits in regions directly or indirectly disrupted by the cortical injury. Although it might be possible to produce significant reorganization of cortical connections in the young brain, the overwhelming evidence in experimental animals is that this is rare and it is just as likely to be associated with abnormal functioning as with recovery.

Second, there could be a generation of new circuitry (Stewart, 1991). Cerebral reorganization can be stimulated either by some sort of behavioral therapy or the application of some sort of pharmacological treatment. In either event, the stimulus could influence reparative processes in the remaining brain or could enhance the production of new circuitry. Once again, it seems most likely that the induced neuronal changes would be in the intrinsic organization of the cortex. One might predict that induced neuronal changes might be more extensive than in the case of endogenous change, in part because the treatment can act upon the whole brain rather than just on affected regions.

Third, there could be a generation of new neurons and glia to replace at least some lost cells. Stem cells that give rise to the neurons and glia of the brain remain active in the brain throughout life (*see* Cognitive Development in Infancy, Neural Mechanisms of; Cortical Development: Transplantation and Rewiring Studies; Hypnosis, Neural Basis of). It is thus possible that neurogenesis could be stimulated after injury, especially during development, and that these new neurons could replace those lost to injury or disease (Kolb et al., 2007). All three of these possibilities for cerebral plasticity do occur, but the likelihood of each type of change varies with age at injury.

Regardless of the proximal mechanism supporting recovery, all mechanisms presumably require changes in gene expression to manufacture the appropriate proteins. There is accumulating epigenetic evidence that a wide variety of experiences, including brain injury, produce regionally specific changes in gene expression (Mychasiuk et al., 2011).

Structural Correlates of Functional Recovery

When the brain is damaged, there is a loss of neurons in the injured area but, in addition, many neurons that are not directly injured lose their normal synaptic inputs. The immediate effects of this synaptic loss are (1) neuronal death, (2) atrophy of the dendritic and/or axonal fields that have lost their connections, and/or (3) subsequent reorganization of the connections. This reorganization can be inferred by using a Golgi stain to visualize the dendritic fields of the remaining neurons. Because the dendritic branches of a neuron provide the substrate for most of a cell's synapses, a decrease (or increase) in the amount of dendritic space available is correlated with a decrease (or increase) in the number of synapses.

The hippocampal formation of adult rats has provided a good model to study synaptic replacement. When input to the hippocampus is reduced by damage to the major input pathway via the entorhinal cortex, there is a massive denervation of the hippocampal formation, which can be seen in the atrophy of the dendritic fields of the hippocampal cells in the week after injury. Beginning at about 7–10 days after the injury, there is then a slow regrowth of the dendritic fields, although the pattern of regrowth is different from the original pattern. This regrowth reflects a reorganization of the circuitry and it is correlated with a partial recovery of function. The restructuring of the synaptic organization of the hippocampus occurs primarily because the remaining inputs to the hippocampus expand to occupy the regions of lost inputs.

A similar process appears to occur after restricted neocortical injury in adulthood as well, although the changes are less dramatic than in the hippocampus and there is much less functional recovery (Kolb, 1995). In contrast, however, there is extensive reorganization after cortical injury during development, and this reorganization is correlated with functional outcome. Golgi analyses of cortical neurons of rats with perinatal lesions consistently show a general atrophy of dendritic fields and a drop in spine density across the cortical mantle. In contrast, rats with cortical lesions around 10 days of age show an increase in dendritic fields and an increase in spine density relative to normal control littermates. Thus, animals with the best functional outcome show the greatest synaptic increase whereas animals with the worst functional outcome have a decrease in synapses relative to control animals.

But infant animals not only show changes in local cortical circuits but also in more distant cortical connectivity. Perhaps the most extensive studies of changes in cortical connectivity are those showing that after unilateral motor cortex lesions in infant rats or cats there is a major expansion of the ipsilateral corticospinal pathway from the undamaged hemisphere, which is correlated with a partial recovery of skilled forelimb use (Castro, 1990). These expanded connections do not reflect the growth of new pathways, however, but rather reflect the failure of superfluous pathways to die off during development. That is, in development there is normally a period during which many pathways are eliminated but after some types of brain injury these pathways are maintained (O'Leary, 1992). The presence of these abnormal pathways may support some types of functional recovery but it may also interfere with the normal functioning of remaining cortical functions. This possibility has been termed 'crowding' to reflect the idea that the normal functions of a cortical region can be crowded out by the development of abnormal connections (Teuber, 1975). Thus, there may be recovery of some types of behavior, such as motor functions, at the expense of other functions. This appears to be especially

likely if the injury occurs during the period of neuronal migration and unlikely if it occurs during the period of maximal synaptogenesis.

As noted earlier, synaptic change may also occur after injury because the brain generates new neurons. Although this is likely to be rare in the adult, it does occur during development, but the generation of these neurons varies with precise embryonic age and location of injury. If it is damaged during the time of neuronal generation, the brain can compensate by generating rather large numbers of new cells that appear to support at least partial restitution of function (Hicks et al., 1984). It is not known if this regenerative process occurs after injury to specific regions of the brain but it is likely a general phenomenon. Thus, if the brain is injured while it is still producing neurons, it appears capable of generating replacements. In contrast, if the brain is injured during neuronal migration there appears to be little generation of new neurons. If the same injury occurs during maximal glial formation, there is neuronal generation. This generation is selective, however, and occurs only if the midline telencephalon is damaged, namely the olfactory bulb or midline cortex (Kolb and Gibb, 2007). The newly generated tissue is functional and animals regain at least partial function.

Modulating Functional Recovery at Different Ages

The functional, molecular, and structural changes discussed to this point have been spontaneous after injury at different ages. It is possible, however, to influence the endogenous changes by various types of experiences. The best-understood experiences include behavioral therapies, pharmacological therapies, and cell-based therapies (Kolb et al., 2010). Behavioral therapies include both specific sensory and motor experiences as well as complex housing, the latter having a more generalized effect on both brain and behavior. Pharmacological treatments range from psychomotor stimulants (amphetamine, nicotine, methylphenidate) to drugs that stimulate axon regeneration (e.g., inosine). These treatments have largely been used in adult animals with little study of their efficacy for developmental injuries. Finally, the cell-based therapies take advantage of the presence of multipotent neural stem cells in the developing and adult forebrain (Reynolds and Weiss, 1992). The goal of these treatments is to stimulate the growth of endogenous stem cells that will migrate to the site of injury and regenerate the lost tissue.

As a general rule of thumb, therapies are most effective both when injuries are not extensive and when the spontaneous recovery has been restricted. For example, when rats have restricted cortical injuries in the first week of life there is less recovery than after similar injuries later in life but the very early injured brain is extremely responsive to a wide variety of treatments. Kolb and Gibb (2010) used tactile stimulation for two weeks after injury to facilitate recovery in adulthood after perinatal injury to prefrontal or parietal cortex. Similar treatment after injury in the second week, which normally leads to good spontaneous injury, had minimal effect on outcome that suggests that the spontaneous recovery had already exhausted the available plasticity. Cell-based therapies have been little used in development, the exception being Fibroblast Growth Factor-2 (FGF-2), a protein implicated in brain development and wound healing. FGF-2 stimulates functional improvement and synaptic changes after cortical injury in the first week of life but has much less impressive effects in adulthood (Comeau et al., 2007). Curiously, whereas FGF-2 induces synaptic change after perinatal injury, it mobilizes stem cells to generate and migrate to the site of injury in the second week of life (Monfils et al., 2008). These cells integrate into the developing circuitry and support both cognitive and motor recovery. This does not happen after FGF-2 treatment for similar adolescent or adult injury.

In sum, there is an age-dependent pattern of both spontaneous and treatment-facilitated recovery after cerebral injury. The age-dependent functional recovery is correlated with differing mechanisms at different ages.

See also: Brain Development: Ontogenetic Neurobiology of; Cerebral Cortex; Childhood and Adolescence: Developmental Assets; Cognitive Development in Infancy, Neural Mechanisms of; Cortical Development: Transplantation and Rewiring Studies; Epigenetics of Brain Plasticity and Behavior; Neural Plasticity; Prefrontal Cortex Development and Development of Cognitive Function; Youth Culture, Anthropology of.

Bibliography

Bachevalier, J., Malkova, L., 2000. Neuropsychological indices of early medial temporal lobe dysfunction in primates. In: Levin, H.S., Grafman, J. (Eds.), Cerebral Reorganization of Function after Brain Damage. Oxford University Press, New York, pp. 27–48.

Crowe, L.M., Catroppa, C., Babl, F.E., Rosenfeld, J.V., Anderson, V., 2012. Timing of traumatic brain injury in childhood and intellectual outcome. Journal of Pediatric Psychology 37, 745–754.

Castro, A.J., 1990. Plasticity in the motor system. In: Kolb, B., Tees, R.C. (Eds.), The Cerebral Cortex of the Rat. MIT Press, Cambridge, MA, pp. 563–588.

Comeau, W., Hastings, E., Kolb, B., 2007. Pre-0 and postnatal FGF-2 both facilitate recovery and alter cortical morphology following early medial prefrontal cortical injury. Behavioural Brain Research 180, 18–27.

Finger, S., Almli, C.R., 1988. Margaret Kennard and her 'principle' in historical perspective. In: Finger, S., Levere, T.E., Almli, C.R., Stein, D.G. (Eds.), Brain Injury and Recovery. Plenum, New York, pp. 117–132.

Hebb, D.O., 1949. The Organization of Behavior: A Neuropsychological Theory. Wiley, New York.

Hicks, S.P., D'Amato, C.J., Glover, R.A., 1984. Recovery or malformation after fetal radiation and other injuries. In: Finger, S., Almli, C.R. (Eds.), Early Brain Damage. Academic Press, Orlando, FL, pp. 127–147.

Kennard, M.A., 1940. Relation of age to motor impairment in man and in subhuman primates. Archives of Neurology and Psychiatry 44, 377–397.

Kolb, B., 1995. Brain Plasticity and Behavior. Erlbaum, Mahwah, NJ.

Kolb, B., Gibb, R., 2007. Brain plasticity and recovery from early cortical injury. Developmental Psychobiology 50, 134–156.

Kolb, B., Gibb, R., 2010. Tactile stimulation facilitates functional recovery and dendritic change after neonatal medial frontal or posterior parietal lesions in rats. Behavioural Brain Research 214, 115–120.

Kolb, B., Morshead, C., Gonzalez, C., Kim, N., Shingo, T., Weiss, S., 2007. Growth factor-stimulated generation of new cortical tissue and functional recovery after stroke damage to the motor cortex of rats. Journal of Cerebral Blood Flow and Metabolism 27, 983–997.

Kolb, B., Teskey, G.C., Gibb, R., 2010. Factors influencing cerebral plasticity in the normal and injured brain. Frontiers of Human Neuroscience 4, 204.

Kolb, B., Whishaw, I.Q., 1989. Plasticity in the neocortex: mechanisms underlying recovery from early brain damage. Progress in Neurobiology 32, 235–276.

Krashen, S.D., 1973. Lateralization, language learning, and the critical period: some new evidence. Language Learning 23, 63–74.

Monfils, M.H., Driscoll, I., Vavrek, R., Kolb, B., Fouad, K., 2008. FGF-2 induced functional improvement from neonatal motor cortex injury via corticospinal projections. Experimental Brain Research 185, 453–460.

Mychasiuk, R., Gibb, R., Kolb, B., 2011. Prenatal stress produces sexually dimorphic and regionally-specific changes in gene expression in hippocampus and frontal cortex of developing rat offspring. Developmental Neuroscience 33, 531–538.

Nemati, F., Kolb, B., 2010. Motor cortex injury has different behavioral and anatomical effects in juvenile and adolescent rats. Behavioral Neuroscience 24, 612–622.

Nemati, F., Kolb, B., 2012. Recovery from medial prefrontal cortex injury during adolescence: implications for age-dependent plasticity. Behavioural Brain Research 229, 168–175.

O'Leary, D.D., 1992. Development of connectional diversity and specificity in the mammalian brain by the pruning of collateral projections. Current Opinion in Neuroscience 2, 70–77.

Rasmussen, T., Milner, B., 1977. The role of early left-brain injury in determining lateralization of cerebral speech functions. Annals of the New York Academy of Sciences 299, 355–369.

Reynolds, B., Weiss, S., 1992. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. Science 255, 1707–1710.

Steward, O., 1991. Synapse replacement on cortical neurons following denervation. In: Peters, A., Jones, E.G. (Eds.), Cerebral Cortex, vol. 9. Plenum, New York, pp. 81–131.

Stiles, J., 2000. Spatial cognitive development following prenatal or perinatal focal brain injury. In: Levin, H.S., Grafman, J. (Eds.), Cerebral Reorganization of Function after Brain Damage. Oxford University Press, New York, pp. 201–217.

Stiles, J., 2012. The effects of injury to dynamic neural networks in the mature and developing brain. Developmental Psychobiology 54, 343–349.

Teuber, H.-L., 1975. Recovery of function after brain injury in man. In: Outcome of Severe Damage to the Nervous System. Elsevier, North-Holland, Amsterdam, The Netherlands. CIBA Foundation Symposium 34.

Re-entry and Reintegration After Incarceration

Christy A Visher, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With few exceptions, men and women who are incarcerated are eventually released. Reentry happens when incarceration ends. These individuals face daunting barriers to their successful reintegration and reconnection with society. Successful reentry requires strong family and community support networks and comprehensive services. Interventions that reflect these principles may hold the greatest promise for success.

Reentry is defined as the inevitable consequence of incarceration. With the exception of those who are sentenced to life without parole, die of natural causes or are executed, everyone placed in confinement is eventually released. Reentry is not a legal status. Indeed, not all prisoners are released to parole supervision; many are released directly into the community with no continuing obligation to observe special conditions of their release. Nor is reentry a new kind of program. Certainly, the pathways of reentry can be influenced by such factors as the prisoner's participation in drug treatment, education programs, religious organizations, or prison industries, but reentry is not a result of program participation. Reentry happens when incarceration ends. In other words, for those who are incarcerated, reentry is not an option. In this broad definition, 'reentry' is experienced by individuals sent to either jail or prison, federal or state facilities, and either as adults or juveniles. Juxtaposed against this fact is the experience of *reintegration*, or the individual's reconnection with the institutions of society, which is both a process and a goal.

Although there has been a lot of attention to the problem of prisoner reentry since 2000, the concept of prisoner reentry is not new. As early as 1970, John Irwin, author of *The Felon*, challenged scholars to tackle the complexities of the reentry phenomenon. Noting the irony that researchers and journalists had extensively examined the reentry challenges facing World War II veterans and Peace Corps returnees, he reached this conclusion:

> [T]here is little indication either from the literature or from interviews of persons involved in dealing with parolees of the existence of any awareness of the broader aspects of the reentry problem. This general blindness seems to be related to formal and informal societal conceptions of the ex-convict. (Irwin, 1970: p. 109)

When Irwin wrote these words, the prison population in the United States was 200 000. Today's intense interest in prisoner reentry is focused on an annual reentry cohort in the United States more than three times larger, and reentry is recognized as a significant societal problem around the world, not just in the United States. Internationally, high imprisonment rates compound the challenge to the criminal justice system that the current system of incarceration and reentry creates.

Unfortunately, a majority of released prisoners will reoffend. Repeated cycles of failure, a link between drug use and crime, and addiction masking mental health problems are common. The social costs associated with this phenomenon are significant for prisoners and their families – including higher penalties for repeated offenders, further detachment from prosocial networks – and government agencies, which include public safety concerns, and growing fiscal pressures. Some of these obstacles are not new; however, their relevance and impact has been amplified by mass imprisonment and more conservative policies and ideologies. In response, practitioners and scholars have reconsidered the logic and structure of the reintegration process for ex-prisoners in order to provide more effective services and interventions (Travis, 2005; Visher and Travis, 2003). A growing body of research has begun to document these reentry initiatives from various perspectives and methodologies.

In the United States, as the rapid expansion of prison populations in the 1980s and 1990s resulted in ever larger numbers of prisoners being released each year, policymakers realized that correctional costs could be reduced by helping individuals make a successful transition. Practitioners and policymakers alike have begun demanding evidence-based strategies that help prisoners succeed after release. Some of these strategies – education, job training, work release – were the focus of offender rehabilitation efforts 30 years ago. However, recent attention to the problem of prisoner reentry is not just focused on the individual offender, but takes a broader approach in conceptualizing the problem and devising solutions.

The Challenges of Reentry

Men and women returning home from prison face formidable and simultaneous challenges, including finding a job and a place to live, locating health services and substance abuse treatment, arranging transportation, and developing prosocial relationships. Aside from individual characteristics, other factors that affect the process of reintegration include an individual's family context, the community to which he or she is returning, and the larger sociolegal environment governed by existing policies and regulations.

Until recently, describing the characteristics of individuals who are exiting prison and the challenges they face was difficult. Prior to 2000, only a handful of small, qualitative studies had examined the lives of individuals released from prison. Moreover, these reentry experiences were often told through the eyes of service providers or parole officers. Our

understanding of the reentry process and the challenges faced by released prisoners has increased significantly through new data sources that rely on reentering individual's reports of their experiences (Lattimore and Visher 2009; Visher, 2007; Visher et al., 2010). Utilizing these self-report data reveals that prisoners' perspectives of the reentry experience differ in important respects from the assumptions shared by many practitioners, policymakers, and researchers. It is also likely that some commonly held views of prisoners are shaped by the experience of working with certain groups rather than with all those who return to society. For example, former prisoners who appear at shelters, soup kitchens, or community mental health clinics are likely to have different needs and characteristics than ex-prisoners who do not need shelters, donated food, or mental health treatment. In addition, former prisoners are not a homogeneous group and they do not have identical experiences in their transition from prison to the community. Length of sentence, experiences in prison, age, and family relationships all affect the reintegration process.

Despite this heterogeneity, numerous social and economic disadvantages characterize the vast majority of individuals who are released from prison, including poor educational attainment and employment histories, poor physical and mental health, and substance misuse (Visher et al., 2010). A small group of prisoners are able to further their education while incarcerated but such opportunities depend on the availability of education programs and inmate eligibility for those programs. Although prisoners typically report working before their incarceration, most have held only temporary jobs for short periods of time. It is well known that a majority of prisoners have extensive substance use histories, yet only a fraction of those who need substance abuse treatment receive it during incarceration (Petersilia, 2003, 2005). Moreover, given high rates of substance use, it is not surprising that exiting prisoners are also likely to suffer from health problems and mental illness (Greifinger et al., 2007; National Commission on Correctional Health Care, 2002). The majority of men exiting prison have a chronic physical or mental health condition, with the most common conditions including depression, asthma, hepatitis, and high blood pressure (Visher and Mallik-Kane, 2007). Prisoners with health problems may have a more difficult transition than others.

This portrait of former prisoners reveals the difficult challenges in their personal lives as they leave prison and return to the community. Not surprisingly, the self-reported needs of exiting prisoners mirror these challenges. The needs of former prisoners range from general needs such as more education to specific needs such as photo identification or a driver's license (Lattimore et al., 2010). Among men, the levels of expressed need are highest for education, job training, and employment, followed by the need for various transitional and health-related services. Almost all soon-to-be-released prisoners report needing more education, job training, and a job. Substantial proportions of exiting prisoners need basic transitional services such as housing, access to clothing and food, and transportation, including the need for a driver's license. A majority of prisoners need medical treatment, and many recognize their need for treatment of substance use and mental health issues. Exiting prisoners also recognize a need for services that will help them change their past behavior patterns as they return to the community, such as mentoring, services to help change criminal attitudes, and anger management classes (Lattimore et al., 2010).

Female prisoners express greater needs for services than men, especially needs for education, employment, and access to clothing and food. Moreover, women need more health services than men, including mental health treatment and substance use treatment. About one-third requested access to support groups for abuse victims. Not surprisingly, need for child-related services among exiting female prisoners is high with the vast majority of women with children needing some type of child-related service (Lattimore and Visher, 2009; Lindquist et al., 2009).

In the months after release from prison, men and women continue to report needs for a variety of services. Indeed, the highest risk of recidivism is in the 6 months after release (National Research Council, 2007). The greatest continued needs for both men and women are more education, financial assistance, and health care insurance. Yet, obtaining services in the community is difficult because of inadequate information on program or treatment availability and eligibility, lack of transportation, and poor coordination among service providers. Frustrated and overwhelmed, individuals returning to their communities often do not receive the services and treatment that they need. Relapse and recidivism is a common result.

Family

The role of family is important to understanding the reintegration of former prisoners, yet we know little about its precise impacts. In criminology, criminal behavior has long been tied to family attributes. Many studies have examined the impact of an individual's early family structure and experiences on criminal activity. However, the topic has been virtually ignored in studies of recidivism, although desistance research indicates that the family may be critical to explaining individual pathways after release from prison. Among the family influences that may be important are nature of family contact (e.g., frequency, antisocial, prosocial) and type of support offered, if any (i.e., emotional, financial, housing). In response to the renewed focus on reentry and reintegration, research attention is being focused on the family factors that lead to the success or failure of former prisoners.

Strong ties between prisoners and their families or close friends during incarceration appears to have a positive impact on postrelease success. The earliest efforts to understand the relationship between family ties and postprison success came from Lloyd Ohlin's efforts to develop a parole success prediction scale for the State of Illinois in the 1950s. Ohlin constructed an 'index of family interest' and compared the number of visits and number of visitors for a sample of releasees from Illinois State Prisons drawn from 1925 to 1935. Inmates who were classified as maintaining active family interest were more likely to be successful on parole, while those who were classified as loners experienced significantly lower rates of parole success (Ohlin, 1954). These findings were replicated in the 1970s and 1980s during a period of renewed attention to the study of family impact on postprison outcomes. Stronger ties

between inmates and families and close friends during incarceration lead to better postrelease outcomes. Prisoners who experience more family contact – whether through visits or mail, or via participation in programs intended to facilitate family contact – experience lower recidivism rates and greater success postrelease (see Holt and Miller, 1972; Hairston, 1988).

More recently, research has identified the kinds of assistance that released prisoners receive from supportive families that appear to affect postrelease outcomes. In 1999, researchers with the Vera Institute of Justice tracked and interviewed 49 people who were released from New York State prisons or New York City jails during the initial 30 days immediately following release. Nelson et al. (1999) found that for these former prisoners, family support played a critical role during that first 30 days. Specifically, the interviews revealed two important areas of family support: emotional support and housing assistance. Researchers found family acceptance and encouragement as well as perceived emotional support from family were both related to postrelease success. The recently released inmates who demonstrated the greatest success in employment and abstinence from drugs, and generally exhibited optimistic attitudes were "the ones that talked most about their family's acceptance of them" (Nelson et al., 1999: p. 10). In addition, those who returned to live with family were less likely to abscond from parole. In one large longitudinal study of released prisoners followed for up to a year, the majority reported that they had at least one close family member, and almost half had four or more close family members (Visher, 2007). Moreover, the majority of returning prisoners live with family members and/or intimate partners immediately after release.

In response to the consistent findings about the importance of family support, some jurisdictions have begun to develop programs for the family members of prisoners while the offender is incarcerated. These programs include components such as marital workshops, family education, and parenting skills; these programs have been shown to improve family relationships and ties as a means of preparing the prisoner for release and reintegration into both the family and the community (Shapiro and Schwartz, 2001). Given the consistent empirical evidence that maintaining and strengthening family ties positively affects postprison outcomes, a logical next-step was for practitioners to develop programs and services for families of inmates, through which healthy ties can be fostered. A few studies, primarily program evaluations, report that providing certain social services to families of recently released prisoners leads to several positive outcomes, including decreased drug use, fewer physical, mental, and emotional problems, and decreased recidivism (Couturier, 1995; Shapiro and Schwartz, 2001; Sullivan et al., 2002). Thus, strengthening the family network and maintaining supportive family contact can improve outcomes after prisoners are released.

Overall, family connections are important for successful reintegration. Returning prisoners who assume conventional roles within their families have greater success in their transition from prison: married men have more successful transitions than single men; men who reside with their wives and children upon release have more successful transitions than those who live alone or with a parent; and recently released inmates who describe their marriage as harmonious do better than those who describe their marriage as one of conflict (Gadsden, 2003; Waller, 1964).

Community

A third set of factors that influence individual transitions from prison to community are the characteristics of the environment to which the ex-prisoner returns. Overall, research estimating the impact of neighborhood factors such as socioeconomic status, racial diversity, and residential stability on individual outcomes shows that these effects are moderate and depend on individual characteristics (Iannotta and Ross, 2002). The study of crime and delinquency has a rich history of ecological research that has focused on the ways in which characteristics of communities may influence rates of crime and violence. Studies of neighborhood-level influences on individual criminal activity and recidivism show that structural features of neighborhoods such as residential stability, rates of organizational participation, and measures of informal and formal social control have either direct or mediated effects on individual criminal activity (see Sampson, 2002, 2012).

Returning prisoners are also increasingly concentrated in central cities and within them, in a relatively small number of neighborhoods that often are characterized by severe poverty, social disorganization, and high crime rates (Lynch and Sabol, 2001; Travis, 2005; Clear, 2007). It makes sense that a place of residence may affect individual behavior, especially for persons returning to a community after a period of absence. In the case of returning prisoners, a variety of neighborhood indicators and resources may affect postrelease outcomes, including housing values and availability, job availability or proximity to jobs, health care services, and substance abuse treatment (Travis, 2005). Moreover, the availability of social services facilitates a smoother reentry and diminishes repeat criminal activity (Freudenberg et al., 2008).

In addition to the structural features of an area, the *social* organization of a neighborhood, including the density of networks among residents, satisfaction with one's community, and the capacity of the community to join together around common goals, is also theoretically linked to rates of crime and disorder (see Sampson, 2002). Sampson et al. (1997) created a new measure of these elements, calling their concept neighborhood 'collective efficacy' – a sense of mutual trust and shared willingness to intervene for the common good. They showed that collective efficacy affected local crime rates independent of neighborhood characteristics such as poverty, residential stability, and immigrant concentration. Conceivably, neighborhood social processes may affect individual transitions from prison to community. Formal and informal social controls involving local social networks are likely to vary across neighborhoods, and these processes may deter or exacerbate individual propensities to commit crime.

Because of the geographical concentration of crime in urban areas, some neighborhoods have been affected in recent years by two types of population change: first, large numbers of men have been removed from these neighborhoods due to incarceration, and second, the return of these men as they are released from prison. The high rates of incarceration concentrated in these neighborhoods may actually increase crime because the removal of men from these areas destabilizes social

networks that promote informal social control among neighbors (Clear, 2007; Lynch and Sabol, 2000). The magnitude of the prisoner reentry phenomenon also profoundly affects other aspects of communities, including public health, homelessness, declines in civic participation, and lost connections among families and within communities. Other criminals may move in to replace the incarcerated residents. Residents in these areas may also suffer financial hardship, further weakening the vitality of the neighborhood.

Former prisoners returning to these neighborhoods may also increase neighborhood crime because the social network of the community and its ability to maintain order is already weakened. Former prisoners may have weak ties to residents, be excluded from neighborhood life, and experience some level of discrimination or stigma because of their ex-convict status (Clear, 2007). Social and structural features of neighborhoods also affect a variety of individual postrelease outcomes (other than criminal activity) that are critical to a successful transition from prison, including finding and holding a job, avoiding use of alcohol and illegal drugs, finding affordable housing, and receiving physical and mental health care (Visher and Farrell, 2005).

Sociolegal Environment

Government policy represents a fourth dimension influencing individual transitions from prison to community. The three dimensions discussed above – individual, family, and community contexts – all operate within an environment that is defined to a large extent by a network of interrelated sociolegal policies.

Although it might be reasonable to assume that individuals who have completed their sentences are free from conviction-related constraints, former prisoners likely experience collateral effects of both the prison experience and a criminal record that creates additional obstacles to a successful transition from prison back to the community. Collateral consequences are the additional penalties tied to a conviction that greatly impact an individual's capacity to engage politically, economically, and socially upon their reentry to society (Travis, 2005). These consequences include barriers to housing, education, and employment, felony disenfranchisement, and ineligibility for public benefits. Collateral consequences are distinct from direct consequences of convictions in that they are not factored into the calculation of punishment or sentencing, and are triggered outside the jurisdiction of the courts.

These collateral effects (sometimes called 'invisible punishments') may include exclusions from certain professions (e.g., realtor or health care), access to public benefits (e.g., student loans, public housing, or food stamps), as well as loss of parental rights (Petersilia, 2003; Hagan and Dinovitzer, 1999; Travis, 2005). In the United States, the federal welfare law imposes a lifetime ban on anyone convicted of a drug-related felony from receiving federally funded food assistance and cash assistance. Unless a state passes legislation opting out of the federal law, individuals with these convictions are permanently barred from receiving benefits even if they have completed their sentence, overcome an addiction, been gainfully employed and subsequently laid off, or earned a certificate of rehabilitation or other form of clemency. Denying persons with felony drug convictions food, clothing, and shelter makes it more difficult for them to support themselves as they leave the criminal justice system and reenter society. These realities, in conjunction with conditions imposed upon their release (e.g., employment, in-person reporting, payment of restitution, fees, and fines), and the need for state-approved identification and transportation, make the first few weeks and months after release especially difficult.

The collateral consequence of a prison sentence on an individual's employment prospects after release can be especially harsh. Conventional wisdom states that finding a job is one of the most important elements for a person to successfully transit from incarceration back into the community. In fact, individuals returning home from prison often identify employment as the most important factor that helped them stay crime free. Yet, most former prisoners face formidable barriers to employment. With the rapidly expanding use of background checks, employers are routinely, and increasing illegally, excluding all job applicants who have criminal records from consideration, no matter how minor or dated their offenses (Holzer et al., 2004; Pager, 2003). Thus ensues a vicious cycle: people with criminal records find it harder to get a job, but not having a job makes it more likely they will reoffend.

Overcoming the sociolegal barriers that former prisoners face in returning to their communities – whether it is finding a job, locating housing, or receiving public assistance – is critical to successful reintegration.

Emerging Policy Lessons

The successful reintegration of former prisoners into society is arguably one of the most important components of an effective criminal justice system. Yet, it seems clear from the available data and discourse that local, state, and federal policymakers are unhappy with the existing approach that relies heavily on the overburdened correctional system and are looking for strategies that will be more successful than those currently in place.

Many scholars agree that the last decade has been a revival of the rehabilitative ideal. The essence of the rehabilitationist ideal, during its heyday a half century ago as well as today, is that it is both desirable and possible to achieve positive behavioral change in individual criminal offenders. Thus, the new version of this correctional principle focuses less on the prisons themselves than on the process of prisoner reintegration into the community. It also is a harder edged, 'evidence-based' rehabilitationism, characterized by cost-benefit assessment, meta-analyses, offender 'accountability,' risk assessment, graduated sanctions, coerced treatment, and an overriding emphasis on public safety. Indeed, today's rehabilitationism is battle hardened from the 'nothing works' attitude of the 1980s and 1990s.

Researchers have developed an impressive body of studies that underscore the modest effectiveness of a variety of interventions. The prevailing opinion is that recidivism can be reduced by 10–20%, depending on program design and implementation integrity. The more successful programs include in-prison and jail drug treatment, especially with a community component; cognitive behavioral therapy;

vocational education and training programs; employment training and job assistance; and adult basic education (Aos et al., 2006; Lipsey and Cullen, 2007; MacKenzie, 2006; National Research Council, 2007). In particular, there is growing consensus that practices focusing on *individual-level change*, including cognitive change, education, and drug treatment, are likely to be more effective than other strategies, such as programs that increase opportunities for work, reunite families, and provide housing (Andrews and Bonta, 2006; Lattimore et al., 2012; MacKenzie, 2006).

The design of these reentry programs is of critical importance. Experts recommend that programs: (1) focus on behavioral outcomes, targeting criminogenic factors and using positive reinforcements; (2) target high risk offenders; (3) use risk assessment instruments; (4) begin treatment in prison and provide continuity in the community; and (5) provide intensive interventions for at least 6 months (Andrews and Bonta, 2006; Solomon et al., 2008). Thus, the most effective strategies focus on dynamic criminogenic factors, are skill-oriented, are based on cognitive/behavioral models, and treat multiple offender deficits simultaneously. Approaches with the largest impact on recidivism, possibly 20% or greater, are likely to require intensive supervision in the community with mandatory treatment programs tailored to individual needs. Further, exiting prisoners are at the highest risk for recidivism immediately after release and could benefit from intensive services during that period (National Research Council, 2007). Yet, service receipt is typically highest during incarceration (Lattimore and Visher, 2009) and the services available in the community are rarely adequate to meet the needs of these individuals in the high risk period after release.

Evaluations also point to what interventions do *not* work. The evidence has been very consistent in establishing that contact-driven supervision, surveillance, and enforcement of supervision conditions have a limited ability to change offender behavior or to reduce the likelihood of recidivism (Aos et al., 2006; MacKenzie, 2006). Individuals placed on parole supervision after prison are no less likely to be rearrested than individuals released with no supervision (Solomon et al., 2005). Other programs that have no impact on recidivism include boot camps, some types of sex offender treatment, electronic monitoring, and life skills education (MacKenzie, 2006).

A properly and fully implemented reentry program may generate improvements in intermediate outcomes, but the impacts on recidivism may be minimal. Other reentry experts have expressed similar views that poor implementation may explain why criminal justice interventions often fail to produce positive impacts on recidivism (Lowenkamp et al., 2006; White, 2007; Wilson and Davis, 2006). Some of the most common implementation problems include ineffective delivery of services, poor matching of individual needs to program content, and failure to incorporate established principles for effective rehabilitative programming. The timing and dosage of service delivery may also be important (Bourgon and Armstrong, 2005; National Research Council, 2007). Thus, greater reductions in recidivism from prisoner reentry programs may only be achieved through attention to implementation issues.

State responses to prisoner reentry have expanded dramatically in the last decade and are embracing promising strategies that create coalitions of community organizations to support returning prisoners (Lattimore, 2007; Visher and Travis, 2011). The emergence of new models of community-based reentry interventions that mobilize the resources of family, community, and positive networks of formerly incarcerated individuals in addition to traditional service delivery systems and criminal justice agencies, hold the best promise for even greater reductions in new crimes. The complexity of the disadvantages and the depths of the needs confronting men and women being released from prison means that the reentry 'problem' cannot be addressed with a single generic program or intervention. Reentry experts are encouraging a broader focus on comprehensive reentry *strategies*, not programs (Lattimore, 2007; Travis, 2005; Visher, 2007). Such strategies usually involve multiple levels of government, coordination of efforts across agencies, and involvement of organizations that are traditionally not part of the reentry discussion (e.g., public health department, the business community, local community colleges). According to best estimates of the research community, if we could implement effective programs for all returning prisoners with all the resources needed, we could expect recidivism reductions on the order of 15–20%; greater reductions may be possible with targeted, community-based strategies.

In sum, our knowledge about what works for successful reintegration has increased dramatically in the last decade. Refocusing the justice system around a reentry perspective represents a fundamental paradigm shift. In the next decade, improvements in strategies to ensure high rates of successful reintegration require rigorous evaluations of promising interventions that combine individual-based and community-based approaches.

See also: Communities and Crime; Crime and Gender; Mass Imprisonment and Its Consequences; Prisons and Imprisonment; Rehabilitation of the offender.

Bibliography

Andrews, D., Bonta, J., 2006. The Psychology of Criminal Conduct, fourth ed. Lexisnexis/Matthew Bender, Newark, New Jersey.

Aos, S., Miller, M., Drake, E., 2006. Evidence-Based Adult Corrections Programs: What Works and What Does Not. Washington State Institute for Public Policy, Olympia.

Bourgon, G., Armstrong, B., 2005. Transferring the principles of effective treatment into a 'real world' prison setting. Criminal Justice and Behavior 32, 3–25.

Clear, T.R., 2007. Imprisoning Communities: How Mass Incarceration Makes Disadvantaged Neighborhoods Worse. Oxford University Press, New York.

Couturier, L.C., 1995. Inmates benefit from family services programs. Corrections Today 57, 100–106.

Freudenberg, N., Daniels, J., Crum, M., Perkins, T., Richie, B.E., 2008. Coming home from jail: the social and health consequences of community reentry for women, male adolescents, and their families and communities. American Journal of Public Health 98 (1), S191–S202.

Gadsden, V. (Ed.), 2003. Heading Home: Offender Reintegration into the Family. American Correctional Association, Lanham, Maryland.

Greifinger, R.B., Bick, J.A., Goldenson, J., 2007. Public Health behind Bars: From Prisons to Communities. Springer-Verlag, Dobbs Ferry, New York.

Hagan, J., Dinovitzer, R., 1999. Collateral consequences of imprisonment for children, communities, and prisoners. In: Tonry, M., Petersilia, J. (Eds.), Prisons. University of Chicago Press, Chicago, Illinois, pp. 121–162.

Hairston, C.F., 1988. Family ties during imprisonment: do they influence future criminal activity? Federal Probation 52, 48–52.

Holt, N., Miller, D., 1972. Explorations in Inmate-Family Relationships. California Department of Corrections, Sacramento, California.

Holzer, H.J., Raphael, S., Stoll, M., 2004. Will employers hire former offenders?: employer preferences, background checks, and their determinants. In: Western, B., Patillo, M., Weiman, D. (Eds.), Imprisoning America: The Social Effects of Mass Incarceration. Russell Sage Foundation, New York.

Iannotta, J.G., Ross, J.L., 2002. Equality of Opportunity and the Importance of Place: Summary of a Workshop. National Research Council, National Academy Press, Washington, DC.

Irwin, J., 1970. The Felon. University of California Press, Berkeley, California.

Lattimore, P.K., 2007. The challenges of reentry. Corrections Today 69, 88–91.

Lattimore, P.K., Visher, C.A., 2009. The Multi-Site Evaluation of SVORI: Summary and Synthesis. RTI International, Research Triangle Park, North Carolina.

Lattimore, P.K., Visher, C., Steffey, D., 2010. Prisoner reentry in the decade of the twenty-first century. Victims and Offenders 5, 1–15.

Lattimore, P.K., Barrick, K., Cowell, A., Dawes, D., Steffey, D., Tueller, S., Visher, C.A., 2012. Prisoner Reentry Services: What Worked for SVORI Evaluation Participants? RTI International, Research Triangle Park, North Carolina. https://www.ncjrs.gov/pdffiles1/nij/grants/238214.pdf.

Lindquist, C.H., Barrick, K., Lattimore, P.K., Visher, C.A., 2009. Prison Reentry Experiences of Adult Females: Characteristics, Service Receipt, and Outcomes of Participants in the SVORI Multi-Site Evaluation. RTI International, Research Triangle Park, North Carolina.

Lipsey, M.W., Cullen, F.T., 2007. The effectiveness of correctional rehabilitation: a review of systematic reviews. Annual Review of Law and Social Science 3, 297–320.

Lowenkamp, C.T., Latessa, E.J., Smith, P., 2006. Does correctional program quality really matter?: the impact of adhering to the principles of effective intervention. Criminology and Public Policy 5, 575–594.

Lynch, J.P., Sabol, W.J., 2000. Prison use and social control. In: Horney, J. (Ed.), Policies, Processes, and Decisions of the Criminal Justice System. U.S. Department of Justice, Washington, DC, pp. 7–44.

Lynch, J.P., Sabol, W., 2001. Prisoner Reentry in Perspective. Urban Institute, Washington, DC.

MacKenzie, D.L., 2006. What Works in Corrections: Reducing the Criminal Activities of Offenders and Delinquents. Cambridge University Press, New York.

National Commission on Correctional Health Care (NCCHC), 2002. The Health Status of Soon-To-Be-Released Inmates: A Report to Congress. http://www.ncchc.org/health-status-of-soon-to-be-released-inmates.

National Research Council, 2007. Parole, Desistance from Crime, and Community Integration. Committee on Community Supervision and Desistance from Crime. Committee on Law and Justice. National Research Council of the National Academies. The National Academy Press, Washington, DC.

Nelson, M., Deess, P., Allen, C., 1999. The First Month Out: Post Incarceration Experiences in New York City. Vera Institute of Justice, New York.

Ohlin, L., 1954. The Stability and Validity of Parole Experience Tables. Doctoral Dissertation. University of Chicago, Chicago, Illinois.

Pager, D., 2003. The mark of a criminal record. American Journal of Sociology 108, 937–975.

Petersilia, J., 2003. When Prisoners Come Home: Parole and Prisoner Reentry. Oxford University Press, New York.

Petersilia, J., 2005. From cell to society: who is returning home? In: Travis, J., Visher, C. (Eds.), Prisoner Reentry and Crime in America. Cambridge University Press, Cambridge, pp. 15–49.

Sampson, R.J., 2002. The community. In: Wilson, J.Q., Petersilia, J. (Eds.), Crime: Public Policies for Crime Control. ICS Press, Oakland, California.

Sampson, R., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, Illinois.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Shapiro, C., Schwartz, M., 2001. Coming home: building on family connections. Corrections Management Quarterly 5, 52–61.

Solomon, A.L., Kachnowski, V., Bhati, A., 2005. Does Parole Work? Analyzing the Impact of Postprison Supervision on Rearrest Outcomes. The Urban Institute, Justice Policy Center, Washington, DC.

Solomon, A.L., Jannetta, J., Elderbroom, B., Winterfield, L., Osborne, J., Burke, P., Stroker, R.P., Rhine, E.E., Burrell, W.D., 2008. Putting Public Safety First: 13 Strategies for Successful Supervision and Reentry. The Urban Institute, Justice Policy Center, Washington, DC.

Sullivan, E., Mino, M., Nelson, K., Pope, J., 2002. Families As a Resource in Recovery from Drug Abuse: An Evaluation of La Bodega de la Familia. Vera Institute of Justice, New York.

Travis, J., 2002. Invisible punishment: an instrument of social exclusion. In: Mauer, M., Chesney-Lind, M. (Eds.), Invisible Punishment: the Collateral Consequences of Mass Imprisonment. The New Press, New York, pp. 1–36.

Travis, J., 2005. But They All Come Back: Facing the Challenges of Prisoner Reentry. The Urban Institute Press, Washington, DC.

Visher, C., 2007. Returning home: emerging findings and policy lessons about prisoner reentry. Federal Sentencing Reporter 20, 93–102.

Visher, C., Farrell, J., 2005. Chicago Communities and Prisoner Reentry. The Urban Institute, Justice Policy Center, Washington, DC.

Visher, C., Mallik-Kane, K., 2007. Reentry experiences of men with health problems. In: Greifinger, R. (Ed.), Public Health Is Public Safety: Improving Public Health through Correctional Health Care. Springer-Verlag, London, pp. 432–447.

Visher, C., Travis, J., 2003. Transitions from prison to community: understanding individual pathways. Annual Review of Sociology 29, 89–113.

Visher, C., Travis, C., 2011. Life on the outside: returning home after incarceration. The Prison Journal 91, 102–119.

Visher, C., Yahner, J., La Vigne, N.G., 2010. Life after Prison: Tracking the Experiences of Male Prisoners Returning to Chicago, Cleveland, and Houston. The Urban Institute, Washington, DC.

Waller, I., 1964. Men Released from Prison. University of Toronto Press, Toronto, Ontario, Canada.

White, T.F., 2007. Creating leadership to facilitate implementation of evidence-based practice. Journal of Community Corrections 17, 13–28.

Wilson, J.A., Davis, R.C., 2006. Good intentions meet hard realities: an evaluation of the Project Greenlight Reentry Program. Criminology and Public Policy 5, 303–338.

Reference and Representation: Philosophical Aspects

Frank Jackson, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concepts of representation and representational content are introduced. Examples of things that represent are barometers, sentences, and beliefs. They represent that things are a certain way. This means that they can be thought of as being about something or as referring to something – namely, the certain way in question. The article addresses the connections between representation and reference, accounts of content, representation in language, and mental representation, and discusses some theories of reference for names. The article concludes with discussions of language of thought versus map views of mental representation, and the debate over representation and context.

Representation: The General Notion

What is common to all cases of representation is a correspondence between various states of what does the representing and various states of what is represented. Thus, the number of tree rings represents the age of a tree by virtue of the possibility of matching different tree ring states with different ages. In this case, the correspondence is a very simple one: *n* tree rings correspond to an age of *n* years. Likewise, there is a simple correspondence between various states of gas gages and various levels of gas in the tank to which the gage is connected. But there are highly complex cases as well. Inscriptions in secure codes of credit card numbers represent those numbers, but via very complex correspondences. When there are pairwise correspondences, there is the possibility of carrying information. This is how the number of tree rings can give information about the age of a tree. But, of course, the nature of the correspondence needs to be known for this possibility to be realized.

In many cases, the relevant correspondence relation has a causal underpinning that runs from what's being represented to what does the representing. For example, the level of gas in the tank causes the reading on the gage, and the orientation of a windsock is caused by the direction of the wind. But structures can represent how things will be; a drawing may represent where a playground will be located but is, obviously, not caused by where it will be located. There is representation in the wide sense when there are nonaccidental correspondences between states of affairs. There will typically be a causal explanation of these nonaccidental correspondences, but its precise nature varies from case to case. (For more on the variety, see Dretske (1995), Chapter 1, but note that he ties representation to purpose.)

Representational Content and Reference in General

Things that represent are said to have representational content, the content being how things are being represented as being. The usual way of capturing representational content is set theoretically: content is given by a division of all the complete ways that things might be into those that accord with how things are being represented as being and those that do not. Complete ways that things might be are the ways things might be in which everything is settled one way or the other. They are often called possible worlds, with exactly one of these possible worlds being the actual world (our world). The content is then the set of possible worlds that accords with how things are being represented as being. For example, if a gas gage represents a gas tank as half-full, its being half-full is the content, and that is identified with the set of possible worlds where the tank is half-full. There are, though, considerable differences of opinion about the nature of the possible worlds appealed to in this style of account. This is an issue in metaphysics that will not be discussed here. (For an accessible discussion of the main views, see Bigelow and Pargetter (1990), Chapter 4.)

One advantage of the set-theoretic approach to content is that it allows a simple account of when things are as they are being represented to be. Things are as they are being represented to be – that is, the representation is true, correct, or accurate – if and only if the way things actually are is a member of the content (i.e., if and only if the actual world is a member of the set of possible worlds that is the content of the representation).

A second advantage is that it allows a simple account of the widest sense in which states that represent are about, or refer to, this, that, or the next thing. Suppose that a barometer reading of X represents the surrounding air pressure as Y. Intuitively, the reading is about the air pressure being Y. This can be spelled out as follows. In every possible world where the reading is correct, the air pressure is Y. States that represent, are about, or refer to what is the case in all the worlds where things are as they are being represented as being. (For more on 'aboutness' and the difference between being about an object and a property, see Dretske (1995), Chapter 1, Section 5.)

A third advantage is that it is easy to extend the account to cover the fact that often what's represented is how things are from a point of view. A photograph represents how things are from where it was taken. Its content can be thought of as a set of centered possible worlds, where each center is as things are being represented to be from where the photograph was taken. If the photograph is of a dog, each centered world has a dog in front of its center.

Mental Representation

A footprint can be thought of as representing the shape of the foot that made it. But equally, it can be thought of as representing the weight of the person who made it, the shape of the

footprint itself half an hour earlier, the location of the person who made it at the time it was made, and so on. Almost any state that represents will have many representational contents.

Among the representational structures of especial interest to humans are their brains, especially those parts of our brains that constitute our powers of mental representation. In particular, humans believe that things are thus and so, and do so by virtue of the processes in their brains that constitute their beliefs representing that things are thus and so. These processes *qua* states of brains will themselves have many contents, including, for example, that the temperature of the surrounding brain fluid is thus and so, but only one will be the content of what is believed. To suppose otherwise would make what is believed absurdly indeterminate. Therefore, a major question in the philosophy of mind is what determines, out of all the contents of the brain structures that are beliefs, the one that is the content of those beliefs.

One popular answer is the functional role the structure plays, where this is understood as the role the structure plays in mediating between inputs to subjects (typically via their sense organs), outputs understood in terms of behavioral responses, and other internal states. However, there are many different ways of fleshing out this answer. (For a survey, see Braddon-Mitchell and Jackson (2007), Chapters 3 and 5.)

Linguistic Representation and Reference

Introduction

Language is an especially powerful representational device. With it, highly detailed claims are possible concerning how things are that far outrun, in terms of detail and precision, what can be done with barometers, windsocks, or gas gages. This is why sentences are such a marvelous source of information.

There is a syntactical distinction between sentences and parts of sentences that roughly corresponds to an important semantic or representational distinction. A sentence makes a claim about how things are. This is how it gets to be true or false: it is true when the way the sentence represents things to be is the way they are, and false otherwise. (This is why the biconditional statement " 'Snow is white' is true if and only if snow is white" is self-evident. The left-hand side is true if and only if things are as 'Snow is white' represents them as being, and a good way of stating how 'Snow is white' represents things as being is to use the sentence itself.) By contrast, parts of sentences, like 'snow' and 'is white,' contribute in a systematic way to how the sentences in which they appear represent things as being, but do not themselves represent how things are. This is why they are not themselves either true or false. 'Snow' and 'is white' can, however, be thought of as referring to, respectively, snow and white things. This is because the contribution that the presence of the word 'snow' makes to the conditions under which 'Snow is white' is true, that is, to how the sentence represents things as being, is such that the sentence cannot be true unless there is some snow. Similarly, the contribution that the presence of 'is white' makes is such that the sentence cannot be true unless there are some white things.

Many recent discussions have focused on the referential properties of names, and some authors restrict the use of the term 'refer' to names. They prefer to talk of the relation between predicates (and, sometimes, common nouns) and things in terms of whether the predicates are *true of* the things or whether the things *satisfy* the predicates.

The next section considers the three main theories of reference for names.

Theories of Reference for Names

A word like 'London' is used for something taken to have certain properties. This is why it is not used for water or Paris – it is known that they do not have the right properties. On the other hand, the large city that is the capital of England does have the right properties. In short, certain properties or descriptions are associated with the word 'London,' and this guides its use. This suggests a simple theory of reference for names in general: 'N' refers to that which satisfies the descriptions associated with 'N'. If we use 'D' for those descriptions, another way of stating this theory is by saying that 'N' is short for the definite description 'the D.'

A famous advantage of the *description theory of reference* is that it can explain why certain identity sentences are informative. Everything is, trivially, self-identical. How then can it be informative to be told that Hesperus is identical to Phosphorus? The answer is that the first name is associated with being the first star seen in the evening and the second name is associated with being the last star to disappear from the night sky as morning approaches, and it is informative to be told that one thing has these two properties. (The classic source for claims of this kind is Frege (1966).)

Recent discussions in the philosophy of language have raised a number of objections to the description theory of reference. (A good account of the major objections can be found in Devitt and Sterelny (1999). The classic source is Kripke (1980).)

First, if 'N' is an abbreviation of 'the D,' it should make no sense to suppose that N might have failed to be D. But, in fact, it typically makes perfect sense. For example, the descriptions associated with 'Shakespeare' include as a major part that he was the author of certain plays. (Surprisingly little is known about William Shakespeare's life.) But it makes perfect sense to suppose that Shakespeare did not write the plays associated with his name, and we might discover that this supposition is true and that Christopher Marlowe, say, wrote them. A similar point can be made for names of natural kinds. The description associated with 'water' is something like potable, clear liquid that fills lakes and falls as rain from the sky – this is roughly what is in a dictionary. However, if on some remote planet – traditionally called 'Twin Earth' by philosophers – there is found clear potable stuff in its lakes that falls from its sky and so on but is not H_2O, it is arguable that the right thing to say would be that this stuff – XYZ, as it is often dubbed – is not water. As might be said, "It has been discovered that water is not the only clear potable liquid that fills lakes, falls from skies etc. in the universe." (The classic source for Twin Earth is Putnam (1975).)

Second, if names are abbreviated definite descriptions, we should be able to shift what they refer to by making appropriate changes to what's supposed to be the case. 'The president of the United States in 1998' refers to Bill Clinton, but had he lost to Bob Dole in 1996, it would refer to Dole and not

Clinton. But it seems that the reference of names cannot be shifted; they are, as it is said, *rigid designators*. 'Bill Clinton' always refers to the very same person.

Description theorists have proposed a number of replies to these (and other) objections. For example, they have urged that it is wrong to conflate the description that settles the reference of a name with the description(s) most famously associated with a name. Although Shakespeare is most famous for the plays, the key description that settles the reference of the name is something like "the person who is a certain kind of causal origin of the use of the term 'Shakespeare.'" The debate over whether or not Shakespeare wrote the plays is the debate over whether the person who is a certain kind of causal origin of the occurrences of the name on the covers of plays and the like did indeed write the plays. (For a catalog of replies that the description theorists might make to various objections, see Jackson (1998, 2010), Chapter 1.)

The causal theory of reference is the best-known alternative to the description theory. The fundamental idea is that 'Aristotle' in our mouths refers to the person it does because of an initial act of naming that tags that person with a name, which then got passed on from speaker to speaker as the name to use for the tagged person until it ended up in our mouths. (The classic source is Kripke (1980).) Of course, the name itself underwent various changes as it passed from Ancient Greek to Modern English. A similar account is suggested to explain why 'water' refers to H_2O. H_2O was named 'water' (in ignorance of the fact that it was H_2O that was being named), and then the word was passed on as the word to use for that which had been so named. This account explains why the clear potable stuff that fills the lakes and falls as rain from the sky on Twin Earth is not water and is not what is referred to when the word 'water' is used, by appealing to the fact that it was not the kind initially named.

An obvious question at this point is how, exactly, the causal theory differs from a description theory that gives a prominent role to a description like "was named … and subsequently used for what had been so named"? (Some of these questions are addressed in Devitt and Sterelny (1999), from a perspective sympathetic to the causal theory, and in Kroon (2009), from a perspective sympathetic to the description theory.)

Finally, there is *the direct reference theory of names*. On this view, a name like 'Aristotle' refers directly to Aristotle. The only semantic property of a name is its bearer. A question for this theory is what makes it true that 'Aristotle' refers to Aristotle and not to Plato or my dog. Typically this question is answered by appealing to the kind of account given by the causal theory. Accordingly, sometimes the direct reference theory and the causal theory are seen as two sides of a single coin: the causal theory gives an account of how come 'Aristotle' refers to Aristotle, and the direct reference theory gives an account of the meaning of 'Aristotle.' An issue for the direct reference theory is what to say about empty names. It has sometimes been speculated that Socrates never existed. 'He' was merely an expository device of Plato. Although this is almost certainly false, it makes sense to suppose that it is true. But is this possible if 'Socrates' has no meaning apart from its bearer? (For references, more discussion, and dissent from the direct reference theory, see Devitt (1996), Chapter 4 and Jackson (2010), and for a defense of the direct reference theory and discussion of the whole debate, see Soames (2002).)

Mental Representation and Linguistic Representation

How Are They Connected?

As noted in this article, our minds represent; in particular, to believe is to represent that things are thus and so. Usually, the easiest way for humans to express how they believe things are is to use a sentence. For example, the sentence 'Clinton will not be impeached' was used in February 1999 by many commentators on American politics to express what they believed. It is obvious that there is an intimate connection between mental and linguistic representation. What is controversial is the extent to which mental representation requires linguistic representation.

One view is that the phenomenon of mental representation in general, and belief in particular, is independent of language; dogs, for example, have beliefs. This view can allow that there are some beliefs one could not have without having a language. It is, for instance, hard to make sense of a creature believing that there are a prime number of apples in the bowl in front of it in the absence of that creature having mastery of a complex coding system (i.e., a language).

However, some hold that belief in general requires language, where language means something substantially more sophisticated than the kind of signaling systems that, for example, dogs have. Sometimes, this seems to be, at bottom, a verbal stipulation about the use of the term 'belief': dogs have quasi-belief, which is very like belief except that it is not expressed in a public language. But sometimes it is clear that the view is that dogs do not have anything like belief. A problem for this view is the close connection between perception and belief. Perceptual experience seems to be shot through with belief-like representational content: the experience of seeing something as moving seems to be a state that by its very nature supports believing that something is moving. The challenge for advocates of the more radical view is to reconcile their view that dogs, for example, lack beliefs with the fact that dogs have perceptual experiences that represent how things are. (For a defense of the view that belief requires language, see Davidson (1984), Chapter 11.)

Content and Logical Equivalence

It was noted in this article that the standard way of capturing representational content is in terms of possible worlds. The problem of the content of logically equivalent sentences and beliefs raises an important question for this view. Intuitively, the sentences 'X is an equilateral triangle in Euclidean space' and 'X is an equiangular triangle in Euclidean space' differ in meaning. Likewise, intuitively, the belief that X is an equilateral triangle in Euclidean space and the belief that X is an equiangular triangle in Euclidean space are different beliefs. Couldn't you have one belief without having the other? But, on standard conceptions of possible worlds, 'X is an equilateral triangle in Euclidean space' and 'X is an equiangular triangle in Euclidean space' are true in exactly the same possible worlds, because the sentences are logically equivalent. The same goes for the corresponding beliefs.

There have been, broadly speaking, three reactions to this problem. One has been to bite the bullet and insist that the contents are the same. (The most worked-out version of this

response can be found in Stalnaker (1984).) A second has been to argue that, for the purposes of capturing the content of, for example, belief and language, we need to bring concepts into the picture. To believe that X is an equilateral triangle in Euclidean space is to believe that X falls under the concept *equilateral triangle in Euclidean space*, whereas to believe that X is an equiangular triangle in Euclidean space is to believe that X falls under the concept *equiangular triangle in Euclidean space*. These are different concepts because the concept *side* differs from the concept *angle*, and we can therefore allow that the beliefs (and the sentences that express them) have different contents despite being true at the very same worlds (see Peacocke, 1983; for an overview of the issues, see Braddon-Mitchell and Jackson, 2007). A third has been to argue that we should allow that there is more than one legitimate way of thinking of content, and meaning more generally. It is a mistake to demand that we make a choice (see Lewis, 1986, Chapter 1, Section 4).

Mental Representation and the Language of Thought

The functionalist answer to how a belief gets to have the belief content that it does was described earlier in this article, and it was noted that there are many views about how to spell out the key functional roles. However, a popular answer holds that an important part of any functionalist account is the way belief and desire work together to produce behavior: roughly, subjects behave in a way such that what they desire is satisfied if what they believe is true (Stalnaker, 1984; Braddon-Mitchell and Jackson, 2007). This is the key to what determines the content of both belief and desire.

An alternative approach – the language of thought hypothesis – appeals to the idea that creatures with beliefs and desires (and contentful mental states more generally) have a language of thought. When subjects believe that snow is white, they have in their heads a sentence in *mentalese* that says that snow is white. (This approach allows that dogs may have beliefs, because it is possible to have an internal language of thought while lacking a public language in which to express them.) This approach gives functional role a place in determining the difference, for example, between belief and desire, but content is treated via a *semantics for mentalese*. Consider, for example, subjects who believe that snow is white. They have in their heads a token that stands for snow (the mentalese word for snow) and a token that stands for whiteness (the mentalese predicate for whiteness), put together in a structure that makes up the mentalese sentence *snow is white*. What makes it true that they believe rather than desire that snow is white is the functional role this structure plays: it serves more of a 'responding to the world' role and less of a 'changing the world' role. (A classic source is Fodor (1988).)

This approach raises many important issues. One is how we should spell out the crucial relation between the words of mentalese and what they stand for. What makes it true that a structure in the brain is the mentalese for snow and not water, for example? One proffered answer is some kind of causal cum informational link between brain token and what is tokened, but the details are controversial. A second issue is whether neurological investigations might refute the language of thought hypothesis. Some cognitive scientists have argued that the way the brain represents the environment is more like the way maps represent than the way sentences represent (see Braddon-Mitchell and Jackson, 2007 and for a reader-friendly discussion of how research into the brain informs views of mental representation, see Clark, 2001).

Representation and Context

Representational content is often thought of as context dependent. For example, what a gas gage represents about the level of gas in the tank depends on the nature of the connection between gage and tank. Connected in one of the more usual ways, the gage having its pointer in the far left position indicates that the tank is empty, but connected differently that pointer position might represent that the tank is full. That is to say, identical gage configurations can have very different representational contents depending on the gage's context. Or, as it is often put, representational content with respect to the amount of gas in the tank does not supervene on a petrol gage's configuration. As it is put, the content is broad, wide, or context dependent.

It is equally possible, however, to think of any given reading as saying: either the gage is connected to the tank in way W_1 and the gas level is L_1, or the gage is connected in way W_2 and the level is L_2, or …. In this way of thinking, a way that is naturally adopted when the nature of the connection between gage and tank is unknown, the content is not dependent on the nature of the connection.

In other words, and speaking more generally, there are two essentially equivalent ways of thinking of the interplay of representational content and context. In the first, it looks like this:

In C_1, S represents that P_1
…
In C_i, S represents that P_i

In the second, it looks like this:

S represents that if C_1, then P_1
…
S represents that if C_i, then P_i

In the first way of thinking, content comes out as context dependent or broad; in the second, it comes out as context independent or, as it is said, narrow.

A pressing, currently much debated question about both mental and linguistic content is whether we should think of it in the context-dependent or in the context-independent way. Or, equivalently, is what a person believes and what a person claims about how things are (e.g., when they speak and write) a broad or narrow property of them?

The traditional view is that mental and linguistic content is narrow. A principal source for this view is a hypothesis often imagined by science fiction writers. Suppose the stimuli actually impinging on subjects via their environments are duplicated exactly, say, by a computer, then the subjects will not be able to tell the difference. This is the core idea behind flight simulators and virtual reality machines. Although these devices fall short of perfectly duplicating the impacts of their target environments on their subjects, their makers take it for granted

that were perfect duplication achieved, subjects would be unable to tell whether, for example, they were really flying or in a simulator. This suggests that subjects' mental lives, including how their mental states represent things as being, supervene on their nature from the skin in (or, perhaps, from their central processors in). How their mental states represent things as being is, therefore, a narrow property of subjects: two subjects that are identical from the skin in are identical in how they take their worlds to be, desire them to be, and so on. Moreover, if the content of sentences is given by the content of the mental states they serve to express, it will follow that linguistic content is narrow.

However, many philosophers, often influenced by the Twin Earth example discussed in this article, repudiate this suggestion. Consider, runs their argument, Mary and Mary's duplicate from the skin in, Twin Mary, on Twin Earth. They both use the word 'water' for the clear, potable stuff around them, but the conditions under which what is said is true, and equally those under which what is thought is true, differ. What Mary says and thinks when she uses the sentence "There is water nearby" is true if and only if there is H_2O nearby, but what Twin Mary says and believes will be true if and only if there is XYZ nearby. Hence, runs the argument, the contents of the words and thoughts of Mary and Twin Mary differ; ergo, content is context dependent.

The issue is complicated, however, by the fact that it is plausible that there is an indexical element to the use of a great many words, including 'water.' It is plausible that the word 'water' is used in part to mean the stuff that the users of the word have come across that is typically clear, potable, and so on. If this is right, a plausible explanation of why Mary and Twin Mary refer to different stuff in thought and talk when they use the word 'water' is that they and their language communities interact with different stuff when the word 'water' is used. That is, Mary refers to the clear, potable stuff that she and her community interact with, whereas Twin Mary refers to the stuff she and her (different) community interact with, and the reason they refer differently is that the stuff they interact with is different. This means that there is a sense in which what they say and think is the same when they say and think, "There is water nearby." Each says and thinks that some of the clear, potable stuff that she herself interacts with is nearby. This does not affect the point that the truth conditions of what they say and think differ, but they differ in the same way that the truth conditions of "I am bald" differ depending on who is speaking. There is a good sense in which each person who says or thinks that he is bald says or thinks the same thing, namely, that they are one of the bald people, despite the fact that the conditions under which what they say and think is true differ. If X says and thinks it, it is true just if X is bald; if Y says and thinks it, it is true just if Y is bald. (There is a large literature on this question; see Pessin and Goldberg (1996) and Braddon-Mitchell and Jackson (2007) for discussion and references.)

See also: Meaning and Rule Following; Methodological Individualism: Philosophical Aspects.

Bibliography

Bigelow, J., Pargetter, R., 1990. Science and Necessity. Cambridge University Press, Cambridge, UK.
Braddon-Mitchell, D., Jackson, F., 2007. The Philosophy of Mind and Cognition, second ed. Basil Blackwell, Oxford, UK.
Clark, A., 2001. Mindware: An Introduction to the Philosophy of Cognitive Science. Oxford University Press, Oxford, UK.
Davidson, D., 1984. Inquiries into Truth and Interpretation. Oxford University Press, Oxford, UK.
Devitt, M., 1996. Coming to Our Senses: A Naturalistic Program for Semantic Localism. Cambridge University Press, Cambridge, UK.
Devitt, M., Sterelny, K., 1999. Language and Reality, second ed. MIT Press, Cambridge, MA.
Dretske, F., 1995. Naturalizing the Mind. MIT Press, Cambridge, MA.
Fodor, J., 1988. Psychosemantics. MIT Press, Cambridge, MA.
Frege, G., 1966. On sense and reference. In: Geach, P., Black, M. (Eds.), Translations from the Philosophical Writings of Gottlob Frege, second ed. Philosophical Library, New York, pp. 56–78.
Jackson, F., 1998. Reference and description revisited. Philosophical Perspectives 12, 201–218.
Jackson, F., 2010. Language, Names, and Information. Wiley-Blackwell, Oxford, UK.
Kripke, S., 1980. Naming and Necessity. Harvard University Press, Cambridge, MA.
Kroon, F., 2009. Names, plans, and descriptions. In: Braddon-Mitchell, D., Nola, R. (Eds.), Conceptual Analysis and Philosophical Naturalism. MIT Press, Cambridge, MA, pp. 139–158.
Lewis, D., 1986. On the Plurality of Worlds. Basil Blackwell, Oxford, UK.
Peacocke, C., 1983. Sense and Content. Oxford University Press, Oxford, UK.
Pessin, A., Goldberg, S. (Eds.), 1996. The Twin Earth Chronicles. M. E. Sharpe, Armonk, NY.
Putnam, H., 1975. The meaning of "meaning". In: Putnam, H. (Ed.), Mind, Language and Reality. Cambridge University Press, Cambridge, UK, pp. 215–271.
Soames, S., 2002. Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity. Oxford University Press, Oxford, UK.
Stalnaker, R., 1984. Inquiry. MIT Press, Cambridge, MA.

Reference Group, History of

Barbara Hoenig, Innsbruck University, Innsbruck, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A person's reference group can be considered either as a source of normative orientation or as a group of social comparison in making a judgment, irrespective of one's group membership. The concept has mainly been applied in two theoretical traditions of research: While symbolic interactionists and social psychologists outlined the normative importance of reference groups for the development of the self, structural functionalist sociologists analyzed how material deprivation is estimated relative and with reference to other groups of comparison in the social structure. The concept has also played an important role in developing the discipline's methodology of social research.

A reference group is a group to whom one orients as a source of individual values and perspectives or with which one compares oneself in making a judgment or evaluation, irrespective of group membership. The concept designates a group of 'significant others' providing standards, expectations, and a frame of reference against which a person judges his or her own situation, and initially has been coined by Herbert H. Hyman in his social psychology of status judgment (Hyman, 1942). Hyman's distinction of a person's identificatory and judgmental group orientations has been analytically specified by Harold H. Kelley (1952) by distinguishing a normative from a comparative function of a reference group: The former indicates that whenever a group is "in a position to award recognition to the person or to withhold it," its function is that of "setting and enforcing standards" for a person motivated to be accepted by a group in showing conformity to its norms and role performance (Kelley, 1952: 412). Conversely, the concept's comparative meaning indicates the perceptual, judgmental aspect of a group's function "serving as or being a standard or comparison point against which the person can evaluate himself and others" (Kelley, 1952: 413). Although the very same reference group might serve both functions (Merton, 1957: 284), subsequent research has taken that distinction mainly as one between different types of reference groups. Turner outlined that the term's normative meaning can be regarded as synonymous for the process of role-taking, in which one adopts the group's standpoint as one's own, whereas interpreting a reference group as a point of a social comparison indicates a broader, more inclusive concept: Merely taking a reference group of others into account, and viewing it from a third-party standpoint, in order to accomplish one's purposes, estimate one's social standing, or decide whether to feel satisfied or not, does not necessarily mean to take the role of a member (Turner, 1956: 327). This enlarged meaning of the concept has been explicated by Robert Merton and Alice Kitt Rossi, particularly indicating the relevance of those structural conditions under which nonmembership groups constitute a significant frame of reference for a person (Merton and Kitt, 1950; Merton, 1957). Much research on reference group behavior has conceptually been clarified by them, in order to explain the selection of normative reference groups, and the consequences of that process in terms of social structure. In drawing such conceptual distinctions, they not only pointed to the structural sources of social comparisons for explaining people's feelings of satisfaction, legitimacy, and justice, but have also generated a distinct theory of reference group behavior. In empirical research, the concept has been applied to an array of social problems, for instance, adolescents' social conformity and deviance, college students' political attitudes and future orientation, immigrants' identity boundaries and social mobility, social scientists' communication patterns, role models' influence in public opinion formation and voting behavior, industrial workers' behavior toward trade unions, and the term's relevance for explaining the genesis and nature of mass revolutionary consciousness. Many of these studies are available in edited books, the volume of Hyman and Eleanor Singer still covering most of them, and the broad empirical use of the concept contributed to developing the methodology of social research as well (Newcomb and Hartley, 1947; Sherif and Sherif 1956; Merton and Lazarsfeld, 1950; Swanson et al., 1952; Eisenstadt, 1954; Lazarsfeld and Rosenberg, 1955; Lipset and Trow, 1957; Blau, 1956; Runciman, 1966; Hyman and Singer, 1968). John Urry (1973) provided a bibliography on reference group theory till the early 1970s, outlining a symbolic interactionist perspective toward the concept's past and future. Although the sociological interest in reference groups seems to have somewhat declined, recently the concept has been applied for analyzing social inequalities beyond national frames of reference (Whelan and Maitre, 2009). The influence of reference groups on an individual's attitude formation and behavior has mainly been examined in two theoretical traditions of research, on which we subsequently focus: On the one hand, symbolic interactionism and social psychology considered the normative importance of reference groups in the process of developing a self. On the other hand, structural and functionalist sociologists have analyzed the role of reference groups when material deprivation is estimated relative to other groups of social comparison. In the last section, the concept's contribution for developing the methodology of social research will be taken into account.

Normative Reference and the Development of the Self

Early anticipations of the reference group concept can be found in pragmatist theory and in symbolic interactionist's notions on the social self (James, 1890; Cooley, 1902; Mead, 1934), in Georg Simmel's studies on social groups (Simmel, 1908), and in William G. Sumner's distinction of in-group and out-group

(Sumner, 1940). As moralist reformer, William James characterized a 'potential' social self appraising itself with reference to others in "the pursuit of an ideal social self, of a self that is at least worthy of approving recognition by the highest possible judging companion, if such companion there be" (James, 1890: vol. 1, 315), highlighting the function of social distance when others influence our world views. Mead has encapsulated his central notion of the significance of a person's "generalized other" as "the generalized standpoint of the social group as a whole to which he belongs" (Mead, 1934: 138). He compared a person's role-taking, as that process generating a person's unity of the self, to the experience in a ball team (Mead, 1934: 154). According to George H. Mead, humans are able through symbols to communicate, to take on the role of the other, and become an object to oneself; this qualifies them to orient toward group norms and role performance when making judgments on their own (Mead, 1934: 138f.). While early in that process parents occupy an influential reference position, this then is shifted to peers and a generalized sense of society (Mead, 1934: 152ff.).

Examining students' development of political attitudes and behavior at Bennington College, Vermont, USA, Theodore M. Newcomb has differentiated positive and negative types of normative reference groups (Newcomb, 1943). The positive type involved the assimilation of group standards as a basis of self-appraisal; the negative one indicated that group norms may also be actively rejected and motivating the formation of counternorms, like in adolescent's rebellion against parents or hostile relations between groups. Newcomb systematically explored the ways in which college students related themselves to their community and with which consequences they did so. He observed that while freshmen tended to conservatism when measured against a number of public issues, this was less likely among junior and senior students; moreover, nonconservatism was more likely among those students strongly involved in college life and of higher status in college (Newcomb, 1943).

Hyman (1960) has emphasized the usefulness of the concept for developing public opinion research. Applying it to reference individuals as role models, he proposed that not only close friends and local neighbors, but also figures of public life as opinion leaders enact considerable influence. Moreover, he critically reflected opinion research's practical research routines that would benefit from a more extended concept of reference group theory. Already the study on voting behavior among citizens in the late 1940s in Elmira, New York, has shown that psychological identification with groups seen as important for oneself often predicts voting better than the mere measure of objective group membership (Berelson et al., 1954; Suchman and Menzel, 1955). This led Hyman to suggest that only the combination of judging one's location in class structure with expressing class awareness and subjective identification with a reference group can be treated as a powerful predictor of data on voting (Hyman, 1960: 394). Adumbrating that among nonmembers the ease of orienting oneself to a reference group norm is dependent on whether the social environment brings one closer to that group or not, and underscoring that in sampling large, heterogeneous groups we already deal with a differentiated structure, a set of norms, and also a group's perceived cohesiveness' (Hartley, 1960): The greater the perceived cohesiveness between norms of an actual membership group and those of an aspired reference group, the more likely an actor accepts the new group as reference group (Hartley, 1960).

From the standpoint of social psychology, Muzafer Sherif and Carolyn W. Sherif have introduced kindred terms such as points, scales, and particularly frames of reference as important in a person's attempt to gain acceptance by a specific reference group, and they applied the concept to issues of marginality and inconsistency in intergroup relations (Sherif and Sherif 1956: 143ff., 619ff.). Tamotsu Shibutani has characterized reference groups as perspectives available to the actor, defining a reference group as "a group whose outlook is used by the actor as the frame of reference in the organization of his perceptual field" (Shibutani, 1955: 565). While he considered the use of symbols, language, and communication in which this process is based, he did not take account of the structural arrangements that might condition these perspectives.

Social Comparisons and Material Deprivation

'The American Soldier' study (Stouffer et al., 1949a) has proved as particularly important for conceptually developing reference group theory. In a large empirical investigation of the life conditions and attitudes of US soldiers during World War II, Samuel A. Stouffer and his colleagues have been confronted with rather unexpected, counterintuitive findings: Soldiers in those branches characterized by less social mobility opportunities, like low-ranked officers, have reported greater satisfaction with their chances for promotion than their colleagues in the Air Corps where, objectively seen, career opportunities were much better. The authors explained this paradox with a 'theory of relative deprivation' proposing that the subjective feeling of deficiencies not only is a reflection of objective conditions but also depends on the respective standard of comparison. They assumed that low-ranked officers experienced greater satisfaction with their social situation because they compared themselves 'relative to others in the same boat with them' (Stouffer et al., 1949b: vol. 1, 251), among colleagues indicating those men of like-longevity, educational status, rank, and branch of Service, having not been promoted toward officership.

A theoretically oriented reexamination of these findings has been provided by Merton and Kitt Rossi (Merton and Kitt, 1950; Merton, 1957), examining the interplay between social theory and applied research from the sociological standpoint of structural functionalism. In systematically reflecting empirical cases from the 'American Soldier' study, the authors demonstrated that the theory of relative deprivation can be interpreted as specific case of a more general theory of reference groups of much broader sociological relevance. They emphasized the relative character of deprivation, where both feelings of dissatisfaction and of satisfaction result from standards of comparison with other groups situated in the social structure. "In general, then, reference group theory aims to systematize the determinants and consequences of those processes of evaluation and self-appraisal in which the individual takes the values or standards of other individuals and groups as a comparative frame of reference" (Merton and Kitt, 1950:

50f.). For Merton and Kitt Rossi, it has not been important that other groups are influential for a person's attitudes and behavior, but under which structural conditions membership and particularly nonmembership groups become influential. According to them, reference group selection is determined by institutional definitions of social structure, providing 'common comparative contexts for definition of the situation' among persons (Merton and Kitt, 1950: 62). At least a minimum of perceived commonalities, implying some knowledge of groups' norms and values, conditions that we take a particular reference group as a significant context for our judgments and evaluations (Merton and Kitt, 1950: 57). Dynamically seen, in case of multiple reference groups, persons have to make choices to resolve possible discrepancies and ambivalences arising out of contradictory norms and standards: In upward social mobility, the perception of, knowledge about, and adaption to norms and values of those groups are required in which one seeks acceptance and wishes to become a future member (Merton and Kitt, 1950: 57). The authors coined that phenomenon as 'anticipatory socialization' (Merton and Kitt, 1950: 87), a function of positive orientation to nonmembership groups, given a society's relatively open social structure. In the 'American Soldier' study, upward-mobility-oriented soldiers did not much conform to their immediate membership group's norms, but rather to the Army's 'official mores' and their more prestigious groups, in fact often having been at odds with in-group norms. Anticipatory socialization's consequences for a person result in finding readier acceptance by a group, making an easier adjustment to it, at the same time alienating from one's in-group. For the in-group to which a person belongs, that might be dysfunctional, thus putting members' positive orientation toward out-group norms under restraint (Merton and Kitt, 1950: 87). For the social system, reference group behavior's consequences are related to the legitimacy ascribed to institutional arrangements, also affecting the range of intergroup and interindividual comparisons (Merton and Kitt, 1950: 90). The authors intended to clarify the social processes sustaining or curbing orientations to nonmembership groups, and detach sociological concepts from a language of sentiment. In this way, Merton and Kitt Rossi have proved how reference group theory can be applied to a variety of areas of social life like class mobility, migration, social organization, and control.

Some years later, Merton (1957) examined empirical research that meanwhile had spread in the relevant field by specifying central concepts of groups and a provisional list of group properties. Drawing a distinction to reference group individuals as role models, he defined groups and group membership in terms of objective and subjective criteria: the former that social interaction among a number of people occurring in some patterned, frequent form, the latter that persons define themselves, and are similarly defined by others, as members of a group (Merton, 1957: 286). Structurally, Merton distinguished groups from collectivities, their members sharing a sense of solidarity while lacking social interaction with each other, and from social categories, persons with like social characteristics and statuses, but not necessarily a common body of norms (Merton, 1957: 299). Dynamically, he considered how boundaries of group membership and degrees of nonmembership can change in response to situational contexts. Concerning criteria for the latter, he drew on Simmel's notion of a group's 'completeness' (Simmel, 1908: 55), measured as a ratio of potential to its actual members, and indicated a direct relationship between a social group's completeness and power. Extensively he elaborated Simmel's notion of a group's 'capability to be surveyed' (Simmel, 1908: 50) as a group property relevant for regulation and control in social organization and the public, particularly when structurally hindered flows of information play an important role (Merton, 1957: 322, 336ff.). Merton's term of social visibility refers to the degree to which a person's status identity or a group's norms and role performance are readily visible or observable for others. Social visibility is also important to understand structural variability in knowledge about group norms, because that knowledge conditions to which extent a group can be taken as frame of normative reference at all (Merton, 1957: 322). Apart from visibility, Merton outlined objective eligibility and subjective attitudes toward becoming a member as criteria for nonmembership, resulting in structurally different kinds of nonmembers, for instance, 'candidates' or 'marginal men,' 'autonomous,' or 'antagonistic nonmembers' (Merton, 1957: 290). He reflected that vernacular language registers ex-members often with affectively toned terms that establish strong in-group versus out-group differentiations. Concerning determinants for the selection of reference groups, Merton has been carefully taking account of an array of empirical studies available on the topic, clarifying their conceptual implications for social structure, and pointing to the need of comparative studies of reference group behavior within disparate structural contexts (Merton, 1957: 306f.). However, he did not much address how his reference group theory is linked to social roles and to his general social theory. While his status- and role-set theory may be taken as part of his reference group theory, in reverse, Merton's reference group theory can also be regarded as "a consequent set of ideas about how status-and-role structures actually work" (Crothers, 1987: 93).

Reference Groups in the Methodology of Social Research

In order to provide an insight into how the concept helped to evolve the methodology of social research, from the abundance of studies on reference group behavior we consider only some, anticipated as combinatory cases along the axes of micro-macro- and structure-process analysis. First, the analysis of structurally induced conflicts and contradictions among multiple reference groups at microsocial level has proved to be important for conceptually developing role theory. In contrast to Merton's tendency to see conflicts as disturbing forces in need to be harmonized by mechanisms of social integration, his former student Ruth Laub Coser underlined the possibility of gaining individual autonomy, cognitive flexibility, and intellectual creativity in that process: Precisely because actors with 'complex' instead of 'simple' parochial role-sets frequently experience ambiguities among them, they develop competencies in dealing with these situations; this grants them higher individual autonomy in socially differentiated modern societies (Coser, 1975). Second, macrostructural conflicts between reference groups recently have found some attention in

analyzing social inequalities at national and supranational level. Still there are lively scholarly debates whether, by increasing globalization in times of crisis, people have begun to perceive themselves as part of larger stratification systems beyond a national framework, or whether their relative estimation of material deprivation rather depends on a national context (for instance, Whelan and Maitre, 2009). Third, taking microdynamics of shifting reference groups during one's lifetime into account, it has been examined how in adolescence, norms of peers as reference groups become more important than those of parents. Conflicting expectations and standards of parents and peers frequently create cross-pressure situations for adolescents; demanding standards of parental norms sometimes are reflected in those of a peer group as well (Rosen, 1955). Fourth, the dynamic macrocase of reference groups concerns the possibility that people dissatisfied with their social situation are motivated to engage in social movements with the potential to transform their social world. Urry (1973) applied reference group analysis to the empirical case of Indonesia in the first half of the twentieth century, in order to theoretically explain why actors experience dissent, organize themselves in social movements, and develop a mass revolutionary consciousness or not. He quoted Henry D. Thoreau who once has noted: "If a man does not keep pace with his companions, perhaps it is because he hears a different drummer" (Thoreau, quoted in Urry, 1973: 17). The need for understanding that drum as well has been convincingly outlined by Urry by emphasizing symbolic interactionism's fruitfulness in exploring the reference group concept. However, some of his suggestions for reformulating reference group theory seem to come closer to a structural functionalist tradition than he might have intended.

Methodologically seen, the reference group concept played an essential role in sociologically reflecting and bridging the two-way relationship of social theory and empirical research. Merton and Kitt Rossi (1950) outlined the concept's potential for structural and functionalist sociology, also distancing it from routines of small group research dominant in much social psychology of their time. The authors have frequently emphasized the concept's importance for comparative studies on social structure and patterns of reference group behavior of like-statused individuals. While they have explored the theoretical relevance of the concept for a broader scope of sociological research, conversely, reference groups also served as part of more abstract concepts' operational definitions. In order to get a clearer prospect of how to conduct research on Parsonian concepts of universalism and particularism in people's action orientation (Parsons and Shils, 1951: 76–88), Peter M. Blau (1962) interpreted them in the language of reference groups, hence necessarily referring to attributes of social structure, in contrast to cultural norms or personal motivational standards. The 'American Soldier' study's tremendous influence on an evolving discipline has also been reflected in the reader *Language of Social Research* (Lazarsfeld and Rosenberg, 1955); it also adumbrates how the broad use of the reference group concept apparently helped to develop procedures for social research. This alludes to generating social indices such as of mobility rates, group cohesion, social differentiation, and integration for studying the bearing of variations of social contexts upon the individuals' behavior situated in that social structure (Kendall, 1955; Merton and Kitt, 1950: 81ff.). The reference group concept also clarified the interplay of objective and subjective group properties in multivariate analyses on voting (Berelson et al., 1954; Suchman and Menzel, 1955) and the analysis of change through time in processes of public opinion and attitude formation (Lazarsfeld et al., 1955). More generally, Merton has taken account of the methodological perspective from which reference groups are examined: Since reference group behavior involves two-sided social relationships, similar to a theory of influence, it required a "simultaneous analysis both of the individuals adopting various reference groups and of the groups which provide these frames of reference" (Merton, 1957: 328). The reference group concept is useful for social research, because it provides an explanation both for the structural conditions under which social groups, particularly nonmembership groups, influence a person's values, attitudes, and behavior, and for the consequences of that process within the larger social structure.

See also: Functionalism, History of; Motivation: History of the Concept; Pragmatist Social Thought, History of; Self: History of the Concept; Social Structure: History of the Concept; Structuralism.

Bibliography

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N., 1954. Voting: A Study of Opinion Formation in a Presidential Campaign. University of Chicago Press, Chicago.

Blau, P.M., 1956. Social mobility and interpersonal relations. American Sociological Review 21 (3), 290–295.

Blau, P.M., 1962. Operationalizing a conceptual scheme: the universalism-particularism pattern variable. American Sociological Review 27 (2), 159–169.

Cooley, C.H., 1902. The Two Major Works of Charles H. Cooley: Social Organization. Human Nature and the Social Order. With an Introduction by Robert Cooley Angell. The Free Press, Glencoe, IL.

Coser, R.L., 1975. The complexity of roles as a seedbed of individual autonomy. In: Coser, L.A. (Ed.), The Idea of Social Structure. Papers in Honor of Robert K. Merton. Harcourt Brace Jovanovich, New York, pp. 237–264.

Crothers, C., 1987. Key Sociologists: Robert K. Merton. Tavistock, London.

Eisenstadt, S.N., 1954. Reference group behavior and social integration: an explorative study. American Sociological Review 19, 175–185.

Hartley, R., 1960. Norm compatibility, norm preference, and the acceptance of new reference groups. Journal of Social Psychology 52, 87–95.

Hyman, H.H., 1942. The psychology of status. Archives of Psychology 269, 47–57.

Hyman, H.H., 1960. Reflections on reference groups. Public Opinion Quarterly 24 (3), 383–396.

Hyman, H.H., Singer, E. (Eds.), 1968. Readings in Reference Group Theory and Research. The Free Press, New York.

James, W., 1890. The Principles of Psychology, 2 vols. Smith, New York.

Kelley, H.H., 1952. The two functions of reference groups. In: Swanson, G.E., Newcomb, T.M., Hartley, E.L. (Eds.), Readings in Social Psychology. Holt, New York, pp. 410–414.

Kendall, P.L., 1955. A review of indicators used in 'The American Soldier'. In: Lazarsfeld, P.F., Rosenberg, M. (Eds.), The Language of Social Resarch: A Reader in the Methodology of Social Research. The Free Press, Glencoe, IL, pp. 37–40.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., 1955. The process of opinion and attitude formation. In: Lazarsfeld, P.F., Rosenberg, M. (Eds.), The Language of Social Research: A Reader in the Methodology of Social Research. The Free Press, Glencoe, IL, pp. 231–242.

Lazarsfeld, P.F., Rosenberg, M. (Eds.), 1955. The Language of Social Research: A Reader in the Methodology of Social Research. The Free Press, Glencoe, IL.

Lipset, S.M., Trow, M., 1957. Reference group theory and trade Union Wage Policy. In: Komarovski, M. (Ed.), Common Frontiers of the Social Sciences. The Free Press, Glencoe, IL, pp. 391–411.

Mead, G.H., 1934. Mind, Self and Society. University of Chicago Press, Chicago.
Merton, R.K., 1957. Continuities in the theory of reference groups and social structure. In: Merton, R.K. (Ed.), Social Theory and Social Structure. The Free Press, Glencoe, pp. 281–386.
Merton, R.K., Rossi, A. Kitt., 1950. Contributions to the theory of reference group behavior. In: Merton, R.K., Lazarsfeld, P.F. (Eds.), Continuities in Social Research: Studies in the Scope and Method of 'The American Soldier'. The Free Press, Glencoe, IL, pp. 40–105.
Merton, R.K., Lazarsfeld, P.F. (Eds.), 1950. Continuities in Social Research: Studies in the Scope and Method of 'The American Soldier'. The Free Press, Glencoe, IL.
Newcomb, T.M., Hartley, E.L. (Eds.), 1947. Reading in Social Psychology. Henry Holt and Company, New York.
Newcomb, T.M., 1943. Personality and Social Change: Attitude Formation in a Student Community. Dryden, New York.
Parsons, T., Shils, E., 1951. Toward a General Theory of Action. Harvard University Press, Cambridge, MA.
Rosen, B.C., 1955. Conflicting group membership: a study of parent-peer group cross-pressures. American Sociological Review 20 (2), 155–161.
Runciman, W.G., 1966. Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth-Century Britain. Routledge & Kegan, London.
Sherif, M., Sherif, C.W., 1956. An Outline of Social Psychology. Revised Edition, with a New Introduction by Garner Murphy. Harper & Row, New York and Evanston.
Shibutani, T., 1955. Reference groups as perspectives. American Journal of Sociology 60 (6), 562–569.
Simmel, G., 1908. Über die quantitative Bestimmtheit der Gruppe. In: Soziologie. Untersuchungen über die Formen der Vergesellschaftung. Duncker & Humblot, Leipzig, pp. 47–133 (1908. Engl. The number of members as determining the sociological form of the group. American Journal of Sociology, vol. 8, pp. 1–46.).
Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A., Williams, R.M., 1949a. The American Soldier: Adjustment during Army Life, vol. 1. Princeton University Press, Princeton.
Stouffer, S.A., Lumsdaine, M.H., Williams, R.M., Smith, M.B., Janis, I.L., Star, S.A., Cottrell, L.S., 1949b. The American Soldier: Combat and its Aftermath, vol. 2. Princeton University Press, Princeton.
Suchman, E.A., Menzel, H., 1955. The interplay of demographic and psychological variables in the analysis of voting surveys. In: Lazarsfeld, P.F., Rosenberg, M. (Eds.), The Language of Social Research: A Reader in the Methodology of Social Research. The Free Press, Glencoe, IL, pp. 148–155.
Sumner, W.G., 1940. Folkways. A Study of the Sociological Importance of Usages, Manners, Customs Mores and Morals. Ginn, Boston, MA.
Swanson, G.E., Newcomb, T.M., Hartley, E.L. (Eds.), 1952. Readings in Social Psychology, revised ed. Henry Holt and Company, New York.
Turner, R.H., 1956. Role-taking, role standpoint and reference group behavior. American Journal of Sociology 61 (4), 316–328.
Urry, J., 1973. Reference Groups and the Theory of Revolution. Routledge, London.
Whelan, C.T., Maitre, B., 2009. Europeanization of inequality and European reference groups. Journal of European Social Policy 19 (2), 117–130.

Reflective Equilibrium

Stelios Virvidakis, University of Athens, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N. Daniels, Volume 19, pp. 12872–12877, © 2001, Elsevier Ltd.

Abstract

The article deals with the notion of reflective equilibrium, understood as a state of coherence of our thoughts about one or more issues, resulting from a deliberative process of mutual adjustment of beliefs, principles, theories, and arguments. It focuses on its use as the central idea of a coherentist method of justification employed mainly in moral philosophy and elaborated in the writings of John Rawls. It discusses a variety of construals of the concept of reflective equilibrium, as involving a normative or descriptive approach, a wide or a narrow scope, a radical or conservative attitude, and a general or individual range of application. It summarizes the main general criticisms of the method, highlighting the problems of eventual circularity, lack of contact with reality and relativism, and more particular ones disputing the supposedly foundational character or questioning the reliability of the considered judgments appealed to in its pursuit, and pointing out the vagueness and the vacuity of the goal it aims at when it is presented in its wide version. A study of some of the main possible answers to these criticisms leads to the conclusion that the quest for a reflective equilibrium may not amount to much more than a regulative ideal, but it may still be regarded as the best methodological approach for ethics and for philosophy in general, when one takes into consideration the foundationalist and scientific alternatives. It seems to be especially useful for the adjudication of debates in applied ethics and especially in bioethics, where it allows us to avoid both deductions from abstract principles and an appeal to casuistic moves or extreme particularism.

The Concept of Reflective Equilibrium

A 'reflective equilibrium' is a state of coherence of our thoughts about one or more issues, resulting from a deliberative process of mutual adjustment of beliefs, principles, theories, and arguments. This deliberative process, involving the balancing and revising judgments at all levels of generality, until coherence is achieved, can be described as the 'method of reflective equilibrium.' It can be employed in any area of philosophical inquiry, but has become prominent in ethics, after its elaboration and application by John Rawls in the development of his moral and political theory. In fact, it can be interpreted as a particular form of a coherentist approach to the justification of beliefs and principles and contrasted to foundationalist accounts. In what follows, we shall try to present the main characteristics, advantages, and shortcomings of the method, focusing on the Rawlsian legacy. We will examine alternative construals of the idea of a reflective equilibrium, and we will take into consideration various positive and negative assessments of its usefulness with a view to providing a critical evaluation of its contribution to philosophical methodology.

The Method and Its Recent History

The term 'reflective equilibrium' was first introduced by John Rawls in *A Theory of Justice* (Rawls 1971, 1999a). However, the first clear description of the method of a mutual adjustment of beliefs and principles had been already provided by Nelson Goodman's in his *Fact, Fiction and Forecast*. Goodman argues that the principles of both deductive and inductive inference are justified through the process of bringing rules and particular inferences into agreement with each other. Thus, 'a rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend' (Goodman, 1979: p. 64). It would be pointless to try to provide a justification appealing to some form of foundation, self-evident starting points or a grounding basis.

Now, Rawls points out that the method of reflective equilibrium can be traced back to Aristotle's dialectical argumentation, which is described in his *Topics*, and according to many scholars is the main methodological approach adopted in the *Nicomachean Ethics*. He also compares it to Henry Sidgwick's argumentative strategy in *The Methods of Ethics* (Rawls, 1999a: p. 45n). Already in his *Outline of a Decision Procedure for Ethics*, published in 1951, he had discussed the problem of the justification of particular moral judgments by appealing to justifiable principles. He had observed that such principles "explicate the considered judgments of competent judges" and that the "reasonableness of a principle is tested by seeing whether it shows a capacity to become accepted by competent moral judges after they have freely weighed its merits by criticism and open discussion, and after each has thought it over and compared it with his considered judgments" (Rawls, 1999b: pp. 10–11). His conception of the method is specified in detail and plays a crucial role in the course of the construction of his theory of justice as fairness and of the justification of its essential components. The appeal to a hypothetical contract situation, which we are supposed to imagine through the thought experiment of an original position of rational contractors, leading to the formulation of basic principles of justice, as well as the use of the ideas of 'public reason' and of a 'well-ordered society,' at further stages of his reasoning, are tested within the broader methodological framework of the pursuit of a reflective equilibrium with our considered judgments about justice.

We are supposed to proceed as follows: We first try to identify a set of considered judgments about justice, that is, confident, stable judgments based on adequate information and arrived at through careful thinking protected from

distortions due to the interference of emotional factors, partiality or self-interest (Rawls, 1999a: p. 42; 1999b: p. 6). In fact, they are grasped and formulated at this stage, more or less immediately, without the appeal to any principles. We then seek principles which best 'account for' these judgments, and we soon find it necessary to engage in a process of mutual adjustment in order to remove divergences and overcome conflicts which we are likely to encounter. We realize that we have to move 'back and forth' between principles and judgments until we reach an optimal state where we no longer need to revise the principles we have come up with, or modify, 'prune,' or even give up some of the judgments the principles fail to account for (Rawls, 1999a: p. 18; Scanlon, 2003: p. 141).

However, the analysis just provided may give the impression that we are simply interested in looking for an equilibrium between a limited number of descriptions which 'more or less match one's existing judgments except for minor discrepancies' and we do not have to take into account 'all possible descriptions to which one might plausibly conform one's judgments together with all relevant philosophical arguments for them' (Rawls, 1999a: p. 43). It is the latter conception of the method which, according to Rawls, we are primarily interested in while engaging in the quest for a satisfactory moral theory, and its application could make possible a 'radical shift' of our judgments or principles. To be sure, it must be admitted that it is probably impossible to carry out fully the philosophical enterprise it seems to entail. The state of reflective equilibrium should be regarded as an unattainable ideal which we could only approximate in our indefinitely continuing work (Rawls, 1996: pp. 97, 385). The distinction between these two different accounts of the equilibrium one should aim at, already sketched in *A Theory of Justice*, is further elaborated in Rawls' later works. It is first presented as an opposition between 'narrow' and 'wide' reflective equilibrium in his *The Independence of Moral Theory*, published in 1974 (Rawls, 1999b), and its importance is emphasized again in *Justice As Fairness: A Reassessment* (Rawls, 2001: pp. 30–31). As we shall see, it is wide reflective equilibrium that becomes the central methodological notion adopted by contemporary philosophers.

Construals of Reflective Equilibrium

In fact, we have to distinguish among different construals of the concept of reflective equilibrium which have already been mentioned, or at least hinted at in our discussion so far. We are interested in highlighting interpretative options which must be taken into account when we try to deal with some of the main criticisms mounted against it.

Normative vs Descriptive

We have pointed out that the quest for a reflective equilibrium is invoked in a process of *justification* of beliefs and principles at all levels of generality and we have spoken of the deliberative process which we follow in our efforts to reach it. We are thus adopting a *normative* or *deliberative* conception, which goes beyond a mere description of our practices and aims at establishing the correctness of the judgments of principles in question (Scanlon, 2003: pp. 40–43). In fact, there are certain passages in *A Theory of Justice* where Rawls seems to be claiming that the principles arrived at in a state of reflective equilibrium help us *describe* our sense of justice in a way analogous to principles of grammar which reflect a sense of grammatical ability enabling a speaker of a language to recognize well-formed sentences (Rawls, 1999a: p. 41). However, the normative aims of moral and political philosophy such as those practiced by Rawls are different from the descriptive and explanatory goals of moral psychology or moral anthropology, even if philosophers should not ignore the results of the empirical investigations of the latter. In fact, this becomes clear when we realize that we have to broaden the scope of our procedure to include much more than the original set of considered judgments and the principles supposedly accounting for them. The descriptive construal seems to be based on a narrow understanding of the equilibrium to be achieved (Daniels, 1979: p. 258n).

Wide vs Narrow

We have seen that Rawls himself favored a *wide* rather than a *narrow* construal of his key idea. This construal has been further elaborated by Norman Daniels in his seminal paper *Wide Reflective Equilibrium and Theory Acceptance in Ethics* (Daniels, 1979, 1996a). Daniels speaks of an ordered triple of sets of beliefs, including sets of considered moral judgments, moral principles and relevant background theories, and describes four different levels of argumentative moves matching judgments, principles, and theories in ways which allow us to test alternatives and the feasibility of the principles selected. Daniels adds an important independence constraint, stipulating that "some nontrivial interesting portions of the set of considered moral judgments that constrains the background theories and of the set that constrains the moral principles should be disjoint" (Daniels, 1979: pp. 258–260).

An even more detailed survey of all the factors playing an important role in the procedure is provided by Michael DePaul in his critical analysis of the method of reflective equilibrium. Thus, we would have to take into account and seek a satisfactory match involving all of the following: (1) the inquirer's considered moral judgments, (2) background philosophical theories, such as theories regarding the nature of persons and personal identity, (3) background social scientific theories, such as theories regarding the role that a moral conception plays in a society, (4) the various alternative moral theories that have been developed, and defended, within the inquirer's cultural tradition, and possibly within alien traditions with which the inquirer's tradition has had to interact, and (5) the various philosophical arguments that have been constructed for and against all these moral conceptions (DePaul, 1993: p. 137).

Radical vs Conservative

As we shall see below, one of the criticisms addressed to defenders of the method of reflective equilibrium is related to a fear of conservatism. In fact, it may be thought that the pursuit of an equilibrium between moral convictions expressed

by considered judgments and moral principles entails an attempt to preserve the former by consciously or unconsciously selecting such principles that are *made* to account for them. Indeed, one may get this impression if one does not go beyond narrow reflective equilibrium. The 'partial,' limited fit, which could be achieved more or less easily at the first stage of matching of levels seems to involve minor adjustments. However, things are very different, if one is committed to the search for a wide equilibrium which, as we have already observed, often requires major revisions. In Rawls' words, "People have considered judgments at all levels of generality, from those about particular situations and institutions up through broad standards and first principles to formal and abstract conditions on moral conceptions… Although in order to get started various judgments are viewed as firm enough to be taken provisionally as fixed points, there are no judgments on any level of generality that are in principle immune to revision. Even the totality of particular judgments are not assigned a decisive role …" (Rawls, 1999b: p. 289). Moreover, philosophers who take seriously the far reaching implications of the commitment to the pursuit of an ideal, most probably unattainable state of wide reflective equilibrium realize its radical thrust which could bring about the reversal of well entrenched beliefs (Daniels, 1996a; Scanlon, 2003). Hence, it turns out that the method elaborated by Rawls and his disciples is quite different from its alleged predecessors, such as Aristotle's dialectic which does not allow for the possibility of a wholesale revision of reputed opinions (*endoxa*).

General vs Individual

As originally presented, the pursuit of reflective equilibrium concerns the individual subject of thought and action seeking to impose a coherent order on his or her beliefs at various levels of generality. Each person, and more specifically each citizen involved in reflection on political conceptions of justice, is invited to engage in a philosophical inquiry which proceeds by a continuous mutual adjustment of his or her considered judgments, principles, theories, and arguments. However, although Rawls defends a form of moral constructivism rather than realism, and despite the fact that, in his later work on political liberalism, he limits the universalist pretensions of his claims about a just society, he does think that we should try to "work up from the fundamental ideas implicit in the political culture, a public basis of justification that all citizens as reasonable and rational can endorse from within their own comprehensive doctrines. If this is achieved we have an overlapping consensus of reasonable doctrines and with it the political conception affirmed in reflective equilibrium" (Rawls, 2001: p. 29). In fact, if we assume that "each citizen in a well-ordered society has achieved wide reflective equilibrium" and since "citizens recognize that they affirm the same public conception of political justice," we will agree that "reflective equilibrium is also general: the same conception is affirmed in everyone's considered judgments. Thus, citizens have achieved general and wide, or what we may refer to as full, reflective equilibrium." In a well-ordered society, "the public point of view from which all citizens can adjudicate their claims is mutually recognized as affirmed by them all in full reflective equilibrium" (Rawls, 2001: p. 31, 1996: pp. 384–385n). As Rawls puts it, in a reply to Habermas, "this equilibrium is fully intersubjective: that is, each citizen has taken into account the reasoning and arguments of every other citizen" (Rawls, 1996: p. 385n).

We thus move to a construal of reflective equilibrium as involving the mutual adjustment of judgments, principles, theories, and arguments at an interpersonal, social level. This essential 'horizontal' expansion of the wide conception to include our fellow citizens, and more generally our fellow human beings, is crucial for overcoming any purely individualistic model of justification. It seems to conform fully to the requirement of reciprocity built in all contractualist normative theories. However, it should be noted that it renders the unattainable ideal of a perfectly satisfactory match or fit even more distant, broadening the 'indefinitely continuing process' (Rawls, 1996: p. 97) in the pursuit of expressions of an overlapping consensus, extending to well-ordered societies on a global scale.

Criticisms of Reflective Equilibrium

General Worries about Coherentism

Some of the general objections to reflective equilibrium are instances of familiar strictures directed against all coherentist conceptions of justification. At this point, we shall present them briefly, summarizing some of the main responses which have been provided by supporters of such conceptions and can also be invoked in the case of reflective equilibrium.

One is often reminded of the *circularity* besetting the reasoning procedure which is employed in the pursuit of reflective equilibrium. It is a well-known problem recognized by defenders of coherentism and by Goodman himself in his original description of the method (Goodman, 1979). However, it can be pointed out that the entire set of interconnected argumentative moves, deployed in all directions, which aim at balancing and matching elements in a nonlinear manner do not form a vicious circle. The justification we are seeking is achieved precisely when we manage to construct a web, the nodes of which hold together in a way that does not require further adjustment.

Another reason for wanting to reject the method is the impression that the entire set of judgments, principles, theories, and arguments, in wide reflective equilibrium does not bear the appropriate relation to the moral reality it is supposed to reveal. Indeed, as is the case with coherent sets of beliefs, a typical complaint is that *no contact with reality* is established through the strengthening of internal relations leading to equilibrium. On the contrary, it seems that the reality itself we are interested in is *constituted* by the entire set of the very elements we try to fit in the best possible way. Such an interpretation of the proposed approach to the various dimensions of moral thinking and acting at all levels seems to be supported by Rawls' defense of a constructivist account of morality and by his reluctance to speak of moral truth. Actually, there is no serious objection to the method, insofar as one may be happy to embrace such a version of antirealist constructivism.

However, it should be pointed out that the conception of reflective equilibrium as such does not necessarily imply antirealism. The state of coherence attained may be regarded as

evidence of truth. In fact, some of the philosophers who have recently endorsed a rather strong, naturalistic version of moral realism do not hesitate to appropriate the method of reflective equilibrium (Boyd, 1988; Brink, 1989). Their work shows clearly that a coherence theory of justification is not always accompanied by a commitment to a coherence theory of truth. Moreover, some philosophers have argued that even the latter is compatible with some form of realism (Dancy, 1985).

A parallel and complementary worry that is usually expressed is based on the fear of the *relativism* that may have to be accepted if we recognize the possibility of the existence of many sets of judgments, principles, theories, and arguments in states of equilibrium, which may all be equally acceptable. In other words, we are confronted by the objection of the possibility of many equally coherent sets of beliefs, quite familiar from debates in the theory of justification. Now, the idea of a point of maximal coherence, or of a perfect mutual adjustment which we may never be able to reach, apparently endorsed by Rawls in his mature work, seems to imply the transcendance of all actual equilibria reflecting the development of subjective or communal viewpoints which most probably require further elaboration. Even if we can never reach it, as Rawls puts it, 'we may get closer to it in the sense that through discussion, our ideals, principles and judgments seem more reasonable to us and we regard them as better founded than they were before'(Rawls, 1996: p. 385). Of course, we strive for consensus, we hope for a convergence of views toward the full and general, wide equilibrium we seek, and we cannot be certain that this can be realized, even if we pursue our investigations indefinitely. Here, all we can say to skeptics who do not seem to acknowledge any signs of such a convergence concerning moral and political matters, is that the pursuit of reflective equilibrium, which we propose to adopt as the most reliable nondogmatic procedure, does not guarantee that our expectations would not be frustrated.

Particular Objections to Reflective Equilibrium

Apart from the above general challenges to the method of reflective equilibrium as a coherentist approach, there are complaints about the particular concepts appealed to in the procedure advocated and followed by Rawls. There are two basic lines of criticism focusing on the notion of considered moral judgments, which seem to point in opposite directions:

On the one hand, it has been claimed that the use of considered judgments hides an implicit dogmatic commitment to intuitive starting points that we wish to preserve and eventually justify by the proper selection of principles. Thus, it is thought that there is a *disguised intuitionism* of subjective character covered behind the requirement to draw on beliefs that we claim to be testing. This accusation has been put forth by philosophers, such as utilitarians, who prefer different, straightforward ways to isolate and defend their basic principles, from which they shall claim to derive rational conclusions, however counterintuitive they may be (Hare, 1973; Singer, 1974, 2005). In fact, we have already referred to the suspicion of conservatism which Rawls was trying to dispel, explaining that the considered judgments in question are not supposed to be treated as secure and are not immune to revision any more than the principles they test and are tested against. Critics here fail to understand the importance of the perspective of wide reflective equilibrium and of its radical character.

On the other hand, we also encounter the complaint that the considered judgments are not secure enough to be trusted in the pursuit of our investigations. Their *degree of credibility* is not sufficient to provide us with a solid basis that will help us to proceed to the search of principles which will account for them and will presumably be accepted after the necessary work of mutual adjustment (Brandt, 1998). Here, we may once more detect foundationalist aspirations involving the requirement of an appeal to 'facts' and 'data.' However, as Daniels correctly points out, considered moral judgments are unfavorably contrasted to 'observation reports' which possess higher initial credibility, and it is not realized that they are more like 'theoretical statements' than observation reports. The parallel between ethics and science is pushed too far and we miss the crucial difference in the degree of theory dependence of the judgments and beliefs we begin with respectively in the two cases (Daniels, 1979: pp. 269–272).

One of the most interesting criticisms of the method is arguably what can be considered to be one of its advantages, namely its flexibility and its extension making it possible, at the end of the day, to accommodate, to integrate and harmonize all kinds of judgments, principles, theories, and arguments in a vast holistic perspective. Even certain self-evident principles inherited from the intuitionist tradition could find a place in the broad set of disparate elements at equilibrium, provided they are introduced without the support of any claim of independent warrant, presumably established before engaging in the balancing and matching procedures that are imposed by the directives of the proposed method. Moreover, although they must be denied the infallibility of necessary truths, we may even be willing to attribute to them some form of relativized necessity, if they have 'earned' it at the state of equilibrium, after having been subjected to testing against considered judgments and other moral and relevant nonmoral principles. In other words, the generality and the vagueness of the concept of equilibrium, the excessive openness, the breadth and the lack of determinacy of directives for the deliberative procedure of reflective 'adjusting,' 'matching,' or 'fitting' – problematic features, analogous to those of the idea of coherence – give rise to worries concerning its correct application. If anything could eventually be adjusted – after appropriate radical revisions – to anything else, within a vast framework of considerations at various theoretical and practical levels, one fears that the very ideal of reflective equilibrium might turn out to be *vacuous* and its usefulness might prove mostly *negative* in blocking the employment of unjustified foundationalist methodological options.

A final criticism which goes along with a recognition of the substantial virtues of a method of radical and wide reflective equilibrium and its advantages over available, traditional alternatives is elaborated by Michael DePaul in his *Balance and Refinement: Beyond Coherence Methods in Ethics* (DePaul, 1993). In fact, although DePaul's approach is not a simple correction or extension of the amended Rawlsian model in its wide conception, it can in a sense be regarded as a descendant. DePaul argues that the method is very *intellectualistic* and

mechanical. It cannot account for an experiential dimension of crucial moral significance that is not captured by the adjustment of judgments, principles, theories and their inferential or evidential links. Their equilibrium does not secure they are always warranted. He proposes to replace it by an analogous conception of justification of beliefs and further guidance of our thought, which partly replaces, rather than supplements, the reasoning procedures of the method of reflective equilibrium by a more general balancing of intellectual and affective factors. He thus stresses the need, not only for a careful rational balancing of elements, but also for the continuous refinement of our sensibility through the enrichment of moral experience afforded by various sources which include fiction and works of art (DePaul, 1993).

Further Applications and Prospects

As we saw, it might be argued that given the generality and the vagueness of the concept of a reflective equilibrium, as well as its unattainable character, it is preferable to speak of a regulative ideal of a method, rather than of a properly specified method with clear directives. To be sure, as we have known since the time of Aristotle, it would be foolish to seek scientific or mathematical determinacy and precision in the ethical domain.

In any case, despite the fact that the worries we have sketched in our survey of objections and criticisms do persist and the responses provided may not be fully satisfactory, the method or methodological ideal is well entrenched, if not dominant, in contemporary moral and political philosophy. It is naturally employed in domains of ethical inquiry which require the combination of different levels and fields, but also call for some form of direct application in practice. We have already noted its popularity in metaethics, more particularly among naturalist defenders of moral realism. We may also acknowledge its importance for the adjudication of methodological debates in applied ethics and especially in bioethics, where it allows us to avoid both deductions from abstract principles and appeal to casuistic moves or extreme particularism. We can thus seek a complex synthesis of 'top-down' and 'bottom-up' approaches, following a wide conception of reflective equilibrium of a variety of considerations at all levels (Daniels, 1996b; Van de Burg and Van Willigenburg, 1998). The lesson we shall draw from recent applications of the method is that answers and solutions at all levels of generality, from the most general metaethical theory to the concrete and immediate judgments supporting practical decisions, are best justified if we take into consideration, directly or indirectly, holistic 'packages.'

Moreover, reflective equilibrium could be regarded the most plausible general coherentist approach to all areas of philosophical thinking, perhaps by default; provided, of course, one is not a skeptic and is also willing to jettison rationalist accounts or scientistic and reductionistic conceptions, invoking either a priori or empirical secure foundations (Cummins, 1998; DePaul, 1998). In any case, those who are eager to challenge it must do better than simply reaffirm a dogmatic commitment to foundationalism of one or another variety.

See also: Law and Morality: An Analytical Perspective; Practical Reasoning: Philosophical Aspects; Public Reason; Rawls, John (1921–2002); Social Work Ethics.

Bibliography

Boyd, R., 1988. How to be a moral realist. In: Sayre-McCord (Ed.), Essays on Moral Realism. Cornell University Press, Ithaka and London.
Brandt, R., 1998. A Theory of the Good and the Right. Prometheus Books, New York.
Brink, D., 1989. Moral Realism and the Foundations of Ethics. Cambridge University Press, Cambridge.
Cummins, R., 1998. Reflection on reflective equilibrium. In: DePaul, M., Ramsey, W. (Eds.), Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry. Rowman and Littlefield, Lanham, pp. 113–127.
Dancy, Jonathan, 1985. Introduction to Contemporary Epistemology. Oxford, Blackwell.
Daniels, N., 1979. Wide reflective equilibrium and theory acceptance in ethics. Journal of Philosophy 76, 256–282 [Reprinted in Daniels, 1996].
Daniels, N., 1996a. Justice and Justification: Reflective Equilibrium in Theory and Practice. Cambridge University Press, Cambridge and New York.
Daniels, N., 1996b. Wide reflective equilibrium in theory and practice. In: Daniels, N. (Ed.), Justice and Justification: Reflective Equilibrium in Theory and Practice. Cambridge University Press, Cambridge and New York, pp. 332–352.
DePaul, M.R., 1993. Balance and Refinement. Beyond Coherence Methods of Moral Inquiry. Routledge, New York.
DePaul, M.R., 1998. Why bother with reflective equilibrium? In: DePaul, M., Ramsey, W. (Eds.), Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry. Rowman and Littlefield, Lanham, pp. 293–310.
Goodman, N., 1979. Fact, Fiction and Forecast, third ed. Hackett, Indianapolis and Cambridge.
Hare, R.M., 1973. Rawls' theory of justice. Philosophical Quarterly 23, 144–155, 241–251. [Reprinted in Daniels, N. (ed.) Reading Rawls, Basic Books, New York, 1975].
Rawls, J., 1951. Outline of a decision procedure for ethics. Philosophical Review 60, 177–197 [Reprinted in Rawls, 1999b].
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Rawls, J., 1974. The independence of moral theory. Proceedings and Addresses of the American Philosophical Association 47, 5–22 [Reprinted in Rawls, 1999b].
Rawls, J., 1996. Political Liberalism. Columbia University Press, New York.
Rawls, J., 1999a. A Theory of Justice, Revised ed. Oxford University Press, Oxford.
Rawls, J., 1999b. In: Freeman, S. (Ed.), Collected Papers. Harvard University Press, Cambridge MA.
Rawls, J., 2001. In: Kelly, E. (Ed.), Justice as Fairness: a Restatement. Harvard University Press, Cambridge MA.
Scanlon, T., 2003. Rawls on justification. In: Freeman, S. (Ed.), The Cambridge Companion to Rawls. Cambridge University Press, Cambridge.
Singer, P., 1974. Sidgwick on reflective equilibrium. The Monist 58, 490–517.
Singer, P., 2005. Ethics and intuition. The Journal of Ethics 9, 331–352.
Van de Burg, W., Van Willigenburg, T. (Eds.), 1998. Reflective Equilibrium: Essays in Honour of Robert Heeger. Kluwer Academic Publishers, Dordrecht.

Reflexivity

Nick Gould, University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reflexivity and reflexive practice are terms that have become pervasive in social work literature in recent years, but that remain contested and ill-defined, often being conflated with reflection and reflective practice. It is argued here that while the reflexivity literature demonstrates a generic concern with the critical analysis of practice in its social context, a finer grained distinction can be made between sociologically informed writing about 'practicing reflexivity,' and a more cognitive and process-oriented literature about 'reflexive practice.' Together, these approaches to practice offer insights that subvert the conventional understandings of expert practice and illuminate the situated nature of social work, but as yet they fall short of a unified or empirically tested theory of practice.

Introduction

> There's a lot of people talking, there's mighty few that know…

These words were sung by the blues artist Sonny Boy Williamson and the same might be said about writers about the concept of reflexivity and its relation to social work. Arguably, social work has a tradition of borrowing concepts from sociology and social theory and applying them to theories of practice, usually somewhat later than they were current in social theory and with some imprecision in their usage (Stenson and Gould, 1986). A review of the use of the concept of reflexivity in social work writing (D'Cruz et al., 2007) found a general lack of clarity in the use of the concept and, in particular, a lack of differentiation from the concepts of 'reflection' and 'critical reflection.' Other social work writers seem either to evade the issue of definition of reflexivity or perhaps assume that readers have an implicit understanding of what is meant; a recent chapter titled reflexive practice does not only omit a definition of reflexivity, but the term also does not occur in the text of the chapter (Doherty and White, 2012). Another definitional maneuver is to argue that in social work reflective practice is associated with professional practice, whereas reflexivity is associated with research methods (Wilson et al., 2008: p. 12). This is tidy but seems difficult to sustain, not least because some writers about reflexivity emphasize the overlap between practice and research methods, particularly those derived from feminist and critical qualitative research (Shaw and Gould, 2001).

An expectation that everyone knows what is meant by reflexivity is particularly problematic as an examination of the social sciences reveals a more diverse and contested usage of the term than the social work literature seems to assume. A few examples illustrate the point. Reflexivity is a widely used concept in economics and is associated with the ideas of economist George Soros (1987), who challenges orthodox theories of market equilibrium, arguing that markets follow self-reinforcing cycles between prices and expectations leading to boom and bust, i.e., disequilibrium. From this reading, reflexivity is not necessarily a benign or progressive force. However, the implication that outcomes cannot be predicted from aggregated macroeconomic data because of the unpredictability of human reactions to events does echo some of the social work critiques of positivism and evidence-based practice (Gray et al., 2009). In anthropology, reflexivity has come to have two distinct meanings, one that refers to the researcher's awareness of and analytic focus on their relationship to the field of study, and the other that attends to the ways that cultural practices involve consciousness and commentary on themselves (Clifford, 1997; Geertz, 1987). These anthropological interpretations of reflexivity may have more apparent relevance to social work, although they do not imply a unified field of discourse.

This article will review the ways reflexivity has been used in the social work literature and attempt a provisional consideration of whether any general synthetic lessons can be drawn. In their review of the literature, D'Cruz et al. (2007) argue that they can identify three 'variations' (their term) in the usage of reflexivity in the social work literature: the first focuses on ideas of self-actualization, the second emphasizes critical awareness of the self in action, and the third majors on the role of the self in the creation of knowledge. This summary does not do justice to the depth of their argument but, it is not clear that their typology is analytically sustainable. There seems to be considerable overlap in these categories around the repeated emphasis on the centrality of self. This seems somewhat ironic given that a central argument of at least some protagonists of reflexive practice is the 'decentering' of self (e.g., Taylor and White, 2000: pp. 43–44). A German social work scholar, Kessl (2009), constructs an alternative typology. He argues that debates within social work about reflexivity are located within analyses that have been emerging since the 1970s around the transformation of the welfare state, and that these contributions reflect contrasting methodological (and inherently epistemological) standpoints. He identifies these standpoints as deriving from critical theory, radical constructivism, and ethnomethodology, although he concedes that the terrain of the debate is complex and not easily amenable to simple categorization. This article will take a different approach, accepting the conclusion of these reviewers that reflexivity remains a contested and vaguely defined term, but that a basic and pragmatic distinction can

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28075-6

be made between those who can be described as 'practicing reflexivity,' which primarily operationalizes the implications of practicing within a sociologically informed understanding of modern reflexivity, and 'reflexive practice,' which situates social work practice within a more psychological, cognitive, and process-oriented approach to expert practice. Again, the boundaries of this distinction may be permeable, but it offers a broadly robust method for categorizing the literature. But first, before one considers what 'practicing reflexivity' and 'reflexive practice' are, what are they not?

Reflexivity and Reflexive Practice: What Are They Not?

The argument that social work requires the use of self within a reflexive awareness of the sociological and structural context of practice has an important provenance, and it would be an error to see it solely as a postmodern phenomenon. Even within the classic 'social casework' tradition, although the specific terminology of reflexivity had not arrived, it is clear that the psychosocial approach dealt with the 'person-in-situation,' a gestalt that indicated individuals' (including the social worker's) transactions with the social systems of which they were a part (Hollis, 1970). Radical social work of various hues, in part a reaction to the psychological individualism of traditional casework, took on and developed the Marxist concept of 'praxis,' the idea of practice being constituted both by action and reflection (Leonard, 1984). Although strongly indebted to the class analysis of Marxism, radical social work also drew on critical theory and aspects of psychoanalysis to theorize social work as engaged in raising, "critical consciousness of the nature of the services and how they define both the recipient of the service and the state worker within it" (Leonard, 1984: p. 215). Praxis was also a defining element of approaches to social work influenced by Paolo Freire (1970), which theorized 'conscientization' as a collective process of consciousness raising. Similarly, the Women's Movement and those who developed feminist standpoints in social work stressed the primacy of connecting the personal with the political to raise consciousness, "the process of transforming the hidden, individual fears of people into a shared awareness of the meaning of them as social problems" (Mitchell, 1975: p. 61). Feminist social work also connected with the emergence of antiracist and antidiscriminatory perspectives that acknowledged the multidimensional nature of oppression and the need for practitioners to acknowledge their own position within relationships of power (Dominelli, 1988, 2002).

More recently, the promotion of reflexivity as a standpoint in social work has been a way of steering a route between the Scylla of reflective practice and the Charybdis of evidence-based practice (Sheppard et al., 2000; Evans and Hardy, 2010). Although there is also a tradition using the terminology of reflection associated with psychodynamic perspectives (Mattinson, 1975; Yelloly and Henkel, 1995), reflective practice as a concept is generally taken to have been imported into social work from the research and writing of Donald Schon (Gould, 1989; Papell and Skolnik, 1992). Schon (1983) offered a bottom-up critique of the technical rational model of professional activity that characterizes practice as deductively drawn from formal, scientific knowledge. Instead, practice is represented as a form of artistry where practitioners iterate their own theory of practice through direct engagement with problems and the conscious deployment of reflection-in-action. The formal canon of knowledge may inform and contextualize practice but theory is understood as inductive and idiographic. In social work, the emergence of reflective practice seemed to gain traction in the 1990s not least from what has been described as the crisis of the professions: growing public skepticism about the infallibility of professionals, a reduction in automatic deference to experts, and sustained criticism of the adequacy of qualifying educational programs to prepare professionals for practice (Gould, 1996).

Aspects of this critique share some common ground with the theories of the risk society and late modernity (e.g., Webb, 2006), which the author will consider in relation to reflexivity, although the intellectual provenance is distinct. Some social work writers have subsequently taken reflective and reflexive practice to be interchangeable terms, and particularly so in regard to 'critical reflection,' a variant that draws heavily from critical theory and is particularly associated with the writing of Fook (e.g., Fook and Askeland, 2006). However, protagonists of reflexive practice have identified shortcomings and problems with reflective practice for which they offer reflexivity as a corrective. Evans and Hardy (2010) describe the reflexive critique of reflection as being the view that the latter is essentially concerned with individualistic psychological processes and overlooks the social, political, and ethical dimensions of social work:

> Thus, the notion of reflection has expanded into the wider notion of reflexivity, the key distinction being between the former's focus on internal self-awareness and the latter's emphasis on ensuring that the role that external, contextual factors play in informing individual subjectivity is specified. (Evans and Hardy, 2010: p. 124)

Thus, reflexivity expands the parameters of reflection. However, reflexive practice positions itself not only in relation to its refutation of aspects of reflection, but also in relation to its counterfactual, evidence-based, or empirical practice (Taylor and White, 2000). Again, there is a rejection of linear procedural and technological solutions to describing what constitutes 'best practice' and, as with the reflective critique of technical rationality there is a concern to identify the qualitative aspects of practice. The character of typical presenting problems confronted by social workers is that they are ambiguous and uncertain (Gould, 1996; Webb, 2006). Consequently, given the socially constructed nature of the 'problem,' formal algorithms derived from experimental research cannot be applied in the prescriptive fashion that is espoused in evidence-based medicine. Protagonists of reflexivity are vocal in their espoused rejection of what is usually described by them as positivism, and they are also usually careful to qualify this rejection

with a confirmation that formal knowledge remains relevant to practice:

> If we are arguing that the knowledge produced by the technical/rational approach and advocated by the EBP movement does not consist of neutral and detached observations of the real world, does this mean that it is worthless and should be disregarded? Most certainly not. (Taylor and White, 2000: p. 189)

The reflexive practice literature refers heavily to the sociology of scientific literature (e.g., Collins, 1985; Gilbert and Mulkay, 1984; Latour, 1999), which shows that the practice of science, including laboratory-based science, involves more than following stepwise scientific procedures, and is characterized by the insertion of judgment, improvisation, and communication through social networks. Reflexive approaches, whatever their differences, share a concern that empirical practice again underestimates the variability and uncertainty of the social environment of social work practice.

Reflexivity and Research

There is a long-standing, and at times fierce, debate in the social work literature as to whether doing direct practice is analogous to the process of doing research (Gilgun, 1994; Padgett, 1998), with specific reference to some of the similarities with qualitative research methods (Sheppard et al., 2000). For example, some writers on social work research position debates about reflexivity within wider postmodernist debates about fragmentation of social structures and the emergence of the 'risk society.' Reflexive modernity involves the management of risky situations under conditions of such complexity that outcomes cannot be predicted. The literature on reflexive practice, although not conflating research and social work intervention as forms of practice, does make explicit reference to the legacy of certain research traditions, particularly feminist and qualitative research. The latter makes little use of standardized instrumentation, the researcher is the main instrument of doing the research, and thus, the subjectivities of the researcher, those being researched, and their interaction are strongly foregrounded. Some sociologists have expressed their concern at the self-indulgence this can generate, as Becker (1993: p. 226) famously commented, "some researchers specialize in philosophical and methodological worry as a profession," and one shall see that some critics of the reflexive turn in social work echo this cautionary note.

Practicing Reflexivity

At the heart of much of the writing about reflexivity in social work is a project to build on the writings about 'reflexive modernity' developed by Beck (1992) with subsequent contributions of other sociologists such as Giddens (1990) and Lash (1994). A key contributor to social work writing about reflexive modernity has been Ferguson (2004) who, drawing specifically on the theoretical contributions of Beck and Giddens, argues that individuals now have unprecedented opportunities to shape their own lives. Reflexivity is characterized as a project of the 'self,' made possible by the fragmentation of the grand narratives that derived from modernity, which enables people under appropriate conditions to reflect on their actions and shape their identities. Ferguson suggests that these opportunities are potentially available to the clients of social workers to define their own needs rather than be subjugated by definitions imposed by experts, and thus are able to influence services to make them more appropriate and responsive to their expressed needs. This analysis is most extensively developed in relation to child protection services (as is most of the reflexive practice literature), in relation to which Ferguson suggests that women and children might be able to use services proactively and creatively to find protection from abusive relationships and avoid the damage and trauma generated by abuse. Ferguson redefines clients as "creative reflexive citizens" and he recasts the responsibilities of social workers and others who are part of child protection services to resist received constructions of the client that emphasize pathologies and deficits, in favor of engagement with strengths.

Some critics (e.g., Scourfield and Welsh, 2003; D'Cruz et al., 2007) criticize Ferguson for holding an almost unsociological view of the individual, presenting a view of social work clients as autonomous agents and downplaying their position vis-à-vis structural relations of power and class inequalities. Critics of Ferguson sometimes cite Kondrat (1999) who had already critiqued reflexive self-awareness as an orientation to practice as it emphasizes individual psychological processes at the expense of wider social relationships. Kondrat drew attention to another aspect of Gidden's (1990) writing, structuration theory, sometimes referred to as the 'double hermeneutic,' in which the goal of critical reflexivity as a form of practice aims for social workers to operate with greater awareness of structural relations and their opportunities to influence them. Ferguson might reasonably feel these criticisms to be overstated or even unfair; his most extended discussion of reflexivity and reflexive modernity and child protection (2004) comments extensively and repeatedly on the influence of class and gender on the nature of child protection practice. His analysis suggests that the challenge for social work is to go beyond compliance with forms of social control that are conventionally scripted by child protection systems, to work alongside clients in what he refers to as 'life planning,' supporting people to overcome the roles and destinies conventionally assigned to them. Reflexivity is thus a strategy that can be imparted to people who receive services to help them achieve emancipation and self-actualization, an ambition that has echoes with the work of radical adult educators such as Friere (1970) on conscientization (see above) and Mezirow (2009) on transformative learning.

Ferguson's writing on reflexive modernity, as already indicated, is primarily located in theoretical contributions to the understanding of the risk society drawn from Beck, Giddens, Lash, and others. This contrasts but overlaps with developments in the theorization of health and social care practice that more particularly focuses on the contributions of Foucault,

poststructuralism, and discourse theory (e.g., White, 1997; Taylor and White, 2000). As would be anticipated from Foucault's preoccupations with knowledge and power, processes of knowledge creation, and disciplinary practices, the work of White and others emphasizes the importance for social work of processes of critical analysis and deconstruction of the scenarios and interactions in which practitioners participate, to reveal terrains of power and influence. This contrasts with Schon (1983) and reflective practice where it is suggested that reflection takes place when unanticipated and nonroutine problems are encountered, so the objective of reflexive practice is to destabilize taken for granted assumptions that accompany and are embedded in routine practice and language. Methodologically, this standpoint draws extensively from ethnography and autoethnography as processes of social analysis. Indebtedness is also demonstrated to feminist perspectives and conversation analysis. The lessons of earlier ethnographies of social work practice are also discernible, for example, Dingwall et al.'s (1983) study of child protection practice, which drew attention to the inherently ambiguous and contested nature of 'evidence' in relation to the detection of child abuse.

Reflexive practice from within this perspective of the examination of making meaning, goes beyond the sense of reflection as disparagingly described by Woolgar (1988) as 'benign introspection.' A point of departure in the approach is that in many of the situations encountered by social workers and those in allied professions, there will be conflicting and disputed accounts of events: a baby is admitted to Accident and Emergency with extensive scalds, is this because the thermostat on the domestic water failed and the parent had not realized that the bath water was scalding, or had the parent deliberately punished the baby by immersing the baby in hot water, as suspected by the health visitor? As Dingwall et al. (1983) had found, physical injuries unless directly witnessed invariably require interpretation and judgment, and do not speak for themselves. In asserting their own versions and claims to truth in relation to these events, actors will draw on a repertoire of rhetorical devices to press their own view. For Taylor and White (2000), this is inherent in the nature of practice, and rather than proceed on the basis of the illusory certainties of evidence-based practice, they enjoin practitioners to cultivate a critical analysis of the forms of making sense they encounter. Reflexive practice as elaborated from this standpoint does not lead to the formulation of particular or distinctive methods of intervention but encourages practitioners to cultivate a skeptical interrogation of the ends that are implied by an intervention and the forms of power that are reproduced. For example, does a particular variant of a parent-training program mobilize a normative and gendered view of the role of a parent, and does it undermine the self-efficacy of a parent who undergoes it? And why?

Reflexive Practice–Heuristic and Process

In contrast to these sociologically oriented versions of reflexivity, some writers have treated reflexivity as a skill or capability, which, to a degree, is a disposition possessed by everyone, but which can also be transmitted and refined into higher levels of competence. Collins and Evans (2007), for example, identify reflexive ability as a skill that is part of a generic skill set that is not dependent on an individual's specialist subject knowledge, but is a precondition to developing expertise. Collins and Evans' field is the sociology of scientific knowledge, but the view of reflexivity as a heuristic (an experientially based strategy for problem solving) emerges in some areas of social work writing about epistemology or processes of knowledge creation. For example, Sheppard has argued in a series of theoretical and empirical contributions that reflexivity includes a set of cognitive skills that can be developed. Again, he draws analogies with the teaching of methodology to research students to equip them to conduct rigorous research so the development of concepts and classifications of reflexivity could enable this to be taught as a discrete knowledge area for social work practice. This process knowledge would inform the conduct and recognition of good practice (Sheppard et al., 2000: p. 469).

For Sheppard, the concept of reflexivity in social work emphasizes two essential conditions, that the social worker is an autonomous thinker who is able to assess, respond, and initiate action, and is aware of the socially situated nature of their relationship to the service user (Sheppard, 1998). His research suggested that practitioners demonstrate differential or variable levels of capability in relation to reflexivity but that an 'ideal type' of reflexive practice would contain a series of cognitive processes, starting from scrutiny of information at an initial referral stage, through critical appraisal of that information to generate hypotheses or questions relating to alternative formulations and ultimately strategies for intervention. Like Ferguson (2004) or White (1997), he stresses the ambiguity of much preliminary information and the need for a process of iterative critical appraisal as a situation evolves and new information is revealed.

Models of reflexive practice that build on theoretical insights to provide guidelines for practice intervention are relatively scarce and undeveloped. One model of reflexive practice has been developed by New Zealand and Australian academics and practitioners specifically geared again to the context of child protection practice (Connolly et al., 2006). They argue that child protection takes place at the intersection of several cultures: cultures of professional identity, family, ethnicity, and so on. Thinking about cultures, they argue, is particularly susceptible to the effects of stereotyping, with negative consequences for children and families. They propose a culturally reflexive model (CRM) to challenge reductionist thinking and to make practice more culturally responsive. Unpacking CRM, it is evident that reflexivity in this context is directly theorized from the approach of Fook and others toward 'critical reflection,' which, as was seen is sometimes taken to be synonymous with reflexive practice. Connolly et al. (2006) describe a three-phase model of practice, which begins with identifying personal, family, and societal cultural attitudes; proceeds through critical reflection of those attitudes to uncover and challenge the assumptions and aspects of power they

embody; and, third, putting into practice new approaches ('reflective outcomes') that are more productive and effective in protecting children. Connolly and her colleagues assert the importance of developing theories such as theirs as resources that social workers can draw on to improve practice.

Payne (2009) suggests that practice can be conceptualized as a 'reflexive cycle' that connects a series of reflexive activities each of which involves a level of engagement with the self: self-reference, self-analysis, self-interpretation, and self-criticism. The cycle begins with observation and appreciation of the context within which the practice takes place, analysis takes place of the interaction and communication between those involved in the situation, the worker supports others to understand how they communicate, and the impact of their interaction and changes take place in communication as a result of the understandings that have emerged. Payne illustrates these generic processes with an example of his own practice with a married couple coming to terms with the husband's serious illness. As Payne himself indicates, an inherent problem with suggesting a model of intervention that claims to be reflexive is that paradoxically it imposes a linear, reified process on activities, which, by definition, are supposed to be flexible and improvisatory (Payne, 2009: p. 169). Payne's example of the reflexive cycle also seems to be predominantly psychological, not dissimilar to models of reflective practice, and lacks the sociologically situated dimension that is so prominent in more theorized accounts of reflexive practice that engage explicitly with social structures and power relationships.

Discussion

It was seen that there is no clear consensus around the meaning of reflexive practice, beyond a claim that practice should be consciously located within an awareness of the impact of the self and/or the broader sociological context of practice. Can this be said to constitute a new paradigm of practice, a distinct break with the traditional concerns of social work with locating the person within their social environment, and cultivation of the use of self within practice? Is the idea of social work promoting in clients the status of 'reflexive active citizens' really very different from the approach to challenging deficit models of practice found in the radical social work literature discussed above and in 'strengths' models of practice (Saleebey, 2009)? Scourfield and Welsh (2003) have argued that the reflexivity thesis (particularly within the 'practicing reflexivity' variant) is simply the 'same old story' presenting well-established truisms about the nature of social work as a form of social control, but presented within new theoretical frameworks. Using the example of child protection practice, and drawing heavily on Scourfield's ethnographical research, they show that the class-based and gendered nature of social work practice is only too manifest and overt, without needing deconstruction through the theoretical prism of reflexivity. Inequalities of power and governance of the family by professional experts, and the sanctions experts can invoke, are well understood by the subjects of social work assessment and intervention; they do not need to have them revealed through processes of 'life-planning,' as advocated by Ferguson. Indeed, Scourfield and Welsh claim that Ferguson misconceives aspects of Beck's theory of the 'risk society,' namely, that Beck implies reflexivity produces 'more of the same,' whereas the term 'reflection' denotes critical appraisal and change, which is an inversion of the critiques of reflective practice proposed by proponents of reflexivity (Scourfield and Welsh, 2003: p. 403).

Conclusions

In this article, 'practicing reflexivity' has been contrasted with 'reflexive practice,' the latter encompassing the social work theorists who emphasize the various cognitive and affective components of micropractice. Here, the critics have contended that such approaches engender a sense of self-importance of the practitioner, and self-justification rather than self-critique. This is associated with a concern that an approach that emphasizes uncertainty and ambiguity in the analysis of social relations can lead to a situation of endless relativism, where nothing can be established or agreed with sufficient certainty to provide a basis for action. Not least, the protagonists for reflexive practice are challenged to develop a research agenda that validates, or otherwise, the effectiveness of its contribution, and can demonstrate that it makes a difference to outcomes for service users. A central plank of the reflexive thesis has been sustained skepticism of evaluation research and the 'what works?' agenda. This has resulted, in so far as reflexive practice has promoted an evidence base, in a preponderance of qualitative, process-oriented studies. The challenge for reflexive social work becomes one of reinterpreting the nature of outcome-based research to support more pluralistic and participatory definitions of what constitutes an outcome, and how and by whom it can be measured (Kessl, 2009).

'Reflexive practice' and 'practicing reflexivity' are works in progress, having entered the lexicon of social work relatively recently. These fields of scholarship are making various contributions to the development of social work that are becoming part of a broader consensus: not least they draw attention to the dangers of taking experts to be infallible and warn against an overemphasis on the attention to psychological aspects of practice without attention to social context. Also, reflexivity attempts to steer a course between the over-subjectivity of reflective practice and the depersonalized reductionism of evidence-based practice (Healy, 2005). Where should reflexive practice look in terms of its further development? Kessl (2009) suggests that the imperative is to develop an understanding of social work that reflects the realities of the postwelfare state that has been emerging since the 1970s. Social work emerged as a component of the welfare state in response to the relationships between the state, civil society, and the private citizen that was characteristic of capitalist societies in the late nineteenth century and twentieth century. Internationally, social work is having to redefine itself in relation to global economic forces that fragment traditional solidarities and introduce

the logic of the market to every aspect of social protection (Dominelli, 1999; Pugh and Gould, 2000). Within this political context, reflexive social work is at least a resource sensitizing social workers against taken-for-granted thinking and being co-opted into implementing social policies that commodify welfare and reinforce social exclusion and inequality.

See also: Child Neglect; Child Protection; Critical Social Work Practice; Fathers and Fathering: Social Work Perspectives; Social Work Theory.

Bibliography

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Becker, H., 1993. Theory: the necessary evil. In: Flinders, D., Mills, G. (Eds.), Theory and Concepts in Qualitative Research: Perspectives from the Field. Columbia University Press, New York.

Clifford, J., 1997. Routes: Travel and Translation in the Late Twentieth Century. Harvard University Press, Cambridge, MA.

Collins, H., 1985. Changing Order: Replication and Induction in Scientific Practice. University of Chicago Press, Chicago.

Collins, H., Evans, R., 2007. Rethinking Expertise. University of Chicago Press, Chicago.

Connolly, M., Crichton-Hill, Y., Ward, T., 2006. Culture and Child Protection: Reflexive Responses. Jessica Kingsley Publishers, London.

Dingwall, R., Eekelaar, J., Murray, T., 1983. The Protection of Children: State Intervention and Family Life. Basil Blackwell, Oxford.

D'Cruz, H., Gillingham, P., Melendez, S., 2007. Reflexivity, its meanings and relevance for social work: a critical review of the literature. British Journal of Social Work 37 (1), 73–90.

Doherty, P., White, S., 2012. Knowledge for reflexive practice. In: Gray, M., Midgley, J., Webb, S. (Eds.), The Sage Handbook of Social Work. Sage, Los Angeles, pp. 211–223.

Dominelli, L., 1988. Anti-racist Social Work. Macmillan, Basingstoke.

Dominelli, L., 1999. Neo-liberalism, social exclusion and welfare clients in a global economy. International Journal of Social Welfare 8 (1), 14–22.

Dominelli, L., 2002. Feminist Social Work Theory and Practice. Palgrave Macmillan, Basingstoke.

Evans, T., Hardy, M., 2010. Evidence and Knowledge for Practice. Polity Press, Cambridge.

Ferguson, H., 2004. Protecting Children in Time: Child Abuse, Child Protection and the Consequences of Modernity. Palgrave Macmillan.

Fook, J., Askeland, G., 2006. The 'critical' in critical reflection. In: Fook, J., White, S., Gardner, F. (Eds.), Critical Reflection in Health and Social Care. Open University Press, Maidenhead, pp. 40–53.

Freire, P., 1970. Pedagogy of the Oppressed. Continuum, New York.

Geertz, C., 1987. Works and Lives: The Anthropologist as Author. Stanford University Press, Stanford.

Giddens, A., 1990. The Consequences of Modernity. Polity Press, Oxford.

Gilbert, G., Mulkay, M., 1984. Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse. Cambridge University press, Cambridge.

Gilgun, J., 1994. Hand in glove: the grounded theory approach and social work practice research. In: Sherman, E., Reid, W. (Eds.), Qualitative Research in Social Work. Columbia University Press, New York.

Gould, N., 1989. Reflective learning for social work practice. Social Work Education 8 (2), 9–20.

Gould, N., 1996. Introduction: social work education and the 'crisis of the professions'. In: Gould, N., Taylor, I. (Eds.), Reflective Learning for Social Work. Arena, Aldershot, pp. 1–10.

Gray, M., Plath, D., Webb, S., 2009. Evidence-based Social Work: A Critical Stance. Routledge, London.

Healy, K., 2005. Social Work Theories in Context: Creating Frameworks for Practice. Palgrave Macmillan, Basingstoke.

Hollis, F., 1970. The psychosocial approach to the practice of casework. In: Roberts, R., Nee, R. (Eds.), Theories of Social Casework. University of Chicago Press, Chicago, pp. 33–76.

Kessl, F., 2009. Critical reflexivity, social work, and the emerging European post-welfare states. European Journal of Social Work 12, 305–317.

Kondrat, M.E., 1999. Who is the 'self' in self-aware? Professional self-awareness from a critical theory perspective. Social Service Review 73 (4), 451–477.

Lash, S., 1994. Reflexivity and its doubles: structure, aesthetics, modernity. In: Beck, U., Giddens, A., Lash, S. (Eds.), Reflexive Modernizaton: Politics, Tradition and Aesthetics in the Modern Social Order. Cambridge, Polity.

Latour, B., 1999. Pandora's hope: essays on the reality of science studies. Harvard University Press, London.

Leonard, P., 1984. Personality and Ideology: Towards a Materialist Understanding of the Individual. Macmillan Education Ltd, Basingstoke.

Mattinson, J., 1975. The Reflection Process in Social Work Supervision. Tavistock, London.

Mezirow, J., 2009. Transformative Learning in Practice: Insights from Community, Workplace and Education. Jossey-Bass Inc., San Francisco.

Mitchell, J., 1975. Psychoanalysis and Feminism. Penguin, Harmondsworth.

Padgett, D., 1998. Does the glove really fit? Qualitative research and clinical social work practice. Social Work 43 (4), 373–381.

Papell, C., Skolnik, L., 1992. The reflective practitioner: a contemporary paradigm for social work education. Journal of Social Work Education 28 (1), 18–26.

Payne, M., 2009. Understanding social work process. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work: Themes, Issues and Critical Debates, third ed. Palgrave Macmillan, Basingstoke, pp. 159–174.

Pugh, R., Gould, N., 2000. Globalization, social work and social welfare. European Journal of Social Work 3 (2), 123–138.

Saleebey, D., 2009. The Strengths Perspective in Social Work Practice, fifth ed. Pearson Education, Inc., Boston.

Schon, D., 1983. The Reflective Practitioner. Basic Books, New York.

Scourfield, J., Welsh, I., 2003. Risk, reflexivity and social control in child protection: new times or same old story? Critical Social Policy 23 (3), 398–420.

Shaw, I., Gould, N., 2001. Qualitative Research in Social Work. Sage Publications, London.

Sheppard, M., 1998. Practice validity, reflexivity and knowledge for social work. British Journal of Social Work 28 (5), 763–781.

Sheppard, M., Newstead, S., Di Caccavo, A., Ryan, K., 2000. Reflexivity and the development of process knowledge in social work. British Journal of Social Work 30 (4), 465–488.

Soros, G., 1987. The Alchemy of Finance: Reading the Mind of the Market. John Wiley, New York.

Stenson, K., Gould, N., 1986. A comment on "a framework for theory in social work" by Whittington and Holland. Issues in Social Work Education 6 (1), 41–46.

Taylor, C., White, S., 2000. Practising Reflexivity in Health and Welfare: Making Knowledge. Open University Press, Buckingham.

Webb, S., 2006. Social Work in a Risk Society: Social and Political Perspectives. Palgrave Macmillan, Bsingstoke.

White, S., 1997. Beyond retroduction? Hermeneutics, reflexivity and social work practice. British Journal of Social Work 27 (5), 739–753.

Wilson, K., Ruch, G., Lymbery, M., Cooper, A., 2008. Social Work: An Introduction to Contemporary Practice. Pearson Longman, Harlow.

Woolgar, S., 1988. New Frontiers in the Sociology of Knowledge. Sage, London.

Yelloly, M., Henkel, M. (Eds.), 1995. Learning and Teaching in Social Work: Towards Reflective Practice. Jessica Kingsley Publishers, London.

Reflexivity in Anthropology

George E Marcus, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reflexivity has been an essential dimension of cultural and social anthropology's emblematic method of inquiry: ethnographic fieldwork and writing. However, this core feature in the making of anthropological knowledge only became theorized and widely discussed in the 1980s and 1990s when an interdisciplinary intellectual movement with many facets and influences intensively assessed all aspects of the culture concept. Distinctive forms of reflexive thinking, analysis, and writing in anthropology have become canonical. This article reviews intellectual roots of this turn, its period of incorporation, and how it is manifested in contemporary anthropological research.

Essentially the capacity of a system of signification to turn back on itself, reflexivity, is the specific modality that an important critique of anthropological method and knowledge took in the 1980s. Reflexive inquiry has long been a potential in the social sciences. It experienced elaborate development in the United States as a result of the intersection of the influence of hermeneutical approaches and a sustained atmosphere of critique following the Vietnam War. Inherent aspects of ethnography were called into question by anthropologists themselves under the influence of trends of critical theory in literary studies and in interdisciplinary movements such as feminism and postmodernism, for which reflexivity was a key technique. The consequences for the process of inquiry in anthropology, represented by fieldwork and the genres of writing from it, have been far reaching and enduring. Reflexive practices are now mainstream in anthropology, and how they have worked their way into the changing conditions and subjects of contemporary ethnography is considered.

Critiques of the social sciences as practiced in the United States emerged in the 1980s in which reflexivity became a keyword, standing for both the means of the critiques and the form that alternative practices might take. The consequences for disciplinary practices of reflexivity as means of critique and alternative form of inquiry were nowhere as great among the social sciences as in social and cultural anthropology.

According to one of the early statements in this critical trend, reflexivity is "… the capacity of any system of signification to turn back upon itself, to make itself its own object by referring to itself: subject and object fuse" (Myerhoff and Ruby, 1982). Myerhoff and Ruby go on to stress that "significant distinctions exist between reflexiveness and related attitudes such as self-regard, self-absorption, solipsism, self-reference, self-consciousness, and autobiography. Reflexiveness does not leave the subject lost in its own concerns; it pulls one toward the other and away from isolated attentiveness toward oneself."

Reflexivity as a practice in the social sciences is, thus, both a distinctive mode of thinking and a strategy of writing that concerns itself with the process of inquiry as an integral part of the claims to knowledge that the inquiry might produce. As a concept, it is related to the idea of feedback in cybernetic models, of metacommunication, and of such functions as deixis and indexicality inherent in the faculty of language use. However, it is the more ideological uses of reflexivity in debates about the revision of anthropology's distinctive methods of fieldwork and ethnography that are of concern here. Debates about the role and purpose of reflexivity in the process of research as well as in the changing form of the ethnographic text became the major specifiable way of assessing the impact upon anthropology of the broader intellectual movement of the 1970s and 1980s known loosely as postmodernism.

In the broad sweep of European philosophy, reflexivity as a category of interest in social theory begins with Kant's distinction between analysis and an analytic in which an analytic is the kind of inquiry where the subject of inquiry is also the object of inquiry. One genealogical stream from this originary idea of reflexive inquiry leads through Hegel to phenomenology and eventually to Heidegger and existentialism in France. The other, and for us, more relevant genealogical stream leads to hermeneutics in the work of Hans-Georg Gadamer and Paul Ricoeur by way of the neo-Kantian writings of Dilthey and the work of Max Weber, in which reflexivity, while implicit, is a strong element of his methodological considerations. In the work of Gadamer and Ricoeur, reflexivity becomes an explicit and crucial intellectual operation in the activity of interpretation whereby the individual ethical consciousness expands reflectively through encounters with the foreign or the different. Thus, while long a potential existing in a sort of minor key to the predominant positivist tendency of social theory that oversaw the post–World War II expansion of sociology and anthropology in the United States, philosophical works of hermeneutics, in which the operation of reflexivity was highlighted, only developed explicitly a strong appeal among sociologists and anthropologists from the early 1980s onward.

Of all of the social sciences, the importance of reflexivity as a style of inquiry has been most natural to precisely those methods of anthropology that have distinguished it as a discipline. Pioneered by Bronislaw Malinowski and Franz Boas, the pursuit of fieldwork investigation through long-term participant observation, working directly in native languages, and the constant interpretation and translation of behaviors and speech encountered on a daily basis, inherently involves on the part of the investigator a keen consciousness of oneself and one's relations with specific others as the medium of all claims to knowledge from this research practice.

For a long time, anthropologists only came to terms with the complexities and problems of this distinctive form of knowledge production informally as a pastime of professional culture and folklore, as something to be communicated between mentors and students. It was only at the end of the 1960s, and then again, and more definitively in the 1980s, that anthropologists examined and developed explicitly the varieties of meanings and implications of reflexivity as a major style of knowledge production for the future practice of their discipline.

Since the 1980s, reflexivity has become a 'keyword' indicating both a self-conscious, mainstream style of work in anthropology and part of its own enduring disciplinary politics of knowledge. Here, it is important to note that reflexivity has only served as such a central category, standing in turn for the most concrete aspect of the controversial impact of postmodernism on anthropological practice, in the recent evolution of the discipline in the United States. It has been discussed and debated from the 1980s on in other anthropological establishments around the world, but mainly through the stimulus of the prestige and diffusion of trends in United States anthropology.

Reflexivity in Two Moments in the Critique of Anthropology

During the late 1960s, US anthropologists became preoccupied openly with a long simmering critique of their discipline. This critique had theoretical and methodological dimensions, but it was primarily political in nature, stimulated by the ferment of domestic resistance to the Vietnam War, renewed discussions of anthropology's historic development within the expansion of European colonialisms, and the interest in new work in Marxist theory. At that time as well, Geertz was developing a broad and deep theoretical foundation for cultural anthropology in hermeneutic philosophy, work that was eventually to serve as both the primary foil and influence for later discussions of reflexivity in anthropological method. (See Geertz, 1968, for his early position on reflexivity in fieldwork.) But in the *de facto* reflexive attempt to undertake a critical anthropology of anthropology in the politicized atmosphere of the 1960s, the issue of reflexivity itself as the major dynamic of what it was to practice critical analysis was little addressed. An important exception is the paper by Scholte (1969), which summarized this period of critique. Important as a figure who introduced currents of European social theory into US anthropology, Scholte foreshadowed what sort of theoretical work and changes in practice would be necessary to transform anthropology into a discipline with a sustained critical purpose beyond that moment of heightened political critique.

In the interval between the two periods of internal disciplinary critique, perhaps the most important work that served as the raw material on which the eventual surfacing of explicit considerations of reflexivity in anthropology depended was the burgeoning of a literature on fieldwork experience, written alongside the production of more formal ethnographies from which such reflections about the process of inquiry were largely absent. This literature, fueled by the publication of Malinowski's diaries in 1967, incorporated the spirit of self-critique of the late 1960s and was produced by anthropologists who were then students and young professors. There emerged a more theoretical strain of this literature in which the reflexive character of fieldwork became the major preoccupation (e.g., Rabinow, 1977) and that led eventually to experiments with the form of the ethnography itself (e.g., Crapanzano, 1980). These works, combined with an interdisciplinary movement, variously characterized as postmodernism and poststructuralism and led by scholars of literature who were experiencing their own disciplinary crisis, stimulated the theoretically deeper, more wide ranging, and sustained internal critique of anthropology of the 1980s.

Incorporating all the dimensions of the earlier critique, the 1980s critique primarily examined the central practices of the discipline, especially in its character as a hermeneutic enterprise, a view for which Geertz had laid the foundation. In particular, it did so by a focus on the rhetoric and forms of the signature textual product of anthropology, the ethnography. (See Clifford and Marcus, 1986; the volume that has stood for this critique in the way that the Hymes volume did for the earlier one.) This critique was, thus, the efflorescence of a comprehensive discussion of all the issues relating to reflexivity that had long been so essential to the practice of anthropology. Influenced by the circulation of much more diverse and sophisticated theoretical resources, anthropologists looked with an unprecedented intensity and thoroughness at the relationships in their work between reflexive modes of knowing, critical and ethical awareness, the political contexts of inquiry, and alternative possibilities for practice.

The basic act of reflexivity that initiated this critique were the probing discussions of the relationships of anthropologists to informants – of the self to other – and the dependence of the production of anthropological knowledge on these relationships, which had so often been effaced in the ways that ethnographies had in fact been written. Evoking alternative modes of representing others, questioning the act of representation itself, suggesting the ways ethnographic knowledge emerges from collaboration and dialogue, and probing the political conditions that have inhibited such possibilities, this moment of critique served both to stimulate degrees of experiment with the ethnographic form and to encourage ramifying discussions about the critical purpose of anthropological analysis (e.g., Marcus and Fischer, 1999). This undertaking was itself an umbrella act of reflexivity, and reflexive practice became the common frame for thinking about past ethnographies, future possibilities, and the form and purpose of anthropological knowledge itself. This field of discussion was highly contested: some felt that the critique by its nature was too hermetic, too literary, and not sufficiently committed to anthropology's study of social struggles in the contemporary world; some thought the attention to reflexive issues was an aesthetic distraction and a major blow to the production of objective knowledge that could be shared by a scholarly community and on which disciplinary progress depended; and contrastingly, for some, the kinds and levels of reflexivity encouraged by the critique were neither reflexive nor critical enough, leading to a spiraling array of discussions about insufficiencies and exclusions that had even more ideal reflexive practice in view.

The immediate effects of this period of critique were the changes in the ways that ethnographies were both written and discussed. But while the critique was developed through attention to texts and the way that anthropologists rhetorically established authority for them, the larger stake was the effort finally to justify them theoretically and ethically and expand anthropology as a critically reflexive field of inquiry. There can be little doubt that this much has been established as at least a long-term mainstream tendency in the discipline. This tendency has placed anthropology in much closer relation to such parallel interdisciplinary fields, themselves rooted in varieties of reflexive inquiry, as media studies, feminist studies, postcolonial studies, and most broadly, cultural studies, than to its historic partners in the other social sciences.

The Incorporation of Reflexivity in the Mainstream of Anthropological Research

Discussions about reflexivity, or at least the evaluation of ethnographic claims to knowledge in terms of the specific conditions and relations of inquiry that produced them from fieldwork, are not only highly visible and commonplace in anthropological discourse of the 1990s, they are *de rigueur*. Further, some form and degree of reflexive commentary is worked prominently into virtually every ethnography written these days. If not enacted in experimental text design, there is the presumption that all ethnography is produced collaboratively, and there are far richer and more probing discussions than ever before about the kinds of contexts that influence ethnographic representations of subjects and shape arguments from ethnographic data (see Marcus, 1994). Concern for reflexivity has had a special appeal for ethnographic filmmakers, and some of the most cogent writing on the importance of reflexivity in the representation of other forms of life have come from this arena (e.g., MacDougall, 1998). Current productions of ethnographic film and media demonstrate most clearly that the practice of reflexivity crucially redefines and adapts the classic conditions of ethnographic inquiry to new conditions of fieldwork where anthropological research projects must negotiate different identities and roles in relation to the highly articulated projects of their subjects in the overlapping contexts of a globally interconnected world.

Variant Ideologies in the Practice and Form of Reflexivity

In sociology, discussion of reflexivity has tended to be integrated formally as a methodological tool or technique (e.g., see Bourdieu, 1990), limiting its ability to affect fundamentally the forms and purposes of sociological research. In anthropology, the outcome has been quite different, generating a number of changes in the ways that ethnographies are written and fieldwork is conceived. This in turn has obviously had considerable implications for changes in both the form and content of the knowledge that anthropology's core research process generates.

In engagement with the terms of the 'writing culture' critique of the 1980s, the largest body of literature stimulated explicitly by reflexive inquiry examines in great detail the self–other relation in classic conditions and specific sites of fieldwork. While some have characterized this literature as narcissistic or merely self quests, in its most accomplished examples (e.g., Behar et al., 1995; Reed-Danahay, 1997; Taylor, 1999; Tsing, 1993), it has addressed cogently issues of contemporary changes in concepts of self, personhood, and identity among many of the peoples who had been the subject of study by previous generations of ethnographers. At the microlevel of fieldwork, it has also clarified aspects of the history of anthropology as an academic enterprise and in relation to the broad range of effects that inquiry itself has on the peoples and subjects studied (e.g., Rosaldo, 1989). While somewhat unruly and less governed by genre rules of form in social science writing, this literature has, in sum, been very useful in deepening and critiquing the complexities of constructs of subjectivity and otherness within their conventional usages in Western intellectual discourse – a long-term, continuing objective of anthropology as a discipline. It has also probed critically ethnographers' positions in relation to initial objects of study so as to transform the objects in novel ways and reconfigure the compass of research in response to radical contemporary changes in both the worlds of the academic and who is identified as subjects of study.

There has also been a strain of reflexive inquiry in anthropology for which reflexivity has been more distinctly a means by which to define an object of study or to circumscribe new temporal and spatial boundaries for a fieldwork project (Marcus, 1998). Key to these works is clarity of the positioning or location of ethnographic inquiry, which reflexivity enables, in relation to other historic and geographical agents in the same social field. Reflexivity as positioning not only serves the creation of new forms of objectivity in anthropology, sensitive to the issues of critique that discussions of reflexivity have raised, but redefines both the scope of ethnography and its arguments. In these works, the limits and reach of once very locally focused ethnographic arguments expand in unpredictable and intriguing ways (for a now classic experiment of this sort, see Taussig, 1987; for more recent examples, see Martin, 1994; Fortun, 2000).

Finally, there is the presence of reflexivity in ethnography because of the evolution of certain realities on the ground over the course of career-long scholarly involvements with certain peoples and places. In recent decades, literally self-conscious cultural awareness has increased markedly and even been mobilized politically in many locales and regions such that the conditions of ethnographic research require a rethinking and reorientation toward one's subjects, now organized into social movements, nongovernmental organizations, and other groups involved in objectifying and representing themselves. Here, reflexive inquiry is the means to a complex and necessary politics of research that not only shapes fieldwork but makes its continuation possible at all (see Warren, 1999; for an exemplar of this type of increasing variant of reflexive inquiry).

Reflexive Modernization and Reflexivity in the Changing Conditions of Anthropological Fieldwork

Broader changes in contemporary social conditions worldwide that stimulate the trend of reflexive inquiry in anthropology is encompassed in the concept of reflexive modernization (Beck et al., 1994), a development in social theory of the 1990s

that responds to the postulation of a present era of post-modernity by trying to specify the characteristics of such an era for the pursuit of empirical inquiry. This construct proposes that the modernity characteristic of industrial society is succeeded by a modernity that involves living with irreducible contingency, living in a more complex and less controllable world. Most importantly, reflexivity as a self-monitoring process on the level of both institutions and individual persons becomes not only a pervasive aspect of society, but also a key focus of social thought itself.

This version of social theory suggests for the proliferation of styles of reflexive inquiry in anthropology that these are not driven only as a result of the internal critique of anthropology, as a seminal countertrend to a preceding era of positivist tendencies, but as a more generally shared condition with the societies and cultures that anthropology makes its subjects of study. This suggestion in turn has a potentially profound and enduring effect on the research protocols of anthropology in recognizing certain affinities and identifications between the reflexive predicaments of ethnographers and those whom they study. Reflexivity is relevant not because it provides truer or more powerful explanatory models with which anthropologists can treat social facts among their subjects, but rather because reflexivity is also practiced differentially 'out there' on a widespread basis (see Marcus, 1999). This fact suggests the breaking of the frame of conventional scholarly work altogether, at least in the anthropological tradition, rather than the demise and replacement of specific concepts and characteristics of this academic frame. Thus, reflexivity is no longer only a moment in the critique of disciplines but becomes more generally the object and medium of social and cultural analysis as well.

Reflexivity into the 2000s

The legacies of the 'reflexive turn' begun in the 1980s, first with attention to the way ethnographies were produced and written, and then in the emergence of a vibrant theoretical apparatus about how to constitute subjects as the foci of contemporary anthropological research, are alive and well into the first and now second decade of the new century. They indeed have become canonical in the methods of ethnographic inquiry and in producing descriptive and analytic results from them. This is true, even in the burgeoning application of ethnographic methods to the study of virtual worlds (Boellstorff et al., 2012).

If there has been a different conceptual twist, or new thinking about reflexivity, it has accompanied a more general shift in anthropologists' interests toward ontological questions, more than epistemological ones. The latter evoke the essence of reflexive practices. These address questions of knowing, experience, meaning, and the relationships between the anthropologist as observing/interpreting agent and his or her subjects as the main focus of inquiry. Ontological questions are evoked and stimulated by the changing objects and environments or circumstances in which ethnographic fieldwork is situated anywhere today. They concern the analysis and description of the emergence of new forms, organizations, processes, and practices. Descriptive and analytic attention to materiality, techniques, invention, and new patterns of relations and transaction takes precedence over symbolic or hermeneutic analysis.

New technologies, global projects, environmental change, disasters, infrastructures, emphases on material conditions of life, and the new levels of awareness and entanglements that these changing conditions of planetary life cause everywhere, all require new thinking about how to situate anthropological research as fieldwork, and how to negotiate conditions of sustained inquiry. For anthropological research projects, this embedding of fieldwork in complex global projects requires unprecedented levels and degrees of collaboration and alliance with assemblages (Ong and Collier, 2005) of experts, activists, and stakeholders, and stimulate new adaptations of established tendencies in conducting fieldwork. Anthropologists of course remain interested in the life worlds of subjects with which they reflexively engage, but in order to even define subjects in this way today they have to conceive of working on, and within, new projects, structures, and organizations of unprecedented complexity and technical constitution, along with a politics of doing research that requires negotiations of traditional categories (such as culture, the human, kinship, or even ethnography itself), which have defined the specialized terms of disciplinary discussion within anthropology itself.

This increased curiosity about the ontology of new objects of study in the world, and being immersed as fieldworkers within them as assemblages of relations in systems of distributed knowledge, requires a kind of modified or morphed reflexive posture. In confronting this renewed trend of working within complex systems (of governance, of markets, of finance, of technoscience), anthropologists have not for the most part returned to the techniques of formal modeling characteristic of the other social sciences, but rather have continued to adapt theoretical innovations to the still largely qualitative nature of ethnographic research. The most influential conceptual paradigm emerging at the turn of the century, supporting this shift toward the ontological in anthropology has been actor–network theory, associated with the work of Bruno Latour (2005). While problems of reflexivity are discernable in this paradigm, they have not been explicitly addressed much by its major proponents.

How then to think about the shift from, or parallel to, reflexivity, when anthropologists become immersed as fieldworkers in complex assemblages of relations, tied to emergent changes and events, and thus return to an interest in describing systems, structures, and macroscale events? In the author's view, the most cogent expression of what the discussion of the concern with reflexivity becomes in anthropology's new entanglements, with an accented ontological interest, in complex systems and distributed knowledge assemblages is Kim Fortun's formulation (2000), in her study of the environmental justice movement following the Bhopal disaster. She tracks her conceptual shift from a stance of reflexivity to recursivity (2000: p. 23):

> Reflexivity calls for the ethnographer to position herself. Recursivity positions her within processes she affects without controlling, within competing calls for response. Reflexivity asks what constitutes the ethnographer as a speaking subject. Recursivity asks what interrupts her and demands a reply. Thinking in terms of recursivity is a way to hold ethnography responsible for advocacy. Attention to recursivity foregrounds how every articulation—whether ethnographic or in direct advocacy—operates on previous articulations, inserting every move and every word within multiple discourses and worlds…

One can see this variant of reflexivity, as recursivity, operating methodologically and analytically in contemporary anthropological research within its other major thematic: that is, to operate in, and speak to, public spheres, to be engaged, and even activist at the same time that it produces research that is empirical and achieves scholarly standards. The intimate, microscale settings of fieldwork by which reflexivity was intersubjectively theorized in the 1980s and 1990s now become a discussion about the ethical and normative commitments of anthropological knowledge, and its accountabilities as such, in relation to specific publics, institutions, and global projects as ethnographers move recursively in their circuits of inquiry, as evoked by Fortun. The original idea of methodological reflexivity in ethnographic inquiry thus now encourages anthropologists to shape and present their findings in relation to a larger community of reception and debate. Beyond a kind of rhetoric of desire and purpose, any fulfillment of this goal of engaged research would entail specific arguments and analyses from the recursions of fieldwork that link anthropologists with subjects through their movements within complex systems or assemblages. Recursivity is thus less a situating of fieldwork practices in relation to subjects, which reflexive practices classically have done, but provide a compass for movement through an unknown terrain of not just places, locales, and situated communities, but concepts, programs, discourses, values, and the communities or publics that produce them. This is what or more, or different, about what reflexivity has become in anthropological research moving into the twenty-first century.

See also: Cultural Critique, Anthropological; Ethnography in Applied Social Research; Ethnography; Fieldwork in Social and Cultural Anthropology; Hermeneutics, History of; Hermeneutics; Interpretation in Anthropology; Knowledge, Anthropology of; Postmodernism: Methodology.

Bibliography

Beck, U., Giddens, A., Lash, S., 1994. Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order. Stanford University Press, Stanford, CA.

Behar, R., Gordon, D.A. (Eds.), 1995. Women Writing Culture. University of California Press, Berkeley, CA.

Boellstorff, T., Nardi, B., Pearce, C., Taylor, T.L. (Eds.), 2012. Ethnography and Virtual Worlds: A Handbook of Method. Princeton University Press, Princeton, NJ.

Bourdieu, P., 1990. The Logic of Practice. Stanford University Press, Stanford, CA.

Clifford, J., Marcus, G.E., 1986. Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA.

Crapanzano, V., 1980. Tuhami: Portrait of a Moroccan. University of Chicago Press, Chicago, IL.

Fortun, K., 2000. Advocacy after Bhopal: Environmentalism, Disaster, New Global Orders. University of Chicago Press, Chicago, IL.

Geertz, C., 1968. Thinking as a moral act: ethical dimensions of anthropological fieldwork in the new states. Antioch Review 28, 139–158.

Hymes, D. (Ed.), 1972. Reinventing Anthropology, first ed. Pantheon Books, New York.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, Oxford.

MacDougall, D., 1998. Transcultural Cinema. Princeton University Press, Princeton, NJ.

Marcus, G.E., 1994. On ideologies of reflexivity in contemporary efforts to remake the human sciences. Poetics Today 15, 383–404.

Marcus, G.E., 1998. Ethnography through Thick and Thin. Princeton University Press, Princeton, NJ.

Marcus, G.E. (Ed.), 1999. Critical Anthropology Now: Unexpected Contexts, Shifting Constituencies, Changing Agendas. School of American Research Press, Santa Fé, NM.

Marcus, G.E., Fischer, M.J., 1999[1986]. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. University of Chicago Press, Chicago, IL.

Marcus, G.E., 2012. The Legacies of writing culture and the near future of the ethnographic form. Cultural Anthropology 27 (3), 427–445.

Martin, E., 1994. Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of AIDS. Beacon Press, Boston, MA.

Myerhoff, B., Ruby, J., 1982. Introduction. In: Ruby, J. (Ed.), A Crack in the Mirror: Reflexive Perspectives in Anthropology. University of Pennsylvania Press, Philadelphia, PA.

Ong, A., Collier, S. (Eds.), 2005. Global Assemblages. Blackwell, London.

Rabinow, P., 1977. Reflections on Fieldwork in Morocco. University of California Press, Berkeley, CA.

Reed-Danahay, D.E. (Ed.), 1997. Auto/Ethnography: Rewriting of the Self and the Social. Berg, Oxford, UK.

Rosaldo, R., 1989. Imperialist nostalgia. In: Culture & Truth: The Remaking of Social Analysis. Beacon Press, Boston, MA, pp. 68–87.

Ruby, J. (Ed.), 1982. A Crack in the Mirror: Reflexive Perspectives in Anthropology. University of Pennsylvania Press, Philadelphia, PA.

Scholte, B., 1969. Toward a reflexive and critical anthropology. In: Hymes, D. (Ed.), Reinventing Anthropology. Random House, New York, pp. 430–457.

Taussig, M., 1987. Shamanism, Colonialism and the Wild Man: A Study in Terror and Healing. University of Chicago Press, Chicago, IL.

Taylor, J., 1999. Paper Tangoes. Duke University Press, Durham, NC.

Tsing, A.L., 1993. In the Realm of the Diamond Queen. Princeton University Press, Princeton, NJ.

Warren, K.B., 1999. Indigenous Movements and Their Critics: Pan-Maya Activism in Guatemala. Princeton University Press, Princeton, NJ.

Reflexivity in Science and Technology Studies

Malcolm Ashmore, Loughborough University, Leicestershire, UK; and Universidad Nacional de Colombia, Bogotá, Colombia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article looks at the history, varieties, influences on, and criticisms of reflexivity in science and technology studies (STS). It deals with the epistemic, aesthetic, practical, and political aspects of reflexivity in this context. First, the meanings of 'reflexivity,' as this term has been understood in STS, are set out. A brief history of the varieties of reflexive concerns in this field is presented, from Bloor's explicit, yet limited, endorsement, to the full-blooded treatments of 'the reflexivists' (Mulkay, Woolgar, Ashmore) and beyond. Criticisms of reflexive writing from within STS (especially from Collins and Latour) are dealt with and responded to. A consideration of the internal role and status of reflexive work in STS and an assessment of the current state and future prospects of reflexivity are undertaken.

Introduction

"Considerations of reflexivity have salience in diverse intellectual practices and traditions: literature, philosophical logic, the visual arts, cybernetics, as well as social theory and sociology. Indeed, it is an issue with potential relevance for all modes of representational and performative practice. In the most general terms, considered as a feature or style of inquiry, reflexivity is a concept that expresses a concern to turn 'subject' into 'object,' to turn the activity of inquiry into a topic for (its own) inquiry. To 'be reflexive' is to bring the process of doing an activity into the purview of that activity as a feature of it. Thus, it implies a certain self-consciousness, and a certain paradoxical, even Zen-like, mode of operation: bootstrapping; attempting to see the flashlight that illuminates the scene as a part of the scene it illuminates; taking account of this moment as one writes of 'this moment.' This article details the ways in which reflexivity has featured in science and technology studies (STS), and speculates on the current and future status of reflexive study in STS.

Alternative Introduction

This heading is not to declare, exactly, that there is anything 'wrong' with that first introductory paragraph. It does the job, and it does it correctly; that is, in accord with the 'Instructions for Manuscript Preparation' for this encyclopedia. It is just that in doing 'just that,' it is unable to acknowledge the conditions under which it does so. Thus, it leaves all those things out of (its) account; which is normal, and normally unobjectionable. Perhaps no other article in this *encyclopedia* makes any reference to such things, preferring instead, perfectly 'obviously,' to concentrate on the primary task of speaking authoritatively (and, importantly, briefly) about its topic. Thus, the conditions of production of these particular texts are constructed as irrelevancies for, and obstacles to, the proper work. Social science inquiry that focuses on conditions of production has always acted (intentionally or not) to 'make trouble' of this kind: Marxist economics for capitalism; sociology of education for teaching; STS for science and technology. Reflexivity adds a twist to this dynamic: in collapsing the external/internal, outsider/insider, on/in relation, the source of trouble, being as much 'here' as it is 'there,' is no longer clearly located or easily attacked. This, then, is the primary role for reflexive inquiry: to make things difficult; to enrich the process of inquiry; to go round and round in circles; to fill out the picture; to frustrate all action; to act authentically and without self-contradiction; to change the subject.

Reflexivity as a Problem

In STS, the impetus for explicitly reflexive studies has been felt most strongly in the important subfield of STS known as the sociology of scientific knowledge (SSK). Here, the reflexive tie between its subject (knowledge about scientific knowledge) and its object (scientific knowledge) is very marked. That is, SSK exhibits a strong 'X of X' structure. Because of this, and because SSK's epistemology is explicitly relativist, reflexivity in SSK has a 'bite' akin to classic philosophical skepticism.

Extremely crudely, and for illustrative purposes only, SSK's relativism with respect to scientific knowledge can be, and indeed has been, taken as equivalent to a statement of this kind: "Scientific knowledge does not constitute truth because it is socially constructed." Read this way, SSK then becomes vulnerable to the standard antirelativist critique, the argument from self-refutation: "If this statement is asserted as a (scientific) truth, then it suffers from performative self-contradiction, and thus refutes itself" (see, for example, Hollis and Lukes, 1982). Note that the self-refutation argument uses reflexivity: it is through the process of turning its target onto itself that the demonstration of self-refutation is effected.

This is indeed one of the commonest modes in which reflexivity operates: as a weapon in the agonistic game of criticism. Thus, for those threatened with attack-by-reflexivity, it becomes a source of vulnerability to be abjured and defended against. Reflexivity becomes a 'problem' to be 'managed.' The various ways in which practitioners of SSK have done this, have made up one of the preoccupations of 'reflexivist' studies (see Section Relativist Consistency, Radical Reflexivity: The Reflexivists).

Logical Consistency: Bloor's Strong Program

David Bloor's *Knowledge and Social Imagery* (1991) is one of the earliest texts to set out the elements of what has become known as SSK. It contains an endorsement of reflexivity as the fourth and last of the 'tenets' for the "strong program in the sociology of knowledge": "In principle [the strong program's] patterns of explanation would have to be applicable to sociology itself. Like the requirement of symmetry, this is a response to the need to seek general explanations. It is an obvious requirement of principle because otherwise sociology would be a standing refutation of its own theories" (Bloor, 1991: p. 7).

Here, Bloor's positive support for reflexivity is a matter of logical consistency and a creative way of managing the self-refutation charge. Rather than deny the reflexive applicability of the strong program, Bloor emphasizes it, secure in his understanding that the socially constructed character of all knowledge, including his own, does not count against its validity. Bloor's reflexivity, however, remains at this 'in principle' level. His negative attitude toward any empirical instantiation of his reflexive programmatics is documented in a text that claims to be precisely that (Ashmore, 1989). Thus for Bloor (as well as for his former 'Edinburgh school' colleague Barry Barnes (1974), Epilogue), reflexivity is an argument used to defend a position; it is certainly not a practice.

Harry Collins: Special Relativism and the Ban on Reflexivity

Not so for, for example, Harry Collins (1985), perhaps the most influential writer through the heyday of SSK (*c.*1975–90). Collins's doctrine of 'special relativism' separates the natural world from the social world and the distinctive tasks and 'natural attitudes' of the natural scientist from those of the social scientist. The sociologist of scientific knowledge should treat the natural world as socially constructed, while treating the social world as 'real' and a source of sound data. However, this sense of 'realism' is taken from a reading of how natural scientists act toward their realm of inquiry, namely, the natural world. In effect, what Collins is doing here is to deny, for purposes of the conduct of research, the practical relevance of the purported reflexive similarity between subject and object in SSK. The result is the short-circuiting of attempts to apply Collins' conclusions about, for example, the replication of experiments to those conclusions themselves (nevertheless, see Mulkay, 1985: Chapters 4 and 5; Ashmore, 1989: Chapter 4). In short, Collins' 'management strategy' for reflexivity is to ban it altogether. This radical solution to the problem of reflexivity has laid Collins open to charges of inconsistency from both wings of the 'realist–relativist' continuum. The recommended moves for repairing this perceived inconsistency differ, however, with the 'realist' suggesting he abandon his problematic relativism about nature, and the 'more-radical-relativist' proposing the dropping of his social realism (*see* Technology, Social Construction of).

Relativist Consistency, Radical Reflexivity: The Reflexivists

In the late 1980s, a small number of 'full-blown' reflexive texts were published (Ashmore, 1989; Mulkay, 1985; Woolgar, 1988). This group of writers became known as 'reflexivists.' Ashmore's *The Reflexive Thesis* is an SSK study of SSK, a general exploration of reflexivity, and an experiment in 'new literary forms.' Mulkay's *The Word and the World* is a series of 'explorations in the form of sociological analysis' that draws upon analyses of biochemists' discourse. Woolgar's *Knowledge and Reflexivity* is a collection of papers from a small group of regular contributors (including Ashmore, Mulkay, and Woolgar) to the Discourse and Reflexivity Workshops held in the United Kingdom in the early 1980s. What links all these texts together, apart from their authors' close socio-cognitive relations to one another, is the common attempt to treat reflexivity less as a problem and more as an opportunity; an opportunity, even, for celebration. What happens when one stops evading reflexivity, when one looks 'the monster' in the eye, when one gives it a place at the heart of the research process? Self-destruction? Massive irrelevance? Third apocalyptic simile?

Bruno Latour's contribution to *Knowledge and Reflexivity* (Latour, 1988) is an exception. Latour argues that the 'meta-reflexive' writing strategies of the reflexivists provide a solution to a false problem: namely, the imagined naiveté of readers who believe too much and too easily. On the contrary, the real problem that (scientific, factual) writers have, is to be believed, or indeed, read at all. Latour further claims that the reflexivists are searching, despite their own gestures at weakening their own authorial privilege, for a final meta-level that leaves no remainder and from which vantage point, everything can be taken into (their) account. As an antidote to such impossible ambitions, Latour recommends a strategy of 'infra-reflexivity': just to tell stories, drawing upon the richness and variety of all the 'ordinary' resources available. Latour's *Aramis* (1992, 1996) – despite its appearance as an experiment in 'new literary forms' – is, perhaps, an example.

Criticisms: Pitfalls and Pratfalls

Reflexivity gets you nowhere. It is not a useful thing to do and its results cannot be used. It is impractical. As an activity, it immobilizes you. It leads to idle navel-gazing and foolish self-absorption. It is 'all pots and pans and no pudding' (Bloor, cited in Ashmore, 1989: p. 20). Logically, it leads to the abyss of the infinite regress or the never-ending search for the perfect meta-level. Reflexivity is politically and morally irresponsible games playing, fiddling while Rome burns. It has no soul (to bare) and no wisdom (to distribute), just an overdose of clever–clever irony. It disables criticism by taking that role on for itself; in its attempt to play all the roles of the academic community at once, and all by itself, it betrays itself as deeply, and ironically, antisocial. Indeed, it appears to seek nothing from its readers but a kind of 'aesthetic' admiration marked by a knowing, smug, and boyish amusement.

However, for the larger STS community, reflexivity has had positive functions and effects. The reflexivist fulfills the important role of the jester at the court of STS. Moreover, the presence of reflexive work as part of the 'canon' of STS, acts as a corrective to a kind of methodological overconfidence resulting from the recent success of the field. Reflexivity

reminds us that the fundamental epistemic questions and quandaries that originally animated SSK are permanent features of the practice of 'inquiry into inquiry' epitomized by STS scholarship. The text that aspires to 'transparency,' that is, blithely unconcerned with the conditions of its own production, is now, at the very least, 'visible,' a member of a marked, rather than unmarked category.

The Current Situation"

Careful or instructed readers will notice that all of the above (starting in the first section, and running through the end of this section's heading) is 'in quotes,' which indicates that the quotation includes almost all of this article, up to 'this' point, as it appeared in the 2001 edition of this *encyclopedia* (Ashmore, 2001). So what, then, now, is the current 'Current Situation' – though not, unfortunately, the (post-publication, 2015 and beyond) current, but the compositional Now of late 2012?

New 'New Literary Forms'?

Though this author, clearly, is still given to playing minor textual games, are these still being played in STS in the post-reflexivist era? Though hardly prominent, and certainly not trumpeted as *new* 'new literary forms,' experimental texts do still appear. Latour's *Aramis* (1992, 1996) (see Section Relativist Consistency, Radical Reflexivity: The Reflexivists) is more a detective story than a conventional technography of the eponymous, failed, rapid-transport system that is its topic; a story that even features the Aramis system accounting for its own death. Mol's *The Body Multiple* (2002) sustains a double text throughout: on top, her ethnography of the disease of atherosclerosis, and on the bottom her relation of her 'relation to the literature' – see also Ashmore (2005) which attempts a pale salute to Mol's achievement.

Radical Reflexivity: Is It All Over Now?

A session at the 4S/EASST (Society for Social Studies of Science/European Association for the Study of Science and Technology) conference held in Vienna in 2000 bore the title now given to this section. Most effectively dramatizing the question, one paper in that session, presented as a letter to its speaker, Steve Woolgar, from the seemingly-apocryphal 'Brian Peabody PhD OBE FREng' raised the question (a question raised before, according to the letter, by one of Woolgar's students) 'What happened?': "Looking for the standard indicators of the success of the field, you might ask where are the new journals of reflexivity, the international association of radical reflexivity (IARR), the annual conferences (Reflex, 2000), the newly established Chairs in Reflexivity and so on" (Woolgar, 2000: p. 3). Peabody/Woolgar then goes on to suggest that reflexivists would be bound to be skeptical of the value and desirability of such indicators, and thus, precisely because of their absence, "… reflexivity has succeeded spectacularly. Reflexivity is a success because it has argued itself out of a job" (Woolgar, 2000: p. 3). Rappert (2007), which includes a commentary on the Woolgar/Peabody 'personal correspondence' of 2000, treats the 'failure' of the reflexivist program as his main example of the difficulties involved in making constructivist generalizations. (Interestingly, Rappert's paper, despite its negative conclusions about the program that introduced such devices, uses an unconventional double text – see Section New 'New Literary Forms'?) Knuuttila (2002) conducts a similar postmortem on STS reflexivism, concluding that "[i]n retrospect, it seems fair to say that radical reflexivity of STS stumbled against its own impossibility" (paragraph 28). Pels (2000) is less apocalyptic in his rejection of reflexivism, which he understands as a matter of the infinite ascent of meta-levels, with no (logical) endpoint. He recommends, both as an alternative to this endless spiral and to the flatness of classical realism, 'one step up,' no more and no less. In this, as he acknowledges, he is close to Latour's (1988) 'infra-reflexivity' (see Section Relativist Consistency, Radical Reflexivity: The Reflexivists).

Social Science's Ubiquitous Reflexivities

Meanwhile, elsewhere, 'reflexive' has, over the last 20 years or so, become a highly desirable attribute of any ethnographic, qualitative study throughout the social sciences (May, 1999). What is meant by this term, though, when lauded so positively, is rarely anything like its understanding in STS reflexivism. Most often, a claim that an author (or text, or work of research) is reflexive means that some notice is taken of the involvement of the 'subject' in the working up of the 'object.' (A representative discussion is Finlay, 2002.) The extent and character of this involvement, as Lynch's (2000) comprehensive typology makes clear, is highly variable, but is often satisfied with an autobiographical gesture or two. As an indicator of the near-ubiquity of reflexive claims in qualitative social science, the online journal *Forum: Qualitative Research* published two special issues on 'Subjectivity and Reflexivity in Qualitative Research' (issue 3(3) in 2002 and 4(2) in 2003) comprising a total of 32 articles. Lynch's (2000) effort is rare enough, these days, in mounting an argument, to quote the title, 'Against Reflexivity as an Academic Virtue and Source of Privileged Knowledge.' As a precursor to mounting this argument he catalogs the reflexivities into six major types: 1. mechanical (with three variants); 2. substantive (two variants); 3. methodological (four variants); 4. meta-theoretical (three variants); 5. interpretative (two variants; 5b being the 'radical reflexivity' associated with STS reflexivism); 6. ethnomethodological (just the one). Lynch's preference is for type 6, as, being a ubiquitous and unavoidable feature of the social world, ethnomethodological reflexivity cannot be a matter of claims-making, and thus the potential gaining of virtue or privilege, at all.

The term 'reflexivity' has also become prominent in social theory, particularly in the influential thesis of 'reflexive modernization' put forward by Beck et al. (1994) – Lynch's (2000) type 2a: substantive/systemic – where 'our' modernity is characterized as pervaded by reflexive monitoring of its own processes in the form of cost–benefit and risk analyses, opinion-polling, and so forth. Another well-known social-theoretic source of reflexivity occurs in and as the work of Bourdieu (Bourdieu, 2001, 2004, 2008; Bourdieu and

Wacquant, 1992; see Maton, 2003 for a general discussion). Bourdieu's 'hyperobjectivist' form of reflexivity (type 4c) – entailing the scientific objectivation of the subject of objectivation – incurs the wrath of Lynch (2000: pp. 30–31); but then, at the end of his life, Bourdieu (2001, 2004) directed his own wrath at 'the new sociology of science' (SSK/STS), a move lovingly described by Mialet (2003). A later posthumous work (2008) is reviewed by Gingras (2010) as demonstrating the superiority of Bourdieu's 'sociological' reflexivity over the 'textual' version homegrown in STS "that rapidly led to an impasse" (Gingras, 2010: p. 619).

STS' Topical Extensions

It has been a long time since STS was mainly SSK together with ANT (actor network theory); indeed this couple largely ceased to cohabit following the polemics published in Pickering (1992). STS has become extraordinarily diverse, topically, conceptually, institutionally, and nationally. Though the acronym STS seems to specify a single (double) topic, that of science and technology, its topics have now expanded into a whole range of what Lynch (1993) has called 'epistopics': law and legal deliberations, politics and policy, engineering and design, medicine and health care, finance and markets; and all the hybridities performed when 'science' and/or 'technology' (technoscience) meet (these and) other institutions and practices – particularly as heterogeneous assemblages forming around public 'matters of concern' as opposed to the private laboratory-based production of 'matters of fact' (Latour, 2004; Latour and Weibel, 2005). Where, if anywhere, is the place of reflexivity in this STS?

The varieties of reflexivities (see section Social Science's Ubiquitous Reflexivities) have made themselves felt in STS, as they have elsewhere. A recent special issue of *Science as Culture* (Cohen and Galusky, 2010) is dedicated to reflexivity conceived as how personal narrative and experience can lead to STS-relevant insights. A rather different sense of the term is present in Wynne's normative call for greater 'institutional reflexivity' (Wynne, 1993); an emphasis that applies the (positive) attribute of reflexivity, not to the STS researcher, but to the researched; that is, to scientific institutions that tend to suffer from a self-awareness deficit. And what of the emphasis of the 1980s reflexivists? It can sometimes seem as if this never existed. Collyer (2011: p. 316) argues that "despite its potential, theorising within the sociology of science and technology is hampered by insufficient attention to … the concept and practice of reflexivity"! An endangered species, evidently (see Section Radical Reflexivity: Is It All Over Now?) but not yet quite dead, though largely confined to a counterintuitive corner of the STS universe – that of business schools and marketing research – courtesy of the Final Reflexivist, Steve Woolgar's elevation to Professor of Marketing at Säid Business School, Oxford University (Neyland and Simakova, 2009).

Thus…

… though the career of 'reflexivity' has never been healthier, it has, in the STS context, lost much of its radicalism, bite, and distinctiveness, much as Pollner (1991) complained, years ago, in the case of ethnomethodology. Whether this, Dear Reader, is any cause for concern, is, as always, a matter in you to decide.

See also: Actor-Network Theory; Human Sciences, History of; Science and Technology Studies, Ethnomethodology of; Scientific Knowledge, Sociology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Strong Program, Sociology of; Technology, Social Construction of; Truth and Credibility in Science.

Bibliography

Ashmore, M., 1989. The Reflexive Thesis: Wrighting Sociology of Scientific Knowledge. University of Chicago Press, Chicago.

Ashmore, M., 2001. Reflexivity, in science and technology studies. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social and Behavioural Sciences. Elsevier, Oxford.

Ashmore, M., 2005. The life inside/the left-hand side (review of A Mol (2002) The Body Multiple). Social Studies of Science 35, 827–830.

Barnes, B., 1974. Scientific Knowledge and Sociological Theory. Routledge & Kegan Paul, London.

Beck, U., Giddens, A., Lash, S. (Eds.), 1994. Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order. Polity Press, Cambridge.

Bloor, D., 1991. Knowledge and Social Imagery, second ed. University of Chicago Press, Chicago.

Bourdieu, P., 2001. Science de la Science et Réflexivité. Editions Raisons d'Agir, Collection 'Cours et Travaux,' Paris.

Bourdieu, P., 2004. Science of Science and Reflexivity. University of Chicago Press, Chicago.

Bourdieu, P., 2008. Sketch for a Self-analysis. University of Chicago Press, Chicago, IL.

Bourdieu, P., Wacquant, L.J.D., 1992. An Invitation to Reflexive Sociology. Polity Press, Cambridge, UK.

Cohen, B.R., Galusky, W., 2010. Guest editorial. Science as Culture 19 (1), 1–14.

Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.

Collyer, F., 2011. Reflexivity and the sociology of science and technology: the invention of "eryc" the antibiotic. The Qualitative Report 16 (2), 316–340. http://www.nova.edu/ssss/QR/QR16-2/collyer.pdf (accessed 24.07.13.).

Finlay, L., 2002. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. Qualitative Research 2 (2), 209–230.

Gingras, Y., 2010. Sociological reflexivity in action (review of P Bourdieu (2008) Sketch for a self-analysis). Social Studies of Science 40 (4), 619–631.

Hollis, M., Lukes, S. (Eds.), 1982. Rationality and Relativism. Blackwell, Oxford.

Knuuttila, T., September 2002. Signing for reflexivity: constructionist rhetorics and its reflexive critique in science and technology studies. Forum: Qualitative Research 3 (3). http://www.qualitative-research.net/index.php/fqs/article/view/828#g23 (accessed 24.07.13.).

Latour, B., 1988. The politics of explanation: an alternative. In: Woolgar, S. (Ed.), Knowledge and Reflexivity. Sage, London, pp. 154–176.

Latour, B., 1992. Aramis ou l'amour des techniques. La Decouverte, Paris.

Latour, B., 1996. Aramis or the Love of Technology. Harvard University Press, Cambridge, MA.

Latour, B., 2004. Why has critique run out of steam? From matters of fact to matters of concern. Critical Inquiry 30, 225–248.

Latour, B., Weibel, P., 2005. Making Things Public: Atmospheres of Democracy. MIT Press, Cambridge, MA.

Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge University Press, Cambridge & New York.

Lynch, M., 2000. Against reflexivity as an academic virtue and source of privileged knowledge. Theory, Culture and Society 17 (3), 26–54.

Maton, K., 2003. Reflexivity, relationism and research: Pierre Bourdieu and the epistemic conditions of social scientific knowledge. Space & Culture 6 (1), 52–65.

May, T., 1999. Reflexivity and sociological practice. Sociological Research Online 4 (3). http://www.socresonline.org.uk/4/3/may.html (accessed 24.07.13.).

Mialet, H., 2003. The 'righteous wrath' of Pierre Bourdieu, review of Pierre Bourdieu (2001) Science de la Science et Réflexivité. Social Studies of Science 33 (4), 613–621.

Mol, A., 2002. The Body Multiple: Ontology in Medical Practice. Duke University Press, Durham, NC.
Mulkay, M., 1985. The Word and the World: Explorations in the Form of Sociological Analysis. George Allen & Unwin, London.
Neyland, D., Simakova, E., 2009. How far can we push sceptical reflexivity? An analysis of marketing ethics and the certification of poverty. Journal of Marketing Management 25 (7–8), 777–794.
Pels, D., 2000. Reflexivity: one step up. Theory, Culture and Society 17 (3), 1–25.
Pickering, A. (Ed.), 1992. Science as Practice and Culture. University of Chicago Press, Chicago.
Pollner, M., 1991. Left of ethnomethodology: the rise and decline of radical reflexivity. American Sociological Review 56, 370–380.
Rappert, B., 2007. On the mid range: an exercise in disposing (or minding the gaps). Science, Technology & Human Values 32 (6), 693–712.
Woolgar, S. (Ed.), 1988. Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge. Sage, London.
Woolgar, S., 2000. Personal Correspondence: Letter from Brian Peabody to Professor Steve Woolgar, June 6, 2000. Paper presented at 4S/EASST conference, Vienna, September 29, 2000.
Wynne, B., 1993. Public uptake of science: a case for institutional reflexivity. Public Understanding of Science 2 (4), 321–337.

Reformation and Confessionalization

Ute Lotz-Heumann, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the historiography and the major historical developments of the Protestant Reformation, Catholic reform, and confessionalization in sixteenth- and seventeenth-century Germany and Europe. In the first part of this article, the historiography of the Reformation and the confessional age is discussed, with particular emphasis on questions of periodization. In the second part, the major historical developments between the late Middle Ages and the middle of the seventeenth century in the German Empire as well as in Europe are summarized.

Historiography and Periodization

The era of the Reformation and Confessionalization encompasses roughly the period between 1500/1517 and 1650/1700, i.e., the first part of the early modern period (if one includes the Enlightenment in early modernity). But even this vague definition has been debated in historiography and there is by no means a consensus on the periodization of the Reformation and the confessional age.

Reformation and Counter-Reformation

In nineteenth-century German historiography, not least through the influence of Leopold von Ranke (Ranke, 1839–47), 'Reformation' and 'Counter-Reformation' ('*Gegenreformation*') were established as two distinct historical periods in the sixteenth century, with the Protestant Reformation occupying the earlier half and the Catholic Counter-Reformation the latter half. Building on earlier Protestant interpretations of the Reformation, Protestant Prussian historiography regarded the Reformation as a breakthrough to modernity, a German event that changed the course of world history by spreading values like rationality. This view was still reflected in the modified subtitle of the leading journal in the field, the *Archive for Reformation History*, after it became a German–United States collaboration in 1951: "international journal concerned with the history of the Reformation and its significance in world affairs" (ARG 42 (1951)). In contrast, 'Counter-Reformation' implied a mere reaction to the new Protestant movements of the early sixteenth century.

Over the course of the nineteenth and twentieth centuries, this view of the sixteenth century with its clear Protestant bias was much debated. With regard to Counter-Reformation, historians pointed to the fact that there were numerous Catholic reform movements during the late Middle Ages, which in turn made it clear that the Counter-Reformation was not only a reaction against Luther and other reformers, but an early modern 'Catholic Reformation' in its own right. This discussion culminated in Hubert Jedin's suggestion for a compromise term in 1946: 'Catholic reform and Counter-Reformation' (Jedin, 1999).

With regard to the Reformation as a marker of a new era and a breakthrough to modernity, twentieth-century scholars cast serious doubt on this conception. Ernst Troeltsch argued in 1906 that the 'Old Protestantism' ('*Altprotestantismus*') of Luther and the other reformers was deeply embedded in the Middle Ages and only the 'New Protestantism' ('*Neuprotestantismus*') of Protestant groups like Baptists and Spiritualists had brought about modernity (Troeltsch, 1906). Later historians emphasized long-term trends and comparatively stable structures in society and economy (for example, the social hierarchy or the household as the basic social and economic unit) which, in their view, minimized the significance of the religious upheavals of the sixteenth century for periodization. For example, the German-American historian Dietrich Gerhard suggested to replace the idea of a historical break around the year 1500, with the Protestant Reformation as the major component, with a period called 'Old Europe' lasting from the eleventh until the end of the eighteenth century (Gerhard, 1962).

Reformation and Confessionalization

Since the 1960s, and especially the 1980s, these discussions have continued, with new historiographical concepts throwing different lights on the question of periodization. In terms of the Reformation as an epochal break, Marxist historians in East Germany – taking Engel's interpretation of the German Peasants' War as a positive revolutionary event as their starting point – considered the Peasants' War with Thomas Müntzer as its leader and the radical Reformation an 'early bourgeois Revolution.' This interpretation of the Reformation largely ended with the fall of the Iron Curtain.

In contrast, it more and more became a consensus among historians in the West (albeit with some variation and different emphases) that the Reformation should be regarded as part of a longer period of change encompassing the late Middle Ages and beginning of the early modern period. Heiko A. Oberman described the Reformation as the culmination of a long, late medieval reform process which, however, did not culminate in the Lutheran, but in the western European, Calvinist Reformations (Oberman, 2001). Heinz Schilling has argued that the Reformation was part of a long-term challenge and response process lasting from the reforms of the late Middle Ages until the age of Confessionalization (Schilling, 1998). And in Berndt Hamm's model of 'normative centering,' the Reformation is seen as a theological break, but, overall, as part of a much broader process of centralization and concentration in all areas

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62009-3

of late medieval and early modern life (politics, society, economy, the arts) (Hamm, 1993).

With regard to the relationship between Protestantism in the second half of the sixteenth century and the 'Counter-Reformation,' Ernst Walter Zeeden fundamentally changed the trajectory of historical research when he observed in 1958 that, during the second half of the sixteenth century, Catholicism, Lutheranism, and Calvinism each started to build a modern, clearly defined church, each of which centered on a specific confession of faith. He called this process 'confession building' ('Konfessionsbildung'), a neutral term which could be applied to all churches (Zeeden, 1958).

In the late 1970s and the early 1980s, two scholars of the next generation, Wolfgang Reinhard and Heinz Schilling, reinforced this historiographical change by developing the concept of 'confessionalization' (e.g., Schilling, 1986; Reinhard and Schilling, 1995). This paradigm dominated historiographical discussions for the next 25 years and resulted in a shift in research interests. While the historiography of the 1960s and 1970s had focused on the first half of the sixteenth century with a particular emphasis on the urban Reformation (e.g., Brady, 1978; Greyerz, 1980; Moeller, 1987), from the 1980s onward, historians shifted their attention to the second half of the sixteenth century and the early years of the seventeenth century.

The concept of confessionalization countered the negative and antimodern implications of the term 'Counter-Reformation' by replacing it with the idea that the confessional churches of the early modern period underwent parallel developments. As a result, Schilling and Reinhard suggested replacing the terms 'Lutheran orthodoxy,' 'Second Reformation' (for Calvinism), and 'Counter-Reformation' with the parallel terms 'Lutheran,' 'Calvinist,' or 'Catholic confessionalization' (Schilling, 1992: pp. 209–210). Nevertheless, the periodization of the era of confessionalization remains unclear. Heinz Schilling has developed a periodization of the process of confessionalization in the Empire between the Peace of Augsburg in 1555 and the Thirty Years' War (Schilling, 1992: pp. 216–232). Schilling sees four phases: first, a 'preparatory phase' starting in the late 1540s; second, the 'transition to confessional confrontation' in the 1570s, when all three churches acquired clear confessions of faith and began a process of boundary and identity building; third, the 'apogee of confessionalization' between the 1580s and 1620s, when the hardening of religious lines continued and was reinforced by political conflicts in the Empire, leading to the outbreak of the Thirty Years' War; and fourth, 'the end of confessionalization under the conditions of war and on the basis of the Peace of Westphalia' starting in the 1620s. In this phase, the devastation and suffering of the war led to the realization, resulting in the Peace of Westphalia, that compromise was the only realistic option.

The debate about the concept of confessionalization has led many scholars, especially those who work on Catholic areas, to doubt Heinz Schilling's periodization. Wolfgang Reinhard has extended the process of confessionalization much further, starting with the Augsburg Confession of 1530, the first Lutheran confession, and ending with the expulsion of the Salzburg Protestants in 1731–32 which was, according to Reinhard, the last confessionally motivated act of expulsion in the early modern Empire (Reinhard, 1995: p. 435). Others have argued that the thesis of parallel developments (and thus a periodization applying equally to all confessional churches) cannot stand, particularly in view of the fact that Baroque Catholicism continued to shape Catholic confessional identity and culture well into the eighteenth century.

Heinz Schilling and Wolfgang Reinhard understand the concept of confessionalization as a broadening of the concept of confession building, which was focused on religious history, into a paradigm of societal history. As a result, they regard confessionalization as a 'fundamental process of society' (Schilling, 1992: p. 209) in which all confessional churches, in close alliance with the early modern state, homogenized the population, intensified social disciplining, and solidified confessional, political, and cultural identities. They argue (in clear opposition to the earlier implications of the term 'Counter-Reformation') that all churches thereby underwent a (partially unintended) process of modernization, and that, rather than focusing on doctrinal differences as the older historiography did, the concept of confessionalization reveals the "functional similarities" (Schilling, 1992: p. 210) between the early modern confessional churches.

In this context, Wolfgang Reinhard has identified seven methods, or mechanisms, of confessionalization, used by church and state to try and establish confessional homogeneity (Reinhard, 1999: pp. 177–182). First, the definition of 'pure doctrine' in a confession of faith meant that a church was defined vis-à-vis other confessional churches. Second, these norms then needed to be imposed on the personnel of the church, for example, through confessional oaths. Third, church and state used the printing press for propaganda and tried to put in place a system of censorship. Fourth, by establishing educational institutions on all levels, church and state attempted to influence the younger generations. Fifth, through different measures of social control like visitations and church discipline, church and state aimed at creating a confessionally homogeneous population. Sixth, religious rituals were cultivated as markers of a confessional church, and participation was monitored so as to ensure the loyalty of the faithful. Seventh, Reinhard refers to the confessionalization of language, for example, the choice of first names in different confessions.

The concept of confessionalization resulted in an intense and long-lasting historiographical debate which started in the early 1980s and is still ongoing, although it has significantly lessened in intensity since the mid-2000s. This debate has focused on elements of the concept that many historians have regarded as too rigid and/or limiting. Above all, the concept of confessionalization has been described as a product of German societal history of the 1970s with its focus on modernization theory following Max Weber. Social and cultural historians have rejected the concept on the grounds that its claim that confessionalization had modernizing effects is teleological and that it does not do justice to the complexities of sixteenth- and seventeenth-century developments. In a similar vein, historians and church historians have criticized the concept for its emphasis on parallel developments in all major confessions. In their view, this detracts from the specific characteristics of each confessional church and reduces religion to its social and political functions.

Another major criticism of the concept pertains to its claim that confessionalization was a fundamental social process, affecting all areas of early modern life. There has been much debate on the question how deeply confessionalization reached into different areas and aspects of life with critics of the concept putting emphasis on its limitations, for example, with regard to changing popular religion on the one end of the social spectrum and the supraconfessional nature of the Humanist republic of letters on the other. Finally, the close relationship between confession building and state formation postulated by the concept of confessionalization has given rise to major discussions. In recent historiography, there is a consensus that Reinhard and Schilling have overemphasized the role of the state in the process of confessionalization, thereby describing it as a top-to-bottom process which does not take into account agency from below, various forms of resistance, and self-confessionalization.

Overall, the social history of the Reformation and confessional age, which – parallel to the debates about the concept of confessionalization – dominated the field in the later twentieth century and emphasized subjects like popular religion and gender (e.g., Scribner, 1981; Roper, 1989; Wiesner, 1993), has developed in the last 20 years into a new cultural history. On the one hand, this new cultural history is concerned with completely new questions, researching emotions, rituals, space, sound, and material culture in the religious movements and churches of the sixteenth and seventeenth centuries (e.g., Karant-Nunn, 1997, 2010; Spicer, 2012). In doing so, this new cultural history has moved beyond the issues that earlier generations of historians struggled with: The term 'Counter-Reformation' is now used as a completely neutral term without its previous negative connotations. The term 'Reformations' or 'age of Reformations' (in plural) is often employed to denote a period encompassing the Reformation and the confessional age. And historians analyze Protestantism and Catholicism 'on the ground' in the early modern period without currently discussing macrohistorical questions of modernization, state formation, or social control.

On the other hand, the 500th anniversary of the Lutheran Reformation in 2017 will not be without consequences for historical research. As this article goes to press, it is not yet clear how this will affect research on religion in the early modern period overall. However, two trends are already discernible: Research interests are clearly shifting back to the period of the early Reformation. And there seems to be a growing interest in the biographies of the central figures of the Reformation, above all, of course, Luther.

Religious, Social, and Political Structures and Developments in the Era of the Reformation and Confessionalization

Electoral Saxony, the territory where the Lutheran Reformation originated, was part of the Holy Roman Empire of the German Nation, a conglomerate of territories of different sizes and political power in central Europe. When the Wittenberg theology professor Martin Luther (1483–1546) published his 95 theses and other writings in an effort to reform the Church, his own fate and the fate of the evangelical movement were determined by the political makeup and developments in the German Empire and its territories. Soon, however, the Protestant Reformation, which could not be repressed in Germany, became a European phenomenon.

Other major reformers emerged, not least Huldrych Zwingli (1484–1531) in Zurich, who deeply influenced Reformed theology in Switzerland and upper Germany, and Jean Calvin (1509–64), the Genevan reformer, whose theology spread across wide swathes of Europe (France, the Netherlands, the German territories, the British Isles, and east central Europe). As a result, Protestant confessional churches were established all over central and northern Europe, each with a different doctrinal basis, ecclesiastical organization, and church–state relationship. At the same time (and only partially as a reaction to the Protestant Reformation), the Catholic Church underwent a fundamental process of reform. The Council of Trent (1545–63) defined the Catholic Church as a confessional church both in theology and church organization.

Basic Structures

The Reformation and the confessional age were characterized by a number of basic religious, social, and political structures that are important to recognize when analyzing the events and developments of the sixteenth and seventeenth centuries. First, it is essential to recognize that there existed in the early modern period a strong ideal of the unity of church and state. In the eyes of contemporaries, one state should have only one religion. As a result, rulers attempted to unify religion and politics and use religion as an instrument of state formation.

Second, this did of course not mean that rulers succeeded and achieved their ideal. Rather, the process of integration of politics and religion was one of negotiation with regional and local elites as well as the populace. Early modern states, with their limited bureaucracies and their resulting difficulties to govern and control their territories, had to rely heavily on the collaboration of regional and local elites, for example, noblemen and urban magistrates. If this collaboration was not forthcoming, religious and political attempts at unification often stalled or failed altogether. As a result, opposition and resistance to the introduction of the Reformation or measures of confessionalization were just as common in early modern Europe as compliance with such measures.

Third, especially with regard to the general population, it is important to differentiate between conversion – understood as the conscious adoption of a set of religious beliefs and practices – and conformity – defined as outward compliance with a church by law established in a certain territory. Although we cannot always tell whether a particular person had converted or 'merely' conformed, the significance of conformity as a group phenomenon cannot be overrated in early modern Europe. While both Protestant and Catholic reformers lamented the existence of various forms of popular religion, this usually did not affect the populace's general conformity with the established church. However, it did make a huge difference whether the majority of the population in a territory conformed with the state church

and participated in its rituals or whether they failed to do so and actively resisted the official church.

The Late Middle Ages

Historians today agree that the late Middle Ages laid important groundwork for the Protestant Reformation. In particular, attempts at reform of the Church through a general council were an important feature of the age as were a number of dissident or, as they were regarded at the time, 'heretical' movements. In addition, a number of different movements, both clerical and lay, attest to the broad reform efforts during the late medieval era.

The conciliar movement of the late Middle Ages aimed at a comprehensive reform of the Church through a general council. Although several great councils took place (above all, Constance, 1414–18, Basel–Ferrara–Florence 1431–49), and the conflict between the papacy and conciliarism was decided in favor of the papacy just before the advent of the Protestant Reformation, it remains an open question in historiography whether the reform of the Church had not sufficiently progressed by 1517 or whether, on the contrary, it had made so much progress that it was regarded as a threat in the more remote areas of the Latin Church such as Saxony.

During the late Middle Ages, there were a number of 'heretical' movements in different parts of Europe, which were of course religious movements in their own right, but were later regarded as precursors of the Protestant Reformation. The Waldensians sprang up in the region around Lyon in France in the twelfth century as a religious movement that shunned wealth and rejected the authority of the Church. After having been declared heretics, they were persecuted by the Inquisition throughout the late Middle Ages. The small surviving group took up the ideas of the Protestant Reformation in the sixteenth century.

In the fourteenth century, John Wyclif started the religious movement of the so-called Lollards, which focused on reading the Bible in the native language, and therefore rejected the veneration of saints and images. The Lollards were declared heretics by several church councils, but the movement survived in England until the sixteenth century when, just as the Waldensians, the remaining Lollards became part of the Protestant movement there.

Finally, Jan Hus, who was influenced by Wyclif's ideas, started the movement of the Hussites in Bohemia. He rejected papal authority, propagated reading the Bible in the native language as well as communion in both kinds (that is, for laypeople to receive the wine in addition to the bread). At the Council of Constance, Hus was convicted of heresy and burned at the stake. Nevertheless, the two branches of the Hussite movement, the Utraquists and the Bohemian Brothers, survived into the early modern period.

Various reform movements were active in the late Middle Ages, among others the Observant movement of the mendicant friars, the Humanist movement at the universities, and the Brethren of the Common Life, a lay movement which originated in the Netherlands in the fourteenth century. Overall, the late Middle Ages saw in intensification of lay piety in the form of the cult of the saints, pilgrimages and processions, and indulgences. While on the one hand, these movements confirm that the Protestant Reformation was part of a long era of change and reform that encompassed the late Middle Ages and the first half of the early modern period, on the other hand, the mixed image of the late medieval Church seemingly caught between crisis and revitalization continues to make it difficult for historians to explain how and why the Protestant Reformation took hold.

Protestant and Catholic Theology in the Era of the Reformation and Confessionalization

The theological ideas of Martin Luther and other reformers ultimately resulted in the division of Latin Christianity into multiple confessional churches. However, one should regard Luther himself as a man of the Middle Ages who wanted to reform, not divide, the Church. Luther's theology built upon medieval foundations, but it also, as Berndt Hamm has formulated, constituted a 'systemic break' in theological terms ('Systembruch') (Hamm, 1993: p. 79). Luther's theology was above all characterized by the so-called *sola* principles: *sola gratia*, *sola fide*, *sola scriptura* – by grace alone, by faith alone, by Scripture alone. Luther rejected the idea of salvation through works and postulated that man was saved only through God's grace. In addition, the Bible was to be the only authority in questions of faith, not the tradition of the Church. This resulted in a sweeping rejection of medieval beliefs and practices: Luther was opposed to indulgences and the veneration of relics and saints, he rejected the role of the priest as a mediator to God and postulated the priesthood of all believers. He also radically reduced the number of the sacraments from seven to two: baptism and communion, which the laity received in both kinds.

The Protestant movements, which emerged over the course of the sixteenth century, differed, above all, in their views of the Lord's Supper. While Luther rejected the idea of transubstantiation, he maintained the real presence of the body and blood of Christ in the bread and wine during communion. In contrast, Zwingli regarded communion as a memorial meal held in commemoration of the last supper of Christ with the elements serving only as symbols. Calvin developed a compromise version between Luther's and Zwingli's understandings of communion, arguing that Christ is spiritually (but not corporeally) present during the Lord's Supper.

In addition to the major confessional churches, radical Protestant groups emerged during the sixteenth century, especially the Anabaptists, a group that originated in Zwingli's Zurich. Anabaptists rejected the baptism of children and practiced adult baptism. Other radical positions, for example their strict interpretation of the priesthood of all believers and their refusal to take oaths, led to their being persecuted by the authorities, especially after the excesses of the Anabaptist Kingdom of Münster in 1534/35.

The Catholic Church responded to the theological challenge of the Protestant Reformation by clearly defining its own doctrines in the decrees of the Council of Trent. The council reaffirmed the doctrine of transubstantiation as well as the seven Sacraments. It stressed the dual nature of justification through both God's grace and good works, and emphasized the validity of tradition against the Protestant principle of *sola scriptura*. The major doctrines were enshrined in a new

confession of faith, the *Professio Fidei Tridentina*. In addition to these doctrinal decrees, the Council of Trent also promulgated a whole range of decrees concerning ecclesiastical reform. The gradual implementation of these decrees, together with the activities of new orders, not least the Jesuits (founded in 1534 by Ignatius of Loyola), resulted in the formation of a confessional church, which was active all over the world.

Germany during the Age of the Reformation and Confessionalization

Why did the Lutheran Reformation survive and even succeed? This question can only be answered by looking at the political, social, and communicative contexts and in particular the unique structures of the Holy Roman Empire of the German Nation. Very soon after Luther had sent his 95 theses against indulgences to the Archbishop of Mainz in order to start a theological disputation in 1517, the Reformation became a media event, thanks to the recently invented printing press with movable type. The 95 theses were translated into German and published, and Luther followed up with further publications. The heresy trial against him was opened in Rome in the summer of 1518.

Shortly afterward, the politics of the Empire first played out to the advantage of the Protestant Reformation. Because Pope Leo X wanted to recruit the Elector of Saxony, Luther's territorial lord, as a candidate for the Emperorship against Charles V of the Habsburg family, a heresy trial seemed inopportune and was suspended. While the Pope was not successful in preventing the election of Charles V, the pattern of provisional measures and repeated suspension of a more forceful policy against the Reformation and its supporters among the imperial princes and cities characterized the entire first half of the sixteenth century and ultimately secured the success of the Reformation.

Charles V was preoccupied with enemies and wars on two fronts: against Francis I of France and against the Ottoman Empire. As a result, he spent many years outside the Empire and was willing to compromise again and again on the question of religion, especially to receive the so-called *Türkenhilfe* (a levy because of the Turkish danger) from the imperial estates. For example, at the imperial diet of Speyer in 1526, it was decided that, in exchange for a *Türkenhilfe*, the imperial estates were allowed to regulate religion in their territories in whatever way they thought best. This and similar moratoria on the enforcement of the Edict of Worms of 1521 (which had declared Luther a heretic) opened the door for princes to establish Lutheran churches in their territories, a step which could not be undone later.

By 1530, the two parties had affirmed their doctrinal positions to such a degree that the first two confessions of faith were presented to the Emperor at the imperial diet of Augsburg: the Lutheran *Confessio Augustana* and the Catholic *Confutatio*. And in 1531, the first Protestant military alliance was founded, the Schmalkaldic League. After the Peace of Crépy with France in 1544, Charles V was finally able to turn his attention to the Empire. However, in spite of his initial success against the Protestants in the Schmalkaldic War of 1546/47 and during the imperial diet of Augsburg in 1548 when he imposed the Augsburg Interim, Lutheranism in the Empire could no longer be suppressed. An alliance of the German Protestant princes with the French king led to the Princes' War of 1552, which the Emperor lost. As a result, Charles V left it to his brother Ferdinand to negotiate a peace with the Protestant princes.

The Peace of Augsburg of 1555 legally established Lutheranism next to Roman Catholicism as a confession in the Empire. The territorial princes gained the right to determine the confession of their territory and their subjects (*ius reformandi*, later enshrined by imperial lawyers in the phrase *cuius regio, eius religio* – whose realm, his religion). In order to ensure that the imperial Church and the electoral college would remain Catholic, the so-called *reservatum ecclesiasticum* (ecclesiastical reservation) was inserted into the peace treaty. It stipulated that if the ruler of an ecclesiastical territory, such as a prince-bishop, converted to Lutheranism, the *ius refomandi* principle would not apply and the convert would lose his position.

In the early years, the Reformation had been a media event as well as a popular movement. Even though A.G. Dickens's dictum that the German Reformation was 'an urban event' (Dickens, 1976: p. 182) might be an overstatement, there can be no doubt that the success of Lutheranism in the towns and especially the imperial cities largely contributed to its overall success. However, in the long run, the cities were more vulnerable than the territorial states. In contrast to the territories, the cities did not gain the *ius reformandi* in the Peace of Augsburg.

The imperial knights and the peasants had proved to be even more vulnerable. The imperial knights, many of them early supporters of Luther, and the peasants who adopted Lutheranism in the early years of the Reformation are examples of the amalgamation of religious and political ideas and demands that would prove to be characteristic for the entire age. However, both groups lost the wars they waged against the rising territorial princes of the Empire (the War of the Imperial Knights in 1522/23 and the Peasants' War in 1524/25).

The Peace of Augsburg brought several decades of peace to the Empire while, at the same time, a civil war over religion was raging in France. As such, it can only be considered a remarkable success. However, the peace treaty eventually caused friction and conflict in the Empire because, in order to find common ground, the treaty contained many dissimulating formulations which, over time, proved to be open to contentious interpretations. Starting in the 1580s, an increasing confessional polarization took hold in the Empire. This was intensified by the implementation of Tridentine reform in the Catholic territories on the one hand and the conversion of a number of German Lutheran princes to Calvinism on the other. The Lutheran church, which had been deeply divided after Luther's death, also united and confessionalized with the Formula and Book of Concord of 1577/1580.

Between 1583 and 1618, several conflicts broke out in the Empire which highlighted the contentious aspects of the Peace of Augsburg (for example, the Cologne War of 1583, the Donauwörth conflict of 1606, the succession conflict in Jülich-Cleve-Berg in 1609). As a result, the imperial institutions were paralyzed and military alliances were founded in 1608 (Protestant Union) and 1609 (Catholic League), respectively. Confessional war was increasingly on the horizon and because of the deep European connections of the Empire and its ruling families, this could only be a European war. It eventually broke out in 1618 in Bohemia, where the Protestant Bohemian

estates had rebelled against their Catholic Habsburg king and elected the Calvinist Elector of the Palatinate, Frederick V, instead. Thus, the Thirty Years' War began.

Europe during the Age of the Reformation and Confessionalization

We have seen above that both the processes of Reformation and confessionalization could result, on the one hand, in the religious and political integration of a city or territory under one confession or, on the other hand, in religious and political resistance and war. In conclusion, a brief 'tour' of early modern Europe serves to illustrate this further.

The religious integration of the recently united kingdoms of Spain was successful because early Protestant stirrings were quickly suppressed with the help of the Inquisition and a strong identity, which combined religious and proto-national elements, was established. Similarly, the Scandinavian kingdoms, Denmark and Sweden, successfully integrated Lutheran church building with state formation. The kingdom of England also belongs to this group. Queen Elizabeth I managed to establish Protestantism as part of national identity by favoring a gradual process of Protestantization in her realm.

However, England was, in fact, a 'composite monarchy' (Elliott, 1992) in the early modern period, which consisted, after 1603, of England, Wales, Ireland, and Scotland. While Wales was successfully protestantized, Ireland was not. Here, a defense of 'old traditions' in the political realm combined with a Catholic identity newly invigorated by Tridentine missionaries to serve as rallying points for resistance against encroachment by the Protestant English state and church. Scotland rebelled against England in the middle of the seventeenth century when it feared that its Calvinist consensus was under threat by the English crown and its state church.

In France, the spread of Calvinist ideas in the sixteenth century led to the establishment of a powerful Protestant (Huguenot) movement, both in political and social terms. The resulting confessional split in French society led to a civil war between 1562 and 1589. With the Edict of Nantes of 1598, King Henri IV (who had converted to Catholicism and to whom the sentence 'Paris is well worth a Mass' has been ascribed) enforced a legal settlement which ensured Catholic-Huguenot coexistence in France. However, King Louis XIV's revocation of the Edict of Nantes in 1685 attests to the fact that the ideal of a kingdom unified in religion (*un roi, une foi, une loi* – one king, one faith, one law) was still very much alive in the late seventeenth century.

And finally, two territories of the Habsburg dynasty, the Netherlands and Bohemia, are striking examples of how the connection between state formation and confessionalization could provoke massive resistance and lead to war. In both dependent territories, oppositional movements of the estates, which originally resisted centralization and state formation, became intertwined with confessional resistance against (re-) catholicization. In the case of the Netherlands, this resulted in the Dutch Revolt and an eighty years' war with Spain. In the end, the Dutch Republic became an independent state with the Calvinist church as its 'public' (i.e., official) church. In the case of Bohemia, the defeat of the Calvinist Elector of the Palatinate and short-term King of Bohemia Frederick V in the Battle of the White Mountain in 1620 resulted in the replacement of the majority of the Bohemian elite and a thorough process of recatholization of the territory.

Only in Switzerland and in the German Empire did the dualistic political structure, which prevailed in both countries, lead to the long-term establishment of multiconfessional political entities. While the existence of more than one confession was legally established in the larger political entities, most of the smaller political units (territories in the Empire, cantons in Switzerland) were mono-confessional. While the Reformation and the confessional age had lead to the division of Latin Christianity into multiple confessions, the ideal of the unity of church and state continued to prevail in Europe until at least the end of the seventeenth century.

See also: Historical Thought and Historiography: Christian Tradition; Modernization and Modernity in History; Political Religions: History of the Concept; Religion and Economic Life; Religion, History of; Religion, Sociology of.

Bibliography

Archiv für Reformationsgeschichte, 2003. Internationale Zeitschrift zur Erforschung der Reformation und ihrer Weltwirkungen/Archive for Reformation History: An international journal concerned with the history of the reformation and its significance in world affairs, 1 (1903/1904), esp. Focal point/Themenschwerpunkt: confessionalization and social discipline in France, Italy, and Spain. Archiv für Reformationsgeschichte 94, 276–319.

Bamji, A., Janssen, G.H., Laven, M. (Eds.), 2013. The Ashgate Research Companion to the Counter-Reformation. Ashgate, Aldershot.

Brady, T.A., 1978. Ruling Class, Regime and Reformation at Strasbourg, 1520–1555. Brill, Leiden.

Brady, T.A., Oberman, H.A., Tracy, J.D. (Eds.), 1994. Handbook of European History, 1400–1600: Late Middle Ages, Renaissance, and Reformation. Structures and Assertions, vol. 1. E.J. Brill, Leiden.

Brady, T.A., Oberman, H.A., Tracy, J.D. (Eds.), 1995. Handbook of European History, 1400–1600: Late Middle Ages, Renaissance, and Reformation. Visions, Programs, and Outcomes, vol. 2. E.J. Brill, Leiden.

Brady, T.A., 2009. German Histories in the Age of Reformations, 1400–1650. Cambridge University Press, Cambridge.

Dickens, A.G., 1976. The German Nation and Martin Luther. Fontana, London.

Ehrenpreis, S., Lotz-Heumann, U., 2008. Reformation und konfessionelles Zeitalter. Kontroversen um die Geschichte, second ed. Wissenschaftliche Buchgesellschaft, Darmstadt.

Elliott, J.H., 1992. A Europe of composite monarchies. Past and Present 137, 48–71.

Forster, M.R., 1992. The Counter-Reformation in the Villages: Religion and Reform in the Bishopric of Speyer, 1560–1720. Cornell University Press, Ithaca, NY.

Forster, M.R., 1998. With and without confessionalization: varieties of early modern German Catholicism. Journal of Early Modern History 1, 315–343.

Gerhard, D., 1962. Alte und Neue Welt in vergleichender Geschichtsbetrachtung. Vandenhoeck & Ruprecht, Göttingen.

Greyerz, K.v., 1980. The Late City Reformation in Germany: The Case of Colmar, 1522–1628. Steiner, Wiesbaden.

Hamm, B., 1993. Von der spätmittelalterlichen reformatio zur Reformation: Der Prozeß normativer Zentrierung von Religion und Gesellschaft in Deutschland. Archiv für Reformationsgeschichte 84, 7–82.

Headley, J.M., Hillerbrand, H.J., Papalas, A.J. (Eds.), 2004. Confessionalization in Europe, 1555–1700: Essays in Memory of Bodo Nischan. Ashgate, Aldershot.

Jedin, H., 1999. Catholic reformation or counter-reformation? In: Luebke, D.M. (Ed.), The Counter-Reformation: The Essential Readings. Blackwell, Oxford, pp. 19–45.

Karant-Nunn, S.C., 1997. The Reformation of Ritual: An Interpretation of Early Modern Germany. Routledge, London, New York.

Karant-Nunn, S.C., 2010. The Reformation of Feeling: Shaping the Religious Emotions in Early Modern Germany. Oxford University Press, New York.

Lotz-Heumann, U., 2004. Confessionalization. In: Hillerbrand, H.J. (Ed.), The Encyclopedia of Protestantism, vol. 1. Routledge, New York, London, pp. 497–501.

Lotz-Heumann, U., Ehrenpreis, S., 4 April 2005. The Concept of Confessionalization as a Research Tool. H-German. Forum: Confessionalization. http://h-net.msu.edu/cgi-bin/logbrowse.pl?trx=vx&list=H-German&month=0504&week=a&msg=OBVg%2bAebgXhn7SqJrbkRFA&user=&pw (accessed 31.05.14.).

Lotz-Heumann, U., 2006. Die deutsche Reformation und die Konfessionalität in Europa. In: Schilling, H., Heun, W., Götzmann, J. (Eds.), Altes Reich und neue Staaten 1495 bis 1806. Sandstein Verlag, Dresden, pp. 145–157.

Lotz-Heumann, U., Pohlig, M., 2007. Confessionalization and literature in the empire, 1555–1700. Central European History 40, 35–59.

Lotz-Heumann, U., 2008. Confessionalization. In: Whitford, D. (Ed.), Reformation and Early Modern Europe: A Guide to Research. Truman State University Press, Kirksville, MO, pp. 136–157.

Lotz-Heumann, U., 2010. Reformation und Konfessionalisierung – Europa. In: Demel, W. (Ed.), WBG Weltgeschichte, Entdeckungen und neue Ordnungen (1200 bis 1800), vol. IV. Wissenschaftliche Buchgesellschaft, Darmstadt, pp. 296–324.

Lotz-Heumann, U., 2013. Confessionalization. In: Bamji, A., Janssen, G.H., Laven, M. (Eds.), The Ashgate Research Companion to the Counter-Reformation. Ashgate, Aldershot, pp. 33–53.

Maltby, W.S. (Ed.), 1992. Reformation Europe: A Guide to Research II. Center for Reformation Research, St. Louis.

Moeller, B., 1987. Reichsstadt und Reformation, second ed. Evangelische Verlags-Anstalt, Berlin.

Oberman, H., 2001. The long fifteenth century: in search of its profile. In: Brady, T.A. (Ed.), Die deutsche Reformation zwischen Spätmittelalter und Früher Neuzeit. Oldenbourg, Munich, pp. 1–18.

O'Malley, J.W. (Ed.), 1988. Catholicism in Early Modern History: A Guide to Research. Center for Reformation Research, St. Louis.

O'Malley, J.W., 2000. Trent and All that: Renaming Catholicism in the Early Modern Era. Harvard University Press, Cambridge, MA.

Ozment, S. (Ed.), 1982. Reformation Europe: A Guide to Research. Center for Reformation Research, St. Louis.

Ranke, L.v., 1839–47. Deutsche Geschichte im Zeitalter der Reformation. Duncker & Humblot, Berlin.

Reinhard, W., 1995. Was ist katholische Konfessionalisierung? In: Reinhard, W., Schilling, H. (Eds.), Die katholische Konfessionalisierung. Gütersloher Verlagshaus, Gütersloh, pp. 419–452.

Reinhard, W., 1999. Pressures towards confessionalization? Prolegomena to a theory of the confessional age. In: Dixon, C.S. (Ed.), The German Reformation: The Essential Readings. Blackwell, Oxford, pp. 169–192.

Reinhard, W., Schilling, H., 1995. Die katholische Konfessionalisierung. Gütersloher Verlagshaus, Gütersloh.

Reinhard, W., 1999. Reformation, counter-reformation, and the early modern state: a reassessment. In: Luebke, D.M. (Ed.), The Counter-Reformation: The Essential Readings. Blackwell, Oxford.

Roper, L., 1989. The Holy Household: Women and Morals in Reformation Augsburg. Clarendon Press, Oxford.

Rublack, H.-C. (Ed.), 1992. Die lutherische Konfessionalisierung in Deutschland. Gütersloher Verlagshaus, Gütersloh.

Schilling, H. (Ed.), 1986. Die reformierte Konfessionalisierung in Deutschland – Das Problem der "Zweiten Reformation". Gütersloher Verlagshaus, Gütersloh.

Schilling, H., 1992. Confessionalization in the empire: religious and societal change in Germany between 1555 and 1620. In: Schilling, H. (Ed.), Religion, Political Culture and the Emergence of Early Modern Society: Essays in German and Dutch History. Brill, Leiden, pp. 205–245.

Schilling, H., 1998. Reformation – Umbruch oder Gipfelpunkt eines Temps des Réformes? In: Moeller, B. (Ed.), Die frühe Reformation in Deutschland als Umbruch. Gütersloher Verlagshaus, Gütersloh, pp. 13–34.

Schmidt, H.R., 1992. Konfessionalisierung im 16. Jahrhundert. Oldenbourg, Munich.

Scribner, R.W., 1981. For the Sake of Simple Folk: Popular Propaganda for the German Reformation. Cambridge University Press, Cambridge, England.

Spicer, A., 2012. Lutheran Churches in Early Modern Europe. Ashgate, Farnham, Surrey, England.

Troeltsch, E., 1906. Die Bedeutung des Protestantismus für die Entstehung der modernen Welt. Historische Zeitschrift 97, 1–66.

Whitford, D. (Ed.), 2008. Reformation and Early Modern Europe: A Guide to Research. Truman State University Press, Kirksville, MO.

Wiesner, M.E., 1993. The reformation of the women. In: Guggisberg, H.R., Krodel, G.G. (Eds.), Die Reformation in Deutschland und Europa: Interpretationen und Debatten. Gütersloher Verlagshaus, Gütersloh, pp. 193–208.

Zeeden, E.W., 1958. Grundlagen und Wege der Konfessionsbildung im Zeitalter der Glaubenskämpfe. Historische Zeitschrift 185, 249–299.

Relevant Website

http://www.h-net.org/~german/discuss/Confessionalization/Confess_index.htm – Forum "Confessionalization" H-German, April 2005.

Refugees in Anthropology

Marita Eastmond, University of Gothenburg, Göteborg, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article outlines the development of 'refugees' as a legal category as well as an academic field of study, in response to global refugee crises in the twentieth century. It also presents some key concepts and themes of research, and demonstrates the particular relevance of anthropology for the study of refugees, including some of the challenges of global movement to the discipline and its ethnographic method.

Refugees as a Field of Study

People forced to leave their homes to settle in new environments is not a new phenomenon. However, 'refugees' as a legal category and as the object of academic research are more recent. This article will outline the development of refugee studies in response to global refugee crises in the twentieth century. It will also present some key concepts and themes of research and demonstrate the particular relevance of anthropology to this field.

The study of refugees emerged in the 1980s as an academic field in its own right, with its own research centers, associations, academic degrees and international journals. Until then, international migration had been the main academic context for research on refugees and displaced people. The establishment of refugee studies was related to the global refugee crisis at the time. By 1981, the escalation of violent conflicts and mass disruptions in many parts of the globe had dramatically doubled the global refugee population in three years. New increases followed the political changes at the end of the Cold War, with large numbers of refugees from upheavals in postcolonial states and from crumbling Communist federations in Europe dominating asylum claims in the 1990s. In 2001, the pivotal events of September 11 and the subsequent wars on terror produced mass refugee movements, notably from Iraq and Afghanistan, while concerns with 'securitizing' borders, especially in northern states, have tended to criminalize and sanction asylum seekers. The numbers of refugees and asylum seekers remain high at the end of 2010, but given the increasingly restrictive migration regime in the north, people in need of safety are increasingly seen to be making their way to other destinations, notably emergent middle-income countries in Africa, Asia, and Latin America (UNHCR, 2011; Van Hear, 2011).

Since its inception, refugee studies have been multidisciplinary, drawing together a range of research from the social and behavioral sciences, such as political science and peace and conflict studies, medicine and psychiatry, and always with a strong presence of international law and anthropology. Refugees have also become an established research topic in recent years in anthropology's own subdisciplines, notably development anthropology, medical anthropology, legal anthropology, and the anthropology of policy. With its clear relevance for the acute issues of many refugee situations, anthropology has engaged with practical application and advocacy, often in cross-disciplinary teams. Its specific contribution in these contexts has been to elicit and promote refugees' own voices, through ethnographically informed analyses of refugees' predicaments and perspectives, in relation to the group's specific history and context of displacement. In the past two decades, however, policy and practice of states and their bureaucracies in dealing with refugees have themselves come under anthropological scrutiny. Moreover, researching dispersed and vulnerable populations across different sites have meant great challenges to studies relying on ethnographic methods. The multidisciplinary character of this field, as well as the increasingly complex geopolitical dynamics and changing migration regimes, spawns ongoing debates about refugees as an object of study.

Who Is a Refugee? Legal Definitions and Academic Concepts

The historical contexts in which refugees emerged as an internationally recognized category were the massive displacements in Europe following two world wars. The Geneva Convention of 1951 is the critical landmark of the establishment of institutions (such as the UNHCR) and legal instruments to protect and assist such populations still valid today. A *convention refugee* is an individual who has crossed a national boundary, and has a well-founded fear of persecution in the country of origin on the basis of race, religious or political affiliation, ethnic or social group. Subsequently, other, regionally based legal conventions, notably in Africa and Latin America, have adopted more inclusive definitions of refugees which better correspond to the manifold types of mass displacements that often follow violent conflicts and other social disruptions in their regions. Also many other states today recognize *people in refugee-like circumstances*, i.e., persons who are outside their country of origin and face protection risks similar to those of convention refugees; although not personally persecuted, they are victims of war or other social disruptions. Categories outside of legal provisions but equally of concern to agencies of protection, as well as to research, are *internally displaced people* (IDPs), those who do not cross an international border but flee within their own countries. Of concern are also the many refugees and IDPs who voluntarily return to their places of origin.

These legal categories notwithstanding, the answer to the question 'Who is a refugee?' remains a crucial one in today's world. Legal categories are subject to the interpretations made

by receiving states and their policies of admission may differ in response to foreign policy interests, public opinion and perceived 'load' of new arrivals. For those being classified, it may be a matter of life or death, because of the exclusive rights and entitlements of humanitarian concern and international protection that come with refugee status.

In the academic field, the question is also a continuing concern as a matter of theoretical and conceptual delineation of its object of study. 'Refugees' as defined in the Geneva convention and based on individual persecution is too narrow, many argue, both as a protection instrument and as a theoretical concept and does not take account of the complexity of many situations which force people to leave their homes and in which violent conflict and insupportable living conditions, the political and the economic, are often interlinked. Thus, according to Van Hear (2011), 'forced migration' would be more useful as an analytical concept in social science rather than 'refugee studies,' as it takes account of the complexity of motives of moving and also recognizes the need to think of refugees as social actors exercising agency and making choices, however, pressing the circumstances. Both terms may be distinguished from what is meant by 'refugee' in public discourse, or emically, as defined by refugees themselves. Others underline the importance of maintaining that which is specific to refugees, in terms of human rights and other humanitarian provisions made in their name (Shami, 1994) and legal researchers are wary that a shift from the study of refugees to the more inclusive one of 'forced migration' will dilute their protection and thus respond to the interests of northern states (Hathaway, 2007, Chimni, 2009 in Van Hear, 2011).

The burning issue of who and what a refugee is can also be addressed by examining the powerful institutional contexts and discourses, national and international, in which the category of 'refugee' is construed and public policy shaped. Important inspiration from anthropology has been drawn from theories concerned with the politics of space and otherness (Gupta and Ferguson, 1992, 1997) and in particular from Liisa Malkki's work which opened up for conceptualizing the anthropological study of displacement in new ways. According to Malkki, the category of 'refugee,' emerging from policy frameworks grounded in a particular history and view of the nation-state order, had been uncritically adopted into research as well as practice. The 'refugee experience,' following this logic, was defined by loss and uprooting, and depicted as an undifferentiated and essentialized condition for which repatriation was the best solution (1995a). In a world where the nation state is seen as the 'natural' basis of belonging, Malkki argued, refugees become an anomaly that challenges the dominant 'national order of things' (Malkki, 1995b). Eschewing such essentializing notions, anthropological research today is more likely to examine not only the political and historical circumstances that force particular people to move but also the meanings they themselves attribute to their displacement.

The Dynamics of Displacement

Forced migration then is inevitably related to change, often acute and disruptive, and any theoretical approach to refugees necessarily engages with process. In some cases, displacement affects most aspects of life and addresses core issues in the anthropological study of communities, such as subsistence and livelihood, kinship and family, social and political networks as well as the cultural meanings which give a sense of coherence and predictability to everyday life. For those displaced, change is often multifaceted and compressed in time, thus also providing a challenge to an older generation of anthropologists more accustomed in the past to studying slower and more cumulative processes of change.

With a shift in recent years to an emphasis on the workings of power in contexts of reception, as well as the agency of refugees themselves, anthropologists have taken a greater interest in uncovering the perspectives and strategies through which displaced groups deal with disruptive change, renegotiate livelihoods, social relations, and identities. A growing body of research in the anthropology of war and violent conflict, population movement, and reconstruction is contributing important theoretical approaches to understanding the interplay between concerns and strategies of people on the move with processes at global and state levels (Jansen and Löfving, 2008).

A Diversity of Refugee Situations and Research Themes

However, these processes unfold differently in different places. The great diversity of conditions, in particular the contrast between southern and northern host states, has meant the development of different emphases in the concerns and theoretical approaches applied to the problem of refugees in different parts of the world. Differences reflect the global distribution and politics in terms of admitting refugees: Today, as before, three-fourths of the world's refugees reside in a country neighboring their own. Only a trickle make it to the more affluent receiving countries of the north, in Europe or in North America, and even fewer are allowed to settle there.

Thus, in *southern* countries, particularly in Africa and Asia, displacement and refugees have often been studied as a problem of assistance and development (Bakewell, 2012). A particular theme concerns the consequences of development schemes and modernization, forcibly displacing large numbers of people within their countries. An early contribution is Elizabeth Colson's classical ethnography of the social and cultural consequences of the relocation of the Gwembe Tonga due to the building of the Kariba hydroelectric dam in the Zambezi River Basin in 1958 (Colson, 1971). Another major research theme in southern contexts has been a concern with *humanitarian assistance* to refugees, and the potential of such programs to have more long-term development effects that also benefit the country and the local communities that host them. One of the first studies of this dynamic interaction between different actors and agencies was Harrell-Bond's illuminating ethnographic study (1986) analyzing the implications of relief operations on refugees and their hosts in Uganda. Neglecting refugees' own concerns and resources, the interventions of relief agencies were found to create conflicts and fragmentation both within refugee collectivities and with the local resident population. If and how displaced people define themselves is also a question of how they perceive their prospects. Malkki's

ethnography (1995b) of displaced Hutu refugees from Burundi settled in Tanzania compares identity and historical memory of refugees in camps to those who have settled in a nearby town, relying on their own creative strategies for survival. Whereas these were pragmatically juggling social identities, in the diversity typical of many African towns, camp refugees were passionately concerned with creating their mythico-history as members of the Hutu nation. The strength of ethnographic fieldwork is that it allows 'a critical examination of camps as organized solutions to refugee crises,' while also dispelling the one-sided image of refugees as dependent and passive victims. An example of this more complex picture is Cindy Horst's analysis of how Somalis cope with refugee life in Kenyan camps. Emphasizing agency rather than victimhood, she describes the variety of social and cultural resources in dealing with camp life; these include a high degree of mobility and strong social networks that stretch far beyond the confines of the camp (2006).

Most of the world's refugees are in protracted situations, reflecting prolonged conditions of violence and instability in their countries of origin. For the many who move back and forth across borders or inside their country in response to new outbreaks, distinctions between flight and homecoming, or refugees and internally displaced lose their meaning (Allen, 1996). Others get stuck in what is often referred to as refugee warehousing, involving the confinement in camps or segregated settlements in remote or near-border areas (USCRI, 2009). Such confinement may last for years (e.g., Afghans in Iran and Pakistan) or even generations (Palestinians in Lebanon, Gaza, the West Bank).

'Going home' has been promoted in international policy as the most durable solution to refugee problems and the naturally desired end of 'the refugee cycle' (Black and Koser, 1999). The return of refugees and IDPs has also become a strong priority of international peace building efforts in war-torn societies in recent decades. The international repatriation discourse has often portrayed return within an essentialized conception of identities as rooted – as people coming back to their 'natural' homes, often predicated on an idealized notion of prewar local life. But rather than assuming rootedness, anthropologists have emphasized the need to see attachment to place as contingent on wider social and historical processes through which people lay claim to a place and call it home (Gupta and Fergusson, 1997). This has constituted an important theoretical framework for the study of the repatriation of refugees, as seen for instance in ethnographies of return (Stefansson, 2006) as well as in studies of the creative ways in which new places are made into home by returnees (Hammond, 2004). Studies in different parts of the world attest to the fact that images of home in exile cannot always be reconciled with the realities of life upon return, and that reemplacement requires negotiating one's entry in a new context of power and difference (Allen and Morsink, 1994; Black et al., 2006; Kamal, 2010; McDowell and Eastmond, 2002). In such conditions, repatriation is almost always a 'struggle for home' (Jansen and Löfving, 2008).

In *northern* contexts, the resettlement and integration of refugees have been predominant research themes, at least until the 1990s. Current frameworks have moved away from more of these conventional community studies that focused on a particular group and its sociocultural integration, often based on the idea of 'culture' as shared tradition (O'Neal, 1999). Subsequently, more processual perspectives have underlined the ways in which 'culture' and 'ethnic identity' are negotiated or reinvented, and the meanings that groups attribute to their exile (DeVoe, 1992). In the conventional migration paradigm, continuing attachments to the country of origin were often interpreted as poor social integration, a notion subsequently dispelled by the transnational frameworks of migration emerging in the early 1990s.

Current research, in anthropology as in other social sciences, reflects a growing concern with the increasingly restrictive refugee regime in northern states, particularly in the EU. The measures undertaken to manage migration include policies to control and deter 'flows' from the outside, for instance by exporting asylum processing or confinement of applicants to the margins of EU, such as Malta (DeBono, 2011) or, by agreement, to transit or source countries outside of EU (Betts and Milner, 2006). Anthropologists' new interest in borders and legality includes exploring states' regulation of movement (Khosravi, 2011). Given that it is nearly impossible for claimants to meet the demands for legal entry, the restrictions give rise to new categories, such as irregular, undocumented, or illegal migrants, thus tending to criminalize asylum seekers and legitimate their forced removal. With respect to the conceptualizations made in the policy field, Turton (2003) reminds us that categories and images are powerful means of constituting, not merely reflecting, reality. The terms by which we describe forced migrants are not innocent but are part of how we interpret and therefore act toward those who are displaced. Zetter's (2007) article on refugee labeling in the context of public policy practices illustrates this process of the making of bureaucratic categories, and the consequences for those categorized, and offers insights into the impacts of institutional and bureaucratic power on the lives of refugees. For this reason, anthropologists' study of public policy, its role, and consequences in contemporary society, including the managing of migration, is a promising line of inquiry (Shore et al., 2011). Public policy research entails scrutinizing the assumptions of public policy 'as a legal-rational way of getting things done' and illuminating the constellation of actors, activities, and influence that shape policy, how they are implemented, and what their consequences are for those concerned (Wedel et al., 2005: 30). In a similar vein, recent ethnographic studies address the procedures of managing unwanted migrants, the discourses that underpin these procedures, and the responses of those they address. Khosravi (2009) has examined a detention center in Sweden and how officials formulate their task in terms of motivating and assisting detainees in returning. Griffiths (2012), doing fieldwork in a detention center in the UK, analyses how failed asylum seekers awaiting deportation negotiate discourses of identification and criminalization. Whyte's (2011) ethnography of life at a Danish reception center describes the techniques of surveillance of asylum-seeker residents as a kind of myopticon, and the strategies of evasion and resistance of the residents. Along similar lines, new interest comes from legal anthropology, directing attention to the legal and bureaucratic systems and discourses through which asylum claims are processed (Good, 2007). The system also enlists anthropologists, in some EU countries at least, in a new role as expert witnesses in asylum courts (Good and Gibb, in press).

Refugee Children

An increasing number of claimants in recent years have been unaccompanied children, three-fourths of whom make their way to Europe, often after a long and hazardous passage from countries of origin in Asia or Africa (Watters, 2008; Kohli, 2006). Reflecting the culture of mistrust of contemporary European reception systems, two conflicting images of these children, similar to those used to depict adult asylum seekers, have emerged: the vulnerable victim or the bogus untruthful opportunist (Giner, 2007; Griffiths, 2012). Age determination procedures, notoriously unreliable, to decide whether applicants qualify as children reflect a reluctant and mistrusting system. The public controversy in Sweden over asylum-seeking children in the early 2000s, which engaged the public and nearly toppled the government, is a case in point. Asylum-seeking children who feared deportation developed a condition of severe apathy which took epidemic proportions but for which there was no clear-cut medical diagnosis. In the public debate and the contrasting images it spawned, children were either seen as victims of the state rejecting them or, as state experts maintained, malingerers seeking permanent residence in Sweden. Eventually, the government introduced a temporary law through which most of the children and their families were eventually granted residence status (Eastmond and Ascher, 2011; Hacking, 2010). The well-being of refugee children has drawn attention from a range of disciplines, not least psychology and pediatrics. Ethnographies of children and youth in camps in southern countries (Atkinson, 2006; Chatty, 2010) or received in the north (Watters, 2008) examine children's experiences and needs, as well as the interventions made on children's behalf in terms of health and social care. Theoretical concerns are notions of childhood and children's agency; other concerns are the different, more experiential, methods required when working with children.

Health problems of refugee children as well as adults have been addressed by medical anthropologists (Watters and Ingelby, 2004). Anthropologists caution against identifying disorders such as traumatization and posttraumatic stress disorder (PTSD) across cultures and call for ethnographically grounded emic understandings of such responses. While diagnostic labels may be convenient categories for clinicians, they may pathologize refugees and their responses to violence (Eastmond, 2005; Parker, 1996). The issue forms part of a broader inquiry into the political and moral issues of reformulating injustice and violence as suffering and trauma (Fassin and Rechtman, 2009).

Refugee Diasporas

Reflecting increasing globalization, and thus eclipsing the north–south distinction, much research is today directed at diasporas, the global networks of refugees and migrants and their transnational practices. Diaspora was introduced as a useful concept in anthropology in the 1990s (Safran, 1991; Clifford, 1994); it referred to populations dispersed by force, maintaining over long-time relations to, or at least a dream of, homeland and return. Today, the term encompasses different categories of migrants (including 'ethnic communities') who retain ties to people in other parts of the world and engage in diasporic practices (Brubaker, 2005). Diasporas and their economic resources today have recognized importance for the global economy, and 'diaspora' has been adopted as a policy term used by governments and development agencies (as well as in migrants' own discourse) to harness the potential that such networks represent (Van Hear, 2011). At the level of the household, remittances have long played a vital role in patterns of support to people 'back home,' including enabling other family members to migrate. Transnational migration and diaspora have thus reshaped the traditional paradigm of migration as a one-way, permanent move. While 'permanent integration in another country' has long been defined by the UNHCR as the second best solution for refugees when return is not possible, such distinctions are no longer that clear. For refugees themselves, transnational lives can resolve the conflict between 'here' and 'there' or between the ideological commitment to return and the practical realities of everyday life (Eastmond, 2006; Fuglerud, 1999).

Refugees: Ethnographic Challenges

Ethnographically, anthropologists would ideally follow the refugees, the routes they take and the connections they make. Such observational strategies are rarely feasibly except in auto-ethnographies (Khosravi, 2011). In most cases, the focus is on one or two of the local contexts in which refugees' lives are embedded, usually the country of asylum and that of origin, and possibly a relevant selection of sites in their global network. Given this dispersal of places and relations in a global field, the greatest challenge is one of access and depth. Still, multisited ethnography is not that novel and much can be learnt from drawing upon research into many other globalized phenomena besides migration (Hannerz, 2003). Some refugee sites are more inaccessible than others, however; camps or detention centers may curtail observation and limit researchers' access to informants. Others may be bureaucratic worlds and legal procedures that may seem as foreign and confusing to the refugee as to the ethnographer (Hannerz, 2003). Such fields may thus entail entering unfamiliar territory 'at home' and require interaction with immigration officers, deportation escorts, and other officials of the state, or appearing as expert witness in asylum courts (Good, 2007). Access and understanding can be difficult also in other ways. To elicit narratives of torture, flight or detention, or other experiences of loss and personal pain require sensitive approaches and strong ethical awareness, especially when working with unaccompanied children (Kohli, 2006).

The Future of Refugee Studies

From having been rather marginal as a research area until the 1980s (in interesting ways reflecting the position of displaced populations themselves), the anthropology of refugees and forced migration is today an established field of research. Linked to and accentuating globalization and global geopolitics, refugees have presented a challenge not only to nation states, but also to the social sciences. Increased mobility and

transnational forms of social organization have required anthropology in particular to reexamine its concepts of culture and community as bounded and territorialized units. Refugee studies have been instrumental in this conceptual move away from the 'people and places' to a greater focus on the processes of making and maintaining connections between them.

While the study of refugees shares some common ground with international migration, it will have to continue addressing the circumstances of force that produce refugees as well as the politics of reception in the context of tightening refugee regimes. With theoretical development in recent decades in the anthropology of conflict, violence, and social memory, globalization and diasporas, security and human rights, the study of refugees and exile as transnational phenomena has gained new relevance and provided fresh theoretical insights. These issues will increasingly engage and infuse new ideas in the domains of moral and political anthropology, including a waxing interest in legal anthropology and in studying public policy 'at home.' The broadening of areas of anthropological inquiry and the increasing relevance of core anthropological concerns signal the coming of age of the study of refugees as an academic field. If refugees are symptomatic of the state of the world, with global mobility, fluidity, and fragmentation, and increasingly global networks and multiple allegiances as well as claims to particular lands and cultural identities, then refugees will continue to be a central area of study in anthropology.

See also: Diaspora; Hazards and Disaster Research in Contemporary Anthropology; Human Rights and Social Work; Immigration, Citizenship, and Integration: Social Work Connections; Multiculturalism.

Bibliography

Allen, T., Morsink, H., 1994. When Refugees Go Home: African Experiences. James Currey, London.

Allen, Tim (Ed.), 1996. War, Flight & Homecoming in Northeastern Africa. James Currey Ltd, London.

Atkinson, L., 2006. From play to knowledge: from visual to verbal? Anthropology Matters 8 (2). ISSN: 17586453 www.anthropologymatters.com.

Bakewell, O., 2012. Migration and Development. Edward Elgar Publishing.

Betts, A., Milner, J., 2006. The Externalisation of EU Asylum Policy: The Position of African States. Working Paper 36. Centre on Migration, Policy and Society, Oxford.

Black, R., Eastmond, M., Ghent, S. (Eds.), 2006. Sustainable Return in the Balkans. International Migration 44 (3), 5–13.

Black, R., Koser, K., 1999. The End of the Refugee Cycle? Refugee Repatriation and Reconstruction. Berghahn Books, Oxford.

Brubaker, R., 2005. The 'diaspora' diaspora. Ethnic and Racial Studies 28 (1), 1–19.

Chatty, D. (Ed.), 2010. Deterritorialized Youth: Afghan and Sahrawi Refugee Young People at the Margins of the Middle East. Berghahn Press, Oxford and New York.

Clifford, J., 1994. Diasporas. Cultural Anthropology 9 (3), 302–338.

Colson, E., 1971. The Social Consequences of Resettlement. Manchester University Press, Manchester.

DeBono, D., 2011. 'Human rights for the Maltese first': irregular migration and human rights in Malta. In: Xuereb, P. (Ed.), Irregular Migration: Rights and Realities. Malta European Studies Association, Malta.

DeVoe, P., 1992. Selected Papers on Refugee Issues. Committtee on Refugee Issues (CORI) American Anthropological Association, Washington.

Eastmond, M., 2005. The disorders of displacement. Bosnian refugees and the re-constitution of normality. In: Jenkins, R., Steffen, V., Jessen, H. (Eds.), Managing Uncertainty: Ethnographic Studies of Illness, Risk and the Struggle for Control. Museum Tusculanum Press, Copenhagen, pp. 149–171.

Eastmond, M., 2006. Beyond exile: Refugee strategies in transnational contexts. In: Crepeau, F., Nakache, D., Collyer, M., Goetz, N.H., Hansen, A. (Eds.), Forced Migration and Global Processes Lexington. Lexington Press, Lanham, pp. 217–236.

Eastmond, M., Ascher, H., 2011. In the best interest of the child? The politics of vulnerability and negotiations for asylum in Sweden. Journal of Ethnic and Migration Studies 37 (8), 1185–1200.

Fassin, D., Rechtman, R., 2009. Empire of Trauma: An Inquiry into the Condition of Victimhood. Princeton University Press, Princeton.

Fuglerud, O., 1999. Life on the Outside. The Tamil Diaspora and Long Distance Nationalism. Pluto Press, London.

Gibb, R., Good, A. Do the facts speak for themselves? Country of origin information in French and British refugee status determination procedures. International Journal of Refugee Law, ISSN 0953-8186, in press.

Giner, C., 2007. The politics of childhood and asylum in the UK. Children & Society 21 (4), 249–260.

Good, A., 2007. Anthropology and Expertise in the Asylum Courts. Routledge-Cavendish, Abingdon.

Griffiths, M., 2012. Anonymous aliens? Questions of identification in the detention and deportation of failed asylum seekers, population. Space and Place 18 (6), 715–727.

Gupta, A., Ferguson, J., 1992. Beyond "culture": space, identity and the politics of difference. Cultural Anthropology 7 (1), 6–23.

Gupta, A., Ferguson, J., 1997. Culture, Power, Place: Explorations in Critical Anthropology. Duke University Press, Durham.

Hacking, I., 2010. Pathological withdrawal of refugee children seeking asylum in Sweden studies. History and Philosophy of Science Part C 41 (4), 309–317.

Hammond, L., 2004. This Place Will Become Home. Refuge Repatriation to Ethiopia. Cornell University Press, Ithaca, New York.

Hannerz, U., 2003. Being there… and there… and there!: reflections on multi-site ethnography. Ethnography 4 (21), 201–216.

Harrel-Bond, B.E., 1986. Imposing Aid: Emergency Assistance to Refugees. Oxford University Press, Oxford.

Hathaway, J., 2007. Forced Migration Studies: Could we Agree Just to 'Date'? Journal of Refugee Studies 20, 349–369, 385–390.

Horst, C., 2006. Transnational Nomads. How Somalis cope with refugee life in Dadaab camps in Kenya. Berghahn, Oxford.

Jansen, S., Löfving, S. (Eds.), 2008. Struggles for Home. Violence, Hope and the Movement of People. Berghahn, Oxford.

Kamal, S., 2010. Repatriation and reconstruction; Afghan youth as a 'burnt generation' in post-conflict return. In: Chatty, D., Finlayson, B. (Eds.), Dispossession and Displacement. Forced Migration in the Middle East and North Africa. Oxford University Press, Oxford.

Khosravi, S., 2009. Sweden: detention and deportation of asylum seekers. Race and Class 50 (4), 38–56.

Khosravi, S., 2011. Illegal Traveller. An Auto-Ethnography of Borders. Palgrave Macmillan, Basingstoke.

Kohli, R., 2006. The sound of silence: listening to what unaccompanied asylum-seeking children say and do not say. British Journal of Social Work 36 (5), 707–721.

Malkki, L., 1995a. Refugees and exile: from "refugee studies" to the national order of things. Annual Review of Anthropology 24, 495–523.

Malkki, L., 1995b. The Purity of Exile. Violence, Memory and National Cosmology Among Hutu Refugees in Tanzania. The University of Chicago Press, Chicago and London.

McDowell, D., Eastmond, M., 2002. Transitions, state-building and the 'residual' refugee problem: the East Timor and Cambodian repatriation experience. Australian Journal of Human Rights 8 (1).

O'Neall, C., 1999. Possibilities for migration anthropology. American Ethnologist 26 (1), 221–225.

Parker, M., 1996. Social devastation and mental health in northeast Africa: some reflections on an absent literature. In: Allen, T. (Ed.), In Search of Cool Ground: War, Flight and Homecoming in Northeast Africa. James Currey, London, pp. 262–273.

Safran, W., 1991. Diasporas in modern sciences: myths of homeland and return. Diaspora 1 (1), 83–99.

Shami, S., 1994. Population Displacement and Resettlement. Development and Conflict in The Middle East. Center for Migration Studies, New York.

Shore, C., Wright, S., Pero, D. (Eds.), 2011. Policy Worlds: Anthropology and the Analysis of Contemporary Power. Berghahn, Oxford.

Stefansson, A., 2006. Homes in the making: property restitution, refugee return and senses of belonging in a post-war Bosnian Town. International Migration 44 (3), 115–139.

Turton, D., 2003. Conceptualising Forced Migration. Working Paper No 12, Refugee Studies Centre, Oxford.

USCRI, 2009. World Refugee Survey: Refugee Warehousing. www.refugees.org.

UNHCR, 2011. Global Trends 2011. United Nations High Commissioner for Refugees, Geneva.

Van Hear, N., 2011. Forcing the issue: migration crises and the uneasy dialogue between refugee research and policy. Journal of Refugee Studies 25 (1), 124.

Watters, C., 2008. Refugee Children: Towards the Next Horizon. Routledge, London.

Watters, C., Ingelby, D., 2004. Locations of care: meeting the mental health and social care needs of refugees in Europe. International Journal of Law and Psychiatry 27, 549–570.

Wedel, J., Shore, C., Feldman, G., 2005. Toward an anthropology of public policy. Annals of the American Academy of Political and Social Sciences 600 (1), 30–51.

Whyte, Z., 2011. Enter the myopticon: uncertain surveillance in the Danish asylum system. Anthropology Today 27 (3), 18–21.

Zetter, R., 2007. More labels, fewer refugees: remaking the refugee label in an era of globalization. Journal of Refugee Studies 20 (2), 172–192.

Refugees: Geographical Aspects

Prem Kumar Rajaram, Central European University, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Bascom, volume 19, pp. 12895–12901, © 2001, Elsevier Ltd.

Abstract

'Refugee' is not a neutral descriptive term denoting people who have fled; it is a legal and political term and the way it is bestowed speaks to placemaking activities of the state. The contribution that geographical approaches to refugees makes is in studying the relationship between 'refugeeness' as a general condition of forced displacement and the legal category refugee bestowed by states. States are manifestations of a territorial take on political life; geographers do well to illumine the connections between territoriality, place making, and refugees.

The term 'refugee' is used by states, international organizations, and aid agencies to recognize and to govern individuals who have fled their countries of origin for specific reasons. In practice, states have exercised considerable leeway in determining who counts as a refugee. The 1951 Geneva Convention on the Status of Refugees defines a refugee as one who has fled his or her home state to another state because of persecution that the state either sponsors or is unable or unwilling to prevent. States are therefore the central parties in determining refugee status. This is despite the existence of the United Nations High Commissioner for Refugees (UNHCR), which acts as an expert and advocacy body conveying best practice for assessing refugee claims, but in most cases has little authority. The authority with the most capacity to determine who may be legally recognized as a refugee and who not is one that is centered on states. In other words, 'refugee' is not a neutral descriptive category; it is a prescriptive term of government. A geographical account of refugees begins at this juncture – between an ostensibly neutral description of individuals who are displaced and a prescriptive account of who counts as a refugee. The former gives an account of numbers and movement across space and territory; the latter focuses on the spatial and territorial practices – practices of making place – that people claiming to be refugees must negotiate, avoid, submit to, or resist. Practically, this involves studying the cultural, economic, and political practices that give territories their specific flavor across different scales and how these influence how states respond to – or fail to respond to – refugee claims.

A geographical treatment of refugees must include, but cannot be limited to, a survey of where refugees are distributed across the world. Central to any understanding of geographical aspects of refugees is an account of how they are governed and the spatial issues involved. As refugees move from a place of persecution to a place of possible protection, they are caught up with different economic, political, and cultural practices that constitute the meaning of the spaces they move in or toward.

Studying this geographical juncture involves looking at the different points at which individual or group displacement meets placemaking activities. One way of doing so is to study the movement of refugees. In a somewhat stylized – but still useful – account, this movement has three stages: flight from places of residence across an international border; capture, reception, and assessment at another international border; and either refugee status bestowed or refused in a host country; the main issues here are deportation and conditions pertaining to the right to remain. This qualitative account seeks to add color and social nuance to a more neutral account of refugees – embodied, for example, in the 1951 Convention and in the UNHCR Charter which I describe later – that say refugees are individuals forced by persecution or social collapse to move outside their country to another country to seek protection. This definition does not focus on the spatial practices of making place that can influence how refugees move and how their reasons for moving are perceived. Unlike the neutral account of refugee movement in the 1951 Convention and the UNHCR Charter, examining actual practice may allow us to read the social geographies that refugees encounter and must deal with.

The International Refugee Regime

The central document influencing refugee recognition and the way states and the UNHCR relate to refugees is the 1951 Geneva Convention on the Protection of the Rights of Refugees. That document outlines obligations states have toward refugees and lists factors to take into account when determining refugee status. The key factors are that refugees must have experienced or be in threat of persecution that the state either sponsors or is unable or unwilling to contain and that refugees must be outside of the country of their origin. Central to the 1951 Convention, and to all subsequent analyses and recognition of refugees, is this state-centered and international aspect. Refugees must move beyond borders to another state because their home state is unable or unwilling to protect them.

Article 1 of the 1951 Geneva Convention Relating to the Status of Refugees applied the term to.

> [Any person who] … owing to a well-founded fear of persecution for reasons of race, religion, nationality, membership of a particular social group, or political opinion is forced to migrate involuntarily across an international boundary and remain outside his country of nationality.

Cause, motive, and political boundary thus became the classic benchmark determiners of refugee status. The 1951 Refugee Convention specifically addressed the status of Europeans displaced by World War II. An amended 1967 Protocol established legal standards for refugees everywhere. Although the 133 signatories are not required to provide asylum to refugees, they are prohibited against refoulement – expelling or returning refugees to countries where their life or freedom would be threatened (Hamilton, 1999).

In 1951, the United Nations (UN) General Assembly created the Office of the UNHCR with a brief 3-year life span in mind. In 1954, however, the Assembly extended UNHCR's mandate for another 5 years. Successive 5-year renewals have followed ever since. UNHCR is charged with refugee protection and the promotion of lasting solutions to their displacement The organization advocates the protection of refugees in would-be host countries and assists and protects returnees who repatriate to their home countries and groups of people who face refugeelike situations within their country of origin. Such assistance may include developing refugee integration programs, managing refugee camps, disseminating country information to states to assist them in making decisions and relief work in refugee situations. The UNHCR is also active in states that are not signatories of the Geneva Convention where they assess asylum claims and provide letters of protection, which states recognize at their prerogative.

The condition of forced migration applies to more than 34 million of the estimated 100 million permanent migrants worldwide. Of these, around 20 million are internally displaced persons (IDPs). IDPs are forced to move due to explicit forms of persecution or armed conflict or because of natural or man-made disasters and are not normally seen as legitimate refugees because of the Convention's bias toward movement across state borders.

Flight

Although geographers, starting with Ravenstein's work in the late 1800s, have long studied voluntary migration, their analysis of forced migration lagged behind. At the 1950 meetings of the Association of American Geographers, Proudfoot introduced a ninefold categorization of refugee population. That paper constitutes an early example of geographic work on the specific topic of refugees. Geographers thus began documenting the flight and distribution of specific refugee populations and have continued to do so.

Refugee research, including that of geographers, has grown substantially during the last 25 years (Hyndman, 1999). Refugee study programs now exist at universities in Canada, Egypt, Great Britain, Kenya, Sweden, and the United States. Academic interest has correlated with the growing magnitude of refugee crises worldwide; numbers of refugees, including IDPs, increased from less than 3.5 million in 1977 to more than 20 million in 2012 but the magnitude of refugees is also cultural, political, and economic and this is often evident in discourses and practices that see refugees as a threat to a national fabric (Bigo, 2002).

During the 1980s, refugee research largely focused on the first and third stages of the refugee experience, the processes of flight and resettlement. Resettled Central Americans and Indo-Chinese in the United States were the focus of study in more developed countries. That East Africa had become the epicenter of refugee crisis in the developing world was reflected in books on Eritrean and Ugandan resettlement (see Rogge, 1985; Harrell-Bond, 1986; Bulcha, 1988). A basic distinction was frequently made between assisted refugees housed in UN-sanctioned camps or settlements and unassisted refugees who either self-settled in rural areas or in urban centers (Wijbrandi, 1990).

Over the past 20 years, considerable attention has been paid to understanding 'flight' as a specific component of the complex movements that refugees undertake. Entwined with attention to movement and mobility are accounts of immobilization. Diken (2004) and Agier (2002) argue that refugees and displaced peoples are subject to concerted efforts by states, security personnel, the UNHCR, and sometimes asylum bureaucracies to contain, direct, or control their flight. As a result, people in flight are directed to camps or various types of holding centers. What occurs then are attempts to quarantine refugees away from the normal social, economic, and political environment. Refugees and displaced people on the move, in flight, are cast as people to be concerned about, the territorialization of identity headed by the state calls for political and social being that is settled and domesticated (Dillon, 1999).

White (2002) argued that the rise of refugee studies as a discipline contributes to a static depiction of refugees characterized by simplified models of community, home, and place. These are models that owe much to the dynamics of aid funding as well as to security discourses. From around the mid-1990s as a response to increased displacement of groups and individuals due to civil strife or environmental disaster, the concept 'complex humanitarian emergency' was articulated and employed to manage and help internally displaced people (Hyndman, 1999). Under conditions of complex humanitarian emergency, many actors are involved in managing displaced people (the state does not necessarily dominate). The UNHCR and international aid agencies frame displaced people as individuals or groups in need of help or assistance.

The consequence, argues White (1999), is that the focus becomes how flight may be straightforwardly managed from a humanitarian emergency perspective. A key assumption is that space is a neutral element and that refugees may be understood holistically and isolated from the spaces in which they move and live. One consequence is that refugees are viewed as a specific population, different from other populations in a given territory. They become a category of humanitarian government with the chaotic reality of living and moving through social space becoming subsumed to a systematic and coherent conceptualization of displaced persons as victims requiring humanitarian aid (and not as political subjects) (Malkki, 1996). The multiple ways in which individuals connect to place and community become secondary, and the differences within refugee populations that mean there are multiple identities and connections stretching across scales become less important. Refugees are understood as self-contained communities and as needy individuals. One result is that differences among refugees – how they move or the extent of their capacity to influence their social

environment – become neglected relative to the emphasis placed on their systematic management.

To recap then, geographers and sociologists have noted that at the moment of flight, people become trapped by a discourse and practice of "complex humanitarian emergency," which simplifies how displaced people relate to space, and the reasons for their displacement. Further, this discourse and practice ends up depoliticizing refugees: the primary way they are seen is as victims of humanitarian emergency; they are not seen as people who may have different degrees and experiences of displacement and of the causes of displacement. Displacement is a political condition: those who flee and are less able to cope with the conditions of displacement are those already marginalized by society, for example, the very poor or in some cases, women.

The UNHCR estimates that in 2012 there were 28.8 million internally displaced people in the world. The UNHCR also estimates that, at the beginning of 2012, there were 10.8 million refugees in the world. Numbers of refugees fell slightly in 2012 but numbers of internally displaced people have steadily increased. In 1997, there were around 17 million internally displaced people. In 2012, around 6.5 million individuals were newly displaced, a significant rise on the 3.5 million of the year before. The majority of displaced people are in sub-Saharan Africa, with the largest number of newly displaced people located in the Middle East and North Africa, a result of recent conflicts in that area. More people are becoming internally displaced, but fewer are crossing international borders. One reason for this may be their treatment as humanitarian victims immobilized in camps, another may be the successes that would-be receiving countries have had in preventing displaced people from crossing borders. This point will be examined in the next section, on borders.

Encountering Borders

In 1990, the locus of the global refugee population was the Global South. Developing countries hosted nearly 97% of all refugees. Conflicts in Afghanistan, Ethiopia, and Mozambique produced populations that each numbered in the millions. The Palestinian community (estimated at up to 4 million) had long since become permanent stateless exiles dispersed throughout the Middle East. Together, these four groups accounted for more than 13 million of the 16.6 million refugees worldwide. Yet, refugees could be found among more than one hundred different countries. And exile periods were lengthening.

During the 1990s, geographers diversified their approach to match the growing complexities of the refugee picture (Black, 1991; Black and Robinson, 1993). Moving beyond a 'refugee-centric' perspective (Chambers, 1986), scholars treated refugee settlement less often as if it were a self-contained process in isolation. Bascom (1993), for example, related the larger processes of rural transformation and agrarian change to that of refugee settlement. Although media images give the impression that refugees are mass 'flows,' it became abundantly clear that ethnic, age, gender, and class differentiation exist among uprooted populations (Bascom, 1998).

More recent scholarship by geographers has focused on the specific imaginings of the border that are discursively repeated and embedded in the state and nonstate institutions that refugees negotiate with. The centrality of the national border to refugee assessment rests on the way it is deployed as a way of imagining and realizing a national unity. Grundy-Warr and Wong (2002) argue that this imagination in Burma led to the displacement of hundreds of thousands of villagers from border areas populated largely by national minorities. In the quest for national unity, the state perceived minorities near the border as threatening. The taint of the border carries, meaning that individuals and groups denoted as not having a right of belonging in a territorially delimited 'national unity' tend to be approached at best with caution and at worst with prejudice by receiving states. Refugees are perceived as entering a preformed space, where they are to "conform to existing patterns of socio-spatial identity" (Newman and Passi, 1998: 190) with consequences for settlement or integration programs, as I will explore in the next section, as well as for asylum assessment policies at the border.

Receiving countries increasingly describe refugees as threats (Mountz, 2010). Boat people are criminalized; camps and holding centers are set up offshore. Over the past decade, Australia has developed offshore processing centers in nearby countries Nauru and Papua New Guinea (Prem Kumar, 2004). Geographers have used these and other examples to argue that borders are not static entities but instruments of control deployed against perceived threats as a way of making good sovereignty (Hyndman and Mountz, 2008).

In studying refugees' varied encounters with potential receiving states, scholars illumine not only the condition of 'refugeeness' (Malkki, 1995), but also how territorialization as the basis of modern politics occurs. Studying refugees is then an opening into studying how politics is constituted. Scholars have argued that territorialization contains human rights within national borders. Some scholars have deduced that refugees are without rights as a consequence, though it is perhaps more correct to say that refugees may be treated by individual states as though they have a lesser complement of rights. As it proclaims a national unity, the border develops a cultural power that state institutions and media deploy to show national belonging and nonbelonging. Refugees then become individuals who at least potentially are figures upon which the state can perform its power.

Some states create formal or informal 'buffer zones' that restrict refugees from entering into its territory. Australia, for example, has exploited its unequal relationship with small Pacific countries to create offshore processing procedures in exchange for aid in Nauru and Papua New Guinea. By creating a buffer zone, Australia effectively divests itself of its responsibility under the Geneva Convention to assess refugee claimants. In the European Union, in order to curb the movement of refugees across the Mediterranean, Italy and Libya cooperate to intercept boats from north Africa, to detain returnees in Libya, and have established Italian police liaison in Libya (Spruce, 2013).

Refugees and Host Countries

Repatriation became a 'lead' research theme in the 1990s (Wood, 1989; Allen and Morsink, 1994). As early as 1984,

UNHCR had identified repatriation as one of three 'durable' solutions for refugees. The *de facto* solution is long-term integration of refugees in the country of first asylum. The exceptional solution, reserved for only a small percentage of asylum applicants, is resettlement to a third country of asylum. The optimum solution of the three is repatriation – the return migration of refugees from exile to their homeland. The UNHCR proclaimed the 1990s as the 'decade of repatriation.'

Repatriation research focused first on legal or logistical matters, then shifted to the socioeconomic and psychological dimensions of repatriations (Rogge, 1991). Mozambique, Eritrea, and Afghanistan were some of the principal countries of concern. In theory, the repatriation decision is meant to be voluntary, fully informed, and to occur only when the conditions that prompted flight no longer exist. Practically, however, repatriations can involve different scenarios such as refugees choosing to return despite ongoing conflict or being forced to return home by host governments. Forced deportation, or refoulement, of refugees is the worst possible violation of the 1951 Refugee Convention. Deportation or repatriation does not necessarily occur because of a neutral assessment of home country conditions. Refugees may be sent to their countries of origin for a number of reasons, including political and economic pressures in receiving countries.

In Europe, the use of temporary protection visas (TPVs) has been on the rise since the Yugoslav wars and has facilitated deportations. A state's responsibility for a refugee under a temporary protection scheme must be renewed at regular intervals. Such renewal is based on assessment of conditions in home countries. People under TPVs experience a general insecurity, which impacts negatively on their capacity to incorporate into the host society. Once a home country is considered safe, TPV holders may be deported with little notice, often to postconflict societies with inadequate social infrastructure to assist reintegration. Children on TPVs may grow up in a host society and have little or no connection with their 'home' countries to which they are suddenly deported.

In Europe, deportations of failed asylum seekers and of refugees whose TPVs are not renewed appear to be more the norm than is the granting of refugee status. In 2004 in Germany, reports estimated that around 50 000 people were removed, and only 960 granted asylum, and in the United Kingdom, estimates in the same year say that only 4% of asylum applicants receive some form of protective status. The United Kingdom has since also passed a law where those already with refugee status may be deported to their home countries, regardless of their safety there, if they commit a 'serious crime,' the definition of which includes trespassing and possession of drugs. This definition appears to be against the Geneva Convention, which states that refugees may only be deported if they commit 'particularly serious crimes,' what the UNHCR calls 'heinous acts' (Fekete, 2005).

The UK example in particular seems to indicate the precarious conditions of refugees. As outsiders, refugees find considerable difficulty in shifting their status to something approaching permanence. In Amartya Sen's language – he applies it to poor people under conditions of food insecurity – refugees have difficulty in exerting command over the rights that they have as well as the rights that they deserve. In another example, legally recognized refugees in the European Union do not have the same right of free movement as other EU residents. They cannot move out of the country in which they were granted asylum, regardless of the existence of better job opportunities or personal relations.

The majority of the world's refugees are in the South. While refugee situations and treatment in the North are given due attention because of the stark inequalities and contradictions of ostensibly liberal-democratic states treating refugees illiberally, the fact is that most refugees in the South relate to, negotiate with, are governed by, and are sometimes persecuted by a range of different groups. The key difference between refugees in the North and those in the South is that refugees in the South encounter states with weaker machinery. Weakness as regards refugees may sometimes be assumed, a state may portray itself as being unable to provide services for refugees or to prevent discrimination or persecution by nonstate parties, thus restricting their responsibility for refugee well-being.

In a sense, because of the territorial nature of modern politics and the consequence that refugees are strangers in a world of nation-states, there are strong similarities in the way refugees are treated in both the North and the South; the key difference is that in the North external and internal pressures – media or international organizations or a tradition of a democracy centered on a state – are strong enough that the dominant relationship that refugees have is with the state, and – largely – with the institutionalized rules that guide its operation. In much of the South, states do not or cannot control the way their authorities treat refugees. In Malaysia, for example, a paramilitary civil volunteer group, RELA, is tasked with hunting down illegal immigrants, leading to allegations of extortion, use of force, and illegal detention (Human Rights Watch, 2007). The outsourcing of asylum management to RELA also has the consequence of distorting differences between migrant workers and asylum seekers, all classed as potentially 'illegal immigrants.' Some asylum seekers in Malaysia, notably those from Burma, are in 'protracted refugee situations,' where refugees are unable to return to their country of origin, integrate into a host state, or move to a third country. The UNHCR estimate that there are some 6 million refugees in such situations – excluding the special case of Palestinian refugees – most of them in Africa. In a protracted refugee situation, host governments are unwilling or unable to countenance the economic and political implications of settlement. The outsider status that sticks to most refugees tends to gain a harder currency. Refugees tend to be criminalized in state policy and in general public discourse, often with devastating consequences. In Kenya, for example, in the 1990s, the Dadaab camps hosting refugees from Somalia were rife with violence, rape, and murder. Refugees in protracted situations are structurally insecure, their condition of nonbelonging becomes ingrained, protection is either withdrawn or unreliable, and various groups – paramilitary organizations like RELA or bandits around Dadaab – prey on them.

Future Directions

Geographical perspectives on refugees center then on assessing the connections between the movement of refugees and the social spaces through which they move. Allied disciplines, such

as social anthropology and sociology, provide in-depth ethnographic accounts of how refugees are embedded in social, economic, and political relations. Such accounts do not always give a systematic account of space and its production. The contribution of geographical approaches to refugees centers on their capacity to read such relations as active moments of placemaking, of transforming space into territory at different scales and with different degrees of intensity. Geographers connect particular and ostensibly localized acts to wider spatial projects, such as that of territorialization.

Refugee situations are likely to increase in complexity. Three dominant strains, exemplary of broader political and economic structures, are worth pointing out and may prove useful areas for future study. One is the increasing number of people displaced by conflict coupled with the curtailing of their possibilities for flight and resettlement as refugees. EU member states supported military intervention during the 'Arab Spring' conflicts – at least in Libya – and have been vocal in criticizing abuses in Egypt, Tunisia, and Syria. At the same time, EU states have opposed flight from these regions to Europe. Second is the increased number of actors who deal with refugees at the different stages of their movement. One reason for this is the withdrawal of the state, what I have described above as the state assuming or taking on weakness. While largely characteristic of the state in Asia and Africa, state weakness is also evident in Europe at different stages of refugee movement. The formal withdrawal of the state in repatriation procedures in the European Union is telling, with private operators managing the process. A related characteristic is the fragmentation of the state when it comes to assessing asylum claims. Unlike citizens, refugees find it generally difficult to address the entirety of the state. States tend not to be coherent, and political departments and politicians as well as the legislature and judiciary are not necessarily in concert over every issue. Citizens are often able to play with these internal disagreements to try to achieve specific aims. Refugees by contrast tend to be dominated by one specific department of the state – whether the police or an asylum bureaucracy – with little leeway to address or strategically bring into play the entirety of the state. A third strain then is the increasingly blurred lines between asylum seekers, working class migrants, and undocumented migrants. The reasons for these blurred lines are varied, and often depend on specific political and economic pressures or conditions in different host states. The refugeelike experience, existing on the margins of society with comparatively restricted capacity to command rights is an increasingly common condition for poor migrants whether documented or not or whether asylum-seeking or not. In 2011, there were more asylum seekers from Zimbabwe than any other country as individuals sought to flee economic collapse. Only 5% of these asylum-seekers were granted refugee status. The conditions of 'persecution' that lead to flight and to people seeking to resettle elsewhere are varied and the categories of forced or voluntary migrant do not do them justice.

UNHCR's statistics point to increasing numbers of displaced people but declining numbers of people making asylum applications. In 2011, the UNHCR estimates that there were approximately 26 million refugees, including IDPs, but in that year only 876 100 asylum applicants were made. A major issue then is the gap between refugee situations and the capacity of individuals in such situations to enter into refugee assessment procedures. The existing refugee regime does not appear able to address the complexity or magnitude of refugee situations. That regime is structured to manage and contain a potential or existing 'refugee problem,' rather than to help people in flight find abiding security. Geographers can shed light on the difficulties of flight by looking at territorialization mechanisms that work in the first instance to contain a refugee problem. Geographers can also shed light on the complex placemaking activities that obscure refugee situations, or simplify these so that they may be adequately managed by states.

See also: Global Migration; Homelessness in the United States; Immigration Policy; Immigration, Applied Research on; Immigration: Political Aspects; Internal Migration: Developing Countries; International Migration; Migration, Theory of; Migration: Anthropological Perspectives; Refugees in Anthropology; Space and Social Theory in Geography.

Bibliography

Agier, M., 2002. Between war and city: towards an urban anthropology of refugee camps. Ethnography 3 (3), 317–341.

Allen, T., Morsink, H. (Eds.), 1994. When Refugees Go Home. Africa World Press, Trenton, NJ.

Bascom, J., 1993. The peasant economy of refugee resettlement in Eastern Sudan. Annals of the Association of American Geographers 83, 320–346.

Bascom, J., 1998. Losing Place: Refugee Populations and Rural Transformations in East Africa. Berghahn Press, New York.

Bigo, D., 2002. Security and immigration: toward a critique of the governmentality of unease. Alternatives 27, 63–92.

Black, R., 1991. Refugees and displaced persons: geographical perspectives and research directions. Progress in Human Geography 15, 281–298.

Black, R., Robinson, V. (Eds.), 1993. Geography and Refugees: Patterns and Processes of Change. Belhaven Press, London.

Bulcha, M., 1988. Flight and Integration: Causes of Mass Exodus from Ethiopia and Problems of Integration in the Sudan. Scandinavian Institute of African Studies, Uppsala, Sweden.

Chambers, R., 1986. Hidden losers? the impact of rural refugee programs on poorer hosts. International Migration Review 20, 245–263.

Collinson, Sarah, 1996. Visa requirements, carrier sanctions, 'safe third countries' and 'readmission': the development of an asylum 'buffer zone' in Europe. Transactions of the Institute of British Geographers 21 (1), 76–90.

Diken, B., 2004. From refugee camps to gated communities: biopolitics and the end of the city. Citizenship Studies 8 (1), 83–106.

Dillon, M.G., 1999. The scandal of the refugee. In: Campbell, D., Shapiro, M. (Eds.), Moral Spaces: Rethinking Ethics and Global Politics. Minnesota University Press, Minneapolis.

Fekete, L., 2005. The deportation machine: Europe, asylum and human rights. Race & Class 47 (1), 64–91.

Grundy-Warr, C., Wong, E., 2002. Geographies of displacement: the Karenni and the Shan across the Myanmar-Thailand border. Singapore Journal of Tropical Geography 23 (1), 93–122.

Hamilton, V., 1999. The 1999 World Refugee Survey. US Committee for Refugees, Washington, DC.

Harrell-Bond, B., 1986. Imposing Aid: Emergency Assistance to Refugees. Oxford University Press, London.

Human Rights Watch, 10 May 2007. Malaysia: Disband Abusive Volunteer Corps. News Release. http://www.hrw.org/news/2007/05/08/malaysia-disband-abusive-volunteer-corps (accessed 16.08.13.).

Hyndman, J., 1999. A post-cold war geography of forced migration in Kenya and Somalia. The Professional Geographer 51, 104–114.

Hyndman, J., Mountz, A., 2008. Another brick in the wall? Neo-*refoulement* and the externalization of asylum by Australia and Europe. Government & Opposition 43 (2), 249–269.

Malkki, L., 1995. Refugees and exile: from "refugee studies" to the national order of things. Annual Review of Anthropology 24, 495–523.

Malkki, L., 1996. Speechless emissaries: refugees, humanitarianism and dehistoricization. Cultural Anthropology 11 (3), 377–404.

Mountz, Alison, 2010. Seeking Asylum: Human Smuggling and Bureaucracy at the Border. University of Minnesota Press, Minneapolis.

Newman, D., Passi, A., 1998. Fences and neighbours in the postmodern world: boundary narratives in political geography. Progress in Human Geography 22 (2), 186–207.

Prem Kumar, R., 2004. "Making place": the "Pacific solution" and Australian emplacement in the Pacific and on refugee bodies. Singapore Journal of Tropical Geography 24 (3), 290–306.

Prem Kumar, R., Arendas, Zs, 2013. Exceeding categories: law, bureaucracy and acts of citizenship by asylum seekers in Hungary. In: Isin, E., Saward, M. (Eds.), Enacting European Citizenship. Cambridge Univesity Press, Cambridge.

Rogge, J., 1977. A geography of refugees: some illustrations from Africa. Professional Geographer 29, 186–189.

Rogge, J., 1985. Too Many, Too Long: Sudan's Twenty-year Refugee Dilemma. Rowman and Littlefield, Totowa, NJ.

Rogge, J., 1991. Repatriation of refugees: a not-so-simple 'optimum' solution. In: Allen, T., Morsink, H. (Eds.), When Refugees Go Home. Africa World Press, Trenton, NJ.

Spruce, D., 2013. From the Pacific solution to the mediterranean solution (Doctoral thesis). University of Technology Sydney.

UNHCR, 1993. The State of the World's Refugees 1993: The Challenge of Protection. Penguin, New York.

UNHCR, 2011. Statistical Yearbook 2011. United Nations, New York.

White, A., 2002. Organic functionalism, 'community' and place: refugee studies and the geographical constitution of refugee identities. Geoforum 33 (1), 73–83.

Wijbrandi, J., 1990. Organized and spontaneous settlement. In: Khulman, T., Tieleman, H. (Eds.), Enduring Crisis: Refugee Problems in Eastern Sudan. Lieden. African Studies Centre, The Netherlands.

Wood, W., 1989. Long time coming: the repatriation of Afghan refugees. Annals of the Association of American Geographers 79, 345–369.

Region: Geographical

Andrew EG Jonas, Department of Geography, Environment and Earth Sciences, University of Hull, Hull, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The regional idea has long contributed to our sense of living in a world of geographical differences. Even as geographers have championed the concept, they have had profound disagreements about how to approach the region. Early disagreements occurred around the extent to which a region's uniqueness reflected the dominance of the physical environment over cultural factors. Later generations rejected the task of carving up the world into formal regions in preference for spatial analysis. Nevertheless the regional concept has demonstrated resilience even in the face of forces of global homogeneity. Regional scholars today examine regions not as neatly bounded territories but instead as semicoherent territorial assemblages of power relations and economic development processes.

Most people are familiar with the idea that the world can be divided up into geographical regions. Whether based on physical features such as topography, ecology, climate, and vegetation, or human characteristics such as culture, language, and religion, regions still have a great deal of resonance in popular culture, tourism, and the media. Notwithstanding today's powerful forces of global economic and cultural integration, the regional idea continues to contribute to our sense of living in a world marked by geographical difference and diversity. It should therefore come as no surprise that geographers have often championed the concept of the region, regarding it as the discipline's *raison d'etre* in much the same way that geologists think of plate tectonics or economists relate to supply-and-demand curves.

While few now question the value of the regional idea, geographers over the years have had some profound disagreements about how to approach the region. Intellectual schisms initially occurred around the extent to which a region's uniqueness – usually taken as given – reflected the relative importance of environmental influences viz-à-viz human or cultural factors. Subsequently geographers wondered whether carving up the world into formal regions based on their unique attributes was adequate to the task of promoting geography as the science of the spatial. In the 1950s and 1960s, regions were analyzed on the basis of spatial interactions (e.g., trade flows or service-area thresholds) rather than classified on the basis of formal attributes such as climate or culture. Spatial scientists looked to distance the discipline of geography from the regional tradition, only themselves later to be challenged by radical geographers who attributed growing regional inequalities to uneven capitalist development. Nevertheless, regional synthesis *qua* cultural geography clung on well into the latter half of the twentieth century, offering a counterfoil to the quantitative reductionism of spatial science.

In the 1980s, a new generation of critical geographers sought to revive the regional concept. Inspired by neo-Marxist studies of uneven regional development, the new regional geography (NRG) promised to deliver theoretical insights into wider processes of post-Fordist industrial restructuring. It led to a resurgence of the regional concept in geography and closely related disciplines, such as planning and economics, well into the 2000s. Arguing that the uniqueness of regions results from a combination of global and local processes, today's critical scholars no longer examine regions as neatly bounded territories (or scales) waiting to be discovered. Instead regions are approached as semicoherent territorial assemblages wherein wider power relations and networks of governance come together – or fragment – often in contingent and unexpected ways.

Traditional Regional Geography: From Environmental Determinism to Regional Synthesis

In the late nineteenth century, geographers were interested in the role of physical and human factors shaping the form and character of regions. At this stage, geography was influenced by neo-Darwinist views on the relationship between the environment and the development of human societies. Environmental determinism held that dominant racial characteristics reflected the overwhelming influence of the physical environment (Huntington, 1924). The task of the geographer was to identify the presence of such characteristics in a regional population and, in so doing, chart regional boundaries. In the 1920s and 1930s, the Royal Geographical Society in the UK commissioned major expeditions to unexplored regions like the Himalayas, laying the groundwork for carving the world up into formal regions, many still recognizable today. In this manner, the study of human–environment relations prepared the ground for the spread of colonialist systems of governance to remote regions of the world well into the mid-twentieth century.

What gave further legitimacy to the regional idea in geography was the analysis of peripheral regions within Europe at a time of growing threats to empire and the nation state. Early advocates of regional analysis, such as Friederick Ratzel in Germany and Halford Mackinder in the UK, were considered experts in the study of geopolitics. In Germany, Ratzelian geopolitics was soon tainted by its association with the rise of national socialism and racialized visions of regional (viz. national) purity. In the UK, however, Mackinder saw in regional synthesis an opportunity to promote geography and overcome the 'great abyss' between it and the great natural science disciplines of the day such as geology (Livingstone, 1992). What was at stake was the public relevance of the

emerging subject of geography as it struggled to find its niche in the British academy. Later on, in the USA the political geographer Richard Hartshorne developed principles of functional regionalism in order to demonstrate how strong systems of regional planning underpinned national territorial integration and geopolitical statecraft (Hartshorne, 1950).

Environmental determinism eventually gave way to regional possibilism and the idea that human adaptation can be an active response to the constraints of the physical environment. In Germany, Alfred Hettner had proposed the concept of chorology to demonstrate the coimportance of human and physical and human factors in explaining the character of a region (Livingstone, 1992).Yet no single unified view of the region took hold. Instead different traditions of regional enquiry flourished, some barely touched by environmental determinism or functionalism. These hybrid regional approaches sought to achieve a 'true' synthesis of environmental, social, cultural, and economic factors.

Many regional scholars were motivated to study regional life on its own terms. The French geographer Paul Vidal de la Blache referred to this in *Principles of Human Geography* (1926) as a region's *genre de vie* – its particular lifestyle resulting from a combination of natural, social, and cultural factors. Vidal is perhaps better known in Anglo-American geography for *Tableau de la Geographie de la France* (1903) in which he captures what the regional idea represented to France prior to the spread of industrial capitalism. If the main concern was to specify the uniqueness of the region, this somehow also spoke to the uniqueness of the nation. It was essential to Vidal's regional method that the material underpinnings of the region were local and rural even if the effects of long distance trade were clearly visible in the landscape (Wrigley, 1970). Where Vidal's vision of regional geography began to encounter difficulties was when these wider economic and political forces began to transform the cultural–environmental integrity of the region. To paraphrase Wrigley (*op. cit.*: p. 10), the arrival of the canals, railways, and new industries did not simply create a new layer in the landscape which meshed with preexisting layers without disrupting the essential integrity of the latter. Rather it actively destroyed the rural past and so transformed the essence of regional life. Nevertheless, the regional approach continued as a form of descriptive synthesis well into the mid-twentieth century.

Meanwhile, in the United States two very different regional traditions competed for acceptance (Pudup, 1988). These were the Berkeley School of Cultural Geography headed by Carl Sauer and areal differentiation as espoused by Richard Hartshorne. Sauer (1963) viewed regions as organic wholes demanding exhaustive investigation and interpretation through intensive fieldwork. In *The Nature of Geography* (1939), Hartshorne rejected the Sauerian notion that regions exist as cultural organic wholes waiting to be discovered. He sought to show that regions could be defined systematically by showing how their particular component elements, such as relief, vegetation, climate, land use, and so forth, related to other elements. For Hartshorne, regional classification called for a formal process of areal differentiation. Although both perspectives were marginalized after World War II, there were various attempts at revival and rapprochement, which eventually floundered in the rising tide of spatial science.

The New Regional Geography

Criticisms leveled at traditional regional geography became more strident in the 1950s and 1960s. Inspired by powerful new computer technologies and the allure of quantitative methods, spatial scientists turned their backs on regional analysis, regarding it as backward, parochial, and devoid of intellectual rigor (see Johnston, 1991). In light of the alleged superiority of spatial modeling, disagreements over the respective merits of regional synthesis versus areal differentiations seemed out of touch. However, the election of John Fraser Hart as the president of the Association of American Geographers led to a brief attempt to engineer a revival of the regional concept in the early 1980s. In his Presidential Address published in the Annals of the Association of American Geographers in 1982, Fraser Hart memorably made the case for regional synthesis, arguing that "We need a unifying theme that ties together all of the widely disparate phenomena with which we deal, and I know of no other theme that is even remotely so satisfying as the idea of the region. All of the different strands of geography converge when we try to understand a place, an area, or a region … and we need the concept of the region in order to understand why we need the diverse and variegated systematic subfields of geography" (Hart, 1982: p. 18). These sentiments were echoed 3 years later by Pierce Lewis in his Presidential Address (Lewis, 1985). Yet regional geography's eventual savior came from a somewhat unexpected quarter, namely, Marxist-inspired radical geography.

Radical geographers were attracted to regional synthesis because of the need to back up their theories of industrial restructuring and uneven development with empirical evidence. What was at issue was not whether regions could (or should) be formally delimited and functionally classified. Rather the focus of the NRG was on the causes of regional industrial transformation. Specifically, could industrial decline be attributed to some particular characteristics unique to regions or did the causes lie elsewhere in the form of regionally nonspecific processes of deindustrialization and technological change? Doreen Massey (1983: p.75) puts the issue succinctly as follows: "What [is seen] are national processes at work producing the uniqueness of local economic and social structures… [But] the vast variety of conditions already existing at [the] local level also affects how these processes themselves operate." Making use of a geological metaphor, it was possible to argue that regional uniqueness resulted from the interactions between recent 'rounds of capital investment' and the respective local 'deposits' (e.g., labor skills, class structures, products, and so forth) left behind in regions by earlier 'rounds of investment' (Warde, 1985).

The NRG-inspired geographers (along with other social scientists) to conduct detailed study of localities. 'Locality studies' soon became a surrogate term for regional synthesis. However, concerns were expressed that such studies represented a return to empiricism (Smith, 1987). In attempt to deflect these criticisms, proponents of locality studies argued that, unlike past approaches to the region, the NRG was informed at the outset by theories of industrial restructuring. It was important to demonstrate – rather than simply assert – how regional features, such as local labor markets and regional class structures, make a difference to industrial

restructuring, for instance, by altering the relative attractiveness of regions to subsequent rounds of capital investment (Warde, *op. cit.*).

In accounting for the revival of the regional concept, there were respectful references to the Vidalian tradition along with a growing awareness that the NRG itself encompassed diverse theoretical approaches (Pudup, 1988; Jonas, 1988; Agnew, 1990). Besides the aforementioned focus on regional industrial restructuring, the region could be examined as representing a meeting ground between human life-paths and wider social structures under different historical modes of production (Thrift, 1983). Moreover, the NRG portended the revival of the study of nature–society relations. However, this was not a return to a form of crude environmental determinism but rather a growing recognition of human adaption to the limits of nature and of capital. For example, work in political ecology has retained a strong regional environmental focus as it seeks to achieve a critical synthesis of culture, local knowledge, nature, and production (Robbins, 2012: p. 28).

In the meantime, the rural bias of traditional regional geography has given way to a plurality of spatial possibilities, including new economic regions such as the Third Italy and 'edge city' regions like Orange County in Southern California. Alongside environmental conditions, regional economic development processes are now accorded far greater importance in regional transformation in contrast to traditional regional geography's emphasis on culture, settlement forms, and landscape. Capitalist development after mass production seems to be producing not so much a uniformly globalized economy but instead a 'world of regions' based upon specialized clusters of economic activities (Storper, 1997).

A further extension of this theme is looking at how globalization occurs around major city-regions, which in turn undermine the integrity of the nation-state, ushering in a new era of regional resurgence. For economic geographers, such as Allen Scott (2001) and Richard Florida (2008), larger spatial economic units in the form of mega-urban regions (e.g., Greater Tokyo in Japan or Boston–Washington DC in the USA) are today's engines of the global economy rather than older territorial units like countries or Fordist industrial heartlands. Despite the privileging of global city-regions, the nation-state remains influential not least because it allocates public expenditures competitively in support of the infrastructural development of such city-regions (Jonas, 2013). Moreover, the rise of new regionalist social movements likewise points to the profound restructuring of citizenship and democracy from below the level of the nation state.

Future Regions: Relational or Territorial?

The regional idea carries such a bruised and battered baggage in geography that over time concepts like place, scale, and territory have achieved far greater prominence (Jessop et al., 2008). Scholars have even suggested that any attempt to try and 'fix' spatial processes around particular geographical scales like the region are both intellectually redundant and politically naive (Marston et al., 2005). Yet the regional concept remains resilient. To be sure, most scholars shun the idea that the region can be preallocated to a given scale. The endgame in regional analysis is not to produce a map of world regions having fixed and permanent boundaries. Rather the aim is to use the regional concept as a basis for investigating other processes – be these networks, flows, or power relations – and thereby reveal how the specific conjuncture of social relations and spatial structures gives regions their particular territorial form.

As Joe Painter (2008) suggests, regional analysis in geography has for too long suffered from a form of 'cartographic anxiety'; it needs to discard old ideas about drawing 'lines on the map' and think about how regions (however these are defined, materially or discursively) serve wider powerful interests. Regional analysis today involves a very different kind of mapping in which the goal is to reveal how the region represents an assemblage of material activities, social structures, and power relations whose domain is not so tightly bounded by the region (Allen and Massey, 1998). Even if the region tends toward some sort of dominant territorial form (e.g., competitive regionalism), this is merely the visible effect of wider state actions and geopolitical practices. Above all else, contemporary regional approaches seek to distance themselves from single slice-through-time descriptions of the region that characterized geography in the past. Most regional scholars now accept that, irrespective of certain commonly held functional or formal attributes, all regions are fluid and historically contingent social constructions.

In conclusion, one can write about the 'localness' or 'globalness' of a region without assuming that the region occupies a fixed scale between the local and the global. Today's regions have a quality of 'betweenness' (Entrikin, 1991) about them, which means that they cannot so easily be separated analytically from 'other' territories, scales, or spatial structures (Jonas, 2006). Martin Jones (2009) has described the region as a kind of 'phase space' in which various processes may combine either to fix the region around a well-defined territory or contribute to its eventual dissolution. It is in this respect that the regional idea remains intellectually resilient even if the material and social objects to which the concept refers often appear fragile, porous, and semicoherent. That regions continue to differ from one another, not just in their climate and settlement patterns, but also their systems of production can no longer be assumed, as was once the case, but has to be explained in a world governed by systemic processes of global capitalism, varieties of state intervention, and diverse social movements. One can anticipate that geographers shall find many ways of 'arguing with regions' (Agnew, 2013) long into the future.

See also: Cultural Geography; Globalization: Geographical Aspects; Local Economic Development; Regional Geography; Rural Geography; Urban Geography.

Bibliography

Agnew, J.A., 1990. Sameness and difference: Hartshorne's the nature of geography and geography as areal variation. In: Entrikin, J.N., Brunn, S.D. (Eds.), Reflections on Richard Hartshorne's the Nature of Geography. The Association of American Geographers, Washington, pp. 121–140.

Agnew, J.A., 2013. Arguing with regions. Regional Studies 47, 6–17.

Allen, J., Massey, D., 1998. Rethinking the Region. Routledge, London.

Entrikin, N., 1991. The Betweenness of Place. Johns Hopkins University Press, Baltimore, MD.
Florida, R., 2008. Who's Your City? How the Creative Economy Is Making where to Live the Most Important Decision of Your Life. Basic Books, New York.
Hart, J.F., 1982. The highest form of the geographer's art. Presidential address. Annals of the Association of American Geographers 72, 1–29.
Hartshorne, R., 1939. The nature of geography. Annals of the Association of American Geographers 29, 1–469 (1961. The Nature of Geography. Edwards, Ann Arbor, MI).
Hartshorne, R., 1950. The functional approach in political geography. Annals of the Association of American Geographers 40, 95–130.
Huntington, E., 1924. The Character of Races, as Influenced by Physical Environment. Natural Selection and Historical Development. C. Scribners Sons, New York.
Jessop, B., Brenner, N., Jones, M., 2008. Theorizing sociospatial relations. Environment and Planning D: Society and Space 26, 389–401.
Johnston, R.J., 1991. A Question of Place. UK: Blackwell, Oxford.
Jonas, A., 1988. A new regional geography of localities? Area 20, 101–110.
Jonas, A.E.G., 2006. Pro scale: further reflections on the 'scale debate' in human geography. Transactions of the Institute of British Geographers NS 31, 399–406.
Jonas, A.E.G., 2013. City-regionalism as a contingent 'geopolitics of capitalism'. Geopolitics 18, 284–298.
Jones, M., 2009. Phase space: geography, relational thinking and beyond. Progress in Human Geography 33, 487–506.
Lewis, P.F., 1985. Beyond description. Annals of the Association of American Geographers 75, 465–478.
Livingstone, D.N., 1992. The Geographical Tradition. UK: Blackwell, Oxford.
Marston, S.A., Jones III, J.P., Woodward, K., 2005. Human geography without scale. Transactions of the Institute of British Geographers NS 30, 416–432.
Massey, D., 1983. Industrial restructuring as class restructuring: production decentralization and local uniqueness. Regional Studies 13, 73–89.
Painter, J., 2008. Cartographic anxiety and the search for regionality. Environment and Planning A 40, 342–361.
Pudup, M., 1988. Arguments within regional geography. Progress in Human Geography 12, 369–390.
Robbins, P., 2012. Political Ecology, second ed. Wiley Blackwell, Chichester.
Sauer, C.O., 1963. Land and Life: A Selection from the Writings of Carl Ortwin Sauer. University of California Press, Berkeley, CA.
Scott, A.J. (Ed.), 2001. Global City-regions: Trends, Theory, Policy. Oxford University Press, Oxford.
Smith, N., 1987. Dangers of the empirical turn: the CURS initiative. Antipode 19, 59–68.
Storper, M., 1997. The Regional World. Guilford Press, New York.
Thrift, N., 1983. On the determination of social action in space and time. Environment and Planning D: Society and Space 1, 23–57.
Vidal de la Blache, P., 1903. Tableau de la Geographie de la France. La Table Ronde, Paris.
Vidal de la Blache, P., 1926. Principles of Human Geography. Lowe and Brydon, London.
Warde, A., 1985. Spatial change, politics and the division of labour. In: Gregory, D., Urry, J. (Eds.), Social Relations and Spatial Structures. Macmillan, London, pp. 190–212.
Wrigley, E., 1970. Changes in the philosophy of geography. In: Chorley, R., Haggett, P. (Eds.), Frontiers in Geographical Teaching. Methuen, London, pp. 3–20.

Regional Geography

John Harrison, Loughborough University, Loughborough, Leicestershire, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Wood, volume 19, pp. 12908–12914, © 2001, Elsevier Ltd.

Abstract

As a field of study, regional geography is concerned with understanding the spatial (regional difference and uniqueness) and temporal (regional change and development) aspects of the increasingly complex world we inhabit. In stark contrast to systematic geography, which is primarily interested in describing the distribution of particular phenomena over the entirety of the earth's surface and uncovering the general rules governing these patterns, regional geography has primary regard for the specificity of regions, places, and locales. As an academic endeavor, regional geography is the backbone of the geographical discipline. Nonetheless, its history is one characterized by ebb and flow, oscillating between periods of strong approval and periods of strong rejection.

As a field of study, regional geography is concerned with understanding the spatial (regional difference and uniqueness) and temporal (regional change and development) aspects of the increasingly complex world we inhabit. In stark contrast to systematic geography, which is primarily interested in describing the distribution of particular phenomena over the entirety of the earth's surface and uncovering the general rules governing these patterns, regional geography has primary regard for the specificity of regions, places, and locales.

As an academic endeavor, regional geography is the backbone of the geographical discipline (Figure 1). In many ways, the history of regional geography reflecting the history of geography – certainly in the mainstream thinking of the twentieth century – has given the former special affinity within geographical circles. This affinity is rooted in understanding how each new branch of modern and postmodern human geographical thought impacted regional geography and concomitantly how advances in regional geography have directly impacted the discipline.

What is important to grasp is despite regional geography being characterized as having a long and distinguished tradition, it is a history which has oscillated between periods of strong approval and periods of strong rejection. This ebb and flow was perhaps best encapsulated by Derek Gregory in his influential book, *Ideology, Science and Human Geography* (1978), where he concluded that:

> Ever since regional geography was declared to be dead – most fervently by those who had never been much good at it anyway – geographers, to their credit, have kept trying to revivify it in one form or another. … This is a vital task. … We need to know about the constitution of *regional* social formations, or *regional* articulations and *regional* transformations.
>
> p. 171

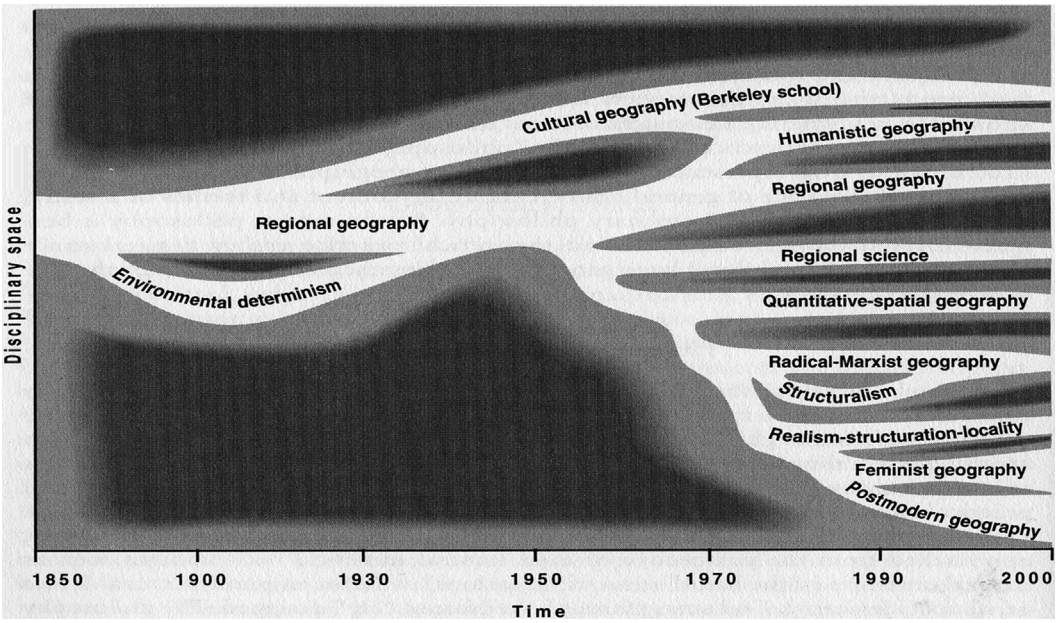

Figure 1 Schools of modern and postmodern human geographical thought. An increasing variety of schools of thought have emerged over time in the discipline of geography. Source: Peet, Richard, 1998. Modern Geographical Thought. Blackwell, Oxford; p. 10.

Thus, the history of regional geography is marked by a series of rises, each sparked by a new 'big idea' which gains traction to become the foundation for a period of 'popular' approval, and falls, as critics come round to deconstructing – and rejecting – the very idea upon which the success of regional geography at that time was ultimately predicated. The following sections of this article deal with how regional geography continually reinvents itself following periods of rejection.

The Birth of Regional Geography

Although regional geography can be traced back to ancient Greece, the 'classical' period of regional geography refers to the late nineteenth and early twentieth centuries. This was the period when geography became established as a university discipline, with regional geography at its core. Nonetheless, the birth of regional geography as an academic endeavor owes much to events taking place in the eighteenth and early nineteenth centuries when rapid colonial exploration and exploitation of the interior of continents necessitated the need for regionally specific geographical knowledge. This ranged from discovering hitherto unexplored parts of the world, surveying and mapping their features (topography, species, climate), and advising on what new infrastructures were needed to establish and protect commercial and political interests there. Naturalists such as Alexander von Humboldt arguably became the early pioneers of regional geography. Traveling extensively to hitherto unexplored parts of South America–Mexico and Cuba especially – between 1799 and 1804, von Humboldt (together with French explorer Aimé Bonpland) undertook observations which led him to conclude that it was the interrelation of all the physical sciences – climate, topography, geology – which produced distinctly 'regional' geographies.

In German-speaking countries, Alfred Hettner (1927) and Viktor Kraft (1929) followed a neo-Kantian line of argument to identify a new type of regional geography. Kraft, an Austrian, subdivided the discipline into 'general' systematic geography (as a nomothetic science) and 'regional' geography (as an idiographic science). Similarly, in Germany, Hettner argued geography must deal with both the unique (regional geography) *and* the universal (systematic geography), but as a chorological science the study of regions must occupy a pivotal position at the core of the discipline.

Meanwhile in the United Kingdom, the Geographical Society of London was established in 1830 (becoming the Royal Geographical Society in 1859) for the stated purpose of the 'advancement of geographical science.' This was allied to 'colonial' explorations principally in Africa, the Indian subcontinent, the polar regions, and central Asia. As well as funding expedition, the Society was also responsible for incorporating the study of (regional) geography in schools and for the first university positions in the discipline at the turn of the twentieth century.

Empire and Environment: The Traditional Regional Geography

The very essence of 'traditional' regional geography was the connection to empire and the environment. It was also characterized by environmental determinism, a period bookended by Charles Darwin's *On the Origin of the Species* (1859) and Richard Hartshorne's *The Nature of Geography* (1939) (see **Figure 1**). The period of traditional regional geography was also the first example of the rise and fall of a dominant orthodoxy – environmental determinism – in (regional) geography.

Traditional regional geography can be divided into three main schools of thought: environmental determinism, possibilism, and cultural traditions. Environmental determinism, geographers' shared conviction that there was direct connection between the nature of the physical environment and the historical activities of man, gained approval as (regional) geography's first 'big idea.' Oxford professor and head of the first UK Geography Department, Halford Mackinder's *Britain and the British Seas* (1902) codified how patterns of war, skills, and the evolution of race are all determined by 'natural features.' Brought to Oxford by Mackinder, Andrew Herbertson (1902) went further, arguing that above all other natural features it is climate that has the greatest influence on man and his work (sic) and proceeded to identify 'natural climatic regions.'

In France, Paul Vidal de la Blache offered his own very influential analysis that allowed for 'possibilism.' Vidalian regional geography does not deny that there are natural limits to human activity but recognizes the significance of humanity's choice of activity as opposed to the natural limits of it. In this way, it is unreasonable to pursue a regional geography which is determined only by environmental factors. For Vidal and those he inspired, the natural landscape cannot be separated from the cultural landscape. Moreover, the two influences – natural and cultural – fuse through regions to produce distinct 'pays' (regional landscapes) and 'genres de vie' (regional lifestyles).

Meanwhile back in the United Kingdom, Herbert J. Fleure injected an even greater interest in human experience into (regional) geography. Rejecting Herbertson's determinist approach, Fleure contended that regional geography owed more and more to the outcome of human activity, especially the exchange of cultural influences as people became less rooted in place. In line with this view, Fleure (1919) proposed 'human regions' and alongside Carl Sauer (Berkeley, California) and Paul Vidal de la Blache (Paris) founded geography's cultural tradition (see **Figure 1**).

Regional Geography as a Science

During the interwar period Anglo-American geography was forced in reinvent itself to uphold its status as a credible scientific discipline. The demise of environmental determinism meant regional geography lacked a 'big idea' while its position as the science of empire was of diminishing use when the major empires were in decline. Regional geography as the science of empire and environment consequently lost ground to emergent, 'scientific,' and 'relevant' disciplines such as economics, sociology, and meteorology.

Richard Hartshorne's *The Nature of Geography – A Critical Survey of Current Thought in the Light of the Past* (1939) represented geography's response to these challenges. Inspired by

Alfred Hettner's (1927) neo-Kantian argument, Hartshorne contended that what made geography 'exceptional' – and therefore useful vis-à-vis other scientific disciplines – was that at its core it was a 'regional' discipline. As the title and subtitle of his book both imply, Hartshorne's approach involved going back to the origins of regional geography to reaffirm geography as the study of what makes some places (or regions) similar and others different. In doing so, he prioritized geographer's use of idiographic methods to deal with the particular and the unique. For Hartshorne, geography was ultimately a study of 'synthesis,' where each region is a unique area defined and distinguished from others by being the product of a distinct blend of systematic geographical factors. Emphasizing the individuality of regions, this type of geography became known as the study of 'areal differentiation.' Perhaps unsurprisingly, attention would often be drawn to the uniqueness of regions, but as Hartshorne himself stated, his approach to regional geography required investigating both similarities and differences over space:

> Three major steps are involved in a study in regional geography. In order to comprehend the actual interrelation of phenomena in specific places it is necessary to consider small subdivisions within each of which the local variations of the various factors are arbitrarily ignored. The second step is to relate the unit areas to each other to discover the structural and functional formation of the larger region. Finally it is necessary to study the arrangement of the regions to each other and the interrelation of phenomena in one region to those in another.
>
> 1939: xv

In the United States especially, the study of areal differentiation was to provide a period of relative calm and stability for the geographical discipline. This was to all change, however, in the 1950s and 1960s. At this time geography's claim to exceptionalism came under sustained attack from a rapidly expanding nomothetic paradigm which was to give rise to regional science and quantitative-spatial geography (see **Figure 1**).

In *Exceptionalism in Geography: A Methodological Examination* (1953), German-born Fred K. Schaefer produced a withering critique of the Hartshorne-inspired practices of regional geography. Practitioners of regional geography were accused of engaging in unscientific working practices, of producing nothing more than woolly empiricist 'descriptions' of phenomena that lacked any theoretical foundations on which an analytical 'explanation' for their empirical evidence could be grounded. As the title implies, Schaefer's article courted controversy in directly attacking the Kant–Hettner–Hartshorne view on what made geography 'exceptional.' A controversy not averted by Schaefer's death prior to the article appearing in print, his piercing academic critique was to pose a serious threat to the regional tradition. At one level, and despite ripostes from Hartshorne (1959) it effectively killed off traditional regional geography as an idiographic discipline. At another level, it argued that for geography to survive it must establish itself as a nomothetic (law-producing) discipline by using the hyptothetico-deductive scientific method.

The law-producing regional geography that emerged in the 1960s owed much to the 'locational school' and its alliance of economics with geography (and planning). In the United States, Walter Isard (1960) was the forerunner in constructing a 'regional science' to counter the a-spatial classical economic theory of the German location school (Johann Heinrich Von Thünen, Alfred Weber, August Lösch, Walter Christaller). His alternative approach sought to move beyond 'explaining' location simply by cataloging geographical factors to a more sophisticated 'analysis' that attempted to fuse these factors together in regional input–output techniques (linear programming). Regional science flourished in the 1950s and 1960s as a nomothetic discourse asserted itself as the mainstream methodological approach in Anglo-American geography. Allied to developments in the United Kingdom, most notably Peter Haggett's (1965) seminal book, *Locational Analysis in Human Geography*, the quest to identify general ('scientific') laws to explain spatial behavior led to a decline in intensely place-based (regional) studies in this period. Yet regions, and regional geography, survived because although regions were often deployed uncritically as spatial backdrops on which human activity took place, in this role they proved to be "one of the most logical and satisfactory ways of organizing geographic information" (Peter Haggett et al., 1977).

By the 1970s regional geography had been pushed to the margins of geography. One commentator went so far as to say that regional geographers should be considered the 'last of the handloom weavers' (John Paterson, 1974) – a group of diminishing relevance, too commonly found looking back to the halcyon days of traditional (regional) geography, and crucially, seen to have little or no contribution to make to a discipline increasingly captivated by developments emerging from a new generation of critical geographers and splintering into new subdisciplinary fields to incorporate new approaches to understanding the spatial construction of social processes. Indeed, despite John Fraser Hart (1982) hailing regional geography as "the highest form of the geographer's art" owing to the rich descriptive narrative a focus on traditional approaches facilitates, it was the emergence of radical Marxist geography and its critical perspectives on the geographies of uneven development that was to prove the torchbearer for a reinvigorated regional geography in the 1980s (see **Figure 1**).

Reinvigorating Regional Geography: The Emergence of 'New Regional Geography'

Regional geography experienced a renaissance in the 1980s owing to two principle developments. On one level, regional geography benefited from the failings of the positivist mode of explanation. Having effectively de-humanized geography by overlooking the active role people play in shaping social processes and practices, the criticism of the nomothetic paradigm in mainstream geography – alongside the crisis of 'grand narratives' of humankind (communism, religion, science) and 'grand theories' in the academic world – saw increased pressure placed on totalizing discourses and led to a reawakening of interest in the special/specific (idiographic). Affectionately referred to as 'the mother' of the new regional geography (Markusen, 1993), Doreen Massey (1984: 51) was one of a growing number of academics to point out how regional

specificity had been firmly reinstated on the methodological agenda:

> … [G]eography matters. The fact that processes take place over space, the facts of distance or closeness, of geographical variation between areas, of the individual character and meaning of specific places and regions – all these are essential to the operation of social processes themselves. Just as there are no purely spatial processes, neither are there any non-spatial social processes. … Geography … is not a constraint on a pre-existing non-geographical social and economic world. It is constitutive of that world.

On another level, a more critical scholarship emerged through the coming together of previously isolated academic approaches – the political-economic, humanistic-phenomenological, and structurationist/critical social theory (cf **Figure 1**) – each seeking common ground through a 'geography of regions.'

The political-economic approach is rooted in Marxist analyses of uneven 'regional' development. It focused on the ways in which regions were seen to be increasingly central to the shifting geographies of capitalism. Following the crisis of the nationally scaled Fordist mode of production and Keynesian regulatory system in the 1970s, the switching and displacement of the crisis tendencies of (global) capitalism into compensating regional configurations provided a significantly new perspective on the region (David Harvey, 1982). Regions were no longer considered to be just a spatial backdrop against which developments occurred but the specific sociospatial, historically contingent context that produces different pathways of capitalist development (Doreen Massey, 1984). Hence, regions are not fixed and easily mapped (as in regional science), but relative and constructed in different ways (in this case, according to the shifting economic geographies of capitalism).

The humanistic-phenomenological perspective adopted a similar position, albeit regions are relative and constructed based on attachment to particular places and regions. Harking back to the Vidalian tradition, regions are a focus of cultural identification (or 'sense of place,' Doreen Massey, 1991) and are constructed through identity formation. In direct opposition to the positivist tradition underlying regional science – where human activity was determined as rational and could therefore be predicted and planned for – the humanist approach placed the intentionally acting human being at the center of their academic inquiry in order to understand the attachment (and sense of belonging) people have to their spatial environment (place, region). Here the emphasis was again placed on understanding regions as historically contingent process, never static but always in a state of formation, reformation, and/or transformation (Allan Pred, 1984). Also important was the analytical distinction made between place, as personal experience structured through one's own individual biography, and region, which embodies a collective dimension of shared experience (Annsi Paasi, 1986).

Derived from Anthony Giddens' (1984) theory of structuration which sought to overcome the traditional binary division between 'structure' and 'agency,' the structurationist perspective is chiefly characterized by consideration of the region as both 'outcome' and 'producer' of social practice. Collectively, the new approaches to understanding the region were thus seen to differ significantly from traditional conceptions because society was included as a key agent in region formation. In short, claims of the birth of a 'new regional geography' were seen to be warranted as the traditional concern with the people–nature relation gave way to the triangular relations between people, society, and nature at this time (Anne Gilbert, 1988). What is more, with the individual and history at its core, this new reconstituted regional geography emphasized the need to understand how regions are socially constructed across space and time. Work in this area is perhaps best showcased by two distinct, yet related, approaches.

According to the Finnish geographer, Anssi Paasi (1986), explaining region formation requires an understanding of regional institutionalization. Regional institutionalization results from the coming together of historical, economic, political, and social processes to produce a specific (and always contested) cultural image. To understand this, Paasi argues that the institutionalization of regions is a process comprising of four interdependent stages: (1) the development of territorial shape (as practices and processes become localized and territorialized); (2) the formulation of conceptual or symbolic shape (e.g., flags, anthems) to establish and reproduce a regions identity; (3) the emergence of institutions (e.g., education, media) to deepen regional practices in politics (e.g., regional government), administration (e.g., regional offices), and culture (e.g., regional events); and, which lead to (4) the establishment of a region in the spatial structure and social consciousness of society.

The impact of this thinking also posed a more substantive methodological challenge for regional geographers. For as John Allen et al. (1998: 2) put it in the context of their own attempts to determine when was, and where is, the southeast region of England:

> Regional studies are always done for a purpose, with a specific view. Whether territorial, political, cultural or whatever, there is always a specific focus. One cannot study everything, and there are always multiple ways of seeing a place: there is no complete 'portrait of a region'. Moreover, 'regions' only exist in relation to particular criteria. They are not 'out there' waiting to be discovered, they are our (and others') constructions.

The new regional geography was therefore more complex, dynamic, and contested than its predecessor, characterized not by a singularity of approach but a plurality of different approaches, each seeking common ground through a (re) new(ed) geography of regions.

From the mid-1980s through to the early 1990s the new regional geography proved fertile terrain for a fairly intensive conceptual debate. Nevertheless the plurality in (conceptual, empirical, philosophical, methodological) approach made it difficult to sustain the new regional geography as a coherent academic endeavor. At one level, academic fatigue began to set in as scholars exhausted certain lines of inquiry and significant steps forward became less common. At another level, inherent shortcomings were being exposed. On the front line for criticism was the major research program on 'localities.'

Inspired by the work of Massey (1984), localities research set out to examine the economic consequences of global capitalist economic restructuring on specific places and regions.

Yet it was not localities as the outcome and producers of spatial processes which were to be understood but social interaction within the boundaries of predetermined study areas ('localities') that left localities research open to such piercing critique (*Antipode*, 1987–91). Add to this further criticism that localities research came dangerously close to returning regional geography to 'exceptionalism' (by systematically failing to engage theoretically or conceptually), never fully defined and grasped the usage of its key concept and term ('locality'), and focused too narrowly on economic factors alone to account for place formation; it is perhaps not surprising that in the face of growing criticism many of those scholars who came together in the 1980s to enthusiastically pursue new academic endeavors through a geography of regions dispersed to pursue other lines of inquiry in the early 1990s rather than undertake the renewal necessary to keep the new regional geography alive.

As the new regional geography started to wane, a new brand of regional geography emerged to reignite enthusiasm for a further shared academic endeavor – the 'new regionalism.'

Regional World(s): The Rise of New Regionalism

The new regionalism rose to prominence in the 1990s as regional geography's latest 'big idea.' Contrary to popular belief held throughout the 1970s and 1980s that globalization was bringing about deterritorialization and an era of global deconcentration (a 'borderless world'), the new regionalism came to represent the realization that for all that globalization was extending our capacity to interact more freely across space, geography still mattered. More substantively, regions were seen to challenge – even replace in some accounts – the nation-state as the prime territorial unit for organizing and structuring economic, social, and political activity in an era of (neo)liberal globalization. From this perspective not only did geography matter, it was a regional geography that appeared to matter most.

Despite interest in the new regionalism centering on a shared endeavor to explain the resurgence of regions in globalization, just like the new regional geography which preceded it, the new regionalism does not exhibit a uniform approach toward method. There are basically two empirical approaches, divided according to whether the research pursued was rooted in the 'geoeconomic' or 'geopolitical' logic for the reemergence of regions.

The Geoeconomics of Regions

One distinctive feature of the new regionalist debate is the realization that localized ('regional') agglomeration economies appeared to foster the conditions, harbor the assets, and generate the capacities upon which increasingly globalized, transnational capital accumulation depended. In this way it formed part of a broader economic development discourse synonymous with accounts documenting how, for all globalization was producing an era of global deconcentration – as capital was freed from the shackles of historically entrenched nationally scaled Keynesian regulatory systems – globalized capital accumulation was increasingly found crystallizing out on the ground in a select number of strategically important local urban and regional production complexes. First among many were the new industrial regions of Silicon Valley and Route 128 in the United States; and Baden-Württemberg, Catalonia, Lombardy; and the Rhône-Alpes, the so-called 'Four Motors of Europe.'

Bucking the trend of national economic decline these new industrial growth areas of the post-Fordist economy became poster-regions for ascendant 'new regionalist' orthodoxy. Constructed around a conjecture that there was something absolute that connected late twentieth-century capitalism to regionalism and regionalization, proponents of the new regionalism appropriated a strong geoeconomic logic for explaining why the post-Fordist growth dynamic was rooted in regional (not national) economies (Michael Storper, 1997). First, economic success is not exclusively related to a narrow set of economic factors but is dependent on social, cultural, and institutional structures. Second, these noneconomic factors see information sharing, collective learning, and the build-up of 'social capital' (e.g., trust, loyalty) replace market-based competition as the basis for economic success and competitiveness in an era of reflexive capitalism. And perhaps most crucially, third, regions (not nations) are the most appropriate scale for spatial clustering, networking, and interaction to take place.

After a fairly intensive period of theorization in the mid-1990s the geoeconomic logic underpinning the resurgence of regions began to capture the mainstream interest of political leaders and policy elites. This led to the formation of a new global policy discourse predicated on a new regionalist orthodoxy broadly claiming that contemporary capitalism and its territorial configurations are best regulated and governed in and through the decentralization of socioeconomic decision-making and associated policy implementation to regional level institutions, frameworks, and supports.

The Geopolitics of Regions

As fundamental to the new regionalist orthodoxy as the strongly appropriated geoeconomic logic is the geopolitical logic which accounts in part for this resurgence of regions in the latter decades of the twentieth century. At the same time a group of primarily economic geographers uncovered new successful forms of production in local industrial districts and selected regional economies which led them to reframe the geographies of the global economy, a coterie of political and cultural geographers (alongside a group of political scientists) observed how a reigniting of interest in nationalist and regionalist territorial identities, politics, and citizenship was reworking the established landscape of contemporary politics and cultural expression. Broadly speaking the geopolitics of regionalism derived from two sources. First, there was the reemergence and affirmation of traditional cultural insignia to distinguish one region from another in an era where American hegemony was seen to threaten erase traditional regionalist and nationalist identities and cultures from the political landscape. Meanwhile second, there was strong political reaction from within national states against the authority of national government's following the failure of the central state to adequately manage and protect the Fordist mode of capital accumulation and Keynesian regulatory system which sparked

calls for the transfer of power to economically advanced territories (e.g., Baden-Württemberg, southeast England) and/or historic territories with separatist movements (e.g., Catalonia, Lombardy, Quebec, Scotland).

The most striking characteristic of this is the way in which the need to distinguish regions from one another resonated with the very essence of the origins of ('traditional') regional geography as the study of what makes regions unique and distinctive. Nevertheless, the more substantive point was actually the placing of regional geography at the center of debates on a purported 'scalar turn' in social scientific thinking. The interest of geographers was arrested by claims the nation-state was being economically and politically challenged by the 'regional state' as a potential successor to its long-held authority as the site and scale at which the economy is managed, social welfare administered, and people are treated as citizens (Kenichi Ohmae, 1995).

The impact of this thinking was particularly evident in Europe, where the rise of regions and regionalism from the 1970s onwards prompted construction of the popular 'Europe of the Regions' discourse which by the early 1990s had advocates suggesting the formation of, and transition to, a federalized European superstate. Yet it was to be the more considered analyses produced at this time that recognized that regions were not replacing the nation-state per se, rather they were becoming competitive and strategic territories within an emerging multilevel system of capital accumulation and governance extending from the global to the local. What this gave rise to was a regional geography heavily focused on the study of state–region relations.

The Critique

There can be little doubt that the new regionalism proved a fertile terrain from which regional geographers were able to once more reassert the central importance of regions to the geographical discipline. But just as with each of regional geography's big ideas that went before, the new regionalism came under intense scrutiny when John Lovering's (1999: 380) contention that proponents of this orthodoxy were narrating an "attractive and persuasive story, but it is largely a fiction" laid bare its inadequacies. Chastising many of those engaged in new regionalist theorizing as guilty of bad abstraction, Lovering raised concern over the exclusive nature of the empirical referents used to construct new regionalist orthodoxy. What is more, in developing this line of argument Lovering was to demonstrate how many proponents of the new regionalism lost sight of the fundamental characteristic of regional geography ('specificity') in the hot pursuit of theorizing regional development as something which is generic and therefore lends itself to be understood in more abstract terms.

Rays of regional sunshine set against an increasingly overcast national economic outlook, on first view it is not surprising these regional industrial production complexes were lauded for spearheading the economic recovery, providing a template for achieving global competitiveness, and being drivers of the national economies within which they were situated. The 1990s was a time when the search for a new post-Fordist, post-Keynesian, post-national blueprint for organizing capitalist development was at its most fervent. That new regionalist accounts appeared to provide such a blueprint catapulted those engaged in new regionalist theorizing to the forefront of academic, but also increasingly political, debate. Captivated by accounts detailing the geoeconomic factors underpinning the growth of Silicon Valley, Baden-Württemberg, and Emilia Romagna, policy elites the world over where mobilized by the belief that the new regionalist consensus provided the policy recipe which would enable them to position their region(s) within national and international circuits of capital accumulation. Furthermore, policy elites mobilized with speed in the knowledge that not only did they need to intervene to create the conditions necessary for capital accumulation, they needed to make their region(s) more attractive to transnational capital than their national and international competitors to be competitive in today's quicksilver global economy.

The problem observed by Lovering was that sophisticated theorizing of actual economic developments and observable outcomes in these prototype regions were to be silenced by more speculative theorizing based on assumptions celebrating what they hoped might be possible in an idealistic scenario. In what he famously referred to as the "policy tail wagging the analytical dog," Lovering detailed how the ensuing period of intense institution-building became the focus of new regionalist theorizing, such that what started out as a warranted academic endeavor to understand the dynamics of regional development was lost to an unwarranted orthodoxy fueled and sustained not by on-the-ground realities, but because they resonated with policy elites who wished to explore the scope for policy interventions at the scale of regions. In a self-perpetuating way, successive rounds of policy intervention at the regional level were purported as evidence of the growing importance of regions as strategic sites in globalization, which in turn fueled the political appetite to explore the scope for further policy initiatives at the regional level.

Where Next for Regional Geography?

Explaining regional change remains one of the great challenges for the social sciences, and geography in particular. Yet as Nigel Thrift (1996) once observed "[o]nly a very few human geographers nowadays would lay claim to being 'regional geographers', and in quite recent times regional geography has been described as if it was in need of exhumation rather than resuscitation." Nearly two decades on regional geography continues to survive. Its prominence is sustained not simply by the constant renewal of ideas, approaches, and perspectives required for any discourse, paradigm, or subdiscipline to evolve and survive for a significant period of time but by a series of major reboots. The history of regional geography is thus dominated by periods where a new discourse or paradigm is first constructed, only to be challenged, damaged, deconstructed, dismantled, and condemned, before then being rebooted, revived, reconstructed, only to be rechallenged (…and so the cycle continues) and where the highs are very high and the lows appear very low.

Notwithstanding the longevity of regional geography within geographical and social scientific debate there remain a series of unresolved issues. To begin with there is the ontological status of its foundational concept and object of inquiry – the region.

Over the years, the region has been variously referred to as an 'embattled,' 'chaotic,' 'mysterious,' and 'vague' concept. The problem is that much work which is purporting to be regional geography fails to adequately define the region to which they are either referring, or are working with. Where this occurs the region is often assumed, thereby failing to recognize the 'multiple ways' of seeing a region (Allen et al., 1998).

This problem relates to another unresolved issue: regional geography does not exhibit a uniform approach to method. Work in regional geography can be categorized as adopting one of three approaches: historically reinterpreting regional developments which have taken place, interpreting regional developments taking place in the present, and projecting regional developments which will take place in the future. This contributes to methodological pluralism which is a defining characteristic of regional geography. More significant than this, however, is the relationship between theory and empirics. This has been at the root of all the episodic rise and fall in regional geography's long history. To illustrate this point we need only look back over the past 30 years at the locality debate and the new regionalism. Focusing exclusively on the locality and region respectively, the theoretical abstraction from empirical description of what appeared to be working in a number of locations often excluded important external factors – national and global political-economic trajectories, the changing role of the nation-state, transnational corporations – which clearly had a major impact on shaping local and regional development. What this points to is a broader challenge which regional geography constantly confronts: the interaction between empirical observations of regional difference ('description' on what is exceptional) and theoretical abstraction ('explanation' of what is typical).

Despite these unresolved issues, the prospects for regional geography ultimately hinge on the ability of its proponents to ensure regional studies remain socially relevant in our increasingly globalized modern world. Indeed, there are a number of new horizons in regional geography which are providing the context for the next generation of regional research.

Theoretically, one of the outstanding arguments centers on the 'regional world(s)' we live in and how best to capture the dynamics of contemporary regional development and change (Martin Jones and Anssi Paasi, 2013). Singular understanding of regions as territorially bounded, homogenous units have by and large been rejected and replaced by relational approaches that view regions as being constructed through wider, 'stretched' social relations. Although this contributed to a back and forth exchange between proponents and engaged critics of the relational approach over the last two decades, this is now being superseded by new conceptual interpretations which consider how different spatial dimensions – territorial, relational-networked, place-based, scalar – actually come together to form complex space-time ('regional') configurations. Singular understandings are thus being replaced by a more plural approach to understanding contemporary societies, with the fusion of the different approaches often animated through a renewed geography of the region.

Socially, there are also a series of turning points which are not only exerting pressures on global society but whose impacts are likely to be highly regionally uneven. The first turning point is the 2008 global financial crisis and ensuing economic crisis. This has already had a major short-term impact on regional development schemes following the withdrawal of credit, but longer-term the cycle of debt repayment and austerity packages will be a serious obstacle to growth and a major impact on public finances. Alongside this the effects of climate change and related pressures in terms of the availability of water, energy, and food are also likely to be highly regionally uneven. Given this, attention has been directed toward consideration of which regions are most/least resilient to these external pressures (Ron Martin, 2012).

Spatially, a deepening of global economic integration alongside accelerated processes of urbanization is having a major impact on the construction and configuration of regions and regional space. One consequence of this was witnessed at the beginning of this century when the new regionalism morphed into a more urban-centric discourse of 'city regionalism.' Purported to be pivotal, societal, and political-economic formations and signifiers of a new phase in capitalist territorial development, one of the dominant characteristics of this focus on globally significant major urban regions is the trend toward identifying ever larger regional spaces. Nowhere is this more evident than in the recent focus on 'megaregions' (Richard Florida et al., 2008) which raises a host of important questions relating to how far the regional concept can be stretched, how plausible it is to conduct empirical research at markedly different regional scales, and what (if anything) marks these spaces out as regional.

All of these unresolved issues and challenges pose searching questions that get to the root of what it means to do regional geography. But it also provides fertile terrain for contemporary regional geography to explore and develop.

See also: Region: Geographical; Regional Planning; Regional Science.

Bibliography

Allen, John, Massey, Doreen, Cochrane, Allan, 1998. Rethinking the Region. Routledge, London.

Fleure, Herbert John, 1919. Human regions. Scottish Geographical Magazine 35, 94–105.

Florida, Richard, Gulden, Tim, Mellander, Charlotta, 2008. The rise of the mega-region. Cambridge Journal of Regions, Economy and Society 1, 459–476.

Fraser Hart, John, 1982. The highest form of the geographer's art. Annals of the Association of American Geographers 72, 1–29.

Giddens, Anthony, 1984. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley.

Gilbert, Anne, 1988. The new regional geography in English and French-speaking countries. Progress in Human Geography 12, 208–228.

Gregory, Derek, 1978. Ideology, Science and Human Geography. Hutchinson, London.

Haggett, Peter, 1965. Locational Analysis in Human Geography. Edward Arnold, London.

Haggett, Peter, Cliff, Andrew, Frey, Allan, 1977. Locational Models. Edward Arnold, London.

Hartshorne, Richard, 1939. The Nature of Geography: A Critical Survey of Current Thought in Light of the Past. Association of American Geography, Lancaster.

Hartshorne, Richard, 1959. Perspective on the Nature of Geography. Association of American Geography, Lancaster.

Harvey, David, 1982. The Limits to Capital. Blackwell, Oxford.

Herbertson, Andrew, 1902. Man and His Work. A. and C. Black, London.

Hettner, Alfred, 1927. Die Geographie: ihre Geschichte, ihr Wesen und ihre Methoden. F. Hirt, Breslau.
Isard, Walter, 1960. Methods of Regional Analysis: An Introduction to Regional Science. Wiley, Cambridge.
Jones, Martin, Paasi, Anssi, 2013. Regional world(s): advancing the geography of regions. Regional Studies 47, 1–5.
Kraft, Viktor, 1929. Die Geographie als Wissenschaft. In: Kraft, Viktor, Lampe, Felix (Eds.), Methodenlehre der Geographie. Franz Deuticke, Leipzig, pp. 1–22.
Lovering, John, 1999. Theory led by policy: the inadequacies of the 'new regionalism' (illustrated from the case of Wales). International Journal of Urban and Regional Research 23, 379–395.
Mackinder, Halford, 1902. Britain and the British Seas. Heinemann, London.
Markusen, Ann, 1993. Classics in human geography revisited – Massey, D. 1984 Spatial divisions of labour. Progress in Human Geography 17, 71–72.
Martin, Ron, 2012. Regional economic resilience, hysteresis and recessionary shocks. Journal of Economic Geography 12, 1–32.
Massey, Doreen, 1984. Spatial Divisions of Labour: Social Structures and the Geography of Production. Macmillan, Basingstoke.
Massey, Doreen, 1991. A global sense of place. Marxism Today 35, 24–29.
Ohmae, Kenichi, 1995. The End of the Nation-State: The Rise of Regional Economies. Harper Collins, London.
Paasi, Anssi, 1986. The institutionalization of regions: a theoretical framework for understanding the emergence of regions and the constitution of regional identity. Fennia 164, 105–146.
Paterson, John, 1974. Writing regional geography: some problems and progress in the Anglo-American realm. Progress in Geography 6, 1–26.
Peet, Richard, 1998. Modern Geographical Thought. Blackwell, Oxford.
Pred, Allan, 1984. Place as historically contingent process: structuration and the time-geography of becoming places. Annals of the Association of American Geographers 74, 279–297.
Schaefer, Fred K., 1953. Exceptionalism in geography: a methodological examination. Annals of the Association of American Geographers 43, 226–249.
Storper, Michael, 1997. The Regional World: Territorial Development in a Global Economy. Guildford Press, New York.
Thrift, Nigel, 1996. Taking aim at the heart of the region. In: Gregory, Derek, Martin, Ron, Smith, Graham (Eds.), Human Geography: Society, Space and Social Science. University of Minnesota Press, Minneapolis, pp. 200–231.

Regional Planning

Alex Lord, University of Liverpool, Liverpool, UK
Mark Tewdwr-Jones, Newcastle University, Newcastle-upon-Tyne, Tyne and Wear, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The case for regional planning turns on the logic that a strategic overview is required to deal effectively with urban and environmental issues that have wide implications. This includes, but is not limited to, transport planning, the management of environmental assets and the location of housing developments. In this article, the authors set out the history of regional planning, contrasting the ways in which the term has been employed in diverse national settings before going on to chronicle the suite of activities that have at various times comprised the activity of regional planning. The authors conclude by taking stock of the position occupied by regional planning in some states at the end of the first decade of the new millennium and ask what turn regional planning will take next.

Introduction

At root, regional planning is about situating the decision-making processes relating to traditional planning activities – transport, housing, environmental impact assessment, and so on – at a strategic scale. In this sense, the 'regional' of regional planning is a byword for a scale of governance at which a wider, more holistic approach can be taken to the management of the urban and natural environments. However, discussing regional planning is problematized by ambiguities in this concept at the core of the expression: what is meant by a 'region' depends fundamentally upon where you are. To the US-ear the idea perhaps speaks most clearly to the idea of governance in or for a metropolitan region (Greater Portland, for example), although this should be clarified to include the origination in the North American context (Gottman, 1961) of the mega region, which might itself be constituted by smaller regional territories not necessarily tied to one specific metropolis. By contrast, in Europe the region is more usually understood to be a constituent subnational territory of varying political status: sometimes a systemically formal component of a federal polity (such as the German *Länder*); occasionally a semiformal territory as with the regional state apparatus devolved to the English regions by the British government in the late 1990s and 2000s; or an informal, tacit recognition in national policy making of centuries-old regional identities built on differences of language, culture, identity, and politics.

Building from this patchwork of various regionalisms nuanced accounts have resulted in an understanding in the West of fluid subnational spaces that might, depending on context, be described as regions, city-regions, or polycentric regions (Davoudi, 2003). In parallel, a different vernacular has emerged elsewhere with the 'mega city' (Hall, 1999; see also, Lin and Ma, 1994; McGee and Robinson, 1995) describing in China and India (see the article 'Regional Planning: China in this encyclopedia) what would probably be thought of as akin to a metropolitan mega region (in the United States) or a European city-region – albeit often on a much grander scale. These differences of scale and size add further complexity to the issue with attempts at creating statistically standardised regional entities, such as the NUTS-3 classification popular in continental Europe, grating against wildly different qualitative scalar understandings of what a region might be ranging from, for example, small Japanese prefectures to the huge oblasts of Russia.

Whilst this discussion of international variation in nomenclature reveals semantic differences perhaps more importantly it illustrates the global significance of a territorial scale, whatever we might call it, between the nation-state and city hall at which academics, policy makers, and business leaders have found it sensible to think about the organization of social, economic, and political life with corresponding implications for the natural and built environments. As a consequence, the vast majority of national governments have at some point in their history chosen to include a regional tier of strategic planning as part of their approach to urban and environmental management: although with differing remits and intentions for what regional planning should seek to achieve from time to time and place to place. In this piece, we set out to provide an overview of this evolution of regional planning in both theory and practice before asking what turns regional planning might take next.

Regional Planning in Historical Focus

Although the practice of acknowledging and responding to regional differences as part of the act of governing heterogeneous societies can be traced back to the ancient world (the Roman Empire being a good example, see Nicolet, 1991), Wannop and Cherry (1994) date the origins of regional planning as a component of a modern statutory planning system to the more prosaic setting offered by South Wales in the United Kingdom. Here, in 1921, a government committee was appointed to consider housing, transport, and the provision of utilities to the area's large coalfields. The committee's report advocated considering the area to be what one might now understand to be a functional economic region comprising 15 individual administrative units. While the proposal was rejected by central government, the idea – the germ of which can be found earlier in Howard (1902) and Geddes' (1915) great early statements of town and

country planning – took hold and can be seen as the genesis of a form of planning for metropolitan regions in cities such as Manchester, Birmingham, and London in the 1920s. The resultant Joint Town Planning Committees as the name implies acted as strategic planning bodies for these cities together with their hinterlands (Hall, 1974).

This concept of planning for metropolitan regions established in the 1920s was reinvigorated following the end of the Second World War and formed a central plank in Britain's postwar reconstruction with many plans from the period taking on an explicitly metropolitan-regional dimension, good examples being the Clyde Valley Plan (1946) and the Greater London Plan (1944). Parallel processes can also be identified in the emergence of strategic regional planning in other Western European contexts at the same time, for example, France (Boudeville, 1966; see also Hudson and Lewis, 1982) and Germany (Brenner, 1997) where the federal system of government invested regions with formal political power and responsibility for developing regional planning policy. Furthermore, the flavor of European regional planning in the aftermath of war came to reflect the politics of the time effectively providing a territorial counterpart to a broad societal consensus built on the twin pillars of social democratic politics and Keynesian macroeconomic policy. A focus on spatially balanced growth, planning new settlements, and demand stimulation in less prosperous regions came to be a core feature of postwar regional planning in Britain and its continental neighbors.

In turn, the diffusion of the concept from Britain to other countries in the Commonwealth such as Canada (Cullingworth, 1987) and Australia (Collits, 2012; Freestone, 2012) gave the idea of regional planning wider international currency. However, the application of 'regional planning' in these settings served primarily to illustrate the malleability of the concept. When overlaid onto the federal system of government in these much larger, more sparsely populated nations, the idea of planning at a regional scale came to inhabit multiple meanings bound up with contextual differences of governance, urban–rural interdependencies/rivalries, and resource allocation. Moreover, the scale at which this form of regional planning took place often spanned territories that dwarfed the original European understanding of what a 'region' could be. The result was a qualitatively different understanding of how regional planning should proceed even between nations in which regions enjoyed the privileged position of being a formal state space, although the kernel of the idea that 'standing back' and taking a strategic, long view remained common.

Elsewhere, this same logic of aligning regional with 'strategic' planning was at work but with quite different effects. Rather than regions representing an amplification of established planning activities overlaid onto preexisting, fractured local authority geographies, in contexts such as the US policy makers looked beneath the formal tier of government offered by the federal state to explore new spatial configurations. Often, the result was the identification of a mega region, such as Southern California, encompassing many cities together with hinterlands boasting important environmental assets. Sometimes, these mega regions subsequently came to be understood as having constituent functional regions within them – for example, 'coastal' and 'desert' in the case of Southern California. The corresponding emergence of planning agencies for these mega regions grew up in parallel to the first identification of this scale as of strategic importance in the early 1960s. For example, again using the same example from the United States, the Southern California Association of Governments has, since 1965, sought to provide a coordinated governmental response to planning issues, such as transport, across a huge territory at the time of writing now encompassing six counties and 191 cities (SCAG, 2013). Moreover, there is evidence to suggest that the identification of US mega regions and the emergence of a state architecture to represent them has been followed by a nascent form of mega-regional identity; for instance, the use of 'SoCal'/'NoCal' to denote an identification with, respectively, Southern or Northern California, has permeated the popular consciousness and is routinely used as a commercial branding device (of clothing, bumper stickers etc.) on the West Coast of the United States.

Academic Interest in Regional Planning and Its Challenges

The resulting impression is that regional planning, as a highly variegated and context-specific set of practices across the globe, is probably best understood as a child of the early twentieth century, symbiotically connected to the rapid and profound social changes brought about by the innovations of that century: mass automobile ownership, commercial aviation, and the 'agricultural revolution.' In the academic world, the differing territorial effects of these wider social and economic trends spawned interest in the challenge of coordinating agencies tied to what had fast become antiquated administrative boundaries and providing an even level of service provision to cities with rapidly increasing urban populations and porous boundaries. Some of these issues had earlier been forecast by the forefathers of planning as being an academic discipline. In this respect, Geddes' (1915) *Cities in Evolution* is often credited with awakening academics to the analysis of the wider (literally, in a spatial sense) implications of social, environmental, and economic changes brought about by Western urbanization (for a review see Glasson and Marshall, 2007). Seminal contributions followed (e.g., Friedman and Alonso, 1964), which in turn spawned a welter of academic interest on all aspects of planning viewed through the prism offered by the region to the extent that by 1979 it was said that "regional planning has come of age" (Friedman and Weaver, 1979: p. 1).

In retrospect, these claims of intellectual maturity may have been precipitous. Glasson (1992: p. 505) identifies a subsequent wane in interest in regional planning from around the time Friedman and Weaver (1979) date its ascension: "from the late 1970s to the start of the 1990s, regional planning was in the doldrums." This decline is perhaps best explained by a shift in academic focus in two directions – to the 'world' or 'global' cities of, for example, London, Los Angeles, New York, Paris, Tokyo (Knox and Taylor, 1995; Sassen, 2001), and the 'shrinking cities' (e.g., Detroit, Leipzig, and Manchester) where postindustrial economic restructuring had left the scars of population loss, unemployment, and social atomization (e.g., Bontje, 2004; Mace et al., 2005; Ryan, 2008). Subsequent years saw academic interest in regions return partly as a result of the

effects of economic globalization and the apparent importance placed by some of the largest transnational corporations on regionally significant variables – such as concentrations of specialist labor – in making their location decisions. The spatial effects of these economic processes were said to be a "pulling up, pushing down, squeezing sideways" (Giddens, 1998: p. 13) of political power that fundamentally 'hollowed out' (Rhodes, 1994) the capacity of central governments to act and ultimately, it was thought, spelt the 'end of the nation-state' (Ohmae, 1996) propelling the region into the position of being the most significant spatial scale at which to make and deliver policy.

The recurring lesson to be taken from much of the work that serves to catalog and chronicle the ideas at the core of the new regionalism is that steadfastly holding to the neat administrative boundaries of local government often inherited from a period outside living memory, and obeying the rigidly defined responsibilities of what are in fact thematically overlapping public service providers results in disjointed and piecemeal planning policy. By extension, dealing with the inherently 'joined-up' nature of the problems with which regional planning engages necessitates cooperation and collaboration among the agencies that have individual responsibilities that might sensibly be aggregated. Stated alternatively:

> There are many different understandings of what regional planning comprises … because of its political roots, regional planning arises where established administrative and planning systems are overtaken by problems they cannot resolve … regional planning and governance can never be perfectly arranged, except in the moment. But since regional issues are recurrent, regional planning and governance are essential arts in continually managing a constantly imperfect system of government. (Wannop, 1995: p. 368)

From this point of view, regional planning is strategic in the sense that it serves to choreograph activities that cut across administrative geographies and institutional lines of accountability. In so doing, the intention is to provide a forward-looking, long-term framework that recognizes interlinkages between and across functional territories: for example, urban–rural interdependencies and transport planning related to extended commuting patterns. Ideally, planning at this strategic scale should, therefore, be better placed to objectively mediate where demand for resources and development pressures conflict with concerns over environmental conservation in a way that prevents local interests, such as opposition to new housing or infrastructure, overriding decision making in the wider public interest. Furthermore, in some contexts where an environmental feature, such as a mountain range, river, coastal waters, or lake, spans several administrative territories regional strategic planning can aid in supporting consistent management of such assets.

However, orchestrating this form of multiagency cooperation over what are by definition large territories has often proved difficult to achieve in practice. Indeed, regional planning's situation between the national (or federal) state and the geography of local government means that it systemically occupies the invidious position of being an intermediary, ideally providing a consistent translation of national policies downward and simultaneously representing regional and local interests upward to inform national policy debates. Research would suggest that this dual accountability coupled with the demands of coordinating the activities of multiple agencies across wide regional territories in a strategic fashion has been very demanding in practice – although by no means impossible (e.g., Glasson et al., 1997; Montgomery, 2011).

In sum, it seems clear that, with caveats about contextual specificities, the concept of regional planning either at a formal regional scale of government or through less formal partnership working across metropolitan regions came to be a relatively stable and enduring feature of planning practice across much of the Western world during the 1950s and 1960s and, in places, well beyond. For example, the previously referred to Clyde Valley Plan of 1946 remained in place until 1982. However, the continued conceptual relevance of regional planning in its original form subsided in light of the social and economic upheavals of the 1970s and 1980s. The stark differentiation between rapidly growing and shrinking urban settlements that characterized this period prompted a host of questions regarding the appropriateness of the metropolitan regional concept of regional planning dating from the 1920s to the1940s. At one extreme, population loss and urban shrinkage as a consequence of industrial collapse militated against the a priori assumption of a fixed or growing metropolis of sufficient scale to underpin a related regional form of politics or a genuine regional identity. In these areas, regional issues were now primarily related to urban regeneration, stimulating economic growth, repopulation, and, in some cases, managing decline. Conversely, exponential growth in other settings created suburban sprawl, the effective merging of neighboring cities into new ever larger regions and the emergence of significant functional economic geographies occupying a position beneath that of the state or nation-state and often representing new spatial configurations starkly at odds with existing political geographies: examples include Silicon Valley in California and Emilia Romagna in Italy.

The Region as a Space of Economic Development

This contrast in fortunes between economically buoyant regions with an acute exposure to the high growth sectors of the global economy (London and England's South East; New York and the tristate area) and those struggling to adjust to the postindustrial turn taken by the global economy pointed to the importance of the region as an economic space. From this starting point, policy makers' attention in the 1990s and 2000s turned to mapping the functional geographies of regional competitiveness in an attempt to catalog assets (airports, universities, and businesses) and identify synergies between them that might support a multiagential strategy for regional economic growth (Etzkowitz and Leydesdorff, 2000). In an academic sense, this understanding of the region as an economic container represents a continuation of a well-established trend among a community of academics working from the 1950s onward in the field of Regional Science. A melange of human geography, economics, and social statistics, regional science continues to be a thriving intellectual community using innovations in, for example, spatial econometrics to further our understanding of the global economy's disaggregation into subnational territories. Often,

the results of research in this field illustrate the anachronistic nature of administrative boundaries and demonstrate the complexities of functional economic geographies that often straddle local boundaries, state lines, and even national borders. Outside the academy, the impact of work of this type has been significant in influencing policy makers to focus attention on planning at a regional scale, even if this understanding of regions turns on an ostensibly apolitical conception of what a region might be in which functional and administrative space are not coterminous. Among the best examples of this stripe of academic work influencing policy can be found in the creation of regional development agencies in England at the end of the 1990s responsible for the development of regional economic strategies, the supervision of regionally significant infrastructure investment, and, more widely, representing their region in the global competition for resources (Deas and Ward, 1999; Webb and Collis, 2000).

Across Western Europe, similar steps were taken to institute regional development bodies (Haiker and Danson, 1997). In many cases, the creation of these institutions was cast as a statutory response to fundamental shifts in the global economy. On one level, the growing significance attached to regions by decision makers in transnational corporations regarding the location of their operations was said to necessitate a corresponding statutory presence at this scale to compete for internationally footloose capital – a form of regional planning bespoke to the neoliberal age of territorial competition. A separate economic logic, built upon the foundations of agglomeration and institutional economics, was also used to encourage a form of regional planning designed to harness the endogenous capacity of functional economic regions themselves (Porter, 1996).

From this perspective, regional planning in the 1990s and thereafter can be seen in a fundamentally different light to that which preceded it. The original concept, rooted in the postwar European ethos of municipal socialism, has usually been understood to have begun from the idea that taking a wider, strategic view was essential in creating plans with welfare objectives at their core. The results of this style of planning – large-scale public housing building programs in the 1950s and 1960s, for example – were often developed with social objectives to the fore and with only passing regard for what an untrammeled free market might have chosen to provide. By stark contrast, the regional planning of the 1990s and thereafter has focused largely on seeking to accommodate and coordinate the preferences of the free market. For example, rather than orchestrating state-led house-building programs, one of the principal responsibilities of the regional development agencies operating across Western Europe at this time has been principally to determine the location of residential development and infrastructure, in virtually all cases the bricks and mortar have been delivered by the private sector.

The Future: The Untimely Death of Regional Planning?

This implied shift in the suite of activities that might be undertaken in the name of regional planning has led to a fundamental reconsideration of the value of having such a tier of statutory planning in some contexts where right of center governments hold office. The palpable shifts in state–market–civil relations that have come to characterize the neoliberal age pose serious questions for an activity that in many settings has seen its role diminished from direct service provider to that of juggling the investment decisions and preferences of both private companies and individuals into a broadly coherent regional pattern of development. In some cases, these questions have manifested themselves as a backlash against both the very idea of 'regions' and the institutions once situated at that scale to provide a strategic planning function. Probably the best example of this at the time of writing is from Britain where the Conservative–Liberal Democrat coalition elected in 2010 made the reform of the English system of regional planning one of its first priorities. In the first instance, this impulse to reform was discursive with all references to 'regions' (or derivations thereof) expunged from official documentation and a promise that:

> Communities will no longer have to endure the previous government's failed Soviet tractor style top-down planning targets … I promised to get rid of them and today I'm revoking regional plans with immediate effect - hammering another nail in the coffin of unwanted and an unaccountable regional bureaucracy. They were a national disaster that robbed local people of their democratic voice, alienating them and entrenching opposition against new development.
>
> I want to make something very clear. Localism means much more than a tug of war of political power between Whitehall and the Town Halls. It's a fundamental shake up of the balance of power in this country. So power goes right back to the people who elected us. (Pickles, 2010)

The subsequent dismantling of England's regional state apparatus and its replacement with an uncoordinated bottom-up 'localism' focused on the smallest territorial scale (save the household), the 'neighborhood,' clearly points to a rejection of the idea that strategic planning at a higher territorial level is in some sense necessary to coordinate interventions across local authority boundaries (Baker and Wong, 2013). By contrast in other contexts, the region remains a fundamentally important tier of government, the strategic significance of which has become more rather than less important particularly in relation to confronting the most recently identified grand challenges of the twenty-first century. In this respect, adaptation to climate change, resource management, and planning for energy and food security are all features that are said to require the strategic scale of regional planning. Moreover, evidence would suggest that the earliest and most comprehensive plans designed to meet the local effects of these global challenges have come from contexts where some form of regional planning is an established part of the governing framework (e.g., Juhola and Westerhoff, 2011).

Given the enormous scale of the issues that our societies are likely to face as the twenty-first century continues, it is difficult to imagine how a comprehensive response could come through anything other than urban and environmental planning organized at a strategic scale. In small nations, this may be possible through centralization of some planning functions at the national level. However, in most cases, the region (widely construed) offers the most likely scale at which this form of strategic planning could be animated. In sum, for so long as

there are planning issues that cut across the administrative geography of local government and require a coordinated response regional planning is likely to remain a prominent feature of the planning polity across the globe.

See also: Development and Urbanization; Environmental Planning; Governance; Regional Planning: China; Urban Geography; Urban History; Urban Megaprojects; Urban Planning: Methods and Technologies; Urban Policy in Europe; Urban Policy in North America; Urbanization.

Bibliography

Baker, M., Wong, C., 2013. The delusion of strategic spatial planning: what's left after the labour government's english regional experiment? Planning Practice and Research 28 (1), 83–103.

Bontje, M., 2004. Facing the challenge of shrinking cities in East Germany: the case of Leipzig. Geojournal 61, 13–21.

Boudeville, J.R., 1966. Problems of Regional Economic Planning. University of Edinburgh Press, Edinburgh.

Brenner, N., 1997. State territorial restructuring and the production of spatial scale: urban and regional planning in the Federal Republic of Germany, 1960–1990. Political Geography 16 (4), 273–306.

Collits, P., 2012. Planning for regions (chapter 9). In: Thompson, S., Maginn, P. (Eds.), Planning Australia. Cambridge University Press, Cambridge, pp. 204–220.

Cullingworth, B., 1987. Urban and Regional Planning in Canada. Transaction Press, New Brunswick.

Davoudi, S., 2003. Polycentricity in European spatial planning: from an analytical tool to a normative agenda. European Planning Studies 11, 979–999.

Deas, I., Ward, K., 1999. The song has ended but the melody lingers: regional development agencies and the lessons of the urban development experiment. Local Economy 14 (2), 114–132.

Etzkowitz, H., Leydesdorff, L., 2000. The dynamics of innovation: from national systems and "mode 2" to a triple helix of university-industry-government relations. Research Policy 29 (2), 109–123.

Freestone, R., 2012. An historical perspective (chapter 4). In: Thompson, S., Maginn, P. (Eds.), Planning Australia. Cambridge University Press, Cambridge, pp. 73–90.

Friedman, J., Alonso, W., 1964. Regional Development and Planning. MIT Press, Cambridge, MA.

Friedman, J., Weaver, C., 1979. Territory and Function. The Evolution of Regional Planning. University of California Press, Berkeley.

Geddes, P., 1915. Cities in Evolution. Williams and Norgate, London.

Giddens, A., 1998. Globalization: The Runaway World. Keynote Address, Prix Latsis Universitaires. Available at: http://www.fondationlatsis.org/plpdf/guest_speakers/Guest_1998.pdf (accessed 27.06.13.).

Glasson, J., 1992. The fall and rise of regional planning in the economically advanced nations. Urban Studies 29 (3–4), 505–531.

Glasson, J., Jennings, S., Wood, D., 1997. Regional planning in Western Australia: new initiatives, but old problems? Urban Policy and Research 15 (1), 31–41.

Glasson, J., Marshall, T., 2007. Regional Planning. Routledege, Oxford.

Gottman, J., 1961. Megalopolis: the Urbanized Northeastern Seaboard of the United States. Twentieth Century Fund, New York.

Haiker, H., Danson, M., 1997. Regional Development Agencies in Western Europe: a survey of key characteristics and trends. European Urban and Regional Studies 4 (3), 243–256.

Hall, P., 1974. The containment of Urban England. The Geographical Journal 140 (3), 386–408.

Hall, P., 1999. Planning for the mega-city: a new Eastern Asian urban form? In: Brotchie, J., Newton, P., Hall, P., Dickey, J. (Eds.), East–West Perspectives on 21st Century Urban Development: Sustainable Eastern and Western Cities in the New Millennium. Ashgate, Aldershot, pp. 3–36.

Howard, E., 1902. Garden Cities of Tomorrow. Swan Sonnenschein & Co, London.

Hudson, R., Lewis, J.R., 1982. Regional Planning in Europe. Pion, London.

Juhola, S., Westerhoff, L., 2011. Challenges of adaptation to climate change across multiple scales: a case study of network governance in two European countries. Environmental Science and Policy 14, 239–247.

Knox, P.L., Taylor, P.J. (Eds.), 1995. World Cities in a World System. Cambridge University Press, Cambridge.

Lin, G.C.S., Ma, L.J.C., 1994. The role of towns in Chinese regional development – the case of Guangdong Province. International Regional Science Review 1, 75–97.

Mace, A., Gallent, N., Hall, P., Porsch, L., Braun, R., Pfeiffer, U., 2005. Shrinking to Grow: The Urban Regeneration Challenge in Leipzig and Manchester. Institute of Community Studies, London.

McGee, T.G., Robinson, I. (Eds.), 1995. The Mega-urban Regions of Southeast Asia. University of British Columbia Press, Vancouver, British Columbia.

Montgomery, C. (Ed.), 2011. Regional Planning for a Sustainable America. Rutgers University Press, Piscataway, NJ.

Nicolet, C., 1991. Space, Geography and Politics in the Roman Empire. University of Michigan Press, Ann Arbor.

Ohmae, K., 1996. The End of the Nation State. Harper Collins, London.

Pickles, E., 2010. Making the Planning System Work More Efficiently and Effectively. Press release, 6 July 2010. Department for Communities and Local Government, London. Available at: https://www.gov.uk/government/news/eric-pickles-puts-stop-to-flawed-regional-strategies-today (accessed 27.06.13.).

Porter, M.E., 1996. Competitive advantage, agglomeration economies and regional policy. International Regional Science Review 19 (1–2), 85–94.

Rhodes, R.A.W., 1994. The hollowing out of the state: the changing nature of the public service in Britain. The Political Quarterly 65 (2), 138–151.

Ryan, B.D., 2008. The restructuring of Detroit: city block form change in a shrinking city 1900–2000. Urban Design International 13, 156–168.

Sassen, S., 2001. The Global City: London, New York, Tokyo. Princeton University Press, Princeton, NJ.

Southern California Association of Governments (SCAG), 2013. Cities, Counties and Sub-regions. SCAG, Los Angeles. Available at: http://www.scag.ca.gov/region/index.htm (accessed 27.06.13.).

Wannop, U., 1995. The regional imperative: regional planning and governance in Britain, Europe and the United States. Jessica Kingsley, London.

Wannop, U., Cherry, G.E., 1994. The development of regional planning in the United Kingdom. Planning Perspectives 9, 29–60.

Webb, D., Collis, C., 2000. Regional development agencies and the 'new regionalism' in England. Regional Studies 34 (9), 857–864.

Regional Planning: China

Cecilia Wong, School of Environment, Education and Development, the University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article highlights the nature and purpose of regional planning and how the concept has evolved when applied in developing countries as regional development planning. It then turns to discuss the knowledge and skills required in the development of the doctrine of regional planning and their relevance to regional planning practice in the development contexts. Regional development planning and its evolution in China is then used as the case study to illustrate the complex relationships between technical rationality and political reality in delivery of different forms of regional planning.

The Introduction of Regional Planning

The concept of regional planning was first introduced by Patrick Geddes who began to develop his broader philosophy of regional planning in his 1915 seminal book, *Cities in Evolution: An introduction to Town Planning Movement and the Study of Civics*. He put forward a synoptic vision of 'place, work, and folk' to explain the symbiotic relationship between economic and sociocultural well-being. Through his planning work of the Edinburgh Zoological Garden, he illustrated the complex interactions between human systems and the natural environment, which led to his idea of developing regional planning models to respond to these complex conditions (Thompson, 2004). To him, the spatial contexts of past and present were critical to planning as city development should be seen as part of a wider spatial network of the city and the surrounding countryside.

Sparked by the ideas of Geddes, Lewis Mumford advanced a conception of regional development by taking an organic, ecological planning approach that aimed to balance the needs of the social world with those of the natural ecosystem. This regional approach to planning led to the establishment of the Regional Planning Association of America and the 'regional city' idea was advocated by Mumford, Clarence Stein, and other members in the 1920s (Meyers, 1998). The notion of a planned, polycentric city was then born, which emphasized on the importance of antisprawl and that individual settlements should be both distinct and interconnected at the regional scale (Talen, 2008).

The term 'regional planning' consists of the spatial entity of 'region' and the practice of 'planning.' The notion of a 'region' varies widely in different contexts and encompasses a whole variety of regions that may cover a third of China to a very tiny region in England. Harvey Perloff (1968) pointed out that regional planning had developed along pragmatic lines with relatively little attention to formal theory and defined it as

> the 'ordering' of activities and facilities in space at a scale greater than a single community and less than a nation, or, in the case of a common market situation where a number of nations are economically integrated, less than that of integrated totality.
>
> (Perloff, 1968: 153)

He also differentiated regional planning that were intended to strengthen national planning from regional planning that has specific objectives and operating components. He then settled on the fact that 'we must deal with a varied assortment of regions which are subsystems of a larger whole and change with the whole over time in both internal structure and external relations' (Perloff, 1968: 153).

Several decades later, Seltzer and Carbonell (2011: 9) view regional planning with similar flexibility: 'rather than encompass a territorial interest that is sharply bounded and represented (or at least claimed) by an institution that holds general governmental powers and responsibilities, it is carried out over a territory that, at least when that planning takes place, is found to share enough characteristics to assert its existence as a region.' Their definition emphasizes on how regional planning should be deployed to address cross-boundary issues across jurisdictions, which differs from the earlier concerns of the size and spatial extent of regions.

The designation of a region, in planning terms, can be based on its administrative or functional boundaries and is likely to include a network or a hierarchy of urban centers and urban and rural activities that have strong linkages and functional ties. The fluidity of the boundaries of a region asserts the crucial importance of interaction. The nature and intensity of interactions depend on what social–economic–physical activities are involved in the functional process (Hall et al., 1973; Parr, 2005; Wong et al., 2006). In this sense, the dynamic interaction between different activities over different spatial areas makes it difficult to delimit a region within a tightly bounded, single-minded physical framework.

Urbanization Challenges and Regional Development Planning

According to the UN Habitat (2012a), the amount of urban land will have to be expanded by 326% in developing countries to accommodate the projected growth of urban dwellers from 2 billion to 4 billion between 2000 and 2050. As highlighted by its Executive Director, Dr Joan Clos, 'Among the greatest challenges of the twenty-first century is the rapid growth of cities … Growing cities and towns face additional challenges, which include: high percentages of people living in slums; expansion and dominance of the informal sector; inadequate urban basic services, especially water, sanitation and energy; unplanned peri-urban expansion; social and political conflict over land resources; high levels of vulnerability to natural disasters; and poor mobility systems' (UN Habitat, 2012a: 3).

Regional planning in the global south context tends to focus on managing challenges brought by the rapid urbanization process

and reducing the chronic spatial inequalities between rural and urban areas. The UN Habitat (1997: ii) strongly promoted regional development planning as a framework for decentralizing the urbanization process and for the development of balanced national urban hierarchies. The most powerful change in the urban system has been continuing spatial decentralization. In the developing world, a defining feature of cities 'is an outward expansion far beyond formal administrative boundaries, largely propelled by the use of the automobile and land speculation… that urban land cover grew, on average, more than double the growth of the urban population' (UN Habitat, 2012b: 28). However, recent UN Habitat reports (2012a,b) have focused the discourse on urban planning and city management. This position has shifted somewhat from its 1997 report of adopting regional development planning as the strategy to manage the dynamic and continuous process of spatial change, to strengthen the linkages between secondary and intermediate cities, small towns and rural service centers, and to develop new growth centers.

The focus on urban development will easily distort the spatial dynamics operating between a city and its wider spatial context. It will depart from the strategic perspective of regional planning that 'was intended to reduce, and in the long run to eliminate, major inequalities of income among regions,' and 'striving for, are a growing awareness of the decisive role of natural resources in sustaining civilized life; a new ecological ethic; greater concern with questions of equity; a deeper understanding of the contradictions between the international division of labor and the territorial aspirations of new nation states; and a reassertion of the principle of self-reliance at all levels of territorial order' (Friedmann and Weaver, 1979: 2).

Planning in the development context tends to take place at the macro national level and the local level and pays little attention to the needs for coordination and mutual adjustment. Regional planning is, therefore, seen as a structure that can help to link the development potentials identified in local plans and initiative to national and sectoral planning (Van Staveren and van Dusseldorp, 1983: 1) by either disaggregating the national development plan and sectoral plans or aggregating local plans, programs, and projects into comprehensive regional development plans (Ahmad and Bajwa, 2005). With the mounting pressure of urbanization and the need to address short-term economic benefits of urban development and the longer term socioenvironmental costs, the concept of development has been broadened and planning has to be more comprehensive (UN Habitat, 2012b). According to Ahmad and Bajwa (2005), countries with centrally planned economies have gradually adopted regional territorial planning to complement their national sectoral planning, whereas countries relying on market mechanisms have also extend their 'master planning' from city proper to the wider urban scales of greater city areas, metropolitan region, and even megalopolis. Both these observed developments can be found in the case of China (to be discussed later in the article).

The Doctrine of Regional Planning

Since a region is very much seen as an open-ended subsystem, the planning of which has to accomplish the objectives of enhancing inter and intraregional development as well as managing spatial organization, the understanding of its wider spatial contexts such as the spatial flows and linkages of investment, people, goods, and other factors is critical to the development of a regional plan (Perloff, 1968; Rodwin, 1978). This means that effective development of regional planning has created a demand for certain types of professional skills and knowledge. As Friedmann and Weaver (1979: 6) commented, "The new doctrine of regional planning came to be supported by a new theory, appropriately labeled regional science. Interest in the philosophical foundations of regional planning had dissipated; marching with the times, planning had become a 'scientific endeavor'." The doctrine of regional science has focused on the problem of spatial organization in relation to urbanization, industrial location, intercity linkages, and redressing spatial imbalances via growth centers to induce growth and 'spread effects' from the growth center to the wider economy.

The assumption that ideas developed for regional planning can be applied in all sorts of regions has gradually changed, especially after gaining experience in less developed countries (Rodwin, 1978). The sophisticated knowledge and methodologies of regional planning have to be adapted to take into account the capacities and available data of different administrative contexts. Perloff (1968: 155) identified three key phases in the entire regional planning process for developing countries: the information analysis phase, the planning–programming phase, and the planning–operational phase. Based on the example of Latin America, he illustrated that each of these three phases of planning had its own requirements in terms of scope, content, and methods. Friedmann (1968) also set out 11 major steps in the formulation and implementation of different types of regional plans, programs, and budgets to achieve national objectives in developing countries. He then further identified 12 types of analyses (Friedmann, 1968: 133–134) to stipulate the technical requirements to be adapted for different regional planning conditions: analysis of spatial organization, locational analysis, urban complex analysis, regional land use analysis, analysis of demographic characteristics and behavior, human resource analysis, natural resource analysis, economic activity analysis, regional trade pattern analysis, regional accounts and development models, project analysis, and regional program budget analysis.

In spite of the progress and development in regional science and analytical techniques (Alden and Morgan, 1974), the findings are not always consistent. As Rodwin (1978: 122) commented, "the differences reflect different paradigms, values, and goals. That is why getting more and better data will not help very much." More importantly, the qualitative and quantitative measures of economic progress tended to show the unanticipated negative effects and the self-defeating effects of regional planning policies (Rodwin, 1978; Stohr and Todtling, 1978; Williamson, 1965). After reviewing the trends of regional planning practice, Rodwin (1978: 127) commented that

> A quarter of a century ago the problems of regional planning were deemed to be largely technical. Today they are not. In the main, the shift has occurred because pressing the political issues has become a way of challenging or even discrediting what is considered a wrong or inadequate technical approach just as raising the technical issue under some circumstances can become a way of skirting or discrediting a wrong or inadequate political process.

Ahmad and Bajwa (2005: 3) also echoed the importance of the political process:

> Like the input/output method, linear programming, or the critical path method in the economic sphere, regional analysis and regional planning cannot substitute themselves for the political process of decision-making.

In order to turn things around, institutional transformation had to be promoted in line with the pursuit of regional development and regional planning objectives. This argument remains true in many regional planning contexts.

Regional Planning in China

Since the Communist Party came to power in 1949, regional planning in China has been characterized by intermittent efforts to define and redefine its concept and scope to meet with different political and economic needs. The changing nature and emphasis of regional planning very much reflects the interaction between the development process and the wider socioeconomic circumstances at different time periods. China is, therefore, used as a case study to illustrate some of the issues identified with regional planning practice in the development context.

There is no official definition of what constitutes a region in China. The entire country is divided into different regions under the national 'five-year plan,' but the division varies over time. The most common classification in the prereform period was the six large cooperative regions (see Goodman, 1989: 188) and they were used for economic and social development purposes. Since the seventh five-year plan, the economic focus has been on three regional belts of east, middle, and west; nonetheless, these are not official regions. It is, therefore, difficult to pin down the spatial scale of 'regional planning' in the Chinese context as its usage and meaning have changed dramatically since the inception of the People's Republic of China. The current wave of regional plans tends to use 'region' loosely to cover all sorts of spatial configuration that are beyond the city plan to address issues that cross administrative boundaries. More importantly, different forms of regional plan are disguised under different labels such as territorial plans, urban system plans, urban cluster/ring plans, and metropolitan plans – they can all be categorized as hybrid 'regional plans.'

Regional Development and Planning: the Historic Context

Planning in China had long been synonymous with the soviet style of economic planning. A useful discussion of the changing physical planning system before and after the reform period has been succinctly provided by Wu (2007). Before 1978, physical planning was merely a subordinating mechanism (Nedovic-Budic, 2001) to allocate land use and infrastructural resources to achieve national economic and social development targets set out in the five-year plans (Leaf, 1998; Ng and Wu, 1995; Xu and Ng, 1998; Yeh and Wu, 1999; Tang, 2000). The preparation of five-year plans has been deeply engrained in the governing institution and has continued in the postreform period. The long-range goals outlined in the national five-year plans will then filter through to different levels of government to prepare their own five-year economic plans through the Development and Reform Commissions.

Besides the national economic planning framework, there are master plans for the built-up areas of cities. However, these statutory land use plans do not have complete coverage of the entire administrative area of cities. The word 'cities' (shi), as administrative units in China, carries a very different connotation from the western usage. Chinese cities have different administrative status, with municipalities at the top rank and under direct jurisdiction of central government (e.g., Shanghai and Beijing), through to provincial capitals and subprovincial cities, prefecture-level cities, and county-level cities. In municipalities, provincial, subprovincial, and prefecture-level cities, the urban area (*shiqu*) covers the urban city district, the adjacent subdistricts, and the suburbs, which is more akin to the metropolitan areas of the West. The only form of urban planning in China is city planning, which is a blueprint style, two-tier planning system with a master plan covering the urban city district and a detailed construction plan.

City planning before the economic reform was simply a technical land use plan prepared by the urban planning bureau to complement the five-year (economic) plan. Following the adoption of a more market-oriented economy in the post-economic reform period, the role of the state in direct resource allocation has changed and so has its city planning system. According to Wu (2007), city plans in China have undergone major changes in orientations in the postreform period. There has been a clear shift from the physical design-oriented approach to focus on the wider economic development objectives. With the accelerating rate of economic development, there has been increasing pressure over land use (Leaf and Hou, 2006). This prompts not only for a more effective control system to curb illegal development and urban sprawl (Yeh and Wu, 1999; Smart and Tang, 2005), but also for more strategic and analytical approaches in preparing city plans to promote urban development. The entrepreneurial ethos has inspired the production of a new form of concept (or strategy) plans to develop more strategic vision over infrastructural development and economic restructuring to increase the competitiveness of cities (Wu and Zhang, 2007). By examining the spatial relationship between places, these concept plans serve to bridge the missing spatial element in the five-year plans.

In spite of the production of strategic concept plans for cities, their restriction to the core built-up area does not provide a strategic spatial horizon to address the gravity of spatial challenges brought by global market forces. Since the Open Door Policy in 1978, foreign direct investment was initiated as part of the reform. Special economic zones were first created along the coastal areas and then along the Yangtze River. This has caused imbalanced growth between the economically buoyant regions along the east coast (particularly in the Pearl River Delta, the Yangtze River Delta and Bo Hai Gulf around Beijing and Tianjin) and those inland and lagging-behind areas in the middle and western regions. Such a differential growth rate has further reinforced the long-standing east–west poverty fault line caused by the differential local resources and historical trajectories of development in China (Edmonds, 1994). This leads not only to major migration flows from the less developed regions to the more economically successful areas, but also to the exploitation of natural resources from the less developed central and western

areas toward the coastal regions. This bears a significant impact on social equity, as well as ecological and environmental degradation problems at a national scale.

Rapid Urbanization and Planning Challenges

The recent relaxation over population registration control gives much greater freedom of rural–urban migration and the traditional administrative bifurcation of urban and rural becomes less important. Surplus rural workforce in less wealth-off locations can move to large cities to seek jobs. In spite of the wealth redistribution brought by the rural–urban migration process, what one witnesses in China is a widening gap between urban and rural locations. According to the National Bureau of Statistics (2005), over 26 million people living in rural areas were deemed as living in absolute poverty. This is a self-perpetuating process because the income gap encourages further rural–urban migration and, in particular, toward the metropolitan cities of Beijing, Shanghai, and Shenzhen. The urban–rural divide, in a strong economic sense, also translates into serious social tension and problems.

The consumption of land has been rapidly increasing following major industrialization and urbanization since the late 1980s. On the economic front, the government and major corporations have been actively adjusting land to construct industrial parks, key transport infrastructure, and public utilities. At the same time, rural–urban migration and the rapid growth of cities have led to rising demands of housing and commodity housing market booms (Tian and Wong, 2007). In order to cope with the development pressure, there has been major planned urban sprawl by local government. The expansion of urban and industrial use of land leads to a continued trend of rural and agriculture land reduction. This has caused stress on the carrying capacity of the environment and has created many unexpected ecological consequences.

These challenges prompt for a more strategic spatial approach for planning in China. President Hu Jiantao has issued clear directions (State Council, 2007) over the reconfiguration of spatial strategies to address key regional development issues of environmental protection, infrastructural provision, land use regulation and economic competitiveness. The concept of 'key functional regions' (*zhutigongnengqu guihua*) are adopted to serve as a framework for sustainable development by taking into account the spatial distribution of population, land use patterns, economic structure, and the urban/rural hierarchy. These new concepts and ideas from the upper level government have since been interpreted by local and provincial governments as 'directions.' Despite the lack of an official definition of regional planning, the government's emphasis on regional development and coordination means that economic regions are flexibly defined according to the characteristics and potentials of the economic, social, and ecological environment of each region.

Three Waves of Regional Planning

Chinese scholars (e.g., Cui et al., 2006; Wang, 2007) suggest that there have been three upsurges of regional planning over the last 60 years (see **Table 1**). Regional planning was first introduced as a successful soviet model in the early years of Mao's regime. The emphasis of development was on the prioritization of industrial growth under the centralized command economy. Regional plans were not merely tools to implement the economic goals of resource exploitation and population distribution, but to direct major industrial projects into the inland areas to realize the first and second 'five-year plans' (Wu, 2006). There was, however, a lack of effective intra-regional planning (Wang and Hague, 1993). Unfortunately, the implementation of these plans did not meet with the government's expectation of achieving economic revival. Following the fracture of the Sino-Soviet relationship in the early 1960s, this form of regional plans fell out of favor in the next 20 years during the turbulent political times.

It was not until the early 1980s, after Mao's regime, that there was a comeback of regional planning. The Open Door Policy has led to the contradictory strategy of adopting the free market mechanism while retaining the strong state control machinery (Lai and Lai, 1991). It also brought major uncertainties over interregional development balance with it as foreign investment was concentrated along the east coast. According to Wang and Hague (1993), it was these contradictory aspirations that motivated the government to adopt comprehensive territorial planning to address the competing imperatives. This new approach to regional planning drew on the experience from Chinese river control and land survey practices and land use planning systems in Japan and Western countries. It also adopted the Geddesian approach of 'survey before plan'.

Territorial planning was still very much grounded on the goal of gaining economic efficiency, particularly during the postreform

Table 1 Three Waves of Regional Planning in China since 1949

Wave 1: 1950s The Soviet model	Productivity distribution planning	Resource exploitation, productivity distribution, location decisions of large industrial projects, cities, and residential locations
Wave 2: 1980s–1990s Open Door Policy and orthodox regional planning paradigm	1980s: Territorial planning The Japanese and Western models	Exploitation, use, and protection of the territorial space – coordinate economic, environment, and population resources (later become land use plan that focuses on farmland preservation)
	1989: Urban system planning (first National Planning Act)	Spatial hierarchy and functional relationship between cities and region – designation of function, size, and allocation of each city and provision of infrastructure at the regional level
Wave 3: 2000 to now Globalization, urbanization, and sustainable development	Experimenting with new forms of regional plans and regional governance	Strategic spatial plans Conceptual plans At different spatial scales (from the whole state to counties) Cross administrative borders

Source: Wong, C., Qian, H., Zhou, K., 2008. In search of regional planning in China: a case of Jiangsu and its surrounding Yangtze Delta. Town Planning Review 79 (2/3), 295–329.

entrepreneurial period, although it tried to strike a balance by maintaining a stable natural and human living environment (Wang and Hague, 1993: 565). This aimed to achieve the 'win–win' solution of sustainable development by having economic growth without damaging its environmental resources and resembled some of the underlying principles of the European Spatial Development Perspective (CEC, 1999) as territorial plans 'were conceived as a means of achieving horizontal integration of proposals from the vertical (multispatial scales) planning of enterprises and functional ministries within the command economy' (Wang and Hague, 1993: 566). Territorial planning also provided a vertical spatial planning framework. At the national level, it focused on identifying the location of key growth centers where human and natural resources could be sustained and at the provincial level, local planning commissions were encouraged to propose new development policies for their regions after the survey stage.

The experiment of the new form of territorial planning, however, ran into trouble as the lack of good-quality statistics and information for the mega regions seriously undermined the 'survey' stage of the plan-making process. As expected, practice tended to fall short of the ideal, the strong 'rational' principles that underpinned territorial planning did not fulfill its promise. More importantly, there was not an institutional structure to implement the strategies and the longer term time horizon did not work well with the pace of market economies. More importantly, even at its height, territorial planning was never a part of the statutory planning system. In 1991, following the reorganization of the central government departments, a new Territorial Ministry was set up to oversee territorial planning, which signified the demise of this strategic form of planning. The new ministry turned territorial planning into physical land use planning with a strong emphasis on farmland preservation, which greatly deviated from the original form of strategic regional planning.

Following the demise of territorial planning, the Urban Planning Department under the National Ministry of Construction set out a new form of urban system planning. This new style of planning was driven by the technical analysis of the relationships between cities and their region to develop a spatial hierarchy to optimize the functions, population size, and allocations of each city and to arrange infrastructural provision of the region. Such analysis was known as the 'three structures' (i.e., the rank-size structure, the functional division of labor structure, and the spatial distribution structure) and 'one network' (i.e., infrastructure network) approach (Gu, 1992; Song and Gu, 1988). This scientific yet simplified form of urban system planning had a big appeal to the government and has become part of the statutory planning system under China's first National Planning Act (1989). Since then, it continued to be the dominant form of regional planning and is implemented across different levels of government. Urban system planning tends to follow a normative paradigm of devising the perfect spatial form and it, therefore, fails to address the complexity of the regional system. More importantly, its bold assumption that regional development problems will be resolved through the three structures and one network approach does not always stand. It is thus criticized as rigid, pseudoscientific, and not implementable (Wang, 2007).

After the Asian financial crisis in 1997, a significant sum of foreign direct investment was redirected from elsewhere in Asia to China. The triple helix of market reforms, globalization, and the extraordinary rate of urbanization are seen as the driving forces that motivate key stakeholders to develop regional plans to express their visions of territorial development and to coordinate regional growth patterns. In spite of the significant efforts made to produce strategic urban plans, they failed to provide a strategic horizon to address the socioeconomic and environmental pressures brought by globalization. This has led to the third regional planning boom at the turn of the millennium. Unlike the city master plans, these regional plans tend to focus on deriving strategic guidance and governance structure to regulate and to coordinate, rather than dwelling on detailed spatial arrangements and the location of land use projects. Li and Wu (2012: 140) argue that 'the ups and downs of regional plans reflect the evolution of state power at the regional scale' and 'the rescaling of governance from the localities to the regional level.'

The Third Wave of Regional Planning

The different labels and names of the current wave of regional plans signify the diversity of their development goals. Examples of five broad types of plan that fall into the genre of regional planning are provided in **Table 2**.

'City-region plans' are a hybrid of traditional regional plans and master plans. These plans provide complete spatial coverage of the metropolitan area and apply the technical approach used in master planning. They focus not only on land use, but also on producing spatial strategies to coordinate development activities for the city-region. 'Cross-boundary regional plans,' however, aim to have some forms of collaborative mechanism in place to address pressing planning issues that transcend administrative boundaries. The move toward open markets means that local authorities are in direct competition with each other within the same region, although they share similar planning concerns. These plans help to improve the coordination of regional development to enhance economic competitiveness. There are also 'urban agglomeration plans' and 'metropolitan plans,' both of which are similar in nature and aim to develop strategies to coordinate the development of a network of cities. The urban agglomeration plan tends to cover a network of similar size cities, while the metropolitan plan tends to involve a major metropolis. Finally, 'urban–rural integration plans' focus on examining the functional complementary between urban and rural areas and to develop strategic frameworks to balance development in these areas.

In spite of the blossoming of different forms of regional plans, as far as the government is concerned, the only statutory regional plans are those called 'urban system plans,' although the nature and character of some of these plans have also changed dramatically in recent times to be more innovative and visionary rather than just about the normative hierarchy of the urban system. More interesting, the once waning territorial planning has been rejuvenated and has resumed its experimental role since 2000 (Niu, 2004).

Li and Wu (2012) distinguish regional plan making according to their top–down and bottom–up approaches. The former includes the Pearl River Delta Regional Plan (Xu and Yeh, 2010, 2011) and the Yangtze River Delta Regional

Table 2 Different types of regional plan

Plan types	*Examples*
Urban-region plans	Jiaxing urban-region master plan
	Ningbo urban-region master plan
Cross-boundary regional plans	Development plan for areas along Yangtze River in Jiangsu Province
	Urban belt plan along Yangtze river in Jiangsu Province
	Yangtze Delta regional plan
Urban agglomeration plans	Peal River Delta strategic plan
	Development strategic plan for urban Agglomeration around Hangzhou Bay area
	Mid-Shaanxi Province urban agglomeration construction plan
Metropolitan plans	Nanjing metropolitan area plan
	Suzhou-Wuxi-Changzhou metropolitan area plan
	Quanzhou half-hour commuting metropolitan area concept plan
Urban–rural integration plans	Nanhu district (Jiaxing city) urban–rural integration plan

Source: Wong, C., Qian, H., Zhou, K., 2008. In search of regional planning in China: a case of Jiangsu and its surrounding Yangtze Delta. Town Planning Review 79 (2/3), 295–329.

Plan (Li and Wu, 2012) initiated by the central government to coordinate and regulate local and regional development, whereas the latter covers regional plans initiated by different layers of local government (Wong et al., 2008) to foster competitive advantage. After an extensive discussion on the emergence of centrally initiated Yangtze River Delta Regional Plan, Li and Wu (2012: 140) argue that regional plans serve as the 'spatialized' economic instrument for the central government to maintain some measures of regulation, both functionally and geographically, without inserting another tier of regional government. To them, regional planning is the result of the rescaling of state governance to manage political fragmentation caused by economic competitiveness of localities.

The analysis of various new regional plans in Jiangsu Province shows that these plans are largely driven by local interests to enhance their global competitiveness (Wong et al., 2008). Although they are not initiated centrally, they tend to be the product of an interactive top–down and bottom–up process involving different layers of government. There is a lot of rhetoric in the vision of these plans and the analysis is filled with fashionable spatial planning concepts and jargon, but with little substantive evidence to back up their usage. While environmental issues and rural areas are featured in the plans, they are predominantly urban based and economic oriented. These forces are shaping the planning process and governance structure as well as the content, rationale, and analytical principles of the plans.

The evidence so far suggests that regional planning in China has very limited actual impact other than promoting partnership working among different stakeholders (Li and Wu, 2012; Wong et al., 2008). Without the statutory status or a regional-tier governance structure to implement the regional plan, the coordination of economic, land use, infrastructure, environment, and other interests is undermined by the complex institutional structure and the jostling of power between different layers of government. It is thus difficult to argue that the latest surge of regional plans have so far achieved more balanced and sustainable development patterns. The casual usage of different terms and labels of the new wave of regional plans in China has obscured the understanding and obstructed meaningful debate over what a regional plan is (Wong et al., 2008). Of the various forms of regional plans, there are only two statutory plans – the urban system plan and the urban master plan. The irony is that there does not seem to be any regulation over the content and format of different types of statutory plans, which allows planning authorities to manipulate the labels and perform the 'new wine in the old bottle' trick. The lack of official status and a suitable institutional framework has resulted in a patchwork quilt of regional plans that are easily subject to the manipulation of local authorities to play one against another when they feel fit (Li and Wu, 2012).

These latest regional plans appear to be more strategic and comprehensive in nature when compared to the dominant form of urban master plans. However, the recent twist to the story is that the traditional urban planning framework was reformed by the enactment of the 2008 Urban and Rural Planning Act. Prior to this Act, the urban core (*zhuchenqu*) and suburban districts (*chengshi jiaoqu*) formed the 'planned urban area,' whereas counties beyond the suburban areas were defined as rural areas (Shieh, 2011). The 2008 Act, replacing the former Urban Planning Act, demarcated the watershed of including rural areas in the formal master planning system and that 'city-regions' become the spatial context for planning (Qian and Wong, 2012). Issues related to sustainable development such as arrangements of development land in urban and rural areas, environmental protection, and basic farmland protection are now required in this new form of master planning. However, the urban master plan can only provide general guidance for rural areas as the detailed planning of rural areas still falls under the township and village plan, which is separated from the mainstream urban plan. The new urban and rural planning system, therefore, cannot be argued as a bona fide integrative regional plan.

Conclusion

Friedmann and Weaver (1979: 6) postulated that 'the doctrine of regional planning will vary not only with national traditions in the practice of planning, but also with ideologies and underlying socioeconomic conditions.' The case of China has demonstrated the changing practice of regional planning to meet with different socioeconomic circumstances and political values and needs over time. While the techniques and methodologies of regional science are well developed, their

applications are highly constrained by the availability of robust data and evidence base. Technical rationality of planning alone cannot address the complex development challenges faced by local and regional actors. It is the political process and institutional changes that affect the ultimate delivery and practice of regional planning. China's two different types of regional planning demonstrate different rationalities. The planning over a larger region such as the Pearl River Delta or Yangtze River Delta is usually directed by the central government, aiming to enhance the coordination over different administrative units, while city-region planning for the core municipality is driven by local government to gain their competitive edges. However, both rationalities serve to provide an arena for different levels of governments to express their own visions and ideas to shape future development across the immediate built-up city areas. Many new concepts and ideas are thus introduced and debated, but more importantly, regional partnerships are emerging to develop new forms of local and regional governance. In order to manage the pressure of rapid urbanization and the socioenvironmental problems brought by globalization and climate change in developing countries, effective forms of regional planning have to continue to innovate the governance and institutional framework, decision-making process, content, and methodology of plan making to adapt to changing conditions.

See also: Planning During Post-socialism; Planning and Development; Regional Planning.

Bibliography

Ahmad, I., Bajwa, I.U., 2005. Regional Development Planning – Issues and Realities. Conference Proceeding of paper presented at the 41st ISoCaRP Congress, 17–20 October, Bilbao http://www.isocarp.net/data/case_studies/649.pdf.

Alden, J., Morgan, R., 1974. Regional Planning: a Comprehensive View. Leonard Hill, Beds.

Cui, G., Wei, Q., Liu, K., 2006. Regional Analysis and Regional Planning. Higher Education Press, Beijing.

Edmonds, R.L., 1994. Patterns of China's Lost Harmony: A Survey of the Country's Environmental Degradation and Protection. Routledge, London.

European Commission (CEC), 1999. European Spatial Development Perspective, towards Balanced and Sustainable Development of the Territory of the European Union. Office for Official Publications of the European Communities, Luxembourg.

Friedmann, J., 1968. Education for regional planning in developing countries. Regional Studies 2 (1), 131–138.

Friedmann, J., Weaver, C., 1979. The Evolution of Regional Planning. University of California Press, Berkeley and Los Angeles.

Geddes, P., 1915. Cities in Evolution. Williams and Norgate, London.

Goodman, D.S.G. (Ed.), 1989. China's Regional Development. Routledge, London.

Gu, C., 1992. Chinese Urban System: History, Current Situation, Prospect. The Commercial Press (in Chinese), Beijing.

Hall, P., Thomas, R., Gracey, H., Drewett, R., 1973. The Containment of Urban England. George Allen and Unwin, London.

Lai, O.K., Lai, T.K., 1991. Socialist state planning reconsidered: regional policy in China, 1949-1989. Environment and Planning C 9, 207–224.

Leaf, M., 1998. Urban planning and urban reality under Chinese economic reforms. Journal of Planning Education and Research 18, 145–153.

Leaf, M., Hou, L., 2006. The "Third Spring" of urban planning in China: the resurrection of professional planning in the post-Mao era. China Information 20 (3), 553–585.

Li, Y., Wu, F., 2012. The emergence of centrally initiated regional plans in China: a case study of Yangtze River Delta Regional Plan. Habitat International 39, 137–147.

Meyers, A.A., 1998. Invisible cities: Lewis Mumford, Thomas Adams, and the invention of the regional city, 1923–1929. Business and Economic History 27 (2), 292–306.

National Bureau of Statistics, 2005. Communiqué on 2004 Rural Poverty Monitoring of China. http://www.stats.gov.cn/english/newsandcomingevents/t20050513_402249491.htm.

Nedovic-Budic, Z., 2001. Adjustment of planning practice to the new eastern European context. Journal of the American Planning Association 67, 38–52.

Ng, M.K., Wu, F., 1995. A critique of the 1989 City Planning Act of the People's Republic of China: a western perspective. Third World Planning Review 17, 279–293.

Niu, H., 2004. National planning, regional planning and urban planning: how to coordinate them in China. City Planning Review 28 (11), 4206.

Parr, J., 2005. Perspectives on the city-region. Regional Studies 39 (5), 555–566.

Perloff, H.S., 1968. Key features of regional planning. Journal of the American Institute of Planners 34 (3), 153–159.

Qian, H., Wong, C., 2012. Master planning under urban–rural integration: the case of Nanjing, China? Urban Policy and Research 30 (4), 403–422.

Rodwin, L., 1978. Regional planning in less developed countries: a retrospective view of the literature and experience. International Regional Science Review 3 (2), 113–131.

State Council, 2007. Report on the Work of the Government. http://www.gov.cn/test/2007-03/17/content_811874.htm.

Seltzer, E., Carbonell, A., 2011. Planning regions. In: Seltzer, E., Carbonell, A. (Eds.), Regional Planning in America: Practice and Prospect. Lincoln Institute of Land Policy, Cambridge, Massachusetts, pp. 1–16.

Shieh, L., 2011. Becoming urban: urban–rural integration in Nanjing, Jiangsu Province. Pacific Affairs 84 (3), 475–493.

Smart, A., Tang, W.S., 2005. Irregular trajectories: illegal building in mainland China and Hong Kong. In: in Ma, L.J.C., Wu, F. (Eds.), Restructuring the Chinese City: Changing Society, Economy and Space. Routledge, London, pp. 80–97.

Song, J.T., Gu, C.L., 1988. Theory and methodology of urban system planning. Acta Geographica Sinica, 1988–2002.

Stohr, W., Todtling, F., 1978. An evaluation of regional policies – experiences in market and mixed economies. In: Hansen, N.M. (Ed.), Human Settlement Systems: International Perspectives on Structure Change and Public Policy. Ballinger, Cambridge.

Talen, E., 2008. Beyond the front porch: regionalist ideals in the new urbanist movement. Journal of Planning History 7 (20), 20–47.

Tang, W.S., 2000. Chinese planning at fifty: an assessment of the planning theory literature. Journal of Planning Literature 14, 347–366.

Thompson, C.W., 2004. Geddes, zoos and the valley section. Landscape Review 10 (1–2), 115–119.

Tian, Y.Y., Wong, C., 2007. Large urban redevelopment projects and socio-spatial stratification in Shanghai. In: Wu, F. (Ed.), China's Emerging Cities: The Making of New Urbanism. Routledge, London, pp. 210–231.

UN Habitat, 1997. Regional Development Planning and Management of Urbanization: Experiences from Developing Countries. United Nations Human Settlements Programme (UN-HABITAT), Nairobi, Kenya.

UN Habitat, 2012a. Urban Planning for City Leaders. United Nations Human Settlements Programme (UN-HABITAT), Nairobi, Kenya.

UN Habitat, 2012b. State of the World's Cities 2012/13: Prosperity of Cities. United Nations Human Settlements Programme (UN-HABITAT), Nairobi, Kenya.

Van Staveren, J.M., van Dusseldorp, D.B.W.M. (Eds.), 1983. Framework for Regional Planning in Developing Countries: Methodology for an Interdisciplinary Approach to the Planned Development of Predominantly Rural Areas. International Institute for Land Reclamation and Improvement, Wageningen.

Wang, H., 2007. Three Booms in Regional Planning in China: Re-approaching the Methodology of Comprehensive Planning? Paper Presented at the International Conference on China's Urban Transition and City Planning, 29–30 June, Cardiff.

Wang, Y.P., Hague, C., 1993. Territorial planning in China: a new regional approach. Regional Studies 27 (6), 561–573.

Williamson, J.G., 1965. Regional inequality and the process of national development. Economic Development and Cultural Change 13 (4), 1–84, 2.

Wong, C., Rae, A., Schultz-Baing, A., 2006. Uniting Britain – The Evidence Base: Spatial Structure and Key Drivers. the Royal Town Planning Institute, London.

Wong, C., Qian, H., Zhou, K., 2008. In search of regional planning in China: a case of Jiangsu and its surrounding Yangtze Delta. Town Planning Review 79 (2/3), 295–329.

Wu, T., 2006. Contemporary Regional Planning in China. Tsinghua University Press, Beijing.

Wu, F., 2007. Re-orientation of the city plan: strategic planning and design competition in China. Geoforum 38, 379–392.

Wu, F., Zhang, J., 2007. Planning the competitive city-region: the emergence of strategic development plan in China. Urban Affairs Review 42 (5), 714–740.

Xu, J., Ng, M.K., 1998. Socialist urban planning in transition: the case of Guangzhou, China. Third World Planning Review 20, 35–51.

Xu, J., Yeh, A.G.O., 2010. Planning mega-city regions in China: rationales and Policies. Progress in Planning 73 (1), 17–22.

Xu, J., Yeh, A.G.O., 2011. Political economy of regional cooperation in the Pan-Pearl River Delta. In: Yeh, A.G.O., Xu, J. (Eds.), China's Pan-Pearl River Delta:, Regional Cooperation and Development. Hong Kong University Press, Hong Kong.

Yeh, A.G.O., Wu, F., 1999. The transformation of the urban planning system in China from a centrally planned to transitional economy. Progress in Planning 51, 67–252.

Regional Planning, History of

David A Johnson, University of Tennessee, Knoxville, TN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 12925–12930, © 2001, Elsevier Ltd.

Abstract

Regional planning may be defined as the integrated management of the economic, social, and physical resources of a spatially bounded area. Regional plans and policies have been proposed and carried out since the beginnings of civilized settlement. In the modern era, regional plans have been promulgated for sub-national, multi-jurisdictional areas such as metropolitan areas. Comprehensive regional plans have also been devised for poly-nucleated urban areas and for open regions such as river basins. Because, for historical reasons, regions often lack governmental organization congruent with their natural or economic characteristics, the realization of regional planning goals, policies, and projects has often been constrained or limited. Nevertheless, significant instances of the successful implementation of regional plans or elements of plans can be cited. Greater London, Paris, New York, and Chicago, for example, all have benefited from past efforts at regional planning. In an increasingly interdependent world faced with environmental degradation, economic development needs, and social inequities, there is a growing awareness of the need for effective regional planning in the form of integrated regional resource management. But pervasive fragmentation of political jurisdictions, dominance of short-term over long-term decision-making, growing inequities between economic classes, and the frequent absence of consensus on goals or future visions still present formidable obstacles to the realization of regional planning in both developed and developing nations. Powerful new tools for data management and monitoring, such as geographic information systems and computer-based visioning simulation, promise to be of considerable help in overcoming barriers to effective regional planning. However, these tools are no substitute for imaginative planning, political will, and a broad agreement on what future regions can and should become.

The Term 'Regional Planning'

'*Gallia est omnia in partes tres*.' All Gaul is divided into three parts. This command, from the reign of Julius Caesar in the first century BC, no doubt is one of the earliest example of regional planning. A region is a district or area defined by generally recognizable attributes, physical cultural, economic, or political. The word 'region' derives from the Latin '*regio*,' meaning direction or boundary, and from '*regere*,' to direct. Thus the concept of region embodies the interaction among place, people, and work or economy. A region may or may not be formally recognized as a political entity or sub-division. More likely a region is simply a shared mental construct of its constituents.

Regional planning then is the effort of the society or elements in the society to intervene in the integrated developmental evolution of the area in question. In ancient kingdoms, the wishes and wisdom of the ruler constituted the regional planning conception for the society. Values and objectives underlying regional planning and foresight particularly included protection from external threats, but also the provision of infrastructure – roads, water, the enhancement and protection of productive fields and forests, provision for population growth and economic development activities such as industry and trade, cultural enrichment, and provision for societal continuity in succeeding generations. These values and practices were often codified in religious beliefs and scripts. These values still underlie and are used to justify contemporary efforts to define and carry out regional planning.

At this point it will be helpful to discuss a few related words or concepts. We need first to distinguish regional from city planning. The concepts are related but city planning implies a *polis* or civic entity governed by elected or other recognized official bodies. A region does not necessarily have such official status. Indeed, often it does not.

Regional analysis is the study of the interaction of key variables – physical, social and economic – in a regional context. Regional science, a relatively recent evolution founded upon economic location theory, seeks to describe and predict systemic regional phenomena.

We need also to distinguish types of regions. The European Community is a region comprised of nation states seeking to develop a commonality of purpose and policy, which presumably will result in superior outcomes for its constituent members. The European Union thus practices a kind of multi-state regional planning, much as the federal government in the United States does for the aggregate of its 50 constituent states. Parts of nation-states also may constitute regions. One thinks of Scotland and Wales in the United Kingdom or the Appalachian Region or the Great Plains of the United States.

Another distinction among regions is between mono-nodal, multi-nodal, and non-nodal regions. Metropolitan regions are examples of mono- and sometimes multi-nodal regions. Natural resource regions are often non-nodal. River basin regions are another category: linear regions.

Regions can also be distinguished as growth regions or as lagging regions. Regions may lag in development behind national averages, due to lack of natural resources, limited accessibility, or cultural or educational factors. Much attention has been given to assisting lagging regions to 'catch up,' with mixed results.

Today, we hear much about 'integrated regional resource management.' and 'sustainable development.' These concepts, while imprecise and difficult to measure, are increasingly shaping the directions toward which regional policy planning is moving.

We should also distinguish between 'regional planning,' and 'regional plan.' A regional plan is a specific document which lays out in physical detail a future desired end-state. It may also contain social and economic policy statements. The regional plan is typically based on a regional survey of social, economic, and physical factors. 'Regional planning' is the process by which regional plans are made. 'Regional policy' is a related concept. Regional policies are, in effect, guidelines to realize or implement regional plans.

Origins and Practice of Regional Planning

The origins of modern regional planning doctrine can be traced to reactions to the oppressive conditions created by the Industrial Revolution in nineteenth century England and France. Robert Owen and Charles Fourier proposed utopian schemes for alternative workers' housing and communities. Piotr Kropotkin's *Fields, Factories and Workshops* (1901) offered insights into the connections of people, land and labor. Charles Booth's *Life and Labor in London* (1892) systematically surveyed the conditions of working class people. Ebenezer Howard's 1902 *Garden Cities of Tomorrow* presented an influential alternative to unplanned metropolitan expansion. Patrick Geddes, a Scottish biologist, published his *Cities in Evolution* in 1915. Geddes' synthesis of the salient connections among people, land, economic activity, and natural systems greatly influenced the American essayist and critic, Lewis Mumford. Mumford, in turn promoted Geddes' integrative approach to social and public policy through such books as *The Culture of Cities* (1938). Through Mumford, Geddes' ideas influenced the Regional Planning Association of America (RPAA), a progressive private group concerned with development which was active in the 1920s and 1930s.

The Regional Planning Association of America

The RPAA embraced Ebenezer Howard's idea to siphon off and channel new growth at the suburban edges of existing large cities into planned garden cities. The group also espoused the notion that economic growth should be decentralized to smaller cities and towns and diverted from large metropolitan areas. The RPAA was especially influential in New York State, where Governor Al Smith appointed Stein and his fellow RPAA colleague, Henry Wright, to a newly-created Commission on Housing and Regional Planning. In 1926, the Commission issued a regional plan for the state that called for a concentrated grid of lines of transportation and communication along major corridors, separation of cities and highways, and the creation of independent communities away from the centralizing tendencies of its major cities. Protection of forest, farmlands and recreation areas were also proposed. The plan had considerable influence on policy and some of its proposals were ultimately carried out. What is notable is that these policies and issues are still under serious debate some 75 years later (Spann, 1996).

The RPAA's conception of regional planning inevitably brought the group into conflict with the well-funded effort to create a Regional Plan for the New York metropolitan area begun in 1921 under the direction of Thomas Adams, a British town planner who had worked on plans for garden cities of Welwyn and Letchworth, and the Greater London region.

The Regional Plan of New York and Its Environs

The monumental Regional Plan of New York and Its Environs, begun in 1921 and published in 1929 (Committee on the Regional Plan of New York and Its Environs (1929)). The plan brought together many of the development trends and proposals of the time and was underwritten by generous grants from the Russell Sage Foundation totaling some $1.2 million. In 1990, this would be equivalent to roughly $20 million. The Regional Plan and Survey was and probably remains the best funded metropolitan regional plan ever undertaken (Hays, 1965).

The plan, encompassing 22 counties surrounding the Port of New York and targeted on the year 1965, incorporated proposals of the Port Authority rail plans, including those for belt line freight rail lines. Schemes for rapid-transit improvements, highway construction, zoning rationalization, and parks and parkways were assembled from a variety of sources and agencies and synthesized into a unified conception. The plan called for a 're-centralization' of industry and business, involving, among other proposals, a shift of manufacturing outward from Manhattan and the core of the region. Whatever its flaws, the 1929 Regional Plan of New York and Its Environs left a significant imprint on Metropolitan New York. The plan's basic physical planning conceptions were reflected in the subsequent development of the region.

The Regional Plan of 1929 clearly had a significant effect on the physical development of the region. But it also brought important changes in the institutional and public bodies concerned with regional development. Numerous planning and zoning agencies were established at the county and municipal levels as a direct result of field work by the staff of the Regional Plan. State policies for regional development were also stimulated by the plan and its advocates. The plan also influenced the Depression-era New Deal policies of the Roosevelt Administration. President Roosevelt was well-aware of the ideas of both the RPAA and the Regional Plan of New York.

Roosevelt spoke at a Roundtable on Regionalism held at the University of Virginia on 11 July 1931 on the need for regional planning and development, aided by the initiatives of the federal government to lead the country out of the economic stagnation it had found itself. Thus Roosevelt was instrumental in getting the thought of Southern Regionalists, such as Howard Odum and Rupert Vance of the University of North Carolina, and the ideas of the RPAA group onto the agenda of the New Deal. This may have been the origin of the idea for a Tennessee Valley Authority (TVA) in Roosevelt's thinking. On a national scale, the New Deal also created a National Resources Committee to survey the needs and resources of the country.

The Tennessee Valley Authority

The Tennessee Valley Authority (TVA) was the first of the river basin or other large regional agencies to attempt to bring together the then-new concepts of integrated resource development, which was central to the conservation ideas of the

time, and regional planning which was a part of the new city and area planning movement.

Roosevelt's early speeches and papers clearly reveal a deep-seated and long-standing concern with farm problems, rural poverty, and the inadequate methods of providing services to rural areas. He also saw these problems as coupled with those of overcrowding and poor living conditions for the working people in the big cities, which he believed were fueled in part by the migration from farms to cities. The emerging concepts of regional planning seemed to Roosevelt to provide the linkage between the two problems. It is within this context that the regional planning ideas of Roosevelt were appended to earlier ideas for a multi-purpose plan for the Tennessee River, which had focused on flood control, navigation, and power generation. The main body of the TVA Act, as introduced on 9 March 1933, consisted of a comprehensive river improvement program plus two sections which provided for surveys of regional development problems and for recommendations to Congress on necessary legislation to solve these problems. The river improvement sections of the TVA Act were very explicit and provided for a navigation channel from Knoxville to the Ohio River, flood control to protect valley cities and to help control floods on the Mississippi River, and power generation. In contrast, the regional planning sections were very general and thus susceptible to either broad or narrow interpretation. TVA's mandate to conduct and implement regional planning was never clear or unambiguous.

The National Resources Committee (later the National Resources Planning Board) envisioned TVA as the first of a nationwide array of river basin planning regions. TVA was viewed as a model for the rest and as a testing ground for new communities as originally proposed by the RPAA group. The first of these was Norris, Tennessee, built near the first hydroelectric dam, Norris Dam, originally to house construction workers, and then to be converted to housing for the local populations dislocated by the rising waters of Norris Lake. Roosevelt's grand vision for river basin regional planning initiatives was thwarted by the interventionist scale of the vision, in a country dominated by unregulated market forces, and by the outbreak of World War Two. TVA nevertheless became a model for numerous grand multi-purpose hydroelectric and irrigation projects around the world, some misconceived and contrary to sound, integrated regional resource planning and management principles. Sadly, TVA itself has relinquished any serious commitment to regional planning leadership and has become, in effect, primarily a publicly-owned wholesale electricity producer.

Appalachian Regional Commission

Besides TVA, America's other important multi-state regional planning effort has been the Appalachian Regional Commission (ARC). Like TVA, the ARC was created to deal with an economically lagging region, the mountainous, relatively inaccessible chain of counties stretching from Alabama and Georgia to the Southern Tier of Upstate New York. The ARC began as an idea of President John Kennedy, who toured West Virginia's poverty-stricken mining towns in his campaign for the White House. The ARC was established in 1965 in the Johnson Administration. The ARC developed a plan for improving access and social services for isolated communities in parts of 13 states. Highway improvement and infrastructure programs were central elements of ARC regional development. Thirty years later, assessments of the effectiveness of federal investments in the region are mixed. But according to one comprehensive study, the federal investment infusion seems to have been effective. Counties in the ARC program were noted on average to have developed faster economically than comparable counties not in the program (Bradshaw, 1992; Isserman and Rephann, 1995). ARC was still functioning in the year 2000 but other federal regional development efforts such as the eight Department of Commerce Title V regional commissions setup in other lagging regions of the US were not so fortunate. All were closed down in 1980.

Metropolitan Regional Planning

The post-war era in the United States, with its spreading suburbs and growing automobile dependency, raised new issues for regional planners, notably in the large metropolitan areas. Federal aid for metropolitan planning became available with the enactment of Section 701 of the Housing Act of 1954, stimulating the creation of planning commissions for the nation's metropolitan areas. By 1960, two-thirds of US metropolitan areas had some form of area-wide planning and, by 1971, thanks to federal grants and statutory requirements, all of the nation's 247 metropolitan areas had adopted official regional planning, mostly through elected, voluntary, Councils of Government (COGs). The A-95 review process, which required regional review and coordination to receive funding for large scale, federally-assisted projects, and Section 208 of the federal Clean Water Act of 1972 further stimulated the creation of regional planning bodies in metropolitan areas.

In the New York Metropolitan Region, the privately-supported Regional Plan Association produced a Second Regional Plan for the New York metropolitan area, released in November 1968. The Second Regional Plan emphasized the creation of regional sub-centers linked by fast rail to Manhattan to expand job opportunities for those trapped in the central cities. For the suburbs, these centers would provide an organizing focus for social and civic life, employment opportunities, and easy access to and from Manhattan. The centers would be composed of 'regional activities' which would have to be built in any event to accommodate the expected additional 12 million people who would join the 18 million already in the 31-county region (expanded from the original 22-county definition) by the year 2000.

Forecasts for a large shift in employment away from manufacturing to service and white-collar jobs suggested the potential for significant office growth, some of which could go to centers outside of Manhattan. The plan proposed massive new office development for Manhattan, both in downtown and midtown but the plan also proposed that regional centers be built, both as redeveloped inner-city sub-centers such as Downtown Brooklyn, Jamaica and Newark, as older city centers in suburban cities, or as entirely new centers on open land. The centers idea took hold in a number of places, notably Jamaica, Queens, Downtown Brooklyn, and Fordham Road in the Bronx. Downtown Newark also experienced new central place

development and expanded centers were developed in the outlying areas of White Plains, Stamford, and New Brunswick.

The Second Regional Plan draft addressed other issues as well. Its housing policies called for increased densities, a more compact region, better related to accessibility to jobs and activities, and more housing opportunities for lower-income and minority families. The draft also called attention to the continuing need to capture more open space and to deal with the increasingly difficult problem of waste management. The plan called for completion of the region's share of the interstate highway network but it also made a strong case for the preservation and enlargement of existing rail transit systems.

In the 1980s, regional planning in the United States suffered a severe setback when the federal government withdrew virtually all its support for regional planning agencies, resulting in the demise of many programs. Only the transportation planning function survived in most metropolitan areas. TVA and the Appalachian programs managed to continue, though their planning functions were greatly reduced. Some states replaced federal assistance for regional planning with state funds. Other regional agencies survived by becoming more entrepreneurial, focusing on service delivery. A few states such as Florida, Georgia and Massachusetts, concerned about the negative aspects of growth, have developed growth management or 'Smart Growth' initiatives, which require regional planning as a framework for development policies, in some cases.

Regional cooperation in politically fragmented metropolitan areas has generally proved difficult to realize in the United States. A number of metro areas, however, have managed to achieve success in facilitating regional planning consensus. The Twin Cities region of Minneapolis/St. Paul, and the Portland Oregon region are often cited as examples of successful regional initiatives. The Minnesota Twin Cities area created in 1967 through state legislation a seven-county Metropolitan Council. The Council was created to address specific region-wide problems such as sewers and transportation. The Council has adopted a development guide for regional and local decision-making that identifies planned urban areas and growth centers. The Twin Cities regional initiative has enjoyed some major success, though its planning function has been scaled back in recent years.

The Portland Oregon three-county region in 1979 authorized by referendum a metropolitan service district. The district provides technical assistance and reviews local plans. In 1992, the Metro, as it is called, was given approval by charter to adopt a regional framework plan to guide regional development. The framework plan will address land use, development densities, housing, transportation, natural areas, and water supply. The Portland process is generally viewed as a successful example of the application of effective regional planning to control the negative aspects of rapid growth, though some concern has been raised about the impact of controls on the housing costs.

Metropolitan regional planning has experienced a resurgence in the 1990s in a number of American urban areas such as San Diego, California and Chattanooga, Tennessee, while, in the New York metropolitan region, the Regional Plan Association released in 1995, a Third Regional Plan, successor to the 1929 Regional Plan of New York and Its Environs, and to its Second Regional Plan of 1968 (Yaro and Hiss, 1996). The Third Plan has focused on ways to help the region recover its momentum so as to maintain its faltering position in the global marketplace. To address the problems facing the region, the Third Regional Plan proposes five ambitious campaigns focused on open and green space, transportation improvements, strengthening urban centers, improving workforce skills, and streamlining governance structures. It remains to be seen whether the Third Regional Plan will have as much impact as its two predecessors.

British and European Regional Planning

The trajectory of regional planning in Britain and Europe has somewhat paralleled that of the United States. Regional planning in post-World War II Britain found new support and momentum. The New Towns idea was re-embraced as a means to rebuild a Britain exhausted by war. Regional plans for the utilization of the old, lagging industrial regions of the North of England, Wales, and Scotland were adopted. These provided for decanting of populations into a new generation of expanded and new towns. Older urban areas would be contained within protected greenbelts.

In the 1960s, regional planning was further spurred by commission reports concerned about fast growing areas, notably in the Southeast of England. Metropolitan regional planning, mainly advisory, was undertaken by county and regional councils. Their mandate was to:

1. work out broad objectives for each region and provide a framework for detailed sectoral decision-making;
2. advise on national policy-making where it could significant affect regions;
3. advise on the application of national policy in their region; and
4. help create a consensus within a region as to its problems and possible solutions.

Regional planning tended to be more favored by the Labor Party, less so by the Conservatives. The Conservatives tended to view taxation for social schemes and infrastructure as a burden on business and regulation and control of development as disincentives to market responses to problems of growth. The Thatcher government in the 1980s reconfigured regional planning bodies and responsibilities: the Greater London Council and other regional councils were abolished and regional planning was transformed into a mechanism for assisting large-scale private redevelopment schemes.

With the return of a more liberal-left government in the 1990s, there has been a modest return to regional planning initiatives. These initiatives, however, have continued the private/public partnership approach promoted by the Thatcher government. The central government places more emphasis on local government bodies. Grand regional planning schemes are not in vogue, as pragmatism and incremental and market approaches are more the order of the day. Whether this approach will enable the UK to meet the challenges of the first decades of the new century remains to be seen. With continuing population growth, a deteriorating environment, limited local resources, and continued un- and under-employment in some regions, together with a need for further integration into the European system, Britain faces hard choices. Robust, effective regional planning and implementation intermediate between

the central government and local governments would appear to be a prerequisite for meeting these challenges. Whether the political climate will permit and support an articulated interventionist approach remains to be seen.

On the European continent, regional planning has been practiced with considerable success in both France and the Netherlands. France, with its strong provincial history and identities, has been able to focus its regional planning efforts on channeling excessive growth away from the Paris region toward second- and third-order city-regions. Meanwhile, the Paris region has benefited from metropolitan region planning and development that has focused on transit-based development and suburban centers. Regional planning in the Netherlands has emphasized a '*Randstadt*' plan which focuses development on the ring of cities surrounding a protected green center in the middle of the country. The European Union has fostered regional cooperation among regions that share common characteristics. An example is the Atlantic Arc Region, the grouping of the regions of EU member country fronting the Atlantic Ocean, which share technologies and cooperative approaches to managing planning and development.

Future Prospects for Regional Planning

To summarize, it is reasonable to say that regional planning in western industrialized countries has historically had a checkered history of meeting the challenges of growth and decay at the sub-national level. But it is also accurate to point out that those areas that adopted and tried to carry out regional planning schemes have often achieved better results than those which simply drifted into incremental, piecemeal development. It is almost always an advantage to have some systematic framework and political consensus on which to build, even if all relevant variables cannot be predicted or controlled.

There is reason to believe that regional planning will enjoy new vitality and acceptance. Powerful new forces, needs, and demands are compelling a renewed look at the desirability of regional planning approaches. Globalization trends are fostering the integration of economies on an international basis. Metropolitan economies are increasing seen as the building blocks of a world trading system. The communication and air travel revolutions have reduced the accessibility advantages enjoyed by a few large cities. Competitiveness will require regions to make the most of their assets, and grow new ones. This will take conscious, deliberate, smart planning. Environmental pressures and the deterioration of air, water, and land have reached serious proportions in many countries. Regional approaches will be needed to address the pollution of air and water sheds, which rarely conform to the boundaries of local jurisdictions. Global warming and the need to reduce carbon emissions will inevitably require regional approaches to the transportation and land–use interface. Sustainability is becoming the norm by which regional development decisions increasingly are being judged. More attention is being paid to the protection of natural and cultural amenities, natural habitat, green areas, wetlands and waterways, which, by their scale, mandate regional approaches. At the same time, the digital revolution has provided policy-makers and the public with new tools to collect, manage, and display vast amounts of relevant data, in ways only dreamed of in the early years of regional planning. Powerful computer-based geographic information systems (GIS) now permit mapping and analysis of trends, projection of impacts, and improved understanding of alternatives by regional bodies. This new capacity will surely enhance regional planning possibilities and effectiveness in many areas and countries.

See also: Planning Issues and Sustainable Development; Regional Geography; Resource Geography; Urban System in Geography.

Bibliography

Boardman, P., 1978. The Worlds of Patrick Geddes: Biologist, Town Planner, Re-educator, Peace-warrior. Routledge & Kegan Paul, London.
Booth, C. (Ed.), 1892. Life and Labour of the People of London. Macmillan, London.
Bradshaw, M., 1992. The Appalachian Regional Commission: Twenty-Five Years of Government Policy. The University Press of Kentucky, Lexington, KY.
Committee on the Regional Plan of New York and Its Environs, 1929. Regional Plan of New York and Its Environs. The Committee on the Regional Plan of New York and Its Environs, New York.
Cullingworth, B., 1999. British Planning: 50 Years of Urban and Regional Policy. Athlone Press, Somerset, NJ.
Downs, A., 1994. New Visions for Metropolitan America. Brookings Institution and the Lincoln Institute of Land Studies, Washington, DC.
Friedmann, J., Weaver, C., 1979. Territory and Function: The Evolution of Regional Planning. University of California Press, Berkeley, LA.
Geddes, P., 1915. Cities in Evolution. Williams and Norgate, London.
Hansen, N., 1968. French Regional Planning. Indiana University Press, Bloomington, IN.
Hays, F., 1965. Community Leadership. Columbia University Press, New York.
Isard, W., 1960. Methods of Regional Analysis. John Wiley & Sons, New York.
Isserman, A., Rephann, T., 1995. The economic effects of the appalachian regional commission: an empirical assessment of 26 years of regional development planning. Journal of the American Planning Association 61 (3), 345–364.
John, D., et al., 1998. Building Stronger Communities and Regions: Can the Federal Government Help? National Academy of Public Administration, Washington, DC.
Johnson, D.A. (Ed.), 1996. Planning the Great Metropolis: The 1929 Regional Plan of New York and its Environs. E & F Spon, London.
Katz, B. (Ed.), 2000. Reflections on Regionalism. Brookings Institution, Washington, DC.
Kropotkin, P., 1901. Fields, Factories and Workshops. S. Sonnenschein, London.
MacFadyen, D., 1970. Sir Ebenezer Howard and the Town Planning Movement. MIT Press, Cambridge, MA.
MacKaye, B., 1962. The New Exploration: A Philosophy of Regional Planning. University of Illinois Press, Urbana, IL.
Meck, S. (Ed.), 1999. Planning Communities for the 21st Century. American Planning Association, Chicago.
Mumford, L., 1938. The Culture of Cities. Harcourt Brace Jovanovich, New York.
So, F., Hand, I., McDowell, B.D. (Eds.), 1986. The Practice of State and Regional Planning. International City Managers Association and the American Planning Association, Chicago.
Spann, E.K., 1996. Designing Modern America: The Regional Planning Association of America and its Members. Ohio State University Press, Columbus, OH.
Wannop, U.A., 1995. The Regional Imperative: Regional Planning and Governance in Britain, Europe, and the United States. Jessica Kingsley Publishers, London.
Yaro, R., Hiss, T., 1996. A Region at Risk: The Third Regional Plan for the New York. New Jersey, Connecticut Metropolitan Area. Island Press, Washington, DC.

Regional Science

Andrew Isserman[†] **and Roger R Stough,** University of Illinois, Urbana, IL, USA; and George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Isserman, volume 19, pp. 12930–12935, © 2001, Elsevier Ltd.

Abstract

This historical assessment of the history of regional science includes a description of its membership as regionally disaggregated association sections were consolidated into a global organization of some 4000 members. The theory, methods, and practices of the field are highlighted along with names of major awards contributors have received.

The origins of regional science are traced from the 1920s, stressing its interdisciplinary nature. The subject matter of regional science is then considered, focusing on location theory, methods of regional and spatial analysis, and regional development and policy. Next, the intellectual progress of the discipline is charted, looking in particular at location and spatial patterns, methods of regional analysis, and the explanation of differences in regional development.

Economists pay too little attention to the location of economic activity and the effects of distance on economic systems. Geographers focus on the spatial dimension but lack rigorous tools to study the effects of location and distance on human behavior or explain why activities are located where they are. Urban and regional planners need a stronger analytical foundation to go beyond the physical and aesthetic aspects of urban form to consider social and economic dimensions of city and region. With such premises, 60 economists, geographers, political scientists, sociologists, and planners gathered at the annual meeting of the Allied Social Science Associations in 1954 and founded the Regional Science Association. Its objective was 'to promote studies focusing on the region.' Envisioned were the use of theoretical frameworks and methods of the social and other sciences and the development of new ones specifically for regional analysis.

This interest in the region as a subject of study had important and immediate political and intellectual antecedents in the United States. In 1927, the regional plan for New York considered economic factors in metropolitan growth and spatial form, largely through the work of Robert Haig. In 1933, the Tennessee Valley Authority became the nation's first major regional development program, arguably since the Western expansion. In 1935, asserting that policies of national development depend on regional factors, the US National Resources Committee asked 'modern students of regional science' (10 regional geographers and 2 regional sociologists) to identify the best type of region to use for regional planning and development. In 1942, Alvin Hansen, perhaps the leading American Keynesian economist and a Harvard professor, advocated a comprehensive postwar expansionary program of regional development on the premise that the effective utilization of resources in any region makes the whole nation richer.

Two organized, interdisciplinary efforts with very different intellectual roots had focused on regions. In the 1920s, Lewis Mumford and his colleagues in the Regional Planning Association of America saw cities as congested and unmanageable. They advocated a mode of thinking and procedural methods aimed at seeing the region as a whole – integrating city and region, and utilizing, not destroying, the natural environment. Closer to the regional scientists because of their emphasis on both science and the region and their scholarly university base were the southern regionalists. Centered at the University of North Carolina, they were well established by 1931 in terms of research achievements, agendas, and infrastructure. Howard Odum, the acknowledged leader of the effort and a sociologist, argued in 1945 that the nation owes the fullest development of the South not only to the South but also to the nation itself. Similar in ambition but predating the Regional Science Association by a quarter century, Odum advocated a regional approach to the scientific study of human society, arguing that it would entail the integration of method and theory, require the joint effort of the traditional social science disciplines, and create fundamental new understanding of all society, not just its regions.

Members of the Regional Science Association from the outset debated whether regional science would become a new discipline or an interdisciplinary field.

Walter Isard, widely recognized as the father of regional science, was the leading proponent of the disciplinary view. In 1956, he became an economics professor at the University of Pennsylvania, where he formed the Department of Regional Science in 1958. By then he had published 23 papers in the leading economics journals (acknowledging the help of Alvin Hansen in the first five). Yet the Regional Science Program would go far beyond economics in its scope and impact. In 1960, the first PhD in Regional Science was awarded to William Alonso, later regional planning professor at Berkeley and Saltonstall, Professor of Population Studies at Harvard. Subsequent doctoral recipients took positions in economics, engineering, geography, planning, and other university faculties as well as public agencies and consulting firms.

Isard's prediction that regional science programs would emerge from geography departments and achieve wide recognition before 2000 was not realized. In the 1970s, Cornell University established the second Regional Science PhD program, and the University of Illinois initiated a multidepartmental program requiring completion of the PhD in a participating department. By 1976, regional science degree programs or specializations existed at 14 universities internationally. The lone department devoted exclusively to regional

[†] Deceased.

science remained the one at Pennsylvania until it was closed during university retrenchment in 1994.

Among the early challengers of the disciplinary view was Lloyd Rodwin, a planning professor at the Massachusetts Institute of Technology. He envisioned the social sciences developing more systematic spatial studies, geography becoming more analytical, and regional science becoming 'a generic name for regional studies.' This vision has become reality, but not fully. Regional science has become a generic name for a certain kind of regional studies, one whose origin is in mathematical abstraction and quantitative modeling.

Today regional science is an international, interdisciplinary endeavor. It brings together economists, geographers, planners, and others whose research has a regional dimension. It is the confluence of many streams, but arguably also the origin of a few. Each discipline brings different theoretical and methodological approaches to regional science, and each typically has different substantive concerns within a given topic. Migration, for example, interests several disciplines. An economist might focus on migration behavior as a way of measuring quality of life in cities, a demographer on regularities in the age pattern of migration, a geographer on turnarounds in core–periphery migration, a planner in forecasting migration for land use and transportation modeling, a sociologist on the effect of kinship links on migration destinations, and a real estate specialist on the relationship between housing prices and migration.

The numerous scholarly regional science journals and annual conferences create effective forums for exchanging ideas among disciplines. Although dominated by the abstraction and modeling inherited from economics, these forums are eclectic and open. They have been helpful incubators for scholarly groups who went on to create their own forums, be they input–output analysts, spatial social theorists, urban economists, or transportation modelers. Collectively regional scientists actively borrow ideas from a wide range of sources and combine, adapt, and extend them to regional research.

The recognition Isard anticipated has materialized in many ways other than through a proliferation of university departments. The Regional Science Association became the Regional Science Association International (RSAI), an umbrella organization overseeing three major supraregional organizations, the North American (today the Americas), European, and Pacific coordinating councils. Most countries in Asia, Europe, and North and South America have active regional science associations, and there are several multinational ones, including the British–Irish, French-language, German-speaking, and Australia–New Zealand organizations. New associations have been established in South America including Argentina, Chile, Columbia, and Brazil. The Mexican Association has existed since the 1990s. Many of the individual organizations like, for example, the Western Regional Science Association have a diverse following of foreign members who often rise to leadership positions. Most of the associations sponsor annual meetings, and several have journals. A list of regional science associations may be found at www.reigonalscience.org.

A widely accepted notion is that almost all regional scientists have dual identities, economist and regional scientist, geographer and regional scientist, and so on. For many, the regional science meetings and journals provide the best audience for honing and distributing their research. Others' participation ebbs and flows with changes in their immediate research agendas. No complete census of this fluid regional science community exists. The associations together, including those in China, India, and Brazil, may have over 10 000 members combined. RSAI itself has about 4000 dues-paying members annually.

Regional science journals have become numerous and well established. Six international ones with regional science in their titles are now published by major commercial presses: *Annals of Regional Science*, *International Regional Science Review*, *Journal of Regional Science*, *Papers in Regional Science*, *Regional Science and Urban Economics*, and *Regional Science Policy and Practice*. Echoing the debate half a century ago about the name 'regional science,' some regional science associations leave it out of their journal titles, for example, *Journal of Regional Analysis and Policy Review*, *Urban and Regional Development Studies*, *Growth and Change*, *Investigaciones Regionales*, *Review of Regional Studies*, and *Applied Spatial Analysis*. The RSAI lists 32 related journals worldwide (see www.regionalscience.org).

Book series, position announcements, honorary degrees, and external research prizes are other marks of regional science's role within the social sciences. Noteworthy, too, are the National Science Foundation's Program in Geography and Regional Science in the US and similar ones in other countries. Perhaps the most spectacular recognition of regional science occurred when the Emperor and Empress of Japan personally welcomed participants to the RSAI World Congress in 1996.

Subject Matter

The international regional science community studies a wide range of topics, sometimes ambitiously defined as those social, economic, political, and behavioral phenomena that have a spatial dimension. A common, narrower view groups regional science research under three themes: location theory, methods of regional and spatial analysis, and regional economic development and policy.

Location theory is the study of what social and economic activities are located where and why. Traditionally, emphasis has been on the location of firms and households. The latter considers interregional and international migration and labor markets, as well as intrametropolitan residential mobility and housing markets. Early contributions focused on the role of transportation costs and proximity to suppliers and markets. Recognizing interdependencies of behavior focuses attention on agglomerations or clusters in space, as well as on questions of land use, urban form, housing markets, city size, hierarchies among and systems of cities, market thresholds, and market boundaries. Recognizing interactions among activities and individuals distributed over space raises questions about commuting, trade flows, congestion, transportation, other infrastructure, facility location, taxes, user fees, local public finance, and diffusion of innovations and growth.

The phrase, methods of regional analysis, describes a growing set of techniques used to identify, analyze, or forecast key regional characteristics or changes and to test

theoretical hypotheses. Among the methods are three invented to study national economies but adapted to regions. The inventors received Nobel Prizes in Economic Sciences: Wassily Leontief (1973) "for the development of the input–output method and for its application to important economic problems," Lawrence Klein (1980) "for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies," and Sir Richard Stone (1984) "for having made fundamental contributions to the development of systems of national accounts." Regional scientists have adapted and extended techniques from demography. The cohort-component model focuses on the components of population change and their probabilities (births, deaths, and migration), grouping the population into cohorts by age, sex, and often race or ethnicity. Regional scientists borrowed and adapted the gravity and entropy models from physics and found that they described certain aggregate human behavior well. Migration, trade, transportation flows, and other interaction between two places tend to increase with the product of their masses (e.g., populations) and decrease with distance between them. Regional scientists have also adapted and applied optimization techniques (e.g., linear programming and its many variants) to contexts in which spatial interaction is important, such as transportation networks, location of facilities, and land use.

Regional scientists use statistical methods extensively. Yet here, too, important adaptations and extensions permit the study of spatial and regional phenomena. Geographers, in particular, have extended standard statistical concepts to situations where observations have locations, for example, to draw random samples over space or to determine the likelihood that an observed locational pattern could have occurred by chance. Spatial statistics, spatial econometrics, geographical information systems, and other methods of spatial statistical analysis have become major parts of regional science – for economists, planners, engineers, and others as well as geographers. In the case of statistical methods widely used to study travel demand, residential location, migration, and other behavior with a spatial dimension, the pioneer was an economist, Daniel McFadden, who won the Nobel Prize in 2000 "for his development of theory and methods for analyzing and modeling discrete choice processes."

Regional economic development and policy focuses on why some regions grow and prosper while others lag and even decline, as well as other things such as innovative cluster dynamics, agglomeration impacts, and entrepreneurship as factors or drivers of development and growth. Research ranges from theories of growth and development to the evaluation of specific policies. Persistent questions deal with the divergence and convergence of regional growth rates and incomes, the relationships between core regions and peripheral regions, and the geography and timing of the spread of growth. Other long-standing questions deal with regional labor markets, including the relationship between growth and unemployment and the persistence of wage differences among regions. Key policy issues concern people prosperity vs place prosperity (should the focus be on helping people directly, for example, through education or out-migration incentives or on helping people indirectly by making the region more competitive through infrastructure, tax incentives, or other actions), ways of making lagging regions more competitive and prosperous or more robust regional economies more stable, the rationales for regional policy, and the roles of different levels of government in policy making and implementation.

Policy evaluation has attracted much attention as well. Commonly the objective is to measure the effects of various (nonregional) national policies on regions, for example, free trade treaties, defense spending, or tax reductions. In other cases the objective is to evaluate regional policies, for example, by estimating what would have happened to lagging regions in the absence of the regional support programs. And sometimes the emphasis is on comparing alternative regional plans or policy instruments before they are implemented, generally using cost–benefit, cost-effectiveness, or goals achievement methods augmented to include policy objectives, distributional issues, and hard to quantify outcomes.

Intellectual Progress

Two easily contested, unavoidable decisions must be made before discussing the intellectual progress of regional science. The first is whose contributions should be counted. The open, fluid nature of the regional science community makes this decision difficult, and the early claims to disciplinary status make it controversial. The perspective here is that the idea itself is more important than the primary tribal self-affiliation of its originator. The key criterion is whether the idea is part of the scholarly conversation in regional science meetings and journals. Mention of an individual's contribution does not imply that the person considers himself or herself a regional scientist, although each person mentioned has participated in regional science meetings or published in regional science journals. Debates about whether particular intellectual contributions were initiated by regional scientists, borrowed by regional scientists, or adapted by regional scientists, and who is really a regional scientist are fruitless. The gravity model provides an early case in point. One observer described it as work that may clearly be termed regional science, while another argued that it had been pioneered by outsiders and applied in many disciplines. By far the most interesting fact here is how understanding and use of an idea have evolved.

The second decision concerns the scope of regional science. Something narrower than human activities and phenomena with an important spatial dimension must bound this discussion, but may offer too incomplete a view. Some describe regional science as a toolbox for practical analysis that lacks rigorous theory, while others decry it as a theoretical never-never land with too little realism. The path chosen here recognizes Isard's deliberate efforts to nurture the field of regional science. He edited several of its journals, commissioned papers, supported dissertations, organized meetings, and shaped an influential book series. Three of those books frame a useful way of discussing regional science: Isard's *Location and Space-Economy* (1956), the Isard-led collaborative effort, *Methods of Regional Analysis* (1960), and a reader edited by John Friedmann and William Alonso and dedicated to Isard, *Regional Planning and Development* (1964). The books'

foci – theory, method, and policy – frame the discussion of intellectual progress well.

Location and the Spatial Economy

The quest here is to model mathematically and show graphically both the determinants and consequences of individual location decisions. Regularities in spatial patterns are readily observable, among them are the emergence and persistence of large cities, hierarchical networks of urban places as envisioned in central place theory, clusters of related industries, relatively homogenous residential areas within cities, and regularities in city size distributions. The goal, chiefly of economists who work in this tradition, is to show how these agglomerations can materialize from the interaction of self-serving decisions of individuals and firms. Isard's book drew on the pioneering German contributions by Johann Heinrich von Thünen, Alfred Weber, August Lösch, and Walter Christaller.

Two works from the birth period of modern regional science were extraordinarily seminal. Alonso's dissertation, *Location and Land Use*, focuses on transportation costs and its role in residential decisions and the resultant land-use patterns within a metropolitan area. Concentric rings of land use emerge from the behavior of utility maximizing individuals. Alonso's starting point, like von Thünen's, is the existence of an agglomeration, in this case the central business district of a monocentric city. Charles Tiebout, a founding member of the Regional Science Association, argued that 'people vote with their feet' by choosing locations within the metropolitan area that offer the combination of revenue and expenditure patterns that appeal most to them. Thus, spatial mobility among diverse communities within an urban area provides a solution to the collective consumption problem posed by Samuelson and another mechanism by which individual preferences create spatial variation. Much of today's field of urban economics builds on these two contributions.

A contemporary contribution of this genre is *The Spatial Economy* (1999) by Masahisa Fujita, Paul Krugman, and Anthony Venables, which they describe as a continuation and possible validation of Isard's project. It derives agglomerations instead of positing them as a starting point (e.g., Alonso's central business district around which commuters organize themselves). In the 'new economic geography,' as Krugman (received the Nobel Prize in economics in 2008 for related work) calls it, agglomerations result from the tension between centripetal forces that promote the concentration of economic activities and centrifugal forces that oppose it. Of key interest is when a spatial concentration of economic activity is sustainable, and when a system without spatial concentration breaks down and concentration emerges. Combining formal theoretical models that include increasing returns to scale and computer simulations to study nonlinear dynamic properties, this work links urban, regional, and trade economics in a way that exceeds the dreams of the formal modelers of half a century ago. As was the case then, however, many geographers and others within the regional science community do not embrace this formal mathematical, theoretical mainstream economics approach. They are far more interested in the social, cultural, and institutional factors involved in spatial development and in the study of real places and real regional histories in contrast to abstract ones. The *Journal of Economic Geography* has served to encourage dialog among the different groups.

Also a recent development in economic growth theory, endogenous growth posited by Paul Romer has found strong interest in regional science. Endogenous theory argues that the legendary Solow residual is explained by local endogenous factors that drive innovation and technology change and thus growth. (The Solow residual refers to the large unexplained variance in his regression of national income on labor and capital. He suggested that this variance was due to technology.) As such it highlights the importance of local regional factors in innovation-driven growth and thus has helped drive a strong and growing interest in entrepreneurship and the innovative clustering of industry as factors that help explain local economic development and growth.

Methods of Regional Analysis

A different early quest of modern regional science was the creation of practical tools for the study of real places, and here regional science has had spectacular success. The input–output model attracted early attention. A series of regional input–output articles appeared in the *Review of Economics and Statistics* between 1951 and 1971. Their foci ranged from the theoretical specifications of the models and the definitions of concepts to studies of the nature of metropolitan economies, the location of industrial complexes, and the effects of arms reduction. This initial period culminated in the creation of two landmark regional input–output models constructed from survey-based data, the Philadelphia model of Isard and Thomas Langford, and the West Virginia model of William Miernyk. The next three decades focused on building and using models, with attention paid to the treatment of households and consumption, economic–demographic linkages, and strategies for building the models in a cost-effective way. Here too, progress exceeded the dreams of the 1950s, with regional input–output models now widely used for impact studies in the private and public sectors in many countries of the world. In fact, this method of regional analysis has become an industry with data and software readily available from commercial firms, most notably the IMPLAN system within the United States.

Econometric modeling similarly saw astonishing progress based on the pioneering work of Jean Paelinck in the Netherlands (noted as the father of spatial econometrics) and others in Europe, Japan, and the US. Interest in this methodology grew, eventually resulting in the formation of the Association of Spatial Econometrics established *c.* 2002 and the *Journal of Applied Spatial Econometrics* in 2005. Parallel and in concert with these developments other scholars such as Lawrence Klein sketched how regional econometric models might be developed as satellites of national models, and his student Norman Glickman (Glickman, 1971) developed a fully operational landmark model of the Philadelphia economy. Subsequently many others built regional econometric models as doctoral dissertations, one becoming the blueprint for the National Regional Impact Evaluation System of the US Department of Commerce. Today most states and many companies maintain such forecasting models. This method, too, has become an industry with many companies selling forecasts to private

subscribers. A particular noteworthy case is the eclectic REMI model of George Treyz, which grew from a design at a regional science meeting in 1978 into a widely used commercial product.

Modeling land use and transportation flows within a metropolitan region proved to be arguably the most ambitious search for practical methods, and the most spatially explicit. The intellectual scope of such modeling efforts is daunting: understanding the determinants of the location of industrial, commercial, residential, and other activities and of the movement of people among those locations. These models combined five elements: the theoretical urban economic framework of Alonso and Lowdon Wingo, mathematical programming methods to make them operational starting with the Herbert–Stevens model of land use, gravity models of spatial interaction eventually enriched by the entropy formulations of Alan Wilson, discrete choice statistical methods eventually incorporating McFadden's work, and the transportation forecasting framework established by Robert Mitchell and Chester Rapkin. For nearly five decades, university research groups, notably those led by Britton Harris, Alan Wilson, and David Boyce, have produced a sustained flow of dissertations and other contributions. Yet, unlike the cases of input–output and spatial econometric models, the pioneering research phase did not lead to widespread adoption in routine public and private practice. Dominant practice remains a set of incremental steps – predict population and employment change, decide (usually unaided by models) where those activities will locate, and then use transportation modeling software to forecast the flows among the locations. The research quest continues, large-scale operational models continue to be built (about a dozen were in use in the early 1990s), and the research leads to improvements in the transportation modeling software and other elements of the incremental approach.

Regional Economic Development and Policy

The basic quest here is to understand differences in regional growth and development and then use this knowledge to shape policy. During the early years of modern regional science these concerns attracted the attention of two more Nobel laureates in economics. Gunnar Myrdal was honored in 1974 for his "penetrating analysis of the interdependence of economic, social, and institutional phenomena." His work on economic theory and underdeveloped regions centered on the principle of cumulative causation, a 'vicious circle' of economic and noneconomic factors that led activities to cluster in certain regions and caused regional inequalities. Douglass North was honored in 1993 for "for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change." He studied regional economic growth using exportable commodities or the export base as the key concept, pointing out the incentives regions have to reduce their processing and transportation costs to better the competitive position of their exports. An influential book by Michael Porter, *The Competitive Advantage of Nations* (1990), builds on these notions of clusters of activities and the competitiveness of export industries and finds a theoretical rationale in endogenous growth theory. Porter argues that national advantage derives from successful firms concentrating geographically within a country, their competitive success stimulated by their local (i.e., endogenous) environment, including via government policy at the regional and local level. An earlier formulation remains popular in regional policy. Based on the growth pole research initiated by Francois Perroux, governments concentrate investment in selected urban centers to stimulate the economies of their surrounding areas or draw population away from the crowded, larger metropolitan areas.

Policy and policy analysis are continuing and important parts of the regional economic development theme and draw not only on the above-referenced approaches to its study but to a large body of empirical research used to help advise policy. While policy has been a major focus of many conceptual and empirical studies in regional science, there was no specific policy-focused regional science journal until recently. In 2008, the RSAI established the *Journal of Regional Science Policy and Practice* as a primary outlet for policy-oriented studies. Further, the rise of importance of endogenous growth theory in regional science has led to growing interest and conceptual support in economic policy and practice at the local regional level and can be seen as one of the driving forces leading to the formation of this journal.

Other

Brief mention of some other work completes a broader picture of regional science. RSAI has awarded its Founder's Medal to Martin Beckmann for research in transportation network theory, location theory, and spatial economics; William Alonso for urban economics, spatial demography, and regional development; Jean Paelinck for methods of regional analysis, spatial econometrics, and regional policy; and David Boyce for transportation network models and urban transportation planning methods; Peter Nijkamp for wide and deep contributions to theory, methods, and policy contributions; Antoine Bailly for diverse contributions to scholarship, research, teaching, institution building, international relations, and leadership; and Arthur Getis for spatial statistics and geographic information science.

The three most cited articles in the *International Regional Science Review* are on population dispersal from major metropolitan regions internationally (Vining and Kontuly, 1978), intraurban residential mobility (Quigley and Weinberg, 1977), and the shift of the US population to nonmetropolitan areas (Beale, 1977). In the *Journal of Regional Science* they are on the distribution of residential activity in urban areas (Herbert and Stevens, 1960), determinants and consequences of migration (Greenwood, 1985), and an efficient algorithm to solve the Weber location problem (Kuhn and Kuenne, 1962). In *Regional Science and Urban Economics*, they are on the baby boom and the housing market (Mankiw and Weil, 1989), the demand for housing characteristics (Follain and Jimenez, 1985), and publicly provided inputs and economic growth (Garcia-Mila and McGuire, 1992). In the *Annals of Regional Science*, they are on spatial statistical analysis (Anselin and Getis, 1992), regional inequality (Menczes et al., 2012), and regional disparities in Europe (Ertur and Koch, 2006). The bibliography includes well-cited books and recent overviews of aspects of regional science.

See also: Cities: Internal Structure; Ecological Economics; Location Theory; Migration, Economics of; Regional Geography; Space and Social Theory in Geography; Spatial Interaction; Spatial Pattern, Analysis of; Transportation Geography; Transportation Planning; Transportation, Supply, and Congestion; Urban System in Geography.

Bibliography

Anselin, L., 1988. Spatial Econometrics. Kluwer Academic Publishers, Boston.

Anselin, L., Getis, A., 1992. Spatial statistical analysis and geographic information systems. Annals of Regional Science 26, 19–33.

Beale, C., 1977. The recent shift of the United States population to metropolitan areas 1970–1975. International Regional Science Review 2 (2), 113–122.

Ertur, C., Koch, W., 2006. Regional disparities in the European Union and the enlargement process: an exploratory spatial data analysis, 1995–2000. Annals of Regional Science 40, 726–765.

Follain, J., Jimenez, E., 1985. Estimating the demand for housing characteristics: a survey and critique. Regional Science and Urban Economics 15, 79–107.

Friedmann, J., Weaver, C., 1979. Territory and Function. University of California Press, Berkeley, CA.

Fujita, M., 1989. Urban Economic Theory. Cambridge University Press, Cambridge, UK.

Garcia-Mila, T., McGuire, T., 1992. The contribution of publically provided inputs to states, economies. Regional Science and Urban Economics 78, 177–180.

Glickman, N., 1971. An econometric forecasting model for the Philadelphia region. Journal of Regional Science 11, 15–32.

Greenwood, M., 1985. Journal of Regional Science 25 (4), 521–544.

Hansen, N., Higgins, B., Savoie, D.J., 1990. Regional Policy in a Changing World. Plenum Press, New York.

Henderson, J.V., 1985. Economic Theory and the Cities, second ed. Academic Press, Orlando, FL.

Herbert, T., Stevens, B., 1960. A model for distribution of residential activism urban areas. Journal of Regional Science 2 (2), 21–36.

International Regional Science Review, 1995. 17 (3) and 18 (2) (Regional Science at 40 issues).

Isard, W., Azis, I.J., Drennan, M.P., Miller, R.E., Saltzman, S., Thorbecke, E., 1998. Methods of Interregional and Regional Analysis. Ashgate Publishing, Aldershot, UK.

Journal of the American Planning Association, 1994, 60 (1), 3–44 (Large Scale Models symposium).

Journal of Regional Science, 1985, 24 (4) (Silver Anniversary issue).

Krugman, P., 1991. Geography and Trade. MIT Press, Cambridge, MA.

Kuhn, H., Kuenne, R., 1962. An efficient algorithm for the numerical solution of the generalized Weber problem. Journal of Regional Science 4, 21–34.

Lahr, M., Miller, R. (Eds.), 2001. Regional Science Perspectives in Economic Analysis. Elsevier, Amsterdam.

Mankiw, N., Weil, D., 1989. The baby boom, the baby bust, and the housing market. Regional Science and Urban Economics 19 (2), 235–258.

Markusen, A., 1985. Profit Cycles, Oligopoly, and Regional Development. MIT Press, Cambridge, MA.

Menczes, T., Silveira-Neto, R.M., Azzoni, C.R., 2012. Demography and evolution of regional inequality. Annals of Regional Science 49 (3), 643–655.

Miller, R.E., Blair, P.D., 1985. Input–Output Analysis Foundations and Extensions. Prentice-Hall, Englewood Cliffs, NJ.

Nijkamp, P. (Ed.), 1986. Handbook of Regional and Urban Economics. North Holland, Amsterdam.

Plane, D.A., Rogerson, P.A., 1994. The Geographical Analysis of Population. Prentice-Hall, Englewood Cliffs, NJ.

Quigley, J., Weinberg, D., 1977. Interurban residential mobility: a review and synthesis. International Regional Science Review 2 (1), 41–66.

Stimson, R., Stough, R., Roberts, R., 2006. Regional Economic Development: Analysis and Planning Strategy, second ed. Springer, Heidelberg, Germany.

Vining, D.R., Kontuly, T., 1978. Population dispersal from major metropolitan regions: an international comparison. International Regional Science Review 3 (1), 49–73.

Wilson, A.G., 1970. Entropy in Urban and Regional Modeling. Pion Press, London.

Regression Discontinuity Design

Joseph C Cappelleri, Pfizer Inc, Groton, CT, USA
William M Trochim, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W. M. Trochim, volume 19, pp. 12939–12945, © 2001, Elsevier Ltd.

Abstract

The regression-discontinuity (RD) research design assigns participants to treatment groups solely on the basis of a pretreatment cutoff score, allowing the relative effect of treatment to be studied in participants who most need or deserve a particular treatment. In this article a brief history of the design is given, the basic RD design structure is described, design considerations and variations are highlighted (on cutoff selection, assignment variations, multiple cutoff points, multiple assignment measures, treatment variations, and posttreatment measurement variations, internal validity, measurement error, statistical power), and recent methodological developments are presented (on nonparametric statistical analysis, the 'fuzzy' RD design, and design variations).

Definition

The regression-discontinuity (RD) research design is a quasi-experimental method that can be used to assess the effects of a treatment or intervention. Unique to the RD design is that participants are assigned to groups solely on the basis of a pretreatment cutoff score. The name 'regression-discontinuity' comes from the fact that a treatment effect appears as a 'jump' or discontinuity at the cutoff point in the regression function linking the assignment variable to the outcome. In its simplest form, the design has a pretest or pretreatment (the assignment variable) measure, two groups (those scoring above and below the cutoff), and a posttest or posttreatment (the outcome) measure. More complex variations are also possible.

To many, the RD design initially seems counterintuitive. Using a cutoff to assign participants to the program (or treatment) group and comparison (or control) group creates a pretreatment group nonequivalence that seems like it should lead to selection bias. While the design does induce initial nonequivalence, it does not necessarily lead to biased treatment effect estimates. Instead of assuming pretreatment equivalence (in measured and unmeasured means and variances) as the randomized experiment does, the RD design assumes the two groups are equivalent in their pre–post regression functions in the absence of a treatment effect. This assumption can be tested.

The major attraction of RD designs is that they can be used to estimate the effects of treatments given to those who most need or deserve them – provided that need or merit is determined using a qualification score on the pretreatment assignment variable (i.e., the cutoff) to experimental conditions, and nothing else. Because the design does not require that some needy or deserving individuals get assigned to a no-treatment or comparison group, in some settings it may have ethical advantages over experiments for assessing treatment effects.

History

In a 2008 special issue on RD methodology in the *Journal of Econometrics*, Cook (2008) provided a detailed and comprehensive history of the RD design in three academic disciplines – psychology, statistics, and economics – which covers the reinvention of the design across these disciplines and the differential waxing and waning by discipline (Cook, 2008). Applications of the RD design have also expanded to the health sciences (Trochim, 1990; Trochim and Cappelleri, 1992; Cappelleri and Trochim, 1994, 1995; Finkelstein et al., 1996a, 1996b; Linden et al., 2006; Zuckerman et al., 2006). Many noteworthy contributions using the RD design have been made in the 1960s (Thistlethwaite and Campbell, 1960), 1970s (Goldberger, 1972; Wilder, 1972; Tallmadge and Horst, 1976; Rubin, 1977), 1980s (Berk and Rauma, 1983), 1990s (Cappelleri et al., 1991; Trochim et al., 1991; Mark and Mellor, 1991; Cappelleri et al., 1994; Reichardt et al., 1995; Aiken et al., 1998; Angrist and Lavy, 1999; Berk and de Leeuw, 1999), and 2000s (Van Der Klaauw, 2002; DiNardo and Lee, 2004; Card and Shore-Sheppard, 2004; Jacob and Lefgren, 2004; Ludwig and Miller, 2007; Matsudaira, 2008; Wong et al., 2008; Van Der Klaauw, 2008; Lalive, 2008; Lee, 2008; Gamse et al., 2008).

A growing number of books have featured the RD design. It was first discussed by Campbell and Stanley (1963) and in greater detail in books on social experimentation (Riecken et al., 1974; Bennett and Lumsdaine, 1975). Later the RD design was extensively discussed in texts on quasi-experimentation (Cook and Campbell, 1979; Shadish et al., 2002) and on evaluation (Judd and Kenny, 1981; Mohr, 1995). Trochim (1984) wrote the first and, to date, only book devoted exclusively to the method. With the dawning of the twenty-first century several books discussed the RD design, presenting updates on its methodology and applications, including (but not limited to) books by Fleiss et al. (2003), Senn (2007), Angrist and Pischke (2009), and Murnane and Willett (2011).

The Basic RD Design Structure

Figure 1 shows an example of the RD design using simulated data to depict the case of a compensatory treatment – perhaps a reading program designed to help students who initially score poorly (defined as below the pretest cutoff) on a standardized reading measure. The figure is a standard bivariate plot of

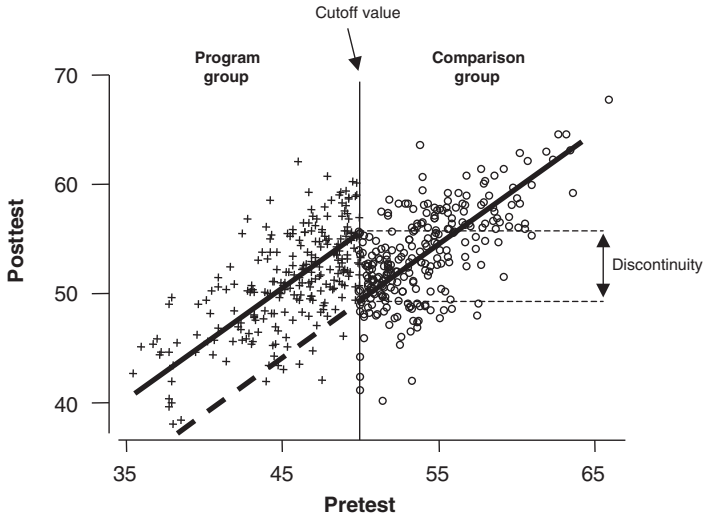

Figure 1 Basic regression-discontinuity design for a compensatory treatment showing a positive effect.

posttest and pretest for 500 simulated participants. Data points for the treatment recipients are represented with '+,' while the comparison cases are 'o.' The cutoff value in this example occurs at the middle of the pretest distribution (pretest = 50) and is indicated by the vertical line. All students scoring below 50 are defined as needing compensatory training and get the treatment, while the remaining (i.e., those to the right of the cutoff) are assigned to the comparison (control) group.

The solid lines represent the straight-line regression of posttest on pretest for each group. In the absence of a treatment effect, the critical assumption is that the regression line in the comparison group would continue to the left of the cutoff into the treatment group region, as indicated by the heavy dashed line. That is, if the reading program does not work, then the observed treatment group regression line would be where the dashed line is. The fact that the observed treatment group line is displaced from this 'expected' line suggests a treatment effect – treatment group participants scored higher on the posttest (i.e., vertically) than would have been predicted from the comparison group pre–post relationship. The size of the treatment main effect is estimated as the vertical jump or *discontinuity* between the *regression* lines at the cutoff, hence the name *regression-discontinuity*. Assuming that higher scores on both measures indicate better performance, one can conclude from the figure that the treatment improved participant performance by the amount of the vertical discontinuity.

In experimental designs random assignment assures that the treatment and comparison groups are initially probabilistically equivalent. Posttreatment differences can be then attributed to the intervention. In the RD design one obviously does not expect this kind of equivalence. Instead, it is assumed that the *relationship* between the assignment variable and the outcome is equivalent for the two groups – that is, the same continuous regression line describes both groups. Thus, interpretation of the RD design depends mostly on two factors: (1) that no alternative cause could induce a discontinuity at the cutoff and (2) that we can perfectly model the pre–post relationship. These issues, particularly the second one, constitute the major problems in the statistical analysis of the RD design (Trochim, 1984, 1990; Winship and Morgan, 1999).

Design Considerations and Variations

Selection of the Cutoff

The RD design requires that the assignment and outcome measures be of sufficient quality of measurement to warrant an appropriate regression analysis. The choice of cutoff value can be made solely on the basis of available resources. For instance, if a program can handle only 25 persons and 70 apply, then a cutoff point can be selected that distinguishes the 25 most 'needy' persons from the rest. The cutoff can also be chosen on substantive grounds. If the assignment measure is an indication of severity of illness measured on a one to seven scale, relevant experts might contend that all those scoring an average of five or more are ill enough to deserve treatment. Therefore, a theoretically justified cutoff of five could be used.

Assignment Variations

A 'sharp' RD design has its pretreatment cutoff value followed perfectly, without exception. While the use of a pretreatment cutoff distinguishes the RD design, it is often difficult to implement. In principle it does not allow room for professional judgment or discretion. In fact, the design does not preclude incorporating such judgment so long as it can be quantified or explicitly accounted for when specifying the cutoff value.

Violation of the assignment rule leads to 'misassignment' and to what is often termed the 'fuzzy' RD design. Typically, misassignment is most prevalent for 'close call' cases that fall near the cutoff. For instance, teachers or parents may successfully argue that students who just miss the cutoff for a novel compensatory tutoring program should be assigned to it based on grounds other than their assignment score. In the fuzzy RD case, standard analyses will yield biased estimates of treatment effect. While statistical procedures for dealing with

misassignment have been suggested (Trochim, 1984), misassignment is better avoided if possible.

Multiple Cutoff Points

RD designs are not limited to a single cutoff value. In the absence of treatment the assumption is that the pre–post relationship can be described as a single continuous regression line extending over the entire range of the assignment scores. Even when multiple cutoff points are used, the comparison group's pre–post relationship is still used as the counterfactual for projecting where other groups' regression lines should be if there is no treatment effect.

The simplest multiple cutoff case involves two cutoff scores. In this variant, persons scoring above the higher cutoff might be assigned to one treatment, those scoring below the lower cutoff to another, and those scoring within the cutoff interval might be controls. The assignment of treatments to each part of the assignment distribution can be made in alternative ways, including by random assignment. For instance, individuals within the interval between the two cutoffs could be randomly assigned to conditions, while those scoring below the lower cutoff and above the upper cutoff are analyzed as from an RD design (Boruch, 1973). Trochim and Cappelleri (1992) discuss four other ways of coupling random assignment to RD design features.

Multiple cutoff points are also involved in designs where the treatment is implemented a series of times (i.e., in phases), perhaps beginning with the most needy before successively giving it to those less needy. In the first phase, one uses a single cutoff value, assigning those most in need to the treatment and then measuring posttreatment performance. Then, the cutoff value would be moved to include the next most needy group on the assignment measure. Eventually, everybody could receive the treatment, but in groups ordered by need and phase. This is particularly useful in organizations seeking to 'roll out' a treatment over time, but eventually serving all.

Multiple Assignment Measures

The requirement of strict adherence to the cutoff criterion entails a single quantitative indicator. When this does not capture the degree of pretreatment need well, Trochim (1990) discusses two multivariate strategies. The first involves the use of several separate measures, each having its own cutoff value. This is most feasible when a prospective participant must meet the different cutoff criteria on measure 1 'and' measure 2 'and' measure 3, and so on. Less feasible is an 'or' rule when assignment depends on meeting a fixed number of criteria from a larger set – say, 5 of 10. The second involves assignment according to a new variable that is a composite of several individual measures. Using multiple measures this way can help incorporate more judgment into a quantitative assignment strategy.

Treatment Variations

When treatment comparisons are absolute, the control (or comparison) group receives no formal or placebo treatment. When they are relative, the comparison group receives an alternative treatment, very often the current standard treatment to which a new one is to be compared.

An advantage of the RD design is that it enables assignment of progressively riskier treatments to those in greater need. Assume three treatment programs: the standard plus two increasingly riskier experimental ones. It is then possible to have two cutoff points and to assign the riskiest treatment to the neediest patients, the least risky (i.e., standard) treatment to the least needy, and the moderately risky treatment to those falling between the two cutoffs. (If no a priori basis exists for making risk judgments, an alternative would be to use a single cutoff and to assign those least in need to standard treatment while randomly assigning those on the other side of the cutoff into one of the two experimental conditions.) RD designs are flexible for examining treatment variations.

Posttreatment Measurement Variations

In RD designs, pre and post measures do not have to be on the same type of measure. For instance, one might assign persons to a health education treatment on the basis of household income but then examine its effects on attitudes toward health; here the assignment variable is an economic indicator and the outcome is a cognitive one.

The RD design is not restricted to a single outcome, and for each posttreatment measure there can be a separate RD design and analysis. However, it will often be useful to create an aggregate outcome or outcomes. For instance, we may have many knowledge items that are scaled to give a total (aggregate) score, with subtest scores for mathematical reasoning, computational skills, verbal reasoning, analogies, and so on. We might then conduct separate analyses for both the total score and each subtest.

Outcome measures need not be continuous normal variables. While such distributions facilitate RD analysis, they are not necessary for it. For example, Berk and Rauma (1983) evaluated the effects of a California law that extended unemployment benefits to released prisoners who had not previously been eligible. In order to qualify, prisoners had to earn at least $1500 working in the prison over the 12 months prior to release. Would extending these benefits reduce subsequent recidivism? A dichotomous outcome was created with 1 indicating recidivism and 0, staying out of prison. To analyze this outcome, Berk and Rauma (1983) relied on a linear random utility model related to the binary logit model.

Internal Validity

Internal validity refers to the degree to which causal inference is reasonable because alternative explanations for an apparent treatment effect can plausibly be ruled out. The RD design is very strong with respect to internal validity because the process of selection into treatments is fully known and only factors that would serendipitously induce a discontinuity in the pre–post relationship at the cutoff can be considered threats. However, validity also depends on how well the analyst can model the true pre–post relationship so that underlying nonlinear relationships do not masquerade as discontinuities at the cutoff point.

The RD design does not yield biased estimates of the treatment effect when the assignment variable is measured with

error (Cappelleri et al., 1991; Trochim et al., 1991). The reason such error does not induce bias here, although it does in other quasi-experiments, is that assignment to treatment is by the fallible assignment variable and not by some latent underlying selection process, as happens in quasi-experiments.

Statistical Power

It takes approximately 2.75 times as many participants in an RD design to achieve statistical power comparable to a simple randomized experiment, assuming equal numbers assigned to the treatment and control groups (Goldberger, 1972). When RD is combined with random assignment within an interval around this cutoff, 2.48 times the number of respondents are needed if 20% of participants are randomized, 1.96 times with 40% randomization, 1.46 with 60%, and 1.14 with 80% (Cappelleri and Trochim, 1994, 1995, 2010).

Parametric Analysis

Assumptions of RD Analysis

Five central assumptions must be made in order for responsible analysis of data from an RD design:

1. The cutoff criterion: The cutoff criterion must be followed without exception. Otherwise, unless some valid adjustment is made, a selection threat arises, and estimates of the treatment effect are likely to be biased.
2. The pre–post distribution: The true pre–post distribution must be known and correctly specified as linear, polynomial, logarithmic, or the like. Reichardt et al. (1995) point out the difficulties in specifying the appropriate model, especially when some curvilinearity exists in the pre-post relationship.
3. Comparison group pretest variance: There must be a sufficient range of pretest values in the comparison group to enable adequate estimation of the pre–post regression line for that group. Variability in the treatment group is also desirable, although not strictly required because one can project the comparison group line even to a single treatment group point.
4. Continuous pretest distribution: Both groups must come from a single continuous pretest distribution. In some cases one might find intact groups that serendipitously divide on a measure so as to imply some cutoff (e.g., patients from two different geographic locations). However, such naturally discontinuous groups would confound geography with treatment and should not be used.
5. Treatment implementation: It is assumed that the treatment is uniformly delivered to all recipients – that is, they all receive the same dosage, length of stay, amount of training, or whatever. If not, it is necessary to model explicitly the treatment as implemented, thus complicating the analysis considerably.

A Model for the Basic RD Design

The model presented here is for the basic RD design. Given a pretreatment assignment measure, x_i, and a posttreatment measure, y_i, the model for the expected value (E) of y_i can be stated as follows:

$$E(y_i) = \beta_0 + \beta_1 z_i + \beta_2 \tilde{x}_i + \beta_3 \tilde{x}_i z_i + \cdots + \beta_{n-1} \tilde{x}_i^s + \beta_n \tilde{x}_i^s z_i$$

where

$\tilde{x}_i$ = pretreatment measure for individual i minus the cutoff value (i.e., $\tilde{x}_i = x_i - x_0$)
y_i = posttreatment measure for individual i
z_i = assignment variable (1 if treatment participant; 0 if comparison participant)
s = the degree of the polynomial for the associated $\tilde{x}_i$
β_0 = parameter for comparison group intercept at cutoff x_0
β_1 = treatment effect parameter
β_2 = linear slope parameter
β_3 = linear interaction parameter
β_{n-1} = parameter for the sth-order polynomial
β_n = parameter for the sth-order polynomial interaction term

The major null hypothesis of interest H_0: $\beta_1 = 0$ is tested against the alternative H_1: $\beta_1 \neq 0$. The model estimates both main and interaction effects at the cutoff point in order to examine changes in means and slopes. It postulates a polynomial pre–post relationship. It also requires subtracting the cutoff score from each pretest score. The term $\tilde{x}_i$ has a superscript tilde to indicate this transformation of the pretest x_i.

Finally, the model allows for any order of polynomial function (although certain restrictions are made in specifying the function as described below). Thus, the true pre–post relationship can in theory be linear, quadratic, cubic, quartic, and so on.

In the simplest model with no interaction term, which includes only treatment and the pretest (pretreatment) assignment covariate as predictors, the treatment effect (β_1) is constant across pretest assignment scores and, therefore, the same at the cutoff score as any other pretest score on the assignment continuum. In a model with a linear treatment-by-pretest interaction, the treatment effect is no longer constant across the (pretest) assignment continuum but depends linearly on the value of the pretest assignment score, allowing for the population slopes of the pre–post relationship to differ on opposite sides of the discontinuity.

Model Specification

The key analytic problem is correctly specifying the model for the data – in this example, as a polynomial model. No simple or mechanical way exists to determine definitively the appropriate model. As is the case with any judicious statistical modeling, the RD analysis requires judgment and discretion, along with conducting multiple analyses based on different assumptions about the true pre–post relationship.

Steps in the Analysis

Recall what data the basic RD design provides on each unit. There is a pretest assignment value x_i in the model. Knowing this and the cutoff value allows creating a new variable z_i, which is equal to 1 in the treatment group and 0 if not. Finally, there is a posttest (or posttreatment) score labeled y_i in the

model. Given the variables x_i, z_i, and y_i, the steps to be followed in the polynomial analysis are as follows:

1. Transform the pretest

The analysis begins by subtracting the cutoff value from each pretest score thus creating the term $\tilde{x}_i$ as in the model. This sets the intercept equal to the cutoff so that estimates of effect are made at the cutoff rather than at $x_i = 0$.

2. Examine the relationship visually

It is important to determine whether there is any visually discernible discontinuity at the cutoff. It could be a change in level, slope, or both. If a discontinuity is visually clear at the cutoff one should not be satisfied with analytic results that indicate no effect. However, if no discontinuity is apparent, variability in the data may be masking an effect, and one must attend carefully to the analytic results.

The second thing to look for is the degree of polynomial that may be required as indicated by the overall slope of the distribution, but particularly in the comparison group part. A good approach is to count the number of flexion points (how often the distribution 'flexes' or 'bends'). A linear distribution implies no flexion points, while a single flexion point would indicate a quadratic function, and so on, such that the estimated polynomial is equal to one order less than the total number of flexion points (i.e., with two flexion points one would hypothesize a cubic or third-order polynomial). This information is used to specify the initial model.

3. Create higher order terms and interactions

Depending on the number of flexion points, create transformations of the transformed assignment variable $\tilde{x}_i$. The rule of thumb is to go two orders of polynomial higher than indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (i.e., no flexion points), one would want to create transformations up to a second-order $(0 + 2)$ polynomial. The first-order polynomial already exists in the model $(\tilde{x}_i)$, and one would only have to create the second-order polynomial by squaring $\tilde{x}_i$ to obtain $\tilde{x}_i^2$. For each transformation of $\tilde{x}_i$ one also creates the interaction term by multiplying the polynomial by z_i. In this example there would be two interaction terms: $\tilde{x}_i z_i$ and $\tilde{x}_i^2 z_i$. If there seems to be two flexion points in the bivariate distribution, one would then create transformations up to the fourth $(2 + 2)$ power and their interactions. This rule of thumb errs toward overestimating the true polynomial function needed, for reasons outlined in Trochim (1984).

4. Estimate the initial model

The true analysis can now begin. One simply regresses the posttest scores, y_i, on $\tilde{x}_i$, z_i and all higher order transformations and interactions created in step 3 above. The regression coefficient associated with the z_i term (i.e., the group membership variable) is the estimate of the main effect of the treatment. If there is a vertical discontinuity at the cutoff it will be estimated by this coefficient. One can test the significance of the coefficient (or any other) by constructing a standard t-test. If the analyst correctly overestimated the polynomial function required to model the distribution at step 3, then the treatment effect estimate will be unbiased. However, by initially including terms that may not be needed in the true model, the estimate is likely to be inefficient, that is, standard error terms will be inflated and statistical significance underestimated. However, if the coefficient for the effect is highly significant in this initial model, it would be reasonable to conclude there is an effect. Interaction effects can also be examined. A linear interaction is implied by a significant coefficient for the $\tilde{x}_i z_i$ term.

5. Refine the model

The procedure described thus far is conservative, designed to reduce the chances of a biased treatment effect estimate even at the risk of increasing the error of the estimate (Trochim, 1984). Therefore, on the basis of the results of step 4 one might wish to attempt to remove apparently unnecessary terms and reestimate the treatment effect with greater efficiency. This is tricky and should be approached with caution lest it introduces bias. One should certainly examine the regression results from step 4, noting the degree to which the overall model fits the data, the presence of any insignificant coefficients, and the pattern of residuals. One might examine the highest order term in the current model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit, one might then drop these two terms and reestimate the model. One would repeat this procedure until (1) all the coefficients are significant, (2) the goodness-of-fit measure drops appreciably, or (3) the pattern of residuals indicates a poor fit.

Analyses with Multiple Cutoff Points

The basic model described above can be applied directly when multiple cutoff points are used and random assignment is followed within cutoff intervals. If assignment within intervals is nonrandom, one must also address the potential for selection bias in the analysis. When multiple cutoffs are used to distinguish separate treatments (i.e., multiple treatments are not assigned within the same cutoff interval), one would have to construct multiple treatment assignment variables for the analytic model (e.g., z_1, z_2, z_3) and all necessary interaction terms. Clearly, as more cutoffs and groups are added, model specification becomes more complex.

Analyses with Multiple Assignment Measures

Imagine having multiple measures, each with its own cutoff. Each assignment measure is then transformed by having its own cutoff value subtracted from it to create $\tilde{x}_i$. The analysis would then include all transformed assignment measures, group membership, higher order terms, and interactions. For the simple first-order (or linear) case, one could use the following model:

$$E(y_i) = \beta_0 + \beta_1 z_i + \beta_2 \tilde{x}_{1i} + \beta_3 \tilde{x}_{2i} + \beta_4 \tilde{x}_{3i} + \beta_5 \tilde{x}_{1i} z_i + \beta_6 \tilde{x}_{2i} z_i + \beta_7 \tilde{x}_{3i} z_i$$

This model does not include any two-way assignment variable multiplicative terms (e.g., $\tilde{x}_{1i} x_{2i}$ or $\tilde{x}_{1i} x_{2i} z_i$) or any three-way assignment variable multiplicative terms (e.g.,

$\tilde{x}_{1i}x_{2i}x_{3i}$ or $\tilde{x}_{1i}x_{2i}x_{3i}z_i$), but such an assumption may be reasonable where primary interest is in estimating β_1, the treatment effect at the cutoff, and not interactions. Clearly, however, the use of multiple assignment variables with higher order polynomial models will quickly lead to an unwieldy analysis.

The situation is simpler when multiple assignment variables are rescaled to reflect the number of assignment variables on which a person meets the criterion (e.g., on 5 of 10). This number then becomes the x_i variable, and one would conduct the basic analysis described earlier. When multiple assignment variables are combined into a single index, the analysis is also a straightforward application of the basic procedures described initially.

Recent Methodological Developments

Nonparametric Statistical Analysis

As noted, the most serious limitation of RD designs is the possible sensitivity of the functional form between the posttest outcome and pretest assignment covariate. Traditionally, parametric models with polynomial regression have been used to analyze RD design. If the extrapolation beyond the cutoff score is not adequate, then what masquerades as a jump due to treatment might simply be a nonlinear relationship showing no real treatment effect.

To reduce the likelihood of such mistakes, economists have proposed and implemented nonparametric approaches that look only at the data in the neighborhood around the cutoff score (Hahn et al., 2001; Ludwig and Miller, 2007; Imbens and Lemieux, 2008; Angrist and Pischke, 2009). Comparisons of average outcome scores in small-enough neighborhoods to the left and right of the cutoff can be estimated in a way that does not depend on the correct functional form. Instead of providing a constant treatment effect generalizable across all values of the pretest assignment covariate (in the case of a model with no pretest-treatment interaction), as occurs with the parametric approach, the average treatment effect in the nonparametric approach is restricted to (and conditional on) the cutoff score.

The nonparametric approach to RD requires good estimates of the mean posttreatment scores in small neighborhoods to the immediate right and left of the cutoff. However, such estimation can be tricky. One notably conspicuous problem is the sparse or limited data in a small neighborhood of the cutoff means. Local linear regression, a nonparametric smoothing technique, has been proposed to address the typically relatively scant data around the cutoff means (Hahn et al., 2001; Imbens and Lemieux, 2008; Angrist and Pischke, 2009).

Nonparametric RD estimation is not without its shortcomings (Bloom, 2012). The method requires very large samples to provide an adequate number of observations in the two bandwidths adjacent to the cutoff point. Another limitation is the potential sensitivity of nonparametric estimations to the choice of bandwidth. A trade-off is needed between, on the one hand, introducing bias from a bandwidth that is too wide and, on the other hand, losing precision from a bandwidth that is too narrow. The most widely used empirical approach for making such a trade-off is cross-validation (Imbens and Lemieux, 2008).

Fuzzy RD Design

As noted previously, a violation of the assignment rule to treatment and comparison groups results in a fuzzy RD design. Here external factors result in individuals being placed in one group when they should have been placed in the other group according to their pretreatment assignment score. A conventional analysis of a fuzzy RD design, which does not account for this misassignment, would result in a biased treatment effect. Under certain circumstances, a potential solution to obtain an unbiased treatment effect in the fuzzy RD design is to apply the technique of instrumental variables (Angrist and Pischke, 2009; Murnane and Willett, 2011). Here the pretreatment (or pretest) assignment measure can serve as a credible instrumental variable for actual group membership if this pretreatment assignment measure predicts actual group membership and, at the same time, is not directly related to the posttreatment measure.

A two-stage least squares regression method (a parametric approach based on ordinary least squares) to instrumental variable estimation has been proposed and applied to analyze the fuzzy RD design (Jacob and Lefgren, 2004; Angrist and Pischke, 2009; Murnane and Willett, 2011). The nonparametric version of fuzzy RD consists of instrumental variable estimation in a small neighborhood around the discontinuity. Hahn et al. (2001) developed a nonparametric procedure for instrumental variables using local linear regression around the cutoff, which was applied to estimate the effect of class size on childrens' test scores (Angrist and Lavy, 1999).

In research related to instrumental variables, Battistin and Retorre (2008) and Bloom (2012) provide additional insights and expanded paradigms for fuzzy RD designs. They provide conditions that are required to make fuzzy RD designs valid in identifying average treatment effects.

Design Variations

Building on previous research on logistical and analytical issues of RD designs with varying amounts of randomization (Trochim and Cappelleri, 1992; Cappelleri and Trochim, 1994, 1995, 2010; Cappelleri et al., 1994), Mandell (2008) provides extended formulations and insights into this hybrid design in which cutoff-based assignment and random assignment are combined depending on values of the pretreatment assignment indicator to experimental conditions. Different hybrid designs can have varying degrees of randomization and discontinuity. They entail smaller sacrifices in statistical efficiency than the basic or pure RD design (which has no randomization embedded within it) and are also likely, in certain circumstances, to be judged considerably fairer than a true randomized experiment and possibly even a pure RD design.

Another variation is the clustered RD design where groups (rather than individuals) are assigned to an intervention. Considerably more methodological work has been performed on the design and analysis of RD designs when treatments are applied to individuals (Shadish et al., 2002;

Imbens and Lemieux, 2008) than to groups. More recently, however, increased methodological research on RD designs has been applied in the context of community or group-level interventions. The clustered RD design, for example, was the primary design to evaluate the federal education program mandated in the No Child Left Behind Act of 2001 (Gamse et al., 2008). Motivated by 'No Child Left Behind,' Schochet (2009) examined the statistical power under such clustered RD designs (without randomization) using techniques from the causal inference and hierarchical linear modeling literature. The main conclusion is that three to four times larger samples are typically required under the clustered RD design than under the clustered randomized design to achieve estimates of effects with the same level of precision.

Pennell et al. (2011) noted that designs in education are often quite different from designs used in public health and investigated the design and analysis of clustered RD designs in which a varying proportion of groups is randomized. These researchers extend upon previous work to consider analysis, statistical power, and sample size implications of RD designs in community-based intervention studies in public health. In doing so they examined the power of these designs as a function of the intraclass correlation, number of groups, and number of members per group and compared the results with the traditional group randomized trial.

Variations on the RD design can also be distinguished by the nature of the assignment. Research on the RD design has typically focused on applications with a single assignment variable. In many settings, however, it may be more appropriate to impose cutoffs on several assignment variables in defining a set of different treatments. For instance, cutoffs on two assignment variables in a basic RD design would result in two dimensions with four treatment regions. By using regression to model the response surface in each region, and obtaining predicted values along each of the discontinuity edges, Papay et al. (2011) show how to generalize the basic RD design to include multiple assignment variables simultaneously and to estimate the effects of several treatments.

Conclusions

From a methodological point of view, inferences drawn from a well-implemented RD design are considered comparable in internal validity to conclusions from randomized experiments. However, the lower statistical efficiency of RD designs, and the resulting sample size demands, may limit the design's utility. From an ethical perspective, RD designs are compatible with the goal of getting the treatment to those most deserving or in need. It is not necessary to deny the treatment from potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design may be directly usable when allocation formulas are the basis for assigning treatment or treatments.

With all these considerations in mind, the RD design must be used judiciously. In general, the randomized experiment is still the method of first choice when assessing causal hypotheses (Cappelleri and Trochim, 2010). However, where they are ruled out as impractical, the RD design should be considered as a practical high-quality alternative.

See also: Causal Counterfactuals in Social Science Research; Comparative Studies: Method and Design; Control Variables in Research; External Validity; Internal Validity; Nonequivalent Group Designs.

Bibliography

Aiken, L.S., West, S.G., Schwalm, D.E., Carroll, J., Hsuing, S., 1998. Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: efficacy of a university-level remedial writing program. Evaluation Review 22, 207–244.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Angrist, J.D., Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114, 533–575.

Battistin, E., Retorre, E., 2008. Ineligibles and eligibles non-participants as a double comparison group in regression discontinuity designs. Journal of Econometrics 142, 715–730.

Bennett, C.A., Lumsdaine, A.A. (Eds.), 1975. Evaluation and Experiment. Academic Press, New York, NY.

Berk, R.A., Rauma, D., 1983. Capitalizing on nonrandom assignment to treatment: a regression discontinuity of a crime control treatment. Journal of the American Statistical Association 78, 21–28.

Berk, R.A., de Leeuw, J., 1999. An evaluation of California's inmate classification system using a generalized regression discontinuity design. Journal of the American Statistical Association 94, 1045–1052.

Bloom, H.S., 2012. Modern regression discontinuity analysis. Journal of Research on Educational Effectiveness 5, 43–82.

Boruch, R.F., 1973. Regression-Discontinuity Designs Revisited. Unpublished manuscript, Northwestern University.

Campbell, D.T., Stanley, J.C., 1963. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L. (Ed.), Handbook of Research on Teaching. Rand McNally, Chicago, IL (Also published as Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally, 1966.).

Cappelleri, J.C., Trochim, W.M.K., 2010. Cutoff designs. In: Chow, S.-C. (Ed.), Encyclopedia of Biopharmaceutical Statistics, third ed. revised and expanded. Informa Healthcare, New York, NY, pp. 386–391.

Cappelleri, J.C., Trochim, W.M.K., 1995. Ethical and scientific features of cutoff-based designs of clinical trials: a simulation study. Medical Decision Making 15, 387–394.

Cappelleri, J.C., Darlington, R.B., Trochim, W.M.K., 1994. Power analysis of cutoff-based randomized clinical trials. Evaluation Review 18, 141–152.

Cappelleri, J.C., Trochim, W.M.K., 1994. An illustrative statistical analysis of cutoff-based randomized clinical trials. Journal of Clinical Epidemiology 47, 261–270.

Cappelleri, J.C., Trochim, W.M.K., Stanley, T.D., Reichardt, C.S., 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. Evaluation Review 15, 395–419.

Card, D., Shore-Sheppard, L.D., 2004. Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansions on low-income children. Review of Economics and Statistics 86, 752–766.

Cook, T.D., 2008. "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics 142, 636–654.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis for Field Settings. Rand McNally, Chicago, IL.

DiNardo, J., Lee, D.S., 2004. Economic impacts of new unionization on private sector employers: 1984–2001. Quarterly Journal of Economics 119, 1383–1441.

Finkelstein, M.O., Levin, B., Robbins, H., 1996a. Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. American Journal of Public Health 86, 691–695.

Finkelstein, M.O., Levin, B., Robbins, H., 1996b. Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. American Journal of Public Health 86, 696–705.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, third ed. John Wiley & Sons, New York, NY.

Gamse, B.C., Jacob, R.T., Horst, M., Boulay, B., Unlu, F., 2008. Reading First Impact Study Final Report (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Goldberger, A.S., 1972. Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations. Discussion Papers, 123–72. Institute for Research on Poverty, University of Wisconsin, Madison, WI.

Hahn, J., Todd, T., Van Der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrics 69, 201–209.

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142, 625–635.

Jacob, B.A., Lefgren, L., 2004. Remedial education and student achievement: a regression-discontinuity analysis. Review of Economics and Statistics 86, 226–244.

Judd, C.M., Kenny, D.A., 1981. Estimating the Effects of Social Interventions. Cambridge University Press, New York, NY.

Lalive, R., 2008. How do extended benefits affect unemployment duration? a regression-discontinuity approach. Journal of Econometrics 142, 785–806.

Lee, D.S., 2008. Randomized experiments from non-random selection in U.S. House elections. Journal of Econometrics 142, 675–697.

Linden, A., Adams, J.L., Roberts, N., 2006. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. Journal of Evaluation in Clinical Practice 12, 124–131.

Ludwig, J., Miller, D.L., 2007. Does head start improve children's life chances? evidence from a regression discontinuity design. Quarterly Journal of Economics 122, 159–208.

Mandell, M.B., 2008. Having one's cake and eating it, too: combining true experiments with regression discontinuity designs. Evaluation Review 32, 415–434.

Mark, M.M., Mellor, S., 1991. Effect of self-relevance of an event on hindsight bias: the foreseeability of a layoff. Journal of Applied Psychology 76, 569–577.

Matsudaira, J.D., 2008. Mandatory summer school and student achievement. Journal of Econometrics 142, 829–850.

Mohr, L.B., 1995. Impact Analysis for Program Evaluation, second ed. Sage Publications, Thousands Oaks, CA.

Murnane, R.J., Willett, J.B., 2011. Methods Matter: Improving Causal Inference in Educational and Social Science Research. Oxford University Press, New York, NY.

Papay, J.P., Willett, J.B., Murnane, R.J., 2011. Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics 161, 203–207.

Pennell, M.L., Hade, E.M., Murray, D.M., Rhoda, D.A., 2011. Cutoff designs for community-based intervention studies. Statistics in Medicine 30, 1865–1882.

Reichardt, C.S., Trochim, W.M.K., Cappelleri, J.C., 1995. Reports of the death of regression-discontinuity analysis are greatly exaggerated. Evaluation Review 19, 39–63.

Riecken, H.W., Boruch, R., Campbell, D.T., Caplan, N., Glenman, T.K., et al., 1974. Social Experimentation: A Method for Planning and Evaluating Social Intervention. Academic Press, New York, NY.

Rubin, D.B., 1977. Assignment to treatment on the basis of a covariate. Journal of Educational Statistics 2, 1–26.

Schochet, P.Z., 2009. Statistical power of regression discontinuity designs in education evaluations. Journal of Educational and Behavioral Statistics 34, 238–266.

Senn, S.J., 2007. Statistical Issues in Drug Development, second ed. John Wiley & Sons, Chichester, UK.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston, MA.

Tallmadge, G.K., Horst, D.P., 1976. A Procedural Guide for Validating Achievement Gains in Educational Projects. Office of Education, Washington, DC.

Thistlethwaite, D.L., Campbell, D.T., 1960. Regression-discontinuity analysis: an alternative to the *ex post facto* experiment. Journal of Educational Psychology 51, 309–317.

Trochim, W.M.K., Cappelleri, J.C., 1992. Cutoff assignment strategies for enhancing randomized clinical trials. Controlled Clinical Trials 13, 190–212.

Trochim, W.M.K., Cappelleri, J.C., Reichardt, C.S., 1991. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. Evaluation Review 15, 571–604.

Trochim, W.M.K., 1990. Regression-discontinuity design in health evaluation. In: Sechrest, L., Perrin, E., Bunker, J. (Eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. U.S. Department of Health & Human Services, Rockville, MD, pp. 119–139.

Trochim, W.M.K., 1984. Research Design for Program Evaluation: The Regression-Discontinuity Approach. Sage, Beverly, CA.

Van Der Klaauw, W., 2008. Breaking the link between poverty and low student achievement: an evaluation of title I. Journal of Econometrics 142, 731–756.

Van Der Klaauw, W., 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. International Economic Review 43, 1249–1287.

Wilder, C.S., 1972. Physician Visits, Volume, and Interval since Last Visit, U.S., 1969. National Center for Health Statistics, Rockville, MD (Series 10, No. 75; DHEW Pub. No. HSM 72-1064).

Winship, C., Morgan, S.L., 1999. The estimation of causal effects from observational data. Annual Review of Sociology 25, 659–707.

Wong, V.C., Cook, T.D., Barnett, W.S., Jung, K., 2008. An effectiveness-based evaluation of five state pre-kindergarten programs. Journal of Policy Analysis and Management 27, 122–154.

Zuckerman, I.H., Lee, E., Wutoh, A.K., Zue, Z., Stuart, B., 2006. Application of regression-discontinuity analysis in pharmaceutical health services research. Health Services Research 41, 550–563.

Regulation and Administration

Bronwen Morgan, The University of New South Wales (UNSW), Sydney, NSW, Australia
Susan Sterett, University of Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Sterett, volume 19, pp. 12945–12951, © 2001, Elsevier Ltd.

Abstract

Regulation is the practice of shaping significant institutions, particularly businesses, so as to take into account social purposes that might not be taken into account in the ordinary process of doing business. According to one definition, it is the imposition of legal orders on those engaged in economic activity. The social purposes of regulation can range from promoting economic efficiency, or protection from harm or risk, to promoting diversity or equity. States regulate capitalist enterprises within capitalist or mixed economies – controlling aspects of, for example, the price of services, entry into occupations, and the products and processes of businesses.

Regulation is the practice of shaping significant institutions – particularly business, but also industry, universities, governments – so as to take into account social purposes. Regulation often aims to promote some social concerns that might not be taken into account in the ordinary process of doing business. According to one definition, it is the imposition of legal orders on those engaged in economic activity (Kagan, 1978). States regulate capitalist enterprises within capitalist or mixed economies, often focusing on the price of services, entry into occupations, and the products and processes of businesses.

Cross nationally and across different policy areas, various institutions take on the task of regulation, including government boards, inspectorates, specialized courts, and independent commissions. How states are organized shapes the processes of regulation: to what extent courts are involved, how open procedures designing regulation are to interest group participation, who may be challenged in litigation. With the increasing proliferation of issues that cross or transcend national borders – immigration, climate change, and financial transactions being only the most obvious – has come global coordination of regulation. While much of this early coordination was informal, more formal systems of global regulation are increasingly emerging in arenas outside the state. Whether global or national, however, much regulation is constituted in the ordinary decisions of case-level officials. The meaning of those decisions requires attention to the organizational context in which they are made, not just the broader choices states have made in statutes or international agreements. In the last 10–15 years, regulation has become more important in developing countries, at global and transnational levels, and in nonstate settings.

Emergence and Justification

Bureaucratic regulation of industries emerged in the nineteenth century with the rise of industrial capitalism and the bureaucratic division of labor. Britain established a factory inspectorate for the textile industry as early as 1832. The use of inspectorates to address the more dangerous conditions at work – in match factories, in tenement labor – expanded dramatically throughout the Western world in the 1890s (Rodgers, 1998).

Regulation has proven to be at least partly reversible: central trends in capitalist economies during the 1970s and 1980s included both privatization of state-owned industries (see e.g., Graham and Prosser, 1988; Pistula and Rasmussen, 1990) and political pressure toward deregulating the already regulated industries (Derthick and Quirk, 1985; Harris and Milkis, 1996).

Economists have thought it important to explain why regulation is necessary within a capitalist economy, given that neoclassical economic theory holds that social purposes are best accomplished by markets. The importance of neoclassical theory has meant that advocates and critics of regulation have also expended effort understanding the justifications (see e.g., Breyer, 1982). Justification has sometimes centered on the concept of 'externalities.' A market requires that business pays for some components, for example, steel in manufacturing a car. However, the manufacturer does not have to buy the air that it pollutes in the process of making cars, and in turn, neither do consumers who drive the cars. In the United States in the 1960s and 1970s in particular, concern about the inability of companies to take into account the social costs of doing business was used to justify regulating products and processes in manufacturing.

In addition to regulating products and processes, governments regulate price and entry: the price that an industry may charge and the requirements to enter into the business. The justification of price and entry regulation does not rely upon the concept of externalities. Rather, the argument is that some industries take on the characteristics of monopoly, whether as a natural monopoly or because there are very high barriers to entry. For example, it could be inefficient to have multiple railways competing against each other since the costs of laying railroad track are very high. What is more, because it is expensive to build a railway, it is extremely difficult to enter the business, which means that the competition that is supposed to ensure that prices and services are responsive to consumer demand does not exist. Without competition, railroads could charge what they would for services. Thus railroads, the first industry subject to national bureaucratic regulation in the United States, would charge more for short-haul than for

long-haul trips because there was much less competition in short haul (Skowronek, 1982). Economists also note the importance of 'network effects' in monopoly. If the exchange of information is facilitated by having compatible computer systems, each person buying a computer system has an incentive to buy one compatible with those people already have, whether or not that is the best system. Advocates assert that professions like medicine and law require regulation because expertise is required to assess whether the profession is being practiced correctly. Consumers cannot assess whether a surgeon knows what he or she is doing until it might be too late.

These justifications for regulation have been subject to serious critique from those who believe that markets should be unregulated and from those who believe that regulation is ineffective because industries have special access to the state not shared by their workers or customers (Derthick and Quirk, 1985). In this view, regulatory bureaucracies, designed to incorporate public concerns into capitalist economies, become 'captured' by the industries they are to regulate. The regulatory bureaucracies then act largely to insulate the industries from competition and protect their interests. Sometimes that has happened by design: railroad regulators meant to maintain and steady the railroad industry. In the late nineteenth century, railroads in the United States wanted to be regulated to moderate the incentives to cut prices that were putting railroads out of business (Skowronek, 1982). In France, *notaires* welcome regulation to preserve their monopoly on drawing up legal contracts; opening that monopoly would have meant decreasing income as a result of competition (Suleiman, 1987).

Economists call the effort to gain protection from the market by legislative or bureaucratic intervention 'rent-seeking behavior' – that is, an effort to gain a return beyond what is required to keep one in the business – and many have argued that rent seeking is the heart of regulation. For example, lawyers benefit from the licensing of lawyers; it restricts the numbers of lawyers available and thereby maintains a higher price than would be the case if anyone could practice law. The claim that regulation protected inefficient industries from competition and used state power to keep prices artificially high provided a substantial part of the support for deregulation in the United States in the late 1970s and 1980s (Derthick and Quirk, 1985; Harris and Milkis, 1996).

While the United States developed regulatory institutions in a context of mainly private ownership of industries, in other nations, critiques of state ownership of industries led to privatization, which in turn entailed regulation (Graham and Prosser, 1988). During the 1980s and 1990s in Europe and the 1990s and 2000s in the developing world (Jordana and Levi-Faur, 2006; Minogue and Carino, 2006; Dubash and Morgan, 2013), regulatory agencies have proliferated. Their emergence and justification have been linked to purposes of both promoting economic efficiency (especially in essential services in both Europe and the developing world) and protection from risk or harm (especially in environmental or health and safety regulation in Europe). The importance of protective regulation can be seen by noting that the strength of the proenvironmental regulation lobby meant that environmental deregulation did not go far even in the United States (Harris and Milkis, 1996). Those advocating a more market-based approach in pollution rights gained recognition in Clean Air Act Amendments, and more recently in emissions trading regulatory schemes, but government officials remain deeply involved in regulation (Harris and Milkis, 1996; Ureta, 2013). Social solidarity and human rights have also provided justifications for the emergence of many regulatory regimes in recent years (Prosser, 2010).

Institutions that Regulate

Regulatory institutions vary considerably and all have different styles. Courts, legislatures, local governments, marketing boards, bureaucratic agencies, and (more indirectly) interest groups all can impose, and have imposed, orders on capitalist enterprise. In the United States before the Civil War, rather than relying on negotiations between business and bureaucrats, or on rules from bureaucrats that were enforced in the courts, the United States relied on courts – largely state courts – to enforce statutes or common law duties. By the late nineteenth century, most US states as well as the United Kingdom had instituted some bewildering array of boards, commissions, and special courts to ensure that a broader public interest was taken into account.

The first US federal regulatory agency, the Interstate Commerce Commission, was created in 1887. It was unusual in that it established an agency independent of immediate control by ministers, responsible for making a framework of policy through trial-type procedures, with decisions that could be appealed to general jurisdiction courts (Baldwin, 1985: p. 7). The United States' official regulatory procedure is more rule oriented than those in other Western industrialized states, and courts are more deeply involved in making policy. The American system also allows for much more public participation and opportunity to argue against regulations both within an administrative agency and through courts after administrative agencies have decided, though recent shifts in both Europe (Cassese, 2005) and the developing world (Thiruvengadam and Joshi, 2012; Uruena, 2012) are narrowing this difference.

After World War II, regulation in the United States became open to wider public participation. Administrative agencies must hold public hearings after announcing the topic of a rule they are considering. They must then justify the rule they issue according to the evidence they have. Often an organization concerned with the rule will bring the regulatory agency to court. Federal courts usually review the processes by which the rules were promulgated and decide whether the rule or decision is justified according to the evidence. What lawyers have won in the United States has been observed sometimes with horror, sometimes with envy in other nations.

Other nations do not rely so extensively on the equivalent of generalist courts to assess the procedures by which rules are enacted. The French have a specialized court that hears challenges to administrative decision making, the *Conseil d'Etat*, or Council of State. The *Conseil d'Etat* is staffed with people who have built careers within administration; the purpose is to have a court that will take into account public purposes, or the interests of the state, in deciding cases brought by individual challengers. In Britain, practitioners in regulation were frustrated enough with how their general jurisdiction courts supervised administrative decisions that the state created a version of a specialized administrative court in the 1970s (Sterett, 1997). Particular judges specialize in hearing

complaints about administrators. The political justification for both courts is that specialization in administrative complaints will lead to sensitivity to public purpose and an awareness of what can go wrong in administration. Germany, also, relies on a specialized administrative court staffed by a trained civil service judiciary.

In most countries, there are specialized bodies between frontline administrators and either generalist courts or upper level specialized administrative courts. These specialized bodies usually include representatives of relevant sectors (e.g., employers and union representatives in labor tribunals) in their staff. Members of specialized bodies are not always legally trained. Their procedures are often less formal than those of generalist courts, and those before them do not always have legal representation. The informality and discretion allowed most tribunals that provides substantial opportunity for a slip between formal statement of government policy and its meaning in particular cases.

Interest groups are also an important source of challenges to the general rules under which administrators work. Such challenges are probably more numerous in the United States than in any other nation (Kagan, 2000). However, social welfare groups in Western Europe have in recent years organized to challenge rules across diverse policy areas, including immigration and environmental regulation (Keleman, 2011). The expansion of standards via European Union directives and European Court of Justice decisions has made it more possible to challenge domestic regulation, or refusals to regulate, in Europe. And in the developing world, interest groups have begun to challenge domestic regulation in health, essential services, education, and water by invoking social and economic human rights commitments (Gauri and Brinks, 2010).

Finally, international and transnational agreements and standards have become increasingly important avenues that shape domestic regulation. For example, international treaties sometimes affect regulation in the same sector (e.g., in ozone regulation (Braithwaite and Drahos, 2000; Canan and Reichman, 1993)) and sometimes in different policy areas (for example bilateral treaties on international investment standards shape domestic environmental regulation (Szablowski, 2007)). Moreover, interest groups are increasingly developing detailed regulatory regimes of transnational standards outside (or in parallel with) states, (e.g., in forestry, finance, and labor protection) (Cashore et al., 2004; O'Rourke, 2003; Shaffer, 2012).

What Does It Cost?

Line items in government budgets for regulatory agencies do not actually state what regulation costs. To comply with regulatory requirements, businesses incur the costs of improving safety or at least documenting that they are trying to comply. Regulation involves trade-offs between effectiveness and other considerations. For example, we might want nursing homes to provide good care for people, but that is hard to measure. To the extent that we try to enforce it, we might direct effort in an organization toward documenting practices that do not quite get at caring, such as providing meals on time (Braithwaite, 1993). A more substantive mandate to care might better capture what we want, but it would be unenforceable.

In addition, different substantive regulations are acceptable to businesses and broader publics in different countries. For example, commitments to nondiscrimination in the United States, which critics see as more extensive than those in most other countries, impose costs through requirements of documentation of fair procedures. Such provisions do not require any particular substantive outcome. Other countries, however, employ approaches that would be unacceptable in the United States. For example, Sweden tried to implement employment equality in a way that could keep administrative costs low while resulting in real change, by setting employment quotas for women in jobs in which they were underrepresented (Brown and Wilcher, 1987). Britain, on the other hand, has relied on loose and open-ended definitions of nondiscrimination implemented through tribunals whose members are not legally trained and whose primary jurisdiction is not nondiscrimination law. The result seems to be a relatively inexpensive system that has done little to challenge discriminatory practices in the workplace (Kenney, 1992).

Business representatives have argued that the processes of American regulation are more resource intensive than those in other Western states: compliance requires more documentation, inspectors more readily impose fines, businesses can expect more public interest group confrontation, hearings, and possibly court battles (Kagan, 2000). However, concerns about the cost of formal regulation are now pervasive, and bureaucrats in both developed and developing countries are now often required to carry out regulatory impact analysis (and sometimes gain the permission of specialized bodies that enforce net benefit standards) before regulatory laws can be approved (Hahn and Muething, 2003; Raedelli and DeFrancesco, 2007; Morgan, 2003).

Despite concerns about the expense of regulation, enterprises sometimes choose to pursue the goals of regulation beyond those governments require, whether because compliance offices sometimes employ people who take the goals of regulation to be their mission, or because large corporations concern themselves with their public image as good citizens (Edelman, 1990; Gunningham et al., 2003). Compliance is not always a matter of direct orders.

Explaining Regulatory Styles: Political Institutions and Political Culture

Different approaches to regulation tend to fall into two very different styles: contentious with an emphasis on public procedures and rules, as against consensual without formal rules. A particularly important set of explanations for such differences stresses the historical emergence of national governance. A historical institutional approach holds that political institutions have shaped the emergence of different regulatory regimes. Those political institutions are in turn the result of large-scale historical processes. Two facets of this history are especially important: how insulated are regulatory institutions from electoral pressures, and how fragmented are political institutions in general?

In much of Western Europe, a career civil service insulated from politics was created in the nineteenth century (Kelman, 1981). Administrative arms of government often started as the personal servants of the monarch; the political environment within which they worked did not include mass electorates. Governments could develop bureaucracies insulated from political pressure, setting in place systems difficult to disrupt, no matter what cultural preferences there might be for a broader or more accessible or more aggressive system of regulation.

The insulation from the electoral politics that so profoundly shapes the civil services includes the judiciary. In civil law systems, judging is a career to which people are trained. The highest level administrative court in France, the *Conseil d'Etat*, is staffed with top level civil servants. Their charge is quite clearly to attend to complaints against administrators with an eye to maintaining the effectiveness of a regulatory regime; it is not a system designed to allow confrontation over the fundamentals of public policy. Even within British-style common law systems, judges in the general legal system are selected from the elite of the bar, who are narrowly trained and disinclined to confront administrators. In developing countries, the relatively weak institutional capacities of the civil service possibly create more contentious involvement of the courts in regulatory regimes (Uruena, 2012), and the relative lack of civil service insulation from politics certainly shapes regulatory politics in the global South, moving it along a spectrum between rules and deals (Dubash and Morgan, 2013).

Fragmentation of political institutions also contributes to contentious regulatory styles. In the United States, for example, the legislatures – Congress and the states – are not wholly within the control of the executive. Those who wish to insist on visions of regulation different from that of the executive can gain leverage in the legislature to do so, not least through mobilizing popular mistrust of closed procedures. The American system of formal rulemaking procedures and some openness in administration resulted first from the American Bar Association's mistrust of regulation and then from the politically distinct mistrust emergent in the 1960s and 1970s. Both gained leverage through Congress and through the courts, which had more generous standards for allowing people to sue than did courts in other nations.

By contrast, parliamentary systems, in which the government controls the legislative agenda, provide the government substantial control over how a regulatory bureaucracy will be organized and the forms accountability will take. Should a government believe it is best to include only responsible seeming or moderate interest groups in their discussion of regulation, or to limit checks such as a right for interest groups to sue administrators, they will find little institutionalized opposition (Vogel, 1986). From the point of view of government ministers, both in charge of government departments and sitting in parliament, such a capability only makes regulation more difficult. Blocking such capabilities leads to more consensual styles of regulation, for example, in the United Kingdom (Vogel, 1986). As with insulation from electoral pressures, political fragmentation has also been important in developing countries: for example, in understanding affirmative action policy in India (Parikh, 1997).

Some accounts of different regulatory styles focus more on cultural predilections (such as individualism or faith in authority) than on the institutional sources of shared meanings. From this perspective, US culture is individualist and assertive, while cultures in other Western countries are more deferential, in part based on a monarchical past and strong social hierarchies (Vogel, 1986). However, there are many ways in which the cultural explanation is incomplete. Deference, for example, hardly characterizes the upheavals of the 1960s and 1970s, pervasive throughout Western industrialized states. Cultural preferences may manifest in different types of legal change: in the United States, formal participation in rulemaking was institutionalized while other nation-states created state institutions such as ombudsmen to listen to complaints about administration in particular cases (Seneviratne, 2002). Cultural preferences are rarely the same across different areas of public life: in Britain, for example, a country cited as having informal regulatory practices in comparison to the United States (Vogel, 1986), land use planning has been contentious, rule bound, and full of litigation. A cultural explanation arguing that the British are more deferential than Americans cannot account for that contentiousness. An alternative institutional and historical explanation is relatively simple. The legal profession long made most of its income in land. By the time bureaucratic land use regulation emerged, the government had to take into account the interests of an already organized legal profession accustomed to litigating land disputes.

Historical institutional explanations are also better able to account for the ways in which supranational institutions harmonize regulatory requirements cross nationally. For example, within the European Union, the European Court of Human Rights and the European Court of Justice make decisions concerning the rules that nation-states use in areas as diverse as admitting immigrants, protecting the environment, providing health services, limiting sex discrimination, and determining the right to a family life or the right to travel. A perspective that takes nation-states as monolithic entities would see the supranational courts as an imposition on national autonomy. Since, however, not everyone within a nation-state agrees – some political activists want stricter regulation of pollution, or more openness to the families of settled immigrants – the introduction of supranational institutions makes for more complex alliances. State officials themselves are often not united, so supranational decisions or networks can strengthen the hand of one administrative office over another: this is particularly common in relation to free trade and competition standards (Maher, 2002; Shaffer, 2012).

Constituting Regulation: Rules and Discretion

No matter which institutions regulate, how many specific rules are made or how much regulatory contention emerges, discretion is an unavoidable part of all regulatory regimes.

National legislatures and executives may state general rules for regulation: that sex discrimination in employment is illegal, that businesses must reduce air pollution. Regulation is a complex task in which goals are often in tension – for example, to clean the air while ensuring that businesses that pollute do not have to close, or without severely curtailing

driving. Multiple and conflicting goals for complex organizations mean that what counts as compliance requires interpretation. Even in the United States, where the courts are in the business of continually clarifying the rules, the facts of a situation can be complicated enough that compliance with a rule is something that will entail negotiation between the regulator and regulated, which a court then only ratifies.

There are multiple levels in any regulatory setting where the exercise of discretion shapes the outcomes. Changing the ordinary practices of an organization is difficult. Where there are applicable rules or direct court orders against an organization, the organization must construe what it means to comply. Organizations have every reason to interpret decisions so that they fit with organizational priorities. For example, a court order to pay women equally with men in comparable jobs still leaves open what count as equivalent jobs, how equal the pay must be, and over how long a period it must be equalized, all of which will be up to an employer to determine (McCann, 1994).

The professional training of decision makers thus shapes what regulations mean. In employment discrimination law in the United States, personnel administrators interpret the law so as to require what they see as good management practices, which requires that people who work together get along. They do not often interpret the law as a matter of legal right, that is, as an enforceable claim that someone can make against his or her employer. Rather, personnel administrators are more likely to interpret problems as matters of personality conflicts (Edelman, 1993). Thus the presence of enforceable rights does not necessarily make regulation very different from, for example, enforcement of water pollution laws in Britain, where there are no clear rules and enforcement officials negotiate and bluff (Hawkins, 1984; Grabosky and Braithwaite, 1986). Even where tribunals or courts play an important role in overseeing the exercise of discretion, case-level administrators often interpret their instructions in light of what regulated businesses ordinarily do, and tribunals that oversee this often support it.

In recognition of the pervasiveness of discretion, scholarship increasingly focuses on both business organizations and administrative officials working in complex environments in which they mutually shape meanings of the law (Edelman, 1993; Harris and Milkis, 1996). Widespread shifts to 'process-oriented regulation' recognize this, and the effect is to place increased prominence on the uncertainty of human relationships within a regulatory environment, especially on the role of trust (Gilad, 2011). The more explicit recognition of the pervasiveness of discretion in regulatory design may also stem from the greater formal involvement of nonstate interest groups, and (perhaps relatedly), the increasing presence of transnational regulatory institutions that was noted above. This context is shaping regulation so as to require analysts to attend to reasons for convergence as much as differences between nation-states. The global context of regulatory implementation might lead to a two-level global economy – relative convergence among Western states (Braithwaite and Drahos, 2000), while developing countries produce goods in conditions much worse for workers. In response to the latter, the growth of certification as a nonstate regulatory regime has been a popular middle road in political terms, aiming to contain discretion that 'leaks' beyond traditional national regulatory regimes (Bartley, 2011). As a mechanism of social regulation, certification has a dual appeal both to companies who prefer private self-regulation and to social movements seeking to build new political institutions at the transnational level. Strong changes in on-the-ground practices as a result of such regimes, however, remain difficult to prove.

See also: Administrative Law: The United States and Beyond; Bureaucratization and Bureaucracy, History of; Courts and Adjudication; Employment and Labor, Regulation of; Environmental Policy: Protection and Regulation; Globalization: Legal Aspects; Occupational Health and Safety, Regulation of; Organizations and the Law; Regulation, Economic Theory of; Regulation, Empirical Analysis of; Regulatory Agencies.

Bibliography

Baldwin, R., 1985. Regulating the Airlines: Administrative Justice and Agency Discretion. Clarendon Press, Oxford, UK.

Bartley, T., 2011. Certification as a mode of social regulation. In: Levi-Faur, D. (Ed.), Handbook on the Politics of Regulation. Edward Elgar Publishing, Cheltenham, pp. 441–452.

Boyle, E., 1998. Political frames and legal activity: the case of nuclear power in four countries. Law and Society Review 32 (1), 141–174.

Braithwaite, J., 1993. The nursing home industry. In: Tonry, M., Reiss, A.J. (Eds.), Beyond the Law: Crime in Complex Organizations, Crime and Justice, vol. 18, pp. 11–54.

Braithwaite, J., Drahos, P., 2000. Global Business Regulation. Cambridge University Press, Cambridge, UK.

Breyer, S., 1982. Regulation and its Reform. Harvard University Press, Cambridge, MA.

Brown, C., Wilcher, S.J., 1987. Sex-based employment quotas in Sweden. In: Brown, C., Pechman, J.A. (Eds.), Gender in the Workplace. Brookings Institution, Washington, DC.

Canan, P., Reichman, N., 1993. Ozone partnerships, the construction of regulatory communities, and the future of global regulatory power. Law and Policy 15 (1), 61–74.

Cashore, B., Auld, G., Newsom, D., 2004. Governing through Markets: Forest Certification and the Emergence of Non-State Authority. Yale University Press, New Haven, CT.

Cassese, S., 2005. Administrative law without the state? The challenge of global regulation. New York University Journal of International Law and Politics 37, 663–694.

Derthick, M., Quirk, P.J., 1985. The Politics of Deregulation. Brookings, Washington, DC.

Dubash, N.K., Morgan, B., 2013. The Rise of the Regulatory State of the South: Infrastructure and Development in Emerging Economies. Oxford University Press, Oxford.

Edelman, L.B., 1990. Legal environments on organizational governance: the expansion of due process in the American workplace. American Journal of Sociology 95 (6), 11401–11440.

Edelman, L.B., 1993. Internal dispute resolution: the transformation of civil rights in the workplace. Law and Society Review 27 (3), 497–514.

Edelman, L.B., Suchman, M.C., 1997. Legal rational myths: the new institutionalism and the law and society tradition. Law and Social Inquiry 22, 906–941.

Einhorn, R., 1991. Property Rules. University of Chicago Press, Chicago.

Gauri, V., Brinks, D., 2010. Courting Social Justice: Judicial Enforcement of Social and Economic Rights in the Developing World. Cambridge University Press, Cambridge.

Gilad, S., 2011. Process-oriented regulation: conceptualization and assessment. In: Levi-Faur, D. (Ed.), Handbook on the Politics of Regulation. Edward Elgar Publishing, Cheltenham, pp. 423–440.

Grabosky, P., Braithwaite, J., 1986. Of Manners Gentle: Enforcement Strategies of Australian Business Regulatory Agencies. Oxford University Press, Melbourne, Australia.

Graham, C., Prosser, T., 1988. Waiving the Rules. Open University Press, Philadelphia.

Gunningham, N., Kagan, R., Thornton, D., 2003. Shades of Green: Business, Environment, and Regulation. Stanford University Press, Stanford.

Hahn, R., Muething, M.B., 2003. The grand experiment in regulatory reporting. Administrative Law Review 55, 607–642.

Harris, R.A., Milkis, S.M., 1996. The Politics of Regulatory Change: A Tale of Two Agencies. Oxford University Press, New York.

Hawkins, K., 1984. Environment and Enforcement: Regulation and the Social Definition of Pollution. Clarendon Press, Oxford, UK.
Jordana, J., Levi-Faur, D., 2006. Towards a Latin American regulatory state: the diffusion of autonomous regulatory agencies across countries and sectors. International Journal of Public Administration 29 (4–6), 335–356.
Kagan, R.A., 1978. Regulatory Justice: Implementing a Wage-Price Freeze. Russell Sage Foundation, New York.
Kagan, R.A., 1997. Should Europe worry about adversarial legalism? Oxford Journal of Legal Studies 17 (2), 165.
Kagan, R.A., 2000. The consequences of adversarial legalism. In: Kagan, R.A., Axelrad, L. (Eds.), Regulatory Encounters: Multinational Corporations and American Adversarial Legalism. University of California Press, Berkeley, CA.
Keleman, D., 2011. Eurolegalism: The Transformation of Law and Regulation in the European Union. Harvard University Press, Cambridge, MA.
Kelman, S., 1981. Regulating America, Regulating Sweden: A Comparative Study of Occupational Safety and Health Policy. MIT Press, Cambridge, MA.
Kenney, S.J., 1992. For Whose Protection? Fetal Protection Policies in the United States and Great Britain. University of Michigan Press, Ann Arbor, MI.
Maher, I., 2002. Competition law in the international domain: networks as a new form of governance. Journal of Law and Society 29 (1), 111–136.
McCann, M., 1994. Rights at Work. University of Chicago Press, Chicago.
Melnick, R.S., 1983. Regulation and the Courts. Brookings Institution, Washington, DC.
Minogue, M., Carino, L., 2006. Regulatory Governance in Developing Countries. Edward Elgar Publishing Ltd, Cheltenham, UK.
Morgan, B., 2003. Social Citizenship in the Shadow of Competition: The Bureaucratic Politics of Regulatory Justification. Ashgate Press, Aldershot.
O'Rourke, D., 2003. Community-driven Regulation: Balancing Development and the Environment in Vietnam. MIT Press, Boston, MA.
Parikh, S., 1997. The Politics of Preferences. University of Michigan Press, Ann Arbor, MI.
Pistula, J.M., Rasmussen, K., 1990. Privatizing a Province: The New Right in Saskatchewan. New Star Books, Vancouver, Canada.
Prosser, T., 2010. The Regulatory Enterprise: Government, Regulation, and Legitimacy. Oxford University Press, Oxford.
Raedelli, C., DeFrancesco, F., 2007. Regulatory Quality in Europe, Concepts, Measures, and Policy Processes. Manchester University Press, Manchester.
Rees, J., 1988. Reforming the Workplace: A Study of Self-Regulation in Occupational Safety. University of Pennsylvania Press, Philadelphia.
Rodgers, D., 1998. Atlantic Crossings: Social Politics in a Progressive Age. Harvard University Press, Cambridge, MA.
Ruggie, M., 1984. The State and Working Women. Princeton University Press, Princeton, NJ.
Seneviratne, M., 2002. Ombudsmen: Public Services and Administrative Justice. Cambridge University Press, Cambridge.
Shaffer, G., 2012. Transnational Legal Ordering and State Change. Cambridge University Press, Cambridge.
Skowronek, S., 1982. Building a New American State. Cambridge University Press, New York.
Sterett, S., 1997. Creating Constitutionalism? University of Michigan Press, Ann Arbor, MI.
Suleiman, E., 1987. Private Power and Centralization in France: The Notaires and the State. Princeton University Press, Princeton, NJ.
Szablowski, D., 2007. Transnational Law and Local Struggles: Mining, Communities and the World Bank. Hart Publishing, Oxford.
Thiruvengadam, A.K., Joshi, P., 2012. Judiciaries as crucial actors in Southern regulatory systems: a case study of Indian telecom regulation. Regulation and Governance 6 (3), 327–343.
Ureta, S., 2013. 'Because in Chile [carbon] markets work!' Exploring an experimental implementation of an emissions trading scheme to deal with industrial air pollution in Santiago. Economy and Society 42 (3). http://dx.doi.org/10.1080/03085147.2012.760347.
Uruena, R., 2012. Judiciaries as crucial actors in Southern regulatory system: a case study of indian telecom regulation. Regulation and Governance 6 (3), 282–299.
Vogel, D., 1986. National Styles of Regulation: Environmental Policy in Great Britain and the United States. Cornell University Press, Ithaca, New York.

Regulation, Economic Theory of

David EM Sappington, University of Florida, Gainesville, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Regulation takes many forms in practice, including price, quality, disclosure, privacy, and externality regulation. Popular methods for regulating the prices charged by public utilities include rate of return regulation, price cap regulation, and yardstick regulation. The optimal design of regulatory policy varies with industry structure and institutions, with regulatory objectives and policy instruments, and with the prevailing information structure.

Regulation is defined as "the act of controlling or directing by a rule, principle, [or] method" (Webster, 1989). This definition encompasses many forms of control by a variety of actors. The ensuing discussion will focus on the important form of economic regulation in which the government controls the activities of producers in order to enhance the well-being of the customers they serve. Particular emphasis will be placed on price regulation of utilities, i.e., suppliers of key services such as communications, electricity, and water.

Regulating with Accurate Information

When the average cost each firm incurs in serving customers (i.e., the firm's cost per customer) declines as the firm serves more customers, industry costs will be minimized when only one producer serves all customers. For example, the cost of providing local telephone service to a community is often lower when a single telecommunications network is constructed and operated than when multiple networks are operated. Similarly, the local distribution of electricity, water, and gas is often provided at minimum cost by a single firm. But if a single profit-maximizing firm faces no competition, consumers will have no alternative source of supply, and so an unregulated sole producer may have considerable freedom to charge high prices for its services. Regulation of the prices charged by a single (monopoly) supplier can help to protect consumers while permitting an industry configuration that minimizes the costs of serving customers.

Regulators typically consider many factors when setting prices. To illustrate, the US Postal Rate Commission is required to design a simple, transparent, just, and reasonable pricing structure for postal services in which the price charged for each service reflects the value of the service and generates revenue in excess of the corresponding production cost. In doing so, the Commission must consider many factors, including "the educational, cultural, scientific, and informational value to the recipient of mail matter" (Postal Accountability and Enhancement Act, 2006: Section 3622(b)).

To assess the basic principles that underlie the design of prices for regulated utilities more generally, consider first the case of an omniscient regulator who knows both how highly consumers value the essential services produced by the regulated firm and the minimum possible costs the firm incurs in serving customers.

Linear Pricing with a Single Output

Initially, suppose the regulated utility produces only one product. Also suppose the regulator desires to set the price for this product so as to maximize the aggregate welfare of all consumers while delivering sufficient profit to the firm to guarantee it will voluntarily serve customers at the established price. Consumer welfare measures the value consumers derive from their purchase and consumption of the product. A firm's profit is the difference between the revenues it receives and the costs it incurs. Profit is a measure of producer welfare. To maximize consumer welfare while delivering at least the minimum required level of producer welfare, the regulator must set the price consumers pay for each unit of the product they purchase so as to deliver the largest possible sum of consumer and producer welfare. This welfare-maximizing price is equal to the extra (or marginal) cost the firm incurs in producing the last unit of the product it sells. This is because consumers will continue to buy more of the product as long as the extra value they receive from doing so exceeds the price they must pay for an additional unit of the product. Therefore, when the price reflects the producer's marginal cost of production, consumers will purchase additional units of the product whenever the value they derive from consumption exceeds the cost of allowing that consumption. Consequently, a regulated price that reflects the producer's marginal cost of production will ensure the level of output that maximizes the combined welfare of consumers and the firm.

However, when unit production costs decline as the firm produces more output, a price set equal to the marginal cost of producing the last unit of output will cause the firm to suffer a financial loss. For all but the last unit produced, the compensation the firm receives (i.e., the price) will be less than the incremental production cost it incurs. To ensure that the total revenue the firm receives does not fall below the total production cost it incurs, the regulator must raise the authorized price above the level of marginal cost. A price equal to the firm's average cost of production at the prevailing level of production will deliver the highest possible level of welfare to consumers while ensuring that the regulated firm's costs do not exceed its revenues.

Nonlinear Pricing

A regulator might contemplate implementing a pricing structure in which customers do not simply pay the same price for

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71034-8

each unit of the product they purchase. For example, each customer might be charged both a fixed participation fee that does not vary with the amount of the product purchased and a constant unit price for each unit purchased. Under such a pricing structure, the unit price might be set equal to the regulated supplier's marginal cost of producing the last unit of output it sells and the participation fee might be set to recover exactly the deficit the firm incurs (i.e., the difference between the total revenue it receives from sales and the total production costs it incurs) when it sells its product at the specified unit price. This pricing structure would deliver the maximum possible level of consumer welfare while ensuring nonnegative profit for the firm, provided the participation fee is not so onerous as to compel some customers to terminate all purchases of the regulated product (Wilson, 1993). In practice, large participation fees are often viewed as unduly onerous for consumers with limited resources. Consequently, particular concern for the well-being of these consumers or general concerns with equity and fairness can limit the use of participation fees and other alternatives to constant unit prices.

Linear Pricing with Multiple Outputs

Now consider the more common setting in which the regulated utility produces multiple products (for example, both rural and urban telephone service). Also suppose the regulator chooses to implement only a constant unit price for each regulated product. Since the prices charged for different products can vary, there are many different ways in which the regulator can raise prices above marginal production costs while ensuring that revenues match total operating costs exactly. To maximize the combined welfare of all consumers, the regulator will increase the price of a regulated product further above its marginal cost of production the less sensitive consumer demand is to the price of the product. In particular, the regulator will set the price of a product close to its marginal cost of production when a higher price would cause consumers to severely curtail their purchases of the product. In contrast, the regulator will raise the price of a product well above its marginal cost of production when doing so does not cause consumers to reduce their purchases of the product significantly. This pricing policy secures the greatest combined welfare for consumers by limiting the aggregate reduction in consumption that arises as prices are raised sufficiently far above marginal production costs to generate revenues that match production costs (Baumol and Bradford, 1970).

Quality Deterioration and Cream Skimming

As noted above, regulators do not always seek solely to maximize the aggregate welfare of consumers. Consequently, in practice, regulated prices are not always raised above marginal costs by greater amounts the less sensitive consumer demand is to the price. In fact, some prices (e.g., those for rural, residential telephone service) may be set below marginal production costs, while other prices (e.g., those for telephone service to urban business customers) may be set well above marginal production costs. Such prices may arise from fairness and equity concerns, for example. Pricing patterns of this sort can introduce two important problems: quality deterioration and cream skimming.

When the price of a product is set below its marginal cost of production, each extra unit of the product the firm sells reduces its profit. Consequently, the firm would secure greater profit if consumers purchased fewer units of the product. One way to induce consumers to purchase fewer units of a product is to reduce the quality of the product. Therefore, when prices are set below marginal production costs, the regulator may need to impose penalties on the firm if the level of service quality it delivers is found to be inadequate.

When prices are set above the cost for some products and below the cost for other products, competitors may find it profitable to supply only those products for which price exceeds the cost and leave the incumbent regulated firm to supply those services for which price is set below the cost. For example, unless they are prohibited from doing so, competitors may provide telecommunications service only to large businesses in urban areas. This practice is known as cream skimming. A regulator can prevent cream skimming by barring potential competitors of the incumbent regulated firm from providing services in the regulated industry. Such entry prohibitions, though, can hinder industry performance by precluding the operation of firms that have lower operating costs than the incumbent supplier and by removing the discipline that the threat of entry can impose on the incumbent firm. Furthermore, competitors are sometimes able to circumvent prohibitions on entry by developing and marketing substitute products. A regulator can limit incentives for cream skimming by setting prices that more closely reflect the incumbent firm's operating costs. When prices reflect costs in this manner, entry will typically be encouraged only when a competitor can satisfy consumers' needs at lower cost than the incumbent supplier can.

Regulation in Two-Sided Markets

Suppliers of network services sometimes serve two distinct groups of constituents. For example, Internet service providers (ISPs) serve both retail consumers and content providers (CPs). An ISP allows retail customers to access the content provided by CPs and simultaneously provides CPs with access to retail consumers. In principle, an ISP could charge both 'sides' of the market, i.e., charge retail customers for Internet access and charge individual CPs for making their content available to the ISP's retail customers. The welfare-maximizing price charged to one side of the market depends on the marginal cost of serving that side of the market and on the extent to which a price above marginal cost reduces Internet use by both retail customers and CPs (Rochet and Tirole, 2006). In practice, an ISP charges its retail customers for Internet access but typically does not charge individual CPs for allowing the ISP's retail customers to access CP content. This asymmetric pricing policy – commonly referred to as net neutrality – often is encouraged on the grounds that it promotes the widespread dissemination of ideas, facilitates the free exchange of information and opinions, and encourages innovation by CPs. However, ISPs contend that an inability to charge CPs limits the ability of ISPs to manage network congestion, reduces the scale and scope of beneficial network investment an ISP can afford to undertake, and

inhibits the ISP's ability to tailor network capabilities to the idiosyncratic needs of diverse CPs (Schuett, 2010).

Regulating with Limited Information: Incentive Regulation

When, as presumed above, the regulator is well informed about the firm's operating capabilities, the regulator can set prices to generate revenues that exactly match production costs when the firm operates efficiently (i.e., when it incurs only the minimum possible cost required to serve customers). When the regulator is not perfectly informed about the firm's capabilities, the regulator cannot simply require the firm to operate efficiently. Instead, the regulator must motivate the firm to reduce its operating costs toward their minimum possible levels. There are a variety of techniques the regulator can employ to do so. These techniques constitute elements of what is termed incentive regulation.

Regulatory Lag

When prices are set to match realized production costs only periodically (e.g., once every 3 or 4 years) rather than continually, the firm derives some financial benefit from reducing its operating costs. The longer the period of time before prices are reduced to match lower realized operating costs, the more pronounced is the firm's incentive to reduce its costs. On the other hand, the longer prices are permitted to remain above operating costs, the longer consumer welfare is held below the level that could be secured given the firm's realized costs. This trade-off determines the optimal length of time between price revisions, i.e., the optimal regulatory lag.

Price Cap Regulation

Price cap regulation has become commonplace in many regulated industries. Price cap regulation incorporates a form of regulatory lag while delegating some pricing discretion to the regulated firm. The typical price cap regulation plan specifies a price cap period (often 4 or 5 years) and a maximum rate at which the regulated firm's inflation-adjusted prices can rise annually, on average, during this period (Sappington and Weisman, 2010). The firm's realized costs and earnings are usually reviewed at the end of the specified price cap period, and often employed to update estimates of the pricing restriction that would allow the regulated firm to secure a reasonable level of earnings in the next price cap period. When the firm's realized performance is employed to set future standards, the firm's incentives for exceptional performance are dulled. However, price cap regulation provides some incentives for cost reduction by instituting an explicit regulatory lag.

By controlling the rate at which regulated prices can increase on average, price cap regulation affords the firm some discretion in setting prices for individual services. This discretion can help the regulated firm to limit undesirable cream skimming by realigning prices to more closely match production costs. However, this discretion can also enable the firm to raise certain prices that the regulator prefers to remain low (e.g., the price of local telephone service to rural residential customers). To limit the firm's ability to implement undesired pricing structures, price cap regulation plans often prohibit the firm from raising significantly the prices of certain specified products.

Earnings Sharing Regulation

By providing incentives to reduce operating costs, price cap regulation can simultaneously provide incentives to reduce service quality, since superior quality is often more costly to provide. By allowing prices to diverge from realized costs for considerable periods of time, price cap regulation also can admit earnings that depart significantly from expected levels. To mitigate these potential problems, an earnings sharing plan is sometimes appended to price cap regulation. An earnings sharing plan requires the firm to share with its customers a fraction of realized earnings above (and often below) specified thresholds. Such sharing helps to insure the firm and the regulator against extremely high (and low) earnings, but at the expense of reduced incentives for exceptional performance (Lyon, 1996).

Yardstick Regulation

In certain settings, yardstick regulation can be employed to provide strong incentives for superior performance while limiting the risk of extreme earnings. Under yardstick regulation, the standards to which one firm is held are based upon the performance levels achieved by other firms. To illustrate, in the British water industry, the prices and service quality standards that are imposed on the monopoly provider in one jurisdiction are linked to the costs and service quality levels achieved by monopoly producers in other geographic regions (Abbott and Cohen, 2009).

When firms face similar operating circumstances, yardstick regulation can enable the regulator to impose a reasonable but challenging performance standard on the regulated firm without basing this standard on the firm's own historic performance. Consequently, yardstick regulation can ensure reasonable earnings for the regulated firm while providing the firm with strong incentives for superior performance. However, these desirable outcomes can only be secured via yardstick regulation if the firms are known to face similar operating circumstances or if all relevant differences in their operating circumstances are well understood. In the absence of this knowledge, the standards imposed under yardstick regulation can result in extreme earnings, much like under price cap regulation. Yardstick regulation may also encourage collusion among the firms that are implicitly being compared. By agreeing to reduce their performance below achievable levels, the firms can effectively relax the constraints imposed by yardstick regulation (Tangeras, 2002).

Franchise Bidding

Even when there is a single regulatory jurisdiction and a single producer in the industry, a regulator may still be able to harness the discipline of competition by structuring competition for the industry. The regulator may allow potential producers to bid for the right to serve as the sole producer in the industry. When

the potential producers have similar abilities and similar information about operating circumstances, intense bidding competition for the market can secure for consumers the low prices and high service quality that competition among producers in the market might otherwise provide. However, successful franchise bidding requires that all relevant performance requirements be specified completely in advance, and that the regulator and firm both adhere precisely to the terms of the franchise contract once a producer has been selected. If the regulated firm can renege on its responsibilities after it is installed as the sole producer, or if the regulator can demand more services from the selected producer than it originally agreed to provide, franchise bidding can fail to deliver the maximum possible attainable level of consumer welfare (Klemperer, 2002).

Regulatory Options

When the regulated firm has better information than the regulator about its operating capabilities, consumers are often best served when the regulator allows the firm a choice among regulatory plans. A careful structuring of the options afforded the firm can induce it to employ its superior information to choose a plan that best serves consumers (Laffont and Tirole, 1993; Armstrong and Sappington, 2007).

To illustrate this point, suppose a regulator decides to impose price cap regulation on a firm rather than allow the firm to choose between price cap regulation and cost-plus regulation, whereby prices are reset frequently to match realized production costs. To limit the likelihood of financial distress, the regulator may not insist that prices decline rapidly. In contrast, when the firm has the option to choose cost-plus regulation, the regulator can impose more stringent limits on price increases under price cap regulation without the fear of causing financial distress. If the firm believes it will be unable to realize adequate earnings under the proposed price cap regime, it will choose to operate under cost-plus regulation instead. But if the regulator has not overestimated the firm's capabilities in setting the challenging requirements of the price cap regime, the firm will choose this regime and deliver its pronounced benefits to consumers.

Regulation in Vertically Related Industries

Regulated firms sometimes supply both inputs and final products. Inputs are products that firms employ to make other products. Final products are those that customers consume, and that are not used to make other products. For example, an incumbent supplier of telecommunications services may both deliver key inputs (e.g., network access and message transport) to rival suppliers and sell telephone service to retail customers. Additional considerations arise in such settings.

Limiting the Disadvantaging of Retail Rivals

A vertically integrated supplier (VIS) may exploit its position as the supplier of key inputs to rivals to hinder their operations. For example, consider a setting where the VIS is a relatively efficient supplier of the retail product, and so would secure the highest level of profit if it were the sole supplier of the retail product. In such settings, the VIS will be inclined to charge a very high price for the inputs it supplies to retail rivals (and/or severely limit the quality of the inputs) in order to induce the rivals to terminate their market activities. Regulation of input prices and/or input qualities that limits such exclusionary behavior can promote competition among retail suppliers in such settings and thereby benefit retail customers.

Although a VIS often will have an incentive to disadvantage retail rivals, this is not always the case for at least two reasons. First, if a VIS derives substantial profit from selling its inputs to retail rivals, activities that limit the sales of retail rivals can reduce their purchases of inputs from the VIS and thereby diminish the profit of the VIS (Sibley and Weisman, 1998). Consequently, regulated input prices above marginal production costs can limit the incentive of a VIS to hinder the operations of retail rivals (Beard et al., 2001). Second, activities that reduce the demand for rivals' retail products can induce the rivals to lower the prices they charge for their products. Such price reductions reduce the retail profit of the VIS, and so the VIS may refrain from activities that reduce the demand for rivals' retail products (Mandy and Sappington, 2007).

Regulating Input Prices

In addition to limiting the incentive of a VIS to hinder the operations of retail rivals, an input price above its marginal cost of production can help to limit cream skimming and ensure that downstream production is carried out by the least-cost suppliers. To illustrate this point, suppose that if the input supplier operates downstream, the prices it charges there are regulated. Further suppose that the price of one downstream product (called product A) is set at a level (p^A) that exceeds the supplier's marginal cost of production (perhaps to facilitate a price below marginal cost for another product). If the price of the essential input is set equal to its marginal cost of production in this setting, a downstream producer may be able to serve all consumer demand for product A profitably, even though its marginal cost of production ($\tilde{c}^A$) exceeds the corresponding cost (c^A) of the input supplier. The higher cost producer can set a price below the regulated price for product A and still secure positive profit if its cost disadvantage ($\tilde{c}^A - c^A$) is smaller than the regulated profit margin on product A ($p^A - c^A$). To preclude operation by higher cost competitors, the price of the essential input can be raised above its marginal cost of production by the amount of profit ($p^A - c^A$) the input supplier foregoes when it provides the input to competitors (Armstrong, 2002).

Regulating Market Structure

The vertical divestiture of a VIS can reduce its incentive to disadvantage retail operators. When the supplier of a key input has no direct financial interest in any retail supplier, the input supplier may find it most profitable to enhance, rather than hinder, the capabilities of all retail suppliers. In many settings, though, the complete separation of the wholesale and retail operations of a formerly integrated firm can increase the firm's production costs, complicate its planning activities, and disrupt industry operations. Therefore, regulators in some industries

(e.g., the telecommunications industry) often view the vertical divestiture of an industry supplier as a last resort (Ofcom, 2005).

Vertical divestiture has been promoted in the electricity industry to help ensure that transmission companies do not provide network access on differentially favorable terms and conditions to affiliated generators. Such discrimination has the potential to raise the prices customers pay for electricity and to discourage investment in generation capacity by independent electricity generators (Hogan, 2002; Joskow, 2005).

Additional Roles of Regulation

Although regulation often focuses on the prices that utilities charge for their services, regulation can assume other forms.

Service Quality Regulation

As noted above, input quality regulation can help to limit the ability of a VIS to disadvantage retail rivals by supplying them with low-quality inputs. Regulation of the quality of retail services also can benefit consumers. For example, when consumers lack the expertise required to evaluate the benefits associated with different levels of service quality (e.g., product safety features), the imposition of service quality standards can help to ensure producers deliver the quality levels that best serve consumers' needs (Viscusi, 2007).

Disclosure Regulation

In settings where consumers understand the benefits of different levels of service quality but have difficulty observing the quality of a product before it is purchased, consumers can benefit from regulation that mandates information disclosure. Disclosure regulation can require producers to inform consumers about key elements of product quality (e.g., the risks associated with certain financial or pharmaceutical products) and thereby help consumers choose the products and services that best serve their needs.

Privacy Regulation

Although consumers can benefit from the mandated disclosure of certain types of information, they may object to the disclosure of other types of information. Consumers reveal information about their preferences and may provide additional personal information (e.g., age, gender, marital status, and residence location) when they engage in commercial transactions (e.g., purchasing a product or applying for credit). Conceivably, advertisers might purchase personal customer information from merchants in order to be better able to tailor advertising activities to the tastes and interests of individual customers. Although such tailoring might both increase the profits of sellers and reduce the amount of irrelevant advertisements consumers receive, some consumers may object to the lack of privacy implied by the sharing of their personal information. Regulation to limit such sharing that is not explicitly authorized by consumers therefore has the potential to enhance consumer welfare in some settings (Evans, 2009).

Regulation of Externalities

In certain instances, production activities can impose negative externalities on society. A negative externality generated by a production activity is a cost the activity imposes on individuals not directly involved in the activity. For example, some production processes may expel harmful pollutants into the surrounding land, water, or air. If no restrictions are placed on producers' rights to expel pollutants into the environment, producers may produce more than the surplus-maximizing level of pollution. This is the level of pollution at which the social damage that would be caused by an additional unit of pollution is exactly equal to the social cost that must be incurred to eliminate that additional unit of pollution. Pollution in excess of the surplus-maximizing level imposes social damage that exceeds the cost of eliminating the extra pollution. In contrast, the cost of reducing pollution below the surplus-maximizing level exceeds the social benefits derived from the reduction in pollution.

Various regulatory policies can be employed to control environmental pollution and other types of negative externalities. For instance, direct taxes (effluent fees) can be imposed on pollution when it is readily observed and measured. By setting the tax on pollution equal to the marginal social damage caused by the pollution, the producer can be induced to realize the surplus-maximizing level of pollution. When the level of pollution produced by a firm is difficult to measure, taxes can be imposed instead on the products whose production causes pollution, or on inputs (such as gasoline) to the activity (automobile driving) that causes pollution. The use of specific pollution abatement technologies can also be mandated.

Some governments have introduced tradable pollution permits to limit pollution. Each permit entitles its bearer to emit one unit of pollution (e.g., 1 ton of sulfur dioxide). Polluting firms are able to buy and sell permits at prices determined by the available supply of and demand for these permits. A key potential benefit of tradable permits is that they promote pollution abatement at minimum cost. Producers that are able to reduce the pollutants they expel at a cost below the prevailing price of a permit will find it profitable to do so and sell any permits they own. In contrast, producers for whom pollution abatement is more costly than the price of a permit will continue to expel pollutants but will pay for the right to do so (Burtraw et al., 2005).

Sector Regulation

While this discussion has focused on price regulation of utilities, other forms of regulation prevail in other industries. To illustrate, regulations designed to avoid systematic industry failure are common in the financial sector. The *Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010*, for instance, limits the ability of banks to invest consumer deposits insured by the US government in risky ventures. The Act also attempts to prevent financial institutions from becoming 'too big to fail', and restricts the use of public funds to rescue private institutions that do fail (Acharya et al., 2011).

In the health care industry, legislation like the *Patient Protection and Affordable Care Act of 2010* requires US citizens to purchase health insurance, provides corresponding

financial subsidies to low-income citizens, attempts to enhance competition among suppliers of health insurance, and limits the ability of these suppliers to discriminate against individuals with preexisting chronic health problems (Kaiser, 2013).

Competing for Regulatory Benefits

This discussion illustrates some of the ways in which the regulation of producer activities can be designed to benefit consumers. It is important to note, however, that regulation may not always serve this role in practice. Because regulators typically have the power to deliver benefits to some parties and impose costs on others, the affected parties will have an incentive to try to secure favorable treatment from regulators. When industry producers are better organized and have more at stake than industry consumers, regulation that benefits producers at the expense of consumers can arise (Bo, 2006). For example, new competitors may be barred from an industry even though their entry and operation would be in the best interest of consumers. These realistic possibilities merit careful consideration when evaluating the benefits and costs of existing and proposed regulations.

See also: Financial Regulation; Governance; Regulatory Agencies.

Bibliography

Abbott, M., Cohen, B., 2009. Productivity and efficiency in the water industry. Utilities Policy 17 (3–4), 233–244.

Acharya, V., Cooley, T., Richardson, M., Walter, I., 2011. Regulating Wall Street: The Dodd-frank Act and the New Architecture of Global Finance. John Wiley & Sons, Hoboken, NJ.

Armstrong, M., 2002. The theory of access pricing and interconnection. In: Cave, M., Sumit, M., Vogelsang, I. (Eds.), International Handbook of Telecommunications Economics. North-Holland Publishers, Amsterdam.

Armstrong, M., Sappington, D., 2007. Recent developments in the theory of regulation. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. 3. Elsevier Science Publishers, Amsterdam.

Baumol, W., Bradford, D., 1970. Optimal departures from marginal cost pricing. American Economic Review 60 (2), 265–283.

Beard, T., Kaserman, D., Mayo, J., 2001. Regulation, vertical integration, and sabotage. Journal of Industrial Economics 49 (3), 319–333.

Bo, E., 2006. Regulatory capture: a review. Oxford Review of Economic Policy 22 (2), 203–225.

Burtraw, D., Evans, D., Krupnick, A., Palmer, K., Toth, R., 2005. Economics of pollution trading for SO_2 and NO_X. Annual Review of Environment and Resources 30, 253–289.

Evans, D., 2009. The online advertising industry: economics, evolution, and privacy. Journal of Economic Perspectives 23 (3), 37–60.

Hogan, W., 2002. Electricity market restructuring: reforms of reforms. Journal of Regulatory Economics 21 (1), 103–132.

Joskow, P., 2005. Transmission policy in the United States. Utilities Policy 13 (2), 95–115.

Kaiser Family Foundation, 2013. Summary of the Affordable Care Act. http://kff.org/health-reform/fact-sheet/summary-of-new-health-reform-law.

Klemperer, P., 2002. What really matters in auction design. Journal of Economic Perspectives 16 (1), 169–189.

Laffont, J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge, MA.

Lyon, T., 1996. A model of sliding-scale regulation. Journal of Regulatory Economics 9 (3), 227–247.

Mandy, D., Sappington, D., 2007. Incentives for sabotage in vertically-related industries. Journal of Regulatory Economics 31 (3), 235–260.

Office of Communications (Ofcom), 2005. Final Statements on the Strategic Review of Telecommunications, and Undertakings in Lieu of a Reference under the Enterprise Act 2002.

Postal Accountability and Enhancement Act, 2006. H.R. 6407 – 109th Congress.

Rochet, J., Tirole, J., 2006. Two-sided markets: a progress report. Rand Journal of Economics 7 (3), 645–667.

Sappington, D., Weisman, D., 2010. Price cap regulation: what have we learned from twenty-five years of experience in the telecommunications industry? Journal of Regulatory Economics 38 (3), 227–257.

Schuett, F., 2010. Network neutrality: a survey of the economic literature. Review of Network Economics 9 (1).

Sibley, D., Weisman, D., 1998. Raising rivals' costs: the entry of an upstream monopolist into downstream markets. Information Economics and Policy 10 (4), 451–470.

Tangeras, T., 2002. Collusion-proof yardstick competition. Journal of Public Economics 83 (2), 231–254.

Viscusi, K., 2007. Regulation of health, safety, and environmental risks. In: Polinsky, M., Shavell, S. (Eds.), Handbook of Law and Economics, vol. 1. Elsevier Science Publishers, Amsterdam.

Webster's Encyclopedic Unabridged Dictionary of the English Language, 1989. Portland House, New York.

Wilson, R., 1993. Nonlinear Pricing. Oxford University Press, Oxford.

Regulation, Empirical Analysis of

Nancy L Rose, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 12957–12963, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Economic regulation injects the state into firms' pricing, entry, investment, and product decisions. The empirical study of economic regulation has evolved considerably since George Stigler's seminal work in 1962. An extensive literature analyzes the effects of regulation in both natural monopoly and structurally competitive industries. This work suggests that regulation can have significant effects on prices, rent creation and distribution, costs, and technological innovation. Current policy and research activity focuses on restructuring and market design in traditional natural monopoly sectors and on the shift toward market-based or incentive regulation. Myriad worldwide policy experiments provide a rich set of observations for empirical analysis. Early results suggest some unanticipated consequences, and highlight the importance of institutional detail in market design.

The economic regulation of firms and industries constitutes an extreme form of government intervention in markets. Government antitrust, competition, tax, and trade policies, as well as most environmental, health, safety, and information regulations, shape but do not replace the market, although the most severe of these policies may blur that distinction. Economic regulation stands in sharp contrast. In industries subject to economic regulation, government agencies or departments exercise considerable control over firms' entry, pricing, investment, and product choice decisions, and administrative decisions replace market outcomes. Intervention may go as far as replacing private ownership with government-owned and -operated enterprises.

The modern field of empirical regulatory economics developed during the 1960s and early 1970s, as economists began "to apply the familiar principles of microeconomics in comprehensive fashion to the subject" (Kahn, 1988: p. xv). George Stigler, whose contributions to the field were recognized in his 1982 Economics Nobel prize citation, sparked much of this growth. Through the 1980s, economists focused on quantifying the effect of regulation on firm behavior, industry outcomes, and consumer welfare. The diversity of regulatory forms and industry characteristics across the transportation, energy, utility, financial, and communication sectors provided a rich set of institutions for investigation. Significant policy reforms, even deregulation, in structurally competitive sectors during the 1970s and 1980s provided new sources of variation in the regulatory environment and invigorated empirical regulatory research. More recent attention has turned to the traditional 'public utility' sector, particularly the telecommunication and electric utility industries. Market-based institutions have begun to replace state-ownership and regulation, as technological change called into question the natural monopoly assumption in many of these industries and policy interest in privatization raised questions of how best to restructure and sell-off vertically integrated monopolies. Analyzing the operation of these, redesigned markets has become a main focus of empirical regulatory research.

This article focuses on empirical analysis of the effects and reform of economic regulation. Section Empirical Methodologies describes the dominant empirical methodologies. Section The Effects of Traditional Economic Regulation sketches the evidence on the effects of traditional economic regulation. Section The Effects of Restructuring Regulated Natural Monopolies discusses recent analyses of market design and regulatory scope. To keep the reference list manageable, readers are referred to surveys and collective volumes wherever possible. These guide readers interested in delving more deeply into particular issues to a wealth of primary studies.

Empirical Methodologies

Measuring the effect of economic regulation requires the establishment of a counterfactual benchmark: what would firm or industry performance look like in the absence of regulation? A variety of approaches to this have been used. The most compelling lesson of the existing literature is that the impact of economic regulation depends critically upon its particular institutional form and the characteristics of the industry under study. Measuring the impact of regulation requires as much attention to the details of how regulators operate in practice as to the prescribed legal form of the regulation. Paul Joskow's (1974) study of electric utility regulation in the US provides a persuasive exposition of this argument.

The most common methodology used to measure regulatory effects compares outcomes in regulated markets with those in unregulated markets. Time series variation may facilitate this comparison; the wave of US deregulation in the late 1970s and early 1980s generated numerous 'before-and-after' studies of regulation. Differences across political jurisdictions, e.g., across states within a country or across countries, may provide another source of variation in regulatory regimes. Empirical studies also may use differences in the intensity or form of regulation to infer its effects, particularly if regulation is likely to bind only under certain well-defined circumstances.

Two conditions limit the usefulness of this approach. First, unregulated comparisons may not exist. Industries that are regulated in one political jurisdiction tend to be regulated or government owned in other political jurisdictions. Regulation may be so long-lived that comparing current regulated performance with historical unregulated performance is meaningless.

Second, this method implicitly assumes that differences in regulatory regimes are exogenous to firm or market performance. This is often false. If, for example, jurisdictions impose electricity regulation in response to high or rapidly increasing local electricity rates, comparing price levels in regulated and unregulated locales might suggest no effect even if regulation does, in fact, constrain prices. Using panel data to estimate effects from 'differences in differences' econometric models may be useful in some cases, as may developing more complete models of the determinants of regulation.

A second methodology estimates a structural model of the market, forms assumptions about behavior in the absence of regulation, and compares simulated unregulated outcomes with those observed under regulation. This technique may be useful even in the absence of regulatory variation. It is most appropriate when regulatory constraints can be clearly specified. Suppose regulators constrain prices but do not influence investment or factor use decisions. Structural estimates of underlying demand and cost functions permit one to predict unregulated profit-maximizing prices and compare these with current regulated prices. If regulated firms do not minimize costs, however, or if regulatory intervention alters investment decisions in a systematic way, then the regulated cost function may be a poor estimate of the unregulated long-run cost function, and the implied price comparison will be misleading.

A third approach uses experimental data to measure regulatory effects. Field experiments, such as the time-of-use electricity pricing studies conducted in the US during the early 1970s, can reveal responses to alternative regulatory or market designs. In the best of circumstances, these can provide information on how market outcomes respond to exogenous variation in regulatory mechanisms, holding other economic conditions constant. Their cost limits their widespread use, however, and their temporary nature may preclude an assessment of long-run equilibrium responses to alternative policies. Laboratory experiments, which present volunteers (often students) with a set of objectives, operating guidelines, and payoffs, also have been used in a limited way to simulate responses to alternative regulatory designs.

The Effects of Traditional Economic Regulation

It is impossible to describe succinctly the full effects of economic regulation across industries, over time, and through space. Regulation may influence prices, costs, and the pace of technological change in an industry. This may lead to the creation or suppression of economic rents, and change the distribution of rents across various stakeholders or interests in the industry. These effects are likely to vary considerably across markets and regulatory forms, and preclude simple generalizations. This article highlights some illustrative results. Further detail is provided in the many reviews referenced at the end of this article and the primary sources they cite.

Empirical Analysis of Natural Monopoly Regulation

Industries such as electricity supply, telecommunications, natural gas transmission and distribution, water and sewer service, cable television service, and to some extent railroads have historically been considered 'natural monopolies.' This applies when total production costs are minimized by providing service to all customers in a given market through a single firm. As economists since Adam Smith have recognized, minimizing production cost may not minimize price, since monopolies use their market power to markup price above competitive levels and earn monopoly rents. This historically engendered one of two policy responses. Government ownership in natural monopoly sectors was typical outside the USA, and was used frequently in the USA for water and sewer service, but less so for electric utilities and cable television service. Government regulation of privately owned firms was the dominant US response in most other cases. In this scheme, regulators typically grant a legal, 'franchise' monopoly to a single provider in a geographic market, but retain control over price to mitigate the monopoly markups that an unconstrained legal monopolist would choose. Prices are determined generally by commissions that review costs, and often investment decisions, and set prices to cover allowable costs and a 'fair rate of return' on 'prudent' capital investments.

Early economic research on the regulated public utility sector was generally theoretical and focused on how to perfect regulatory pricing, treating as given the desirability and general effectiveness of government regulation. Optimal pricing questions also arose for state-owned utility enterprises, and French and British economists were in the vanguard of research in this area (Kahn, 1988; Laffont, 1994; *see* Poverty: Measurement and Analysis). Stigler and Friedland's (1962) study of the early electricity industry was one of the first empirical analyses of the actual effect of utility regulation, and concluded that regulation did little to reduce prices below unregulated monopoly levels. While the particulars of this research generated controversy at the time, and subsequent work suggests that electricity prices in the Friedland–Stigler sample were, in fact, lower for regulated firms (Peltzman, 1993), the notion that regulation might not improve upon market outcomes even in those sectors with the strongest public interest rationale stimulated considerable additional research.

The literature subsequent to Stigler and Friedland (1962) suggests generally that price regulation, particularly in the electricity and cable television industries, has constrained average prices below the level an unconstrained legal monopolist would choose. Analysis of average price levels masks the effect of regulation on price structures, however. Price structure analyses in electricity and telephone industries suggest redistributive rather than Ramsey-like efficient pricing objectives, often with considerable transfers from customers with relatively elastic demands to those with relatively inelastic demands. For example, regulatory cross-subsidies from long-distance telephone rates to basic residential telephone service and from urban to rural telephone customers were a major force driving new entry into the US long-distance sector, leading ultimately to the 1984 divestiture of AT&T and the regulatory reform of the telephone industry. The tenacity of these cross-subsidies more than 15 years later testifies to the strength of political interests supporting these policies. Some economists have estimated the social welfare cost of the residential service cross-subsidy to be as high as $7.0 billion per year (in Peltzman and Winston, 2000: p. 78).

The effect of natural monopoly regulation on static production costs is less well measured. One of the most developed theoretical and empirical literatures on incentives for excessive capital investment in rate-of-return regulated industries proved largely inconclusive (Averch and Johnson, 1962; Joskow and Rose, 1989). There is some evidence of higher wages in firms that expect regulators to flow increased labor costs through to higher consumer prices, though wages do not tend to be higher overall in regulated sectors (Hendricks, 1975; Peoples, 1998). In the railroad industry, low regulated profit rates led to underinvestment in track maintenance and repair on many lines, reducing efficiency and increasing average transit time over the track.

Of particular concern is the effect of regulation on the pace of technological innovation. This may occur from direct regulatory constraints on new process or product introductions, such as the Federal Communications Commission's delay in allocating radio spectrum to wireless telephone systems in the USA and its restrictions on enhanced telephone service offerings. Indirect regulatory effects may also influence the pace of technological change by altering expected payoffs to investment in new technology. While often difficult to measure, these costs can be enormous. For example, Hausman estimated lost surplus from the delayed introduction of wireless phone service in the USA at more than $30 billion annually (in Peltzman and Winston, 2000: p. 80).

The ambiguity of cost and innovation effects may arise in part from the difficulty of establishing an appropriate empirical counterfactual benchmark to regulation, an essential prerequisite to measuring either static or dynamic efficiency effects. Nevertheless, ongoing policy concerns about the efficiency of production, investment incentives, and the pace of technological change in natural monopoly sectors have fueled much of the recent movement toward market-based reforms, discussed in Section The Effects of Restructuring Regulated Natural Monopolies.

Empirical Analysis of Regulation in Multifirm Industries

A substantial portion of the regulatory apparatus built during the twentieth century involved multifirm industries, such as trucking, airlines, natural gas and oil production, property and liability insurance, and banking. Regulation in many of these was rationalized by the deleterious effects of excessive, rather than insufficient, competition, particularly during the depression of the 1930s. In others, such as airlines, regulation was advocated to promote the orderly development of an emerging commercial industry. Many of these regulatory systems were dismantled or reformed substantially between 1975 and 1985, reflecting a broad consensus on its social costs.

Multifirm price regulation usually requires all firms serving a given market to adhere to a standard tariff. In some industries, this tariff was set by the regulatory commission, as the Civil Aeronautics Board did for the US airline industry. In others, tariffs were set collectively by the firms and ratified by the regulatory commission, as the Interstate Commerce Commission did for trucking rate bureau filings. Regulation generally discourages entry or market extension. By eliminating the ability of existing firms to discount rates and the ability of new firms to enter the market, regulation limits competition and tends to raise prices significantly. For example, economists estimate that regulated prices in the late 1970s were at least 15–20% above competitive levels in the trucking industry and 25–30% above unregulated prices in the airline industry.

High prices generate significant potential rents. The value of these rents is embodied frequently in the market value of regulatory assets, such as taxicab medallions, operating certificates for regulated trucking firms, or state liquor licenses. Asset values may reflect only a portion of the rents generated by regulation, however. In some sectors, organized labor was able to garnish a substantial share of regulatory rents (Rose, 1987; Peoples, 1998). Members of the Teamsters Union, for example, were a major beneficiary of trucking regulation, with wages 50% above those for nonunion truck drivers and drivers outside the regulated industry. In other industries, regulatory rents were dissipated through various rent-seeking behaviors. This seems to have been particularly acute in the airline industry, where nonprice competition repeatedly eroded load factors and profit margins despite increasing regulated prices. In the rare situations in which regulatory policy effectively constrains maximum prices and exit from markets, rents may be negative. In the short run, this can transfer quasi-rents from owners of regulated capital to consumers of their products. Over the longer run, its adverse effect on investment tends to erode both the quantity and quality of service provided. The rash of bankruptcies in the US railroad industry during the 1970s and early 1980s, and natural gas shortages during periods of below-market ceiling prices in the 1960s and 1970s, were consequences of this type of policy.

As in the natural monopoly sectors, regulation influences the structure of prices as well as the average price level. Economic regulation often leads to a more uniform price structure than in unregulated markets. Prices tend to be less responsive to cost differences across consumers or over time, due in part to cross-subsidies for favored higher-cost consumer groups and to aversion to congestion-based or peak-load pricing. This distorts both capacity utilization and firms' investment incentives. These effects have been analyzed in the context of traffic mix distortions between railroads and trucking firms, provision of air service to small communities, and excessive pooling in insurance markets.

Postderegulation experience suggests that regulation often suppressed variation due to price discrimination, relative to the unregulated market. Airlines are perhaps the best example of this. It is important to recognize that price discrimination, while controversial, may enhance overall welfare relative to a uniform price structure; see the literature on Ramsey–Boiteaux pricing (Kahn, 1988). In airline markets, for example, the pervasive use of low discount fares to fill seats that would otherwise depart vacant increases substantially the population for whom air travel is viable, and raises airline revenue per available seat-mile. The net welfare effect of current levels of price discrimination is difficult to measure, and has not been directly calculated for any deregulated market.

Economic regulation in multifirm industries has been associated with both direct and indirect cost effects. Regulatory restrictions and rigidities may prevent firms from optimizing production in both static and dynamic contexts. For example, entry restrictions in the trucking industry led to high rates of partial or empty backhauls and circuitous routings, raising the

cost of transporting a given commodity flow. Unit banking regulations, which restrict a firm to a single physical location in one state, both limit realizations of economies of scale in provision of service and insulate less efficient banks from takeovers by more efficient competitors. Inefficiencies, such as these generate potentially huge regulatory deadweight losses (Winston, 1998).

The welfare effects of cost increases may be difficult to quantify when associated with offsetting changes in the product provided to consumers. For example, the suppression of price competition may encourage nonprice competition, increasing both costs and quality relative to outcomes in unregulated markets. In the regulated airline industry, increased flight frequency and rapid adoption of new aircraft increased both cost and value to passengers. Measuring cost or productivity effects without controlling for the value of higher service quality leads one to overstate the adverse welfare effects of regulation in situations like this. Recent advances in structural demand estimation may help to quantify the value of particular quality attributes, though both data and modeling limitations continue to make this one of the weaker empirical links in most assessments of the overall effects of regulation.

The Effects of Restructuring Regulated Natural Monopolies

The focus of regulatory policy and research changed sharply in the late 1980s and 1990s, with the ascendancy of market-based institutions over state-ownership and regulation in even traditional 'natural monopoly' sectors. Restructuring, requiring firms to divest along vertical segment lines, has become common. Segments viewed as structurally competitive, such as electricity generation, have been increasingly deregulated. Network segments that continue to operate under natural monopoly cost conditions, such as electricity transmission grids or distribution systems, generally are required to provide access to competitive suppliers, often at terms subject to continued regulatory oversight. Where economic regulation has been retained, it has become more market-oriented. This reflects both dissatisfaction with recent performance under regulation and a response to theoretical research emphasizing attractive properties of incentive-based regulatory schemes (*see* Poverty: Measurement and Analysis).

These changes raise new questions about market design and regulatory interfaces. The emerging empirical literature has focused on three main issues: the impact of privatization, the performance of restructured sectors, and the effects of incentive-based regulation. Policy in this area is very much in flux, as the initial wave of enthusiasm for market solutions confronts unanticipated consequences from their implementation.

The Impact of Privatization

Privatization has been motivated by a range of concerns including the cost of service, pace of technological innovation, and impact of continuing subsidies on government budget deficits. A review of the voluminous literature on privatization is beyond the scope of the present article; see Vickers and Yarrow (1991), La Pörta and López-de-Silanes (1999), and their references. Because privatization has been so closely linked with restructuring and regulatory reform initiatives over the past quarter-century, however, it is useful to highlight key empirical results.

One of the most robust empirical results from this literature is the dramatic impact privatization typically has on cost reduction. This is particularly evident in labor costs. Substantial labor shedding has accompanied privatization almost everywhere, with labor force reductions of 50% or more a common result. Where comparisons with privately owned firms are feasible, studies suggest that postprivatization labor rates approach those at firms that were never government owned. This suggests considerable excess labor in state-owned enterprises. This may reflect the use of state-owned enterprise to promote certain social policies as well as direct operating inefficiencies, suggesting that excess labor costs may include both transfers as well as social deadweight loss.

Privatization often leads to both price and output increases. Rising output despite higher prices suggests the elimination of artificial constraints, such as rationing, or higher quality service for which consumers are willing to pay more. In numerous studies of utility sectors, waiting time for service falls and investment and service quality increase following privatization. This seems most common in sectors that were capital-starved or in which service was underpriced under state-ownership. Even in countries where state-owned utilities were perceived to be functioning relatively well, however, increased investment and innovation may be important benefits of privatization.

One caveat is in order; the frequent concurrence of privatization with changes in the competitive environment can make it difficult to empirically disentangle the relative contributions of ownership change, hard budget constraints, and increased competition to improved performance. Moreover, it appears that the benefits of privatization may be greater when accompanied by increased competition or deregulation. Studies vary considerably in their recognition and successful treatment of this problem.

Performance in Restructured Natural Monopoly Sectors

Vertical restructuring has become common in natural monopoly industries worldwide. In telecommunications, equipment provision and long-distance service were among the first to be carved out of vertically integrated telecommunications firms and opened to competitive suppliers. Power generation has been divested from transmission and distribution functions and opened to competition in many electricity systems. Much of this policy movement appears stimulated by concerns that prices charged by monopoly providers were too high, relative to those attainable by privately owned firms operating in competitive sectors. In some cases, this reflects real operating inefficiencies and reduced innovation under monopoly. In other cases, apparently high prices may in part reflect perverse regulatory pricing schemes. Action on these concerns was facilitated by technological innovations that reduced scale economies at certain production stages and enabled decentralized operations to achieve many of the economies of scope previously attained by vertically integrated firms.

Economists have been especially interested in assessing the competitiveness of restructured markets. Empirical research suggests these markets have yet to converge to competitive ideals. In the US long-distance telephone market, for example, AT&T lost its sanctioned monopoly provider status in 1984, following the earlier technological erosion of its 'natural monopoly.' While long-distance prices have declined, the convergence to competitive norms has been relatively slow, with average per-minute prices more than double estimated long-run incremental costs as late as 1998 (Peltzman and Winston, 2000: p. 77). Some economists have suggested that this reflects relatively weak competition among the small and decreasing number of existing long-distance providers.

Analyses of electricity markets echo this theme. In most restructured electricity markets, unregulated generators bid to supply electricity to the system operator, which determines the dispatch order of generators to supply the electricity demanded at each moment in time. There is considerable evidence of generator market power in many markets, with generators bidding above their marginal generation costs and strategically withholding capacity from the market to increase prices (see Wolfram, 1999; Borenstein et al., 2000; Joskow in Peltzman and Winston, 2000; Joskow and Kahn, 2001).

While many restructured markets have been prone to supracompetitive prices, the extent to which price exceeds cost varies tremendously across different settings. There is growing evidence that performance variations are heavily dependent upon the institutional details of the market's design. These include the way the competitive sector is structured, the terms on which suppliers interface with their customers and the network, and the management of network investments. In electricity markets, for example, the relative sizes of different suppliers, their inframarginal capacity, the use of long-term hedging contracts between generators and their customers, and the degree to which customers see and can respond to real-time price variation have all been demonstrated to influence the exercise of market power.

While competition may enhance cost efficiency in the deregulated sector, restructuring may increase other costs. For example, substitution of market mechanisms for internal decision making or 'command and control' by a vertically integrated firm may increase coordination costs in electricity markets (Joskow in Peltzman and Winston, 2000). In California, for example, reserve margins and hence system operating costs have increased as a result of moving to a dispatch market. Local distribution companies may face difficulties when their requirement to serve, which was imposed in exchange for franchise monopoly in many utility sectors, is maintained, but restructuring eliminates their direct control over electricity supplies, as evidenced by California utilities in 2000–01.

In telecommunications, access issues focus on pricing and interconnection requirements for the 'local loop,' the set of wires connecting individual subscribers to the local network. These problems are magnified by the potential for local exchange carriers to compete with other firms for services beyond the local loop. The determination of access terms has substantial effects for investment incentives by incumbent exchange companies and potential competitors, entry into local, long distance, and enhanced services, and the performance of competitive markets where entry occurs. This is the focus of much current regulatory policy and some empirical analysis (Crandall and Hausman in Peltzman and Winston, 2000).

Effects of Incentive-Based Regulation

While regulatory reform over the past two decades has been substantial, economic regulation continues to play an important role in many sectors. Economic regulation in the new millennium is more market-based than its counterpart of 1970, but less market-driven than current rhetoric would suggest. Policy innovation has been particularly notable outside the USA, perhaps because these systems were being designed *de novo* rather than patched on to existing regulatory practices. Regulation has moved away from traditional rate-of-return regulation, which determined prices through cost-based mark-up rules (Kahn, 1988). More common are higher-powered incentive schemes, such as price caps, which fix price levels or rates of change, and allow firms to keep the surplus generated from unanticipated improvements in efficiency, at least until the next review of the cap or its adjustment factor. The models behind these schemes are discussed elsewhere (*see* Poverty: Measurement and Analysis). It is important to recognize that implementation, however, generally deviates considerably from the theoretical construct.

Many researchers have attempted to measure the impact of incentive-based regulation, particularly in the telecommunications sector. This literature broadly suggests that incentive regulation is accompanied by lower prices. Kridel et al. (1996) suggest that these effects may reflect coincident policies that increase competition more than the incentive effects of incentive regulation itself, however. This interpretation is particularly suggested by the difficulty in finding consistent effects of incentive regulation on costs or productivity, although the short time span of many studies may impede detection of such effects. As earlier empirical work in regulation suggests, one should exercise great caution in affirming the superiority of incentive-based policies based on the theoretical properties of idealized schemes.

Conclusion

The reforms of the 1990s have created a vast new set of institutions and regulatory policies. This is an area in which empirical analysis greatly lags behind the theoretical literature, and the returns to additional careful work are likely to be significant. Empirical research in regulatory economics, argued by some in the 1980s to belong to the domain of economic history, remains vital to guiding, understanding, and evaluating policy initiatives today.

See also: Economic Theory of Regulation; Economics and Ethics.

Bibliography

Averch, H., Johnson, L.L., 1962. Behavior of the firm under regulatory constraint. American Economic Review 52, 1052–1069.

Borenstein, S., Bushnell, J., Wolak, F., 2000. Diagnosing Market Power in California's Restructured Wholesale Electricity Market. National Bureau of Economic Research Working Paper 7868, Cambridge, MA.

Hendricks, W., 1975. The effect of regulation on collective bargaining in electric utilities. Bell Journal of Economics and Management 6, 451–465.

Joskow, P.L., 1974. Inflation and environmental concern: structural change in the process of public utility price regulation. Journal of Law and Economics 17, 291–327.

Joskow, P.L., Kahn, E.P., 2001. A Quantitative Analysis of Pricing Behavior in California's Wholesale Electricity Market during Summer 2000. National Bureau of Economic Research, Cambridge, MA.

Joskow, P.L., Noll, R.G., 1994. Economic regulation during the 1980s. In: Feldstein, M. (Ed.), Economic Policy during the 1980s. University of Chicago Press, Chicago.

Joskow, P.L., Rose, N.L., 1989. The effects of economic regulation. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 1449–1506.

Kahn, A.E., 1988. The Economics of Regulation: Principles and Institutions. MIT Press, Cambridge, MA.

Kridel, D.J., Sappington, D.E.M., Weisman, D.L., 1996. The effects of incentive regulation in the telecommunications industry: a survey. Journal of Regulatory Economics 9, 269–306.

La Pörta, R., López-de-Silanes, F., 1999. The benefits of privatization: evidence from Mexico. Quarterly Journal of Economics 114, 1193–1242.

Laffont, J.-J., 1994. The new economics of regulation ten years after. Econometrica 62, 507–537.

Peltzman, S., 1993. George Stigler's contribution to the economic analysis of regulation. Journal of Political Economy 101, 818–832.

Peltzman, S., Winston, C. (Eds.), 2000. Deregulation of Network Industries: What's Next? AEI-Brookings Joint Center for Regulatory Studies, Washington, DC.

Peoples, J., 1998. Deregulation and the labor market. Journal of Economic Perspectives 12, 111–130.

Rose, N.L., 1987. Labor rent-sharing and regulation: evidence from the trucking industry. Journal of Political Economy 95, 1146–1178.

Stigler, G.J., 1971. The theory of economic regulation. Bell Journal of Economics and Management 2, 3–21.

Stigler, G.J., Friedland, C., 1962. What can regulators regulate? The case of electricity. Journal of Law and Economics 5, 1–16.

Vickers, J., Yarrow, G., 1991. Economic perspectives on privatization. Journal of Economic Perspectives 5, 111–132.

Winston, C., 1998. US industry adjustment to economic deregulation. Journal of Economic Perspectives 12, 89–110.

Wolak, F., 2001. An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market. National Bureau of Economic Research, Cambridge, MA.

Wolfram, C.D., 1999. Measuring duopoly power in the British electricity spot market. American Economic Review 89, 805–826.

Regulation, Political Economy of

Nancy L Rose, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 12967–12970, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Government regulation of firms uses the 'coercive power' of the state to alter firms' pricing, entry, production, investment, and product choice decisions. An extensive empirical literature analyzes the effects of 'economic regulation' of price and entry as well as environmental, health, safety, and information regulation. Results of this work suggest that regulation cannot be understood simply as an efficient intervention to correct market failure. Research on the political economy of regulation seeks instead to understand the origin, structure, and reform of regulatory policy as an outcome of rent-seeking behavior by interest groups mediated through the political process. These models have considerable empirical power in explaining variation in support for particular regulations, but fall short in explaining why some industries are regulated while others are not, and why policy in this area tends to occur in waves.

Government regulation of firms uses the 'coercive power' of the state to alter firms' pricing, entry, production, investment, or product choice decisions. An extensive empirical literature analyzes the effects of 'economic regulation' of price and entry as well as environmental, health, safety, and information regulation. These analyses amassed substantial evidence of high prices, excess profits, and production inefficiencies under economic regulation (see Joskow and Rose, 1989; Joskow and Noll, 1994; and Regulation, Empirical Analysis of). Studies of environmental, health, and safety regulation revealed substantial deviations from least-cost solutions and dramatic variation in measures such as expected cost per life saved across regulatory interventions (see Viscusi, 1994; Gruenspecht and Lave, 1989). These outcomes, so systematic and pervasive, were difficult to reconcile with 'public interest' models of regulatory intervention. Examination of regulatory systems further suggested that they often deviated from cost minimization or efficiency objectives by design. This begged for an explanation beyond bureaucratic incompetence or malfeasance, common to earlier Marxian or 'capture' theories of government. Stigler's (1971) *Economic Theory of Regulation* provided the conceptual underpinnings of what has become a vast enterprise among economists and political scientists to understand the origin, objectives, maintenance, and reversal of regulatory policy in the context of rational choice models of political behavior. These simple insights stimulated a substantial theoretical and empirical literature (Noll, 1989; Peltzman, 1993).

Theories of legislative and bureaucratic behavior have become increasingly sophisticated since the 1980s. The literature has moved well beyond Stigler's simple model to incorporate the role of competing interests, policy entrepreneurs, majoritarian political institutions, legislative oversight of regulatory agencies and commissions, and asymmetric information (see Noll, 1989; Joskow and Noll, 1994; Laffont, 1999, and Regulatory Agencies). The empirical literature has struggled to keep pace with these developments.

Common to virtually all empirical studies in this area is the importance, and difficulty, of defining well-specified alternatives to the political economy model under investigation. To test a given model of political behavior, or to interpret results as consistent with a given model, one must understand how predicted outcomes differ between competing hypotheses of behavior and the model being tested. For example, there is a natural tendency to identify rent conveyance with rent-seeking activity. In some cases, however, increased rents for regulated firms may be consistent with both successful rent-seeking behavior by those firms and elimination of some market inefficiency. It is important, though often quite difficult, for empirical analyses to identify and attempt to distinguish alternatives such as these. Studies vary considerably in the degree to which they meet this standard.

This article introduces the main empirical methodologies used to explore the political economy of economic, environmental, health and safety regulation. Examples of their applications are briefly discussed. Interested readers are urged to pursue the surveys listed in the references for more complete discussions of this literature.

Measuring Distributional Effects

Empirical studies of political economy generally begin with estimates of the distributional effects of a given policy. While many empirical methods may be used (see Joskow and Rose, 1989 and Regulation, Empirical Analysis of), asset market 'event studies' have been particularly popular. These measure price movements in capital assets, such as company share prices, at the time policy changes are first announced. If the announcement is not previously anticipated, and asset markets efficiently process information, these price changes provide estimates of the policy's effect on the capitalized value of expected rent streams. This method may be used to measure the effects of the policy change on firms within a regulated industry, corporate customers and suppliers, and possible competitors. The distributional implications for individual or small business customers, employee groups, and other noncorporate interests usually must be measured using alternative methods.

The simplest political economy studies use measured distributional effects to infer the relative influence of different interest groups in the political process. These implicitly assume that the observed changes in rent distribution are anticipated and intentional, and that larger rent

gains reflect greater political power. Consider, for example, the extensive debate over the origin of the first federal economic regulatory commission in the US, the Interstate Commerce Commission created in 1887 to regulate railroads. Was this legislation intended to solve an inefficiency in an industry characterized by large sunk costs and natural monopoly conditions? If the legislation was a response to interest group politics, did it respond to agrarian interests and other shippers' demands to constrain monopoly pricing by railroads, or was it a railroad-sponsored effort to establish a government-enforced cartel? Did legislation intend one outcome, only to be thwarted by the regulatory agency's subsequent behavior? There have been numerous empirical studies attempting to answer these questions, many using railroad share price responses to various regulatory 'events' (e.g., Prager, 1989; Mullin, 2000). In these, railroad share price gains in response to the Interstate Commerce Act, for example, have been interpreted as evidence that the regulation was intended to benefit railroads, supporting Stiglerian models of regulatory politics. While it may be reasonable to assume that *ex post* gains identify *ex ante* intended beneficiaries, this should be made explicit and considered against alternative explanations. To move convincingly from a positive statement about the distributional consequences of a policy to an inference about its political motivation requires an understanding of distributional effects under leading alternative models.

Estimating Political Support for Regulation

A second empirical methodology estimates the determinants of regulatory incidence or political support as a function of underlying economic interests. These studies generally involve three steps. First, *ex ante* winners and losers of a regulatory policy are identified, using either prior estimates or a model of the regulation's effect. For example, Stigler's (1971) analysis of truck weight laws modeled state truck weight and size limits as raising the cost of trucking service, particularly over longer hauls. Truckers therefore lose from lower limits, as do their current customers, businesses using their own trucks, and potential customers for whom trucking rates constrain alternative transportation rates. Competitors to trucks, such as railroads, gain from tighter restrictions. There may be diffuse 'public interests' as well, for example, in limiting damage to roadbeds and reducing congestion, though these may well be endogenous.

Second, the importance of the identified economic interests within a given jurisdiction or at a particular point in time is measured. There is considerable variation in how this is done. For example, the number of trucks, scaled or unscaled, the dollar volume of activity in trucking, and the fraction of freight hauls susceptible to potential truck competition are all possible measures of the importance of trucking interests in a jurisdiction.

Finally, the incidence or intensity of regulation, or the political support for regulation, is estimated to be a function of the measured economic interests and any other relevant factors. Stigler used this method to explain differences in the intensity of truck weight limitations across states, for example. Other studies apply this technique to model variation in political support, such as voting records, as a function of constituent economic interests in the policy, *inter alia*.

Most studies in this area find that constituent interests are significant determinants of regulatory policy, though interests must be defined broadly to include constituent preferences (such as conservation or environmental preferences) as well as direct economic interests. Detailed studies of voting behavior suggest that the underlying models may be quite complex, however; see, for example, Joskow and Schmalensee (1998) on the US acid rain program. The significant impact of legislators' ideology on their voting behavior (e.g., Kalt and Zupan, 1984) suggests possible slack in the agency relationship between constituents and their legislators, as well support for more sophisticated models of voting in majoritarian systems across issues and over time.

Modeling Regulatory Agency Behavior

A final strand of empirical work in political economy begins with the recognition that most regulatory legislation leaves considerable discretion over its implementation to regulatory agencies or commissions. Early work on regulatory 'capture' assumed that industry groups could influence regulators to use this discretion to benefit the industry in direct contravention to legislative intent. As economists and political scientists began to explore the extent to which apparent regulatory capture was instead a reflection of legislative intent, they confronted the questions of how much discretion regulators retain and how regulators use that discretion to form policy (e.g., McCubbins et al., 1987). This led to an extensive literature on the operation and oversight of the administrative process, most of it within political science. This area remains one of the least-developed empirical literatures in regulatory economics.

While some early analyses explored the responsiveness of regulatory agency decision-making to congressional preferences, most economics research in this area analyzes the responsiveness of agency decision-making to interest groups and various measures of efficiency or social welfare. This is methodologically quite similar to the legislative studies described in Section Estimating Political Support for Regulation, with individual regulatory decisions as the unit of observation. As in those studies, the results suggest a role for concentrated economic interests to influence regulatory outcomes. This does not itself demonstrate that regulators respond to interest groups in contravention with broader social objectives, however. For example, studies have found that despite significant overall delays in the process by which the Food and Drug Administration (FDA) controls entry of new pharmaceutical products, the FDA responds more quickly to drugs with greater market potential (Dranove and Meltzer, 1994). In this case, both producer and social interests are aligned with reducing delay of an innovation.

Conclusion

While research on the political economy of regulation has made significant progress since the 1970s, fundamental

questions remain unanswered by either theoretical or empirical work (Peltzman, 1993; Joskow and Noll, 1994). One of the most basic is why some industries are regulated (or deregulated) and others are not. Another is why regulatory policy seems characterized by 'waves' or cycles: changes tend to occur across broad ranges of industries and sectors within fairly short periods, separated by longer periods of relative inactivity. Considerable scope remains for additional work in this area.

Editor's Note

Dr. Nancy Rose's current thinking on the political economy of regulation can be found in N.L. Rose, 2014.

See also: Antitrust Policy: Lessons from the US; Broadcasting: Regulation; Environmental Economics; Occupational Health and Safety, Regulation of; Regulation, Economic Theory of; Regulation, Empirical Analysis of; Regulatory Agencies; Safety, Economics of.

Bibliography

Dranove, D., Meltzer, D., 1994. Do important drugs reach the market sooner? RAND Journal of Economics 25, 402–423.

Gruenspecht, H.K., Lave, L.B., 1989. The economics of health, safety, and environmental regulation. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 1507–1550.

Joskow, P.L., Noll, R.G., 1994. Economic regulation during the 1980s. In: Feldstein, M. (Ed.), Economic Policy During the 1980s. University of Chicago Press, Chicago.

Joskow, P.L., Rose, N.L., 1989. The effects of economic regulation. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 1449–1506.

Joskow, P.L., Schmalensee, R.L., 1998. The political economy of market-based environmental policy: The US acid rain program. Journal of Law and Economics 41, 37–83.

Kalt, J.P., Zupan, M.A., 1984. Capture and ideology in the economic theory of politics. American Economic Review 74, 279–300.

Laffont, J.-J., 1999. Political economy, information and incentives. European Economic Review 43, 649–669.

McCubbins, M., Noll, R., Weingast, B., 1987. Administrative procedures as instruments of political control. Journal of Law, Economics and Organizations 3, 243–277.

Mullin, W., 2000. Railroad revisionists revisited: Stock market evidence from the progressive era. Journal of Regulatory Economics 17, 25–47.

Noll, R.G., 1989. Economic perspectives on the politics of regulation. In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam.

Peltzman, S., 1993. George Stigler's contribution to the economic analysis of regulation. Journal of Political Economy 101, 818–832.

Prager, R., 1989. Using stock price data to measure the effects of regulation: The Interstate Commerce Act and the Railroad Industry. RAND Journal of Economics 20, 280–290.

Stigler, G.J., 1971. The theory of economic regulation. Bell Journal of Economics and Management 2, 3–21.

Viscusi, K., 1994. Health and safety regulation. In: Feldstein, M. (Ed.), Economic Policy During the 1980s. University of Chicago Press, Chicago.

Rose, N.L., 2014. Learning from the past: insights for the regulation of economic activity. In: Rose, N.L. (Ed.), Economic Regulation and Its Reform: What Have We Learned? NBER and University of Chicago Press, Chicago.

Regulatory Agencies

Christopher H Foreman, Maryland School of Public Policy, University of Maryland, College Park, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'regulatory agency' is attached mainly to administrative entities intended to monitor and influence business activity in response to varieties of market failure. Observers have undertaken four general lines of inquiry into, and commentary about, such agencies: (1) creation dynamics, (2) types and tactics of influence by governmental principals and by attentive nongovernmental players, (3) circumstances that may undermine agency existence, and (4) how well regulatory agencies serve their assigned purposes. The United States and Europe manifest significant similarities and differences.

Broadly defined, regulation includes any governmental effort to control behavior by other entities, such as business firms, subordinate levels of government, nonstate entities (such as political campaigns or civil society actors), or individuals. Sovereign governments may cooperate under regimes of mutual constraint that yield rules or decisions characterized as regulations. Most political and scholarly discourse on regulation addresses programs or policies intended to monitor and restrict commercial activity. Regulatory agencies are, generally speaking, those public entities holding delegated authority to implement such programs and policies.

Persons not thoroughly familiar with comparative regulatory policy are cautioned regarding important, and potentially confusing, differences in terminology between the United States and Europe. The terms 'regulation,' 'directive,' and 'decision' are not synonyms in European Union (EU) parlance where each has a distinctive connotation and is regularly described as a type of 'legislation' (see e.g., http://ec.europa.eu/legislation/index_en.htm). In the United States, one never describes the national bicameral Congress as in itself a 'regulatory agency' but rather as a legislative body that authorizes, appropriates funding, for and oversees activity within such agencies. In the EU, a 'regulation' is 'similar to a national law with the difference that it is applicable within all EU countries,' while a 'directive' establishes 'general rules to be transferred into national law by each country' and a 'decision' addresses 'a particular issue and specifically mentioned persons or organizations.'

Within business regulation in the United States four primary types may be distinguished (Breyer, 1982). First, regulation may control market entry (as with occupational licensing or air carrier certification). Second, it may (as in the case of electric utilities and other 'natural monopolies') establish directly the prices that firms can charge customers. Third, it may limit the competitive practices available to firms, especially as regards price setting (e.g., the law of antitrust), advertising, and disclosure. Fourth, regulation may aim to improve and standardize (through industry-wide rules, inspection regimes, or premarket approval processes – e.g., commercial pharmaceuticals) the quality of specific work practices or products, often to protect human health and safety or the environment. The first and fourth types may overlap in practice; carrier certification, for example, necessarily creates entry barriers in the service of air safety. These four types are sometimes subsumed under two more general headings: 'economic' (i.e., price, entry, and competitive practices) and 'social' (meaning generally health, safety, and environmental) controls. More refined policy rationales for each type of regulation, broadly described as varieties of 'market failure' (e.g., curbing monopoly and cartel power, moral hazard, planning on behalf of future needs, imperfect and asymmetric information, provision of public goods, control of negative externalities) may be identified (Baldwin and Cave, 2012).

Social scientists and historians have posed three primary empirical questions about regulatory programs. First, what forces lead to their creation (Kolko, 1965; Wilson, 1980)? Second, political scientists and economists have asked what factors influence agencies, prompted by a desire to fix more precisely, in the face of governmental fragmentation, the true institutional locus of political control over policy-relevant behavior by agencies (Wood, 1988; Moe, 1982, 1985) or the precise role of supervisory tools such as budgets (Carpenter, 1996) or legislative oversight and procedural constraints (McCubbins et al., 1987)? Third, given the inertial forces that regularly contribute to the durability of many public programs (i.e., combinations of concentrated support, diffuse criticism, and widespread indifference) under what conditions may regulatory programs be scaled back or terminated (Derthick and Quirk, 1985)?

Normative questions also proliferate. Much business regulation, especially in the United States, has been prone to intense controversy. Both business representatives and policy-focused academics have pursued long-running debates about the merits of various regulatory programs, the quality of their specific decisions, and reforms to enhance accountability, independence, expertise, administrative efficiency, and sensitivity to the costs that regulation imposes on specific business sectors and on society generally. With the creation and evolution of the EU arose inevitable questions of the correct balance to be struck among a plethora of public values, including the tension between established national preferences and practices on the one hand and the need for effective supranational coordination and consistency on the other (Gerardin et al., 2005).

Empirical Inquiry

Students of regulation have long found the formally stated public goals of regulation inadequate to explain program creation and maintenance. From the 1950s to the 1970s, working within such disciplines as economics, political science, and historical analysis, commentators viewed regulatory programs frequently as being prone to 'capture' by regulated interests (Huntington, 1952; Bernstein, 1955; Kolko, 1965; Stigler, 1971). By some accounts, regulation was said to have arisen because a diffuse majority favoring regulation (and motivated by a 'public interest' rationale, such as the elimination of predatory pricing or consumer fraud) succeeded, at least briefly, in obtaining legislative consensus, often in the wake of well-publicized scandal. Over time, however, according to this 'life cycle theory' of regulation posited by Bernstein (1955), waning public attention and media scrutiny was said to have allowed resourceful regulated interests, whose attention to regulatory program activity did not wane, to exercise a strong (even dominant) influence over the program, thus thwarting its original intent.

A different version of capture analysis suggested that such programs were sometimes created, at least partially, at the behest of economic interests, perhaps as an exercise in 'rent-seeking' behavior (Dudley and Brito, 2012). Agencies might thus, in effect, be 'born captured' due to business incentives to seek stable and reliably profitable operating environments. The capture model of regulation is in some respects akin to a traditional political science metaphor, widely employed during the same period, describing the politics of narrowly distributive public policies in terms of 'iron triangles' composed of interest groups, legislative committees, and agencies. The similarity is striking, even though an analytic scheme widely appropriated among political scientists (Lowi, 1964) highlighted a distinction between 'distributive' and 'regulatory' politics. From the 1970s onward, however, scholars of public policy challenged increasingly the descriptive accuracy of the 'capture' and 'iron triangle' metaphors (Heclo, 1978; Wilson, 1989).

In the United States especially, this challenge stemmed importantly from the creation of regulatory programs (e.g., in environmental, consumer, and worker protection) where lack of media scrutiny, public indifference, and business dominance plainly did not apply over the long run (Wilson, 1980). The 1960s and the 1970s were notable (Sunstein, 1990) for a number of "bold regulatory initiatives in a number of new areas" (p. 25). Indeed, the creation of these programs often represented sustained triumphs by idea-driven policy entrepreneurs over business interests that had to adjust to the presence of long-term political competitors. In one prominent instance, the regulation of nuclear power generation, a policy milieu changed dramatically from one to which the term 'capture' could arguably have applied in the 1950s. By the 1990s, the politics of nuclear power in the United States had evolved into a far more complicated set of power relations, in which environmental advocates and citizen groups competed more effectively with the nuclear industry (Duffy, 1997).

Throughout the 1980s, political science commented increasingly about the inadequacy of simple generalizations regarding regulatory politics as a whole. The concentration of costs and benefits prevailing in regulation were found to differ importantly among agencies, and across issues within a single agency (Wilson, 1980). The politics of broadcast regulation, pursued within a markedly vague standard of 'public interest, convenience, and necessity,' clearly varied between technical industry structure matters and issues of program content (Krasnow et al., 1982). Within public utility regulation, issues were determined to differ in complexity and conflict, with significant implications for the power of different kinds of advocates (Gormley, 1983), and the characteristics of commission personnel were also found to vary in ways important to generating distinctive regulatory outputs (Berry, 1984). Moe (1985) found decision making by the National Labor Relations Board to stem from quite complex causal influences including an array of external and internal institutional actors and exogenous conditions.

Regulatory policy became the staging ground for scholarly argument over the relative independence of agencies. A conventional position among journalists and political scientists had for many years held that Congress did little to oversee bureaucratic behavior, largely because of the modest political incentives said to prevail. In arguing against what he called 'interest group liberalism,' Lowi (1969) bemoaned the vagueness that had crept into legislative delegations for regulatory and other programs. A later principal-agent perspective on the legislative–executive relationship argued for the existence of a more robust oversight mechanism grounded in a rational legislative approach to structure and procedure that appeared superficially to be weak, a system that, paradoxically, worked precisely by appearing to have failed (McCubbins and Schwartz, 1984). Examination of agency case histories determined that widely trumpeted accusations of 'runaway bureaucracy' often directed at regulation were misplaced (Foreman, 1988).

The classic political science question 'who governs?' (Dahl, 1961) has been answered repeatedly over many years for regulatory agencies in numerous ways by observers employing various methods and intellectual vantage points. Because regulation is politically contentious and generates substantial numbers of comparable implementation and enforcement decisions over time, political scientists (often working within the tradition of rational choice models and principal-agent theory) have gravitated to regulation as a window on relative institutional influence more generally. Some analysts have emphasized the power of legislative principals (activated by the material stakes of affected parties) over agency behavior (Weingast and Moran, 1983; Weingast, 1984). Moe (1982), examining time series data drawn from three US regulatory commissions concluded that presidential power deserved emphasis. On the other hand, some analysts have looked inside regulatory bureaucracies to the professional worldviews and incentives of staff members as key drivers of agency behavior. In antitrust enforcement, economists and attorneys are among those professionals whose views have been found to have a profound impact on agencies, both in case studies (Katzmann, 1980; Weaver, 1977) and in quantitative analysis (Eisner and Meier, 1990). The theory of 'overhead democracy' embraced by Wood and Anderson (1993) ascribed variation in both the level and substance

of antitrust regulation to a complex set of institutional forces; changing presidential administrations, congressional committee preferences, and court decisions all helped to mold antitrust activity.

The US Food and Drug Administration (FDA), responsible for premarketing the clearance of drugs and medical devices as well as scrutiny of the food supply (save meat, eggs, and poultry, which fall under the purview of the US Department of Agriculture) epitomizes the diversity of considered views that have evolved regarding power relations. Shipan (2004) develops a spatial model of political oversight, applies it to FDA, and finds that the degree and locus of legislative influence varies, with the agency acting autonomously in some instances. Carpenter (2001, 2010) has emphasized agency 'reputation' and executive entrepreneurship as the major drivers of substantial bureaucratic autonomy in both the FDA and other administrative contexts. Alternatively, observers beyond political science often appear more wary of corporate influence in the FDA's political environment (Hilts, 2003; Nestle, 2002). Both mocked by business-friendly politicians as the supreme 'job killer' in the United States (Carpenter, 2010: p. 731) and challenged by consumer activists and environmental enthusiasts to strive ever harder on behalf of the public's health and safety, the FDA is inherently controversy prone.

Deregulation is rarer, and politically more problematic, than regulatory initiatives. As regards social regulation, such goals as a clean environment, and safe workplaces and pharmaceutical products, are popular with the public. In economic regulation, organized stakeholders would be expected to resist unfavorable change (i.e., alteration of profitable regulated environments). Given such forces, it is worth emphasizing that deregulation has occurred on occasion, and asking how this has come about. Derthick and Quirk (1985) point to the power of ideas and effective leadership. They observe that, by the late 1970s, economists had mounted a sustained, coherent, and nearly unanimous attack on the underlying justifications for price and entry regulation in the transportation sector. This procompetition critique gained added political impetus from the increasing visibility of economists in national policy debates, from the infiltration of economists into leadership of the regulatory apparatus, and from the transformation of deregulation into an antiinflation and consumerist cause that both liberals and conservatives could support. Strong presidential and legislative leadership also proved to be pivotal. One dramatic result was the phased legislative elimination, culminating in 1984, of the Civil Aeronautics Board, which had, since the 1930s, managed pricing and entry for commercial interstate aviation in the United States.

Normative Assessments

Regulatory programs regularly inspire supporters and detractors in the social sciences on grounds of efficiency (whether administrative or economic), equity, and democratic accountability. Even Sunstein (1990), a strong academic defender of regulation (who would also serve as top regulatory policy reviewer for a US president) acknowledges many ways that regulation may fail. Unlike price and entry regulation, which regularly has found advocates of abolition in the economics profession, other types of regulation enjoy strong academic support on principle (e.g., recognition of 'market failure'), though often not in specific detail. Leading the charge for varieties of regulatory reform have been economics-oriented analysts associated with policy research institutions or 'think tanks.' Criticisms of environmental policy stress varieties of putative inefficiency or irrationality. For many years, critics have argued that costs often exceed benefits, are higher than necessary to achieve a given level of environmental benefit, and perhaps are masked altogether by inadequate economic analysis (Kraft, 1996: pp. 165–166; Litan and Nordhaus, 1983; Lave, 1981). Inefficiency is sometimes argued to have been caused by the logic of political compromise (Ackerman and Hassler, 1981) or by unwillingness to impose overtly visible losses on large numbers of voters (Nivola and Crandall, 1995). Inadequate monitoring has long been said to bedevil the assessment of pollution control efficacy (Crandall, 1983; Davies and Mazurek, 1998). A broader, efficiency-based critique encompasses additional concerns, including the discordant, meandering, and arguably panic-stricken character of regulatory politics and policy (Breyer, 1993). Critics point to a chronic inability to set risk priorities, to confront risk–benefit and risk–risk trade-offs (Graham and Wiener, 1995), and to reconsider regulatory statutes in the light of such concerns.

Especially in the United States support for risk assessment and cost–benefit analysis have in recent decades somewhat displaced precautionary 'zero-risk' stances in regulatory reform. The Delaney Clause enacted in the US Food, Drug, and Cosmetic Act in 1958, one of the earliest and best-known examples of the zero-risk mind-set, held any food additive intolerable in any amount in the food supply if determined by any study to cause cancer in humans or animals (Vogel, 2012). After many years of discrediting criticism, the provision was finally replaced in food safety law in 1996 by a standard of 'reasonable certainty of no harm.' Moreover, US presidents of both parties since the early 1970s, and especially since the Reagan era, have embraced formal and centralized regulatory review processes – note Sunstein's role alluded to above – grounded in cost–benefit and cost-effectiveness analyses.

In a quite different vein, however, environmental, agricultural, energy, and other related policies (and thus the regulatory agencies central to forging and implementing them) have encountered varieties of 'grassroots' community mobilization and critical advocacy, creating an evolving discourse, simultaneously scholarly and public, under such equity-centered rubrics as 'environmental justice' (Foreman, 1998; Shrader-Frechette, 2005), 'food justice' (Gottlieb and Joshi, 2010), and 'climate justice' (Posner and Weisbach, 2010; Tokar, 2010). These framing rubrics emphasize that low-income communities in the industrial West, and the global poor generally, face disproportionate negative externalities, and enjoy less-than-commensurate gain, from many commercial pursuits. Such insurgent activist-generated frames naturally challenge established regulatory agency mandates and cultures. They also often depart from a 'disinterested politics' grounded in scientific rationality and formalized risk assessment in favor of the presumed 'experiential knowledge'

of health impacts associated with residence in a particular locale (Tesh, 2000). Overall, these competing discourses regularly exhibit considerable polarization, with efficiency-based arguments generally finding little support among the champions of such equity-based critiques, and vice versa.

The sheer inability of regulation to keep pace with the innovation it purportedly oversees is a common criticism. Pharmaceuticals and toxic chemicals have long been targets of concern. More recently, the global Great Recession triggered by a series of financial shocks in the market for an array of underregulated (and often poorly understood) financial instruments produced significant (if to many observers problematic) new regulatory legislation in the United States. It also led political scientists to diagnose a systemic problem of 'political bubbles' (grounded in combinations of ideology, interest, and institutional forces) that promote, rather than curb, a damaging inclination to speculative frenzies (McCarty et al., 2013). Buiter (2008) suggests the term 'cognitive regulatory capture' as applying to the relationship between Wall Street and the US Federal Reserve in the period leading up to the Great Recession.

International and Comparative Perspectives

All developed nations regulate but vary substantially in both regulatory structure and regulatory political culture. The United States has often been portrayed as having a markedly more adversarial and legalistic system than other countries (Kagan, 2003). One study concluded that while "regulatory power has been directed at very similar ends (i.e., control of carcinogens)" in Britain, France, Germany, and the United States, "the openness, complexity and divisiveness of American decision making have no parallel in Europe" (Brickman et al., 1985: pp. 301–302). It is not clear, however, that the American style of regulation is necessarily more stringent than that of other nations. Conceding the more open regulatory style of the United States in comparison with Canada, Harrison and Hoberg (1994), nevertheless, do not find the latter consistently less stringent than the former.

The most up-to-date and wide-ranging comparison of health, safety, and environmental regulation in the United States and Europe (Vogel, 2012) argues that, since about 1990, the EU has embarked on a path of notably more stringent regulation than the United States in such areas as chemical hazards, greenhouse gas emissions, cosmetics, and various areas of food safety including hormone-treated beef, antibiotics in animal feed, and genetically modified organisms. The United States has in recent years retreated considerably from its former inclination to embrace the 'precautionary principle' (holding that a product or process should not be deemed safe until demonstrated to be so) and instead elevated risk assessment and cost–benefit analysis as regulatory decision tools. Europe, on the other hand, has moved in a very different direction, explicitly privileging precautionary thinking and acceding to the demands of member governments, nongovernmental organizations (NGOs), and popular sentiment in favor of regulatory stringency. The key supranational regulatory agency of the EU, the European Commission, wielding powers of both policy initiation and executive implementation, and "where most European legislation is drafted before being submitted to the Council of Ministers and the European Parliament" has been proved to be especially open to "the expertise of environmental NGOs in formulating and drafting new regulatory proposals" (Vogel: p. 245).

Lacking both the heritage of nationalized industry and the stronger commitment to welfare state provision common in Europe, it was perhaps natural that the United States would sooner and more strongly adhere to a path of regulatory agency policing of the private sector. Thatcher (2005) observes that historically "very few [independent regulatory agencies] existed in Europe – they were very much an American institution" (p. 50). However, in the years since the formal creation of the EU in 1993, the continent has increasingly been described in terms of a 'regulatory state' (Majone, 1996). Ambition to create a single European market, manifest in the 1957 Treaty of Rome (and superseded most recently by the 2009 Treaty of Lisbon) necessitated, through regulatory innovation, the elimination of barriers to economic exchange among member nations, though this was impeded by "a proliferation of national rules as the member states increasingly adopted national measures to protect their industries and to respond to new concerns about consumer and environmental protection in the late 1970s and early 1980s" (Young, 2005: p. 96). Even today the emergence and continued strengthening of the EU has not obliterated national variation and vested interests. Though notably more resistant overall to genetically modified organisms than the United States, the political efficacy of opponents varies among European nations and can be blunted by a strong biotechnology sector (Kurzer and Cooper, 2007). Commission efforts to pursue tobacco advertising regulation prompted intense and long-running disagreement grounded in varying national policy outlooks (Duina and Kurzer, 2004). In telecommunications, the Commission, "mainly because of opposition from regulators in the EU member states, tried but failed to create an EU agency" but succeeded in aviation safety (Levi-Faur, 2011). Indeed, channels of consultation and cooperation (formal and informal) among the key EU institutions, an array of nongovernmental interests, and the European member governments are pervasive with policy results varying across regulatory issues and sectors.

Observers of the European regulatory milieu commonly point to its proclivity toward "a particularly intense form of multilateralism" and western Europe's "apparent predisposition to engage in cross-border regime building" as driven significantly by "specific features of history and geography" [and]… "a political culture of investing in institutionalized cooperation with neighbors and partners, at least in the period since the second world war" (Wallace et al., 2005: p. 5). Multilateral regulatory policy innovation has faced much rougher sledding in the United States. Vogel (2012: p. 6) notes that while "every major environmental agreement supported by the United States has been ratified by the member states of the EU and/or the EU itself, since the early 1990s the United States has not ratified twelve important international environmental agreements ratified by the EU and/or it members states."

Globalization has also prompted a major voluntary firm-level effort to promote harmonization under the ISO 14001 framework. This establishes criteria by which a firm or an organization may seek certification for its 'environmental management system' (Delmas, 2004). Hence, governmental regulatory agency mandates are by no means the sole formal mechanism for attempting to create incentives and processes supportive of a significant category of regulatory policy goals.

Strong divisions of formal authority among sovereign national and subnational governmental layers found in the United States have no exact analog in the EU context (or in most others for that matter), although diffusion of responsibility and significant informal interaction characterize both regimes. The EU maintains a legislative arrangement embracing (at the time of this writing) a 766-member popularly elected European Parliament, a 28-seat Council of the EU (also known as the Council of Ministers) representing member states, and a 28-seat European Commission charged with initiating legislation. America's bicameralism is only broadly similar with a 100-seat Senate (each state represented by two popularly elected senators) and 435-seat House of Representatives, a structure recognizably mirrored in all but one of its 50 subnational state legislatures (i.e., the unicameral exception of Nebraska). The United States and the EU are alike, however, in their respective pursuits of democratic management and oversight of the use of authority delegated to agencies with varying regulatory mandates and degrees of independence.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Interest Groups, History of.

Bibliography

Ackerman, B.A., Hassler, W.T., 1981. Clean Coal/Dirty Air: Or How the Clean Air Act Became a Multibillion Dollar Bail-Out for High-Sulfur Coal Producers and What Should Be Done About It. Yale University Press, New Haven.

Baldwin, R., Cave, M., 2012. Understanding Regulation: Theory, Strategy and Practice, second ed. Oxford University Press, Oxford.

Bernstein, M.H., 1955. Regulating Business by Independent Commission. Princeton University Press, Princeton, NJ.

Berry, W.D., 1984. An alternative to the capture theory of regulation: the case of state public utility commissions. American Journal of Political Science 28, 524–558.

Breyer, S., 1982. Regulation and Its Reform. Harvard University Press, Cambridge, MA.

Breyer, S., 1993. Breaking the Vicious Circle: Toward Effective Risk Regulation. Harvard University Press, Cambridge, MA.

Brickman, R., Jasanoff, S., Ilgen, T., 1985. Controlling Chemicals: The Politics of Regulation in Europe and the United States. Cornell University Press, Ithaca, NY.

Buiter, W.H., 2008. Lessons from the North Atlantic financial crisis. In: Paper Presented at the Conference on "the Role of Money Markets". Columbia Business School and the Federal Reserve Bank of New York.

Carpenter, D.P., 1996. Adaptive signal processing, hierarchy, and budgetary control in federal regulation. American Political Science Review 90, 283–302.

Carpenter, D.P., 2001. The Forging of Bureaucratic Autonomy: Reputations, Networks and Policy Innovation in Executive Agencies, 1862–1928. Princeton University Press, Princeton, NJ.

Carpenter, D.P., 2010. Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA. Princeton University Press, Princeton, NJ.

Crandall, R.W., 1983. Controlling Industrial Pollution: The Economics, and Politics of Clean Air. Brookings Institution Press, Washington, DC.

Dahl, R., 1961. Who Governs? Democracy and Power in an American City. Yale University Press, New Haven and London.

Davies, J., Mazurek, C., 1998. Pollution Control in the United States: Evaluating the System. Resources for the Future, Washington, DC.

Delmas, M.A., 2004. Environmental management standards and globalization. In: Vogel, D., Kagan, R.A. (Eds.), Dynamics of Regulatory Change: How Globalization Affects National Regulatory Policies. University of California Press, Berkeley, Los Angeles and London.

Derthick, M., Quirk, P.J., 1985. The Politics of Deregulation. Brookings Institution Press, Washington, DC.

Dudley, S.E., Brito, J., 2012. Regulation: A Primer, second ed. Mercatus Center and George Washington University Regulatory Studies Center, Washington, DC.

Duffy, R.J., 1997. Nuclear Politics in America: A History and Theory of Government Regulation. University Press of Kansas, Lawrence, KS.

Duina, F., Kurzer, P., 2004. Smoke in your eyes: the struggle over tobacco control in the European Union. Journal of European Public Policy 11, 57–77.

Eisner, M.A., Meier, K.J., 1990. Presidential control versus bureaucratic power: explaining the Reagan revolution in antitrust. American Journal of Political Science 34, 269–287.

Foreman Jr., C.H., 1988. Signals from the Hill: Congressional Oversight, and the Challenge of Social Regulation. Yale University Press, New Haven and London.

Foreman Jr., C.H., 1998. The Promise and Peril of Environmental Justice. Brookings Institution Press, Washington, DC.

Gerardin, D., Munoz, R., Petit, N., 2005. Regulation through Agencies in the EU: A New Paradigm of European Governance. Edward Elgar, Cheltenham, UK.

Gormley, W.T., 1983. Policy, politics, and public utility regulation. American Journal of Political Science 27, 86–105.

Gottlieb, R., Joshi, A., 2010. Food Justice. The MIT Press, Cambridge, MA.

Graham, J.G., Wiener, J.B. (Eds.), 1995. Risk versus Risk: Tradeoffs in Protecting Health and the Environment. Harvard University Press, Cambridge, MA.

Harrison, K., Hoberg, G., 1994. Risk, Science, and Politics: Regulating Toxic Substances in Canada and the United States. McGill-Queen's University Press, Montreal.

Heclo, H., 1978. Issue networks and the executive establishment. In: King, A. (Ed.), the Political System. American Enterprise Institute, Washington, DC, pp. 87–124.

Hilts, P.J., 2003. Protecting America's Health: The FDA, Business, and One Hundred Years of Regulation. Alfred A. Knopf, New York.

Huntington, S.P., 1952. The Marasmus of the ICC. Yale Law Journal 61, 467–509.

Kagan, R.A., 2003. Adversarial Legalism: The American Way of Law. Harvard University Press, Cambridge, MA.

Katzmann, R.A., 1980. Regulatory Bureaucracy: The FTC and Antitrust Policy. The MIT Press, Cambridge, MA.

Kolko, G., 1965. Railroads and Regulation, 1877–1918. Norton, New York.

Kraft, M.E., 1996. Environmental Policy and Politics: Toward the Twenty-First Century. HarperCollins, New York.

Krasnow, E., Longley, L., Terry, H., 1982. The Politics of Broadcast Regulation. St Martin's Press, New York.

Kurzer, P., Cooper, A., 2007. What's for dinner?: European farming and food traditions confront American biotechnology. Comparative Political Studies 40, 1035–1058.

Lave, L.B., 1981. The Strategy of Social Regulation: Decision Frameworks for Policy. Brookings Institution Press, Washington, DC.

Levi-Faur, D., 2011. Regulatory networks and regulatory agencification: towards a Single European Regulatory Space. Journal of European Public Policy 18, 810–829.

Litan, R.E., Nordhaus, W.D., 1983. Reforming Federal Regulation. Yale University Press, New Haven and London.

Lowi, T.J., 1964. American business, public policy, case studies and political theory. World Politics 16, 677–715.

Lowi, T.J., 1969. The End of Liberalism: Ideology, Policy and the Crisis of Public Authority. Norton, New York.

Majone, G., 1996. Regulating Europe. Routledge, London and New York.

McCarty, N., Poole, K.T., Rosenthal, H., 2013. Political Bubbles: Financial Crises and the Failure of American Democracy. Princeton University Press, Princeton and Oxford.

McCubbins, M., Schwartz, T., 1984. Congressional oversight overlooked: police patrols vs. fire alarms. American Journal of Political Science 28, 165–179.

McCubbins, M.D., Noll, R.G., Weingast, B.R., 1987. Administrative procedures as instruments of political control. Journal of Law, Economics and Organization 3, 243–277.

Moe, T.M., 1982. Regulatory performance and presidential administration. American Journal of Political Science 26, 197–224.

Moe, T.M., 1985. Control and feedback in economic regulation: the case of the NLRB. American Political Science Review 79, 1094–1116.
Nestle, M., 2002. Food Politics: How the Food Industry Influences Nutrition and Health. University of California Press, Berkeley and Los Angeles.
Nivola, P.S., Crandall, R.W., 1995. The Extra Mile: Rethinking Energy Policy for Automotive Transportation. Brookings Institution Press, Washington, DC.
Posner, E.A., Weisbach, D., 2010. Climate Change Justice. Princeton University Press, Princeton, NJ.
Shipan, C.R., 2004. Regulatory regimes, agency actions, and the conditional nature of congressional influence. American Political Science Review 98, 467–480.
Shrader-Frechette, K., 2005. Environmental Justice: Creating Equality, Reclaiming Democracy. Oxford University Press, New York and London.
Stigler, G., 1971. The theory of economic regulation. Bell Journal of Economics and Management Science 2, 3–21.
Stone, A., 1982. Regulation and its Alternatives. Congressional Quarterly Press, Washington, DC.
Sunstein, C., 1990. After the Rights Revolution: Reconceiving the Regulatory State. Harvard University Press, Cambridge and London.
Tesh, S.N., 2000. Uncertain Hazards: Environmental Activists and Scientific Proof. Cornell University Press, Ithaca and London.
Thatcher, M., 2005. Independent regulatory agencies and elected politicians in Europe. In: Gerardin, D., Munoz, R., Petit, N. (Eds.).
Tokar, B., 2010. Toward Climate Justice: Perspectives on the Climate Crisis and Social Change. Communalism Press, Porsgrunn, Norway.
Vogel, D., 2012. The Politics of Precaution: Regulating Health, Safety, and Environmental Risks in Europe and the United States. Princeton University Press, Princeton and Oxford.
Wallace, H., Wallace, W., Pollack, M.A. (Eds.), 2005. Policy-Making in the European Union, fifth ed. Oxford University Press, Oxford and New York.
Weaver, S., 1977. Decision to Prosecute: Organization and Public Policy in the Antitrust Division. The MIT Press, Cambridge, MA.
Weingast, B.R., 1984. The congressional bureaucratic system: a principal-agent perspective (with applications to the SEC). Public Choice 44, 147–191.
Weingast, B.R., Moran, M., 1983. Bureaucratic discretion of congressional control: regulatory agency policy making at the FTC. Journal of Political Economy 91, 765–800.
Wilson, J.Q., 1989. Bureaucracy: What Government Agencies Do and Why They Do It. Basic Books, New York.
Wilson, J.Q. (Ed.), 1980. The Politics of Regulation. Basic Books, New York.
Wood, B.D., 1988. Bureaucrats, principals, and responsiveness in clean air enforcements. American Political Science Review 82, 215–234.
Wood, B.D., Anderson, J.E., 1993. The politics of US antitrust regulation. American Journal of Political Science 37, 1–39.
Young, A.R., 2005. The single market: a new approach to policy. In: Wallace, H., Wallace, W., Pollack, M.A. (Eds.), Policy-Making in the European Union, fifth ed. Oxford University Press, Oxford and New York.

Rehabilitation (Cardiac): Psychosocial Aspects

Martyn Jones and Karen M Smith, School of Nursing and Midwifery, University of Dundee, Dundee, UK
Oliver Herber, Institute of General Practice, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D.W. Johnston, volume 19, pp. 12985–12988, © 2001, Elsevier Ltd.

Abstract

Cardiac rehabilitation (CR) is a clinical intervention aimed at improving the health and well-being of people with cardiac disease. CR programs combine key components of exercise, education, behavior change, and psychosocial support. An overview of the influence of major psychosocial factors on people's recovery is provided. Despite overwhelming evidence of the effectiveness of CR, uptake and adherence still remains low. A better understanding of both patterns of attendance and predictors of noncompletion and nonattendance are required to inform interventions to increase uptake of CR and maintain or improve the adherence of people with long-term cardiac conditions.

Background

Cardiovascular disease (CVD) remains the leading cause of mortality in many industrialized countries, including North America, Europe, and the UK (WHO, 2011). Improvements in acute and longer-term clinical management have increased the numbers of patients surviving cardiac events. Within the UK, an estimated 2.5 million people live with coronary heart disease (CHD) generating health care costs in the region of £1.8 billion in 2009 in the UK alone (British Heart Foundation (BHF), 2012).

Cardiac rehabilitation (CR) is a clinical and cost-effective intervention, routinely offered to patients following hospital admission with acute coronary syndrome (ACS). ACS is a term used to describe acute myocardial infarction (MI) and unstable angina. The primary aim of CR is to improve the health and well-being of people who have heart problems such as ACS to allow a return to premorbid levels of function. Behavioral change, a key element of CR, aims to promote improved diet, to reduce smoking, and to increase activity levels in order to reduce the risk of recurrent cardiac events.

Cardiac rehabilitation is defined as:

> The coordinated sum of activities required to influence favorably the underlying cause of cardiovascular disease, as well as to provide the best possible physical, metal and social conditions, so that the patients may, by their own efforts, preserve or resume optimal functioning in their community and through improved health behavior, slow or reverse progression of disease.
>
> (British Association for Cardiovascular Prevention and Rehabilitation (BACPR), 2012: p. 2)

CR may be offered in a range of forms, delivered in health care, community, or home-based settings. CR programs, once exclusively exercise based, are now more comprehensive combining the key components of exercise, education, behavior change, and other psychosocial support interventions. CR is increasingly tailored and delivered to meet individual need (Heran et al., 2011). CR forms part of an integrated cardiology service that positively influences patient participation, adherence, and long-term behavior change and is an obligatory element of any cardiovascular health care service (BACPR, 2012).

Current CR provision is delivered in the context of a dynamic clinical environment. Diagnostic definitions have changed with the use of new highly sensitive and specific biomarkers. Early interventional procedures, integral to medical management, have led to very short periods of hospitalization. Early discharge results in limited time for patients to adjust to their cardiac event. Changing demographic profiles also mean patients may be older with more complex comorbidities, therefore CR needs to be further tailored and to be sufficiently accessible to meet the needs of these individuals.

Development of Cardiac Rehabilitation

Early CR programs developed in the 1960s targeted patients following MI or cardiac surgery were mainly exercise based but have since been expanded to include cardiac patients with angina, heart failure, postintervention, and following device implantation (BACPR, 2012). Current more comprehensive menu-based programs including educational, exercise, psychosocial components as well as health-related behavior change strategies have been developed in response to an increased understanding of key issues including the underlying pathophysiology, traditional risk factors such as smoking, poor diet, physical inactivity, and more recently important psychological determinants, such as anxiety, depression, anger, and misconceptions regarding coronary heart disease.

Education is an essential component of CR, designed both to address misconceptions about ACS or CR itself (French et al., 2006) and to facilitate patients to make informed choices about their ongoing management. Education is critical to the development of improved patient understanding regarding issues surrounding risk factor reduction and is essential in improving patient reports of perceived personal control and facilitating confidence in managing their cardiac condition (BACPR, 2012).

Given the complexity of patients' needs and the potential variation in modes of delivery of CR programs there is currently no standard CR program. There is a move away from the provision of CR in distinct phases or stages to the design and delivery of individualized programs that meet patient needs from the point of the acute event and thereafter within their long-term care. While exercise continues to be a core component of CR there is significant variation in service provision. In the US, an outpatient CR program typically consists of 36 sessions, with 2–3 per week over 12–18 weeks (Oldridge, 2012). In contrast, many UK outpatient programs start around 4–6 weeks following discharge and are offered once to twice weekly over 8–12 weeks. Patient journeys through CR are increasingly individualized and based upon needs assessment, making comparisons between CR in settings and within patients difficult.

Alternative, effective models of CR have been developed that support patients in remote and rural settings, or to support those who prefer not to participate in-group sessions (Higgins et al., 2005). More recent developments include the use of cellular Web-based electronic service delivery further extending the reach of such CR services and increasingly allowing remote telemonitoring of patient symptoms and interactive support of progress (Neubeck et al., 2009; Varnfield et al., 2011). Home-based programs, often based on standardized manuals such as the Heart Manual show comparative psychological benefit compared with hospital-based programs (Lacey et al., 2004). The primary advantage of home-based programs relate to improved access. Otherwise, the effects are similar on cardiac risk factors, health-related quality of life (HRQoL), mortality, and clinical events.

Risk Factor Reduction: Life Style and Behavioral Risk

The individual and synergistic roles of cardiovascular risk factors including smoking, poor diet, increased cholesterol, obesity, and physical inactivity in the development of cardiac disease are well recognized. The prevalence of each of these risk factors can be influenced by age, sex, socioeconomic status, ethnicity, and geographical region (BHF, 2012).

Psychoeducational interventions in CR have been used to reduce conventional risk factors, many of which are related to behavior, including elevated blood pressure, raised blood cholesterol levels, overweight, obesity, and diabetes (Whalley et al., 2011). Health behavior change is central to CR, and may be facilitated by the use of motivational interviewing, relapse prevention strategies, goal setting and pacing techniques, as well as offering partner support (Moser and Dracup, 2004). CR programs develop and apply tailored lifestyle risk factor management plans following comprehensive assessment of a range of lifestyle behaviors. The positive benefits of increasing physical exercise and activity levels, in combination with a healthy, cardioprotective, Mediterranean style diet have been established in terms of weight management and avoiding obesity (NICE, 2007). Health behavior change targeting smoking cessation, alcohol reduction, and relapse prevention are now part of UK CR standards (BACPR, 2012). Taxonomies of behavior change techniques have been developed in areas of physical activity, eating, and in smoking (Michie et al., 2011).

Psychosocial Determinants and Impact on CHD

More recently, the role of psychological factors, such as anxiety, depression, and anger have been recognized as independent risk factors in the development and maintenance of chronic disease (Khayyam-Nekouei et al., 2013; Suls, 2013). Cardiac misconceptions, illness, and treatment beliefs may also influence recovery and outcome (French et al., 2006; Furze et al., 2005). The psychosocial elements of CR include the assessment and management of emotion and mood, including anxiety, depression, and anger. Specific illness perceptions, perceptions of social support, self-efficacy, and outcome expectancy beliefs regarding treatment can support behavioral change (BACPR, 2012).

Anxiety and Depression

Depressed patients are 64% more likely to develop ACS (Lett et al., 2004) and to recover less well in the short and longer term (Barth et al., 2004). Development of depression following a cardiac event has been associated with an increase in serious subsequent cardiac events (Dickens et al., 2008), more frequent hospital admissions, and an increase in all-cause morbidity and cardiac mortality (Barth et al., 2004; Connerney et al., 2010).

The presence of both anxiety and depression in cardiac patients may be the result of a reaction to hospitalization, the disease itself, or may represent an underlying chronic condition (Kornerup et al., 2011). The prevalence of anxiety in cardiac patients has been found to be as high as 32% (Kornerup et al., 2011) and significant levels of anxiety and depression may persist up to 12 months post-ACS event (Thombs et al., 2006). Anxious patients have a 36% increased risk of adverse cardiovascular outcome compared with nonanxious patients (Roest et al., 2010). However, studies comparing anxiety and clinical endpoints such as MI and cardiovascular death do not show consistent results. For example, the likelihood of an ACS patient having a further event in a 5-year follow-up is not associated with anxiety or depression measured at either baseline or 1 year (Kornerup et al., 2011).

Anger and Hostility

While studies such as the Recurrent Coronary Prevention Project (Friedman et al., 1986) show that reductions in Type A behavior (i.e., individuals that are impatient, ambitious, highly status-conscious, sensitive, rigidly organized) lead to lower rates of recurrent MI, anger, and hostility are now thought to be a more reliable predictor of CVD. Anger and related emotions are associated with increased prevalence of CVD, subclinical disease, and life-threatening arrhythmias (Suls, 2013). Patients with high hostility scores are up to 2.5 times more likely to have restenosis following angioplasty (Kuper et al., 2002). Few intervention studies exist targeting anger in cardiac patients and their results are unequivocal (Suls, 2013).

Misconceptions, Illness Beliefs, and Self-Efficacy

Cardiac misconceptions are erroneous beliefs about heart disease, often associated with maladaptive coping strategies,

which may provoke anxiety (Furze et al., 2005; Roesch and Weiner, 2001) and negatively influence a person's decision to adopt or reject lifestyle modification to reduce future coronary risk (Darr et al., 2008). Misconceptions can negatively influence treatment adherence, attendance at CR, mood, physical functioning, HRQoL, and return to work, and can lead to more admissions to hospital (Le Grande et al., 2012). A systematic review (French et al., 2001) found that participants reported stress to be the most important cause for CHD in 41% of studies, rather than classical risk factors of lifestyle behaviors, which were cited as principal source in only 31% of studies. This lack of concordance between patients' actual coronary risk factors and their causal attributions for CHD is commonly reported (Astin and Jones, 2004). Misconceptions are also reported by peers and family members (Roebuck et al., 2001) resulting in unhelpful behaviors such as over protection (Roebuck et al., 2001). Clinical staffs who hold erroneous beliefs may also offer inappropriate advice for recovery, perhaps inadvertently influencing patients to adopt avoidant behaviors and sedentary lifestyles following their acute event (Lin et al., 2008).

Patient beliefs about their illness may also influence coping, subsequent behavior, and cardiovascular outcome (Hagger and Orbell, 2003). For example, patients who believe that their illness was controllable at the time of their cardiac event report high self-efficacy in their ability to maintain exercise 3 years later (Lau-Walker, 2007). Patients with acute MI tend to perceive their condition to have less severe consequences than those with heart failure (HF). MI patients report shorter duration (timeline) and lower identity scores than HF or angina patients (Le Grande et al., 2012). This may reflect patient underestimation of the severity of their condition and the future effects of their illness. Early intervention procedures such as angioplasty, a shorter hospital stay coupled with patient anxiety, may negatively impact on information retention and patient recovery (Rushford et al., 2007).

Changes from an acute to a more chronic perception of illness in patients post-MI is likely to result in decreased personal and treatment control beliefs and is associated with increased fatigue and poorer HRQoL (Alsén et al., 2010). While multifaceted interventions combining motivational interviewing techniques with cognitive behavior therapy (CBT) may change maladaptive beliefs significantly and positively (Lau-Walker, 2007), the effects of these changes on behavioral, psychological, and functional outcomes require further investigation (Goulding et al., 2010).

Cardiac Rehabilitation: Is It Effective?

There is strong, consistent, and positive evidence from systematic reviews and meta-analyses of randomized controlled trials that comprehensive CR programs facilitate physical and psychological recovery following acute cardiac events (Oldridge, 2012).

Morbidity and Mortality

Cardiac rehabilitation is associated with a reduction of cardiac mortality by 26% and all-cause mortality by 13–26% (Heran et al., 2011; Lawler et al., 2011; Taylor et al., 2004; Oldridge, 2012). Only 86 patients need to be treated to achieve benefit in all-cause mortality, and even fewer, 37 patients, to achieve these gains in cardiac mortality (Oldridge, 2012). CR also leads to a 38% reduction in nonfatal MI (Clark et al., 2005) as well as to a 28–56% decrease in unplanned readmissions (Davies et al., 2010a), thus yielding significant cost savings. Patients who participate in CR are up to 3 times less likely to die or have a further cardiac event as compared to those who do not participate (Chow et al., 2010; Hammill et al., 2010).

Life Style and Behavioral Risk

The modification of cardiovascular risk and adherence to lifestyle change during CR, combining a healthy diet, physical exercise, weight management with the avoidance of obesity and smoking is associated with improvement in cardiovascular health and all-cause mortality (Chow et al., 2010). Attending CR also positively influences functional capacity and the development of self-management skills supporting an early return to work (Yohannes et al., 2010).

Brief interventions designed to motivate lifestyle change can also have a modest positive effect (Fernandez et al., 2007). Short-term programs supported by CBT-based interventions targeting negative emotions, feelings, and cardiac misconceptions, enhancing self-management and self-regulatory processes (i.e., goal setting) have reduced cardiovascular risk and increased physical activity at 6-month follow-up (Clark et al., 2011; Ferrier et al., 2011; Lewin et al., 2002; Zetta et al., 2011).

Such positive changes may not endure, however. In the OASIS trial, 65% of current smokers had quit at 30 days but only 25% of the quitters were still not smoking at 6 months (Chow et al., 2010). Recommended secondary prevention targets for blood pressure, waist circumference, cholesterol, or physical activity are not always achieved in the short-term, emphasizing the importance of secondary prevention initiatives (Murphy et al., 2011).

Psychosocial Risk

A recent meta-analysis has disaggregated the effects of the psychosocial element of CR (Whalley et al., 2011) and compared interventions combining psychosocial and exercise-based strategies with those based on psychosocial approaches alone. This revealed that psychosocial intervention, delivered by staff trained in psychological intervention, had only a modest effect on cardiac mortality, no reduction of risk in terms of total mortality, and no impact on revascularization rates or nonfatal infarction. This was in contrast to findings from an earlier systematic review (Welton et al., 2009), which included mixed intervention studies, for example, combining psychological intervention with exercise. Interventions based upon client-led discussion of cardiovascular risk and emotional support or including family members in the therapy were negatively associated with depression. Psychosocial intervention have little effect on HRQoL, perceived stress, Type A behavior, anger, and perceived exhaustion (Whalley et al., 2011). Psychological interventions targeting Type A behavior are most effective in treating or reducing depression.

Nonpharmacological interventions may improve symptoms of depression. For example, a 12-week depression care management program for hospitalized cardiac patients showed benefit at 6 and 12 months post-ACS event (Stafford et al., 2009). A shorter 6-week CBT program, demonstrated similar benefit in depressive symptoms (Hambridge et al., 2009); however, this brief intervention was delivered to an established, well-motivated group already taking part in CR and was delivered by skilled clinical psychologists (Hambridge et al., 2009) creating challenges in the transferability of this approach. Programs based on the principles of CBT can facilitate patient skills to manage depression and make positive lifestyle changes (Berkman et al., 2003).

Educational Intervention

Systematic reviews of CR have explored the effectiveness of the educational component of CR alone rather than as part of a CR program in combination with an exercise or psychological intervention (Brown et al., 2011). Patients/families receiving educational sessions in the form of instructional activity from a health care professional, ranging from 2 short periods of contact to a 4-week residential placement, were no different than those receiving standard care in terms of outcomes including cardiac and total mortality, revascularization, hospitalizations, and readmissions. Improvements in HRQoL and reductions in health care costs suggest that education is still a vital component of CR, but by itself there is little evidence of effect beyond these outcomes (Brown et al., 2011).

Exercise

The exercise component of CR has been compared to comprehensive CR (Heran et al., 2011). Exercise only CR is highly effective in reducing overall and cardiac mortality and reducing the number of hospital admissions. This form of CR does not reduce the risk of subsequent MI, coronary artery bypass grafting or *percutaneous transluminal coronary angioplasty*. Exercise-based CR led to an increase in HRQoL in 7/12 trials examined. This positive effect may result from a direct effect of exercise on physiological outcome, or may be mediated by improvement in the risk factors for atherosclerosis, e.g., reductions in lipids, smoking, and blood pressure. Home-based exercise, often based around increasing activities such as walking, is equally effective compared with center-based programs (Taylor et al., 2010).

Understanding and Improving Uptake and Adherence

Despite CR remaining the main stay of treatment following an acute coronary event, only 42% of those eligible actually take part (BACPR, 2012). Reasons for low participation are multifactorial. Four broad factors are associated with cardiac rehabilitation attendance. These include sociodemographic factors such as age (Artham et al., 2008), diagnosis (Turner et al., 2002), infrastructure-related issues such as transportation or referral (Glazer et al., 2002; Jackson et al., 2005), and psychological or cognitive variables associated with CR attendance and adherence including depression and anxiety (Whitmarsh et al., 2003).

While patients with elevated depression scores are as likely as nondepressed to enroll in CR (Farley et al., 2003), they are 2.2 times less likely to complete CR after controlling for age, gender, BMI, and employment status (Casey et al., 2008). Ongoing anxiety and depression can lead to reports of poorer health status (Pedersen et al., 2006) and reduced compliance with treatments, behavior, and lifestyle modification (Kronish et al., 2006).

Perceptions of high controllability, severity, and the significant consequence of ACS are particularly strong predictors of CR attendance (Taylor et al., 2011). Beliefs about the necessity of CR and concerns regarding CR are also predictive (Cooper et al., 2007). High self-efficacy has been strongly associated with cardiac rehabilitation attendance (Jackson et al., 2005). Patients receiving practical support were 3.6 times more likely to adhere to medication than those with no support and were more likely to attend CR (DiMatteo, 2004; Molloy et al., 2008). While national guidelines suggest that the involvement of a significant supporting other (e.g., friend, partner, or family member) is as important as the prescription of cardioprotective medication and behavioral and medical risk factor management (BACPR, 2012), a recent meta-analysis concluded that CR based on therapeutic approaches including families had little effect (Whalley et al., 2011).

An intervention based around the use of motivational letters using a theory of planned behavior approach improved attendance (Mosleh et al., 2009), while a liaison nurse improved attendance rates, this approach failed to deliver any reduction in cardiovascular risk factors (Davies et al., 2010b). Evidence of improved adherence to exercise within a CR program is limited. Increased adherence to unsupervised exercise has been demonstrated in patients already attending a structured CR (Duncan and Pozehl, 2002) but this may be limited to patients who are likely to attend and complete CR. A combination of action planning and coping resulted in greater exercise adherence than for patients exposed to coping, or action planning alone (Sniehotta et al., 2006).

Despite these impressive initial improvements attained while participating in CR, achieved benefits may not be maintained over time. Adherence to CR is critical since patients who do not engage or dropout prior to completion are at greater risk of experiencing further cardiac events than those who attend all recommended sessions (Taylor et al., 2011).

Conclusion

This article has presented considerable evidence of the effectiveness of CR as a psychosocial intervention. Not surprisingly, the greatest evidence of effect of CR is seen in combined comprehensive programs. Psychological risk factors, such as cardiac misconceptions, illness, and treatment representations are likely to affect uptake and adherence to CR (BACPR, 2012).

Despite this evidence of benefit, uptake, and adherence to CR remains relatively low. Attempts to understand patterns of attendance of CR are critical in building an evidence base to improve uptake and adherence. Current studies exploring these issues are often based on sound psychological models, and

attempt to measure dynamic psychosocial processes, but do so using retrospective methods of data collection prone to recall bias, and by the use of designs limited by two or three longitudinal data collection points. There is a need, therefore, for additional longitudinal studies with several data collection points to establish predictors of attendance at CR measuring motivation, mood, illness, and treatment beliefs (e.g., Herber et al., 2012). Only in this way will we understand changes in patterns of the predictors of noncompletion and nonattendance allowing the development of interventions to increase uptake and thereafter adherence.

See also: Cardiac Disease, Coping with; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Control Beliefs: Health Perspectives; Coronary Heart Disease: Psychosocial Aspects; Gender and Physical Health; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Behaviors; Health Education and Health Promotion; Health Interventions, Community-based; Health Outcomes, Assessment of; Health Psychology; Hospitals and Health: Psychosocial Aspects; Illness Behavior and Care-Seeking; Patient Adherence to Health Care Regimens; Personalized Medicine; Physical Activity and Health; Smoking and Health; Social Support and Recovery from Disease and Medical Procedures.

Bibliography

Alsén, P., Brink, E., Persson, L.-O., Brändström, Y., Karlson, B.W., 2010. Illness perceptions after myocardial infarction: relations to fatigue, emotional distress, and health-related quality of life. Journal of Cardiovascular Nursing 25 (2), E1–E10. http://dx.doi.org/10.1097/JCN.1090b1013e3181c1096dcfd.

Artham, S.M., Lavie, C.J., Milani, R.V., 2008. Cardiac rehabilitation programs markedly improve high-risk profiles in coronary patients with high psychological distress. Southern Medical Journal 101 (3), 262–267. http://dx.doi.org/10.1097/SMJ.1090b1013e318164dfa318168.

Astin, F., Jones, K., 2004. Heart disease attributions of patients prior to elective percutaneous transluminal coronary angioplasty. The Journal of Cardiovascular Nursing 19 (1), 41–47.

BACPR, 2012. Cardiovascular Disease Prevention and Rehabilitation 2012. British Cardiovascular Society, London.

Barth, J., Schumacher, M., Herrmann-Lingen, C., 2004. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosomatic Medicine 66, 802–813.

Berkman, L.F., Blumenthal, J., Burg, M., et al., 2003. Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) Randomized Trial. Journal of the American Medical Association 289, 3106–3116.

BHF, 2012. Coronary Heart Disease Statistics: A Compendium of Health Statistics, 2012 ed. British Heart Foundation, London.

Brown, J., Clark, A., Dalal, H., Welch, K., Taylor, R., 2011. Patient education in the management of coronary heart disease. Cochrane Database Systematic Review 12.

Casey, E., Hughes, J.W., Waechter, D., Josephson, R., Rosneck, J., 2008. Depression predicts failure to complete phase-II cardiac rehabilitation. Journal of Behavioural Medicine 31 (5), 421–431.

Chow, C., Jolly, S., Rao-Melacini, P., Fox, K., Anand, S., Yusuf, S., 2010. Association of diet, exercise, and smoking modification with risk of early cardiovascular events after acute coronary syndromes. Circulation 121 (6), 750–758.

Clark, A., Hartling, L., Vandermeer, B., McAlister, F., 2005. Meta-analysis: secondary prevention programs for patients with coronary artery disease. Annals of Internal Medicine 143 (9), 659–672.

Clark, M., Kelly, T., Deighan, C., 2011. A systematic review of the heart manual literature. European Journal of Cardiovascular Nursing 10 (1), 3–13.

Connerney, I., Sloan, R.P., Shapiro, P.A., Bagiella, E., Seckman, C., 2010. Depression is associated with increased mortality 10 years after coronary artery bypass surgery. Psychosomatic Medicine 72 (9), 874–881.

Cooper, A.F., Weinman, J., Hankins, M., Jackson, G., Horne, R., 2007. Assessing patients' beliefs about cardiac rehabilitation as a basis for predicting attendance after acute myocardial infarction. Heart 93 (1), 53–58.

Darr, A., Astin, F., Atkin, K., 2008. Causal attributions, lifestyle change, and coronary heart disease: Illness beliefs of patients of South Asian and European origin living in the United Kingdom. Heart & Lung: The Journal of Acute and Critical Care 37 (2), 91–104.

Davies, E., Moxham, T., Rees, K., Singh, S., Coats, A., Ebrahim, S., et al., 2010a. Exercise training for systolic heart failure: Cochrane systematic review and meta-analysis. European Journal of Heart Failure 12 (7), 706–715.

Davies, P., Taylor, F., Beswick, A., Wise, F., Moxham, T., Rees, K., et al., 2010b. Promoting patient uptake and adherence in cardiac rehabilitation. Cochrane Database of Systematic Reviews (7). http://dx.doi.org/10.1002/14651858.CD007131.pub2. Retrieved from: http://summaries.cochrane.org/CD007131/promoting-patient-uptake-and-adherence-in-cardiac-rehabilitation.

Dickens, C., McGowan, L., Percival, C., Tomenson, B., Cotter, L., Heagerty, A., et al., 2008. Negative illness perceptions are associated with new-onset depression following myocardial infarction. General Hospital Psychiatry 30 (5), 414–420.

DiMatteo, M., 2004. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychology 23, 207–218.

Duncan, K., Pozehl, B., 2002. Staying on course: the effects of an adherence facilitation intervention on home based exercise participation. Progress in Cardiovascular Diseases 17 (2), 59–65.

Farley, R., Wade, T., Birchmore, L., 2003. Factors influencing attendance at cardiac rehabilitation among coronary heart disease patients. European Journal of Cardiovascular Nursing 2 (3), 205–212.

Fernandez, R., Griffiths, R., Everett, B., Davidson, P., Salamonson, Y., Andrew, S., 2007. Effectiveness of brief structured interventions on risk factor modification for patients with coronary heart disease: a systematic review. International Journal of Evidence-Based Healthcare 5 (4), 370–405.

Ferrier, S., Blanchard, C.M., Vallis, M., Giacomantonio, N., 2011. Behavioural interventions to increase the physical activity of cardiac patients: a review. European Journal of Cardiovascular Prevention & Rehabilitation 18 (1), 15–32.

French, D.P., Cooper, A., Weinman, J., 2006. Illness perceptions predict attendance at cardiac rehabilitation following acute myocardial infarction: a systematic review with meta-analysis. Journal of Psychosomatic Research 61 (6), 757–767.

French, D.P., Senior, V., Weinman, J., Marteau, T.M., 2001. Causal attributions for heart disease: a systematic review. Psychology & Health 16 (1), 77–98.

Friedman, M., Thoresen, C.E., Gill, J.J., Ulmer, D., Powell, L.H., Price, V.A., et al., 1986. Alteration of type A behavior and its effect on cardiac recurrences in post myocardial infarction patients: summary results of the recurrent coronary prevention project. American Heart Journal 112 (4), 653–665.

Furze, G., Lewin, R.J.P., Murberg, T., Bull, P., Thompson, D.R., 2005. Does it matter what patients think? The relationship between changes in patients' beliefs about angina and their psychological and functional status. Journal of Psychosomatic Research 59 (5), 323–329.

Glazer, K., Emery, C., Frid, D., Banyasz, R., 2002. Psychological predictors of adherence and outcomes among patients in cardiac rehabilitation. Journal of Cardiopulmonary Rehabilitation 22 (1), 40–46.

Goulding, L., Furze, G., Birks, Y., 2010. Randomized controlled trials of interventions to change maladaptive illness beliefs in people with coronary heart disease: systematic review. Journal of Advanced Nursing 66 (5), 946–961.

Le Grande, M.R., Elliott, P.C., Worcester, M.U.C., Murphy, B.M., Goble, A.J., Kugathasan, V., et al., 2012. Identifying illness perception schemata and their association with depression and quality of life in cardiac patients. Psychology, Health & Medicine 17 (6), 709–722.

Hagger, M.S., Orbell, S., 2003. A meta-analytic review of the common-sense model of illness representations. Psychology and Health 18 (2), 141–184.

Hambridge, J.A., Turner, A., Baker, A.L., 2009. Braveheart begins: pilot results of group cognitive behaviour therapy for depression and anxiety in cardiac patients. Australian and New Zealand Journal of Psychiatry 43 (12), 1171–1177.

Hammill, B.G., Curtis, L.H., Schulman, K.A., Whellan, D.J., 2010. Relationship between cardiac rehabilitation and long-term risks of death and myocardial infarction among elderly medicare beneficiaries. Circulation 121 (1), 63–70.

Heran, B., Chen, J., Ebrahim, S., Moxham, T., Oldridge, N., Rees, K., et al., 2011. Exercise-based cardiac rehabilitation for coronary heart disease. Cochrane Database Systematic Review (Online) (7), CD001800.

Herber, O.R., Jones, M.C., Smith, K., Johnston, D.W., 2012. Assessing acute coronary syndrome patients' cardiac-related beliefs, motivation and mood over time to predict non-attendance at cardiac rehabilitation. Journal of Advanced Nursing 68 (12), 2778–2788.

Higgins, R.O., Murphy, B.M., Grande, M.R.L., Parkinson, A., Worcester, M.U.C., Goble, A.J., 2005. Expressed preferences for health education of patients after percutaneous coronary intervention. European Journal of Cardiovascular Prevention & Rehabilitation 12 (6), 572–579.

Jackson, L., Leclerc, J., Erskine, Y., Linden, W., 2005. Getting the most out of cardiac rehabilitation: a review of referral and adherence predictors. Heart 91 (1), 10–14.

Khayyam-Nekouei, Z., Neshatdoost, H., Yousefy, A., Sadeghi, M., Manshaee, G., 9 January 2013. Psychological factors and coronary heart disease. Advanced Research Yields in Atherosclerosis (ARYA) Atherosclerosis 9 (1), 102–111. North America. Retrieved from: http://arya.mui.ac.ir/index.php/arya/article/view/510.

Kornerup, H., Zwisler, A.-D.O., Prescott, E., 2011. No association between anxiety and depression and adverse clinical outcome among patients with cardiovascular disease: findings from the DANREHAB trial. Journal of Psychosomatic Research 71 (4), 207–214.

Kronish, I., Rieckmann, N., Halm, E.E.A., 2006. Depressed patients are less likely to follow recommended risk reducing behaviors after acute coronary syndromes. Journal of General Internal Medicine 21, 1178–1183.

Kuper, H., Marmot, M., Hemingway, H., 2002. Systematic review of prospective cohort studies of psychosocial factors in the etiology and prognosis of coronary heart disease. Seminars in Vascular Medicine 2, 267–314.

Lacey, E., Musgrave, R., Freeman, J., Tod, A., Scott, P., 2004. Psychological morbidity after myocardial infarction in an area of deprivation in the UK: evaluation of a self-help package. European Journal of Cardiovascular Nursing 3 (3), 219–224.

Lau-Walker, M., 2007. Importance of illness beliefs and self-efficacy for patients with coronary heart disease. Journal of Advanced Nursing 60 (2), 187–198.

Lawler, P., Filion, K., Eisenberg, M., 2011. Efficacy of exercise-based cardiac rehabilitation post-myocardial infarction: a systematic review and meta-analysis of randomized controlled trials. American Heart Journal 162 (4), 571–584 e572.

Lett, H.S., Blumenthal, J.A., Babyak, M.A., Sherwood, A., Strauman, T., Robins, C., et al., 2004. Depression as a risk factor for coronary artery disease: evidence, mechanisms, and treatment. Psychosomatic Medicine 66 (3), 305–315.

Lewin, R.J.P., Furze, G., Robinson, J., Griffith, K., Wiseman, S., Pye, M., et al., 2002. A randomised controlled trial of a self-management plan for patients with newly diagnosed angina. British Journal of General Practice 52, 194–201.

Lin, Y.-P., Furze, G., Spilsbury, K., Lewin, R.J.P., 2008. Cardiac misconceptions: comparisons among nurses, nursing students and people with heart disease in Taiwan. Journal of Advanced Nursing 64 (3), 251–260. http://dx.doi.org/10.1111/j.1365-2648.2008.04802.x.

Michie, S., Ashford, S., Sniehotta, F., Dombrowski, S., Bishop, A., French, D., 2011. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. Psychology and Health 26 (11), 1479–1498.

Molloy, G.J., Perkins-Porras, L., Bhattacharyya, M.R., Strike, P.C., Steptoe, A., 2008. Practical support predicts medication adherence and attendance at cardiac rehabilitation following acute coronary syndrome. Journal of Psychosomatic Research 65 (6), 581–586.

Moser, D.K., Dracup, K., 2004. Role of spousal anxiety and depression in patients' psychosocial recovery after a cardiac event. Psychosomatic Medicine 66 (4), 527–532.

Mosleh, S.M., Kiger, A., Campbell, N., 2009. Improving uptake of cardiac rehabilitation: using theoretical modelling to design an intervention. European Journal of Cardiovascular Nursing 8 (3), 161–168.

Murphy, B.M., Worcester, M.U.C., Goble, A.J., Mitchell, F., Navaratnam, H., Higgins, R.O., et al., 2011. Lifestyle and physiological risk factor profiles six weeks after an acute cardiac event: are patients achieving recommended targets for secondary prevention? Heart, Lung and Circulation 20 (7), 446–451.

Neubeck, L., Redfern, J. u., Fernandez, R., Briffa, T., Bauman, A., Freedman, S.B., 2009. Telehealth interventions for the secondary prevention of coronary heart disease: a systematic review. European Journal of Cardiovascular Prevention & Rehabilitation 16 (3), 281–289.

NICE, 2007. NICE Clinical Guideline 48. Secondary Prevention in Primary and Secondary Care for Patients Following a Myocardial Infarction. Royal College of General Practitioners, London.

Oldridge, N., 2012. Exercise-based cardiac rehabilitation in patients with coronary heart disease: meta-analysis outcomes revisited. Future Cardiology 8 (5), 729–751.

Pedersen, S.P., Denolet, J., Spindler, H., Ong, A.T.L., Serruys, P.W., Erdman, R.A.M., van Domburg, R.T., 2006. Anxiety enhances the detrimental effect of depressive symptoms on health status following percutaneous coronary intervention. Journal of Psychosomatic Research 61, 783–789.

Roebuck, A., Furze, G., Thompson, D.R., 2001. Health-related quality of life after myocardial infarction: an interview study. Journal of Advanced Nursing 34 (6), 787–794.

Roesch, S.C., Weiner, B., 2001. A meta-analytic review of coping with illness: do causal attributions matter? Journal of Psychosomatic Research 50 (4), 205–219.

Roest, A.M., Martens, E.J., de Jonge, P., Denollet, J., 2010. Anxiety and risk of incident coronary heart disease: a meta-analysis. Journal of the American College of Cardiology 56 (1), 38–46.

Rushford, N., Murphy, B.M., Worcester, M.U.C., Goble, A.J., Higgins, R.O., Grande, M.R.L., et al., 2007. Recall of information received in hospital by female cardiac patients. European Journal of Cardiovascular Prevention & Rehabilitation 14 (3), 463–469.

Sniehotta, F., Scholz, U., Schwarzer, R., 2006. Action plans and coping plans for physical exercise: a longitudinal interventional study in cardiac rehabilitation. British Journal of Health Psychology 11 (1), 23–37.

Stafford, L., Berk, M., Jackson, H.J., 2009. Are illness perceptions about coronary artery disease predictive of depression and quality of life outcomes? Journal of Psychosomatic Research 66 (3), 211–220.

Suls, J., 2013. Anger and the heart: perspectives on cardiac risk, mechanisms and interventions. Progress in Cardiovascular Diseases 55 (6), 538–547.

Taylor, G., Wilson, S., Sharp, J., 2011. Medical, psychological, and sociodemographic factors associated with adherence to cardiac rehabilitation programmes. A systematic review. Journal of Cardiovascular Nursing 26 (3), 202–209.

Taylor, R., Brown, A., Ebrahim, S., Jolliffe, J., Noorani, H., Rees, K., et al., 2004. Exercise-based rehabilitation for patients with coronary heart disease: systematic review and meta-analysis of randomized controlled trials. American Journal of Medicine 116 (10), 682–692.

Taylor, R., Dalal, H., Jolly, K., Moxham, T., Zawada, A., 20 January 2010. Home-based versus centre-based cardiac rehabilitation. Cochrane Database Systematic Review (1): CD007130. http://dx.doi.org/10.1002/14651858.CD007130.pub2.

Thombs, B.D., Bass, E.B., Ford, D.E., et al., 2006. Prevalence of depression in survivors of acute myocardial infarction. Journal of General Internal Medicine 21 (1), 30–38.

Turner, S., Bethell, H., Evans, J., Goddard, J., Mullee, M., 2002. Patient characteristics and outcomes of cardiac rehabilitation. Journal of Cardiopulmonary Rehabilitation 22 (4), 253–260.

Varnfield, M., Karunanithi, M.K., Särelä, A., Garcia, E., Fairfull, A., Oldenburg, B.F., et al., 2011. Uptake of a technology-assisted home-care cardiac rehabilitation program. Medical Journal of Australia 194 (4), 15–19.

Welton, N.J., Caldwell, D., Adamopoulos, E., Vedhara, K., 2009. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. American Journal of Epidemiology 169 (9), 1158–1165.

Whalley, B., Rees, K., Davies, P., Bennett, P., Ebrahim, S., Liu, Z., West, R., Moxham, T., Thompson, D.R., Taylor, R.S., 2011. Psychological interventions for coronary heart disease. Cochrane Database of Systematic Reviews (8).

Whitmarsh, A., Koutantji, M., Sidell, K., 2003. Illness perceptions, mood and coping in predicting attendance at cardiac rehabilitation. British Journal of Health Psychology 8 (2), 209–221.

WHO, 2011. Global Atlas on Cardiovascular Disease Prevention and Control. World Health Organization, Geneva.

Yohannes, A.M., Doherty, P., Bundy, C., Yalfani, A., 2010. The long-term benefits of cardiac rehabilitation on depression, anxiety, physical activity and quality of life. Journal of Clinical Nursing 19 (19–20), 2806–2813.

Zetta, S., Smith, K., Jones, M., Allcoat, P., Sullivan, F., 2011. Evaluating the angina plan in patients admitted to hospital with angina: a randomized controlled trial. Cardiovascular Therapeutics 29 (2), 112–124.

Rehabilitation of the Offender

Friedrich Lösel, University of Cambridge, Cambridge, UK; and University of Erlangen-Nuremberg, Erlangen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Programs of offender rehabilitation aim to reduce reoffending and promote a noncriminal lifestyle. In contrast to the 'nothing works' doctrine of the 1970s, research on 'what works' has shown a desirable mean effect of 10–20% less reoffending. Cognitive–behavioral programs, hierarchical therapeutic communities, multisystemic therapy, education, vocational training, and restorative justice measures are supported by evidence. Counseling, low-structured case work, mentoring, sports programs, electronic monitoring, and medication for specific offender groups are promising, but need more sound evaluations. Purely deterrent and punitive measures show no or even slightly negative effects. Programs that adhere to the Risk-Need-Responsivity Model reveal particularly positive outcomes. Currently, there is a trend to integrate findings from desistance research and other concepts. More attention should be paid to the many moderators of effectiveness such as characteristics of the program, institutional context, participants, and evaluation methods.

The reactions of societies to those who break criminal laws vary across cultures and times. Over the last century in many countries corporal punishment and death penalty have been replaced by fines, prison sentences, and alternative sanctions such as community service or social training courses. Incarceration rates differ substantially and are, for example, 10 times higher in the United States than it is in Scandinavia. Such figures are only loosely related to national crime rates but depend, among other reasons, on a different emphasis given to the penal aims of compensation of guilt, protection of the public, general or specific deterrence, retribution, and the topic of this article: rehabilitation.

In general the term 'rehabilitation' refers to the regeneration of a previous state. With regard to offenders it means a process of reintegration into society, although those who come from very deprived backgrounds may have never been well integrated before. Related terms are 'resocialization' and 'correctional treatment.' The latter does not imply that most offenders need to be cured of a mental health problem. The interventions may include psychotherapy or medical treatment as well as education, vocational training or supervision. In practice, rehabilitation refers to all measures and processes that aim to reduce reoffending. This aim may be achieved via pathways such as adequate accommodation, education and employment, mental and physical health, reduced alcohol and drug problems, solid financial resources, stable family relations, and noncriminal attitudes, thinking, and behavior (Home Office, 2004).

The idea of offender rehabilitation has already been emphasized in the eighteenth and nineteenth centuries by Cesare Beccaria and Franz von Liszt who pleaded for sanctions that enable offenders to desist from crime. During the 1960s and early 1970s there was not only great therapeutic optimism, but also critique from different political directions that advocated tough punishment, wanted to avoid the costs of rehabilitation, feared a forced personality change of offenders or aimed for 'just deserts' punishment. The therapeutic optimism came to a halt when Lipton et al. (1975) carried out a systematic review and Martinson's (1974) prepublication of this study led to the widely cited doctrine of 'nothing works' (for details see Cullen, 2013; Lösel et al., 1987; Palmer, 1992).

From 'Nothing Works' to 'what Works'

The main message of Lipton et al. (1975) was not a proof for 'nothing works,' but a lack of methodologically sound evaluations. Furthermore, the authors only counted significant versus nonsignificant results and did not calculate comparable effect sizes. In contrast, early quantitative meta-analyses of controlled studies showed significant desirable effects of correctional interventions (Andrews et al., 1990; Lipsey, 1992a; Lösel et al., 1987). Since then dozens of meta-analyses of hundreds of primary evaluations revealed mean positive effects of programs for general, violent, sexual, and other offender groups. Some meta-analyses are listed in **Table 1**.

The mean effects vary between *d*-coefficients of 0.20 ± 0.10. They are small but by no means trivial effects. If, for example, the recidivism rate in the control group is 55%, the respective rate in the treated group would be 45%, a reduction of 10% points or 18%. As the base rates in sexual recidivism are lower, the percentage reduction can be larger in studies of sex offender treatment.

Overall, the effects in offender treatment are smaller than in various other fields of psychosocial intervention (Lipsey and Wilson, 1993). This is insofar plausible as offender treatment is more or less coerced in nontherapeutic milieus, many offenders have accumulated problems (i.e., substance misuse, low education) and evaluations use official recidivism after several years as the main outcome criterion. Although the latter is a distal and not a very sensitive indicator of treatment impact (particularly in case of minor offences) it is most important for policy making.

Different Types of Program

Nearly all meta-analyses on offender rehabilitation have shown differences in effect size between types of programs (Lipsey and

Table 1 Average effects of various meta-analyses on offender treatment

	Number of studies	*Effect size (d)*
Programs for general (i.e., violent, acquisitive) offenders		
Andrews et al. (1990): Various program types (i/a)	154	0.20
Dowden and Andrews (1999): Programs for female offenders (i/a)	45	0.28
Landenberger and Lipsey (2005): Cognitive–behavioral programs (i/a)	58	0.24
Lipton et al. (2002a): Therapeutic communities and milieu therapy (i/a)	42	0.28
Lösel et al. (1987): Treatment in social–therapeutic prisons (i)	18	0.22
Petrosino (1997): Various program types, randomized studies (i/a)[a]	115	0.20
Redondo et al. (1999): Various program types in Europe (i/a)	32	0.24
Programs for juvenile offenders		
Gottschalk et al. (1987): Various types of programs (a)	101	0.12
Koehler et al. (2013b): Various program types in Europe (i/a)[b]	25	0.16
Latimer et al. (2003): Various program types (i/a)	176	0.18
Lipsey (1993): Various program types (i/a)	397	0.10
Lipsey and Wilson (1998): Various program types, serious and violent offenders (i/a)	200	0.12
Whitehead and Lab (1989): Various program types (i/a)	50	0.24
Programs for sexual offenders[c]		
Aos et al. (2001): Various program types (i/a)	14	0.19
Hall (1995): Various program types (i/a)	12	0.24
Hanson et al. (2002): Psychosocial interventions (i/a)	43	0.12
Lösel and Schmucker (2005): Psychosocial and biological interventions (i/a)	80	0.28
Hanson et al. (2009): Psychosocial interventions (i/a)	23	0.23
Schmucker and Lösel (2013): Psychosocial interventions (i/a)	28	0.20

Note: i: Institutional/custodial treatment, a: Ambulatory/community treatment, d: Cohen's *d*-coefficient, partially calculated from Phi coefficients or odds ratios.
[a]Effect without weighting for sample size, smaller effect when *n*-weighted (0.03).
[b]Including young adult offenders up to age 25 years.
[c]Recidivism for sexual offences only.

Cullen, 2007; Lösel, 2012a; MacKenzie, 2006). In the following, various treatment modes are briefly discussed.

Cognitive–Behavioral Treatment

Cognitive–behavioral treatment (CBT) has the broadest empirical support. Such programs are based on theories of social learning and contain sessions on self-reflection, anger management, self-control, social skills training, interpersonal problem solving, moral reasoning, and perspective taking/empathy (Andrews and Bonta, 2010; Hollin and Palmer, 2009; Lipsey et al., 2007). CBT programs are typically carried out in a group format and vary with regard to structure, content, intensity, and focus on cognitions versus behavior. They can reduce reoffending by approximately 10–30% and are suitable for general, violent, juvenile, sexual, drug-addicted, female, and other offenders. Modules on anger management, interpersonal problem solving, and cognitive skills training seem to be particularly important (Jolliffe and Farrington, 2009; Landenberger and Lipsey, 2005).

Therapeutic Communities, Milieu Therapy and Social-Therapeutic Prisons

Therapeutic communities (TCs) and milieu therapy are more comprehensive than CBT programs. TCs originally placed emphasis on democratic processes and self-governing, but many are now hierarchic and structured (Cullen et al., 1997). In custody, they contain a therapeutic social climate, relatively large numbers of therapeutic staff, intensive contact with the inmates, appropriate responsibilities, various modes of therapy, and a stepwise opening to the community. Social-therapeutic prisons in Germany incorporate elements of hierarchical TCs, but have similar organizational structures as in regular prisons (Lösel and Egg, 1997). TCs and milieu-oriented programs are appropriate for difficult groups such as personality-disordered, violent and drug-addicted offenders. Overall the mean effects are positive (Aos et al., 2001; Holloway et al., 2008; Lösel, 1995b; Lipton et al., 2002b), but smaller than those of successful CBTs.

Multisystemic Therapy and Family-Oriented Programs

Multisystemic therapy (MST) and related programs are intensive, short-term, family- and community-based interventions for juvenile offenders (Henggeler et al., 2009). MST works with not only young people and their families, but also with schools, peers, neighborhoods, communities, and health care services. The approach is structured and manual-based, but not standardized. It uses broader therapeutic principles: for example, focus on the juvenile's social context, strengths and not only deficits, present- and action-orientation, development of appropriate behavior, continuity in effort, ongoing evaluation, and accountability. The interventions depend on the respective case. Meta-analyses have shown substantial reductions in recidivism (Aos et al., 2001; Curtis et al., 2004; Schmucker and Lösel, 2013), although there are some nonsignificant findings and methodological problems (Littell, 2006).

Education and Vocational Programs

Many offenders have deficits in basic education, low vocational qualifications, and histories of unemployment. As stable integration in the labor market is important for desistance from crime (Farrall and Calverley, 2006; Laub and Sampson, 2007), basic education and vocational programs are important. However, there is not much controlled evaluation on this topic. Academic education and vocational training programs reduce recidivism (MacKenzie, 2006; Wilson et al., 2000), but the effect of prison work and noncustodial employment programs are not yet well proven (MacKenzie, 2006; Visher et al., 2006). Such measures may be successful only if there is a demand in the labor market and if the offenders acquire personal skills to retain a job. Therefore, a combination with psychosocial interventions seems to be important.

Low-Structured Therapy and Counseling

These measures are often based on psychodynamic and humanistic concepts or contain eclectic case work. They aim to realize basic therapeutic principles such as emotional acceptance, empathy, openness, mutual bonding, and support. Psychodynamic approaches use a more educative format than in the traditional 'talking cure.' Some meta-analyses did not support these types of programs (e.g., Andrews et al., 1990), but others showed more desirable effects (Lipsey and Wilson, 1998; Pearson et al., 2002; Redondo et al., 1999). The relatively small and mixed effects may partly be due to the heterogeneity and low structure that can reduce program integrity. However, the basic therapeutic philosophies are relevant in many interventions (Lipsey, 2009; Lösel and Egg, 1997).

Mentoring Programs

Counseling is also a part of mentoring programs. These involve a helping, nondeviant reference person who forms a relationship with a young offender, gives advice, and functions as a role model. Meta-analyses show small positive effects on aggression and delinquency (Tolan et al., 2008), but there is not yet sound evidence that mentoring reduces reoffending in the longer term (Jolliffe and Farrington, 2008). It seems to be most promising in the context of more comprehensive interventions and for young offenders with no protective social network.

Restorative Justice Programs

Restorative justice (RJ) aims for conflict resolution and reparation of the harm caused to victims and other stakeholders. The offenders should show insight into the harm they have done, truly regret, and be willing to take concrete steps to compensate (where appropriate). RJ programs range from victim-offender-mediation and conferencing to restitution and community services (Ahmed et al., 2001; Sherman and Strang, 2007). Evaluations of various British RJ schemes showed a reduction in the frequency of reoffending and a financial payoff, but no significant effect on offense seriousness (Shapland et al., 2008). International reviews suggest desirable effects on recidivism (Sherman and Strang, 2007), in particular for face-to-face conferences convened by professionals and involving offenders, victims, and other stakeholders. Offender–victim mediation is not equally meaningful for all offender groups, that is, personality disordered or drug-addicted offenders would need more intensive treatment. However, as successful RJ contains elements of CBT more integration of RJ and treatment could be worthwhile.

Sports and Adventure Programs

Sport may help (young) offenders to learn rules and discipline, strengthen achievement motivation, cooperate with others, become healthy, satisfy needs for thrill and action, engage in nondeviant behavior, and increase self-efficacy and self-worth. As a consequence, these experiences may contribute to rehabilitation. Although sport is popular among offenders, there is little well-controlled research on its effect on reoffending. Correlational studies on sports and delinquency showed no or even a risk relation, particularly for boxing and martial arts (Endresen and Olweus, 2005; Lösel and Bliesener, 2003). In contrast, evaluations of sports programs suggested a reduction in offending (e.g., Nichols, 2007), but most studies used pre–post-designs and not equivalent control groups. Systematic reviews of adventure, challenge, and wilderness programs for young offenders suggest a desirable effect (Wilson and Lipsey, 2000). More well-controlled evaluations are needed.

Biological Treatment

Neurobiological research on serious offenders has made substantial progress (Raine, 2013). In forensic hospitals medication plays a key role with regard to underlying mental illnesses such as depression or schizophrenia, but biological measures are also promising in other fields of offender treatment. Only a few examples can be mentioned here: Substitution programs for drug-addicted offenders worked in reducing recidivism (Koehler et al., 2013). There were also positive findings on anti-androgen treatment for subtypes of sexual offenders (Lösel and Schmucker, 2005). Selective serotonin reuptake inhibitors (SSRIs) showed desirable effects in personality-disordered clients (Bellino et al., 2011). In a large study of patients with Attention-Deficit Hyperactivity Disorder (ADHD) medication strongly reduced the risk of later criminality (Lichtenstein et al., 2012). In a randomized experiment, a diet that aimed for healthy brain functioning reduced antisocial behavior in prisoners (Gesch et al., 2002). More well-controlled and large-scale evaluations of biological interventions are needed, including data on potential negative side effects. Biological interventions would not imply that crime is no longer seen as a primarily social phenomenon, but more integration with psychosocial approaches would follow a model that is well-established in the treatment of depression.

Deterrence and Punishment

Most offender rehabilitation programs are embedded in the basically punitive context of criminal justice. Therefore, it is important whether punishment alone has a deterrent effect and reduces the offender's recidivism. Many people say 'yes,' however, already first-time imprisonment seems to have no deterrent but a slightly criminogenic effect (Durlauf and Nagin, 2011). Comprehensive research reviews revealed no difference or even somewhat higher reoffending rates in comparisons of custodial versus community sanctions or longer versus shorter imprisonment (Killias et al., 2006; Nagin, 2013; Smith et al., 2002). The findings on 'shock incarceration' or 'scared straight' programs are also negative (MacKenzie, 2006; Petrosino et al., 2003). Similarly, boot camps with tough military-like regimes showed a zero mean effect (Aos et al., 2001; MacKenzie, 2006). Only a combination with educational and treatment components may have a rehabilitative impact. Of course, as mentioned, rehabilitation is not the only aim of custody and a proportion of ex-inmates do not reoffend, for example, due to self-critical reflection or getting older. However, negative aspects of incarceration (Listwan et al., 2013) are not only relevant for 'pure punishment,' but

also for rehabilitation-oriented custody, for example, contamination effects, break-up of social relations, and stigmatization after release. Accordingly, treatment in the community seems to be more effective than in custody (Lipsey and Cullen, 2007; Schmucker and Lösel, 2013).

Intensive Supervision

Some reviews of intensive supervision, intermediate sanctions or 'smart punishment' in the community showed no or small positive effects in comparison to routine parole, probation, fines, or diversion (Aos et al., 2001, 2006; Pearson et al., 2002; Smith et al., 2002). However, criminogenic effects have also been recorded (Gendreau et al., 2000). Intensive community supervision only seems to reduce reoffending when accompanied by treatment or counseling (Aos et al., 2006). The potential of intensive supervision may be counteracted by an enhanced risk of failure due to the closer monitoring and, thus, more visible violation of licensing orders. The same applies to electronic monitoring that showed a mean zero effect in a meta-analysis (Aos et al., 2006). Although there are some promising results, more well-controlled outcome and process evaluations are needed (Renzema and Mayo-Wilson, 2005).

The RNR Model and Its Implementation in Practice

Based on the evidence from 'what works' the Risk-Need-Responsivity (RNR) model of Andrews et al. (1990) became particularly influential in practice. According to this concept appropriate treatment must match the offender's risk for reoffending (R), address his/her specific criminogenic needs (N), and use methods that fit to the individual's learning style (R). Meta-analyses showed that appropriate programs reduced reoffending by up to 60% (Andrews et al., 1990) or – more realistically – by approximately 20–30% (Andrews and Bonta, 2010). Treatment effects systematically increased with the number of principles realized (Andrews and Bonta, 2010; Hanson et al., 2009; Koehler et al., 2013b), whereas programs that met none of the three conditions even had a slightly negative effect.

The social learning theory background of the RNR model particularly supports CBT, but it is not limited to such programs. The model has been expanded by emphasizing organizational factors, staff skills, therapeutic relationships, offenders' strengths, and linkages to the community (Andrews and Bonta, 2010; Andrews et al., 2011; Lösel, 1995a). Such broader perspectives go beyond the content of single programs. In various countries they were transformed into accreditation criteria. For example, the Correctional Services Advisory and Accreditation Panel of England and Wales applies the following criteria:

1. clear model of change;
2. thorough selection of offenders;
3. targeting a range of dynamic risk factors;
4. effective learning and teaching methods;
5. skill-orientation;
6. adequate sequence and duration;
7. promotion of offender motivation;
8. continuity of services;
9. ensuring program integrity;
10. ongoing evaluation.

Within a decade many programs for various offender groups have been accredited. Measures of systematic offender assessment, staff training, supervision, and auditing (mainly in prisons) have been implemented to ensure quality. The National Offender Management Service (NOMS) has been established for more integration of prison and probation services. The fast roll-out of offending behavior programs inevitably led to some implementation problems and, in particular, to deficits in controlled evaluation (Maguire et al., 2010). Quasi-experimental studies showed some encouraging findings, but also mixed results (Harper and Chitty, 2005; Hollin, 2008). However, in a large study of more than 17 000 participants of the custodial Enhanced Thinking Skills program there was a reduction in reoffending by 12% in comparison to a national cohort of offenders matched for risk and sentence length (Travers et al., 2013b). The overall decrease in reoffending in England and Wales during the last decade (Ministry of Justice (2013)) may partially be attributed to the large-scale implementation of rehabilitation programs; however, this is only indirect and not robust evidence.

The Good Lives and Desistance Concepts

Not surprisingly, there are also critical discussions of the evidence-oriented 'what works' approach (Brayford et al., 2010; McNeill, 2006; Ward and Maruna, 2007). These authors note a one-sided focus on standardized CBT and RNR programs, a danger of reduced creativity in practice, too demanding procedures of accreditation and audit, and not enough attention to systems and institutional issues. Particularly influential arguments came from Ward and colleagues (Ward and Maruna, 2007). Although they acknowledged the evidence for 'what works,' they felt that such measures do not inform offenders about the rewards of desisting from crime, neglect the role of the offender's self-directed actions, ignore the individual's need for specific experiences, use a psychometric model, do not address the role of treatment alliances, and often contain a 'one-size fits all' approach (Ward and Brown, 2004). As an alternative, Ward proposed the Good Lives Model (GLM). This integrates ideas from positive psychology, humanistic therapy, social learning, and strain theory. It assumes that offenders – like all human beings – want to achieve legitimate goals. This requires competences and resources to reach basic goods, that is, healthy functioning, knowledge, excellence in work and play, autonomy and self-directedness, inner peace and freedom of stress, friendship and community, spirituality, happiness, and creativity. Rehabilitative approaches should be designed to help the offender reach such good goals.

Although the GLM is not explicitly related to resilience research it is in accordance with its empirical findings on protective factors (Lösel and Bender, 2003; Rutter, 2012).

However, the above arguments against the 'what works' approach contain some stereotypes of the broad range of evidence-oriented programs (Lösel, 2012a). Expansions of RNR have already addressed various critical issues (Andrews et al., 2011), although therapeutic relationships and individualization indeed require more attention. Instead of conceptual polarization it needs to be seen whether the GLM achieves a robust empirical basis.

Another important widening of the 'what works' perspective comes from research on criminal careers and desistance. In a 'desistance paradigm' (McNeill, 2006) the focus switches from individual-oriented programs to social institutions and protective factors in the offenders' natural development. For example, marriage to a nondeviant partner, stable employment, social support from a prosocial network, active coping processes, experiences of self-efficacy, realistic future planning, and other social and personal resources can contribute to turning points in life and desistance from crime (Farrall and Calverley, 2006; Laub and Sampson, 2007; Lösel and Farrington, 2012; Maruna, 2001; Theobald and Farrington, 2009). The focus on desistance is not in conflict with the 'what works' approach. Although more weight is given to community services (including probation), social resources and individual strengths, programs that aim for reducing reoffending can easily be seen as measures that may help to start or continue processes of desistance in the natural context (Lösel, 2012a; Porporino, 2010). Correctional treatment programs also remain important to cope with internal or external obstacles to desistance (Bottoms and Shapland, 2011) and help to leave a criminal pathway as early and as long as possible.

'What works' and desistance research should become more integrated to benefit from each other and improve correctional practice. For example, evaluations of treatment programs can apply sound (quasi-) experimental designs that provide stronger evidence for causal processes than the correlational designs on protective factors (Lösel and Farrington, 2012). When the latter are replicated they need to be targeted in rehabilitation programs. MST, TCs, or social therapy are examples in which this is already taken into account. Another example is the inclusion of motivational modules and cognitive interviewing in CBT and similar programs (McMurran, 2002). A third example is measures that can help to strengthen protective family bonds during resettlement (Lösel et al., 2012).

Multiple Influences in Offender Rehabilitation

The above discussion shows that the focus on isolated rehabilitation programs should be widened to broader characteristics of 'what works.' This is also suggested by substantial outcome variation of different implementations of the same or rather similar programs (Lipsey and Wilson, 1998; Schmucker and Lösel, 2013; Tong and Farrington, 2006). Nearly all meta-analyses show that the program content explains only a part of the variance in effect size and many other factors can have an impact. **Figure 1** contains a model that summarizes some potentially relevant factors.

Program Factors

Program Content

This type of program is relevant for its effectiveness, but most studies refer to the concept and not to its factual implementation (Lösel, 1995a; Shaffer and Pratt, 2009). Therefore, it is important to assess what was *de facto* implemented as part of the descriptive validity of evaluations (Lösel and Köferl, 1989). Better the documentation, larger the effects (Schmucker and Lösel, 2013).

Quality of Program Implementation

Even the best rehabilitation programs may show no effect when they are not properly delivered. Accordingly, effects are stronger when there is sound integrity in implementation (Andrews and Dowden, 2006; Lipsey et al., 2007). Careful staff selection, training and supervision are only some of the factors that enhance Integrity.

Program Intensity

The RNR model suggests that the intensity of an intervention should match the offenders' risk of recidivism. Accordingly, more intensive programs (longer duration, more sessions) tend to show better effects in groups at higher risk (Latimer et al., 2003; Lipsey et al., 2007).

Individualization

As mentioned above a fully standardized approach may not be optimal because one size does not fit all. Although there is a lack of research on this issue, some degree of individualization seems to provide better effects in sexual offenders (Schmucker and Lösel, 2103). The positive findings on MST point in the same direction.

Control Condition

The effects in evaluations also depend on the conditions in the untreated control groups (Lösel and Egg, 1997). As routine reactions in the criminal justice system have improved in many countries, this may lead to reduced effect sizes in more recent evaluations (Jolliffe and Farrington, 2009; Lösel and Schmucker, 2005; Tong and Farrington, 2006).

Context Factors

Custody versus Community

Various meta-analyses showed larger effects in ambulatory than in custodial treatment (Andrews and Bonta, 2010; Koehler et al., 2013b; Lipsey and Cullen, 2007; Schmucker and Lösel, 2013). This may partially be due to the negative influences of incarceration. However, basic difficulties of knowledge transfer from a closed institution to the community and resettlement problems also play a role. Therefore, measures to support the resettlement process are highly relevant (Maguire and Raynor, 2006; Taxman and Belenko, 2012).

Institutional Climate

A positive organizational climate can promote behavior in prisons. This is not only relevant for the amount of conflicts,

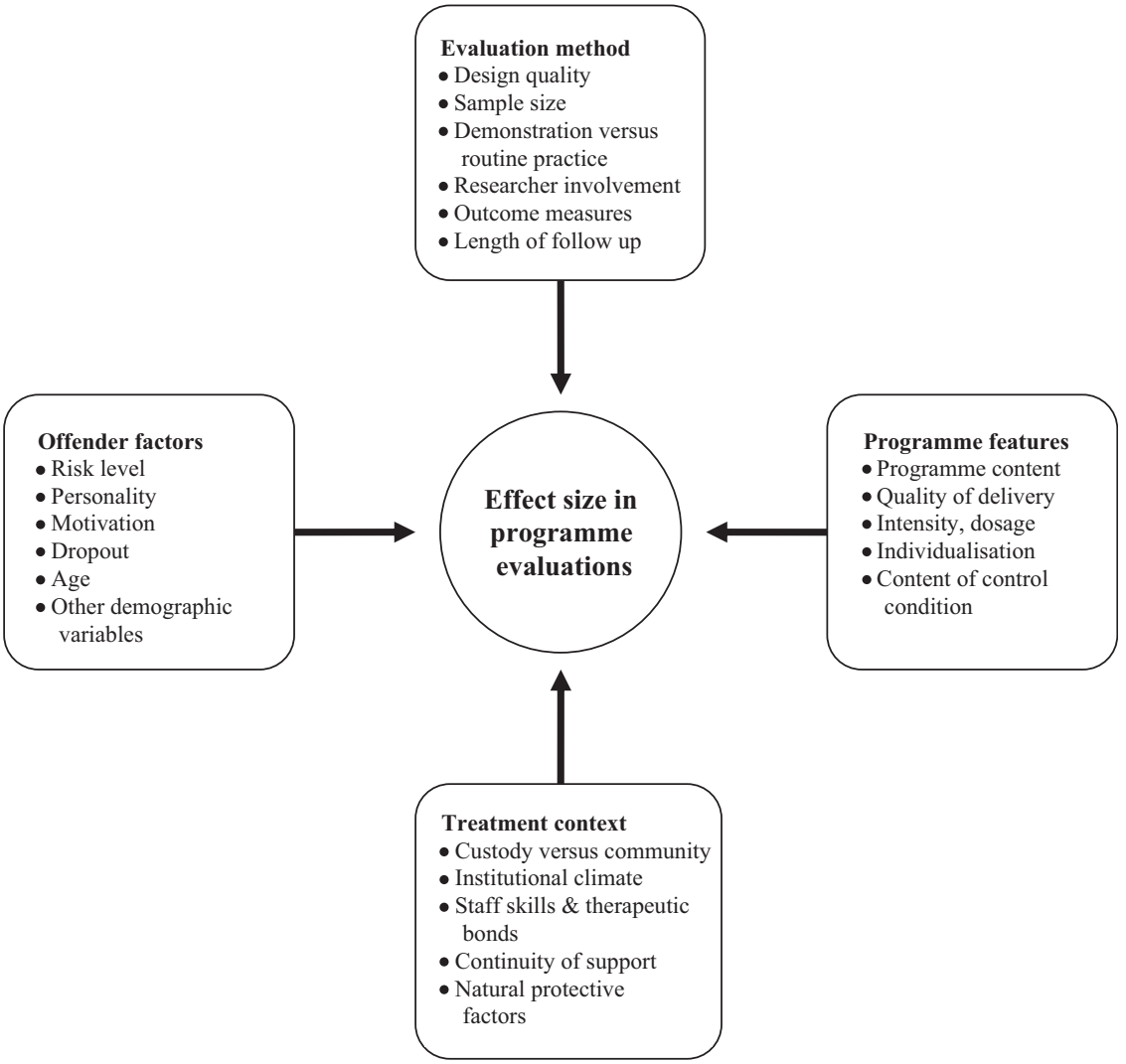

Figure 1 A model of potential influences on the effects of correctional treatment. From Lösel, F., 2012b. What works in correctional treatment and rehabilitation for young adults? In: Lösel, F., Bottoms, A.E., Farrington, D.P. (Eds.), Young Adult Offenders: Lost in Transition? Routledge, Milton Park, UK, pp. 74–112.

violence, or suicides (Cooke et al., 2008; Liebling, 2005), but may also influence the impact of treatment programs.

Staff Factors

Staff–client alliances are highly relevant in psychotherapy. Accordingly, correctional treatment is more effective when staff have good relations with the clients, adequate reinforcement practices, and social competences (Dowden and Andrews, 2004). Therefore, staff selection, training, and supervision are essential for effective rehabilitation (Gendreau et al., 1999).

Continuity of Care

Custodial programs are more effective when there is adequate care after release (Farrall and Calverley, 2006; Maguire and Raynor, 2006). Similarly, measures of vocational training, employment, and accommodation should be related to psychosocial interventions to improve their impact.

'Natural' Protective Factors

'Natural,' personal, and social resources are relevant for desistance from crime (Farrall and Calverley, 2006; Lösel and Farrington, 2012; Maruna, 2001). Such resources are targeted in MST and Functional Family Therapy; however they should also be addressed in more circumscribed structured programs.

Offender Factors

Risk Level

Treatment effects are larger in medium and high risk offenders than in low risk groups who do not need a program to desist (Andrews and Bonta, 2010; Koehler et al., 2013b; Lipsey, 2009; Lipsey and Wilson, 1998; Lösel and Schmucker, 2005). However, very high risk offenders such as psychopaths or professional criminals are less responsive to treatment so that an inverse U-shaped relation between risk level and effect size seems to be plausible (Lösel, 1996).

Personality Factors

Psychopathic offenders are particularly difficult to manage (Lösel, 1998; Salekin et al., 2010). Similarly, types of sexual offenders, for example, with or without a deviant sexual preference (paraphilia) or an antisocial personality disorder have to be taken into account. Personality factors and crime-related attitudes are also likely to play a role when prisoners with burglary or robbery index offences do not benefit from CBT (Travers et al., 2013a,b).

Treatment Motivation

The motivation to change is now seen as a dynamic process and not as dichotomous static characteristic (McMurran, 2002; Porporino, 2010). Therefore, the implementation of motivational interviewing (Miller and Rollnick, 2002) and other strategies to cope with temporary lapse and relapse are important to enhance treatment effectiveness. Accordingly, mandatory treatment does not seem to be more effective than is voluntary participation (Koehler et al., 2013b; Schmucker and Lösel, 2013).

Treatment Drop Out

Offenders who do not regularly terminate a program often have relatively high reoffending rates (Koehler et al., 2013b; Lipsey and Cullen, 2007; Travers et al., 2013a,b). In addition to personal problems this can be due to an inadequate selection process or difficulties in program implementation.

Age of Offenders

According to the age–crime curve young offenders have a higher base rate of reoffending. However, there is a lack of direct comparisons in the effects of treatment programs at different ages. Overall, there seems to be a tendency of larger effects in studies of young offenders (Lipsey and Cullen, 2007; Redondo et al., 2002).

Other Biographical Factors

Most research on offender rehabilitation has been carried out in males. Basic principles and effects are similar in female offenders (Andrews and Bonta, 2010; Dowden and Andrews, 1999; Landenberger and Lipsey, 2005), although sexual abuse and intimate relationships are more important for women's delinquency. Ethnic minorities in the United States seem to benefit from the same programs as others (Landenberger and Lipsey, 2005; Wilson et al., 2003). More research on such issues in other cultures is necessary.

Methodological Factors

Design Quality

Contrary to other fields of criminology, it is not yet clear whether randomized experiments on offender rehabilitation reveal smaller effects than do weaker studies (Jolliffe and Farrington, 2009; Koehler et al., 2013b; Lipsey and Cullen, 2007; Lipsey and Wilson, 1998; Schmucker and Lösel, 2013). As evaluation quality has various aspects (Lösel, 2007) other moderators may interfere.

Sample Size

Studies of larger samples often show larger effects than small scale evaluations (Koehler et al., 2013b; Lipsey et al., 2007; Lipsey and Wilson, 1998; Schmucker and Lösel, 2013). This can be due to a 'publication bias' that favors significant effects. However, smaller studies enable a better monitored implementation that can lead to stronger effects.

Demonstration Studies versus Routine Practice

There is a tendency of larger effects in demonstration or model projects than in evaluations of routine practice (Andrews and Bonta, 2010; Koehler et al., 2013b; Lipsey and Landenberger, 2006; Lipsey and Wilson, 1998; Lösel and Schmucker, 2005). Therefore, one has to distinguish between program 'efficacy' and 'effectiveness.' Better monitoring and more selectivity are some of the factors that may contribute to larger efficacy than effectiveness.

Researcher Involvement

Researcher involvement in the development or implementation of a program often goes along with a tendency of larger effects. (z.B. Lipsey and Landenberger, 2006; Lipsey and Wilson, 1998; Lösel and Schmucker, 2005). This may be due to better quality management, but there is also a risk of conflict of interest (Petrosino and Soydan, 2005). Therefore, independent evaluations are necessary.

Outcome Measures

Official recidivism as an outcome measure contains various problems (Maguire, 2012), but it is the key criterion for policy and practice. Some studies showed stronger treatment effects in more serious offences, frequency of reoffending, or self-reported crime (Ministry of Justice (2013); Lösel and Schmucker, 2005). Personality questionnaires also indicate larger effects, but due to social desirability and other influences these are only weakly related to reoffending (Lipsey, 1992b; McDougall et al., 2009).

Length of Follow-Up

Many Anglo-American studies use rather short follow-up periods of 1–2 years. However, there is a trend of smaller effects in evaluations with a longer follow up (Lösel, 1995a; Lösel and Schmucker, 2013). This may indicate a decrease in program impact as well as deficits in aftercare and stressors in the community.

Conclusions and Perspectives

The 'what works' approach in offender rehabilitation has made substantial progress. Carefully implemented evidence-based programs such as cognitive–behavioral treatment, structured therapeutic communities and social therapy, multisystemic therapy, basic education, vocational training, restorative justice, and other approaches can reduce reoffending, whereas pure deterrence and punishment seem to be not effective. However, there are many characteristics of program, context, participants, and evaluation methods that have an impact on effectiveness. This complexity should be taken more into account to strengthen the 'what

works' in research and practice (Lösel, 2012a). Inter alia, this requires the following:

- differentiation of the theoretical bases of rehabilitation programs;
- more controlled evaluations of daily routine (instead of demonstration projects);
- increased process evaluation on program implementation and quality assurance;
- more systems orientation instead of isolated single programs;
- analysis of interactions between specific programs and the institutional context;
- increased awareness of interpersonal relationships and therapeutic alliances;
- more differentiation and individualization (instead of 'one size fits all');
- integration of natural protective factors in treatment concepts;
- more attention to neurobiologial factors that contribute to criminal behavior;
- more frequent direct comparisons of measures in community versus custody;
- increased sensitivity to issues of cost–utility;
- more sensitivity to national and cultural differences;
- more integration of developmental prevention and offender treatment.

If such strategies are followed step by step one can expect that a future-oriented and effective rehabilitation policy will make further progress.

See also: Community Supervision and Diversion in the United States; Deterrence; Evidence-Based Psychotherapy; Prisons and Imprisonment; Psychological Treatment, Effectiveness of; Recidivism; Repetitive Sex Offenders, Treatment of; Supervision, Abusive; Systematic Reviewing and Meta-Analysis.

Bibliography

Ahmed, E., Harris, N., Braithwaite, J., Braithwaite, V. (Eds.), 2001. Shame Management through Integration. Cambridge University Press, Cambridge.

Andrews, D.A., Bonta, J., 2010. The Psychology of Criminal Conduct, fifth ed. Anderson, Cincinatti, OH.

Andrews, D.A., Bonta, J., Wormith, S., 2011. The risk-need-responsivity (RNR) model: does adding the good lives model contribute to effective crime prevention? Criminal Justice and Behavior 38, 735–755.

Andrews, D.A., Dowden, C., 2006. Risk principle of case classification in correctional treatment: a meta-analytic investigation. International Journal of Offender Therapy and Comparative Criminology 50, 88–100.

Andrews, D.A., Zinger, I., Hoge, R.D., Bonta, J., Gendreau, P., Cullen, F.T., 1990. Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. Criminology 28, 369–404.

Aos, S., Phipps, P., Barnoski, R., Lieb, R., 2001. The Comparative Costs and Benefits of Programs to Reduce Crime. Washington State Institute of Public Policy, Olympia, WA.

Aos, S., Miller, M., Drake, E., 2006. Evidence-Based Public Policy Options to Reduce Future Prison Construction, Criminal Justice Costs, and Crime Rates. Washington State Institute of Public Policy, Olympia, WA.

Bellino, S., Bozzatello, P., Rinaldi, C., Bogetto, F., 2011. Paliperidone ER in the treatment of borderline personality disorder: a pilot study of efficacy and tolerability. Depression Research and Treatment 2011, 7. http://dx.doi.org/10.1155/2011/680194.

Bottoms, A., Shapland, J., 2011. Steps towards desistance among young male adult offenders. In: Farrall, S., Hough, M., Maruna, S., Sparks, R. (Eds.), Escape Routes: Contemporary Perspectives on Life after Punishment. Routledge, Milton Park, UK.

Brayford, J., Cowe, F., Deering, J. (Eds.), 2010. What Else Works? Creative Work with Offenders. Willan, Cullompton, UK.

Cooke, D.J., Wozniak, E., Johnstone, L., 2008. Casting light on prison violence in Scotland: evaluating the impact of situational risk factors. Criminal Justice and Behavior 35, 1065–1078.

Cullen, F.T., 2013. Rehabilitation: beyond nothing works. In: Tonry, M. (Ed.), Crime and justice in America, 1975–2025. Crime and Justice: A Review of Research, vol. 42. The University of Chicago Press, Chicago, pp. 299–376.

Cullen, E., Jones, L., Woodward, R. (Eds.), 1997. Therapeutic Communities for Offenders. Wiley, Chichester, UK.

Curtis, N.M., Ronan, K.R., Borduin, C.M., 2004. Multisystematic treatment: a meta-analysis of outcome studies. Journal of Family Psychology 18, 411–419.

Dowden, C., Andrews, D.A., 1999. What works for female offenders: a meta-analytic review. Crime and Delinquency 45, 438–452.

Dowden, C., Andrews, D.A., 2004. The importance of staff practices in delivering effective correctional treatment: a meta-analysis of core correctional practices. International Journal of Offender Therapy and Comparative Criminology 48, 203–214.

Durlauf, S.N., Nagin, D.S., 2011. Imprisonment and crime: can both be reduced? Criminology & Public Policy 10, 13–54.

Endresen, I.M., Olweus, D., 2005. Participation in power sports and antisocial involvement in preadolescent and adolescent boys. Journal of Child Psychology and Psychiatry 46, 468–478.

Farrall, S., 2002. Rethinking what Works with Offenders: Probation, Social Context and Desistance from Crime. Willan, Cullompton, UK.

Farrall, S., Calverley, A., 2006. Understanding Desistance from Crime: Theoretical Directions in Resettlement and Rehabilitation. Open University Press, Maidenhead, UK.

Gendreau, P., Goggin, C., Cullen, F.T., Andrews, D.A., 2000. Does "getting tough" with offenders work? The effects of community sanctions and incarceration. Forum on Corrections Research 12, 10–13.

Gendreau, P., Goggin, C., Smith, P., 1999. The forgotten issue in effective correctional treatment: program implementation. International Journal of Offender Therapy and Comparative Criminology 43, 180–187.

Gesch, C.B., Hammond, S.M., Hampson, S.E., Eves, A., Crowder, M.J., 2002. Influence of supplementary vitamins, minerals and essential fatty acids on the antisocial behaviour of young adult prisoners. British Journal of Psychiatry 181, 22–28.

Gottschalk, R., Davidson II, W.S., Gensheimer, L.K., Mayer, J.P., 1987. Community-based interventions. In: Quay, H.C. (Ed.), Handbook of Juvenile Delinquency. wiley, New York, pp. 266–289.

Hall, G.C.N., 1995. Sexual offender recidivism revisited: a meta-analysis of recent treatment studies. Journal of Consulting and Clinical Psychology 63, 802–809.

Hanson, K., Burgon, G., Helmus, L., Hodgson, S., 2009. The principles of effective correctional treatment also apply to sexual offenders: a meta-analysis. Criminal Justice and Behavior 36, 865–891.

Hanson, R.K., Gordon, A., Harris, A.J.R., Marques, J.K., Murphy, W., Quinsey, V.L., Seto, M.C., 2002. First report of the collaborative outcome data project on the effectiveness of psychological treatment for sex offenders. Sexual Abuse: A Journal of Research and Treatment 14, 169–194.

Harper, G., Chitty, C. (Eds.), 2005. The Impact of Corrections on Re-offending: A Review of 'What Works'. Home Office Research Study 291. Home Office, London.

Henggeler, S.W., Schoenwald, S.K., Borduin, C.M., Rowland, M.D., Cunningham, P.B., 2009. Multisystemic Treatment of Antisocial Behavior in Children and Adolescents, second ed. Guilford Press, New York.

Hollin, C.R., 2008. Evaluating offending behaviour programmes: does only randomization glister? Criminology and Criminal Justice 8, 89–106.

Hollin, C., Palmer, E.J., 2009. Cognitive skills programmes for offenders. Psychology Crime and Law 15, 147–164.

Holloway, K., Bennett, T.H., Farrington, D.P., 2008. Effectiveness of Treatment in Reducing Drug-related Crime. Swedish National Council for Crime Prevention, Stockholm.

Home Office, 2004. Reducing Reoffending: National Action Plan. Home Office, London.

Jolliffe, D., Farrington, D.P., 2008. The Influence of Mentoring on Reoffending. Swedish National Council for Crime Prevention, Stockholm.

Jolliffe, D., Farrington, D.P., 2009. Effectiveness of Interventions with Adult Male Violent Offenders. Swedish National Council for Crime Prevention, Stockholm.

Killias, M., Villettaz, P., Zoder, I., 2006. The Effects of Custodial versus Non-custodial Sentences on Reoffending: A Systematic Review of the State of Knowledge. www.campbellcollaboration.org/reviews_crime_justice/index.php.

Koehler, J.A., Humphreys, D.K., Akoensi, T.D., Sánchez de Ribera, O., Lösel, F.A., 2013. Systematic review and meta-analysis on the effects of European drug treatment programmes. Psychology, Crime and Law. http://dx.doi.org/10.1080/1068316X.2013.804921.

Koehler, J.A., Lösel, F., Humphreys, D.K., Akoensi, T.D., 2013b. A systematic review and meta-analysis on the effects of young offender treatment programs in Europe. Journal of Experimental Criminology 9, 19–43.

Landenberger, N.A., Lipsey, M.W., 2005. The positive effects of cognitive-behavioral programs for offenders: a meta-analysis of factors associated with effective treatment. Journal of Experimental Criminology 1, 451–476.

Latimer, J., Dowden, C., Morton-Bourgon, K.E., 2003. Treating Youth in Conflict with the Law: A New Meta-analysis. Research Report 03YJ-3e. Department of Justice, Ottawa.

Laub, J.H., Sampson, R.J., 2007. Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Harvard University Press, Cambridge, MA.

Lichtenstein, P., Halldner, L., Zetterqvist, J., Sjölander, A., Serlachius, E., Fazel, S., Långström, N., Larsson, H., 2012. Medication for attention deficit–hyperactivity disorder and criminality. New England Journal of Medicine 367, 2006–2014.

Liebling, A., 2005. Prisons and Their Moral Performance: A Study of Values, Quality, and Prison Life. Oxford University Press, Oxford.

Lipsey, M.W., 1992a. Juvenile delinquency treatment: a meta-analytic inquiry into variability of effects. In: Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.L., Louis, T.A., Mosteller, F. (Eds.), Meta-analysis for Explanation. Russell Sage Foundation, New York, pp. 83–127.

Lipsey, M.W., 1992b. The effect of treatment on juvenile delinquents: results from meta-analysis. In: Lösel, F., Bender, D., Bliesener, T. (Eds.), Psychology and Law: International Perspectives. de Gruyter, Berlin, pp. 131–143.

Lipsey, M.W., 2009. The primary factors that characterize effective interventions with juvenile offenders. Victims and Offenders 4, 124–147.

Lipsey, M.W., Cullen, F.T., 2007. The effectiveness of correctional rehabilitation: a review of systematic reviews. Annual Review of Law and Social Science 3, 297–320.

Lipsey, M., Landenberger, N., 2006. Cognitive-behavioral interventions. In: Welsh, B., Farrington, D. (Eds.), Preventing crime: What Works for Children, Offender, Victims, and Places. Springer, Great Britain.

Lipsey, M.W., Landenberger, N.A., Wilson, S., 2007. Effects of Cognitive-behavioral Programs for Criminal Offenders. Systematic Review for the Campbell Collaboration. www.campbellcollaboration.org/reviews_crime_justice/index.php.

Lipsey, M.W., Wilson, D.B., 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. American Psychologist 48, 1181–1209.

Lipsey, M.W., Wilson, D.B., 1998. Effective intervention for serious juvenile offenders. In: Loeber, R., Farrington, D.P. (Eds.), Serious and Violent Juvenile Offenders. Sage, Thousand Oaks, CA, pp. 313–345.

Lipton, D.S., Martinson, R., Wilks, J., 1975. The Effectiveness of Correctional Treatment. Praeger, New York.

Lipton, D.S., Pearson, F.S., Cleland, C.M., Yee, D., 2002a. The effects of therapeutic communities and milieu therapy on recidivism. In: McGuire, J. (Ed.), Offender Rehabilitation and Treatment. Wiley, Chichester, UK, pp. 39–77.

Lipton, D.S., Pearson, F.S., Cleland, C.M., Yee, D., 2002b. The effects of cognitive-behavioural treatment methods on recidivism. In: McGuire, J. (Ed.), Offender Rehabilitation and Treatment. Wiley, Chichester, UK, pp. 79–112.

Listwan, S.J., Sullivan, C.J., Agnew, R., Cullen, F.T., Colvin, M., 2013. The pains of imprisonment revisited: the impact of strain on inmate recidivism. Justice Quarterly 30, 144–168.

Littell, J.H., 2006. The case for multisystemic therapy: evidence or orthodoxy? Children and Youth Services Review 28, 458–472.

Lösel, F., 1995a. The efficacy of correctional treatment: a review and synthesis of meta-evaluations. In: McGuire, J. (Ed.), What Works: Reducing Reoffending. Wiley, Chichester, UK, pp. 79–111.

Lösel, F., 1995b. Increasing consensus in the evaluation of offender rehabilitation? Lessons from research syntheses. Psychology, Crime and Law 2, 19–39.

Lösel, F., 1996. Changing patterns in the use of prisons: an evidence-based perspective. European Journal on Criminal Policy and Research 4, 108–127.

Lösel, F., 1998. Treatment and management of psychopaths. In: Cooke, D.J., Forth, A.E., Hare, R.B. (Eds.), Psychopathy: Theory, Research and Implications for Society. Kluwer, Dordrecht, pp. 303–354.

Lösel, F., 2007. Counterblast: the prison overcrowding Crisis and some constructive perspectives for crime policy. The Howard Journal of Criminal Justice 46, 512–519. http://dx.doi.org/10.1111/j.1468-2311.2007.00497.x.

Lösel, F., 2012a. Offender treatment and rehabilitation: what works? In: Maguire, M., Morgan, R., Reiner, R. (Eds.), The Oxford Handbook of Criminology, fifth ed. Oxford University Press, Oxford, UK, pp. 986–1016.

Lösel, F., 2012b. What works in correctional treatment and rehabilitation for young adults? In: Lösel, F., Bottoms, A.E., Farrington, D.P. (Eds.), Young Adult Offenders: Lost in Transition? Routledge, Milton Park, UK, pp. 74–112.

Lösel, F., Bender, D., 2003. Protective factors and resilience. In: Farrington, D.P., Coid, J. (Eds.), Prevention of Adult Antisocial Behaviour. Cambridge University Press, Cambridge, UK, pp. 130–204.

Lösel, F., Bliesener, T., 2003. Aggression und Delinquenzunter Jugendlichen. Untersuchungen von kognitiven und sozialenBedingungen [Aggression and delinquency among adolescents: A study on cognitive and social origins]. Luchterhand, Neuwied, Germany.

Lösel, F., Egg, R., 1997. Social-therapeutic institutions in Germany: description and evaluation. In: Cullen, E., Jones, L., Woodward, R. (Eds.), Therapeutic Communities in Prisons. Wiley, Chichester, UK, pp. 181–203.

Lösel, F., Farrington, D.P., 2012. Direct protective and buffering protective factors in the development of youth violence. American Journal of Preventive Medicine 43 (2S1), 8–23.

Lösel, F., Köferl, P., 1989. Evaluation research on correctional treatment in West Germany: a meta-analysis. In: Wegener, H., Lösel, F., Haisch, J. (Eds.), Criminal Behavior and the Justice System. Springer, New York, pp. 334–355.

Lösel, F., Köferl, P., Weber, F., 1987. Meta-Evaluation der Sozialtherapie (Meta-evaluation of social-therapeutic prisons). Enke Verlag, Stuttgart.

Lösel, F., Pugh, G., Markson, L., Souza, K., Lanskey, C., 2012. Risk and Protective Factors in the Resettlement of Imprisoned Fathers with Their Families. Research Report. Ormiston Children and Families Trust, Norwich, UK.

Lösel, F., Schmucker, M., 2005. The effectiveness of treatment for sexual offenders: a comprehensive meta-analysis. Journal of Experimental Criminology 1, 117–146.

Lösel, F., Schmucker, M., 2013. Treatment of sex offenders. In: Weisburd, D., Bruinsma, G. (Eds.), Encyclopedia of Criminology and Criminal Justice. Springer, New York.

MacKenzie, D.L., 2006. What Works in Corrections? Reducing the Criminal Activities of Offenders and Delinquents. Cambridge University Press, Cambridge, UK.

Maguire, M., Grubin, D., Lösel, F., Raynor, P., 2010. 'What works' and the correctional services accreditation panel: taking stock from an inside perspective. Criminology and Criminal Justice 10, 37–58.

Maguire, M., Morgan, R., Reiner, R. (Eds.), 2012. The Oxford Handbook of Criminology, fifth ed. Oxford University Press, Oxford, UK, pp. 986–1016.

Maguire, M., Raynor, P., 2006. How the resettlement of prisoners promotes desistance from crime: or does it? Criminology and Criminal Justice 6, 19–38.

Martinson, R., 1974. What works? Questions and answers about prison reform. The Public Interest 35, 22–54.

Maruna, S., 2001. Making Good: How Ex-convicts Reform and Rebuild Their Lives. American Psychological Association, Washington, DC.

McDougall, C., Perry, A.E., Clarbour, J., Bowles, R., Worthy, G., 2009. Evaluation of HM Prison Service Enhanced Thinking Skills Programme. In: Ministry of Justice Research Series 3/09. Ministry of Justice, London.

McGuire, J., 2002. Integrating findings from research reviews. In: McGuire, J. (Ed.), Offender Rehabilitation and Treatment. Wiley, Chichester, UK, pp. 4–38.

McMurran, M., 2002. Motivation to change: selection criterion or treatment need? In: McMurran, M. (Ed.), Motivating Offenders to Change. Wiley, Chichester, UK, pp. 3–13.

McNeill, F., 2006. A desistance paradigm for offender management. Criminology and Criminal Justice 6, 39–62.

Miller, K.E., Melnick, M.J., Barnes, G.M., Sabo, D., Farrel, M.P., 2007. Athletic involvement and adolescent delinquency. Journal of Youth and Adolescence 36, 711–723.

Miller, W., Rollnick, S., 2002. Motivational Interviewing: Preparing People for Change, second ed. Guilford, New York.

Ministry of Justice, 2013. Proven Re-offending Statistics Quarterly Bulletin: July 2010 to June 2011. Ministry of Justice, England and Wales. London.

Mitchell, O., Wilson, D.B., MacKenzie, D.L., 2007. Does incarceration-based drug treatment reduce reoffending? A meta-analytic synthesis of research. Journal of Experimental Criminology 3, 353–375.

Nichols, G. (Ed.), 2007. Sport and Crime Reduction. Routledge, London.

Nagin, D.S., 2013. Deterrence in the twenty-first century. In: Tonry, M. (Ed.), Crime and Justice in America, 1975–2025. Crime and Justice: A Review of Research, vol. 42. The University of Chicago Press, Chicago, pp. 199–263.

Palmer, T., 1992. The Re-emergence of Correctional Intervention. Sage, Newbury Park, CA.

Pearson, F.S., Lipton, D.S., 1999. A meta-analytic review of the effectiveness of correction-based treatments for drug abuse. Prison Journal 79, 384–410.

Pearson, F.S., Lipton, D.S., Cleland, C.M., Yee, D.S., 2002. The effects of behavioral/cognitive-behavioral programs on recidivism. Crime and Delinquency 48, 476–496.

Petrosino, A., 1997. "What Works?" Revisited Again: A Meta-analysis of Randomized Field Experiments in Individual-level Interventions. Doctoral dissertation. Rutgers University, School of Criminal Justice, Newark, NJ.

Petrosino, A., Soydan, H., 2005. The impact of program developers as evaluators on criminal recidivism: results from meta-analyses of experimental and quasi-experimental research. Journal of Experimental Criminology 1, 435–450.

Petrosino, A., Turpin-Petrosino, C., Buehler, J., 2003. Scared straight and other juvenile awareness programs for preventing juvenile delinquency: a systematic review of randomized experimental evidence. Annals of the American Academy of Social and Political Science 589, 41–62.

Porporino, F., 2010. Bringing sense and sensitivity to corrections: from programmes to 'fix' offenders to services to support desistance. In: Brayford, J., Cowe, F., Dering, J. (Eds.), What Else Works? Creative Work with Offenders. Willan, Cullompton, UK, pp. 61–85.

Prochaska, J.O., Levesque, D.A., 2002. Enhancing motivation of offenders at each stage of change and phase of therapy. In: Mc Murran, M. (Ed.), Motivating Offenders to Change. Wiley, Chichester, UK, pp. 57–73.

Raine, A., 2013. The Anatomy of Violence: The Biological Roots of Crime. Random House, New York.

Redondo, S., Sánchez-Meca, J., Garrido, V., 1999. The influence of treatment programmes on the recidivism of juvenile and adult offenders: a European meta-analytic review. Psychology, Crime and Law 5, 251–278.

Redondo, S., Sánchez-Meca, J., Garrido, V., 2002. Crime treatment in Europe: a review of outcome studies. In: McGuire, J. (Ed.), Offender Rehabilitation and Treatment: Effective Programmes and Policies to Reduce Re-offending. Wiley, Chichester, UK, pp. 113–141.

Renzema, M., Mayo-Wilson, E., 2005. Can electronic monitoring reduce crime for moderate to high-risk offenders? Journal of Experimental Criminology 1, 215–237.

Rutter, M., 2012. Resilience as a dynamic concept. Development and Psychopathology 24, 335–344.

Salekin, R.T., Worley, C., Grimes, R.D., 2010. Treatment of psychopathy: a review and brief introduction to the mental model approach for psychopathy. Behavioral Sciences and the Law 28, 235–266.

Schmucker, M., Lösel, F., 2013. The Effects of Sexual Offender Treatment on Recidivism: An International Meta-analysis of Sound Quality Studies (submitted).

Shaffer, D.K., Pratt, T.C., 2009. Meta-analysis, moderators, and treatment effectiveness: the importance of digging deeper for evidence of program integrity. Journal of Offender Rehabilitation 48 (2), 101–119.

Shapland, J., Atkinson, A., Atkinson, H., Dignan, J., Edwards, L., Hibbert, J., Howes, M., Johnstone, J., Robinson, G., Sorsby, A., 2008. Does Restorative Justice Affect Reconviction? Ministry of Justice Research Series No. 10/08. National Offender Management Service, London.

Sherman, L.W., Strang, H., 2007. Restorative Justice: The Evidence. The Smith Institute, London.

Smith, P., Goggin, C., Gendreau, P., 2002. The Effects of Prison Sentences and Intermediate Sanctions on Recidivism: General Effects and Individual Differences. Research Report JS42–103/2002. Public Works and Government Services Canada, Ottawa.

Taxman, F.S., Belenko, S., 2012. Implementing Evidence-based Practices in Community Corrections and Addiction Treatment. Springer, New York.

Theobald, D., Farrington, D.P., 2009. Effects of getting married on offending: results from a prospective longitudinal survey of males. European Journal of Criminology 6, 496–516.

Tolan, P., Henry, D., Schoeny, M., Bass, A., 2008. Mentoring Interventions to Affect Juvenile Delinquency and Associated Problems. Campbell Collaboration. www.campbellcollaboration.org/reviews_crime_justice/index.php.

Tong, L.S.J., Farrington, D.P., 2006. How effective is the reasoning and rehabilitation programme in reducing offending? A meta-analysis of evaluations in four countries. Psychology, Crime and Law 12, 3–24.

Travers, R., Mann, R.E., Hollin, C.R., 2013a. Who Benefits from Cognitive Skills Programs? Differential Impact by Risk and Offense Type. Unpublished paper. NOMS, London.

Travers, R., Wakeling, H.C., Mann, R.E., Hollin, C.R., 2013b. Reconviction following a cognitive skills intervention: an alternative quasi-experimental methodology. Legal and Criminological Psychology 18, 48–65.

Visher, C.A., Winterfield, L., Coggeshall, M.B., 2006. Systematic Review of Non-custodial Employment Programs: Impact on Recidivism Rates of Ex-offenders. www.campbellcollaboration.org/reviews_crime_justice/index.php.

Ward, T., Brown, M., 2004. The good lives model and conceptual issues in offender rehabilitation. Psychology, Crime and Law 10, 243–257.

Ward, T., Maruna, S., 2007. Rehabilitation: Beyond the Risk-paradigm. Routledge, London.

Whitehead, J.T., Lab, S.P., 1989. A meta-analysis of juvenile correctional treatment. Journal of Research in Crime and Delinquency 26, 276–295.

Wilson, D.B., Gallagher, C.A., MacKenzie, D.L., 2000. A meta-analysis of correction-based education, vocation, and work programs for adult offenders. Journal of Research on Crime and Delinquency 37, 347–368.

Wilson, S.J., Lipsey, M.W., 2000. Wilderness challenge programs for delinquent youth: a meta-analysis of outcome evaluations. Evaluation and Program Planning 23, 1–12. http://dx.doi.org/10.1016/S0149-7189(99)00040-3.

Wilson, D.B., Lipsey, M.W., Soydan, H., 2003. Are mainstream programs for juvenile delinquency less effective with minority youth than majority youth? A meta-analysis of coutcomes research. Research on Social Work Practice 13, 3–26.

Reification: History of the Concept

Frederic Vandenberghe, Institute for Social and Political Studies, State University of Rio de Janeiro, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of reification (*Verdinglichung*) is a central category of critical social theory. In the Hegelian-Marxist tradition, from Marx via Lukács to the Frankfurt School, it is used to denounce the alienation and dehumanizing impact of social systems under capitalism. Occasionally, it is also used to criticize the hypostasis of concepts and the transformation of humans into things by naturalist approaches of the social.

Until fairly recently, reification was a central diagnostic concept of critical social theory and social philosophy. Due to its heavy metaphysical baggage and its grounding in an obsolescent philosophy of history, the theory of reification has lost much of its credibility and prestige. To explain, describe, and criticize various forms of dehumanization in modern capitalist societies, it is occasionally rediscovered, refurbished, and actualized by authors within the Marxist tradition of Left Hegelianism. Through a grand narrative of reification, systemic processes of commodification, exploitation, and domination that lead to a loss of community (anomie), meaning (disenchantment), and freedom (domination) are connected to a phenomenological description of the alienation of the modern self from itself, others, and the world. As a critical category, reification squarely ascribes the blame of alienation to the system. The denunciation of reification is paradoxical, however, to the extent that it presupposes that the object is really a subject, it denies what it affirms (that the world is inhuman) and affirms what it denies (namely that there still is a subject that can act and change the world).

Literally, *reification* (*Verdinglichung*) refers to the transformation of human properties, relations, processes, actions, concepts, etc. into things. As a technical term, the term reification emerged in the English language in the 1860s out of the contraction of the verb *facere* (to make) and the substantive *res* (thing), which can refer both to concrete and empirically observable things (*ens*) and to abstract indeterminate things (*aliquid*). As a synonym of 'thingification,' the inverse of personification, reification metaphorically refers to the transformation of human properties, relations, processes, actions, concepts, etc. into *res*, into things that act as pseudopersons, endowed with a life of their own. Depending on the grammatical subject of reification – who reifies what: is it the analyst who reifies the concepts or is it society that alienates the subjects? – the transformation of human properties, social relations, abstract concepts, etc. into things and types can operate both on a methodological and on a social level.

The concept is never a neutral one. Positive instances of reification (Gehlen, Latour, and Virno) are rather rare, though. Usually, the concept is used polemically to denounce the 'violence of abstractions,' either of conceptual abstractions (*Denkabstraktionen*) that suppress the reflexive embeddedness of concepts into their social context, treat social facts as things, and transform metasubjects into megasubjects, or of real abstractions (*Realabstraktionen*) that strip individuals of their autonomy and reduce them to cogs of an abstract social machinery.

As a normative–descriptive concept that denounces the dehumanizing impact of social systems, such as the market, the state, or the bureaucracy, reification shows a family resemblance with the Marxist concept of alienation and with the Weberian concept of formal rationalization. Although the dialectical idea of objectivation (*Entäusserung*)–alienation (*Entfremdung*)–reappropriation (*Aneignung*) is already present in Hegel, the real history of the concept really begins with Marx and with Georg Lukács's Hegelian interpretation of Marx in "Reification and the Consciousness of the Proletariat," the central chapter of *History and Class Consciousness* (1923). In Marx, the concept is used in the context of the critique of the fetishism of commodities, to denounce the transformation of social relations into things. In his classic formulation of the theory of reification, which constitutes the paradigmatic 'hard core' (Brunkhorst, 1983) of the critical theory of the Frankfurt School, Lukács generalized Marx's concept of commodity fetishism, and fused it with Max Weber's concept of formal rationalization and Georg Simmel's concept of the tragedy of culture. Replacing the Marxist theory of class-consciousness with a Freudian theory of repression, the members of the Frankfurt School – Max Horkheimer, Herbert Marcuse, and especially Theodor W. Adorno – radicalized, universalized, and totalized the Lukács's theory of reification. Criticizing the Frankfurt School's identification of rationalization and reification, Jürgen Habermas reformulated the theory of reification in his theory of communication in terms of the colonization of the life-world by the subsystems of the economy and the administration of the state. Axel Honneth has also revisited the theory of reification and, inspired by the phenomenology of Sartre, he has reformulated it in terms of a theory of nonrecognition that foregrounds the lack of an existential relation between self and other. Although reification has received the greatest attention in Western Marxism, and above all in Lukács, it is important not to restrict the use of the concept to that tradition but to see that the concept and the word can also and already be found in the work of Kant, Hegel, Nietzsche, Dilthey, Husserl, Heidegger, Simmel, and Max Weber to criticize the dehumanizing, rationalizing, and alienating tendencies of modernity.

Methodological Reification – or the Critique of Conceptual Abstractions

In the philosophy of the social sciences, the concept of reification is used to denounce (1) the hypostasis of

concepts (nominalist critique of realism), (2) the naturalization of the subject and the life-world (humanist critique of naturalism), and (3) the ideological justification of the status quo (dialectical critique of fetishism).

Nominalist Critique of Realism

In the case of the critique of naive conceptual realism (or reism), the notion of reification of concepts is used to denounce, from a nominalist, vitalist, or criticist perspective, the categorical error of transforming abstractions (notions, representations, and concepts) into a material reality, in a concrete object 'out there.' Reification is here understood as a synonym of the hypostasis of concepts, analytical constructs, and ideal types. It occurs when one slides 'from the substantive to the substance' and identifies the categorical thing with the 'thing in itself.' This is, for instance, the case when scientific concepts (e.g., general intelligence, the social system, and modernity) are identified with concrete, empirical entities or when macrosociologists transform their own conceptual constructs or those of the actors ('the State,' 'the Bourgeoisie,' 'the Proletariat') into historical subjects capable of agency and of determining their own ends ('the State decides,' 'the Anglican Church resists,' 'the glorious Proletariat triumphs,' etc.). It should, however, be noted that due to the absence of a consensus on the ultimate referents of reality and the fact that one can always submit the concepts of the scientist to a neo-Kantian critique of ontology, the charge of reification is almost inevitable. Given that one's typification is another's reification, the critique of the 'fallacy of misplaced concreteness' (Whitehead, 1926) is endemic in sociology.

Humanist Critique of Naturalism

The critique of positivist naturalism in terms of reification of the subject is linked to the series of methodological disputes (*Methodenstreit*) which, since the double foundation of sociology by Comte and Dilthey in the nineteenth century, have opposed the partisans of the method of causal explanation (*Erklären*) to the partisans of the interpretative methods (*Verstehen*). Drawing on Vico's principle of the *verum factum (verum et factum convertuntur)* – according to which we can understand the sociohistorical reality because it is a human product, but not nature which is a divine product – humanists claim that the appropriate method of sociology is interpretative in that it aims to understand – by means of a phenomenological and hermeneutic reconstitution of the meaning of action – the social-historical world (Hegel's objective spirit) as an objectivation of subjective actions. Social facts thus have a meaning and cannot be treated 'as if they were things' (Durkheim). The naturalistic elimination of the meaningfulness of action through statistical observation is reifying, in that it transforms psychic acts into pseudophysical facts and reduces culture to (second) nature. Against Durkheim and his fellow 'factists' who 'change the subject' of the human sciences by substituting factors for actors, humanists thus argue that social facts are not things but that things are social facts whose meaning can be understood and which can be interpreted as an 'ongoing accomplishment of the concerted activities of daily life' (Garfinkel, 1967: p. vii).

Dialectical Critique of Fetishism

The dialectical critique of fetishism offers a metatheoretical critique of the ideological implications of *bourgeois* theories and methodologies of the social that – due to a lack of reflexivity on their context of genesis and application – legitimize the status quo. Dialecticians accept the limits of 'hermeneutic idealism.' When social relations have crystallized into a 'second nature' and social subsystems follow their own pseudonatural laws, 'dehumanizing' theories (e.g., structuralism, functionalism, and systems theory) and methods (e.g., linear modeling and statistical regression) can and have to be applied. But if one does not want to fall prey to a 'reification of the second order' and give a 'reified perception of the reifying' (Adorno) that willy-nilly endorses the reality it registers, the observed facts have to be 'mediated by the totality' (Lukács) and defetishized in such a way that the tension between the real and the possible, between what is and what could or should be, becomes perceptible within the facts themselves.

Social Reification – or the Critique of Real Abstractions

In German social philosophy and critical social theory, the concept is generally used to diagnose, i.e., describe and denounce, the pathological autonomization of social structures, systems, and subsystems – which, although man-made, are out of control – follow their own laws and alienate the actors, dominating them as if they were natural forces.

In the tradition of Western Marxism, Marx's theories of alienation and the fetishism of commodities are combined with Hegel's dialectical phenomenology of spirit, Georg Simmel's theory of the tragedy of culture, and Max Weber's theory of formal rationalization to form a critical theory of society. The concept of reification is used to refer to the relatively autonomous, alienating, and alienated functioning of the social and cultural (sub-) systems of modern capitalist societies that impose their constraints on individuals, limit their freedom, and tend to reduce them to powerless 'carriers' or passive 'executioners' of the system. As products of praxis, institutions and organizations are human objectivations, but in the course of their development, the cultural and social (sub)systems have been complexified, formalized, rationalized, and depersonalized to such an extent that eventually they have been transmuted into self-referentially closed systems that function independently of the will and the consciousness of individuals, thwart their plans, threaten their autonomy, and perhaps even their existence. The critique of reification is dialectical and thus somewhat paradoxical: the insistence on the alienating autonomy of the system aims to reactivate the autonomy of the individuals and to overcome their alienation.

Although the concept of reification (*Verdinglichung*) can already be found in Hegel's *Philosophy of Right*, the real history of the concept begins with Karl Marx and with Georg Lukács's Hegelian interpretation of Marx. The origins of the theory of reification are usually found just where the word itself is absent, namely, in the famous section on the fetishism of commodities (Chap. 1, Sect. 4) of *Das Kapital*. Analyzing

capitalism as a system of generalized exchange, Marx notes that the commodity has become the universal form of the product of labor, with the result that the exchange value of the commodity supplants the use value. Consequently, the exchange value appears to those who exchange goods as a property of the commodity itself, whereas in reality, it is the result of the labor that is incorporated in the commodity and that expresses itself as a quantitative relationship between the exchanged goods. "It is nothing but the definite social relation between men themselves which assumes here, for them, the phantasmagoric form of a relation between things" (Marx, 1869: p. 86). This inversion of humans and things is not simply an illusion, however, but an expression of the real nature of social relations in a competitive market environment. In the absence of a central organism that regulates both the production and the distribution of the products of labor, the social integration of humans is imposed from without by the systemic interconnection of things.

In "Reification and the Consciousness of the Proletariat," the central chapter of *History and Class Consciousness*, Lukács, a Hegelian-Marxist who was once a student of Simmel and Weber, presents the classic formulation of the theory of reification. Synthesizing Weber's theory of formal rationalization with Marx's theory of commodity fetishism, Lukács generalizes the theory of commodity fetishism beyond the sphere of circulation. In the problem of fetishism, which he immediately identifies with the phenomenon of reification, he discovers the "central, structural problem of capitalist societies in all its aspects" (Lukács, 1923/1968: p. 257). The universality of the commodity form, conceived as the prototype of all the forms of objectivity that follow their own laws and dissimulate the traces of human relations that subtend them, affects the life of everybody, both in its objective and subjective manifestations. Objectively, individuals are confronted with a second nature of pseudothings against which they are powerless; subjectively, they are estranged from their own activity, apprehending the products of their own activity in an alienated mode – 'as if they were something else than human products' (Berger and Luckmann, 1966). Moving from the sphere of circulation to the sphere of production, Lukács rediscovers the theory of the alienation of labor which the young Marx had developed but not published in the *Parisian Manuscripts* of 1844 (see Alienation, Sociology of). In the sphere of material production, reification expresses itself most clearly in the reduction of labor power to a commodity and of the laborer to an appendix of the machine. In capitalism, reification is generalized and the fate of the worker becomes paradigmatic of the fate of everyone. Expressing the messianism of the oppressed, Lukács eventually reintroduces the Proletariat as the 'identical subject–object' of history, whose revolutionary actions overcome alienation and reification, and thus realizes the Hegelian dream of the restoration of the 'beautiful totality.'

The development of the so-called Critical Theory of the Frankfurt School can best be understood as the result of a progressive disillusion with revolutionary expectations. Eliminating the Hegelian-Marxist dialectic of consciousness, which they replace by a Freudian account of sublimation and repression, Horkheimer, Marcuse, and especially Adorno, who also gave a Nietzschean twist to the concept of reification, radicalize the Weberian-Marxist strand in Lukács' theory. Universalizing and totalizing reification to the point that it appears as an ontological feature of human civilization, they almost end up indicting Reason as such. Indeed, to explain totalitarianism (in its fascist, communist, and capitalist variants), Horkheimer and Adorno develop a negative philosophy of history which uncovers, in the first protohistorical attempts to dominate nature, the origin of the fatal unfolding of a diabolic logic of increasing reification that will find its culmination (but not its endpoint) in the death camps. In his *Theory of Communicative Action*, Jürgen Habermas (1981), the main representative of the second generation of critical theory, reformulates the theory of reification in terms of the paradigm of language. In this perspective, reification is no longer associated with rationalization as such, as was the case with Max Weber and the Frankfurt School, but reconceptualized in terms of the 'colonization of the life-world' by the subsystems of the economy and the administration of the state. When the mechanisms of systemic integration (money and power) force back the forms of social integration from those domains that can only be integrated through language, a reification ensues that leads to a pathological deformation of the life-world. In a short essay on reification, Axel Honneth, the-main representative of the third generation of the Frankfurt School, has taken a new look at an old idea. By means of a linkage between his theory of recognition and a phenomenological theory of alienation, he disconnects the concept from its Lukácsian origins and predominantly conceives of it as a deficient relation to self and others. Understood as a pathological form of life in which participative forms of knowledge (recognition, reciprocity, and care) are systematically replaced and repressed by instrumental, representative, and manipulative approaches of the world, the other, and the self, reification is redefined in terms of forgetting of primordial way of being in the world.

See also: Alienation, History of; Alienation, Sociology of; Critical Theory: Contemporary; Dehumanization; Fetishism; Marx, Karl (1818–83); Marxism in Contemporary Sociology.

Bibliography

Arato, A., Breines, P., 1979. The Young Lukács and the Origins of Western Marxism. Seabury Press, New York.

Berger, P., Luckmann, T., 1966. The Social Construction of Reality. A Treatise in the Sociology of Knowledge. Anchor Books, Garden City.

Berger, P., Pullberg, S., 1965. Reification and the sociological critique of consciousness. History and Theory 4 (2), 196–211.

Brunkhorst, H., 1983[1998]. Paradigm-core and theory dynamics in critical social theory. Philosophy & Social Criticism 24 (6), 67–110.

Chanson, V., Cukier, A., Monferran, F. (Eds.), 2014. La réification. Histoire et actualité d'un concept critique. La dispute, Paris.

Courtine, J.F., 2006. Res, ens. Radical Philosophy 39, 36–50.

Gabel, J., 1962. La fausse conscience. Essai sur la réification. Editions de Minuit, Paris (1975, False consciousness. An Essay on Reification. Blackwell, Oxford).

Garfinkel, H., 1967. Studies in Ethnomethology. Prentice-Hall, Englewood Cliffs, NJ.

Habermas, J., 1981. Theorie des kommunikativen Handelns, vols. 1 and 2. Frankfurt am Main: Suhrkamp (1984. Theory of Communicative Action. Polity Press, Cambridge, UK).

Honneth, A., 2005. Verdinglichung. Frankfurt am Main: Suhrkamp (2008. Reification: A New Look at an Old Idea. Oxford University Press, Oxford).

Lukács, G., 1923/1968. Geschichte und Klassenbewusstsein. In: Werke, vol. 2. Luchterhand, Neuwied (1971. History and Class Consciousness. Merlin, London).
Marx, K., 1869/1972. Das Kapital, vol. I. In: Marx Engels Werke, vol. 23. Dietz Verlag, Berlin, 1974. Capital, vol. 1. Lawrence & Wishart, London.
Pitkin, H., 1987. Rethinking reification. Theory and Society 16 (2), 263–293.
Rose, G., 1978. The Melancholy Science: An Introduction to the Thought of Theodor W. Adorno. Macmillan, London.
Thomason, B.C., 1982. Making Sense of Reification: Alfred Schutz and Constructionist Theory. Humanities Press, Atlantic Highlands, NJ.
Vandenberghe, F., 1997–1998. Une histoire critique de la sociologie allemande. In: Aliénation et réification, 2 vols. Editions la Découverte, Paris (2009. A Philosophical History of German Sociology. Routledge, London).
Whitehead, A.N., 1926. Science and the Modern World. Cambridge University Press, Cambridge.

Reinforcement, Principle of

Ian M Evans, Massey University, Wellington, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The principle of reinforcement is central to many psychological interventions designed to change behavior. All behaviors are influenced by their consequences, and when these are positive for the individual in some way, the response that preceded them is strengthened, that is, will be more probable when the antecedent conditions reoccur. Positive reinforcements are intrinsically valued, similar to rewards. Negative reinforcement occurs when aversive experiences, including distressing emotions, are reduced by some action, which is then strengthened. While not necessary for learning to occur, material and social rewards provide feedback essential for the shaping and maintenance of new behavior.

Reinforcement refers to two slightly different phenomena depending on the type of learning involved. In instrumental (operant) conditioning it is the presentation of any (usually desirable) event or object following a response, which results in an increased likelihood of that response reoccurring. In classical (Pavlovian) conditioning, reinforcement refers to the presentation of the stimulus that elicits the reflexive response (unconditioned response) that comes to be elicited by a new, previously neutral stimulus. With its evolutionary overtones (behaviors should be repeated or become more probable when they enhance the organism's adaptation or survival prospects), it is one of the most fundamental principles of learning. It is also the cornerstone of many practical interventions (psychological treatments) designed to change behavior, including skill development. Thus, the principle of reinforcement has wide application in clinical, educational, organizational, and community contexts. Recognition that many behaviors are determined by their consequences is a fundamental tenet of behavior therapy (*see* Behavior Therapy: Psychiatric Aspects), as well as for understanding psychological disorders.

The term 'reinforcement' can denote (1) a principle of learning; (2) a procedure designed to modify the frequency of an approved behavior; or (3) the consequence itself (also 'a reinforcer'). The term connotes 'strengthening,' and what is strengthened is the likelihood of a behavior reoccurring in a specific context. A stimulus is judged to be a positive reinforcement to the extent that it actually increases the likelihood of the behavior that preceded it. Thus, a reinforcement is not quite the same as a reward, since what constitutes reinforcement is defined in terms of what 'works' to regulate behavior, whereas a reward is identified by what an individual expresses a desire for, values, or will work to obtain. In practical applications, however, one selects reinforcements on the basis of prior experience, observation, and intuition that the planned consequence will be desired.

When the consequence (food, praise, stimulation, or even apparently negative outcomes such as experiencing pain, or annoying someone else) results in the behavior being more probable (likely to be repeated), it is referred to as positive reinforcement. When an organism works actively to eliminate unpleasant events currently in existence (such as being cold, teacher demandingness, loneliness, boring activities), it is known as negative reinforcement. When a consequence results in a behavior being less probable in the future it is known as punishment. Any previously neutral stimulus that signals the future likelihood of reinforcement comes to have reinforcing properties of its own, and is described as a secondary (or conditioned) reinforcement. Socially, money is the classic example. In applied contexts, a stimulus that has no value other than a designated rate of exchange with something else that is desired is called a token reinforcement – sometimes these are symbols such as poker chips, stars on a chart, or points that can later be exchanged for treats and privileges.

History of the Concept

Reinforcement's origin as a formal principle of modern psychology derives from Thorndike's (1927) Law of Effect: the occurrence of a 'satisfying' event strengthens the association between the setting and the behavior that was instrumental in producing it. Hull's (1943) influential learning theory also proposed that what is strengthened is the association between a prior stimulus (cue or context) and the response, with the mechanism underlying the reinforcement effect being the reduction in a basic biological drive, such as hunger or thirst. However, many stimuli can serve intrinsically as reinforcement when they have only incentive value (such as a sweet but nonnutritive substance). Intrinsic reinforcement can also be thought of as satisfying secondary, social needs, for example, security, acceptance by others, entertainment, or feedback that your response was correct in some way – the latter being the intrinsic reinforcement in many electronic games and puzzles.

The amount, frequency, and quality of the reward influences its reinforcing function, but not in exactly the same way across different animal species. Much past research has been devoted to the question of whether reinforcement is necessary for learning to occur – many early theorists suggesting that 'latent' learning takes place when no reinforcement is present, although the evidence for this is ambiguous (for a discussion, see Jensen, 2006). The question, however, may be a spurious one, since it is difficult to imagine a behavior that has no consequences, and the positive nature of the outcome might be as simple as reducing constraints on the individual, or predicting future pleasures.

In the classical conditioning paradigm (*see* Classical Conditioning Methods in Psychotherapy; Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of), Pavlov also used the term to refer to the unconditioned stimulus (UCS): the stimulus that evokes the

reflexive response. When this stimulus is regularly paired with and thus predicted by a neutral stimulus, the latter comes to evoke a somewhat similar response. The UCS is thus considered reinforcement, since its occurrence increases the tendency to emit a response in the presence of a previously neutral stimulus. Of considerable theoretical and practical importance is the phenomenon that when the UCS is removed and the previously neutral stimulus presented repeatedly, the conditioned response eventually ceases to occur. This phenomenon, as well as the procedure of removing the UCS, is known as extinction. It has an exact parallel in instrumental (operant) conditioning: if the activity no longer produces the positive outcome, its probability decreases.

Although central to both conventional instrumental and classical learning situations, the concept of reinforcement has acquired much of its practical, applied implication from Skinner's (1969) instrumental conditioning paradigm known as operant conditioning. The typical laboratory procedure is for a specific 'operant' response, such as depressing a lever or pushing a button, to result in the automatic delivery of a small amount of food for a hungry animal. The operants studied have been discrete, simple responses that can be repeated easily by the subject. A vast body of experimental research has demonstrated that the form, intensity, and temporal pattern (rate) of responding is a function of the rule specifying the relationship between the operant response, its consequence, and the context in which the consequence occurs (referred to as the 'three-term contingency'). In applied work, the analysis of these three components is often referred to as an A-B-C analysis: specifying the antecedents (situation, context, or triggering stimulus), the behavior, and its consequences.

Contingency Principles

The effectiveness of this procedure for developing or maintaining a specified behavior is influenced by the immediacy and predictability with which the reinforcement is delivered. Although it will slow the initial acquisition of new responses, the reinforcement need not be delivered after every response, an arrangement called a 'partial schedule of reinforcement.' The most common are ratio schedules, in which reinforcement is delivered after a specified number of responses, and interval schedules, in which reinforcement is delivered to the first response occurring after a predetermined period of time has elapsed. In the laboratory, these different schedules of reinforcement produce characteristic patterns of response rate. Analogies are often made with workplace contingencies; for example, if one is paid on a piecework basis it is a ratio schedule, whereas being paid every month is an interval schedule of reinforcement. Reinforcement can also be delivered intermittently, that is to say after a randomly variable number of responses or periods of time – gambling on a slot machine is an example.

Of importance for application is that intermittent reinforcement results in a behavior being more resistant to extinction, meaning that the organism will persist in the behavior for longer periods of time even when the reinforcement has been withdrawn (extinction). In clinical situations, one might think that a behavior is maladaptive when it is apparently not being reinforced and should therefore extinguish. If the individual's history of reinforcement were known, however, it might be recognized that this behavior has been reinforced intermittently in the past.

The specific form of any behavior can be determined by differential reinforcement, in which the criteria for precisely what response will result in reinforcement are carefully specified. By changing or narrowing these criteria gradually, a behavior can be 'shaped' into a particular configuration. Similarly, it is quite common in practice to reinforce differentially, decreases in an undesirable behavior, or to reinforce only an alternative, more appropriate, behavior. Reinforcement is most effective in shaping behavior when it follows the response immediately. Thus, there is also considerable overlap in practice between the concept of reinforcement and that of feedback – providing people with information regarding the quality of their performance. Feedback can, however, impair performance by directing attention to self rather than the task at hand. Similarly, reinforcements elicit responses often incompatible with the response being reinforced – for instance, when given a pellet of food, the hungry animal stops what it is doing (lever pressing) and consumes the food. As a practical example, praising a child for playing quietly (social reinforcement) may disrupt that activity and elicit approach to the parent delivering the praise, the exact opposite of what the parent hoped to achieve.

Reduction in aversive or unpleasant conditions can strongly reinforce any behavior that results in such reductions. As many negative emotions, such as fear, shame, and jealousy, are highly aversive to most people, behavior (including passive behavior such as avoidance), which reduces these negative internal states will be reinforced. This phenomenon is often used to explain human behavior that seems maladaptive but which serves the function of reducing negative feelings, such as using drugs to reduce anxiety, escaping from threatening situations, or forming inappropriate romantic relationships to reduce loneliness.

Constraints in Generalizing from the Animal Laboratory

Despite the substantial body of knowledge derived from studies of animal behavior, the analogy between the operant laboratory and real life has limitations, as can be seen from the practical examples already given. Even for animals, laboratory contingencies are contrived – unlike foraging in the wild, there is no intrinsic relationship between a lever press and obtaining food. Behaviors that are typical of the species are more easily established by reinforcement than others that are not (Breland and Breland, 1961). There are some artificial contingencies operating in everyday life too, particularly work settings, in which a person is reinforced with money, food, or favors for activities desired by those controlling these reinforcements. But more commonly our activities are intrinsically and socially rewarding, for instance, reading a book is reinforced by enjoyment of the story as well as by teacher approval, or climbing a mountain may be reinforced by a sense of accomplishment in addition to the admiration of friends. It has long been debated that if an activity which is usually

reinforced intrinsically is placed under an artificial external contingency (such as a monetary payment), will this later interfere with the intrinsic reward, such that once the artificial reward is withdrawn, the frequency of the behavior will drop below its previous rate? While early research suggested this phenomenon might occur, resulting in some authors claiming that rewards can be detrimental (Kohn, 1993), a meta-analysis of the literature established that reinforcement does not harm an individual's intrinsic motivation (Cameron and Pierce, 1994).

Another major difference between reinforcement studied in the laboratory and the principles affecting human social behavior is that the verbal specification of a contingency can function much like an actual contingency. Thus, for example, people can be promised a reward for an activity and this will change their behavior, particularly if compliance with commands and directives has been reinforced in the past. Skinner referred to this as "rule-governed behavior" whereas more cognitively oriented theorists prefer to talk of expectancies. Bandura (1977), for example, argued that performance on a task depends partly on the expectation that a certain reward will be forthcoming (outcome expectancy). Another way an individual can learn about a potential contingency between a behavior and its consequences is to observe someone else being reinforced for an activity; this is known as vicarious reinforcement.

Most general behavior theories argue that there is a strong link between reinforcement and emotion. For example, Staats' (1996) A-R-D model proposes that stimuli serving as reinforcements (R) have simultaneous attitudinal (A) value (can elicit emotion) and discriminative stimulus (D) functions (can serve as a cue for approach). When environmental conditions deprive an organism of the opportunity to perform a behavior that would likely occur in a free choice situation, it will perform other activities in order to gain access to that behavior, a phenomenon referred to as the Premack (1965) principle.

Where reinforcement is artificial, such as a socially determined prize for a job well done, humans learn from prior experience the standard of performance that deserves to be rewarded. Even young children are unlikely to appreciate a reward given for performance that is below their own standard. Thus, although self-reinforcement is often included as a treatment component for enhancing self-control, it is unlikely to be influential over and above the intrinsic satisfaction of accomplishing one's own goals. Furthermore, if the outcomes achieved after much effortful behavior are less positive than expected, people tend to reevaluate the outcome in an upward direction, an effect known as cognitive dissonance (Lawrence and Festinger, 1962).

Motivation, Reward, and Individual Differences

Although the principle of reinforcement is strictly related to the contingency between a behavior and its consequences, much recent research has focused on the properties of events that are clearly identifiable as rewards, that is, the kinds of reinforcements individuals strive to obtain. Generally, rewards satisfy biological or social needs, but can also be thought of as providing pleasure (Berridge and Kringelbach, 2013). It is now known that active approach behavior and reward systems ('pleasure centers' in the brain) are linked in terms of neural processes and pathways. The earlier idea that biological needs such as hunger and thirst motivate behavior and reinforcement is that which reduces these drives has given way to more hedonic models in which reward elicits motivation and the pursuit of goals. Neural processes involving the mesocorticolimbic dopamine system are increasingly understood within neuroscience, and provide clues to such phenomena as the special nature of addiction; however these insights are beyond the scope of this article (see, e.g., Schultz et al., 1997). Thus, there are plausible biological mechanisms for personality differences in the degree to which a given individual is responsive to positive outcomes (Corr and McNaughton, 2008).

When individuals are particularly sensitive to reward ('reward dominant') it means that they are much more likely to be influenced by positive reinforcement contingencies than by punishment, and will judge not receiving a deserved reward more unfair than being punished for something they did not do. Relative sensitivity to reward over punishment may also be a characteristic of individuals exhibiting certain clinical syndromes, such as Attention Deficit Hyperactivity Disorder, conduct disorder, and recent onset schizophrenia. There are also individual differences in preference for a small but immediate reward as opposed to a larger but more delayed one – a pattern that is generally considered to be maladaptive and detrimental to individual success. Humans are more influenced by rewards they have chosen than those arbitrarily selected by someone else.

Reinforcement as Behavioral Regulation

Much current research in the operant tradition is concerned with predicting the effects of contingencies on the distribution of available activities. Neobehavioral theories that relate reinforcement to motivation (e.g., need reduction) have given way to economic-type theories that consider the sum total of potential behaviors in a situation as well as the sum total of reinforcements. One of the most important observations emanating from this approach is that if there is more than one operant being reinforced (a concurrent schedule), animals will distribute their responses more or less in proportion to the amount of reinforcement available to each one. This is known as Herrnstein's matching law (see Noll, 1995 for a discussion of its implications).

In answer to the apparent anomalies discussed in Section Constraints in Generalizing from the Animal Laboratory, contemporary reinforcement theory stresses homeostatic principles, referring to how an organism must adjust its actions to meet the demands of the environment. An organism will normally distribute its activities so as to minimize the demands (e.g., effortfulness of the response) and maximize opportunities, and this optimum – when there is unconstrained choice – is known as the behavioral 'bliss point' (Kessler, 2009). Once reinforcement is seen in this light, it gives credence to applied approaches in which there is an emphasis on enhancing choices and opportunities for

individuals in otherwise deprived environments, or those with limited responses in their repertoire.

Theories emphasizing behavioral regulation propose that contingencies serve to constrain the organism's free flow of behavior. Given the wide range of behavioral choices available to individuals in natural situations, it is logical that removing reinforcement for one behavior will not be successful in reducing this behavior unless another, more socially desirable behavior, can be reinforced. Clinical research has repeatedly demonstrated the value of reinforcing more appropriate alternatives. Since a systems model emphasizes that increases in one behavior must inevitably be accompanied by decreases in others, extinguishing undesirable behavior and reinforcing appropriate responses may be two sides of the same coin (Evans, 2013).

Conclusions

While context/behavior/outcome (the three-term) contingencies are essential elements in the shaping and development of new skills, the reinforcing event itself has complex influences on human behavior and often serves as a cue initiating other activities. The effects of reinforcement are relative – related to the levels of reinforcement for other possible behaviors – which has important implications for designing environments that will support socially desirable activities. That there are motivational and emotional properties of reinforcement should not be underestimated, such as experiencing success, the relief of reduction in anxiety, or the satisfaction of a craving. There are also cognitive properties, with the history of availability being judged by the individual and influencing expectancies regarding the relative outcomes of different actions. Past experience with certain types of contingencies and deprivation experiences influence the way an individual will respond to a given reinforcement. The relatively simple idea that all behavior has a function – achieves some outcome for an organism – and that function regulates or controls the behavior, is a critical insight for analyzing any behavior that one wishes to change. However, the reverse of this principle – that we can artificially manipulate behavior by managing its consequences – has proved more complex than it appears, partly perhaps because humans at least, resist the experience that they are being deliberately controlled.

See also: Classical Conditioning Methods in Psychotherapy; Reinforcement: Neurochemical Substrates.

Bibliography

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavior change. Psychological Review 84, 191–215.

Berridge, K.C., Kringelbach, M.L., 2013. Neuroscience of affect: brain mechanisms of pleasure and displeasure. Current Opinions in Neurobiology 23, 294–303.

Breland, K., Breland, M., 1961. The misbehavior of organisms. American Psychologist 16, 681–684.

Cameron, J., Pierce, W.D., 1994. Reinforcement, reward, and intrinsic motivation: a meta-analysis. Review of Educational Research 64, 363–423.

Corr, P.J., McNaughton, N., 2008. Reinforcement sensitivity theory and personality. In: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, Cambridge, UK, pp. 155–187.

Evans, I.M., 2013. How and Why People Change: Foundations of Psychological Therapy. Oxford University Press, New York, NY.

Hull, C.L., 1943. Principles of Behavior: An Introduction to Behavior Theory. Appleton-Century-Crofts, New York, NY.

Jensen, R., 2006. Behaviorism, latent learning, and cognitive maps: needed revisions in introductory psychology textbooks. Behavior Analysis 29, 187–209.

Kessler, D.A., 2009. The End of Overeating: Taking Control of the Insatiable American Appetite. Macmillan, New York, NY.

Kohn, A., 1993. Punished by Rewards. Houghton Mifflin, Boston, MA.

Lawrence, D.H., Festinger, L., 1962. Deterrents and Reinforcement: The Psychology of Insufficient Reward. Stanford University Press, Stanford, CA.

Noll, J.P., 1995. The matching law as a theory of choice in behavior therapy. In: O'Donohue, W., Krasner, L. (Eds.), Theories of Behavior Therapy: Exploring Behavior Change. American Psychological Association, Washington, DC.

Premack, D., 1965. Reinforcement theory. In: Levine, D. (Ed.), Nebraska Symposium on Motivation, vol. 13. University of Nebraska Press, Lincoln, NE, pp. 123–180.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.

Skinner, B.F., 1969. Contingencies of Reinforcement: A Theoretical Analysis. Appleton-Century-Crofts, New York, NY.

Staats, A.W., 1996. Behavior and Personality: Psychological Behaviorism. Springer, New York, NY.

Thorndike, E.L., 1927. The law of effect. American Journal of Psychology 39, 212–222.

Reinforcement: Neurochemical Substrates

Larry Stein, University of California at Irvine, School of Medicine, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by J.D. Belluzzi, L. Stein, volume 19, pp. 12996–12999, © 2001, Elsevier Ltd.

Abstract

Successful goal-directed actions are strengthened and controlled by their favorable outcomes or positive stimulus consequences. This process, obviously critical for survival, is termed 'behavioral reinforcement.' Evidence reviewed here suggests that this important process is mediated by specialized dopamine, norepinephrine, opioid peptide, and cannabinoid brain systems, which produce behavioral reinforcement primarily by modification of individual neuronal firing rates rather than by rearrangement of cellular circuitries.

Introduction

The brain's reinforcement systems, when activated by natural rewards or cues that predict rewards, necessarily release their reinforcing transmitters unselectively at all target sites. How then do these systems discriminate between correct and irrelevant responses to selectively strengthen only correct responses? It is convenient to assume that the unit of reinforcement is the target neuron itself, and not, as commonly believed, the complex circuits thought to underlie "whole" responses. A key idea is that, although all target cells receive chemical reinforcement messages indiscriminately, only recently-active target cells are eligible to be modified by them. Since correct responses always precede activations of the reward system, while irrelevant behaviors do not, only those target neurons whose activity mediates correct responses will be consistently reinforced. Finally, a molecular hypothesis of cellular reinforcement is described, which presents these ideas at a biochemical level.

Three Modes of Behavioral Control – Reflex Operant, and Habit

According to Skinner (1953), behavior is that part of the functioning of an organism which engages with or acts upon the environment. Although behavior is continuous, it is often convenient to regard the organism's activity as a succession of relatively discrete responses embedded in a stream of environmental stimuli. These stimuli may act in fundamentally different ways to control responses, depending on whether the controlling stimuli precede or follow the response.

In reflexes, the important controlling stimuli precede the response and elicit it. Some reflexes are inborn (unconditioned) and others are acquired through learning (conditioned). In unconditioned reflexes, the eliciting stimulus is innately effective (e.g., food in the mouth automatically elicits salivation). In conditioned reflexes, on the other hand, an originally ineffective stimulus becomes effective through a process of associative learning or stimulus substitution. For example, after several paired presentations of a bell and food, salivation comes to be elicited by presentation of the originally ineffective bell alone. Although unconditioned and conditioned reflexes may differ in certain ways, they share the same defining property. In all reflexes, conditioned and unconditioned, the direction of information flow in the brain is conventional – afferent (sensory) to efferent (motor).

Conversely – in a second mode of regulation termed 'operant conditioning' by Skinner – the critical stimulus follows the response and may increase or decrease its subsequent probability. Such stimuli are commonly called rewards or punishments, respectively, and the resulting behavior is variously described as reinforced, goal-directed, instrumental, or operant. Unlike reflexes, operant behaviors seem to require an unconventional neuronal circuitry in which efferents (response elements) are activated before afferents (reward-detecting sensory elements). But the brain mechanisms for operant conditioning do not require circuits that link efferent to afferent substrates. This is because operant-response neurons do not directly activate reward-detecting sensory neurons. Rather, the correct response operates on the environment to produce the goal object, and it is this environmental change that activates the reward-detecting afferent systems. Thus, although the reinforcement mechanism does not require efferent-to-afferent circuitry, it must recognize efferent–afferent firing contingencies and be operated selectively by them. Put another way, the operant mechanism must cause behavioral reinforcement only when the neuronal substrates of the correct response and rewarding goal object, in that order, fire off sequentially.

Up to this point, operant conditioning has been described without mentioning any stimulus which acts before the response is made. Most frequently, however, environmental stimuli may be discriminated which signal the availability of a reward if a particular behavior is performed. (The dinner bell is a familiar example of such a 'discriminative stimulus.') After extensive training with reward-signaling discriminative stimuli, goal-directed operant behaviors may become 'habitual.' Now vigorous responses with short latencies are consistently evoked by the discriminative stimulus. This observation and much other evidence, including recent work on the anatomy and physiology of the basal ganglia, support an intriguing idea – that after extensive discriminative training, reward-signaling stimuli may acquire reflex-like (response-eliciting) properties, which serve to transform goal-directed actions into stimulus-driven habits (Yin and Knowlton, 2006; Everitt and Robbins, 2005). If so, habits may be formed by a third kind of brain reinforcement mechanism, which acts by 'stamping-in' sensory-motor linkages, as originally conceived by Thorndike (1911). A similar 'stamping-in' reinforcement process also

may be operative in the learning of skills. In this case, after sufficient skill training and practice, selective reinforcement of increasingly accurate behavioral variations may assemble or link together specific motor units, which then discharge in precise and coordinated sequences.

The problem of characterizing the operant reinforcement mechanism has two main parts. First, one must identify the neurons or neuronal systems that perform the reinforcing function (reinforcing substrate). Second, one must identify the neurons or neuronal systems that are modified by the reinforcement process (target substrate). Despite species-specific constraints, which may favor some response-reinforcer pairings over others, it is generally accepted that any goal object may reinforce any behavior in an animal's repertoire. Such flexibility at the behavioral level suggests that the brain circuitry for operant conditioning is arranged to provide (1) convergence of sensory information from all goal objects to the reinforcing substrate and (2) divergence from this substrate of reinforcing output to all target or behavioral substrates. Here I shall briefly summarize the substantial published literature that has been devoted to the first question – the identification of the reinforcing substrate. Until recently, little consideration has been given to the identification of the target substrate, but it may be useful to review some initial findings and to consider the directions that research on this question may take.

Reinforcing Neurons – Brain Self-Stimulation Evidence

The idea that behavioral reinforcement is specialized neurochemically and is mediated by brain catecholamine systems has guided research in this field for 50 years (Stein, 1962, 1968). More specific hypotheses – that dopamine, norepinephrine, and opioid-peptide brain cells may serve as reinforcing neurons – extend this idea and are supported by evidence from brain self-stimulation and drug self-administration experiments (Olds, 1975; Crow, 1972; Stein, 1964, 1978; Wise, 1982, 2004, 2008). In self-stimulation experiments, animals work to deliver electrical stimulation to their own brains through permanently indwelling electrodes much as hungry or thirsty rats work to obtain natural reinforcers, such as food or water (Figure 1). In the absence of other sources of reward, the reinforcement for self-stimulation behavior must arise from the neuronal activity that is excited by the electrical stimulus. Accordingly, it is tempting to assume that some of the neurons under the electrode tip actually are the reinforcing neurons that mediate the effects of natural rewards, or at least are neurons that directly excite them.

High self-stimulation rates are observed when electrodes are implanted in regions containing dopamine or opioid peptide cell bodies or pathways. In particular, self-stimulation tightly overlaps the distribution of dopamine cells in the ventral tegmental area and substantia nigra. Self-stimulation closely follows the anteriorly projecting dopamine fibers through the hypothalamus, but it correlates somewhat less closely with the dopamine terminal fields in the forebrain. Similarly, mapping of opioid peptide sites for self-stimulation is consistent with the idea that certain beta-endorphin and dynorphin neurons are involved in self-stimulation. Involvement of norepinephrine

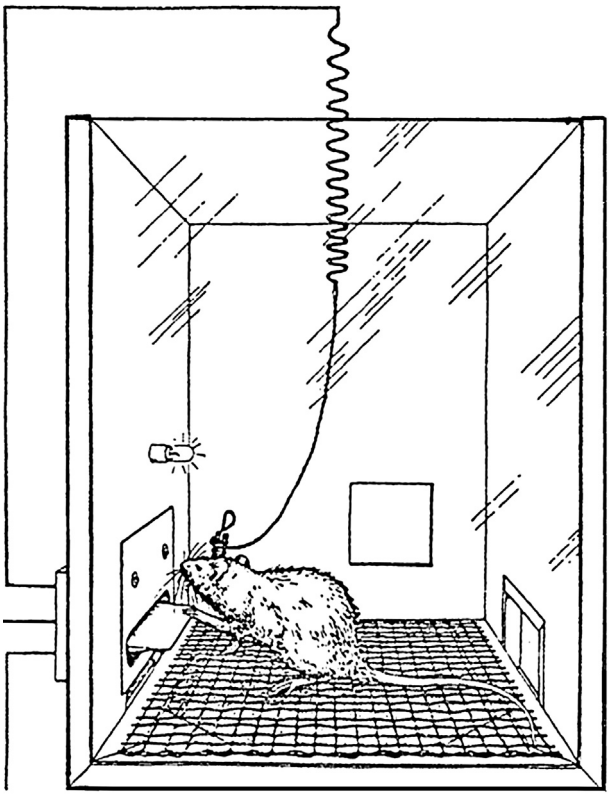

Figure 1 Diagram of self-stimulation experiment. When the rat presses on lever, it triggers a reinforcing electrical stimulus to its brain. Reprinted with permission from Olds, J., 1956. Pleasure centers in the brain. Scientific American 195, 105–116.

neurons in self-stimulation is more controversial. Although many laboratories report self-stimulation from sites in or near the locus coeruleus and its dorsal bundle projection, there is no general agreement that the noradrenergic neurons which make up this nucleus are responsible for the reinforcing effect. However, noradrenergic neurons in the locus coeruleus and other nuclei innervate the ventral tegmentum and provide excitatory drive to midbrain dopamine neurons. Thus, electrical stimulation of these noradrenergic sites, at the very least, could reinforce behavior via a dopaminergic pathway.

The catecholamine-opioid peptide reinforcement hypothesis also is firmly supported by pharmacological experiments. Significantly, selective antagonists of dopamine and opioid peptide receptors, such as pimozide and naloxone, respectively, reduce self-stimulation rates in patterns that resemble the extinction of behavior after reward omission (Wise, 2004; Trujillo et al., 1989). Thus, the antagonist drugs appear to act by blocking or weakening a chemically coded reinforcement message. In the disputed case of norepinephrine, selective inhibition of its biosynthesis by disulfiram and related agents decreases or eliminates self-stimulation over the same time course as the depletion of norepinephrine stores. Furthermore, intraventricular infusions of *l*-norepinephrine restore normal rates of self-stimulation within a few minutes, presumably after the exogenous transmitter diffuses to and replenishes reinforcement-relevant stores (Wise and Stein, 1969). Similar infusions of the unnatural isomer *d*-norepinephrine or

dopamine (whose stores are not depleted by disulfiram) fail to restore self-stimulation. An alternative interpretation of these results (Roll, 1970) – that norepinephrine is not a reinforcement transmitter, but acts merely to arouse disulfiram-sedated rats – has gained a widespread currency (e.g., Wise, 2008). But the arousal hypothesis is challenged by much other evidence. First, sedation itself is not adequate to reduce self-stimulation behavior; in fact, heavily sedative doses of barbiturates or benzodiazepines do not suppress, and sometimes even increase, rates of self-stimulation. Second, physical handling of the rats, priming trains of stimulation given by the experimenter, and control infusions of saline or other substances, which also produce arousal, fail to restore normal rates of self-stimulation. Indeed, even response-independent noradrenergic receptor stimulation by infusion of norepinephrine agonists (e.g., clonidine and oxymetazoline) is ineffective. The agonist infusions might be arousing, but nevertheless do not restore self-stimulation (arguably, because response-independent activation of noradrenergic receptors confuses the obligatory contingency relationship between self-stimulation responses and the electrically induced release of reinforcement transmitter).

Reinforcing Neurons – Drug Self-Administration Evidence

In self-administration experiments, operant behavior is reinforced by systemic or central injections of neurotransmitters or drugs. Of thousands of chemical substances available, animals and humans avidly self-administer only a few. These self-administered substances may properly be termed 'pharmacological reinforcers,' and it is interesting to ask why these compounds are selectively associated with behavioral reinforcement. Surely, it is no coincidence that all of the most notorious drugs of abuse (heroin, cocaine, methamphetamine, and marijuana) have the ability to mimic, release, or otherwise facilitate one or more of the natural reinforcement transmitters. Thus, many dopamine, opioid peptide, and cannabinoid receptor activators are known to support self-administration behavior (e.g., Self and Stein, 1992; Wise, 2004). In particular, agonists acting at dopamine D_1-like and D_2-like receptors, such as apomorphine (D_1/D_2), piribedil (D_2), and SKF 82958 (D_1) are self-administered. Similarly, selective agonists of mu and delta opioid receptors function as positive reinforcers as do the naturally occurring opioid peptides, beta-endorphin and dynorphin. The patterns of self-administration of many of these receptor agonists (particularly, D_1 and mu agonists) closely resemble the patterns observed when cocaine or heroin are self-administered. Hence, it is reasonable to suppose that self-administration of stimulant or opiate drugs may largely depend on their ability to activate or to enhance the activity of dopamine or opioid peptide reinforcement systems, respectively. As expected, dopamine and opioid peptide antagonists appropriately block, not only the reinforcing effects of dopamine and opioid receptor agonists, but also the reinforcing effects of stimulant and opiate drugs. Finally, a significant role for norepinephrine in stimulant drug abuse is suggested by clinical evidence – the most promising agent for the treatment of cocaine addiction is disulfiram, a selective inhibitor of norepinephrine synthesis (Weinshenker and Schroeder, 2007). Animal models of relapse following withdrawal suggest a possible explanation of disulfiram's effectiveness in patients. In disulfiram-treated rats, the rapid reinstatement of extinguished cocaine-seeking behavior, which is characteristically induced by drug-associated stimuli, stressors, or injections of cocaine itself, is greatly attenuated by norepinephrine depletion (Schroeder et al., 2010).

Self-administration methods also have been used in attempts to identify the sites of action of stimulants and opiates. In some studies, animals are trained to self-inject cocaine and morphine directly into the brain to find preferred sites for their reinforcing effects. In other studies, dopamine or opioid receptor antagonists are injected into different brain regions in attempts to block systemic self-administration of drugs of abuse. Results suggest that important sites of action for both stimulant and opiate drugs are concentrated in the nucleus accumbens, but reinforcing actions are also reported after direct injections of cocaine in the prefrontal cortex, morphine in the ventral tegmental area, and dynorphin in the hippocampus. Evidently, target sites for the reinforcing actions of drugs are widely distributed in the brain.

Methodological Caveats and Terminology

Although these anatomical and pharmacological findings support the idea that operant reinforcement is mediated by specialized neurotransmitter systems, some important problems remain. Administration of pure substances in the self-administration experiments strictly controls the chemical nature of the reinforcing injections, but because the distribution of injected transmitters to active sites cannot exactly duplicate that of naturally released transmitters, the ensuing pattern of receptor activation could be artifactual or misleading. On the other hand, although electrical activation of transmitter pathways during self-stimulation presumably releases the reinforcing chemical messenger in a relatively natural distribution at appropriate postsynaptic sites, the electrical stimulus also must cause the simultaneous release of many irrelevant transmitters and modulators. Thus in self-stimulation, identification of which cells and which transmitters are relevant to the reinforcement process is largely a matter of inference. Solutions based on mapping studies demonstrate self-stimulation in anatomically coherent systems, and those based on pharmacologic studies implicate specific neurotransmitters. However, the mapping studies often are insufficiently detailed, and the pharmacological studies can be criticized because they frequently fail to distinguish specific drug effects on reinforcement from nonspecific effects on motor performance or attention. Despite such difficulties, the dopamine–norepinephrine–opioid peptide–cannabinoid theory of reinforcement and drug addiction enjoys an increasingly widespread acceptance (for a discerning review, see Weinshenker and Schroeder, 2007). Of course, other neurotransmitters (including some not yet discovered) also may be found eventually to participate in reinforcement functions.

As used in this article, the terms 'reinforcement' and 'positive reinforcement' refer to a behavior-strengthening brain

process – the mechanism by which positive outcomes or favorable stimulus consequences increase the probability of operant responses. The term 'negative reinforcement' refers to a similar behavior-strengthening process, but one which is produced instead by the avoidance or termination of a noxious consequence. (Note: It is common to misuse 'negative reinforcement' – and to confuse it with 'punishment' – to refer to the suppression of behavior by the presentation of noxious consequences.) In the literature, positive reinforcement is often used interchangeably with the terms 'reward' and 'pleasure.' The term 'pleasure,' as used here, refers to a sweet or pleasant feeling that is experienced subjectively and thus is difficult to measure accurately for scientific purposes. As a consequence, we remain uncertain whether or not catecholamine, opioid peptide, or cannabinoid neurotransmission is required for the mediation of pleasurable feelings. 'Reward' is a term in the vernacular that may refer either to the reinforcing stimulus or goal object itself (noun), or to the process of reinforcement (verb). Its availability as a synonym of reinforcement is often stylistically convenient. Other terms, 'incentive' or 'incentive salience,' often are used to refer to neutral cues whose properties are changed in instrumental learning after pairings with biologically significant stimuli. According to some investigators, incentives facilitate behavioral performance by inducing an increase in dopamine-mediated 'motivation' or 'wanting' (Berridge, 2007; but see Wise, 2008). In any case, what seems undisputable – from research on Parkinson's disease in man and extensive related work in many species – is the crucial role of dopamine systems in movement and motor function. Evolution, it appears, has long used dopamine as a facilitator of action in animals. Accordingly, in the natural selection of a behavior-facilitating reinforcement mediator, brain dopamine may have been an obvious candidate.

Reinforcing Neurons – Electrophysiological Evidence

Electrophysiological evidence supports the idea that dopamine activity may be more directly related to behavioral facilitation than to pleasurable emotions. Schultz (2000) measured the firing of dopamine neurons in the ventral tegmental area of monkeys during studies of operant conditioning. Before training, the dopamine neurons discharged at high rates only when the food was touched or tasted. In well-trained monkeys, however, the dopamine neurons discharged rapidly just before or during the performance of the operant response, but not when animals actually ate the food. If it is correct to assume that pleasurable feelings would be most strongly evoked by the taste of the food, then, in well-trained animals, the correlation between dopamine activation and pleasure seems to weaken and disappear. Similar findings were obtained in stimulus discrimination experiments. At the start of training, reward-signaling clues did not elicit high-rate dopamine cellular responses. After many correctly discriminated responses, however, each presentation of the positive clue triggered a brief phase of elevated dopamine firing. Again, and significantly, the touching and tasting of food by these highly trained monkeys no longer excited an increase in dopamine firing rates. Schultz's results may explain how goal-directed operant actions may be transformed into habitual responses by discrimination training (as noted above, Section Three Modes of Behavioral Control – Reflex Operant, and Habit). Specifically, the shift in the phasic activation of dopamine neurons from food itself to a discriminative stimulus could mediate the neurobehavioral shift from operant response to habit. According to this idea, early in training, the food presentations activate reinforcing dopamine cells to elevate the probability of any just-completed operant response. Later in training, when food no longer excites such dopaminergic firing, the less precise operant process is deactivated. Now the habit formation process is predominant because reinforcing dopamine cells are activated only by the discriminative stimulus. This mode of reinforcement presumably stamps-in a more discerning and useful association – the linkage between a reliable reward-signaling stimulus and a well-documented correct response.

Target Neurons – Theoretical Considerations and Evidence from Cellular Operant Conditioning

As noted, little theoretical or experimental effort has been devoted to the identification of the substrate that is modified by the operant reinforcement process. What are the neuronal targets of the reinforcing neurons? Because it is operant behavior that is reinforced, it is plausible to assume that behavioral substrates are major targets. It is commonly believed that operant conditioning involves the strengthening or reorganization of complex (but yet to be identified) neuronal circuits, which presumably form the substrates of reinforceable responses. According to this conventional view, the whole response is the unit of reinforcement. However, where in the rat's brain, for example, might one find the circuit for a lever press response? And how might this circuitry change when lever pressing is reinforced? Such anatomical and physiological considerations, as well as difficulties at the behavioral level associated with response definition, challenge the idea that elaborate neuronal circuits are the functional units for reinforcement. According to Skinner (1953), a more useful conceptual scheme recognizes the ultimately continuous nature of operant behavior. In this scheme, as noted, operant behavior is construed as a stream of actions and interactions with the environment, which arises from the intrinsic activity and constantly shifting activation of infinitesimal behavioral elements. It is these elements, or 'behavioral atoms,' and not their manifestation in combination as whole responses that are the units actually strengthened by reinforcement. If so, and if atoms of behavior can be represented by the firing of individual neurons, then it could be argued that it is the behavior of individual neurons, and not of neuronal circuits, which is directly modified by reinforcing signals (Klopf, 1982).

The cellular reinforcement hypothesis has been tested in my laboratories by use of a novel '*in vitro* reinforcement' paradigm (Stein et al., 1993, 1994). The method uses training procedures closely analogous to those of behavioral operant conditioning to answer the question: Can the activity of relatively isolated brain cells in hippocampal slices be reinforced by local micropressure injections of dopamine or other reinforcing transmitters or drugs? My colleagues and I found that the bursting responses of individual pyramidal neurons in the CA1

layer are in fact progressively increased by burst-contingent applications of dopaminergic and cannabinoid receptor agonists (Figure 2). The bursting responses of CA3 units are similarly increased by dynorphin, an opioid receptor agonist. The same microinjections, administered independently of cellular activity, fail to facilitate and frequently suppress hippocampal bursting. This finding, and the observation that glutamate (an excitatory transmitter that is not associated with the behaviorally reinforcing effects of drugs) fails to increase bursting, together indicate that general stimulation of hippocampal activity is an unlikely explanation of the facilitatory action of the burst-contingent injections. Reinforcement delays exceeding 200 ms largely eliminate the reinforcing action of dopaminergic agonists. Taken together, these initial findings are consistent with the suggestion that the reinforcing actions of dopaminergic, opioid, and cannabinoid drugs may be modeled *in vitro* at the cellular level.

A Molecular Hypothesis of Reinforcement

Surprisingly, only burst firing of hippocampal cells – and not single-spike firing – is amenable to *in vitro* reinforcement. This unexpected observation led to the formulation of a detailed molecular hypothesis of cellular and behavioral operant conditioning (Stein, 1997). The model is based on the premise that the phosphorylation of L-type calcium channels is the ultimate step in the reinforcement of hippocampal bursting

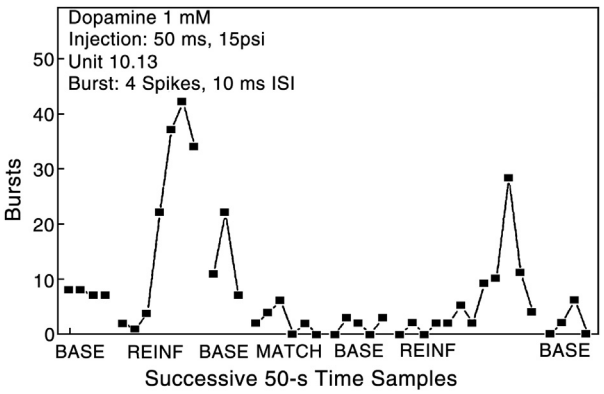

Figure 2 Operant conditioning of the activity of a CA1 pyramidal cell in a slice of dorsal hippocampus with local injections of dopamine used as reinforcement. The bursting activity of the unit throughout seven phases of a complete experiment is shown. Each point shows the number of bursts in successive 50-s time samples. Prior to the first baseline phase, a burst criterion of four or more spikes was selected. This criterion gave a stable burst rate for this unit that varied between 7 and 8 per 50 s in the initial baseline period (BASE). In the reinforcement period (REINF), dopamine HC1 (1 mM in 165 mM saline) was applied for 50 ms immediately after each burst; *after an initial period of suppression*, the burst rate eventually increased to a maximum of 42 per 50 s. Following a second baseline period, the same number of dopamine injections were delivered independently of the unit's behavior as a control for possible stimulant effects (MATCH). The number of injections was matched to that earned during the last six 50-s periods of the reinforcement phase. Bursting rates were increased by contingent dopamine injections during the reinforcement periods, but were not increased when the same injections were administered noncontingently in the matched-injection period.

responses. Hippocampal bursts are made up of a few initial sodium spikes followed by a succession of calcium spikes. The sodium spikes induce a brief membrane depolarization, which may trigger a subsequent series of calcium spikes via the voltage-gated L-type calcium channels, but only if the calcium channel protein is in its active (phosphorylated) configuration. Any useful molecular hypothesis of *in vitro* reinforcement must provide a plausible explanation of the burst-dependent (response-contingent) nature of dopamine's reinforcing action. Just as in behavioral reinforcement, response contingency in cellular reinforcement is obligatory. As noted, we indeed find that burst-independent microinjections of dopamine are ineffective as reinforcements in cellular operant conditioning. According to my working hypothesis, this important observation indicates, at the molecular level, that burst-independent dopamine microinjections fail to promote the phosphorylation of L-type calcium channels.

What is needed most critically to complete the hypothesis is a molecular coincidence detector able to respond selectively to the conjunction of bursting activity (calcium influx) and dopamine microinjections (D_1/D_5 receptor activation). Interestingly, a brain enzyme has been described which exhibits the precise biochemical properties and subcellular localization required for the molecular coincidence detector. *Type VIII adenylyl cyclase* efficiently catalyzes the phosphorylation of calcium channels, and it is present in high concentrations in the hippocampal CA1 region. More particularly, it is activated synergistically by the joint action of two regulatory molecules: calcium ion and dopamine (which acts via a stimulatory G protein subunit). In the absence of one or both regulatory molecules, Type VIII adenylyl cyclase is essentially inactive. Thus, according to my hypothesis, for successful cellular operant conditioning, dopamine must be microinjected just after a bursting response (when local calcium ion concentrations are briefly elevated) to effectively activate the reinforcing type VIII enzyme. Notably, the failure of burst-independent dopamine to support cellular reinforcement has a related explanation. Burst-independent dopamine applications do not activate the type VIII coincidence detector because the requisite local calcium concentrations are insufficient when neurons are quiescent or firing only in single spikes.

The foregoing mechanism, based on second-messenger activation of L-type calcium channels, is a short-term process appropriate for temporary reinforcing effects. For long-term reinforcement effects, which presumably require new protein synthesis and structural changes, the genetic machinery of the reinforced target cells and their synapses must be engaged. Intriguingly, the same second messengers involved in the above-hypothesized short-term reinforcement mechanism – calcium ion and cAMP – can also activate the important transcription factor cAMP response element binding protein (CREB), which has long been implicated in long-term memory formation. Hence, selective activation of CREB in reinforced target cells, triggered by rises in calcium and cAMP, could plausibly initiate the new protein synthesis and structural modifications necessary for long-term operant conditioning. Recent evidence from David Self's laboratory on CREB overexpression in cocaine self-administering rats provides experimental support for this suggestion (Larson et al., 2011).

See also: Brain Stimulation Reward; Consciousness, Neural Basis of; Emotion, Neural Basis of; Learning and Memory, Neural Basis of; Limbic System; Neural Plasticity; Neurons and Dendrites, Integration of Information In; Neurotransmitters; Synaptic Mechanisms and Protein Synthesis in Memory.

Bibliography

Berridge, K.C., 2007. The debate over dopamine's role in reward; the case for incentive salience. Psychopharmacologia 191, 391–431.

Crow, T.J., 1972. A map of the rat mesencephalon for electrical self-stimulation. Brain Research 36, 265–273.

Everitt, B.J., Robbins, B.W., 2005. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience 8, 1481–1489.

Klopf, A.H., 1982. The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence. Hemisphere Press, New York.

Larson, E.B., Graham, D.L., Arzaga, R.R., et al., 2011. Over-expression of CREB in the nucleus accumbens shell increases cocaine reinforcement in self-administering rats. Journal of Neuroscience 31, 16447–16457.

Olds, J., 1956. Pleasure centers in the brain. Scientific American 195, 105–116.

Olds, J., 1975. Reward and drive neurons: 1975. In: Wauquier, A., Rolls, E.T. (Eds.), Brain-Stimulation Reward. North-Holland, Amsterdam, pp. 1–27.

Roll, S.K., 1970. Intracranial self-stimulation and wakefulness: effect of manipulating brain catecholamines. Science 168, 1370–1372.

Schultz, W., 2000. Multiple reward signals in the brain. Nature Reviews Neuroscience 1, 199–207.

Schroeder, J.D., Cooper, D.A., Schank, J.R., et al., 2010. Disulfiram attenuates drug-primed reinstatement of cocaine seeking via inhibition of dopamine-β-hydroxylase. Neuropsychopharmacology 35, 2440–2449.

Self, D.W., Stein, L., 1992. Receptor subtypes in opioid and stimulant reward. Pharmacology and Toxicology 70, 87–94.

Skinner, B.F., 1953. Science and Human Behavior. The Free Press, New York.

Stein, L., 1962. Effects and interactions of imipramine, chlorpromazine, reserpine and amphetamine on self-stimulation: possible neurophysiological basis of depression. In: Wortis, J. (Ed.), Recent Advances in Biological Psychiatry. Plenum Press, New York, pp. 288–308.

Stein, L., 1964. Self-stimulation of the brain and the central stimulant action of amphetamine. Federation Proceedings 23, 836–850.

Stein, L., 1968. Chemistry of reward and punishment. In: Efron, D.H. (Ed.), Proceedings of the American College of Neuropschopharmacology. US Government Printing Office, Washington, DC, pp. 105–123.

Stein, L., 1978. Reward transmitters: catecholamines and opioid peptides. In: Lipton, M.A., DiMascio, A., Killam, K.F. (Eds.), Psychopharmacology: A Generation of Progress. Raven Press, New York, pp. 569–581.

Stein, L., 1997. Biological substrates of operant conditioning and the operant-respondent distinction. Journal of the Experimental Analysis of Behavior 67, 246–253.

Stein, L., Xue, B.G., Belluzzi, J.D., 1993. A cellular analogue of operant conditioning. Journal of the Experimental Analysis of Behavior 60, 41–53.

Stein, L., Xue, B.G., Belluzzi, J.D., 1994. In vitro reinforcement of hippocampal bursting: a search for Skinner's atoms of behavior. Journal of the Experimental Analysis of Behavior 61, 155–168.

Thorndike, E.L., 1911. Animal Intelligence: Experimental Studies. Macmillan, New York.

Trujillo, K.A., Belluzzi, J.D., Stein, L., 1989. Opiate antagonists and self-stimulation: extinction-like response patterns suggest select reward deficit. Brain Research 492, 15–28.

Weinshenker, D., Schroeder, J.P., 2007. There and back again: a tale of norepinephrine and drug addiction. Neuropsychopharmacology 32, 1433–1451.

Wise, C.D., Stein, L., 1969. Facilitation of brain self-stimulation by central administration of norepinephrine. Science 163, 299–301.

Wise, R.A., 1982. Neuroleptics and operant behavior: the anhedonia hypothesis. Behavior and Brain Science 5, 39–87.

Wise, R.A., 2004. Dopamine, learning and motivation. Nature Reviews Neuroscience 5, 483–494.

Wise, R.A., 2008. Dopamine and reward: the anhedonia hypothesis 30 years on. Neurotoxicity Research 14 (2–3), 169–183.

Yin, H.H., Knowlton, B.J., 2006. The role of the basal ganglia in habit formation. Nature Reviews Neuroscience 7, 464–476.

Relapse Prevention Training

Peter H Wilson, Southern Cross University, Coffs Harbour, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 13002–13005, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Relapse prevention training involves the use of a set of procedures that are designed to minimize relapse following a psychological intervention. Marlatt and Gordon (1985) were responsible for developing a set of relapse prevention strategies for addictive disorders, but the approach has been extended to a broad range of psychological problems including obesity, depression, and criminal behavior. The approach includes the identification of (and preparation for) high-risk situations, rehearsal of coping skills, relapse-crisis debriefing, education, and lifestyle modification. Although there are some encouraging findings in the treatment outcome literature, empirical support for relapse prevention has been mixed and many experimental studies have methodological limitations which prevent strong conclusions at this stage.

Many psychological and psychiatric problems are characterized by a temporal pattern in which symptoms are either chronic or vary in severity over a long period of time, with oscillations between remissions and return of symptoms. Substance-related disorders such as alcohol dependence are often relatively enduring, and mood disorders typically follow the second pattern of alternating symptom presence and remission. The empirical demonstration of the efficacy of psychotherapeutic interventions requires substantiation of the long-term benefits of therapies for such disorders (*see* Psychological Treatments, Empirically Supported). Accordingly, there has been a growth of interest in the development of therapeutic approaches that are designed to minimize relapse (or, conversely, maximize the maintenance of therapeutic gains) over extended periods of time. Relapse prevention refers to a set of procedures that designed to obtain long-lasting benefits. Marlatt and Gordon (1985) published the seminal work in the area of relapse prevention, in which they described a set of relapse prevention strategies for excessive alcohol consumption and other addictive disorders. This approach applies a cognitive–behavioral perspective to the problem of relapse and involves the employment of a specific set of treatment techniques. Other researchers have focused on relapse prevention for a diverse range of problems, adapting the original approach accordingly. Empirical support for the approach has been mixed, but there are some encouraging findings in relation to some specific problems. Methodological issues in the research place limitations on the conclusions that can be reached at this stage.

Conceptual Basis of Relapse Prevention

The term Relapse Prevention (RP) has been associated with a particular set of techniques as described by Marlatt and Gordon (1985). It can be argued that any approach which has as its clear goal the minimization of relapse (or enhancement of maintenance) might be construed as RP. Few approaches other than Marlatt and Gordon's RP, however, have been embedded in a coherent theoretical framework that is intended to be generalizable across a variety of psychological or psychiatric problems. Cognitive and behavioral theories provide both the framework for the development of RP techniques and the conceptual basis for understanding why a particular person relapses and how best the problem might be ameliorated. Specifically, problem behaviors are viewed as arising from the influence of both learning principles (classical and operant conditioning) and cognitive appraisal processes (attributions, predictions, and other judgments). For example, a person's alcohol consumption is regarded as being (a) elicited by environmental cues (presence of stimuli associated with drinking such as a glass, bottle, certain companions, time of day) and (b) reinforced by its consequences (stress reduction or mood enhancement). Apart from environmental cues, a considerable amount of attention is given the role of cognitive factors in RP. Cognitive factors may include, amongst others, causal attributions or outcome expectations. Marlatt and Gordon's RP approach draws heavily upon the concept of self-efficacy (Bandura, 1977) in its theoretical foundation (*see* Self-Efficacy). Thus, methods which enhance a person's confidence that they will be able to perform the necessary behaviors and the belief that their performance will lead to an effective outcome, are thought to be the ones which are most likely to prevent relapse. A good deal of the research literature on predictors of relapse has focused on self-efficacy. For example, low self-efficacy about the performance of particular behaviors has been found to be predictive of relapse across several problem areas such as alcohol use, obesity, depression, and smoking. The underlying philosophy of RP emphasizes the notion that the maintenance period is an opportunity for the practice of newly acquired self-control strategies. Any 'lapses' that occur are viewed as being useful to provide information about the ways in which people identify, contribute to the occurrence of, or respond to, high risk situations. A 'lapse' may involve a return to the problem behavior on a single, specific occasion (or a small number of occasions), but a 'relapse' involves a repeated accumulation of lapses that approximate the level of severity of the original problem. From this standpoint, 'lapses' are seen as providing positive experience from which to draw important information that might assist in preventing future lapses, and, by implication, future relapse.

Relapse Prevention Strategies

In principle, RP procedures may be implemented at one of three different points in time, either as: (a) an integral part of the initial treatment program; (b) a discrete set of sessions that occur towards the end of treatment; or (c) a set of sessions that are spaced across a period of time after the end of initial treatment. In the case of the latter approach, sessions may sometimes be more properly regarded as 'booster sessions,' especially if the predominant approach involves a continuation of the basic components of the original treatment or the use of prompts for the person to employ the techniques used in the original treatment. Most commonly, Marlatt and Gordon's RP is viewed as an integrated treatment which focuses on the concept of prevention of 'lapses' from the outset, and continues this emphasis beyond the normal period of treatment into a 'maintenance phase.' In practice, studies of the efficacy of RP differ in the precise details of the timing of the delivery of the specific components. These components include (but are not limited to) the identification of high-risk situations, preparation for any high-risk situations that have been identified, explicit rehearsal of coping skills, relapse-crisis debriefing, education (e.g., about the effects of drugs), and lifestyle modification.

The identification of high-risk situations represents an essential component of the RP approach. For example, it might be found that resumption of alcohol consumption occurs when the person has had a clash of opinion with a colleague or manager at work. A list of idiosyncratic high-risk situations can be obtained, based either on knowledge of the person's previous history or on the responses to self-report inventories of typical situations in which the person is asked to indicate (a) how often these situations occur, and (b) how difficult it is to engage in the desired behavior or not engage in the undesired behavior when such a situation arises. In the case of cigarette smoking, typical high-risk situations could be (a) having a cup of coffee; (b) sitting down to study at the desk; or (c) seeing someone else smoke. Learning to recognize that a high-risk situation is occurring or is imminent is a crucial part of the RP approach. People are taught to recognize the importance of 'apparently irrelevant behaviors,' exemplified by the gambler who chooses to drive home from work via a more scenic route (justified by the better view) when the route just happens to pass the casino. On the other hand, other high-risk situations may be more obvious, even to the individual, such as the person who travels business class by plane and is, therefore, likely to be offered free alcoholic drinks. RP is also connected with the concept of the 'abstinence violation effect.' This effect is thought to underlie the occurrence of a relapse in a person who has merely lapsed by breaking a preplanned limit on daily alcohol or cigarette consumption. It is argued that once a person has broken the limit, they may fail to exert any further control over their behavior. RP decreases this emphasis on 'rule violation' by focusing on self-control and deriving useful information from the occurrence of any lapses.

RP involves teaching people not only to identify such situations as risky, but also to prepare themselves for reducing the likelihood of a lapse. Simulations of high-risk situations can be used to provide opportunities to practice behavioral responses under supervision of the therapist. Many relapses are thought to occur following the occurrence of negative emotional states such as anger, depressive mood, and anxiety. Thus, RP involves the identification of such emotion–behavior linkages and the use of cognitive–behavioral strategies for the management of these negative states. Other lapses may be triggered by interpersonal conflict and may necessitate a heavier therapeutic focus on helping people to develop better communication and interpersonal problem-solving skills. In the case of addictive behaviors, there may be social pressure to engage in smoking, drinking or other drug taking. Preparation for dealing with such social pressure through assertiveness training may be incorporated into the treatment. Lifestyle modification is also employed in order to increase the frequency of engagement in pleasant, relaxing activities that may also be incompatible with the addictive behaviors.

Efficacy of RP

RP has been the subject of a number of experimental evaluations since it was first proposed as a treatment for alcohol and other addictive problems. Wilson (1992) reviewed the existing RP literature in the areas of alcohol problems, smoking, and obesity. Carroll (1996) also published a review of RP to a wide range of problems. Twenty-six studies, on substance use only (covering the years 1978 to 1995), were the subject of a meta-analysis conducted by Irvin et al. (1999). Even before considering the results of the empirical literature, the point needs to be made that there is surprisingly little research published on RP despite its presence as a concept in the treatment-outcome literature for about two decades. The drawing of firm conclusions is hampered by this relatively small body of literature, especially given the diverse types of psychological problems to which RP has been applied and the variety of procedural modifications that have been made in the published studies. A second cause for concern is the relatively short length of follow-up periods, especially given the explicit focus on relapse as an empirical issue. While it is likely that maintenance diminishes as time increases, and this appears to be the case (Irvin et al., 1999), it is not clear what type and level of resources need to be put into RP-based therapeutic endeavors to bring about meaningful results in the public health sense. With these, and other limitations in mind, the following tentative conclusions can be drawn.

The earliest controlled studies of RP were conducted on alcohol problems and it continues to be the main focus of attention for RP researchers. Varying degrees of success have been reported for RP in these studies. A number of methodological explanations may be offered for some failures to find differences between RP and control groups in some studies, including low statistical power, ceiling effects associated with the success of the initial intervention, the brevity of the follow-up, and the sharing of common elements in the treatments. Irvin et al. (1999) conclude that, amongst the substance use disorders, the strongest positive effects for RP have been found for problem drinking, although even within this group of studies, the outcome variable that is associated with the best response is psychosocial adjustment rather than drinking behavior *per se*. Like problem drinking, relapse rates following

smoking cessation programs are notoriously high. While some success has been achieved using booster sessions, these effects frequently are lost after the eventual termination of the booster sessions. Studies of relapse prevention strategies for smoking cessation have also yielded inconsistent results. In their meta-analysis, Irvin et al. (1999) draw attention to the relatively poor outcomes with RP in the area of smoking cessation. It should be noted that a few studies have been conducted on RP for illegal drug use. Cocaine use has been the subject of a small amount of research. As with smoking, evaluations of RP for these addictions have not been noted for their successful outcomes. Irvin et al. (1999) also conclude that the effects for these problems are weaker than those for alcohol problems.

The maintenance of weight loss following cognitive–behavioral treatment was the subject of a series of studies reported by Perri and co-workers in the 1980s (see Perri et al., 1992). In this research, behavior therapy plus relapse prevention training and contact after treatment produced the largest effects. Despite the sometimes positive outcome of longer-term treatment for obesity using adaptations of RP approaches, it is clear that the overall gains associated with treatments for obesity are often modest and that many people regain the weight loss during an initial treatment period.

Other psychological problems for which an attempt has been made to develop and evaluate RP approaches include depression (Wilson, 1992) and sexual offending (Laws, 1989). Although it is well established that relapse following an episode of depression is as high as 50 percent over a two-year period, there has been little systematic exploration of specific, psychologically informed RP methods for this problem. It is possible that focus on RP for some types of problems such as depression is misdirected because the major impact on maintenance may be derived from the potency of the initial cognitive–behavioral intervention.

Methodological Issues

There are several methodological problems in the study of RP and its effectiveness. RP clearly is concerned with long-term maintenance, but the relatively brief length of follow-up periods in many studies places some limitations on the conclusions that can be drawn. The RP approach contains a number of specific components, but these components are not necessarily employed in all studies and it is not clear which components are necessary for the best outcome. In many studies, RP procedures have been mixed into a broader treatment, making it difficult to separate the effects of RP from those of other components. With some problems, such as depression, it would be preferable for research on RP to be conducted using samples of people who are at high risk for a subsequent episode rather than for all people who complete an initial intervention. In some studies, especially when the RP is being delivered following the completion of an initial broad-based treatment approach (e.g., cognitive–behavior therapy for depression or anxiety), there is a need to separate initial treatment responders from nonresponders. There is little point in examining the impact of a post-treatment RP on nonresponders to an initial treatment program because they have yet to demonstrate responsiveness to treatment. These methodological issues pose particularly difficult problems for which the solutions are generally costly and time consuming to implement.

Conclusion

Despite nearly two decades of research on RP, the research produced has been quite meager and many questions remain to be addressed or resolved. The studies of RP for any given problem are so few in number that it is difficult to draw any firm conclusions at this stage. Scattered through the literature, there is some encouraging support for RP in relation to alcohol-related problems, but strong conclusions are limited due to methodological shortcomings and the relative scarcity of studies. There is a need for a large-scale evaluation of RP with addictive disorders. There has also been some progress on the development of techniques for other problems such as obesity, depression, sexual offending, and health problems.

The field of RP research may benefit from greater attention to improving our understanding of the specific ways in which RP affects basic psychological processes (i.e., investigations of the mechanism of change). Further research is also needed on both general and problem-specific predictors of relapse for particular types of symptoms so that the RP approaches can be tailored more specifically to suit the features of specific disorders. The timing of the implementation of RP procedures also needs to be considered in relation to knowledge about the temporal patterns of symptom presence and remission that are characteristic of different disorders. Continued efforts to refine and develop novel variation on RP approaches to chronic mental health and physical problems should remain a high priority for research.

Editor's Note

Since Wilson's essay was first published, the treatment community continues to search for evidence of successful relapse prevention strategies in areas such as nicotine cessation, weight loss, alcohol abuse, and others. What Wilson wrote 15 years ago, remains true today: "Despite nearly two decades [now nearly four decades – JDW] of research, the research produced has been quite meager and many questions remain to be addressed or resolved."

For a recent review of evidence-based relapse prevention approaches, see Witkiewitz and Marlatt (2007). A useful review of nicotine-cessation relapse prevention research is by Hajek et al. (2013). In the case of alcohol and drug relapse, a similarly useful review is Hendershot et al. (2011).

Two relatively recent emphases in the research literature are pharmacological strategies for relapse prevention in alcohol abuse and 'mindfulness' as a prevention approach. On the former, see Spanagel and Vengeliene (2013). 'Mindfulness' derives from Buddhist meditation practices and involves focusing the patient's mind to accept one's experiences (of craving, withdrawal, anxiety, etc.) but to recognize 'mindfully' that all such conditions are transient or impermanent. The use of the strategy in relapse prevention for substance abuse is described in Witkiewitz et al. (2013); for depression, in Piet and Houggard (2011); for smoking, Brewer et al. (2011).

See also: Behavior Therapy: Psychiatric Aspects; Psychological Treatment, Effectiveness of; Psychotherapy Integration; Psychotherapy Process Research.

Bibliography

Bandura, A., 1977. Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review 84, 191–215.

Brewer, Judson A., Mallik, Sarah, Babuscio, Theresa A., et al., 2011. Mindfulness training for smoking cessation: Results from a randomized controlled trial. Drug and Alcohol Dependence 119 (1–2), 72–80.

Carroll, K.M., 1996. Relapse prevention as a psychosocial treatment: A review of controlled clinical trials. Experimental and Clinical Psychopharmacology 4, 46–54.

Hajek, P., Stead, L.F., West, R., Jarvis, M., Hartmann-Boyce, J., Lancaster, T., 2013. Relapse prevention interventions for smoking cessation: A review. Retrieved June 24, 2104 from http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD003999.pub4/pdf/standard.

Hendershot, Christian S., Witkiewitz, Katie, George, William H., Marlatt, G. Alan, 2011. Relapse prevention for addictive behaviors. Substance Abuse Treatment, Prevention, and Policy 6 (17), 1–17.

Irvin, J.E., Bowers, C.A., Dunn, M.E., Wang, M.C., 1999. Efficacy of relapse prevention: A meta-analytic review. Journal of Consulting and Clinical Psychology 67, 563–570.

Laws, D.R. (Ed.), 1989. Relapse Prevention with Sex Offenders. Guilford Press, New York.

Marlatt, G.A., Gordon, J.R., 1985. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. Guilford Press, New York.

Perri, M.G., Nezu, A.M., Viegener, B.J., 1992. Improving the Long-Term Management of Obesity: Theory, Research, and Clinical Guidelines. Wiley, New York.

Piet, Jacob, Houggard, Esben, 2011. The effect of mindfulness-based cognitive therapy for prevention of relapse in recurrent major depressive disorder: A systematic review and meta-analysis. Clinical Psychology Review 31 (6), 1032–1040.

Spanagel, Rainer, Vengeliene, Valentina, 2013. New Pharmacological Treatment Strategies for Relapse Prevention. Current Topics in Behavioral Neurosciences 13, 583–609.

Wilson, P.H. (Ed.), 1992. Principles and Practice of Relapse Prevention. Guilford Press, New York.

Witkiewitz, Katie, Marlatt, G. Alan (Eds.), 2007. Therapist's Guide to Evidence Based Relapse Prevention. Elsevier, San Diego.

Witkiewitz, Katie, Bowen, Sarah, Douglas, Haley, Hsu, Sharon, 2013. Mindfulness-based relapse prevention for substance craving. Addictive Behaviors 38 (2), 1563–1571.

Relational Social Work

Fabio Folgheraiter, Università Cattolica del Sacro Cuore, Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Relational social work is a paradigm for which the process of existential problem solving, in whatever setting it takes place, presupposes free encounter and synergy among diverse individual efforts and points of view. It is not the outcome of unidirectional cause–effect intervention. According to this conception of welfare, well-being is not achieved through the manipulative procedures of welfare professionals and systems. On the contrary, it maintains that freedom of cognitive elaboration and the serious commitment of people in community networks can be crucial resources for the functioning, and the purpose itself, of welfare organizations and policies.

Introduction

It is a given that there were innumerable effective helpers – some of them extraordinarily skilled – in the long prescientific past of humanity. Although in itself banal, this is, nevertheless, surprising, almost inexplicable, to modern eyes. If social work is considered, like other modern professions, it would consist of the direct application of scientific knowledge to the problems of human life, making it difficult to understand how a substantial core of helping practices could exist long before the advent of modernity. Nor can one understand – after more than a century of constant technical scientific and managerial progress – why social work today is at risk of deteriorating and regressing rather than improving (Lymbery, 2001; Marston and McDonald, 2012; Rogowski, 2010). Social problems have been made increasingly interconnected and extensive by globalization (Dominelli, 2010), but they are no more severe than they were at the beginning of the profession (Bortoli, 2012). Nevertheless, today's social work seems more uncertain and debilitated than it was at that time (Bamford, 1990; Folgheraiter, 2012; Jones, 2012).

'Relational social work' (Folgheraiter, 2004, 2007, 2011) is a new approach to professional practice that harmonizes these two contradictory sets of evidence. Among the many possible causes of the current crisis, the relational paradigm points to the 'hidden vice,' which has gradually become embedded in the collective unconscious – as Jung would put it – of professional practice, namely, an ever greater technicism (Dominelli, 2010). Like other welfare professionals, social workers have been induced to overestimate the capacity of scientific knowledge, and then of objective procedures, to produce (or repair) social well-being. They have neglected other formidable logics of care that have existed and have fortunately persisted beneath the structures – even the rubble – of modernization. However, the relational paradigm conducts this critique with caution. The role of scientific power is rethought rather than rejected, because rigorous scientific reflection can reveal the limitations of positivist technoscientism in the human professions.

Copernicus understood that it was necessary to reverse the fundamentals of classical astronomy to explain celestial motion. He succeeded in this endeavor by recasting the role of the human observer, whom he ideally located outside the stellar system, and therefore no longer on an Earth conceived as the center of the universe. The relational paradigm does largely the same thing in that it no longer places the professional expert, with his or her exclusive abilities, at the center of 'clinical' action (and 'diagnostic' observation). At the same time, neither does it center on the so-called help recipients envisaged by the famous 'client-centered' approach developed by Carl Rogers (1951). The relational theory locates the crux midway between the professional social worker and his or her interlocutors based in 'lifeworlds' (Husserl, 1959) – that is, in the human communities where people and families live and interact.

Theoretical Bases and Key Concepts of the Relational Approach

As its name suggests, the relational paradigm considers helping (and the consequent well-being) as a largely imponderable process arising from well-intentioned social relations. This definition highlights the ethically driven nature of social work, which is conceived as a profession that accompanies the 'free doing of good' in human societies. One accordingly understands that this relational conception does not coincide with the well-known systemic approach of family therapy (Minuchin, 1974; Pincus and Minahan, 1973; Selvini-Palazzoli et al., 1978), which also emphasizes social relations but does so negatively and somehow mechanically (Seikkula and Arnkil, 2006) by viewing interpersonal relations as 'encysted' causes of social problems that must be repaired with specific action. Nor should the relational approach be confused with 'relational psychoanalytically informed social work' (Ornstein and Ganzer, 2005; Pozzuto and Arnd-Caddigan, 2006) which, as the name suggests, is based on a clinical matrix extraneous to the essence of the social logic of helping (Folgheraiter, 2012). The core concept of the relational approach is a nonintuitive idea of the social relationship developed in the sociological reflections of Donati (2010, 2011) and in all the principal human sciences, especially philosophical (Levinas, 1961), economic (Sen, 1980), and anthropopedagogical (Illich, 1973; Nussbaum, 2006).

Respecting Otherness

For every person, who is, nevertheless, irreducibly alone, being at ease in the world during the life journey is conditioned by the

encounter with the 'other,' Levinas writes: "We are surrounded by beings and things with which we maintain relationships. Through sight, touch, sympathy and cooperative work, we are with others" (Levinas, 1979: pp. 15–16). Realizing that the humanity of others is bound up with one's own – understanding that there is a meaningful bond between people – is to become aware that individual lives also have sense external to the self – that is, the function of social responsibility (Jonas, 1979). Perceiving 'sense' for oneself is synonymous with achieved well-being (which literally means 'feeling well in oneself'). But that sense arrives when one's being goes beyond itself and 'overflows': I am entirely well only when I open up and become 'my brother's keeper' (Bauman, 2000). In so far as I feel that I live beyond my mere functioning, I am a moral being (Barnes, 2012; Sevenhuijsen, 1998; Tronto, 1993) and I am proud of my identity.

Nevertheless – Levinas maintains – "all these relationships are transitive: I touch an object, I see the other. But I am not the Other" (Levinas, 1979: pp. 15–16). There is a minimum of two beings involved in every relationship, and by definition, they are distinct: they remain what they are in essence. In their encounter, they remain, as Levinas puts it, 'the same.' If the outcome were a merger or incorporation between them, a molding of the one to the other, the relationship would be dead. Levinas again writes: "If one could possess, grasp and know the other, it would not be the Other." Levinas takes this principle to its extreme: even the simple endeavor to know the other, because of mere desire for knowledge or functional purposes (e.g., to make a diagnosis), is subtly to pollute the relationship (Rossiter, 2011). "Knowledge is always … assimilation … with everything that goes with the 'prehend' in 'comprehending'. The most audacious and distant knowledge does not put us in communion with the real other; it does not replace sociality: it is still and always solitude" (Levinas, 1982: p. 75).

Turning to social work, or more generally to every personal helping process, it could be said that a relationship undertaken with the praiseworthy purpose of changing the other for the better, or with the intent to make the other 'conform to my idea' of him or her (Levinas, 1982), puts the social worker in a state of solitude as a 'solver.' A certain distress, at times even a burnout, results from the impossibility of being in authentic contact with the person being helped. Although this is formally the 'object' of social work, he or she is a subject. There is always an element of violence in every technique applied to the helpee, and this lack of respect clashes with relational principles. The technique undermines the effectiveness of helping processes by suffocating both the practitioner's and the recipient's energy. It produces conditions in which difficulties tend to 'decay' rather than produce real change (Hare, 2004). If, however, a technique proves effective in what Luhman calls 'people processing,' the other is reduced to a 'thing' and is therefore degraded as a human being. He or she would be unknowingly or unwillingly violated in his or her essence, in exchange for having achieved or perhaps reacquired some better functionality. These are not just academic hypotheses or logical considerations: consider, for example, behaviorism, or Skinner's lucidly delirious *Walden Two*. With the sincere intent to 'do good' (but also at the bottom, to 'clear up' a unilaterally defined disorder in the other person's life), the technical intervention generates malaise and expands it.

Relational social work neither consists simply in being good hearted nor is it merely action to produce a generic pleasantness in human contacts. Even less does it involve the use of strategies or tactics in human relations. Social workers may try to be pleasant and empathetic with their users because this is what theories tell them to do. But they often act in this way for the covert instrumental purpose of asserting their will. It is convenient for the expert to be polite when entering the world of others because this induces the trust necessary in the latter so that they behave as the expert wants. A professional of this kind resembles more a huckster or even a cheat – although, of course, acting in good faith – than a social worker.

In short, the first, and counterintuitive, principle of relational social work is that social problems may often have solutions, but they can never be 'solved.' People may change (as they constantly do) but they cannot ever 'be changed' (Prochaska et al., 1992). A social worker is not allowed to manipulate people so that their lives comply with want she or he wants them to be (Seikkula and Arnkil, 2006; Folgheraiter, 2004). (It is not allowed either methodologically – because in this way the intervention does not work – or ethically – because in this way the principle of self-determination is violated.) No professional can unilaterally eliminate problems in other people's lives just because she or he knows what needs to be done. Nor is this possible even if the expert shrewdly adopts the tactic of associating with other colleagues in a 'multidimensional team' – which is like forming a phalanx to deal with social problems. The expertise derived from scientific knowledge is only one of the factors – and not the most important – in determining the effectiveness (measurable or otherwise) of professional helping practices. When concluding a technical social work intervention, a certain effectiveness has come about notwithstanding the exercise of that technicality. It has occurred because the two people involved in the relationship have 'taken each other by the hand' and together generated – each remaining distinct – a certain additional human energy (greater than the sum of the energies of the two parties), which gradually, and unpredictably, alters the situation and produces the shared value that Donati calls a 'relational good' (Donati and Solci, 2011).

Reciprocity and Relational Empowerment

A second principle, which is also counterintuitive, is that generating well-being in complex existential situations requires each person involved to leave the role of user or client (i.e., the role of manipulated) to assume that of therapist or helper (Folgheraiter, 2000). When one is immersed in a compromised or objectively destabilized (and destabilizing) situation and begin a process of recovery that eventually makes one feel well, it is always because one has been able to exit from himself or herself as an individual beset by problems, and because one has been able to perform a useful and wide-ranging role. As long as one considers oneself a victim assisted by superior beings – however, compassionate or, as denounced by Ivan Illich (Illich et al., 1977), conceited or self-interested – one will never be able to 'feel well.' The

principle of reciprocity (or of parity or mutuality) – which is the core of the relational approach – states that users can receive true help (as prescribed by their role) only if they can give themselves to those from whom they accept it (Freeberg, 2007; Petterson and Hem, 2011). Likewise, social workers can give help (as prescribed by their role) only if they know how to ask for and receive it, primarily from their 'needy' interlocutors. Literally, 'helping relation' means that the help arises from a relation: that is, from a synergy between two or more agents engaged with equal commitment and dignity in achieving shared improvement (Folgheraiter, 2004).

This claim recalls the concept of relational empowerment. This is a rebalancing of therapeutic and manipulative power in which the party with most of it (generally the professional practitioner) cedes some to the less empowered interlocutor so that she or he becomes more autonomous and active in dealing with the situation. This operation is not zero-sum. The power surrendered is not lost but 'invested' in social relations so that it yields high-interest returns for the social worker.

Self-Help and Mutual Help

Another important principle is that of self-help. The assumption of responsibility for others has an automatic beneficial effect on the person who offers the help, in conformity with the helper therapy principle (Riessman, 1965), which says that, when I help another person with a problem, I also help myself. But there is a deeper lying relationship involved: that with oneself. The Alter for whom responsibility is assumed by a certain Ego may coincide with that same Ego. This happens whenever an Ego that feels inadequate decides to become different from what it is now (i.e., to change). In this case, we may say that Ego enters a relationship with an otherness (other from the self), which is what Ego wants or hopes to become. Ego enters a relationship with his or her life by cognitively observing it from outside as if it were another person's life and reflexively interacting with it (Archer, 2003). In this internal game of mirrors, the person motivated to reorganize his or her life – to 'redesign' it (Ferguson, 2001) – works on himself or herself, but this self-generated impetus may take the person out of himself or herself toward truly 'external' relationships. If a person is motivated to change (Prochaska and DiClemente, 1982), and if the dynamic of that expected change is effectively triggered in that person's life, by definition, she or he takes responsibility for himself or herself simultaneously with taking it for others. By assuming responsibility for others, that person simultaneously accepts that others assume responsibility for him or her. This reciprocal reliance of persons on one another is what is commonly called 'mutual help.' At its basis there must be trust, which is the 'glue' of social relations and which, by cumulating, creates what is generally termed 'social capital' (Folgheraiter and Pasini, 2009).

The Coping Network and Its Freedom of Action

Action directed toward other persons constructs relational patterns of varying complexity. If they are truly dialogical and not merely functionalist meetings of various kinds (like those of the above-mentioned multidisciplinary teams, or case management panels, or technical planning boards), we may call them 'coping networks' (Folgheraiter, 2011). A network is – to use Levinas's expression – a "pluralism which does not fuse into unity" (1979: p. 15): a pluralism of voices and intentions in which each is respected, a contrivance that does not attempt to coordinate or hierarchize or to impart orders or engineer individual abilities, and so on. A network is such when the dispersed units of which it is composed (people in the present case) interact without 'condensing' to become a coherent 'whole.' In a network, there is no preestablished coherence but freedom (Sen, 1999), indeed a freedom squared, of higher order, because it is the amplified result of multiple freedoms. The only coherence, if everything goes well, is with respect to the coping aim that unites all the network's members in its pursuit. In social work practice, aims are by definition profound and broad because they involve the entire lives of the people in the network. They are, in Archer's (2007) terminology, 'ultimate concerns.' The freedom and creativity that spring from the unpredictable are therefore necessary if one is not to remain caught in *ex ante* schemas, which may be intelligent but are more often proved wrong by concrete experiences.

The greatest degree of freedom is perhaps given by a certain degree of 'ignorance.' Not having the initially precise knowledge that constrains future action is, paradoxically, 'freedom.' More specifically, when the lack of knowledge is due to pure negligence – that is, if the precise knowledge is available but not known – that lack is deplorable. But when the knowledge does not exist, and cannot exist, or when it exists in abstract but has no reference to the specific contingency, then admitting ignorance is not just an honorable surrender. It is a necessary strategy.

Pretending to know is the frequent unconscious bluff pulled by modern experts. Admitting *not* to know like Socrates, as the necessary premise for mobilizing every resource in search of a truth or a certain good, is the authentic moral (and therefore operational) basis of social work. The term 'coping' denotes the effort made to resist and actively combat an evil without knowing 'how to do so' in advance. By managing as best as they can, but with an open mind, all the network's members hope to learn together along the way (Folgheraiter, 2011).

Methodological and Technical Principles

According to relational theory, therefore, a social worker is a professional who seeks to produce as yet unknown solutions by associating the people motivated to seek them. She or he endeavors to create or strengthen trusting relations that are sufficiently 'intimate' and profound to help her or him to pursue the elusive aim that she or he is paid to achieve: the 'creation' of human well-being. In negative terms, a social worker who merely redistributes the standard provisions of an organized welfare system is not a relational practitioner. The classic functions of redistribution and social control are important, but they can only be effective if they are performed 'relationally': that is, by creating and supporting associated and cooperative action among the subjects involved (Folgheraiter and Raineri, 2012).

The Facilitation of Coping Networks

Social ties, still *weak* in Granovetter's (1973) sense, can create well-being in two ways. They can do so in all those circumstances in which suffering is due to loneliness and marginalization, so that the expected well-being coincides entirely with some form of conviviality (Illich, 1973) or sociability (Donati, 2011). Or they can do so by generating transformative agency (Archer, 2003). If the aim of the associated action is to produce more than mere sociality (more than just 'being together') and people must rearrange or change the organization of their lives, networked social ties – besides providing the protective nourishment that is social support – make it possible to engender the additional mental strength necessary to remove chronic problems. This requires cohesion, determination, and courage; but above all, it requires original thought, which may conflict with the everyday routines that have caused the malaise. This both disruptive and constructive reflexivities develop through the action that Habermas (1981) calls 'communicative' but that in social work is better termed 'dialogical' (Seikkula and Arnkil, 2006).

An Optimistic Look Forward

In general, therefore, relational social workers are facilitators of human relations – expressed in networks with varying degrees of organization – able to produce the ethically directed thought typical of every deliberate coping action. Their work is second hand, so to speak: they do not act directly, but instead facilitate the action of others. By acting in this way, a social worker does not look back to identify the causes of present dysfunctions. The worker looks forward to an open future and guides/stimulates people to do likewise. A clinical expert must take time to look backward in order to define a problem objectively and determine its causes. A relational social worker regards this determinism with a certain detachment: the worker connects with the malaise and seeks to make it the focus of the raw and approximate human energies able to cope with it. This is not a 'negative' or 'malignant' approach (Kitwood, 1997) intended to emphasize negativities and almost sadistically bring them out in the open. Rather, the relational social worker seeks to fluidify the positive elements embedded in difficult or deteriorated situations, however scarce they may be.

Who Are the Members of a Coping Network?

Relational social workers interact with networks in two ways: they either seek to catalyze them *ex novo* if they (formal coping networks) do not yet exist or they approach them when they have already formed spontaneously (natural coping networks). In both cases, the most common maneuver is to gather the persons concerned around a table and accompany and support them so that each has a voice and open dialogue can develop (Seikkula and Arnkil, 2006; Morris and Burford, 2007). But who should be invited to these reflexive meetings in which each participant is on a par with the others? The answer – easy to say, but more difficult to put into practice – is this: all those who, in various capacities (users, caregivers, neighbors, volunteers, citizens, social workers, social service managers, and policy makers, among others), are already involved not in the problem as such, but in the will to cope with it (Hall et al., 2012; Raineri, 2004).

Being involved in coping with a problem means first, being able to perceive it. At times, the people involved in serious problems do not see them or deny their existence. They are, therefore, subjectively extraneous to those problems, even though they suffer from them. Second, it means that they must be worried about the problem, seeing it as an evil for themselves and wanting to eliminate it. At times, people see an evil but feel that it is none of their business. Third, it means having the internal motivation or the mental strength to address the problem (Bandura, 1997). At times, people worry about a situation and consider it their business, but do not believe that they are able to do anything about it. In conclusion, we may say that the human resources that the relational social worker looks for in distressful situations and – having found them – gathers around a table so that they can help him or her address their shared problem through discussion are those summed up in the broad concept of 'care.' Relational social workers are experts in discerning that evanescent 'essence' within others and within themselves. Their assessment skills are concentrated on identifying not so much the canonical 'needs' or 'risks' as the people willing to engage in first person in the quest for their common good.

What Does a Facilitator Essentially Do?

A facilitator gathers the motivated people together and, on an equal footing, encourages them to interact and take decisions. To 'facilitate' is not to lead, coordinate, or command. A facilitator accompanies the action of these people and supports them in all the displays that she or he deems able to mitigate the problem. Even more so, she or he must be able to support action in directions that she or he would never have envisaged alone. A relational social worker sometimes sees the emergence of decisions or opinions that she or he thinks are wrong or ethically debatable. But as she or he is also a member of the network, although performing the role of facilitator, she or he does not directly dispute those decisions or opinions. Rather, she or he 'acts as the devil's advocate' by stimulating further and better discussion on the matter. In all these cases, the social worker does not provide answers and does not give advice – not even when requested to do so – but provides reflexive feedback by referring to the network everything that she or he sees happening to it and that instead the network does not perceive. The most important elements to monitor during a network session are the dominance of one particular personality or the marginalization of another, reluctance and fear to talk, covert or overt conflicts, the loss of logical coherence, time-wasting chatter, and so on. In all these cases the facilitator, by using the basic techniques of narrative counseling (Parton and O'Byrne, 2000; Milner and O'Byrne, 2002), refers what she or he observed to the network so that it addresses every single issue.

A Graphic Representation of Relational Coping: The Social Pentagram

In a network, multiple free agents express themselves with their own voices and their own visions of an open future. All together,

they may express themselves with a single voice. A social worker who wants to support a relational process of this kind may need a conceptual tool that visually represents it, for instance, the one shown in Figure 1. Because this scheme is constructed on five coordinated logical elements, comparing social action to a musical melody, it can be called the 'social pentagram' (Folgheraiter, 2012).

Time

The first element of the scheme is the time axis pointing out the moment (t_0) when the relationship between the facilitator and the coping network starts. The abstract scheme shows that the coping process moves toward the future and never has a rigidly preestablished conclusion – as indicated by the symbol t_x, suggesting that social coping is indefinite (and therefore not amenable to managerial control). An effective coping process may require just one meeting or even 20 of them: the facilitator is free to optimize this variable as matters proceed.

The Shared Aim

The second element of the scheme is the aim of the coping process, which is the unifying cognitive purpose present – more or less clearly – in the minds of the network's members. It is the desirable state that is the reverse of the problem whose solution the network is seeking. The need to foster the maximum freedom of expression of the people concerned, and therefore their creativity, entails that the aim must be formulated in 'open' and broad terms. If the aims of professional social work are specific and circumscribed, then strictly speaking, they are objectives rather than aims. Objectives can be achieved through technical actions, that is, through chains of preestablished acts. In the case of coping, however, situations are by definition broader in their extent because they entail changes in the life courses of all those involved. Belonging in this 'open' category – in which people associate together and are free to set their own aims – are the three standard levels of social work: the management of particular situations (casework), the activation of social or self/mutual-help groups (group work), and the organization of community projects (community work).

The Network of Motivated People

The third element consists in the set of people in relation involved in the network's action. As said, they are comprised in the ideal 'container' represented by the small innermost circle in the scheme, because they are worried by the problem and motivated to attain the shared aim. By definition, the elements of a coping network are always physical people in flesh and blood, not 'abstract' things like services or organizations. Although different, they all have equal right to free speech and free reflection.

The Facilitator or Relational Guide

The fourth element of the scheme is a facilitator willing to support and accompany the network as it progresses toward achievement of its aim. The facilitator is not only an external observer of the network but also an internal member if we consider a larger network comprising a member able to conceptualize the unitariness of an associated action that is itself fragmented. This broader network capable of self-observation is represented in the scheme by the large external circle. This is called the 'formal coping network' to indicate that, because it has a facilitator (although it remains free and fluid), it is a more structured and organized entity. The two-headed arrow connecting the facilitator with the network indicates that the facilitator does not command the network but instead relates to it.

The Surrounding Environment

The fifth element of the scheme is in fact external to the dynamic of action. It generically represents everything that is in close contact with the relational coping network and affects it, for better or worse, but does not belong to it. Significant elements in this 'environment' are the various people somehow connected to the problem, even if they are not actively involved in coping with it. Some of these people may be obstacles to the shared action (e.g., a manager who

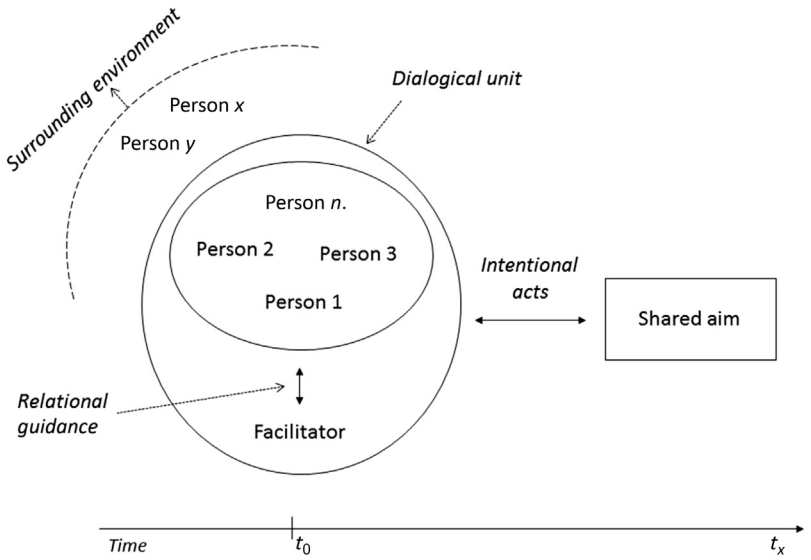

Figure 1 The social pentagram for relational social work about here.

disapproves of discussion meetings among families who have children in child protection); others may be potential resources (e.g., a disabled man's brother whom he currently does not look after). Other elements are laws, procedures, and guidelines, as well as the prejudices and inertias of institutions. All these are structural conditionings that the network's otherwise free action must somehow take into account – although the network may sometimes act to mitigate such conditionings or even, if necessary and possible, demolish them.

See also: Choice, Empowerment, and Social Work; Critical Social Work Practice; Social Work Theory.

Bibliography

Archer, M., 2003. Structure, Agency and the Internal Conversation. Cambridge University Press, Cambridge.

Archer, M., 2007. Making Our Way through the World: Human Reflexivity and Social Mobility. Cambridge University Press, Cambridge.

Bamford, T., 1990. The Future of Social Work. Macmillan, Basingstoke.

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. W.H. Freeman & Co., New York.

Barnes, M., 2012. Care in Everyday Life: An Ethic of Care in Practice. The Policy Press, Bristol.

Bauman, Z., 2000. Am I my brother's keeper? European Journal of Social Work 3 (1), 5–11.

Bortoli, B., 2012. I giganti del lavoro sociale. Grandi donne (e grandi uomini) nella storia del welfare (1526–1939). Erickson, Trento.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Donati, P., 2010. Relational Sociology: A New Paradigm for the Social Sciences. Routledge, London–New York.

Donati, P., 2011. Modernization and relational reflexivity. International Review of Sociology/Revue Internationale de Sociologie 21 (1), 21–39.

Donati, P., Solci, R., 2011. I beni relazionali. Che cosa sono e quali effetti producono. Bollati Boringhieri, Torino.

Ferguson, H., 2001. Social work, individualization, and life politics. British Journal of Social Work 31 (1), 41–55.

Folgheraiter, F., 2000. L'utente che non c'è. Lavoro di rete ed empowerment nei servizi alla persona. Erickson, Trento.

Folgheraiter, F., 2004. Relational Social Work: Toward Networking and Societal Practices. Jessica Kingsley, London.

Folgheraiter, F., 2007. Relational social work: principles and practices. Social Policy and Society 6 (2), 265–274.

Folgheraiter, F., 2011. Fondamenti di metodologia relazionale: La logica sociale dell'aiuto. Erickson, Trento.

Folgheraiter, F., 2012. The Mystery of Social Work. Erickson, Trento.

Folgheraiter, F., Pasini, A., 2009. Social capital and self-help groups: new directions in welfare policies. Social Work Education 28 (3), 253–267.

Folgheraiter, F., Raineri, M.L., 2012. A critical analysis of the social work definition according to the relational paradigm. International Social Work 55 (4), 473–487.

Freeberg, S., 2007. Re-examining empathy: a relational-feminist point of view. Social Work 52 (3), 251–259.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78 (6), 1360–1380.

Habermas, J., 1981. Theorie des kommunikativen Handelns. Suhrkamp, Frankfurt am Main.

Hall, G., Boddy, J., Chenowetha, L., Davied, K., 2012. Mutual benefits: developing relational service approaches within Centrelink. Australian Social Work 65 (1), 87–103.

Hare, I., 2004. Defining social work for the 21st century: the International Federation of Social Workers' revised definition of social work. International Social Work 47, 407–424.

Husserl, E., 1959. Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. Martinus Nijhoff's Boekhandel, L'Aja.

Illich, I., 1973. Tools for Conviviality. Harper and Row, New York.

Illich, I., McKnight, J., Zola, I.K., Caplan, J., Shaiken, H., 1977. Disabling Professions. Marion Boyars, London.

Jonas, H., 1979. Das Prinzip Verantuwortung. Suhrkamp, Frankfurt am Main.

Jones, R., 2012. The best of times, the worst of times: social work and its moment. British Journal of Social Work, 1–18. http://dx.doi.org/10.1093/bjsw/bcs157.

Kitwood, T., 1997. Demetina Reconsidered: The Person Comes First. Open University Press, Buckingham.

Levinas, E., 1961. Totalité et infini, Essai sur l'extériorité. M. Nijhoff, La Haye.

Levinas, E., 1979. Le temps et l'autre. Fata Morgana, Montpellier.

Levinas, E., 1982. Ethique et infini. Dialogues avec Philippe Nemo. Fayard, Paris.

Lymbery, M., 2001. Social work at the crossroads. British Journal of Social Work 31, 369–384.

Marston, G., McDonald, C., 2012. Getting beyond 'heroic agency' in conceptualising social workers as policy actors in the twenty-first century. British Journal of Social Work 42, 1022–1038.

Milner, J., O'Byrne, P., 2002. Brief Counselling: Narratives and Solutions. Palgrave, Basingstoke–New York.

Minuchin, S., 1974. Families and Family Therapy. Harvard University Press, Cambridge, MA.

Morris, K., Burford, G., 2007. Working with family networks-building opportunities for change? Social Policy and Society 6 (2), 209–217.

Nussbaum, M., 2006. Frontiers of Justice. Disability, Nationality, Species Membership. The Belknap Press of Harvard University Press, Cambridge, MA–London.

Ornstein, E.D., Ganzer, C., 2005. Relational social work: a model for the future. Families in Society 86 (4), 565–572.

Parton, N., O'Byrne, P., 2000. Constructive Social Work: Towards a New Practice. Palgrave, Basingstoke.

Petterson, T., Hem, M.H., 2011. Mature care and reciprocity: two cases from acute psychiatry. Nursing Ethics 18 (2), 217–231.

Pincus, A., Minahan, A., 1973. Social Work Practice: Model and Method. Peacock, Itasca.

Pozzuto, R., Arnd-Caddigan, M., 2006. Commentary on relational social work: a model for the future. Families in Society 87 (2), 301–302.

Prochaska, J.O., DiClemente, C.C., 1982. Trans-theoretical therapy – toward a more integrative model of change. Psychotherapy: Theory, Research and Practice 19 (3), 276–288.

Prochaska, J.O., DiClemente, C.C., Norcross, J., 1992. In search of how people change: applications to addictive behaviors. American Psychologist 47, 1102–1114.

Raineri, M.L., 2004. Il metodo di rete in pratica. Studi di caso nel servizio sociale. Erickson, Trento.

Riessman, F., 1965. The helper therapy principle. Social Work 10 (2), 27–32.

Rogers, C.R., 1951. Client-Centered Therapy: Its Current Practice, Implications and Theory. Constable, London.

Rogowski, S., 2010. Social Work: The Rise and Fall of a Profession? Policy Press, Bristol.

Rossiter, A., 2011. Unsettled social work: the challenge of Levinas's ethics. British Journal of Social Work 41, 980–995.

Seikkula, J., Arnkil, T.E., 2006. Dialogical Meetings in Social Networks. Karnac Books, London.

Selvini-Palazzoli, M., Boscolo, L., Cecchin, G., Prata, G., 1978. Paradox and Counter-paradox. Jason Aronson, New York.

Sen, A., 1980. Equality of what? In: McMurrin, S. (Ed.), The Tanner Lectures on Human Values. Cambridge University Press, Cambridge, pp. 197–220.

Sen, A., 1999. Development as Freedom. Oxford University Press, Oxford.

Sevenhuijsen, S., 1998. Citizenship and the Ethics of Care: Feminist Considerations on Justice, Morality, and Politics. Routledge, New York.

Tronto, J., 1993. Moral Boundaries: A Political Argument for an Ethic of Care. Routledge, London–New York.

Relativism, Cognitive

Diederick Raven, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Bunge, Volume 19, pp. 13009–13012, © 2001, Elsevier Ltd.

Abstract

Cognitive relativism is the (philosophical) position that claims there are no absolute standards against which knowledge claims can be calibrated. Two key aspects of the relativist position are: (1) epistemological skepticism regarding a realist understanding of knowledge claims, and (2) acceptance that competing theories can be incommensurable. Linguistic monism, a position that holds that the world is constituted in and through discourse, is rejected. Instead a minimalist realist position is proposed that holds that the 'world out there' has independent causal effectiveness. Skepticism is maintained about our theories telling the truth about it. The caprice problem is tackled in terms of the political constitution of science. Relativism begs the question, relative to what? Are knowledge claims relative to a group of experts or civilization as a whole? Science as reason's exemplar is a specifically Western institution. The hostility of Islamic traditionalists toward Western rational science is explained in terms of encroachment of the realm of divine authority and usurpation of Allah's authority. In Islam, any legitimate knowledge endeavor must be premised upon and prefaced by transcendental truths. Hence, unlike the Western tradition, within Islam man's inherent ratiocinating capacities can never be a legitimate source of knowledge.

Introduction

Cognitive or epistemological relativism claims there are no absolute standards, criteria, practices etc. against which knowledge claims can be calibrated to assess their truth value. The only viable way to access knowledge claims is relative to a theoretical framework or practice. Apart from vague general statements such as 'knowledge is whatever the norms of the local culture believe it to be,' no agreement ever emerged as to what the unit of analysis should be to which knowledge claims are said to be relative. In part, this reflects the negative nature of the position itself: the rejection of the basic conviction that there is or must be some permanent, ahistorical matrix to which one can ultimately appeal in determining the nature of knowledge, truth, reality, or rationality – a position best christened absolutism. Because the dominant temperament of our age is fallibilistic, the position can also be referred as objectivism.

Critics of cognitive relativism in general are philosophers of knowledge – philosophy of science is included here – who try to construe an ahistorical matrix from which all evaluation of knowledge can proceed and support this find with the strongest possible reasons. Their worry is that unless such a matrix is found to ground philosophy, knowledge, or language in a rigorous manner, radical skepticism cannot be avoided. Given the self-proclaimed job description of philosophers of knowledge as guardians of truth, rationality, and objectivity, it should not come as a surprise that much of their energy is spent on reigning in what is deemed the precarious nature of the cognitive relativist's position. To its critics relativism seems to imply an 'anything goes' stance on knowledge – there are no defenders of relativism who ever claimed such a silly position – and hence is looked upon as a poor man's epistemology. All they claim is that there is nothing to be said about either truth or rationality beyond descriptions of the legitimate procedures of justification a given society uses in an area of enquiry.

Proponents of relativism in general are empirical-minded students of knowledge practices who take as their starting point the anthropological truism 'that elsewhere things are different' and argue that knowledge practices are no exception. The unfortunate effect of this deep and long-running split is that the at times highly charged exchange between the two groups has generated more heat than light. The contours of some kind of understanding between the two have remained mainly in a zone of shade obscurity.

Critique of absolutism normally takes either of two forms:

1. The criteria set to evaluate knowledge practices cannot be met – basically because they are set too high – which is the classical skeptical take on this issue.
2. There are culturally specific assumptions underpinning the idea of absolute criteria – originating within the Western intellectual tradition – implying that the position the philosophers defend is a culturally specific one.

Prima facie, the latter kind of critique is the more fundamental because it is a direct challenge to the integrity of a transcendental realm of the philosophical a priori on which so much philosophical thinking depends.

Epistemological Skepticism as the Basis for Relativism

Skepticism as a matter of course does not involve a substantive and constructive position of its own. It is parasitical on the opponent's position. In the context of the cognitive relativist's position, this raises the crucial question 'Skepticism with regard to what?' If knowledge practices differ across time, place, and culture, as clearly anthropological and historical studies suggest they do, the skepticism it yields seems to be directed toward a realistic understanding of knowledge claims. Realism understood here as the claim that our (scientific) theories are able to pick out the building blocks of the 'world out there' is

made of. Beware, however, that (scientific) realism is a double-barreled position and consists of two separate claims: (1) there is a world of objective reality that exists independently of us, and (2) this independently existing world has a determinate nature that is within reach of our cognizing powers. Central to this double-barreled or maximalist position is the correspondence theory of truth. The minimalist position preferred by relativists only acknowledges the existence of an independent ontological realm – a realm about nothing much substantive can be said except that it exists. Applying Ockham's razor suggests this independent ontological realm is redundant. But is it?

In its most extreme form relativism presents itself as linguistic monism: *il n'y a pas de hors-texte*. The reasoning behind the claim that 'there is nothing outside the text' is: the 'objective world' is constituted in and through discourse that there is no reality independent of the words (text, signs, documents, and so on); the organization of discourse is the object. If discourse and object are identical then it is tenable to say that the collection of words acquires a currency in discourse, which creates a referent in the real world. But is it tenable to say that real worldly objects are created by virtue of a discourse? After all, by drinking water there is a chance of satisfying thirst – with accounts of drinking water, there is not.

Another way to see the implausibility of the linguistic monistic position is to ask the obvious question of why people go to such extraordinary lengths to find out how the 'world out there' reacts to all kinds of experimental interventions. The independent causal effectiveness of the 'world out there' seems a pretty good reason. Implied in this argument is that this independent ontological realm isn't redundant after all. The causal effects of the whole in the ozone layer don't disappear when we stop talking about it! Sure enough it is correct to say that people actively have to intervene in nature to come up with some kind of knowledge about it. However, it is quite another thing to say that all there is to the world is how it gets wrapped up in a web of significance spun by them. In short, it does not follow from this that the *phenomena* scientists investigate and attempt to manipulate and control are created by them. The *entities* manifest in those phenomena are not their creation. Never should we infer from the fact that our practices are necessary for access to scientific entities to the conclusion that those entities are *constituted* by our (linguistic) practices.

What made cognitive relativism such a life issue in the debates from say the time of the 1960s is that there are well-founded philosophical grounds to be skeptical about (scientific) realism. When it comes to realism, relativists are skeptics about the claim that our theories can tell the truth about this independent ontological realm. What they reject is the idea "that reality has an idiom in which it prefers to be described, that its very nature demands we talk about it without fuss – a spade is a spade, a rose is a rose – on pain of illusion, trumpery, and self-bewitchment" (Geertz). Our theories aren't in any obvious way about the ontology of the world; at best we develop a discourse ontology to talk about it.

The recalcitrance of the 'world out there' in no way explains why a particular theory is held at one time or another. At best it explains why the history of knowledge is a graveyard of rejected theories. Formulated in a more general way, the relativist claim that what philosophers take as explanatory notions for why a specific theory is held at one time or another, such as logic, methodology, rationality, are taken on their own or taken collectively, is insufficient to explain the content of a (scientific) theory. Numerous case studies have made clear that there is no such thing as a single conceptualization of let's say the idea of methodology that can account for the many uses that is being made of it over time. Feyerabend's famous and notorious remark 'anything goes' needs to be read as a *reductio ad absurdum* of his opponents who put forward a view of science based on a restricted set of methodological rules and principles. His point is that if you want a general rule that is historically adequate, i.e., fits all instances of historical development, you should go for 'anything goes.' Please notice the conditional formulation here. He is not offering this as his own constructive position, although of course his diagnoses of the ills of his opponent's views do influence his alternative of methodological pluralism: there is no single rationality, no unique way of attaining knowledge, and no single body of truth to be thereby attained.

Relativism and Its Key Concept: Incommensurability

An obvious response to this line of argument is that scientific knowledge is generated by a combination of different kinds of rules, principles, or practices, be they methodological or epistemological as well as ontological assumptions. True as that may be, this doesn't in any way reign in the room relativists claim for their own. Quite the opposite, it enhances their position. They claim that with the demise of the logical-empiricist view of knowledge, the idea that methodological rules are independent from say epistemological ones or ontological assumptions doesn't hold any longer. Since this argument leads directly to what is the key central claim of relativism – different theories, paradigms, research programs, are incommensurable – it is best to explain what the idea behind incommensurability is before proceeding with my exposition.

Suppose you're living in England in 1845 and interested in 'natives.' What would you start reading? In all likelihood you would end up with a book called *The Natural History of Man* by J.C. Pritchard MD, FRS (1786–1848) and just on the market in an enlarged second edition. If lucky, you would be able to find his *Researches into the Physical History of Man* written in 1813 as well. In case you have never heard of Pritchard – founding father of ethnology and the notion of 'diminished responsibility' can be traced back to him – his ideas are used as the main conceptual tool to present the many different peoples that make up the empire to the visitors of that "*realisation of the unity of mankind*" (Albert, Prince Consort): the great exhibition of 1851 at Crystal Place. Fast forward to 1865. Which book would you pick up? Most likely *Ancient Law*, by H.J.S. Maine (1822–88).

Significant about the change from Pritchard to Maine is that for Pritchard the central question is the challenge set by the emergence of comparative anatomy and its tendency toward a polygenist view of mankind. Dark-skinned people were taken as a separate species of mankind, and so were the Indians, etc. This runs counter to the monogenist view as presented in the bible and very well defended by the Society of Friends member Pritchard. Pritchard's main research question is a physical one.

But Maine's research question is totally different. He is interested in the social organization of savages. Apart from how this change from physical anthropology to social anthropology came about, and why it happened at the time it did – in France the change only comes into full effect after World War I – the question does present itself how best to pinpoint in any reasonable detail what is going on in this change.

We have two approaches: one physical and one social. Incommensurability is at play here and refers to specific types of difference between the two approaches. (Please note that incommensurability doesn't mean two approaches can't be compared; the point of its introduction is to clarify what is involved when we do compare alternative and rival approaches.) Prima facie each approach invokes different kind of properties and regularities, the notions of explanatory adequacy differ, and in a sense both approaches, because there doesn't seem to be much common ground between them, are in a deep sense dealing with a different problem. That's because the two approaches start from irreconcilable initial assumptions yielding irreconcilable practices of explanation.

With this in hand it is easy to understand what incommensurability is about. Incommensurability points to differences in, among other things, standards of explanatory adequacy, object domain, etc. What makes rival paradigms incommensurable is not their mere difference but the fact that these differences are (1) irreconcilable and (2) incorporated into the practices each paradigm sets for itself. Incommensurability can exist between two successive paradigms as well as between two or more competing paradigms. For the latter, think about the disagreements between behaviorists, cognitive psychologists, and psycho-analyses.

Let this suffice as an explanation of the concept of incommensurability. What are the implications of the fact that there are commensurable paradigms? Philosophically the most disturbing one is that no theory or paradigm can be privileged over another. If one wants to familiarize oneself with a paradigm one has to come to terms with an inextricable mixture of assumptions, notions, working practices, axiological orientations, etc. The implications of this are far reaching. Arguments to justify a paradigm tend to become close to circular arguments. There is no obvious way to deny that there is an element of self-vindication when a paradigm is justified in terms of its being able to account for empirical results. The key objection to relativism can now be formulated in a concise way: what to do with the Caprice problem?

Suppose Caprice decides to form a community of self-appointed experts on say climate change. Suppose that having gone over all the evidence they reject the anthropogenic thesis of climate change. Is there any (independent) argument available to rule out the social monad created by Caprice as unscientific? Is there any argument to prevent science as being portrayed as any activity of a group of self-validating experts? The sting of the Caprice problem is that there is no standard higher than the assent of the relevant community. If there is such an argument it in all likelihood is tied to the independent causal effectiveness of the 'world out there,' for example, 'long term proven reliability.'

Take, for example, chicken sexing, a manual skill that involves a lot of tacit knowledge. A company that depends for its economic survival on this knowledge needs to get sex determination to be as close as possible to a 100% right. (As it turns out, 2 out of a 100 are wrong.) Clearly the long-term success of the business is an independent test of the chicken-sexing process. If there is a clear and independent pragmatic criterion available, then obviously the answer is affirmative and one could conclude the case. But does this apply to the case of the climate deniers? For sure, of all the scientific papers published on climate change less than a quarter of one percent challenge the anthropogenic thesis of climate change. Of course this fact need not have much impact on the Caprice group – nor that the scientific consensus on this research is that it is of dubious quality.

What the Caprice example brings into to the fray is the thorny issue "What does it take to convince this group that climate change is indeed happening?" and "How long may it take to arrive at this conclusion?" The crux in this example is that if by 2050 all the evidence is finally deemed to be conclusive, it probably is too late to do anything about it. Acting now however involves making fundamental adjustments in our way of living and for that reason is politically a highly sensitive conclusion.

One way of tackling the Caprice problem is in terms of the political constitution of science. Are there paramount values that make up the science game? Its silly to say there are none. A sensible approach seems to be (1) to allow for a certain amount of variation over time of these values and (2) to allow for some variance of emphasis between different domains of knowledge – experimentation is sensible for certain domains in psychology but not for anthropology. If this is done an affirmative answer is defensible. Traditionally, what is described as central to scientific endeavors is that researchers develop hypotheses, systematically test them through data collection and analysis, and produce publishable results while archiving the data for replication. Triangulation – the use of two or more different methodological approaches to study the same phenomenon; the aim is to make sure you get independent corroboration of the data and/or method you're investigation – needs to be added here. Philosophers traditionally have talked about criteria a theory has to live up to in terms of values like accuracy, consistency, scope, simplicity, and fruitfulness. Given the impossibility to solve the Popperian demarcation problem, it is an illusion to think that however we made up this 'political constitution of science,' this will forever separate real science form the crypto- and proto-science. As case studies have shown, those trying to show that parapsychology is a truly scientific endeavor apply these paramount values much stricter than is being done in standard science. This is all to no avail; they are still not included in mainstream science. However, if we look at the sometimes-heated discussion about the scientific states of psychoanalysis, the striking point is not that those paramount values do not yield a unanimous outcome – that is more or less to be expected, experts have a habit of disagreeing among themselves over nearly everything – but the very limited number of values that are invoked at all. In short that it makes sense to take a restricted set of paramount values as the core of the political constitution science.

The suggestion to be entertained here is that cognitive relativism is in the end a much less radical thesis than claimed

by its critics. As with any social practice, the freedom there is a constrained one.

People who are usually called 'relativists' are those who say that the grounds for choosing between incommensurable opinions are less algorithmic than had been thought, not that they are made in complete freedom nor that they are made without justifications. What relativists argue is that scientific rationality in action amounts to an informed opinion, derived from a particular point of view on the issue in question, a reasoned judgment. Judgment, here, is defined as a temporary and fallible verdict that articulates the assessment of the available and relevant information. A judgment is made and informed by the political constitution of science.

Relativists Are (Social) Constructivists

Relativists are philosophically best classified as post-empiricist: they all claim the logical empiricist view is found to be wanting. It's because of the flaws in the philosophical elucidation of knowledge practices that the proponents of cognitive relativism were able to claim the integrity of an intellectual space of their own. No longer did they have to labor under the assumption that a sociological understanding of knowledge was only possible if scientific knowledge was excluded. Even the founding fathers of the sociology of knowledge, such as Weber, Scheler, and Mannheim, at the beginning of twentieth century, subscribed to that view. For relativists the gaps in the philosophical understanding of knowledge had to be plucked by means of sociological mechanisms and principles like interest and/or tradition.

One obvious drawback of this line of reasoning is that the sociological argument is a kind of 'philosophy of gaps' argument: only at those places where the traditional philosophy of science falls short of its explanatory power is there room for a sociological argument. Up to a point this criticism is correct. The sharper criticism is that the sociological theory that the relativists propose is very much of a short-range nature. It deals mainly with the local squabbles within a highly specialist community.

The scientists themselves may feel they are contributing to the solution of the bigger problems in society but their wrangling over the correct interpretation of nature is with other specialist. It is this wrangling or negotiation that is at the core of the sociological analyses of science. Having arrived at this point, a distinction needs to be made between (social) constructionism and constructivism. (If everything is subject to a social process of negotiation, adding the word *social* in one's description of the constructionism loses much of its edge, so the notion of 'social' might as well be left out of it altogether.)

Constructionism aims at showing how and why particular concepts like, for example, 'the noble savage,' or 'religion,' came into being. The idea is to show how it is the outcome of a historical process and the stress is on the contingent social historical determinants of that process. Implication of that argument is that X need not have existed, or need not be at all like it is today. In a way constructionism is trivially true, but let no one lose sight of what the strength of the position is: it asks important questions such as who determines what the problem is and who determines what the solution is? Is the current economic crisis like the one in the 1980s? In which case a monetarist stance is the solution, i.e., get the deficit under control by slashing government spending and deflating wages and house prices. Or is it a debt crisis? In which case classical austerity measures are only going to exacerbate the financial hole governments and the public are in, and more emphasis on short-term measures to stimulate the economy, via increasing the debt ratio, and getting it back on the path to growth seem to be preferable. Obviously the answer is hard fought over and is a highly charged political affair.

Constructionism explores the essential contestability of any social ontology and as such needs to be sharply distinguished from constructivism, which holds that the only reality with which an inquirer can have any cognitive commerce is reality as he constructs it to be: we only know X when we can construct X. This is the tradition of the maker's conception of knowledge that received its full philosophical articulation in Vico's *verum ipsum factum* (the truth is what is made). The attractiveness of the constructivist position is that it in a direct manner gets rid of the evidence-transcendent meaning surplus of absolutist notions like truth, rationality, and reality.

Constructivism sidesteps the skepticism by arguing that knowledge is not 'grounded' in experiences but in non-verbalized and hence noncognitive, routines, and actions. Practice of course directly brings with it the suggestion of routine and lack of questioning. Learning it requires a familiarization with the ways of manipulating the resources that are taken for granted. Again we arrive at the point that what is central to a relativistic understanding of knowledge is that the locus of knowledge is the community of expert practitioners.

The West versus the Rest

'Relative to our culture' is equivocal: culture can either refer to expert group or civilization. When you train as anthropologist you have to pass a series of the tests set by the community of experts in that field: write an ethnography based on your fieldwork experience among a particular group of people, say beekeepers, in which you provide thick description of closely studied phenomenon. You describe their way of living, their engagement with the bees, the skills required etc. If you train as a psychologist you learn to do experiments, how to assess the data and test different hypotheses via various complex statistical technics, etc. Relativism on the professional level means that every expert group sets its own requirements to become a qualified (licensed) member of it. In some cases, the medical profession for example, there is a legal framework in place for this.

Relativism on civilizational level is different: it reflects the prestige science as a knowledge-generating activity has. Relative to 'our culture' in the West means relative to what 'science' – or given that it comprises of many different knowledge domains – relative to the 'sciences' may be the more appropriate way of putting it. This reflects that in the West science is the only legitimate knowledge-generating activity, that science is reason's exemplar. Substitute Islam for the West and the picture radically emerges. Seyed Attas, in his book *Islam and Secularism*,

has a chapter called 'Dewesternization of Knowledge,' and the opening reads as follows:

> Many challenges have arisen in the midst of man's confusion throughout the ages, but none perhaps more serious and destructive to man than today's challenge posed by western civilization. (…) the greatest challenge that has surreptitiously arisen in our age is the challenge of knowledge, indeed, not as against ignorance; but as knowledge as conceived and disseminated throughout the world by Western civilisation; knowledge whose nature has become problematic because it has lost its true purpose due to being unjust conceived, and has been brought about the chaos in man's; life instead of, and rather than, peace and justice; knowledge which pretends to be real but which is productive of confusion and scepticism, which has elevated doubt and conjecture to the 'scientific' rank in methodology and which regards doubt as an eminently valid epistemological tool in the pursuit of truth; knowledge which has, for the first time in history, brought chaos to the Three Kingdoms of Nature, the animal, vegetable and mineral. (…) What is formulated and disseminated is knowledge infused with the character and personality of that civilisation – knowledge presented and conveyed as knowledge in that guise so subtly fused together with the real so that others take it unawares in *toto* to be the real knowledge *per se*. (Emphasis in the original.)

This is, indeed, a pronounced, elegant, and powerful critique of Western science. To understand the reasoning behind it, it needs to be contextualized.

The hostility of an Islamic traditionalist like al-Ghazzali (1058–1111) or his modern-day equivalent Qutb (1906–66) against knowledge that is acquired through the application of reason and logic is easy to understand if we start with the word الإسلام (Islam), which means (voluntary) submission to Allah. To be a follower of Allah implies accepting القرآن (Qur'an) as Allah's direct word, the willingness to live in accordance with the Qur'anic law, and accepting the shari'ah as the sole source of regulating and legislating all aspects of human live. This obedience to Allah is embodied in the Qur'anic verse "I created jinn and mankind only that they might worship Me" (51:56). The education that the Prophet Mohamed received through the Qur'an is from an Islamic perspective the ideal model to emulate. The Prophet is the most perfect example of *al-insan al kamil* (The Perfect Man – perfect because of the near complete realization of the Divine in worldly realities) and the function of education is to strive to produce men and women resembling him (the Prophet) as nearly as possible. A comprehensive and integrated approach to education in Islam strives to produce a balanced growth of the total personality through training man's spirit, intellect, rational self, feelings, and bodily senses such that faith is infused into the whole personality. The aim of Islamic educational theory is the actualization and perfection of all dimensions of the human being. But a balanced growth of the human character requires that both the heart (*qalb*) as the seat of the spirit and the intellect (*ᶜagl*) should receive equal attention.

The glorious Qur'an is the foundation of Islamic education; it is an immutable source of fundamental tenets of Islam, of its principles, ethics, and culture. The élan of the Islam is moral. The crucial point is that the idea of Islam, of submission, has quite radical implications for epistemology and rationality. Any source of authority is to be grounded in Allah. The implication of "Verily with God is full knowledge and He is acquainted (with all things)" (31:34) is that the Qur'an contains all of knowledge in principle and stands at the apex of modes and sources of knowledge.

Because it provides the essential foundation for all forms of knowledge, revelatory knowledge is the most elevated form of knowledge. The insistence on this divine sovereignty implies a rebuttal of any rationalist theory in which reason alone becomes both method and justification of human knowledge. Such a rationalist theory of knowledge is objectionable because it represents human encroachment on the realm of divine authority and is usurpation of Allah's rightful authority and yields a corrupted concept of the world:

> To favour reason at the expense of spirituality hampers balanced growth. Exclusive training of the intellect, for example, is inadequate in developing and refraining elements of love kindness, compassion and selflessness, which have an altogether spiritual ambiance and can only be appealed to by processes of spiritual training. Separating the spiritual development of the human being from the ration, temporal aspects of the same person, (…) is the main cause for the disintegration of the human personality. (Cook, 'Islamic versus Western Conceptions of Education': p. 346.)

To ascertain truth by complete reliance on reason alone is restrictive since both spiritual and temporal reality are two sides of the same sphere. Indeed, the highest form of knowledge is the perception of God (*idrak*), which cannot be realized in any other way than through faith (*iman*).

An important point here is that there is no room for legitimacy of whatever kind to knowledge claims based upon human reason alone. What is denied is that the world is ascertainable through purely rationalist method. This of course does not mean no knowledge could be produced that way, but only that it is without legitimacy. There is no room for an epistemology that assumes that truth about the world can be reached solely through the human faculties. For that to be the case any knowledge endeavor most be premised upon divine transcendental truths, most accept that knowledge starts by accepting the undiluted supremacy of divine authority and the eternal truths that are implied by this acceptance. The dominance of rote learning in the Islamic civilization signals the restricted source of the legitimacy that is allotted to human reasoning, and it reflects that legitimacy is not placed with man's inherent ratiocinating capacities. (It is this line of thinking that is behind the naming of Boko Haram, a radical Islamist group in Nigeria. In Hausa '*Boko Haram*' means 'Western education is sinful,' from the Hausa word *Boko*, western education, and the Arabic word *haram*, sin, literally forbidden.)

What this example shows is two things. Science as reason's example is a culturally specific idea that presented itself in the West, but there is no reason to assume that it is a viable option elsewhere. Secondly, attempts to critique relativism because it naturalizes reason, i.e., all that is left of notions of truth, rationality, and knowledge is a nonnormative *in situ* description, begs the question. Of course any human being has the capacity to reason, and without a doubt in any kind of reasoning there is a critical or transcendent function that can be used to criticize existing norms and practices. The question at

issue is what is the legitimacy of the intellectualistic knowledge produced in this way? The Islamic case makes clear that it is relative to the tradition one is part of.

See also: Phenomenology in Sociology; Rationality in Society; Sociology, Epistemology of; Values, Sociology of.

Bibliography

al-Attas, S.M.N., 1993[1978]. Islam and Secularism. International Institute of Islamic Thought and Civilisation (ISTAC), Kuala Lumpur.
Bernstein, R.J., 1985. Beyond Objectivism and Relativism. University of Pennsylvania Press, Philadelphia.
Cook, B.J., 1999. Islamic versus western education of education: reflections on Egypt. International Review of Education 45 (3/4), 339–357.
Davidson, D., 1974. On the very idea of a conceptual scheme. Proceedings and Addresses of the American Philosophical Association 47, 5–20.
Gellner, E., 1985. Relativism and the Social Sciences. Cambridge University Press, Cambridge.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Goodman, N., 1978. Ways of Worldmaking. Hackett, Indianapolis.
Hollis, M., Lukes, S. (Eds.), 1982. Rationality and Relativism. Blackwell, Oxford.
Horton, R., Finnegan, R. (Eds.), 1973. Modes of Thought; Essays on Thinking in Western and Non-Western Societies. Faber, London.
Jarvie, I.C., 1984. Rationality and Relativism: In Search of a Philosophy and History of Anthropology. Routledge & Kegan Paul, London.
Krausz, M., 1989. Relativism: Conflicts and Confrontations. University of Notre Dame Press, Notre Dame.
Krausz, M., Meiland, J.W. (Eds.), 1982. Relativism: Cognitive and Moral. University of Notre Dame Press, Notre Dame.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.
Margolis, J., 1991. The Truth about Relativism. Blackwell, Oxford.
Putnam, H., 1981. Reason, Truth and History. Cambridge University Press, Cambridge.
Putnam, H., 1987. The Many Faces of Realism. Open Court, La Salle, Illinois.
Quine, W. Van Orman, 1969. Ontological Relativity and Other Essays. Columbia University Press, New York.
Raven, D., et al. (Eds.), 1992. Cognitive Relativism and Social Science. Transaction Publishers, New Brunswick, NJ & London.
Rorty, R., 1980. Philosophy and the Mirror of Nature. Blackwell, Oxford.
Rorty, R., 1991. Objectivity, Relativism, and Truth: Philosophical Papers, vol. 1. Cambridge University Press, Cambridge.
Whorf, B.L., 1956. Language, Thought and Reality. M.I.T. Press, Cambridge, MA.
Winch, P., 1958. The Idea of a Social Science and Its Relation to Philosophy. Routledge & Kegan Paul, London.
Wilson, B., 1970. Rationality. Blackwell, Oxford.

Relativism: Philosophical Aspects

Maria Baghramian, University College Dublin, Dublin, Ireland

Published by Elsevier Ltd.

Abstract

Relativism starts with the premise that standards of truth, rationality, and ethical right and wrong vary greatly, and there are no universal criteria for adjudicating between them. Faced with intractable differences and disagreements the relativist argues that what is true or right can be determined only within (variable) contexts of assessment and therefore there could be more than one standard of correctness. Several interrelated doctrines are subsumed under the single heading of 'relativism': cognitive, conceptual, ethical, and semantic relativisms are some main subdivisions. The most common criticism is the charge of incoherence, for if 'all is relative' then so is relativism. Despite frequent attacks, relativism retains its appeal because it captures an essential insight about the unavailability of a univocal solution to most human problems.

Definition

Relativism is the view that cognitive, moral, or aesthetic norms vary with and are dependent on the specific social, conceptual, or linguistic contexts that underpin them, and consequently a neutral standpoint for evaluating such norms is not available. Relativists usually rely on the observation that opinions and judgments on matters of fact and value vary greatly between individuals, societies, and historical epochs and that there are no universal criteria for deciding between them. In effect, relativism can be seen as a strategy for overcoming intractable disagreements on matters where adjudication between strongly held views seems impossible. Faced with incompatible beliefs and norms, held with equal conviction, and in the absence of an overarching or universal justificatory scheme, we are pushed to the two extremes of skepticism (the Pyrrhonian option) or relativism (the Protagorean option). To put it slightly differently, presented with the contested pair of beliefs P and its negation not P, the skeptic abandons all claims to knowledge, while the relativist accepts the truth of both by making truth dependent on a hidden parameter such as culture, a conceptual scheme, or semantic context. In addition, relativists assume that different perspectives cannot converge, and in this they stand in opposition to universalist and absolutist aspirations to truth and knowledge.

Historical Background

The earliest documented source of relativism in the Western intellectual tradition is Plato's account of the Sophist philosopher Protagoras and his famous dictum, "Man is the measure of all things; of the things that are, that they are; and of the things that are not, that they are not" (Plato, 1961). The trigger for relativistic tendencies in ancient Greece, as in later times, was the increased contact between people of different cultures, customs, and beliefs, and the realization that other societies arrange their social and ethical lives in radically different ways. Relativism was subsumed under skepticism by the Pyrrhonian skeptics who in turn influenced Michel De Montaigne (1533–92), a key figure in the emergence of modern philosophy. Contemporary relativists, however, are influenced by the perspectivism of Friedrich Nietzsche (1844–1900) and the historicism of Wilhelm Dilthey (1833–1911).

According to Nietzsche, all claims of knowledge are representations of reality only from a certain standpoint. Reports of so-called 'facts' are interpretations that could always be supplemented or replaced by other interpretations. Language is not a means of describing what there is, but rather, it imposes its own interpretation on our thinking. Thus, the very idea of truth and reality is called into question by Nietzsche. Whether Nietzsche's perspectivism should be equated with relativism is open to debate, but his writings have become a source of inspiration for postmodernist and relativist thinkers such as Michel Foucault, Jacques Derrida, and Jean François Lyotard. Like Nietzsche, postmodernism, a dominant theoretical approach in literary theory, sociology, social anthropology, cultural studies, and feminist theories since the 1980s, scorns the quest for universal cognitive and moral values as a manifestation of the 'will to power.' Postmodernism, with its attendant relativistic approach on the other hand, is presented as an ally in the fight for emancipation from tyrannies of all sorts, including the hegemonic tyranny of science with its universalist aspirations.

Contemporary relativism has also been shaped by Wilhelm Dilthey's view that each nation can be seen as a self-contained unit with its own conception of reality and system of values. Different historical periods, Dilthey maintains, produce different norms and values, each of which is taken to be unconditional or universal. We cannot gain a correct understanding of social and intellectual conditions if we do not understand or take into account the inevitable historicity of all thought. Dilthey's views have been particularly influential among the advocates of cultural relativism (see Section Cultural Relativism).

Motivations

The contemporary appeal of relativism arises out of certain features of our experiences of the world, as well as a number of intellectual currents.

The Situatedness of All Judgments

Relativists frequently justify their position by pointing out the many ways in which our judgments and claims to knowledge are bound by particular situations and contexts. Many of our

assertions refer to events that happen at a particular time and place, and are experienced by a particular individual. Such assertions seem to be true only at the time, place, or circumstances of their utterance, and in that sense are relative to these parameters. Moreover, our beliefs, it is claimed, are always from a point of view, a perspective. We cannot step out of our skin, so to speak, and experience reality as it is. A God's-eye view or a view from nowhere is not available to mere humans.

Cultural Diversity

Increasing awareness of the extent of diversity of beliefs, practices, and the customs of different cultures and historical epochs has cast doubt on the universality and objectivity of judgments. Most significantly, the reports by anthropologists about the thinking habits of remote peoples have led to the suggestion that all normative judgments, whether cognitive or ethical, may have only a local validity. Western ethnocentrism and the intellectual legacy of the Enlightenment, it is charged, prevent us from appreciating or even seeing the uniqueness of different cultures and modes of thought.

The Disappearance of Old Certainties

The disappearance of old certainties, both in the religious and scientific arenas, has been instrumental in the increasing prominence of relativism. The questioning of a religiously motivated cosmology, which fixes the position of individuals within an immutable framework, helped to bolster a climate that was conducive to relativistic outlooks. In science, the discovery of the possibility of non-Euclidean geometries in the nineteenth century and Einstein's theory of relativity in the early twentieth century eroded confidence in classical geometry and Newtonian physics. New theories in quantum mechanics and particle physics, Heisenberg's uncertainty principle in particular, are also cited as evidence that universal, absolute laws are rarer than it was once imagined. Darwin's evolutionary theory also eroded the belief in the uniqueness of the human point of view. With the disappearance of the old certainties, the very idea of objectivity came into question.

Varieties of Relativism

Relativism comes in great variety. One way to distinguish between its different forms is to consider the questions:

1. What is it that is being relativized?
2. What is it being relativized to?

These dual questions lead to the following main subdivisions of relativism:

Cultural Relativism

This, probably the most influential form relativism, was a methodological assumption of much of the social sciences from the outset of their emergence as a discipline. Auguste Comte, the founder of the discipline of sociology, is responsible for the catchphrase "all is relative." William Graham Sumner, the first person to give a course titled 'Sociology' in the United States, defended cultural relativism in his book *Folkways* (1906). Social anthropologists who conducted fieldwork among tribal people adopted it as part of their methodology as well as their ethical outlook toward the world. The anthropologist M.J. Herskovits defined cultural relativism as:

> The principle of cultural relativism, briefly stated, is as follows: Judgments are based on experience, and experience is interpreted by each individual in terms of his own enculturation. … Even the facts of the physical world are discerned through the enculturative screen, so that the perception of time, distance, weight, size, and other 'realities' is mediated by the conventions of any given group. (Herskovits, 1960: p. 61)

The doctrine relies on several key assumptions.

1. A methodological assumption to the effect that human behavior and thought should be explained in cultural rather than biological or individualistic psychological terms.
2. The empirical claim that cultures vary greatly in their judgments and values (known as descriptive relativism). Ethnographic data, the claim goes, show that various groups of people hold systems of belief that according to Western standards are false, irrational, and even self-contradictory. The Nuer's belief that twins are birds or the Azande belief in witchcraft and oracles are well-known examples.
3. A philosophical assumption that there is no Archimedean point, or a view from nowhere, to adjudicate between different cultural perspectives.
4. A normative injunction to the effect that we should accept cultures and their practices on their own terms, avoid ethnocentrism, and show tolerance (prescriptive cultural relativism).

Cognitive Relativism

Cognitive relativism is the view that epistemic norms such as truth, rationality, or validity can vary between cultures and conceptual frameworks and that we have no transcultural or framework-transcendent criteria for adjudicating between them. Some main subdivisions of cognitive relativism are relativism about truth (alethic relativism), relativism about rationality and logic, and conceptual relativism.

Relativism about Truth

Relativism about truth, or alethic relativism, is the claim that what is true for one individual or social group may not be true for another and that every effort to adjudicate between context-dependent standards of truth and falsity is bound to be futile. The truth of an assertion, so the claim goes, is relative to the beliefs, attitudes, and other psychological idiosyncrasies of the individuals, or more generally to the conceptual, or cultural, background of the persons making them.

Relativism about truth is the core of all relativistic positions since other subdivisions of cognitive relativism, and even moral relativism, are reducible to it. For instance, relativism about rationality can be expressed in terms of whether there are true (universal) standards of rationality. Relativism about logic may be restated as the question whether logical truths are relative to specific cultures or cognitive schemes or whether

they are in any sense universal. Ethical relativism can be seen as the claim that the truth of ethical judgments is relative to the context or the culture in which these particular judgments are being made. In this sense, alethic relativism is the most general and strongest form of relativism and the target of the most trenchant criticisms of relativism (see 4.1). A semantic version of relativism about truth, known as truth-relativism, has been proposed in recent years and will be discussed in Section Semantic Relativism.

Epistemic Relativism

Epistemic relativism is established through three interconnected claims: (1) beliefs can be justified only relative to systems or frameworks of beliefs; (2) there are many different systems of beliefs – some of which are radically different; and (3) it is impossible to demonstrate (in a noncircular fashion) that our system of beliefs is superior to all others (Williams, 2007). The conclusion drawn is that either we can chauvinistically maintain that our belief system is superior to all or accept the equal legitimacy of all such systems.

Peter Winch's position is an example of this type of relativism. Winch has argued that since standards of rationality in different societies do not always coincide, we should use only contextually and internally given criteria of rationality in our assessment of the systems of belief of other cultures and societies. Winch, who based his arguments in reaction to E.E. Evans-Pritchard's account of the Azande's beliefs in witchcraft and magic and who was greatly influenced by the work of the philosopher Ludwig Wittgenstein, argues that it does not make sense to speak of a universal standard of rationality because what is rational is decided by norms of a given language and the form of life that provides a backdrop to it. We, as outside observers, are therefore not in a position to impute irrationality or illogicality to the Azande or other groups whose practices differ from ours.

Rationality invariably involves the idea of having good reasons, which in turn involves the application of rules of logic and questions of warrant and justification. These components of 'rational thought' have also become targets of relativistic treatment.

Relativism about Logic

Discussions of the Azande belief system, coupled with a strongly relativistic interpretation of Wittgenstein's views, have led Barry Barnes and David Bloor to argue that not only standards of rationality but also laws of logic, often seen as the necessary conditions for rational thought, are defined by and hence are relative to their social context. According to them, the very distinction between valid and invalid modes of argument is relative to the practices prevalent in a given linguistic community or society and cannot in any sense be seen as universal. Bloor argues that there are no culture-transcendent rules of validity and inference. Different societies may have incompatible, but internally coherent, systems of logic, and the norms of logical reasoning are relative to a background of cultural conditioning. According to Bloor, "The Azande have the same psychology as us but radically different institutions. If we relate logic to the psychology of reasoning we shall be inclined to say that they have the same logic; if we relate logic more closely to the institutional framework of thought then we shall incline to the view that the two cultures have different logics" (Bloor, 1976: pp. 129–130).

Relativism about Justification and Warrant

A rational person is expected to have justification for her views and warrant for her beliefs. The relativist argues that the legitimacy of a justificatory system and the perceived strength of epistemic warrant are only decided locally. Richard Rorty, who has been accused of relativism, by among others Hilary Putnam and Paul Boghossian, has made the influential claim that "there is nothing to be said about either truth or rationality apart from descriptions of the familiar procedures of justification which a given society – ours – uses in one or another area of inquiry" (Rorty, 1991: p. 23). The belief in the necessarily parochial scope of any justification or reason we may offer is a hallmark of relativistic thinking.

Conceptual or Ontological Relativism

Conceptual or ontological relativism is motivated by philosophical considerations about the relationship between human thought and the world rather than mere cultural and historical differences. The roots of conceptual relativism rest with Immanuel Kant's (1724–1804) distinction between the data of our sense experiences and the principles of organization or the categories we use to organize them. Kant, unlike the relativist, believed that all thinkers necessarily apply the same categories or conceptual schemes to the undifferentiated flow of their experiences. However, once the distinction between a conceptual scheme and its content was introduced, it was easy to accept that there could be more than one system or scheme of organization – thus the idea of conceptual relativism was born. Conceptual relativists argue that, at least in principle, there is a multiplicity of ways of carving up the world, and the so-called 'facts' are delineated by human interests and modes of thought rather than the impersonal hand of nature. One of the most influential versions of conceptual relativism in the twentieth century was proposed by W.V.O. Quine, according to whom, to be able to talk about the world and to cope with our stream of sensory experience, we must impose on it a conceptual or linguistic scheme or theory. However, given the underdetermination of theory by available data, it is possible to envisage a plurality of conceptual or theoretical frameworks, all of which explain and predict our sensory experiences equally well. There are no determinate facts to enable us to choose between these varying frameworks; hence what Quine has called 'ontological relativity' ensues.

The influence of Quine's views is widespread. We can, in particular, see its impact on the writings of Nelson Goodman and Hilary Putnam who proposed very sophisticated forms of conceptual relativism. Putnam, for instance, has denied that there is a fact of the matter as to what counts as an object outside the standards set by a conceptual scheme. One and the same situation, according to Putnam, can be described "as involving entirely different numbers and kinds of objects" (1992: p. 122), and we do not have an account of which of these is the right description external to and independent of all conceptual schemes. It is important to note that most conceptual relativists, Quine and Putnam in particular, wish

to distance themselves from cultural relativism, which they think is fundamentally incoherent. Their brand of relativism is based on their theories of meaning and content rather than the purported differences between cultural traditions.

Relativism about Science

Conceptual relativism also informed the work of Thomas Kuhn and Paul Feyerabend, who emphasized the formative role of different paradigms and models of reasoning at various periods in the history of science. In their different ways, they have argued that scientific theories, developed within differing paradigms, are incommensurable or incomparable. Each paradigm acts as a unique and untranslatable language or conceptual framework, and since rival paradigms do not have access to a common criterion for adjudication, they remain self-contained and beyond the critical reach of their presumed rivals. Kuhn's and Feyerabend's writings, in turn, had a formative role on the Strong Programme of Sociology of Science of Barnes and Bloor (see 3.2.3), and the so-called 'Science Studies.' Conceptual relativism also informed the work of Thomas Kuhn and Paul Feyerabend, who emphasized the formative role of different paradigms and models of reasoning at various periods in the history of science. In their different ways, they have argued that scientific theories, developed within differing paradigms, are incommensurable or incomparable. Each paradigm acts as a unique and untranslatable language or conceptual framework, and since rival paradigms do not have access to a common criterion for adjudication, they remain self-contained and beyond the critical reach of their presumed rivals. Kuhn's and Feyerabend's writings, in turn, had a formative role on the Strong Programme of Sociology of Science of Barnes and Bloor (see 3.2.3), as well as on Bruno Latour and other social constructionists. According to the social constructionists, scientific facts, and even reality – or 'the world' with its objects, properties, and categories – are not simply out there to be discovered by scientists, rather they are constructed via communal norm-governed processes and practices such as negotiations, interpretations, and manipulation of data (as well as accidental and opportunistic developments). Scientific discoveries and theoretical knowledge are the products of socially sanctioned norms and practices and are guided by projects that are of cultural, economic, or political importance. As Latour and Woolgar put it, "Our point is that 'out-there-ness' is the consequence of scientific work rather than the cause" (1986: p. 180). The high point of the debate on relativism about science, or the 'Science Wars,' was the publication of the infamous 'Sokal Hoax.' In 1996, the postmodernist journal *Social Text* published an article by the physicist Alan Sokal where he mimicked relativist and postmodernist objections to the objectivity of science. Sokal then went on to confess that his paper was full of non sequiturs and was but a sham. The ensuing 'science wars' colored much of the discussions of relativism at the turn of the twenty-first century. The debates are still continuing, but they have lost their 'war-like' tone (see Ashman and Baringer, 2000).

Moral Relativism

Moral or ethical relativism is simultaneously the most influential and the most reviled of the relativistic doctrines. Supporters see it as a harbinger of tolerance, open mindedness, and antiauthoritarianism. Detractors, including Pope Emeritus Benedict XVI, think it undermines the very possibility of ethics and is a sign of either confused thinking or moral turpitude. Briefly stated, moral relativism is the view that right and wrong, good and bad are relative to and vary with individual or cultural perspectives and frameworks. The first position, where the individual's preferences and judgmental framework provide the context of relativization is known as subjectivism, while the latter, where cultural and social frameworks are the context of relativization, is moral relativism proper. Moral relativism can and often is embraced independently of cognitive relativism, for ethical relativists may agree that convergence between diverse worldviews and theoretical frameworks is possible in the natural sciences but deny that this is the case for values. There is a fundamental cleavage between scientific and moral enquiries, they argue, because moral precepts and judgments, unlike the posits of science, are not part of the natural furniture of the universe; they are man-made and would not exist independently of human actions, beliefs, and customs.

Moral inquiry is concerned with how life should be lived and how human beings should behave toward each other and their world. Moral relativists start from the premise that moral values are grounded on societal conventions, historical conditions, metaphysical beliefs, etc., which vary from one society or social grouping to another and argue that there are no neutral standards available to us for adjudicating between such competing claims. The truth or falsity, the appropriateness or inappropriateness of ethical beliefs, are relative to their socio-historical backgrounds, and moral beliefs cannot be assessed independently of their social framework. In that sense, it could be seen as a subspecies of cultural relativism (see Section Cultural Relativism).

Moral relativism is usually subdivided into descriptive, ethical and metaethical varieties. Descriptive moral relativism is an empirical doctrine based on observation of diversity in moral values across different societies and historic epochs. There is widespread disagreement among individuals and cultures on issues ranging from capital punishment to abortion, from human rights to animal rights. Descriptive relativists go so far as to accept that there may be broad similarities in the core beliefs of different societies, for instance, injunctions against the killing of innocents, but argue that such shared values are given very different and even incompatible interpretations by different cultures.

Ethical relativism is a normative position, informed by descriptive moral relativism, claiming that "it is wrong to pass judgment on those who have substantially different values, or to try to make them conform to one's values, for the reason that their values are as valid as one's own" (Wong, 1993: p. 442). In this, ethical relativism has much in common with cultural relativism and is similarly associated with tolerance and open mindedness (see Section Cultural Relativism). It is sometimes argued that ethical relativism is itself a normative doctrine insofar as it prohibits interference with the ways of life and moral practices of other cultures. But this degree of tolerance can also lead to moral indifference and even nihilism. A relativist does not have the certitude to condemn genocide, torture, or child abuse, for example, so long as such actions are seen as

part of the fabric of the social life of some cultures and peoples. Furthermore, if ethical relativism is true, then even tolerance and noninterference could not be seen as universal ethical values and could not be recommended to those who do not already embrace them.

Moral diversity is of course compatible with the possibility that there are right and wrong answers to ethical questions, but finding such answers would depend on having a dependable path toward the discovery of ethical truths or the establishment of a universal framework for arbitrating between incompatible moral claims. Metaethical relativists deny this possibility because they claim that the truth or the justifiability of ethical judgments is internal to the ethical framework agents embrace. Conflicting moral judgments, they argue, can be equally correct or equally justified because what decides their correctness is the moral standard to which their judgment is indexed (see, for instance, Gilbert Harman, 2010; Dreier, 2006).

Semantic Relativism

Semantic relativism, also known as 'new relativism' or 'truth-relativism,' has arisen out of work in the philosophy of language in the analytic tradition. Semantic relativism is the view that some statements can be characterized as true only relative to a context of assessment. John MacFarlane, a champion of this new approach, claims that the *truth*-conditions of utterances for a certain class of statements, e.g., disputes about taste and values, future contingent propositions, and attributions of knowledge, may be appropriately treated as true relative not just to their context of use but also to the *contexts of assessment*. A statement can be true relative to one 'context of assessment' and false relative to another (MacFarlane, 2005, 2007). This sort of relativism is seen as useful for explaining away what has become known, after Max Kölbel, as instances of faultless disagreement. Kölbel (2004) motivates truth-relativism by finding faultless disagreements in contexts that lack objectivity but which are nevertheless truth-apt. A disagreement is faultless where one person asserts a proposition p and the other denies or disputes it and asserts negation or not p (the condition of disagreement), but neither party could be said to have made a mistake in their assertions (the condition of faultlessness). In such circumstances, we have to say that both parties are correct, and therefore both their beliefs p and not p must be true. The truth-relativists' way to deal with this scenario is to say that each belief is true relative to a perspective. To take an example, in a situation where A asserts that "rhubarb pies are delicious" and B disputes this by asserting the contrary, i.e., "Rhubarb pie is not delicious," both A and B could be thought of as saying something true if truth is relativized to a taste parameter. A is asserting something true with respect to her standard of taste and false relative to B's (MacFarlane, 2007; Wright, 2008). One strength of truth-relativism is that it retains our intuition that some judgments are subjective and yet open to genuine disagreement, in so far as they have shared content. The position, however, has come under attack, among other reasons, for its failure to respect deep-seated intuition about truth as a norm of assertion (Greenough, 2010; Baghramian and Hamilton, 2011). The position has also come under attack by those with equally deep-seated intuitions that truth is fundamentally a monadic and nonrelative property of propositions (e.g., Cappelen and Hawthorne, 2009).

Arguments against Relativism

Critics of relativism have found the doctrine incoherent. The charge of incoherence has been leveled in different ways, the most common of which is the charge of inconsistency. The argument goes back to Plato whose paradigmatic argument against the relativist Protagoras amounted to the claim that relativism, at least in its stronger global form of relativism about truth, is either inconsistent or incoherent because of the dubious status of the claim that "truth is relative." If "truth is relative" is itself true unconditionally, it is argued, then there is at least one truth that is not relative, and hence relativism is not true. If, on the other hand, "truth is relative" is true only relative to the relativist's framework, then the relativist's advocacy of his position is devoid of significance. To avoid inconsistency a relativist should accept that "truth is relative" is itself only relatively true, i.e., true only for people who endorse it, but then she or he would not be in a position to convince an antirelativist of the legitimacy of the claim. The relativist is condemned to silence.

Variants of the charge of incoherence have been proposed. Hilary Putnam, for instance, has argued that relativism is ultimately incoherent because it cannot distinguish between being right and thinking that one is right. And without such a distinction, thinking and speaking become impossible (Putnam, 1981). Paul Boghossian (2006a,b), on the other hand, has focused largely on social constructionist arguments for epistemic relativism. He argues that relativism fails because of the incoherence of its claim of "equal validity," the claim that "There are many radically different, yet 'equally valid' ways of knowing the world, with science being just one of them" (Boghossian, 2006a,b: p. 2). To accept the doctrine of "equal validity," Boghossian argues, is to give up on all our pretheoretic ground level intuitions on truth, knowledge, and justification.

A different version of the accusation of incoherence is directed at the intelligibility of the relativist's claim that one and the same statement may be true for one linguistic community and false for another. Donald Davidson, the main advocate of this position, argues that all types of relativism rely on the assumption that there can be radically divergent conceptual schemes or systems of belief. But in order to be in a position to make the claim that a system of belief or cognitive scheme diverges from ours in a significant manner, we first need to identify such a scheme and the beliefs that result from it. But in order to identify something as a belief, Davidson argues, we must be able to interpret it (or translate the sentence expressing that belief into our language). To translate a sentence or to assign meaning to it, Davidson maintains, we will have to regard that sentence, or the belief system underpinning it, as true or in agreement with us. A correct understanding of the beliefs and other propositional attitudes of a person, Davidson maintains, leads to the conclusion that most of her beliefs must be true. So the relativist cannot get its claim of incommensurability or radical diversity among different belief systems off the ground, because the recognition of such

diversity requires charity in interpretation and presupposes that there are some identifiable commonalities between different languages and belief systems. If this argument is right, then relativism cannot even get started, because arguments for relativism are built on the premise that there are, or can be, fundamental cultural and conceptual differences between human beings.

Relativism has survived 2000 years of attempts to bury it. Its perennial appeal and periodic fashionability are due to its ability to capture and express an essential insight, be it in an exaggerated fashion, about the human condition. The insight that we live in a world of pluralities and do not have access to univocal answers to our diverse problems.

See also: Ethics and Values; Evans-Pritchard, Sir Edward E. (1902–73); Fieldwork in Social and Cultural Anthropology; Irrationality: Philosophical Aspects; Kant, Immanuel (1724–1804); Kuhn, Thomas S. (1922–96); Nietzsche, Friedrich (1844–1900); Pluralism; Wittgenstein, Ludwig (1889–1951).

Bibliography

Ashman, K., Baringer, P. (Eds.), 2000. After the Science Wars. Routledge, New York, London.
Baghramian, M., 2004. Relativism: The Problems of Philosophy Series. Routledge, New York, London.
Baghramian, M., 2013. Relativism about science. In: The Routledge Companion to Philosophy of Science, second ed. Routledge, New York, London.
Baghramian, M., 2010. Relativism: a brief history. In: Krausz, M. (Ed.), Relativism: A Contemporary Anthology. Columbia University Press, New York.
Baghramian, M., Hamilton, R., 2011. Relativism and the norm of truth. Trópo, Rivista di Ermeneutica e Critica Filosofica III (1), 33–51.
Barnes, B., Bloor, D., 1982. Relativism, rationalism and the sociology of knowledge. In: Hollis, M., Lukes, S. (Eds.), Rationality and Relativism. Basil Blackwell, Oxford, pp. 21–47.
Bloor, D., 1976. Knowledge and Social Imagery. Routledge & Kegan Paul, London.
Boghossian, P., 2006a. Fear of Knowledge: Against Relativism and Constructivism. Oxford University Press, Oxford.
Boghossian, P., 2006b. What is relativism? In: Greenough, P., Lynch, M.P. (Eds.), Truth and Realism. Oxford University Press, Oxford, pp. 13–37.
Cappelen, H., Hawthorne, J., 2009. Relativism and Monadic Truth. Oxford University Press, Oxford.
Carter, J.A. Epistemic Relativism. Palgrave MacMillan, Basingstoke, Hampshire, forthcoming.
Code, L., 1993. Taking subjectivity into account. In: Alcoff, L., Potter, E. (Eds.), Feminist Epistemologies. Routledge, New York, London, pp. 15–48.
Davidson, D., 1984. Inquiries into Truth and Interpretation. Oxford University Press, Oxford.
Dreier, J., 2006. Moral relativism and moral nihilism. In: Copp, D. (Ed.), The Oxford Handbook of Ethical Theory. Oxford University Press, New York, pp. 240–264.
Evans-Pritchard, E.E., 1993. Witchcraft, Oracles and Magic among the Azande. Clarendon Press, Oxford.
Feyerabend, P., 1987. Farewell to Reason. Verso, London.
Feyerabend, P., 1988. Against Method. Verso, London.
Foot, P., 1982. Moral relativism. In: Krausz, M., Meiland, J. (Eds.), Relativism: Cognitive and Moral. University of Notre Dame Press, Notre Dame, pp. 152–166.
Garcia-Carpintero, M., Kölbel, M. (Eds.), 2008. Relative Truth. Oxford University Press, Oxford.
Goldman, A., 2010. Epistemic relativism and reasonable disagreement. In: Feldman, R., Warfield, Ted A. (Eds.), Disagreement. Oxford University Press, Oxford, pp. 187–215.
Greenough, P., 2010. Relativism, assertion and belief. In: Brown, Cappelen (Eds.), Assertion. Oxford University Press, Oxford.
Haack, S., 1998. Reflections on relativism: from momentous tautology to seductive contradiction. In: Manifesto of a Passionate Moderate: Unfashionable Essays. University of Chicago Press, Chicago, IL.
Hales, S., 2006. Relativism and the Foundations of Philosophy. The MIT Press, Cambridge, MA.
Hales, S., 2011. A Companion to Relativism. Blackwell, Oxford.
Harman, G., Thomson, J.J., 1996. Moral Relativism and Moral Objectivity. Blackwell, Oxford.
Harman, G., 2010. Moral relativism defended. In: Krausz, M. (Ed.), A Contemporary Anthology. Columbia University Press, New York, pp. 31–50. Relativism.
Harre, R., Krausz, M., 1996. Varieties of Relativism. Blackwell, Oxford.
Hollis, M., Lukes, S. (Eds.), 1982. Rationality and Relativism. Basil Blackwell, Oxford.
Kölbel, M., 2002. Truth without Objectivity. Routledge, London.
Kölbel, M., 2003. Faultless disagreement. Proceedings of the Aristotelian Society 104, 53–73.
Krausz, M. (Ed.), 1989. Relativism: Interpretation and Confrontation. University of Notre Dame Press, Notre Dame, IL.
Krausz, M., Meiland, J.W. (Eds.), 1982. Relativism: Cognitive and Moral. University of Notre Dame Press, Notre Dame, IL.
Kuhn, T.S., 1962, 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago, IL.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton, NJ.
Laudan, L., 1990. Science and Relativism: Some Key Controversies in the Philosophy of Science. University of Chicago Press, Chicago, IL.
MacFarlane, J., 2005. Making sense of relative truth. Proceedings of the Aristotelian Society 105, 321–339.
MacFarlane, J., 2007. Relativism and disagreement. Philosophical Studies 132, 17–31.
Margolis, J., 1991. The Truth about Relativism. Blackwell, Oxford.
Moser, P.K., Carson, T.L. (Eds.), 2001. Moral Relativism: A Reader. Oxford University Press, Oxford.
Norris, C., 1997. Against Relativism: Philosophy of Science, Deconstruction and Critical Theory. Blackwell, Oxford.
O'Grady, P., 2002. Relativism. Acumen, Chesham.
Plato, 1961. Theaetetus. In: Hamilton, E., Cairns, H. (Eds.), The Collected Dialogues of Plato (F.M. Cornford, Trans.). Bollingen Series. Pantheon Books, New York, pp. 845–919.
Popper, K., 1994. The Myth of the Framework: In Defence of Science and Rationality. Routledge, London.
Putnam, H., 1981. Reason, Truth and History. Cambridge University Press, Cambridge.
Putnam, H., 1987. The Many Faces of Realism. Open Court, La Salle.
Putnam, H., 1992. Renewing Philosophy. Harvard University Press, Cambridge.
Quine, W.V.O., 1969. Ontological Relativity. In: Ontological Relativity and Other Essays. Columbia University Press, New York.
Quine, W.V.O., 1984. Relativism and absolutism. Monist 67, 293–296.
Rorty, R., 1982. Pragmatism, relativism and irrationalism. In: Rorty, R. (Ed.), Consequences of Pragmatism. University of Minnesota Press, Minneapolis, MN, pp. 160–175.
Rorty, R., 1991. Solidarity or Objectivity. In: Objectivity, Relativism, and Truth: Philosophical Papers, vol. 1. Cambridge University Press, Cambridge, UK.
Siegel, H., 1987. Relativism Refuted: A Critique of Contemporary Epistemological Relativism. D. Reidel Publishing Company, Dordrecht, The Netherlands.
Sokal, A., 1996. Transgressing the boundaries: towards a transformative hermeneutics of quantum gravity". Social Text 46/47, 217–252.
Sumner, W.G., 1906. Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals. Ginn and Co, Boston, MA.
Williams, B., 1975/2010. The truth in relativism. In: Krausz, M. (Ed.), Relativism: A Contemporary Anthology. Columbia University Press, New York.
Williams, M., 2007. Why (Wittgensteinian) contextualism is not relativism. Episteme, 4.
Winch, P., 1958. The Idea of a Social Science and Its Relation to Philosophy. Routledge & Kegan Paul, London.
Winch, P., 1964. Understanding a primitive society. American Philosophical Quarterly 1, 307–324.
Wong, D.B., 2006. Natural Moralities: A Defense of Pluralistic Relativism. Oxford University Press, Oxford.
Wright, C., 2007. New age relativism and epistemic possibility: the question of evidence. Philosophical Issues 17 (1), 262–283.
Wright, C., 2008. Relativism about truth itself: haphazard thoughts about the very idea. In: Garcia-Carpintero, M., Kölbel, M. (Eds.), Relative Truth. Oxford University Press, Oxford, pp. 157–186.

Reliability and Validity Assessment: New Approaches

Duane F Alwin, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The quality of measurement is considered the *sine qua non* of research in all sciences. In the social sciences, the concern with measurement quality is critical because theoretical constructs are unobserved and often elusive. In order to evaluate the quality of measurement in these fields, it is necessary to, first, develop a model that specifies the linkage between observations and the latent variable(s) of interest; second, to develop a research design that permits the estimation of parameters of such a model; and third, to develop an interpretation of these parameters that is consistent with the concepts of measurement quality. Several traditions exist for the formulation of such models that relate observed to unobserved variables, and the concepts of *reliability* and *validity* are often used to interpret the quality of measurement from such designs. This article discusses the traditional techniques of reliability and validity assessment and briefly presents some new approaches to evaluating the properties of individual measures that are aimed at measuring theoretical constructs of interest.

Introduction

Measurement represents the link between theory and research in all sciences, and the quality of measurement is the linchpin upon which rest our inferences about patterns and processes in the particular subject matter of interest. Almost everyone agrees that the quality of measurement is especially critical in the social and behavioral sciences because variables often lack precise definition and suitable indicators are often lacking. Indeed, it can be argued that without reliable and valid measurement, the contributions of these human sciences would be of little value, given that our inferences depend intimately on having high-quality tools for measurement. The examination of the question of the patterns of measurement error in such social data, often obtained through sample surveys and interview reports, and an evaluation of the quality of such data in general are strongly warranted.

In the social and behavioral sciences, the psychometric concepts of *reliability* and *validity* are essential ingredients to the assessment of data quality, and this is the focus of this article. Validity refers to *what we are measuring*, and it emphasizes the gathering of consensus surrounding 'validity claims' concerning the quality of indicators for assessing essential constructs. Reliability, on the other hand, refers to *how well we are measuring* whatever it is that our indicators specify as constructs of interest. The study of patterns of measurement reliability and validity typically requires repeated measurement (i.e., within-subjects designs), although many things can be learned from between-subjects designs as well. The present article is, however, limited to a focus on assessing reliability and validity using within-subjects measurement, specifying some newer emphases in these matters. I begin, however, with a brief discussion of the concepts of reliability and validity, and the traditional approaches to assessing the quality of measurement.

The Concepts of Reliability and Validity

The following discussion of the concepts of reliability and validity relies in part on a presentation of similar content in Alwin (2009, 2010, 2011). There are two important levels of discourse involved in the conceptualization of reliability and validity – *observed* (or manifest) variables and *unobserved* (or latent) variables. For our purposes, it is important to think of a response to a question in a measurement device to be an 'observed' variable that reflects some 'latent' unobserved variable or construct of interest. We may define the response variable as a random variable (Y), observed in some population of interest, which has a latent counterpart (T), which is also a random variable in the population, and which in part is responsible for the production of the response, i.e., $T \rightarrow Y$. In this sense, we may think of T as a part of (or a component of) Y. There is no necessary identity between the two, however, as we know there are a number of types of measurement errors, which we can represent globally using the notation E, also considered to be a part of Y. In this sense, we can conceptualize Y as having two major parts that contribute to the response, namely, what we are measuring and the errors that impede our ability to do so, i.e., $T \rightarrow Y \leftarrow E$.

Note that while Y is observed, both T and E are unobserved, or latent, variables. This type of formulation allows us to think in terms of two general considerations when examining the variation observed in responses to measures; one is the variation in the underlying phenomenon one is trying to measure, e.g., income or employment status, and the other is the variation contributed by errors of measurement. The consideration of each of these parts – the part concerning the construct one is trying to measure as well as the part concerning measurement errors – is the focus of discussions of reliability and validity.

It is important to note here that our formulation is intended to be very general, allowing for a range of different types of Y and T random variables, Y and T variables that represent discrete categories, as well as Y and T variables that are continuous variables. There are a number of different cases or types of situation involving different types of variables at each level. If Y and T are continuous random variables in a specified population of interest, for example, and if it is assumed that E is random with respect to the latent variable, that is, T and E are uncorrelated determinants of Y, we have the case that is represented by *classical true-score theory* (CTST) (see Lord and Novick, 1968; Alwin, 2007), where the T latent variable is

referred to as a true score and measurement error is simply defined as the difference between the true score and the observed score, i.e., $E = Y - T$ (see Groves, 1989).

In general, there is no constraint on the covariance properties of E and T, although in the CTST case, E is considered to be random, i.e., uncorrelated with T, and this assumption underlies the development of psychometric approaches to estimating reliability. Although we generally find the CTST to be quite valuable in many situations, it has certain limitations, especially when it comes to assessing the reliability and validity of discrete data, e.g., event history data (Belli et al., 2009). For this reason, for purposes of this discussion, we insist on a more general conceptualization, which permits T to be either a set of latent nominal classes, a set of latent ordered classes, or a latent continuous variable, *and* a conceptualization that permits Y to also be a nominal, ordinal, or interval response variable (see Alwin, 2007: pp. 128–130).

Before developing the idea of reliability further, however, it is important to add more complexity, in order to eventually incorporate the concept of reliability into the discussion of validity. Indeed, in addition to reliability, and perhaps of greater interest, is the concept of validity – the extent to which the investigator is measuring the theoretical construct of interest. In a more general sense, measurement validity refers to the extent to which measurement accomplishes the purpose for which it is intended: that is, is one measuring what one intends to measure? Validity is a somewhat more elusive concept than reliability, given the absence of adequate criteria for its evaluation, historically (see AERA/APA/NCME, 1999). Although matters of validity are often thought to be capable of solutions on an abstract level, as with the notion of 'content' or 'face' validity, on an empirical level the issue of validity can only be defined with respect to some criterion that is related to the purpose of measurement, and we often must rely on 'validity claims,' supported by consensus in the scientific community that some X is a valid measure of construct C. When the validity of measurement is at issue, this if often assessed with respect to its correlation with criterion variables, often outside the particular research design, that are linked to records of factual information. In addition, validity is often defined at the level of the construct, that is, the relationship between T (or the latent variable being measured) and the construct of interest, C, i.e., $C \rightarrow T$. Of course, without a direct measure of this 'construct of interest,' one often must fall back on notions of *criterion validity*.

I should note at this point that, although some will mistake it for validity, when traditional psychometric approaches refer to the reliability of measurement *on a conceptual level*, they refer to the closeness of the correlation between Y and T. Thus, when one is dealing with measurements that are reliable, it is assumed that the measurement error part of Y is minimal, and in a 'variance accounting' framework, mostly due to latent or 'true' variation. By contrast, when measures are unreliable, it is because there is a great deal more 'error of measurement' than 'true variation,' that is, much more 'noise' than 'signal,' to use a metaphor from radio telegraphy. Validity, on the other hand, refers to the extent of correlation of between T and C. High reliability is important in part because validity depends on reliability – without highly reliable measures, high validity is impossible. Reliability of measurement is a *sine qua non* of all social science research, and it is readily shown that reliability is a necessary although not sufficient condition for validity (Lord and Novick, 1968; see also Alwin, 2007: pp. 291–292). Specifically, it can be shown that the *criterion validity* of a measure – its correlation with another variable – cannot exceed the reliability of either variable. There is a mathematical proof for this assertion (Alwin, 2007: pp. 291–292), but the logic that underlies the idea is normally accepted without formal proof: if our measures are unreliable, they cannot be trusted to detect patterns and relationships among variables of interest, including the criterion variables one might use to assess validity.

To summarize, for our present purposes we take reliability to refer to the relationship between Y and T (as defined above), whereas validity has to do with the relationship between T and C, that is, between the latent variable being measured and the theoretical construct of interest. When one has a direct record of the variable or a 'gold standard' for this theoretical construct (e.g., employer records of earnings), then one can examine the relationship between T and C. Such designs are relatively rare, but when available, they can be very useful (I return to a brief discussion of these below). Both measures of 'gold standards' and of 'alternative methods' contain measurement error, so such correlations or measures of agreement must always be interpreted in light of reliability of measurement in both variables.

Designs for Assessing Validity and Reliability

Most people agree that there are two basic approaches to dealing with errors of measurement resulting from invalidity and unreliability: (1) one can try to reduce errors in the collection of data, by developing more valid indicators and using more precise and consistent methods of gathering information, and/or (2) one can try to develop models of the processes that produce error. The purpose of the second strategy of modeling error is ultimately to provide feedback on how to reduce errors in data collection, by altering the questionnaire strategies used to collect data, or by abandoning measurement approaches that are fraught with error. The present article focuses solely on the second strategy, namely, question evaluation methods as they are applied in survey research (see Madans et al., 2011).

More than 50 years ago, psychologists Donald Campbell and Donald Fiske (1959) proposed some ideas that continue to express some fundamental principles with respect to assessment of reliability and validity, which are still useful. They argued that both concepts of reliability and of validity require a high level of agreement between measurements – *validity* is supported when there is correspondence between two efforts to measure the same trait through *maximally different* methods; *reliability* is demonstrated when there is agreement between two efforts to assess the same thing using *maximally similar*, or replicate, measures (Campbell and Fiske, 1959: p. 83). Thus, both types of studies involve assessing the correspondence between measurements – studies of reliability focus on the consistency of repeating the measurement using replicate measures, whereas researches on validity are concerned with the correspondence of a given measure to some criterion of interest, taking into account the reliability of either measure.

Articulating these ideas further, Alwin and Krosnick (1991: p. 141) suggested three desiderata for assessing the reliability and validity in the case of survey data:

- First, a model is needed that specifies the linkage between true and observed variables.
- Second, a research design is needed that permits the estimation of parameters of such a model.
- Third, one must be able to offer an interpretation of these parameters that is consistent with the concepts of measurement quality (reliability and validity).

We have already discussed above the general ideas involved in specifying the linkage between true and observed variables. In the following discussion, we focus specifically on the types of research designs that are used to permit the estimation of the parameters of models for reliability and validity assessment.

Validation Studies

Establishing the validity of survey measurement is difficult because within a given survey instrument, there is typically little available information that would establish a criterion for validation. However, there is a well-known genre of research situated under the heading 'record check studies,' which involves the comparison of survey data with information obtained from other record sources. Although rare, such studies can shed light on survey measurement errors; however, as noted earlier, correlations among multiple sources of information are limited by the level of reliability involved in reports from either source (Alwin, 2007: pp. 48–49). As noted earlier, in the psychometric tradition, the concept of validity has mainly to do with the utility of measures with respect to getting at particular theoretical concepts. This concept of validity is difficult to adapt to the case of survey measures, given the absence of well-defined criteria representing the theoretical construct, but several efforts have been made. One important design that has been used for studying validity involves record check studies, where the investigator compares survey reports with a set of record data that may exist for the variables in question in a given population (e.g., Bound et al., 1990; Marquis, 1978).

Another design that has been used in the study of nonfactual content, e.g., attitudes, beliefs, and self-perceptions, involves the application of multitrait-multimethod (MTMM) measurement designs, where the investigator measures multiple concepts using multiple methods of measurement. In recent years, a great deal of attention has been paid to a model originally proposed by Campbell and Fiske (1959), involving the measurement of multiple concepts or 'traits' each measured using multiple methods – hence, the MTMM measurement design (see Alwin, 2007; Saris and Gallhofer, 2007). Using this design, methods have been developed for separating validity and invalidity in survey reports. Its use has been limited to the study of components of variance in measures of attitudes, beliefs, and self-descriptions primarily because there are often not multiple measurement formats for gathering factual data. I return to a discussion of this approach later in this article.

A third type of validation study is what has been referred to as 'gold standard' studies, where an alternative approach to measurement, e.g., an event history calendar, is compared to data gathered on the same people using the method that represents the *gold standard* or the accepted or standard approach to measuring particular content (see, e.g., Belli et al., 2001). All of these qualify under the definition of validity studies, in that they all three are aimed at examining the correspondence between responses to a particular survey question and the 'true' or 'best' indicator of the construct it is intended to measure. All of these validation designs embody the principle articulated by Campbell and Fiske (1959), that the best evidence for validity is the convergence of measurements employing maximally different methods.

Studies of Reliability

With respect to studies of reliability, there are several standard methods of reliability estimation in common use in survey research; however, many of these approaches are inadequate for purposes of dealing with the evaluation of the reliability of specific measures, as will be discussed below. Strategies of reliability estimation for continuous variables are well developed, and we briefly consider these models here – they are discussed in detail elsewhere (see Alwin, 2010). The basic model is derived from a set of definitions about the true score of a fixed person as an expected value of that person's propensity distribution (see Lord and Novick, 1968). The model is typically written as follows:

$$y_{ij} = \tau_{ij} + \varepsilon_{ij}$$

where y_{ij} is an observed score for person i on measure j, τ_{ij} is the true-score counterpart for person i on measure j, and ε_{ij} is measurement error, i.e., the difference between y_{ij} and τ_{ij} for person i on measure j. The assumption of this model is that in a population of persons for whom this model holds, measurement error is random, i.e., ε_{ij} and τ_{ij} are uncorrelated. Stated more formally, $E(\varepsilon_{ij}) = E(\varepsilon_{ij},\tau_{ij}) = 0$. The *reliability of measurement* for y_j in a fixed population, S, is defined as the ratio of the true variance to the observed variance, i.e., $\mathrm{VAR}(\tau_{ij})/\mathrm{VAR}(y_{ij})$, that is, the proportion of the observed variance that is true variance (see Alwin, 2007).

Simple Reliability Models

The classical literature on reliability estimation refers to "the correlation between truly parallel measurements taken in such a way that the person's true score does not change between them is often called the *coefficient of precision*" (Lord and Novick, 1968: p. 134). In this case, the only source contributing to measurement error is the unreliability or imprecision of measurement. The assumption here, as is expected in all such designs, is that "if a measurement were taken twice and if no practice, fatigue, memory, or other factor affected repeated measurements," the correlation between the measures reflects the precision, or reliability, of measurement (Lord and Novick, 1968: p. 134). For two measures, it is possible to relax the assumption of parallel measures and assume only that tau-equivalence holds in order to obtain an estimate of reliability for each of the two measures. Using the above formulation, the reliability of measurement is defined as $\mathrm{COV}(y_1y_2)/\mathrm{VAR}(y_1)$ and $\mathrm{COV}(y_1y_2)/\mathrm{VAR}(y_2)$ in the case of two tau-equivalent

measures, y_j (where $j = 1$ and $j = 2$). The problem, of course, is to be able to design studies that permit the interpretation of this quantity in terms of the parameters of the model.

Internal Consistency Reliability

Cronbach (1951) developed a measure for a composite score of what is often referred to as 'internal consistency reliability' (ICR), which estimates the lower bound of reliability for a set of univocal measures that are tau-equivalent in structure. Let Y symbolize such a linear composite defined as the sum $Y_1 + Y_2 + \cdots Y_g + \cdots + Y_G$, where G is the number of measures. The formula for the ICR (denoted as α) is given as follows:

$$\alpha = \frac{G}{G-1}\left[1 - \frac{\sum_g \mathrm{VAR}(Y_g)}{\mathrm{VAR}(Y)}\right]$$

Jöreskog (1971) reformulated this coefficient for the case of congeneric measures, and Greene and Carmines (1979) generalized all of these models to the general case of linear composites (see also Heise and Bohrnstedt, 1970). It can be shown (although we do not do so here) that this formula is derived from the definition of a classical true-score model of the ratio of the true variance to the observed variance of a composite variable, i.e., VAR(T)/VAR(Y) (see Alwin, 2007: pp. 51–53). Despite their popularity, these ICR models are not technically estimates of reliability in models where univocity does not hold, and given this relatively stringent set of requirements on the design of survey measures, they are often unrealistic for purposes of assessing the reliability in survey measurement. These methods assume the existence of multiple univocal measures of a single underlying variable and independence in the errors of measurement. It is difficult to obtain multiple measures of the same construct in the typical cross-sectional survey, and even when researchers fashion their measures as suitable for the application of psychometric models for reliability estimation, it is difficult to demonstrate that the assumptions required for these models are met (Alwin, 2007).

Common Factor Model Representation of CTST

It is a straightforward exercise to express the basic elements of CTST as a special case of the metric (unstandardized) form of the common factor model and to generalize this model to the specification of K sets of congeneric measures, which makes coefficient α somewhat more comprehensible. Consider the following common factor model:

$$Y = \Lambda T + E$$

where Y is a $(G \times 1)$ vector of observed random variables, T is a $(K \times 1)$ vector of true-score random variables measured by the observed variables, E is a $(G \times 1)$ vector of error scores, and Λ is a $(G \times K)$ matrix of regression coefficients relating true and observed random variables. The covariance matrix among measures under this model can be represented as follows:

$$\Sigma_{YY} = \Lambda\Phi\Lambda' + \Theta^2$$

where Σ_{YY}, Φ, and Θ^2 are covariance matrices for the Y, T, and E vectors defined above, and Λ is the coefficient matrix as defined above.

For purposes of this illustration, we consider all variables to be centered about their means. Here we take the simplest case where $K = 1$, that is, all G variables in Y are measures of T; however, we note that the model can be written for the general case of multiple sets of congeneric measures. In the present case, the model can be represented as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_G \end{bmatrix} = \begin{bmatrix} \lambda_{1T} \\ \lambda_{2T} \\ \vdots \\ \lambda_{GT} \end{bmatrix} \times [T] + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_G \end{bmatrix}$$

For G congeneric measures of T, there are $2G$ unknown parameters in this model ($G\lambda_{gT}$ coefficients and G error variances, θ_g^2) with degrees of freedom (df) equal to $0.5G(G+1) - 2G$. In general, this model requires a scale be fixed for T, since it is an unobserved latent random variable. Two options exist for doing this: (1) the diagonal element in Φ can be fixed at some arbitrary value, say 1.0, or (2) one of the λ_{gT} coefficients can be fixed at some arbitrary value, say 1.0. For G measures of T, the tau-equivalent model has $G + 1$ parameters with d$f = 0.5G(G+1) - G + 1$. For a set of G parallel measures, there are two parameters to be estimated, VAR[T] and VAR[E], with d$f = 0.5G(G+1) - 2$. Note that both the tau-equivalent measures and parallel measures form of this model invoke the assumption of tau-equivalence. This is imposed on the model by fixing all λ_{gT} coefficients to unity. In order to identify the tau-equivalent or parallel measures model, observations on two measures, Y_1 and Y_2, are sufficient to identify the model. For the congeneric model, G must be greater than or equal to 3. It should be clear that the congeneric measures model is the most general and least restrictive of these models, and the tau-equivalent and parallel measures models simply involve restrictions on this model.

What has been stated for the model in the above paragraphs can be generalized to any number of G measures of any number of K factors. The only constraint is that the assumptions of the model – univocity and random measurement error – are realistic for the measures and the population from which the data come. Although there is no way of testing whether the model is correct, when the model is overidentified the fit of the model can be evaluated using standard likelihood ratio approaches to hypothesis testing within the confirmatory factor analysis framework. There is, for example, a straightforward test for whether a single factor can account for the covariances among the G measures. Absent such confirming evidence, it is unlikely that a simple true-score model is appropriate, and a different interpretation must be given to the ICR estimate.

To summarize, there are two questionable assumptions of the CTST approach. The first is that the measures are *univocal* measures of a single underlying variable. The second is that the errors of the measures are independent of one another. These assumptions rule out, for example, the operation of memory in the organization of responses to items in a series of questions. Obviously, respondents are fully cognizant of the answers they have given to previous questions, so there is

the possibility that memory operates to an unknown degree to distort the degree of consistency in responses. These assumptions for reliability estimation in cross-sectional studies also rule out the operation of other types of correlated errors, for example, the operation of systematic method factors. Thus, in cross-sectional data it may be impossible to assume that measurement errors in a set of measures are independent, since similar questions are often given in sequence or at least included in the same battery. Given the shortcomings of the ICR approaches, attention has turned to more closely examining the sources of variation at the level of the individual survey question (see Alwin, 2010). Two variants on the basic CTST model have been developed for this purpose, which I briefly summarize in the following: the MTMM measurement design using cross-sectional data and the quasi-simplex approach for longitudinal data.

Reliability Models for Single Measures

Increasingly, research on survey methodology has been moving in the direction of assessing the reliability of single survey questions rather than composite variables (e.g., Groves, 1989; Saris and Andrews, 1991; Saris and van Meurs, 1990; Saris and Gallhofer, 2007; Scherpenzeel and Saris, 1997; Alwin, 1989, 1992, 1997, 2005, 2007; Alwin and Krosnick, 1991). Second, the idea of using coefficients that assume a composite measure of the underlying variable is unrealistic, and certainly so for assessing the reliability (or validity) of many types of latent variables. Standard internal consistency approaches, e.g., coefficient α, make no sense in most cases as an estimate of reliability of measurement.

MTMM Models

There is an increasing amount of support for the view that shared method variance inflates ICR estimates. One approach to dealing with this is to reformulate the CTST along the lines of a multiple-factor approach and to include sources of systematic variation from *both* true variables and method factors. With multiple measures of the same concept, as well as different concepts measured by the same method, it is possible to develop an *MTMM matrix* (see **Table 1**) and formulate a model that permits the decomposition of the observed variance in a unique trait–method combination into those parts that can be interpreted as *validity*, *invalidity*, and random error, or *unreliability*. The design requirements are stringent, in that one needs the measurement of K traits measured by each of M methods (generating $G = KM$ observed variables) (see Campbell and Fiske, 1959; Alwin, 1974). If such a design can be implemented, this allows the specification of a model such as that given in the following.

In order to specify the model for the MTMM design, we can formulate an extension of the common factor representation of the CTST given above as follows:

$$Y = \Lambda_{T^*} T^* + \Lambda_M M + E$$

where Y is a ($G \times 1$) vector of observed random variables, T^* is a ($K \times 1$) vector of 'trait' true-score random variables, M is an ($M \times 1$) vector of 'method' true-score random variables, and E is a ($G \times 1$) vector of error scores (see **Figure 1**). The matrices Λ_{T^*} and Λ_M are ($G \times K$) and ($G \times M$) coefficient matrices, respectively, containing the regression relationships between the G observed variables and the K and M latent trait and method latent variables. Note that with respect to the CTST model given above, $\Lambda T = \Lambda_{T^*} T^* + \Lambda_M M$. The covariance structure for the model can be stated as:

$$\Sigma_{YY} = [\Lambda_{T^*} | \Lambda_M] \Phi_T [\Lambda_{T^*} | \Lambda_M]' + \Theta^2$$

where Φ_T has the following structure:

$$\Phi_T = \left[\begin{array}{c|c} \Phi_{T^*} & 0 \\ \hline 0 & \Phi_M \end{array} \right]$$

Note that the specification of the model places the constraint that the trait and method factors are uncorrelated. The estimation of this model permits the decomposition of reliable variance in each observed measure into valid and invalid parts, which we do not have the space to discuss further here. Suffice it to say that there is a growing literature surrounding the application of these methods (see Alwin, 2007, 2011; Saris and Galhofer, 2007; Scherpenzeel and Saris, 1997).

Test–Retest Reliability

In past years, the 'test–retest' method was considered a valuable design for assessing reliability (see Siegel and Hodge, 1968), although early on there have been serious issues raised about its utility (Moser and Kalton, 1972). In some cases, it can be a quite valuable design; however, when the reinterview time period is so short that respondents can remember their earlier response, or when the reinterview period is so long that the respondents change, there are problems (see Alwin, 2007: pp. 96–101). It is probably worth noting in this regard that the test–retest design is not mentioned in the current set of standards for educational and psychological testing (AERA/APA/NCME, 1999). Investigators interested in the study of measurement error in survey data using longitudinal methods have turned to alternative designs and other models for reliability estimation, as we discuss in the following section.

Longitudinal Methods

Despite the critiques of the test–retest approach, it can be argued that for purposes of assessing the reliability of survey data, longitudinal data provide an optimal design (see Alwin, 2007). Indeed, the idea of replication of questions in panel studies as a way of getting at measurement consistency has been present in the literature for decades (see Heise, 1969; Wiley and Wiley, 1970; Jöreskog, 1970; Saris and Andrews, 1991). The limitations of the test–retest design are well-known, but they can be overcome by incorporating three or more waves of data separated by lengthy periods of time (see Alwin, 2007: pp. 96–116). The multiple-wave reinterview design discussed in this article goes well beyond the traditional test–retest design (see Moser and Kalton, 1972: pp. 353–354), and specifically, by employing models that permit change in the underlying true score (using the

Table 1 Multitrait-multimethod matrix for three traits and three methods

		Method 1			Method 2			Method 3		
	Trait	X	Y	Z	X	Y	Z	X	Y	Z
1	X	$r_{X_1 X_1}$								
	Y	$r_{X_1 Y_1}$	$r_{Y_1 Y_1}$							
	Z	$r_{X_1 Z_1}$	$r_{Y_1 Z_1}$	$r_{Z_1 Z_1}$						
2	X	$\mathbf{r_{X_1 X_2}}$	$r_{X_2 Y_1}$	$r_{X_2 Z_1}$						
	Y	$r_{X_1 Y_2}$	$\mathbf{r_{Y_1 Y_2}}$	$r_{Y_2 Z_1}$						
	Z	$r_{X_1 Z_2}$	$r_{Y_1 Z_2}$	$\mathbf{r_{Z_1 Z_2}}$						
3	X				$\mathbf{r_{X_3 X_2}}$	$r_{X_3 Y_2}$	$r_{X_3 Z_2}$	$r_{X_3 X_3}$		
	Y				$r_{X_2 Y_3}$	$\mathbf{r_{Y_2 Y_3}}$	$r_{Y_3 Y_2}$	$r_{X_3 Y_3}$	$r_{Y_3 Y_3}$	
	Z				$r_{X_2 Z_3}$	$r_{Y_2 Z_2}$	$\mathbf{r_{Z_2 Z_3}}$	$r_{X_3 Z_3}$	$r_{Y_3 Z_3}$	$r_{Z_3 Z_3}$

Monomethod blocks

Heteromethod blocks

Note: Values in validity diagonals (MTHM) are in boldface type.

Source: Alwin, D.F., 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: Costner, H.L., (Ed.), Sociological Methodology 1973–74. Jossey-Bass, San Francisco, pp. 79–105.

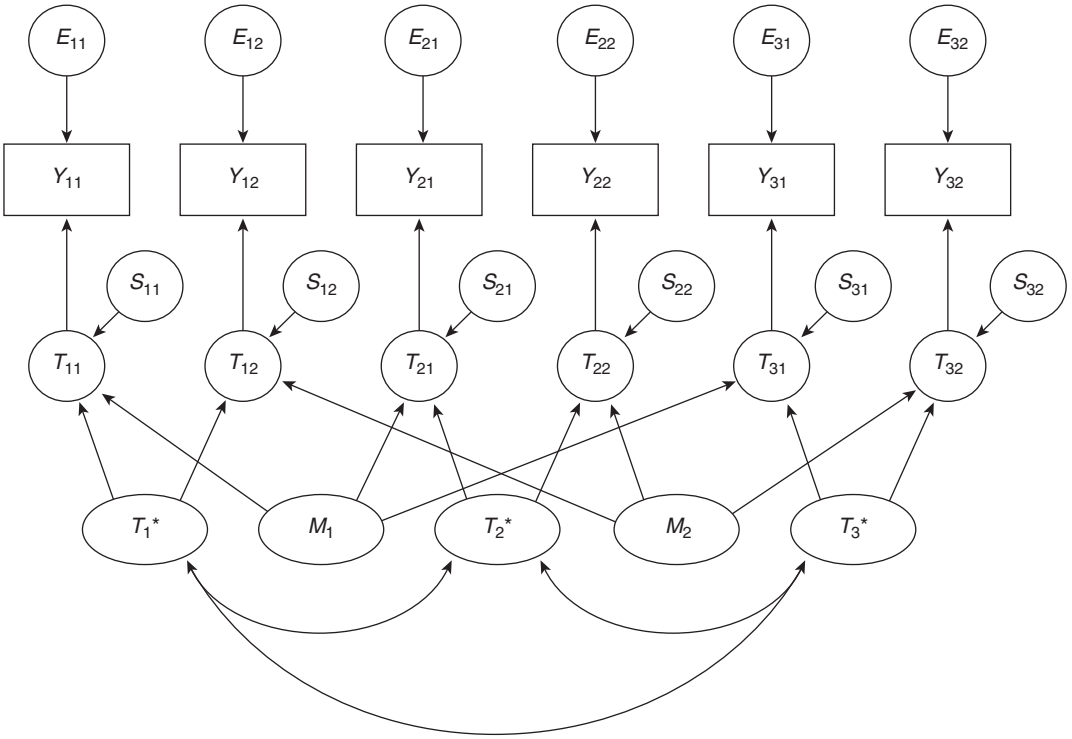

Figure 1 Path diagram for a common factor representation of the multitrait-multimethod design for six observed variables. Source: Alwin, D.F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. John Wiley and Sons, New York.

quasi-Markov simplex approach) allows us to overcome one of the key limitations of the test–retest design. And through the use of design strategies with relatively distant reinterview intervals, e.g., 2-year intervals, the problem of consistency due to retest effects or memory can be remedied, or at least minimized. There are two main advantages of the reinterview design for reliability estimation. First, the estimate of reliability obtained includes all reliable sources of variation in the measure, both common and specific variance. Second, under appropriate circumstances it is possible to eliminate the confounding of the systematic error component discussed earlier, if systematic components of error are not stable over time. In order to address the question of stable components of error, the panel survey must deal with the problem of memory, because in the panel design, by definition, measurement is repeated. So, while this overcomes one limitation of cross-sectional surveys, it presents problems like if respondents can remember what they say and are motivated to provide consistent responses. If reinterviews are spread over months or years, this can help rule out sources of bias that occur in cross-sectional studies. Given the difficulty of estimating memory functions, estimation of reliability from reinterview designs makes sense only if one can rule out memory as a factor in the covariance of measures over time, and thus, the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory.

The analysis of panel data in the estimation of reliability must be able to cope with the fact that people change over time, so that models for estimating reliability must take the potential for individual-level change into account (see Coleman, 1964, 1968; Goldstein, 1995; Wiggins, 1973). Given these requirements, techniques have been developed for estimating measurement reliability in panel designs where there are three or more waves, wherein change in the latent variable is incorporated into the model. With this approach, there is no need to rely on multiple indicators within a particular wave or cross-section in order to estimate the measurement reliability, and there is no need to be concerned about the separation of reliable common variance from reliable specific variance. That is, there is no decrement to reliability estimates due to the presence of specific variance in the error term; here specific variance is contained in the true score. This approach is possible using modern structural equation modeling (SEM) methods for longitudinal data and is discussed further in related literature (see Alwin, 1989, 1992, 2007; Alwin and Krosnick, 1991; Saris and Andrews, 1991).

In the general case of the quasi-simplex model for a single variable assessed in a multiwave panel study can be defined as:

$$Y = \Lambda_Y T + E$$

$$T = BT + Z$$

Here Y is a $(P \times 1)$ vector of observed scores, T is a $(P \times 1)$ vector of true scores, E is a $(P \times 1)$ vector of measurement errors, Z is a $(P \times 1)$ vector of disturbances on the true scores, Λ_Y is a $(P \times \mathrm{P})$ identity matrix, and B is a $(P \times P)$ matrix of regression coefficients linking true scores at adjacent time points. The general path diagram for the case of P overtime measures is given in Figure 2.

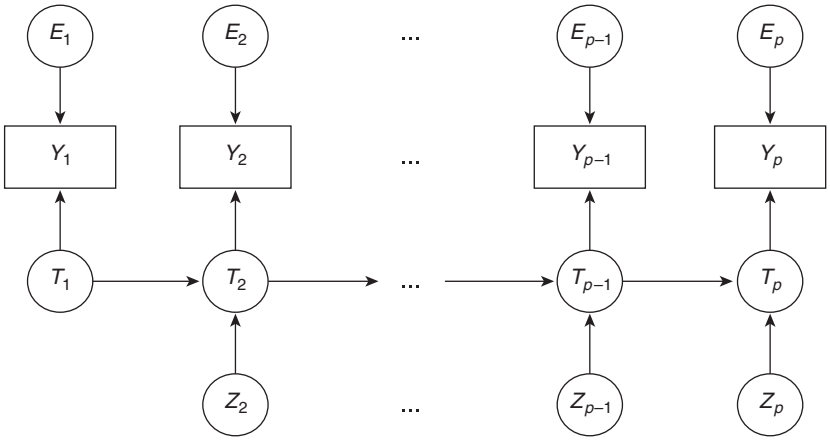

Figure 2 Path diagram of the quasi-Markov simplex model – general case ($P > 4$). Source: Alwin, D.F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. John Wiley and Sons, New York.

For the case where $P = 4$, we can write the matrix form of the model as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_{41} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{pmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -\beta_{21} & 1 & 0 & 0 \\ 0 & -\beta_{32} & 1 & 0 \\ 0 & 0 & -\beta_{43} & 0 \end{bmatrix} \times \begin{pmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix}$$

The reduced form of the model is written as:

$$Y = \Lambda_Y [I - B]^{-1} Z + E$$

and the covariance matrix for Y as:

$$\Sigma_{YY} = \Lambda_Y [I - B]^{-1} \Psi [I - B]^{-1} \Lambda_Y + \Theta^2$$

where B and Λ_Y are of the form described above, Ψ is a $(P \times P)$ diagonal matrix of variances of the disturbances on the true scores, and Θ^2 is a $(P \times P)$ diagonal matrix of measurement error variances. This model for multiple-wave data (where $P \geq 3$) can be estimated using any SEM approach (see Heise, 1969; Wiley and Wiley, 1970; Alwin, 2007).

Conclusions

The purpose of this article was to provide an overview of some of the newer strategies for assessing the validity and reliability of social science data. There are some traditional notions about implementing assessments of data quality based on composite scores (e.g., Cronbach's α coefficient) that have most likely outworn their usefulness in the modern era. More recently, researchers have become oriented to the evaluation of the reliability and validity of specific questions, and other tools are now viewed as more useful. Although this article has pointed the reader in the direction of some of the available techniques, there are a number of practical considerations that must be addressed before any method can be used to adduce interpretations of the reliability and validity of measurement.

See also: Classical (Psychometric) Test Theory; Content Validity; Measurement Theory: History and Philosophy; Psychometrics: Classical Test Theory; Reliability: Measurement; Structural Equation Modeling.

Bibilography

Alwin, D.F., 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: Costner, H.L. (Ed.), Sociological Methodology 1973–74. Jossey-Bass, San Francisco, pp. 79–105.

Alwin, D.F., 1989. Problems in the estimation and interpretation of the reliability of survey data. Quality and Quantity 23, 277–331.

Alwin, D.F., 1992. Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. In: Marsden, P.V. (Ed.), Sociological Methodology 1992. American Sociological Association, Washington, DC, pp. 83–118.

Alwin, D.F., 1997. Feeling thermometers vs. seven-point scales: which are better? Sociological Methods and Research 25, 318–340.

Alwin, D.F., 2005. Reliability. In: Kempf-Leonard, K., et al. (Eds.), Encyclopedia of Social Measurement. Academic Press, New York, pp. 145–166.

Alwin, D.F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. John Wiley and Sons, New York.

Alwin, D.F., 2009. Assessing the validity and reliability of timeline and event history data. In: Belli, R.F., Stafford, F.P., Alwin, D.F. (Eds.), Calendar and Time Diary Methods in Life Course Research. SAGE Publications, Thousand Oaks, CA, pp. 277–307.

Alwin, D.F., 2010. How good is survey measurement? Assessing the reliability and validity of survey measures. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research. Emerald Group Publishing, Ltd., Bingley, UK, pp. 405–434.

Alwin, D.F., 2011. Evaluating the reliability and validity of survey interview data using the MTMM approach. In: Madans, J., Miller, K., Maitland, A., Willis, G. (Eds.), Question Evaluation Methods – Contributing to the Science of Data Quality. John Wiley and Sons, Hoboken, NJ, pp. 265–293.

Alwin, D.F., Jackson, D.J., 1979. Measurement models for response errors in surveys: issues and applications. In: Schuessler, K.F. (Ed.), Sociological Methodology 1980. Jossey-Bass, San Francisco, pp. 68–119.

Alwin, D.F., Krosnick, J.A., 1991. The reliability of survey attitude measurement: the influence of question and respondent attributes. Sociological Methods and Research 20, 139–181.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Eduction, 1999. Standards for Educational and Psychological Testing. American Educational Research Association, Washington, DC.

Belli, R.F., Shay, W.L., Stafford, F.P., 2001. Event history calendars and question list surveys: a direct comparison of interviewing methods. Public Opinion Quarterly 65, 45–74.

Belli, R.F., Stafford, F.P., Alwin, D.F. (Eds.), 2009. Calendar and Time Diary Methods in Life Course Research. SAGE Publications, Thousand Oaks, CA.

Bound, J., Brown, C., Duncan, G.J., Rodgers, W.L., 1990. Measurement error in cross-sectional and longitudinal labor market surveys: validation study evidence. In: Hartog, J., Ridder, G., Theeuwes, J. (Eds.), Panel Data and Labor Market Studies. Elsevier Science Publishers.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 6, 81–105.

Coleman, J.S., 1964. Models of Change and Response Uncertainty. Prentice-Hall, Englewood Cliffs, NJ.

Coleman, J.S., 1968. The mathematical study of change. In: Blalock Jr., H.M., Blalock, A.B. (Eds.), Methodology in Social Research. McGraw-Hill, New York, pp. 428–478.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Goldstein, H., 1995. Multilevel Statistical Models, second ed. Arnold, London.

Greene, V.L., Carmines, E.G., 1979. Assessing the reliability of linear composites. In: Schuessler, K.F. (Ed.), Sociological Methodology 1980. Jossey-Bass, San Francisco, pp. 160–175.

Groves, R.M., 1989. Survey Errors and Survey Costs. John Wiley and Sons, New York.

Heise, D.R., 1969. Separating reliability and stability in test-retest correlation. American Sociological Review 34, 93–191.

Heise, D.R., Bohrnstedt, G.W., 1970. Validity, invalidity, and reliability. In: Borgatta, E.F., Bohrnstedt, G.W. (Eds.), Sociological Methodology 1970. Jossey-Bass, San Francisco, pp. 104–129.

Jöreskog, K.G., 1970. Estimating and testing of simplex models. British Journal of Mathematical and Statistical Psychology 23, 121–145.

Jöreskog, K.G., 1971. Statistical analysis of sets of congeneric tests. Psychometrika 36, 109–133.

Lord, F.M., Novick, M.L., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Madans, J., Miller, K., Maitland, A., Willis, G., 2011. Question Evaluation Methods–Contributing to the Science of Data Quality. John Wiley and Sons, Inc., Hoboken, NJ.

Marquis, K.H., 1978. Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations. The Rand Corporation, Santa Monica, CA.

Moser, C.A., Kalton, G., 1972. Survey Methods in Social Investigation, second ed. Basic Books, New York.

Saris, W.E., Andrews, F.M., 1991. Evaluation of measurement instruments using a structural modeling approach. In: Biemer, P.B., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (Eds.), Measurement Errors in Surveys. Wiley, New York, pp. 575–597.

Saris, W.E., Gallhofer, I.N., 2007. Design, Evaluation, and Analysis of Questionnaires for Survey Research. John Wiley and Sons, Hoboken, NJ.

Saris, W.E., van Meurs, A., 1990. Evaluation of Measurement Instruments by Metaanalysis of Multitrait Multimethod Studies. North-Holland, Amsterdam.

Scherpenzeel, A.C., Saris, W.E., 1997. The validity and reliability of survey questions: a meta-analysis of MTMM studies. Sociological Methods and Research 25, 341–383.

Siegel, P.M., Hodge, R.W., 1968. A causal approach to the study of measurement error. In: Blalock Jr., H.M., Blalock, A.B. (Eds.), Methodology in Social Research. McGraw-Hill, New York, pp. 28–59.

Wiggins, L.M., 1973. Panel Analysis: Latent Probability Models for Attitude and Behavior Processes. Elsevier Scientific Publishing Company, New York.

Wiley, D.E., Wiley, J.A., 1970. The estimation of measurement error in panel data. American Sociological Review 35, 112–117.

Reliability: Measurement

David Rindskopf, CUNY Graduate Center, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reliability of measurement refers to the dependability or repeatability of the measurement process: Does one get about the same answer when one measures a construct several times, or using similar methods of measurement, or using different people? Relevant theory comes from classical test theory, item response theory, generalizability theory, and related statistical methods. Classical test theory and generalizability theory are theories of the performance of a whole test or instrument (e.g., questionnaire), whereas item response theory deals with the behavior of individual test items as well as the whole test.

Introduction

Whenever an aspect of human behavior is measured, the question that always arises is: How good is the measurement process? In the behavioral sciences, conceptual and statistical models for answering that question are generally divided into two major categories: *reliability* and *validity*. As a first approximation, one may think of validity as being concerned with whether the right construct is being measured, and reliability as being concerned with the issue of how well we are measuring whatever it is that is being measured (regardless of whether or not it is the right quantity to measure). When choosing a measurement method, one must consider both of these issues; the total error of measurement comes from both the unreliability and invalidity of a measure. Although this article concentrates primarily on reliability, it also discusses the measurement process as a whole and the relationship between reliability and validity issues.

As with many terms, the scientific and psychometric uses of the word 'reliability' derive from, but are different from, the everyday use of the term. In everyday life, a car would be considered reliable if it started every time. If the word were used in the technical sense, one would also say that a car was reliable if it never started; it would be unreliable only if it sometimes started and sometimes did not. The technical term, in other words, is more concerned with the *consistency* of behavior and its measurement, regardless of where the behavior is on a scale. The main question is whether one gets the same answer every time a measurement of an individual is made. Whether that measurement is then useful is a question of validity; a measurement procedure could be perfectly reliable but useless.

Reliability theory generally begins with a conceptual framework, which is then translated more formally into a statistical model. Most reliability theory is concerned with a particular type of measurement instrument, multiple-item tests, and some theory is applied in an even more limited setting of items that are scored dichotomously as true or false. The three most commonly used models are (1) classical test theory (CTT), based on factor analysis; (2) generalizability theory (GT), based on analysis of variance (ANOVA); and (3) item response theory (IRT), based on latent trait models. Other statistical methods usually are needed for dealing with measurements that are either not tests, such as observations of behavior, or are not multiple-item tests. Much of measurement theory (and terminology) derived originally from consideration of ability tests, but the theory applies equally well to many other areas of behavior such as personality assessment.

Classical Test Theory

CTT is based on the assumption that a single underlying dimension (a trait or skill) is being measured, and that every person has a single true score, T, on that dimension. The underlying dimension or test domain may be more or less 'pure'; for example, one might think of skill in (1) doing multiplication of 3-digit numbers; (2) solving algebra word problems; or (3) solving college mathematics problems. The important consideration is that all items on the test are measuring this skill or trait, and that the skill or trait is well defined. Each person's observed test score, X, is assumed to be determined by his or her true score T plus some error of measurement, e, which would either raise or lower the observed score relative to the person's true score. The basic equation of CTT is, therefore,

$$X = T + e \qquad [1]$$

in which the quantities are as defined above, and the errors are assumed to have a mean of zero and to be uncorrelated with (unrelated to) T.

What exactly does T represent? Conceptually, the definition requires a bit of imagination and tolerance for ambiguity. Imagine a person being tested over and over again, using different sets of items selected from the collection of all items designed to measure the underlying construct. (A test may consist of any number of items, including the special case of a single item, such as an essay.) The average of all these scores (more technically, the expected value) is the person's true score T.

The first major consequence of this model is that the variation among people on observed test scores (σ_X^2) is the sum of the variation in true scores (σ_T^2) and the variation in the errors of measurement (σ_e^2); that is,

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2. \qquad [2]$$

The optimal case occurs when the error for each person is zero; then there is no variation in errors, so $\sigma_e^2 = 0$. Consideration of this equation leads to the definition of reliability as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.44050-X

the proportion of total variance in scores that is attributable to true variation rather than error variation; that is,

$$\text{Reliability} = \sigma_T^2/(\sigma_T^2 + \sigma_e^2) = \sigma_T^2/\sigma_X^2. \quad [3]$$

For example, if the true score variance were 8 and the error variance 2, the total variance would be 10, and the reliability would be $8/10 = .80$. One useful interpretation of this number is that it is the square of the correlation between true scores (T) and observed scores (X).

Among the interesting implications of the CTT model is that a test can be made more reliable by making it longer; adding items increases the true score variance at a faster rate than the error variance. One aspect of test design is determining how long a test must be to have a reliability great enough for the intended uses of the test.

Another implication of the model is that the reliability of a test is not a single number but depends on how individuals who took the test were sampled. If a broad range of people are sampled, so that the true score variance is large, then the reliability will be greater than if people with a more limited range of true scores are sampled. (This is because the error variance would be the same in each case.) For example, if a test were given to all high school seniors, it would have a higher reliability estimate than if it were given only to seniors who planned to attend a highly selective college or university.

Estimation of Reliability

Unfortunately, because true score and error variation are not directly observable, the definitional eqn [3] above cannot be applied directly. However, this conceptualization and the accompanying statistical model lead to several possible methods of estimating reliability. First, one could randomly divide any multiple-item test into two halves and see how the scores on one half relate to the scores on the other half. If a test were perfectly reliable, each person would get the same score on both halves of the test. Of course, the correlation is never perfect; the observed correlation allows the calculation of an estimate of the reliability of the whole test. Other methods use the correlations among the items to estimate reliability. These methods, including Cronbach's alpha (developed from an idea of Louis Guttman), are based on the same conceptual and statistical principles as the split-half coefficient.

Because each person is measured at one time point, any variability due to changes in that person across time are ignored by these methods. Here is where the definition of true score becomes critical. One can think of 'true' changes in true score (due to learning or forgetting in between testing times); one can also think of more temporary changes, due to illness, noise, or distractions in the testing room, and so on. In addition, some characteristics are expected to be variable over time, including measures of mood. In personality psychology, the more stable characteristics are called *traits*, while the more labile characteristics are called *states*. In CTT, the temporary effects are included in errors of measurement, whereas changes in traits generally are not. In spite of this conceptual uncertainty, this suggests another commonly used strategy for assessing reliability: Give the same test to a group of people at two different times and see whether their scores are closely related. For obvious reasons, this is called test-retest reliability. Of course, one assumes that the first measurement does not affect the second measurement, which may or may not be the case and which must be considered. To minimize the effects of 'true' changes in true scores over time, the time periods between testings must be kept relatively short.

A related method uses two or more forms of a test, with each form being constructed in an identical manner, so that the forms will be as nearly alike as possible. If in fact they have the same means, variances, and reliabilities, then they are said to be *parallel tests*. If one administers two such forms to the same group of people (e.g., on two consecutive days), the correlation between scores on the forms is called a parallel test estimate of the reliability.

In addition to measuring the overall performance of the test, one is also interested in how well the score of each individual has been estimated. The *standard error* of the test score is such a number; if small, the score is a more accurate estimate of ability than if the standard error is large. As an example, suppose a person has a score of 120 on an IQ test, and the test has a standard error of 5 points. Then we can construct a 95% confidence interval around the estimate by adding and subtracting 2 standard errors (10 points, in this case); here the interval is from 110 to 130. This means, roughly speaking, that there is a 95% chance that the interval will contain the person's true score.

Besides considering the reliability of the test as a whole, one must also consider how good each item on the test is. The most commonly used method to assess this in CTT is the item total correlation, which measures how well the item discriminates among people of varying skill or ability levels. To calculate this, one correlates scores on the total test (usually with the particular item omitted) with scores on the item (with 0 used for an incorrect response and 1 for a correct response). In other words, an item is good if people who score high on the test tend to get the item right more often than do people who score at lower levels on the test. An item would be considered bad if people with low scores on the test are just as likely to get the item right as people with high scores on the test. In the test construction process, items that have low-item total correlations are generally omitted or rewritten.

Another common item statistic is the proportion of test takers who get the item right; this is usually called the *difficulty* of the item (though the term *facility* would be more appropriate). Although it is commonly reported, it is of little relevance to reliability; item difficulties are more important when trying to tailor a test for various ability levels or to compare the characteristics of tests that are supposed to be equivalent (parallel, in technical terms).

Weaknesses of CTT

CTT has several weaknesses that have led to the development of other models for test scores. First, the concept of reliability is dependent on the group used to develop the test. If the group has a wide range of skills or abilities, the reliability will be higher than if the group has a narrow range of skills or abilities. Thus, reliability is not invariant with respect to the sample of test takers and is therefore not a characteristic of the test itself. In addition, neither are the common measures of item discrimination (such as the item total correlation) or item difficulty (percent getting the item correct). As if this were not bad enough, the usual assumption that the standard error is the same for test takers at

all ability levels is usually incorrect. (In some extensions of CTT, this assumption is dropped, but these extensions are not well known or widely used.) CTT also does not adequately account for observed test score distributions that have floor and ceiling effects, where a large proportion of test takers score at either the low or the high end of the test score range.

CTT has difficulties in handling some typical test-development problems, horizontal and vertical equating. The problem of *horizontal equating* arises when one wishes to develop another test with the same properties as (or at least with a known relationship to) an existing test. For example, students who take college admissions tests such as the ACT or SAT should get the same score, regardless of which version of the test they take. *Vertical equating* involves developing a series of tests that measure a wide range of ability or skill. For example, although one could develop completely independent tests of arithmetic for each elementary school grade, we might instead want to link them to have one continuous scale of mathematical skill for grades 1 through 6. Vertical and horizontal equating are not impossible within CTT, but they are much more straightforward with IRT.

These problems of CTT are caused partly by some fuzziness in the theory (the population to be sampled is usually not considered in any detail in the theory). But they are also due to the failure of most data collection to be carried out on a random sample of any population, let alone a population considered appropriate for the test being investigated. In practice, convenience samples are used. They generally fit some criteria that the investigator specifies but are otherwise taken as they are available to the researcher. Finally, CTT as originally conceptualized was never intended to address some of the practical testing problems described above.

Generalizability Theory

CTT provided a framework that has proved useful in a number of situations, but psychometricians realized many of its limitations even as they were developing the theory. Most important, they realized that a wide variety of factors can influence observed test scores, including how the items were generated, how the tests were administered, and how the items were scored (especially if they were free responses scored by raters). In CTT, all these factors were included in the (one and only) error term in the statistical model. To include these factors in their statistical models, psychometricians developed what is known as GT, which included more than one source of error in test scores.

GT is based on the statistical methods of ANOVA. Factors thought to have major influences on test scores are included in a design, and the analysis allows the researcher to determine how much influence each factor has on variation in test scores. GT allows the computation of various estimates of reliability, depending on which factors are considered important for a particular use of the test.

As an example, consider a test of writing in which each student writes three brief essays, each of which is scored by two raters. GT allows the estimation of how much variation in test scores is due to variation in essay topics and how much is due to variation in raters.

Like CTT, GT is concerned primarily with the behavior of the test as a whole, rather than the performance of components, such as subscores or items.

For further details on GT, see the article in this volume. The classic source on GT is the book by Cronbach et al. (1972). Brennan (2011) and Shavelson et al. (1989) provide a less technical overview.

Item Response Theory

The past few decades have seen a steady rise in the use of IRT to solve some of the problems that CTT could not deal with adequately. IRT, like CTT, is based on an assumption that a single underlying dimension is being measured. But the basic conceptual idea behind IRT is the relationship between the score (usually right/wrong) on an item and the underlying dimension, whereas CTT is based on the relationship between total test score and the underlying dimension. This relationship, represented by an item characteristic curve, or ICC, shows how the probability of getting the item right (for dichotomous items) changes with the underlying trait. The probability cannot be below zero or greater than one, so the ICC generally has an elongated S shape. In three-parameter models, the items may differ in whether the curve starts at zero for people of very low ability (no guessing) or above zero (if guessing correct is likely); how fast the curve rises (discrimination); and whether the curve is to the left (easy), in the middle, or to the right (hard). In two-parameter models, only discrimination and difficulty are used; guessing is fixed at zero. And for one-parameter (Rasch) models, only difficulty is used; guessing is fixed at zero and discrimination at one for all items.

IRT models solve many problems that were difficult for CTT, including the ability to construct a different test for each test taker, so that students at all levels of ability can be assessed with items appropriate to their status. Such adaptive tests, generally administered by computers, are becoming standard in many large-scale testing situations in some countries, including the Graduate Record Examination and licensing examinations in many fields.

In IRT, the traditional concept of reliability has no place. Instead, the emphasis is on the more basic concept of standard error of measurement; that is, how accurately has a test taker's true score been estimated? For a fixed test given to a group of examinees, the test characteristic curve shows the standard error for each level of ability. Unlike the simplifying assumption of CTT, the standard error can vary with ability. For example, if a test is relatively easy, the standard error will be smaller for students of low ability levels because the test is better matched to their abilities; students with higher abilities will not be well measured by an easy test.

For more information about IRT, see Thissen and Steinberg (2009).

Reliability in Criterion-Referenced Measurement

In some measurement situations the primary interest is in determining whether a person meets a predetermined standard or

criterion. These judgments are sometimes made for individuals (e.g., Is my student Karol reading at grade level?) and at other times for groups (e.g., What proportion of eighth-graders is reading at or above grade level?) In everyday life, a driver's license test provides a familiar example of a criterion-referenced test. In education, the theory of criterion-referenced measurement was developed to provide a suitable technical scaffolding for making such judgments.

Again using the concept of consistency as a basis for conceptualization, reliability of criterion-referenced measures is concerned with issues such as whether an individual would receive the same classification if tested by different raters, on a different form of the test, or on different days. It is clear that the consistency will depend on how close the individual's true score is to the cut-point determining mastery. If a person is well above or below the cut-point, decisions will tend to be more consistent than for people near the cut-point. In the extreme, if a person's true score is exactly at the cut-point, then that person will pass half the time and fail the other half, no matter how well the test is constructed.

Because of the 'yes/no' nature of scoring, criterion-referenced test theory is somewhat different from the usual CTT but only because some different statistical methods are appropriate for these types of data. Criterion-referenced measurement is also more obviously concerned with decisions made on the basis of a test; other test theory models are only implicitly concerned with such decisions.

For more information about criterion-referenced measurement, see Hambleton et al. (1978), Harris et al. (1974), and Popham (1978).

Issues for Observational Measures

Many characteristics can be assessed by tests and questionnaires, but others are better assessed through observation of behavior. Behavior can be assessed by direct observation, or it can be recorded for later coding into categories. The recording can be more or less direct, ranging from video and audio devices to the memory of observers about past behavior (where the recording is in the mind of the observer).

Behavioral observations tend to take place in 'real-world' settings and generally (but not always) in uncontrolled conditions. Unlike tests, when those taking them know and expect that their behaviors are being observed, in natural settings one can imagine a range of possibilities for reactivity to the act of being observed. The more unobtrusive the observations, the more confidence one has in the quality of the measurement (though this generally would be classified as affecting validity rather than reliability). The classic reference on unobtrusive measurement is Webb et al. (1966).

Some issues tend to arise much more frequently in cases of observational measures than in tests and questionnaires. One such issue is the decision about whether to consider a characteristic being assessed as being categorical or continuous. For example, one might want to say that a person is either depressed or not, rather than being given a rating on a (nearly) continuous scale from not at all depressed all the way through extremely depressed. In such a case, the underlying trait is categorical, and different statistical methods are required. The most commonly applied method in this case is latent class analysis, which has been used in settings as diverse as determining stage of development of children, psychiatric diagnosis, and medical diagnosis.

One also tends to see different types of observed measures in observational data. Sometimes a variable has unordered categories, as would be the case if one classified a student in a class as either sitting quietly, sitting but making noise, or being out of his or her seat. Other times a variable may be a count or a rate, such as the number of times a student is out of his or her seat or the number of such times divided by the amount of time during which he or she was observed.

Typically, there are multiple observers of behavior or multiple coders of recorded behavior or both. (This can, of course, occur for certain types of tests, not just in behavioral observations. Essay tests are a common example, because multiple raters are often used. Similarly, portfolio assessments of students' work is often based on the judgment of multiple raters.)

These considerations present additional problems that affect the reliability of observational measures. Although consistency remains the primary concern, and the reliability (and validity) of behavioral observations depends on many of the same factors as for tests, additional factors of observational methods include the length of the period of observation, the conditions under which observations are made (including irrelevant stimuli), and the observers, raters, and coders used.

As one might infer from the discussion of GT, observational measures usually have more than one possible source of error, unlike that which is assumed in CTT. In addition to interobserver agreement problems, observers can make their observations at different times, in different settings, and under different conditions; all of these contribute to the error of measurement. GT allows researchers to estimate how much variability in the observations is due to each of these potential sources of error.

These differences from typical test data have additional implications for the statistical theory and data analysis methods generally used with observational data. When data come from observers, the issue of reliability is primarily a matter of whether two or more observers would come to the same conclusion (or, if a rating, the degree to which they agree) and whether the conclusion depends on when and for how long the observation occurs. For variables that fall into categories, one apparently simple method is to calculate the proportion of instances in which two raters or observers agree with each other about the category into which a particular observation should be placed. Unfortunately, the simplicity is only apparent, because for behaviors that are either rare or very common, a high percentage of agreement can occur by chance. For example, if a student is classified as being out of seat on 10% of occasions by each observer (and in seat the other 90%), the observers would agree 82% of the time, even if they were responding independent of the actual behavior. Therefore, such measures are typically corrected for chance agreement. Additional complications arise when there are three or more observers and when the observers do not rate or observe for the same period of time.

Details on statistical methods suitable for analyzing observational data are found in Suen and Ary (1989), and issues of

reliability and validity in ethnographic research are discussed in LeCompte and Goetz (1982).

Relationship of Reliability to Validity

In CTT, one cannot have a validity of a test that is higher than the square root of the reliability. This has led some to conclude that the reliability is more important than the validity of a measure because a test cannot be valid if it is not reliable. Although a test cannot be valid if its reliability is zero or extremely low, it is possible for one test to have a higher validity coefficient than another, even though the first is less reliable. Suppose that two tests (or other measures) are not actually measuring the same true score because they are measuring different traits. Then the one with lower reliability can have a higher validity if the trait it measures is more closely related to the outcome that the tests are meant to predict. For example, a paper-and-pencil test of rock-climbing skills might have greater reliability than a performance test done on a climbing wall in a gym, but the latter might still have greater validity as a predictor of performance in an actual rock-climbing situation.

The (invalid) conclusion about the relationship of reliability and validity has often caused strain between quantitative researchers and qualitative researchers. Qualitative researchers often use measurement procedures that have lower reliability than the procedures devised by quantitative researchers, but the qualitative researchers prefer these less reliable measures because they are measuring what the researchers want to measure. It is quite possible that in many cases they are right, and their less reliable measures have greater validity than certain measures that are more reliable.

Suggestions for Further Reading

Martin and Bateson (1986) is a short introductory work on measuring behavior. The two most popular undergraduate texts on measurement are those of Anastasi (1988) and Cronbach (1990). Nardi (2006) discusses reliability in a less technical manner. At a higher level are texts by Crocker and Algina (1986), Allen and Yen (1979) and Alwin (2007, 2010), and the general reference work edited by Linn (1989). McDonald (1999) is a newer text that provides a unified treatment of many issues in test theory. At a very advanced level, Lord and Novick (1968) and Bohrnstedt (2010) are texts on both CTT and IRT. A more modern text on IRT is Hambleton and Swaminathan (1984); simpler versions include Hambleton et al. (1991) and Sharkness and DeAngelo (2011).

See also: Classical (Psychometric) Test Theory; Data Collection: Interviewing; Dimensionality of Tests: Methodology; Generalizability Theory.

Bibliography

Allen, M., Yen, W., 1979. Introduction to Measurement Theory. Brooks/Cole, Monterey, CA.
Alwin, D.F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. Wiley-Interscience, Hoboken, NJ.
Alwin, D.F., 2010. How good is survey measurement? Assessing the reliability and validity of survey measures. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research, Second ed. Emerald, United Kingdom, pp. 405–434.
Anastasi, A., 1988. Psychological Testing, sixth ed. MacMillan, New York.
Bohrnstedt, G.W., 2010. Measurement models for survey research. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research, Second ed. Emerald, United Kingdom, pp. 347–404.
Brennan, R.L., 2011. Generalizability theory and classical test theory. Applied Measurement in Education 24, 1–21.
Crocker, L.M., Algina, J., 1986. Introduction to Classical and Modern Test Theory. Holt, Rinehart, and Winston, New York.
Cronbach, L.J., 1990. Essentials of Psychological Testing, fifth ed. Harper & Row, New York.
Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements. Wiley, New York.
Hambleton, R.K., Swaminathan, H., 1984. Item Response Theory: Principles and Applications. Kluwer-Nijhoff, Boston, MA.
Hambleton, R.K., Swaminathan, H., Algina, J., Coulson, D.B., 1978. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research 48, 1–47.
Hambleton, R.K., Swaminathan, H., Rogers, H.J., 1991. Fundamentals of Item Response Theory. Sage, Newbury Park, CA.
Harris, C.W., Alkin, M.C., Popham, W.J. (Eds.), 1974. Problems in Criterion-Referenced Measurement. Center for the Study of Evaluation. UCLA, Los Angeles.
LeCompte, M.D., Goetz, J.P., 1982. Problems of reliability and validity in ethnographic research. Review of Educational Research 52, 31–60.
Linn, R.L. (Ed.), 1989. Educational Measurement, third ed. Macmillan, New York.
Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.
Martin, P., Bateson, P., 1986. Measuring Behaviour: An Introductory Guide. Cambridge University Press, Cambridge, MA.
McDonald, R.P., 1999. Test Theory: A Unified Treatment. Lawrence Erlbaum, Mahwah, NJ.
Nardi, P.M., 2006. Doing Survey Research: A Guide to Quantitative Methods, Second ed. Pearson Education Inc., Boston, MA.
Popham, W.J., 1978. Criterion-referenced Measurement. Prentice-Hall, Englewood Cliffs, NJ.
Sharkness, J., DeAngelo, L., 2011. Measuring student involvement: a comparison of classical test theory and item response theory in the construction of scales from student surveys. Research in Higher Education 52, 480–507.
Shavelson, R.J., Webb, N.M., Rowley, G.L., 1989. Generalizability theory. American Psychologist 44, 922–932.
Suen, H.K., Ary, D., 1989. Analyzing Quantitative Behavioral Observation Data. Lawrence Erlbaum, Mahwah, NJ.
Thissen, D., Steinberg, L., 2009. Item response theory. In: Millsap, R., Maydeu-Olivares, A. (Eds.), The Sage Handbook of Quantitative Methods in Psychology. Sage Publications, London, pp. 148–177.
Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago, IL.

Religion and Abuse

Nancy Nason-Clark, University of New Brunswick, Fredericton, NB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Abuse occurs both in families of faith and within families holding no expressed religious convictions or practices. When abuse and religion are linked, the experience is profound for those who are victimized – whether it is a battered wife or a boy who is sexually abused by his parish priest. This entry explores some of the dynamics of abuse in families of deep religious faith, as well as the resources available to assist those who have been impacted.

According to the 'Religion and Violence e-Learning' web-based training, (www.theraveproject.org), abuse is behavior that shatters dreams and causes pain. It can involve physical violence like hitting, kicking, or choking. It can include name-calling or harsh words that are meant to hurt. Sometimes, it involves limiting or denying access to food, or money, or other family resources. Abuse can involve demands for sexual activity. However, abuse always involves responses that are meant to control, shame, or humiliate another person through intimacy and shared experience. Peace and safety for all family members cannot exist when abuse is present. Spiritual abuse – the attempt to impair or belittle a person's spiritual life or spiritual well-being – includes preventing him or her from performing spiritual acts or causing him or her to transgress spiritual obligations or prohibitions (Dehan and Levi, 2009; Sevcik et al., forthcoming). For highly religious women, spiritual abuse can accompany physical, emotional, and other forms of abusive behavior (Nason-Clark et al., 2013).

Victims write about abuse as a pattern of coercive controlling behavior that is directed at one intimate partner by the other (Owens, 2008). While women are its most common victims, at least in heterosexual relationships, any person can be a victim. Domestic violence is rooted in the desire of one partner to gain and maintain, power and control over the other. Many survivors refer to the betrayal and emotional abuse as the most damaging to them (Nason-Clark and Kroeger, 2004). Long after the broken bones have healed, the angry harsh words continue to erode the victim's sense of self-worth. Domestic violence that strikes at home is the training ground for future abuse. Children who grow up in families where they witness or experience abuse, are far more likely than those who have not suffered this experience, to be victims or perpetrators of similar abuse as adults.

The Prevalence and Severity of Abuse

Approximately one in four women around the globe has been physically abused by an intimate male partner, such as a husband or ex-husband. In fact, a woman is more likely to be injured, raped, or physically threatened by a current or former partner than by a stranger or another person. Studies differ in their funding base, their research design, the exact wording of the questions, and how the research instruments are administered; yet in every country of the world where data have been collected, the evidence is overwhelming: large numbers of women suffer physical violence from the men with whom they have exchanged marriage vows or shared intimacy and residence (Kroeger and Nason-Clark, 2010).

The US Centers for Disease Prevention and Control (CDC) report on their website that among victims of intimate partner violence, about one in four women and one in seven men have experienced severe physical violence by an intimate partner (e.g., hit with a fist or something hard, beaten, slammed against something) at some point in their lifetime. This has led Dr. Howard Spivak, the Director of Violence Prevention at the CDC to state: "Intimate partner violence, rape, stalking – all of these forms of violence can create toxic stress on the body that is long-lasting and cumulative, and can negatively impact a person's health and well-being for the rest of their life." (http://www.cdc.gov/violenceprevention/nisvs/index.html)

In Canada, (between 2000 and 2005), 653 000 women reported being victims of spousal violence, with 26% of these women being assaulted more than 10 times. Statistics Canada (2005) data reveal that younger women are at higher risk: women between the ages of 15 and 24 reported the highest one-year rates of spousal violence. Aboriginal women are also more likely to experience abuse in an intimate relationship; they are three times more likely to be victims of spousal violence than nonaboriginal women (Statistics Canada, 2008).

There is a tendency for many victims to minimize or deny the abuse, leading to the assumption that the abuse reported in country by country surveys is just the tip of the iceberg. Many staff in policing services and in the 'transition house' movement believe that the reporting of domestic violence in the United States (and in other Western countries) may have increased in recent years, largely because of enhanced training for first responders (like the police) and greater awareness amongst the public. There is ample evidence that in countries around the world, abuse impacts many women – and they often suffer in silence, experience deep shame, and harbor a sense of hopelessness about the future (Nason-Clark and Kroeger, 2004).

Understanding the Dynamics of Abuse

Fear is the most common response amongst victims of domestic violence. There is fear that the violence could escalate and fear of reprisal if they should tell someone. Fear leads to

secrecy. So much energy can be taken up by trying to pretend that everything is alright at home that there is little energy left over to take action. Many women are economically dependent on the men who abuse them. Others cling to the promise of changed behavior. Some women feel that they must stay in the marriage no matter what the cost. Religious women who are Christian sometimes believe it is their personal 'cross to bear' or that they are called upon to be a 'suffering servant' (Sevcik et al., forthcoming). Many women report that they remain in a marriage, despite the abuse that is suffered, on account of the children. Others lack adequate support systems if they decide to flee violent homes. For aboriginal women, immigrant women, and especially African American women in the United States, the fear of incarceration of their partner deters them from reporting the abuse or leaving a violent relationship. On the whole, women who are abused want the violence to stop but not for the relationships to come to an end. Faith traditions are sometimes guilty of encouraging women to return to relationships – even violent ones – after being offered advice that their changed behavior will lead to cessation of the abuse.

Violence is a learned behavior, and a substantial proportion of abusive religious men have experienced, or witnessed abuse in their childhood homes (Nason-Clark and Fisher-Townsend, forthcoming). As a result, the cycle of violence is passed from one generation to the next. Substantial numbers of the next cohort of abusive men are now young boys watching their mothers suffer. Children cannot flourish where abuse abounds. It affects how they perform at school, relate to friends, view family life, and feel about themselves. Abuse in the home creates an environment that threatens the physical and emotional health of all family members and leaves children especially vulnerable. Some abusers are determined to alter their violent ways and seek, with the support of the wider community, to change their behavior; and many abusive partners use religious language in an attempt to convince their wives, probation officers, pastors, or facilitators in batterer intervention programs (BIPs) that they are serious about altering their abusive behavior. Often, there are other issues, like substance or alcohol abuse, which need to be part of the perpetrator's overall plan for changed behavior.

Domestic Violence and Communities of Faith

When abuse strikes within religious families, many women seek the help of their religious leaders or other women of like faith. This is true within the Christian community, for women as diverse as Baptists, Pentecostals, Methodists or Roman Catholics; but it is also true for Jewish women and for Muslims. It is often a tightly guarded secret that a woman carries within herself, bearing the shame and the fear and other emotions created by living in a home environment where her safety and emotional health cannot be assured. Some abused women disclose their problems to their religious leaders in the privacy of the pastoral study, but many religious women choose other women of faith as their first point of contact in revealing the story of domestic violence.

Governments around the globe are recognizing the devastating consequences of abuse – especially violence against women and children – and public money is beginning to be directed toward a greater understanding of the problem of abuse, reforming judicial systems to respond more quickly, and providing health and other social services to those impacted. Yet, many religious organizations – churches, mosques, synagogues, and other houses of worship – have been slow to recognize and respond to the call for faith communities to be part of the solution to domestic violence, rather than be part of the problem. A so-called 'holy hush,' pervades many religious organizations (Nason-Clark, 2000). It can be observed when religious leaders do not name violence in the family context directly and choose rather, to call it family conflict, disagreements, or problems of communication. It occurs when religious leaders are unaware of the resources in their own local region to assist a woman in crisis – whether those resources are linked to the criminal justice system or advocacy agencies, like transition houses or shelters. Most religious leaders have never preached a message that explicitly condemns abuse in the family and many pastors, priests, and other faith leaders do not include any information about abuse in the information they offer to couples about to be married. Youth leaders within religious organizations too, often neglect to encourage young men and women to identify and then practice healthy interpersonal encounters with others of their own age. Finally, when women or men who have been victimized come to their faith communities for help, pastors, rabbis, and others in leadership roles are often reluctant to refer them to outside community resources, fearing that the experts outside the faith might challenge the faith perspective rather than focus on the cessation of abuse and the journey toward healing and wholeness.

Yet, when religious leaders employ both the language of their faith tradition and the contemporary cultural language of safety, empathy, and justice to condemn abuse and respond with compassion to victims, the result could be very effective. Religious leaders are able to speak words of comfort to victims and words of accountability to offenders that have the potential to offer help beyond that available at the level of community-based agencies. When priests, ministers, rabbis, and imams speak out against violence, they use their moral authority to bring healing to victims and call those who act abusively to accountability. This is a crucial aspect of ending the 'holy hush,' which denies the presence, potential, and long-term consequences of violence in families of faith.

Unique Vulnerabilities Facing Religious Battered Women

Religious women – Christian, Jew, and Muslim alike – tend to be very reluctant to see themselves as battered wives, even when they experience repeated and severe physical abuse at the hands of their intimate partners (Alkhateeb and Abugideiri, 2007; Dehan and Levi, 2009; Sevcik et al., forthcoming). Like most women, they feel afraid for themselves and their children and fear reprisal if they leave or seek assistance. But religious women also tend to be very forgiving, are more prone to blame themselves, and believe often that they have promised God (and their extended family network) that they will stay married until death. They also experience both the fear and

reality of rejection in their religious communities when they disclose the abuse and when attempts to repair the relationship fail.

The resources that abused religious women seek in the aftermath of domestic violence also differentiates them from other abused women. Since many communities of faith cling to traditional notions of gender and family life, and celebrate 'family values,' religious battered women have to take even more responsibility than other women for attempting to hold the family together no matter what the cost to them personally. Sometimes this is linked to notions of family honor, for others, it is intertwined with notions of the cost of discipleship. Victims are often reluctant to seek secular, community-based sources of support, preferring to look to others who share their faith perspective when the crisis comes home.

As a result, it is imperative that those in the helping professions (like therapists, domestic violence advocates, the police, and legal advisors, to name but a few) understand some of the specific needs of clients who are very religious. Often spiritual needs are primary in the aftermath of violence at home. And there are many questions that surface, questions that require answers from someone with either spiritual credentials or knowledge of the specific religious tradition with which the family is connected. Whether a religious victim of abuse first seeks help in her faith community, or from an agency in the broader society of which she is a part, she should be able to expect that her story of abuse is taken seriously; that she is given accurate, practical advice; that her safety and security are paramount; and that her faith is understood and respected. For highly religious women, the experience of domestic violence is so intertwined with the experience of personal faith that it would be impossible to understand one as separate from the other. A successful outcome will need to take into account both the experience of violation that a victim has endured and the victim's deeply held faith and practices. Cultural and religious competency is needed from those who work in supportive services, but this is rarely extended to abused women from any of the major faith traditions, who are very religious.

Responding to Abusive Religious Men

Imperative features of any intervention services offered to men who abuse their wives or intimate partners are the issues of accountability, justice, changed beliefs, and behavior. Most men who attend BIPs, come because the criminal justice system has mandated that they do so. It usually occurs as part of the sentencing after a conviction for domestic violence. A few men are referred to a program by a therapist, a religious leader, or their partners, as a final stage in the process of considering the end of the relationship.

Religious women who were abused are often very hopeful that intervention programs and services will change the violent men with whom they are in relationship. Since many abused religious women want the relationship to continue and avoid divorce, they often have faith that these interventions will change their abusive partners. While there are very few faith-based or faith-enriched batterer intervention services, the scant evidence that has been collected suggests that when religious leaders hold religious men accountable for their abusive acts and encourage them to follow through on court-mandated treatment options, those men are more likely, than men not so encouraged, to complete the program in its entirety (Nason-Clark et al., 2004). This is no small matter as program completion in all BIPs is an enormous challenge.

In any intervention services for religious men who act abusively, it is important that at least one of the facilitators or staff at the agency involved, is able to respond using the language of the participant's faith traditions. This enables facilitators to counter any claims made by program participants that their sacred books or traditions justified the abuse or violent acts toward their partners. This ability could help programs hold religious men accountable for their actions. Religious faith can be an important variable in helping violent men become accountable and do the work required for change; and faith leaders are an important, yet unrecognized-resource in community-based responses to domestic violence (Nason-Clark and Fisher-Townsend, forthcoming).

While scholars have argued for some time that it is critical to see the centrality of the religious belief system in the lives of many men who are violent, it has not become a key component of service delivery. Yet, faith is a central ingredient in understanding the assumption of male entitlement, power, and control that fosters abuse, and it must be part of the solution to domestic violence.

The Challenge for Religious Leaders in Responding to Abuse

Studies that investigate the response of religious leaders to abuse have reported that Christian pastors and priests claim that their training has not adequately prepared them to understand the prevalence and severity of abuse, its long-term consequences, or the spiritual dimensions that often surface in the life of a victim or an abuser. The same can be said for religious leaders in other faith traditions. Even seminarians currently enrolled in higher education programs note that they do not feel equipped to respond to families impacted by abuse (McMullin and Nason-Clark, 2011, 2013). Many victims, as well as perpetrators, look to their faith communities in the aftermath of crises at home: religious leaders, then, become the gatekeepers to sources of support and referral options. If the clergy denies or minimizes the abuse, the woman or man who has come for help often will not seek assistance elsewhere. Members of the clergy who are aware of the issue of domestic violence and are connected with a broad range of community-based resources, can identify and use their faith traditions to help both survivors and those who act abusively.

Religious leaders who lack knowledge about abuse are less likely to make referrals and more likely to see the problems in spiritual terms alone. This has the potential to increase the safety risk for a victim and often fails to bring the abuser to justice and accountability. Many religious leaders – Christian, Jewish, and Muslim – are strong supporters of family connectedness and they are reluctant to see any couples divorce or separate. Yet, often, living apart at least temporarily, is exactly what needs to happen to ensure the safety of all. Because the clergy tends to have little training

on the issue of abuse; often mistrusts broader community-based resources; and lacks specific faith-based resources; it often misunderstands the relationship dynamics that lead to abuse. Moreover, for some religious leaders, translating the rhetoric of happy family living into practical assistance for families in crisis seems overwhelming. It can be difficult for untrained members of the clergy to know how to harness their religious traditions to promote safety, justice, and accountability.

The Challenge for Community-Based Agencies in Responding to Religious Clients and the Power of Collaboration

Many workers in community-based agencies find it difficult to intervene in cases where clients are very religious. Many staff members do not have the specific knowledge that would allow them to work in a culturally competent manner with people from diverse religious backgrounds. Thus many victims of abuse may not feel safe in disclosing to a community-based agency that they are highly religious. In part, some of the challenges for secular workers intervening in cases involving highly religious women, may be the result of workers blaming the religious tradition for the violence or being unable to challenge religious beliefs used to justify abuse in particular faith traditions. Building bridges between community agencies and religious congregations is a very important initiative for responding to the prevalence and severity of abuse in the context of highly religious families. A coordinated approach between criminal justice workers (like the police, the courts, probation services), therapeutic staff (like social workers), domestic violence advocates, and religious leaders helps to ensure that the response to domestic violence takes into account cultural and religious characteristics leading to safe outcomes.

There are many web-based and print resources, as well as video documentaries, that have been created to assist religious women who are victims of domestic violence. Examples include The Religion and Violence e-Learning Project (www.theraveproject.org), the Faith Trust Institute (www.faithtrust.org), and Jewish Women International (www.jwi.org). Excellent documentaries include those produced by Diva Communications (www.divacommunications.com/programs/i-believe-you/) and Day of Discovery (www.rbc.org). Many of these organizations, including the Muslim-initiated Peaceful Families Project, produce brochures for washrooms located in houses of worship, shoe cards identifying how to contact a help line or shelter, bulletin inserts for Family Violence Awareness month, materials for religious leaders to use in premarital counseling, as well as curriculum for youth meetings.

When Religious Leaders Abuse

For the last 25 years, the public has been inundated with press stories of religious leaders who have abused the trust and respect of those they were supposed to serve. These abusive religious professionals claimed a divine calling to a ministry of compassion, usually at the parish or congregational level. The story line is familiar: a beloved priest or pastor; a victim initially grateful for the extra time or attention the religious leader offered; an unsuspecting congregation; secrecy and denial; an unrepentant religious hierarchy intent on concealing any wrongdoing on the part of its ordained staff; confusion and mixed messages between judicial and religious inculpatory processing of the allegations (with the later possibility of charges); very little emotional (not to mention financial) support for victims, their families and the congregations directly impacted by the tragedies; mounting legal costs; and long-term consequences for those who have been violated and betrayed.

The impact on victims and their families cannot be overstated. At the deepest level, for the adolescent victims, it was a betrayal of trust that had physical, emotional, sexual, and spiritual implications. Parents, relatives, and the broader Roman Catholic community shared secondary victimization of the clergical malfeasance (Shupe, 1995). For many years after the judicial sentencing and the Commissions and government-sponsored inquiries that followed, the parishes where priests had been charged faced the aftermath of the scandal. Often congregational members had been vocal supporters of individual priests and hostile critics of those who disclosed the abuse. Once the allegations turned to charges and then to convictions, there was extreme embarrassment and outrage directed toward the church and its leaders. Particularly painful was the fact that often Bishops and others in the clerical hierarchy knew of the proclivity of a particular religious leader to be errant. When given a choice, those clerical dignitaries ignored the pain and suffering of victims, their families, and the congregations where the errant priests were placed in ministry; and chose instead to protect the reputation of the malfeasant priest and the broader institution of which he was a representative. Often the priest was moved from one parish to another without even a hint of his sexually abusive behaviors toward minors being noted.

Disclosures of sexual abuse at the hands of Roman Catholic priests have been mounting since the 1980s. This is true across Canada, the United States, Ireland, Germany, and in many other parts of the developed world. There is mounting evidence that there were strategic cover-ups by the Roman Catholic religious hierarchy, and a lack of compassion and justice for victims and their families. But, sexual abuse and sexual misconduct, has not been linked only to the Roman Catholic religious elite.

Diana Garland has conducted a major research initiative to explore sexual misconduct in the clergy more fully (www.baylor.edu/clergysexualmisconduct/). She argues that most of the studies regarding this issue have focused on the motives of the offender, the nature of the abusive relationship, and the long-term consequences for both primary and secondary victims. Prevention strategies are rarely considered. Scarce too is information on the steps to follow if you are a victim of clergical sexual misconduct. Garland argues that it is erroneous to consider a sexual relationship 'an affaire' when it occurs between a member of the clergy and a congregational member, and that such relationships are rightfully viewed as clerical sexual misconduct. Providing a proper code of ethics in a congregational context is imperative. Both clergy and

parishioners need to understand that clerical sexual misconduct is inappropriate.

In a national random survey conducted in 2008 with 3558 respondents, Garland and colleagues found that approximately three percent of women who had attended a congregation in the preceding month reported that they had been the object of some form of clerical sexual misconduct in their adult lives. The overwhelming majority of the sexual advances (92%) were made in secret rather than in open dating relationships, and most of the offenders (67%) were married to someone else at the time of the advance. Representing a variety of different Christian and Jewish affiliations, 8% of those surveyed reported having known about clerical sexual misconduct occurring in a congregation they had attended. One of the very important points Garland and colleagues (Chaves and Garland, 2009; Garland, 2009) raised in their work, is the high level of self-disclosure that can occur in congregational life. Garland refers to this as 'trust in the sanctuary' – meaning that congregations are considered sanctuaries, literally safe places, where a person's normal attentiveness to self-preservation need not be a priority. With limited oversight of pastors' work and behavior, clerical sexual misconduct can develop and flourish. Vigilance is necessary both by religious leaders and those who faithfully support them through participation in congregational life.

See also: Catholicism and Gender; Judaism and Gender; Organizations and Culture; Religion and Health: Sociological Considerations; Religion, Family, and Kinship; Women's Religiosity.

Bibliography

Alkhateeb, M., Abugideiri, S., 2007. Change from within: Diverse Perspectives on Domestic Violence in Muslim Communities. Peaceful Families Project, Great Falls, VA.

Chaves, M., Garland, D., 2009. The prevalence of clergy sexual advances toward adults in their congregations. Journal for the Scientific Study of Religion 48, 817–824.

Dehan, N., Levi, Z., 2009. Spiritual abuse: an additional dimension of abuse experienced by abused Haredi (ultraorthodox) Jewish wives. Violence Against Women 15 (11), 1294–1310.

Garland, D., 2009. The Prevalence of Clergy Sexual Misconduct with Adults: A Research Study Executive Summary. http://www.baylor.edu/clergysexualmisconduct/index.php?id=67406 (accessed 17.06.03.).

Kroeger, C., Nason-Clark, N., 2010. No Place for Abuse: Biblical and Practical Resources to Counteract Domestic Violence, second ed. InterVarsity Press, Downers Grove, IL.

McMullin, S., Nason-Clark, N., 2011. Seminary students and domestic violence – applying sociological research. In: Nason-Clark, N., Kroeger, C., Fisher-Townsend, B. (Eds.), Responding to Abuse in Christian Homes: A Challenge to Churches and Their Leaders. Wipf and Stock, Eugene, OR, pp. 231–246.

McMullin, S., Nason-Clark, N., 2013. Preparing seminaries for collaborative work. In: Nason-Clark, N., Fisher-Townsend, B., Fahlberg, V. (Eds.), Strengthening Families and Ending Abuse: Churches and Their Leaders Look to the Future. Wipf and Stock, Eugene, OR, pp. 220–241.

Nason-Clark, N., 1998. The impact of abuses of clergy trust on female congregants' faith and practice. In: Shupe, A. (Ed.), Wolves among the Fold: Religious Leadership and Abuses of Power. Rutgers University Press, New York, pp. 85–100.

Nason-Clark, N., 2000. Making the sacred safe: woman abuse and communities of faith. Sociology of Religion 61, 349–368.

Nason-Clark, N., Fisher-Townsend, B., forthcoming. Men Who Batter. Under contract with Oxford University Press.

Nason-Clark, N., Kroeger, C., 2004. Refuge from Abuse: Hope and Healing for Abused Christian Women. InterVarsity Press, Downers Grove, IL.

Nason-Clark, N., Murphy, N., Fisher-Townsend, B., Ruff, L., 2004. An overview of the characteristics of the clients at a faith-based batterers' intervention program. Journal of Religion and Abuse 5, 51–72.

Nason-Clark, N., Fisher-Townsend, B., Fahlberg, V. (Eds.), 2013. Strengthening Families and Ending Abuse: Churches and Their Leaders Look to the Future. Wipf and Stock, Eugene, OR.

Owens, J., 2008. A survivor looks back: what I wish pastors had known when I was looking for help. In: Kroeger, C., Nason-Clark, N., Fisher-Townsend, B. (Eds.), Beyond Abuse in the Christian Home: Raising Voices for Change. Wipf and Stock, Eugene, OR, pp. 1–23.

Sevcik, I., Rothery, M., Nason-Clark, N., Pynn, R., forthcoming. Secular and Sacral Communities: Working Together to Address Intimate Partner Violence. University of Alberta Press, Edmonton, AB.

Shupe, A., 1995. In the Name of All That's Holy: A Theory of Clergy Malfeasance. Praeger, Westport, CT.

Statistics Canada, 2005. Family Violence in Canada: A Statistical Profile. Statistics Canada, Ottawa.

Statistics Canada, 2008. Family Violence in Canada: A Statistical Profile. Canadian Center for Justice Statistics, Ottawa.

Shupe, A., Stacey, W., Darnell, S. (Eds.), 2000. Bad Pastors: Clergy Misconduct in Modern America. New York University Press, New York.

Relevant Websites

http://www.theraveproject.com/index.php – The Religion and Violence e-Learning Project is a web-based series of resources and online training for religious leaders and other secular workers that assist religious victims (or religious perpetrator) of domestic violence (accessed 17.06.03.).

http://www.baylor.edu/clergysexualmisconduct/ – Information on the Clergy Sexual Misconduct Project is available, as are resources, such as a Code of Ethics (accessed 17.06.03.).

http://www.faithtrustinstitute.org – Faith Trust Institute offers a wide range of resources and training on issues of abuse, and clergy misconduct (accessed 17.06.03.).

http://www.jsafe.org – Jewish Institute: Supporting an Abuse-Free Environment (accessed 17.06.03.).

When Love Hurts©, 2007When Love Hurts, © 2007 by RBC Ministries. This documentary considers various aspects of abuse in families of faith and the response of the church to victims and to perpetrators. For copies, contact RBC Ministries, 3000 Kraft SE, Grand Rapids, Michigan 49512.

http://www.divacommunications.com/programs/i-believe-you/ – Diva Communications – Documentary Program, I Believe You, an interfaith look at the issue of domestic violence in religious families and the role of religious leaders in responding to this crisis (accessed 17.06.03.).

http://www.jwi.org/page.aspx?pid=1276 – Jewish Women International – Domestic Violence Prevention and Training. A wide variety of resources and suggestions are available on this site for victims, their families and friends, and rabbis (accessed 17.06.03.).

http://www.cdc.gov/violenceprevention/nisvs/index.html – US Centers for Disease Control and Prevention – offers up-to-date data on intimate partner violence and other forms of abuse in the family context (accessed 18.06.03.).

Religion and Economic Life

Richard H Roberts, University of Stirling, Stirling, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.H. Roberts, volume 19, pp. 13028–13034, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Religious factors have played a considerable role in economic systems throughout history. Here, the relationship between religion and economic systems are examined beginning with the archaic mode of production and continuing through feudalism, early capitalism, late capitalism, and in the present global economy. Marxist and Weberian theories of religion and economic life are presented, along with more modern theories of the complex connection between religion and economics.

Investigation of the relation between religion and economic life has long been a central concern of sociological reflection, but the topic has been largely defined through debates on the relationship of religion and capitalism, the latter being the specific and dominant expression of economic life in modernity. Marxist and Weberian perspectives on religion and capitalism inevitably overshadow the discussion of both the history and the contemporary forms of the ambiguous and changing relation between religion and economic life. First, this article will review precapitalist configurations of religion and economic life. Second, the emergence of modern capitalism is considered. Third, reflective religious responses to nineteenth- and twentieth-century capitalism are outlined. Fourth, the impact of religion on economic life and economic life on religion is examined under contemporary capitalism, which brought with it globalization, commodification, and managerialism, and a wide range of contrasting religious responses. In the concluding section, future avenues for research are considered.

Precapitalist Configurations of Religion and Economic Life

The relationship between religion to economic life depends on the theoretical assumptions of the inquirer. The most relevant founding fathers of sociology, Marx, Weber, and Simmel, sought to develop grand theory. In general terms, sociologists of religion have devised and reinforced relatively homogenous accounts of the modernization process and attendant phenomena of secularization, and they have used particular economic theories in order to inform and justify their approaches. By contrast, under the influence of Marcel Mauss' account of the gift and using exchange theory (Gregory, 1994), anthropologists have generally paid more attention to specific human ecologies and reciprocal community economies as originating in "Stone Age Economics" (Sahlins, 1974). Some anthropologists have correspondingly defended distinctive and reciprocal human ecologies and cultures against the ravages of economic modernization in ways contrary to orthodox sociologists.

Archeological and textual records indicate that ancient Egypt, Assyria, Babylon, and Israel all had contextually relevant, practical financial institutions, but there has been limited attention to how religious institutions factored into the economic operations of archaic empires. The history of Western economic thought is normally traced back to the agrarian world of the fourth- and fifth-century BC, which had a well-defined merchant class. The representative unit of the economy was the household and commodities were exchanged in free markets. In the Graeco-Roman world early 'economic' reflection was based on the *nomos* or law and rule of the household *oikos*, and the practical wisdom of household management (*oeconomicos*) was the source from which early economic conceptions were derived. In the setting of the city-state (*polis*), the universal existence of slavery ensured that labor was relatively fixed and immobile. Religion and economic life as part of the cultural superstructure were conditioned by the dependence of ancient Greek and Roman life upon a slave economy; correspondingly, manual labor was regarded as menial. This classical tradition provided some of the core values of Western culture for two millennia in the West.

Central to Greek thought were the good life and reflection on the nature of virtue: seen thus, economic activity amounted to a mere external factor. Plato and Aristotle were, however, aware of such concepts as the distinction between use value and exchange value, the role of money, and the division of labor. In *The Republic*, the latter corresponded with the human types of ruler, soldier, and worker, and each class pursued its proper, specialized goals. The ideal rulers, the philosopher kings, would ostensibly be free from vulgarity of wealth. In the classical tradition, arts of commerce and the practice of usury were condemned, above all the latter, which makes gain out of money itself and not from the natural object. Money was to be understood as a medium of exchange, a unit of account, and a store of value, but in itself money was barren, and usury the oppressive exploitation of misfortune. The classical Christian doctrine of the Trinity, couched in terms of the 'economy' of God's self-revelation of his 'essential' being, employed a metaphor with an enduring importance. Later, the evolving theology of the atonement repeatedly exploited certain economic metaphors, not least in the conceptuality of payment of debt and redemption from sin.

In prefeudal societies throughout the world, religious institutions were undifferentiated from the state and from economic institutions, yet systematic studies of the influence of

religious institutions on economic life in pre-Christian and non-Christian societies are lacking. Economies based upon slavery were questioned in principle by early Christianity (and probably in other traditions as well), but the slave–master system of the archaic economy was generally supported in all major religious traditions, providing political legitimacy for economic relations in a system where religion, the state, and the economy were intimately intertwined. According to Schumpeter there was little substantial interest in Western Christianity in what was to become the dismal science of economics, for "the Christian Church did not aim at social reform in any sense other than that of moral reform of individual behavior," nor, moreover, "did the Church attempt a frontal attack on the existing social system or on any of its more important institutions." Indeed, "It never promised economic paradise or, for that matter, any paradise this side of the grave" (Schumpeter, 1954: p. 72). Michael Mann argued that "Christianity was a form of ideological power …. Only by becoming a Christian could one live a truly meaningful life" and that its "…power resides originally in the fit between the Christian message and the motivations and needs of the converted" (Mann, 1986: p. 302). The question as to the 'fit' or affinity, between individual Christian belief and spirituality upon individual motivation and economic life has remained a matter of central importance in Weberian scholarship. In contrast, Marxist analysis implied that the religious ideology that emerged in the theological debates of the patristic theologians of the first four centuries in the West was but a reflex of the material base; this, however, does less than adequate justice to the complexity of the factors involved.

After the fading of the eschatological expectation of the earliest Christian communities and abandonment of the primitive communism, monasticism preserved the radical exemplar of the economic ethic of the Synoptic Gospels (Troeltsch, 1931). The religious orders that emerged from the practice of the Desert Fathers provided both the basis for the continuity of a religious culture and an economic model compatible in key respects with the warrior culture of feudalism. Monastic communities and the religious orders came to be centers of rational economic organization in ways that prefigure the business corporation. Monastic organization created wealth, which, in turn, led to inner contradictions tackled by attempts at institutional reform – and which later became the target of acquisitive envy during the Protestant Reformation.

At the high point of scholastic theology and in the context of feudalism, a crucial debate took place, which connects the early economic ideas of Aristotle (and of his wide philosophical concerns) with the early modern loosening of the constraints upon economic activity that facilitated the growth of international trade, and the justification of such expansion in mercantilism. In the Middle Ages, Europe was characterized by a relatively stable agrarian, largely subsistence economy with slow urban growth. Central to urban productive activity was the guild, a form of association, which ensured corporate monopolies, the protection of members, standards of quality, and price stability. A similar form of urban economic organization prevailed in the Islamic world (Kuran, 2011). The presence of the Church in relation to all social institutions ensured an unparalleled degree of cultural interfusion between religion and economic life. Social status and social function implied definite religious duties in a divinely ordered governance that extended throughout the body politic. Nineteenth century guild socialism attempted to recover something of this integration of meaning into 'menial' work.

St Thomas Aquinas, the leading theologian of scholasticism, defended the just price and the prohibition of usury, the core principles of the attempted control by the Church of economic life. Estimating the just price based on moderate gain was a difficult task: the merchant was justified in trading at a price, which ensured the life of his household in its accustomed style. Following Aristotle's account of money as barren, St Thomas condemned usury, the lending of money at interest, arguing that given the true role of money solely as a medium of exchange, it did not exist in the way that a rented house continued to exist when occupied by a tenant. St Thomas' counterrealist account of money applied more to the ethics of the personal loan rather than to large-scale financial transactions. During the course of the fourteenth and fifteenth centuries the increasing complexity of major financial transactions made it increasingly difficult to fit new practices within the constraints of Church law and this was reflected in modifications of St Thomas' rigorist views.

The prohibition of usury was circumvented through the creation of an international system identified with the Jews that was capable of supporting the accumulation of the major loans required for effective warfare. A similar role was also undertaken for a limited period by the Templars. The pariah status of Jews and their indispensable role in facilitating financial transactions and more widely in commerce came to reinforce negative attitudes, which provided a long-enduring basis for anti-Semitism (Stern, 1977: Chap. 18). The role of such Jewish thinkers as Marx and Sombart in the theorization of capitalism in the nineteenth century and beyond reinforced the impression that Jews played an essential role in the development of modern capitalism (Paul Morris in Roberts, 1995: Chap. 5). In contrast to the shifting perspectives on finance and usury in the Christianity, Islamic empires remained committed to prohibitions, and allowed Christians and Jews to have advantageous access to developing mercantile (Choudhury and Malik, 1992; Kuran, 2011).

In the early modern period, the growth of international trade, the breakdown of unity within Christendom following the Reformation, and the destruction of stable communities unified by shared religious values were factors associated with the rise of mercantilism. During the course of the sixteenth, seventeenth, and eighteenth centuries, economic analysis gradually gained increasing autonomy, throwing off the remaining constraints of scholasticism in favor of market-based mercantilism and physiocratic economic theories.

The Emergence of Modern Capitalism and Its Critique

The advent of early industrialization and the development of the factory system brought about unprecedented social disruption, breaking many ancestral ties: this process required both explanation and legitimation. In *The Wealth of Nations* (1776), Adam Smith articulated a world view in which virtue is equated with profit. Smith's analysis of the rational division of

labor and its necessary place in the making of profit provided the justification for the removal of remaining constraints. The quasi-providential 'invisible hand' purportedly provided a spontaneous connection between individual wealth creation and the common good of society. Political economy attained a high status among an ascendant entrepreneurial class as it emancipated economic practice and provided the means for the elimination of embedded social, ethical, and cultural inhibitions, including those traditionally endorsed by religion in preindustrial agrarian societies.

Marx regarded the intellectual destruction of religion by the eighteenth century Enlightenment as the axiom underpinning the true emancipation of humankind through subsequent social, political, and economic critique, and for Marx there was no basis for any mutual reflective encounter between religion and economic life. The representation of the proletariat in Marxist theory crystallized the fears of churches already confronted with their loss of influence over a growing urban working class and its anticlericalism.

The uprooting of agrarian workers from an ancestral status sanctioned by divine order and their regimentation into a disciplined, wage-earning working class differentiated by a new division of labor created a dilemma for the churches. On the one hand, the Anglican clergyman Thomas Malthus endorsed population regulation in accordance with economic needs. On the other, the living conditions of the working class were an affront to those sensitive to the historic Christian teaching on wealth and poverty. Such figures as the Scot, Thomas Chalmers, English Christian socialists, and continental religious socialists struggled to find ways to redefine the role of religion under industrialized conditions. An unchurched, brutalized proletariat represented a threat to the role of Christian clerical and intellectual elites across Europe. In the English religious establishment, serious critical engagement with 'scientific socialism' or political economy was minimal. Methodism was exceptional in that it successfully harnessed religion and industrial work in ways, which may have spared England a revolution.

Max Weber was fully aware of the threat, which Marx's thought and Marxist socialism created for Liberal Protestant European bourgeois culture and society. According to Weber, the *mentalité* associated with the capitalist mode of production had historical connections and elective affinities with 'this-worldly' Protestant asceticism. It is therefore customary (but not undisputed) to link the nascent 'spirit of capitalism' with the much disputed 'Protestant ethic' particularly in the Reformed Christian tradition. Weber explored the affinities between the single-minded pursuit of God and an equally focused desire for profit, but he also applied his analysis to the major world religions, classically, for example, to the religion of India (D. Gellner in Roberts, 1995: Chap. 1). Marx and Weber agreed on the centrality of acquisitiveness to the bourgeois mind: political economy both explained and justified single-minded acquisitiveness, and divorced economic relations from human needs. In fact the greatest good of this ethic, the earning of more and more money combined with strict avoidance of all spontaneous enjoyment of life, ignores the production of happiness as a measure of the merits of an economic system. Economic acquisition is no longer subordinated to man as the means for the satisfaction of his material needs (Weber, 1930/1976: p. 53).

In *Religion and the Rise of Capitalism*, R.H. Tawney applied Weber's insights to seventeenth-century England, and provided an indirect critique of the *laissez-faire* liberal capitalism, which underwent economic and social crisis during the period between World War I and World War II. Later, further British explorations of the processual metaphor of the 'rise' of capitalism were extended by V.A. Demant to its 'decline' and its 'persistence' by R.H. Preston. Christian thinkers such as the Anglican William Temple relied upon the emerging Keynesian standpoint in advocating 'middle axioms' as the basis for socioethical intervention. Following the World War II, the Keynesian program of partly state-managed mixed economy (and declining) capitalism underpinned the ideology and implementation of the 'welfare state' for 30 years until the 'triumph of capitalism' associated with the Reagan–Thatcher era. The reassertion of neoclassical economics (Friedman, 1962) and the ideology of the New Christian Right (Green, 1987) stripped away both tacit belief in the 'hidden hand' and all ascetic hesitance with regard to hedonism (Roberts, 1992). Mainline religions in Anglo-American Western countries were split between localized variants of the 'option for the poor' and outright endorsement of the 'enterprise culture.' Those who advocated the option for the poor were depicted as naive and ill-informed practitioners of the 'kindness that kills.' Conversely, New Christian Right ideologues proclaimed the virtues of greed, the adequacy of the 'trickle down' theory of welfare, and the necessity of a docile, 'flexible' labor force willing to be uprooted and exploited.

Reflective Religious Critiques of the Relation of Religion and Economic Life

The most significant mainline religious response to industrialization and to the growth of capitalism and Marxist socialism is to be found in Roman Catholic social teaching. After lengthy consultation provoked by the changing socioeconomic conditions of the nineteenth century, the papal encyclical *Rerum Novarum* was published in 1891. Aware of the fate of the Catholic Church in revolutionary France (and then later of the Orthodox Church in Soviet Russia), papal teaching represented the historic struggle between capital and labor as conflict between the competing rights and obligations of owners and of workers, in relation to which the Church was to exercise a moderating role. Both parties were to recognize the mutuality of their rights and obligations, and to relativize these through acknowledgment of the beatific vision as the sole and legitimate goal of humankind. For a century, the Church developed this doctrine through a series of encyclicals culminating in Pope John Paul II's *Laborem exercens* (*On Human Work*) (1981). Following the collapse of Marxist socialist states in 1989/90 the changing nature of capitalism was recognized in the encyclical *Centesimus Annus* (1991). The latter contains a cultural rather than a politico-legal analysis and critique of both socialism and capitalism in which the integrity of the human agent is asserted against the contradictions of a destructive, secularizing modernity in need of comprehensive reevangelization. This tradition of Catholic opposition to inequality and support for the lower classes is strongly ascendant in the pronouncements of Pope Francis.

Direct intellectual contact between Christians and Marxist socialists did not take place on a significant scale until after the Second Vatican Council (1962–65) and took the form of Christian–Marxist dialogue during the following decade, but even this discussion of the nature of 'true humanism' would seem to have had minimal long-term influence. At the same time 'political theology' embodying a synthesis between Protestant theology and various strands of revisionist Marxism also developed in Germany. Insights from political theology were transplanted to Central and Latin America, where, after the Second Vatican Council and the Medellin Bishops' Conference of 1968, 'liberation theology' emerged in the context of rapid industrialization and urbanization. This theology drew critical perspectives from Marxist analysis of society but side-stepped the atheistic aspects of Marxism, as was to be emphasized in the critical response of the Sacred Congregation in Rome. In Central and Latin America liberation theology's stress upon the 'option for the poor' has tended to exacerbate a divergence between Roman Catholicism and Pentecostal Protestantism, where the latter has performed a role analogous to that of Methodism during the European industrial revolution (Martin, 1990: pp. 27ff.). The most radical theological critique of capitalism through the representation of the imbalanced distribution of wealth in the global economy was a confessional issue (Duchrow, 1987).

The Roman Catholic apologist and critic of liberation theology Michael Novak has written of a 'magic capitalism.' Humankind has to abandon the naive and infantile 'dream' (the 'false consciousness') of socialism, and to recognize the universal necessity of alienation at the heart of a vital economic reality. Every individual must pass through a rite of passage in which each encounters the 'empty shrine' of 'democratic capitalism': this is the contemporary "dark night of the soul" (Novak, 1982/1991: p. 53). Others, like Peter Berger, deny that capitalism has such mythicizing power (Berger, 1987). Additional accounts of religion and economic life are to be found in Protestant Christian theologies of 'God' the Economist (Meeks, 1989).

Recently, the globally influential 'Third Way' politics of the center left advocated by some has been accompanied by a revival in Christian socialism and continuing campaigns to remoralize the individual. This post-Keynesian and post-communist active capitalism seeks to combine the ideology of entrepreneurial zeal with managerialism and relentless socio-cultural 'modernization.' The 'Third Way' contributes to the increasing salience of the religious factor in economic life, albeit with paradoxical consequences.

Contemporary Capitalism: Globalization, Commodification, and Managerialism

Contemporary capitalism has diversified into a plurality of capitalisms, each with its contextually relevant elective cultural affinities between religion and economic life. These capitalisms have become consolidated through globalization into an infinitely complex system. The representation of globalization as a new paradigm for the sociological interpretation of religion (Robertson, 1992) has encouraged investigation of the struggle both between and within religions for power and influence in the context of the asymmetrically distributed economic opportunities and outcomes in an ever more competitive core–periphery world system (Beyer, 1995). Moreover, the problematic interface between the global and the local has become the focal point of social identity formation (Castells, 1996: vol. II, Chap. 1). Commodification, the ascription of exchange and monetary value to all aspects of the world and human life, has advanced inexorably with the result that the economic function is central, and natural or intrinsic values become ever more marginal and are reduced to mere market values. Global managerialism (Entemann, 1994), mediatized marketing techniques, and the application of 'neo-Fordist' organization to manage work (Drucker, 1993) have accelerated the homogenization of human beings through the replacement of critical education with rational training. These are now processes replicated within some influential strands of both mainline religion and alternative spiritualities.

Information technology has increased the velocity of circulation of all forms of 'capital' to the point that space–time compression is having pervasive cultural consequences in new 'economies of signs and space' (Lash and Urry, 1994). For those at the heart of a core–periphery economic system, capitalism has become 'soft' and infused with transformative cultural power (Thrift, 1997). Under these conditions, religious language and theological vocabulary have become resources for the expression of economic and cultural power. Thus, terms like 'omnipresence,' 'omnipotence,' 'omniscience,' and 'instantaneity' formerly attributed to the 'God' who subsisted at the core of Western consciousness are perhaps the only adequate conceptual categories for such transmutations. Spiritual practices are now recruited into management training in order to assist through the provision of peak experiences the strongest possible integration of individual motivation with corporate goals (Roberts in Roberts, 1995: Chap. 9). The religious field may be remapped in relation to these ongoing transformations of economic life. The application of rational choice theory to religious behavior, a syncretic global marketplace of New Age 'self-religions' and spiritual body practices, the managerialization and marketization of mainline Christianity, Christian fundamentalist 'prosperity theology,' neo-pagan 'nature religion,' and ecological piety are all indications that religion and economic life are in active and sometimes mutually contradictory synergies. Consequently, there is no one 'spirit of capitalism' in advanced modernity – but a multiplicity of affinities and collusions in context marked by growing differentiation between 'religion' and 'spirituality.'

Conclusion: Future Prospect – Research Options

In the modern world, the modern and the postmodern conditions sometimes coexist in close, yet problematic contiguity (Harvey, 1989). From being marginal to the order of things, conscious economic rationality is now for many 'life' itself. 'Economic life' tends to absorb, transmute, and meld 'reality' into a globalized *maya*, outside of which there would appear to be no fulcrum for a critical ideology or politics of resistance (Baudrillard, 1981). After the so-called 'End of History,' leading sociologists of religion such as Steven Bruce and Bryan Wilson continue to represent religion as subject to terminal attrition by an inevitable process of secularization.

Regarded thus, religion can no longer be a resource for the culturally significant critique of total human submission to the dictates of 'economic life.' Yet, this view of the role of religion in economic life ignores the resurgence of religious control over economies in most of the middle east, and the overarching political power of conservative Christianity in the United States and Russia, the rising dominance of conservative Hindu nationalism in India, pan-Islamic movements throughout Asia, and Buddhist control over the political economy of Sri Lanka and much of Southeast Asia.

Future research could productively focus on the intersection between modern religious movements throughout the world and their impact on economic relations. The political economy of the contemporary world is dominated by the impact of mass political movements, which regulate economies, determine the fortunes of individuals and groups, and fuel economically devastating wars. Religious movements will also play a substantial role in determining responses to global climate change, and it is unclear how different religious movements will respond. For example, conservative Christian movements in the United States have played a key role in deregulating the economy and denying that human activities have consequences for the environment and climate. And, it appears that Hindu nationalism is playing a similar role in India. Religious movements are also responsible for undermining the economic fortunes of women, denying them access to education, health care, transportation, and basic freedoms – with considerable negative economic consequences for nations and for the global economy.

See also: Catholicism; Economic Anthropology; Economic Sociology; Religion: Mobilization and Power.

Bibliography

Baudrillard, J., 1981. For a Critique of the Political Economy of the Sign. Telos Press, St. Louis, MO.

Berger, P.L., 1987. The Capitalist Revolution. Fifty Propositions about Prosperity, Equality, and Liberty. Wildwood House, London.

Beyer, P., 1995. Religion and Globalisation. Sage, London.

Castells, M., 1996. The Power of Identity, the Information Age, vol. II. Blackwell, Oxford, UK.

Choudhury, M.A., Malik, U.A., 1992. The Foundations of Islamic Political Economy. Macmillan, London.

Drucker, P.F., 1993. Post-Capitalist Society. Butterworth-Heinemann, London.

Duchrow, U., 1987. Global Economy: A Confessional Issue for the Churches? WCC, Geneva.

Entemann, W.F., 1994. Managerialism: The Emergence of a New Ideology. University of Wisconsin Press, Madison, WI.

Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.

Green, D.G., 1987. The New Right: The Counterrevolution in Political, Economic and Social Thought. Harvester Wheatsheaf, London.

Gregory, C.A., 1994. Exchange and reciprocity. In: Ingold, T. (Ed.), Companion Encyclopaedia of Anthropology. Routledge, London, pp. 911–937.

Harvey, D., 1989. The Condition of Postmodernity. Blackwell, Oxford, UK.

John Paul II, Pope, 1991. Centesimus Annus. Origins, Washington, DC.

Kuran, T., 2011. The Long Divergence: How Islamic Law Held Back the Middle East. Princeton University Press, Princeton.

Lash, S., Urry, J., 1994. Economies of Signs and Space. Sage, London.

Mann, M., 1986. The Sources of Social Power History of Power from the Beginning to AD 1760, vol. I. Cambridge University Press, Cambridge, UK.

Martin, D., 1990. Tongues of Fire: The Explosion of Protestantism in Latin America. Blackwell, Oxford, UK.

Meeks, M.D., 1989. God the Economist: The Doctrine of God and Political Economy. Fortress Press, Philadelphia, PA.

Novak, M., 1982/1991. The Spirit of Democratic Capitalism. Institute for Economic Affairs, London.

Roberts, R.H., 1992. Religion and the 'enterprise culture': the British experience in the Thatcher era (1979–1990). Social Compass 39 (1), 5–33.

Roberts, R.H. (Ed.), 1995. Religion and the Transformations of Capitalism: Comparative Perspectives. Routledge, London.

Robertson, R., 1992. Globalization: Social Theory and Global Culture. Sage, London.

Sahlins, M., 1974. Stone Age Economics. Tavistock, London.

Schumpeter, J.A., 1954. History of Economic Analysis. Oxford University Press, New York. Allen & Unwin, London.

Simmel, G., 1978. The Philosophy of Money. Routledge and Kegan Paul, London.

Stern, F., 1977. Gold and Iron: Bismarck, Bleichroder, and the Building of the German Empire. Penguin, Harmondsworth, UK.

Thrift, N., 1997. Soft Capitalism. Sage, London.

Troeltsch, E., 1931. The Social Teachings of the Christian Churches. Allen and Unwin, London.

Weber, M., 1930/1976. The Protestant Ethic and the Spirit of Capitalism. Allen and Unwin, London.

Religion and Health: Clinical Considerations and Applications

Harold G Koenig, Duke University Medical Center, Durham, NC, USA; and King Abdulaziz University, Jeddah, Saudi Arabia
Faten N Al Zaben and Saad Al Shohaib, King Abdulaziz University, Jeddah, Saudi Arabia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.L. Idler, volume 19, pp. 13037–13040, © 2001, Elsevier Ltd.

Abstract

This article provides a brief review of research on religion and health and suggested clinical applications. Most of the research reviewed was published between 1960 and 2010 and is based on a systematic review, although more selected recent research has also been included. The majority of the research summarized here indicates that religious involvement is related to better mental health, social health, health behaviors, and physical health overall. We examine here why this might be the case, and then make recommendations on what health professionals should do differently based on these research findings.

Religion and health have been of considerable interest to humans throughout recorded history, although until recently the relationship between the two had not been carefully examined through scientific research. This has changed in the past 25 years, as increasing attention has been paid to the connections between religion and health by psychological, social, behavioral, and medical researchers. Most of the world's population's religious beliefs are often central to how the people view life, and in particular, how they make sense of trauma, loss, and physical illness.

Based on an international survey of 51 927 persons from 57 countries, only 13% of the world's population are atheists, and most of these individuals are concentrated in China (47% atheist) and Japan (31% atheist) (WIN-Gallup International, 2012). Furthermore, only 23% of the world's population are not religious (again, most are concentrated in East Asia). Hindus are most likely to say they are religious (82%) compared to Christians (81%), Muslims (74%), and Jews (38%). Religiosity is highest among the poor (66% of those in the lowest quintile of income) and uneducated (68% of those with less than a secondary education).

The most religious of the 57 countries surveyed were Ghana, Nigeria, Armenia, Fiji, and Macedonia (all over 90%). The least religious were Sweden, Turkey, the Czech Republic, Japan, and China (under 30%). In Turkey (a Muslim country), only 23% described themselves as religious. In the Czech Republic (a Christian country, but until recently Communist), only 20% described themselves as religious. Globally, there was a 9% decline in religiosity from 2005 to 2012. During this time, the countries showing the greatest decline in religiosity were Vietnam (−23%), Switzerland (−21%), France (−21%), South Africa (−19%), and Iceland (−17%). Countries showing an increase in religiosity were Pakistan (6%), Macedonia (5%), Moldova (5%), Serbia (5%), and Malaysia (4%). Interestingly, those *under age* 30 and over age 65 were most likely to describe themselves as religious (60% and 66%, respectively).

What Is Religion?

According to the definition provided in the *Handbook of Religion and Health*, "[Religion] Involves beliefs, practices, and rituals related to the *transcendent*, where the transcendent is God, Allah, HaShem, or a Higher Power in Western religious traditions, or to Brahman, manifestations of Brahman, Buddha, Dao, or ultimate truth/reality in Eastern traditions. This often involves the mystical or supernatural. Religions usually have specific beliefs about life after death and rules about conduct within a social group. Religion is a multi-dimensional construct that includes beliefs, behaviors, rituals, and ceremonies that may be held or practiced in private or public settings, but are in some way derived from established traditions that developed over time within a community. Religion is also an organized system of beliefs, practices, and symbols designed (a) to facilitate closeness to the transcendent, and (b) to foster an understanding of one's relationship and responsibility to others in living together in a community"(Koenig et al., 2012: p. 45). Religion is distinguished from the term "spirituality," which is now understood as a broader concept than religion (although may suffer from measurement contamination when conducting research) (Koenig, 2008).

Is Religion Related to Health?

We examine the research findings related to religion/spirituality (R/S) and health in four sections: mental health, social health, behavioral health, and physical health. This summary is based on a systematic review of peer-reviewed quantitative, original research reports published largely from 1960 through mid-2010. This research is described in more detail in the two editions of the *Handbook of Religion and Health* (Koenig et al., 2001, 2012).

Besides summarizing the research findings, the Handbooks also provide quality ratings for each of the approximately 3300 studies. These quality ratings range from 0 (low) to 10 (high) and are based on six criteria: study design, sampling method, number of R/S measures, quality of measures (both religious predictors and health outcomes), inclusion of control variables, and statistical method. We summarize the research findings below in terms of the total number of studies and of the studies with quality ratings of 7 or higher.

Mental Health

Most of the research on religion and health (80%) focuses on mental health. Of primary importance is that religion is often

used to cope with difficult life circumstances out of the individual's control. We identified 454 such studies that explored the role that R/S played in helping people cope with conditions as diverse as general medical illness, chronic pain, renal disease, diabetes, lung disease, malignancy, disorders of the blood, heart and other cardiovascular diseases, vision and dental problems, neurological syndromes, acquired immunodeficiency syndrome, systemic lupus erythematosus, inflammatory bowel disease, musculoskeletal disease, caregiver burden, chronic psychiatric illness, loss of loved ones, conditions at the end-of-life, natural disasters, war, acts of terrorism, and a wide range of other challenging traumatic events and difficult circumstances. In most of these studies, people reported that R/S was helpful in coping. In some studies of medically ill adults, 90% said religion enabled them to cope at least to a moderate degree and over 40% said it was the most important factor that kept them going.

Mental Disorders

By improving coping, R/S beliefs and practices could influence the development and course of common mental disorders, including depression, suicide, anxiety, psychosis, and substance use disorders.

Depression

At least 444 studies had examined the relationship between R/S and depression prior to mid-2010. Among those, 272 (61%) reported significant inverse relationships between R/S and depression (including nine studies at a statistical trend level) and 28 (6%) reported greater depression in those with more belief in R/S (2 studies at a statistical trend level). Of the 178 studies rated at 7 or higher on quality, 119 (67%) reported inverse relationships and 13 (7%) positive relationships (i.e., greater depression among those who were more religious). At least 30 clinical trials examined the effects of an R/S intervention on depression compared with either standard treatment or a control group. Of those, 19 (63%) reported a significant decline in depressive symptoms in response to R/S interventions compared to controls, whereas 2 studies (7%) found standard treatments superior to R/S interventions.

More recently, psychiatric epidemiologists at Columbia University have published a series of reports on a cohort of children at low or high risk for depression because of parents with depressive disorder. These studies have found that religious involvement may prevent the development of depression in high-risk persons (Miller et al., 2012, 2014; Kasen et al., 2012).

Suicide

The research above is consistent with three-quarters of published research on R/S and suicide attempts, completed suicide, or attitudes toward suicide, which find lower risk among those who are more R/S. Of 141 studies, 106 (75%) have found inverse relationships with R/S, whereas 4 (3%) have reported positive relationships. As the methodological rigor of the studies improves, the likelihood of finding inverse relationships increases from 75% to 80% (i.e., 39 of 49 studies).

Anxiety

Our review identified 299 studies that had examined relationships between R/S and anxiety, of which 147 (49%) found inverse associations with R/S (three at a trend level) and 33 (11%) found greater anxiety. Most of the latter (94%), however, were cross-sectional studies, making it impossible to determine whether R/S was the cause of the greater anxiety or whether great anxiety was the cause for R/S involvement. Quality of the study did not seem to influence the findings. Fifty-five percent (38 of 67 studies with quality ratings of 7 or higher) reported inverse relationships, while 10% (7 of 10) reported greater anxiety among the more R/S. Included here were 32 randomized clinical trials, of which 22 (69%) found that R/S interventions were more effective in reducing anxiety than standard treatment or control conditions (only one study reporting worse outcomes in the R/S intervention group and that was among individuals suffering from severe alcohol dependence).

Psychosis

Relationships with psychosis are less clear. Only about one-third (14 of 43) of studies have found inverse relationships between R/S and psychotic symptoms, whereas one-quarter (10 of 43) of studies have reported positive relationships. Two studies that reported inverse relationships between R/S and psychosis were prospective in nature, where those who had more belief in R/S had better outcomes. The reader is referred elsewhere for a more comprehensive discussion of R/S and psychotic disorders (and ways that clinicians can distinguish the two) (Koenig, 2011).

Alcohol and Drug Use/Abuse

While the relationship between R/S and psychotic symptoms may be somewhat unclear, there is no doubt about associations with alcohol and drug use. Of the 278 studies that examined relationships between alcohol use/abuse and R/S, 240 (86%) have found inverse relationships, whereas only 4 studies reported a positive relationship. Studies with the best methodology report the same findings (i.e., 90% of 145 studies with quality ratings of 7 or higher report inverse relationships with R/S). Relationships between R/S and drug use/abuse are similar. Of 185 studies, 84% have found inverse relationships, whereas only 2 studies reported positive relationships. Again, the findings were not influenced by the quality of the study. Note that most of these studies were conducted on high school and college students, which are a particularly vulnerable group and this is the time when youth are beginning to establish substance use patterns that will influence the rest of their lives.

Positive Emotions

Although of less interest to clinicians, the experience of positive emotions is of considerable importance to individuals in the general population, who may be quite concerned about the relationship between R/S involvement and emotional states such as well-being, happiness, optimism, hope, meaning and purpose, and self-esteem, which often make life worth living (or not).

Well-being

Among the 326 studies that examined relationships between R/S and well-being, happiness, or life satisfaction up through mid-2010, nearly 80% (256 studies) reported significant positive relationships (eight at a statistical trend), compared to <1% that found inverse relationships. Again, quality of the study had little effect on the findings, such that 82% of the 98 best studies reported positive relationships and only one study reported a negative relationship.

Hope

The majority of studies found that R/S individuals are more hopeful. Among the 40 studies we identified in our systematic review, 29 (73%) reported significant positive relationships and no studies found inverse relationships.

Optimism

Of the 32 studies examining relationships between R/S and optimism, 26 (81%) reported significant positive relationships and no studies reported inverse relationships.

Meaning and Purpose

Sigmund Freud said that "…only religion can answer the question of the purpose of life. One can hardly be wrong in concluding that the idea of life having a purpose stands and falls with the religious system" (Freud, 1930). On this point Freud was probably right. Of the 45 studies that have examined relationships with R/S, 42 (93%) reported significant positive relationships, and of the 10 studies with quality ratings of 7 or higher, all (100%) reported significant positive associations.

Self-esteem

Although psychiatrists have argued that R/S may adversely affect self-esteem because it fosters humility and discourages pride (Watters, 1992), the research published to date does not support this claim. Of the 69 studies that have examined relationships with self-esteem, 42 (61%) have reported *greater* self-esteem among the more R/S and only two studies have found lower self-esteem. As the quality of the study increases, the results remain favorable toward R/S, where 17 of 25 (68%) reported greater self-esteem. These findings are consistent with those reported between R/S and depression or suicide, which often result from a loss of self-esteem.

Social Health

Social health involves social support, marital stability, absence of crime and delinquent activities, altruistic activities, and social capital. R/S is related to every one of these, and the direction of those relationships is consistent.

Social Support

We identified 74 studies examining associations between R/S and social support, including a number of social contacts, frequency of interaction with family and friends, and subjective sense of satisfaction with that support. Of those studies 61 (82%) found that R/S was associated with significantly higher social support, and none found lower social support. Nearly all of the highest quality studies (27 of 29) found greater social support among the more religious. This is especially true for older adults in whom members of the faith community are often the most important source of support outside of family members (Cutler, 1976; Koenig et al., 1988).

Marital Stability

Of the 79 studies we identified up through mid-2010, 68 (86%) reported significant relationships between R/S and lower rates of divorce, higher rates of two-parent homes, greater marital satisfaction, and lower rates of spousal abuse. Among studies with the best methodology (ratings of 7 or higher), 35 of 38 studies reported this finding. These results are consistent with meta-analytic reviews of the literature by others (Mahoney et al., 2001) and by more recent research published in this field (Spilman et al., 2013; David and Staffor, 2013).

Delinquency/Crime

Of the 104 studies that examined relationships with R/S, again the vast majority (79% or 82 or 104) found inverse relationships with delinquency during youth and crime during adulthood. Only three studies have reported the opposite. Of the 60 methodologically most rigorous studies, the same proportion (82% or 49 of 60 studies) found less delinquency and crime among the more religious.

Altruistic Activities

Our systematic review of R/S and altruism/volunteerism found that 70% (33 of 47 studies) reported significant positive relationships, and if organ donation studies are excluded (since some religions prohibit this practice), the percentage of positive studies increases to 79%.

Social Capital

Social capital is a term used to describe the overall health of a community, and is often assessed by measuring level of community participation, trust, reciprocity between community members, rates of volunteerism, and proportion of the community that participates in civic, political, or social justice organizations. Again, the vast majority of these studies (79% or 11 of 14 studies) reported that R/S of members in a community relates to significantly greater levels of social capital.

Behavioral Health

Religious beliefs and doctrines may influence lifestyle choices related to physical health by emphasizing responsibility to care for and nourish the body (while avoiding gluttony). Behaviors that harm the body are discouraged, since good health enables the religious person to serve God and others. We review here studies that have examined relationships

between R/S and cigarette smoking, exercise, weight, diet, and sexual behavior.

Cigarette Smoking

At least 137 studies that examined the relationship with cigarette smoking, and of those, 90% (123 studies) reported that R/S is related to significantly less cigarette smoking (three at a trend level). No studies found positive relationships with smoking. Likewise 75 of the 83 highest quality studies reported this finding. Many of these were among youth. Consider the effects that this single health behavior is likely to have, given the fact that cigarette smoking is associated with 30% of all cancers, chronic lung disease, and increased rates of coronary artery disease, hypertension, and stroke later in life.

Exercise

In general, those who are more R/S also tend to be more physically active. We found that two-thirds of studies (25 of 37) examining this relationship found higher rates of exercise or greater physical activity among the more R/S, compared to six studies (16%) that found less activity. Of the 21 best studies, 16 (76%) reported positive associations and 2 (10%) found negative associations.

Weight

Unfortunately, greater exercise and less cigarette smoking do not translate into lower weight among the religious. Of all the health characteristics, this one tends to be problematic among the religious. We identified 36 studies that examined the relationship between weight (body mass index) and R/S involvement, of which 14 (39%) found that religiosity was related to *greater* weight, whereas only 7 studies reported lower weight (19%). The findings do not change among the higher quality studies (44% greater weight, 20% lower weight). Eating seems to be the one sin that is not only allowed but encouraged in faith community settings. The Amish, Jews, and Buddhists tend to have lower weight than Christians and Muslims.

Diet

Although weight is a problem, at least the composition of food intake on average tends to be better among the more R/S. Of the 21 studies that have examined relationships with diet, R/S were more likely (62% of studies) to have increased intake of fiber, green vegetables, fruit, and fish; low intake of snacks, processed foods, and fat; regular vitamin intake; frequent eating of breakfast; or overall better nutrition. Among the highest quality studies, 7 of 10 found a better diet among the more religious. Cholesterol levels also tend to be lower. Of the 23 studies that examined this relationship, 12 found lower cholesterol levels among the more R/S, whereas 3 found higher cholesterol.

Sexual Behavior

Given the relationship between extra-marital sexual activity and a host of sexually transmitted diseases, this is another behavior likely to influence health status and susceptibility to disease (syphilis, gonorrhea, herpes, chancroid, chlamydia, viral hepatitis, human papillomavirus, human immunodeficiency virus, etc.). Of the 95 studies that examined relationships between R/S and sex outside of marriage (premarital, extra-marital, number of partners, etc.), 82 (86%) found lower and 1 study (1%) found higher rates of risky sexual activity among the more R/S. Among the 50 highest quality studies, the findings were similar (84% inverse relationships, 0% positive relationships).

Physical Health

Given the well-known relationships between physical health and mental, social, and behavioral factors, a relationship between R/S and physical health would be expected. Such relationships, however, are often – if not always – indirect and act through psychosocial–behavioral pathways. Here we examine the relationships between R/S and coronary heart disease (CHD), blood pressure (BP)/hypertension, cerebrovascular disease, dementia, immune and endocrine functions, cancer, and overall mortality.

Heart Disease

Psychosocial and behavioral factors are known to influence the development and course of CHD. We would expect R/S involvement to do likewise. We identified 19 studies that had examined associations between R/S and CHD up through 2010. Nearly two-thirds (12 studies or 63%) indicated lower rates among those who were more R/S, whereas one study found a significant positive association. Similarly, nine of the 13 most rigorously designed studies (69%) reported less CHD among the more R/S. This research is supported by studies examining relationships between R/S and a wide range of cardiovascular functions or outcomes, including cardiovascular reactivity, heart rate variability, cardiac surgery outcomes, and other vascular functions. Eleven of 16 such studies found positive relationships between R/S and these functions or inverse relationships with inflammatory markers (C-reactive protein) or coagulation factors (fibrinogen) that increase the risk of vascular disease.

Hypertension

Our systematic review identified 63 studies examining associations between R/S and BP or rates of hypertension. As expected, the majority (36 studies or 57%) reported inverse relationships, compared to 7 studies (11%) that found significantly higher BP or hypertension. Among the 39 best designed studies, 24 (62%) reported lower BP among those who are more R/S or in response to an R/S intervention.

Cerebrovascular Disease

Since vascular disease and hypertension are strong predictors of stroke and other forms of cerebrovascular disease, R/S involvement should likewise influence these disorders. Of the

nine studies we identified that examined this, four reported a lower risk of stroke among the more R/S, whereas one paradoxically found greater carotid artery thickening (increasing stroke risk) (Feinstein et al., 2010).

Alzheimer's Disease and Dementia

Vascular disease also increases the risk of vascular dementia and the risk of Alzheimer's disease in older adults. Depressive disorder and high stress may also contribute to the development of dementia acting through elevated cortisol levels that adversely affect large neurons in the temporal lobes (Betz cells) of the brain. By reducing vascular disease and emotional stress, R/S might influence the course of memory loss later in life. Furthermore, the higher cortical functions activated as a result of religious beliefs, reading, and thinking about theological issues, may help to preserve these functions. Our systematic review identified 21 studies that examined associations between R/S and cognitive function in older adults. Half of these studies (10 or 48%) found better cognitive functioning in the more R/S and three reported worse cognitive functioning. Higher quality studies, including several that examined these relationships prospectively, were also likely to report such findings (8 of 14 studies or 57%). Recent research has continued to find better cognitive functioning among older adults who are more religious (Coin et al., 2010; Zhang, 2010; Inzelberg et al., 2013).

Immune Function

The immune system is responsible for preventing disease and preserving internal homeostasis. Our systematic review identified 27 studies that examined relationships with R/S. The majority (15 of 27 or 56%) reported positive associations or improved immunity in response to an R/S intervention. Only one study reported an inverse association. Likewise, 10 of the 14 highest quality studies reported better immune function among the more R/S. Better immune function is also reflected in susceptibility to infection. We identified an additional 12 such studies that examined R/S and infection rates or virus concentrations in blood (viral load). Two-thirds (67%) of these studies reported lower infection rates or viral load in those who were more religious, and the higher quality studies reported similar results (70%).

Endocrine Function

Endocrine functions of particular interest here are the stress hormones – cortisol, epinephrine, norepinephrine. Among 31 studies that examined levels of R/S or R/S interventions, 23 (74%) reported inverse associations or positive effects (most studies examining Eastern meditation) on stress hormone levels. No studies reported higher levels of stress hormones in those who are more religious or in response to R/S interventions.

Cancer

Among the 29 studies examining relationships between R/S and the development or course of cancer, 16 (55%) reported lower rates or better prognosis among those who had more belief in R/S, whereas 2 studies reported worse prognosis. Similar findings were reported among the methodologically more rigorous studies (60% better outcomes, none with worse). These findings may be explained in part by better health behaviors (less cigarette smoking, alcohol intake, etc.), although not entirely.

Mortality

If R/S involvement is related to better mental, social, behavioral, and physical health over the lifetime, these effects should manifest themselves in terms of longevity. We identified 121 studies examining the ability of R/S to predict survival (most being prospective cohort studies). Of those, 82 (68%) found greater longevity among the more religious (three at a trend level), whereas 6 studies (5%) found a worse survival rate. Of the 63 best designed studies (largest samples, best R/S measures, longest follow-up), 47 (75%) reported greater survival among the more religious. These findings have been supported by at least three other systematic reviews (McCullough et al., 2000; Powell et al., 2003; Chida et al., 2009). The effect on longevity has been compared to the effect of cholesterol lowering drugs or exercise-based cardiac rehabilitation on survival following myocardial infarction (McCullough, 2001).

How Religion Affects Physical Health

We have already hinted at the mechanisms likely responsible for the effects that R/S has on physical health and longevity. We summarize and integrate these psychosocial, behavioral, and physiological pathways here.

Religious beliefs and behaviors are important in helping people to cope with adversity and life stress, giving meaning to loss and change. This reduces the likelihood of developing stress reactions, depression, anxiety, and resorting to alcohol and drugs for comfort, and increases the likelihood of experiencing positive emotions such as hope, optimism, and well-being, enhancing outlook and self-esteem. Religious involvement also increases social contacts and may enhance the quality of social interactions, since religious doctrines promote the love of neighbor, altruistic acts, and the practice of forgiveness. The psychological and social benefits of religion likely translate into consequences for physical health as well. Physiological changes in immune, endocrine, and cardiovascular functions help to preserve homeostatic mechanisms which help to prevent disease or facilitate recovery. In addition, religious doctrines that discourage behaviors that adversely affect health, such as cigarette smoking, excessive alcohol use, drug use, risky sexual behavior, and encourage those behaviors that preserve health, such as physical activity and attention to diet, have physical health consequences.

There are other pathways by which R/S likely influences health, and by which mental and physical health probably influence R/S belief and activity. There is no doubt genetic and developmental factors influence health and well-being, and there is growing evidence that these factors may influence religious beliefs, practices, and experiences as well (Dew and Koenig, 2014), adding complexity to the religion–health relationship. Personality or temperament (strongly

influenced by genetic factors and early childhood experiences) may help to determine whether a person engages in religious activities, and may also influence psychological, social, and behavioral factors that affect physical health and susceptibility to disease. These factors must be taken into account when examining the religion–health relationship, including the fact that those with the least resources (the poor, uneducated, marginalized minority groups) are often the most religious. As Karl Marx often quoted, religion is the "opiate of the masses" (i.e., those with the most disadvantages).

Clinical Applications

Clinical applications that result from the relationships found here are vast. The primary application at the patient level is that health professionals should be willing to identify and discuss patients' religious or spiritual beliefs with them. This is done primarily by taking a spiritual history from patients with chronic or serious medical or psychiatric illness, and then ensuring that spiritual needs that arise during such discussions are addressed by those trained to do so (typically chaplains or pastoral counselors). Resources now exist that will help guide health professionals in this task (Koenig, 2013).

Summary

Religious beliefs and practices are often used to cope with mental and physical illness. Studies show that those who are more R/S experience better mental health and adapt better to medical illness, stress, and trauma. They also have larger and more supportive social networks, and live healthier lifestyles, avoiding behavior likely to adversely affect their health. These psychological, social, and behavioral influences are also likely to have consequences for physical health, susceptibility to disease, and ultimately on survival and mortality, as the vast majority of research now demonstrates. These relationships with health, along with the widespread presence of spiritual needs during serious or chronic medical illness, have implications for clinical practice. Health professionals should begin to regularly inquire about the religious or spiritual beliefs of patients and be willing to discuss these issues as they relate to the health care being provided.

See also: Bereavement; Healing Practices: Alternative and Complementary; Stress, Coping and Health.

Bibliography

Chida, Y.C., Steptoe, A., Powell, L.H., 2009. Religiosity/spirituality and mortality: a systematic quantitative review. Psychotherapy and Psychosomatics 78, 81–90.

Coin, A., Perissinotto, E., Najjar, M., Girardi, A., Inelmen, E.M., Enzi, G., Manzato, E., Sergi, G., 2010. Does religiosity protect against cognitive and behavioral decline in Alzheimer's dementia? Current Alzheimer Research 7 (5), 445–452.

Cutler, S.J., 1976. Membership in different types of voluntary associations and psychological well-being. The Gerontologist 16, 335–339.

David, P., Staffor, L., 2013. A relational approach to religion and spirituality in marriage: the role of couples' religious communication in marital satisfaction. Journal of Family Issues [E-pub ahead of print].

Dew, R.E., Koenig, H.G., 2014. Religious involvement, the serotonin transporter promoter polymorphism, and drug use in young adults. International Journal of Social Sciences Studies 2 (1), 98–104.

Feinstein, M., Liu, K., Ning, H., Fitchett, G., Lloyd-Jones, D.M., 2010. Burden of cardiovascular risk factors, subclinical atherosclerosis, and incident cardiovascular events across dimensions of religiosity: the Multi-Ethnic Study of Atherosclerosis. Circulation 121, 659–666.

Freud, S., 1930. Civilization and its discontents. In: Strachey, J. (Ed. and Trans.), Standard Edition of the Complete Psychological Works of Sigmund Freud. Hogarth Press, 1962, London, p. 25.

Inzelberg, R., Afgin, A., Massarwa, M., Schechtman, E., Israeli-Korn, S., Strugastsky, R., Amin, A., Efrat, K., Lindsay, A.F., Friedland, R., 2013. Current Alzheimer Research 10 (3), 340–346.

Kasen, S., Wickramaratne, P., Gameroff, M.J., 2012. Religiosity and resilience in persons at high risk for major depression. Psychological Medicine 42 (3), 509–519.

Koenig, H.G., Moberg, D.O., Kvale, J.N., 1988. Religious activities and attitudes of older adults in a geriatric assessment clinic. Journal of the American Geriatrics Society 36, 362–374.

Koenig, H.G., 2008. Concerns about measuring "spirituality" in research. Journal of Nervous & Mental Disease 196 (5), 349–355.

Koenig, H.G., McCullough, M.E., Larson, D.B., 2001. Handbook of Religion and Health, first ed. Oxford University Press, New York.

Koenig, H.G., 2011. Schizophrenia and other psychotic disorders (chapter 2). In: Peteet, J., Lu, F.G., Narrow, W.E. (Eds.), Religious and Spiritual Issues in Psychiatric Diagnosis: A Research Agenda for DSM-V. American Psychiatric Association Press, Arlington, VA, pp. 31–52.

Koenig, H.G., King, D.E., Carson, V.B., 2012. Handbook of Religion and Health, second ed. Oxford University Press, New York.

Koenig, H.G., 2013. Spirituality in Patient Care, third ed. Templeton Press, Conshohocken, PA.

Mahoney, A., Pargament, K.I., Tarakeshwar, N., Swank, A.B., 2001. Religion in the home in the 1980s and 1990s: a meta-analytic review and conceptual analysis of links between religion, marriage, and parenting. Journal of Family Psychology 15 (4), 559–596.

McCullough, M.E., Hoyt, W.T., Larson, D.B., Koenig, H.G., Thoresen, C., 2000. Religious involvement and mortality: a meta-analytic review. Health Psychology 19 (3), 211–222.

McCullough, M., 2001. Religious involvement and mortality. In: Plane, T.G., Sherman, A.C. (Eds.), Faith and Health: Psychological Perspectives. Guilford Press, New York, NY, pp. 53–74.

Miller, L., Wickramaratne, P., Gameroff, M.J., Sage, M., Tenke, C.E., Weissman, M.M., 2012. Religiosity and major depression in adults at high risk: a ten-year prospective study. American Journal of Psychiatry 169 (1), 89–94.

Miller, L., Bansai, R., Wickramaratne, P., Hao, X., Tenke, C.E., Weissman, M.M., 2014. Neuroanatomical correlates of religiosity and spirituality in adults at high and low familial risk for depression. Journal of the American Medical Association-Psychiatry 71 (2), 128–135.

Powell, L.H., Shahabi, L., Thoresen, C.E., 2003. Religion and spirituality. Linkages to physical health. American Psychologist 58 (1), 36–52.

Spilman, S.K., Neppl, T.K., Donnellan, M.B., Schofield, T.J., Conger, R.D., 2013. Incorporating religiosity into a developmental model of positive family functioning across generations. Developmental Psychology 49 (4), 762–774.

Watters, W., 1992. Deadly Doctrine: Health, Illness, and Christian God-Talk. Prometheus Books, Buffalo, New York.

WIN-Gallup International, 2012. Global Index of Religiosity and Atheism. Gallup International Asociation, Washington, DC. http://www.wingia.com/en/news/win_gallup_international_ae_religiosity_and_atheism_index_ao_reveals_atheists_are_a_small_minority_in_the_early_years_of_21st_century/14/.

Zhang, W., 2010. Religious participation, gender differences, and cognitive impairment among the oldest-old in China. Journal of Aging Research 2010, 160294.

Religion and Health: Sociological Considerations

Ellen L Idler, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Empirical studies and theoretical reasoning about religion and health began with the origins of the discipline of sociology. Theoretical frameworks draw on Durkheim's concepts of social integration and regulation to understand religion as a determinant of health, and on Weber's concept of *verstehen* to understand religion as a source of meaning in illness. Physical health measures include mortality, morbidity, and disability; mental health measures include well-being, disorders of depression, substance abuse, anxiety, and psychoses. Multiple explanatory mechanisms for the relationship, including reduced risk behaviors, enhanced social support, and effective coping strategies, have been proposed.

There are two distinct ways in which religion may be related to health. Religions – complicated systems of symbols, beliefs, practices, and the social institutions associated with them – may be considered among the psychosocial factors that have apparent influence on health status. In this view, religiousness may be investigated as a protective or preventive factor in reducing mortality or morbidity. A second view reverses this causal order, as religious beliefs and practices are often turned to by individuals and families faced with serious illness, or at the end of life, and religiousness can be affected by changes in aging or health status. Research approaching religion and health from both perspectives has increased dramatically since the 1980s. The comprehensive Handbook of Religion and Health, first published in 2000, was revised in 2012 (Koenig et al., 2012); the authors estimate that there are over 2800 original data-based quantitative studies of some aspect of religion/spirituality and health or clinical practice now published in peer-reviewed journals.

Religion as a Factor Affecting Health

> The earliest examples of practices designed to promote the public health are to be found, among primitive peoples, inextricably mingled with the ritual of religion.

These are the opening words of C-E. A. Winslow in the entry for 'Public Health' in the original 1935 Encyclopedia of the Social Sciences, an early precursor to the present volume. Winslow, professor of public health at the Yale School of Medicine and known for his studies of the decline in infectious disease mortality, drew attention to early quarantine efforts in ancient Persia and among the Biblical Hebrews as effective public health practices, motivated and given form by religious beliefs. The burial of the dead, the separation of lepers from the community, or the ritualistic preparation of meat for consumption are but a few of the early examples of practices carried out for religious reasons that had unintended public health consequences. In general, this characteristic of being unintended is typical of the work discussed under this heading – religious practices or beliefs seen to have associations with health are not adopted for the purpose of improving individual or community health, but rather, are valued ends in themselves.

The social scientific origins of both theoretical and empirical work on religion and health can be found in Emile Durkheim's (1951) *Suicide*. Durkheim demonstrated, with the use of European national and regional statistical records, that suicide rates varied dramatically and predictably with certain social characteristics, among them the rates of Protestants, Catholics, and Jews in a population. His finding that suicide rates were lower in regions with larger Catholic and Jewish populations led him to denote types of suicides (anomic and egoistic) in modern societies. Suicides were prevented in religious and other social groups that regulated or constrained the behavior of individual members and also provided them with integrative and socially supportive relationships.

Despite this foundation in late nineteenth-century classical sociological thought, research on religion was largely absent from social scientific work in health until the later half of the twentieth century. When such work did appear, it was limited to studies of small, peripheral subpopulations, or to suicide as the only cause of death. Beginning in the 1960s, studies finding lower rates of mortality and morbidity (particularly cancer and hypertension) for clergy, members of religious orders, and certain well-established sectarian groups including Seventh-Day Adventists, Mormons, and Amish, appeared in US medical and public health literature (Jarvis and Northcott, 1987). Work on religiousness and suicide rates also appeared steadily, with generally consistent support for Durkheim's original findings of differences by religious affiliation and also lower rates for those with any religious affiliation (Stack, 2000). All of these studies were based on the comparison of group rates.

An individual-level approach to these questions first came in 1979, when the well-known Alameda County (California, USA) analysis of social networks and all-cause mortality appeared (Berkman and Syme, 1979). It was quickly followed by similar findings from large population studies in Tecumseh County, Michigan (USA), and Evans County, Georgia (USA). In each of these prospective, longitudinal studies, social ties – including ties to religious congregations – were associated with lower overall mortality risks.

These studies were important for many reasons. Until this time, the social environment, long a member of the agent–host–environment triad in multifactorial models of disease causation, had primarily been thought of as having negative influence on health. Social environments were considered sources of stress; social conditions were frequently operationalized as poverty, urban crowding, time-pressured working conditions, or the stress of modern life. The social network studies opened for consideration the possibility that the social environment also provided positive or protective factors in health. The first social network studies were also important because they included multiple types of social ties, including some measure of affiliation with a religious group or attendance at religious services.

Since the 1980s, studies primarily from the United States, but also from Israel, Denmark, and Taiwan, focused specifically on religion as a mortality risk factor; most showed a substantial and significant protective effect of religious participation, even after adjustments for possible selection factors, confounders, and mediators (Idler, 2011). For example, Hummer et al. (1999) found, in a US nationally representative sample, that frequent attendance at religious services reduced mortality risk such that 7 years was added to life expectancy at age 20. These findings were different from the earlier group-level studies of homogeneous religious groups or clergy, in that they represented general populations, with their natural heterogeneity in both types and levels of religious behaviors and beliefs. Like the individual-level social network studies that included other types of social ties, they were longitudinal, and thus were able to adjust for the health status of respondents at the start of the study. Data from prospective observational studies with well-designed health status assessment at baseline is critical, particularly when measurement of the religion variable is limited to the indicator of attendance at religious services. The potential selection effects of healthier members of the population attending religious services more frequently, and being at simultaneously lower risk of mortality, are obvious. Additional important selectivity variables included in these analyses include age, sex, socioeconomic status, education, ethnicity, and marital status, as religious participation is generally higher for women, married couples, blacks, older people, and those with higher education and income. This research is comprehensively summarized by Koenig, King, and Carson; based on evidence from three meta-analyses of the research, they estimate a 37% higher probability of survival for religious participants compared with nonparticipants, after selectivity variables are included (2012, p. 478).

However, a chief criticism of the studies in this rapidly growing field is that most contain only one or two indicators of religiousness, an inadequate level of conceptualization and measurement for such a highly complex set of practices, beliefs, and institutions. Most frequently, attendance at religious services has been the only available measure. When some measure of private religious feelings or practices has been included and compared, attendance is usually the measure with the largest protective effects (Chida et al., 2009). Better measurement of the independent variable is clearly necessary before a larger understanding of explanatory mechanisms can be gained. At the present stage of research, however, several mechanisms through which religiousness might affect mortality have been described.

Reduction of Behavioral Risk Factors

The health economist Victor Fuchs observed in the early 1970s that religious teachings regulate and constrain human behaviors in numerous ways that have known health consequences (Fuchs, 1974, pp. 52–55). He noted the enormous differences in infant, child, adolescent, and adult mortality rates in the two adjacent states of Utah and Nevada. Population density, climate, health care, income, and education levels in the two states were rather similar; any advantage went to Nevada. Yet death rates in Nevada for males aged 40–49 were 54% higher, and for females 69% higher than in Utah. When the causes of death were limited to cirrhosis of the liver and cancer of the respiratory system, Fuchs found that the corresponding figures were 111% and 296%. He attributed the dramatic differences to the 'different lifestyles' of the residents of the two states, Utah being primarily inhabited by Mormons who do not drink alcohol or smoke tobacco and 'in general lead stable, quiet lives' with high marriage and fertility rates and low divorce rates. The situation 40 years later is not different; Utah is the state with by far the lowest smoking rate in the United States and with one of the longest life expectancies.

Most religious groups, however, do not have specific, doctrinal beliefs regarding these behavioral health risk factors, and yet rates for cigarette smoking and substance use are significantly lower, and health promotion activities including exercise and the use of preventive medical care are higher among those who attend religious services regularly, even in studies of religiously heterogeneous populations (Koenig et al., 2012). A US national study of high school seniors found that the more frequent attenders, and those who thought religion was important in their lives, had lower rates of smoking, alcohol, and marijuana use, carrying weapons, getting into fights, and driving while drinking. They also had higher rates of positive health practices such as seat belt use, and eating fruits, vegetables, and breakfast (Wallace and Forman, 1998).

Religious involvement among adolescents that lowers many behavioral health risks is especially important because it represents a lowering of lifetime risks for behaviors (like smoking) that are most likely to begin in adolescence. Moreover, an analysis of data from the Alameda County study shows that religious attenders who do smoke are more likely than nonattenders to quit smoking during the course of a follow-up study, and hence further reduce their mortality risk (Strawbridge et al., 1997). Thus one of the mechanisms through which religion influences health consists of the constraints placed on behaviors that are sensation-seeking, self-stimulating, and that increase risk to the self and others. The reduction in lifestyle risk factors results in lowered rates of lifestyle-related causes of death, and in at least one study a temporal pattern of religious practices preceding risk reduction behaviors has been established.

Enhancement of Social Support Networks

A second mechanism by which the effects of religious involvement may operate is through the expansion and maintenance of social support networks. The study of religion and health is in some ways derivative of the study of social networks and health, and indeed the latter studies provided the earliest modern findings, in Alameda County and elsewhere, as noted above. This mechanism represents the parallel strain in the Durkheimian theoretical tradition, that religious institutions provide both regulation for the behavior of members, as well as integration, or feelings of belonging in a social group. Religious congregants report larger social networks of both friends and relatives, and also more social support from them, compared with nonattenders. There are several reasons why religious congregations would serve as centers of effective social support for members of all ages, and especially for the elderly. Congregations offer a mix of strong and weak ties, and can thus offer a broad range of emotional and instrumental supports. Congregations of all major faiths share values and beliefs about the importance of helping others who are less fortunate and also provide organizational structure for mounting such efforts. Religious institutions provide both support and opportunities to support others, and reciprocal or bidirectional social support may be the most beneficial. Finally, as a corollary of this, members of religious groups who have served others in the past, and who have seen others receiving help, have reason to anticipate support in the future in the event of need.

Population studies of religiousness and health outcomes, particularly mortality, have generally included some assessment of these two explanatory mechanisms (e.g., Hummer et al., 1999; Strawbridge et al., 1997). Usual findings have been that (1) religious involvement, particularly public participation in services, is associated both with better health practices and with higher levels of social support, and (2) these known behavioral and psychosocial risk factors explain some proportion (but not all) of the effect of religiousness on later health outcomes. In studies of healthy populations, then, religion (conventionally operationalized as attendance at religious services) appears to be related to a lower risk of mortality after adjustment for known confounders and selectivity variables, and the effect is partly explained by existing explanatory mechanisms and partly independent of them. This is still an emerging field, however (Ellison and Hummer, 2010). The number of prospective, longitudinal, population-based studies is relatively few, the measurement of the religion variable is unsophisticated, the research has been largely done in the United States in regional samples with Judeo-Christian beliefs, and other explanatory mechanisms should be tested.

Illness and the Search for Meaning

The first part of this article has been concerned with the unintended health consequences of religious practices, or measurable changes in health that result from religious practices undertaken for religious – not health-related – purposes. Equally important for a social scientific study of religion and health, however, are religious beliefs and practices that are engaged in as an individual and social response to illness. Religious groups may offer not only spiritual and tangible support to the sick, including prayers and comforting rituals, but also practical help such as meals, visitation, childcare, and financial support.

Research on the effects of religious coping has frequently taken indicators of mental health and well-being as outcomes. There is a voluminous literature on the types of religious coping, both positive and negative, undertaken by persons with physical illnesses (especially cancer, cardiovascular, HIV/AIDS, autoimmune disorders) and mental illnesses (especially depression, anxiety, substance abuse, schizophrenia) (Koenig et al., 2012). Importantly, not all forms or styles of religious coping may be benign, for example, when the cause of an illness is attributed to moral failure; research is just beginning on maladaptive religious coping and religious struggle (Pargament, 1997).

Classical sociological theories of religion again anticipate these functions of religion in health. Weber's concepts of *verstehen* (understanding the meaning of an action from another's point of view) and theodicy (religious explanations for suffering in the world) provide historical models for the articulation of the problems of individuals with religious language. Religious individuals can draw on the experiences of fellow believers or the stories of their faith tradition to put their own lives into a larger context, to learn lessons from others who have faced similar circumstances, to reduce fear, or to gain hope for the future. Religious ritual practices for mourning or for healing can assist individuals in making transitions to new life stages in which adaptation to loss is required. Religious language and interpretations appear prominently in the narratives of individuals faced with their own aging, or serious illness, or the illness or death of a loved one, situations over which the individual has little direct control, and which may lack clear explanations or challenge beliefs that the world is just (Pargament, 1997). In contrast with the population studies described above, in which attendance at worship services was most predictive of later health outcomes, studies of religious coping frequently find that subjective religious beliefs or nonorganizational forms of religious practice (e.g., private prayer) play an important role.

Religious beliefs and practices are especially prominent in the care for and concerns of individuals at the end of life (Kellehear, 2007). The modern hospice, begun in England in the 1960s, was modeled after medieval Christian institutions that cared for the dying, and the spiritual concerns of patients are still a central aspect of the specialty of palliative medicine. There is also some evidence that the observance of religious holidays may alter the timing of death for elderly individuals; multiple studies in both Jewish and Christian populations have observed short delays before death that permit the celebration of a last sacred ritual (Idler and Kasl, 1992).

Despite a background in the Durkheimian and Weberian traditions of the sociology of religion, the field of research in religion and health remained mostly unexplored through the century, but began growing quite rapidly in the 1990s, especially in the United States, where levels of interest in religion

are generally higher than in Europe. One may anticipate further growth in this field, as the measurement and conceptualization of the relevant domains of religiousness and spirituality improve, and as the population ages and the prevalence of chronic illness increases.

See also: Adolescent Health and Health Behaviors; Aging and Health in Old Age; Bereavement; Chronic Illness, Psychosocial Coping With; Depression, Pessimism, and Health; Health Behavior, Psychosocial Theories of; Religion, Phenomenology of; Religion, Psychology of; Religion, Sociology of; Religion: Psychiatric Aspects; Social Epidemiology; Social Integration, Social Networks, and Health.

Bibliography

Berkman, Lisa F., Syme, S. Leonard, 1979. Social networks, host resistance, and mortality. American Journal of Epidemiology 109, 186–204.

Chida, Yoichi, Steptoe, Andrew, Powell, Lynda H., 2009. Religiosity/spirituality and mortality: a systematic quantitative review. Psychotherapy & Psychosomatics 78, 81–90.

Durkheim, Emile, 1951. Suicide. Free Press, Glencoe, IL.

Ellison, Christopher G., Hummer, Robert A., 2010. Religion, Families, and Health: Population-Based Research in the United States. Rutgers, New Brunswick, NJ.

Fuchs, Victor R., 1974. Who Shall Live? Health, Economics, and Social Choice. Basic Books, New York.

Hummer, Robert A., Rogers, Richard G., Nam, Charles B., Ellison, Christopher G., 1999. Religious involvement and US adult mortality. Demography 36, 273–285.

Idler, Ellen L., 2011. Religion and adult mortality: group- and individual-level perspectives. In: Rogers, Richard G., Crimmins, Eileen M (Eds.), International Handbook of Adult Mortality. Springer, Dordrecht, pp. 345–377.

Idler, Ellen L., Kasl, Stanislav V., 1992. Religion, disability, depression, and the timing of death. American Journal of Sociology 97, 1052–1079.

Jarvis, George K., Northcott, Herbert C., 1987. Religion and differences in morbidity and mortality. Social Science and Medicine 25, 813–824.

Kellehear, Allan, 2007. A Social History of Dying. Cambridge University Press, Cambridge.

Koenig, Harold G., King, Dana E., Carson, Verna B., 2012. Handbook of Religion and Health, second ed. Oxford, New York.

Pargament, Kenneth I., 1997. The Psychology of Religion and Coping. Guilford, New York.

Stack, Stephen, 2000. Suicide: a 15-year review of the sociological literature. Part I: cultural and economic factors. Suicide and Life-Threatening Behavior 30, 145–162.

Strawbridge, William J., Cohen, Richard D., Shema, Sarah J., Kaplan, George, 1997. Frequent attendance at religious services and mortality over 28 years. American Journal of Public Health 87, 957–961.

Wallace, John M., Forman, Tyrone, 1998. Religion's role in promoting health and reducing risk among American youth. Health Education and Behavior 25, 721–741.

Religion and New Media

Heidi A Campbell, Texas A&M University, College Station, Texas, TX, USA
Louise Connelly, University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Understanding how new media (virtual worlds, the Internet, blogs, social networking, and mobile phones) are being used by religious groups, organizations, and individuals can highlight how social and behavioral changes are manifesting within the twenty-first century. It can be suggested that the prolific uptake of new media has had a direct impact on religious identities, communities, and practice both online and offline. For many, the use of new media enables them to engage with religion on a daily basis. However, it is also being argued that religious use of new media can impact upon users' traditional religion, attitudes, and behaviors.

Introduction

The increasing uptake and integration of new media (such as blogs, social networking, and mobile phones) in the Western society has impacted many facets of contemporary culture including the practices of various religious groups, organizations, and individuals. Most notably this impact can be seen in how religious rituals, identities, and communities are performed and practiced, both online and offline. We argue that by exploring the practices of religion online we can more clearly understand how religion is manifested and conceived of in the wider society.

Over the last 20 years scholars have been paying increasing attention to the relationship between new media, religion, and digital culture, considering how various religious traditions have engaged with online mediums including Paganism (Cowan, 2005), Christianity (Cheong et al., 2012), Islam (Bunt, 2009), new religious movements (Baffelli et al., 2011), and Buddhism (Grieve, 2010). Such examinations of religion on the Internet have been undertaken by scholars of various fields such as psychology, sociology, anthropology, media studies, cultural studies, and within other fields such as religious studies. These investigations have raised a number of questions about the nature of religion and religious practice in digital contexts, as well as highlighting areas needing further examination, such as the four key themes of authority, community, identity, and ritual.

What we are concerned with here are current trends and findings within the study of religion and the Internet and, based on this scholarship, claims about how new media use influences, religious organizations, and individuals. For many, the use of new media enables them to more easily integrate religious practice into their daily lives, for example, accessing the *Qu'ran Reader* for the iPhone, reading the Dalai Lama's tweets or participating in a virtual church, such as the Anglican Cathedral in the online world of *Second Life*. It is also argued that the use of new media can have a direct impact on traditional religious practices and behaviors. Therefore, by analyzing who is using the Internet we can consider whether there are and could be changes to religious authority, community, identity, and ritual – both online and offline.

The Use of New Media and Religion

New media are being used by individuals, organizations, and communities in order to engage with religion both online and offline. The expectation or belief that online religion is distinct from offline religion is no longer accepted as the "Internet constitutes the space where individuals and groups live out their social and spiritual lives, and offline boundaries and relations often inform the online sphere" (Campbell and Lovheim, 2011: p. 1083). Exploring how individuals and groups are using new media can provide an insight into the relationship between culture, new media, and religion within everyday life.

The focus here will be on the exploration of the themes of authority, community, identity, and ritual in relation to specific religions. A number of questions will be considered, such as what are the implications of transporting the spiritual into technological practices; how is religion online reflected in larger societal trends toward lived religion and; how are sources/authorities informing online–offline religious practices in relation to Christianity, Islam, Hinduism, and Buddhism.

Authority

Recent studies have given much attention to the subject of authority and authoritativeness. Scholars have explored how sources and authorities inform online–offline religious practices, as well as highlighting whether we should be concerned that new media could potentially effect religious policy and theology and therefore, challenge authority and authoritativeness (Bunt, 2009; Cheong et al., 2012; Campbell, 2010). Examples of this include Israeli Haredi Jewish communities who have implemented filters and degrees of control on the Internet usage so that it does not violate the moral boundaries of their tradition. Further examples of how religious communities are negotiating the Internet can be found in the Muslim online communities and spaces such as the virtual Great Mosque of Mecca (Ka'ba Shrine) in *OnIslam* in *Second Life* (slurl.com/secondlife/OnIslam/10/54/21). Here, offline cultural behaviors and attitudes are transferred online. For instance, in 2007 when the virtual Mosque came under attack, *OnIslam* responded by putting a sign at the gates of the virtual hajj (pilgrimage site) detailing rules of acceptable behavior in

this virtual space, including no drinking, abusive language, or immodest dress by visiting avatars. In addition, there has been much discussion around the issue of online *fatwas*, which are religious interpretations of the scriptures (Bunt, 2009). The issue of who has the authority to translate the Qu'ran is considered offline but how can this be monitored in the online context? Online there are individuals providing *fatwas*, when offline, they would not be considered a source of authority or allowed to do this. Consequently, online *fatwas* can be perceived by some as less authoritative and therefore challenge 'authentic' Islam and authority. In contrast, the Muslim leader, Seik Al-Qaradawi, maintains that the Internet can 'foster religious identity' and so he has used the Internet to maintain his authority and identity both online and offline (Campbell, 2010: p. 35). Moreover, it has been argued that new media, such as blogs and forums can provide an opportunity for individuals to attain a position of power and authority and become influencers of religious hierarchies and power structures. In some instances, use of new media such as blogs can result in bloggers acting as "armchair theologians or cultural critics" (Campbell, 2012: p. 72).

This paradoxical view of how the Internet can be used to strengthen as well as challenge religious authority is explored by other scholars, such as Cheong, when she considers the paradox of how the Internet simultaneously supports and undermines traditional religious leaders (in Campbell, 2013). This is also highlighted by Baffelli et al. in the context of Japanese new religious movements' use of video sharing sites and interactive Web sites and how they are being used to communicate with followers and challenge other religious movements (2011). What is evident from these studies is the impact that new media is having on various individuals and communities, and it highlights that we need to consider who has the "true legitimate voice for a particular religious tradition or community" (Campbell, 2012: p. 76).

Community and Identity

In recent days, scholars have argued that it is important to understand trends within religion online in light of how this reflects larger societal trends toward lived religion, where every day lived religion is often more dynamic and fluid than religiosity prescribed in traditional contexts. This is clearly shown in how individuals create or negotiate religious community and identity online. Networked communities and associated identities developed online are often associated with the character of those rooted in offline community, such as virtual churches drawing their architecture or patterns of life from offline structures. *Second Life* (Anglican Church), blog communities, or on social media sites (e.g., YouTube, Facebook, Twitter). By studying networked religious communities we are able to see how "people in contemporary society live in religious social networks that are emergent, varying in depth, fluid, and highly personalized" (Campbell, 2012: p. 71) as they draw from multiple sources in the creation of online religious communities. Furthermore, the construction of identity may vary and allow for "dynamic opportunities for self-expression" (Campbell, 2012: p. 71). This is clearly evident as individuals may create and maintain multiple online personas on different digital media platforms for different motivations. A prominent example of the establishment of an online persona is the creation of an avatar in a virtual world. The avatar is a flexible identity that may engage in religious or secular activity and can change their appearance to reflect different identities, such as the wearing monastic robes or animal likenesses, thus demonstrating the ability of users to freely explore and construct unique religious identity presentations online. Another example of new media assisting self-expression is blogging. Blogs provide a platform in which religious imagery or language can be used to present and represent the belief structure of the author or enable them to self-reflect on different expressions of spirituality within the wider participatory culture of the blogosphere. Understanding the nuances and complexities of religion online and how new media can assist with creating and developing religious communities and identity can provide a clearer description of religion today.

If we look at this more closely in relation to Christianity it is possible to highlight examples of these trends within online religious communities and identities. Initially it was perceived that the use of the Internet by Christians might become a substitute for offline religious engagement, however, research has shown religious users involved in the formation of communities and online Christian identity to do this to supplement offline religious involvements. For example, blogging has become away for individuals to connect with like-minded communities where they can construct a shared religious identity via stories and behaviors. Also Christian communities have actively embraced new media to spread the Word and engage with other Christians from around the world, linking the online and offline context through integrated patterns of ministry and worship.

This is illustrated within the Anglican Cathedral in *Second Life*, a community of 500 members attending multiple weekly services in a virtual cathedral on Epiphany Island (Campbell, 2010: p. 126). The cathedral "was birthed through the vision of a small group of Episcopalians and Anglicans from the United Kingdom, New Zealand, and the United States, who met informally in Second Life" and who are potentially "reconfiguring religious practice online" (Campbell, 2010: p. 123). This online community holds services, which draw from both "traditional and contemporary styles of worship," in addition to taking into account the different time zones from around the world as well as designing the cathedral to replicate an offline cathedral, possibly to further legitimate the authenticity (2010: p. 123). One of the areas for concern is that "the online cathedral challenges traditional notions of authority and accountability" (2010: p. 124). These challenges are further highlighted, when in 2008, Mark Brown became a licensed minister for this virtual community, possibly providing the online community with "offline legitimacy" (Campbell, 2010: p. 126). The use of *Second Life* to expand the Anglican community is not without its challenges, although there are also opportunities, such as the potential for creating a virtual theology and establishing virtual Christian communities with a wide outreach, to members of the congregation across the world.

Ritual

When considering the implication of transporting the spiritual into technological spaces, a number of considerations arise,

such as the extent to which religious practice can be adapted to the Internet; the authenticity of a blended form of religion online; and how adaptations of religious behaviors may coalesce or undermine given belief systems. New media provide a way in which individuals can engage with and practice religion and in some instances transform ritual. However, unlike some aspects of traditional religion, new media enables the user easy access to the religion and to personalize it to their own needs. This personalization can result in what is seen as a 'pick and mix' or individualistic style of religion, and consequently, behaviors and belief systems can become merged, altered, or completely changed.

New media ritual can come in many guises including, Buddhist meditation, Hindu puja (worship), pilgrimage (e.g., in OnIslam in *Second* Life), and prayer facilitated via smartphone apps. The study of ritual online has been undertaken by Helland (2007), who argues that ritual can be both a religious and a secular act and therefore rituals are meaning-making systems. His examination of early USENET groups aimed to determine why such groups form and what they can provide for many individuals who are "geographically dispersed" (2007). Early Web sites were limited to textual interaction or broadcasting live events (e.g., online events, festivals, or worship). However, with the introduction of new media, there is the potential for visual, audio, and virtual selves engaging with others from around the world and actively participating in religious rituals.

Helland also examines Hindu rituals such as puja online and argues that although participants "are not there in person, their Web presence is felt, and they are conducting an authentic religious activity and receiving the benefits of that activity" (2007). He also considers three questions: Is the Hindu ritual, traditionally found in the home, moving to the home computer? Is the person "undertak[ing] an authentic and significant religious activity?" and Is online ritual assisting the wider diaspora? (2007). Answers to such questions may vary from religion to religion but for many, the ritual via new media is authentic and part of their everyday lives. This is further illustrated by Wagner who explores the use of smartphone mobile apps and how they can provide a means of engaging with religious ritual whenever the user wishes. She argues "that we can learn something about our streaming selves by looking at how we use, display, and interact with religion via our mobile devices" (2012: p. 105).

The authenticity of online ritual as well as the negotiation between offline–online spaces is acutely seen in *Second Life*. An avatar can prostrate on entering the virtual Buddhist temple, present an offering (donation) of Linden dollars (*Second Life* currency), and spin a virtual prayer wheel, thus suggesting an authentic experience by imitating offline practices. This raises questions as to whether or not a Buddhist virtual ritual can result in the same meritorious acts found offline. For some of the creators of Buddhist virtual objects and places, they argue that as long as the ritual is done with the same intention found offline, then it can result in merit for the online participant (Connelly in Campbell, 2013). It is clear that how people engage with and practice religion can change the dynamics of traditional religion (Campbell, 2012: pp. 76–77), but it can also provide us with an insight into how religious cultures are changing and how individuals are using new media for religious rituals.

The Emergence of the Study of Religion and the Internet

The academic study of religion and the Internet can be traced back to the mid-1990s when early works documented the emergence of religion online and specific practices such as Stephen O'Leary's (1996) observance of the Internet rituals enabling the re-imagination of religious sensibilities online. These early works explored a number of significant questions, such as 'Who is using the Internet?,' 'What are they using it for?,' and 'How are they using it to access religion or religious information?' Scholars, such as Dawson, noted the majority of research in the early stages focused on "identification and measurement" (Dawson in Cowan and Hadden, 2000: p. 26) and found a number of common research areas including the study of 'the nature and quality of people's experiences' in terms of religion, attempts to understand the negotiation of religious practice offline and online; identifying the implications of technology on religion; and considering whether one type of religion is more suited to the Internet than others (2004: pp. 10–11).

These research areas were explored in detail in three important edited collections. *Religion on the Internet: Research Prospects and Promises* (Cowan and Hadden, 2000) offered a comparative approach gathering key case studies within the initial wave of Internet scholarship, offering primarily descriptive research often contextualized within a utopian or dystopian Internet paradigm. Other important descriptive studies include Cowan's comprehensive account of paganism online in *Cyberhenge: Modern Pagans on the Internet* (Cowan, 2005), where he illustrates how Pagans use the Internet for the formation of cybercovens and the impact of online forums and informational Web sites on pagan practices. This documentary work highlights important research themes of the nature of identity, community, and authority online. Cowan explains that online pagans challenge previous notions of who or what is considered authoritative or authentic within pagan traditions. Thus initial, or what is called first wave research, sought to define and document the rise of different features of religion online.

The second collection is *Religion Online: Finding Faith on the Internet* (Dawson and Cowan, 2004), which marked a move toward more detailed analysis and refined methodological approaches. Questions were raised about particulars of studying religion online, such as what were the common forms of religious Internet practice and unique relationship between religion online and offline. Within this second wave there was a move toward categorization as an important study such as Helland's typology of religious Web sites as either 'religion-online' (informational) or 'online-religion' (participatory) (2000). There is a general consensus that there is no longer such a clear demarcation between the two. Yet it served an important purpose providing scholars with critical categories for evaluation of religious presentations online.

In the third collection *Religion and Cyberspace* (Hojsgaard and Warburg, 2005), the concept of three waves of research was introduced and later refined as "the descriptive, the categorical, and the theoretical" trends in scholarship (Campbell, 2013). Hojsgaard and Warburg used the metaphor of the third wave to highlight the theoretical and methodological turn

in scholarship of religion on the Internet. Here scholars have tried to consider what studies of religion and the Internet demonstrate about the nature of social relations and the practices of religious communities within wider cultural spheres. We also see movement toward providing theoretical interpretive frameworks, such as the religious social-shaping of technology (Campbell, 2010) to explain religious decision-making regarding online and technology practices.

The study of religion online continues to be given attention by different fields of scholarship and organizations such as the Pew Internet and American Life Project (http://www.pewinternet.org/). They published two key reports on this area including *Cyberfaith: How Americans Pursue Religion Online* report (Larsen, 2001) and *Faith Online*, which found that 64% of Americans have used the Internet for religious purposes (Hoover et al., 2005). This growing interest has also resulted in a number of dedicated journals or special issues such as *Religion* (2002), *Journal of Computer-Mediated Communication* (2007), and *Information, Communication and Society* (2011). In 2005, the Heidelberg *Journal of Religions on the Internet* (http://online.uni-hd.de/) was founded to consider issues of methodology, aesthetics online, and religious rituals on the Internet. For example, in 2010, the special issue focused on aesthetics online and articles in this issue examined the online Jewish prayer wall, Buddhist temples in *Second Life*, online churches, bloggers, and Hindu worship (*puja*). This dedicated focus shows that the study of religion and new media is increasingly recognized as a valid and vibrant subfield of Internet studies within the new field of media, religion, and culture studies.

Current Approaches to the Study of Religion on the Internet

This section presents the current approaches to the study of religion, highlighting the key questions while referring to the four themes: authority, community, identity, and ritual. A number of questions are raised and areas considered include, how new media are being embedded in everyday (religious) life; the concerns or challenges arising from using new media; and the unpacking of the online/offline context (Campbell and Lovheim, 2011; Wagner, 2012). Campbell and Lovheim claim that "after a decade and a half of published research on religion online, we are finally able to begin to make some educated claims about the impact of the Internet on religious culture and social forms" (2011: p. 1). The early waves of research primarily focused on descriptive accounts but recent research has begun to include more refined methodologies and typologies for interpreting and categorizing religion online, and offline – which some claim is the emergence of the fourth wave of research (Campbell and Lovheim, 2011).

Current approaches to the study of religion and new media include more comparative work across different traditions and cultural contexts, and a push toward more nuanced cultural considerations of dynamics and the relationship of religion online and offline. One area that has been given much attention is the online/offline demarcation of religious behavior and practice (Dawson and Cowan, 2004; Helland, 2000; Campbell and Lovheim, 2011; Campbell, 2012, 2013). There is now consensus that there is no longer a clear demarcation between offline and online and that new media are enabling an integration of the offline and online religious contexts, while potentially challenging traditional models of religion.

Furthermore, in *When Religion Meets New Media*, Campbell argues that there are two trends emerging, namely the growth in "new media alternatives" and the response from "conservative communities" and their use of new media (Campbell, 2010: pp. 191–193). Prominent examples of 'new media alternatives' are 'godcasting' (religious podcasting, www.godtube.com), Christian blogs (www.christianblog.com/), Islamic forums ("ask the Iman," www.islam.tc/ask-imam), and online Hindu worship known as *puja* (Rudra Centre, http://www.rudraksha-ratna.com/). The consideration of 'conservative communities' use of new media includes religious authorities engaging with new media in order to stay relevant; communities and authorities trying to take control of new media in order to 'support their position'; and offline communities evaluating the impact of new media on their tradition (Campbell, 2010: pp. 192–193). Examples of this include the Vatican adoption of a YouTube channel and Japanese new religious movements using video sites to engage with followers; yet at the same time seeking to maintain tight control over how the medium is used, such as restricting or disabling the comments feature (Campbell, 2010: p. 192; Baffelli et al., 2011). This clearly highlights the contention between traditional religion, authority, and the use of new media and that "authority is clearly challenged by emergent digital cultures" (Cheong et al., 2012: p. xi)

The relationship between cyberculture and religion has been considered by Campbell and Lovheim who emphasize that "religion has not only been impacted by, but also impacts" (2011: p. 1084) and they consider whether technology shapes religion or if religion shapes technology (2011: p. 1088). The themes of authority, community (or belonging), and religious practice can be explored and answers sought to how individuals, communities, and groups are "negotiating between the requirements of establishing a successful online presence and traditional criteria for legitimate knowledge, expertise or sincere commitment, often based in the conditions of offline contexts and sources" (Campbell and Lovheim, 2011: p. 1089). Moreover, studies of how new media are being used and the impact they are having and how religious culture may change are being given considerable attention (Campbell, 2013; Wagner, 2012).

This impact as well as religious and technological influence can be illustrated by looking at the user of mobile and smartphones. With the significant increase in the user of smartphones, we can consider the use of the religious app. In the context of identity, Wagner examines the use of smartphone apps to engage with religion and create religious identities. She presents six types of apps, which are (1) prayer apps, (2) ritual apps, (3) sacred text apps, (4) social media apps, (5) self-expression apps, and (6) focusing/meditation apps (2012: pp. 102–105). She argues that the religious app enables a more individualistic style of religion and the collection of apps on the smartphone relates to our own unique religious identity. Furthermore, she maintains that the ritual via the app is not facilitated by 'recognized religious authorities' but by anyone who owns a smartphone (2012: p. 105). This form of new media differentiates itself from other forms,

where the intention is to build online communities or be an extension of an offline community. Instead, the mobile app provides access to religious prayers, scriptures, rituals, and information as and when the user wants to engage with their religious practice.

Future Areas of Research

As new media emerges, changes, and grows in popularity, there is scope for further research and the refinement of existing approaches to the study of religion and new media. There are a number of areas, which need to be considered within the wider social and behavioral context and some of which align with Dawson's initial research concerns (2004: pp. 10–11). This would include, but is not limited to, the impact of those who have Internet access and those who do not and whether there is a growing digital divide between different religions or individuals use of new media to engage with religion; how new media are being used to assist with marketing and selling of religion, for example, the use of social networking sites to advertise nontraditional as well as traditional forms of religion; unauthorized surveillance of religious groups or restricting access to certain Internet sites (e.g., in China) (Campbell and Lovheim, 2011: p. 1093); the impact of changing technology on religion – good and bad – for example, the rise of the mobile apps to engage with religion 24/7 and which could be perceived as a more personal form of religion (Wagner, 2012); and understanding the manifestation of religious experience within the online context, such as engaging in virtual ritual in *Second Life* and this resulting in an authentic religious experience for the participant.

Another area which has been specifically highlighted for further investigation is Asian religions on the Internet, especially Buddhism (Campbell, 2010: pp. 190–191). Some work is emerging in this area and focusing on Buddhism in the online world of *Second Life* (see Grieve, 2010; Connelly in Campbell, 2013) and new religious movements (Baffelli et al., 2011). There are two other significant emerging areas, including the proposal of a subfield, 'digital religion' (Campbell, 2013) and the theoretical concept 'networked religion' (Campbell, 2012). In *Digital Religion: Understanding Religious Practice in New Media Worlds* (2013), Campbell argues that the subfield of 'digital religion' can be defined as the social and technological space, which is evoked or created when we consider how online and offline religious spheres merge and interconnect within digital culture. Thus digital religion can be used as a term to describe to the connection and extension of online religious practice and spaces into offline religious contexts (2013). An example of digital religion can be found in *Second Life*, where there is the simultaneous practice of Buddhist meditation offline and online. This manifests via the instruction of a Zen Buddhist priest online, however, there is the expectation that the avatar in *Second Life* is meditating offline, in accordance with the online instruction (Connelly in Campbell, 2013).

In addition to the proposal of the subfield there is also the introduction of the concept of 'networked religion.' This is used to describe how religious practice online parallels with social and religious practices offline, and represents the "social reality of life in a networked society" (Campbell, 2012: p. 66). Significantly, Campbell maintains that the naïve assumption that religion online ultimately and totally transforms religion offline is too simplistic, instead a nuanced exploration of the interrelationship of religion online and offline is required. Networked religion is characterized by five traits: convergent practice, multisite reality, storied identity, networked community, and shifting authority. "Together, these characteristics present an understanding of religion that is informed by a dynamic and integrated relationship between the sacred and the mundane online" (Campbell, 2012: p. 65). Understanding the complexity of the relationships, behaviors, and experiences offline can explain the choice of technologies and behaviors, experiences, and religious communities used online. This highlights how religion online shapes and reflects a nuanced view of religious practice in contemporary society.

The future direction and continued study of religion and new media is also suggested by the establishment of the *Network for New Media, Religion and Digital Culture Studies* in 2011 and its associated Web site launched in 2012 (http://digitalreligion.tamu.edu/) and the *Journal of Religion, Media and Digital Culture (JRMDC)* (*http://jrmdc.com/*) launched in 2012, both which seek to connect interdisciplinary scholars and promote the work of the study of religion and new media. The development of networks, journals, and conferences further emphasizes that the relationship between new media, religion, and culture is of significant interest if we are to understand the implications this may have on wider society and specifically religious authority, community, identity, and ritual.

Conclusion

What has become apparent in recent years is that there are a number of parallels between the offline and online religious culture, behaviors, and practices. This is clearly evident when discussed in relation to the four common themes of authority, community, identity, and ritual. The cultural contextualization for understanding how new media are being used, for what and by who, is similar to the approach for understanding the nuances of online religious culture, individuals, religious groups, and organizations who are embracing or rejecting new media as a medium for religious practice, information sharing, and communication. We have highlighted a number of issues, including the challenge to traditional religion and ideologies; the potential change from a communal to an individualistic form of religion, such as those engaged with via the religious app (Wagner, 2012); and the issue of wider social or political constraints placed on using different technology, for example, access to particular Web sites or social media in countries where some new media are banned (e.g., Facebook in China). What is not clear is whether new technologies and the constantly changing landscape of new media will radically alter religion.

By focusing on the relationship between religion and new media it enables a greater understanding of how to approach the examination of this topic, including the need to adapt existing methodologies or develop new ones; as well as highlighting what areas require further exploration. The study of media, religion, and culture has been framed within the three

and four wave context and in doing so this has helped to pin point key developments in the history of this field. We have also seen that fully understanding the relationship between new media and religion requires interdisciplinary working between a variety of fields and areas of inquiry (see Hojsgaard and Warburg, 2005: p. 9; Campbell and Lovheim, 2011: p. 1093)

The overview provided here illustrates that this topic can be discussed within the theoretical framework of 'networked religion' and positioned within the emerging subfield of 'digital religion.' Future research may address Dawson's research areas (2004: pp. 10–11), including more in-depth comparative approaches and longitudinal studies, in order to gain further insight into whether religion is being impacted by new media. This constantly changing landscape and the increasing introduction of new technologies provide new possibilities and potential challenges to religion, behavior, and culture, how we explore and negotiate these changes needs further consideration.

See also: Buddhism; Community and Media; Community and Society: History of the Concepts; Culture, Sociology of; Identity Offline and Online; Internet and Culture; Internet and Social Media: Anthropological Aspects; New Media and Social Capital; New Media and the Digital Divide; New Religious Movements; Pilgrimage; Public Sphere and the Media; Religions of East Asia; Ritual; Secular Religions; Social Media.

Bibliography

Baffelli, E., Reader, I., Staemmler, B., 2011. Japanese Religions on the Internet: Innovation, Representation, and Authority. Routledge, London.

Bunt, G., 2009. iMuslims: Rewiring the House of Islam. Hurst, London, UK.

Campbell, H., 2010. When Religion Meets New Media. Routledge, Oxon.

Campbell, H., March 2012. Relationship between religion online and offline. Journal of the American Academy of Religion 80 (1), 64–93.

Campbell, H., Lovheim, M., 2011. Rethinking the online-offline connection in religion online. Information, Communication & Society 18 (4), 1083–1096.

Campbell, H.A. (Ed.), 2013. Digital Religion. Understanding Religious Practice in New Media Worlds. Routledge, London.

Cheong, P.H., Fisher-Nieleen, P., Gelgren, S., Ess, C., 2012. Digital Religion, Social Media and Culture. Peter Lang Publishing, New York.

Cowan, D.E., 2005. Cyberhenge: Modern Pagans on the Internet. Routledge, New York; London.

Cowan, D.E., Hadden, J.K., 2000. Religion on the Internet: Research Prospects and Promises. JAI, New York.

Dawson, L.L., Cowan, D. (Eds.), 2004. Religion Online. Routledge, London.

Grieve, G., 2010. Virtually embodying the field: silent online meditation, immersion, and the cardean ethnographic method. Heidelberg Journal of Religions on the Internet 4 (1), 35–62.

Helland, C., 2000. Online religion/religion online and virtual communitas. In: Cowan, D.E., Hadden, J.K. (Eds.), Religion on the Internet: research prospects and promises, religion and the social order. JAI Press, New York, NY, pp. 205–223.

Helland, C., 2007. Diaspora on the electronic frontier: developing virtual connections with sacred homelands. Journal of Computer Mediated Communication 12 (3) (Special Issue on Religion and the Internet). Available from: http://jcmc.indiana.edu/vol12/issue3/ (accessed 12.06.12.).

Hojsgaard, M.T., Warburg, M. (Eds.), 2005. Religion and Cyberspace. Routledge, Oxon.

Hoover, S., Clark, L.S., Rainie, L., 2004. 64% of wired Americans have used the Internet for spiritual or religious information, Pew Internet and American Life Project. Faith online. http://www.pewinternet.org/reports/toc.asp?Report=119.

Hoover, S., Clark, L.S., Rainie, L., 2005. Faith Online. PEW Internet and American Life Project 2004. www.pewinternet.org/ (accessed 01.12.10.).

Larsen, E., 2001. Cyberfaith: How Americans Pursue Religion Online. PEW Internet and American Life Project 2001. Available from: www.pewinternet.org/ (accessed 01.12.10.).

O'Leary, S., 1996. Cyberspace as sacred space: communicating religion on computer networks. Journal of the American Academy of Religion 64, 781–808.

Wagner, R., 2012. Godwired. Religion, Rituals & Virtual Reality. Routledge, London.

Relevant Websites

http://www.beliefnet.com/ – Beliefnet.

http://cmrc.colorado.edu/ – Center for Media, Religion and Culture.

http://jrmdc.com/ – Journal of Religion, Media and Digital Culture.

http://digitalreligion.tamu.edu/ – Network for New Media, Religion, and Digital Culture Studies.

http://www.online.uni-hd.de/ – Online – Heidelberg Journal of Religions on the Internet.

www.patheos.com/ – Patheos-Hosting the Conversation on Faith.

Religion and Politics in the United States

Paul A Djupe, Denison University, Granville, OH, USA
Andrew R Lewis, University of Cincinnati, Cincinnati, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The character of religious engagement in U.S. politics is powerfully shaped by the First Amendment to the Constitution that established a religious free market. Competition encouraged a vibrant, inclusive, comprehensive organizational presence, in which religious organizations face internal disagreement as well as external value threats. Neither is stable, which means religious organizations continually battle a two-front war, and will come in and out of politics. Because of this, we should expect a dynamic, variable presence of religion in American politics.

Although most advanced industrial democracies have a public square denuded of religion, the United States is exceptional. The vast majority of Americans are religious by at least minimal measures, and substantial pluralities are religious in a more robust sense. Religious symbols are expected to be invoked by public officials, and they are. Yet, the connections between religion and politics are diverse, often contradictory, and undermine confidence about speaking uniformly of any particular effect religion has. This article examines these varied connections to understand religion's connection to democracy in America.

Definitions

Religion is a multilevel concept that gains shape through the often-federated organizational structure of recognized religions; the beliefs, values, and practices of individuals and social collectives; and the symbols (or cues) that people use in everyday communication. Religious actors can be average citizens, clergy, organizational elites, and office holders. Politics is individual and collective commitments that bear on government decision making; it is not just the exercise of power in everyday life, but rather it concerns the opinions people have as well as the behaviors people and organizations engage in to affect their interests through government. Politics in the United States often aspires to be democratic in the sense that it supports the maintenance of a limited state, but it also functions under a set of rules establishing the equality of participants and a near-universal ability of citizens to participate.

Democratic Roles for Religion

If the central democratic problem is how to resolve difference peaceably and not fall prey to the violence of faction, then scholars have rightly devoted their time to resolving it. The primary author of the U.S. Constitution and future president James Madison advocated a pluralist solution in Federalist 10 – the 85 *Federalist Papers* were written for New York newspapers from 1787 to 1789 to encourage states to ratify the drafted Constitution. Madison's solution was to embrace difference and rely on the checks and balances that competing interests will place on each other to achieve stability. In essence, diverse nations will aggregate diverse representatives who will not suffer proposals that cut against their interests. Religious diversity is important in this scheme because many see their aims as nonnegotiable, thus their vigilance will remain high to identify efforts that deviate from intractable religious scripts.

Tocqueville documented a considerably different phenomenon. Instead of a nation cleaved by difference, Tocqueville observed a nation united in good citizenship – personal responsibility was prized and local associations of all kinds, religious and not, were formed to resolve collective action problems. Importantly, religion undergirded this system, which enabled government to innovate on the security that the citizenry was essentially self-sufficient and without the strife that overt religious participation in politics had brought to European nations. If a pluralistic religious society meant stability through disagreement and discord to Madison, for Tocqueville, it meant precisely the opposite.

These two solutions also found their locus in vastly different mechanisms. Madison's approach vested institutions with the power to resolve democracy's contradictions; Tocqueville, on the other hand, clearly saw the solution rooted in value systems and social control outside of politics (see also Durkheim, 1965/1915). These divergent perspectives eventually coalesced into a deliberative theory of democracy (Gutmann and Thompson, 1996), which provides institutional elements that consciously aggregate different views among people committed to working together for a common solution. It is easy to imagine representatives of different religious groups sitting at the negotiating table, but there are features of the U.S. system that promote exposure to and engagement with difference within religious organizations as well. Although there are other approaches to the relationship between religion and democracy, including militant separation and a gentler sense of religion as a refuge (Fowler, 1989), these three perspectives (Madison's pluralism, Tocqueville's communitarianism, and deliberative democracy) encapsulate the style and content of the political *engagement* of religious interests.

It is remarkable that, by at least one measure, the United States has been a Christian nation only relatively recently. Church adherence was quite low at the founding (about 17%) and steadily increased to a plateau for the past 40 years around

62% (Finke and Stark, 2005: 23). With the ratification of the Bill of Rights in 1789, the U.S. Constitution established structural boundaries for the religious marketplace in America, at least on the federal level. With the First Amendment's religion clauses, the Constitution prevented the U.S. Congress from making laws 'respecting an establishment of religion' or 'prohibiting the free exercise thereof.' These two requirements, in addition to the prohibition of religious tests in Article VI of the Constitution, created a federal order much different than the European models and established a religious free market. Christian religions flourished in many forms in this unregulated space (and other religions did later). To effectively compete, religious entrepreneurs appealed to a particular community's needs and developed comprehensive and diverse sets of social, spiritual, and service programs to help meet them. Even if successful religious groups maintained high boundaries with the world, the drive to grow congregations dictated an essentially inclusivist orientation. Together, this placed religion as central organizations in people's lives, imbued many aspects of life for adherents with religious symbols, and gave adherents organizational experience, all of which have political consequences.

Constitutional Bounds

The market is not entirely free and the particular boundaries of it have shifted over time. Because of the federalist relationship between the national and the state governments, the constitutional structure protected state establishments of religion and state oaths of office that included religious items, some of which lasted into the nineteenth century (Curry, 1987; Dreisbach, 1996). Since the Constitution's religion clauses limited their application to the federal government, the U.S. Supreme Court did not adjudicate religious matters until the twentieth century, with a few exceptions. Per the constitutional order, there was no threat of a national church, although a generalized version of Protestant Christianity dominated American culture and politics for the first 150 years (Hamburger, 2002). By the mid-twentieth century (see *Cantwell v. Connecticut*, 1940; *Everson v. Board of Ed.*, 1947), the Court had applied its incorporation doctrine to the religion clauses. This applied the religion clauses to the states and initiated more than 70 years of intense debate regarding the prudential role of religion in public life.

The debate centers on competing views of religion's role in society that roughly parallel the tension between Madison's and Tocqueville's views. In free exercise case law, the Court must decide whether to permit practices that stray outside the mainstream, including anti-Catholic speech (*Cantwell*, 1940), keeping children from public schools (*Yoder*, 1972), peyote ingestion (*Smith*, 1990), and animal sacrifice (*Lukumi Babalu*, 1993). In establishment cases, the Court considers the opposite – state support for the majority. The dividing line lies at the intersection of pluralism, in which diverse perspectives keep society in check, and communitarian impulses that generate concern for maintaining a core to American culture.

In the modern period, free exercise decisions were weighted in favor of pluralism through the use of the 'compelling state interest' test, which places a heavy burden on state action that restricts religious practice. State support for religion was not given the same scrutiny. The Court promoted a strong antiestablishment perspective and initiated a 'wall between church and state...[that] must be kept high and impregnable' (see *Everson*, 1947; see also *McCollum*, 1948). At times, the wall did not look impermeable. The Court frequently allowed indirect aid to religion (e.g., *Everson*, 1947; *Walz*, 1970), while typically opposing direct aid or preference for religion (e.g., *Engle v. Vitale*, 1962; *Nyquist*, 1973). The wall that separated church and state reached its apex with *Lemon v. Kurtzman* (1971).

In the past 25 years, there have been revisions to both doctrines, effectively weakening both the free exercise and the establishment clauses. In *Oregon v. Smith* (1990), the Court dramatically changed the direction of free exercise law by refusing to carve out exceptions to a neutral law, with the exception of the right to hire and fire on the basis of religious belief (*Tabor v. EEOC*, 2012).

Free exercise rights were limited by a single case, but the limiting of the separation of church and state was a slower process as the Court steadily built precedent sufficient to allow state support for everyday religious organizations involved with secular programing, especially education. The shift sidestepped the establishment clause, arguing that the freedoms of speech and association prevent government institutions from discriminating against religious groups. Doing so has gained religious entities equal access to public facilities and public funds for students to attend private religious schools, among other things, that led one prominent legal scholar to declare that 'the constitutional era where separation is the dominant theme appears to be over' (Lupu, 1991, p. 557).

Religious Institutions and Politics

If Tocqueville was ever right that organized religion was content to remain outside of politics (see Hatch, 1989), the observation did not stand for long. Religious interests have been primary advocates in not just constitutional battles, but in a very broad set of political debates from issues as mundane as zoning, to self-interested concern for public support for social services, to cultural issues like same-sex marriage (see Djupe and Olson, 2007). Baptists and other religious minorities encouraged federal and state governments to formally disestablish religion in the founding period (McLoughlin, 1971). Jehovah's Witnesses and Mormons brought some of the early suits to extend free exercise rights, and most major religious organizations continue to work together to promote broad religious freedoms. Jewish organizations, along with Baptist and Mainline Protestant denominations, have been active in promoting the separation of church and state, whereas Catholics and evangelicals have championed the non-preferential and accommodationist interpretations of the establishment clause (Teles, 2007; Jelen, 2000).

The interests of religious organizations are much more extensive than the First Amendment and precisely because of it. That is, houses of worship and denominations developed comprehensive social service programs, which naturally invested them in governmental development of the welfare state. In the twenty-first century, very many groups maintain permanent representation before government, though often tempering

their advocacy because of internal disagreement among members of faith groups (Hertzke, 1988; Hofrenning, 1995). Moreover, members often know little about the work of denominational lobbies, even as they are careful to populate their agenda with broadly supported issues.

Clergy might also be thought to play a representational role, although there is considerable variance in whether they represent the faith or the faithful. When their congregations are underrepresented in the community, they tend to take a more active political role encouraging political engagement (Djupe and Gilbert, 2003). There is still debate about whether congregational clergy act in prophetic ways to challenge their memberships or whether they hew their agenda closely to safe positions. Clergy typically are able to find reference groups supportive of their positions (Calfano, 2009), while they are also sensitive to maintain member support, which is largely driven by a satisfactory experience with congregational programing. Studies that attempt to unravel the interplay of these reference systems are particularly needed.

Given that congregations face competitive pressures, they cast a wide net for new members and often aggregate a wide range of people (Wald et al., 1988). Thus, across denominations, clergy are by default inclusive and respectful of individual differences (to varying degrees, of course). Clergy typically face a significant degree of diversity in their congregations and often employ democratic norms (equality, mutual respect, and full participation) to govern congregational groups, even if they do not share that commitment privately or in their public speech. While many clergy, value democracy in its own right, it is easy to find examples of clergy promoting intolerance of a wide range of groups. The conditions under which clergy become 'republican schoolmasters' (Franklin and Liane, 1989) are deserving of further study.

Religion and Political Movements

In certain periods, religion has helped spark political movements in America, not merely political advocacy. These movements are defined by 'the mobilization of previously unorganized or non-political challengers' (McAdam, 1999: xxv), whereas ordinary political advocacy occurs through established lobbying organizations and networks. In the history of American democracy, religion has played an important role in many social and political movements. In the modern era, nowhere has this been more evident than the civil rights and the Christian Right movements.

To be successful, movements often appropriate existing organizational structures, and religious organizations, as the most widespread social organizations in America provide effective grassroots networks (Wald and Calhoun-Brown, 2011). Churches and churchgoers are not, by their nature, typically engaged in political advocacy. At certain times, however, some have become central to political movements. The civil rights movement was rooted predominantly in local African-American churches, which previously has been a conservative force in the black community and opposed to political activity (Morris, 1984). Evangelicals, the prominent religious group in the Christian Right, were also initially reluctant to engage in politics. Their beliefs about a supposed cultural and moral decline, including expanded abortion rights, acceptance of nontraditional sexual practices, and the decreased role for public religion spawned mobilization. These culture war issues brought together committed, orthodox faith communities (Hunter, 1991) – spurred by prominent pastors, entrepreneurial laypeople, local church networks, and grassroots organizations (Green, 1997) – and made a central connection to the Republican Party. Cultural politics can provide an important opportunity for the minority party, and Republicans utilized the discontent among religious conservatives to their advantage (Leege et al., 2002).

Religious political movements have had an important impact on American public policy and the political landscape. Religious activists in the civil rights movement achieved legal, political, and cultural change. The Christian Right movement has kept their issues on the political agenda, although there was limited success in changing state or national policy (Green et al., 2003) until recently. The Christian Right arguably has been most effective in securing a 'seat at the table' for conservative Christians (Wilcox and Larson, 2006), especially within the Republican Party, expanding the rights of the religious and mobilizing in support of ballot initiatives.

Like all social movements, religious political movements face difficulties in formation and maintenance. The civil rights movement waned when windows of opportunities closed, organizations fractured, and resources decreased (McAdam, 1999). The Christian Right has been plagued by internal conflict, lack of cooperation between groups, difficulty in mobilization, an intense countermobilization by opponents (Green, 1997), and steady integration into regular, Republican politics. The movement's diversity and complexity have inhibited scholarly assessments of Christian Right mobilization. Some suggest that psychological attachment and collective identity drive mobilization (Wilcox et al., 1993). Others have argued that Christian Right mobilization transcends denominational identities and conforms to a broader culture wars perspective (Wuthnow, 1988).

Because the key factor of a movement is organizing the previously unorganized, movement participation provides political skills and experiences, expands social networks, and engages people with political organizations. In this way, both the civil rights movement and the Christian Right movement have expanded the bounds of our representative democracy (but see Claassen and Povtak, 2010). At the same time, religious movements have a tendency to inflame conflict. Cultural politics, a key focus of the Christian Right, may be detrimental to democracy by promoting polarization (Layman, 2001) and restricting compromise and other democratic norms (Wilcox and Larson, 2006). This culture clash is largely absent from the mass public (Fiorina et al., 2006), but its presence in elite and niche politics may counteract some of the benefits of enhanced political participation that religious movements provide. Continued research comparing the positive and negative influences of religion on democratic participation is important.

Religion and Political Behavior

While religion has been tied to political movements, it is necessary to understand how this relationship occurs. Religious

influences on citizen political behavior are varied, and sociological and psychological theories contend to explain the connections between religion and political behavior. Religion often provides local organization in which people learn about the political choices of fellow citizens and of religious leaders (Djupe and Gilbert, 2009). Religion is a set of beliefs and values that individuals can use to cognitively structure their political choices (Barker and Carman, 2000). Religion also can be an identity that allows individuals to parse the world into ingroups and outgroups (Jelen, 1993), which helps establish trust in an information source. Religion also can be a set of cues that elites use to communicate ingroup solidarity and political content to potential followers (Leege et al., 2002). There is no reason to assume that these religious mechanisms combine to produce consistent attitudes and behaviors, but examining the conditions when they do is an important question.

The study of religious influence on environmental attitudes offers an ideal vantage point to observe these three perspectives at work. The earliest work by Lynn White argues that Christians believe that humans have dominion over the Earth and therefore are more indifferent to environmental protection. Subsequent research has refined and qualified this finding, with substantial disagreement depending on the measures used. Opposition to environmentalism has been limited largely to conservative Protestants, a finding that supports many interpretations, including the identity-based treatment that environmentalists are not Christians and environmental initiatives place nature over humans (which earned environmentalism the label satanic from the late Rev. Falwell and others). Strikingly, religiosity has been found to have positive effects on environmentalism especially when the specific belief that believers are stewards of the Earth is examined (Sherkat and Ellison, 2007). In the late 1990s, opposition to environmental measures was correlated with biblical literalism, as a proxy for a belief in human dominion over nature and as a belief in the immanence of the end times (Guth et al., 1995), but the effect faded to insignificance once more salient and particular beliefs were considered.

Environmental attitudes have changed considerably over the past 40 years in all sectors of society, including among religious traditions at the same rate. Since religious beliefs and practices have not changed much within religious traditions, something else must explain the shifts. Complicating the relationship, we know that religious beliefs relevant to shaping environmental attitudes come from a house of worship, which also conveys political opinion content through sermons and other sources directly to members. This raises the specter that religious belief effects are, in fact, spurious and dissipate once actual political communication on the environment is measured. Moreover, evangelicals are willing to become more environmentally friendly when cued by an identifiable ingroup elite (such as Rev. Rich Cizik, formerly of the National Association of Evangelicals). Taken together, religion exerts a variety of forces on opinion and also shows that connections between beliefs and opinions can be reinterpreted in the face of social (and elite) pressures. Few other opinions have been as thoroughly investigated as environmental support, but we suspect that similar, fundamental disagreements between theoretical perspectives would arise elsewhere.

There is less disagreement in the literature on religion and political activism. Most would agree in broad brush with Leege that "People (who) share a sacred community … develop the self-confidence to act in concert. Religion specifies what action to take, and religious beliefs (and religious institutions) create the obligation to act" (Leege and Kellstedt, 1993: 10). However, the precise mechanisms for each component and the degree to which participation is coordinated and aggregated by the congregation are less well known. And, curiously, little to no work has investigated religious coordination of the content of political activity.

Research acknowledges the importance of the organizational church, which offers opportunities to develop motivation, practice politically relevant skills, and be asked to take part. Studies locate organizational influences not in worship attendance but small group participation (Leege, 1988), where individuals meet face to face and enjoy leadership opportunities. Other literature, particularly in studies of the black church, places particular emphasis on religiosity as a mechanism that develops efficacy (Calhoun-Brown, 1996). Progress in this area will come from understanding why particular congregations develop a participatory infrastructure and asking whether resources from particular congregations translate to political activity at different rates. One promising avenue is to explore the link between participation and political communication in the congregation – initial work finds that civic skills practiced in nonpolitical churches have no relationship with political participation (Djupe and Gilbert, 2009: Chapter 5). Others should explore the long-asserted, little-tested links between congregations and interest groups, and community and congregation, especially if congregations help magnify parties' efforts to get out the vote.

Conclusion

Leege concluded the previous article in this volume with "Basic problems of conceptualization and measurement, underspecified models, endogeneity, failure to anchor puzzles in strong social theory, and inappropriate samples continue." Many of these problems persist after a decade but much abated. If religious influence is composed of the full range of social and psychological mechanisms and a multitude of actors, then existing studies of religion and political behavior are still highly fragmented, typically examining just one at a time. The unifocal character of most studies limits the development of covering theories and weakens our understandings of the sum and consistencies of religious influence. Two paths forward are recommended. One promising opportunity is greater use of experiments that allow for the controlled introduction of information from an actor in novel settings to help resolve the ambiguity of causal claims (see Djupe and Calfano, 2013). Equally promising are studies that combine data from across levels of analysis, such as congregations within denominations, clergy and congregations (Djupe and Gilbert, 2009), members and clergy within denominations (Lewis, 2011), clergy and communities (Olson, 2000), and individuals within states. Such studies naturally invite broader theorizing and should help religion and politics engage social theory of wider interest.

See also: Christianity in the English-Speaking West; Civil Religion; Evangelical, Revivalist, and Pentecostal Religions; Religion and Economic Life; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religiosity; Religious Organizations.

Bibliography

Barker, David C., Carman, Christopher J., 2000. The spirit of capitalism? Religious doctrine, values, and economic attitude constructs. Political Behavior 22 (1), 1–27.

Calfano, Brian R., 2009. Choosing constituent cues: reference group influence on clergy political speech. Social Science Quarterly 90 (1), 88–102.

Calhoun-Brown, Allison, 1996. African-American churches and political mobilization: the psychological impact of organizational resources. Journal of Politics 58, 935–953.

Claassen, Ryan, Povtak, Andrew, 2010. The Christian right thesis: explaining longitudinal change in participation among evangelical Christians. Journal of Politics 72, 2–15.

Curry, Thomas J, 1987. The First Freedoms: Church and State in America to the Passage of the First Amendment. Oxford University Press, New York.

Djupe, Paul A., Calfano, Brian R., 2013. God Talk: Experimenting with the Religious Causes of Public Opinion. Temple University Press, Philadelphia.

Djupe, Paul A., Gilbert, Christopher P., 2003. The Prophetic Pulpit: Clergy, Churches, and Communities in American Politics. Rowman & Littlefield, Lanham, MD.

Djupe, Paul A., Gilbert, Christopher P., 2009. The Political Influence of Churches. Cambridge University Press, New York.

Djupe, Paul A., Olson, Laura R, 2007. Religious Interests in Community Conflict: Beyond the Culture Wars. Baylor University Press, Waco, TX.

Dreisbach, Daniel L, 1996. The constitution's forgotten religion clause: reflections on the Article VI Religious Test Ban. Journal of Church and State 38 (2), 261–295.

Durkheim, Emile, 1965/1915. The Elementary Forms of Religious Life. Free Press, New York.

Finke, Roger, Stark, Rodney, 2005. The Churching of America 1776–2005: Winners and Losers in Our Religious Economy. Rutgers University Press, New Brunswick.

Fiorina, Morris P., Abrams, Samuel J., Pope, Jeremy, 2006. Culture War? The Myth of a Polarized America, second ed. Pearson Longman, New York.

Fowler, Robert B, 1989. Unconventional Partners: Religion and Liberal Culture in the United States. Eerdmans, Grand Rapids.

Franklin, Charles H., Kosaki, Liane C., 1989. Republican schoolmaster: the U.S. supreme court and public opinion. American Political Science Review 83 (3), 751–771.

Green, John C, 1997. The Christian right and the 1994 elections: an overview. In: Rozell, Mark J., Wilcox, Clyde (Eds.), God at the Grassroots, 1996: The Christian Right in the 1996 Elections. Rowman & Littlefield, Lanham, MD.

Green, John C., Rozell, Mark J., Wilcox, Clyde, 2003. The Christian Right in American Politics: Marching to the Millennium. Georgetown University Press, Washington, DC.

Guth, James L., Green, John C., Kellstedt, Lyman A., Smidt, Corwin E., 1995. Faith and the environment: religious beliefs and attitudes on environmental policy. American Journal of Political Science 39 (2), 364–382.

Gutmann, Amy, Thompson, Dennis, 1996. Democracy and Disagreement. Belknap Press, Cambridge.

Hamburger, Philip, 2002. Separation of Church and State. Harvard University Press, Cambridge, MA.

Hatch, Nathan O, 1989. The Democratization of American Christianity. Yale University Press, New Haven, CT.

Hertzke, Allen D, 1988. Representing God in Washington: The Role of Religious Lobbies. University of Tennessee Press, Knoxville, TN.

Hofrenning, Daniel J. B (Ed.), 1995. Washington But Not of It: The Prophetic Politics of Religious Lobbyists. Temple University Press, Philadelphia, PA.

Hunter, James D, 1991. Culture Wars: The Struggle to Define America. Basic Books, New York.

Jelen, Ted G, 1993. The political consequences of religious group attitudes. Journal of Politics 55 (1), 178–190.

Jelen, Ted G, 2000. To Serve God and Mammon: Church-State Relations in American Politics. Westview Press, Boulder, CO.

Layman, Geoffrey C, 2001. The Great Divide: Religious and Cultural Conflict in American Party Politics. Columbia University Press, New York.

Leege, David C, 1988. Catholics and the civic order: Parish participation, politics, and civic participation. The Review of Politics 50 (4), 704–737.

Leege, David C., Kellstedt, Lyman A., 1993. Rediscovering the Religious Factor in American Politics. M. E. Sharpe, Armonk, NY.

Leege, David C., Wald, Kenneth D., Krueger, Brian S., Mueller, Paul D., 2002. The Politics of Cultural Differences: Social Change and Voter Mobilization Strategies in the Post-New Deal Period. Princeton University Press, Princeton, NJ.

Lewis, Andrew R, 2011. The Southern Baptist Church-State 'Culture War:' The Internal Politics of Denominational Advocacy. Ph.D. diss., Political Science, American University, Washington, DC.

Lupu, Ira C, 1991. Reconstructing the establishment clause: the case against discretionary accommodation of religion. University of Pennsylvania Law Review 140, 555–612.

McAdam, Doug, 1999. Political Process and the Development of Black Insurgency, 1930–1970, second ed. University of Chicago Press, Chicago, IL.

McLoughlin, William, 1971. New England Dissent 1960–1833: The Baptists and Separation of Church and State. Harvard University Press, Cambridge, MA.

Morris, Aldon, 1984. The Origins of the Civil Rights Movement: Black Communities Organizing for Change. Free Press, New York.

Olson, Laura R, 2000. Filled with Spirit and Power: Protestant Clergy in Politics. State University of New York Press, New York.

Sherkat, Darren E., Ellison, Christopher G., 2007. Structuring the religion-environment connection: identifying religious influences on environmental concern and activism. Journal for the Scientific Study of Religion 46 (1), 71–85.

Teles, Steven M, 2007. The Rise of the Conservative Legal Movement: The Battle for Control of the Law. Princeton University Press, Princeton, NJ.

Wald, Kenneth D., Calhoun-Brown, Allison, 2011. Religion and Politics in the United States. Rowman & Littlefield, Lanham, MD.

Wald, Kenneth D., Owen, Dennis, Hill, Samuel, 1988. Churches as political communities. American Political Science Review 82, 531–548.

Wilcox, Clyde, Jelen, Ted G., Leege, David C., 1993. Religious group identifications: toward a cognitive theory of religious mobilization. In: Leege, David C, Kellstedt, Lyman A (Eds.), Rediscovering the Religious Factor. M.E. Sharpe, Armonk.

Wilcox, Clyde, Larson, Carin, 2006. Onward Christian Soldiers: The Religious Right in American Politics. Westview Press, Boulder, CO.

Wuthnow, Robert, 1988. The Restructuring of American Religion. Princeton University Press, Princeton, NJ.

Religion and Sciences

Gary B Ferngren, Oregon State University, Corvallis, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The intersection of religion and science has been both more complex and less conflictive than is frequently depicted by those who cite the trial of Galileo, the dispute over the theological implications of Charles Darwin's theory of evolution by natural selection, and the Scopes trial as evidence of the historical incompatibility of religious beliefs and science. This article suggests that a complexity thesis better accounts for the relationship than the more familiar conflict thesis. It surveys the history of the relations between religion and science in the Western tradition from the fifth century BC to the present, with special reference to Christianity.

The intersection of religion and sciences has been both more complex and less conflictive than is frequently depicted by those who cite the trial of Galileo, the dispute over the theological implications of Charles Darwin's theory of natural selection, and the Scopes trial as evidence of the historical incompatibility of religious beliefs and scientific endeavor. This article will first examine the conflict thesis, and then survey the history of the relations between religion and science in the Western tradition from the fifth century BC through the first decade of the twenty-first century with special reference to Christianity.

The Historiography of Religion and Sciences

The thesis that Christianity has been historically hostile to science was popularized by two influential works that appeared in the late nineteenth century: John William Draper's (1811–82) *History of the Conflict between Religion and Science* (1874) and Andrew Dickson White's (1832–1918) *A History of the Warfare of Science with Theology in Christendom* (1896). White argued that Christianity had a long history of opposing scientific progress in the interest of dogmatic theology. Wedding a triumphalist view of science to a dismissive view of theology, his thesis struck a responsive chord in American thought, which was at the turn of the twentieth century increasingly committed both to a secular outlook and to recognizing the central role that the sciences played in modern society. The Draper–White thesis amplified a positivist outlook, viewing science as continually progressing and overcoming the entrenched antagonism of religious opinions, which invariably retreated before its advance. This thesis became enormously influential in the United States both among academics and in the popular mind.

Some historians regarded the Draper–White thesis as oversimplifying and distorting a complicated relationship. Thus, Burtt (1932) suggested that, far from being hindered by religion, science was in fact indebted to theological ideas. Alfred North Whitehead (1926) rejected the view that in controversies between science and religion, "religion was wrong, and … science was always right. The true facts of the case are very much more complex, and refuse to be summarized in these simple terms." A number of important studies by Robert K. Merton, Herbert Butterfield, and Alexandre Koyré, among others, also challenged the conflict thesis. Some historians of science went even further to argue that specific components of Christian theology were responsible for the scientific revolution. Following Max Weber and Robert Merton, some scholars argued that Calvinism's accommodationist understanding of Scripture and its voluntaristic theology encouraged scientific endeavor, and others maintained that theism encourages a belief in a rational creator, which may precondition scientific reasoning (Jaki, 1974).

Several scholars have provided a systematic reevaluation of the conflict thesis, particularly Lindberg and Numbers (1986) and Brooke (1991). Lindberg and Numbers argued for a historical picture that recognized a complex and diverse interaction that defies reduction to simple 'conflict' or 'harmony' to replace the largely polemical one drawn by Draper and White. Brooke objected to imparting essentialist understandings to words like 'science' and 'religion,' which had many different meanings over time and could not be reduced to simple static concepts. 'Natural philosophy,' for example, denoted something very different in the seventeenth century from 'science' in the twentieth. The growing acknowledgment that in their relations, religion and the sciences have exhibited a multiplicity of attitudes, reflecting local conditions and particular historical circumstances, led Brooke to suggest that a 'complexity thesis' was a more accurate model than the familiar conflict thesis. The views of Brooke and of Lindberg and Numbers have gained increasing acceptance among many historians of science.

Greece and Rome

Western science as an explanatory enterprise began with the Greeks. It was in the sixth century BC that pre-Socratic philosophers like Thales (fl. 585 BC), Anaxagoras (*c.* 500–428 BC), and Anaximander (610–546 BC) challenged the mythopoeic picture of the natural world. The pre-Socratics understood the world in nonmythological terms, viewing it as the product of natural forces. In the late fifth century BC, they experienced some popular opposition in Athens. In *c.* 437–436 BC, Anaxagoras was prosecuted on a charge of impiety and his books publicly burned.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84030-1

By the end of the fifth century BC, natural philosophy (physical science) had come to be recognized by the Greeks as offering the most suitable means of understanding the world of nature, and religious issues played a limited role in either hindering or encouraging scientific endeavor in the classical world after that time. This does not mean that religious ideas were of no account in formulating worldviews for the Greeks or for the Romans, who took over the philosophical and scientific culture of the Greeks. The Epicureans explicitly repudiated a religious view of the universe, while for the Stoics, science (especially cosmology) and religious thought were intimately related. Moreover, the Greek view of the world was predominantly an organicist or vitalist one. Philosophical naturalism, which was espoused by the pre-Socratic philosophers, the sophists, Aristotle (384–22 BC), and the Epicureans, always remained a minority opinion in Greco-Roman thought. In general, the Greeks conceived of nature as a divinity. Plato (*c.* 427–347 BC) believed that a world soul (*psyche*) animated the universe, in which the sun, the moon, the stars, and the planets were gods. A vitalist world picture, stripped off its pagan elements, was transmitted by the Romans to Western Europeans, who continued to hold it well into the Middle Ages.

Late Antiquity and the Middle Ages

Christianity spread rapidly throughout the Roman Empire owing to its extensive missionary activity. In the second century AD, Christian apologists (theologians who defended their faith philosophically against pagan critics) began the process of harmonizing Christian revelation with Greco-Roman philosophy. Apologists like Justin Martyr (*c.* 100–65), Tertullian (*c.* 160–220), and Irenaeus (*c.* 130–200) held ambiguous positions on natural philosophy. While they often denounced it for its pagan religious or anti-Christian elements, they employed its methods, thought forms, and vocabulary – and drew on it when it was useful in the rational defense of Christianity. This hellenized Christianity by taking elements of classical culture and incorporating them into the Christian worldview. Tertullian, for example, who asked, "What indeed has Athens to do with Jerusalem? What harmony is there between the Academy and the church?" is sometimes cited as an example of early Christian antiintellectualism. Yet he often borrowed from Greek natural science and medicine to buttress his theological arguments. Not all Christian apologists accepted Greek science, whilst a few, influenced by ascetic ideals or concerns about potential demonic influences, were hostile in their approach to Greek pagan natural philosophy.

Augustine of Hippo (354–430) shared the reservations of some of the apologists about the value of natural philosophy, but he recognized its utility. Like other apologists, he rejected *curiositas*, the desire to pursue knowledge for the sake of knowing, which he regarded as a characteristic feature of pagan learning. But Augustine believed that the study of natural philosophy was legitimated when it served higher purposes, such as aiding in the understanding of Scripture or lending support to Christian theology. By becoming the handmaiden of Christian theology, natural philosophy was assured a place in the emerging Christian culture, one that was to last through the Middle Ages and well into the early modern period.

The early Middle Ages (*c.* 500–1000) was a period of disintegration that followed the collapse of the Roman Empire in the West under repeated attacks and occupation by Germanic tribes. As towns became villages, public services ceased to exist, and roads became impassable, the cultural continuity with the classical world was nearly severed. Educational institutions disappeared, and learning in the West was confined to Christian monasteries. A few scholars attempted to preserve natural philosophy during this period, including Cassiodorus (*c.* 485–580); Isidore, bishop of Seville (*c.* 560–636); the Venerable Bede (*c.* 673–735); and John Scotus Erigena (*c.* 810–77), but they did so as clerics, within a monastic framework, and with the confidence that scientific knowledge, as a handmaiden to theology (*philosophia ancilla theologiae*), was worthy of being transmitted to future generations.

The eleventh and twelfth centuries experienced the revival of trade between western Europe and the eastern Mediterranean, together with the concomitant renewal of urban life, the growth of manufacturing, and the beginning of the gradual transformation of Europe from a rural to an urban economy. Cathedral schools sprang up in cities and gradually replaced monastic schools. In the twelfth century, natural philosophers avidly studied Plato's *Timaeus* and attempted to harmonize its *cosmology* with the Creation account in Genesis. The Arabs had preserved Greek learning, and works by Greek authors on natural philosophy were translated into Latin from Arabic, thus making available much new scientific material. Muslim contributions (e.g., in optics) were significant as well.

There was much in Aristotle's writings that was useful to European Christians, but there were also areas of potential conflict, such as Aristotle's belief that the world is eternal, that the soul is not immortal, and that the universe can be accounted for by naturalistic and even deterministic explanations apart from divine providence. These and other incompatible doctrines caused the teaching of Aristotle's natural philosophy to be banned from the University of Paris in 1210 and 1215, but by 1255 his writings had gained official sanction, and they came to be widely studied. A number of distinguished scholars, including Roger Bacon (1213–91), Albert the Great (1200–80), and his student Thomas Aquinas (*c.* 1224–74), attempted to create a synthesis of Aristotelian philosophy and Christian theology that would reconcile the Christian faith and classical reason. The existence of a more radical faction at Paris, headed by *Siger of Brabant* (*c.* 1240–84), led the bishop of Paris, Etienne Tempier, to condemn, at the request of traditional theologians, many of the philosophical propositions espoused by Siger. The condemnations represented a conservative backlash that placed limits on speculation in natural philosophy. In the long run, however, they provided incentives to explore other scientific options that lay outside Aristotelian natural philosophy, particularly by way of thought experiments that entertained hypothetical possibilities, such as the existence of other worlds or of a vacuum.

The Copernican Revolution and Galileo

The geocentric theory of the solar system, as refined by Ptolemy of Alexandria (second century AD), was almost universally accepted in the West and taken to be compatible with their

theology by Jews, Muslims, and Christians until the seventeenth century. The first major challenge to the Ptolemaic theory came with the publication of Nicolaus Copernicus' (1473–1543) *De revolutionibus orbium coelestium* (*On the Revolutions of the Heavenly Bodies*), which appeared after the author's death in 1543. Copernicus advocated a heliocentric model of the solar system, and was encouraged to publish it by the Lutheran astronomer Georg Rheticus (1514–74); and while Lutherans at the *University of Wittenberg* initially rejected heliocentric cosmology, several, including Philip Melanchthon (1497–1560), were attracted to the mathematical convenience of the new theory. However, most natural philosophers rejected heliocentrism, and in 1600 there were probably no more than 10 Copernicans in all Europe.

One scientist who was quick to embrace the new theory was *Galileo Galilei* (1564–1642), a young professor of mathematics at the University of Padua. Although secretly a Copernican by 1597, Galileo did not publicize his views until 1613, when in *Letters on Sunspots* he declared his acceptance of the heliocentric system. It was, however, his *Letter to the Grand Duchess Christina* (1615), in which Galileo argued that Scripture did not intend to impart scientific truth but accommodated itself to popular language in describing the natural world, that provoked the Catholic Church's response. Although a layman, Galileo had undertaken to interpret the Bible. In the following year (1616), he appeared before the Congregation of the Holy Office (the Inquisition). The Church declared the theory 'false and contrary to Holy Scripture' and recommended that *De revolutionibus* be 'suspended until corrected.' Galileo was warned not to teach or defend heliocentrism. In 1632, Galileo published his *Dialogue on the Two Principal World Systems*, an ostensibly even-handed debate between proponents of the Ptolemaic and Copernican theories in which the arguments for the Ptolemaic theory were made to look feeble. Galileo was summoned to Rome, where he stood trial in 1633. Compelled to abjure his belief that the earth moved, he was confined to a country estate until his death in 1642. The strong reaction against Copernicanism (and Galileo's espousal of it) came from the Aristotelians of the scientific establishment, some of whom were determined to entangle Galileo with the Church. In particular, Galileo aroused the ire of a loosely organized movement of traditional scientists called the *Liga*, which succeeded in introducing theological issues into scientific discourse and attacked Galileo for his views. The Church was also spurred by the Protestant Reformation to prove its good faith in interpreting Scripture literally in the face of Protestant attacks on its allegorizing tendencies in biblical exegesis. Aristotelianism became, in its Thomistic form, a rigid dogma, and the papal court became, after the Council of Trent (1545–63), more authoritarian. The Inquisition and the Index of Prohibited Books, introduced into Italy by forces of the Counter-Reformation, were employed against Galileo.

The Mechanical Philosophy and Natural Theology

By the mid-seventeenth century, the Copernican theory had come to be widely accepted, and a mechanical model of the universe was formulated to account for all natural phenomena in terms of matter and motion without reference to action at a distance. The mechanical philosophy had its roots in corpuscular materialism, which explained physical phenomena by the interaction of tiny particles of matter called atoms. While elements of the model had been adopted by several of the leading figures of the scientific revolution, its most influential advocates were René Descartes (1596–50) and Pierre Gassendi (1592–55), who sought to devise a new philosophy that would replace Aristotelianism. While Descartes and Gassendi grounded their mechanical philosophies in theological considerations, critics feared that their views would lead to materialism, as they did in the case of Thomas Hobbes (1588–1679), or to deism, as they did in the case of Voltaire (1694–1778).

Mechanical philosophy received its greatest synthesis at the hands of Isaac Newton (1642–1727) in his *Principia Mathematica Naturalis Philosophiae* (*Mathematical Principles of Natural Philosophy* (1687)). By his discovery of the principle of universal gravitation, Newton succeeded in harmonizing celestial and terrestrial mechanics, bringing for the first time the whole of nature into a precise rational interpretation. Newton's picture of a universe governed by uniform laws that could be described without recourse to primary causes resulted in the widespread belief that nature was self-governing, and aided the 'natural religion' of deism and its idea of a 'watchmaker' God. On the basis of his synthesis of the mechanical philosophy, David Hume (1711–76) developed a materialistic philosophy in England, while in France philosophers like Julien Offray de La Mettrie (1709–51) and Baron d'Holbach (1723–89) embraced atheistic materialism.

The popularity of the mechanical philosophy led in the latter half of the seventeenth century to the creation of a natural theology that continued into the nineteenth century. Natural theology seeks to argue for the existence and attributes of God by means of knowledge derived from natural reason rather than from special revelation. There were several grounds for its attraction: the wonders of scientific discoveries, many of which could be seen for the first time with the invention of the microscope; the desire to secure an adequate foundation for a theology of divine providence that would confront the perceived dangers of materialistic atheism; and the enormous influence of Newton's mechanical picture of the world. The argument from design became the cornerstone of a natural theology, in which God was described as a craftsman, a geometer, or an architect. Natural theology reached its apex in William Paley's (1743–1805) influential *Natural Theology, or Evidences of the Existence and Attributes of the Deity* (1802).

Geology and Darwinian Evolution

The influence of the Enlightenment, with its exaltation of reason and rejection of special revelation, can be observed in the developments that were occurring in science, particularly in the increasing tendency to seek naturalistic explanations and to divorce science altogether from theological considerations. In the seventeenth and eighteenth centuries, theories of the earth had sought to weave together geology and theology in what was termed 'physico-theology.' Many of these theories calculated the age of the earth by using biblical chronologies in a way that could be harmonized with physical principles. Invariably, they postulated a young earth. While most

attributed organic fossils to the Noachian Flood, by the eighteenth century it was becoming increasingly difficult to explain the global distribution of fossils by a single event (the Deluge) that had taken place a few thousand years ago. The French naturalist *Georges Louis Leclerc, Comte de Buffon* (1707–88), a deist, proposed a much older earth, publicly of 75 000 years and privately perhaps as old as 3 million years, and suggested that each of the six biblical days of Creation might correspond to an immense length of time. Georges Cuvier (1769–1832) demonstrated the existence of fossil mammals unlike any in his own day, and argued that a series of geologic catastrophes had eliminated species in past ages. Cuvier's catastrophes were not global and in his case did not derive from his Christian convictions. But they were later read as such and were appropriated in English natural theology to support catastrophism, which became a popular form of physico-theology that attributed geological changes to great catastrophes of the past, the most recent being the biblical Flood.

The Scottish geologist Charles Lyell (1797–1875) rejected providential interventions in geologic history. In his *Principles of Geology* (3 vols., 1830–33), he succeeded in severing geology from theology. Lyell rejected diluvialism (catastrophism) for what became known as uniformitarianism: the theory that geological changes on the earth's surface have occurred slowly over long periods of time and that in explaining change we refer to processes that we currently observe. Previous proponents of *diluvialism*, like William Buckland (1784–1856) and Adam Sedgwick (1785–1873), quickly abandoned catastrophism and adopted the new geology, which Christians harmonized with the Genesis account of Creation in a relatively painless fashion. A group of laymen who called themselves Mosaic or Scriptural geologists continued to take their geology from Genesis, but they were marginalized since nearly all professional geologists abandoned the Noachian Flood as an organizing idea in accounting for changes in the earth's history.

Charles Darwin (1809–82) had, as a young man, been strongly influenced by Paley's natural theology. During his voyage aboard the H.M.S. *Beagle*, however, he read Lyell's work and began to think that slow changes had produced the geological phenomena that he observed. His theory of evolution by natural selection, published in *Origin of Species* (1859), made use of Lyell's theories that the earth was very old and the fossil record is imperfect. Evolutionary ideas had preceded Darwin. Originating in the Enlightenment, they had already been proposed in the eighteenth century by Charles's grandfather, Erasmus Darwin (1731–1802), and by Jean-Baptiste Lamarck (1744–1829), the latter proposing the theory that acquired characteristics were inherited. By the mid-nineteenth century, the belief that transmutationism (evolution) represented a divine plan was widespread. In advancing natural selection as a mechanism for evolution, Darwin argued that in the 'struggle for existence,' those individuals with even a slight advantage would survive and pass on their advantage to their offspring. Herbert Spencer (1820–1903) referred to it as the principle of the 'survival of the fittest.'

Darwin's theory met with a mixed response that cut across conventional professional and religious boundaries. Some scientists initially rejected it, while a number of clergymen accepted it. Conservative anti-Darwinists, including several leading naturalists like Louis Agassiz of Harvard (1807–73), geologists like the Canadian William Dawson (1820–99) of McGill, and theologians like Charles Hodge (1797–1878) of Princeton Seminary, rejected it, fearing the break with final causes. A second group of Christians accepted evolutionary theory as compatible with a biblical faith. They included the Scot James McCosh (1811–94), president of Princeton University, and the American botanist Asa Gray (1810–88) of Harvard. A third group came to believe that evolution had rendered supernatural religion untenable and opted (most often) for some form of philosophical materialism.

By the 1870s, however, most scientists and a large body of the educated public had accepted some form of evolution (though not necessarily by natural selection). Full acceptance of Darwin's views on human nature, as set forth in his *Descent of Man* (1871), was more difficult for Christians. Darwin maintained that moral values were rationalizations of social instincts that had developed by natural selection, thus asserting a full-blown materialism. Alfred Russel Wallace (1823–1913), the co-discoverer of the theory of evolution by natural selection, rejected materialism, maintaining that the creation of the human mind required nonmaterial forces.

In the 1920s, evolution was one of several theological issues that divided the major Protestant denominations of North America in the modernist–fundamentalist controversy. While many Protestants (as well as Roman Catholics and Jews) had come to believe in an old earth and to harmonize belief in evolution with a nonliteral interpretation of the Genesis account of Creation, conservatives emphasized in their rejection the atheistic and materialistic dangers inherent in a naturalistic theory of human origins. But the Scopes trial , held in Dayton, TN in 1925, symbolized the inability of biblical literalists to withstand the force of Darwinism, which had come to be taken for granted by the intellectual elites of Europe and America. Antievolutionism was pilloried in the American press as a rear-guard reaction to scientific progress, and it was associated with nativism and a rural and uneducated religious fundamentalism. Although biblical catastrophism was revived in the 1920s, it enjoyed favor only in conservative and sectarian Protestant circles in North America before it exploded into a popular creationist movement (called 'flood geology') in the 1960s.

Twentieth-Century Physics and Cosmology

From about 1895, modern physics began to deal with issues that went beyond those of classical physics, including quantum mechanics and special and general relativity. The introduction of quantum mechanics challenged the determinism of the mechanical picture of Newtonian physics. The central features of quantum theory were developed separately in 1926 by Werner Heisenberg (1901–76) and Erwin Schroedinger (1887–1961). The indeterminacy principle of Werner Heisenberg demonstrated that one could not know perfectly both the position and momentum of a subatomic particle. Heisenberg's principle gave rise to much discussion of its wider philosophical and theological implications. Theologians such as John Polkinghorne have argued that the

discovery that physical processes, both at the microscopic level of quantum theory and at the macroscopic level of chaos theory, are unpredictable brought about the demise of a mechanical view of the universe. Theologians argued that this offered space for God's immanence to operate through natural law in providential action, in place of the deistic belief in a God who is not involved in superintending its creation (Polkinghorne, 2005: x–xi).

Discoveries were also made in cosmology that caught the interest of theologians. Edwin Hubble's (1889–1953) discovery that the universe was expanding at a tremendous rate led the Belgian Catholic priest and physicist *Georges Lemaître* (1894–1966) to propose that the universe had begun as a dense 'primeval atom' that had exploded in the distant past. The English astrophysicist Fred Hoyle (1915–2001) dismissively termed it the 'big-bang' theory. Hoyle challenged the theory by offering an alternative steady-state theory, according to which the universe is eternal but matter is being continuously created in space. In the late 1960s, several observational discoveries led to the almost universal acceptance of the big-bang theory, which some scientists with religious convictions have argued supports the notion that the universe was created by God.

Several cosmological theories have developed to attempt to counter materialistic explanations for the origins of the universe and of humans. Among them are the anthropic principle and intelligent design (ID). The anthropic principle asserts that the emergence of human beings in the universe is not an accidental by-product of purposeless cosmic evolution. According to this theory, the universe was so finely tuned during its earliest moments that if any one of several physical parameters had been slightly different it would have been unsuitable for the creation and survival of human life. Hence, it is argued that the inhabited universe is anthropocentric, having been designed for human beings. The anthropic principle reintroduced into cosmological speculation issues that scientists had laid to rest, including questions of teleology and divine providence. Much more controversial was the ID movement that arose in the 1990s. It extended the argument from design to biology, claiming that intelligent causes are required to account for complex structures in biology and that they can be empirically detected.

Eugenics and Biotechnology

The term 'eugenics' was coined by the English scientist Francis Galton (1822–1911), cousin of Charles Darwin. The eugenics movement sought to improve the human race by eliminating mental illness, retardation, and criminal behavior through such means as compulsory sterilization and restrictions on marriage. In the early decades of the twentieth century, the movement attracted broad scientific and medical support in Europe and North America from well-known liberal Protestants who lobbied for legislation to require eugenic sterilization. Religious bodies and leading clergy in England and on the Continent supported national eugenics legislation. In 1930, Pope Pius XI, in an encyclical on marriage, condemned eugenics. In doing so, he was criticized by scientists and some Protestant clergy, who believed that scientific research supported it. Roman Catholic and some conservative Protestant opposition helped to defeat eugenics laws in several American states in which the Catholic population was large.

In the latter third of the twentieth century, biotechnology raised new ethical issues that elicited responses from religious groups. These included medical experimentation on human subjects, artificial insemination, *in vitro* fertilization, organ transplantation, cloning, and stem cell research. On genetic engineering, widespread concern existed across the theological spectrum that genetic experimentation might usurp God's creative role and that transgenic animal research and human cloning could alter the genetic makeup of humans. On the other hand, some theologians, such as Ronald Cole-Turner (1993) and Ted Peters (1997), argued that genetic engineering offered people opportunities to become co-creators with God.

Ecology and the Environmental Movement

Other scientific matters drew the attention of religious communities in the late twentieth century. Prominent was the Protestant theme (although it is by no means exclusively Protestant) of man's stewardship over nature, which involved the care of creation that God had entrusted to the human race. The duty of Christians to care for the environment, with special concern for other creatures, extended from the Puritans to John Wesley and was earlier advocated by St. Benedict (*c.* 480–543) and St. Francis of Assisi (1182–1226). It undergirded physicotheology in the eighteenth and nineteenth centuries and exploded into the ecological movement in the late twentieth century. In 1967, Lynn White Jr. famously argued that the marriage of science and technology was responsible for environmental devastation in the West and that its ideological origin lay in the Christian view that God had given man dominion over nature. The White thesis provoked considerable controversy, in which alternative explanations were offered. They included the democratization that followed the French Revolution and the concept of a limitless frontier that had influenced American history. Keith Thomas asserted that White had overestimated the influence of religion on the motivation of human behavior and that it was man's stewardship over nature, not domination over it, that was in fact a biblical theme. Indeed, in the last third of the twentieth century, many Christian churches adopted environmental concerns with enthusiasm in what David Livingstone has termed 'the greening of theology.' Ecological themes have entered Christian discourse and influenced its hymns, liturgies, and creeds, while a spate of books representing almost every religious point of view has joined the chorus of environmental concern. They include a variety of pantheistic and organicist views, such as eco-feminism and James Lovelock's Gaia hypothesis, which has enjoyed popularity.

Conclusion

By the 1940s, the modern synthesis of genetics and Darwin's theory of natural selection, which accounted for random variation by genetic mutation, made it possible for biological scientists to adopt a fully naturalistic position in both

method and metaphysics. Naturalistic evolution, it was alleged, was able to offer a complete explanation for the origin and development of the universe and for all forms of life, as well as for social factors like language, religion, and ethics. While some scholars considered materialism to be the only philosophical position capable of scientific support, this view is by no means uniform in the scientific community. The last decade of the twentieth century, moreover, saw a resurgence of interest in the relationship of religion to the sciences. This interest arose from a variety of sources both within and outside the scientific community, including those of Christian and other theistic traditions (who most often asserted the complementarity of religion and the sciences), process philosophers, proponents of New Age and pantheistic spiritualism, and non-Western religious traditions. Perhaps it reflected the widespread recognition that, for all its achievements, science does not provide ultimate answers to questions regarding the meaning of the universe and of life itself.

Bibliography

Barbour, I.G., 1966. Issues in Science and Religion. Harper & Row, New York.

Barbour, I.G., 1997. Religion and Science: Historical and Contemporary Issues. HarperCollins, San Francisco, CA.

Barbour, I.G., 2000. When Science Meets Religion. HarperOne, New York.

Brooke, J.H., 1991. Science and Religion: Some Historical Perspectives. Cambridge University Press, Cambridge.

Brooke, J.H., Cantor, G., 1998. Reconstructing Nature: The Engagement of Science and Religion. The 1995–1996 Gifford Lectures at Glasgow. T & T Clark, Edinburgh.

Brooke, J.H., Numbers, R.L. (Eds.), 2011. Science and Religion around the World. Oxford University Press, New York.

Burtt, E.A., 1932. The Metaphysical Foundations of Modern Physical Science, revised ed. Harcourt, Brace, New York.

Clayton, P., Simpson, Z. (Eds.), 2006. The Oxford Handbook of Religion and Science. Oxford University Press, New York.

Dixon, T., 2008. Science and Religion: A Very Short Introduction. Oxford University Press, Oxford.

Draper, J.W., 1928 (1874). History of the Conflict between Religion and Science. Appleton, New York.

Drees, W., 1999. Religion, Science and Naturalism. Cambridge University Press, Cambridge.

Ferngren, G. (General Ed.), 2000. The History of Science and Religion in the Western Tradition: An Encyclopedia. Garland, New York.

Ferngren, G. (Ed.), 2002. Science and Religion: A Historical Introduction. Johns Hopkins University Press, Baltimore and London.

Ferngren, G., 2014. Medicine and Religion: A Historical Introduction. Johns Hopkins University Press, Baltimore, MD.

Haag, J.W., Peterson, G.R., Spezio, M.L. (Eds.), 2012. The Routledge Companion to Religion and Science. Routledge, London and New York.

Harrison, P. (Ed.), 2010. The Cambridge Companion to Religion and Science. Cambridge University Press, Cambridge.

Haught, J., 1995. Science and Religion: From Conflict to Conversation. Paulist Press, New York.

Jaki, S.L., 1974. Science and Creation: From Eternal Cycles to an Oscillating Universe. Scottish Academic Press, Edinburgh.

Koyré, A., 1957. From the Closed World to the Infinite Universe. Johns Hopkins Press, Baltimore, MD.

Lindberg, D.C., Numbers, R. (Eds.), 1986. God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA.

Lindberg, D.C., Numbers, R. (Eds.), 2003. When Science and Christianity Meet. University of Chicago Press, Chicago and London.

Numbers, R., Lindberg, D. (Eds.), 2003. When Science and Christianity Meet. University of Chicago Press, Chicago, IL.

Richardson, W.M., Wildman, W. (Eds.), 1996. Religion and Science: History, Method, Dialogue. Routledge, New York.

Rolston, H., 2006. Science and Religion: A Critical Survey, new ed. Templeton Foundation Press, Philadelphia, PA.

Smith, H., 2001. Why Religion Matters: The Fate of the Human Spirit in an Age of Disbelief. HarperSanFrancisco, San Francisco, CA.

Stenmark, M., 2004. How to Relate Science and Religion: A Multidimensional Model. Eerdmans, Grand Rapids, MI.

White, A.D., 1896. A History of the Warfare of Science with Theology in Christendom, vols. 2. Appleton, New York.

Theology and Science

Barbour, I., 1990 (Gifford Lectures 1989–1991). Religion in an Age of Science, vol. 1. HarperOne, San Francisco, CA.

Barbour, I., 2002. Nature, Human Nature, and God. Fortress Press, Minneapolis, MN.

Clayton, P., 1997. God and Contemporary Science. Eerdmans, Grand Rapids, MI.

Edwards, D., 2004. Breath of Life: A Theology of the Creator Spirit. Orbis, Maryknoll, NY.

Ellis, G., Murphy, N., 1996. On the Moral Nature of the Universe. Eerdmans, Grand Rapids, MI.

Gregersen, N.H., van Huyssteen, J.W. (Eds.), 1998. Rethinking Theology and Science: Six Models for the Current Dialogue. Eerdmans, Grand Rapids, MI.

McGrath, A.E., 2009. Science and Religion: A New Introduction, second ed. Wiley-Blackwell, Chichester, UK.

Murphy, N., 1990. Theology in the Age of Scientific Reasoning. Cornell University Press, Ithaca and London.

Murphy, N., Ellis, G.F.R., 1996. On the Moral Nature of the Universe. Fortress Press, Minneapolis, MN.

Pannenberg, W., 1976. Theology and the Philosophy of Science (F. McDonagh, Trans.). Westminster Press, Philadelphia, PA.

Peacocke, A., 1971. Science and the Christian Experiment. Oxford University Press, London.

Peacocke, A., 1993. Theology for a Scientific Age: Being and Becoming – Natural, Divine, and Human. Fortress Press, Minneapolis, MN.

Peters, T., Bennett, G. (Eds.), 2003. Bridging Science and Religion. Fortress Press, Minneapolis, MN.

Polkinghorne, J., 1994. Science and Christian Belief. Society for Promoting Christian Knowledge, London.

Polkinghorne, J., 2001. The Work of Love: Creation as Kenosis. Eerdmans, Grand Rapids, MI.

Polkinghorne, J., 2005. Science and Providence: God's Interaction with the World. Templeton Foundation Press, Philadelphia and London.

Shults, F.L., 2003. Reforming Theological Anthropology: After the Philosophical Turn to Relationality. Eerdmans, Grand Rapids, MI.

Shults, F.L., 2008. Christology and Science. Ashgate, Aldershot, UK.

Southgate, C. (Ed.), 2005. God, Humanity and the Cosmos: A Companion to the Science-Religion Debate, second ed. T & T Clark, London.

Torrance, T.F., 1969. Theological Science. Oxford University Press, London.

van Huyssteen, J.W., 1999. The Shaping of Rationality: Toward Interdisciplinary in Theology and Science. Eerdmans, Grand Rapids, MI.

van Huyssteen, J.W., 2006. Alone in the World?: Human Uniqueness in Science and Theology. Eerdmans, Grand Rapids, MI.

Ward, K., 2008. The Big Questions in Science and Religion. Templeton Press, West Conshohocken, PA.

Galileo and the Scientific Revolution

Biagoli, M., 1993. Galileo Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago, IL.

Fantoli, A., 1994. Galileo: For Copernicanism and for the Church (V. George, S.J. Coyne, Trans.). Vatican Observatory, Vatican City.

Feldhay, R., 1995. Galileo and the Church: Political Inquisition or Critical Dialogue? Cambridge University Press, Cambridge, UK.

Finnocchiaro, M.A. (Ed.), 1989. The Galileo Affair: A Documentary History. University of California Press, Berkeley, CA.

Harrison, P., 1998. The Bible, Protestantism, and the Rise of Natural Science. Cambridge University Press, Cambridge.

McMullin, E., 2005. The Church and Galileo. University of Notre Dame Press, Notre Dame, IN.

Victorian Science and Religion

Cantor, G., 1991. Michael Faraday: Sandemanian and Scientist. Macmillan, London.
Fyfe, A., 2004. Science and Salvation: Evangelical Popular Science Publishing in Victorian Britain. University of Chicago Press, Chicago, IL.
Lightman, B., 2007. Victorian Popularizers of Science: Designing Nature for New Audiences. University of Chicago Press, Chicago, IL.
Turner, F.M., 1974. Between Science and Religion: The Reaction to Scientific Naturalism in Late Victorian England. Yale University Press, New Haven, CT.
Turner, F.M., 1978. Victorian conflict between science and religion: a professional dimension. Isis 49, 356–376.

Darwin and Evolution

Artigas, M., Glick, T.F., Martinez, R.A., 2006. Negotiating Darwin: The Vatican Confronts Evolution, 1877–1902. Johns Hopkins University Press, Baltimore, MD.
Durant, J. (Ed.), 1985. Darwinism and Divinity. Blackwell, Oxford/New York.
Haught, J., 2007. God after Darwin, second ed. Westview Press, New York.
Hefner, P., 2000. The Human Factor. Augsburg, Minneapolis, MN.
Johnson, P.E., 1991. Darwin on Trial. Regnery Gateway, Washington, DC.
Larson, E.J., 1998. Summer for the Gods: The Scopes Trial and America's Continuing Debate over Science and Religion. Basic Books, New York.
McMullin, E. (Ed.), 1985. Evolution and Creation. Universtiy of Notre Dame Press, Notre Dame, IN.
Miller, K., 1999. Finding Darwin's God: A Scientist's Search for Common Ground between God and Evolution. HarperCollins, New York.
Moore, J.R., 1979. The Post-darwinian Controversies: A Study of the Protestant Struggle to Come to Terms with Darwin in Great Britain and America, 1870–1900. Cambridge University Press, Cambridge.
Numbers, R.L., 1998. Darwinism Comes to America. Harvard University Press, Cambridge, MA and London.
Peters, T., Hewlett, M., 2003. Evolution from Creation to New Creation: Conflict, Conversation, and Convergence. Abingdon Press, Nashville, TN.
Ruse, M., 2005. The Evolution-Creation Struggle. Harvard University Press, Cambridge, MA.

Creationism and Intelligent Design

Behe, M., 1996. Darwin's Black Box: The Biochemical Challenge to Evolution. Free Press, New York.
Dembski, W.A., 1998. The Design Inference: Eliminating Chance through Small Probabilities. Cambridge University Press, Cambridge.
Numbers, R.L., 2006. The Creationists: From Scientific Creationism to Intelligent Design, second ed. Harvard University Press, Cambridge, MA.
Whitcomb, Jr., J.C., Morris, H.M., 1961. The Genesis Flood: The Biblical Flood and Its Scientific Implications. Presbyterian and Reformed Publishing Co, Philadelphia, PA.

Catholicism and Science

Harris, S., 1996. Confession-building, long-distance networks, and the organization of Jesuit science. Early Science and Medicine 1, 287–318.
Heilbron, J.L., 1989. Science in the church. Science in Context 3, 9–28.
Heilbron, J.L., 1999. The Sun in the Church: Cathedrals as Solar Observatories. Harvard University Press, Cambridge, MA.
Kragh, H., 1987. The beginning of the world: George Lemaître and the expanding universe. Centaurus 32, 114–139.
O'Leary, D., 2006. Roman Catholicism and Modern Science: A History. Continuum, New York.

Judaism and Science

Cantor, G., Swetlitz, M. (Eds.), 2006. Jewish Tradition and the Challenge of Darwinism. University of Chicago Press, Chicago, IL.
Efron, N., 2007. Judaism and Science: A Historical Introduction. Greenwood Press, Westport, CT and London.
Fisch, M., 1997. Rational Rabbis: Science and Talmudic Culture. Indiana University Press, Bloomington, IN.
Hollinger, D., 1996. Science, Jews, and Secular Culture: Studies in Mid-Twentieth-Century American Intellectual History. Princeton University Press, Princeton, NJ.
Peters, T., Iqbal, M., Nomanul Haq, S. (Eds.), 2003. God, Life, and the Cosmos: Christian and Islamic Perspectives. Ashgate, Aldershot, UK.
Ruderman, D.B., 1995. Jewish Thought and Scientific Discovery in Early Modern Europe. Yale University Press, New Haven, CT.
Samuelson, N., 2008. Jewish Faith and Modern Science: On the Death and Rebirth of Jewish Philosophy. Rowman & Littlefield, Lanham, MD.
Turner, R. (Ed.), 2008. Design and Destiny: Jewish and Christian Perspectives on Human Germline Modification. MIT Press, Cambridge, MA.

Islam and Science

Dallal, A., 2000. Science, medicine, and technology: the making of a scientific culture. In: Esposito, J.L. (Ed.), The Oxford History of Islam. Oxford University Press, New York, pp. 155–213.
Edis, T., 2007. An Illusion of Harmony: Science and Religion in Islam. Prometheus Books, New York.
Hoodbhoy, P., 1991. Islam and Science: Religious Orthodoxy and the Battle for Rationality. Zed Books, London & Atlantic Highlands, NJ.
Huff, T.E., 1993. The Rise of Early Modern Science: Islam, China, and the West. Cambridge University Press, Cambridge.

Cosmology

Barrow, J.D., Tipler, F.J., 1986. The Anthropic Cosmological Principle. Clarendon Press, Oxford.
Drees, W.B., 1990. Beyond the Big Bang: Quantum Cosmology and God. Open Court, La Salle, IL.
Gingerich, O., 2006. God's Universe. Belknap Press of Harvard University Press, Cambridge, MA.
Godart, I., Heller, M., 1985. Cosmology of Lemaitre. Pachart, Tucson, AR.
Jaki, S.L., 1989. God and the Cosmologists. Scottish Academic Press, Edinburgh.
Matt, D., 1996. God and the Big Bang: Discovering Harmony between Science and Spirituality. Jewish Lights, Woodstock, VT.
Primack, J.R., Abrams, N.E., 2006. The View from the Center of the Universe: Discovering Our Extraordinary Place in the Cosmos. Riverhead, New York.
Russell, R.J., 2008. Cosmology: From Alpha to Omega. Fortress Press, Minneapolis, MN.

Physics

Brooke, J.H., Osler, M.J., van der Meer, J. (Eds.), 2001. Science in Theistic Context: Cognitive Dimensions. University of Chicago Press, Chicago, IL.
Davies, P.C., 1983. God and the New Physics. Simon and Schuster, New York.
Russell, R.J., Stoeger, W., Coyne, G. (Eds.), 1988. Physics, Philosophy, and Theology: A Common Quest for Understanding. Vatican Observatory, Vatican City.
Russell, R.J., Murphy, N., Peacock, A., 1995. Chaos and Complexity: Scientific Perspectives on Divine Action. Vatican Observatory, Vatican City.
Tipler, F., 1994. The Physics of Immortality. Doubleday, New York.
Worthing, M.W., 1996. God, Creation, and Contemporary Physics. Fortress, Minneapolis, MN.

Quantum Theory

Jaki, S.L., 1986. Chance or Reality and Other Essays. University Press of America, Lanham, MD.
Jaki, S.L., 1987. The Road of Science and the Ways to God. University of Chicago Press, Chicago, IL.
Russell, P.C., Wegter-McNelly, K., Polkinghorne, J. (Eds.), 2001. Quantum Mechanics: Scientific Perspectives on Divine Action. Vatican Observatory Publications and Center for Theology and the Natural Sciences, Vatican City State and Berkeley, CA.

Genetics and Biotechnology

Cole-Turner, R., 1993. The New Genesis: Theology and the Genetic Revolution. Westminster/John Knox Press, Louisville, KY.
Cole-Turner, R. (Ed.), 1997. Human Cloning: Religious Responses. Westminster/John Knox Press, Louisville, KY.
Holland, S., Lebacqz, K., Zoloth, L. (Eds.), 2001. The Human Embryonic Stem Cell Debate: Science, Ethics, and Public Policy. MIT Press, Cambridge, MA.
Peterson, J.C., 2001. Genetic Turning Points: The Ethics of Human Genetic Intervention. Eerdmans, Grand Rapids, MI.
Rolston, H., 1999. Genes, Genesis and God. Cambridge University Press, Cambridge.
Waters, B., Cole-Turner, R. (Eds.), 2003. God and the Embryo: Religious Voices on Stem Cells and Cloning. Georgetown University Press, Washington, DC.

Ecology and the Environment

Attfield, R., 1983. Christian attitudes to nature. Journal of the History of Ideas 44, 369–386.
Gottlieb, R.S., 2006. The Oxford Handbook of Religion and Ecology. Oxford University Press, Oxford.
Kalland, A., 2005. The religious environmentalist paradigm. In: Taylor, B.R. (Ed.), The Encyclopedia of Religion and Nature, vol. 2. Continuum, London & New York, pp. 1367–1371.
Leopold, A., 1966. A Sand County Almanac, with Essays on Conservation from Round River. Ballantine Books, New York.
Livingston, D.N., 2000. Ecology and the environment. In: Ferngren, G. (Ed.), The History of Science and Religion: An Encyclopedia, pp. 429–434.
Lovelock, J., 1979. Gaia: A New Look at Life on Earth. Oxford University Press, Oxford, UK.
McGrath, A.E., 2002. The Retrenchment of Nature: The Denial of Religion and the Ecological Crisis. Doubleday, New York.
Thomas, K., 1984. Man and the Natural World: Changing Attitudes in England, 1500–1800. Penguin, London.
White, Jr., L., 1967. The historical roots of our ecologic crisis. Science 155, 1203–1207.
Worster, D., 1977. Nature's Economy: A History of Ecological Ideas. Cambridge University Press, Cambridge, UK.

Neuroscience and the Soul

Brown, W.S., Murphy, N., Malony, H.N. (Eds.), 1998. Whatever Happened to the Soul? Scientific and Theological Portraits of Human Nature. Fortress Press, Minneapolis, MN.
Clayton, P., 2004. Mind and Emergence: From Quantum to Consciousness. Oxford University Press, New York.
Clayton, P., Davies, P. (Eds.), 2006. The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion. Oxford University Press, New York.
Newberg, A., d'Aquili, E., 2002. Why God Won't Go Away: Brain Science and the Biology of Belief. Ballantine Books, New York.
Russell, R.J., Murphy, N., Meyering, T.C., Arbib, A.E. (Eds.), 1999. Neuroscience and the Person: Scientific Perspectives on Divine Action. University of Notre Dame Press, Notre Dame, IN.

Religion and Stratification

Lisa A Keister and Brad Fulton, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B.S. Turner, Volume 19, pp. 13136–13139, © 2001, Elsevier Ltd.

Abstract

There is a strong and well-documented relationship between religion and social stratification. Social scientists have studied this connection for decades, and the field has recently gained considerable momentum. We survey recent theoretical approaches to studying religion and stratification. We then provide an overview of some of the most important empirical contributions of contemporary literature. We focus on religion's role in shaping family processes, education, work, income, wealth, immigration, and health. We end with a discussion of nonreligious people. Our conclusion highlights gaps in the literature.

Religion and Stratification

Religion plays an important role in creating and maintaining social and economic inequality, and this relationship attracted considerable attention in the early days of social science (Durkheim (1912)1954; Sombart 1911; Weber (1905)1930). In the 1960s, researchers revived questions about religion and stratification, but the research lost momentum when debates about socioeconomic status convergence between mainline Protestants (MPs; members of a religiously liberal Protestant church, who typically do not accept the Bible literally and tend to be more socially liberal than conservative Protestants; CPs) and Catholics came to dominate the literature, and survey data at the time were unable to adjudicate competing claims (Glenn and Hyland, 1967; Lenski, 1961; Roof and McKinney, 1987). The study of religion and inequality has again begun to thrive, and contemporary research is extremely rich, both theoretically and empirically. In this article, we survey recent theoretical approaches to studying religion and stratification. We then provide an overview of some of the most important empirical contributions of contemporary literature. We focus on religion's role in shaping family processes and education, work and income, and wealth (wealth is usually defined as net worth or total assets, less total debts). We then address the important role that immigration and health play in these relationships. We end our article with a discussion of processes unique to nonreligious people. Our conclusion highlights gaps in the current literature.

Theoretical Approaches

Much of contemporary research on religion and stratification draws on the status attainment and life course traditions in the social sciences. Status attainment refers to the process by which individuals arrive at socioeconomic standing during their lives, and the status attainment approach has become one of the most widely used theoretical perspectives in sociological research on social and economic well-being. The theoretical foundation of this approach is Blau and Duncan's (1967) seminal research on occupational attainment. This groundbreaking work demonstrated that achieved status (i.e., education and prior occupational prestige) was the most important determinant of attainment (i.e., occupational prestige), even controlling for ascribed status (i.e., parents' status). An extensive literature has subsequently extended the basic status attainment model to explain educational attainment, income, wealth, and other measures of well-being, usually as a function of family background and individual traits – like education – that are acquired over time (Campbell and Henretta, 1980; Looker and Pineo, 1983). Recently, scholars of religion have shown that religious beliefs are an important part of the attainment process. Religious beliefs from both childhood and adulthood can affect the degree to which attainment of certain traits (e.g., education, family size) are important, and religion appears to be highly correlated with the likelihood that individuals attain these states (Keister, 2012).

The life course perspective is another common approach used to understand the relationship between religion and inequality. Life course research represents an important theoretical approach used to understand changes that occur in individual lives over time, and it has been used effectively to understand how religion affects inequality. Life course ideas are frequently used to understand changes that occur in individual lives over time. Research in this tradition typically approaches the individual life course as a coherent entity with 'multiple antecedent-consequent linkages that give it shape and substance' (Kerckhoff, 1993: p. 3). This starting point leads life course researchers to talk about trajectories or pathways over the lives of individuals and to mark these trajectories or pathways by the individuals' ages at key turning points (O'Rand, 1996; Warren et al., 2002). Life course research also typically assumes that there are patterns or regularities across individual life courses that can be identified, explained, and used to understand other outcomes (Kerckhoff, 1993). The notion of a life course captures the nuances and shapes of the trajectory within which religion influences status. Religious beliefs are dynamic and can either create or result from important turning points, and ideas from life course research have the potential to capture patterns in the resulting paths. In the following sections, we elaborate on how these ideas have come to shape the literature concerning religion and stratification.

Childhood Family Processes

Modern data from the United States provide powerful evidence that religious affiliation is correlated with various family processes and human capital. Childhood religion and childhood family traits, most notably family size, are strongly associated – for example, CPs, members of a religiously conservative Protestant church). CPs typically accept the Bible as the literal word of God and are more socially conservative than MPs, and Mormons (LDSs) tend to come from relatively large families. Today's adult white and Hispanic Roman Catholics (RCs) also both tend to come from large families, although white RC family sizes have become smaller in recent decades (D'Antonio et al., 2007; Keister, 2007). In contrast, MPs and Jews tend to come from much smaller families. Family size is an important determinant of adult attainment and well-being: those from larger families tend to have lower adult education levels, incomes, and wealth because larger families have fewer resources, both material and nonmaterial, to devote to children's intellectual development and education (Blake, 1989; Downey, 1995).

Adult Family: Marriage and Fertility

There are strong associations between religious affiliation and adult family processes, including marriage and fertility. Religious beliefs affect orientations toward the desirability of marriage, the age at which people should first marry, and related decisions about cohabitation (Hammond et al., 1993; Lehrer, 2004b, 2008; Mosher et al., 1992). Decisions about marriage and cohabitation, in turn, are closely linked to orientations and decisions regarding family, education, and work. In relatively conservative faiths, women are encouraged to focus on home and family activities, are encouraged to have large families, and are exposed to others who marry early and focus on home activities. As a result, early marriage may be perceived as both desirable and acceptable by members of these faiths (Lehrer, 2004b). Women in conservative faiths also tend to have low education levels, which limits job prospects and increases the appeal of early marriage (Hammond et al., 1993; Lehrer, 2004b, 2008; Mosher et al., 1992). In more liberal faiths, early marriage is uncommon, while cohabitation is relatively common, and educational and career attainment are encouraged, so early marriage is less appealing. Religion also affects the choice of spouse, including whether the couple practices the same religion (homogamy) (Lehrer, 1998; Sherkat, 2004); marital stability and satisfaction (Lehrer, 1996; Lehrer and Chiswick, 1993); and the likelihood of divorce (Call and Heaton, 1997; Lehrer, 2008). The selection of a marriage partner is often motivated by homogamy (e.g., education, background, and other traits), and religion is a critical dimension according to which people select a partner. Religious homogamy is an important predictor of marital stability and satisfaction, and religious heterogamy can be destabilizing and lead to divorce (Lehrer, 2004b, 1998). Couples who share the same religious beliefs can share spiritual experiences, participate jointly in religious observances and activities, both at home and in other settings, and can develop overlapping social relationships originating in religious groups. Religious homogamy also increases the likelihood that couples are similar on a large number of demographic behaviors and processes, including education, childrearing, work patterns, and decisions about finances.

Religious beliefs – separate from the effects of religious homogamy – can also affect the likelihood of divorce. In faiths in which divorce is explicitly prohibited, the social costs associated with marital dissolution are high, and the decision to divorce can generate severe social and spiritual consequences. Catholicism, for instance, has a well-known prohibition against divorce. Catholic doctrine regards marriage as a sacrament, and any valid marriage between two baptized Catholics cannot be dissolved (D'Antonio et al., 2007; Tropman, 2002). Until relatively recently, divorce was very rare among Catholics, adding social pressure to the spiritual pressure to remain married. Although divorce has increased among Catholics in recent decades, particularly among white Catholics in the United States, the likelihood of divorce is still lower for Catholics than for those from other spiritual traditions with comparable educations, incomes, and other demographic traits (D'Antonio et al., 2001, 2007).

Marriage and divorce are consequential for stratification outcomes. Marriage increases earnings and wealth, whereas divorce has a negative impact on fortunes, particularly for women (Waite and Lehrer, 2003; Keister, 2005). Part of the increase in wealth reflects the tendency of married couples to combine their assets and to begin treating formerly separate assets as joint property (Keister, 2005). Couples join savings and checking accounts and combine investments, they purchase homes rather than rent, and they otherwise consolidate finances into jointly owned property. Marriage allows couples to pool risks (e.g., if one person is unemployed, the other can continue working), creates economies of scale (e.g., in housing costs), and allows people to take advantage of a division of labor (Waite and Lehrer, 2003). Marriage also creates common goals (e.g., children and children's educations, home improvements and upgrades, and retirement objectives) that encourage couples to save. Early marriage reduces attainment of other outcomes such as education, but early marriage also might increase the likelihood of early saving and home ownership.

Fertility behavior is an important part of the causal processes linking religion, education, and family processes with inequality. It is well documented that religion influences orientations toward premarital sex and the onset of sexual activity, attitudes regarding birth control and the use of contraception, whether a person or couple has any children, the age at which people have their first child (i.e., age at first birth), family size, and even behaviors such as taking virginity and secondary abstinence pledges (Lehrer, 1996, 2004b; McQuillan, 2004; Sherkat and Ellison, 1999). CPs, Muslims, Hindus, and Catholics tend to have more traditional approaches to sexual activity and family formation. Some religious teachings and traditions include social and psychological rewards for having large families. In such pronatalist faiths, early fertility is approved of and even encouraged both by formal teachings and rules and through norms, and having many children can provide considerable social status (Lehrer, 2004b; McQuillan, 2004). LDS fertility

rates have been notably high in the United States, at least in part in response to such incentives (Lehrer, 2004b, 2008; Stark and Finke, 2000). Catholicism also strongly discourages contraceptive use and abortion, and Catholics have larger families than Protestants in many nations, though in the United States, white Catholics' family sizes have declined since the 1970s, while family sizes remain large among Hispanic Catholics (Lehrer, 2004b; Mosher et al., 1992). There is considerable variation in pronatalism across Protestant denominations, particularly in the United States, and CPs have strong norms that encourage large families when compared to MPs, Jews, and other religious groups.

Fertility behaviors have important effects on education, income, savings, and wealth. Having children early in life is particularly difficult because having a family can make it difficult to finish or complete schooling, makes career development more challenging, and can reduce initial saving and investing that can contribute to life-long asset appreciation (Keister, 2005). Studies have found that the timing of family formation helps to explain religious differences in educational attainment in the United States (Fitzgerald and Glass, 2008). Members of faiths in which early childbearing and large families are encouraged or are simply common generally wind up on the bottom of the stratification hierarchy.

Human Capital

There is extensive literature documenting the important role that religion plays in human capital acquisition in the United States (Burstein, 2007; Fitzgerald and Glass, 2008; Glass and Jacobs, 2005; Lehrer, 2004a; Sherkat and Darnell, 1999), and education is one of the strongest predictors of income and wealth accumulation. Table 1 shows that there are notable patterns in educational attainment according to religious groups in the United States. Educational attainment is particularly high among those raised in Jewish families: both years of education and the percentage holding an advanced degree (beyond a bachelor's degree) are exceptionally high for Jews. MPs also stand out in their educational attainments. White Catholics are similar to MPs in the number of years of education they attain, but they still lag behind in completing advanced degrees, a pattern that is consistent with knowing that education levels have only recently begun to increase for white Catholics. Table 1 also shows that CPs have relatively low educational attainment (Darnell and Sherkat, 1997; Fitzgerald and Glass, 2008; Glass and Jacobs, 2005; Greeley and Hout, 2006; Smith and Faris, 2005). In contrast, LDS respondents have notably higher levels of education, a pattern that sets them apart from other CP groups. Black Protestants (BPs; members of a traditionally African American Protestant church), black Catholics, and Hispanic Catholics report the lowest levels of educational attainment, particularly completion of advanced degrees. Americans who hold identifications with 'other religions' (mostly Buddhists, Hindus, Moslems, and Sikhs) also have high education levels.

Table 1 Childhood religion and educational attainment

	Years of education	*Having advanced degree (%)*
Protestant		
Conservative Protestant (CP)	13.2	10.7
Black Protestant (BP)	12.7	4.2
Mainline Protestant (MP)	14.3	19.5
Catholic	13.6	11.0
White	13.9	12.3
Hispanic	12.5	6.4
Black	13.4	4.0
Jewish	16.0	33.4
Other religion	14.9	25.2
No religious preference	12.7	9.5
Full sample	13.6	12.5

Total $n = 4369$.
Reproduced from Keister, Lisa A., 2012. Faith and Money: How Religious Belief Contributes to Wealth and Poverty. Cambridge University Press, Cambridge.

There are at least three reasons for the effects of religion on education. Some researchers emphasize the importance of constraints such as family background in shaping educational decisions and outcomes (Sherkat and Darnell, 1999). Others emphasize individual preferences (demand for education) and contextual or macrolevel conditions (supply of education) that determine options (Lehrer, 1999, 2004a, 2004b). Still others focus on particular characteristics of religious groups to explain patterns in educational attainment and human capital acquisition. CPs, for example, tend to be skeptical of the approaches taken in secular schools and universities that propagate secular humanist values and promote scientific investigation rather than acceptance of divine truths (Darnell and Sherkat, 1997; Sherkat, 2009; Sherkat and Darnell, 1999; Sikkink, 1999). These unique orientations contribute to low levels of education among CPs (Fitzgerald and Glass, 2008; Glass and Jacobs, 2005; Sherkat and Darnell, 1999). In contrast, particular traits have also been cited as significant factors in producing high levels of educational attainment among Jews (Burstein, 2007) and LDSs (Shaefer and Zellner, 2007). Burstein (2007) points out that cultural emphasis on pursuits in this life are important contributors to educational success for Jews. Shaefer and Zellner (2007) discuss the role played by the educational efforts of the LDSs, including a system of seminaries and Brigham Young University, in producing well-educated, committed LDS adults.

Work, Income, and Occupations

Religious commitments also impact occupational attainment and labor-force participation, and they have concomitant effects on personal and household incomes. One of the most pronounced relationships occurs between religious affiliation and gender differences in work behavior. There are important differences between men and women in various religious groups in terms of patterns of working full-time jobs. Women in more conservative religious traditions are less likely to work full time, whereas men in these groups are as likely as the general population to work full time. In the United States, white Catholic women of prior generations were more traditional and less likely than others to work full time, but today's

white Catholic women are nearly as likely as white Catholic men to work full time. Jewish women tend to work full time until their children are born; then they are more likely to stay at home while the children are young and return to full-time work after the first few years of the children's lives (Keister, 2012).

Although researchers have largely ignored the role that religion plays in shaping occupation and income, recent evidence shows that there are important connections. In the United States, CPs have the most low-prestige occupations, even relative to their low rates of education attainment. White RCs are moving into occupations that put them close to parity with MPs, a pattern that would have been unheard of in prior generations. At the other end of the spectrum, Jews and nonreligious people have considerable occupational advantages (Sherkat, 2012). Similarly, religion has important effects on income. Because religion is correlated with family background, adult family processes, human capital, work, and occupation, it is no surprise that many of the same patterns hold in the relationship between religion and income. CPs, BPs, and Hispanic RCs tend to have relatively low income, are more likely than others to have been raised in poverty (i.e., to have household incomes below the poverty line), are more likely to be in poverty as adults, are more likely to receive government transfer payments, and are more likely to experience extended spells of unemployment (Keister, 2012). In contrast, white RCs, MPs, and Jews have significant advantages in each of these dimensions (Keister, 2012). These findings are consistent across studies (Lehrer, 2010; Steen, 1996). Research shows that the ranking holds when various individual and family traits are controlled for, particularly the magnitude of the Jewish advantage (Lehrer, 2010; Steen, 1996).

Wealth: Assets and Debts

Wealth ownership is one of the most important components of social and economic stratification, and researchers have shown that wealth inequality is extreme in the United States. Wealth, or net worth, is total household assets less total liabilities. Wealth is relatively enduring – both within and across generations – and it is related in some way to most other measures of achievement. Wealth can enhance educational attainment, occupational opportunities, political power, and social influence. Religious groups with higher concentrations of wealth can solidify their advantages over other groups, and members of those groups will be better suited to deal with income interruptions, medical emergencies, and other crises such as war, accidents, and natural disasters. Wealth can create more wealth when it is reinvested, and it can generate income in the form of interest or dividends. At high levels of wealth, the income it generates can make paid employment unnecessary. Yet even a small amount of savings can improve security, mitigate the effects of job loss and other financial shocks, and improve well-being dramatically.

Religious commitments help to stratify societies in terms of wealth through their impact on the attainment processes described above, and religion appears to affect wealth ownership directly. In the United States, the resulting patterns are quite pronounced. CPs, BPs, and Hispanic RCs have very low levels of total net worth, real assets (e.g., tangible assets such as the home and other real estate), and financial assets (e.g., relatively liquid assets, such as stocks, bonds, mutual funds, and bank accounts), and they are very likely to have zero or negative net worth. In addition, members of these groups tend to accumulate assets relatively slowly across the entire life course, and they rarely enter high levels of the wealth distribution. In contrast, white RCs, MPs, and particular Jews are much more advantaged on each of these measures. Table 2 shows the extent of net worth differences by religious affiliation in the United States.

Immigration

Religion plays a particularly important, and theoretically interesting, role in the lives of immigrants, and religious stratification is very often a product of the variable fortunes of immigrants from different parts of the world. Social scientists in the United States have paid particular attention to religion's role in immigrant assimilation into the host country, with an emphasis on post-1965 immigrants using religion to navigate the transition process and achieve economic mobility (Hagan and Ebaugh, 2003; Hirschman, 2004). Immigrant religious organizations may also be important because they can facilitate cultural assimilation by simultaneously functioning as cultural havens and as acculturation resource centers. These organizations can ease the shock of entry into a new culture by providing space for the expression of immigrant customs, language, and group solidarity. At the same time, they can facilitate incorporation by providing resources for adapting to the new culture and navigating the naturalization process (Warner and Wittner, 1998). Similarly, religious participation can facilitate economic incorporation. Immigrant congregations often offer classes that help immigrants develop language skills, acquire professional certifications, and obtain work permits. They also provide immigrants with opportunities to develop leadership skills, build self-confidence, and expand their social networks, all of which can facilitate their searches for employment (Connor,

Table 2 Religion and wealth: median net worth

	Median net worth ($)
Protestant	
Conservative Protestant (CP)	53 800
Black Protestant (BP)	8000
Mainline Protestant (MP)	145 250
Catholic	141 500
White	333 219
Hispanic	187 188
Black	91 438
Jewish	392 500
Other religion	64 200
No religious preference	70 000
Full Sample	86 534

Total $n = 4369$.
Reproduced from Keister, Lisa A., 2012. Faith and Money: How Religious Belief Contributes to Wealth and Poverty. Cambridge University Press, Cambridge.

2011; Foley and Hoge, 2007). Although religion's role in facilitating incorporation can promote religious participation among first-generation immigrants, there is an ongoing debate regarding the long-term effects on immigrants. Commitments to religion may well foster segmentation rather than acculturation, and this may hinder attainment processes, perhaps especially for women brought up in patriarchal religious traditions.

Health

There is considerable evidence that religious involvement promotes both physical and mental health (Chatters, 2000), which could, in turn, influence educational, occupational, and wealth attainment. Scholars point to the social and psychological resources acquired through religious participation to explain this robust relationship, and they argue that religion affects health by operating through these resources. Being involved in a religious congregation can provide health-promoting social resources. Although social isolation can increase the risk for morbidity, people who attend religious services on a regular basis tend to have larger social networks and stronger network ties, which reduce the risks for social isolation and increase the opportunities to receive social support (Connor, 2011; Ellison and George, 1994). Despite the numerous health benefits associated with religion, religion can also undermine health. Some religious groups discourage seeking professional health care assistance, and some religious teachings proscribe specific medical procedures that are designed to promote health. Furthermore, religious belief can adversely affect health by producing serious psychological strain in adherents who wrestle with religious doubts or guilt (Krause and Wulff, 2004).

Nonreligious People

The number of nonreligious people has grown in most nations of the world, prompting scholars to look closely at this unique group who are often called 'nones.' Between 1990 and 2010, the number of people claiming no religious preference in the United States increased from 7 to 18%. In general, those who reject religious identification are more educated and also more likely to hold professional occupations and to earn higher salaries, particularly when compared to members of conservative religious groups. In the United States, those who reject religious identification are on par with MPs and white Catholics in terms of status attainment (Sherkat, 2012). However, it is unclear whether nonidentification is as strongly linked to stratification outcomes in nations where secularism is more common. Further, in highly religious nations with less meritocratic stratification systems, failure to identify as a religious person may have a negative impact on educational attainment, occupational attainment, and wealth accumulation. Finally, there is an important social-scientific question about the relationship of the association: does nonidentification or secularity promote attainment, or do education, income, and wealth motivate individuals to disavow religious commitments?

Conclusion

Renewed interest in religious stratification has led to several theoretical and methodological developments that have improved scholars' understanding of how religion creates and maintains social inequality through its influence on dynamics, such as family processes, human capital acquisition, occupational attainment, household income, wealth accumulation, immigrant transition, and health-related behavior. Scholarship in this area could be advanced by combining social capital and social network theory to analyze how the content and structure of networks interact to with religious social ties to affect attainment. Further, measures of religiosity are often limited to broad categories of identification, whereas the impact of religious beliefs over the life course is less well developed. This is particularly true for religious commitments and beliefs outside of the European Christian tradition. Most of the research into stratification and religion has used samples from the United States, and more comparative research is needed. A particularly exciting area of inquiry is the potential impact of religion on stratification for immigrants in Europe, where there has been a large influx of immigrants from Muslim nations. Until recently, widespread use of these approaches had been hindered by data limitations; however, as data sets have begun to include reliable social-network information and comprehensive measures of religiosity along with attainment-related variables, scholars now have the opportunity to explore more thoroughly the complex relationship between religion and social stratification.

Bibliography

Akresh, Ilana R., 2011. Immigrants' religious participation in the United States. Ethnic and Racial Studies 34, 643–661.

Baker, Joseph O., Smith, Buster G., 2009. The nones: social characteristics of the religiously unaffiliated. Social Forces 87, 1251–1263.

Bartkowski, John P., Ellison, Christopher G., 1995. Divergent perspectives on child-rearing in popular manuals: conservative Protestants vs. the mainstream experts. Sociology of Religion 56, 21–34.

Blake, Judith, 1989. Family Size and Achievement. University of California Press, Berkeley, CA.

Burstein, Paul, 2007. Jewish educational and economic success in the United States: a search for explanations. Sociological Perspectives 50, 209–228.

Cadge, Wendy, Ecklund, Elaine Howard, 2007. Immigration and religion. Annual Review of Sociology 33, 359–379.

Call, Vaughn R.A., Heaton, Tim B., 1997. Religious influence on marital stability. Journal for the Scientific Study of Religion 36, 382–392.

Campbell, Richard, Henretta, John, 1980. Status claims and status attainment: the determinants of financial well-being. American Journal of Sociology 86, 618–629.

Chatters, Linda M., 2000. Religion and health: public health research and practice. Annual Review of Public Health 21, 335–367.

Chiswick, Barry R., 1993. The skills and economic status of American Jewry: trends over the last half century. Journal of Labor Economics 11, 229–242.

Cnaan, Ram, Boddie, Stephanie, 2002. The Invisible Caring Hand: American Congregations and the Provision of Welfare. New York University Press, New York.

Connor, Phillip, 2011. Religion as resource: religion and immigrant economic incorporation. Social Science Research 40, 1350–1361.

D'Antonio, William V., Davidson, James D., Hoge, Dean R., Meyer, Katherine, 2001. American Catholics: Gender, Generation, and Commitment. Altamira Press, Walnut Creek, CA.

D'Antonio, William V., Hoge, Dean R., Davidson, James D., 2007. American Catholics Today: New Realities of Their Faith and Their Church. Rowman and Littlefield, Lanham, MD.

Darnell, Alfred, Sherkat, Darren E., 1997. The impact of Protestant fundamentalism on educational attainment. American Sociological Review 62, 306–315.

Downey, Douglas B., 1995. When bigger is not better: family size, parental resources, and children's educational performance. American Sociological Review 60, 746–761.
Durkheim, Emile, (1912)1954. The Elementary Forms of Religious Life. Free Press, New York.
Ellison, Christopher G., George, Linda K., 1994. Religious involvement, social ties, and social support in a southeastern community. Journal for the Scientific Study of Religion 33, 46–61.
Ellison, Christopher, Sherkat, Darren, 1995. Is sociology the core discipline for the scientific study of religion? Social Forces 73, 1255–1266.
Filsinger, Erik E., Wilson, Margaret R., 1984. Religiosity, socioeconomic rewards, and family development: predictors of marital adjustment. Journal of Marriage and the Family 46, 663–670.
Finke, Roger, Stark, Rodney, 2005. The Churching of America, 1776–2005: Winners and Losers in Our Religious Economy. Rutgers University Press, Rutgers, NJ.
Fitzgerald, Scott T., Glass, Jennifer, 2008. Can early family formation explain the lower educational attainment of US conservative Protestants? Sociological Spectrum 28, 556–577.
Foley, Michael, Hoge, Dean R., 2007. Religion and the New Immigrants: How Faith Communities Form Our Newest Citizens. Oxford University Press, New York.
George, Linda K., Ellison, Christopher G., Larson, David B., 2002. Explaining the relationships between religious involvement and health. Psychological Inquiry 13, 190–200.
Glass, Jennifer, Jacobs, Jerry, 2005. Childhood religious conservatism and adult attainment among black and white women. Social Forces 83, 555–579.
Glenn, Norval D., Hyland, Ruth, 1967. Religious preference and worldly success: some evidence from national surveys. American Sociological Review 32, 73–85.
Gordon, Milton, 1964. Assimilation in American Life: The Role of Race, Religion, and National Origins. Oxford University Press, New York.
Greeley, Andrew, Hout, Michael, 2006. The Truth about Conservative Christians: What They Think and What They Believe. University of Chicago, Chicago.
Hagan, Jacqueline, Ebaugh, Helen Rose, 2003. Calling upon the sacred: migrants' use of religion in the migration process. International Migration Review 37, 1145–1162.
Hammond, Judith A., Cole, Bettie S., Beck, Scott H., 1993. Religious heritage and teenage marriage. Review of Religious Research 35, 117–133.
Hirschman, Charles, 2004. The role of religion in the origins and adaptation of immigrant groups in the United States. International Migration Review 38, 1206–1233.
Hollinger, David A., 2004. Rich, powerful, and smart: Jewish overrepresentation should be explained instead of avoided or mystified. The Jewish Quarterly Review 94, 595–602.
Hout, Michael, Fischer, Claude S., 2002. Why more Americans have no religious preference: politics and generations. American Sociological Review 67, 165–190.
Hout, Michael, Morgan, William R., 1975. Race and sex variations in the causes of the expected attainments of high school seniors. American Journal of Sociology 81, 364–394.
Keister, Lisa A., 2005. Getting Rich: America's New Rich and How They Got that Way. Cambridge University Press, New York.
Keister, Lisa A., 2007. Upward wealth mobility: exploring the Roman Catholic advantage. Social Forces 85, 1195–1226.
Keister, Lisa A., 2011. Faith and Money: How Religious Belief Contributes to Wealth and Poverty. Cambridge University Press, New York.
Keister, Lisa A., 2012. Faith and Money: How Religious Belief Contributes to Wealth and Poverty. Cambridge University Press, Cambridge.
Kerckhoff, Alan C., 1993. Diverging Pathways: Social Structure and Career Deflections. Cambridge University Press, Cambridge.
Kosmin, Barry A., Keysar, Ariela, 2009. American Nones: The Profile of the No Religion Population. Trinity College, Hartford, CT.
Krause, Neal, Wulff, Keith M., 2004. Religious doubt and health: exploring the potential dark side of religion. Sociology of Religion 65, 35–56.
Lehrer, Evelyn, 1996. The determinants of marital stability: a comparative analysis of first and higher order marriages. In: Schultz., T.P. (Ed.), Research in Population Economics. JAI, Greenwich, CT, pp. 91–121.
Lehrer, Evelyn, 2004a. Religiosity as a determinant of educational attainment: the case of conservative Protestant women in the United States. Review of Economics of the Household 2, 203–219.
Lehrer, Evelyn, 2004b. The role of religion in union formation: an economic perspective. Population Research and Policy Review 23, 161–185.
Lehrer, Evelyn, 2008. Religion as a Determinant of Economic and Demographic Behavior. Routledge, New York.
Lehrer, Evelyn, 2010. Religious affiliation and participation as determinants of women's educational attainment and wages. In: Ellison, C.G., Hummer, R.A. (Eds.), Religion, Families, and Health: Population-Based Research in the United States. Rutgers University Press, New Brunswick, NJ, pp. 186–205.
Lehrer, Evelyn L., 1998. Religious intermarriage in the United States: determinants and trends. Social Science Research 27, 245–263.
Lehrer, Evelyn, 1999. Religion as a determinant of educational attainment: an economic perspective. Social Science Research 28, 358–379.
Lehrer, Evelyn L., Chiswick, Carmel U., 1993. Religion as a determinant of marital stability. Demography 30, 385–404.
Lenski, Gerhard, 1961. The Religious Factor: A Sociological Study of Religion's Impact on Politics, Economics, and Family Life. Doubleday, Garden City, NY.
Looker, E. Dianne, Pineo, Peter C., 1983. Social psychological variables and their relevance to the status attainment of teenagers. The American Journal of Sociology 88, 1195–1219.
Marwell, Gerald, Demerath, N.J., 2003'. 'Secularization' by any other name. American Sociological Review 68, 314–316.
McQuillan, Kevin, 2004. When does religion influence fertility? Population and Development Review 30, 532–543.
Mooney, Margarita A., 2009. Faith Makes Us Live: Surviving and Thriving in the Haitian Diaspora. University of California Press, Berkeley.
Mosher, William D., Williams, Linda B., Johnson, David P., 1992. Religion and fertility in the United States: new patterns. Demography 29, 199–214.
O'Rand, Angela M., 1996. The precious and the precocious: understanding cumulative advantage over the life course. The Gerontologist 36, 230–238.
Otto, Luther B., Haller, Archibald O., 1979. Evidence for a social psychological view of the status attainment process: four studies compared. Social Forces 57, 887–914.
Roof, Wade Clark, McKinney, William, 1987. American Mainline Religion: Its Changing Shape and Future. Rutgers University Press, New Brunswick, NJ.
Sampson, Robert J., Laub, John H., 1996. Socioeconomic achievement in the life course of disadvantaged men: military service as a turning point, circa 1940–1965. American Sociological Review 61, 347–367.
Sewell, William H., Hauser, Robert M., Wolf, Wendy, 1980. Sex, schooling, and occupational status. American Journal of Sociology 86, 551–583.
Shaefer, Richard T., Zellner, William W., 2007. Extraordinary Groups: An Examination of Unconventional Lifestyles. McMillan, New York.
Sherkat, Darren, 2009. Religion and verbal ability. Social Science Research 10, 2–13.
Sherkat, Darren, 2012. Religion and the American Occupation Structure. In: Keister, L.A., McCarthy, J., Finke, R. (Eds.), Religion, Work, and Money. Emerald, London.
Sherkat, Darren E., 2004. Religious intermarriage in the United States: trends, patterns, and predictors. Social Science Research 33, 606–625.
Sherkat, Darren E., Darnell, Alfred, 1999. The effects of parents' fundamentalism on children's educational attainment: examining differences by gender and children's fundamentalism. Journal for the Scientific Study of Religion 38, 23–35.
Sherkat, Darren E., Ellison, Christopher G., 1999. Recent developments and current controversies in the sociology of religion. Annual Review of Sociology 25, 363–394.
Sikkink, David, 1999. The social sources of alienation from public schools. Social Forces 78, 51–86.
Smith, Christian, Faris, Robert, 2005. Socioeconomic inequality in the American religious system: an update and assessment. Journal for the Scientific Study of Religion 44, 95–104.
Sombart, Werner, 1911. The Jews and Modern Capitalism. M. Epstein, Trans.. E. P. Dutton and Company, New York.
Stark, Rodney, Finke, Roger, 2000. Acts of Faith: Explaining the Human Side of Religion. University of California Press, Berkeley.
Steen, Todd P., 1996. Religion and earnings: evidence from the NLS Youth Cohort. International Journal of Social Economics 23, 47–58.
Tropman, John E., 1995. The Catholic Ethic in American Society: An Exploration of Values. Jossey-Bass, San Francisco, CA.
Tropman, John E., 2002. The Catholic Ethic and the Spirit of Community. Georgetown University Press, Washington, DC.
Waite, Linda, Lehrer, Evelyn L., 2003. The benefits from marriage and religion in the United States: a comparative analysis. Population and Development Review 29, 255–275.
Warner, Stephen, Wittner, Judith, 1998. Gatherings in Diaspora: Religious Communities and the New Immigration. Temple University Press, Philadelphia.
Warren, John Robert, Hauser, Robert M., Sheridan, Jennifer T., 2002. Occupational stratification across the life course: evidence from the Wisconsin longitudinal study. American Sociological Review 67, 432–455.
Weber, Max, (1905)1930. The Protestant Ethic and the Spirit of Capitalism. Harper-Collins, New York.
Wilder, Esther I., Walters, William H., 1997. American Jewish household income, 1969 and 1989. Journal of Economic and Social Measurement 23, 197–212.

Religion and Youth

Lisa D Pearce, University of North Carolina, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Collins, volume 19, pp. 13051–13054. © 2001 Elsevier Ltd.

Abstract

The study of religion and youth continues to grow and improve, fostering the development of new theories and data. Research on youth and religion generally fits six themes: (1) trends in religiosity and spirituality; (2) how religious or spiritual identity develops and changes; (3) how social networks, institutions, and contexts both shape and are shaped by youth religiosity and spirituality; (4) youth agency in the development of their religious or spiritual identity; (5) the degree to which religiosity or spirituality promotes well-being and prevents risky behavior; and (6) when religion has negative consequences for youth. In this article, I discuss theories, current knowledge, data, and methods common to these themes.

The study of religion in the period of the life course considered 'youth' has unique features and challenges, because this span of life (including childhood, adolescence, and emerging adulthood) encompasses immense physical and cognitive development, along with changes in social ties, autonomy, and societal expectations. Much of the research on youth and religion falls under six main themes. First are examinations of trends in religious affiliation, beliefs, practices, and salience. Second is a focus on how religious or spiritual identity develops and changes as children become adolescents and then adults. A third focus has been on how social networks, institutions, and contexts both shape and are shaped by youth religiosity and spirituality. A fourth theme emerging is the study of youth agency in the development and maintenance of their religious or spiritual identity. Fifth, research has highlighted the degree to which religiosity or spirituality promotes health (physical and mental) and achievement as well as prevents risky behavior. And a sixth theme is the study of the negative consequences of religion for youth. This article summarizes theoretical frameworks, methodological issues, and the current state of knowledge regarding youth and religion.

What Religion, Religiosity, and Spirituality Mean for Youth

Much has been written about the definition of religion and the meaning of spirituality to all ages, especially in relation to religiosity (cf Hill and Pargament, 2003; Roof, 2003; Wuthnow, 1998). Definitions of religion often involve an aspect of organization (e.g., 'a system'), and therefore refer to an external structure (e.g., a set of beliefs or an institution) that may or may not be internalized by individuals in a variety of ways (Geertz, 1993). It is also important to recognize that religions are not passively internalized (Verbit, 1970). They are in turn shaped by the human involvement they inspire. Thus, when we talk about religion in the lives of youth, it is a social institution that both influences and responds to them in interesting ways. Being 'religious,' then, refers to some manifestation of religion, and one's 'religiosity' or 'religiousness' suggests a level of engagement or identification with religion.

There are important nuances to theory about the sources and consequences of religiosity in the lives of youth depending on which dimensions of religiosity are the focus (Pearce and Denton, 2011). And multiple dimensions of religiosity are often packaged by youth in seemingly inconsistent ways (e.g., high centrality of religion and levels of private practice, but low public religiosity), which encourages examining religiosity as more of a bricolage than a linear variable that runs from low to high. Another issue is spirituality, which has very different meanings across contexts. Some connect spirituality to alternative or Eastern religious practices like meditation and may consider it wholly distinct from organized religion. Others view spirituality as a way of practicing their religious faith in ways that eschew what is viewed as the vapid routine of religious ritual. Some Christians call their spirituality a personal relationship with Jesus Christ. Some argue that these views of spirituality share a core element: a significant, personal connection to the sacred, a divine source of meaning (Hill and Pargament, 2003). Some youth find ways to enact their spirituality on their own, but many link their personal spirituality to participation in a religious community. Only a small minority of youth in the United States identify as spiritual but not religious (8% of 13–17 year olds and 12% of 16–21 year olds) – and similar results hold in qualitative studies (Denton et al., 2008; Pearce and Denton, 2011).

Trends in Youth Religiosity over Time

Studying trends, or stability and change, in youth religiosity and spirituality over time has taken two main forms – analyzing survey data for time trends, and rich observations of particular generations that thickly describe unique practices and understandings of religiosity and spirituality in youth.

Data from the Monitoring the Future Study (Bachman et al., 2000) of high school seniors in the United States show that although general levels of religious affiliation, attendance, and importance remain relatively high, there have been noticeable decreases in all three dimensions of religiosity over time since 1976. Since that time, more youth are claiming identification other than Protestant, Catholic, or

Jewish, and more are rejecting religious identification. Both groups doubled in size, from 11 to 22% and from 7 to 14%, respectively (see Figure 1).

There has also been interesting change in 12th graders' religious service attendance. Figure 2 shows that in 1976, the modal category by far was 'once or twice a month.' In 2006, attending 'rarely' became the most common response. The percentage attending weekly or more has dropped from 41% in 1976 to 30% in 2010. The percentage never attending rose from 11% in 1976 to 21% in 2010. Studies of youth religious involvement in the United Kingdom and Australia show much lower religious participation than in the United States and more steep declines in the adolescent years and beyond (Francis and Kay, 1995; Mason et al., 2007).

The importance of religion to 12th graders has not changed as much over time as the other religion measures. As seen in Figure 3, the percentage of youth finding religion 'very important' has remained fairly steady. It was 29% in 1976 and 28% in 2010. There were spikes up to 33% in the early 1980s and late 1990s. The percentage saying either 'pretty important' or 'a little important' has declined a little over time, and there has been a sizable increase in the percentage stating that religion is 'not important,' up from 13% in 1976 to 21% in 2010.

When it comes to religious beliefs, we have less trend data to work with, but recent studies give us a good sense of the contemporary situation. Data from the first wave (2002–03) of the National Study of Youth and Religion show that 84% of American teenagers (ages 13–17) believe in God, 12% are unsure, and only 3% do not believe in God (Smith and Denton, 2005). Between four and six out of 10 American teenagers say they believe in angels, demons, divine miracles, and life after death (Smith and Denton, 2005). In comparison, Francis and Kay (1995) find that among a representative sample of 13–15 year olds attending school

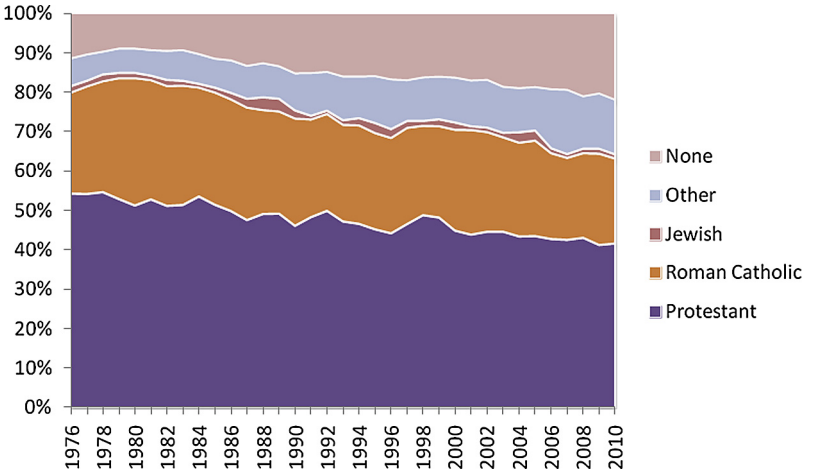

Figure 1 Religious affiliation of 12th graders in the United States, 1976–2010. Monitoring the Future.

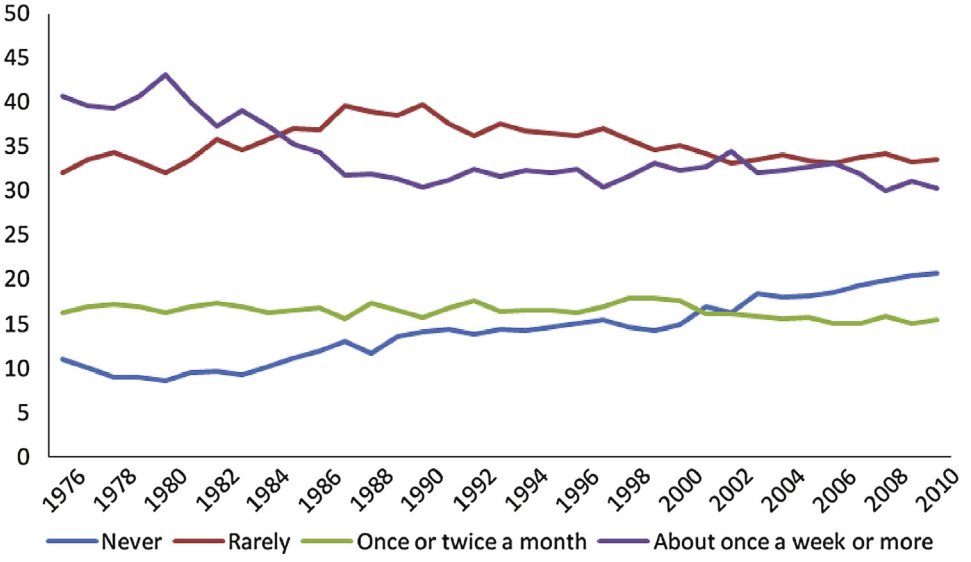

Figure 2 Religious service attendance of 12th graders in the United States, 1976–2010. Monitoring the Future.

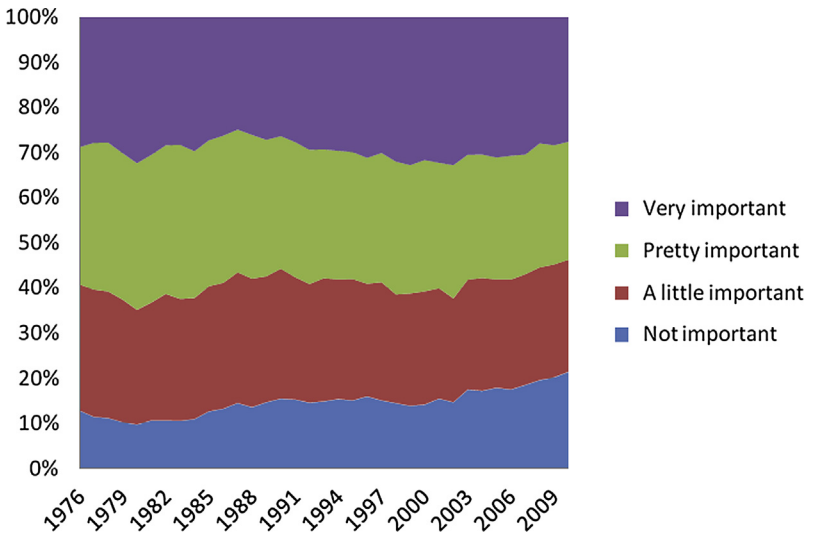

Figure 3 Importance of religion to 12th graders in the United States, 1976–2010. Monitoring the Future.

in England and Wales, 39% believe in God and 41% believe in life after death. Mason et al. (2007) report that just over one half of the Australian youth born between 1981 and 1995 believe in God, and this is a major decrease from the 81% of their parents' generation who reported believing in God at the same age.

Crockett and Voas (2006) find a great deal of change in religiosity across generations in Britain, with young British adults being only a little more than half as religious as their parents in terms of affiliation (59%), attendance (54%), and belief (56%). Although religious change has not been as dramatic in the United States, several studies point to the unique qualities of religious identity and practice for recent generations. Flory and Miller (2008) connect the unique experiences of the 'postboomer' generation to their style of religiosity, one that incorporates a spiritual quest, often within a religious community. Their findings parallel Lytch's (2004) discovery of high levels of personal autonomy among youth heavily engaged in conventional Christian youth groups. They label the mix 'expressive communalism.'

This balance and integration of one's own individualistic search for meaning together with a deep desire to do so in the company of others are themes repeatedly underlying many examinations of the character of youth religiosity in a variety of countries with different religious landscapes (Day, 2010). The institutional locations of religious activity are not always conventional religious buildings. In a study of Scottish youth, skate parks, pubs, classrooms, and youth festivals all emerge as sites of revelation and transformation (Vincett et al., 2012). Some are established by adults looking to increase the relevance of religious practice to youth; others collectively emerge from the efforts of youth themselves desiring ways to experience their faith more authentically (Hopkins et al., 2011). Harris (2010) shows how both spiritual growth and social connection are achieved among British youth on pilgrimages to the Marian Shrine at Lourdes, France. Virtanen (2010) explains how contemporary shamanic practices in the Brazilian Amazon are a means of building and maintaining trusting relationships. And Cush (2010) shows how teenage witches and Pagans emphasize their own autonomy and search for meaning, all the while building self-esteem that helps them be proud of being 'different' and therefore making them more confident in belonging and interacting with others. When contemporary youth are religious, they emphasize faith practiced in a way that connects them to others and reinforces a sense of both belonging and meaning in their lives.

Religious and Spiritual Development and Trajectories in Youth

Another central focus of the study of youth and religion is its developmental aspect, or how religiosity and spirituality change with age regardless of the historical context. Much theory in this area focuses on the cognitive development taking place during youth and how that might correspond to how youth think about and engage in religious or spiritual matters. For years, research has suggested that religious thinking between ages 12 and 18 becomes less literal and more abstract (Elkind, 1971; Nelsen et al., 1977; Potvin et al., 1976). Hoge and Petrillo (1978) find that abstract religious thinking among adolescents is negatively associated with creedal assent and religious practice. The initiation of abstract thinking is one reason typically cited for why religious practices and strength of belief tend to decrease during adolescence and emerging adulthood.

An additional theoretical frame for the study of religious or spiritual development has been identity theory. Foundational research in this area by Marcia (1966) and Erikson (1968, 1969) stresses stages of establishing a mature religious identity. Findings in this area have been mixed because of incompatible definitions of mature religion and incomparable data (Hood et al., 2003). More recent work in this area has emphasized the interaction and hierarchy of a variety of identities, including religion, to understand how

adolescents create multifaceted identities and the extent to which religion is a salient part of that identity (Spencer et al., 1997; Vaisey, 2009). It is important that the study of religion in youth continues to draw on the most recent developments in identity theory at large to refine theory and to improve measurement (Hemming and Madge, 2012). As it becomes increasingly clear that a hierarchy of one's identities informs the consequences of each feature of identity, we should be measuring religious importance in a way that uses other social identities as reference points.

One way in which the management of multiple identities is being studied is through interview studies focused on how immigrant youth develop and integrate destination culture and expectations with their native beliefs, practices, and symbols, such as Sikh youth wearing turbans (Singh, 2010). Also, Clydesdale (2007) discusses how youth who have just finished high school in the United States manage their various identities in the midst of dramatic change and transition. He finds that they use an 'identity lockbox' in which their religiosity (among other aspects of identity) is stored away and not tinkered with given everything else that is happening.

It is commonly argued that disengagement from religion occurs most often between the late teens and early 20s (Hood et al., 2003). Factors often invoked to explain this disengagement include life transitions and cognitive-development issues (Schweitzer, 2000). Many religious institutions maintain a focus on bringing in members when they marry and/or have young children, but as the span of years between adolescence and family formation widens, more distance develops between youth and their childhood religious institutions. This all reinforces the extent to which the life course perspective is a highly useful guide for examining religious change and stability during youth. The life course perspective integrates consideration of developmental processes (i.e., trajectories) and changes (i.e., transitions) that may alter patterns of development with insight into how the social context of early adolescence and life transitions throughout childhood, adolescence, and young adulthood shape trajectories of religious involvement (Elder, 1998; Petts, 2009).

Multiple studies have suggested that the frequency of religious service attendance tends to decline between the 8th and 12th grades (Potvin et al., 1976; Benson et al., 1989; Roehlkepartain and Benson, 1993; Wallace et al., 2003). A longitudinal study of adolescents in Iowa found that while frequency of religious attendance dipped during the high school years, levels of participation in other church activities tended to rise over the same period (King et al., 1997). Smith and Snell (2009) demonstrate that the period of emerging adulthood following adolescence entails a great many transitions, disruptions, and distractions. They find that a little over one-half of emerging adults in the late 2000s in the United States remain fairly consistent in their strength of belief and practices (whether high, medium, or low). The other 40% experiences changes, mostly modest and decreasing.

One recent advance in the literature on change in dimensions of religiosity from childhood through emerging adulthood is the availability of more waves of data and the application of more sophisticated longitudinal analyses such as growth curve or trajectory models (Desmond et al., 2010; Smith and Snell, 2009). Using group-based trajectory modeling, Petts (2009) finds six common trajectories in religious participation from ages 10 to 25. The three stable trajectories are those who attend weekly all along (3%), those who attend occasionally all along (23%), and those who never attend (7%). The three patterns of change are those who experience gradual decline in attendance (26%), those who experience early decline (16%), and those who experience late decline (24%). Although about one-third of youth maintain a relatively stable level of religious participation, the other two-thirds experience a decline (Petts, 2009). As more data become available, increased application of these cutting-edge statistical methods will provide useful details on the rate and timing of change in different aspects of religiosity and spirituality across childhood, adolescence, and emerging adulthood.

While the bulk of studies of religious or spiritual change in youth focus on incremental change over time (the most common type of change), dramatic religious change, often in the form of conversion or total apostasy, has drawn some interest as well. It is agreed that the peak age of conversion falls somewhere between the ages of 13 and 16 (see Gillespie, 1991; Beit-Hallahmi, 1997; Hood et al., 2003). Regnerus and Uecker (2006) find that very few adolescents in the United States experience dramatic religious change, and that personality and behavioral factors are consistent predictors of rapid decrease in religious practice and importance, while demographic, family, and peer network effects are predictive of changes in religiosity among youth.

Little is known about religious belief, practices, and meaning in childhood (Beste, 2011). This is in part due to the challenges that come in collecting data directly from children (Levey, 2009). Yet a few studies have examined qualitative data on childhood socialization (Gilkes, 2009; Haight, 2002), primarily through observational methods and parent interviews. Bales (2005) used observational methods to study children's first communion experiences from their perspective, finding that children do think for themselves, assert agency, show resistance, and have religious perspectives that are distinct from those of adults. Coles (1991) used interviews with children of all different faiths around the world to reveal how school-aged children process and express their religious and spiritual beliefs and understandings. Beste (2011) showed how children actively co-construct meaning and reality as opposed to merely absorbing and internalizing the religious teachings of adults (Beste, 2011). Rather than passively going through the motions, many second graders value this sacrament because it provides a concrete way to feel forgiven by God and absolved of wrongdoing (Beste, 2011). Researchers have tended to ascribe agency more to adolescents experiencing a higher level of cognitive development, but clearly the study of religious development could further benefit from a focus on younger children.

Agents of Religious Socialization and Youth

A major line of research in the area of religion and youth is the importance of influential others in shaping one's religious beliefs and practices, otherwise known as religious

socialization theory (Sherkat, 2003). When it comes to the religious socialization of youth, we most often think of parents and peers. Studies repeatedly find parents to be the most important religious influence in the lives of youth (e.g., Cnaan et al., 2004; Pearce and Denton, 2011; Smith and Denton, 2005; Spilka et al., 1985). And there is considerable research identifying the impact of peers on religious commitments (Desmond et al., 2010; Gunnoe and Moore, 2002; Regnerus et al., 2004). Social institutions are also important for religious socialization, and considerable research has examined the importance of educational institutions and religious organizations for influencing religious commitments (Sherkat, 2003).

From childhood through emerging adulthood, one's religious affiliation is highly correlated with the tradition in which one was raised (Sherkat, 1998; Smith and Snell, 2009). Religious beliefs and practices are also closely associated with those that parents modeled and encouraged during childhood (Gunnoe and Moore, 2002; Ozorak, 1989; Smith and Denton, 2005; Wuthnow, 1998). When parents cultivate an understanding that the family attends religious services, youth tend to believe they should attend (Lytch, 2004). Having religious parents makes one less vulnerable to dramatic religious declines during youth (Regnerus and Uecker, 2006). Furthermore, any decline in youth religiosity happens later for youth whose parents provide consistent religious messages (Petts, 2009). Although both parents matter, studies suggest that mothers' beliefs and practices are more influential than fathers' (Acock and Bengtson, 1978; Benson and Eklin, 1990). Research suggests that religious socialization tends to be most successful when families are close and relationships are harmonious, when youth live with two parents, and when parents share the same religious affiliation, beliefs, and practices (Myers, 1996; Petts, 2009; Sherkat, 1991). Part of how religious parents influence their children is by encouraging them to choose more religious peers and be involved in religious institutions (Erickson, 1992; Sherkat, 2003). The influence of religious socialization from parents takes on special importance in the case of immigrant families, in which immigrant or second-generation youth must contend with multiple cultures and identities (Warner and Williams, 2010). Not surprisingly, socialization and religious influence do not just work in one direction. Children also influence their parents. Sherkat (1991) finds that there are reciprocal relationships between parent and child religious participation and belief from childhood through adulthood.

Peers also play a role in the socialization of religious commitments; since much of youths' lives are spent in educational or extracurricular activities segregated by age, their identities evolve in some measure according to what they perceive as acceptable to their peers. When it comes to peers and religion, having schoolmates or peers with higher levels of religious affiliation, belief, practice, and salience is, not surprisingly, related to higher levels of these dimensions of religiosity in youth (Desmond et al., 2010; Gunnoe and Moore, 2002; Regnerus et al., 2004). Also, having more religious schoolmates and peers is protective against dramatic religious decline and apostasy (Regnerus and Uecker, 2006). Pearce and Denton (2011) find that adolescent peers do not report talking to each other about religion much, so either less direct methods of socialization are occurring, or these associations may be spurious, explained by other characteristics that peers tend to have in common.

Schools

Studies have found that, independent of any influence that peers have on each other, the religious composition of one's whole school is related to an individual student's religious affiliation, beliefs, and practices and to changes in these over time (Barrett et al., 2007; Regnerus et al., 2004; Regnerus and Uecker, 2006). Other research has focused on parochial education and how such experiences shape youth religiosity and spirituality, and most studies find a positive effect of religious education on religious commitments. Ethnographic studies of fundamentalist Protestant schools find that a major objective is to shelter children from the secular world, to 'keep them out of the hands of Satan' (Peshkin, 1986). Much research has focused on Catholic schools, and these studies show that the number of years of Catholic schooling is positively related to a range of proreligion beliefs and practices (Greeley and McCready, 1976; Greeley and Rossi, 1966; Guerra et al., 1989; Uecker, 2008). Youth who enrolled in Protestant schools or were homeschooled had similar levels of involvement in their local congregation as youth who attended public schools (Uecker, 2008).

Another line of research considers how college experiences shape religiosity in emerging adulthood. Sociologists of religion have long linked educational attainment to religious decline (Sherkat, 1998). However, one recent study contends that emerging adults who do *not* attend college are the most likely to curtail religious commitments in early adulthood (Uecker et al., 2007).

Religious Institutions

A few studies have examined features of religious congregations across a variety of denominations to identify features that promote religious commitment (Benson and Donahue, 1990; Benson and Eklin, 1990; Kelly et al., 1986). Research finds that congregations that offer a sense of belonging, a sense of meaning based in religious principles, and opportunities to develop competencies in the transition to adulthood do best at retaining youth members (Lytch, 2004). Certain religious traditions motivate higher levels of religiosity from both youth and emerging adults. Smith and Denton (2005) find youth who identify as Latter-Day Saints report the highest levels of religiosity across dimensions, followed by conservative and black Protestants, mainline Protestants, and Catholics, with Jewish and unaffiliated teens at the lowest levels of all dimensions of religiosity.

How Personal Characteristics and Agency Relate to Religiosity

Although a great deal of attention has been paid to theorized causes of religiosity and religious change in the lives of youth, increasingly research considers the extent to which children,

adolescents, and emerging adults exercise agency in their religious identity and actions (Beste, 2011). Youth often negotiate, resist, engage with, modify, and/or reject religious ideas and practices presented to them (Hopkins et al., 2011). Some youth actively challenge the ways in which their parents had or had not tried to transmit certain views or perspectives about religion (Hopkins et al., 2011). This can result in conflict and disagreement among parents and youth, affecting parent–child relationship quality (Pearce and Axinn, 1998).

There is evidence that certain personality characteristics are associated with different levels of religious belief and practice. Risk-aversive people report higher levels of religiosity than those who like to take risks (Miller and Hoffmann, 1995). Miller and Stark (2002) suggest that risk proclivity alone may account for gender difference in religiousness. Other personality features, such as having a strategic or planful orientation, may be at work as well. Regnerus and Smith (2005) found that strategic adolescents are significantly more likely to attend religious services frequently and report religion as being very important. Studies of genetic and birth order influences on religiousness are illuminating and invite further the study of genetic roles in religious and spiritual identity and practice (Heath et al., 1999; Saroglou and Fiasse, 2003; Bradshaw and Ellison, 2008).

Another way in which youth may exercise agency is to choose to limit or dissolve their religious identity when other identities or preferences conflict with what their religious faith prescribes. In the case of youth, it is often premarital sex, drinking alcohol, or substance use that becomes at odds with their religious beliefs and practices (Benda and Corwyn, 1997; Regnerus and Uecker, 2006; Thornton and Camburn, 1989; Thornton et al., 1992).

The Effects of Religion

Another common approach in the social scientific study of religion is to study ways in which religion may or may not be healthy or protective in the lives of young people. Research has focused on prosocial behaviors, subjective well-being, and risk behaviors in the lives of children, adolescents, and emerging adults.

Parent religious affiliation, beliefs, practice, and salience have all been linked to variance in parenting styles and methods of discipline (Bartkowski, 1995; Ellison and Sherkat, 1993; Wilcox, 2004), and therefore child outcomes (Bartkowski et al., 2008). When parents share the same religious affiliation, children have higher levels of behavioral, emotional, and cognitive development, but when parents argue about religious matters, children's developmental outcomes are lower (Bartkowski et al., 2008). Some studies have focused on how religiosity is related to the internalization and externalization of problem behaviors in children (Kim, 2008; Petts, 2011).

In studies of adolescents, their own religious beliefs, practice, and salience are often shown to be positively related to overall subjective well-being (Benson and Donahue, 1995; Markstrom, 1999), positive life attitudes, satisfaction, and hope for the future (Francis and Kay, 1995; Smith and Faris, 2002a,b), sexual behavior (Regnerus, 2007), as well as a host of other normative outcomes. Religion may provide a variety of developmental assets that lead to these positive associations, such as social support from family, friends, and nonrelative adults; empowerment; clear boundaries and expectations; having constructive uses of time; commitment to learning; positive values; social competencies; and positive identity, which likely mediate the influence of religion (Benson et al., 1997; Benson et al., 2012; Roehlkepartain, 1998; Scales et al., 2003; Smith, 2003; Wagener et al., 2003).

Of course, the positive effects of religious institutions depend on an assumption that religious communities are composed of positive interactions and role models. This is not always the case. Evidence suggests that religious adults and institutions often socialize girls to limit their educational and career aspirations and planning in anticipation of wife and mother roles that may or may not come to fruition or take up as much of the life course as expected (Pearce and Halliday, 2012). Heynes (2010) demonstrates that it is predominantly male heroes that are represented in religious education materials and that girls notice. Therefore, the benefits theorized to accrue from religious institutions may advantage boys more than girls.

A Note on Race and Ethnicity, Class, and Gender

There are three other social statuses or identities that rival, if not surpass, the salience of religion in the lives of youth: race or ethnic identity, social class, and gender identity. Most of the survey or quantitative research cited here looks at differences across these groups or at least 'controls' for them in regression analyses. Usual findings are that minority groups tend to exhibit higher levels of religiosity across a variety of dimensions, that there are class differences in denominational affiliation but few discernable patterns in practice or belief, and that girls tend to be more religious across dimensions than boys. These are all differences noted in adult studies as well, so they are not particularly insightful findings for the unique experience of religion or spirituality in youth. Studies involving more ethnographic methods such as observations and/or interviews are able to focus on certain subgroups within these statuses, delving deeper to provide rich detail on experiences, but few existing studies of youth and religion tackle the intersections of these identities (Hemming and Madge, 2012). As our theories expand and develop to consider the complexity of identities and how they are managed, our research questions and methods should adapt to collect and analyze empirical data that can speak to the ideas.

Future Directions

While the study of youth and religiosity has flourished in the United States and parts of Western Europe, the field is weakest in studying the religious lives of youth in non-Western countries where Christianity is not the dominant religion. It is important that our theories and methods be applied in new locales to (1) expand our knowledge to new settings, and (2) further realize the limitations and ethnocentricity of our current approaches. A great strength of the field of youth and

religion studies is the diversity of methods being applied and the growing number of studies that mix methods collecting both qualitative and quantitative data. This leads to a balance of broad versus deep research, showing us patterns in religious and spiritual beliefs, practices, salience, and experience across the globe as well as rich ethnographies opening our eyes to specific rituals, unique subgroups of the youth population, and distinctive religious institutions. Going forward, more cross-fertilization of the findings from disparate methods (whether in the same project or not) will advance the field even further. For example, we could use more ethnographic studies of youth to unpack the mechanisms through which religion either protects or limits youth well-being. To date, this stream of the research has been heavily tilted toward the use of quantitative data and analyses.

Finally, one of the most recent and potentially significant lines of research is that which illuminates the agency of youth in their religious and spiritual lives. For too long, the field has assumed a passive stance of youth, and we have much to learn by (1) continuing to develop and use methods to observe and interview children regarding their agency in religious and spiritual development, and (2) employing research methods, such as taking a phenomenological approach or employing statistical methods that carefully account for reciprocal relationships between factors typically thought to mostly influence youth rather than the other way around.

See also: Adolescence, Sociology of; Adolescent Health and Health Behaviors; Delinquency, Sociology of; Religion and Health: Sociological Considerations; Religion, Family, and Kinship; Religion, Sociology of; Religiosity; Socialization in Infancy and Childhood.

Bibliography

Acock, A., Bengtson, V., 1978. Relative influence of mothers and fathers – covariance analysis of political and religious socialization. Journal of Marriage and the Family 40 (3), 519–530.

Bachman, J., Johnston, L.D., O'Malley, P.M., 2000. Monitoring the Future. University of Michigan Institute for Social Research, Ann Arbor, MI.

Bales, S.R., 2005. When I Was a Child: Children's Interpretations of First Communion. University of North Carolina Press, Chapel Hill, NC.

Barrett, J.B., Pearson, J., Muller, C., Frank, K., 2007. Adolescent religiosity and school contexts. Social Science Quarterly 88 (4), 1024–1037.

Bartkowski, J.P., 1995. Spare the rod..., or spare the child? Divergent perspectives on conservative Protestant child discipline. Review of Religious Research 37 (2), 97–116.

Bartkowski, J.P., Xu, X., Levin, M.L., 2008. Religion and child development: evidence from the early childhood longitudinal study. Social Science Research 37 (1), 18–36.

Beit-Hallahmi, B., 1997. The Psychology of Religious Behaviour, Belief and Experience. Routledge, London, New York.

Benda, B., Corwyn, R., 1997. A test of a model with reciprocal effects between religiosity and various forms of delinquency using 2-stage least squares regression. Journal of Social Service Research 22 (3), 27–52.

Benson, P.L., Donahue, M.J., 1990. Valuegenesis, North American Division Office of Education. Seventh-Day Adventist Church, Silver Spring, MD.

Benson, P.L., Donahue, M.J., 1995. Religion and the well-being of adolescents. The Journal of Social Issues 51 (2), 145–160.

Benson, P.L., Eklin, C.H., 1990. Summary Report: The National Study of Mainline Protestant Denominations. Search Institute, Minneapolis, MN.

Benson, P.L., Donahue, M.J., Erickson, J.A., 1989. Adolescence and religion: a review of the literature from 1970 to 1986. Research in the Social Scientific Study of Religion 1, 153–181.

Benson, P.L., Masters, K.S., Larson, D.B., 1997. Religious influences on child and adolescent development. In: Alessi, N.E. (Ed.), Handbook of Child and Adolescent Psychiatry. Wiley, New York, pp. 206–219.

Benson, P.L., Leffert, N., Scales, P.C., Blyth, D.A., 2012. Beyond the "village" rhetoric: creating healthy communities for children and adolescents. Applied Developmental Science 16 (1), 3–23. Reprinted from: Applied Developmental Science 2, 138, (1998).

Beste, J., 2011. Children speak: catholic second graders' agency and experiences in the sacrament of reconciliation. Sociology of Religion 72 (3), 327–350.

Bradshaw, M., Ellison, C.G., 2008. Do genetic factors influence religious life? Findings from a behavior genetic analysis of twin siblings. Journal for the Scientific Study of Religion 47 (4), 529–544.

Clydesdale, T., 2007. The First Year Out: Understanding American Teens after High School. The University of Chicago Press, Chicago, IL.

Cnaan, R., Gelles, R., Sinha, J., 2004. Youth and religion: the gameboy generation goes to "church". Social Indicators Research 68 (2), 175–200.

Coles, R., 1991. The Spiritual Life of Children. A Peter Davison Book/Houghton Mifflin Company, Boston, MA.

Crockett, A., Voas, D., 2006. Generations of decline: religious change in twentieth-century Britain. Journal for the Scientific Study of Religion 45 (4), 567–584.

Cush, D., 2010. Teenage witchcraft in Britain. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 81–88.

Day, A., 2010. Believing in belonging: an exploration of young people's social contexts and constructions of belief. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 97–104.

Denton, M.L., Pearce, L.D., Smith, C., 2008. Religion and Spirituality On the Path through Adolescence, Research Report Number 8, National Study of Youth and Religion. University of North Carolina, Chapel Hill, NC.

Desmond, S.A., Morgan, K.H., Morgan, K.H., Kikuchi, G., 2010. Religious development: how (and why) does religiosity change from adolescence to young adulthood? Sociological Perspectives 53 (2), 247–270.

Elder Jr., G.H., 1998. Life course theory and human development. Sociological Analysis 1 (2), 1–12.

Elkind, D., 1971. The development of religious understanding in children and adolescents. In: Strommen, M.P. (Ed.), Research on Religious Development: A Comprehensive Handbook. Hawthorne, New York.

Ellison, C.G., Sherkat, D.E., 1993. Obedience and autonomy: religion and parental values reconsidered. Journal for the Scientific Study of Religion 32 (4), 313–329.

Erickson, J.A., 1992. Adolescent religious development and commitment: a structural equation model of the role of family, peer group, and educational influences. Journal for the Scientific Study of Religion 31, 131–152.

Erikson, E.H., 1968. Identity: Youth and Crisis. W.W. Norton & Company, Inc., New York, NY.

Erikson, E.H., 1902–1994, 1969. Gandhi's Truth: On the Origins of Militant Nonviolence. Norton, New York.

Flory, R., Miller, D.E., 2008. Finding Faith: The Spiritual Quest of the Post-boomer Generation. Rutgers University Press, Piscataway, NJ.

Francis, L.J., Kay, W.K., 1995. Teenage Religion and Values. Gracewing, Leominster, Herefordshire, England, UK.

Furrow, J.L., King, P.E., White, K., 2004. Religion and positive youth development: identity, meaning, and prosocial concerns. Applied Developmental Science 8 (1), 17–26.

Geertz, C., 1993. Religion as a cultural system. In: The Interpretation of Cultures: Selected Essays. Fontana Press, pp. 87–125.

Gilkes, C.T., 2009. The Black church and children. In: Browning, Don S., Miller-McLemore, Bonnie J. (Eds.), Children and Childhood in American Religions. Rutgers University Press, New Brunswick, NJ, pp. 85–102.

Gillespie, V.B., 1991. The Dynamics of Religious Conversion. Religious Education Press, Birmingham, Ala.

Greeley, A.M., McCready, W.C., McCourt, K., 1928–1975, 1976. Catholic Schools in a Declining Church. Sheed and Ward, Kansas City, MO.

Greeley, A.M., Rossi, P.H., 1928–1966, 1966. The Education of Catholic Americans. AldinePub. Co, Chicago.

Guerra, M., Benson, P.L., Donahue, M., 1989. The Heart of the Matter: Effects of Catholic High Schools on Student Values, Beliefs, and Behaviors. Catholic Educational Association, Washington, DC.

Gunnoe, M., Moore, K., 2002. Predictors of religiosity among youth aged 17–22: a longitudinal study of the national survey of children. Journal for the Scientific Study of Religion 41 (4), 613–622.
Haight, W.L., 1958–2002. African-American Children at Church: A Sociocultural Perspective. Cambridge University Press, Cambridge.
Harris, A., 2010. A place to grow spiritually and socially: the experiences of young pilgrims to Lourdes. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 139–148.
Heath, A.C., Madden, P.A., Grant, J.D., McLaughlin, T.L., Todorov, A.A., Bucholz, K.K., 1999. Resiliency factors protecting against teenage alchohol use and smoking: influences of religion, religious involvement and values, and ethnicity in the Missouri Adolescent Female Twin Study. Twin Research 2, 145–155.
Hemming, P.J., Madge, N., 2012. Researching children, youth and religion: identity, complexity and agency. Childhood 19 (1), 38–51.
Heynes, J., 2010. Engaging with teenage girls' understandings of religion and gender. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 123–130.
Hill, P.C., Pargament, K.I., 2003. Advances in the conceptualization and measurement of religion and spirituality: implications for physical and mental health research. American Psychologist 58 (1), 64–74.
Hoge, D.R., Petrillo, G.H., 1978. Determinants of church participation and attitudes among high school youth. Journal for the Scientific Study of Religion 17 (4), 359–379.
Hood Jr., R.W., Hill, P.C., Spilka, B., 2003. The Psychology of Religion: An Empirical Approach. Guilford Press, New York.
Hopkins, P., Olson, E., Pain, R., Vincett, G., 2011. Mapping intergenerationalities: the formation of youthful religiosities. Transactions of the Institute of British Geographers 36 (2), 314–327.
Kelly, F.D., Benson, Peter L., Donahue, M.J., 1986. Toward Effective Parish Religious Education for Children and Young People: A National Study. National Catholic Educational Association, Washington, DC.
Kim, J., 2008. The protective effects of religiosity on maladjustment among maltreated and nonmaltreated children. Child Abuse & Neglect 32 (7), 711–720.
King, V., Elder Jr., G.H., Whitbeck, L.B., 1997. Religious involvement among rural youth: an ecological and life-course perspective. Journal of Research on Adolescence 7 (4), 431–456.
Levey, H., 2009. "Which one is yours?": children and ethnography. Qualitative Sociology 32 (3), 311–331.
Lytch, C., 2004. Choosing Church: What Makes a Difference for Teens. John Knox Press, Westminster.
Marcia, J.E., 1966. Development and validation of ego-identity status. Journal of Personality and Social Psychology 3 (5), 551–558.
Markstrom, C., 1999. Religious involvement and adolescent psychosocial development. Journal of Adolescence 22 (2), 205–221.
Mason, M.A., Singleton, A., Webber, R., 2007. The spirituality of young Australians. International Journal of Children's Spirituality 12 (2), 149–163.
Miller, A.S., Hoffmann, J.P., 1995. Risk and religion: an explanation of gender differences in religiosity. Journal for the Scientific Study of Religion 34, 63–75.
Miller, A.S., Stark, R., 2002. Gender and religiousness: can socialization explanations be saved? American Journal of Sociology 107, 1399–1423.
Myers, S.M., 1996. An interactive model of religiosity inheritance: the importance of family context. American Sociological Review 61 (5), 858–866.
Nelsen, H.M., Potvin, Raymond H., Shields, Joseph, 1977. The Religion of Children. United States Catholic Conference, Washington, DC.
Ozorak, E.W., 1989. Social and cognitive influences on the development of religious beliefs and commitment in adolescence. Journal for the Scientific Study of Religion 28 (4), 448–463.
Pearce, L.D., Axinn, W.G., 1998. The impact of family religious life on the quality of mother-child relations. American Sociological Review 63 (6), 810–828.
Pearce, L.D., Denton, M.L., 2011. A Faith of Their Own: Stability and Change in the Religiosity of America's Adolescents. Oxford University Press, New York.
Pearce, L.D., Halliday, J.H., 2012. Religion's role in the development of girls' occupational aspirations. In: Trommsdorff, G., Chen, X. (Eds.), Values, Religion, Culture, and Adolescent Development. Cambridge University Press, Cambridge, UK, pp. 271–289.
Peshkin, A., 1986. God's choice: the total world of a fundamentalist christian school. Educational Leadership 43 (4), 36–41.
Petts, R.J., 2009. Trajectories of religious participation from adolescence to young adulthood. Journal for the Scientific Study of Religion 48 (3), 552–571.
Petts, R.J., 2011. Parental religiosity, religious homogamy, and young children's well-being. Sociology of Religion.
Potvin, R.H., 1976. Religion and American Youth, with Emphasis on Catholic Adolescents and Young Adults. Publications Office, The Conference, Washington.
Regnerus, M.D., 2007. Forbidden Fruit: Sex and Religion in the Lives of American Teenagers. Oxford University Press, New York.
Regnerus, M.D., Smith, C., 2005. Selection effects in studies of religious influence. Review of Religious Research 47 (1), 23–50.
Regnerus, M.D., Uecker, J.E., 2006. Finding faith, losing faith: the prevalence and context of religious transformations during adolescence. Review of Religious Research 47 (3), 217–237.
Regnerus, M.D., Smith, C., Smith, B., 2004. Social context in the development of adolescent religiosity. Applied Developmental Science 8 (1), 27–38.
Roehlkepartain, E.C., 1998. Building Assets in Congregations: A Practical Guide for Helping Youth Grow Up Healthy. Search Institute, Minneapolis, MN.
Roehlkepartain, E.C., Benson, P.L., 1993. Youth in Protestant Churches: A Special Search Institute Report. Search Institute, Minneapolis, MN.
Roof, W.C., 2003. Religion and spirituality: toward an integrated analysis. In: Dillon, M. (Ed.), Handbook of the Sociology of Religion. Cambridge University Press, Cambridge, UK, pp. 137–151.
Saroglou, V., Fiasse, L., 2003. Birth order, personality, and religion: a study among young adults from a three-sibling family. Personality and Individual Differences 35 (1), 19–29.
Scales, P.C., Peter Benson, L., Mannes, M., 2003. Grading Grown-Ups 2002: How Do American Kids and Adults Relate? A National Study. Search Institute, Minneapolis, MN.
Schweitzer, F., 2000. Religious affiliation and disaffiliation in late adolescence and early adulthood: the impact of a neglected period of life. In: Francis, L.J., Katz, Y.J. (Eds.), Joining and Leaving Religion: Research Perspectives. Gracewing, Leominster, Herefordshire, England, pp. 87–101.
Sherkat, D.E., 1991. Leaving the faith: testing theories of religious switching using survival models. Social Science Research 20 (2), 171–187.
Sherkat, D.E., 1998. Counterculture or continuity? Competing influences on baby-boomers' religious orientations and participation. Social Forces 76 (3), 1087–1114.
Sherkat, D.E., 2003. Religious socialization: agents of influence and influences of agency. In: Dillon, M. (Ed.), Handbook for the Sociology of Religion. Cambridge University Press, Cambridge, UK, pp. 151–163.
Singh, J., 2010. British Sikh youth: identity, hair and the turban. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 131–138.
Smith, C., 2003. Theorizing religious effects among American adolescents. Journal for the Scientific Study of Religion 42 (1), 17–30.
Smith, C., Denton, M.L., 2005. Soul Searching: The Religious and Spiritual Lives of American Teenagers. Oxford University Press, New York.
Smith, C., Faris, R., 2002a. Religion and American Adolescents Delinquency Risk Factors and Constructive Social Activities. The National Study of Youth and Religion, Chapel Hill, NC.
Smith, C., Faris, R., 2002b. Religion and Life Attitude and Self-images of American Adolescents. The National Study of Youth and Religion, Chapel Hill, NC.
Smith, C., Snell, P., 2009. Souls in Transition: The Religious Lives of Emerging Adults in America. Oxford University Press, New York.
Spencer, M., Dupree, D., Hartmann, T., 1997. A phenomenological variant of ecological systems theory (PVEST): a self-organization perspective in context. Development and Psychopathology 9 (4), 817–833.
Thornton, A., Camburn, D., 1989. Religious participation and adolescent sexual behavior and attitudes. Journal of Marriage and Family 51 (3), 641–653.
Thornton, A., Axinn, W.G., Hill, D.H., 1992. Reciprocal effects of religiosity, cohabitation, and marriage. American Journal of Sociology 98 (3), 628–651.
Uecker, J.E., 2008. Alternative schooling strategies and the religious lives of American adolescents. Journal for the Scientific Study of Religion 47 (4), 563–584.
Uecker, J.E., Regnerus, M.D., Vaaler, M.L., 2007. Losing my religion: the social sources of religious decline in early adulthood. Social Forces 85 (4), 1667–1692.
Vaisey, S., 2009. Motivation and justification: a dual-process model of culture in action. American Journal of Sociology 114 (6), 1675–1715.
Verbit, M.F., 1970. The components and dimensions of religious behaviour: toward a reconceptualization of religiosity. In: Hammond, P.E., Johnson, B. (Eds.), American Mosaic: Social Patterns of Religion in the United States. Random House, New York, pp. 24–39.
Vincett, G., Olson, E., Hopkins, P., Pain, R., 2012. Young people and performance christianity in Scotland. Journal of Contemporary Religion 27 (2), 275–290.

Virtanen, P.K., 2010. Shamanic practices and social capital among native youths in the Brazilian Amazon. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 73–80.

Wagener, L., Furrow, J., King, P., Leffert, N., Benson, P., 2003. Religious involvement and developmental resources in youth. Review of Religious Research 44 (3), 271–284.

Wallace, J., Forman, T., Caldwell, C., Willis, D., 2003. Religion and US secondary school students - current patterns, recent trends, and sociodemographic correlates. Youth & Society 35 (1), 98–125.

Warner, R.S., Williams, R.H., 2010. The role of families and religious institutions in transmitting faith among Christians, Muslims, and Hindus in the USA. In: Collins-Mayo, S., Dandelion, P. (Eds.), Religion and Youth. Ashgate, Surrey, UK, pp. 159–166.

Wilcox, W.B., 2004. Soft Patriarchs, New Men: How Christianity Shapes Fathers and Husbands. University of Chicago Press, Chicago, UK.

Wuthnow, R., 1998. After Heaven: Spirituality in America Since the 1950s. University of California Press, Berkeley and Los Angeles.

Relevant Websites

http://www.thearda.com/ – The Association of Religion Data Archives.

http://youthandreligion.org – The National Study of Youth and Religion.

http://www.uic.edu/depts/soci/yrp/comp/pages/home.html – Youth and Religion Project.

Religion, Anthropology of

Simon Coleman, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the history of the anthropology of religion from the nineteenth century to the present. It argues that a focus on such questions as rationality and ritual was central to the emergence of the discipline. These themes, along with topics such as witchcraft, belief, language, and the body, have remained of perennial interest. More recently, focus has also been placed on the anthropologies of world religions such as Islam, Christianity, Hinduism, and Buddhism; on religion in relation to globalization and diaspora; and on cognitive approaches to the workings of the human mind.

Emergence of the Subfield

The comparative study of religion formed a central building block of anthropology as the discipline emerged in the nineteenth century and early twentieth century. In the light of social evolutionary models of human development, religious practice was perceived as providing a powerful index of the mental and moral levels of so-called primitive peoples. James Frazer's *The Golden Bough*, first published in 1890, traced magical and religious threads throughout history and weaved them into a pattern depicting the past and future progress of humanity, claiming to discern shifts from magical manipulation toward religious devotion and then ultimately in the direction of purely scientific modes of engaging the world. Inherent in Frazer's work was also a juxtaposition that has reemerged, albeit in very different form, in contemporary writings (e.g., Cannell, 2006): Christianity as an object of study but also a mode of thought that has itself framed anthropological understandings of religion, temporality, and culture.

The use of religion as a key site for the examination of human rationality permeates E.E. Evans-Pritchard's *Witchcraft, Oracles and Magic among the Azande* (1937). In contrast to Frazer's evolutionary assumptions and reliance on scattered examples drawn from around the world, Evans-Pritchard focuses on a single African case study, showing interrelations among religious, social, and political aspects of Azande life. The book assesses the logic and consistency of Azande modes of thought, and indicates how they might be translated into the understandings of a Western readership. Evans-Pritchard's implicit interlocutor was the French ethnologist and philosopher, Lucien Lévy-Bruhl, who had highlighted what he saw as the 'primitive' mind's inability to distinguish the supernatural from reality. While refuting Lévy-Bruhl's view, Evans-Pritchard's book also implies a comparison with Christianity: his discussion of Azande explanations of misfortune echoes biblical scholars' attempts to explain the presence and direction of evil in the world.

A further foundational strand in the anthropological study of religion has been the investigation of the relationship between religion and social order. In such work, the discipline's Durkheimian inheritance has come to the fore, in particular the view of ritual as expression and promoter of societal unity, alongside the more general assumption that religious ideas provide the key to socially shared categories of understanding.

Reflecting on the history of the anthropology of religion, Michael Lambek (2002: p. 4) characterizes contemporary research as drawing on a number of early sources: Franz Boas's tracing of connections between religion and language; Émile Durkheim's emphasis on the importance of the social; Karl Marx's pointing to forms of alienation, mystification, and power; and Max Weber's analysis of the place of religion in transitions to modernity.

Definitions

Attempts to produce a sustainable, universal definition of religion have prompted much debate. Not all scholars believe that a definition is possible. Saler (2000: p. ix) asserts that "Religion is a Western folk category that contemporary Western scholars have appropriated." A similarly skeptical view is maintained by Maurice Bloch, who notes (2010: p. 4): "'Religion' is a word that can only refer to a series of historically created situations which, although continually changing, have unique and specific genealogies closely linked to the Abrahamic religions… In other words religion is not a natural kind, by which I mean a category that has a basis other than that given by an arbitrary definition."

The earliest influential attempt at a definition was provided by British ethnologist Edward Tylor, himself born into a Quaker family. In *Primitive Culture*, published in1871, Tylor summed up religion as "belief in spiritual things." His characterization is terse in comparison with Durkheim's assertion in his book *The Elementary Forms of Religious Life*, published in 1912, that a religion can be seen as "a unified set of beliefs and practices relative to sacred things, that is to say, things set apart and forbidden – beliefs and practices which unite [into] one single moral community called a Church, all those who adhere to them" (Durkheim, 1912: p. 44). We see here some characteristic Durkheimian juxtapositions, such as the sacred and nonsacred, and belief balanced by practice; but also the claim that religion must be seen as linked to a social formation, a 'Church.' This definition of religion as a certain kind of object of study therefore also points toward a specific method of study: the focus is on what can be physically observed, and eschews assessment of the truth value or otherwise of any given religion.

Durkheim's depiction of social order proved highly influential for the British anthropologist Alfred Radcliffe-Brown,

who in the middle years of the century drew on such ideas in developing the notion of 'structural-functionalism' – a view of society as made up of, and stabilized by, interlocking and complementary components, including religious institutions. The emphasis on observable reality also informed Bronislaw Malinowski's stress not on the evolution of religion, but on the importance of fieldwork in discerning the contemporary societal and psychological rationale behind ritual and magic. At the same time, a Durkheimian approach still raised questions as to (1) the worth of assuming that rigid distinctions between sacred and profane existed cross-culturally, (2) the focus on order and stability as a feature of social institutions such as religion, and (3) the universality of any characterization of religion.

In the 1970s, the American anthropologist Clifford Geertz combined a Durkheimian understanding of religion as a collective social act with a more Weberian emphasis on meaning and experience. For Geertz (1973: p. 4), religion was "(1) a system of symbols which acts to (2) establish powerful, pervasive, and long-lasting moods and motivations in men [sic] by (3) formulating conceptions of a general order of existence and (4) clothing these conceptions with such an aura of factuality that (5) the moods and motivations seem uniquely realistic." Geertz's humanistic, 'interpretive' anthropology created intellectual distance from such figures as Claude Lévi-Strauss, whose French 'structuralism' focused on very broad cultural meanings divorced from the actions and interpretations of specific people.

A later period of anthropology would express worries over the cross-cultural validity but also the intellectual and cultural politics of the very act of making definitions. The best-known contributor to such debate has been Talal Asad (1993). Brought up in Pakistan, Asad spent some time working with Evans-Pritchard at Oxford before eventually moving to the United States. Influenced by Michel Foucault and Edward Said, he is acutely aware of the power relations involved in representing religion, and focuses on the need to trace 'genealogies' of both religious and secular, Western and non-Western, ways of viewing the world. For Asad, Geertz's (1973: p. 29) definition runs into problems in its attempt to capture a panhuman phenomenon divorced from particular cultural, social, and political contexts. In his view, the very act of defining must be seen as the historical product of ideologically charged, discursive processes. Similarly, the notion of religion as an autonomous activity is regarded as emerging from a unique, Western, post-Reformation history.

What, then, might be the solution to such dilemmas? Bloch (2010) turns to ritual rather than religion in articulating his comparative approach, arguing that the former can be found in all types of society, and is a specific type of modification of the way human beings in general communicate. Saler (2000: p. x) takes a more pragmatic approach, drawing in part on Wittgenstein's discussions of 'family resemblances' to argue that the different instantiations of what is called religion need not all share one feature, or a specific conjunction of features, but may possess certain overlapping similarities. In the following, I adopt another pragmatic, inductive approach, taking the anthropology of religion to be what scholars actually do, no matter what definition of the subject they endorse.

Enduring Themes

The following highlights some of the most prominent themes in the anthropology of religion that have remained topics of interest throughout the discipline's history.

Belief

Asad's skepticism as to the utility of belief as an analytical term is not unique. Malcolm Ruel points to "the monumental peculiarity" (1982: p. 100) and historical instability of Christian notions of belief. He notes that both the original Greek verb *pisteuo* and the Hebrew root *mn* express ideas of trust or confidence in an agreement, indicating a fundamentally social orientation. The noun *pistis* then acquires a special twist in the apostolic writings of the New Testament, where it is often used in the sense of being converted, denoting the 'belief' held collectively by early Christians as a common conviction in the resurrection of Jesus. Subsequently, the Protestant Reformation hinges on a stress on the inward totality of Christian belief. However, it is difficult to find equivalents to Protestant notions of belief in Judaism, Islam, Buddhism, or Hinduism.

Similar worries are taken up by Rodney Needham in *Belief, Language and Experience* (1972). Needham discusses the problems involved in discussing the statement "I believe in God" with one of his Indonesian informants, who belongs to an ethnic group called the Penan. As Needham puts the problem: "The Penan had no formal creed, and…they had no other conventional means for expressing belief in their God. Nevertheless, I had been accustomed to say…that they believed in a supreme god….Yet it suddenly appeared that I had no linguistic evidence at all to this effect" (Needham, 1972: pp. 1, 2). One of the issues that Needham touches on is crucial for much comparative ethnography: whereas Western Christianity is premised on the possibility of opting out of religious institutions or indeed denying belief, the kind of religion described by Evans-Pritchard for the Azande or Needham for the Penan is not a matter of choice for informants: it is an integral part of life.

What, however, of contexts where informants themselves maintain a strong and self-conscious sense of 'belief'? Joel Robbins (2004, 2007) examines the conversion to Pentecostal Christianity of the Urapmin of Papua New Guinea, and shows how even the apparent assimilation of Christian categories can conceal a more complex relationship to belief than might at first appear. The pigeon word 'bilip' is widely used by Urapmin, but it means something specific. For the Urapmin, Christian belief is not about mentally assenting to a set of propositions about divinity, but rather a form of trusting God to do what He promised.

Ritual

Anthropologists have often emphasized the importance of taking into account embodied, ritual activity. Alexander Henn (2008) notes that while Émile Durkheim emphasized the need to explore functional aspects of ritual in the making of human sociality and institutions, Clifford Geertz encouraged scholars to interpret rituals as akin to texts that could be 'read' by informants and analysts, and Victor Turner theorized ritual as

a dialectical process, moving between structure and anti-structure, secular and sacred. These approaches Henn characterizes as "normative," "intelligible" and "dialectic," respectively (Alexander Henn, 2008: p. 11). They share a tendency to emphasize the capacity of ritual to generate feelings of certainty and continuity. Roy Rappaport has also produced an influential discussion of ritual that sees it as primarily as "the performance of more or less invariant sequences of formal acts" (1999: p. 24). Yet, questions of how informants' interpretations relate to those of the analyst remain, including debates over how intentionality is linked to ritual.

Many analysts are still influenced by the work of a contemporary of Durkheim, Arnold Van Gennep. Van Gennep argued that ritual has particular significance during critical periods of transition in the life-cycle, such as attainment of adulthood, marriage, and death. In his view, it helps the participant to adjust to his or her change in role while also publicly announcing such a change. Furthermore, such rites follow a strikingly standard pattern across cultures, with the person's separation from everyday society followed by transition and then incorporation back into the social group, now bearing a changed social status. Turner finds particular inspiration in considering the 'liminal,' threshold-like section of rites of passage, arguing that they represent reversals of everyday structure and draw on powerful ritual symbolism, such as that of death, to symbolize the end of the old status. He notes that in Ndembu initiation rites in Zambia, circumcision of boys becomes a metaphor for killing, since it destroys the childhood status of the initiate (Turner, 1967).

Bloch (1992) is also interested in the symbolic and literal violence involved in many rites of passage, but applies a broadly Marxist frame of interpretation, highlighting the coercive powers of ritual. He agrees with Van Gennep concerning the existence of a basic grammar underlying ritual across cultures, and suggests that an irreducible core of the ritual process invokes a violent conquest of the present world by the transcendental, divine realm. Separation rites and rites of incorporation expunge the initiate's original vitality; but what Bloch calls the rebounding violence of transcendental conquest is contained in the return to everyday life, as the person is transformed from being prey into being a hunter – into a state where the powerful transcendent element inherited from the ritual dominates the person's identity. For instance, in the Orokaiva ritual involving the initiation of children in Papua New Guinea, participants are told they are dead, and are then taught to play sacred flutes and bullroarers that represent the voices of the spirits. After a time of seclusion, they return to the village, shouting "bite, bite, bite." They are now transformed beings.

Such analysis emphasizes the political implications of ritual experience in the service of order and hierarchy. More recently, in a number of influential books on the Tshidi of South Africa, Jean Comaroff and John Comaroff (1985) have analyzed ritual that involves complex forms of resistance quite as much as submission. Examining interactions between 'traditional' and Christian missionary beliefs and practices in colonial and postcolonial contexts, they show how Zionist churches have provided Tshidi with forms of identity, realized in part through new Christian idioms, which have allowed them to adapt or subvert structures of colonial and capitalist authority.

Anthropologists have worked more and more in urban and Western contexts often associated with deritualization and secularization, and with decreased need for common rites of passage to define accession to social roles. Nonetheless, the transformation and pluralization of ritual forms does not necessarily mean that they are disappearing. Tomas Gerholm's (1988) 'postmodern' view of a Hindu funeral ritual in Trinidad examines fragmentation of meaning and diversity of experience. His piece describes a ritualized tribute paid by the writer V.S. Naipaul to his deceased sister – an improvised response taking place thousands of miles away from the official funeral in the writer's country of origin, and yet still capable of carrying significance. If Gerholm suggests that it is often difficult to locate the core of ritual practice in postmodern contexts, Catherine Bell's highly influential approach in *Ritual Theory, Ritual Practice* (1992) moves the focus away from seeing ritual as an autonomous activity, and proposes instead a notion of 'ritualization' that takes into account links with strategic activities in social life in general.

Rationality

Frazer's radical move in *The Golden Bough* was ultimately to juxtapose Christianity with science, and to suggest the incompatibility of the former with the latter. More recent work has taken the opposite step of proposing that Western science has significant parallels with 'traditional' religious thinking. Robin Horton's "African Traditional Thought and Western Science" (1970) argues that African religions link causes and effects in much the same way as scientific practices. Religion in this sense can be seen as deploying spirits as explanatory principles much as a scientist might use atoms or molecules. Horton's assumption is that religion is fundamentally *about* explanation, a so-called 'intellectualist' stance toward religion that places him in the genealogy of both Tylor and Frazer. The key contrast here is with so-called symbolist approaches, which draw on a Durkheimian inheritance to emphasize how religion, and in particular ritual, should primarily be interpreted as made up of representations of the social order.

Despite such differences of opinion, the basic consensus among anthropologists is that human capacities to think are universally the same the world over. Another book published in the 1960s, Lévi-Strauss's *La Pensée Sauvage* (1966: 'The Savage Mind'), denies that non-Western mental abilities are inherently inferior to those of Western readers, even if attitudes toward abstract thought may differ.

The Body

Durkheim's and Radcliffe-Brown's functionalist approaches drew on a model of society as equivalent to a body, made up of interlinked and mutually dependent parts. Other authors in the Durkheimian school provided significant perspectives on the physical workings and symbolism of the body. Marcel Mauss, Durkheim's nephew, developed the notion of 'habitus,' the idea that certain elements of culture (habits, tastes, skills) become rooted in the body over time. Within Anglo-Saxon anthropological circles, Mary Douglas's combination of British empiricism and French structuralism resulted in some highly influential work on the function of the body as social and

religious symbol. In *Purity and Danger* (1966), Douglas (a student of Evans-Pritchard at Oxford, and like him a Roman Catholic) explored Western as well as non-Western contexts. In her view bodies, like societies, can be seen as bounded systems, whose integrity and boundaries often need to be protected. What is considered pure and impure, however, varies cross-culturally, and reflects the social experience of the social group involved. Thus in Hindu society, high-caste Brahmins and low-caste untouchables are often interdependent in economic and social terms, but idioms of purity and pollution regulate their behaviors toward each other in very specific ways – for instance, through preventing them from sharing food. A further implication of this argument is that symbols or actions that defy categories or bridge them may be seen as powerful, and under certain circumstances even sacred, rather than polluting. We see here parallels with the potent middle section of the rite of passage, where initiates are passing 'in between' statuses.

Douglas has little to say about the actual physical experiences of having a body that possesses senses and moves through the world. A leading figure in the move toward a more phenomenological view of the body has been the American anthropologist Thomas Csordas (1997), in his studies of healing, language use, and spatial orientation among charismatic Catholics in America. Csordas has presented a sense of the body as both 'self' and 'not-self,' both subject and object, but also the very grounding of the human experience of culture. Important work has also been carried out on the relationship between gendered experience and religious engagement (e.g., Austin-Broos, 1997).

Language

Language mediates much religious activity, ranging from spells to prayers to texts. Some religious language seems not merely to describe reality, but also to have 'performative' dimensions, in other words to create material or social change through being articulated by authoritative figures. An early piece by Bloch (1974) explores his interests in the relationships between ritual and authority by arguing that the religious oratory of the Merina of Madagascar is expressed in a language that is so formalized that it is difficult ever to argue against: we might compare it with a Latin Mass, for instance, or with the inauguration speech of a President.

Some of the broader issues relating to religious language are explored by Keane (1997). For him, a fundamental question is: What, if anything, is particular to religious language? He suggests that such language may raise a particular kind of problematic: the effort to know and interact with an 'otherworld' tends to demand highly marked uses of linguistic resources. In religious contexts the sources of words, as well as the identity, agency, authority, and presence of nonhuman as well as human participants in an interaction, can be especially problematic.

Witchcraft

Witchcraft has proved a resonant subject in part because it invokes many of the themes evident in other areas of the subfield. The witch represents an 'impure' challenge to social order as well as to physical well-being, and embodies some of the key anxieties of a given society. Evans-Pritchard's classic study of the Azande retains much of its power because his argument is rooted in analysis of both social relations and modes of thought. We follow Azande investigating why misfortune appears to strike certain people at certain times, while accusations are linked to tensions in relations among equals. Subsequent work on witchcraft from the 1940s onwards often examined it in the context of changing social relations and periods of communal stress.

Later work has examined the question of rationality and witchcraft under very different social conditions. Tanya Luhrmann's *Persuasions of the Witch's Craft* (1991) examines the role of witchcraft – or Wicca – among middle-class Londoners in the light of theories of the sociology of knowledge as well as of conversion. Luhrmann traces the gradual emergence of commitment to magical ideas among practitioners who have easy access to alternative modes of thought, and who may deploy very different systems of rationality in their work lives.

Some of the most recent work on witchcraft and wider notions of evil and the occult links these again with forms of anxiety rising from shifting and uncertain social, political, and economic circumstances. Meyer (1999) analyzes notions of evil in the context of the emergence of local Christianity and its relation to changing social, political, and economic formations among the Peki Ewe in Ghana. Her main argument is that, for the Ewe, the Devil forms a hybrid figure, expressing people's obsessions with occult forces as a way to mediate the attractions and discontents of modernity. Writing of the postcolonial period in South Africa but also highlighting resonances with other postrevolutionary societies, Jean Comaroff and John Comaroff (1999) explore occasions when vast wealth appears to concentrate in the hands of just a few citizens, so that the market seems to contain mysterious mechanisms of accumulation and distribution. In such contexts, they argue, disenfranchised people imagine new, magical means to attain otherwise unattainable ends, even as they mistrust those who appear to enrich themselves through illegitimate deployment of the forces of production and reproduction.

Newer Themes

Despite predictions from secularization theorists that the significance of religion would weaken around the world along with modernization, religion has retained and even increased its profile in many public as well as private contexts. To some degree anthropologists have turned their attention to secularism itself as a particular, historically constructed category. The study of religion has generated new topics and methods as well as building on more established ones.

Anthropologies of 'World' Religions

Although most anthropologists feel uneasy with the idea that so-called world religions such as Christianity, Islam, Judaism, Buddhism, and Hinduism can be regarded as autonomous systems, there has been a move in recent decades for researchers to identify themselves as ethnographers of a particular religion. This development is a result of a number of factors: the shift of fieldwork toward more

urban, plural contexts, where self-conscious religious identification is occurring; increasing interest in past encounters between Western missions and colonized peoples; a new focus on the affinities of different religious forms with 'modernity'; the relative success of Islam and Christianity as missionary forces; and the self-conscious identification of many informants with religions that have transnational, including diasporic, referents.

A major theme in the anthropology of Islam that has recently emerged relates to the emergence of 'Muslim publics' that have been energized by migration, challenges to Western hegemony, and access to new media technologies. Eickelman and Anderson (1999: p. 1) refer to the emergence of a new sense of public throughout Muslim-majority states as well as Muslim communities elsewhere. They propose the existence of a distinctly Muslim public sphere located at the intersections of religious, political, and social life, operating outside formal state control. The role of this sphere in the emergence of the 'Arab Spring' from the end of 2010 remains to be analyzed. Hirschkind (2006) focuses on the role of the cassette sermon in promoting Islamic revival. His argument that listening to sermons links to ethical self-improvement forms a bridge to another significant theme in the analysis of Islam: the cultivation of piety. Mahmood's *Politics of Piety* (2005) focuses on a women's movement in the mosques of Cairo. While members cultivate embodied practices of personal piety that appear to submit to the patriarchal logic of certain forms of Islam, Mahmood indicates how informants are motivated by dissatisfaction with both secularization and Westernization. At the same time, her analysis critiques secular liberal assumptions concerning the universality of aspects of feminist theory.

Anthropologists have studied examples of Christian worship ever since the early days of the discipline, yet it has received less attention than other religious expressions. One reason may be the ironic one that Christianity is too close to the culture of Western anthropologists, and so it was avoided or perhaps simply not noticed as a valid topic of study. More recently, the emergence of Christianity as an explicit object of study has contributed to and resonated with many current concerns of anthropology, including globalization, reflexivity, and postcolonialism. A particularly fertile strand of research has been a focus on materiality and its relationship with notions of both transcendence and modernity (Keane, 2007).

Globalization and Diaspora

Some of the most effective work in this area has looked at past processes of colonization, where religion has formed part of the tangled encounter between colonizers and colonized, 'modernity' and 'tradition' (Van der Veer, 1996). More recently, religious communities have taken advantage of, and helped to constitute, the globalization of culture. The years following World War II have seen the reemergence of evangelical forms of Christianity in many parts of the world. Although the term 'diaspora' has in the past most often been used to describe the dispersion of Jewish populations away from the land of Israel, in the last two decades in particular it has been taken to refer to the creation of large South-Asian communities in European and North American contexts (Vertovec, 2000), as well as the movement of African migrants in search of both economic mobility and the possibility of 'reverse mission' within the secularized West.

Processes of globalization and migration, combined with tourism, have ensured that pilgrimage has become an increasingly visible dimension of religious activity in all faiths, and also one that has attracted more anthropological attention. Victor and Edith Turner's seminal work, *Image and Pilgrimage in Christian Culture* (1978), which deployed a rite-of-passage model to emphasize the idea that sacred shrines should be seen as liminal spaces, has been challenged by Eade and Sallnow's (1991) argument that such shrines are in fact deeply enmeshed in everyday political and economic processes. Coleman and Eade (2004) place anthropological studies of pilgrimage in the wider frame of analyzing physical and metaphorical aspects of movement.

Cognitive Approaches

In recent years, research on religion as a means to understand human thought has received a striking new impetus within cognitive approaches to the field. This work has blended the findings of evolutionary and cognitive psychology with those of anthropology. In contrast to the work of anthropologists who emphasize the historical contingency of any given definition of religion, researchers in this subfield have argued that there is a biological basis to religious activity. An important strand has also focused on how counterintuitive or surprising experiences are recalled and transmitted over time (e.g., Boyer, 2001).

Salazar (2010) discerns two main orientations among cognitive researchers. The 'adaptationist' argument states that no human society can survive without language or religion, and so there must be both a language instinct and a religious instinct. The alternative approach is to see religion more as a by-product of evolution. For instance, the so-called Theory of Mind Mechanism can be seen as a cognitive tool enabling humans to interpret a living organism's movements in terms of its inferred intentions: such construal of what is going on in other people's brains is clearly an adaptive skill. As a result of this skill, humans have perhaps also become prone to seeing intention in events, to see them as guided by a controlling force, such as a divinity.

Overall

Both in the past and the present, the ethnographic study of religion has contributed to numerous mainstream analytical concerns, ranging from classic approaches to rationality, symbolism, and social order to more recent studies of transnationalism, materiality, and both cultural and evolutionary change. The study of religion remains in a vibrant state, even as scholars still wrestle with key questions of whether and how religion can be seen as an autonomous area of activity in its own right.

See also: Belief, Anthropology of; Body: Anthropological Aspects; Carnival; Cognitive Anthropology; Death, Anthropology of; Diaspora; Fetishism; Liminality; Magic, Anthropology of; Ritual; Sacrifice; Shamanism; Syncretism; Taboo; Witchcraft; Worldview.

Bibliography

Asad, T., 1993. Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam. John Hopkins University Press, Baltimore, MD.

Austin-Broos, D., 1997. Jamaica Genesis: Religion and the Politics of Moral Orders. University of Chicago Press, Chicago.

Bell, C., 1992. Ritual Theory, Ritual Practice. Oxford University Press, New York.

Bloch, M., 1974. Symbols, song, dance, and features of articulation or is religion a extreme form of traditional authority? Archives Européenes de Sociologie 15, 55–81.

Bloch, M., 1992. Prey into Hunter. The Politics of Religious Experience. Cambridge University Press, Cambridge.

Bloch, M., 2010. Bloch on Bloch on 'Religion'. Religion and Society: Advances in Research 1, 4–13.

Boyer, P., 2001. Religion Explained: The Evolutionary Origins of Religious Thought. Basic Books, New York.

Cannell, F. (Ed.), 2006. The Anthropology of Christianity. Duke University Press, Durham.

Coleman, S., Eade, J. (Eds.), 2004. Reframing Pilgrimage: Cultures in Motion. Routledge, London.

Comaroff, J., 1985. Body of Power, Spirit of Resistance: The Culture and History of a South African People. University of Chicago Press, Chicago.

Comaroff, J., Comaroff, J., 1999. Occult economies and the violence of abstraction: notes from the South African postcolony. American Ethnologist 26, 279–303.

Csordas, T., 1997. The *Sacred Self*: A Cultural Phenomenology of Charismatic Healing. California University Press, Berkeley.

Douglas, M., 1966. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Routledge, London.

Durkheim, E., 1915. The Elementary Forms of the Religious Life (J.W. Swain, Trans.). Allen & Unwin, London.

Eade, J., Sallnow, M. (Eds.), 1991. Contesting the Sacred: The Anthropology of Christian Pilgrimage. Routledge, London.

Eickelman, D., Anderson, J., 1999. Redefining Muslim publics. In: Eickelman, D., Anderson, J. (Eds.), New Media in the Muslim World: The Emerging Public Sphere. Indian University Press, Bloomington and Indianapolis, pp. 1–18.

Evans-Pritchard, E., 1937. Witchcraft, Oracles and Magic among the Azande. Oxford University Press, Oxford.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Gerholm, T., 1988. On ritual: a postmodernist view. Ethnos 53, 190–203.

Henn, A., 2008. Introduction: beyond norm, text and dialectics. Ritual as social praxis. In: Henn, A., Koepping, K.-P. (Eds.), Rituals in an Unstable World: Contingency-Hybridity-Embodiment. Peter Lang, Frankfurt, pp. 9–28.

Hirschkind, C., 2006. The Ethical Soundscape: Cassette Sermons and Islamic Counterpublics. Columbia University Press, New York.

Horton, R., 1970. African traditional thought and western science. In: Wilson, B. (Ed.), Modes of Thought. Blackwell, Oxford, pp. 131–171.

Keane, W., 1997. Religious language. Annual Review of Anthropology 26, 47–71.

Keane, W., 2007. Christian Moderns: Freedom and Fetish in the Mission Encounter. University of California Press, Berkeley.

Lambek, M., 2002. A Reader in the Anthropology of Religion. Blackwell, Oxford.

Lévi-Strauss, C., 1966. The Savage Mind. Weidenfeld and Nicolson, London.

Luhrmann, T., 1991. Persuasions of the Witch's Craft: Ritual Magic in Contemporary England. Harvard University Press, Cambridge.

Mahmood, S., 2005. Politics of Piety: the Islamic Revival and the Feminist Subject. Princeton University Press, Princeton.

Meyer, B., 1999. Translating the Devil: Religion and Modernity Among the Ewe in Ghana. Edinburgh University Press, Edinburgh.

Needham, R., 1972. Belief, Language, and Experience. Blackwell, Oxford.

Rappaport, R., 1999. Ritual and Religion in the Making of Humanity. Cambridge University Press, Cambridge.

Robbins, J., 2004. Becoming Sinners: Christianity and Moral Torment in a Papua New, Guinea Society. University of California Press, Berkeley.

Robbins, J., 2007. Continuity thinking and the problem of Christian culture: belief, time and the anthropology of Christianity. Current Anthropology 48, 5–38.

Ruel, M., 1982. Christians as believers. In: David, J. (Ed.), Religious Organization and Religious Experience. Academic Press, London, pp. 9–31.

Salazar, C., 2010. Anthropology and the cognitive science of religion: a critical assessment. Religion and Society: Advances in Research 1, 44–56.

Saler, B., 2000. Conceptualizing Religion: Immanent Anthropologists, Transcendent Natives, and Unbounded Categories. Berghahn, Oxford.

Turner, V., 1967. The Forest of Symbols: Aspects of Ndembu Ritual. Cornell University Press, Ithaca.

Turner, V., Turner, E., 1978. Image and Pilgrimage in Christian Cultures. Columbia University Press, New York.

Van der Veer, P. (Ed.), 1996. Conversion to Modernities: The Globalization of Christianity. Routledge, London.

Vertovec, S., 2000. The Hindu Diaspora: Comparative Patterns. Routledge, New York.

Religion: Evolution and Development

Robert N Bellah, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 13062–13066, © 2001, Elsevier Ltd.

Abstract

Since the idea of religious development merges into the much larger field of the history of religions, this article focuses on religious evolution. Contrary to common belief, it is to the writings of Max Weber that a defensible conception of religious evolution owes the most. According to Weber, Protestantism, though occurring in only one tradition, was, indirectly but crucially, an indispensable precondition for the emergence of modernity cross-culturally, and thus can legitimately be considered an instance of religious evolution. In this comparative studies in the sociology of religion Weber was particularly interested in the emergence of what he called the world religions in the first millennium BC. By calling these new symbolic forms 'rationalized' Weber was pointing to the fact that they were more coherent, more cognitively and ethically universalizing, more potentially self-critical (reflexive), and more disengaged from the existing society than what preceded them. Just as Protestantism emerged out of one of the world religions, so they themselves emerged out of societies organized in terms of kinship and neighborhood in which religion was primarily magical, by which Weber meant, one in which ritual means were used to attain utilitarian ends. This three-stage typology remains fundamental although it can be significantly refined.

Both evolution and development are ways of looking at the history of religion. Development implies some kind of directional change in the history of a particular religion, whereas evolution implies comparable directional changes in a number of religious traditions. Since histories usually describe directional change of some sort, the idea of religious development merges into the much larger field of the history of religions. Therefore this article will focus on religious evolution.

The Term 'Religious Evolution'

The idea of religious evolution, like the idea of social evolution, has been much criticized in the twentieth century, although neither idea has been entirely abandoned. If the notion of religious evolution is to be salvaged it must be dissociated from several preconceptions of its nineteenth century proponents. These preconceptions include the idea that all religions must be fitted into a single (unilineal) pattern, that progression through a fixed series of stages is inevitable, and that more advanced stages are 'higher' (ethically and spiritually better) than earlier ones (Comte, 1830–42; Tylor, 1871; Spencer, 1876–1997; Marett, 1914, etc.). It is the notion of inevitable progress that has been rejected by twentieth century writers; the idea that religious change correlates with social change in comparable ways across traditions can still be defended as 'religious evolution.' Contrary to common belief, it is to the writings of Max Weber (1920–1921, 1921–1922/1978) that a defensible conception of religious evolution owes the most.

The Protestant Ethic

Weber began his study of religion with his famous *The Protestant Ethic and the Spirit of Caplitalism* (1904–05/1930). At first glance, it might seem that this book is simply a developmental history focusing on a dramatic change within a single religious tradition. However, it is clear from the place of the Protestant ethic argument in *Economy and Society* (1921–22/ 1978) and in the *Gesammelte Aufsätze zur Religionssoziologie* (1920–21) that Weber believed that ascetic Protestantism was an indispensable catalyst for the emergence of a new form of society, which he called modern capitalism. Never before the Reformation in the West was the hold of agrarian society, organized through patrimonial (including patrimonial bureaucratic) and feudal political forms, broken through. Once established, however, capitalism became a worldwide phenomenon, even though taking different forms in different civilizational areas. Thus Protestantism, though occurring in only one tradition, played, indirectly but crucially, a developmental role cross culturally, and thus can legitimately be considered an instance of religious evolution.

Since Weber's time there has been a search for functional equivalents of Protestantism in other traditions that prepared some societies better than others and some groups in some societies better than others for the transition to capitalism, or, perhaps better, to modernity (Bellah, 1957; Eisenstadt, 1968). Ernest Gellner made the interesting suggestion that there were elements in all the great traditions, Islam in particular, that might have led to a Protestant-like breakthrough, but that a combination of fortuitous circumstances in Western Europe made it possible for it to occur there (1988). This raises the interesting question of what gave the great religious traditions a rationalizing potentiality even if it was only fully realized in one case.

The Axial Age

In his comparative studies in the sociology of religion (1920–21) and in the sociology of religion section of *Economy and Society* (1921–22/1978) Weber pointed out that in each of the world religions, sometime in the first millennium BC, there emerged prophets or saviors who radically rationalized

previous forms of what he tended to call magical religion. In each case the emergent figure (Confucius, the Buddha, the Hebrew prophets, Socrates, Jesus) preached a systematic form of ethical conduct quite different from the diffuse ritual and sacramental practices that preceded them. By calling these new symbolic forms 'rationalized' Weber was pointing to the fact that they were more coherent, more cognitively and ethically universalizing, more potentially self-critical (reflexive), and more disengaged from the existing society than what preceded them.

Weber's friend, Karl Jaspers, spoke of this period as the Axial Age (1948), and S. N. Eisenstadt has contributed significantly to the further study of what he calls Axial civilizations (1986 and elsewhere). Following Weber one can characterize each of the Axial breakthroughs (first millennium BC China, India, Israel, Greece) as involving, to different degrees and in different ways, an element of world rejection (Weber, 1920/1946). Religious renouncers (Dumont, 1980) in each case called the empirical world into question and, by invoking a transcendental realm, provided a significant degree of tension between worldly interests and religious demands that enhanced the possibility of social change. Although Weber himself contributed to the stereotype of static traditional societies, in his studies of particular traditions he discerned significant possibilities of social transformation (Weber, 1920–1921, 1921–1922/1978).

Subsequent scholars have pointed out the dynamic potentialities in all the great traditions. What scholars such as Gellner have emphasized is that a radically rationalized ethic of conduct, such as that found in Protestantism, was not in itself enough. As long as the existing military, political, and religious hierarchies of agrarian societies could contain these rationalizing potentialities within monastic orders or private conventicles, a new type of society could not emerge. What made Protestantism successful was not just its unique ethic but the possibility of gaining sufficient political influence (in Geneva, the Netherlands, England, and New England) to institutionalize that ethic in the major spheres of social life.

The Shape of Religious Evolution

If Weber had at least two stages of religious evolution, namely an Axial stage correlating with the emergence of historic agrarian societies, and a modern stage for which ascetic Protestantism was an indispensable precondition, did he have any more stages? In looking at his historical sociology as a whole it is clear that he did. The world religions were preceded by societies organized primarily in terms of kinship and neighborhood and in which religion was primarily magical, by which Weber meant, one in which ritual means were used to attain utilitarian ends such as rain, good harvests, human fertility, etc. While such a three-stage typology is for many purposes adequate, it does seem possible to refine it further.

Bellah, in his essay *Religious Evolution* (1964), suggested a further differentiation of the early stages. He distinguished between a 'primitive,' now better, 'tribal', phase, and an archaic phase. Myth and ritual characterize tribal religion – indeed they may be coterminous with hominization itself – but they are myth and ritual of a particular type. The spiritual beings whose activities are narrated in the myths are not gods and are not worshipped. These half-human, half-animal, often ancestral, beings express in archetypal fashion the social and ecological life of the people. Ritual involves the reenactment of the primordial events such that mythical time becomes present, not the worship of distanced deities. Tribal religion is most characteristic of hunter-gatherers, but is also common in horticultural societies. With agriculture and the emergence of complex chiefdoms, and particularly with the early state, archaic religion comes into being. While in tribal societies every clan can be traced to an ancestral being, in archaic societies only the aristocratic lineages claim divine descent. A hierarchic pantheon of gods appears, quite different from the mythical beings of tribal religion. The gods require worship, and, significantly, sacrifice, in many cases human sacrifice, and a differentiated class of priests who supervise the worship. In archaic societies the ruler often claims descent from the highest god, even the claim to be a living god himself. In general, archaic religion is characterized by a fusion of political and religious hierarchy largely missing in tribal society. Examples of archaic societies, often identified as Bronze Age societies, would include third and second millennium BC Mesopotamia, Egypt, the Indus Valley, and North China, but also early first millennium AD Japan, and first millennium AD Meso-America and Peru.

It was in the midst of or on the peripheries of such archaic societies that the axial religions of the first millennium BC emerged. They took the hierarchical tendencies of archaic religion in a radically transcendental direction. All value was seen to reside in a realm beyond this world and thus this world and all the worldly hierarchies within it were relativized and potentially delegitimized. But in most cases the new Axial religions either remained marginal to the centers of power or, if they were recognized by the ruling center, they made compromises with it so that they largely legitimated that center rather than challenged it, thus allowing a quasi-archaic situation to continue. This is not to minimize, however, the importance of Axial religions in the elite education and culture of each of the great traditions. Also, the critical potential of the Axial religions was continually reasserted in the form of heresies, uprisings, and rebellions (Eisenstadt, 1986, 1999), but these were normally either destroyed, marginalized or, if successful, coopted by the existing structures of power.

A simplified way of looking at the shape of religious evolution was proposed by Bellah (1964). He argued that whereas tribal and archaic religions were primarily this-worldly in orientation, which is what Weber, perhaps unwisely, meant by the word magical, the Axial religions were world-rejecting (and thus, for Weber, very importantly, magic rejecting), although they differed as to whether this rejection was to be worked out within the world (ethically), or as far as possible outside the world (mystically). It should be remembered that in its radical consistency Axial religion was never more than the religion of a minority; the majority continued to entertain beliefs and practices continuous with archaic or even tribal religion, which is what Weber meant by the return to the garden of magic (1920–21).

With the Protestant Reformation, however, the belief in a radically transcendent God had dramatic this-worldly

consequences: the consistent demands of an Axial ethic were to characterize every sphere of daily life. But in the subsequent development of modernity, though the this-worldly dimension remained dominant, its transcendental basis became transformed into immanentism, thus returning the modern world in a much different way to this-worldly immanentism of pre-Axial times. Weber clearly observed this transition, but viewed it almost entirely negatively. The modern world of rationalization would run on its own bureaucratic and economic energies without any transcendental sanction, would become an iron cage (1904–05/1930). Charles Taylor has described the same transformation in less somber hues as 'the affirmation of ordinary life' and has stressed growing individuation (1989). In any case, though the world of the Protestant Reformation and the Western world since the eighteenth century (and, more recently, much of the rest of the world, as well) are clearly different, it is doubtful whether the Protestant phase needs to be designated as a separate 'early modern stage' as Bellah did in 1964, or whether it should be seen as merely the opening transition toward modern religion, the contours of which, though it has been developing for several centuries, are still far from clear.

The shape of religious evolution should not obscure its many irregularities, its lack ofunilineal consistency. Regression was always a possibility: Axial Greek civilization did not emerge from the womb of archaic Minoan and Mycenean societies, but only after the breakdown of those societies and the return to tribalism. Stages were sometimes skipped, especially on the peripheries of advanced societies: the Israelite Axial breakthrough began in a tribal confederation on the boundaries of archaic empires; Muhammad grew up in the tribal city of Mecca but founded an Axial civilization, to be sure under Jewish and Christian influence. Not only was every religion from tribal to modern full of internal tensions and divisions but all of them were open to continuous external influence.

The Causes of Religious Evolution

To move beyond description and consider causation is always difficult in historical sociology; nevertheless, how one answers the question of causation, even tentatively, will determine how one understands the entire process. Religious evolution is certainly embedded in social evolution and cannot be understood apart from it. Long-term tendencies, such as the increasing division of labor, the increasing capacity of states to mobilize social power, the development of media such as money and writing, the differentiation of elites of various types as carriers of distinct cultural traditions, the tendency of pre-modern societies to oscillate between political centralization (patrimonialism) and decentralization (feudalism) all influenced and were influenced by religious beliefs and practices.

There has been a tendency among supporters of the idea of social evolution, Marxist and non-Marxist alike, to emphasize economic factors as causal, and religious developments as epiphenomenal. It has been common to believe that Max Weber reversed this understanding and held that religion was primary in social causation. A reading of Weber's entire work, which puts *The Protestant Ethic and the Spirit of Capitalism* in its proper perspective, makes it clear that such was not his belief. Rather, Weber argued for the relative independence of the religious dimension as a factor in social life, but one that always interacts with other factors, notably economic, political, and legal. S. N. Eisenstadt rephrases Weber by showing that it is not just a matter of emphasizing multi-causality: while never sufficient as a casual factor in itself, religion provides the indispensable element of *cultural codes*, not only for the institutionalization of new social forms but, when internalized in personality systems, for new forms of individual motivation as well (Eisenstadt, 1999).

Religious elements have played a critical role in all the great social transitions. Terrence Deacon has even argued that the origin of language among the early hominids was probably facilitated by what he specifically calls ritual (1997). Some primates today can learn words, but only under the most intense pressure of human teaching. Deacon hypothesizes that the bodily and repetitive nature of ritual may have been a necessary precondition for the counterintuitive, from an early primate point of view, shift to symbolic representation that is the basis of language. In any case, Paleolithic evidence, difficult to interpret though it is, suggests that ritual played a central role in creating the social solidarity and respect for the normative order of kinship relations that differentiate humans from other primates.

The transition to archaic society is associated almost everywhere with monumental temple architecture, large-scale sacrifice, and divine kingship. It would seem that the move from the basic egalitarianism of tribal society to the intensely hierarchical and centralized order of archaic societies was only possible with the institutionalization of new religious codes, in association with an unprecedented concentration of social power.

While causal linkages are far from clear, the first millennium BC did see a number of developments that might have prepared the ground for the rise of the Axial religions. These include the beginnings of a market economy, an increase in bureaucratic and military power, and increasing differentiations of social classes, all of which could have given rise to Weber's 'problems of meaning' (1921–1922/1978). But an indispensable precondition for these radical new religious developments was the existence of cultural elites at least to some degree independent of central political power. These include the Hebrew prophets, the Greek philosophers, the Chinese sages, and the Indian ascetics and mystics. It was among these groups that archaic symbol systems were radically reinterpreted in a transcendental ethical and spiritual direction.

The influence of the Protestant Reformation on the emergence of modern society has been the subject of a vast literature. A significant recent development, however, has been the recognition of the importance of a 'disciplinary revolution,' particularly associated with the Calvinist wing of the Reformation (Gorski, 1993). The disciplinary revolution operated at a number of levels: a new code of systematic ethical action expected from believers, new institutions capable of enforcing the code, and new solidary organizations strong enough to influence political and economic developments. The transformation of these new codes under a variety of influences and pressures into such tendencies as utilitarianism, liberalism, Jacobinism, and socialism, and more recently,

individualism and fundamentalism, has significantly affected the shape of and the tensions within modern society.

Thus, viewing the process as a whole, it appears that religious evolution is not only a subject of great interest in itself, but is a critical dimension of social evolution.

See also: Civil Religion; Comte, Auguste (1798–1857); Reformation and Confessionalization; Religion, Phenomenology of; Religion, Sociology of; Religious Fundamentalism: Cultural Concerns; Secularization; Spencer, Herbert (1820–1903); Weber, Max (1864–1920).

Bibliography

Bellah, R.N., 1957. Tokugawa Religion. Free Press, Glencoe Il.
Bellah, R.N., 1964. Religious evolution. American Sociological Review 29, 358–374.
Comte, A., 1830–42. Cours de Philosophie Positive. Bachelier, Paris.
Deacon, T.W., 1997. The Symbolic Species: The Co-evolution of Language and the Brain, first ed. Norton, New York.
Dumont, L., 1980. Homo Hierarchicus. University of Chicago Press, Chicago.
Eisenstadt, S.N., 1968. The Protestant Ethic and Modernization: A Comparative View. Basic Book, New York.
Eisenstadt, S.N. (Ed.), 1986. The Origins and Diversity of Axial Age Civilizations. State University of New York Press, Albany, NY.
Eisenstadt, S.N., 1999. Axial age sectarianism and the antinomies of modernity. In: McKnight, S.A., Pierce, C.I. (Eds.), Politics, Order and History. Sheffield Academic Press, Sheffield, UK.
Gellner, E., 1988. Plough, Sword and Book. University of Chicago Press, Chicago.
Gorski, P.S., 1993. The Protestant ethic revisited: disciplinary revolution and state formation in Holland and Prussia. American Journal of Sociology 99, 265–316.
Jaspers, K., 1948. The Origin and Goal of History. Routledge and Kegan Paul, London.
Marett, R.R., 1914. The Threshold of Religion, second ed. Methuen, London.
Spencer, H., 1876–1897. The Principles of Sociology. D. Appleton, New York.
Taylor, C., 1989. Sources of the Self. Harvard University Press, Cambridge, MA.
Tylor, E.B., 1871/1974. Primitive Culture. Gordon, New York.
Weber, M., 1904–1905/1930. The Protestant Ethic and the Spirit of Capitalism [trans. Talcott Parsons]. Scribner, New York.
Weber, M., 1920–1921. Gesammelte Aufsätze zur Religions-soziologie. J. C. B. Mohr (Paul Siebeck), vols. 3. Tubingen, Germany.
Weber, M., 1920/1946. Religious rejections of the world and their directions. In: From Max Weber: Essays in Sociology [trans. Gerth H H, Wright Mills C]. Oxford University Press, New York.
Weber, M., 1921–1922/1978. The sociology of religion. In: Roth, G., Wittich, C.M. (Eds.), Economy and Society. University of California Press, Berkeley, CA.

Religion, Family, and Kinship

Annette Mahoney, Bowling Green State University, Bowling Green, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines contemporary research regarding the connection between religious and spiritual beliefs and practices and family structure and functioning. Religion and family are strongly connected, and religious institutions and beliefs influence the formation of families, fertility, childrearing behaviors, divorce, and coping with family conflict. Contemporary research is unfortunately dominated by samples from the US, and more diverse studies are needed.

There are entrenched historical ties between religion and family life across the globe. In the twenty-first century, diverse religious communities on every continent continue to offer their adherents guidance on virtually every aspect of forming and sustaining intimate bonds within family and kinship circles across the life span, including courtship, cohabiting, mating, childbearing, and nurturing family members of every age. Yet, contemporary social scientists have primarily focused on individual religiosity rather than collective commitments when investigating influences between religion and family or kinship networks. That is, most empirical attention is directed toward positive and negative roles that religion can have on individuals' health and psychosocial functioning, rather than on how religion might influence the structure and quality of intimate, close relationships. The relative paucity of rigorous scientific studies on religion within the context of family and kinship relations is illustrated by comparing the number of peer-reviewed studies focused on marriage versus on religion and marriage. When the word 'marriage' is entered as a search term into the three fields of Abstract or Keyword or Subject for the years 1980–2009, the PsycINFO search engine yields 10 322 studies and SocINDEX yields 11 252 articles published in behavioral and social science journals in English. Yet across two reviews that spanned this 30-year period (Mahoney et al., 2001; Mahoney, 2010), Mahoney and colleagues, were able to locate only 131 peer-reviewed studies linking religion and marriage. In this article, I summarize the nature of scientific research conducted on links between these two domains and scientific findings.

To provide a broad perspective on cross-cultural issues that shape peer-reviewed empirical studies on faith and family, the following issues need to be recognized. First, similar to scholarly debates about the ambiguous boundaries surrounding definitions of 'family' and 'kinship,' controversy exists among social scientists about defining 'religion' versus 'spirituality' for the purpose of scientific studies. Clarifying boundaries between religion and spirituality seems pressing given that affiliation rates and public attendance at well-established religious institutions are on the decline in North America, Europe, and Australia. Further, in scientific literature dominated by researchers working in Western societies, 'being religious' is increasingly portrayed as involving membership in a given religious community, adherence to institutional dogma, worship or rituals, and external pressure. By contrast, 'being spiritual' is increasingly framed as a personal connection to what an individual holds as sacred; a private search for enlightenment, purpose, meaning, or virtues and internal motivation. However, organized religion remains the primarily social institution that attempts to promote spirituality, with nearly all Americans who self-describe as 'religious' also describing themselves as 'spiritual.' More pragmatically, academic debates over whether it is religion or spirituality that matters for family relationships turns out to be largely moot in this niche since most studies rely heavily on single-item questions to assess either domain. For example, the vast majority of studies conducted on religion or spirituality and family life from 1980 to 2009 asked if a given household member self-identifies as affiliated with a general religious tradition (e.g., Catholic, Protestant, Muslim, Jewish), attends religious services, or reports that religion or spirituality was personally important to his or her own life. Relatively few studies asked family members about specific religious or spiritual beliefs or practices focused on family relationships that are promoted by diverse religious groups. As a result, it is easy to argue that a global religious variable, such as greater religious attendance, is tied to family functioning for reasons that have little or nothing to do with the content or substance of religious scriptures or teachings, beliefs, practices, or coping strategies that involve supernatural forces. For example, a correlation between higher religious attendance and lower divorce risk can be attributed to social coercion and/or support that spouses receive to remain married because their lives are embedded in a social network that discourages marital dissolution, not because of specific religious beliefs about marriage promoted by religious groups. Social membership in a cohesive secular group that likewise discourages divorce without religious beliefs about marriage could presumably operate similarly to lower divorce rates. Alternatively, the link between greater religious attendance and lower divorce rates could be attributed to the degree to which spouses internalize religious beliefs or behaviors promoted by religious social networks. For example, diverse religions teach that marriages can be imbued with sacred qualities (e.g., everlasting, transcendent) and manifest the actions of a Higher Power. Single-item measures of religious involvement obscure the extent to which secular and religious mechanisms work independently or interactively to foster psychosocial outcomes. However, as highlighted later, some specific religious beliefs and behaviors promoted by some

This article is based on two of the authors' previous exhaustive reviews of the literature on religion, family and kinship: Mahoney 2010; Mahoney, Pargament, Swank, and Tarakeswar, 2001. These peer-reviewed journal articles may be consulted for detailed citations and discussion to support the statements made in the present article.

religious groups have consistently been linked to family structures and processes.

Second, although diverse definitions and compositions of family and kinship circles exist cross-culturally, most empirical studies in the past three to four decades have focused on the role of religion for so-called nuclear family units comprising married, heterosexual couples with children. For example, although numerous small-scale qualitative studies have focused on the role of religion in family life with people with ethnic or religious minority background (e.g., Latter-day Saint, Jewish, Muslim), most of these participants were married heterosexuals with children. Such studies leave many questions open about the relevance of religion for cohabiting, childless or older couples regardless of their sexual orientation, or families with minors headed by single, foster, step, or divorced parents or coparents, grandparents, and same-sex couples. Further, most quantitative surveys of large samples have involved predominantly Christian participants drawn from the United States and have focused on the marital satisfaction of heterosexual couples or their parenting practices. Relatively few large-scale studies have involved people who belong to other types of families and/or non-Christian groups, leaving open much room for rigorous scientific studies on various religious or spiritual beliefs and practices about family life across diverse religious and kinship networks.

Finally, methodological issues constrain inferences that can be drawn from available empirical studies on faith and family life. Namely, most studies rely on cross-sectional rather than longitudinal or experimental designs. This issue makes drawing causal conclusions about the influences between family and religious variables difficult because the relationships may be confounded by other factors such as socioeconomic status, education, and personality variables. Most studies also rely on self-reports from one given family member, rather than multiple family members, for both religious and family variables, and direct observation of family interactions is very rare. These two limitations raise concerns about self-report biases driving associations between religious and family factors.

Relational Spirituality: Organizing What We Know for Family Relationships

Despite the aforementioned challenges in interpreting available empirical research, social scientists have fruitfully examined many family topics in connection with religion. These topics include union formation; maternal fertility; paternal time with offspring; spousal egalitarianism; work–home balance; divorce risk; marital satisfaction; risk of domestic violence and infidelity; pregnancy; discipline, warmth, and physical abuse toward children; parenting style and relational quality with adolescents; and domestic violence and post-divorce adjustment. To organize and clarify consistent findings across this wide range of topics, Mahoney (2010) developed a relational spirituality framework that differentiated three general stages of family relationships: (1) discovery – forming and structuring a particular set of family or kinship relationships, (2) maintenance – sustaining processes in established relationships, and (3) transformation – coping with changes in the structure and/or processes of relationships. This schema has three advantages. First, controversies about religion and family life reflect underlying theological disagreements between socially conservative versus socially progressive religious or secular subgroups over the perceived moral legitimacy of diverse family structures. For instance, within and across religious subcultures, people hold conflicting theological stances about the morality of same-sex marriages, nontraditional gender roles, and the ideal composition of household units. Second, these conflicts between religious subgroups make it easy to lose sight of the fact that many religious subgroups promote similar virtues for how family members should behave toward members of their chosen family and kinship circles (e.g., display love and commitment). Third, it is possible that certain religious beliefs reinforce destructive family processes within subsamples where people are struggling with family-related stressors or dysfunction, yet this also might obscure evidence that religious involvement tends to lower the risk of family-related stressors. For example, although a minority of individuals may draw upon religious beliefs to defend acts of domestic violence or child physical abuse within families where such family problems exist, higher religious attendance generally lowers the risk of exposure to both forms of family aggression based on large heterogeneous samples of people.

Relational Spirituality: Forming Family Relationships

Getting Married

Cross-culturally, adults have historically and continue to rank religious similarity as an important factor when searching for a mate. Men and women express a similar desire for a potential spouse to belong to the same religious tradition (e.g., 'none,' Catholic, Protestant). Couples also decide early in their courtship whether (non)religious compatibility matters. That is, the percentage of couples who have the same religious affiliation does not change across the stages of dating, being sexually intimate, cohabiting, or marrying. Conservative Protestants, Catholics, Latter-day Saints, and Jews are the most likely to marry partners from their religious affiliation compared to others. In addition, those who view religion as more important to their daily life and who more often attend religious services are less likely to cohabit. Similarly, people from all denominations who see religion as highly important more often get married in their early 20s rather than delaying marriage. Higher religious attendance also increases the odds that unmarried, pregnant women will marry after giving birth. Analogously, gay and lesbian individuals who view religion as important more often engage in ritualized commitment ceremonies with same-sex partners.

Maternal Fertility

For centuries, religions have encouraged married couples to procreate. This message continues to sway women. Those who view religion as important to their lives are more likely to give birth after age 24 years. This is consistent with diverse religious teachings that motherhood should be delayed and ideally reserved for marriage. Further, women who report on surveys that religion is unimportant are more likely to have

unplanned births, especially during adolescence, or to remain childless into middle age. Also, regardless of how often they attend services or their particular faith tradition, women's plans to have children are also tied to greater importance of religion in their lives. This finding applies to women who do or do not hold socially conservative attitudes toward feminism or family life emphasized by some religious groups. Overall, adult women who value their religious faith are more likely to have children and want to be mothers. Further efforts are needed, however, to pinpoint specific spiritual beliefs or practices centered on motherhood that motivates this life goal.

Paternal Involvement

Researchers examining married men in the United States have found that conservative Protestant fathers do not spend more time than other fathers with their children. In fact, unmarried American fathers are more likely to avoid offspring born out of wedlock if they are conservative Protestants. On the other hand, across religious denominations, fathers who attend religious services more often are more likely to spend time playing with their biological children. This implies that more intensive religious involvement across the spectrum of religious traditions helps motivate fathers to invest time in creating father–child relationships but more research on this topic is merited.

Home and Work Balance

Couples face complex choices in deciding who does what in and out of the home. Many couples cannot afford for either spouse to be a full-time homemaker, and most heterosexual American wives are employed before and after having children. Thus, couples face a difficult juggling act in assigning the roles of breadwinner and homemaker. Yet Americans from conservative Protestant or Muslim families report diverse values about how couples should negotiate such issues. That is, conservative Protestants and Muslims define equality within families through the lens of their religious beliefs. Contrasting interpretations of the Bible in evangelical Christian circles urge husbands to devote more time to their 'manly' segregated household tasks or to pull more weight across the board. Yet neither message seems to have hit home based on US national data. Conservative Protestant wives do more of both traditional 'female' and 'male' housework than other wives, and conservative Protestant husbands exert the same effort as other husbands on both types of housework. Nevertheless, more progressive Christian couples are not markedly more egalitarian in marital decision making. That is, as a group, conservative Protestant couples are equally likely as other couples to share decisions about financial matters, child-rearing, and who should work outside the home. Thus, although more conservative Christian couples say they believe more in traditional gender roles than mainline/liberal couples, the actions of both groups tell a different story. There is considerable evidence that conservative Protestantism reduces female employment and earnings even more than it does for males, and this reinforces traditional gender investments is careers and family.

Relational Spirituality: Sustaining Family Relationships

Despite theological divisions within and across religious subcultures about whether an ideal family structure is that of married heterosexuals with children, most religious traditions encourage similar virtues (e.g., love, self-sacrifice, commitment) to facilitate the stability and quality of family relationships that do create.

Enhancing Marital Quality

Higher marital satisfaction has been tied repeatedly to each spouse's general religiousness (e.g., religious attendance rates, importance of religion) and husband–wife similarity on such factors. Emerging research offers insights into three specific religious or spiritual constructs that may enhance marital dynamics. First, spouses can view their marriage as having spiritual significance, either by seeing the union as having sacred qualities (e.g., is sacred; is part of a larger spiritual plan) or being a manifestation of God (e.g., God plays a role in; is a reflection of God's will). Several studies show that greater belief in the sanctification of one's union predicts more marital love, satisfaction, and constructive communication processes; such effects are not merely the result of other positive elements of the marriage or spouses' overall religious involvement. Second, couples can engage in intimate discussions about spirituality. Such dyadic activity is tied to higher marital satisfaction and better communication processes. Analogously, college students and mothers who mutually discuss their spiritual journeys tend to have more harmonious relationships and resolve conflict collaboratively. Third, family members can privately pray for the well-being for other family members. Rigorous experimental and longitudinal small-scale studies have found that prayer for a romantic partner or spouse improves relationship quality. In qualitative surveys, married couples report that turning to God as a resource helps them to resolve conflict by encouraging mutual self-reflection, patience, empathy, and accountability. On the other hand, family members may use religious beliefs to escalate distance and conflict between family members. For example, couples who deeply disagree in their interpretation of the Bible experience more conflict.

Decreasing the Risk of Domestic Violence and Sexual Infidelity

According to multiple studies, more frequent attenders of religious services are much less likely than infrequent attenders to perpetrate, or to be the victim of, domestic violence in marital, cohabiting, or dating relationships. Further, conservative Protestant identification, fundamentalist beliefs about the Bible, and interfaith marriage does not increase the risk of domestic violence except when biblically conservative men are married to more liberal wives, which increases the risk of domestic violence.

Individuals who more often attend religious services are also less likely to engage in extramarital sexual relationships. The odds of infidelity paradoxically increase for high attenders who do not feel close to God and for low attenders who do feel close to God. Also, purely private spiritual experiences (e.g., prayer, subjective spiritual growth) make no difference for the likelihood of having an affair. Thus, sexual fidelity seems to depend both on religious

communal support for, and the personal internalization of, a spiritual commitment to sexual fidelity within marriage. Further, sexual infidelity is more likely to trigger a subsequent divorce for couples who frequently attend religious services.

Disciplining and Nurturing Children

Parents with conservative Christian identities (e.g., affiliation, Biblical conservatism) tend to spank children more than other parents. This does not mean, however, that conservative Christian parents are excessively harsh. In fact, in national US surveys, this religious subgroup of parents report yelling less and being more physically affectionate with their children than other parents. Further, in one in-depth study, conservative Protestant parents were no more likely to spank their preschoolers when stressed than other parents and were equally likely to use nonpunitive disciplinary techniques. Further, conservative Protestant fathers report giving their children more affection and supervision than other fathers. Overall, conservative Protestant parents appear, on average, to blend firmness and warmth in child-rearing in a manner consistent with their religious views of parenting. Moreover, some studies have suggested that corporal punishment is less harmful in the normative context of conservative Protestantism.

Much less is known about the ways religion may promote parenting by those who do not identify with conservative Protestantism. Yet higher religious attendance and personal importance of God or spirituality has been repeatedly correlated with positive parenting by unmarried mothers especially at the risk of having difficulties. Namely, these factors predict more satisfaction, self-confidence, authoritativeness, and consistency in parenting as well as less parental distress and risk of child abuse by young, single mothers with limited financial resources and/or from ethnic minority groups within the United States.

Risk of Child Physical Abuse

Multiple peer-reviewed studies show that frequent religious attendance by parents substantially decreases their risk of being physically abusive to their children over time. Further, despite concerns that religiously conservative parents may justify child physical abuse on spiritual grounds, this question has rarely been directly studied. In two related studies, the risk of being abusive toward children was related tied to extrinsic religiousness where people report participating in organized religion for instrumental and utilitarian reasons (e.g., gain status, social support). In contrast, researchers have not found higher rates of child maltreatment to be associated with any particular religious tradition, attendance, orthodoxy, or an intrinsic religious orientation where people report religion is central to their identities and well integrated into their personal values.

Relational Spirituality: Coping with Family Problems

Because more religious individuals more often tend to seek out and enjoy harmonious family relationships, they may be especially ill equipped to handle situations that violate their religious beliefs of how to form and maintain family bonds. However, scarce research exists on the ways people may draw on religion to cope with family problems. Examples of situations where specific religious or spiritual beliefs or practices may impact coping include relying on religion to cope with a romantic partner who will not commit to marriage, coping with a nonmarital birth, dealing with infidelity, responding to domestic violence, or serious dysfunction in marital or parent–child relationships. Such stressful situations may trigger spiritual struggles that contribute to personal and relationship distress. Or, religion may be a well-spring of resilience in times of family trouble. Research on ways people use religion and spirituality to cope with divorce illustrates the potential for religion to help or harm when family difficulties occur.

Coping with Divorce

Americans who often attend religious services and couples who are religiously similar are less likely to divorce in the future than others. Nevertheless, significant percentages of people get divorced, including those affiliated with a religious group. For example, one US survey found that approximately 28% of Catholics, 34% of Protestants, and 33% of 'born again' Christians reported having had a prior divorce. Thus, religion is associated with being less likely to have had a divorce, but many people affiliated with a religion do experience divorce. This raises important questions about religious or spiritual coping strategies and postdivorce adjustment. Initial research on this topic has found that divorcing adults as well as college students with divorced parents often interpret this event as the loss or violation of something sacred and experience spiritual struggles about it. Such religious or spiritual distress exacerbates emotional distress. On the other hand, divorcing adults who more often turn to God and a religious community for support are less depressed over time. The picture is less clear for parental divorce. Young adults who recalled trying to use such spiritual resources at the time of the divorce reported current adjustment. National surveys also show that parental divorce triggers switching of denominations or disengagement from organized religion by offspring. Furthermore, the most highly religiously engaged youth are most likely to experience religious or spiritual struggles and psychological distress over time due to their parents' divorce, whereas those with moderately strong ties to religion who are somewhat engaged in a religious tradition appear more able to draw on religious or spiritual resources and thus may experience less distress over time due to parental divorce.

Relational Spirituality: Deepening Our Understanding of Prevention and Intervention

Overall, scientific evidence to date suggests that religion tends to promote the formation and maintenance of family ties, and thus may help to prevent family difficulties. Yet, much empirical work remains to identify specific resources and struggles promoted by religion that could impact family relationships across diverse types of families and societies, especially for subsamples of people who have become embroiled in stressful family situations. Social scientists could productively build more collaborative relationships with diverse people of faith to conduct more basic and applied research on religious or spiritual beliefs and practices that can help or harm family and

kinship bonds during daily life and times of family crisis. In particular, almost no controlled empirical research exists on the efficacy of incorporating religious or spiritual concepts into established prevention or intervention programs aimed at family issues. Thus, while higher involvement in religion appears to lower the risk of various family difficulties, it is far from clear how to translate such basic research findings into practical applications to help prevent family distress and intervene effectively when family difficulties do occur.

See also: Religion and Health: Clinical Considerations and Applications; Religion and Stratification; Religion and Youth; Religiosity; Women in Buddhism.

Bibliography

Denton, M.L., 2004. Gender and marital decision making: negotiating religious ideology and practice. Social Forces 82, 1151–1180.

Edgell, P., 2005. Religion and Family in a Changing Society. Princeton University Press, Princeton, NJ.

Fincham, F.D., Beach, S.R.H., Lambert, N.M., Stillman, T., Braithwaite, S., 2008. Spiritual behaviors and relationship satisfaction: a critical analysis of the role of prayer. Journal of Social and Clinical Psychology 27, 362–388.

Gallagher, S.K., 2003. Evangelical Identity and Gendered Family Life. Rutgers University Press, Piscataway, NJ.

Glass, J., Fitzgerald, S.T., 2014. Conservative protestants, normative pathways, and adult attainment. In: Keister, L., Sherkat, D. (Eds.), Religion and Inequality in America. Cambridge University Press, pp. 97–145.

Mahoney, A., 2010. Religion in families 1999–2009: a relational spirituality framework. Journal of Marriage and Family 72, 805–827.

Mahoney, A., Krumrei, E.J., 2012. Questions left unaddressed by religious familism: is spirituality relevant to non-traditional families? In: Miller, L. (Ed.), The Oxford Handbook of the Psychology of Spirituality. Oxford University Press, Oxford, UK, pp. 165–181.

Mahoney, A., Pargament, K.I., Swank, A., Tarakeshwar, N., 2001. Religion in the home in the 1980s and 90s: a meta-analytic review and conceptual analysis of religion, marriage, and parenting. Journal of Family Psychology 15, 559–596.

Myers, S.M., 2006. Religious homogamy and marital quality: historical and generational patterns, 1980–1997. Journal of Marriage & Family 68, 292–304.

Religion, History of

Friedrich W Graf and Astrid Reuter, University of Munich, Munich, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 13071–13077, © 2001, Elsevier Ltd.

Abstract

This article discusses the way in which scholars, since antiquity, have examined their own and foreign religions and, within this framework, have responded to the pressing cultural and social questions of their day and age. Its aim, therefore, is not to reconstruct the history of religious phenomena as such, but rather the historical developments within the historiography of religion. After an introduction to the latter's terminological history, the article sketches its themes, central concepts, and methods up to the year 2001. This is done as they have emerged from the dialogue between such disciplines as philosophy, theology, philology, history, and ethnology, and with reference to its historically determined cultural, scientific, and institutional contexts. The article closes with a summary of the current trends and perspectives within the historiography of religion, as exemplified by research interests in France, Germany, the UK, and the USA.

Historical Background to the Term 'History of Religion'

The Latin term *historia religionis* first appeared in the sixteenth century and was employed alongside related terms such as *historia ecclesiastica* to entitle chronicles of important events in church history, and the life histories of influential theologians and witnesses of the Christian faith. The term was also applied to accounts on the origin and development of Judaism and portrayals of foreign religions. In his *Historia Religionum* (1697), Gebhard Theodor Meier used the concept in the sense of a comparative study, aimed at portraying a history of the three monotheistic religions. Compared with the more descriptive treatises on religion, comparative *historiae religionis* drew on theological distinctions, such as that between God's chosen people and heathens, to the end of justifying the exclusive truth claims of Christianity. In his *Abris einer Geschichte der Religionspartheyen* (1755), the Protestant theologian and student of Christian Wolff, Sigmund Jacob Baumgarten (1706–57), examined these different methods and aims of religious historiography, explicitly distinguishing between a *usus historicus* and a *usus polemicus*.

From the 1880s onwards we find the introduction of new compound terms of religious historiography which competed with *history of religion*(*s*), *Religions-geschichte* or *histoire des religions* and reflected a heightened interest in scientifically recording and interpreting religious phenomena. The appearance of the terms 'philosophy of religion,' 'science of religion,' 'religious studies,' and – later – 'sociology of religion,' 'psychology of religion,' and 'phenomenology of religion' demonstrated the increasing differentiation within the scholarship dedicated to scientifically interpreting religious culture. However, none of these new academic subjects succeeded in establishing itself as a dominant or authoritative discipline for investigating and interpreting religious systems of belief. Academic research into religion is characterized, rather, by a high plurality of divergent and competitive disciplines that have been institutionalized quite differently, mainly in the countries of Europe and in the USA. This rivalry between disciplines also applies to the historical inquiry into religions: besides professional historians, who sought to render their discipline the new leading human science of the nineteenth century, other cultural scientists and theologians also participated in the historical study of religion and contributed significantly to its advances. As a result, most of the 'classical' works in nineteenth century religious historiography were not compiled by historians but by theologians (or former theologians), philosophers, ethnologists, philologists, sociologists, and professors in the nascent university disciplines, Sinology, Indology, and Egyptology.

In the course of these developments, the term 'history of religion' remained controversial. Friedrich Max Müller (1823–1900), who is regarded as the founder of the 'comparative science of religion' (*vergleichende Religionswissenschaft*), abandoned the term 'history of religion' in favor of the 'science of religion' in order to emphasize the scientific aims of an inquiry directed toward all historically given religious phenomena and traditions, and which strives to emancipate itself from the prejudiced approach of both theology and philosophy. Other researchers of religion, such as Cornelius Petrus Tiele (1830–1902), preferred the term 'comparative religion.' Nevertheless, common to all these different terms and their respective definitions of approach was the motivation to dispense with theologically orientated normative constructions in the study of religion. In Anglo-Saxon countries, this movement found its most important source of support from the so-called 'Lectures,' an institution backed by the private endowments of wealthy patrons (in 1878 Friedrich Max Muller inaugurated the 'Hibbert lectures' and, in 1888, the first 'Gifford lectures' were given). The first professorship for the 'General History of Religion' was established in Geneva, in 1873; in 1877 the universities of Leiden and Amsterdam followed suit. In 1879 France's first chair in *Histoire des religions* was founded at the *Collége deFrance* and in 1886 the Graduate School of Sorbonne, *Ecole Pratique des Hautes Etudes*, instituted its *Section des sciences religieuses*. However, this appearance of several chairs in the 'history of religion' and 'religious studies' did not achieve the full emancipation repeatedly postulated by Christian theologians. In Germany, for instance, most religio-historical chairs were set up within evangelical-theological faculties; and even where this was not the

case and provision made for institutional independence, as in The Netherlands, the majority of scholars appointed to such chairs were theologians.

History of Religious Historiography

The intellectual reflection on religion(s) can be traced back to scholars in Greek and Roman antiquity. Here we find the origination of several important topoi that continue to influence the study of religion (such as the critique of religion, theories on the origin of religion and its social function, the comparison between polytheism and monotheism). Notwithstanding this degree of continuity, the ancient world's approach to the subject of religion cannot be termed 'historical.' Beyond the context of Graeco-Roman traditions, we can also find evidence of the examination of religious life in early Arabic, Persian, and South-East Asian scholarship. However, the influence of these traditions on European historiography of religion has been minimal. One central impetus for broadening Europe's religio-historical horizon was provided by the overseas conquests of the fifteenth and sixteenth centuries. What conquerors and travelers had to report on the foreign religious ideas and practices of the far-off lands they had visited led scholars to question the cultural self-evidence of Christianity, particularly after its confessional pluralization following the Reformation. In an attempt to intellectually overcome religious pluralism as well as the denominational conflicts within Christianity, E. Herbert of Cherbury (1581–1648) focused on a 'natural religion of reason' which he believed to be the true essence of all the different guises of positive religion. In doing so, this British deist introduced a distinction that would later be appropriated by Enlightenment theologians.

However, strictly speaking, it was the seventeenth century that gave birth to the historiography of religion proper. In the wake of the ecclesiastical schism and resultant religious wars, the Christian 'culture of unity' (Troeltsch) began to crumble and was soon replaced by an absolutist state which succeeded in politically neutralizing the vying religious parties by way of an absolute leader ruling over all religious parties. At the same time, the religious conflicts of the seventeenth century gave rise to the interest in a history of religion which was not conceived theologically, but had recourse to an impartial 'third' standpoint beyond confessional polemics. This was found in the 'historico-philological method' based on the principles formulated by Jean Mabillon (1632–1707) in his *De re diplomatica*, which was one of the first comprehensive treatises on methods of documentation and the study of hand-written manuscripts. The publication of this work in 1681 can be seen as the birth of textual criticism that played such a central role in the development of religious historiography. At about the same time, Richard Simon (1638–1712) applied literary-critical methods to the study of the Pentateuch in his *Histoire critique de I'Ancien Testament* (1678) and, in doing so, was one of the first to expose the religious history of the Judaeo-Christian tradition to the process of historicization. This rigorous distinction between the theological and historical historiography of religion also formed the guiding precept of Johann von Gro(e)ing's *Historie der heutigen Religionen* which appeared in Hamburg in 1702. In dismissing every form of *theologia polemica*, this author declared his commitment to a historical approach orientated toward the facts.

That the historian had not irrevocably displaced Christian apologetics, however, is evidenced by Jacques-Bénigne Bossuet's (1627–1704) work *Discours sur l'Histoire universelle* (1681) which appeared in the same year as Mabillon's important treatise and adhered to an interpretation of world history stressing God's saving grace. In opposition to Bossuet, Montesquieu (1689–1755) banished the guiding hand of divine providence from history and declared mankind to be history's focal point. The human being, expounded Montesquieu in *Esprit des lois* (1748), is determined by lots of factors – climate, laws, customs, and traditions as well as religion–the combination of which gives rise to the particular intellectual characteristic of a nation and its underlying 'spirit.' The work of the Göttingen exegete Johann David Michaelis (1717–91) illustrates how Montesquieu's point of departure could be used to widen the religio-historical perspective. According to his own statements, Michaelis sought to view the Old and New Testament 'through the eyes of Montesquieu.' To this end he reconstructed in his six-volume *magnum opus*, *Mosaisches Recht* (1770–75), the structure and contents of the Mosaic codex from the climatic, economical, political, and socio-cultural preconditions of its natural environment, and thereby offered an interpretation of this codex in the spirit of Montesquieu and in reply to the demands of his day.

Alongside this *historico-philological* examination of religion, we also find the appearance of the *psychological* approach to religion, as represented by David Hume (1711–76) in his treatise *The Natural History of Religion*, published in 1757. Hume was not primarily concerned with explaining religion historically, but rather with accounting for religious ideas and practices as phenomena of the human psyche under the dictates of irrational emotions such as fear and hope. Because Hume saw reason as having gradually overcome these irrational forces in the course of history, however, this Scottish philosopher had also furnished the historiography of religion with an evolutionist perspective that would then feature as a key concept in the second half of the nineteenth century.

With the conception of world history as a developmental process from lower to higher, that is to say, from uncivilized to rational forms of life, religion came to represent not an exception to, but a manifestation of these historical dynamics. The evaluation of religion's status within this process of advance was divided: either one adopted a religio-critical stance, and assumed that reason would eventually replace religion (Voltaire, 1694–1778), or one took up a more positive, religio-historical position, and integrated religion into historical progress. One leading exponent of the latter position was Charles de Brosses (1707–77). In his study *Du culte des dieux fétiches* (published in 1760) he employed a comparison between ancient and contemporary African cult ritual to support his claim that fetishism – understood as the worship of arbitrarily chosen natural objects or living creatures – is to be identified as the original form of all religions. However, in the course of mankind's historical advance this has been successively superseded by the worship of celestial bodies, by polytheism, and then finally by monotheistic religions, he contended.

In his *Grundrß der Geschichte aller Religionen* (1787), Christoph Meiners (1747–1810) also construed the history of

religion as part of a general historical process of cultural advance. As a result of this pronounced tendency in the eighteenth century to assimilate the portrayal of religion into an all-encompassing cultural historiography, the traditional anchor of religious history in the theological field of church history began to corrode, causing a telling shift of emphasis. Johann Christoph Gatterer (1727–99), for instance, dealt with church history within the framework of a universal history of religion. In a similar vein, Christian Wilhelm Flügge approached church history as an aspect of religious history that he conceived as a branch of the general cultural history of the human spirit.

Resistance against this linear and optimistic idea of historical progress had already begun to emerge in the eighteenth century. For Jean-Jacques Rousseau (1712–78), scientific and technical progress amounted to a moral step backwards behind the original and good 'natural state.' Even religion, he argued, has not escaped this cultural decline, but degenerated from an honest 'affair of the heart' into an institutionally depraved ecclesiasticism no longer capable of integrating the community. Nonetheless, because Rousseau considered religious feeling to be the strongest social bond, in *Contrat Social* (1762) he spoke up for a religious 'occupation' of the values which hold together the political community, viz. the establishment of a 'civil religion.'

Similar ideas were to be voiced by Alexis de Tocqueville (1805–59) almost a century later who, in his famous work *De la démocratie en Amérique* (1834/40), prognosticated that the a politico-moral consensus within democratic societies would adopt the socially integrative role which religion had played in the social system of the Middle Ages. In eighteenth-century Germany, Johann Gottfried Herder (1744–1803) likewise objected to the belief in progress of Enlightenment thinkers, supporting the Romantic conception of history as a process of organic growth. In doing so, Herder influenced the organo-historical ideas of classical German historicism. This school of thought also rejected the Enlightenment's linear concept of history and belief in universal-historical progress, viewing history as the unfolding of individual cultural units.

The turn of the century into the 1900s was also a transitionary period for European historical consciousness which had begun to move away from universal generality and attach greater importance to that which is particular and individual. This change of perspective not only enhanced the status of the historical religions, which had been neglected by the concepts of historical progress of the eighteenth century, but also paved the way for the empirical research into the history of religion, especially the intensive analysis of ancient religious traditions.

Among others in Germany, Otto Jahn (1813–69) and Karl Otfried Müller (1797–1840) devoted their scholarly activities to investigating ancient mythology. In his influential study *La cité antique* (1864), the French historian N.D. Fustel de Coulanges (1830–89) examined the stabilizing function of religion for the Greek polis and Roman system of rule. Furthermore, this new interest in empirical research into religion led to what Edgar Quinet described in his work *Génie des religions* (1842) as an 'Oriental Renaissance': the first translations of India's holy scriptures had already begun to appear at the end of the eighteenth century.

In 1808, Friedrich Schlegel published his writings on the language and wisdom of India (*Über die Sprache und Weisheit der Inder*) whose conception of religious history diametrically opposed that of the Enlightenment. In India's ancient wisdom, Schlegel saw the realization of a unity between nature and spirit (Geist) which 'enlightened Europe' had forfeited and India itself submerged, and which he strove to reconstruct from the oldest of India's religio-historical sources. Drawing on Schlegel, Georg Friedrich Creuzer (1771–1858) and Joseph Görres (1776–1848) called into question the origins of antiquity's mythical tradition, claiming that Greek myths originally derive from the wisdom of Indian priests. With this Romantic enthusiasm for oriental languages and literature, the availability of translations of the latter and, not least, the advances in the field of 'historical and comparative linguistics since the turn of the nineteenth century,' the historiography of religion in the nineteenth century was afforded a wide range of new stimuli.

The endeavors of Franz Bopp (1791–1867) to distill a common Indo-European proto-language from a grammatical comparison between Sanskrit, Greek, Latin, Persian, and German, for instance, gave the comparative study of religions new impetus, temporarily pushing universal interpretations of history into the background. Aside from the work of Eugene Burnouf (1801–52), a prominent figure in the historiography of Iranian religion, it was above all the services of his, and Bopp's student, Friedrich Max Müller, which rendered the philological approach beneficial to religious historiography.

Analogous to the Romantics' interest in the origins of religious history, Müller linguistically classified the Indo-European languages, particularly with reference to their respective names for God, to the end of reconstructing a primal religion common to all these peoples, which he saw as having once flourished in the advanced civilization of ancient India. Hermann Usener's (1834–1905) investigations into ancient religious history were also based on a combination of philologico-comparative studies and theoretical-explanatory ambitions. With his programmatic work on an interdisciplinary ancient religious history – which his students Eduard Norden (1868–1941) and Ulrich von Wilamowitz-Moellendorff (1848–1931) advanced – Usener not only influenced the study of ancient religion, but also classical philology, art history and oriental studies, not least cultural theory until well into the twentieth century.

Besides the philological exploration of religious traditions, which chiefly focused on antiquity and the Orient, the second half of the nineteenth century also witnessed the proliferation of 'ethnological' research into contemporary, non-European religions which had not been taken into consideration by the romantically inspired religious historiography of such scholars as F.M. Müller. That these religions had become the focal point of attention within nineteenth-century religio-historiographical discourse was due to a revival of the evolutionist theory, aimed at assimilating the individual, historical, and divergent forms of religion into a universally applicable model of cultural advance.

In his major work *Primitive Culture* (1871), Edward Burnett Tylor (1832–1917) held the view that the religions of contemporary tribal societies represented the beginnings of religious history and therefore attached great importance to the investigation of such communities. However, Tylor's claim that animism (the attribution of a living soul to plants, inanimate

objects, and natural phenomena) constituted the primal form of religion was soon superseded by other theories on the origins of religion, the most influential of which was William Robertson Smith's (1846–94) totemism theory tracing religion back to the worship of ancestors symbolized by totemic emblems of plants and animals. Emile Durkheim's (1858–1917) writings on the elementary forms of religious life *(Les formes élémentaires de la vie religieuse*, 1912) can be seen as the most prominent adaptation of Smith's propositions. Religio-historical evolutionism proved to be an extremely resilient paradigm, providing the working hypothesis for James George Frazer's (1854–1941) reconstruction of religious history on the basis of contemporary ethnological material in his famous twelve-volume opus *The Golden Bough* (1890–1915).

In the latter half of the nineteenth century, ethnological and philological investigations into the history of religion found enormous support within academia and aroused ever greater interest among the general public. This trend was evidenced in the UK by the positive response to lectures in this field of study (cf Section Historical Background to the Term 'History of Religion') and the huge popularity of Ernest Renan's (1823–92) *La vie de Jesus* (1863) in France and David Friedrich Strauss' (1808–1974) works *Das Leben Jesus* (1835/36) and *Der alte und der neue Glaube* (1872), as well as Hermann Oldenberg's book *Buddha* (1890), in Germany. *Historical* scientists, by contrast, showed very little interest in religio-historical phenomena; the majority of professional historians adhered to a narrow politico-national approach to their subject in which religio-historical questions remained peripheral.

The few exceptions, held in high esteem today, were then decisive outsiders, such as Jules Michelet (1798–1874), who in his 17 volumes on the history of France *(Histoire de France*, 1833–67) drew attention to the religious folklore of the Middle Ages and, in doing so, became a predecessor of the *histoire des mentalités* (cf Section Overview of Recent Research), or Jacob Burkhardt (1818–97) who in *Weltgeschichtliche Betrachtungen* (1868–73) and *Die Kultur der Renaissance in Italien* reflected on religion as a third historical 'power' (*Potenz*) alongside culture and politics.

The cultural-historical revival at the turn of the twentieth century effectively placed religion at the forefront of scholarly discourse throughout Europe and the USA. The rapid erosion of time-honored cultural traditions, the growing insight into the social costs of capitalist modernization, and the new disintegrative tendencies in many Western societies provoked heated debates on the relationship between the religious history of European countries (as well as that of the USA) and their individual formation processes as modern societies. In defiance of traditional disciplinary borders, the historical research into religion became a central medium for the self-reflection of Western societies. From a systematic perspective, three closely connected issues came to play a decisive role in this intensive exploration of religious history around 1900. These issues included: the participation of religious traditions in the genesis of modern occidental rationalism; the socially integrative potential of religion in functionally differentiated and politically fragmented societies; and the cultural relevance of differences between Christianity's denominations on the one hand, and non-Christian religions, especially those of the East, on the other.

First and foremost, this discourse applied to Protestant theology which had largely viewed itself as a historico-cultural science of Christianity since the mid-eighteenth century and, since the 1880s, became increasingly aware of the interrelations between the Christian religion and its non-Christian environment. Above all, it was the proponents of the so-called 'History-of-Religions School' (*Religionsgeschichtliche Schule*) within German Protestant theology who, in classical exegetic and religio-historical treatises, analyzed the interconnections between Jewish antiquity and the religious worlds of the ancient Orient, and the syncretistic character of Christianity. Although religio-historically minded scholars of the Old and New Testament, such as Hermann Gunkel (1862–1932) and Wilhelm Bousset (1865–1920), as well as the systematic theologian and cultural philosopher Ernst Troeltsch (1865–1923) – named the 'systematist of the History-of-Religions School' – viewed 'religion' as a relatively autonomous cultural power *sui generis*, they nevertheless gave greater priority to investigating the interactions between religious and other spheres of culture. Ernst Troeltsch's 'The Social Teaching of the Christian Churches' (*Die Soziallehren der christlichen Kirchen und Gruppen*, 1912) is regarded as the most important work of this religio-historical group. In this publication, Troeltsch embarked on a cultural-historical reconstruction of the formative influence of Christianity's denominational churches and sects on social ideas and the effects of social structures on the formation of religious ideas and communities. Appropriating the typological distinction between 'church' and 'sects,' developed by his friend Max Weber (1864–1920), Troeltsch differentiated between 'ecclesiasticism,' 'sectarianism,' and 'mysticism' as specific forms of religious socialization. With Weber, Troeltsch shared the ambition to demonstrate the significance of religious ideas and practices for the emergence of the modern world. In *Die Bedeutung des Protestantismus für die Entstehung der modernen Welt* (1906), he focused on Protestantism and Calvinism, in particular.

While Troeltsch transformed church history into the cultural and social history of Christianity, Weber expanded his pioneering study 'The Protestant Ethic and the Spirit of Capitalism' (*Die protestantische Ethik und der Geist des Kapitalismus*, 1904/05) into an extremely elaborate enterprise of categorically comparing the economic ethics of the world religions (*Die Wirtschaftsethik der Weltreligionen*, 1915). In 'The Protestant Ethic,' Weber had inquired into the effects of a religiously shaped *habitus* in areas of life and culture far removed from those commonly associated with religion in the narrower sense of the term – above all, in the economic sphere. In his comparative analyses of the economic ethos of the world religions, however, he primarily concentrated on occidental rationalism and sought to grasp its particularity and formative power on cultural life. Today, scholars are more skeptical of the details of Weber's ideas. Nonetheless, it cannot be denied that the worldwide reception of his theories and derivations in the study of religion has scarcely been matched by any other oeuvre.

Overview of Recent Research

Since the 1960s, the renaissance of religious movements in many societies, the upward trend of 'fundamentalism' and

Protestant sectarianism, the spread of charismatic Christianity, and the manifold processes of a rapidly changing religious landscape (conversions, new syncretism) have attracted the attention of historians and led to a critical revision of the 'secularization theory,' which equated social modernization with a general decline in religion. Religion has been rediscovered as a relatively autonomous system of communication and symbolism which shapes the *habitus* of individuals or identity constructions of social groups, and exerts a positive or negative influence on the development of societies. As a result, the old debates on the definition of religion and master narratives on modernization de-Christianization or secularization are being steadily replaced by micro-studies into the diverse religious milieus of the modern age. While religion has become a central topic in the historical discourse of all countries generally, several national-cultural particularities deserve mention.

In France, the work of Lucien Febvre (1878–1956) and Marc Bloch (1886–1944), together with the journal *Annales, Histoire, Sciences Sociales* (originally entitled *Annales d' histoire économique et sociale)* they founded in 1929, supplied the historiography of religion with an innovative approach which broke with the politico-historiographical tradition of the nineteenth century. Their objective was to uncover a temporal stratum of *longue durée* (Fernand Braudel, 1902–85), that is to say, extensive periods of time which formatively shape not only political and socio-economical history, but also the concrete experiential world of the general population. Concepts stemming from *Annales*-historiography, such as 'structure' and '*longue durée*,' were then applied heuristically to the broader discussion of collective 'mentalities,' from which the historiography of religion profited in particular.

Although the term *histoire des mentalités* soon became a key concept of *Annales*-historiography, no agreement had been reached on the question as to what 'mentality' exactly means. These definitional problems notwithstanding, *histoire des mentalités* can be generally described as an historical discourse directed toward the collective consciousness of an epoch, i.e., its cognitive, ethical, and affective dispositions, which form the unquestioned cultural basis for the daily life of individuals and social groups, thus constituting their identity. In its application to the history of religion, this approach initiated a change in perspective away from the narrow argumentation of ecclesiastical historiography: religious history came to be seen as a history of religious consciousness whose carriers remain for the most part anonymous. Following this, historians of religion in France began to attach less importance to major events, personalities, and institutions, and focus on 'structural-historical' and, in particular, 'anthropological' lines of questioning.

Here their investigations have been concerned with historically varying attitudes to birth and childhood, mortality and death, corporeality and sexuality, nature and the environment, God and the Church, heaven and hell (cf Ph. Ariés, M. Vovelle, J. Le Goff, J. Delumeau). To date, however, exponents of *histoire des mentalités* have scarcely researched the religious culture of more recent history, concentrating rather on the Middle Ages and early Modern Age, and – in co-operation with experts in ancient history such as J.-P. Vernant, M. Detienne and P. Vidal-Naquet since the1970s – on the religious history of antiquity. Further drawbacks of applying this historical approach to religious historiography arise from its emphasis on the alleged durability of mental dispositions, its disregard of the processes of change within religious mentalities, and its neglect of the conflicts caused by differing religious backgrounds (resulting from the acculturation of foreign ideas and practices or due to inner-social differences).

In Germany, the 'socio-historical turn' of the 1960s and 1970s as well as the institutional separation of church history (pursued within theological faculties) and general history encouraged scholars to view religion and confession as nothing other than marginal factors in the history of the modern age. Since the 1970s, however, several attempts have been made to integrate religio-historical perspectives into social historiography (cf W. Schieder, R. van Dülmen, R. von Thadden) and to open up historical research in church history and theology to the methodological debates of cultural and social scientists (cf M. Greschat, K. Nowak, F.W. Graf). Numerous younger historians are now conducting research into a range of religious milieus, such as ecclesiastical organizations and festivities, theological controversies and quarrels over religious symbolism as well as the tense coexistence of Catholics, Protestants, and Jews. The collapse of East Germany, moreover, has generated a new sensibility for the political role of churches and their partly antagonistic, partly affirmative relationship to the 'political religions' of the twentieth century. And last, but by no means least, the holocaust initiated worldwide intensive research into the cultural climate of German Judaism and the conflict-ridden history that finally led to the destruction of the 'German-Jewish synthesis.'

In the UK, discussions continue to be determined by the weight of social anthropology (cf B. Malinowski, A.R. Radcliffe-Brown). Here, from the outset – unlike in Germany – 'society' and 'culture' were seen as one, and consideration given to the class-specific religion of the workforce and the processes of religious change in this country's urban centers (cf H. McLeod). Particularly pioneering were the studies on the connection between religious radicalism and revolution (cf Ch. Hill) and between Methodism and the emerging labor movement (cf E.P. Thomson). Research into the processes of religio-theological differentiation in the Anglican churches has also drawn on the interpretative approach of *histoire des mentalités.*

In the USA, historians of religion since the 1980s have been strongly influenced by the 'interpretative' ethnology (stressing the concept of 'understanding' rather than that of 'explanation') and cultural anthropology of C. Geertz. Beyond the classical self-reflective discourse on the religion of native Americans, the faith of the Pilgrim Fathers, pluralism within Protestant denominationalism, the emergence of a specifically American Catholic self-understanding in the late nineteenth century and a 'civil religion' (N. Bellah) capable of integrating religious pluralism, in the most recent American historiography of religion themes such as gender, race, language, and the religious forms of minority groups have played an important role. In addition, the ethnological debate on 'Writing Culture' (viz. the constructive character inherent in self- portrayal and the description of that which is foreign; cf F. Clifford/G.E. Marcus), the comparative analysis of fundamentalist movements in the

world religions (M.E. Marty), the interpretation of Latin America's religious developments (D. Martin) and the interrelation between law and religion (W. Fallers Sullivan) have featured as central concerns of recent research. Current North American religious historiography has also profited greatly from the application of eco-nomico-theoretical approaches (Stark and Bainbridge, 1985; Stark and Finke, 2000).

Outlook

Religion has gained in importance worldwide and this situation has accordingly afforded the research into religion greater significance. At the forefront of this discourse we find the questions concerning the inter-relatedness of religion and ethos, the political relevance of religion, the construction of religious identities of minority and the conflicts between members of different religions and confessions. The debates on migration, religious conversion and new syncretism have shown, however, that historians of religion find it difficult to precisely grasp the long-term changes of religious mentalities. The correlation of conflict-ridden modernization, on the one hand, and the renaissance of strictly binding religion on the other, promises to become an issue of particular concern in the future: for the rapidly advancing globalization of capitalism, together with its constraints on heterogeneity, will only fuel those religious movements intent on communicating a more fulfilled form of human life which transcends an existence solely defined by the utilitarianism of economic rationality.

See also: Christianity in Latin America; Christianity in the English-Speaking West; European Religion from the Atlantic to the Black Sea; Historical Thought and Historiography: Jewish Tradition; Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Judaism; Near Middle East/North Africa Studies: Religion; Religion: Evolution and Development; Religions of East Asia; Religions of India.

Bibliography

Despland, M., 1979. La Religion en Occident. Evolution des Idées et du Vécu. Fides, Paris.
Dubuisson, D., 1998. L'Occident et la Religion. Mythes, Science et Idéologic. éditions Complexe, Brussels, Belgium.
Feil, E., 1986. 1997 Religio, 2 vols. Vandenhoeck & Ruprecht, Göttingen, Germany.
Hjelde, S., 1994. Die Religionswissenschaft und das Christentum. Eine historische Untersuchung über das Verhältnisvon Religions-wissenschaft und Theologie. Brill, Leiden, Germany.
Idinopulos, T.A., Wilson, B.C. (Eds.), 1998. What Is Religion? Origins, Definitions, and Explanations. Brill, Leiden, Germany.
Jastrow, M., 1981. The Study of Religion. Reprint. Scholars Press, Atlanta, GA.
Jordan, L.H., 1986. Comparative Religion. Its Genesis and Growth. Reprint. Scholars Press, Atlanta, GA.
Kemper, D., 1992. Religionsgeschichte I. In: Ritter, J., Gründer, K. (Eds.), Historisches Wörterbuch der Philosophie. Schwabe, Basel.
Kippenberg, H.G., 1997. Die Entdeckung der Religionsgeschichte. Religionswissenschaft und Moderne. Beck, Munich.
Lehmann, E., 1925/[1887–1889]. Zur Geschichte der Religionsgeschichte. In: Bertholet, A., Lehmann, E. (Eds.), Lehrbuch der Religionsgeschichte, fourth ed. Mohr (Siebeck), Tübingen, Germany.
McCutcheon, R.T., 1997. Manufacturing Religion. The Discourse on Sui Generis Religion and the Politics of Nostalgia. Oxford University Press, Oxford, New York.
Molendijk, A.L., Peter, P. (Eds.), 1998. Religion in the Making: The Emergence of the Sciences of Religion. Brill, Leiden, Germany.
Pailin, D.A., 1984. Attitudes to Other Religions. Comparative Religion in 17th and 18th Century Britain. Manchester University Press, Manchester, UK.
Pinard de la Boullaye, H., 1922–1925. L'Étude Comparée des Religions. Essai Critique. G. Beauchesne, Paris.
Platvoet, J.G., Molendijk, A.L. (Eds.), 1999. The Pragmatics of Defining Religion. Contexts, Concepts and Contests. Brill, Leiden, Germany.
Puech, H.C. (Ed.), 1970/1972/1976. Histoire des Religions. Encyclopédie de la Pléiade, 3 vols. Gallimard, Paris.
Rudolph, K., 1992. Geschichte und Probleme der Religions-wissenschaft. Brill, Leiden, Germany.
Saler, B., 1993. Conceptualizing Religion. Immanent Anthropologists, Transcendent Natives, and Unbounded Categories. Brill, Leiden, Germany.
Schlesier, R., 1994. Kulte, Mythen und Gelehrte. Anthropologie der Antike seit 1800. Fischer, Frankfurt, Germany.
Sharpe, E.J., 1997/1975. Comparative Religion. A History, fifth ed. Open Court, Chicago and La Salle, IL.
Smart, N., Clayton, J., Sherry, P., Katz, S.T., 1985. In: Nineteenth Century Religious Thought in the West, 3 vols. Cambridge University Press, Cambridge, London, New York.
Stark, R., Bainbridge, W.S., 1985. The Future of Religion: Secularization, Revival, and Cult Formation. University of California Press, Berkeley, CA.
Stark, R., Finke, R., 2000. Acts of Faith. Explaining the Human Side of Religion. University of California Press, Berkeley, CA.

Religion: Mobilization and Power

Darren E Sherkat, Southern Illinois University, Carbondale, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Percy, volume 19, pp. 13077–13082, © 2001, Elsevier Ltd.

Abstract

Religion has always been a strong force for collective mobilization that influences control of state resources. In earlier eras, religious control over the state was direct, and there was little differentiation between the dominant religious institutions and state power. At the same time, alternative religious expressions and institutions were brutally repressed, and mobilization in favored religious groups made compulsory. The differentiation of religious and political institutions in modern societies has meant that religious groups seek power through the political process, and they are often advantaged by having premobilized resources and a common cultural orientation motivating their pursuit of collective goods. Yet, in diverse societies the collective goods sought by one religious movement often harm members of other religious groups or people who reject religion.

Religion has always been a powerful social institution with the capacity to mobilize resources of societies and use these resources to further particular ends. The way that religion obtains and maintains power – defined as control over collective resources – varies over time and space. Prior to the rise of modern, capitalist nation-states, religion and the state were unified, which provided considerable power for dominant religious traditions and minimal power for minority faiths. In most modern societies, religion has become separate from the state, and religious movements wield power by mobilizing their adherents in the political process.

Religion is a social institution that provides explanations about the supernatural (Sherkat, 2005). The collective nature of religions allows religious institutions to accumulate resources – temples, priests, treasures, and adherents are all physical resources enjoined in religious institutions. Religions also produce potent ideological resources that can legitimate collective action, define collective goods, and motivate the use of social resources (Zald, 2000). Religious ideologies are readily transposed into the political realm, and can justify economic activities and relations of production (Sherkat, 2005). Religious ideological resources are both extensive – they have considerable appeal across time and space – and intensive – they enable control over understandings and behaviors (Young, 2002). Religious schemata are almost always the most powerful ideological elements in any culture, and they help to define collective goods (and collective insults), and direct the propriety of collective action (Snow et al., 1986). Classical treatments often assumed that otherworldly religion would stifle political mobilization; however, contemporary research has shown that religious beliefs, values, and institutions are crucial for supporting contentious tactics (Sherkat, 2005; Zald, 2000; Young, 2002).

Religious Mobilization and Power in Precapitalist States

Before the rise of the modern nation-state, religious institutions were directly involved in the administration of the state and in economic production. Religious institutions were the principal supplier of education and cultural goods (music, art, literature), and they actively funded public works projects, such as roads, bridges, and waterworks. Agents of religious institutions helped codify and adjudicate laws, and they also often served as tax collectors and accountants of public resources (Weber, 1978). Religious institutions from most of the major religious traditions (particularly Christianity, Islam, Hindu, and Buddhism) also controlled large landholdings and were important economic engines for the production of food and other commodities. The overlapping structures of church and state gave religious institutions incredible power in precapitalist societies.

Because of the role of religion in precapitalist societies for the state and economy, religious movements among elites were often sources of political conflict. Religious movements could seek greater power through adjudicating laws in a particular way to favor certain economic or political interests, and they could lobby for alternative taxing arrangements. Yet, radical religious movements that bucked established procedures could run the risk of being branded heretical and squelched. Further, the alignment of religious institutions to existing political and economic arrangements tied the dominant religious groups to the precapitalist political economy (Braudel, 1982). As political economic systems rose and fell, so too did the power of particular religious institutions.

The alignment of dominant religious institutions with political and economic elites often meant that the religious desires of the peasantry were ignored (Stark and Finke, 2000). Peasants were often largely unaware of the supernatural explanations produced by dominant religious institutions, and the masses developed their own folk religions. Sometimes these religious expressions were also violently repressed as heretical movements, as in the frequent witch-burning crusades in Europe, and the repression of Sufism and Mahayana Buddhism in the East (Jensen, 2007). However, often religious minorities were tolerated in precapitalist societies, and many minority religious groups came to fill vital roles in politics and the economy as capitalism developed (Braudel, 1982). This is particularly true for Armenian Christians and Jews in various

societies, Jain and Zoroastrians in India, and Copts in Egypt. As capitalism developed, many of these outsider religious groups gained economic power and pushed for political advantages, and this was particularly effective in Islamic nations where religious strictures prevented Muslims from engaging in finance or brokering business deals that might conflict with Islamic law (Kuran, 2010; Poulson, 2006).

Religion and Social Movements in Modern Societies

The differentiation of political, economic, and religious institutions had a profound impact on how religious institutions engage the state, and how religious groups mobilize resources for their own operations and to impact broader society. The separation of dominant religious organizations from political control of the state and control over the means of economic production secularized political and economic authority (Chaves, 1994); however, this also freed religious institutions from the shackles of political conflict and economic vicissitudes. Disestablishment allows religion to be removed from political and economic problems, and a more diverse set of religious expressions have the opportunity to mobilize as the dominant religious institutions lose the power to repress religious alternatives (Stark and Finke, 2000). Further, developing states have a strong incentive to legitimate their authority using religious institutions (Gill, 1998), perhaps especially when nations are plagued by glaring economic and political inequalities across class and ethnic groups.

The loss of dominant religious groups' political and economic privileges spurred the political mobilization of religious resources. As modern polities began to use the collective resources of the state to produce collective goods like public education, care for the poor and infirmed, adjudicating marriage, divorce, and inheritance, and redistributing lands to the former peasantry, religious movements from the dominant institutions often opposed these efforts (Jurgensmeyer, 2008; Gill, 1998). Religious groups also mobilized to press for the political regulation of free markets for capitalist goods, and religious mobilization has been most animated about regulating the production and sale of proscribed commodities (alcohol, beef, pork, and other drugs), and cultural goods deemed blasphemous or profane (literature, art, music, theater, film) (Sherkat, 2005). Many nations actively tried to reduce the power of feudal religious institutions through anticlerical laws and other proscriptions, such as bans on religious clothing and mandatory secular education. In the nineteenth and early twentieth century such moves were met with resistance by the formerly dominant religious traditions – whether Catholic, Islamic, Buddhist, or Hindu. However, many current controversies over the secularization of the public sphere have mobilized the faithful from minority religious groups, such as sectarian Protestants in the United States, or predominately Muslim immigrant religious groups throughout Europe.

One key focus of religious mobilization has been the regulation of sexuality, particularly women's sexuality. Religious groups from many traditions lobby nation-states to inhibit women's movement, proscribe women from political participation, promote sex segregation in schools and public places, prevent access to contraception and abortion, and criminalize nonmarital sexuality (including premarital, extramarital, and homosexual sex). Religious groups also mobilize resources to ensure that the state promotes their preferred style of dress, usually with the goal of reducing women's sexual expression. Religious movements also lobby modern states to maintain their right to impose forced marriage, and to prevent intermarriage across religious or caste groups. The issue of same sex marriage has created considerable mobilization among religious groups in Western democracies in the last decade, and notably this often pits liberal religious groups and secular movements against varied conservative religious groups who oppose extending marriage rights to same-sex couples (Sherkat et al., 2011). The regulation of sexuality is a prominent feature in religious movements seeking to influence cultural markets like literature, film, and the Internet.

Education is another issue that fosters mobilization of religious resources in the political realm. While feudal societies granted the dominant religious groups control over most educational instruction, advances in science and cultural production gave modern states a stake in making sure that educational attainment was widely available and attentive to scientific and cultural progress. The development of widespread or universal public education presents a clear threat to the hegemony of religious institutions, since scientific discoveries and theories call into question supernatural explanations for the origin and purpose of life, and religious narratives regarding the role of the supernatural in this world (e.g., faith healing, miracles, and explanations for catastrophes). Further, teaching about cultural advances in the humanities and social sciences often grates against the hegemony of religious particularisms, exposing students to alternative values, discourses, and religions, and pointing to flaws and contradictions in traditional beliefs, understandings, and aesthetics (Sherkat, 2005). Religious movements have often been quite successful in co-opting public education, particularly by ensuring that dominant religious traditions are placed into public school curricula, and/or by lobbying the state to support independent religious instruction through religious schools.

Religious movements also have a powerful effect on defining the boundaries of nation-states, and determining citizenship rights. Religious movements throughout the world have been central for motivating political upheavals and revolutions, and have been critical for delineating the geography of states emerging from colonial rule – perhaps the most dramatic example being the division of British colonial control into Pakistan, India, Sri Lanka, and Bangladesh. Within the borders of nations, religious movements often seek to control who is considered a citizen of the state and given privileges of enfranchisement, the right to hold political office or state supported employment, or enjoy collective goods such as education and health care. Religious movements are critical for justifying political exclusion of religious minorities, and this is quite evident in places like Israel, Islamic Iran, and Apartheid South Africa (Dubow, 1992; Sherkat, 2005). Often, religiously motivated political exclusion is furthered by creating restrictive immigration policies that prevent people from alternative religious traditions from entering the nation, and encourage

immigration by coreligionists (as in Israel, Saudi Arabia, and the United States before 1965).

Religious groups also mobilize to maintain their privileged ability to mobilize resources, often by encouraging repression against alternative religious groups (Stark and Finke, 2000; Gill, 1998). Even in many highly secularized nations, state-sponsored religious groups have been granted the privilege of state supported resource acquisition through the collection of religious taxes. Maintaining this requires constant mobilization by religious elites to ensure that the franchise is not lost or divided out to competing religious groups. Even in societies like the United States, where state-sponsored religion ended in the early nineteenth century, religious groups lobby for important advantages like exemptions from taxation for property and religious professionals, discounted mail, and diplomatic support for missions. Notably, in most Western democracies the power of religious movements remains so strong that such privileges are relatively unchallenged in the political process.

The Political Mobilization of Religious Resources

The collapse of feudal and colonial societies placed religious groups at a competitive advantage for mobilizing resources in modern nation-states, and the organizational structure of religious institutions played a strong role in helping religious groups engage the state. Because religion is a collective pursuit, religious institutions enjoyed the benefit of preexisting social networks, as well as the cumulative advantage of premobilized collective resources. In contrast, secular political movements based on social class, regional economic interests, or philosophical orientations would have to recruit members *de novo* and coax them to provide resources for the movement.

Religious groups bring people together for regular ritual interactions, and these interactions take place in defined settings controlled by religious movements and usually staffed and directed by professional religious cadre workers (McCarthy and Zald, 1977). Because all modern religious groups have enjoyed a fairly long period of institutional development, they have accumulated substantial resources, which can be committed to obtaining political advantages. And, the rise of democratic political systems favors religious groups that have access to large numbers of adherents who could be mobilized to co-opt or form political parties, run for political office, and participate in the political process.

At the individual level, religious mobilization is advantaged because social networks among coreligionists provide built-in selective incentives for participating in political movements (McCarthy and Zald, 1977). Most important in these are the sense of solidarity and camaraderie that is fostered in religious groups, which can motivate people to volunteer for political and religious causes. Further, those same networks can enforce norms of participation by ostracizing or otherwise sanctioning free riders. If a political cause is sanctioned by religious institutions, then failing to participate marks one as an infidel and causes a loss of social standing. The importance of religious groups for fostering social incentives for political participation is especially acute in nations with considerable religious diversity and for religious groups with little political power. Powerful religious groups are able to garner resources directly from the state, but minority religious groups or those that lack power must mobilize all of their resources directly from constituents.

The norm of giving time, money, and goods to religious institutions provides a stable stream of resources for religious groups, and these resources accumulate over time, even for religious groups that serve disadvantaged minority groups. The gradual development of segregated African-American Christian denominations in the nineteenth and twentieth century were pivotal for providing resources for political movements seeking civil rights for African-Americans (Morris, 1984). Even for this oppressed minority, regular contributions from individuals gradually led to the establishment of well-funded denominations like the National Baptist Convention and the African Methodist Episcopal Church. Religious groups often become critical indigenous institutions for oppressed populations, and they can provide both tangible and ideological resources to promote political and economic gains for their constituents (Billings, 1990).

Religious organizations have control over valuable physical resources, including buildings, schools, temples, printing presses, cars and buses, and communication equipment, such as radio and television studios. Religious schools and temples provide control over space where congregants and leaders can mobilize for political action, plan strategies for action, and retreat from repression. Schools and temples also serve as centers for communicating religious narratives, and these are often fused with the framing of political issues (Poulson, 2006; Morris, 1984). Schools for religious professionals are also critical for mass mobilization potential. First, seminaries for religious professionals are less likely to come under the control of the state, and it is a direct assault on a religious group if a political movement seeks to take over religious schools. Second, seminaries promote social network ties among religious professionals, and they forge a web of like-minded ministers who could be called upon to deliver congregational resources in the future. Third, schools for religious professionals are distinctive in that they train individuals to become leaders of organizations where the mobilization of resources is crucial for day-to-day functioning. Priests, ministers, imams, and gurus are professional social movement cadre workers, and since religious movements have political interests these professional activists are ready-made leaders for political movements.

Religious Ideologies and the Framing of Political Issues

Religious ideological resources are a definitive source of power for religious institutions, and religious understanding help shape the framing of political issues. Religious narratives allow social movements to define problems in sacred terms, and to link solutions and potential outcomes to prophetic narratives that resonate with potential constituents. Religious framing also helps social movements to couch the opposition in an unfavorable light, and to justify extreme tactics against the state or other social movements (Jurgensmeyer, 2008). Because religious narratives are widely understood and often strongly

held (Young, 2002), they serve as a useful tool for mobilizing adherents and for legitimating policy.

Sacred texts from all religious traditions contain narratives that explain how human societies operate in the real world. Often such narratives explain the causes of human suffering, and propose supernatural solutions, which also often have material consequence. Because religious narratives are fluid and can be interpreted through the lens of distinctive faith commitments, the transposition of religious schemata into the political realm is often contest. Indeed, some of the most serious religiously motivated political conflicts arise within religious traditions (Sunni vs Shi'ite in Islam, and Catholic vs Protestant in Christianity). Narratives of suffering and subjugation are juxtaposed with theodicies of release and redemption, and these are commonly used to motivate mobilization.

For example, in the Abrahamic tradition, the story of bondage in Egypt is resolved through the Passover narrative where an angel kills Egyptian children, and Moses leads the faithful out of Egypt after Pharaoh's army is destroyed by God. Such narratives provide political movements with a divinely inspired hope for emancipation, and provide divine sanction for genocidal treatment against opponents. The messianic prophesies of many religious traditions lend a similar justification for establishing theocratic rule and fulfilling the revelations (Jurgensmeyer, 2008). Other narratives point more explicitly to longstanding inter- or intrareligious conflict, such as the veneration of the martyrdom of Hussein in Shi'ite Islam, which provides a resonant narrative to their claim to legitimate rule and the illegitimacy of the Sunni Caliphate (Poulson, 2006). Religious framings help to amplify beliefs about the seriousness of political issues, the identification of unsavory targets for blame, and also about the propriety of mobilizing for political action (Snow et al., 1986). The issue of the propriety of collective action often requires theological innovation for otherworldly religious traditions, since mobilizing for political gain in this world runs contrary to religious emphases on otherworldly salvation (Morris, 1984; Sherkat, 2005).

Current Trajectories of Religious Mobilization and Power

Classical theories of secularization predicted that religious disestablishment would lead to a decline in the importance and power of religious institutions; however, the formal disconnection of religion from politics and economics instead has made religion a key player, particularly in developing democracies and postcolonial societies. Even in highly secularized nations where religion had lost power, immigrant religious groups, particularly Muslim immigrants, have mobilized to garner additional rights like religious schools and legal systems, and to confront threats against religious particularisms, such as antiveiling laws and building codes that prevent the construction of mosques (Just et al., 2014). Indeed, after the fall of the Soviet Union there was a flourishing of religious mobilization among Orthodox Christians and Muslims (Froese, 2008), and religiously inspired political movements have given considerable power to religious groups throughout the former Soviet Union. The rise of a politically powerful Russian Orthodox Church and political Islam in many former Soviet Republics has resulted in the harsh repression of minority religious movements and secular movements. Similar religiously inspired regimes have come to power in many nations throughout the world and across a variety of religious traditions, Buddhism in Southeast Asia, Hinduism in India, Islam in much of Africa and Asia, and Christians in parts of Africa and North America (Jurgensmeyer, 2008). Powerful religious movements from disparate traditions continue to foment devastating violent conflicts throughout Africa, the Middle East, and Southeast Asia.

The political mobilization of religion remains substantial, and the power of these movements to impact social life is considerable in most parts of the world. Further, in a globalized society, religiously inspired political movements and conflicts spillover into international politics and affect the operation of the global economy – almost always in a negative way (Beyer, 1994). Religious institutions have become increasingly international or global in scope, and this has given rise to politicized movements that operate worldwide, particularly in Christianity and Islam.

See also: Civil Religion; Islam: Middle East; Islamic Law; Religion and Economic Life; Religion: Nationalism and Identity.

Bibliography

Beyer, Peter, 1994. Religion and Globalization. Sage, London.

Billings, Dwight B., 1990. Religion as opposition: a Gramscian analysis. American Journal of Sociology 96, 1–31.

Braudel, Fernand, 1982. Civilization and Capitalism, 15th-18th Century: The Wheels of Commerce, vol. 2. University of California Press, Berkeley, CA.

Chaves, M., 1994. Secularization as declining religious authority. Social Forces 72, 749–774.

Dubow, Saul, 1992. Afrikaner nationalism, apartheid and the conceptualization of 'race'. The Journal of African History 33, 209–237.

Froese, Paul, 2008. The Plot to Kill God: Findings from the Soviet Experiment in Secularization. University of California Press, Berkeley, CA.

Gill, Anthony, 1998. Rendering unto Caesar: The Catholic Church and the State in Latin America. University of Chicago Press.

Jensen, Gary, 2007. The Path of the Devil: A Study of Early Modern Witch Hunts. Rowman & Littlefield, Lanham, MD.

Jurgensmeyer, Mark, 2008. Global Rebellion: Religious Challenges to the Secular State, from Christian Militia to Al Qaeda. University of California Press, Berkeley, CA.

Just, Aida, Sandovici, Maria Elena, Listhaug, Ola, 2014. Islam, religiosity, and immigrant political action in Western Europe. Social Science Research 43, 127–144.

Kuran, Timur, 2010. The Long Divergence: How Islamic Law Held Back the Middle East. Princeton University Press, Princeton, NJ.

McCarthy, John, Zald, Mayer, 1977. Resource mobilization and social movements. American Journal of Sociology 82, 1212–1241.

Morris, Aldon, 1984. The Origins of the Civil Rights Movement: Black Communities Organizing for Change. Free Press, New York.

Poulson, Stephen, 2006. Social Movements in Twentieth-Century Iran: Culture, Ideology, and Mobilizing Frameworks. Lexington Books, Lanham, MD.

Sherkat, Darren E., 2005. Religion, social movements, and politics. In: Ebaugh, Helen Rose (Ed.), Handbook on Sociology of Religion and Social Institutions. Kluwer Press, pp. 1–20.

Sherkat, Darren E., Powell-Williams, Melissa, Maddox, Gregory, Vries, Kylan de, 2011. Religion, politics, and support for same-sex marriage in the United States, 1988-2008. Social Science Research 40, 167–180.

Snow, David A., Rochford, E. Burke, Worden, Steven K., Benford, Robert D., 1986. Frame alignment processes, micromobilization, and movement participation. American Sociological Review 51, 464–481.

Stark, R., Finke, R., 2000. Acts of Faith: Explaining the Human Side of Religion. University of California Press, Berkeley, CA.

Weber, Max, 1978. Economy and Society. University of California Press, Berkeley, CA.

Young, Michael P., 2002. Confessional protest: the religious birth of U.S. National social movements. American Sociological Review 67, 660–688.

Zald, Mayer N., 2000. Ideologically structured action: an enlarged agenda for social movement research. Mobilization 5, 1–16.

Religion: Morality and Social Control

John P Hoffmann, Brigham Young University, Provo, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is revision of the previous edition article by C. Davies, volume 19, pp. 13082–13085, © 2001, Elsevier Ltd.

Abstract

Religion, morality, and social control have been inexorably linked for centuries. In fact, some argue that morality and social control are derived fundamentally from religion. Dominant religious bodies in many societies define correct behavior, codify conduct, and establish social control systems designed to enforce systems of behaviors. This article begins with a discussion of morality, including its definition in the social sciences and how it both influences and is influenced by society. Social control is treated similarly, with attention to different types of control, distinct types of social control organizations, whether social control is unduly influenced by elite members of society, and whether its differential impact on people is distinguished by class, race/ethnicity, and gender. The role of religion in defining and being affected by morality and social control is presented next. This section addresses how various aspects of religion affect moral behaviors and serve as social control mechanisms. The final section discusses how religion is sometimes cast as immoral and deviant.

Religion, morality, and social control have been inexorably linked for centuries. Sacred mandates such as the biblical Ten Commandments, the *pañcasila* (five precepts) of Buddhism, and the Nine Beliefs of Hinduism were used – and are still used – to guide appropriate conduct and organize proper societies. Each includes moral guidelines that adherents are expected to observe. Moreover, the broader literature in which they are found outline systems of social control that are designed to discourage adherents from stepping outside moral boundaries (e.g., Deut. 22:24; Qu'ran 5:38).

In fact, some have argued that morality and social control are derived fundamentally from religion. Dominant religious bodies in many societies define correct behavior, codify conduct, and establish social control systems designed to enforce systems of behaviors. This may occur even if government sanctioned social control mechanisms do not consider a behavior as improper. Moreover, although the direct influence of religious organizations on social life has waned in most modern societies, remnants of their moral guidelines and control functions remain.

There is a persistent unease in the social sciences, however, with how morality and social control are defined and what these definitions imply for our understanding of religion, social structure, culture, and myriad other social phenomena. It is therefore important to review historical and contemporary understandings of morality and social control before addressing their associations with religion.

Morality

Several of the founders of the social and behavioral sciences had a keen interest in morality. It may be said that this interest was fundamental since one of the founders of sociology, Émile Dürkheim (1973 [1925]), sought to establish a discipline by focusing on how morality was a reflection of the nature and organization of a society. Sigmund Freud (1989 [1930]) discussed how ignoble human desires, if left unchecked, would lead to annihilation, so laws guiding moral sensibilities – fostered early in life – were needed to quell this tendency. Bronislaw Malinowski (1926), based on his fieldwork among the Trobriand Islanders, argued that life among so-called savages was not anarchic; rather, their lives were guided by moral rules that imparted practical guidance and served to bind them together in a community. Unfortunately, the examination of the origins of morality lapsed as a topic of interest for several decades.

There has been a resurgence of academic interest in morality in recent years, especially in sociology and psychology (Black, 2011; Haidt, 2007; Hitlin and Vaisey, 2010). This revivification has focused on (1) defining morality more precisely; (2) identifying the sources of morality at the micro and macrolevels; and (3) understanding the cultural implications of moral meanings and judgments (Hitlin and Vaisey, 2010). In its simplest form, morality involves 'rules of conduct' or what have commonly been called 'norms' (Black, 2011, p. 11). Yet, this term has also been distinguished by an emphasis on (1) behaviors that are deemed right or wrong, good or bad, or acceptable or unacceptable; or (2) as a way to judge the appropriateness of certain behaviors. Some suggest that these are interdependent since judgments of appropriateness – if they are truly 'moral' – must occur in some general sense and not be based simply on individual preferences or wishes (Smith, 2003). The evaluation of moral actions is therefore built on general rules of conduct (which include a sense of duty or obligation) that are shared by some set of people and apply in particular situations. Thus, it is contingent since moral actions and judgments often depend on contextual factors, for example, what is appropriate conduct at home may not be appropriate at school or in the workplace.

Observers have also argued that the sources of morality may be found at both the micro and the macrolevels of analysis. At the microlevel, individuals are not atomistic, making choices unfettered by others. Instead, they are part of a hierarchy of individuals and groups that define what is desirable and the accepted means to achieve it. The common sociological and psychological concern with 'socialization' illustrates how important others are in inculcating a sense of what is moral. At the macrolevel, differential group resources and authority serve

to establish what is acceptable and unacceptable. Some conflict theorists contend that powerful groups disproportionately define morality in most societies. Social movement theorists suggest that shared interests among group members galvanize definitions of acceptable vs unacceptable actions. However, many social movements are animated by a shared sense of morality that precedes group formation.

Morality is also influenced by culture and gives rise to worldviews and meaning structures that result in self-identification and salience. 'Moral meanings' affect (1) group solidarity; (2) judgments and even condemnation of outsiders; and (3) narratives designed to explain one's life and identity (Haidt, 2007; Hitlin and Vaisey, 2010). These narratives and associated worldviews support group formation and are encouraged by group membership. For example, conservative and liberal worldviews are centered on different models of moral meanings (Lakoff, 2002). When generalized moral judgments are classified into five domains (harm/care, fairness/reciprocity, in-group/loyalty, authority/respect, and purity/sanctity), liberals tend to endorse harm/care and fairness/reciprocity. Yet conservatives tend to weigh each of these equally. This attention to how morality is cognitively organized and used helps explain the many recalcitrant political, social, and economic disputes in the United States (Graham et al., 2009).

Social Control

The term social control has had a similar conceptual life course as the term morality. Early definitions focused on premodern groups and held that social control consisted of shared norms and practices designed to penalize deviants (Horwitz, 1990). A more general definition expressed that social control entails social relationships, processes, and intentional actions that sustain social conformity, thus regulating deviant behavior (Liska, 1992). This includes relationships, processes, and actions that take place in families, the workplace, schools, and religious and legal institutions (from police to prisons). Moreover, it may include any number of actions, including mild verbal correction, humiliation, ostracism, banishment, fines, beatings, incarceration, and execution.

Attention to this concept intensified during the early decades of the twentieth century as sociologists attempted to study what was happening to social relationships as large numbers of people moved from rural areas and small towns to urban areas. Social scientists investigated whether the primary relationships, processes, and actions that sustained conformity in small towns were transferable to – or as effective in – urban areas where social relationships tended to be more fleeting and not as intimate (Meier, 1982). Subsequently, researchers tended to address other topics, such as (1) different types of control, such as external vs internal or formal vs informal; (2) distinct forms of social control organizations; (3) whether social control was disproportionately established and unduly influenced by elite members of society; and (4) the differential impact of social control agencies on people distinguished by class, race/ethnicity, and gender.

The emphasis on external vs internal and formal vs informal control led to numerous studies of adolescent and adult deviance, in particular criminal behaviors, sexual deviance (a rapidly changing concept), self-destructive behaviors, and mental health problems. Some scholars distinguished between internal forces, such as personal beliefs and values, and external forces, such as parental discipline. A recent type of internal control that has gained significant attention in studies of crime and delinquency involves self-control. Individuals lacking self-control tend to be "impulsive, insensitive, physical (as opposed to mental), risk taking, [and] shortsighted" (Gottfredson and Hirschi, 1990: 90). Its advocates maintain that a parsimonious explanation of deviant behavior is low self-control, which heightens the risk of inappropriate conduct when opportunities are present. Whereas this is a highly individualistic form of internal control, others have examined informal and external mechanisms of control by family members, teachers, and peer groups, and formal mechanisms of control by the police, the courts, or the mental health system.

Conflict theorists were instrumental in drawing attention to how elite segments of society influenced the development and execution of social control mechanisms. They also argued that social control processes disproportionately affect less powerful segments of society (Liska, 1997). Studies demonstrating a positive association between the concentration of poverty or the percent of ethnic minorities and police arrest rates have illustrated that low status groups, as well as low status communities, tend to be unduly targeted – explicitly or implicitly – by systems of social control (Eitle and Monahan, 2009). Similarly, neighborhoods with a high concentration of minority groups or a high degree of poverty are often labeled as areas in need of social control, even after taking into account rates of actual deviance in these neighborhoods (Sampson and Raudenbush, 2004). However, others have pointed out that social movements, which may or may not represent elite interests, are particularly instrumental in influencing social control (e.g., prohibition). Moreover, a key hypothesis that emerged a couple of decades ago focuses on how perceived social threats from deviant behaviors or even from groups seen as involved in these behaviors lead to an escalation in social control efforts, such as more arrests, mental health admissions, and curfew laws (King and Wheelock, 2007; Liska, 1997).

How is Religion Related to Morality and Social Control?

Although the previous sections were presented independently, it should be clear that morality and social control are inexorably connected. After all, the main focus of social control – deviance – is based on judgments and viewpoints about morality. Moreover, any maintenance of morality must be structured by some type of social control mechanism, whether it involves familial or community rules of conduct, or state-run organizations designed to enforce laws and regulations.

In fact, Dürkheim's early work posited that conformity resulted from external sanctions, whereas his later work emphasized that moral behavior was guided by internalized rules or controls. These could be so powerful that they inhibited even the calculation of deviant behavior

(see Wikström (2006) for a contemporary version of this argument). Yet, the distinction between morality and social control is not clear in these works. Are moral meanings or viewpoints simply a type of internal social control? Present-day scholars studying morality and social control have tended to distinguish the two, with moral meanings making control either more effective or unnecessary (Wikström, 2006).

Nevertheless, Dürkheim, Freud, and others did address the role that religion played in the development of morality and social control. This concern led to different conclusions, with Dürkheim seeing some forms of religion as offering sources of common moral standards, thus enhancing the integrative functions of a society; and Freud outlining some harmful aspects of religiosity and its influences on social control.

Contemporary research on religion, morality, and social control addresses (1) the sources of moral beliefs and meanings in religion; (2) the effects of religious beliefs and practices – and how these may serve as social control mechanisms – on moral beliefs, attitudes, and behaviors; and (3) the contingent and macrolevel influences of religion on moral behaviors.

It is evident that throughout history moral beliefs and meanings have been found in religious teachings and practices. Sacred texts are replete with instructions on what is permissible or encouraged and what is prohibited or discouraged. As noted earlier, many of the sacred texts of the world's major religions, such as the Torah, the Old and New Testaments, the Qu'ran, the Vedas, and many of the sayings attributed to the Buddha, describe specific prohibitions, and some outline the risks inherent in particular behaviors (e.g., eternal torment, stoning, banishment). Several of the prohibitions and sanctions taught by these religious traditions are still in force, even though, in quite a few nations, the behaviors they seek to influence are no longer forbidden under formal systems of social control (e.g., adultery, premarital sex, homosexuality).

Nevertheless, it is important to acknowledge that moral beliefs and meanings are not consistently derived from religious institutions or teachings; they may be spring from various other sources, including rational evaluations of one's own and others' actions, altruistic and empathetic tendencies, and even enlightened self-interest (Baggini, 2003). Moreover, contemporary studies suggest that general moral rules are less influential in predicting behaviors than are immediate concerns, emotions, and heuristic tools. Thus, we find that people naturally search for a moral explanation following some behavior, even if it is not generally considered moral by one's peers (Bloom, 2012; Haidt, 2007). This may help explain why religious beliefs and practices are not consistently associated with moral acts. Consider that the Inquisition and the institution of slavery, as well as Mother Teresa's work with lepers in India and Martin Luther King's calls for racial equality and justice, had religious underpinnings.

A vast number of studies have examined the effects of religious affiliation, beliefs, and behaviors on attitudes toward moral issues and the performance of moral and immoral behaviors (the definition of morality used in these studies is not consistent, however). In other words, religion as a form of social control – a topic with a long history of academic interest (Ross, 1896) – has become focused on how particular aspects of religiosity affect individual-level attitudes and behaviors. Empirical research suggests that belonging to a conservative Christian denomination, believing in the inerrancy of the Bible, seeing God as a father figure or a severe disciplinarian, and participating more often in religious services are associated with stricter or more traditional views about abortion, pornography, same-sex relations, corporal punishment, capital punishment, euthanasia, and several other social, political, and cultural issues (Greeley and Hout, 2006).

The association between these religious characteristics and actual behaviors is less certain. One consistent finding of this research involves sexual behavior among young people: those who are highly religious tend to delay sexual initiation, have fewer sexual partners, and are more likely to be monogamous; even though they are, in general, as likely to engage in sex as other youths (Bader and Finke, 2010; Regnerus, 2007). Similarly, most studies have found that stronger religious salience and greater participation are negatively associated with illegal drug use (e.g., Bahr and Hoffmann, 2008). Less consistent, though, is the association between religion and criminal or delinquent behavior, suicidal tendencies, and mental health problems. Early studies found weak to no associations between aspects of religiosity and these outcomes. Subsequent research has found stronger evidence of associations, but they tend to be contingent on other factors (Hoffmann and Bahr, 2005).

The most common finding of recent studies is that participating in a religious community, such as by attending worship services, volunteering for a religious organization, or being part of religious youth and adult groups, is negatively associated with engaging in deviant behaviors. The effect of religious beliefs is substantially weaker, however. Thus, some observers claim that it is enmeshment in religious social networks and a community of religious adherents that serves as a social control against what members see as immoral activities (Graham and Haidt, 2010). It also encourages moral behaviors, such as altruism, charitable giving, and perhaps even healthy physical activities (Bloom, 2012; Hill et al., 2003).

Another important finding involves macrolevel religious influences. This has been approached in two ways. First, the moral communities' hypothesis contends that individual-level religion exerts the strongest negative influence on deviant behaviors, especially delinquent behaviors, in communities relatively high in religiosity (Stark and Bainbridge, 1997). Presumably, when communities are high in religiosity, this reinforces the religiosity of the individual and thus behaviors are affected more.

Second, several recent studies examine whether macrolevel religious influences positively affect aggregate levels of moral behaviors, or have a direct impact on individual-level moral behaviors. For example, cities with higher rates of church attendance tend to have lower crime and suicide rates (Stark and Bainbridge, 1997). Areas that have a higher degree of religious homogeneity (the extent to which residents are members of a single or a smaller number of religious faiths) have lower rates of suicide, divorce, and violent crime, and more religious areas produce a lower risk of

suicide and suicide attempts (Ellison et al., 1997; Trawick and Housen, 2006; Lee and Bartkowski, 2004; Maimon and Kuhl, 2008).

It appears, therefore, that religious adherence and participation at the aggregate level serves as a form of social control – or social integration mechanism – that reduces the risk of some forms of deviant behavior. Presumably, this reflects some enhanced level of morality tied to religion, at least in the United States (although there is also evidence of this in Muslim nations; Huang, 1996), yet it is not the direct function of religious beliefs per se. In other words, religious beliefs, although they typically indicate some generalized support for morality, do not translate directly into moral behaviors. Instead, it is enmeshment in moral communities – aggregates of religiously involved people – that enhance generally accepted moral behaviors and diminish deviant behaviors.

Religion as Immorality and Deviance

A disparate thread of social scientific attention is to religion as deviance and immorality. Moral crusades have been undertaken countless times against religious groups. The Crusades, the many periods of persecution of Jews, religious conflicts between Catholics and Protestants in sixteenth- and seventeenth-century Europe, the execution of hundreds of Christians in seventeenth-century Japan, the Mormon expulsion from Missouri and Illinois in the 1840s, the Branch Davidian conflagration, the massacre of Muslims in Kosovo, and the continuing conflicts between Muslim Pakistan and Hindu India are but a few examples of how perceptions of outsider religious groups can turn to violence. Although persecution, mass executions, and ethnic cleansing are extreme forms of social control used against religious groups, many more subtle methods have also been used.

The pejorative use of the term cult brings to mind how some religious groups are seen as deviant and often labeled as immoral due to their beliefs and behaviors. Yet virtually all major religious traditions began as deviant groups, at odds with the prevailing culture and acting in a manner that was easily termed abnormal or illicit. For example, the early Christian church was illegal during certain periods of the Roman occupation and persecution – by Romans, Jews, and others – for perceived immoral practices was rampant. Not only were they considered deviant for not worshipping Roman gods and for failing to require converts to be circumcised, but they also were the object of rumors of child sacrifice and cannibalism (Gaddis, 2005).

Moreover, religiousness has often been linked to mental illness. Whereas many early psychological and psychiatric researchers assumed that religious believers were likely delusional or prone to mental illness (and thus labeled as deviant by the prevailing medical standards of the time), more recent scholarship has disproved this assumption. Most studies find that those who are enmeshed in a faith tradition are less likely to experience a diagnosable mental disorder, and that they report higher levels of life satisfaction and happiness than others (Ellison, 1991; Koenig et al., 2012; Stark and Bainbridge, 1997). Once again, this may reflect the beneficial aspects of belonging to a community of like-minded people who provide social support and sustenance, rather than a unique effect of religious participation or beliefs.

See also: Morality: Evolution of; Religion and Economic Life; Religion, Psychology of; Religion, Sociology of; Religiosity; Technology and Social Control.

Bibliography

Bader, C.D., Finke, R., 2010. What does God require? Understanding religious context and morality. In: Hitlin, S., Vaisey, S. (Eds.), Handbook of the Sociology of Morality. Springer, New York, pp. 241–254.

Bahr, S., Hoffmann, J.P., 2008. Religiosity, peers, and adolescent drug use. Journal of Drug Issues 38, 743–770.

Baggini, J., 2003. Atheism: A Very Short Introduction. Oxford University Press, New York.

Black, D., 2011. Moral Time. Oxford University Press, New York.

Bloom, P., 2012. Religion, morality, evolution. Annual Review of Psychology 63, 179–199.

Dürkheim, E., 1973 [1925]. Moral Education. Free Press, New York.

Eitle, D., Monahan, S., 2009. Revisiting the racial threat thesis: the role of police organizational characteristics in predicting race-specific drug arrest rates. Justice Quarterly 26, 528–561.

Ellison, C.G., 1991. Religious involvement and subjective well-being. Journal of Health and Social Behavior 32, 80–99.

Ellison, C.G., Burr, A.L., McCall, P.L., 1997. Religious homogeneity and metropolitan suicide rates. Social Forces 76, 273–299.

Freud, S., 1989 [1926]. Civilization and Its Discontent. Norton, New York.

Gaddis, M., 2005. There Is No Crime for Those Who Have Christ: Religious Violence in the Christian Roman Empire. University of California Press, Berkeley, CA.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Graham, J., Haidt, J., 2010. Beyond beliefs: religions bind individuals into moral communities. Personality and Social Psychology Review 14, 140–150.

Graham, J., Haidt, J., Nosek, B., 2009. Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology 96, 1029–1046.

Greeley, A., Hout, M., 2006. The Truth about Conservative Christians: What They Think and What They Believe. University of Chicago Press, Chicago, IL.

Haidt, J., 2007. The new synthesis in moral psychology. Science 316, 998–1002.

Hill, T.D., Ellison, C.G., Burdette, A.M., Musick, M.A., 2003. Religious involvement and healthy lifestyles: evidence from the survey of Texas adults. Annals of Behavioral Medicine 34, 217–222.

Hitlin, S., Vaisey, S., 2010. Back to the future: reviving the sociology of morality. In: Hitlin, S., Vaisey, S. (Eds.), Handbook of the Sociology of Morality. Springer, New York, pp. 3–14.

Hoffmann, J.P., Bahr, S.J., 2005. Crime/deviance. In: Ebaugh, H.R. (Ed.), Handbook of Religion and Social Institutions. Springer, New York, pp. 241–263.

Horwitz, A.V., 1990. The Logic of Social Control. Plenum, New York.

Huang, W.C., 1996. Religion, culture, economic and sociological correlates of suicide rates: a cross-national analysis. Applied Economics Letters 3, 779–782.

King, R.D., Wheelock, D., 2007. Group threat and social control: race, perceptions of minorities and the desire to punish. Social Forces 85, 1255–1280.

Koenig, H.G., King, D.E., Carson, V.B., 2012. Handbook of Religion and Mental Health, second ed. Oxford University Press, New York.

Lakoff, G., 2002. Moral Politics, second ed. University of Chicago Press, Chicago, IL.

Lee, M., Bartkowski, J.P., 2004. Love thy neighbor? Moral communities, civic engagement, and juvenile homicide in rural areas. Social Forces 83, 1001–1035.

Liska, A.E., 1992. Social Threat and Social Control. State University of New York Press, Albany, NY.

Liska, A.E., 1997. Modeling the relationships between macro forms of social control. Annual Review of Sociology 23, 39–61.

Maimon, D., Kuhl, D.C., 2008. Social control and youth suicidality: situating Durkheim's ideas in a multilevel framework. American Sociological Review 73, 921–943.

Malinowski, B., 1926. Crime and Custom in Savage Society. Routledge and Kegan Paul, London.

Meier, R.F., 1982. Perspectives on the concept of social control. Annual Review of Sociology 8, 33–55.
Regnerus, M.D., 2007. Forbidden Fruit: Sex and Religion in the Lives of American Teenagers. Oxford University Press, New York.
Ross, E.A., 1896. Social control: v. religion. American Journal of Sociology 2, 433–445.
Sampson, R.J., Raudenbush, S.W., 2004. Seeing disorder: neighborhood stigma and the social construction of "broken windows". Social Psychology Quarterly 67, 319–342.
Smith, C., 2003. Moral, Believing Animals: Human Personhood and Culture. Oxford University Press, New York.
Stark, R., Bainbridge, W.S., 1997. Religion, Deviance, and Social Control. Routledge, New York.
Trawick, M.W., Housen, R.M., 2006. Crime and community heterogeneity: race, ethnicity, and religion. Applied Economics Letters 13, 341–345.
Wikström, P.H., 2006. Individuals, settings, and acts of crime: situational mechanisms and the explanation of crime. In: Wikström, P.H., Sampson, R.J. (Eds.), The Explanation of Crime: Context, Mechanisms, and Development. Cambridge University Press, New York, pp. 61–107.

Religion: Nationalism and Identity

John Myhill, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the last decade, it has come to be generally recognized in Western countries that religious identity has not been eclipsed by national identity but remains as a powerful political force, interacting with national identity in complex ways and sometimes superseding it. Much of the most obvious evidence for this comes from an increased awareness of and contact with Islam, particularly because so many Muslims believe that loyalty to the Muslim *'umma*, i.e. the community of all Muslims, is more important than loyalty to the nation or state. This is reflected in the behavior of Muslim immigrants to Western countries, the powerful effect of religion on democratic elections in Muslim countries, and violent conflicts between Muslims belonging to different sects but having the same nationality. But similar interaction can also be found in non-Muslim societies. Religiously-motivated Christians, Hindus, and Jews in Western countries, India, and Israel have become increasingly sophisticated in terms of using democratic institutions to influence national policy and decision marking in accordance with their own religious convictions.

The understanding of the relationship between religion and nationalism has changed dramatically in the last decade or two. Prior to this, the modern understanding had been that religion and nationalism, and religious and national identity, were naturally opposed in the sense that to the extent that states and individuals emphasized one, they deemphasized the other, and that furthermore there is a general evolutionary trend in human societies toward national allegiance at the expense of religious allegiance. But toward the end of the last century, an increasing amount of research began to suggest that this was not necessarily the case, and scholars began to reevaluate and rethink the relationship between religion and nationalism. Today the tide has turned entirely, in terms of both intellectual developments and real-world events, and the question is not whether religion and nationalism are in conflict or not but rather what the nature of the relationship between them is and how it can vary from case to case. Scholars no longer ask whether religious allegiance is politically relevant in the modern world, but rather how it is politically relevant. So in the last decade it has become increasingly mainstream to discuss theories of the relationship between religion and nationalism (e.g., Bader, 2007; Brubaker, 2013; Buruma, 2010; Casanova, 2009; Gorski and Altinordu, 2008; Martin, 2005; Turner, 2011) and cases of interaction between religion and nationalism in constructing political identities in the past (e.g., English, Dutch, and Swedish nationalism in Ihalainen (2005); Polish nationalism in Zubrzycki (2006); Slovenian nationalism in Mihelj (2007); Sinhalese nationalism in DeVotta (2007); Japanese nationalism in Fukase-Indergaard and Indergaard (2008); Finnish nationalism in Tilli (2009); etc.). This is no longer seen as revolutionary or even controversial.

For these reasons, the present article will focus on recent events that show something about the relationship between religion and nationalism in the present day, which should give some idea of what we might expect in the foreseeable future. Most of the events which will be discussed here have taken place in the last decade, although a few took place a bit earlier. In terms of recent events, it has certainly been the connection between Islam and nationalism that has brought the most attention to the question of the relationship between religion and nationalism in the modern world, and so this will be the first topic to be addressed here. This has taken several interrelated forms.

First, in the second half of the twentieth century, as a result of the catastrophe of the two world wars and the end of the age of colonialism, nationalism and racism were ideologically discredited, and this, combined with the apparent success of the American melting pot experiment, led Western European and Scandinavian countries to adopt immigration policies allowing people of diverse ethnic backgrounds to move to, live in, and in many cases become citizens of their countries. The assumption was that in the modern world national identity in the form of citizenship would trump other factors – ethnicity, race, and religion – so that even when a state has taken what was proclaimed to be a multicultural approach to immigration, the implicit or explicit understanding was that it would be the immigrants who would accommodate the state that they have chosen to become citizens of, not the other way around. But in the last few decades it has become apparent that things are not necessarily so simple, and while complications in this regard have been perceived with a number of immigrant groups, there is no question that Muslims have been understood to be the most generally problematic in this regard, particularly in terms of publicly expressing the expectation that they will be able to live under Islamic law in Western countries, and in some prominent cases even carrying out violent and sometimes fatal politically motivated attacks against non-Muslim citizens of the countries they live in. In terms of the issues of identity which this review focuses on, there is a perception (right or wrong) that for many Muslims, contrary to the assumptions that lay behind post-World War II immigration policies, religious identity as a member of the worldwide Muslim *'umma* is more important than citizenship in the state they have immigrated to, so that they assume that it is the state that must accommodate to Islam and not the other way around. This has resulted in a prominent public dialogue on issues related to immigration, multiculturalism, and cultural accommodation, couched, in principle, in general terms but tending to focus on the place of Muslims in Western societies in particular (see for example Bawer, 2006; Bruckner, 2007; Buruma, 2006; Caldwell, 2009; Lentin and Titley, 2011; Saunders, 2012; Triandafyllidou et al., 2011).

Second, there has been a recognition, particularly in the last decade, that the understanding of the *'umma* in the Muslim world is generally a challenge to the assumption that citizenship-based national identity must trump religious identity in the modern world, not only in terms of how Muslims relate to Western states but also in terms of what happens in countries in which Muslims are demographically and politically dominant. The *'umma* can be perceived as either being transnational, with loyalty to Islam transcending loyalty to the nationality, or alternatively itself constituting a nationality of a type radically different from that found today in Western societies, but in either case, understanding it – let alone dealing with it – requires some sort of rethinking of the traditional Western modernist conceptualization of the relationship between religion and nationalism (see e.g., Jones and Mas, 2011; Mandaville, 2011; Saunders, 2008; Zubaida, 2004).

Third, the general assumption of modern Western society has been that the establishment and spread of democratic institutions must be associated with secularization, so that, in the terms of this review, the right to participate in democratic elections motivates people to identify more with their national identity as citizens of the country whose leadership they have actively participated in choosing and less with their own religious affiliation. Until relatively recently, countries in which Muslims are politically dominant did not provide evidence to the contrary, because they generally had not developed democratic institutions. It thus seemed natural to assume that they would go through the same evolutionary process when they developed such institutions. But since the end of the Cold War, which gave a big push to the general spread of democratic institutions to countries all around the world, and increasingly in the last decade, it has become apparent that things do not necessarily work this way in Muslim countries – in many cases, it seems that it is particularly when Muslim citizens are allowed to vote and their votes mean something that their countries become more overtly Islamic. The most prominent example of this is Turkey, where it is clear in retrospect that it was only repeated coups by the avowedly secular military against democratically elected governments that were perceived as too Islamic, that made it possible to maintain the façade of a country that was Muslim, democratic, and secular – in the last decade. As the military has been effectively eliminated as a political force and the Turkish people have been free to choose their own government through democratic elections, the country has become markedly more Islamic (see e.g., Brockett, 2011; Solmazturk, 2008; Tibi, 2009; Turan, 2006; Ünver, 2013).

Recent developments in Egypt have pointed to the same conclusion. Until 2011, Egypt was a relatively secular state with few pretensions of being democratic (while Turkey had been prior to the last decade a secular state with the façade of democracy), and the fall of Mubarak led to democratic elections in which the two strongly Islamist blocs – the Muslim Brotherhood and the Salafists – won more than 70% of the elected seats in Parliament and took power (see e.g., Farahat, 2011; Stock, 2012). In 2013, backed by a petition whose organizers, the Tamarod movement, claimed to have 22 million signatories, the military took power again (the only estimates of the number of signatories or supporters of the Tamarod movement come from the military and Tamarod itself) and established more or less the status quo ante, and although there are claims that one day elections will be held again, this does not seem likely to happen in the foreseeable future. One may call the military's reassertion of control a coup or something else, but it clearly was not done through normal democratic procedure. As with developments in Turkey, this similarly suggests that it is naïve to believe that, in Muslim countries at least, secularism and democracy necessarily go together or that the expression of national identity through democratic elections naturally motivates people to downplay the religious aspect of their identity. On the contrary, the pattern in at least some Muslim societies seems to be that democratic institutions reflecting popular voting strengthen the religious aspect of the state at the expense of its secular aspect, so that in such societies one might have to choose between secularism and democracy (or, alternatively, redefining democracy in such a way as to not directly reflect popular voting).

It seems that national identity in Muslim countries can be strengthened by expressions of religious identity, so that there is a perception that e.g., if a Turk is more of a Muslim, she/he is also more of a Turk. Many, if not most, Muslims apparently do not distinguish between being nationalist and being Islamist. Far from setting up their state in opposition to religious institutions, modern Muslim states seem to be using Islam as best they can to legitimize themselves, and this has become more important and complicated in recent times as the general trend toward democratic elections has grown stronger.

It is of course true that not all states dominated by Muslims clearly fit this pattern. One might identify states such as Malaysia, Bangladesh, and Albania as cases that better fit the Western model, where the establishment of democratic institutions has not resulted in significant popular support for Islamism. It is worth noting that these countries correspond much more closely to the Western ideal of the nation-state – the overwhelming majority of citizens belong to a particular language-based nationality, there are no significant conflicts with members of other language-based nationalities (as between the Turks and the Kurds in Turkey, a point discussed below), and there is no pan-national identity at stake (as in the case of Egypt and Arab nationalism). It may well be that the cases in which Muslims generally do not want to put their national identity clearly ahead of their religious identity are specifically those cases in which their national identity is not associated with an existing nation-state; this is a matter which needs to be investigated.

A fourth way in which the relationship between Islam and national identity has developed in a way counter to Western assumptions is in cases where Islam is used in an attempt to neutralize the distinction between different linguistic groups in the same country by referring to their common religion. Whether or not this puts religious identity in opposition to national identity depends upon whether one defines national identity in terms of language or in terms of citizenship – if national identity is defined in terms of language, then the idea is to subordinate national identity to religion in the sense of saying that common religion trumps different language, while if national identity is defined in terms of citizenship, the idea is to strengthen national identity by associating it with a common religious identity. To be sure, Western governments did try things like this in the past, most obviously the attempts by the

Austro-Hungarian government to hold their empire together by appealing to the common Catholicism of the different linguistic-national groups, but these attempts were completely unsuccessful. But thus far this is generally not the case in Muslim countries today – on the contrary, a crucial aspect of the rise of Islamic identity in Turkey today has been efforts in the last decade by Islamically oriented Turks and Kurds to bridge the gap between the different linguistic groups by emphasizing their common Islamic identity (see Sarigil and Fazlioglu, 2013). In fact, the rise of the AKP as the dominant party in Turkey coincided exactly with the end of the Kurdish–Turkish conflict, which took more than 35 000 lives and ended in 2000, with the AKP being founded in 2001 and taking almost 2/3 of the seats in the 2002 elections. Turkey appears to have solved its problem of having two linguistically divided nationalities, at least for the time being, by turning to Islamism. Similar motivations can be found for the appeal of Islamism in countries with radically mixed linguistic populations such as Iran, Pakistan, Indonesia, and Sudan, this in contrast to the limited appeal of Islamism in more prototypical nation-states like Malaysia, Bangladesh, and Albania. (To be sure, such tactics are by no means limited to Muslims – Tuttle (2005) shows how efforts to incorporate Tibet into China have been associated with appeals to their common Buddhist tradition.)

Fifth, in the last decade radical splits have developed between different Islamic factions, particularly Sunnis on one side and Shiites (and in Syria also Alawites) on the other side, most dramatically in Iraq and Syria but also in other countries such as Bahrain and Yemen. This split is occurring particularly in Arabic-speaking areas, although there are increasing tensions between these groups in other areas as well. To a significant extent the armed conflict in Arabic-speaking areas has become particularly serious in the last decade because non-Arab states, Iran and the United States, have materially helped the Arabic-speaking Shiites, who had been politically powerless in the area since the twelfth century. This created a more balanced playing field (viewed positively) and an arena for conflict (viewed negatively). Of course, the United States did not enter the arena in Iraq to help the Shiites there, but because they deposed the Sunni ruler Saddam Hussein. This action established a framework for democratic elections, which the Shiites – by far the largest group – would obviously win, and this was a predictable result of their involvement. But whatever outside powers have done, it is still undeniable that Arabic-speaking Sunnis and Arabic-speaking Shiites/Alawites are systematically killing each other because of their religious affiliation, and if Arab national identity is associated with speaking Arabic, this means that they are evidently treating religious identity as being more important than national identity.

But in fact it turns out that that Arab national identity is not, and has never been, associated only with speaking Arabic. The Arab League has defined Arab nationality as follows: "An Arab is a person whose language is Arabic, who lives in an Arabic-speaking country, who is in sympathy with the aspirations of the Arabic-speaking peoples." The assumption here is that having Arab nationality involves not only speaking Arabic (that is, a typical language-based criterion for nationality) but also a particular political orientation. Although this definition does not explicitly refer to religious affiliation, in practice the applicability of the term 'Arab' does appear to depend upon religion. Before the last decade, the only Arabic-speaking group that was more or less universally considered not to be Arab was Arabic-speaking Jews, that is, a group defined by religious affiliation (see e.g., Shenhav, 2006). If this were the only case of this type, it might be ascribed to something about Jews in particular as opposed to religion in general. However, there were also the cases of the Maronites of Lebanon, another religiously defined group for whom there was evidence that the overwhelming majority do not consider themselves to be Arabs (Khashan, 1990) – although this was mitigated by the fact that non-Maronites did not generally share this belief – and also Israeli Druze, another religious group, whose men serve in the Israeli army and whom the Israeli government does not officially consider to be Arabs.

But in the last decade, the cases of Arabic-speaking people of particular religious affiliations who are not considered to be Arabs have proliferated to the point where this can no longer be explained away as unrelated to religion, particularly in the context of conflicts between Sunnis, Shiites, and Alawites, which have made it undeniable that Arab nationality is directly tied to religion, specifically to Sunni Islam. The American invasion of Iraq in 2003 accelerated this process, and in the subsequent American-sponsored elections of January 2005, Shiite parties won a resounding victory. This was expected, given that the Arabic-speaking Shiites are 60% of the population of the country, but the Arabic-speaking Sunni minority, constituting 20% of the population, had been politically dominant since Iraq became independent in 1932. Indeed, Sunni Islam had been the politically dominant religion in what is now Iraqi territory since the seventh century (aside from a brief period under Mongol rule), so that the Shiite takeover was a radical challenge to the Arabic-speaking Sunnis' understanding of the natural political order of things. They responded by initiating a series of attacks against Shiite targets, which produced predictable responses, and on 6 April 2006, President Hosni Mubarak of Egypt (a Sunni, like almost all Egyptian Muslims) stated in a televised interview on al-Arabiyya that 'Shiites are mostly always loyal to Iran and not the countries where they live,' in effect excommunicating Arabic-speaking Shiites from the Arab nationality on the basis of their religion. As political scientist Vali Nasr interpreted this statement in an interview with *Der Spiegel* on 25 October 2006: "When President Hosni Mubarak of Egypt says that the Shiites are more loyal to Iran than their own countries, he is saying that the Shiites are not genuine Arabs, that they are somehow Iranian tools."

Since then, it has become common to refer to the different Arabic-speaking groups in Iraq as being 'Shiites' and 'Sunni Arabs.' For example, in the title of Mohammed M.A. Ahmed's 2009 book *America Unravels Iraq: Kurds, Shiites, and Sunni Arabs Compete for Supremacy*. Any random Google search will produce countless references of this type, contrasting 'Shiites' and 'Sunni Arabs' (but never 'Sunnis' and 'Shiite Arabs'). A similar situation has developed for Alawites in Syria, particularly since the beginning of the uprising against the Alawite regime of Bashir al-Assad in 2011, e.g., a video clip on YouTube entitled 'Syrian Sunni Arab citizen soldiers blow up Alawite outpost' [sic]. In connection to this, in a questionnaire study done in 2012 regarding how much different Arabic-speaking groups are perceived as being real Arabs, Ayyad (2014) found that Palestinians and Egyptians gave each other average Arabness ratings of 4.5 on a 5-point scale but rated Iraqi Shiites and Alawites as averaging lower

than 2, barely higher than Maronite Christians, Israeli Druze, Moroccan Jews, and Arabic-speaking Israeli Jews.

From these data, it appears that what is going on is that although Arab nationality is, in principle, defined in terms of the language that one speaks, in practice religion plays a significant role in that certain religious groups – Jews, Israeli Druze, Maronite Christians, Iraqi Shiites, and Alawites – are now widely considered to be excluded from Arab identity. Why these groups in particular? What all of these groups have in common, and what distinguishes them from other Arabic-speaking groups, is that they have all taken up weapons to fight against Arabic-speaking Sunnis. To be sure, this is not necessarily the reason which is overtly given in each case – it is usually stated in terms of loyalty to the Arab cause (like 'in sympathy with the aspirations of the Arabic-speaking peoples' in the definition of the Arab League). But there is apparently an implicit understanding that the Arab cause is the Sunni cause, that the aspirations of the Arabic-speaking people are for Sunni rule everywhere in Arabic-speaking territory, whether or not Sunnis are the local majority. To be sure, this is specifically the Sunni understanding of Arab identity – it is much less widespread among non-Sunnis. But Sunnis are the overwhelming majority of Arabic speakers overall, and so it is their understanding that basically controls public discourse; non-Sunnis never deny the Arab identity of Arabic-speaking Sunnis – at most they try to assert that they too are Arabs.

A final point to be made with regard to the effect of Islam upon the understanding of the relationship between nationality and religion today is that recent developments in contacts with Muslims have motivated some Westerners to adopt a political worldview which is in some sense parallel to that of Muslims in terms of assuming that religious identity as a Christian strengthens one's national identity as a citizen of a state politically dominated by Christians. The 9/11 attacks and smaller-scale attacks by Muslims against Western targets, as well as demands for accommodations by some Muslim immigrants to Western countries, have produced an anti-Muslim backlash by many Westerners, which has taken the form of politically rallying around Christianity – there were of course elements of Western society which in any case would have taken such a view, but there is no question that behavior by Muslims has increased both the number of Westerners who feel this way and the intensity of their feelings (see Goldberg, 2007). George Bush instinctively used the word 'Crusade' in speeches on the day of the 9/11 attacks and subsequent times to refer to military actions the United States would take to prevent further attacks of this type, and the connection to earlier conflicts with Muslims was undeniable and resonated with many Americans and Western Christians. Samuel Huntington's 1996 book *The Clash of Civilizations and the Remaking of World Order* was interpreted as prophetic in this context, and although Huntington used the term 'Western' rather than 'Christian' or 'Protestant/Catholic' to define what he saw as his own civilization, the fact that he used religious terms to define both the civilization which Western civilization is in direct conflict with at the present time and also the civilization which it had been in direct conflict with immediately before ('Muslim' and 'Orthodox' respectively) strongly implied that in terms of clashes, the really relevant parameter defining Western civilization is Western Christianity.

Recent events have also suggested that radicalization of any religion can be used to serve the cause of a nation associated with that religion, or at least to do something that radicals will perceive as serving such a national cause (see Greenfeld, 1996). Arabs, for example, appear to have been motivated to promote world-wide Islam by the fact that for historical reasons they can claim that Islam is 'their' religion and that the goals of Arab nationalism should be adopted by the *'umma* in general. Perhaps the most graphic example of this in the last decade has been the genocidal policies of the Arab government of Sudan toward the non-Arab Muslims of Darfur, which has resulted in about half a million civilian deaths, supported by an ideology according to which, although the non-Arabs of Darfur are Muslims, they are not as good Muslims as the Arabs are, so that it is in the interests of the Muslim *'umma* that they be subdued by whatever means necessary. As Gerard Prunier formulated this thinking, "Darfur was poor and backward because it was insufficiently Arabized. It had missed out in the great adhesion to the Muslim *'umma* because its Islam was primitive and insufficiently Arabic" (Prunier, 2005: p. 162), and accordingly in 2007 Osama bin-Laden called for a jihad in Darfur against Sunni Muslim non-Arabs in support of the Khartoum government. Similarly, as the conflict between Sunni and Shia Islam has heated up in recent years, it is natural that Turkey and Iran would try to increase their national influence by presenting themselves as the champions of Sunni and Shia Islam respectively, Turkey by emphasizing the role of the Turks in their Ottoman and post-Ottoman guise, Iran by virtue of being the only Shia power (see Louër, 2008). For that matter, perhaps George Bush the evangelical Christian was in initiating a Crusade in response to the 9/11 attacks actually motivated by George Bush the oil man, who recognized his country's national need for oil.

In the same way, although Jewish nationalism was originally not only secularist but opposed by the religious Jewish establishment, the situation has changed radically since the capture of the Old City of Jerusalem in 1967. On the one hand, religious Jews in Israel have (with the exception of the Naturei Karta, who remain anti-Zionist on the grounds that the messiah has evidently not yet come) not only embraced Zionism but have even managed to change the understanding of what Zionism is – whereas Zionism originally called for *a* Jewish state without any particular borders, it is now widely understood to require the inclusion in this state of areas beyond the 1967 borders of the state, particularly those deemed of religious significance to Judaism. This view motivates the establishment of Jewish settlements in these areas (increasingly focusing on certain areas in the West Bank, excluding Gaza). While this might appear to be a case of the religious tail wagging the nationalist dog, things are not so clear. Secular Jewish nationalists, opposed to territorial concessions to the Palestinians, have taken advantage of the situation by forming political alliances with representatives of the settler movement. Indeed, considering the fact that only one-third of Israeli Jews define themselves as being religious, and the expense and danger as well as the external criticism resulting from the settler movement, it is clear that many Israeli Jews must feel a non-religious, presumably nationalist, reason for supporting this movement. This is particularly clear in the case of Avigdor Lieberman's avowedly secular Yisrael Beiteinu party, supported particularly by Russian immigrants. In one way or another, this thinking pervades Israeli politics, a striking

recent example being the alliance instrumental in forming the present government between Yair Lapid's centrist Yesh Atid party, representing the secular middle class, and Naftali Bennett's HaBayit HaYehudi party, representing religious Zionism. While such alliances may seem bizarre or even incomprehensible to non-Israelis, they illustrate the extent to which it is not clear in Israel whether nationalism is the tool of religion or whether the reverse is the case.

One of the most striking examples of religion being used as a tool for nationalism involves the Lord's Resistance Army (LRA), a rebel group originally formed in Uganda in order to fight against the army of the national government there. The ostensible goal of the LRA has been to establish a Christian theocratic state based upon the Ten Commandments, and its leader Joseph Kony claims to be a spirit medium who reveals God's word. The depredations of the LRA are widely known today, but what is much less known is the relationship between the group and Acholi nationalism, the Acholis being a Nilotic group constituting about 5% of the population of Uganda whose homeland is in the north–central part of the country and whose members have constituted the overwhelming majority of those fighting for the LRA. The roots of the LRA can be traced back to British colonial policy in Uganda, which resulted in the armed forces, including the officers' corps, having a radically disproportionate number of Acholis, a situation which continued after decolonization. This resulted in significant tensions between the largely Acholi military and non-Acholi rulers of Uganda (the Kakwa Idi Amin (ruled 1971–79) and the Lango Milton Obote (ruled 1979–85)), including purges against Acholi officers, leading to a coup in July 1985 led by two Acholi officers, Gasilio Olara-Okello and Tito Okello. Soon after, in January 1986, the National Resistance Army (NRA), led by the Banyankole Yoweri Museveni, overthrew the Acholi clique, and Museveni has remained in power to the present day. But the Acholi soldiers did not give up the struggle – rather, they reconstituted themselves as the Uganda People's Defense Army (UPDA) and attempted to evict the NRA from northern Uganda. Although the UPDA began to run out of weapons and ammunition within a year, Acholi nationalism quickly took a new form – the Christian faith healer Alice Aum, who claimed to be possessed by the Holy Spirit Lakwena, founded the Holy Spirit Movement, a military organization which fought based upon Christian principles (singing while fighting and not taking cover), and this movement drew the most dedicated Acholi fighters from the UPDA. Aum soon gave up the struggle and fled to Kenya, but beginning in April 1987, Joseph Kony organized the Lord's Salvation Army, later the LRA, made up overwhelmingly of Acholis, based upon Aum's ideas, and they began to fight a guerrilla war against the NRA, and eventually against other groups that opposed them. Failing to make significant progress against the NRA, the LRA's main target became other Acholis, particularly for the purposes of kidnapping children to serve in their army. Nevertheless, throughout its history, the LRA's leaders and followers have always understood that they are fighting for the Acholi people, regardless of whether or not other Acholis support them and their actions. Religious – specifically Christian – fervor has played a crucial role in convincing the members of the LRA of the rightness of their actions and given them the courage to risk death and reject normal human morality (see Le Sage, 2011; Noll, 2006; van Acker, 2004).

A strikingly similar case is that of the Taliban and Pashtun nationalism. The Pashtuns constitute about a third of the population of Afghanistan and about 15% of the population of Pakistan, and prior to the Soviet invasion of Afghanistan in 1979 they had been the politically dominant group in Afghanistan for hundreds of years. After the Soviets withdrew in 1989 and the communist government they had installed fell in 1992, the different Afghan political parties reached an agreement, known as the Peshawar Accords, which created the Islamic State of Afghanistan and an interim government representing different ethnic groups. The situation was still unstable, however, and the Taliban, which had begun to develop in 1991 in the Pashtun areas of southern Afghanistan based upon a radical fundamentalist interpretation of Islam that had been spread by the mujahedeen who had fought against the Soviets, soon spread their influence, drawing a disproportionate number of recruits from the Pashtuns. They began military operations in 1994 and captured Kabul in 1996, establishing the Islamic Emirate of Afghanistan, but the conflict continued, as the deposed leaders of the Islamic State of Afghanistan reconstituted their forces and formed the Northern Alliance to fight against the Taliban. The Northern Alliance was initially too weak to defeat the Taliban, but aided by the American invasion of Afghanistan in 2001, they were able to seize control of the central government, although they have still not managed to secure the entire country. Although in principle neither the Taliban nor the Northern Alliance were restricted to particular ethnic groups but differed rather in their interpretations of Islam, with the Taliban adopting a much more radical fundamentalist version, in practice the Taliban has been overwhelmingly Pashtun while the Northern Alliance was composed of other groups – Tajiks, Uzbeks, Hazaras, and other groups, but with very few Pashtuns compared to their share of the overall population of Afghanistan. The rise of the Taliban can therefore be understood as a radicalization of Islam by elements of the Pashtun community in an effort to restore the Pashtuns to their traditional position of political dominance in Afghanistan, just as the rise of the LRA in Uganda was the result of the radicalization of Acholi soldiers who feared that they would lose their influence under the Museveni regime. As the Taliban has lost influence in the last decade, Pashtuns have increasingly turned to cooperating with the central government (for example, Hamid Karzai, the President of Afghanistan since 2001, is a Pashtun), but the story of the Taliban remains as a recent example of how religion can be used in an attempt to further national interests in the modern world. Of course, in all of these cases (Arab, Jewish, Acholi, and Pashtun) and similar ones, many members of the national group would deny that this use of religion is consistent with the essence, interests, or values of the nationality, but there are members of the nationality who have a different interpretation of the situation and use religion as a way to direct and motivate national endeavors in a way that they see fit.

Nationalism and religion interact today in myriad ways in different contexts, and to be realistic this is likely to continue in the indefinite future, although there will certainly be trends in how they interact. The clearest trend in this respect is that the interaction between nationalism and religion is increasingly being played out in the politics of democratic institutions,

as religiously inclined people are becoming increasingly sophisticated at playing the game of democracy and maneuvering to improve their position. This can be seen not only in the rise of, for example, the religious right in the United States, the AKP in Turkey, the BJP in India, and religious parties in Israel, but also in the manner in which religious fundamentalists in countries like Egypt and Syria used initially secular street protests, which created the appearance of popular resistance to a non-democratic government, as a means of destabilizing that government for the purposes of eventually taking control of the country. But it is not only religious people who are coming to grips with this reality, as governments and non-religious people are trying to figure out how to deal politically with the situation – for example, secular Israelis in choosing which secular party to vote for will naturally consider statements from secular party leaders regarding whether or not they will sit in a government including one or another religious party, and they may remember whether or not particular leaders have kept such promises in the past (see e.g., articles in Foret and Itçaina, 2012 about the political role of religion in Europe today). But at the same time that religion is coming to be understood as a normal part of democracy, it is also coming to have an increasingly prominent role in motivating violence, which can undermine and circumvent democracy, as religiously motivated people may believe in a law that transcends laws established by democratic institutions, and they may not particularly fear human punishment for transgressing these laws. But whereas in earlier times such efforts were motivated by purely religious conviction, in the present day there is an increasing tendency to use them to advance national interests in the eyes of the actor and to circumvent democratic processes in determining national policy.

See also: European Religion from the Atlantic to the Black Sea; Globalization and Religion; Islam: Asia; Islam: Middle East; Judaism; Religion and Politics in the United States; Religion, Sociology of; Religion: Evolution and Development; Religion: Mobilization and Power; Religiosity; Secularization; Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns.

Bibliography

Ahmed, Mohammed M.A., 2009. America Unravels Iraq: Kurds, Shiites, and Sunni Arabs Compete for Supremacy. Mazda Pub, Costa Mesa, CA.
Ayyad, Mariam, 2014. The Identity of the Israeli Arabs: Between the Hammer and the Anvil. MA thesis. University of Haifa, Israel.
van Acker, Frank, 2004. Uganda and the Lord's resistance army: the new order no one ordered. African Affairs 103 (412), 335–357.
Bader, V., 2007. Secularism or Democracy: Associational Governance of Religious Diversity. Amsterdam University Press, Amsterdam.
Bawer, Bruce, 2006. Crisis in Europe. Hudson Review 1.
Brockett, Gavin D., 2011. How Happy to Call Oneself a Turk: Provincial Newspapers and the Negotiation of a Muslim National Identity. University of Texas Press, Austin.
Brubaker, Rogers, 2013. Language, religion and the politics of difference. Nations and Nationalism 19 (1), 1–20.
Bruckner, Pascal, 2007. Enlightenment Fundamentalism or Racism of the Anti-racists? signandsight.com.
Buruma, Ian, 2006. Murder in Amsterdam: The Death of Theo Van Gogh and the Limits of Tolerance. Penguin, New York.
Buruma, Ian, 2010. Taming the Gods: Religion and Democracy on Three Continents. Princeton University Press, Princeton, NJ.
Caldwell, Christopher, 2009. Reflections on the Revolution in Europe: Immigration, Islam and the West. Doubleday, New York.
Casanova, J., 2009. Religion, Politics and Gender Equality: Public Religions Revisited. Gender and Development Programme Paper Number 5. United Nations Research Institute for Social Development, Geneva.
DeVotta, Neil, 2007. Sinhalese Buddhist Nationalist Ideology: Implications for Politics and Conflict Resolution in Sri Lanka. Policy Studies No. 40. East-West Center, Washington, DC.
Farahat, Cynthia, 2011. The Arab upheaval: Egypt's Islamist shadow. Middle East Quarterly 18 (3), 19–24.
Foret, François, Itçaina, Xabier (Eds.), 2012. Politics of Religion in Western Europe. Modernities in Conflict? Routledge, New York.
Fukase-Indergaard, F., Indergaard, M., 2008. Religious nationalism and the making of the modern Japanese state. Theory and Society 37 (4), 343–374.
Goldberg, Michelle, 2007. Kingdom Coming: The Rise of Christian Nationalism. W. W. Norton & Company, New York.
Gorski, P.S., Altinordu, A., 2008. After secularization? Annual Review of Sociology 34, 55–85.
Greenfeld, Liah, 1996. The modern religion? Critical Review 10, 2.119.42.
Ihalainen, Pasi, 2005. Protestant Nations Redefined: Changing Perceptions of National Identity in the Rhetoric of the English, Dutch and Swedish Public Churches, 1685–1772. Brill, Leiden.
Jones, Carla, Mas, Ruth, 2011. Transnational conceptions of Islamic community: national and religious subjectivities. Nations and Nationalism 17 (1), 2–6.
Khashan, Hilal, 1990. The political values of Lebanese maronite college students. Journal of Conflict Resolution 34 (4), 723–744.
Lentin, Alana, Titley, Gavan, 2011. The Crises of Multiculturalism: Racism in a Neoliberal Age. Zed Books, London.
Le Sage, Andre, 2011. Countering the Lord's resistance army in central Africa. Institute for National Strategic Studies: Strategic Forum 270, 1–16.
Louër, Laurence, 2008. Transnational Shia Politics: Religious and Political Networks in the Gulf. Columbia, New York.
Mandaville, Peter, 2011. Transnational Muslim solidarities and everyday life. Nations and Nationalism 17 (1), 7–24.
Martin, D., 2005. On Secularization: Towards a Revised General Theory. Ashgate, Aldershot.
Mihelj, Sabina, 2007. 'Faith in nation comes in different guises': modernist versions of religious nationalism. Nations and Nationalism 13 (2), 265–284. Journal of Conflict Resolution.
Noll, Christian, 2006. The Betrayed: An Exploration of the Acholi Opinion of the International Criminal Court. UMI Dissertation Publishing, Ann Arbor.
Prunier, Gérard, 2005. Darfur: The Ambiguous Genocide. Cornell University Press, Ithaca, NY.
Sarigil, Zeki, Fazlioglu, Omer, 2013. Religion and ethno-nationalism: Turkey's Kurdish issue. Nations and Nationalism 19 (3), 551–571.
Saunders, Doug, 2012. The Myth of the Muslim Tide. Random House of Canada, Toronto.
Saunders, Robert A., 2008. The ummah as nation: a reappraisal in the wake of the 'Cartoons Affair'. Nations and Nationalism 14 (2), 303–321.
Shenhav, Yehuda, 2006. The Arab Jews: A Postcolonial Reading of Nationalism, Religion, and Ethnicity. Stanford University Press, Stanford, CA.
Solmazturk, E. Haldun, 2008. Turkish secularism is democratic. Middle East Quarterly 15 (3), 67–70.
Stock, Raymond, 2012. The Donkey, the Camel, and the Facebook Scam: How the Muslim Brotherhood Conquered Egypt and Conned the World. Foreign Policy Research Institute, Philadelphia.
Tibi, Bassam, 2009. Islamists approach europe. Middle East Quarterly 16, 47–54.
Tilli, Jouni, 2009. An interpretation of Finnish religious nationalism: the four topoi of theological depoliticisation. Nations and Nationalism 15 (4), 597–615.
Totten, Michael J., 30 April 2009. The Taliban and Pashtun Nationalism. Commentary.
Turan, Berna, 2006. Between Islam and the State: The Politics of Engagement. Stanford University Press, Stanford, CA.
Turner, B.S., 2011. Religion and Modern Society. Cambridge University Press, New York.
Triandafyllidou, Anna, Modood, Tariq, Meer, Nasar (Eds.), 2011. European Multiculturalisms: Cultural, Religious and Ethnic Challenges. Edinburgh University Press, Edinburgh.
Tuttle, Gray, 2005. Tibetan Buddhists in the Making of Modern China. Columbia University Press, New York.
Ünver, H. Akin, 2013. The forgotten secular Turkish model. Middle East Quarterly 20 (1), 57–64.
Zubaida, Sami, 2004. Islam and nationalism: continuities and contradictions. Nations and Nationalism 10 (4), 407–420.
Zubrzycki, G., 2006. The Crosses of Auschwitz: Nationalism and Religion in Post-Communist Poland. University of Chicago Press, Chicago, IL.

Religion, Phenomenology of

John R Hall and Joshua D Hayes, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H. Knoblauch, volume 19, pp. 13093–13096, © 2001, Elsevier Ltd.

Abstract

The phenomenology of religion has uneven relationships to philosophical phenomenology and social-scientific research. Descriptive, analytic, and social phenomenological approaches can be distinguished as ideal types. Descriptive phenomenology elides the philosophical distinction between existence and appearances, thereby bypassing issues of ontology and epistemology. Nevertheless, historical and ethnographic descriptions have been important to the development of the two other approaches, and they yield substantive studies of religion. Analytic phenomenology encompasses structural, hermeneutic, and formal possibilities. Formal phenomenology and Schütz's lifeworldly extension clarified the basis for a social phenomenology of religion, both on a general basis and in comparative and historical phenomenology.

Introduction

The phenomenology of religion refers to a range of approaches that address religion as apprehended, religious experience, and religious worldviews. These methods tend to emphasize understanding religious phenomena according to the subjective meanings and actions of individuals engaging in religious experience, rather than on the basis of neutral observations. Approaches to the phenomenology of religion can be mapped in relation to three ideal-typical alternatives: (1) *Descriptive* attempts to identify the essence of religion in empirical phenomena and to develop typologies of religiosity based on essential characteristics; (2) *Analytical* efforts to use explicit methodologies to develop accounts of phenomena; and (3) *Social* accounts of religious phenomena in their lifeworldly manifestations. Each of these approaches has developed over time (1) through reflexive engagement with issues in the phenomenology of religion itself, and (2) often through engagement with other fields of inquiry such as philosophy, anthropology, history, psychology, and sociology.

The Descriptive Approach

During the eighteenth and nineteenth centuries, phenomenology as a distinctive theme emerged at the intersection of philosophy and religion in the works of Lambert, Kant, and Hegel as they grappled with understanding consciousness and lived experience (Knoblauch, 2001). In the nineteenth century, religious studies came into its own as a distinctive enterprise, in the German *Religionswissenschaft* – the analytic science of religion founded by Max Müller. Differentiated from theology and dogmatics based on anthropological and historical interests, it became known in the United Kingdom under the rubric of comparative religion and in the United States as the history of religions. However, like subsequent formulations of religious studies as an interdisciplinary enterprise, the *Religionswissenschaft* lacked any single established methodology, and its practitioners drew freely from archeological, ethnographic, and historical approaches. Their generic project was to use descriptions of religious phenomena to create typologies of subjective experience through systematic, nonreductionist, and comparative inquiry.

Features of the descriptive phenomenology of religion can be found in the works of William James and Emile Durkheim, though neither construed himself as a phenomenologist. Both authors drew on empirical observations from disparate sources. Durkheim's (1995 [1912]) characterization of 'collective effervescence' – the shared psychic energy experienced in moments of ritualized communion – and James's (1997 [1902]) concepts of the 'sick soul' and 'mysticism' both involved efforts to capture and catalog religious experience in its basic features. Results of these works were expected to be applicable across place, time, and culture.

Two other scholars are especially notable for their more self-consciously phenomenological descriptions. In the *Idea of the Holy*, Rudolf Otto (1923 [1917]) sought to define and describe a specific kind of religious phenomenon. Characteristic of the descriptive approach, Otto avoided normative assessment of religious experience. Instead, he sought to identify the experience of the 'holy,' devoid of any commonly assumed moral underpinnings. For Otto, the 'numinous,' or experience of the divine, was an empirically given and fundamental feature of religion. He held that, though universal, experience of the holy transpires in myriad ways. Variations in specific historical instances do not contradict the numinous as a single irreducible feature of religious experience, he argued. Rather, each variation can be identified as a manifestation of the numinous as a fundamental phenomenon.

Gerardus van der Leeuw's (1963a [1933]) *Religion in Essence and Manifestation* is also emblematic of the descriptive approach. Van der Leeuw created a broad basis for understanding religious phenomena by systematically considering their forms, for example, seeking to identify the essence of angels, power, and sacrifice by comparison across religious traditions.

The descriptive phenomenological project was rigorous, but like nineteenth-century anthropological work, it was largely ahistorical and simultaneously historicist in its avoidance of engagement with epistemological or methodological issues. There was an underlying assumption that by basing analysis on diverse observations, the phenomenology of religion could

divorce itself from the biases and influences of place, time, and method. From diverse scholarly standpoints and at various junctures over the course of the twentieth century, such an assumption came to be regarded as unreflexive and naive.

Analytical Approaches to Phenomenology

Analytical approaches to phenomenology give increased emphasis to methodological and theoretical frameworks. They largely emerged in reaction to the shortcomings of descriptive phenomenology. However, analytical perspectives did not replace descriptive phenomenology. Some authors, intentionally or unintentionally, blended description and analysis. Nor were analytic approaches all of a piece (or equally engaged in the investigation of religion). The major analytical tendencies that emerged can be described in relation to three general approaches: (1) structuralist phenomenology, (2) hermeneutics, and (3) formal phenomenology.

Structuralist Phenomenology

Structuralist phenomenology involves an attempt to focus on observable phenomena, often seeking generalization through induction. This perspective can be seen in scholars such as van der Leeuw (1963a), whose descriptive efforts led him toward categorization of religious phenomena, and Mircea Eliade (1954; 1959) who, for example in *The Sacred and The Profane*, identified a vast array of cases that incorporate a center to sacred space, the *axis mundi*. After analyzing each case in relation to the others, Eliade deduced general rules or 'law-like' systems. The generalizations were then imposed back onto each case as the true meaning or essence of the phenomenon. In this approach, as with Durkheim, experience and meaning were eclipsed, and phenomenology began to approximate other movements toward structuralist analysis of social forms independently of meaning.

Hermeneutics

An alternative approach, hermeneutics, looks to the distinctive ways of understanding and meanings that individuals employ to apprehend and act upon the worlds that they experience. Hermeneutics as a practice is, of course, deeply grounded in the analysis of religious texts. However, in the late eighteenth century, Schleiermacher and, in the early twentieth century, Dilthey shifted hermeneutics into the broader analysis of social and historical texts. Max Weber (1978 [1922]) adopted Dilthey's principle of *verstehen* (lit. 'understanding'; Dilthey used it to signify the analyst's effort to make sense of the perspective that actors have on their situations and actions), and tied it methodologically to the construction of ideal types, yielding a quasi-phenomenological sociology of subjectively meaningful action and its organizational and processual correlates. Weber worked to preserve the significance of subjectivity and meaning that are lost in structural reduction to social forms. However, he did not seek to reduce all meaningful phenomena to their historical specificities, which would remove any hope of generalizability. He therefore employed ideal types to ferret out the degree to which a given case approximated various generic structures of meaning. Thus, a phenomenon such as 'the Protestant Ethic' could be identified as an ideal type, or general sociohistorical model, even if it is never observed completely in a single instance but only in sectarian alternatives that approximated it in one way or another, and to a greater or lesser degree.

Formal Phenomenology

Early descriptive approaches to phenomenology involved efforts to apprehend appearances, and thereby to identify the properties of those things apprehended, posited to have ontological existences on the basis of which perception as phenomena was possible. Modern phenomenologists, by contrast, took more seriously the problem of phenomenal appearances, and their formal turn involved interrogating the character of consciousness that makes sensate appearances possible (for an overview, see Smith, 2011). In modern formal phenomenology, the point of departure is similar to that of social hermeneutics, centered on individuals' understandings of their own and others' meaningful actions. However, the strong formal move, taken by Husserl, was to ground all knowledge in understanding the character of subjective consciousness per se, and to do so through an 'epoché' that 'brackets' the contents of all perceptions and sensations and any assumptions concerning the existence of external objects (e.g., a 'red chair') that might conventionally be taken for granted as giving rise to them. In this 'transcendental reduction,' the objects of perception are no longer assumed to reflect any ontological status. Contrary to the structuralist effort to identify the essential character of experiences (or even, in looser phenomenologies, the objects themselves), Husserl's formal phenomenological approach takes consciousness as the object of phenomenological investigation, most successfully in his *Phenomenology of Internal Time Consciousness* (1964 [1928]).

Scholars responded to Husserl's phenomenology of consciousness by either attempting to further his investigations or criticizing his conclusions as untenable. For social inquiry, the most significant development came in the work of Alfred Schütz, who argued that a social phenomenology could only proceed after first clarifying the nature of consciousness via Husserl's transcendental investigations. As Husserl (1964) showed, individual consciousness transpires as a stream of temporally sequential events. However, Schütz (1967) noted, the cognized and constructed world that an individual experiences in the 'natural attitude' in effect disappears in Husserl's transcendental reduction, leaving no basis for a specifically social phenomenology. Put differently, individuals generally experience the world that they apprehend, including the setting, other people, families, and organizations, as really being there. Thus, a social phenomenology may be based upon Husserl's transcendental investigations of consciousness, but it must proceed by investigating the structure of the lifeworld as apprehended by the individual under the dispensation of the natural attitude, in which the historicity and social relations of the world that appear in consciousness are taken to be real.

The other analytical approaches to phenomenology (not to mention descriptive phenomenology) were not fully

phenomenological, at least in the modern philosophical sense of the term. For its part, Husserl's formal turn in phenomenology occluded any investigation of the objects of consciousness, including ones deemed 'religious.'

Social Phenomenology and Religion

Prior to Schütz, most phenomenologically inspired study of religion took place within fields that eventually became consolidated as the (inter)discipline of Religious Studies, or they were underwritten by an ethnographic anthropology. The exceptions were Emile Durkheim, whose study of religion, however, was only implicitly phenomenological, and Max Weber, whose emphasis on meaningful action opened up the possibility of a comparative hermeneutic sociology of religion. Indeed, Weber's monumental comparative sociology of the world religions and his famous studies of the Protestant ethic can be read in a phenomenological light.

Here, Schütz's intervention was important. By moving from the pure, transcendental phenomenology of Husserl to a phenomenology in the 'natural attitude' of the individual located in the natural and social world, Schütz (1967 [1932]) was able to offer a phenomenological critique of Weber, whom he regarded as offering an observer's typification of meaning, action, and action structures rather than a phenomenology of the social world radiating out from an actor's consciousness. Schütz, with his collaborator Thomas Luckmann (1973), in turn offered a formal phenomenology of the lifeworld. However, Schütz himself only touched upon the subject of religion in the most glancing ways. His project thus left unfulfilled the possibility of a phenomenologically grounded social-scientific approach to religion.

Since Schütz, a specifically phenomenological social-scientific approach to religion has been taken up in two interconnected programs: (1) a general phenomenology of religious experience and meaning, and (2) an historical and comparative phenomenology of religion in relation to alternative patterns of social action and organization.

The General Phenomenology of Religion

As early as 1963, Thomas Luckmann and Peter Berger began to connect the study of religion to issues in the sociology of knowledge, themselves amenable to a phenomenological approach. From there, individually and together, the two scholars elaborated a general approach to the social phenomenology of religion, based in what was to become their extension of phenomenology, designated as the 'social construction of reality'.

Luckmann's (1967 [1963]) analysis locates religion phenomenologically in ways directly tethered to the work of Schütz, but moving beyond it. He argued that religion is, at base, a symbolic universe of meaning, and that concrete historical forms of religion – as church, or sect, or ancestor worship, or so forth – are simply instantiations of that more general phenomenon. Yet religion is not simply one of many such symbolic universes. For Luckmann, it is a primordial one connected to the social formation of the self. Religion, phenomenologically understood, is a fundamental and pervasive social phenomenon, but on a somewhat different basis than that argued by Durkheim. For Luckmann, a transcendent and publicly objectivated worldview, constructed in important ways as a sacred cosmos maintained through ritual action, is the ultimate basis of selfhood. From this phenomenological point of departure, Luckmann was able to theorize about the general character of religious institutions and about modernizing societal developments of social differentiation and 'secularization,' in relation to both religious institutions and individual religiosity.

In turn, Peter Berger (1967, 1969) grounded a general sociological analysis of religion both in relation to classic phenomenological scholars on religion such as van der Leeuw or Eliade, and in the more formal but nonreligious social phenomenology of Schütz (Schütz, 1967; Schütz and Luckmann, 1973). On this basis, Berger especially developed the general phenomenological analysis of religion's capacity to locate a sacred cosmos that simultaneously grounds the social world ontologically and provides a legitimation of it in everyday life, especially important when the reality of that life is challenged by 'marginal situations' such as death or war that might otherwise be taken to throw the otherwise taken-for-granted social world into doubt. Like Luckmann, Berger was particularly concerned with 'secularization' and the rise of social orders grounded outside religion.

The scholarship of Luckmann and Berger provide a key bridge that links the explicitly social phenomenology of Schütz with the classic phenomenological treatments of religion. Their work thus infuses general social theory with a phenomenology of religion.

The Comparative and Historical Phenomenology of Religion

Social science is concerned not only with general accounts of the social but also with the specificities of social constructions of self and identities, meaningful patterns of social action, institutions, organizations, and movements. Thus, a general social phenomenology of religion would need to be elaborated in relation to the substantive constructions of historically given social worlds in order to connect more richly with social-scientific analysis. Already, Berger and Luckmann's interests in modernity and secularization signal this possibility. Similarly, Luckmann (1967: 50) cautioned against defining religion as dealing with the 'supernatural,' and this point accords with much empirical work on religion, from that of Max Weber onward, which demonstrates that the focus of religious sacrality can become oriented to the most diverse loci of identity construction and social action – from mystical experience to ascetic engagement in economic action.

Thus, insofar as phenomenology is concerned with the lifeworld and its myriad possibilities of instantiation, a general social phenomenology of religion is rightly to be supplanted by a comparative phenomenology of lifeworlds. Such a project has been pursued in historicist and ethnographic studies. In addition, Hall (1977; 1981) called for a comparative phenomenology of the lifeworld, and he (Hall, 2009) began to specify this approach in relation to religion by describing a general analytic framework for phenomenological comparison and developmental historical analysis of forms of social process and organization (bureaucracy, competition, community, etc.) in relation to a phenomenology of temporally specified social

action (in the here-and-now, diachronic, and apocalyptic orientations). Such an approach offers a basis to extend the phenomenology of religion in ways that affirm Luckmann's concern with moving analysis beyond the supernatural, at the same time, linking with Weber's recognition that sacrality can be directed to the most diverse spheres of social life, even to economically rational action.

Conclusion

There is no single methodological or epistemological foundation for the phenomenology of religion. Even the major alternatives – descriptive, analytical, and social approaches – are neither mutually exclusive nor contradictory. Rather, they amount to strategies for engaging phenomenological investigation in relation to understandings of essences and lifeworldly enactments, all in relation to social theory. Moreover, the alternative approaches can increasingly be understood in relation to one another. Scholars draw on descriptive, analytical, and social approaches to the phenomenology of religion in varying degrees and complex articulations. Thus, the phenomenology of religion, grounded in an understanding of structures of consciousness and the lifeworld, offers a way of mapping religiosity in its social formations that should become increasingly important to the pursuit of diverse research programs.

See also: Charisma and Charismatic; Globalization and Religion; Millennialism; Mission: Proselytization and Conversion; Mysticism; Myth in Religion; Prophetism; Religion, Sociology of; Religion: Evolution and Development; Religiosity; Religious Organizations; Secularization.

Bibliography

Berger, Peter L., 1967. The Sacred Canopy: Elements of a Sociological Theory of Religion. Doubleday, Garden City, NY.

Berger, Peter L., 1969. A Rumor of Angels: Modern Society and the Rediscovery of the Supernatural. Doubleday, Garden City, NY.

Durkheim, Emile, 1995 [1912]. The Elementary Forms of Religious Life. The Free Press, New York, NY.

Eliade, Mircea, 1954. The Myth of the Eternal Return. Harper & Row, New York, NY.

Eliade, Mircea, 1959. The Sacred and the Profane. Harper & Row, New York, NY.

Hall, John R., 1977. Alfred Schütz, his critics, and applied phenomenology. Cultural Hermeneutics 4, 265–279.

Hall, John R., 1981. Max Weber's methodological strategy and comparative lifeworld phenomenology. Human Studies 4, 153–165.

Hall, John R., 2009. Apocalypse: From Antiquity to the Empire of Modernity. Polity Press, Cambridge, UK.

Husserl, Edmund, 1964 [1928]. The Phenomenology of Internal Time Consciousness. Indiana University Press, Bloomington, IN.

James, William, 1997 [1902]. The Varieties of Religious Experience. Touchstone, New York, NY.

Knoblauch, H., 2001. Religion, phenomenology of. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social & Behavior Sciences. Pergamon, Oxford, UK, pp. 13093–13096.

Luckmann, Thomas, 1967[1963]. The Invisible Religion: The Problem of Religion in Modern Society.

Otto, Rudolf, 1923 [1917]. The Idea of the Holy. Oxford University Press, New York, NY.

Schütz, Alfred, 1967 [1932]. Phenomenology of the Social World. Northwestern University Press, Evanston, IL.

Schütz, Alfred, Luckmann, Thomas, 1973. Structures of the Lifeworld. Northwestern University Press, Evanston, IL.

Smith, David Woodruff, 2011. Phenomenology. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy (Fall). http://plato.stanford.edu/archives/fall2011/entries/phenomenology/.

Van Der Leeuw, Gerardus, 1963a [1933]. Religion in Essence and Manifestation, vol. 1. Harper & Row, New York, NY.

Van Der Leeuw, Gerardus, 1963b. Sacred and Profane Beauty. 1872. Abingdon Press, Nashville, TN.

Religion: Psychiatric Aspects

Sheila M LoboPrabhu, Michael E. DeBakey V.A. Medical Center, Houston, TX, USA; and Baylor College of Medicine, Houston, TX, USA
Kenneth I Pargament, Institute for Spirituality and Health at the Texas Medical Center, Houston, TX, USA; and Bowling Green State University, Bowling Green, OH, USA
James W Lomax, Baylor College of Medicine, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G.T. Harding, volume 19, pp. 13096–13102, © 2001, Elsevier Ltd.

Abstract

Although psychiatry and religion have a long and troubled history with each other, there are good reasons for a rapprochement. Empirical studies have shown that religion can be both a resource for the mentally ill and a source of problems. Patients also generally prefer religiously sensitive care. In recent years, the field has made promising theoretical advances, including work in the areas of religious coping and attachment. Research studies have also underscored the value of spiritually integrated approaches to psychiatric diagnosis and treatment. The article concludes with a description of promising new directions for theory and practice in psychiatry and religion.

Psychiatry and Religion: A Long and Complicated History

Psychiatry and religion have a long and complicated relationship with each other. In ancient times, signs of mental disorder were understood as reflections of imbalanced body fluids, supernatural agents, or a troubled relationship between the individual and God. During the Middle Ages and early Modern period in Europe, mental illness was often interpreted as a manifestation of diabolical forces, evil spirits, or possession. The Modern era and Enlightenment heralded a shift from harsh religious explanations of madness to more biological interpretations and more humane approaches to treatment.

In the twentieth century, the relationship between psychiatry and religion was strongly influenced by Freud's psychological theories. Freud (1927) famously conceptualized religion as a response to childish feelings of helplessness in a world of chaotic forces and need for a father-figure who would provide protection and care. He maintained that religion was illusory and he suggested that it would be better for individuals to face the realities of life rather than to rely on unrealistic, childlike wishes of having a God. Freud's harsh perspective on religion was tempered by some of his followers, such as Jung, Rizzuto, and Winnicott. Jung treated religious beliefs and behaviors in a more positive manner and provided a psychological understanding of terms such as 'soul,' 'sacred,' 'God,' 'transcendence,' and 'evil' (Palmer, 1997). Rizzuto (1979) outlined how the development of the relationship with a primary object (such as a parent) correlates with the transitional object that represents God. She took detailed life histories and spiritual histories from her patients to learn more about their understandings of God. Winnicott (1971) treated 'illusion' as a developmentally necessary bridge between the infant's self-absorption and later involvement in reality. He viewed illusion as an essential part of human experience by which humans meet their deeper needs for artistic, symbolic, and religious meaning and involvement. He posited that religion can be viewed as a transitional phenomenon, neither subjective nor objective. Rather than automatically diminishing or dismissing religious experience as subjective or hallucinatory, merely because physical or objective evidence is absent, Winnicott's writings set the stage for a more constructive view of religion.

In spite of these more benevolent psychiatric perspectives on religion, some tension remains between psychiatry and religious viewpoints (Shuman, 2009). The tension is, in part, a reflection of psychiatric theory critical of religion, such as the work of Freud. However, other factors contribute to this antagonism. Psychiatrists, like other mental health professionals, tend to be less religious than their patients. For instance, 38% of psychiatrists sampled from the American Psychiatric Association reported that religion is very important to them in comparison with 58% of the American population (Shafranske, 2000). As a result of their relatively lower religiousness, psychiatrists may underestimate the salience of religion among patients and fail to appreciate or attend to the religious significance of illness to the patient and family. Furthermore, only 27% of US psychiatry residency programs train residents to work with religious patients (Bowman, 2009); therefore US psychiatrists may be ill-equipped to deal with religious issues arising in treatment (Curlin et al., 2005). In addition, psychiatrists may lack knowledge about faith-based systems that offer resources for treatment, religious and spiritual factors that exacerbate psychiatric symptoms, and religious beliefs and practices that affect treatment acceptance. Psychiatrists may be wary about discussing religion for other reasons as well, such as time constraints, fears that the patient will not be receptive, or fear of being accused of violating boundaries.

Finally, although psychiatry and religion share a concern with understanding and enhancing the health and well-being of people, they offer at least somewhat different perspectives on the causes and remedies of illness. Tensions between psychiatry and religion can grow out of these differences in world views

(Shuman, 2009; Hughes and Wintrob, 2000). For example, the medical community uses the diagnosis–treatment approach to illness, whereas many faith communities hold that symptoms originate from taboo violations and resultant guilt, and believe that confession is the appropriate response. The world views of psychiatry and religion are not necessarily contradictory or incompatible. Psychiatrists and clergy alike often keep both naturalistic and religious explanations in mind, and have used both understandings to formulate approaches to treatment and change. However, the world views of psychiatry and religion can become polarized, and when they do, antagonism and mistrust may occur.

The implications of this tension are significant. Patients seeking psychiatric care may fear being labeled neurotic, pathological, or even psychotic for their religious beliefs. As a result, they may not raise religious issues during psychiatric treatment. Some religious individuals may decline or discontinue psychiatric treatment out of concern that the plausibility of their faith may be challenged or threatened. In this vein, Keating and Fretz (1990) found that more religiously committed patients had more reservations about therapy, fearing that the therapist would ignore spiritual concerns, find them bizarre, reject the idea of communicating with a higher power, or assume that the patient shared the therapist's own religious beliefs. Psychiatrists and clergy may also keep their distance from each other. For example, clergy refer less than 10% of their clients to mental health clinicians; and mental health professionals refer even fewer of their patients to clergy (Mollica et al., 1986; Larson et al., 1988).

Rationale for Rapprochement between Psychiatry and Religion

Even though there are significant historically based tensions between psychiatry and religion, there are also strong empirically based reasons for a closer working relationship between these two domains. The need for rapprochement is rooted in the centrality of religion as a cultural institution, and because the mentally ill often use religion as a resource. These conditions make sensitivity to religious factors an ethical mandate for psychiatrists.

Religious beliefs, practices, and communities are deeply interwoven into the cultural fabric of the United States and other nations in the world. In the United States, for example, over 147 million people are members of religious congregations, and there are approximately 335 000 congregations in the United States (Lindner, 2010). According to the Pew US Religious Landscape Survey (2008), 58% of Americans report that they pray daily, 74% believe in life after death, 68% believe in angels and demons, 59% believe in hell, and 57% agree that it is necessary to believe in God to be moral and have good values. It is important to recognize that religion in the United States is dominated by various denominations of Christianity, although the culture is becoming increasingly pluralistic religiously. Within other countries, religion takes different forms, but it remains a critical, defining aspect of these cultures. Psychiatry recognizes that the development and course of mental illness cannot be understood outside of its larger cultural context. Religion is one key ingredient of this context (Pargament et al., 2013).

Religious beliefs and practices are also common among people with mental illness (see Mohr, 2013, for review). Kroll and Sheehan (1989) studied 52 psychiatric inpatients in Minnesota and found that 94% believed in God, 67% in the devil, 53% prayed or consulted the Bible, and 51% attended church weekly. In a survey of the alternative health practices of people with serious mental illness (SMI), religious/spiritual activities (50%) and meditation (40%) were the most frequently reported practices (Russinova et al., 2002). Elsewhere, Pieper (2004) studied 118 patients in a Protestant mental hospital in the Netherlands, and found that 92% participated regularly in private religious activities and religious services.

Many patients with mental illness indicate that religion offers valuable resources in the process of coping with their major stressors. Tepper et al. (2001) reported that 80% of a sample of 406 psychiatric outpatients with SMI in Los Angeles used religion to cope. In a comparative study of European psychiatric patients and a nonpsychiatric control group, Neeleman and Lewis (1994) found that the psychiatric patients reported a larger number of religious beliefs and practices that offered comfort during stressful life experiences. Religious professionals are another resource for people with mental illness. SMI individuals are as likely to seek treatment from clergy as from a mental health professional (Larson et al., 1988). It is important to add that religion can serve as a resource to caregivers of SMI patients. In a survey of caregivers to people with SMI, many participants reported that God is a source of comfort and strength (Murray-Swank, 2006).

There are ethical mandates to consider religion when providing psychiatric care (American Psychiatric Association Principles of Medical Ethics, 2013). For example, a committee of the American Psychiatric Association (2006) stated that: "Psychiatrists should foster recovery by making treatment decisions with patients in ways that respect and take into meaningful consideration their cultural, religious/spiritual, and personal ideals." Some writers have cautioned against health professionals overstepping their boundaries by addressing religious issues in the context of treatment (Sloan et al., 1999). They maintain that clergy are better suited than healthcare professionals to manage patients' spiritual concerns. Ethical standards in psychiatry, however, explicitly attend to these risks: "Psychiatrists should not impose their own religious/spiritual, antireligious/spiritual, or other values, beliefs and world views on their patients, nor substitute such commitments or religious/spiritual rituals for professionally accepted diagnostic methods or therapeutic practice" (American Psychiatric Association, 2006).

Religion Is Related to Health and Well-Being

Religion can serve a number of valuable psychological functions for people with mental illness. In one project, Mohr et al. (2006) found that among Swiss schizophrenia patients religion provided a number of functions, such as hope, comfort, meaning of life, enjoyment of life, love, compassion, self-respect, and self-confidence; and one-third

of these patients reported that they actually received social support from a religious community. In a study of 115 psychiatric outpatients in Geneva, Mohr et al. (2010) found that people who accessed positive religious coping resources experienced fewer negative symptoms and better social adaptation and quality of life 3 years later. These findings have been mirrored in other investigations of people with mental illness, in which higher levels of religiousness have been associated with the provision of meaning, hope, peace, and comfort (e.g., Fallot, 2007; Kelly and Gamble, 2005).

Several studies have shown that higher levels of religious commitment and involvement are associated with lower prevalence of mental illness and more favorable mental health outcomes (Koenig et al., 2012; Shreve-Neiger and Edelstein, 2004). Religiousness has been robustly linked with lower risk of substance abuse (Johnson, 2013) and successful recovery (Simoni et al., 2002; Pence et al., 2008). Religion can provide patients with guidelines for living a sober life, a sense of power through identification with a larger spiritual force, and a new healthier source of meaning and purpose. In addition, religious commitment may lower suicide rates (Huguelet et al., 2007). Huguelet et al. (2009) reported that one-third of their cases of people with SMI believed that religion offered protection against suicide attempts by providing meaning in life, a way of coping with despair, ethical condemnation of suicide, and positive religious experiences. Studies also suggest that the involvement of clergy in patient care results in increased patient satisfaction, increased medical knowledge, and improved treatment compliance (Parkum, 1985; Saunders and Kong, 1983; Bay et al., 2008; Iler et al., 2001). Finally, in contrast to the view that religion interferes with treatment adherence, one study of 52 patients with psychosis found that religiousness was associated with greater adherence to medication (Kirov et al., 1998). Similarly, Harris et al. (2006) reported that among those with SMI, religious attendance and importance of religious beliefs were both positively correlated with outpatient mental health service use and medication use.

Mental Illness Can Be Associated with Problematic Forms of Religiousness

Mental illness impacts people not only psychologically, socially, and physically, but spiritually as well. A significant number of people experience religious struggles as a result of their illness. Religious struggles reflect tension and doubt about spiritual issues within oneself, with a religious community, or with God (Pargament et al., 2005). For example, a mental illness may lead to feelings of anger toward, abandonment by, or being punished by God. In a survey of a national sample in the United States, McConnell et al. (2006) found that religious struggles were associated with a wide range of mental health symptomatology, including anxiety, depression, paranoid ideation, compulsiveness, and somatization. Although religious struggles may be triggered by psychiatric illness (i.e., secondary religious struggles), there is also evidence that religious struggles exacerbate mental health and physical health problems (i.e., primary religious struggles). Empirical studies of a variety of religious groups in a variety of contexts have tied religious struggles to declines in mood, quality of life, physical health, and even greater risk of mortality (Cummings and Pargament, 2010). Mohr (2013) elaborates on religious struggles and their consequences among Swiss patients with schizophrenia. These include feelings of religious alienation as a result of lack of support from patients' religious communities; a negative sense of self marked by feelings of guilt, despair, anger, fear, and suicidality; and vulnerability to the development of addictions in response to the sense of spiritual loss and emptiness.

Although there is little evidence that religiousness in and of itself generally causes delusions or hallucinations, people may express psychiatric illness through a religious idiom or filter (Greenberg and Witztum, 2001). Religious delusions occur in 36% of schizophrenic patients, 33% of bipolar patients, 17% of patients with substance-induced psychosis, and 26% of patients with other psychoses (Appelbaum et al., 1999). Some patients with religious delusions have committed self-harm, suicide, or homicide (Mohr and Pfeifer, 2009). However, the psychiatrist's task is complicated by the fact that hallucinations can be a normal part of religious experience in certain contexts, and are then experienced as positive. Nevertheless, hallucinations that scare, upset, or disrupt an individual's life are considered pathological. Pathological hallucinations are most often auditory, but can also be visual, tactile, olfactory, or gustatory (Mohr and Pfeifer, 2009).

Sims' (1992) three criteria help to differentiate between religious belief and religious delusion. (1) The experience reported by the patient gives the impression of a delusion rather than a mere religious belief. Both the observed behavior and subjective experience are in concordance with the psychiatric symptom. (2) The presence of other psychiatric symptoms such as hallucinations or mood or thought disorder make it more likely that the individual has a religious delusion rather than mere religious belief. (3) The outcome of the experience seems more in keeping with the natural history and course of a mental disorder rather than a life-enhancing experience.

Patients with Mental Illness Want Religiously Sensitive Care

Perhaps it should not be surprising that many patients would like to receive religiously sensitive healthcare, given the importance of religion as a resource and its significant implications for health and well-being. In one study, 83% indicated that their religious beliefs are closely tied to their state of mental and emotional health, and 75% said it is important to see professionals who integrate their values and beliefs into counseling (American Association of Pastoral Counselors Samaritan Institute Report, 2000). It is important to note that concerns about the openness of mental health professionals to religion and spirituality was the second most common reason for not seeking out help from this group; 15% said that they 'fear that (their) spiritual values and beliefs may not be respected and taken seriously.' African Americans and devout Evangelicals were particularly likely to voice these concerns. Other surveys of individuals

in the United States dealing with physical and emotional problems lead to the same conclusion – a majority of people would like to see religion and spirituality integrated into their treatment in a sensitive and respectful fashion (e.g., Lindgren and Coursey, 1995; Stanley et al., 2011). In spite of this expressed desire, religion is not often addressed in health care. Huguelet et al. (2006) reported that, although a majority of study patients with schizophrenia rated religion as important, only 36% discussed this issue with their clinicians.

In sum, there are a variety of sound empirically based reasons to justify a rapprochement between psychiatry and religion. In fact, over the past 75 years, theorists have made significant advances in conceptualizing the interplay between religion and mental health from a more balanced, nuanced perspective. These theoretical developments have led, in turn, to progress toward more spiritually sensitive psychiatric care.

Promising Theoretical Advances

Recently, two theoretical approaches have been advanced that hold particularly important implications for psychiatry: coping theory and attachment theory. Coping theory, derived from ego psychology (Lazarus and Folkman, 1984), emphasizes the role of religion in helping people understand and deal with major life stressors (Pargament, 1997). Religion provides individuals with coping resources (e.g., spiritual support from a sacred entity, religious reframing of suffering, confession) to assist them in coming to terms with life events, particularly those that reveal human finitude and frailty. In this sense, religion serves an empowering, ego-strengthening function. This may help to explain the high levels of religious involvement among people grappling with mental illness, as noted above. However, people vary according to the breadth and depth of their religious resources. Some are equipped with religious frameworks of meaning, ritual, and connectedness that allow them to withstand and even transform themselves in the face of significant stressors, including mental illness. Others have religious orientations that rest on a less benevolent view of the universe, other people, and the individual's place in the world; this group is more likely to experience profound religious struggles that compound the harsh impact of mental illness.

The effectiveness of coping from the perspective of this theory has to do not only with the individual's breadth and depth of coping resources, but also with the degree to which coping methods are well integrated with the demands of the situation, the needs and aspirations of the individual, and the larger social context. For example, active surrender of control to God's will has very different clinical implications for the individual who must take active steps to manage an emotional illness and the individual in hospice who has exhausted all treatment alternatives. The pursuit of religious growth and development can facilitate mental health and physical health more generally or interfere with those ends when it becomes overly scrupulous. Seeking out religious support for anxiety about gender identity will have different implications depending on whether the individual is part of a conservative or liberal religious tradition.

Attachment theory builds on the work of object relations theory and evolutionary theory. Attachment theorists have emphasized the role of religion in fostering a meaningful relationship between the self and the other (Granqvist et al., 2010). For most religious individuals, this attachment seeking involves a connectedness with a Divine other. People who pursue spirituality outside of traditional religious settings may seek a particular type of relationship with other entities or objects deemed sacred, such as nature, eminent figures, social justice, or work.

Several theorists have elaborated on this notion. Ana-Maria Rizutto emphasized the importance of learning about an individual's picture or image of God. She contends that this God image is heavily influenced by the internal representations of self and important others in the individual's relationship world, especially parents and subsequent key attachment figures (Rizzuto, 1979). Winnicott developed the notion of transitional objects as a key intermediary reflection of how a young child begins to translate experience with early attachment figures into predictions about others when engaging new individuals in the child's relational world (Winnicott, 1953). William Meissner extrapolated Winnicott's notion of the transitional object into his concept of creative illusion formation as a developmental capacity that plays an important positive and healthy role in the meaningfulness of religion participation (Meissner, 1984). Meissner described how growth-producing illusions were responses to objective reality that represented the person's capacity to transform (mere) reality into something reflective of inner significance and hope.

Consistent with this theory, emerging research has found a correspondence between the quality of an individual's parental attachments and attachments to God (Granqvist and Kirkpatrick, 2013). Thus secure attachments to parents are often tied to a secure felt connection to the Divine. Conversely, insecure attachments to parents are linked to anxious, avoidant, and ambivalent attachments to the Divine. There is also evidence that some people form a secure relationship with God in compensation for insecure relationships to parents, as in the case of religious conversions. These findings and the underlying theory on which they are based have important practical implications for psychiatry.

Promising Therapeutic Directions

Psychiatrists and other mental health professionals have begun to integrate greater religious and spiritual awareness and sensitivity into various treatment orientations (Pargament, 2007). Empirical evaluations of the effectiveness of these spiritually integrated treatments have shown encouraging results. Worthington et al. (2011) conducted a meta-analysis of 51 samples from 46 studies, and compared the outcomes of religious accommodative therapies and nonreligious therapies. They found that religious/spiritual psychotherapies (R/S) were as effective as secular psychotherapies with respect to the psychological

outcomes of patients. However, in comparison to secular psychotherapies, R/S therapies were tied to greater improvement among patients with respect to spiritual outcomes. They concluded that, for patients and contexts in which spiritual outcomes are highly valued, spiritually integrated therapies may be a treatment of choice.

Although a variety of spiritually integrated treatments have been developed, most have a cognitive-behavioral therapy orientation. These treatments have been applied to several patient populations, with promising results (Propst et al., 1992; Rosmarin et al., 2010; Harris et al., 2011; Hardman et al., 2003; Margolin et al., 2007, Richards et al., 2006). Even though much of the work on spiritually integrated treatments has grown out of a cognitive-behavioral orientation, practitioners from other therapeutic orientations have also begun to integrate greater spiritual sensitivity into their treatments. These treatment modalities have in common the attempt to provide a healthy revisiting of critical periods of development within the therapist–patient relationship. There is reason to believe that such experiences occur fairly often across healthcare disciplines. However, the impact of these 'sacred' or at least 'important' moments experienced in early development and brought into treatment awaits study (Lomax and Pargament, 2011).

New Directions

In recent years, great progress has been made in clarifying the interface between psychiatry and religion. Yet significant questions remain. First, because much of the work in this area has focused on western religion and medicine, it is important to extend research and practice to other cultures, particularly non-western contexts. Second, research in this arena is moving toward studies of more specific aspects of religiousness as they express themselves in specific forms of psychiatric illness (Hill and Pargament, 2003; Peteet et al., 2011). For example, it is important to consider whether mystical experiences and religious conversion contribute to destructive dissociation or positive integration among patients with significant mental illness. Similarly, research should explore the roles that patients' understandings of 'evil' and 'the Devil' play in their illness. There is some evidence that patients who try to make sense of their illness by invoking diabolical divine agents may be at risk for mental health problems and have increased risk of mortality (Pargament et al., 2001). The language of evil and the demonic may also offer a way to find meaning in otherwise meaningless suffering. Third, longitudinal studies are needed to disentangle the complex links between religion and psychopathology; in particular, the ways in which SMI may impact religious life and the ways religious problems and resources may foster or ameliorate psychopathology. Fourth, to foster more spiritually sensitive care and respond to the dearth of education about religious issues within psychiatric training programs, new models of psychiatric training must be developed and tested (Blass, 2007). Finally, further studies are needed to determine whether spiritually integrated treatments add to the effectiveness of psychiatric treatments. It is also important to learn how psychiatric treatments, including psychotropic medications, impact the spiritual lives of patients. These represent just a few of the many promising areas of study open to exploration by researchers in the twenty-first century.

See also: Buddhism; New Religious Movements; Religion and Health: Sociological Considerations; Religion and Sciences; Religion, Phenomenology of; Religion, Psychology of; Religion, Sociology of; Science and Religion.

Bibliography

American Association of Pastoral Counselors Samaritan Institute Report, November 8, 2000. Greenberg Quinlan Research Inc., Washington, DC.

Appelbaum, P.S., Robbins, P.C., Roth, L.H., 1999. Dimensional approach to delusions: comparisons across types and diagnoses. American Journal of Psychiatry 156, 1938–1943.

Bay, P.S., Beckman, D., Trippi, J., Gunderman, R., Terry, C., 2008. The effect of pastoral care services on anxiety, depression, hope, religious coping, and religious problem solving styles: a randomized controlled study. Journal of Religion and Health 47, 57–69.

Blass, D.M., 2007. A pragmatic approach to teaching psychiatry residents the assessment and treatment of religious patients. Academic Psychiatry 31, 25–31.

Bowman, E.S., 2009. Teaching religious and spiritual issues. In: Huguelet, P., Koenig, S. (Eds.), Religion and Spirituality in Psychiatry. Cambridge University Press, New York, p. 338.

Cummings, J.P., Pargament, K.I., 2010. Medicine for the spirit: religious coping in individuals with medical conditions. Religions 1, 28–53. 40–41.

Curlin, F.A., Lantos, J.D., Roach, C.J., Sellergren, S.A., Chin, M.H., 2005. Religious characteristics of U.S. physicians. Journal of Internal Medicine 20, 629–634.

Fallot, R.D., 2007. Spirituality and religion in recovery: some current issues. Psychiatric Rehabilitation Journal 30, 261–270.

Freud, S., 1927. The future of an illusion. In: Strachey, J. (Ed.), Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 21. Hogarth Press, London, pp. 1–56.

Granqvist, P., Kirkpatrick, L.A., 2013. Religion, spirituality, and attachment. In: Pargament, K.I., Exline, J., Jones, J. (Eds.), 2013. APA Handbook of Psychology, Religion, and Spirituality, vol. 1. APA Press, Washington, DC, pp. 139–156.

Granqvist, P., Mikulincer, M., Shaver, P.R., 2010. Religion as attachment: normative processes and individual differences. Personality and Social Psychology Review 14, 49–59.

Greenberg, D., Witztum, E., 2001. Sanity and Sanctity: Mental Health Work among the Ultra-Orthodox in Jerusalem. Yale University press, New Haven, CT.

Hardman, R.K., Berrett, M.E., Richards, P.S., 2003. Spirituality and ten false beliefs and pursuits of women with eating disorders: implications for counselors. Counseling and Values 48, 67–78.

Harris, K.M., Edlund, M.J., Larson, S.L., 2006. Religious involvement and the use of mental health care. Health Services Research 41, 395–410.

Harris, J.I., Erbes, C.R., Engdahl, B.E., et al., 2011. The effectiveness of a trauma focused spiritually integrated intervention for veterans exposed to trauma. Journal of Clinical Psychology 67, 425–438.

Hill, P.C., Pargament, K.I., 2003. Advances in the conceptualization and measurement of religion and spirituality. American Psychologist 58, 64–74.

Hughes, C.C., Wintrob, R.M., 2000. Psychiatry and religion in cross-cultural context. In: Boehlein, J. (Ed.), Psychiatry and Religion: The Convergence of Mind and Spirit. American Psychiatric Press Inc., Washington, DC, pp. 27–32.

Huguelet, P., Borras, L., Gillieron, C., Brandt, P.Y., Mohr, S., 2009. Influence of spirituality and religiousness on substance misuse in patients with schizophrenia or schizo-affective disorder. Substance Use & Misuse 44, 502–513.

Huguelet, P., Mohr, S., Brandt, P.-Y., Borras, L., Gillieron, C., 2006. Spirituality and religious practices among outpatients with schizophrenia and their clinicians. Psychiatric Services 57, 366–372.

Huguelet, P., Mohr, S., Jung, V., et al., 2007. Effect of religion on suicide attempts in outpatients with schizophrenia or schizoaffective disorders compared with inpatients with non-psychotic disorders. European Psychiatry 22, 188–194.

Iler, W.L., Obenshain, D., Camac, M., 2001. The impact of daily visits from chaplains on patients with chronic obstructive pulmonary disease. Chaplain Today 17, 5–11.

Johnson, B.R., 2013. Addressing religion and spirituality in correctional settings: The role of faith-based prison programs. In: Pargament, K.I., Mahoney, A., Shafranske, E. (Eds.), 2013. APA Handbook of Psychology, Religion, and Spirituality, vol. 2. APA Press, Washington, DC, pp. 543–560.
Keating, A.M., Fretz, B.R., 1990. Christians' anticipations about counselors in response to counselor descriptions. Journal of Counseling Psychology 37, 293–296.
Kelly, M., Gamble, C., 2005. Exploring the concept of recovery in schizophrenia. Journal of Psychiatric and Mental Health Nursing 12, 245–251.
Kirov, G., Kemp, R., Kirov, K., David, A.S., 1998. Religious faith after psychotic illness. Psychopathology 31, 234–245.
Koenig, H.G., King, D., Carson, V., 2012. Handbook of Religion and Health, second ed. Oxford Press, New York.
Kroll, J., Sheehan, W., 1989. Religious beliefs and practices among 52 psychiatric inpatients in Minnesota. American Journal of Psychiatry 146, 67–72.
Larson, D.B., Hohmann, A.A., Kessler, L.G., et al., 1988. The couch and the cloth: the need for linkage. Hospital and Community Psychiatry 39, 1064–1069.
Lazarus, R.F., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.
Lindgren, K.N., Coursey, R.D., 1995. Spirituality and serious mental illness: a two-part study. Psychosocial Rehabilitation Journal 18, 93–111.
Lindner, E.W., 2010. Yearbook of American and Canadian Churches 2010. Abingdon Press, Nashville, TN.
Lomax, J.W., Pargament, K.I., 2011. Seeking 'Sacred Moments' in psychotherapy and in life. Psyche & Geloof 22, 79–90.
Margolin, A., Schuman-Olivier, Z., Beitel, M., et al., 2007. A preliminary study of spiritual self-schema therapy for reducing impulsivity in HIV-positive drug users. Journal of Clinical Psychology 63, 979–999.
McConnell, K., Pargament, K.I., Ellison, C.G., Flannelly, K.J., 2006. Examining the links between spiritual struggles and symptoms of psychopathology in a national sample. Journal of Clinical Psychology 62, 1469–1484.
Meissner, W.W., 1984. Psychoanalysis and Religious Experience. Yale University Press, New Haven, CT.
Mohr, S., 2013. Religion, spirituality, and severe mental disorder: from research to clinical practice. In: Pargament, K.I., Mahoney, A., Shafranske, E. (Eds.), 2013. APA Handbook of Psychology, Religion, and Spirituality, vol. 2. APA Press, Washington, DC, pp. 257–274.
Mohr, S., Borras, L., Rieben, I., et al., 2010. Evolution of spirituality and religiousness in chronic schizophrenia or schizo-affective disorders: a 3-years follow-up study. Social Psychiatry and Psychiatric Epidemiology 45, 1095–1103.
Mohr, S., Brandt, P.-Y., Borras, L., Gillieron, C., Huguelet, P., 2006. Toward an integration of spirituality and religiousness into the psychosocial dimension of schizophrenia. American Journal of Psychiatry 163, 1952–1959.
Mohr, S., Pfeifer, S., 2009. Delusions and hallucinations. In: Huguelet, P., Koenig, H.G. (Eds.), Religion and Spirituality in Psychiatry. Cambridge University Press, New York, pp. 82–83. 88–89.
Mollica, R.F., Streets, F.J., Boscarino, J., et al., 1986. A community study of formal pastoral counseling activities of the clergy. American Journal of Psychiatry 143, 323–328.
Murray-Swank, A.B., 2006. Religiosity, psychosocial adjustment, and subjective burden of persons who care for those with mental illness. Psychiatric Services 57, 361–365.
Neeleman, J., Lewis, G., 1994. Religious identity and comfort beliefs in three groups of psychiatric patients and a group of medical controls. International Journal of Social Psychiatry 40, 124–134.
Palmer, M., 1997. Freud and Jung on Religion. Routledge Press, London, UK.
Pargament, K.I., 1997. The Psychology of Religion and Coping: Theory, Research, Practice. Guilford Press, New York.
Pargament, K.I., 2007. Spiritually Integrated Psychotherapy: Understanding and Addressing the Sacred. Guilford Press, New York.
Pargament, K.I., Exline, J., Jones, J., et al. (Eds.), 2013. APA Handbooks in Psychology: APA Handbook of Psychology, Religion, and Spirituality, vols. 1 and 2. APA Press, Washington, DC.
Pargament, K.I., Koenig, H.G., Tarakeshwar, N., Hahn, J., 2001. Religious struggle as a predictor of mortality among medically ill elderly patients: a two-year longitudinal study. Archives of Internal Medicine 161, 1881–1885.
Pargament, K.I., Murray-Swank, N., Magyar, G., Ano, G., 2005. Spiritual struggle: a phenomenon of interest to psychology and religion. In: Miller, W.R., Delaney, H. (Eds.), Judeo-Christian Perspectives on Psychology: Human Nature, Motivation, and Change. APA Press, Washington, DC, pp. 245–268.
Parkum, K., 1985. The impact of chaplaincy services in selected hospitals in the eastern United States. Journal of Pastoral Care 39, 262–269.
Pence, B.W., Thielman, N.M., Whetten, K., et al., 2008. Coping strategies and patterns of alcohol and drug use among HIV-infected patients in the United States southeast. AIDS Patient Care and Sexually Transmitted Diseases 22, 869–877.
Peteet, J.R., Lu, F.G., Narrow, W.E. (Eds.), 2011. Religious and Spiritual Issues in Psychiatric Diagnosis: A Research Agenda for DSM-V. American Psychiatric Association, Arlington, VA.
Pew Forum on Religion and Public Life, 2008. US Religious Landscape Survey. Pew Research Center, Washington, DC.
Pieper, J.Z.T., 2004. Religious coping in highly religious psychiatric inpatients. Mental Health, Religion & Culture 7, 349–363.
Propst, L.R., Ostrom, R., Watkins, P., Dean, T., Mashburn, D., 1992. Comparative efficacy of religious and nonreligious cognitive-behavioral therapy for the treatment of clinical depression in religious individuals. Journal of Consulting and Clinical Psychology 60, 94–103.
Richards, P.S., Hardman, M.E., Berrett, R.K., Eggett, D.L., 2006. Comparative efficacy of spiritual, cognitive, and emotional support groups for treating eating disorder inpatients. Eating Disorders 14, 401–415.
Rizzuto, A., 1979. The Birth of the Living God: A Psychoanalytic Study. University of Chicago Press, Chicago, IL.
Rosmarin, D.H., Pargament, K.I., Pirutinsky, S., Mahoney, A., 2010. A randomized controlled evaluation of a spiritually integrated treatment for subclinical anxiety in the Jewish community, delivered via the Internet. Journal of Anxiety Disorders 24, 799–808.
Russinova, Z., Wewiorski, N.J., Dane, C., 2002. Use of alternative health care practices by persons with serious mental illness: perceived benefits. American Journal of Public Health 92, 1600–1603.
Saunders, T., Kong, W., 1983. The role for churches in hypertension management. Urban Health 12, 49–51.
Shafranske, E.P., 2000. Religious involvement and professional practices of psychiatrists and other mental health professionals. Psychiatric Annals 30, 525–532.
Shreve-Neiger, A.K., Edelstein, B.A., 2004. Religion and anxiety: a critical review of the literature. Clinical Psychological Review 24, 379–397.
Shuman, J.J., 2009. Theological perspectives on the care of patients with psychiatric disorders. In: Huguelet, P., Koenig, S. (Eds.), Religion and Spirituality in Psychiatry. Cambridge University Press, New York, pp. 19–26.
Simoni, J.M., Martone, M.G., Kerwin, J.F., 2002. Spirituality and psychological adaptation among women with HIV/AIDS: implications for counseling. Journal of Counseling Psychology 49, 139–147.
Sims, A.C.P., 1992. Symptoms and beliefs. Journal of the Royal Society of Health 112, 42–46.
Sloan, R.P., Bagiella, E., Powell, T., 1999. Religion, spirituality, and medicine. The Lancet 353, 664–667.
Stanley, M.A., Bush, A.L., Camp, M.E., et al., 2011. Older adults' preferences for religion/spirituality in treatment for anxiety and depression. Aging and Mental Health 15, 334–343.
Tepper, L., Rogers, S.A., Coleman, E.M., Malony, H.N., 2001. The prevalence of religious coping among persons with persistent mental illness. Psychiatric Services 52, 660–665.
Winnicott, D.W., 1953. Transitional objects and transitional phenomena. International Journal of Psychoanalysis 34, 89–97.
Winnicott, D.W., 1971. Playing and Reality. Basic Books, New York.
Worthington, E.L., Hook, J.N., Davis, D.E., McDaniel, M.A., 2011. Religion and spirituality. Journal of Clinical Psychology: In Session 67, 204–214.

Relevant Websites

American Psychiatric Association, 2013. The Principles of Medical Ethics with Annotations Specifically Applicable to Psychiatry. Available at: http://www.psychiatry.org/practice/ethics/resources-standards.
American Psychiatric Association, 2006. Religious/spiritual commitments and psychiatric practice resource document. Available at: http://www.psych.org/departments/edu/library/apaofficialdocumentsandrelated/resourcedocuments/200604.aspx.
Newport, F., 2007. Americans more likely to believe in God than the Devil, Heaven more than Hell. Gallup News Service. Available at: http://www.gallup.com/poll/27877/Americans-More-Likely-Believe-God-Than-Devil-Heaven-More-Than-Hell.aspx.

Religion, Psychology of

Fraser N Watts and Ryan J Williams, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Watt, volume 19, pp. 13102–13106, © 2001, Elsevier Ltd.

Abstract

From the end of the nineteenth century, psychologists have had a keen interest in the empirical study of religion. Today, the psychology of religion is one among a family of social scientific approaches to the study of religion. Early interests in religious change through conversion and religious experience have continued in the contemporary psychology of religion, and new interests have emerged that have followed the trajectory of developments in broader psychology, including the relationship between religion and brain processes that are investigated through medical technology. The psychology of religion continues to bring into sharp relief issues in the study of religion, including methodological issues related to reductionism, quantitative vs qualitative approaches, and the relationship between the individual and the social environment, and also relates to broader interests in the formation and consequences of human beliefs and behavior.

The psychology of religion was a central part of psychology at its inception as an independent empirical discipline at the end of the nineteenth century. Psychology is also one of the families of human sciences that has made an important contribution to the study of religion. Argyle (2000) and Loewenthal (2000) provide accessible introductions. This presentation will look first at some key historical figures, then at contemporary research contributions, and finally at critical current issues. For an approach focusing on key figures in the history of psychology of religion see Wulff (1997).

History

Early work on the psychology of religion included quantitative methods, such as the questionnaire research of Edwin Starbuck on types of conversion. However, the most durable masterpiece of the early period was undoubtedly the qualitative research of William James. The psychology of religion continued to be a vigorous part of the discipline of psychology through the first quarter of the twentieth century, and the most widely read summation of that period was J.B. Pratt's *The Religious Consciousness* (see Wulff, 1997).

From about 1925, there followed a fallow period in the psychology of religion, in part reflecting the behaviorist direction taken by the discipline generally. However, with the growing emancipation of psychology that took place in the 1960s, there was a revival in the psychology of religion. This recent period has not seen as much development of major psychological theories of religion, indeed many would doubt whether they are appropriate. However, there has been a wealth of investigations in many aspects of religious experience and behavior.

William James

For his *Varieties of Religious Experience*, written at the turn of the century, James (1985) assembled a large quantity of autobiographical material about religious experience. Several of his conceptualizations of this material have become permanent contributions to the psychology of religion, such as his distinctions between 'once-born' and 'twice-born' conversion types, and between 'sick-soul' and 'healthy-soul' types. His description of mystical experience as having the hallmarks of ineffability, a noetic quality, transiency, and passivity has also stood the test of time.

However, James' approach has been criticized (e.g., Proudfoot, 1985) for his overemphasis on the private nature of religious experience as 'the feelings, acts, and experiences of individual men in their solitude.' Whereas James saw religious doctrines and church practices as arising out of individual experience, others have emphasized how much individual experiences are shaped by the religious tradition. These criticisms of James have been reinforced in recent years by the growing awareness of the extent to which experience is 'socially constructed.' There has also been concern about James' overreliance on vivid but unusual experiences and on the special phenomenological qualities of religious experience. Others would want to place the emphasis on how ordinary religious people interpret everyday events in religious terms, emphasizing that the hallmark of religious experience is not its unusual phenomenal quality, but simply the way in which a religious frame of interpretation is applied to experiences of an ordinary kind.

Sigmund Freud

Another rather separate strand in the psychology of religion that goes back to the early twentieth century is represented by the contribution of Freud, and the psychoanalytic tradition stemming from him. Freud published several books on religion, but his most enduring contribution has probably been *The Future of an Illusion* (1962), originally published in 1927. There, he argues that religion is a form of wish fulfillment (which is what he means by 'illusion'), and that God is a projection of the human mind. These were not altogether new ideas, and a similar line had been taken by Feuerbach. However, the fact that Freud formulated them in the context

of a well-articulated general theory of mental functions gave them a new authority.

Though Freud's views have been very influential, there has been a growing recognition that they are highly speculative and less authoritative than was initially supposed. It has also become clear that an alternative, more sympathetic view of religion can be developed within the framework of psychoanalytic theory, something that Freud conceded might be possible. A reconciliation between psychoanalysis and religion was begun by Freud's colleague Oskar Pfister, a Lutheran pastor. More recently it has been shaped by Winicott's concept of the 'transitional' world that was neither wholly internal nor external, but an illusionistic world that plays a constructive role in adaptation (see Meissner, 1984). Though Freud's own work on religion was surprisingly weak in case studies, others have now conducted rich case studies documenting the close parallelism that can often be found between relationships to parents and to God. Of course, this need not be interpreted as showing that God is nothing but a projection of the mind, just that concepts of God are often shaped by personal factors.

C.G. Jung

Jung took an unusual interest in religious symbols and doctrines. Indeed, one of the reasons for his break from Freud was that he was unhappy at what he saw as Freud's disparaging treatment of religion. Jung interpreted religious symbols in relation to the personal journey of 'individuation,' a journey from egocenteredness toward the 'Self,' involving the integration of the unconscious with the ego. Symbols are important both in representing the goal of this journey and also in facilitating it. Central to Jung's approach to religious symbols is his theory of the 'collective unconscious' and its archetypes; Jung does not regard religious symbols as just being features of the personal unconscious.

Jung had a strong interest in symbols of the Self, and saw God and Christ as being such symbols. However, he laments the extent to which religious symbols have now lost their psychological power and are no longer felt to be relevant to the journey of individuation. His work on Christianity included extended essays on the symbolism of the Trinity and the Mass (see Jung, 1969). He also took an extensive interest in Eastern religion, where he also found symbols of individuation, though he was opposed to Westerners taking up Eastern practices such as yoga.

Jung's view of religion is markedly more sympathetic than Freud's, though his reformulation of religion is at points decidedly heterodox. He tended to broaden religion to include anything 'numinous.' Also he emphasized that as a psychologist he could concern himself, not with religious metaphysics, but only with psychological realties (see Palmer, 1997). Though Jung saw 'God' and 'Christ' as psychic images and symbols of individuation, whether there was a 'metaphysical' God was something about which he had no official view. However, it would be a misreading of Jung to imagine that he thought people could work out their own salvation with the resources of the ego. He thought the ego was quite unable to secure 'individuation,' and emphasized the radical otherness of the 'Self,' of which God is a symbol.

The Contemporary Psychology of Religion

Much contemporary psychology of religion employs empirical research methods, and there follows a brief review of representative research findings. For a comprehensive guide to the empirical literature see Hood et al. (2009).

Individual Differences

The study of individual differences has long been an important part of the psychology of religion. The search for a particular kind of personality associated with religious allegiance has generally not been successful, largely because of the marked diversity of religious people. However, there is reliable evidence supporting an association between different personality types and different approaches to religion. This emerged first in connection with social prejudice, where 'intrinsic' religious people (those whose dominant motivation was religion) were found to be less prejudiced on average than extrinsically religious people (who might be involved in religion for more extraneous reasons). The same is true of committed vs consensual religious people (another set of terms for a related distinction). There has also recently been considerable interest in the relation of Jung's psychological typology, as measured by instruments such as the Myers–Briggs Indicator, to various aspects of religious life such as prayer and church leadership.

Mental Health

Mental health (at least as measured by self-report instruments) has been found to show the same relationship as social prejudice to types of religiousness, with intrinsic religious people having good mental health, but extrinsic religious people having relatively poor mental health. Even where religious people are not subdivided, religiousness tends to be associated with better than average scores on dimensions such as Eysenckian 'psychoticism.' When more objective measures are considered, religious people are often found to have lower levels of formal mental illness and abuse of drugs and alcohol. There is also growing evidence that religion has a positive relationship to ways of coping (see Pargament, 1997). There are probably a variety of explanations of the relationship between religion and health. These include 'selection' factors, such as alcoholics being less likely to go to church. However, there may be various ways in which religion can improve health and happiness, such as the therapeutic value of prayer, and the social support provided by the churches (see Loewenthal, 1995).

Developmental Psychology

Children's intellectual understanding of religion has been found to follow the usual path, documented by Piaget for many areas of cognitive development, from concrete to abstract modes of understanding. Early research showed that this applied to the interpretation of Bible stories, but the same general pattern seems to apply to other aspects of religious understanding, such as prayer and sacraments. There have been attempts to extend such 'stages of development' models to other aspects of religion, and to take them into adulthood.

The best-known work of this kind is Fowler's theory of 'faith development.' However, many critics have expressed caution about the proper limits of stage models of religious development. Stage models seem to work best for development in the intellectual understanding of religion, less well for other aspects; they also only seem to work better for religious development in children than in adults. The key issue is whether religiousness changes in a sufficiently predictable and universal way for the concept of development to be appropriate. It has been suggested that Fowler's 'stages' of faith development may be better seen as religious styles. There has been growing recognition that children's spiritual experience runs ahead of their ability to conceptualize it.

Religious Experience

There has been extensive empirical research on mystical and religious experience, though it is beset by the problem of how such experience should be defined. The initial question used in the research of Hay (1987) is characteristic: "Have you ever been aware of, or influenced by, a presence or power, whether you call it God or not, which is different from your everyday self?" Most empirical surveys find roughly a third of the population respond positively to such questions. Interestingly, the proportion is far from zero in atheists and agnostics. The experiences often occur at times of stress and generate feelings of peace and happiness. One of the most debated matters is whether there is a common core to such experiences that is independent of religious tradition and culture. Psychometric research by Hood suggests that there are two groups of descriptors, one focusing on the experience of unity, the other set of descriptors more influenced by religious background.

Attitudes

The measurement of religiousness is an aspect of the study of attitudes and involves the vexed question of the relationship between self-report and behavioral measures. It has been claimed by some that religious believing currently exceeds belonging. Though, on some measures, this appears to be the case, it depends on exactly what measures of each are used. One of the methodological principles to have emerged in the general literature on attitude measurement is that single-item measures of expressed attitudes are a poor predictor of behavior, whereas fuller self-report measures predict better, and there is reason to think that this applies to religious attitudes. Attitude change can be studied in relation to religious conversion.

Conversion

Conversion has long been a central topic in the psychology of religion. It is a multifaceted, and by no means uniform phenomenon. There are well-established differences between sudden and gradual conversion types, with the former being more conservative in religion and in other matters. It has also long been noticed that adolescence is the peak time for conversion, but the interpretation of this needs to take account of the fact that it is probably the peak time for apostasy as well. Conversion often does not represent as complete or permanent a change as is supposed. Also, many of those experiencing a conversion are returning to previously held attitudes; conversions among those with no previous religious affiliation are relatively rare. In recent years, there has been particular interest in conversions to new religious movements. In fact, these often seem to be associated with improvements in the sense of well-being. There is some evidence of mental health problems in those who leave new religious movements, though this by no means always occurs. There has been a growing interest in deconversion as a phenomenon that may parallel conversion.

Charismatic Religion: Speaking in Tongues

Speaking in tongues (glossolalia) is much the best-investigated charismatic phenomenon (Malony and Lovekin, 1985). Psycholinguistic research generally indicates that it is to be regarded as some kind of 'ecstatic utterance' rather than the use of an identifiable language. Social learning is apparently required before people are able to speak in tongues; this is suggested by the way in which people first watch it before joining in, and also get more proficient at it with practice. It has been suggested that it involves some kind of dissociated state of consciousness, perhaps a trance state or a regressive state, but there is as yet no clear support for such hypotheses. It has been difficult to identify personality characteristics of those who speak in tongues. This may perhaps be because it occurs in different ways in different settings, but there is as yet no generally accepted classification. However, there is some evidence that it is associated with positive personality changes, though it is not unique in having such beneficial effects.

Prayer and Meditation

There has been a growing interest in psychological perspectives on prayer (Spilka and Ladd, 2013). For example, aspects of prayer may provide for a kind of cognitive retraining; for example, thanksgiving seems to involve a change in the pattern of attributions people use. There is growing evidence that persistent prayer may result in altered patterns of experience. There has long been evidence that meditation alters people's levels of arousal, though probably not in any unique or distinctive way when compared to other procedures such as relaxation. However, meditation has notable benefits seen, for example, in emotional experience and reduced use of drugs and alcohol.

Perspectives from Evolution, Brain, and Cognitive Science

Alongside the kind of detailed approach to religious phenomena illustrated in the above section, a more general approach to the explanation of religion has emerged from cognitive science.

Cognitive Science

The cognitive science of religion (CSR) has multidisciplinary roots in cognitive psychology, evolutionary biology, neuroscience, and philosophy, and focuses on a variety of religious phenomena, including religious rituals, belief in gods and spirits, and the afterlife. The background assumption is that

religion is a predictable consequence of how normal cognition is structured and organized and is, in that sense, 'natural.' Though there is considerable diversity in CSR, the standard model focuses primarily on supernatural agents that, it is assumed, are overperceived in religion. Much CSR favors explanations in terms of specific mental devices. Some, such as Justin Barrett (Barrett, 2004), attribute perception of supernatural agents to a 'hypersensitive agency detection device' though, without independent evidence of such a device, it is questionable how far that provides an actual explanation. CSR sets out to explain 'religion' as an all-encompassing phenomenon in human history, whereas much current psychology of religion sees that as too broad and focuses instead on more specific phenomena.

Evolution

There has recently been much interest in the evolution of religion (Bulbulia, 2008). Most see religion as being a spin-off of other evolutionary developments, though there is a minority view that argues that religion has adaptive value. Most recent work on the evolution of religion takes a CSR perspective, though some argue that CSR is unnecessarily constrained in its assumptions and that a broader range of factors should be taken into account (Watts and Turner, 2014). Rather than looking for specific explanations of how religion evolved, it may be better to focus on the restructuring of cognitive capacities that led to the 'cultural explosion' of the Upper Palaeolithic period.

Brain

The study of the brain processes involved in religion promises to become a lively research area (Jeeves and Brown, 2009). There has been considerable interest in a possible link between the brain processes involved in religious experience and temporal lobe epilepsy. Though there are similarities in the two sets of experiences, there are also marked dissimilarities, and critics have pointed out that evidence that the same brain mechanisms are involved is weak (Jeeves, 1997). Some have suggested that the frontal lobes are equally important in religion, though more in belief and practice than in experience. Much recent evidence has focused on assessment of brain activity (assessed through fMRI scanning or in some other way) in religious practices such as meditation. Perhaps the most interesting conclusion so far is that there seem to be no regions devoted solely to religion. As evidence accumulates about the brain processes involved in religion, there is growing debate about whether or not they offer a complete explanation of religion.

General Issues

The psychology of religion involves a number of general issues about methodology and interpretation, including the following.

Relation to General Psychology

The psychology of religion currently lacks integrative conceptual frameworks and general theories. A psychology of religion that is no more than a fragmented study of various particular religious phenomena will never look impressive scientifically. The other pitfall to be avoided is, of course, offering a psychological theory of religion that is too simplistic to handle the relevant phenomena adequately. The kind of theory Freud offered, in terms of religion meeting psychological needs, proved neither convincing nor scientifically fruitful, and is not a good precedent to follow. However, the problems faced in studying religion psychologically are not unique. The study of emotion, for example, faces similar problems, in that emotion is as multifaceted a phenomenon as religion. Broadly conceived 'cognitive' approaches to religion may give useful coherence, as they are currently doing in the psychology of emotion. The psychology of religion interfaces with many other areas of general psychology. Indeed, it can be argued that the subject is pursued most fruitfully when its links with general psychology are strong.

Relation to Sociology

The separation between sociology and psychology is in some ways an artificial one, and affects the study of many social phenomena, including religion. Most religious phenomena have both social and psychological aspects. For example, conversion and speaking in tongues have been studied fruitfully from both perspectives. However, social and psychological studies often proceed in isolation from each other; it seems unlikely that either alone will be able to offer an adequate explanation of any religious phenomenon though there is a strong case for the psychology of religion having stronger links with the sociology of religion. This tends to raise the question of whether religion is to be understood in social or individualistic terms. Both perspectives are no doubt needed for an adequate understanding of religion, and the psychology of religion does not exclude a social approach. Though the increasing interest in the brain processes in religion are likely to pull the psychology of religion in a biological direction, the psychology of religion, like other aspects of psychology, will need to remain both a social and a biological science, and to hold together the different methodologies involved.

Relation to Theology and Religious Studies

Finally, what should the relationship be between the psychology of religion and theology? Different countries have different traditions about the relationship between theology and religious studies, some maintaining a sharp distinction between the 'insiders'' perspective of theology and the 'outsiders'' perspective of religious studies (including the psychology of religion); others have tried to maintain links between them. This issue is affected by the postmodern turn in philosophy, which is casting doubt on the possibility of any wholly detached, objective study of anything, including religion. The psychology of religion has been slow to accommodate itself to these general intellectual developments, though it is hard to see that it can stand aside from them indefinitely. One particular issue is whether there is scope for insiders' and outsiders' explanations to run alongside one another. Charismatic phenomena represent an example. It is not clear that the study of the psychological processes involved in charismatic religion excludes a complementary religious account in terms of the activity of the Holy Spirit.

The Nature of Explanation: Beyond Reductionism

Many of these issues are brought together in questions about the nature of explanation in the psychology of religion. Explanations in psychology are seldom complete explanations, and are certainly unlikely to be such in the psychology of religion. This means that religion will not be reduced to psychology, or 'explained away.' Equally, religion is not a 'natural kind,' and all explanations in the psychology of religion will therefore be culturally and historically bound and depend on interpretations of the nature of religion; the psychology of religion can never hope to provide a science of religion that is a 'view from nowhere.' Nevertheless, there is no sharp boundary between the explanation and interpretation of religion, and it is inherent in the nature of religion that it will adopt a reflexive approach.

See also: Human Sciences, History of; Mission: Proselytization and Conversion; Mysticism; Religion and Health: Clinical Considerations and Applications; Religion and Sciences; Religion, Phenomenology of; Religion: Evolution and Development; Religion: Psychiatric Aspects; Religiosity; Religious Fundamentalism: Cultural Concerns; Science and Religion; Women's Religiosity.

Bibliography

Argyle, M., 2000. Psychology and Religion: An Introduction. Routledge, London.
Barrett, J., 2004. Why Would Anyone Believe in God? AltaMira Press, Walnut Creek, CA.
Bulbulia, J., 2008. The Evolution of Religion: Studies, Theories & Critiques. Collins Foundation Press, Santa Margarita, CA.
Freud, S., 1962. The Future of an Illusion, rev ed. Norton, New York.
Hay, D., 1987. Exploring Inner Space: Scientists and Religious Experience, second ed. Mowbay, London.
Hood, W., Spilka, B., Hunsberger, B., Gorsuch, R., 2009. The Psychology of Religion: An Empirical Approach, third ed. Guilford Press, London.
James, W., 1985. The Varieties of Religious Experience. Harvard University Press, Cambridge, MA.
Jeeves, M.A., 1997. Human Nature at the Millennium. Baker Books, Grand Rapids, MI.
Jeeves, M., Brown, W., 2009. Neuroscience, Psychology, and Religion: Illusions, Delusions, and Realities about Human Nature. Templeton Foundation Press, West Conshohocken, PA.
Jung, C.G., 1969. Psychology and Religion: West and East. Princeton University Press, Princeton, NJ.
Loewenthal, K.M., 2000. The Psychology of Religion: A Short Introduction. Oneworld, Oxford.
Loewenthal, K.M., 1995. Mental Health and Religion. Chapman and Hall, London.
Malony, H.N., Lovekin, A.A., 1985. Glossolalia: Behavioral Science Perspectives on Speaking in Tongues. Oxford University Press, New York.
Meissner, W.W., 1984. Psychoanalysis and Religious Experience. Yale University Press, New Haven, CT.
Palmer, M., 1997. Freud and Jung on Religion. Routledge, London.
Pargament, K., 1997. The Psychology of Religion and Coping: Theory, Research, Practice. Guilford Press, London.
Proudfoot, W., 1985. Religious Experience. University of California Press, Berkeley, CA.
Spilka, B., Ladd, K., 2013. The Psychology of Prayer: A Scientific Approach. Guildford Press, London.
Watts, F., Turner, L. (Eds.), 2014. Cognitive Approaches to the Evolution of Religion: Critical Perspectives. Oxford University Press, Oxford.
Wulff, D.M., 1997. Psychology of Religion: Classic and Contemporary Views, second ed. Wiley, New York.

Religion, Sociology of

David Voas, University of Essex, Colchester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by R. O'Toole, volume 19, pp. 13106–13112, © 2001, Elsevier Ltd.

Abstract

The sociology of religion is a branch of social science that aims to describe and explain religion in society through empirical investigation and testable theory. It is concerned with individual beliefs and behavior on the one hand and norms and institutions on the other, and with their reciprocal influences. Sociologists seek to explain how religion is affected by social forces and how religion affects other domains in society.

Introduction

Disciplinary Boundaries

Sociology is concerned with the way that people organize themselves in groups, how those groups influence individual beliefs and behavior, and how different kinds of collectives interact in a society. Some of these groups are defined by their attention to gods or things that are sacred and so are of particular interest to sociologists of religion.

The sociology of religion is a social science; unlike theology or the philosophy of religion, it does not seek to evaluate doctrines, arguments, practices, or ways of life. Nor is it concerned with the history, scriptures, and comparative nature of different religions, as the discipline of religious studies would be, except to the extent that these features are relevant to sociological questions. Instead, it sets out to describe and explain religion in society through empirical investigation and testable theory (Aldridge, 2013; Beckford and Demerath, 2007; Clarke, 2009; Emerson et al., 2010; Fenn, 2000; Roberts and Yamane, 2011; Turner, 2010).

The boundaries between different aspects of the scientific study of religion are not always clear. Political scientists are typically concerned with the influence of religion on political choices, institutions, or practices; geographers may be interested in the spatial distribution of religion. The economics of religion is a relatively new subdiscipline that generally involves the application of rational choice theory to religious phenomena; the metaphor of a market for religion has been widely adopted. Anthropologists are demarcated from sociologists more by method (for example, ethnography) than by subject matter, although in the past they were usually concerned with traditional, non-Western societies. Psychologists are typically interested in individual personality, emotional needs and responses, and cognition, and hence explanations tend to be at the individual level. By contrast, sociologists are more concerned with how social factors affect religious attitudes and behaviors (and vice versa).

Defining Religion

Describing the sociology of religion requires knowing what one means by religion as well as sociology. Religion can be characterized in terms of what it is (substantive definitions) or what it does (functional definitions). Substantive definitions are generally easy to grasp; religion involves beliefs, practices, or institutions that are based on gods or supernatural forces. The difficulty is that not everyone will agree on what is 'supernatural.' and moreover some groups (for example, some intellectual varieties of Confucianism or Buddhism) seem religious without relying on the supernatural. Functional definitions sidestep these problems by stipulating that religion provides meaning, promotes social cohesion, or performs other functions for adherents and society. Almost inevitably, however, some groups or ideologies that few people regard as religious fulfill the same functions. Interesting questions about the extent to which environmentalism, for example, resembles religion then become uninteresting ones about whether the movement really is a religion. In practice, the difficulties around substantive definitions are minor; it may not be possible to set out necessary and sufficient conditions for something to count as a religion, but family resemblances enable us to identify members of the class.

Scholars wrestle with a number of related concepts that are distinct from religion, in particular magic, spirituality, the sacred, and the secular. A good deal of attention has been devoted to the cultural specificity of these terms. Ultimately what matters is utility: one wants to understand each concept, argue about what is important, and use vocabulary that is suited to the task.

The Founders

Religion was a central preoccupation of the founders of sociology, in particular Emile Durkheim (1858–1917) and Max Weber (1864–1920). It also received attention from Georg Simmel (1858–1918) and to a lesser degree Karl Marx (1818–83). The influence of the sociological classics has been large and enduring, arguably to an unhealthy degree. The humanities are in continuous conversation with the past, while science quickly leaves its past behind; the sociology of religion sometimes seems more akin to the former than the latter.

One of the reasons that Durkheim and Weber are so often cited in this context is that they jointly cover religion's role in society. Durkheim stressed its function as social glue; for Weber it could act as social solvent. As the complementary ideas of social cohesion and cultural transformation persistently arise in writing about religion and society, it is natural for scholars to genuflect in the direction of the classics.

Key Debates

The Social or Evolutionary Basis of Religion

Sociologists treat religion as a social phenomenon. For early social scientists, the innovation was to treat it as a human rather than a divine creation. For contemporary sociologists, the challenge comes from a Darwinian direction, via nature–nurture debates. Are humans naturally religious, and if so, why? Is there a hereditary predisposition to be religious?

Evolutionary psychologists and cognitive scientists have offered a number of theories to explain the existence of belief in supernatural entities, the creation of ritual systems, and other features of religion. Some suggest that religion provides direct benefits to individuals and groups in adapting to their environment – for example, in promoting cooperation – and so will be favored by natural selection. Others doubt that religion itself is adaptive, but view it as a by-product of other dispositions that do increase the chances of survival. People may come to believe in the existence of supernatural agents, for example, as a result of developing the capacity to trace effects to their causes, or to recognize the intentions of other beings (Bulbulia et al., 2008).

If demand for religion is universal, then cross-national differences in religious involvement must be explained by variations in supply; this idea is one of the foundations of the market model of religion, at least in its dominant supply-side version. It is clear, however, that many people now have no interest in religion, and it does not seem axiomatic that religion is a necessary feature of life for individuals or societies.

Secularization

Sociology of religion came into being with the question of how modernization affects religion. At the level of social institutions, the answer is clear: religion loses power and authority. The term 'differentiation' describes the separation of government, law, education, medicine, and other realms of life from the control of religious organizations. What remains disputable is the impact on individual self-identification, belief, and practice. Proponents of the secularization thesis assert that various aspects of modern life tend to bring about a decline in the prestige and popularity, as well as the power, of religion (Bruce, 2011); they point to the clear inverse relationship between socioeconomic development and religious involvement. Opponents cite the United States as a counterexample and the persistent religiosity of societies that are now modernizing (Stark and Finke, 2000).

Although the debate is frequently depicted in black and white, much of the recent research uses various shades of gray. It is commonly conceded that traditional religious organizations and orthodoxies are struggling, but many scholars claim that religion is being transformed rather than eroded. Belief and practice, in their view, are individualized, possibly privatized, but still with us. For secularization theorists, the issue is whether these new, less formal belief systems have the personal and social significance that characterized old-time religion.

Fundamentalism and the Persistence of Conservative Religion

As modernization began to undermine support for traditional beliefs, some religious leaders reacted by defending what they regarded as core principles. This movement became known as fundamentalism, using the label adopted around the turn of the twentieth century by American Protestant theologians who sought to identify the fundamental tenets of the Christian faith. Its contemporary common usage tends to be pejorative, particularly as fundamentalists from different religions (Hindu, Jewish, and Muslim as well as Christian) have become active in politics. For many sociologists of religion, however, 'fundamentalism' remains a useful technical term to refer to an important strand of doctrinal conservatism, typically involving literal interpretation of scripture and adherence to traditional values.

The importance of fundamentalism demonstrates the complexity of social change; however strong the current, not everyone is carried in the same direction. Understanding the impact of modernization involves studying reactions against it as well as forms of adjustment to it (Marty and Appleby, 2004).

Sects and New Religious Movements

The history of religion, like the history of thought, is marked by a flowering of creative energy in particular times and places. The Reformation brought a proliferation of new religious groups, some of which established themselves as powerful national churches while others remained resolutely outside the mainstream. The late nineteenth century in the United States was a fertile time for sects, groups in tension with society that claimed to be in unique possession of the truth. The 1960s, 1970s, and 1980s saw a considerable rise in interest in – and also concern about – a variety of 'cults' (a term that can be used in a technical as well as an unflattering sense), some with roots in Hinduism and Buddhism, others in evangelical Christianity, and yet others (like Scientology) that were completely new creations. New religious movements (NRMs) is a generic descriptor of sects and cults (Hammer and Rothstein, 2012).

Innovation is an important topic in sociology, and religion experiences flux as well as persistence. The study of NRMs is, in part, an investigation of how, when, and where society changes. The challenge for researchers in this field is to move from the particular (and many groups have been described in fascinating ethnographic detail) to the general. One needs to know not just how NRMs fit into the story of religious change in the modern world, but also what they tell us about the culture generally. Because the number of people involved is relatively small, 'very little' is a possible answer.

Alternative Spirituality

The diverse currents of belief and practice previously described as 'New Age' and now more often known as 'alternative spirituality' attracted increasing attention from sociologists of religion in the latter part of the twentieth century. Some suggested that this belief system was so well adapted to contemporary conditions that it would displace conventional, institutional religion. Certainly, the focus on the self as a trinity of mind, body, and spirit was particularly attractive to Western

individualists, especially when an ideology of oneness with nature, humanity, or the universe helped to leaven the appearance of solipsism (Lynch, 2007). Alternative spirituality has, however, followed the pattern seen for NRMs, important to a small minority and ridiculed or ignored by most.

The extraordinary variety of ideas that fall under this heading points to a key problem for nonorganized religion. Individuals may want to pick and choose their beliefs, but if the list of options is very long, it becomes hard to identify with any common cause.

International Migration and Religious Diversity

Apart from secularization, the major source of religious change in developed countries has been immigration. The religious landscapes in North America, Europe, and Australia have been transformed by the arrival of millions of people of different religions (Eck, 1997). The growing number of Muslims in Western Europe has been a particularly important development (Open Society Institute, 2010). Religion is coming to be seen as a key social division, and the social scientific study of religion, ethnicity, and migration now overlap. Public concerns about social cohesion and political violence rooted in religious difference have helped to bring the study of religion back from the margins of scholarly research.

Theory and Methods

Levels of Analysis and what Is to be Explained

By its nature, sociology is concerned with relationships between individuals, groups, and society as a whole. Analysis may focus on one of these levels or how they are linked. At one extreme, there are macrosocial theories of church–state relations, or the differentiation of law, medicine, education, and other institutions from religion; at the other, there is microsocial research on conversion, religious practice, and so on. At an intermediate level, one finds studies of religious organizations, for example, charismatic leadership or congregational dynamics.

Religion can be affected by other social forces, and it can itself affect individuals and society. For sociologists, then, religious participation may be what is to be explained or what does the explaining. For quantitative researchers, religious involvement may be the dependent or the independent variable.

As suggested above, especially in relation to debates over secularization, a principal preoccupation of sociologists of religion has been to explain the relative prevalence of religion or religious commitment in different times and places. Many of the main theories of religion aim to account for the success or failure of the religious enterprise in relation to other factors. Researchers study continuity and change; religious growth and decline; the connection between believing and belonging; nontraditional spirituality; the influence of age, gender, and socioeconomic characteristics; intergenerational transmission; cross-national comparisons; and much else besides.

There is much more to the sociology of religion than research on variations of religious commitment. Social scientists are more often concerned with the impact of religion on other aspects of life. In these studies, religion is treated as an explanatory variable. As discussed in more detail below, religion can affect anything from health to family size to trust and tolerance.

Religion versus Religiosity

Measurement of religion in a narrow sense would be simply a matter of recording affiliation (as Catholic or Buddhist, for example). For some people, however, affiliation is purely nominal; others will have a serious personal commitment, seeing faith as important in their lives. What matters may be not only or even mainly one's notional identity or affiliation, but instead one's degree of religious commitment, or 'religiosity.' Religiosity is bound up with attitudes, behavior, and values, while religion per se is arguably more like ethnicity, something that for most people is transmitted to them rather than being chosen by them.

There is no single marker of having a religion or of being religious. A number of factors may be relevant: belief, practice, membership, affiliation, ritual initiation, doctrinal knowledge, cultural affinity, moral sense, core values, external perception, or something else. 'Having a religion' is an ambiguous notion, and 'being religious' is even more difficult to define or observe.

In recent years, it has become conventional to focus on three aspects of religious involvement: belief, practice, and affiliation. The first two dimensions seem fundamental, representing the distinction between the internal (belief in creeds, knowledge and acceptance of doctrine, affective connection) and the behavioral (participation in services, private devotion, and communal activity).

Methods and Scope

The sociology of religion, like the wider discipline, tends to be divided as much by method as by topic. Some researchers specialize in the quantitative analysis of large data sets from social surveys; others use qualitative methods such as personal interviews or participant observation. Although some research projects combine the two, a degree of mutual suspicion is common. Qualitative researchers tend to regard questions with precoded responses as superficial and the statistical analysis of data as inaccessible; survey researchers often see limited interest in small-scale studies producing findings that cannot be generalized.

Survey research on religion is generally concerned with individual faith and practice. The analysis aims to identify the effects of particular factors, controlling for potentially confounding variables. Some studies use aggregate data (for example, the total number of church weddings in a particular year or the breakdown of the Muslim population by ethnicity). Work on religion is being influenced by growth in data processing power and the development of increasingly sophisticated statistical techniques.

Although the sociology of religion has its origins in the West, its current scope is international and cross-cultural. While its concerns are dominated by the role of religion in contemporary developed societies (leaving historians and anthropologists to work on the past and traditional religion), the

discipline does not confine itself to any place or period. Although a large proportion of the research effort has focused on the United States, reflecting the American preponderance in the field, public interest in Eastern spirituality, Islam, the Pentecostal movement, fundamentalism and political extremism, and other transnational issues has encouraged a broadening of geographical scope.

How Religion Is Supplied

Types of Religious Organizations

It is common for sociologists to divide religious organizations into four types: churches, sects, denominations, and cults. Although the terms are Christian, the typology can be applied in other cultural contexts. A church (or within any particular society, the Church) represents the dominant religious tradition; its authority often extends beyond the religious sphere. The sect is more fervent and reforming; it may also claim to hold a monopoly on religious truth, but only a fraction of the population is prepared to recognize those claims. As sects become more worldly and churches are obliged to accommodate the existence of religious competition, they become denominations, mainstream groups that comfortably coexist. Finally, the cult has little respect for religious tradition, and in turn receives little respect from outsiders. The strength of the societal reaction to cults is such that the term itself is pejorative in common use, and sociologists of religion tend to refer instead to 'NRMs.'

This typology may lose its utility as people become less deferential toward religious groups and less concerned about the differences between them. Claims to unique legitimacy are viewed with derision in modern pluralistic societies, and increasing tolerance in conjunction with decreasing orthodoxy means that sects and cults are no longer regarded as especially deviant. All religious groups hover in status between the old categories of denomination and cult, receiving respect or disrespect according to circumstances.

Local Religious Organizations

The organization of religious involvement varies across groups and between societies. The idea of a relatively stable worshipping community that meets for religious services is characteristic of the main monotheistic religions (Judaism, Christianity, and Islam). Activity in other religious traditions is often organized differently, with individuals making visits to shrines or temples at irregular times, on special occasions, or for festivals. The tendency of Hindus, Buddhists, and other non-Christian immigrants in Western countries to develop congregations and national denominational bodies is of considerable interest to sociologists of religion.

Many Christian denominations are organized hierarchically, with local churches grouped into districts (dioceses, circuits, etc.), which in turn report to a provincial or national leadership. The degree of centralized authority may be high or low. Complete localism is achieved by independent or congregational churches that are entirely self-governing. Some 'house churches' eschew the conventional form completely, in that collective worship does not depend on having a dedicated space or direction from religious professionals.

Large, so-called megachurches have become an important part of the Western religious landscape. Many have been created by – and are dependent on – the charismatic efforts of a founding leader. Some set out to bring in 'seekers,' people who are unenthusiastic about conventional churches but hope to find a group or idea that will meet their needs (Thumma and Travis, 2007). The size of the megachurch makes it possible to provide high-quality services, both in terms of religious and secular content.

Religious Specialists

Sociologists study the way that institutions act through people, and the human face of religion is often that of the prophet or priest. These specialist roles depend on some combination of personal charisma and official authority. Even where clergy are in a sense religious functionaries, employed largely to conduct rituals for local congregations, their success may be dependent on individual qualities. Conversely the prophet, shaman, or magician must quickly acquire some kind of public standing or risk being treated as a madman.

While there is a considerable amount of cultural and religious specificity in the deployment of religious specialists, a number of roles are found widely. The prophet/entrepreneur has new ideas to promote, either within an existing tradition or completely outside of them. The evangelist seeks to spread a recognized form of belief or practice. Clergy are responsible for dispensing services on behalf of a religious institution; magicians and shaman are less able to call on external sources of authority, although they may operate within a well-understood system of practice. An intriguingly general type is the monk, nun, or theological scholar, a religious professional who to a greater or lesser extent is separated from ordinary social life.

Ritual and Religious Activity

The nature and timing of religious activity shows both continuity and change. On the one hand, rites of passage (ceremonies marking key life events such as birth, coming of age, marriage, and death) have clearly existed for millennia and continue to be important elements of religious practice. On the other hand, daily or weekly observance might once have been left to religious professionals, but now one sees routine acts of worship involving the community of followers.

Ritual may be highly formalized, with set texts (liturgies) and special clothing for the clergy, but it can also be informal and even spontaneous to some degree. Much religious practice is individual (for example, private prayer) rather than collective, however. The debates over 'alternative spirituality' partly relate to whether the balance is shifting in the direction of privatized activity as well as belief.

Religion and the Media

Since the invention of the printing press, new forms of communication have been used to disseminate religious teaching. Religious broadcasting on radio and television is

a staple in many countries, some of which explicitly require broadcasting companies to devote air time to religion.

The relaxation of regulation in the United States led to the 'televangelism' phenomenon, with a number of conservative Protestant pastors leading the way. It is an empirical question whether such broadcasting attracts new participants, draws the faithful (or their financial giving) away from conventional churches, or merely supplements the religious involvement of churchgoers. There is little evidence that television has revitalized religion in the West, but it has helped to launch some religious leaders into politics.

The Internet offers new opportunities for religious organizations, entrepreneurs, and evangelists. It also produces new threats, by making critical information and commentaries readily available, by showcasing the sheer diversity of possible worldviews, and by providing access to myriad alternative uses of time and attention. It remains to be seen how well religion responds to these challenges.

How Religion Is Consumed

Religion and the Life Course

Whether or not humans are inclined by nature to be religious, it is clearly the case that taste for religion is cultivated within families and the social environment. If one wants to predict someone's religion and religiosity, two pieces of information will take one most of the way: where the person grew up and how religious his or her parents were. The amount of effort parents invest in the religious socialization of their children is strongly associated with the importance they themselves attach to the religion.

As children emerge into adulthood, they may diverge from the family path; in the Western world, the motor of declining affiliation and attendance has been a drift away from religion at this point. Within any given generation, the average level of religiosity appears to be fairly stable over the life course; contrary to common belief, there is little evidence that people become markedly more religious with age. Exceptions are most likely to occur as a result of family formation, when people use their own upbringing as a model for how children should be raised.

Gender Roles

It is well known that women tend to be more religious than men by virtually any of the measures commonly used. The only exception is found in particular groups (for example, Orthodox Jews) where the responsibility for public practice falls on men.

Sociologists typically reject the idea that differences between men and women in religious belief and behavior are largely determined by biology. Boys and girls have traditionally been socialized to be different; girls are trained to be more submissive and compliant, and accordingly are more likely to embrace or at least fail to reject religion. Differential socialization also affects education, labor market participation, and caring choices, which in turn affect religious behavior.

Religion may serve particular functions for women because of their social location. The conservative ethos of religious organizations validates and reinforces the choice to be a homemaker, and the 'love and compassion' narrative of many religions provides meaning and psychological support to those tasked with caring. To the extent that mothers are responsible for the moral socialization of their children, they may find that religious concepts, private observance, and attendance at services assist them in that task.

Ethnicity, Nationalism, and Religion

Religion and tribal identity were bound up in traditional societies, and they remain closely connected in many modern ones. Although the rise of 'universal religions' such as Christianity, Islam, and Buddhism helped to separate the concepts of religion and ethnicity, the separation is far from complete. It is now common for sociologists to discuss 'ethnoreligious groups' (rather than 'race,' for example) as related forms of social identity and potential social division. And because many churches or mosques are associated with specific immigrant groups, scholars studying religion are interested in ethnicity.

Religion continues to be pressed into service in support of nationalism. It helps to define the boundary between one group and another, and ironically appears to be more 'rational' for these purposes than characteristics such as language, race, or customs. The honorary kinship between believers helps to create the 'imagined community' of the nation. The influence runs both ways; a demand for national or ethnic self-defense helps to encourage religious participation, where religion serves as an identity marker.

Mysticism and Religious Experience

Not all religion is organized. In addition to mundane private religiosity, some people have religious or spiritual experiences that are out of the ordinary. The nature of these experiences can be highly varied, and include visions of the divine, mystical enlightenment, extraordinary feelings of contact with a different reality, and so on. These sensations may be consciously sought or unexpected, and welcomed or rejected by institutional religion.

Although most research on this topic has been undertaken by psychologists rather than sociologists, the character and incidence of these experiences are historically contingent and socially patterned. One might even suggest that such apparently spontaneous and subjectively interpreted experiences amount to learned behavior, with the culture providing the repertoire. Different religious traditions encourage different kinds of visions or insights, and the individualism of the modern age is conducive to everything from 'born again' to Buddha-like enlightenment.

Conversion, Switching, Apostasy

As societies have become more religiously diverse, the number of people changing their religious affiliations has tended to increase. The changes may be from one world religion to another or from one denomination to another; these moves are often known as 'conversion' and 'switching' (or 'reaffiliation') respectively, but the terminology can vary. The growing acceptability of having no religious affiliation means that changes into and out of religion are also important.

Sociologists are interested in when, why, and among whom such changes occur. Marriage is often the occasion, as spouses seek to avoid the problems that different religious identities, beliefs, and practices might produce, not least in raising children. Geographical moves may encourage leaving, joining, or switching. The strictness or distinctiveness of the religion of origin has received considerable attention as a possible inhibitor of religious mobility.

The impact of religious diversity on religious involvement remains an open question. Are people more religious when a single church is dominant and defines a taken-for-granted worldview, or are they more likely to participate when numerous alternative churches are available, each working hard to gain adherents?

Religion/Religiosity as Explanatory Factors

As well as being interested in religion for its own sake, sociologists are concerned with the impact of religion on other aspects of life – what are often referred to as 'outcomes.' Religion can affect health, drug and alcohol use, crime and deviance, fertility, mortality, marital stability, social and political attitudes, morality and altruism, tolerance, voting, civic participation, trust and social capital, education, economic activity, happiness, and any number of other areas.

By far the largest literature is in health, where there are journals specifically dedicated to work in this field and thousands of articles have been published (Koenig et al., 2012). The evidence suggests that the positive effects of religion (e.g., on lifestyle, emotional stability, and in social support) dominate the potential negative ones (anxiety, fatalism, etc.), but the balance will depend on the particular case.

There is a strong association between homogamy (marriage within the same group) of all kinds and marital stability: the more similar spouses are in religion (but also in age, ethnicity, education, social class, and so on), the more likely they are to stay married. A further issue is whether some religions are better for marriage than others (or none). The answer is probably affirmative – the theology and communal structures of some groups strongly discourage separation and divorce – although selection bias may also be relevant (if, for example, people with no religion tend to be less traditional and more susceptible to partner change than others).

The influence of religion on fertility has been a standard topic in demography, for example, on its historical role in the European fertility transition, the changing importance of 'Catholic' fertility, or its contemporary significance in less developed countries. Specifically religious norms may matter little unless the religious institutions supporting them play a significant role in the lives of their members.

Religious affiliation may be associated with socioeconomic position for many reasons, including not only group values but also historical advantage, social networks, and preferential treatment. Education is an important determinant of economic success, and educational attainment has historically been associated with religion. An important influence of religion on socioeconomic status may be a protective one relating to religiosity rather than affiliation. There is a substantial literature on the inverse association between religious involvement and criminal activity, drug abuse, and a range of risky or antisocial behavior.

Religious beliefs are likely to influence other social attitudes. In previous decades, it was commonplace for political parties in continental European countries to be associated with particular religious groups (or alternatively to appeal to the self-consciously secular), but these links are rapidly disappearing. Religiosity has become one of the best predictors of voting intentions in the United States, however.

Religion and Public Policy

Religion intersects public policy in a number of areas: church establishment and state regulation of religion; faith-based initiatives; official protection of religious belief, for example, through laws on blasphemy, or protection against discrimination; laws on incitement to religious hatred; the accommodation of religious differences in education and employment; criminal justice; housing and urban planning; immigration; and health care and social services.

Religious groups are major providers of social services, and government-backed 'faith-based initiatives' support such roles. Church action may be direct (for example, in providing sheltered accommodation and hospice care) or indirect (for example, in promoting charitable giving, voluntary service and informal networks of care). Common issues concern the extent to which publicly funded groups engage in religious activity and whether such groups deny their facilities to people of whom they disapprove.

Protection of religious beliefs (in particular through legal prohibition of blasphemy) has tended to be replaced by protection of persons (by outlawing incitement to religious hatred). Freedom of belief is protected by legal barriers to discrimination based on religious or nonreligious views. Conflicts arise when religious individuals claim the right to discriminate against others, with gender and sexuality being the usual grounds.

Religion enters into a number of debates over public regulation of education. Church-run schools – in many instances state-supported – exist in virtually all countries, but the arrival of religious minority groups seeking their own schools has led to fears that separate education will increase social segregation. Pressures from religious minorities may also promote the existence of single-sex schools or increase the cultural sensitivity of some aspects of the curriculum in areas as disparate as art, biology, sport, and of course religion itself. Conflicting pressure groups in Western countries attempt to make state-funded education more or alternatively less Christian. The general issue of accommodating the religious requirements of minority students (for example, for prayer facilities or special holidays) is one that schools share with employers. Finally, dress codes in school have been highly contentious, particularly in France (where years of debate culminated in an outright ban on head scarves and other overt symbols of faith in state schools) and also elsewhere.

Religion can be a sensitive issue in the workplace; issues include equal opportunity in recruitment, work schedules, food supplied on site, holidays, uniforms or dress codes, social activities that tend to exclude some employees (for example,

because they do not drink alcohol), provision of prayer rooms, and offensive behavior directed at members of certain groups. A great deal of attention has been given to providing 'reasonable accommodation' of religious requirements.

Religion affects policy and practice in the areas of policing and criminal justice. Security profiling is by design a form of discrimination, and members of religious minorities (especially Muslims) are disproportionately likely to be identified as suspect. Similarly, police stop-and-search activities have created a good deal of ill feeling. Prejudice or intergroup conflict can lead to the commission of hate crimes against people from particular religions. In prisons, there are religious issues concerning food, chaplaincy, and abusive behavior.

Immigration policies sometimes show favoritism toward people who share the country's dominant religion. Conversely, it can be difficult to obtain visas for minority religious leaders. Authorities may treat arranged marriages unfavorably, which tends to affect particular minority groups.

Housing policies that relate to extended families can have a disproportionate impact on some religious groups. Urban planning problems are generally related to equal treatment of minority applications (for example, permission to build temples and mosques in historically Christian countries, or churches in Muslim ones). A considerable amount of symbolism is involved in marking space with public buildings, particularly ones that seem large and exclusive. There are also issues of fairness; church bells are commonly accepted in European countries, but amplification of the Muslim call to prayer is much more likely to be restricted.

A host of cultural issues with religious roots are significant for providers of health and social services. Some medicines, procedures, or immunizations are not favored by certain religious groups; mixed wards may not be acceptable to everyone; hospitals must accommodate a variety of dietary restrictions; hospital chaplains need to serve diverse groups; postmortem requirements are variable. Similar issues arise in residential care. The take-up of contraceptive and some screening services may vary by religion, and how or by whom female patients are treated is occasionally an issue. Whether religiously motivated circumcision should be performed is also debated. The religion (as well as the race) of prospective adoptive parents has provoked discussion. Members of NRMs or neopagan groups complain that they are sometimes the target of unjustified child protection orders.

Groups concerned with animal welfare have objected to methods used by kosher and halal butchers (who slaughter animals without first stunning them). Jewish and Muslim organizations have protested that their methods of slaughter are humane, and the governments have typically resisted imposing restrictions. Pressures from religious and animal welfare groups will not be easy to reconcile, however, and it seems unlikely that the issue can be settled definitively.

See also: Globalization and Religion; New Religious Movements; Religion and Economic Life; Religion and Health: Sociological Considerations; Religion and Stratification; Religion and Youth; Religion, Family, and Kinship; Religion: Evolution and Development; Religion: Nationalism and Identity; Religiosity; Religious Organizations; Secularization; Women's Religiosity.

Bibliography

Aldridge, A., 2013. Religion in the Contemporary World: A Sociological Introduction, third ed. Polity Press, Cambridge.

Beckford, J.A., Demerath, N.J., 2007. The Sage Handbook of the Sociology of Religion. Sage, London.

Bruce, S., 2011. Secularization: In Defence of an Unfashionable Theory. Oxford University Press, Oxford.

Bulbulia, J., Sosis, R., Harris, E., Genet, R., Genet, C., Wyman, K., 2008. The Evolution of Religion: Studies, Theories, and Critiques. The Collins Foundation Press, Santa Margarita, CA.

Clarke, P., 2009. The Oxford Handbook of the Sociology of Religion. Oxford University Press, New York.

Eck, D.L., 1997. A New Religious America: How a "Christian Country" Has Become the World's Most Religiously Diverse Nation. HarperSanFrancisco, San Francisco, CA.

Emerson, M.O., Mirola, W.A., Monahan, S.C., 2010. Religion Matters: What Sociology Teaches Us about Religion in Our World. Pearson, Prentice Hall, Upper Saddle River, NJ.

Fenn, R.K., 2000. The Blackwell Companion to the Sociology of Religion. Wiley-Blackwell, Oxford.

Hammer, O., Rothstein, M., 2012. The Cambridge Companion to New Religious Movements. Cambridge University Press, Cambridge.

Koenig, H., King, D., Carson, V.B., 2012. Handbook of Religion and Health, second ed. Oxford University Press, New York.

Lynch, G., 2007. New Spirituality: An Introduction to Belief beyond Religion. I.B. Tauris, London.

Marty, M.E., Appleby, R.S., 2004. Fundamentalisms Comprehended. University of Chicago Press, Chicago.

Open Society Institute, 2010. Muslims in Europe: A Report on 11 EU Cities. Open Society Institute, Budapest.

Roberts, K.A., Yamane, D.A., 2011. Religion in Sociological Perspective, fifth ed. Sage, Thousand Oaks, CA.

Stark, R., Finke, R., 2000. Acts of Faith: Explaining the Human Side of Religion. University of California Press, Berkeley, CA.

Thumma, S., Travis, D., 2007. Beyond Megachurch Myths: What We Can Learn from America's Largest Churches. Jossey-Bass, San Francisco, CA.

Turner, B.S., 2010. The New Blackwell Companion to the Sociology of Religion. Wiley-Blackwell, Oxford.

Relevant Websites

http://www.thearda.com/ – The Association of Religion Data Archives.
http://hirr.hartsem.edu/ – The Hartford Institute for Religion Research.

Religions of East Asia

Ian Reader, Lancaster University, Lancaster, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 6, pp. 3947–3953, © 2001, Elsevier Ltd.

Abstarct

This article will present an overview of the main religious forms in East Asia (China, Taiwan, Korea, and Japan). It will outline the main religious traditions, dealing both with organized and established traditions such as Buddhism, Taoism, Confucianism, Shinto, and the New Religions of the region, as well as the amorphous but influential folk religious traditions that form the bedrock to all religious activity in the region. Attention will be paid within this wider folk tradition to beliefs and practices relating to shamanism and to the role of charismatic figures, healing, and millennialism. Issues such as the relationship of religion and nationalism (especially critical in Japan in relation to Shinto) and the integrative functions of religion in East Asia will be discussed, while consideration will be given to new religious movements, especially Japan, as manifestations of a religious response to modernity and as a product of the tense interactions between modern change and adherence to cultural traditions and beliefs. In addition, reference will be made to the scholarship produced within the region, and the significance this has for the wider field of the social sciences.

Comprising China, the Korean peninsula, Taiwan, and Japan and containing a sizeable proportion of the world's population and home to some of the oldest and most advanced cultural systems in the world, East Asia has produced a number of major religious traditions such as Taoism, Confucianism, and Shinto, as well as developing its own distinctive regional forms of Buddhism. Besides these religions there exists also a pervasive folk religious tradition linked to customs, practices, and beliefs that forms the underlying bedrock to the entire religious structure of the region. Especially since the nineteenth century, too, numerous new religious movements have developed in the region, in response to rapid social change and modernity, but often reasserting many of the themes of folk religion and of the traditional religious environment within new forms adapted to the modern age. The region has also developed its own scholastic traditions which are essential for any understanding of its religious dynamics and that provide important new perspectives and alternative voices from those derived from Western academia.

East Asia in Religious Contexts: an Overview

East Asia represents a problematic area of analysis because of its size, vast and varied population, historical and social complexity, the differing historical and social experiences of each country, and its multiplicity of religious forms. This, and the existence of religious expressions associated with particular countries of the region, raises questions about how far one can talk about the region as a whole in religious or cultural terms. An example here would be Shinto, which is specific to the Japanese cultural sphere and which raises questions about the use of religious ideology in the formation of nation states, about cultural identity, and about the relationship between religion and politics. However, while Shinto provides meaningful examples relating to such questions, it cannot necessarily be considered as an exemplar of the patterns of East Asian religion in general, given that no similarly identifiable national religious traditions exist in other countries in the region. Similarly, the religious responses to rapid social change and modernity, which have seen the formation of new religious movements in different parts of the region, cannot necessarily be considered as indicative of a region-wide commonality. Many of the new movements that have arisen in Korea, e.g., draw heavily on evangelical Christian themes as well as the shamanic beliefs and practices which are central to the Korean tradition, while Japanese new religions have drawn much more heavily on Japanese cultural traditions based on Buddhism, Shinto, and folk religion. Indeed, scholars from the region have been reluctant to draw many parallels between the countries of East Asia. Japanese scholars, e.g., prefer to discuss the Japanese new religions in the comparative context of new religions in the West rather than in the Asian context. Rather than expressing similarities and commonalities, indeed, the academic traditions of each country in the region tend to emphasize concepts of cultural uniqueness and difference and where they have sought a comparative focus have looked outside the region.

Religious Traditions of East Asia: Buddhism, Confucianism, Taoism

The East Asian religious spectrum includes historical religions that have influenced the whole region and localized traditions that are influential in some, or one, area but that cannot claim cross-border significance. Three major historical traditions have been influential across boundaries and formed part of the religious constitution of the region as a whole: Buddhism, which although coming from outside, has taken on particular regional characteristics in China, Korea, Taiwan, and Japan, and Taoism and Confucianism, both of which originated in China but have also been influential in Korea, Taiwan, and Japan.

Buddhism

Buddhism entered East Asia along the major trade routes, supported by the patronage of ruling elites who saw it as a means of strengthening the ethos and power of the state. This was particularly so in Japan, where it was adopted as an arm

and pillar of the state from the seventh century onwards. While its growth and development of an extensive monastic system owed much to its promises, expressed in popular Buddhist texts such as the Lotus Sutra, of spiritual emancipation for all, and while it has developed monastic forms and meditative traditions (most famously Zen in Japan) with a primary orientation towards renunciation and spiritual transcendence, Buddhism's main orientation in East Asia has been social, familial, and mundane. As elsewhere, it has used magical symbols and rituals for the amelioration of the living, both individually and in social contexts, and operated as a religion dealing with death and the after-life, offering an eschatological framework of other-worldly salvation through a range of memorial and funerary services which aid the dead spirits in their journey to the other world. It has been through Buddhism that the spirits of the dead have been transformed into ancestors who guard over their living kin and in this way it offers solace to the bereaved while reinforcing the solidarity and continuity of the household.

Confucianism

While the thoughts of Confucius (551–479 BC) may be regarded as the foundations of a moral and philosophical tradition emphasizing filial piety and the moral duties of rulers and subjects, Confucianism has developed into a broader cultural and religious form that has influenced the development of all East Asian societies. While Confucianism has taken on overt religious manifestations in the ritual veneration of Confucius, its primary influences have been ethical and relating to the affirmation of social codes that underpin the existing order of society. Confucianism, besides emphasizing classical learning and the veneration of tradition, affirms the basic goodness of human nature and asserts that the ways individuals conduct themselves affects the basic social order at familial and wider social levels. This notion of individual spiritual responsibility for one's actions, and of the individual as a microcosm, in spiritual terms, of society at large has been central also to many of the East Asian new religions.

Taoism

Although originally a mystical tradition of meditation and passive withdrawal from worldly affairs, Taoism became associated with the search for longevity and the overcoming, through ritual and mystical practices, of the physical limitations of the body. Taoist practitioners in China sought, through the study of alchemy, to find the elixir of life, and they formed cults venerating that deities symbolized longevity. Taoism also developed complex systems of divination and fortune telling which were central to its appeal beyond China. Taoist ideas spread across the region in tandem with other Chinese cultural influences, including Confucianism, Buddhism, and the Chinese writing system. In Japan, Taoist influences were subsumed within the broader cultural sweep of continental culture and it never assumed its own distinct religious form there. However, its influences were pervasive, especially in terms of divination and magical practices. The early Japanese Imperial Court, e.g., developed a Bureau of Divination using Taoist techniques to predict and govern state affairs, while popular divination practices derived from Taoism remain widespread in Japan.

Indigenous Traditions: The Case of Shinto

Although Shinto, literally 'the way of the gods,' is often termed an indigenous Japanese religion, it is more accurately considered as a localized tradition that developed in conjunction with and in antithesis to the foreign religion of Buddhism which entered Japan from the sixth century onwards. The native religious tradition of Japan prior to the entry of Buddhism centered around the worship of local gods such as clan tutelary deities who were guardians of communities and clan lineages. This loose-knit local tradition acquired a specific identity with the entry of Buddhism. Even the term Shinto first came into existence after the incursion of Buddhism, as a means of differentiating the local from the imported tradition. Shinto is, thus, an example of a tradition that developed in response to (and hence in conjunction with) another, external religion, a point that has been emphasized consistently by influential Japanese scholars such as Kuroda Toshio.

Shinto is based in concepts relating to the *kami* – a term usually translated as god or deity, but implying a spiritual life-giving force. *Kami* may be associated with specific natural phenomena or landscape (for in Shinto, the entire world of Japan is considered to be the creation of the *kami* who give life to and animate it) but they may be spiritual entities that embody particular virtues or capacities and that can, if suitably beseeched and venerated, bestow blessings. The Shinto pantheon contains many thousands of *kami* from localized deities to ones of national influence. *Kami* may also be the spirits of humans of note or special accomplishments, such as Sugawara no Michizane, the ninth century scholar whose spirit is venerated as Tenjin, the *kami* of learning.

Shinto, Buddhism, and Religious Integration in East Asia

Interaction rather than conflict between Buddhism and Shinto has been a core theme in Japanese religious history. Buddhism's strategies of assimilation and accommodation have been nowhere more striking than in its dealings with Shinto and its *kami*. Since the *kami* were the guardians of the earth and the landscape, Buddhism adopted them as spiritual protectors of Buddhist temples: the striking vermilion Shinto gateway, representing entry to the *kami*'s sacred area, is rarely absent from Japanese Buddhist temple courtyards and, especially up to the nineteenth century, many religious centers were not so much Buddhist temples or Shinto shrines, but shrine-temple complexes combining both traditions. Many deities are also shared between the traditions, while this complementarity is manifest also in the roles Shinto and Buddhism play in the life cycles of individuals and communities. While Buddhism deals with death rituals and the memorialization of ancestors, Shinto is associated with beginnings and growth. The Shinto gods guard over the beginnings of life and are called upon for protection and support when babies are born and as infants grow: the *miyamairi* rite in which newly born children are taken to shrines to be placed under the protection of the Shinto gods,

remains widely followed in Japan, as is the *shichigosan* festival in November when children of seven, five, and three are given further ritual blessings.

This functional integration, in which Shinto deals with birth and life and Buddhism with death and the hereafter, has strengthened both traditions in the Japanese system. It is not a pattern limited only to Japan, for a dominant characteristic of the religious structures of the entire region has been that of integration of a number of traditions in the lives of the people. Thus, a state official in premodern China might be involved in Confucian rituals at work, Taoist divination rites and practices in his family life, and Buddhist funerary rituals when someone died, while in Korea, too, similar patterns (Confucian public rituals, the use of shamans and diviners for personal concerns, and Buddhist death rituals) could be discerned. It has only been in recent times, with the emergence of exclusive religious groups, such as the evangelical Korean protestant sects and a number of Japanese new religions that refuse to condone traditional religious practices, that this integrative pattern has been challenged. For the majority, however, it remains dominant.

Festivals and the Ritual Calendar

Shinto illustrates the close interaction of religion, festivity, and the calendrical cycle – a theme that permeates much of East Asian religion. Its festivals (*matsuri,* meaning both prayer and celebration) were traditionally a central element in the lives of local communities, and in modern Japan they remain a colorful and vibrant element in the calendar, involving a mixture of sacred ritual and spectacular festive entertainment that may include processions, dances, fairs, tourist attractions, and much celebratory eating and drinking. The most widespread of all such festivals is New Year when people make a first visit of the year to a shrine (*hatsumôde*) to thank the gods for blessings received in the past year and for protection and good fortune in the coming year. Many people also visit Buddhist temples at this time. *Hatsumôde* is highly popular and, since the 1970s, has been performed by a growing number of Japanese: numbers have risen from around 30 million in the late 1950s to around 85 million (out of a population of 120 million) in the 1990s. The growth of *hatsumôde* numbers is not isolated: Japanese scholars have noted a growth in participation in customary religious ritual activities between the 1950s and the early 1990s, and this has led them to question the validity of secularization theories and to show that these do not necessarily have universal validity.

However, such growth masks a shifting pattern within Shinto, in which small, rural shrines have declined due to rural depopulation and because people have preferred to visit better known and more identifiably national shrines further a field. Partially this has been a result of the development of an advanced transportation system but also because, as localized community structures have broken down or been transformed due to changing life patterns, people have sought their primary focus of identification in the national community. Hence, while they once would have performed *miyamairi* and *hatsumôde* at local shrines, tying themselves to local communities and deities, they now do these at shrines of national significance, an act that confirms their identification with the wider national community and reaffirms, in an age of internationalization in which the primary forms of comparison in terms of identity are across national rather than regional or local lines, their sense of Japanese identity. Many nationally well-known shrines receive many millions of worshippers at festival times: Meiji shrine in Tokyo, for instance, receives around four million visitors at New Year.

Shinto, Mythology, National Identity, Nationalism, and the State

The connection between Shinto and national identity indicated by recent patterns of shrine visiting points to a deep-seated characteristic of religion in the region: the close connection of religion, state, and identity. Here, again, Shinto is a striking (if particularized) example. Its myths attribute the creation of Japan and the origins of the Japanese people and its Imperial dynasty to the *kami*. In these myths the Imperial family are ordained divinely to oversee the nation, as they are descended from Amaterasu, the Sun Goddess, who is venerated at the great shrines of Ise, the most important shrines in Japan. The *kami* are the procreators of Japan and its people, who thus exist in a special, unique relationship to their gods and land. This special relationship is central to Japanese concepts of identity and has been associated closely with Japanese nationalism since the Meiji Restoration of 1868, when Japan developed a modern-style sense of nationhood by using Shinto and its mythological links to the Emperor as a focus of national unity and as a means of emphasizing national uniqueness as a chosen people and land. The association of Shinto and the state helped create a myth of racial superiority which legitimated Japan's colonization of Taiwan and Korea, involved the suppression of other religions and played a major part in Japan's fascism and war mongering in the first half of the twentieth century.

In the postwar era, official links between Shinto and the state have been barred under the Constitution, Japan has become a secular state, and Shinto exists predominantly as a shrine-centered tradition based around life cycle and other rituals and festivals. Nevertheless, there remain many people (especially of right-wing persuasion) and many Shinto priests who aspire to a return to a formal association between Shinto and the state. Many of the discussions of Shinto in the modern age thus center around issues of religion, nationalism, and politics.

Underlying Structures and Common Themes – Folk Religion

Although the above-mentioned religions are the most visibly identifiable older traditions of the region, they are underpinned by a pervasive folk tradition. It is a common argument of scholars from the region (e.g., Miyake Hitoshi in Japan) that this amorphous folk religion of shared customs, attitudes, beliefs, and practices deeply rooted in everyday lives and concerns of the people, forms the foundational religious structure of the region. Indeed, many of the ideas and themes that are central to the new religions derive directly from these folk religious foundations, while the new religions themselves

have been analyzed widely as modern articulations of the folk tradition. The folk tradition has also been the filter through which the historical religions of the region have been adopted by and adapted to local cultures. Those religions that have best developed strategies to accommodate or work with this tradition have been the ones which have transplanted most successfully to or developed in the region: examples include Buddhism, whose strategies and doctrines of accommodation have enabled it to absorb and work with local customs, and (in a reverse case) Christianity, whose inability to accommodate or accept folk religious beliefs and customs, notably relating to the ancestors, has proved a major impediment to its growth in much of the region.

The importance and vitality of this folk tradition is such that it has been a major focus of academic study in the countries of the region. Folk studies (or folk religious studies) are a primary academic discipline in Korea (where shamanism, as the cornerstone of the folk and native religious structure, has received especial attention) and Japan, and is generally accorded greater weight as an academic discipline than in the West. It is not just concerned with the indigenous traditions of folk religion; Japanese scholars such as Sakurai Tokutarô (shamanism), Miyake Hitoshi (mountain religion), and Shinno Toshikazu (pilgrimage and legends) have been concerned to map out and discuss the interactions of folk religion and Buddhism, and to demonstrate that 'folk religion' as manifest in Japan should not be discussed solely in Japanese terms but placed in a wider context. Making use also of Western academic studies of religion, scholars such as Miyake have demonstrated the potential universality of folk religious models, as well as broadening the comparative and potentially universalizing dimensions of Japanese scholarship of religion.

Themes of East Asian Folk Religion

The varying elements associated with the East Asian folk religious tradition(s) include emphases on calendar and ritual cycles relating to individuals and communities, the prevalence of shamanic beliefs and concepts relating to the role of spirits (including the spirits of the dead), spirit possession, healing, and the analysis of misfortune and provision of healing, the veneration of ancestors, and various this-worldly beliefs and customs relating to seeking assistance and amelioration in this world.

All these exist alongside or interact with the major traditions or are manifest in and expressed by them. Concepts relating to the spirits of the dead and the ancestors are deeply rooted in the folk religious arena, but the ritual services related to death and the ancestors are most commonly performed through the medium of Buddhism, whose success in East Asia has depended greatly on its ability to adapt to and accommodate folk beliefs in this area. The importance of the ritual calendar and the notion of the yearly cycle of festivals and rituals, originally based in the folk tradition and centered around agricultural rites, has already been mentioned in the context of Shinto in Japan, but it extends throughout the region in differing forms.

Perhaps the most predominant beliefs in the folk tradition relate to psychic interpretations of causation (of illness, misfortune, and, indeed, good fortune). Beliefs that this world does not exist in isolation but is influenced by other realms, and that those living in this world can be affected by the actions of spirits, mean that, in folk terms, any problem or event may be seen as having spiritual as well as physical causes. Thus, an illness may have psychic causes (e.g., the actions of an unhappy spirit) as well as physical (e.g., the effects of germs) ones, and hence solutions to problems require spiritual attention (e.g., ritual performances, acts of penance) to deal with its spiritual aspects. A strong shamanic tradition has existed throughout East Asia, in which spiritual healers (very often female, and with claims to be able to either travel between this and the spiritual world, or to make contact with it and find ways of appeasing the spirits who are afflicting the living) play a major role. This shamanic tradition is perhaps the most ubiquitous element in the Korean religious universe, but it also remains important in the rest of the region. Studies of Japanese religious behavior in modern times show that people are far more likely to turn to a local healer or diviner than to a Buddhist priest when they have personal problems. This loose-knit tradition of healers and diviners serves also as a counterbalance to the formal religious systems of Buddhism and Shinto, whose priesthoods are almost entirely male, and provides a mode of religious expression for spiritually inclined females, who have remained marginalized in the formal traditions. By contrast, the vast majority of shamanic spiritual healers in Korea and Japan are female.

Millennial Themes in East Asian Religious Culture

An important theme in East Asian religion that shows how religious ideas may be expressed across the various traditions and cultures of the region relates to millennial beliefs and expectations of future worldly salvation. These are found throughout the region, articulated through the folk religious traditions, Taoism, Buddhism, and, more recently, the new religions. Expectations of a new dawn, the arrival of a spiritual kingdom on earth, and/or the appearance of a celestial or spiritual savior who will bring justice, redeem the sick, punish the unworthy, and transform or renew the world have been especially potent in China, sometimes with Taoist influences but often associated with Buddhism, but they have emerged in Korea and Japan as well. In China, various Taoist-inspired movements such as the second century CE Way of Great Peace movement have rebelled against the state in the hope of establishing a new paradise on earth, while Buddhist-inspired movements have also been inspired by the concept of imminent salvation on earth, often centered on the future Buddha Maitreya (whose cult also inspired similar millenarian impulses in Japan). In China, too, there have been repeated uprisings of the millennial White Lotus Society, which combined Pure Land Buddhist ideas along with folk, Taoist, and other elements, while the Pai Shan-ti Hui movement, which incorporated Christian ideas and sought to create a heavenly kingdom of eternal peace on earth, caused the bloodiest and most dramatic of all such millennialist uprisings, the Tai Ping rebellion of 1849–64 which claimed millions of lives.

Millennial themes are prevalent in many of the new religions. In Korea, numerous religious movements have fused extreme forms of evangelical Protestantism with dramatic

millennial visions. The Tami Church (known also as the Mission for the Coming Days) attracted attention when it announced that the apocalypse would begin on October 28, 1992 and that its members would be raptured up to heaven. While this case of failed prophecy damaged the Tami Church, it has not dampened the emphasis on apocalypticism, which remains a central theme in much of Korean Protestantism, and which forms one of its cardinal attractions for the local populace.

The most striking manifestation of millennialism in Japan has been in the new religions that have arisen since the mid-nineteenth century, initially through the actions of charismatic female religious leaders such as Nakayama Miki (1798–1887) of Tenrikyo and Deguchi Nao (1837–1918) of Ômotokyô, but most recently among the 'new' new religions of the late twentieth century, many of which have emphasized that the world is at a critical turning point in which the old, materialist civilization will be swept away as a new spiritual world dawns. While many of the adherents of earlier millenarian movements were poor and disadvantaged, and the thrust of their movements was initially towards seeking social justice, the beliefs of the latest wave of new religions have been conditioned more by concerns that this world is facing extinction because of nuclear or environmental problems, and that a spiritual revolution is essential in order to avoid destruction. The most dramatic and famous of these later millennial movements was Aum Shinrikyô, which believed that a final war between good and evil was inevitable and in which the vast majority of humanity would be wiped out prior to the advent of a more spiritually oriented world.

New Departures: The New Religions

Since the mid-nineteenth century massive economic development, the impact of Western influences, and immense social change and population movement have produced recurrent tensions between traditional social, cultural, and religious structures, and the individualizing patterns of modernity. In religious terms, the most striking sign of such tensions and developments has been the emergence of new religious movements throughout the region. This phenomenon has been most pronounced in Japan but is discernible also in the rest of the region.

The new religions have generally been eclectic, drawing on the primary religious inspirations of the societies in which they have emerged: Korean new religions, e.g., might draw on Buddhism, shamanism, and Confucianism, while Shinto (along with Buddhism and the folk tradition) is a source of inspiration for many Japanese new religions. In very recent times, influences have come from outside the region as well, with evangelical Christian movements especially prominent in Korea and new-age themes appearing in the most recent new religions in Japan.

The new religions have been especially effective in dealing with the tensions between modernity and tradition, through their articulation of many aspects of traditional folk religious belief (e.g., concepts of psychic causation) in ways that fit with the needs of individuals in the modern world. They have flourished mostly in the growing conurbations of East Asia, providing a new sense of belonging and community identity for populations that have moved from rural areas to the cities for work reasons. While providing individual solutions to problems and focusing on understandings and definitions of the self in modern terms, they have equally emphasized tradition, reaffirming many of the social parameters of traditional culture and morality. The solutions they pose to individual problems center around shamanic concepts of spirit possession and magically-charged healing rituals, but combined also with reaffirmations of traditional moral structures and obligations, notably Confucian ethical values of filial piety. A typical example taken from a new religion in Japan might be that a personal misfortune or illness is the result of having neglected to perform appropriate rituals for one's ancestors and of causing their wrath: this breach of ritual propriety is a moral lapse which needs to be rectified through repentance and self-reflection, as well as participating in healing rituals. As such many of the new religions display a pronounced interweaving of magical practices alongside affirmations of morality; as various Japanese scholars have noted (e.g., Shimazono 1992), rather than magical practices decreasing as ethical and moral considerations become more strongly emphasized, the Japanese new religions show that the two often operate in tandem, the latter increasing and being articulated more pronouncedly as magical practices become more widespread.

The region's new religions almost invariably owe their origins to the charisma and inspirational leadership of individuals who claim to have direct contacts with other realms or to bring new revealed truths to the world. While serving as a medium for a newly revealed deity, however, a common pattern is that the charismatic leader gradually displaces the deity as a source of inspiration, authority, and worship, often becoming regarded as living deities and venerated accordingly. Their charisma, however, may continue after death, a point often emphasized by Japanese scholars interested in the study of charisma: authority and succession to the leadership generally (following traditional East Asian familial patterns) passes through the family lineage to a son or daughter, suggesting that charisma is in part related to the concept of blood and lineage, and the deceased founder becomes an even more central figure of worship, with their mausoleum often becoming a pilgrimage center for the religion.

Aum Shinrikyô: New Religions and Violence

In 1995 world attention was directed towards the millennial Japanese new religion Aum Shinrikyô and its blind charismatic leader, Asahara Shôkô, because of its nerve gas attack on the Tokyo subway. While Aum was not the first or only religious movement in East Asia in modern times to commit acts of violence the sensational nature of the subway attack, Aum's use of chemical weapons, and the subsequent evidence showing that it had engaged in a protracted series of crimes emanating from its religious activities and beliefs, caught the attention of defense and law agencies as well as scholars world wide.

The Aum affair has special significance for the sociological study of religion for many reasons. It showed similarities with

patterns of violence that manifest in other religious movements in the West (e.g., the Peoples Temple, the Rajneesh movement, and the Order of the Solar Temple) and appeared, in an extreme form, to fit into a wider pattern of the production of violence within certain forms of newly formed charismatic, world rejecting millennial movements. It provides a dramatic case study for those interested in the relationship between religion and violence, and several of those who have studied Aum have interpreted it as an example of the violence-producing dimensions of religion (e.g., Shimazono 1997, Reader 2000). As an example of how new religions can produce violence primarily because of their doctrinal orientations and internal dynamics, the case of Aum provides a significant counter-example to the arguments widely held by many scholars of new religions in the West, that movements that engage in violence do so primarily in response to excessive external pressures placed on them by a hostile society.

Scholarship on the New Religions and the Problem of East Asia

Mention has been made of the scholastic traditions in the region, especially of Japan, and some of the analyses and ideas they have transmitted. The new religions of Japan provide a case in point where work done by Japanese scholars has added to the wider understanding of new religions across the globe. Japanese scholarship generally has examined the 'new' new religions in the context of late twentieth century urban society and modernity, and as such has thus recognized, either explicitly or implicitly, the parallels that could be drawn between Japan as a first-world, modern, and technological society, and Western countries. Western scholars specializing in new religions have also been drawn to comparative studies of the Japanese new religions because of these recognizably common influences and factors. Much of the best Japanese scholarship on the new religions of Japan has taken account of Western theoretical perspectives, while many of the scholars in Japan have worked closely with Western scholars of new religions, mutually influencing and being influenced by them.

This has placed contemporary Japanese scholarship on the new religions in an international framework, and has made Western scholars more than ever aware of the importance, when seeking to develop comprehensive or universal models and categories in the study of religion, of the Japanese situation. Certainly no sociologist of religion concerned with the rise of new religions in the modern world can ignore the work being done in this field in and relating to Japan, just as no student of religious violence or of millennial movements could afford not to take note of Aum Shinrikyo. However, the corollary of this development has been to separate out Japan from the rest of East Asia in terms of the study of new religions: rather than viewing parallels within the region, e.g., Japanese scholars (and Western scholars working on Japan) have tended to examine Japanese new religions in the comparative context of Western new religions. This, however, means that the (potential) parallels between Japanese new religions and those of the rest of East Asia are focused on less clearly.

In other words, the problematic question raised earlier about how far one can discern or discuss the region in broad terms, is in some areas being answered in practical terms by contemporary scholastic work in and centered on Japan, which associates Japan far more with the industrialized West than, and thereby draws implicit boundaries with, the rest of geographical East Asia. While this does not invalidate the conceptual category or notion of East Asian religions, it does indicate the problematic aspects of talking of the region in any uniform sense. Not only do the traditions themselves take on different guises in the different parts of the region, but the manner in which they may be studied and the lessons that may be drawn from them may also highlight questions of difference rather than commonality. To what extent future studies will further add emphasis to the issue of difference and challenge the viability of treating the region as an entity is a moot question of great interest and significance to the field as a whole.

See also: Buddhism; Charisma and Charismatic; Folklore; New Religious Movements; Religion, Sociology of; Religion: Nationalism and Identity; Sociological Traditions; Tradition: Social; Youth Movements.

Bibliography

Buswell, R.E., 1992. The Zen Monastic Experience: Buddhist Practice in Contemporary Korea. Princeton University Press, Princeton, NJ.
Ching, J., 1993. Chinese Religions. Macmillan, Basingstoke, UK.
Faure, B., 1991. The Rhetoric of Immediacy: A Cultural Critique of Chan/Zen Buddhism. Princeton University Press, Princeton, NJ.
Grapard, A.G., 1992. The Protocol of the Gods: a Study of the Kasuga Cult in Japanese History. University of California Press, Berkeley, CA.
Hardacre, H., 1986. Kurozumikyo and the New Religions of Japan. Princeton University Press, Princeton, NJ.
Hardacre, H., 1989. Shinto and the State. Princeton University Press, Princeton, NJ.
Hori, I., 1968. Folk Religion in Japan. University of Chicago Press, Chicago.
Jones, C.B., 1999. Buddhism in Taiwan: Religion and the State, 1660–1990. University of Hawaii, Honolulu, HI.
Kendall, L., 1985. Shamans, housewives and other restless spirits: women in Korean ritual life. University of Hawaii Press, Honolulu, HI.
Kuroda, T., 1975. Nihon no chûsei no kokka to shûkyô [The Medieval Japanese State and Religion]. Iwanami Shoten, Tokyo.
Kuroda, T., 1981. Shinto in the history of Japanese religion. Journal of Japanese Studies 7 (1), 1–21.
Kyburz, J.A., 1987. Cultes ou Croyances au Japon: Kaida, une Commune dans les Montagnes du Japon Central. Maisonneuve à Larose, Paris.
Maspero, H., 1981. Taoism and Chinese Religion (trans. Kierman F A). University of Massachusetts Press, Amherst, MA.
Miyake, H., 1974. Nihon shûkyô no kôzô [The Structure of Japanese Religion]. Keiô Tsûshinsha, Tokyo.
Miyake, H., 1989. Shûkyô Minzokugaku [Folk Religious Studies]. Tôkyô Daigaku Shuppankai, Tokyo.
Ooms, E.G., 1993. Women and Millenarian Protest in Meiji Japan:Deguchi Nao and Ômotokyô. Cornell University, Ithaca.
Overmeyer, D.L., 1976. Folk Buddhist Religion: Dissenting Sects in Late Traditional China. Harvard University Press, Cambridge, MA.
Overmeyer, D.L., 1986. Religions of China: The World as a Living System. Harper and Row, San Francisco.
Reader, I., 2000. Religious Violence in Contemporary Japan: The Case of Aum Shinrikyô. Curzon and University of Hawaii, London and Honolulu, HI.

Sakurai, T., 1978. Nihon shûkyô no fukkô kôzô. Kôbundô, Tokyo.
Shimazono, S., 1992. Gendai kyûsai shûkyôron [The Study of This-worldly Salvationist Religions]. Seiyusha, Tokyo.
Shimazono, S., 1997. Gendai shûkyô no kanôsei: Oumu to bôryoku [The Potential of Religion in the Modern World: Aum and Violence]. Iwanami Shoten, Tokyo.
Shinno, T., 1991. Nihon yugyô shûkyôron [Studies of Japanese Ascetic Religious Travel]. Yoshikawa Kobunkan, Tokyo.
Smyers, K.A., 1998. The Fox and the Jewel: Shared and Private Meanings in Contemporary Japanese Inari Worship. University of Hawaii, Honolulu, HI.
Teiser, S.F., 1988. The Ghost Festival in Medieval China. Princeton University Press, Princeton, NJ.

Religions of India

Peter van der Veer, Max Planck Institute for the Study of Religious and Ethnic Diversity, Göttingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A large number of Hindu religious traditions refer, positively or negatively, to a set of religious books, called the Vedas. These scriptures were, originally, the intellectual property of a priestly caste of Brahmans and the basis of their rituals and philosophies. Buddhism, Jainism, and Sikhism are major Indian traditions which explicitly reject the authority of the Vedas and the social and ritual hegemony of Brahmans. Islam and Christianity are major traditions in India which have been increasingly stigmatized by Hindus as 'foreign' to Indian civilization. In the modern period, the expansion of the colonial state is crucial for defining the location of religion in India. The general pattern in all Indian religions is that modern movements attack traditional priestly leadership, so-called 'backward' popular religion, and emphasize free, lay access to scripture and to religious debate. In this way, the foundation of religious authority changes and becomes mediated by education and new forms of communication, such as print, and later radio and television. This development also entails a growing participation of religious movements in modern mass politics. The greatest competition in the political arena is between Hindus and Muslims and has led to the partition of British India into India and Pakistan. Hindu nationalism needs to construct a Hindu majority in India, but runs counter to a number of anti-Hindu and anti-Brahmanical movements. Especially in the 1980s and 1990s, it has been successful in launching campaigns against Muslims and Christians to bolster 'Hindu unity.' Not only Hindus, but also Muslims and Sikhs, have been increasingly active in the political arena.

Indian civilization contains a great number of religious traditions, such as Hinduism, Buddhism, Jainism, Islam, Christianity, and Sikhism. All these traditions have influenced each other and have become part and parcel of Indian civilization in one way or another. A large number of Indian traditions refer, positively or negatively, to a set of religious books, called the Vedas. These scriptures were, originally, the intellectual property of a priestly caste of Brahmans and the basis of their rituals and philosophies. Brahmans as a caste form the apex of a hierarchy of status categories which can be seen as an early form of the later much more elaborate Indian caste system. Buddhism and Jainism emerged around 500 BC in protest against the Brahmanical monopoly of Vedic ritual. They are essentially ascetic religions which have much in common with other speculative philosophies which continued to accept the authority of the Vedas. Buddhism spread from India to the rest of Asia, but became marginal in India, while Jainism did not expand outside of India, but continued in India as the religion of, primarily, businessmen. Christianity came to Kerala in the first centuries of the Common Era and in other varieties with the Portuguese to Goa in the sixteenth century and with other Europeans in the centuries thereafter. Islam came to India at the end of the first millennium, along the coast with traders and in the northern plains as an element of the military and religious expansion of a variety of groups from West and Central Asia.

Hinduism

The majority of the Indian population today calls itself Hindu. Hinduism is a term that only in the eighteenth century came to be used for a variety of religious traditions which dominate the Indian religious landscape. When they are Brahmanical, these traditions refer to the Vedas, but when they are not they can refer to a number of religious texts, oral or scriptural. Gods like Vishnu and his reincarnations Rama and Krishna are very popular. Likewise, a god like Shiva, often represented in an aniconic, phallic form, is worshipped all over India. In many localities, one finds local manifestations of gods like Vishnu or Shiva. A special and important category of gods are Mother Goddesses which often protect against evil and diseases. The typical form of worship of images of gods is the offering of gifts like flowers or sometimes goats or buffaloes. These temples are serviced by Brahman or ascetic priests. Temples are everywhere, but especially in sacred places, which are somehow connected to the story of the god. Hindus make pilgrimages to these places, especially at auspicious times in the lunar calendar. There is also an elaborate complex of rituals which are performed in the domestic context, often related to moments in the life cycle. Finally, there are groups of priests, specialized in warding off evil and witchcraft and performing mortuary rites.

In terms of the social organization of religion, there are five elements in Hindu traditions. First, there is caste. Brahman groups have continued from the earliest times to dominate a large part of religious beliefs and practices. Ritual purity and sacred knowledge are the foundations of their supremacy. Other castes have also created cults of their own, but often they are seen of lower status than the Brahmanical cults. Second, there are rituals of kingship. There is the conception of a generic relationship between royal deity and human king which is reflected in ritual practice. In a large number of regions in India one finds a network of temples and palaces which are mutually interdependent. Third, there is the level of the village. The local community is often ritually represented in a number of festivals. Fourth, there is the notion of world renunciation. Ascetics have for a long time up to the twenty-first century built monastic orders and drawn lay support for their religious and worldly activities. They are innovators of religious traditions, but also, as free agents in village society, important long-distance traders, money brokers, and military entrepreneurs until the coming of the colonial state. Fifth, there is the concept of bhakti (devotion). One cult after another sprang up in India

during the second millennium, channeling waves of devotional enthusiasm. Often ascetic leaders were involved in this and in a number of cases they revolted against the Brahmans and the caste system. By that token they were sometimes able to find a distinct religious tradition, such as the religion founded by Guru Nanak, which evolved gradually into Sikhism and defined itself as different from Hinduism.

Hinduism is a term which has developed from 'Hindu,' a term used by Muslims who came to the river Indus or Sindh (from Sanskrit Sindhu) in the second half of the first millennium. 'Hindu' designates the inhabitants of this area of Al-Hind or Hindustan and their customs and beliefs. The term Hindu was made into an 'ism' by the British in the eighteenth century. The conceptualization of Hinduism as one particular religion among other religions and theoretically comparable to them can be understood in relation to a number of philosophical arguments about the universality of the category of 'religion' in the European Enlightenment. The conceptualization of their religious traditions by Hindus themselves had evolved along lines of philosophical argumentation about dharma (the sacred order of things) by religious scholars who are easiest divided in followers of the God Shiva and followers of the God Vishnu. Their philosophical arguments were often accompanied by ritual injunctions, especially for temple worship. Below this level of high philosophical argumentation there were a number of devotional movements which can broadly be divided into those which focus on the worship of images and those which have an aniconic image of God. The British introduced the essentialized and unified concept of Hinduism as a religion in an already existing arena of lively polemics about religious difference.

Modern Period

In the modern period, the expansion of the colonial state is crucial for defining the location of religion in India. In the eighteenth century, the British East-India Company followed a traditional policy of religious patronage benefiting the religious establishments in the areas which came under their growing influence. This policy was further supported by a strong interest in Hindu scripture from a viewpoint of comparative linguistics and comparative religion. This so-called 'orientalism' of the late-eighteenth century was vehemently attacked at home and in the colony by utilitarian administrators and evangelical missionaries who were united in their rejection of anything Hindu. In education and religious worship, the colonial administration was forced to withdraw from giving its patronage to religious establishments and a policy of 'religious neutrality' was established gradually in the nineteenth century.

The response of Indian religious thinkers, institutions, and movements to colonial interference under the aegis of religious neutrality has been central to the formation of modern religions in India during the nineteenth century. Especially attacks by evangelical missionaries on religious practices and social customs led to both reformist and conservative responses. In Bengal, Rammohan Roy (1772–1833) studied Christianity and felt great affinity with the rational critique of religious orthodoxy launched in Britain by Unitarians. In 1827, he founded the British Indian Unitarian Association, but his study of the Vedas and Upanishads quickly led him to explore the limits of Unitarian universalism. In 1828, he decided that a universal, rational religion had to be based upon the Vedas and Upanishads and created the Brahmo Samaj, a small, but influential elite movement, propagating a deist and universalist kind of religion, based upon Hindu sources. The movement is particularly opposed to 'superstitious customs' of 'ignorant people,' 'deceived' by their Brahman leaders.

Rammohan was in the forefront of public action. In January 1830 he, together with 300 residents of Calcutta, presented a petition to Governor-General William Bentinck in support of government prohibition of widow immolation, a practice found in some parts of India when after the demise of a husband the widow throws herself on the pyre of her husband to be burned with him. Rammohan rejected this practice on the basis of his reading of Hindu scripture. He distinguished authoritative sources, such as the Vedas, from other sources. It is significant that he did not refer to authoritative interpretations by learned Brahmans, but relied entirely on his own judgment. This is an important step toward the laicization of Hinduism, that is to say, the formation of religious movements, carried by lay believers and relatively independent of traditional, sacred interpreters of the Hindu canon. The Brahmo Samaj was the cradle of an intellectual Hinduism which was at the end of the century personified by Swami Vivekananda (1863–1902), the founder of the Ramakrishna movement.

Movements like the Brahmo Samaj are sometimes called Neo-Hindu because they have been clearly influenced by European thought and Christianity. On the other hand, however, they are also in direct interaction with more traditional thinkers and movements. A good example is the illiterate saint Ramakrishna (1836–88), a priest of a Kali temple near Calcutta whose devotional practices, ecstasies, and unconventional behavior was very attractive to the intellectual Brahmo Samaj adherents, because he seemed to signify a certain untainted Hindu authenticity. The great reformer Vivekananda took Ramakrishna as his spiritual guide.

Vivekananda became famous after his visit to the World Parliament of Religions which was organized in Chicago in 1893. He offered a packaged Hinduism, in which spirituality and yoga (mental and physical discipline) were keywords, first to a world audience and later to an audience at home. Central to Vivekananda's religious message is the notion of 'practical Vedanta,' in which he combined the Vedanta (= philosophical commentaries on the Vedas) with participation in social activism rather than with ascetic withdrawal from the world, as had been common in Hindu traditions. It is this message which could be tied up with nationalism and political activism as was done by political leaders such as Mohandas ('Mahatma') Gandhi (1869–1948).

Also outside of Bengal we see a number of movements which interact both with Hindu tradition and with modernity, as brought by colonial rule. The Arya Samaj is the most important among them. Founded by Dayananda Saraswati (1824–83), an ascetic who had been initiated in one of the most austere monastic orders in Hinduism, it is an iconoclastic movement which identifies image worship with the decline of an earlier pure Vedic religion. It attacked also the caste system and its system of marriage and the traditional Brahmanical priesthood. The Arya Samaj became and remains one of the central religious pillars of Hindu nationalism.

Public Sphere

It is not only Hinduism which is developing modern movements in the new public sphere created under colonial rule. The general pattern is an attack on traditional priestly leadership, on 'backward,' popular religion and an emphasis on free, lay access to scripture and to religious debate. In this way, the foundation of religious authority changes and becomes mediated by education and new forms of communication, such as print, and later radio and television. Muslim movements such as the Deobandis, the Ahmadiyyas, and the modernist Aligarh movement share with the Hindu movements of the period the defense of their religion against the onslaught of missionary criticism and the attempt to formulate an answer to modern intellectual challenges. Clearly, they are all deeply political since they mobilize people against certain aspects of colonial rule as well as against aspects of their religious traditions and traditional leadership. They are the basis for the formation of the Muslim political party, the Muslim League, and for Muslim participation in the Congress Party. Also the Sikhs have to respond to colonial modernity. The British gave the Sikhs preferential treatment in their recruitment for the army, which made the rather fluid Sikh identity both more valuable and more fixed. The Singh Sabha movement is especially important, since it challenges the control over the centers of Sikhism, the temples or gurdwaras. These centers were owned by hereditary priests, but following the general pattern of laicization the Singh Sabha movement demanded that they came under the control of the lay community of Sikhs. Control over the committee which governs these temples is still one of the biggest prizes in Sikh politics. It is also the basis of the power of the Sikh political party, the Akali Dal.

The twentieth century witnessed a further politicization of religious identity. The greatest competition was between Hindus and Muslims. The colonial state reinforced this competition by creating separate job opportunities, separate civil laws, and separate electorates. The advent of mass politics brought with it the mass mobilization of people on the basis of reified identities. Mohandas Gandhi, the great leader of the Congress party, was brilliant in his strategies to dismantle British power, but these strategies were based explicitly upon Hindu practices. Both he and, to a lesser extent, Jawaharlal Nehru were perceived by Muslims as Hindu despite their open secular politics. Radical movements emerged which challenge the idea of India as a plural civilization. 'To be an Indian is to be a Hindu' was a slogan of V. D. Savarkar (1883–1966), one of the founding ideologies of Hindu nationalism. Movements such as the Rashtriya Swayamsevak Sangh (RSS) and the Hindu Mahasabha wanted to establish Hindu hegemony in independent India. The assassins of Gandhi in 1947 emerged from such organizations. On the Muslim side, the Muslim League became gradually more and more radical until it demanded Pakistan, a homeland for Muslims, in 1942 and ultimately forced the Partition of India. The tragedy of the violently enforced migration of Hindus to India and Muslims to Pakistan, the division of East and West Pakistan which led in 1971 to the foundation of Bangladesh, the territorial dispute about Kashmir, and the threat of warfare between these two nuclear powers were all results of the political mobilization of religious difference on the Indian subcontinent.

Nationalism

Hindu nationalism needs to construct a Hindu majority among the Indian population to realize its claims to political hegemony. It runs not only against the resistance of non-Hindu religious communities, such as Muslims or Sikhs, but also against that of so-called 'untouchables' and 'tribals' who form a large minority in that section of the population which has to be constructed as a Hindu majority, but who are severely discriminated within Hindu practices. There are a number of anti-Brahman, and sometimes anti-Hindu, movements among these underprivileged groups which demand special protection and positive discrimination from the state. Most of these demands concern education and employment, but participation in Hindu rituals and access to temples are also important in the mobility of these groups. One of the most important movements among 'untouchables' was led by Bhim Rao Ambedkar (1891–1956), a Mahar untouchable from West India who has a doctorate in law from Columbia University. It was Ambedkar, one of the principal authors of the Indian constitution, who late in his life concluded that the untouchables would not be equal citizens in India as long as they continued to be second-rate Hindus. He led a mass conversion to Buddhism, primarily by his own community of Mahars, but also followed by other untouchable groups in India. The conversion to Buddhism, however, has not solved the discrimination of these groups in India and more recently a unifying tactic has been to claim a more secular, militant identity in terms of race rather than religion.

In Indian Islam, one of the major developments in the twentieth century was the founding of the Tablighi Jama'at in 1927 by Maulana Muhammad Ilyas (1886–1944). In 1934, Ilyas started a great revivalist campaign among Muslims of Mewat, an area near Delhi. Going from person to person, he called upon them to follow their Muslim duties, such as praying in the mosque. Gradually he formed groups of laymen who went for fixed periods on tours to invite their fellow Muslims to perform their duties. This created a snowball effect whereby more and more people became involved in this activity. The groups went around first in the area of Mewat, then in other places in India and Pakistan, and, finally, in the 1950s abroad to countries such as Nigeria, Morocco, Britain, Canada, and France. The movement thus spread from South Asians to other Muslims and is now one of the most successful Islamic movements in the world. The Tablighi Jama'at is professedly apolitical and in that way very different from the major Islamist movement of the subcontinent, the Jama'at-i-Islami. This movement was founded in 1941 by a former journalist, Maulana Maududi (1903–79), one of the most influential Islamist thinkers of the twentieth century. It is a much smaller movement than the Tablighi Jama'at, but highly political and organized as a tightly and hierarchically run revolutionary force which aims to capture the state in order to transform society according to Islamic principles. In Pakistan, it has a considerable influence in politics and also in the rest of the Muslim world revolutionary movements of the Islamist persuasion have taken it as an inspiring example.

The 1980s saw the development of some ultranationalist religious movements in India. The Vishwa Hindu Parishad, founded in Bombay in 1964 by some religious leaders with the help of the RSS, organized a country-wide protest against

the conversion to Islam of a group of 'untouchables' in the South Indian village of Meenakshipuram in 1981. The success of this protest led to the staging of another All-India campaign in 1984 to demand the demolition of a sixteenth century mosque in Ayodhya, in North India, which had allegedly been built on the site of a temple commemorating the birthplace of the God Rama. This campaign became one of the most important issues in Indian politics in the decade following its start. In 1992, the mosque was demolished with the tacit acceptance of the Indian state.

The 1980s also saw the development of a radical Sikh separatism in the Punjab, led by Sant Jarnail Singh Bhindranwale (1947–84), a young religious leader. His militant career started with intersectarian strife among the Sikhs, but developed into straightforward confrontation with the Indian state. In 1984 the most important Sikh shrine, the Akal Takht in Amritsar, was captured by the militants who demanded Khalistan, a separate Sikh state, imitating the earlier demand for Pakistan by the Muslims of South Asia. In the same year the Indian army decided to attack the Golden temple and chase the militants. The action caused great disaffection with the state among Sikhs worldwide and led to the assassination of Mrs Indira Gandhi (1917–84), the then prime minister of India.

During the twentieth century, there was a growing nationalization of Indian religions, which makes it difficult to separate them from processes of state formation. The migration of South Asians outside India has led to these forms of religious nationalism being spread to the rest of the world.

Also the relation between India and Pakistan is an important element in the development of religious nationalism. Especially the contested status of Muslim-majority Kashmir leads to continuous violence. In 2008 the Laskar-e-Taiba, a militant group based in Pakistan, executed a well-coordinated attack on Mumbai, killing 164 people and drawing international attention. One of the most important challenges of India today is to improve Hindu–Muslim relations.

See also: Asia, Sociocultural Overviews: South Asia; Buddhism; Conflict: Anthropological Aspects; Marx, Karl (1818–83); Revolution, Theories of; Revolution; Revolutions, Sociology of.

Bibliography

Appadurai, A., 1981. Worship and Conflict under Colonial Rule: A South Indian Case. Cambridge University Press, Cambridge, UK.

Dirks, N.B., 1987. The Hollow Crown: Ethnohistory of an Indian Kingdom. Cambridge University Press, Cambridge, UK.

Dumont, L., 1970. Homo Hierarchicus: The Caste System and Its Implications. Weidenfeld and Nicolson, London.

Fuller, C.J., 1992. The Camphor Flame: Popular Hinduism and Society in India. Princeton University Press, Princeton, NJ.

Metcalf, B.D., 1993. Living Hadith in the Tablighi Jama'at. Journal of Asian Studies 52 (3), 584–608.

Thapar, R., 1997. Syndicated hinduism. In: Sonntheimer, G.D., Kulke, H. (Eds.), Hinduism Reconsidered. Manohar, New Delhi.

van der Veer, P., 1994. Religious Nationalism: Hindus and Muslims in India. University of California Press, Berkeley, CA.

Weber, M., 1967(1917). The Religion of India: The Sociology of Hinduism and Buddhism. Free Press, Glencoe, IL.

Religiosity

Darren E Sherkat, Southern Illinois University, Carbondale, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K.J. Christiano, volume 19, pp. 13115–13119, © 2001, Elsevier Ltd.

Abstract

Religiosity, in the social sciences, is a term used to differentiate and describe both individual and collective attachments to religion. Religiosity identifies connections to religious traditions and differentiates subgroups or denominations within religious traditions to describe the personal attachments of individuals or the collective commitments of groups or nations. The diversity or homogeneity of religiosity in nations or groups can foster both solidarity and conflict. Individuals' religious commitments are associated with a host of social scientific outcomes, including social mobility, ethnic solidarity, and political commitments.

Religiosity is a characteristic of individuals and collectivities that displays various features of beliefs about the supernatural and how individuals and social groups engage in behaviors related to the supernatural. For individuals, supernatural explanations about the meaning and purpose of life and the trajectory of individuals and groups are often comforting and valuable (Stark and Bainbridge, 1987; Stark and Finke, 2000), and groups with common beliefs can enjoy social solidarity that enables collective action – whether for producing collective goods through social capital or eliminating competition through repressing those whose religiosity conflicts with the dominant group (Sherkat, 2005).

Individuals are socialized as young children by families and religious groups to understand and believe supernatural explanations about the meaning and purpose of life (Sherkat, 2003). In modern societies, these explanations are proffered by and connected to formal religious organizations, and individuals grow to feel a sense of common identification with the religious groups. The process of instilling beliefs in supernatural explanations also involves formal collective behaviors. Individuals participate in religious organizations and ritual actions that help reinforce supernatural beliefs and explanations and generate a sense of individual identification and collective identity.

Collective identifications with religious organizations and traditions provide definition of the religiosity of social groups, for both areal units (nations, regions, municipalities, and the like) or subgroups in a population, particularly ethnic groups. Social scientists often describe nations or regions as, for example, 'Christian,' 'Buddhist,' 'Muslim,' or 'Sikh' as a discernment of the type of religion predominating. Levels of religious participation and commitment are also definitive of collective religiosity, and are often used to characterize the quality and vitality of religious institutions. The importance of participation for institutional vitality makes it a useful characteristic of the concept of religiosity applied by social scientists to religious traditions, religious organizations, or areal units. Levels of religiosity are consequential for the mobilization of resources by religious groups, and these resources are used not only for religious purposes such as constructing religious buildings and staffing religious institutions with professionals and support personnel, but also for broader collective actions such as public works and charity, political mobilization, and religiously motivated social control (Sherkat, 2005).

Religious Traditions and Denominations

Religious traditions are constellations of beliefs, practices, and institutions used to describe a common type of religiosity. Religious traditions are broad understandings of the supernatural produced in societies and practiced by groups and individuals. Traditions espouse distinctive explanations, and in contemporary religious traditions these are codified in texts considered sacred by individuals and institutions that adhere to and produce that form of religion. Contemporary religious traditions include Baha'i, Buddhism, Christianity, Confucianism, Druze, Hindu, Islam, Jain, Judaism, Mormon, Sikh, Sufism, Taoism, and Zoroastrianism. In the history of religious traditions, most distinctive supernatural explanations have been extirpated as many cultures were destroyed through conquest – indeed, even in fairly recent history, we have minimal understanding of what the religious traditions were in many societies such as the Indigenous populations of the Americas and Australia, the non-Christian religions in early Europe, or the traditions we now call Hindu. Many traditions are formed by adding to the sacred explanations of an existing tradition. Social scientists define these movements as cult movements or new religious movements (Stark and Bainbridge, 1985, 1987), and so Christianity, Islam, and Mormon sprang as new traditions from Judaism, and Buddhism was a distinctive strand of Hindu. Other new religious movements are syncretic, and combine explanations of the supernatural from two or more traditions (e.g., Sikhism melds together Islam and Hindu, and Baha'i and the Unification Church syncretize from many traditions). Of course, over time it no longer makes sense to identify new religious movements as cults, and so traditions like Mormon and Baha'i are independent traditions. However, this is often contested, as early Christians identified as Jews, most Mormons consider themselves Christians, and Sufis identify as Muslims.

Beyond these broad traditions are more idiosyncratic interpretations of religious explanations within traditions. In Western social sciences, these forms of religiosity are conventionally referred to as denominations of a religious tradition. For example,

even the relatively new Mormon tradition is denominated by at least four organizations: (1) the Church of Jesus Christ of Latter Day Saints, (2) the Reorganized Church of LDS (now called the Community of Christ), (3) the Church of Jesus Christ, and (4) the Fundamentalist LDS. Denominations are formed through differentiation and schism over interpretations of supernatural explanations and conflicts over the control and use of religious resources. Denominational diversity is often difficult to assess in many religious contexts because local religious organizations may operate independently to interpret religious explanations. This is particularly true in Islam, Judaism, Taoism, Confucianism, and other religious traditions where religious leaders are free to interpret religion as they wish, and where there is minimal hierarchical control over the type of religious explanations put forward at the congregational level. Religious diversity is often examined using denominations as the unit of religiosity being examined (Finke and Stark, 1992; Iannaccone, 1991), and religious conflict is very often spurred by competition for control of traditions among competing denominations (e.g., Catholic–Protestant in Christianity, Mahayana–Theravada in Buddhism, and Sunni–Shiite in Islam).

At both the aggregate and individual levels, social scientists often use religious traditions to describe the religious commitments that predominate in nations, regions, or smaller geographic units or are held by individuals. Generally, however, this is a matter of convenience caused by a dearth of data that would enable a more exacting portrait of denominational distributions and identifications. Denominational data simply do not exist for most nations and areas, and many studies of individual religious commitments are limited to identification with religious traditions as the only indicator of religiosity.

Religiosity as Religious Behavior: Participation and Devotion

Religious collective action is a definitive element of religiosity. Religiosity is very often linked to levels of religious behavior in areas or by individuals. Religious behaviors can be public, as in participating in religious rituals, making pilgrimages, or donating time and money to religious groups; or they can be private, such as prayer, meditation, and reading sacred texts. In many religious traditions, public religious participation is routinized in regular religious rituals performed at a temple. This is particularly true for faiths springing from the Abrahamic tradition, where weekly ritual participation is prescribed for devotees and supported by institutional resources. In other traditions, however, ritual activities are mostly confined to the priestly class and lay members are expected to participate only on special holy days and festivals (as in Hindu and Buddhism).

Religious traditions or denominations that emphasize more irregular collective activities for lay members tend to amplify the value of engaging in personal prayer and devotion, giving money to temples and religious authorities, making pilgrimages to holy sites, and prescribing household shrines to facilitate religiosity. Sociologists often classify these types of traditions as 'devotional' rather than 'congregational' in terms of their form of religiosity. Devotional forms of worship are effective for satisfying religious desires when religious institutions lack the facilities and staff to serve the needs of large populations or devotees in rural areas. Congregational forms of religious organization become more common in urban areas and when religious resources are plentiful.

Religious traditions and denominations also vary in how they approach outsiders, with some groups having a conversion-oriented approach amplifying proselytization of the faith to outsiders, while other groups advocate separation from those who do not adhere to the faith. Proselytization is a form of religious behavior, and some groups (particularly the Jehovah's Witnesses, Mormons, Soka Gakkai Buddhists, and Bhakti Hindu groups) incorporate proselytization into their prescribed religious behaviors.

Religious Beliefs, Identities, and Personal Religiosity

Religious beliefs are a cognitive component of religiosity, are definitive of distinctions between different religious traditions, and are often consequential for distinguishing denominations within traditions. Beliefs as a component of religiosity often reflect the degree to which people believe in supernatural explanations – such as creation myths, gods and spirits, beliefs about prophesy, supernatural rewards and punishments, and the accuracy or divine authority of sacred texts. Beliefs that adhere strictly to particular interpretations of sacred texts and that claim exclusive understandings of supernatural explanations are indicative of fundamentalist religiosity. Social scientists are often interested in how beliefs change over time, and in the secularization of beliefs about the supernatural.

Studies of religious beliefs are problematic because of the peculiarity of religious beliefs within traditions and the gulf between traditions. There are commonalities in beliefs across the Abrahamic tradition – Jews, Christians, Muslims, and Mormons share core sacred texts with common creation myths, gods, and prescriptions and proscriptions for religiously sanctioned behaviors. Yet, Christians almost uniformly reject beliefs in the propriety of religious rules regarding dress, animal slaughter, and dietary proscriptions.

Religiosity is often linked to politicized or collective identities. Here, specific religious movements link to a more collective identity such as 'born-again Christian,' 'fundamentalist,' 'Evangelical,' or 'ultra-Orthodox.' These identifications may be associated with institutional ties, but individuals can choose to embrace or reject the identification without disrupting connections to a religious group. Politicized identities are often fickle, and in the United States we have seen 'fundamentalist,' 'born again,' and now 'evangelical' identities tied to religiosity and politics. Many treatments examine religiosity in terms of its relative degree of orthodoxy or fundamentalism. Here, researchers attend to whether religious beliefs adhere to literal textual interpretations of sacred texts in the tradition – or at least to contemporary interpretations of some aspects of those texts.

Contemporary research on European and American Christianity has argued that Western Christians increasingly 'believe but don't belong' or are becoming 'spiritual but not religious' (Davie, 1990). While this thesis has been given considerable play by journalists and religious pundits, social scientific investigations have consistently shown that people who do not belong to religious organizations tend not to believe, and

that many who do belong do not believe core tenets of the faith (Voas and Crockett, 2005).

Because of variations in religious beliefs across and within traditions, most examinations of beliefs focus on common indicators, such as the belief in gods, self-perceptions as being a religious person, or the importance of religion in people's lives. Studies find considerable diversity regarding beliefs about gods and personal religiosity. In the Czech Republic only 16% believe in a personal god, while the figure is 19% in Sweden and 24% in Japan. In contrast, 68% in the United States believe in a personal god, while the number is 72% of Chileans and 92% of Filipinos (Smith, 2012).

Macrolevel Dimensions of Religiosity

Across geographic space and over time, religiosity varies, and social scientific theories and analyses typically address how and why it varies from place to place and time to time. Of most interest have been secularization theories, which posit that religious institutions will lose influence as societies modernize, and that consequently indicators of personal religiosity will also decrease – such as religious participation, identification with religious groups, and beliefs in gods and the veracity of sacred texts (Stark and Bainbridge, 1985, 1987). More recently, religious market theories have focused on the importance of state-level regulation for influencing religiosity, and on the importance of normative pressures to consume religion (Finke and Stark, 1992; Iannaccone, 1991; Sherkat and Wilson, 1995). Finally, religiosity is often intertwined with ethnicity and language, and ethnic nationalism and regionalism are often prime motivations for collective religiosity (Beyer, 2006).

Secularization theorists have long argued that, as societies became more differentiated and more modern, widespread education would undermine beliefs in supernatural explanations, thus leading to declines in religiosity (Berger, 1965). In Europe and Japan, this appears to have taken place, and religious hegemonies have been shattered. However, elsewhere in the world, religion remains a powerful institution and religiosity remains high. Some nations were deliberately secularized through state control and regulation of religious institutions, particularly in the former Soviet Bloc and China. However, the breakup of the Soviet Empire and the fall of socialist states led to decreased religious regulation and an increase in religiosity in those nations (Froese, 2008; Froese and Pfaff, 2001).

Supply-side religious market theorists have explained the rise and fall of religiosity by arguing that the state regulation of religious firms and state sponsorship of religion have decreased religiosity (Iannaccone, 1991; Stark and Finke, 2000). From this perspective, religiosity is motivated by having access to high-quality religious goods – religious explanations imbued with certainty instead of doubt, and enthusiastic religious participation rather than formulaic ritual. Supply-side theorists argue that when states constrict religious products or attempt to produce a homogeneous religion, it undermines the quality of religious goods and reduces overall rates of religiosity. At the individual level, supply-side market theorists argue that religiosity is a function of religious human capital, defined as an understanding of how religion is produced in particular settings. When individuals have high levels of religious human capital, they are able to collectively produce valued religious goods; however, if their religious capital is incompatible with existing religious resources (i.e., a Hindu in a Catholic setting), there will be minimal religious production and low levels of religiosity (Iannaccone, 1990).

Ethnicity and nationalisms also influence and are influenced by religiosity. Many ethnic groups across varied religious traditions embrace religious participation and the expression and promulgation of religious beliefs in order to sustain ethnic identity and motivate political salience (Beyer, 2006). In multiethnic nations, religiosity is often used as a marker of ethnic commitment, and in the presence of political persecution religiosity is often an oppositional activity motivated by ethnic commitments and political ties as much as by an individual or collective desire for supernatural explanations, rewards, or compensators (Froese, 2008; Froese and Pfaff, 2001).

Microlevel Processes and Religiosity

Most social scientific studies of religion focus on individual religiosity and the microlevel processes that inform the intergenerational transmission of religiosity, variations in religiosity over the life course, and the consequences of personal religiosity across a range of social behaviors and institutions. Life course processes are an important focus of research examining variations in religiosity, and the connections between religion and marriage, fertility, and career trajectories are the focus of a substantial body of research, particularly in the United States and Europe.

Religious participation and devotion are often linked to life course events such as marriage, childbirth, childrearing, and death. The formation of a family through procreation is sacralized through religious rituals of marriage, and religiously sanctioned marriage is an aspect and indicator of religiosity. Children are often dedicated to religious traditions at birth or in early childhood through ceremonies of baptism or ritual cleansing, and through training and confirmation of religious fealty. Life course transitions like school graduations, obtaining employment, and purchasing a home or a car are also often marked by religious rituals. Religious funerals are another indication of religiosity, and many religious traditions have particular prescriptions and proscriptions regarding the burial or disposal of dead bodies of the faithful.

The socialization of children often gives rise to higher levels of religious engagement, particularly in traditions where congregational styles of worship predominate. Longitudinal studies show that marriage and childrearing have a positive impact on religious participation, particularly when the age of marriage and the timing of children follow a normative pattern for these life course events (Stolzenberg et al., 1995; Sherkat, 2003). Religious identifications, beliefs, and patterns of participation are passed from parents to children, and the family of origin is the most important source of religious socialization (Sherkat, 2003; Sherkat and Wilson, 1995). Religious socialization also happens within religious institutions, and some sectarian religious groups demand higher levels of public religious expression and more exclusive religious beliefs.

Retirement and the departure of adult children from the household may also provide opportunities for religious participation and pilgrimage, and even to join religious

societies or train for the clergy. Aging also appears to be positively associated with the certainty of beliefs about gods (Sherkat, 2010); however, it is uncertain whether this is true outside of the United States.

Individual-level religiosity is also structured by ethnicity and gender. In many areas of the world, religion is a cultural marker of ethnic status, and religious participation is often highly valued or mandated in those groups (Ellison and Sherkat, 1995). In many religious traditions, women are substantially more active than men, particularly in Western Christianity (Sherkat, 2003), and there has been considerable debate in the literature about the origins of gender differences in religiosity.

Educational institutions and attainment processes also influence religiosity and religious beliefs. In the United States, higher levels of education are associated with higher rates of participation and less exclusivist and more secular religious beliefs (Sherkat, 2003). Religious schools play a role in securing particular religious commitments through childhood socialization and the homogenization of social networks. Individual or family religiosity will influence educational choices, and can lead to higher or lower levels of attainment depending on how religiosity is seen to conflict or correspond with educational goals (Sherkat, 2010). In the United States, sectarian Protestants consistently show educational disadvantages, while Jews and liberal Protestants have high levels of attainment (Lehrer, 2004; Fitzgerald and Glass, 2008). These educational trajectories also influence occupational attainment, income, and wealth, and are linked to patterns of early marriage, patriarchal gender norms, large family size, and anti-intellectualism (Keister, 2011; Darnell and Sherkat, 1997; Sherkat, 2010).

Religiosity also plays an important role in political engagement, structuring preferences for political goods and influencing ties to political organizations. Individuals' social networks are strongly influenced by religiosity, and the pattern and strength of these ties can be used as resources by political movements (Sherkat, 2005). Religious beliefs are increasingly important for informing beliefs about political goods associated with modern states – particularly the value of education; equality across ethnic, religious, and gender groups; and the sanctity of religious strictures and regulations on social life (blasphemy laws, regulation of food and drugs, clothing regulations, limitations on the depiction of sex, and the like) (Jurgensmeyer, 2008).

Future Directions in Religiosity Research

The study of religiosity has been enlivened by the continued importance of religiosity for modern political and social conflicts, and by the increasing diversity of religious expressions throughout the world. The globalization of religious diversity is evident even in heavily regulated religious environments like Saudi Arabia. Yet, the study of religiosity has not kept pace with this global diversification. Indeed, we lack adequate classification systems to describe religious diversity in many environments – we cannot distinguish diverse types of Muslims, Buddhists, or Hindus, for example. Religiosity has not been adequately examined from a social scientific perspective in most parts of the world, and most social scientific data collection efforts have given only perfunctory attention to how to measure religious commitments. The collection of data on religion remains controversial and potentially deadly in much of the world, and this hampers our understanding of how religiosity plays a role in social life in those areas. The problem is not solved even in well-studied environments like the United States. Many studies and data collection efforts give limited attention to religiosity and provide scant information on religious commitments and beliefs.

See also: Catholicism; Christianity in Asia; Christianity, Origins of; Globalization and Religion; Nationalism: General; Religion, Morality, and Social Control; Religion, Sociology of.

Bibliography

Berger, Peter, 1965. The Sacred Canopy. Beacon Press, Boston, MA.
Beyer, Peter, 2006. Religions in Global Society. Routledge, London.
Darnell, Alfred, Sherkat, Darren E., 1997. The impact of fundamentalism on educational attainment. American Sociological Review 62, 306–315.
Davie, Grace, 1990. Believing without belonging: is this the future of religion in Britain? Social Compass 37 (4), 455–469.
Ellison, Christopher G., Sherkat, Darren E., 1995. The semi-involuntary institution revisited: regional variations in church participation among black Americans. Social Forces 73, 1415–1437.
Finke, Roger, Stark, Rodney, 1992. The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy. Rutgers University Press, New Brunswick, NJ.
Fitzgerald, Scott T., Glass, Jennifer, 2008. Can early family formation explain the lower educational attainment of U.S. conservative Protestants? Sociological Spectrum 28, 556–557.
Froese, Paul, Pfaff, Steven, 2001. Replete and desolate markets: Poland, East Germany and the new religious paradigm. Social Forces 80 (2), 481–507.
Froese, Paul, 2008. The Plot to Kill God: Findings from the Soviet Experiment in Secularization. University of California Press, Berkeley.
Iannaccone, Laurence, 1990. Religious practice: a human capital approach. Journal for the Scientific Study of Religion 29 (3), 297–314.
Iannaccone, Laurence, 1991. The consequences of religious market structure: Adam Smith and the economics of religion. Rationality and Society 3 (2), 156–177.
Jurgensmeyer, Mark, 2008. Global Rebellion: Religious Challenges to the Secular State, from Christian Militia to Al Qaeda. University of California Press, Berkeley.
Keister, Lisa A, 2011. Faith and Money: How Religion Contributes to Wealth and Poverty. Cambridge University Press, New York.
Lehrer, Evelyn L, 2004. Religiosity as a determinant of educational attainment: the case of conservative protestant women in the United States. Review of Economics of the Household 2 (2), 203–219.
Sherkat, Darren E., Wilson, John, 1995. Preferences, constraints, and choices in religious markets: an examination of religious switching and apostasy. Social Forces 73, 993–1026.
Sherkat, Darren E, 2003. Religious socialization: agents of influence and influences of agency. In: Dillon, Michele (Ed.), Handbook for the Sociology of Religion. Cambridge University Press, pp. 151–163.
Sherkat, Darren E, 2005. Religion, social movements, and politics. In: Ebaugh, Helen Rose (Ed.), Handbook on Sociology of Religion and Social Institutions. Kluwer Press, pp. 1–20.
Sherkat, Darren E, 2010. The religious demography of the United States. In: Hummer, Robert, Ellison, Christopher G. (Eds.), Religion, Family, and Health. Rutgers University Press, pp. 403–430.
Smith, Tom W, 2012. Beliefs about God across Time and Countries. University of Chicago, NORC.
Stark, Rodney, Bainbridge, William Sims, 1985. The Future of Religion: Secularization, Revival, and Cult Formation. Univ. of California Press, Berkeley.
Stark, Rodney, Bainbridge, William Sims, 1987. A Theory of Religion. Rutgers University Press, New Brunswick, NJ.
Stark, Rodney, Finke, Roger, 2000. Acts of Faith. University of California Press, Berkeley.
Stolzenberg, Ross M., Blair-Loy, Mary, Waite, Linda J., 1995. Religious participation in early adulthood: age and family life cycle effects on church membership. American Sociological Review 60, 84–103.
Voas, David, Crockett, Alasdair, 2005. Religion in Britain: neither believing nor belonging. Sociology 39, 11–28.

Religiosity and Prejudice against Minorities

Peer Scheepers and Rob Eisinga, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The authors delineate two broad streams of research, one more psychological focusing on intrapersonal religious orientations and the other more sociological focusing on extrapersonal religious dimension, both concerned with the religion and prejudice against ethnic minorities. The authors discuss robust relationships that in the course of nearly 60 years of research have been discovered. They focus on and find dimensions of religiosity that 'make and unmake prejudice,' in line with the classic and paradoxical statement of Allport.

Introduction

Just after World War II, a crucial research question was developed and answered by Allport and Kramer (1946) on the relationship between religiosity and intolerance toward minorities. They found some evidence that the more religious people supported prejudice. This finding was considered surprising but also distressing, particularly in the wake of the war that had been accompanied by widespread prejudices against different minorities, both ethnic (particularly Jews) as well as sexual (particularly homosexual) minorities and had culminated in the genocide of so many innocent victims during the Holocaust. Ever since then, this socially relevant question has been on the research agenda of generations of researchers, particularly in the decades following the war, has and increasingly as patterns of immigration have brought diverse religious groups together.

Allport summarized numerous findings on this relationship in his contributions in the late 1950s and early 1960s, stating: "… there is something about religion that makes prejudice and something…that unmakes prejudice…" (1958,1966: p. 447). At regular intervals, overviews of the empirical relationship between religiosity and prejudice were published and several of these were frequently cited (Gorsuch and Aleshire, 1974; Chalfant and Peek, 1983; Batson et al., 1993). Five decades after the initial study, Hood et al. (1996: p. 359) again concluded that the massive evidence justified a broad generalization: "…the more religious an individual is, the more prejudiced that person is…"; however, they followed this generalization with many qualifications based on different dimensions of religiosity as well as by varied targets of prejudice.

After analyzing the numerous contributions, also including recent overviews, the authors distinguish two broad disciplinary traditions that have focused on these different dimensions of religiosity. In a more psychological tradition, researchers have focused predominantly on (intrapersonal) religious orientations, whereas in a more sociological tradition, researchers have focused predominantly on (extrapersonal) religious dimensions. The authors review and describe these theoretical considerations and empirical findings as related to intolerance toward ethnic (and sexual) minorities.

Psychological Studies

Religiosity in this tradition goes back to religious orientations, initially introduced by Allport and Ross (1967) as extrinsic versus intrinsic orientations. They found that those people with an intrinsic orientation, for whom religiosity served as a dominant master motive in their daily lives, were less prejudiced against minorities than those with an extrinsic orientation, for whom religiosity served as a means to an end. Intrinsically motivated religious people were considered to be more tolerant because they 'lived' or at least supported religious teachings on brotherhood of man and rejected contempt of other groups in society. The initial findings have been replicated numerous times, at least until the mid-1980s (Spilka et al., 1985). These religious orientations were additionally refined by Batson (1976) who introduced the quest orientation, an orientation to be ready to face existential questions and acknowledge religious doubts. People who were motivated by a quest orientation were considered to be more tolerant because they were ready to search for independence of (informal) religious teachings or shared collective religious norms on other ethnic groups. Thereafter, this tradition was enriched by theoretical notions on religious fundamentalism, developed by Altemeyer and Hunsberger (1992). They proposed that religious fundamentalism, an orientation reflecting an unquestioning and unwavering certainty in basic religious truths, would encourage fundamentalists to discriminate between those living according to these truths versus those who do not, and, moreover, to condemn and reject members of other religious or ethnic groups.

Recent meta-analytic overviews by Hunsberger and Jackson (2005) and Hall et al. (2010) showed highly consistent results: religious identification, religious fundamentalism, and extrinsic orientations show strong positive associations with ethnic intolerance, providing evidence to the above-mentioned theoretical propositions. On the other hand, they found that intrinsic and quest orientations showed negative associations, also in line with the above-mentioned theoretical expectations. In sum, these findings are evidence for the classic claim by Allport about religiosity inducing and reducing intolerance. Moreover, additional evidence showed that the relationship between religious fundamentalism and intolerance was explained by authoritarianism (Eisinga et al., 1999; Hall et al., 2010; Johnson et al., 2011).

In their meta-analysis of numerous studies over several decades, Hall et al. (2010: p. 131) also estimated the overtime variation in the association and actually found that "the positive relation between extrinsic religiosity and racism decreased significantly in recent years," and this also held for the relationship between religious fundamentalism and racism (2010: p. 132). A serious drawback of these findings is that they have been based predominantly on small, nonnational, and selective samples, and many of the studies are of college students, and this truncated factors such as educational attainment and other determinants that have been shown to be strongly related to prejudice, and this reduces their generalizability (Scheepers et al., 2002).

Sociological Studies

Religiosity in sociological studies goes back to dimensions of religiosity (Glock and Stark, 1965; Stark and Glock, 1968): practice, belief, experience, knowledge, and consequences. Most dimensions have been covered in individual publications, although overviews and syntheses are available in some studies (Gorsuch and Aleshire, 1974; Scheepers et al., 2002). In the sociological literature, there is substantial evidence based on large representative national samples that some dimensions of religiosity increase prejudice while other dimensions decrease prejudice, although some scholars find this evidence less conclusive (Ceobanu and Escandell, 2010).

Religious practice relates to denominational membership and church attendance. People considering themselves to be belonging to Christian denominations were rather consistently found to hold prejudices more than nonmembers and members of non-Christian denominations (Scheepers et al., 2002). Lasting controversies on the relationship between church attendance and prejudice have developed, and some contend that the relationship is curvilinear rather than linear. Gorsuch and Aleshire (1974) developed the argument that moderately frequent church attendees were likely to conform to a dominant church culture in which strong sentiments about white supremacy prevail that would induce prejudice; whereas, conversely, nonattendees and frequent attendees were more likely to dissociate themselves from such sentiments that would reduce prejudice. Hood et al. (1996) considered the evidence for the curvilinear relationship not very convincing, and Scheepers et al. (2002) found strong evidence for the linear rather than for the curvilinear nature of the relationship. Before the turn of the century, the more frequent people in European countries attended church meetings, the stronger they supported prejudice. Yet, after the turn of the century, the evidence for a linear relationship seems to have become nonsignificant in many Central and Eastern European countries (Strabac and Listhaug, 2008; Coenders et al., 2009) or even seems to have reversed for anti-immigrant attitudes in general (Bohman and Hjerm, 2013) as well as for anti-Muslim prejudice (Savelkoul et al., 2011).

Regarding religious or doctrinal beliefs, researchers have failed to develop convincing theoretical arguments as to why these beliefs would be related to prejudice: why would substantial beliefs in God be related to prejudice? However, there is clear evidence from many European countries, before and after the turn of the century, that religious doctrinal beliefs reduce prejudice (Scheepers et al., 2002; Bohman and Hjerm, 2013). Yet, there have been overriding arguments as to why another dimension of doctrinal beliefs, i.e., religious particularism, the belief that there is only one true religion, could be related to prejudice (Stark and Glock, 1968). Religious particularism predisposes people to strong religiocentric stances, and induces unfavorable attitudes toward other religions, other religious out-groups, and toward ethnic out-groups that are majority non-Christian. There is strong empirical evidence for this crucial proposition in many European countries: the stronger people support religious particularism, the stronger their prejudice against ethnic minorities (Scheepers et al., 2002).

There are relatively few studies examining the effect of religious experiences and knowledge on prejudice (Hood et al., 1996). However, one recent study using data from several European countries showed that "people with experiences of a spiritual, supernatural nature, may put inner-worldly tensions between in-groups and out-groups into a perspective of mutual tolerance…which would reduce prejudice" (Scheepers et al., 2002: pp. 246–247), However, no studies have examined the impact of religious knowledge on prejudice.

Studies have also related the concept of religious saliency to prejudice (Stark and Glock, 1968), and this sociological concept is strongly related to the psychological dimension of intrinsic religiosity (Allport and Ross, 1967). Those people for whom religiosity serves as a dominant master motive in their daily lives, who take the consequences of their religiosity seriously into account in their daily lives and in their everyday decisions, may be expected to live by one of the commandments to 'love thy neighbor' and consequently be less prejudiced to out-groups. There is strong evidence for this proposition for European countries before the turn of the century (Scheepers et al., 2002).

Just like in the psychological literature, there is also evidence that some relationships between religiosity and intolerance (toward Jews and Muslims) can be explained by authoritarianism (Konig et al., 2000) and, moreover by perceived threat, another intermediary determinant derived from crucial propositions in ethnic competition theories (Savelkoul et al., 2011). And just like in the psychological literature, there are indications that the relationships between religiosity and prejudice have decreased over time. Building on Hall et al. (2010), Bohman and Hjerm report that: "…the strength of the correlation between church attendance and belief and anti-immigration attitudes declines somewhat between the ESS rounds, even though the change is not so dramatic. So, perhaps we have already come to the situation in Europe today, especially in some countries, where we are witnessing a shift in the meaning of being religious…" (2013: p. 12).

Synthesis

Surveying and analyzing the conceptual distinctions and empirical findings of both streams or traditions of research, one tends to conclude that the overall picture has become quite clear as to which dimensions of religion "…make and unmake prejudice…" to rephrase Allport's (1958) paradox. Religious

fundamentalism (reflecting an unquestioning certainty in basic religious truths) focused on in the psychological tradition, is conceptually and operationally akin to the sociological dimension of religious particularism (reflecting the belief that there is only one true religion). Both may tend to encourage (some) religious people to categorize between those living according to these religious truths versus those who do not. There is compelling evidence that those who subscribe to religious fundamentalism, and to particularism, tend to be prejudiced against ethnic minorities. This has been shown to hold for prejudice against ethnic minorities and may also hold for prejudice against religious minorities like Muslims as well as for other (sexual) minorities, who overall deviate more or less strongly from the commandments of their own true religion (van den Akker et al., 2013). From a theoretical perspective, this mere categorization may be related to processes and mechanisms of positive group distinctiveness described in social identity theory (Tajfel, 1982).

Intrinsic religiosity (dominant in religious people for whom religiosity serves as a dominant master motive in their daily lives) is conceptually and operationally akin to the sociological concept of religious saliency (dominant in religious people who take the consequences of their religiosity seriously into account in their daily lives). And compelling evidence shows that these intrinsic orientations seem to promote prosocial orientations and decrease prejudice toward ethnic out-groups. However, there has been no investigation into how intrinsic religiosity or salience might influence prejudice against other groups such as religious and sexual minorities.

From a longitudinal perspective, it seems to have become all the more urgent to focus on changing relationships between dimensions of religiosity, more particularly extrinsic religiosity respectively church attendance, and prejudice against ethnic minorities and migrants. Both psychological and sociological research provides some evidence that these relationships seems to have diminished; however, it is not clear from the empirical evidence what kind of mechanisms are at work. The relationship between church attendance and prejudice may diminish due to different processes: frequent church attendees have over time decreased their prejudices and have become more like non-attendees who have been shown to be less prejudiced or vice versa, nonattendees have over time increased their prejudices and have become more like frequent attendees who have been shown to be more prejudiced. Considering the amount of cross-national, high-quality empirical data available for secondary analyses, these contrasting hypotheses can be tested rigorously in many different countries (http://www.europeansocialsurvey.org/). These data also provide the scientific community with possibilities to test the differential relationships between contextual level religiosity, individual level religiosity, and exclusionist stances toward minorities in their countries. Since many of these European societies have become more diverse in terms of ethnicities, not so many researchers have realized that these countries have simultaneously become more diverse in terms of religiosity: many immigrants consider themselves to belong to other than Christian denominations. In terms of theories on competition on religious markets (Stark and Bainbridge, 1996) as well as in terms of ethnic competition theories (Savelkoul et al., 2011), these demographic changes in the religious landscape need to be urgently tested.

See also: Globalization and Religion; Religion, Sociology of; Religion: Mobilization and Power; Religion: Morality and Social Control; Religion: Nationalism and Identity; Religiosity; Secularization.

Bibliography

Allport, G., 1958. The Nature of Prejudice. Doubleday and Company, New York.

Allport, G., 1966. The religious context of prejudice. Journal for the Scientific Study of Religion 5, 447–457.

Allport, G., Kramer, B., 1946. Some roots of prejudice. Journal of Psychology 22, 9–39.

Allport, G., Ross, M., 1967. Personal religious orientation and prejudice. Journal of Personality and Social Psychology 5, 432–443.

Altemeyer, B., Hunsberger, B., 1992. Authoritarianism, religious fundamentalism, quest, and prejudice. International Journal for the Psychology of Religion 2, 113–133.

Batson, C.D., 1976. Religion as prosocial: agent or double agent. Journal for the Scientific Study of Religion 15, 29–45.

Batson, C.D., Schoenrade, P., Ross, M., Ventis, L., 1993. Religion and the Individual. A Social Psychological Perspective. Oxford University Press, New York/Oxford.

Bohman, A., Hjerm, M., 2013. How the religious context affects the relationship between religiosity and attitudes towards immigration. Ethnic and Racial Studies. http://dx.doi.org/10.1080/0149870.2012.748210.

Ceobanu, A., Escandell, X., 2010. Comparative analyses of public attitudes toward immigrants and immigration using multinational survey data. A review of theories and research. Annual Review of Sociology 36, 309–328.

Chalfant, H., Peek, C.W., 1983. Religious affiliation, religiosity and racial prejudice. A new look at old relationships. Review of Religious Research 25, 155–161.

Coenders, M., Lubbers, M., Scheepers, P., 2009. Opposition to civil rights for racial minorities in Central and Eastern Europe. East European Politics and Societies 23 (2), 146–164.

Eisinga, R., Billiet, J., Felling, A., 1999. Christian religion and ethnic prejudice in cross-national perspective. A comparative analysis of the Netherlands and Flanders (Belgium). International Journal of Comparative Sociology 40, 375–393.

Glock, C.Y., Stark, R., 1965. Religion and Society in Tension. Rand McNally and Company, Chicago.

Gorsuch, R., Aleshire, D., 1974. Christian faith and ethnic prejudice. A review and interpretation of research. Journal for the Scientific Study of Religion 13, 281–307.

Hall, D., Matz, D., Wood, W., 2010. Why don't we practice what we preach? A meta-analytic review of religious racism. Personality and Social Psychology Review 14, 126–139.

Hood, R., Spilka, B., Hunsberger, B., Gorsuch, R., 1996. The Psychology of Religion, an Empirical Approach, second ed. The Guilford Press, New York/London.

Hunsberger, B., Jackson, M., 2005. Religion, meaning and prejudice. Journal of Social Issues 61, 807–826.

Johnson, M., Rowatt, W., Barnad-Brak, L., Patock-Peckham, J., Labouff, J., Carlisle, R., 2011. A meditational analysis of the role of right-wing authoritarianism and religious fundamentalism in the religiosity-prejudice link. Personality and Individual Differences 50, 851–856.

Konig, R., Eisinga, R., Scheepers, P., 2000. Explaining the relationship between Christian religion and anti-Semitism in the Netherlands. Review of Religious Research 41, 373–393.

Savelkoul, M., Scheepers, P., Tolsma, J., Hagendoorn, L., 2011. Anti-Muslim attitudes in the Netherlands. Tests of contradictory hypotheses derived from ethnic competition theory and intergroup contact theory. European Sociological Review 27, 741–758.

Scheepers, P., Gijsberts, M., Hello, E., 2002. Religiosity and prejudice in Europe. Test of a controversial relationship. Review of Religious Research 43, 242–265.

Spilka, B., Kojetin, B., McIntosh, D., 1985. Forms and measures of personal faith. Questions, correlates and distinctions. Journal for the Scientific Study of Religion 24, 437–442.

Stark, R., Bainbridge, W.S., 1996. A Theory of Religion. Rutgers University Press, New Brunswick, NJ.

Stark, R., Glock, C.Y., 1968. American Piety, Patterns of Religious Commitment. University of California Press, Berkely/Los Angeles.

Strabac, Z., Listhaug, O., 2008. Anti-Muslim prejudice in Europe. A multi-level analysis of survey data from 30 countries. Social Science Research 37, 268–286.

Tajfel, H., 1982. Social psychology of intergroup relations. Annual Review of Psychology 33, 1–39.

van den Akker, H., van der Ploeg, R., Scheepers, P., 2013. Disapproval of homosexual lifestyles in Europe 2002–2008. International Journal of Public Opinion Research 25 (1), 64–86.

Religious Belief, Evolution of

Armin W Geertz, Aarhus University, Aarhus, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

With the rise of the comparative sciences in the middle of the nineteenth century, major evolutionary schemes were posited. During the first decades of the twentieth century, these theories were rejected in anthropology, sociology, and the historical sciences. Evolutionary approaches, however, continued in isolation, but grew stronger toward the 1970s. In the historical and social sciences, evolutionary theories gained importance in relation to the Axial Age. The cognitive science of religion, founded in 1990 by scholars of religion, developed hypotheses conceived within evolutionary frameworks. These hypotheses are being tested in the laboratory, in the field, in simulation models, and in the development of historical databases.

Introduction

Religion has generally posed a problem for evolutionary theories. What advantages could beliefs and behaviors that are based on postulated supernatural agents, often with dangerous side-effects and practices, possibly have? What selective advantages have secured the survival of religious behavior ever since the appearance of *Homo sapiens*, behavior that was most likely already in play even before the species arose? Much of the debate on the evolution of religion has been based on what professional scholars of religion critically refer to as the '*sui generis* assumption,' in other words, the assumption that religious behavior is unitive and set apart from all other behaviors, thus requiring special feelings, experiences, brain areas, acts, and, subsequently, special methods for studying it. Contemporary scholarship, on the other hand, understands religion to be a cultural and social phenomenon that can be studied and explained using the same methodologies as in the study of any other social behaviors and institutions. It makes sense, therefore, to study religion within an evolutionary framework.

Religion is understood here as "a cultural system and a social institution that governs and promotes ideal interpretations of existence and ideal praxis with reference to postulated transempirical powers or beings" (Geertz, 1999, p. 471). Religion consists, therefore, of a world of ideas (and/or beliefs) about the natural and supernatural world and about humans and other creatures and beings that inform social behavior and social institutions. It is a powerful instrument that must have had selective advantages. The British scholar of religion, Ninian Smart, has sketched out seven dimensions of religion that may be useful here. Religion, he pointed out, consists of the practical and ritual, the experiential and emotional, the narrative or mythic, the doctrinal and philosophical, the ethical and legal, the social and institutional, and the material dimensions (Smart, 1989, p. 12–21). Religion is obviously a complex phenomenon, and, furthermore, its many dimensions are more or less identical to other cultural institutions as well as secular worldviews (such as nationalism, communism, humanism, and atheism). Religions differ around the world so that some of these dimensions might be missing or downplayed and others overemphasized. There is quite often confusion about religion, not the least, the confusion between religion in the real world and religion as a scientific concept. Religion is as difficult to capture analytically as culture is. And yet, we know that they have causal effects on genes, individuals, societies, and the material and natural world.

Because of the complexity of religious behavior, ideas, and institutions, whole disciplines have arisen dedicated to the scientific study of religion each with their own methods, theories, and perspectives. This has led to a division of labor with the unfortunate side-effect of disciplinary isolation. The professional disciplines dedicated to the science of religion are history of religions (often called comparative religion), the sociology of religion, the psychology of religion, and, quite recently, the cognitive science of religion. Theology and the philosophy of religion have mostly pursued their own a priori interests in theology and belief. Attempts have been and are being made to liberate the philosophy of religion from its (mostly Christian) theological framework. The term "religious studies" is used in the United States and England as a cover term for all of the above.

Early Theories

The nineteenth century was rife with theories about the evolution of religion. For both theological and ideological reasons, most thinkers distinguished between the religions of contemporary indigenous tribes living in the colonies and frontiers of European empires and the religions of 'advanced civilizations,' such as and especially Christianity in its official forms. Darwin followed this practice as well, carefully distinguishing between the social and moral faculties of the "ape-like progenitors of man" and contemporary savage tribes on the one hand and those of "civilized nations" on the other (Darwin, 1874, p. 147–165). Darwin had no doubt that social and moral faculties had selective advantage. But his ideas on religion were clearly based on the discussions and debates among leading intellectuals of the new comparative sciences, such as comparative religion, psychology, sociology, and history, living in the cosmopolitan centers of the empires. Their many schemes and ideas about the evolution of religion were not based on carefully collected data that biologists and geologists were accustomed to. There was no undisputed,

accumulated evidence waiting for a theory that would "transform the obvious" (Wilson, 2008, p. 23). The information they had on religion was being produced in the wilds of the empires, based on interactions often of a violent nature and misunderstandings between uninformed military personnel, colonists, traders, and more or less informed missionaries, on the one hand, and the hapless, highly resistant indigenes on the other (Chidester, 1996). It was not until the introduction of ethnographic fieldwork at the beginning of the twentieth century that the science of culture could begin. Thus, the many evolutionary schemes of the nineteenth century were speculations based on faulty data. Such schemes as those based on comparative religion scholar Friedrich Max Müller's idea that religion arose through enactments of natural phenomena and mythology as a disease of language, comparative religion scholar Sir James George Frazer's theory on the evolution of magic, religion, and science, anthropologist Edward Tylor's theory of the origin of primitive religion in animism, classicist Theodor Mommsen's stages of Roman religious development, sociologist Herbert Spencer's elaborate social evolutionary schemes, beginning with an error of thought, followed by ancestor worship, polytheism and ending with monotheism, sociologist Émile Durkheim's theory of the origin of primitive religion in totemic ritual practices, psychologist Wilhelm Wundt's scheme from errors in causal thinking to various forms of religion, psychologist Sigmund Freud's patricide and the Oedipus complex as the source of religion, jurist Johann Jacob Bachhofen's theory of the primordial matriarchy and its mother goddess worship as the source of religion, and so on, were all more or less rejected during the first decades of the twentieth century. This does not mean, however, that some of their ideas are not useful today.

Theories of the Twentieth Century: Anthropology

Throughout the twentieth century, most scholars in the humanities explicitly rejected evolutionary theory for a variety of reasons. As the accompanying chart indicates, studies on the evolution of religion have only recently (and dramatically) begun to appear (**Figure 1**). Scholars of religion rejected evolutionary theory because of the above-mentioned lack of trustworthy data and also because they were mostly interested in religious texts rather than religious behavior. Most ethnographers under the influence of Franz Boas, Bronislaw Malinowski, and A.R. Radcliffe-Brown had also rejected the evolutionary framework and concentrated their research on intensive, fieldwork-based cultural and social studies of particular groups of people. The common stance for ethnographers and scholars of religion was antitheoretical and dominated by historical and ethnographic particularism.

There were, however, a small group of anthropologists who maintained an evolutionary perspective. Robert L. Carneiro (2003) has provided a useful overview of evolutionism in cultural anthropology. The early stages of the reemergence of evolutionism in the twentieth century were represented especially by Leslie A. White. He argues that the errors of nineteenth-century thinking could not justify the rejection of the concept of cultural evolution. He and other colleagues like him, for instance Alexander Lesser, were a decided minority. After an impassioned talk on the evolutionary approach at the 1939 meeting of the American Anthropological Association before an audience that included Boas himself, White reports that the chair of the session, Ralph Linton, said to him that he would be given "the courtesy extended to suspected horse thieves and shady gamblers in the days of the Wild West,

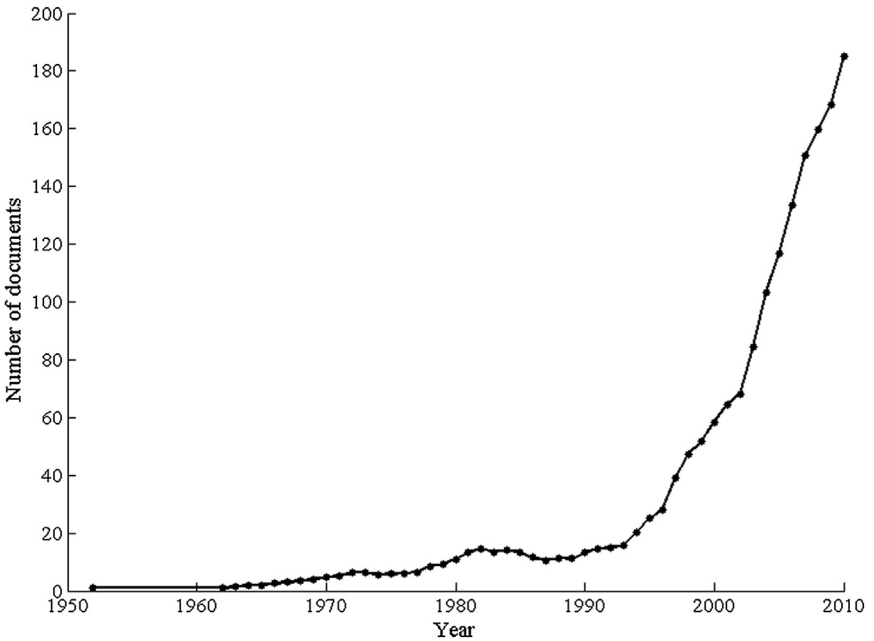

Figure 1 Graph showing number of articles and books in the Scopus database that contains the terms 'religion' and 'evolution' in title, abstract, or keywords.

namely, to allow them to get out of town before sundown" (White, 1987, p. 14; Carneiro, 2003, p. 101).

Support from other ethnologists who also took cultural evolution seriously grew slowly. These include Julian H. Steward's work on the Great Basin tribes and his editorship of *Handbook of South American Indians* (Washington, 1940–47) which was influenced by the evolutionary paradigms of archaeologists working in Mesoamerica and South America. The British archaeologist V. Gordon Childe with his studies of European prehistory was also crucial in reintroducing cultural evolutionism.

From the 1950s, evolutionary theory began to make an impact on cultural anthropology. Such scholars as Robert Redfield, Kalervo Oberg, Marshall Sahlins, Irving Goldman, and Walter Goldschmidt were all promoting evolutionary approaches in their areas of study. Carneiro notes that 1959, the centennial year of *On the Origin of Species*, was a pivotal point in the return of evolutionary theory in cultural anthropology. He mentions White's book *The Evolution of Culture* (1959) and the Anthropological Society of Washington's collected essays *Evolution in Anthropology: A Centennial Appraisal* (1959). He did not, however, note an equally important collection of essays edited by J.N. Spuhler entitled *The Evolution of Man's Capacity for Culture* (1959). In this latter collection of essays by archaeologists and anthropologists is found an explicit connection between prehistorical archaeology, physical anthropology, and the origins and evolution of culture. That collection furthermore played a profound role in the founding father of symbolic anthropology Clifford Geertz's understanding of culture, evolution, and human psychology.

Not many evolutionary anthropologists were interested in religion. Notable exceptions were neo-Marxists Michael Kearney and Marvin Harris. Kearney's analysis of worldview conceptions among Californian Indians and Mexican peasants is placed within a historical materialist framework (Kearney, 1984). Harris' strident ideology-critical study of the history of anthropological theory is a classic (Harris, 1969), but his smaller collections of essays on such topics as cargo cults, potlatch, messiahs, witch-crazes, and the origin of sacred cows are masterful analyses of religious behavior and ideas within the framework of technological, environmental, and economic development (Harris, 1975, 1977).

The late 1980s and early 1990s witnessed a few signal anthropological studies of religion and evolution. Two warrant special mention. The first, Roy A. Rappaport's *Ritual and Religion in the Making of Humanity* (1999), had a significant impact on a range of disciplines interested in religion. Published posthumously, his complex study argues that religion and especially ritual provided adaptive advantages to humankind and were coextensive with language and culture. For him, the defining characteristic of humanity is that it lives in terms of meanings "it must construct in a world devoid of intrinsic meaning but subject to physical law" (1999, p. 1) and thus he claims that without religion, humanity could never have emerged from its pre- or proto-human condition. Group selection, according to Rappaport, was essential to human evolution. Human systems, he wrote, are "cultural-organic systems constituted by symbolic (linguistic) as well as genetic information" and the "transformation from organic to cultural-organic must have been strongly selected for" (1999, p. 10). The role of religion in nature consists of the relationship between Ultimate Sacred Postulates and the conventional world. He likens this relationship to those of thermodynamics to the physical world (1999, p. 407). He argues that sanctity is significant "both among the adaptive *properties* of the species as a whole and in the adaptive *processes* of the social units into which the species is organized" (1999, p. 414). Sanctity is "the functional replacement for genetic determination of patterns of behavior" (1999, p. 418) because it "forms an unchanging ground upon which all else in adaptive social structures can change continuously without loss of orderliness" (1999, p. 427). In this sense, religions are cybernetic systems.

Rappaport's impact on the humanities consisted more of his analysis concerning the nature of religion and the logical, social, and subjective effects of ritual (which concerns the bulk of the book) rather than the adaptive role that religion played in the emergence of humanity. Even at this late date, the majority of researchers in the humanities and some social scientists were still "evolution-blind" mostly for ideological reasons. The anthropological-sociological team of Jonathan H. Turner and Alexandra Maryanski spelled out more specifically in their book *On the Origin of Societies by Natural Selection* (2008) what social evolution consists of. Their goal is to make the relationship between "human biology and sociocultural arrangements more explicit and systematic by exploring the ways that biology, social structure, and culture interact" (2008, p. 2). They posit five forces of the social universe that generate selection pressures: population, production, distribution, regulation, and reproduction (2008, p. 125). These forces are systematically analyzed in terms of sociocultural systems: hunter-gatherer, horticultural, agrarian, industrial, and postindustrial societies. Emphasis is placed on the ecology, demography, economy, kinship system, religion, polity systems, social institutions, and stratification systems of each social type. The result is a finely detailed analysis of the role that religion and ritual has played in interplay with the other selective forces. Thus, in this very hands-on approach, Turner and Maryanski have produced a highly readable application of evolutionary theory for cultural scientists. Their argument is clear: "we … need to be attuned to the fact that human biology not only makes human culture and social structure possible …, but the viability of culture and social structure over the long term is affected by the biology of humans as an evolved primate" (2008, p. 2).

Theories of the Twentieth Century: Sociology

Sociologists, with their interest in social structures and institutions, have a strong evolutionary tradition. All of the nineteenth and early twentieth century sociological thinkers assumed evolutionary schemata, such as August Comte's law of history, Herbert Spencer's detailed evolutionary scheme, Lewis H. Morgan's technology-driven evolutionary eras, the historical materialism of Karl Marx and Friedrich Engels, Max Weber's individual drive for domination and control, and Émile Durkheim's theory of the mechanical and organic solidarity and the division of labor. Sociologists also experienced an antievolutionary movement, and according to Pierre L. van den Berghe still were in 1990: "Sociological resistance to biological

thinking is in large part *trained* incompetence, not simply garden-variety anthropocentrism" (Berghe, 1990, p. 177).

Talcott Parsons' work on social evolution came to have an influence on mid-twentieth century and later sociology. Talcott Parsons completed his PhD and Dr Phil in Heidelberg and was therefore exposed not only to Karl Jaspers but also to Alfred Weber (Max Weber's brother), Edgar Salin, Emil Lederer, Karl Mannheim, and Ernst D. Troeltsch. Especially with the founding of the Department of Social Relations at Harvard, which was an institutional attempt to bring psychologists, sociologists, and anthropologists together, Parsons had a profound influence on many scholars such as Robert K. Merton, Clifford Geertz, Harold Garfinkel, and Robert N. Bellah. In developing his systems theory, Parsons was influenced by Alfred Emerson's idea that the symbol was the functional equivalent of the gene. Parsons envisioned three stages of evolution: the primitive, archaic, and modern.

Of the sociologists who have made an impact on the study of religion within an evolutionary perspective, Robert N. Bellah warrants special mention. He and Parsons had known each other for many years and had regularly corresponded. Already in 1964, Bellah introduced his theory of social evolution in the *American Sociological Review* which was largely ignored. It was based on ideas presented during a seminar on social evolution which Bellah gave together with Talcott Parsons and S.N. Eisenstadt at Harvard University in the spring of 1963 (Bellah, 1964, p. 358). In an earlier essay on Durkheim and history, Bellah emphasized, "I agree with Durkheim that the problem of evolution, including even social origins, is central for sociology as a science. To be convincing, this view must be backed by research – a challenge not to be evaded" (1959, p. 460 n.62).

The prime exponent of the analytical turn in comparative religion, Jonathan Z. Smith, remarked on Bellah's article in one of his publications that it was "irresponsibly crude" (Smith, 1987, p. 52). Smith was not against evolutionary theory, but he was concerned that the article did not specify mechanisms or carriers of evolution. Fortunately, Bellah took on the topic once again, and partially met the challenge raised in his 1959 article by publishing shortly before his death in 2013 a monumental study entitled *Religion in Human Evolution: From the Paleolithic to the Axial Age* (2011). Smith subsequently acknowledged Bellah's achievement at a panel during the American Academy of Religion 2011 annual meeting. In an unpublished but widely circulated, handwritten copy of his paper, Smith retracted his previous criticisms and noted that "the burdens of mechanism, agency, bearer et al. now fall to the biological. It is the genus *Homo* and not generic 'religion' that is the subject of evolution" (Smith, 2011, p. 4).

Smith was not entirely correct in his initial characterization of Bellah's earlier paper. Having been influenced by Parsons' systems theory, the mechanisms undergoing evolution according to Bellah are systems, whether human, animal, bacterial, or any other systems, persisting through time and adapting to change. The abstract to his article explains his understanding of evolution:

> Evolution in the sphere of religion is traced on three levels. First and most central is the evolution of religious symbol systems which are described as moving from "compact" to "differentiated." In close conjunction with this evolution, religious collectivities become more differentiated from other social structures and there is an increasing consciousness of the self as a religious subject. Five ideal typical stages of development are posited but it is recognized that these stages are not inevitable, that there is a wide variety of types within each stage, and that actual cases present many important features which cannot be neatly characterized in terms of any one stage. The close connection between religious evolution and other aspects of socio-cultural evolution is assumed but not explored.
>
> Bellah, 1964, p. 358

Thus, Bellah's interest is in the evolution of symbol systems. In his opus, that interest has not changed. His vision, however, did. Smith was correct in that evolution for Bellah "is part of the human story all the way through, though quite a long time ago it gave rise to culture and has coevolved with it ever since" (Bellah, 2011, p. xi). Bellah's story literally begins with the Big Bang (2011, p. 50ff.), but his abiding interest is still in symbol systems and in the way that religion shapes the social order and everyday life. Bellah draws on psychologist Merlin Donald's three evolutionary stages of human culture: mimetic, mythic, and theoretic that occurred during the past 2 million years (Donald, 1991, 2001). All of the phases presuppose the subsequent phases and are still present in those phases: Bellah's maxim here is "Nothing is ever lost" (Bellah, 2005, p. 72). The phases of mimetic and mythic are evident in tribal religions. As societies became more complex, chiefdoms and kingships arose in the archaic phase with hierarchical social classes supported by myth and ritual pageantry. During the first millennium BCE, theoretic culture appears giving rise to the critical questioning of the old myths and sterile ritualism of earlier forms. These cultures and their religions, Buddhism, Ancient Greece, China, Judaism, and India, during the so-called axial age, promote an ethical and spiritual universalism (2011, p. xix). Bellah's book consists of an erudite fleshing out of these three phases: Tribal, Archaic, and Axial. As he notes in the introduction, his book ends where the last 2000 years begin. So the last two of the original five phases that he sketched out in his 1964 article were left untouched: Primitive, Archaic, Historic, Early Modern, and Modern. But others in the history of religions have picked up where Bellah left off.

Theories of the Twentieth Century: Comparative Religion

As mentioned, most scholars in the humanities, including historians of religion, explicitly rejected evolutionary theory. The common stance in comparative religion was historical and ethnographic particularism. For decades, there was a struggle between particularism and the broader more theoretical need for comparativism. Evolutionary theory, however, had no place in this struggle. As Robert Bellah opined in the above-mentioned 1964 article:

> … by the third decade of the 20th century the evolutionary wave was in full retreat both in the general field of science of religion and in the sociology of religion in particular. Of course, this was only one aspect of the general retreat of evolutionary thought in social science,

> but nowhere did the retreat go further nor the intensity of the opposition to evolution go deeper than in the field of religion.
>
> Bellah, 1964, p. 358

There were, however, a few scholars who were not anti-evolutionists. One warrants mention here, namely the German Egyptologist Jan Assmann.

The term 'axial age' ('Achsenzeit,' i.e., 'axial period') is widely held to have been introduced by the German existentialist philosopher Karl Jaspers in his book *Vom Ursprung und Ziel der Geschichte* ["On the Origin and Goal of History"] (1949). Hans Joas, however, has pointed out that Georg Simmel in his *Lebensanschauung* (1918) used the term axis before Jaspers, although Jaspers does not refer to him. Jaspers developed his ideas in dialogue with and against Hegel, but was most likely influenced by Max Weber, his brother Alfred Weber, and their colleagues Ernst Troeltsch and Wilhelm Bousset. They were in turn influenced by the historian of religions Hermann Siebeck and indologist T.W. Rhys Davids, so, Joas notes, "one could already speak of an 'Axial Age debate' around 1900" (Joas, 2012, p. 11). But they were not the first ones. Jaspers mentions sinologist Victor von Strauss and classicist and philosopher Ernst von Lasaulx. Lasaulx, on the other hand, mentioned Abraham-Hyacinth Anquetil-Duperron (1731–1805) as a forerunner. The founding father of Iranian studies, Anquetil-Duperron claimed in 1771 that Zoroaster's religious ideas in the sixth century BCE were a part of "a more general 'revolution' in different parts of the world" (Joas, 2012, p. 11).

The term was meant to cover the major religious and philosophical movements in the first millennium BCE throughout Eurasia. Jaspers did not look for the material conditions behind these independent movements, but posited a spiritual change that was independent of social and material conditions. This claim was closely bound to his ethical, political, and philosophical attempts to formulate a postwar philosophy of history. Bellah, Shwartz, and Eisenstadt, however, brought the term into the realms of natural history. Whereas Bellah focused more on the moral and practical dimensions of the axial age and the effects that it had on human cognition, sociologist Eisenstadt was concerned with the role that intellectuals had in the transformation. His understanding of the term was that axial age civilizations were a type of social order and thus not necessarily bound to an actual historical time period. This approach – which Eisenstadt promoted through a series of conferences with subsequent influential publications (1986, 1987, 1992) – brought Christianity and Islam into the picture (Boy and Torpey, 2013, p. 253). It should be noted, however, that Eisenstadt was convinced that the rise of Axial civilizations was a revolutionary rather than an evolutionary process (reiterated in Eisenstadt, 2012, p. 277). This was because of the social disruption that characterized all of the Axial civilizations.

As John D. Boy and John Torpey note in their detailed history of the term:

> The full-blown shift to a notion of "axiality" emancipated from the specific epoch delineated by Jaspers comes in two variants: that seeking to move "axiality" to an earlier time period, and that seeing it as an occasional phenomenon in human history with no prescribed temporal coordinates whatsoever.
>
> Boy and Torpey, 2013, p. 253

Eisenstadt represents the latter, whereas Jan Assmann's work clearly represents the former. Jan Assmann was a professor of Egyptology in Heidelberg. He is especially important because he consistently placed his work in relation to broader cultural theories and to cultural evolution. He understood the cultural evolution of religion to consist of the transition from 'primary religions' to 'secondary religions'; in other words, from cultures totally immersed in religious worlds to cultures in which religious worlds were relative to one another and could therefore be the objects of individual choice. The latter distinction led to the idea of 'true' and 'false' religions and the consequent idea of sanctions against those who hold the false religions. Assmann considered Egyptian religion as an example of a primary religion and Israelite religion as an example of a secondary religion, also characterized by Assmann – rightly or wrongly – as the transition from polytheism to monotheism (Assmann, 2003). In relation to the Axial debate, Assmann argued that the transition already occurred with the invention of writing (especially the culture of writing) and thus pushed axiality further back in history.

Furthermore, Assmann takes issue with Jaspers' understanding of preaxial and postaxial worlds. Through an analysis of the Egyptian term *ma'at*, 'truth, order, justice,' Assmann argues that the breakthrough of transcendence, which Jaspers considered to be characteristic of the Axial Age, was already represented in Ancient Egypt (Assmann, 1990, p. 121). In his contribution to the anthology on the Axial Age, edited by Bellah and Joas, Assmann argued that the Axial Age is "a creation of philosophers and sociologists, not of historians and philologists," in the sense that the latter are concerned with the past as far as it can be discerned whereas the former are concerned with the roots of modernity. He calls this attempt the quest for beginnings and in this sense exhibiting mythical qualities. "Myths," he argues, "tend to construct beginnings" (Assmann, 2012, p. 366). Assmann's criticism is well-placed and needed, although he is not the only friendly critic of the Axial debate. He rightly points out that the debate began during transitional periods in the nineteenth-century paradigm of progressive evolution and re-invoked in the social upheavals of twentieth-century Europe (2012, p. 369). This should give pause for thought. Assmann is not against the idea of cultural evolution. He simply points out that cultural evolution also involves conscious processes and manipulations by the populations undergoing cultural change (Assmann, 2012, p. 368).

The Axial Age debate is important because it is perhaps the main hypothesis that has succeeded in bringing historians of religions and other historians more or less directly into evolutionary theory. Furthermore, new developments in quantitative approaches to history have targeted, among others, the Axial Age hypothesis as worthy of testing. As will be mentioned below, zoologist Peter Turchin, biologist David Sloan Wilson, and cognitive scientists of religion Harvey Whitehouse and Edward Slingerland are all involved in developing world history databases that can serve as quantitative testing grounds of the historical evidence.

Theories of the Twentieth Century: The Cognitive Science of Religion

Although the cognitive science of religion (CSR) grew out of religious studies, it was also significantly influenced by psychology. It should be noted, however, that it was not general psychology, but a specific branch of the field, namely, evolutionary psychology, that influenced the early phases of the CSR.

As with all of the main sciences mentioned here, history, sociology, comparative religion, and so on, scientific psychology arose with Wilhelm Wundt in Leipzig during the Darwinian revolution of the latter decades of the nineteenth century. It was clearly also the inheritor of the many philosophies of mind prior to that century (Hume, Descartes, Aristotle, and so on). As Plotkin notes in his brief history of evolutionary theory in psychology, the difference between the new psychology and former philosophies of mind was an interest in psycho-physiology and the possibility of quantifying mental phenomena (Plotkin, 2004, p. 29). Wundt, however, also envisioned a supplement to his experimental psychology which was concerned with cultural factors and was like cultural anthropology. He had a significant impact on almost all the main intellectuals of the time – especially Americans (psychologists such as William James, Stanley Hall, Edwin Starbuck, and James Leuba). Sigmund Freud, despite the problems with his theory and methods, situated his ideas within the framework of evolutionary theory (Plotkin, 2004, p. 33). William James, the founder of American psychology, conceived of human psychology in his two-volume *Principles of Psychology* (1890) within the framework of Darwinian theory. However, neither James nor other later influential psychologists, such as John Dewey and B.F. Skinner, actually used evolutionary theory as an analytical tool in empirical studies (Plotkin, 2004, p. 46–47). An exception was James Mark Baldwin who, unfortunately, plagued by scandal, was forced to retire from his academic positions and left the United States. Plotkin argues that Baldwin was a 'real' evolutionary psychologist, that he had an impact on evolutionary theory, that he discovered cognitive developmental stages in humans, that he arguably anticipated the Theory of Mind, and that he contributed significantly to evolutionary epistemology (Plotkin, 2004, p. 70). He preceded by decades the theories of Jean Piaget and Lev Vygotsky. Behaviorism drastically changed the American scene. Although it represented the triumph of empiricism, "Nurture and blank slates ruled, evolution was gone, and so too was the rationalism of Plato and Descartes" (Plotkin, 2004, p. 62).

The cognitive revolution of the 1950s changed all of that especially under the influence of linguist Noam Chomsky along with the other fathers of the revolution, namely George Miller, Johan McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon. Although the break with behaviorism allowed the scientific study of the mind, the main paradigm was of course the mind as an information processor. As philosopher Paul Thagard has shown in his history of the cognitive sciences, there are six competing, independent intellectual traditions that have so far failed to produce a unified theory of mind (Thagard, 1996, p. 128).

The real breakthrough in the sciences of the mind came in the late 1980s with the founding of evolutionary psychology out of the ashes of sociobiology. The latter was an attempt at applying animal ethology to human behavior, most spectacularly demonstrated in biologist Edward O. Wilson's *Sociobiology: The New Synthesis* (1975). As Kevin N. Laland and Gillian R. Brown note in their research history of evolutionary perspectives on human behavior, Wilson's book opened "a land rush of biologists into the territory of the human sciences, where they received an extremely hostile reception" (Laland and Brown, 2002, p. 71). This hostility, however, brought a well-deserved backlash most persuasively led by the founders of the first Center for Research in Evolutionary Psychology, anthropologist John Tooby and psychologist Leda Cosmides. Tooby and Cosmides formulated their thesis in direct opposition to what they polemically called the Standard Social Science Model (SSSM). They posit what they call the Integrated Causal Model (ICM). In their groundbreaking chapter in *The Adapted Mind* (1992), Tooby and Cosmides argued that the cognitive tasks in what they termed the "environment of evolutionary adaptedness (EEA)," roughly covering the stone age, could not have been dealt with through socialization processes. Our ability to adapt to the challenges of the environment was ensured by genetically evolved domain-specific modules. They thus rejected the idea that the brain is a domain-general learning mechanism. In their view:

> … culture is the manufactured product of evolved psychological mechanisms situated in individuals living in groups. Culture and human social behavior is complexly variable, but not because the human mind is a social product, a blank slate, or an externally programmed general-purpose computer, lacking a richly defined evolved structure. Instead, human culture and social behavior is richly variable because it is generated by an incredibly intricate, contingent set of functional programs that use and process information from the world, including information that is provided both intentionally and unintentionally by other human beings.
>
> Tooby and Cosmides, 1992, p. 24

Although they succeeded in bringing the mind back into evolutionary theory, their adamant modularity has been criticized by many, and Laland and Brown argue that evolutionary psychologists should pay more attention to evolutionary biologists who have developed rigorous methods for analyzing evolutionary processes (Laland and Brown, 2002, p. 194).

Tooby and Cosmides were not interested in religion, but they had a decided impact on the cognitive science of religion. In fact, the first to consciously formulate the new cognitive science of religion, scholars of religion E. Thomas Lawson and Robert N. McCauley (1990), did so as an alternative to the dominant, mostly postmodern and antiscience, cultural studies that also came to dominate religious studies first in the United States and later in Europe. Attempts by historian of religions Jonathan Z. Smith to transform religious studies to an analytical and critical endeavor during the 1970s and 1980s foreshadowed the birth of the cognitive science of religion, although Smith himself never dealt with the topic of cognition except to say that "the burgeoning enterprise of cognitive theories of religion has begun to provide theoretically interesting and important empirical observations toward an evolutionary account of

religion…[but] not having yet fully mastered the relevant literature has prevented me from doing much more, in public speech, than calling attention to this development and signaling my interest in it" (Smith, 2011, p. 3).

There were precedents to the CSR of course. One of the most obvious was anthropologist Stewart Guthrie's groundbreaking 1980 article "A Cognitive Theory of Religion" which was published in the prestigious anthropological journal *Current Anthropology*. It was published as a target article with peer response. In it, Guthrie argues that "religious and nonreligious belief and behavior can best be understood as closely related variants of a single human cognitive process, not as polar opposites" (Guthrie, 1980, p. 181), a statement addressed to the rationality debate at the time and not to evolutionary theory. He was also mainly concerned with defining religion. His main hypothesis, however, which he developed in a major book 13 years later (1993) is that religion is a form of anthropomorphization. This hypothesis has been widely acknowledged as a key contribution to an evolutionary theory of religion. The gist of the hypothesis is that religion is "systematic anthropomorphism: the attribution of human characteristics to nonhuman things or events" (1993, p. 3). The strategy of attributing human like characteristics to nature helps us navigate in an uncertain world. Even though we often err, the times we are right more than justify the strategy in terms of survival. Thus, Guthrie's theory is both a cognitive and an evolutionary one. Guthrie draws on cognitive theories of perception which indicate that perception is predisposed to anthropomorphism. Although he formalizes his theory in terms of five propositions and two informal deductions, Guthrie does not systematically explore anthropomorphism in terms of evolutionary theory. He does, however, state our capacity for anthropomorphization is built in: "not mere opportune exploitation of anatomy and neurology, but a biologically broadly based, deeply integrated system" (1980, p. 189). The challenge of formulating his hypothesis in evolutionary terms was later taken up by psychologist Justin Barrett.

Barrett performed tests with Frank C. Keil on university students' tendency to anthropomorphize in their concepts of God (Barrett and Keil, 1996). Later, he postulated a cognitive mechanism he termed 'hyperactive agency detection device' (HADD) which he claims is an essential element in the origin of religious thought, more specifically, belief in supernatural beings (systematically argued in his popularized book *Why Would Anyone Believe in God?*, 2004). By device, he means evolved, specialized mental tools in the brain that help us survive in the world. He argues that these tools consist of categorizers, describers, and facilitators. The categorizers consist of among other things an agency detection device. Among describers, one finds Theory of Mind (ToM), the ability to attribute mental states to other people. These two in combination are essential to the evolution of God beliefs. Three facilitators are also essential to the development of religious beliefs, namely, social exchange regulator, social status monitor, and intuitive morality (Barrett, 2004, p. 3–5). Barrett is agnostic about whether these mental tools are hardwired in the brain and thus leaves open the question of whether these tools really are evolutionary adaptations.

Lawson and McCauley formulated their dissatisfaction with cultural studies by arguing that their approach is explanatory rather than interpretive. Without regard to cognitive constraints, they argue, interpretations will be misguided. Drawing inspiration from the competence approach in linguistics (Chomsky), they claim that participants' representations of ritual acts constrain both the form and types of acceptable ritual behavior in a given religious system (Lawson and McCauley, 1990, p. 8; 78). Their theory has three major components: (1) an action representation system, (2) a religious conceptual scheme, and (3) universal principles of religious ritual structure (1990, p. 121). Rituals are thus analyzed much along the same lines as in generative linguistics. The universal principles, they argue, are the core of their theory and "offer some basis for formulating a genuine explanation of these phenomena" (1990, p. 121). Although they are simply postulated, the principles hinge on their action representation model. The first principle is the Principle of Superhuman Agency and the second is the Principle of Superhuman Immediacy. Their proximity and position in ritual action determines particular types of rituals (1990, p. 125). Lawson and McCauley do not specifically address evolutionary issues.

In a second book, which is mostly concerned with Harvey Whitehouse's modes of religiosity hypothesis, they add the further elements of sensory pageantry, memory, and motivation to their ritual form hypothesis in order to reveal "larger patterns in the evolution of religious ritual systems…that cut across cultures and historical epochs" (McCauley and Lawson, 2002, p. 7). Their goal is to delineate "the cognitive architecture of *Homo religiosus* – not merely to understand well-known historic patterns in religious systems better but also to *explain* them" (McCauley and Lawson, 2002, p. 8). But they are more specific in relating their model to evolutionary theory. Especially in terms of memory, their interest is on what psychological mechanisms ensure the faithful transmission of religious and ritual representations across generations. They argue:

> … if religious systems, in fact, evolve in the direction of greater memorability, at least in the absence of special cultural tools for the faithful preservation of such representations, then the relevance of these matters to the transmission of religious ritual should be clear …. [M]ost religious systems, including all of the great world religions, emerged among populations that were mostly illiterate…. Thus, it should come as no surprise that religious systems and ritual systems, in particular, have evolved so as to exploit variables that facilitate memory.
>
> McCauley and Lawson, 2002, p. 85

These issues are systematically explored in their analyses of the evolution of particular rituals and religious systems in New Guinea.

The issue of memorability is the main point of departure for the other founding father of CSR, anthropologist Pascal Boyer. Already in his earliest volume, *Tradition as Truth and Communication* (1990), Boyer argues that anthropology needs a strong psychological theory if it wishes to explain tradition. He rejects the usual explanations of socialization, cultural models, and/or meaningful symbols so common in anthropological research. Instead, he posits in his 1994 book *The Naturalness of Religious Ideas* that the success of religious

ideas is that they are well-formed in terms of evolved cognitive domains or repertoires, which he later termed 'templates' in his 2001 book *Religion Explained*. In this sense, religious ideas are natural. He notes that religious ideas have three features: they contain counterintuitive claims, inscrutable concepts and inferential gaps (1994, p. 53). Thus religious ideas build upon and use normal cognitive domains. These domains supply information intuitively and effortlessly. Inference systems, he notes, direct our knowledge acquisition and assist us in surviving in our environments. To this more or less ready information is added a slight tweak or counterintuitive property, such as tigers that fly, statues that listen, or humans that rise from the dead, which make the property cognitively salient and thus more easily remembered and passed on. Boyer neatly sums up what has become the standard understanding of religion by CSR scholars:

> (1) Most of the relevant mental machinery is not consciously accessible. People's explicitly held, consciously accessible beliefs, as in other domains of cognition, only represent a fragment of the relevant processes. Experimental tests show that people's actual religious concepts often diverge from what they believe they believe. This is why theologies, explicit dogmas, scholarly interpretations of religion cannot be taken as a reliable description of either the contents or the causes of people's beliefs; (2) What makes religious thoughts 'natural' might be the operation of a whole collection of distinct mental systems rather than a unique, specific process; (3) In each of these systems religious thoughts are not a dramatic departure from, but a predictable by-product of, ordinary cognitive function.
>
> Boyer, 2003, p. 119

Thus, the by-product argument came to influence the burgeoning new cognitive science of religion. This has become a point of disagreement in relation to later approaches. It should be noted that Boyer recently has developed a more sophisticated argument concerning religious ideas based on experimental evidence. Thus he argues in his article "Functional Origins of Religious Concepts" (2000) that religious ideas are not solely salient because of their counterintuitive properties. This is the so-called "Mickey Mouse problem": not all counterintuitive agents are religiously salient. Those agents that are religiously salient are connected to morality, group identity, ritual, and emotion, and, furthermore, have perfect access to socially strategic information. The mental mechanisms that our evolutionary heritage has given to social interaction are a hypertrophy of social intelligence, a capacity to evaluate potential cooperators and to detect potential defectors, a repertoire of moral feelings, the stability of some unfakeable (or difficult to fake) signals, easily produced self-deception, and emotional rewards for gossip (Boyer, 2000, p. 203–204).

As noted above, most cognitive scientists of religion accepted the computer metaphor of the mind and rejected any causal influence that culture might have on cognition (Geertz, 2004, 2010b). Furthermore, they assume all religious thought and behavior to be by-products of naturally selected cognitive mechanisms. Others take issue with this assumption. Anything as costly as religious behavior must have had selective advantages, if not on an individual level, surely on a group level. Furthermore, the interrelations between cognition and culture are mutual. The whole area of epigenesis is currently being explored by many biologists, philosophers, and others who point out that the relations between genes and culture are not at all as clear-cut as has been assumed (Richerson and Boyd, 2005; Tomasello, 1999). Sociocultural anthropologist Benjamin Purzycki and behavioral ecologist Richard Sosis succinctly stated:

> … religion is a transsomatic adaptation. At the genic level, the religious system constitutes an extended phenotype that has been fashioned by natural selection to overcome socioecological challenges inherent in human sociality, primarily problems of cooperation and coordination. At the collective level, the religious system constitutes a cognitive niche.
>
> Purzycki and Sosis, 2013, p. 99

As a historian of religion, I have argued that we need to push further back into our evolutionary sapient history in order to understand the origins of religion. I have long argued that culture and cognition became so closely meshed, even before the appearance of *Homo sapiens* that clear-cut causal relations are almost impossible to distinguish in both natural and experimental contexts. It seems fairly certain that tool-making was one of the prime causal mechanisms behind the expansion of the brain and other physiological changes in the hominin line. Fire control and cooking were most likely also contributing factors to the development of our cognitive abilities. The use of symbolic objects and possible symbolic behavior seems to be fairly ancient, most likely during the time of *Homo erectus*. The kinds of minds we must look for in the archaeological record would have a finely honed social cognition, a drive to communicate and cooperate, a self-deceptive brain, and a superstitious brain prone to unusual mental and/or emotional experiences (Geertz, 2013, p. 31). These characteristics are based on recent neurological and neuropsychological insights. Brains are prediction organs busily constructing the world in order to help us negotiate and survive in natural, social, and psychological environments. They are creative organs, filling in all of the missing data in the sensed world. One side effect, however, is that they are also highly vulnerable to influence when manipulated in specific ways. I therefore argue:

> We are intelligent apes that are highly emotional, easily spooked, very superstitious, extremely sensitive to social norms and virtual realities, and equipped with nervous systems that are vulnerable to influence from conspecifics and their symbolic worlds. These traits are prerequisites for religious behaviour.
>
> Geertz, 2013, p. 19

Religious behavior would have evolved gradually, but it evolved with an evolving brain that evolved in particular ways, especially in the prefrontal cortex where the executive functions developed. Planning, conscious deliberation, identity, self-reflection, and communication are all closely interwoven in these areas. And they coevolved with culture. Thus, the result is

that we are biocultural creatures (drawing on Merlin Donald; Geertz, 2010a) that are born unfinished but ready to adapt to the cultural environment within which the brain grows and matures. Our brains are helpless without cultural frameworks, role models, norms and rules as well as an intense interaction of feeling felt by emotional conspecifics (Siegel, 2001; Turner and Maryanski, 2008). Religion, like other cultural traits, would thus have a selective advantage for such brains.

One of the more robust adaptive hypotheses was developed by Richard Sosis and scholar of religion Joseph Bulbulia. Drawing on the costly signaling theory of Amotz Zahavi and Avishag Zahavi, Sosis and Bulbulia argue that religious behavior can be defined as "a costly signal capable of authenticating religious commitment" (Bulbulia, 2004, p. 669). In a comparative study of 200 religious and secular communities in nineteenth-century United States, Sosis found that religious communities lasted four times longer than secular ones even though they imposed twice as many costs (Sosis, 2000; Sosis and Bressler, 2003). The use of such signals seems to solve the problems of establishing trust among strangers, detecting free-riders and defectors, and generally providing a means of predicting behavior: "We believe that cultural evolutionary scholars are correct to notice…that norms evolve to manage problems of cooperative prediction among strangers, the solution of which is a condition for the possibility of cooperation in large social worlds" (Bulbulia and Sosis, 2011, p. 373–374).

A variant of this hypothesis, proposed by anthropologist Joseph Henrich, concerns learners' prestige-bias. The hypothesis is that learners have "evolved to attend to *credibility enhancing displays* (CREDS)" that "provide the learner with reliable measures of the model's actual degree of commitment to (or belief in) the representations that he has inexpensively expressed symbolically (e.g., verbally)" (Henrich, 2009, p. 244–245). In this sense, CREDs are cultural immune systems that have spread by cultural group selection.

One of the problems confronting cognitive scientists of religion is the inability to test their evolutionary hypotheses. Since quantifiable historical data are either lacking or are of questionable quality, the only other option is to perform psychological experiments in the laboratory or in the field. The logic is that if psychological tests can demonstrate intuitive responses from participants, then it is likely that the responses are universal traits, constraints, mechanisms, and so on, that again, most likely, were selected for during our evolution. Since most of the participants are Western, Educated, Industrialized, Rich, and Democratic (dubbed WEIRD by Henrich et al., 2010), it is surprising that psychologists haven't been more self-critical. Fortunately, cross-cultural studies are being performed, but a large number of them consist of self-report surveys, which of course is almost as problematic. Then there is the whole problem of priming (Bargh, 2006; Doyen et al., 2012) and, finally, the most worrying tendency of skipping replication (or ignoring nonreplication) and accepting 'consensus' as a kind of fact. These are not weaknesses affecting the CSR in particular. It is a general problem in experimental psychology and in social priming (Kahneman, 2012). CSR scholars are, however, plagued by these problems as well.

There is a growing number of articles claiming support, or not, of the main CSR hypotheses. One obvious candidate for testing is the claim that religions, and especially religious rituals, promote prosociality. Realizing that the CSR has mostly remained on a metatheoretical level, younger generations have become quite adept at designing psychological experiments, computer simulations, and field experiments. Various teams have performed field experiments among people who use firewalking and/or piercing rituals. The main hypotheses that have been tested concern whether extreme rituals promote prosociality and solidarity (Konvalinka et al., 2011; Xygalatas et al., 2013a; Fischer et al., 2013), whether extreme rituals stimulate episodic memory (Xygalatas et al., 2013b), and whether rhythm and synchrony promote prosociality (Cohen et al., 2010; Reddish et al., 2014). All of these hypotheses are well-known in the science of religion literature and in psychological literature.

In their review article, scholars of religion Panagiotis Mitkidis and Gabriel Levy have shown, however, that the evidence on whether religion promotes prosociality is contradictory and some of the positive results may be due to other factors. Studies showing that religious individuals are more altruistic, generous, cooperative, etc. are plagued by limitations and biases. Survey and correlational studies are often inadequate in determining causality or in identifying confounding factors. Self-report studies may be influenced by participants' social desirability biases. They argue that "despite the occasional study that has shown a weak or ambiguous correlation between aspects of religiosity and (actual) prosocial behavior…, the bulk of the evidence suggests that such a correlation is absent … or even negative" (Mitkidis and Levy, 2014, p. 3–4). Neither religion nor religious rituals promote prosociality on their own. Religious institutions have, however, organized "a matrix of other actions and exposures that do promote prosociality" and bring people into situations exposing them to effective primes (Mitkidis and Levy, 2014, p. 6). An example of the latter is the use of eyes, whether in paintings, icons, images, or symbols. The experimental evidence indicates that the mere presence of eyes or eye-like symbols affect participants' behavior (e.g., Haley and Fessler, 2006). They conclude:

> In other words, in our understanding religion is not a powerful social force so much because of what it does but because of what people think it does. Religion is a set of institutionalized attitudes (such as beliefs) and practices organized around one or more superhuman agents and branded as a guarantee of moral norms. Religious institutions surround themselves with priming technologies that often produce prosociality – they attempt to monopolize these technologies – but religions are in no way the only institutions that do so. In this sense, any particular prosociality due to religion is more of an accident than a product of design.

Another strong candidate for an evolutionary hypothesis is that religious behavior promotes better health. This area, however, is particularly plagued with faulty designs and overblown conclusions. A recent study has shown that prayer does not relieve pain physiologically. Religious participants, however, self-report a subjective reduction in pain intensity and pain unpleasantness (Jegindø et al., 2013a). A follow-up

test using fMRI technology repeated the results despite the fact that some of the religious participants were administered naloxone, an opioid-blockage medicine (Jegindø et al., in review). In extreme rituals, performed on Mauritius, the team found that dissociation experienced during the rituals were effective ways of coping with the pain (Jegindø et al., 2013b). The fact that placebo may be at play is another example of how religious institutions may have benefited by manipulating expectations and biological mechanisms (Tracey, 2010; Vase et al., 2009, 2011). Despite claims of a consensus on the positive effects of religious behavior and belief on health (for instance Koenig et al., 2001; see critique of the field by Sloan, 2006), the most one can claim is that religious behavior may help patients cope with their afflictions (Pargament, 1997). Even meditation studies and mindfulness studies are inconclusive (Cahn and Polich, 2006; Goyal et al., 2014). There is no doubt that religious institutions have throughout history dealt with health, illness, and death. But again, these institutions have learned to use endogenous mechanisms already in place.

A promising development in testing evolutionary hypotheses of religion is the use of simulations and databases. Examples of the use of simulations to test hypotheses is well-known in biology but not in the humanities (Nielbo et al., 2012). Anthropologist Harvey Whitehouse has attempted to use these new tools in testing his modes of religiosity hypothesis (2000; 2004). Briefly, Whitehouse hypothesizes that the dichotomy of religious experience (i.e., the ecstatic and the more mundane) often noted by scholars of religion can be explained in terms of why people remember their beliefs and rituals and why they are motivated to pass them on. The two types of religiosity, which he calls the doctrinal and the imagistic modes of religiosity, are caused by the ritual stimulation of semantic memory for the former and episodic memory for the latter. Thus, in the doctrinal mode, frequent repetition activates semantic memory for religious teachings and implicit memory for religious rituals. This type of religiosity is usually found in large, anonymous communities led by centralized religious leadership. In the imagistic mode, on the other hand, infrequent repetition accompanied by high arousal activates episodic memory. Since such religious behavior is normally localized in small groups without centralized leadership, the individual is moved by his or her own spontaneous reflections, thus leading to a diversity of religious representations. The modes theory posits psychological causes of sociopolitical features (2002, p. 308–309). Because such features are long term and involve larger social dynamics, testing the theory is a formidable problem. Thus there is a need for simulations and large-scale databases.

Whitehouse has been at the forefront of pursuing both types of testing which is concurrent with developments in quantitative and experimental approaches to history. In using simulations, Whitehouse discovered aspects in the modes theory that were underspecified, thus generating new hypotheses for future research (Whitehouse et al., 2012). These included better predictions of the effects of ritual frequency and systemic rigidity; the hidden complexities of frequency and intuitiveness; discoveries on the relationships between frequency, arousal, and motivation; a more nuanced insight into imagistic splintering and insights into the between-group competition among doctrinal traditions (Whitehouse et al., 2012, p. 197–198).

Perhaps one of the most significant advances in testing evolutionary theories of religion (or any theories of religion for that matter) is efforts in building databases that are longitudinal and fine-grained. Three separate yet linked databases are being developed by teams under the leadership of scholar of religion Edward Slingerland, Harvey Whitehouse, and Peter Turchin. In a signal article from 2012, they have sketched out the usefulness and design of such a database (Turchin et al., 2012). The goal of such databases is to test rival theories with historical data. This involves the following procedure:

- Specify the main question and identify empirical patterns that need to be explained.
- Develop a set of competing hypotheses that make different empirical predictions. This phase may require building mathematical models as an intermediate step, for example, when theories postulate a complex set of interrelated processes.
- Code the data and perform statistical analyses of the resulting database to determine which hypotheses most parsimoniously explain the empirical patterns. (Turchin et al., 2012, p. 273)

Turchin and colleagues have recently demonstrated the usefulness of this approach in testing his theory on the relations between war and the development of complex societies (Turchin et al., 2013).

Conclusion

Thus, there is an endless variety of evolutionary theories of religion, none of which have attained the status of fact. Many of the theories are plausible; however, very few are amenable to experimental design. Furthermore, the burgeoning field of the cognitive science of religion is still in the process of struggling with the same problems found in other psychological sciences, such as design, priming, replication, and cross-cultural validity. In the long run, the future is promising. With the development of longitudinal and fine-grained databases, it may become possible for scholars to test competing theories and reach more definitive conclusions, at the very least on which ideas should be rejected.

See also: Cognition, Evolution of; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Darwinism; Epigenetic Inheritance; Evolution, History of; Evolutionary Developmental Psychology; Evolutionary Social Psychology; Gene–Culture Coevolution; Human Behavioral Ecology; Religion and Health: Sociological Considerations; Religion, Psychology of; Sociobiology, History of.

Bibliography

Assmann, Jan, 1990. Ma'at: Gerechtigkeit und Unsterblichkeit im Alten Ägypten. C.H. Beck, München.

Assmann, Jan, 2003. Die Mosaische Unterscheidung oder der Preis des Monotheismus. Hanser, München.

Assmann, Jan, 2012. Cultural memory and the myth of the axial age. In: Bellah, Robert N., Joas, Hans (Eds.), The Axial Age and its Consequences. The Belknap Press of Harvard University Press, Cambridge, London, pp. 366–407.

Bargh, John A., 2006. What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. European Journal of Social Psychology 36 (2), 147–168.

Barrett, Justin L., 2004. Why Would Anyone Believe in God? AltaMira Press, Walnut Creek.

Barrett, Justin L., Keil, Frank C., 1996. Conceptualizing a nonnatural entity: anthropomorphism in god concepts. Cognitive Psychology 31, 219–247.

Bellah, Robert N., 1964. Religious evolution. American Sociological Review 29 (3), 358–374.

Bellah, Robert N., 2005. What is axial about the axial age? Archives Européennes de Sociologie 46 (1), 69–89.

Bellah, Robert N., 2011. Religion in Human Evolution: From the Paleolithic to the Axial Age. Harvard University Press, Cambridge, London.

Bellah, Robert N., Joas, Hans (Eds.), 2012. The Axial Age and its Consequences. The Belknap Press of Harvard University Press, Cambridge, London.

Berghe, Pierre L. van den, 1990. Why most sociologists don't (and won't) think evolutionarily. Sociological Forum 5 (2), 173–185.

Boy, John D., Torpey, John, 2013. Inventing the axial age: the origins and uses of a historical concept. Theoretical Sociology 42, 241–259.

Boyer, Pascal, 1990. Tradition as Truth and Communication: A Cognitive Description of Traditional Discourse. Cambridge University Press, Cambridge.

Boyer, Pascal, 1994. The Naturalness of Religious Ideas: A Cognitive Theory of Religion. University of California Press, Berkeley, Los Angeles, London.

Boyer, Pascal, 2000. Functional origins of religious concepts: ontological and strategic selection in evolved minds. The Journal of the Royal Anthropological Institute 6 (2), 195–214.

Boyer, Pascal, 2001. Religion Explained: The Evolutionary Origins of Religious Thought. Basic Books, New York.

Boyer, Pascal, 2003. Religious thought and behaviour as by-products of brain function. TRENDS in Cognitive Sciences 7 (3), 119–124.

Bulbulia, Joseph, 2004. The cognitive and evolutionary psychology of religion. Biology and Philosophy 19, 655–686.

Bulbulia, Joseph, Sosis, Richard, 2011. Signalling theory and the evolution of religious cooperation. Religion 41 (3), 363–388.

Bulbulia, Joseph, Sosis, Richard, Harris, Erica, Genet, Russell, Genet, Cheryl, Wyman, Karen (Eds.), 2008. The Evolution of Religion: Studies, Theories, & Critiques. Collins Foundation Press, Santa Margarita.

Cahn, B. Rael, Polich, John, 2006. Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychological Bulletin 132 (2), 180–211.

Carneiro, Robert L., 2003. Evolutionism in Cultural Anthropology: A Critical History. Westview Press, Boulder.

Chidester, David, 1996. Savage systems: Colonialism and Comparative Religion in Southern Africa. University Press of Virginia, Charlottesville, London.

Cohen, Emma E.A., Ejsmond-Frey, Robin, Knight, Nicola, Dunbar, R.I.M., 2010. Rowers' high: behavioural synchrony is correlated with elevated pain thresholds. Biology Letters 6, 106–108.

Darwin, Charles, 1874. The Descent of Man, and Selection in Relation to Sex. Reprinted from the Second English Edition, Revised and Augmented, second ed. John Murray, London. 1871; reprint. 1874. A.L. Burt, New York.

Donald, Merlin, 1991. Origins of the Modern Mind: Three States in the Evolution of Culture and Cognition. Harvard University Press, Cambridge, London.

Donald, Merlin, 2001. A mind so rare: the evolution of human consciousness. W.W. Norton & Company, New York, London.

Doyen, Stéphane, Klein, Olivier, Pichon, Cora-Lise, Cleeremans, Axel, 2012. Behavioral priming: it's all in the mind, but whose mind? PLoS ONE 7 (1), 1–7.

Eisenstadt, Shmuel N. (Ed.), 1986. The Origins and Diversity of Axial Age Civilizations. State University of New York Press, Albany.

Eisenstadt, Shmuel N. (Ed.), 1987. Kulturen der Achsenzeit I: Ihre Ursprünge und ihre Vielfalt. Suhrkamp, Frankfurt am Main.

Eisenstadt, Shmuel N. (Ed.), 1992. Kulturen der Achsenzeit II: Ihre institutionelle und kulturelle Dynamik. Suhrkamp, Frankfurt am Main.

Eisenstadt, Shmuel N., 2012. The axial conundrum between transcendental visions and vicissitudes of their institutionalizations: constructive and destructive possibilities. In: Bellah, Robert N., Joas, Hans (Eds.), The Axial Age and its Consequences. The Belknap Press of Harvard University Press, Cambridge, London, pp. 277–293.

Fischer, Ronald, Callander, Rohan, Reddish, Paul, Bulbulia, Joseph, 2013. How do rituals affect cooperation? An experimental field study comparing nine ritual types. Human Nature 24, 115–125.

Geertz, Armin W., 1999. Definition as analytical strategy in the study of religion. Historical Reflections/Reflexions Historiques 25 (3), 445–475.

Geertz, Armin W., 2004. Cognitive approaches to the study of religion. In: Antes, Peter, Geertz, Armin W., Warne, Randi R. (Eds.), New Approaches to the Study of Religion. Volume 2: Textual, Comparative, Sociological, and Cognitive Approaches. Walter de Gruyter, Berlin, New York, pp. 347–399.

Geertz, Armin W., 2010a. Brain, body and culture: a biocultural theory of religion. Method and Theory in the Study of Religion 22 (4), 304–321.

Geertz, Armin W., 2010b. Too much mind and not enough brain, body and culture: on what needs to be done in the cognitive science of religion. Historia Religionum. An International Journal 2, 21–37.

Geertz, Armin W., 2013. Whence religion? How the brain constructs the world and what this might tell us about the origins of religion, cognition and culture. In: Geertz, Armin W. (Ed.), Origins of Religion, Cognition and Culture. Acumen Publishing Limited, Durham, pp. 17–70.

Geertz, Armin W. (Ed.), 2013. Origins of Religion, Cognition and Culture. Acumen Publishing Limited, Durham.

Goyal, Madhav, Singh, Sonal, Sibinga, Erica M.S., Gould, Neda F., Rowland-Seymour, Anastasia, Sharma, Ritu, Berger, Zackary, Sleicher, Dana, Maron, David D., Shihab, Hasan M., Ranasinghe, Padmini D., Linn, Shauna, Saha, Shonali, Bass, Eric B., Haythornthwaite, Jennifer A., 2014. Meditation programs for psychological stress and well-being: a systematic review and meta-analysis. JAMA Internal Medicine 174 (3), 357–368.

Guthrie, Stewart Elliott, 1980. A cognitive theory of religion. Current Anthropology 21 (2), 181–203.

Guthrie, Stewart Elliott, 1993. Faces in the Clouds: A New Theory of Religion. Oxford University Press, Oxford, New York.

Haley, Kevin J., Fessler, Daniel M.T., 2006. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evolution and Human Behavior 26, 245–256.

Harris, Marvin, 1969. The Rise of Anthropological Theory: A History of Theories of Culture. Routledge & Kegan Paul, London.

Harris, Marvin, 1975. Cows, Pigs, Wars & Witches: The Riddles of Culture. Hutchinson & Co. Ltd, London.

Harris, Marvin, 1977. Cannibals and Kings: The Origins of Cultures. Vintage Books, New York.

Henrich, Joseph, 2009. The evolution of costly displays, cooperation and religion: credibility enhancing displays and their implications for cultural evolution. Evolution and Human Behavior 30, 244–260.

Henrich, Joseph, Heine, Steven J., Norenzayan, Ara, 2010. The weirdest people in the world? Behavioral and Brain Sciences 33, 61–135.

James, William, 1890. The Principles of Psychology. In Two Volumes. Henry Holt and Company, New York, 1890; reprint, 2012. University of Michigan Library, Ann Arbor.

Jaspers, Karl, 1953. Vom Ursprung und Ziel der Geschichte. Zürich: Artemis/Munich: Piper. English: The origin and goal of history, trans. Michael Bullock. Routledge & Kegan Paul/New Haven, London, 1953. Yale University Press,.

Jegindø, Else-Marie Elmholdt, Vase, Lene, Skewes, Joshua Charles, Terkelsen, Astrid Juhl, Hansen, John, Geertz, Armin W., Roepstorff, Andreas, Jensen, Troels Staehelin, 2013a. Expectations contribute to reduced pain levels during prayer in highly religious participants. Journal of Behavioral Medicine 36 (4), 413–426.

Jegindø, Else-Marie Elmholdt, Vase, Lene, Jegindø, Jens, Geertz, Armin W., 2013b. Pain and sacrifice: experience and modulation of pain in a religious piercing ritual. International Journal for the Psychology of Religion 23 (3), 171–187.

Jegindø, Else-Marie Elmholdt, Skewes, Joshua, Dietz, Martin, Møller, Arne, Roepstorff, Andreas, Wiech, Katja, Jensen, Troels Staehelin. Pain modulation by religious prayer is associated with reduced BOLD signal in Parietofrontal networks, in review.

Joas, Hans, 2012. The axial age debate as religious discourse. In: Bellah, Robert N., Joas, Hans (Eds.), The Axial Age and its Consequences. The Belknap Press of Harvard University Press, Cambridge, London, pp. 9–29.

Kahneman, Daniel, 2012. A proposal to deal with questions about priming effects. E-mail letter to students of social priming widely circulated through the electronic media.

Kearney, Michael, 1984. World View. Chandler & Sharp Publ., Inc, Novato, California.

Koenig, Harold G., McCullough, Michael E., Larson, David B., 2001. Handbook of Religion and Health. Oxford University Press, Oxford.

Konvalinka, Ivana, Xygalatas, Dimitris, Bulbulia, Joseph, Schjødt, Uffe, Jegindø, Else-Marie, Wallot, Sebastian, Van Orden, Guy, Roepstorff, Andreas, 2011. Synchronized arousal between performers and related spectators in a fire-walking ritual. Proceedings of the National Academy of Sciences 108 (20), 8514–8519.

Laland, Kevin N., Brown, Gillian R., 2002. Sense and Nonsense: Evolutionary Perspectives on Human Behaviour. Oxford University Press, Oxford.

Lawson, E. Thomas, McCauley, Robert N., 1990. Rethinking Religion: Connecting Cognition and Culture. Cambridge University Press, Cambridge.
McCauley, Robert N., Lawson, E. Thomas, 2002. Bringing Ritual to Mind: Psychological Foundations of Cultural Forms. Cambridge University Press, Cambridge.
Mitkidis, Panagiotis, Levy, Gabriel. False advertising: the attractiveness of religion as a moral brand. In: Slone, J., van Slyke, J. (Eds.), Religion and Sexual Selection: The Attraction of Religion. Bloomsbury Publishing, London, pp. 1–19, in press.
Nielbo, Kristoffer L., Braxton, Donald M., Upal, Afzal, 2012. Computing religion: a new tool in the multilevel analysis of religion. Method and Theory in the Study of Religion 24 (3), 267–290.
Pargament, Kenneth I., 1997. The Psychology of Religion and Coping: Theory, Research, Practice. The Guilford Press, New York, London.
Plotkin, Henry, 2004. Evolutionary Thought in Psychology: A Brief History. Blackwell Publishing, Oxford.
Purzycki, Benjamin, Sosis, Richard, 2013. The extended religious phenotype and the adaptive coupling of ritual and belief. Israel Journal of Ecology & Evolution 59 (2), 99–108.
Rappaport, Roy A., 1999. Ritual and Religion in the Making of Humanity. Cambridge University Press, Cambridge.
Reddish, Paul, Bulbulia, Joseph, Fischer, Ronald, 2014. Does synchrony promote generalized prosociality? Religion, Brain & Behavior 4 (1), 3–19.
Richerson, Peter J., Boyd, Robert, 2005. Not by Genes Alone: How Culture Transformed Human Evolution. The University of Chicago Press, Chicago, London.
Siegel, Daniel J., 2001. Toward an interpersonal neurobiology of the developing mind: attachment relationships, "mindsight," and neural integration. Infant Mental Health Journal 22 (1–2), 67–94.
Sloan, Richard P., 2006. Blind Faith: The Unholy Alliance of Religion and Medicine. St. Martin's Press, New York.
Smart, Ninian, 1989. The World's Religions: Old Traditions and Modern Transformations. Cambridge University Press, Cambridge, 1989, paperback edition 1992.
Smith, Jonathan Z., 1987. To Take Place: Toward Theory in Ritual. University of Chicago Press, Chicago, London.
Smith, Jonathan Z. JZS Panel Presentation AAR 11-20-11, 'Conversation with Robert Bellah on Religion in Human Evolution.' Unpublished paper, 9pp.
Sosis, Richard, 2000. Religion and intragroup cooperation: preliminary results of a comparative analysis of utopian communities. Cross-Cultural Research 34 (1), 77–88.
Sosis, Richard, Bressler, Eric R., 2003. Cooperation and commune longevity: a test of the costly signaling theory of religion. Cross-Cultural Research 37 (2), 11–39.
Thagard, Paul, 1996. Mind: Introduction to Cognitive Science. The MIT Press, Cambridge, London.
Tomasello, Michael, 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, London.
Tooby, John, Cosmides, Leda, 1992. The psychological foundations of culture. In: Barkow, Jerome H., Cosmides, Leda, Tooby, John (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford, New York, pp. 19–136.
Tracey, I., 2010. Getting the pain you expect: mechanisms of placebo, nocebo, and reappraisal effects in humans. Nature Medicine 16, 1277–1283.
Turchin, Peter, Whitehouse, Harvey, François, Pieter, Slingerland, Edward, Collard, Mark, 2012. A historical database of sociocultural evolution. Cliodynamics: the Journal of Theoretical and Mathematical History 3 (2), 271–293.
Turchin, Peter, Currie, Thomas E., Turner, Edward A.L., Gavrilets, Sergey, 2013. War, space, and the evolution of old world complex societies. Proceedings of the National Academy of Sciences 110 (41), 16384–16389.
Turner, Jonathan H., Maryanski, Alexandra, 2008. On the Origin of Societies by Natural Selection. Paradigm Publishers, Boulder, London.
Vase, Lene, Petersen, G.L., Riley III, J.L., Price, D.D., 2009. Factors contributing to large analgesic effects in placebo mechanism studies conducted between 2002 and 2007. Pain 145, 17–25.
Vase, Lene, Nørskov, K.N., Petersen, G.L., Price, D.D., 2011. Patients' direct experiences as central elements of placebo analgesia. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 366, 1913–1921.
White, Leslie A., 1987. In: Dillingham, Beth, Carneiro, Robert L. (Eds.), Ethnological Essays. University of New Mexico Press, Albuquerque.
Whitehouse, Harvey, 2000. Arguments and Icons: Divergent Modes of Religiosity. Oxford University Press, London, New York.
Whitehouse, Harvey, 2002. Modes of religiosity: towards a cognitive explanation of the sociopolitical dynamics of religion. Method & Theory in the Study of Religion 14 (3/4), 293–315.
Whitehouse, Harvey, 2004. Modes of Religiosity: A Cognitive Theory of Religious Transmission. AltaMira Press, Walnut Creek.
Whitehouse, Harvey, Kahn, Ken, Hochberg, Michael E., Bryson, Joanna J., 2012. The role of simulations in theory construction for the social sciences: case studies concerning divergent modes of religiosity. Religion, Brain & Behavior 2 (3), 182–224.
Wilson, Edward O., 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge.
Wilson, David Sloan, 2008. Evolution and religion: the transformation of the obvious. In: Bulbulia, Joseph, Sosis, Richard, Harris, Erica, Genet, Russell, Genet, Cheryl, Wyman, Karen (Eds.), The Evolution of Religion: Studies, Theories, & Critiques. Collins Foundation Press, Santa Margarita, pp. 23–29.
Xygalatas, Dimitris, Mitkidis, Panagiotis, Fischer, Ronald, Reddish, Paul, Skewes, Joshua, Geertz, Armin W., Roepstorff, Andreas, Bulbulia, Joseph, 2013a. Extreme rituals promote prosociality. Psychological Science prepub 1–4.
Xygalatas, Dimitris, Schjoedt, Uffe, Bulbulia, Joseph, Konvalinka, Ivana, Jegindø, Else-Marie, Reddish, Paul, Geertz, Armin W., Roepstorff, Andreas, 2013b. Autobiographical memory in a fire-walking ritual. Journal of Cognition and Culture 13 (1), 1–16.

Religious Fundamentalism: Cultural Concerns

Martin E Marty, University of Chicago, Chicago IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Religious fundamentalisms, reactions to the real or perceived threats to religion, have appeared in many nations, societies, and religions. Not merely individual responses to change, fundamentalisms appear in cultural forms, which means that they cohere as assemblages of beliefs and behavior patterns that become instruments for dealing with change and assaults. The content of each fundamentalistic response will be distinctive since it takes its character from existing but differing religions, but forms and practices appear to have some common characteristics across the boundaries of the religions.

'Fundamentalism' is the name applied to characterize a variety of modern religious movements internationally, but its use is contested in some cases. The key feature in these movements is their reactive character. That is, they are not simply traditional, classic, conservative, or orthodox. Scholars and others usually speak of fundamentalism only in cases that find sets of religious people responding to certain kinds of challenges. The challenges may vary, but they tend to be variations on what the reactors or observers ordinarily would call 'the modern,' 'modernity,' 'modernization,' or 'modernism.'

Fundamentalist Cultures

Fundamentalism, however defined, can be viewed and appraised as a deeply personal combination of beliefs and behaviors. Fundamentalism is ordinarily perceived as a social and cultural phenomenon, a religious presence or movement within particular denominations or religious traditions that attracts and inspires collectives of like-minded people who make up cultures or, more often, subcultures. To speak in these terms invites and demands some attention to what is meant by 'cultures'.

Philip Bagby, who gave much attention to comparative cultures, defined a culture as manifesting "regularities in the behaviour, internal and external, of the members of a society, excluding those regularities which are clearly hereditary in origin." (Philip Bagby, 1958. Culture and History: Prolegomena to the Comparative Study of Civilizations. University of California Press, Berkeley, p. 93.) Therefore 'culture' refers to elements in our experience. Thus one observes regularities among motorcyclists, who form a 'cyclist culture,' among addicted people in a 'drug culture,' among devotees of classical music who participate in a 'high culture' or who have their counterparts in 'mass culture.' Bagby goes on to define a culture also as a "distinctive assemblage of culture-traits and complexes regularly found together" (p. 93). Such distinctive assemblages can be visible, especially in the form of what people wear. Members of Amish culture clothe themselves quite differently than do the members of a 'country-club culture'. The culture traits and complexes found together will almost always be noted in distinctive languages – colloquial or academic and the likes – that separate the participants in a culture from others.

Observing such distinctive sets of traits and complexes is especially important in a phenomenon like fundamentalism. This is so because the boundaries between fundamentalist subcultures and people around them are on one hand very defined. Those who belong to them can discern almost instantly who 'belongs' and who is a stranger. Sometimes a label or a phrase, such as 'born again' among some sets of American Protestant fundamentalists, will serve as a signal. Orthodox Jewish cultures are very distinctive and easily discerned, while liberal or 'reform' Jewish cultures are less defined but their signals and traits are very important to participants as they define themselves.

The Term 'Fundamentalism'

One of the most controversial features of fundamentalist movements is the term itself. People in statecraft and scholars have reservations about the use of the term 'fundamentalist.' First, some question the accuracy and fairness of designating worldwide movements in many religions by using the name derived from one reactive movement in American Protestantism in the early twentieth century. Second, in the eyes of many, the transfer of this term to other religious contexts looks like an example of a western linguistic imperialism that needs to be resisted. Still others rightfully contend that features peculiar to a specific Protestant confiict differ in so many respects from those in other reactive movements that only confusion could result from the use of the term.

Another major reason why so many members of these cultures are reluctant to accept the term as their self-description or for scholarly purposes has to do with the fact that it can be used pejoratively and stigmatically. In a way, this resistance is ironic, since in the American context and, by analogy, in other religions and cultures, those who formed and originally adhered to such separatist and often militant movements originally *wanted* to be stigmatized, and still may want distinction, but on their own terms.

The American Protestant fundamentalists in the 1920s did not want to be confused with conservatives, regularly complaining that while the majority of their fellow-believers claimed to be conservative, they did not do battle for the divine cause. Because other Christians were too gentle and compromising, it was argued, a new term had to be coined and applied. This term would help provide boundaries around the movement and create distance between it and conventional conservatisms – or against anything else that implied a potential for compromise

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84051-9

and comfortable adaptations to other cultures. But because the strategy of setting boundaries and deliberately seeking to offend others was often carried to extremes, many became discontented with the fundamentalists' self-chosen name. Some sought moderating terms, and in the American case, 'neo evangelicalism' and then 'evangelicalism' served this purpose.

The term 'fundamentalism' will not be found in dictionaries or encyclopedias, even encyclopedias of religion in America, before the time of a fundamentalist– modernist controversy in the United States, one that came to a climax around 1925. When the word did make its way into reference books after that period, for example in German or French dictionaries, all the meanings of fundamentalism were reduced to something about a battle over the verbal inerrancy and literal interpretation of the Christian Bible – a reduction at odds with some identifiers and confusing to observers of Islamic, Jewish, or other religions.

Given these burdens associated with the term, there is good reason to ask about its continued usefulness and appropriateness. Several reasons are cited in defense, among them the following: first, the word has made its way into common speech. So widespread is the use among reporters, scholars, and politicians that it is not likely to disappear because of discontent over its use by many. Better to clarify what is meant, it is said, than to invent new coordinating terms that will meet similar resistance in due course. The second reason has to do with the need for some such coordinating designations. To mention that need is, of course, to open the door to the many controversies over comparative methods. James (1902) typically made a defense of the practice of typifying and classifying. Without classifying, it is hard to understand many features of the phenomena. At the same time, James warned against the danger of losing the individual and particular character of the objects studied.

In the case of fundamentalism, the interest in coordination is almost entirely formal, not substantive. That is, one notices aspects of certain religious movements by studying characteristic features while being well aware that these movements all differ from each other in respect to content. Indeed, a Muslim fundamentalist will be 'doctrinally' more distant from a Protestant fundamentalist than modernists in both religions would be to each other. So there is no point in using the word 'fundamentalist' to characterize the central teachings of these religious expressions.

Even here, however, the line is a bit fuzzy. While fundamentalists reject modernity, paradoxically many are adept at using modern media such as radio, television, film, and the Internet to hold their groups together, seek converts, and express themselves to outsiders. That observation grows out of formal comparison. But a second generalization – that all the fundamentalisms are not only patriarchal but that in their talk about God, they stress God's maleness while moderate and modernist movements play down masculine references to deity – does show that there can be certain kinds of substantive similarities.

A third reason for using the term, say its defenders, is that no clear alternative is apparent. Some reporters, trying to avoid the word, will use something like 'militants,' appropriate for one dimension of some of their lives. But these words do not signal anything of cultural context: one can be militant about revenge, the search for territory, or any number of themes that have nothing to do with religious reaction. Further defense of the term comes from those who say that all words travel, and globally people would make little sense of each other if they did not have terms that could be reappropriated in a variety of cultures. Thus words like 'nationalism,' 'liberalism,' 'colonialism,' 'conservatism,' 'radicalism,' 'democracy,' and 'republicanism' were born within western contexts and get translated for usage elsewhere. The People's Republic of China means something very different with the term 'republic' than do those who speak of the United States as a republic. Yet there are reasons for people to use the term in both contexts – and then to clarify what they mean and point to differences.

Attempts to replace the term usually are specific and cannot be used for coordinative or comparative purposes. Thus, one might speak of 'radical revolutionary nationalist Islamic reform,' using other borrowed terms, and then find that they do not relate to phenomena in the Catholic or Hindu world, so it could not serve as a coordinating term. Mention of such instances suggests why many choose to use 'fundamentalism': it has acquired some rather specific meanings, meanings that deserve examination.

The Originating Instance in America as a Case Study

American Protestantism, having been divided into many denominations from the beginning, experienced some trans-denominational conflicts in the nineteenth century. The first great breach occurred between the North and the South in the Civil War, when churches on each side inspired, monitored, provided morale for, and negatively defined the other side in religious terms. But all through the century another split was taking shape, thanks to latent tensions between theologically more liberal and more conservative parties.

These tensions increased with the arrival in the United States after the middle of the nineteenth century of Darwinian evolution, the higher criticism of the Bible, and various forms of political progressivism. Many conservatives vehemently resisted theories of evolution, in part because these were seen as assaults on the Bible. They all rejected biblical criticism as a threat to the foundations of the faith. And progressivism did not do justice to apocalyptic messages in the Bible. Meanwhile, modernist parties began to develop toward the turn of the twentieth century. These were strongest at the prestigious theological schools such as Harvard, Yale, and Chicago – but rather dramatically not at Princeton University and Princeton Theological Seminary, from which some intellectual impetus for fundamentalism was to come.

Between 1910 and 1915, the conservative parties issued an influential set of booklets called *The Fundamentals*. From then on, the accent on 'fundamentals' grew. In 1919, a World's Christian Fundamentals Association was born. In 1920, during a battle for power in the Northern Baptist Convention, a Baptist editor proudly chose and publicized the term 'fundamentalist' for his faction. The Presbyterians, and to a lesser extent other mainstream Protestant denominations, entered into the battle. The celebrated Scopes trial in Tennessee in 1925, one that featured conflict over the teaching of evolution in public schools, was the best-known event in the formation and publicizing of fundamentalism.

This is not the place to trace in detail the outcomes of the struggles. Fundamentalists lost the battles for control, but

inventively went their way, developing their own Bible colleges, seminaries, mission boards, and, thanks to schisms, denominations. They were a presence but not a vivid one during this period. In 1941, some among them formed an American Council of Christian Churches, which engaged in rivalry with the more moderate National Association of Evangelicals born in 1942. Still regarded by nonfundamentalists as marginal, eccentric, and withdrawn, fundamentalists remained generally aloof from politics. However, during the 1970s most of them reversed their stance and began to adopt political interventionist postures. They have been a prominent presence ever since.

To detail this case study further might well bias the case of those who charge that global and interreligious use of the term 'fundamentalism' is too reliant on the American experience. But one could study cognate movements elsewhere – such as the Muslim Brotherhood in Egypt – and find similar reasons for Islamic conservative groups there to move into reaction. A description of elements that make up what is called 'fundamentalism' in any religion will be more profitable, however, than would extending the number of case studies.

A Phenomenological Approach to Fundamentalist Features

By 'phenomenological' here is meant a pattern of observation of cultures in which the analysts 'bracket' their presuppositions – in this case, for example, their assumptions about the positive or negative aspects of fundamentalism – and to ask: 'What stands out in the phenomena that have been called fundamentalist?' Another way to put it is: what phenomena developed in the twentieth century that occasioned the need for fresh naming – since words like 'conservative' and 'traditional' had been easily available? These questions enable one to escape the charge that a priori definitions color the observations. These are some of those features. First, while fundamentalisms originally were dismissed by others as 'the old-time religion' and while some parade themselves as traditional, they differ in significant respects from conservatisms. In the eyes of doctrinal, political, and sometimes even armed militant partisans, conservatives may be too passive. They are too content to let people in the world go about their business, without reckoning with them. If in any particular instance the cause being challenged is seen as divine – and we are here dealing only with religious fundamentalisms – for anyone to be content with tradition is a betrayal.

Fundamentalisms rise when historically conservative religious movements experience something that challenges them to their core. In Iran and elsewhere in the Middle East after the late 1970s, the challenger might be 'the Great Satan' of western imperialism, specifically, in the eyes of its enemy, in its American form. In hyper-Orthodox Israel, the threat to Jewish identity itself appears to parties such as Gush Emunim to derive from nonobservant, compromising Jews who might even let some of the land of Israel slip away in diplomatic negotiations. In the United States, the challenge in the 1920s came in more doctrinal ways, through contests over biblical literalism. But in all these cases, the threat had to be seen as total; fundamentalists do not bother much about trivial issues.

Because the challenges are encompassing, the response of fundamentalist movements has to be drastic. Fundamentalisms characteristically are reactive. They rise among communities that have been acted upon, and they act back. Very often they use the very instruments that had presented the challenge. If a nation–state is the ominous force, fundamentalists make efforts to win power, through the ballot if possible and through weaponry if necessary. Let it be noted that not all fundamentalisms take on any form of military or terrorist guise. Some even shun the political arena if they feel that they can convert sufficient recruits from that worldly zone or win their way simply through cultural militancy, as with disputes over 'values' and media. Again, fundamentalists tend to use some of the weapons that have brought change whenever, for example, they take instruments of mass media, the agents that penetrated their homes and consciousness, and become sophisticated in using these to fight back.

Somehow, in every case, they react. Not to be reactive in these contexts is to be not fundamentalist. George Marsden, the foremost historian of the American version of fundamentalism, describes a fundamentalist as 'an evangelical who got mad.' In reaction, the fundamentalist seeks ammunition for defense and aggression. Here is where the fundamentals come in. Fundamentalists reach back to foundations, some in the Qur'an, some in the Torah, some in the New Testament, still others in classic events of the past, and claim them for their own. In every case in their eyes there had to be some sort of golden moment or precious book to which arguing parties can resort to in efforts to determine winners and losers.

Acting upon these fundamentals, members of the group then engage, as noted, in setting boundaries. They are often described as bearing a Manichaean cast. This means that the cosmos is divided into God versus anti-God, good versus evil, our people versus everyone else. They may not necessarily withdraw into enclaves, but they will tend to use code words, slogans, and informal testing devices that help them screen out the most dangerous persons and forces, namely the moderates. The total outsider, the infidel, the Great Satan, is useful for mobilizing forces, but is not a threat in the way a compromising modernizer is seen to be.

With the setting of boundaries comes the claim that participants are an elect people, God's own instrument. They know something of where history is going; it has been revealed to them by sacred books in which they find the fundamentals. The elect people, called to represent God, must keep themselves pure, try to recruit others, and, to use the words from the American struggle in the 1920s, 'do battle for the Lord.' It is easier to be fundamentalist in 'religions of the book' such as Judaism, Christianity, and Islam, than it is in faiths such as Hinduism and Buddhism, where no single divine scriptural revelation is foundational. The battles may vary in character. Thus in most cases in the Islamic world the enemy is not external, but instead a regime seen as corrupt, secularizing, and compromising even though all its leaders are themselves practicing Muslims. In Israel, the parties called fundamentalist may engage in warfare against Arab Muslims, but they take the most pains in differentiating themselves from compromising Jews. Action may be hortatory, prophetic, political, military, or terrorist, though it is often only the latter two that draw the world's attention.

Accounting for Fundamentalisms in Culture

No scholar would claim that modern fundamentalisms have no counterparts or analogs in the past within the various religions. All of them have at times manifested forces of traditionalist reaction and conservative aggression. It is valuable to study these analogs. But the word 'fundamentalism' has more appropriately been reserved for responses to what has been seen as modern. One way to put it: these are reactive movements that appeared after the modern academy thought there would be no more like it, that the religions that survived would be small, weak, and always compromising.

In recent years, one of the most focused endeavors to understand fundamentalisms has been in the hands of psychologists, especially social psychologists. They have on occasion assessed persons in the movements or the movements as cultures, and generalized about their mentality. Participants, they note, tend to be people who cannot tolerate ambiguity, paradox, contradiction, or doctrinal and moral 'fuzziness.' There is evidence for this. Yet many scholars caution against what can be superficial assessments. It is, they observe, difficult to reduce fundamentalisms to this psychological cast, given the fact that people with such predilections in one culture turn religious fundamentalist and those in others do not. So to note the psychology of fundamentalists is a valuable bur insufficient measure unless it is directed to the peculiarities of cultures and their historic experiences. To define fundamentalisms as intense literalist devotion to sacred texts such as the Qur'an, the Torah, or the Christian Bible will be of little help in understanding fundamentalist-like movements, movements bearing family resemblances to fundamentalism in, say Buddhist societies and cultures whose religions are less determined by texts than are those characterized as 'people of the book,' such as Judaism, Islam, and Christianity.

The most helpful aspect of psychological studies of fundamentalist cultures is to try to determine what moves people to leave behind other religious preferences and to 'turn fundamentalist.' Numbers of studies focused on how religious texts are put to work or, to turn this around, how psychological interests turn people to texts to reinforce their searches and to help discredit others. Political and social scientists perceiving fundamentalisms in culture often relate the rise of such movements to class conflict. Marxist interpretations, for example, can see fundamentalisms as futile resistances by people who take resort in otherworldly solace when they can do little to gain in this world. It is true that in many instances it is the poor and apparently powerless who do turn to fundamentalism. But it is also true that those who react may do so from the comfortable confines of a Princeton Theological Seminary and Princeton University in the 1920s, as did the intellectual definers of early American fundamentalism. And one will find fundamentalists of wealth in many of the American cities and note that some among them extravagantly fund fundamentalist projects. Class analysis does contribute to understanding the phenomena, but most scholars again resist reductionism of a sort that fundamentalists are 'nothing but' this or that. Instead, they give evidence that fundamentalists experience complex responses and reactions to many kinds of cultural stimuli, and as such manifest features that are best appraised with a variety of perspectives and disciplines. Among these would be humanistic efforts to bring to bear history, philosophy, and religious studies.

One way to approach the study is to listen as anthropologists do for 'agent's description.' Someone who pays attention to, records, and coordinates the testimony of fundamentalists in many religious contexts will find that the issues that evoke response begin very close to the self of the believer. Diplomats, reporters, and scholars often begin at the furthest extensions of fundamentalisms, as when some engage in terrorism or revolutionary military action.

Fundamentalist believers begin by defending a worldview, including one that has been recently changed, thanks to conversion. The world looks different to someone who is an initiate, 'born again' in some versions, than it does to someone else who is distant from or quiet about religious impulses. Very close to that egocenter is the question of personal, social, and cultural identity. To whom do I belong? Whom can I trust? With whom can I check signals and from whom do I draw inspiration? Those who ask such questions in the erosive climate and culture of late modernity find that identity becomes a major question. They also note that it is threatened by forces such as travel, modern mass communications, and mass higher education, especially where the challenges of pluralism and relativism are fierce.

As close to the center of the person as are worldview and identity is the issue of gender. Fundamentalists preoccupy themselves with what it means to be a man or a woman. The Manichaean differentiation is strong: there is no room for ambiguity in defining gender. The fundamentalisms, for example, all reject homosexual expression. From gender they move to defense of family. In all the cultures studied by fundamentalist scholars, 'family values' and shoring up the family, however defined, are central to the program.

From the orbits of gender and family one moves to other challenges. Thus, one's mind and habits are shaped by mass communications, and fundamentalists tend to react by trying to screen out uncongenial images, to engage in boycotts, or attempt legally to censor and exclude broadcast or published materials that they believe to be evil or subversive of their values.

Education is another zone in which fundamentalists do battle. They start their own schools to protect and nurture their own young and often engage in efforts to subvert the character of others' schools, especially the public's schools, when their teaching and habits are offensive. Similarly, many of them have challenged modern medical practices, and make the clinic their arena of conflict. They organize against abortion, euthanasia, and many forms of treatment that violate their view of the body as God's temple.

In the sphere of economics it is harder to generalize. Some religious texts favored by fundamentalists – notably those detailing *shari'ah*, or Islamic law – may be explicit as when they proscribe the taking of interest. But some fundamentalist movements promote communal, one might say communistic, economic patterns, while others, notably in North America, often align themselves with *laissez-faire* economic forces. But when they perceive economic circumstances as a threat to them, they may adopt any number of strategies, and find their holy books less defining and definitive in setting boundaries and distancing others.

When it comes to efforts to change politics, for example, it is through the use of the instruments of political propaganda,

the ballot, and constitutional revision that fundamentalists find themselves reckoned with most strenuously. Where they cannot take over a regime, they may in God's name undertake revolutionary activity. It is activity of that sort that elicits headlines about fundamentalist militants around the world.

Curiosity about the work of historians and social or cultural analysts of fundamentalisms might have been confined to theology schools, strategists in the institutional life of religious organizations, or curators of religious museums or libraries. However, a turn that modern fundamentalisms took in the last decades of the twentieth century changed all that. Frustrated or ambitious fundamentalists who claimed that they knew that God was on their side, lost patience waiting for victory and took history into their own hands. Although Hindu fundamentalists came to power in a variety of parties and movements around 1925 in India and Muslims in Egypt formed a Brotherhood in 1928, there as elsewhere the political stirrings did not evolve into military action, until the Post-Cold War era, in a world often described as 'secular,' when some fundamentalists turned militant, as in the Iranian Revolution in 1979–80. Similarly extreme fundamentalist movements developed in Saudi Arabia and elsewhere in the Middle East. Most notable was the formation of al-Qaeda, which became the best-known terrorist organization, and which waged subversive activities and war in the name of Allah. What began as a search for meaning became an assertion of power.

Although many fundamentalist movements began among people who wanted to be distanced from others, including from the religious movements from which they sprang or with which they might be confused, in a drastic turn observable globally, numbers of these changed completely. Where fundamentalists had once been otherworldly, now they wanted to take over portions of the world. Fundamentalists had separated from the state, but in a later version they wanted to take over the state. Whereas they had once tried to hide from 'the world,' now they aspired to transform the world in images the fundamentalists chose. Observers and leaders of state and the economy frequently ask if the fundamentalists will win. That is impossible to answer, but the fact that some aggressive fundamentalist movements have experienced setbacks in cultural battles indicates that they are not all-powerful. Many other factions among them, in the course of time, grew weary or chose to compromise and adapt to elements in the surrounding cultures. In every case, however, their very presence would have been a surprise to people who, a century before they appeared, had not envisioned that in a world of technology and devotion to reason, such intense and dedicated religions might arise, to put on the defensive many secular and moderate movements. The key test in all cases in the future will be: does the movement possess sustaining religious energies, or is it almost entirely a reaction to modernity with its many faces and challenges?

See also: Islamic Fundamentalism.

Bibliography

Almond, G., Appleby, R., Sivan, E., 2003. Strong Religion: The Rise of Fundamentalisms around the World. University of Chicago Press, Chicago.

Ammerman, N.T., 1987. Bible Believers. Rutgers University Press, New Brunswick, NJ.

Appleby, R., 2000. The Ambivalence of the Sacred. Rowman and Littlefield, London.

Arjomand, S.A. (Ed.), 1984. From Nationalism to Revolutionary Islam. State University of New York Press, Albany, NY.

Arjomand, S.A., 1988. The Turban for the Crown: The Islamic Revolution in Iran. Oxford University Press, New York.

Armstrong, K., 2000. The Battle for God. Knopf, New York.

Bagby, P., 1958. Culture and History: Prolegomena to the Comparative Study of Civilizations. University of California Press, Berkeley.

Boone, K., 1989. The Bible Tells Them So: The Discourse of Protestantism Fundamentalism. State University of New York Press, Albany.

Bruce, S., 2000. Fundamentalism. Polity Press, Cambridge.

Caplan, L. (Ed.), 1987. Studies in Religious Fundamentalism. State University of New York Press, Albany, NY.

Cohen, N.J. (Ed.), 1990. The Fundamentalist Phenomenon: A View from within; A Response from without. Eerdmans, Grand Rapids, MI.

Dekmejian, R.H., 1985. Islam in Revolution: Fundamentalism in the Arab World. Syracuse University Press, Syracuse, NY.

Dessouri, A.E.H. (Ed.), 1982. Islamic Resurgence in the Arab World. Prager, New York.

Emmons, R., 1999. The Psychology of Ultimate Concern: Motivation and Spirituality in Personality. Guilford Press, New York.

Esposito, J.L. (Ed.), 1983. Voices of Resurgent Islam. Oxford University Press, New York.

Friedman, M., 1992. The Haredi (Ultra-Orthodox) Society: Sources, Trends and Processes. Jerusalem Institute for Israel Studies, Jerusalem.

Heilman, S., 1992. Defenders of the Faith: Inside Ultra-Orthodox Jewry. Schocken, New York.

Hood, R., Hill, P., Williamson, W. (Eds.), 2005. The Psychology of Religious Fundamentalism. Guilford, London.

James, W., 1902. The Varieties of Religious Experience. Longmans, Green, New York.

Juergensmeyer, M., 1993. The New Cold War? Religious Nationalism Confronts the Secular State. University of California Press, Berkeley, CA.

Kepel, G., 1985. Muslim Extremism in Egypt: The Prophet and the Patriarch. University of California, Berkeley, CA.

Lawrence, B., 1989. Defenders of God: The Fundamentalist Revolt against the Modern Age. Harper & Row, San Francisco.

Lefebure, L., 2000. Revelation, the Religions, and Violence. Orbis, Markynoll, NY.

Lustick, I.S., 1988. For the Land and the Lord: Jewish Fundamentalism in Israel. Council on Foreign Relations, New York.

Marsden, G.M., 1980. Fundamentalism and American Culture. Oxford University Press, New York.

Martin, D., 1990. Tongues of Fire: The Explosion of Protestantism in Latin America. Blackwell, Oxford, UK.

Marty, M.E., Appleby, R.S., 1991. Fundamentalisms Observed. University of Chicago Press, Chicago.

Marty, M.E., Appleby, R.S., 1993a. Fundamentalisms and the State: Remaking Polities, Economies, and Militance. University of Chicago Press, Chicago.

Marty, M.E., Appleby, R.S., 1993b. Fundamentalisms and Society: Reclaiming the Sciences, the Family, and Education. University of Chicago Press, Chicago.

Marty, M.E., Appleby, R.S., 1994. Accounting for Fundamentalisms: The Dynamic Character of Movements. University of Chicago Press, Chicago.

Marty, M.E., Appleby, R.S., 1995. Fundamentalisms Comprehended. University of Chicago Press, Chicago.

Mitchell, R.P., 1969. The Society of Muslim Brothers. Oxford University Press, London.

Piscatori, J., 1991. Islamic Fundamentalisms and the Gulf Crisis. American Academy of Arts and Sciences, Chicago.

Rahman, F., 1982. Islam and Modernity: Transformation of an Intellectual Tradition. University of Chicago Press, Chicago.

Sivan, E., 1985. Radical Islam. Yale University Press, New Haven, CT.

Strozier, C., 1994. Apocalypse: On the Psychology of Fundamentalism in America. Beacon, Boston.

Strozier, C., Terman, D., Jones, J. (Eds.), 2005. The Fundamentalist Mindset. Oxford University Press, New York.

Tibi, B., 1988. The Crisis of Modern Islam. University of Utah Press, Salt Lake City, UT.

Ward, K., 2007. Is Religion Dangerous? Eerdmans, Grand Rapids, MI.

Religious Nationalism and the Secular State

Mark Juergensmeyer, University of California, Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T.N. Madan, volume 19, pp. 13123–13127, © 2001, Elsevier Ltd.

Abstract

In the global era of the post–Cold War world, secular nationalism is in peril. As demographic shifts, transnational economic networks, and transnational forms of communication have undermined traditional notions of nationalism, local leaders have searched for new anchors to ground their social identities and political loyalties. Many have turned to ethnicity and religion. Although many of the framers of the new nationalisms have reached back in history for ancient images and concepts that will give them credibility, theirs are not efforts simply to resuscitate old ideas, these are contemporary ideologies that meet present-day social and political needs. In the modern context the notion that indigenous culture can provide the basis for new political institutions, including resuscitated forms of the nation-state, is revolutionary. And for this reason, the encounters between religious nationalism and the secular state are sometimes violent.

The late twentieth century and the first decade of the twenty-first was a remarkable era of strident religious protest around the world. Beginning with the Ayatollah Khomeini's Islamic revolution in Iran in 1979 and the Khalistani movement of Sikhs in northern India in the 1980s, movements of religious politics asserted themselves in all major religious traditions: Jewish militants in Israel, Hindu nationalists in India, Christian militia in the United States, Buddhist extremists in Sri Lanka, and Muslim activists in the Middle East. In some cases, such as the jihadi terrorists associated with al-Qaida, the goal was transnational. In most other instances of religious politics, including the Iranian and Sikh cases, the goal was a new form of religious nationalism. Both rejected the secular nationalism that had been the central feature of European Enlightenment since the eighteenth century.

The new religious movements were largely reactions to the worldwide spread of secular modernity. They were responses to the insufficiencies of what is often touted as the world's global political standard: the secular, Westernized constructs of nationalism that are found not only in the West but also in many parts of the former Third World as vestiges of European colonialism. These secular nationalisms, weakened by globalization, have been under siege. Their vulnerability has enabled new ethnic and religious politics to step into the breach and shore up national identities in their own distinctive ways. Yet these identities have often had international and transnational aspects of their own. Thus some forms of ethnic and religious politics have been global, some virulently antiglobal, and still others stridently nationalistic. Yet all have preyed upon the weakened state of secular nationalism in the present period of late modernity.

Religious Challenges to the Secular State

The idea of the modern nation-state, the political legacy of the European Enlightenment, is at once profound and simple: the state is created by the people within a given national territory. Secular nationalism – the ideology that originally gave the nation-state its legitimacy – contends that the authority of a nation is based on the secular idea of a social compact of equals rather than on ethnic ties or sacred mandates. It is a compelling idea, one with claims to apply universally. It reached its widest extent of worldwide acceptance in the first half of the twentieth century with the creation of scores of new independent nations at the end of the era of European colonialism.

But the second half of the century turned out to be a different story. The secular nation-state proved to be a fragile artifice, especially in those many areas of the world where nations had been created by retreating colonial powers – in Africa by Britain, Portugal, Belgium, and France; in Latin America by Spain and Portugal; in South and Southeast Asia by Britain, France, the Netherlands, and the United States; and in Eurasia by the Soviet Union. In some cases boundary disputes led to conflicts between neighboring nations. In others the very idea of the nation caused suspicion.

Many of these 'imagined nations,' as Benedict Anderson (1983) has called them – some with invented names such as Pakistan, Indonesia, and Yugoslavia – were not accepted by all within their territory. In other cases the tasks of administration became difficult to perform. In many cases the newly created nations had only brief histories of prior colonial control to unite them, and after independence they had only the most modest of economic, administrative, and cultural infrastructures to hold the disparate regions within them together.

By the 1990s these ties had begun to fray. The global economic market undercut national economies, and the awesome military technology of the United States and NATO reduced national armies to border patrols. More significant, the rationale for the nation-state came into question. With the collapse of the Soviet Union and the postcolonial, post–Vietnam critique of Western democracy, the secular basis for the nation-state seemed increasingly open to criticism. In cases such as Yugoslavia, when the ideological glue of secular nationalism began to dissolve, the state fell apart.

The effect of what I have elsewhere called 'the loss of faith in secular nationalism' was devastating (Juergensmeyer, 2008). It seemed that nationalism was being challenged everywhere, and the scholarly community joined in the task of trying to

understand the concept in a post–Cold War and transnational era (Kotkin, 1994; Smith, 1995). Part of the reason for the shaky status of nationalism was that it was transported to parts of the world in the cultural baggage of what Jürgen Habermas (1997) has called 'the unfinished project of modernity' – a trust in reason that was by now considered obsolete. In a multicultural world, where a variety of views of modernity are in competition, the very concept of a universal model of secular nationalism became highly debatable.

Globalization has challenged the modern idea of nationalism in a variety of ways. These challenges have been varied because globalization is multifaceted. The very term 'globalization' refers not to any one thing but to a series of processes. It embraces not only the global reach of transnational businesses but also their labor supply, currency, and financial instruments. In a broader sense globalization also refers to the planetary expansion of media and communications technology, popular culture, and environmental concerns. Ultimately, it also includes a sense of global citizenship and a commitment to world order.

Globalization has affected societies in differing ways. Some countries that have been brought into contact with economic globalization – by supplying labor for the commodity chains of globalized production – have not experienced the globalization of culture and citizenship. The advent of economic globalization has threatened local identities in such a way as to encourage the protection of local cultures and social identities, sometimes in hostile and defensive ways. This effect has been observed by Benjamin Barber, who notes that the 'McWorld' of contemporary Westernized culture has assaulted the public consciousness in various parts of the world. In extreme forms the reaction has been the 'jihad' of militant tribalism (1995).

Some of the most intense movements of ethnic and religious nationalism have arisen in nations which have felt themselves exploited by the global economy, such as in Iran and Egypt, or else have believed that somehow the benefits of economic globalization are passing them by (Juergensmeyer, 2003, 2008). The global shifts in economic and political power that occurred following the breakup of the Soviet Union and the sudden rise and subsequent fall of Japanese and other Asian economies in the last decade of the twentieth century have also had significant social repercussions. The public sense of insecurity that came in the wake of these changes was felt especially in areas economically devastated by the changes, including those nations that had been under the dominance of the Soviet Union.

These shifts led to a crisis of national purpose in less developed nations as well. Leaders such as India's Jawaharlal Nehru, Egypt's Gamal Abdel Nasser, and Iran's Riza Shah Pahlavi had tried to create their own versions of America – or else a cross between America and the Soviet Union. But a new, postcolonial generation no longer believed in the Westernized vision of Nehru, Nasser, or the Shah. Rather, it wanted to complete the process of decolonialization by asserting the legitimacy of their countries' own traditional values in the public sphere and by constructing a national identity based on indigenous culture. This eagerness became all the more keen with the global media assault of Western music, videos, and films – an assault that has threatened to obliterate local and traditional forms of cultural expression.

In other cases it was a different kind of globalization – the emergence of multicultural societies through global migrations and the suggestion of global military and political control in a 'new world order' – that caused fear. Perhaps surprisingly, this response was most intense in the high-income countries, which in other ways seemed to be the paradigm of globalization. In the United States, for example, the Christian Identity movement and militia organizations were fueled by fears of a massive global conspiracy of liberal American politicians and the United Nations. In Japan a similar conspiracy theory motivated leaders of the Aum Shinrikyo movement to predict a catastrophic World War III, which their nerve gas assault in the Tokyo subways was meant to emulate.

As farfetched as the idea of a 'new world order' of global control may be, there is some truth to the notion that the globalization of culture has brought the world closer together. Although it is unlikely that a cartel of malicious schemers plotted this global trend, its effect on local societies and national identities has nonetheless been profound. Globalization has undermined the modern idea of the nation-state by providing nonnational and transnational forms of economic, social, and cultural interaction. The global economic and social ties of the inhabitants of contemporary global cities are linked together in a way that supersedes the Enlightenment notion that peoples in particular regions are naturally linked together by a social contract. In a global world it is hard to say where one region ends and another begins. In fact, it is hard to say how the 'people' of a particular nation should even be defined.

Religion and ethnicity have stepped in to redefine public communities. The fading of the nation-state and of old forms of secular nationalism has produced both the opportunity and the need for new nationalisms. The opportunity has arisen because the old orders have become so weak. The need for national identity persists because no single alternative form of social cohesion and affiliation has yet appeared to dominate public life, the way the secular nation-state did in the twentieth century. In a curious way traditional forms of social identity have helped to rescue the idea of national societies. In the continuing absence of any other demarcation of national loyalty and commitment, these old staples – religion, ethnicity, and traditional culture – have become resources for national identification.

Nationalism Defined by Religion

Today religious rather than secular nationalism has provided a solution to the problem of Western-style secular politics in a non-Western and multicultural world. As secular ties have begun to unravel in the post–Soviet and postcolonial era, local leaders have searched for new anchors to ground their social identities and political loyalties. Many have turned to ethnicity and religion. What is ideologically significant about these ethnic and religious movements is their creativity. Although many of the framers of the new nationalisms have reached back in history for ancient images and concepts that will give them credibility, theirs are not efforts simply to resuscitate old ideas. These are contemporary ideologies that meet present-day social and political needs.

In the modern context the notion that indigenous culture can provide the basis for new political institutions, including resuscitated forms of the nation-state, is revolutionary. Movements that have supported ethnic and religious nationalism have therefore often been confrontational and violent. They have rejected the intervention of outsiders and at the risk of being intolerant have pandered to indigenous cultural bases and have enforced traditional social boundaries. It is no surprise, then, that they have gotten into trouble with one another and with defenders of the secular state. Yet even these conflicts with secular modernity have served a purpose for the movements. They have helped define who they were as people and who they were not – for example, not secularists.

Since secularism has often been targeted as the enemy, that enemy has most easily been symbolized by things American. The United States has taken the brunt of religious and ethnic terrorist attacks in recent years, in part because it so aptly symbolizes the transnational secularism that the religious and ethnic nationalists loathe and in part because the United States does indeed promote transnational and secular values. For instance, the United States has a vested economic and political interest in shoring up the stability of certain regimes around the world. This interest often puts the United States in the position of defending secular governments. Moreover, the United States supports a globalized economy and a modern culture. In a world where villagers in remote corners of the world increasingly have access to MTV, Hollywood movies, and the Internet, the images and values that have been projected globally have been American.

It is thus understandable that since the United States is seen by many in traditionally religious countries as promoting a global secular agenda it is often disdained by. The demonization of the United States by many ethnic and religious groups fits into a process of delegitimizing secular authority. In order to delegitimize the public authority, the religious groups have had to shore up their own pillars of authority. In doing so, they have appropriated traditional religious images, especially that of cosmic war. In this scenario, competing religious groups are seen as foes and scapegoats, and the secular state is viewed as the enemy of religion. This 'satanization' is aimed at reducing the power of opponents by discrediting them. By humiliating them – by making them subhuman – religious groups assert the superiority of their own moral power.

During the early days of the Persian Gulf War in 1991, the Hamas movement issued a communique stating that the United States 'commands all the forces hostile to Islam and the Muslims' and singled out then-President George Bush as not only 'the leader of the forces of evil' but also 'the chief of the false gods.' As late as 1997, Iranian politicians, without a trace of exaggeration, described America as the 'Great Satan.' This rhetoric first surfaced in Iran during the early stages of the Islamic revolution, when both the Shah and President Jimmy Carter were referred to as Yazid ('agent of Satan'). 'All the problems of Iran,' the Ayatollah Khomeini elaborated, are "the work of America" (Khomeyni, 1977: p. 3). He meant not only political and economic problems but also cultural and intellectual ones, fostered by "the preachers they planted in the religious teaching institutions, the agents they employed in the universities, government educational institutions, and publishing houses, and the Orientalists who work in the service of the imperialist states" (Khomeini, 1985: p. 28).

Religious Nationalism in the Global Era

When members of the extreme religious right in the United States recoil in horror over what they call 'the New World Order,' what they and members of many other radical religious groups around the world fear is not just globalization, but also the secular aspects of globalization. They are afraid that global economic forces and cultural values will undercut the legitimacy of their own bases of identity and power. By contrast, other aspects of globalization – technical and economic – are often perceived as neutral and even useful. Osama bin Laden exploited to spread his messages, and members of the al-Qaida network have at times relied on encrypted Web sites for communication.

Some groups have a global agenda of their own, a transnational alternative to political nationalism. Increasingly, terrorist wars have been waged on an international and transnational scale. When the World Trade Center was demolished in the dramatic aerial assaults of 11 September 2001, it was not just the United States that was targeted but also the power of the global economic system that the buildings symbolized. Osama bin Laden's al-Qaida network was itself a global structure. The Gamaa i-Islamiya group, related to al-Qaida, literally moved its war against secular powers abroad when its leader, Sheik Omar Abdul Rahman, moved from Egypt to Sudan to Afghanistan to New Jersey. It was from the Jersey City location that his followers organized the 1993 bombing attack on the World Trade Center that killed six and injured a thousand. Osama bin Laden's operative, Ramzi Youssef, also convicted of complicity in the 1993 World Trade Center bombing, masterminded the 'Bojinka Plot' that would have destroyed a dozen US commercial planes over the Pacific during the mid-1990s. Youssef moved from place to place throughout the world, including Pakistan and the Philippines. Algerian Muslim activists brought their war against secular Algerian leaders to Paris, where they have been implicated in a series of subway bombings in 1995. Hassan Turabi in Sudan has been accused of orchestrating Islamic rebellions in a variety of countries, linking Islamic activists in common cause against what is seen as the great satanic power of the secular West. Osama bin Laden, from his encampment in Afghanistan, was alleged to have ordered many acts of terrorism around the world.

These worldwide attacks may be seen as skirmishes in a new Cold War or, more apocalyptically, a 'clash of civilizations,' as Samuel Huntington has termed it (see Huntington, 1996). It is possible to imagine this clash if one assumes that Islam and other religions are civilizations comparable with the modern West, or if one regards secular nationalism as, in the words of one of the leaders of the Iranian revolution, 'a kind of religion.' Those religious opponents of secular nationalism who regard it as religious often describe it as a religion that is peculiar to the West, a point echoed by leaders of the Muslim Brotherhood in Egypt.

Behind this image of a clash of cultures and civilizations is a vision of social reality, one that involves a series of concentric

circles. The smallest are families and clans; then come ethnic groups and nations; the largest and most important are religions. Religions here are not just bodies of doctrine and communities of believers but shared worldviews that span great expanses of time and space. They are global civilizations. Among these are Islam, Buddhism, and what some who hold this view call 'Christendom,' 'Western civilization,' or 'Westernism.' The so-called secular culture of places such as Germany, France, and the United States stand as subsets of Christendom/Western civilization. Similarly, Egypt, Iran, Pakistan, and other nations are subsets of Islamic civilization. From this vantage point it is both a theological and a political error to suggest that Egypt or Iran should be thrust into a Western frame of reference. By this view of the world they are intrinsically part of Islamic, not Western, civilization, and it is an act of imperialism to think of them otherwise. Proponents of Islamic nationalism often see themselves as a part of a larger, global encounter among Western, Islamic, and yet other cultures. This clash of civilizations view is not confined to the imaginations of either Samuel Huntington or a small number of Islamic extremists but has animated much of the political unrest at the dawn of the twenty-first century.

An even more extreme version of this global cultural clash has been an apocalyptic one, in which contemporary politics has been seen as fulfilling an extraordinary religious vision. Some Messianic Jews, for instance, think the Kingdom that will arise with the coming of the Messiah is close at hand. It will occur when the biblical lands of the West Bank are returned to complete Jewish control and when the Jerusalem Temple described in the Bible is restored on its original site, one presently occupied by the Muslim Dome of the Rock. Several Jewish activists have been implicated in plots to blow up the shrine in order to hasten the coming of the Kingdom. One who served time in prison for his part in such a plot said that the rebuilding of the Temple was not just a national obligation but an obligation for redemption of the world.

Religious activists such as Millennarian Christians and Shi'ite Muslims, who have a strong sense of the historical fulfillment of prophecy, also look to a religious apocalypse that will usher in a new age. American Christian political activists such as Pat Robertson and the late Jerry Falwell are animated by the idea that the political agenda of a righteous America will help to usher in an era of global redemption. The leader of Aum Shinrikyo, borrowing Christian ideas from the sixteenth-century French astrologer Nostradamus, proclaimed the coming of Armageddon in 1999. Those who survived this World War III – mostly members of his own movement – would create a new society in 2014, one led by Aum-trained 'saints.'

Activists in other religious traditions have seen a righteous society established in a less dramatic manner. Some Sunni Muslims, Hindus, and Buddhists have in their own ways articulated hopes for a political fulfillment of a religious society. They believe that "dhammic society can be established on earth," as one activist Buddhist monk in Sri Lanka puts it, by creating a religious state. These forms of religious politics are more than nationalist because they envision the world as caught up in a cosmic confrontation, albeit one that will ultimately lead to a peaceful world order constructed by religious nations. The result of this process will be a global order radically different from secular versions of globalization, but it will still be an ideological confrontation on a global scale.

The Future of Religious Nationalism

The goal of other religious activists is the revival of a nation-state that avoids the effects of globalization. Where new religious states have emerged, they have tended to be isolationist. In Iran, for instance, the ideology of Islamic nationalism that emerged during and after the 1979 revolution from the Ayatollah Khomeini and his political theoretician, Ali Shari'ati, was intensely parochial. It was not until some 20 years later that new movements of moderate Islamic politics encouraged its leaders to move out of their self-imposed international isolation. The religious politics of Afghanistan, especially after Taliban militants seized control in 1995, were even more strongly isolationist. Led by members of the Pathan ethnic community who were former students of Islamic schools, the religious revolutionaries of the Taliban established a self-contained autocratic state with strict adherence to traditional Islamic codes of behavior. Only after the collapse of the Taliban in 2001, following the al-Qaida terrorist attacks, did Afghanistan become more open to the wider world.

Other movements of religious nationalism have not been quite so isolationist. In India, when Hindu nationalists in the Bharatiya Janata Party (BJP), or 'Indian People's Party,' came to power in 1998 – a victory that was consolidated in the national elections of 1999 – some observers feared that India would become isolated from world opinion. The testing of nuclear weapons as one of the BJP's first acts in power did little to dispel these apprehensions. But in many other ways, including its openness to economic ties and international relations, the BJP has maintained India's active role in the world community. Credit for this openness may go in part to the moderate leadership of the former BJP Prime Minister, Atal Bihari Vajpayee, one of the country's most experienced and temperate politicians.

It is an open question whether movements of religious and ethnic nationalism elsewhere in the world, if they ever come into power, will behave like the Taliban or like the BJP. The movements in Pakistan, Egypt, and Algeria could go either way. When Abdurrahman Wahid, a Muslim cleric, edged past the daughter of Indonesia's founder to become the country's Prime Minister in 1999, observers wondered whether he would usher in an era of religious nationalism. In this case the fears were unfounded. The actions of his government showed Wahid's brand of Islam to be moderate and tolerant, one committed to bringing Indonesia into the world community and the global economic market.

In other regions of the world it has not been the creation of new religious states that has been at issue but instead the breakdown of old secular states with no clear political alternative. In some instances religious activists have contributed to these anarchic conditions. In the former Yugoslavia, for instance, the bloodshed in Bosnia and Kosovo was caused as much by the collapse of civil order as by the efforts to create new ethnic and religious regions. Because these situations have been threats to world order, they have prompted the

intervention of international forces such as NATO and the UN. It is, however, world order itself that many of these religious nationalists have opposed. They note that the increasingly multicultural societies of most urban communities around the world have undermined traditional cultures and their leaders. They have imagined the United States and the United Nations to be agents of an international conspiracy, one hell-bent on forming a homogenous world society and a global police state. It was this specter, graphically described in the novel *The Turner Diaries* by Andrew Macdonald, that one of the novel's greatest fans, Timothy McVeigh, had hoped to forestall by attacking a symbol of Federal control in America's heartland, the Oklahoma City Federal building, killing hundreds. In Norway, in July 2011, Anders Breivik undertook a massacre of young people in a summer camp sponsored by a liberal political party; in his manifesto, Breivik identified multiculturalism as the enemy. These, and other terrorist attacks around the world such as Osama bin Laden's bombing of US embassies in Africa in 1998 and of the *USS Cole* in Yemen in 2000, were acts of what might be considered to be a kind of 'guerrilla antiglobalism.'

Ultimately, however, it seems likely that despite these efforts to ignore or reject the forces of globalization, transnational cultures will expand, and that among them will be elements of religion and ethnicity. One future form of religious transnationality may emerge from the international relations of kindred religious states. According to one theory of global Islamic politics that circulated in Egypt in the 1980s and 1990s, local movements of Muslim politics were meant to be only the first step in creating a larger Islamic political entity – a consortium of contiguous Muslim nations. In this scenario, religious nationalism would be the precursor of religious transnationalism. Transnational Islam would lead to Islamic versions of such secular consortia as NAFTA and the European Community. In this Islamic model, however, the divisions among states would eventually wither away, and a greater Islamic union would arise.

A second kind of transnational association of religious and ethnic activists has developed in the diaspora of cultures around the world. Rapid Internet communication allows members of ethnic and religious communities to maintain a close association despite geographic dispersion. These 'e-mail ethnicities' are not limited by political boundaries or national authorities. Expatriate members of separatist communities such as India's Sikhs and both Sinhalese and Tamil Sri Lankans have provided funding for their compatriots' causes. In the case of Kurds their 'nation' is spread throughout Europe and the world, united through a variety of modern communications. In some cases these communities long for a nation-state of their own. In other cases they are prepared to maintain their nonstate national identities for the indefinite future.

Each of these futures harbors a paradoxical relationship between the national and the globalizing aspects of religious politics. This relationship suggests that there is a symbiotic relationship between certain forms of globalization and religious nationalism. It may appear ironic, but the globalism of culture and the emergence of transnational political and economic institutions enhance the need for local identities. They create the desire for a more localized form of authority and social accountability. In an era of globalization the crucial problems are identity and control. The two problems are linked, in that a loss of a sense of belonging leads to a feeling of powerlessness. At the same time what has been perceived as a loss of faith in secular nationalism is experienced as a loss of agency as well as of identity. For all these reasons the assertion of traditional forms of religious identities is tied to attempts to reclaim personal and cultural power. In this sense the various movements of religious nationalism that have arisen at the turn of the twenty-first century can be seen as attempts to regain social control.

See also: European Religion from the Atlantic to the Black Sea; Nationalism, Sociology of; Religion and Economic Life; Religion and New Media; Religion and Politics in the United States; Religion and Youth; Religion: Nationalism and Identity.

Bibliography

Anderson, Benedict, 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Barber, Benjamin R., 1995. Jihad vs. McWorld. Times Books, New York.

Calhoun, Craig, Juergensmeyer, Mark, Van Antwerpen, Jonathan (Eds.), 2011. Rethinking Secularism. Oxford University Press, New York.

Casanova, José, 1994. Public Religions in the Modern World. University of Chicago Press, Chicago.

Habermas, Jürgen, 1997. Modernity: an unfinished project. In: d'Intreves, M.P., Benhabib, S. (Eds.), Habermas and the Unfinished Project of Modernity. MIT University Press, Cambridge, MA, pp. 38–55.

Hedetoft, Ulf, Hjort, Mette, 2002. The Postnational Self: Belonging and Identity. University of Minnesota Press, Minneapolis.

Huntington, Samuel P., 1996. The Clash of Civilizations and the Remaking of World Order. Simon and Schuster, New York.

Juergensmeyer, Mark, 2008. Global Rebellion: Religious Challenges to the Secular State. University of California Press, Berkeley.

Juergensmeyer, Mark, 2003. Terror in the Mind of God: The Global Rise of Religious Violence, third ed. University of California Press, Berkeley.

Juergensmeyer, Mark (Ed.), 2005. Religion in Global Civil Society. Oxford University Press, New York.

Khomeini, Imam [Ayatollah], 1985. Islam and Revolution: Writings and Declarations [1981] (Hamid Algar, Trans.). Routledge and Kegan Paul, London.

Khomeyni, Imam [Ayatollah], 1997. Collection of Speeches, Position Statements [Translations from "Najaf Min Watha 'iq Al-Imam Al-Khomeyni did al-Quwa al Imbiriyaliyah wa al-Sahyuniyah wa al-Raj'iyah" ("From the Papers of Imam Khomeyni Against Imperialist, Zionist, and Reactionist Powers")]. Translations on Near East and North Africa, Number 1902. Joint Publications Research Service, Arlington, VA.

Kippenberg, Hans, 2012. Violence as Worship: Religious Wars in the Age of Globalization. Stanford University Press, Stanford.

Lie, John, 2004. Modern Peoplehood. Harvard University Press, Cambridge, MA.

Lincoln, Bruce, 2002. Holy Terrors: Thinking about Religion after 11 September 2001. University of Chicago Press, Chicago.

Roy, Olivier, 2006. Globalized Islam: The Search for a New Ummah. Columbia University Press, New York.

Rudolph, Susanne Hoeber, Piscatori, James (Eds.), 1997. Transnational Religion and Fading States. Westview Press, Boulder, CO.

Toft, Monica Duffy, Philpott, Daniel, Shah, Timothy, 2011. God's Century: Resurgent Religion and Global Politics. W.W. Norton, New York.

Wellman, James K. (Ed.), 2009. Belief and Bloodshed: Religion and Violence across Time and Tradition. Rowman and Littlefield, Lanham MD.

Westerlund, David (Ed.), 1996. Questioning the Secular State: The Worldwide Resurgence of Religion in Politics. Hurst, London.

Religious Organizations

James A Beckford, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Religious organizations are diverse and distinctive while sharing characteristics with other types of organization. Pioneering studies identified types and subtypes, but more recent research has focused on the dynamics of religious organizations in respect of their forms of authority, governance, and leadership. There are wide variations in leadership roles, local settings, congregational forms, and patterns of recruitment and training. Religious organizations raise questions about the growing visibility of religious international nongovernmental organizations at the UN; the development of Buddhist, Hindu, Muslim, and Sikh organizations in liberal democracies; and the salience of legal questions arising from the growth of religious diversity.

Definitions of religion are varied and contentious. The notion of religious organizations is equally challenging because it seems to assume that the distinction between religion and nonreligion is analytically useful and can be drawn clearly. In fact, ambiguity and uncertainty cast doubt on the distinction, especially if the term 'spirituality' is also taken into account. A better approach than attempting a clear distinction is to regard the boundary between religious, spiritual, and nonreligious organizations as flexible, porous, and shifting. This approach places religious organizations on a sliding scale: at one end of the scale, they share much in common with other organizations or movements while, at the other, they retain characteristics which mark them out as significantly different. If the defining characteristic of religions is their assumption that deities, supernatural orders, or superhuman beings are active agents in the human world, then religious organizations are likely to be considered highly distinctive. If, on the other hand, religions are defined in terms of more inclusive assumptions about ultimate truth, the 'felt whole,' human flourishing, or the sacredness of the human world and its natural environment, then religious organizations are less likely to appear markedly different from other voluntary and communal organizations.

In short, attempts to draw a sharp distinction between religious and nonreligious organizations divert our attention from relevant overlaps between these two types of organizations (Demerath and Schmitt, 1998). Research on religious organizations therefore seeks to identify the extent to which they are significantly different from other organizations in terms of their personnel, goals and values, production processes, products, resources, authority structure, and networks. Examples include religious codes of law, charismatic religious leadership, and the potential of religious convictions for motivating donations of time and money to faith-based organizations. At the same time, the distinction between religious and nonreligious organizations may be important for legal or material reasons. For example, taxation advantages or exemption from antidiscrimination laws are available to organizations defined as religious in some countries. The consequences of failing to meet official criteria defining religious organizations can be grave in respect of groups claiming conscientious objection to compulsory military service or eligibility to provide chaplains in hospitals, prisons, or military establishments. Scholarly interest in questions about law and regulation of religious organizations increased strongly in the first decade of the twenty-first century, as did corresponding concerns with the freedom of religion.

Diversity and Distinctiveness

Religious organizations take many different forms, although the focus of their activities is typically on worship, prayer, meditation, teaching, healing, and spiritual well-being. These core activities are rooted in the revelations, texts, codes, laws, and principles which are regarded as authoritative by the people who identify or associate with particular religious organizations. Some of these people are, if we use the metaphor of markets, merely 'clients' who consult healers, spirit mediums, gurus, seers, or shamans. But others have strong and long-lasting commitments to transnational organizations such as the Roman Catholic Church, the Gülen movement, or Soka Gakkai International. Even the most formal, hierarchically structured religious organizations provide space for informal activities and relationships.

Religious organizations are also highly diverse in terms of their 'products,' and production processes. They typically produce sacred knowledge, sacramental acts, transcendental experiences, prophecies, worship and meditation, healing rituals, rites of passage, religious ethics, and social welfare programs. These products are delivered by professional and lay specialists, monks and nuns, recruiting missions, education programs, publishing and broadcasting activities, political and social movement campaigns, social service agencies, and local fellowships. The effectiveness of these activities depends largely on the capacity of religious organizations to secure adequate resources, to motivate, regulate, and train their personnel and followers, to coordinate their programs, to cope with dissent and indifference, to communicate with their environments, to defend their market share, and to deter free riders. It is typical of religious organizations to pursue their interests by establishing local fellowships, professional and lay ministries, regional, national, and international sources of authority, seminaries, schools, hospitals, welfare agencies, chaplaincies, publishing enterprises, and missionary orders.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.84044-1

In addition, many religious organizations participate in ecumenical or interfaith activities that cut across the boundaries of individual organizations, giving rise in some cases to higher-level organizations. In turn, interfaith bodies can function not only as sites of discussion but also as representative agencies that may be consulted by public authorities at local and national levels. Indeed, the public significance of interfaith organizations increases in countries where there is a growing diversity of religions. Nevertheless, the question remains open of whether official recognition of interfaith organizations undermines the authority of each separate religion and leads to a form of 'generic' religion. This problem is less likely to occur when religious organizations form alliances with one other for practical purposes such as challenging laws that are perceived to be damaging to shared interests, or joining *amicus curiae* submissions to courts of law in cases where the issues appear to affect religious organizations as a category.

History

The study of religious organizations has developed in two main streams. The first is concerned primarily with their distinctiveness and their subtypes. The second deals with their internal dynamics and their varying fortunes in different environments.

Types and Subtypes

Relations between religion and modernity were at the root of the pioneering studies that Max Weber (1864–1920) and Ernst Troeltsch (1865–1923) made of religious organizations. Weber highlighted the crucial contributions made by Ancient Judaism and Christianity toward the rationalization of the modern world and the differentiation of religion from other institutions. He argued that secularization, bureaucratization, and the disenchantment of the modern western world had all reduced the social significance of religious organizations. Although church-type organizations (large, territorially based, and inclusive in membership terms) still retained the potential to exercise power, Weber thought that their ties to the secular world hindered the translation of their Christian ethics into effective programs of action. By contrast, he believed that sect-type organizations (voluntaristic, relatively small, and exclusive in membership terms) which retained distinctive and radical versions of Christian theology and ethics were nevertheless unable to influence the world around them because they kept their distance from it. A major challenge for religious organizations is to achieve a suitable balance between the search for power and the retention of ideological purity.

The process of rationalization was of less interest to Troeltsch. His concern was with the historical evolution of the social forms of 'church,' 'sect,' and 'mysticism' to which the New Testament's ethics had given expression. The main question for Troeltsch was whether liberal Protestantism was sufficiently robust and well adapted to modernity to act as a social and political vehicle for Christian ethics in a secularizing world.

Subsequent generations of sociologists have used 'church,' 'sect,' and 'cult' – as well as variants such as 'established sect' and 'denomination' – to analyze patterns of continuity and change among religious organizations not only in Christianity but also in other faith traditions. Despite the difficulties of adapting these concepts to traditions such as Hinduism, Islam, and Buddhism, the concept of sect is still employed in analyses of religious movements in many parts of the world. Bryan Wilson's (1970) scheme of seven subtypes of sects remains especially influential. The concept of 'cult' has come to figure prominently in popular, as well as scholarly, discussions of religious movements such as the Unification Church, the Hare Krishna movement, and the Church of Scientology, all of which became controversial in the 1970s.

Beginning in the 1980s, the concepts of 'sect' and 'cult' underwent further development as proponents of rational choice – or religious economies – theory offered novel explanations of the capacity of 'high demand' religious organizations to attract and retain members. The theory held that certain people are rationally willing to incur the relatively heavy costs of participating in strict religious organizations such as sects and cults in order to obtain supernatural compensators which purport to give convincing answers to questions of ultimate meaning. It also depicted churches as offering answers that conformed to prevailing societal norms and values and were therefore unable to offer the selective benefits that sects and cults provided exclusively for their members in return for incurring high costs. Rational choice theorizing thereby tried to account for the historical formation of sects (by schism) and of cults (by innovation) as a never-ending process of competition in religious markets between religious organizations standing in different degrees of tension with culture and society (Iannaccone, 1988).

The basic concept of 'church' has undergone relatively little refinement since the early twentieth century, although the notions of 'house church' and 'megachurch' were developed in the 1980s. Similarly, little progress has been made in applying or refining the closely related concepts of 'ecclesiola in ecclesia,' 'fraternitas,' and 'religious order' that Joachim Wach (1898–1955) and others had elaborated. Instead, the late twentieth century saw the development of other concepts designed to capture the organizational forms taken by Christianity in Sub-Saharan Africa, Latin America, and the former Soviet Union. They included 'independent churches,' 'African initiated churches,' 'base ecclesial communities,' and 'underground churches.' But in North America greater analytical attention was given to 'denominations' and their constituent 'congregations' at the local level. Research has centered on the social and ethnic patterning of denominations, the conditions favoring schism and merger, the ambiguous status of theological seminaries, the position of female clergy, the complex webs of 'interorganizational fields' in which denominations operate, the polarization of conservative and liberal factions and the associated threats to denominational integrity, the growth of special purpose groups, and tensions between religious hierarchies and functional agencies (Chaves, 2004; Ammerman, 2009; Roozen, 2011). Another rich vein of research has investigated the contribution of denominations toward ideologies of pluralism, pluralistic provision of social welfare, social movement campaigns, and the Third Sector of community and voluntary work (Chaves and Tsitsos, 2001). There is also strong evidence of growing isomorphism between the structures of religious organizations involved in social service delivery and the dominant forms of nonreligious, not-for-profit organizations – no doubt partly as a consequence of

tax codes that 'flatten out' differences between all tax exempt organizations.

The study of religious organizations has developed more slowly in relation to religions that are less formally structured than are most Christian organizations. Thus, although the authority structures and principles of membership may be less formal than in Christianity among indigenous spiritualities, popular religiosity, and religious traditions such as Buddhism, Hinduism, Islam, Jainism, Judaism, Sikhism, Taoism, and Shintoism, they nevertheless have distinctive forms of organization. For example, Muslim organizations in Muslim-majority countries as well as in the countries of Western Europe, Australasia, and North America where growing numbers of Muslims are living as both migrants and settlers include 'brotherhoods,' mosque committees, seminaries, '*biraderi*' (clan networks), and '*waqf*' (charitable endowments). Similarly, Buddhist monastic institutions (*sangha*) and transnational movements of 'engaged' or 'compassionate' Buddhism have adapted to widely differing cultural, social, and political circumstances (King, 2009). Moreover, Hindu movements actively promote their values through transnational programs of school-level education and political mobilization. Yet, research has also underlined the extent to which the organization of non-Christian religious activities in Europe, North America, and Australasia is increasingly following the path of denominationalization on the basis of local groupings that closely resemble congregations (Yang and Ebaugh, 2001).

Dynamics of Religious Organizations

The confluence of two intellectual developments beginning in the 1950s helped to stimulate new studies of organizational dynamics in religions. One was the growing salience of 'organization theory' in social sciences, and the other was a concern with 'ministry' as a modern profession in religious organizations experiencing rapid social change (Beckford, 1973). Although studies of religious organizations made relatively little impact on organizational theories in general they chimed well with emerging interests in organizational climate and culture in the 1970s. They also overlapped with newsworthy research on communes and social movement cultures which highlighted the importance of strong commitments and normative beliefs. Subsequent findings about the importance of rituals, retreats, and emotion management in all kinds of organizations only enhanced the interest that religious organizations came to have for organizational analysts (DiMaggio, 1998). The rapprochement between the study of formal organizations and the study of religious organizations was further boosted in the 1980s when so-called New Age and quasireligious and parareligious businesses began to attract attention for trying to combine profit-seeking activity with ethical and religious practices. 'Selling spirituality' (Carrette and King, 2005) has also elicited skepticism and criticism.

Indeed, general questions about the resourcing of religious organizations do not figure as prominently in research as they should. Competition for members and the material resources that they bring to religious groups is a feature of societies such as the United States and Japan where segmented markets exist within the broad categories of Christian, Buddhist, and Shinto organizations. In many other countries, however, 'religious markets' are much less vibrant except among the new or marginal movements which compete for immigrants or drop-outs from the locally dominant religions. Rates of 'switching' between religious organizations also vary widely between countries or regions with different degrees of competition between a monopoly, majorities, pluralities, and minorities.

The economic underpinning of religious organizations is no less varied than are their other characteristics. Some organizations such as the Catholic Church and the Orthodox Church in parts of Central and Eastern Europe have acquired substantial amounts of land, property, and wealth over the course of centuries. Others such as Buddhist organizations in Asia possess priceless buildings and artifacts. Meanwhile, the assets of some of Islam's charitable foundations and brotherhoods contribute strongly to local economies. The generation of economic resources can therefore take many different forms, although variants on 'enrollment economies,' 'freewill offering,' and tithing are widespread. Moreover, the tax systems of some states confer fiscal advantages on religious organizations.

But very few large religious organizations – unlike their secular counterparts – engage in outsourcing or subcontracting their central operations to other suppliers. The main reason for this is that it would be difficult for religious organizations to protect their all-important sense of ideological or theological integrity if they had to depend on agencies outside their control. In extreme cases, sectarian and communal organizations that insist on distancing themselves from 'the world' may choose to make their own furniture, grow their own food, make their own clothes, and print their own publications. By contrast, the revenue of other religious organizations can come from the sale of products and services to the public, including access to sites of pilgrimage and worship, the provision of healing rituals, and the conduct of rites of passage. Indeed, such organizations also enter into contracts with suppliers of, for example, sacred artifacts, *kosher* or *halal* food, musical instruments, and ritual clothing. But controversies tend to engulf those newcomers to the religious marketplace which solicit direct donations from the public, encourage followers to make bequests and legacies, or charge fees for participation. Equally controversial is the fact that some religious organizations expect their members to work for them without payment and without sickness or retirement benefits.

Two dimensions of religious organizations are crucial for their dynamics: authority and leadership. This does not mean that other dimensions are unimportant but it signifies that organizations purveying allegedly sacred truths and the highest values show a disproportionately strong concern with establishing, reproducing, and defending their teachings and practices against perceived corruption, subversion, or apathy.

Authority

Not only are religious organizations in competition with purveyors of rival versions of ultimate truths but they also have to monitor and regulate their own members or followers in order to deter heresy, apostasy, and abuse. This is why questions of authority are crucial for religious organizations and why change

often meets with strong resistance. Religious organizations are therefore distinctive, if not unique, for the effort they put into preserving and elaborating the connections between their founding truths and their organizational structure. The authority of leaders depends heavily on their success in constantly rehearsing these connections in rituals, symbols, exegesis of sacred texts, preaching, praying, teaching, and meditating.

Difficulties can occur, however, in religious organizations that have a division of labor between a clerical hierarchy and a parallel set of agencies responsible for the management of finances, property, human relations, publishing, education, and healthcare. The delivery of welfare services by faith-based initiatives in some religious organizations is a further source of potential difficulties, especially if it exacerbates tension between local groups and central bureaucracies. Competing sources of authority between these different organizational structures can generate friction and conflict. Nevertheless, religious organizations remain more numerous and more active than any other American nonprofit organizations in providing social services (Cadge and Wuthnow, 2006).

Sources of Authority

Christianity has taken many organizational forms that are unique among the world's religions not only for being highly centralized and formal but also for cutting across ethnic, tribal, and political groupings. Indeed, some churches in medieval and early modern Europe operated as part of the machinery of state. By contrast, Hinduism, Judaism, and Shinto crystallized slowly out of predominantly caste-based, tribal, and ethnic cultures, thereby giving rise to temples, synagogues, and shrines as the locations for rites and ritual worship rather than as agencies of civil administration. Diasporic Judaism, lacking the central authority previously vested in the Temple at Jerusalem and its hereditary priesthood, regrouped around rabbis as teachers of the law and around synagogues as ritual centers. Communal organizations continue to play a major role in preserving the ethnic and religious identity of many Jews, but would-be central agencies for controlling Jewish beliefs and practices do not command universal compliance.

Buddhism and Islam both emerged from deliberate attempts to cultivate philosophical and religious systems that were partly at odds with their social context but they have also formed close alliances with states and political powers at various times. Buddhism in Japan and Thailand, for example, has sometimes functioned through parish systems. Since 1945 massive, hierarchical movements of lay Buddhists such as Reiyūkai, Risshō Kōseikai, and Sōka Gakkai have flourished in Japan. In many regions of Asia, large-scale and well-organized movements of 'compassionate' or 'engaged' Buddhism are currently mobilizing their followers in campaigns for justice, peace, and environmental sustainability (King, 2009).

The organizational forms of Islam are subtle, multistranded, and variable by region of the world. Although Islam lacks a centralized authority structure, it can be effectively interwoven with states, political parties, and liberation movements. Indeed, the complex organizational tapestry of Islam consists of multiple – and sometimes competing – sources of theological, legal, spiritual, political, and moral authority. The authoritative figures include learned experts, jurists, hereditary sheikhs, leaders of brotherhoods, and clergy (among the Shi'ites). Now that tens of millions of Muslims from different national and ethnic backgrounds live as religious minorities in North America, Australasia, and Europe, some governments have encouraged the formation of representative bodies of Muslims that could help to resolve questions about, for example, the licensed slaughter of animals for meat, the provision of burial spaces, and the management of Muslim chaplaincies in hospitals, prisons, and military institutions.

Forms of Governance

The implementation of religious authority in forms of governance can take many different forms in religious organizations. At one extreme are loosely structured fellowships, and at the other extreme, are large, complex, and formal organizations which require their members to accept centrally imposed dogmas and practices mandated by leaders. Most religious organizations fall on a continuum between these two extremes, although their precise location may change in response to internal or external factors. The time-honored distinction between church-type and sect-type organizations is no longer a good guide to these variations in degrees of complexity, centralization, and formality.

Differences among Orthodox, Episcopal, Presbyterian, and Congregational forms of governance reflect differing interpretations of Christian scripture. Similarly, the distinction between the authority exercised by ordained clergy, monks, and nuns in religious orders, and laity has theological justification. But social scientific research examines how well the official authority structures actually inform the actions of members of religious organizations and how far external factors shape organizational cultures. For example, the Episcopal form of governance is interpreted and implemented in widely differing ways in churches in the worldwide Anglican communion. Presbyterian and congregational forms of governance have also come under pressure to adapt their exercise of authority to the sociocultural norms and values prevailing in different societies. In particular, critics of patriarchalism have effectively challenged the gendered distribution of authority in churches and tried to infuse them with feminist, antihierarchical values (Nesbitt, 1997). Moreover, tension between congregational autonomy and pressure to conform to the authority at the regional and national levels of religious organizations is common.

Fundamentalism

Questions about religious authority have dominated analyses of the forms of religious fundamentalism that became influential in various faith traditions in the closing decades of the twentieth century. Fundamentalism centers on unswerving faith in the authority of fixed principles, usually enshrined in sacred texts, to dictate categorical rules for the conduct of personal and public life. It is also associated with assertive leadership and deep distrust of life outside its groups of committed believers. Christian fundamentalism thrives in local congregations, religious broadcasting media, online networking, and social movement campaigns against, for example, abortion and gay

rights. The 'vehicles' of Islamic fundamentalism also include the full gamut of new communications technologies, especially dedicated websites, as well as conservative institutions of Islamic learning. Hindu fundamentalism is closely associated with the political mobilization of resources in India and in diasporic communities in support of Hindu nationalism. Fundamentalism is not, therefore, antimodern, traditionalist, or otherworldly. On the contrary, it seeks to transform the world in accordance with its reading of authoritative texts, and this is compatible with forms of organization that are effective, flexible, and responsive to local contexts.

Leadership

Roles

Religious organizations vary widely in their leadership roles, structures, and training (Hoge, 2009). Ideas about the source and nature of sacred power shape the precise pattern of roles at all organizational levels. Leaders of organizations that identify sacred power exclusively with scriptures and laws derive their authority mainly from their acquired capacity to read and interpret them. Scholars, teachers, and jurists exemplify this pattern of leadership in Judaism and Islam. The various forms of priesthood that are characteristic of Hinduism, Shinto, Christianity, and Sikhism combine reverence for sacred texts with belief in the capacity of priests to perform rituals allowing human beings to interact with divinities and sacred power. The effectiveness of this power is believed to be conditional on the proper conduct of rituals by priests. In contrast, prophets and charismatic leaders are more likely to aim at reforming religions by opposing their institutionalized practices and seeking to replace them with radically new beliefs and practices. Different again is the position of ministers in many Protestant traditions whose leadership consists mainly of preaching, exposition of scripture, pastoral care, and – at least in evangelicalism – efforts to persuade followers of their sinfulness and need for redemption. Some other religious traditions have 'virtuosi' such as monks, nuns, and 'holy men' who exercise leadership on the strength of their discipline, asceticism, devotion, and selflessness. For example, the *sangha*, or monastic community, remains at the heart of Buddhism in many parts of the world; monastic institutions were the bulwark of Christianity in the early European Middle Ages; Catholic religious orders have been in the forefront of missionary work in Africa, Asia, and Latin America; and monks continue to occupy the highest leadership positions in some Orthodox churches. But strong notions of religious leadership tend to meet resistance in traditions that locate authority in the spiritual freedom of individuals to find their own way to the sacred, sometimes under the guidance of spiritual masters, teachers, or gurus.

Although the founders of some new religions in Japan, of spiritist movements in the Caribbean region and Central America, and of new spiritualities in the West have been women, the pattern in many religions is either to exclude women from positions of leadership or to prevent them from gaining access to the highest positions. The ordination of female priests and the consecration of women bishops have been highly contentious in the worldwide Anglican communion. Pressure to admit women to the Roman Catholic priesthood has been mounting since the 1960s when a shortage of priests became apparent in western countries, and large numbers of nuns began to leave their orders and their work in schools, orphanages, and hospitals.

Local Settings

Religious leaders are usually attached to particular buildings or social centers where they conduct teaching, preaching, meditation, worship, prayer, or pastoral functions. Some also have spiritual oversight of institutions such as schools, Sunday Schools, youth organizations, seminaries, orphanages, hospitals, refuges, pilgrimage sites, prisons, and military establishments. Many Christian organizations deliver religious, pastoral, and welfare services in geographical areas known as parishes representing the local units of national or international churches. The idea that residents of a locality would belong to its parish and form a congregation of worshippers under the leadership of local clergy was especially widespread in parts of Europe where the dominant church was closely aligned with the state. Japanese Shinto also has a system of parishes linked to local shrines.

Congregations

Congregations, as the basic unit of most Christian organizations, are the social groups with which members identify most strongly (Ammerman, 2009). Their precise status varies with the theology and authority structure of the broader churches or denominations to which they belong. The congregations of hierarchical churches are local outlets of ideas and practices controlled by regional, national, and international authorities such as bishops, archbishops, cardinals, and the Pope. By contrast, local congregations in religious organizations with little or no theological justification for hierarchy or centralized control tend to enjoy relative autonomy. Examples include independent fellowships of conservative Protestants, the Reform Synagogues of Great Britain, and the Christadelphians. The balance between centralization and local autonomy in religious organizations is an important determinant of their capacity to generate resources, mobilize members, deliver services, respond to internal problems, and influence public life.

As Hinduism, Islam, and Sikhism have never been structured in the regions of the world where they originated in the form of inclusive, unitary organizations, their local groups have not tended to function as congregations or parishes. Although temples, mosques, and gurdwaras undoubtedly operate as ritual and pastoral centers for people in their vicinity, it is rare for them to be strongly integrated into higher-level structures. They are accountable to their private owners or to their management committees. Consequently, local groups select and retain their religious leaders as professionals or qualified volunteers rather than as employees or agents of religious organizations at a higher level.

Nevertheless, the growing numbers of Hindus, Muslims, and Sikhs now residing in Europe and North America are adopting forms of organization that more closely resemble Christian patterns of denominations and congregations (Yang and Ebaugh, 2001). Their increasing involvement in chaplaincies in healthcare, military, and prison establishments is another

indication of their ability to adapt their organizational forms to local contexts.

Recruitment and Training

The major Christian churches and denominations in advanced industrial societies are experiencing varying degrees of difficulty in recruiting and retaining their clergy (Carroll and McMillan, 2006). At the same time, the social characteristics of recruits have also changed since the mid-twentieth century: women are now much more numerous; a career in the clergy is a second career for a growing number of recruits; ethnic diversity is more apparent; and some churches actively seek recruits from developing countries. But despite the 'feminization' of religious leadership (Nesbitt, 1997), the upper echelons of many churches are still under the control of men.

It is mainly in the economically developing countries of Africa, Asia, Eastern Europe, and Latin America that the supply of young women and men for Christian ministry training remains relatively buoyant. It is no accident that these regions of the world are also those where participation in churches and denominations has also been increasing strongly in recent decades. Similarly, the production of religious scholars and imams in *madrassas* in the heartlands of Islam in the Middle East, North Africa, and Asia has expanded to meet the demand for qualified men to serve in the growing number of mosques and chaplaincies in Europe and North America. Globalization affects religious organizations no less powerfully than other kinds of organization.

Questions About the Future

Religious organizations – notably the Roman Catholic Church – were among the major contributors to the emergence of civil society at the international level in the nineteenth century. Indeed, religious international nongovernmental organizations (RINGOs) still constitute one of the largest categories of lobbyists at the United Nations. Recognition of the important roles played by religious organizations in articulating, aggravating, and solving international problems has increased sharply since the 1980s. But it remains to be seen how the growing strength of China, India, and Indonesia in international politics and economics will affect the balance of power among the RINGOs.

A closely allied issue for the future arises from the fact that processes of globalization have fostered the growth of Buddhist, Hindu, Muslim, and Sikh organizations in the liberal democracies of Western Europe, North America, and Australasia – as well as in the regions of the world where they had long been powerful. Transnational religious movements continue to play leading roles in establishing the organizational infrastructures for the practice and the promotion of Buddhism, Hinduism, Islam, and Sikhism. At the same time, agencies of the state have responded to these developments with varying degrees of enthusiasm. One of the questions for the future is, therefore, how well these recently implanted religious organizations will adapt to their legal, political, and cultural circumstances. Another question is whether their forms of organization will follow the logic of isomorphism with other organizations in the voluntary sector of their respective civil societies. A third issue is about their prospects for influencing the organizational forms of Buddhism, Hinduism, Islam, and Sikhism – and other religions – both in their countries of origin and in their countries of settlement.

Finally, religious organizations in the West seem increasingly likely to be affected by legal issues arising from the growth of religious diversity occasioned mainly by globalization. The implementation of international codes of human rights and of domestic laws against discrimination on the grounds of religion (or in protection of religious freedom) increasingly encounters the objection that there must be limits to the accommodation of religious diversity. The very understanding of 'religion' is germane to many legal cases concerning the right to manifest or to practice religion in public. As a result, the eligibility of religious organizations for tax privileges, for involvement in public education, and for the status of contractors to public authorities is increasingly at stake in legal disputes. For example, chaplaincies in the military, prisons, and hospitals have become sites of contention not only for the advocates of religious freedom but also for secularist opponents to any form of official recognition or resourcing of religious organizations.

See also: Buddhism; Catholicism; Christianity in Asia; Christianity in Latin America; Christianity in the English-Speaking West; Evangelical, Revivalist, and Pentecostal Religions; Islam: Asia; Islam: Middle East; Islam: Sub-Saharan Africa; Judaism; New Religious Movements; Religion, Sociology of; Religion: Mobilization and Power; Religious Fundamentalism: Cultural Concerns; Secularization.

Bibliography

Ammerman, Nancy T., 2009. Congregations. Local, social, and religious. In: Clarke, Peter B. (Ed.), The Oxford Handbook of the Sociology of Religion. Oxford University Press, Oxford, pp. 562–580.

Beckford, James A., 1973. Religious organization. Current Sociology 21 (2), 1–170.

Cadge, Wendy, Wuthnow, Robert, 2006. Religion and the nonprofit sector. In: Powell, Walter W., Steinberg, Richard (Eds.), The Non-profit Sector, second ed. Yale University Press, New Haven, CT, pp. 485–505.

Carrette, Jeremy, King, Richard, 2005. Selling Spirituality. The Silent Takeover of Religion. Routledge, London.

Carroll, Jackson W., McMillan, Becky R., 2006. God's Potters: Pastoral Leadership and the Shaping of Congregations. Wm. B. Eerdmans Publishing, Grand Rapids, MI.

Chaves, Mark, 2004. Congregations in America. Harvard University Press, Cambridge, MA.

Chaves, Mark, Tsitsos, William, 2001. Congregations and social services: what they do, how they do it, and with whom. Nonprofit and Voluntary Sector Quarterly 30 (4), 660–683.

Demerath III, N.J., Schmitt, Terry, 1998. Transcending sacred and secular: mutual benefits in analyzing religious and nonreligious organizations. In: DemerathIII, N.J., Hall, Peter D., Schmitt, Terry, Williams, Rhys H. (Eds.), Sacred Companies. Organizational Aspects of Religion and Religious Aspects of Organizations. Oxford University Press, New York, pp. 381–400.

DiMaggio, Paul, 1998. The relevance of organization theory to the study of religion. In: DemerathIII, N.J., Hall, Peter D., Schmitt, Terry, Williams, Rhys H. (Eds.), Sacred Companies. Organizational Aspects of Religion and Religious Aspects of Organizations. Oxford University Press, New York, pp. 7–23.

Hoge, Dean R., 2009. The sociology of the clergy. In: Clarke, Peter B. (Ed.), The Oxford Handbook of the Sociology of Religion. Oxford University Press, Oxford, pp. 581–596.
Iannaccone, Laurence R., 1988. A formal model of church and sect. American Journal of Sociology 94 (Suppl.), 241–268.
King, Sallie B., 2009. Socially Engaged Buddhism. University of Hawai'i Press, Honolulu.
Nesbitt, Paula D., 1997. The Feminization of the Clergy in America. Oxford University Press, New York.
Roozen, David A., 2011. A Decade of Change in American Congregations, 2000–2010. Hartford Institute for Religion Research, Hartford, CT.
Wilson, Bryan R., 1970. Religious Sects. Weidenfeld & Nicolson, London.
Yang, Fenggang, Ebaugh, Helen Rose, 2001. Transformations in new immigrant religions and their global implications. American Sociological Review 66 (2), 269–288.

Relevant Websites

http://www.thearda.com/ – Association of Religion Data Archives.
http://www.brin.ac.uk/ – British Religion in Numbers.
http://www.cesnur.org/ – Cesnur. Centre for Studies of New Religions.
http://hirr.hartsem.edu/ – Hartford Institute for Religion Research.
http://www.pewforum.org/ – Pew Forum on Religion & Public Life.

Remedies and Damages: Legal Aspects

Theodore Eisenberg, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Damages levels vary across countries. The size of damages awards in the United States became an issue of debate stemming from asserted civil liability crises. Little evidence exists that damages awards have substantially increased over time and one longitudinal 40-year data set shows no real growth over four decades. Punitive damages in the United States may be the most controversial area of damages. The best available evidence is that punitive damages awards are infrequent: punitive damages are requested in about 10% of completed trials and awarded in about one-third of the cases in which plaintiffs won and they were requested. Growth in punitive damages award levels has not followed a clear trend. The relation between punitive and compensatory awards is highly statistically significant, contrary to claims that punitive damages amounts are random.

Remedies can be categorized as monetary or nonmonetary. Although much litigation leads to monetary awards, many cases end without a monetary award. Injunctive relief, orders to specifically perform contracts, divorce decrees, probate of wills, and other orders do not necessarily involve damages or monetary payments.

Nonmonetary Relief

Several categories of nonmonetary relief exist and their use varies across countries. For example, three different approaches exist to the availability of specific performance – a particular act that a party is ordered to perform by a court – for contract breach: (1) some countries, such as Germany and socialist systems, make specific performance generally available, subject to exceptions, (2) other countries, such as France, make specific performance available for specific obligations, while (3) common law countries, including England and the United States, regard specific performance as an exceptional remedy, available only in limited circumstances. In the United States, evidence exists that sophisticated parties often contractually provide for specific performance (Eisenberg and Miller, 2014).

Nonmonetary relief to prevent injury can be more important than redressing injuries already done. Common law countries address this need by allowing injunctive relief in cases in which irreparable harm can be demonstrated. Other legal systems, such as Germany, Switzerland, Austria, and others influenced by German law, have a general class of actions to prevent threatened injury. Courts in such systems have less discretion to deny preventive remedial relief than do common law courts. Some systems allow remedial relief but lack a coercive component, such as a penalty for violating an injunction, to enforce it (Stoll, 1983). Prejudgment nonmonetary remedies, such as attachment and preliminary injunction or its equivalent, are widely available.

Damages Rules across Countries

With respect to monetary relief, one cannot completely separate the question of the quantum of damages from the question of liability. The English jurist William Blackstone (1723–80), in writing about the common law, did not regard damages as a separable ending of the legal process. The right to damages arose the instant the victim was injured. Nevertheless, one can intelligently address many damages issues independently of the question whether any damages should be awarded.

In the last quarter of the twentieth century, damages issues became a focus of the tort reform movement in the United States, alarming foreigners doing business in the United States. In many respects, however, damages rules are not strikingly different between the United States and other countries. All countries seek to compensate victims fully, though most depend more on social compensation systems than does the United States. Most countries allow consideration of the injured person's fault, a consideration similar to the United States doctrine of comparative negligence, and a few (the Netherlands and Sweden) allow for mitigation of damages in view of the parties' financial circumstances and insurance coverage. European and Japanese damages rules governing injury, though less feared than United States rules, are not necessarily simpler. They have been said to contain confusing variation both in the constituents of noneconomic damages and in measuring them (Pfennigstorf, 1991).

There are, however, differences in the size of awards. Non-United States systems produce lower and more predictable within-country awards. Some of this predictability stems from written and unwritten guidelines. In Japan and Sweden, for example, expert commissions establish consistent guidelines for automobile accidents. Sweden extends this practice to injuries caused by medical mistreatment, by pharmaceutical drugs, and by environmental pollution. Some of the perceived award consistency may be due to the absence of studies. For example, 38%–59% variation in Austrian awards for pain and suffering has been reported based on the valuation methodology courts use (Flatscher-Thöni et al., 2013). Across European countries, awards are not consistent. One study estimates that in a wrongful death case involving a 20-year old victim with no dependents, awards across the European Union would vary from £176 368 in Italy to £1250 in England; it also finds that pain and suffering awards vary considerably across countries (McIntosh and Holmes, 2003).

Social Scientific Study of Compensatory Damages

Media reports, which understandably emphasize extreme awards, promote exaggerated perceptions of large United States damages awards. The media overreport products liability and medical malpractice cases, exaggerate plaintiff win rates, and distort impressions about the actual distribution of jury awards. In addition, special interest groups that seek to minimize the liability of large corporations promote inaccurate descriptions of the legal system that reinforce the media's tendency to emphasize large recoveries. Interest groups such as the US Chamber of Commerce and the American Medical Association exaggerate the impact of damages awards and the legal system's effect on primary actors' behavior (Eisenberg, 2009). Social scientific assessment of damages is critical in helping to ameliorate the effect of both the media reporting tendencies and the effect of interest groups' exaggerations.

General Adequacy of Damages

Despite agreement that awards are higher in the United States, little evidence exists to support the claim that injury victims are systematically overcompensated in any country. For example, in the United States, evidence exists that plaintiffs with small losses tend to be overcompensated but plaintiffs with large losses tend to be undercompensated (Vidmar, 1998). An English study found that serious injuries may not result in substantial recoveries and that minor injuries that do lead to recoveries may have unexpectedly severe consequences (Law Commission, 1994). Both findings suggest systematic undercompensation. Evidence from medical malpractice studies establishes underdeterrence of negligent behavior by hospitals and physicians (Mello et al., 2007). The massive scope of medical malpractice has not materially diminished over time (Landrigan et al., 2010; National Health Care Quality Report, 2007; National Health Care Quality Report, 2009).

Award Levels and Trends

The most studied system of damages is the tort system in the United States. Concerns about the level of awards, time trends in awards, and the relative competence of judges and juries have generated several studies. Data gathered by the National Center for State Courts from 45 of the largest 75 United States counties, under grants from the Bureau of Justice Statistics (BJS) (1995, 1999, 2004, 2008), provide the least biased sample of awards in litigation. The data allow for meaningful comparisons across case categories. Moreover, since the same 45 counties were surveyed in 1991, 1996, 2001, and 2005 (with 110 additional small counties surveyed in 2005), the data also allow meaningful comparisons over time. **Table 1** summarizes the pattern of jury and judge compensatory damages awards for the combined BJS data sets, in thousands of 2005 dollars by case category and by trial mode (judge or jury).

Table 1 Mean and median compensatory damages award, state court trials, by case category trial mode, 1991–2005, in thousands of 2005 dollars

	Mean	*Median*	N	*Mean*	*Median*	N
Case type	*Compensatory award, juries*			*Compensatory award, judges*		
Motor vehicle tort	284	22	5533	340	17	490
Premises liability	586	92	1493	384	47	137
Products liability, asbestos	1602	492	170	164	164	1
Products liability, other	3402	500	110	196	73	17
Intentional tort	596	50	403	166	38	112
Medical malpractice	1953	455	769	284	228	26
Professional malpractice, other	1137	173	125	454	55	39
Slander/libel	935	35	68	561	19	21
Animal attack	477	96	160	78	38	10
Conversion	477	115	22	199	20	31
False arrest, impris., prosec.	364	187	14	40	31	3
Other or unknown tort	5667	109	528	179	38	68
Fraud	1107	79	453	451	44	350
Seller plaintiff	375	64	604	141	34	1487
Buyer plaintiff	400	58	656	145	17	493
Employment dispute, discrimination	456	175	257	96	60	21
Employment dispute, other	1759	86	222	78	28	114
Rental/lease agreement	218	47	101	96	24	319
Tortious interference	2193	143	98	118	32	62
Partnership dispute	1504	127	24	197	87	20
Other or unknown contract	738	83	138	211	38	85
Subrogation	45	9	22	29	8	23
Other or unknown real property	114	4	84	263	24	114
Unknown	400	49	22			0
Total	830	50	12 076	207	29	4043

Note: Unweighted results reported.

Table 1 shows that damages awards vary substantially by case category. Median compensatory damages awards are highest in cases involving products liability and medical malpractice. Median awards in these case categories are substantial multiples of the median award in automobile accident cases. The levels of awards are generally modest and are consistent with other comprehensive studies of particular jurisdictions.

Reliability of Compensatory Awards

In studies that include independent assessments of the defendant's behavior, damages awards correlate with the nature and degree of the wrong. Studies of medical malpractice cases show that those who behaved the worst and caused the most harm are likely to pay damages and likely to pay the most damages (Cheney et al., 1989; Cranberg et al., 2007; Farber and White, 1991; Peeples et al., 2002; Sloan et al., 1993; Spurr and Howze, 2001; Taragin et al., 1992). Studies of securities law cases also find that case outcomes correlate with the merits of claims (Cox et al., 2008). Experimental studies confirm these findings (Wissler et al., 1999). Osborne (1999) uses a different methodology but reaches similar results. His study includes many substantive areas of law and examines the relation of court awards to the pretrial expectations of litigants and their attorneys and to measurable, economically relevant damages. He finds that court awards are highly predictable, as variance in expectations explains much of the variance in awards. The awards also significantly relate to both medical costs and property damage. He rejects the hypothesis of a highly unpredictable court system.

Compensatory Damages Awards – Time Trends

Table 2 allows study of award trends over time. A popular perception exists that damages awards increase substantially over time, and some empirical studies support the claim. But these studies have been questioned on methodological grounds (Vidmar, 1998). Table 2 shows little systematic evidence of increasing trial awards. A 40-year longitudinal study of awards within a jurisdiction showed no real increase in award levels over time (Seabury et al., 2004).

Table 2 is limited to state court cases. Awards in federal court cases tend to be higher, at least partially because of jurisdictional amount requirements in federal court. Median trial awards in federal court products liability cases did increase from the early 1980s to the mid-1990s. The increase also occurred in judge-tried cases and in contracts cases.

Table 2 Mean and median compensatory award 1991 to 2005, state court trials, in thousands of 2005 dollars

	Mean	*Median*	N	*Mean*	*Median*	N
Year	*Compensatory award, juries*			*Compensatory award, judges*		
1991	510	61	3053	Data on judges not gathered		
1996	1351	58	3005	197	32	1377
2001	737	42	3086	205	26	1204
2005	723	40	2954	218	29	1470
Total	829	50	12 098	207	29	4051

Note: Unweighted results reported.

Punitive Damages

Punitive or exemplary damages are sometimes awarded in addition to the monies given to make the victim whole. But normally punishment is left to the state, not to private civil litigation. Thus, many legal systems do not allow damages beyond those necessary to compensate the victim. For example, little authority to grant punitive damages exists in Roman law and in some of the continental European systems that derived from it. But many countries allow punitive damages or their equivalent (exemplary damages, moral damages) in some classes of cases. Recently, the European Court of Justice has ruled that national courts must be willing to award exemplary damages pursuant to claims or actions founded on European Community law, if such damages may be awarded pursuant to similar claims based on domestic law. Nevertheless, punitive damages are substantially less important outside the United States.

The most discussed questions in the recent United States literature on damages relate to punitive damages. Understandably skewed media coverage emphasizing high awards, combined with advocacy by lobbying groups, has generated enormous pressure on punitive damages law and has helped to shape professional and public impressions. Business lobbying groups have enjoyed some success, with several states enacting provisions raising the burden of proof for punitive damages claims, placing monetary ceilings on awards, and otherwise limiting the likelihood and size of punitive awards.

Social science research on punitive damages expanded as a result of the 1994 jury award of $5 billion in punitive damages against Exxon Corporation (now Exxon Mobil) for the 1989 Exxon Valdez oil spill. Exxon-funded several leading academics, which produced an impressive body of work that is critical of punitive damages and supports 'reform.' But some of the work is inconsistent with prior theoretical discussions of punitive damages (Eisenberg, 1998; Vidmar, 1998), some of the work is methodologically unsound (Eisenberg and Wells, 2006), and none of the work seeks to reconcile its claims of incoherent punitive damages awards with real-world evidence of their coherence (Eisenberg et al., 2002). In 2008, the United States Supreme Court reduced the punitive damages award to $500 million but expressly refused to rely on the social science research funded by Exxon and noted that claims of out-of-control punitive damages awards were refuted by empirical evidence (Exxon Shipping Co. v. Baker, 2008).

Rate of Punitive Awards

Several large studies of punitive damages awards, such as Moller (1996), Daniels and Martin (1995), data commissioned by the BJS (1995, 2004, 2008), and Eisenberg et al. (2006), cover multiple United States jurisdictions and include case categories that are sufficiently well defined to warrant comparison. Each of the sources reveal overall punitive damages award rates of less than 10% in successful jury trials. In two of the areas of greatest interest to policymakers, products liability and medical

Table 3 Mean and median punitive award 1991 to 2005, state court trials, by trial mode, in thousands of 2005 dollars

	Mean	*Median*	N	*Mean*	*Median*	N
Year	*Punitive award, juries*			*Punitive award, judges*		
1991	749	702	177			0
1996	2418	65	121	743	39	55
2001	6980	83	155	169	55	48
2005	2967	147	141	251	38	45
Total	3241	83	594	407	47	148

malpractice litigation, punitive awards are especially rare. Except in case categories dominated by fraud and employment law issues, all punitive award rate levels are likely less than 10%, and many are less than 5%.

For example, in the BJS data from 2005, punitive damages were sought in about 10% of the cases that reached trial, and were awarded in about 30% of the cases won by plaintiffs in which they were sought (Eisenberg et al., 2011). These social scientific analyses undermine the belief that punitive damages are frequently awarded.

Level of Punitive Damages Awards

Large, publicized punitive awards, such as that against Exxon, create the impression that punitive awards are on average much larger than they truly are. BJS data allow one to assess the level of punitive damages awards. Table 3 summarizes punitive damages awards, in thousands of 2005 dollars, for the years 1991, 1996, 2001, and 2005. Median punitive damages awards have been more modest than the headline-grabbing awards. Mean punitive awards are substantially affected by extremes. Excluding the most extreme punitive award in the 1991 and 1996 reduces the mean by over 25% in one year and by over 50% in the other. All studies confirm that the bulk of punitive awards are modest.

Relation between Punitive and Compensatory Awards

A criticism of punitive damages is that they bear no relation to the compensatory award – the punishment inflicted therefore is not proportional to the harm caused. The United States Supreme Court has indicated that a punitive damages award normally should bear a reasonable relation to the compensatory award. Social scientific analysis suggests that punitive awards do relate to compensatory awards. Figure 1 shows the relation (in logs) between compensatory and punitive awards in the BJS studies, by year and adjudicator. It shows an unmistakable correlation. The constancy of the relation indicates that the pattern is not a consequence of cases in a particular year or of cases tried before a particular adjudicator.

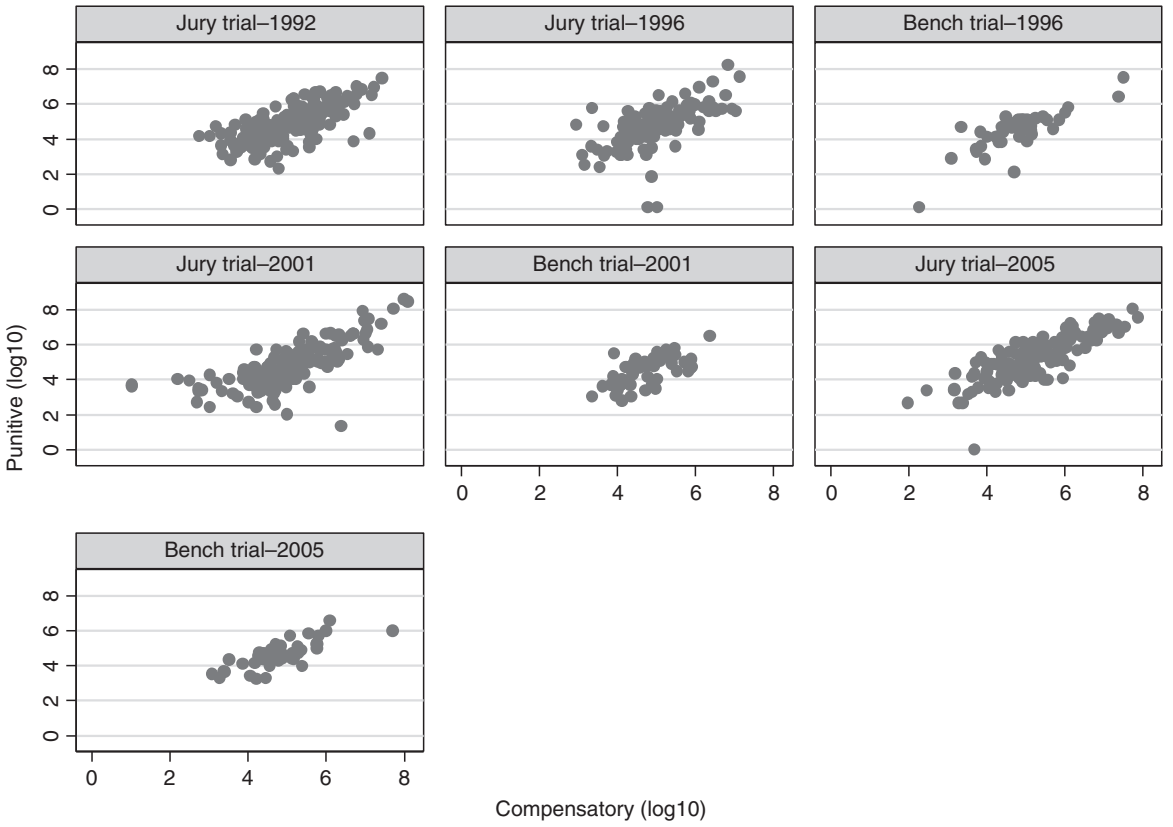

Figure 1 The relation between punitive and compensatory damages over time, by adjudicator.

Punitive damages awards are thus rarer, smaller, and more closely related to compensatory awards than is widely believed.

Punitive Damages Awards – Time Trends

Punitive damages, like compensatory damages, are often said to be ever increasing in number and size. The BJS data allow one to compare the rate and size of punitive awards. Table 3 above shows that median punitive awards were substantially lower in 1996 than 1991. Mean awards were higher in 1996, an increase that is a function of a few large cases. Median awards increased from 1996 to 2005 but mean awards first increased and then decreased. Overall there is no consistent pattern of increase. Table 3 also shows no evidence of consistent increases in the frequency of punitive damages awards over time. Thus, even minimal social scientific analysis provides no support for the belief that punitive damages awards are ever increasing in size or frequency.

The Importance of the Adjudicator

It is widely believed that the choice of adjudicator – judge, jury, arbitrator – substantially affects the damages awarded. Higher and more variable awards in the United States are attributed to the greater role of jurors as adjudicators. While real differences exist across adjudicator groups, the social scientific evidence supports skepticism about claims of massive differences across adjudicators. The classic Kalven and Zeisel study used questionnaires to presiding judges in about 4000 actual civil jury trials in the 1950s (Kalven and Zeisel, 1966). Their data showed a 78% agreement rate between judge and jury on liability, with no revealing pattern in the cases of disagreement. The rate of judge–jury agreement is high compared to other human endeavors. They also showed that jurors awarded on average 20% more than judges said they would have awarded. The modest judge–jury differences reduced interest in attacks on juries.

Renewed attacks on the jury's competence to assess damages, and especially punitive damages, reemerged in the 1980s and 1990s. Research using interviews and experiments find that different decisionmakers are quite similar and predictable with respect to their judgments of injury severity (Wissler et al., 1999). Such research also finds that jurors' awards lose consistency in translating severity of injury into dollar awards. Other studies suggest somewhat lesser judge–jury differences. Vidmar and Rice (1993) report results of experiments in which subjects were jurors awaiting jury duty and legal professionals. The outcome of interest was the level of pain and suffering awards. In a medical malpractice burn case, they found that six-person juries would give more variable awards than 12-person juries but that both groups would be less variable than individual legal professionals. Their second experiment involved an automobile accident case and a medical malpractice case. They found no differences between jurors and legal professionals.

Studies of actual cases, which lack the controls of experimental settings but provide the relevant setting, suggest modest adjudicator differences. Osborne's (1999) study compares the predictability of judge and jury awards based on litigant expectations. He finds that one can reject the hypothesis that jury trials are largely random events. Litigant expectations explain more than a third of the variance in awards. And he finds that the behavior of judges is not significantly more predictable than that of juries. Helland and Tabarrok (2000) conclude that most differences between judge and jury awards can be explained by differences in the sample of cases coming before judges and juries but that some suggestive differences exist. Eisenberg et al. (2006) find little meaningful difference in the pattern of judge and jury punitive damages awards.

Exxon-funded research finds experimental evidence of vast judge–jury differences. For example, Hastie and Viscusi (1998) report mock jurors to be much more likely than judges to impose punitive sanctions in a hypothetical railroad accident case. Vast judge–jury differences on liability in these studies, which contrast with Kalven and Zeisel's and other findings, calls into question the experimental designs' ability to detect socially meaningful punitive damages differences between judges and juries.

Overall Assessments and Comparisons of Damages Awards

Greater injuries lead to greater damages awards in all systems for which data are available. With respect to variability in the United States, where variability is regarded as greatest, from half to almost three-fourths of awards' variation can be accounted for by measures of injury severity, as well as by other predictors, but a substantial random component also exists. Legal systems that rely on schedules of damages for various injuries have lower variability though not necessarily more complete compensation.

Prospects for definitive social scientific comparison of damages rules' operation are dim. Some inherent limitations apply to comparing performance of legal rules across state and national boundaries. The number of differing legal systems will tend to be small – there are, for example, only 50 states, and gathering useful damages information about that many countries is an enormous task. Important differences, such as varying reliance on social compensation systems, different mixes of cases reaching court systems, different methods for compensating attorneys, and varying standards of living and purchasing power need to be considered. For the foreseeable future, reasoned analysis of data available for individual jurisdictions and informal comparison across jurisdictions will have to suffice in assessing performance of the damages component of remedial systems.

See also: Deterrence Theory: Crime; Deterrence; Deterrent Effect of Police and Prisons; Litigation; Punishment: Social and Legal Aspects; Torts in Law.

Bibliography

Bureau of Justice Statistics Bulletin: Civil Justice Survey of State Courts, 1999. Civil Trial Cases and Verdicts in Large Counties, 1996.

Bureau of Justice Statistics Bulletin: Civil Justice Survey of State Courts, 2004. Civil Trial Cases and Verdicts in Large Counties, 2001.
Bureau of Justice Statistics Special Report, 1995. Civil Justice Survey of State Courts, 1992: Tort Cases in Large Counties.
Bureau of Justice Statistics Special Report, 2008. Civil Justice Survey of State Courts, 2005. Civil Bench and Jury Trials in State Courts, 2005.
Cheney, F.W., Posner, K., Caplan, R.A., Ward, R.J., 1989. Standard of care and anesthesia liability. Journal of the American Medical Association 261, 1599–1608.
Cox, J.D., Thomas, R.S., Bai, L., 2008. There are plaintiffs and... There are plaintiffs: an empirical analysis of securities class action settlements. Vanderbilt Law Review 61, 355–386.
Cranberg, L.D., Glick, T.H., Sato, L., 2007. Do the claims hold up? A study of medical negligence claims against neurologists. Journal of Empirical Legal Studies 4, 155–162.
Daniels, S., Martin, J., 1995. Civil Juries and the Politics of Reform. Northwestern University Press, Evanston, IL.
Eisenberg, T., 1998. Measuring the deterrent effect of punitive damages. Georgetown Law Journal 87, 347–357.
Eisenberg, T., 2009. U.S. Chamber of Commerce Liability survey: inaccurate, unfair, and bad for business. Journal of Empirical Legal Studies 6, 969–1002.
Eisenberg, T., Miller, G.P., 2014. Damages versus specific performance: lessons from commercial contracts. Journal of Empirical Legal Studies 11.
Eisenberg, T., Wells, M.T., 2006. The significant association between punitive and compensatory damages in blockbuster cases: a methodological primer. Journal of Empirical Legal Studies 3, 175–195.
Eisenberg, T., Rachlinski, J.J., Wells, M.T., 2002. Reconciling experimental incoherence with real-world coherence in punitive damages. Stanford Law Review 54, 1239–1271.
Eisenberg, T., Heise, M., Waters, N.L., Wells, M.T., 2011. The decision to award punitive damages: an empirical study. Journal of Legal Analysis 2, 577–620.
Eisenberg, T., Goerdt, J., Ostrom, B., Rottman, D., Wells, M.T., 1997. The predictability of punitive damages. Journal of Legal Studies 26, 623–661.
Eisenberg, T., Hannaford-Agor, P.L., Heise, M., LaFountain, N., Munsterman, G.T., Ostrom, B., Wells, M.T., 2006. Juries, judges, and punitive damages: empirical analyses using the civil justice survey of state courts 1992, 1996, and 2001 data. Journal of Empirical Legal Studies 3, 263–295.
Exxon Shipping Co. v. Baker, 2008. 554 U.S. 471.
Farber, H.S., White, M.J., 1991. Medical malpractice: an empirical examination of the litigation process. RAND Journal of Economics 22, 199–217.
Flatscher-Thöni, M., Leiter, A.M., Winner, H., 2013. Pricing damages for pain and suffering in court: the impact of the valuation method. Journal of Empirical Legal Studies 10, 104–119.
Hastie, R., Viscusi, W.K., 1998. What juries can't do well: the jury's performance as a risk manager. Arizona Law Review 40, 901–921.
Helland, E., Tabarrok, A., 2000. Runaway judges? Selection effects and the jury. Journal of Law Economics Organization 16, 306–333.
Kalven, H., Zeisel, H., 1966. The American Jury. University of Chicago Press, Chicago, IL.
Landrigan, C.P., Parry, G.J., Bones, C.B., Hackbarth, A.D., Goldmann, D.A., Sharek, P.J., 2010. Temporal trends in rates of patient harm resulting from medical care. New England Journal of Medicine 363, 2124–2134.
Law Commission for England and Wales, 1994. Personal Injury Compensation: How Much Is Enough? A Study of the Compensation Experiences of Victims of Personal Injury. Law Com. No. 235. HMSO, London.
McIntosh, D., Holmes, M., 2003. Personal Injury Awards in EU and EFTA Countries. Kluwer Law International, The Hague.
Mello, M.M., Studdert, D.M., Thomas, E.J., Yoon, C.S., Brennan, T.A., 2007. Who pays for medical Errors? An analysis of adverse event costs, the medical liability system, and incentives for patient safety improvement. Journal of Empirical Legal Studies 4, 835–860.
Moller, E.K., 1996. Trends in Civil Jury Verdicts Since 1985. RAND, Santa Monica, CA.
National Health Care Quality Report, 2007. Agency for Healthcare Research and Quality, Rockville, MD.
National Health Care Quality Report, 2009. Agency for Healthcare Research and Quality, Rockville, MD.
Osborne, E., 1999. Courts as casinos? an empirical investigation of randomness and efficiency in civil litigation. Journal of Legal Studies 28, 187–203.
Peeples, R., Harris, C.T., Metzloff, T.B., 2002. The process of managing medical malpractice cases: the role of standard of care. Wake Forest Law Review 38, 877–902.
Pfennigstorf, W., 1991. A Comparative Study of Liability Law and Compensation Schemes in Ten Countries and the United States. Insurance Research Council, Oak Brook, IL.
Seabury, S.A., Pace, N.M., Reville, R., 2004. Forty years of civil jury verdicts. Journal of Empirical Legal Studies 1, 1–25.
Sloan, F.A., Githens, P.B., Clayton, E.W., Hickson, G.B., Gentile, D.A., Partlett, D.F., 1993. Suing for Medical Malpractice. University of Chicago Press, Chicago.
Spurr, S.J., Howze, S., 2001. The effect of care quality on medical malpractice litigation. Quarterly Review of Economics and Finance 41, 491–513.
Stoll, H., 1983. Consequences of liability: remedies. In: Tunc, A. (Ed.), International Encyclopedia of Comparative Law 11. J. C. B. Mohr, Tübingen, Germany.
Taragin, M.I., Willett, L.R., Wilczek, A.P., Trout, R., Carson, J.L., 1992. The influence of standard of care and severity of injury on the resolution of medical malpractice claims. Annals of Internal Medicine 117, 780–784.
Vidmar, N., 1998. The performance of the American civil jury: an empirical perspective. Arizona Law Review 40, 849–899.
Vidmar, N., Rice, J., 1993. Assessments of noneconomic damage awards in medical negligence – A comparison of jurors with legal professionals. Iowa Law Review 78, 883–911.
Wissler, R.L., Hart, A.J., Saks, M.J., 1999. Decisionmaking about general damages: a comparison of jurors, judges, and lawyers. Michigan Law Review 98, 751–826.

Remote Sensing

Assaf Anyamba, NASA/Goddard Space Flight Center and Universities Space Research Association, Biospheric Sciences Laboratory, Greenbelt, MD, USA
J Estes, K Kline, and E Collins[†], University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Estes, K. Kline, E. Collins, volume 19, pp. 13144–13150, © 2001, Elsevier Ltd.

Abstract

Remote sensing is one of a suite of geospatial technologies that are having a growing impact in a wide variety of areas from commerce to science to public policy. The field of remote sensing evolved from the interpretation of aerial photographs to the analysis of satellite imagery, and from local area studies to global analyses, with advances in sensor system technologies and digital computing. Today remote sensor systems can provide data from energy emitted, reflected, and/or transmitted from all parts of the electromagnetic spectrum. Examples of applications of these data include population and demography studies, study of archaeological sites, energy studies using hydrological models, urban planning, environmental monitoring, environments, treaty enforcement, land use/land cover planning, weather forecasting, and agricultural production estimation, just to name a few. The material that follows provides a brief overview of the historical development of remote sensing, emphasizing the increasing complexity of platforms, systems, and tasks.

Remote sensing is one of a suite of geospatial technologies that are having a growing impact in a wide variety of areas from commerce to science in public policy. While remote sensing's origins and technological advance are an outgrowth of military and intelligence applications, commercial and scientific applications of the technology have been growing. The field of remote sensing evolved from the interpretation of aerial photographs to the analysis of satellite imagery, and from local area studies to global analyses, with advances in sensor system technologies and digital computing. Today remote sensor systems can provide data from energy emitted, reflected, and/or transmitted from all parts of the electromagnetic spectrum. These data can then be transformed into information products either by manual or machine-assisted interpretation and employed in an analytical or management decision-making context. Examples of applications of these data include population and demography studies, study of archaeological sites, energy studies using hydrological models, urban planning, environmental monitoring, environments, treaty enforcement, land use/land cover planning, weather forecasting, and agricultural production estimation. This article provides a brief overview of the historical development of remote sensing, emphasizing the increasing complexity of platforms, systems, and tasks.

The definition of remote sensing used here, provided by the American Society for Photogrammetry and Remote Sensing, is as follows:

> In the broadest sense, the measurement or acquisition of information of some property of an object or phenomena, by a recording device that is not in physical or intimate contact with the object or phenomenon under study; e.g., the utilization at a distance (as from aircraft, spacecraft, or ship) of any device and its attendant display for gathering information pertinent to the environment, such as measurements of force fields, electromagnetic radiation, or acoustic energy. The technique employs such devices as the camera, lasers, and radio frequency receivers, radar systems, sonar, seismographs, magnetometers, and scintillation counters.
> (Reeves et al., 1975)

[†] Deceased

What follows is a brief summary of various sensor systems within the context of an historical background of remote sensing, and a short discussion of the future of remote sensing with some reference to social and behavioral sciences.

Sensor Systems

Many systems are used to detect energy patterns, which provide information about the Earth's atmosphere, oceans, and land surfaces. These systems range from cameras in aircraft to advanced sensor systems in Earth-orbiting satellites. Most satellite remotely sensed data are acquired by sensor systems such as multispectral scanners, linear array devices, and synthetic aperture radar systems that operate in the visible, near infrared, thermal infrared, and microwave portions of the electromagnetic spectrum.

These remote sensor systems can be broadly categorized into passive systems and active systems. A camera is an example of a passive system as its film records light energy reflected from an existing source, the sun. Add a flash attachment to the camera and it becomes an active system, since it now provides its own source of illumination. Passive systems record solar radiation reflected from the Earth's surface. Data derived from passive systems such as multispectral scanners provide information on, among other things, land use, vegetation types, distribution and condition of landforms, soils, surface waters, and river networks. Active microwave (radar) systems are used commonly in geological and hydrological applications. Because active systems do not depend on reflected energy from the sun, they are able to acquire data during the day or night. In addition, because the atmosphere does not attenuate microwave radiation as much as shorter wavelength energy, radar systems allow data acquisition in cloudy and rainy weather. This ability to operate day or night, and to create images of the Earth's surface in spite of cloud cover, makes radar sensors particularly attractive tools for improving understanding of tropical and polar areas.

Thermal infrared data, which represent a record of emitted energy from surfaces, have been particularly useful in

monitoring fires and land surface moisture conditions and in improving understanding of volcanic and geothermal activity. Ocean surface temperatures are related to the dynamics of coastal waters and ocean currents. Over land, plant water stress induces changes in canopy temperatures, which can be detected by thermal sensors. Thermal imagery from satellites has been used to record the nighttime lights of the globe, thereby indicating the spatial distribution of population. Other remote sensing systems detect the earth's magnetic and gravitational fields; these tools are used extensively in oil and mineral exploration and more recently ground water storage. Employing imagery from more than one portion of the electromagnetic spectrum can increase the information derived from an analysis, and is typically referred to as multispectral image analysis. Analyses of such multispectral data can, if properly designed, increase both the quantity and quality of information for given applications.

Image Interpretation

Interpretation of remotely sensed data is part science and part art. Manual interpretation of remotely sensed data is fundamental to all image interpretation. Even with the ever-increasing sophistication of machine-assisted image interpretation software, people must ultimately label the feature or phenomenon being identified. Not all individuals have equal interpretive skills with respect to aerial photographs or images acquired by remote sensor systems. It is axiomatic, however, that the more one knows about the elements, techniques, and methods of image analysis the better an interpreter one will be. In addition, the more one knows about the systems that acquire the data one is interpreting and the subject or discipline area most closely related to the interpretation, the better an interpreter one will make. It is seldom that an interpreter uses only the information in the imagery being analyzed as the basis for an interpretation. Ancillary data in the form of interpretation keys, maps, diagrams, textual material, and other forms of site-specific literature often aid in interpretations. Fieldwork is also a key element in image interpretation, and the collection of what is termed 'ground truth' should be accomplished whenever possible. Such ground truth validation is now increasingly assisted by around the clock digital cameras mounted strategically at field sites.

Resolution

Resolution is at the core of the successful use of remote sensing data. Resolution, as typically applied by a user, normally involves the question 'can I identify what I need to identify on this imagery?' The answer involves the use of the term ground resolved distance (GRD). GRD is used to describe the smallest object that can be unambiguously identified on the ground given good object-to-background contrast. Object-to-background contrast is important here. While it may be possible to see white lines on a black road on one set of images, it may not be possible to see the same size white lines on a concrete stretch of the same road. Resolution is thus dependent upon many things.

Indeed, the level of matching between the spatial, spectral, and temporal resolutions of the measurements and needs of the investigation will determine the usefulness of the derived data/information. Spectral range and resolution as used here refer to the portion of the electromagnetic spectrum being used in the measurement. It must be appropriate to the questions being addressed, for example, thermal range for fire detection, visible or infrared for vegetation health and condition. Spatial resolution determines the level of detail that can be extracted concerning objects in a given scene. As a general rule, it must correspond to the sizes of typical objects that must be separately identified. A resolution of a meter or less is needed to identify a single tree, a resolution of 5 m for a single-family residence, a resolution of 50 m might be sufficient for identification of most agricultural fields in North America, while a resolution of 500 m to a kilometer may be enough for studying ocean surface features. A necessary compromise between area to be covered and ground resolution of the measurement must take into account the volume of data that will be generated by, say, a large area coverage – millions of square kilometers – using high resolution data – 1–30 m.

The frequency of the data acquisition must also match the natural frequency of change in the landscape and the physical, social, or cultural phenomena of interest. For traffic studies, time frames of minutes can be important. Hourly observations might be of interest for certain meteorological conditions. Daily observations may be needed for an assessment of plant evapotranspiration associated with irrigated agricultural crops, while observation once every year could be sufficient for land cover change assessment.

Geographic Information Systems

The synergy between geographic information systems (GIS) and remote sensing comes into play here. To be interpreted accurately, remotely sensed data are often supplemented with other data. Often, these ancillary geospatial data can be found or included in a GIS for analysis. But to be valuable in decision-making contexts, GIS data layers should be as up-to-date as is practical. Remotely sensed data are a key technology for updating many types of GIS data. Thus, when environmental planners, resource managers, and public policy decision makers want to measure, map, monitor, or model future scenarios in order to facilitate better management decision making, remote sensing is being increasingly employed within the context of a GIS as a decision support system.

Historical Development

In the most general sense, the field of remote sensing begins with the invention of photography. Joseph Nicéphore Niepce took the first photograph of nature in 1827. It is a picture of the courtyard of his home in the village of Saint Loup de Varenne, France, with an exposure time of 8 h. In 1829, Niepce and Louis M. Daguerre signed a partnership agreement to work on a practical method of taking photographs. Niepce had been working on what he called heliography, or sun drawing, and Daguerre on dioramas, which he constructed with the aid of a camera obscura. The process they pioneered was called direct

positive image. The result of their collaboration, called the daguerreotype, was finalized in 1839. Also in 1839, an Englishman named William H. Fox-Talbot and a member of the Royal Academy invented the calotype (from the Greek *kalos* for beautiful and *typos* for impression). Talbot's positive–negative process is essentially the same as that used today, yet at the time the images produced by the two processes were not comparable.

In 1858, Gaspar Felix Tournachon, also known as Nadar, a Parisian photographer and balloonist, took the first known aerial photograph from a captive balloon at an altitude of 1200 feet over Paris. During the late 1870s, German foresters used aerial photography taken from balloons to map and measure the aerial extent of tree species. This is an example of the earliest known application of both aerial photographic interpretation and photogrammetry (the science of making precise measurements from remotely sensed data).

The use of rockets to launch cameras first occurred in 1906 when Albert Maul launched a rocket propelled by compressed air that was equipped with a camera. The camera took a picture from an altitude of 2600 feet and the camera was then returned to the ground by parachute. In 1909, Wilbur Wright took one of the first photographs from an airplane in Centrocelli, Italy. By 1915, cameras devoted to aerial photography were being produced. World War One saw an explosion in the use of aerial photographs. During the Meuse Argonne Offensive some 55 000 aerial photographs were taken and delivered to Allied Expeditionary Forces over a 4-day period. During the war, thousands of photo interpreters were trained. These interpreters laid the groundwork for the applications of aerial photographic interpretation that began to expand after the end of World War One.

In the early 1920s, the first books and articles on the applications of aerial photo interpretation began to be published. Lee's (1922) *The Face of the Earth as Seen From the Air* 'shows familiar scenes from a new angle,' and contains chapters on mountain features, coastal mud flats, plains, and submerged landforms. Joerg's (1927) 'The use of the air plane in city geography,' published in the *Annals of the Association of American Geographers*, focused on the application of aerial photography for urban studies. In 1927, Bergen's 'Aerial surveys for city planning' came out in the *Transactions of the American Society of Civil Engineers*, and in 1930 the American Geographical Society published the work by Platt and Johnson, '*Peru from the Air*.' While providing a regional overview of the country, this work also focuses on archeological sites. Of a more general nature is the work by Fairchild (1927) in the May issue of the *Annals of the American Academy of Political and Social Sciences* entitled, 'Aerial photography, its development and future,' which discusses potential applications of aerial photography from highway planning to water supply management. During this period, however, by far the most attention was given to the uses of aerial photography in oil exploration, structural and engineering geological analyses, land feature mapping, and soils mapping.

In the late 1920s and early 1930s, agencies of the US Government began to make significant use of aerial photography for operational applications. At this time, the Agricultural Adjustment Administration began to systematically photograph farm and ranch lands across the entire United States. This use of aerial photography for documenting agriculturally active areas has continued at 5- to 10-year intervals. The program now called National Agriculture Imagery Program (NAIP) is mandated to acquire peak growing season 'leaf-on' imagery and deliver this imagery to US Department of Agriculture (USDA) County Service Centers in order to maintain the common land unit boundaries and assist with farm programs. The goal of NAIP is to collect 1 m imagery for the entire conterminous United States. The imagery is either natural color or 'four band' imagery, and is delivered in the year of acquisition. This historical record has been significant in many types of studies, particularly in legal actions where the documentation of land use and/or land cover conditions and changes through time are in question. At the same time, the US Forest Service began using aerial photography to document much of the nation's timber reserves. Similar studies began in other nations as well, notably Canada. The US Geological Survey began to photograph many areas of the country in order to make topographical and geological maps based upon photogrammetric techniques and interpretation. The Tennessee Valley Authority and other regional agencies began to use aerial photography for planning purposes, as did state and metropolitan planning agencies. The *Master Plan of Residential Land Use*, published in 1943 by the Chicago Planning Commission, was based largely on data interpreted from aerial photographs.

World War Two led to developments in aerial photography, with a special interest in new and higher resolution film emulsions and thermal infrared imagery. Kodak patented the first false-color infrared sensitive film in 1946. The 1950s saw developments in technology that allowed multispectral imagery to be taken. Thermal infrared and side-looking airborne radar systems began to be acquired by the US military. Around this time, research personnel at the US Office of Naval Research coined the term 'remote sensing' to cover the expanding range of sensor systems and technologies for acquiring images of the world around us. This was the same year that the Soviet Union launched Sputnik, the first Earth-orbiting satellite. Civilian use of remotely sensed data during this period produced some of the first studies on the application of aerial photography for population estimation.

As early as 1946, the US Government began to pursue an operational space-based imaging system when the US Air Force asked the RAND Corporation to explore the possibility of launching an object into orbit. An early experiment in satellite land remote sensing was conducted from the Explorer 6 platform in 1959. The synoptic images of Earth's weather patterns acquired by the Television and Infrared Observational Satellite, launched in April of 1960, captivated the American public and demonstrated the capability and necessity of earth observation from space. Also in 1960, the first experimental spy satellite, a component of what would later become known as the Corona Program, was launched and, shortly thereafter, classified as secret. These data were withheld from the public until 1995, when the imagery and documentation were released to the public under Executive Order 12951. In 1965, National Aeronautics and Space Administration (NASA) initiated the Earth Resources Survey (ERS) program. This early program sought to develop methods and applications for remote sensing. The Department of Agriculture researched the use of

remote sensing for geological, hydrological, geographical, and cartographical applications. In 1966, the Department of Commerce began to participate in the ERS program by establishing the Environmental Sciences Group within the Environmental Sciences Administration (the predecessor to National Oceanographic and Atmospheric Administration (NOAA)). Also in 1966, the Department of the Interior initiated the multiagency Earth Resources Observation Satellite program, which studied the potential of Earth-observing satellites for natural resources applications.

During the 1960s and into the mid-1970s, NASA and a number of other Federal Agencies focused a significant amount of attention on applied remote sensing research. This was, in part, aimed at justifying the need for civil land satellite remote sensing systems and cataloging the variety of fundamental and applied science questions that could be addressed by such systems. During this period, efforts began to develop computer-assisted image interpretation techniques. Results of these activities, and the state-of-the-art practice of remote sensing during this time period, are documented in the American Society of Photogrammetry's first edition of the *Manual of Remote Sensing* (1975), as well as other sources.

The first Earth Resources Technology Satellite (ERTS-1) was designed originally as a research tool. When it was launched in July of 1972, however, this satellite ushered in a brief era of emphasis by NASA on the demonstration of operational applications using this series of satellites. Through 1974, ERTS-1 (later renamed Landsat 1) transmitted over 100 000 terrestrial images covering 75% of the Earth's land surface. Between 1972 and 1978, major research efforts utilizing this new civil land remote sensing data were conducted in agriculture, forestry, water resources, and geology. Other notable regional-scale applications research in the 1970s included the Large Area Crop Inventory Experiment, Agricultural and Resource Inventory Surveys through Aerospace Remote Sensing, and the Western Snow Melt Application System Verification Test.

In 1975, Landsat 2 was launched and the Ford Administration approved budget appropriations for Landsat 3, which was launched in 1978. In March of 1979, President Carter announced the Administration's commitment to maintaining the continuity of civilian terrestrial satellite remotely sensed data and recommended the transition of the program to the private sector. This initiated the process of program commercialization that would come to fruition in the mid-1980s. In 1981, the Reagan Administration accelerated the pace of Landsat commercialization by essentially rejecting the Carter Administration's commitment to an operational civilian land remote sensing program and advocating the immediate end of government funding for the Landsat Program.

By the middle of the 1980s, it was clear that satellite remote sensing could provide globally consistent, spatially disaggregated, and temporally repetitive coverage of the land surface of the Earth. Landsat 4, with the first Thematic Mapper instrument, was launched in 1982. Higher spatial resolution multispectral imagery was now publicly available and could be applied by scientists and professionals to the assessment of terrestrial ecosystems and land cover. A number of researchers began to develop techniques to employ the spectral information from Landsat to measure green vegetation status, and phenology, as well as interannual variation in these parameters. These studies then led researchers to the development of vegetation measures that provide an index of condition and, in part, describe absorption of sunlight, photosynthetic capacity, and evaporation rates. With this research, the science community began to understand that space-based Earth imaging can and does provide information on the spatial and temporal patterns associated with land processes. During this period of capability demonstration, researchers noted that the possibility of developing fully integrated land–ocean–atmosphere monitoring and modeling capabilities were essentially realized.

In March 1984, Landsat 5 was launched and in July of that year, the US Congress provided the legislative authority to transfer the Landsat program to the private sector. Thus began the nearly decade long period of 'commercialization' of the US civil land remote sensing program. Commercialization (and the subsequent price increases for Landsat data) caused a shift in scientific research emphasis. Researchers began to move away from Landsat imagery and began to design experiments using NOAA polar orbiting meteorological satellites and their Advanced Very High Resolution Radiometer sensor system as a primary data source for scientific research. This shift was the result of increased costs, greater restrictions on sharing, and more limited availability of Landsat data to the basic and applied land remote sensing communities.

The US Government would not regain operational control of the Landsat program until the late 1990s. Although the law changed in 1992, government control of any orbital assets did not occur until Landsat 7 was launched in April 1999. The commercial pricing and data-sharing restrictions prior to this transition back to government pricing had restricted research and applications because of high data costs. Not until the twenty-first century did the Landsat program again come under the control of an operational land-oriented federal agency, the United States Geological Survey (USGS).

Despite these difficulties, research using Landsat data during this time (1984–92) included work in many disciplines and application areas. Organizations important in formulating the types of research conducted at this time include the Intergovernmental Panel on Climate Change, the International Council of Scientific Unions, the International Geosphere Biosphere Program, and the Human Dimensions of Global Change project. Particularly significant was research that validated the use of multitemporal Landsat and other data to monitor and assess tropical deforestation and changes occurring in every area of the world.

Other countries entered the space imaging arena with the launch of the French Système Probatoire de la Observation de la Terre (SPOT) satellite in 1986 and the Indian Remote Sensing Satellite in 1988. In 1990, the French launched SPOT 2; a year later the European Radar Satellite (ERS-1) was launched, aimed primarily at oceanographic applications, followed shortly by the Russian radar satellite Almaz; 1992 saw the launch of the first Japanese Earth Resources Satellite. In 1995, ERS-2 was launched and Canada launched Radarsat, with a primary objective of aiding navigation of shipping in the far north. In March 2002, Environmental Satellite (Envisat) was launched by the European Space Agency (ESA) with a suite of sensors to study atmospheric chemistry, ozone depletion,

biological oceanography, ocean temperature and color, wind waves, hydrology (humidity, floods), agriculture and arboriculture, and natural hazards interferometry. The mission lasted until May 2012.

The most recent advances at the end of the twentieth century involved the US launch of Landsat 7 with the Enhanced Thematic Mapper (ETM+) system, the Shuttle Interferometeric Synthetic Aperture Radar (IFSAR) mission, Earth Observing System (EOS)-1 (now named Terra) and EOS-2 (now named Aqua) currently in operation having gathered 10+ years of 250-m spatial resolution data. In addition, in March 1998, a consortium of European nations launched SPOT 4 satellite, the first truly global 1-km spatial resolution polar orbiting satellite, followed by SPOT 5 in May 2002. This mission delivers measurements specifically tailored to monitor vegetation and other land surface parameters with a frequency of about once a day. The role of SPOT-VEGETATION has been taken over by ESA's technologically advanced PROBA-V (Probing Vegetation) mission from June after the expiry of the VEGETATION mission. This marks the transition to a new set of sensors to be launched in the next several years. The 1990s also saw the entry of high-resolution commercial satellites to collect panchromatic images with 1-m resolution and multispectral imagery with 4-m resolution. The suite of commercial satellites known as GeoEye delivers data targeted at commercial and government sectors and is used mostly in intelligence and precision agriculture. Available since 2009 is commercial 22-m spatial resolution imagery from Astrium, targeted at natural resources and agricultural applications from Deimos-1 and Disaster Monitoring Constellation satellites. The Landsat-like data with a combined short repeat cycle (~4 days) is now being used by USDA's National Agricultural Statistics Service Information (NASS) to monitor agricultural growing conditions, improve agricultural land cover classification, and update the crop data layer on an annual basis. The Landsat ETM+ data are already proving useful in a variety of studies from fire fuels mapping to deforestation and urban change monitoring. IFSAR data are helping to improve the quality of topographical maps between about 60° N and 60° S of the Equator. Terra and Aqua data provide continental to global-scale land cover change information with a temporal frequency never before available. IKONOS and GeoEye data enable civil researchers to see features from space that could only be seen if aerial photography or if intelligence satellite data were available to the user.

These satellites are adding immeasurably to our ability to measure, map, monitor, and model our Earth system as never before. The launch of the Landsat Data Continuity Mission (LDCM), renamed Landsat 8, in February 2013 extends the almost 40-year Landsat data archive with image sufficiently consistent from legacy Landsat to allow long-term studies of regional and global land cover change. The two instruments onboard LDCM, the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), will acquire coincident images that will be well calibrated, coregistered, and archived together. This significant improvement of Landsat data will allow for the use of well-calibrated vegetation (from OLI) and land surface temperature (from TIRS) for comprehensive studies of land surface phenomena. At medium-to-coarse resolution the Suomi National Polar-orbiting Partnership launched in October 2011 continues NASA's EOS mission and provides striking new insights into many aspects of Earth, including its clouds, oceans, vegetation, ice, and atmosphere. This mission, the operational successor to Terra and Aqua and jointly operated by NOAA and NASA, enhances the continuity missions under the Polar-orbiting Operational Environmental Satellites umbrella in partnership with European Organization for the Exploitation of Meteorological Satellites, which constantly operate two polar-orbiting satellites. These polar missions will be continued under the next-generation Joint Polar Satellite System program between NOAA and NASA.

Unusual among remote sensing missions of the last decade is the Gravity Recovery and Climate Experiment (GRACE) launched on 17 March 2002. The GRACE mission consists of two identical spacecraft that detect changes in Earth's gravity field by monitoring the changes in distance between the two satellites as they orbit earth 220 km apart. This technique provides critical input to many scientific models used in oceanography, hydrology, geology, and related disciplines, and the derived data are now being used in a variety of applications including measuring the changing mass of polar ice caps, measuring changes in water resources on land, understanding sea level change resulting from ocean temperature and water mass changes, understanding the forces that generate Earth's geomagnetic field, and understanding internal Earth forces that move tectonic plates and result in earthquakes and volcanic eruptions. This along with other future missions including ESA's Sentinel mission and NASA's Global Precipitation Measurement will set new standards for mapping and monitoring facets of the land surface and precipitation environments, respectively, from space.

Future Directions

As of 2013, some 20 governments and organizations have launched civilian and commercial high-spatial resolution satellites in an attempt to capture a share of the growing market for remote sensing imagery. All nations are vested in understanding changes in their resource base and many seek to become global leaders in space imaging. Examples include the China–Brazil Earth Resources Satellite program, which has launched satellites since 1999, and the Indian Space Research Organization, which has launched earth resource satellites almost every 3–5 years since 1998. Japan, China, Pakistan, Argentina, South Korea, Thailand, Turkey, Indonesia, and Nigeria have aggressively pursued satellite remote sensing, with systems with spatial resolutions that range from several kilometers to less than a meter. Developments such as these suggest that a number of the long-standing restrictions that the military and intelligence communities placed on satellite remote sensing are being put aside as the technology becomes more accessible, creating greater transparency among nations of the world. With this increasing transparency comes the opportunity for both positive and negative uses of remotely sensed data. On the one hand, it may be easier to detect violations of international agreements, while on the other it may be easier for states to gain tactically important information on other states.

Remotely sensed data have attracted considerable interest among social, economic, and behavioral scientists. This interest stems from the realization that, although remotely sensed data cannot provide all the data/information needed to answer given questions, it may, when combined with other information, allow data to be more reliably interpreted across space and through time. Indeed, an important aspect of remote sensing is its value as a historical record; applications in legal proceedings involving environmental contamination and ownership liability are but one example. Historical data records from remotely sensed measurement are now required and critical for use in assessment of the impacts of a changing climate on natural resources. Other applications include estimating populations, providing famine early warnings and disaster assessments, inventorying industrial capacity, assessing transportation infrastructure, and monitoring environmental laws and treaties.

Still, users interested in questions in the physical and biological sciences have made more use of remotely sensed data than have users in the social, economic, and behavioral sciences. The reasons for this disparity are many and varied. They range from the inability of remote sensing to provide needed data/information to practitioners in the social, economic, and behavioral sciences to unwillingness on the part of users in these areas to take the time to learn the capabilities and limitations of the information that can be extracted from remotely sensed data. For some applications, the spatial and temporal scales of data acquisition are not appropriate; for others, the models to translate the data collected into meaningful information specific to a given need are yet to be developed. However, changes such as the emergence of distributed data centers, the growth of global Internet networks, and widespread cell phone technology are prompting the use of remotely sensed derived data and maps by all segments of society. Although, remote sensing cannot see what is in the human mind, it can document actions that people have taken in response to some (e.g., environmental) stimuli. For example, remote sensing cannot predict what crop a farmer will plant in a given year, but it can, given appropriate dates of image acquisition and timely interpretation, tell us what has been planted before a crop is harvested. Indeed, given adequate ancillary data and models, satellite imagery can indicate how good the farmer's harvest may be.

As the ability to collect data at higher spatial, spectral, and temporal resolutions increases, so too does the ability to derive useful information from remote sensing data. Remote sensing will never replace other, more traditional methods of data collection, such as census and survey instruments. Yet it is a proven, powerful method of collecting data, which, when properly interpreted and evaluated, can provide a wide variety of useful information. This information, when taken alone or when augmented by other data collected by more traditional means, can create useful data sets for social, economic, and behavioral scientists especially when analyzed together in GIS. The threats posed by climate change and the recognition that these threats are common to all nations have stimulated cooperation among nations in remote sensing missions, research, operations, and applications.

Organizations That Provide Remotely Sensed Data

- Astrium
- ESA
- US NASA
- USGS
- SPOT Image
- GlobeXplorer
- USDA
- US NOAA
- Space Imaging
- DigitalGlobe
- Vision on Technology

Bibliography

Bergen, E.T., 1927. Aerial surveys for city planners. Transaction of the American Society of Civil Engineers 90, 627–679.

Chicago Planning Commission, 1943. Master Plan of Residential Land Use. Chicago Planning Commission, Chicago.

Estes, J.E., Ehlers, M., Malingreau, J.-P., Noble, I., Raper, J., Sellman, A., Star, J.L., Weber, J., 1992. Advanced Data Acquisition and Analysis Technologies for Sustainable Development. UNESCO, Paris. MAB Digest 12.

Fairchild, S.M., May 1927. Aerial photography, its development and future. Annals of the American Academy of Political and Social Sciences 131.

Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next landsat satellite: the landsat data continuity Mission. Remote Sensing of Environment. http://dx.doi.org/10.1016/j.rse.2011.08.026.

Jensen, J.R., 1996. Introductory Digital Image Processing: A Remote Sensing Perspective, second ed. Prentice Hall, Upper Saddle River, NJ.

Jensen, J.R., 2000. Remote Sensing of the Environment: An Earth Resource Perspective. Prentice Hall, Upper Saddle River, NJ.

Joerg, E.L.E., 1923. The use of the air plane in city geography. Annals of the Association of American Geographers 13 (4).

Lee, W.T., 1922. The Face of the Earth as Seen from the Air: A Study in the Application of Airplane Photography to Geography. American Geographical Society, New York.

Liverman, D., Moran, E.F., Rindfuss, R.R., Stern, P.C. (Eds.), 1998. People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Washington, DC.

National Research Council, 1970. Remote Sensing with Special Reference to Agriculture and Forestry. National Academy of Sciences, Washington, DC.

Philipson, W. (Ed.), 1997. Manual of Photographic Interpretation, second ed. American Society for Photogrammetry and Remote Sensing, Bethesda, MD.

Platt, R.R., Johnson, G.R., 1930. Peru from the Air. American Geographical Society, New York.

Reeves, R.G., Anson, A., Landen, D. (Eds.), 1975. Manual of Remote Sensing. American Society of Photogrammetry, Falls Church, VA.

Relevant Websites

http://www.esa.int/Our_Activities/Operations/GMES_Sentinels – Global Monitoring for Environment and Security

http://pmm.nasa.gov/ – Global Precipitation Measurement

http://earthobservatory.nasa.gov/Features/GRACE/page3.php and http://www.nasa.gov/mission_pages/Grace/index.html – Gravity Recovery and Climate Experiment Mission

http://landsat.gsfc.nasa.gov/ – The LANDSAT Program

http://www.fsa.usda.gov/FSA/apfoapp?area¼home&subject¼prog&topic¼nai – National Agriculture Imagery Program

http://www.vgt.vito.be/ – Probing Vegetation

Renaissance

Bernd Roeck, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article discusses the uniqueness and exceptionality of the European Renaissance and the multitude of its implications for the emergence of the Modern World. Notwithstanding developments elsewhere, it was only in (Western) Europe where the richness of antiquity was harnessed fully and paved the way for revolutionary breakthroughs in discourse, science, and industry.

In this article the concept of Renaissance will be developed in the sense of a history of ideas, by pointing especially to Jacob Burckhardt's groundbreaking concept of the genesis of the individual, on the philosophy of Nietzsche and the racism of Gobineau, together with Walter Paters aestheticism. In this context, the social interrelations between the functions of the bourgeois Renaissancism (*bildungsbürgerlicher Renaissancismus*) against the background of the growing industrialized mass society will be looked at. This article follows the history of research: i.e., the revisions of Aby Warburg and the 'revolution of the medievalists' (Ferguson). The article then focuses on single aspects of historical research, such as the network of patronage, but more generally on the interrelations between historical fundamental processes. The description of the modern analysis of the Renaissance starts with the discussion of Hans Baron's 'civic humanism'. Additionally, new concepts are discussed. Still central is, in the author's view, the old debate about the genesis of a modern subjectivism and the genesis of the Renaissance in general. The article ends with 'Manierism' and the 'manieristic Renaissance'.

Two nineteenth-century historians were instrumental in giving the term 'Renaissance' the content that it retains to this day. In his 1855 *Histoire de France*, Jules Michelet (1798–1874) chose the word for the period of the *découverte de l'homme et du monde*, the sixteenth century. But the Swiss Jacob Burckhardt (1818–97) was even more important. His famous book *Die Cultur der Renaissance in Italien. Ein Versuch* (1860) became a milestone of research. Burckhardt sketched a monumental fresco that located the origin of modernity in fifteenth-century Italy – in conscious distinction from Hegel's philosophy of history, which posited that the German Reformation marked the beginning of the Modern Age. Burckhardt referred to the conditions in the Italian ministates as 'works of art', in which he saw the preconditions for the 'emergence of the individual' as the most important and far-reaching development of the Modern Age. He presented the Renaissance, with the many-faceted Humanistic movement as its intellectual substrate, as a period that forced the pace of modernization and secularization and as the beginning of the Modern World in which he himself lived and whose negative sides he analyzed perhaps more acutely than any other thinker of his time.

The Development of the Concept of the Renaissance

Renaissance

Burckhardt's thought-provoking image of the epoch exerted substantial influence, not only on the study of history, but also on literature and philosophy, in particular on Nietzsche's thinking. It remains the more or less explicit background of discussions of the term 'Renaissance', though it soon roused dissension. For example, in the book *Franz von Assisi und die Anfänge der Renaissance in Italien* (1885), Henry Thode (1857–1920) attempted a romantic reinterpretation, placing the *poverello* and the religious movement he inspired at the beginning of the epoch. Thode and his followers saw the Renaissance more as a reform movement on a medieval-clerical foundation, eliminating its character as an epoch of its own. Thode rejected the idea that classicism and rationality fundamentally constituted the Renaissance and divided it from the Middle Ages. He posited an individualism of piety and sensitivity to beauty, in contrast to Burckhardt's subjects becoming aware of themselves.

But the prospect of another, religious side of the Renaissance had already existed before Thode and did not completely contradict Burckhardt's view. In his influential essay collection *The Renaissance* (1873), Walter Pater (1839–94), an admirer of Ruskin presented several artists and thinkers he considered representative of the Renaissance. The essay on Botticelli became essential to the incomparable cult of this painter, which thrived throughout Victorian society, and not solely among the Pre-Raphaelites. Botticelli's art seemed to correspond to a 'metaphysical need' in the modern, industrialized world of Victorian England.

The inventions of the historians and literary writers stimulated broad segments of the educated bourgeois class. Florence became the preferred destination for an international cultural tourism. People built in the Renaissance style, surrounded themselves with splendid interiors in the style of the fifteenth and sixteenth centuries, and purchased original works of art. Especially in the second half of the nineteenth century, collections of important Renaissance art were amassed from Boston to Berlin. The 'Renaissance' was not only a focus of scholarly interest, but also part of a lifestyle.

Around 1900, the international Renaissancism reached a zenith. Burckhardt's classic was reprinted again and again, and

novels like Heinrich Mann's 'Göttinnen' depicted the sixteenth century as a foil for the contemporary 'decadent' *fin de siècle*. The closing nineteenth century saw itself not only in a strange spiritual kinship with the Renaissance: the fiction of a 'lascivious', 'unscrupulous' epoch paradoxically appeared to offer counterimages to the usual bourgeois existence, taking shape in princes of commerce, patrons, *condottieri*, and courtesans.

At this time, the term 'Renaissance' was loaded with a plethora of sometimes contradictory meanings, triggering associations of 'gold and purple', as Johan Huizinga said. The patronage of the Medici, the magnificence of the papal court, and the art of the Renaissance with its heroes Leonardo, Raphael, and Michelangelo provided new cult objects for religious worship in a disenchanted world. At the same time, the word recalled the art of power and *Realpolitik*, as Machiavelli had analyzed in his trailblazing work. Racist ideas of a victory of hard men and cunning women over decline and enfeeblement were inherent in this image of the epoch. A widely read literature in the succession of Carlyle, Count Gobineau, Nietzsche, and others popularized the cliché of the 'blond beast', the amoral 'Renaissance man' up to mischief.

The myth of the Renaissance thus offered models for Caesarism and imperialism. It was no coincidence that a nationally oriented writing of history took up Italy's *Signorie* and identified them with the precursors of the modern military and power-wielding state. But the iridescent construct of the 'Renaissance' also offered fragments for a historical foundation of modern civic societies. As early as the beginning of the nineteenth century, the Swiss Sismonde de Sismondi had described the Italian city republics in this way, and travelers from England and the USA thought they were tracing the ideal roots of their own nations when they came to Florence. It is hardly surprising that the inghlesi were the largest group among the tourists. The Florence of the Medici was long a primarily Anglo-American myth.

Examples of the ideological interpretation of the Renaissance with an unmistakable relationship to current history are also found in the works of Hans Baron, a German Jew who emigrated to the USA. In the essay he published in the middle of World War II, he saw the crisis of Italian 'freedom' caused by the aggression of the Visconti *signoria* in the early Quattrocento as the precondition for the formation of a modern, pluralistic state system on the Italian peninsula. In Baron's view, the high degree of citizen participation in Florentine politics was a precondition for this most important result of the struggle against the tyrants: He saw it as a kind of Italian America winning the upper hand over Visconti/Hitler (Baron, 1955).

The Revolution of the Medievalists

More recent research on the Renaissance has tried to free itself of mythologizing. Insight has been sought into the clientele systems of the clan that enabled Cosimo de' Medici and above all Lorenzo the Magnificent to exercise an authoritarian power behind a republican façade (Wackernagel, 1938; Thomas, 1995), whereby special attention was paid to the patrons commissioning the art (Esch and Frommel, 1995). But art history tended to go its own way. Wölfflin and others separated Renaissance art from the cultural history context and treated it primarily as a problem of style, and thus of aesthetics.

The conception of the art historian Aby Warburg (1866–1929) remained fragmentary and accessible only in a very few essays. It aimed once again at the great synthesis and still provides impetus for new research. He located the epoch in a long-term process of civilization that experienced its climax in the Apollonian beauty of the art of the High Renaissance. In the triumph of the Olympic gods, Warburg saw an important step toward overcoming 'oriental', 'medieval' irrationality. He was interested in the difficult breakthrough to the Enlightenment, humanity, and reason in the history of the West. Warburg saw works of art as documents of the states of the human psyche, and he believed that, through subtle analysis, he could uncover the *Zeitgeist* preserved in them. Warburg's works addressed aspects previously not or only marginally considered. The role of the patron became a theme, along with what Michael Baxandall (1972) calls 'troc': the market of intellectual and material goods where what we call 'Renaissance art' took shape. Warburg sought to illuminate the dark side of the epoch – its interest in divination, astrology, and magic, to which Burckhardt had devoted only a few passages. Equally important was that he conceived the emergence of Renaissance culture as a transfer process and thus examined not only the Orient, but also the culture of the North – the courts of Burgundy, France, and the Netherlands. This underscored what is now the general possession of research: the cultural physiognomy of the Renaissance as an amalgamation of varying influences. Italy remained important, but it was placed in the context of a network cast across the continent (Porter and Teich, 1992).

This geographical expansion, the focus on the 'occult Renaissance', and above all his constant question of the beginnings led Warburg to avoid drawing the boundary between the 'Renaissance' and the 'Middle Ages' too sharply. Although his works had no discernible importance for it, he prepared the ground for what Wallace K. Ferguson has termed a 'revolt of the medievalists' (Ferguson, 1948). This debate, characteristic of the 1920s, adopted aspects of the image of the Renaissance as sketched by Thode, Sabatier, and other researchers. It pointed out, for example, that the tradition of antiquity was never completely buried and recalled that various 'renaissances' were palpable between the ninth and thirteenth centuries, in which ancient forms clearly stand out as elements shaping culture. Some researchers, including Etienne Gilson and Johann Nordström saw the Renaissance as a mere continuation of currents existing in the twelfth century. In this famous book *The Waning of the Middle Ages* (1919), the Dutch writer Johan Huizinga drew an image of the fifteenth century diametrically opposed to Burckhardt's conception. Contemplating the courtly world of Burgundy, France, and the Netherlands, he depicted the epoch as a flowering 'autumn' of burgeoning color, rather than as the spring of a new age.

A Movement of the Happy Few: The Sociohistorical Dimension

Discussion thus ensued on the boundaries of epochs; it was the hour of the deperiodizers, who made the Renaissance their preferred target. George M. Trevelyan went as far as to locate the turn from the Middle Ages to the Modern Age in the epoch of the Industrial Revolution, the eighteenth century; and for other

researchers as well (like Delio Cantimori), the Renaissance of the fourteenth and sixteenth centuries lost its meaning as a significant turning point. A more differentiated sociohistorical research underscored that Renaissance societies were more 'medieval' than widely conceived. Additionally, this culture had been borne by a tiny minority – one of the reasons Ernst H. Gombrich no longer speaks of the Renaissance as an epoch, but as a movement.

It is indeed indisputable that epochs named for styles can describe little more than the cultural stances of an often compoundly interwoven elite. Social history has provided glimpses of the world of the majority of the 'Renaissance people'; these reveal little 'Renaissance' quality (e.g., Herlihy and Klapisch-Zuber, 1978; Davis, 1982; Squatriti and Merideth, 1999). Recently, the theme of gender has also been considered more intensely (see King, 1991, with comprehensive bibliography); here, research focuses not only on the 'Renaissance women', the princesses and courtesans who populate the mythological scenario, but also on other women who lived at the time – witches and saints, nuns, artists, authors, and women in working life in general (Ferguson, 1986).

'Modern Individualism' and the Concept of the Renaissance Today

Art Historical Findings

Despite the still current discussion of the term 'Renaissance', the concept has prevailed as the name of a cultural epoch (cf, e.g., Burke, 1972; Delumeau, 1967; Hale, 1993). What arguments can be adduced for or against the concept of the Renaissance? The most conspicuous are the findings of art history, and they were what gave the essential impetus to construe the epoch in the first place. Thus, it is beyond doubt that the painting of Giotto and Massaccio and the sculpture of the Pisani in the fourteenth and fifteenth centuries marked a fundamental transformation. What was decidedly new in painting was that dependence on the Byzantine model was overcome. Here, Flemish influences played a major role, and the Renaissance in general increasingly appears as a complex amalgamation of transfer processes embracing Europe, the Middle East, and finally the New World. The early Quattrocento brought the introduction of a central perspective formulated with geometric precision, a Florentine invention. The goal was to create the illusion of 'real' living beings, objects, and spaces.

One of the most important arguments for Burckhardt's old 'individualism hypothesis' is still that not only did artists' signatures appear even more frequently, but also that, in the fifteenth century, initially in the Netherlands and Italy, portraits were painted that appear closer to nature than anything since classical antiquity. Since the Trecento, sculpture once again clearly approached antiquity's conceptions of the body, and, increasingly frequently, portrait busts clearly strove to convey the individuality of the person portrayed. While it cannot be claimed that the 'modern individual' was invented in the fifteenth century, if we also adduce literary testimony in the broadest sense, for example Ghiberti's *Autobiography* and Giovanni Ruccelai's unique *Zibaldone*, we do note that reflection on the self-gained decisively in importance (Davis, 1986, is highly differentiating; see also Schulze, 1996; Porter, 1997; Schweikhart, 1998).

In architecture, the building forms of classical antiquity reappeared first and initially exclusively on the Italian peninsula. Only at the beginning of the sixteenth century was anything remotely comparable found north of the Alps, starting with the buildings commissioned by the Fuggers in Augsburg. Regional variant forms emerged rapidly. Their common denominator was a more or less clearly articulated reference to antiquity. With mannerism, the Renaissance flows into a typically European style (cf Section The 'Mannerist' Renaissance).

The Revolution in Discourse and Renaissance Humanism

The art of the Renaissance developed in a complex content with a theoretical discourse that, on the one hand, used texts from antiquity, for example Vitruvius' theory of architecture. This was new and genuinely Italian. In the fifteenth and sixteenth centuries, no other European country produced a theory and critique of art whose quality and differentiation were even remotely comparable. Italy was also the first place with a palpable group of collectors, art agents, and patrons who, for aesthetic considerations, acquired real or putative ancient objects. These inspired an almost romantic veneration and learned discussion and correspondence. One of the earliest representatives of this modern phenomenon of collecting was the Florentine Niccolò Niccoli (1364–1437).

Interest in ancient and contemporary art was part of the humanistic culture that was closely tied to the Renaissance. The term *studia humanitatis* is of ancient origin and refers to education in the broadest sense, the acquisition of knowledge as well as of 'good' ways of living. The fifteenth and sixteenth centuries saw the emergence of the social type of the *humanista*, the scholar concerned with the canon of humanistic studies – grammar, rhetoric, poetics, history, and moral philosophy – with recourse to the authors of classical antiquity. Humanistic scholars and humanistically educated elites became the real bearers of the Renaissance. They gave art, architecture, and literature essential elements of form and content (see, e.g., Wind, 1968; Panofsky, 1962). Modern research sees the first beginnings in upper Italy in the second half of the thirteenth century. But Francesco Petrarca (1304–74) was the first thinker to unite in himself the essential aspects of the humanistic movement. The older idea that the humanists were a group of traveling scholars staging a revolution against scholastic theology and church doctrines is no longer tenable. Social history cannot view the host of humanists as homogeneous, even as early as the fifteenth century. The majority of the humanists of Florence, the 'capital' of the Renaissance, were economically well off (Martines, 1968). Monks, prelates, princes, university scholars, urban jurists, and many others spread the *studia humanitatis* in Europe, and in many places humanism established itself as a national movement with national characteristics.

Humanism, and with it the culture of the Renaissance, were in no way pagan. The vast majority of humanists were firmly based in Christianity, and this did not change with the Reformation (Verdon and Henderson, 1997). Significantly, pictorial subjects remained mostly religious during the Renaissance. In the sixteenth century, however, we note their gradual

dwindling (Burke, 1972), a telling indication that Renaissance humanism contained substantial potential for secularization. The great expansion of the spectrum of questions discussed and written about in the fifteenth and sixteenth centuries was of long-term importance. Theology and religion still dominated, but texts on every possible aspect of life increasingly moved to the foreground; finally, and again with recourse to ancient thinkers, a new theory of politics developed (Pocock, 1975). This 'differentiation of discourse' did not depend on the Gutenberg Revolution, but began much earlier. In a certain way, the invention of the letterpress with movable type can be viewed as a response to a market ravenous for written material.

The Research Problem of the Genesis of the Renaissance

A question equally fascinating and controversial is why the beginnings of humanism and the Renaissance lay in Italy and what led to the formation of these movements. As Burckhardt had already underscored, there was the multiplicity of *signoria* and city republics, whose political rivalry sought artistic representation. Ancient monuments, statues, and other relics also provided models. Italy and Flanders, the regions in which the new culture developed, were probably the richest in Europe at the time, so here is where the economic prerequisites for extensive patronage were given. But the connection between art and economy should not be absolutized. There is no necessary parallel between the unfolding of Renaissance culture and a general economic upswing; Roberto S. Lopez (1963) even attempts to interpret the Renaissance as an attempt to compensate for crisis-like developments. The financial position of the patrons was decisive (see, e.g., Roover, 1963). The *condottieri* wars in fifteenth-century Italy resulted not merely in a destruction of resources, but also in processes of redistribution from the great centers to small seats of courts like Urbino and Bergamo, where successful military commanders spent astronomic sums on books and in patronizing the arts. It is difficult to estimate the significance of the dynamics of economic shrinkage triggered by plague epidemics and failed harvests in the first half of the fourteenth century. Along with the resulting acceleration of urbanization, the consolidation of capital in the hands of the survivors contributed to the financing of early Renaissance art – in part as an 'investment in saving one's soul', which must have suggested itself and seemed necessary against the backdrop of the crisis of church hierarchy and the experiences of the plague era. The Renaissance, humanism, and the Reformation thus appear as varying answers to one and the same problem, namely the determination of the human position in a fundamentally changed world. The writings of Giannozzo Manetti and Pico della Mirandola, the political theories of Machiavelli, the primary Reformation writings of Luther and the invention of the realistic portrait and of central perspectives thus all had the same preconditions.

The 'Mannerist' Renaissance

The question of the end of the Renaissance is no less complex and controversial than that of its beginnings. Art history sees indications of a metamorphosis to the 'Mannerist Renaissance' or to mannerism as early as the second decade of the sixteenth century (Shearman, 1967; Arasse and Tönnesmann, 1997). Literary history usually locates the end of the epoch around 1600. And music history, too, has its difficulties with the question of the definition and temporal boundaries of mannerism (important but controversial, see Maniates, 1979). National differences in particular must be differentiated. Setting the boundaries of the epoch is also made difficult by the fact such important intellectual currents constitutive of the Renaissance as humanism and such aesthetic ideals as the orientation toward the standards of Greek classicism lived on after 1600 and experienced their own renaissances.

There has been no lack of attempts to set the stylistic transformation of the sixteenth century in direct relationship to political and societal changes (the best known is Arnold Hauser, 1979). Structural changes at least in the art market and thus in the relationships between artists and patrons probably played an important role. Europe's princes, the rulers of the rising early modern states, moved increasingly into the foreground as patrons. They used the arts to buttress their legitimacy and to display their grandeur: now, in the sixteenth century, the art of the power availed itself of the power of art more than hardly ever before. Political competition also manifested itself as competition among patrons of the arts. The artists' virtuosity, or *maniera*, had to prove itself in this complicated field.

Over the past decades historiography has increasingly focused on the European Renaissance of the fifteenth and sixteenth centuries in a broad chronological and intercultural comparison. On the one hand, the main interest revolved around its antecedents in the Middle Ages and antiquity; among these efforts were, for instance, Erwin Panofsky's argument of the medieval 'revival' known as the Carolingian Renaissance or the various references to antiquity, usually summed up and discussed under the terms 'Proto-Humanism' or 'Proto-Renaissance' (Panofsky, 1960; Masse, 2010). The wide variety of cultural physiognomies associated with the Italian 'Renaissances' and its many centers has been systematically elaborated on recently (Crouzet-Pavan, 2007), focusing on topics as diverse as the importance of tradition, religiosity, fear of suffering in the afterlife, and the new idea of man. On the other hand, the search for 'Renaissances' in other civilizations continues. For instance, Peter Burke ended his study on the Renaissance with a comparative perspective on the Netherlands and Japan (Burke, 1972). Consequently, these approaches lead to the questioning of the prerequisites of distinct cultural epochs, the uniqueness of the European Renaissance (Goody, 2010; Huff, 2011), their implications for the Scientific Revolution of the Modern Age (Cohen, 2010), and their role in the emergence of the Industrial Revolution. In this context the comparison with China under the Song Dynasty (approximately tenth to thirteenth centuries) promises valuable insights; during its rule, the Middle Kingdom achieved outstanding accomplishments well in excess of contemporaneous developments in Latin Europe. Similarly of importance are questions on the origins and consequences of the widening gap between Latin Europe and the Muslim world, especially after the end of the Fatimid Caliphate (Saliba, 2007). A combination of the following factors may have contributed

to the path taken by the Latin Europe: the multitude of states in Europe and the multiple possibilities of patronage for both artists and scholars they offered; the existence of distinctively separate legal entities such as universities, relatively autonomous urban communes, and their creative middle strata; above all, however, Latin Europe profited from the containment of ecclesiastical power. The basis for these developments was the existence of the rich legacy of Hellenistic Greece. Only in Latin Europe with its widening possibilities of (secular) discourse were these riches of the past harnessed and developed further; the topography of innovation mirrors these facts (Murray, 2003). Other important prerequisites of the Scientific Revolutions included the existence of appropriate means of communication, exemplified by the printing press and the more liberal approaches to illustrating (e.g., the human body) practiced in the Renaissance West. The European Renaissance was quintessential for the emergence of the Modern Age and has especially with respect to the pluralistic effects of discourse and innovation, no parallel in human history.

See also: Art History; Bourgeoisie and Middle Classes, History of; Civilization, Concept and History of; Cultural History; Intellectual History; Middle Ages, The; Science, History of; State, History of.

Bibliography

Arasse, D., Tönnesmann, A., 1997. La Rénaissance Maniériste. Gallimard, Paris.

Baron, H., 1955. The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny. Princeton University Press, Princeton, NJ.

Baxandall, M., 1972. Painting and Experience in Fifteenth Century Italy: A Primer in the Social History of Pictorial Style. Oxford University Press, Oxford, UK.

Brucker, G., 1977. The Civic World of Early Renaissance Florence. Princeton University Press, Princeton, NJ.

Burke, P., 1972. Culture and Society in Renaissance Italy. B.T. Batsford, London.

Burke, P., 1998. The European Renaissance. Blackwell, Oxford, UK.

Cohen, H.F., 2010. How Modern Science Came into the World: Four Civilizations, One 17th Century-Breakthrough. Amsterdam University Press, Amsterdam, The Netherlands.

Crouzet-Pavan, E., 2007. Renaissances Italiennes 1380–1500 Éditions. Albin Michel, Paris.

Davis, N.Z., 1982. Women in the crafts in sixteenth-century Lyon. Feminist Studies 8, 47–80.

Davis, N.Z., 1986. Boundaries and the sense of self in early modern France. In: Heller, T.C., et al. (Eds.), Reconstructing Individualism. Stanford University Press, Stanford, CA, pp. 53–63, 332–336.

Delumeau, J., 1967. La Civilisation de la Renaissance. Editions Arnaud, Paris.

Esch, A., Frommel, C.L. (Eds.), 1995. Arte, Committenza ed Economia a Rome e nelle corti del Rinascimento. Giulio Einaudi, Turin, Italy.

Ferguson, W.K., 1948. The Renaissance in Historical Thought: Five Centuries of Interpretation. Houghton Mifflin/The Riverside Press, Cambridge, MA.

Ferguson, W.K. (Ed.), 1986. Rewriting the Renaissance: The Discourses of Sexual Difference in Early Modern Europe. University of Chicago Press, Chicago, IL.

Goldthwaite, R.A., 1993. Wealth and the Demand for Art in Italy, 1300–1600. Johns Hopkins University Press, Baltimore, MD.

Goody, J., 2010. Renaissances: The One or the Many? Cambridge University Press, Cambridge.

Hale, J., 1993. The Civilization of Europe in the Renaissance. HarperCollins, London.

Hauser, A., 1979. Der Manierismus. Die Krise der Renaissance und der Ursprung der modernen Kunst. Beck, Munich, Germany.

Herlihy, D., Klapisch-Zuber, C., 1978. Tuscans and Their Families: A Study of the Florentine Catasto of 1427. Yale University Press, New Haven, CT.

Holmes, G., 1986. Florence, Rome and the Origins of the Renaissance. Clarendon Press, Oxford, UK.

Huff, T.E., 2011. Intellectual Curiosity and the Scientific Revolution. Cambridge University Press, Cambridge, New York.

King, M.L., 1991. Women of the Renaissance. Chicago University Press, Chicago, IL.

Kristeller, P.O., 1961/1965. Renaissance Thought, vols. I and II. Harper Torchbook, New York.

Kristeller, P.O., 1972. Renaissance Concepts of Man and Other Essays. Harper Torchbook, New York.

Lopez, R.S., 1963. Hard times and investment in culture. In: Ferguson, W.K. (Ed.), The Renaissance: A Symposium. Seemann, New York, pp. 29–54.

Maniates, M.R., 1979. Mannerism in Italian Music and Culture 1530–1630. Durham, NC.

Martines, L., 1968. The Social World of the Florentine Humanists, 1390–1460. Princeton University Press, Princeton, NJ.

Masse, M.-S. (Ed.), 2010. La Renaissance? Des Renaissances? (VIIIe-xvi Siècles). éditions Klincksiek, Clamecy.

Murray, C., 2003. Human Accomplishment: The Pursuit of Excellence in Arts and Sciences, 800 B.C. to 1950. HarperCollins, New York.

Panofsky, E., 1960. Renaissance and Renascences in Western Art. Almquist & Wiksell/Gebers Förlag AB, Stockholm, Sweden.

Panofsky, E., 1962. Studies in Iconology. Harper Torchbook, New York.

Panofsky, E., 1969. Renaissance and Renaissances in Western Art. Almquist & Wiksell/Gebers Förlag AB, Stockholm, Sweden.

Pocock, J.G.A., 1975. The Machiavellian Moment. Princeton University Press, Princeton, NJ.

Porter, R., 1997. Rewriting the Self, Histories from the Renaissance to the Present. Routledge, London.

Porter, R., Teich, M., 1992. The Renaissance in National Context. Cambridge University Press, Cambridge, UK.

Roover, R.de, 1963. The Rise and Decline of the Medici Bank, 1397–1494. Harvard University Press, Cambridge, MA.

Rubinstein, N., 1997. The Government of Florence under the Medici (1434–1494). Clarendon Press, Oxford, UK.

Saliba, G., 2007. Islamic Science and the Making of the European Renaissance. Cambridge University Press, Cambrigde, MA/London.

Schmitt, C.B., Skinner, Q. (Eds.), 1988. The Cambridge History of Renaissance Philosophy. Cambridge University Press, Cambridge, UK.

Schulze, W. (Ed.), 1996. Ego-Dokumente. Annäherungen an den Menschen in der Geschichte. Akademie-Verlag, Berlin.

Schweikhart, G. (Ed.), 1998. Autobiographie und Selbstporträt in der Renaissance. Walther König, Cologne, Germany.

Shearman, J., 1967. Mannerism. Penguin, London.

Squatriti, P., Merideth, B., 1999. The Renaissance in the Fields: Family Memoirs of a Fifteenth-Century Tuscan Peasant. Penn State University Press, Philadelphia, PA.

Thomas, A., 1995. The Painter's Practice in Renaissance Tuscany. Cambridge University Press, Cambridge, UK.

Verdon, T., Henderson, J. (Eds.), 1997. Christianity and the Renaissance: Image and Religious Imagination in the Quattrocento. Syracuse University Press, Syracuse, NY.

Wackernagel, M., 1938. Der Lebensraum des Künstlers in der florentinischen Renaissance. Aufgaben and Auftraggeber, Werkstatt und Kunstmarkt. Leipzig, Germany.

Wind, E., 1968. Pagan Mysteries in the Renaissance. Faber & Faber, London.

Repartnering and Stepchildren

Frances K Goldscheider, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 13160–13164, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article examines three issues. These include (a) the role of children in the likelihood their parents repartner (considering different effects for men and women); (b) the role of such children on the fertility of the unions formed, and (c) the effect of stepchildren on union dissolution.

Introduction

Sexual unions create children and hence, parenthood. Union dissolution creates single parents and absent parents. When these parents enter a new partnership, they may have a new kind of children – stepchildren – children who are not the joint concern of the couple at the start of the union, as at least one member enters the union as a step-parent. This asymmetry in parenthood is likely to shape the partner dynamics that lead to union formation and hence to affect the calculations of those with children (whether coresident or not) and any partners who might join them in a new union. This asymmetry affects their lives as a couple, particularly whether they have additional, joint children, and also affects whether their union survives. This article on children and new partnerships examines three issues: (a) the role of children in the likelihood that their parents repartner (considering different effects for men and women); (b) the role of such children on the fertility of the unions formed; and (c) the effect of stepchildren on union dissolution.

Background

The rapid rise in divorce in most industrialized countries together with the rise in cohabiting unions, which tend to be more unstable than marriages, is creating a new demographic situation for men and women. A union disruption places them in the partner market again, which is particularly challenging for those who are the parents of young children. Although historically, high mortality often left young parents widowed, divorce creates twice as many new prospective partners. Further, whatever wisdom had accumulated for those facing this challenge, has likely been lost. The divorce revolution that began in the 1960s and 1970s in the industrialized countries in the west followed a half century or more of rapidly declining mortality, which sharply reduced widowhood for both sexes and concentrated it at ages when children are older, or even adults. In addition, the early years of the increase in divorce rates disproportionately involved couples without children far more than has been the case in the 1980s and 1990s (Andersson, 1997). Hence, the effect of existing children on new partnerships is not well understood, either in the community of scholars of the demography of the family or in the communities experiencing the new situation. Little is known about how children affect their parents' repartnering, and not much more is known about those partnerships as they develop over time in terms of their fertility and stability. The research community has been further hampered by a lack of appropriate data to study these issues.

The numbers involved are substantial. In the United States, it is estimated that 40–50% of all marriages dissolve and a substantial minority of all children spend some time in a single-parent family (Bumpass and Raley, 1995). Half of both previously married cohabiting couples and remarried couples have children in the household and many of the rest have children living elsewhere (Bernhardt and Goldscheider, 2001). Although such rates are higher in the US than in other industrialized countries, they are rising rapidly outside the US (*see* Second Demographic Transition), while US levels appear to have reached a plateau in the last two decades of the twentieth century. At whatever level, these trends mean that a substantial proportion of adults are at risk of entering a new partnership, and many of those who do will have children, either living with them or dependent upon them financially. Those considering forming a union have to take their potential partner's children into account.

A critical question might be whether the presence of children increases the likelihood that the unions formed are cohabiting ones. There is some evidence that this is, in fact, the case. Although the rise in cohabitation has not fully offset the decline in proportions married among the never married, the offset is total among the formerly married, suggesting that remarriage is being replaced by cohabitation (e.g., *see* Second Demographic Transition). However, most research, and most of the data on which research could be based, has made it difficult to study cohabiting partnerships that involve children. Research on 'single-parent' families has frequently studied only 'unmarried' parents. The most common control for resources in families was 'family income,' which typically excluded the income of 'non-family' members, i.e., cohabiting partners. New surveys, such as the Family and Fertility Surveys (FFS), should overcome these difficulties.

Repartnering of Parents

Nearly all research on repartnering when pre-existing children are involved has considered only women with coresident

children, focusing almost entirely on remarriage and on the US. These studies have typically found that the presence and sometimes the number of children reduce women's likelihood of marrying, with effects often depending on the ages of the children and of the mother. One of the few studies that considers cohabitation suggests that children reduce the partners' commitment to the union, decreasing the risk of getting married but increasing the risk of cohabiting (Bennett et al., 1995).

There is much less evidence on how children affect *men's* likelihood of union formation. A major problem for men is distinguishing between coresidential and non-co-residential children, a distinction that has been less important for women. One study examined widowers, where presumably fewer of the children lived with other relatives (Smith et al., 1991). That analysis showed that having dependent children reduces middle-aged (<60) men's likelihood of remarriage after widowhood, although the reduction was much less than is the case for women with children. However, more recent research on both the US and Sweden, which has included divorced and never-married men with coresident children, as well as widowed men, has found that such men enter partnerships significantly *more* quickly than men without children (Goldscheider and Sassler, 2000; Bernhardt and Goldscheider, 1999). This suggests that the effects of coresidential children on union formation are quite different for men than for women.

There is even less research on how non-coresidential children affect union formation, although this is a particularly important issue for men (and the women who consider entering unions with them). The studies mentioned above also examined this issue, and found that in both the US and Sweden, men with non-coresident children are significantly *more* likely than otherwise comparable men with no coresident children to enter a union than to remain outside a partnership. In Sweden, such men are also more likely to enter a union with either a woman who has co-resident children or one who does not than to remain unpartnered; in the US this effect does not appear for union entry for women without children. The few women who have non-coresident children are also more likely to enter a union than to remain outside a partnership. In this case, however, at least in Sweden, they are much more likely to choose a partner with no coresident children, although this effect may reflect the small numbers of nonpartnered men who live with their children.

Perhaps a more important question, however, is the broader one: what factors affect entry into a union with a partner who has children? Research on the US and Sweden shows that two family-related factors distinguish men by whether the partnerships they enter into include children: their parents' marital history and their own marital history (Bernhardt and Goldscheider, 1999; Goldscheider and Sassler, 2000). Men who experienced a nontraditional family structure in childhood are significantly more likely to form a marriage that converts them into stepfathers than one that does not, and the same general difference appears for cohabitation.

US men who have been previously married resemble those who are already parents, in being very likely to enter a union (Goldscheider and Sassler, 2000). If they marry, there appears to be no difference in the likelihood that their partner has children or not. If the new union is a cohabiting one, however, they are more likely to enter one with children. In fact, divorced men are more likely either to marry or cohabit with children than to remain unpartnered; and they are significantly less likely to cohabit without children than to remain outside a partnership. Similar results were obtained for Sweden (Bernhardt and Goldscheider, 1999).

Men who enter unions with children are not simply different in terms of their family histories, however; the evidence for both Sweden and the US suggests that they are also drawn from among the less educated, but the studies do not show why. In Sweden, those with a post-gymnasium education are the most likely to enter a union with a partner without children but the least likely to enter a union where the new partner has children (Bernhardt and Goldscheider, 1999). This pattern is the same for US men; college graduates are more likely to enter a union than to remain outside a partnership if the woman does not have children. They are more likely, however, to remain outside a partnership than to enter a union with a women who has children (Goldscheider and Sassler, 2000).

These results suggest that there are important differences in the likelihood of repartnering for those with and without children, with patterns that differ between men and women. Children, whether co-resident or not, increase repartnering for men, but only non-coresident children do so for women; for them coresident children have the opposite effect. This suggests that surveys need to go beyond providing information in their partnership histories about whether partners brought children into the union and at least also provide information on whether partners have children living elsewhere. Such children clearly affect union formation; it is likely that they would also affect partnership dynamics (fertility and union stability), though this cannot yet be assessed fully, as we shall see in the next section. It is also likely that other characteristics of the children, such as their sex, might affect their parents' repartnering. No studies to date, however, have assessed whether the gender of the child(ren) matters. This question will take on increased importance later in this article.

More fundamentally, it becomes increasingly important to understand the reasons why children reduce the likelihood of repartnering, which we now see is only the case when they are coresident children and are children of the woman and not the man (unless he is older and widowed). How important is the fact that, although in most industrialized countries stepfathers have few legal rights or responsibilities vis-à-vis their stepchildren, most men realize that the children will be claimants on their incomes, at least while they remain in the household? The effect of having coresident children appears to be less for Swedish women than for women in the US (cf Bernhardt and Goldscheider, 1999; Bennett et al., 1995). It may be that the greater support provided to children of all kinds by the Swedish state offsets some of the perceived liability for prospective stepfathers.

A perhaps equally important problem is ambiguity over the nature of stepfather–stepchildren relationships. Stepfathers are discouraged from assuming parental responsibilities for their stepchildren's lives by the children, their mothers, and by the larger society in most western, industrialized countries. Paternal involvement is an important factor in family dynamics, particularly for union stability, and we will examine this topic below.

A further interesting question is how much of the decline in marriage formation is a result of the increase in potential partners with children? Most studies that have examined whether the quality of the partner market has declined have considered women's decisions and have examined trends in the earnings of unmarried men who would be their potential partners. It might be equally interesting to study men's decisions and examine whether trends in men's marriage likelihood are influenced by the changing composition of potential female partners, as increasing proportions have children living with them.

Fertility of Couples with Children from Previous Relationships

There have been more studies on the effects of children from a previous union on childbearing in a new union than on repartnering. Few, however, distinguish between coresident and non-coresident children, and none have taken into account the selectivity of repartnering when children are already present as revealed in the studies reviewed above. Nevertheless, these studies have assembled an impressive set of results for a wide range of industrialized countries in Europe, as well as in the US.

Not surprisingly, the general finding is that couples in later unions have lower fertility within those unions than do couples in otherwise comparable first unions, which indicates clearly that many individuals have childbearing goals they do not want to exceed. Their previous children are considered in their decisions to continue childbearing. This result has appeared in studies in Germany, Czechoslovakia, the US, France, and Sweden (cf references in Vikat et al., 1999).

How much previous children matter and whether it makes a difference whether they are coresident or not remain open questions. They are much more difficult questions to assess for the moment, as there is rarely information on previous children for both members of a couple, particularly on their non-coresident children. The most complete study is a recent analysis of Swedish data (Vikat et al., 1999), which, although it only had data on non-co-resident children for one member of the couple, produced a number of interesting findings.

First, the study found that the effect of previous children is definitely not total, as couples in Sweden appear to have at least one shared biological child independently of previous childbearing by either partner. The authors interpret this result as indicating that joint children are an important element of a union. Second, the authors were able to rule out some alternative reasons for childbearing in later unions. There was little evidence that the risk of having a first child in a union was higher when one of the partners was childless, suggesting that Swedes do not have a child simply as an indicator of adult status. Further, there was little evidence the couples have a child to produce a sibling for a pre-union child, although couples did appear to avoid having only one shared child. This was interpreted as suggesting that it is full siblings that are considered important by Swedish couples in new unions, not half-siblings. Research on the US indicates that while stepsiblings remain fairly close to each other in adulthood, they are much less close than full siblings (White and Riedmann, 1992), although it is unclear how much this difference reflects their different histories prior to coresidence and how much it reflects problems in their relationships while sharing a home. It was not possible to assess how much fertility in repartnerships is affected by coresident as opposed to non-coresident stepchildren, as full information on the latter was not available.

Fertility in repartnerships is clearly a complex subject, and one that might depend not only on the numbers of children each of the partners has had previously and where they live, but also the genders of those children. Strong gender effects appear in research on stepfamily relationships, and there is evidence that gender of children affects union dissolution (Brooks-Gunn, 1994). No studies to date, however, have assessed this issue, as no research has addressed the selectivity of repartnering.

Stepchildren and Union Dissolution

One of the most widely observed characteristics of second partnerships is that they are less stable than first partnerships. Although many have speculated that this pattern was the result of one or more elements reducing commitment (from religious prohibitions to personal orientations toward divorce), a landmark study found that most of the difference in marital satisfaction and in stability could be attributed to the presence of stepchildren. Previously childless persons who repartnered had very little greater risk of union dissolution than those in first partnerships (White and Booth, 1985). Studies of union dissolution since then have generally included indicators of the presence of non-joint children as well as joint children, and found that whatever protective effect joint children might have does not characterize stepchildren, who increase the risk of separation. Little research has been done, however, on what types of stepchildren contribute more to this effect. Are the ages of the children a factor? Is their sex? Does the involvement of the step-parent and/or of the non-coresident parent matter? An even more important element might be the parenting practices of the new partners. Two studies of paternal involvement that was restricted to first partnerships showed that greater involvement of a father with his children increased the wife's marital satisfaction (Harris and Morgan, 1991) and their children's well-being (Amato, 1994); stepfathers are normally less involved with their stepchildren than are biological fathers (Marsiglio, 1991).

Related to this issue is a startling finding – that couples with daughters are more likely to divorce than couples with sons (Morgan et al., 1988). The key to this result appears to be paternal involvement with children. Men tend to spend more time with their sons than their daughters, both in one-on-one conversation and in terms of participation in *family* activities, sharing more meals, leisure time, and home activities with them (Cooksey and Fondell, 1996). In a sense, stepchildren may be like daughters for many men, children who they wanted less, who they do not relate to as closely, and who they can leave more easily. This may also account for the commonly observed phenomenon that sons often benefit more by their mother's remarriage than daughters (Brooks-Gunn, 1994). Given the increased importance of repartnering in the lives of children and their parents in western, industrialized societies, understanding this phenomenon needs to be a high priority.

See also: Children and Law; Divorce, Sociology of; Family Business; Family Theory: Economics of Marriage and Divorce; Fatherhood; Marital Status, Divorce, and Child Development; Marriage; Motherhood; Parenting Attitudes and Beliefs across Cultures.

Bibliography

Amato, P.R., 1994. Father-child relations, mother-child relations, and offspring psychological well-being in early adulthood. Journal Marriage and the Family 56, 1031–1042.

Andersson, G., 1997. The impact of children on divorce risks of Swedish women. European Journal of Population 13, 109–145.

Bennett, N.G., Bloom, D.E., Miller, C.K., 1995. The influence of nonmarital childbearing on the formation of first marriages. Demography 32, 47–62.

Bernhardt, E., Goldscheider, F., 2001. Men, resources and family living: the determinants of union and parental status in the US and Sweden. Journal of Marriage and the Family 63, 793–803.

Bernhardt, E., Goldscheider, F., 1999. In: Children and Union Formation in Sweden. Paper Presented at the European Population Conference, August. The Hague, The Netherlands.

Brooks-Gunn, J., 1994. Research and policy issues in the study of stepparenting. In: Booth, A., Dunn, J. (Eds.), Stepfamilies: Who Benefits? Who Does Not? L. Erlbaum Associates, Hillsdale, NJ, pp. 167–189.

Bumpass, L.L., Raley, R.K., 1995. Redefining single-parent families: cohabitation and changing family reality. Demography 32, 97–109.

Bzostek, Sharon H., McLanahan, Sara S., Carlson, Marcia J., 2012. Mothers' repartnering after a nonmarital birth. Social Forces 90 (3), 817–841.

Cooksey, E.C., Fondell, M.M., 1996. Spending time with his kids: effects of family structure on fathers', and children's lives. Journal of Marriage and the Family 58, 693–707.

Goldscheider, F., Sassler, S., 2000. Men, Children, and Union Formation, unpublished paper.

Harris, K.M., Morgan, S.P., 1991. Fathers, sons, and daughters: differential paternal involvement in parenting. Journal of Marriage and the Family 53, 531–544.

Kalmijn, Matthijs, 2013. Relationships between fathers and adult children: the cumulative effects of divorce and repartnering. Journal of Family Issues 20 (10), 1–23.

Kreider, Rose M., Lofquist, Daphne A., 2014. Adopted Children and Stepchildren, 2010. Current Population Reports. US Census Bureau, Washington DC, pp. 20–572.

Marsiglio, W., 1991. Paternal engagement activities with minor children. Journal of Marriage and the Family 53, 973–986.

Morgan, S.P., Lye, D., Condran, G., 1988. Sons, daughters and the risk of marital disruption. American Journal of Sociology 94, 110–129.

Noël-Miller, Claire M., 2013. Repartnering following divorce: implications for older fathers' relations with their adult children. Journal of Marriage and Family 75 (3), 697–712.

Smith, K., Zick, C.D., Duncan, G.J., 1991. Remarriage patterns among recent widows and widowers. Demography 28, 361–374.

Sweeney, Megan M., June 2010. Remarriage and stepfamilies: strategic sites for family scholarship in the 21st century. Journal of Marriage and Family 72 (3), 667–684.

Vikat, A., Thomson, E., Hoem, J., 1999. Stepfamily fertility in contemporary Sweden: the impact of childbearing before the current union. Population Studies 53, 211–225.

White, L.K., Booth, A., 1985. The quality and stability of remarriages: the role of stepchildren. American Sociological Review 50, 689–698.

White, L.K., Riedmann, A., 1992. When the Brady Bunch grows up: step/half sibling and full sibling relationships in adulthood. Journal of Marriage and the Family 54, 197–208.

Repetitive Sex Offenders, Treatment of

Robert A Prentky and Raina V Lamade, Fairleigh Dickinson University, Teaneck, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.A. Prentky, B.K. Schwartz, volume 23, pp. 15893–15897, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of sex offender legislation, cases, treatment, and research, including juvenile sex offenders and the new challenging area of Internet sex offenders. Laws pertaining to sex offenders have waxed and waned throughout the years. Legislation enacted in response to notable crimes and public response include: the Jacob Wetterling Crimes Against Children and Sexually Violent Offender Registration Program, also known as 'Megan's Law,' the Pam Lynchner Sexual Offenders Tracking and Identification Act, the Adam Walsh Child Protection and Safety Act. Treatment models for sex offenders typically employ a relapse prevention (RP)/cognitive behavior therapy (CBT) model and frequently target impaired intimate relationships, victim empathy, anger management, impulse control, antisocial personality features, lifestyle impulsivity, cognitive distortions, sexual fantasy, and deviant sexual arousal. In addition to RP and CBT, treatment modalities include behavioral modification (covert sensitization), pharmacological interventions, dialectical behavior therapy, restorative justice model, risk-needs-responsivity, and the good lives model. Treatment efficacy along with challenges to treatment and management of sex offenders are discussed.

Designation of a subclass of sex offenders as 'repetitive' derives from sexual psychopath legislation that sought to confine sex offenders for indeterminate periods through civil commitment. In response to horrendous crimes committed by a few sex offenders released from prison, states passed a series of laws providing for the involuntary, civil commitment of individuals identified as 'sexually dangerous persons,' 'sexual psychopaths,' or 'mentally disordered sex offenders.' Thirty states passed such laws. During the 1970s and 1980s these laws were successfully challenged in court and many were repealed. By the mid-1980s, only a half dozen states remained with such laws. In 1990, a sex offender recently released from a prison in Washington state abducted, raped, and castrated a 7-year-old boy. An outraged public demanded a response, and involuntary commitment was resurrected. The legality of the Washington statute, as well as other similar laws, were immediately challenged, and a Kansas statute, identical to Washington's statute, reached the US Supreme Court. In what would become a landmark case, the Supreme Court upheld the constitutionality of the Kansas statute (*Kansas vs Hendricks*, see below).

Sex Offender Legislation

Beginning in the mid-1990s, Congress passed many laws governing the management of sexual offenders, most of which focused on community notification. These laws included: (1) The Wetterling Act, passed in 1994 as part of the Violent Crime Control and Law Enforcement Act (42 U.S.C. 14071); (2) the amended Wetterling Act of 1996, officially called the Jacob Wetterling Crimes Against Children and Sexually Violent Offender Registration Program, also known as 'Megan's Law' (P.L.104–105); (3) The Pam Lynchner Sexual Offender Tracking and Identification Act (42 U.S.C. Cardiac Disease, Coping with) of 1996; (4) The Wetterling Improvements Act of 1997 (P.L. 105–119); and (5) The Adam Walsh Child Protection and Safety Act of 2006 (P.L. No. 109–248, 42 U.S.C. 16901).

The amended Wetterling Act of 1996 ('Megan's Law') encouraged all states to institute a procedure for community notification, and The Adam Walsh Act (AWA) established a comprehensive national registry of sex offenders. Title I of AWA, entitled the Sex Offender Registration and Notification Act provided minimum guidelines for sex offender registration and notification, and specified minimum registration time frames (i.e., 15 years, 25 years, lifetime) based on the convicted sexual offense, without consideration of psychological status or risk level. The Pam Lynchner Sexual Offender Tracking and Identification Act established a national database of sex offenders to assist law enforcement in tracking sex offenders across state lines.

The *Hendricks* Era

The single most influential court decision in the decade of the 1990s was *Kansas vs Hendricks*, U.S., 117 S. Ct. 2072 (1997). The Kansas statute established procedures for civil commitment of persons who, due to a 'mental abnormality' or a 'personality disorder' were likely to engage in 'predatory acts of sexual violence' (*Kan. Stat. Ann. Sec.* 59-29a01 1994; *Kansas vs Hendricks*). Kansas invoked the act for the first time to commit Leroy Hendricks, a 60-year-old man with a long history of taking 'indecent liberties' with children, and who was scheduled for release from prison shortly after the act became law. According to Justice Breyer, in his dissenting opinion, Hendricks had a 'serious mental disorder' (pedophilia) that rendered him dangerous because of a 'specific, serious, and highly unusual inability to control his actions' (Cornwell, 1998: p. 397). In a subsequent landmark case (*Kansas vs Crane*, 534 U.S. 407 (2002)), the US Supreme Court upheld the Kansas Sexually Violent Predator Act. At present, 20 states have statutes patterned after the Kansas statute. All of these statutes include the word *predispose* or *likely* (i.e., the mental abnormality makes the person *likely* to engage in sexually violent behavior or

predisposes the person to engage in sexually violent behavior). Although *Hendricks* affirmed the importance of volitional impairment, the subsequent *Crane* decision clarified that *Hendricks* did not have complete lack of control, only some degree of impaired control. Precisely what constitutes 'lack of control' in general and partial lack of control (*some degree of*) in particular has never been clarified.

The clear implication is that sex offenders, to be eligible for civil commitment, must be incapable of controlling their impulses to act out sexually in unlawful ways. Indeed, in those cases where diagnostic criteria may be inferred, the predominant *behavioral* criterion for the old sexual psychopath statutes and the current sexual predator statutes involves some form of volitional impairment (i.e., an inability to control impulses to engage in acts of sexual aggression). Impaired impulse control is, at least in principle if not in reality, associated with 'repetitive' offenses. Thus, the 'modal' individual committed under these civil commitment statutes is a repetitive or habitual sex offender who has an inability, or an impaired ability, to control his impulses to offend sexually. Although the use of the word 'repetitive' in statutory lexicon tells us relatively little about the individual, clinically or dynamically, it does inform us about risk. Repetitiveness clearly is a risk factor for reoffense. Thus, although all offenders with multiple known sexual crimes may be classified as 'repetitive' and arguably pose a greater risk for reoffense, those same offenders may be highly heterogenous with respect to etiology and clinical/treatment issues. One of the resulting conundrums in designing treatment programs is that many clinical issues that would seem logical targets for treatment (e.g., victim empathy, remorse, denial, and cognitive distortions) are not significantly related to reoffense.

Treatment Models for Sex Offenders: Relapse Prevention

Models for treating these 'repetitive' sex offenders are no different from models used to treat sex offenders not classified as predators or as repetitive. Historically, the dominant model for treating sex offenders has been relapse prevention (RP), a cognitive-behavioral approach that was adapted from an intervention intended for use with alcoholics (Marlatt and Gordon, 1985). The theory guiding its adaptation for sex offenders assumed that sexual offending behavior has identifiable triggers (with cognitive, emotional, and behavioral aspects) that must be identified and dealt with in healthy ways to interrupt the offense cycle (Pithers et al., 1983). The RP model and its many elaborations and revisions has been the subject of numerous articles, chapters, and books over the past three decades (Laws, 1989; Schwartz and Cellini, 1995, 1997; see *Sex Offenders, Clinical Psychology of*). Modifying any unwanted behavior stems from the informed consideration of those factors that are most importantly associated with the emergence and the sustenance of that behavior. From the ample research on addictive disorders, particularly alcoholism, smoking, and obesity, there has emerged a substantial literature on preventing relapse (Brownell et al., 1986). As noted, RP, as a model, was revised and adapted for use with sex offenders (Laws, 1989; Laws et al., 2000). Although sexual aggression derives from a substantially more complex amalgam of factors than most addictive disorders and typically reflects a chronic pattern of maladaptive behaviors, the underlying principles remain similar. That is, before designing strategies for modifying sexually aggressive behavior, those factors that are most importantly related to the behavior (i.e., determinants of relapse) must be identified. Although the factors discussed here are, by no means, inclusive, they do represent the core components of most contemporary treatment programs.

Core Elements of Most Treatment Programs for Sex Offenders

The following core elements occur within a cognitive-behavioral therapy (CBT) framework that may be administered on an individual and/or group format.

Impaired Adult Relationships

Impaired relationships with adults reflect a broad dimension of social competence that is temporally stable and multiply determined. Although social skills training is one of the most frequently included components of treatment for sex offenders, its role in relation to sexual aggression remains unclear. Because there are many facets to social competence, specifying what aspect one is assessing is critical. For instance, although there is reasonable support for the conclusion that rapists have assertiveness problems, differences in other areas of social competence are less evident. Likewise some child molesters, such as 'pedophiles,' have a variety of social and interpersonal inadequacies.

Although the question of the etiologic importance of social competence remains uncertain, the inclusion of techniques to address social skills deficits is still clinically appropriate. In addition to social skills training, the modalities that are most often used to improve relationships with adults include assertiveness training, relaxation training, systematic desensitization, sex education, and self-esteem enhancement.

Lack of Empathy

In all domains of interpersonal violence, a general lack of empathic relatedness for one's victim can be regarded as a powerful disinhibitor. Alternatively, the presence of empathic concern may serve to inhibit aggression. There is a clear clinical rationale for assuming the presence of such deficits and targeting interventions to enhance empathy.

Hildebran and Pithers (1989) described the importance of developing empathy for all victims of sexual abuse as an essential first stage of treatment, followed by the development of empathy for one's own victims. At this point most sex offender treatment programs include a separate component for increasing victim empathy. Indeed, Knopp et al. (1992) found in their survey of treatment programs in North America that 94% included victim empathy training. In addition to the standard exercises and tapes (video and audio) used in victim empathy training, expressive therapy may be used to increase the offender's emotional or affective

response to the distress of the victim. Some programs introduce victim advocates, victim counselors, and occasionally victims to increase further the emotional impact. Moreover, increasing the offender's affective appreciation of his own childhood experiences of victimization can instill a greater awareness of his victim's experience of abuse.

Degree and Nature of Anger

The degree and nature of anger evidenced in the assault has long been assumed to differentiate between offenders who were – and those who were not – willing to use extreme force to gain victim compliance (Prentky and Knight, 1991). Sex offenders differ considerably in the nature and extent of their anger and the ways that anger manifests itself in their crimes. Sex offenders who intend only to force victim compliance are likely to vary widely in the amount of aggression evident in their offenses. When the aggression clearly exceeds what was necessary to force victim compliance, the motivation and manifestation may also vary considerably. In some cases, the anger may be undifferentiated with respect to victims (i.e., the offender may be globally angry and express his anger at any available target), and in other cases the offenders may be focally angry at women (i.e., misogynistic anger) and displace their anger at women in general or on specific groups of women (e.g., prostitutes or the elderly). Similarly, child molesters who intentionally inflict considerable physical injury on their victims may be focally angry at children.

The recognition of the critical importance of anger as a driving force in sexual offenses has resulted in the inclusion of treatment techniques to reduce and contain anger. The most commonly employed of these techniques is an anger management group that uses cognitive-behavioral strategies to increase self-control as well as the timely and appropriate expression of angry feelings. In addition, RP, which also focuses on increasing self-management skills, and stress management can assist the offender to gain control over chronic and situationally induced anger. Finally, early life experiences of victimization can fuel lifelong anger that is periodically triggered by real or imagined provocations. A group that focuses on childhood victimization can help the offender to master these traumatic events.

Cognitive Distortions

Cognitive distortions are 'irrational' ideas, thoughts, and attitudes that serve to: (1) perpetuate denial around sexually aggressive behavior, (2) foster the minimization and trivialization of the impact of sexually aggressive behavior on victims, and (3) justify and sustain further sexually aggressive behavior. Cognitive distortions are presumed to be learned attitudes that are instilled at an early age by caregivers, reinforced by peers during childhood and adolescence, and further strengthened in adulthood by the prevailing social climate.

The social and cultural forces that have been hypothesized to contribute to sexual violence include the permissive responses of a wide variety of social systems and institutions that function to perpetuate rape myths and misogynistic attitudes, the objectification and exploitation of children and women in pornography, and the often similar but more subtle messages conveyed in advertising that support, or at least condone, sexual harassment (Stermac et al., 1990). The importance of cognitions in moderating sexual arousal has been repeatedly demonstrated. Moreover, clinical observations have suggested that many sexual offenders harbor offense-justifying attitudes and that these attitudes are importantly related to the maintenance of the 'sexual assault cycle.' Thus, the modification of irrational attitudes has been a major focus of treatment intervention. Although there are a variety of treatment modalities that may impact these distortions, the most commonly employed technique is cognitive restructuring. For cognitive restructuring to be most effective, it is critical that cognitive and affective components be addressed. That is, it is insufficient merely to confront the 'distorted' nature of the attitudes, to discuss the role that such distortions play for the individual, or to provide accurate information about sexual abuse (all cognitive components). It is equally important to create discomfort by focusing on the victim's response (e.g., fright, pain, humiliation) – the affective component. This latter exercise is also integral to victim empathy training.

In addition to cognitive restructuring, a group that focuses on childhood victimization can also be very helpful. Since the origin of these distorted attitudes is often a primary caregiver who was an influential role model, as well as exposure to peer role models, often in institutional settings, a group that focuses on these early adverse life experiences can help to trace the cognitions to their source, thereby challenging the generality of these attitudes, diminishing their logic or their 'truth.'

Sexual Fantasy and Deviant Sexual Arousal

Sexual fantasy refers to cognitive activity that focuses on thoughts and images having sexual content, and deviant sexual arousal refers to an arousal response that is prompted by thoughts, stimuli (visual or auditory), or behaviors that are defined as unconventional or antisocial by society. Studies investigating the relation of sexual fantasy to sexual aggression have typically employed plethysmography to assess sexual arousal to auditory and visual stimuli that purportedly tap specific sexual preferences. The guiding premise of these studies has been that deviant sexual fantasy is highly correlated with deviant sexual arousal and that both deviant fantasy and deviant arousal patterns are important precursors of deviant sexual behavior (Prentky and Knight, 1991).

The frequent targeting of such fantasies for therapeutic intervention reflects the widely held belief that the modification of deviant sexual fantasies and deviant sexual arousal patterns is critical for the successful treatment of sex offenders. Indeed, the presence of deviant sexual fantasies does appear to increase the likelihood of subsequent deviant sexual behavior. Moreover, the moderate success at increasing nondeviant arousal and behavior by applying techniques aimed solely at modifying arousal to deviant sexual fantasies supports the hypothesis that deviant fantasies not only lead to and maintain deviant sexual behavior but also impede normal sexual adaptation.

Behavioral techniques for modifying sexual arousal patterns are grouped into two categories, those that decrease deviant arousal (e.g., covert sensitization, aversion,

masturbatory satiation) and those that increase appropriate arousal (systematic desensitization, fantasy modification and orgasmic reconditioning). Although numerous behavioral techniques have been reported in the literature, the most widely used method has involved some variant of aversive therapy.

In addition to the repertoire of behavioral interventions, somatic treatment has become increasingly popular as a complement to psychological treatment. These organic or drug treatments consist primarily of antiandrogens (Bradford, 1989) and antidepressants (Kafka, 1991). The antiandrogens (e.g., medroxyprogesterone acetate and cyproterone acetate) reduce sexual drive by reducing the level of testosterone. The antidepressants that are used are primarily the selective serotonin reuptake inhibitors, such as fluoxetine. Although the neuroregulation of sexual drive remains unclear, there is some evidence that enhanced central serotonin neurotransmission inhibits sexual arousal. In addition, there is clinical evidence that uncontrollable sexual urges and compulsive sexual behaviors are associated with dysthymia and major depression.

Antisocial Personality/Lifestyle Impulsivity

In research on sexual offenders, impulsive, antisocial behavior has proven to be a critical risk factor for recidivism (Hall, 1989; Prentky and Knight, 1986). There is increasing evidence even within noncriminal samples (e.g., college students) that the likelihood of engaging in sexually aggressive behavior is greater among those who are more impulsive (Prentky and Knight, 1991).

Clinicians have long recognized the importance of impulsivity for relapse and have introduced self-control and impulsivity management modules into treatment. In addition to groups that focus specifically on impulse control, most treatment programs include components of RP. RP begins by identifying the chain of events and emotions that lead to sexually aggressive behavior. Once this 'assault cycle' is described, two interventions are employed: (1) strategies that help the offender avoid high-risk situations, and (2) strategies that minimize the likelihood that high-risk situations, once encountered, will lead to relapse. This is an 'internal self-management' system (Pithers, 1990) that is designed to interrupt the seemingly inexorable chain of events that lead to an offense. RP is potentially helpful for interrupting patterns of behavior that eventuate in specific outcomes, such as sexual assault, as well as patterns of behavior that are more global, such as impulsive, antisocial behavior.

Beyond CBT: Alternative Treatment Models

There are other treatment models in addition to RP/CBT.

The *restorative justice* model has its roots in indigenous cultures in the United States, Canada, and Australia. The principle of restorative justice is that when someone harms another, it has an effect not only on the victim but also on the extended family, on the community at large, and the person who caused the harm. Unlike all other treatment models, which focus on rehabilitating the offender and managing his risk, restorative justice programs emphasize restoring a sense of well-being not only to those who were harmed, but to the offender, and to all affected members of the community. Through mediation between victims and offenders, reparations, conferencing and negotiation, restorative justice moves justice out of the courtroom and into the community (Rodriguez, 2007). One of the most notable restorative justice programs implemented with sex offenders is RESTORE, designed by Dr Mary Koss (Koss et al., 2003). RESTORE is intended for those who have committed date and acquaintance rape. Restorative justice models have also been used with incest offenders and with juvenile sex offenders.

The *Risk-needs-responsivity* (RNR) model is an offender rehabilitation model developed by Andrews et al. (1990). The RNR model adheres to three important principles: *risk principle* (the most intensive services should target the highest risk offenders), *need principle* (treatment should target criminogenic needs, which are almost entirely dynamic in nature), and *responsivity principle* (treatment should be provided in a way that is most responsive to the offender's abilities). In brief, risk addresses *who* should be treated; need addresses *what* should be treated; responsivity addresses *how* treatment should be provided. As described by Andrews and Bonta (2010), criminogenic needs captures three of the 'big four' risk factors (procriminal attitudes, antisocial personality traits, and procriminal peers). The fourth factor, criminal history, is largely static. The RNR model has been implemented in numerous offender programs in Canada, the United Kingdom, Australia, and New Zealand. RNR has been found to have applicability for female offenders and juvenile offenders as well. RNR has not been widely adopted in the United States and has not been widely used with sex offenders. In one metaanalysis, Hanson et al. (2009) reported that interventions that follow the principles of the RNR model are applied to sex offenders and recidivism rates (sexual and general) are significantly reduced.

The *good lives* model is a 'strength-based' approach that complements CBT with a comprehensive risk-focused intervention strategy for managing sex offenders (Ward et al., 2007). Good lives model emphasizes the development and enhancement of individual strengths while simultaneously recognizing the need to manage risk to reoffend. With at least four recent books highlighting good lives, this model has rapidly achieved widespread recognition (Laws and Ward, 2011; Ward and Maruna, 2007; Yates and Prescott, 2011; Yates et al., 2010). Perhaps a mark of this recognition is competitiveness with RNR. Andrews et al. (2011) questioned whether the good lives model adds anything beyond what RNR already provides, and Ward et al. (2012) responded in kind.

Dialectical behavior therapy (DBT), although not a comprehensive treatment model, has been making slow inroads into sex offender therapy. DBT was originally developed by Dr Marsha Linehan (1993) as a treatment for borderline personality disorder (BPD) and clients presenting with high-risk, self-harming behaviors. Shingler (2004) pointed out that both those diagnosed with BPD and sex offenders exhibit interpersonal, behavioral, emotional, cognitive, and self-dysregulation. The twin rivers corrections center (TRCC) (Monroe, WA) sex offender treatment program is one example

of the adaptation of DBT for use with sex offenders. In the second phase of a two-phase treatment program, the TRCC implemented DBT group skills training in a community setting and found that it was effective in reducing impulsivity and increasing therapeutic alliance in high-risk offenders but not lower-risk offenders (Berzins and Trestman, 2004).

Although treatment programs for juvenile sex offenders have traditionally been modeled on the adult programs, focusing on CBT and RP, a strength-based approach has also been adopted by many juvenile sex offender treatment providers. The Oregon Youth Authority (Dr Keith Kaufman), for example, has designed a variation of this theme, using psychoeducational classes that target competencies, strengths, and healthy adaptations in youth. As noted, restorative justice programs have also been used successfully, though much less frequently, with juveniles.

Challenging Problems in Offender Treatment

The question of treatment effectiveness is a deceptively difficult one to answer due, principally, to the constraints imposed on treating forensic clients in general and sex offenders in particular. The numerous 'intervening' obstacles, which have been elsewhere discussed that may arise when treating sex offenders: (1) in a prison, the highest premium is placed on security, not therapy, and conditions for therapy often are suboptimal (e.g., severe space limitations, limited scheduling options, canceled therapy sessions on a moment's notice, maintaining staff morale in an unsupportive environment); (2) expecting 'clients' to fully disclose and 'come clean' in an environment where it may be dangerous to be known in population as a sex offender, (3) expecting offenders to engage in treatment when confidentiality cannot be assumed or ensured and when the offenders are reluctant 'customers' who have asked to be in therapy, (4) where therapists often are inadequately trained and supervised, (5) where program-related resources are limited and often restricted, and (6) where the 'outcome' variable (survival after release) is determined by a complex legal proceeding, *not* by programmatic considerations (Prentky et al., 2011). Although there are some more encouraging reports, the only randomized clinical trial of sex offender treatment remains that of Dr. Janice Marques (Marques et al., 2005). A prison-based RP program followed those assigned to RP, a volunteer control group and a nonvolunteer control group for an average of eight years at risk. There were no significant differences between the three groups in sexual recidivism. All three groups had roughly a 20% recidivism rate over the follow-up period. Explaining this important outcome study is tricky. It likely speaks more to the difficulties of conducting effective treatment in a penal environment than to the inefficacy of the treatment itself.

Management in the Community

In addition to registration and community notification, residency restriction is another management strategy for offenders living in the community. Residency restrictions often prohibit offenders from living less than 500 feet from where children congregate and therefore severely limit housing options for offenders in the community. In addition to being isolated from the community, additional consequences of residency restriction include limited access to employment and health care, particularly mental health services at a time when they are perhaps the most crucial.

Internet Sex Offenders

The Internet provides a medium through which one can target and groom victims for a hands-on offense, distribute child pornographic images, or find like-minded individuals who share distorted beliefs about child sexuality and child sexual offenses. The combination of *accessibility*, *affordability* and *anonymity*, dubbed the 'Triple-A Engine' by Cooper (1998), make the Internet an attractive, powerful, and efficient vehicle for gratifying sexual interests, including child pornography. Recent advances in smart phone technology allow for practically unlimited Internet access at any time. This has opened the gateway to a new 'type' of child molester: the Internet sex offender who possesses, distributes, trades, and may produce Internet child pornography. Wolak et al. (2005) found that 39% of a sample of child pornography offenders were so called 'dual offenders' (having both child pornography and contact sexual offenses against minors).

Although still in its infancy, the empirical literature suggests that 'Internet-only' offenders (sex offenders who have committed an Internet sexual offense, but not a contact offense) may be a unique subgroup of pedophiles, as they seem to be differentiated from contact offenders. The consistent differentiating factor is antisociality, with contact offenders being significantly more antisocial (Eke et al., 2011; Seto and Eke, 2005; Webb et al., 2007; Lee et al., 2012). Internet sex offenders also appear to be less deviant, less physically aggressive and less impulsive than contact offenders (Tomak et al., 2009). Webb et al. (2007) found that while contact sexual offenders were more likely to 'act out' and break rules, Internet offenders reported more psychological difficulties in adulthood and fewer prior sexual convictions.

The essential question for law enforcement is who, among all those found in possession of Internet child pornography, is at high risk of committing a contact offense against a child, or has already done so and not been caught. Endrass et al. (2009) found that for those individuals without previous convictions for hands-on offenses, accessing Internet child pornography alone is not a risk factor for committing a contact sexual offense against a child. This does not address questions about how Internet child pornography might operate in tandem with or indirectly through other antecedent and risk factors. Seto et al. (2001) recommended that future research look at potential moderators, such as hostility and paraphilic interest.

The Internet can also serve as an environment to promote pedophilic beliefs, bringing together like-minded individuals. For example, Holt et al. (2010) looked at attitudes and beliefs displayed in Web forums for pedophiles. They found that these forums exhibited high levels of cognitive distortions around the following themes: (1) marginalization (i.e., they are victims and not understood); (2) sexuality (i.e., children are sexual beings); (3) law (i.e., information about legal statutes); (4) security (i.e., information about how to evade

the law, how to protect oneself from getting caught). This suggests that the Internet provides a cultural context by which beliefs and ideas about behavior (e.g., sex with children is normal) can be reinforced. A corollary of this is that there may be an interaction between the Internet and other factors (i.e., cognitive distortions).

Conclusion

The primary concerns regarding sex offenders involve community safety, accurate assessment of risk and methods for reducing risk. However, as mentioned, the advent of the Internet brings a host of new concerns to the forefront that must be addressed and are still in nascent stages of investigation. Given the failure of more traditional correctional remedies, such as deterrence and incapacitation, for reducing the level of sexual violence in society, other interventions have been actively sought. One potentially effective intervention for known offenders is treatment. As noted, the verdict as to the efficacy of treatment for sexual offenders is inevitably a complex one that addresses: (1) optimal treatment modalities for specific subtypes of offenders, (2) optimal conditions under which treatment and follow-up should occur, (3) selection (or exclusion) criteria for treatment candidates, and (4) fidelity of the treatment program. At the present time, the most informed and dispassionate conclusion must be that the jury is still out.

The essential point, of course, is the legislative intent of laws that commit sexual offenders for indefinite periods of treatment. If the primary intent of the law is preventive detention, as Pacht (1976) argued 35 years ago, and treatment is merely thrown in so that the law will pass constitutional muster, then it is very unlikely that the law will promote a marked diminution in recidivism among released offenders. The task of effecting behavior change in seriously characterologically disordered individuals is, in itself, considerable. When these individuals have not chosen treatment but are placed in treatment by the court (as is often the case), and when the treatment occurs in the context of a dangerous prison environment where the goal is survival not therapy (as is often the case), then the task becomes immeasurably more difficult. When the treatment provided is suboptimal, the conditions (i.e., prison) are highly problematic, and the therapists are inadequately trained and supervised, the task may well be impossible. Under these onerous conditions it would be an error to draw any conclusions about the presumptive inefficacy of treatment with sex offenders.

See also: Rape and Sexual Coercion; Rehabilitation of the offender.

Bibliography

Andrews, D.A., Bonta, J., 2010. Rehabilitating criminal justice policy and practice. Psychology, Public Policy, and Law 16 (1), 39–55.

Andrews, D.A., Bonta, J., Hoge, R.D., 1990. Classification for effective rehabilitation: rediscovering psychology. Criminal Justice and Behavior 17, 19–52.

Andrews, D., Bonta, J., Wormith, S., 2011. The risk-need-responsivity (RNR) model does adding the good lives model contribute to effective crime prevention? Criminal Justice and Behavior 38 (7), 735–755.

Berzins, L.G., Trestman, R.L., 2004. The development and implementation of dialectical behavior therapy in forensic settings. International Journal of Forensic Mental Health 3 (1), 93–103.

Bradford, J.M.W., 1989. The antiandrogen and hormonal treatment of sex offenders. In: Marshall, W.L., Laws, D.R., Barbaree, H.E. (Eds.), Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender. Plenum, New York, pp. 297–310.

Brownell, K.D., Marlatt, G.A., Lichtenstein, E., Wilson, G.T., 1986. Understanding and preventing relapse. American Psychologist 41, 765–782.

Cooper, A., 1998. Sexuality and the internet: surfing into the next millennium. CyberPsychology and Behavior 1 (2), 187–193.

Cornwell, J.K., 1998. Understanding the role of the police and parens patriae powers in involuntary civil commitment before and after Hendricks. Psychology, Public Policy, and Law 4, 377–413.

Eke, A., Seto, M., Williams, J., 2011. Examining the criminal history and future offending of child pornography offenders: an extended prospective follow-up study. Law and Human Behavior 35 (6), 466–478.

Endrass, j., Urbaniok, F., Hammermeister, L.C., Benz, C., Elbert, T., Laubacher, A., Rossegger, A., 2009. The consumption of internet child pornography and violent sex offending. BMC Psychiatry 9, 43–50.

Hall, C.C.N., 1989. Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. Journal of Consulting and Clinical Psychology 56, 773–775.

Hanson, R.K., Bourgon, G., Helmus, L., Hodgson, S., 2009. The principles of effective correctional treatment also apply to sexual offenders: a meta-analysis. Criminal Justice and Behavior 36 (9), 865–891.

Hildebran, D., Pithers, W.D., 1989. Enhancing offender empathy for sexual abuse victims. In: Laws, D.R. (Ed.), Relapse Prevention with Sex Offenders. Guilford Press, New York, pp. 236–243.

Holt, T.J., Blevins, K.R., Burkert, N., 2010. Considering the pedophile subculture online. Sexual Abuse: A Journal of Research and Treatment 22 (1), 3–24.

Kafka, M.P., 1991. Successful antidepressant treatment of nonparaphilic sexual addictions and paraphilias in men. Journal of Clinical Psychiatry 52, 60–65.

Kan. Stat. Ann. Sec. 59-29a01, 1994.

Knopp, J.H., Freeman-Longo, R.E., Stevenson, W., 1992. Nationwide Survey of Juvenile and Adult Sex-offender Treatment Programs. Safer Society Press, Orwell, Vermont.

Koss, M.P., Bachar, K.J., Hopkins, C.Q., 2003. Restorative justice for sexual violence: repairing victims, building community, and holding offenders accountable. Annals of New York Academy of Science 989, 384–396.

Laws, D.R. (Ed.), 1989. Relapse Prevention with Sex Offenders. Guilford Press, New York.

Laws, D.R., Hudson, S.M., Ward, T. (Eds.), 2000. Remaking Relapse Prevention with Sex Offenders. Sage, Thousand Oaks, CA.

Laws, D.R., Ward, T., 2011. Desistance from Sexual Offending: Alternatives to Throwing Away the Keys. Guilford, New York, NY.

Lee, A.F., Li, N.-C., Lamade, R., Schuler, A., Prentky, A., 2012. Predicting hands-on child sexual offenses among possessors of internet child pornography. Psychology, Public Policy and the Law. http://dx.doi.org/10.1037/a0027517.

Linehan, M.M., 1993. Cognitive Behavioral Treatment of Borderline Personality Disorder. Guilford Press, New York.

Marlatt, G.A., Gordon, J.R. (Eds.), 1985. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. Guilford Press, New York, NY.

Marques, J.K., Wiederanders, M., Day, D.M., Nelson, C., van Ommeren, A., 2005. Effects of a relapse prevention program on sexual recidivism: final results from California's sex offender treatment and evaluation project (SOTEP). Sex Abuse: A Journal of Research and Treatment 17 (1), 79–107.

Pacht, A.R., 1976. The rapist in treatment: professional myths and psychological realities. In: Walker, M.J., Brodsky, S.L. (Eds.), Sexual Assault: The Victim and the Rapist. Lexington Books, Lexington, MA.

Pithers, W.D., 1990. Relapse prevention with sexual aggressors: a method for maintaining therapeutic gain and enhancing external supervision. In: Marshall, W.L., Laws, D.R., Barbaree, H.E. (Eds.), Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender. Plenum, New York, pp. 343–361.

Pithers, W.D., Marques, J.K., Gibat, C.C., Marlatt, G.A., 1983. Relapse prevention with sexual aggressive: a self-control model of treatment and maintenance of change. In: Greer, J.G., Stuart, I.R. (Eds.), The Sexual Aggressor: Current Perspectives on Treatment. Van Nostrand Reinhold, New York, NY, pp. 214–239.

Prentky, R.A., Gabriel, A., Coward, A., 2011. Treatment for perpetrators of sexual violence (invited Chapter). In: White, J.W., Koss, M.P., Kazdin, A.E. (Eds.), Violence

Against Women and Children: Navigating Solutions. American Psychological Association, Washington, D.C, pp. 115–135.
Prentky, R.A., Knight, E.A., 1986. Impulsivity in the lifestyle and criminal behavior of sexual offenders. Criminal Justice and Behavior 13, 141–164.
Prentky, R.A., Knight, R.A., 1991. Identifying critical dimensions for discriminating among rapists. Journal of Consulting and Clinical Psychology 59, 643–661.
Rodriguez, N., 2007. Restorative justice at work: examining the impact of restorative justice resolutions on juvenile recidivism. Crime Delinquency 53 (3), 355–379.
Schwartz, B.K., Cellini, H.R. (Eds.), 1995. The Sex Offender. Corrections, Treatment and Legal Practice. Civic Research Institute, Kingston, NJ.
Schwartz, B.K., Cellini, H.R. (Eds.), 1997. The Sex Offender. New Insights, Treatment Innovations, and Legal Developments. Civic Research Institute, Kingston, NJ.
Seto, M.C., Eke, A.W., 2005. The criminal histories and later offending of child pornography offenders. Sexual Abuse: A Journal of Research and Treatment 17 (2), 201–209.
Seto, M.C., Maric, A., Barbaree, H.E., 2001. The role of pornography in the etiology of sexual aggression. Aggression and Violent Behavior 6, 35–53.
Stermac, L.E., Segal, Z.V., Gillis, R., 1990. Social and cultural factors in sexual assault. In: Marshall, W.L., Laws, D.R., Barbaree, H.E. (Eds.), Handbook of Sexual Assault: Issues, Theories, and Treatment of the Offender. Plenum, New York, pp. 143–159.
Tomak, S., Weschler, F.S., Ghahramanlou-Holloway, M., Virden, T., Nademin, M.E., 2009. An empirical study of the personality characteristics of internet sex offenders. Journal of Sexual Aggression 15 (2), 139–148.
Ward, T., Mann, R.E., Gannon, T.A., 2007. The good lives model of offender rehabilitation: clinical implications. Aggression and Violent Behavior 12, 87–107.
Ward, T., Maruna, S., 2007. Rehabilitation: Beyond the Risk Paradigm. Routledge, London.
Ward, T., Yates, P.M., Willis, G.M., 2012. The good lives model and the risk need responsivity model: a critical response to Andrews, Bonta, and Wormith (2011). Criminal Justice and Behavior 39, 94–110.
Webb, L., Craissati, J., Keen, S., 2007. Characteristics of internet child pornography offenders: a comparison with child molesters. Sex Abuse 19, 449–465.
Wolak, J., Finkelhor, D., Mitchell, K.J., 2005. Child Pornography Possessors Arrested in Internet-related Crimes: Findings from the National Juvenile Online Victimization Study (NCMEC 06-05-023). National Center for Missing and Exploited Children, Alexandria, VA.
Yates, P.M., Prescott, D.S., 2011. Building a Better Life: A Good Lives and Self-regulation Workbook. Safer Society Press, Brandon, VT.
Yates, P.M., Prescott, D., Ward, T., 2010. Applying the Good Lives and Self-regulation Models to Sex Offender Treatment: A Practical Guide for Clinicians. Safer Society Press, Brandon, VT.

Replication

Lowell L Hargens, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Rosenthal, volume 19, pp. 13164–13167, © 2001, Elsevier Ltd.

Abstract

The ability to successfully replicate previous work has long been identified as the foundation of scientific progress. Exact replication, in which researchers attempt to duplicate a previous study in every respect, is rare compared to approximate replication, which may produce knowledge about the generalizability of a previous finding or the mechanisms that produced it. In the social and behavioral sciences, replication is potentially an important antidote to publication bias caused by researchers' reliance on statistical hypothesis testing. Few replications are published in those fields, however, because their members tend to value novelty and view replication as pedestrian.

The ability of researchers to repeat previous work and obtain the same results has long been viewed as the foundation of scientific progress (Rosenthal, 1991). To foster this, research articles in the empirical sciences usually include a 'methods and materials' or 'data and methods' section that, at least in principle, provides enough information to enable other researchers to replicate the findings those research articles report (this possibility is often termed 'intersubjective testability' or 'reproducibility'). Researchers in fields where inquiry consists of experiments are expected to have replicated their work before publishing it, and experimentalists who publish results that others cannot reproduce usually suffer damage to their professional reputations.

Exact and Approximate Replication

Replications vary in the degree to which they employ exactly the same methods and materials of an earlier study. Exact replications carried out by researchers trying to test the validity of others' work are rare for several reasons. First, they are often costly and difficult to carry out. Second, exact replications that succeed in producing previously reported results are usually not publishable because they contribute little new knowledge. Third, the status of an attempted exact replication that fails to produce previously reported results is often ambiguous because the researcher carrying out the replication may not have succeeded in using exactly the same methods and materials as the original study (Collins, 1985). Thus, attempts at exact replication tend to be limited to cases where the work to be replicated is highly important theoretically or monetarily.

Approximate replications, wherein researchers modify one of more of the procedures or conditions of the original study, are more frequent and are valuable because they may give information about the generalizability of the original results or the mechanisms that produced them (Aronson et al., 1998; Tsang and Kwan, 1999). In some experimental fields, researchers routinely carry out approximate replications of others' work in order to build directly upon it. Routine successful replication of others' work enables a community of scholars to build a scaffold of knowledge made up of interrelated elements produced by different members of the community. A happy coincidence of frequent successful routine replication is that it gives researchers confidence in the validity of their joint enterprise.

Replication in the Social and Behavioral Sciences

Replication is infrequent in fields, including most of the social and behavioral sciences, where researchers do not typically build directly on previously published results. Novel and counterintuitive findings are often much prized in these fields (Guetzkow et al., 2004), so studies that attempt to build on past results, are usually judged to be uninteresting and pedestrian (pejoratively referred to as being only 'normal science'). In addition, many believe that exact replication is impossible in these fields because changing social and cultural conditions beyond the control of researchers may produce differences in the results obtained by studies even if they do use identical methods. The meaning of a failure to replicate previous results is therefore more ambiguous in the social and behavioral sciences than in the natural sciences (Schuman and Presser, 1981).

Many have argued that dearth of replication in the social and behavioral sciences is an important factor limiting the progress of those sciences. This is because researchers in these fields often rely on the results of statistical hypothesis testing as an important determinant of whether their findings are worthy of publication. Specifically, research projects that do not produce results that pass statistical significance tests are not represented in the research literature because editors and reviewers require statistically significant results, and also because authors, anticipating that their work will be rejected if they submit it for publication, decide to not write papers about their results if the results do not pass statistical significance tests. This pattern is referred to by such terms as 'publication bias' and 'the file drawer problem' (Rosenthal, 1991). In contrast, results that do pass statistical tests, even if they are a result of Type-I error, achieve publication. Many have noted that this system raises the possibility that the research literature of the social and behavioral sciences consist disproportionately

of papers that report misleading results (Sterling, 1959), a condition that hinders knowledge accumulation. Consistent with this possibility, surveys of attempted replications that have been carried out in the social and behavioral sciences report high rates of failure (Tsang and Kwan, 1999). A telling exception to this generalization is that studies of the prevalence of replication studies in social and behavioral science journals uniformly report their absence (e.g., Gaston, 1979; Tsang and Kwan, 1999).

In order to increase researchers' motivation to carry out replications and thereby reduce the pernicious effects of publication bias, some argue that scholarly journals in the social and behavioral sciences should reserve a certain proportion of their pages to papers that report the results of replication attempts. Surveys of editors of such journals, however, find that they, like their colleagues, also view replication studies as pedestrian, and therefore unworthy of publication (Madden et al., 1995).

See also: Hypothesis Testing in Statistics; Resampling Methods of Estimation; Systematic Reviewing and Meta-Analysis.

Bibliography

Aronson, E., Wilson, T.D., Brewer, M.B., 1998. Experimentation in social psychology. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed. McGraw-Hill, New York, pp. 99–142.

Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.

Gaston, J., 1979. The big three and the status of sociology. Contemporary Sociology 8, 789–793.

Guetzkow, J., Lamont, M., Mallard, G., 2004. What is originality in the humanities and social sciences? American Sociological Review 69, 190–212.

Madden, C.S., Easley, R.W., Dunn, M.G., 1995. How journal editors view replication research. Journal of Advertising 24, 77–87.

Rosenthal, R., 1991. Replication in behavioral research. In: Neuliep, J.W. (Ed.), Replication Research in the Social Sciences. Sage, Newbury Park, NJ, pp. 1–30.

Schuman, H., Presser, S., 1981. Mysteries of replication and non-replication. In: Schuman, H., Presser, S. (Eds.), Questions and Answers in Attitude Surveys. Academic Press, New York, pp. 317–324.

Sterling, T.D., 1959. Publication decisions and their possible effects on inferences drawn from tests of significance-or visa versa. Journal of the American Statistical Association 54, 30–34.

Tsang, E.W.K., Kwan, K.-M., 1999. Replication and theory development in organizational science: a critical realist perspective. Academy of Management Review 24, 759–780.

Representations, Social Psychology of

Ivana Marková, University of Stirling, Stirling, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The theory of social representations studies the formation and transformation of meanings and activities of complex social phenomena like health and illness, political problems or environmental issues in and through language and communication, history and culture. There are two mutually interdependent meanings of social representations. The first meaning concerns the theory of social representations as an interactional theory of knowledge. It refers to networks of concepts and figurative schemes that are generated in and through tradition, common sense, daily knowledge, and communication; these are shared by particular groups and communities. The main features of this theory are the Ego–Alter–Object, the field, the interdependence of asymmetries and symmetries, ethics, figurative scheme, and cognitive polyphasia. Second, social representations refer to concrete social phenomena and to forms of apprehending and creating social realities in and through communication, experience, social practices, and interventions. Human thinking is characterized by the capacity to make distinctions and understand phenomena as dyadic antinomies or themata. Thematization of dyadic antinomies is linked with anchoring and objectification, through which social representations are formed and transformed.

What Are Social Representations?

'Representation' is one of the most common concepts in psychology. Like many significant concepts that have been part of human thought for a long time, the meaning of 'representation' has radically changed and diversified throughout history. In particular, during the last century or so it has become the basis of numerous theories in cognitive sciences, sociology, and psychology. Adjectives that characterize a representation as mental, collective, or social, identify very different theoretical perspectives. This fact is often ignored or misunderstood, and this leads to substitution of one kind of representation for another one and so to a confusion of their meanings. In order to understand the concept of social representations, we need to clarify how it differs from other kinds of representations.

Mental Representations

The seventeenth-century philosophers René Descartes and John Locke laid down theories of knowledge in which a mental representation played a significant role; these theories of knowledge still steer contemporary conceptions of mental representations. René Descartes' philosophy conceived of the universe as being guided by mechanistic laws of physics, geometry, and mathematics. The same laws of mechanics also pertained to biological organisms and even to language. For Descartes, the self's rational thinking was the only source of certain knowledge. According to John Locke, individuals acquired knowledge by mirroring elements of concrete reality and so constructing more or less correct mental representations. In these ideas of Descartes and Locke, we already find germs of the main characteristics of mental representations as currently studied in cognitive science. The leading cognitive scientists like Chomsky and Fodor, too, presuppose that the mind (indeed the mind/brain) is a mechanism. Terms like the 'mind', 'mental representation', and 'mental computation' among others refer to capacities of the brain to mirror external reality. Operations of the brain are understood in terms of causes and effects or of purposes, governed by rules or by algorithms, for which no historical explanation is required. Today, the study of mental representations is vast and diverse, with many assumptions built into manifold theories, posing different kinds of questions in psychology, philosophy, and cognitive science. Mental representations are conceived as formalizations of rational individuals. In contrast, common sense concepts are conceived as inferior forms cognition; it is believed that with scientific education they will gradually disappear.

Collective Representations

Emile Durkheim's concept of collective representations that he developed in the later part of the nineteenth century is at a crossroad between the static and the mechanistic mental representation and the interactive and dynamic social representation. Like Descartes and Locke, Durkheim was preoccupied with the question as to how one acquires knowledge of the external world. In contrast to Descartes and Locke, he adopted the point of view that the world cannot be understood through representations of single minds, but through collective representations. Collective representations are socially produced phenomena that circulate, and are shared in society, for example, as religious, mythical, or scientific representations. They arise directly from social structures and they constitute social reality just like physical facts constitute physical reality. Durkheim called them social facts; they are external to individuals, who do not contribute toward their formation. Collective representations impose irresistible pressure on individuals who, yielding to their coercion, internalize and perpetuate these social forms of acting, thinking, and feeling. Durkheim conceived the sociology of knowledge as firmly based on the concept of stable collective representations that reproduce themselves through the process of internalizations in single individuals. He insisted that knowledge must be certain and unchanging, otherwise, it would not be knowledge; therefore, collective representations, in order to have the status

of knowledge, they must be relatively stable. Nevertheless, throughout historical development, collective representations slowly change and become more scientific. Like many researchers of his time, Durkheim privileged scientific knowledge over common sense knowledge and thus he perpetuated the idea that common sense thinking is erroneous and trivial.

Social Representations

The theory of social representations studies the formation and transformation of meanings, knowledge, and activities of complex social phenomena in and through language and communication, history and culture. Examples of such phenomena are health and illness (e.g., AIDS, mental illness, or diabetes), political problems (e.g., democracy, totalitarianism, trust and distrust of institutions), environment, and so on. Such phenomena may present danger, threat, as well as hope, for individuals and groups; they invigorate, innovate, and transform society; they create conflict as well as facilitate conformity. The multifaceted characteristics of such phenomena reveal themselves in and through language and communication of various kinds, e.g., in the media, the public discourse, or interpersonal dialogues.

Serge Moscovici proposed and developed the theory of social representations, both conceptually and empirically, in 'La Psychanalyse: Son Image et Son Public' (Moscovici, 1961/1976; English edition in 2008). This classic explored, in a specific sociopolitical culture in the late 1950s in France, transformations of professional and scientific knowledge of psychoanalysis into everyday thinking and discourses of various social groups and the mass media reporting. It would be wrong to assume that transformation of professional and scientific knowledge into everyday thinking is achieved by developing simplified lay theories. Instead, it is accomplished by, and enriched through, diverse means of communication and images, arguments based on trust and distrust of others, collective memories, conscious and unconscious beliefs, myths and metaphors, fears and hopes. Moscovici's work was driven by questioning the prevalent view that ordinary people do not think rationally (Moscovici and Marková, 2000). He objected to the view that a social scientist should study rational thought, which, in the future, would replace inferior common sense thinking. In contrast, Moscovici asserts that a social scientist should study the daily thinking of citizens and the ways of their creating social reality using images, beliefs, language, and communication. Among the forms of daily thinking, common sense is a legitimate field of study for social scientists. Although Moscovici has not explicitly specified what constitutes the epistemology of common sense, one can suggest that this epistemology is carried and transmitted in and through social interaction, language, and communication. It draws on daily thinking and imagination, on figurative schemes, innovation, and ethics.

Two Meanings of Social Representations

We can distinguish two fundamental meanings of social representations. First, the theory of social representations is an interactional theory of social knowledge – or interactional epistemology. (The term 'interactional epistemology' fits with the aim of this article to present a general conception of social representations. Elsewhere when I present my own position with respect to social representations, I use the term 'dialogical epistemology'.) As such, it institutes networks of concepts and figurative schemes that are generated in and through tradition, common sense, daily knowledge, and communication; these are shared by particular groups and communities. The theory of social knowledge enables the researcher to define research problems. The second meaning of social representations is related to the study of concrete social phenomena, to their apprehension, creation and change in and through communication, experience, social practices, and interventions.

Social Representations as Interactional Epistemology

As a theory of social knowledge based on interactional epistemology social representations involve a number of specific characteristics, among which the following appear to be the most significant.

Ego–Alter–Object

The relation between the Ego–Alter–Object constitutes the pillar of the epistemology of social representations: knowledge and beliefs (Object) are generated both by the self (Ego) and by others (Alter). Therefore, the minimum unit in the formation of social knowledge and beliefs is a triadic relation, the Ego–Alter–Object (Moscovici, 1972/2000, 1984; Bauer and Gaskell, 1999; Marková, 2003/2005; Jesuino, 2008). In this generalized model, the relation between the Ego and the Alter stands for all kinds of selves and others; however, in concrete and contextualized situations of knowledge production there is always the specific Ego and the specific Alter. For example, the Ego can be a child and the Alter can be his or her parents; moreover, this dyad, i.e., child–parents, could be conceived as an Ego within a specific culture (Alter), and so on. Thus, the formation and maintenance of social representations can be conceived as multifaceted structures of knowledge and beliefs produced by embedded kinds of the Ego–Alter. This epistemology presupposes that there can be no single mind without other minds: they dialogically coconstitute one another. I am using the term 'dialogically' with reference to 'dialogicality'. Dialogicality is the fundamental capacity of the Ego to conceive, create, and communicate about social realities in terms of the Alter. What the human individual has become through the work of the past, and what his or her prospects are for the future, results from dialogicality (Marková, 2003/2005). Such language-based interdependencies make the dyadic relations between the Ego–Alter dynamic, with implicit and explicit meanings affecting their discourses and contributing to transformation of representations. Participants contribute to different dialogical perspectives and create tensions among them. Neutrality can be only artificially imposed but speech is always judgmental, evaluative, and oriented toward creating new meanings.

The Field

In interactional epistemology of social representations, social reality is not generated from behavior of single individuals. Instead, it results from interactive forces binding the Ego and

the Alter. Such interactive forces constitute a field. Within the field the parties exert a mutual influence on one another, and they jointly generate an object, that is, new patterns of knowledge, beliefs, and images. The field involves forces (e.g., attraction, volition to communicate, the interest in a particular subject matter, repulsion, coordinated attention and intentions, and so on) between communicants (the Ego–Alter) with respect to the object of study. This is why triangularity in the theory of social representations is the basic presupposition; it separates this theory from individualistic theories of knowledge.

Within the epistemological triad of Ego–Alter–Object, relations between these components can take on different forms and commitments. For example, if the individual searches for knowledge of an Object, he or she might be more strongly committed to that Object than to the Alter. In this case, the individual would examine dispassionately the object of knowledge, even ignoring the general consensus with respect to the accepted knowledge.

Another kind of relation within the triangularity of the Ego–Alter–Object could be based on a strong relation between the Ego and Alter, whereas the relation between the Ego and the Object could be treated as secondary. In this case, knowledge/beliefs could range from irresistible to those that would function as constraints – be it compliance, conformity, or obeisance. Irresistible beliefs created by interdependence between the Ego–Alter can hardly be changed by evidence or by persuasion to their contrary. They can lead to self-sacrifice and other-sacrifice of individuals and groups, rather than to their change. Such strong beliefs within a group, e.g., religious or political, are often based on trust and trustworthiness of the other; convictions, rituals or ideas can take possession of us and they take control or our reasoning (Moscovici, 1993: p. 50).

The presupposition of triangularity, of course, will lead to the inclusion of other components in the process of generating social knowledge, e.g., time (Bauer and Gaskell, 1999), location, cultural resources (Valsiner, 2003; Chryssochoou, 2004), and so on.

Asymmetries and Symmetries

Interactional epistemology of social representations is an epistemology of changes between two opposite tendencies, for example: asymmetries and symmetries, disequilibria and equilibria, creativity/innovation, and conformity/consensus.

It has been well established that changes in physical and biological systems take place through alterations between symmetrical and asymmetrical states. Likewise, formation of knowledge, interactions, and relations between social agents involve, on the one hand, an effort for consensus. Individuals conform to one another make compromises with respect to their opinions and choices and, in general, shift their positions toward convergence in order to reduce conflicts. On the other hand, they may favor opposite tendency toward change rather than stability: breaking the symmetry works through tension, divergence, and differentiation between people, and even through conflict. Interaction breaks invariances. Individuals have a capacity to switch from asymmetry to symmetry, from equilibrium to disequilibrium, from conformity to innovation, and vice versa. The breakdown of symmetry initiates a dynamic order, an order that is evolving, encouraging innovative ways of thinking, differentiating and generating new patterns of knowledge, beliefs, images, and activities (e.g., Galam and Moscovici, 1994). Moscovici argues that the precondition for rationality and coherence of thought is not only collective consensus but that dissensus and contradiction are the other side of rational thought. Therefore, social psychologists must study not only conflict resolution, but also conflict arousal (Moscovici and Marková, 2006: p. 173).

In the study of psychoanalysis (Moscovici, 1961/1976/2008), asymmetries strongly dominate systems of communication like propaganda, diffusion, and propagation. For example, Moscovici's (1961/1976/2008: p. 334) concept of 'semantic barriers' retains specific asymmetries in contents and processes of diverse representations interacting with one another. Moscovici explored semantic barriers particularly in the relation between language and action in propaganda. In order to achieve transformation of beliefs in individuals propaganda creates a specific language by using qualified expressions and so changing meanings of words. Words that originally referred to certain representations as, e.g., psychoanalysis as a therapy, then referred, with a qualified expression 'American', to an alternative representation 'American psychoanalysis', which had an ideological and negative meaning. By distorting meanings of words, propaganda not only erects semantic barriers among groups, but it affects actions, creates conflicts, and postulates new ideologies (see Gillespie, 2008).

Heterogeneity and Cognitive Polyphasia

Language and communication as a point of departure in epistemology of social representations has yet another significance: to communicate means to take diverse routes, leading once to intersubjective understanding between individual and group, once to conflict, or to negotiation and compromise, or to a firm self-positioning. This implies transformation of one kind of knowledge into another one. Transformation of knowledge and messages is pertinent to specific social, historical, and cultural conditions.

One of the basic features of the theory of social representations is its focus on the dynamic coexistence of distinct modalities of thinking, which are present in communicative actions. They contribute to viewing the issue in question from diverse perspectives, which enable formulation of different arguments. Such communication-centered thinking involves a number of directions: it judges, evaluates, criticizes, makes proposals for action and raises controversial issues. Moscovici (1961/1976/2008) calls these diverse modalities of thinking and communicating 'cognitive polyphasia'. Cognitive polyphasia is characterized by tension, conflict, and constraint rather than by equilibrium and adaptation.

The concept of cognitive polyphasia is inherently dialogical. The divergent modalities of thinking are expressed as specific Ego–Alter communications to follow diverse mental paths. Cognitive and emotional goals employ heterogeneous modes of thought which may range from scientific to religious, from literal meanings to metaphoric interpretations, from jokes to formal expressions, and so on. They are suited to and articulated in contexts of which they are parts and in which speakers create links to others' communications, anticipating their responses, reactions, and feelings.

Following the original publication of 'La Psychanalyse' (1961/1976/2008), other researchers have presented many examples of cognitive polyphasia in common sense thinking (for reviews see, e.g., Duveen, 2007; Jovchelovitch, 2007; Wagner and Hayes, 2005). These include cognitive polyphasia in different cultures, daily situations, and various contexts.

Ethics

The theory of social representations places emphasis on humans as ethical beings who pursue passions and make ethical choices. Focus on making ethical choices (Moscovici, 1961/1976/2008) distinguishes the theory of social representations from other current social psychological theories as well as from natural sciences and from those social sciences that are based on deterministic and 'objective' theories.

Only if we describe humans in dialogical terms of the Ego–Alter, can we reasonably speak about relational notions like freedom, justice, solidarity, dignity, or forgiveness. Ego–Alter relations. They are characterized by openness toward multifaceted and heterogeneous realities and their relationships. These relations are never neutral: they are ethical, that is, evaluative and judgmental. This conception therefore preclude the separation of 'facts' and values. Moscovici (2011) traces historical attempts to rationalize and mechanize interpersonal relations, as well as the efforts to draw objectified and mechanistic rules into ethics.

In order to construct knowledge about one another, speakers complement their partial knowledge by imagining what the other could be like and what might be expected from him or her. Imagining could be also involved in an internal dialogue with the self, others, or with circumstances. Moreover, we act on the basis of imaginations created in our internal dialogues. I imagine something and get angry: imagination is communicative, it is dialogical. Words may convey specific images over centuries.

Ethical choices are particularly visible in ceremonies of name giving, bringing into the focus of attention ways, in which humans reinforce the importance of creations of new social realities and fix them in and through language; giving something a name means giving the thing a new content and meaning. This is an imaginative enterprise expressing ethical choices of the namer: it conveys hopes, it judges and condemns others, it evaluates. Names not only stabilize social representations but also initiate new ones. Moreover, by naming we impose limits, properties, and structures on the named object, person, or phenomenon. In order to amplify the impact of names we use metaphors as the ways of objectifying ideas and turning them to reality.

Figurative Scheme

A social representation is structured around a figurative scheme that underlies images, concepts, or judgments that a group or society has generated over time. A figurative scheme often takes the form of a dyadic antinomy. For example, we understand freedom in contrast to what we consider to be a lack of freedom; justice is understood through absence of justice; logos is contrasted with mythos, and so on. Antinomies are features of thinking, language, and communication in all cultures, but different cultures and societies employ their capacity of making distinctions and thinking in antinomies in specific ways. We find them throughout eons of human history both in scientific and in common sense thinking, although very often they are present implicitly without becoming an explicit discursive topic. Sociocultural changes, however, may bring implicit antinomies to the public awareness and into discourses, reflecting societal tensions and conflicts. This means that from that moment on, antinomies turn into 'themata', whether in scientific thinking where, according to Geoffrey Holton, they generate scientific theories, or in common sense thinking where they generate social representations (Moscovici and Vignaux, 1994/2000). Analyzing a figurative scheme in the case of ethnic minorities, e.g., gypsies, Moscovici (2011) shows it is articulated around the thema nomadic/sedentary. He argues that this thema is as basic as, for example, the thema left/right in social representations of political parties. A figurative scheme associates an ethnic minority with a number of themata that form mutually interdependent networks. For example, the thema pure/impure (clean/dirty) defines the minority's existence as a deviance within the native population. When the Spanish Inquisition introduced the racial concept of 'purity of blood', gypsies were regarded as a 'wicked nation of beggars and idlers' (thema diligent/lazy). However, they were Christians (Christians/Jews thema) and therefore, they were not expelled from Spain; but because they were impure, they could live only in specific places and were not allowed to leave those: "Blood was no longer evoked, but purity and impurity became concrete realities" (Moscovici, 2011: p. 455) and this was why they could not mix with the native majority.

Social Representations as Phenomena and as Interventions

The second meaning of social representations concerns concrete social phenomena that are part of daily life, of public discourses and that human agents continuously make and remake. In other words, in the process of re-presenting, individuals and groups act and intervene in life practices. This means that the theory of social representations assumes that when interacting, humans use resources from culture and traditions as well as from their living experience. Historical/cultural and contemporary resources may be in conflict, leading to disruption of routines, turning them upside down, and calling for action. Forces (e.g., commitments, attractions, etc. see above) between different kinds of Ego and Alter constitute a field for re-presenting complex social phenomena that are in the center of social life and social disputes, e.g., political, ecological, or health related. Such complex phenomena obtain their meanings in interdependence with and in relation to other representations within that culture and community. For example, the representation of democracy would be interdependent with other social representations within that particular culture, like representations of minority groups, of freedom of speech, economic situation of the country, and so on.

The relation between social representations and intervention practices is itself an object of research practice (Abric, 1994), e.g., of health research (Jodelet, 2006; Marková and Farr, 1995), education (Garnier and Rouquette, 2000), or politics. Intervention should allow for exchanges between traditional and new forms of knowledge, the preservation of culture, and its negotiation with emerging alternatives in

society (Jodelet, 2013). Doise (2002) regards social representations of human rights as interventions into social relations, whether these concern relations between individuals and groups or individuals and institutions. Human rights must be clearly defined precisely because they are interventions of one kind or other. Intervention strategies for the improvement of the quality of life would depend on a number of issues like the meaning of well-being, of life satisfaction, social welfare, as well as economic possibilities. Throughout his career, Moscovici has persistently insisted that the theory of social representations is – or should be treated – as anthropology of contemporary culture. Cultural anthropologists are concerned with the totality of life of social groups under study – that is, with beliefs and knowledge, myths, images, as well as with social practices in daily living. To understand these phenomena, it is necessary to study them in relation to one another, like meaningful wholes, rather than as independent elements.

Formation and Transformation of Social Representations

Let us now evoke the two basic processes that generate social representations: anchoring and objectification (Moscovici, 1961/1976/2008, 1984). In and through anchoring and objectification individuals and groups conventionalize objects, persons, and events by categorizing them and giving them distinct and shared meanings.

Anchoring is the process that familiarizes the strange phenomenon by classifying and naming it. Naming and categorization is not a neutral process but it involves judgment, evaluation and imagination; in other words, it is ethical. It builds on past knowledge, beliefs, and images. By anchoring humans reveal their theories about others, e.g., about a particular ethnic group.

Objectification is a more in-depth process, infusing anchored phenomena with a concrete reality. The idea of objectification is effectively expressed in Moscovici's claim that the phenomenon that is originally perceived in the process of anchoring becomes, in the process of objectification, conceived, that is, it becomes conceptualized as real. The interdependence and transformation of the anchored, i.e., the perceived phenomenon into an objectified, i.e., a conceived phenomenon, can be observed particularly in language. Naming embraces both anchoring and objectification. For example, the term like Freud's 'Oedipus complex' may be originally perceived as a strange and intellectual phenomenon and becomes associated with, or anchored to, the boy's jalousie of his father. Subsequently, through the use of 'Oedipus complex' in daily language, the term becomes objectified. It turns into designated ways in which relations between parents and children are organized, and it helps to interpret certain tensions in such relations and to refer to them as pathological symptoms. In other words, elements of scientific terminology turn into signs that correspond to a particular reality and become reified within that reality. Thus we may say about someone that he/she is stubborn or quarrelsome; through repeated use in language this label becomes conceived as a real feature of that person and subsequently, this 'reality' becomes incorporated into habitual social practices with respect to that individual.

Graphic representation is another form of objectifying ideas into entities. Written texts, posters, or pictures, all turn an idea into real graphic objects. Visual images in the press, advertisements, and campaigns are used to influence or change social representations of political or health issues. Visual images in the press have been particularly influential in staged photographs capturing public images about genetic engineering as injecting tomatoes with genes that make them grow bigger. Wagner and Hayes (2005: p. 181) comment that viewers associate images of genetically modified tomatoes with inoculation and with injecting foreign materials into human bodies as they know it from medicine and chemistry. There is also a related belief of infection that passes from one organism to another. These examples show that a new reality and new social representations are created by means of semiotic signs and symbols and their transformation.

We have already noted that one of the fundamental features of human thinking is making distinctions and understanding phenomena as antinomies, and that a figurative scheme of the representation often takes the form of a dyadic antinomy – or a thema. Themata that generate most social representations are those pertaining to the Ego–Alter, like private/public, morality/immorality, justice/injustice, and freedom/oppression, among others. Such themata are in the heart of social sciences, and they generate social representations of phenomena like democracy, citizenship, quality of life, and health and illness, to name but a few. Here we come to the point that links themata with anchoring and objectification. Many antinomies are implicitly present in common sense thinking for centuries, and they may never be brought to explicit awareness. However, when a new and strange phenomenon is perceived it may become anchored to such an implicit dyadic antinomy. For example, the emergence of HIV/AIDS in the 1980s anchored this disease to morality/immorality and led to the creation of a new figurative scheme. The existing relational category of morality/immorality became rethematized and it objectified new contents, values, images, and has generated a social representation of a newly emerged object, i.e., of HIV/AIDS. A social representation is rarely generated from a single thema; in the case of HIV/AIDS, rethematization of morality/immorality has been associated with rethematization of social values related to sexuality, promiscuity in the general public, discrimination of minorities, and social recognition, among other issues. Although the antinomy morality/immorality itself has not been questioned, the content and context of morality/immorality has been differently thematized in different fields (see above) in which the social representation of HIV/AIDS has been engaged. For example, the question of personal and social responsibility, medical confidentiality, and human rights all became part of discourse in such specific fields. Communicative processes, through which these changes in meanings are usually achieved, carry symbols and images, which not only circulate in public discourses but also organize and generate discourses. They shape common thinking, language, and behavior, and provide grounds for the formation of new social representations.

Different Approaches to the Study of Social Representations

Following the publication of 'La Psychanalyse', social representing has been studied in various social, political, health-related, and other kinds of phenomena circulating in public discourse (for a comprehensive review, see Wagner and Hayes, 2005). Explorations of social representations have become widely differentiated. As a result, researchers now speak about social representational approaches – or schools of social representations – rather than about a single theory. For example, there is the Aix-en-Provence school based on structuralistic approach that emphasizes central nucleus and periphery of representations (e.g., Abric, 1994; Guimelli, 1994), whereas the Genevean school of Doise specifies organizing principles of social representations (Doise, 1986). Other approaches are anthropological and cultural (e.g., Arruda and de Alba, 2007; Chryssochoou, 2004; Jodelet, 1989/1991). Still others (Wagner and Hayes, 2005; Duveen, 2007) bring to attention the role of social construction and discourse; and Valsiner draws on the role of semiotic mediation and social experience (e.g., Valsiner, 2003). In addition, one can hardly discuss social representations and culture without foregrounding language, communication, and, more specifically, dialogicality as a major feature of the relation between social representations and culture (Marková, 2003/2005; Valsiner, 2003).

Methods of Studying Social Representations

The growing interest in theoretical and empirical research in social representations also highlights its challenges. Many studies, in order to explore social representations of specific phenomena, use *inductive and deductive* methods, and they assimilate social representations to research based on traditional epistemology and on using statistical methods by proving of disproving hypotheses. In contrast, Moscovici argues that surprising findings in research come from discovery and invention rather than from inductive methods by proving hypotheses (Moscovici, 1992).

The difference between method of proof and method of discovery draws attention to the fact that in the latter, the researcher must examine the social representation of the phenomenon (e.g., of psychoanalysis, of AIDS) as interdependent with the relevant situation (e.g., social, historical, political). If we consider Moscovici's study of psychoanalysis, we find that historical, political, and social forces of these studies are just as important as interviews and observations of activities in which the participants were involved. They both constitute the field (see above); they are studies of single cases. Generalization based on single case studies is of a different nature than inductive generalization. In using a single case study one does not generalize the findings to a population but to the theory of the studied phenomenon. Moscovici (1961/1976/2008: p. 175) speaks about 'analogical generalization', which does not refer to generalization to populations as in inductive research. Instead, it refers to the possibility of transformation of one kind of knowledge into another one, using a preliminary theory of the phenomenon in question. The study of psychoanalysis captured social representations during a cultural fight between ideas expressed by the Catholic Church and the strong Communist Party in France in the late 1950s and early 1960s. In this single historical event, it was the coexistence of strong contradictory forces that interacted with common sense thinking that generated social representations of psychoanalysis. However, the findings from this single historical event with respect to different kinds of intellectual polemics, forms of thinking and the clash with established values are transferable to other kinds of social representations – or we could say – to the class of events represented by that case. The forces that form the field do not leave any of the components stable over time, whether the data gathered from participants (interviews, the media) or the data constituted by the relevant historical, political situations, and cultural diversities. Of course, this does not mean that the researcher should not study attitudes or opinions using inductive techniques. Researchers have many choices and it is fully legitimate to use survey, experiments, or any other methods in order to study specific questions, as long as the researcher realizes that it could be problematic to call such research a study in social representation. Implicit knowledge and unconscious processes are characteristics of the human mind just as much as explicit knowledge and conscious processes; equally, implicitness and unconsciousness are features of dialogicality and of the theory of social representations.

See also: Levels of Analysis in Social Psychology; Risk: Social Psychological Perspectives; Social Psychology.

Bibliography

Abric, Jean-Claud (Ed.), 1994. Pratiques sociales et représentations. Presses Universitaires de France, Paris.
Arruda, Angela, de Alba, Martha (Eds.), 2007. Espacios imaginarios y representaciones sociales. Aportes desde Latinoamérica. UAM-Anthropos, Barcelona.
Bauer, Martin, Gaskell, George, 1999. Towards a paradigm for research on social representations. Journal for Theory of Social Behaviour 29, 163–186.
Chryssochoou, Xenia, 2004. Cultural Diversity. Blackwell, Oxford and Malden.
Doise, Willem, 1986. Levels of Explanation in Social Psychology. Cambridge University Press, Cambridge.
Doise, Willem, 2002. Human Rights as Social Representations. Routledge, London.
Duveen, Gerard, 2007. Culture and social representations. In: Valsiner, Jaan, Rosa, Alberto (Eds.), The Cambridge Handbook of Sociocultural Psychology. Cambridge University Press, Cambridge, pp. 543–559.
Galam, Serge, Moscovici, Serge, 1994. Towards a theory of collective phenomena. II: conformity and power. European Journal of Social Psychology 24, 481–495.
Garnier, Catherine, Rouquette, Michel-Louis (Eds.), 2000. Représentations sociales et éducation. Editions Nouvelles, Montreal.
Gillespie, Alex, 2008. Social representations, alternative representations and semantic barriers. Journal for the Theory of Social Behaviour 38, 375–391.
Guimelli, Christian (Ed.), 1994. Structures et transformations des représations sociales. Delachaux et Niestlé, Neuchâtel.
Jesuino, Correia Jorge, 2008. Linking science to common sense. Journal for the Theory of Social Behavior 38, 393–409.
Jodelet, Denise, 1989/1991. Madness and Social Representations. In: Duveen Gerard (Ed.) (T. Pownall, Trans.). Harvester Wheatsheaf, London.
Jodelet, Denise, 2006. Culture et pratique de santé. Nouvelle Revue de Psychosociologie 1, 219–239.
Jodelet, Denise, 2013. Interconnections between social representations and intervention. In: Silvana de Rosa, Annamaria (Ed.), Social Representations in the Social Arena. Routledge, London, pp. 77–88.

Jovchelovitch, Sandra, 2007. Knowledge in Context: Representations, Community and Culture. Routledge, London & New York.

Marková, Ivana, 2003/2005. Dialogicality and Social Representations. Cambridge University Press, Cambridge.

Marková, Ivana, Farr, Rob (Eds.), 1995. Representations of Health, Illness and Handicap. Harwood, Chur.

Moscovici, Serge, 1961/1976/2008. La psychanalyse: son image et son public. PUF, Paris. (M. David, Trans.). Psychoanalysis: its image and its public. Polity Press, Cambridge.

Moscovici, Serge, 1972/2000. Society and theory in social psychology. In: Joachim, Israel, Tajfel, Henry (Eds.), The Context of Social Psychology. Academic Press, London & New York, pp. 17–68. Reprinted from: Social representations: explorations in social psychology. In: Duveen Gerard (Ed.). Polity Press, Cambridge, pp. 78–119 (2000).

Moscovici, Serge, 1984. The phenomenon of social representations. In: Farr, Rob, Moscovici, Serge (Eds.), Social Representations. Cambridge University Press, Cambridge, pp. 3–69.

Moscovici, Serge, 1988. La Machine à faire des dieux: sociologie et psychologie. Fayard, Paris. (W.D. Hall as The invention of society, Trans.). Polity Press, Cambridge, 1993.

Moscovici, Serge, 1992. The discovery of group polarization. In: Granberg, D., Sarup, G. (Eds.), Social Judgement and Intergroup Relations. Springer, New York, pp. 107–127.

Moscovici, Serge, 1993. The return of the unconscious. Social Research 60, 39–93.

Moscovici, Serge, 2011. An essay on social representations and ethnic minorities. Social Science Information 50, 441–461.

Moscovici, Serge, Vignaux, George, 1994. Le concept de thêmata. In: Christian, Guimelli (Ed.), Structures et transformations des représentations sociales. Delachaux et Niestlé, Neuchatel, pp. 25–72. Reprinted from: Social representations. In: Gerard, Duveen (Ed.). Polity Press, London, pp. 156–183 (2000).

Moscovici, Serge, Marková, Ivana, 2000. Ideas and their development: a dialogue between Serge Moscovici and Ivana Marková. In: Moscovici Serge (Ed.), Social Representations. In: Gerard, Duveen (Ed.). Polity Press, London, pp. 224–286.

Moscovici, Serge, Marková, Ivana, 2006. The Making of Modern Social Psychology. Polity, Cambridge.

Valsiner, Jaan, 2003. Beyond social representations: a theory of enablement. Papers on Social Representations 12 (7.1–7.16).

Wagner, Wolfgang, Hayes, Nicky, 2005. Everyday Discourse and Common Sense. The Theory of Social Representations. Palgrave, Basingstoke.

Reproduction and Cultural Anthropology

Elizabeth Roberts, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reproduction is central to contemporary anthropological theory, intersecting with studies of kinship, biomedicine, exchange, sex, gender, politics, nature, and capitalism. The production of reproduction as a category and its domaining as feminine, private, and biological was central to the modern Euro-American separation of industrial wage labor from domestic labor and care, that is, between reproduction and production and between nature and culture. Feminist anthropology of the 1970s called for a focus on women's lives, laying the foundation for an explosion of interest in reproductive experience, as well as critique of the category itself, as both natural and naturally private and female.

Reproduction is central to contemporary anthropological theory, intersecting with studies of kinship, medicine, science, exchange, sex, gender, politics, nature, disability, capitalism, life, humanity, morality, and temporality, to name a few. This has not always been the case. Until recently, reproduction was dismissed as apolitical and irrelevant to the 'big ideas' of anthropological theory, associated as it is with women – what their bodies and biologies do to have and care for children – through menstruation, conception, gestation, childbirth, nursing, and menopause (Taylor, 2004). In Europe and North America, these reproductive processes have tended to be understood as making individual, new people, within the private, intimate domain of heterosexual coitus, separate from political and economic life.

The production of reproduction as a category and its domaining as feminine, private, and biological was central to the eighteenth- and nineteenth-century Euro-American separation of modern, industrial wage labor from domestic labor and care, that is, between reproduction and production and between nature and culture. In the late nineteenth century, this productive/reproductive division shaped the ideal unit into a couple, consisting of an active, public, productive male and a coy, domestic, reproductive female whose intimate reproductive capacities required intensive surveillance and regulation. Although the domaining of reproduction as private, feminine, and biological makes it seem nonpolitical, some of the fiercest political battles of the nineteenth, twentieth, and now twenty-first centuries such as population control, the regulation of prostitution, homosexuality, abortion, and the beginnings of 'life' have been reproductive.

Early anthropologists' investigations of what we now call reproduction were part of the study of kinship among primitives. With hindsight, it is apparent that these investigations said as much about the Euro-American organization of reproduction (as private, biological, and female) as they did about the lives of their subjects and objects of study. Kinship then, through the mid-twentieth century, was the study of how primitive peoples culturally organized their natural reproductive processes, reinforcing the divide between nature and culture, which informed so much of the twentieth-century anthropological project.

Feminist anthropology of the 1970s called for an increased focus on women's lives, laying the foundation for an explosion of interest in reproductive experience, as well as critique of the category itself, as both natural and naturally private and female. More recently, cultural, medical, and feminist anthropologists have begun to trace the entanglements of the material world and symbolic processes; thus the study of reproduction has expanded to include informatics and artificial life, regenerating earlier connections between reproduction and production. These studies are part of a larger shift in cultural anthropology toward including the material and biological world as important objects of study, challenging the assumption that there is a multiplicity of cultural means to organize a singular set of underlying universal and fixed biological processes. This article, then, is about the anthropology of reproduction, but also, importantly, about how both anthropology and reproduction have been constitutively shaped as categories within a modern cosmology that claims that the world is divisible into separate spheres.

The Nature of Reproduction and the Culture of Kinship

Feminist historians and anthropologists describe how the concept of reproduction is thoroughly embedded in the immense shifts wrought by the industrialization of Europe and North America, which radically reconstituted relations and essentialized, as natural and apolitical, new distinctions between humans and nonhumans, women and men, adults and children, as well as classes and races (Daston, 1998; Foucault, 1990; Latour, 1993). The word 'reproduction' came into common use in the eighteenth century, and emphasized abstracted, non-agentive, biological processes as distinct from production, which was increasingly understood as agentive, mechanical, and male (Jordanova, 1995). Through the development of the biological sciences, reproduction came to be naturalized and universalized as a set of processes that take place in individual bodies, especially the bodies of women. The nuclearization and privatization of family groupings through capitalist wage labor resulted in the conflation and marginalization of women's reproductive abilities and sexuality, as well as the association of children with mothers instead of fathers, and the exclusion of other important kin relationships (like siblings) from productive–reproductive processes. Although the domestic reproductive sphere was crucial for the continuance of capitalist production,

and the domain through which private property was transferred, the family, especially women and children, became the sentimental 'haven in a heartless world,' which required protection from the economic rapaciousness of the (male) market (Lasch, 1977). The productive–reproductive division had far-reaching effects, such as making economic calculations in relation to family life offensive to bourgeois propriety, and making women, as newly noneconomic actors, into nonpolitical subjects (Jordanova, 1995; Kant, 1949).

One of the key projects of early twentieth-century anthropology was to understand how 'natives' organized themselves when they seemed to have no political systems: a lack that served as a rationale for the colonial civilizing project and a means for further anthropological distinctions between the West and the rest. Thus, instead of political systems, anthropologists found kinship systems among the primitives they studied (Franklin, 1997). A central puzzle became why these kinship systems did not match the 'facts' of biological reproduction, or biogenetic assumptions as they had been established in eighteenth- and nineteenth-century Europe and codified through Darwinian evolutionary theory. Within nineteenth century and early twentieth century models of unilinear cultural evolution, then, accurate knowledge of biological reproduction, most importantly paternity, was a step toward civilization. Primitive people, the story went, organized their kinship culturally, while Euro-Americans recognized their relatedness through natural biological principles, furthering the formative anthropological divide between nature and culture (Degler, 1991; Franklin, 1997; Weiner, 1995). This distinction between the nature of reproduction and the culture of kinship played out in Bronislaw Malinowski's classic ethnographic work in the Trobriand Islands (Malinowski, 1932). Malinowski argued that Trobrianders did not understand the true link between coitus and conception or recognize biogenetic paternity, believing instead that spirits inseminated women, a formulation that shaped the twentieth-century anthropological division between natural and social facts (Weiner, 1995). Malinowski's account set off the heated mid-twentieth century 'virgin birth debates' about the ability or inability of Trobriand natives to understand reproductive processes. While anthropologists like Edmund Leach and Melford Spiro argued about what the natives knew, they never examined how situated the biological facts of life were within the social, political, and economic conditions of late nineteenth-century Europe and North America (Franklin, 1997). As several recent feminist anthropologists have pointed out, Malinowski's claim was based on his assumption that coitus is the one and only reproductive act, which was itself different from earlier European models of conception, that for instance saw female orgasm as essential to conception or conception as a more environmental process (Weiner, 1995). Malinowski's Victorian–Darwinian model of conception also assumed maternity as much as paternity, whereas in the Trobriands neither men nor women were seen as generators of children, although both contributed to their being. Furthermore, while Malinowski's Trobriand work incorporated theories of exchange as well as reproduction, his male-centered biogenetic assumptions prevented him from noticing how women's power came through siblings, especially sister/brother dyads, that linked reproduction to economic and cosmological relations. As Weiner argues, in the Trobriands, the organization of these relations was not about nature but about power, and gender-based power was not only or even primarily constituted through the intergenerational transmission of substance, but through siblings' exchange relationships (Weiner, 1995).

The model undergirding Malinowski's analysis, where kinship is the cultural means to organize biological reproduction, is omnipresent in some of the twentieth century's most celebrated kinship studies (Franklin, 1997). To name just two: Lévi-Strauss's model of exchange as the origin of social life never departs from the biogenetic assumption that defines kinship and society as the exchange of control over women's biologically ordained procreative role (Lévi-Strauss, 1969). This model posits women as only reproducers and ignores how intergenerational relationships can be produced outside of heterosexual marriage through sibling relations and larger kin groupings (Rubin, 1975; Weiner, 1995). And even while David Schneider's account of American kinship was crucial to the emergence of critical feminist studies of kinship and reproduction, his model preserved the distinction between fixed natural reproductive facts and malleable social facts (Schneider, 1980). An assumption underlying both of these classic kinship studies is that the biological nature of reproduction remains outside the purview of scholarship in cultural anthropology. Reproductive processes remained under-theorized on two counts; as both biological and female, reproduction was not cultural or political.

The Anthropology of Reproduction: 1980–2000

Feminist anthropologists from the 1970s on shifted anthropology's attention to women's lives and experiences, which facilitated a florescence of research on reproduction in cultural anthropology. For the most part, these studies have demonstrated how reproduction – which can seem so private, personal, intimate, medical, natural, and biological – is thoroughly political and economic, serves as a key site of global processes, and is saturated with stratifications of race, labor, class, and sex (Jordanova, 1995). 'Making reproduction central to social theory,' which is exactly what Rayna Rapp and Faye Ginsburg did in their formative edited volume *Conceiving the New World Order*, allowed for a reexamination of some of the most long-standing theoretical debates in anthropology: the relationship of nature to culture, of production to reproduction, between the individual body and the body politic, and the examination of how power works, making reproduction a vital category for anthropological understanding of the world (Ginsburg and Rapp, 1995: p. 1).

Late twentieth-century anthropological studies of reproductive processes and practices – such as pregnancy, abortion, the treatment of infertility, birth control, adoption, conception, miscarriage, sterilization, the collection of vital statistics, the analysis of demographic trends, the codification of the categories of maternal and infant mortality, birth, menstruation, menopause, the medicalization and individualization of reproductive processes, reproductive technologies, parenthood (both same-sex and heterosexual), and child-rearing – used several intertwined theoretical approaches, especially political-economic and poststructuralist accounts that examine

configurations of power, knowledge, and subjects. Anthropologists have shared these approaches with historians, sociologists, literary critics, and political scientists, as they serve to construct critical genealogies of the Euro-American historical formations that produced the emergence of the category of reproduction (and anthropology, for that matter). In addition, anthropological accounts of reproduction have also tended to be comparative, a method specific to anthropology, which has served to denaturalize the category of reproduction itself and its near-automatic association with the female, the private, and the biological.

Political-economic approaches to reproduction often focus on the transformation of Western Europe to modern industrial capitalism, which reorganized social life through distinctions between production and reproduction, private and public, the economic and the domestic, and the intensification of the globally unequal distribution of material resources. Much of the recent literature on the relationship of contemporary consumptive practices to the reproduction and personhood of children takes its cue from the historical shift between productive children, who contributed to household economies, and priceless children, who are consumed for by their parents (Anagnost, 1995; Layne, 1999; Taylor et al., 2004; Zelizer, 1994). Emily Martin's cultural analysis of reproduction linked narratives and experiences of conception, birth, and menstruation in the United States to contemporaneous modes of production throughout the twentieth century (Martin, 1992). One of the principal political economic concepts in the literature has been that of 'stratified reproduction,' a term coined by Shellee Colen to describe "the power relations by whom some categories of people are empowered to nurture and reproduce while others are disempowered" (Ginsburg and Rapp, 1995: p. 3). The concept has been taken up widely, used within Colen's own work on West Indian nannies working for affluent families in New York City and in recent work on transnational surrogacy in India (Colen, 1995; Whittaker, 2010).

The development of medical anthropology in the late twentieth century is crucial to understanding the emergence of the anthropology of reproduction. Medical anthropology began as an applied endeavor, in which anthropologists assisted biomedical, public health, and development experts, both internationally and domestically, in understanding how cultural context contributes to underlying biological etiologies of ill-health (Foster and Anderson, 1978). Over time, more critical strains of medical anthropology developed. One critical framework examines the political-economic relations of the global distribution of health care (Singer and Baer, 1995). The analysis of maternal mortality as a key indicator of social and economic inequality is central to this approach, which seeks to address public health and biomedical experts within their own language and categories, in order to affect policy and reduce health inequalities (Berry, 2010; Janes and Chuluundorj, 2004). Another critical medical anthropological framework has been to examine the production of these powerful categories themselves, such as maternal and infant mortality, women's reproductive health, and reproduction, using both political economic analysis and poststructuralist approaches to the production of knowledge claims and their effects in the world (Adams, 2000; Adams and Pigg, 2005; Scheper-Hughes, 1992). Currently, medical anthropology is constituted through tensions between these two related but disparate frameworks (Fassin, 2012). Although the former framework undoubtedly has more effects on health policy outside of anthropology, the rest of this article will focus primarily on the latter because it offers a better means to examine the elaboration of the category of reproduction within anthropology itself.

A vast number of anthropological studies of reproduction combine political economy with poststructuralist analysis of how biology came to have no politics, producing subjects whose roles were naturalized in the political-economic order. By the late nineteenth century in Western Europe and North America, the domain of what we think of as reproduction was formulated and regulated in the name of increasing the life and health of populations (as working and fighting forces) through heterosexual reproduction, understood as natural biological processes. Practices that were understood to impede the right kind of reproduction, such as childhood masturbation, prostitution, homosexuality, abortion, and the use of birth control, were codified, managed, or outlawed for the first time. To account for these transformations, the social theorist Michel Foucault developed the term 'biopower' to describe how scientific medicine deemed as apolitical became a powerful tool for governance in post-Enlightenment Europe. Through the codification of sex and race as innate and natural, the newly biological body became embedded within increased regimes of surveillance (Foucault, 1990). As the philosopher Judith Butler contends, "The materialization of bodies as matter-prior-to-power is in fact power's greatest effect" (cited in Murphy, 2011: p. 24). These insights about power as deriving from knowledge of the natural world, deemed apolitical, have greatly influenced many of the scholars working on the regulation of reproduction through the vital politics of abortion, embryos, population control, eugenics, birth control, women's reproductive health, and life itself (Casper, 1998; Erikson, 2001; Ginsburg, 1998; Landecker, 2003; Mamo and Fishman, 2001; Morgan, 2009; Rapp, 1999). Many of these studies analyze these practices through the framework of medicalization, which describes how processes come to be understood and managed through a primarily medical lens, which brackets the political and social world.

While Foucault delineated biopower through sex, several theorists have pointed out that sex as a domain is somewhat different now than in the late nineteenth century, partly because biological reproduction can occur through other means than heterosexual coitus (e.g., *in vitro* fertilization (IVF), artificial insemination) and there are now more means to prevent conception within coitus (Rabinow and Rose, 2003; Strathern, 1992a). Given the Euro-American truism that coitus is essential for reproduction, there is surprisingly little overlap in the study of sex practices in relation to reproduction. Perhaps the paucity of focus on sex acts in the anthropology of reproduction arises from the way power imbalances across sex, race, and class have contributed to many forms of troubling heterosexual sex, like rape, that also trouble reproduction (for a rare exception, see Weismantel, 2004). While there are relatively few studies that investigate sex acts of any kind in relation to reproduction, the critique of the naturalization of reproduction to women – for example, the collapse of all of 'women's health' to 'reproductive health' – has compelled a burgeoning literature on male

reproduction, connected to larger moves within gender studies that examine how men continue to be the unmarked sex (Dudgeon and Inhorn, 2004; Gutmann, 2007; Inhorn, 2006).

Contestations over homosexuality (understood as nonreproductive sex) and assisted reproduction (understood as nonsexual reproduction) in Europe and North America in the last decades facilitated a resurgence of interest in kinship in anthropology (Franklin and McKinnon, 2001). Queer kinship and same-sex parenting, in connection with the recently expanding literature on care, have contributed to anthropologists' ability to analyze the specificity of heterosexual, genealogical reproduction (Borneman, 2001; Lewin, 1993). Likewise, reproductive technologies like IVF and egg donation allow anthropologists to examine reproduction outside of heterosexual intercourse as well as investigate how assisted reproduction has combined both the natural and the artificial, established, and distinct domains since the eighteenth and nineteenth centuries (Strathern, 1992b; Thompson, 2005).

In the industrialized world, consisting of settler nations and nation-states made wealthy through colonial capitalism, the wrong and right kinds of reproduction were often defined through biological race, which continues to shape policy and practice within nations and transnationally, especially in the realm of international population policy and migration (Fraser, 1995; Goldade, 2007; Morgan and Roberts, 2012; Pashigian, 2009). Nineteenth- and twentieth-century battles about slavery, miscegenation, immigration, population control, female sexuality, eugenics, and Nazism, which played out on national and international stages, intertwined through apprehensions about the reproduction of cheap labor, the distribution of resources, and the intractable or mutable nature of groups deemed to be fit or unfit reproducers (Douglass, 2005; Horn, 1994; Nouzeilles, 2003; Stepan, 1991). Even though race is by no means a universal form of human organization, the ubiquity of governance through biological reproduction tends to avail itself of the specific stratifications that maintain the hierarchies at work within and between nation-states. And while having and caring for children has not been universally practiced as a private domestic matter governed through medicine, increasingly states outside of North America and Western Europe codify and govern reproduction as a biological matter of great interest to the nation (Greenhalgh, 2008; Tarlo, 2003; Unnithan-Kumar, 2004).

Along with historians, sociologists, and political theorists, anthropologists have illuminated how reproduction became a crucial site of governance within Euro-American politics and economics that simultaneously produced individual bodies filled with apolitical, universal biological processes. In addition, the anthropology of reproduction engages in comparisons demonstrating how having and caring for children has been organized and experienced outside of a Euro-American reproductive framework, which provide a means to see how categories, institutions, and practices arise in particular times and places. Anthropologists usually make these comparisons between people in the nonindustrialized or the developing world and Euro-Americans, although increasingly they make comparisons outside of this particular dyad. Comparisons are fraught with difficulties because of the tendency to see differences as natural, timeless, and fixed, and most problematically as arising separately from one other – as though the historical, economic, and environmental processes that shape one place do not shape another. Through comparison, though, some of the twentieth century's most heated controversies involving reproduction have been made explicable and local, allowing for recognition that there could be other ways of organizing the world.

One of the central insights of these studies is that while reproduction seems to be located in individual families in Western Europe and North America, it is often organized as explicitly collective in nonindustrialized locations, whether on the scale of the extended family, village, community, region, or nation. Another more recent insight parses how key theoretical innovations in the mid-twentieth century, such as gift exchange (Levi-Stauss, Mauss) and the symbolic ritual process (Victor Turner, Van Ganep), were based on practices narrated by male anthropologists as magico-religious childbirth and conception rituals, understood as nonpolitical. With reanalysis, these accounts demonstrate the absolute centrality of these processes to political life in nonindustrialized settings.

Anthropologists have analyzed practices like adoption, abortion, birth, and assisted reproduction to great comparative effect. The comparative examination of adoption and child circulation has contributed to insights into the Eurocentric privileging of intergenerational biogenetic ties, whereas the circulation of children often reinforces material care relations between a larger group of adults as much as the reproduction of children (Leifsen, 2010; Leinaweaver, 2008; Modell, 1995). Pregnancy and childbirth are two other realms where comparison demonstrates how a process deemed natural and universal in Euro-American contexts is thoroughly shaped within the material semiotic and relational context where it is situated (Boddy, 1998; Ivry, 2010; Jordan, 1997; Morgan, 1989; Pinto, 2008; Van Hollen, 1998). The anthropological literature on abortion and personhood provincializes North American debates about when life begins as local and specific, since personhood in different non-Western sites can commence at various times, often post-birth, and the concept of individual life is not relevant to how children are made at all (Browner, 1976; Conklin and Morgan, 1996; Kaufman and Morgan, 2005; Picone, 1998).

New reproductive technologies have been an especially rich site for comparative work on reproduction. Starting in the early 1990s, Marilyn Strathern used her ethnographic delineation of relational personhood in New Guinea to analyze how the Euro-American use of assisted reproductive technologies (ARTs) is based on (1) the formulation of a domain called reproduction within the separation of the natural and the artificial (2) configured through persons as (3) unique individuals (Strathern, 2005). Strathern's work opened up a vast array of comparative questions for current ethnographers of reproduction (Bamford, 2007; Franklin, 1997; Roberts, 2012a). Ethnographers who study the use of ARTs outside of Europe and the United States, and those who have made comparisons within Europe, have contributed to an increasing sense of the specificity of reproduction within particular political, economic, and religious contexts (Bharadwaj, 2002, 2006; Inhorn, 2003; Kahn, 2000; Melhuus, 2003).

As a whole, anthropological comparison demonstrates that the biological processes involved in having children are not universally organized under the heading of reproduction.

This insight resonates with recent critical work in medical anthropology and science technology studies that diffuse and redistribute the self-evident nature of reproduction by examining how biological processes themselves are shaped through historic, economic, and political realities.

Reproduction Constructed

The cultural anthropological work of the late twentieth century, which sought to bring reproduction into the heart of anthropological social theory, delineated the political and economic processes that affect reproduction and the centrality of reproduction to social life. While much of this scholarship made an implicit distinction between biological and cultural processes, this work also made it possible to trouble the categories of nature and culture that shaped how reproduction was framed through the twentieth century (Franklin, 1995). This troubling has taken on new forms in the last decades through the work of scholars who examine the history, politics, and economics of biological processes themselves and by decentering the individual body as the central site of reproduction. These scholars, whom I loosely categorize as constructionists, in distinction to social or cultural constructionists, do not assume that diverse malleable cultural processes sit on top of fixed and universal biological processes, but instead insist on parsing the specific, material, semiotic relations between people and the worlds they inhabit without seeking to classify them a priori as nature or culture.

Margaret Lock's seminal book *Encounters with Aging* (Lock, 1993) was an early example of a constructivist approach. Through an intensely comparative study of menopause in Japan and Canada, Lock observed that while women in mid-twentieth- century Canada and women in mid-twentieth century-Japan both cease to menstruate in midlife, the endocrine changes associated with menopause in North America are by no means universal. At the time of her study, women in Japan did not experience hot flashes. Lock argued that women residing in these different nations possessed 'local biologies,' constructed within fields of thick historical and economic relations (for further uses of 'local biologies,' to explicate reproductive processes, see Behague, 2002; Roberts, 2012b).

In a related vein, Sarah Franklin draws our attention to how embryos produced in laboratories, along with heart valves and skin, are not only 'born and bred, or born and made, but made and born' through rigorous quality control that takes away impurities, making these lab-made entities as 'good as nature' (Franklin, 2006: p. 171). In moving away from the "dichotomizations of social and natural facts" (Franklin and Ragoné, 1998: p. 3), scholars have begun to link reproduction to informatics, biocapital, intellectual property, and genetic engineering (Helmreich, 1998; Waldby and Mitchell, 2006). These linkages between human reproduction and the contemporary means to engineer biological processes collapse distinctions between nature/culture and production/reproduction.

Collapsing nature/culture works in conjunction with other collapses like that of the surety that reproduction works in individual bodies. As the feminist philosopher of science Donna Haraway argues, reproductive technologies are not only expensive technomedical procedures that treat individual women's bodies, but complex economic and biological processes that make diaper pins, home birth, accounting techniques, population politics, breast milk, and baby formula (Haraway, 1997; Haraway and Goodeve, 2000). Taking her cue from Haraway, feminist historian of science Michelle Murphy argues that if we ask "Where does biological reproduction reside?" the answer "the body" is simply not up to the task (Murphy, 2011). Murphy's framework of 'distributed reproduction' allows us to examine "what counts as biological reproduction by tracking the dispersion of sexed living beings into their infrastructural and political economic milieu." The framework of individual bodies and universal biological processes is not up to the task, for instance, of understanding why among the Aamjiwnaang, First Nations peoples living near the chemical alley of Sarnia in eastern Canada, only 35 boys are born for every 100 girls. To grasp reproduction in this case, we need to trace the specific tangles of technoscience, nation-states, capital, the malleable and vulnerable endocrine processes of fish and people, the fluid dynamics of rivers, the techno-assemblages of chemical manufacture, population politics, and tax breaks. While twentieth-century anthropology excelled at describing the collective nature of cultural practices that coalesced around reproductive processes outside the West, much of the recent constructionist work on reproduction expands the members of the collectives everywhere beyond people to include policies, chemicals, waterways, and fish in all their relational ontologies. Through this lens, reproduction becomes a useful and constructed domain requiring a broad analytic lens to more fully account for the wide array of actors and processes involved in its production.

Conclusion

Within anthropology before the last decades of the twentieth century, "reproduction could be described as invisibly central (along Darwinian, patriarchal, and biologically determinist lines), while remaining visibly marginalized (for example, in terms of childbirth practices)" (Franklin and Ragoné, 1998: p. 3). The situation has changed. Foregrounding reproduction as a useful conceptual domain, instead of as a universal biological process residing in unique, bounded, individuals, has been a key anthropological innovation, allowing anthropologists to transform how we constitute our objects of study, what questions we ask, and how we collaborate across fields if productive/reproductive and nature/culture divides no longer configure the boundaries of our discipline.

See also: Birth, Anthropology of; Capitalism: The History of the Concept; Comparative Method in Anthropology; Cultural Critique, Anthropological; Culture: Contemporary Views; Economic Anthropology; Exchange in Anthropology; Family: The Anthropology of the Concept and Its History; Feminism in Anthropology; Gender: Anthropological Aspects; Health: Anthropological Aspects; Kinship in Anthropology; Modernity: Anthropological Aspects; Political Economy in Anthropology; Science and Technology, Anthropology of.

Bibliography

Adams, Vincanne, 2000. Particularizing modernity: Tibetan medical theorizing of women's health in Lhasa Tibet. In: Connor, L., Samuel, G. (Eds.), Healing Powers and Modernity: Traditional Medicine, Shamanism and Science in Asian Societies. Bergin and Garvey, London, pp. 222–246.

Adams, Vincanne, Pigg, Stacy Leigh, 2005. Sex in Development: Science, Sexuality, and Morality in Global Perspective. Duke University Press, Durham.

Anagnost, Ann, 1995. A surfeit of bodies: population and the rationality of the state in post-Mao China. In: Ginsburg, F., Rapp, R. (Eds.), Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley.

Bamford, Sandra C., 2007. Biology Unmoored: Melanesian Reflections on Life and Biotechnology. University of California Press, Berkeley.

Behague, Dominique, 2002. Beyond the simple economics of cesarean section birthing: women's resistance to social inequality. Culture, Medicine and Psychiatry 26 (4), 473–507.

Berry, Nicole, 2010. Unsafe Motherhood: Mayan Maternal Mortality and Subjectivity in Post-War Guatemala. Berghahn, New York, Oxford.

Bharadwaj, Aditya, 2002. Conception politics: medical egos, media spotlights, and the contest over test-tube firsts in India. In: Inhorn, M., Balen, F.V. (Eds.), Infertility Around the Globe. University of California Press, Berkeley, pp. 315–333.

Bharadwaj, Aditya, 2006. Sacred conceptions: religion and the global practice of IVF – Special Issue. Culture Medicine and Psychiatry 30 (4).

Boddy, Janice, 1998. Remembering Amal: on birth and the British. In: Lock, M.M., Kaufert, P.A. (Eds.), Pragmatic Women and Body Politics, Cambridge Studies in Medical Anthropology, vol. 5. Cambridge University Press, New York, pp. 28–57.

Borneman, John, 2001. Caring and being cared for: displacing marriage, kinship, gender and sexuality. In: Faubion, J.D. (Ed.), The Ethics of Kinship: Ethnographic Inquiries. Rowman & Littlefield, Lanham, MD, pp. i29–i46.

Browner, Carole, Dec 1976. Poor Women's Fertility Decisions: Illegal Abortion in Cali, Colombia (Ph.D. thesis in Anthropology). University of California, Berkeley.

Casper, Monica J., 1998. The Making of the Unborn Patient: a Social Anatomy of Fetal Surgery. Rutgers University Press, New Brunswick, NJ.

Colen, Shellee, 1995. 'Like a mother to them': stratified reproduction and West Indian child care workers and employers in New York. In: Ginsburg, F.D., Rapp, R. (Eds.), Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley, pp. 78–102.

Conklin, B.A., Morgan, L.M., 1996. Babies, bodies, and the production of personhood in North America and the native Amazonian society. Ethos 24 (4), 657–694.

Daston, L., 1998. The nature of nature in early modern Europe. Configurations 6 (2), 149–172.

Degler, Carl N., 1991. In Search of Human Nature: the Decline and Revival of Darwinism in American Social Thought. Oxford University Press, New York.

Douglass, Carrie B., 2005. Barren States: the Population "Implosion" in Europe. Berg, Oxford, New York.

Dudgeon, Matthew, Inhorn, Marcia, 2004. Men's influences on women's reproductive health: medical anthropological perspectives. Social Science & Medicine 59 (7), 1379–1395.

Erikson, Susan, 2001. Maternity care policies and maternity care practices: a tale of two Germanys. In: DeVries, R., Benoit, C., Van, E., Teijlingen, S.W. (Eds.), Birth by Design: Pregnancy, Maternity Care, and Midwifery in North America and Europe. Routledge, London, New York, pp. 203–217.

Fassin, Didier, 2012. That obscure object of global health. In: Inhorn, M.C., Wentzell, E.A. (Eds.), Medical Anthropology at the Intersections: Histories, Activisms, and Futures. Duke University Press, Durham, pp. 95–115.

Foster, George McClelland, Anderson, Barbara Gallatin, 1978. Medical Anthropology. Wiley, New York.

Foucault, Michel, 1990. The History of Sexuality. Vintage Books, New York.

Franklin, Sarah, 1995. Science as culture, cultures of science. Annual Review of Anthropology 24, 163–184.

Franklin, Sarah, 1997. Embodied Progress: a Cultural Account of Assisted Conception. Routledge, London; New York.

Franklin, Sarah, 2006. The cyborg embryo: our path to transbiology. Theory, Culture and Society 23 (7–8), 167–187.

Franklin, Sarah, McKinnon, Susan, 2001. Relative Values: Reconfiguring Kinship Studies. Duke University Press, Durham, NC.

Franklin, Sarah, Ragoné, Helene, 1998. Reproducing Reproduction: Kinship, Power, and Technological Innovation. University of Pennsylvania Press, Philadelphia.

Fraser, Gertrude J., 1995. Modern bodies, modern minds: midwifery and reproductive change in an African American community. In: Ginsburg, F.D., Rapp, R. (Eds.), Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley, pp. 42–58.

Ginsburg, Faye D., 1998. Contested Lives: the Abortion Debate in an American Community. University of California Press, Berkeley.

Ginsburg, Faye D., Rapp, Rayna, 1995. Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley.

Goldade, Kate, 2007. Reproducción transnacional: la salud reproductiva, las limitaciones y las contradicciones para las migrantes laborales nicaragüenses en Costa Rica. In: García, C.S. (Ed.), El Mito Roto: Inmigración y Emigración en Costa Rica. UCR Press, San José, pp. 233–260.

Greenhalgh, Susan, 2008. Just One Child: Science and Policy in Deng's China. University of California Press, Berkeley.

Gutmann, Matthew C., 2007. Fixing Men: Sex, Birth Control, and AIDS in Mexico. University of California Press, Berkeley.

Haraway, Donna Jeanne, 1997. Modest Witness@Second Millennium. Female Man Meets Oncomouse: Feminism and Technoscience. Routledge, New York.

Haraway, Donna Jeanne, Goodeve, Thyrza Nichols, 2000. How Like a Leaf: an Interview with Thyrza Nichols Goodeve. Routledge, New York.

Helmreich, Stefan, 1998. Silicon Second Nature: Culturing Artificial Life in a Digital World. University of California Press, Berkeley.

Horn, David, 1994. Social Bodies: Science, Reproduction, and Italian Modernity. Princeton University Press, Princeton, NJ.

Inhorn, Marcia, 2006. Defining Women's Health. Medical Anthropology Quarterly 20 (3), 345–378.

Inhorn, Marcia Claire, 2003. Local Babies, Global Science: Gender, Religion, and In vitro Fertilization in Egypt. Routledge, New York.

Ivry, Tsipy, 2010. Embodying Culture: Pregnancy in Japan and Israel. Rutgers University Press, New Brunswick, NJ.

Janes, Craig R., Chuluundorj, Oyuntsetseg, 2004. Free markets and dead mothers: the social ecology of maternal mortality in post-socialist Mongolia. Medical Anthropology Quarterly 18 (2), 230–257.

Jordan, Brigitte, 1997. Authoritative knowledge and its construction. In: Sargent, C.F., Davis-Floyd, R. (Eds.), Childbirth and Authoritative Knowledge: Cross-Cultural Perspectives. University of California Press, Berkeley p. xii, 510.

Jordanova, Ludmilla, 1995. Interrogating the concept of reproduction in the eighteenth century. In: Ginsburg, F.D., Rapp, R. (Eds.), Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley, pp. 369–386.

Kahn, Susan, 2000. Reproducing Jews: a Cultural Account of Assisted Conception in Israel. Duke University Press, Durham.

Kant, Immanuel, 1949. Foundation of the Fundamental Principles of the Metaphysics of Morals (T.K. Abbott, Trans.). Library of Liberal Arts, Indianapolis.

Kaufman, Sharon, Morgan, Lynn, 2005. The anthropology of the beginnings and ends of life. Annual Review of Anthropology 34, 317–341.

Landecker, Hannah, 2003. On beginning and ending with apoptosis: cell death and biomedicine. In: Franklin, S., Lock, M.M. (Eds.), Remaking Life & Death: Toward an Anthropology of the Biosciences. School of American Research Press, Santa Fe.

Lasch, Christopher, 1977. Haven in a Heartless World: the Family Besieged. Basic Books, New York.

Latour, Bruno, 1993. We Have Never Been Modern. Harvester Wheatsheaf, New York.

Layne, Linda L., 1999. Transformative Motherhood: on Giving and Getting in a Consumer Culture. New York University Press, New York.

Leifsen, Esben, 2010. Child welfare, biopower and mestizo relatedness in Quito, Ecuador. In: Thelen, T., Haukanes, H. (Eds.), Parenting after the Century of the Child: Traveling Ideals, Institutional Negotiations and Individual Responses. Ashgate, Surrey, England; Burlington, VT, pp. 103–121.

Leinaweaver, Jessaca B., 2008. The Circulation of Children: Kinship, Adoption, and Morality in Andean Peru. Duke University Press, Durham.

Lévi-Strauss, Claude, 1969. The Elementary Structures of Kinship (Les Structures Élémentaires de la Parenté). Eyre & Spottiswoode, London.

Lewin, Ellen, 1993. Lesbian Mothers: Accounts of Gender in American Culture. Cornell University Press, Ithaca, NY.

Lock, Margaret M., 1993. Encounters with Aging: Mythologies of Menopause in Japan and North America. University of California Press, Berkeley.

Malinowski, Bronislaw, 1932. The Sexual Life of Savages in North-Western Melanesia: an Ethnographic Account of Courtship, Marriage, and Family Life among the Natives of the Trobriand Islands. G. Routledge & Sons, ltd, London.

Mamo, Laura, Fishman, Jennifer, 2001. Potency in all the right places: Viagra as a technology of the gendered body. Body and Society 7 (4), 13–35.

Martin, Emily, 1992. The Woman in the Body: a Cultural Analysis of Reproduction. Beacon Press, Boston.

Melhuus, Marit, 2003. Exchange matters: issues of law and the flow of human substances. In: Eriksen, T.H. (Ed.), "Globalisation" Studies in Anthropology. Pluto Press, London; Sterling, VA, pp. 39–57.

Modell, Judith, 1995. "Nowadays everyone is Hanai": child exchange in the construction of Hawaiian urban culture. Société des Océanistes 100–101, 201–219.

Morgan, Lynn, 1989. When does life begin? A cross-cultural perspective on the personhood of fetuses and young children. In: Doerr, E., Prescott, J.W. (Eds.), Abortion Rights and Fetal Personhood. Centerline Press, Long Beach, California, pp. 89–107.

Morgan, Lynn, 2009. Icons of Life: a Cultural History of Human Embryos. University of California Press, Berkeley.

Morgan, Lynn, Roberts, Elizabeth F.S., 2012. Reproductive governance in Latin America. Anthropology and Medicine 19 (2), 241–254.

Murphy, Michelle, 2011. Distributed reproduction. In: Casper, M.J., Currah, P. (Eds.), Corpus: an Interdisciplinary Reader on Bodies and Knowledge, first ed. Palgrave Macmillan, New York, pp. 21–39.

Nouzeilles, Gabriela, 2003. Hysteria in turn-of-the-century Buenos Aires. In: Armus, D. (Ed.), Disease in the History of Modern Latin America: from Malaria to AIDS. Duke University Press, Durham, pp. 51–75.

Pashigian, Melissa, 2009. Inappropriate relations: the ban on surrogacy with in vitro fertilization and the limits of state renovation in contemporary Vietnam. In: Birenbaum-Carmeli, D., Inhorn, M.C. (Eds.), Assisting Reproduction, Testing Genes: Global Encounters with New Biotechnologies. Berghahn Books, New York, pp. 164–188.

Picone, Mary, 1998. Infanticide, the spirits of aborted fetuses, and the making of motherhood in Japan. In: Scheper-Hughes, N., Sargent, C.F. (Eds.), Small Wars: the Cultural Politics of Childhood. University of California Press, Berkeley, pp. 37–57.

Pinto, Sarah, 2008. Where There is No Midwife: Birth and Loss in Rural India. Berghahn Books, New York.

Rabinow, Paul, Rose, Nikolas, 2006. Bio-Power Today. Biosocieties 1, 195–217.

Rapp, Rayna, 1999. Testing Women, Testing the Fetus: the Social Impact of Amniocentesis in America. Routledge, New York.

Roberts, Elizabeth F.S., 2012a. God's Laboratory: Assisted Reproduction in the Andes. University of California, Berkeley.

Roberts, Elizabeth F.S., 2012b. Scars of nation: surgical penetration and the Ecuadorian state. Journal of Latin American and Caribbean Anthropology 17 (2), 215–239.

Rubin, Gayle, 1975. The traffic in women: notes on the "political economy" of sex. In: Reiter, R.R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York, pp. 157–210.

Scheper-Hughes, Nancy, 1992. Death without Weeping: the Violence of Everyday Life in Brazil. University of California Press, Berkeley.

Schneider, David, 1980. American Kinship: a Cultural Account. University of Chicago Press, Chicago.

Singer, Merrill, Baer, Hans, 1995. Critical Medical Anthropology. Baywood Publishing Co., Amityville, NY.

Stepan, Nancy, 1991. The Hour of Eugenics: Race, Gender, and Nation in Latin America. Cornell University Press, Ithaca.

Strathern, Marilyn, 1992a. After Nature: English Kinship in the Late Twentieth Century. Cambridge University Press, Cambridge, New York.

Strathern, Marilyn, 1992b. Reproducing the Future: Essays on Anthropology, Kinship, and the New Reproductive Technologies. Routledge, New York.

Strathern, Marilyn, 2005. Kinship, Law and the Unexpected: Relatives are Always a Surprise. Cambridge University Press, New York, NY.

Tarlo, Emma, 2003. Unsettling Memories: Narratives of the Emergency in Delhi. University of California Press, Berkeley.

Taylor, Janelle, 2004. Big ideas: feminist ethnographies of reproduction. American Ethnologist 31 (1), 123–130.

Taylor, Janelle S., Layne, Linda L., Wozniak, Danielle F., 2004. Consuming Motherhood. Rutgers University Press, New Brunswick, NJ.

Thompson, Charis, 2005. Making Parents: the Ontological Choreography of Reproductive Technologies. MIT Press, Cambridge.

Unnithan-Kumar, Maya, 2004. Reproductive Agency, Medicine and the State: Cultural Transformations in Childbearing. Berghahn Books, New York.

Van Hollen, Cecilia Coale, 1998. Birth on the Threshold: Childbirth and Modernity among Lower Class Women in Tamil Nadu. University of California Press, Berkeley.

Waldby, Cathy, Mitchell, Robert, 2006. Tissue Economies: Blood, Organs, and Cell Lines in Late Capitalism. Duke University Press, Durham.

Weiner, Annette, 1995. Reassessing reproduction in social theory. In: Ginsburg, F., Rapp, R. (Eds.), Conceiving the New World Order: the Global Politics of Reproduction. University of California Press, Berkeley.

Weismantel, Mary, 2004. Moche sex pots: reproduction and temporality in ancient South America. American Anthropologist 106 (3), 495–505.

Whittaker, Andrea, 2010. "Cycling overseas": care, commodification, and stratification in cross-border reproductive travel. Medical Anthropology 29 (4), 363–383.

Zelizer, Viviana A. Rotman, 1994. Pricing the Priceless Child: the Changing Social Value of Children. Princeton University Press, Princeton, NJ.

Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective

Sonia Corrêa, The Brazilian Interdisciplinary from HIV/AIDS, Sexuality Policy Watch, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article argues that a not so sharp distinction between affluent and developing nations in terms of understanding of and challenges for the realization of reproductive rights and examines intersections and disjunctions in relation to sexual rights. It charts the many challenges in contemporary politics of human rights in the domains of reproduction and sexuality in terms of enabling conditions required for the realization of these rights, the growth of religious dogmatic reactions to these claims, the scalar developments in biomedicine and biotechnology.

Introduction

In 1994, the Program of Action of the International Conference of Population and Development (ICPD Cairo, 1994) defined reproductive rights as:

> [Embracing] certain human rights that are already recognized in national laws, international human rights documents and other consensus documents. These rights rest on the recognition of the basic right of all couples and individuals to decide freely and responsibly the number, spacing and timing of their children and to have the information and means to do so, and the right to attain the highest standard of sexual and reproductive health. It also includes their right to make decisions concerning reproduction free of discrimination, coercion and violence.
>
> Paragraph 7.3, United Nations, 1994

The ICPD Program of Action also defined abortion as a major public health problem and strongly recommended that when legal it should be safe and accessible (Paragraph 8.25). A year later, the Platform of Action of the IV World Women Conference (IV WOW Beijing) reaffirmed these ICPD principles, adding the recommendation that countries should revise existing punitive legislation on abortion. The Beijing Platform of Action also spelled out the substantive contents of women's sexual rights:

> The human rights of women include their right to have control over and decide freely and responsibly on matters related to their sexuality, including sexual and reproductive health, free of coercion, discrimination and violence. Equal relationships between women and men in matters of sexual relations and reproduction, including full respect for the integrity of the person, require mutual respect, consent and shared responsibility for sexual behavior and its consequences.
>
> Paragraph 96, United Nations, 1995

In 20 years since these definitions were adopted the language referring to reproductive rights and abortion has not been substantially altered. On the other hand the interpretation, use, and application of sexual rights have decidedly expanded beyond women's human rights and the heteronormative imprint of the second sentence of Paragraph 96 has been contested and deconstructed. Although no new text has been agreed upon in intergovernmental negotiations, the work of human rights bodies and authoritative experts gradually construed a revised definition of sexual rights that reads as follows:

> Sexual rights include the right of all persons to express their sexual orientation, with due regard for the well-being and rights of others, without fear of persecution, denial of liberty or social interference.
>
> Paul Hunt, Special Rapporteur on the Right to the Highest Standard of Health – United Nations E/CN.4/2004/49, 16 February 2004, paragraph 54

This article focuses fundamentally on reproductive rights and does not examine exhaustively the contents and challenges more specific to sexual rights in a broader sense. But it charts aspects of the trajectory of sexual rights, particularly in relation to sexual orientation and gender identity as they relate to parallel developments in reproductive rights, and it also briefly explores the intersections and distinctions between the two concepts. However, important aspects of sexual rights such as those related to sexual violence in its various forms, genital mutilations or sex work are not examined in depth.

Historical Trajectories: Longer Cycles and Close Antecedents

In the sober rooms of the International Conference of Population and Development in Cairo in 1994, many state actors were taken by surprise when rights claims in relation to 'sex' erupted. However, intertwined linkages between sexuality, gender, and reproduction have always been central to societal regulation across time and cultures. While reproductive and sexual rights belong to the liberal legacy of human and citizenship rights, they also call upon the foundational premises of freedom, equality, privacy, nondiscrimination, and dignity to contest the persistent patriarchal, androcentric, heteronormative bias of the modern social contract (Corrêa et al., 2008).

The meanings of sexual and reproductive rights claims must be examined in relation to the other forms of social

control and discipline that underlie the modern social contract, the apparatuses of societal regulation based on demography, biomedicine, criminology, and economics that also took form in the transition to Western modernity. Reproductive and sexual rights are intrinsically linked to biopower and biopolitics, as defined by Michel Foucault (1978). Early concerns with population size and dynamics, in particular the eighteenth-century Malthusian thesis on high fertility among the poor, would be persistently reactivated in the course of twentieth-century debates, including the Cairo Conference; its imprints have not vanished from contemporary policy concerns (Corrêa et al., 2013).

Biopolitics also implied concerns with the 'quality' of populations in particular in relation to 'race,' a notion that was made scientific in the nineteenth century and thereafter occupied a nodal place in the ideologies of nation building, homogeneity, and colonialism in Europe and the United States. Techniques of social and biological 'improvement' construed sex, gender, and reproduction as key pivots of regulation.

These modern juridical, legal, and infrajuridical apparatuses of regulation and discipline worldwide would be transplanted, at varying paces and degrees of intensity, to colonial territories. Colonial laws criminalizing abortion and sodomy that remain in the books in a large number of Southern postcolonial states are glaring examples of this transplantation (Sanders, 2009; ILGA, 2012). The profusion of hygiene and public health norms and interventions promoted by colonial administrations from the mid-nineteenth century has formed the basis for public health discourses and institutions incorporated by the early independent states in Latin America and later on by those emerging from the post-World War II wave of decolonization.

In the early nineteenth century the writings and experiments of Utopian socialists first combined economic and sexual justice. Later Engels (1884) analyzed the articulation between sex and economics in *The Origin of the Family, Private Property, and the State*, opening a pathway followed by socialist feminist pioneers like Emma Goldman and reaching as far as the 1960s feminist wave.

Concurrently, under the impact of Darwin's (1871, 1859) theories on sex reproduction and evolution of species, the 'science of sex' was invented and with it the concepts of sexual instinct and homosexuality. While most new 'sexology' experts and later psychoanalysts contested religious doctrines on reproduction and sexual behavior and argued against the criminalization of same sex relations, the new science of sex also crafted the powerful concept of a universal human sexual instinct that glossed over the wide variability of cultural meanings in respect to sexuality, gender, and reproduction, placed reproductive behavior at the top of a taxonomic hierarchy of sexual practices, and overlapped with racial and class views associated with eugenics (Rubin, 1984).

Paradoxically, these political and scientific ideas on 'sex,' gender, and reproduction both inspired many emancipatory projects, such as those propelled by feminist pioneers' and progressives' fight for birth control and sexual reforms, and fed a vast arsenal of ideas of social control, discipline, hierarchizing, and marginalization. And they continue to influence political debates, policy propositions, and controversies around the regulation of fertility. The paradox is illustrated in the biography of Margaret Sanger, the early twentieth-century North American radical advocate of contraception and abortion. From the late 1920s, Sanger established connections with neo-Malthusian and eugenic groups and created the International Birth Control League, later renamed the International Planned Parenthood Federation (IPPF). This institution worked with colonial administrations and propelled the emergence of local birth control associations in the Southern hemisphere, which would play a major role in the fertility control initiatives of the 1960s (Chesler, 1992).

The biopolitical tragedy of Nazi-Fascism imposed silence on the most radical and salient pre-World War II eugenic proposals. But ideas and policy proposals around birth control did not die away and by the 1950s linked up with a new wave of population concerns now crafted in planetary terms and articulated with economic development theory: excessive or too rapid population growth would impair state economies. By then Sanger was a sponsor of research leading to the invention of the pill, which would become the icon of women's sexual and reproductive autonomy – even of sexual liberation more broadly speaking. Yet the pill would also become a key element in the post-World War II transnational machinery of population control.

In the 1960s, geopolitical tensions of the Cold War were at their height when a youth Cultural Revolution flared across industrialized societies, provoking as well eruptions in the socialist world and Latin America. The 1960s sexual rebellion inaugurated the politics of gender, sexuality, and reproduction, as we know it today. It preceded the 1970s and 1980s legal reforms in the realms of gender equality, abortion, and contraception and also the decriminalization of sex relations (mostly but not exclusively in affluent societies). It also inspired the contemporary wave of sexuality theorizing that challenged the normalizing and colonizing effects of positivist and naturalist tenets of nineteenth-century sex science (Correa et al., 2013).

But, in 1968, a book titled *The Population Bomb* (Ehrlich and Ehrlich, 1968) was also published. Its Cold War rhetoric about high fertility rates in the regions south of the equator fueled concern about mass starvation and societal upheavals. Although not all actors engaged in population research and related debates called for coercive measures of fertility control, the debate took flagrant colonial, racist and, quite often, eugenic turns (Hartmann, 1995).

At the first UN Conference on Population, in Bucharest in 1974, Third World countries reacted to population control rationales by crafting the proposition: "Development is the best contraceptive." However, in many cases this rhetoric simply concealed deeply rooted pronatalist stances. Yet not all developing nations remained faithful to their Bucharest commitment. By the late 1970s, draconian population control was already in place in Asia, particularly in China, and India and Mexico had also adopted a national policy based on fertility targets.

From the late 1970s, these politically intricate paradoxes began to be critically deconstructed through debates around gender equality, reproductive autonomy, and sexual freedom transnationally and in a series of United Nations processes and conferences. As early as the late 1960s feminists

had begun to criticize international population policies, fertility targets, and coercive contraceptive provisions. Feminist groups that began blossoming South of the Equator most often supported the 1974 Bucharest 'Third World development agenda' but they also criticized population control's negative effects on women's health, lives, and autonomy as well as the pronatalist views of their governments and of religious institutions. In the case of Latin America feminist critiques targeted in particular the Catholic Church that viewed women as reproducers and called for quality contraception and legal abortion. They also addressed the racialized biases of fertility control globally and within countries – as exemplified by South African apartheid population policies – and raised sharp critiques in respect to the bad quality of care provided by both mother and child programs, which dominated states' responses to women's health needs all across the global South (Corrêa and Reichmann, 1994).

Later in 1984 in Amsterdam, the first International Reproductive Rights Conference sponsored by the International Campaign on Abortion, Sterilization and Contraception gathered feminists from all over the world. On this occasion the Women's Global Network on Reproductive Rights (WGNRR) was created and a transnational consensus was reached around the concept of reproductive rights defined as "women's right to decide whether, when and how to have children – regardless of nationality, class, race, age, religion, disability, sexuality or marital status – in the social, economic and political conditions that make such decisions possible." One key feature of this consensus entailed the broadening of reproductive autonomy tenets beyond abortion, contraception, and sterilization, to include maternal mortality, prenatal and obstetric care, breast and cervical cancer, and adolescent and postmenopausal health. Southern feminists also strongly emphasized that women's labor rights and social reproduction roles should be part of the reproductive rights agenda (Dixon-Mueller, 1993).

The subsequent WGNRR-sponsored Women's Health Conference held in San Jose, Costa Rica (1987), began an international campaign to fight maternal mortality and inspired the creation of the Latin American Campaign for the Decriminalization of Abortion. In Manila, in 1990, at the next Women's Health Conference, the proposal was made that the international reproductive rights movement should dare to conceptualize a feminist population policy.

These debates and processes must also be situated in relation to the HIV/AIDS epidemics. Recognized in the early 1980s in the West among primarily gay men and later as a generalized global epidemic, the epidemic brought with it a human tragedy of gigantic proportions and propelled establishment of new global policy frames that would have impact on national sexual and reproductive laws and policies. AIDS has also, and paradoxically, opened key inroads for public and political discourses on sexuality-related discrimination and stigma to blossom worldwide (Altman, 2013; Parker, 2001).

The United Nations Policy Arenas

The first UN International Conference on Human Rights held in Tehran in 1968 approved a principle ensuring couples the "basic human right to decide freely and responsibly on the number and spacing of their children and a right to adequate education and information in this respect." In 1974 in Bucharest, despite the geopolitical tensions and skirmishes, this definition was rephrased to extend these rights to individuals (not just couples) and to ensure access to the means of exercising them.

Concurrently gender equality concerns in relation to dominant family planning frameworks were evolving. In 1975, in the First United Nations sponsored International Women's Conference held in Mexico City, coercive measures of population control were openly criticized and the final document claimed respect for the bodily integrity of women as 'a fundamental element of human dignity and freedom.' In 1979, the Convention on the Elimination of All Forms of Violence against Women (CEDAW) affirmed women's rights to family planning, equal right (with men) to decide on the number and spacing of children, the right to maintain their jobs while pregnant, and to have access to maternity leave and childcare. In 1981, an UN Symposium on Population and Human Rights held in Vienna also declared that "Both the compulsory use of abortion and its unqualified prohibition would be a serious violation of human rights."

Despite this emerging consensus, the global policy climate in relation to reproductive autonomy was far from favorable. At the International Population Conference of Mexico in 1984, while developing countries expressed a broader acceptance of women's right of access to contraception, an alliance between the United States' Reagan Administration and the Vatican openly targeted family planning programs, particularly abortion; the United States announced that it would suspend funding for these programs, including to the United Nation Population Fund (UNFPA), a policy that would remain in place until 1993. On the other hand, by the end of the 1980s, the World Health Organization framed and adopted the notions of reproductive and sexual health that would later be incorporated by other UN institutions as well as by feminist and HIV/AIDS activists (Dixon-Mueller, 1993).

Then in the 1990s, in the Cairo and Beijing Conferences, UN member states agreed upon the definitions of reproductive rights and of sexual rights of women. These groundbreaking agreements were the result of debates that took place in a series of conferences sponsored by the United Nations in the 1990s: Environment and Development (Rio 1992), Human Rights (Vienna 1993), Population and Development (Cairo 1994), Social Development (Copenhagen 1995), and finally Women (Beijing, 1995). For example, the 1993 Vienna Conference legitimized the premises of women's rights in public and private spheres – regardless of cultural relativity – and established the indivisibility of civil, political, economic, and social rights, both definitions being crucial for the further the development of reproductive and sexual rights.

Whereas in the 1970 and 1980s, both Northern and Southern feminist intellectual and political endeavors had evolved largely disconnected from international policy arenas, in the 1990s, feminists were directly engaged in these UN negotiations and the presence and impact of global South feminist perspectives were palpable. The Cairo Conference and its reaffirmation and expansion in Beijing were correctly interpreted as a major paradigm shift from long-standing instrumental population concerns toward a perspective based

on gender equality, human rights, and health. On the other hand these negotiations were contentious. Although many countries from the economic South – mostly from Latin America, but also South Africa, India, and even Egypt – moved toward supporting gender equality, reproductive rights, and even sexual rights, the conservative resistance was fierce and systematic. Initially clustered around the Vatican and its allied Latin American and African countries, this hub of resistance expanded to include Islamic states and also some Asian and African countries.

Although this systematic opposition – named by feminists as the Unholy Alliance – was not able to block the political consensus in the agreed documents, it was the origin of a consistent platform of resistance to the application of human rights to sexuality and reproduction that continues to operate in all UN global policy arenas. The tensions at play in these early 1990s UN disputes were hotter than similar controversies then underway at national levels, with the exception of the United States. These difficult UN negotiations anticipated the harsh political climate that, in the 2000s, also prevails at national levels, across continents, in relation to reproductive and sexual rights claims (Sen and Corrêa, 1999).

The 2000s Policy and Political Landscape

The first decade and a half of the twenty-first century has seen the expansion and greater visibility of rights claims in the domain of reproductive and sexual rights, and a diversification of voices, subjects, and demands (Ilkkaracan and Jolly, 2007). Political mobilizing for abortion rights mushroomed in almost all countries, while the women's reproductive and sexual agenda continued to address issues of contraception and expanded to encompass HIV/AIDS prevention and treatment, radically transforming the global map of reproductive politics (Harcourt, 2009). Arguably the 2000s have also been the era of sexual rights broadly speaking, as claims of equality, nondiscrimination and protection from violence of persons whose sexual orientation and gender identity differ from the dominant norms have blossomed nationally and globally, as incalculable choreographies of sexual and gender diversity and malleability have erupted, so to speak, in terms of political and media visibility (Richardson, 2000; Thoresen, 2009). Concurrently the political demands of sex workers have also become more palpable and transnationally connected (Kempadoo and Doezema, 1998).

These dynamics are decidedly linked to enormous changes in communication technologies allowing for ideas and practices to cross state boundaries, to the deep and wide impacts of market forces in terms of recreating sexual cultures and reproductive practices, and to intensified migration (Altman, 2002; Parker, 1999; Tan, 2014). However, mobilization and advocacy at both local and global levels around reproductive and sexual rights since the 1990s and the cultural and normative changes resulting from these efforts are also key factors. Since 1995 UN language and norms spread across many policy levels and propelled the imagination of those whose rights are violated in the most diverse settings. Even though these definitions were not legally binding, they have become ethical and normative parameters on the basis of which existing laws and norms even within private institutions, such as biomedical procedures, could be contested and their detrimental effects rejected.

For example, in 1994, the UN Committee on Human Rights in the case of *Toonen v. Australia* concluded that Toonen's right to privacy was violated by the sodomy law remaining in the books of the state of Tasmania, and called Australia to review the legislation. In 2003, Brazil tabled the first resolution presented to the UN explicitly addressing sexual orientation and human rights at the UN Human Rights Commission. Although the Brazilian initiative faced great resistance from Islamic states and was not successful, it propelled a sustained global political mobilization around the rights of sexually nonconforming persons in the context of UN policy arenas, in particular the UN Human Rights Council (HRC) established in 2005 (Girard, 2007).

Then, in 2006, an independent international group of experts and activists elaborated the Yogyakarta Principles for the Application of Human Rights in relation to Sexual Orientation and Gender Identity. This document anchored the human rights claims related to sexual orientation and gender identity in consecrated premises of international law, and became an intellectual inspiration, both globally and nationally, for activists as well as jurists, legislators, and policy makers (O′ Flaherty and Fisher, 2008). Since 2008 a series of declarations and one resolution on sexual orientation, gender identity, and human rights have been presented and adopted by the UN HRC and there are strong indications that in the next few years substantial resolutions on human rights applied to sexual orientation and gender identity will be adopted and permanent mechanisms for monitoring these rights will be established. In parallel, since 1994 numerous sexual orientation and gender identity cases have been examined with positive results by UN human rights bodies as well as by the European and Inter-American judicial systems (Saiz, 2004; International Commission of Jurists, 2010).

These international debates were surrounded by national debates and law reforms in relation to LGBTQI rights in a wide range of countries. While legal reforms regarding same sex civil union and marriage equality tend to capture greater media and public attention, groundbreaking antidiscrimination legislation and specific laws concerning gender identity have also materialized in recent years in diverse settings (Currah et al., 2006). Not less importantly, political struggles, legal suits, and legislative debates against the criminalization of same sex relations have also multiplied in the most diverse settings.

Legal gains in this domain achieved since the 1990s have not been confined to so-called affluent nations. New constitutions in Ecuador, South Africa, and Fiji in the 1990s included sexual orientation in their prohibition of discrimination; in 2008 Nepal's Supreme Court declared full fundamental human rights for 'sexual and gender minorities.' Criminal laws prohibiting sodomy were declared unconstitutional not only in the United States (2003) but also in Nicaragua (2008), Panama (2008), and the groundbreaking Delhi High Court decision on article 377 of the Indian Penal Code in July 2009, which would be,

however, struck down by the Indian Supreme Court on December 2013 (Boyce and Khanna, 2011; Tamale, 2009). In 2012, a new Argentinean law on gender identity broke ground by not requiring a psychiatric diagnosis before obtaining a new social identity and body modification procedures (Cabral, 2012).

Reforms in relation to same sex unions or marriage either through legislation or court decisions have also not been limited to the global North. The list includes various U.S. states, European countries, such as France, Spain, and Portugal, and New Zealand. But reforms have also been approved in Argentina, Brazil, Colombia, Mexico, South Africa, and Uruguay, and there are indications that they may soon take place in Vietnam and Thailand. The extension of marriage equality to same sex couples is a relevant development regarding sexual rights, but it also has important reproductive rights implications.

Turning specifically to the politics of reproduction, beginning in the early 2000s, reproductive rights cases began to be more systematically examined in international and regional human rights systems. In 2003, the UN Human Rights Committee in the *KL v. Peru* case condemned Peru for not providing abortion when the young woman was carrying a baby with grave fetal abnormality. In 2005, in the *Paulina v. Mexico* case scrutinized by the Inter-American Human Rights Commission, Mexico was condemned for not providing access to abortion when the woman had been raped and the state law permitted abortion in this case. In 2011, the CEDAW committee examining a maternal mortality case – the *Alyne Pimentel Case* – condemned Brazil for having infringed the fundamental rights of Alyne as defined by the Convention (Garita et al., forthcoming).

A major breakthrough was the adoption in November 2005 of the African Protocol on Women's Rights known as the Maputo Protocol. According to the Center for Reproductive Rights it is the first legally binding human rights instrument to fully articulate women's reproductive rights as human rights, and to expressly guarantee a woman's right to control her fertility, affirming women's right to reproductive choice and autonomy. The Protocol is critical not only for Africa but it also sets a parameter for struggles and claims elsewhere. Then in 2008 a resolution adopted by the Human Rights Council for the first time defined maternal mortality as a human rights violation. Although the text does not explicitly mention abortion, it called for the Office of the High Commissioner of Human Rights to study these violations. The resulting report made public in 2012 identifies illegal and unsafe abortion as one main cause of maternal mortality as a human rights violation (Center for Reproductive Rights, 2006).

Since 1995, 16 countries have reduced legal barriers to abortion, most principally on the basis of health reasons, but in some cases law reforms have also included rape, economic necessity, and fetal malformation as justified reasons for the termination of pregnancy. Most of these countries are in Africa: Benin (2003), Burkina Faso (2003), Chad (2002), Ethiopia (2004), Guinea (2000), Mali (2000), Saint Lucia (2004), Swaziland (2005), and Togo (2007). The list also includes Saint Lucia (2004), Bhutan (2003), and the 2006 groundbreaking decisions of the Colombian Constitutional Court, considered by some analysts as a full fledge abortion reform. Most significantly, in nine countries recent reforms occurred that made abortion legally available without restriction within the first 12–14 weeks of pregnancy: Albania (1996), Ethiopia (2005), Cambodia (1997), Nepal (2002), Portugal (2007), Spain (2010), South Africa (1996), Switzerland (2002), and Uruguay (2012); and state-level reforms have also taken place in Australia (2007) and in the Federal District of Mexico (April 2007). As the list shows, the large majority of progressive abortion reforms have taken place in the global South where legislation was previously more restrictive (Center for Reproductive Rights, 2013).

What were the conditions that enabled these recent developments? Since the 1990s, the poorly understood and poorly regulated forces of globalization had been sweeping the world. These forces deeply altered cultures by favoring the transformation of sexual and gender orders, women's labor patterns, and household structures in ways that potentially favored sexual and reproductive autonomy and greater moral pluralism in societies (Petchesky and Judd, 1998). They also propelled across continents the absorption and use of the lingua franca of human rights, including in relation to matters related to sexuality and reproduction (Cheah, 1997; Pecheny and de la Dehesa, 2014). On the other hand, global transformations propel novel and more complex patterns of social and economic inequalities, within and across countries, and immense financial instability across boundaries and geoeconomic shifts. For example, the 2008 financial crisis in the United States and Europe had devastating effects on employment, inequality, and poverty and led to the repositioning of former peripheral big economies – Brazil, Russia, India, China, and South Africa as the new emerging global players. The realization of reproductive and sexual rights is necessarily affected by worsening inequalities and their future developments will have to be mapped and understood in relation to these global shifts.

Furthermore globalization, as it is widely known, went hand in hand with neoliberal policy frameworks based on market-oriented policy models, including the continuing privatization of health systems, either through public–private partnerships, or in more extreme forms. The existence of functional, accessible, and affordable or free health services is a key enabling condition for the exercise of reproductive rights. Decriminalizing abortion is not sufficient; safe abortions require the support of health services. Similarly, access to quality care health services constitutes an important element of the enabling conditions for sexual rights, including HIV/AIDS prevention and care, and the medical needs of trans- and intersex persons.

Yet, since the 1990s, partial or total privatization of health care has increased worldwide, including in Europe where welfare states are being dismantled by stringent austerity fiscal measures; or even in Brazil where the system established in the 1980s was premised on principles of universal coverage and gratuity. The 2011 U.S. health care reform may appear exceptional, but it was designed to ease the sharp detrimental effects of extreme privatization that prevailed in the American system since the 1980s. In any case, if the ICPD policy agenda was premised, to a large extent, on the existence of publicly funded, universal and free of charge health systems, this aspiration has not materialized South or North of the Equator.

In 2010 a group of feminist researchers and activists organized a meeting in Lankawi, Malaysia, to critically charter key

factors negatively impacting on the implementation of the reproductive health and rights agenda that had emerged from the 1990s UN cycle of conferences. They concluded that "Inequities in access to health care due to the increasing privatization and commodification of health services and health insurance, the introduction of internal markets into health care delivery, and decades-old policies of charging user fees in the public health system on top of out-of-pocket costs, have undermined the myriad commitments made by governments and other institutions of power to achieving health for all, including sexual and reproductive health." (Berer, 2011, page 3).

Furthermore, despite positive developments since the 1990s in terms of international norms, national legal reforms, and wide cultural transformations, reproductive and sexual rights have been and continue to be fiercely resisted. Some states – such as China, India, Indonesia, and Vietnam – had been implementing for decades top-down and coercive population policies that would not be easily transformed. In some of these cases, governments, despite having signed the agreement, had no real interest in implementing the required changes; in others, obstacles to transforming policies were not easily overcome.

In China, for example, the strict one child policy remained in place until recently when demographic and economic reasons justified its flexibilization. Quality of care in reproductive care gained space in official policy discourses, but the widespread practice of coercive abortion and contraception was not eliminated. In India, under the pressure of women's organizations, new reproductive health guidelines based on the ICPD premises were adopted by the central government, but reframing the fertility control machinery proved to be daunting. Another example is Vietnam where the early openness to the Cairo agenda would be left behind by the 2000s when the government became anxious about the results of the new market-oriented economic policy then being implemented.

In summary, either because population control models are too resilient or because the ambitious rights agendas of ICPD and Beijing were not absorbed uniformly across the world, almost 20 years later problems related to coercive fertility control practices have not disappeared. A number of recent press reports and academic research findings made visible episodes of female sterilization abuse in contexts as varied as in India, where women are still dying after sterilization camps' procedure, and Slovakia, where Roma women were submitted to forceful tubal ligations; or the 2012 scandal of forced female sterilization scandal in Uzbeskisthan that reached the front pages of the global press.

The refractory policy structures of population control are not the only or even major challenge. Another major obstacle is what the French philosopher Jacques Derrida called 'the return of religion' (Derrida, 1998). Contemporary religious dogmatic revivals that began to be visible in the 1970s have gained scalar proportions in the past 20 years. Today dogmatic visions are manifest in all dominant religious traditions and, practically everywhere, politically organized religious extremists operate to influence legal norms and policies, or take state power through electoral politics. Some authors have interpreted these extremist revivals as a reaction to the growing uncertainties of late capitalism. They pose unpredicted challenges with respect to the foundations and meanings of secularity and everywhere they target transformations underway in sexual, gender, and reproductive orders as one main strategy to recreate 'tradition' and enlarge the space for dominant heteronormativity and sexism to regain social and political legitimacy (Corrêa et al., 2008).

These forces have impacted on the Cairo and Beijing rights agenda and related processes in many ways. For example, following George W. Bush's election in 2000, his administration reinstated Reagan's Global Gag Rule, prohibiting overseas recipients of United States Agency for International Development (USAID) funds to research, provide, or advocate for. Later in 2003 the President's Emergency Plan for AIDS Relief (PEPFAR) was enshrined into a law and began globally promoting abstinence and monogamy as prevention. One year later Congress approved the Trafficking Victims Protection Reauthorization Act (TVPRA) that prohibits organizations receiving U.S. funds to recognize and advocate for the rights of sex workers.

Although since 2009 the Obama administration has either suspended or eased these directives, the effects of the Bush era American policy are still felt in many Southern countries. Furthermore, in the past 20 years U.S. conservative religious forces and advocates have begun to export their agendas worldwide.

The presence of U.S. Evangelicals in sub-Saharan Africa has been triggering trails of sexual panic, violence, murder, and incarceration of gays, lesbians, and trans persons, but also the proposal of extremely regressive legislation in respect to both homosexuality and abortion (Kaoma, 2012). U.S. extremist Evangelism has also established connections with groups in Latin American countries, such as Brazil, where the number of Evangelicals has significantly increased, as well as in China. The rise of active antiabortion forces in the United Kingdom also articulates with similar U.S. initiatives, at least tactically, such as rescue operations in front of clinics.

Religious dogmatic attacks on reproductive and sexual rights in the United States and elsewhere, while strongly grounded in Biblical Evangelism, are also promoted by the Catholic leadership, intellectuals, and groups of faithful. Catholic conservative voices are active in Ireland, the Philippines, Poland, Portugal, Spain, a number of African countries, all of Latin America, and even France, a country known for its strong attachment to secularism and laicité. Mobilizing civil society and media, they sometimes work to elect antiabortion and antihomosexuality politicians to directly influence legislative and judiciary bodies. In contrast to other dogmatic religious voices, contemporary conservative Catholic arguments and strategies are no longer exclusively grounded on faith and theological morality. Rather they are framed in terms of natural law premises and revised forms of constitutionalism and, most significantly, they also resort to bioscientific arguments – in relation to the beginning of life or binary conception of sex and gender to construct political arguments against abortion and homosexuality (Lemaitre, 2012).

When the lens is shifted toward the Islamic world we see the caveats of Islamic cultural and legal norms in respect to gender and sexuality. Yet Islamic doctrine with respect to abortion and contraception has sometimes proved more flexible than Christian extremist canons. Abortion has long been legal in

Turkey (1980s) and Tunisia (1970s), and Iran had until recently a national family planning program and allowed abortion for health reasons.

But once again, since Cairo and Beijing, Islamic states and leaders have also moved toward more radically regressive positions in relation to gender, sexuality, and reproductive matters. In many contexts in Africa, the Middle East, and Asia, Islamic extremists quite often revive the historical racist bias of population control policies to portray and condemn contraception as another Western strategy to reduce the number of Muslims. In 2012 the Turkish president took feminist groups by surprise when he declared the right to abortion should be abolished. In the aftermath of the Arab Spring, new threats to gender equality and related claims in the realms of sexual and reproductive autonomy are also taking place in Egypt and Tunisia after Islamist parties gained election. Positions against reproductive autonomy can also be identified in other non-Christian religious cultures as exemplified by Sri Lankan conservative Buddhists' public positions against abortion. Current debates around sex-selective abortions in India and China – which are blatant examples of violation of women's equality – are now interwoven with slippery discourses on female feticide and the right of girls to be born that easily feed antiabortion stances.

These cross-boundary and multilayered regressive politics have had detrimental effects worldwide. In the heated debates around the 2011 U.S. health reform public funding for abortion procedures was limited. In 2010, in Kenya, religious dogmatic pressures led to the inclusion of a right to life from conception in the text of the new constitution. In the Philippines, a new reproductive health law was voted by the House in 2012 – after many years of feminist struggle – which was once again heavily attacked by the Catholic Church. As of early 2014 final approval of the law was still pending. In Portugal and Spain, Catholic conservatives who systematically attacked recent abortion legal reforms have gained the support of conservative governments elected after 2011.

In Latin America and the Caribbean, Nicaragua (2008) and Dominican Republic (2011), abortion has been totally banned as in Chile (adding to the list of the extremely restrictive laws of El Salvador and Honduras). Strong Catholic pressures delayed the approval of the Uruguayan abortion law for 8 years. In Colombia the 2007 decision of the Constitutional Court allowing access to safe and legal abortion to preserve women's health is resisted at the level of health services. In Mexico, after legal abortion was granted in the Federal District, a number of states adopted a right to life premise in their constitutions. In Brazil, the alliance between conservative Catholics and Evangelicals has been propelling flagrant policy regressions, such as the 2010 law provision to guarantee personhood status to the unborn. Everywhere, South and North of the Equator, conscientious objection is being diffused creating obstacles to access abortion in contexts where it is partially or totally legal.

The impacts of the 'return of religion' are also identified in global policy arenas, in particular United Nations bodies and forums. Since the early 2000s, when the Bush administration strongly pressured for and succeeded in deleting ICPD reproductive health targets and indicators from the Millennium Development Agenda (the targets would be restored in 2005), UN agencies and others have gradually tended to fall back toward a more conventional maternal health and maternal morality, far from sufficient from the point of view of a comprehensive reproductive (and sexual) rights perspective. Furthermore, the Vatican, some Catholic states, and Islamic countries organized around the Organization of the Islamic Conference, and many faith-based religious organizations are now much more present in negotiations involving women's rights, particularly sexual and reproductive rights, than in the 1990s.

In each new UN debate on these topics, a renewed battle erupts making it more and more difficult to preserve the premises of these rights as defined in the 1990s intergovernmental consensus. Despite gains such as the *KL v. Peru* and *Paulina v. Mexico* cases, great difficulties are now experienced with respect to language on abortion rights. For example, it was impossible to explicitly include abortion as a main cause of maternal mortality in the resolutions adopted by the Human Rights Council in 2008.

Alice Miller and Mindy Roseman in a recent assessment of key challenges in sexual and reproductive rights work correctly observe that though both LGBTQI rights and abortion rights are under fierce attack in all UN negotiations and processes in terms of normative developments a flagrant asymmetry is observed. During the past 10 years important forward steps have occurred in respect to granting rights related to sexual orientation and gender identity on the grounds of equality, non-discrimination, right to privacy, and autonomy. In contrast, 20 years after Cairo and Beijing, no international human rights surveillance committee nor UN high-level expert – except the Special Rapporteur on the Right to the Highest Standard of Health – had dared to frontally addressed the decriminalization of abortion on the same grounds or even as a core component of the right to health (United Nations, 2012).

The development of sexual and reproductive rights must also be situated in relation to major developments underway in the vast domain of biomedicine and biotechnology. This dimension had always a central place in earlier debates on population control and sexual and reproductive health and rights, as exemplified by long-standing discussions on contraception and abortion technologies, including those related to sex-selective abortion and HIV/AIDS prevention and treatment, to mention a few. However, the magnitude of scientific and technological breakthroughs of the past 30 years is of a new scale.

These transformations are at the background of the revival of conceptual and policy frames that increasingly privilege technological means as the easiest and quickest solution to complex social problems over the long-term and more uncertain processes of political and cultural transformation. This tendency that, in the 1980s and 1990s debates on population, was criticized by feminists as simplistic tech fix solutions is once again blossoming in diverse areas related to sexual and reproductive health and rights. One example is the recent development of new medical technologies of HIV prevention – such as pre- and postexposure prophylaxis and male circumcision. These new strategies of prevention tend to be disseminated through vertical delivery programs, very similar to those adopted by past fertility control programs, that have incurred in many distortions, such as uncontrolled

side effects and lack of attention to cultural dimensions. Most compelling perhaps is the 'new' global initiative on family planning, launched in 2012 that is supported by the Gates Foundation and DFID and focuses mainly in sub-Saharan Africa. While portrayed by the global mainstream media as a policy proposition that challenges religious views on contraception, the initiative dismisses the Cairo and Beijing comprehensive rights approach and, not surprisingly, has been criticized by many reproductive rights research and advocacy voices from both North and South of the Equator.

Recent scientific and technological developments in the biomedical field are deeply imbricated with neoliberal mechanisms of governamentality that devolve and disperse regulatory functions from the state toward self-regulatory mechanisms of private institutions or consumers. Under these rapidly shifting conditions, the further development, elaboration, diffusion, and political mobilization around reproductive and sexual rights claims are becoming exceedingly more complex, fragmented, and contradictory. Individuals, specific social groups and societies at large, in all continents are both absorbing and negatively reacting to the intense medicalization of practices and the aggressive commercialization of pharmaceutical and biotech products and services, in a world that remains traversed by sharp national and global economic inequities.

On the one hand, millions of people around the world are still coping with 'old' biomedical and technological issues such as the limited options for and poor quality of contraception, restrictions in the access to medical abortion technologies, the high costs of ARVs and the basic obstacles experienced by trans persons in accessing safe body modification and sex reassignment procedures. At the same time the transnationalization of new reproductive technologies including surrogacy or the skyrocketing global marketing for sexual enhancers such as Viagra is now commonplace, not to mention a wide range of novel techniques of body modification including sophisticated plastic surgery of genitals. These are the paradoxical frontlines of what Nikolas Rose – is his chartering of contemporary trends in medicalization, biotechnology, and bio-capitalism – defines as biological citizenship struggles (Rose, 2006).

Political and Conceptual Inquiries

Reproduction, Sexuality and Human Rights: Intersections and Distinctions

Both in theory and in practice, reproductive and sexual right both intersect and are distinct in a variety of complex ways, in the same manner sexuality and gender can overlap or diverge as domains of personal and social relations. It might be useful for reproductive rights advocates to consider how diverse sexualities and reproduction can be both intertwined and delinked and to recognize that sexual and gender diversities raise new challenges for reproductive locations, aspirations, and politics.

Likewise, sexual rights advocates might begin to examine the reproductive potentials of nonnormative sexual and gender identities at the same time they continue to explore the transgressive (as well as accommodative) potentials of every variation in the sex–gender rainbow. They would also do well to recognize that the reproductive realm is not exempt from transgressive practices, the most evident example being the widespread recourse to abortion in contexts where it is criminalized.

These analytical efforts may be guided by confronting such issues as sexual hierarchies; sexual pleasure as a right of heterosexuals and sexual dissidents alike; reproductive and sexual health services that go beyond the dominant gender binary that prevails in biomedical interventions; condemnation of forced sterilization of not only women, men, and disabled persons but also trans persons; reproductive rights as a frame not only for the right of infertile women to become pregnant, but also for the choice of lesbians to be pregnant and to bear and raise children and the complex situations in which single men or gay couples resort to surrogate mothers to be able to biologically reproduce; rights of gay men, trans and intersex people, and sex workers to adopt and raise children; family-related benefits within nontraditional households and kin networks, for persons rather than couples.

Engaging such questions should move us toward greater inclusivity in how we conceptualize both reproductive and erotic justice.

The Indispensability and Insufficiency of Rights

An enabling environment – democratic conditions, social and economic well-being, gender equality, and a climate of ethical pluralism – is a prerequisite for the realization and full exercise of reproductive and by extension sexual rights. Reproductive and sexual rights claims imply both positive entitlements – such as reproductive health care, employment, social security support – and negative rights, such as protection against rape and discrimination, regardless of gender, race ethnicity, or sexual orientation or gender identity and expression. There are no significant signs that these prerequisites have been fulfilled or even significantly improved in the past decade.

However, it is also evident that everywhere, including in the global South, gays, lesbian, and even trans persons who have access to economic, knowledge, and symbolic resources may live their lives without experiencing high levels of discrimination or violence; and affluent women can access safe abortion procedures even where it is illegal. The skyrocketing expansion of LGBT tourist routes and of medical tourism in the area of sexual and reproductive health-specialized services (such as insemination, IFV, surrogacy, and body modifications) is other striking examples of how persons with access to resources may be sexually and reproductively entitled. Yet millions who also have needs and aspirations in the realms of reproduction and sexuality suffer inequalities, lack of resources, threats to personal safety, and human rights violations. The requisite of enabling conditions is therefore, perhaps, more urgent today than in the past.

The trajectory of the past two decades reveals that addressing reproductive and sexuality claims through the liberal paradigm of inclusion through citizenship and human rights had unequivocal gains. Human rights language and interpretative work have been decidedly engendered and sexualized and there have been achievements in relation to nondiscrimination, equality under the law, right to privacy, self-determination, and freedom of expression. This is critical because the realization of reproductive and sexual rights

cannot be attained if premises of personal and individual agency are not guaranteed. However, the androcentric, individualistic, and contractual nature of the dominant Western liberal theory and practice of rights must also be challenged.

World realities make it more compelling today to continue developing reproductive and sexual rights as social and collective entitlements. Furthermore the conversation around these rights must also be situated in relation to militarization and securitization trends in many places. The 'states of exception' that prevail in these circumstances and sites turn rights language and claims into meaningless resources.

With the justice framework proposed by Nancy Fraser (1997) in mind, articulating principles of recognition and redistribution, how can we soberly assess, for example, recent legal victories in relation to same sex marriage in those settings that remain fraught by sharp economic inequalities, where poor trans persons are targets of homophobic murder and no prospect exists for poor women to have access to safe and legal abortion in the short term? We need to critically examine the many pitfalls of the Indian surrogacy industry after access to these services for foreign single individuals and same-sex couples was prohibited in 2013 both from the point of view of poor women who rent their wombs to sustain their family livelihoods and that of those persons who may be said to have been discriminated against by the new rules. What can we say about the total absence of sexual and reproductive entitlements of millions of persons incarcerated worldwide (particularly in the United States, Brazil, and China that comprise almost 80% of the global inmate population), or of refugees from conflicts in Syria and elsewhere?

To once again return to Jacques Derrida, rights work in relation to reproduction and sexuality cannot avoid threading the thin line between, on the one hand, the appraisal and enlargement of rights language as indispensable and, on the other, the recognition that rights are also insufficient to meet expanded demands of justice (Derrida, 1992).

Religious Dogmatism, Secularity, and Science: Troubled Crossroads

As increasingly extremist forms of religion directly attack gender equality and reproductive and sexual freedoms, secularity and *laicité* has gradually gained an aura of either a lost golden age, or a key pathway to overcome the threats of dogmatism. We must be cautious with this schematic division locating religion on the side of gender domination and sexual repression and secularity as the realm of gender equality and sexual freedom and nondiscrimination. The historical experience and current realities are much more complex and hazy. Even in a number of countries hailed for respect for secularity, the separation between state and church were not are as sharp and clear as suggested. Both in the past and today secular states institutions and norms have curtailed and continue to curtail reproductive and sexual autonomy, either because they remained influenced by religious views, or because they are informed by strictly secular, in fact, scientific rationales of control and discipline. The criminalization of abortion may be impregnated with religious morality, but it has well-served states' natalist objectives. Sodomy laws were much more about the secular morality of the male bourgeois and colonizer and scientific notions of sexual deviance than about sin. The sexual ideologies of communist states where secularity was compulsory were not exactly liberal.

Secularity and secularization are without doubt enabling conditions or prerequisites of religious freedom as well as of moral plurality required for the realization of gender, reproductive, and erotic justice. But secularity should neither be mystified as a silver bullet that will resolve the conflicts and controversies of sexual and reproductive politics today, nor spared from critical assessments. Quite possibly for the foreseeable future reproductive and sexual rights activists and intellectuals will be challenged to reengage with religion without 'returning' there. This implies recognizing religion as a perennial, perhaps, but also changing aspect of political and social reality, not it's 'opposite.' This kind of critical engagement with religions means contesting the injustices perpetrated in their names, while avoiding the demonization of religious communities as a whole – as occurs in the case of Islamophobia – while at the same time opening spaces and establishing connection with the voices of plurality and respect to sexual and reproductive freedom that exist in all religious traditions.

Caution is also required in respect to science (and technology) as the main strategic intellectual and pragmatic resource to countervail religious dogmatism. The core tenets of scientific inquiry and the valuing of systematic doubt and empirical evidence are certainly important sources to guide arguments in favor of reproductive and erotic justice in the difficult and complex conversations underway in the most diverse public spheres torn apart today between religious and nonreligious visions on these matters. However we must remember that dominant scientific conceptions of 'sex,' gender, and reproduction also place these dimensions of human experience squarely in the laws of nature hampering transformed visions that emphasize the contingency and plasticity of these domains, which are critical to the contemporary politics of reproduction and sexuality. The naturalizing imprint of scientific conceptions and discourses is what allows them to be increasingly used by dogmatic religious actors in their political arguments around the right to life or unshakable natural imprint of gender, sexual attraction, and love.

Lastly but not less relevant scientific rationales have been and remain one key source of ideologies of social engineering – and within it of extreme forms of sexual and reproductive disciplining or even triage. The ethical risks and potential human rights violations implied by the uses and abuses of scientific instrumental rationality for societal administration are greater and more palpable today than ever. The future development of sexual and reproductive rights is therefore challenged to increasingly engage with the novel trends and scientific production as to make sure that science and technology will be guided by the ethical premises of freedom of will, expression and consciousness, equality, privacy, and nondiscrimination.

See also: Biocitizenship; Biomedical Sciences, History and Sociology of; Civil Liberties and Human Rights; Eugenics as

a Basis of Population Policy; Genomics, Ethical Issues in; Globalization and Religion; Human Rights: Legal Aspects; Postcolonial Law; Religion: Mobilization and Power; Sexual Orientation and the Law.

Bibliography

Altman, D., 2013. Power & Community. Routledge, London/New York.

Altman, D., 2002. Global Sex. University of Chicago Press, Chicago.

Berer, M., 2011. Repoliticising sexual and reproductivehealth and rights. Reproductive Health Matters 19 (38), 4–10.

Boyce, P., Khanna, A., 2011. Rights and representations: querying the male-to-male sexual subject in India. Culture, Health & Sexuality 13 (1), 89–100.

Cabral, M., 2012. La ley de identidad de género. http://www.sxpolitics.org/es/?p=1967 (retrieved from 25.04.12.).

Center for Reproductive Rights, 2006. The Protocol on the Rights of Women in Africa: An Instrument for Advancing Reproductive and Sexual Rights – A Briefing Paper, New York.

Center for Reproductive Rights, 2013. The World Abortion Laws, 2013. http://worldabortionlaws.com/ (retrieved from 15.03.12.).

Cheah, P., 1997. Positioning human rights in the current global conjuncture. Public Culture 9 (2), 233–266.

Chesler, E., 1992. Woman of Valor: Margaret Sanger and the Birth Control Movement in America. Simon & Schuster, New York.

Corrêa, S., with Reichmann, R., 1994. Population and Reproductive Rights: Feminist Perspectives from the South. Zed Books, London.

Corrêa, S., Parker, R., Petchesky, R., 2008. Sexuality, Health & Human Rights. Routledge, London, New York.

Corrêa, S., Richard-Davis, B., Parker, R., 2013. Sexuality and Globalities. In: Tolman, Deborah L., et al. (Eds.), APA Handbook of Sexuality and Psychology. American Psychological Association, Washington (Contextual Approaches, v. 2).

Currah, P., Juang, R.M., Minter, S.P. (Eds.), 2006. Transgender Rights. University of Minnesota Press, Minneapolis.

Darwin, C., [1859]2009. The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life. In: Bynum, W.F. (Ed.). AL Burt.

Darwin, C., [1871]2004. The Descent of Man and Selection in Relation to. Publishing. SexDigireads.com.

Derrida, J., 1992. Force of law: "the mystical foundation of authority". In: Cornell, D., Rosenfeld, M., Carlson, D.G. (Eds.), Deconstruction and the Possibility of Justice. Routledge, New York and London.

Derrida, J., 1998. Faith and reason: the two sources of "religion" at the limits of reason Alone. In: Derrida, J., Vattimo, G. (Eds.), Religion. Stanford University Press, Stanford, pp. 1–78.

Dixon-Mueller, R., 1993. Population Policy and Women's Rights: Transforming Reproductive Choice. Praeger, Westport.

Engels, F., 1884. The Origin of the Family, Private Property and the State. Available online at www.marxists.org/archive/marx/works/1884/origin-family/index.htm (accessed 30.10.14.).

Foucault, M., 1978[orig. 1976]. The History of Sexuality: An Introduction, vol. 1. Pantheon, New York.

Fraser, N., 1997. Justice Interruptus: Critical Reflections on the 'Postsocialist' Condition. Routledge, London, New York.

Fried, S.T., Landsberg-Lewis, I., 2000. Sexual rights: from concept to strategy. In: Askin, K., Koenig, D. (Eds.), Women's Human Rights Reference Guide. Transnational Press, New York, pp. 91–122.

Garcia, J., Parker, R., 2006. From global discourse to local action: the makings of a sexual rights movement. Horizontes Antropológicos 12 (26), 13–42. http://dx.doi.org/10.1590/S0104-71832006000200002.

Girard, F., 2007. Negotiating sexual rights and sexual orientation at UN. In: Parker, R., Petchesky, R., Sember, R. (Eds.), SexPolitics: Reports from the Frontlines. http://www.sxpolitics.org/frontlines/home/index.php (retrieved from 18.03.12.).

Harcourt, W., 2009. Body Politics in Development – Critical Debates in Gender and Development. Zed Books, London.

Hartmann, B., 1995. Reproductive Rights and Wrongs. South End Press, Boston.

Hunt, P., February 16, 2004. Economic, Social and Cultural Rights: The Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health, Report of the Special Rapporteur. UN Commission on Human Rights, 60th Session.

ILGA, 2012. State-sponsored Homophobia Report and Gay and Lesbian Rights Maps. http://old.ilga.org/Statehomophobia/ILGA_State_Sponsored_Homophobia_2011.pdf (retrieved from 30.04.12.).

Ilkkaracan, P., Jolly, S., 2007. Gender and Sexuality: Overview Report. Bridge, Institute of Development Studies. Brighton, UK. http://www.bridge.ids.ac.uk/reports/CEP-Sexuality-OR.pdf (retrieved from 22.11.11.).

International Commission of Jurists, 2010. Sexual-orientation and Gender Identity in Human Rights Law References to Jurisprudence and Doctrine of the United Nations Human-rights-system. http://www.icj.org/sexual-orientation-and-gender-identity-in-human-rights-law-references-to-jurisprudence-and-doctrine-of-the-united-nations-human-rights-system-4th-edition/ (retrieved from 01.11.12.).

Kaoma, K., 2012. Colonizing African Values: How the US Christian Right Is Transforming Sexual Politics in Africa. Political Research Associates, Somerville, MA.

Kempadoo, K., Doezema, J., 1998. Global Sex Workers: Rights, Resistance, and Redefinition. Routledge, NewYork and London.

Lemaitre, J., 2012. By reason alone: Catholicism, constitutions and sex in the Americas. International Journal of Constitutional Law 10, 493–511.

Miller, A., Roseman, M., November 2011. Sexual and reproductive rights at the United Nations: frustration or fulfilment? Reproductive Health Matters 19 (38), 102–118. http://dx.doi.org/10.1016/S0968-8080(11)38585-0.

O'Flaherty, M., Fisher, J., 2008. Sexual orientation, gender identity and international human rights law: contextualizing the Yogyakarta principles. Human Rights Law Review 8, 207–248.

Parker, R., 2001. Sexuality, culture and power in HIV/AIDS research. Annual Review of Anthropology 30, 163–179.

Pecheny, M., de la Dehesa, R., 2014. Sexuality and politics in Latin America: Outline for a Discussion. In: Corrêa, Sonia, de la Dehesa, Rafael, Parker, Richard (Eds.), Sexuality and Politics: Regional Dialogues from the South. Sexuality Policy Watch, Rio de Janeiro. Accessed at http://www.sxpolitics.org/sexuality-and-politics/pdfs/volume1/4.pdf.

Petchesky, R., Judd, K. (Eds.), 1998. Negotiating Reproductive Rights: Women's Perspectives across Countries and Cultures. International Reproductive Rights Research Action Group.

Richardson, D., 2000. Rethinking Sexuality. Sage, London.

Rose, N., 2006. The Politics of Life Itself: Biomedicine, Power and Subjectivity in the Twenty-first Century. Princeton University Press, Princeton, NJ.

Rubin, G., 1984. Thinking sex: notes for a radical theory of the politics of sexuality. In: Vance, C. (Ed.), Pleasure and Danger: Exploring Female Sexuality. Routledge & Kegan Paul, Boston.

Saiz, I., 2004. Bracketing sexuality: human rights and sexual orientation – a decade of development and denial at the United Nations. Health and Human Rights 7, 48–81.

Sanders, D., 2009. 377 and the unnatural afterlife of British colonialism in Asia. Asian Journal of Comparative Law 4. Article 7. http://www.sxpolitics.org/wp-content/uploads/2009/04/new-version_377_sander.pdf (retrieved from 05.08.13.).

Sen, G., Corrêa, S., 1999. Cairo plus five: moving forward in the eye of the storm. In: Social Watch Report, 1999 Edition. http://old.socialwatch.org/en/informesTematicos/36.html (retrieved from 10.08.13.).

Tamale, S., 2009. A human rights impact assessment of the Ugandan Anti-homosexuality Bill 2009. The Equal Rights Review 4, 49–57.

Tan, M., 2014. Reflectionson digital information and communication technologies, gender and sexualities in Asia. In: Corrêa, S., de la Dehesa, R., Parker, R. (Eds.), Sexualities and Politics: Regional Dialogues from the Global South. http://www.sxpolitics.org/sexuality-and-politics/pdfs/volume1/7.pdf (accessed on 29.10.14.).

Thoresen, R., 2009. Queering human rights: the Yogyakarta Principles and the norm that dare not speak its name. Journal of Human Rights 8, 323–339. http://dx.doi.org/10.1080/14754830903324746.

United Nations, 1994. Program of Action of the International Conference on Population and Development. United Nations, New York.

United Nations, 1996. Beijing Declaration and Platform for Action, Adopted by the Fourth World Conference on Women (1995). United Nations, New York.

United Nations, 2004. Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental (E/CN.4/2004/49).

United Nations, 2011. Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental (A/66/254). http://daccess-ddsny.un.org/doc/UNDOC/GEN/N11/443/58/PDF/N1144358.pdf?OpenElement.

United Nations, 2012. A/HRC/14/20 Report of the Special Rapporteuron the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health. Anand Grover. http://www2.ohchr.org/english/bodies/hrcouncil/docs/14session/A.HRC.14.20.pdf (accessed 30.10.14.). (retrieved from 25.07.13.).

United Nations, 2012. A/HRC/14/20 Report of the Special Rapporteuron the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health. Anand Grover (accessed 30.10.14.). http://www2.ohchr.org/english/bodies/hrcouncil/docs/14session/A.HRC.14.20.pdf.

Relevant Websites

http://www.creaworld – CREA
http://www.reproductiverights.org – Center for Reproductive Rights (CRR)
http://www.ighhrc.org – International Gay and Lesbian Human Rights Commission (IGLHRC)
http://ilga.org – International Lesbian, Gay and Transexual Association (ILGA)
http://www.ipas.org – IPAS
http://www.rhmjournal.org.uk/ – Reproductive Health Matters
http://www.sxpolitics.org – Sexuality Policy Watch
http://www.ypinaction.org – Yogakarta Principles in Action

Reproductive Rights in Affluent Nations

Roxanne Mykitiuk, Osgoode Hall Law School, York University, Toronto, ON, Canada
Robyn Lee, Banting Postdoctoral Fellow, Brock University, St. Catharines, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Reproductive rights have historically focused on access to safe, legal contraception and abortion. More recently, reproductive rights have come to include preservation of reproductive health and access to assisted reproductive and genetic technologies that facilitate procreation. Recognition of additional barriers to reproductive self-determination experienced by members of marginalized groups has led to demands for reproductive justice that address the concerns of poor women, racialized women, women with disabilities, and LGBTQ individuals.

Introduction

Beginning in the nineteenth century and gaining strength in the twentieth century, the struggles for access to safe, legal contraception and abortion have been the most dominant issues in the pursuit of reproductive rights in developed nations. Recent advances in technology have meant that assisted reproduction and genetic screening and testing have also become contested sites in struggles over reproductive rights. Earlier accounts of reproductive rights were largely concerned with women's abilities to make choices to avoid unwanted procreation, whereas in the contemporary context, reproductive rights have been argued to include a spectrum of rights aimed at preserving women's reproductive health, enabling procreation, and possibly determining the characteristics of one's future children. Whereas earlier struggles over contraception and abortion were framed as upholding women's reproductive autonomy, recent debates on reproductive justice also include the right to bear and raise children and the determination of genetic characteristics of those children.

Complicating this view is the linkage of reproductive rights with social justice developed by organizations advocating for the rights of women of color beginning in the United States during the late 1980s. The concept of reproductive justice arose from recognition of the substantial barriers poor women, racialized women, women with disabilities, and LGBTQ individuals have always encountered in exercising their reproductive rights. This broader conception of reproductive justice, which recognizes that women from marginalized groups face additional obstacles to reproductive self-determination, has since expanded beyond the U.S. context to address obstacles faced by women worldwide.

Reproductive Rights

Reproductive rights have been grounded in the recognition that individuals are entitled to bodily autonomy and the ability to make their own reproductive choices. Autonomy is conventionally defined in terms of individual control over one's body and behavior, analogous to ownership of one's self. This narrow understanding of autonomy and choice obscures the extent to which behaviors and actions are profoundly affected by broader social influences. Relational autonomy offers an alternative understanding of autonomy and selfhood that recognizes that we are both shaped through our relationships with others, while at the same time we are autonomous in not being determined by our relationships (Nedelsky, 2011). Under a relational autonomy framework, reproductive rights cannot be limited to protections for individuals from interference from the state, but must also include necessary supports for bearing and raising children. However, reproductive rights in developed nations do not generally follow a relational autonomy framework.

Conventional debates on reproductive rights have been animated by a number of distinctions that have contributed to informing their scope and content. Central has been the distinction between negative and positive rights and the distinction between procreative and parental rights. More recently, international law sources and norms have also informed the content and scope of reproductive rights.

Traditionally, negative rights are understood as civil and political rights, considered to be guarantees of freedom from state interference. Positive rights are understood in terms of social and economic rights, as affirmative obligations on the state to enable certain rights. In the context of reproductive rights, negative rights would include protections against forced sterilization, for example, while positive rights would require the state to provide access to a minimum standard of sexual and reproductive health care.

It is important to distinguish between procreative and parental rights, since parental rights do not necessarily follow from procreative rights. Definitions and protections for both of these types of rights vary widely in different jurisdictions. For example, procreative rights include whether, how, when, and with whom to procreate. While it is often assumed that parental rights flow automatically from the exercise of procreative rights, in some jurisdictions they do not. This has become especially apparent in the case of determining parental status when assisted reproductive technologies (ARTs) are used and multiple parties are claiming parental status. Biological progenitors and/or surrogates who have an embodied relationship with the resulting child may be denied parental rights in relation to the child with whom they have a procreative relationship. Even in circumstances where ARTs are not used, in some jurisdictions the doctrine of best interests of the child takes precedence over parental rights; for instance, in the case of child abuse, parental substance abuse or incarceration, and in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86143-7

disputes over child custody in the event of divorce (Meyer et al., 2010; Scott and Emery, 2013). For example, in Canada, the *Divorce Act* mandates that the best interests of the child be the principal consideration in determining child custody arrangements (Branch, 2007). On the other hand, Australia passed the *Family Law Amendment (Shared Parental Responsibility) Act* (2006), which emphasizes parental rights (in the form of shared custody after divorce) above the best interests of the child.

Reproductive rights encompass both procreative and parental rights. Reproductive rights for women include the right to decide whether or not to have children, how many children they will have, and the spacing and timing of pregnancies (Cook et al., 2003). Reproductive rights also include the right to be free of illnesses and other conditions that might interfere with sexual activity and reproduction (Cook et al., 2003). The right to found a family and the right to raise healthy children through providing for their physical, social, emotional, and educational needs are also important reproductive rights (Washenfelder, 2004).

There is increasing awareness within the international human rights community about the need to protect the sexual and reproductive health of women. Women's rights to intimate relations and reproductive choice have been expressly recognized in international human rights law – particularly in the Convention on the Elimination of all Forms of Discrimination against Women (CEDAW) adopted in 1979 by the United Nations General Assembly – in such areas as protections for marital status (CEDAW, Article 16), access to contraception and family planning (CEDAW, Article 12), and pre- and postnatal care (CEDAW, Article 12). Sexual and reproductive rights are protected in various internationally recognized principles of human rights relating to health and self-determination.

Since rights depend on the observance of other rights, this collection of international human rights principles must be interpreted and applied interactively in order to advance women's sexual and reproductive health. Some of the international human rights principles that apply to legal, social, economic, and political rights and which may consequently provide legal grounding for human rights to parenting and reproductive health include: the right to life, liberty, and security of person (Universal Declaration of Human Rights, Article 3); the right to equality and nondiscrimination (Universal Declaration of Human Rights, Article 7); the right to marry and have a family (Universal Declaration of Human Rights, Article 16); the right to the highest standard of attainable health (Universal Declaration of Human Rights, Article 25); the right to reproductive health, understood to include family planning and maternal health services (CEDAW, Article 12); and the right to information and education concerning sexual health, family planning, and reproductive services (CEDAW, Article 12; Article 14b).

Contraception and Abortion

The struggle for reproductive rights in the developed world historically has centered on access to contraception and safe, legal abortions. Contraception in the West has largely been framed as an issue of morality. Support for contraception was equated with support for extramarital sex and from the late 1800s onward, contraception was illegal in Canada and the United States. Contraception was legalized in the United States in 1965, in France in 1967, and in in Canada in 1969. Since then contraception has become increasingly available worldwide (Bullough, 2001).

Contemporary barriers to accessing contraception include cost, requirements for parental consent, differential prescription practices for different groups of women depending on race and socioeconomic status (Borrero et al., 2013), need for prescriptions rather than over-the-counter access, and conscientious objection among prescribers and health care professional service providers (McLeod, 2010). Current controversies in contraception include access to the morning-after pill, safety concerns, and coerced contraception based on race and socioeconomic status as in the case of the Norplant implant controversy. Throughout the 1990s, U.S. legislators attempted to provide financial incentives to welfare recipients who received the Norplant implant. Some judges offered the implants to women as an alternative to jail terms – particularly racialized women and women on welfare. Teenaged girls in predominantly Black schools in the United States were also targeted by campaigns to promote use of the Norplant implant (Roberts, 1997).

Abortion is an enduring source of controversy in the West and its legality varies widely among countries and over time. It has been a rallying point for women's reproductive rights, restricted historically, but with access expanding over time. Despite the trend to liberalized abortion, barriers to access continue to be introduced in many countries, including mandatory waiting periods, biased counseling requirements, and unrestricted conscientious objection among health care providers (Finer and Fine, 2013).

In many European countries, abortion is allowed only during the first trimester; abortions during later stages of pregnancy being allowed only for specific reasons, such as the life or health of the pregnant woman, the occurrence of a sexual offense including rape or incest, or fetal impairment (Boland, 2010). For example, Germany requires a mandatory 3-day waiting period and a counseling session and forbids abortions after the first 12 weeks of pregnancy except in cases of grave threat to the pregnant woman's physical or mental health. Similarly, the Netherlands requires a 5-day waiting period, after which abortion is permitted until fetal viability, legally defined as 24 weeks but usually interpreted as 22 weeks. A notable exception is Northern Ireland, where abortion at any stage of pregnancy remains illegal unless the life of the pregnant woman is at immediate risk or there is a long-term or permanent risk to her physical or mental health (Daniels et al., 2013).

In the United States, abortion law depends on the duration of pregnancy. Abortion is permitted until the point of fetal viability, variously interpreted in different states, with Alabama holding the strictest definition of 19 weeks (Arzuaga and Lee, 2011). Many U.S. states have unsuccessfully attempted to pass laws banning abortion, but have imposed mandatory restrictions such as waiting periods, counseling, and ultrasounds. Canada is unusual in having no criminal laws pertaining to abortion.

Those who oppose abortion generally justify their position through an appeal to the rights of the embryo or fetus.

Supporters of rights to abortion appeal to women's rights to self-determination. Reproductive freedom was foundational to the modern feminist movement of the 1960s and 1970s. Access to contraception and abortion helped to liberate women from unwanted pregnancy. The conventional understanding is that the right to contraception and abortion is fundamentally a negative right based on noninterference in women's reproductive choices by the state. However, this is too narrow a conception of reproductive rights and limits safe access to abortion. As Rodgers and Downie (2006) point out, even in countries where abortion is legal, access can vary widely depending on cost, geographical location, and access to health services.

Fetal Personhood and Maternal Agency

Although laws vary globally, in Australia, the United Kingdom, Canada, and most parts of the United States, legal personhood is only accorded to a child once it exists independently from the mother; the fetus has no legal personhood. However, in many jurisdictions the fetus is granted special legal status as a result of being viewed as a potential person (Harris-Short, 2009). Debates over the status of the fetus occur between those who attribute personhood to the fetus at varying stages of development and those who deny that it has any legal status until it is born, with a range of positions in between.

The debate over fetal protection was active in many countries in the late 1980s and early 1990s and continues today. Women's rights were overwhelmed by the so-called 'protective incarcerations' of pregnant, drug-addicted women, some of whom received nonconsensual medical treatments. Some women whose infants were born testing positive for drugs were prosecuted for harm alleged to have been caused *in utero*. Poor, racialized women have been particularly targeted (Roberts, 1991). Such prosecutions continue in some U.S. states, in addition to repeated unsuccessful attempts to establish legal personhood status for fetuses.

The Supreme Court of Canada has to date refused to assign liability to pregnant women for harms to the fetus before birth. In the case of *Dobson (Litigation Guardian of) v. Dobson* (1999), the court considered whether a fetus born alive and viable could bring a tort action against his biological mother for injuries allegedly caused while *in utero* as a result of her negligent operation of a motor vehicle. Distinguishing between a legal action brought against a pregnant woman by her born alive child and one brought against a third party, Justice Corey wrote that the imposition of liability "may make life for women who are pregnant or who are merely contemplating pregnancy intolerable." The court held that the liberty and equality rights of women would be violated by a finding of liability in these circumstances and that their privacy and autonomy rights would be jeopardized. In *Winnipeg Child and Family Services (Northwest Area) v. D.F.G.* (1997), the Supreme Court of Canada considered whether a pregnant aboriginal woman (Mrs G) addicted to glue sniffing could be involuntarily detained until the birth of her child and forced to undergo treatment for her addiction. A majority of the court held that the fetus has no legal rights until born alive and therefore there was no legal person on whose behalf Mrs G could be detained. Neither the law of tort nor the court's *parens patriae* jurisdiction can be used to detain and treat a pregnant woman for the purpose of preventing harm to an unborn child. Both of these cases reject fetal personhood and reassert women's reproductive autonomy and security of the person.

The concept of fetal viability in law is used to limit reproductive choices for women. Fetal viability refers to a developmental stage in which the fetus can survive outside the womb if it were prematurely delivered. In the groundbreaking U.S. Supreme Court abortion case (*Roe v. Wade*, 1973), the fetus was determined to have limited legal personhood at the point of viability. In *Roe v. Wade* (1973), states could regulate to limit abortions in the third trimester except where they were necessary to save the mother's life. But in the later Supreme Court case (*Casey v. Planned Parenthood*, 1992), the emphasis on the third trimester was changed. Instead, the court ruled that state laws were not to place any undue burden on women seeking abortions up to the point of viability. Viability could arise before the third trimester given advances in technology. As a consequence, fetal viability has been constructed as the moment when a fetus gains personhood status under the law, rather than personhood beginning with birth.

Although fetal personhood and maternal agency are often viewed as competing interests under a traditional liberal conception of subjectivity, a relational understanding of subjectivity allows us to see that the bodily boundaries of fetus and pregnant woman are actually intermingled. Reconceptualizing autonomy through a relational lens allows for the development of laws that recognize a fuller understanding of selfhood as constituted through relationships with other people and social institutions, bodily practices, and beliefs (Downie and Llewellyn, 2011; Nedelsky, 2011). Such a view of the relationship between pregnant woman and fetus is also consistent with postmodern feminist descriptions of the body in which self and other are never fully separated (Haraway, 2013; Shildrick, 1997). However, as Ordolis points out, a relational approach must be reconciled with respect for women's reproductive autonomy. This requires going beyond the inevitably adversarial approach of the courts, recognizing the systemic causes of maternal substance abuse, including discrimination, poverty, and violence, and promoting maternal and fetal health by supporting women's capacity to make healthy choices without eroding their bodily integrity (Ordolis, 2008).

Reproductive Justice

The concept of reproductive justice was developed and principally advocated for by organizations of women of color who were critical of the narrow focus on access to contraception and abortion and called for a broader analysis of the racial, economic, cultural, and structural constraints on women's reproductive choices. These choices for marginalized women have been forcefully curtailed as a result of systematic oppression. According to the reproductive justice movement, the pro-choice paradigm of reproductive rights fails to recognize broader infringements of reproductive rights faced by women of color and aboriginal women, poor women, women with disabilities, and women from other marginalized communities

(Smith, 2005). In this way, the concept of reproductive justice goes beyond choice and promotes women's reproductive self-determination more broadly, recognizing the systematic forms of oppression to which many women are subject.

Women of color, lesbian, bisexual and queer women, poor and disabled women, and women from other marginalized groups have faced obstructions of their reproductive rights, including forced sterilization, restrictions on entering into a legally valid marriage, limitations on access to work to protect reproductive capability, and birth registration rules that impose paternity and fail to recognize maternity (Rodgers, 2007). Sexual sterilization was (and still is) carried out without consent on many groups of women, including women of color and aboriginal women, poor women, women with mental and physical disabilities, and women suspected of being criminally deviant or promiscuous. Historically, eugenics programs attempted to remove unwanted or undesirable characteristics, thought to be hereditary, from society, and operated in a larger context of colonial racism and xenophobia. For example, in Canada, the *Sexual Sterilization Act* in Alberta came into effect in 1928 and was not repealed until 1972 (Grekul et al., 2004). What was widely referred to as the Eugenics Board authorized 4725 sterilizations and carried out 2822 of them (Caulfield and Robertson, 1996). Aboriginal and Métis people were, in particular, targeted by the Eugenics Board.

Women with Disabilities

Women with disabilities are often prevented from participating in sexual, reproductive, and nurturing activities. They often experience double discrimination; regarded by many as asexual and not aspiring to motherhood. Reproductive rights to which women with disabilities face significant barriers include sexual health education, the right to intimate relationships, the right to marry, the right to sexual and reproductive health (including contraception, obstetrical care, and reproductive assistance), and the right to found a family.

Women with disabilities are protected under the United Nations Convention on the Rights of Persons with Disabilities, which guarantees "the same range, quality and standard of free or affordable health care and programmes as provided to other persons, including in the area of sexual and reproductive health and population-based public health programmes" (Article 25a). However, these services are often unavailable (Gibson and Mykitiuk, 2012).

Frequently girls and women with disabilities do not have access to nonjudgmental information and counseling regarding their sexual and reproductive health needs (Quinn et al., 2011). They are often considered not to require comprehensive sexual and reproductive health education because of social prejudice (Esmail et al., 2010). Generic sexual and reproductive health education materials may not provide accurate information about bodily differences in women with disabilities and may not be available in accessible formats. As well, girls and women with intellectual disabilities may not receive adequate information as a result of being infantilized and overprotected (Isler et al., 2009). Health care providers often lack training in sexual health issues relevant to persons with disabilities, medical facilities are often physically inaccessible, and services are often provided in an inaccessible format (Gibson and Mykitiuk, 2012; Gibson and O'Connor, 2010). Additionally, there are obstacles to women with disabilities gaining access to contraceptives and receiving accurate knowledge about their contraceptive options.

Women with disabilities face significant obstacles to social participation, which limit their ability to enter into and enjoy social relationships and express themselves sexually (Lafferty et al., 2012). Women with disabilities who live in institutional settings are often blocked from sexual expression and developing intimate relationships, while those who live independently but require personal assistance may have their privacy negatively impacted. As a result of prevailing social prejudices, most service agencies for people with disabilities have overlooked sexual issues or have deliberately refused to address the sexual needs of people with disabilities (Couldrick et al., 2010).

Similarly, women with disabilities face substantial barriers to parenting. They face difficulty accessing both obstetrical care and ARTs due to discriminatory assumptions about their ability to become parents. Women with disabilities are often viewed as less maternal and unfit to parent, criticized for becoming pregnant, and assumed to be incapable of caring for their children (Silvers, 1998). Mothers with disabilities are often subject to increased surveillance from the child protection system because they are often assumed to be unable to care for their children. Child protection workers may have negative assumptions about women with disabilities due to the prevalence of myths and stereotypes about disability and the child protection system may not be sensitive to the socioeconomic context of some women with disabilities, which may include a lack of disability-related supports and services, under or unemployment, and abuse.

LGBTQ Reproductive Rights

ARTs allow for a broader range of choice in ways of reproducing and in the creation of new family forms. Lesbian, gay, bisexual, and queer individuals often rely on ARTs in order to have genetically related children in the absence of procreative sex. However, they face barriers to accessing ARTs, including prejudice against homosexuality and the high cost of the procedures (Corbett et al., 2013; Greenfeld and Seli, 2013). In addition, laws determining parent/child relationships, where disputes arise in the context of the use of ARTs, often reinscribe a heterosexual nuclear family form and fail to recognize the parental rights of queer family members (Dempsey, 2010).

Lesbians have been able to reproduce, often without medical intervention, through the process of artificial insemination using donor sperm from either anonymous or known donors. Since gay men cannot biologically bear a child, they require the participation of one or sometimes two women in order to produce a child to whom they are genetically related. In some cases, where it is legal, one woman may consent to donor insemination with the sperm of the intended gay father and consent to act as a surrogate mother to enable him to have a child genetically related to him. In other cases, one woman may undergo the medical procedures and risks incurred in *in vitro* fertilization (IVF) in order to produce ova for fertilization,

while a second woman will gestate the embryo(s) and give birth to the baby. In order to have a genetically related child without a female coparent, it is necessary for an intended gay father to procure both an egg donor and surrogate or someone willing to do both often at costs that greatly limit their accessibility (Mitchell and Green, 2007). Adoption may be a less expensive route to gay paternity, but adoption is often more difficult for gay parents, particularly in the case of international adoption (Brodzinsky and Pertman, 2011; Lewin, 2006). The use of known donors (both sperm and ova) and coparenting among lesbian, gay, and bisexual parents of opposite sexes may also be carried out in order to form chosen families, transforming gendered parental roles (Padavic and Butterfield, 2011).

Lesbian and bisexual women in comparison with gay and bisexual men face different obstacles to procreation because acquiring access to donor sperm is frequently easier than accessing donated ova and finding a woman who is willing to gestate. In addition, legal regimes may also construct rules restricting the procreative choices of gay, lesbian, and bisexual parties. For example, in some cases, lesbian women have had their reproductive choices thwarted by demands for legal paternal recognition by sperm donors (Dowd, 2013). Despite the frequent use of assisted reproduction by LGBTQ people, barriers persist, including heteronormative attitudes and discrimination, difficulty accessing genetic material, lack of legal protections for same-sex marriage and parental rights of nonbiological parents, and high costs (Creating Our Families, 2010; Rank, 2010).

Transgender and gender-nonconforming individuals face additional obstacles in exercising reproductive rights. Trans parents challenge conventional gender roles and assumptions, requiring a reconceptualization of pregnancy discrimination law in order to protect the reproductive rights of trans individuals (Karaian, 2013). Since Thomas Beattie entered the public eye as a pregnant trans man in 2008, trans parents have become more visible, but continue to face substantial challenges in exercising reproductive rights due to discrimination and ignorance (Ryan, 2009). They face significant barriers in accessing health care (Bauer et al., 2009), and although trans parenting is underresearched, preliminary studies indicate that ART service providers are unprepared to meet the needs of trans persons (James-Abra, 2012).

Transgender people are also subject to legally prescribed, state-enforced sterilization (Castagnoli, 2010). Sixteen European countries currently do not permit trans persons to legally change their sex without undergoing sterilization; Sweden has recently repealed this requirement (European Union Agency for Fundamental Rights, 2010). Thus, reproductive rights are deeply gendered across a multiplicity of gender identities.

Assisted Human Reproduction

Assisted reproduction allows for the deliberate manipulation of the materials and processes of reproduction in the absence of sexual intercourse. Examples of ARTs include IVF, intracytoplasmic sperm injection (ICSI), insemination with donor sperm, and reliance on third-party reproductive materials and capacities, including sperm and egg provision and gestation. The use of ARTs allows infertile individuals and couples, single women, and people who are lesbian, gay, or transsexual the opportunity to have children to whom they are genetically related. ARTs may also be used, often in combination with preimplantation genetic diagnosis (PGD), by individuals or couples who carry genetic conditions and seek to avoid passing those conditions to their children, or who desire to pass on specific genetic conditions or characteristics to their children.

The use of ARTs expands the realm of reproductive choices and options for some individuals and couples, providing increased opportunities to have a child to whom one is genetically related or with particular genetic characteristics. However, the social, legal, and economic context in which decisions about the use of ARTs are determined, who gets access to them and under what circumstances, raises questions about the nature and scope of reproductive rights. Central to this is an examination of the right to reproduce as a positive right, and what this means not only when the state is called upon to facilitate access to ARTs, but also when exercising one's right to reproduce may involve the participation of another human being. The practice of using ARTs also invites discussion of the reproductive health needs and choices of women who donate ova or gestational services and the concomitant issue of payment for reproductive labor and tissues.

The most well-known reproductive technology is IVF in which a woman's ovaries are stimulated with menotropins to produce multiple oocytes which are removed through transvaginal ultrasonographic-guided needles. Oocytes are then fertilized in petri dishes and one or two transferred to the woman's uterus. Any remaining embryos are cryopreserved for future use, or donated to research, donated for implantation, or discarded once it has been determined that remaining embryos will no longer be needed for an individual's or couple's own reproductive purposes.

Use of embryos in research requires safeguards in order to protect women, who are most directly affected by ARTs. Given the risks of egg retrieval, additional embryos created through IVF should be frozen (cryopreservation) for possible use in future IVF cycles (Nelson et al., 2008). The decisions to end cryopreservation and to donate or discard the embryos should be made separately. Informed consent must be obtained from both the gamete and embryo providers before using embryos for research purposes. Potential donors should be provided with complete information about the research project for which their donated embryos will be used and all details about what will be done with the donated embryos. Potential donors should also be informed about the potential risks and benefits of participating in the research project, including reduction of the embryos available to them for reproductive purposes, the need to potentially undergo another egg donation retrieval cycle, and possible psychological implications of donating embryos. Finally, donors should also be informed that they may withdraw their consent at any time before the embryos are thawed for use in research, and that their medical care will not be affected by their decision about donating embryos for use in research.

Legal definitions of embryos as entities distinct from the bodies of women who may gestate them also raise issues for women's reproductive decision-making abilities. There are limits on how long embryos can remain in storage in the

United Kingdom and Australia; consequently, women have been forced to make disposition decisions about their embryos before they were ready to (Millbank et al., 2012). Options for disposition include destruction, donation for others' reproductive use, and donation for research purposes. Decisions regarding disposition is a significant and frequently unresolved issue for couples with cyropreserved embryos, and very personal conceptualizations of embryos contribute to ambivalence, uncertainty, and difficulty in reaching a decision (Nachtigall et al., 2005).

In circumstances where a woman's eggs are not viable or where she does not have any, donor eggs are required for pregnancy. Countries in the West have divergent regulation on this issue. In Austria, Italy, and Germany egg donation is illegal, while in Canada and France the altruistic donation of eggs is legal. In the United States and in the United Kingdom egg donors may be compensated. However, in the United Kingdom the rate of compensation is fixed and regulated by the Human Fertilization and Embryo Authority, which regulates widely in the realm of assisted reproduction, while in the United States free market principles determine the rate of compensation for egg donation and the practice is widely unregulated. The intimate, time-consuming, and medically risky nature of egg donation has long made egg donors hard to come by. There is, however, an emerging market in human eggs in which women are paid – legally in some jurisdictions, illegally in others – to provide their eggs to intended parents. In order to avoid legal restrictions, women may be paid in one country and undergo egg retrieval in another or, alternatively, intended parents may travel long distances. The market in reproductive tissue raises ethical concerns about the potential exploitation of women who put their health at risk by donating eggs.

Surrogacy is a long-standing ART. Traditional surrogacy (in which the surrogate mother is genetically related to the baby she is carrying) is still sometimes practiced, but in combination with IVF, the use of gestational surrogates is becoming increasingly popular. Gestational surrogacy involves a surrogate mother carrying an embryo produced from the genetic material of the intended mother and father. Surrogacy may be used in cases where an intended mother has uterine difficulties (or no uterus) that prevent her from carrying a child to term, or in the case of gay couples relying on a combination of both an egg donor and a gestational surrogate (or simply a traditional surrogate).

Like egg donation, surrogacy is a controversial practice. It is illegal in many European countries including Germany, France, Spain, and Italy, while commercial surrogacy is illegal in the United Kingdom and Canada. Commercial surrogacy is, however, legal in the United States (Norton et al., 2013). In many of these jurisdictions, there is an unregulated, underground market in surrogacy, and violations of the law are rarely enforced. People seeking surrogates in the West in jurisdictions, where surrogacy is illegal or strictly regulated (or where it is very expensive), may travel to countries including India or Thailand to find commercial surrogates (Hudson et al., 2011).

The use of ARTs by couples and individuals seeking to have children has been increasing since the 1980s. However, barriers to equitable access exist in some jurisdictions due to the high cost of the procedures, formal and informal eligibility criteria, and discriminatory clinic screening practices (Gurmankin et al., 2005; Luce, 2010; Peterson, 2005). In the absence of public funding for ARTs, access is generally restricted to women and couples who have the economic means to pay for the procedures and the associated drug regimen. A number of countries, including Australia, Israel, the United Kingdom, and the province of Quebec in Canada, do publicly fund IVF and ICSI, for example, but there may be restrictions placed on the number of cycles, the age of the woman, or the number of children for which public funding is available.

Ability to pay is not the only impediment to accessing ARTs. Clinics and physicians use a number of criteria to determine eligibility that focus on the potential benefits and risks to the health and safety of the participants based on medical factors, including the participants age and cause of infertility. However, nonmedical criteria have also been used to restrict access to ARTs for persons from a number of marginalized groups, based on disability, race and ethnicity, marital status, and sexual orientation (Kissil and Davey, 2012).

Fertility Preservation

Increased use of ARTs is in many cases a response to infertility. Healthy pregnancies and reproductive lives are part of the larger picture of reproductive freedom, alongside access to abortion care and contraception. Women's fertility may be curtailed by various factors including sexually transmitted infections, age, toxic exposures, smoking, and others. Growing awareness of the impact of chemical exposures on pregnancy is leading to calls for chemical policy reform that protects women's reproductive health and fertility from toxic chemicals (ACOG, 2013; RCOG, 2013). Women from marginalized groups are likely to be substantially more affected by exposures to toxic chemicals (Sze, 2007). Environmental health concerns are an integral and sustainable part of reproductive rights. Stronger regulation of toxic chemicals is required in order to improve reproductive health outcomes.

Reproductive Choice and Prenatal Genetic Testing

Reproductive rights in the developed world have largely been grounded in the recognition that individuals are entitled to bodily autonomy and the ability to make their own reproductive choices. While earlier accounts of reproductive rights were primarily concerned with women's abilities to make safe, legal choices to avoid unwanted procreation, reproductive justice in the contemporary context has been argued to comprise a spectrum of rights that include not only the right to bear and raise children, but also the right to determine the genetic characteristics of those children.

Over the last 20 years there has been a proliferation of diagnostic technologies aimed at identifying and eliminating potential 'abnormalities' in future children (Karpin and Savell, 2012). Such technologies, and their attendant practices, raise questions about the ways in which individuals define and understand concepts of disability and normalcy. In Western culture, a medical model of disability has long dominated mainstream thinking, characterizing disability as not only objectively ascertainable, but also inherently undesirable (Karpin and Savell, 2012). Negative assumptions about

disability have particular relevance to our understanding of the social context in which reproductive choices about the use of prenatal screening and testing technologies are made. In a society where health is highly valued, the conflation of disability with ill health has increased the responsibility of women, and to a lesser extent men, to become informed about the available screening and testing and to evaluate their own, as well as their future progeny's, projected health status (Mykitiuk, 2002).

In order to determine whether a fetus is at risk for diseases involving chromosomal abnormalities, postconception screening tests (usually maternal blood tests or ultrasounds, including nuchal translucency scans) may be used. Conditions frequently tested for include aneuploidy, such as Down's Syndrome, Edwards Syndrome, and Patau Syndrome, as well as spina bifida (Karpin and Savell, 2012). However, the false positive rates of the current screening methods are approximately 5%; thus, in order to obtain a more definitive diagnosis, fetal cells must be acquired through amniocentesis or chorionic villus sampling and then analyzed (Barber, 2013). Each of these methods of prenatal diagnosis (PND) is invasive and carries the risk of miscarriage (Barber, 2013).

PGD provides an alternative to current postconception diagnostic procedures. The technique involves taking a biopsy sample (removing one or two cells) from an embryo that has developed to a minimum of four cells and conducting a genetic analysis thereafter. To date, over 350 chromosomal and genetic abnormalities can be tested for using PGD, including cystic fibrosis, hemophilia, some hereditary cancers, diabetes, and deafness (Karpin and Savell, 2012). Sex and some physical characteristics can also be selected for using PGD, raising concerns that clinicians and patients might try to use the technology to create the so-called 'designer babies.' Nevertheless, the decision to use PGD is usually prompted by a patient query regarding concern over the presence of a hereditary genetic condition in his or her family (Karpin and Savell, 2012). Because PGD is both expensive and invasive, the technology is currently only used by a small number of people worldwide and is often subject to restrictive regulatory regimes that further limit its access.

The legality of PGD testing varies in different countries. Many countries and regions including the United Kingdom, South Australia, New Zealand, and eight European countries (Greece, Spain, France, Denmark, Sweden, Norway, Germany, and the Czech Republic) stipulate that PGD may only be used in the case of 'serious' disease or disability. However, what constitutes a 'serious' disease or disability is often ambiguous in legislation and is typically left up to the determination of clinicians (Karpin and Savell, 2012). Western Australia only permits the use of PGD when it is therapeutic for a particular egg or embryo, while the Victoria region of Australia allows clinicians to determine its use on a case-by-case basis. In the United States no uniform laws exist regulating PGD testing, although health professional organizations have developed best practice guidelines regarding its use. Finally, in the United Kingdom sex selection through PGD for social or cultural reasons is prohibited, but is allowed in order to avoid sex-linked disorders.

Despite regulatory barriers to access, the number of individuals who use PND and PGD is growing. Unlike earlier eugenic practices that were coercive and operated at the level of populations, genetic testing operates at the level of individual pregnancies and is voluntary. Prenatal tests are often promoted as conveying neutral information to prospective parents to facilitate 'informed choices' about their pregnancies. These choices include deciding whether to terminate a pregnancy if testing indicates a risk for a genetic disease, taking measures (where available) during pregnancy to reduce the effects of a genetic disease, and preparing for the birth of child with a genetic disease or disability. Far from neutral, though, a woman's 'choice' cannot be severed from the discourses of risk that surround pregnancy and motherhood (Weir, 2006). Routine screening and testing create an environment in which women are pressured not only to become 'responsible' reproducers, but also to manage their pregnancies to ensure the 'best' possible opportunity for a 'healthy,' 'normal' outcome (Karpin and Savell, 2012). Consequently, there is a danger that those who challenge dominant attitudes about risk and disability may be viewed as reckless if they choose to bring a disabled child to term – complicating the idea that testing is freely chosen (Botkin, 2002; Hensel, 2005).

In this way, the growth of genetic testing increasingly makes pregnant women and their health care providers responsible for deciding whether lives are worth living. It is therefore necessary to recognize how genetic testing, as it is presently practiced, reinforces problematic views of disability as abnormal, undesirable, and as an individual medical problem. Decisions about genetic testing must involve genuine choices by prospective parents that are neither constrained by health care funding nor oversimplified understandings of particular genetic conditions. Otherwise, diagnostic technologies have the potential to hinder, rather than promote, reproductive justice.

Conclusion

Reproductive rights remain contested in developed countries. Technological advancements produce new challenges to reproductive rights. Access to contraception and abortion remain important issues in reproductive rights, along with new issues such as genetic testing and ARTs. It is increasingly recognized that reproductive rights must be understood in the context of broader social justice issues. Consequently, the organization Asian Communities for Reproductive Justice (recently renamed Forward Together), one of the original groups promoting reproductive justice, argues that three frameworks are necessary to address reproductive oppression: reproductive health, dealing with service delivery; reproductive rights, addressing legal issues; and reproductive justice, focusing on movement building (www.forwardtogether.org). Although these frameworks are distinct in their approaches, all three are necessary to uphold the reproductive self-determination of all individuals and to ensure the protection of meaningful reproductive choice for all.

See also: Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective.

Bibliography

ACOG, 2013. Exposure to Toxic Environmental Agents. The American Congress of Obstetricians and Gynecologists.

Arzuaga, B.H., Lee, B.H., 2011. Limits of human viability in the United States: a medicolegal review. Pediatrics 128, 1047–1052.

Barber, Paul L., 2013. Prenatal diagnosis: an ethical and a regulatory dilemma. Houston Journal of Health Law and Policy 13, 329–351.

Bauer, G.R., Hammond, R., Travers, R., Kaay, M., Hohenadel, K.M., Boyce, M., 2009. "I Don't think this is theoretical; this is our lives": how erasure impacts health care for transgender people. Journal of the Association of Nurses in AIDS Care 20, 348–361.

Boland, R., 2010. Second trimester abortion laws globally: actuality, trends and recommendations. Reproductive Health Matters 18, 67–89.

Borrero, S., Zhao, X., Mor, M.K., Schwarz, E.B., Good, C.B., Gellad, W.F., 2013. Adherence to hormonal contraception among women veterans: differences by race/ethnicity and contraceptive supply. American Journal of Obstetrics Gynecology 209, 103.e1–11.

Botkin, J.R., 2002. Prenatal diagnosis and the selection of children. Florida State University Law Review 30, 265.

Branch, L.S., 2007. Consolidated Federal Laws of Canada, Divorce Act [WWW Document]. http://laws-lois.justice.gc.ca/eng/acts/D-3.4/ (accessed 11.15.13.).

Brodzinsky, D.M., Pertman, A., 2011. Adoption by Lesbians and Gay Men: A New Dimension in Family Diversity. Oxford University Press, New York.

Bullough, V.L. (Ed.), 2001. Encyclopedia of Birth Control. ABC-CLIO, Santa Barbara, CA.

Castagnoli, C., 2010. Transgender Persons' Rights in the EU Member States. Directorate General for Internal Policies, European Parliament.

Caulfield, T., Robertson, G., 1996. Eugenic policies in Alberta: from the systematic to the systemic? Alberta Law Review 35, 59–79.

Cook, R.J., Dickens, B.M., Fathalla, M.F., 2003. Reproductive Health and Human Rights: Integrating Medicine, Ethics, and Law: Integrating Medicine, Ethics, and Law. Oxford University Press.

Corbett, S.L., Frecker, H.M., Shapiro, H.M., Yudin, M.H., 2013. Access to fertility services for lesbian women in Canada. Fertility Sterility 100, 1077–1080.

Couldrick, L., Sadlo, G., Cross, V., 2010. Proposing a new sexual health model of practice for disability teams: the recognition model. International Journal of Therapy and Rehabilitation 17, 290–299.

Creating Our Families: A Pilot Study of the Experiences of Lesbian, Gay, Bisexual and Trans People Accessing Assisted Human Reproduction Services in Ontario [WWW Document], 2010. Researching for LGBTQ Health. http://lgbtqhealth.ca/projects/creatingourfamilies.php (accessed 11.20.13.).

Daniels, P., Campbell, P., Clinton, A., 2013. The current state of abortion law and practice in Northern Ireland. British Journal Nursing 22, 326–328.

Dempsey, D., 2010. Conceiving and negotiating reproductive relationships: lesbians and gay men forming families with children. Sociology 44, 1145–1162.

Dobson, 1999. (Litigation Guardian of) V. Dobson.

Dowd, N., 2013. Sperm, testosterone, masculinities and fatherhood. Nevada Law Journal 13, 438–457.

Downie, J., Llewellyn, J.J., 2011. Being Relational: Reflections on Relational Theory and Health Law. UBC Press, Vancouver, BC.

Esmail, S., Darry, K., Walter, A., Knupp, H., 2010. Attitudes and perceptions towards disability and sexuality. Disability and Rehabilitation 32, 1148–1155.

European Union Agency for Fundamental Rights, 2010. Homophobia, Transphobia and Discrimination on Grounds of Sexual Orientation and Gender Identity. Luxembourg.

Finer, L., Fine, J.B., 2013. Abortion law around the world: progress and pushback. American Journal of Public Health 103, 585–589.

Fried, M.G. (Ed.), 1990. From Abortion to Reproductive Freedom: Transforming a Movement. South End Press, Boston.

Gibson, B.E., Mykitiuk, R., 2012. Health care access and support for disabled women in Canada: falling short of the UN convention on the rights of persons with disabilities: a qualitative study. Womens Health Issue. Official Publication Jacobs Institue of Women's Health 22, e111–118.

Gibson, J., O'Connor, R., 2010. Access to health care for disabled people: a systematic review. Social Care and Neurodisability 1, 21–31.

Greenfeld, D.A., Seli, E., 2013. Assisted reproduction in same sex couples. In: Sauer, M.V. (Ed.), Principles of Oocyte and Embryo Donation. Springer, New York, pp. 289–302.

Grekul, J., Krahn, A., Odynak, D., 2004. Sterilizing the "Feeble-minded": eugenics in Alberta, Canada, 1929–1972. Journal of Historical Sociology 17, 358–384.

Gurmankin, A.D., Caplan, A.L., Braverman, A.M., 2005. Screening practices and beliefs of assisted reproductive technology programs. Fertility and Sterility 83, 61–67.

Haraway, D.J., 2013. Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York.

Harris-Short, S., 2009. Regulating reproduction: frozen embryos, consent, welfare and the equality myth. In: Smith, S.W., Deazley, R. (Eds.), The Legal, Medical and Cultural Regulation of the Body - Transformation and Transgression. Ashgate, London, pp. 47–76.

Hensel, W., 2005. The disabling impact of wrongful birth and wrongful life actions. Harvard Civil Rights-Civil Liberties Law Review 40, 141.

Hudson, N., Culley, L., Blyth, E., Norton, W., Rapport, F., Pacey, A., 2011. Cross-border reproductive care: a review of the literature. Reproductive Biomedicine Online 22, 673–685.

Isler, A., Tas, F., Beytut, D., Conk, Z., 2009. Sexuality in adolescents with intellectual disabilities. Sexuality and Disability 27, 27–34.

James-Abra, S., 2012. Access to Assisted Human Reproduction (AHR) Services for Trans People in Ontario (Thesis). University of Toronto, Toronto.

Karaian, L., 2013. Pregnant men: repronormativity, critical trans theory and the re(conceive)ing of sex and pregnancy in law. Social & Legal Studies 22 (2), 211–230. http://dx.doi.org/10.1177/0964663912474862.

Karpin, I., Savell, K., 2012. Perfecting Pregnancy: Law, Disability, and the Future of Reproduction. Cambridge University Press, New York.

Kissil, K., Davey, M., 2012. Health disparities in procreation: unequal access to assisted reproductive technologies. Journal of Feminist Family Therapy 24, 197–212.

Lafferty, A., McConkey, R., Simpson, A., 2012. Reducing the barriers to relationships and sexuality education for persons with intellectual disabilities. Journal of Intellectual Disabilities 16, 29–43.

Lewin, E., 2006. Family values: gay men and adoption in America. In: Wegar, K. (Ed.), Adoptive Families in a Diverse Society. Rutgers University Press, New Brunswick, NJ, pp. 129–145.

Luce, J., 2010. Beyond Expectation: Lesbian/bi/queer Women and Assisted Conception. University of Toronto Press.

McLeod, C., 2010. Harm or mere inconvenience? Denying women emergency contraception. Hypatia 25, 11–30.

Meyer, A.S., McWey, L.M., McKendrick, W., Henderson, T.L., 2010. Substance using parents, foster care, and termination of parental rights: the importance of risk factors for legal outcomes. Children and Youth Service Review 32, 639–649.

Millbank, J., Stuhmcke, A., Karpin, I., Chandler, E., 2012. Enhancing Reproductive Opportunity: A Study of Decision-making Concerning Stored Embryos. UTS, Sydney.

Mitchell, V., Green, R.-J., 2007. Different storks for different folks. Journal of GLBT Family Studies 3, 81–104.

Mykitiuk, R., 2002. Public bodies, private parts: genetics in a post-Keynesian Era. In: Cossman, B., Fudge, J. (Eds.), Privatization, Law, and the Challenge to Feminism. University of Toronto Press, Toronto, pp. 311–354.

Nachtigall, R.D., Becker, G., Friese, C., Butler, A., MacDougall, K., 2005. Parents' conceptualization of their frozen embryos complicates the disposition decision. Fertility and Sterility 84, 431–434.

Nedelsky, J., 2011. Law's Relations: A Relational Theory of Self, Autonomy, and Law. Oxford University Press, Oxford, New York.

Nelson, E., Mykitiuk, R., Nisker, J., et al., 2008. Informed consent to donate embryos for research purposes: SOGC clinical practice guideline. Journal of Obstetrics and Gynaecology Canada 215, 824–829.

Norton, W., Hudson, N., Culley, L., 2013. Gay men seeking surrogacy to achieve parenthood. Reproductive Biomedicine Online 27, 271–279.

Ordolis, E., 2008. Maternal substance abuse and the limits of law: a relational challenge. Alberta Law Review 46, 119.

Padavic, I., Butterfield, J., 2011. Mothers, fathers, and "mathers": negotiating a lesbian co-parental identity. Gender & Society 25 (2), 176–196.

Peterson, M.M., 2005. Assisted reproductive technologies and equity of access issues. Journal of Medical Ethics 31, 280–285.

Quinn, C., Happell, B., Browne, G., 2011. Talking or avoiding? Mental health nurses' views about discussing sexual health with consumers. International Journal of Mental Health Nursing 20, 21–28.

Rank, N., 2010. Barriers for access to assisted reproductive technologies by lesbian women: the search for parity within the healthcare system. Houston Journal of Health Law and Policy 10, 115.

RCOG, 2013. Chemical exposures during pregnancy (Scientific impact paper 37). Royal College of Obstetricians and Gynaecologists.

Roberts, D.E., 1991. Punishing drug addicts who have babies: women of color, equality and the right to privacy. Harverd Law Review 104, 1419–1482.

Roberts, D.E., 1997. Killing the Black Body: Race, Reproduction, and the Meaning of Liberty. Vintage books, New York.

Rodgers, S., 2007. Women's reproductive equality and the supreme court of Canada. In: Downie, J., Gibson, E. (Eds.), Health Law at the Supreme Court of Canada. Irwin Law, Toronto, pp. 189–218.
Rodgers, S., Downie, J., 2006. Abortion: ensuring access. Canadian Medical Association Journal 175, 9.
Ryan, M., 2009. Beyond Thomas Beatie: trans men and the new parenthood. In: Epstein, R. (Ed.), Who's Your Daddy?: And Other Writings on Queer Parenting. Sumach Press, Toronto, pp. 139–150.
Scott, E., Emery, R., 2013. Gender politics and child custody: the puzzling persistence of the best interest standard. Columbia Public Law & Legal Theory Working Papers. Retrieved from http://lsr.nellco.org/columbia_pllt/9200.
Shildrick, M., 1997. Leaky Bodies and Boundaries: Feminism, Postmodernism and (Bio)ethics. Routledge, New York.
Silvers, A., 1998. Reprising women's disability: feminist identity strategy and disability rights. Berkeley Women's Law Journal 13, 81–116.
Smith, A., 2005. Beyond pro-choice versus pro-life: women of color and reproductive justice. National Women's Studies Association Journal 17, 119–140.
Sze, J., 2007. Noxious New York: The Racial Politics of Urban Health and Environmental Justice. MIT Press, New York.
United Nations, 1945. Charter of the United Nations [WWW Document]. http://www.un.org/en/documents/charter/ (accessed 11.11.13.).
United Nations General Assembly, 1948. Universal Declaration of Human Rights.
United Nations General Assembly, 1974. Declaration on Social Progress and Development.
United Nations General Assembly, 1979. Convention on the Elimination of All Forms of Discrimination against Women.
Washenfelder, C., 2004. Regulating a revolution: the extent of reproductive rights in Canada. Health Law Review 12, 44–52.
Weir, L., 2006. Pregnancy, Risk and Biopolitics: On the Threshold of the Living Subject. Routledge, New York.
West, R., 2008. From choice to reproductive justice: de-constitutionalizing abortion rights. Yale Law Journal 118, 1394–1433.
Winnipeg Child and Family Servisssces, 1997. (Northwest Area) v. D.F.G.

Relevant Websites

http://www.bionews.org.uk/home – BioNews.
http://reproductiverights.org/ – Center for Reproductive Rights.
http://www.astra.org.pl/ – Central and Eastern European Women's Network for Sexual and Reproductive Rights and Health.
http://lgbtqhealth.ca/projects/creatingourfamilies.php – Creating Our Families: A pilot study of the experiences of lesbian, gay, bisexual and trans people accessing assisted human reproduction services in Ontario.
http://www.geneticsandsociety.rsvp1.com/index.php?mgh=http%3A%2F%2Fwww.geneticsandsociety.org&mgf=1 – Genetics and Society.
http://www.ijfab.org/blog/ – International Journal for Feminist Approaches to Bioethics blog.
http://libraryeuroparl.wordpress.com/2013/01/19/sexual-and-reproductive-health-and-rights/ – Library of the European Parliament.
http://www.protectchoice.org/ – The Pro-Choice Public Education Project.
http://www.reproductivechoiceaustralia.org.au/ – Reproductive Choice Australia.
http://www.sistersong.net/ – Sistersong: Women of Colour Reproductive Justice Collective.
http://www.unfpa.org/rights/rights.htm – United Nations Population Fund: Reproductive Rights.

Republicanism: Philosophical Aspects

Mortimer NS Sellers, University System of Maryland, MD, USA; and University of Baltimore Center for International and Comparative Law, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Republicanism is the doctrine that public power should always serve the common good of all those subject to its rule. This raises the question how to do so most effectively, either through particular policies or through constitutional structure ('the republican form of government'). The republican philosophical tradition began with Plato and Aristotle, flowered in the writings of Marcus Tullius Cicero, and reappeared with the revival of learning in such authors as Niccolò Machiavelli, James Harrington, John Adams, and Immanuel Kant. More recently Philip Pettit, Jürgen Habermas, and others have returned to the republican conception of liberty as nondomination, and how to secure this through the rule of law, popular sovereignty, and the checks and balances of well-designed deliberative politics. Republicanism seeks freedom and justice through law and government in pursuit of the common good.

'Republicanism' as a philosophical doctrine begins with the axiom that law and government should always serve the '*res publica*' or common good of the people. 'People' in this context signifies all citizens or (more recently) all participants in the relevant community. The *locus classicus* and most famous manifesto of republican thought is in the philosophical writings of Marcus Tullius Cicero (106–43 BC), above all *de officiis*, *de legibus*, and *de re publica*, in which he defined a republic as the property and project of the '*populus*', when the people come together in pursuit of justice and the common good (*de re publica*, xxxv.39). This standard definition of a republican, repeated by St. Augustine (354–430) (*de civitate Dei* XIX.xxi), John Adams (1735–1826) (1787: p. I.xviii), and many others should be given in full and in Latin, because it remains the foundation and persistent common thread in all republic thought: "*res publica res [est] populi, populus autem non omnis hominum coetus quoquo modo congregatus, sed coetus multitudinis juris consensu et utilitatis communione sociatus*". Which is to say: "a republic is the property of the people – not just of any group of persons jumbled together – but of a *people*, properly speaking united in pursuit of justice and the common good of society as a whole.

While Cicero and the Latin language inaugurated the philosophical discourse of republicanism, Cicero himself (*de officiis*, I.xxv.85) attributed the insight to Plato (423–347 BC), who had argued in his *Politeia* that rulers exist to serve the welfare of all citizens alike (*Politeia*, I.xv. 342 E). Aristotle (384–322 BC) made the same point when he equated justice with the common good (*Politica*, III.iv.7; VII.ii.10), and this became the dominant conception of government, law, and justice through most of European history (Sellers, 1998). Cicero wrote his dialogues *de re publica* and *de legibus* to celebrate and modernize Plato's work on the state and the laws (*de legibus*, II.v.14), and the practical aspects of republican philosophy have remained dominant ever since. It is not enough simply to declare the abstract purpose of serving the common good; one must also propose practical institutions to realize the common good in fact. This 'republican form of government' became the primary object of subsequent republican discourse (Adams, 1776).

Discussion of republican principles and the republican form of government has had a revolutionary impact wherever it has taken hold. By asking 'the great question', "what form of government will compel the formation of good and equal laws, an impartial execution, and faithful interpretation of them, so that citizens may constantly enjoy the benefit of them, and be sure of their continuance" (Adams, 1788), philosophers, lawyers, and statesmen have shaken the governments of Italy, the Netherlands, England, North America, France, and many other nations with demands for 'liberty', 'the rule of law', 'popular sovereignty', 'checks and balances', and other basic requirements of republican government, as established already by Cicero in Rome (Sellers, 1998). Subsequent republicans would add representation, the separation of powers, and other new structural devices to protect public liberty ('*libertas*') and avoid Rome's eventual descent into popular tyranny and military despotism (Hamilton et al., 1787). Contemporary philosophers such as Philip Pettit (1945–) (Pettit, 1997) and Jürgen Habermas (1929–) (Habermas, 1992) have renewed this tradition, with further proposals directed at protecting people everywhere against arbitrary power.

The Republican Tradition

'*Res publica*' was the Romans' term for their state, its public business, all public property, and the purposes these served. The word notoriously evades translation, most often appearing in English as 'commonwealth', or simply (more recently) 'republic', the use preferred here. Cicero and Titus Livius (Livy) (59 BC–AD 17) constructed the first and most influential comprehensively republican ideology, to try to explain how and why the Roman republic had failed to serve the public good. Both agreed that republican institutions collapsed when military expansion and party conflict upset traditional checks and balances between the senate, the magistrates, and the people of Rome.

Cicero and Livy inaugurated a republican tradition of 'liberty' that fortified principled resistance to demagogues, emperors, and kings for the next 2000 years. Niccolò Machiavelli (1469–1527) did the most to revive this republican tradition in Italy, in his *Discorsi sopra la prima deca di Tito Livio* (1517).

The resistance of the Swiss, various Italian cantons, and the United Provinces of the Netherlands to imperial control added practical models for republican liberty, as did the constitutional and theoretical writings of various English authors, in their efforts to restrain or remove their kings during the Civil War and Commonwealth (1642–60), the Glorious Revolution (1688), and the extended British controversies over American Independence (1763–83).

The Republican Revival

Republican political and legal philosophy provides an enduring model for philosophers and lawyers who oppose the domination of arbitrary power. The constitutional controversies of the French and American revolutions left institutional relics, which established republican ideals in the structures of Western politics, with a pervasive commitment to 'liberty' and even (in the United States), a federal 'guarantee' that every state in the Union must enjoy 'a republican form of government' (*United States Constitution*, 1787: Article IV, Section 4). Similar provisions requiring senates and independent judges give republican theorists a practical foundation in the constitutions of most Western democracies. Lawyers such as Frank Michelman (1936–) and Cass Sunstein (1954–) revived republican doctrine in law schools, to justify judicial intervention for the common good, against partisan legislation and political corruption (*Symposium*, 1988).

Philip Pettit inaugurated a similar revival among philosophers, embracing three central claims of the republican tradition: that government must serve the common good; that certain institutional arrangements will do so; and that liberty will be the result (Pettit, 1997). Republicanism, since Cicero has offered a political epistemology for finding the common good, or approximating it as closely as possible. The republican question has not been (in the first instance): 'What is the common good?'; but rather: "Which political procedures will best find and protect the common good, given the possibilities and limitations of history and human nature?" (Sellers, 1991).

Liberty as Nondomination

Republican liberty consists in not being interfered with by other persons or the state, except as regulated by the common good of the people, established by republican politics (Harrington, 1656). This conception of liberty as 'nondomination' contradicts a widely held (but more recent) understanding of liberty as not being interfered with at all. Public and private actions that constrain or influence a person's activities do not restrict republican liberty unless they violate the law or republican principles established to serve the common good of the people as a whole. Republican liberty cannot exist without a republican state because nonrepublics will always dominate their subjects, as masters dominate their slaves.

Domination marks the end of liberty, in republican theory, because domination subjects one person to the arbitrary will of another, without regard for the common good. The unchecked power of the *dominus* (or master) makes his subordinates slaves, whether the master chooses to exercise power or not. Slaves to good masters remain slaves, and dominated, because they have no security against their master's will. States dominate citizens when constitutions give legislatures or public officials the arbitrary power to act without regard for the public welfare. Private individuals dominate others when laws and public officials do not protect citizens against private oppression.

The common good is the measure of domination, and therefore of liberty, in republican political vocabulary, because only the common good distinguishes 'arbitrary' interference from justified state action. *Libertas* (liberty) in Rome signified the status of a free citizen in a republican state. The *res publica* (republic) represented every citizen's common interest in the public good. Restraining citizens equally with just laws in pursuit of the common good enhances their liberty by restraining arbitrary power. Neorepublican doctrine revives liberty as a worthy object of public policy by rescuing the sense in which the word first developed as a political ideal. Liberty is not, as Thomas Hobbes (1588–1679) tendentiously redefined it, the unfettered license to do what one wants, but rather a public shield against uncontrolled license in others. Republican liberty consists in equal subjection to the rule of laws, made for the common good, through the carefully balanced mechanism of a republican form of government. For further references *see* Freedom: Political.

The Common Good

The first difficulty republicans face in finding the common good lies in defining the relevant community. Ancient republics closely restricted their citizenship, excluding women, immigrants, and slaves. Most republicans would now extend citizenship and membership in the *populus* or people to all inhabitants of a given territory. This leaves open the question of boundaries. Cicero endorsed a universal republic or society of all humanity, but also recognized the value of local politics. Cultures and human relationships develop and solidify locally. Therefore, the republican interest in harmony and community implies small homogenous republics, where people will have more in common (Montesquieu, 1748). But the natural diversity of human talents and interests guarantees that small communities will find their own internal minorities to dominate and oppress. Larger republics with greater diversity better protect minorities, by making dominant factions harder to assemble (Hamilton et al., 1787: p. x).

The republican solution to the problem of republican boundaries has been to encourage different levels of nested republics, for different purposes, with different capacities. Immanuel Kant (1724–1804) advocated a federation of national republics, assembled in perpetual peace (Kant, 1795). The formula suggested by the United States and modern European Union has been to entrust the protection of individual human rights and commerce, which local majorities might threaten, to federal suprarepublican authorities, while delegating social and cultural concerns to smaller, more homogenous republics. When regional authorities consistently abuse their powers to oppress local minorities, their oppression implicitly recognizes such minorities as separate people, entitled to separate republican homelands, within the republican federation (Sellers, 2006). For further references *see* Public Interest.

Popular Sovereignty

Republicans realized very early in Rome, and have maintained the principle ever since, that the common good will never be found or maintained without the *imperium populi* or popular sovereignty to support it (Tullius Cicero *Cat*. IV.14; *Phil*. IV.iv.8). Machiavelli credited Cicero with the insight that although the people may be ignorant, they are capable of grasping truth, when good men place the truth before them (Machiavelli, 1517: I.iv.10). The purpose of republican popular sovereignty is not that the people should govern everyday, but rather, as Benjamin Rush (1745–1813) suggested in advocating an American 'republic', that the people should select their rulers themselves. The people exercise their power on election days, then defer to magistrates that they themselves have selected (Rush, 1777).

This republican reliance on popular sovereignty distinguishes republican government from liberalism, which diverged from the republican tradition at the beginning of the nineteenth century (Sellers, 1998). Early liberals rejected popular sovereignty in the wake of the French revolution, while retaining republican commitments to liberty, the rule of law, and certain specific rights against the government (Constant, 1819). Liberals divorced liberty from its basis in public deliberation by distinguishing 'political liberty' from personal independence. This left liberals dependent on judges and magistrates to verify their assertions of rights, and protect individual liberty against the state. Liberalism started as a flight from politics, but found that this implied a new definition of liberty as the ability to do what one wants, without state regulation (Pettit, 1997).

Contemporary liberalism has largely abandoned the fear of democracy that separated liberalism from republicanism in the first place, and most liberals now endorse the widespread political participation of citizens in elections (Rawls, 1993). Republican popular sovereignty never signified democracy in any case, but rather a carefully balanced and controlled mechanism for gathering the insights of all members of society, through a process of public deliberation, in pursuit of the common good. As modern liberals have sought a method of democratic deliberation to legitimate their sense of community, and regulate conflicting private interests, they have increasingly returned to the republican principles and institutions already embedded in Western constitutional democracy.

Checks and Balances

The primary, process-centered challenge for republicans has been to identify those institutions that find and secure the common good best. Popular sovereignty provides access to every citizen's interests and insights, but democracy would become an elective despotism, without checks and balances to restrain it (Pettit, 1999). Cicero blamed the Greeks' misfortunes on the turbulence of their popular assemblies, lacking internal balance or a senate to control them (Tullius Cicero *Flacc*.). Republics since Rome have maintained bicameral legislatures to prevent self-seeking in either assembly (Harrington, 1656: p. 22). Democracy is one of several foundation stones of liberty, not an end in itself.

The necessary checks and balances of republican government prevent public officials from making themselves and not the *res publica* the object the state (Paine, 1792: p. 168). The dispersion of power through bicameralism, federalism, and the separation of powers makes it harder for any one person or faction, including the majority, to wield arbitrary power over others (Pettit, 1997: p. 177). American republicans such as John Adams (1787–8: p. I.ii–iii) and Alexander Hamilton et al. (1787: IX) added representation and the life tenure of judges to bicameralism and checks on governmental powers as basic desiderata of balanced republican institutions. If the people and judges are outside government, they can better control their government's mistakes (Hamilton et al., 1787: p. LXIII).

Modern republicans such as Philip Pettit differ from democrats in viewing democracy as a derivative value, in service to the broader ideal of balanced or 'contested' government (Pettit, 1997: p. 187). All government decisions should be subject to challenge by institutions that prevent private inclinations from ruling public interests and ideas. Republicans have proposed bills of rights, public hearings, and many other devices designed to constrain and channel public decision making, so that ordinary citizens may influence their government's decisions, without diverting the public purposes of the state. For further references *see* Constitutionalism.

The Rule of Law

The '*imperia legum*' or 'rule of laws and not of men' protects republican liberty by forestalling arbitrary power. Republican authors from Livy to Pettit have protected the moral element in the republican concept of 'law' against the simplifications of legal positivists, beginning with Thomas Hobbes. Laws secure liberty against domination by delimiting and protecting the line across which private or state behavior violates individual autonomy against the common good. Cicero insisted that such laws must serve the public welfare (*populi utilitas*) not the public will (*populi voluntas*) (Tullius Cicero *Sulla* 25) because votes cannot alter the natural laws of justice (Tullius Cicero *Leg*. I.xvi. 44). Republics require institutions that will find justice by securing the common good.

The empire-of-law condition of republican government entails that laws should be publicly promulgated, intelligible, consistent, stable, general, and apply to everyone, including the legislators themselves. This prevents arbitrary decisions by public officials, by subjecting their will to known constraints and purposes. Republics maintain a general presumption that government action, when needed, will operate by law, not by ad hoc or *ex post* decisions. All agencies of government must act through principled, regulated structures, maintaining the 'due process' of laws to prevent the abuse of governmental power (Pettit, 1997: pp. 174–175).

Codifying the law in every detail will not be possible or desirable. Some discretion must remain, but subject to the dispersion-of-power and popular sovereignty procedures that generate republican laws in the first place. Discretion subject to constraints against arbitrariness may sometimes secure the common good, but unlimited discretion produces mistakes about justice, through the natural partiality and limited viewpoint of any individual decision maker. The republican rule of law protects liberty by respecting the welfare of every member of society, pooling their insights to secure the common good of

the people. Legislation is the public reason of the republic, protecting citizens' liberty against private power (Harrington, 1656: pp. 19–20).

Neorepublican Ideals

Contemporary philosophical republicanism reasserts the republican conception of liberty, proposing new republican institutions to support republican liberty in the modern multicultural state. Philip Pettit has given the most detailed proposal for neorepublican politics, supplanting the Benthamite liberal opposition of 'negative' and 'positive' liberty with older republican conceptions of liberty as nondomination, and protection against arbitrary power (Pettit, 1997). Pettit understands arbitrary power ('*arbitrium*') to include any action against the 'interests and ideas' of those who suffer interference. This surpasses the old republican standard, which measured arbitrary power by its violation of an objective public, common, or collective good.

Neorepublican deference to private interests and ideas reflects the residual liberal commitment to moral pluralism and cultural diversity. This enduring hesitation to overrule anyone's personal ideas makes the revised standard of nondomination very strong. Interference or influence over others' choices becomes domination unless it tracks their private values. This introduces a subjective element into republican doctrine. Older republican ideals tolerated hierarchy and private influences when toleration served the common good of all citizens.

Neorepublican rhetoric describes the state as seeking to *promote* freedom as nondomination, where older republicans would have understood republican pursuit of the common good simply to *be* freedom as nondomination. The difference lies in the pluralistic language of the newer theories. Some suggest that immigrant groups and minority cultures will maintain their separate identities in republics, because republics help to preserve hereditary differences within the larger culture (Pettit, 1997: p. 144). This represents a significant departure from traditional doctrine. Republics exist precisely to create and maintain an overarching culture that embraces all citizens. This necessarily influences and discourages dissenting subcultures by promoting an image of national unity, which will tend to assimilate minority perspectives.

Neorepublican Institutions

The neorepublican commitment to pluralism encourages a greater emphasis on counter-majoritarian aspects of republican popular sovereignty (Pettit, 1997; *Symposium*, 1988). Neorepublicans embrace the empire-of-law condition of republican government and the dispersion-of-power constraint, but they often seek to displace public decision making onto independent panels of social scientists or courts (Pettit, 1997; Sunstein, 1993; *Symposium*, 1988). This favors 'contestability' over consensus to limit the weight of unified public opinion.

Neorepublicans do not endorse interest-group pluralism in the classical amoral sense of liberal democracy, but their concept of representation favors cultural minorities. Some believe that public policy should be able to be justified according to the lights of every member of society (Pettit, 1997: p. 169). This overstates the value of mistaken perceptions of the good. Republican government differs from liberal democracy precisely in that it offers a deliberative technique for overruling mistaken views, while allowing constructive dissent by those who oppose the consensus.

Traditional republics maintained geographical representation to encourage the development of local communities. Most recent proposals prefer religious, cultural, or sexual diversity. Popular elections now seem too crude to identify administrative competence, or to guarantee minimum statistical representation for every stakeholder grouping (Pettit, 1997: p. 192). Neorepublicans often propose deliberation among statistically selected 'representative' officials, in place of elected representatives or public assemblies. This leads to difficulties in identifying which subcultures or stakeholder 'groups' to represent. Most people have many such partial or self-defined sources of identity.

The People

The fundamental republican conception of the common good encourages social solidarity by seeking a community of interest among the *populus* or people of any given republic. Cicero described republics as the property of a people (*res populi*), in the most influential and frequently repeated passage in the republican canon. A *populus* is not just any collection of humans, but a partnership about justice, in pursuit of the common good (Tullius Cicero *Rep*. I.xxv.39). There will be no republic without people and no people (in the fullest sense) without republic, which is to say the mutual recognition of common citizenship (fraternity) and the equal importance (equality) of all citizens in the eyes of the state.

The concept of peoples represents the most contested element in contemporary republican doctrine, and the most likely source of disagreement between neorepublican philosophers. The United Nations Charter recognized the republican principle of self-determination of peoples (Article I (2)), implying a people's right 'freely (to) determine their political status' (International Covenant on Civil and Political Rights 1966, Article 1). This gives the republican identity a practical importance in international politics, implying that recognized status as 'people' entail some degree of political separation. The salient examples of liberated 'peoples' have been the inhabitants of colonial territories, who constitute peoples in international law for the purpose of pursuing their independence from imperial domination.

Some neorepublicans would separate the national and territorial principles to create politically separate republican peoples on the basis of ethnic, religious, or linguistic affinities within larger multinational empires. Different laws would apply to different subjects of the overarching state, according to their different internal 'citizenship' or status. This violates the republican principle of equal citizenship to achieve the republican ambition of social solidarity among state-maintained subgroups of the population. Traditional republican doctrine would insist that all citizens enjoy the same privileges and immunities in their public capacities, while permitting private initiatives to pursue elective affinities. This protects (but may subtly discourage) private diversity, because the people, as public citizens, will be the same under law.

Republican Philosophy

Republican philosophy since Cicero has sought to construct a harmony of interests and common sense of justice among citizens through the 'empire of laws and not of men'. Republican laws draw the line between liberty and license, in pursuit of the common good. Republican theory seeks to find and establish good laws, by discovering the principles and basic structure that serve the *res publica* best. So although republican philosophy begins, as Thomas Paine (1737–1809) put it, by making the '*res publica*, the public affairs, or the public good' the object of all governments, and 'republican government is no other than government established and conducted for the interest of the public' (Paine, 1792: p. 168), the idea of the republic entails a constellation of political structures to secure republican legislation, embedded in 2000 years of republican tradition, derived from Rome.

The republican commitment to political institutions that protect liberty challenges the later 'liberal' flight from politics after the French revolution. Former republicans such as Benjamin Constant (1767–1830) hoped by securing certain rights or liberties to protect their own private liberty against the state. But republican philosophy recognizes the futility of 'rights' without power. Republican doctrine offers a political epistemology to discover and protect public justice and fundamental human rights. Republicans believe that without popular sovereignty, the rule of law, deliberative senators, elected executives, independent judges, a representative popular assembly, and proper checks and balances, the people cannot know or enjoy their rights and duties to each other, or to the state. The basic desiderata of republican government may appear more (or less) important in different contexts, so that even a monarchy could seem nearly 'republican' when limited by two independent branches in the legislature, and subject to the rule of law (Adams, 1787–8: pp. I. xxi–xxii). Contemporary republican philosophers adapt republican principles to new circumstances to protect public liberty against improper domination ('*dominium*') by private interests, or the arbitrary government ('*imperium*') of the state (Pettit, 1997). What makes such theories republican is their fundamental commitment to the common good, and to the political structures that support the *res publica* best. Republican philosophy is a theory of freedom and government, or rather of freedom through government, to secure a shared sense of justice, in pursuit of the common good. Republicans believe that there can be no justice without community, no liberty without the law.

See also: Justice: Philosophical Aspects.

Bibliography

Adams, J., 1776. Thoughts on Government [1983. In: Hyneman, C.S., Lutz, D.S. (Eds.), American Political Writing During the Founding Era 1760–1805, 2 vols. Liberty Press, Indianapolis, IN].

Adams, J., 1787–8. A Defence of the Constitutions of Government of the United States of America, 3 vols. C. Dilly, London.

Barnes, J., Griffin, M., 1989. Philosophia Togata I: Essays on Philosophy and Roman Society. Clarendon Press, Oxford, UK.

Barnes, J., Griffin, M., 1997. Philosophia Togata II: Plato and Aristotle at Rome. Clarendon Press, Oxford, UK.

Baron, H., 1966. The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in the Age of Classicism and Tyranny. Princeton University Press, Princeton, NJ.

Bleicken, J., 1972. Staatliche Ordnung und Freiheit in der römischen Republik. Lassleben, Kallmunz, Germany.

Bock, G., Skinner, Q., Viroli, M. (Eds.), 1990. Machiavelli and Republicanism. Cambridge University Press, Cambridge, UK.

Constant, B., 1819/1988. De la liberté des anciens comparée à celle des modernes. In: Fontana, B. (Ed.), Political Writings. Cambridge University Press, Cambridge, UK.

Cornelius, T., 1906. In: Fisher, C.D. (Ed.), Ab Excessu Divi Augusti Annalium Libri. Oxford University Press, Oxford, UK [1974. The Annals. Penguin, Harmondsworth, UK].

Cornelius, T., 1911. In: Fisher, C.D. (Ed.), Historiarum Libri. Oxford University Press, Oxford [1964. The Histories. Penguin, Harmonsworth, UK].

Dagger, R., 1997. Civic Virtues: Rights, Citizenship, and Republican Liberalism. Oxford University Press, Oxford, UK.

Fink, Z.S., 1962. The Classical Republicans: An Essay in the Recovery of a Pattern of Thought in Seventeenth-Century England, second ed. Northwestern University Press, Evanston, IL.

Fisher, H.A.L., 1911. The Republican Tradition in Europe. G. P. Putnam's Sons, New York.

Fontana, B. (Ed.), 1994. The Invention of the Modern Republic. Cambridge University Press, Cambridge, UK.

Gordon, T., 1728–31. The Works of Tacitus with Political Discourses Upon that Author, 2 vols. Woodward and Peele, London.

Gordon, T., 1746. The Works of Sallust Translated into English with Political Discourses upon that Author. Woodward and Peele, London.

Habermas, J., 1992. Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats. Suhrkamp, Frankfurt, Germany [1996. Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy. MIT Press, Cambridge, MA].

Hamilton, A., Jay, J., Madison, J., ['Publius'] 1787. The Federalist. [1987. In: Kramnick, I. (Ed.), The Federalist Papers. Penguin, Harmondsworth, UK].

Harrington, J., 1656. [1992. In: Pocock, J.G.A. (Ed.), The Commonwealth of Oceana. Cambridge University Press, Cambridge, UK].

Hobbes, T., 1651. Leviathan, or the Matter, Form, and Power of a Commonwealth Ecclesiasticall and Civil [1991. Tuck, R. (Ed.), Cambridge University Press, Cambridge, UK].

Honohan, Iseult, 2002. Civic Republicanism. Routledge, London, UK.

Kant, I., 1795. Zum Ewigen Frieden. F. Nicolovius, Königsberg.

Koenigsberger, H.G. (Ed.), 1988. Republiken und Republikanismus im Europa der frühen Neuzeit. R. Oldenbourg, Munich, Germany.

Laborde, Cecile, Maynor, John, 2008. Republicanism and Political Theory. Blackwell, Malden, MA.

Livius, T., 1919. Ab urbe condita VI–X. In: Walters, C.F., Conway, R.S. (Eds.). Oxford University Press, Oxford, UK [1982. Rome and Italy. Penguin, Harmondsworth, UK].

Livius, T., 1974. Ab urbe condita I–V. In: Ogilvie, R.M. (Ed.). Oxford University Press, Oxford, UK [1971. The Early History of Rome. Penguin, Harmondsworth, UK].

Machiavelli, N., 1517. Discorsi sopra la prima deca di Tito Livio. [1984. In: Inglese, G. (Ed.), Discorsi Rizzoli Editore, Milano] [The Discourses. Penguin, Harmondsworth, UK].

Montesquieu C. de Secondat. baron de la Brède et ce, 1734. Considérations sur les causes de la grandeur des Romains et de leur décadence. [1967. Truc, G. (Ed.), Editions Garnier, Paris] [1968. Considerations on the Causes of the Greatness of the Romans and Their Decline. Cornell University Press, Ithaca, NY].

Montesquieu C. de Secondat. baron de la brède et ce, 1748. De l'Esprit des lois. [1973. Derathé, R. (Ed.), 2 vols. Editions Garnier, Paris] [1989. The Spirit of the Laws. Cambridge University Press, Cambridge, UK].

Nicolet, C., 1982. L'idée républicaine en France (1789–1924): Essai d'histoire critique. Gallimard, Paris.

Onuf, N.G., 1998. The Republican Legacy in International Thought. Cambridge University Press, Cambridge, UK.

Paine, T., 1792. The Rights of Man, Part II [1989. In: Kuklick, B. (Ed.), Thomas Paine: Political Writings. Cambridge University Press, Cambridge, UK].

Pettit, P., 1997. Republicanism: A Theory of Freedom and Government. Oxford University Press, Oxford, UK.

Pettit, P., 1999. Postscript to the Paperback Edition of Republicanism: A Theory of Freedom and Government. Oxford University Press, Oxford, UK.

Plutarchos Bioi paralleloi. [1949–1971. In: Lindskog, C., Ziegler, K. (Eds.), Plutarchi Vitae Parallelae. Teubner, Leipzig] [1914–26. The Parallel Lives. Harvard University Press, Cambridge, MA].

Pocock, J.G.A., 1975. The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition. Princeton University Press, Princeton, NJ.

Polybios Historia. [1866–1868. Dindorf, L. (Ed.), 4 vols. Teubner, Leipzig] [1922–7. The Histories. Harvard University Press, Cambridge, MA].
Powell, J.G.F. (Ed.), 1995. Cicero the Philosopher: Twelve Papers. Clarendon Press, Oxford, UK.
Rahe, P.A., 1992. Republics Ancient and Modern: Classical Republicanism and the American Revolution. University of North Carolina Press, Chapel Hill, NC.
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York.
Robbins, C., 1959. The Eighteenth-Century Commonwealthman: Studies in the Transmission, Development and Circumstance of English Liberal Thought from the Restoration of Charles II until the War with the Thirteen Colonies. Harvard University Press, Cambridge, MA.
Rousseau, J.-J., 1762. Du contrat social. [1973. Guillemin, H. (Ed.), Union Générale d' Éditions, Paris] [1997. The Social Contract. Cambridge University Press, Cambridge, MA].
Rush, B., 1777. Observations upon the Present Government of Pennsylvania in Four Letters to the People of Pennsylvania. Steiner and Cist, Philadelphia, PA.
Sallustius Crispus C., Bellum Catilinae. [1991. In: Reynolds, L.D. (Ed.), Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana. Oxford University Press, Oxford, UK] [1921. The War with Catiline. Harvard University Press, Cambridge, MA].
Sallustius Crispus C., Bellum Iugurthinum. [1991. In: Reynolds, L.D. (Ed.), Catilina; Jugurtha; Historiarum Fragmenta Selecta; Appendix Sallustiana. Oxford University Press, Oxford, UK] [1921. The War with Jugurtha. Harvard University Press, Cambridge, MA].
Sandel, M.J., 1996. Democracy's Discontent: America in Search of a Public Philosophy. Belknap Press of Harvard University Press, Cambridge, MA.
Sellers, M.N.S., 1991. Republican impartiality. Oxford Journal of Legal Studies 11, 273–282.
Sellers, M.N.S., 1994. American Republicanism: Roman Ideology in the United States Constitution. Macmillan, Basingstoke, UK.
Sellers, M.N.S., 1998. The Sacred Fire of Liberty: Republicanism, Liberalism and the Law. Macmillan, Basingstoke, UK.
Sellers, M.N.S., 2003. Republican Legal Theory: The History, Constitution and Purposes of Law in a Free State. Palgrave Macmillan, Basingstoke, UK.
Sellers, M.N.S., 2006. Republican Principles in International Law: The Fundamental Requirements of a Just World Order. Macmillan, Basingstoke, UK.
Sidney, A., 1798. Discourses Concerning Government. [1990. West, T.G., (Ed.), Liberty Classics, Indianapolis, IN].
Skinner, Q., 1998. Liberty Before Liberalism. Cambridge University Press, Cambridge, UK.
Spitz, J.-F., 1995. La liberté politique: Essai de généalogie conceptuelle. Presses Universitaires de France, Paris.
Sunstein, C.R., 1993. The Partial Constitution. Harvard University Press, Cambridge, MA.
Symposium: The Republican Civic Tradition, 1988. The Yale Law Journal 97 (8).
Trenchard, J., Gordon, T., 1723. Cato's Letters or Essays on Liberty, Civil and Religious, and Other Important Subjects, 4 vols. [1995 Liberty Fund, Indianapolis, IN].
Tullius Cicero M. De legibus. [1914. Mueller, C.F.W. (Ed.), Teubner, Leipzig] [1998. The Laws. Oxford University Press, Oxford, UK].
Tullius Cicero, M.M., Antonium orationes Philippicae. [1986. Fedeli, P. (Ed.), second ed. Teubner, Leipzig] [1926. Philippics. Harvard University Press, Cambridge, MA].
Tullius Cicero, M. De officiis. [1994. Winterbottom, M. (Ed.), Oxford University Press, Oxford, UK] [1991. On Duties. Cambridge University Press, Cambridge, UK].
Tullius Cicero M. Orationes in L. Catilinam. [1901. Clark, A.B. (Ed.), Oxford University Press, Oxford, UK] [1976. In Catilinam. Harvard University Press, Cambridge, MA].
Tullius Cicero M. Pro L. Flacco Oratio. [1906. In: Mueller, C.F.W. (Ed.), M. Tulli Ciceronis orationes, Pro P. Sulla, Pro A. Licinio Archia poeta, Pro L. Flacco. Teubner, Leipzig] [1976. Pro Flacco. Harvard University Press, Cambridge, MA].
Tullius Cicero M. Pro P. Sulla oratio. [1906. In: Mueller, C.F.W., (Ed.), M. Tulli Ciceronis orationes, Pro P. Sulla, Pro A. Licinio Archia poeta, Pro L. Flacco. Teubner, Leipzig] [1976 Pro Sulla. Harvard University Press, Cambridge, MA].
Tullius Cicero M. De re publica. [1969. Ziegler, K. (Ed.), Teubner, Leipzig] [1998. The Republic. Oxford University Press, Oxford, UK].
Viroli, M., 1999. Repubblicanesim – Una nuova utopia della libertà. Lütevza, Roma. [2002. Republicanism [S. Anthony Trans.]. Hill and Wang, New York].
Washington, G., 1789. The First Inaugural Speech. [1985. In: Allen, W.B. (Ed.), George Washington: A Collection. Liberty Classics, Indianapolis, IN].
Wirszubski, C., 1960. Libertas as a Political Idea at Rome during the Late Republic and Early Principate, second ed. Cambridge University Press, Cambridge, UK.

Reputation

Kurt Lang, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A reputation exists in the form of a shared image that develops within networks. It can vary along three dimensions: from high to low, in size, and in longevity. Difference can be estimated either from survey responses or from indicators in documents. Reputations reduce uncertainty about what to expect from those who hold them. It influences both a person's own behavior and one's treatment by others. They follow trajectories of different length, some of which extend beyond a lifetime.

A reputation exists in the form of an image shared within some public. Applied to a person, it identifies "not an intrinsic quality he possesses but rather the opinion other people have of him [or her]" (Bailey, 1971, p. 2). One can also refer to an institution, organization, professional group (or its leadership), and by extension to just about anything rated or ranked like a dining establishment, a vintage year of wine, a neighborhood, or a resort in which to take a vacation along with much else. All these reputations are socially constructed. They may or may not be deserved and are always open to challenge. But, as long as they endure, they are real in their consequences.

Overview

Over 300 years ago, when John Locke referred to the "law of opinion and reputation," he highlighted how much reputations were a part of the small politics of everyday life ([1690] 1894, p. 479). To keep them in good reputation, people do accommodate to the opinions and rules subscribed to by those with whom they converse or otherwise interact. Adam Smith, a century later, argued in a similar vein when he located the love of virtue in the desire for "being what ought to be approved of." The success of persons who possess none of the privileges associated with noble lineage or inherited wealth "almost always depends upon the favor and good opinion of their neighbors and equals. Without a tolerably regular conduct, these can very seldom be obtained" (1804, p. I, 123 f). From a strictly economic perspective, moreover, a reputation constitutes a valuable property. It can grow or diminish but its value is real enough for anyone who feels his or her good name has been unjustly sullied by malicious or slanderous statements can seek redress through the courts even if not always successfully. And tax laws in some countries allow the corporate write-off of expenditures for building and protecting the reputation of their company or its brand name. From a broader social and cultural perspective, a reputation can be conceptualized, as Bourdieu (1992) has, as a form of symbolic capital validated against the perceptions of those in the know. The criteria by which they are judged vary among fields. The main one in economics is reliability or trustworthiness; in politics reputations involve the ability to mobilize support; what counts most in the social–cultural area is the quality of the achievement attributed to the target.

A significant body of research, much of it by economists on the gains an organization reaps from a good reputation and policies to preserve it but also by game theorists in probes of how an opponent's reputation influences an experimental subject's own strategy, has been published in recent years. An alternative approach to the subject is that of Kenneth H. Craik, a social psychologist who, as an outgrowth of the work of his institute on personality assessment, is more focused on how individuals develop reputations and how these evolve. The conceptual model presented in his book (2009), meant as a road map, takes as its point of departure the often ignored fact that people's appraisal of someone's beliefs and abilities, not necessarily of someone they know personally or based on what they actually know, are not just based on their own observations but often even more on what they hear others espouse until accepted as common 'knowledge.'

The Craik model incorporates three components. The first is visibility. One can speak of a reputation only when its carrier has a recognizable identity as a person (and I would add, organization, work of art, or other object) that is shared. This visibility, regardless of how restrictive or widely diffused, remains abstract, without content, and does not amount to a judgment in any sense. Reputational networks, the channels through which information about someone or something circulates, is the second critical component. Usually they extend beyond the circle of mutually known acquaintances to include people whose sole source is what they heard about from others, read about, or followed on television. Each individual participates in a network of his or her own making. The impressions they form, and consequently the reputation based on them, are apt to be greatest among people connected to the same network. This takes us to the third component, the distributive rather the discursive aspect, of the model. The contents that flow through the network end up in personal 'storage bins,' as Craik calls them, from which they can pass on whatever seems relevant to them as well as what others have to offer. After the requisite sifting, collation, and corrections people form the image we commonly designate as reputations. They remain subject to change as new information becomes available.

Reputations vary along three dimensions: Most attention has gone to the position of person or object's position on a hierarchical scale. Those ranking high have gained general respect, enjoy credibility, are free to exercise authority, admired

for their exceptional skill, and so forth. The list is long. Reputations, high or low, are also subject to lateral ranking with placement determined by the range within which a name is both visible and associated with certain qualities or achievements. Some terminological distinctions between these two aspects are called for. To meet this need, Lang and Lang (1990) proposed that reputations within some kind of in-group, such as a circle of personal acquaintances, a local neighborhood, professional peers, or other in-group, be designated 'recognition' and that those and the extent of its circulation throughout society 'renown.' While terms imply positive or favorable rankings, some earn a reputation not for good deeds or desirable traits but for something repellent and therefore to be avoided. To designate these two types on the negative side of the hierarchical scale but otherwise equivalent, we add another pair of terms to the lexicon: 'disrepute' for a bad reputation among peers and 'notoriety' that bad reputation is more widely diffused within a larger population.

The third temporal dimension indexes the duration of a reputation over time. Not only is its content apt to change as new information surfaces but revelations that no one ever expected can cause an idol to fall and enable some despised villains to redeem themselves. Most lifetime renown and notoriety suffer gradual decay because of competition from new entrants into severely limited public space ever more crowded because of the inexorable entry of new generations. As the surviving number of people with personal memories of predecessors passes away, they are bound to lose out to successors that compete for their place. Sooner or later, their reputation comes to depend exclusively on the preservation of the secondary evidence contained in written documents and other material textual residues of the past to become part of an established canon clamoring for recognition from practically everybody. The few that, once they have faded from the collective memory, are later unearthed to recover their former place, are indeed the exceptions.

Reputations, in short, follow different trajectories. At one extreme of the continuum are the 'celebrities' that suddenly burst on the scene apparently from nowhere to enjoy instant fame only to sink back quickly into obscurity, after which they are no longer of interest to anyone except for historians professionally specialized to dig into the past in order to ferret out what has been forgotten. The life span of a celebrity is short. The level of 'eminence' attained by the truly great who proved to be groundbreaking eludes them. The counterparts of both celebrity and eminence, which occupy the positive side of the hierarchical dimension, are 'ignominy,' such as the short-lived outrage triggered by something heinous, denotes a short public reaction. It represents the negative counterpart of celebrity and contrasts with the longer lasting 'infamy' associated with historical memories stigmatized as evil or as undeserved tragedies (cf Fine, 2001). The categories are arrayed in the accompanying diagram (see **Figure 1**).

Ascertaining a Reputation

A theory and a set of categories, if they are orient researchers, must be accompanied by indicators that signal the variable characteristics of the subject to which they apply. For reputations this requirement calls for methods and tools by and to locate or at least estimate positions on the separate scales in terms of who shares the reputation and the qualities they refer to. The two aspects are often fused.

The 'Who?' question demands adequate definition of the social circle within which the reputation is shared. Size does make a difference here. The reputation of the members of a family or a village small enough for every member to know everyone else is based on mutual recognition of what praiseworthy attributes of the object and of its defects draw disdain. The renown or notoriety acquired within a multitude too large and too dispersed for much interaction among those involved develops without, or even in spite of, the personal acquaintance a few among them may have with the object. Because reputations often differ between from social circle to circle, according to the knowledge available to each, a measure not explicit about who shares it – whether it is by people in general or only within a delimited group of insiders, within some self-selected category of connoisseurs, fans, or followers distinguishable only by their unusually high interest in the particular field or object being rated – is of little use.

The 'What?' question refers more to the vertical dimension. The reputation of an object evolves from the circulation of information not about some abstract quality but some particular mode of behavior, skill, deed, or achievement people attribute to it. Where the criterion for what is best, greatest, or most entitled to admiration are left vague, as it has occasionally been, the meaning of the findings can be hard to interpret.

Indicators used for measurement in past studies are essentially of two kinds: those tapping directly into the just mentioned storage bins by surveying the opinions within the relevant social circle and attempts at the quantification of evidence in documents of what these opinions are. To specify the quality on which the objects are ranked or rated as positive or negative is always helpful. The reliability of the findings obtained by either method will be greatest for people coming in some contact with one another. This holds even for reputations not actually earned but based on malicious rumors refuted too late to save an innocent from being pilloried. A method developed by J. L. Moreno, which he named sociometry, relied on peer ratings by members of some group as the best on something or whom they collectively would prefer for

	Lateral (expansion)	**Vertical (status)**	**Temporal (duration in time)**	
High	*Recognition*	*Renown*	*Celebrity*	*Eminence*
Low	*Disrepute*	*Notoriety*	*Ignominy*	*Infamy*

Figure 1 Types by Reputational Dimension

some role. Widely used in the selection of leaders, the method proved its value as a predictor of future performance. The selections by the members of groups more dispersed but frequently trading information, such as the colleagues in some field, also exhibit a fairly high level of convergence, but leave more room for strong disagreement by a few deviants.

The significance of this kind of informal consensus becomes more problematic the farther removed those queried are from the ones on whom they pass judgment, of which the national polls conducted for years asking respondents to name the man and also the woman they most admired, provide an extreme example. Given the freedom of respondents to nominate anyone alive, their choices have always been widely scattered but with the president and the first lady almost always heading the list. Their consistently high ranking, as well as of other highly placed names, has as much to do with the prominent coverage of these two in the news media as with the degree of respect they enjoy in the public sphere or, for that matter, among the political activists who constitute their particular reputational community. It is, however, a measure – although perhaps a somewhat crude one – of where the White House occupants place among a very small number of celebrities, or infamous villains, about whom just about everyone is prepared to express an opinion and, although correlated with measures of presidential approval and performance by this same public, should be mistaken for their equivalent. An open question that leaves respondents free to mention whoever comes to mind is, to say the least, of dubious value as a measure of public reputation.

This caveat does not extend to responses to similar queries addressed to highly qualified experts. Even when the sample of respondents is small and purposively selected, it does represent an effort to replicate what the providers of popular entertainment began to probe decades ago when they first sponsored audience surveys that asked people to state their 'favorites' among the diverse offerings and sources available to them. When seeking a meaningful appraisal of historical figures, there really is no-good alternative to relying specialists. A pioneer of this genre is the study by Arthur M. Schlesinger, a Harvard historian, of American presidents, conducted in the Roosevelt era when he sent a questionnaire to 55 colleagues, all them experts on the presidency, calling on them to place each occupant of the office into one of five categories ranging from 'great' to 'failure.' In the intervening decades, his study has been closely replicated by his historian son (Schlesinger, 1997). Still others contacted larger samples with more detailed questionnaires and more refined measures only to find that the same three – Washington, Lincoln, and Franklin D. Roosevelt – consistently placed at the top, although not always in the same order. Choices of the best and worst presidents have also remained fairly stable with more fluctuation among those ranked in-between, mostly incumbents who neither weathered a serious national crisis nor failed in some obvious way. Most striking among the changes has been the steady rise in reputation of Eisenhower as new information about his presidency has come to light.

More studies have been turning to nominations and/or ratings by respondents with some qualification as insiders to estimate reputations in more restricted realms. They include the rankings of major corporations have earned among business leaders by their exemplary performance, of judiciary figures considered outstanding by their peers, and of the top academic departments as rated by other academics in the same discipline, all of them with better access to facts than the rest of the population. As the practice has spread, the method has been refined. In a recent survey, underwritten by the federally sponsored National Foundation of Science, the open question was replaced by a meticulously designed questionnaire calling on academics in each of several to rank departments in their own discipline. All recipients were urged to respond but given a "don't know" option for any about which they felt insufficiently informed. The resulting rankings correlated with productivity, in terms of research productivity, but also turned up some discrepancies between this one and several other 'objective' indicators of departmental quality (*see* Scientometrics).

Many prefer documentary indicators of current reputation as also the signals of past fame to be found not only in surviving residues from the past, mostly in writing, but also in surviving shrines, statues, sacred objects, and even coins that celebrate a deceased. Insofar writers have been the primary purveyors of fame, how often a member of some group or category of subjects turns up in the literary sources scanned seems acceptable as a measure of the public regard. To be sure, characters not at all honorable or likable, can also arouse interest in their life histories but are less likely to be the subject of full-length biographies than those recognized as heroes and high achievers. There is more information on them and there are more past acquaintances happy to provide personal testimony. Rarely do less than stellar figures offer quite this enticement. Hence, the number of full-length biographies does at best identify only the most famous. Too many of the rest would fall into the 'none' or 'only one category' for this ranking to be of much use. Obviously, no single documentary indicator can possibly fit all. To cast one's net widely is always a good idea.

Documentary indicators discriminate best when limited to a specific field of endeavor or some recognized category of practitioners with some limits set on the time frame. When James R. Beniger (1990), a sociologist, scanned the authoritative encyclopedia of communication with contributions by the established experts in its subfields to estimate greatness from the number of mentions and pages devoted to each, the ranking he came up with turned out to be mildly surprising. A handful from the past, several from long ago, came out ahead of even the most distinguished scholars in this burgeoning field. Randall Collins (1998), another sociologist, in a truly magisterial study, used a similar method, to sort out the philosophers that were the real luminaries with influence on their contemporaries and whose ideas continue to be of great interest from the large number of lesser lights working in roughly the same time period.

For creative artists, an indicator as discriminating, and possibly more than the writings about them, is the fate of their works. Paintings, sculptures as well as prints, and even drawings from the hand of the 'great' masters usually occupy the most prominent places in exhibitions. The museums that own them devote a good part of their conservation budgets to protect and preserve them for posterity. In this they there motivated partly by their rarity and the high interest in them.

Whenever one appears on the art market, curators and wealthy collectors vie for it. They drive up the price to a level out of line with the practical utility of the object or with the labor and value of the materials that went into it. The high price the work commands is determined almost completely by the reputation of its creator. If ever a presumed masterpiece is exposed as a fake, as sometimes happens, or reattributed to one of the master's pupils, its market value plummets. The price premium exacted by the original over the same work, identical with the former in every other respect, disappears. This phenomenon testifies to the validity of reputational rankings of artists derived from the prices they achieve on the art market.

The performing arts allow for no similar distinction between an original premiere and a repeat performance. The longer a play runs or the more often a musical piece is performed the greater the boost to the reputation of its creator. The better indicator here is to be found in the frequency with which a work by a composer finds a place in the repertory of major orchestras or a playwright in the repertories of Broadway, London, and other centrally located theaters. Fashions come and go, and repertories change accordingly. The fairly few works granted canonical status may move down a couple of notches but are unlikely to disappear into the abundant archives of epigonal work, some of which may have had a highly successful premiere prior to receding into oblivion. Repertories often do, however, involve a tacit compromise between the appraisals shared by professionals and demands for work high enough in popular appeal to be a magnet strong enough to draw audiences of a size sufficient to assure the financial survival of the organization on whom they themselves depend. This need has been only somewhat less of a problem in the museum world, where art experts have more control over choices but where it still takes a blockbuster exhibition to bring in the crowds to impress sponsors and attract donors but where high costs, traceable to the scarcity of original masterpieces, are the primary constraint.

Performances, in contrast to the visual arts, do not even have to be 'live.' In fact, sales of recordings, not just of every hit but of the more established composers, playwrights, and performers, not only provide a meaningful measure of their contemporary renown. These data can also lead to potentially interesting findings about the few who survive in the collective memory.

The most 'democratic' of indicators are the frequencies of citation. It is widely used in the knowledge industry, where reference by authors to prior findings relevant to their own work, is an established practice. The utility of this measure for ranking leading scholars in the various academic fields has been pretty much institutionalized. Tenure committees, department heads, and deans in universities refer to it as one among other criteria for deciding who is deserving of a permanent appointment or the best on a short list of highly qualified applicants for an opening. Several indices listing all citations in publications of the professional literature in most disciplines available for half-a-century or more have since gone online for easier access.

Formal honors, such as prizes and election to honorary societies, are less satisfactory indicators of reputations. Some are simply too exclusive to pick more than a small number from an overwhelmingly large pool. Such are the Nobel prizes. Whenever awarded, which they are but once a year, they attract worldwide attention. Since only the still living qualify, the prematurely and unexpectedly deceased loose out, not for lack of the necessary qualifications but on grounds unrelated to the informal honors peers have bestowed on them in the course of an extended career. Selection into the Academy of France, with places for a maximum of 50 members, is similarly restricted to none but the most highly esteemed scientists as they are, in various ways, for many other honors granted only for the highest achievements in the given field within a predefined period. These awards are of no help in discriminating among the runner ups. In addition there is the multitude of potential aspirants whose names never reach the judges who make the final selections, often among nearly equally qualified candidates. Not having a truly objective measuring rod. Immediate unanimity is rare and the final selection often the result of negotiations and compromise.

Nor can their individual judgments be totally free from idiosyncratic preferences – political or personal – of their own, such as may be only tangentially related to what the award is intended to honor. Nor can one even expect these judges to peruse and review personally the qualifications of every prospective candidate. Most pools are simply too large for their limited time and capabilities. To whittle down the numbers to something manageable, most juries rely on preliminary screening by assumedly qualified gatekeepers. This may work for some but for some awards, such as the Oscars annually dispensed by the American movie industry, the process can become hectic. The initial selections are ballot. Some hundreds and more are eligible because of their professional roles in the making of films. Whatever the specific procedures, they leave little room for discussion among screeners to give them opportunity to fill any gaps in their knowledge and correct what misconceptions they may have. When the final jury meets, the best it can do is adjudicate among the remaining candidates. Meanwhile, the recent proliferation of all kinds of awards, many funded by donors eager for publicity, is not very effective as a countertrend. Many of them are so specific and made by so small a constituency that the sheer number amassed by any has lost most of its value as an indicator of the recipient's general standing.

Effects of Reputation

A wide range of empirical work highlights how having a reputation affects that person's own behavior as well as its influence on the behavior of others toward that same person. Fear of doing damage to one's own reputation acts as a deterrent – against selling shoddy goods, harming others while pursuing one's own short-term interest, and all kinds of wrongdoing that most people prefer not to be caught in. Its deterrent influence garners strength when the same two parties engage in repeated dealings over an extended time and, not surprisingly, is generally more effective on people involved in dense social networks or, better yet, among the members of an intimate group, in which they are well-known and with whom they have built warm relationships that they treasure. It certainly is more effective in such a setting than under conditions of partial anonymity. Further and to the extent that a person's reputation has been incorporated into their personal identity, they will be

even more motivated to go out of their way to live up to it by refraining from anything that could raise doubt in the mind of other whether are really the kind of person they want them to think that they are. Nevertheless and contrary to the contention of Adam Smith, which has been seconded by many economists, a reputation is not invariably an incentive for 'good' behavior, not because the process they identified does not work as they see it but rather because of the occasional existence of motives in competition with those posited by them. Thus, for a tyrant or a bully the ability to instill fear can to override concern that this might never strike people as morally upright. Coercive measures that bring desirable results can indeed be more gratifying than blind adhesion to the moral prescriptions most people support as a matter of principle without regard for what the consequences.

Persons greatly concerned about the reputation they project, especially performers and professionals whose fortune depends on their being in the limelight, may be moved to call on spokespersons with special expertise. This same motive could also tempt some to making false claims and to doctor their records which, if their fabrications were ever to come out would bring them infamy instead of approbation for which they are thirsting. Behavior toward adversaries is governed by much the same logic. A state involved in a dispute with another will be disinclined to yield "on a series of issues when matters at stake are not critical... may [make] it difficult to communicate to the other just when a vital issue has been reached" but inclined to act forcefully for no other reason than to retain its "reputation for action" (Schelling, 1966, p. 124). A more recent researcher concluded on the basis of an elaborate statistical analysis of more than 1300 international crises accompanied by threat of force that a state's reputation for honesty or bluffing was an influence on its reputation (Sartori, 2005).

The effect of one's own reputation on the behavior of others has been subject to intensive experimental examination, most often by game theorists guided by formal models of a prisoner's dilemma. Just how well a game repeatedly played in social psychological laboratories played by the same two players approximates the complexities of real life must remain an open question. Whatever the final answer, these experiments at least point in the right direction. By showing how subjects confronted with a deliberately ambiguous situation – due to their uncertainty over whether the other 'prisoner' will confess – this body of research contains ample documentation of the signal importance of a reputation for trustworthiness whether explicitly assigned by the experimenter or one that the subjects themselves inferred from the pattern of behavior they were able to observe on their own. The development of trust, even when there seems no rational basis for it, fosters cooperation.

The above is a guiding principle in a good part of the world of commerce, where people have little or no personal knowledge of the trustworthiness of the many strangers they deal with. All too often reputations, whether by word-of-mouth or formal endorsements, are the main source of the reassurance for people in deciding what to buy, how much to pay for it, safe places to invest, the college likely to offer the best educational experience, and so forth. Mass advertising via the media and, more recently, over the Internet also has an important part in the process even for people not aware of its influence. A good reputation or being associated with a respected band name is as much of an asset for a company as name recognition and an appealing personality are for a political leader. Groups and institutions are under the same duress. Nor are interactions between enemies trading threats instead of goods and votes unconcerned over how they are perceived by the other. In the Cold War period, the two major powers were maneuvering to gain an advantage but anxious to avoid a catastrophic outcome, they gave a lot of weight in planning their every move both to the reputation they had constructed from their opponent's past behavior in similar situations and to how their own moves might affect the image of themselves they had been trying to convey to their opponent. Neither wanted to act out character and risk a wrong response.

On the Internet, where the parties to a transaction are anonymous, an entirely different situation prevails. The participants in Internet auctions are not at all connected so as to be able to circulate the information about sellers needed to develop a good or bad reputation. For buyers on eBay the main source of reassurance is the posted feedback previously left by other customers of the seller and, vice versa, by sellers about buyers. They are, however, a very imperfect proxy for reputations. Their greatest plus is for researchers. They give them access to a database of a sufficient magnitude for statistically reliable findings. Studies, among them a randomized and controlled field experiment of transactions, confirm that sellers with near-perfect feedback ratings generally realize higher prices, especially for more expensive items and where the buyer also incurred some risk. There are other studies that, as far as they go, confirm that negative ratings, while harming a seller in the short run, make little difference over the longer run and that a very positive reputation, which takes longer to build, has more staying power.

An increasing number of smaller but highly competitive businesses – book sellers, hotels, restaurants, etc. have likewise been exploiting the improved versatility of the Internet with invitations for customer feedback that end up on a website, not necessary their own but those of rating agencies. Reputations derived without interaction among customers oare clientele are less reliable than reputation grounded in pooled experience. Participants in eBay, for example, as free to opt out and re-register under a new alias. Nor are yet the means to prevent anyone from mobilizing friends, relatives, or employees to sing their praise on Internet sites like Angies List or Yelp, two agencies whose activities border too closely on advertising in the form of the Good Housekeeping seal of approval during earlier period. A genuine reputation in the commercial realm functions much like an established brand name that allows its owner to extract some price premium over a generic product.

On the other hand, the advantages of a reputation once established tend to be cumulative. This was shown by Robert K. Merton (1967) to hold for a large part of the knowledge industry when he observed that the more distinguished scientists receive a disproportionate amount of the credit for discoveries that result from jointly published research conducted in close collaboration with others. Another researcher found evidence in his study of information exchange among the scientists employed in the same company that, when one of them contacts another to request for information meant to

advance a project of his or her own, the one more renowned stands a better chance to receive a helpful answer than a lesser one (Ensign, 2009). On the institutional level, elite universities are better able to recruit higher a quality faculty, not only because of its high salaries and ample resources but equally because applicants expect more from the presence of especially distinguished colleagues, which in turn acts as a magnet for prospective donors, the best students, thereby adding to its prestige. Nor can we entirely rule out its influence on reviewers of grants. Submissions by scientists to the Australian Research Council received somewhat higher ratings (Jayasinghe et al., 2003, pp. 279–300). In these contexts, the multiple advantages of a reputation are of quite a different kind (for a discussion of academic reputation *see* Scientometrics).

But nowhere are the returns from a good reputation greater than in the arts, where values are somewhat intangible, subject to fashion, highly dependent on critical and public endorsement (Moulin, 1987). When paintings certified by art experts as originals by the seventeenth Dutch master Jan Vermeer were finally exposed as a modern forgery by Han van Meegeren, they lost nearly all value. The name attached to a work affects not only its price but also the attention it gets. A change in the attribution of the Man in with the Golden Helmet, previously celebrated as a painting by Rembrandt, to one of his students, led to its removal from the permanent exhibition in the German National Gallery in Berlin.

Reputational Trajectories

Propositions about the development of recognition derive largely from life histories, usually of successful individuals, treated by their authors as slated for success. Some endowment or ability is an obvious prerequisite but not enough by itself. Reputations diffuse through of social networks and nothing really counts unless enough people take notice. Admirable deeds, repeatedly honest performance, even a remarkable achievement or transcending some overpowering obstacle if by a habitual recluse is easily ignored while comparable ones by more gregarious individuals are much more likely to be fully acknowledged. Collective approval by peers and members in the same in-group, obtained more easily than more general renown, is the usual first step up the ladder but still not so high that others have to look way up for a good view. Farther ascend to the level of the renown becomes contingent on the ability of the individual to extend these home-grown networks. Some who achieved fame had indeed everything going for them from early on – a stable family, early signs of talent nurtured by elders, a background and sufficiently well connected to the right background and connections that can help, backed by the confidence in their ability to promote themselves. All these are great reputational assets but being deficient in one or more of them asset may be an obstacle but not always an insurmountable barrier that prevents otherwise endowed individuals from developing a reputation and from ascending to even the pinnacles of fame. Great strength in one asset can even substitute for weakness in another. For Albert Einstein, probably the most renowned among the ones with an apparent handicap, it was failing a major school examination; for Glenn Cunningham, the champion miler for several years, it was a childhood accident, in which he suffered such serious burns on his legs that his doctors were not at all sure whether he would ever walk again. And, quite apart from the proverbial late bloomers, dire poverty in childhood was a serious obstacle for innumerable others.

The reputation of a person, institution, or product is the image that other individuals develop from a combination of their own observations but only to the extent that they are confirmed and corrected. The qualities people ascrib to someone are not always be the ones initially possessed by that person or not in the amount credited. They therefore have to be earned. A good example is the struggle confronting an American president notwithstanding the recognition any incumbent commands immediately upon taking office. From here on how much power his office is able to wield, according to Richard Neustadt, a scholar of the American presidency, depends on how well the incumbent measures up to the expectations political insiders have of him. Neustadt locates the source of that power, not just in the Constitution and in the mandate supposedly bestowed by an electoral victory, but in what others in the Washington political community think and expect of the incumbent. Justified or not, "it is the reality on which they act, at least until their calculations turn out wrong …." (1990, p. 51). What the president must demonstrate in the many different games he plays, often all at the same time, is his ability and his will to make use of any bargaining advantages he has. Success breeds success but a pattern of repeated failure is bound to result in a loss of faith in his reputation for effectiveness within the restricted network of political insiders, which would will impair his power to persuade.

A repeatedly outstanding exercise of power or the admiration granted for some act or performance, like a deed previously considered beyond human capability or actions worthy of sainthood, can be a first step toward renown but still leave many quite sterling individuals by the wayside. Their achievements may have come too late in life or they never aspired to wider recognition. Only rarely is renown conferred posthumously on people overlooked in their lifetime. Those intent on building a reputation tend to start early in their career and hew to the right track. One of the smoothest roads toward advancement is to address a problem on whose importance there is general agreement and to exploit it for the attention and contacts such effort is bound to bring. The results may be disappointing but the respect it commands from associates could still be a springboard from which to launch a campaign for greater renown. Attaching oneself to important people is a second strategy, equally appropriate for a career in business, politics, the sciences and the arts. All in all, people who do not stray too far off the mainstream are best positioned to outpace their contemporaries at least in the early phase of their career but not necessarily in the competition for the highest honors, which ultimately are more likely to go to the innovators, who may at first have been given a hard time by a complacent and overly conservative establishment in control of all the good places but who over the longer run the renown that eludes most of its members.

Whether recognition within any group converts into renown is highly contingent on distribution and on publicity (Becker, 1982). On this front, each field of endeavor makes its

demands. Writers and publishers need booksellers; playwrights, performers and a suitable stage; and the performers themselves need recordings and bookings on major outlets, and visual artists exhibition space for public display of their often unwieldy accomplishments. Upcoming painters stymied in their ambitions by an entrenched art establishment in control of scarce public facilities and the best private outlets have formed associations to be better able *collectively* pursue the recognition so far denied them but to which they felt their accomplishments entitled them. The evident appeal of this strategy is what led the painters denigrated as 'impressionists,' an epithet for their apparent disdain for truly 'serious' subjects, to found their own exhibition group. Many more since have chosen the same course to gain visibility.

An important facilitator of anyone's rise in a world where so many interactions are with strangers is locating a devoted patron, hopefully one with the influence and means to become an effective mediator of their reputation. Willingness to invest in this role can be motivated by a desire to be in on the latest, by the prestige that being the first to spot a rising star may bring, by a charitable impulse to eradicate poverty in some deserving cultural realm, or simply by the requirements of a journalist's or a teacher's job. Then as a reputation gains momentum, it builds on itself until it starts to decline, its time having passed and new arrivals move into the limelight. Mediators are in a position to do things that most creators, lacking in the requisite skills, do not do very well for themselves and who are therefore dependent, to this extent, on the intervention of who depend to this extent, on the assistance and intervention of wealthy collectors and influential critics. Their praise, especially when consensual, has the power to stimulate museum purchases (Moulin, 1987). It can also persuade theater circuits to reserve a permanent place for a work in their repertories and boost a reputation in other ways whereas lack of attention, even more than sharp criticism, condemns an artist or a work to a marginal existence.

Courting popularity can also backfire. Insiders in one of the more esoteric fields of the arts or of knowledge are apt to spurn rather than to embrace the renown acquired by publicity hungry colleagues. To protect their elite status, they insist on a clear separation between real quality and mass appeal. The problem for Alfred Hitchcock, a highly successful Hollywood director of many popular films, was that not all film genres were equally ordained by the elite establishment. To scale this barrier, he adopted the new and more acceptable idiom and transformed himself into a genuine *auteur* (Kapsis, 1992). It is of course possible that the public has bad judgment that it is unqualified to select what is truly deserving of special approval. But this hardly explains all differences in reputations. There is, for example, little room for ambiguity in singling out the top performers in most sports, where achievements are measurable in terms of scores, wins, speed, height, weight, or distance. As a result, no more than a highly select few winners and record holders ever achieve fame beyond the fans of their sport. The critical factor here is once again the size of the talent pool. It has grown too large for a completely dispassionate and objective sorting out in terms of performance. It has given rise to a star system and celebrity cult that, although still unmistakably oriented toward performance criteria, is less than coldy objective. The media have played their part in anointing a relatively small number of outstanding athletes, not always unequivocally the best. Once so identified, they receive a huge excess of acclaim compared to nearly equal competitors who perform in relative obscurity but on whom the appeal of the game or contest for spectators depends. The consequences of this 'winner take all' situation (Goode, 1978; Frank and Cook, 1995) are distinctly unfair, especially" when the highly acclaimed achievements, from which a few profit excessively, depend on cooperation, as it does very clearly in some team sports, where the contribution of supporting players remains largely unnoticed. The problem in other areas is actually worse. Most discoveries and inventions would have been impossible without the work predecessors and current achievements, such as in rocketry, where Wernher von Braun stands out as a heroic figure, have become ever more dependent on team effort. Yet only a small number, even among scientists, would even recognize the names of the many highly competent engineers who have make space flight possible.

The survival of all reputations is equally problematical. Over time, as memories fade, the lives of most everyone, except for a diminutive number of canonical figures, end up in the dust bins of history. There is a continuous sorting out of those deemed truly memorable. As to likely survival in the collective memory, posthumous reputations are the combined effect are sustained, on the one hand, by the deceased's lifetime accomplishments on society or on some domain and, on the other, by the extent to which, even if for altogether different reasons, an accomplishment still touches the imagination of posterity. When it does, 'greatness' can be retrospectively conferred as validation of or correction to earlier valuations but this hardly ever happens. Notwithstanding these advantages, it is more likely that individuals once recognized for their accomplishments will sink into oblivion soon after their death than for some hardly knowns to be posthumously be discovered.

The renown achieved in a person's lifetime enhances the chance for reputational survival insofar as it is likely to have encouraged more contact with others uniquely qualified to serve as mediators of fame across generations (Zilsel, 1972). Kings, popes, presidents, generals, and executives benefit from the archivists at their command but those without an official position have to link up with the appropriate networks. Colleagues and subordinates (students, employees, or loyal followers might be ready at hand but would still have to be linked to the gatekeepers in control of the portals into and out of the archives. Should the advocates represented in some network be dissatisfied by their progress, they may join in a society of 'friends' to prepare the way for some kind of a monument or a museum, in which to store the papers plus all sorts memorabilia of and about the deceased or, at the minimum, to convert a former residence, birthplace, lab, or studio into a shrine. To succeed such a venture would above all have to attract funds.

The already famous, because of their wider contacts, already have an advantage over the less known. They leave more traces with more people. This can happen all on its own, without any doing on their part, as a consequence of their lifetime fame, which lends value to anything connected with them and which provides an extra incentive to collectors and archivists to go after anything of theirs, especially pictures, diaries, and correspondence, can put people into vicarious touch with the

deceased. Because the paths of the already famous will also have crossed those of other famous contemporaries, it increases the likelihood that similar material will also have been housed in other archives. Its availability means that the scholar on the hunt for data does not have to start from scratch.

Subjects themselves do, however, have some, although limited, control over the record on which their reputational fate comes to depend. Those who take care to keep and organize records strengthen their claim on the memory of posterity. A practice that has become standard for American presidents is developing plans for a library to house his papers, despite the fact that most would be saved in any event. Some of them as well as other famous people are known to have taken the further step of purging their files of items inconsistent with the image they prefer to project. Writers, too, have burned manuscripts they believed detrimental to their reputation, and there are painters who systematically destroyed all pictures executed in a style or of a quality for which they do not want to be remembered. Theodore Roosevelt did the reverse when he dictated letters intended for posterity. Thomas Hardy actually controlled the biography his wife published under her own name following the author's death (Hamilton, 1994). The same impulse shows up in instructions to estate executors to convert the residence of the deceased into a landmark or to use remaining assets to preserve his or her reputation. Not that their wishes are always followed or followed more than half-heartedly. Most notorious is Richard Nixon's effort to retain control over his secretly recorded tapes that was obviously thwarted by other interests that prevailed.

Factors affecting the survival of reputation beyond the control of the individual are, there are circumstances of birth and death. Prospective candidates for a claim on symbolic immortality can be to born into a time or to live when or where there are none to record their deeds. It so happens that the Chinese precursor of Gutenberg, allegedly the first to set moveable type, remains unknown and that of Columbus, instead of Leif Ericson, is hailed as the European discoverer of America. As to the timing or circumstances of death, a long productive life full of honors is an obvious reputational asset. So is to be struck down in the prime of life by an assassin, a tragic illness, or unfortunate accident, which has the potential to transform the victim into an icon. No one can possibly know how John F. Kennedy would have been remembered had he served out his terms or had also been reelected. Yet the halo that surrounds him because of his tragic and untimely death overshadows the modest achievements and all too apparent failures of his administration. A still more striking example is that of Anne Frank. Would her diary, translated into many languages, have had the same impact if she had survived the Holocaust? And had she survived, would she have tolerated the publication by her father of this very private document from which he excised passages that displeased him.

Having survivors who, like Anne Frank's father, have a personal stake in preserving memory of a deceased is affected by such demographic factors as marriage, fertility, and longevity. Lang and Lang (1990) in their study of posthumous reputations of British etchers found that the revival of that art form in the late nineteenth and early twentieth century had attracted an unusually large number – for that time period – of talented women, well recognized during their lifetimes and voted into the honorary society for the highly versed practitioners of that medium. Today they are much less remembered than their male peers. Contributing to the reputational disparity between the two groups was first the extraordinary longevity of these 'lady etchers' and second the large number of them who remained single or, if they ever married, had no children. They often Outlived not only their spouses but also their colleagues, close friends, publishers, dealers, admiring collectors as well as the end of the etching cycle, often by decades, so that by the time they died there was practically no one left to mourn their death and write the obligatory obit. An additional handicap was not having a mate or a daughter to whom they could have delegated the tedious task of record keeping to serve later as a source for a future biography or for a definitive catalog raisonné of the deceased printmaker's work. For many of them there is none and never will be one.

All reputational communities have boundaries. The heroes celebrated in one may be largely unknown in another. Nowhere is this more obvious than in the different statesmen, military heroes, literary figures and so forth the different nations celebrate. National perspectives even intrude into the allocation of credit for achievements universally recognized as a breakthrough. Obviously, the pioneers of space exploration who have received the greatest attention in America differ from those celebrated in the former Soviet Union. The inventor of the incandescent lamp, as remembered by Americans, is Thomas Edison, who was in fact preceded by Joseph Swan, the English inventor who exhibited the first successful carbon-filament light bulb in England, where he had the greater impact. His factory near Newcastle rivaled Edison's in prosperity. But whereas Edison immediately patented everything, Swan was slow to do so. Nor did he have a wealthy sponsor like Henry Ford, who thought him the greatest man alive and started collecting Edison memorabilia as early as 1905. Ultimately, by transporting, piece by piece, and reassembling his New Jersey laboratory in Dearborn, Michigan, Ford erected a shrine for the inventor. For swan's reputation to grow across cultural and geographic boundaries and be an inspiration, as Edison was to become, he or some sponsor had to make his achievement appear unique and the inventor behind it a veritable genius.

Time creates its own boundaries. With only so much room in an ever more crowded field, what Collins (1998) refers to as the 'law of small numbers' comes into play. Secondary figures are pushed into the background as new interests, ideas, experiences, and those who espouse them come to the fore. Within these limits, posterity still has some leeway in how it views the past, not only in selecting but also by continuously reinterpreting the achievements of works by people with a firmly established place in the canon. It reinterpets and selects anew the work of writers and artists. As Jane Tomkins (1996) pointed out with respect to Nathaniel Hawthorne, the writings most admired by his contemporaries are not the same as those that touch us most strongly today. Similarly, according to George Boas (1940), the qualities for which Da Vinci's *Mona Lisa* has been admired have significantly changed from century to century. Long-term eminence, to put it another way, derives not solely from the stubborn persistence of beliefs about persons and their deeds. It has to be multivalent and cut across domains. The larger the number of domains touched on, the greater its resilience to gradual decay.

The development, change, preservation, and decline of reputations as well as how they function have to be understood as part of an ongoing social process. Experiments and statistical studies provide useful information, mostly about outcomes, but require supplementation with case material to illuminate the extent to which reputations and their trajectories are socially determined. Qualitative evidence, especially when derived from more than one case and subjected to systematic analysis, is not to be discounted as merely anecdotal.

See also: Advertising: Effects; Collective Memory, Social Psychology of; Culture, Sociology of; Scientometrics.

Bibliography

Braudy, L., 1986. The Frenzy of Renown. Oxford University Press, New York.
Bailey, F.G., 1971. Gifts and poison. In: Gifts and Poison: The Politics of Reputation. Blackwell, Oxford, pp. 1–25.
Becker, H., 1982. Art Worlds. University of California Press, Berkeley.
Beniger, J.R., 1990. Who are the most important theorists of communication. Communication Research 17, 698–715.
Boas, F., 1940. Race, Language and Culture. The Macmillan Co., New York.
Bourdieu, P., 1992. Les règles de l'art. Editions du Seuil, Paris.
Cooley, C.H., 1927. Fame. In: Social Process. Scribner's, New York, pp. 112–124.
Collins, R., 1998. The Sociology of Philosophy: A Global Theory of Intellectual Change. Belknap Press, Cambridge.
Craik, K.H., 2009. Reputation: A Network Interpretation. Oxford University Press, New York.
Ensign, P.C., 2009. Knowledge Sharing Among Scientists: Why Reputation Matters for R&D in Multinational Firms. Palgrave Macmillan, New York.
Fombrun, C.J., 1996. Reputation: Realizing the Value from the Corporate Image. Harvard Business School Press, Boston.
Fine, G.A., 2001. Difficult Reputations: Collective Memories of the Evil, Inept, and Controversial. University of Chicago Press, Chicago.
Frank, R.H., Cook, P.J., 1995. The Winner-take-all Society: How More and More Americans Compete for Ever Fewer and Bigger Prizes, Encouraging Economic Waste, Income Inequality, and an Impoverished Cultural Life. Free Press, New York.
Goode, W.J., 1978. The Celebration of Heroes: Prestige as a Social Control System. University of California Press, Berkeley, Los Angeles.
Hamilton, I., 1994. Keepers of the Flame: Literary Estates and the Rise of Biography from Shakespeare to Plath. Faber and Faber, Boston.
Jayasinghe, U.W., Marsh, H.W., Bond, N., 2003. A multilevel cross-classified modeling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. Journal of the Royal Statistical Society: Series A 166, 279–300.
Klein, D.B. (Ed.), 1997. Reputation: Studies in the Voluntary Elicitation of Good Conduct. University of Michigan Press, Ann Arbor.
Kapsis, R.E., 1992. Hitchcock: The Making of a Reputation. University of Chicago Press, Chicago.
Lang, G.E., Lang, K., 1990. Etched in Memory: The Building and Survival of Artistic Reputation. University of North Carolina Press, Chapel Hill.
Lange-Eichbaum, W., Kurth, W., 1956. Genie, Irrsinn und Ruhm: Eine Pathographie des Genies (Transl. from the German and a thoroughly revised 4th edition). English Books, Munich, Basel.
Merton, R.K., 1967. The Matthew effect in science. Science 159, 56–63.
Moreno, J.L., 1947. Sociometric Control Studies of Grouping and Regrouping: With Reference to Authoritative and Democratic Groupings. Beacon House, New York.
Moulin, R., 1987. The French Art Market : A Sociological View. Rutgers University Press, New Brunswick.
Neustadt, R.E., 1990. Presidential Power and the Modern Presidents: The Politics of Leadership from Roosevelt to Reagan. Free Press, New York.
Sartori, A.E., 2005. Deterrence by Diplomacy. Princeton and Oxford: Princeton: Princeton University Press.
Schelling, T.C., 1966. Arms and Influence. Yale University Press, New Haven.
Schlesinger Jr., A.M., 1997. Rating the presidents: Washington to Clinton. Political Science Quarterly 112, 179–190.
Smith, A., 1804. The Theory of Moral Sentiments. Cadell & Davis, London.
Tomkins, C., 1996. Duchamp: A Biography. Henry Holt, New York.
Zilsel, E., Maus, H., Olms, G., 1972. Die Enstehung des Geniebegriffes: ein Beitrag zur Ideengeschichte der Antike und des Fruehkapitalismus. Hildesheim.

Resampling Methods of Estimation

Bradley Efron, David Rogosa and Robert Tibshirani, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 19, pp. 13216–13220, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

In this article, a brief description of the bootstrap method for data analysis given. The use of the bootstrap for estimating standard errors and confidence intervals is reviewed and illustrated with an application to growth curves for educational testing data.

Most scientists occasionally or frequently face problems of data analysis: what data should I collect? Having collected my data, what does it say? Having seen what it says, how far can I trust the conclusions? Statistics is the mathematical science that deals with such questions. Some statistical methods have become so familiar in the scientific literature, especially linear regression, hypothesis testing, standard errors, and confidence intervals, that they seem to date back to biblical times. In fact most of the 'classical' methods were developed between 1920 and 1950, by scientists like R. A. Fisher, Jerzy Neyman and Harold Hotelling who were senior colleagues to statisticians still active today.

The 1980s produced a rising curve of new statistical theory and methods, based on the power of electronic computation. Today's data analyst can afford to expend more computation on a single problem than the world's yearly total of statistical computation in the 1920s. How can such computational wealth be spent wisely, in a way that genuinely adds to the classical methodology without merely elaborating it? Answering this question has become a dominant theme of modern statistical theory.

The Bootstrap

The following fundamental question arises in almost every statistical data analysis: on the basis of a data set $\mathbf{z}$ we calculate a statistic $s(\mathbf{z})$ for the purpose of estimating some quantity of interest. For example, $\mathbf{z}$ could be the nine cholesterol reduction scores shown in **Table 1**, and $s(\mathbf{z})$ their mean value $\bar{z} = 28.58$, intended as an estimate of the true mean value of the cholesterol reduction scores. (The 'true mean value' is the mean we would obtain if we observed a much larger set of scores.) How accurate is $s(\mathbf{z})$?

This question has a simple answer if $s(\mathbf{z})$ is the mean $\bar{z}$ of numbers $z_1, z_2, \ldots, z_n$. Then the *standard error* of $\bar{z}$, its root mean square error, is estimated by a formula made famous in elementary statistics courses.

$$\left[\sum_{i=1}^{n}(z_i-\bar{z})^2/(n\cdot(n-1))\right]^{\frac{1}{2}} \qquad [1]$$

For the nine numbers in **Table 1**, Eqn. [1] gives 10.13. The estimate of the true cholesterol reduction mean usually would be expressed as 28.58 ± 10.13, or perhaps $28.58 \pm 10.13 \cdot s$, where s is some constant like 1.645 or 1.960 relating to areas under a bell-shaped curve. With $s = 1.645$, interval (2) has approximately 90% chance of containing the true mean value.

Table 1 Nine men's cholesterol reduction scores after taking cholostyramine; the nine scores are an ordered random sample from the scores of 164 men (data from the Stanford arm of the LRC-CPPT experiment, courtesy of Drs D. Feldman and J. Farquhar, Stanford University (see Efron and Feldman, 1991)

−21.0	3.25	10.75	13.75	32.50	39.50	41.75	56.75	80.0

The bootstrap, (Efron, 1979), was introduced primarily as a device for extending Eqn. [1] to estimators other than the mean. For example, suppose $s(\mathbf{z})$ is the '25 percent trimmed mean' $\bar{z} = \{0.25\}$, defined as the average of the middle 50% of the data: we order the observations $z_1, z_2, \ldots, z_n$, discard the lower and upper 25% of them, and take the mean of the remaining 50%. Interpolation is required for cases where $0.25 \cdot n$ is not an integer. For the cholesterol data, $\bar{z} = \{0.25\} = 27.81$.

There is no neat algebraic formula like Eqn. [1] for the standard error of a trimmed mean, or for that matter almost any estimate other than the mean. That is why the mean is so popular in statistics courses. In lieu of a formula, the bootstrap uses computational power to get a numerical estimate of the standard error. The bootstrap algorithm depends on the notion of a *bootstrap sample*, which is a sample of size n drawn *with replacement* from the original data set $\mathbf{z} = (z_1, z_2, \ldots, z_n)$. The bootstrap sample is denoted $\mathbf{z}^* = (z_1^*, z_2^*, \ldots, z_n^*)$. Each z_i^* is one of the original z values, randomly selected, perhaps $z_1^* = z_7, z_2^* = z_5, z_3^* = z_5, z_4^* = z_9, z_5^* = z_7$, etc. The name *bootstrap* refers to the use of the original data set to generate 'new' data sets $\mathbf{z}^*$.

The bootstrap estimate of standard error for $\bar{z}\{0.25\}$ is computed as follows: (a) a large number 'B' of independent bootstrap samples each of size n, is generated using a random number device; (b) the 25% trimmed mean is calculated for each bootstrap sample; (c) the empirical standard deviation of the B bootstrap trimmed means is the bootstrap estimate of standard error for $\bar{z} = \{0.25\}$. **Figure 1** shows a schematic diagram of the bootstrap algorithm, applied to a general statistic $s(\mathbf{z})$.

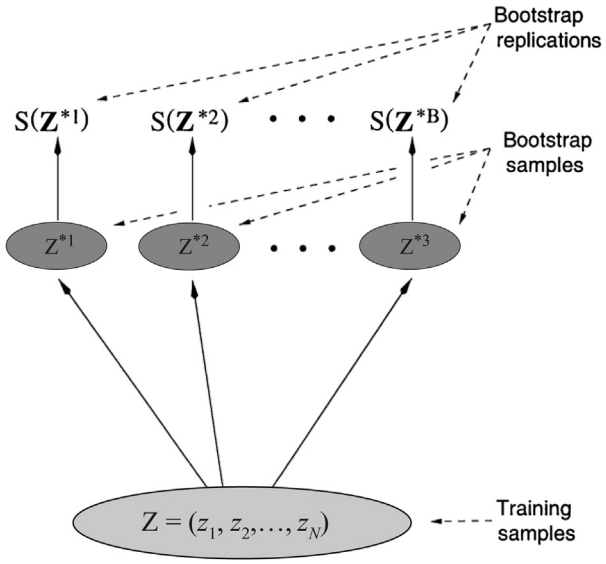

Figure 1 A diagram of the bootstrap algorithm for estimating the standard error of a statistic $s(\mathbf{z})$; each of the B bootstrap samples is a random sample of size n drawn with replacement from the original data set $(z_1, z_2, \ldots, z_n)$.

Trying different values of B gave these bootstrap estimates of standard error for the 25% trimmed mean applied to the cholesterol data

B:	25	50	100	200	400
Bootstrap estimate of standard error:	12.44	9.71	11.50	10.70	10.48

Ideally we would like B to go to infinity. However, it can be shown that randomness in the bootstrap standard error that comes from using a finite value of B is usually negligible for B greater than 100. 'Negligible' here means small relative to the randomness caused by variations in the original data set $\mathbf{z}$. Even values of B as small as 25 often give satisfactory results. This can be important if the statistic $s(\mathbf{z})$ is difficult to compute, since the bootstrap algorithm requires about B times as much computation as $s(\mathbf{z})$.

The bootstrap algorithm can be applied to almost any statistical estimation problem. (a) The individual data points z_i need not be single numbers; they can be vectors, matrices, or more general quantities like maps or graphs. (b) The statistic $s(\mathbf{z})$ can be anything at all, as long as we can compute $s(\mathbf{z}^*)$ for every bootstrap data set $\mathbf{z}^*$. (c) The data set '$\mathbf{z}$' does not have to be a simple from a single distribution. Other data structures, for example, regression models, time series, or stratified samples, can be accommodated by appropriate changes in the definition of a bootstrap sample. (d) Measures of statistical accuracy other than the standard error, for instance biases, mean absolute value errors, and confidence intervals, can be calculated at the final stage of the algorithm. These points are discussed in Efron and Tibshirani (1993). The more complicated example below illustrates some of them.

There is one statistic $s(\mathbf{z})$ for which we don't need the computer to calculate the bootstrap standard error, namely the mean $\bar{z}$. In this case it can be proved that as B goes to infinity, the bootstrap standard error estimate goes to $\sqrt{(n-1)/n}$ times Eqn. [1]. The factor $\sqrt{(n-1)/n}$, which equals 0.943 for $n = 9$, could be removed by redefining the last step of the bootstrap algorithm, but there is no general advantage in doing so. For the statistic $\bar{z}$, using the bootstrap algorithm is about the same as using Eqn. [1].

At a deeper level it can be shown that the logic that makes Eqn. [1] a reasonable assessment of standard error for $\bar{z}$, applies just as well to the bootstrap as an assessment of standard error for a general statistic $s(\mathbf{z})$. In both cases we are assessing the standard error of the statistic of interest by the 'true' standard error that would apply if the unknown probability distribution yielding the data exactly equalled the empirical distribution of the data. The efficacy of this simple estimation principle has been verified by a large amount of theoretical work in the statistics literature of the 1990s (see the discussion and references in Efron and Tibshirani, 1993).

Why would we consider using a trimmed mean rather than $\bar{z}$ itself? The theory of *robust* statistics, developed since 1960, shows that if the data $\mathbf{z}$ comes from a long-tailed probability distribution, then the trimmed mean can be substantially more accurate than $\bar{z}$, that is, it can have substantially smaller standard error (see, e.g., Huber, 1964). The trouble, of course, is that in practice we don't know *a priori* whether or not the true probability distribution is long-tailed. The bootstrap can help answer this question.

Figure 2 shows the bootstrap estimates of standard error for 5 different trimmed means $\bar{z}\{p\}$, 'p' being the proportion of the data trimmed off each end of the sample before the mean is taken. (So $\bar{z}\{0\}$ is $\bar{z}$, the usual mean, while $\bar{z}\{0.5\}$ is the

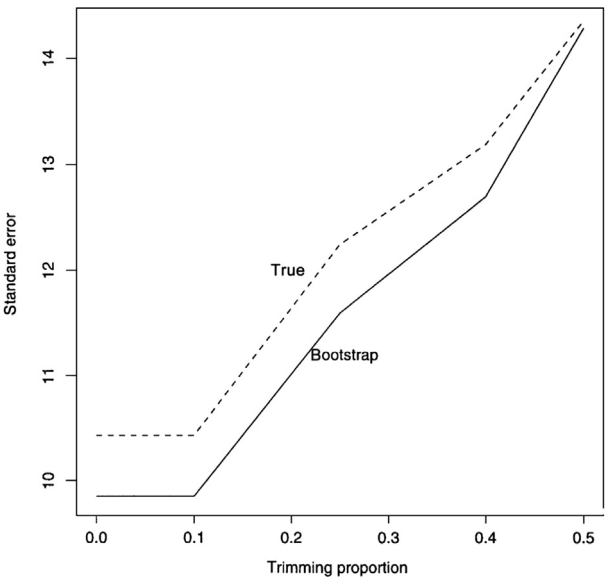

Figure 2 Bootstrap estimates of standard error for five different trimmed means $\bar{z}\{p\}$, $p = 0.$ 0.10, 0.25, 0.40, 0.5, applied to the cholesterol data of **Table 1**. Based on $B = 400$ bootstrap samples. Also shown is the true standard error of $\bar{z}\{p\}$, obtained by taking random samples of size nine from the population of 164 cholesterol reduction scores. In this case the bootstrap correctly indicates that $\bar{z}\{0\}$, the ordinary mean, gives the smallest standard error.

median.) These were computed using the bootstrap algorithm in Figure 1, $B = 400$, except that at step 2 five different statistics were evaluated for each bootstrap sample $\mathbf{z}^*$, namely $\bar{z}\{0\}$, $\bar{z}\{0.10\}$, $\bar{z}\{0.25\}$, $\bar{z}\{0.40\}$, and $\bar{z}\{0.50\}$.

According to the bootstrap standard errors in Figure 2, the ordinary mean has the smallest standard error among the five trimmed means. This seems to indicate that there is no advantage to trimming for this particular data set.

In fact the nine cholesterol reduction scores in Table 1 were a random sample from a much bigger data set: 164 scores, corresponding to the 164 men in the Stanford arm of a large clinical trial designed to test the efficiency of the cholesterol-reducing drug called cholestyramine (Efron and Feldman, 1991). With all of this extra data available, we can check the bootstrap standard errors. The solid line in Figure 2 indicates the true standard errors for the five trimmed means, 'true' meaning the standard error of random samples of size nine taken from the population of 164 scores.

We see that the true standard errors confirm the bootstrap conclusion that the ordinary mean is the estimator of choice in this case. The main point here is that the bootstrap estimates use only the data in Table 1, while the true standard errors require extra data that usually isn't available in a real data analysis problem.

Growth Curves

Longitudinal data are often modeled using growth curves. Our example uses educational testing data from students in North Carolina (see Williamson et al., 1991; Rogosa and Saner, 1995). The data we use for the example are eight yearly observations on achievement test scores in mathematics Y for 277 females, each followed from grade 1 to grade 8. The straight-line growth curve model for each individual can be written as

$$Y_p(t) = \xi_p(0) + \theta_p t + \varepsilon_{pt} \qquad [2]$$

where the observable scores Y contain error of measurement indicated by ε_{pt}. Here t denotes time and p denotes person. The constant rate of change for each person is indicated by θ_p; θ_p has a distribution over persons with mean μ_θ and variance σ^2_θ.

To illustrate bootstrap estimation we pick out two of the questions that are investigated using the growth curve models:

a. Can individual change be measured reliably? In behavioral and social science research the reliability index for the estimated rate of change, $\rho(\widehat{\theta})$, is the parameter of interest.
b. Is there a relation between initial standing (math achievement in grade 1) and the rate of improvement? The parameter to be estimated is correlation between θ and initial status (over the population of individuals), written as $\rho_{\theta\xi(1)}$.

Maximum-likelihood estimates of $\rho(\widehat{\theta})$ and $\rho_{\theta\xi(1)}$ can be obtained from standard mixed model analyses. Standard errors and interval estimates for these quantities are not available from widely-used computer packages.

Figure 3 shows histograms for 4000 bootstrap estimates of $\rho(\widehat{\theta})$ and $\rho_{\theta\xi(1)}$. Bootstrap samples were constructed by resampling individuals (where the data for each individual is a vector of the eight observations). The 90% confidence intervals were obtained using the so-called 'percentile method': the endpoints are the 0.05 and 0.95 percentiles of the bootstrap histograms. (The percentile method is the simplest of a number of methods for confidence interval construction via the bootstrap: a generally better method is the 'BC_a'; see Efron and Tibshirani, 1993, chap. 12–14. The results in Table 2 show good precision for estimation of the reliability, but relatively less precision for the correlation parameter.

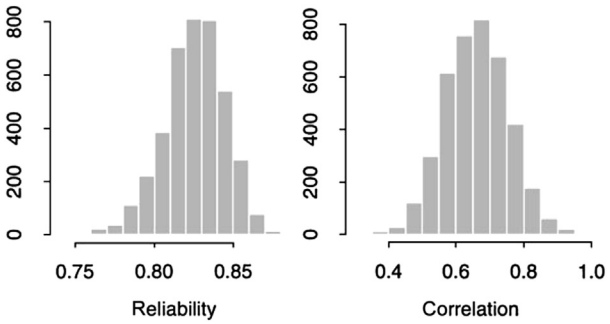

Figure 3 Histograms of 4000 bootstrap estimates of $\rho(\theta)$ (left) and $\rho_{\theta\chi(1)}$ (right).

Related Methods

There are a number of important related methods, with strong similarities to the bootstrap. The *jackknife* systematically deletes one or more observations from the dataset, and applies the statistic to the remaining data. By doing this for each observation or groups of observations, one can obtain estimates of standard errors and bias that roughly approximate to the bootstrap estimates.

Cross-validation is similar to the jackknife but is used to assess the prediction error of a model. It systematically excludes one or more observations from the dataset, fits the model to the remaining data, and then predicts the data that was left out. Application of the bootstrap for estimating prediction error is not straightforward and requires some special modifications. With these, the bootstrap and cross-validation have similar performance for estimating prediction error. Details and references are given in Efron and Tibshirani (1993, 1997).

There is a connection of the bootstrap to Bayesian inference. Roughly speaking, the bootstrap distribution approximates a nonparametric posterior distribution, derived using a non-informative prior. Rubin (1981) gives details.

A difficult theoretical question is: when does the bootstrap work? There is much research on this question (see, e.g., Hall, 1992). In 'regular' problems where asymptotic normality can be established, first and higher order correctness of the

Table 2 Results for growth curve example

Parameter	*Estimate*	*Standard error*	*90% CI*
Reliability $\rho(\widehat{\theta})$	0.828	0.019	(0.792, 0.855)
Correlation $\rho_{\theta\xi(1)}$	0.651	0.093	(0.510, 0.814)

bootstrap (and modifications) have been proven. However, the bootstrap can fail, for example, when it is used to estimate the variance of the maximum or minimum of a sample. More generally, it is difficult to prove theoretically that the bootstrap works in complicated problems: but these are the very settings in which the bootstrap is needed most. Eventually, theory will start to catch up with practice. In the meantime, the bootstrap, like any statistical tool, should be applied with care and in the company of other statistical approaches.

Some popular statistics packages have bootstrap functions, including S-PLUS and Resampling Stats. In most packages that facilitate some customization by the user, it is easy to write one's own bootstrap function.

Conclusions

The bootstrap is a powerful tool for data analysis, with many applications in the social sciences. It is being used more and more, as fast computational tools and convenient software become widely available. Good general references on the bootstrap include Efron and Tibshirani (1993), Davison and Hinkley (1997), Hall (1992), Shao and Tu (1995), and Westfall and Young (1993).

See also: Exploratory Data Analysis: Univariate Methods; Functional Data Analysis.

Bibliography

Abadie, A., Imbens, G.W., 2008. On the failure of the bootstrap for matching estimators. Econometrica 76 (6), 1537–1557.

Alfaro, Michael E., Zoller, Stefan, Lutzoni, François, 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution 20 (2), 255–266.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.

Efron, B., Feldman, D., 1991. Compliance as an explanatory variable in clinical trials. Journal of the American Statistical Association 86, 9–26.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman and Hall, London.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: The 632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

Efron, B., 1979. Bootstrap methods: Another look at the jack-knife. Annals of Statistics 7, 1–26.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer-Verlag, Berlin.

Huber, P., 1964. Robust estimation of a location parameter. Annals of Mathematics and Statistics 53, 73–101.

MacKinnon, David P., Lockwood, Chondra M., Williams, Jason, 2004. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivariate behavioral research 39 (1), 99–128.

Molinaro, A.M., Simon, Richard, Pfeiffer, R.M., 2005. Prediction Error Estimation: A Comparison of Resampling Methods. Bioinformatics 21 (15), 3301–3307.

Rogosa, D.R., Saner, H.M., 1995. Longitudinal data analysis examples with random coefficient models. Journal of Educational and Behavioral Statistics 20, 149–170.

Rubin, D., 1981. The Bayesian bootstrap. Annals of Statistics, 130–134.

Shao, J., Tu, D., 1995. The Jack-knife and Bootstrap. Springer-Verlag.

Westfall, P.H., Young, S.S., 1993. Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment. Wiley, New York.

Williamson, G., Appelbaum, M., Epanchin, A., 1991. Longitudinal analyses of academic achievement. Journal of Educational Measurement 1, 61–76.

Research and Development in Organizations

Keith Pavitt[†], University of Sussex, Brighton, UK
Mei-Chih Hu, National Tsing Hua University, Hsinchu, Taiwan

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by K. Pavitt, volume 19, pp. 13220–13224, © 2001, Elsevier Ltd.

Abstract

The steady increase in research and development (R & D) activities in modernizing societies reflects increasing specialization and professionalization in the creation of the knowledge that underlies inventions and innovations in products, processes, services, and in related systems of organization. Functionally specialized R & D departments emerged in business firms in the nineteenth century, as did complementary and more theoretical R & D activities and training in universities and related institutions. National R & D expenditures have increased with national income per head, as has the share funded and performed by business firms. However, these R & D activities are only one aspect of the invention and innovation process; a variety of other science and technology indicators have been developed and are now applied in order to reduce risks associated with bringing a product through the stages of R & D to market. Major contemporary problems in R & D relate to increasing complexity, increasing internationalization and localization, the changing role of academic research, and the barriers imposed by intellectual property protection (especially patents).

Introduction

The growth of research and development (R & D) activities reflects the increasing division of labor in the production and use of scientific and technological knowledge – a process foreseen by Adam Smith at the beginning of the Industrial Revolution. Since then, professional education, the establishment of laboratories, and improvements in techniques of measurement and experimentation have progressively increased the efficiency of discovery, invention, and innovation (Mowery and Rosenberg, 1989). While the notion of the National Innovation System largely depends on how actors, which include governments, firms, bridging institutes, and other contributing institutions, function and interact with each other to develop and apply innovative knowledge, the functioning of R & D systems for innovation has been addressed in different ways in the literature in the contexts of both advanced and latecomer countries (e.g., Hobday, 1995; Furman et al., 2002; Hu and Mathews, 2005). (Freeman (1987) is the first to propose the concept of a National Innovation System in his study of Japan's technological development, defining it as the network of institutions in the public and private sectors whose activities and interactions initiate, import, modify, and diffuse new technologies, later studies built on this with similar definitions and generally recognized this concept as comprising a complex of functions and interactions among various actors and institutions.) Despite different approaches and designs, it is widely agreed that patterns of specialization are critical for linking complex networks of innovation systems, wherein the various degrees of interaction between and among public R & D, private R & D, and university R & D are important. In particular, the influence of public R & D on specialization in high-tech industries is largely effective in East Asian latecomer countries, as they make the transition from imitation to innovation.

Building on the current literature, this article aims to discuss the influence of R & D management and its derived problems and challenges at the country, industry, and firm levels. We begin by highlighting the increased specialization in knowledge production and use, followed by the influence of R & D patterns and activities, along with problems in the processes of R & D commercialization. To relieve these problems, some of the applied methodologies and tools, as well as their economic impacts, are also elaborated on. We then conclude our findings with a discussion of the contemporary debates in R & D management, for future studies.

As addressed below, three complementary forms of specialization have developed in parallel. They are specialization by discipline, specialization by corporate function, and specialization by institution. Each contributes to the growth of R & D, as well as shaping innovative capacity at the country, industry, and firm levels.

Specialization by Discipline: The Growth of Science and Engineering

Progress and specialization in scientific disciplines accelerated in the nineteenth century with the development of more powerful research tools and techniques. Increasing opportunities for application of scientific knowledge led to the emergence of the engineering disciplines: hence chemical engineering in addition to applied chemistry, and electrical engineering in addition to applied physics (Rosenberg and Nelson, 1994). The differences in timing of the emergence of new opportunities for innovation reflected radical but uneven improvements in the knowledge base underlying technological change, and in particular the emergence of new technological paradigms (Freeman et al., 1982; Dosi, 1982).

The mechanical paradigm was the basis of the Industrial Revolution. It did not grow out of contemporaneous scientific advances, but of improvements in the quality of metals and the precision with which they could be formed and machined. These enabled the *design, development, and testing* of prototypes of families of machines with greatly improved performance based on materials of homogeneous and predictable quality,

[†] Deceased.

and on shapes with compatible sizes. Similarly, initial improvements in metal processing technologies depended less on scientific understanding than on *development* activities, experiments with different alloys and processing conditions in pilot plants. With the wide application of useful tools such as TRIZ and process simulation software, such development activities in prototype and pilot plants are typically accelerated and account for about three-fourths of the R & D expenditures of private business firms and for about two-thirds of total R & D. (TRIZ (the theory of inventive problem solving) is a problem-solving, analysis, and forecasting tool, deriving from the study of patterns of invention in the global patent literature. It was developed by the Soviet inventor and science fiction author Genrich Altshuller and his colleagues, beginning in 1946. The theory includes a practical methodology, tool sets, a knowledge base, and model-based technology for generating new ideas and solutions for problem solving. For more details, please see: http://www.triz-journal.com.)

Specialization by Corporate Function: The Growth of Business R & D

The science-based chemical paradigm (based increasingly on synthetic chemistry) and the electrical paradigm (based on electromagnetism and radio waves) laid the basis for the development of the industrial R & D laboratory, initially in the chemical and electrical firms of Germany and the USA. Many of the initial applications – including the use of techniques of chemical analysis to control the quality and composition of materials in the increasingly large-scale processing industries – were relatively mundane, but produced very profitable cost reductions. Management learned progressively that the science-based technological opportunities were applicable over a range of existing and new product markets, and therefore opened up opportunities for what is now called 'related' product diversification. Mowery and Rosenberg (1989) have described the spread of the establishment of R & D laboratories in the USA from the chemical industry to other sectors, and from larger to smaller firms. Similar processes have been at work in Europe, and more recently in East Asia (Caron et al., 1995; Nelson, 1993).

Thus, in addition to the benefits of the cognitive division of labor into more specialized fields, the functional division of labor within business firms also augmented the rate of technical change. Corporate R & D laboratories and similar groups, devoted full-time to inventive and innovative activities, provided an effective method for creating, combining, and coordinating increasingly specialized knowledge. They provided improved and specialized instrumentation, which enabled firms to monitor and benefit more systematically and effectively from advances in specialized academic disciplines. They also created techniques in the development and testing of laboratory concepts and prototypes, and translated these into commercialized products. Firms were able to experiment with a wider range of products and processes than had previously been possible when constrained by established products and production lines. In fields rich in technological opportunity, firms have consequently broadened their scopes, creating multiple products by the use of a variety of technologies.

Technological paradigms have been associated with the emergence of large dynamic firms that have been successful in exploiting new opportunities. The largest R & D spenders today are in companies that grew with the emergence of the mechanical (and automobile), biochemical, and electrical–electronic paradigms. Aerospace is a special case, having been technologically force-fed, especially since World War II, by government R & D subsidies and procurements linked with military requirements. The fastest growing R & D spenders today are those closely associated with biotechnology, ICTs (information and communication technologies), and software technology.

Specialization by Institution: The Growth of Academic Research

In the nineteenth century, de Tocqueville foresaw that the dynamics of capitalist competition would greatly stimulate the development of innovative activities that showed the prospect of a commercial return. He also argued that public authorities would need to support complementary public research of a more fundamental nature, in order to avoid diminishing returns, to open up new opportunities, and to provide trained researchers. This has happened too. In both advanced and latecomer countries, governments have become the main source of the funding of research activities, and related postgraduate training, in universities and similar institutions. Corporate R & D laboratories have come to depend increasingly on a supply of scientists and engineers aware of the latest research results and trained in the latest research techniques (Salter and Martin, 2001; Lundvall et al., 2011). In particular, based on the US Bayh–Dole Act of 1980, many East Asian latecomer countries have adopted an equivalent model since the late 1990s. This has made universities, in coordination with public research institutes, act as critical sources of knowledge production and diffusion for the industrial development of the nation (Mathews and Hu, 2007).

R & D and Modernization

This historical pattern of specialization and growth in knowledge production has been broadly repeated in the processes of modernization of latecomer countries, and is observable in countries at different levels of development today. R & D expenditures as a percentage of Gross Domestic Product (GDP) increase with GDP per head, both in specific countries over time, and in cross-sections of countries at any point in time. In successful modernizing countries, R & D is preceded by systematic business investments in improvement-generating activities: in particular, in production engineering, quality control, and design activities, initially making minor modifications and improvements, and later becoming the basis for indigenous innovative activities (Lall, 1992; Bell, 1984; Hu and Mathews, 2005). At this stage, business-funded R & D increases rapidly, accompanied by the growth of university-based research in underlying disciplines, with both reflected in the growth of international patents granted and scientific papers published. This pattern can be observed most clearly in both

South Korea and Taiwan during the 1980s and 1990s, and China since the 2000s.

The centrally planned Soviet model of modernization, practiced in the former USSR and imposed or adopted elsewhere, gave high priority to R & D. By the late 1950s and early 1960s, R & D expenditures in Central and Eastern European countries were apparently higher than those in Western Europe and the USA. Certain Western observers, therefore, concluded that the Soviet system was superior in promoting R & D and technical change. However, Soviet economic performance subsequently deteriorated, and it later became clear that a very high proportion of Soviet R & D was oriented toward weapons development, and that the government-established R & D laboratories established for each industry were decoupled from the requirements of producers and consumers (Hansen and Pavitt, 1987). The major reductions in R & D activities since their collapse in 1989 can be seen as painful adjustments to make R & D activities – and the underlying activities in production engineering, quality control, and design – become an integral part of a process of economic modernization (Dyker and Radosevic, 1999).

The Soviet system also had inadequate linkages with technical advance in the rest of the world economy. One important feature distinguishing today's modernizing countries from those of the nineteenth century is the availability of more productive technologies in more advanced countries. Countries successful in assimilating these more advanced technologies have had two characteristics (Hobday, 1995): firstly, strong linkages with the more advanced countries' technology, whether through inward direct investment (e.g., Singapore), inward technology licensing by locally owned firms (e.g., Japan, South Korea), or subcontracting agreements from advanced countries to local firms (e.g., Taiwan); secondly, the development of local, change-generating activities, culminating in R & D activities that were essential inputs to effective imitation, as well as to innovation (Cohen and Levinthal, 1989). In particular, the involvement of public R & D in the emerging technologies not only essentially catalyzed and induced the specialization of corporate R & D activity but also acted as a facilitator, associated with university R & D, to bridge the knowledge gap between advanced countries and latecomer countries in the transitional process from imitation toward innovation (Hu and Mathews, 2005).

Patterns of R & D Activity

By the beginning of the 1960s, initiatives by the US National Science Foundation and the Organization of Economic Co-operation and Development had established common definitions of R & D activities, and led to the collection by governments of systematic data on R & D activities. These have since been complemented by privately funded surveys comparing the R & D activities of individual business firms. They show the following common and largely invariant features of R & D activities.

1. In industrially advanced countries, business and government are the main sources of funding for R & D activities. Business funding is in general larger and spent in-house, mainly on *applied research and development* activities. Government funding is divided between *basic research* performed mainly in university-type institutes, and various types of R & D associated with health, the environment, defense, and the support of industry and agriculture. This R & D is performed mainly in the laboratories of government agencies and business firms.
2. In industrially advanced countries, the share of GDP typically spent on R & D varies between 1.5 and 3.0%, compared to 0.5 and 1.5% in the newly industrializing countries, and less than 0.5% in the rest. The share of national R & D funded and performed by business firms tends to positively correlate with increasing GDP per head and is widely recognized as an indicator of the degree of industrialization and modernization pursued by the industrially latecomer countries.
3. More than 60% of all business-funded R & D is typically performed in the chemical-, biotechnology-, electrical-, electronic-, and automobile-related industries. Currently, the largest individual corporate R & D spenders are the world's leading automobile, biochemical, electrical, and ICT firms.

The Economic Determinants and Impact of R & D Activities

The improved quality of R & D statistics has enabled economists and other analysts to more deeply understand both the determinants and the economic impacts of R & D activities, at the country, industrial, and firm levels. A number of analysts have been able to show that, among the industrially advanced countries, differences in the levels and rates of growth of national R & D activities have a significant influence on differences in national performance in exports and productivity (Fagerberg, 1987; Furman et al., 2002). In the industrially latecomer countries such as the East Asian Tigers, studies of R & D conducted by public research institutions, such as ITRI for Taiwan or KIET for Korea, made it clear that they had indeed been the architects of the technological catch-up strategies and had been responsible for acquiring the technologies that were then rapidly diffused throughout the business sector (Hu and Mathews, 2005). This was a feature not observable in studies of industrially advanced countries, however.

At the industry level, considerable attention has been given to the effects of firm size and industrial structure on R & D activities. Given the dependence of specialization on scale, the proportion of firms performing R & D increases with firm size. For example, by using USPTO patent data, Hu (2012) demonstrated that R & D activity and degrees of specialization vary with firm size and industrial structure, in the development of the TFT-LCD (thin-film transistor-liquid crystal displays) industries in Japan, Korea, and Taiwan. Specifically, advanced Japanese firms formed inter-*keiretsu* collaborations to increase both technological scale and efficiency (i.e., both R & D scale and intensity) while protecting itself from the risks of a technological and capital-intensive industry. To the extent that technological opportunity was available, Korean conglomerate latecomers increased their *scale* of technology specialization (i.e., R & D scale) to improve their market performance, while Taiwanese small to

medium-sized latecomer firms pursued *efficiency* of technology specialization (i.e., R & D intensity) by utilizing their limited R & D resources. This is compelling evidence that variations between industries in both the concentration and intensity of R & D are determined, jointly, by a third factor, which concerns the differences in technological opportunities available between industries (Levin et al., 1985).

Less progress has been made so far in measuring the effects of differences in R & D expenditures on company performance. This is partly because comprehensive and comparable R & D data at the company level are only now slowly becoming available. Also, the intangible effect of R & D spillovers is often induced and moderated by external factors such as industry-level technological opportunity and institutional design at the country level (Coe et al., 1997, 2009). It is also because of the difficulty in defining a proper measure of corporate performance. Those used include the R & D production function, stock market evaluation, and long-term growth (Patel, 2000).

Limitations of R & D Statistics

As shown above, R & D activities can be a useful proxy measure for innovation, but these have their limitations. The most often mentioned of which is that they measure inputs and not outputs. Equally, if not more, important are the following:

1. In large firms, R & D statistics do not measure the often considerable expenditures on outsourced or complementary activities in production and marketing that are necessary to transform the outputs of R & D into commercially successful innovations.
2. In small firms, they do not measure innovative activities conducted by part-time, external manpower, which are particularly important in the machinery and software sectors.
3. For service sectors such as finance and distribution, R & D activities are very imperfect proxies, because a growing volume of innovative activities in software are now being performed in firms. In such sectors, however, quantitative outputs performed in the market are largely mediated by qualitative indicators such as qualified manpower and management skills.

To improve the limitations of traditional R & D statistics, indicators of knowledge-generation and innovative activities have been developed. The most applicable of these have been counts of patents and scientific papers, as well as their citations, technology licensing and royalty revenues, and the number of new and innovative products and services (Hu et al., 2014). By using patent information, the impact of R & D activity can, to some extent, be evaluated and measured, by examining the patenting activity of each organization in comparison with its competitors and collaborators. We addressed this as follows.

Patent Information versus R & D Activities

In knowledge-intensive economies, intellectual property rights (and especially patents) have become valuable assets for R & D activities and decisive strategic factors. They protect R & D achievements, raising research performance, creating profit and maintaining competitive advantages for countries, industries, and firms. In particular, the patent document is considered one of the major indicators of cutting-edge technology, and holds great commercial value. The WIPO (World Intellectual Property Organization) reported that over 90% of R & D achievements were disclosed in patent gazettes, 80% of these not being revealed in academic journals or the technical literature. Therefore, one of the clearest indicators of R & D performance is the rate of take-up of patents issued by the patent offices, although the precise rate of technological innovation is not measurable. Based on this, different indicators are being developed and used to predict the value of a particular patent, or an organization's R & D strength.

The standard patent indicators most utilized are patent number and the growth rate of patents in a given organization. Since patents can vary enormously in their importance or value, simple patent counts are unlikely to capture the full force of the innovative output of the sector (see, for example, Jaffe and Trajtenberg, 2002). Thus, the alternative is patent citation rate, which is recognized as a proxy reflecting the impact and depth of R & D activity. While research has established that highly cited patents represent economically and technically important inventions (e.g., Griliches, 1990), we list five of the most popular patent citation indicators, invented by the ipIQ, to evaluate R & D activities. (ipIQ (formerly CHI Research) is one of the foremost intellectual property consultants, providing technology-oriented services, including IP-related industry surveys and reports to firms contemplating acquisitions. The company has formulated various measures of patenting quality, which can be deployed through own calculations.)

1. Forward citation rate – a count of the citations received by an organization's patents from subsequent patents. This helps to evaluate the technological impact of patents. High citation counts are often associated with important inventions, ones that are fundamental to future inventions and that may have more competitive advantages in a given technological field.
2. Backward citation rate – a count of the citations to prior patents made reference by an organization's patents. This helps to trace the source of innovation/knowledge as well as the developmental trajectory of innovation capability in the organization.
3. Current Impact Index – represents the number of times in which an organization's previous 5 years of patents are cited in the current year; this shows how frequently the organization's R & D outcomes were used as foundations for other inventions.
4. Technology Cycle Time (TCT) – indicates speed of innovation or how fast the technology is turning over, defined as the median age of the patent cited on the front page of a patent document. The TCT is measured in years. This indicator can measure the pace of technological progress. Shorter cycle times reflect faster substitutions, indicating faster progress; while longer cycle times reflect slower substitutions, indicating slower progress.
5. Science Linkage Indicators – the counts of patent references citing to scientific papers. Increasingly, patents are citing nonpatent documents as prior art, and many of these are

papers in scientific journals. High science linkage thus indicates that a patent is building its technology based on advances in science.

By making use of patent information, decision-makers are thus able to assess their own R & D portfolio in comparison with their competitors within and across the industry/country. Patent analysis enhances the ability to assess the attractiveness of technologies, especially those posing a threat, or new opportunities for the existing business; recognize strategic changes in the competitive environment; identify and assess external sources for knowledge generation (e.g., mergers and acquisitions (M & A) and R & D alliances); assess the status of patent appropriation in new business areas; evaluate important market partners, especially customers and suppliers, to determine whether the firm's R & D strategy is in alignment with that of its major customers and competent suppliers; and improve human resource management pertaining to leading inventors in specific technological fields. In summary, integrated patent information can act as a quick and efficient monitoring system for capturing R & D trends in the firm/industry/country by the timely review and updating of patenting portfolio analysis.

Conclusion: Contemporary Debates in R & D Management

Since the nineteenth century, many of the characteristics and problems of managing corporate R & D have remained until today. Of these, the most important issues are as follows

1. The need to orchestrate and integrate specialized knowledge and skills across disciplines, professions, functions, and divisions.
2. Progressive improvements in fundamental scientific understanding, accompanying the increasing systemic complexity of innovations – a process that will continue with the more widespread application of the techniques emerging from ICTs.
3. The difficulty of making reliable predictions, demonstrations, and measurements about the success or otherwise of specific innovations, especially major ones.

Various types of methodologies and open source/proprietary software (such as the above-mentioned TRIZ, business modeling, and knowledge integration simulations) have been developed to help overcome these problems that affect the journey of products from the stages of R & D to market.

At the same time, new problems and challenges have emerged.

1. With increasing specialization in both the production of knowledge, and of the components of increasingly complex systems, firms are faced with increasingly difficult choices about what products and knowledge to outsource, and what to retain in-house. An important part of the in-house R & D function is now devoted to the coordination of innovative and production activities external to the firm through examination and regular monitoring of patenting activity within the nation and throughout worldwide patent data bases, including the WIPO, Trilateral patent offices (including US Patent and Trademark Office, European Patent Office, and Japan Patent Office), and China's State Intellectual Property Office.
2. While the risks and costs of developing cutting-edge products are rapidly increasing, efforts to secure intellectual property appropriation by means of IP licensing and litigation have reached a critical stage, and now face challenges in the deployment of business/industrial/national strategies. In many instances, especially in the industrially latecomer countries, impositions of high licensing royalties and threats of patent infringement lawsuits become significant barriers to entry, together with transaction costs for business expansion and industrial development in a nation.
3. In addition to the well-established function of supporting foreign production, large firms are performing a growing share of their R & D outside their home country, in order to tap into the increasingly numerous international sources of leading-edge scientific and technical advances, as well as in local markets. This poses new challenges for managers in integrating skills and knowledge over long distances and diverse culture (Niosi, 1999), and for national policy-makers in assessing the location and institutional benefits of public investments and resources allocation. Consequently, a multidisciplined and well-trained workforce has become one of the most valuable assets for companies, industries, and nations as a whole.
4. The evidence has shown that academic research is becoming increasingly linked to commercial R & D activities, wherein academic R & D performance is measured by not only publications but also technology licensing revenues, and transfer and number of start-ups. For some, this is the consequence of an unwelcomed financial and institutional pressure from governments on universities to trade-off their forward-thinking 'basic research' with short-term oriented 'applied research and experiments.' For others, it is the consequence of fundamental changes to the role of universities, now supplying a well-educated workforce for knowledge exploitation. This cultural transition of universities challenges policy-makers in the management of academic research.

This study revisits the often-discussed topic of the conceptualized influence of R & D management at the country, industry, and firm levels. The relationship between R & D activity and market performance in each organization is associated with different resource allocations, which affect and are affected by specialization levels, technological opportunity, industrial-specific concentrations, and country-specific investment and institutional preferences. Patenting information provides specific and reliable indicators for R & D evaluation. Furthermore, the contemporary debates and challenges in R & D management indicate that other non-R & D measures, such as inter- and intraknowledge sharing and integration, linkage between basic research and applied research, strategic business and cross-licensing alliances, and fluid manpower flows are crucial complementary drivers that reduce transaction costs in the commercial processes of R & D activities.

See also: Corporate Culture; Design and Form: Organizational; Information and Knowledge, Organizational; Innovation and Technological Change, Economics of; Innovation, Theory of; Intelligence, Prior Knowledge, and Learning; Learning: Organizational; Organizational Behavior, Psychology of; Organizational Decision Making; Organizations: Unintended Consequences; Technology Transfer; Technology and Organization.

Bibliography

Bell, M., 1984. Learning and the accumulation of technological capacity in developing countries. In: Fransman, M., King, K. (Eds.), Technological Capability in the Third World. St Martin's Press, New York.

Caron, F., Erker, P., Fischer, W., 1995. Innovations in the European Economy between the Wars. W de Gruyter, Berlin.

Coe, D.T., Helpman, E., Hoffmaister, A.W., 2009. International R&D spillovers and institutions. European Economic Review 53 (7), 723–741. Scopus Exact.

Coe, D.T., Helpman, E., Hoffmaister, A.W., 1997. North-South R&D spillovers. Economic Journal, Royal Economic Society 107 (440), 134–149.

Cohen, W.M., Levinthal, D.A., 1989. Innovation and learning: the two faces of R & D. Economic Journal 99, 569–596.

Dosi, G., 1982. Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. Research Policy 11, 147–162.

Dyker, D., Radosevic, S. (Eds.), 1999. Innovation and Structural Change in Post-socialist Countries: A Quantitative Approach. Kluwer, Dordrecht, The Netherlands.

Fagerberg, J., 1987. A technology gap approach to why growth rates differ. In: Freeman, C. (Ed.), Output Measurement in Science and Technology: Essays in Honour of Y. North-Holland, Fabian. Amsterdam.

Freeman, C., 1987. Technology and Economic Performance: Lessons from Japan. Pinter, London.

Freeman, C., Clark, J., Soete, L., 1982. Unemployment and Technical Innovation: A Study of Long Waves and Economic Development. Greenwood, Westport, CT.

Furman, J., Porter, M.E., Stern, S., 2002. The determinants of national innovative capacity. Research Policy 31 (6), 899–933.

Griliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 25, 1661–1707.

Hansen, P., Pavitt, K., 1987. The Comparative Economics of Research Development and Innovation in East and West: A Survey. Harwood, Chur, Switzerland.

Hobday, M., 1995. Innovation in East Asia. Elgar, Brookfield, VT.

Hu, M.C., Mathews, J.A., 2005. National innovative capacity in East Asia. Research Policy 34 (9), 1322–1349.

Hu, M.C., 2012. Technological innovation capabilities in the thin film transistor-liquid crystal display industries of Japan, Korea, and Taiwan. Research Policy 41 (3), 541–555.

Hu, M.C., Wu, C.Y., Lee, J.H., Lu, Y.C., 2014. The influence of knowledge source and ambidexterity in the thin film transistor and liquid crystal display industry: evidence from Japan, Korea, and Taiwan. Scientometrics 99 (2), 233–260.

Jaffe, A., Trajtenberg, M., 2002. Patents, Citations & Innovations: A Window on the Knowledge Economy. The MIT Press. Massachusetts Institute of Technology.

Lall, S., 1992. Technological capabilities and industrialisation. World Development 20, 165–186.

Levin, R., Cohen, W.M., Mowery, D.C., 1985. R & D appropriability, opportunity and market structure: new evidence on some Schumpeterian hypotheses. American Economic Review 75 (2), 20–24.

Lundvall, B.Å., Joseph, K.J., Charminade, C., Vang, J., 2011. Handbook of Innovation Systems and Developing Countries: Building Domestic Capabilities in a Global Setting. Edward Elgar Publisher, Cheltenham, U.K.

Mathews, J.A., Hu, M.C., 2007. In: Yusuf, S., Nabeshima, K. (Eds.), Universities and Public Research Institutions as Drivers of Economic Development in Asia, Chapter 5, How Universities Promote Economic Growth. World Bank, Washington DC. USA.

Mowery, D., Rosenberg, N., 1989. Technology and the Pursuit of Economic Growth. Cambridge University Press, Cambridge, UK.

Nelson, R. (Ed.), 1993. National Systems of Innovation: A Comparative Study. Oxford University Press, Oxford, UK.

Niosi, J., 1999. The internationalisation of industrial R & D: from technology transfer to the learning organisation. Research Policy 28, 107–117.

Patel, P., 2000. Technological indicators of performance. In: Tidd, J. (Ed.), Measuring Strategic Competencies. Imperial College Press, London.

Rosenberg, N., Nelson, R., 1994. American universities and technical advance in industry. Research Policy 23, 323–348.

Salter, A., Martin, B., 2001. The economic benefits of publicly funded basic research: a critical review. Research Policy 30, 509–532.

Research and Theory Building in Social Work

Jane F Gilgun, School of Social Work, University of Minnesota, Twin Cities, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Grounded theory and deductive qualitative analysis share an intellectual heritage with social work. Common roots include a focus on lived experiences of research participants, who in social work typically are service users and service providers. They also incorporate other qualities associated with social work, such as understanding persons in context, action orientation, and focus on understanding multiple perspectives through immersion in the field. Research findings, therefore, are useful and usable to practitioners.

Introduction

Grounded theory (GT) and deductive qualitative analysis (DQA) are important ways to do social work research because of their fit with the mission and definitions of social work. Like social work, both focus on human interactions in social environments, multiple meanings and multiple dimensions of human phenomena, immersion in field settings, interpretations grounded in understandings of human experience, and action orientations. Social work's emphasis on social justice springs to life when researchers seek the meanings that research participants attribute to issues that compose their lived experiences. The emphasis of GT and DQA on multiple meanings and perspectives fits not only with human experience in general but also with the worlds of practitioners (Gilgun, 2005a, 2012a,b). Those worlds typically are complicated, untidy, confusing, and often intrusive and traumatizing (O'Sullivan, 2011). The goodness of fit between practice and these two forms of research raises the likelihood that social workers will find the results of GT and DQA research immediately useful.

This article will demonstrate the relevance of GT and DQA to the development of knowledge important to social work practice. Although the focus is on practice, the ideas may be tested for fit with other domains of social work. This article begins with definitions and discussions of social work-specific research and related terms, moves to an overview of the procedures for carrying out these two forms of qualitative research, attempts to clarify confusions over what GT is, and shows the kinds of products that researchers can develop from these two approaches. While there are many ways to develop theory, this article focuses on two, both of which have roots in the Chicago School of Sociology. Social workers made significant contributions to the research methods and methodologies of the Chicago School (Bulmer, 1984; Deegan, 1990; Forte, 2004; Gilgun, 2012a,b; Oliver, 2012). This school flourished in the first part of the twentieth century, and its legacy endures. It is no coincidence that GT and DQA fit assumptions of social work.

Definitions

Terms relevant to this article are social work-specific research, focused research on social work practice, theory, theory building, hypotheses, concepts, conceptual frameworks, evidence, critical realism, deduction, induction, and practice. While there are many definitions for each of them, this article only defines terms as they are used in the present discussion.

There are two broad types of social work research. One is *social work-specific research* and the other is *focused research on social work practice*. The emphasis of published focused research on social work practice is on researcher-defined questions and interpretations; surveys, questionnaires, and other close-ended ways of gathering data; and sophisticated analysis of data that few practitioners and many researchers understand or want to understand. While audiences may have difficulties with the mathematics of this research, it is important research to do because it shows that in specified situations with specified populations certain interventions work. Some call this type of research *empirical*, but there are many definitions of this word, including the idea that any research that is based on data is empirical. Mathematics does not make research empirical. Data are materials that take the form of words and numbers and represent the experiences of research participants.

The following articles exemplify focused research on social work practice: Byrne et al. (2013), Katz and Herskowitz (2012), and Thyer (2001). Many articles published in two US-based publications, *Research on Social Work Practice* and *Social Work Research*, are of this type. Since social work is an intervention-based discipline, establishing a mathematical foundation for its interventions is essential.

Social work-specific research incorporates principles from definitions and missions of social work that are present in codes of ethics and discussions of social work's methodological principles that appear in published articles. These typically focus on the perspectives of participants, the questions become participant defined, and researchers seek to understand multiple perspectives within person–environment interactions. Additionally, this type of research may also be based on researcher immersion into social settings, be explicit about ethics and values such as antioppressive practice and social justice, and be clear regarding the links between general statements about findings and the data on which these are based. This research contributes much to social work's knowledge base, but cannot establish incidence and prevalence, nor can it offer mathematical interpretations. See as examples of the ideas behind this type of research, the codes of ethics of social work in various countries and the work of Dominelli (1996), Forte (2004), Saleebey (2004), and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28086-0

Strier (2007). Articles like these appear regularly in journals such as the *British Journal of Social Work*, *Australian Journal of Social Work*, *Families in Society*, and *Qualitative Social Work*.

In this article, theories are conceptualized as transportable sets of ideas that social workers can apply to particular cases or situations, whether these situations involve individuals, families, groups, communities, nations, policy, or advocacy efforts. Theories illuminate social processes and help observers notice aspects of phenomena that they might otherwise have overlooked. Concepts and their relationship compose theory. Theories are provisional and subject to further testing and modification. Theories typically are based on multiple sources of information, such as research, theory, practitioner expertise, personal experience, experiences of populations of persons whom social work serves, and personal and professional values.

Hypotheses are statements of relationships between concepts. The term is often used interchangeably with *theory* in GT and DQA. In the process of doing the research, hypotheses are called working hypotheses in that they are the ideas with which researchers work as they do the research. Researchers test them for fit and modify them when warranted. The products of a GT or DQA study that seek to develop theory are hypotheses. Typically, researchers discuss the hypotheses, a process often called interpretations, and then they situate the hypotheses and their interpretations within scholarly traditions, showing how the hypotheses fit with, add to, and undermine what is already known.

Concepts are the components of hypotheses and are abstractions from concrete descriptions of processes and other phenomena. Concepts are often sensitizing in that they "suggest directions along which to look" (Blumer, 1986: p. 148) as researchers engage in data collection, analysis, and interpretation. Sensitizing concepts are defined in open-ended ways and are flexible in order to incorporate new meanings as evidence is identified to redefine them. In contrast, definitive concepts have clearly fixed meanings and attributes. Theory-guided research like DQA always uses sensitizing concepts, sometimes as a guide to research and sometimes for the purposes of developing theory.

Conceptual frameworks are a synthesis and integration of existing research and theory relevant to the topic of the research. They are the perspectives with which researchers begin most studies, including DQA studies. Often researchers make statements about their experiences, values, and expectations that are relevant to their focus. In these statements, called reflexivity statements, researchers discuss their personal and professional perspectives and attempt to account for the possible perspectives of research participants, including issues related to power between researchers and research participants (D'Cruz et al., 2007; Gilgun, 2008). These statements typically follow reviews of the literature. Conceptual frameworks are the sources of hypotheses and concepts that researchers use to provide focus and guidance to research. Often researchers show how their reflexivity statements play a part in developing a focus and hypotheses and concepts. The components of theory that researchers use when conducting research are called working hypotheses and working concepts.

Evidence is not raw data, but data that researchers filter through their own worldviews, assumptions, and experiences. This is a critical realist stance that also recognizes that what composes data exists independently of anyone's experience and interpretations of it. Data, as defined earlier, are composed of words and numbers that represent the experiences of research participants including those of researchers. Thus, both DQA and contemporary understandings of GT fit assumptions of critical realism, which is an emerging perspective in social work (Longhofer and Floersch, 2012). See Houston (2001) and Oliver (2012) for further discussion of critical realism in general and in social work in particular.

The term *knowledge* in this article refers to information and understandings that researchers and practitioners develop from their professional practices, from their personal and professional values and personal experiences, and from their understandings of research and theory. Knowledge is information and understandings that researchers and practitioners have used, tested in various situations, and modified to fit new learnings.

Theory building is composed of any number of procedures that lead to the production of theory, where theories are sets of transportable ideas that can be applied to specific situations. Theory building is especially important in social work because of the need to understand complex social situations. Theories and the concepts that compose them can illuminate these situations, even as practitioners are alert for dimensions of situations that theories do not cover and thus seek to individualize interventions.

The word *deduction* has many meanings. For the purposes of this article, deduction refers to processes that involve the use of prior theory, where the concepts and the hypotheses of the theory are the starting points of research. The processes of deduction involve understanding the various connotations of theoretical material and capacities for identifying concrete phenomena such as actions, behaviors, and beliefs as encoded in words and other texts, and as instances of cases that fit the theory and its concepts. A second understanding important to this article refers to, in the words of Dewey (1910), 'complete thinking,' which involves both deduction and induction. Deduction is a process of testing working hypotheses, for the purposes of "confirming, refuting, and modifying" (p. 82) them. Hypotheses are not fixed conclusions, but are open to further testing and change. This second definition fits research whose purpose is theory building and not primarily descriptive.

Induction is the movement from data to concepts, which is difficult to do in a 'pure' way but does represent an attempt to put aside researchers' own views and do active listening and open-minded observations in interviewing, observations, and document analysis in order to discover aspects of phenomena that researchers may not have anticipated.

Practice is any form of social work that involves face-to-face interaction with service users. This includes work in child welfare, children's mental health, shelters, group homes, and residential treatment programs, therapy with families, couples, and individuals, group work, community organizing, and in the many other domains in which social workers ply their craft.

Since GT, DQA, and the perspectives that are part of social work's self-definition have common roots, the kinds of research that GT and DQA are compatible with are social work specific and probably not focused research on practice. DQA and GT,

however, can contribute to focused research by providing information on the meanings of interventions to practitioners and research participants, experiences related to compliance with and fidelity to research protocols, the various adjustments practitioners make to established protocols in order to fit the situations of service users, and the various other conditions under which interventions take place. This is now the well-established view – that mixed methods, which used to be called multiple methods, can provide comprehensive views of social phenomena, including interventions (Creswell and Clark, 2010).

Deductive Qualitative Analysis

Deductive qualitative analysis is a form of research that (1) begins with theory, (2) recommends reflexivity statements, (3) uses negative case analysis as part of its sampling plan, (4) allows choice in terms of how researchers do coding, (5) recommends group analysis of data, (6) organize findings by themes, hypotheses, or concepts, and (7) does documentation and analysis of findings through a three level system of presentation of excerpts from data, researchers' interpretations, and analysis of how findings contribute to what is already known (Gilgun, 2005a, 2012a,b).

Initial Theory

The initial theory is either hypotheses or sensitizing concepts. The theory is typically developed through reflection on conceptual frameworks as defined earlier. By beginning with theory, researchers seek to develop further the theory or use the concepts of the theory to focus the research. The initial theory, whether hypotheses or concepts, is sensitizing in that it illuminates directions along which to look, while at the same time researchers purposefully seek information that will modify the concepts and hypotheses with which they began. They want the products of their research to fit the data. Both concepts and hypotheses, therefore, are provisional, subject to change and modification when evidence warrants it.

Many researchers assume that to do qualitative studies, they have to begin their research with no hypotheses to test and, in so doing, bypass hypothesis testing and theory-guided research. This widespread impression arose from the principles and procedures of early forms of GT where Strauss, Glaser, and Corbin (Glaser and Strauss, 1967; Strauss and Corbin, 1998; Corbin and Strauss, 2008) advised researchers not to begin with theory, but to begin research inductively, with the assurance that eventually theory would emerge through the procedures that they delineated. The developers of GT were concerned that researchers would impose theory on findings, rather than develop theory from findings.

There is good reason to begin research with initial concepts and theory. Researchers interested in particular theoretical models cannot and should not be expected to start anew, or act as if they do not already know something about their areas of interest. They can test theories qualitatively or use concepts to guide their research. Furthermore dissertation committees are unlikely to approve studies that do not build on what is already known nor are funders inclined to commit money to such studies. Even Strauss (1987) softened his earlier stance when he advised researchers who do GT to conduct preliminary studies to write proposals that they intend to submit for funding.

DQA recognizes these issues and provides guidelines for doing research that begins with an initial conceptualization that can range from a parsimonious model or theory to a rather loose set of ideas. I coined the term *deductive qualitative analysis*, after trying out many others, because my experience led me to conclude that having a prior conceptual framework is an important way to do qualitative research (Gilgun, 2005a, 2012a,b). I pieced together the procedures of DQA from earlier work on analytic induction (Cressey, 1953; Lindesmith, 1947; Znaniecki, 1934) and the thinking of philosophers John Dewey (1910) and Karl Popper (1969).

Reflexivity Statements

As discussed earlier, reflexivity statements show respect for researchers' personal and professional experiences from which they develop compelling research questions. Lindesmith (1947) makes an excellent case for the role of experience. They have an important role to play in evidence-based practice (Gilgun, 2005b).

Negative Case Analysis

Imposing theory on findings is poor practice. Beginning a study with no theory is one way to attempt to avoid this, but it is not foolproof. Researchers can add theory to their study at any time and force data to fit the categories of the theory at any point in the research process. DQA uses a procedure called negative case analysis to avoid forcing. Negative case analysis guides researchers purposefully to seek data that undermine their developing theories, rather than seek data that support the theory with which they began. Cressey's (1953) analysis of embezzlement illustrates how to do negative case analysis. Negative case analysis fits with Popper's (1969) idea of conjectures, refutations, and reformulations that Popper believes is the fundamental process of science.

Coding, Analysis, and Interpretation

In DQA, coding, analysis, and interpretation can be done any number of ways, such as the generic three-level codes described in Strauss (1987) and Strauss and Corbin (1998) or the kinds of analysis associated with interpretive phenomenology (Benner, 1994), which includes several kinds of analysis such as the identification of themes and the construction of exemplars and paradigm cases. Through experience, I have found group analysis of data to be effective for testing emerging understandings against the interpretations of others who have different life experiences, training, knowledge, and other perspectives. Group analysis also brings in multiple perspectives. This strategy is traditional, going back at least to Booth's studies of the London poor at the end of the nineteenth century (Webb and Webb, 1932). Olesen et al. (1994) provided an account of group analysis within the Chicago tradition. Group analysis helps to ensure that findings will account for the complexity and variations of social phenomena.

Some Background

In earlier writings, I used the term *analytic induction* to describe what I am now calling DQA (Gilgun, 1995, 2012b). Analytic induction is the procedure that researchers used on classic studies, such as the work of Becker et al. (1961), Cressey (1953), and Lindesmith (1947). I prefer the term *deduction* because induction is not possible as an initial step in the doing of research. In order to sort particular observations from innumerable possible observations, researchers do not begin as blank slates. Something helps them notice particular phenomena and not others. Even Glaser and Strauss (1967), widely considered premier and pioneering inductive qualitative researchers, acknowledge this in a footnote in their well-known text. Induction, then, is a misnomer in this type of research.

I also decided to put together procedures that I call DQA to invite researchers trained in logico-deductive methods to do qualitative research. For too long, the idea that qualitative research does not begin with a conceptual model has been dominant. Social work researchers have a wealth of theory- and research-based knowledge that can be put to use in qualitative studies. Their results will be closely linked to respondents' experiences and may be useful to policy, programs, prevention, and intervention.

Furthermore, I encourage DQA in my teaching and writing to debunk the widespread perception that analytic induction is based on obsolete philosophies of science that guide research toward seeking causal and universal hypotheses that are invariant. The term has a 'spoiled identity' (Gilgun, 2005a, 2012b). Some commentators (Robinson, 1951; Vidich and Lyman, 2000) dismissed analytic induction for this reason. As a result, the term *analytic induction* has rarely been used for about 50 years. Researchers who want to carry out theory-guided qualitative research have few if any guidelines for doing so (Gilgun, 2012b, 2005a).

Grounded Theory

GT has so many forms that researchers must specify the type that they are using. Originally, GT was a method for developing theory that begins in an open-ended way where researchers enter the field in order to identify theoretical concepts that emerge from data (Glaser and Strauss, 1967). The originators were clear that researchers must not impose theory on data but allow the theory to emerge. They also took a stand against 'grand theories' that they viewed were detached from the empirical world. They coined the term *grounded theory* to designate a kind of middle range theory (Merton, 1968) where researchers develop theory so that its links to concrete data are explicit. Theoretical sensitivity, theoretical sampling, and the constant comparative method were and, still are, part of GT. Theoretical sensitivity involves capacities for identifying concrete instances in data as part of preexisting theory (Glaser, 1978). Theoretical sampling involves choosing the next case to investigate on the basis of its promise to expand upon the emerging theory. Constant comparison involves comparing data bits and associated concepts within and across cases in order to build and ground the theory.

As methodologists continued to develop procedures for GT, they made efforts to clarify these procedures. Eventually, doing GT began to appear prescriptive, such as the use of the conditional matrix that Corbin and Strauss (2008) described as "a coding device" (p. 183) and "a conceptual guide" (p. 193) that directed researchers to account for processes such as actions, interactions, and the conditions and consequences of social processes. The matrix was intended to help researchers to account for person–environment interactions and to make decisions about which analytic threads to follow and which situations and persons to sample, in other words, as a guide to theoretical sampling and the development of core categories that became the concepts of the theory under development.

Strauss and Corbin specified other procedures as well, such as the three-level coding scheme of open, axial, and selective coding (Strauss, 1987; Strauss and Corbin, 1998; Corbin and Strauss, 2008). They have an extensive set of terms, which can take time and practice to understand and master. Schneider's (2012) work on women's experiences of childbirth is an example of a traditional approach to analysis using concepts from the conditional matrix. She coded for "properties, dimensions, and relationships between concepts" (p. 217). How she collected her data was not traditional GT because her data set was composed of responses to open-ended questions on an online survey.

By the early 1990s, Glaser (1992) broke away from the directions that Strauss and colleagues were taking. He became concerned that the newer procedures were doing exactly what the first iteration of GT as set forth in the original work (Glaser and Strauss, 1967) was meant to combat, which was imposing theory on data and not allowing theory to emerge. Other methodologists, too, were concerned about the prescriptive nature of GT. Denzin (1997), for example, abandoned GT for decades, because he became swamped in its technicalities. He followed instructions that Strauss delivered in person about the constant comparative method, comparisons across field sites, and the search for emerging theories, indicators of concepts, and links to theory. Denzin instead followed other Chicago School traditions related to the less prescriptive and more descriptive style of research that Chicago methodologists called *fieldwork*.

After 40 years, Denzin (2010) came out as an enthusiastic promoter of a particular form of GT: collaborative, constructivist, and critical. After all those years, Denzin saw that GT does not have to be prescriptive, and researchers can use it in adaptable and open-ended ways to develop descriptive and conceptually rich understandings of human situations. Its open-endedness permits researchers to adapt it to their own particular methodologies and conscious and unconscious biases. Denzin adapted GT to serve his commitment to social justice issues in research, which includes researcher collaborations with participants, the importance of local knowledge, and, once the research is completed, advocacy for social change.

Finally, there are confusions about GT as method and as product. Bryant and Charmaz (2007) distinguished between grounded theory methodology (GTM) and GT. The former marks the use of generic procedures associated with GT and the latter is GT as product. The generic procedures include, among others, the coding scheme and the constant comparative

method. The term *grounded theory* has become a big tent, but it is confusing to label procedures as GTM when they simply are strategies for doing qualitative research and are generic. As Bryant and Charmaz (2007) pointed out, researchers commonly label as GT work done using generic procedures, such as coding or cross-case comparisons (constant comparative method). In the most recent edition of their coauthored book, Corbin and Strauss (2008) shifted toward a more middle position when they stated that use of prior theory is acceptable in GT, as long as the theory is provisional. The authors did not specify how to use prior theory in GT.

Fieldwork and the Chicago Tradition

Glaser, Strauss, and Corbin did not explore or explain the intellectual roots of GT. The brief discussions they had of the Chicago tradition were dismissive, such as disparaging negative case analysis while giving a superficial account of it (cf Glaser and Strauss, 1967). What Denzin now calls GT is a good old-fashioned Chicago School of Sociology methodology. Members of the Chicago School did not name the overall approach to research except to call it fieldwork (Strauss, 1992), although they did name the procedures of analytic induction including negative case analysis and group analysis of data.

Fieldwork is generally descriptive, although infused with conceptual material, and characterized by multiple methods and the other qualities associated with Chicago research, as already discussed. Examples abound: Frazier (1932) and Thomas and Znaniecki (1918–1920/1927). Bogdan and Biklen (2007), among others, have been a force in keeping the traditions of Chicago descriptive research alive.

Applications to Practice Research

GT and DQA can facilitate the development of theory and conceptually rich descriptive research useful in many types of social work practice settings. The kind of research that these approaches facilitate is social work-specific research, which incorporates principles associated with social work's self-definitions and that in turn has connections to the Chicago School of Sociology.

I searched the database, *Social Work Abstracts*, to see what products researchers have developed using GT and DQA. I used the search term *grounded theory*, and the search resulted in more than 200 articles. The term *deductive qualitative analysis* yielded naught. Yet, as I read the articles, I saw that several researchers who said they did GT actually used procedures that fit definitions of DQA, while others did descriptive research that in the Chicago tradition is fieldwork. Some were unclear about methods to the point that it was difficult to know the procedures they followed, although they wrote that GT was part of the mix.

In this section, I note this state of confusion and attempt to clarify. At the same time, I want to show what DQA and GT bring to social work research. This is a dual agenda, but I could not figure out how to talk about products without also noting whether DQA or GT is associated with the research I discuss.

The following citations are representative articles from my search using the term *grounded theory*. These articles presented descriptive research whose purpose was to understand social workers' experiences with specific populations (Saini et al., 2012), the experiences of specific populations who are in the care of social workers (Barney and Buckingham, 2012), concept development (Leece and Peace, 2010), and research whose purpose is to influence public policy and programming (Leece and Peace, 2010). These articles exemplify social work-specific research and are not focused research on social work interventions.

Some of the articles fit DQA better than GT. Others were labeled GT, but the researchers did not produce a theory and most did not report beginning the research in an open-ended way, both defining hallmarks of GT. Few if any used the terms *theoretical sampling* and *constant compassion*. They used coding schemes associated with GT, but coding schemes do not make a GT study (cf Bryant and Charmaz, 2007).

In some cases, researchers reported that they did a GT study, but their methods are exemplars of DQA. For example, Leece and Peace (2010) began their research with concepts and sought to further develop these concepts. They called their study *grounded theory* but its features are closer to definitions of DQA than GT. Their work fits definitions of DQA and not GT because they began with theory, their purpose was to develop theory, they focused on the theory throughout the research process, and their findings were elaborated theoretical concepts. They did not do negative case analysis, which is important to do when beginning with theory, and they did not have a reflexivity statement, which is recommended in DQA.

There are many other examples of social work research that authors label as GT when they actually use the procedures of DQA, e.g., Barney and Buckingham (2012) who sought to develop the concept of spirituality. The confusion between GT and GTM as Bryant and Charmaz (2007) noted is present in social work research as is confusion between DQA, GT, and GTM.

Conclusion

GT and DQA are important ways to do social work-specific research. Their products are closely linked to the lived experiences of service users and service providers, making findings useful and usable to practitioners. The usefulness of GT and DQA is related to an intellectual heritage they share with social work.

See also: Antisocial Behavior; Critical Social Work Practice; Social Work Ethics; Social Work Theory.

Bibliography

Barney, R.J., Buckingham, S.L., 2012. HIV/AIDS in a South African township. Journal of Religion and Spirituality in Social Work 31, 51–66.

Becker, H.S., Geer, B., Hughes, E.C., Strauss, A.L., 1961. Boys in White. University of Chicago, Chicago.

Benner, P. (Ed.), 1994. Interpretive Phenomenology. Sage, Thousand Oaks, CA.
Blumer, H., 1986. What is wrong with social theory? In: Blumer, H. (Ed.), Symbolic Interactionism. University of California Press, Berkeley, pp. 140–152.
Bogdan, R.C., Biklen, S.K., 2007. Qualitative Research for Education, fifth ed. Pearson, Boston.
Bryant, A., Charmaz, K. (Eds.), 2007. Sage Handbook of Grounded Theory. Sage, Thousand Oaks, CA, pp. 1–28.
Byrne, S., Salmela-Aro, K., Read, S., Rodrigo, M.J., 2013. Individual and group effects in a community-based implementation of a positive parenting program. Research on Social Work Practice 23 (1), 46–56.
Bulmer, M., 1984. The Chicago School of Sociology. University of Chicago Press, Chicago.
Charmaz, K., 2006. Constructing Grounded Theory. Sage, Thousand Oaks, CA.
Corbin, J., Strauss, A., 2008. Basics of Qualitative Research, third ed. Sage, Thousand Oaks, CA.
Cressey, D., 1953. Other People's Money. Wadsworth, Belmont, CA.
Creswell, J.W., Clark, V.L.P., 2010. Designing and Conducting Mixed Methods Research. Sage, Thousand Oaks, CA.
D'Cruz, H., Gillingham, P., Melendez, S., 2007. Reflexivity, its meanings and relevance for social work: a critical review of the literature. British Journal of Social Work 37 (1), 73–90.
Denzin, N.K., 1997. Coffee with Anselm. Qualitative Family Research 11 (1&2), 1–4. http://www.scribd.com/doc/27352636/Coffee-with-Anselm. Downloaded 4 February 2014.
Denzin, Norman K., 2010. Grounded and indigenous theories and the politics of pragmatism. Sociological Inquiry 80 (2), 286–312.
Deegan, M.J., 1990. Jane Addams and the Men of the Chicago School of Sociology, 1892–1918. Transaction, New Brunswick.
Dewey, J., 1910. How We Think. Prometheus, Amherst, NY.
Dominelli, L., 1996. Deprofessionalizing social work: anti-oppressive practice, competencies and postmodernism. British Journal of Social Work 26, 153–175.
Forte, J.A., 2004. Symbolic interactionism and social work: a forgotten legacy. Part 2. Families in Society 85 (4), 421–530.
Frazier, E.F., 1932. The Negro Family in Chicago. University of Chicago Press, Chicago.
Gilgun, J.F., 1995. We shared something special: the moral discourse of incest perpetrators. Journal of Marriage and the Family 57, 265–281.
Gilgun, J.F., 2005a. Qualitative research and family psychology. Journal of Family Psychology 19 (1), 40–50.
Gilgun, J.F., 2005b. The four cornerstones of evidence-based practice in social work. Research on Social Work Practice 15 (1), 52–61.
Gilgun, J.F., 2008. Lived experience, reflexivity, and research on perpetrators of interpersonal violence. Qualitative Social Work 7 (2), 181–197.
Gilgun, J., 2010. Methods for enhancing theory and knowledge about problems, policies, and practice. In: Briar, K., Orme, J., Ruckdeschel, R., Shaw, I. (Eds.), Sage Handbook of Social Work Research. Sage, London, pp. 281–297.
Gilgun, J.F., 2012a. Enduring themes in qualitative family research. Journal of Family Theory and Review 4, 80–95.
Gilgun, J.F., 2012b. Hand into glove: grounded theory, deductive qualitative analysis and social work research and practice. In: Fortune, Anne E., Reid, William, Miller, Robert (Eds.), Qualitative Methods in Social Work, second ed. Columbia University Press, New York, pp. 107–134.
Glaser, B., 1978. Theoretical Sensitivity. Sociology Press, Mill Valley, CA.
Glaser, B., 1992. Basics of Grounded Theory Analysis. Sociology Press, Mill Valley, CA.
Glaser, B., Strauss, A.L., 1967. The Discovery of Grounded Theory. Aldine, New York.
Houston, S., 2001. Beyond social constructionism: critical realism and social work. British Journal of Social Work 31, 841–861.
Katz, C., Herskowitz, I., 2012. Repeated interviews with children who are the alleged victims of sexual abuse. Research on Social Work Practice 23 (2), 210–218.
Leece, J., Peace, S., 2010. Developing new understandings of independence and autonomy in the personalised relationship. British Journal of Social Work 40, 1847–1865.
Lindesmith, A.R., 1947. Opiate Addiction. Principia, Bloomington, IN.
Longhofer, J., Floersch, J., 2012. The coming crisis in social work: some thoughts on social work and science. Research on Social Work Practice 22 (5), 499–519.
Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.
Olesen, V., Droes, N., Hatton, D., Chico, N., Schatzman, L., 1994. In: Bryman, A., Burgess, R.G. (Eds.), Analyzing Qualitative Data. Routledge, London, pp. 111–128.
Oliver, C., 2012. Critical realist grounded theory. British Journal of Social Work 42, 371–387.
O'Sullivan, T., 2011. Decision-Making in Social Work, second ed. Palgrave-Macmillan, Basingstoke.
Popper, K.R., 1969. Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge and Kegan Paul, London.
Robinson, W.S., 1951. The logical structure of analytic induction. American Sociological Review 16, 812–818.
Saini, M., Black, T., Lwin, K., Marshall, A., Fallon, B., Goodman, D., 2012. Child protection workers' experiences in working with high conflict separating parents. Children and Youth Services Review 34, 1309–1316.
Saleebey, D., 2004. 'The power of place': another look at the environment. Families in Society 85 (1), 7–16.
Schneider, D.A., 2012. The miracle bearers: narratives of birthing women and implications for spiritually informed social work practice. Journal of Social Service Research 38, 212–230.
Strauss, A.L., 1987. Qualitative Analysis for Social Scientists. Cambridge University Press, New York.
Strauss, Anselm, 1992. A personal history of grounded theory. Qualitative Family Research 5 (2), 1–2. http://www.scribd.com/doc/44659994/Anselm-Strauss-Writes-A-Personal-History-of-Grounded-Theory-Other-Articles. Downloaded 4 February 2014.
Strauss, A., Corbin, J., 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, second ed. Sage, Thousand Oaks, CA.
Strier, R., 2007. Anti-oppressive research in social work: a preliminary definition. British Journal of Social Work 37 (5), 857–871.
Thomas, W.I., Znaniecki, F., 1918–1920/1927. The Polish Peasant in Europe and America, vol. 1–2. Knopf, New York. First published in 1918–1920.
Thyer, Bruce, 2001. What is the role of theory in research on social work practice? Journal of Social Work Education 37 (1), 9–25.
Vidich, A.J., Lyman, S.M., 2000. Qualitative methods: their history in sociology and anthropology. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research, second ed. Sage, Thousand Oaks, CA, pp. 37–84.
Webb, S., Webb, B., 1932. Methods of Social Study. Longman, Greens.
Znaniecki, F., 1934. The Method of Sociology. Farrar and Rinehart, New York.

Research Ethics, Cross-Cultural Dimensions of

Leonardo D de Castro, Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
Peter A Sy, College of Social Sciences and Philosophy, University of the Philippines, Quezon City, Philippines
Nur IB Soonaan, Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Tuzin, volume 3, pp. 13231–13235, © 2001, Elsevier Ltd.

Abstract

Collaboration across countries poses challenges relating to the understanding of the objectives and outcomes of research studies as well as the ethical issues arising from the engagement of researchers and study populations with varying cultural backgrounds. Cross-cultural dimensions pose difficulties for researchers' understanding and observance of ethical and legal considerations. These considerations may apply to ways of dealing with research subjects, handling data, research design and methodology, and negotiating authorship credits. As concepts such as autonomy, privacy, and informed consent take on variable meanings, researchers should take the variability into account in order to minimize ethical complications.

Introduction

Research ethics encompasses norms of conduct and concomitant discourse pertaining to the delineation of good or acceptable from bad or unacceptable behavior in research. Some norms (e.g., respect for life) are widely regarded as universal, as they cut across different societies. A few are arguably more pervasive and more commonly observed (e.g., 'human rights' as articulated in the Universal Declaration of Human Rights) in some cultures than in others. A study of these norms – as they are practiced in, or applied to research – is also within the scope of research ethics.

Particular cultures are usually defined in terms of socially transmitted behavioral patterns, including beliefs, arts, religions, institutions, ways of life, thinking and knowing, and other products of human labor. With boundaries that are flexible and permeable and features that are dynamic, one culture can neither be adequately described without implicating another, especially today in an era of globalization and ubiquitous information and communication technologies (Brody, 2009), nor is a culture adequately describable without implicating other societal domains like politics and economics. Cultures of large populations are so complex that they also support diverse ethnicities and dynamic subcultures. Thus, 'cross cultural' could refer to cultures not only across but also within societies.

In this article, cross-cultural dimensions of research ethics are described in three different but interrelated domains: contextual concerns, conduct of research and research dissemination, and international initiatives addressing cross-cultural aspects of research.

Contextual Concerns

Differences among researchers, among research participants, and between researchers and participants are complex, multidimensional, and polycontextual. Such differences are amplified by the growth in international research as well as by long-standing divides of culture, research and scientific traditions, and diverse regulatory regimes of research.

Growth of International Collaborative Research

The increasing number of research publications with multicountry authorship clearly indicates the correlative growth of multicountry research collaboration. In turn, the growth of research collaboration across countries poses challenges relating to the understanding of the outcome of the studies as well as the ethical issues that may arise because of the engagement of researchers and study populations belonging to different nationalities with varying cultural backgrounds. The cross-cultural dimension makes it imperative for researchers to be able to navigate cultural boundaries in a way that enables them to conform to ethical and legal considerations. In practice, these considerations may apply to ways of dealing with research subjects, handling data, deciding on research design and methodology, managing workloads and working habits, and negotiating authorship credits, especially when different research teams and different research sites are involved.

For example, research design can pose many difficulties, including those that pertain to the meaning of concepts and the language that is used. Thus, in health-related social research, there may be discrepancies in the meaning given to words like 'disease' or the names given to particular diseases. There are bound to be variations in the understanding of conditions such as dementia, leprosy, or HIV/AIDS, particularly because of the emotive (if not pejorative) content that these words can have in different social contexts. Certain scientific concepts (like placebo and randomization) can also defy translation into local languages. Insistence on translation may even complicate communication between researchers and research participants (Tindana et al., 2006).

Cultural and linguistic differences as well as the ethical and legal issues that these give rise to can have profound roots. Moral values grow together with a society's sense of right and wrong. A particular society's values and sense of right and wrong are informed by the cultural context in which they grow. Researchers need to be mindful not only of the differences in cultural contexts in which research is carried out but also of the deeper roots that these cultures have in moral values and a communal sense of right and wrong.

East–West

One contrast that is often cited to illustrate cultural differences is that between East and West. There are those who hold the view that what may be regarded as Western medical research has been guided by prevailing moral philosophies in the West; Eastern medical research has had roots in the prevailing moral philosophies in the East. Put in this way, the view is obviously oversimplified. However, it could be useful in paving the way for a more specific and nuanced deconstruction of complicated circumstances.

Cultural values affect ethical standards not only by virtue of their having a Western or Eastern character but also in many other respects. The boundaries between Western and Eastern values are seldom clear insofar as they apply to research subjects or research contexts. Geographical boundaries have proven to be very permeable and cultural influences have penetrated boundaries very easily, courtesy of travel and communication technologies. Western moral philosophies have influenced Eastern ethical standards and Eastern moral philosophies have influenced Western ethical standards. Moreover, cultural values and traditions are emerging from a collective sense of right and wrong in communities and social groups that are not necessarily bound geographically, although geographical proximity is still a factor that can influence the values that people acquire.

Insofar as research and science have to contend with value differences between and among researchers and subjects or participants, they are inevitably cross-cultural. While there are value differences, there are also commonalities. Given the porous, dynamic, and nonimmutable nature of cultures, we are seeing more and more cultural hybridization. For the sake of naming convenience and simplification, however, references here are made to 'Western' and 'non-Western' cultures more as dominant tendencies rather than as hard and fast categories. Complicating the understanding of these terms are economic and technological dimensions of expressions like 'developed' and 'developing' as well as 'first world' and 'third world.' While not interchangeable with 'Western' and 'non-Western,' 'developed' and 'first world' countries often subscribe to dominant 'Western' standards. These are to be understood not so much as oppositional terms but rather as dynamic tendencies in light of cultural hybridization in the social world.

Research and Science as Culture

The US Code of Federal Regulations (CFR) Title 45 Part 46 defines 'research' as "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge." Operationally 'research' is inseparable from science as both have to do with knowledge of the world and procedures for obtaining such knowledge. However, the definition of 'scientific research' is far from neutral. Insofar as research and science are highly organized human social activities, they take off from certain cultural or ethnic assumptions. Among these assumptions are values usually deemed 'universal' by the dominant culture. But some practitioners have pointed to the ethnographic features of science (as championed in many research institutions today) marked by specific conventions, boundary techniques, and values (Franklin, 1995). In the face of differences and variations in practice, instruments of control in the form of laws, declarations, conventions, and guidelines have been promulgated. In turn, these have given rise to behavioral expectations for the conduct of research throughout the world.

However, the resulting ethical codes and guidelines (and the research paradigms that inform them) are increasingly deemed inadequate in addressing the issues arising from cultural differences in research practice (Marshall and Batten, 2003). "Research ethics" norms (as currently articulated in international instruments) are facing the challenge posed by local communities that see research norms as a form of 'local knowledge.' Thus, some of these norms are encountering resistance as they are applied beyond the 'local' culture and regulators or practitioners seek their broader implementation. When international instruments are presumed to have universal applicability or absolute character, tension with the unique cultural features that form part of a local culture's identity arises (Christakis, 1992).

Internationalization and Regulation of Research

To many, research ethics signifies regulation. If one is to be an ethical researcher, one has to be able to comply with regulations intended to ensure ethical conduct. Conversely, there is the view that the lack of regulation has allowed researchers to behave unethically. Regulation in the way that it is known today was an offshoot of the findings at the Doctors Trial of 1946–47, which happened within the context of the Nuremberg Trials for Nazi war criminals. The resulting document, the Nuremberg Code (1949), outlines the ethical principles for research involving human subjects, including voluntariness, informed consent, beneficence, and nonmaleficence. Succeeding codes or regulations, like the World Medical Association's (2013) Helsinki Declaration and the US Code of Federal Regulations (CFR) Title 45 Volume 46, have been influenced by the Nuremberg Code. The cultural import of such development lies in the adversarial context of the origins of these regulations in the West. Its narrative has been about unscrupulous researchers doing experiments that sacrifice human concerns in the fulfillment of debatable scientific goals. A set of research ethics codes is therefore needed to help prevent these researchers from harming human subjects. The narrative forms part of the "human subjects protection" ethos of international regulation of research. Essentially researchers are to demonstrate before ethics committees that, in part, they are not out to harm people. While the injunction against harming research subjects may be inescapable in international research regulation, its overarching narrative, and adversarial context of origin could be antithetical to relationship building in some non-Western cultures.

Initial efforts at the ethical regulation of research showed a tendency for consensus building. However, as implementation of international declarations and guidelines spread beyond the need to address the catastrophic consequences of war-time research involving human subjects, the scientific community gradually realized the necessity of coming to terms with the cultural nuances that made universal norms complicated in their practical application.

A major concern in the internationalization and regulation of research is the applicability of foreign codes and guidelines

to local settings or their compatibility with local laws, rules, practices as well as with local institutional mandates and priorities. For example, genetics research projects have been undertaken in rural areas in developing countries that encounter difficulties when observing U.S. Institutional Review Board (IRB) guidelines because of differences in sociocultural norms, religious and other traditions, language, and geography (Sampson et al., 2013). Such differences challenge the applicability of IRB concepts and principles in non-Western societies. In comparison, codes and guidelines that are intended to apply globally such as those originating from the World Health Organization – Council for International Organization of Medical Sciences, the United Nations Educational, Scientific and Cultural Organization, and the World Medical Association (2013) are usually subjected to broad international consultations that take cultural differences into account before they are issued. Nevertheless, occasional challenges can still be encountered in some cultural contexts.

Some of the challenges have to do with the application of guidelines pertaining to privacy. International guidelines and codes of ethics require health research to respect the confidentiality of health information. However, rigid observance of privacy requirements in patient-healer relationships has been shown in ethnographic research to lead to discomfort among research participants (Monshi and Zieglmayer, 2004). Insistence on 'privacy' could reflect an ethnocentric bias that does not carry well in non-Western contexts where what would otherwise be regarded as 'private' and 'confidential' information could readily be shared according to traditional practices.

There could also be a limited understanding of the relationship between ethical guidelines and local research ethics committee regulations (Bhutta, 2002). Where there is a limited capacity among the locals to conduct research, international researchers are more likely to dictate the research agenda. Hence, research questions could easily be more about problems of people from sponsoring countries and less about raising local standards of care. It is easy for such research to gloss over the significance of local community participation and the sharing of benefits with the host community.

Conduct of Research and Research Dissemination

Informed consent is generally regarded as the bedrock of permissible conduct of Western scientific research involving human subjects. It is important to understand the cultural context where informed consent is sought and given. Because informed consent has been a crucial factor in the ethical implementation of research, it has been the focus of much discourse pertaining to the significance of cultural differences in appreciating research ethics. For example, a lot has been said about "inadequacies" in the implementation of informed consent requirements in some developing countries. The contrast between 'developed' and 'developing' countries is ordinarily understood in relation to economic progress. Yet, the economic differences are also intertwined with cultural features, some of which have economic roots themselves. Informed consent procedures often are institutionalized and well-regulated in developed countries. Regulations can be so specific that researchers are left with little room to improvise when informed consent procedures, tried and tested in developed countries, have little chance of success in developing countries. Problems have arisen in situations where developing countries or communities are determined to preserve crucial aspects of their identity and are thus unwilling to yield to foreign informed consent conventions that tend to undermine their culturally determined values and practices.

Autonomy is a crucial factor in determining the validity of informed consent. The exercise of autonomy is generally regarded as a necessary condition for the exercise of free consent. It is in relation to the individual subject that autonomy is commonly gauged. The individual person is the subject of assessment when efforts are made to determine whether there is sufficient understanding of relevant information given. However, the individual is not always regarded as the basic social unit for decision-making in every given community or society, even with respect to matters regarding a particular individual as a subject of research. Thus, developed country researchers may be making a mistake when they invoke autonomy in behalf of an individual in a developing country where the family – rather than the individual – is regarded as the basic unit of society.

In developed countries, the informed consent process requires the involvement of individual participants, except where these participants are not legally competent. Even when representatives of a community are involved (e.g., in population studies), individual consent is still a requirement. The rule generally observed is that community consent cannot override individual consent. Community consent is often regarded as a necessary but not a sufficient condition to validate consent. However, in many developing country contexts, the starting point for the informed consent process has to be a social unit that is bigger than the individual. In resource-challenged developing country situations, individual lives could be so interdependent (e.g., economically, in matters of home-based health care, or in many aspects of day-to-day activities) that people feel they cannot afford to be making decisions on their own. Individual participants may find themselves confronted with a research project that has been approved in their behalf already by their representatives. In theory, the participants' consent may be individually required; in practice, they either defer to the community decision as they are used to, or they find community pressure irresistible, but also as an inevitable feature of communal life.

Overlapping with the developed–developing country duality is the contrast often made between Western and non-Western moral philosophies. The individual is regarded as the moral agent who is expected to make decisions, make those decisions on the basis of reasons, and assess those reasons as an act of self-determination. This is thought to provide the framework not only for evaluating the ethics of research but also for undertaking research itself.

On the other hand, many non-Western societies have situated informed consent within the framework of a community decision-making process. This is why decisions to participate in biomedical research have had to be made by families rather than solely by an individual.

Cultural nuances are also significant in dealing with documentary requirements and signatures for informed consent. In resource-challenged societies where illiteracy is common, the

value of written documents and signatures is limited. The assertion of an individually centered decision to participate or not in research could result in problems that may linger long after a research project has been completed and the researcher has left the research site. Researchers may find it difficult to insist that informed consent documents be signed individually. Lengthy and complicated forms are difficult to process for those who may be unduly pressured to affix signatures. One could see how such a value (or disvalue) does not contribute to the flourishing of an individual or of the community to which that individual belongs.

Moreover, signatures and the ritual of signing have specific meanings for people in some developing country contexts. One must understand those meanings and consider alternative procedures for consent taking that are sensitive to those cultural peculiarities. Various ethnic groups may perceive the signature requirement as a challenge to a cultural perspective that is founded on the reliability of the spoken word.

Human subjects may also have other reasons to be intimidated by the idea of having to sign consent forms. There have been cases where people belonging to certain communities could not readily accept the idea of signing consent documents because they had previously lost parcels of land by affixing their signatures to a document (Creed-Kanashiro et al., 2005). There are also accounts of people developing a fear of signing documents because of negative consequences that doing so entailed when they were still living under communist regimes (Liamputtong, 2010).

In relation to these examples, being developed or developing are used with reference to economic conditions rather than to cultural or social values. It would be easy, but wrong, to understand these terms to refer to ethical values or even to research capacity. The research practices associated with developing countries are not necessarily lacking in development. Neither are the cultural values that may be reflected in those practices. The rejection of written documentation among small ethnic groups should not be treated condescendingly as a manifestation of backwardness.

In addition to the examples cited above, the refusal to provide acceptable documentation of consent could be indicative of a high regard for the spoken word ("I need not affix my signature if I have already given my word") or of a fear of the possibility of the signed document being used against the signatory in a negative way in future. In these contexts, 'developed' is not necessarily better than 'developing.' 'Developing' is not necessarily just a stage leading to the 'developed' world, as those in developed countries know it at present. There are many regulations that cannot simply be applied and many practices that cannot simply be adopted in developing countries just because these have been successful in developed countries. Adaptation is necessary and researchers have to be creative enough to try alternative procedures. In dealing with illiterate populations, researchers have used audio-visual aids such as photographs, diagrams, and video clips to explain a research project's aims, methods, and procedures. Video recordings have also been used to validate verbal consent when the latter is obtained in lieu of a signed written document. Taking steps such as these helps ensure that challenges to the cultural aspects of ethical research are overcome.

Content and Understanding of Consent Documents

There are also problems that pertain to the content and understanding of consent documents. Obviously, it is important for the informed consent process to be documented using the signature of the participant and under the guidance of ethics review committees. However, the information given, especially across cultures, can be misunderstood or misinterpreted. The translation from one language to another could distort meaning or significance and give rise to a decision regarding participation that is based on distorted information or on a misguided assessment of benefits and risks. Studies have already shown that even within a single country using a single language the information regarding a research proposal could easily be misunderstood. This danger is increased when research is carried out across national boundaries and participants have to try to understand a translation.

Research Publication and Utilization

The conduct of research and research dissemination are essentially two sides of the same scientific enterprise. Once research is conducted, the expectation is for researchers to publish study results in peer-reviewed journals. In this 'assembly line' of knowledge production, journal publishers become *de facto* knowledge gatekeepers.

Journal Publishers as Gatekeepers

Medical journal editors have played the role of ethics gatekeepers by requiring that authors of articles submitted for publication should declare that their research project had prior approval by an ethics review committee and that the investigators complied with the guidelines expressed in the Helsinki Declaration. By requiring compliance with international ethical guidelines intended to protect human research subjects, editors and publishers have encouraged investigators to observe culturally sensitive ethics requirements, while at the same time supporting international collaboration indirectly. Some journals have promoted culturally sensitive manuscript submission policies by requiring consent for publication of case reports or case series regardless of the country from which the manuscript was submitted, giving human subjects a chance to review information pertaining to them before being released for publication (Bevan and Miller, 2005).

While investigators seem to be able to respect confidentiality of information regarding individuals without much difficulty, they could more easily overlook confidentiality requirements for information about groups of individuals or communities. For instance, one may come across a study reporting that a worryingly high prevalence rate for urinary tract infections has been observed among women in a particular community or that lorry drivers plying a certain route in one country have been found to visits brothels at an alarming rate during overnight stops.

In response to the above concerns, some have suggested that sensitive research data should not be published when there is a request to that effect from the community. An example is given of a study among the Hmong diaspora in Australia where an investigator was asked not to write and elaborate about certain aspects of rituals that could harm the cultural integrity of the community. A suggestion is thus made for researchers to

avoid publishing sensitive identifiable information that may put the participants at risk of unwarranted scrutiny or even physical and social harm by authorities or security forces. The investigator further recommends the use of pseudonyms or the adoption of fictional names for places in order to help protect subjects. In some cases, the omission of some research data from a publication may be necessary to minimize danger to research participants (Liamputtong, 2010).

Assessing Benefits of Research: Cross-Cultural Aspects

In research involving human subjects, the expectation does not end in publication. Direct benefits to study participants are looked into.

The contrast between developed and developing country cultures shows up in the assessment of benefits resulting from health-social science research. One study done in Mexico revealed that while research subjects derive benefits from the studies in which they assume the risks, these benefits are minimal when compared with health, financial, and academic gains that accrue to sponsors and future patients in resource-rich countries. The interaction between developed country sponsors and investigators and developing country research subjects reveals a culture of exploitation (Páez and Garcia De Alba, 2009).

In this particular example, the investigators reviewed 34 multinational, multicenter research protocols and found that the protocols did not specify benefits for the host institution or for the host community. There was also no reference to post-trial access to successfully researched therapies. These were expected to be too expensive for the Mexican population in general, making them culturally irrelevant. The authors also claimed that the studies did not address the greatest health priority for the country, which was insufficient health care coverage (Páez and Garcia De Alba, 2009).

The significance of culturally mismatched benefits has been observed elsewhere, leading to the observation that relevant and meaningful health research in developing countries has to be able to promote equity and develop local bioethics capacity, especially in the face of continuing globalization and the genomics revolution (Bhutta, 2002).

International Initiatives Addressing Cross-Cultural Concerns of Research

Given the complexities involved in social and behavioral research, it would be almost impossible to find clear-cut solutions to the cross-cultural concerns discussed above. However, there have been various initiatives to develop international consensus at the level of general principles. The most prominent of these have been associated with guidance mechanisms to help promote a culturally sensitive implementation of research projects and protect the welfare of research participants across societies.

International Instruments

Faced with cultural concerns complicating the assessment of ethical issues in research, international agencies and organizations have worked out guidelines to try to address the difficulties from a multicountry perspective. Some of the issues that have been identified are being addressed by regulatory instruments such as UNESCO's Universal Declaration on Bioethics and Human Rights (2005), which deals with ethical issues related to medicine, life sciences, and associated technologies as applied to human beings, taking into account their social, legal, and environmental dimensions. Providing guidance to UNESCO's member states, it provides guidance regarding decisions or practices of individuals, groups, communities, institutions and corporations, public and private. Article 12 highlights the importance of giving due regard to cultural diversity and pluralism while at the same time warning against invoking such considerations in a manner that infringes upon human dignity, human rights, and fundamental freedoms.

The Declaration also provides that multicountry research should be the object of an appropriate level of ethical review in the host countries and the country in which the funder is located. In addition, the Declaration maintains that transnational health research should be responsive to the needs of host countries, and that the importance of research contributing to the alleviation of urgent global health problems should be recognized.

The assessment of research benefits could also be distorted by a failure to take into account cross-cultural differences and disparities among research populations. Thus, we can understand the Universal Declaration's position that terms for collaboration and agreement on the benefits of research should be established with equal participation by researchers, research communities, funders, and any other party to the negotiation.

The International Ethical Guidelines for Biomedical Research Involving Human Subjects (2002) issued by the Council for International Organizations of Medical Sciences (CIOMS) is also particularly sensitive to cultural differences relating to the ethical practice of research. Dealing primarily with biomedical research, the guidelines aim to reflect the conditions and the needs of low-resource countries involved in multinational or transnational research. CIOMS emphasizes the need to obtain research permission from a community leader, a council of elders, or another designated authority and thereby show respect for local customs. The Guidelines also note that since many people in all cultures are unfamiliar with, or do not readily understand, scientific concepts such as those of placebo or randomization, sponsors, and investigators should develop culturally appropriate ways to communicate information that is necessary for adherence to the standard required in the informed consent process, and should describe and justify in the research protocol the procedure they plan to use in communicating information to subjects. Referring specifically to collaborative research in developing countries, CIOMS says that the research project should include the provision of resources to ensure that informed consent can indeed be obtained legitimately within different linguistic and cultural settings.

As for research in populations and communities with limited resources, CIOMS provides that in a population or community with limited resources, the sponsor and the investigator must make every effort to ensure that:

- the research is responsive to the health needs and the priorities of the population or community in which it is to be carried out; and

- any intervention or product developed, or knowledge generated, will be made reasonably available for the benefit of that population or community.

Responding to Local Healthcare Needs

Also reflecting an emerging international consensus regarding research involving people in low and middle income countries, the Wellcome Trust Guidance Notes (2013) on Research Involving People in Low and Middle Income Countries reinforce the view that investigators must demonstrate that their research is responsive to the healthcare needs or priorities within that country. Taking the example of malaria, tuberculosis and HIV/AIDS, the Trust says that the outcomes of such research should be likely to lead to relevant and sustainable health benefits to people in those areas where the research is undertaken. In this regard, it considers research to have local applicability where:

- insofar as it can be estimated, the prospective treatment will or might foreseeably become available to local patients on a sustainable basis;
- the prospective treatment will or might foreseeably become affordable locally; or
- prospective treatments could be delivered within existing structures, or structures that are to be (or could readily be) developed. For example, if a prospective treatment involved administering medication via a particular medical procedure, local healthcare staff would need to have appropriate expertise. If such expertise were neither available nor going to become available in the foreseeable future, careful consideration would need to be given as to how appropriate training should be arranged.

Also relevant in this respect are guidelines issued by UNESCO's Management of Science and Technology Program (MOST) (1994) aiming to draw researchers' attention to certain areas in which conflicts between ethical principles and aims of the research might arise, provides that research should be carried out in full compliance with, and awareness of local customs, standards, laws and regulations. MOST stresses the need for all researchers to show familiarity with, and respect for the host culture. Where researchers are studying cultures, countries, and ethnic groups other than their own, they should make their research objectives particularly clear and remain aware of the concerns and welfare of the individuals or communities to be studied. Additionally, participants should be offered access to research results, presented in a manner and language they can understand.

See also: Asia, Sociocultural Overviews: Korea; Bioethics; Common Rule: US Regulatory Protections for Human Research Participants; Cultural Critique, Anthropological; Cultural Relativism, Anthropology of; Culture: Contemporary Views; Development: Social-Anthropological Aspects; Diaspora; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Family: The Anthropology of the Concept and Its History; Fieldwork in Social and Cultural Anthropology; Fieldwork: Ethical Aspects; Folklore; Gender and Social Research; Health: Anthropological Aspects; Identity in Anthropology; Inequality: Comparative Aspects; Informed Consent of Research Subjects; Japan: Sociocultural Aspects; Objectivity of Research: Ethical Aspects; Poverty, Culture of; Privacy of Individuals in Social Research: Confidentiality; Research Ethics: Research; Research Funding: Ethical Aspects; Research Publication: Ethical Aspects; Southeast Asia: Sociocultural Aspects; Third World; Values, Anthropology of; Youth Culture, Anthropology of.

Bibliography

Bevan, J.C., Miller, D.R., 2005. Medical journals and cross-cultural research ethics. Canadian Journal of Anaesthesia 52 (10), 1009–1016. Available at: http://link.springer.com/article/10.1007%2FBF03021597.

Bhutta, Z.A., 2002. Ethics in international health research: a perspective from the developing world. Bulletin of the World Health Organization 80 (2), 114–120. PubMed Central PMCID: PMC2567726.

Brody, H., 2009. The Future of Bioethics. Oxford University Press, New York.

Christakis, N., 1992. Ethics are local: engaging cross-cultural variation in the ethics for clinical research. Social Science & Medicine 35 (9), 1079–1091. http://dx.doi.org/10.1016/0277-9536(92)90220-K.

CIOMS, 2002. International Ethical Guidelines for Biomedical Research Involving Human Subjects. Available at: http://www.cioms.ch/publications/layout_guide2002.pdf.

Creed-Kanashiro, H., Oré, B., Scurrah, M., Gil, A., Penny, M., 2005. Conducting research in developing countries: experiences of the informed consent process from community studies in Peru. Journal of Nutrition 135 (4), 925–928. Pubmed PMID: 15795463.

Franklin, S., 1995. Science as culture, cultures of science. Annual Review of Anthropology 24 (1), 163–184. http://dx.doi.org/10.1146/annurev.an.24.100195.001115.

Liamputtong, P., 2010. Moral and ethical perspectives. In: Liamputtong, P. (Ed.), Performing Qualitative Cross-Cultural Research. Cambridge University Press, Cambridge, pp. 31–57. http://dx.doi.org/10.1017/CBO9780511812705.

Marshall, A., Batten, S., 2003. Ethical issues in cross-cultural research. Connections 3, 139–151. Available at: http://xa.yimg.com/kq/groups/22129499/384089802/name/10Marshall105.pdf.

Monshi, B., Zieglmayer, V., 2004. The problem of privacy in transcultural research: reflections on an ethnographic study in Sri Lanka. Ethics & Behavior 14 (4), 305–312. Pubmed PMID: 16622990.

Páez, R., Garcia De Alba, J.E., 2009. International research and just sharing of benefits in Mexico. Developing World Bioethics 9 (2), 65–73. http://dx.doi.org/10.1111/j.1471-8847.2008.00228.x.

Sampson, D.A., Caldwell, D., Taylor, A.D., Taylor, J.Y., 2013. Blending genetics and sociocultural historical inquiry: ethics, culture, and human subjects protection in international cross cultural research. Yale Journal of Biology and Medicine 86 (1), 89–98. Pubmed Central PMCID: PMC3584499.

Tindana, P.O., Kass, N., Akweongo, P., 2006. The informed consent process in a rural African setting: a case study of the Kassena-Nankana District of Northern Ghana. IRB 28 (3), 1–6. Pubmed Central PMCID: PMC3099424.

UNESCO, 1994. Ethical Guidelines for International Comparative Social Science Research in the Framework of MOST. Available at: http://www.unesco.org/most/ethical.htm.

UNESCO, 2005. Universal Declaration on Bioethics and Human Rights. Available at: http://unesdoc.unesco.org/images/0014/001461/146180E.pdf.

United Nations General Assembly, 1948. Universal Declaration of Human Rights. Available at: http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=eng.

US Government Printing Office, 1949. Nuremberg code. Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law 10 (2), 181–182.

US Government Printing Office, 2014. Title 45 US Code of Federal Regulation Pt 46: Protection of Human Subjects. Available at: http://www.ecfr.gov/cgi-bin/text-idx?SID=f2cee9b62e70880a8264860e65340ee5&tpl=/ecfrbrowse/Title45/45cfr46_main_02.tpl.

Wellcome Trust, March 2013. Guidance Notes on Research Involving People in Low and Middle Income Countries. Available at: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD015295.htm.

World Medical Association, October 2013. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Available at: http://www.wma.net/en/30publications/10policies/b3/.

Research Ethics: Research

Elizabeth Yuko and Celia B Fisher, Center for Ethics Education, Fordham University, Bronx, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.E. Sieber, volume 19, pp. 13235–13240, © 2001, Elsevier Ltd.

Abstract

Since the adoption of The Nuremberg Code (1949) following Nazi medical experimentation in World War II, the ethical issues involved in research have been identified and are constantly developing and evolving. This article begins by providing a history of international research ethics regulations and guidelines and their development. It then provides a brief overview of research ethics issues such as ethics review procedures, and issues related to conflicts of interest. The article then explains and provides examples of mechanisms for the application of the three primary ethical principles involved in research ethics: respect for persons, beneficence, and justice. We conclude with a discussion of research ethics in an international and cross-cultural context.

The History of Research Ethics Regulations and Their Implementation

The first international regulation involving research ethics – The Nuremberg Code (1949) – was created in response to the atrocities in human experimentation conducted on prisoners by German Nazi doctors during World War II. Although the Nuremberg tribunal believed that the principles articulated in the Code were already widely accepted by the medical and scientific community, by codifying these principles, they established stringent and novel ethical regulations of human subjects research (Steinbock et al., 2003). The first principle emphasizes the tribunal's faith in informed and voluntary consent as the best means of protecting participant rights. Additional standards reflected the obligation of scientists to justify the scientific and social value of a study, avoid unnecessary physical and mental harms, and put in place adequate protections against harm. Almost two decades later, the World Medical Association published the Declaration of Helsinki – an international set of guidelines emphasizing the physician-scientist's primary obligation to promote the health, well-being, and rights of participants in medical research. This duty included protecting the life, health, dignity, integrity, right to self-determination, privacy, and confidentiality of personal information of research participants. For more information on international research ethics guidelines, *see* Ethics for Biomedical Research Involving Humans: International Guidelines.

Although professional organizations, such as the American Psychological Association (APA), began including general principles of research ethics in their codes of conduct as early as the 1950s, the Nuremberg Code and the Declaration of Helsinki had little impact on the regulation of medical and social science research until the 1970s following revelation, and subsequent public outrage, over the government-sponsored Tuskegee syphilis study. This study – which started in 1932 and involved 400 rural African-American men – was designed to determine what happens when syphilis remains untreated. Over the course of the 30-year study, the participants were neither informed nor treated when penicillin was discovered to be an effective treatment. When the details of the Tuskegee study were made public in 1972, public pressure led to US Public Law 93-348 calling for the formation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1974. After 5 years, the Commission produced the Belmont Report (DHEW, 1979), which established three ethical principles that became the foundation for the 1981 Code of US Federal Regulations 45CFR46 on the protection of human subjects in medical and behavioral research regulations. The Belmont principles include: (1) respect for persons, embodied in regulatory requirements for informed consent; (2) beneficence, embodied in regulations on the obligation to minimize risks and maximize research benefits; and (3) justice, reflecting in regulations requiring fair access of all populations to the benefits and burdens of research participation. For more information on research regulations in the US, *see* The Common Rule: US Regulatory Protections for Human Research Participants.

Shortly thereafter, in 1982, the Council for International Organizations of Medical Services (CIOMS), in cooperation with the World Health Organization (WHO) published the Proposed International Ethical Guidelines for Biomedical Research Involving Human Subjects, that demonstrated how ethical principles outlined in the Declaration of Helsinki could be applied, particularly in developing countries. The two subsequent revisions in 1993 and 2002 expanded on these guidelines in response to the globalization of medical and social science research. With advances in genomic science, the Council of Europe adopted the Convention on Human Rights and Biomedicine (also known as the Oviedo Convention) in 1997, which contains a chapter on the ethics of scientific research, in particular, the protection of persons undergoing research, the protection of persons unable to consent to research, and research on embryos *in vitro*.

Although many of the research ethics guidelines have originated in Europe and North America, WHO – the Special Programme for Research and Training in Tropical Diseases (TDR) in particular – established the Strategic Initiative for Developing Capacity in Ethical Review (SIDCER) in 2001, which was created to carry out its activities regionally, and has committees in Africa, Asia and the Western Pacific, Latin America, and the countries of the former Soviet Union.

The most recent internationally agreed-upon instrument on human subjects research is found in United Nations Education, Scientific and Cultural Organization's (UNESCO) Universal

Declaration on Bioethics and Human Rights (2005). One of the aims of the declaration is "to recognize the importance of freedom of scientific research and the benefits derived from scientific and technological developments, while stressing the need for such research and developments to occur within the framework of ethical principles set out in this Declaration and to respect human dignity, human rights and fundamental freedoms" (UNESCO, 2005: Art. 2(d)). The declaration lists 15 ethical principles, all of which can be applied to human subjects research, and many of which involve the principle of justice, including benefit sharing and equality.

Recognizing the strong biomedical basis for many of the international guidelines governing research, social science organizations began to develop specific standards for the conduct of social and behavioral research. As mentioned previously, the APA established their first ethics code in 1953, which was followed by the adoption of the American Anthropological Association's Statement on Problems of Anthropological Research and Ethics in 1967 (Hill, 1987). These early social science codes were written in aspirational language that was often challenging to apply uniformly to enforce research ethics practices (Fisher, 2013; Fisher and Anushko, 2008). It was not until the 1990s that specifically worded enforceable standards of conduct for research, teaching, and professional practice were included in regulations such as the British Psychological Society *Ethical Principles for Conducting Research with Human Participants*, the Canadian Psychological Association *Canadian Code of Ethics for Psychologists*, and the APA Ethics Code (BPS, 2010; Canter et al., 1994; CPA, 1991).

Specific ethical standards for social science research continue to be grounded in broad ethical principles. For example, Australia included social science provisions in their revised (1999) and current (2014) versions of their National Statement on Ethical Conduct in Research Involving Humans, which highlights principles such as research merit and integrity, justice, beneficence, and respect (NHMRC, 1999, 2014). The APA Ethics Code currently begins with a section outlining five general ethical principles: beneficence and nonmaleficence, fidelity and responsibility, integrity, justice, and respect for people's rights and dignity. Other codes, such as the Canadian Psychological Association's Code of Ethics, attempts to assist researchers prioritize ethical principles when they come into conflict, using the following order: (1) respect for the dignity of persons; (2) responsible caring; (3) integrity in relationships; and (4) responsibility to society (CPA, 2000). For more information on ethical dilemmas, research and treatment priorities, *see* Ethical Dilemmas, Research, and Access to Treatment.

The social science ethics codes continue to be revised in response to contemporary social issues. For example, in response to psychologists' involvement in harsh prisoner interrogations following the 11 September 2001 attacks, the most recent revision of the APA Ethics Code (2010) includes enforceable standards requiring psychologists to refrain from activities that could be used to justify or defend violation of human rights (Fisher, 2013).

Research Ethics Review

The regulations and guidelines discussed above stipulate that research involving human participants must undergo a review of its ethical acceptability prior to beginning the study. Research Ethics Committees (RECs) – referred to as Institutional Review Boards (IRBs) in the United States – are charged with the task of ensuring that the rights and welfare of the research participants are protected (Fisher, 2013). This includes achieving an appropriate balance between research risks and benefits; adequate informed consent procedures; fair procedures for the selection of participants; and special protections for vulnerable populations. Across the globe there are varying levels of review at the institutional, national, and regional levels, and having a good systems approach and clear guidelines will greatly help facilitate international research in the future (WHO, 2011). For example, in the European Union, it is legally required for all clinical trial research to be reviewed by an REC prior to the commencement of the study, in which it "comes to the conclusion that the anticipated therapeutic and public health benefits justify the risks and may be continued only if compliance with this requirement is permanently monitored" (EU Directive 2001/20/EC).

Conflicts of Interest

The trust and validity of research may be compromised by conflicting professional, personal, financial, or legal interests that can impair the objectivity of data collection, analysis, and interpretation (Fisher and Anushko, 2008). It is therefore necessary to establish ethical guidelines to avoid potentially harmful or exploitative conflicts of interest, as the impairment of objectivity can be detrimental to participants, the public, the institution, the funders, the investigator, and the integrity of social science. Such guidelines exist on the international (e.g., the Declaration of Helsinki and the CIOMS Guidelines) and national level (e.g., the National Institutes of Health (NIH) Office of Extramural Research's requirements in the United States). National regulations, such as the ones issued by the NIH, require every institution receiving federal research funds to have written guidelines for the avoidance and institutional review of conflict of interest (Fisher, 2013). Similarly, transnational regulations, such as the European Research Council (ERC) Scientific Misconduct Strategy, issued by the Standing Committee on Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME), exist to ensure that "grants are awarded on the sole criterion of scientific excellence," and that any incidences of misconduct are widely and publicly reported, in order for the ERC to "uphold high ethical standards in all its activities" (ERC, 2012; *see* Research Funding: Ethical Aspects).

Ethical Principles

Although regulations and guidelines vary in their emphasis on ethical principles necessary to ensure the protection of research participants' rights and welfare, three continue to serve as the most important and inclusive: respect for persons, beneficence, and justice. As social science research and regulations have developed, there has been a shift in the importance of the principles. The primary concern of the Nuremberg Code, for example, was the need for informed consent – which falls under the principle of respect for persons. But in recent times,

with advances in benefits derived from pharmaceutical and behavioral risk prevention research, there has been increasing emphasis placed on the principle of justice, particularly in terms of access to the benefits of research for vulnerable populations in economically developed and developing countries. This section will address each of the three principles in order of historical importance, along with the many mechanisms of application and the concepts that fall under each principle.

Respect for Persons

The ethical principle of respect for persons incorporates a respect for human rights and human dignity. More specifically, it involves treating individuals as autonomous agents, and ensuring that those with diminished capacities are entitled to added protections. Respect for persons is most commonly embodied in the requirements of informed consent and protection of confidential data.

Informed Consent

There are three components of informed consent: information, comprehension, and voluntariness (DHEW, 1979).

Information

Investigators are ethically obligated to provide research participants with the information necessary to ensure they can make an informed decision to agree or disagree to research participation. In the Declaration of Helsinki, for example, this information includes the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher; the anticipated benefits and potential risks of the study and the discomfort it may entail; the post-study provisions; the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal; and any other relevant aspects of the study.

The exploratory and open-ended nature of participant observation, ethnographic, and other types of qualitative research raises questions about whether truly informed consent can be obtained for such studies (Marshall, 1992). In such cases, neither the investigator nor the participant may fully be aware of information that will arise during the course of discussion or observations. One way to address this issue is to clarify in advance that if unanticipated and sensitive issues emerge during the course of observation or discussion, a mutually negotiated re-consent procedure will be implemented (Fisher, 2004; Haverkamp, 2005). Additionally, researchers can view and seek informed consent as an ongoing, mutually negotiated process, rather than a single event (Grafanaki, 1996; Haverkamp, 2005; McLeod, 1996; Munhall, 1989; Raudonis, 1992; Rossiter et al., 1996; Smyth and Murray, 2000).

Comprehension

It is important that research participants not only receive accurate and necessary information about the study, but that the format and context in which the information is provided facilitates and ensures their understanding. For example, information should be presented to the research participants in a clear and organized manner, at an appropriate pace, and allowing for enough time for consideration and for participants to ask questions (DHEW, 1979).

In addition, the legal status and consent capabilities of research participants must be taken into consideration. Regulations and international and organizational guidelines require that investigators protect the rights of children who have not attained the legal status to consent, and adults with intellectual disabilities and other consent-compromising conditions. For example, guardian permission is required to protect children from consent vulnerabilities related to immature cognitive skills and lack of experience in health-care decision-making. At the same time, as articulated in the UN Convention on the Rights of the Child (United Naitons, 1989) respect for children's developing autonomy means that their assent to participate is also required. Under US regulations, guardian permission may be waived, however, if there is serious doubt that the guardian's interests adequately reflect the child's interests (e.g., research on child abuse), or in the case of adolescents, when research is related to health services for which they have attained the legal right to access independent of guardian permission. A report from the Ethics Working Group of the Confederation of European Specialists in Paediatrics (CESP) takes the stance that medical research involving children is not only acceptable, but necessary, as long as the following six guidelines are met: (1) the research is "important for the benefit of all children and should be supported, encouraged and conducted in an ethical manner"; (2) children have a unique set of interests; (3) research should only take place if comparable research on adults could not answer the same question; (4) if a research procedure is nontherapeutic then it must be likely to yield generalizable knowledge of vital importance; (5) all research proposals involving children should be submitted to an REC with experts in pediatric research; and (6) informed consent from a child, parent, or guardian should be obtained, as appropriate; when parental consent is obtained, the assent of school-age children who participate in research should also be requested by researchers (Sauer, 2002).

Similar protections are required for adults with health disorders that have required appointing a legal guardian. In many cases, however, adults with Alzheimer's disease or other progressive or developmental intellectual deficits may not have a legal guardian. In such cases, a trusted adult may serve as a consent advocate and provide support for the prospective participant's consent decision, which can protect against the potential for triggering a legal competency review for adults with disabilities who do not have a legal guardian (Fisher, 2002a, 2013). When investigators are unsure of the research participants' capacity to consent, they may use any number of assessment tools to determine the ability of the prospective participants to comprehend the nature of and their rights in the research. For example, the MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR) measures four increasingly complex aspects of research decision-making competence: the ability to (1) express a participation choice; (2) understand the purpose and nature of the research; (3) appreciate how participation will affect them personally; and (4) make a reasoned participation decision based on balancing risks and benefits (Grisso and Applebaum, 2001). For research

presenting only minimal risks to participants, the ability of a participant to understand and make a participation choice may be sufficient protection for ensuring they are making a choice in their own best interests. The latter two criteria may be more important as the risk of research participation increases.

For both children and adults with impaired cognitive capacity, participation decisions can each be enhanced through a model of ethical decision-making called goodness-of-fit ethics (GFE) (Fisher and Ragsdale, 2006; Masty and Fisher, 2008; Fisher and Goodman, 2009). According to GFE, consent vulnerability must be understood in terms of how well-informed consent procedures fit certain participant characteristics. GFE thus requires researchers to identify the research population's consent strengths and vulnerabilities, and fit the consent or assent procedures in ways that minimize vulnerabilities, enhance consent strengths, and provide consent supports when feasible (Fisher, 2013).

Therapeutic Misconception

A consent comprehension challenge arising in intervention research is therapeutic misconception (TM). First identified by Applebaum, Roth and Lidz in 1982, TM refers to two common, but incorrect beliefs held by research participants in randomized, controlled intervention research: (1) that their individual needs will be taken into consideration in group assignment and (2) that it is very likely that they will benefit therapeutically from the research participation. The TM tendency among prospective participants obligates investigators to clarify during informed consent the differences between the roles of a practitioner and those of an investigator, and the goals of treatment versus research. In addition, TM may be exacerbated by therapeutic mistrust in socially vulnerable or marginalized populations (Fisher et al., 2008). When working with social minorities, researchers must anticipate and effectively address the personal and collective history of abuses that these populations have experienced as research subjects (Fisher et al., 2002; Reverby, 2001). For more information on ethical implications and assessment and application of therapeutic effectiveness, *see* Therapeutic Effectiveness Assessment: Approaches in Contemporary Research.

Voluntariness

Informed consent must clarify the voluntary nature of participation, as well as an individual's right to withdraw his or her participation at any time during the study. Participants recruited from treatment centers, educational settings, social service agencies, or prisons may be understandably concerned that a refusal to participate in research may result in a denial of services, or other institutional retaliation. Investigators conducting research in such settings need to be vigilant in ensuring that such services will not be deprived, and clarifying such protections to prospective participants during consent.

While offering some form of compensation is often both ethical and necessary to fairly compensate individuals for their research participation time and effort and to guarantee an appropriately sized and representative research sample, it is possible – and crucial – for researchers to distinguish between 'due' and 'undue' inducements (Dickert and Grady, 1999; Macklin, 1999; VanderWalde and Kurzban, 2011). According to Emanuel (2005), compensation for research participation is only coercive when each of these four elements are satisfied: (1) the prospective participant is offered something that is perceived as valuable or desirable; (2) the incentive is so significant or in excess that it is irresistible in the context of the participation request; (3) the offer leads prospective participants to exercise poor judgment in an important decision, such as during recruitment or informed consent; and (4) the individual's poor judgment leads to the sufficiently high probability that the participation decision will lead to a harmful experience that 'seriously contravenes his or her interests' (Emanuel, 2005, p. 9). It is, however, important to note that strict prohibitions on undue inducements can lead to a paradoxical result: the less economically developed a community is, the fewer benefits can be offered, so as not to unduly influence the decision to participate (Schroeder and Gefenas, 2012). This paradox provides research sponsors with an 'ethical' argument for limiting benefits given to study participants in developing countries, and a challenge for RECs that must make decisions about this type of proposed research (Ballantyne, 2008; Schroeder and Gefenas, 2012).

There are several strategies for selecting an appropriate rate of compensation. For example, many institutions have adopted a set level of compensation for all research participants, while others have stipulated that a noncoercive financial incentive is the amount that an average, healthy volunteer would lose in work and/or travel time by the fair market value for the work involved (Fisher, 2013). Furthermore, consulting with members of the population that will be recruited for research participation about appropriate levels of compensation may assist researchers and their RECs determine whether the compensation is fair or coercive (Fisher, 2003; Flicker et al., 2007; Oransky et al., 2009).

Similar to capacity to consent, instruments exist to measure voluntariness. For example, Applebaum et al. (2009) identified offers, pressures, or threats related to research, as the types of influence most likely to reflect mechanisms by which one person can exert influence on the decisions of another, and constructed a brief instrument to assess voluntariness and motivation to participation.

Deception in Research

Deception in research involves many ethical challenges to adhering to the principle of respect for persons. One challenge is what Fisher (2005) has described as the 'consent paradox,' reflecting the moral paradox involved in intentionally deceiving a prospective participant about the nature of the research, thereby compromising the individual's ability to make a fully informed decision regarding research. In these situations, researchers conducting deception research intentionally give participants false information about the purpose and/or nature of their study, therefore leading individuals to erroneously believe that they have the autonomy to determine the type of experimental procedures in which they agree to participate; furthermore, the participants may experience

potential discomfort following the study when they are notified of the deception (Fisher and Fyrberg, 1994; see also Fried, 2012; McCambridge et al., 2013; Sieber, 1982).

The most recent revision of the APA Ethics Code (2010) includes a more protective stance on deception in research, prohibiting the use of deceptive procedures if: (1) effective nondeceptive alternative methods are feasible or (2) the research is reasonably expected to cause participants physical pain or severe emotional distress. The standard also stipulates that researchers must debrief participants about the deception as soon as possible and permit them to withdraw their data following debriefing. For more information on informed consent, *see* Informed Consent of Research Subjects.

Protecting Participant Privacy

Social scientists respect the privacy and dignity of persons by protecting confidential information obtained through the course of research. Standard procedures for protecting confidentiality include use of participant codes on all data-collected materials, keeping password-protected files, limiting access to personally identifiable information to research staff members, and attaining competence in protecting information collected over the Internet, mobile phones, and other new technologies (Fisher, 2013). Investigators must also ensure that recruitment and research procedures do not in themselves inadvertently reveal confidential information. For example, street recruitment for a study on drug use and sexual risk behaviors can inadvertently alert police to the identity of prospective participants who are using or selling illegal drugs (Fisher, 2004). When publishing or otherwise disseminating research findings, investigators need to consider special confidentiality protections when dealing with unnamed but small, unique, or easily identified samples (i.e., participants living on American Indian or Alaskan Native tribal grounds) (Fisher, 2013).

The APA Ethics Code, the British Psychological Association Code of Human Research Ethics, the National Committees for Research Ethics in Norway's Guidelines for Research Ethics in the Social Sciences, Law and the Humanities, and other social science research guidelines recognize instances in which investigators may need to reveal confidential information to protect individuals from harm to themselves or others. For example, a study of school-bullying may reveal that a student is planning to or has already violently attacked another student. Or, during an interview study, a person addicted to drugs comes to the research site in a state suggesting a serious drug overdose. Investigators conducting research on parenting or related behaviors may also be mandated child abuse reporters. Under such conditions, informed consent must alert participants of the investigator's reporting responsibilities if child abuse is mentioned during the course of research interviews or other procedures. Fisher and Goodman (2009) propose taking the following steps in making ethical decisions regarding disclosure of confidential information: (1) anticipate disclosure challenges based on the nature of the sample and data collected; (2) investigate one's own legal responsibilities as well as community resources if an instance emerges in which disclosing information to protect a participant or other is required; (3) based on the above, select a specific disclosure policy before implementing the study; (4) train research staff in reporting procedures; and (5) include a description of the specific disclosure policy in the informed consent. For more information on privacy and confidentiality in research, *see* Privacy of Individuals in Social Research: Confidentiality.

Beneficence

In every research study, there is a continuum of benefits and risks. It is the responsibility of the researcher to maximize benefits, minimize risks, and achieve equitable balance between the two. Consequently, the ethical principle of beneficence is comprised of two complementary actions: beneficence ('to do good') and nonmaleficence ('do no harm'). In applying these principles, investigators are required to maximize research benefits and minimize research risks. Beneficial research outcomes can be achieved directly through clinical trials designed to test the efficacy of an educational, behavioral, or medical intervention, or indirectly through descriptive research identifying the individual, interpersonal, social, economic, or other contexts influencing behavior or health.

Scientific merit is a minimal requirement for determining whether possible research risks to persons or groups are ethically justified, based on the study's proposed scientific validity and value. A poorly designed study holds no possibility of benefits. Research must be designed to produce reliable information according to accepted principles of research. However, a study may be well-designed, but the benefits are not balanced by research risks if the hypotheses themselves do not meaningfully enhance understanding of a social or psychological phenomenon, or help to ameliorate an educational, behavioral, or health-related problem (The Nuremberg Code, 1949). The threshold for scientific validity and social value is higher when research is conducted with vulnerable groups such as social minorities or historically oppressed groups and requires: (1) an awareness of the importance of scientific, social, and political factors governing definitions of race, ethnicity, and culture; (2) an understanding of between- and within-group differences; (3) familiarity with and skills in constructing and implementing culturally valid and language-appropriate assessment instruments; and (4) knowledge of the cultural and political circumstances of participants' lives (Fisher et al., 2002). Social value has gained increased attention as a key requirement for research, and is becoming central to the current revision of CIOMS (Bahn, 2014).

Participant Perspectives

In recent years, there has been a movement to incorporate the views of participant communities in the assessment of research risks and benefits on the assumption that investigators' and RECs' experiences may not be similar to those of participants, and therefore may have different perspectives and perceptions of research harms and benefits (Fisher, 1999, 2011; Fisher and Wallace, 2000). As detailed in several reports by the Nuffield Council on Bioethics, engagement with the community and participants plays an important role in conducting research and forming policy (Nuffield, 2002, 2012).

In both underserved populations in the US and in developing countries, there has also been a movement toward community engaged research (CEnR) (Friedman Ross et al., 2010) that includes community members as equal partners in the development and evaluation of human subjects' protections and research design, implementation, analysis, interpretation, and dissemination. CEnR principles bring mutual respect, community voice, and ways of knowing how to guide research agendas and develop interventions from cultural perspectives that resonate with the study population (Cochran et al., 2008; Minkler and Wallerstein, 2008; see also Alvarez et al., 2006; Boga et al., 2011; Dickert and Sugarman, 2005; Newman et al., 2011; Solomon et al., 2014; Strauss et al., 2001).

Clinical Equipoise

Beneficence also guides investigators in determining whether there is a fair distribution of research risks and benefits in the different conditions to which participants will be randomly assigned in a clinical trial. The effectiveness of a new prevention or treatment study is often evaluated by comparing it to a control condition (i.e., a placebo, treatment as usual). This type of research is often called a randomized clinical trial (RCT) because standards of research design require that participants be randomly assigned to the treatment or control condition. This means that individuals who meet inclusion criteria are assigned to a condition according to chance, rather than by their own choice, or as dictated by the usual response to their diagnosis or prognosis (*see* Placebo-Controlled Trials, Ethics of; Cluster Randomized Trials, Ethical Implications of).

It is unethical to assign subjects to either a treatment or control condition if one of those conditions is known to be inferior. Before beginning a treatment-control group study, the investigator must be honestly able to state that sufficient scientific or clinical evidence does not yet exist to conclude that subjects in the treatment condition will differ in outcome from subjects treated in the control condition. According to the concept of clinical equipoise, a treatment-control group study may be justified where there is a current or likely dispute among experts as to which of the two or more conditions is superior (Freedman, 1987; Weijer and Miller, 2003).

Justice

The principle of justice addresses two primary questions: who should bear the burdens of research, and who should receive the benefits? Another way of thinking about justice is "who is obligated to provide what to whom and why" (Pratt and Loff, 2011, p. 75). The principle of justice has emerged as a moral imperative in human rights debates regarding access to research and the individual and group benefits it may generate. An injustice occurs when disproportionate burden in research is imposed upon a group, population, or individual. For example, this occurs when institutionalized persons are recruited as samples of convenience for studies designed to benefit the larger population. Injustices also occur when specific social groups are excluded from the benefits of research; for example, when studies on educational or mental health interventions are conducted with middle-class ethnic majority participants with the results generalized to services delivered to poorer ethnic minority citizens.

One application of justice is the investigator's responsibility to ensure fair procedures and outcomes in the selection of human research subjects. An ethical issue arises when racial or ethnic minority, disadvantaged, or disenfranchised populations are recruited to participate in studies that do not include consideration of the unique population characteristics that may reduce their ability or capacity to receive the benefits of the research, or expose them to greater risk or financial burden (Fisher, 1999; Trimble and Fisher, 2006). Over the past decades, for example, US federal regulations have been modified to encourage inclusion of social minorities in research, and to require investigators to justify the exclusion of women, ethnic/racial minority groups, or children from research applications for federal funding.

Distributive Justice

In addition to the issues discussed above, there are also issues of distributive justice in research. Who or what should determine the priority and focus of research? In other words, should priority be given to research involving those most in need of improved health (Dresser, 2001)? How should funding agencies decide who is most deserving of their budget? For example, a funding body may have to determine whether to spend money on research on a less serious, but widespread health problem like arthritis, or on studies examining more serious but uncommon conditions. Current debates include whether an 'affirmative action' approach should ever be taken to research funding, and if so, whether current research focus should on populations that were previously excluded from or underrepresented in research (Dresser, 2001).

Justice in International Research

As research becomes more globalized, certain ethical challenges arise that are specific to international and cross-cultural research, particularly in situations where the researcher or funding institution from one country is conducting research in another country. Researchers may decide to conduct research abroad in order to study the behavior a specific population, to help the host country address a public health problem, or because a certain location is a more convenient, efficient, or less troublesome site for conducting a particular clinical trial (NBAC, 2001).

When conducting clinical trials in developing countries, risk – benefit assessment takes on different dimensions, because the access to health care outside of the context of research may be limited (Macklin, 2004). The principle of justice requires an understanding of the health needs and priorities of the host community or country; providing the participants with fair benefits during and after the research (including access to treatments or practices that may have been developed by the study); and strengthening the research capacity and ethical review process in the host communities or countries (Pratt and Loff, 2011). Ensuring that these ethical obligations are met is important in the promotion of

global justice, and to fill research gaps in underperforming countries, and to reduce health disparities between nations (Pratt and Loff, 2011).

In some countries, investigators may need to consider state-sanctioned discrimination or disempowerment of certain groups. For example, traditions in some localities may require an investigator to obtain permission from a woman's husband, father, or other male relative before approaching her for research participation. In such situations, these traditions may be respected, but it is ethically imperative that protections are put in place to ensure that the woman's consent is informed and voluntary (NBAC, 2001; Pratt and Loff, 2011). Moreover, research should never be conducted in communities where political structure or power imbalance based on gender would, in any way, compromise the voluntariness of participation by pressuring women to participate, or when such procedures will justify or defend the violation of human rights (Fisher, 2013). For more information on cross-cultural dimensions of research ethics, *see* Research Ethics, Cross Cultural Dimensions of; Fieldwork: Ethical Aspects; Gender and Social Research.

In international social science research, the researcher and participant communities may have different concepts of physical, informational, and decisional privacy (Goldman and Choy, 2001). These differences can manifest themselves in various ways. For example, in societies with living arrangements and cultural values that emphasize community openness and family independence, participants may question the motives of a researcher who attempts to conduct research interviews in the absence of family members (Monshi and Zieglmayer, 2004). As a result, when conducting international and/or cross-cultural research, it is important for the researcher to determine how the participant population defines and values privacy, and then design confidentiality procedures to reflect these values (Fisher, 2013). In localities affected by war, or in which vulnerable populations such as gay and lesbian individuals or drug users are subject to imprisonment, investigators must ensure that their recruitment procedures do not identify individuals in ways that place them at risk of violence or incarceration (Wolfe and Cohen, 2010).

Conclusion

Over time, there has been a shift in emphasis of the ethical principles of research ethics, from an approach primarily focused on informed consent, to one that incorporates increased consideration to justice. Additionally, increasing numbers of international research studies – a by-product of our ever-globalizing world – brings along with it numerous ethical challenges. It is crucial that researchers see the importance and value of adhering to the aforementioned principles in their work in an effort to maximize the benefits of research for our global society.

See also: Cluster Randomized Trials, Ethical Implications of; Common Rule: US Regulatory Protections for Human Research Participants; Ethical Dilemmas, Research, and Access to Treatment; Ethics for Biomedical Research Involving Humans: International Guidelines; Fieldwork: Ethical Aspects; Gender and Social Research; Informed Consent of Research Subjects; Placebo-Controlled Trials, Ethics of; Privacy of Individuals in Social Research: Confidentiality; Research Ethics, Cross-Cultural Dimensions of; Research Funding: Ethical Aspects; Therapeutic Effectiveness Assessment: Approaches in Contemporary Research.

Bibliography

Alvarez, R.A., Vasquez, E., Mayorga, C.C., Feaster, D.J., Mitrani, V.B., 2006. Increasing minority research participation through community organization outreach. Western Journal of Nursing Research 28 (5), 541–560.

American Psychological Association, 1953. Ethical Standards of Psychologists.

American Psychological Association, 2010. Ethical Principles of Psychologists and Code of Conduct, Including 2010 Amendments. http://www.apa.org/ethics/code/index.aspx (accessed 07.04.14.).

American Sociological Association, 1999. Code of Ethics. American, Washington, DC.

Appelbaum, P.S., Roth, L.H., Lidz, C., 1982. The therapeutic misconception: informed consent in psychiatric research. International Journal of Law and Psychiatry 5 (3), 319–329.

Applebaum, P.S., Lidz, C.W., Klitzman, R., 2009. Voluntariness of consent to research: a preliminary empirical investigation. IRB: Ethics & Human Research 31 (6), 10–18.

Australian National Health and Medical Research Council, 2014. National Statement on Ethical Conduct in Human Research updated March 2014. https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72_national_statement_march_2014_140331.pdf (accessed 04.04.14.).

Bahn, A., 2014. The CIOMS guidelines in low and middle income countries. IAB Congress, Symposium on the Revision of the CIOMS Guidelines. http://www.cioms.ch/images/stories/CIOMS/Bhan_CIOMSwg_2014Mexico.pdf (accessed 02.09.14.).

Ballantyne, A., 2008. Benefits to research subjects in international trials: do they reduce exploitation or increase undue inducement? Developing World Bioethics 8 (3), 178–191.

Boga, M., Davies, A., Kamuya, D., Kinyanjui, S.M., Kivaya, E., Kombe, F., Mwalukore, S., 2011. Strengthening the informed consent process in international health research through community engagement: the KEMRI-wellcome trust research programme experience. PLoS Medicine 8 (9), e1001089.

British Psychological Society, 2010. Code of Human Research Ethics updated 2010. http://www.bps.org.uk/sites/default/files/documents/code_of_human_research_ethics.pdf (accessed 28.08.14.).

Canadian Psychological Association, 1991. Canadian Code of Ethics for Psychologists. Ottawa, ON.

Canadian Psychological Association, 2000. Canadian Code of Ethics for; Psychologists, third ed. http://www.cpa.ca/cpasite/userfiles/Documents/Canadian%20Code%20of%20Ethics%20for%20Psycho.pdf (accessed 04.04.14.).

Canter, M.B., Bennett, B.E., Jones, S.E., Nagy, T.F., 1994. Ethics for Psychologists: A Commentary on the APA Ethics Code. American Psychological Association, Washington, DC.

Cochran, P.A., Marshall, C.A., Garcia-Downing, C., Kendall, E., Cook, D., McCubbin, L., Gover, R.M., 2008. Indigenous ways of knowing: implications for participatory research and community. American Journal of Public Health 98 (1), 22–27.

Council of Europe, 1997. Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine. http://conventions.coe.int/Treaty/en/Treaties/Html/164.htm (accessed 07.04.14.).

Council for International Organizations of Medical Sciences and the World Health Organization, 1982. Proposed International Ethical Guidelines for Biomedical Research Involving Human Subjects.

Department of Health, Education, and Welfare, 1979. The Belmont Report. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html (accessed 07.04.14.).

Dickert, N., Grady, C., 1999. What's the price of a research subject? Approaches to payment for research participation. New England Journal of Medicine 341, 198–203.

Dickert, N., Sugarman, J., 2005. Ethical goals of community consultation in research. American Journal of Public Health 95 (7), 1123–1127.

Dresser, R., 2001. When Science Offers Salvation: Patient Advocacy and Research Ethics. Oxford University Press, USA.

Emanuel, E.J., 2005. Undue inducement: nonsense on stilts? The American Journal of Bioethics 5 (5), 9–13.

European Union, 2001. Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the Approximation of the Laws, Regulations and Administrative Provisions of the Member States Relating to the Implementation of Good Clinical Practice in the Conduct of Clinical Trials on Medicinal Products for Human Use. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:121:0034:0044:en (accessed 26.08.14.).

European Research Council (ERC) Scientific Misconduct Strategy, 2012. http://erc.europa.eu/sites/default/files/document/file/ERC_Scientific_misconduct_strategy.pdf.

Fisher, C.B., 1999. Relational Ethics and Research with Vulnerable Populations in Reports on Research Involving Persons with Mental Disorders that May Affect Decision-Making Capacity. Commissioned Paper by the National Bioethics Advisory Commission, vol. 2. National Bioethics Advisory Commission, Rockville, MD, pp. 29–49.

Fisher, C.B., 2002. A goodness-of-fit ethic for informed consent. Fordham Urban Law Journal 30, 159–171.

Fisher, C.B., 2003. Adolescent and parent perspectives on ethical issues in youth drug use and suicide survey research. Ethics & Behavior 13 (4), 303–332.

Fisher, C.B., 2004. Ethics in drug abuse and related HIV risk research. Applied Developmental Science 8, 90–102.

Fisher, C.B., 2005. Deception research involving children: ethical practices and paradoxes. Ethics & Behavior 15 (3), 271–287.

Fisher, C.B., 2011. Addiction research ethics and the Belmont principles: do drug users have a different moral voice? Substance Use and Misuse 46 (6), 728–741.

Fisher, C.B., 2013. Decoding the Ethics Code, third ed. Sage, London.

Fisher, C., Anushko, A., 2008. Research ethics in social science. In: Alasuutari, P., Bickman, L., Brannen, J. (Eds.), The SAGE Handbook of Social Research Methods. Sage, London, pp. 95–110.

Fisher, C.B., Fried, A.F., Goodman, S.J., Germano, K.K., 2009. Measures of mentoring, department climate, and graduate student preparedness in the responsible conduct of psychological research. Ethics & Behavior 19, 227–252.

Fisher, C.B., Fyrberg, D., 1994. Participant partners: college students weigh the costs and benefits of deceptive research. American Psychologist 49, 417–427.

Fisher, C.B., Goodman, S.J., 2009. Goodness-of-fit ethics for non-intervention research involving dangerous and illegal behaviors. In: Buchanan, D., B.Fisher, C., Gable, L. (Eds.), Research with High-Risk Populations: Balancing Science, Ethics, and Law. APA Books, Washington, DC, pp. 25–46.

Fisher, C.B., Hoagwood, K., Boyce, C., Duster, T., Frank, D.A., Grisso, T., Macklin, R., Levine, R.J., Spencer, M.B., Takanishi, R., Trimble, J.E., Zayas, L.H., 2002. Research ethics for mental health science involving ethnic minority children and youth. American Psychologist 57, 1024–1040.

Fisher, C.B., Oransky, M., Mahadevan, M., Singer, M., Mirhej, G., Hodge, D., 2008. Marginalized populations and drug addiction research: realism, mistrust, and misconception. IRB: Ethics & Human Research 30, 1–9.

Fisher, C.B., Ragsdale, K., 2006. A goodness-of-fit ethics for multicultural research. In: Trimble, J., Fisher, C.B. (Eds.), The Handbook of Ethical Research with Ethnocultural Populations and Communities. Sage, Thousand Oaks, CA, pp. 3–26.

Fisher, C.B., Wallace, S.A., 2000. Through the community looking glass: re-evaluating the ethical and policy implications of research on adolescent risk and psychopathology. Ethics & Behavior 10, 99–118.

Flicker, S., Travers, R., Guta, A., McDonald, S., Meagher, A., 2007. Ethical dilemmas in community-based participatory research: recommendations for institutional review boards. Journal of Urban Health 84 (4), 478–493.

Freedman, B., 1987. Equipoise and the ethics of clinical research. The New England Journal of Medicine 317 (3), 141–145.

Fried, A.L., 2012. Ethics in psychological research: guidelines and regulations. In: Cooper, H., Camic, P., Long, D., Panter, A.T., Rindskopf, D., Sher, K. (Eds.), APA Handbook of Research Methods in Psychology, Foundations, Planning, Measures, and Psychometrics, vol. 1, pp. 55–73.

Friedman Ross, F., Loup, A., Nelson, R., Botkin, J.R., Kost, R., Smith, G., Gehlert, S., 2010. Human subjects protection in community-engaged research: a research ethics framework. Journal of Empirical Research on Human Research Ethics 5 (1), 5–11.

Goldman, J., Choy, A., 2001. Privacy and confidentiality in health research. Paper commissioned by the National Bioethics Advisory Commission.

Grafanaki, S., 1996. How research can change the researcher: the need for sensitivity, flexibility and ethical boundaries in conducting qualitative research in counselling/psychotherapy. British Journal of Guidance and Counselling 24, 329–338.

Grisso, T.H., Appelbaum, P.S., 2001. MacArthur Competence Assessment Tool for Clinical Research (MacCAT-cr). Professional Resource Press, Sarasota, Fla.

Haverkamp, B.E., 2005. Ethical perspectives on qualitative research in applied psychology. Journal of Counseling Psychology 52 (2), 146.

Hill, J.M., 1987. The committee on ethics: past, present and future. In: Cassell, J., Jacobs, S.E. (Eds.), Handbook on Ethical Issues in Anthropology. American Anthropological Association, Washington, DC.

Macklin, R., 1999. Moral progress and ethnical universalism. In: Macklin, R. (Ed.), Against Relativism: Cultural Diversity and the Search for Ethical Universal in Medicine. Oxford University Press, New York, pp. 249–274.

Macklin, R., 2004. Double Standards in Medical Research in Developing Countries. Cambridge University Press, New York.

Marshall, P.A., 1992. Research ethics in applied anthropology. IRB: Ethics & Human Research 14, 1–5.

Masty, J., Fisher, C.B., 2008. A goodness of fit approach to parent permission and child assent pediatric intervention research. Ethics & Behavior 13, 139–160.

McCambridge, J., Kypri, K., Bendtsen, P., Porter, J., 2013. The use of deception in public health behavioral intervention trials: a case study of three online alcohol trials. The American Journal of Bioethics 13 (11), 39–47.

McLeod, J., 1996. Qualitative approaches to research in counselling and psychotherapy: issues and challenges. British Journal of Guidance and Counselling 24, 309–316.

Minkler, M., Wallerstein, N., 2008. Introduction to CBPR: new issues and emphases. In: Minkler, M., Wallerstein, N. (Eds.), Community-Based Participatory Research for Health: From Process to Outcome. John Wiley & Sons, San Francisco, CA.

Monshi, B., Zieglmayer, V., 2004. The problem of privacy in transcultural research: reflections on an ethnographic study in Sri Lanka. Ethics & Behavior 14 (4), 305–312.

Munhall, P.L., 1989. Ethical considerations in qualitative research. Western Journal of Nursing Research 10, 150–162.

National Bioethics Advisory Committee, 2001. In: Ethical and Policy Issues in International Research: Clinical Trials in Developing Countries, vol. 1 (accessed 07.04.14.). https://bioethicsarchive.georgetown.edu/nbac/clinical/Vol1.pdf.

National Committees for Research Ethics in Norway, 2006. Guidelines for Research Ethics in the Social Sciences, Law and the Humanities. https://www.etikkom.no/Documents/Publikasjoner-som-PDF/Guidelines%20for%20research%20ethics%20in%20the%20social%20sciences,%20law%20and%20the%20humanities%20(2006) (accessed 28.08.14.).

Newman, S.D., Andrews, J.O., Magwood, G.S., Jenkins, C., Cox, M.J., Williamson, D.C., 2011. Community advisory boards in community-based participatory research: a synthesis of best processes. Preventing Chronic Disease: Public Health Research, Practice, and Policy 8 (3), 1–12.

Nuffield Council on Bioethics, 2002. The Ethics Related to Healthcare Research in Developing Countries. http://nuffieldbioethics.org/wp-content/uploads/2014/07/Ethics-of-research-related-to-healthcare-in-developing-countries-I.pdf (accessed 02.09.14.).

Nuffield Council on Bioethics, 2012. Emerging Biotechnologies: Technology, Choice and the Public Good. http://nuffieldbioethics.org/wp-content/uploads/2014/07/Emerging_biotechnologies_full_report_web_0.pdf (accessed 02.09.14.).

Oransky, M., Fisher, C.B., Mahadevan, M., Singer, M., 2009. Barriers and opportunities for recruitment for non-intervention studies on HIV risk: perspectives of street drug users. Substance Use & Misuse 44, 1642–1659.

Pratt, B., Loff, B., 2011. Justice in international clinical research. Developing world bioethics 11 (2), 75–81.

Raudonis, B.M., 1992. Ethical considerations in qualitative research with hospice patients. Qualitative Health Research 2 (2), 238–249.

Raudonia, B.M., 2002. Pearls, pith, and provocation: ethical considerations in qualitative research with hospice patients. Qualitative Health Research 2, 238–249.

Reverby, S., 2001. More than fact and fiction: cultural memory and the Tuskegee Syphilis Study. Hastings Center Report 31 (5), 22–28.

Rossiter, A., Walsh-Bowers, R., Prilleltensky, I., 1996. Learning from broken rules: individualism, bureaucracy, and ethics. Ethics & Behavior 6, 307–320.

Sociological Association. http://www.asanet.org/images/asa/docs/pdf/CodeofEthics.pdf (accessed 04.04.14.).

Sauer, P.J., 2002. Research in children: a report of the ethics working group of the CESP. European Journal of Pediatrics 161, 1–5.

Schroeder, D., Gefenas, E., 2012. Realizing benefit sharing – the case of post-study obligations. Bioethics 26 (6), 305–314.

Sieber, J.E., 1982. Deception in social research I: kinds of deception and the wrongs they may involve. IRB: Ethics and Human Research 4, 1–5.

Smythe, W.E., Murray, M.J., 2000. Owning the story: ethical considerations in narrative research. Ethics and Behavior 10, 311–336.

Solomon, S., Eakin, B., Kirk, R., Piechowski, P., Thomas, B., 2014. Piloting a nationally disseminated, interactive human subjects protection program for community partners: design, content, and evaluation. Clinical and Translational Science 7 (2), 177–183.

Steinbock, B., Arras, J.D., London, A.J., 2003. Ethical Issues in Modern Medicine, sixth ed. McGraw-Hill, New York.

Strauss, R.P., Sengupta, S., Quinn, S.C., Goeppinger, J., Spaulding, C., Kegeles, S.M., Millett, G., 2001. The role of community advisory boards: involving communities in the informed consent process. American Journal of Public Health 91 (12), 1938–1943.

The Nuremberg Code, 1949. http://www.hhs.gov/ohrp/archive/nurcode.html (accessed 07.04.14.).

Trimble, J.E., Fisher, C.B., 2006. The Handbook of Ethical Research with Ethnocultural Population and Communities. Sage Publications, Thousand Oaks, CA.

United Nations Educational, Scientific and Cultural Organization, 2005. Universal Declaration on Bioethics and Human Rights. http://portal.unesco.org/en/ev.php-URL_ID=31058&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed 07.04.14.).

United Naitons, 1989. Convention on the Rights of the Child. http://www.ohchr.org/EN/ProfessionalInterest/Pages/CRC.aspx (accessed 07.04.14.).

U.S. Public Law 93–348. http://history.nih.gov/research/downloads/PL93-348.pdf (accessed 07.04.14.).

U.S. Code of U.S. Federal Regulations 45CFR46, 1981, revised 2009.

VanderWalde, A., Kurzban, S., 2011. Paying human subjects in research: where are we, how did we get here, and now what? The Journal of Law, Medicine & Ethics 39 (3), 543–558.

Weijer, C., Miller, P.B., 2003. Rehabilitating equipoise. Kennedy Institute of Ethics Journal 13 (2), 93–118.

Wolfe, D., Cohen, J., 2010. Human rights and HIV prevention treatment and care for people who inject drugs: key principles and research needs. Journal of Acquired Immune Deficiency Syndrome 55 (Suppl.), 56–62.

World Health Organization, 2011. Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants. http://whqlibdoc.who.int/publications/2011/9789241502948_eng.pdf (accessed 07.04.14.).

World Medical Association, 1964. Declaration of Helsinki. http://www.wma.net/en/30publications/10policies/b3/ (accessed 07.04.14.).

Research Funding: Ethical Aspects

Virginie Pirard, Institut Pasteur, Paris, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by Ørjar Øyen, volume 19, pp. 13240–13242, © 2001, Elsevier Ltd.

Abstract

Ethical concerns related to the research funding process are usually connected with the question of influence between the funder and the beneficiary. This influence can be clearly demonstrated whether the funder is a profit-seeking company or not. However, because of its possible negative outcomes on paramount ethical values such as freedom of research, distributive justice, and scientists' responsibility toward community needs and goals, this issue should be carefully assessed at every stage of the research process. A crucial way to assess these links is to maximize transparency and to raise both the scientific community and civil society's awareness.

Sources of Ethical Concerns

Ethical concerns occur when some ethical principles might be affected in a given situation. What kind of principles and situations are involved regarding research on the one hand and funding on the other?

Freedom of research appears as one of the most well-known fundamental principles in science because of its crucial link with the claimed goal of Science: seeking truth on the basis of facts.

This 'simple' goal has caused a lot of controversy in the last centuries about the meaning of 'truth,' 'facts,' and more recently, 'objectivity' in science. These debates are beyond the scope of this article. However, it appears that production of a valid scientific knowledge requires strong and defined methodologies, allowing replication of the experiments and verifiability (or falsifiability, following K. Popper's philosophy) of the experiments' conclusions, as well as protection against any undue influence. In that perspective, scientific knowledge is socially considered as objective, even if, from an epistemological point of view, objectivity in research seems more like an ideal than an always-present-reality.

If the pursuit of objectivity in science requires scientific researchers to be free from any undue influence, the funding process can however be a source of influence between the funder and the beneficiary. This is why the funding issue should be carefully assessed at all stages of a research project. Indeed, undue influence could affect the selection of the research area to be investigated as well as the research process itself: methodologies to be used, production of the results, dissemination and exploitation of the results, etc.

Another ethical concern must be considered when discussing the influence of research funding on the selection of the research area to investigate: could influence harm the principle of social and distributive justice which should regulate the allocation of limited resources? This is particularly critical for research related to health, especially when research concerns neglected areas such as rare diseases, pediatric treatments, or health needs in developing countries. Research funding in these areas generates debates on opportunity cost and equality of opportunity for all (Pinxten et al., 2012). It means that when a (public or private) organization decides – or not – to fund a certain type of project, its decision has an impact on the rest of the research funding issues, at least on a theoretical level. In other words, research funding contributes to the establishment of a scale of values which has consequences far beyond the relation between the funder and the beneficiary and leads to maximizing or minimizing existing inequalities in the world. This should be kept in mind by all.

This second ethical concern, less obvious than the first one – freedom of research – leads us to addressing a third point of concern: the responsibility of scientists and funders. The idea of research as an isolated activity, with no link to the actual needs of societies, appears definitely outdated (Marchant and Bird, 2009), even when speaking about basic research. While public funders are more and more inclined to question the social desirability of a proposal before deciding whether they want to fund it or not, private funders are under scrutiny to justify their choice(s) beyond the question of their own interests. Scientists are required to identify and motivate how their project, if funded, would contribute to the progress and well-being of society. This again shows how asking for or granting funding bears responsibility toward community needs.

Outcomes of Influence

Freedom of Research

If funding – whether granted by a public institution or a private organization – seems to generate the possibility of influence between a funder and its beneficiary, this potential influence is particularly problematic when scientific researchers are working on topics connected with general interest (e.g., public health) although their research are, at the same time, funded by a profit-seeking organization or group. Diverging but coexisting interests are mixed in these situations and the capacity of scientists to preserve the objectivity of the research process and the integrity of their research results is questionable. A typical example of this situation is provided when research is funded by gambling, alcohol, or tobacco companies, as recently shown by the addiction industry studies – this new type of studies focused on the relationships that determine policy and regulation in the area of addiction (Adams, 2013).

When Research Is Funded by Companies Which Sell Products with Addiction Potential

These studies reveal some links between scientists and gambling, alcohol, or tobacco companies. These links have the potential to bring an ethical tension between current mixed and diverging interests in research related to addictive products. The sensitive question would then be to determine the nature of these links, and their interactions with scientific objectivity and the stakes relevant for public health.

Of course, the relation between research and addiction companies is paradigmatic of the ethical concern related to research funding because, *in casu*, the discrepancy between the general interest (public health interest) and the for-profit interest is obvious.

Although selling addictive products with potential negative impact on health, private companies aim to show that they care about public health and wish to assume their social responsibility (Babor and Robaina, 2013). This is the reason why they develop 'social aspects/public relation' organizations (SAPROs). These SAPROs, which present themselves as independent and not-for-profit organizations, take part in public debates with policy-makers about public health issues related to tobacco, alcohol, or gambling, and provide limited research project grants to young scientists. Even though they claim themselves as an 'evidence-based organization,' it appears after strong review (Miller et al., 2011), that the research projects funded by SAPROs tend to: (1) deflect attention away from other research areas usually considered as promising on a public health level (Adams, 2013); (2) promote basically only industry-friendly interventions; and (3) avoid to compare, on a scientific basis, these interventions with other actions positively assessed by independent peer-review research. In addition, if it happens that SAPROs are provided with 'nonindustry-friendly' results from a research they funded, evidence is shown that they usually do not refer to these results in their communication campaign.

As a consequence, SAPROs are active in several critical areas for science: sponsorship of scientific research, influence on public perceptions of scientific findings, dissemination of scientific information, and support to public policy actions. However, "case studies and ethical analyses have suggested that active collaboration with the alcohol industry contributes little to the advancement of science or public health" (Babor and Robaina, 2013: p. 211).

This shows the risk for research and scientists when funded by private and profit-seeking companies with interests diverging from the general interest: the potential loss of the capacity to produce 'objective' knowledge. Even if the production of knowledge remains objective, there is a risk of misuse and nondissemination of the research results with consequence on peer-reviewed and evidence-based science as well as for public policies.

Distributive Justice

Funding a research is not an isolated action: it can indeed lead to preoccupation while also contributing to maintaining others silent, given that the funding possibilities are, sadly, limited. Therefore, the question is whether the allocation of these restricted sources takes into account social and distributive justice or contributes to increasing existing inequalities.

When Research Is Funded by Pharmaceutical Industries

Pharmaceutical drugs are supposed to bring at least one of these clearly identified public health's benefits: prevention, relief, cure, or treatment. This should guarantee at least a minimal convergence between the interests involved. In that perspective, the ethical tension observed regarding funding research by addiction companies appears to be lower when pharmaceutical companies are the funders, whether they fund research to develop their own products or are involved in sponsorship activities.

However, pharmaceutical research is currently regulated by the patenting system which "attempts to promote pharmaceutical research by rewarding pharmaceutical companies with market exclusivity for their product. Such a system does not provide enough market incentives to develop innovative, effective, and affordable medicines or health technologies because it focuses on the market competition between pharmaceutical companies on bargaining for global monopolies for their products, (…)" (Oprea et al., 2013: p. 311).

The pharmaceutical research consequently still focuses on 'me-too' medicines (noninnovative drug, similar to an already existing drug, with small health improvements, dedicated to a large population) rather than to take responsibility to create innovative drugs for neglected diseases or for pediatric populations. Even if some major pharmaceutical groups have taken decisive actions to assume their social responsibility and correct that trend, the patent system still contributes to separate the agenda of pharmaceutical research and the public health priorities that replicates and increases inequalities between developed countries and low- and middle-income countries.

This situation has been deeply analyzed by the Consultative Expert Working Group on Research and Development (CEWG) established by WHO in 2010 and a panel of solutions was examined by this Group in order to correct it, at least partially. As this Group showed, this kind of situation also exists when grants are provided by public institutions: "public investment is also dominated by the rich world and its own health needs" (WHO, 2012: p. 1). As a consequence, some research areas are neglected (and not funded) because the definition by the developed countries of public health priorities in the low- and middle-income countries is shaped according to certain perceptions and data: current funding targets research on communicable disease at the expense of noncommunicable disease and other research areas such as resuscitation research (Aufderheide et al., 2013). In that perspective, research funding issues have a political dimension and appear to be connected with responsibility of scientists and science.

Responsibility of Scientists and Science

The idea that the community has ability to hold scientists to account for their research is quite old as Galileo Galilei's case shows it for the worse. In response, the conception of freedom of research and science grew during the seventeenth and eighteenth centuries. This would imply that science should always have sufficient autonomy from any moral, religious, political, or social constraints. However, except for very limited areas of

basic research, the knowledge for its own sake no longer seems relevant. Since many atrocities were carried out in the name of science during the twentieth century – and especially during World War II – the idea that science is always and by itself a factor of progress for humanity is no longer acceptable. In addition to complying with the legal framework, scientists must now justify the interest of their research from a public interest point of view. This trend is clear when observing the criteria used by public organization or agencies to provide funding even when they support fundamental and nonmedical research, as it is the case for the National Science Foundation (NSF). This agency of the U.S. federal government supports fundamental research and education in all nonmedical fields of science and engineering (Davis and Laas, 2013). In its process to grant funding, this agency assesses the kind of 'broader impacts' a research project could bring. By doing this, the funding agency values the advancement of scientific knowledge and activities that contribute to the achievement of societally relevant outcomes. These outcomes may include, among others, the full participation of women; persons with disabilities and underrepresented minorities in science, technology, engineering, and mathematics. Another desirable outcome would be to improve the well-being of individuals in society (NSF, 2013). This criterion also converges with the 'Responsible Research and Innovation' one. Both criteria emphasize the potential desirability of a research for a given society. As a consequence, scientists should not only comply with (1) the respect of national and international regulations, but also (2) the deontological rules of their corporation: they also have to face the ethical necessity to take into account any possible outcome of their planned activity, even if this outcome originates from a nonscientific use of the research results. Due to this evolution, scientists are weighed down by a heavy responsibility – to which scientific studies prepare only poorly, even if there are current debates about how a Scientist's Oath (Revill and Dando, 2006) could be an interesting tool to raise young scientists' awareness of their responsibility. This responsibility concerns as much the risks to avoid as the benefits to expect.

Risks to Avoid – The Example of Dual Use

If scientists are, in principle, trained to take a certain number of precautions when carrying out laboratory/research experiments, things are not as straightforward when considering risks generated by an intentional misuse of a technology or a scientific discovery. The possibility for any knowledge or scientific savvy to be dealt within a dual use must therefore be carefully assessed.

In the United States, the National Science Advisory Board for Biosecurity defined dual-use research as "the development of new technologies and the generation of information with the potential for both benevolent and malevolent purposes" (NSABB, 2007). In Europe, the European Commission chose in 2009 to provide the public with a list of potential dual-use items and defined these as "items, including software and technology, which can be used for both civil and military purposes, and shall include all goods which can be used for both non-explosive uses and assisting anyway in the manufacture of nuclear weapons and other nuclear explosive devices" (Council Regulation (EC) n°428/2009).

The risk of dual use is potentially present in all life sciences research but especially feared in bacteriology and virology research when research experiments could lead to increasing the pathogenicity of viral agents and toxins. The identification and management of this kind of risks require that scientists proceed to a particularly difficult evaluation: beyond their own goal and the scope of their research activities, could their research lead to an additional risk of misuse? In that kind of assessment, also required by the European Commission as Research Funding Agency, scientists have to imagine all the possibilities – even those completely independent of their will and scientific goal – that could lead to a potential misuse such as bioterrorism. Besides the specific difficulty of such evaluation, the scientific community is currently divided on how to deal with a research that could lead to such risks.

While some scientists wish to ban all experiments which could possibly lead to such potential risks (Wain-Hobson, 2013), part of the scientific community considers that these experiments should be conducted if they have a decisive public health interest and if an adequate design and a strong follow-up of this risk during the project's life are provided. While the controversy within scientific community is still ongoing (Casadevall and Shenk, 2012; Webster, 2012), public research funding agencies seem to agree with the later position, at least for the moment. It is worth mentioning that this type of assessment and balance between risk and benefit of research projects with potential dual-use concern is considered as an 'ethical dilemma' by the European Commission. This dilemma requires that scientists should never be left alone with such assessment and responsibility. They must be accompanied by an independent ethics board that could require some specific measures to be taken in order to limit the risk(s): for example, it could recommend the restricted dissemination or even the nonpublication of the research results, which is of course a very difficult situation knowing the existing pressure on scientists to publish their research findings.

The Benefits to Expect – Material and Immaterial Outcomes

As already said, project leaders should be able to anticipate the kind of benefits to expect from a planned project. This requirement seems more adapted to applied science than to basic research. However, the evaluation of the benefits is required even by some agencies (see the example of NSF in the United States) which do fund basic research. The benefits could concern a fully materialized finding (a new technology, a know-how, etc.) or the promotion by and during the research of some socially desirable values.

This points out that, even when research is reputed for being free, knowledge produced by science is not value-free: as noted by Drenth (2006: p. 15), "Scientists cannot avoid the meta-scientific question of whether what they pursue is worth knowing. They have to justify – not only to themselves but also publicly if the taxpayer or a sponsor's money is involved – why scientific issues need to be addressed. In essence, this justification implies non-scientific choices and decisions."

After World War II's atrocious experiments conducted by Nazi doctors on prisoners and detainees, a trial – the Doctors' Trial – was held before the U.S. Military Court in Nuremberg. A part of the sentence given by the law court is well-known as the 'Code of Nuremberg.' This 'Code' sets out the minimal ethical

conditions that any scientific experiment involving human subjects should meet. This list has been adopted and reviewed by several International Ethical Guidelines such as, among others, the WMA Declaration of Helsinki (World Medical Association, Ethical Principles for Medical Research Involving Human Subjects, 1964) and the CIOMS guidelines (Council for International Organizations of Medical Sciences, International Ethical Guidelines for Biomedical Research Involving Human Subject, 1982).

It is interesting to note that, at least since these events, there is an international ethical consensus to consider that a research involving human beings should always prospect to discover new ways of benefiting people's health by improving preventive, diagnostic, or therapeutic interventions. In other words, a research project involving human subjects, conducted for pure scientific sake without any possibility to contribute to these three goals (prevention, diagnostic, treatment) would be unethical. Sometimes, the lessons of history are so strong that no responsible science should ignore them.

How to Deal with Influence: The Ways to Increase Transparency

As previously described, influence is clearly a part of the research funding reality between the provider and the beneficiary. The consequences of this influence may be various for several crucial parameters of the research: from the selection of a scientific question or research area to investigate to the possibility to produce 'objective' knowledge and data, including the way – and sometimes the simple possibility – to disseminate the obtained scientific results. As seen, influence is not systematically a bad thing: indeed, public funders force the scientists who apply to their call to take into account any possible outcome of the research designed in the proposal and, even if it is difficult for the researchers, this is a positive inducement toward a responsible science.

It would however be considered as taking a shortcut to conclude that influence is positive when coming from a public funder, but rather negative when stemming from a profit-seeking organization or company. The issue is complex and no efficient solution can be found if we just look at the source of the funding, although this is of course an important point. In the face of the situation where the influence's paths and nature can vary from one research to the other, all efforts should be made to increase access to information regarding the links between funding providers and beneficiaries. If public funding providers usually require that the scientists disclose the origin of a research funded by them in each paper to publish, things are less clear-cut when the funding providers are for-profit companies. Usually, obscurity is the trend. Following a movement initiated by Ethical Guidelines and, later, by the International Committee of Medical Journal Editors (ICMJE), public authorities in the United States and Europe (see for France, the "LOI n° 2011–12 du 29 décembre 2011 relative au renforcement de la sécurité sanitaire du médicament et des produits de santé") have implemented – or are on the point of doing it – strict regulations increasing the disclosure of links of interests between on the one hand drug manufacturers and, on the other, physicians and researchers.

The Example of the Physician Payment Sunshine Act

Since the beginning of 2013, the rules implemented by the Physician Payment Sunshine Act (2010) have become effective (42CFR parts 402 and 403). This Act sets forth that pharmaceuticals companies, medical device companies, and other manufacturers of medical supplies must collect, gather, and disclose information on any value transfer they provided to physicians when these transfers are over $10 each, or when several payments under $10 to the same physician the same year account for $100 total.

These payments concern, among others, research grants.

The relevant information is to be reported to the Centers for Medicare and Medicaid Service (CMS) which will register, compile, and publish them on a public database by September 2014 (Social Security Administration, 2013).

The obvious goal of this regulation is to increase transparency to the public, a critical point given that data suggest that there is a relation between exposure to pharmaceutical company information – even when this information is claimed as strictly scientific and educational – higher prescribing frequency and higher costs (Spurling et al., 2010). Therefore, what about the influence toward money, gifts, inducements, and grants from these companies when speaking about researchers? As noted by Sismondo (2013), the payments made to researchers are intended to lead medical opinion toward a certain direction, the companies' favorite one. Companies aim to enroll researchers, or at least some of them, especially because they are considered as key opinion leaders (KOLs). Funding their research could contribute to shaping the scientific opinion and, by doing so, to getting influence on other members of the medical community. The tricky point is that this influence is empowered by the reputation of integrity, freedom, and independence of research and scientists, who are supposed to speak only for the sake of science. This situation should not lead to considering that any information provided by a KOL researcher is false or incorrect. However, it underlines how influence plays its game in the process of producing science and medical opinion. This is why some authors call for the 'sequestration' of clinical trial research on drugs or complete separation between pharmaceutical research and marketing unit (Sismondo, 2013: p. 641).

International Ethical Guidelines

Regarding the research involving human subjects, it is interesting to point out that the way of transparency, which a few legislative systems implement and make at present mandatory, was anticipated a long time ago by the International Ethical Guidelines to the research participants' benefit. Indeed, following the WMA Declaration of Helsinki, "each potential subject must be adequately informed of the (…) sources of funding, any possible conflicts of interests, institutional affiliations of the researcher (…)" (WMA Declaration of Helsinki, point 26). Similar requirement is stated in the CIOMS guidelines. Transparency is required in these international guidelines, especially for the research participants. Participants are provided with the relevant information on these issues, and therefore in a position to give their informed consent. Indeed, the disclosure of such information

allows the participants to check if their moral right to choose the kind of research they accepted to support is respected. Participants, yet more than anybody else, have the right to know if private companies will be the sponsor and for which purpose the trial will be held. This may be particularly important for healthy volunteers, who are often driven by altruistic intentions. Their way to conceive their contribution to science could be in complete opposition with some for-profit goals. Transparency should allow these volunteers and any other participants to a research to check if the research project they planned to be involved in is consistent with their moral values.

ICMJE's Recommendations

At the end of the 1970s, a group of editors decided to issue recommendations for the publication of manuscripts in biomedical journals in English. These recommendations, initially focused on formal aspects and standardization, have since been reviewed several times. Issues related to potential conflicts of interest were mentioned for the first time in the 1988 update, in the 'Acknowledgments' section of the recommendations. The recommendation was that "at an appropriate place in the article" should one or more statements specify acknowledgments of financial and material support. The nature of the support and "financial relationships that may pose a conflict of interest" should be specified (ICMJE, 1988).

In the 2004 update, instead of being a part of the Acknowledgments section, the issues related to conflicts of interest were included in the section dedicated to 'Ethical considerations in the conduct and reporting of research.' This section mentioned several rules regarding potential conflict of interest related to project support:

- Researchers should not enter into agreements that could interfere in any ways with their capacity to conduct independently a research and to proceed to its publication.
- The role of the study sponsor(s) should be described at every step of the research process and Editors "may request that authors of a study funded by an agency with a proprietary or financial interests in the outcome sign a statement in which he/she declares that he had full access to all of the data in this study and that he takes complete responsibility for the integrity of the data and the accuracy of the data analysis."
- In addition, Editors should review the protocol and the contracts associated with the research project under review. They may choose not to publish an article if a sponsor has asserted control over the authors' right to publish (ICMJE, 2004).

The updating process of the ICMJE guidelines reveals how the issues related to the conflicts of interest have been regarded by scientific community throughout the last decades. At the beginning, it was considered nearly as an endogamous question, which emphasized the links within the scientific community and considered that these links deserve acknowledgment. Now, the disclosure process of the conflicts of interest enables readers as well as the entire civil society to check how a research project has been funded and conducted from its start to the publication process. Regarding this, it is crystal clear that the ethical concerns related to freedom of research must be shared and managed by both the scientific community and civil society.

See also: Ethics for Biomedical Research Involving Humans: International Guidelines; Informed Consent of Research Subjects; Objectivity of Research: Ethical Aspects.

Bibliography

Adams, P.J., 2013. Addiction industry studies: understanding how proconsumption influences block effective interventions. American Journal of Public Health 103, e35–e38.

Aufderheide, T.P., Nolan, J.P., Jacobs, I.G., et al., 2013. Global health and emergency care: a resuscitation research agenda part 1. Academic Emergency Medicine 20, 1289–1296.

Babor, T.F., Robaina, K., 2013. Public health, academic medicine, and the alcohol industry's corporate social responsibility activities. American Journal of Public Health 103, 206–214.

Casadevall, A., Shenk, Th, 2012. The H5N1 moratorium controversy and debate (editorial). mBio 3 (5), e000379–12, 1–2. http://dx.doi.org/10.1128.

Council regulation (EC) n°428/2009 of 5 May 2009 setting up a community regime for the control of exports transfer, brokering and transit of dual-use items. Official Journal of the European Union, 05 29, 2009. L134/1. Available from: http://trade.ec.europa.eu/doclib/docs/2009/june/tradoc_143390.pdf (accessed 30.01.14.).

Davis, M., Laas, K., 24 October 2013. "Broader impacts" or "responsible research and innovation"? A comparison of two criteria for funding research in science and engineering. Science and Engineering Ethics. http://dx.doi.org/10.1007/s11948-013-9480-1 [PubMed].

Drenth, P.J.D., 2006. Responsible conduct in research. Science and Engineering Ethics 12, 13–21.

International Committee of Medical Journal Editors (ICMJE), 1988. Style Matters. Uniform requirements for manuscripts submitted to biomedical journals. British Medical Journal 296, 401–405. http://www.icmje.org/recommendations/archives/1988_urm.pdf.

International Committee of Medical Journal Editors (ICMJE), 2004. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication. ICMJE, 1–15. http://www.icmje.org/recommendations/archives/2004_urm.pdf.

Marchant, G.E., Bird, S.J., 2009. Editors' overview: forbiding science? Science and Engineering Ethics 15, 263–269.

Miller, P.G., de Groot, F., McKenzie, S., Droste, N., 2011. Vested interests in addiction research. Alcohol industry use of social aspect public relations organizations against preventive health measures. Addiction 106, 1560–1567.

Oprea, L., Braunack-Mayer, A., Gericke, C.A., 2013. Ethical issues in funding research and development of drugs or neglected tropical diseases. Journal of Medical Ethics 35, 310–314.

Pinxten, W., Denier, Y., Dooms, M., Cassiman, J.-J., Dierickx, K., 2012. A fair share for the orphans: ethical guidelines for a fair distribution of resources within the bounds of the 10-year-old European orphan drug regulation. Journal of Medical Ethics 38, 148–153.

Revill, J., Dando, M.R., 2006. A Hippocratic oath for life scientists. EMBO Reports 7, S55–S60.

Sismondo, S., 2013. Key opinion leaders and the corruption of medical knowledge: what the sunshine act will and won't cast light on', institutional corruption and the pharmaceutical industry. The Journal of Law, Medicine & Ethics 41 (3), 635–643.

Spurling, G.K., Mansfield, P.R., Montgomery, B.D., Lexchin, J., Doust, J., et al., 2010. Information from pharmaceutical companies and the quality, quantity, and cost of physicians' prescribing: a systematic review. PLoS Medicine 7 (10), e10000352. http://dx.doi.org/10.1371/journal.pmed.1000352.

Wain-Hobson, S., November 2013. Pandemic influenza viruses: time to recognize our inability to predict the unpredictable and stop dangerous gain-of-function experiments. EMBO Molecular Medicine 5 (11), 1637–1641.

Webster, R.G., 2012. Mammalian-transmissible H5N1 influenza: the dilemma of dual-use research. mBio 3 (1). http://dx.doi.org/10.1128/mBio.00005-12.

Relevant Websites

http://www.cioms.ch/publications/layout_guide2002.pdf – Council for International Organizations of Medical Sciences (CIOMS), 1982. International Ethical Guidelines for Biomedical Research Involving Human Subject (Last Update, 2002) (accessed 30.01.14.).

http://www.icmje.org – International Committee of Medical Journal Editors (ICMJE). Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals (Last Update, 2013) (accessed 30.01.14.).

http://oba.od.nih.gov/biosecurity/biosecurity_documents.html – National Science Advisory Board for Biosecurity (NSABB), June 2007. Proposed Framework for the Oversight of Dual Use Life Sciences Research: Strategies for Minimizing the Potential Misuse of Research Information (accessed 30.01.14.).

http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#IIC2d – National Science Foundation (NSF). Grant Proposal Guide (accessed 30.01.14.).

http://www.ssa.gov/OP_Home/ssact/title11/1128G.htm – Social Security Administration (SSA). Transparency Reports and Reporting of Physician Ownership or Investments Interest, Sec. 1128G. [42 U.S.C. 1320a-7h] (accessed 30.01.14.).

http://www.wma.net/en/30publications/10policies/b3/index.html – World Medical Association (WMA), 1964. Ethical Principles for Medical Research Involving Human Subjects (Last Update, 2013) (accessed 30.01.14.).

http://www.who.int/phi/cewg_report/en/index.html – World Health Organization (WHO), 2012. Report of the Consultative Expert Working Group on Research and Development: Financing and Coordination (accessed 30.01.14.).

Research Publication: Ethical Aspects

Federico Di Trocchio, Università di Roma Sapienza, Roma, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethical aspects in research publication are all the moral issues and problems raised by any behavioral misuse or abuse of the system for communicating scientific results and information within the international scientific community. Misuses or abuses include plagiarism, self-plagiarism, fraudulent and repetitive publication, unethical or fabricated data publication, censoring negative data, guest-, ghost-, and gift-authorship in multiauthored papers, conflict of interests involving authors, editors, or referees, censoring and rejecting publications on biased judgment, selective reporting, copyright breaches, confidentiality and informed consent infringements.

Such unethical behaviors are tightly related to the misconduct phenomenon, the common origin of both lying on the conflict between the founding ideals of the scientific community and the actual scientific practice, which is far from follow the Merton's norms. Here proposed explanations are discussed as well as ways to prevent or stop the phenomenon.

The intellectual style and the social structure of modern scientific community trace back to sixteenth-century *Respublica litterarum* and eighteenth-century *République des lettres*. This was a virtual supranational community of intellectuals embracing the whole world, all nationalities, all social classes and whose main concern was the advancement of learning. Members conceived of knowledge as 'common property' according to the natural right. Cosmopolitanism, integrity, impartiality, and independence from politics, religion, and economic constraints were its essential and ideal characteristics; so that the republic went eventually to be steadily grounded at least on three of the four Mertonian norms: universalism, communality, and disinterestedness. There was no competition other than for recognition of intellectual excellence and priority.

This was however mainly a literary enterprise and the scientific one was born as a specialized province of it for which M. Polanyi (1962) coined the new sociological concept of 'republic of science.' Specialization was acquired mainly through scientific method by which the ethical norm of organized skepticism was also firmly stated. Scientists developed an explicit conception of experimental procedures, use of controls, repetition of experiments, exhibition of numerical results of actual measurements as well as those calculated from theory, recognition of approximations, and estimations of experimental errors. On this basis scientific knowledge came to be conceived as community property unceasingly increased by individuals or groups in mutual cooperation and fair and honest competition; and researchers are intended as primarily engaged in the disinterested pursuit of truth and in the quest of new insights and discoveries. Honesty toward oneself and toward others is considered as a fundamental condition for cooperation and fair competition, since researchers depend on each other and cannot be successful unless they are able to trust each other and their predecessors.

Crucial was the radical innovation of the communication system. The printed book and scientific correspondence were the essential means of detailed informal exchanges of ideas and information within the republic of letters. Within the republic of science the function of private correspondence was taken over and that of books modified by the professional scientific journal as a medium for reporting experimental research into specific problems. The original and main aim of a scientific paper was to give details of how to repeat experiments (repeatability). Today merely to rediscover the results of already published investigations is considered quite trivial: we mainly rely upon the honesty and competence of the original observers. In other words, the critical role of each scientist with respect to the experimental work of his/her colleagues has been somewhat abrogated but still exists in principle: there is always the likelihood that experiment will be repeated.

To warrant this moral and professional commitment and, at the same time, ensure clarity and quality, the scientific paper has been submitted to the peer review process, and progressively restyled in a standardized format. The process of soliciting peers (experts) to evaluate scholarly work prior to publication was initiated in 1731 by the *Medical essays and observations* of the Royal Society of Edinburgh but became of general use as a full structured practice only in the middle of the twentieth century. In the meantime, the article evolved from the letter form and purely descriptive style in the seventeenth century, to a very standardized structure in four main sections: introduction, methods, results, and discussion (IMRAD format) which began to be used in the 1940s and was adopted as a majority in the 1970s. Publications came to be the primary medium through which scientists give an account of their work, cooperate, contribute to the advancement of knowledge, and claim the rights to be recognized for their discoveries. Therefore, dishonesty in publication not merely throws research open to doubt but destroys the very fundament of science.

On the basis of these ethical and professional premises the scientific enterprise grew up as an effective and reliable system for the advancement of learning from sixteenth to nineteenth century.

Problems began to rise in the 1940s when modern science policy was established in view of the expected contributions of the scientific community to national defense and industrial progress. Paradoxically enough it was in these very years that Merton (1942) delineated the 'ethos of science,' a set of descriptive norms (communism or communality, universalism, disinterestedness, organized skepticism) that govern

scientists' relationship with each other. It has been noted (Macfarlane and Cheng) that in the 1940s as today Merton's norms were anachronistic and no more descriptive of actual scientific practice. It has been even asked if would it not be appropriate for a scientist to promote the norms of capitalism, particularism, and interestedness as key elements during the daily struggle for reputation, funding, and research grants. Especially disinterestedness as a scientific norm loses support and aligning research with funding opportunities has become common. It seems apparent that professionalization, industrial and military involvement, and the emerging of science policy in Western world created a general condition of potential conflict between the actual practice and the founding ideals of the scientific community. Contemporary escalation of misconduct and ensuing infringements of publication ethics are surely to be linked to this conflictual condition.

The new reward and funding system enhanced the traditional competition between scientists and it was then, in 1942, that the sociologist L. Wilson apparently coined the expression 'publish or perish.' It could hardly be said however that in the first 30 years competition did in some way sensibly promote fraudulent behaviors or the pollution of scientific literature. In fact the turning point has been stated in the late 1970s when the social structure of the scientific community underwent some crucial changes due mainly to the exponential increase of the number of people involved in research and the proportional (though not absolute) decreasing rate in assigning funds to research activity. The ensuing economic constraint (upon which the resources poured in by private companies had no or little effect) led to a situation of even stronger competition among scientists for academic rewards, position, or research funds. One crucial decision was that of the adoption of publication as the main performance indicator and of measuring productivity as the number of publications per length of time. The additional tool of citation analysis which was introduced in those very years (Garfield, 1972, 1979) and based on the key concept of 'impact factor' (as the ratio of the number of citations a journal receives to the number of papers published over a particular time period), and then extended to calculate the impact of the work of individuals, groups, or departments generated a race to produce in the shortest possible time the least publishable unit in a journal with the highest impact factor. Currently over 7.1 million researchers in the world compete to have their research published in over 80 000 scientific journals. At the same time the growth of the scientific population implied a statistical decline in ethical standards and in the quality of production. Lotka's law of productivity in fact states that in a certain field half of the scientific literature is produced by a population of the square root ($\sqrt{N}$) of all N scientists and De Solla Price pointed out that the exponential growth of the scientific population implies a proportional increase in the number of scientists able to author professional papers but not important and innovatory articles. From a statistic compiled by the Institute for Scientific Information (ISI) it resulted for example, that 55% of the scientific papers published between 1981 and 1985 in journals indexed by the Institute received no citation at all (Hamilton, 1990), leading some students to draw the conclusion that more than half or perhaps more than three-fourth of the scientific literature is worthless; while others reacted against the notion that most uncited papers are without value. As for the use of citations as a measure of impact and of the quality of a publication it has been questioned whether this may not lead to abuses such as citation cartels and it was estimated by the same ISI that self-citation accounts for between 5 and 20% of all citation.

It is from, or in any case within, those changes that, according to the current views (see, for example, Petersdorf, 1986) the pathological phenomenon arose of scientific misconduct to which the major ethical issues in scientific publishing are tightly related. As a consequence of the analysis of these issues, the evaluation of the frequency and gravity of the transgressions and the discussions on how to manage the problems they pose have been conducted as part of a general study on scientific misconduct.

The publication of a fraudulent article (the very outcome of a scientific fraud consisting in the fabrication, falsification, or in 'cooking' and manipulating statistical data) remains the most serious transgression of ethics of scientific publication and is commonly considered as an essential part of the scientific misconduct as well as plagiarism, the deliberate presentation of another's ideas or texts as one's own. Compared to these, repetitive publications (publishing the same article in different journals), sequential publication (reporting follow-up of the same study with additional subjects but without new results), 'salami slicing' (dividing up different parts of the same study for publication in different journals) and guest- and ghost-authorship (giving or not giving authorship credit to persons who did or did not take part in the research or in drafting or revising the article) are all considered as minor transgressions resulting in a huge amount of unwieldy and polluted literature which distorts the direction of science. In the 1980s it was roughly calculated that flawed papers published on the 5600 scientific journals indexed at the time by the Science Citation Index amounted to about 10 000 each year. It results from the PubMed database that of the 9 398 715 articles published between 1950 and 2004, 596 were retracted, and it has been calculated (Van Noorden, 2011) that in the past decade the number of retractions has shot up 10-fold while the literature has expanded by only 44%. According to a recent analysis (Fang et al., 2012), about 67% of 2047 studies retracted from biomedical and life science journals resulted from scientific misconduct, while about 21% of the retractions were attributed to a scientific error.

This raises the problem of the reliability of the peer review system designed to assess the quality of scientific literature, and that of the fate of the invalid ideas and data generated by fraudulent publication (Friedman, 1990).

Good scientific journals today publish original articles only after they have been examined by competent reviewers for their validity and originality. The system is credited to have greatly contributed to the efficiency and professional organization of the scientific enterprise but since the 1980s it has been severely criticized not only for its enormous costs in time and money but also for the progressive loss of accuracy (Lock, 1986). Reviewers have indeed some moral obligation to make every necessary effort to recognize manipulations and falsifications, but to expect irregularities to be reliably detected would however be misguiding: the original data are not available to reviewers, and if they were they would not have time to

replicate experiments and observations. This is why, according to many, peer review is so vulnerable to dishonest conduct. Other flaws have been detected in negligence (it resulted from inquiries that the higher the academic status and age of the reviewer, the lower the quality of the review), in vested interests of the referees and in many forms of bias (against particular individuals, subjects, or institution) the worst of which is that against the innovatory. Some critics maintain that the peer review system favors mediocre basic science and fails to recognize and promote creativity, originality, and innovation (Horrobin, 1990; Campanario, 1996). Even the anonymity of the referees has been questioned for its finality as contrary to the moral norms of universalism, communality, disinterestedness, and organized skepticism that are thought to characterize the ethic of science. In fact, the peer review system holds some risks for the authors because ideas, research findings, and texts which are unprotected by patents or intellectual property rights are submitted to persons whose identity is unknown to the authors and who may happen to be their direct competitors.

As for the fate of invalid data generated by transgressions to the ethics of scientific communication, the scientific community has yet to agree on the level of inaccuracy required to mandate a retraction versus an erratum or who may retract an article; and though early in 1987 the International Committee of Medical Journal Editors published guidelines for the format and terminology of retraction notices, articles known to be fraudulent may be not retracted at all or retracted with ambiguous retraction while in other cases only part of the data or conclusions are retracted. Editors of journals have little leeway for action since publishing corrections is fraught with juridical risks unless they are jointly signed by all authors. Omitted or ambiguous retraction however are not the only reasons of invalid work not being effectively purged from the scientific literature: it resulted from inquiries on U.S. authors (Pfeifer and Snodgrass, 1990) that after retraction the invalid information they produced continues to be used since the majority of retractions are not indexed by their own journals and no reliable source exists to locate fraudulent or erroneous work.

Some ethical relevance has also the soft censoring activity by which administrations and companies tend to control and restrict the dissemination of sensitive scientific and technological knowledge on the grounds of national security and economic reasons (Relyea, 1994).

Unfortunately, despite increasing publicity during the past 25 years, attempts to assess the quantitative dimensions of the problem of scientific misconduct did not lead to conclusive results: its magnitude is unknown and its detection largely serendipitous. Guidelines and actions adopted to manage scientific and (as a consequence) publication frauds and other abuses vary according to which of the two main different approaches they are based on. Sociologists and the vast majority of scientists share the opinion that the frequency and incidence of misconduct did not sensibly change during the twentieth century and that the impression that fraud has vastly increased may be the result of the heightened self-consciousness in society coupled with strong moral expectation that scientists in the pursuit of reliable knowledge will live up to higher standard of probity. This view seems to be substantiated both by the reports of the two most important U.S. authorities responsible for dealing with misconduct cases, the Office of Inspector General of the National Science Foundation, and the Office of Scientific Integrity of the Public Health Service and by that of analogous authorities in Denmark, Norway, Finland, and Australia. The ensuing suggested strategy is firstly and mainly intended to save the autonomy right and self regulation privilege of the scientific community from legal or government unnecessary intervention, and recommends that inquiries be made primarily by local academic bodies, stressing at the same time the crucial role of education to improve the ethical standards of researchers.

Opposite is the view according to which, though there are no official data on the scale of the problem, the reports suggest that only a comparatively small proportion of cases of misconduct is actually detected and that what emerges is just the tip of an iceberg (Lock, 1993; Fanelli, 2009). According to this view self regulation has proven to be ineffective and unsuitable: scientists are not trained in conflict and conflict of interests resolution; their intuitive response is usually wrong and they tend to set up shaky ad hoc procedures that do not guarantee the accused notice of all the charges, the opportunity to respond to all the charges and to the evidence, and a decision based on rigorous standards. In this view, the setting up of national agencies rather than local committees is recommended and the attention is focused on the economic and social background considering that those transgressions simply reflect the problems of a science which is today too big, too entrepreneurial and too competitive to guarantee its very scope: the cultural and material improvement of mankind, in due and reasonable proportion to the intellectual and economic resources and to a rate comparable to the past one.

As far as scientific publication is concerned this second approach led the Committee on Publication Ethics (COPE), created in 1997 in the UK, to ask for an independent agency to detect, handle, and prevent research publication abuses. COPE provides guidelines and a forum for its member journals (today the great majority) to seek advice from other members on troublesome cases. The establishment of COPE has increased awareness and education but it resulted from a recent analysis of cases managed between 1997 and 2010 (Foo and Wilson, 2012) that the number of ethical cases is not diminishing.

There is however no general agreement on the view which led some of its proponents to recommend for some radical adjustment in the scientific policy, first of all in connection to the ratio between scientific population and funds. While those proposals have been opposed tacitly or patently by universities, scientific societies, funding agencies, and governments, even their advocates acknowledge the importance of improving the ethical standards via the teaching of research ethics which is currently considered by competent institutions and single students besides and before any disciplinary, economic, or legal action. Likewise all agree on the necessity of suitable ethical and professional guidelines and measures which have in fact been proposed or adopted, though independently and according to different experiences and aims, by learning societies, agencies, journals, or individual students.

The primary importance of ethical (coupled with methodological) education emerges particularly in connection to some of the problems of scientific publishing which seem less directly linked to the social and economic structure of the scientific system than to the peculiarity of research areas which involve cultural diversity or human and animal rights. In fact though science is largely transcultural the human community has not yet developed a completely corresponding body of transcultural ethics. In some areas this may affect both methodology and ethics: uncritical and unwitting reliance on the researcher's own culturally received principles and values regarding personal and physical identity, health, loving, suffering, death, etc., may generate potentially distorted initial observations which can lead to systematic divergence from truth in anthropological, psychological, sociological, or medical inquiries (where also systematic infringements of human or animal rights or the exploitation of subjugated populations may occur). Scientific publications based on such inquiries are, in principle, to be considered as poorly or improperly designed and performed studies but current ethical and methodological standards do not include clear and effective rules for authors, editors, and peer-reviewers though the problems they raise have been debated in connection to some controversial cases.

A first general attempt to mitigate the pressure to publish that lies behind most deviant practices is, according to many, to place a ceiling on the number of publications that can be considered in evaluating a candidate for promotion or funding, but some experts have also suggested that the number of scientific journals be drastically reduced (Bracey, 1987). On the same line; although the use of quantitative indicators of scientific performance seems unavoidable, it has been recommended that impact factors shall not by themselves serve as the basis for funding decisions and that the quality of publication must be the primary consideration (Bucholz, 1995). While no effective measures (other than severe sanctions) seem foreseeable against major authorship transgressions, as plagiarism or the publication of fabricated data, it has been recommended to journals to cooperate in cross-checking control activities and to research institutes and universities to store the primary data on which publications are based. All sources and methods used to obtain and analyze those data should be, when required, fully disclosed and the disappearance of primary data from a laboratory considered as an infraction of the basic principles of careful scientific practice. On the other hand, since there is no universally agreed definition of authorship, specific criteria and requirements have been invoked as an acknowledgment or footnote in every coauthored paper to state the exact contribution by each author or the signing of a form where each author affirms having seen the final version and agrees to its publication. According to those proposed criteria the award of authorship should balance intellectual contributions to the conception of studies or experiments, to the generation, analysis, and interpretation of the data, and to prepare the manuscript against the collection of data and other routine work. If there is no task that can reasonably be attributed to a particular individual, then that individual should not be credited with authorship. In principle all authors must take public responsibility for the content of their paper and honorary or guest authorship is in no way acceptable. Against repetitive publication since 1969 the so-called 'Ingelfinger rule' has been proposed (Angell and Kassirer, 1991) according to which papers are submitted to journals with the understanding that they (or their essential part) have been neither published nor submitted elsewhere nor communicated to the media. Reviewers are supposed to vigilate upon transgressions and it is stated in some journals' guidelines that if they suspect misconduct, they should write in confidence to the editor. As for the same reviewers it is reminded that they are bound to the duty of confidentiality in the assessment of a manuscript and asked to disclose conflict of interests and to provide speedy, accurate, courteous, unbiased, and justifiable reports. The submitted manuscript should not be retained or copied and both editors and reviewers should not make any use of the data, arguments, or interpretations, unless they have the authors' permission. Editors should regard their function as ensuring that authors are treated fairly and their decisions to accept or reject a paper for publication should be based only on the paper's importance, originality, and clarity. To avoid the suppression of originality and creativity it has also been recommended that studies that challenge previous work should be given an especially sympathetic hearing and that independent ombudsmen for journals be appointed. Some few suggestions have also been advanced on the subject of external pressures as for example, that editorial decisions must not be influenced by advertising revenue or reprint potential, but it seems that little can be done for the secrecy imposed by industries or governments on sensitive research fields.

Envisaged actions and sanctions rank from a letter of reprimand and warning as to future conduct or publication of a notice of redundant publication or plagiarism, to formal withdrawal or retraction of the paper from the scientific literature, informing other editors and the indexing authorities or reporting the case to the authorities or organizations which can investigate and act with due process.

There is however no suitable insight on the impact those guidelines and measures had or may have in the long term.

See also: Animal Rights in Research and Research Application; Censorship in Research and Scholarship; Communalism; Cultural Relativism, Anthropology of; Informed Consent of Research Subjects; Objectivity of Research: Ethical Aspects; Peer Review: Organized Skepticism; Research Ethics, Cross-Cultural Dimensions of; Research Funding: Ethical Aspects; Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct.

Bibliography

Angell, M., Kassirer, J.P., 1991. The ingelfinger rule revisited. New England Journal of Medicine 325, 1371–1373.

Bracey, G.W., 1987. The time has come to abolish research journals: too many are writing too much about too little. Chronicle of Higher Education 30, 44–45.

Bucholz, K., 1995. Criteria for the analysis of scientific quality. Scientometrics 32, 195–218.

Budd, J.M., Sievert, M.-E., Schultz, T., 1998. Reasons for retraction and citations to the publications. Journal of the American Medical Association 280, 296–297.

Campanario, J.M., 1996. Have referees rejected some of the most-cited papers of all times? Journal of the American Society for Information Science 47, 302–310.
CIOMS, 1991. International guidelines for ethical review of epidemiological studies. Law, Medicine & Health Care 19, 3–4 (Appendix I): 247–258.
Committee on Publication Ethics, 1998. Annual Report 1998. BMJ Publishing Group, London.
De Solla Price, J.D., 1963. Little Science, Big Science. Columbia University Press, New York.
De Solla Price, J.D., 1969. Measuring the size of science. Proceedings of the Israel Academy of Science and Humanities 4, 98–111.
Fanelli, D., 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 4, e5738.
Fang, F.C., Steen, R.G., Casadevall, A., 2012. Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Science 109, 17028–17033.
Foo, J.Y.A., Wilson, J.S., 2012. An analysis on the research ethics cases managed by the committee on publication ethics (COPE) between 1997 and 2010. Science and Engineering Ethics 18, 621–631.
Friedman, P.J., 1990. Correcting the literature following fraudulent publication. Journal of the American Medical Association 263, 1416–1419.
Garfield, E., 1972. Citation analysis as a tool in journal evaluation. Science 178, 471–479.
Garfield, E., 1979. Citation Indexing: Its Theory and Application in Science, Technology and Humanities. Wiley, New York.
Hamilton, D.P., 1990. Publishing by – and for? – the numbers. Science 250, 1331–1332.
Horrobin, D.F., 1990. The philosophical basis of peer review and the suppression of innovation. Journal of the American Medical Association 263, 1438–1441.
LaFollette, M., 1992. Stealing into Print: Fraud, Plagiarism and Misconduct in Scientific Publishing. University of California Press, Berkeley, CA.
LaFollette, M., 2000. The evolution of the "scientific misconduct" issue: an historical overview. Proceedings of the Society for Experimental Biology and Medicine 224, 211–215.
Lock, S., 1986. A Difficult Balance: Editorial Peer Review in Medicine. ISI Press, Philadelphia, PA.
Lock, S., 1993. Research misconduct: a résumé of recent events. In: Lock, S., Wells, F. (Eds.), Fraud and Misconduct in Medical Research. British Medical Journal, London, pp. 5–24.
Macfarlane, B., Cheng, M., 2008. Communism, universalism and disinterestedness: re-examining contemporary support among academics for Merton's scientific norms. Journal of Academic Ethics 6, 67–78.
Merton, R.K., 1973[1942]. The normative structure of science. In: Merton, R.K., Storer, N.W. (Eds.), The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago, IL.
Petersdorf, R.G., 1986. The pathogenesis of fraud in medical sciences. Annals of Internal Medicine 104, 252–254.
Pfeifer, M.P., Snodgrass, G.L., 1990. The continued use of retracted, invalid scientific literature. Journal of the American Medical Association 263, 1420–1423.
Polanyi, M., 1962. The republic of science: its political and economic theory. Minerva 1, 54–74.
Relyea, H.C., 1994. Silencing Science: National Security Controls and Scientific Communication. Ablex, Norwood, NJ.
Serebnick, J., 1991. Identifying unethical practices in journal publishing. Library Trends 40, 357–372.
Shankman, P., 1996. The history of Samoan sexual conduct and the Mead-Freeman controversy. American Anthropology 3, 555–567.
Sollaci, L.B., Pereira, M.G., 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of Medical Library Association 92, 364–367.
Sponholz, G., 2000. Teaching scientific integrity and research ethics. Forensic Science International 113, 511–514.
Stossel, T.P., 1985. SPEED: an essay on biomedical communication. New England Journal of Medicine 313, 123–126.
Van Noorden, R., 2011. Science publishing: the trouble with retractions. Nature 5 (478), 26–28.
Whiteley, W.P., Rennie, D., Hafner, A.W., 1994. The scientific community's response to evidence of fraudulent publication. Journal of the American Medical Association 272, 170–173.

Residential Child Care

Andrew Kendrick, School of Social Work and Social Policy, University of Strathclyde, Glasgow, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Residential child care has proved to be one of the most difficult and complex interventions in the lives of children around the globe. There are polarized views about residential care and this is reflected in ambivalence in legislation, policy, and practice. Residential child care, however, continues to have a positive role in the care of children and young people when they are unable to remain in their own families. High-quality and effective residential care should provide an environment that enhances development, promotes resilience, and builds the confidence and self-esteem of children and young people.

Introduction

Residential child care has proved to be one of the most difficult and complex interventions in the lives of children and young people around the globe. It has been used to save and protect children from the consequences of natural disaster, war, disease, and famine. Children have gained respite from abuse and neglect inflicted by their own families. However, residential care has also been used as a weapon of cultural genocide, and it has stripped the children of indigenous peoples of their language, dress, and culture. Children and young people have suffered physical, sexual, and emotional abuse at the hands of carers and other young people. It is not surprising, then, that there are polarized views about the role of residential child care in the range of care services for children and young people and this is reflected in the ambivalence present in legislation, policy, and practice.

In this context, international comparison of policy and practice in residential care is important because it can provide information on the range of problems that is addressed by the use of residential care, as well as highlighting contrasting views of how to respond to particular social challenges. I explore these complex issues in this article, drawing together research and evidence from widely different societies and cultures that can challenge deep-seated assumptions about effective practice with children to find the most effective way to intervene to protect their best interests (Francis et al., 2004).

With increasing globalization and the movement of children across international boundaries as migrants, refugees, and unaccompanied asylum seekers, learning from international comparison becomes important in guiding development in our own countries; highlighting the necessity of dealing effectively with cultural and ethnic diversity (Kendrick, 2008; del Valle and Bravo, 2013).

There are, however, significant issues in carrying out international comparison. Basic information about residential care may be lacking in developing and, indeed, developed countries; and there is certainly a lack of reliable, cross-national statistical information (Francis et al., 2004; del Valle and Bravo, 2013). There are issues of terminology and definition (Ainsworth and Thoburn, 2014). Even basic terms such as 'residential child care' or 'institutional care' are open to debate, and there will be variation in how different forms of care are categorized. A clear example of this is difference in the use of the term 'foster care' in the United States and the United Kingdom. In the United States, 'foster care' includes group homes as well as foster-family care, while in the latter, foster care refers specifically to family care settings and is contrasted with residential care in group settings. The UN Guidelines for Alternative Care for Children provides a useful definition of residential care as:

> … care provided in any nonfamily-based group setting, such as places of safety for emergency care, transit centres in emergency situations, and all other short and long-term residential care facilities, including group homes.
>
> United Nations, 2010; cited in Cantwell et al., 2012.

Residential child care cannot be considered in isolation from its wider context; and countries and localities will have particular economic, political, and cultural histories which affect the legal, policy, and administrative frameworks of child welfare (Courtney and Iwaniec, 2009; Ainsworth and Thoburn, 2014). These will affect the way in which residential care is used, its relationship with other child care services, (such as family-based care), the populations of children and young people placed in care, and specific approaches to care, therapeutic interventions, and practice. Given these issues, practitioners must be at least a little cautious when applying knowledge and practices from other countries to their own residential child care settings.

Despite this complexity and the ambiguity and ambivalence about residential care, there continues to be need for residential child care to meet the needs of particular children and young people, not least because sometimes they themselves have identified their preference for residential care (Kendrick, 2013). This does mean, however, that there is an urgent need to improve standards in residential child care in order to provide high-quality care which will provide a positive and enriching experience for children and young people.

Ambivalence in International Policy

Markedly different experiences of residential child care have resulted in a cross-national ambivalence around residential child care and its role in the continuum of child care services (Anglin and Knorth, 2004; Courtney and Iwaniec, 2009). This

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28077-X

has manifested itself in a range of ways: an unbalanced focus on the negative aspects of residential child care meaning that children can remain in dangerous family situations; the tension between residential care as a 'positive choice' or a 'last resort' and attempts to get rid of residential child care at either local or national level.

It goes without saying that children and young people should grow up in a family environment wherever possible and that this should apply in terms of the child's own family and in the context of foster care and other family-based care. This does not mean, however, that residential care can never be the most appropriate form of care to meet the best interests of the child.

Anglin and Knorth (2004) highlighted how contrasting views have been expressed in the context of two international declarations on residential care. The Malmo Declaration (1986) was drawn up by the European organization FICE (Fédération Internationale des Communautés Educatives). It highlights positive developments in residential care, focusing on new models of 'community' residential care which are treatment-oriented, interdisciplinary, and work in partnership with parents, social networks, and the community. Residential child care is seen not as a last resort but as an important, available intervention. The Malmo Declaration states that residential care should continue to provide a positive environment where children and young people have the opportunity to create their own network of dependable social relationships (Malmö Declaration, 1986, in Anglin and Knorth, 2004).

In contrast, the Stockholm Declaration of 2003, written largely from the perspective of nongovernmental organizations (NGOs) working in developing countries, starts by stating that "[t]here is indisputable evidence that institutional care has negative consequences for both individual children and for society at large" (Stockholm Declaration, 2003, in Anglin and Knorth, 2004). It calls for reduction in the use of residential care in favor of community and family settings, and suggests that residential care should only be used as a last resort. From this perspective, there is a clear suggestion that the ultimate goal would be an end to the practice of placing children in residential care.

Interestingly, the differing contexts of the authors of the two declarations led to markedly different conclusions. These contrasting views highlight the ambivalence about the role of residential child care, which is apparent in policy, practice, and legislation in different national contexts.

UN Guidelines on Alternative Care for Children and Their Implementation

A more balanced approach to the current position of residential child care in the range of options for alternative care for children away from home is in the work that has been done under the auspices of the United Nations. The starting point for this is the 1989 UN Convention on the Rights of the Child (UNCRC) which sets out the basic human rights that children everywhere have: the right to survival; to develop to the fullest; to protection from harmful influences, abuse, and exploitation; and to participate fully in family, cultural, and social life.

Placement of children in out-of-home care is primarily addressed in Article 20 of the UNCRC and it states that children who are deprived of their family environment, or who cannot be allowed to remain there, are entitled to special protection and assistance from the state. They are entitled to alternative care that "could include, *inter alia*, foster placement, kafalah of Islamic law, adoption or if necessary placement in suitable institutions for the care of children" (UNICEF, n.d.: p. 7).

In 2009, the United Nations General Assembly adopted the *Guidelines for the Alternative Care of Children* in order to provide more detailed, internationally recognized guidance and to enhance the implementation of the UNCRC (Cantwell et al., 2012). The 'Guidelines' set out desirable policy and practice approaches to support efforts to keep children in their families and, where this is not possible, to ensure that the best interests of the children are upheld in the provision of alternative care, including residential care. Cantwell et al. (2012) set out three basic levels of preventive action given in the 'Guidelines' which involve: the general population's access to basic services and social justice; targeted support to the families of children at risk; and support for the reintegration of children with their families at the earliest opportunity.

The 'Guidelines' acknowledge that all decisions concerning alternative care should minimize disruption and facilitate reintegration with family. Removal of a child from the care of the family should be seen as a measure of last resort and should be, whenever possible, temporary and of the shortest possible duration. A range of alternative care options should exist and while priority should be given to family and community-based care, placement in residential care should be made where such a setting is specifically appropriate, necessary, and constructive for the individual children concerned and in their best interest.

A number of reasons are identified that can make residential care the best option for children. Negative experiences in family settings can mean that foster care is not the best option for some children and young people. Residential care can also provide the best placement to keep groups of siblings together. It can also provide specialized care and treatment for children who have experienced trauma and abuse. In addition, some children and young people prefer residential care to family-based placements.

The 'Guidelines' outline the requirements for high-quality standards in residential child care. It should be based on 'small-group' living that promotes appropriate relationships between children and carers, taking into account the importance of attachment and supporting relationships with specific carers to children's well-being. While residential care should be seen as a temporary measure with the aim of returning children to their families, it may be required for a longer time for the small number of children who cannot return to their families and for whom foster care or family-based placement is not appropriate. There needs to be a rigorous framework for the regulation, monitoring, and inspection of residential settings. Children's rights need to be protected and these include: participation in decisions affecting them; contact with their parents, wider family, and community; complaints procedures; and the monitoring and review of outcomes. There is also a need for high-quality staff in residential care, with strategies to recruit, select, and accredit carers and provide support and training for them (Cantwell et al., 2012).

'Deinstitutionalization' and the Move Away from Residential Child Care

The Guidelines on Alternative Care of Children also comment on the issue of deinstitutionalization, and this is important in the development of positive approaches to residential child care. Cantwell et al. (2012) point out that in this context, it is important to distinguish between 'residential facilities' and 'institutions,' and it is the latter that need to be targeted through a deinstitutionalization strategy. The issue with such settings is not simply size but rather the nature of the care regime – the institutional culture that depersonalizes and isolates children and young people, and fails to address their individual and emotional needs. van Ijzendoorn et al. (2011) note the following negative characteristics of institutional care:

- Group sizes tend to be larger and the number of children per caregiver is large.
- Groups tend to be homogeneous in terms of age and disability.
- Caregivers for any single child tend to change constantly because of staff turnover, shift patterns, and rotas.
- Other adults tend to come and go in children's lives.
- Caregivers are poorly trained and their work is focused on physically caring for the children.
- Caregiving tasks tend to be carried out with a lack of warmth, sensitivity, or responsiveness.

It is institutions with such harsh regimes and negative outcomes that are the focus of the Stockholm Declaration and they are rightly the focus of the deinstitutionalization agenda.

Where 'institutions' do still exist, the 'Guidelines' stress that alternatives should be developed which will allow for their progressive elimination. This needs a national, strategic plan for deinstitutionalization and the development of a range of alternative services including small-group homes, foster care, support services for parents and extended family, and support for young people to live independently. Planning of alternative services is especially important in addressing the needs of disabled children and those under the age of 3 years.

The process of deinstitutionalization, however, is not an easy one, and there are a range of issues that need be addressed. Negoita (2010), for example, raises serious concerns about the process of deinstitutionalization in Romania and, in particular, the pace of the reform changes driven by a wider political agenda. She argues that although the initial results of the movement of children out of large institutions into the community and to smaller residential homes were considered successful, there were longer-term negative consequences. For example, for many of the children who were returned to their families, there have been continuing issues of lack of support. There have also been major concerns about abuse in foster care and the ramifications that this has had in terms of audit and bureaucracy. Frimpong-Manso (2013) found that there were similar issues in the deinstitutionalization program that was set up in Ghana. While there were promising developments in family- and community-based care, the reform program struggled to achieve its aims. Children returned to families which did not have the resources or support to care for them, and some were at risk of abuse; there was also a lack of financial and human resources to implement the goals of the reforms.

In many countries, large institutional care settings are a thing of the past – the workhouses, orphanages, and industrial schools of nineteenth- and twentieth-century Britain, for example, have long gone, although their shadow still falls on current perceptions of residential child care (Kendrick, 2008). The process of deinstitutionalization in Western Europe, North America, and Australia is linked to the general reduction in the use of residential care. Along with the negative perceptions of institutional care (a focus on poor outcomes and scandals of abuse), there has been a move toward prevention and support in the community, and family-based forms of care. The extent of the reduction and the pace of change, however, have varied a great deal (Kendrick, 2008; Courtney and Iwaniec, 2009; del Valle and Bravo, 2013).

The consequences of deinstitutionalization have perhaps been most striking in Australia where it has been suggested that the scale of the reduction of the residential sector has led to major disadvantages for children and young people (Ainsworth and Hansen, 2005). Driven by the negative impact of experiences of Aboriginal children and child migrants in residential care, along with the financial and ideological drivers that have given priority to foster and family-based care, Australian state governments reduced the use of residential care drastically. Ainsworth and Hansen (2005) highlight, however, that appropriate alternative forms of care have not been developed. They detail how homelessness programs, the juvenile justice system, and foster care have to deal with disturbed and vulnerable children and young people who would be better served in appropriate residential care.

> The dream – no more residential care – has gone disastrously wrong.
> Ainsworth and Hansen, 2005: p. 197

The Use of Residential Child Care

Already in this article, there have been indications of the range of factors that will determine the development and use of residential care in different countries. These will include the complex interplay of historical, economic, political, ideological and cultural factors (Courtney and Iwaniec, 2009; Ainsworth and Thoburn, 2014). This has led to a marked diversity in the use of residential child care across the countries of the world, and there is often rapid change.

Ainsworth and Thoburn (2014) provide an overview of the use of residential child care in 21 countries. First, there are major differences in the rate of children in care in the population. Countries that were part of the former Soviet Union have high rates of children in care (Russian Federation, 305 per 10 000 aged 0–17 years; Lithuania, 246; Ukraine, 194). A cluster of countries have rates between 50 per 10 000 and 80 per 10 000 (Australia, Germany, Ireland, Spain, Sweden, England, Scotland, and the United States). Of the countries included in the study, Japan has the lowest rate of children in care, only 17 per 10 000. There are further significant differences in the proportion of children in care who are placed in residential care. At the extremes, Australia has only 6% of children in care, in residential care, compared with Armenia,

which has over 95% in group care. (Only Australia and Ireland have less than 10% in residential placements.) England, the United States, Hungary, Scotland, Spain, and Sweden have between 10 and 30% in residential placements. France, Romania, Denmark, Italy, Poland, and the Russian Federation have between 30 and 50% in residential care. Germany, Lithuania, Ukraine, Czech Republic, and Israel have between 50 and 80% in residential care. Finally, Japan has 92% of its children in care, in residential placements (Ainsworth and Thoburn, 2014: p. 18).

The Communist states of Eastern Europe evidence the role of political ideology in the development of residential institutions after 1945, and the significant move away from institutional care following the breakup of the old Soviet empire. Economic development and industrialization can create a demand for residential child care because of the breakdown of traditional family and community structures. More recently, however, the impact of recession and financial austerity has placed families under stress and disadvantage, which may lead to abuse, neglect, and the breakup of families. This can also create a need for residential care. In China, the birth planning policy has been linked to the eightfold increase in the number of children in child welfare institutions between 1984 and 2004, although other social and economic factors have undoubtedly contributed (Apter, 2014). The role of religion and religious organizations has also had a major influence in the development of residential care in particular countries. In Ireland, however, a backlash has been seen in the almost entire eradication of residential care due to the scandals of abuse in Catholic residential establishments. Broader cultural norms can affect the balance between residential care and foster care. For example, in Japan, traditional views of the family have led to the marked predominance of residential over foster care.

Major events such as disaster and disease have had a significant impact on the development of residential care in some areas. The largest child welfare problem in many developing countries is that of children who are orphaned (Abebe, 2009). A great deal of the residential child care that exists in some countries caters to children who are in need as a consequence of disease (notably HIV/AIDS), natural disasters such as the South Asian tsunami of 2004, and war. There has been a significant involvement of NGOs in the establishment and delivery of residential provision in Africa. This is mirrored in other parts of the world, particularly in countries such as Cambodia, where political and social upheaval has left a continued legacy of orphaned, abandoned, and vulnerable children (Emond, 2010).

Models of residential care, unsurprisingly, differ widely, from large-scale institutions providing basic care and education for children to specialist, small-scale provision offering therapeutic services. There has been a general tendency for residential homes to become smaller, and, in the United Kingdom, it is not unusual to have homes for three or four children. Indeed, there has been a recent rise in the number of residential settings for individual young people. This involves individual young people being supported by a group of residential staff members and often these placements arise out of crisis situations. In 2013, there were over 100 registered homes with just one place in England (OFSTED, 2013). The significance of this development is the move away from the 'group care' element of residential care.

Although the general trend has been toward smaller residential establishments, there continue to be larger establishments that do not fall into the category of the institutions described above. The campus or village model is found in both the developed and developing world. In the United States, for example, Boys Town has villages across the country. The largest is the Village of Boys Town in Nebraska which is a municipality in its own right – with its own police and fire departments. At any given time, some 550 children and young people live in this village. The Boys Town approach is for an integrated continuum involving community support services; in-home family services; foster family services; treatment family homes; intervention and assessment services; specialized treatment group homes; and intensive residential treatment (Ivey, 2000). Another example of the village model is the SOS Children's Villages movement, with some 500 villages in 124 countries around the globe. Children are placed with permanent resident SOS mothers in 'family groups' within each village (Abebe, 2009).

Effectiveness, Quality, and Standards in Residential Child Care

There continues to be discussion and debate about the effectiveness of residential child care and the outcomes for children and young people. As has been highlighted above, residential child care covers a wide range of settings and environments across different countries. Care is needed in making broad generalizations about the effectiveness of residential care. Adopting a more nuanced approach about what works for children and young people in particular circumstances is necessary. Research has identified that young people often leave residential care and move to independence with poor outcomes (Stein and Munro, 2008). However, this should not necessarily be considered a failure of residential care, but a failure of the child welfare system as a whole. Children and young people are often placed in residential care because of the range and level of their difficulties, either in terms of their behavior or the complexity of their needs, and this has often followed extensive involvement with a range of other services. Research has found that when the complexity of needs is taken into account, residential care is just as effective as other forms of care. Recent research carried out in Cambodia, Ethiopia, India, Kenya, and Tanzania concluded that in terms of health, emotional and cognitive functioning, and physical growth, outcomes were no worse for children based in institutions than those experiencing community living (Whetten et al., 2009). Wolff and Fesseha's (2005) study of Eritrean war orphans compared the outcomes for children in four settings: home-reared children with their own mothers; reunified orphans and their host mothers; orphans in small group homes with their host mothers; and orphans in large institutions. Those placed in smaller group homes had fewer signs and symptoms of emotional distress and greater adaptive skills not only compared with those in institutions, but also compared with those who had been reunified.

Where research has focused specifically on the outcomes of residential child care, it has generally identified positive

outcomes. A review and meta-analysis of the outcomes of residential care for children and youth concluded that, on average, after a period of residential care, children improve in their psychosocial functioning. It found that behavior-modification and family-focused components achieved positive results, as did training focused on social-cognitive and social-emotional skills (Knorth et al., 2008). This is consistent with other research reviews that have identified the positive features of residential care (Clough et al., 2006) and concluded that generally children and young people do better than they were doing beforehand, following a period in residential care (Forrester, 2008).

An important issue affecting children and young people who have been in residential child care concerns their transition out of care back into the family or on to independence. In order to maintain the benefits of time in residential care, it is important that children, young people, and their families have ongoing support. Research has highlighted how crucial it is for young people to have plans in place in terms of accommodation, education or employment, finances, and support networks (Stein and Munro, 2008).

A number of factors have been identified in contributing to high-quality and effective residential child care. The quality of the leadership and management of the homes is crucial and research has highlighted the importance of the role of the manager of the residential establishment (Hicks et al., 2007). The residential home should have clear functions and objectives and the manager of the establishment needs to have a clear philosophy on how the home can meet these objectives. It is also important that the head of the care home has some autonomy in decision making to ensure stability in the running of the home. The manager should have a positive leadership style that is participative, and he or she should work closely, side by side with staff, respecting them and knowing their strengths and vulnerabilities. The manager should also have a good grasp of the specific responsibilities of management, including dealing with crisis situations. Positive leadership and management should enable a strong, supportive staff culture where staff members are agreed on how the establishment should be run and are able to offer consistency in their relationships with children and young people, particularly in the management of behavior. Just as the manager is participative, staff members should focus on the experience of children and young people and spend time with them both individually and in groups. Staff should be committed in supporting children to succeed and developing their emotional resilience and self-esteem.

Children and young people have frequently identified positive relationships with staff members as central to their experiences of residential child care, and the network of relationships is one of the distinctive features of group care. The quality of the interaction between staff members and children and young people in their care is crucial to the success of residential child care. Key features of this interaction should be the staff members' readiness to listen (to the momentous and to the mundane); their readiness to praise success and good behavior; their empathy, approachability, reliability, and persistence. Children and young people can experience such positive relationships as akin to being in their own families. Relationships between staff members and young people, then, are central to defining the residential experience as negative or positive (Kendrick, 2013).

Given the central importance of the quality of staff members in ensuring high standards in residential care and positive outcomes for children and young people, it is of little surprise that high priority should be given to the selection, training and supervision of staff in residential care. In many countries, the education and training of residential care staff has been a neglected area. Even in England, residential care workers require much lower qualifications than social workers, even though it has long been recognized that they are working with some of the most vulnerable and difficult children and young people. In many European countries, people who work in residential child care are social pedagogues; trained to a degree level in the education of children in the broadest sense, promoting well-being through wide-ranging socio-educational approaches (Cameron and Moss, 2007). In North America and some other parts of the world, residential child care falls within the child and youth care profession, which is also a discipline distinct from social work and primarily focused on the growth and development of children and young people in a holistic way, based on direct, day-to-day work with children and young people (Curry et al., 2010). In the United Kingdom, the education and training of residential child care staff is located under the broader umbrella of social work and some have argued that it has suffered because of this (Milligan, 1998).

Key to all training is the focus on relationships, direct work with children and young people in the life-space, supporting the emotional and behavioral development of children, promoting educational achievement (in its widest sense), attachment, and resilience.

Conclusion

This article has aimed to provide an overview of the issues facing residential child care in the context of its diversity and complexity around the world. I have highlighted the continuing ambiguity and ambivalence about the role of residential child care stemming from scandals of abuse and poor experiences of children and young people. However, I have also stressed that residential child care does have a positive place in the continuum of care services for children and young people when, for whatever reason, they are unable to remain in their own family. The future of residential care must be based on the research evidence of what works in achieving positive outcomes for children and young people. There is clear evidence that residential care can be effective in addressing the needs of vulnerable children and young people. The key resources are the staff members who care for children, and it is essential that they receive training and support to engage effectively with children and build positive and nurturing relationships. In this way, children and young people will experience residential care as an environment that enhances development, promotes resilience, and builds confidence and self-esteem.

See also: Adoption: Domestic, International and Global Perspectives; Child Neglect; Child Protection: International

Issues; Child Protection; Children and Families in Social Work; Indigenous Child Welfare Practice: The Canadian Case; Motherhood; Risk Assessment in Social Work; Social Work Theory; Social Work and Disasters; Social Work and Refugees.

Bibliography

Abebe, T., 2009. Orphanhood, poverty and the care dilemma: review of global policy trends. Social Work and Society 7 (1).

Ainsworth, F., Hansen, P., 2005. A dream come true – no more residential care: a corrective note. International Journal of Social Welfare 14 (3), 195–199.

Ainsworth, F., Thoburn, J., 2014. An exploration of the differential usage of residential childcare across national boundaries. International Journal of Social Welfare 23, 16–24.

Anglin, J., Knorth, E., 2004. Competing declarations on residential care for children and youth—Stockholm versus Malmö: international perspectives on rethinking residential care. Child & Youth Care Forum 33 (3), 141–149.

Apter, N.D., 2014. All in the family: new approaches to child relief in post-Mao China. Modern China 40 (1), 3–39.

Cameron, C., Moss, P., 2007. Care Work in Europe: Current Understandings and Future Directions. Routledge, London.

Cantwell, N., Davidson, J., Elsley, S., Milligan, I., Quinn, N., 2012. Moving Forward: Implementing the 'Guidelines for the Alternative Care of Children'. Centre for Excellence for Looked after Children in Scotland, Glasgow.

Clough, R., Bullock, R., Ward, A., 2006. What Works in Residential Child Care: A Review of Research Evidence and the Practical Considerations. National Children's Bureau, London.

Courtney, M.E., Iwaniec, D. (Eds.), 2009. Residential Care of Children: Comparative Perspectives. Oxford University Press, Oxford.

Curry, D., Eckles, F., Stuart, C., Qaqish, B., 2010. National child and youth care practitioner professional certification: promoting competent care for children and youth. Child Welfare 89 (2), 57–77.

Emond, R., 2010. Caring as a moral, practical and powerful endeavour: peer care in a Cambodian orphanage. British Journal of Social Work 40 (1), 63–81.

Forrester, D., 2008. Is the care system failing children? The Political Quarterly 79 (2), 206–211.

Francis, J., Kendrick, A., Poso, T., 2004. On the margin: residential child care in Scotland and Finland. European Journal of Social Work 10 (3), 337–352.

Frimpong-Manso, K., 2013. From walls to homes: child care reform and deinstitutionalisation in Ghana. International Journal of Social Welfare. http://dx.doi.org/10.1111/ijsw.12073.

Hicks, L., Gibbs, I., Weatherley, H., Byford, S., 2007. Managing Children's Homes: Developing Effective Leadership in Small Organisations. Jessica Kingsley Publishers, London.

van Ijzendoorn, M.H., Palacios, J., Sonuga-Barke, E.J.S., Gunnar, M.R., Vorria, P., McCall, L., Bakermans-Kranenburg, M.J., Dobroval-Krol, N.A., Juffer, F., 2011. I. Children in institutional care: delayed development and resilience. Monographs of the Society for Research in Child Development 76 (4), 8–30.

Ivey, J., 2000. Boys Town: The Constant Spirit. Arcadia Publishing, Mount Pleasant, SC.

Kendrick, A., 1995. The integration of child care services in Scotland. Children and Youth Services Review 17 (5/6), 619–635.

Kendrick, A. (Ed.), 2008. Residential Child Care: Prospects and Challenges. Jessica Kingsley Publishers, London.

Kendrick, A., 2013. Relations, relationships and relatedness: residential child care and the family metaphor. Child and Family Social Work 18 (1), 77–86.

Knorth, E.J., Harder, A.T., Zandberg, T., Kendrick, A., 2008. Under one roof? A review and selective meta-analysis on the outcomes of residential child and youth care. Children and Youth Services Review 30 (2), 123–140.

Milligan, I., 1998. Residential child care is not social work! Social Work Education 17, 275–285.

Negoita, M., 2010. A model in the desert: modernization, advanced liberalism, and child protection reform in postcommunist Romania. Politics & Society 38 (1), 95–117.

OFSTED, November 2013. Children's Social Care Providers and Places. OFSTED, London.

Stein, M., Munro, E.R. (Eds.), 2008. Young People's Transitions from Care to Adulthood: International Research and Practice. Jessica Kingsley Publishers, London.

UNICEF, n.d. The United Nations Convention on the Rights of the Child. UNICEF, London, UK.

del Valle, J.F., Bravo, A., 2013. Current trends, figures and challenges in out of home care: an international comparative analysis. Psychosocial Intervention 22, 251–257.

Whetten, K., Ostermann, J., Whetten, R.A., Pence, B.W., O'Donnell, K., Messer, L.C., Thielman, N.M., 2009. A comparison of the wellbeing of orphans and abandoned children ages 6–12 in institutional and community-based care settings in 5 less wealthy nations. PLoS ONE 4 (12), 8169. http://dx.doi.org/10.1371/journal.pone.0008169.

Wolff, P.H., Fesseha, G., 2005. The orphans of Eritrea: what are the choices? American Journal of Orthopsychiatry 75 (4), 47–84.

Residential Segregation in the United States

Joanna Marie Pinto-Coelho and Camille Z Charles, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by D.S. Massey, volume 19, pp. 13263–13266, © 2001, Elsevier Ltd.

Abstract

The relationship between racial residential segregation and integration since 1960 has been neither simplistic nor linear, but has actually grown more complex over time. For example, while overall segregation levels have declined since their peak in 1960, progress has slowed significantly since 1980, and the segregation of Latinos and Asians is actually increasing (Rugh and Massey, 2013). Many demographic factors besides race, including nativity, linguistic ability, and socioeconomic status, are critical to the processes that distribute individuals and groups across social and physical space and thus either facilitate or frustrate their upward mobility. It is critical to understand modern dynamics of segregation and integration as they have significant effects on prospects for reducing racial inequality.

Introduction

There is a plethora of evidence that "where we live is consequential for our life chances" (Charles, 2006: p. 4). Inequality in locational attainment, or residential segregation, is a key aspect of and a mechanism that contributes to racial inequality across countless other measures of social welfare (Taeuber and Taeuber, 1965; Massey and Denton, 1985). According to Sampson (2012), our neighborhoods are not just physical places in which social actors react and interact according to regimented social processes, but are actually powerful determinants of those macro- and microlevel social processes in and of themselves. In other words, the spatial is not second to the social; the social is second to the spatial. Therefore the racialized social structure is perhaps best represented by the distribution of racial and ethnic groups across social space, or the phenomenon of racial residential segregation.

One of the first explicit mentions of racial segregation in scholarly literature was by W.E.B. Du Bois, whose landmark work *The Philadelphia Negro* (1899) illustrated how neighborhood racial segregation in turn-of-the-century Philadelphia had critical consequences for the city's Black population. Wirth (1928) and Cox (1948) identified segregation as a primary tool in the maintenance of racial oppression. In their landmark work *American Apartheid*, Massey and Denton explicitly stated that "a racially segregated society cannot be a race-blind society; as long as U.S. cities remain segregated—indeed, hypersegregated—the United States cannot claim to have equalized opportunities for Blacks and whites" (1993: p. 148). The racial segregation of neighborhoods directly leads to the racial segregation of public schools, and schools that have a majority of minorities often suffer from lower quality facilities, less qualified teachers, fewer resources, and lower achievement on tests and grades (Coleman et al., 1966). In fact, racial gaps in educational aspirations, achievement, and attainment persist to this day (Kao and Thompson, 2003).

The relationship between segregation and integration since 1960 has been neither simplistic nor linear, but has actually grown more complex over time. Many demographic factors besides race, including nativity, linguistic ability, and socioeconomic status, are critical to the processes that distribute individuals and groups across social and physical space and thus either broaden or constrain their choices. It is critical to understand modern dynamics of segregation and integration as they have significant effects on prospects for racial inequality.

A Brief History of Racial Residential Segregation in the United States

Somewhat paradoxically, while racial residential segregation became openly entrenched in the American legal system at the turn of the twentieth century, scholarly literature on the subject and public perception of it as a social problem virtually disappeared. National discourse resumed in 1944, when Swedish academic Gunnar Myrdal published *An American Dilemma*, providing an outsider's account of this social problem in the United States. He concluded that residential segregation was widespread, deeply entrenched, and a major barrier to racial equality. Myrdal was also concerned that the phenomenon was self-perpetuating because the racial isolation it produced caused misunderstandings and misperceptions to grow between the races, which, in turn, reified both *de jure* and *de facto* rationales for strict racial separation.

Residential segregation was not truly considered a social problem by the white American political elite until the race riots of the late 1960s saw the nation's ghettos erupt in chaos and frustration. In 1967, President Lyndon Johnson formed a National Advisory Commission on Civil Disorders in order to learn more about the problem and to formulate solutions. In February of 1968, the commission released its report, warning that "our nation is moving toward two societies, one black, one white—separate and unequal" (Kerner et al., 1968).

Coming at a time of confusion, frustration, and fear, the Kerner Report's strong language and bold stance brought popular and prominent attention to residential segregation. Indeed, it was the catalyst for the formulation and passage of the Fair Housing Act later in 1968. The Home Mortgage Disclosure Act of 1975 and the Community Reinvestment Act of 1977 followed, both of which aimed at making sure potential homebuyers of color had access to loans and properties in predominantly white communities. Then, in 1988, the Fair Housing Act was amended, lengthening the

statute of limitations, providing for the payment of legal fees by prevailing plaintiffs, increasing fines for violations, and enlarging the Justice Department's role in prosecution (Massey and Denton, 1993).

Segregation and Integration over the Last 50 Years

In general, at the national level, residential segregation has declined since its peak in 1960. After a brief spike in integration, however, progress slowed significantly between 1980 and 2010. Since 1980, aggregate segregation measures for Blacks have gone down, while Asian and Latino segregation has been on the rise (Rugh and Massey, 2013). Wealthier Asians and Latinos are moving into predominantly neighborhoods, somewhat obscuring the worsening situations of their poorer coethnics (Frey and Farley, 1996). Krysan and Bader (2007) have noted that the 'chain migration' patterns typical for immigrants from Mexico and Central America have made the Latino segregation situation worse than before. Despite these two trends, Blacks always have been and continue to be more segregated than Latinos and Asians, both in the cities and in the suburbs (Massey and Denton, 1989; Charles, 2003).

More has changed in the sociodemographic landscape since the 1960s than these general statistics show. After all, it is important not to think of the American segregation story as linear, moving directly and inexorably total segregation to total integration. First, gentrification and other 'revitalization' efforts have already begun to change cities, especially inner cities, and these areas are often less segregated than others (Frey and Farley, 1996). Gentrification is a controversial endeavor, however, because, while it produces revenue for local economies and encourages multiracial communities, it also destroys existing predominantly Black and Latino neighborhoods, thus displacing those populations of color (Lee et al., 1985).

Another instance of the progressively more complex nature of this story has to do with the emergence of suburbs, or ever-larger rings of residential areas that serve as middle-grounds, even buffers, between city centers and rural areas (Massey and Denton, 1987). Patterns of racial residential segregation are different, not absent, in the suburbs, and not all suburban neighborhoods are created equal (Adelman et al., 2001; Logan et al., 2004). Many scholars have noted that Blacks of higher socioeconomic status are moving from the cities to the suburbs (Alba et al., 1999). These middle-income Blacks live with more whites in the suburbs than they would live with in cities, but they are more likely than suburban whites with the same income to have white neighbors who are less affluent than they are (Alba et al., 2000). Some segregation researchers have documented that crime is on the rise in suburban neighborhoods with higher concentrations of families of color. This leads to white flight, thus racially segregating formerly mixed suburban neighborhoods and concentrating higher rates of crime in neighborhoods in which people of color live (Liska et al., 1998). Other researchers are currently discussing the recent emergence of 'ethnoburbs,' or 'affluent, suburban, ethnic neighborhoods' outside major American cities (Wen et al., 2009).

Suburbs are also becoming the first destinations of immigrants, especially Latino immigrants (Alba et al., 1999). Race, national origin, and English language fluency exert multiple segregative forces on immigrants of color. Foreign-born Latinos and Asians are more segregated from whites than their native-born coethnics are (Iceland and Scopilliti, 2008; Kim and White, 2010). Immigrant families tend to live in neighborhoods of lower quality compared to those of their native-born peers, and their communities tend to experience higher rates of teen pregnancy, higher percentages of students underperforming in school, and higher rates of residents receiving funds from the Aid to Families with Dependent Children (AFDC) program (Rosenbaum and Friedman, 2001).

Zoning, both in the city and the suburbs, has also been a tool for maintaining and increasing racial residential segregation in the post-Civil Rights era. Rothwell and Massey (2009) contend that antidensity zoning in predominantly white areas reduces the availability of affordable, high-quality housing in those areas, thus reducing Black and Latino access to those neighborhoods. Other policies, specifically the trend in subprime mortgage lending among banks in recent years, have also hit families of color particularly hard. More specifically, banks that rated Black homebuyers as risky borrowers because they were purchasing homes in racially segregated neighborhoods penalized them with risky loans, thus putting them in greater jeopardy for astronomical debt and disproportionate rates of foreclosure (Rugh and Massey, 2010). Sharkey (2013) takes a longer view of Black poverty and segregation, describing them as being 'passed down' from one generation to the next. He asserts that it is critical to examine racial inequality in a multigenerational context in order to fully understand its scope, effects, and prospective solutions.

The Consequences of Racial Isolation

Life outcomes for whites in all-white and integrated multiracial neighborhoods are 'statistically indistinguishable' from one another (Cutler and Glaeser, 1997). But vast evidence shows that racial residential segregation has extremely negative outcomes for persons of color. Living in racially segregated and economically vulnerable neighborhoods exposes residents of color to chronic joblessness and underemployment (Santiago and Wilder, 1991; Wilson, 1997); welfare dependency (Massey et al., 1987); low-quality public education (Farley and Taeuber, 1974); more severe and exponentially higher numbers of crimes (Krivo and Peterson, 1996; Liska et al., 1998); and higher rates of asthma, heart disease, and cancer (Waitzman and Smith, 1998; Morello-Frosch and Jesdale, 2006). It is also critical to note that racial residential segregation leads to negative life outcomes, not the reverse, thus debunking cultural explanations for concentrated poverty and segregation (Cutler and Glaeser, 1997).

Explaining the Persistence of Racial Residential Segregation

The question remains as to how segregation has lingered, and sometimes even strengthened, when the practices that facilitate it are illegal and the social acceptance of it has waned. There are three main schools of thought explaining the persistence of

racial residential segregation in the United States. The psychosocial model postulates that individuals want to live near people of their same racial group (Crowder et al., 2006). Evidence from extensive attitudinal surveys suggests that members of all racial minority groups prefer integrated neighborhoods to those dominated by members of their own race. These same surveys, however, indicate that many whites prefer to live exclusively with other whites (Bobo and Zubrinsky, 1996; Charles, 2000, 2006).

The spatial assimilation or socioeconomic model argues that individual gains in education and income enable people to move to higher-quality and less-segregated neighborhoods regardless of their race (Massey and Denton, 1985; South and Crowder, 1997; Crowder et al., 2006). There is substantial evidence in support of this model to explain the residential outcomes of Latinos and Asians, but not those of Blacks; unlike Asians and Latinos, Blacks have a much harder time converting socioeconomic gains into residence in integrated neighborhoods (Massey and Denton, 1987, 1988). And, furthermore, there is significant socioeconomic variation within neighborhoods, regardless of their racial composition (Logan and Alba, 1993; South and Crowder, 1997).

The third explanation for continued racial residential segregation, the place stratification model, attributes more importance to the persistence of racial prejudice in social institutions like real estate agencies, banks, and neighborhood associations (Massey and Denton, 1993; Yinger, 1995). Racist institutional practices such as steering, redlining, and blockbusting actively maintain segregation and constrain the opportunities of Blacks and Latinos to turn their socioeconomic gains into homes in higher-quality neighborhoods (Logan and Alba, 1993). Although this approach takes into account the institutional and prejudicial forces at work in maintaining residential segregation, it has done so at the expense of individual-level factors that are necessary for a thorough and complex explanatory model.

Measuring Segregation and Its Future Study

In 1989, Massey and Denton pioneered the statistical measurement of the concept of 'hypersegregation,' or extreme racial segregation. They sought to compare and contrast the racial residential segregation of Blacks and Latinos in the United States. In order to facilitate this comparison, they identified five dimensions of hypersegregation: (1) evenness, or how uniformly members of a racial minority group are spread across a geographical area relative to whites; (2) exposure, or the degree of contact that members of a racial minority group could have with whites; (3) clustering, or the extent to which neighborhoods with high concentrations of racial minorities are adjacent to other minority-dominated neighborhoods; (4) centralization, or the degree to which minority populations live in and around the centers of urban spaces; and (5) concentration, or the amount of space in a given geographical area occupied by a racial minority group.

The index of dissimilarity, represented by Massey and Denton's measure of evenness, has been and continues to be the most commonly used indicator of segregation. Furthermore, scholars have measured residential segregation using statistical indices of segregation that could be computed for only two groups at a time, usually Blacks and whites. In fact, the vast majority of segregation literature focuses on the Black-White binary. Over the past two decades, however, scholars have begun incorporating Asians and Latinos into their analyses (e.g., Frey and Farley, 1996; Maly, 2005; Charles, 2006). Quantitative researchers have also developed new multigroup indicators of segregation (e.g., Zuberi, 2001; Reardon and Firebaugh, 2002; Iceland, 2004; Reardon and O'Sullivan, 2004), and have pioneered ways of quantifying 'neighborhoods' as units of analysis other than the census tract (e.g., Lee et al., 2008). Still others have been testing new methodologies for studying segregation, including sociodemographic mapping (e.g., Charles, 2006; Frey, 2006; Krysan and Bader, 2009; Pinto-Coelho and Zuberi, n.d.) and qualitative techniques such as ethnography and interviewing (e.g., Maly, 2005). Some scholars are turning their attention to the definition, measurement, and future of stable racial residential integration (Ellen, 2000; Pinto-Coelho and Zuberi, n.d.).

See also: Concentrated Disadvantage; Crime, Immigration, and Ethnicity; Inner City; Migration to and from Latin America; Minorities; Neighborhood Effects; Racial Relations; Segregation Indices; Social Stratification; Spatial Mismatch; Underclass; Urban Sociology.

Bibliography

Alba, R.D., Logan, J.R., Stults, B.J., 2000. How segregated are middle-class Blacks? Social Problems 47 (4), 543–558.

Alba, R.D., Logan, J.R., Stults, B.J., Marzan, G., Zhang, W., 1999. Immigrant groups in the suburbs: a reexamination of suburbanization and spatial assimilation. American Sociological Review 64 (3), 446–460.

Bobo, L., Zubrinsky, C., 1996. Attitudes on residential integration: perceived status differences, mere in-group preference, or racial prejudice? Social Forces 74 (3), 883–909.

Charles, C., 2000. Neighborhood racial-composition preferences: evidence from a multiethnic metropolis. Social Problems 47 (3), 379–407.

Charles, C.Z., 2003. The dynamics of racial residential segregation. Annual Review of Sociology 29 (1), 167–207.

Charles, C., 2006. Won't You Be My Neighbor? The Russell Sage Foundation, New York.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. U.S. Government Printing Office, Washington, DC.

Cox, O.C., 1948. Caste, Class and Race: A Study in Social Dynamics. Doubleday, New York.

Crowder, K., South, S., Chavez, E.E., 2006. Wealth, race, and inter-neighborhood migration. American Sociological Review 7 (1), 72–94.

Cutler, D.M., Glaeser, E.L., 1997. Are ghettos good or bad? The Quarterly Journal of Economics 112 (3), 827–872.

Du Bois, W.E.B., 1899. The Philadelphia Negro. Benjamin Bloom, New York.

Ellen, I.G., 2000. Sharing American's Neighborhoods: The Prospects for Stable Racial Integration. Harvard University Press, Cambridge, MA.

Farley, R., Taeuber, A., 1974. Racial segregation in the public schools. The American Journal of Sociology 79 (4), 888–905.

Frey, W.H., Farley, R., 1996. Latino, Asian, and Black segregation in U.S. metropolitan areas: are multi-ethnic metros different? Demography 33 (1), 35–50.

Frey, W., 2006. Diversity Spreads Out: Metropolitan Shifts in Hispanic, Asian, and Black Populations since 2000. Brookings Institution, Washington, DC.

Iceland, J., Scopilliti, M., 2008. Immigration and residential segregation in U.S. metropolitan areas, 1990–2000. Demography 45 (1), 79–94.

Iceland, J., 2004. The Multigroup Entropy Index (Also Known as Theil's H or the Information Theory Index). The United States Census Bureau, Washington, DC.

Kao, G., Thompson, J., 2003. Race and ethnic stratification in educational achievement and attainment. Annual Review of Sociology 29, 417–442.
Kerner, O., et al., 1968. Report of the National Advisory Commission on Civil Disorders. Bantam Books, New York.
Kim, A.H., White, M., 2010. Panethnicity, ethnic diversity, and residential segregation. American Journal of Sociology 115 (5), 1558–1596.
Krivo, L.J., Peterson, R.D., 1996. Extremely disadvantaged neighborhoods and urban crime. Social Forces 75 (2), 619–648.
Krysan, M., Bader, M., 2007. Perceiving the metropolis: seeing the city through a prism of race. Social Forces 86 (2), 699–733.
Krysan, M., Bader, M., 2009. Racial blind spots: Black-White-Latino differences in community knowledge. Social Problems 56 (4), 677–701.
Lee, B.A., Spain, D., Umberson, D., 1985. Neighborhood revitalization and racial change: the case of Washington, DC. Demography 22 (4), 581–602.
Lee, B., Firebaugh, G., Matthews, S., Reardon, S., Farrell, C., O'Sullivan, D., 2008. Beyond the census tract: patterns and determinants of racial segregation at multiple geographic scales. American Sociological Review 73 (5), 766–791.
Liska, A.E., Logan, J.R., Bellair, P.E., 1998. Race and violent crime in the suburbs. American Sociological Review 63 (1), 27–38.
Logan, J.R., Alba, R.D., 1993. Locational returns to human capital: minority access to suburban community resources. Demography 30 (2), 243–268.
Logan, J.R., Stults, B., Farley, R., 2004. Segregation of minorities in the metropolis: two decades of change. Demography 41 (1), 1–22.
Maly, M.T., 2005. Beyond Segregation: Multiracial and Multiethnic Neighborhoods in the United States. Temple University Press, Philadelphia.
Massey, D., Denton, N., 1985. Spatial assimilation as a socioeconomic outcome. American Sociological Review 50 (1), 94–106.
Massey, D., Condran, G.A., Denton, N., 1987. The effect of residential segregation on Black social and economic well-being. Social Forces 66 (1), 29–56.
Massey, D., Denton, N., 1987. Trends in the residential segregation of Blacks, Hispanics, and Asians: 1970–1980. American Sociological Review 52, 802–825.
Massey, D., Denton, N., 1988. Suburbanization and segregation in U.S. metropolitan areas. The American Journal of Sociology 94 (3), 592–626.
Massey, D., Denton, N., 1989. Hypersegregation in U.S. metropolitan areas: Black and Hispanic segregation along five dimensions. Demography 26 (3), 373–391.
Massey, D., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.
Morello-Frosch, R., Jesdale, B.M., 2006. Separate and unequal: residential segregation and estimated cancer risks associated with ambient air toxics in U.S. metropolitan areas. Environmental Health Perspectives 114 (3), 386–393.
Myrdal, G., 1944. An American Dilemma: The Negro Problem and Modern Democracy. Harper and Row, New York.
Pinto-Coelho, J.M., Zuberi, T., n.d. Segregated Diversity.
Reardon, S.F., Firebaugh, G., 2002. Measures of multigroup segregation. Sociological Methodology 32, 33–67.
Reardon, S.F., O'Sullivan, D., 2004. Measures of spatial segregation. Sociological Methodology 34, 121–162.
Rosenbaum, E., Friedman, S., 2001. Differences in the locational attainment of immigrant and native-born households with children in New York City. Demography 38 (3), 337–348.
Rothwell, J., Massey, D., 2009. The effect of density zoning on racial segregation in U.S. urban areas. Urban Affairs Review 44, 779–806.
Rugh, J., Massey, D., 2010. Racial segregation and the American foreclosure crisis. American Sociological Review 75, 629–651.
Rugh, J., Massey, D., 2013. Segregation in post-Civil Rights America: stalled integration or end of the segregated century? Du Bois Review, 1–29.
Sampson, R., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.
Santiago, A.M., Wilder, M., 1991. Residential segregation and links to minority poverty: the case of Latinos in the United States. Social Problems 38 (4), 492–515.
Sharkey, P., 2013. Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality. University of Chicago Press, Chicago.
South, S.J., Crowder, K., 1997. Escaping distressed neighborhoods: individual, community, and metropolitan influences. The American Journal of Sociology 102 (4), 1040–1084.
Taeuber, K.E., Taueber, A., 1965. Negroes in Cities: Residential Segregation and Neighborhood Change. Aldine Publishing Company, Chicago, IL.
Waitzman, N.J., Smith, K.R., 1998. Separate but lethal: the effects of economic segregation on mortality in metropolitan America. The Milbank Quarterly 76 (3), 341–373.
Wen, M., Lauderdale, D.S., Kandula, N.R., 2009. Ethnic neighborhoods in multi-ethnic America, 1990–2000: resurgent ethnicity in the ethnoburbs? Social Forces 88 (1), 425–450.
Wilson, W.J., 1997. When Work Disappears. University of Chicago Press, Chicago.
Wirth, L., 1928. The Ghetto. University of Chicago Press, Chicago.
Yinger, J., 1995. Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination. Russell Sage Foundation, New York.
Zuberi, T., 2001. The population dynamics of the changing color line. In: Anderson, E., Massey, D. (Eds.), Problem of the Century: Racial Stratification in the United States. Russell Sage Foundation, New York, pp. 145–167.

Residential Segregation: Geographic Aspects

Susan J Smith, University of Edinburgh, Edinburgh, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 13258–13263, © 2001, Elsevier Ltd.

Abstract

Residential segregation refers to the practices and ideologies by which socio-economic characteristics are expressed in, and shaped through, the organization of residential space. This article outlines the forces at work in the division of residential space and considers the problem of measuring and interpreting patterns of segregation. It explores the relationship between segregation and inequality, focusing particularly on the ways in which residential has been implicated in both the expression of racism and the articulation of resistance.

Residential segregation refers to the practices and ideologies by which socio-economic characteristics are expressed in, and shaped through, the organization of residential space.

The Division of Residential Space

Residential space is a geographically uneven resource, and the land and housing which constitute it vary in quality, character, condition, and value. Land values vary according to demand and have traditionally been highest in city centres. Housing characteristics vary according to the circumstances of housing production and use, including the macroeconomic climate, the legislative and regulatory framework, the fortunes of the construction industry, the availability of loans and subsidies, changing architectural fashions, and so on.

A primary determinant of access to housing (and neighborhood) is cost, but the markets and institutions that make up the housing system also allow a range of social characteristics to mediate the process of inclusion within, or exclusion from, different parts of the residential environment. It is this interaction of the spatial structure of the housing stock with social structures, property values and with the operation of commercial and public institutions that gives rise to the geography of residential segregation. These processes are, furthermore, embedded in, and legitimized through, the wider set of entitlements, obligations, and ideologies built into the constitutions of states and societies.

Housing and the residential space it comprises is both a commodity and a welfare resource. Housing is bought and sold in the market place, but housing policy can also be a means by which states intervene to meet social needs and alleviate social inequality. Both market forces and governmental priorities therefore play a part in structuring residential space; and such space is fractured in many ways. This fracturing occurs because of differences in 'need' and in 'ability to pay,' and it occurs too because the processes governing access to housing tend to discriminate on the basis of social characteristics. Direct and indirect discrimination has been documented in all sectors of the housing system. In the housing market, access to housing opportunities generally, and to housing finance in particular, may be dictated as much by who people are (or are presumed to be), by what they do, and by where they currently live, as by their actual ability to pay. In the public or social sector, the principle of allocating homes according to 'need' is always hard to operationalize and in practice may be compromised by, for example, subjective discretionary judgments about what types of people 'deserve' what kind of housing, and where.

As a consequence of all this, while class, income, wealth, and other economic indicators provide one set of fault lines in the residential landscape, ideas about 'race,' ethnicity, religion, and language provide another. Health inequalities are also expressed and reproduced in the residential environment; and although most research on segregation refers to the experience of 'households' (implicitly lumping men, women, and children together, and often referring mainly to the experience of men), the division of residential space is also fundamentally gendered and sexualized.

Measurement: the "Index Wars"

A strong empirical tradition underpins the literature on residential segregation, and this has become the province of geographers (*see* Social Geography) and of those whom Peach (1975) terms 'spatial sociologists.' The idea implicit in all this work is that there is some kind of link between society and space; that social processes literally 'take place'; and that where people live is both an expression of who they are and a factor affecting who they can become.

Empirical studies of residential segregation are dogged by the problems of measurement and interpretation. The problem of measurement is apparent from the large literature comprising the so-called index wars. One debate concerns the appropriate index with which to measure segregation. This has revolved around a group of indices of segregation and dissimilarity, which were favored for their simplicity and apparent (though questionable) scale independence. The index of dissimilarity (ID), for example, measures the difference in the distribution of any two population groups over a given number of areal units (e.g., census districts). ID is calculated very simply as half the sum of the difference in the proportion of each group in each areal unit. The index varies from 0 (complete similarity) to 100 (complete segregation) and the value of the index refers to the proportion of either group that would have to move from one areal unit to another in order for the two distributions to be the same.

ID is a special form of the more general index of segregation (IS) which measures the residential dissimilarity of one group from the rest of the population. Several other indices can be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72118-0

derived from this to capture the spatial segregation and isolation of one or more population groups from one another. Perhaps the best known of these is the asymmetric index of segregation, P*, which describes the spatial isolation of a social group and provides a measure of potential interaction between that group and another (Lieberson and Carter, 1982).

A second point of dispute in the measurement of residential segregation concerns whether the degree of segregation should be gauged against a norm of evenness or randomness. If populations were randomly distributed across space, some clustering would be expected as part of this random pattern. The index of dissimilarity refers to the proportion of a population that would need to move to achieve evenness, and it may therefore exaggerate the amount of segregation that is enforced (or chosen). There is still no consensus on the best way forward here. Nevertheless, all these debates are dwarfed by the problems of how to conceptualize the groups in the study, of whether to standardize (e.g., for income when measuring 'racial' segregation), and, crucially, of how to interpret the results of applying the measures.

Interpretation: the 'Choice-Constraint' Debate

A major debate over the interpretation of measures of residential segregation concerns the question of causality. It hinges around the extent to which people choose to live separately from one another, and the extent to which segregation is forced upon them. This has become known as the 'choice-constraint' debate. In the UK, for example, all the major cities display enduring patterns of residential segregation between the white, African-Caribbean and (South) Asian populations. However, within the Asian populations, Sikhs, Hindus, and Muslims tend to live in different neighborhoods to one another, while West Indians in London are segregated according to the island of origin of those who are migrants. Given these findings, it might plausibly be argued that the residential separation of Hindus and Muslims, or of Jamaicans and Trinidadians, is a matter of choice. Indeed in the USA, Clark (1992) draws on Schelling's (1971) ideas about the aggregate effects of small variations in residential preferences to argue that, in the wake of declining institutional discrimination, the accumulation of own-group residential choices is a key factor sustaining current patterns of segregation in US cities.

While most people do have some choice over the precise space they occupy in the residential environment, the fact that, for the poorest people, these choices are highly limited, leads many authors to focus more on sources of 'constraint' than on elements of 'choice.' Smith (1987, 1989), for example, has argued that while it might well be the case that intragroup segregation is the result of household preferences, it is hard to sustain the argument that a very substantial proportion of such a diverse group as, say, South Asians in Britain, would choose to live in neighborhoods that are on average cheaper, poorer, less well serviced, and more deprived than any others. The demonstrable links between segregation and inequality, the marked extent to which poverty overlaps with (or is constructed through) these spatial processes of marginalization and exclusion, suggests that households segregated into the least desirable residential locations have very little room for maneuver.

Despite a number of conceptual and methodological problems, the empirical tradition has established beyond question that residential segregation is a key feature of the residential landscape, and that, preferences notwithstanding, patterns of residential segregation tend more often than not to be linked systematically with patterns of social inequality (*see* Spatial Equity). Some have suggested that this means that residential segregation has the capacity to act as a catalyst in the process of class formation or as a springboard for collective action more generally. However, in one of the most influential theoretical accounts of the relationships between class structures and residential differentiation, David Harvey (1985) argues that residential segregation is part of the reproduction of social relations within capitalist societies. From this perspective the fracturing of populations into residential communities mainly serves to fragment the class consciousness which might otherwise be used to improve the quality of local life.

Some of these fragmenting, isolating effects of segregation are uncovered by Pratt and Hanson (1994) in a study of women workers in Worcester, Massachusetts. They show how residential neighborhoods are differentiated by the distinctive mix of 'race,' class, gender, and work traditions within them. This mix influences the locational decisions of employers, who work with stereotypes about the differences between men and women's work, and about the suitability of different ethnic groups for particular types of task. Employers' locational strategies thus reinforce a spatial division of labor which already emphasizes the gender divide. But because this division of labor also operates along class and ethnic lines, the potential for solidarity among women is undermined, despite their common interests (*see* Gender and Feminist Studies in Geography). Residential segregation is thus experienced as isolating and exclusionary rather than as a potential resource.

Segregation and Inequality

The literature on residential segregation explores the many social dimensions embedded in the division of residential space. While there is some interest in the location of elites, and in the continuing residential segregation of predominantly "white" ethnic groups (for example Poles, Jews, and Italians in US cities) despite their economic integration, perhaps the bulk of writing focuses on disadvantaged and marginalized minorities. There are therefore studies of residential segregation among people with mental health problems, the old and impoverished, disabled people, single mothers, and so on. However, the main research effort has been concerned with the impact of segregation on the lives of marginalized racialized groups in Western societies (*see* Ethnic Conflict, Geography of). This perhaps reflects the magnitude of the challenge of "otherness" that most developed countries face in a postcolonial world, where 'people of color in the historically white counties of the West are treated as second-class citizens' (Bonilla-Silva, 2000, p. 194). Failure to acknowledge this – that is, failure to recognize the intensely racialized nature of much residential segregation in the Western world – has been criticized by

Wacquant (1997) who (writing with specific reference to the USA) is concerned that by reducing the problem of racial segregation to one of poverty and disadvantage, analysts underestimate the power of history and ideology.

A growing literature on the racial dimension of residential segregation shows that segregation is not just an economic process and a material fact of life, but that it also has powerful ideological underpinnings. Historians George Fredrickson (1981) and John Cell (1982), for example, show that segregation is part of the changing character of racism. The idea of 'race' may have arisen from the notion that particular places (climates and environments) produced different kinds of people ("races"), but the tenacity of this idea is now secured by an unequal distribution of power which allows particular kinds of people to be restricted to particular types of place. Thus racisms become institutionally normalized in and through the spatial configuration of the residential environment (cf Goldberg, 1993).

So, whereas many writers once assumed that racial residential segregation is a declining legacy of a racist, colonial past, Cell and Fredrickson show how segregationism emerged during the early twentieth century as a new and powerful way to manage power relations. In the US South and South Africa, this was initially legally enforced through the Jim Crow laws and apartheid, respectively. But what Cell's analysis vividly shows is that one of the key characteristics of segregationism is its tenacity, flexibility, and ability to survive as a practice even with the declining respectability of racism and even with the end of *de jure* racial separatism. Indeed, Cell goes so far as to define segregation as one of the most successful political ideologies of the twentieth century. 'In a segregated society,' he argues, 'we expect job restrictions and unequal wages; inferior housing, education, transportation, health and social services; prejudiced courts and police forces.' And to the extent that all these forms of exclusion are enacted through the practice of residential differentiation, segregation represents what he calls 'the highest stage of white supremacy.'

Residential Segregation and Racial Inequality

In the USA, the tenacity of these discriminatory and exclusionary forms of segregation has been closely documented by Massey and Denton (1993). Their analysis of the 1990 US census identified 20 metropolitan areas in which racial segregation was so complete and so expressive of the social and financial exclusion of black Americans that they merit the label 'hypersegregated.' Massey and Denton argue for the case of the USA that 'residential segregation lies beyond the ability of any individual to change: it constrains black life chances irrespective of personal traits, individual motivations or private achievements' (1993, pp. 2–3). A startling demonstration of this is given by Polednak (1997) in his study of the relationship between segregation and health which shows, among other things, that mortality rates and infant mortality rates in segregated communities in the USA can exceed those in parts of the less developed world.

In a series of related studies, Massey and colleagues suggest that not only is segregation problematic because it creates spaces of deprivation, exclusion, and neglect, but it is problematic too because the fact of residential segregation acts to concentrate poverty during times of macroeconomic change. So, while all poor people have been affected by those structural changes in the economy which increase income inequality, the most highly segregated communities – in the USA, African Americans – far worst because of the way that residential segregation interacts with economic restructuring to concentrate urban poverty (Massey and Eggers, 1990, Massey et al., 1991).

This work has been contested on both theoretical and methodological grounds by authors who place more emphasis on 'class' than 'race' in explaining recent increases in the spatial concentration of poverty among black Americans. However, Massey and Fisher (2000) have responded with a persuasive empirically-based argument that the same socio-economic changes in a given society have radically different consequences for the concentration of poverty as segregation rises. Residential segregation, they argue, 'is the key factor differentiating the experience of racial and ethnic groups in the United States, dividing them into two basic categories: those whose poor experience a high degree of class isolation in neighborhoods characterized by unusally high concentrations of poverty, and those whose poor do not' (p. 689). Because black Americans are by far the most highly segregated of the studied populations, 'the population of poor experiencing high concentrations of poverty is overwhelmingly black' (p. 689).

This kind of finding is by no means only a feature of societies with a historical legacy of *de jure* segregationism. In Britain, with no history of domestic slavery or legally enforced racial separatism, the residential segregation of racialized minorities has been a persistent theme for more than half a century (Smith, 1989). There has, it is true, been a fierce empirical debate over whether British inner cities are as segregated as many US ghettos. This has centered on whether, for example, an ID of 60 calculated using British enumeration districts is somehow 'better' or less problematic than an ID of 80 for US census districts. However, Smith (1987) argues that while segregation persists at any of these levels, it is the form not the intensity of the phenomenon that is at issue. Whether or not ghettoes are the same thing in all places, that higher levels of segregation are associated with higher rates of poverty, unemployment, poor health, and so on, is indisputable. Indeed, Smith (1989) goes on to argue that segregationism is as effective in Britain as in the USA as an interlocking system of economic institutions, social practices, and political strategies designed by one group to keep another in its place. Residential segregation, then, is a material fact of the lives of marginalized, racialized groups in Britain, and it goes with poor housing, limited employ-
ment prospects, limited service provision, poorer health prospects, and a range of other inbuilt disadvantages,

A key finding in several studies of racial segregation is that, once established, residential segregation is not just an expression of the marginality of racially 'othered' groups, but that it is also a determinant of their future life chances and opportunities. So the iniquitous form that segregation takes not only expresses marginality, but also reinforces it.

This reinforcement takes place in a number of ways. In the USA, for example, it has been suggested that there is a spatial mismatch between those who are unemployed and the jobs they need to survive (*see* Spatial Mismatch). The argument

here is that the spatial restructuring of the economy has taken jobs and incomes away from the areas in which black Americans live (*see* Economic Restructuring: Geographic Aspects). Residential segregation can, in short, act as a constraint on labor market participation. In the UK, a similar process has been observed – there too, entrapment in spaces of deprivation can prevent already disadvantaged groups from seizing opportunities opened up by the restructuring of the economy. In the British example, moreover, this process is overlaid by what might be thought of as a 'welfare mismatch' as those most dependent on welfare find themselves excluded spatially and economically from the range of commodities, insurances, and financial services that have expanded with the 'rolling back' of the welfare state (Smith, 1989).

So residential segregation both expresses social inequalities, and materially affects the participation of spatially excluded groups in social, political, and economic life. But because of its ideological dimension (noted above) segregation is not just a material fact, it is also a political idea – and one which often bears little resemblance to the lived experience of those it refers to. Smith (1988, 1993), for example, shows how, in Britain, 'racial segregation' as a political construction has helped sustain common-sense assumptions about the natural origins of racial differentiation without tackling the problems of racism and discrimination which underpin the inequalities embedded in the structure of residential space. The importance of ideas and images (ideologies) about residential space as a factor structuring people's life chances and opportunities is further illustrated by Anderson (1991) in her work on Canadian constructions of Chinatown. She shows how that segregated space known as 'Chinatown' grew out of, and came to structure, a politically divisive system of racial discourse that justified domination over people of Chinese origin in late nineteenth-century Vancouver. In a similar vein, Wilson (1996) illuminates the power of metaphor to drive the process of uneven development between segregated segments of urban Indianapolis in the USA. His aim is to illustrate what potent political instruments metaphorical symbols can be. In the example of Indianapolis, USA, 'they injected a constellation of representations into everyday thought that rendered a form of growth as value-free and neutral that was, in fact, directly adverse to the interests of the black inner city poor' (p. 96).

Because of its ideological underpinnings and material effects, residential segregation is something that those who are marginalized and excluded by it can have little control over. However, iniquitous forms and practices of segregation can be, and are being, resisted in a number of ways. There is a policy dimension to this, of course, the relevance of which is firmly established by Bashi and Hughes (1997). In a thoughtful account of the likely implications of globalization for the character and intensity of residential segregation by race they argue that the main problems and their likely solutions are located locally (nationally) rather than globally. 'On the topic of segregation,' they note, 'the real problem is less one of sustained explanation than one of sustained action' (p. 120). This adds new significance to the work of those who argue that the best weapon with which to tackle segregation is the law (Goering, 1986, McGrew, 1997).

Other ways of illuminating and challenging the inequalities embedded in residential segregation range from urban unrest to artistic expression. The first of these is extensively documented. The second is less fully recognized, though a glance at the work of artists like photographer Ingrid Pollard, painter Jean-Michel Basquiat, or filmmaker Spike Lee soon illustrates its potency. Ingrid Pollard's exhibition 'Pastoral Interludes,' for example, locates a series of 'black' figures in the 'white' English countryside. This contrasts with the 'usual' positioning of black Britons within the segregated inner city, and challenges the taken-for-granted assumption that it is somehow normal for black people to be absent from rural areas. The other side of the coin – a stark vision of (and dissatisfaction with) the experiences of confinement within residentially segregated urban spaces – finds expression in the stick-like black figures which characterize so many of Basquiat's paintings and in the rise of rap and hip-hop which was used so effectively in Spike Lee's *Do the Right Thing* to speak about life in the segregated space of New York's Brooklyn.

Overview

To summarize, residential segregation is both an expression and a mediation of the ideologies and practices which structure social life. It is a product of the uneven development of residential space and of the selective operation of the markets and institutions that govern access to housing. Residential space is a gateway to services, educational, and employment opportunities, health, and social services. Where people live is therefore a reflection of who they are, but it is also an influence on who they can become. It has therefore been argued that tackling the inequalities embedded in residential segregation is a key priority for public policy and an urgent focus for social concern.

See also: Ethnic Conflict, Geography of; Gender and Feminist Studies in Geography; Population Geography; Residential Segregation in the United States; Segregation Indices; Social Geography; Spatial Equity; Spatial Mismatch.

Bibliography

Anderson, K., 1991. Vancouver's Chinatown: Racial Discourse in Canada 1865–1980. McGill-Queen's University Press, Montreal.

Bashi, V.I., Hughes, M.A., 1997. Globalization and residential segregation by race. Annals of the American Academy of Political and Social Science 551, 105–120.

Bonilla-Silva, E., 2000. 'This is a white country': the racial ideology of the western nations of the world system. Sociological Inquiry 70, 188–214.

Cell, J.W., 1982. The Highest Stage of White Supremacy: The Origins of Segregation in South Africa and the American South. Cambridge University Press, Cambridge, UK.

Clark, W.A.V., 1992. Residential preferences and residential choices in a multiethnic context. Demography 29, 451–466.

Fredrickson, G., 1981. White Supremacy: A Comparative Study in American and South African History. Oxford University Press, New York.

Goering, J.M., 1986. Housing Desegregation and Federal Policy. University of North Carolina Press, Chapel Hill, NC.

Goldberg, D.T., 1993. Racist Culture: Philosophy and the Politics of Meaning. Blackwell, Oxford, UK.

Harvey, D., 1985. The Urbanization of Capital: Studies in the History and Theory of Capitalist Urbanization. Johns Hopkins University Press, Baltimore.

Lieberson, S., Carter, D., 1982. A model for inferring the voluntary and involuntary causes of residential segregation. Demography 19, 511–526.

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Massey, D.S., Eggers, M.L., 1990. The ecology of inequality: minorities and the concentration of poverty. American Journal of Sociology 95, 1153–1189.

Massey, D.S., Eggers, M.L., Gross, A.B., 1991. Segregation, the concentration of poverty, and the life chances of individuals. Social Science Research 20, 397–420.

Massey, D.S., Fischer, M.J., 2000. How segregation concentrates poverty, 1970–1980. Ethnic and Racial Studies 23, 670–691.

McGrew, T., 1997. The history of residential segregation in the United States and Title VIII. The Black Scholar 27, 22–30.

Peach, C. (Ed.), 1975. Urban Social Segregation. Longman, London.

Polednak, A., 1997. Segregation, Poverty and Mortality in Urban African Americans. Oxford University Press, New York and London.

Pratt, G., Hanson, S., 1994. Geography and the construction of difference. Gender, Place and Culture 1, 5–29.

Schelling, T., 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1, 14–86.

Smith, S.J., 1987. Residential segregation: A geography of English racism? In: Jackson, P. (Ed.), Race and Racism: Essays in Social Geography. George Allen and Unwin, London, pp. 25–49.

Smith, S.J., 1988. Political interpretations of racial segregation in Britain. Environment and Planning D: Society and Space 6, 423–444.

Smith, S.J., 1989. The Politics of 'Race' and Residence: Citizenship, Segregation and White Supremacy in Britain. Polity Press, Cambridge.

Smith, S.J., 1993. Residential segregation and the politics of racialisation. In: Cross, M., Keith, M. (Eds.), Racism, the City and the State. Unwin Hyman, London, pp. 128–143.

Wacquant, L.J.D., 1997. Three pernicious premises in the study of the American ghetto. International Journal of Urban and Regional Research 21, 341–353.

Wilson, D., 1996. Metaphors, growth coalition discourses and black poverty neighborhoods in a US city. Antipode 28, 72–96.

Residential Segregation: Recent Trends

William AV Clark, University of California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Residential segregation continues to be an important topic of research within studies of the spatial structure of metropolitan areas. Although segregation has declined from the high levels of the 1960s and 1970s, considerable racial and ethnic segregation in US and European cities remains, especially, in large metropolitan areas. New methods of measuring spatial segregation are being developed, and a more nuanced approach to understanding the reasons behind continuing residential segregation is emerging. Once, discrimination was invoked as the primary explanation for continuing separation in the residential mosaic, but now most studies emphasize that the explanation is more likely to be a result of the combined effect of incomes, resources, and preferences. Varying combinations of access to resources, preferences for varying proportions of other races and ethnicities, and the urban structure underpin the continuing patterns of spatial separation. Models using agent-based formulations have expanded understanding of how segregation occurs and why it continues in an era of increasing tolerance.

Introduction

Immigration from the Middle East, North Africa, Mexico, and Central America has transformed the urban structure of cities in Europe and North America, and immigration from Asia has had similar impacts in Australia and New Zealand. An outcome of these flows is the emergence of strong ethnic groupings, an outcome that is nearly universal across large metropolitan areas in all these countries. In this context there has been renewed interest in residential segregation, in both its measurement and interpretation. Whereas two or three decades ago concern focused on how to adequately measure residential segregation and whether income and resources mattered in explaining patterns of segregation, now there are real advances in both the measurement and interpretation of segregation. That is not to say that there is universal agreement about how segregation is created or of the relative roles of income, resources, and preferences in shaping observed patterns of segregation; measurement problems also remain. Still, scholars have moved beyond the simplistic argument that discrimination alone accounts for the ethnic and racial patterns so evident in the global city. Now the focus is on building interpretive models of segregation formation and measuring the relative roles of inequality and poverty as fundamental forces underpinning patterns of segregation.

This extension of the first International Encyclopedia of Social and Behavioral Sciences (IESBS) edition entry on residential segregation (reprinted in this edition) focuses on recent trends and takes up five topics central to recent studies of segregation – (1) what are the current trends in patterns of segregation, (2) what are the implications of mixed-race and multiethnic households for the patterns of segregation, (3) how has agent-based modeling improved our understanding of segregation patterns, (4) what do new methods of measuring segregation tell us, and (5) how are changing patterns of segregation by class changing the way we think about segregation.

Measuring Segregation Patterns

How do we know if, when, and where segregation exists? That simple question has created an extensive literature on various indices of segregation and launched a debate about which index to use and which ethnic and or racial groups to include. The residential segregation entry in the first edition of this encyclopedia referred to the 'index wars' as competing groups claimed status for their approaches. Nevertheless, two indices continue to dominate the evaluation of separation in the residential fabric – the dissimilarity index and the exposure index. (Dissimilarity is the proportion of one of any two groups that would need to move in order to create a uniform distribution of the population. Exposure is the probability that a member of one group would meet or interact with a member of another group. See the Residential Segregation entry in the IESBS first edition for formal definitions.) These measures are single indices and use some form of administrative areal unit – for example, blocks, tracts, cities, or counties. Of course the size of spatial unit influences the statistical outcome: the index will be higher with smaller geographic units. In addition, different spatial units in different contexts make comparability problematic.

The dissimilarity and exposure indices both have serious weaknesses; each creates a single summary measure for the metropolitan area, city, or county being examined; and each lacks an adequate theoretical basis. For example, what is the theoretical justification for expecting a totally mixed population of any kind (the definition of the absence of segregation)? Or for an index that requires half the population of one group to move to create an even (unsegregated) distribution of the population across whatever spatial unit is selected? Even the generalized multiple group versions of the dissimilarity and exposure indices still present single measures for metropolitan and regional units. Most critically, neither dissimilarity nor exposure index captures the spatial nature of segregation. This article will later describe recently developed, more sophisticated measures of spatial segregation.

Patterns and Change in Segregation

Residential segregation refers to the spatial distribution of different population groups across neighborhoods and

communities within cities. The 2010 US Census provided support for the continuing decline in segregation levels overall (Scommegna, 2011) as well as in Black–White segregation though some authors have interpreted the new data to show little change. As expected, the changes in segregation vary by ethnic group and by metropolitan area. Overall, America's neighborhoods continue to be segregated, with Hispanic and Asian populations as segregated today as they were 30 years ago although they are much less segregated than are African-Americans (Logan and Stults, 2011).

The 2010 census data analyzed by Logan and Stults shows that the average non-Hispanic White person continues to live in a neighborhood that is quite different racially from the neighborhoods where the average Black, Hispanic, or Asian person lives. Thus, despite a substantial shift of minorities from city to suburbs and despite continued mixing in city neighborhoods, minority populations are still not relatively well integrated in metropolitan American communities. Using binary measures, the dissimilarity index and the exposure index, research by Logan and others shows that Whites still live in neighborhoods with low minority representation whereas Blacks and Hispanics live in neighborhoods with high minority representation and relatively few White neighbors. Still, Black–White segregation trends as measured by the dissimilarity index have declined from a peak in the 1960s of close to 0.82 to around 0.62 in 2010. At the same time exposure has declined much less. To some extent the interpretation is in the context of a glass half full or half empty (I use measures in the range 0 (least segregated) to 1.00 (most segregated) although they can also be reported as percentage values between 0 and 100). Do we focus on the gains or on the continuing relatively high levels of separation? Clearly, change has occurred, but that change has been slower than that expected from the initial desegregation in the 1980s and 1990s (Logan and Stults, 2011).

Is Segregation Over?

Changes in segregation have generated a spirited debate about whether or not 'the segregated century has ended' (Glaeser and Vigdor, 2012). These authors use the dissimilarity index to suggest that the national level of segregation has declined, from about 0.80 in 1980 to approximately 0.55 in 2010. They conclude that American cities are now more integrated than they have been since 1910 and that all-White neighborhoods are effectively extinct. Not only do they argue that in general segregation has declined, especially in the first decade of the twenty-first century; they also document that the separation of Blacks from individuals of other races has declined in all 85 of the largest US metropolitan areas. In the housing markets tracked by the US Census Bureau, segregation is now lower than the average level segregation in 1970.

The explanation for this significant change, they suggest, is changes in federal housing policy that have shifted from actions promoting or perpetuating segregation to actions that diminish segregation. Fair housing legislation as part of the Civil Rights legislation of the 1960s and 1970s affected the ability of minority households to buy and rent housing; additionally, the demolition of large-scale housing projects in major cities increased the long process of population shifts away from inner city 'ghetto' neighborhoods. Also documented has been a change in White attitudes and increased tolerance in society as a whole (Goldman, 2012). The combination of policy changes and attitudinal changes has gone a long way to altering the playing field in which minority households seek housing. Increases in mortgage availability, in Black incomes, and in the number of African-American professionals have influenced the housing opportunity matrix. Still as noted later, the post-2008 housing crisis may have hurt as much as helped Black households.

While the Glaeser and Vigdor (2012) report provides a persuasive case for the end of the segregated century, questions remain about how far segregation has declined and whether the report adequately measures the change in segregation. Using the New York metropolitan region as an example, Alba and Romalewski (2013) suggest that a more nuanced story can be told about segregation and its changes in the past decade. Whereas neighborhoods with mixed ethnic and racial combinations are certainly more numerous, Alba and Romalewski point out that substantially Black, Hispanic, White, and Asian neighborhoods remain in the New York metropolitan region and have changed little between 2000 and 2010. But is New York a special case, and are the measures that Glaeser and Vigdor use comparable to other analyses of segregation?

First, it does appear that large cities that attract large numbers of immigrants are likely to have somewhat higher levels of segregation. This outcome, however, is likely the result of the surge in new – especially Hispanic – immigration to Los Angeles, New York, Chicago, Boston, and other large cities during the 1990s. These new immigration concentrations have been well documented and are often strongly homogeneous clusters of people with similar ethnic backgrounds. That said, increasing evidence shows that these groups are becoming less segregated as their income and resources increase (Clark and Blue, 2004).

Second, unlike most previous studies of segregation, which used the dissimilarity index between non-Hispanic Whites and non-Hispanic Blacks, Glaeser, and Vigdor use non-Hispanic Black against *all* other races which is a measure of separation that differs from that used by Logan and Stults (2011), Farley and Frey (1994) and other analysts. To illustrate the potential difficulty with defining populations differently in the dissimilarity index, I simply note, following Alba and Romalewski (2013), that aggregating Hispanics, non-Hispanic Whites, Asians, and 'other' races, and measuring the relationship with non-Hispanic Black populations confuses the situation in regions and cities where there are large Black and Hispanic populations who are living in more integrated settings. For example in Los Angeles the dissimilarity index is 0.54 in the Glaeser–Vigdor measure while the Logan–Stults calculation is 0.65. In fact for nine of the ten largest cities the Glaeser measure is significantly lower than it is in those studies that use the standard dissimilarity measure based on non-Hispanic Whites and non-Hispanic Blacks.

The continuing debate about levels of segregation between two groups misses the most significant and continuing change in American metropolitan areas: the emergence of substantial mixed-race neighborhoods and increasing diversity more generally.

Mixed-Race, Multiethnic Households, and Diversity

The shift from a majority White population to a population that will be proportionately majority-minority will occur in the US sometime about midcentury. This shift, sometimes called the third demographic transition, of continuing new immigration, high minority fertility, and an overall decrease in the White population has far-ranging implications for measuring segregation. Overall, the US will change from a low-fertility, native-born majority population to a relatively high-fertility, racial, and ethnic immigrant, or children of immigrants, population. This change, as Lichter (2013) has documented, is occurring first and foremost in the children and youth of the US population – diversity from the bottom up. While Whites are still a majority of the older age cohorts in the United States, they are already a minority population among children and younger age cohorts.

It is immigration that is changing diversity and its patterns. Flows of immigrants have integrated and changed formerly minority and White neighborhoods to mixed neighborhoods of White, Hispanic, Asian, and Black. More Americans today, compared to their parents' generation, live in places where Whites, Hispanics, or Blacks represent a modest majority and two or three other groups each enjoy significant presence (Lee et al., 2012).

Another window on changing patterns of diversity comes from the growing number of mixed-race persons (persons reporting that they are multiethnic) and the increase in number of interrace and interethnic marriages (Ellis et al., 2007). An aggregate study of more than 300 metropolitan US areas reported that mixed-race individuals were more likely to live in integrated neighborhoods. The average neighborhood for a mixed-White/Black person is 61% White and 19% Black while the average neighborhood for a person who identifies as Black-alone is only 29% White and 54% Black. The average neighborhood for a person who identifies as White-alone is majority White and only modestly Black. These findings are consistent with a general trend to greater ethnic mixing, and Fasenfest et al. (2004) show that both Whites and Blacks are increasingly living among people of other ethnicities and not just with each other.

A central element of understanding the new patterns is to recognize that minorities, including new immigrants, translate their economic gains into residential movements away from concentrated racial or ethnic neighborhoods as immigrant groups did in previous decades. South and Crowder (1998), for example, show that both Blacks and Whites attempt to convert socioeconomic resources, such as income and education, into residence in whiter and ostensibly, higher status neighborhoods. If, as recent research has suggested, in fact multirace households are likely to be younger, have higher educational levels, and higher earnings levels, then one could hypothesize that these households are more likely to live in integrated settings and will ultimately change patterns of separation in the urban fabric.

Explaining Continuing Segregation

During the 1980s and 1990s, the primary explanation offered for continuing segregation was discriminatory behavior on the part of Whites. Indeed, some sociologists argued that the actions of Whites alone created the segregation of Blacks from Whites: "high levels of racial segregation in the contemporary United States … cannot be explained away by Black and White income differences or… by a resort to notions of personal choice … residential segregation is the direct and continuing *result* of racism" (Denton, 1996: pp. 811, 905). Similarly, a review of residential segregation patterns in a National Academy of Sciences report concluded that extreme geographic segregation is largely unrelated to economic status (and) is not explained by residential preferences (National Research Council, 2001). For much of the past three decades discrimination has been the accepted explanation for residential segregation among Black and White households with similar observable household characteristics (Dawkins, 2004).

But high and growing levels of segregation between Whites and Hispanics, between Asians and Blacks, and between Hispanics and Blacks raised the question of how discrimination was relevant for groups who were newly arrived in the United States and who were living in concentrated ethnic and racial clusters. Was such racial/ethnic clustering created by discrimination or by ethnic preferences and in-group selection? These questions are still being debated, but it is increasingly clear that ethnic preferences and socioeconomic status do play roles in creating residential separation. Several studies have shown that socioeconomic status is relevant in the shifting patterns of separation (Clark and Blue, 2004; South et al., 2005; Quillian, 2002). Research on five gateway immigrant cities suggests that segregation does decline by education as well as by income, though more for the former than for the latter (Clark and Blue, 2004). Indeed, in the inner cities of Miami and New York, segregation does not decline by income, which demonstrates that the picture is at the very least complicated and varies by region and city context.

Still, Iceland et al. (2005) come down on the side of race while conceding that class deserves increasing attention (p. 264). Along with many recent commentators, they are more nuanced in their interpretations and agree that higher status African-Americans generally live in more integrated neighborhoods. Discriminatory lending practices by banks were widely cited as a powerful force in the creation of segregated residential patterns, but such practices are not especially useful in explaining concentrations of gated Black communities in Atlanta or concentrations of wealthy Koreans and Chinese in the suburbs of Los Angeles. Altogether, research now recognizes that money does matter as we will show in the discussion of inequality later in this article.

With respect to preferences and choice – the other area of research on explanations for residential segregation – a specific study of expressed preference showed how preferences can play a powerful role in the choices that households make and in the resulting patterning in the residential fabric (Clark, 1992). These and more recent studies draw on the work of Thomas Schelling (1971), whose influential paper, *Dynamic Models of Segregation*, sought to explain why groups clustered together, whether in the lecture hall or in residential neighborhoods. Schelling argued that small differences in the 'preference' of an individual to be with others of a similar type (ethnicity, for

example) could lead to quite distinct patterns of separation in the population. The paper drew on the individual preferences tradition in economics to show how microlevel voluntary choices and economic competition can create or maintain macrolevel patterns of residential segregation along ethnic and socioeconomic dimensions.

A body of research has since shown that in general significant differences exist among the preferences of Whites, Blacks, Hispanics, and Asians (Clark, 2002, 2009; Clark and Fossett, 2008; Farley and Krysan, 2002; Fossett, 2006). The preferences of Black and White households do differ: in general, African-American households express a stronger desire than do White households to be in a mixed neighborhood. Groups other than Whites also expressed strong own race-ethnicity preferences (Clark, 2009); Hispanics and Asians routinely state preferences for neighborhoods in which they are a majority presence. How should we interpret this outcome of stated preferences? On the one hand, Charles (2000) interprets this finding as revealing not neutral ethnocentrism but negative stereotyping of 'other' races and ethnicities by Whites. On the other hand, Clark (2009) and Fossett (2006) emphasize the role of social distance and neutral ethnocentrism.

The work on agent-based modeling of segregation provides strong support for the Schelling finding that a small difference in preferences for one location or another, or one group or another, can lead to quite strong patterns of segregation and separation (Zhang, 2004; Fossett, 2006; Clark and Fossett, 2008). An important finding by Zhang (2004) shows how separation (segregation) can develop without discriminatory processes. In fact the theoretical work suggests that segregation will emerge from the behavior implied in the Schelling model, even if all parties desired integration. The conclusion that separation can exist across neighborhoods even if people prefer integrated neighborhoods is a fundamental finding because it explains continuing separation in a world of increasing tolerance and a world where opportunities are greater than in previous periods when discriminatory behavior was enshrined in policy.

The research using agent-based modeling has drawn discussions of the distributional forms of the preference function, the nature of the neighborhood, and the role of selection (Bruch and Mare, 2006; Van Rijt et al., 2009). Clearly, it is theoretically possible for the form of the preference function in agent models to influence the outcome with respect to the level of segregation. While some of the debate has been about linear versus threshold functions, it appears that evidence supports Schelling's views. Van Rijt et al. (2009) in fact argue that if anything Schelling understated the tendency to segregation.

Current discussions accept that residential segregation is not the outcome solely of income, affordability, social status, or preferences; it is the outcome of a mixture of these factors tempered by the myriad daily decisions in the rental and owner markets of American cities (Clark, 2002). As has been stated elsewhere, to assert that income does not explain residential segregation is only to assert the obvious: No one factor explains the complex residential patterns so evident in our metropolitan areas, and while discrimination may still play a role in some cases, blaming discriminatory behavior alone is simply not supportable by the research evidence.

Spatial Measures of Segregation

The sociological and geographical communities interested in the spatial structure of segregation have proposed a variety of solutions to the aspatial nature of segregation measures. Wong (2005), Getis and Ord (1992), and Matthews and Yang (2012) have tested alternative ways of bringing a formal spatial function into the measurement of segregation. Each entails innovative thinking about how to measure spatial patterning in the urban fabric. Much of the work builds on research involving a spatially weighted matrix to reflect the extent of contact between spatial units (Getis and Ord, 1992). The most sophisticated of these attempts, uses a spatially weighted matrix to create a modified dissimilarity or exposure index (Wong, 2005); but despite their improved representation of spatial structure, these spatial measures – perhaps because of their complexity – have not supplanted the general use of the simpler, aspatial dissimilarity, or exposure measures.

More recent work aims to generate measures that are truly spatial in that they represent the underlying spatial patterns of the groups being studied and can provide both indices and spatial representations of the underlying patterns. In many ways these new measures build on the ideas of earlier studies that tried to bring distance into the index function. A recent model uses distance and density to generate a mapping of the amount of separation between racial groups. The technique involves computing a measure of segregation at various scales when the scales are defined by local distance radii. Considering increasing radii of 500–4000 m and within each specified radius computing a spatial information theory index comes closer to providing a spatial measure of segregation (Reardon et al., 2008). The index is based on a distance decay function that weights nearby locations more heavily than more distant locations in computing the racial composition in each local environment. Using this tool, one can map levels of segregation, a marked advance over previous attempts at spatial measures of segregation.

A perhaps simpler and more intuitive spatial measure has been proposed that counts 'nearest neighbors' who are similar to or different from the selected respondent. In this scenario a set of differing nearest neighbors – 100, 1000, 10 000 – are examined to see how many are of the same ethnicity, same income group, or some other measurable characteristic. This formulation and the associated methodology overcome the scale problem (the dependence of the segregation index on size of administrative units) and provide measures of separation that focus on the probability of contact with like persons across a set of nearest neighbor distances. Specifically, the technique asks, what is the probability of an individual of one group meeting individuals from a different group (Östh et al., 2013)? The ability to provide a visual picture of segregation, which the Reardon et al. (2008) index also does, is a real advance in incorporating spatial pattern in segregation measures.

Race, Class, and the Changing Urban Mosaic

While ethnic and racial segregation has declined modestly over the past three decades, segregation based on family

income has increased during the same period. In 2010 approximately 28% of low-income households were located in mostly lower income census tracts, an increase from 23% in 1980 (Taylor and Fry, 2012). Over the same time period upper income households became far more likely to live in an upper income census tract (Reardon and Bischoff, 2011). This increase in income segregation is related to the long-term rise in national income inequality, including a decline in the number of middle-class or mixed-income neighborhoods.

Studies from the 2010 Census conclude that mixed-income neighborhoods have become less prevalent whereas affluent and poor neighborhoods have become more common, with the result that the residential isolation of both poor and affluent families has increased over the last four decades (Reardon and Bischoff, 2011). The proportion of families living in poor or affluent neighborhoods doubled from 15 to 31% while the proportion of families living in middle-income neighborhoods declined. The interest in income segregation arises from the argument that neighborhood context can directly affect a person's social, economic, or physical outcomes.

Observations and Conclusions

Understanding of the patterns and causes of segregation has grown significantly in the past two decades. We have a more balanced view of what drives segregation – a combination of income differences, varying resources and preferences, although we also know that private discriminatory practices still occur. Better tools have also emerged to measure the extent and patterns of segregation although the dissimilarity and exposure indices still hold considerable sway.

While the study of ethnic segregation is still important, interest has increased in understanding economic and income segregation. Much of the new concern with understanding economic segregation has revolved around the arguments that segregation, especially of the poor, leads to negative outcomes in quality of life and poorer access to high quality resources, such as good schools. This growing interest in social and economic inequality is also providing an impetus for studies of segregation. These issues continue to drive the interest in, and research on patterns and causes of residential segregation.

See also: Cities: Internal Structure; Electoral Geography; Neighborhood Effects; Race and Racism, Geography of; Residential Segregation: Geographical Perspectives; Spatial Pattern, Analysis of; Urban Geography.

Bibliography

Alba, R., Romalewski, S., 2013. The End of Segregation? Hardly. A More Nuanced View from the New York Metropolitan Region. Center for Urban Research. www.urbanresearch.org.

Bayer, P., McMillan, R., Rueben, K., 2004. What drives racial segregation? New evidence using census micro data. Journal of Urban Economics 56, 514–535.

Bruch, E., Mare, R., 2006. Neighborhood choice and neighborhood change. American Journal of Sociology 112, 667–709.

Charles, C.Z., 2000. Residential segregation in Los Angeles. In: Bobo, L., Oliver, M., Johnson, J., Valenzuela, A. (Eds.), Prismatic Metropolis: Inequality in Los Angeles. Russell Sage Foundation, New York, pp. 167–219.

Clark, W.A.V., 1992. Residential preferences and residential choices in a multi-ethnic context. Demography 30, 451–466.

Clark, W.A.V., 2002. Ethnic preferences and ethnic perceptions in multi-ethnic settings. Urban Geography 23, 237–256.

Clark, W.A.V., 2009. Changing residential preferences across income, education, and age: findings from the multi-city study of urban inequality. Urban Affairs Review 44, 334–355.

Clark, W.A.V., Blue, S., 2004. Race, class and segregation patterns in U.S. immigrant gateway cities. Urban Affairs Review 39, 667–688.

Clark, W.A.V., Fossett, M., 2008. Understanding the social context of the Schelling segregation model. Proceedings of the National Academy of Sciences 105, 4109–4114.

Dawkins, C., 2004. Recent evidence on the continuing causes of Black-White residential segregation. Journal of Urban Affairs 26, 379–400.

Denton, N., 1996. The persistence of segregation links between residential segregation and school segregation. Minnesota Law Review 80, 795–824.

Ellis, M., Holloway, S.R., Wright, R., East, M., 2007. The effects of mixed-race households on residential segregation. Urban Geography 28 (6), 554–577.

Farley, R., Frey, W., 1994. Changes in the segregation of whites from blacks during the 1980s: small steps toward a more integrated society. American Sociological Review 59, 23–45.

Farley, R., Krysan, M., 2002. The residential preferences of blacks: do they explain persistent segregation? Social Forces 80 (3), 937–980.

Farley, R., Fielding, E., Krysan, 1997. The residential preferences of black and whites: a four metropolis study. Housing Policy Debate 8, 763–800.

Fasenfest, D., Booza, J., Metzger, K., 2004. Living Together: A New Look at Racial and Ethnic Integration in Metropolitan Neighborhoods, 1990–2000. The Living Cities Series. The Brookings Institution, Washington, DC, pp. 1–18.

Fossett, M., 2006. Ethnic preferences, social distance dynamics, and residential segregation: theoretical explorations using simulation analysis. Mathematical Sociology 30, 185–274.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24, 198–206.

Glaeser, E., Vigdor, J., 2012. The End of the Segregated Century: Racial Separation in America's Neighborhoods, 1890–2010. Manhattan Institute for Policy Research. Civic report 66 January.

Goldman, S., 2012. Effects of the 2008 Obama presidential campaign on white racial prejudice. Public Opinion Quarterly 76, 663–687.

Iceland, J., Sharpe, C., Steinmetz, E., 2005. Class differences in African American residential patterns in U.S. metropolitan areas: 1990–2000. Social Science Research 34 (1), 252–266.

Ihlanfeldt, Keith R., Scafidi, Benjamin, 2001. Black self-segregation as a cause of housing segregation: evidence from the multi-city study of urban inequality. Journal of Urban Economics 51 (2), 366–390.

Lee, B., Iceland, J., Sharp, G., 2012. Racial and Ethnic Diversity Goes Local: Charting Change in American Communities over Three Decades Brown University Used US 2010 Project.

Lichter, D.T., 2013. Integration or fragmentation? Racial diversity and the American future. Demography 50, 359–391.

Logan, J.R., Stults, B., 2011. The Persistence of Segregation in the Metropolis New Findings from the 2010 Census. Census Brief Prepared for Project US 2010.

Matthews, S.A., Yang, T.C., 2012. Mapping the results of local statistics: using geographically weighted regression. Demographic Research 26 (Article 6), 151–166. http://dx.doi.org/10.4054/DemRes.2012.26.6.

National Research Council, 2001. In: Smelser, N.J., Wilson, W.J., Mitchell, F. (Eds.), America Becoming – Racial Trends and Their Consequences. National Academy Press, Washington, DC.

Östh, J., Malmberg, B., Andersson, E. Analysing segregation with individualized neighbourhoods defined by population size. In: Lloyd, C.D., Shuttleworth, I., Wong, D. (Eds.), Social-Spatial Segregation: Concepts, Processes and Outcomes. Policy Press, forthcoming.

Quillian, L., 2002. Why is black white residential segregation so persistent: evidence from three theories from migration data. Social Science Research 31, 197–229.

Reardon, S., Bischoff, K., 2011. Growth in the Residential Segregation of Families by Income, 1970–2009. US2010 Discover America in a New Century. Brown University.

Reardon, S., Matthews, S., O'Sullivan, D., Lee, B., Firebaugh, G., Farrell, C., Bischoff, K., 2008. The geographic scale of metropolitan segregation. Demography 45, 489–514.

Schelling, T., 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1, 143–186.

Scommegna, P., 2011. Least Segregated US Metros Concentrated in Fast Growing South and West. Population Reference Bureau, Washington, DC.

South, S., Crowder, K., 1998. Housing discrimination and residential mobility: impacts for blacks and whites. Population Research and Policy Review 17, 369–387.

South, S., Crowder, K., Chavez, E., 2005. Exiting and entering high poverty neighborhoods: Latinos Blacks and Anglos compared. Social Forces 84, 873–900.

Taylor, P., Fry, R., 2012. The Rise of Residential Segregation by Income. Pew Research Center, Social and Demographic Trends, Washington, DC.

Van Rijt, A., Siegel, D., Macy, M., 2009. Neighborhood chance and neighborhood change: a comment on Bruch and Mare. American Journal of Sociology 114, 1166–1180.

Wong, D.W., 2005. Formulating a general spatial segregation measure. Professional Geographer 57 (2), 285–294.

Zhang, J., 2004. Residential segregation in an all integrationist world. Journal of Economic Behavior and Organization 54 (4), 533–550.

Resilience

Christoph Steinebach, School of Applied Psychology, Zürich University of Applied Sciences, Zürich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The capacity of people to withstand obstacles in their lives consistently impresses. What distinguishes those who demonstrate 'resilience?' And, how can we promote this capacity in education and therapy? The first section of this article gives an overview of the complexity of relevant questions. There are overlaps with the theory, research, and practice of positive psychology, as well as the concepts of optimal development and health. These are discussed in the second section, where explanations of resilience are given from different perspectives and research approaches. These largely theoretical considerations refer to the basics of theory and practice of promoting resilience. Subsequently, diagnosis, prevention and intervention, and evaluation are described. The final section provides a view on future developments.

Many myths, heroic legends, and biblical stories tell of people who faced special challenges, who mastered the most significant stresses with their own power, and who emerged stronger from crises. Given this historical interest in the theme, it is all the more surprising that research in the behavioral and social sciences has only relatively recently taken up the theme of 'resilience.' Today, theories and models of resilience are discussed with respect to situations when there is a need to regain balance in the face of stress or crisis. Resilience, in fact, has two related components: (1) the act of trying to handle stress or difficulties and (2) the gain in stress management skills and personal development that may occur through this process of learning to cope with difficulties. This concept relates not only to individuals but also to groups, organizations, communities, and societies (Fletcher and Sarkar, 2013). It relates more generally to all creatures, plants, and landscapes (e.g., under the stresses of climate change) (Béné et al., 2012; McElwee, 2007), as well as to objects and materials under severe conditions.

Development of a Program

As convincing as personal or anecdotal experiences may be, 'resilience' is difficult to define scientifically. Lacking a clear description of the term, one happily turns to metaphors and references the football or the comb made of horn which regain their shape and "… who will 'bounce back' from earlier dysfunction" (Luthar, 2006: 745). Indeed, research has shown that people respond very differently to severe stress (Rutter, 2013; Werner and Smith, 2001), with some carrying on in the face of adversity, while others suffering severe disruption to their lives. The first step toward developing a deeper understanding of resilience has been to describe the resiliency of individuals, beginning with the study of resilience in young people. The focus of this research was first on individual characteristics, then on external factors, and finally on the process of dealing and coping with critical life events and circumstances (Davis et al., 2009). The questions of how individual and environmental factors directly influence the way in which a person copes with stressful events and living conditions, as well as the more complex and indirect processes by which individual competencies and strengths influence the systemic interactions between the individual and his or her environment have become increasingly important (McElwee, 2007; Ungar et al., 2013; Fletcher and Sarkar, 2013).

Questions and Observations

Issues related to resilience are diverse. Research and practice on 'resilience' must, therefore, meet a range of requirements. Kaplan (2005) argues that it is necessary to discuss whether the concept and research results have any specific meaning, and if they will endure over time. From this point of view, he critiques the breadth of the approach and the weaknesses of the measurements.

Factors: What are the risk factors and protective factors that influence the vulnerability of a person or a social system? In her early work, Emmy Werner pointed out the importance of social support for positive development despite exposure to stress (Werner and Smith, 1982; Rutter, 2006). Being part of a family might provide help and integration (e.g., social support, McElwee, 2007) and emotional support (e.g., social bonds, McElwee, 2007). Conversely, legal separation and divorce are risk factors that can be connected to particular pressures. The ways in which these protective and risk factors influence outcomes depends on the context in which they are present. For example, for children it is crucial whether the separation of the parents was preceded by long-term serious conflicts (Rutter, 2013). In addition, Rutter further shows that there are substantial interactions between genetics and environment.

Genes and environment: As current behavioral genetics shows there are diverse and fundamental interactions between the genetic makeup of a person (genes, G) and the risk factors in the environment (environment, E) of that individual. "G × E relies on studying an environmental risk factor for which there is good evidence of substantial risk, and of environmental mediation of that risk, as well as heterogeneity in outcome" (Rutter, 2006: 4). Such relationships are discussed inter alia for antisocial behavior as a function of monoamine oxidase (MAO) activity and neglect in childhood, the effects of abuse in childhood and vulnerability to depression moderated by the serotonin transporter 5-HTT gene (Rutter, 2006; Feder et al., 2009), and "the valine variant of the COMT gene in relation to the effects of early heavy use cannabis,

schizophrenia being the outcome variable" (Rutter, 2006: 4; catechol-O-methyltransferase COMT, Feder et al., 2009; see the next section). As these results show, genetic effects can be partially demonstrated only under specific environmental conditions. The effects of such G × E interactions are rather specific. Hence, there cannot be a single 'resilience trait' (Rutter, 2006: 6). Rather it has been suggested that G and E "share the same pathophysiological pathway" (Rutter, 2006: 6). Genetic variation is not alone, in and of itself, a risk or protective factor (Rutter, 2006). "Adverse or positive early life experiences lead to 'biological embedding' via gene–environment interplay, and shape brain and body; in the end they bias the individual to react in certain ways to stressors generated both externally and internally" (Karatoreos and McEwen, 2013: 341). In addition to pregnancy, the neonatal period, and the childhood, the adolescent period is considered to be particularly sensitive to 'biological embedding' by affecting the process of allostasis. Allostasis refers to the ways in which people cope with stress on a muscular, neuronal, and biochemical level, on the emotional and cognitive level, and the ways in which they use bodily and behavioral changes to adjust appropriately and to brace themselves for new burdens. Therefore, allostasis consists not only of attaining an internal equilibrium, but it also entails establishing new defaults that represent an improvement in the ability to calm down and cope with stress, allowing for greater stability than previous levels ('stability through change,' Karatoreos and McEwen, 2013: 338).

Psychophysiology and brain: Different hormonal substances, neurotransmitters, and neuropeptides all play an important role in maintaining and regaining allostasis (Wu et al., 2013; Feder et al., 2009): the hypothalamic–pituitary–adrenal (HPA) axis influences susceptibility to anxiety disorders after childhood abuse (Wu et al., 2013; Feder et al., 2009); the noradrenergic and dopaminergic systems influence the risk of serious posttraumatic disorders (Wu et al., 2013); and serotonergic systems in interaction with environmental conditions may increase the individual's vulnerability to stress and depression. In addition, the dopaminergic systems may increase the risk of reduced resilience and stress management. Neuropeptide Y (NPY) is associated with increased susceptibility to anxiety disorders after neglect and abuse in childhood (Wu et al., 2013). In addition, the brain-derived neurotrophic factor (BDNF) has been shown to interact with stress and to promote depression and anxiety (Wu et al., 2013; Feder et al., 2009). Epigenetic mechanisms also determine hardiness to stress (Feder et al., 2009; Wu et al., 2013). The "neural circuitry of fear, reward, and emotion" (Feder et al., 2009: 7–10; Wu et al., 2013) makes a decisive contribution to the development of functional stability of the biophysical system. All these findings "underscore the view that resilience is not simply the absence of maladaptive changes that occur in vulnerable individuals; rather it is mediated by a unique set of adaptive changes" (Feder et al., 2009: 6). The "adult as well as [the] developing brain has considerable structural and functional plasticity and capacity for resilience" (Karatoreos and McEwen, 2013: 340).

Environment: The assumption that risk and protective factors are of equal importance worldwide is certainly wrong. As Rutter (2013) shows, with respect to marriage and divorce, the "main clinical implication [is] the requirement to assess individual needs in relation with particular circumstances, rather than assume that all risk and protective factors have similar effects in all conditions in all people" (Rutter, 2013: 476). Meanwhile, a variety of environmental characteristics that promote resilience have been identified. But what distinguishes an environment that supports resilience from one that does not? Certain elements are of special importance and many are context dependent. For example, in the case of an earthquake, social support and integration, cooperation and networking of services, structure and leadership of the community, and the commitment to the community are of particular importance (Norris et al., 2008). Crucially, the resources themselves must be resilient. They must be robustly designed (robustness), established, and managed, so that they do not collapse when problems arise. The strategies and resources must be 'redundant,' that is they must be available, and, if they become unavailable, backup substitute resources should keep the process of adaptation in motion. Additionally, the system must be built to be able to identify problems and activate more resources when it is threatened (resourcefulness). At a more individual level, the goals that one is striving for must be attainable within a reasonable time frame (rapidity) (Norris et al., 2008: 134, with reference to Bruneau et al., 2003).

Crises: Which challenges and stressors, crises or traumas are considered so serious that they require resilience? "Problems in life are inevitable. From daily hassles to major traumatic events, considerable research effort has focused on the potential short- and long term negative consequences for mental and physical health" (Davis et al., 2009: 1637). Type, quality, and timing of stress may be crucial. These receive their real meaning only when they are measured relative to the coping resources of the individuals in their environments. What is a great burden for one person may hardly pose a challenge to another, because they have enough personal resources or can rely on the support in their environment (Steinebach, 2012). Often, smaller everyday challenges offer the chance to steel oneself for times of trouble (Rutter, 2013).

Lifemanagement: What are the critical signs for coping and growth? "[G]ood coping, opportunities for growth, beating the odds, and inoculation against future stress" (McElwee, 2007: 59) are related to resilience. Aggression, withdrawal, and excessive control are seen as 'unhealthy resilience' (McElwee, 2007: 59). Even when 'healthy' and unhealthy ('catastrophe') reactions are well differentiated, the full process of coping must be observed. It may be that after a period of hapless attempts to cope, the goal is finally achieved. This process of learning and struggling through difficulties can result in an increase in overall personal resources and further positive development. The assessment of whether resilience is present requires an evaluation of the coping process over time (Steinebach, 2012; e.g., for the detection of a 'steeling' effect, Rutter, 2006: 2).

Concepts and Definitions

The roots of the word 'resilience' are the Latin words 'salire,' to jump, and 're,' back (Smith et al., 2013: 167). Correspondingly, 'resilience' at its root means that an initial state is regained: subjects regain their earlier form, the equilibrium in systems or rebalances in humans, and stress and trauma are replaced by feelings of well-being. Increasingly, however, several other

aspects of resilience are discussed: (1) Resilience is reflected in the interplay of the complex structure of interrelated environmental factors (e.g., Ungar et al., 2013) whose significance at first glance might be seen as quite neutral or even 'risky' (Rutter, 2013: 477). (2) Resilience is understood not only as a possible response to crises, but also as a positive response to the challenges of everyday life (Masten, 2001; 'chronic adversity versus resilience following single-incident traumas,' Bonanno and Diminich, 2013: 378; 'inoculation and coping,' Rutter, 2013: 477). (3) Resilience suggests positive development beyond mere readjustment (e.g., Davis et al., 2009). Accordingly, there "are several existing definitions of resilience that share a number of features all relating to human strengths, some type of disruption and growth, adaptive coping, and positive outcomes following exposure to adversity" (Prince-Embury, 2013: 9). (4) Resilience is also considered as a dynamic feature of groups, organizations, and larger communities (e.g., Zautra, 2009). (5) Resilience can be observed over the whole life span (e.g., Ong et al., 2009). (6) Finally, it is noted that in all these processes of individuals dealing with stress, cultural differences are taken into account (e.g., Zautra, 2009), as well as the biophysical processes on the part of the individual (e.g., Rutter, 2013; Wu et al., 2013).

'Resilience' is, therefore, seen overall as the positive adaptation and sustainable development of a system to respond to short- or long-term everyday challenges or severe stress. Based on internal system processes and through dealing with the environment, the system defines new reference values and develops required competencies, and the ability to cope with future stresses improves (Ong et al., 2009; Bonanno and Diminich, 2013; 'transformative capacity,' Béné et al., 2012: 21; 'new normal' and 'adaptive trajectory' vs 'growth,' Norris et al., 2008: 132; 'ecological resilience,' Norris et al., 2008: 130).

Boundaries and Interfaces

The following three subsections briefly outline the conceptual relationships of 'resilience' to 'positive psychology' and to theories and models of 'optimal development' and 'health.' Additionally, differences between these concepts also become apparent.

Positive Psychology

A particular concern of positive psychology is overcoming the negative perspective in psychological research, theory, and practice. Similar proposals were initially formulated in the early 1950s. Positive psychology and resilience therefore have common roots (see McElwee, 2007). Since the beginning of this millennium, 'positive psychology' is understood as 'the scientific study of what goes right in life' (Peterson, 2009: 3). Thus, topics of interest include character strengths, happiness, courage, mindfulness, humor, confidence, and optimism. In addition to such individual character attributes, strengths of families, groups and teams, organizations, communities, and societies are examined (Brendtro and Steinebach, 2012; Steinebach, 2012). The strengths addressed in the theories and models of resilience are seen as protective factors that are features of the person or their environment (e.g., Smith et al., 2013). On the other hand, these strengths are considered criteria for successful adaptation and coping with stress, crises, and traumas. Accordingly "the study of resilience … [has] overturned many negative assumptions in deficit-focused models" (Goldstein and Brooks, 2013: 6). However, following Luthar (2006), we must point out several significant differences between positive psychology and 'resilience': (1) In addition to the positive factors of behavior and experience, resilience takes into account difficulties and risk factors. Resilience is evident only in dealing with stress. (2) The sustainable development of resistance is revealed only over the course of an individual's development. For this reason, development over the whole life span must be observed. (3) Resilience is developed, implemented, and promoted in a complex interchange of environmental, cognitive-emotional, and biophysiological processes. The biological basis of this process has, to date, been insufficiently accounted for in positive psychology. (4) Positive psychology does not examine negative aspects of development; however, the concept of resilience does address these aspects (Ungar and Lerner, 2008). Increasingly, it becomes clear that while wishing to emphasize strengths, as demanded by positive psychology, negative aspects of individual characteristics and the environment must also be taken into account. When integrating strengths into such an approach, as promoted by positive psychology, negative aspects of the individual and their environment must also be considered to reflect the complexity of human experience and behavior. This is particularly true because after an event, whether a personality trait or an environmental factor was positive or negative and whether the resilient coping with stress will endure are only revealed over time. Finally, we cannot assume that gains in resilience will be lifelong (Rutter, 2006).

Optimal Development

In struggling with tasks, challenges, stresses, and crises, the strengths and weaknesses of individuals and their environments become clear. This raises the question of how individuals can become competent to manage their development independently and 'optimally,' given all the challenges. Starting from the cybernetic theory of adaptive self-optimizing systems (Steinebach, 2000), flexible goal attainment by individuals and improvement of their own resources prove to be an important guarantee for optimal development. These elements are also an expression of sustainable resilience, which not only restores the original state, but also promotes successful further development. An adaptive self-optimizing system has several elements, which in their interplay are also able to define new reference values. Corresponding ideas and processes are considered in different models: for example, the model of selection, optimization, and compensation (Baltes and Baltes, 1990); the two-process model of assimilation and accommodation (Brandtstädter, 2009); or in the life course theory of control (collectively, Lang and Scheel, 2012). "We may ask ourselves how we can sustain a good life under adversity, what aims in life we should fight hardest to preserve, which voice within do we listen most, and how do we set our priorities for the future in the face of limited resources" (Zautra, 2009: 1938). We assume that comparable processes also occur on biochemical, neural–muscular, emotional, and cognitive levels and

contribute in terms of semiautonomous system processes to a successful interplay between person and the environment; for example, in dealing with chronic stress to maintain allostasis (see Section Questions and Observations). In order to appropriately map the various environmental aspects, the systematization of Bronfenbrenner (1989) is particularly useful. He distinguishes between the micro-, meso-, exo-, and macro- and the chrono-systems. He describes ecological transitions as the central driving force for development and calls for research, which has ecological validity and that takes into account this diversity of systems. As the work of Ungar shows (Ungar et al., 2013), this model allows for a multi-systemic understanding of resilience, which reflects the heterogeneity of the environment of the individual. Furthermore, the cybernetic theory of adaptive self-optimizing systems points out the importance of having competencies in the accurate perception of external and internal processes, the importance of differentiated self- and environmental models, and a sensitive perception of an individual's own needs and motives. This theory also points out the importance of effective executive functions (Hughes, 2011).

Health

Health is affected not only by biological factors but also by emotional and cognitive processes. Several elements are crucial for facilitating positive health behaviors, including attitudes and a supportive environment (e.g., the family and the workplace), which includes culture and society at large. Therefore, both health and resilience are seen as proof of a successful interplay of personal and environmental conditions (Ungar and Lerner, 2008). The Ottawa Charter of 1986 strongly shaped the current model of health promotion. In focusing on health instead of disease and the contribution of the individual and the society for the preservation and promotion of health it rejects the biomedical model. If health is more than the absence of diseases, the focus falls on general well-being (World Health Organization, 1986, 2009). As Antonovsky (1996) asked which factors make health and well-being possible, despite both everyday and occasional extreme stress. Accordingly, the Ottawa Charter defined health promotion for individuals as the "… process to increase control over, and to improve, their health" (World Health Organization, 2009: 1). In this perspective, health promotion, according to the Bangkok Charter, provides "… a positive and inclusive concept of health as a determinant of the quality of life and encompassing mental and spiritual well-being" (World Health Organization, 2009: 24). Well-being as a central component of health does not mean that environmental aspects are overlooked. Health promotion targets both humans and environments. The environment is understood as a system of various components, from the interpersonal level to politics, society, and culture. Communities as life-spaces are seen as especially relevant.

Motivation is seen as crucial factor for healthy behavior. When are people motivated to take an active role in the direction of their lives and to take care of their own health? Following Ryan and Deci (2000), we can assume that three needs are guiding individuals: competence, autonomy, and relatedness. If an action serves these needs, people become active, committed, and feel comfortable. The experience of belonging (e.g., to a group) is an essential component of identity. The autonomy from requirements of others allows people to react according to their own interests, expertise, and experience which ensures self-confidence.

Successful self-control has a positive effect on well-being (Bandura, 1998; Sternberg and Spear-Swerling, 1998). This is also reflected at the level of psycho-physiological processes. To the extent to which positive affect is associated with health, those positive feelings are considered as protective factors (Steptoe et al., 2009). Thus, we can find biological correlates of positive affect that are of great importance for health and life satisfaction.

Principles and Practice of Promoting Resilience

The definitions of 'resilience' highlight the important elements of the concept: confrontation with adversity and successful management of both the self and/or the environment. Decisions about proper diagnosis, the appropriate practice of preventive services, and intervention and evaluation design are strongly influenced by the underlying theory defining the preferred model.

Assessment

Assessment in resilience promotion requires the differentiated monitoring and evaluation of the strengths of individuals, their environments, and the existing interactions between person and environment. The question of objective, reliable, and valid assessment of resources ('assessment challenge,' Prince-Embury, 2013: 13) is crucial. The focus on resource assessment aims to supplement the deficit-related diagnosis and helps to identify target groups for prevention and intervention. However, for many problems there are no models with high explanatory value and reliable assessment methods (Goldstein and Brooks, 2013). Differentiated profiles of strengths and risks from individuals to culture would facilitate designing interventions (Rutter, 2013). Finally, only repeated assessment enables the reliable monitoring of gradients of improvements and effects of interventions. To evaluate interactions between individuals and their environment as elements of the assessment process, personal traits and features of the environment as well as the competency of individuals in shaping their own interactions with the environment are assessed (Petermann and Lohbeck, 2013). In the multimodal resources assessment questionnaire, interviews, checklists, and behavioral observations are used. Individualized diagnosis may be necessary, which has, to a large extent, a situation-specific implementation and which might be supplemented by physiological parameters (Zautra, 2009; possibly supplemented by 'ecological momentary assessment,' Steptoe et al., 2009: 1748).

Prevention and Intervention

Assessment is a good starting point for prevention and intervention, especially when also trying to evaluate causal relationships. The assessment process also takes into account those conditions that are important for the implementation of the intervention (Rutter, 2013). What distinguishes effective

interventions? Luthar (2006) identifies the following features of effective interventions: (1) interventions are likely to be effective if they are aimed at a target group with a high-risk potential; (2) they should be aimed at aspects of behavior and experiences that are relatively easy to change; (3) interventions should be geographically close to the targeted people, easily accessible, and reliable; (4) interventions should not be too short-term, rather they should last longer in order to support people in everyday life; (5) interventions should support the development of new resources and competencies, and the activation of available strengths. In addition, the conceptual broadness of 'resilience' seems to recommend an approach that (6) takes into account both the strengths and weaknesses of the individuals and their environments. Such offers are aimed at (7) changes at several levels (e.g., physical, emotional, and cognitive) and in several environmental areas (e.g., family, school, and workplace, for instance, Ungar et al., 2013; 'Multidimensional Treatment Foster Care,' Leve et al., 2009: 1869) in order to ensure sustainability. In addition to this interventions should be offered early. They both reduce risk factors, and expand and stabilize strengths. For example, to help build the 'Five Cs' (competence, confidence, connection, character, and caring for; Lerner et al., 2005: 12), which are the core concepts of positive youth psychology. "… [S]uch programs are most likely to result in the development of these Cs when they involve sustained positive adult–youth relationships, youth skill-building activities, and opportunities for youth participation in and leadership of community-based activities" (Lerner et al., 2005: 12).

Evaluation

The evaluation of measures to promote resilience is not easy. Rigorous evaluation requires differentiated, mostly quasi-experimental designs, with large sample sizes. The factors and processes to be considered are highly complex and dynamic (Ungar et al., 2013). If interventions are to be ethically and professionally appropriate, an evidence-based practice is essential. Evaluation is not only useful for research and theory development, but it also serves to help in the optimization of practice. Promotion of resilience in this sense is understood as evidence-based practice. The evaluation of practice then provides important insights for the improvement of theory and how to transfer this same theory into practice (e.g., Leve et al., 2009). Due to the complex nature of the interventions, evaluation is an interdisciplinary concern. With the variety of results from disciplines such as medicine, psychology, educational science, and sociology, the task is to continually integrate the most important evaluation and research results. To date, we summarize, based on the work of Rutter (2006), the main findings of resilience research: (1) resilience can develop in dealing with challenging environmental conditions – even more than when risks are avoided; (2) protective factors can prove to be neutral or risky in the absence of threat or stress; (3) protection can also arise from what people do to deal with stressful situations; (4) protection may arise from conditions that emerge long after the exposure to stress. Therefore, longitudinal multifocal and multi-methodological evaluations are required. Short-, medium-, and long-term effects should be considered in different situations and life stages (e.g., Leve et al., 2009).

Outlook

We defined 'resilience' as the positive adjustment and sustainable development of a system to cope with short-term or longer lasting everyday challenges or severe stress. Resilience is understood as the result of internal system processes in dealing with the environments of the system. Resilience can lead to adaptation or to the development of new reference values and skills to facilitate stress management, which helps to face upcoming loads. Thus, resilience is understood as a systemic process of development, and as an expression of a highly complex systemic action. In this complex subject, those interventions that are effective act at different levels at the same time, consider different methods to address the person in terms of a holistic empowerment and promote positive environmental aspects.

In the future (1) it will be important to develop more interdisciplinary, participatory, and sustainable models of practice. In these models, the focus is on the development of skills of self-control (e.g., 'awareness' and 'mindfulness,' Methods of Cognitive–Behavioral Therapy) as well as on environmental measures (promotion of cohesion in the community, commitment, and transparency in planning processes).

Applied research faces the challenge of appropriately evaluating the different interventions to ensure the comparability of research processes and methods despite the diversity of practice. (2) Substantial insights can be gained if – in addition to large-scale evaluations – the scientific evaluation of smaller projects becomes a matter of course. In addition, it will be important to connect processes and changes on neurophysiological, emotional, and cognitive levels with changes in environmental systems.

Resilience is about learning and development in everyday life. In interaction with the environment, skills and resources are developed. The person is an active actor, as is the environment. The environment provides resources that offer the individual support and assistance. The ability to secure resources in the appropriate environment is an important competence. In addition, all people have a duty to support the resilience of others. (3) We are asked to provide resources and develop our own as competencies for our own and for the sake of others. Cultures may vary in their modes of nurturing resilience. The promotion of resources to cope in everyday life with all its hardships and crises is certainly a strength with benefits for everyone in every culture and society.

See also: Coping with Job Stress; Environmental Stress and Health; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Behaviors; Health Education and Health Promotion; Health Interventions, Community-based; Health Promotion in Schools; Health Psychology in Childhood; Health Psychology; Health: Definitions; Human Development and Health; Occupational Health; Self-Efficacy and Health; Social Integration, Social Networks, and Health.

Bibliography

Antonovsky, A., 1996. The salutogenic model as a theory to guide health promotion. Health Promotion International 11 (1), 11–18.

Bandura, A., 1998. Health promotion from the perspective of social cognitive theory. Psychology & Health 13 (4), 623–649.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, pp. 1–34.

Béné, C., Wood, R.G., Newsham, A., Davies, M., 2012. Resilience: New Utopia or New Tyranny? Reflection about the Potentials and Limits of the Concept of Resilience in Relation to Vulnerability Reduction Programs. IDS Working Paper, Volume 2012, Number 405. CSP Working Paper Number 006. Institute of Development Studies, Brighton.

Bonanno, G.A., Diminich, E.D., 2013. Annual research review: positive adjustment to adversity – trajectories of minimal–impact resilience and emergent resilience. Journal of Child Psychology and Psychiatry 54 (4), 378–401.

Brandtstädter, J., 2009. Goal pursuit and goal adjustment: self-regulation and intentional self-development in changing developmental contexts. Advances in Life Course Research 14, 52–62.

Brendtro, L.K., Steinebach, C., 2012. Positive Psychologie für die Praxis. In: Steinebach, C., Jungo, D., Zihlmann, R. (Eds.), Positive Psychologie in der Praxis. Anwendung in Psychotherapie, Beratung und Coaching. Beltz, Weinheim, pp. 18–26.

Bronfenbrenner, U., 1989. Ecological systems theory. Annals of Child Development 6, 187–249.

Bruneau, M., Chang, S., Eguchi, R., Lee, G., O'Rourke, T., Reinhorn, A., Shinozuka, M., Tierney, K., Wallace, W., von Winterfelt, D., 2003. A framework to quantitatively assess and enhance the seismic resilience of communities. EERI Spectra Journal 19 (4), 733–752.

Davis, M.C., Luecken, L., Lemery-Chalfant, K., 2009. Resilience in common life: introduction to the special issue. Journal of Personality 77 (6), 1637–1644.

Fletcher, D., Sarkar, M., 2013. Psychological resilience. A review and critique of definitions, concepts, and theory. European Psychologist 18 (1), 12–23.

Feder, A., Nestler, E.J., Charney, D.S., 2009. Psychobiology and molecular genetics of resilience. Nature Reviews Neuroscience 10 (6), 446–457.

Goldstein, S., Brooks, R.B., 2006. Preface. In: Goldstein, S., Brooks, R.B. (Eds.), Resilience in Children. Springer, New York, pp. xiii–xixv.

Goldstein, S., Brooks, R.B., 2013. Why study resilience? In: Goldstein, S., Brooks, R.B. (Eds.), Handbook of Resilience in Children. Springer, New York, pp. 3–14.

Hughes, C., 2011. Changes and challenges in 20 years of research into the development of executive functions. Infant and Child Development 20, 251–271.

Kaplan, H.B., 2005. Understanding the concept of resilience. In: Goldstein, S., Brooks, R.B. (Eds.), Handbook of Resilience in Children. Springer, New York, pp. 39–47.

Karatoreos, I.N., McEwen, B.S., 2013. Annual research review: the neurobiology and physiology of resilience and adaptation across the life course. Journal of Child Psychology and Psychiatry 54 (4), 337–347.

Lang, F., Scheel, J., 2012. Modelle der erfolgreichen Entwicklung. In: Steinebach, C., Jungo, D., Zihlmann, R. (Eds.), Positive Psychologie in der Praxis. Anwendung in Psychotherapie, Beratung und Coaching. Beltz, Weinheim, pp. 27–34.

Leipold, B., Greve, W., 2009. Resilience. A conceptual bridge between coping and development. European Psychologist 14 (1), 40–50.

Lerner, R.M., Almerigi, J.B., Theokas, C., Lerner, J.V., 2005. Positive youth development. A view of the issues. Journal of Early Adolescence 25 (1), 10–16.

Leve, L.D., Fisher, P.A., Chamberlain, P., 2009. Multidimensional treatment foster care as a preventive intervention to promote resiliency among youth in the child welfare system. Journal of Personality 77 (6), 1869–1902.

Luthar, S.S., 2006. Resilience in development: a synthesis of research across five decades. In: Cicchetti, D., Cohen, D.J. (Eds.), Developmental Psychopathology, Risk, Disorder, and Adaptation, vol. III. Wiley, Hoboken, NJ, pp. 739–795.

Masten, A.S., 2001. Ordinary magic. Resilience processes in development. American Psychologist 56 (3), 227–238.

McElwee, N., 2007. A focus on the personal and structural: resilience explored. Child & Youth Services 29 (1/2), 57–59.

Norris, F.H., Stevens, S.P., Pfefferbaum, B., Wyche, K.F., Pfefferbaum, R.L., 2008. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. American Journal of Community Psychology 41, 127–150.

Ong, A.D., Bergeman, C.S., Boker, S.M., 2009. Resilience comes of age: defining features in later adulthood. Journal of Personality 77 (6), 1777–1804.

Petermann, F., Lohbeck, A., 2013. Diagnostik: Stärken sehen lernen. In: Steinebach, C., Gharabaghi, K. (Eds.), Resilienzförderung im Jugendalter. Springer, Heidelberg, pp. 33–49.

Peterson, C., 2009. Positive psychology. Reclaiming Children and Youth 18, 3–7.

Prince-Embury, S., 2013. Translating resilience theory for assessment and application with children, adolescents, and adults: conceptual issues. In: Prince-Embury, S., Saklofske, D.H. (Eds.), Resilience in Children, Adolescents, and Adults: Translating Research into Practice. Springer, New York, pp. 9–15.

Prince-Embury, S., Saklofske, D.H., 2013. Translating resilience theory for application: introduction. In: Prince-Embury, S., Saklofske, D.H. (Eds.), Resilience in Children, Adolescents, and Adults: Translating Research into Practice. Springer, New York, pp. 3–7.

Rutter, M., 2006. Implications of resilience concepts for scientific understanding. Annals of the New York Academy of Sciences, 1–12.

Rutter, M., 2013. Annual research review: resilience – clinical implications. Journal of Child Psychology and Psychiatry 54 (4), 474–487.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Smith, B.W., Epstein, E.M., Ortiz, J.A., Christopher, P.J., Tooley, E.M., 2013. The foundations of resilience: what are the critical resources for bouncing back from stress? In: Prince-Embury, S., Saklofske, D.H. (Eds.), Resilience in Children, Adolescents, and Adults: Translating Research into Practice. Springer, New York, pp. 167–187.

Steinebach, C., 2000. Entwicklungspsychologie. Klett-Cotta, Stuttgart.

Steinebach, C., 2012. Resilienz. In: Steinebach, C., Jungo, D., Zihlmann, R. (Eds.), Positive Psychologie in der Praxis. Anwendung in Psychotherapie, Beratung und Coaching. Beltz, Weinheim, pp. 95–101.

Steptoe, A., Dockray, S., Wardle, J., 2009. Positive affect and psychobiological processes relevant to health. Journal of Personality 77 (6), 1747–1775.

Sternberg, R.J., Spear-Swerling, L., 1998. Personal navigation. In: Ferrari, M., Sternberg, R.J. (Eds.), Self-awareness: Its Nature and Development. Guilford Press, New York, pp. 219–245.

Ungar, M., Ghazinour, M., Richter, J., 2013. Annual Research Review: what is resilience within the social ecology of human development? Journal of Child Psychology and Psychiatry 54 (4), 348–366.

Ungar, M., Lerner, R.M., 2008. Introduction to a special issue of research in human development: resilience and positive development across the life span: a view of the issues. Research in Human Development 5 (3), 135–138.

Werner, E.E., Smith, R.S., 1982. Vulnerable but Invincible: A Study of Resilient Children. McGraw-Hill, New York.

Werner, E.E., Smith, R.S., 2001. Journeys from Childhood to Mid-life: Risk, Resilience and Recovery. Cornell University Press, Ithaca, NY.

World Health Organization, 1986. Ottawa-Charta. http://www.euro.who.int/__data/assets/pdf_file/0004/129532/Ottawa_Charter.pdf (accessed 15.09.13.).

World Health Organization, 2009. Milestones in Health Promotion: Statements from Global Conferences. World Health Organization, Geneva.

Wu, G., Feder, A., Cohen, H., Kim, J.J., Calderon, S., Charney, D.S., Mathé, A.A., 2013. Understanding resilience. Frontiers in Behavioral Neuroscience 7, 1–15.

Zautra, A.J., 2009. Resilience: one part recovery, two parts sustainability. Journal of Personality 77 (6), 1935–2019.

Relevant Websites

http://www.apa.org/helpcenter/road-resilience.aspx – American Psychological Association.

http://www.resalliance.org/ – ResilienceAlliance. Research on resilience in social-ecological systems - a basis for sustainability.

http://www.resilience.org/ – Resilience.

http://www.resilienceproject.org/ – The Resilience Research Centre.

http://www.stockholmresilience.org/ – Stockholm Resilience Centre.

Resilient City

Danny MacKinnon, Newcastle University, Newcastle Upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a review of the idea of the resilient city, which has become influential in recent years. The concepts of resilience and urban resilience are defined in the introductory section. The article then proceeds to outline the influential ecological conception of resilience and its application to cities, before highlighting some recent criticisms of this approach. The remainder of the article discusses developments in three main areas of research: adaptation to climate change and natural disasters, urban governance and institutions, and urban economic resilience. Finally, a brief summary highlights the main points of the article.

Introduction

The concept of resilience has attracted growing interest in urban and regional studies and public policy in recent years. It can be defined as the ability of an entity or system to recover its form and function following a disturbance or disruption of some kind (Simmie and Martin, 2010: 28). As such, the resilient city refers to its capacity to deal with external sources of stress and maintain or recover normal functioning in terms of dimensions such as economic production, transportation, housing, energy transmission, and social interaction. Key sources of disturbance include natural disasters such as Hurricanes Katrina or Sandy, climate change, international terrorism, and economic downturns such as that resulting from the financial crash of 2007–08. Such events focus considerable attention on the responsive capacities of cities to anticipate, manage, and adapt to change (Dawley et al., 2010). There are close parallels between resilience and the concept of sustainability, particularly in terms of the underlying imperative for cities and communities to become more resilient in the face of a range of externally derived environmental, economic, or political threats, which seems uncontentious and commonsensical (Swyngedouw, 2007). Indeed, Foster (2010: 3) claims that resilience is to the 2000s and 2010s what sustainability was to the 1980s and 1990s.

Resilience is becoming 'a pervasive idiom' of urban governance (Walker and Cooper, 2011: 144), requiring strategies for disaster mitigation and the protection of critical infrastructures as well as the development of procedures to anticipate and preempt key threats. The context for such strategies is derived from natural disasters such as Hurricane Katrina, which devastated much of New Orleans in 2005, and terrorist incidents like the attacks on New York in 2001 and Mumbai in 2008. Levels of resilience reflect a city's vulnerability to surprise and shock in the sense that high vulnerability is associated with low resilience, implying a strong degree of exposure to potential threats and a lack of anticipation and preparedness (Foster, 2010: 4–5). The term is also invoked by progressive social and environmental movements that are aiming to cultivate alternative social relations based upon environmental sustainability and social well-being rather than economic growth. From this perspective, resilience has connotations of ecological and social stability and integrity that are consistent with initiatives to localize the supply of food and energy, contrary to the environmentally destructive practices of global capitalism (Hudson, 2010). As such, the resilient city is not only a popular concept, but also a contested one that is capable of being deployed and mobilized by different groups in different ways.

Resilience and Cities

The concept of resilience has roots in both physics and mathematics, where it refers to the capacity of a system or material to recover its shape following a displacement or disturbance (Norris et al., 2008), and ecology where it emphasizes the capacity of an ecosystem to absorb shocks and maintain functioning (Holling, 1973). Subsequent applications to a number of objects from the built environment to individuals, social systems, and communities have spawned a range of definitions (Table 1). The work of the American ecologist C.S. Holling (1973, 2001) has proved particularly influential, not least through his role in groups such as the Resilience Alliance of scientists and the Stockholm Resilience Centre, a high-profile think tank. In the ecological literature, two types of resilience are commonly identified. The first is 'engineering resilience,' which is concerned with the stability of a system near an equilibrium or steady state, where resilience is defined in terms of elasticity, which emphasizes resistance to disruption and speed of return to the preexisting equilibrium. Second, 'ecological resilience' refers to external disturbances and shocks that result in a system becoming transformed through the emergence of new structures and behaviors. As these concepts suggest, resilience is based upon systems thinking and particularly the notion of complex adaptive systems based on nonlinear dynamics and self-organization, meaning that large-scale structures tend to emerge spontaneously from the interactions between microlevel processes (Holling, 2001).

Resilience is closely associated with the ecological idea of 'panarchy,' which emphasizes the evolutionary character of adaptive cycles seen as nested within each other across time and space (Holling, 2001). Four key phases of the evolution of an ecosystem such as a forest are identified, each of which corresponds with a particular level of resilience (Figure 1). The first 'exploitation' phase, which equates to the early phase of development, is defined by rapid growth and high

Table 1 Selected definitions of resilience

Author, date	*Discipline*	*Level of analysis*	*Definition*
Gordon, 1978	Physics	Physical system	The ability to store energy and deflect elasticity under a load without breaking or being deformed
Holling, 1973	Ecology	Ecological system	The persistence of relationships within a system; the ability of systems to absorb change and still persist
Walker et al., 2004	Ecology	Ecological system	The capacity of a system to absorb disturbance and reorganize while undergoing change so as to retain the same function, structure, and feedbacks
Egeland et al., 1993	Psychology	Individual	The capacity for successful adaptation and functioning despite high risk, stress, or trauma
Agder, 2000	Geography	Social	The ability of communities to withstand external shocks to their social infrastructure
Katz, 2004	Geography	Social	Ways in which people adapt to changing circumstances to get by and 'make do' through the exercising of autonomous initiative
Hill et al., 2008	Interdisciplinary social science	Region	The ability of a region to recover successfully form shocks to its economy
Martin, 2012	Geography	Region	Region's resistance to, and recovery from, major economic shocks

Reproduced from: MacKinnon, D., Derickson, K.D., 2013. From resilience to resourcefulness: a critique of resilience policy and activism. Progress in Human Geography 37, 253–270. http://dx.doi.org/10.1177/0309132512454775. http://phg.sagepub.com/content/early/2012/08/08/0309132512454775; Adapted and extended from: Norris, F.H., Stevens, S.P., Pfefferbaum, B., Wyche, K.F., Pfefferbaum, R.L., 2008. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. American Journal of Community Psychology 41, 129.

resilience as resources are accumulated and openness and emergence promote adaptability. A successful forest moves into the 'conservation' phase as it matures, resulting in reduced resilience as internal relationships become fixed and rigid. This renders the ecosystem vulnerable to external shock and stress, which propels the transition to a third 'release' phase based on crisis and destruction, involving the expulsion and exhaustion of accumulated resources (Holling, 2001). This facilities a fourth phase of 'reorganization' through experimentation and renewal as new arrangements begin to emerge and take hold. Resilience is initially low in the release phase, but begins to increase as the remnants of the system become sensitive to external forces and reorganization gathers momentum. The underlying idea is that ecosystems maintain their resilience by undergoing such cycles, with destruction essentially purging the system of its vulnerability and paving the way for renewed growth (Swanstrom, 2008).

In recent years, the ecologically rooted concept of resilience has rapidly infiltrated the fields of urban governance and planning, based on concerns about security, critical infrastructure management, public health, economic development, and planning. Walker and Cooper (2011: 2) argue that the success of resilience in 'colonizing multiple arenas of governance' reflects its ideological fit with neoliberalism as the dominant contemporary approach to economic policy that espouses individual self-reliance, competition, and private property rights. In an urban context, neoliberalism is associated with a framework of spatial competition in which cities compete for investment, markets, and resources. Enhanced urban competitiveness is seen as the underlying source of economic success and prosperity, based upon the

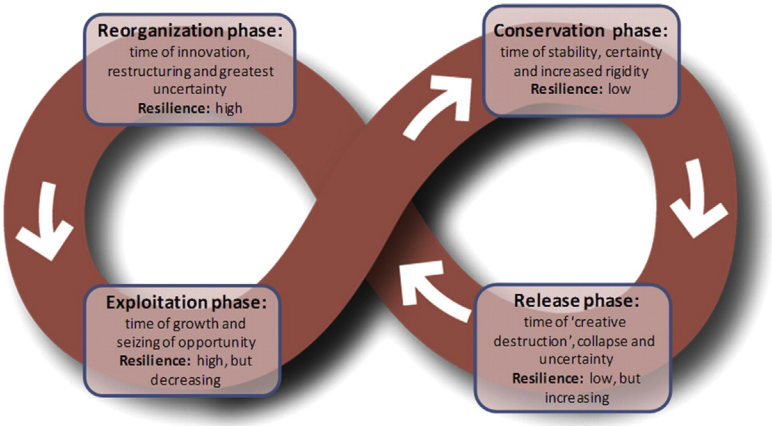

Figure 1 Ecological studies of resilience propose a four-phase cycle of system adaptation and change. Adapted from: Holling, C.S., Gunderson, L.H., 2002. Resilience and adaptive cycles. In: Gunderson, L.H., Holling, C.S. (Eds.), Panarchy: Understanding Transformations in Human and Natural Systems. Island Press, Washington DC, pp. 63–102.; Pendall, R. et al., 2010. Cambridge Journal of Regions, Economy and Society 3, 71–84.

harnessing of local resources and assets through initiatives that seek to upgrade workforce skills, stimulate the formation of new firms, and develop new facilities and cultural attractions (e.g., shopping centers, conference venues, hotels, and museums). Resilience thinking derived from ecology serves to reinforce and extend the focus on cities and regions as units of competition, emphasizing the need for flexibility and responsiveness in the face of a range of external threats.

Resilience can be seen as the latest in a long line of naturalistic metaphors to be applied to cities (Gandy, 2002). Organic conceptions of cities as systems displaying natural traits such as growth, competition, and self-organization have proved particularly influential, informing the urban ecologies of Patrick Geddes, Ebenezer Howard, and the Chicago School (Evans, 2011). The use of natural metaphors had implications for the governance and management of cities into the twentieth century, as Gandy (2002: 8) observes:

> In the twentieth century, a range of technological advances facilitated a new mediation between organic metaphors and the production of urban space. In 1965, for example, the engineer Abel Wolman outlined "the complete metabolism of the modern city" as the culmination of advances in the technical organisation of urban space. Yet these metabolic metaphors treat the city as a discrete physical entity. The "body of the city" is considered in isolation from wider determinants of urban form, and the social production of space is downplayed in relation to the technical mastery of cities.

In the context of scientific efforts to create more resilient urban infrastructures, Evans (2011: 224) suggests that ecology may come to play a key role in the shaping of twenty-first century cities. Informed by theories of ecosystems as complex adaptive systems, this new urban ecology conceives of the city as a social–ecological system, in which biophysical and social factors are linked by multiple feedback loops and exhibit the common properties of resilience and complexity (Evans, 2011: 229). The effect is to naturalize cities as self-contained systems by divorcing them 'from wider determinants of urban form' such as flows of capital and modes of state regulation (Gandy, 2002: 8).

In their critique of the notion of resilient cities and regions, MacKinnon and Derickson (2013) argue that the ecological concept is conservative when applied to the social sphere, referring to the stability of a system against interference. This is invariably defined in terms of recovery, rebound, and return to a preexisting set of norms, notwithstanding notions of 'bounce forward,' which commonly project these existing norms into the future. The implication of the extension of ecological thinking to the social sphere is that human society should mimic the decentralized and resilient processes of nature (Swanstrom, 2008: 15). Deference to 'the emergent order of nature' (Swanstrom, 2008: 16) is implicitly extended to society as existing social networks and institutions are taken for granted as 'natural' and harmonious. As a result, the existence of social divisions and inequalities tends to be glossed over. Ecological models of resilience are fundamentally antipolitical, viewing adaptation to change in terms of decentralized actors, systems, and relationships and failing to accommodate the critical role of the state and politics (Evans, 2011). This prompts Swanstrom (2008: 16) to argue that the ecological model of resilience is 'profoundly conservative' when it is exported into a social context. This conservatism is reinforced by the normative aspect of resilience, which is assumed to be always a positive quality, imbued with notions of individual self-reliance and triumph over adversity.

Urban Resilience, Climate Change, and Disaster Management

Research on urban ecological resilience tends to operationalize the term 'resilience' as it pertains to the ability of cities to both withstand sudden dramatic shocks such as a major weather event and adapt to more long-term disruptions such as those related to climate change, which refers to a long-term shift in underlying weather patterns that results from a range of natural and human-induced factors. The Community and Regional Resilience Initiative in the Southeast of the United States, for instance, defines resilience as "a community's capability to prepare for, respond to and recover from significant multi-hazards threats with minimum damage to public safety and health, the economy and national security" (Colten et al., 2008: 20). Enhancing a community's resilience involves a fourfold process of anticipating major multihazard threats, reducing its overall vulnerability to hazard events, responding to such events when they occur, and longer term recovery from them (Colten et al., 2008). This framework is then used to assess the resilience of New Orleans in the face of Hurricane Katrina. The likelihood of such an event had been widely anticipated following experience of previous hurricane strikes and the accurate tracking of Katrina across the Gulf of Mexico, but planning did not anticipate the widespread and massive levee failure that occurred, resulting in the inundation of about 80% of the city. As has been widely documented, the response was slow and inadequate. Subsequent recovery has proved highly uneven with the economy recovering fairly well, while repopulation and rehousing has been slower, particularly in the most severely affected areas of Eastern New Orleans, which are predominantly African-American. Reduction of the city's vulnerability to future events has focused on the hurricane protection systems of levees and pumps and a revamping of emergency response planning, although progress has been slow and the fundamental sources of geophysical and social vulnerability in New Orleans – its below sea level position; accelerating subsidence; rising sea levels; exposure to storm surges and hurricanes, together with development in low-lying areas; poverty; aging segments of the population; and large groups that lack access to private transport – remain in place.

Climate change poses a variety of challenges to cities across the world, especially those in coastal locations, prompting city leaders and governments to assess a range of vulnerabilities, impacts, and adaptation options (Leichenko and Thomas, 2012). In this context, resilience spans the ability to resist or withstand impacts and the capacity to recover and reorganize so as to restore previous levels of functioning (Brown et al., 2012: 534). A number of studies have shown how socio-economic conditions interact with physical factors to shape levels of vulnerability and resilience (Leichenko and Thomas, 2012). One stand of research on the vulnerability of coastal regions of the United States, for example, emphasizes that

high levels of poverty, low levels of education, and low median house prices accentuate urban communities' exposure to shocks and undermine their capacity to respond, thus reducing overall resilience (Leichenko and Thomas, 2012). The importance of such socioeconomic factors means that cities in the developing world are often particularly vulnerable to the impacts of climate change (Brown et al., 2012). For instance, cities in Latin America are "not only growing sources of greenhouse gases, but … are also hotspots of vulnerability to floods, heat waves, and other hazards that climate change is expected to aggravate" (Hardoy and Romero Lankao, 2011: 158). Urban growth and fragmentation are key drivers of increased emissions in the region, particularly in terms of transportation, which highlights the contradiction between responsibility for contributing to the causes of climate change and exposure to its impacts. In Mexico City, for instance, private cars, used predominantly by higher income groups, contribute only 18% of daily trip segments but 40.8% of CO_2 equivalent emissions, while public transport accounts for 82% of daily trip segments but 29.5% of CO_2 emissions (Hardoy and Romero Lankao, 2011: 159).

In an effort to enhance the adaptive capacities of urban populations and governments, the Rockefeller Foundation has supported the Asian Cities Climate Change Resilience Network (ACCCRN) since late 2007. ACCCRN is focused on so-called second-tier cities in which more than 60% of the projected increase in population between 2012 and 2050 is expected to occur, initially selecting 10 cities across four countries (India, Indonesia, Thailand, and Vietnam). It has identified 10 key action areas for cities to address in order to increase their resilience to climate change (Brown et al., 2012):

- Climate sensitive land use and urban planning
- Institutional coordination mechanisms and capacity support
- Drainage, flood, and solid waste management
- Water demand and conservation systems
- Emergency management and early warning systems
- Responsive health systems
- Resilient housing and transport systems
- Strengthening ecosystem services
- Diversification and protection of climate-affected livelihoods
- Education and capacity building of citizens

These key themes were identified through an inclusive and iterative process of analysis, planning, and engagement that involved multiple stakeholders, resulting in a locally owned strategy for each city. This is seen as a model for other cities to adopt, with successful implementation also requiring expanded national and international policy support and financing, alongside increased public awareness and the establishment of new national and local institutions responsible for urban climate change resilience.

Several crosscutting issues emerge from the studies of urban climate change resilience reviewed in this section. First, the political economy of urban development based upon capitalist land markets means that public policies related to climate change adaption are shaped and mediated by myriad private interests (Brown et al., 2012). As such, long-term goals of security and sustainability are often crowded out by short-term commercial and fiscal considerations. Second, questions of equity and social and spatial justice are paramount as low-income groups in cities tend to be most exposed to the impacts of climate change, but least responsible for generating emissions. This requires the adoption of pro-poor strategies based on the improvement of housing conditions, infrastructure, and services. Third, governance and the development of institutional capacity and coordination remain a key challenge since most city governments in the developing world lack the power and authority to develop and implement effective climate change adaptation strategies (Brown et al., 2012). Addressing these deficits requires support from higher levels of government and international organizations and nongovernmental organizations.

Resilience and Urban Governance

Work in this area examines the effects of different institutional arrangements on resilience and assesses how resilience can inform the development of new governance strategies for adapting to external stresses and shocks (Leichenko, 2011). In this context, urban governance refers to the range of institutions and agencies that are involved in governing a particular city.

Research typically stresses the need to anticipate external events and to respond flexibly and efficiently, based upon polycentric and accountable institutions that operate in a transparent and inclusive manner. For instance, Lebel et al. (2006) identify three key dimensions: public participation and deliberation, which are seen as essential to building trust and shared understanding; polycentric and multilayered institutions that support local knowledge and local monitoring and facilitate interaction and feedback between the different levels; and accountable and just institutions that pursue socially just distributions of benefits and risk among the populations they serve. The emphasis on participation, local accountability, and polycentric institutions not only echoes the approach to 'good governance' adopted by prominent international organizations like the World Bank over the past couple of decades, but also reflects the properties of self-organization and decentralization that characterize the ecological concept of resilience (Holling, 1973). The associated suspicion of central authority in resilience thinking has much in common with neoliberal principles of limited government and the 'enabling state' (Walker and Cooper, 2011). This means that the emphasis on self-organization and polycentric institutions remains curiously divorced from the sociopolitical realities of state authority through which resilience policies have often been formulated and implemented, while the emphasis on participation and inclusion raises broader issues of equity and social justice that resilience thinking is poorly equipped to address (MacKinnon and Derickson, 2013).

Much like sustainability, resilience is externally defined as a crucial imperative by state agencies and expert commentators in spheres such as security, emergency planning, economic development, and urban design (Walker and Cooper, 2011). With regard to security, for instance, the identification of 'new' global risks, coupled with political leaders' claims of 'unique' and 'classified' knowledge of potential threats, justified 'the implementation of a raft of resilience policies

without critical civic consultation' following the events of 11 September 2001 (Coaffee and Rogers, 2008: 115). The development and implementation of policies to combat an unpredictable terrorist threat requires what has been termed a 'multiscale governance fix' (Coaffee and Murakami Wood, 2006: 509). This involves a partial rescaling of security from the national to the urban levels of governance through a range of physical and technological controls at specific sites (e.g., visible policing and security patrols, closed-circuit television), often supported by legislative measures and regulatory guidance. These site-specific controls are often temporarily utilized to reborder cities through the hosting of major cultural, sporting, or political events and conferences. Securitization has become increasingly linked to urban economic development through urban marketing strategies stressing the 'safety' and 'security' of cities as places to conduct business (Coaffee and Murakami Wood, 2006).

The emerging literature on resilience in the context of urban economic development emphasizes the importance of strategic planning and 'place-renewing leadership' as "a form of public–private strategic leadership that empowers institutional or social forms of decision-making to absorb and adjust (proactively and reactively) to path-breaking economic change" (Bailey et al., 2010: 462). The role of urban policy makers is to foster the adaptive capacity to enable the renewal and 'branching out' of economic activity from existing local assets, echoing the processes that seem to have underpinned the development of successful regions such as Cambridge (Dawley et al., 2010; Simmie and Martin, 2010). As such, resilience is serving to reinforce and extend existing trends in urban development policy toward increased responsiveness to external conditions, strategic management, capacity building, and the harnessing of local assets. This helps to account for its adoption by urban and regional policy makers, providing a somewhat more muted successor to the 'creative cities' craze of the mid-2000s (Florida, 2002). Like the creative cities script, resilience is a mobilizing discourse, confronting individuals, communities, and organizations with the imperative of ongoing adaptation to the challenges of an increasingly turbulent environment (Peck, 2010: 221). Yet the 'pseudoscientific discourse' of resilience presents something of a paradox of change, emphasizing the prevalence of turbulence and crisis, but accepting them passively and placing the onus on communities to get on with the business of adapting (Evans, 2011: 234). The effect is to naturalize crisis, obscuring the social roots of 'shocks' such as economic downturns, ongoing processes of climate change and terrorist attacks, as well as the social sources of vulnerability to natural disasters such as hurricane strikes and tsunamis.

Urban governance agencies have placed increasing emphasis on the social and community aspects of resilience in recent years as part of an effort to raise public awareness of potential threats and to encourage increased 'responsibilization' by involving citizens and communities in their own risk management (Coaffee and Rogers, 2008: 113). In the United Kingdom, community resilience has been redefined in terms of 'communities and individuals harnessing local resources and expertise to help themselves in an emergency' (Cabinet Office, 2011: 4) as part of the Coalition Government's 'Big Society' program of renewed localism and voluntary action since 2010 (Featherstone et al., 2012). This fosters an internalist understanding of the local, which sees the sources of resilience as lying within the particular city, neighborhood, or community in question, particularly through the identification of specific assets, capacities, and resources to be cultivated and mobilized. The notion of community resilience has also been adopted by a range of community groups and initiatives such as the rapidly growing Transition Towns movement (Carnegie UK Trust, 2011). While couched within a tradition of radical localism and community action, these 'bottom–up' initiatives echo the language of self-help and self-reliance that animates mainstream policy. The 'Big Society' context in the United Kingdom indicates that arguments for greater urban and community resilience can be utilized to support a program of fiscal austerity involving reductions in public expenditure (Featherstone et al., 2012). In the context of uneven geographical development and entrenched social inequality, there is a sense in which the promotion of decentralized governance and participation as a means of enacting 'resilience' risks simply reinvigorating the trope of individual responsibility with a renewed 'community' twist. Moreover, viewing the sources of resilience as lying within a particular community or neighborhood can all too easily foster a deficit model of policy, which identifies particular areas (often low-income ones) as problematic due to an apparent lack of internal capacity, leadership, or social capital (Mohan and Mohan, 2002).

The Resilience of Urban Economies

A third key strand of research is concerned with the resilience of urban economies (Leichenko and Thomas, 2012). Definitions of urban economic resilience in the literature span three interrelated dimensions: the ability of an urban economy to withstand external pressure (resistance), its capacity to respond positively to external change, and its longer term adaptability or learning capacities (Davies, 2011; Hudson, 2010). Martin (2012) applies this idea of economic resilience by examining the responses of UK cities and regions to the major recessions of 1979–83, 1990–93, and 2008–10. The first recession had the greatest impact on the old industrial regions of the peripheral and northern United Kingdom, whereas the 1990–93 downturn affected the greater South East and Midlands most severely, and the effects of the post-2008 recession have been less clearly differentiated, although Northern and peripheral regions such as Wales, Northern Ireland, Yorkshire–Humberside, and Scotland have been disproportionately affected.

In a preliminary empirical exploration, Simmie and Martin (2010) apply the panarchy model of adaptive cycles (Figure 1) to the UK cities of Cambridge and Swansea. They identify that the former has undergone three main phases since the 1960s of 'reorganization,' 'exploitation,' and 'conservation,' while Swansea has experienced six: 'release' in terms of the decline of traditional extractive industries; the reorganization, exploitation, conservation, and release phases of a model based on the attraction of foreign direct investment (FDI); and recent efforts to establish a new growth path. The Cambridge economy proved far more resilient in response to the early

1980s and 1990s recessions with University-related research providing a strong platform for the growth of a high-tech innovative economy. In Swansea, by contrast, manufacturing employment never fully recovered from the shock of the early 1980s recession and FDI in consumer electronics was unable to provide a sufficient basis for sustained growth. Prosperous cities such as Greater London or Cambridge invariably emerge as more resilient than less favored ones like Swansea or Newcastle from such analyses, reflecting their greater capacities for adaption and diversification over time. These differences in performance are not simply the result of divergent endogenous capacities for innovation, capacity building, and leadership as ecological conceptions of resilience would suggest, but are bound up with the operation of a range of wider political and economic relations that have positioned cities such as London as leading world cities and old industrial cities like Swansea as economically marginal.

Other research has examined the economic restructuring and adaption of industrial city-regions in Europe and beyond since the 1980s (Beatty et al., 2006; Grabher, 1993). Different forms of recovery and reorientation have been identified, particularly in terms of the contrast between industrial cities in France and Germany, which have tended to focus on diversification into new but related technologies and their British equivalents, which largely relied upon the attraction of unrelated industries and services through FDI. The former group has certainly proved more resistant to deindustrialization, although not at the expense of longer term adaptability, while many former industrial cities in the United Kingdom have become highly dependent on public services, leaving them exposed to future cutbacks in expenditure. Research on the renaissance of so-called phoenix cities in Europe, which had a high proportion of manufacturing jobs in the 1970s, has identified the key factors behind their emergent recovery in the 1990s and 2000s. These include institutional leadership and public funding, physical restoration, a focus on employment and skills, neighborhood renewal, social inclusion, and public investment (Power et al., 2008). Recent analyses of the urban and regional effects of the post-2008 recession in Europe indicate that manufacturing-based cities have often been worst hit (Davies, 2011). In the United Kingdom, research has shown that the impact of the initial 2008–09 recession has been the greatest in those local authority areas in which unemployment was already high in March 2008 when the recession began, meaning that Northern city regions with the weakest labor markets have generally been the hardest hit (Dolphin, 2009).

In another study, Swanstrom et al. (2009) examine how metropolitan areas in the United States have responded to the foreclosure crisis in recent years, defining resilience in terms of three main processes: the redeployment of assets or alteration of organizational routines; collaboration within and across public, private, and nonprofit sectors; and the mobilization and capturing of resources from external sources. They focus on resilience and foreclosure prevention in six metropolitan areas: St. Louis, Cleveland, East Bay (California), Riverside, Chicago, and Atlanta. The first two are defined as weak housing market areas, the second two as traditionally stronger market areas, and the last two as mixed. On the basis of these paired comparisons, they found that Cleveland was more resilient than St. Louis, Riverside performed better than the Bay Area, and Chicago was more active than Atlanta (Swanstrom et al., 2009: 46). Interestingly, while 'horizontal' collaboration between public, private, and nonprofit actors within metropolitan areas was important, Swanstrom et al. (2009) argue that the effects of this would remain limited without support from 'vertical' policies emanating from the state and federal scales of government. Only these institutions can provide the necessary level of resources to support local foreclosure prevention and neighborhood recovery activities. Contrary to much resilience thinking, this indicates that external institutions and relationships can play a crucial role in fostering urban resilience in addition to internal factors.

Summary

As this article has emphasized, the idea of the resilient city has proved highly influential in recent years, referring to the ability of a city to respond to external sources of disturbance and maintain its basic functioning. Notions of urban resilience are derived largely from ecology and systems theory, which emphasize the self-organizing properties of complex, adaptive systems in responding to external disturbances and threats. As such, resilience can be seen as the latest in a line of ecological concepts to be applied to cites, viewing them as self-contained and autonomous entities or units analogous to ecosystems (Evans, 2011). Critics have argued that the ecological model has conservative connotations in the social sphere, accepting existing social relations and networks as natural and worthy of preservation, while having little conception of conflictual social relations, politics, and the state (MacKinnon and Derickson, 2013; Swanstrom, 2008). In an urban context, the ecologically rooted notion of resilience stresses the need for decentralized and polycentric institutions to foster adaptive and responsive forms of governance. At the same time, studies of economic resilience focus attention on the endogenous capacities of cities, often overlooking the role of external linkages and networks. Accordingly, the ecologically based notion of the resilient city tends to foster an internalist approach, which sees the sources of resilience as lying within a particular city, community, or neighborhood. This can easily slide into a deficit model that effectively blames particular groups, areas, and individuals as responsible for a lack of resilience, although some research has pointed to the importance of external relations and institutions (Swanstrom et al., 2009). In response, there is a need for scholars and policy makers to adapt a more comprehensive and critical approach to urban resilience that moves beyond the confines of ecology and systems thinking to accommodate the role of wider social, economic, and political relations in shaping the resilience of cities.

Bibliography

Agder, W.N., 2000. Social and ecological resilience: are they related? Progress in Human Geography 24, 347–364.

Bailey, D., Bellandi, M., Caloffi, A., De Propris, L., 2010. Trajectories of change for mature manufacturing regions in Europe: the role of place-renewing leadership. Policy Studies 31, 457–474.

Beatty, C., Fothergill, S., Powell, R., 2006. Twenty years on: has the economy of the UK coalfields recovered. Environment and Planning A 39, 1654–1675.

Brown, A., Dayal, A., Del Rio, C.R., 2012. From practice to theory: emerging lessons from Asia for building climate change resilience. Environment and Urbanization 24, 531–556.

Cabinet Office, 2011. Strategic National Framework on Community Resilience. UK government, London. http://www.cabinetoffice.gov.uk/sites/default/files/resources/Strategic-National-Framework-on-Community-Resilience_0.pdf (accessed 24 October 2011).

Carnegie UK Trust, 2011. Exploring Community Resilience in Times of Rapid Change. Fiery Spirits Community of Practice, Dunfermline.

Coaffee, J., Murakami Wood, D., 2006. Security is coming home: rethinking scale and constructing resilience in the global urban response to terrorist risk. International Relations 20, 503–517.

Coaffee, J., Rogers, P., 2008. Rebordering the city for new security challenges: from counter-terrorism to community resilience. Space and Polity 12, 101–118.

Colten, C.E., Kates, R.W., Laska, S.B., 2008. Community Resilience: Lessons from New Orleans and Hurricane Katrina. Community and Regional Research Initiative (CARRI) Report 3. National Security Directorate, Oak Ridge National Laboratory, Oak Ridge, TN.

Davies, S., 2011. Regional resilience in the 2008–2010 downturn: comparative evidence form European countries. Cambridge Journal of Regions, Economy and Society 4, 362–389.

Dawley, S., Pike, Tomaney, J., 2010. Towards the resilient region? Local Economy 25, 650–667.

Dolphin, T., 2009. The Impact of the Recession on Northern City-Regions. IPPR North, Newcastle.

Egeland, B., Carlson, E., Sroufe, L., 1993. Resilience as process. Development and Psychopathology 5, 517–5238.

Evans, J.P., 2011. Resilience, ecology and adaptation in the experimental city. Transactions of the Institute of British Geographers 36, 223–237.

Featherstone, D., Ince, A., MacKinnon, D., Cumbers, A., Strauss, K., 2012. Progressive localism and the construction of political alternatives. Transactions of the Institute of British Geographers NS 37, 177–182.

Florida, R., 2002. The Rise of the Creative Classes. Basic Books, New York.

Foster, K., 2010. Regional Resilience: How Do We Know It When We See It? University at Buffalo Regional Institute. http://www.gwu.edu (accessed 26 November 2012).

Gandy, M., 2002. Concrete and Clay: Reworking Nature in New York City. The MIT Press, Cambridge, MA.

Gordon, J., 1978. Structures. Penguin, London.

Grabher, G., 1993. The weakness of strong ties: the lock-in of regional development in the Ruhr Area. In: Grabher, G. (Ed.), The Embedded Firm: On the Socio-Economics of Industrial Networks. Routledge, London, pp. 255–277.

Hardoy, J., Romero Lankao, P., 2011. Latin American cities and climate change: challenges and options to mitigation and adaptation responses. Current Opinion in Environmental Sustainability 3, 158–163.

Hill, E.W., Wial, H., Wolman, H., 2008. Exploring Regional Resilience. Working Paper 2008-04. Macarthur Foundation Research Network on Building Resilient Regions, Institute for Urban and Regional Development. University of California, Berkeley, CA.

Holling, C.S., 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4, 1–23.

Holling, C.S., 2001. Understanding the complexity of economic, ecological and social systems. Ecosystems 4, 390–405.

Holling, C.S., Gunderson, L.H., 2002. Resilience and adaptive cycles. In: Gunderson, L.H., Holling, C.S. (Eds.), Panarchy: Understanding Transformations in Human and Natural Systems. Island Press, Washington, DC, pp. 63–102.

Hudson, R., 2010. Resilient regions in an uncertain world: wishful thinking or a practical reality? Cambridge Journal of Regions, Economy and Society 3, 11–25.

Katz, C., 2004. Growing Up Global: Economic Restructuring and Children's Everyday Lives. University of Minnesota Press, Minneapolis.

Lebel, L., Anderies, J., Campbell, B., Folke, C., Hatfield-Dodds, S., Hughes, T.P., Wilson, J., 2006. Governance and the capacity to manage resilience in regional social-ecological systems. Ecology and Society 11, 19. http://www.ecologyandsociety.org/vol11/iss1/art19/.

Leichenko, R., 2011. Climate change and urban resilience. Current Opinion in Environmental Sustainability 3, 164–168.

Leichenko, R., Thomas, A., 2012. Coastal cities and regions in a changing climate: economic impacts, risks and vulnerabilities. Geography Compass 6, 327–339.

MacKinnon, D., Derickson, K.D., 2013. From resilience to resourcefulness: a critique of resilience policy and activism. Progress in Human Geography 37, 253–270. http://dx.doi.org/10.1177/0309132512454775. http://phg.sagepub.com/content/early/2012/08/08/0309132512454775.

Martin, R., 2012. Regional economic resilience, hysteris and recessionary shocks. Journal of Economic Geography 12, 1–32.

Mohan, G., Mohan, J., 2002. Placing social capital. Progress in Human Geography 26, 191–210.

Norris, F.H., Stevens, S.P., Pfefferbaum, B., Wyche, K.F., Pfefferbaum, R.L., 2008. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. American Journal of Community Psychology 41, 127–150.

Peck, J., 2010. Constructions of Neoliberal Reason. Oxford University Press, Oxford.

Power, A., Plöger, J., Winkler, A., 2008. Transforming Cities across Europe, CASE Report 49. Centre for Analysis of Social Exclusion, LSE, London.

Simmie, J., Martin, R., 2010. The economic resilience of regions: towards an evolutionary approach. Cambridge Journal of Regions, Economy and Society 3, 27–44.

Swanstrom, T., 2008. Regional Resilience: A Critical Examination of the Ecological Framework. Working Paper 2008–07. Macarthur Foundation Research Network on Building Resilient Regions, Institute for Urban and Regional Development, University of California, Berkeley, CA.

Swanstrom, T., Chapple, K., Immergluck, D., 2009. Regional Resilience in the Face of Foreclosures: Evidence from Six Metropolitan Areas. Working Paper 2009–05. Macarthur Foundation Research Network on Building Resilient Regions, Institute for Urban and Regional Development, University of California, Berkeley, CA.

Swyngedouw, E., 2007. Impossible/undesirable "sustainability" and the post-political condition. In: Krueger, R., Gibbs, D. (Eds.), The Sustainable Development Paradox. Guildford Press, New York, pp. 13–40.

Walker, J., Cooper, M., 2011. Genealogies of resilience: from systems ecology to the political economy of crisis adaptation. Security Dialogue 43, 143–160.

Walker, B., Holling, C.S., Carpenter, S.R., Kinzig, A., 2004. Resilience, adaptability and transformability in social–ecological systems. Ecology and Society 9 (2), 5. http://www.ecologyandsociety.org/vol9/iss2/art5/.

Resource Geography

Roger Hayter, Simon Fraser University, Burnaby, BC, Canada
Jerry Patchell, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Bridge, volume 19, pp. 13266–13269, © 2001, Elsevier Ltd.

Abstract

This article examines and proposes connections between resource and economic geography from an institutional perspective. Economic geography urgently needs to more substantively embrace resources from its theoretical purview while resource geography's established approaches can be enriched by economic geography's focus on space–place relations, institutional interests and power structures, and the influence of markets on resource use. The article progressively elaborates on the themes of natural resource governance, property rights, evolutionary and spatial dynamics, resource markets, their coordination in value chains and cycles, and the nature of resource peripheries and their remapping that is driven by conflicting institutional considerations.

Contemporary definitions of (natural) resources typically take their cue from Zimmerman's (1933: pp. 3–9) view that "resources are not they become." Thus for Bridge (2000: p. 13266) resources are not 'givens' but "culturally mediated appraisals of the physical environment" that are shaped by economic factors, political institutions, social attributes, and belief systems (see also Mitchell, 1989: p. 1; Hanink, 2000: p. 227). From this perspective, the task of resource geography is truly monumental: to explain how these mediations differentiate and integrate the global economy, and examine their well-being and environmental outcomes. As Geography developed along nomothetic lines after the 1950s, resource or environmental geography became a specialty subject focused largely on concrete problems and policy responses (Bridge, 2000: p. 13266). This applied, ecological emphasis on exploring the relationships between environment, resources, and society remains important. In textbook format, it has been elaborated thematically (e.g., Mather and Chapman, 1995; Mitchell 1989) and with respect to national jurisdictions such as Canada (e.g., Dearden and Mitchell, 2005; Draper, 1998). Research, on the other hand, has emphasized the local expressions in these relationships around the world (O'Riordan, 2001). Moreover, there is growing concern to more fully incorporate resources within economic geography's orbit. This article reflects on these attempts.

While economic geography textbooks invariably recognize the fundamental importance of resources to the global economy, resource exploitation has been peripheral to the subdiscipline's substantive concerns. The rationale for economic geography to address this neglect is urgent and severalfold. First, agglomerations and associated global or local production, value chains, or networks simply cannot exist without resources. Second, the relentlessly powerful centripetal tendencies of economic development simultaneously promote centrifugal tendencies in the search for resources in the remotest parts of the earth. Third, in recent decades paradigmatic shifts in society attitudes and 'cultural mediations' of resources have increasingly emphasized their environmental values and concern for the negative externalities of their exploitation, massively evidenced by climate warming, biodiversity loss, and large-scale degradation. Fourth, resource peripheries have become increasingly contested places around the globe and in effect center stage to battles over resource values and use. Fifth, resource-based MNCs remain among the biggest corporations in the world, and resource-based trade and foreign direct investment vitally important to global integration.

Although economic geography's reemerging interest in resources is evident from a variety of perspectives, this article is informed by an institutional approach (Bridge, 2008; Clapp, 1998; Essletzbichler, 2012). This approach is based on the reality that resources become rather than derive as 'gifts' of nature. Indeed, resources offer multiple (market and nonmarket) values causing use and abuse to invoke deeply felt public concerns and sentiments that require governments to pay special attention to the social acceptability of resource policies. The institutional reality is highlighted by property rights regimes that shape spatial patterns of resource use, conflicts driven by alternative value systems and power structures, and the integration of resources into the operations of corporations, value chains, and other institutions. In brief, the article broadly explores the relationships between location, organization, and markets of resource exploitation.

The remainder of the article progressively elaborates on the themes of natural resource governance and evolutionary and spatial dynamics, resource markets, and their coordination in value chains and cycles. The concluding sections focus on resource peripheries and their remapping, a global trend epitomized by complex resource conflicts driven by the intersection of competing economic, geopolitical, environmental, and cultural considerations. Given space limitations, the article is conceptual, refraining from discussing individual resource sectors nor engaging with ideas of human and capital resources. We also leave agriculture and resource-based tourism to their own articles.

Resources are extremely varied, often summarily classified as biotic or mineral, geographically as point and diffuse, and dynamically as renewable and nonrenewable depending upon whether the rate of depletion is less or more than the rate of replenishment. Resources typically require collection and some initial processing before private consumption or further industrial processing. Species of biotic resources such as forests

or fish vary enormously while even so-called resource commodities such as oil, coal, and lumber are differentiated in character with important implications for use. The measurement of the size and distribution of resources, long the subject of inquiry within economic geography, is not straightforward. Even visible resources such as trees are difficult to count even with modern-day technologies, discovery is an ongoing uncertain process, the viability of resources is enhanced and diminished by price and technological changes, substitution possibilities among resource inputs are considerable, and public policy can redefine resource values at the stroke of a pen. The Athabasca oil/tar sands, for example, had been used by indigenous people for several centuries to waterproof canoes and its recent massive commercial exploitation as a source of 'heavy' oil had to await technological innovations and oil price increases. Its continued expansion, perhaps even survival, now depends on public policy decisions in Canada and the United States that are seeking to balance wealth generating effects with environmental and cultural implications of its processing and distribution.

Ownership and Control of Resources (Property Rights)

Resource exploitation is rooted in the rules of governance and associated assignment of property rights, which, in general, are claims to benefits or income streams that imply exclusivity, security, and transferability. Exclusivity assigns sole rights of use to (resource) property, security is provided by law or social sanction, and transferability means that owners can sell or trade their property rights. Resource rights are held privately by individuals or corporations, in common by specified groups of people, and by governments; conversely 'open access' resources lack property rights. To crudely generalize, historically common property resources were widespread and underpinned different forms of aboriginal and peasant cultures that were rooted in tradition, custom, and the assignment of rights by elites. However, the onset of industrialization at the end of the eighteenth century, Polanyi's (1944) 'Great Transformation' toward market economies, privileged (legally enshrined) private resource property rights, often imposed at the expense of tradition and common property arrangements. This transformation involved enclosure movements in Europe and settler colonies in the New World, the former helping to stimulate the (massive) migration flows to the latter in the nineteenth century with the latter displacing indigenous populations through technology and military might, disease, and treaties (Diamond, 1997). The new property rights regime installed itself in an ad hoc, power-based manner. In Europe property rights maintained heredity titles, shaped by marriage arrangements and primogeniture (e.g., the UK's Royal family remain the largest private landowners), while European colonists stimulated settlement through various land grant systems (e.g., land given to railroad companies to fund their building and stimulate settlement).

Property rights assigned to land ownership, the template for property rights generally, can involve complex bundles of rights, obligations, and limits on resource extraction. The right to exploit resource above or below the ground can be allocated irrespective of land ownership and in the case where resources are mobile (e.g., oil, water, and fish), exploitation can be limited to prevent impacts outside the spatially designated area of property rights. Fundamentally, resource rights stipulate the right to withdrawal as set by institutional conditions that vary from place to place. These place-bound institutions are called resource regimes and refer to both the property rights regarding the use and transfer of resources and the (formal) de jure and (informal) *de facto* rules that govern these rights and their uses (Schlager and Ostrom, 1992; Vatn, 2007). The distribution of private rights can lead to direct conflict, such as when subsurface rights compromises access to surface rights. More generally, as illustrated by widespread environmental problems, private resource owners and users can have substantial indirect impacts on common property and open access resources.

Resource regimes are highly varied, and if the privatization of the commons is a relentless historical theme, there are many resources that remain publicly owned while privately controlled. For example, staking mineral rights have been used as a way of stimulating mineral exploration in remote areas in North America, allowing people to not only explore but also exploit mineral deposits on government or private land. During the 'Fordist' decades of the twentieth century resource exploitation escalated massively, and the so-called 'seven sisters' (oil MNCs) grew to dominate global petroleum reserves by the 1960s (O'Dell, 1963) by formalizing property rights within 'resource contracts' with host governments such as Saudi Arabia. These contracts or bargains were often quasiconfidential, but nevertheless took on a variety of forms that specified obligations, rights, and ownership conditions, the latter including outright privatization of resources but also more limited rights to resource use. In British Columbia, forests were initially sold into private ownership but this practice was stopped in 1912 following excessive speculation and overcutting. After 1947, to promote development throughout the province, the cutting rights of large forest tenures were leased or licensed over varying periods of time to large companies at relative low-stumpage or taxation levels in return for major capital investments in factories using local timber supplies and other (and evolving) obligations (Prudham, 2007).

For governments, the assignment and pricing of property rights are often controversial even when resources are primarily seen in industrial or economic growth terms. The nature, size, and temporal structure of royalties or taxes imposed on owners and/or controlling organizations of resource exploitation, along with the extent to which resource revenues should pay for existing public sector programs or be saved for use by future generations, are major issues. Important decisions also relate to the timing and speed of developments, and the extent and nature to which resource ownership or rights should be privatized among small firms, domestic champions, joint ventures, and MNCs. Thus the World Bank has been a major support of resource contracts with MNCs who they argue provide expertise, access to markets and finance, and abilities to promote development in remote areas. However, driven by fears over loss of domestic control, profits, and diversification opportunities, resource nationalism emerged as an important response as several developing countries, but also Canada, chose to return some resources to public control or at least establish champion domestic resource companies. For American MNCs

resource nationalism implied an 'obsolescing bargain' (Vernon, 1971). Indeed, in the contemporary oil sector, the biggest companies are now state-controlled, and several of the seven sisters have disappeared through acquisitions. The ownership of water resources has similarly become controversial as privatization is both seen as a way to promote efficient delivery and needless alienation of a gift of nature essential to life, and traditionally under public control (Bakker, 2005). Water use is also an issue in the spectacular expansion of oil shale fracking as a source of natural gas and oil that may threaten water tables in Dakota and other American states. The expansion has occurred on private farms that have no control over subsurface rights, unable to stop the fracking process by those who control these rights, often obtained decades ago.

Resource exploitation also occurs in the absence of property rights. In the coal mining region of Raniganj in India, for example, Lahiri-Dutt (2003) estimates that there are around 200 000 people involved in informal coal mining in which individuals collect coal from the tailings of wastes of the large mines and transport their coal, by foot or bicycle, to markets such as households and shops. If such behavior has social virtue levels of illegal poaching, for example, of endangered animals, fish, and trees, are massive, estimates (for 2009) placing the values between US$25 and 40 billion a year (Traffic, 2013: p. 8)

Resource Cycle Dynamics

Legal or illegal, embeddedness in nature, coupled with the limits to resource exploitation in any site, is the distinguishing basis for the evolutionary and spatial dynamics of resource exploitation. According to the resource cycle thesis, exploitation follows a life cycle curve based on discovery, slow initial growth, rapid growth, and a plateau effect as the region matures, followed by depletion and industrial decline (Clapp, 1998). Resource peripheries are exploited for the potential of new rich, low-cost resources – host governments encourage investment in social and economic infrastructure, while entrepreneurs and corporations compete to access the resource and provide services to resource producers. These actions compel labor to migrate for typically high-income jobs. The potential to large internal and external economies of scale is great and accompanying individual (marginal) increments such as new mines, pulp and paper mills, or oil drilling platforms are also large. Moreover, since the best, most accessible resources are typically extracted or harvested first, costs of exploitation typically increase. Meanwhile downward pressure on resource prices, for example, from increased competition, can imply a long-term cost-price squeeze for established operations (Freudenburg, 1992). In any given periphery, with depletion and competition from elsewhere, the resource cycle plateaus over time. While for nonrenewable resources, continued depletion inevitably leads to collapse, crises of deforestation, overfishing, and dustbowls indicate similar failures to renew renewable resources.

Geographically, resource cycles are constrained by nature and are powerful centrifugal forces in opening up remote 'new economic spaces.' Market prices in relation to production and transportation costs remain a fundamental point of departure for understanding the location dynamics of resource cycles (Hay, 1976). In conventional location theory terms, while transportation costs typically increase with distance from markets and the nature of transportation networks, processing costs are more irregular, dependent primarily upon size and quality of the resource and its ease of working. In this model, spatial pattern of mining reflects underlying cost structures and, for example, remote mines can operate when nearby mines are closed by substituting lower per capita processing costs for higher transportation costs. Indeed, the dispersal of mining (and other resource-based operations) operations to the distant corners of the globe is closely associated with increasing operating size that realize economies of scale, as well as investments in specialized transportation systems that reduce freight rates.

In practice, the nature of resource cycles in given peripheries are highly varied in duration, form, fluctuations, and intermittency. In the short run, operations can move in and out of production according to business cycle effects. Over the longer term, resource booms are reinforced by host governments seeking to encourage development by various incentives and favorable property rights conditions and by governments of major consuming countries or regions, such as the United States, Japan, EU, and now China seeking to secure low resources from a diversity of sources to ensure supply stability and pressure on prices (e.g., Gunton, 2003). Similarly, security interests underlying 'strategic resources' can help stimulate domestic supplies, as witnessed by recent ethanol, natural gas, and fracking booms in the United States (Kedron and Bagchi-Sen, 2011) or the search for alternatives as did Japan when China restricted its sales of rare earth minerals. Rich governments in particular have options to subsidize and prolong already mature resource cycles or, as happened in the rapid demise of the British coal industry in the 1980s, can chose to phase out operations for political as well as economic reasons. Notwithstanding the latter case, in general government policies that stimulate resource developments and encourage firms to seek secure supplies encourages more geographically diversified (and cyclical) patterns of exploitation than anticipated by cost considerations alone. Corporate rivalries in accessing resources reinforce this point (O'Dell, 1963). Further, the nature of resource cycles over the long run is highly varied, epitomized by technological changes that open up vast new resources, prolong the viability of existing operations or threaten them with obsolescence, and with possibilities for substitution. Cost reduction is not the only concern, as exemplified by public policy demand to develop renewable or less carbon-intensive sources of energy to replace conventional fossil fuels in homes, transportation, and business.

Resource Markets

The simplicity of competitive bidding and asking among many independent buyers and sellers for resources in some locations, such as in Japanese log auctions, is still evident (Reiffenstein and Hayter, 2006). Markets for most resources while retaining competitive price setting, however, no longer exist at a local basis and operate in an abstract manner. The prices for internationally traded, standardized commodities, are exchanged in highly competitive and liquid markets, but often in the form of

complex financial instruments (e.g., futures, contracts, swaps, and derivatives) (Zalik, 2010). The base price of fossil fuels, precious metals, and other minerals are all set on global markets, in the case of oil according to various regional and quality standards, such as West Texas Intermediate, North Sea Bent Crude Resource prices were once set in particular commodity exchanges, such as the Chicago Board of Trade, and then relayed around the globe, are increasingly set in electronic exchanges with instantaneous inputs from around the globe.

The viability of any resource exploitation must be measured against the prices set by these international markets, which react swiftly to political, economic, and technological changes and completely disregard production cost differences and any investments made by companies or countries. The relatively inexpensive exploitation of shale gas and oil is disrupting enormous investments and socioeconomic expectations in regions as diverse as Alberta and Nigeria. At the same time, many resource markets are strongly regulated at global, national, regional, and local levels. The Organization of Petroleum Exporting Countries (OPEC) has attempted to control its members export levels to maintain prices, local authorities have established environmental taxes, and Alberta's exports to the United States are sold at substantial discounts in part for quality reasons but also because of a lack of alternative markets and control over transfer prices by MNCs. In the timber trade, construction-grade Canadian lumber exports to the United States have faced tariff, export taxes, and quota limitations since the mid-1980s (Edenhoffer and Hayter, 2013a). Nationally and regionally the prices of utilities such as water and electricity are often set by bureaucracies or monitored closely by authorities with oversight. In addition, competition, land and labor costs, or other conditions may cause prices to vary somewhat from the global. A few commodities, such gravel or water, tend to be marketed more on a local or regional basis because of high-transportation costs.

Resource markets can be highly controversial, perhaps because of their derivation from nature with its connotations of public ownership and because of their role in providing essentials of life. In recent debates, these controversies have been especially intense when formerly public resources have been privatized. Within economic geography, these trends have been regarded as a neoliberalization or privatization of nature (Bridge and Jonas, 2002; Castree, 2008). Bakker (2010a,b), for example, has examined how water has been transformed from a natural, commonly held good by increasing marketization, that is, a shift from reliance on public provision to private provision, and allowing for predominantly locally governed resource to one infiltrated by global capital. Relatedly, agricultural commodities as well are becoming virtual water commodities as they embody the vast amount of water used in cultivation. Moreover, the marketization of ecosystem services (such as carbon sequestration, water retention, and soil formation), which although advocated by environmentalists, governments, and business as a means to rescue nature, have incurred unintended social and environmental consequences, such inequitable distribution of benefits and undervaluation of resources (Matulis, 2013; McAfee, 2012).

Carbon pricing is an emerging vital issue that usually imposes direct taxes on carbon produced or indirectly as a cap and trade system to quantify and internalize the environmental costs of green-house gas (GHG). Carbon markets especially target resource industries, and have been attempted in various jurisdictions but so far with uncertain, variable, and contentious effects at global, supranational, national, and regional levels (Bailey and Maresh, 2009; Bailey et al., 2011; Boyd, 2009; Boyd et al., 2011; Bumpus, 2011; Bumpus and Liverman, 2008; Ciscell, 2010; Demeritt, 2001; Gutiérrez, 2011; Knight, 2011; Lovell and Mackenzie, 2011; Robertson, 2011). Some of these regulations are perceived as a form of protection for domestic industries. For example, under present world trade rules, nations that have domestic GHG legislation have the right to impose tariffs on imports from jurisdictions that do not have this legislation. Such tariffs, already introduced by Japan and California and are anticipated in China, South Korea and for the United States as a whole, effectively reduce the returns of (foreign) resource suppliers. In California's case, while existing and proposed GHG tariffs occur on a wide range of resource products, in-state producers (including of heavy oil) are released from such taxes. In response to such trade protection, foreign resource peripheries need to reduce GHGs prior to export.

Value Chain Coordination of Markets

Although most commodities are priced at some point through competitive markets where there is reasonably symmetrical information among many buyers and sellers, these prices are influenced to a great degree by the internal and quasi-internal markets of global value chains (GVCs) in three ways. First and relating to the discussion of commodity markets, global and national resource industries are dominated by oligopolies and monopolies that have a large impact on the setting of prices. Indeed, antitrust (antimonopoly) laws were initiated to break Standard Oil control over petroleum prices in the late-nineteenth century, then the seven sisters effectively set a world price with its 'Gulf plus' agreement, which was superseded by OPEC's cartel.

Second, MNCs are powerful agents that organize GVCs or global production chains (GPNs) in terms of production, distribution, and recycling of resources. Within GVCs, the MNC and firms can exploit several opportunities to extract a fraction of the margin of the ultimate price of the commodity, or use the commodity as vehicle to market other goods and services. These GVCs are controlling resources such as petroleum or coal that after a relatively low level of refining/processing are marketed and consumed as a finished commodities that will be consumed by end users, Bridge (2008) has mapped out the contours of these global commodity chains using oil as an example. The stages of production are exploration, extraction, refining, distribution, consumption, and carbon capture. The backbone of this commodity chain is comprised of the major international oil companies, and more recently, the national oil companies such as Aramco, which integrate several stages of the chain. However, technical reasons for such integration only occurs at particular links in the chain, and at each stage there is room for competition and outsourcing; indeed, the major oil companies have increased outsourcing. In a finished commodity GVC, value capture, that is, the establishment of prices internal to the GVC, depends on the (1) benefits to internal integration from economies of scope and transaction

cost reductions; (2) demand–supply balance causing swings from consumer and to producer and back; (3) skills of firms at different stages of the production (e.g., by developing cost reducing technologies or innovative exploration methods); (4) trading on extraction rights between countries/owners and companies; and (5) imposition of carbon or other externality taxes. The oil GVC supplies a template for gas and coal-finished commodities although conditions from extraction to carbon sequestration differ significantly.

Third, GVCs incorporate myriad resources or embodied commodities (materials) such as hydrocarbons (plastics), metals, wood fibers, water, and other commodities in goods and services production. Interestingly, GVC analysis of manufactured goods (and industrial and economic geography generally) has largely ignored resource inputs, an increasingly untenable neglect because of environmental considerations. Resource geographers have focused primarily on sites of extraction and not delved into downstream concerns, while industrial geographers have simply assumed that commodity inputs were there. However, as ecological economists (Lovins et al., 1999) remind us, all manufactured capital depend on natural capital. Therein lies the rub because although goods and services cannot be produced without natural resource inputs, their value added within the GVC is minimal. Yet the logic of the GVC/GPN literature (Coe et al., 2008; Gereffi et al., 2005) is that the spatial distribution of commodity inputs is not only global, but value capture (and conversely value minimization) is negotiated through diverse institutions in different places. Moreover, the necessity to recover, recycle, and revalorize resources when products are discarded is recognized in the GVC/GNP literature, and reinforced by geographical contributions to analysis of industrial ecology (Deutz and Lyons, 2008; Gibbs et al., 2005; Lyons, 2007).

Urban Markets

Ultimately, demand for resources depends on urban markets, which directly and indirectly are the largest consumers of materials and fuels (Wackernagel and Rees, 1996). Global markets may set global prices, but it is the relative consumption, competition, and conservation of resources by cities that determines demand. Cities may appear dependent on resources, and may suffer from short-term scarcities, but because they can readily switch both location of supply and substitute most resources with other types of inputs, cities can act with high levels of impunity against resource scarcities. Moreover, urban systems are engines of innovation, constantly improving productivity and adding value to what they produce or changing what they produce, thereby improving levels of income in absolute and relative terms. On the other hand, the constant or relative value of resources has declined over time irrespective of increasing scarcity and nominally increasing costs of extraction, processing, and transportation. The terms of trade of the periphery are often in a constant, if irregular, state of decline.

Urban resource markets can be differentiated into public and private consumption. Governments or private companies provide infrastructure, buildings, and so on which consume large quantities of mineral, fuels, and water (concrete, asphalt, etc.), which are then consumed in a shared manner by consumers, companies, etc. Private consumption is when households and companies do not share the use of finished or embodied commodities (e.g., gasoline and cars as opposed to roads). The public and private realms are of course highly integrated, especially through the use of infrastructure and the vagaries of pricing and subsidizing the use of which (e.g., water, transportation and road subsidies, taxes, and tolls) bear upon how markets set prices for resources. Increasingly, cities are seen as the key to resource conservation and therefore reduced environmental impacts. Geographers are keenly involved in the reconfiguration of land use, infrastructure provision, and governance needed to transform the city's ecological shadow (Bulkeley, 2010, 2013). The recovery, recycling, and revalorizing of resources once treated as waste (metals, paper, plastics, etc.) exemplify how in addition to demand sinks, cities have become resource generators. In association, these changing demands and attitudes toward resources have important implications for resource peripheries.

Resource Peripheries

While resources are produced and consumed everywhere, resource peripheries gain their identity from significant dependence on resources for their well-being. Across the global economy, resource peripheries (and communities) have developed mainly to exploit and export resources for consumption in major markets (Auty, 2001; Hayter and Barnes, 1990).

Export-dependent resource peripheries are typically highly vulnerable to the fluctuations of the business cycle, moving in and out of production as demand (and price) rises and falls. In this regard, resource peripheries are the geographical equivalent of the marginal unit in mainstream economics, which is the last to be added and the first to be withdrawn in responding to, and 'settling down' price changes. For resource peripheries, this vulnerability to demand changes is powerfully reinforced by additional institutional forces that impose longer term booms and busts, each of which can be 'interrupted,' even masked, by short-term business cycle effects (Edenhoffer and Hayter, 2013a). Indeed, booming and busting is a widely proclaimed characteristic of export-driven resource peripheries that can be summarily explained with reference to resource cycle dynamics that incorporate the implications of depletion, technological change, and competition (Freudenburg, 1992; Clapp, 1998). These implications are further elaborated by considerations of geopolitics, corporate strategies, and host region development aspirations.

Typically, resource peripheries feature high levels of internal and external competition. In this regard, suggestions to add value to resources, differentiate production, and develop flexible specialized structures have helped mitigate competition (Reiffenstein et al., 2002; Patchell and Hayter, 1997). However, potentials in this regard seem limited, as attempts to add value and increase prices face resistance from cores, the policies of MNCs as well as market size limitations. Collusion among peripheries to control supply and prices also seems to have limited potentials, OPEC's example notwithstanding. Rather, as indicated by the current 'rush' in several jurisdictions to build massive new LNG plants, resource peripheries seek

competitive advantages through large-scale, low unit-cost production, and by penetrating booming markets as early as possible. In turn, such strategies ensure resource cycle effects. Admittedly, competition between new low cost and old high cost peripheries is not straightforward. Mature resource regions draw upon established capital and human resources, may develop more valuable products, and think up ways to reduce costs while the opening of new resource spaces costs requires major investments in productive capital, infrastructure and the hiring, training, and housing of workers.

The ability of governments in export-dependent resource peripheries to protect producers as they mature is limited by international trading rules and domestic market size. Such protection nevertheless occurs, including by rich core countries. Moreover, rich countries may cite security as well as social concerns in support of higher cost domestic operations that replace cheaper imports. In this regard, it is interesting to wonder why the United States and Western powers allowed rare earth minerals to become concentrated in China, given their evident vital importance in contemporary industrial processes. Rare earths, usually mined in very small quantities, have vital roles in a wide range of industrial processing and sophisticated green technology applications; for example, neodymium is used in wind turbines and lithium in batteries for electronic vehicles and other products. However, this dependence is changing.

For many resource peripheries and communities, the staple trap or curse is a haunting reality, as diversification potentials are limited by remoteness from markets, specialized infrastructures, ingrained export-based resource cultures, small population base, specialized labor skills, and so on (Watkins, 1963; Gunton, 2003). Resource corporations are also often highly conservative, locked-in to established attitudes, knowledge, technologies, and missions (Edenhoffer and Hayter, 2013b). Resource wealth has provided the basis for highly diversified, rich, urbanized economies (Walker, 2001). However, within in regions such as California, British Columbia, and Quebec there are resource peripheries and communities struggling to become more resilient (Markey et al., 2012; Polèse and Shearmur, 2006). For mature, booming, and new resource towns alike, however, plans for growth or revitalization can no longer think of resources solely in industrial terms.

Remapping Control

As revealed by the core-periphery dilemma, host countries and regions are vexed by the need to control their resource industries to enhance the welfare of its people. This struggle for control has undergone a remapping over the last few decades from a dyad between state and corporation to multifaceted power struggles among a diversity of local and global stakeholders. This remapping also reflects a shift from market to nonmarket valuation of resources.

In brief, in the evolution of development thinking in the 1950s and 1960s, resource peripheries were conceptualized as 'empty lands' within the formulation of export-based 'North American' models, in comparison 'European' development grew out of established societies (Thomas, 1969). At that time, and since, within the classic market valuation dyad, the role of MNCs in resource exploitation has been the principal lightening rod for debate. Different perspectives on this role are readily provided by Canadian and Norwegian approaches to promoting the oil industry. Thus in Norway, its oil resources in the North Sea have remained under public control, it has privileged a state-owned champion, and has favored high levels of taxation to generate wealth to be saved for future generations. Indeed, Norway's Heritage Fund had reached $664 billion by 2012, mainly invested off-shore to reduce Dutch Disease (currency inflation) problems and the interest on which can fund social programs permanently. Canada's approach is more short-term, encourages ownership and control by MNCs (and now state-owned companies from elsewhere), and emphasizes low taxation levels that mainly fund government programs now. Alberta, as Canada's main oil producer, has a modest heritage fund of around $16 billion. Even so, its oil fields, now dominated by the Athabasca oil/tar sands, have served as a major boost to the provincial and national economy by providing export income, demands for a wide range of goods and services, massive job opportunities, a major source of 'transfer payments' that are sent by the federal government to poorer provinces (in 2012 this meant all provinces but Alberta, BC, and Newfoundland), and if the necessary distribution systems are built potentially providing for domestic energy self-sufficiency and/or export diversification away from the United States.

It is interesting to recall that in the formulation of export-based models in which resource endowment provided the basis of competitive advantage, nature's nonindustrial benefits were largely tangential, recognized mainly in terms of amenity values. Meanwhile, the empty land assumption meant indigenous populations could be similarly ignored in the rich country (North American) model, or treated as 'inhibiting constraints' in the context of export-driven growth in poorer countries (Hirschman, 1958). These environmental and cultural blinkers can be accepted no longer. Over the past several decades environmental and indigenous peoples' voices have become increasingly insistent and resource policies have increasingly reflected their concerns. If these voices are related they are not the same, and are highly differentiated. In addition, albeit in decline, there are still a few subsistence economies located in peripheries in which resource harvesting is locally consumed. The trends are related: conservation areas are a response to the spread of resource-based production that continues to threaten surviving subsistence cultures but more broadly indigenous groups that are drawn from these cultures. Nevertheless, across the globe, environmental and aboriginal concerns are exerting a profound influence on resource peripheries, especially in relation to the once dominant emphasis on industrialization.

Moreover, in recent decades, as an important feature of contemporary globalization, urban centers are no longer simply demanding resources from the periphery, but are expecting peripheries to conserve resources for regional, national, and even global ecological benefits (Zimmerer, 2006). These demands are vocalized through many environmental nongovernment organizations (ENGOs) such as Greenpeace and WWF. Sometimes the ENGOs act in concert with indigenous and local people, but sometimes not, and their urban idealizations of the periphery can also conflict with the expectations of local labor industry and community

interests in rich and poor countries (Hayter and Soyez, 1996; Nietschmann, 1997).

The conflicts and compromises arising from the confrontation between not only the host-corporation dyad, but also demands for indigenous and environmental rights are causing remapping of resource peripheries. This reterritorization is rooted in a redefinition of property rights away from vested industrial interests in favor of hitherto marginalized actors (Brogden and Greenburg, 2003; Clapp, 2005; Hayter, 2003). However, remapping processes are inherently normative and in practice have been marked by considerable conflict with 'resource wars,' metaphorical and real, occurring around the world. Analytically, these conflicts have been interpreted by a four-legged stakeholder model that stresses interactions among differing industrial, political, environmental, and cultural institutions (Affolderbach et al., 2012; Hayter et al., 2003) and by a related three-legged model framed by corporate, community, and government actors with ENGOs as a fourth estate (Ballard and Banks, 2003). With especial reference to ENGOs, Affolderbach (2012) interpreted forestry conflicts in terms of environmental bargaining between ENGOs and key decision-makers while Affolderbach et al. (2012) reveal the importance of 'boundary organizations' that involve scientific inputs that are acceptable to all parties, in resolving such disputes.

Within resource peripheries around the world, in rich and poor countries alike, aboriginal rights have become increasingly more high profile, symbolized on a global scale by the 'UN Declaration on the Rights of Indigenous Peoples' in 2007. While often joining with ENGOs to oppose large-scale industrialization they are not the same, especially when the latter seek to limit control by indigenous peoples over 'their' resources. Whether indigenous peoples are seeking to stop resource industrialization to protect traditional ways of life or become more significant stakeholders in resource wealth generation in several new world economies, such as Australia, Canada, and New Zealand, and developing countries such as Brazil and Indonesia resource contracts are having to take into account aboriginal concerns far more seriously than in the past.

From industry's perspective, the proliferation of resource wars are not simply a passing annoyance, even if in the 1980s this was how they were often perceived, but have encouraged attitudes and policies that are redefining the global role of resource peripheries in much broader terms than imagined in original export-based theories. This remapping is most evident in the considerable expansion of conservation areas from 2 million km in 1965 to 14 million km in 2006, much located in the Boreal forest region, along with an expansion of protected marine areas (Affolderbach et al., 2012; Zimmerer, 2006). In these conservation areas, resource supplies to industry is either stopped or greatly modified. Further, a battery of environmental legislation has required reductions in air and water pollution that have significant costs especially on existing resource industries that have been forced to modify long-established practices. While new projects can often, more cheaply incorporate state of the art technology from start-up, they must embrace costly, time-consuming environmental impact assessment in their planning, even if such procedures are highly controversial. In addition, various forms of certification that apply across industries, such as ISO 1400, and to specific industries, such as forest certification (Stringer, 2006), are providing increasingly strict vigilance on firms and supply chains that have increasing recognition among consumers.

Conclusion

Resource geography shares economic geography emphases on space–place relations, global–local dynamics, examination of the role and diversity of institutions, and evaluation of the positive and negative impacts of markets. It cannot be said, however, that resource geography informs economic geography to a great extent, a surprising phenomenon given the size of the resource sector, its indispensability and integration with the rest of the economy, and the linkage with geography's environmental concerns. The remapping of the resource sector, however, is an on-the-ground reality that declares this neglect is untenable By more fully incorporating resources into its theorizing mainstream, economic geography will not only be returning to its roots (Zimmerman, 1933), but also strengthen its explanatory power.

See also: Economic Geography; Energy, Geography of; Geographies of Food: Politics, Place, Power Reprised; Nature and Society in Geography; Organization, Geography of; Population Geography.

Bibliography

Affolderbach, J., 2011. Environmental bargains: power struggles and decision-making over British Columbia and Tasmania's old growth forests. Economic Geography 87, 181–206.

Affolderbach, J., Clapp, A., Hayter, R., 2012. Environmental bargaining and boundary organizations: remapping British Columbia's Great Bear Rainforest. Annals of the Association of American Geographers 102 (6), 1391–1408.

Auty, R.M., 2001. Resource Abundance and Economic Development. Oxford University Press, Oxford.

Bailey, I., Maresh, S., 2009. Scales and networks of neoliberal climate governance: the regulatory and territorial logics of European Union emissions trading. Transactions of the Institute of British Geographers 34 (4), 445–461.

Bailey, I., Gouldson, A., Newell, P., 2011. Ecological modernisation and the governance of carbon: a critical analysis. Antipode 43, 682–703.

Bakker, K., 2005. Neoliberalizing nature? Market environmentalism in water supply in England and Wales. Annals of the Association of American Geographers 95, 542–565.

Bakker, K., 2010a. Debating green neoliberalism: the limits of "neoliberal natures". Progress in Human Geography 34, 715–735.

Bakker, K., 2010b. Privatizing Water: Governance Failure and the World's Urban Water Crisis. Cornell University Press, Ithaca.

Ballard, C., Banks, G., 2003. Resource wars: the anthropology of mining. Annual Review of Anthropology 32, 287–313.

Boyd, E., 2009. Governing the clean development mechanism: global rhetoric versus local realities in carbon sequestration projects. Environment and Planning A.A 41, 2380–2395.

Boyd, E., Boykoff, M., Newell, P., 2011. The "new" carbon economy: what's new? Antipode 43, 601–611.

Bridge, G., 2000. Resource geography. International Encyclopedia of the Social and Behavioral Sciences, 13266–13269.

Bridge, G., 2002. Grounding globalization: the prospects and perils of linking economic processes of globalization to environmental outcomes. Economic Geography 78, 361–386.

Bridge, G., 2008. Global production networks and the extractive sector: governing resource-based development. Journal of Economic Geography 8, 389–419.

Bridge, G., Jonas, A., 2002. Governing nature: the reregulation of resource access, production, and consumption. Environment and Planning A 34, 759–766.
Brogden, M.J., Greenburg, J.B., 2003. The fight for the west: a political ecology of land use conflicts in Arizona. Human Organization 62 (3), 289–298.
Bulkeley, H., 2010. Cities and the governing of climate change. Annual Review of Environment and Resources 35, 229–253.
Bulkeley, H., 2013. Cities and Climate Change. Routledge, London.
Bumpus, A., 2011. The matter of carbon: understanding the materiality of tCO2e in carbon offsets. Antipode 43, 612–638.
Bumpus, A., Liverman, D., 2008. Accumulation by decarbonization and the governance of carbon offsets. Economic Geography 84, 127–155.
Castree, N., 2008. Neoliberalising nature: the logics of deregulation and reregulation. Environment and Planning A 40, 131–152.
Ciscell, G.M., 2010. Beyond CO2lonialism: the potential for fair trade certification to embrace voluntary carbon offsets. International Journal of Sustainable Society 2, 121–132.
Clapp, R.A., 1998. The resource cycle in forestry and fishing. The Canadian Geographer 42, 129–144.
Clapp, R.A., 2005. Wilderness ethics and political ecology; remapping the great bear rainforest. Political Geography 23, 839–862.
Coe, N.M., Dicken, P., Hess, M., 2008. Global production networks: realizing the potential. Journal of Economic Geography 8, 271–294.
Dearden, P., Mitchell, B., 2005. Environmental Change and Challenge. Oxford University Press, Toronto.
Demeritt, D., 2001. The construction of global warming and the politics of science. Annals of the Association of American Geographers 91, 307–337.
Deutz, P., Lyons, D.I., 2008. Editorial: industrial symbiosis – an environmental perspective on regional development. Regional Studies 42, 1295–1298.
Diamond, J., 1997. Guns, Germs and Steel: The Fates of Human Societies. W. W. Norton, New York.
Draper, D., 1998. Our Environment: A Canadian Perspective. ITP Nelson, Scarborough.
Edenhoffer, K., Hayter, R., 2013a. Restructuring on a vertiginous plateau: the evolutionary trajectories of British Columbia's forest industries 1980–2010. Geoforum 44, 139–151.
Edenhoffer, K., Hayter, R., 2013b. Organizational restructuring in British Columbia's forest industries 1980–2010: the survival of a dinosaur. Applied Geography 40, 222–231.
Essletzbichler, J., 2012. Renewable energy technology and path creation: a multi-scalar approach to energy transition in the UK. European Planning Studies 20 (5), 791–816.
Freudenburg, W.R., 1992. Addictive economies: extractive industries and vulnerable localities in a changing world economy. Rural Sociology 57, 305–332.
Gereffi, G., Humphrey, J., Sturgeon, T., 2005. The governance of global value chains. Review of International Political Economy 12, 78–104.
Gibbs, D., Deutz, P., Proctor, A., 2005. Industrial ecology and eco-industrial development: a new paradigm for local and regional development? Regional Studies 39, 171–183.
Gunton, T.G., 2003. Natural resources and regional development: an assessment of dependency and comparative advantage paradigms. Economic Geography 79, 67–94.
Gutiérrez, M., 2011. Making markets out of thin air: a case of capital involution. Antipode 43, 639–661.
Hanink, D.M., 2000. Resources. In: Sheppard, E., Barnes, T.G. (Eds.), A Companion to Economic Geography. Blackwell, Oxford, pp. 227–241.
Hay, A., 1976. A simple location theory for mining. Geography 71, 65–76.
Hayter, R., 2003. The war in the woods': post-fordist restructuring, globalization and the contested remapping of British Columbia's forest economy. Annals of the Association of American Geographers 93, 706–729.
Hayter, R., Barnes, T., 1990. Innis' staple theory, exports and recession: British Columbia 1981–86. Economic Geography 60, 156–173.
Hayter, R., Barnes, T., Bradshaw, M., 2003. Relocating resource peripheries to the core of economic geography's theorizing: rationale and agenda. Area 35, 15–23.
Hayter, R., Soyez, D., 1996. Clearcut issues: German environmental pressure and the British Columbia forest sector. Geographische Zeitschrift 84, 143–156.
Hirschman, O.E., 1958. The Strategy of Economic Development. Yale University Press, New Haven.
Kedron, P., Bagchi-Sen, S., 2011. A study of the renewable energy sector within Iowa. Annals of the Association of American Geographers 101, 1–15.
Knight, E., 2011. The economic geography of European carbon market trading. Journal of Economic Geography 11, 817–841.
Lahiri-Dutt, K., 2003. Informal coal mining in Eastern India: evidence from the Raniganj coalbelt. Natural Resources Forum 27, 68–77.
Le Billon, P., 2001. The political ecology of war: natural resources and armed conflict. Political Geography 20, 561–584.
Lovell, H., Mackenzie, D., 2011. Accounting for carbon: the role of accounting professional organisations in governing climate change. Antipode 43, 704–730.
Lovins, H., Lovins, A., Hawken, P., 1999. Natural Capitalism. Little, Brown and Company, New York.
Lyons, D.J., 2007. A spatial analysis of loop closing among recycling, remanufacturing, and waste treatment firms in Texas. Journal of Industrial Ecology 11, 43–54.
Markey, S., Halseth, G., Mansen, D., 2012. Investing in Place: Economic Renewal on North British Columbia. University of British Columbia Press, Vancouver.
Mather, A.S., Chapman, K., 1995. Environmental Resources. John Wiley, New York.
Matulis, B., 2013. The narrowing gap between vision and execution: neoliberalization of PES in Costa Rica. Geoforum 44, 253–260.
McAfee, K., 2012. The contradictory logic of global ecosystem services markets. Development and Change 43, 105–131.
Mitchell, B., 1989. Geography and Resource Analysis. Longman, New York.
Nietschmann, B., 1997. Protecting coral reefs and sea territories, Miskito Coast, RAAN, Nicaragua. In: Stevens (Ed.), Conservation through Cultural Survival: Indigenous Peoples and Protected Areas. Island Press, Washington, DC, pp. 193–224.
Odell, P.R., 1963. An Economic Geography of Oil. Bell, London.
O'Riordan, T. (Ed.), 2001. Globalism, Localism and Identity: Fresh Perspectives on the Transition to Sustainability. Earthscan, London.
Patchell, J., Hayter, R., 1997. Japanese precious wood: the paradox of value added. Geographical Review 87, 375–395.
Polanyi, K., 1944. The Great Transformation. Rinehart, New York.
Polèse, M., Shearmur, R., 2006. Why some regions will decline: a Canadian case study with thoughts on local development strategies. Papers in Regional Science 85, 23–46.
Prudham, W.S., 2007. Sustaining sustained yield: class, politics and post-war forest regulation in British Columbia. Environment and Planning D Society and Space 25, 258–283.
Reiffenstein, T., Hayter, R., 2006. Domestic timber auctions and flexibly specialized forestry in Japan. The Canadian Geographer 50, 503–525.
Reiffenstein, T., Hayter, R., Edgington, D.W., 2002. Crossing cultures, learning to export: making houses in British Columbia for export in Japan. Economic Geography 78, 195–220.
Robertson, M., 2011. Measurement and alienation: making a world of ecosystem services. Transactions of the Institute of British Geographers 37, 386–401.
Schlager, E., Ostrom, E., 1992. Property-rights regimes and natural resources: a conceptual analysis. Land Economics 68, 249–262.
Stringer, C., 2006. Forest certification and changing commodity chains. Journal of Economic Geography 6, 701–722.
Thomas, M.D., 1969. Regional economic growth: some conceptual aspects. Land Economics 46, 43–51.
Traffic, 2013. Wildlife Trade: What Is It?. http://www.traffic.org/trade/ (accessed 20.06.13.).
Vatn, A., 2007. Resource regimes and cooperation. Land Use Policy 24, 624–632.
Vernon, R., 1971. Sovereignty at Bay: The Multinational Spread of US Enterprises. Basic Books, New York.
Wackernagel, M., Rees, W.E., 1996. Our Ecological Foot-Print: Reducing Human Impact on the Earth. Gabriola Press New Society Publishing, Gabriola, BC.
Walker, R., 2001. California's golden road to riches: natural resources and regional capitalism, 1848–1940. Annals of the Association of American Geographers 91, 167–199.
Watkins, M., 1963. A staple theory of economic growth. The Canadian Journal of Economics and Political Science 29, 141–148.
Zalik, A., 2010. Oil 'futures': Shell's scenarios and the social constitution of the global oil market. Geoforum 41, 553–564.
Zimmerer, K. (Ed.), 2006. Globalization and New Geographies of Conservation. University of Chicago, Chicago.
Zimmerman, E.W., 1933. World Resources and Industries. Harper and Brothers, New York.

Respiratory Disorders: Psychosocial Aspects

Andreas von Leupoldt, University of Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Asthma and chronic obstructive pulmonary disease (COPD) are prevalent respiratory diseases and are associated with great individual and socioeconomic burden. Psychosocial aspects play an important role in the course and management of these diseases. This article provides an overview of important psychological and social aspects in asthma and COPD and how these are related to morbidity. In addition, respective treatment approaches that include psychosocial aspects are presented and the urgent need for future research in this area is emphasized.

Introduction

Asthma and chronic obstructive pulmonary disease (COPD) are not only the most prevalent respiratory diseases, but also belong to the most common chronic diseases in general. For example, asthma is the most common chronic disease in children whereas COPD is the fifth-leading cause of death worldwide (Global Initiative for Asthma (GINA, 2012); Global Initiative for Chronic Obstructive Lung Disease (GOLD, 2013)). Both diseases are characterized by chronic inflammatory intrapulmonary processes, which lead to reduced lung function due to bronchoconstriction, mucus production, and/or permanent remodeling and damage of lung tissues. Resulting symptoms are episodes of breathlessness, cough, and wheezing, which can be accompanied by various other comorbid symptoms, especially in COPD (GINA, 2012; GOLD, 2013). Although there is some overlap between asthma and COPD, both diseases have distinct characteristics regarding onset, pathophysiological mechanisms, and treatment. For example, whereas asthma is often diagnosed in childhood with symptoms usually being present episodically, COPD is typically diagnosed in older individuals, often after a long history of heavy smoking, with symptoms usually being chronically present. Both diseases require close monitoring of symptoms and regular medication intake including short- or long-acting bronchodilators and anti-inflammatory corticosteroids, preferably administered by inhalation. Asthma and COPD are associated with considerable individual as well as socioeconomic burden. In addition to prominent physiological symptoms, psychosocial aspects play an important role in the course and management of these diseases. However, compared to other chronic diseases, psychosocial aspects in asthma and COPD have not received the same level of systematic research activity or public interest. Therefore, this article provides an overview of important psychological and social aspects in asthma and COPD, on potential links with morbidity as well as on respective treatment approaches. Moreover, needs for future research in this area are pointed out.

Comorbid Psychological Symptoms in Asthma and COPD

Comorbid psychological symptoms, especially anxiety and depression, are highly prevalent in patients with asthma and COPD. In asthma patients, prevalence rates for anxiety and depression of up to 45 and 41%, respectively, have been reported for some inner-city samples (Feldman et al., 2005). In COPD patients, a meta-analysis revealed prevalence rates for anxiety and depression of 36 and 40%, respectively (Yohannes et al., 2000), which can reach up to 80% in specific subgroups of patients.

A substantial number of studies have demonstrated that comorbid psychological symptoms are associated with a less favorable course of asthma and COPD, respectively. This includes more frequent exacerbations, longer and more frequent hospitalizations, increased inflammatory status, persistent smoking, increased symptom burden, inappropriate medication use, worsened physical and social functioning, decreased quality of life, reduced activity levels, less favorable outcome of pulmonary rehabilitation, and even increased mortality rates (Ritz et al., 2013; von Leupoldt et al., 2012; von Leupoldt and Kenn, 2013). These negative relationships seem particularly prominent in severe and not well-controlled forms of the diseases, but are already of importance in milder forms. Not surprisingly, comorbid psychological symptoms were found to be related to increased health-care costs in patients with asthma and COPD. For example, a large retrospective cohort study estimated that the average annual all-cause cost per COPD patient was $28.961 for those with comorbid depression and anxiety compared with $22.512 for those without these comorbidities (Dalal et al., 2011). Notably, psychological comorbidities remain undetected and untreated in more than 50 and 70% of patients with asthma and COPD, respectively (Katon et al., 2006; Kunik et al., 2005). Therefore, the early detection and adequate treatment of comorbid psychological symptoms in patients with asthma and COPD remains a major clinical target requiring intensified research on the development and implementation of respective strategies.

While the exact reasons for the high prevalence of comorbid psychological symptoms in patients with asthma and COPD are largely unknown, they are most likely multifactorial with considerable interindividual variability. Some findings suggest that the high burden associated with these chronic diseases including activity limitations and/or repeated experiences with threatening episodes of breathlessness is responsible for the later development of psychological comorbidities, especially in patients who tend to overinterpret bodily sensations (Livermore et al., 2010). Longitudinal data in patients with

asthma and COPD support this view by showing that a diagnosis of asthma and COPD predicted the later onset of symptoms of anxiety or depression, respectively (Hasler et al., 2005; Schneider et al., 2010). Nevertheless, some of these studies also found that preexisting symptoms of anxiety and depression were associated with the subsequent onset of asthma or COPD (Hasler et al., 2005; Scott et al., 2008; Schneider et al., 2010), suggesting a reciprocal association between these respiratory and mental disorders. In addition, it might be speculated that a common third factor is underlying both the development of the respiratory disease and psychological symptoms, for example, smoking behavior, systemic inflammatory processes, or genetic influences (Mrazek, 2003; Van Lieshout et al., 2009; Goodwin et al., 2012). The limited availability of respective studies necessitates future multidisciplinary research activities to better understand and treat these high prevalence rates of psychological symptoms in patients with asthma and COPD.

Social Aspects in Asthma and COPD

Social aspects including cultural, environmental, and family factors have been demonstrated to greatly impact respiratory diseases, especially asthma (Ritz et al., 2013). For example, minority populations in the USA including African-Americans and Puerto Ricans have not only higher prevalence rates of asthma than those with European ancestry, but also show lower asthma control, fewer prescriptions of inhaled corticosteroids, more frequent hospitalizations, and emergency department visits as well as more fatal asthma attacks (Leong et al., 2012). Moreover, living conditions in urban environments can contribute to higher asthma morbidity and mortality due to several factors. These include increased indoor allergen exposure, air pollution, and viral infection, but also social stressors such as poverty, high crime rates, domestic violence, and disrupted family constellations (Wright and Subramanian, 2007).

Family variables play a critical role in asthma, particularly in children with asthma. Factors such as the parent–child relationship, parenting difficulties, family conflicts, and family organization have all been demonstrated to impact asthma management and clinical outcomes (Ritz et al., 2013). Family factors are already of importance during the prenatal development. For example, high maternal stress levels are associated with the later onset of asthma and atopy in the offspring (Wright, 2011).

In patients with COPD, research on social factors is still at an early stage. Some findings suggest that supportive social relationships are associated with a more favorable course of disease including lower hospital readmission rates, breathlessness, or depressive symptoms as well as improved exercise tolerance and survival rates, although conflicting results have also been reported (Kaplan and Ries, 2002; von Leupoldt et al., 2012). In addition, further aspects such as peer group attitudes and behavior, dynamics of self-evaluation, stigmatization of asthma or COPD as a disease as well as social comparison processes have been shown to affect disease outcomes (Ritz et al., 2013; Petersen et al., 2012). Notably, some studies show that not only patients with COPD but also their partners have reduced quality of life and elevated symptoms of anxiety and depression, which might interact with the patients' level of symptoms (von Leupoldt et al., 2012). These findings indicate that partners of COPD patients can be substantially affected by the disease of their ill partners, hence suggesting a need for interventions that include partners. In general, future research is needed to refine our understanding on how social aspects impact on asthma and especially on COPD and how these aspects could be used to improve their treatment.

Links between Psychosocial Aspects and Disease Processes

Both direct physiological and indirect behavioral pathways as well as their interactions can potentially link psychosocial aspects with pathophysiological processes in asthma and COPD and contribute to worse or improved course of disease.

Evidence for direct pathways comes mostly from animal and human studies on asthma, which demonstrated direct psychosocial influences on the pulmonary system, on autonomic and endocrine regulation, gene expression as well as on inflammatory and immune processes (Ritz et al., 2013). For example, negative affective states or stress lead to vagally mediated bronchoconstriction (Ritz et al., 2010) and increased airway inflammatory responses in patients with asthma (Chen et al., 2010). Elevated stress levels during pregnancy have been suggested to alter the normal development of organ systems such as immune and respiratory system and to prime the neuroendocrine, autonomic, and immune systems of the offspring toward the later development of allergic diseases including asthma (Wright, 2011). Emerging neuroimaging studies in healthy individuals illustrate that negative affect impacts on the neural processing of respiratory sensations (von Leupoldt et al., 2013), presumably via brain areas with shared relevance for both emotion processing and the processing of respiratory symptoms (e.g., insular cortex, anterior cingulate cortex, amygdala) (Davenport and Vovk, 2009). First findings suggest that patients with asthma might show disease-specific patterns in neural activation and brain structures, which might be related to disease duration, near-fatal attacks, or airway and inflammatory responses (von Leupoldt et al., 2013; Ritz et al., 2013).

Indirect pathways linking psychosocial factors and morbidity in asthma and COPD are through self-management and health behaviors. Both diseases require constant symptom monitoring, regular medication intake, and timely initialization of adequate actions in case of disease exacerbations, which necessitate various behavioral skills of the patients. Results from several studies suggest that negative clinical outcomes in asthma and COPD are related to chronically poor medication adherence, insufficient inhaler technique, reduced levels of physical activity, and unwillingness to quit smoking (Ritz et al., 2013; von Leupoldt et al., 2012; von Leupoldt and Kenn, 2013). These behaviors are greatly influenced by patients' subjective beliefs about the disease, key symptoms, and treatment (side)effects. For example, fearful overinterpretations of breathlessness and/or beliefs about effects of exercise have been shown to be associated with reduced exercise performance, lower quality of life, and higher breathlessness ratings in COPD patients (Janssens et al., 2011). The presence of more general

comorbid psychological symptoms such as anxiety and depression presumably amplifies these unfavorable associations. For example, in patients with COPD, depression was found to be related to reduced medication adherence (Khdour et al., 2012). Moreover, the subjective perception of respiratory symptoms can greatly impact on the course of respiratory diseases via subsequent health behaviors. In this regard, both over- and underperception of breathlessness have been linked to morbidity and mortality, especially in asthma, due to either exaggerated (self)treatments and activity avoidance or due to delayed and insufficient treatments (Ritz et al., 2013; von Leupoldt et al., 2012; von Leupoldt and Kenn, 2013). A variety of psychosocial aspects might further impact on respiratory symptom perception including negative affect, focus of attention, learning, context, and social variables (Janssens et al., 2009; von Leupoldt et al., 2013).

Although more refined future research is certainly required, it seems that many social aspects likely operate via indirect pathways. High levels of social support and beneficial family and environmental constellations have been suggested to be effective in absorbing distress caused by the deleterious consequences of the illness, in enhancing perceived self-efficacy, and/or in promoting positive adaptive health behaviors including treatment adherence (von Leupoldt et al., 2012). For example, a more favorable level of family routines has been shown to be associated with decreased inflammatory status in young asthma patients, which was mediated by its impact on medication use (Schreier and Chen, 2010). Notably, the various levels of the social environment may affect respiratory disease morbidity through both direct and indirect pathways in the same patients. In this regard, Chen and co-workers (Chen et al., 2007) demonstrated in the same sample of adolescent asthma patients that family factors affected the adolescents' asthma via direct physiologic changes (inflammatory status), whereas community factors operated via indirect behavioral pathways (smoking behavior). However, intensified future research efforts on both direct and indirect pathways in patients with asthma and especially in patients with COPD are necessary in order to improve our understanding of the multiple systems involved and their impact.

Interventions Including Psychosocial Aspects in Asthma and COPD

Effective management and treatment of asthma and COPD aims at several goals, including symptom relief, prevention of disease progression/inflammation, improvement of activity/exercise levels, health status, and quality of life as well as prevention and treatment of complications and exacerbations (GINA, 2012; GOLD, 2013). Patients with asthma and COPD need to be treated by a physician, preferably with specific pulmonary training. Regular intake of usually inhaled bronchodilators and anti-inflammatory corticosteroids forms the mainstay of treatment, often accompanied by further medications. However, many aspects of these conditions necessitate a more comprehensive treatment approach according to the biopsychosocial model of disease, which requires a multidisciplinary approach combining various forms of treatment. These have to be tailored to individual patient's needs and should continuously be monitored (GINA, 2012; GOLD, 2013). Ideally, asthma and COPD care should be performed by a multidisciplinary team including physicians, nurses, physiotherapists, exercise physiologists, kinesiologists, occupational therapists, dietitians, social workers, mental health specialists, behavior nurse therapists, clinical psychologists, or liaison psychiatrists. In particular, behavioral medicine approaches offer a variety of relevant techniques and interventions that can be effective in alleviating the burden of asthma and COPD by addressing psychosocial aspects and by supporting important behavioral change in patients (Ritz et al., 2013; von Leupoldt et al., 2012). These interventions are often based on elements of cognitive-behavioral therapy extended by additional disease-specific elements such as behavioral activation, skill-building exercises, cognitive restructuring, psycho-education, exposure to feared situations, problem-solving techniques, thought stopping, self-instructional training, breathing trainings as well as nonspecific relaxation and stress management techniques or biofeedback trainings. Most of the techniques listed below are part of comprehensive treatment programs (e.g., pulmonary rehabilitation) or used in various combinations, which makes it often difficult to evaluate their distinct individual effects.

Self-management and patient education programs play an important role in the treatment of asthma and COPD and are recommended core components of pulmonary rehabilitation. Ideally, the character of these programs is interactive and focused on the teaching of self-management skills, emphasizing illness control through modifications of patients' health behavior. Typical components include the pathophysiology of lung disease, proper use of medication, benefits of exercise and physical activities, irritant avoidance/smoking cessation, prevention and early treatment of exacerbations, coping with the disease, control of negative affect, and breathing strategies (Ritz et al., 2013; von Leupoldt et al., 2012). Positive effects of self-management and patient education programs have been demonstrated in patients with asthma and COPD, including improvements in hospitalizations, quality of life, symptoms, and exercise capacity. However, future randomized controlled trials with long-term follow-ups are needed to develop clear recommendations regarding the form and specific contents of these programs. An interesting alternative to conventional programs might be programs using new media such as Internet, mobile phones, and smart phones in which patients can participate from home.

Recommended pulmonary rehabilitation programs entail comprehensive interventions, which can be delivered in inpatient or outpatient settings, usually in a group format. A large body of evidence illustrates the beneficial effects of pulmonary rehabilitation in patients with asthma and COPD including improvements in exercise capacity, symptoms, quality of life, hospitalizations, and negative affect (Spruit et al., 2013). Duration and components of pulmonary rehabilitation vary between programs. In addition to monitoring and optimization of pharmacological treatments, the common elements of these programs include exercise training, self-management and patient education, behavioral training, nutrition counseling, psychosocial support, smoking cessation support, breathing therapy, and respiratory muscle training (Spruit et al., 2013; von Leupoldt et al., 2012). The optimal duration of these programs and the most effective mixture of implemented

interventions remain to be established, but some evidence suggests that longer programs compared to shorter ones are more effective. Because benefits achieved during pulmonary rehabilitation wane after completion, intensified research efforts are necessary in order to improve the maintenance of these benefits once patients have completed these programs.

In general, exercise training is recommended for most patients with asthma and COPD and underlines the need for long-term behavior change. Documented benefits include improvements in cardiopulmonary fitness, skeletal muscle function, ventilation, symptoms, and quality of life (Spruit et al., 2013). Training elements vary considerably between programs and include among others lower extremity exercise (e.g., stationary cycle ergometer, walking) and upper extremity exercise (e.g., arm cycle ergometer, free weights). Training at higher exercise intensity seems to be associated with greater physiological benefits than lower-intensity training and a combination of endurance and strength training is probably the best strategy to improve peripheral muscle function. In many patients, bronchodilator therapy is given prior to exercise to reduce exercise-induced bronchoconstriction and breathlessness and to improve exercise tolerance.

Tobacco smoke is a main risk factor for the development and/or poor prognosis of respiratory diseases, especially in COPD. However, many patients continue to smoke, rendering smoking a major target for behavioral interventions (GINA, 2012; GOLD, 2013). Smoking cessation has been shown to improve lung function, symptoms, exacerbations, and mortality rates (Scanlon et al., 2000). Several interventions are available to motivate and support patients to quit smoking, often as part of comprehensive treatment programs such as self-management and patient education or pulmonary rehabilitation. These range from low-intensity interventions (e.g., brief advice by physicians, self-help materials, telephone counseling) to more-intensive and effective interventions (e.g., group counseling) (Ritz et al., 2013; von Leupoldt et al., 2012). Recent guidelines recommend the combination of psychosocial interventions and pharmacotherapy (e.g., nicotine replacement therapy) to support smoking cessation in patients with COPD (GOLD, 2013). However, typical success rates are usually poor and underline the urgent need for improved interventions aimed at supporting smoking cessation in patients with asthma and COPD.

Respiratory muscle weakness is often observed in respiratory diseases and contributes to blood gas anomalies, breathlessness, and reduced exercise performance. Inspiratory muscle training is aimed at increasing inspiratory muscle strength, usually by having patients breathe through defined resistances. Respective trainings vary with respect to frequency, duration, and type of resistances used, but continuous training in everyday life is mandatory to maintain achieved benefits. In patients with COPD, beneficial effects include improvements in inspiratory muscle strength and endurance, functional exercise capacity, symptoms, and quality of life (von Leupoldt et al., 2012). Future well-controlled studies are needed to examine whether similar effects can be achieved in patients with asthma (Silva et al., 2013).

Breathing training includes several controlled breathing exercises aimed at improving different pathophysiological aspects of lung diseases, such as reducing lung hyperinflation and improving regional ventilation, gas exchange, and thoracoabdominal coordination. Prominent examples are pursed-lip breathing, active expiration, and body positioning, which should be tailored to individual patient's needs. Recent studies demonstrated improvements in quality of life, mood, and symptoms by some of these exercises in patients with asthma and COPD (Ritz et al., 2013; von Leupoldt et al., 2012). However, future controlled trials are necessary in order to better understand which specific variant of breathing training is effective in which subgroup of patients.

The negative effects of anxiety and depression in patients with asthma and COPD necessitate adequate treatment, for example, with psychotherapeutic interventions. These are usually performed in groups and are based on techniques of cognitive-behavioral therapy extended by additional elements. Respective interventions include – among other elements – cognitive restructuring, mindfulness-based interventions, behavioral activation, skill-building exercises, psychoeducation, and exposure to feared situations. The duration of reported interventions greatly varies ranging from a single 2-h meeting up to programs lasting several weeks. Recent studies have shown some promising anxiolytic or antidepressant effects and improvements in quality of life, functional capacity, and hospitalizations in patients with asthma and COPD (Ritz et al., 2013; von Leupoldt et al., 2012). A small number of findings suggest some positive effects of family therapy for the management of asthma in children. These family therapy programs for asthma most likely do not only address emotional aspects of the family interaction, but also more general asthma management skills and their optimal implementation in the family constellation. In general, more and larger randomized controlled trials are necessary to draw definite conclusions about the beneficial effects of certain psychotherapeutic techniques for specific subpopulations of patients with asthma and COPD (i.e., differential indication), and the required duration and frequency of these treatments. In addition, the high psychological burden in partners or family members of patients with asthma and COPD requires intensified research activities on respective interventions.

A variety of further techniques targeting psychosocial aspects have been examined in recent years in patients with asthma and COPD including symptom perception training, relaxation training, biofeedback training, yoga therapy, singing training, and dietary interventions (Ritz et al., 2013; von Leupoldt et al., 2012). However, although promising effects were observed in some studies, the overall picture is that of mixed results in too few studies, which requires further randomized controlled studies on these as well as innovative new interventions. For example, the use of new media such as the Internet and smart phones might provide valuable options for existing as well as new forms of interventions.

Summary

Various psychological and social aspects have been demonstrated to play an important role in asthma and COPD. Both direct physiological and indirect behavioral pathways link psychosocial aspects with pathophysiological processes in these respiratory diseases. However, there is a considerable

need for further multidisciplinary research on these mechanisms and their specific impact on the course and management of asthma and COPD. Several interventions in asthma and COPD include psychosocial elements. While the beneficial effects of some interventions such as pulmonary rehabilitation and self-management programs are already established, distinct components of these interventions need further elaboration. Preliminary findings on other interventions are often encouraging, but require future randomized controlled studies in order to recommend their regular use. Moreover, additional research on the optimal transfer of beneficial treatment effects in clinical settings into long-lasting improvements in the everyday life of patients with asthma and COPD is necessary. In this context, the use of new media might be valuable, but requires further evaluation.

See also: Behavioral Medicine; Caregiver Burden; Chronic Illness, Psychosocial Coping with; Co-morbidity of Mental and Physical Conditions; Mindfulness; Personalized Medicine; Race: Ethnicity and Health; Rehabilitation (Cardiac): Psychosocial Aspects; Smoking and Health; Socioeconomic Status and Health; Symptom Perception, Awareness and Interpretation.

Bibliography

Chen, E., et al., 2007. The role of the social environment in children and adolescents with asthma. American Journal of Respiratory and Critical Care Medicine 176 (7), 644–649.

Chen, E., et al., 2010. Socioeconomic status associated with exhaled nitric oxide responses to acute stress in children with asthma. Brain Behavior and Immunity 24, 444–450. http://dx.doi.org/10.1016/j.bbi.2009.11.017.

Dalal, A.A., et al., 2011. Clinical and economic burden of depression/anxiety in chronic obstructive pulmonary disease patients within a managed care population. COPD 8 (4), 293–299.

Davenport, P.W., Vovk, A., 2009. Cortical and subcortical central neural pathways in respiratory sensations. Respiratory Physiology & Neurobiology 167 (1), 72–86.

Feldman, J.M., et al., 2005. Psychiatric disorders and asthma outcomes among high-risk inner-city patients. Psychosomatic Medicine 67 (6), 989–996.

Global Initiative for Asthma (GINA), 2012. Global Strategy for Asthma Management and Prevention. Retrieved from: http://www.ginasthma.org/.

Global Initiative for Chronic Obstructive Lung Disease (GOLD), 2013. Global Strategy for Diagnosis, Management, and Prevention of COPD. Retrieved from: http://www.goldcopd.com.

Goodwin, R.D., et al., 2012. Depression, anxiety, and COPD: the unexamined role of nicotine dependence. Nicotine and Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco 14 (2), 176–183.

Hasler, G., et al., 2005. Asthma and panic in young adults: a 20-year prospective community study. American Journal of Respiratory and Critical Care Medicine 171 (11), 1224–1230.

Janssens, T., et al., 2011. Dyspnea perception in COPD: association between anxiety, dyspnea-related fear, and dyspnea in a pulmonary rehabilitation program. Chest 140 (3), 618–625.

Janssens, T., et al., 2009. Inaccurate perception of asthma symptoms: a cognitive-affective framework and implications for asthma treatment. Clinical Psychology Review 29 (4), 317–327.

von Leupoldt, A., et al., 2012. Behavioral medicine approaches to chronic obstructive pulmonary disease. Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine 44 (1), 52–65.

von Leupoldt, A., et al., 2013. Emotions and neural processing of respiratory sensations investigated with respiratory-related evoked potentials. Psychosomatic Medicine 75 (3), 244–252.

von Leupoldt, A., Kenn, K., 2013. The psychology of chronic obstructive pulmonary disease. Current Opinion in Psychiatry 26 (5), 458–463.

Kaplan, R.M., Ries, A.L., 2002. Chronic obstructive pulmonary disease: behavioural assessment and treatment. In: Respiratory Disorders and Behavioral Medicine. Martin Dunitz Ltd., London, pp. 85–116.

Katon, W.J., et al., 2006. Quality of mental health care for youth with asthma and comorbid anxiety and depression. Medical Care 44 (12), 1064–1072.

Khdour, M.R., et al., 2012. Potential risk factors for medication non-adherence in patients with chronic obstructive pulmonary disease (COPD). European Journal of Clinical Pharmacology 68 (10), 1365–1373.

Kunik, M.E., et al., 2005. Surprisingly high prevalence of anxiety and depression in chronic breathing disorders. Chest 127 (4), 1205–1211.

Leong, A.B., Ramsey, C.D., Celedón, J.C., 2012. The challenge of asthma in minority populations. Clinical Reviews in Allergy and Immunology 43 (1–2), 156–183.

Livermore, N., Sharpe, L., McKenzie, D., 2010. Panic attacks and panic disorder in chronic obstructive pulmonary disease: a cognitive behavioral perspective. Respiratory Medicine 104 (9), 1246–1253.

Mrazek, D.A., 2003. Psychiatric symptoms in patients with asthma causality, comorbidity, or shared genetic etiology. Child and Adolescent Psychiatric Clinics of North America 12 (3), 459–471.

Petersen, S., et al., 2012. Social comparison and anxious mood in pulmonary rehabilitation: the role of cognitive focus. British Journal of Health Psychology 17 (3), 463–476.

Ritz, T., et al., 2010. Airway response to emotional stimuli in asthma: the role of the cholinergic pathway. Journal of Applied Physiology 108 (6), 1542–1549.

Ritz, T., et al., 2013. Psychosocial factors and behavioral medicine interventions in asthma. Journal of Consulting and Clinical Psychology 81 (2), 231–250.

Scanlon, P.D., et al., 2000. Smoking cessation and lung function in mild-to-moderate chronic obstructive pulmonary disease. The lung health study. American Journal of Respiratory and Critical Care Medicine 161 (2 Pt 1), 381–390.

Schneider, C., et al., 2010. COPD and the risk of depression. Chest 137 (2), 341–347.

Schreier, H.M.C., Chen, E., 2010. Longitudinal relationships between family routines and biological profiles among youth with asthma. Health Psychology 29 (1), 82–90.

Scott, K.M., et al., 2008. Childhood adversity, early-onset depressive/anxiety disorders, and adult-onset asthma. Psychosomatic Medicine 70 (9), 1035–1043.

Silva, I.S., et al., 2013. Inspiratory muscle training for asthma. The Cochrane Database of Systematic Reviews 9, CD003792.

Spruit, M.A., et al., 2013. An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary rehabilitation. American Journal of Respiratory and Critical Care Medicine 188 (8), e13–64.

Van Lieshout, R.J., Bienenstock, J., MacQueen, G.M., 2009. A review of candidate pathways underlying the association between asthma and major depressive disorder. Psychosomatic Medicine 71 (2), 187–195.

Wright, R.J., 2011. Epidemiology of stress and asthma: from constricting communities and fragile families to epigenetics. Immunology and Allergy Clinics of North America 31 (1), 19–39.

Wright, R.J., Subramanian, S.V., 2007. Advancing a multilevel framework for epidemiologic research on asthma disparities. Chest 132 (Suppl. 5), 757S–769S.

Yohannes, A.M., Baldwin, R.C., Connolly, M.J., 2000. Mood disorders in elderly patients with chronic obstructive pulmonary disease. Reviews in Clinical Gerontology 10 (02), 193–202.

Relevant Websites

http://www.ginasthma.org/ – Global Initiative for Asthma.
http://www.goldcopd.org/ – Global Initiative for Chronic Obstructive Lung Disease.

Responsibility in Social Work

Viviene E Cree, The University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Responsibility is central to social work wherever it is practiced in the world, and yet it is a concept that is most often taken-for-granted, rather than afforded serious examination. This article argues that the idea of responsibility must be interrogated both vertically and horizontally, if we are to make sense of the complex dilemmas and contradictions at the heart of social work practice. The article concludes that social workers have particular responsibility to themselves, to others, and to the world in which people live by virtue of social work's unique role.

Introduction

A child cries, alone and hungry. An older person collapses in the street. A convicted sex offender commits suicide. Who is responsible? Social workers are the professional group most likely to be pilloried for doing too much or doing too little, accused of not dealing adequately with risk and for failing to prevent harm, "damned if they do and damned if they don't" (Cree, 2013). And yet the task of social worker is to take responsibility for those who are unable or unwilling to take care of themselves. Social workers are located between the individual and society, negotiating and at times, in conflict with both (Cree, 1995). This suggests that social work is, then, by its very nature, a 'responsible' profession.

This article will begin by offering a dictionary definition of responsibility, arguing that this must be understood in its historical, social, and cultural context. Responsibility is not a concept that is agreed by everyone across time. Instead, accepted ideas of what it is to be responsible have changed over the years and are currently understood differently in developing and developed world contexts. This does not mean that it has no specific meaning for social work. On the contrary, social work relies on a particular interpretation of responsibility that is enshrined in legislation, in its international definition, and its codes of ethics. These will be explored as this article unfolds.

The article will first introduce different types of responsibility vertically, exploring, in turn, ideas of individual, family, social, and collective responsibility. It will then examine responsibility horizontally, giving attention to moral and professional responsibility, that is, the kinds of responsibility that run throughout social work practice wherever it is located. All ways of thinking about responsibility reflect crosscutting themes of rights, risk, duty, and blame. These come together in a final discussion that turns to conclusions and implications for the future.

Defining Responsibility

The Oxford English Dictionary (2012) definition of responsibility is as follows:

1. having a duty to deal with something or of having control over someone;
2. being accountable or blamed for something;
3. the opportunity or ability to act independently and take decisions without authorization.

This definition throws up a number of key ideas, including 'duty,' 'control,' 'blame,' 'independence.' All are central to a wider examination of responsibility, both vertically and horizontally, and all are open to negotiation and likely to change over time and in different contexts and countries.

Vertical Strands

Individual Responsibility

At its most simple, responsibility is concerned with individual or personal responsibility: individuals are responsible for themselves, their successes and failures. People also may hold responsibility for the action (or inaction) of others who, by virtue of their age and vulnerability, need another person to take some responsibility for them (including very young and very old family members, those who are sick or disabled and in need of care).

This understanding of responsibility is filled with historical resonance. From the seventeenth century onward, ideas of individualism and liberalism were at the heart of political, economic, and social life in the then industrializing world. 'Laissez-faire' principles maintained that the market should be free to regulate itself; that government intervention in economics and in society would do more harm than good. At the same time, it was believed that anything was possible for those who worked hard and showed initiative; 'heaven helps those who help themselves' was the familiar mantra of the day. Self-help was seen as the best way forward, and any assistance given to needy individuals should be that which encouraged independence and individual responsibility, not dependence and sloth (Fraser, 2009). The idea of individual responsibility is illustrated in the workings of the UK Poor Law (Cree and Myers, 2008). Those in need were classified into two groups: the 'deserving' and 'undeserving' poor; the deserving poor were those who were prepared to take responsibility and act to improve their situation, while the undeserving poor were characterized as idle, feckless, and irresponsible. When the Charity Organisation Society was formed toward the end of the nineteenth century, the principles of 'scientific charity' again highlighted the value of individual responsibility (Seed, 1973: 17).

A very similar presentation of responsibility is found across the world, demonstrated in the popular slogan, "If you give a man a fish you feed him for a day. If you teach a man

to fish you feed him for a lifetime" (Kiewra, 2010: 78). Encouraging individual responsibility is not, however, always positive, for the individual or society. Individualism may bring with it selfishness, acquisitiveness, and a lack of concern for others. It may also encourage a culture of blame. Webb points out that "although we are supposed to increasingly take responsibility for our own actions, there is plenty of evidence suggesting we wish to point the finger of blame elsewhere" (2006: 12). He cites obesity as an example of this; people prefer to blame the 'fast food' industry for our obese society, rather than confront the more difficult issue of individual choices in eating habits. They are also quick to deny any personal responsibility for 'bad' things such as atrocities or torture. Individuals often prefer to acquiesce, obey, and deny responsibility, rather than face up to the challenge of expressing disagreement with others (Cohen, 2000).

Ideas of individual responsibility infuse two current strands within social work: the idea of 'personalization' and the notion of 'responsivity.' Personalization is a policy initiative that emerged within social care. The not-so-radical idea is that people should be responsible for their own support services, able to determine and manage their own care, rather than fitting in with existing models of service provision. This, of course, reflects notions of individualism discussed above. It also owes allegiance to the idea of 'client self-determination,' which came out of casework practice in the 1940s and 1950s (Biestek, 1951) as well as pointing to current ideas about individual rights, illustrated in both Human Rights and Disability Discrimination legislation. Personalization has been viewed by some commentators (for example, Ferguson, 2007; Lymbery, 2012) as a double-edged sword: although it is, in principle, a positive development for service users, personalization moves responsibility for care from service providers to service users, allowing the state to abdicate responsibility and, at the same time, transferring any risks associated with this; the market, meanwhile, is free to dominate the supply of care services (Cree, 2009).

Responsivity also draws on ideas of individual responsibility. The term first emerged in criminal justice literature in North America (Andrews et al., 1990). The responsivity principle suggests that the delivery of programmes of intervention should be matched to the characteristics of offenders and their needs; the relationship between the offender and his/her social worker or probation officer is identified as central to this (Whyte, 2009: 62). Proponents argue that variations in reoffending can be related principally to risk, need, and responsivity (Andrews et al., 1990). Critics, however, assert that this is oversimplistic; that programmes must take account of social and cultural contexts and must build on offenders' strengths, rather than focusing only on their weaknesses and problems (Whyte, 2009). This is a reminder that focusing on the individual and on individual responsibility is not enough; rather that issues of society and structure remain of major importance.

Family Responsibility

An idea that is as old as the idea of individual responsibility is that of family responsibility: the family is here held to be the first port of call for anyone in need. An illustration is found in the early Elizabethan Poor Laws in the UK, which stated:

> It should be the duty of the father, grandfather, mother, grandmother, husband or child of a poor, old, blind, lame or impotent person, or other poor person, not able to work, if possessed of sufficient means, to relieve and maintain that person.

When people later came together to campaign for the establishment of welfare services in the UK in the early years of the twentieth century, critics argued that such provision would undermine family responsibility and make it too easy for people to abdicate their responsibilities for each other (Fraser, 2009), thus demonstrating again the close connection between ideas of individual and family responsibility.

The principle of the paramountcy of family responsibility is one that runs through all social welfare today, wherever it exists in the world. Charters that uphold individual rights are commonly prefaced by ideas of family duty. For example, the UN Convention of the Rights of the Child (1989) and the African Charter on the Rights and Welfare of the Child (1990) both state that parents have primary responsibility for the upbringing of their children, although, interestingly, the African charter suggests that children have responsibilities too, to their families and communities:

> Every child shall have responsibilities towards his family and society, the State and other legally recognized communities and the international community (Article 31).

The idea of family responsibility is also central to world religions. For example, an Islamic prophet wrote: "Every one of you (people) is a shepherd. And every one is responsible for whatever falls under his responsibility. A man is like a shepherd of his own family, and he is responsible for them."

The issue of family responsibility is one that, traditionally, has had particular meaning for women. Numerous feminist studies from the 1970s onward pointed out that taking responsibility for family members usually falls to women. Graham (1983) went so far as to argue that caring was a fundamental part of women's self-identity. More recently, it has been acknowledged that issues such as 'race'/ethnicity and social class are also significant; it is often the labor of black and working-class women that allows middle-class parents to work outside the home (Graham, 2007). Research has also drawn attention to the reality that family responsibility is not, in practice, held only by parents. Grandparents and young people are also strongly implicated in family care (Gray, 2005; Cass, 2007).

Radical voices in social work in the twenty-first century have returned to the potential for misuse of the notion of family responsibility. It is argued that foregrounding family responsibility as a policy initiative allows the state and professionals to blame victims for their plight, meanwhile the structural failings of society remain obscured (see Ferguson and Woodward, 2009).

Social Responsibility

The notion of social responsibility, like that of individual and family responsibility, is an ancient one. Within the

Judeo-Christian tradition, the idea that human beings are responsible for each other is found in the parable of the "Good Samaritan" (Luke 10: 29–37). This tells the story of a traveler who is beaten and robbed; the person who rescues him is not the priest or a member of his own social group but a Samaritan (Samaritans were despised by Jewish people). The parable draws us to reflect on "who is my neighbour?" The Qu'ran similarly enjoins people to look after the weak and vulnerable or else expect condemnation on the day of judgment. It asks: "Have you seen him who denies the Day of Judgement? That is he who repulses the orphan harshly and urges not on the feeding of the poor" (Qu'ran 107: 1–3).

The idea of social responsibility has taken practical form in different ways at different times. For example, Thomas Chalmers, a local clergyman, established a welfare scheme in the 1820s in the East End of Glasgow, Scotland. He withdrew his parish (local community) from statutory poor law measures and instead, instigated a system of support delivered for, and by, other community members. He believed that by removing external (state) aid, community members would take responsibility for themselves and each other. Chalmers's scheme was a source of inspiration to the Charity Organisation Society that emerged at the end of the nineteenth century in the UK and the US (Fraser, 2009). It was also demonstrated in the UK Prime Minister, David Cameron's vision of a 'big society':

> We want to give citizens, communities and local government the power and information they need to come together, solve the problems they face and build the Britain they want. We want society – the families, networks, neighbourhoods and communities that form the fabric of so much of our everyday lives – to be bigger and stronger than ever before. Only when people and communities are given more power and take more responsibility can we achieve fairness and opportunity for all.
>
> 2010: 1

This is, of course, the only one way of thinking about social responsibility. The so-called 'postwar consensus' in the UK after the end of the Second World War illustrated a very different conceptualization. According to this, it was generally accepted that social responsibility could be relegated to the state; a new understanding promised that living standards could be improved for all through the state taking responsibility for key aspects of social life including housing, unemployment, and health. Social services meanwhile would provide a 'safety net' for those unable to take advantage of the mainstream services. There was an expansion of statutory social work provision from 1945 onward and the strengthening of social work as a profession. At the same time, voluntary agencies came to be seen as an 'add-on,' dependent on statutory agencies to pay for the services they provided (Cree, 1995; Parry and Parry, 1979). The UK today has a mixed economy of care; social responsibility is shared among the state, the voluntary sector, private providers, and individuals and families working alongside one another. There is no shared view about this (if there ever was one), and instead, arguments for and against the provision of social responsibility bubble underneath the surface all the time, erupting at key moments such as parliamentary elections and at times of social unrest and economic recession.

Other parts of the developed world such as Scandinavia have chosen a very different path toward social responsibility. Here, members of society have been willing to pay higher taxes so that there is a more generous provision of social services for all. This development has not, however, been welcomed unreservedly by all. For example, Wolfe (1989a,b) argues that a consequence of Scandinavian welfare states' high levels of responsibility for others is that individuals no longer accept personal responsibility for others; their understanding of moral responsibility, he asserts, has diminished. This viewpoint shows again the contentiousness of ideas of social responsibility, as well as the crossover with notions of moral responsibility, to be discussed more fully below.

Collective Responsibility

The idea of collective responsibility is very close to that of social responsibility, but in this case, responsibility is envisaged as held by the whole, not just by a particular social group within society. From this perspective, it might be said that a government or local authority or private sector agency is responsible, for example, for a collective failure to protect and care adequately for vulnerable people. Social workers working in a particular agency may hold individual, moral, and professional responsibility for such a failure in practice (see Ife, 2012).

Collective responsibility is not just limited to organizational responsibility, however. Writing about 'green' social work, Dominelli (2012) draws attention to social work's commitment as a profession that claims to intervene to enhance people's well-being. The international definition of social work also highlights this:

> The social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilising theories of human behaviour and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work.
>
> Adopted by the IFSW General Meeting in Montréal, Canada, July 2000, IFSW 2013

In spite of this broad claim to action, Dominelli points out that social workers have played little more than a low-key role in environmental issues that increasingly impact on people's well-being, both locally and globally, although social workers across the world continue to play a key role in the aftermath of environmental disasters, whether caused by climate change, industrial accidents, or human conflict. Shaw similarly asserts that social work must wake up to its collective responsibilities. She suggests that by failing to engage with ecological issues, our current social policies are "at best not sustainable, and at worst dangerous for our continued social well-being" (2011: 1). Social workers, she advises, have an important role to play because of social work's understanding of the interrelationship between people and the environment, and social workers' ability to integrate environmental issues into their social work practice, and advocate for vulnerable populations. This leads us readily into a discussion of moral responsibility.

Horizontal Strands

Moral Responsibility

All types of responsibility have a moral dimension. Taking this further, moral responsibility should be understood to be the responsibility that underpins all others. Moral and ethical obligations make us responsible for ourselves, for others, for society, and for the planet.

Unsurprisingly, we find discussion of moral responsibility going back centuries. In the Bible, in Genesis 4: 1–16, after killing his brother Abel, Cain angrily asks God, "Am I my brother's keeper?" (The story of Cain and Abel is also found in Islamic tradition in the Qu'ran.) Cain's question has come to signify people's reluctance to accept responsibility for the welfare of their fellow human beings, that is, their 'brothers,' in the broader meaning of the term. As already highlighted, scholars such as Wolfe have argued that modern liberal democracies, such as the US and Scandinavia, have 'broken with traditional sources of morality,' relying instead on 'economic and political frameworks to define their obligations to one another.' Wolfe calls for a sense of community to be reinvigorated and with it, "a sense of obligation to the larger society" (1989b: 374). Bauman is similarly critical of what he sees as a 'modernist' approach to ethics, in which individual responsibility is exchanged for legislation, rules, and codes of behavior. He turns to the work of the philosopher Levinas (1985), who argues that ethics must begin with 'the Other,' with the knowledge that we can never fully understand or explain another person. Picking up this idea, Bauman asserts that moral relationship comes before 'being.' He writes, "I am a moral person because I recognize that dependence and accept the responsibility that follows …. My brother's dependence is what makes me an ethical being" (2000: 5). The first step, he argues in an earlier book, is to "take responsibility for one's responsibility" (1993: 185).

This draws us into literature on care ethics, and the work of Joan Tronto. Building on the insights of Gilligan (1982), Tronto points out that care is a practice, not a set of rules and principles; "It involves both particular acts of caring and a 'general habit of mind' to care that should inform all aspects of a practitioner's life" (1994: 127). Taking the example of the concept of protection, Tronto points out that the whole idea of protection demonstrates a very specific approach to human relationships; it assumes 'bad intentions,' in comparison with the idea of caring, which involves "taking the concerns and needs of the other as the basis for action" (1994: 104). Tronto observes that 'care-giving' is not necessarily the same as 'taking care of' someone. An individual or service may fulfill the first obligation but not the second. She identifies four ethical elements in care: attentiveness, responsibility, competence, and responsiveness, where responsibility is understood as responsibility to act and responsibility to take care. Tronto notes that parents are responsible for the care of their children, but responsibility may go far beyond this, as exemplified in the behavior of some Europeans during the Second World War when they tried to rescue Jews and others from Nazis. Tronto reminds us that assumptions of responsibility may be culturally and socially specific, impacted by gender, class, and family status. She proposes that we adopt a flexible approach to responsibility, rather than one that is rule bound and fixed for all time.

Professional Responsibility

Professional responsibility, along with legal and statutory responsibility, may be viewed as the other side of the coin to moral responsibility. While moral responsibility begins with a series of questions and is open to negotiation, professional responsibility tends to be presented as a set of answers. It is suggested that if practitioners meet their professional/legal/statutory obligations under legislation, practice guidelines, and given codes of practice, they will be able to sleep easy in their beds, sure in the knowledge that they have met their professional responsibilities to those with whom they are working.

Ethical codes of practice were first introduced by social work's newly formed regulatory bodies (the Northern Ireland and Scottish Social Services Council (SSSC) in Scotland and what was then the General Social Care Council (GSCC) in England and Wales) in the UK in 2002. Other countries have similar codes, as do the International Association of Schools of Social Work and the International Federation of Social Work, which jointly produced and agreed an ethical code in 2004 (see http://ifsw.org/policies/statement-of-ethical-principles/).

The 2002 UK codes of practice encompassed two sets of rules and regulations: one for social service employers, the other for social service workers. The Code of Practice for Social Service Employers set out the responsibilities of employers in the regulation of social service workers; significantly, this code of practice has not been retained by the body that replaced the General Social Care Council (GSCC) in England, that is, the Health and Care Professions Council (see http://www.hpc-uk.org). The Employers' Code in Scotland is still in place. It states that "employers are responsible for making sure that they meet the standards set out in Codes of Practice, provide high quality services and promote public trust and confidence in social services. To meet their responsibilities in relation to regulating the social service workforce, social service employers must:

- make sure people are suitable to enter the workforce and understand their roles and responsibilities
- have written policies and procedures in place to enable social service workers to meet the Scottish Social Services Council (SSSC) Code of Practice for Social Service Workers
- provide training and development opportunities to enable social service workers to strengthen and develop their skills and knowledge
- put in place and implement written processes and procedures to deal with dangerous, discriminatory, or exploitative behaviour and practice
- promote the SSSC's Code of Practice to social service workers, service users, and carers and co-operate with SSSC's proceedings" (see http://www.sssc.uk.com/Codes-of-Practice/sssc-codes-of-practice-for-social-service-workers-and-employers.html).

The Code of Practice for Social Service Workers describes the standards of professional conduct and practice required of social workers themselves. It forms part of the wider package of legislation, practice standards, and employers' policies and procedures that social service workers must meet. Social service

workers are held to be "responsible for making sure that their conduct does not fall below the standards set out and that no action or omission on their part harms the well-being of service users. Social service workers must:

- protect the rights and promote the interests of service users and carers
- strive to establish and maintain the trust and confidence of service users and carers
- promote the independence of service users while protecting them as far as possible from danger or harm
- respect the rights of service users while seeking to ensure that their behaviour does not harm themselves or other people
- uphold public trust and confidence in social services
- be accountable for the quality of their work and take responsibility for maintaining and improving their knowledge and skills" (see http://www.sssc.uk.com/Codes-of-Practice/sssc-codes-of-practice-for-social-service-workers-and-employers.html).

Parallel to the regulatory bodies' codes of practice, British Association of Social Workers has published its own *Code of Ethics for Social Work*, revised in 2012. This firstly affirms social work's "respect for the inherent worth and dignity of all people as expressed in the United Nations Universal Declaration of Human Rights (1948) and other related UN declarations on rights and the conventions derived from those declarations." It goes on to state that "social workers have a responsibility to promote social justice, in relation to society generally, and in relation to the people with whom they work," as well as "a responsibility to respect and uphold the values and principles of the profession and act in a reliable, honest and trustworthy manner." The fourth injuncture suggests that "social workers have a responsibility to apply the professional values and principles set out above to their practice. They should act with integrity and treat people with compassion, empathy and care." Two principles for ethical practice are of particular relevance to a discussion of professional responsibility: (14) "Taking responsibility for their own practice and continuing professional development," and (16) "Taking responsibility for the professional development of others."

There is nothing in any of the codes of practice with which we might wish to disagree. For the most part, the codes reflect current ideas of good practice, and represent a fairly broad consensus about what social work is and might be. The codes perform a useful function, in that they set down minimal expectations and obligations about the care individuals may expect to receive from agencies and social workers themselves. The question remains, however, about how useful these really are in the messy, 'real world' of practice; in what Schön (1983) has called the 'swampy lowlands' of professional practice, where competing ideas of responsibility may clash, and where a social worker may have to determine their own path, while coping at the same time with inadequate information and contradictory evidence. The reality is that ethical dilemmas remain, however clearly defined policies and procedures are. It is because of this that Bauman and Tronto's demand for the primacy of moral responsibility comes to the fore.

Conclusion and Implications for the Future

This article has explored a range of interpretations of the concept of responsibility. It has acknowledged that, while remaining a fundamental idea at the heart of social work, it is nevertheless a contested term, and one that benefits from reflection and analysis. Through this process, we can see that individualistic, narrowly defined, rights-based notions of responsibility may work against the very principles on which the profession of social work claims to be based. If practitioners were to truly embrace ideas of human rights and social justice, then this will only be through engagement with the reality that "I am my brother's keeper." Social workers have responsibility, as individuals, family members, members of society, professionals and, ultimately, as human beings, to act; and that must be done.

See also: Antisocial Behavior; Social Work Theory.

Bibliography

Andrews, D.A., Bonta, J., Hoge, R.D., 1990. Classification for effective rehabilitation: rediscovering psychology. Criminal Justice and Behavior 17, 19–52.
Bauman, Z., 1993. Postmodern Ethics. Blackwell, Oxford.
Bauman, Z., 2000. Am I my brother's keeper? European Journal of Social Work, Special Essay 3 (1), 5–11.
Biestek, F., 1951. The Principle of Client Self-determination in Social Casework. Catholic University of America Press, Washington, DC.
Cass, B., 2007. Exploring social care: applying a new construct to young carers and grandparent carers. Australian Journal of Social Issues 42 (2), 241–254.
Cohen, S., 2000. States of Denial: Knowing about Atrocities and Suffering. Polity Press, Cambridge.
Cree, V.E., 1995. From Public Streets to Private Lives. The Changing Task of Social Work. Avebury, Aldershot.
Cree, V.E., 2009. The changing nature of social work. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Social Work. Themes, Issues and Critical Debates, third ed. Palgrave, Macmillan, Basingstoke.
Cree, V.E., 2013. New practices in empowerment. In: Gray, M., Webb, S. (Eds.), New Politics of Critical Social Work. Palgrave Macmillan, Basingstoke.
Cree, V.E., Myers, S., 2008. Social Work: Making a Difference. Policy Press/BASW, Bristol.
Dominelli, L., 2012. Green social Work: From Environmental Crises to Environmental Justice. Wiley, London.
Ferguson, I., 2007. Increasing user choice or privatizing risk? the antinomies of personalization. British Journal of Social Work 37 (3), 387–403.
Ferguson, I., Woodward, R., 2009. Radical Social Work in Practice: Making a Difference. Policy Press/BASW, Bristol.
Fraser, D., 2009. The Evolution of the British Welfare State: A History of Social Policy since the Industrial Revolution, fourth ed. Palgrave Macmillan, Basingstoke.
Gilligan, C., 1982. In a Different Voice, Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.
Gladstone, D., 1999. The Twentieth-Century Welfare State. Palgrave Macmillan, Basingstoke.
Graham, H., 1983. Caring: a labour of love. In: Finch, J., Groves, D. (Eds.), A Labour of Love. Women, Work and Caring. Routledge and Kegan Paul, London, pp. 13–30.
Graham, H., 2007. Unequal Lives: Health and Social Inequalities. Open University Press, Maidenhead.
Gray, A., 2005. The changing availability of grandparents as carers and its implications for childcare policy in the UK. Journal of Social Policy 34 (4), 557–577.
Ife, J., 2012. Human Rights and Social Work: Towards Rights-Based Practice, third ed. Cambridge University Press, Cambridge.
Jordan, B., 2012. Making sense of the 'big society': social work and the moral order. Journal of Social Work 12 (6), 630–646.
Kiewra, K.A., 2010. How classroom teachers can help students learn and teach them how to learn. Theory Into Practice 41 (2), 71–80.

Levinas, E., 1985. Ethics and Infinity. Conversations with Philippe Nemo (R.A. Cohen, Trans.). Duquesne University Press, Pittsburgh.
Lymbery, N., 2012. Social work and personalisation. British Journal of Social Work 42 (4), 783–792. http://dx.doi.org/10.1093/bjsw/bcs027.
Oxford English Dictionary, 2012. Oxford University Press, Oxford.
Parry, N., Parry, J., 1979. Social work professionalism and the state. In: Parry, N., Rustin, M., Satyamurti, C. (Eds.), Social Work, Welfare and the State. Edward Arnold, London, pp. 21–47.
Schön, D.A., 1983. The Reflective Practitioner: How Professionals Think in Action. Basic Books, New York.
Seed, P., 1973. The Expansion of Social Work in Britain. Routledge and Kegan Paul, London.
Shaw, T.V., 2013. Is social work a green profession? an examination of environmental beliefs. Journal of Social Work, 1–27. http://dx.doi.org/10.1177/1468017311407555 published online 2 June 2011.
Tronto, J., 1994. Moral Boundaries. A Political Argument for an Ethic of Care. Routledge, London.
Webb, S., 2006. Social Work in a Risk Society, Social and Political Perspectives. Palgrave Macmillan, Basingstoke.
Whyte, B., 2009. Youth Justice in Practice. Making a Difference. Policy Press/BASW, Bristol.
Wolfe, A., 1989a. Welfare state and moral obligation: the case of Scandinavia. Politica 21 (2), 149–164.
Wolfe, A., 1989b. Whose Keeper? Social Science and Moral Obligation. University of California Press, Berkeley, CA.

Relevant Websites

http://cdn.basw.co.uk/upload/basw_112315-7.pdf – BASW. 2012. Code of Ethics for Social Work.
http://ifsw.org/policies/definition-of-social-work/; http://ifsw.org/policies/statement-of-ethical-principles/ – IFSW. 2013. International Definition of Social Work.
http://www.sssc.uk.com/about-the-sssc/codes-of-practice/what-are-the-codes-of-practice – Scottish Social Services Council.

Responsibility: Philosophical Aspects

Marina AL Oshana, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Three analyses of the judgment that a person is a morally responsible agent are discussed. One is the reactive-attitudes analysis, whereby a person is responsible when the person is liable to a response on the part of others marked by attitudes such as gratitude and resentment. A second analysis focuses on the phenomenon of holding an agent answerable for his or her actions. The third analysis holds that judgments of responsibility are essentially statements of attribution, which focus on the ethical appraisal of conduct that can be imputed to an individual as the author or adopter of ends.

The Concept of Responsibility

The focus of this article is moral responsibility. Although the article begins with an examination of the concepts of causal responsibility and of responsibility as duty, the discussion in these sections is intended to orient the reader to the distinctive character of the concept of moral responsibility and to related philosophical ideas about what is meant when moral responsibility is ascribed to a human agent for some act or state of affairs. Attributing responsibility to an agent as to a cause is logically and conceptually distinct from claiming that the person is morally responsible for an event. Equally, discerning those tasks that a person has a responsibility or a duty to perform is different from claiming that the person bears moral responsibility for the performance of those tasks. In short, the *concept* of responsibility has a number of distinct applications, the analyses of which point to the variety of *conceptions* of responsibility.

Causal Responsibility

One interested in action theory may investigate causal responsibility. Thus, in claiming "Alice is responsible for spilling the coffee on the floor," we may be saying that Alice spilled the coffee as a result of some action on her part. In a causal context, responsibility ascriptions credit an entity with a role in bringing about an event or state of affairs, and nothing more. Assigning causal responsibility consists in identifying a salient factor that must actually operate in order for an event to transpire, and not in evaluating the moral quality of an event with which the causal factor is instrumentally associated.

Moral responsibility, and responsible agency, cannot be identified with causal responsibility. This is true for at least three reasons. First, that someone or something is causally responsible for some event or state of affairs is not a sufficient basis for moral evaluation. That the agent is causally responsible for Φ (or for Φing) is not, in and of itself, a reason to judge that the agent is good or bad, or praiseworthy or blameworthy, for having brought about Φ. It is not enough, that is, to assess the quality of the agent's will. But judgments of moral responsibility are typically of this sort.

Second, a person need not be causally responsible for the choices, acts, and states of affair for which he or she is held morally responsible – causal responsibility is not necessary for moral responsibility. For example, parents are vicariously responsible, legally and perhaps morally, for at least some of the acts for which their nonadult children bear direct causal responsibility.

Third, causal responsibility can be ascribed to events, objects, nonhuman animals, and irrational or underdeveloped humans, as well as to fully mature, rational persons. That someone or something is causally responsible for some event or state of affairs is not sufficient for the cause to be an *agent* who can be held morally responsible. An agent is a being who deliberates, reflects, decides, intends, and makes states of affair come about. An agent (ideally) establishes standards of behavior for himself and is in control of himself, unlike a patient, or a being that is acted upon, managed, or caused to assume various states. Only *qua* agent is a person in a position to realize his goals by his own effort. Even less does an attribution of causal responsibility single out a class of things that might properly be described as *moral agents*, or *members of a moral community*. When we hold a person such as Alice morally responsible for something, it is in light of certain expectations we have of Alice, given a maturity we assume of her, given what we have reason to believe she knew herself to be doing, and given the circumstances that led her to act. But the conditions for responsible agency met by Alice, and considered more fully below, are not met by very young children, animals, trees, or inanimate phenomena. Thus, while we might assert that faulty construction is to blame for the collapse of a building, what we mean is that a salient cause of the collapse was poor construction. We might allege that the baby was causally responsible for the broken lamp. But neither the baby nor the faulty construction is morally responsible for the states they cause because neither are moral agents.

Duty and Obligation

In describing a person as 'responsible' for some act or state of affairs, we may be expressing the belief that the person is bound by duty to ensure the occurrence of the act. Here, the judgment 'X is responsible for Y' is intended as a statement to the effect that X has certain responsibilities to others: tasks or obligations pertaining to Y, usually in light of X's social or role-given position. Claiming that a person is responsible for Y in this sense is to credit responsibility *prospectively*; it is to look forward to what the person is duty-bound or obligated to do. This idea of having responsibilities is analogous to Kurt Baier's notion of

'task-responsibility,' or the 'forward-looking aspect' of agent-responsibility (Baier, 1972). In its forward-looking sense, agent-responsibility consists in being socially entrusted, on customary, legal, or moral grounds, with a specified event, state of affairs, or duty. Thus, parents are responsible for overseeing the well-being of their children, adult wage earners are responsible for paying taxes, attorneys and psychiatrists are responsible for respecting the confidentiality of their clients, and so forth. In this forward-looking sense, the claim that Alice is responsible for giving Bernie the medicine would mean, at a minimum, that Alice has a duty to give Bernie medicine.

Assertions of responsibility are thus ambiguous between (1) the sense in which they attribute a prospective duty, (2) the sense in which they attribute mere causal responsibility, and (3) the sense in which they attribute moral responsibility, and perhaps imply praise or blame. In the last sense, to be morally responsible is not a matter of having or owing an obligation to another. A person may be morally responsible for those things she had a moral obligation to do, such as caring for her children. But the things for which a person can be held morally responsible in the sense of being praiseworthy or blameworthy are not restricted to things the person had an obligation or a responsibility to do. I might be held (morally) responsible for my slovenly habits, or for my acts of bravery, where neither of these attend the discharging of an obligation. Equally, the fact that an act is one I am obliged to perform does not decide whether I am morally blameworthy or praiseworthy for performing it or for failing to do so. I may have a duty as a parent to seek legal representation for my criminally errant daughter, but am I necessarily blameworthy – is it to my discredit – if I refuse to do so?

Having excluded causal responsibility and responsibility in the sense of duty as the types of responsibility with which an analysis of moral responsibility is concerned, we can turn to the latter idea. What we are interested in are the ideas of being responsible and, more specifically, of being morally responsible.

Being a Responsible Agent and Being Held Responsible

The Conditions for Responsible Agency

An account of moral responsibility must make sense of moral approval or disapproval, and it must be restricted to the activity of agents. Even if the act for which we deem a person responsible is morally neutral – that is, the act is neither praiseworthy nor blameworthy – we appraise the moral quality of the person who acts, and not simply the state of affairs he initiates. Generally speaking, when philosophers describe someone as a morally responsible agent, they mean that the person is someone whom we recognize as bearing characteristics of the sort that makes him a member of the moral community. The person is someone with whom morally significant relationships are possible, and he is a suitable subject of moral appraisal. Philosophers have offered different theories about the conditions for membership in the moral community. In Sections Holding Responsible and Being Responsible, Responsibility as Susceptibility to Reactive Attitudes, Responsibility as Attribution, and Responsibility as Accountability, three standard accounts of the characteristics that morally responsible agents must possess will be discussed. But these accounts are united by points of commonality. Notably, most philosophers agree that membership typically requires that an individual satisfy certain *epistemic* conditions and certain social and psychological conditions of *control*.

The epistemic conditions assure us that the responsible agent is self-aware, that he is means – end rational, and that he is not ignorant of the circumstances in which he acts. In addition, membership requires that the actor is acquainted with and is able to act within established moral guidelines, and that he is responsive to reasons to adjust or amend his behavior in light of these guidelines. In order to be held responsible, the moral agent must know that doing a particular act (or an act of a given type) or cultivating a particular trait of character (say, jealousy, rage, or bigotry) is right or wrong. The moral agent may be held responsible if, suffering from none of the conditions that exempt a person from responsible agency, he should have known the nature of the act or trait, and could have been motivated by the relevant moral guidelines. One's ignorance in this case might be called culpable. By contrast, one should not be held responsible if one did not know, and should not be expected to know, that the act is of an undesirable moral quality.

The control conditions ensure that the agent acts freely. The morally responsible agent is not hostage to debilitating and overwhelming neuroses of a sort that incapacitate his resolve; that is, he can govern the manifestations of neuroses by medical means or by strength of will. A person who is prone to neurosis of a certain type or scope, or who finds himself the victim of coercion or manipulation, may be a responsible agent but one whom we excuse from responsibility for specific acts when it is plain that on a particular occasion coercive or manipulative elements played a decisive role in action. Satisfaction of the control conditions for responsible agency ensure that the agent, rather than a mechanism foreign to him, actually guides or executes the agent's actions and those traits of character for which the agent is morally responsible. A mechanism of action is foreign to a person when it is one to which the person does not assent as a force by which he wishes to be moved. Although the responsible person may lack the power to alter the mechanism that leads him to act – to do otherwise than he in fact does – he is moved by his own lights. At a minimum, the morally responsible agent is capable of deliberate, intentional behavior.

Absent some of the epistemic and control conditions, a person will be a less probable candidate for responsible agency and for membership in the community of moral agents. The satisfaction of some set of these will mean that the entity is capable of responsible agency. There is obviously a threshold that decides when a person is a responsible agent, but it is not within the scope of this discussion to explore the specifics of the threshold. Let it suffice to say that the responsible agent is someone who acts knowingly and freely, and the actions and characteristics for which he is responsible are 'his own.' The account of responsibility laid out here follows closely the account defended by John Martin Fischer. One who is responsible must meet conditions for free agency. These conditions are met when the mechanism that leads the agent to act is moderately reasons-receptive and weakly reasons-responsive.

Fischer's is a 'semicompatibilist account' of responsibility, in that the question of compatibility of moral responsibility and causal determinism can be separated from questions of compatibility of causal determinism and freedom to do otherwise (see Fischer and Ravizza, 1998).

Holding Responsible and Being Responsible

While there is a fair degree of unanimity regarding the basic conditions of epistemic competence and control expected of morally responsible beings, disagreement has transpired over the proper analysis of the judgment that a person is morally responsible for his character, choices, actions, and omissions. In part, the disagreement points to the need to separate questions of what it is to *be* responsible from questions of what it means to *hold* a person responsible. Let us begin with the question of what it means to judge that a person is responsible. Three analyses, in particular, bear mention.

Responsibility as Susceptibility to Reactive Attitudes

One possibility is that in claiming that a person is responsible, we might simply be signaling that an actor, having met certain conditions for responsible agency and having displayed in his conduct a distinctively interpersonal quality of will, is a suitable candidate for a particular quality of regard and treatment forthcoming on the part of others and, perhaps, self-reflexive regard as well. This approach has been coined the reactive-attitudes analysis of responsibility ascriptions. The idea is that when we say a person is responsible, we mean that the person is rightly liable, through his behavior, to a range of responses on the part of others marked by attitudes such as gratitude and resentment, and that the person is an appropriate object of such attitudes in light of certain conditions for agency that he has met. On this analysis, the claim 'X is responsible for Y' (or 'for Y-ing') will amount to the conjunction of the claims "X possesses the epistemic and control abilities of a responsible agent," "X did (or is, or caused to come about) Y," and "X is rightly subject to the favorable or unfavorable responses of others as a result." The reactive-attitudes analysis focuses on the emotional dynamics at stake where decisions about responsibility transpire. That a person is appropriately subject to certain attitudes and practices indicates our prior assessment of the person as a free, competent agent, and offers further reason to assess the person as good or bad, praiseworthy or blameworthy.

Responsibility as Attribution

A second possibility is that judging a person responsible for some act, omission, state of affairs, or trait of character essentially is a judgment of attribution. What we are doing is appraising actions and choices that express attitudes and dispositional states attributed to the person as his own. Gary Watson labels this the *aretaic* face of responsibility for its focus on ethical appraisal of conduct that can be imputed to an individual as its agent and thus as the author or adopter of ends. Because these ends are self-adopted, acting upon them discloses the quality and character of the agent's life. Likewise, Angela Smith argues that the judgment 'P is responsible for Φ' amounts to the claim that 'Φ can be attributed to P as a basis for moral appraisal.' Smith follows T.M. Scanlon in suggesting that attributability implies answerability for those attitudes that are expressions of the actor's character and will. To simplify, to be attributable-responsible is to have judgments made about our moral character given our attitudes and actions, where the latter are connected to our underlying evaluative judgments, that is, to reasons we treat as justifications for what we did. (See Smith, 2007, 2008.)

The attributability approach does far more than assign causal responsibility, since its focus is on the intentional behavior of a human agent, his character, and his ends. At the same time, a judgment of attributability may fail to provide as much as we require of an account of responsibility. To attribute responsibility is to indicate some type of motivationally efficacious relation between the agent and his actions and attitudes, and while this relation tends to shed light upon the normative character of the agent, attribution need not involve the registration of emotions such as resentment or gratitude. Nor need it be true that an explanation of the act in question is anticipated of the actor, although, as we have seen, attributability may well be accompanied by an expectation of answerability.

Responsibility as Accountability

A third possibility is that, in judging a person responsible, we may be saying that we hold the person accountable for his behavior. Construing judgments about moral responsibility as judgments about accountability is appropriate for a number of reasons. One reason is that holding persons responsible is part of a social practice of ascribing responsibility, first by making clear any moral expectations we have of persons vis-à-vis behavior and character, and then by having in place certain social instruments that ensure adequate redress where necessary. Making moral expectations clear presumes the existence of a shared set of beliefs and a system of norms against which actions and traits of character are assessed.

This notion of accountability is variously understood. With Scanlon, David Shoemaker describes accountability as an enhanced species of *answerability*. Answerability for Φ turns on the fact that Φ must be connected to the actor's evaluative judgments, those that express reasons the actor takes to justify his actions. Thus, answerability targets the deliberative, reason-tracking aspect of the agent (Shoemaker, 2011: 616). To be *accountable*, Shoemaker argues, signals the further and distinct idea that the actor is open to sanction and susceptible to being called to account publically for his behavior because the actor has flouted a relationship-defining demand, or because a claim to goodwill has been overlooked. Gary Watson refers to the aspect of responsibility that pertains to accountability as the second of responsibility's two faces, the *culpability* face. For Watson, to be accountable-responsible is to be culpably open to blame. We judge P culpable or open to legitimate criticism because P's choices and actions reveal attitudes that express an inadequate level of regard for others, given P's relations to others (see Watson, 1996).

Others (myself included) construe accountability differently. Rather than incorporate assumptions about the responses of morally affected others into an assessment of

accountability, judgments of accountability emphasize the dimension of answerability, understood as interlocutionary competence. The focus is on the actor and the audience far less than on a judge. To be accountable signals nothing more than the fact that it is appropriate that the actor offer an explanation of his intentions or beliefs about his behavior. To be morally responsible just is to be accountable, and to be accountable just is to be answerable. It is to be positioned, in virtue of one's normative capacities in conjunction with the conditions under which one acts, to give an account of one's intentions and beliefs about one's behavior. The accountable agent must be attentive to the presence of reasons for action that others would accept are on offer and must be capable of managing her behavior in light of these. Given this narrower construction of accountability analysis, to be blameworthy would call for the further fact that one has acted in a way that reveals an inadequate level of regard for others and signals impairment of relations with others.

Note that even on the more expansive construction of accountability, blameworthiness need not involve readiness to respond with reactive attitudes. Blameworthiness and praiseworthiness call for awareness of reasons that are operative and that are standardly regarded by others as reasons to act or to refrain from acting. Oddly, perhaps, blameworthiness might not necessitate an appreciation of the distinctively moral dimensions of one's conduct (see Greenspan, 2003; Oshana 2004; Scanlon, 2008; Talbert, 2008).

Case Studies

Extreme Evil

Two varieties of cases illustrate the differences among these three analyses. First, consider examples of extreme and, some would contend, irremediable evil. If we consider the behavior of someone such as Hitler, it may well appear that he is morally 'beyond the pale' so that none of the standard reactive attitudes is an appropriate response to his 'Final Solution.' Hitler and likeminded persons are ostensibly incapacitated for ordinary adult interpersonal relationships. Such persons simply will not be responsive to the demands of membership in our moral community. At the same time, the heinous acts they have committed are properly attributed to them as perpetrators. The attributability account allows for Hitler, and similarly sociopathic or psychopathic persons, to be deemed responsible for their immoral behavior, yet permits the hesitation some might have in regarding such actors as moral interlocutors, as the accountability approach demands, or as appropriate objects of any of the ordinary moral attitudes, as the reactive-attitudes analysis would require.

Private Moral Failures

One merit of the attributability approach is that it enables us to credit persons with responsibility for matters that are essentially private, matters that are the affair of none but the actor herself. It seems unreasonable to reply to my slovenly garb, or foolish financial practice, with resentment or indignation, just as it seems unreasonable to expect that I account for these practices if they affect no one but myself. Where these practices reveal something about ends I have chosen for myself, they speak to my character – they express, as Watson claims, my own evaluative commitments and my identity as an agent. Thus, on the attributability account, they can be attributed to me as practices I bear responsibility for.

Responsibility and Free Will

Causal Determinism and Moral Responsibility

An adequate discussion of the central problems faced by any account of moral responsibility will include the problem that causal determinism poses to the idea that we have freedom of will. In brief, the problem is that, if our present choices and conduct are causally determined by past events and states, then it appears not the case that we have any freedom with respect to the performance or occurrence of our present choices and conduct. The past is something over which we have no control. And it is equally true that we have no control over the laws of nature operative in the world. Together, these claims suggest that we have no control over the present.

It is a matter of lively debate among philosophers whether or not the truth, or the likelihood, of causal determinism would affect our status as responsible agents. Some philosophers argue for the incompatibility of determinism and responsibility; if either obtains, the other does not. We cannot be both determined and responsible for our conduct. Others argue that the conditions for responsible agency are independent of determinism. On this view, the truth or plausibility of determinism aside, we can control our conduct in a manner sufficient for responsible agency. At the very least, to deny responsibility of persons whom we regard as participants in our moral framework would be to strain the limits of our natural practices. It is undeniable that persons treat one another as agents, as individuals answerable for their conduct, and the alternative – that persons cultivate an attitude of detachment from the conduct of their human associates merely because of the truth of determinism – would signal an impoverished state of human interaction that few, if any, could bear.

See also: Autonomy, Philosophy of; Ethics and Values; Free Will and Action; Identity and Identification, Social Psychology of; Intentionality and Rationality.

Bibliography

Baier, Kurt, 1972. Types of responsibility. In: French, Peter A. (Ed.), Individual and Collective Responsibility. Schenkman Publishing, Cambridge, MA.

Fischer, John Martin, Ravizza, Mark, 1998. Responsibility and Control. Cambridge University Press, Cambridge, UK.

Greenspan, Patricia S., 2003. Responsible psychopaths. Philosophical Psychology 16 (3), 417–429.

Oshana, Marina, 2004. Moral accountability. Philosophical Topics 32 (1–2), 255–274.

Scanlon, T.M., 2008. Moral Dimensions: Permissibility, Meaning, and Blame.

Shoemaker, David, April 2011. Attributability, answerability, and accountability: toward a wider theory of moral responsibility. Ethics 121 (3), 602–632.

Smith, Angela, 2007. On being responsible and holding responsible. The Journal of Ethics 11, 465–484.

Smith, Angela, 2008. Control, responsibility, and moral assessment. Philosophical Studies 138, 367–392.
Talbert, Matthew, 2008. Blame and responsiveness to moral reasons: are psychopaths blameworthy? Pacific Philosophical Quarterly 89, 516–535.
Watson, Gary, 1996. Two faces of responsibility. Philosophical Topics 24 (2), 227–248.

Relevant Website

http://plato.stanford.edu/entries/moral-responsibility/ – Stanford Encyclopedia of Philosophy.

Retirement and Encore Adulthood: The New Later Life Course

Phyllis Moen and Jack Lam, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Retirement became an institutionalized, primarily male, later life-course transition by the middle of the twentieth century; an expected, routine exit from a lifetime of paid work that was also seen as the passage to old age. However, large-scale social forces (globalization, aging populations, women's labor force participation, reduced social protections) have upended conventional expectations about retirement and the later life course. The period roughly between ages 55 and 75 years has arguably become a new life course stage, an 'encore' to conventional adulthood occurring typically after the career- and family-building years, but prior to the infirmities associated with old age.

Introduction

Both the later life course and the singular transition in it – retirement – are increasingly destandardized, risky, and ambiguous, given the large-scale social forces of globalization, aging populations (low fertility, medical advances, and extended life expectancy), and scale-backs in social protections, along with women's employment and the rise of dual-earner couples. We argue that the early twenty-first century is witnessing the emergence of a new life stage, encore adulthood, fueled by these social and economic transformations, and especially the increase in healthy life expectancy that is delaying the infirmities associated with old age. Life in this later life-course stage has become a 'project' to be figured out; even retirement is no longer a taken-for-granted routine. Moreover, apart from the Scandinavian countries, in much of Europe and North America, this is the first generation where large numbers of women have jobs they are retiring from, and large numbers of dual-earner couples are facing two retirements, 'his' and 'hers,' requiring an examination of how this experience differs for men and women, as well as whether couples plan and make decisions about retirement together. This is increasingly an unsettled and uncharted period of the life course, suggesting that past research on retirement and other experiences may well be out of date.

A Brief Historical Overview of Retirement

The life course as role trajectories (such as employment patterns) and transitions (such as retirement) is socially constructed, a product of policies and practices in education, employment, retirement, and social welfare regimes (see Kohli, 2007; Mortimer and Shanahan, 2003; *see* Motivation: Life Course and Sociological Perspectives). This highlights the shared understandings and taken-for-granted rules, roles, relationships, resources, and risks triggered at different ages and life stages, but in different ways for men and women (Moen et al., 2013; O'Rand, 2011). Age and gender are both durable predictors of life-course expectations and experiences.

The second half of the twentieth century witnessed the development of cultural norms about retirement as a typical, legitimated, and almost inevitable status passage occurring among men around age 65 years (Costa, 1998). This fostered a 'standard' education, work, and retirement (leisure) lockstep that became the institutionalized life course (Kohli, 2007), even though there were trends toward early exits (Blossfeld et al., 2011; Burkhauser et al., 1996; Ebbinghaus, 2006; Guillemard and Rein, 1993). This sequence was further legitimated by career and retirement mystiques (and the policies reinforcing them) emphasizing that success and fulfillment could be obtained only through continuous investment in occupational mobility, or at least seniority, thereby earning the full-time leisure associated with retirement (Moen and Roehling, 2005; *see* Age Policy). These career/retirement myths typically characterized the experiences, in the United States at least, of only white, middle-class, and unionized blue-collar men in the middle of the twentieth century – and not all of those. Nevertheless, the schema of continuous careers culminating in one-way, one-time retirement exits became institutionalized, accepted as both the way things are and the way things should be. As workers moved into the roles of pensioners, retirees, and seniors, they exited their status and identities as teachers, electricians, laborers, managers, technicians, doctors, nurses, or civil servants.

The early twenty-first century witnessed large numbers of women in career jobs from which they expect to retire (He et al., 2005; Purcell, 2005). Still, institutional inertia in retirement policies and practices means that social security and pension policies reward those following the continuous full-time employment assumed in the (male) career mystique (Moen and Roehling, 2005) and not the uneven employment experiences of most women, immigrants, and minorities. Moreover, now even older men in 'good' jobs are often experiencing unexpected dislocations. For example, some adults are finding themselves suddenly 'retired' as a result of layoffs or buyouts, with few reemployment prospects (Moen et al., 2010; Sweet and Moen, 2012) or else working under chronic job insecurity (Burgard et al., 2009, 2012). The risks embedded in transitional global labor markets are increasingly obvious, affecting even older workers, previously protected by their seniority (Kalleberg, 2011; Schmid, 2006; Supiot, 2001).

The first decades of the twenty-first century also witnessed a move away from retirement as an age-mandated exit (through public and organizational policies, practices, and age norms),

and away from the mid-twentieth-century trend toward progressively earlier retirement (Blekesaune and Solem, 2005; Blossfeld et al., 2011; Ebbinghaus, 2006; Gendell, 2008; Purcell, 2005). Retirement as a traditional status passage is unraveling, and retirement along with the years of encore adulthood from around 1950s to around the 1970s are becoming problematic, often requiring deliberate choices by individuals and families, a significant departure from simply following taken-for-granted scripts from full-time work to full-time leisure at a prescribed age.

In addition, there is greater heterogeneity in when and why people leave the workforce (*see* Aging and the Labor Market; Old Age Work Participation). Men and women of particular socioeconomic and racial backgrounds at particular ages at particular time periods, living in particular communities, and in particular countries are simultaneously confronting retirement-related decisions within the contexts of both their own situational biographies and a society in flux. For instance, retirement is increasingly perceived as less of a choice by workers in the USA. A study of a national sample of 1024 US adults over age 50 years found a third of retirees report they did not feel like they had a choice over the decision. This was more common among younger respondents (under 65 years); over half (54%) report they had no choice but to retire. In contrast, only 23% of those over 65 years report feeling retirement was forced upon them (Benz et al., 2013).

An Institutional Gendered Life-Course Approach

A life-course approach is useful for studying changes in retirement and the later life course in that it situates individuals in specific life stages and historical contexts, paying particular attention to their present circumstances as a consequence of both personal decisions and institutionalized options and constraints (*see* Motivation: Life Course and Sociological Perspectives). These constrained choices operate to shape peoples' role, relationship, and resource trajectories, and transitions. Thus, the life course consists of distinctive (and unequal) patterns of socially defined, age-graded experiences over time, varying across cohorts, gender, race/ethnicity, education, family background, nativity, health conditions, and historical time period (Levy and Widmer, 2013; Moen, 2014; Mortimer and Shanahan, 2003; Shanahan and Macmillan, 2007; Willson et al., 2007; *see* Life Course and Event History Analysis).

Since gender as well as government, business, and community policies and practices are major shapers of how roles, relationships, and resources are allocated, we propose the value of an institutional gendered life-course theoretical approach to promote understanding of labor market processes and inequalities characterizing the later life course and retirement (Moen, 2013). Importantly, research is necessary to capture the ways work and retirement

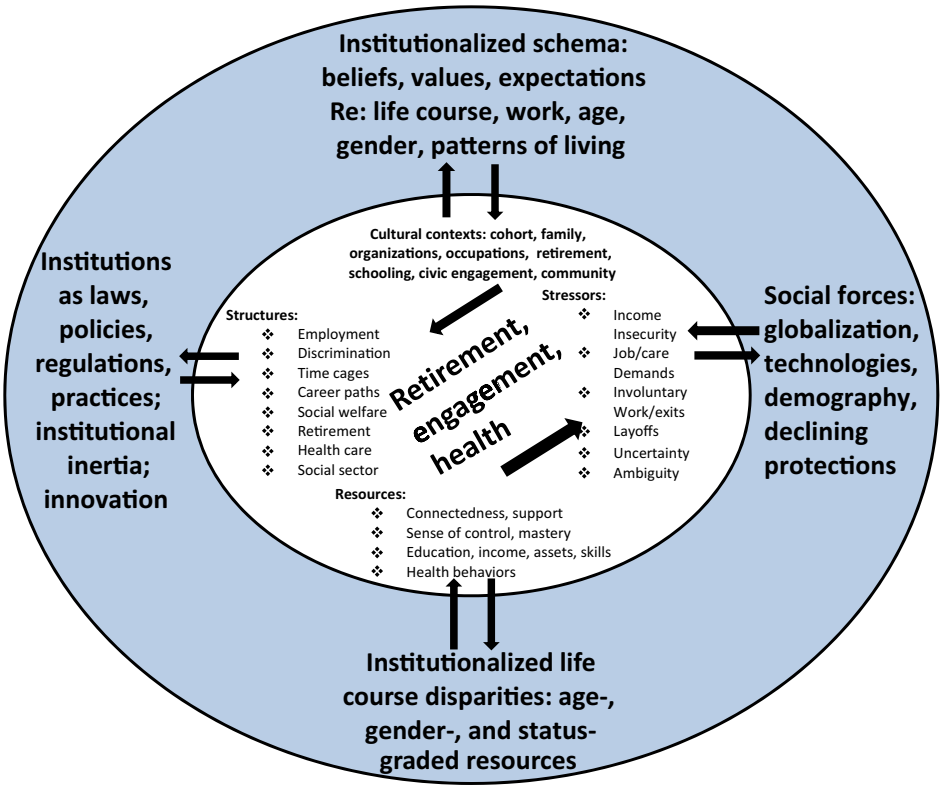

Figure 1 An institutional life-course model of retirement and encore adulthood. Re, regarding. Adaptedand modified from Moen, P., 2013. Constrained choices: the shifting institutional contexts of aging and the life course. In: Waite, L., (Ed.). Panel on New Directions in Social Demography, Social Epidemiology, and the Sociology of Aging, Perspectives on the Future of the Sociology of Aging. National Research Council, Washington, DC, pp. 81–119.

experiences result in contemporary men and women in different countries and different social classes arriving at and moving through the later life course equipped with very different sets of resources, constraints, and options related to employment, retirement timing, family lives, caregiving, health, and lifestyles (see **Figure 1**).

An institutional-gendered life-course theoretical frame underscores the historical times as well as the structural and cultural dynamics in which lives unfold, the uneven rates of social and institutional change, and the complex interplay between biography and history. The early twenty-first century constitutes a transformational historical period, replete with demographic changes (in the form of population aging resulting from extended life expectancy, lower fertility, medical advances), and the aging of the large Boomer cohort (born at different times in different countries but in the USA between 1946 and 1964); economic changes (in the form of global economic turbulence, new technologies, and a global labor force); and institutional changes (dismantling the social contracts equating seniority with job security, reductions and delays in public and private pensions and income). New communication technologies and investment practices have fostered a volatile global economy, with rising employer flexibility in the ability to dismiss high-cost (including older) employees and seek low-cost labor across the globe (Kalleberg, 2011; Sweet and Meiksins, 2013). Moreover, workers' incomes in retirement are contingent on their prior career trajectories, with women and men following different pathways, given the employment breaks or shifts to part-time work that women traditionally take in order to raise children, follow their husbands' job opportunities, or care for infirm relatives, such as aging parents.

Taken together, these social forces are converging to challenge conventional, widely accepted norms and expectations regarding work, retirement, age, gender, and the later life course. Contemporary retirement and later life-course decision-making is taking place in individual lives even as established, standardized cultural scripts, and institutions are being upended.

A New Later Life-Course Stage

In Europe and North America, populations and workforces are aging, yet medical and technological advances are postponing the debilitating conditions associated with old age (*see* Age Stratification). This has paved the way for refiguring the later life course. Scholars are coming to define a new life stage seen as occurring after traditional adulthood but before traditional old age. As longevity and healthy life expectancy increase, individuals in their 50s, 60s, and early 70s are finding themselves in the middle rather than at the end of the contemporary adult course, facing the challenge of having to make up this new life phase as they move along (*see* Retirement and Encore Adulthood: The New Later Life Course).

Originally named the third age (Laslett, 1989), this new stage was depicted as a period of possibility after the first age of childhood, and the second age of adulthood, but before the fourth age of infirmity. We prefer to conceptualize this life phase as 'encore adulthood,' a time to reset priorities and sometimes start new endeavors after the years of family and career building but before the frailties of old age (Moen, 2013, 2014; Moen and Flood, 2013; see also Freedman, 2008, 2012). Retirement still occurs during this new life stage, but as a project that may involve decisions and adaptations to unexpected exigencies around timing, reentries, and possibly several exits from the labor market. The vast Boomer cohort and the World War II cohort just preceding it are moving through this encore adulthood period. But they are without blueprints, confronting uncertainty and ambiguity given shifting risks, protections, norms, and even definitions of retirement (O'Rand, 2011).

There is little existing evidence as to the choices or derailments (such as by layoffs) of contemporary older workers and retirees as they progress through the encore adult years. How do peoples' social structural location, their prior trajectories and transitions, and their self-definitions of being retired (or not) shape their decisions and experiences in encore adulthood? What are the implications of choices (and options or barriers) to continue some form of ongoing public engagement in paid or unpaid work for health, life quality, and even longevity? How is this first large group of older women workers and dual-earner couples adapting to the shifting challenges of retirement?

Encore Adulthood and Retirement in the Twenty-First Century

Retirement is increasingly a phase or project enacted over a period of time, rather than a single event, as people in the encore adult years begin thinking about and planning (or else are thrust into, due to layoffs or a health emergency) life beyond their career jobs. Research is needed on the timing and voluntariness, as well as various pathways, of retirement and other encore adult transitions in various contexts. Some have greater capabilities, opportunities, and advantages in customizing encore adulthood. Education, health, and a sense of mastery or control are critical resources enabling older workers to have more choices regarding the nature and timing of retirement, such as whether to pursue bridge jobs or civic engagement as a way of continuing some form of social participation (e.g., Cahill et al., 2006; Damman et al., 2011; Kaskie et al., 2008). But sometimes older workers exit their jobs because of a push, rather than the pull of retirement (Siegrist et al., 2007; von Bonsdorff et al., 2010). In a study drawing on the 'Survey of Health, Ageing and Retirement in Europe' (SHARE), in 10 European countries, Siegrist et al. (2007) find that poor quality of work (such as low control and effort-reward imbalance) is associated with intended early retirement. Another 11-year study of Finnish municipal employees by von Bonsdorff et al. (2010) found adverse work environments and poor health in midlife predicts subsequent early retirement intentions. Also a study of 1229 Dutch male older (50 plus) workers showed midlife health problems predict early retirement intentions and actual exits (Damman et al., 2011).

However, more broadly disparities in opportunities for public engagement are a result of: (1) the cumulation of

advantages or disadvantages that are associated with class, gender, race/ethnicity, nativity, nationality, and other social markers (Willson et al., 2007); (2) global, economic, and technological forces upending job and retirement security; and (3) the aging of the population, which has raised issues about traditional safety nets (such as Social Security and health care). Some find themselves 'retired' through forced buyouts and layoffs, even as age discrimination prevents their reemployment (Roscigno et al., 2007; Sweet and Meiksins, 2013; Sweet et al., 2007). A 2012 study of a sample of 1502 US adults aged 45–74 years who were working full time or part time, self-employed, or looking for work, finds that almost two-thirds (64%) reported having seen or experienced age discrimination at the workplace. Additionally, one in five (19%) reported that they were not hired for a job they applied for, over one in 10 (12%) reported they were passed up for a promotion, and almost one in 10 (9%) reported being laid off or fired because of age (AARP, 2013a).

Others find they cannot afford to leave their (often demanding) full-time career jobs owing to financial reasons. A US study finds that 4 in 10 retirees over age 50 have problems with debt, while 10% describe their debt as a major problem (AARP, 2013b). This could become an increasingly important issue for the next cohort, given changes to pension schemes and savings patterns. A study of a national sample of 1000 Gen X'ers (born between 1965 and1976) in the USA finds that only a third are on track or have already reached their retirement savings goal, while 7% have not started saving, and another 11% do not have any retirement savings goals (Metlife, 2013).

There is evidence of a lack of well-defined roles for older workers and those who have already exited their career jobs (Casner-Lotto, 2007; Hutchens and Papps, 2005), even though human resource executives anticipate an eventual labor shortage (Ernst and Young, 2006; Leibold and Voelpel, 2006). There is also evidence of greater family engagement and expectations, especially for women, in the form of caring for older parents, spouses, adult children, and grandchildren (Chesley and Moen, 2006; Meyer, 2014; Shahly et al., 2013). Yet, remarkably little attention has been paid to the convergence of dislocational social forces reshaping the experience of those in encore adulthood, approximately 20–25 years between the life stages of building families and careers and the frailties of old age. What is known is that retirement, as we know it, is in disarray.

In the face of an aging population, current public policy goals in the USA and in Europe only seek to delay the transition from full-time work to full-time leisure by several years (in the USA, until age 70 years), or else to encourage phased retirement with work-hour reductions, often on the career job (Delsen and Reday-Mulvey, 1996; Esping-Andersen et al., 2002; Munnell and Sass, 2008; Quinn, 1999; Ruhm, 1996; Taylor, 2002). In other words, they are retaining the standard lockstep life course, only shifting the timing of transitions; few policymakers are considering innovations making ongoing or postretirement employment more meaningful and attractive or facilitating unpaid civic engagement (Freedman, 2008; Kaskie et al., 2008; McNamara and Williamson, 2013; Esping-Andersen, 2009).

The exclusion of older people from ongoing engagement is woven into the fabric of both labor markets and welfare states (see Esping-Andersen, 2009). Moreover, there are increasingly different institutional logics (taken-for-granted beliefs, values, and expectations) – some overlapping, some conflicting – within society regarding retirement and the later life course (Moen, 2013).

Research suggests that many in the encore adult years want something different than continuing their existing (often demanding) career jobs or else moving into total (nonwork) retirement. Moen and Flood (2013) find that significant numbers of Americans in their 50s, 60s, and early 70s are pursuing alternatives to the standard life course, often working in different ways – part-time, self-employed, pursuing second or third careers with new employers and/or in different occupations, and/or formally volunteering for community organizations. However, such forms of social inclusion are available for those who are in a more advantageous position in terms of education and health.

Social inclusion or exclusion of older adults may well have important consequences for health and health behaviors. There is evidence that older workers report both more smoking and more leisure-time physical activity than same-age nonworkers (Fleming et al., 2007), and that older women who leave the workforce are more likely to gain weight than those still employed (Forman-Hoffman et al., 2008). Involuntary job loss among older workers has been found to reduce physical functioning and mental health, whereas reemployment has positive health impacts (Bender, 2012; Gallo et al., 2000, 2006). One study (Bender, 2012) using the 2010 Health and Retirement Study following older Americans reports that respondents who were 'forced' to retire had significantly lower well-being.

Directions for Future Research

This is an exciting time to study retirement and the changing later life course, what we call the encore adult years, occurring prior to the infirmities traditionally associated with old age. Research is needed on the different pathways and possibilities for encore adulthood and retirement for those in different cultural and policy contexts and with different material, social, and personal resources. Are there identifiable patterns of retirement exit based on the economic and policy climates of particular nations and regions and what accounts for these different patterns (see Kinsella and Phillips, 2005; Slack and Jensen, 2008)? What pathways and transition patterns appear to be most conducive to health and subjective, as well as financial, well-being? How important is anticipatory socialization, in the form of both formal and informal retirement discussions and plans (Curl and Ingram, 2013)? Clearly this is a fertile and important area for future research.

See also: Age Policy; Age Stratification; Age Structure; Age, Sociology of; Aging and Health in Old Age; Aging and the Labor Market; Life Course and Event History Analysis; Life Course: Sociological Aspects; Linked Lives; Old Age Work Participation; Old Age and Centenarians: The Human

'Warranty Period'; Old Age: Definitions, Theory, and History of the Concept; Population Aging: Economic and Social Consequences; Population Forecasts; Retirement and Health; Retirement, Economics of; Welfare State, History of.

Bibliography

AARP, 2013a. Staying Ahead of the Curve 2013: AARP Multicultural Work and Career Study. Perceptions of Age Discrimination in the Workplace – Ages 45–74. AARP, Washington, DC. Retrieved from: www.aarp.com.

AARP, 2013b. 2013 Retirement Confidence Survey: A Secondary Analysis of the Findings from Respondents Age 50+. AARP, Washington, DC. Retrieved from: www.aarp.com.

Bender, K.A., 2012. An analysis of well-being in retirement: the role of pensions, health, and 'voluntariness' of retirement. The Journal of Socio-Economics 41 (4), 424–433.

Benz, J., Sedensky, M., Tompson, T., Agiesta, J., 2013. Working Longer: Older Americans' Attitudes on Work and Retirement. The Associated Press and NORC. Retrieved from: www.apnorc.org.

Blekesaune, M., Solem, P.E., 2005. Working conditions and early retirement: a prospective study of retirement behavior. Research on Aging 27, 3–30.

Blossfeld, H., Buchholz, S., Kurz, K., 2011. Ageing Populations, Globalization and the Labor Market: Comparing Late Working Life and Retirement in Modern Societies. Edward Elgar, Northampton, MA.

von Bonsdorff, M.E., Huuhtanen, P., Tuomi, K., Seitsamo, J., 2010. Predictors of employees' early retirement intentions: an 11-year longitudinal study. Occupational Medicine 60 (2), 94–100.

Burgard, S., Brand, J., House, J., 2009. Perceived job insecurity and worker health in the United States. Social Sciences & Medicine 69 (5), 777–785.

Burgard, S.A., Kalousova, L., Seefeldt, K.S., 2012. Perceived job insecurity and health: the Michigan recession and recovery study. Journal of Occupational and Environmental Medicine 54 (9), 1101–1106.

Burkhauser, R.V., Couch, K.A., Phillips, J.W., 1996. Who takes early social security benefits: the economic and health characteristics of early beneficiaries. Gerontologist 36, 789–799.

Cahill, K.E., Giandrea, M.D., Quinn, J.F., 2006. Retirement patterns from career employment. Gerontologist 46 (4), 514–523.

Casner-Lotto, J., 2007. Boomers are Ready for Nonprofits but Are Nonprofits Ready for Them? The Conference Board, New York.

Chesley, N., Moen, P., 2006. When workers care: dual-earner couples' caregiving strategies benefit use, and psychological well-being. American Behavioral Scientist 49 (9), 1248–1269.

Costa, D., 1998. The Evolution of Retirement: An American Economic History, 1880–1999. University of Chicago Press, Chicago.

Curl, A.L., Ingram, J.G., 2013. Anticipatory socialization for retirement: a multilevel dyadic model. Clinical Gerontologist 36, 375–393.

Damman, M., Henkens, K., Kalmijn, M., 2011. The impact of midlife educational, work, health, and family experiences on men's early retirement. Journal of Gerontology: Social Sciences 66B (5), 617–627.

Delsen, L., Reday-Mulvey, G., 1996. Gradual Retirement in the OECD Countries. Ashgate, Aldershot, UK.

Ebbinghaus, B., 2006. Reforming Early Retirement in Europe, Japan and the USA. Oxford University Press, Oxford.

Ernst & Young, 2006. The Aging of the U.S. Workforce: Employer Challenges and Responses. Ernst & Young, New York.

Esping-Andersen, G., 2009. The Incomplete Revolution: Adapting to Women's New Role. Polity, Cambridge, UK.

Esping-Andersen, G., Gallie, D., Hemerijck, A., Myles, J., 2002. Why We Need a New Welfare State. Oxford University Press, Oxford.

Fleming, L.E., Lee, D.J., Caban Martinez, A.J., LeBlanc, W.G., McCollister, K.E., Chung Bridges, K., et al., 2007. The health behaviors of the older US worker. American Journal of Industrial Medicine 50, 427–437.

Forman-Hoffman, V.L., Richardson, K.K., Yankey, J.W., Hillis, S.L., Wallace, R.B., Wolinsky, F.D., 2008. Retirement and weight changes among men and women in the health and retirement study. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 63, S146–S153.

Freedman, M., 2008. Encore: Finding Work That Matters in the Second Half of Life. Public Affairs, New York.

Freedman, M., 2012. The Big Shift: Navigating the New Stage beyond Midlife. Public Affairs, New York.

Gallo, W.T., Bradley, E.H., Siegel, M., Kasl, S.V., 2000. Health effects of involuntary job loss among older workers: findings from the health and retirement survey. Journals of Gerontology Series B 55, S131–S140.

Gallo, W.T., Teng, H.M., Falba, T.A., Kasl, S.V., Krumholz, H.M., Bradley, E.H., 2006. The impact of late career job loss on myocardial infarction and stroke: a 10-year follow-up using the health and retirement survey. Occupational and Environmental Medicine 63 (10), 683–687.

Gendell, M., January 2008. Older workers: increasing their labor force participation and hours of work. Monthly Labor Review 41–54.

Guillemard, A.M., Rein, M., 1993. Comparative patterns of retirement: recent trends in developed societies. Annual Review of Sociology 19, 469–503.

He, W., Sengupta, M., Velkoff, V., DeBarros, K., 2005. 65+ in the United States: 2005 (U.S. Census Bureau), Current Population Reports No. P23-209. U.S. Government Printing Office, Washington, DC.

Hutchens, R., Papps, K.L., 2005. Developments in phased retirement. In: Clark, R.L., Mitchell, O.S. (Eds.), Reinventing the Retirement Paradigm. Oxford University Press, New York, pp. 133–160.

Kalleberg, A.L., 2011. Good Jobs, Bad Jobs: The Rise of Polarized and Precarious Employment Systems in the United States. Russell Sage, New York.

Kaskie, B., Imhof, S., Cavanaugh, J., Culp, K., 2008. Civic engagement as a retirement role for aging Americans. Gerontologist 48, 368–377.

Kinsella, K., Phillips, D.R., 2005. Global aging: the challenge of success. Population Bulletin 60 (1). Population Reference Bureau, Washington, DC.

Kohli, M., 2007. The institutionalization of the life course: looking back to look ahead. Research in Human Development 4, 253–271.

Laslett, P., 1989. A Fresh Map of Life: The Emergence of the Third Age. Harvard University Press, Cambridge.

Leibold, M., Voelpel, S., 2006. Managing the Aging Workforce: Challenges and Solutions. Publicis and Wiley, Erlangen, Germany.

Levy, R., Widmer, E., 2013. Gendered life courses between individualization and standardization. In: A European Approach Applied to Switzerland. LIT Verlag, Wien.

McNamara, T., Williamson, J., 2013. Unequal Prospects: Is Working Longer the Answer? (Framing 21st century social issues) Routledge, New York.

Metlife, 2013. The Metlife Study of Gen X: The MTV Generation Moves into Mid-Life. Metlife Mature Market Institute, New York. Retrieved from: www.metlife.com.

Meyer, M.H., 2014. Grandmothers at Work: Juggling Families and Jobs. NYU Press, New York.

Moen, P., 2013. Constrained choices: the shifting institutional contexts of aging and the life course. In: Waite, L. (Ed.), Panel on New Directions in Social Demography, Social Epidemiology, and the Sociology of Aging, Perspectives on the Future of the Sociology of Aging. National Research Council, Washington, DC, pp. 81–119.

Moen, P., 2014. Boomers on the Edge: Improvising Encore Adulthood. Oxford University Press, Oxford.

Moen, P., Flood, S., 2013. Limited engagements? Women's and men's work/volunteer time in the encore life course stage. Social Problems 60 (2), 1–28.

Moen, P., Lam, J., Jackson, M., 2013. Aging families and the gendered life course. In: Treas, J., Scott, J., Richards, M. (Eds.), The Wiley Blackwell Companion to the Sociology of Families. Wiley Blackwell, Oxford, pp. 444–463.

Moen, P., Roehling, P., 2005. The Career Mystique: Cracks in the American Dream. Rowman & Littlefield Publishers, Inc., Boulder.

Moen, P., Sweet, S., Hill, R., 2010. Risk, resilience, and life course fit: older couples' encores following job loss. In: Fry, P.S., Keyes, C.L.M. (Eds.), New Frontiers in Resilient Aging. Cambridge University Press, New York, pp. 283–309.

Mortimer, J., Shanahan, M., 2003. Handbook of the Life Course. Kluwer Academic/Plenum Publishers, New York.

Munnell, A.H., Sass, S.A., 2008. Working Longer: The Solution to the Retirement Income Challenge. Brookings Institution Press, Washington, DC.

O'Rand, A.M., 2011. 2010 SSS presidential address: the devolution of risk and the changing life course in the United States. Social Forces 90 (1), 1–16.

Purcell, R., 2005. Working in the Community: Perspectives for Change. Lulu Press, Raleigh.

Quinn, J.F., 1999. Retirement Patterns and Bridge Jobs in the 1990s. EBRI Issue Brief no. 206. Employee Benefit Research Institute, Washington, DC.

Roscigno, V.J., Mong, S., Byron, R., Tester, G., 2007. Age discrimination, social closure and employment. Social Forces 86 (1), 313–334.

Ruhm, C.J., 1996. Gender differences in employment behavior during late middle age. Journal of Gerontology: Social Sciences 51B, S11–S17.

Schmid, G., 2006. Social risk management through transitional labour markets. Socio-Economic Review 4, 1–33.

Shahly, V., Chatterji, S., Gruber, M.J., Al-Hamzawi, A., Alonso, J., Andrade, L.H., Kessler, R.C., 2013. Cross-national differences in the prevalence and correlates of burden among older family caregivers in the World Health Organization World Mental Health (WMH) Surveys. Psychological Medicine 43 (4), 865–879.

Shanahan, M.J., Macmillan, R., 2007. Biography and the Sociological Imagination. Norton, New York.

Siegrist, J., Wahredorf, M., von dem Knesebeck, O., Jurges, H., Borsch-Supan, A., 2007. Quality of work, well-being, and intended early retirement of older employees – baseline results from the SHARE study. European Journal of Public Health 17 (1), 62–68.

Slack, T., Jensen, L., 2008. Employment hardship among older workers: does residential and gender inequality extend into older age? Journal of Gerontology: Social Sciences 63, S15–S24.

Supiot, A., 2001. Beyond Employment. Changes in Work and the Future Labour Law in Europe. Oxford University Press, Oxford.

Sweet, S., Meiksins, P., 2013. Changing Contours of Work: Jobs and Opportunities in the New Economy, second ed. (Sociology for a New Century Series) Pine Forge Press, Thousand Oaks.

Sweet, S., Moen, P., 2012. Dual earners preparing for job loss: agency, linked lives, and resilience. Work and Occupations 39 (1), 35–70.

Sweet, S., Moen, P., Meiksins, P., 2007. Dual earners in double jeopardy: preparing for job loss in the new risk economy. Research in Sociology of Work 17, 437–464.

Taylor, P., 2002. New Policies for Older Workers. Transitions after 50 Series. Policy, Bristol, UK.

Willson, A.E., Shuey, K.M., Elder, G.H., 2007. Cumulative advantage processes as mechanisms of inequality in life course health. American Journal of Sociology 112 (6), 1886–1924.

Relevant Websites

www.aarp.com – AARP.

www.apnorc.org – The Associated Press-NORC Center for Public Affairs Research.

www.bc.edu/research/agingandwork/ – The Center on Aging & Work at Boston College.

www.encore.org.

www.metlife.com – MetLife, Inc. is a global provider of insurance, annuities and employee benefit programs.

Retirement and Health

James Banks, School of Social Sciences, University of Manchester, Manchester, UK
Tarani Chandola and Katey Matthews, Cathie Marsh Centre for Census and Social Research, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Whether theoretical or empirical, studies considering the links between retirement and health need to be precise about the meaning of both terms. Retirement experiences are now diverse and may have positive or negative effects on various different domains of health. Strong empirical correlations have been established in many settings but the findings of more robust causal analyses seeking to test the effect of retirement on future health are weaker and somewhat mixed. Recent and upcoming policy reform, coupled with ongoing interdisciplinary data collection and harmonization activities, mean that the potential for future research in this area is great.

Introduction

Retirement has become an increasingly important phenomenon in the life course, whether measured in terms of the proportion of life spent in retirement or in terms of the salience of retirement in people's plans and experiences. Within the context of research, the term 'retirement' is now being used to refer to many different events and processes. When defined on the basis of paid work it may refer to total exit from the labor market (e.g., a "definitive withdrawal from all labour market activity with the intention of staying out permanently," Lazear, 1987) or alternatively to ending one's main career (e.g., "the exit from an organisational path or career of considerable duration, taken by individuals after middle age, and taken with the intention of reduced psychological commitment to work thereafter," Feldman, 1994). But nonemployment-based definitions of retirement are also frequently used. Some relate to the receipt of, or eligibility for, disability or pension benefits. Others simply consider self-reported subjective economic status.

The empirical relationship between these various conceptualizations of retirement has become increasingly complex in recent cohorts of older adults. Research has documented considerable heterogeneity of individual retirement trajectories (e.g., Hans and Moen, 1999), and current and future cohorts of retirees look very different to those retiring in the mid- to late twentieth century who were more likely to have many or all these events occurring simultaneously. Further distinctions, such as whether retirement is viewed as a single life event or a gradual process, or whether retirement is viewed as voluntary or involuntary, are also perhaps more important now than they have been in the past.

Research that seeks to understand the effects of 'retirement' on 'health' needs to be just as clear about what type of retirement is being considered as it is about what type of health outcomes are the object of study. This is for two reasons. First, it is important to be clear about both concepts when thinking about the potential theoretical mechanisms and processes that might link retirement and health. Second, empirical research that seeks causal estimation of the effect of retirement on health has to address the potential endogeneity of the retirement decision, which can only be done with a theoretical discussion of the determinants of retirement set in the context of the nature of the variation available in the data. The nature of this endogeneity, and its empirical implications, will be very different for different conceptualizations of retirement. The same argument would relate to health for those attempting to understand the causal effects of health on retirement.

Theoretical Mechanisms

Leaving aside issues of causality for the time being, a large number of studies have suggested an association between retirement (variously conceptualized) and health, whether defined as physical health, cognitive function, or mental health and well-being. There are a number of processes and mechanisms that could lead to such an association, some of which are reviewed below.

If retirement is conceptualized as an exit from paid work, an understanding of the health implications of such a transition can be informed by consideration of the health risks and protections associated with work as well as the link between income (socioeconomic status) and health. Such effects will clearly differ across individuals according to the nature of the health risks in their jobs and also according to the degree to which preretirement earnings are replaced by pensions and other benefits postretirement. Furthermore, paid work fulfills important psychosocial needs in societies where employment is the norm and is central to individual identity, social roles, and social status, suggesting a withdrawal from paid work may impact negatively on mental health and well-being.

Financial considerations are also important for other reasons since they simultaneously drive the relationship between health and retirement in the other direction. Health status affects retirement both directly, through reduced productivity and wage options (e.g., French, 2005), and indirectly through financial incentives created by health insurance, entitlement to disability, and other related benefits (e.g., French and Jones, 2011). In addition, pension wealth and incentives

 http://dx.doi.org/10.1016/B978-0-08-097086-8.14045-0

have been shown to be a key driver of retirement behavior (e.g., Gruber and Wise, 1999, 2004; French, 2005), and are likely to be correlated with health through unobserved factors that drive both life cycle savings choices and health behaviors across the life course. Extensive literature exist on these effects and are summarized elsewhere (see Lindeboom, 2012 for a recent example) rather than discussed further in what follows.

Selective retirement from work (whether related to health or otherwise) can also be linked to the disengagement theory of aging. Proponents of the disengagement theory view old age as a period during which elderly individuals gradually recede from active involvement in community life that may include retirement from employment (Moody, 2010). Such disengagement is seen as a 'natural' aging process, resulting from declining physical and physiological functioning, when people voluntarily relinquish social responsibilities and expectations in order to engage more fully in personally rewarding leisure activities. Although the loss of social roles can have a negative impact on a person's well-being, this is often counterbalanced by the loss of stressors and responsibilities associated with social roles. Such a framework may be particularly applicable when thinking about retirement from one's primary career or occupation independently of retirement from the labor market entirely.

Furthermore, disengagement enables people to pursue rewarding activities, potentially increasing their well-being as a result. In this sense the process of disengagement leads to a loss of stressful roles, and an increase in well-being activities may lead to positive health benefits for retired people. Drentea (2002), for example, associates retirement with reduced stress and anxiety dependent on the ability to carry out pleasurable activities thereafter. Potential well-being enhancing activities in retirement might include greater opportunities for leisure and physical activity, voluntary activities, cognitively stimulating and learning activities, and having more time for health promotion activities such as health checkups.

Related to this, activity theories suggest that remaining active and engaged in social roles is key to successful aging and positive health and well-being outcomes. Within the context of retirement, this may include the replacement of the work role with other roles that enable the achievement of self-worth, that provide opportunities for social interaction, and that can be an outlet for creative expression. The replacement of lost roles and activities with new ones is necessary in order to maintain a healthy self-concept and a high degree of life satisfaction. Similarly, continuity theory suggests that elderly individuals might try to maintain as much of their earlier identities, activities, and style of life as they can (Moody, 2010). Both theories argue that lower levels of activity in old age are more often the result of health problems and disability than they are the outcome of a process of disengagement. Proponents of both continuity and activity theories focus on the benefits of continued social role involvement for older people as opposed to any specific job or labor market role.

While the above theoretical mechanisms and processes all provide an important way of conceptualizing potential health effects, it seems clear that some potential for heterogeneity should be acknowledged. Certain mechanisms and motivations will be more important for some individuals than for others, whether this is a systematic difference across groups such as gender and class, for example, or within-group heterogeneity. Additionally, the degree to which work, or retirement, provides the type of health benefits or risks suggested in the above theories will vary according to individual circumstances and choices. In reality there are jobs that are protective for health and others that cluster occupational risk factors for ill-health. Similarly there are economic and social circumstances in retirement that are good and bad for health. And there may plausibly be differences in role-related factors that depend on the timing of the retirement relative to the public pension or retirement age. Hence it may be important to examine subjective or benefit-related definitions of retirement as well as those pertaining to specific jobs or labor market activity. At the individual level, therefore, the effects of retirement on health will ultimately depend on which type of job and which type of retirement are relevant for the individual in question.

Empirical Evidence

Given the complexity of the possible theoretical links between retirement and health, particularly when coupled with the possibility of extensive individual heterogeneity in the effects, it is no surprise that, despite an extensive empirical literature on the subject, there is no clear consensus as to the overall effects of retirement on health. This vast majority of the literature documents significant associations between retirement (variously defined) and different aspects of physical health, mental health, and cognitive function. The subset of this literature that provides convincing empirical evidence for the existence of causal relationship from retirement to either good or bad health is much smaller. We begin by summarizing the key associations found in the relevant dimensions described above, and conclude with a briefer discussion of the degree to which these associations are found to be robust when examined in more causal frameworks.

A key review on the evidence on the health effects of work (Waddell and Burton, 2006) concluded that paid employment is necessary for obtaining adequate economic resources and income, which are essential for a person's well-being and health. It is important here, however, to distinguish between overall lifetime levels of employment, labor market participation, or economic resources and levels at any particular age or, more specifically, in older ages.

Waddell and Burton (2006) also highlighted that there is little evidence that continuing paid work at older ages is harmful for health. This view contradicts perceptions of older workers being at increased occupational risk due to declines in their cognitive and physical capacities. Indeed, for those in work, older workers are generally less likely to have accidents than their younger counterparts (Whysall and Ellwood, 2006). While the rate of workplace injury is higher among younger workers, workplace accidents involving older workers tend to result in injuries that are more serious. In terms of absence due to sickness, younger workers tend to be absent more often, but for shorter periods of time, whereas older workers are more likely to be off work for a whole week or more when they are sick (Benjamin and Wilson, 2005).

Westerlund et al. (2010) find an improvement in levels of physical and mental fatigue following retirement, especially among those who were previously working with the burden of long-standing illness. Similarly, Coursolle et al. (2010) finds retirement is associated with a reduction in depression when it is from previously stressful employment, and Mein et al. (2003) find the same association among those who retire from high quality and well-paid employment, suggesting links with both relief from work stressors and ability to enjoy retirement financially. 'Early retirement' from the workforce is commonly shown to be associated with poorer mental health outcomes (Buxton et al., 2005) and levels of self-reported health (Van Solinge, 2007), although in some instances the strength of the association diminishes as normal retirement age is approached (Jokela et al., 2010; Butterworth et al., 2006; Melzer et al., 2004). Retirement is also commonly associated with a decline in cognitive function scores when compared to later-life employment (Adam et al., 2007; Roberts et al. 2011).

Studies of physical activity among older people suggest that they encounter significant barriers to physical activity and that the period of retirement does not necessarily liberate them to become more active. Indeed, there is some evidence that retirement results in a reduction in physical activity from work-related transportation that is not compensated for by an increase in sports participation or an increase in nonsports leisure-time physical activity (Slingerland et al., 2007).

Finally, and relating to potential changes in social networks and social relationships, a study by Szinovacz and Davey (2004) finds men who retire before their partners see an increase in depression, while those with partners who have retired previously and with whom they enjoy shared activities see a reduction.

Most studies of retirement and health have focused on men, presumably because of historically stronger links to the labor market. While women are still more likely to exit the labor force at a younger age compared to men, this is changing in more recent cohorts and particularly so in the countries where female pension and retirement are rising to equal those for men. Where outcomes have been compared by gender, associations between retirement and well-being are often stronger among men with significant detrimental effects of retirement on depression and anxiety among men but not women (Butterworth et al., 2006; Melzer et al., 2004; Szinovacz and Davey, 2004; Kim and Moen, 2002). Much more research on gender differences will need to be done in future as more and more women retirees are observed with work histories covering the full life course.

Empirical studies identifying causal effects have typically tried to exploit the differences between involuntary and voluntary retirees or else used variation in pension or disability benefit eligibility ages or incentives, either across people or over time, in order to control for the endogeneity of retirement decisions. Each method has its strengths and weaknesses and these need to be argued on a case-by-case basis depending on the particular definitions of retirement and health that are being considered. Nevertheless, broad findings are that the strong associations and correlations between retirement and health are far less evident.

With regard to physical health and mortality, the results are particularly mixed. While Tsai et al. (2005) found an association between post-65 mortality rates at early retirement, Hernaes et al. (2013) find no causal effect between retirement and mortality when using pension reforms as an instrument for retirement in Norway. Jokela et al. (2010) find no association between retirement and better physical functioning for involuntary retirements while others find that retirement has a detrimental effect on risk of specific illness such as cardiovascular disease and cancer (Moon et al., 2012; Behncke, 2012), and self-reported health (Behnke, 2012).

Studies seeking to use subjective assessments of general health, either on their own or in comparison to more objective measures, typically identify additional problems in causal analysis due to potential reporting biases and effects (see for example, Bound et al., 2010). Newman (2007) shows differing effects of retirement on subjective and objective health measures in an instrumental variables framework, and while Coe and Zamarro (2011) exploit cross-EU variation in pension ages to show that early retirement may be good for health although, they also show that such effects are only temporary in the case of self-assessed health.

Effects of retirement on broader mental health and cognitive outcomes may be part of the mechanism by which subjective health assessments respond differently to retirement than do more objective measures. Mandal and Roe (2008) for example, find normal age retirement to cause improvements in depression and Charles (2004) finds positive effects on well-being and life satisfaction, as do Calvo et al. (2013). Early involuntary retirement leads to a greater decrease in purchasing of antidepressant medication when compared to normal retirement (Laaksonen et al., 2012; Oksanen et al., 2011), and a study examining the cognitive function of involuntary early retirees also reports poorer outcomes than those who continue to work (Rohwedder and Willis, 2010).

Once more, however, different approaches to the identification of causal effects can lead to differing conclusions, even for the same population over the same period of time. Using variation in pension eligibility ages as an instrument for retirement age, Bonsang et al. (2012) provide evidence of a negative effect of retirement on cognitive functioning. Coe et al. (2012), however, find that there is no evidence of a causal effect when using individual level pension fund early retirement options (sometimes called pension 'windows') as the instrument for retirement outcomes.

Conclusions

The relationship between retirement and health is complex both from a theoretical and an empirical perspective. Effects may be multidimensional and almost certainly differ across individuals. The increasing heterogeneity in retirement trajectories for recent cohorts means that taking into account such complexities has become an important issue. Clarity is needed over the precise timing and nature of the events being studied, whether in relation to retirement or to health, and also over the specific mechanisms by which they may be related. While existing empirical literature has yet to establish any conclusive or systematic causal links from retirement to health, there are certain types of circumstances and situations where effects have been found to exist, and others where their existence can be rejected.

The prospects for future work in this area are particularly good. Recent years have seen increased discourse between the various disciplines engaged in the modeling of health and retirement, and the theoretical models and mechanisms studied by empirical researchers are considerably more interdisciplinary than in the past. More importantly, perhaps, the recent advent of large-scale public-use longitudinal studies covering all aspects of socioeconomic, health, and psychosocial circumstances over long periods of time both before and after retirement age (such as the US Health and Retirement Survey, the English Longitudinal Study of Ageing or the Survey of Health Ageing and Retirement in Europe) means that empirical researchers have an unprecedented evidence base with which to carry out future investigations. The fact that these data sets collect data using comparable measurements and definitions means that international differences in economic, social, and health-care environments can be exploited in order to identify causal relationships as opposed to more simple empirical associations (see, for example, Banks and Smith, 2012).

See also: Aging and Health in Old Age; Aging and the Labor Market; Coping with Job Stress; Health Economics; Occupational Health; Occupational Stress Management; Old Age Work Participation; Pension Policy; Retirement and Encore Adulthood: The New Later Life Course; Retirement, Economics of; Socioeconomic Status and Health; Unemployment and Mental Health.

Bibliography

Adam, S., Bonsang, E., Germain, S., Perelman, S., Unit, N., CREPP, H., 2007. Retirement and Cognitive Reserve: A Stochastic Frontier Approach Applied to Survey Data. HEC-ULg, CREPP working papers, 4.

Banks, J., Smith, J., 2012. International comparisons in health economics: evidence from ageing studies. Annual Reviews of Economics 4, 7–81.

Behncke, S., 2012. Does retirement trigger ill health? Health Economics 21 (3), 282–300.

Benjamin, K., Wilson, S., 2005. Facts and Misconceptions about Age, Health Status and Employability. Health and Safety Laboratory, Buxton. Report HSL/2005/20.

Bonsang, E., Adam, S., Perelman, S., 2012. Does retirement affect cognitive functioning? Journal of Health Economics 31 (3), 490–501.

Bound, J., Stinebrickner, T., Waidmann, S., 2010. Health, economic resources and the work decisions of older men. Journal of Econometrics 156, 106–129.

Butterworth, P., Gill, S.C., Rodgers, B., Anstey, K.J., Villamil, E., Melzer, D., 2006. Retirement and mental health: analysis of the Australian national survey of mental health and well-being. Social Science & Medicine 62 (5), 1179–1191.

Buxton, J.W., Singleton, N., Melzer, D., 2005. The mental health of early retirees. Social Psychiatry and Psychiatric Epidemiology 40, 99–105.

Calvo, E., Sarkisian, N., Tamborini, C.R., 2013. Causal effects of retirement timing on subjective physical and emotional health. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 68 (1), 73–84.

Charles, K.K., 2004. Is retirement depressing?: labor force inactivity and psychological well-being in later life. Research in Labor Economics 23, 269–299.

Coe, N.B., Lindeboom, M., 2008. Does retirement kill you: evidence from early retirement windows. CentER 93.

Coe, N.B., Zamarro, G., 2011. Retirement effects on health in Europe. Journal of Health Economics 30 (1), 77–86.

Coe, N.B., von Gaudecker, H.M., Lindeboom, M., Maurer, J., August 2012. The effect of retirement on cognitive functioning. Health Economics 21 (8), 913–927. http://dx.doi.org/10.1002/hec. 1771. Epub 04.08.11. PMID: 21818822.

Coursolle, K.M., Sweeney, M.M., Raymo, J.M., Ho, J.H., 2010. The association between retirement and emotional well-being: does prior work–family conflict matter? The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 65 (5), 609–620.

Drentea, P., 2002. Retirement and mental health. Journal of Aging and Health 14 (2), 167–194.

Feldman, D.C., 1994. The decision to retire early: a review and conceptualisation. Academy of Management Review 19, 285–311.

French, E., 2005. The effects of health, wealth, and wages on labor supply and retirement behavior. Review of Economic Studies 72, 395–427.

French, E., Jones, J.B., 2011. The effects of health insurance and self-insurance on retirement behavior. Econometrica 79 (3), 693–732.

Gruber, J., Wise, D. (Eds.), 1999. Social Security and Retirement around the World. Chicago University Press.

Gruber, J., Wise, D. (Eds.), 2004. Social Security Programs and Retirement around the World: Micro-estimation. Chicago University Press.

Hans, S.-K., Moen, P., 1999. Clocking out: temporal patterning of retirement. American Journal of Sociology 105 (1), 191–236.

Hernaes, E., Markussen, S., Piggott, J., Vestad, O., 2013. Does retirement age impact mortality? Journal of Health Economics 32, 586–598.

Jokela, M., Ferrie, J.E., Gimeno, D., Chandola, T., Shipley, M.J., Head, J., Kivimäki, M., 2010. From midlife to early old age: health trajectories associated with retirement. Epidemiology (Cambridge, MA) 21 (3), 284.

Kim, J.E., Moen, P., 2002. Retirement transitions, gender, and psychological well-being a life-course, ecological model. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 57 (3), 212–222.

Laaksonen, M., Metsä-Simola, N., Martikainen, P., Pietiläinen, O., Rahkonen, O., Gould, R., Lahelma, E., 2012. Trajectories of mental health before and after old-age and disability retirement: a register-based study on purchases of psychotropic drugs. Scandinavian Journal of Work Environment and Health 38 (5), 409.

Lazear, E.P., 1987. Retirement form the labour force. In: Ashenfelter, O., Layard, R. (Eds.), Handbook of Labor Economics, vol. 1. Elsevier, pp. 305–355.

Lindeboom, M., 2012. Health and work of older workers. In: Jones, A.M. (Ed.), The Elgar Companion to Health Economics, second ed. Edward Elgar.

Mandal, B., Roe, B., 2008. Job loss, retirement and the mental health of older Americans. The Journal of Mental Health Policy and Economics 11, 167–176.

Mein, G., Martikainen, P., Hemingway, H., Stansfeld, S., Marmot, M., 2003. Is retirement good or bad for mental and physical health functioning? Whitehall II longitudinal study of civil servants. Journal of Epidemiology and Community Health 57 (1), 46–49.

Melzer, D., Buxton, J., Villamil, E., 2004. Decline in common mental disorder prevalence in men during the sixth decade of life. Social Psychiatry and Psychiatric Epidemiology 39 (1), 33–38.

Moody, Harry R., 2010. Aging: Concepts and Controversies, sixth ed. Pine Forge Press, Thousand Oaks, CA.

Moon, J.R., Glymour, M.M., Subramanian, S.V., Avendaño, M., Kawachi, I., 2012. Transition to retirement and risk of cardiovascular disease: prospective analysis of the US health and retirement study. Social Science & Medicine 75 (3), 526–530.

Newman, K., 2007. Quit your job and get healthier? The effect of retirement on health. Journal of Labor Research 29 (2), 177–201.

Oksanen, T., Vahtera, J., Westerlund, H., Pentti, J., Sjösten, N., Virtanen, M., Kivimäki, M., 2011. Is retirement beneficial for mental health? Antidepressant use before and after retirement. Epidemiology (Cambridge, MA) 22 (4), 553.

Roberts, B.A., Fuhrer, R., Marmot, M., Richards, M., 2011. Does retirement influence cognitive performance? The whitehall II study. Journal of Epidemiology and Community Health 65 (11), 958–963.

Rohwedder, S., Willis, R.J., 2010. Mental retirement. Journal of Economic Perspectives 24 (1), 119–138.

Slingerland, A.S., van Lenthe, F.J., Jukema, J.W., Kamphius, C.B.M., Looman, C., Giskes, K., Slingerland, A.S., van Lenthe, F.J., Jukema, J.W., Kamphius, C.B.M., Looman, C., Giskes, K., et al., 2007. Aging, retirement, and changes in physical activity: prospective cohort findings from the GLOBE study. American Journal of Epidemiology 165, 1356–1363.

Szinovacz, M.E., Davey, A., 2004. Honeymoons and joint lunches: effects of retirement and spouse's employment on depressive symptoms. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 59 (5), 233–245.

Tsai, S.P., Wendy, J.K., Donnelly, R.P., de Jong, G., Ahmed, F.A., 2005. Age at retirement and long term survival of an industrial population cohort study. British Medical Journal 331 (7523), 995.

Turner, Bryan S., 1989. Ageing, status politics and sociological theory. The British Journal of Sociology 40 (4), 588–606.

Van Solinge, H., 2007. Health change in retirement a longitudinal study among older workers in the Netherlands. Research on Aging 29 (3), 225–256.

Waddell, G., Burton, A.K., 2006. Is Work Good for Your Health and Well-being? The Stationery Office, London.

Westerlund, H., Vahtera, J., Ferrie, J.E., Singh-Manoux, A., Pentti, J., Melchior, M., Kivimäki, M., 2010. Effect of retirement on major chronic conditions and fatigue: French GAZEL occupational cohort study. BMJ: British Medical Journal 341.

Whysall, Z., Ellwood, P., 2006. HSE Horizon Scanning Intelligence Group Demographic Study Report. http://www.hse.gov.uk/horizons/demographics.pdf.

Retirement, Economics of

Courtney C Coile, Wellesley College, Wellesley, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.L. Gustman, T.L. Steinmeier, volume 19, pp. 13289–13294, © 2001, Elsevier Ltd.

Abstract

This article summarizes theory and evidence regarding the economics of retirement, a field that has experienced a surge of new research during the past two decades, due in part to the availability of new longitudinal data. Topics reviewed include the definition of retirement, retirement trends and patterns, and economic factors affecting retirement, including public and private pensions, wealth and savings, health and health insurance, and labor demand. The findings on these topics will be of use to policy makers around the world as they confront the opportunities and challenges posed by an aging population.

Retirement is often thought of as life's third act, coming after one's childhood and work life. Yet retirement is actually a fairly modern concept – in 1880, more than three-quarters of U.S. men, aged 65 and older, were in the labor force (Costa, 1998), meaning that many men worked until death. Since that time, rising incomes and the introduction of public and private pensions have made it possible for people to leave the labor force at younger ages. Increasing longevity and better health at older ages have also allowed people to spend more years in retirement and to receive more enjoyment from this phase of life.

As retirement has become a more significant phase of life, so too has it become the focus of more economic research. Over the past 20 years, more than 2000 articles on the economics of retirement have been published in academic journals indexed by *EconLit*. Much of the recent U.S. research makes use of the Health and Retirement Study (HRS), a longitudinal panel study of individuals over the age of 50 that began in 1992 and is designed to explore changes in labor force participation and health in late middle age and beyond. The success of the HRS has encouraged the creation of similar surveys in other countries, including the Survey of Health, Ageing and Retirement in Europe (SHARE) as well as aging studies in England, Ireland, Israel, Mexico, Japan, Korea, China, and India. With this rich data at their disposal, researchers are more active than ever in exploring issues in the economics of aging, including the economics of retirement.

Defining Retirement

While retirement may seem like a straightforward concept, it is surprisingly difficult to define. Perhaps the simplest option is to define retirement as occurring when an individual stops working for pay. This definition works well when a worker with a full-time career job that he or she has held for many years leaves this job and never works again. However, a majority of U.S. workers with career jobs retire gradually, in stages, often working for a period of time at a bridge job that offers a part-time schedule, more flexibility, or fewer responsibilities (Cahill et al., 2006). While these individuals are considered workers under the standard retirement definition, they may have more in common with fully retired individuals than they do with full-time workers.

Alternatively, retirement could be defined as occurring when an individual leaves his or her career job. But not all workers have career jobs, and it is not clear how to classify workers who switch to part-time status at their same job. A third option is to allow individuals themselves to define whether or not they are retired. This sidesteps some issues, but raises the concern that individuals with the same level of labor force participation could be classified differently. A final option is to focus on the initial claim of public or private pension benefits. While obviously an important milestone, benefit receipt need not be related to work, since individuals may be able to claim pension benefits while they are still working or to delay benefit receipt after they have stopped working.

In the end, most analysts define retirement as withdrawal from the labor force, as this is conceptually straightforward and easy to implement using most survey data. Analysts also tend to treat retirement as an absorbing state and ignore the possibility of labor force reentry, though Maestas (2010) shows that a sizeable share of U.S. workers does return to the labor force after retirement.

Retirement Trends and Patterns

The period since World War II has been characterized by a dramatic decline in the labor force participation of older men in the U.S. Data from the U.S. Bureau of Labor Statistics shows that the share of men aged 55–64 in the labor force fell from 90% in 1948 to 66% in the mid-1990s, indicating that men were retiring earlier. Since then, labor force participation in this age group has risen slowly but steadily to its current value of 70%. For older women in the U.S., labor force participation has risen continuously over the past 60 years, from 24% in 1948 to 59% in 2014. The societal trend of greater labor force participation by women of all ages appears to have swamped any movement toward earlier retirement for women.

A second notable aspect of retirement behavior is the pattern of retirement by age. Coile and Levine (2010) use data from the Current Population Survey to calculate the retirement hazard, or probability that a worker still in the labor force at a given age will exit the labor force that year, for U.S. workers.

When workers are in their mid-50s, the retirement hazard is about 5%. This rises slowly with age then doubles at age 62, the age at which Social Security benefits are first available. The hazard rate drops somewhat at ages 63 and 64 then spikes again at age 65. This is the age of eligibility for Medicare, a government health insurance program for the elderly, and was until recently also the normal retirement age for Social Security.

Though perhaps less well known, there are distinct differences in retirement rates by education, also documented by Coile and Levine (2010). While the retirement hazard by age follows the same pattern for all education groups, the retirement rate at every age is higher for less educated workers. Averaging across ages 55–69, the annual retirement rate in 2007, just prior to the recent recession, was 5% for college graduates, 7% for high-school graduates, and 9% for high-school dropouts, indicating earlier retirement by the less educated.

Retirement behavior in other developed countries displays many of the same patterns seen in the U.S. Gruber and Wise (1999) tabulate data for nine European countries as well as the U.S., Canada, and Japan. They document a decline in men's labor force participation in Europe during the postwar period that was similar to, if not steeper than, that in the U.S. By 1995, only 20–30% of men aged 60–64 were still in the labor force in Belgium, France, Italy, and the Netherlands, as compared with 50 to 60% in the U.S., UK, and Sweden; labor force participation was highest in Japan, at 75%. The retirement hazard by age in other countries is also similar to that in the U.S. but exhibits even larger spikes at the early and normal ages for public pension plans, with as many as 60–80% of remaining workers retiring when they reach these critical ages, as opposed to about 20% in the U.S.

Modeling Retirement Behavior

The economist envisions that a rational individual approaches the retirement decision by comparing the utility he or she would receive by working this year to the utility received by retiring. In such a model, wages will matter, as will potential sources of retirement income, such as public and private pensions. Economists generally believe that leisure time is a desirable good, so there will be additional utility associated with being retired (or, to put it differently, disutility from work), which may depend on the individual's health and age.

In addition, it may be appropriate to think of retirement in a household rather than an individual context, as the value of being retired may be greater if one's spouse is also retired. Naturally retirement will depend on all of the household's resources, not just pension benefits, and it may be appropriate to recognize the possibility that wealth and retirement are determined simultaneously, with a household saving more if its members wish to retire early, for example. If health insurance markets are imperfect and employment is a key means of obtaining insurance, this may also affect retirement decisions. Finally, separate from an individual's potential desire to continue working, the demand for labor from employers may affect an individual's ability to do so.

It should be evident from the preceding discussion that it will be very difficult to construct a single model of retirement that incorporates all of the factors that are potentially important and deals with each in its full complexity. Instead, researchers have tended to focus on one key piece at a time, such as public pensions, and model its effect carefully. Some researchers favor a structural approach, where the researcher posits a specific form for the individual or household utility function and uses data to estimate the parameters that define, for example, how patient the individual is when considering tradeoffs between present and future income. One advantage of this method is that the estimated parameters can be used to simulate the effect of policies outside the scope of real world experience, such as how retirement behavior would change if public pensions were eliminated. Other researchers favor making use of natural experiments – policy changes or other real-world events that have exposed groups of individuals to different conditions, such as different early and normal retirement ages in the pension plan – to learn the effect of these factors on retirement. An advantage of the latter approach is that it does not rely on correctly specifying the form of the utility function and its key assumptions are transparent – generally this boils down to demonstrating that the groups facing different conditions are either identical or can be treated as such after employing strategies to control for any differences.

Economic Factors Affecting Retirement

Public and Private Pensions

Some vocabulary is necessary to discuss the effect of pensions on retirement. A pension plan can be either a defined benefit (DB)- or defined contribution (DC)-type plan. In a DB plan, benefits are determined by a formula – for example, final earnings times years worked times 2%, so a worker with 30 years of experience receives a benefit of 60% of final salary – though public pension plans generally use complex formulas that depend on the worker's entire earnings history. In a DC plan, it is the contribution that is specified – for example, a worker might be required to contribute 10% of wages to the plan. The benefit in a DC plan depends on contributions, asset allocation and investment returns, and asset draw-down in retirement.

A pension plan can also either be fully funded, where a worker's contributions are saved to pay his or her own benefits, or pay-as-you-go, where today's contributions pay for the benefits of today's retirees. A DC plan is always fully funded, while a DB plan may be either fully funded or pay-as-you-go. The U.S. and most developed countries have pay-as-you-go, DB public pension plans, though some countries have more recently added a DC tier. Employer-sponsored private pension plans in the U.S. and elsewhere have shifted over time from DB- to DC-type plans (Munnell, 2006).

In theory, a public pension plan may affect retirement behavior if the plan results in income or substitution effects. An income effect occurs if retirement benefits are greater or less than payroll tax contributions plus a market rate of return. If benefits are greater than taxes, the individual may retire earlier, choosing to spend some of the surplus income on leisure. A substitution effect occurs if the worker is able to increase the

total value of lifetime benefits by continuing to work; if so, the pension plan effectively augments the wage and the worker may choose to work longer as a result.

Public pension systems often feature both intercohort and intracohort redistribution that create income effects. The introduction of programs like Social Security in the U.S. as pay-as-you-go plans advantaged early beneficiaries, who did not contribute over their full working lives, at the expense of later cohorts (Leimer, 1999). Within a cohort, low earners often receive higher benefits relative to contributions than do high earners because many countries incorporate progressivity into the benefit formula. Dependent spouse and survivor benefits also create intracohort redistribution, as a married, one-earner family receives more benefits than a single worker with the same payroll tax contributions.

Even in the absence of redistribution, an income effect may occur due to forced saving. Consider two types of individuals who save little for retirement: a rational, forward-looking individual who gets little utility from retirement consumption and a myopic individual who fails to plan. For both types, the pension plan raises retirement income, leading to earlier retirement. Of course, since pension income is not available until the early retirement age and workers cannot borrow against this income, they may be unable to retire until they reach this age.

The process by which a public pension plan creates substitution effects is complex. By working one more year, the worker makes payroll tax contributions and gives up a year of benefits (once past the early retirement age). However, the extra year of work is incorporated into the benefit calculation, which may raise the benefit amount. Further, the benefit amount is often raised for delaying benefit claim beyond the early eligibility age. If that adjustment is actuarially fair, lifetime benefits are similar regardless of claiming age, but if it is more (or less) than fair, this aspect of the system rewards (or penalizes) additional work. Whether the pension plan as a whole rewards additional work depends on the net of all these factors, and may vary across individuals depending on their earnings history and other characteristics.

Though not always included in standard calculations of the return to work, a final potentially important factor is that public pension plans offer a real annuity, as beneficiaries generally receive inflation-adjusted benefits that last as long as they live. With private annuity markets plagued by adverse selection (Finklestein and Poterba, 2004), workers may not have access to annuities outside the plan and may delay retirement in order to purchase a larger annuity.

With employer-sponsored private pensions, there is no redistribution aspect, since the pension is simply part of the worker's total compensation package. But DB-type private pensions typically do contain strong incentives to work to particular ages – Stock and Wise (1990) note that many pensions provide a large bonus if the employee works to the plan's early retirement age and a substantial inducement to retire thereafter. As firms shift from DB to DC plans, private pensions may have less of an impact on retirement decisions, since DC plans do not include the same types of incentives to work to or retire at particular ages.

Economists have been very interested in estimating the effect of public pensions on retirement, in part because the retirement patterns discussed above are highly suggestive of a relationship. The postwar decline in men's labor force participation occurred over the same period that workers were becoming eligible for public pensions in the U.S. and Europe and that benefits were becoming more generous. The reversal of this labor force decline in the U.S. and other countries (Coile et al., 2014b) over the past 20 years is also consistent with a role for pensions, as many countries trimmed their programs during this period. Moreover, the large spikes in the retirement hazard at the early and normal retirement ages of public pension plans do not easily lend themselves to explanations other than the obvious one, that public pensions affect retirement. Burtless and Moffitt (1986) show that the spike in retirement at age 62 in the U.S. only emerged once age 62 became the early eligibility age for Social Security. Mastrobuoni (2009) finds that as the normal retirement age for U.S. workers was raised from 65 to 66, the retirement age for affected cohorts rose by about half as much.

Researchers have also tried to directly estimate the effect of pensions on retirement, using microdata sets to compute the return to work for individual workers and making use of policy changes or differences in earnings histories or other characteristics that create variation in pension incentives across individuals. The more recent work in this literature has recognized that it is not sufficient to look at the single-year accrual, or change in lifetime pension benefits arising from an additional year of work. In the context of employer-provided DB pensions, Stock and Wise (1990) note that by working this year, the worker is effectively retaining the option of retiring in a future year with potentially higher benefits; they develop an option value model in which the worker compares the utility of retiring today to that available at all future retirement dates. Focusing on the U.S. Social Security system, Coile and Gruber (2007) show that workers are less likely to retire when they have greater gains from continued work, as measured by the option value and by a companion peak value measure that captures the financial (as opposed to utility) gain from continued work.

International comparisons provide further evidence of the effect of public pensions on retirement. Gruber and Wise (1999) calculate the financial incentive for a typical worker in each of a dozen developed countries to continue working at ages 55–69, given the features of their country's public pension plan. They show that this measure varies markedly across countries, with large penalties on additional work in Belgium, France, Italy, and the Netherlands. This correlates strongly with labor force participation rates of older men – indeed, 80–85% of the differences in participation across countries can be explained by this measure. In subsequent work, Gruber and Wise (2004) use microdata to estimate the return to work for individual workers, and show that the option value and peak value measures are key predictors of retirement in virtually all countries in their sample.

Finally, studies that estimate structural models of retirement are also relevant. Pensions enter these models as a key source of retirement income. The model is validated based on its ability to match observed retirement behavior, and model estimates may be then used to infer the role of pensions

through simulations. Gustman and Steinmeier (1986) estimate a structural retirement model that replicates the peaks in the retirement hazard at ages 62 and 65 for U.S. workers. In Gustman and Steinmeier (2005), they predict that raising the Social Security early entitlement age to 64 would cause many people to delay retirement from age 62 to 64. By contrast, French (2005) estimates that moving the early entitlement age would have little effect on retirement, but that eliminating the earnings test (which prevents individuals from receiving benefits while working) would lead individuals to work longer.

Wealth and Savings

The retirement decision is likely to depend on all of the worker's financial resources, not only on pension benefits. Economists generally believe that leisure is a normal good, that is, a good for which demand increases as income rises, and thus expect wealthier individuals to retire earlier, ceteris paribus. One challenge in testing this hypothesis empirically is that wealth is determined by the savings decisions individuals make over time. If wealthier workers retire earlier, this might reflect the fact that these workers wanted to retire early and saved accordingly rather than a causal effect of wealth on retirement. More generally, wealthy individuals may differ from less wealthy individuals in ways that are unobservable to the researcher, creating bias in the estimate.

To surmount these challenges, the primary strategy adopted by researchers who favor a natural experiments approach is to identify unanticipated changes to wealth and test whether these wealth shocks are associated with an increased probability of retirement. Researchers have made use of several sources of wealth shocks. Imbens et al. (2001) show that large lottery gains lead to a reduction in labor supply, particularly for those near retirement age. Brown et al. (2010) find that the receipt of an inheritance increases the probability of retirement, particularly when the inheritance is unexpected and thus more likely to represent a true shock to wealth.

Another source of wealth shocks is the boom–bust cycles in the stock market over the past 20 years, which have created large, unexpected fluctuations in the value of equities. Evidence from this literature is more mixed. Focusing on the late 1990s and early 2000s, Hurd et al. (2009) find no evidence that workers in households with large stock market gains retired earlier than expected, as compared with those without such gains. Coile and Levine (2006) find no evidence that changes in the stock market drive aggregate trends in labor supply, which they attribute to the fact that few workers have large stock holdings. However, Coile and Levine (2011a) find that college-educated workers age 62 and above – who are more likely to own stocks and may be more easily persuaded to retire – do reduce retirement in response to long-term declines in the value of stocks.

On the structural side, researchers who estimate structural models of retirement have augmented their models to incorporate savings decisions as well. Rust and Phelan (1997), French (2005), and Van der Klaauw and Wolpin (2008) are examples in this vein. These authors continue to primarily use their estimates to assess the effect of public pensions on retirement, but do so in a more realistic setting where workers consider retirement and savings decisions simultaneously.

Health and Health Insurance

At the simplest level, a decline in health status is expected to increase the disutility of work and therefore increase the worker's probability of retirement. There may be other effects as well – a decline in health status may lead the worker to update his or her mortality expectations, require an increase in medical expenditures, or change how much the worker values consumption in retirement. Any of these could also affect the retirement decision.

Researchers have used several approaches to assess the effect of health on retirement. One method is to include self-reported health status or work limitations in the retirement model. Studies using this method, such as Diamond and Hausman (1984), find large effects of health on retirement. However, this method is subject to the concern that self-reported health is a subjective measure, as individuals' judgments of what constitutes poor health may vary substantially. Moreover, this measure may be influenced by retirement status if people justify being retired by reporting a health problem. A second method is to use objective measures of health, such as diagnoses of particular diseases. Studies using this method, such as Anderson and Burkhauser (1985), find smaller effects, but these measures are only imperfectly correlated with working capacity. A third method is to combine subjective and objective data. However, Bound (1991) documents that the third approach can lead to bias, and also confirms concerns about the first two methods.

A final approach is to focus on negative shocks to health. As in the wealth shocks literature, the appeal of this approach is that it exploits the arrival of unexpected new information. By focusing on changes in health, this approach avoids the concern that people who are generally in worse health may differ from those in better health in ways that are unobservable to the researcher. McClellan (1998) uses this approach and finds that workers who experience a health shock such as a heart attack or new cancer diagnosis are much more likely to retire, particularly when the shock is accompanied by a large decline in functioning.

Access to health insurance may also play a role in retirement decisions, particularly in the U.S., where most individuals rely on their employer or a family member's employer for access to insurance. Madrian et al. (1994) and Rogowski and Karoly (2000) conclude that workers are much more likely to retire when their employer offers health insurance for retirees. Gruber and Madrian (1996) find that continuation of coverage mandates that allow retirees to purchase insurance through a previous employer for a period of time after leaving the firm raise retirement rates substantially, particularly for workers close to Medicare eligibility.

On the structural side, a number of authors have explored the role of health insurance. Rust and Phelan (1997) find that Medicare plays an important role in explaining the spike in retirement at age 65. French and Jones (2011) estimate that raising the Medicare eligibility age would lead workers to work longer, though Blau and Gilleskie (2006) suggest that the effect would be relatively small.

Retirement in a Family Context

While the preceding discussion largely treats retirement as an individual decision, many workers today are part of married, dual-career couples. It seems likely that workers in such families might make retirement decisions jointly or at least that each spouse might be influenced to some extent by the other spouse's decision, for several reasons. First, individuals in a married couple have access to common financial resources. Second, individuals who are married may have similar preferences – for example, both may be patient with respect to tradeoffs between present and future income. Third, each spouse's enjoyment of his or her leisure time may depend on whether the other spouse is present. One key empirical fact motivating the literature on couples' retirement is that many spouses do retire together – Coile (2004) finds that in approximately one-third of working couples, both spouses retire within 1 year of each other.

Several authors estimate structural models of joint retirement by husbands and wives, including Hurd (1990) and Gustman and Steinmeier (2000). These authors employ different models, but all find that complementarity of leisure – the idea that leisure time is more enjoyable if one's spouse is present – is a key factor driving retirement decisions. Other studies, such as Coile (2004) and Baker (2002) test for joint decision-making more indirectly, by exploring whether public pension incentives for one spouse also affect the retirement behavior of the other spouse; both find that they do.

Labor Demand

The discussion to this point has focused on factors affecting the individual's willingness to supply labor, including public and private pension incentives, wealth, health and health insurance, and a spouse's retirement status. Yet one should not assume that all older individuals who wish to continue working will be able to do so. During the Great Recession, nearly one in six U.S. workers reported losing a job (Farber, 2011). While older workers experienced lower rates of job loss, displaced older workers also experienced longer spells of unemployment and lower reemployment rates, as compared with their younger peers. Even during more normal economic times, there will be older job losers who struggle to find new work.

The effect of labor market conditions on retirement is a relatively new branch of the retirement literature. Chan and Stevens (2001) find that older U.S. workers who experience a job loss have substantially lower employment rates and earnings relative to similar nondisplaced workers. Coile and Levine (2007) show that downturns in the U.S. labor market increase retirement transitions, particularly for workers who are eligible to claim Social Security benefits. Using industry-level variation in employment in Sweden, Hallberg (2011) similarly finds that retirement is countercyclical. Late-career job loss has implications beyond retirement – Coile and Levine (2011b) show that experiencing weak labor market conditions near the age of retirement lowers retiree income, while Sullivan and von Wachter (2009) and Coile et al. (2014a) find that job loss and recessions are associated with higher long-term mortality.

Conclusion

Around the world, many countries are experiencing dramatic population aging as a result of rising longevity and falling fertility rates. This phenomenon has led to a surge of interest in the economics of aging, including the economics of retirement. The introduction of surveys such as the HRS in the U.S. and sister studies in numerous countries has enabled researchers to make tremendous strides in understanding retirement behavior and offers the promise of much exciting work yet to come. The next phase of research in the economics of retirement will be of great value to policy makers facing the difficult challenge of promoting the financial security of the elderly while ensuring that the public programs that support them are financially sustainable.

See also: Health Economics; Labor Market Policy; Labor Supply; Pension Policy; Welfare and Work.

Bibliography

Anderson, Kathryn H., Burkhauser, Richard V., 1985. The retirement-health nexus: a new measure of an old puzzle. Journal of Human Resources 20 (3), 315–330.

Baker, Michael, 2002. The retirement behavior of married couples evidence from the spouse's allowance. Journal of Human Resources 37 (Winter), 1–34.

Blau, David M., Gilleskie, Donna, 2006. Health insurance and retirement of married couples. Journal of Applied Econometrics 21 (7), 935–953.

Bound, John, 1991. Self-reported objective measures of health in retirement models. Journal of Human Resources 26, 103–138.

Brown, Jeffrey R., Coile, Courtney C., Weisbenner, Scott J., 2010. The effect of inheritance receipt on retirement. The Review of Economics and Statistics 92 (2), 425–434.

Burtless, Gary, Moffitt, Robert, 1986. Social security, earnings tests, and age at retirement. Public Finance Quarterly 14, 3–27.

Cahill, Kevin E., Giandrea, Michael D., Quinn, Joseph F., 2006. Retirement patterns from career employment. The Gerontologist 46 (4), 514–523.

Chan, Sewin, Stevens, Ann Huff, 2001. Job loss and employment patterns of older workers. Journal of Labor Economics 19 (2), 484–521.

Coile, Courtney C., 2004. Retirement incentives and couples' retirement decisions. The B.E. Press Journal of Economic Analysis & Policy 4 (1), 1–30.

Coile, Courtney C., Gruber, Jonathan, 2007. Future social security entitlements and the retirement decision. The Review of Economics and Statistics 89 (2), 234–246.

Coile, Courtney C., Levine, Phillip B., 2011a. The market crash and mass layoffs: how the current economic crisis may affect retirement. The B.E. Press Journal of Economic Analysis & Policy 11 (1), 1–22.

Coile, Courtney C., Levine, Phillip B., 2011b. Recessions, retirement, and social security. American Economic Review 101 (3), 23–28.

Coile, Courtney C., Levine, Phillip B., 2010. Reconsidering Retirement: How Losses and Layoffs Affect Older Workers. Brookings Institution Press, Washington, DC.

Coile, Courtney C., Levine, Phillip B., 2007. Labor market shocks and retirement: do government programs matter? Journal of Public Economics 91 (10), 1902–1919.

Coile, Courtney C., Levine, Phillip B., 2006. Bulls, bears, and retirement behavior. Industrial and Labor Relations Review 59, 408–429.

Coile, Courtney C., Levine, Phillip B., McKnight, Robin, 2014a. Recessions, older workers, and longevity: how long are recessions good for you? American Economic Journal: Economic Policy 6 (3), 92–119.

Coile, Courtney C., Milligan, Kevin S., Wise, David A., 2014b. Introduction and summary. In: Wise, David A. (Ed.), Social Security Programs and Retirement around the World: Disability Insurance Programs and Retirement. University of Chicago Press, Chicago.

Costa, Dora L., 1998. The Evolution of Retirement: An American Economic History, 1880–1990. University of Chicago Press, Chicago.

Diamond, Peter, Hausman, Jerry, 1984. The retirement and unemployment behavior of older men. In: Burtless, Gary, Aaron, Henry (Eds.), Retirement and Economic Behavior. Brookings Institution Press, Washington, DC, pp. 97–129.

Farber, Henry S., 2011. Job Loss in the Great Recession: Historical Perspective from the Displaced Workers Survey, 1984–2010. NBER. Working Paper #17040.
Finklestein, Amy, Poterba, James, 2004. Adverse selection in insurance markets: policyholder evidence from the U.K. annuity market. Journal of Political Economy 112 (1), 183–208.
French, Eric, 2005. The effects of health, wealth, and wages on labor supply and retirement behavior. The Review of Economic Studies 72 (2), 395–427.
French, Eric, Jones, John Bailey, 2011. The effects of health insurance and self-insurance on retirement behavior. Econometrica 79 (3), 693–732.
Gruber, Jonathan, Madrian, Brigitte, 1996. Health insurance and early retirement: evidence from the availability of continuation coverage. In: Wise, David A. (Ed.), Advances in the Economics of Aging. University of Chicago Press, Chicago, pp. 115–143.
Gruber, Jonathan, Wise, David A., 2004. Social Security Programs and Retirement around the World: Microestimation. University of Chicago Press, Chicago.
Gruber, Jonathan, Wise, David A., 1999. Introduction and summary. In: Gruber, Jonathan, Wise, David A. (Eds.), Social Security and Retirement around the World. University of Chicago Press, Chicago, pp. 1–35.
Gustman, Alan L., Steinmeier, Thomas L., 2005. The social security early entitlement age in a structural model of retirement and wealth. Journal of Public Economics 89 (2–3), 441–463.
Gustman, Alan L., Steinmeier, Thomas L., 2000. Retirement in dual-career families: a structural model. Journal of Labor Economics 18 (3), 503–545.
Gustman, Alan L., Steinmeier, Thomas L., 1986. A structural retirement model. Econometrica 54 (3), 555–584.
Hallberg, Daniel, 2011. Economic fluctuations and retirement of older employees. Labour 25 (3), 287–307.
Hurd, Michael D., Reti, Monika, Rohwedder, Susann, 2009. The effect of large capital gains or losses on retirement. In: Wise, David A. (Ed.), Developments in the Economics of Aging. University of Chicago Press, Chicago.
Hurd, Michael D., 1990. The joint retirement decision of husbands and wives. In: Wise, David A. (Ed.), Issues in the Economics of Aging. University of Chicago Press, Chicago, pp. 231–258.
Imbens, Guido W., Rubin, Donald B., Sacerdote, Bruce, 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. American Economic Review 91 (4), 778–794.
Leimer, Dean R., 1999. Lifetime redistribution under the social security progam: a literature synopsis. Social Security Bulletin 62 (2), 43–51.
Madrian, Brigitee, Burtless, Gary, Gruber, Jonathan, 1994. The effect of health insurance on retirement. Brookings Papers on Economic Activity 1994 (1), 181–252.
Maestas, Nicole, 2010. Back to work: expectations and realizations of work after retirement. Journal of Human Resources 45 (3), 718–748.
Mastrobuoni, Giovanni, 2009. Labor supply effects of the recent social security benefit cuts: empirical estimates using cohort discontinuities. Journal of Public Economics 93 (11–12), 1224–1233.
McClellan, Mark B., 1998. Health events, health insurance, and labor supply: evidence from the health and retirement survey. In: Wise, David A. (Ed.), Frontiers in the Economics of Aging. University of Chicago Press, Chicago.
Munnell, Alicia H., 2006. Employer-sponsored plans: the shift from defined benefit to defined. The Oxford Handbook of Pensions and Retirement Income 13, 359–380.
Rogowski, Jeannette, Karoly, Lynn, 2000. Health insurance and retirement behavior: evidence from the health and retirement survey. Journal of Health Economics 19 (4), 529–539.
Rust, John, Phelan, Christopher, 1997. How social security and medicare affect retirement behavior in a world of incomplete markets. Econometrica 65 (4), 781–831.
Stock, James, Wise, David, 1990. Pensions, the option value of work, and retirement. Econometrica 58, 1151–1180.
Sullivan, Daniel, von Wachter, Till, 2009. Job displacement and mortality: an analysis using administrative data. The Quarterly Journal of Economics 124 (3), 1265–1306.
Van der Klaauw, Wilbert, Wolpin, Kenneth I., 2008. Social security and the retirement and savings behavior of low-income households. Journal of Econometrics 145 (1), 21–42.

Retrieval

Asaf Gilboa, Baycrest, Toronto, ON, Canada; and University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Retrieval refers to a set of processes that support the recovery of stored memories. At its heart is the ecphoric process that involves an interaction between memory cues and memory traces, and the conversion of ecphoric information into conscious awareness. Retrieval also involves a set of preecphoric processes that support the search for appropriate memory cues and memory traces, and a set of postecphoric monitoring and control processes that evaluate the veracity and appropriateness of retrieved memories and endorse report decisions.

General Introduction

Retrieval is the process by which stored memories are recovered in response to memory cues that are either external or internally generated. The product of a successful retrieval process is an awareness of the target memory or information. The recovery of stored information is a basic component of many cognitive and behavioral activities. It is particularly central for functions that depend on semantic and episodic memory. For example, in the semantic domain retrieval is key to functions such as word finding during a conversation or the interpretation of written language; in the episodic memory domain, retrieval processes are required for accomplishments such as remembering where you placed your keys last night, or recollecting the night of your graduation.

The most straightforward construal of successful retrieval is that a *memory cue* activates or triggers the preexisting *memory representation* (*trace*, *engram*). Two factors that critically predict successful retrieval by this construal are the quality of the retrieval cue and the strength of the memory trace. However, theoretical models of retrieval processes have moved away from the idea that during retrieval memory traces or engrams are just 'triggered,' 'activated,' or 'brought to life'; instead, most theories put forward models in which memory cues and memory representations interact in complex ways that allow the construction/reconstruction of the experience of remembering. Thus, the success of the retrieval process depends on numerous factors, many of which may be independent of how well the information was encoded, the strength of the memory trace or the quality of the cue (**Figure 1**).

In what follows, I will outline some of these factors, as well as the processes that constitute retrieval. The overall framework presented below classifies retrieval into two primary kinds: (1) direct retrieval that involves an automatic, 'ballistic' process of the interaction between a memory cue and a memory representation and (2) indirect or strategic retrieval in which processes akin to problem-solving procedures are invoked in order to arrive at a solution to a memory problem. Both kinds of retrieval may involve postretrieval monitoring and verification processes that can be automatic or more effortful. These monitoring processes will be discussed together with the strategic retrieval components, as they tend to be invoked more during indirect retrieval. Finally, after describing this process-component model of retrieval, some discussion will be devoted to reconstruction as a central neurocognitive concept that is at the core of retrieval as well as other important constructive neurocognitive processes.

Two Routes Retrieval

At its core, retrieval involves an interaction between memory cues and memory traces or engrams. *Ecphory* is a useful term used to describe this interaction. It was originally coined by Richard Semon (1921[1904]) to denote a simple triggering or activation of an otherwise dormant memory trace (Semon, 1921[1904]). Tulving then reintroduced the term Ecphory to denote the process of activation of an engram by retrieval information provided by retrieval cues (Tulving, 1983a,b). In Tulving's model, a key component in the retrieval process is the association or correlation between the information provided by the memory cue and the information stored by the memory engram. Close correlation between these two sources of information may lead to the conversion of ecphoric information into awareness making it available to the rememberer.

An automatic memory retrieval process arises when a specific, proximal cue, one that is highly correlated with the engram, interacts with the information stored in memory immediately leading to the reconstruction of parts or the whole memory. Alternatively, the recovered product of cue–engram correlation can be the result of a strategic memory search, and can act as fuel for subsequent strategic retrieval processes.

Direct Retrieval

Direct or automatic retrieval occurs when a memory cue directly interacts with a memory representation leading to the recovery or reactivation of the target memory and bringing it into awareness. These retrieval processes operate effortlessly, annulling the need for preecphoric top-down, deliberate strategic retrieval processes. Direct retrieval processes are thought to be the bases of several nondeclarative or nonconscious memory phenomena such as perceptual and conceptual implicit memory. Thus, for example, it has been argued that priming occurs through automatic retrieval and represents a

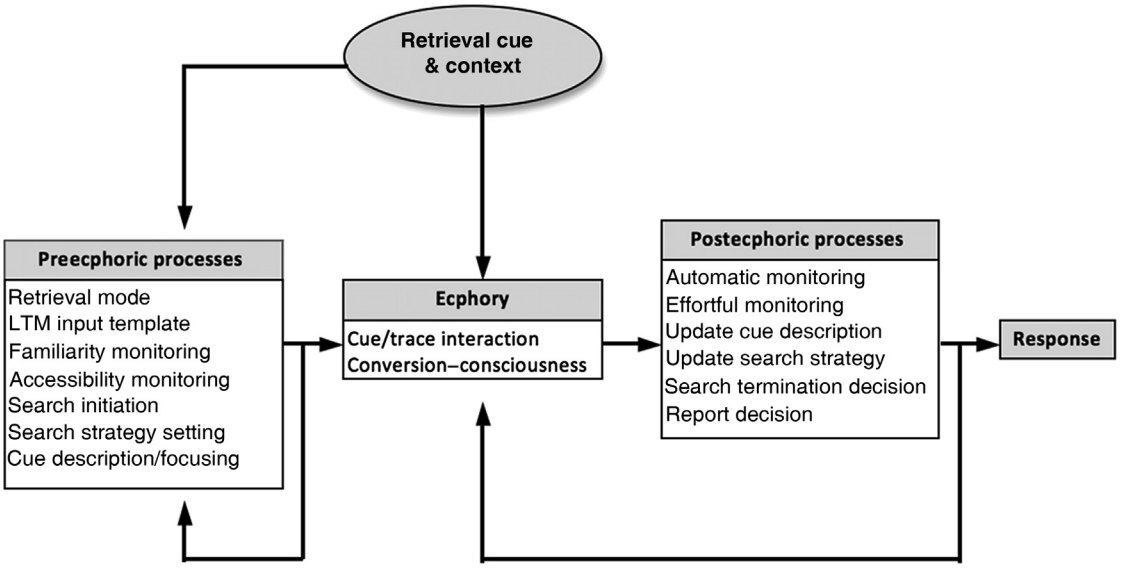

Figure 1 A simplified model of some of the processes that may be invoked during different phases of retrieval. Retrieval cues may evoke ecphory directly or initiate preecphoric processes. Preecphoric processes involve tasks such as comprehension of the retrieval task itself through activation of long-term Memory (LTM) semantic nodes, monitoring of the familiarity and/or accessibility of the memory (e.g., feeling of knowing), initiation of a search process, setting up a search strategy and description of the requirements of the memory task. This is a reiterative process until ecphory occurs or until retrieval attempt is aborted. Ecphory itself involves both a correlation between the memory cue and memory trace and a conversion process making it available to consciousness. The product of the ecphoric process is subject to several postecphoric processes including automatic 'feeling of rightness' monitoring, effortful editing, and if needed updating of the cue description and search strategy. Control processes act on the product of the monitoring processes to decide whether to report, reinitiate search, or abort retrieval attempts altogether.

relatively pure form of these processes (Graf and Mandler, 1984). By this view, all explicit, or declarative, memory retrieval is 'contaminated' by more controlled, conscious, retrieval processes. However, automatic processes may also play an important role in retrieval of certain types of conscious information. That is, some aspects of declarative memory appear to require little attentional resources and are thus considered relatively automatic or bottom-up processes and are supported by different mechanisms than controlled, strategic retrieval. For example, retrieval of associative semantic information is thought to occur through *spreading activation* of network nodes that represent different aspects of the information. This spreading activation process occurs very rapidly after the presentation of the cue and that does not appear to significantly tax attentional resources (Miyashita, 2004). In episodic memory too, some retrieval process appear to be driven in a bottom-up fashion, leading to the automatic capture of attention by retrieved memory content that mediate the conscious awareness of that information (Ciaramelli et al., 2008). Direct retrieval of information depends on two primary variables: the strength of the memory representation or memory trace and the strength or efficiency of the retrieval cue, as described below.

Early models of memory emphasized the importance of the strength of memory representation as the prime determinant of the dynamics and success of retrieval processes. By this view, competition among different memory traces as well as the decay functions of the memory traces themselves were the key factors to be considered in accounts of memory retrieval. Stronger memory traces are more likely to gain access to consciousness; memory strength is determined at encoding and is influenced by the interval elapsed between the encoding phase and time of attempted retrieval. Different such models have been proposed over the years, and most of them have in common the idea that every item (whether presented or not) possesses certain strength in memory that a memory trace's potency is determined by the efficiency of encoding and the decay over time of that strength.

Significant portions of the empirical evidence from memory research can be interpreted in light of *memory trace strength account*. For example, the manner in which an item is encoded has dramatic effects on the likelihood and manner in which that item can be retrieved which supports the premise that memory encoding determines the initial representational strength. This is most eloquently argued and convincingly demonstrated by the *levels of processing framework* (Craik and Lockhart, 1972). Incidental encoding of items that involves deep (meaningful) processing such as placing them in a sentence leads to better recall of these items compared with shallow processing such as counting the number of phonemes. Similarly, when subjects use visual imagery during presentation of names of objects, later recall for the imagined items is greatly enhanced. These examples are compatible with the idea that initial strength of the memory trace influences its later retrieval. The idea that retention intervals and memory strength decay functions determine retrieval has also been supported by research. For example, patterns of decrease in retrieval success as a function of increase in retention interval may reflect a gradual decrease in memory trace strength. Moreover, fluctuations in trace strength are thought to be responsible for invoked as an explanation why memories that were inaccessible can sometimes be retrieved at a later time.

The second factor that significantly influences direct retrieval is the relevance or *quality of the retrieval cue*. Retrieval cues that are closely related to the memory trace are more effective for successful retrieval. Such close relationships are at the heart of the influential encoding specificity principle. *The encoding specificity principle* proposes that greater similarity between the conditions present at encoding and the retrieval cues available at recovery generates the most effective retrieval cues and increases the likelihood of remembering (Tulving and Osler, 1968). The correspondence between encoding and retrieval contexts ensures that the retrieval cues are encoded with the memory trace and are therefore closely related and relevant to it at retrieval. By manipulating encoding specificity, studies have produced counterintuitive results that are best explained in terms of the correspondence between encoding and retrieval environments. For example, when participants study target words (e.g., tree) paired with loosely related cues (e.g., cloud), and are later cued either with the loosely related cue or a highly related cue (e.g., pine), the loosely related words serve as better cues for retrieval. Thus, while the trace-dependent view of retrieval focuses on the strength dimension of memory, the cue-dependent view focuses on it content and its more qualitative aspects.

Both the trace-dependent and cue-dependent views of retrieval capture important aspects of retrieval and in particular of the ecphoric process; but neither view, nor the two together, tells the complete story. The state of the memory trace, regardless of how well it is described, and the quality of the retrieval cue, irrespective of how accurately characterized, are not the only factors that influence retrieval. In what follows, some of the other critical factors will be illustrated.

Indirect Retrieval

Rememberers cannot control the ecphoric operation itself, which is considered to be automatic or ballistic. However, if external or internal cues do not elicit memory directly through ecphory, control may be exerted through strategic manipulation of the production process, biasing the processing of environmental cues or generating cues internally. These 'working with memory' procedures (Moscovitch and Winocur, 2002) involve *preecphoric* strategic processes that influence the way environmental cues are processed, those that are designed to assess the likelihood of successful retrieval and those that aim to create the optimal conditions for successful retrieval. *Postechporic* processes involve different types of memory monitoring, updating search parameters if need be as well as decisions to report or abort the memory search.

Preretrieval

A directed, strategic, process of retrieval needs to be initiated with the establishment of a *retrieval mode*. A prerequisite of successful effortful retrieval is that the cognitive system is dedicated to act in the service of retrieval. Retrieval mode is a cognitive state that ensures that external or internal events are treated as retrieval cues. Consider, for example, examining an old picture of a child. Then comes along a friend and says, "You know who that is." At that moment, the picture ceases to be merely an object of perceptual examination or aesthetic satisfaction and begins to serve as a cue for memory retrieval; the cognitive orientation toward it changes. Tulving (1983b) considered retrieval mode to constitute a tonically maintained state for the duration of a retrieval task. Once in retrieval mode, a set of metacognitive evaluative processes takes place, designed to assess the likelihood a memory would be available given the cues and context. These operate even before an actual retrieval attempt is made and certainly before retrieval occurs and give rise to a 'feeling of knowing' (Hart, 1965) or knowing that you don't know. 'Feeling of Knowing' refers to the subjective experience that a particular piece of information is available in memory despite not being currently available to recall; individuals experiencing 'feeling of knowing' would think that the information would be recognized if presented or be available to recall at a later time. Several mechanisms are hypothesized to support such preretrieval metamemory processes. For example, familiarity with the cue itself or the subject matter of retrieval can determine metamemory judgments, or the accessibility of partial cue-related information from memory can be used to determine the likelihood of the target memory itself being retrieved. Moreover, these mechanisms may interact and contribute to feeling of knowing in a cascade under different conditions and at different delays (Koriat and Levy-Sadot, 2001).

Other strategic processes are used to initiate the memory search, frame the memory query, and constrain the memory survey until local associative/cue-dependent processes can produce a possible answer. Establishing an efficient *search strategy* is often key to successful retrieval. Several retrieval strategies have been described in the literature. Prominent strategies during recall are *direct retrieval*, *plausibility judgment*, *source-constrained retrieval*, and *generate-recognize*. Direct retrieval involves direct comparison of the information provided by the retrieval cue with information in memory, in search for a close match. Plausibility judgment describes inferential decisions that are based on the plausibility and general feature overlap of the retrieval cue itself with respect to general memory representations rather than a search for specific details in memory. Source-constrained retrieval describes a mental reinstatement of the encoding strategy or conditions during retrieval so that items encoded in a particular source are preferentially recognized. Finally, generate-recognize involves an initial recall of multiple candidates and then using recognition memory to determine which candidate best answers the memory problem. The choice of optimal search strategy is sensitive to features such as the delay since encoding, demands of the retrieval task (e.g., amount of detail required), personal characteristics of the rememberer (e.g., age) and so forth.

The choice of search strategy can have dramatic effects on retrieval outcome. Consider for example the study described above using weak associates at encoding (cloud-tree) where retrieval cues are either reinstated weak cues or extralist strong ones. The encoding specificity principle would predict an advantage for the reinstated cues over the extralist ones. However, when participants are given information about the associative strength of each retrieval cue and instructed to use direct retrieval for weak associates but generate-recognize for strong associates, the advantage of reinstatement is eliminated (Higham and Tam, 2005). Thus, it can be argued that the quality of the retrieval cue, as described by the

encoding specificity framework, is not an objective measure of the goodness of fit between memory cue and memory trace, but rather is influenced by the production strategies exercised. Optimizing the search strategy to the kinds of cues available changes their efficacy as retrieval cues.

One of the primary functions of selecting an appropriate search strategy is to support the preliminary process of *description or focusing*. The purpose of focusing is to create a refined retrieval description or query of the sought-for information (Burgess and Shallice, 1996; Schacter et al., 1998). For example, in focusing the retrieval query posed by the friend with respect to the old picture, one might consider whether the child in the picture is likely to be a common childhood friend, or a childhood picture of a current common acquaintance. The search process and range of possible memory answers would dramatically change based on the outcome of this initial focusing or further description of the memory problem. Generating additional internal cues as input to production is a central mechanism in the service of cue description preecphorically. Individuals tend to produce large amounts of incidental information to the main retrieval task that serve to probe memory for additional clues and specify new search contexts. For example, in attempting to determine the identity of the child in the picture, one might try to recall details about one's primary school, extracurricular activities, or summer camps. The purpose of description processes is often to provide a correspondence or alignment between the requirements of the memory task and the content of memory and specify the type of trace that would satisfy the retrieval demands.

Postretrieval

An additional set of strategic processes operates postecphorically to assess the outcome of retrieval through various monitoring and verification processes and to determine whether to report that outcome, reinitiate a search process, or abort the search altogether. Note that while cue-dependent and trace-dependent views of memory focus on quantitative aspects of memory, here the focus is on the accuracy of the memory. Accordingly, in addition to studying hits and omissions that reflect memory quantity, accuracy-oriented approaches to memory also focus on commission errors. Postretrieval monitoring and control processes help edit out retrieved memories that are likely to be wrong and also support decisions about the amount detail and specificity required and so they influence both the accuracy and quantity of the information.

Postecphoric monitoring and verification involve both automatic and controlled processes. Thomas Reid observed that intuitive conviction is a fundamental of human memory (Reid, 1969[1710–1796]). The idea that certain memories carry with them an intrinsic sense of truthfulness has been particularly influential in theories of self-related memories. Compared with other types of information, autobiographical memories carry with them an extraordinary sense of confidence in their veracity, regardless of whether this sense is justified or not (Brewer, 1986). The mechanisms underlying this kind of 'felt rightness' *automatic monitoring* are poorly understood (Moscovitch and Winocur, 2002) but there is evidence for its rapid and automatic influence over memory (Gilboa et al., 2009). Thus, 'felt rightness' refers to the ability to rapidly or intuitively appreciate the appropriateness, suitability, and accuracy of a response with relation to the goals of the memory task and/or current personal goals. It precedes any elaborate cognitive verification of the truthfulness of the memory and the context in which it is retrieved. It has been hypothesized that intuitive 'felt rightness' is the result of an early categorical selection ('true'/'false') of memory cues based on their compatibility with general cognitive schemata that guide or serve as scaffold for memory reconstruction. The intensity of the intuition is a function of the strength of the schema and the extent of compatibility or deviation of the memory from it. Strong representations endow memory traces with more intense 'felt rightness.' The reason autobiographical memories produce the strongest sense of confidence is that the most salient, rich and robust cognitive schema is that of the self.

Early, rapid, and emotionally based decisions to endorse or reject an item may be followed by a more thorough, cognitive assessment of the memory's plausibility whether, for example, the initial response is incompatible with other knowledge or memories. Retrieved memories are constantly evaluated for their veracity by comparing them with other retrieved content, available information and with the memory task or cue descriptors by processes that are akin to problem-solving procedures (Burgess and Shallice, 1996). These have been dubbed 'editor processes' and they heavily rely on working memory processes for holding in mind multiple pieces of information, and performing logical operations for deciding whether various retrieved aspects are compatible with each other and with the memory task. They also heavily rely on processes such as conflict detection and conflict resolution.

One influential memory framework that emphasizes postecphoric processes in memory retrieval is the *source-monitoring framework* (Johnson et al., 1993). When retrieving a memory, contextual information such as when and where the item was encountered is sometimes as important as recalling the item itself. Postecphoric processes play a central role in attributing the retrieved information to a particular source, placing it in the appropriate spatiotemporal context. Retrieval of the source is often a decisive factor in making the decision of whether the information retrieved is the one being sought and even whether the information is a memory rather than a thought or fantasy.

Reconstruction and Rebinding in Retrieval

The above discussions of direct retrieval and of the component process model of strategic working with memory might leave the impression that the information 'resides' in the memory trace waiting to be 'activated' by the correct key introduced by internal or external memory cues. As mentioned in the introduction, retrieval should in fact be thought of as a reconstructive process where the fidelity between the original events and the reconstructed experience is imperfect at best (Bartlett, 1920; Schacter et al., 1998). The issue of reconstructive memory is detailed elsewhere in this volume. However, in the present context it is important to mention it because many so called memory errors that reflect the reconstructive

nature of memory arise at retrieval. For example, the content of retrieval cues can bias or distort the content of the memories that are retrieved in many different ways. Some of these include the influence of expectations during retrieval on the content of memory (suggestibility), source misattributions, gap filling, and the misinformation effect (Loftus, 2005). While the reconstructive nature of retrieval has primarily been investigated by paradigms amplifying the fallibility of memory and its proneness to error, it is important to note that this characteristic of memory is crucial for some of its primary functions. Specifically, it is the reconstructive nature of memory that allows it to be used flexibly in multiple contexts and for multiple purposes including planning for possible future scenarios and imagining possible futures.

The neurobiological instantiation of memory reconstruction during retrieval is captured by ideas that describe memory in terms of binding. Binding involves linking together representations of different kinds of information that are processed by different neural substrates. These mechanisms are important over the short term as in during perception or working memory, but are also critical when fragments of information from long-term storage need to be linked together. During encoding binding mechanisms are important for understanding how different aspects of the memory trace are connected and how memory cues might be bound within the memory trace. At retrieval, they are critical for the reinstatement of patterns of activity through processes such as pattern completion as part of a cue–trace interaction. In that sense, memories are not located anywhere in particular within the brain. Instead, they are distributed across large expansions of neural substrates that represent different features of a prior experience. They involve almost all parts of the association cortex and limbic system. Critical elements of the component process model of retrieval described above may be broadly described in terms of the neural substrates that support them. The memory engram and memory cue are properties of items and contexts that are represented in a distributed fashion across relevant unimodal, transmodal, and heteromodal association cortical areas. Limbic areas, and in particular the hippocampus, are critical for binding together these disparate representations into coherent ensembles during encoding and to pattern-complete these ensembles during retrieval. Finally, structures within the prefrontal cortex and possibly posterior heteromodal cortical regions interact with these limbic-neocortical ensembles to guide, organize, and regulate the rebinding process.

Conclusion

The notion that retrieval is the result of a simple ballistic process of an interaction between a memory cue and a memory trace only describes a very limited aspect of a much more elaborate and complex process. Accordingly, describing retrieval only in terms of cue and trace characteristics is an insufficient account of it. Instead, retrieval is a complex reconstructive, reiterative process involving many neurocognitive systems. While schematic models such as the one presented here are extremely useful in formulating ideas about how such complex systems might work and interact, they are also necessarily a gross simplification and are not to be taken at face value. The implementation of each of the components described by the model is considerably more complex both at the cognitive and the neural level. Some of these components, such as ecphory and retrieval mode, are better understood than others, such as automatic monitoring or cue description. Future research in cognition and cognitive neuroscience will undoubtedly elucidate these processes or perhaps prove some of them wrong or redundant.

See also: Autobiographical Memory; Confabulation; Context Dependent Memory; Declarative Memory, Neural Basis of; Episodic Memory; Eyewitness Testimony; False Memories; Flashbulb Memories; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Memory Binding; Proactive Interference; Recognition; Reconstructive Memory, Psychology of; Retroactive Interference; Semantic Memory; Short-Term Memory: Psychological and Neural Aspects; Source Monitoring; Working Memory, Psychology of.

Bibliography

Bartlett, F.C., 1920. Some experiments on the reproduction of folk stories. Folklore 31 (1), 30–47.
Brewer, W.F., 1986. What is autobiographical memory? In: Rubin, D.C. (Ed.), Autobiographical Memory. Cambridge University Press, New York.
Burgess, P.W., Shallice, T., 1996. Confabulation and the control of recollection. Memory 4 (4), 359–411.
Ciaramelli, E., Grady, C.L., Moscovitch, M., 2008. Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. Neuropsychologia 46 (7), 1828–1851.
Craik, F.I.M., Lockhart, R.S., 1972. Levels of processing: a framework for memory research. Journal of Verbal Learning and Verbal Behavior 11, 671–684.
Gilboa, A., Alain, C., He, Y., Stuss, D.T., Moscovitch, M., 2009. Ventromedial prefrontal cortex lesions produce early functional alterations during remote memory retrieval. Journal of Neuroscience 29 (15), 4871–4881.
Graf, P., Mandler, G., 1984. Activation makes words more accessible, but not necessarily more retrievable. Journal of Verbal Learning and Verbal Behavior 23, 553–568.
Hart, J.T., 1965. Memory and the feeling-of-knowing experience. Journal of Educational Psychology 56 (4), 208–216.
Higham, P.A., Tam, H., 2005. Generation failure: estimating metacognition in cued recall. Journal of Memory and Language 52, 595–617.
Johnson, M.K., Hashtroudi, S., Lindsay, D.S., 1993. Source monitoring. Psychological Bulletin 114 (1), 3–28.
Koriat, A., Levy-Sadot, R., 2001. The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing [Research Support, Non-U.S. Gov't]. Journal of Experimental Psychology: Learning, Memory, and Cognition 27 (1), 34–53.
Loftus, E.F., 2005. Planting misinformation in the human mind: a 30-year investigation of the malleability of memory [review]. Learning & Memory 12 (4), 361–366. http://dx.doi.org/10.1101/lm.94705.
Miyashita, Y., 2004. Cognitive memory: cellular and network machineries and their top-down control. Science 306 (5695), 435–440.
Moscovitch, M., Winocur, G., 2002. The frontal cortex and working with memory. In: Stuss, D.T., Knight, R. (Eds.), Principles of Frontal Lobe Function. Oxford University Press, New York, pp. 188–209.
Reid, T., 1969[1710–1796]. Essays on the Intellectual Powers of Man. III: Concerning Memory. MIT Press, Cambridge, MA.
Schacter, D.L., Norman, K.A., Koutstaal, W., 1998. The cognitive neuroscience of constructive memory. Annual Review of Psychology 49, 289–318.
Semon, R., 1921[1904]. The Nmeme. George Allen & Unwin, London.
Tulving, E., 1983a. Ecphoric processes in episodic memory. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 302 (110), 361–370.
Tulving, E., 1983b. Elements of Episodic Memory. Clarendon Press, Oxford.
Tulving, E., Osler, S., 1968. Effectiveness of retrieval cues in memory for words. Journal of Experimental Psychology: Applied 77 (4), 593–601.

Retroactive Interference

Michael Craig, Human Cognitive Neuroscience, Psychology, University of Edinburgh, UK
Michaela Dewar, Psychology, School of Life Sciences, Heriot-Watt University, Edinburgh, UK; and Human Cognitive Neuroscience, Psychology, University of Edinburgh, UK
Sergio Della Sala, Human Cognitive Neuroscience, Psychology and Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Retroactive interference (RI) refers to when newly acquired information impairs the retention of previously acquired information. RI is observed in both the declarative and nondeclarative divisions of the long-term memory system. RI is observed across the life span. However, susceptibility to RI is heightened in the elderly and patients with memory impairments. RI is believed to occur via two mechanisms: (1) retrieval interference, where two or more memories sharing the same/highly similar retrieval cue compete for retrieval, and (2) consolidation interference, where new incoming information disrupts the memory strengthening of recently acquired memory traces. Consolidation RI appears to be the main contributor of RI-induced forgetting, the effects of which can be reduced if consolidation RI is delayed and/or minimized. Here we discuss representative behavioral and neuroscience research demonstrating the effects of RI.

Background

Imagine you are sat comfortably on your sofa, chatting to an old friend on the telephone when there is a knock at the door. You answer the door finding a salesman attempting to convince you to buy his products. Five minutes later, when you finally get a word in, you politely decline, returning indoors to your phone conversation … but you can no longer recall the details of your recent phone conversation, "What were you saying about your father's birthday party?" This is a fictional example of retroactive interference (RI), whereby listening to the salesman interfered with your retention of your friend's birthday story.

RI is a phenomenon that occurs when new information impairs the ability to recall previously acquired material. This is the opposite of proactive interference (PI) where previously acquired material interferes with the ability to recall new information.

Historical Overview

RI has been documented for over a century by multiple memory researchers (e.g., Bigham, 1984; Müller and Pilzecker, 1900; Skaggs, 1925; McGeoch and McDonald, 1931). However, it is Georg Elias Müller and his student Alfons Pilzecker who are considered to be the pioneers of RI research with their seminal work reported in *Experimentelle Beiträge zur Lehre vom Gedächtniss* (*Experimental Contribution to the Science of Memory, 1900*).

Müller and Pilzecker (1900) conducted a series of experiments investigating memory for learned paired associations when they incidentally discovered RI. This led Müller and Pilzecker (1900) to investigate further this phenomenon, ultimately providing quantitative evidence for RI for the first time. Müller and Pilzecker (1900) also provided the first suggestion of memory consolidation, i.e., the strengthening of new memories over time; a theory that is gaining increasing acknowledgment in the fields of neuropsychology and neuroscience. For a review of Müller and Pilzecker's (1900) research, see Dewar et al. (2007).

In the wake of Müller and Pilzecker's (1900) findings, further researchers began to acknowledge that forgetting could be induced by subsequent interference (e.g., Jenkins and Dallenbach, 1924; Skaggs, 1925). In consideration of such research, McGeoch and McDonald (1931) proposed that the then widely accepted Decay Theory of forgetting should be replaced by an Interference Theory of forgetting. In contrast to the Decay Theory, which proposes that memories are lost due to the mere passage of time, the Interference Theory proposes that forgetting occurs when memories interfere with, and disrupt, one another (Baddeley, 1999). As a result of RI research in the fields of neuropsychology and neuroscience (e.g., Briggs, 1954; Underwood and Postman, 1960; Cowan et al., 2004; Wixted, 2004, 2005, 2011; Dewar et al., 2007), the Interference Theory is now considered to be the most prominent theory of forgetting.

RI in Different Memory Systems

The currently dominant theoretical account of human memory divides memory into different systems (Tulving, 1985), one of which is long-term memory (LTM). LTM can be further subdivided into declarative (facts and knowledge) and nondeclarative (implicit and procedural) memory (Tulving, 1972); RI is observable in both subdivisions of LTM. Here, we discuss representative research that demonstrates the effects of RI within the declarative and nondeclarative human memory systems.

RI in Declarative Memory

In Tulving's (1972) taxonomy, declarative memory is subdivided into episodic and semantic memory. Episodic memory provides the ability to recall detailed information and mentally 'relive' a past episode or event from our lives. For example, if you attempt to recall a special day from your life, whether it be a wedding or birthday, it is likely that you can recall a vast quantity of detailed information from this particular episode in the form of *what* occurred during the

episode, *where* the episode took place, and *when* the episode occurred (Tulving, 1983). In contrast, semantic memory is typically responsible for the recall of factual information, deprived of contextual details. For example, you are likely to know that London is the capital of the United Kingdom but it is unlikely that you can specifically recall the contextual details of *where* and *when* you learned this fact.

In experimental neuropsychology paradigms, RI in episodic memory is typically observed when new learning is followed by further incoming information, thus generating interference and impairing the ability to recall the initially learned material (e.g., Müller and Pilzecker, 1900; Skaggs, 1925; Dewar et al., 2007). In Müller and Pilzecker's (1900) seminal research this was demonstrated by probing retention of newly learned material following one of two critical delay periods: a delay filled with further stimuli or a delay devoid of further stimuli. Retention was markedly lower following the filled delay than the unfilled delay, thus revealing a clear effect of RI.

In Müller and Pilzecker's (1900) early experiments, the same type of material (paired syllable lists) was used as the 'to-be-retained' and interpolated 'interference' material. Thus, the question was proposed that RI may be due to a material similarity effect. However, Müller and Pilzecker (1900) demonstrated this to not to be the case; RI remained even when the filled delay comprised of material (to-be-described landscape pictures) that was highly dissimilar to the initially learned material (paired syllable lists). Müller and Pilzecker (1900) concluded that RI was due to general cognitive engagement during the postlearning delay, and not due to material similarity. Using a modern adaption of Müller and Pilzecker's (1900) paradigm, Dewar et al. (2007) investigated the possible differences in the level of RI generated by different types of postlearning activities. The authors compared the effects of five distinctively different 8-min filled delay activities and an 8-min unfilled delay on the retention of wordlist material. The five filled delay activities applied were as follows: solving short mathematical problems, playing a spot the difference game, attending to an audio recording, watching 'real-world' video clips, and detecting tones. It was found that each of the different filled delays was capable of generating significant levels of RI; retention of wordlist material was lower following each of the five *filled* delay activities than following the unfilled delay. It is hypothesized that RI was observed across all filled delay activities due to additional, new incoming information during activity performance, and that when minimizing this additional, new incoming information, as in the unfilled delay, retention can be improved. In addition to new incoming information from the external environment causing RI, Craig et al. (2014) demonstrated that internal autobiographical memory activities also cause RI. Further to this, Dewar et al. (2012) demonstrated that the RI effect persists over the long term: stories learned prior to a 10-min spot the difference game were retained less well over 7 days than were stories learned prior to a 10-min unfilled delay.

In addition to material similarity not being necessary for RI to occur, the acoustic similarity of initially learned and interpolated material also appears to have little influence on RI generation (Baddeley and Dale, 1966). In contrast, the semantic similarity of initially learned and interpolated material does appear to be an important variable in the level of RI observed (Skaggs, 1925; Barnes and Underwood, 1959; Baddeley and Dale, 1966). Using an AB–AC paradigm, Barnes and Underwood (1959) demonstrated that when learned wordlists shared a high level of semantic similarity, a higher level of RI was observed than when learned AB–AC wordlists shared a low level of semantic similarity.

RI in Nondeclarative Memory

Nondeclarative memory is comprised of implicit and procedural memory. Implicit memory is memory acquired through the senses and serves to identify objects and the structure of language, often in the recognition of themes or patterns. In contrast, procedural memory allows for the learning and long-term retention of specific motor actions and responses, e.g., riding a bicycle or tying a shoelace. RI is observed in both implicit and procedural memory.

RI in Implicit Memory

The susceptibility of implicit memory to RI has been an unresolved debate for some time, with many researchers favoring the interpretation that implicit memory is immune to RI (see reviews by Anderson and Neely, 1996; Schacter and Cooper, 1993). This conclusion was based on the findings of several influential studies (e.g., Graf and Schacter, 1987; Sloman et al., 1988) that did not find evidence of RI in implicit memory. However, the results of several other investigations have demonstrated the presence of RI in implicit memory (Lustig and Hasher, 2001; Nelson et al., 1989; Winocur et al., 1996). As a result of this research, implicit memory is not considered to be immune to RI.

The effects of RI can be observed within different modalities of implicit memory. For example, RI is observable in tone pitch memory. Using a delayed-comparison task of recognition memory for pitch, Massaro (1970) investigated whether a 'blank' period (akin to the previous discussed 'unfilled delay') after the presentation of nonverbal stimuli would function to improve retention for the sensory traces in memory. Results demonstrated that the 'blank' period after tone presentation was sufficient to improve retention when the tone's duration was brief (~2 s). Thus, Massaro (1970) demonstrated that the avoidance of postlearning RI was capable of producing a benefit in the long-term retention of audible sensory information.

In addition, memory for a novel scent is also susceptible to RI (Walk and Johns, 1984; Köster et al., 2002; Olsson et al., 2009; Morrin et al., 2011), despite having previously been thought to be immune (Engen and Ross, 1973; Lawless and Engen, 1977). Walk and Johns (1984) investigated RI in scent memory by exposing participants to two food scents, followed by (or not – control condition) further potentially interfering 'unhelpful' stimuli (a third scent; the name of a third scent), or 'helpful' stimuli (the name of one of the original two scents). Participants were then presented with four scents of which one was the original scent that they were asked to try to correctly recognize. Results demonstrated a clear effect of RI: identification of the original scent was significantly higher when the original scent was followed by the control condition, than when it was followed by the further 'unhelpful' stimuli.

RI in Procedural Memory

The effects of RI in procedural memory can be observed in the learning of a novel motor skill. Acquisition of a new motor skill

can be disrupted if one attempts to acquire a further, different motor skill shortly after learning the initial skill (Robertson et al., 2004). Using a force-field task requiring dynamic transformations, i.e., the learning of new muscle forces, Brashers-Krug et al. (1996) found that when learning of one pattern of forces (force A) was followed by training on a second opposing pattern of forces (force B), recall of the initially encoded pattern (force A) was impaired after a 24-h delay. In addition, Shadmehr and Brashers-Krug (1997) demonstrated that two motor maps could be learned and retained (for at least 5 months), only if training sessions were separated by a minimum of 5 h. If this was not so, significant RI was observed in the retention of force A. Interestingly, Kremer et al. (2011) found that a similar level of RI was generated when acquisition of a new motor skill was followed by both highly similar and highly dissimilar internal practicing of a new, different motor skill, i.e., no physical movement was involved, only internal imagery. The level of RI generated by internal practicing was also highly similar to an external unrelated reading task.

RI in Different Populations

Developmental Effects

Although the majority of research concerning RI has focused on adults, evidence suggests that RI can be observed in early human development. For example, infants demonstrate RI in visual recognition tasks (Turati, 2008) and mobile reinforcement paradigms (Rossi-George and Rovee-Collier, 1999). However, the trajectory of RI susceptibility during development is less clear with research suggesting that susceptibility does not alter during development from childhood (~4 years of age) to adulthood (~18 years of age) (Howe, 1995; Lee and Bussey, 2001; Darby and Sloutsky, 2013).

Healthy Aging

From adolescence onward, susceptibility to RI increases with age (Hartman and Hasher, 1991; Shimamura and Jurica, 1994; Winocur et al., 1996; Ebert and Anderson, 2009; Solesio et al., 2009; Solesio-Jofre et al., 2011). Using an AB–AC paradigm, Ebert and Anderson (2009) compared susceptibility to RI in younger and older adults. Participants first learned a list (A–B) of 12 semantically related word pairs (e.g., knee–bone). Twenty minutes later, a second list (A–C) was learned in which the same stimulus words were paired with new response words (e.g., knee–bend). Both lists were learned to criterion. Postlearning, participants were asked to recall the first (A–B) list. Results demonstrated an increased susceptibility to RI *and* PI (interference by A–B learning during later A–C learning) in older adults than younger adults.

Patient Populations

Some patient populations are profoundly vulnerable to RI. This is particularly evident in those with memory impairment, e.g., dense focal amnesia, amnestic mild cognitive impairment (aMCI), and Alzheimer's disease (AD). A hallmark of these conditions is an acute forgetfulness for newly acquired information (anterograde amnesia) and loss of previously learned information (retrograde amnesia).

Cowan et al. (2004) investigated the effects of RI in patients with amnesia as a result of head injury. In two experiments, patients were presented with lists of words (Experiment 1) and stories (Experiment 2). These stimuli were to be recalled immediately and/or after an extended test delay (10 min in Experiment 1; 1 h in Experiment 2). Results demonstrated an especially high level of forgetting when presentation was followed by performing a series of cognitive tasks than by an unfilled delay. Interestingly, in Experiment 2, when the unfilled delay was increased to 1 h, a substantial memory improvement was observed. Patients who had previously retained an average of 7% of material after the filled delay were now able to successfully retain an average of 79%. In fact, the authors found that three of the four patients who were unable to retain any material (i.e., 0%) when learning was followed by a filled delay were now able to retain 78, 85, and 90% of learned material.

Della Sala et al. (2005) expanded on the findings of Cowan et al. (2004) by assessing RI in patients with aMCI (a prodrome of AD). In their study, Della Sala et al. (2005) presented patients with prose passages that were followed by a 1 h filled delay (comprising of various psychometric tasks) or an unfilled delay of equivalent duration. Results demonstrated a striking memory improvement when learning was followed by an unfilled delay (55%) compared to a filled delay (20% retention). Dewar et al. (2009, 2010a,b, 2012) provided further evidence for the benefit of minimizing RI in amnesic patients, including patients with mild AD. Thus it is suggested that an increased susceptibility to RI is a key cognitive mechanism underlying memory impairment in many patients with amnesia (Cowan et al., 2004; Della Sala et al., 2005; Dewar et al., 2009, 2010a,b, 2012). Further more, these findings provided an unexpected insight into the spared capacity to retain new information in amnesic patients when RI is minimized, an attribute that was previously thought not to be possible.

Mechanisms Underlying RI

Two key mechanisms have been proposed to account for RI: retrieval interference, where two or more memories that share the same/highly similar cues compete during retrieval, and consolidation interference, where the consolidation of memory traces is disrupted by new incoming information.

Retrieval Interference

Tulving and Thomson (1973) suggested that when memories are retrieved in the same/highly similar contextual settings that they were encoded in (e.g., location, auditory, and visual context), recall performance will be superior than retrieval that occurs within a highly dissimilar contextual setting (Mensink and Raajmakers, 1988; Anderson et al., 1994). This was most famously demonstrated by Godden and Baddeley (1975) who found that when deep-sea divers learned a list of words on land or under water, recall performance was superior when retrieval was performed within the same contextual setting as during learning. Contextual components of an encoding environment can act as cues, aiding the retrieval of memory

traces that were encoded in such a context. However, when different memory traces share the same or, highly similar cue, competition for retrieval can occur, potentially inhibiting the retrieval of the desired memory, i.e., retrieval interference.

Retrieval interference is typically observed in the lab using the AB–AC paradigm (Skaggs, 1925; Postman and Alper, 1946; Briggs, 1954; Wixted, 2004; Dewar et al., 2007). In the AB–AC paradigm, participants are presented a list of word pairs – e.g., cat–glass, tree–car, sky–train – i.e., the 'A–B' list. During the recall test, the first word of each pair ('A' words) is presented to cue retrieval of the paired second word ('B' words). Following a delay, a further list of word pairs is presented in which the first words remain the same as in list A–B, but the second words are replaced with new words – e.g., cat–sun, tree–shoe, sky–book – i.e., the 'A–C' list. During a subsequent A–B recall test, the presentation of the cue word, e.g., 'cat' ('A' word), can result in interference at retrieval of the associated 'B' word (glass) due to competing retrieval of the associated 'C' word (sun).

Incisa della Rocchetta and Milner (1993) found that patients with left frontal lobe removal showed increased RI during a cued recall task. The authors suggested that the integrity of the left frontal lobe is vital in strategic retrieval and for the suppression of potentially interfering items in verbal memory.

While retrieval interference can be observed reliably in the laboratory, it is still unknown whether memory traces for interpolated material are simply stronger than older traces, or whether their learning/retrieval may actually weaken old memory traces (Dewar and Della Sala, 2013).

Critically, retrieval interference cannot account for all aspects of interference-induced forgetting. For example, even when interpolated task material is highly dissimilar to the previously learned material and unlikely to have similar retrieval cues, interference can still be observed. In such cases, it is likely that an alternative mechanism of RI is responsible – consolidation interference.

Consolidation Interference

On the acquisition of new information, it is believed that new weak memory traces are formed. Over time these traces are strengthened by consolidation processes. Consolidation takes place in the background, without intentional memory rehearsal, and ultimately results in the ability to recall recently acquired information at a later point in time. Two stages of consolidation have been postulated: 'synaptic' and 'systems' consolidation (Dudai, 2004). Synaptic consolidation is believed to occur in the minutes/hours that follow new learning, boosting memory strength at the cellular level. In contrast, systems consolidation is believed to occur in the hours/years that follow new learning, involving the reorganization and storage of the new memory in the neocortex.

Despite early recognition of the notion of memory consolidation (Ribot, 1881, 1882; Burnham, 1903), or memory 'preservation' as termed by Müller and Pilzecker (1900), it has been mostly ignored within modern neuropsychology other than a few notable exceptions (e.g., Gaskell and Dumay, 2003; Bosshardt et al., 2005a,b; Wixted, 2004; Dewar et al., 2007, 2012). However, an increasing volume of behavioral research now provides considerable evidence for the existence of memory consolidation. With this comes a growing acknowledgment of 'consolidation interference.'

Wixted (2004) suggests that consolidation interference occurs due to consolidation 'resources' not being infinite. Wixted (2004) suggests that when new learning is directly followed by further information, consolidation resources are divided between the two sets of material, leading to a lower level of consolidation of the initially learned material, and thus resulting in a lower level of retention. Carefully designed experiments provide sound evidence that it is indeed consolidation that is hampered by new incoming information: since consolidation strengthens newly formed memory traces over time, a core prediction is that consolidation interference should be especially detrimental if it occurs immediately after new learning, i.e., when the newly formed memory trace has had little chance to strengthen. If, on the other hand, consolidation interference occurs following a delay, the new memory trace should have had sufficient time to be strengthened, thus being less vulnerable to subsequent interference, resulting in increased retention. Such a 'temporal gradient' is indeed observed, as demonstrated initially by Müller and Pilzecker (1900). This effect has been widely observed, spanning declarative (Müller and Pilzecker, 1900; Skaggs, 1925; Dewar et al., 2009) and nondeclarative (Brashers-Krug et al., 1996; Shadmehr and Brashers-Krug, 1997; Valeria Della-Maggiore, 2005) memory, as well as animal neuroscience (see later).

In relation to the discussed patient populations and their increased susceptibility to RI, one possible explanation for this increased susceptibility is an expansion on Wixted's (2004) suggestion that consolidation 'resources' are not infinite. Indeed, as found by Cowan et al. (2004), Della Sala et al. (2005), and Dewar et al. (2007, 2009, 2010a,b) when postlearning RI is minimized, a higher level of material can be retained in amnesic, aMCI, and early AD patients. Moreover, Dewar et al. (2009) demonstrated a temporal gradient of RI, i.e., the longer the delay between new learning and RI onset, the greater the level of newly learned material retained. This suggests that consolidation resources are indeed still available in patients (if not in the same quantity as in healthy matched controls) and some consolidation can occur when provided with optimal conditions, i.e., minimizing RI (Dewar et al., 2009, 2010a,b, 2012). It should be noted that not all patients show a benefit of minimizing RI (Cowan et al., 2004; Della Sala et al., 2005) and that some patients, e.g., those in the moderate stages of AD, show a reduced benefit (Dewar et al., 2012). In this case, it could be that consolidation resources are depleted to the point that new memories are not able to be consolidated, potentially due to marked damage to the hippocampus (Cowan et al., 2004; Della Sala et al., 2005; Dewar et al., 2012).

Sleep memory research also provides evidence for consolidation interference as a cause of RI in declarative (Buzsáki, 1998; Gottselig et al., 2004; Ellenbogen et al., 2006a,b) and nondeclarative (Diekelmann et al., 2009; Robertson et al., 2005) memory. Given that new sensory information is minimal during sleep, sleep should also provide conditions conducive to consolidation (i.e., minimal RI), thus enabling the processes to occur. Ellenbogen et al. (2006b) used the AB–AC paradigm as a means for examining sleep-related consolidation of newly acquired declarative memories. In this research, participants learned a first list of paired material (A–B) prior to a 12-h

period of sleep or wakefulness. Following the 12-h delay, participants were presented with a second list of paired material (A–C), which shared the same cue items, i.e., 'A.' The level of RI generated from A–C material during subsequent retrieval of A–B material was then used as a measure of the strength of A–B memory traces, i.e., the greater the level of consolidation, the stronger the A–B trace, and the less susceptible to A–C interference the trace should be. Ellenbogen et al. (2006b) demonstrated a significant benefit of sleep in A–B list retention. The authors suggested that consolidation processes during sleep were able to strengthen the new memory traces, increasing their immunity to subsequent RI from similar traces.

In addition, a beneficial effect of postlearning sleep can also be observed in nondeclarative memory whereby newly learned motor skills are more stable and resistant to subsequent interference if training is followed by a period of sleep (Walker et al., 2003). Further to this, postlearning sleep leads to enhanced speed and accuracy of skill performance at a delayed retesting if postlearning sleep occurs in the absence of further training during the retention period. Walker et al. (2003) suggested that the long-term retention of newly learned motor skills is highly dependent on processes that occur during sleep, ensuring memory traces are made resistant to subsequent RI.

The association between sleep and consolidation has been demonstrated to hold a great depth with both slow wave sleep and rapid eye movement sleep associated with the consolidation of declarative (Empson and Clarke, 1970; Tilley and Empson, 1978; Rasch et al., 2007; Lahl et al., 2008; Diekelmann et al., 2009) and nondeclarative memories (Diekelmann et al., 2009; Walker et al., 2003; Maquet et al., 2003), respectively. Indeed, it is still being debated whether sleep-related improvements are related to minimal interference during sleep and/or due to more active neural processes present during sleep (Ellenbogen et al., 2006a).

As a result of the discussed research, Dewar et al. (2007) proposed a novel model of forgetting in which consolidation interference is a key mediator in the generation of RI, and thus the forgetting of recently learned memories. Moreover, the authors suggested that when the immediate postlearning delay is devoid of further information, a higher level of consolidation can occur, and thus a greater level of learned material can be retained. For details see Dewar et al. (2007).

RI and Neuroscience

As in neuropsychological research, RI is readily observable in animal neuroscience. In such investigations, RI is typically induced by the administration of protein synthesis inhibitors (usually toxins or antibiotics) shortly after new learning has taken place (Davis and Squire, 1984; Dudai, 2004; Mednick et al., 2011; Wixted, 2004). These drugs are believed to disrupt the consolidation of recently acquired memory traces (Dudai, 2004). When consolidation is disrupted shortly after learning, animals retain very little of what they have just learned. However, when consolidation is disrupted after a delay, animals are able to retain substantially more of what they have just learned (Dudai, 2004). This temporal gradient of pharmacological RI is reminiscent of that seen in neuropsychological research (see earlier) and shows that consolidation interference is particularly detrimental to new, fragile memories which have not yet benefitted from initial strengthening (Izquierdo et al., 1999; Mednick et al., 2011).

In neuroscience, the key mediator of RI is believed to be the inhibition of synaptic consolidation in the highly plastic hippocampal circuit (Carr et al., 2011), which occurs prior to the much longer process of systems consolidation where memories may become hippocampus-independent (Dudai, 2004). Specifically, it appears that RI occurs during synaptic consolidation as a result of disruptions of hippocampal long-term potentiation (LTP). LTP is a long-lasting enhancement of synaptic transmission that occurs in response to brief, high-frequency stimulation of postsynaptic neurons (Wixted, 2004). It is induced during encoding of a new memory and maintained during subsequent consolidation. New LTP induction and memory encoding both interfere with the maintenance of LTP, and thus the consolidation of recent memories, if they occur shortly after new learning (Mednick et al., 2011). According to Mednick et al. (2011) new encoding interferes with the consolidation of recent memories since the induction of new LTP (associated with new encoding) inhibits the maintenance of recent LTP (associated with consolidation). This suggestion mirrors findings and hypotheses in neuropsychological research (e.g., Cowan et al., 2004; Della Sala et al., 2005; Dewar et al., 2009, 2010a,b, 2012).

In addition to LTP inhibition, the disruption of 'offline replay' is also believed to play an important role in RI. Offline replay is the sequential reactivation of hippocampal representations. For example, in rodents, hippocampal activity associated with recent activity (e.g., walking along a track) is actually replayed during subsequent periods of wakefulness. This process is highly suited to promote memory consolidation (Carr et al., 2011).

Animal model lesion studies strongly associate the hippocampus with consolidation, whereby an increased susceptibility to RI, and impaired retention, is observed when a lesion in the hippocampus is made (Jarrard, 1975; Bartko et al., 2010). It is also of note that patient populations who present with hippocampal degradation (e.g., aMCI and AD) demonstrate an increased susceptibility to RI. Thus, while RI mechanisms are not fully understood at present, it is likely that by interfering with hippocampal LTP and offline replay, synaptic consolidation interference occurs, resulting in the impaired retention of new memories.

Neuroscience research also provides evidence for the biological basis of retrieval interference. A large body of research (see, e.g., Miller and Cohen, 2001) suggests that the prefrontal cortex is involved in guiding the online dynamics of competition, thus ensuring that the correct response 'wins' and is successfully retrieved. Shimamura et al. (1995) demonstrated that patients with frontal lobe lesions exhibit increased retrieval interference susceptibility. Moreover, Johansson et al. (2007) demonstrated that sustained prefrontal event-related potentials (ERPs) were related to the level of observed retrieval interference. In contrast, Tendolkar et al. (1997) found no evidence of prefrontal activity in an ERP study investigating retrieval interference. More recently, in an animal model, Bartko et al. (2010) demonstrated an increased susceptibility to retrieval RI in rats with perirhinal and postrhinal cortical lesions.

Thus, while a strong relationship between the hippocampus and consolidation RI is observed, the biological basis of retrieval interference requires further investigation.

Summary

In summary, the neuropsychological and neuroscience research discussed in this article demonstrates the detrimental effects of RI on LTM retention in both declarative and nondeclarative memory. The detrimental effects of RI are visible across the life span, from infants to the elderly. While susceptibility to RI is stable during development, from adolescence, susceptibility increases throughout healthy aging. Further to this, many patients with memory impairment are particularly vulnerable to RI, and this vulnerability is hypothesized to underpin the memory problems observed in such patients. RI is believed to occur via two mechanisms: retrieval interference and consolidation interference. Consolidation interference appears to be the main cause of RI-induced forgetting, the effects of which can be reduced if consolidation interference is delayed and/or minimized. This benefit is believed to be due to enhanced LTP and offline replay which occur shortly after learning.

Acknowledgments

Thanks go to the Alzheimer's Society which currently sponsors Michael Craig as part of a PhD studentship awarded to Michaela Dewar and Sergio Della Sala, and to the Royal Society of Edinburgh/Lloyds TSB Foundation for Scotland which has funded Michaela Dewar's Personal Research Fellowship.

See also: Amnesia: General; Episodic Memory; Hippocampal Amnesia; Non Declarative (Procedural) Memory; Proactive Interference; Semantic Memory; Synaptic Mechanisms and Protein Synthesis in Memory.

Bibliography

Anderson, M.C., Neely, J.H., 1996. Interference and inhibition in memory retrieval. In: Bjork, E.L., Bjork, R.A. (Eds.), Memory. Handbook of Perception and Cognition, second ed. Academic Press, San Diego, CA, pp. 237–313.

Anderson, M.C., Bjork, R.A., Bjork, E.L., 1994. Remembering can cause forgetting: retrieval dynamics in long-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 1063–1087.

Atkinson, R.C., Shiffrin, R.M., 1968. Chapter: human memory: a proposed system and its control processes. In: Spence, K.W., Spence, J.T. (Eds.), The psychology of learning and motivation, vol. 2. Academic Press, New York, pp. 89–195.

Baddeley, A.D., 1999. Essentials of Human Memory. Psychology Press, Hove.

Baddeley, A.D., Dale, H.C.A., 1966. The effect of semantic similarity on retroactive interference in long- and short-term memory. Journal of Verbal Learning and Verbal Behaviour 5, 417–420.

Barnes, J.M., Underwood, B.J., 1959. "Fate" of first-list associations in transfer theory. Journal of Experimental Psychology 58 (2), 97–105.

Bartko, S.J., Cowell, R.A., Winters, B.D., Bussey, T.J., Saksida, L.M., 2010. Heightened susceptibility to interference in an animal model of amnesia: impairment in encoding, storage, retrieval–or all three? Neuropsychologia 48, 2987–2997.

Berman, M.G., Jonides, J., Lewis, R.L., 2009. In search of decay in verbal short-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 317–333.

Bigham, J., 1984. Memory. The Psychological Review 1, 453–461.

Bliss, T.V., Lomo, T., 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. Journal of Physiology 232 (2), 331–356.

Bosshardt, S., Degona, N., Schmidt, C.F., Boesiger, P., Nitsch, R.M., Hock, C., Henke, K., 2005a. One month of human memory consolidation enhances retrieval-related hippocampal activity. Hippocampus 15, 1026–1040.

Bosshardt, S., Schmidt, C.F., Jaermann, T., Degonda, N., Boesiger, P., Nitsch, R.N., Hock, C., Henke, K., 2005b. Effects of memory consolidation on human hippocampal activity during retrieval. Cortex 41, 486–498.

Brashers-Krug, T., Shadmehr, R., Bizzi, E., 1996. Consolidation in human motor memory. Nature 382, 252–255.

Briggs, S.E., 1954. Acquisition, extinction, and recovery functions in retroactive inhibition. Journal of Experimental Psychology 47 (5), 285–293.

Burnham, W.H., 1903. Retroactive amnesia: illustrative cases and tentative explanation. American Journal of Psychology 14, 118–132.

Buzsáki, G., 1998. Memory consolidation during sleep: a neurophysiological perspective. Journal of Sleep and Rest 7, 17–23.

Carr, M.F., Jadhav, S.P., Frank, L.M., 2011. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nature: Neuroscience 14, 147–153.

Conrad, R., 1964. Acoustic confusions in immediate memory. British Journal of Psychology 55, 75–84.

Cowan, N., Beschin, N., Della Sala, S., 2004. Verbal recall in amnesiacs under conditions of diminished retroactive interference. Brain 27, 825–834.

Craig, M., Della Sala, S., Dewar, M., 2014. Autobiographical thinking interferes with episodic memory consolidation. PLOS ONE 9 (4), 1–9.

Cushman, L.A., Como, P.G., Booth, H., Caine, E.D., 1988. Cued recall and release from proactive interference in Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology 10, 685–692.

Dale, H.C.A., 1964. Retroactive interference in short-term memory. Nature 203, 1408.

Darby, K., Sloutsky, V.M., 2013. Proactive and retroactive interference effects in development. In: Knauff, M., Pauen, M., Sebanz, N., Wachsmuth, I. (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 2130–2135.

Davis, H.P., Squire, L.R., 1984. Protein synthesis and memory: a review. Psychological Bulletin 96 (3), 518–559.

Della Sala, S., Cowan, N., Beschin, N., Perini, M., 2005. Just lying there, remembering: improving recall of prose in amnesic patients with mild cognitive impairment by minimizing retroactive interference. Memory 13, 435–440.

Della-Maggiore, Valeria, 2005. Motor memory consolidation, night and day. The Journal of Neuroscience 25, 9067–9069.

Dewar, M., Della Sala, S., 2013. Interference with memory. In: Pashler, H. (Ed.), The Encyclopaedia of the Mind. SAGE, Thousand, Oaks, CA.

Dewar, M.T., Cowan, N., Della Sala, S., 2007. Forgetting due to retroactive interference: a fusion of Müller and Pilzecker's (1900) early insights into forgetting and recent research on anterograde amnesia. Cortex 43, 616–634.

Dewar, M., Fernandez-Garcia, Y., Cowan, N., Della Sala, S., 2009. Delaying interference enhances memory consolidation in amnesic patients. Neuropsychology 23, 627–634.

Dewar, M., Cowan, N., Della Sala, S., 2010a. Forgetting due to retroactive interference in amnesia. Findings and implications. In: Sala, Della (Ed.), Forgetting. Psychology Press, Brighton, pp. 185–209.

Dewar, M., Della Sala, S., Beschin, N., Cowan, N., 2010b. Profound retroactive interference in anterograde amnesia: what interferes? Neuropsychology 24, 357–367.

Dewar, M., Pesallaccia, M., Cowan, N., Provinciali, L., Della Sala, S., 2012. Insights into spared memory capacity in amnestic MCI and Alzheimer's disease via minimal interference. Brain and Cognition 78, 189–199.

Diekelmann, S., Wilhelm, I., Born, J., 2009. The *whats* and *whens* of sleep-dependant memory consolidation. Sleep Medicine Reviews 13, 309–321.

Dudai, Y., Morris, R.G.M., 2000. To consolidate or not to consolidate: what are the questions? In: Bolhuis, J.J. (Ed.), Brain, Perception, Memory. Advances in Cognitive Sciences. Oxford University Press, Oxford, pp. 149–162.

Dudai, Y., 2004. The neurobiology of consolidations, or, how stable is the engram? Annual Review of Psychology 55, 51–86.

Ebert, P.L., Anderson, N.D., 2009. Proactive and retroactive interference in young adults, healthy older adults, and older adults with amnestic mild cognitive impairment. Journal of International Neuropsychological Society 15, 83–93.

Ellenbogen, J.M., Payne, J.D., Stickgold, R., 2006a. The role of sleep in declarative memory consolidation: passive, permissive, active or none? Current Opinion in Neurobiology 16, 716–722.

Ellenbogen, J.M., Hulbert, J.C., Stickgold, R., Dinges, D.F., Thompson-Schill, S.L., 2006b. Interfering with theories of sleep and memory: sleep, declarative memory, and associative interference. Current Biology 16, 1290–1294.

Empson, J.A.C., Clarke, P.R.F., 1970. Rapid eye movements and remembering. Nature 227, 287–288.
Engen, T., Ross, B.M., 1973. Long-term memory of odors with and without verbal descriptions. Journal of Experimental Psychology 100, 221–227.
Gaskell, M.G., Dumay, N., 2003. Lexical competition and the acquisition of novel words. Cognition 89, 105–132.
Glenberg, A.M., Swanson, N.G., 1986. A temporal distinctiveness theory of recency and modality effects. Journal of Experimental Psychology: Learning, Memory, and Cognition 12, 3–15.
Godden, D.R., Baddeley, A., 1975. Context-dependent memory in two natural environments: on land and underwater. British Journal of Psychology 66, 325–331.
Goedert, K., Willingham, D., 2002. Patterns of interference in sequence learning and prism adaptation inconsistent with the consolidation hypothesis. Learning and Memory 9, 279–292.
Gottselig, J.M., Hofer-Tinguely, G., Borbély, A.A., Regel, S.J., Landolt, H.P., Rétey, J.V., Achermann, P., 2004. Sleep and rest facilitate auditory learning. Neuroscience 127, 557–561.
Graf, P., Schacter, D.L., 1987. Selective effects of interference on implicit and explicit memory for new associations. Journal of Experimental Psychology: Learning, Memory, and Cognition 13, 45–53.
Hanseeuw, B.J., Seron, X., Ivanoiu, A., 2012. Increased sensitivity to proactive and retroactive interference in amnestic mild cognitive impairment: new insights. Brain and Cognition 80, 104–110.
Hartman, M., Hasher, L., 1991. Aging and suppression: memory for previously relevant information. Psychology of Aging 6, 587–594.
Howe, M.L., 1995. Interference effects in young children's long-term retention. Developmental Psychology 31, 579–596.
Hupbach, A., Gomez, R., Hardt, O., Nadel, L., 2007. Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. Learning and Memory 14, 47–53.
Incisa della Rocchetta, A., Milner, B., 1993. Strategic search and retrieval inhibition: the role of the frontal lobes. Neuropsychologia 31, 503–524.
Izquierdo, I., Schröder, N., Netto, C.A., Medina, J.H., 1999. Novelty causes time-dependent retrograde amnesia for one-trial avoidance in rats through NMDA receptor- and CaMKII-dependent mechanisms in the hippocampus. The European Journal of Neuroscience 11, 3323–3328.
Jarrard, L.E., 1975. Role of interference in retention by rats with hippocampal lesions. Journal of Comparative Physiology and Psychology 89, 400–408.
Jenkins, J.B., Dallenbach, K.M., 1924. Oblivescence during sleep and waking. American Journal of Psychology 35, 605–612.
Johansson, M., Aslan, A., Bäuml, K., Gäbel, A., Mecklinger, A., 2007. When remembering causes forgetting: electrophysiological correlates of retrieval-induced forgetting. Cerebral Cortex 17, 1335–1341.
Köster, E.P., Degel, J., Piper, D., 2002. Proactive and retroactive interference in implicit odor memory. Chemical Senses 27, 191–206.
Kremer, P., Spittle, M., Malseed, S., 2011. Retroactive interference and mental practice effects on motor performance: a pilot study. Perceptual Motor Skills 113, 805–814.
Lahl, O., Wispel, C., Willigens, B., Pietrowsky, R., 2008. An ultra-short episode of sleep is sufficient to promote declarative memory performance. Journal of Sleep and Rest 17, 3–10.
Lawless, H., Engen, T., 1977. Associations to odors: interference, mnemonics, and verbal labelling. Journal of Experimental Psychology 3, 52–59.
Lee, K., Bussey, K., 2001. Children's susceptibility to retroactive interference: the effects of age and degree of learning. Journal of Experimental Child Psychology 80, 372–391.
Lewandowsky, S., Brown, G.D.A., Wright, T., Nimmo, L.M., 2006. Timeless memory: evidence against temporal distinctiveness models of short-term memory for serial order. Journal of Memory and Language 54, 20–38.
Lustig, C., Hasher, L., 2001. Implicit memory is not immune to interference. Psychological Bulletin 127, 618–628.
Lynch, M.A., 2004. Long-term potentiation and memory. Physiological Reviews 84, 87–136.
Maquet, P., Laureys, S., Peigneux, P., 2003. Experience-dependant changes in cerebral activation during human REM sleep. Nature: Neuroscience 3, 831–836.
Massaro, D.W., 1970. Consolidation and interference in the perceptual memory system. Perception and Psychophysics 7, 153–156.
McGaugh, J., 2000. Memory – a century of consolidation. Science 287, 248–251.
McGeoch, J.A., McDonald, W.T., 1931. Meaningful relation and retroactive inhibition. American Journal of Psychology 43, 579–588.
Mednick, S.C., Cai, D.J., Shuman, T., Anagostaras, S., Wixted, J.T., 2011. An opportunistic theory of cellular and systems consolidation. Trends in Neuroscience 34, 504–514.
Mensink, G., Raajmakers, J.G., 1988. A model for interference and forgetting. Psychological Review 95, 434–455.
Miller, E.K., Cohen, J.D., 2001. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24, 167–202.
Morrin, M., Krishna, A., Lwin, M.O., 2011. Is scent-enhanced memory immune to retroactive interference? Journal of Consumer Psychology 21, 354–361.
Morris, R.G.M., 2003. Long-term potentiation and memory. Philosophical Transactions of the Royal Society of London 358, 643–647.
Müller, G.E., Pilzecker, A., 1900. Experimentelle Beiträge zur Lehre vom Gedächtniss. In: Zeitschrift für, Psychologie. Ergänzungsband, vol. 1, pp. 1–300.
Nelson, D.L., Keelman, P.D., Negrao, M., 1989. Word-fragment cueing: the lexical search hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 388–397.
Norman, K.A., Newman, E., Detre, G., 2007. A neural network model of retrieval-induced forgetting. Psychological Review 114, 887–953.
Olsson, M.J., Lungren, E.B., Soares, S.C., Johansson, M., 2009. Odor memory performance and memory awareness: a comparison to word memory across orienting tasks and retention intervals. Chemical Perception 2, 161–171.
Osgood, C.E., 1949. The similarity paradox in human learning: a resolution. Psychology Review 56, 132–143.
Oudiette, D., Antony, J.W., Creery, J.D., Paller, K.A., 2013. The role of memory reactivation during wakefulness and sleep in determining which memories endure. The Journal of Neuroscience 33, 6672–6678.
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Archives of Neurology 56, 303–308.
Postman, L., Alper, T.G., 1946. Retroactive inhibition as a function of the time of interpolation of the inhibitor between learning and recall. The American Journal of Psychology 59, 439–449.
Rasch, B., Buchel, C., Gais, S., Born, J., 2007. Odor cues during slow-wave-sleep prompt declarative memory consolidation. Science 315, 1426–1429.
Ribot, T., 1881. Les Maladies de la Memoire (Diseases of Memory). Appleton-Century-Crofts, New York.
Ribot, T., 1882. Diseases of Memory: An Essay in Positive Psychology. Kegan Paul, Trench, London.
Robertson, E.M., Press, D.Z., Pascual-Leone, A., 2005. Off-line learning and the primary motor cortex. The Journal of Neuroscience 25, 6372–6378.
Robertson, E.M., Pascual-Leone, A.P., Miall, R.C., 2004. Current concepts in procedural consolidation. Nature Reviews: Neuroscience 5, 1–7.
Rossi-George, A., Rovee-Collier, C., 1999. Retroactive interference in 3-month-old infants. Developmental Psychobiology 35, 167–177.
Schacter, D.L., Cooper, L.A., 1993. Implicit and explicit memory for novel visual objects: structure and function. Journal of Experimental Psychology: Learning, Memory, and Cognition 19, 995–1009.
Shadmehr, R., Brashers-Krug, T., 1997. Functional stages in the formation of human long-term motor memory. Journal of Neuroscience 17, 409–419.
Shadmehr, R., Holcomb, H., 1997. Neural correlates of motor memory consolidation. Science 277, 821–825.
Shimamura, A.P., Jurica, P.J., Mangels, J.A., Gershberg, F.B., Knight, R.T., 1995. Susceptibility to memory interference effects following frontal lobe damage: findings from tests of paired-associate learning. Journal of Cognitive Neuroscience 7, 144–152.
Skaggs, E.B., 1925. A discussion on the temporal point of interpolation and degree of retroactive inhibition. The Journal of Comparative Psychology 16, 411–414.
Sloman, S.A., Hayman, C.G., Ohta, N., Law, J., Tulving, E., 1988. Forgetting in primed fragment completion. Journal of Experimental Psychology: Learning, Memory, and Cognition 14, 223–239.
Solesio, E., Lorenzo-López, L., Campo, P., Lopez-Frutos, J.M., 2009. Retroactive interference in normal gaining: a magnetoencephalography study. Neuroscience Letters 456, 85–88.
Solesio-Jofre, E., Lorenzo-López, L., Gutiérrez, R., López-Frutos, J.M., Ruiz-Vargas, J.M., Maestú, F., 2011. Age effects on retroactive interference during working memory maintenance. Biological Psychology 88, 72–82.
Tendolkar, I., Doyle, M.C., Rugg, M.D., 1997. An event-related potential study of retroactive interference in memory. Neuroreport 8, 501–506.
Tilley, A.J., Empson, J.A.C., 1978. REM sleep and memory consolidation. Biological Psychology 6, 293–300.
Traykov, L., Raoux, N., Latour, F., Gallo, L., Hanon, O., Baudic, S., Bayle, C., Wenisch, E., Remy, P., Rigaud, A.S., 2007. Executive functions deficit in mild cognitive impairment. Cognitive Behavioural Neurology 20, 219–224.
Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–402.
Tulving, E., 1983. Elements of Episodic Memory. Oxford University Press, London.

Tulving, E., 1985. How many memory systems are there? American Psychologist 40, 385–398.
Tulving, E., Thomson, D.M., 1973. Encoding specificity and retrieval processes in episodic memory. Psychological Review 80, 352–373.
Turati, C., 2008. Newborn's memory processes: a study on the effects of retroactive interference and repetition priming. Infancy 6, 557–569.
Underwood, B.J., Postman, L., 1960. Extra-experimental sources of interference in forgetting. Psychological Review 67, 73–95.
Walk, H.A., Johns, E.E., 1984. Interference and facilitation in short-term memory for odors. Perception & Psychophysics 36, 508–514.
Walker, M.P., Brakefield, T., Seidman, J., Morgan, A., Hobson, J.A., Stickgold, R., 2003. Sleep and the time course of motor skill learning. Learning and Memory 10, 275–284.
Wickelgren, W.A., 1965. Acoustic similarity and retroactive interference in short-term memory. Journal of Verbal Learning and Verbal Behaviour 4, 53–61.
Winocur, G., Moscovitch, M., Bruni, J., 1996. Heightened interference on implicit, but not explicit, tests of negative transfer: evidence from patients with temporal lobe lesions and normal old people. Brain and Cognition 30, 44–58.
Wixted, J.T., 2004. The psychology and neuroscience of forgetting. Annual Review of Psychology 55, 235–269.
Wixted, J.T., 2005. A theory about why we forget what we once knew. Current Directions in Psychological Science 14, 6–9.
Wixted, J.T., 2011. An opportunistic theory of cellular and systems consolidation. Trends in Neuroscience 34, 505–514.

Revolution

Richard Stahler-Sholk, Eastern Michigan University, Ypsilanti, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'revolution' has evolved from a general notion of political 'turning-over' to refer more specifically to a rapid, fundamental transformation of the state and social structure, accompanied by mass uprising from below. Theories of revolution that emphasize structural conditions tend to focus on the revolutionary seizure of state power, while theories emphasizing actor motivations view revolution as an open-ended process of social transformation. This article examines the definitions and alternative explanatory frameworks, and considers the future of revolution in a changing global context.

In the 150 years since the *Communist Manifesto*, revolution has mainly occurred not in the heart of industrial capitalism, but in countries at the periphery, with significant peasant participation. Theories of revolution in the twentieth century (*see* Revolutions, Theories of; Revolutions, History of) evolved from descriptive 'natural cycles' of revolution, to explanations emphasizing social-psychological factors, to structuralist theories, adding state and international structures to Marxian class analysis of history. Fourth-generation theories re-emphasize agency, drawing on cultural studies to examine ideological frameworks and subjective motivations of actors. The end of the Cold War contributed to new interest in a broader conception of the revolutionary potential within civil society, including the mainly nonviolent 'people's power' social movements that toppled Communist Party states in Eastern Europe and dictatorship in the Philippines; the 'negotiated revolutions' in Central America that opened space for political democratization; the Zapatista call for an awakening of Mexican civil society; the 'Arab spring' uprisings of 2011; and resistance against globalization, suggesting the intriguing prospect of an emerging global civil society.

Conceptualizing 'Revolution'

The earliest studies of comparative politics noted recurring conflict between elites and populace, sometimes leading to power 'revolving' from one group of rulers to another within a society. After the French Revolution of 1789 (*see* French Revolution, The), the term 'revolution' took on new meaning as a radical overthrow of the social and political order itself (Goldstone, 1998). The nineteenth-century European challenges to monarchic authority by the rising bourgeoisie and the class conflicts of industrial capitalism, punctuated by Marx and Engels' [1848] (1998) publication of *The Communist Manifesto*, focused attention on historical processes in analyzing revolution. After Lenin and Mao adapted Marxist theory to the Russian and Chinese revolutions (*see* Russian Revolution, The; Chinese Revolutions: Twentieth Century), the decades following World War II saw a wave of national liberation movements against colonialism and neo-colonialism (*see* Third World; Colonization and Colonialism, History of; Imperialism, History of). Contrary to Marx's expectation that revolution would occur in the most industrialized capitalist countries, led by the proletariat, revolutions in the twentieth century occurred in the global periphery through mobilization of peasants and artisan workers.

The Mexican Revolution of 1910 (*see* Mexican Revolution, The) and the revolution that eventually brought the Communist Party to power in China in 1949 (*see* Chinese Revolutions: Twentieth Century) were not specifically against colonial rule, since Mexico like most of Latin America had won independence in the early 1800s and China was never fully colonized. In both of these major revolutions, the peasantry played a key role, and anti-imperialist nationalism was an important unifying ideology in mobilizing a much broader base than just the incipient industrial working class.

Common Usage

The eclectic 'socialism' of many Third World revolts and nationalist decolonization movements (*see* National Liberation Movements), and their use of unorthodox tactics of guerrilla warfare – analyzed by practitioners such as Mao Zedong, Che Guevara, and Vo Nguyen Giap – led those characteristics to be generically associated with 'revolution.' This usage was reinforced by Cold War counterinsurgency strategies aimed at forestalling radical change in national political systems, which were seen as extensions of the global East–West contest. The collapse of most colonialism after World War II (*see* Colonization and Colonialism, History of) was sometimes a revolutionary moment, in those cases where it entailed not only reformulating the state to throw off external rule but also fundamentally reordering internal social relations. The large number of newly independent states with a history of having been colonized and/or lagging in development were collectively labeled the 'Third World'; and the Cold War imposed a simplistic dichotomy in which they either remained in the Western camp (where the U.S. had replaced the European colonizers as hegemon), or underwent socialist revolutions and aligned with the Soviet Union. This framework never accurately captured the complex and varied reality of the two-thirds of the world's countries that were put into this box, but it did focus attention on the potential for radical change on the periphery of the global system. National liberation movements that posed radical challenges to the domestic social order, often long after decolonization, included notable cases such as Cuba (*see* Cuban Revolution, The) in 1959, southeast Asia (*see* Vietnamese Revolution, The) and southern Africa in the 1960s and 1970s, and Central America in the 1970s and 1980s.

The term 'revolution' moved away from this historically specific usage after the conservative religious turn in the 1979 Iranian revolution, the Eastern-Central European uprisings against Communist Party states after 1989 (*see* Revolutions of

1989–90 in Eastern Central Europe), and the end of the Cold War and other global power shifts at the dawn of the twenty-first century.

Sometimes the term 'revolution' is used loosely to refer to any kind of paradigmatic shift, as in 'the Industrial Revolution,' 'the Reagan revolution,' or 'the behaviorist revolution' in social sciences. In another variant usage, 'revolution' is contrasted with reformism to connote a radical political program, short time horizon, and willingness to use extreme (violent) means; as in Malcolm X's call for African-American liberation "by any means necessary," or J.F. Kennedy's warning that "those who make peaceful revolution impossible will make violent revolution inevitable" in Latin America.

Political Revolution and Social Revolution

The term 'political revolution' is employed to specify the application of the concept to a national political system. This usage does not include issue-specific social movements, nor all sudden changes of political power through *coup d'état* or civil war (*see* Civil Wars; Insurgency; Social Protest; Social Movements and Political Violence; Social Movements, History of: General). However, 'political revolution' can refer to a wide range of revolts, rebellions, and mobilizational seizures of power – including those led from above such as the Japanese Meiji Restoration of 1868 (*see* Meiji Revolution, The), Nasser's 1952 coup in Egypt, or the 1944–54 period in Guatemala – if they propose major lasting changes in the political order (Goldstone, 1998; DeFronzo, 2006).

Skocpol's influential work (1979: p. 4) proposed a more restrictive focus on 'social revolution':

> Social revolutions are rapid, basic transformations of a society's state and class structures, and they are accompanied and in part carried through by class-based revolts from below. Social revolutions are set apart from other sorts of conflicts and transformative processes above all by the combination of two coincidences: the coincidence of societal structural change with class upheaval, and the coincidence of political with social transformation.

By this narrower definition, social revolution is a relatively rare phenomenon.

Alternative Ways of Understanding Revolution

Goldstone (1980) usefully grouped explanations of revolution into three generations of theory, with a fourth generation emerging since the late twentieth century. In the 1920s and 1930s, a first generation of 'natural histories' of revolutions used techniques of 'thick description' and inductive logic based on empirical observation of major Western revolutions. These efforts to identify key actors, factions, sequences, and patterns helped establish the phenomenon of revolution as a distinctive subject of comparative study.

From Description to Explanation

In the 1950s and 1960s, the upheaval associated with decolonization led to a second generation of 'general theories of political violence.' Some focused on the individual psychological motivations of people experiencing frustrated expectations, as in the 'relative deprivation' approach. Frantz Fanon's *The Wretched of the Earth* (1965) interpreted political violence as redemption, healing the psychological wounds of colonization. Another branch of general theories sought explanations of political violence at the societal rather than individual level. Huntington's (1968) influential work drew on modernization theory, attributing revolutionary upheaval to economic modernization in developing countries, which raised expectations faster than they could be accommodated by the lagging development of political institutions (*see* Modernization, Sociological Theories of).

Critics of modernization theory in the late 1960s and 1970s pushed general theories of political violence in new directions. Tilly (1978) challenged the assumption that social systems were in natural equilibrium, arguing that conflict was normal, so explanations of political violence should focus on the relative resource mobilization capacity of power contenders. Moore (1966), in a sweeping macrohistorical comparative work, showed that there is no single, unilinear process of modernization. Extending Marx's class analysis to a more detailed classification of landlord–peasant relations, he combined this variable with distinct political forms of class struggle in agrarian societies undergoing commercial modernization, to map out different possible revolutionary outcomes.

Moore's (1966) work was the point of departure for Skocpol's (1979) groundbreaking study most associated with a third generation of 'structural theories' of revolution (see also her critique of Moore in Skocpol, 1994). Shifting from the actor-centered focus that characterized general theories of violence, these theories highlighted the structural relations between units that created revolutionary opportunities. They built multicausal explanations based not only on class dynamics, but also on other aspects of the structures of society, the state, and the international system. Structural theories identified revolutionary conditions in the combination of administrative breakdown of the state, dissident elite political movements, a permissive world context, and widespread rural rebellion. Some focused on the types of state structures most vulnerable to revolutionary overthrow, including politically exclusionary authoritarian regimes, neopatrimonial dictatorships, and directly ruled colonies. Goldstone (1991) examined the interaction of factors such as population growth with political institutions and social structure, developing quantitative indicators of revolutionary conditions. Structuralists defined revolution in terms of successful outcomes, distinguishing revolutionary seizure of state power from cases in which one or more necessary conditions were absent.

Structure versus Agency: Emerging Fourth-Generation Theory

An alternative to defining revolution as a set of structural conditions is to conceptualize it in terms of distinct systems of ideas and beliefs that subjectively mobilize actors. An emerging fourth generation of 'actor-centered' theories of revolution emphasize the defining role of ideology, culture, and identity of the subjects who make revolution. To borrow Mao's metaphorical observation that a single spark can ignite a prairie fire, these theorists shifted attention from the wind, temperature, and dryness of the grasslands to the spark itself. This contrasts

with Skocpol's (1979) quotation of Wendell Phillips's statement that "revolutions are not made; they come."

Fourth-generation theories inspired new debates about whether revolutions can be accurately predicted on the basis of structural conditions, particularly in light of the surprising developments in the Iranian and Eastern European revolutions of the late twentieth century. Structuralists began re-examining the role of ideology in revolutions and the interaction of states with other domestic and international structures (Skocpol, 1994). New insights from cultural studies and feminist theory challenged the more state-centered concepts of revolution (Foran et al., 2008). These new perspectives defined revolution not as an outcome achieved by overthrow of the state, but as an ongoing process of transformation of society, involving both the institutionalization of a new order and the consolidation of a new ideology, guided with different degrees of success by different styles of leadership. The emphasis on process rather than outcome seemed more fruitful in light of the ambiguous and indeterminate course of some revolutions, including the negotiated end in the late 1990s to armed conflicts in Central America.

Islamism and Other Revolutions of Values

The 1979 Islamic revolution in Iran highlighted the growing influence of Islamism, also known as political Islam, as an ideological impetus for mass mobilization around radical social and political change. The revolutionary forces arrayed against the Shah of Iran were a diverse coalition, but the religious current led by Ayatollah Ruhollah Khomeini succeeded in gaining power and imposed a theocracy based on the radical clerics' interpretation of Islamic religious law, or shariah. Elsewhere in the Middle East and beyond, a number of groups emerged or expanded that sought a guiding role for Islam in social and political life, though they varied widely in both their strategies for achieving that end and their interpretations of the religion (*see* Islam: Middle East). They ranged from reformist variants such as the Justice and Development Party in Turkey and the Muslim Brotherhood that won Egyptian elections in 2012 after the ouster of Hosni Mubarak, to radical groups such as the Taliban in Afghanistan or the Al Qaeda terrorist network responsible for the 11 September 2001 attack on the World Trade Center. While the Iranian version of Islamism supported a revolution, the Al Qaeda network took aim primarily at U.S. global power, and the Wahhabi movement supported the continuation of conservative monarchy in Saudi Arabia.

Among possible explanations for the rise of radical and sometimes revolutionary political expressions of Islam, political scientist Samuel Huntington posited that the end of the Cold War would see a 'clash of civilizations' in which culture and religious values would be the new axis of fundamental global conflict. This interpretation found support in the analysis of cross-national values surveys (Inglehart and Welzel, 2005; Welzel, 2013) that revised modernization theory to emphasize the way national cultural traditions shaped attitudes and values toward emancipation. Others saw the Islamist phenomenon as fueled by multiple factors, including a backlash against globalization and the corresponding spread of Western culture; moral outrage over corrupt and oppressive regimes in the Arab world; frustration among educated, unemployed youth; and a rejection of U.S. hegemonic pretensions in general and policies in the Middle East in particular.

Religion can be an ideological framework encouraging acquiescence to the status quo – the 'opiate of the masses,' in Marx's famous dictum – or it can provide a unifying message and values framework for potential agents of revolutionary change to mobilize against social injustice. For example, Catholicism in Latin America once served to rationalize the Spanish invasion and centuries of racial/class stratification. Then the Vatican II reforms of the 1960s opened the door for Liberation Theology, a movement advocating a 'preferential option for the poor' within Latin American Catholicism, which influenced revolutionaries in the region as well as a variety of other grassroots initiatives in the succeeding decades. By the late 1980s, the Liberation Theology influence was waning, as Pope John Paul II launched a theological counteroffensive and as more conservative evangelical Protestant religions spread in Latin America.

Another example of fundamental ideological challenge to social hierarchies came from indigenous movements in Latin America. In the poor southern Mexican state of Chiapas, the 1994 Zapatista uprising of Maya peoples rejected the economic impact of neoliberal, free-market policies on the peasantry as well as historic suppression of indigenous rights. The Zapatistas formed their own autonomous governing structures, and their discourse called for a political model of 'governing by obeying' and 'a world in which many worlds may fit.' A mass movement in Bolivia, also motivated by a critique of both global capitalism and ethnic exclusion, overthrew the government and brought the election in 2005 of Evo Morales, the first indigenous president in the country where indigenous peoples are a majority. Some of that government's policies reflected the indigenous cosmovision, including groundbreaking legislation that granted legal rights to Mother Nature. Similarly in Ecuador, the mass mobilization of indigenous people led to a new constitution recognizing the rights of nature and the ancestral communitarian principle of 'Buen Vivir' (the concept of the right to a good life).

Fourth-generation theorists of revolution highlighted what Selbin (2010) calls 'story,' the narratives that have powerful subjective appeal and can generate mass mobilization. The shift of focus from the state to the power of ideas in society, or 'cultures of revolution,' calls attention to the myths, historical figures, and symbols that stir collective memories and action (Foran et al., 2008). Venezuelan President Hugo Chávez invoked the historical figure of the anticolonial liberator Simón Bolívar, along with socialist ideology, to proclaim a 'Bolivarian revolution.' After his election in 1999, Chávez combined radical redistribution, mobilization of the poor, and anti-imperialist rhetoric. These policies represented a radical change in class relations, but they were carried out through state reforms (concentrating executive power) backed by an electoral mandate, generating sharply polarized discourses over whether Chávez should be considered a revolutionary, a populist, or a dictator.

Democratic Revolutions of the Post-Cold-War Era

Tectonic shifts in power configurations at the international level may create historic openings for revolution. The end of the Cold War, punctuated by the taking down of the Berlin Wall in 1989 and the collapse of the Soviet Union in 1991, may be one

such moment. A number of new states and new regimes emerged out of the collapse of erstwhile Communist political systems across East-Central Europe. Ironically, regimes founded on an official ideology that espoused revolution had fallen to popular uprisings. After Mikhail Gorbachev, the last General Secretary of the Communist Party of the Soviet Union, renounced the use of Soviet military force to protect other Communist regimes, the collapse came quickly and in most cases with little violence, suggesting that legitimacy had long since eroded. The Soviet collapse reverberated across other parts of the world that had once been Cold War hot spots, contributing to state collapse in Somalia and economic crisis in Cuba. At the same time, the end of an era of superpower sponsorship of authoritarian client states created a potential opening for democratization elsewhere.

Viewed in historic perspective, some saw the transitions from Communism as a continuation of what Huntington (1991) had identified as a 'third wave' of democracy that began with the mid-1970s collapse of fascism in Portugal and Spain and spread to Latin America and beyond (*see* Democratic Transitions). Notwithstanding some self-congratulatory triumphalism in the West, the fall of Communism had to do with a combination of intrepid civil society activities in Europe, a series of economic and political liberalization initiatives by Gorbachev that revealed the extent of discontent with a stagnant system, and a rapidly shifting calculus by individual actors of their options for 'exit' and 'voice' (Hirschman, 1993; Opp, 1994).

The Overthrow of Communism

While atypical of revolutions in many respects – for example, in their restorative character (Ash, 2009) – the uprisings that overthrew Communist regimes did involve mass uprisings from below, as well as a shift in the founding basis of state power and the accompanying ideology that had rationalized one-party rule. On the other hand, it was not entirely clear that the social basis of power had been fundamentally transformed, at least not in the ways envisioned by those who heralded these transitions as democratic revolutions. The shifts from centrally planned to market economies were often accomplished by opaque and corrupt processes of privatization, leaving concentrations of market power in the hands of the old party and state elites. In the former Yugoslavia, the collapse of the federal state was accompanied by the rise of ultranationalist demagogues and massive ethnic violence. The breakup of the former Soviet Union saw the emergence of corrupt and repressive autocracies in some of the Central Asian and Baltic states, violent conflict over ethnic separatist movements in Chechnya and Georgia, and in Russia itself a new centralization of power and restriction of liberties by Vladimir Putin through his United Russia party. Given these medium-term outcomes, it was debatable whether the uprisings that overthrew Communist regimes should be classified as revolutions, much less democratic revolutions.

Electoral ('Color') Revolutions

In the wake of the democratic setbacks in many post-Communist states, a revived wave of pro-democracy uprisings swept East-Central Europe, the Balkans, and the former Soviet Union beginning around 1996 and continuing into the 2010s. Variously labeled 'electoral revolutions' or 'color revolutions' after the emblematic colors that many of the movements adopted, these uprisings were characterized by coordinated mass protests against fraudulent elections and semi-authoritarian incumbents. They used innovative political symbolism to undermine the legitimacy of transitional post-Communist regimes that were still not fully institutionalized, and to signal (domestically and across countries) the degree of latent discontent in order to encourage others to overcome the barrier of fear. These movements received extensive external support and funding, notably from U.S. governmental, quasi-governmental, and nongovernmental 'democracy promotion' programs, leaving them open to nationalist accusations of Western meddling. The 2012 Russian election, in which Vladimir Putin claimed a sizable victory amid fraud and intimidation, once again raised questions about the scope and durability of the so-called color revolutions. The narrow focus on elections also cast doubt on whether the movements should be considered revolutions.

Rediscovering the Power of Nonviolence

One of the interesting developments in revolutionary movements of the late twentieth and early twenty-first centuries was the renewed salience of nonviolence as a guiding philosophy and/or strategy. Nonviolence political action had a long history, and it played a prominent role in successful social movements such as Mohandas K. Gandhi's struggle against British colonial rule in India in the early twentieth century and the U.S. civil rights movement led by Dr. Martin Luther King, Jr. in the 1960s. The techniques of civil disobedience and other nonviolence resistance were used by part of the South African anti-apartheid movement, particularly from the 1970s through the early 1990s; the Philippine 'People Power Revolution' against the dictatorship of Ferdinand Marcos in 1986; many of the uprisings against Communist rule in Eastern-Central Europe, such as the 1989 'Velvet Revolution' in the former Czechoslovakia; and as a defining feature of the 'color revolutions' from the late 1990s onward. The widespread use of nonviolent civil disobedience in the latter cases may reflect a process of political learning or diffusion of knowledge about the specific vulnerabilities of authoritarian regimes. Indeed, the lessons were deliberately transmitted by one of the foremost contemporary experts on nonviolent political action, whose primer on how to bring down a dictatorship became essential reading for protagonists of the 'color revolutions' (Sharp, 2010).

The notion of a nonviolent revolution seems counterintuitive, since the scope of social and political change inherent in the concept of revolution is almost certain to be intensely resisted by entrenched interests. However, the logic of nonviolent action is not passivity, but rather active confrontation against unjust authority, using techniques such as mass noncooperation to expose the latent structural violence of a system and thereby delegitimize it. Such action may in fact provoke state violence, in which case the opposition may refuse to respond in kind, calculating that the resulting increase in delegitimation will lead to regime collapse. Maintaining a disciplined commitment to nonviolence in such circumstances may be difficult, particularly in the absence of a unifying moral framework of principled nonviolence (such as Gandhi's teachings in India). In fact, many

nonviolent mass uprisings occur in a mixed context in which others have chosen to take up arms.

The post–World War II international system has been based on the principle of state sovereignty and noninterference in the domestic affairs of states, but shifting norms since the late twentieth century may be weakening the protections for authoritarian regimes and creating new openings for mass opposition mobilizing. One such shift is the expanding discourse and institutional framework of universal human rights, most actively promoted by international nongovernmental organizations. Another is the establishment in 2002 of the International Criminal Court to prosecute individual political leaders responsible for genocide, war crimes, or crimes against humanity. A third is the 2005 United Nations endorsement of a 'Responsibility to Protect,' a principle supporting international intervention against regimes committing atrocities against their own people. Given that structuralist theorists of revolution posit an international opening as one of the preconditions for revolution, these shifts along with the end of the Cold War could be part of the explanation for the apparent upsurge in democratic revolutions.

The 'Arab Spring'

Beginning in the spring of 2011, a wave of uprisings in the Arab world rocked some of the most seemingly entrenched autocracies in the Middle East and North Africa. Quickly labeled the 'Arab spring,' 'Arab uprisings,' or 'revolutions,' the protests were initially triggered by the actions of Mohamed Bouazizi, a 26-year-old Tunisian street vendor who set himself on fire after his fruit cart was seized by police. Massive demonstrations subsequently led to the ouster of the regimes of Zine al-Abidine Ben Ali in Tunisia, Muammar Qaddafi in Libya, and Hosni Mubarak in Egypt, as well as ongoing challenges to autocratic rule in Syria, Yemen, Bahrain, and elsewhere in the region.

In seeking to explain these events, structuralist theorists of revolution highlighted the particular susceptibility of personalist or 'sultanistic' regimes to mass opposition organizing. The rapid demographic growth of young, urban, educated populations who were unemployed or underemployed (like Bouazizi, in the informal sector) was cited as another causal factor, aggravated by rising food prices. The class dimension of these uprisings was most evident in Egypt, where workers played a key role in organizing strikes and protests. One structural variable affecting regime survival was the extent of defections among marginal elites, particularly the military, who were sometimes offended by the extreme concentration of wealth and the personal corruption of the rulers.

While the clustering of these uprisings suggested some common causal factors, the motivating grievances and ideological framework differed considerably among the participating groups. There were significant differences between militaries, secular liberal democrats, and Islamists (e.g., Egypt); as well as divided ethnic or religious sectarian (Syria) and tribal (in Libya and Yemen) loyalties and coalitions. Further complicating the outcomes, external actors had varying agendas and degrees of involvement. The U.S., for example, quietly supported NATO intervention on behalf of rebels in Libya, acquiesced to the overthrow of its longtime client Mubarak in Egypt, but backed Saudi use of force against the opposition in Bahrain. Russia and China used Security Council vetoes to block UN coercive action against the Syrian regime's use of violence against opponents. Typically of revolutions, the outcomes were indeterminate: not all resulted in successful transfer of state power, and new regimes did not necessarily reflect a stable consensus of values among those who had joined the mass mobilization. Once again, democracy was not a guaranteed outcome.

One notable feature of the Arab uprisings was the significant participation of youth and women. Women made up about a third of the direct participants in the protests that brought down Mubarak's regime in Egypt. A Yemeni democracy activist, Tawakul Karman, in 2011 became the first Arab woman to win a Nobel Peace Prize. High unemployment was one factor motivating the extensive participation of youth in the protests. Another characteristic of the uprisings was the prominent use of social media (such as blogs, Facebook, and Twitter) for communication and coordination among protesters and their networks of supporters (*see* New Media and Democracy in the Arab World). These technological innovations were more facilitators than causes and did not erase national cultural specificities (Norris and Inglehart, 2011), but they did highlight the ways in which rapid interconnectedness has eroded the relevance of many boundaries and allowed more fluidity in the creation of cognitive communities (Castells, 2012). The imagery, branding, and tactics of the protests spread with startling speed. At the same time, the rapidity of communication can lead to misperceptions, and the plugged-in social sectors may not be representative of the general population. Moreover, just as the rapid transmission of information can fuel a spreading wave of revolution, it can also generate a backlash or preventive measures (for example, in the responses to the Arab uprisings by the governments of Russia, China, and Saudi Arabia).

Rethinking Revolution in the Context of Globalization

Globalization, or the compression of time and space, has reshaped many aspects of the contemporary political world. Alongside states, other axes of power include transnational corporations, global and regional trade regimes, and international financial institutions. Counter-hegemonic or antisystemic globalization-from-below now includes a variety of transnational social movement networks, as well as shifting focal points for the growing 'alter-globalization' movement, such as the World Social Forum of civil society organizations inaugurated in Brazil in 2001, and global protests such as the clashes at the 1999 World Trade Organization (WTO) ministerial meeting in Seattle, and the 2001 summit of the Group of 8 (G-8) industrialized countries in Genoa. The shift in the capitalist world system from the Fordist regime of capital accumulation based on industrial mass production, to a regime of flexible accumulation and globally mobile capital, has been accompanied by the rise of a new regulatory framework of global institutions governed by neoliberal ideology and the emergence of a transnational capitalist class. These changes call for rethinking the concept of revolution (Foran, 2003), as the spaces of contestation move from the state to the world system.

World-systems theory suggests that revolutions should be defined not as discrete phenomena of nation-states, but rather as transformations in the world system (Boswell and Chase-

Dunn, 2000). From this perspective, an interconnected wave of upheavals such as the Arab spring or democratic revolutions can be understood as part of an uprising against the global system. In an influential work, Hardt and Negri (2000) argue that the era of competition among imperialist powers has been replaced by a unitary structure of power they call Empire, a kind of transnational network capitalism which has increasingly privatized public space and reduced the scope for conventional political participation. They argue that Empire actually creates greater potential for revolution, but that instead of a class rising up to seize state power, it will take the form of 'the multitude' resisting the disciplinary regime of global capital through alternative cultural and economic movements.

A possible example of this new way of conceiving revolution would be the Zapatista rebellion of 1994, triggered in part by Mexico's integration into the global capitalist economy as symbolized by entry into the North American Free Trade Agreement. While claiming the right to community-based autonomy from the state, the seemingly local Zapatista movement also challenged the larger logic of subordination to the dictates of global capital, in ways that inspired global networks of resistance. The Zapatista focus on direct participatory democracy, and on connecting such concepts as community-based autonomy, solidarity, and the commons to a critique of global capitalism, served as an inspiration for the emblematic 1999 Seattle WTO protest and many other 'alter-mundista' or 'alter-globalization' movements.

Such antisystemic movements focus more on transforming social relations than on overthrowing the state. Holloway (2002) has suggested a new meaning of revolution, arguing that it is possible and necessary to change the world without taking power, at least in the traditional state-centric and hierarchical understandings of power. Postcolonial and feminist theories have introduced the notion of empowerment as opposed to power-over, and also questioned the hidden power relations in the way concepts are constructed and knowledge is produced. Holloway's thesis remains controversial as states continue to wield conventional forms of power, but it is evocative. At the end of the Cold War when Nicaragua's revolutionaries made history by turning over the state power they had won through insurrection in 1979 after losing elections in 1990, Sandinista leader Daniel Ortega vowed they would 'govern from below.' This vision, albeit unrealized, suggested a new way of thinking about where the essence of a revolution is found, i.e., in empowerment that is constructed within society. The demise of state socialism in Eastern-Central Europe also fueled the global quest for radical democratic alternatives, rooted in autonomous action within civil society.

As globalization undermined territorial definitions of social and political organization, shifting decision-making authority to transnational structures, revolution seemed less about seizing state power. From the Internet-savvy Zapatista rebels of Chiapas, Mexico, to the transnational anti-sweatshop movement and the 1999 Seattle protest against the WTO, there were intriguing signs of an emerging global civil society. The global economic crisis beginning in 2008, with a potentially catastrophic environmental crisis looming in the background, was accompanied by a series of worldwide protests including the '*indignados*' of Spain, anti-austerity uprisings in Greece, student protests in Chile and Mexico, and the Occupy movement that began on Wall Street and quickly spread. All of these mobilizations were directed at global forces and not just the nation-state, and all sought to reclaim the commons from the encroachment of centuries of global capitalism. These new forms of resistance against corporate-dominated globalization stretched the boundaries of existing concepts of 'revolution.'

See also: American Revolution, The; Chinese Revolutions: Twentieth Century; Civil Wars; Colonization and Colonialism, History of; Cuban Revolution, The; English Revolution, The; French Revolution, The; Meiji Revolution, The; Mexican Revolution, The; National Liberation Movements; Revolution, Theories of; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, History of; Revolutions, Sociology of; Russian Revolution, The; Social Movements and Political Violence; Social Movements, History of: General; Vietnamese Revolution, The.

Bibliography

Ash, T.G., 3 December 2009. Velvet revolution: the prospects. The New York Review of Books.
Boswell, T., Chase-Dunn, C., 2000. The Spiral of Capitalism and Socialism. Lynne Rienner Publishers, Boulder, CO.
Castells, M., 2012. Networks of Outrage and Hope: Social Movements in the Internet Age. Polity Press, Cambridge and Malden, MA.
DeFronzo, J.V., 2006. Revolutionary Movements in World History, 3 vols. ABC-CLIO, Santa Barbara.
Fanon, F., 1965. The Wretched of the Earth. Grove Press, New York.
Foran, J. (Ed.), 2003. The Future of Revolutions: Rethinking Radical Change in the Age of Globalization. Zed Books, London and New York.
Foran, J., Lane, D., Zivkovic, A. (Eds.), 2008. Revolution in the Making of the Modern World. Routledge, London and New York.
Goldstone, J.A., 1980. Theories of revolution: the third generation. World Politics 32, 425–453.
Goldstone, J.A., 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley, CA.
Goldstone, J.A. (Ed.), 1998. The Encyclopedia of Political Revolutions. Congressional Quarterly, Washington, DC.
Hardt, M., Negri, A., 2000. Empire. Harvard University Press, Cambridge, MA.
Hirschman, A.O., 1993. Exit, voice, and the fate of the German democratic republic. World Politics 45, 173–202.
Huntington, S.P., 1968. Political Order in Changing Societies. Yale University Press, New Haven, CT.
Huntington, S.P., 1991. The Third Wave: Democratization in the Late Twentieth Century. University of Oklahoma Press, Norman.
Holloway, J., 2002. Change the World without Taking Power: The Meaning of Revolution Today. Pluto, London.
Inglehart, H., Welzel, C., 2005. Modernization, Cultural Change, and Democracy: The Human Development Sequence. Cambridge University Press, Cambridge, UK.
Marx, K., Engels, F., 1998. The Communist Manifesto. Verso, London, 150th anniversary edn. with essay by Hobsbawm, E.
Moore Jr., B., 1966. Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World. Beacon Press, Boston, MA.
Norris, P., Inglehart, R., 2011. Cosmopolitan Communications: Cultural Diversity in a Globalized World. Cambridge University Press, Cambridge, UK.
Opp, K.-D., 1994. Repression and revolutionary action: East Germany in 1989. Rationality and Society 6, 101–138.
Selbin, E., 2010. Revolution, Rebellion, Resistance: The Power of Story. Zed Books, London and New York.
Sharp, G., 2010. From Dictatorship to Democracy: A Conceptual Framework for Liberation, fourth ed. The Albert Einstein Institution, East Boston.
Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. Cambridge University Press, Cambridge, UK.
Skocpol, T., 1994. Social Revolutions in the Modern World. Cambridge University Press, Cambridge, UK.
Tilly, C., 1978. From Mobilization to Revolution. Addison-Wesley, Reading, MA.
Welzel, C., 2013. Freedom Rising: Human Empowerment and the Quest for Emancipation. Cambridge University Press, Cambridge, UK.

Revolution, Theories of

Rod Aya, Amsterdam, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Revolution has different meanings, and theories of revolution do not all explain the same facts. Some theories explain revolution in the sense of a revolutionary movement – an attempt to change government, regime, society, or all three by violence. Others explain revolution in the sense of a revolutionary outcome – a successful change of government, regime, society, or all three by violence. Still others explain revolution in the sense of a revolutionary situation – a contest for state power by violence. But all theories of revolution shortlist explanatory causes and validate the explanation with case histories or statistics to show by the comparative method that where these causes occur, a revolution (by some definition) occurs as well and that where they do not occur, revolution (by that definition) does not occur either. Grouped by causes they cite, theories of revolution make up four types – psychological, functional, structural, and political. Psychological and functional theories mostly explain revolutionary movements; structural theories mostly explain revolutionary outcomes; and political theories mostly explain revolutionary situations. Theories of revolution are contentious, but few controvert the ancient wisdom that revolution needs hope of success and that hope feeds on a weak government torn by faction.

Theories of revolution are (or purport to be) causal explanations of revolution. They say if certain sufficient conditions (the causes) occur, then revolution (the effect) will occur as well. But *revolution* has different meanings, and theories of revolution do not all explain the same facts. Different theories explain different facts as well as cite different causes – different circumstances or situations that (by hypothesis) motivate or constrain people to commit revolutionary violence (Aya, 2001). Before analyzing theories of revolution and the why-question they address, therefore, one must distinguish between the different social facts those theories would explain – the different historical events *revolution* refers to.

Definition

Revolution (as used in the theories-of-revolution literature) has three different meanings – it denotes three different kinds of social fact or historical event that theories of revolution would explain: a revolutionary movement, a revolutionary outcome, and a revolutionary situation. Examples document the claim that *revolution* is defined these three ways.

Revolution in the sense of a revolutionary movement is a concerted effort to change government, regime, society, or all three by violence. By one definition it is "internal war" (Eckstein, 1965) – an attempt by "subordinate groups" to change government, policy, regime, or society by violence (Zagorin, 1982, 1, p. 17). By another definition it is an attempt to "establish a new moral and material world and to impose, or evoke, radically new patterns of day-to-day conduct," again by violence: "A holy commonwealth, a republic of virtue, a communist society – these are the goals revolutionaries seek" (Walzer, 1998, p. 128). Replacing "subordinate groups" with "revolutionaries," this definition includes new rulers like the Committee of Public Safety behind the Terror when the revolution was "in the government, not in the populace" (Palmer, 2005, p. 294). With that, *revolution* meant new regime (Malia, 2006) – save in Japan where the new regime called itself the Restoration.

These definitions highlight revolutionary violence as the chosen means to a wanted end in accord with the view that social action makes sense no other way (Weber, 1968; Coleman, 1990; Popper, 2002). But by neither definition must a revolutionary movement succeed in order to be called a revolution – it can fail like the Committee of Public Safety doomed by faction ramified through the Convention, the revolutionary legislature that elected it (Palmer, 2005).

Revolution in the sense of a revolutionary outcome is a successful change of government, regime, society, or all three by violence. By a venerable definition it is a "violent transference of power and property," that is, a "seizure of power and redistribution of property" (Syme, 1939, pp. vii, 2). By newer, more ambitious definitions it is a "rapid, complete, and violent change in values, social structure, political institutions, governmental policies, and social-political leadership" (Huntington, 1968, p. 344); a "violent, fundamental, and abrupt change of incumbent elites, status and class relations, institutions, values, symbols, and myths" (Mayer, 1971, p. 47); or a violent change of "status and power, … economic organization and property ownership, and … symbols and beliefs that justify the distribution of power, status, and wealth" (Goldstone, 1991, p. 11). And by a narrower definition confined to "successful social revolutions," it is "rapid, basic transformations of a society's state and class structures … accompanied and in part carried through by class-based revolts from below" (Skocpol, 1979, p. 4).

Success is crucial – revolution is a misnomer unless revolutionaries "win state power and use it to effect a major property transfer" (Gouldner, 1979, p. 9) plus whatever else the definition of revolution as a revolutionary outcome stipulates, say, a "complete redesign of state, society, economy, and culture" (Pipes, 1991, p. xxi).

Revolution in the sense of a revolutionary situation is a contest for state power by violence. By one definition it is "dual power," that is, a "trial by battle" where government and

opposition "split the state superstructure" (Trotsky, 1977, pp. 223–232). By another definition it is "multiple sovereignty," which starts when the opposition rules people nominally subject to the government and ends when it defeats, succumbs to, or cuts a deal with the government (Tilly, 1978, pp. 189–222). By other, convergent definitions it is a "revolutionary political crisis" (Skocpol, 1979) or "state breakdown" (Goldstone, 1991) where civil war decides sovereignty.

A revolution in one sense can occur without a revolution in both other senses occurring as well. A revolutionary movement can occur without a revolutionary outcome occurring because revolutionaries can lose like the Taiping Tianguo whose civil war with the Qing regime killed about 20 million people (Jen, 1973). A revolutionary outcome can occur without a revolutionary situation occurring because the new government in power can make a "revolution from above" like Stalin whose collectivization of farming and party purge also killed about 20 million people (Conquest, 1990) and because revolutionaries can take power legally like Hitler (Turner, 1996) whose agents killed about that many civilians – roughly a third of them Jews (Davies, 1997) – in World War II, showing what a "fanatical movement" atop a "well-organized state" can do (Sheehan, 2008, p. 154). And a revolutionary situation can occur without a revolutionary outcome occurring because the government can win or else make a deal as in Central America where "socialist revolution from below" led to a "transition to democracy" (Paige, 1997, p. 358). But neither a revolutionary situation nor a revolutionary outcome can occur without a revolutionary movement occurring because they both entail revolutionary violence, which is organized violence, that is, collective action – no big-time dirty work without motivated, coordinated perpetrators (Aya, 2001).

These three meanings of revolution accord with social science usage, which calls a change of government by violence a coup d'état; a change of regime as well as of government by violence a political revolution; a change of society as well as of government and regime by violence a social revolution; and a contest for state power by violence a civil war (Huntington, 1968; Skocpol, 1979; Walzer, 1998).

But social science usage is misleading if it suggests the social facts or historical events referred to are mutually exclusive. A coup d'état can become a political revolution, and a political revolution can become a social revolution – if the new government decides to make "revolution from above" (Colburn, 1994). Indeed, a social revolution is a change of society by violence the new government itself mainly commits: Red Terror, War Communism, collectivization, and the Great Terror in Soviet Russia were all "willed by the rulers" or their dominant faction (Kolakowski, 1978, 3, p. 157); ditto land reform through revolutionary terror, collectivization, the Great Leap Forward, and the Cultural Revolution in Communist China (Schurmann, 1968). And a civil war that gives both radical revolutionaries and parochial rebels a fighting chance – and is won by whatever side best exploits what Machiavelli called *virtù* and *fortuna* – can start when counterrevolutionaries (Mayer, 1971) subvert or usurp the government "not to make revolution but to avert one" (Pipes, 1995, p. 490).

Theories of revolution that would explain a revolutionary movement or a revolutionary situation need not explain a revolutionary outcome, but theories that would explain a revolutionary situation, a revolutionary outcome, or both have to explain a revolutionary movement as well. And whatever the facts that theories of revolution would explain, those facts entail violence – the willful infliction of physical harm. Burke and the Bolshevik ideologists who dismissed nonviolent revolution as a contradiction in terms conformed to social science usage – no violence, no revolution, by any definition (Johnson, 1982). Theories of revolution converge on the modus operandi – revolutionary violence. The question they would answer is what motivates revolutionary violence and what consequences – anticipated and intended or not – follow from committing it.

Explanation

Theories of revolution in Western social science date back to the Greeks, who viewed revolution as the worst kind of war (Plato *Laws* 629d). Thucydides blamed revolutionary situations with or without revolutionary outcomes on revolutionary movements or conspiracies (as he called them) and blamed those conspiracies on hope of success (*Peloponnesian War* 3.45). Plato agreed and blamed hope of success on weak government torn by faction (*Republic* 465b, 545d), which he would eliminate from his ideal polity by abolishing the family and private property among the rulers – no faction, no revolution (*Laws* 683e). Aristotle added on the widest though not deepest comparative study that since hope along with faction springs eternal, no regime is immune to revolution (*Politics* 5).

The theory that people attempt revolution because they have hope of success, and have hope of success because they think the government is vulnerable, lived on in classic political sociology as an explanation of what to avoid and how to avoid it. Machiavelli, who wrote in effect the first handbook on coup d'état and counterinsurgency, told usurpers how to use force and fraud in order not to be overthrown in turn (*Prince* 17–19). Hobbes, trusting more to organization than to statecraft, designed a regime that would prevent "change of governments by civil war," that is, "attaining sovereignty by rebellion," by repressing "factions or conspiracies" (*Leviathan* 31, 15, 22) – small wonder his first published work was a translation of Thucydides. His take on the English Revolution *Behemoth* asked: Better not to kill 1000 seditious preachers than suffer 100 000 fatalities? *The Federalist* echoed Hobbes on sovereignty but held that checks and balances could avert faction and rebellion as well as tyranny. Burke – who predicted the regicide two years, the Terror three years, and the military dictatorship nine years in advance (O'Brien, 1993) and who in *Reflections on the Revolution in France* excoriated the revolution in its "moderate" phase – observed that the revolutionary legislature usurped a king out for compromise: "Their resistance was made to concession; their revolt was from protection; their blow was aimed at an hand holding out graces, favours, and immunities." Tocqueville summed up the whole line of thought: "The evil you would suffer patiently as inevitable seems unbearable as soon as you conceive the idea of escaping it" (*Old Regime and the Revolution* 3.4).

Not all theorizing on revolution was counterrevolutionary. "Violence is the midwife of every old society pregnant with a new one," wrote Marx (*Capital* 1. 24), though his instant

histories of revolutionary situations autopsied abortions and his class-conflict theory of revolutionary movements and their abortionists proved false – street fighters on both sides of the barricades in June 1848 in Paris, for instance, not only had the same class background but were a cross section of the manual workforce (Traugott, 2002). The theory explained history no better elsewhere – industrial workers neither launched the revolutionary movement nor achieved the revolutionary outcome the theory predicted. In Germany they did not even try, and in Russia they acquiesced as a revolutionary movement of Marxist intellectuals achieved a revolutionary outcome in their name by a military coup d'état – the October Revolution was so low profile that the author of *Ten Days That Shook the World* slept through it – followed up with Red Terror and Civil War, which plus famine and disease killed almost as many people as all belligerents lost soldiers in World War I (Moore, 1978; Pipes, 1991, 1995; Ferguson, 1999). But for the vicarious, posthumous fame Marx got from that much-imitated revolutionary movement whose ideology and terror made him a household word – the body count of Communist regimes may exceed that of both world wars combined (Courtois, 1999) – only intellectual historians might have heard of him today.

Whether counterrevolutionary or revolutionary, theories of revolution were handmaidens to political sociology and historiography before the twentieth century, when – with the revolutions attending both world wars – they hived off as answers to the why-question of revolutionary violence. The chief theorists were not activists but academics because activists propounded only trite theories of revolution consisting mainly of prophecy and propaganda: Trotsky (1977) – a seeming exception much admired by political sociologists (Stinchcombe, 1978; Burawoy, 1989) – stands out through historical detail (tendentiously reported and teleologically framed) plus brazen avowal of force and fraud; his prophecies, moreover, all proved false (Kolakowski, 1978). When activist theorizing rose above agitprop to consider strategic means and ends, it rehearsed the classic wisdom in reverse and advised revolutionaries to exploit faction in order to weaken the government and thereby raise hope of success. Whether more activists became revolutionary dictators (or their adjutants) than academics became theorists of revolution is moot – academics launched a social science growth industry (Gurr, 1980; Rule, 1988; Kimmel, 1990; O'Kane, 2000; Malia, 2006; DeFronzo, 2007; Zimmermann, 2011).

Academic theories of revolution, in the standard pilgrim's progress view, went through three stages of development since the 1920s. First there were natural histories (Edwards, 1970; Brinton, 1965) explaining recurrent sequences of similar events in "great" revolutions by the reasons that experts say motivated the people who made them but not generalizing the reasons into a causal theory. Then there were general theories (Davies, 1971; Feierabend et al., 1972; Smelser, 1962; Johnson, 1982) explaining revolutionary violence by discontent – due to frustration, relative deprivation, strain, or dysfunction – but explaining history ad hoc if at all. Now there are comparative-historical theories explaining select revolutions by "structural conditions" of state and society (Skocpol, 1979, 1994; Goldstone, 1991; Wickham-Crowley, 1992) – admixed with political culture (Goodwin, 2001; Foran, 2005) – cogently enough that older theories look passé.

The standard view is not totally false but ignores obvious continuities of theory and method. Theories of revolution since classical antiquity all cite a short list of sufficient conditions or explanatory causes – situational constraints or reasons that motivate people to commit revolutionary violence – and validate (or rather illustrate) the explanation with factual evidence consisting of case histories or statistics laid out to show by the comparative method (Ragin, 1987) that where the explanatory causes occur, revolution (by some definition) occurs as well and that where the explanatory causes do not occur, revolution (by that definition) does not occur either (Aya, 2006). The same goes for theories of collective violence generally as one lucid analyst (Rule, 1988) makes clear.

Grouped by the explanatory causes they cite, theories of revolution form four families – psychological, functional, structural, and political theories. As a rule, psychological and functional theories explain revolutionary movements – they explain why people attempt to change government, regime, society, or all three by violence; structural theories explain revolutionary outcomes – they explain why people successfully change government, regime, society, or all three by violence; and political theories explain revolutionary situations – they explain why people fight civil wars for state power. Psychological and functional theories that explain revolutionary movements mostly leave revolutionary situations and revolutionary outcomes to historians while structural theories that explain revolutionary outcomes also explain (or at least note) revolutionary situations and revolutionary movements along the way, and political theories that explain revolutionary situations explain (or at least note) revolutionary outcomes and revolutionary movements in passing.

Psychological theories explain revolutionary movements by discontent due to frustration alias relative deprivation – defined as perceived discrepancy between what people want and expect and what they get. These theories all assume the frustration–aggression principle that people who are frustrated will commit aggression against the perceived source of frustration, but they disagree about what social facts arouse frustration – one blames steady improvement followed by a sudden setback that raises expectations but reduces attainments; another blames rapid social change that raises expectations beyond attainments; yet another blames whatever raises expectations, lowers attainments, or both. Few psychological theories explain revolutionary violence by discontent alone, however – most add insurgent strength and government weakness to their short list of sufficient conditions (Davies, 1971; Gurr, 1970; Feierabend et al., 1972).

Functional theories explain revolutionary movements by discontent due to strain, dysfunction, or disequilibrium – defined as discrepancy or dissociation between socially prescribed ends, goals, or values and institutionally available means, procedures, or norms for attaining them (Merton, 1968). When social change creates a gap between ideology and experience that the government is unwilling or unable to close with reform, people turn to revolutionaries for salvation. If the government looks weak or loses control of its armed forces so the revolutionaries think they can

overthrow it, revolutionary violence erupts (Smelser, 1962; Johnson, 1982).

Structural theories explain revolutionary outcomes piecemeal – they explain first why power elites subvert or usurp governments, then why subalterns rebel, then why radical vanguards seize power, and then why new rulers reconstruct state and society. The first two events constitute revolutionary situations; the last two constitute revolutionary outcomes. Each event depends on the one that precedes it – path dependency is the buzzword: no subversion or usurpation, no rebellion; no rebellion, no power transfer; no power transfer, no revolutionary reconstruction. These events comprise collective actions and reactions, which the theories explain as rational efforts to satisfy interests (Coleman, 1990) under structural constraints like government weakness, folk community organization, vanguard politico-military competence, and geopolitical exigency. Ideology is a charter for vested interests – the motivating situational constraints are structural (Skocpol, 1979, 1994; Goldstone, 1991).

Political theories likewise explain revolutionary situations piecemeal – they explain first why insurgents claim sovereignty, then why certain people choose to support them, and then why the government fails to repress the insurgency. Revolutionary situations do not all have revolutionary outcomes but some do, and political theories likewise explain these outcomes piecemeal – they explain first why government supporters defect, then why insurgents gain armed force, then why government forces defect, and then why insurgents oust incumbents. Popular support and armed force are much the same thing – revolution is not an election but a gang war where all sides exchange protection for obedience (Aya, 2005). These events again comprise collective actions and reactions, which political theories explain as rational efforts to satisfy interests (Coleman, 1990) under political and strategic constraints like organization, armament, threat, and opportunity. Ideology again is a charter for interest – the motivating situational constraints are political and strategic. Whether and to what extent new rulers reconstruct state and society is another question, which political theories mostly leave to historians (Tilly, 1978, 1993). Nasty aftereffects of revolution like war and genocide are explained as a rule in political terms (Walt, 1996; Melson, 1992).

Theories that explain revolution piecemeal are cumulative causal models (Stinchcombe, 1978) whose elements are people out to make the best of their situation as they see it over time. Previous choices constrain subsequent choices like moves in a game – *échanges de coups* (Dobry, 2009) – save revolution is a most dangerous game played for keeps where players get killed. Between revolutionary situations and revolutionary outcomes involving revolutionary movements falls the shadow of contingent history comprising strategic interactions.

These family portraits of theories of revolution are not just composite sketches, they are caricatures. But caricatures accentuate real features: psychological and functional theories explain revolutionary violence in general; structural and political theories explain revolutionary violence in history and the consequences of committing it – whether anticipated and intended or not. Matching general theory and historical fact (Merton, 1968) is vexing enough that a standard work on violent strife in "developing societies" (Huntington, 1968) cites psychological theories of revolution in one chapter but sets out a political theory of revolution in another that ignores them completely, using "frustration" to mean thwarted satisfaction of specified vested interests.

To sum up: revolution (the effect to be explained) is a social fact or historical event that (by whatever definition) entails violence used to change government, regime, society, or all three; the conditions or circumstances under which it occurs (the causes) are the situational constraints or reasons that motivate people to commit revolutionary violence; and theory of revolution (like all empirical theory) needs independent evidence besides the fact or event to be explained or else the explanation is circular – it doesn't do to say something is the cause of an effect when the only evidence for the explanation is the effect itself (Popper, 1979).

Validation

Theories of revolution are contentious and have been ever since Aristotle (*Politics* 5) attacked Plato's ideal-type model of regime change, which applies his faction-in-government theory of revolution – based on Thucydides and captive fieldwork (Aya, 2005) – to ideal cases: aristocracy, timocracy, oligarchy, democracy, tyranny (*Republic* 8). Aristotle rejected the model but took the theory for granted. Plato's relentless modern critic Popper (2002) thought the model a brilliant situational logic of power politics (especially tyranny) and the theory true. And so it goes. Disputes over theories of revolution are like strife among factions where (as Gibbon says) the contenders suffer what they would inflict.

Structural theory advocates point out that psychological and functional theories explain neither revolutionary situations that start because power-elite infighting paralyzes governments nor revolutionary outcomes that see new regimes more bureaucratic and tyrannical than the old (Skocpol, 1979, 1994; Goldstone, 1991). Political theory advocates point out that revolutionary outcomes depend on the fortunes of war in revolutionary situations – structural *virtù* proposes, political *fortuna* disposes (Tilly, 1993). Skeptics add that since revolutionary outcomes owe to actions revolutionaries take because they expect these actions to yield great benefits that often do not materialize (Hirschman, 1977) no theory that downplays ideology can explain revolution credibly (Aya, 1990, 2001, 2005). How else to explain strategic decisions like Lenin's based on false expectations of imminent world revolution, the rage for socialism in "developing" countries ruled by Marxist radicals, and the resort to mass terror by revolutionary regimes in proportion to the transformation they attempt and the resistance they meet (Pipes, 1991, 1995; Colburn, 1994; O'Brien, 1993; Mayer, 2000)?

Nor is this all. Revolutionary situations involve *mobilisations multisectorielles* (Dobry, 2009) where "grass roots" join the fray for reasons of their own and may even fight the new regime as in France and Russia (Pipes, 1991, 1995; Mayer, 2000). Mobilizing "grass roots" rank and file in revolutionary movements can be difficult, moreover, as in China and Vietnam (Hofheinz, 1977; Benton, 1992, 1999; Race, 2010; Popkin, 1979). In revolution as in war, the multi-person prisoner's dilemma alias free-rider problem where each is tempted to let everyone

else do the dirty work makes individual rewards for helping out and punishments for holding out imperative (Olson, 1971; Lichbach, 1995).

Structural and political theories mention but do not always explain these facts, which embarrass psychological and functional theories. Another embarrassment is that revolution is the work of militant minorities (Coleman, 1990, p. 489) including "grass roots" whose opportunity owes to government weakness, toleration, or support. Political opportunity *structure* – the trademark idea of the social movements literature (Tarrow, 1998) – is really political opportunity *conjuncture*: it can change overnight. The *sans-culottes* (iconic social movement and "grass roots" militant minority par excellence) are paradigmatic:

> If, for a very short period, the *sans-culottes*, taken together, represented a national force, it was because the Jacobins needed such a force for their own purposes. The true isolation of the *sans-culottes* is demonstrated in the conditions of the year III, when, far from enjoying the support of the governing class, they were actively persecuted by the Thermidorians. The *sans-culottes* were then revealed as completely powerless and also very unpopular.
>
> Cobb, 1970, p. 335

If the volcanic model of revolutionary violence as an eruption of anger or mass temper tantrum (Aya, 1990) were remotely true, this item of strategic analysis by a virtuoso historian would make no sense at all. Revolutionary violence entails armed force whose management decides revolutions – two cheers for Chorley (1973).

The bane of explaining revolution, however, is circularity. A methodological object lesson – the sea is rough today because Neptune is angry as shown by its being rough as ever when Neptune is angry (Popper, 1979, p. 192) – applies to theories of revolution where the fact to be explained is the only evidence for the explanation. The fallacy involves either imputing motivation for revolutionary violence without evidence that people who gang up and kill have this motivation or else claiming that invidious conditions of state and society *cause* revolutionary violence without evidence that these conditions motivate people to gang up and kill – no evidence that those who commit the violence do it in order to remedy these conditions.

Psychological and functional theories are flagrant offenders. Psychological theories explain revolution by discontent due to frustration alias deprivation but cite no evidence of the alleged "root cause" except aggregate data on social conditions presumed to create discontent. Functional theories explain revolution by discontent due to strain, dysfunction, or disequilibrium but cite no evidence of the alleged "root cause" except reports on social conditions presumed to create discontent, which is anyway not enough to motivate revolutionary violence (Aya, 1990).

Structural theories of revolution are scarcely immune. One blames "state breakdown and revolution" on "political stress," the combined result of "state fiscal distress," "elite mobility and competition," and "mass mobilization potential," all due to "population pressure." Rapid population growth causes price inflation, straining government budgets and reducing real wages; intensifies job competition among power elites; fills urban areas with young people – and so stresses political institutions till revolutionary violence erupts (Goldstone, 1991). The facts cited to support the theory – aggregate data on social conditions that allegedly cause revolutionary violence – include no evidence that those who commit the violence do so in order to change these conditions – no evidence that they gang up and kill in order to get jobs or boost wages (Aya, 2001).

Another validation issue is empirical – whether case histories put forward to confirm some theoretical explanation by the comparative method actually do so. Take the "most influential book in the field and for many already a classic" (Malia, 2006, p. 310) – to quote a dissenting historian. It says "successful social revolutions" have "distinctive, long-term, structural causes," namely, "state breakdown" plus subaltern "solidarity and autonomy" – so when elites usurp the government, subalterns rebel, letting radicals seize power and reorganize society (Skocpol, 1979, pp. 5, 295; 1994, p. 5). The second "structural" cause does not turn up in all cases examined but what does turn up – radicals seizing power and reorganizing society – is neither "distinctive," "long-term," nor "structural." Worse, "successful social revolutions" since World War II share no prior "structural conditions" besides weak government torn by faction but see radical vanguards take what they think is a shortcut to power and prosperity (Colburn, 1994). With that, the comparative method empirically refutes the theoretical explanation it supposedly confirms (Aya, 2006).

"Boolean" charts associating hypothesized causes and reported effects – their presence or absence indicated by 1 or 0 (Ragin, 1987; Wickham-Crowley, 1992; Foran, 2005) – leave these validation issues untouched.

Theories of revolution remain contentious, but few controvert Thucydides that revolutionary movements need hope of success or Plato that hope feeds on weak government torn by faction. A landmark treatise on social theory has a chapter on revolution (Coleman, 1990, pp. 466–502) that reinvents their explanation and applies it variously without acknowledgment, or even awareness. The wisdom of the ages on explaining revolution is no less wise for being ancient.

See also: Conflict and Consensus; Insurgency; Revolution; Revolutions, History of; Revolutions, Sociology of; Social Movements and Political Violence; Social Protest; Urban Conflicts.

Bibliography

Aya, R., 1990. Rethinking Revolutions and Collective Violence: Studies on Concept, Theory, and Method. Spinhuis, Amsterdam.

Aya, R., 2001. The third man; or, agency in history; or, rationality in revolution. In: Shaw, D.G. (Ed.), Agency after Postmodernism: History and Theory Theme Issue 40, pp. 143–152.

Aya, R., 2005. Theory of revolution from Thucydides to Tilly. In: Kousis, M., Tilly, C. (Eds.), Economic and Political Contention in Comparative Perspective. Paradigm, Boulder, pp. 89–98.

Aya, R., 2006. Theory, fact, and logic. In: Goodin, R.E., Tilly, C. (Eds.), Oxford Handbook of Contextual Political Analysis. Oxford University Press, Oxford, pp. 114–128.

Benton, G., 1992. Mountain Fires: The Red Army's Three-year War in South China. University of California Press, Berkeley.

Benton, G., 1999. New Fourth Army: Communist Resistance along the Yangtze and the Huai, 1938–1941. Curzon, Richmond, Surrey.
Brinton, C., 1965. The Anatomy of Revolution, third ed. Vintage, New York.
Burawoy, M., 1989. Two methods in search of science: Skocpol versus Trotsky. Theory and Society 18, 759–805.
Chorley, C., 1973/1943. Armies and the Art of Revolution. Beacon, Boston.
Cobb, R.C., 1970. The Police and the People: French Popular Protest, 1789–1820. Oxford University Press, Oxford.
Colburn, F.D., 1994. The Vogue of Revolution in Poor Countries. Princeton University Press, Princeton.
Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge.
Conquest, R., 1990. The Great Terror: A Reassessment. Oxford University Press, New York.
Courtois, S. (Ed.), 1999. The Black Book of Communism: Crimes, Terror, Repression. Harvard University Press, Cambridge.
Davies, J.C. (Ed.), 1971. When Men Revolt and Why: A Reader in Political Violence and Revolution. Free Press, New York.
Davies, N., 1997. Europe: A History. Pimlico, London.
DeFronzo, J., 2007. Revolutions and Revolutionary Movements, third ed. Westview, Boulder.
Dobry, M., 2009. Sociologie des crises politiques: La dynamique des mobilisations multisectorielles, third ed. Presses de Sciences Po, Paris.
Eckstein, H., 1965. On the etiology of internal wars. History and Theory 4, 133–163.
Edwards, L.P., 1970/1927. The Natural History of Revolution. University of Chicago Press, Chicago.
Feierabend, I.K., Feierabend, R.L., Gurr, T.R. (Eds.), 1972. Anger, Violence, and Politics: Theories and Research. Prentice-Hall, Englewood Cliffs.
Ferguson, N., 1999. The Pity of War. Basic, New York.
Foran, J., 2005. Taking Power: On the Origins of Third World Revolutions. Cambridge University Press, Cambridge.
Goldstone, J.A., 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley.
Goodwin, J., 2001. No Other Way Out: States and Revolutionary Movements, 1945–1991. Cambridge University Press, Cambridge.
Gouldner, A.W., 1979. The Future of Intellectuals and the Rise of the New Class. Continuum, New York.
Gurr, T.R., 1970. Why Men Rebel. Princeton University Press, Princeton.
Gurr, T.R. (Ed.), 1980. Handbook of Political Conflict: Theory and Research. Free Press, New York.
Hirschman, A.O., 1977. The Passions and the Interests: Political Arguments for Capitalism before Its Triumph. Princeton University Press, Princeton.
Hofheinz Jr., R., 1977. The Broken Wave: The Chinese Communist Peasant Movement, 1922–1928. Harvard University Press, Cambridge.
Huntington, S.P., 1968. Political Order in Changing Societies. Yale University Press, New Haven.
Jen, Y.W., 1973. The Taiping Revolutionary Movement. Yale University Press, New Haven.
Johnson, C., 1982. Revolutionary Change, second ed. Stanford University Press, Stanford.
Kimmel, M.S., 1990. Revolution: A Sociological Interpretation. Temple University Press, Philadelphia.
Kolakowski, L., 1978. Main Currents of Marxism, 3 vols. Oxford University Press, Oxford.
Lichbach, M.I., 1995. The Rebel's Dilemma. University of Michigan Press, Ann Arbor.
Malia, M., 2006. History's Locomotives: Revolutions and the Making of the Modern World. Yale University Press, New Haven.
Mayer, A.J., 1971. Dynamics of Counterrevolution in Europe, 1870–1956. Harper & Row, New York.
Mayer, A.J., 2000. The Furies: Violence and Terror in the French and Russian Revolutions. Princeton University Press, Princeton.
Melson, R., 1992. Revolution and Genocide: On the Origins of the Armenian Genocide and the Holocaust. University of Chicago Press, Chicago.
Merton, R.K., 1968. Social Theory and Social Structure, third ed. Free Press, New York.
Moore Jr., B., 1978. Injustice: The Social Bases of Obedience and Revolt. Macmillan, London.
O'Brien, C.C., 1993. The Great Melody: A Thematic Biography and Commented Anthology of Edmund Burke. Minerva, London.
O'Kane, R.H.T. (Ed.), 2000. Revolution: Critical Concepts in Political Science, 4 vols. Routledge, London.
Olson, M., 1971. The Logic of Collective Action: Public Goods and the Theory of Groups, second ed. Harvard University Press, Cambridge.
Paige, J.M., 1997. Coffee and Power: Revolution and the Rise of Democracy in Central America. Harvard University Press, Cambridge.
Palmer, R.R., 2005/1941. Twelve Who Ruled: The Year of the Terror in the French Revolution. Princeton University Press, Princeton.
Pipes, R., 1991. The Russian Revolution. Vintage, New York.
Pipes, R., 1995. Russia under the Bolshevik Regime. Vintage, New York.
Popkin, S.L., 1979. The Rational Peasant: The Political Economy of Rural Society in Vietnam. University of California Press, Berkeley.
Popper, K.R., 1979. Objective Knowledge: An Evolutionary Approach, second ed. Oxford University Press, Oxford.
Popper, K.R., 2002. The Open Society and Its Enemies, fifth ed. Routledge, London.
Race, J., 2010. War Comes to Long An, second ed. University of California Press, Berkeley.
Ragin, C.C., 1987. The Comparative Method. University of California Press, Berkeley.
Rule, J.B., 1988. Theories of Civil Violence. University of California Press, Berkeley.
Schurmann, F., 1968. Ideology and Organization in Communist China, second ed. University of California Press, Berkeley.
Sheehan, J.J., 2008. The Monopoly of Violence: Why Europeans Hate Going to War. Faber & Faber, London.
Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge University Press, Cambridge.
Skocpol, T., 1994. Social Revolutions in the Modern World. Cambridge University Press, Cambridge.
Smelser, N.J., 1962. Theory of Collective Behavior. Free Press, New York.
Stinchcombe, A.L., 1978. Theoretical Methods in Social History. Academic, New York.
Syme, R., 1939. The Roman Revolution. Oxford University Press, Oxford.
Tarrow, S., 1998. Power in Movement: Social Movements and Contentious Politics, second ed. Cambridge University Press, Cambridge.
Tilly, C., 1978. From Mobilization to Revolution. Addison-Wesley, Reading.
Tilly, C., 1993. European Revolutions, 1492–1992. Blackwell, Oxford.
Traugott, M., 2002/1985. Armies of the Poor: Determinants of Working-class Participation in the Parisian Insurrection of June 1848. Transaction, New Brunswick.
Trotsky, L.D., 1977/1931–33. The History of the Russian Revolution. Pluto, London.
Turner Jr., H.A., 1996. Hitler's Thirty Days to Power: January 1933. Addison-Wesley, Reading.
Walt, S.M., 1996. Revolution and War. Cornell University Press, Ithaca.
Walzer, M., 1998. Intellectuals, social classes, and revolutions. In: Skocpol, T. (Ed.), Democracy, Revolution, and History. Cornell University Press, Ithaca, pp. 127–142.
Weber, M., 1968. Gesammelte Aufsätze zur Wissenschaftslehre, third ed. Mohr, Tübingen.
Wickham-Crowley, T.P., 1992. Guerrillas and Revolution in Latin America: A Comparative Study of Insurgents and Regimes since 1956. Princeton University Press, Princeton.
Zagorin, P., 1982. Rebels and Rulers, 1500–1700, 2 vols. Cambridge University Press, Cambridge.
Zimmermann, E., 2011/1983. Political Violence, Crises, and Revolutions: Theories and Research. Routledge, London.

Revolutions of 1989–90 in Eastern Central Europe

Timm Beichelt, Europa-Universität Viadrina, Frankfurt/Oder, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. von Beyme, volume 20, pp. 13306–13310, © 2001, Elsevier Ltd.

Abstract

The revolutions in Eastern Europe were different from former revolutions because – with the exception of Romania – never have such radical changes been achieved with so little violence from below and from above. Eastern Europe proved to be a deviant case in the history of revolutions for other reasons too, which are discussed here. The special nature of the 'velvet revolutions' is discussed, touching on the role of elites, the international situation, and the loss of regime legitimacy. Three models of system change that occurred during 1989 are offered, and consolidation of these peaceful revolutions is discussed.

A Deviant Type of Revolution

The revolutions in Eastern Europe were no conventional changes of the system that can be dealt with in the tradition of sequence models. There was no uniform anatomy of the changes, though never in European history have so many regimes collapsed at the same time and for similar reasons. Revolution has traditionally been linked to violent change that made it different from coups d'état. Nevertheless, the depth of their change, made the events between 1989 and 1991 both a political and a social revolution at the same time. In this respect it was comparable only to the Bolshevik revolution of 1917 and its later repercussions in the formation of 'people's democracies' after 1945 in the realms under the influence of the Red Army.

No former system's change was as burdened with the same problems of institution-building as this one. Authoritarian and Fascist dictatorships may have changed political institutions, but they hardly touched the system of stock markets and banks. There was no need in post-authoritarian regimes to create offices to administer unemployment insurance or institutions for social security.

This process of transformation was so unique that the old notions of revolution and reform tend to be inappropriate. Perestroika in the Soviet Union was frequently compared to the politics of Peter the Great, the Meiji Revolution, or the innovations of the Young Turks in the Ottoman Empire. Reform was defined as an eclectic 'un-ideological' process of 'un-dogmatic' actors. Its consequences were contradictory measures and even 'messiness' of reform. Because it was initiated from the top level of Soviet power, Perestroika can also be classified as a reform from above, steered by the ideological thrust of a great leader.

In Gorbachev's self-perception, Perestroika was a revolution. The old Soviet Union under Brezhnev started to accept rather loose talk about revolutions. Revolution lost its connotation of a process initiated by political elites and by political means. It was watered down to processes that the Central Committee and the Politburo of the party were no longer able to control, and thus demographic, environmental, and technological revolutions entered the discourse in Communist countries. Gorbachev tried to use the concept of revolution in a more political way. Perestroika was dubbed a 'second revolution' and Gorbachev hinted at Lenin, who had already envisaged several revivals of revolutionary activities in the tradition of the great French Revolution that failed to implement its goals in one coup and had to be repeated in 1830, 1848, and 1871. Western analysts, on the other hand, never used the term 'revolution' in describing the process of Perestroika.

The reforms in Eastern Europe were different from former revolutions because – with the exception of Romania – never had such radical changes been achieved with so little violence from below and from above. Eastern Europe proved to be a deviant case in the history of revolutions for several reasons:

Political modernization was achieved in two different phases. The system's change was a second attempt at modernization after the Socialist attempts in 1917 and 1945.

The collapse of Socialism resulted from a threefold dissatisfaction in East European populations: dissatisfaction with a repressive political system, with an inefficient economic system, and with the dominance of hegemonic ethnic groups within a federation (Russians, Serbs, and Czechs) or within the Socialist camp (the Soviet Union vis-à-vis its satellites).

There was not the same continuity of periods as in Southern Europe in the 1970s with phases such as liberalization, democratization, and consolidation.

The new process of transition was burdened with the unique double task of transforming both the political and the economic system at the same time, something which has never before happened in world history.

The new transition process was hampered by ethnic strife. Spain had experienced some of it, but managed to solve the problem by various concessions of autonomy to ethnic groups.

In former transitions, a new elite was ready to take over power. There was no real preparation for the new task in Eastern Europe. Even non-Marxist scholars sometimes believed in Marxist 'periodization' schemes. They held that a system needed to develop an economic basis and a recruiting ground for the new political elite. The old regime developed the preconditions of change. Revolution was only the last blow for the old system. The simplicity of this concept of transitions did not even work in the great French Revolution. Historians have shown that the French Revolution did not abolish feudalism or establish a market society throughout the country. This process of accomplishing a true market society was not successful before the beginning of the Third Republic (1875ff.).

Reasons for the Velvet Revolutions

In the history of changes of regimes, the leadership groups have frequently become disunited, and one part of the power elite then started mobilizing the masses. In Eastern Europe, this model did not apply. The elites remained quiet until the last moment. The technical intelligentsia did not develop into an opposition to the 'strategic clique' in power. The technical intelligentsia increasingly de-ideologized and created more room for maneuvering in its own field, independent of political leadership. It avoided conflict, however.

Reformist elites in those Socialist systems where the regime eroded slowly, negotiated to initiate revolution and thus avoided the necessity of appeal to the masses. In Hungary, the negotiated revolution took place with complete discipline and there was no unified counter-elite. In Poland, Solidarity could have been the counterforce, but in the last stage it was pressed by the people rather than by a leadership directing the masses into conflict with the 'nomenclatura' leadership.

Only in Romania – in the third model of a collapse of the system, but under control of a middle-level elite of the old regime – was there conflict among elite sectors. The military, after some wavering, finally sided with the protest movement, not so much because of ideological sympathies with the revolutionaries, but because of its rivalry with the Securitate armed forces. The military played the role of a rear guard – derogated by Lenin as '*chvostizm*' – but apparently the capacity to wait and see and to grasp the right moment for action in the rear guard is not always a negative feature, as Lenin supposed.

The European Fascist dictatorships collapsed in 1945 because they were defeated in war. In Southern Europe and Latin America external factors contributed to the collapse of regimes (Greece, Portugal, and Argentina). In Eastern Europe, changes in the international situation facilitated the peaceful velvet revolutions. The Soviet Union still existed as a hegemonic power, but Gorbachev had indicated that the satellites could no longer rely on Soviet help in a case of internal strife. It was less the active behavior of a superpower, and more the 'non-decisions' of the hegemonic bloc power that did not intervene, which caused the collapse of the whole alliance. Theoretically, each national elite could have followed the Chinese example of bloody suppression. Why did this not happen?

The most important reason is again the international factor: the 'Brezhnev doctrine' was meant to strengthen weak regimes within the bloc. But the bloc center at Moscow was not prepared for a situation where all the leaderships of her satellites entered into troubles at the same time. The Brezhnev doctrine proved its dysfunctional consequences. It had spoiled the satellite leaderships who were always able to rely on Big Brother's tanks. Romania was the only deviant case which did not recognize the doctrine and therefore had to provide for her own security. All of the other governments were not accustomed to organizing big actions for their own security and caved in as soon as the masses confronted them directly.

The old regime in its last defensive stage pretended that the turmoil was caused by subversion from the 'imperialist powers'. This was wrong, since the Western powers were so accustomed to détente that they avoided direct actions for destabilizing the Communist systems. It was not international but rather transnational policies which contributed to the destabilization of Communism from outside. The protest movements in the streets of Leipzig did not receive weapons, but instead got information from the Western media about the situation in the country, the counter-reactions of their leadership, and the points of rally of the protest movement.

The international factor could be the explanation for some parameters of the development, but nobody would have dared to forecast that the Soviet Union would become so occupied with her internal problems that she had no capacity for intervention abroad.

The most complex explanation for the regime breakdown is the loss of legitimacy. In Socialist countries legitimacy could not be measured by surveys. The mass media in the West, such as Radio Free Europe/Radio Liberty, tried to measure legitimacy through the instrument of mail from the East, but results were unreliable. In the early 1980s, it was calculated that the Communists in free elections would not get more than 3 to 5% of the votes, whereas they actually received 10 to 15% in the most developed regimes and much more in the less-developed Communist regimes in the Southeastern parts of Europe.

Was Socialism ever legitimate? Its foundation by force was not forgotten by large parts of the population in the satellite countries. Acceptance of the regime was concentrated in the cadres. With growing success and material benefits, however, legitimacy grew also under Communist conditions.

Paradoxically, Communism collapsed due to its partial success, especially in the sphere of producing intellectually trained elites. There was a mobilization for education, which hardly had an equivalent in the Western world. Communist systems produced a surplus of elites, whom they were unable to funnel into meaningful positions. Western democracies do the same, as periodical overproduction of elites seems to be as inevitable as the famous pig cycle that causes economic slumps in economic and political conflicts. Western systems, however, can offer jobs for new elites in the third 'service' sector when they cannot be integrated into the bureaucracy. This tertiary sector was chronically underdeveloped in Socialist systems. Only the systems that were able to co-opt the dissatisfied young elites, such as Hungary, experienced a slow transition to the new regime. Here the rebellion could be dubbed 'the revolution of the deputy heads of administrative units'.

It would not be 'historical' to look for a single legitimacy for the whole life span of Socialism. Legitimacy also has ideological foundations, and beliefs in legitimation are subject to change. In the early days of revolutionary socialism, the system claimed a kind of teleological legitimation because the party pretended to know the outcome of history. In the Soviet Union, this was combined with charismatic legitimation in the time of Lenin. As time went on, however, both withered away, charisma as well as the capacity to predict the development of history. Quite frequently, the people experienced false prognoses that undermined their belief in the ideology. As Max Weber predicted, the personal charisma of the leader did not survive the founding father. *Veralltäglichung*, the erosion of charisma, was the necessary consequence. Stalin tried to

substitute for his lack of charisma through organization and ideological leadership in all fields – from literature to biology. But this kind of ersatz legitimation was increasingly built on force. With diminishing capacities in the regime for repression and mobilization, new forms of legitimation had to be found. Communist repression had to be substituted by legal security. In the early days, huge groups of defeatist and subversive elements were discovered. After Stalin, nobody believed in this kind of singling out of groups any more. Only in the 'parasitism' and '*chuliganstvo*,' spheres where the Soviet citizen felt endangered for his own security, was he ready to buy from the Communist propaganda.

Communism failed to establish a certain balance between the four basic principles of legitimization in Western systems. The legal state was violated by arbitrary decisions of special hierarchies in the party, in the security forces, and in the planning administration. The nation-state was increasingly invoked by official propaganda in order to reinforce Socialist ideology. However, the hegemony of the Soviet Union prevented the development of authentic, independent, national states, even if more systems tried to embellish the Socialist dogma in national colors. The democratic state was proclaimed by state propaganda, but in reality, for the citizens it was only manipulated participation without alternatives in the political arena. Certain ultra-democratic requisites such as 'recall' and 'imperative mandate' operated only within the aura of the Party, never because of a free decision of the citizens. Thus, the welfare state was used to legitimize the late Socialist systems. In the early days, communist propaganda thought that a social state was unnecessary, because a socialist state is socially minded per se. But the leveling down of all social differences was no longer accepted by the leadership.

During the years of economic growth, Hungarian Goulash Communism or Polski Fiat Communism developed a certain legitimization of the system, but in the 1980s it became clear that the Communist countries were falling further behind the West. Communication with Western countries grew and contributed to a growing capacity in the people to compare East and West. During the 1980s, the system lost its legitimacy due to the economic crisis of Socialism, which was felt in most of these countries.

Transition studies on a quantitative basis confronted us with the finding that democratization is impossible as long as the power resources of a regime are highly concentrated. This assumption was not correct in the case of the Communist countries. All of them had the power to suppress the opposition and had in the 1980s increased their services in both staff and technological equipment. Paradoxically, some of the Socialist countries, such as the German Democratic Republic (GDR) invested most of their technological innovation in the social control of the population instead of investing it in production. Nevertheless, the system collapsed. The enormous security equipment did not help the system survive – as in the cases of the Chinese Wall, the Limes of the Roman Empire, or France's Maginot line in 1940. Unlike these previous cases, the fabulous equipment was not even applied in the last conflict. The index of power resources of the old regimes proved to be of minor importance in the fourth wave of democratization. On the contrary, the better equipped a communist regime was with repressive capacities, the more radical was the change.

Table 1 Pathways to democracy

	Pragmatic piecemeal engineering	*Ideological innovation*
Control from above	Opening of the communist regime (Romania, Bulgaria)	Innovation of socialism (Perestroika, Yugoslavia)
Pressures from below	Erosion of socialism (Poland, Hungary)	Collapse of socialism and takeover by the opposition (GDR, CSSR)

Three models predominated in the system's change in 1989 (see Table 1):

1. (Table 1). System change through the opening of the old regime was achieved by the old cadres of the second rank who kept control of the transition – Romania and Bulgaria (until 1991), Albania (until 1992), Serbia, Croatia, and most of the Commonwealth of Independent States (CIS) (through the first part of the 1990s).
2. (Table 1). Change by innovation was the characteristic of Gorbachev's Perestroika. This type of change proved to be the most complicated. The disintegration of the Soviet Union and Yugoslavia showed elements of a 'bargained revolution'. Bargaining, however, was done not by functional groups but by territorially old and new governments. In no case did the old federation survive and in no case was a new federation feasible after the collapse of the old regime.
3. (Table 1). The erosion of the regime by corporate bargains between the power center and the opposition in a negotiated revolution, took place in Hungary and Poland. This type of transition involved a low rate of participation from among the people, because the change was negotiated by elites without the participation of the citizen. Many voters apparently did not even realize the importance of the change and the voter turnout stagnated between 50 and 60% (Poland, Hungary).
4. (Table 1). The implosion of the Communist regime (Czechoslovakia, GDR). In this model, the control of the 'ancien regime' was least developed. A new elite took power. Continuity of elites and institutions was lower than in the first path to democracy.

Transition theories – in the tradition of theories about the developments of the 1990s – were most interested in the dichotomy between pact and dictate. This dichotomy was confronted with reform and revolution in a matrix (Karl and Schmitter, 1991).

Democratization and Consolidation of the Peaceful Revolutions

'Transitology' as a new branch of knowledge dealing with the peaceful revolutions since the 1970 and 1980s, originally used a sequence model along the lines of liberalization, democratization, and consolidation. Liberalization only took

place in the context of models of innovation as well as erosion of Socialism. The period of democratization is normally identified with institution-building (constitutional power and secondary institutions, such as parties and interest groups). This process was successful in most Central European countries but hardly in the successor states of the Soviet Union. Consolidation of elite attitudes (the 'only game in town' criterion developed by Juan Linz) and the citizens has been incomplete in most post-Soviet countries, upto the present.

Things are different in Central Europe. All countries that became EU members in 2004 (the Czech Republic, Estonia, Latvia, Lithuania, Hungary, Poland, Slovakia, Slovenia) and 2007 (Bulgaria, Romania) are today listed as democracies. However, we have also witnessed substantial backslides. Most often, they had to do with conflicts that date back to the Socialist era or the transition phase. Typically, nationalist governments use non-Democratic means to attack what they see as corruptive networks of elites with a post-Communist background. This has led to polarized politics in Poland, Hungary, and Romania that hollow out the state of law and/or the freedom of the media. In general, however, all the new EU member states have been consolidated by several free elections that reduced the volatility of votes; and they have avoided violence and reduced the weight of extremist parties by alternating governments that included acceptance of even a comeback of the post-Communists, who turned in many countries to Democratic Socialism (Lithuania 1992; Poland 1993; Hungary 1994). In retrospective, the type of revolution has played a limited role for the path of democratic consolidation. More decisive were historic experience with democracy and geographic proximity to Western Europe and the EU. Therefore, the revolutions of 1989/90 have turned into a subject of contemporary history rather than a continued field of interest in social science.

See also: Democratic Transitions; Eastern European Studies: History; Eastern European Studies: Society; Political Parties: Communist; Socialism: Historical Aspects; Socialist Societies: Anthropological Aspects.

Bibliography

Ágh, A., 1998. Emerging Democracies in East Central Europe and the Balkans. Elgar, Cheltenham, UK.

Antohi, S., Tismaneanu, V., 2000. Between Past and Future: The Revolutions of 1989 and Their Aftermath. Central European University Press, New York.

von Beyme, K., 1996. Transition to Democracy in Eastern Europe. Macmillan, Basingstoke, UK.

Dawisha, K., Parrott, B. (Eds.), 1997. The Consolidation of Democracy in East-Central Europe. Cambridge University Press, Cambridge, UK.

Elster, J., Offe, C., Preuss, U.K., 1998. Institutional Design in Post-Communist Societies. Cambridge University Press, Cambridge, UK.

Huntington, S.P., 1991. The Third Wave. Democratization in the Late 20th Century. University of Oklahoma Press, Norman, OK.

Karl, T., Schmitter, P., 1991. Transition to Democracy in Eastern Europe. Macmillan, Basingstoke, UK.

Kenney, P., 2002. A Carnival of Revolution–Central Europe 1989. Princeton University Press, Princeton, NJ.

Linz, J., Stepan, A., 1996. Problems of Democratic Consolidation. Johns Hopkins University Press, Baltimore, MD.

Merkel, W., Puhle, H.-J., 1999. Von der Diktatur zur Demokratie. Westdeutscher Verlag, Opladen, Germany.

Oksenberg, M., Dickson, B.J., 1991. The origins, processes and outcomes of great political reform. In: Rustow, D.A., Erickson, K.P. (Eds.), Comparative Political Dynamics. Harper Collins, New York.

Sarotte, M.E., 2009. 1989: The Struggle to Create Post-Cold War Europe. Princeton University Press, Princeton, NJ.

Sebestyen, V., 2009. Revolution 1989: The Fall of the Soviet Empire. Pantheon Books, New York.

Whitefield, S., 1993. The New Institutional Architecture of Eastern Europe. St. Martin's Press, New York.

Zecchini, S. (Ed.), 1997. Lessons from the Economic Transition. Central and Eastern Europe in the 1990s. Kluwer, Dordrecht, The Netherlands.

Revolutions, History of

Jim Leith, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the terminology, taxonomy, etiology, morphology, and the human cost of revolutions from the Renaissance to the Arab Spring. More specifically, the author explores the historical trajectories of different revolutions including the English Revolution, the American Revolution, the French Revolution, the Russian Revolution, and the Mexican Revolution, as well as the recent revolutionary upheavals in Tunisia, Libya, Egypt, Bahrain, and Syria.

This brief article on the general problems of studying revolutions is divided into six sections. The first section discusses the term 'revolution' as an historical phenomenon that has changed its meaning over time with important consequences. The next section discusses the issue of taxonomy or classification, attempting to exclude upheavals that are not really revolutionary, and to distinguish different types of genuine revolutions. The subsequent section discusses the etiology or origins, examining the attempts of historians, political scientists, and sociologists to find common causes of revolution. The next section discusses the morphology or the possibility of finding a paradigm. This section also discusses the problems of drawing a left-to-right graph that has validity for all the great revolutions of modern times and questions the appropriateness of trying to place the course of major revolutions on a grid showing extreme right, right, center, left, and extreme left. The following section discusses the cost of revolutions in terms of human lives. The final section comprises some concluding remarks.

The Evolution of the Term Revolution

Only a handful of historians and social scientists have paid much attention to the history of the term 'revolution,' despite the fact that this history casts light on important developments since the Renaissance. The term emerged from an obscure Latin word meaning the periodic return of a moving object to its point of origin. Throughout the Middle Ages, it was used in its Latin, French, and English forms to describe the supposed cyclical movement of celestial bodies around the Earth. Then it was popularized by heliocentrics following Copernicus's *On the Revolution of the Heavenly Orbs* in 1543. Galileo and his followers used the term repeatedly. Contemporary dictionaries reflected the domination of the term by scientists. In his famous French–English dictionary published in 1611, Randle Cotgrave defined the word as "a full compassing; rounding, turning back to its first place, or point, th'accomplishing of a circular course." Significantly, he gave no nonscientific meanings.

Gradually the meaning of the term began to expand. Poets and playwrights had already begun to appropriate the word. In *The Compleynt of Mars* (1.03), written *c.*1378, Chaucer speaks of 'hevenish revolucioun.' Later, in the beginning of the seventeenth century, Shakespeare preempted the word to describe his view of death. In *Hamlet*, when the grave-digging clown unearths a skull, the prince of Denmark reflects that it might have belonged to a politician, a courtier, or a lord: "here's a fine revolution, an we had the trick to see't (vol. 1)." Shakespeare, thus, used the word to mean the return to one's original condition, earth to earth, ashes to ashes, dust to dust.

While literary men were using the term 'revolution' to mean a cycle, political commentators were already using it to mean a political upheaval. Several Italian authors used the word during the Renaissance to describe political turns of fortune, the replacement of one prince by another, in the city-states of Italy. Then, in 1674, Alessandro Giraffi published a book entitled *Le Rivolutioni di Napoli* to describe the rebellion of the Neapolitans against the Spanish viceroy. Another Italian and a Frenchman used the term to describe the same event. Moreover, Giraffi's book was translated into English within a few years.

In the United Kingdom, most writers referred to the struggle in the middle of the seventeenth century between parliament and the king as the Great Rebellion, but some authors on both sides referred to it as a revolution, for instance the radical parliamentarian Anthony Ascham and the royalist Matthew Wren. In a letter to the Earl of Dorset in January 1646, the royal historiographer James Howell used the word to mean both a political upheaval and the cycle of the years: "for within these 12 years there have been the strangest revolutions and the horridist things happen'd not only in Europe, but all the World Over, that have befallen mankind, I dare say since Adam fell, in so short a revolution in time" (James Hollowell, 1890; *Epistolae Ho-elianae*, Bk. III, 512). His contemporary John Milton also used the word in these two senses (*Paradise Lost*, II, Lines 597–603; VIII, Lines 15–31).

Although the radical parliamentarians in mid-seventeenth-century England claimed that they were restoring ancient rights, they already had some inkling of a new beginning. After the overthrow of the monarchy they spoke of Year I of Liberty, Year II, Year III, and so on for a few years. Some radicals even had a vision of society completely transformed. "But why may we not have our Heaven here," Gerrard Winstanley asked, "and Heaven thereafter to?" (Sabine, 1941: p. 409). Nevertheless, the idea of reversion persisted: even Winstanley was thinking of a return to a primitive simplicity. The royalists also used the word to mean a return to a previous state. For William Temple the return of the Stuart dynasty in 1660 and the Orange dynasty in the Netherlands in 1672 were revolutionary events in the cyclical sense of the word. Although in the second of his *Two*

Treatises on Government, published in 1690, Locke defended the right to revolution, he used the word only twice, each time to mean a return via the state of nature to a previously existing constitutional position (1964, Cambridge edn. paragraphs 223 and 225, 432–433).

Oddly enough it was not the Great Rebellion of mid-seventeenth-century England which made the word an important term in political vocabulary. Only after the Glorious Revolution in 1688 did lexicographers assign the word a political meaning. The use of the word to describe the overthrow of James II and the signing of the Bill of Rights by William and Mary gave the word a wider meaning. It was not quite a restoration of a previously existing order, rather it consolidated and advanced rights for which parliament had fought for the preceding century. Nevertheless, it was a very conservative event, entrenching the power in the parliament of aristocrats, landed gentry, and rich merchants. Also, the Revolution was not seen as something with a universal application; it was viewed as English, unique, and unrepeatable. The Bill of Rights of 1688 was not for all humanity.

Then in the last quarter of the eighteenth century the term came to mean a profound rupture with the past and the beginning of a new era. On the Great Seal of the United States, we see 1776 associated with the Latin motto *Novus Ordo Seclorum*, 'a new order in the ages.' Then the French Revolution consolidated and intensified the idea of a new beginning in history. Like the Commonwealth-men earlier, they renumbered the years starting in 1789. When the Republic was established in 1792 they began to number the years Year 1 of the Republic and so on. Finally, in October 1793 the Convention introduced a new calendar numbering the years, not from the birth of Christ, but from the inauguration of the Republic on September 22 of the previous year. Even more significant was the fact that the French revolutionaries believed that their ideals would have a universal appeal. The three Declarations of Rights – 1789, 1793, and 1795 – were to apply to all men, not just Frenchmen.

Because of the cluster of revolutions in the late eighteenth and early nineteenth centuries, revolution became recognized as a recurrent phenomenon. This in turn led to the appearance of a new character on the world stage: the professional revolutionary. There was an apostolic succession of such characters – Filippo Buonarroti (1761–1837), Auguste Blanqui (1805–81), Mikhail Bakunin (1814–67), Piotr Kropotkin (1842–1921), Victor Serge (1891–1947), Sun Yat-sen (1866–1925), Mao Zedong (1893–1976), and many others. These life-long revolutionaries anticipated revolution, created organizations to bring it about, and formulated plans about what to do when it arrived. Thus, were born the revolutionary political parties which changed the course of modern revolutions. Once such parties capture power, they are able to suppress opposition for very long periods. Until these parties are finally dislodged from power changes in policy can only take place by decisions within the party leadership.

Taxonomy

Many upheavals are labeled revolutions, but few actually qualify for the term. Most *coups d'état* do not qualify since participation in them is limited, and their result is often confined to a change in government personnel. So-called 'palace revolutions' usually involve struggles with factions or families within the government. Rebellions should probably be excluded, not only because they do not succeed, but also because they usually lack a clear program and are often led by nobles who do not really want to destroy the existing political and social order. Even wars of independence are dubious candidates for the term, contrary to conventional usage. Even the American Revolution scarcely deserves this title. Although the War for Independence saw individual rights entrenched, considerable exchange of property, and a large emigration, the internal institutions of the new states were not radically different from those of the former colonies. Some African struggles in the twentieth century against European imperialist powers had arguably deeper internal repercussions.

Even among upheavals that deserve the title there are many degrees of revolution. There are political revolutions such as the Revolution of 1830 in France, which produce a change in the groups enjoying power, but do not shake society up very much. The great vertical upheavals or social revolutions such as the French Revolutions of 1789, the Russian Revolution of 1917, or the Chinese Communist Revolution of 1949, have led to deep social changes. The latter two, however, are not only political and social revolutions, but socialist ones as well. The Babouvist Conspiracy of Equals in France in 1796 was probably the first case of a group, in which some members supported a primitive communist ideal, which actually tried to seize power. Genuine socialist revolutions, attempting to control and redirect the means of production, could obviously not occur until socialism developed in the nineteenth and twentieth centuries.

Revolution in the deepest sense involves not only a change in the groups exercising power, profound changes in the social structure, and in modern times an attempt to change the control over the means of production, but also an attempt to change the way citizens think, feel, and act. There are glimmerings of such an attempt in the Hussite and Puritan Revolutions, but the French Revolution of 1789–99 was the first to try to implement it in a large way. At the peak of the Revolution in Year II the *gouvernement révolutionnaire* tried to 'regenerate' citizens through an attempted mobilization of all the existing media: art, architecture, festivals, music, newspapers, theater, and children's books and games. Such use of the media has since become a feature of major revolutions on both the left and right, such as Nazi Germany, Bolshevik Russia, and Communist China. Comparative studies of revolutions seldom discuss this added dimension to revolution.

Etiology

There has been great attention to the question of the origins of revolution. If it were possible to establish common causes for revolutions, it would greatly simplify the study of the great revolutions since the seventeenth century, each of which has produced thousands of studies. Once common causes were identified, this knowledge would also be useful both to those anxious to promote revolutions and those intent on preventing them. The hope of applying social science for such purposes

has produced such books as Michael Velli's *Manual for Revolutionary Leaders* published in 1972, and on the other side the collection edited by Harry Eckstein (1963), intent on curbing such internal violence.

Since Aristotle, there has been wide discussion of the etiology of upheavals, rebellions, and revolutions. Most historians, however, do not find the literature comparing revolutions of much use in understanding the revolutions in which they specialize. Many of them point to the problem of assessing the weight to be given to various factors at work, before and during revolutions. Unlike scientists who can run experiments in which all factors except one are held constant, historians cannot rerun events to test the role of certain factors. An historian, for example, cannot demonstrate precisely how the closing decades of the eighteenth century in France would have been different if reformers had been able to solve the financial problems, if those problems had not been exacerbated by intervention in the American War for Independence, if there had been a different King from Louis XVI, if he had not called royal troops into Paris in July 1789, if he had not tried to flee the country in June 1791, if the Girondins had not led the country into war in 1792, or if Robespierre had never been born. Similar questions can be posed about all subsequent revolutions. Specialists try to weigh various factors by getting as deeply as possible into the complexities of each upheaval.

Aristotle argued that rival concepts of justice led to upheaval. Modern theoreticians such as Karl Marx, Max Weber, Pitirim Sorokin, Lyford Edwards, Jaroslav Krejci, Mark Hagopian, Peter Calvert, Chalmers Johnson, and a host of others have focused on increasing tensions in societies leading to revolutions, provided there are precipitants and triggers. These tensions are described in various ways including class conflicts, other contradictions of various kinds, frustration of the expectations of certain groups, dissynchronization between social development and existing values, and other sources of disequilibrium. The fact that there have to be severe tensions in society in order to have an internal war seems a truism. The difficulty with such generalizations is in applying them to very different circumstances. Take, for example, the varying role of war. The American Revolution was made possible in part because the threat of invasion from French territories to the north had been removed by the British victory in the Seven Years War. France was not at war when the revolution broke out in 1789. In contrast, the government in Russia had been severely undermined by war by 1917. Also in China the long struggle with Japan contributed to the collapse of the Guomindang and to the Communist victory in 1949.

Another important variation among revolutions is their very different objectives. As was mentioned earlier, there were no socialist revolutions before the twentieth century. Twentieth-century revolutions in Russia, China, Cuba, and elsewhere have attempted to seize and redirect the means of production. Related to this has been the attempt of countries such as Russia, China, and North Korea to force development in order to catch up with Western Europe and North America. Such a drive was not present in the Hussite, English, or French Revolutions; in fact the hope for a sudden increase in productivity could scarcely have arisen before industrialization, steam power, electrification, and the introduction of internal combustion engine. Comparing revolutions over five or six centuries seems to minimize new ideologies, in fact novelty in human history.

One way of looking at the objectives of various revolutions is to study how they were symbolized. In the French Revolution, the dominant symbols were the figure of Liberty, which gradually evolved into the figure of the Republic, the Liberty Bonnet, and the Liberty Tree, none of which reappeared in the Bolshevik Revolution. Russian revolutionary symbolism featured giant figures of workers holding Hammers and Sickles, workers depicted as part of industrial machinery, electric lights in modernized villages, and of course red stars and flags. In Maoist China there were no symbols of liberty, but thousands of images of Mao like a red sun in the sky, numberless little red books of Mao's writings, and endless images of peasants using muscular power to create terraced hillsides, build dams, canalize rivers, and irrigate the lands. In fact, their terraced hillsides became a symbol of using intensive labor to 'conquer nature.' Machinery was relatively inconspicuous until after Mao's death in 1976. The figure of Liberty was absent until students made her prominent briefly in the abortive democracy movement of 1989. Comparative studies of revolution have so far largely ignored different allegorical figures and symbols.

Morphology

There are similar problems with the attempt to find common patterns in the course of revolutions. One of these is that most historians, sociologists, and political scientists give much more attention to the origins of revolutions than to the inner dynamics that drive them forward, producing the curves so dear to social scientists. A further problem is for how long a period one considers a curve. Jaroslav Krejci extends the Hussite Revolution over 50 years, the English Revolution over 60 years, and the French Revolution over 100 years. Can one speak of a single French Revolution over a century which saw two empires, three forms of monarchy, five revolutions (if one considers 1792 a revolution within the revolution of 1789–99), and numerous coups? Also, if you compare his graphs of the six revolutions that he covers – the Hussite, the English, the French, the Russian Bolshevik, the Turkish, and the Chinese Communist Revolutions they vary considerably from each other (1983: pp. 35, 67, 82, 117, 149). Moreover, his curves vary considerably from those published by Mark Hagopian (1974: p. 245). Also, it is strange that Hagopian shows the Chinese Revolution dipping in 1958, the precise date of the Great Leap Forward. Obviously, reducing great revolutions to scientific-looking graphs is a risky endeavor.

A further problem arises from placing revolutions on a grid marked extreme right, right, left, and extreme left. These terms arose from the accidental seating of conservative deputies to the right of the speaker in the Constituent Assembly in the French Revolution of 1789. In the liberal democratic revolutions in Europe and North America the left has usually meant those who supported genuine parliamentary government, political opposition, and defense of human rights. Can one, however, put the English, the American, and the French Revolutions on the same right–left grid as the Russian, Chinese, and Cuban Revolutions? Placing modern communist revolutions on the left distorts the original meaning of the term. These

twentieth-century revolutions have introduced single-party regimes, have created assemblies that approve decisions already made by the leaders, and have violated human rights. In fact, the way they have operated is very similar to governments on the extreme right, *'les extrêmes qui se touchent,'* as the French put it. This is another example of the perils of applying the same grid to revolutions with very different objectives and consequences.

If the etiology and morphology of revolutions were clear, we should have been able to predict the upheavals in Tunisia, Libya, Egypt, Bahrain, and Syria in 2011–12 and their likely outcomes. Social scientists should be very modest, however, because very few of us predicted these uprisings. At the same time, all the literature on pattern in revolution does not enable us to predict with any certainty the outcome of any of these insurgencies. The metaphor used by western journalists, 'the Arab Spring,' may prove to be completely misleading. In all these countries, the Islamic parties that have been held in check by authoritarian governments, appear ready to move into the vacuum left by the expulsion or liquidation of leaders such as Ben Ali in Tunisia, Muanamar Gaddafi in Libya, and Hosni Mubarak in Egypt.

The future of the unrest in Bahrain is uncertain because already Saudi tanks have rolled over the causeway linking it to Saudi Arabia to support a fellow Sunni ruler, Hamad al-Khalifa, against a Shia majority. Further, the outcome of the bloody uprising against Bashir Al-Assad and his Alawite supporters in Syria remains uncertain. If Assad eventually falls, the Saudi Arabians would prefer the Sunnis to take over, which would not be in the interests of the Shias and Christians who have enjoyed a degree of toleration. John R. Bradley, one journalist who *did* not predict revolution in Egypt, warns in *After the Arab Spring*, Palgrave, 2012, that these uprisings may be the prelude, not to liberal democracy, but to the rule of radical Islamists.

The Cost of Revolution in Human Lives

In the comparative study of revolutions scant attention has been given to their cost in human lives. Whatever benefits revolutions bring, these have to be balanced against their toll in human suffering and death. Calculating this cost is difficult because many revolutions involve civil war and international conflict, plus the question of whether or not to include famines that may have been precipitated by political actions. There is, however, a clear trend: as revolutions have moved from being primarily political upheavals to great social upheavals their toll in human lives has mounted dramatically. The English Civil War 1642–46 between the Puritans in parliament and the High-Church supporters of Charles I, which many historians label a revolution, saw the death of about 190 000 Englishmen, including in the end the king himself in 1649. This upheaval culminated in the Protectorate of Oliver Cromwell who ruled the country until his death in 1658. His invasion of Ireland, where the Catholics were predominantly royalist, resulted in around 400 000 deaths.

The American Revolution, which as we have seen led to an expansion of the meaning of the term revolution, was principally a struggle against English imperial control. Most of its victims died in this war, about 25 000 American servicemen out of around 200 000 who were under arms at one time or the other. No one was executed because of his or her loyalty to George III, although a few died in the aftermath of tarring and feathering, and some were executed as suspected spies. The French Revolution, 1789–99, saw many more deaths. These victims are engraved in our memories because of the images that have come down to us of the condemned being transported in tumbrils to the guillotine. Overall about 40 000 were executed or died in prison, while around 250 000 died in the suppression of counterrevolutionary revolts within France. Then the revolutionary wars against much of Europe in 1792–99 evolved into the Napoleonic wars lasting until 1815 in which about 3 000 000 soldiers and around 1 000 000 civilians died.

The Mexican Revolution, roughly a war between large landowners on one side and small ranchers, unemployed cowboys, and landless peasants on the other, lasted from 1910 till 1920 and cost about 2 000 000 lives. The bloodshed in the Russian Revolution of 1917 resulted in around 1 000 000 military deaths, about 5 000 000 who died in famines, around 2 000 000 victims of disease, plus more deaths by terror or crossfire. By 1928, Joseph Stalin had seized power, retaining it until his death in 1953. His ruthless rule led to about 20 000 000 deaths including liquidation of the Kulaks, purges of possible rivals, prosecution of alleged counterrevolutionaries, famines induced by political policies, and atrocities during the Second World War. These infamous events have been immortalized in literature by authors such as Alexander Solzhenitsyn, *The Gulag Archipelago*, published in the West in 1973.

Modern China, however, breaks the record for revolutionary deaths in modern times. The Taiping Rebellion, which lasted for 14 years, 1850–64, and covered an area the size of western Europe, never was awarded the title of 'revolution' despite costing about 20 000 000 lives, probably because it never succeeded in overthrowing the Qing dynasty. The Communist Revolution in China was arguably the bloodiest upheaval in history. The Communist struggle for power against the Nationalists cost around 7 000 000 lives, 5 000 000 in the first phase, 1926–37, another 2 000 000 in the second phase, 1945–49, not including the 10 000 000 killed in the war with the Japanese. Then the regime led by Mao Zedong from establishment of the People's Republic in 1949 until his death in 1976, was arguably the bloodiest in history. This bloodshed was the result of Mao's effort to purge the privileged members of the old order and then periodically to push the revolution forward – the liquidation of landlords during the Agrarian Reform 1950–51; the Three-Antis campaign 1951–53; the Five-Antis campaign 1952–53 (they overlapped); the Anti-Rightist campaign in 1957, which followed the 100 Flowers Movement encouraging intellectuals to criticize the Party; the Great Leap Forward 1958–62, which led to a disastrous famine; and the Great Proletarian Cultural Revolution 1966–76. Experts differ in estimating the number of victims of these spasmodic purges and lurches forward. The Walker report to the US Congress in 1971 estimated that some 32 to 59.5 million people were killed. Stephane Cortois et al. (The Black Book of Communism, Harvard, 1999) estimated 65 million deaths under Mao (Jung Chang and John Holliday, Mao: The Unknown Story, Knopf, 2005), set the total at 70 million. Only when the Chinese archives are opened will we have firmer

figures, but there is no doubt they will be high, probably at least 40 million.

Conclusion

While it seems very difficult to establish a common taxonomy, etiology, and morphology covering all the major revolutions in the past five centuries, this does not mean that the comparative study of revolutions is not worthwhile. The history of the term 'revolution' demonstrates that it has taken on important new dimensions since it first became politicized in Renaissance Italy. The taxonomy of revolutions can exclude upheavals that are not full revolutions and can distinguish among political, social, and socialist revolutions, or combinations of all three. The etiology of revolutions is particularly problematic because the conditions that occur vary widely; nevertheless the idea that disequilibrium of some kind precedes revolution seems a valuable one. In many cases revolutions occur when society is changing while the regime stands still, but in the case of the Iranian Revolution of 1979 the reverse seems true – disequilibrium resulting from a westernizing Shah moving more quickly than powerful conservative groups, especially religious ones, were willing to accept. On the other hand, the collapse of the satellite Communist regimes in eastern Europe after 1989 seems to be the result of the failure of those governments to satisfy the needs of their citizens and the weakness of systems imposed by external force. As for establishing a common pattern or morphology, the differences in the graphs that scholars have produced warn of the dangers in such attempts. Those historians, however, who steep themselves in just one revolution, are very unlikely to recognize what is unique about the particular upheaval in which they specialize.

Since revolution has been a recurrent phenomenon, and may be so in the future, social scientists will doubtlessly attempt to reach general conclusions about them. In such future attempts more attention needs to be given to the internal dynamics of revolutions once they have broken out. The idea of attempting to create a new citizenry by mobilizing all the available media, which has marked revolutions since 1789, deserves further study. So too does the use of symbols, allegorical figures, and political images that have distinguished revolutions in modern times. Moreover, graphs based on a left-to-right scale need to be modified since the meaning of left and right has shifted drastically over time. Finally, the comparison of revolutions must bring out the unique dimensions of each revolution instead of focusing mainly on similarities.

See also: Communism, History of; Conflict: Anthropological Aspects; Crowds in History; Marx, Karl (1818–83); Revolution, Theories of; Revolution; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, Sociology of; Social Movements and Political Violence; Social Movements, History of: General.

Bibliography

Bradley, J.R., 2012. After the Arab Spring. Palgrave-MacMillan, Hampstead.
Brinton, C., 1960. The Anatomy of Revolution. Vintage Books, New York.
Brunno, O., von Conze, W., Koselleck, R., 1972. Geschichtliche Grundebegriffe. Klette, Stuttgart, Germany.
Close, D., Bridge, C. (Eds.), 1985. Revolution: A History of the Idea. Croom Helm, London.
DeFronzo, J., 1999. Revolutions and Revolutionary Movements, second ed. Westview, Bolder, CO.
Dunn, J., 1989. Modern Revolutions: An Introduction to the Analysis of a Political Phenomenon. Cambridge University Press, Cambridge, UK.
Eckstein, H. (Ed.), 1963. Internal War. Problems and Approaches. Free Press, Glencoe, NY.
Ellul, J., 1971. Autopsy of Revolution. Knopp, New York.
Goldstone, A., 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley, CA.
Hagopian, M.N., 1974. The Phenomenon of Revolution. Dodd Mead, New York.
Johnson, C., 1964. Revolution and the Social System. Hoover Institution Studies 13, Stanford University, Stanford, CA.
Krejci, J., 1987. Great Revolutions Compared. Harvester Press, Brighton, UK.
Leith, J.A., 1974. Media and Revolution. Canadian Broadcasting Corporation, Toronto, Canada.
Lipsky, W.E., 1976. Comparative approaches to the study of revolution: a historiographical essay. Review of Politics 38 (4), 494–509.
Pettee, G., 1938. The Process of Revolution. Harper and Row, New York.
Sabine, H. (Ed.), 1941. Works of Gerrard Winstanley. Cornell University Press, Cornell, NY.
Sorokin, P., 1967. The Sociology of Revolution. Howard Fertig, New York.
Velli, M., 1972. Manual for Revolutionary Leaders. Blade and Red, Detroit, MI.
Zagorin, P., 1973. Theories of revolution in contempory historiography. Political Science Quarterly 88 (1), 23–52.

Revolutions, Sociology of

John Markoff, University of Pittsburgh, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The modern concept of revolution comes out of one particular experience at the end of the eighteenth century, an experience that led some to hope for and others to fear similar occurrences. Reflections on the nature, causes, unfolding, and consequences of revolution were for a long time dominated by the French Revolution. As other cases came to preoccupy social scientists, particularly in the wake of the varied social upheavals of the late twentieth century, scholarly reflection on revolution flourished as it developed in new directions. These included greater attention to culture, more emphasis on contingency and agency, and incorporation of ideas from the study of social movements. In the early twenty-first century, new social movements were again getting scholars to rethink not just their approaches to the study of revolution, but to reconsider what forms revolutions might take.

An Old Word Takes on a New Meaning

'Revolution' (from Latin *revolutio*) initially had the sense of a return to a point of origin. In political discussion, it might be used to characterize a dynastic restoration. By the seventeenth century, it had also taken on the sense of an alteration in fortune and might be employed to describe some significant improvement or deterioration in political affairs. In this latter sense, revolutions were often seen as frequent occurrences. One sees this double set of meanings in later reflections on dramatic seventeenth-century events. Beginning in the 1640s, conflict between England's king and parliament led to civil war, the king's execution, widespread social upheaval during which movements championed radical causes, and new forms of government. For some eighteenth-century writers, the installation of a new monarch in England in 1688 was a great singular revolution in the older sense of a return to an initial state of order after several tumultuous decades; for others, it was but one of the many revolutions – in the sense of another dramatic turn of political fortune – that had marked all of English history (Rey, 1989).

A century after this 'Glorious Revolution' in England, the geographic range and social diversity of French turbulence in 1789 and beyond (rural and urban insurrections throughout the country; military mutinies; elite challenges to royal authority; priests defying bishops; a tidal wave of legislative enactments) provided many revolutions (in the plural) for commentators to reflect upon. But, as these upheavals continued to multiply and as various contenders sought to channel the dramatic events in one or another direction, participants began to speak of these many separate events as components of something singular, the French Revolution, that had a structure and causes, that was an ongoing process, and that rival contenders sought to understand and control. Some participants who were disappointed by specific events in France could even imagine that this particular process was but one instance of possible revolutions and could look forward, as one radical group declared in 1794, to "another, far greater, far more solemn revolution" of which "the French Revolution is but the forerunner" (Furet and Ozouf, 1988: 203). So some sought to make new (and perhaps differently constituted) revolutions and others to avert them; and theorists of either (or neither) persuasion sought to understand them.

Until 1917, the French instance was the central source of empirical evidence of what a revolution might be like. Since the French Revolution inspired some to work for that revolution of the future, either to replicate or to improve upon that model, one of the sources for thinking about the causes and consequences of revolution came from those engaged in trying to bring them about. But others, hoping they could be prevented, also sought to understand their deep causes and their immediate precipitants. Much nineteenth-century reflection on the potential for collective action of the new industrial working classes and the expanded urban populations of Europe and North America – one of the wellsprings of the emerging discipline of Sociology – was precipitated by fears and hopes surrounding revolution.

Much sociological reflection was profoundly stimulated by spectacular failures of radical causes to seize and hold power, despite episodes of impressive popular mobilization such as the widespread European turbulence of 1848, the Paris Commune of 1871, or the Russian upheaval of 1905. But theorists were paying relatively little attention to social conflicts in lesser powers or far from Europe. Even well into the twentieth century, other revolutions, both imagined and experienced, were apt to be understood with concepts derived from the French experience. Military terminations of popular upheaval were held to be 'Bonapartist'; Russian revolutionaries accused each other of being 'Thermidorians'; it has remained very difficult to think of political conflict without invoking notions of 'left' and 'right' (initially, spatial locations in revolutionary France's legislature).

The Russian Revolution of 1917 added some new possibilities to stimulate imaginations. Attention turned to the role of revolutionary parties as organizers of mass discontent or fomenters of it, as refiners of popular grievances, or as carriers of visionary ideologies not necessarily aligned with mass concerns. The formation of a transnational revolutionary organization, the Comintern, brought to the fore a host of questions about the interplay of national and global processes. The revolutionary seizure of power by a party claiming roots in Marxism raised important sociological questions. Marx had

analyzed the material circumstances within which social classes were formed and acted. He argued that the modern industrial working class was to be the prime mover of the future. But Bolshevik success in 1917 could hardly be reduced to the actions of the 'working class.' Nor did it seem possible to account for their triumph as the automatic playing out of environing material circumstances without considering their ideas as revolutionaries and their improvised tactics (Haimson, 1955; Rabinowitch, 1976). A debate about determinism and voluntarism had long preoccupied Marxists, one that has echoes in early twenty-first century preoccupations of scholars of revolution with 'structure' and 'agency.'

Academic scholarship on revolution commonly focused on France and Russia and a small number of other, the so-called 'great,' revolutions and sought to discern their common processes or causes, more rarely their outcomes. In an influential treatment, historian Crane Brinton (1938) sought out the common causes of four revolutions and then developed empirical generalizations about commonalities in their unfolding, arguing, for examples, that 'moderates' are superseded by 'radicals,' but that radicalization eventually generates a 'Thermidorian reaction,' as part of a 'natural history' of revolution. This approach was superseded by sociologists who sought more theoretically grounded accounts of structures, which nurtured revolution. Skocpol (1979) sought the commonalities and differences among three cases and held that a combination of class-based revolt from below, particularly in the countryside, and a state unable to raise needed resources was explosive. State incapacity was explained as the inability to raise the resources demanded by interstate military competition or actual warfare, particularly when state initiatives were stymied by the successful opposition of dominant strata. Different as their specific theses are (Brinton pays little attention to the countryside and a lot to revolutionary ideologies, emphases reversed in Skocpol) they both take up the essential instances of France and Russia, on everyone's short list of revolutions.

But the post-World War II era, with its many anticolonial struggles, civil wars in which at least one party has claimed itself to be revolutionary, and regime transitions of many sorts, has provided even more food for thought about the causes, dynamics, and consequences of revolution and suggested new approaches (Goldstone, 2009). By the early twenty-first century, the French dramas from 1789 onward no longer stood as the central model for theoretical contemplation.

Revolutions in China, Vietnam, or Central America clearly seemed to have as much to do with structures of global domination as they did with the national particularities of social relations. Scholarly attention was soon directed toward a wider canvas. This is not only because those instances from the second half of the twentieth century were far from Europe, but also because some of those instances involved revolutionaries who hoped to challenge the world order, not just the locally dominant strata. They also suggested that there was a history of revolution, not just a history of particular revolutions, as revolutionaries adapted lessons learned from previous episodes to the current moment. Revolutionaries on a now geographically expanded stage claimed to be the legatees of a contested tradition that led from France to Russia, then to China, and then to quite a variety of other places.

But the Iranian Revolution that began with the fall of the Shah in 1979 was hard to situate within this particular historical story since its leaders claimed a different heritage. Theoretically, the Iranian events stimulated reflection on the cultural aspects of revolution, on the ways in which revolutionaries understood themselves, and their role in advancing toward a more desired future state of affairs. The rapid overthrow of the Soviet-style regimes of Eastern Europe and then the Soviet Union itself starting in 1989 raised even more questions, since the oppositional leaderships and those who ascended to power did not claim to be the heirs of an alternate revolutionary tradition and sometimes did claim to be rejecting the idea of revolution altogether. If the later-twentieth century might be called the true 'age of revolution' (Goodwin, 2001: 3), it also ushered in new scholarly emphases.

An Elusive Definition

Despite two centuries of reflection and an extremely vigorous body of comparative scholarship on revolution since the last quarter of the twentieth century, the very definition remained in dispute. First of all, to the extent that notions of revolution were abstracted from that singular French experience rather than from some sense of the commonalties of a large number of experiences, observers could readily differ on what they thought were defining features. Some of the elements that, in varying combinations, entered into various definitions might include:

- Illegal seizure of national political power.
- Extensive property transfers.
- Radical alteration in the form of government.
- Pervasive change in a wide variety of institutions.
- Extensiveness of popular mobilizations.
- Resistance to change by the previously privileged.
- Substantial violence.
- New formulas for the legitimation of authority.
- A new elite's claim to be effecting a radical break in the course of history.
- A new elite's claim to be advancing forward to a new level of human progress.
- Widespread sense of profound change, including at the centers of national power, in local arenas, and reaching into daily life.
- The development of new civic rituals and new ways of justifying rule.

This definitional debate opened up a whole host of fruitful questions about how these phenomena might be connected to one another.

Like many other concepts used by political actors, the term revolution has continued to be a term of legitimation and delegitimation rather than simply analysis. All sorts of things have been labeled (or not labeled) as revolutions in order to honor (or dishonor) them. This has been terribly confusing. A grammatically simple sentence, 'Revolutions entail considerable violence' might be any of at least four very different sorts of statement: part of a *definition* of revolution; a *proposition* about revolution (when defined without reference to violence); an *exhortation* to violent tactics (when uttered among

committed revolutionaries); or a *condemnation* of revolution (when uttered among people who abhor violence).

One reason such debates about defining conditions were ongoing was that they were abstracted from a small number of cases, raising theoretical questions about which subset of conditions would be necessary to qualify another event as a proper revolution. The various armed insurgencies of Southeast Asia, Africa, and Latin America, whether successful, stalemated, or defeated did not reframe the issues. But the serial regime overthrows in the European communist zone were conceptually challenging. They involved deep changes in the formulas of political legitimation, ushered in new forms of political life, reconfigured international relationships, and saw enormous transfers of property. But in 1989, apart from Romania, nonviolent mobilizations and negotiated rather than resisted transitions were the norm. They also lacked the vision of a world-transforming, new social order that would set history on a different course. The East German mobilizations of 1989 aroused great scholarly interest for their combination of rapid expansion to enormous numbers of participants and very little organizational infrastructure (Opp et al., 1995; Hirschman, 1993). Many of the new leaderships took as models of economy and polity the capitalist democracies of the West and mounted no ideological challenge to transnational structures of domination with the result that they were supported rather than opposed by wealthy and powerful states. Some would, therefore, not judge them instances of revolution.

But other scholars have suggested that a new model of revolution was, at hand, largely characterized by mass mobilizations of an overwhelmingly nonviolent character accompanied by negotiated transitional arrangements and urge us to engage in 'redefining revolution' (Garton Ash, 2009: 375) or understand the specific processes of 'nonviolent revolutions' (Nepstad, 2011). Scholars soon noted that possibly relevant instances of this apparently new model were far more numerous than the events of 1989 itself and had happened earlier. There was the Philippine termination of the Marcos regime in 1986 and the even earlier Portuguese 'Revolution of the Carnations' of 1974, for example. Indeed, new scholarship strongly indicates that mass nonviolent protest has been more successful in impelling a political transition than has violent opposition since at least the early twentieth century (Chenoweth and Stephan, 2011).

All of this suggested to some scholars a radical rethinking of the subject matter, away from the conception of revolution as the rare successful, violent struggle to set history on a new course to returning to something like the pre-1789 conception of a radical change in political fortunes that happens with some frequency (Goldstone, 2009). If these are not everyday occurrences, revolutions thus understood, in Tilly's (1995) enumeration, have happened hundreds of times in Europe alone.

Major Research Traditions

Divergent emphases fueled overlapping but distinguishable traditions of research on popular collective mobilization, regime overthrow, political violence, elite dissidence, governmental legitimation and delegitimation, institutional change, and class conflict, and a good deal of reflection on how such processes affected each other. Beginning in the nineteenth century, a well developed Marxist tradition examined ways in which long-term economic transformations created new group identities and interests, while altering political structures. Revolution was seen as a traumatic readjustment of power relations, at bottom, driven by the shifting economic interests of rational actors (Marx and Engels, 1955[1848]; McLellan, 2007). In an important variant, capitalism was seen as generating profoundly destabilizing cyclical economic crises. Economic historian Camille-Ernest Labrousse (1944) argued that cyclical food crises were triggers of major insurrectionary events in the past, including the French Revolution, and other writers, from Marx on, argued that industrial capitalism from the nineteenth century on cyclically generates vast misery. The important theoretical issues for scholars in this tradition were how to connect this mass suffering with the popular insurrection it only sometimes triggered, and where and when it did so to connect that insurrection with the specific programs and tactics of revolutionaries. One important element sustaining this approach to the problem was that it was a theme for continual debate among revolutionaries themselves, especially those who identified with one or another Marxist variant. As scholars come to pay attention to global contexts, some began to consider the geographic aspects of cyclical processes as well and examined how forms of conflict moved in space as well as time (Silver, 2003).

A theoretically more diffuse body of work looked for strains of various sorts produced by industrialization, urbanization, and secularization and tried to explain collective outbursts as responses to such strains. In such interpretations, revolutions were apt to be seen as 'moments of madness,' as frenzied and doomed responses to the forces of modernity, for example, in Durkheimian notions of 'breakdown' or in the 'collective behavior' research tradition (Durkheim 1966[1897]; Smelser (1962)). This approach made revolutions a by-product of 'modernization.'

Tocqueville's thesis linked the growing might of a rationalizing state with a broad array of cultural transformations that undermined the legitimation of aristocratic privilege by transferring to the state activities previously carried out by the lords; indeed, the lords might themselves become oriented to serving the new state bureaucracy. As states did more and lords did less, aristocratic privilege was no longer experienced as a by-product of public service but as theft and oppression. This attack on the problem lay relatively undeveloped in twentieth-century social science until revived by a renewed sociological emphasis on the state from about the 1970s, when sociologists like Theda Skocpol (1979) and Jack Goldstone (1991) saw both the growing reach of states over historical time and moments of seriously diminished state capacities as key sources of revolutionary processes.

The Post-World War II Era

The social upheavals of the post-World War II era brought many new instances of collective action, political violence, and regime overthrow into play. Anticolonial movements, some defeating great powers (as in Vietnam), long-term rural warfare with divergent outcomes (as in Guatemala and Nicaragua), and

Third World state breakdowns (some leading to radically new regimes, as in Iran), all added new cases.

The vastness of human suffering so evident in the twentieth century increasingly suggested attention to opportunity and resources, as well as grievance in the genesis of revolt and the success of revolution. Since revolt was so much more scarce than misery, it was evident that something other than (or at least in addition to) suffering was needed to explain dramatic struggle for change and social scientists became adept at identifying structures and moments that furnished opportunities and resources, including organizational capacities, that might make mobilization more likely. If one peasant community rather than another supported a revolutionary movement, social scientists now sought an understanding of what resources in the daily lives of these (but not other) peasants might account for why these (but not other) peasants revolted. If one revolutionary party in a small, poor country (for example, Nicaragua) succeeded in coming to power, while another revolutionary party in another small, poor country (for example, El Salvador) succeeded only in sustaining a bitter, stalemated war with the government, social scientists began to explore the differential vulnerabilities of different kinds of regimes. One argument in this tradition held that a more broadly based regime was harder to overthrow than one that essentially served a very narrowly constituted group such as a tyrant's friends and family (Goodwin, 2001; Wickham-Crowley, 1992). Distinguishing long-standing differences in resources available to rebels from temporary opportunities, and distinguishing stable structures of elite domination from specific policies, became very important for this research tradition. A 'structural' school of explaining revolution was emerging.

One very important circumstance that commonly emerged in such accounts was diminished state capacities. Skocpol (1979) argued that state incapacity to raise the resources to deal with a dangerous international environment had multiple ramifications that could interact with sources of mobilization from below. Goldstone (1991) extended this by showing that political upheavals commonly featured the mutually exacerbating combination of three processes: a fiscally strained state struggling ineffectively to raise resources, intraelite conflict, and popular revolt. This was a very general model that could guide research on particular instances by examining the pressures on and capacities of a particular state, the sources of solidarity and division among a particular elite, and the abilities to resist heightened demands upon them of particular popular actors.

But accounts of favorable structures soon seemed incomplete as well. The Iranian Revolution of 1979, in which many leading participants spoke a religious language at radical variance with the language of those who had come to power in previous, paradigmatic cases, brought many scholars to reconsider the role of cultural elements in revolutionary processes. This instance of revolution, moreover, caught the attention of social scientists at a moment when the social sciences were engaging in what many called 'the cultural turn,' which soon informed the sociology of revolution as it had other fields of inquiry. A new group of scholars was stressing the ways in which social movements might frame issues in culturally resonant ways or even forge what John Foran (2005) called "political cultures of opposition." Eric Selbin (2010) has argued that revolution is experienced as participating in a story that connects present struggles with understandings of the pasts from which they spring and the futures toward which they strive. One such story connects one revolution to another as imagined futures are glimpsed in France, reemerge in Russia, picked up in China, then Vietnam and Cuba, and elsewhere. This particular story is disrupted by Iran, which does not fit, and by 1989, which terminates a big part of that narrative in a very different place than anyone dreamed of in 1917.

Many instances in the late twentieth century and beyond were occurring in lesser powers on the world stage, in places obviously subject to military, economic, or ideological domination by more powerful states, and whose upheavals often involved attempts (some partially successful) to throw off some of these forms of domination. One would hardly try to understand anticolonial revolutions without paying attention to the policies and institutions of colonial rule, social upheavals in countries exporting tropical crops without paying attention to the world economy, and the resources available to governments and rebels in small, poor, and weak countries without paying attention to the capacities of both sides to elicit support from great powers.

Insights garnered in examination of the newer instances began to suggest different emphases in new examination of older cases. The French or Russian revolutions were looked at as cultural processes; the embeddedness of the 'great' revolutions of the past in transnational connections was noted; conflicts beyond the few previously paradigmatic instances entered the comparativists' field of vision. Scholars began to explore the French Revolution, for example, in the context of the global struggle among Europe-centered empires for world domination. The very variety of participants in the newer cases expanded the intellectual horizons of researchers. As varying combinations of participants in the newer cases became evident, comparative scholars were recognizing how multisided were the conflicts in past revolutions as well. The view of revolutions as the displacement of the domination of one reasonably clearly defined social class by another, was increasingly replaced by the view that revolutions involved actions by multiple social strata, some perhaps defined by class, and some with other sorts of identities, including nationality, neighborhood, ethnicity, gender, region, religion, and age, not to mention identities like 'activist' and 'revolutionary.'

Revolutions, both past and recent, were increasingly likely to be seen as complex affairs in which governing authority failed in many different ways, elites divided in many different ways, and rebels mobilized in many different ways. Methodologically, this meant that social scientists were likely to be looking at configurations: at complex intersections and interactions of processes that mobilized collective actions of varying kinds, produced elite conflict, and weakened governing institutions. Theoretically, this meant that the search for generalizable statements about necessary or sufficient conditions for revolution needed to be rethought.

Emerging Directions in the Sociology of Revolution

The issues engaging comparative scholars in the early twenty-first century meshed with some of the large issues engaging

macrosociology generally, indeed, were suggesting revolutions as a fertile terrain for engaging those issues. Among the issues generating much discussion:

- *how to integrate cultural accounts with structural ones*: Cultural issues were an emerging agenda in various ways. Social historians were making enormous advances in understanding popular culture in varied times and places. A new generation of scholars was examining the overt cultural politics of revolutionary elites. The direction of change in such places as post-Shah Iran and post-communist Eastern Europe seemed altogether impossible to discuss without considering legitimating ideas. Resistance to oppression was seen as having a significant cultural dimension and revolutionary visions were analyzed as stories (Foran, 2005; Selbin, 2010).
- *how to integrate accounts of individual action with macrostructural processes*: As elsewhere in sociological theory, this at first constituted a major challenge and occasioned much debate. Proceeding from the assumption of individual rational actors, how could one account for the forms assumed by collective action, especially in light of the often evident individual risk assumed in participation in revolutions in combination with their uncertain and often undesired outcomes? Lichbach (1995) has shown that there are a variety of plausible answers.
- *adequate attention to the visions of revolutionary elites, the wellsprings of popular mobilizations, and the interaction of the two*: Some structural accounts seemed to make the visions and intentions of revolutionary leaderships inconsequential; some other accounts seemed to make them everything and by implication ignored the motivations, intentions, and wishes of ordinary participants. Some urged us to consider how it was that some revolutionary parties managed to win the adherence of large numbers of followers; others urged us to see popular participation in revolution as rooted in the structures and culture of the popular classes and therefore quite separate from the concerns expressed by revolutionary elites. The empirical difficulty in getting these connections right remained a challenge in studies of particular revolutions, let alone efforts at comparative statement. Poorer, anonymous participants not only often lack the means to make their own wishes widely known (peasants seizing land are less likely to hold press conferences or write memoirs than are revolutionary agriculture ministers), but are, also as usual, often highly motivated to conceal their intentions (especially in highly charged situations like revolutions). It is a major methodological problem that revolutionaries who claim to speak on behalf of, say, Central American peasants, are rarely peasants themselves. Ethnographic fieldwork seems a promising methodology for contemporary cases in which researcher access is possible, but in revolutionary circumstances it is not always possible and it will not help a bit for past instances.
- *attention to the outcomes of revolutions*: The consequences of revolution for economic well-being, state power in the international arena, the rationalization of administration, national identities, citizenship rights, and democratization remain intensely controversial matters about which relatively little systematic investigation has been done by comparativists. In contrast to approaches that place enormous weight on the ideologies revolutionaries take into revolution as blueprints for the future or that see the weight of preexisting structures as deeply constraining, Arthur Stinchcombe (1999) has considered revolution as a process in which new claimants to rulership face challenges that emerge as the old order breaks down and the new one is under construction. Such challenges include the frequently dubious character of their new legitimacy claims; the general climate of defiance and mobilization over which they are trying to preside; the significant role often played by armed force, including with some frequency the hostile attention of dangerous neighboring states; the poor functioning of the state they have conquered; and the widespread anxiety-generating uncertainty about what tomorrow may bring. Such circumstances, Stinchcombe argues, are hardly conducive to the give-and-take of democratic politics and far more likely to generate some form of authoritarian rule, despite the democratic claims some revolutionaries carry with them into the revolution. More recent research suggests that democratic outcomes of political transitions are far more likely when the prime mover of change is massive nonviolent opposition than in cases of violent seizures of power against a background of state collapse or victory in prolonged armed struggle (Chenoweth and Stephan, 2011). There seems a broad agreement among scholars that revolutions commonly result in a strengthened state, a deeply ironic generalization since this has not been a common component of revolutionaries' visions (Goldstone, 2009).

  Regardless of the precise outcome, in light of the great diversity of interests and views of the multiple social groups that recent scholarship sees participating in revolutionary processes, those outcomes are highly likely to be very different from what many hoped for, a very simple explanation for the common occurrence of claims that our revolution has been hijacked. The terms 'success' and 'failure' commonly enter into discussion of revolutions. But if diverse participants have their own criteria for success, and if, moreover, these shift over time, the meanings of these terms might vary among actors and over the months and years. This suggests paying more attention to outcomes, to the consequences of revolutions in all sorts of realms and considered on a range of time scales. Just what the Russian Revolution brought about looked different in 1925, 1935, 1975, and 1995, for example. The outcome of revolution remains a wide open and still understudied field of inquiry.
- *renewed attention to process, at both micro- and macrolevels*: As scholars have become more critical of structural accounts, they have been more inclined to attend to human agency. They have therefore been examining individual participation in revolutionary events not as some automatic consequence of some fixed set of interests nor as the imperative of some preestablished ideology, but as a series of decisions that in a moment of crisis might even be taken day by day on whether to hang back or join in, decisions taken in response to swiftly changing decisions by other actors.

Kurzman (2004) has provided such an analysis of the mobilization of oppositional crowds that propelled the departure of the Shah of Iran. Markoff (1996) showed how during the French Revolution, the interaction of village mobilizations and legislative enactments radicalized both parties over a period of months.

On a more macrolevel, students of many particular revolutions have long noted the centrality of warfare to the dynamics of that revolution. In the revolutionary state, conflict among contenders to power may feed into the strengthening of military capacities and the development of an apparatus of surveillance and control. Revolutionary ideals may support extending the benefits of the new order abroad. Relations with other states may be managed by people inexperienced in such affairs. For their part, the dominant strata in neighboring states, fearful of the revolutionary vision and seeing the turmoil of the revolution, are led to be both motivated to crush it and to wildly underestimate the difficulty of doing so. Long-standing interstate conflict may take on a new dimension when one of the states has a new revolutionary regime. Tensions may be exacerbated if émigré groups have successfully sought refuge in neighboring states and are working to reverse the revolution and if opposition movements in those states seek the support of the revolutionary neighbors (Walt, 1996). The resulting warfare in turn often powerfully shapes the revolution in ways with little connection to the initial dreams of revolutionaries.

Some scholars, therefore, see revolution as a process in which interests are reshaped and ideas emerge in response to unanticipated situations that follow the collapse of the established order. Attention needs to be paid to the contingencies of rapidly developing situations and not just the structures and cultures developed over past decades or even centuries. This means as well that a sociology of revolution needs to recognize 'the power of events' (Sewell, 1996: 262).

- *how to integrate accounts of revolution with other kinds of political conflict*: Scholars of revolution were increasingly making use of tools that resembled those used by other students of politics: attending to the play of threat and opportunity, the variety of resources, the forging of identities in struggle, the irony of unintended consequence, the gaps between elites and ordinary participants (and the interaction between the two). Increasing interchange with scholars of other modes of conflictual action (urged by McAdam et al., 2001) was emerging. Jack Goldstone (2009) noted the convergence of social movement studies and studies of revolution. This convergence may be redefining the subject matter away from an exclusive attention to political transformations accompanied by violence toward a broader subject matter: the efforts of human beings to deliberately redesign their institutions; the great variety of means deployed in such efforts; and the contingent outcomes, both desired and undesired.
- *how to make superior use of comparative methodologies*: With increasing recognition of the complexity and diversity of revolutionary processes, the very logic of comparison was coming under heightened scrutiny. If revolutions are complex amalgams of multiple conflicts, some comparativists suggest that rather than searching for some single general cause, or some single general process, or some set of causes whose separate contributions to revolution may be simply added together, there may be multiple configurations of causal processes that produce revolutions, which themselves unfold in a variety of ways. Some scholars sought to distinguish varieties of revolution about which valid generalization might be made. Others suggested that the way forward, theoretically, was to seek generalizations about much smaller processes, which were components of revolutions, but which combined together in so many distinct ways as to defy generalization on that concatenated level. Some studies have made some use of formal methods for studying varying configurations (Wickham-Crowley, 1992; Foran, 2005).
- *change and continuity*: Revolutions are commonly experienced as moments in which people attempt to deliberately improve the workings of social institutions. They, therefore, provide a collection of instances in which to assess the extent to which those institutions are effectively altered in accord with deliberate design, remain unaltered despite efforts to do so, or are altered in unintended ways, thanks to the complexity of revolutionary processes. Opponents of revolution have commonly asserted both the futility of deliberately trying to reshape institutions and the perverse effects of efforts to do so (Hirschman, 1991).

  By the late twentieth century, some scholars were arguing that the dynamics of transnational structures of domination meant that the projects of revolutionaries who had managed to gain control of particular nation-states were so deeply constrained by those transnational structures that the effective levers for generating deep change were not to be found at the national level. Strong support for this view is provided by the world-system perspective developed by Immanuel Wallerstein and others (Wallerstein, 1990). Some scholarship on revolution was adding to the typical methods of comparative work that is built around resemblances and differences among distinct national cases, by examining as well the interaction of events and processes in different places, especially in the concept of an 'Atlantic Revolution' of the late eighteenth and early nineteenth centuries within which a wide variety of interacting European and American struggles were components. This approach, associated with the mid-twentieth-century work of Palmer (1959, 1964), was increasingly influential in the early twenty-first century (e.g., Adelman, 2006). Students of the social movements of the twenty-first century were also paying attention to processes that transcended the boundaries of individual states, as with the 'Arab Spring' of 2010–11.
- *rethinking the concept of revolution for the twenty-first century*: Along with revisiting the revolutions of the past, and rediscovering the complexity of their conflicts, their embeddedness in transnational processes, their visions, and their cultures, some were asking with Callinicos (2008), "What does revolution mean in the twenty-first century?" Equating revolution with the violent seizure of power in a crumbling state or with a military victory after protracted

armed struggle seemed inadequate for the emerging trends. The frequency with which political regimes are overturned by largely nonviolent mass civil opposition, whether quickly mobilized or long developing, suggested the study of a much more diverse array of instances. The increasing propensity of activists to mobilize transnationally to confront global processes suggested that the achievement of national power might come to be displaced as a defining characteristic, too. The often disappointing character of revolutions past was, by the turn of a new century, leading students of revolution to speculate on "How Might the Revolutions of the Future Have Better Outcomes?" (Foran, 2003: 284).

> Social activists in the early twenty-first century were increasingly thinking in terms of issues that transcended boundaries and seemed increasingly likely to think that the route, if there is a route, to transcending globalized capitalism lay in concerted action across rather than within national frontiers. To the extent that they were envisioning a radically different global order, they were suggesting that if there were to be a revolution for the new century that it would need to be envisioned on a larger geographical scale than the nation-state (Smith and Wiest, 2012). The revival of anarchist traditions, a striking presence in many of the social movements of the early twenty-first century, also suggested ways of thinking about revolutionary transformation other than seizing the state (Bamyeh, 2009). Wright (2010: 273–365) proposes directing theoretical attention more broadly to the variety of transformative processes. These might be 'ruptural' in the sense of a sudden access to power, very likely generating considerable violence and including common notions of revolution inherited from the past. But they also might be 'interstitial,' the development of new institutions and practices that bypass the state, or 'symbiotic,' carried out in partnership with an existing state, the sorts of programs commonly thought of as reforms. So, while many scholars had returned to a notion of revolution as something that, if not an everyday experience, had happened many times in many places, others were trying to discern the form of a singular large but multifaceted process transcending national boundaries that may lie ahead.

Conclusion

More than two centuries of reflection since 1789 have hardly exhausted the stimulation provided by revolution to the thinking of social scientists; if anything, since the 1970s there has been a tremendous intellectual reinvigoration. This is likely to continue, for revolutions seem a particularly privileged terrain for many significant issues in macrosociology, among which I indicate four:

- the intermingling of change and persistence;
- the interplay of long-run and short-run processes;
- the relationship of local, national, and transnational processes; and
- the collision of hopes for change and recalcitrant realities.

Revolution brings to the fore all the many issues that have swirled around the interplay of 'structure' and 'agency.' Whatever else they may be, they are social processes that strikingly juxtapose the capacity of human beings to construct for themselves previously unrealized and unexpected patterns alongside the incapacity of human beings to construct their world as they please.

See also: Capitalism; Chinese Revolutions: Twentieth Century; Cuban Revolution, The; English Revolution, The; French Revolution, The; Political Sociology; Power in Society; Revolution, Theories of; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, History of; Russian Revolution, The; Social Movements: A Social Psychological Perspective; Vietnamese Revolution, The.

Bibliography

Adelman, J., 2006. Sovereignty and Revolution in the Iberian Atlantic. Princeton University Press, Princeton, NJ.
Bamyeh, M., 2009. Anarchy as Order: The History and Future of Civic Humanity. Rowman and Littlefield, Lanham, MD.
Brinton, C., 1938. The Anatomy of Revolution. Vintage Books, New York.
Chenoweth, E., Stephan, M., 2011. Why Civil Resistance Works. The Strategic Logic of Nonviolent Conflict. Columbia University Press, New York.
Callinicos, A., 2008. What does revolution mean in the twenty-first century? In: Foran, J., Lane, D., Zivkovic, Z. (Eds.), Revolution in the Making of the Modern World, pp. 151–164.
Durkheim, E., 1966[1897]. Suicide, a Study in Sociology. Free Press, New York.
Foran, J. (Ed.), 2003. The Future of Revolutions: Rethinking Radical Social Change in the Age of Globalization. Zed Books, London and New York.
Foran, J., 2005. Taking Power: On the Origins of Third World Revolutions. Cambridge University Press, Cambridge, NY.
Furet, F., Ozouf, M., 1988. Dictionnaire critique de la Révolution française. Flammarion, Paris.
Garton Ash, T., 2009. A century of civil resistance. In: Garton Ash, T., Roberts, A. (Eds.), Civil Resistance and Power Politics. Oxford University Press, Oxford, UK.
Goldstone, J.A., 1991. Revolution and Rebellion in the Early Modern World. University of California Press, Berkeley, CA.
Goldstone, J.A., 2009. Revolution. In: Landman, T., Robinson, N. (Eds.), Sage Handbook of Comparative Politics. Sage, Beverly Hills, CA, pp. 319–347.
Goodwin, J., 2001. No Other Way Out: States and Revolutionary Movements, 1945–1991. Cambridge University Press, Cambridge, NY.
Haimson, L., 1955. The Russian Marxists and the Origins of Bolshevism. Harvard University Press, Cambridge, MA.
Hirschman, A.O., 1991. Rhetoric of Reaction. Perversity, Futility, Jeopardy. Belknap Press, Cambridge, MA.
Hirschman, A.O., 1993. Exit, voice, and the fate of the German Democratic Republic: an essay in conceptual history. World Politics 45 (2), 173–202.
Kurzman, C., 2004. The Unthinkable Revolution in Iran. Harvard University Press, Cambridge, MA.
Labrousse, C.-E., 1944. La crise de l'économie française à la fin de l'ancien régime et au début de la Révolution. Presses Universitaires de France, Paris.
Lichbach, M.I., 1995. The Rebel's Dilemma. University of Michigan Press, Ann Arbor.
Markoff, J., 1996. The Abolition of Feudalism. Peasants, Lords and Legislators in the French Revolution. The Pennsylvania State University Press, University Park, PA.
Marx, K., Engels, F., 1955[1848]. Manifesto of the communist party. In: Marx, K., Engels, F. (Eds.), Selected works, vol.1. Foreign Languages Publishing House, Moscow, pp. 21–65.
McAdam, D., Tarrow, S., Tilly, C., 2001. Dynamics of Contention. Cambridge University Press, Cambridge, NY.
McLellan, D., 2007. Marxism after Marx. Palgrave Macmillan, New York.
Nepstad, S.E., 2011. Nonviolent Revolutions. Civil Resistance in the Late 20th Century. Oxford University Press, New York.
Opp, K.-D., Voss, P., Gern, C., 1995. Origins of a Spontaneous Revolution: East Germany, 1989. University of Michigan Press, Ann Arbor, MI.

Palmer, R.R., 1959, 1964. The Age of the Democratic Revolution. A Political History of Europe and America, 1760–1800, 2 vols. Princeton University Press, Princeton, NJ.
Rabinowitch, A., 1976. The Bolsheviks Come to Power: The Revolution of 1917 in Petrograd. Norton, New York.
Rey, A., 1989. "Révolution", Histoire D'un Mot. Gallimard, Paris.
Selbin, E., 2010. Revolution, Rebellion, Resistance. The Power of Story. Zed Books, London and New York.
Sewell Jr., W.H., 1996. Three temporalities: toward an eventful sociology. In: McDonald, T. (Ed.), The Historic Turn in the Human Sciences. University of Michigan Press, Ann Arbor, MI, pp. 245–280.
Silver, B., 2003. Forces of Labor. Workers' Movements and Globalization since 1870. Cambridge University Press, Cambridge.
Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, Cambridge.
Smelser, N.J., 1962. Theory of Collective Behavior. Free Press, New York.
Smith, J., Wiest, D., 2012. Social Movements in the World-System: The Politics of Crisis and Transformation. Russell Sage Foundation, New York.
Stinchcombe, A.L., 1999. Ending revolutions and building new governments. Annual Review of Political Science 2 (1), 49–73.
Tilly, C., 1995. European Revolutions, 1492–1992. Blackwell, Oxford.
Tocqueville, A de, 1952–53. L'Ancien Régime et la révolution, 2 vols. Gallimard, Paris.
Wallerstein, I., 1990. Antisystemic movements: history and dilemmas. In: Amin, S., Arrighi, G., Frank, A.G., Wallerstein, I. (Eds.), Transforming the Revolution: Social Movements and the World-System. Monthly Review Press, New York, pp. 13–53.
Walt, S.M., 1996. Revolution and War. Cornell University Press, Ithaca, NY.
Wickham-Crowley, T.P., 1992. Guerrillas and Revolution in Latin America. A Comparative Study of Insurgents and Regimes since 1956. Princeton University Press, Princeton, NJ.
Wright, E.O., 2010. Envisioning Real Utopias. Verso, London and New York.

Rhetorical Analysis

Peter Larsen, University of Bergen, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rhetoric is the old *art of oratory*. Originally a theory of speaking in public with special focus on argumentation and persuasion, it gradually became associated with poetics and stylistics of written texts. During the last century several attempts have been made to revive rhetoric, making it a part of a general communication theory. The first part of the article traces the history of rhetoric from classical antiquity to the late nineteenth century. The second part of the article presents four major trends in modern rhetoric. (1) *Rhetoric and semiotics*; (2) *Visual Rhetoric*; (3) *Metaphor, metonymy, and mental mapping*; and (4) *Rhetoric, argumentation, and thinking*.

Rhetorical Analysis

Rhetoric, the *art of oratory*, was developed in ancient Greece and Rome. Both as a general communication theory and as a collection of practical advices on how to convince an audience, rhetoric exerted decisive influence on the history of Western civilization and culture. In the period from antiquity until the nineteenth century, rhetoric was the single most important subject taught in Western educational institutions. During these more than 2000 years it changed both in scope and perspective. Corresponding with the growth in written communication, it was gradually transformed into an *art of writing*. In later centuries it became closely associated with the writing of literature, while less emphasis was put upon argumentation and persuasion.

Rhetoric deals with communicative situations. Speeches and texts are always considered in their relation to the aims and intentions of the speaker/writer, the subject matter in question, the general mood and prior knowledge of the expected audience, etc. Correspondingly, any kind of analysis that sets out to describe the ways in which a given discourse is determined or affected by its immediate communicative context could be said to be rhetorical in nature. Rhetorical analyses of this kind have always played an important role in literary studies, but during the last few decades similar methods have also been applied to nonliterary texts and even to nonverbal means of expressions, for instance, in connection with studies of political propaganda, advertising strategies, as well as mass media in general.

The History of Traditional Rhetoric

Classical Antiquity

Some ground rules of oratory were developed and taught by Sicilians in 485 BC in connection with legal disputes about real estate. This rudimentary rhetoric quickly spread to Athens where a more comprehensive body of knowledge about the art of public speaking was developed from the mid-fifth century and onward. Greek rhetoric was introduced in Rome in the first-century BC as part of the education of older children.

In the Greek period, the so-called Sophists – persons who taught rhetoric, politics, and mathematics for money – played an important role in the development of rhetoric as a formal discipline. They were attacked by Plato for putting more emphasis on style than on substance. Plato's critique raises a question, which has followed rhetoric ever since: Is rhetoric a method for discovering the truth, or merely an effective manipulation technique? Is rhetoric being used in order to convince or to persuade the audience?

There is, however, no clear answer to these questions. The old Greco-Roman rhetoric – as described by Aristotle (384–322 BC; *see* Rhetoric), Cicero (106–143 BC; De Oratore), and Quintilian (*c*. AD 35–95; Institutio Oratoria) – was a complex social system of interrelated, and in some cases even contradictory, practices (cf Barthes, 1970; Vickers, 1988). First of all, rhetoric was an art in the classical, Aristotelian sense of the word, i.e., *a craft or technique*, a collection of rules and advices for speakers who want to convince or persuade an audience. These rules were taught in and transmitted through strong educational institutions. Thus, rhetoric was also a discipline, a *subject*, a body of knowledge. In addition, this body of knowledge constituted a *theory*, or at least a semitheoretical field; it described and classified some important effects of verbal language.

The Rhetorical System

The old manuals divided rhetoric into five parts: *Inventio*, *Dispositio*, *Elocutio*, *Memoria*, and *Actio*. Each part covers a specific field of knowledge and aims at training certain skills related to the field in question.

- *inventio* (finding what to say) is the part dealing with *argumentation*. Speakers learn to analyze a given subject matter, to find relevant information, and to determine which types of information are the most useful for their purposes. They also learn to convince an audience with quasilogical proofs (*logos*), and to move the audience by appearing to be good and trustworthy (*ethos*) or by stirring the audience's emotions (*pathos*).
- *dispositio* (organizing what one has found) deals with *composition*. Speakers learn to organize their speeches in clearly separated sections.
- *elucutio* (choosing the proper words) is the part that deals with verbal *ornamentation*. Some elementary rules are presented together with examples of striking images and other verbal devices, which can be used to color a text.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95031-1

- *memoria* (remembering what to say) is an introduction to *mnemonics*.
- *actio* (performing the speech) contains practical advices on how to *deliver* a text before an audience.

The Reduction of Rhetoric

The memoria and actio parts were rather sketchy in the old manuals, and because they were directly related to the practice of oratory, they disappeared as soon as rhetoric became a theory of written discourse. During the Middle Ages the tripartite system of inventio/dispositio/elocutio was taught together with grammar and logic. This system collapsed during the seventeenth century, and rhetoric ended up being a part of poetics. From now on it was further reduced: inventio and dispositio, the parts most closely associated with logic and argumentation, lost in importance and in the end all that remained of the old system was elocutio, the part dealing with questions of stylistics.

The *elocutio* sections of the old manuals taught the reader that significant persuasive or poetic effects can be obtained by selecting the proper words and combining them in the right manner. A flat statement of facts can be transformed into a strong, colorful, and convincing message by the use of *tropes* or *figures*, i.e., by substituting single words (*tropes*) or by manipulating phrases consisting of several words (*figures*). However, when eighteenth-century linguists and literary theorists began to study and classify the many varieties of discursive ornamentation, they found that most of these poetic effects could be described as variations of two semantic relations, viz. *similarity* and *contiguity*. The similarity relation finds its clearest expression in the *metaphor*, while *metonymy* and *synecdoche* are well-known examples of ornaments based on causal and spatial contiguity relations. From the eighteenth century and onward, metaphor gradually became a kind of supertrope representing all the other tropes and figures based on similarity relations, while metonymy correspondingly became the representative of all the various ornaments based on contiguity relations (Genette, 1970).

Modern Rhetoric

Rhetoric and Semiotics

During the twentieth century several writers sought to incorporate this reduced version of traditional rhetoric into a broader, more general framework. One of them was the Russian linguist Roman Jakobson who in 1956 presented a structuralist interpretation of *elocutio*, arguing that the metaphor–metonymy dichotomy corresponds with a general bipolar structure in language (Jakobson, 1956). All discursive production is the result of two fundamental linguistic processes based on similarity and contiguity. These metaphoric and metonymic poles in language can be observed in all normal linguistic behavior but, Jakobson argued, they are also manifest in any symbolic process, either intrapersonal or social (for a fuller discussion of the metaphor–metonymy dichotomy, see the following section).

Jakobson thus opened the field of rhetoric toward sign systems other than language. His own examples included the dream work studied by psychoanalysts, the magic rites described by anthropologists, as well as nonverbal forms of art like cubist painting, the films of Chaplin and Eisenstein, etc.

Jakobson exerted decisive influence on French structuralists and semioticians during the 1950s and 1960s, among them Roland Barthes who in a seminal article, *Rhetoric of the Image*, proposed to rethink classical rhetoric in structural terms (Barthes, 1964; *see* Semiotics). Like Jakobson, he argued that the immense variety of tropes and figures categorized by traditional rhetoric are merely surface manifestations of a few formal relations that can be shown to be the structural armature in all kinds of verbal discourse as well as in several nonverbal forms of expression. He predicted that it would be possible to establish a general rhetoric, which would be valid for all types of discourses, for texts as well as for sound and images.

Since the 1960s there have been several attempts at implementing Barthes' project.

The most comprehensive contribution is the work of the so-called *Groupe* μ, a group of Belgian researchers, who presented a general description of rhetorical tropes and figures on a semiotic/structuralist basis (cf Groupe μ, 1970).

The semiotic-structuralist interpretation of traditional rhetoric was mainly a theoretical project, but it had important practical and analytical consequences as well. In his 1964 article, Barthes argued that there is a close connection between rhetoric and ideology, a view supported by the fact that the classical manuals usually were more concerned with probabilities than with the actual truth of the matter. In order to be convincing one must know one's audience, know what people will or will not accept as probable, in short one has to respect *doxa*, i.e., public opinion, the stereotypical forms of thought that members of the audience rely on in their everyday lives. This means that the rhetorical devices used in a text can be understood as discursive expressions – or exploitations – of dominant ideologies within a given society and that such ideologies can be brought to light by means of rhetorical analysis.

That rhetorical analysis can be used as a decisive critical tool had already been proven by the American critic, Kenneth Burke who in 1939 wrote an influential article on the argumentative strategy in Hitler's *Mein Kampf* (Burke, 1941; see also Burke, 1962). From the 1960s and onward, Barthes and many other writers followed Burke's example and made rhetorical analysis a part of a general critique of ideology. Using the insights gained by the fusion of semiotics and rhetoric they made systematic studies of the patterns and systems of belief that are presented as 'natural' or 'evident' in newspapers, fashion magazines, popular films, television programs, etc. This form of analysis was particularly efficient in studies of advertising (for an overview see Dyer, 1982).

Visual Rhetoric

The question of visual rhetoric is central to many works in the semiotic/structuralist tradition. One early example is Jacques Durand's study of advertising in newspapers and magazines (Durand, 1970[[1983]], 1987). Durand reorganized the classic list of linguistic tropes and figures by defining them as transformations of 'simple' propositions into 'figurative' ones. He argued that these building blocks of traditional rhetoric could

be classified by taking two parameters into account: (1) the particular type of transformation – addition, suppression, substitution, and exchange – and (2) the relations – of identity, similarity, difference, and opposition – between the original and the transformed elements. Using this classificatory grid while analyzing a comprehensive corpus of advertisement he showed that it was possible to find modern, visual parallels to all the linguistic tropes and figures described in the classic rhetorical manuals.

Several other scholars have followed Durand's example and proposed similar or more differentiated classifications and typologies (see for example McQuarrie and Mick, 1999, 2003; Philips and McQuarrie, 2004). In a similar perspective, the Belgian Groupe μ has presented a general description of all conceivable types of visual signs aiming at establishing an exhaustive 'rhetoric of the image' (Groupe μ, 1992).

Durand's classifications of visual tropes and figures as well as some of the later typologies have been criticized for using linguistic rhetoric as the basic model, for not taking the special nature of images into account, and for not focusing on the special argumentative force of images. In his work on visual persuasion the US media scholar, Paul Messaris has, in contrast, emphasized the differences between visual and other forms of communication (Messaris, 1997). Describing the role of images in persuasive communication, Messaris points to three fundamental characteristics:

- *Iconicity*: Images can simulate visual reality and evoke attitudes and emotions associated with real people, objects, places, and situations.
- *Indexicality*: Due to the nature of analogue recording, photographic images can function as proof of and support for linguistic statements.
- *Syntactic indeterminacy*: In contrast to linguistic statements, images cannot propose explicit arguments. Since visual arguments are indirect and blurry, it is difficult to argue against them.

These features give images a particular rhetorical strength: While linguistic statements can be checked and tested, images can suggest and hint at propositions that cannot be formulated explicitly. The special relations between text and images in advertising have been analyzed by Leiss et al. in a study of two Canadian magazines from the 1890s to the present (Leiss et al., 2005). Using a combination of semiotic textual analysis and quantitative content analysis the authors study the changing rhetorical strategies in advertisements during the period. They show that not only has there been a major shift from text to images during this period, but also the rhetorical function of the images has changed considerably. The rhetoric of the early phase was characterized by a primarily textual, quasilogical praising of the products and their qualities with images being used as mere illustrations. During the 1920s, as potential buyers became more sophisticated readers of advertisements, this strategy gave way to an indirect, but still primarily textual, argumentation based on suggestions and associations that situate the products in networks of abstract, symbolic qualities. The images became the dominating element during the 1950s and were used to imply that products can change the way in which buyers are perceived by other people. From the 1960s and onward the chief function of images have been to associate the products in question with attractive social groups and lifestyles.

Metaphor, Metonymy, and Mental Mapping

Elocutio was the only part that was left of the old rhetorical system by the beginning of the twentieth century. There was, however, a growing interest in this surviving part and particularly in metaphor. Not only did metaphor become one of the most studied subjects in literary theory, but also the interest spread to several other areas. At present the study of metaphor is one of the most fertile fields in the humanities and the social sciences. It plays a crucial role in academic disciplines as diverse as linguistics, film and media studies, constructivist psychology, anthropology, and cognition theory.

During the last century the conception of metaphor has changed dramatically. From being a poetic device producing a specialized type of discursive effect, it is now generally regarded as a very fundamental mental tool: "the essence of metaphor is understanding and experiencing one kind of thing in terms of another," as George Lakoff and Mark Johnson put it in their book *Metaphors We Live By* (1980). Metaphor is the kind of understanding, which is produced when a set of relations is displaced from one kind of thing to another. Or to put it another way: Metaphor is a type of *mental mapping*. When people say that 'time is money,' they establish a simple relation of similarity between two elements but this process entails a further mapping of a whole series of relations from the time-field onto the money-field: if time is money, then time must also be valuable, scarce, etc.

According to Lakoff and Johnson there are three major types of metaphor. One is the *structural metaphor*, the 'time is money' variety just mentioned: a restricted, almost local, structuration of one field in terms of another. Another one is the *orientational metaphor* which is used to give larger systems of concepts a spatial orientation. When for example happiness is described as 'being up' and sadness as 'being down,' a spatial relation (up/down) is mapped onto an emotional relation (happy/sad). Such orientational metaphors are clearly grounded in very concrete, physical experiences of the body and its position in space. According to Lakoff and Johnson all metaphors do actually have some kind of concrete experiential basis. One example is the third major type, the *ontological metaphor*, which is based on the everyday experience of handling physical objects. Ontological metaphors make it possible for an individual to understand – and thereby to deal with – events, activities, emotions, ideas, etc. as if they were objects, entities, or substances.

Metonymy, the other supertrope, does also have important cognitive functions. Its basic principle is that one entity *stands for* another. In the phrase 'sails on the water,' a part (sail) is used as a shorthand reference to a totality (boat). Thus, a metonymy emphasizes or highlights certain aspects or features within a given totality. In the very act of referring, metonymy thereby serves some of the same purposes as metaphor does.

Metaphor and metonymy are mental tools which enable people to focus their everyday experiences and to understand the world they live in. They are, as Lakoff and Johnson point out, part of the ordinary way in which people think, act, and

talk. Sometimes we use these tools in new and original ways, but most of the resemblances we perceive and most of the shorthand references we make in our everyday lives are the result of relatively conventional metaphors and metonymies that have become part of our conceptual system. In later years, there have been a growing number of studies of such conventional metaphor/metonymy-based systems and the role they play in public debates and political communication. One striking example is George Lakoff's own analysis of the political rhetoric used by US politicians in the weeks before the war in the Gulf in 1991 (Lakoff, 1991). In a later work, Lakoff argues that the metaphors used in political discourse function as 'frames' that actively shape and influence the ways in which an electorate thinks about the issues in question (Lakoff, 2004).

Rhetoric, Argumentation, and Thinking

Traditional rhetoric originated in relatively small, closed societies characterized by limited flows of information. Politicians and lawyers who should speak in public were faced with the task of convincing or persuading uneducated and incompetent audiences. The manuals told them how to do that.

Communication today is still a question of finding and interpreting relevant information and presenting it in effective ways. Nevertheless, the general communicative situation has changed radically since the days of classical antiquity. The most important communication is mediated by radio, films, television, the Internet, etc. – media characterized by what has been called a new type of orality (Ong, 1982). Information is distributed at an incredible speed, and the total amount of available information is growing rapidly. Even though the general level of education is considerably higher than any time before in the history of the world, a modern mass audience is a rather heterogeneous phenomenon consisting of people with very varying degrees of knowledge and competence.

The fear that this complex situation with its rapid flow of information may result in massive manipulation of the audiences provides a significant background for many of the newer attempts at reviving rhetoric. One example is the so-called *New Rhetoric* which was developed in the mid-twentieth century (Perelman and Olbrects-Tyteca, 1958; Toulmin, 1958). The New Rhetoric was initially an attempt at reestablishing the total rhetorical system in order to be able to cope with all aspects of the new, media-driven types of communication. However, from the very beginning the strongest emphasis was put upon *inventio*, and more specifically on *argumentation*, one of the parts of traditional rhetoric that more or less had disappeared when rhetoric became associated with literature and poetics.

In the philosophical tradition argumentation has usually been associated with logic and the techniques of deduction and induction. What the New Rhetoricians have in mind, however, is not this strict, formal, and logical type of argumentation, but the 'fuzzy' thinking, the jumping-to-conclusions known from everyday life. Aristotle excluded this kind of thinking from logic proper but included it in his Rhetoric and actually considered it to be one of its most important parts.

Rhetorical argumentation is a 'situated' form of argumentation; it cannot be performed or understood without taking the total communicative situation into consideration. It is dependent on the subject matter, the level of available information, the preexisting attitudes and prejudices of the speaker and the audience, the type of media which is used, etc. In short, what is at stake here and what the New Rhetoricians as well as many other, later writers (cf for example Billig, 1987) set out to describe and analyze, are the various ways in which practical, everyday reasoning is determined by its immediate context.

Viewed like this rhetoric is much more than a theory of oratory or a collection of simple rules to be used by poets – it is a discipline which studies the underlying principles governing all human discourse, the way in which people argue with themselves and with others, the way in which they make value judgments, etc. Rhetorical analyses of such arguments and of how judgments are made can provide a better understanding of how opinions and social attitudes are formed in complex modern media-driven societies.

See also: Advertising: Effects; Advertising: General; Jakobson, Roman (1896–1982); Mass Media, Representations In; Media Effects; Metaphor, Role in Social Thought: History of the Concept; Political Communication; Popular Culture; Semiotics.

Bibliography

Barthes, R., 1964. Rhétorique de l'image. Communications 4 [1977 Rhetoric of the image. In: Image Music Text. Fontana Press, London].

Barthes, R., 1970. L'ancienne rhétorique. Aide-mémoire. Communications 16 [1988 the Old Rhetoric. An aide-mémoire. In: The Semiotic Challenge. Basil Blackwell, Oxford, UK].

Billig, M., 1987. Arguing and Thinking. A Rhetorical Approach to Social Psychology. Cambridge University Press, Cambridge, UK.

Burke, K., 1941. The rhetoric of Hitler's battle. In: The Philosophy of Literary Form. Studies in Symbolic Action. Louisiana State University Press, Baton Rouge, LA.

Burke, K., 1962. A Grammar of Motives and a Rhetoric of Motives. World Publishing Company, Cleveland, OH.

Durand, J., 1970[1983]. Rhétorique et image publicitaire. Communications 15 [Rhetoric and the Advertising Image. In: Australian Journal of Cultural Studies 1(2)].

Durand, J., 1987. Rhetorical figures in the advertising image. In: Umiker-Sebeok, J. (Ed.), Marketing and Semiotics. New Directions in the Study of Signs for Sale. Mouton de Gruyter, Berlin, New York, Amsterdan, pp. 295–318.

Dyer, G., 1982. Advertising as Communication. Methuen, London.

Genette, G., 1970. La rhétorique restreinte. Communications 16, 158–171.

(Dubois, J., Edeline, F., Klinkenberg, J-M., Minguet, P.) Groupe, μ, 1970. Rhétorique générale. Libraire Larousse, Paris [1981 A General Rhetoric. Johns Hopkins University Press, Baltimore, MD].

(Edeline, F., Klinkenberg, J-M., Minguet, P.) Groupe μ, 1992. Traité du signe visuel: pour une rhétorique de l'image. Seuil, Paris.

Jakobson, R., 1956. Two aspects of language and two types of aphasic disturbances. In: Jakobson, R., Halle, M. (Eds.), Fundamentals of Language. Mouton, The Hague.

Lakoff, G., 1991. Metaphor and war: the metaphor system used to justify war in the Gulf. In: Hallet, B. (Ed.), Engulfed in War: Just War and the Persian Gulf. Matsunaga Institute for Peace, Honolulu, HI.

Lakoff, G., 2004. Don't Think of an Elephant! White River Junction. Green Publishing, VT: Chelsea.

Lakoff, G., Johnson, M., 1980. Metaphors We Live by. University of Chicago Press, Chicago, IL.

Leiss, W., et al., 2005. Social Communication in Advertising. Consumption in the Mediated Marketplace. Routledge, New York.

McQuarrie, E.F., Mick, D.G., 1999. Visual rhetoric in advertising: text-interpretive, experimental, and reader-response analyses. The Journal of Consumer Research 21 (1).

McQuarrie, E.F., Mick, D.G., 2003. The contribution of semiotic and rhetorical perspectives to the explanation of visual persuasion in advertising. In: Scott, L., Batra, R. (Eds.), Persuasive Imagery: A Consumer Response Perspective. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 191–221.

Messaris, P., 1997. Visual Persuasion. Sage, Thousand Oaks.

Ong, W.J., 1982. Orality and Literacy. The Technologizing of the Word. Methuen, London.

Perelman, C., Olbrects-Tyteca, L., 1958. Traité de l'argumentation; la nouvelle rhétorique, first ed., vols. 1–2. Presses universitaires de France, Paris. [1969 The New Rhetoric. A Treatise on Argumentation. University of Notre Dame Press, Notre Dame, IN].

Philips, B., McQuarrie, E.F., 2004. Beyond visual metaphor: a new typology of visual rhetoric in advertising. In Marketing Theory 4 (1/2).

Toulmin, S.E., 1958. The Uses of Argument. Cambridge University Press, Cambridge, UK.

Vickers, B., 1988. In Defence of Rhetoric. Clarendon Press, Oxford, UK.

Rheumatoid Arthritis: Psychosocial Aspects

Christina Bode and Erik Taal, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rheumatoid Arthritis (RA) is a chronic inflammatory disease, which is accompanied by swollen and painful joints, fatigue and limitations in physical functioning. Psychological reactions to disease symptoms and related attitudes and psychological characteristics of patients clearly influence the impact of disease symptoms on psychological distress and well-being, on daily activities and on social participation. In addition to pharmacological treatment of disease activity, regular monitoring of patient-reported symptoms and timely psychological stepped care approaches seem the best way to support and empower patients to successfully adapt to RA and to live a meaningful life with RA.

Rheumatoid Arthritis: Etiology, Prevalence, Course and Disease Management

Rheumatoid Arthritis (RA) is a chronic inflammatory disorder based on an autoimmune process. The immune tolerance for the own body is broken down and the (chronic) inflammatory reaction affects cartilage, joints and bones, and possibly other organs.

The course of the disease is heterogeneous. It ranges from mild courses without bone erosion and spontaneous remissions to severe, rapidly progressive forms with damage of joints, disability, and high negative impact on quality of life.

Relevant factors discussed for the etiology of RA are genetic predispositions in interaction with environmental factors that determine the immune response and inflammatory reaction. As important environmental factors mainly two aspects have been discussed: infections might start the autoimmune response and cigarette smoking increases the risk of developing RA and is associated with more severe forms of the disease. In addition to environmental factors, hormonal and reproductive factors may play a role in the development and severity of RA, suggested by the fact that clearly more women than men are affected by RA (Kvien et al., 2009).

RA affects 0.5–1.0% of adults in developed countries; it is more frequent in women (3:1) and prevalence rises with age (Scott et al., 2010).

Pharmacological treatment aims at the control of disease activity, in the best case this means remission to prevent joint damage and minimize pain, fatigue, disability, and reestablish social and work participation. If remission is not achievable, the pharmacological treatment aims to reach and maintain the lowest possible level of disease activity as soon as possible in every patient. As long as this target has not been reached, treatment should be adjusted by frequent and strict monitoring (European League Against Rheumatism recommendation for management of RA: Smolen et al., 2010).

Nonpharmacological treatment for RA includes patient education, exercise therapy, joint protection, psychological support, and foot care. Ideally, multidisciplinary teams of rheumatologists, nurses, physiotherapists, occupational therapists, psychologists, and podiatrists deliver holistic care for patients with RA. In reality, however, only very few specialized hospitals offer multidisciplinary team care for RA patients.

Psychophysical Aspects of RA: Pain, Fatigue, and Physical Functioning

Pain

Patients with RA perceive joint pain as one of the most important problems of their disease (Taal et al., 1993; Walsh and McWilliams, 2012). Pain is caused by inflammation and structural damage of the joints. In a 10-year longitudinal study (with measurements at baseline and after 1, 2, 5, and 10 years) of patients with recently diagnosed RA the mean pain score of the patients on a visual analog scale from 0 (no pain) to 100 (severe pain) was between 28 and 33 on all measurements over time. At all time points, 30–35% of the patients had clinically relevant pain scores $\geq$40 (Odegard et al., 2007). Patients with RA see pain as their highest priority for improvement, although it may improve over time (Ten Klooster et al., 2007). Modern early and intensive pharmacological treatments lead to successful suppression of inflammation and improved levels of pain (Walsh and McWilliams, 2012).

Next to inflammation and structural damage, psychological factors may influence the experience of pain (Walsh and McWilliams, 2012). Odegard et al. (2007) showed that over 10 years anxiety was longitudinally associated with the course of pain. Schieir et al. (2009) found that also baseline depression predicted the course of pain from baseline to 6 months in patients with early inflammatory arthritis. In a 1-year longitudinal study feelings of helplessness and passive coping significantly mediated the relationship between physical disability and future pain in patients with RA (Covic et al., 2003). In a study with weekly measures over a duration of 11 weeks anxiety, depression, and negative and positive affect predicted weekly changes in pain in patients with RA and osteoarthritis. The effects of anxiety on pain were much stronger than the effects of depression. Remarkably, the effect of anxiety on pain was mediated by negative affect and the effect of depression was mediated by positive affect (Smith and Zautra, 2008). Lower social support was related to increased pain after 3 and 5 years in patients with recently diagnosed RA. Contrary to expectations passive pain-coping was not related to increased pain levels (Evers et al., 2003).

The role of psychosocial factors in the experience of pain indicates that, next to pharmacological treatment, there is a role for psychological interventions for the management of pain.

Self-efficacy beliefs of patients (that they can manage their pain) and the actual use of pain-coping strategies influence pain experiences and also the impact of pain on physical function and psychological distress (Keefe et al., 2008). Several studies have shown that pain-coping skills training and cognitive behavioral therapy can be successful in reducing pain. Furthermore, it has been shown that these interventions also lead to improvements in anxiety, depression, and physical function (Keefe et al., 2008).

Fatigue

Severe fatigue is commonly reported by patients with RA. Prevalence of severe fatigue differs depending on studied sample and measurement instrument; 40–80% of patients diagnosed with RA report persistent fatigue symptoms (Wolfe et al., 1996; Pollard et al., 2006). Qualitative research from different countries has shown that patients experience fatigue as a multidimensional, annoying symptom with far-reaching consequences (Hewlett et al., 2005; Repping-Wuts et al., 2008; Nikolaus et al., 2010). Physical, cognitive, and emotional aspects of fatigue have been described, and also perceived causes and consequences of fatigue. Additionally, interindividual and intraindividual variations in fatigue severity and the uncontrollability of fatigue symptoms were found as important dimensions from the patient's perspective.

Fatigue influences fulfillment of social roles and participation negatively: functioning as parent, partner, friend, and also in the work domain. Fatigue was identified as an important predictor of work absenteeism and early retirement (De Croon et al., 2005). In contrast to the far-reaching consequences and the priority from patient perspective, fatigue plays no role in treatment decisions and is mainly neglected in the communication between health care providers and patients. This might be explained by the fact that the knowledge about the etiology of fatigue symptoms is very limited.

A systematic review (Nikolaus et al., 2013) provides an overview of the existing literature on possible causes and consequences of fatigue in RA. Studies report possible causes for fatigue among illness-related aspects (pain), physical functioning (sleep quality, sleep disturbances, global health ratings, disability), cognitive/emotional functioning (depression, anxiety, perceptions as self-efficacy), social/environmental aspects (interpersonal events, social support), and female gender. However, none of these variables showed profound and stable relationships with fatigue across studies. Inconsistent results were especially reported with regard to inflammatory activity (e.g., ESR, DAS28, flares). Moreover, it is unclear how the proposed directions between fatigue and the other variables correspond with the real processes. Probably, multiple, circular processes underlie fatigue in RA (Hewlett et al., 2011), so that it is not possible to identify single causes of fatigue without considering interactions with other potential pathways. Nicassio et al. (2012) found some cross-sectional evidence for an indirect relationship of disease activity and fatigue through its effects on mood disturbances and poor quality of sleep. Based on the systematic review of Nikolaus et al. (2013), three variables have the highest probability to be involved in the complex process of fatigue in RA: pain, physical functioning, and depressive mood.

Fatigue has scarcely been studied as primary outcome of pharmacological treatment. A meta-analysis by Chauffier et al. (2012) has demonstrated that the effects of biotherapies on fatigue in RA are very limited and that they fall short of the expectations with regard to the fatigue-reducing potential of new pharmacological treatment options. A review on non-pharmacological interventions provides some evidence that physical activity and psychosocial interventions have positive effects on self-reported fatigue in adults with RA, for other nonpharmacological interventions there is currently insufficient evidence for their effectiveness (Cramp et al., 2013).

Physical Functioning

The inflammation and structural damage of joints result in restrictions in physical function in patients with RA (Hazes, 2003). Physical function is the ability of a person to perform activities of daily living (ADL). In interviews with RA patients, functional disability ('not being able to do things,' 'mobility restrictions') was most often reported as one of the main problems of having RA (Taal et al., 1993). Recent studies showed that many concerns about the future of patients with RA are related to restrictions in physical functioning such as dependency on others regarding activities such as shopping or self-care, no longer being able to walk or to do leisure activities (Buitinga et al., 2012).

The golden standard to assess restrictions in physical function in patients with RA is the Health Assessment Questionnaire – Disability Index (HAQ-DI), which measures difficulties in activities that involve both the lower and upper extremities. It has questions in eight categories of ADL's that can be affected by RA: dressing and grooming, rising, eating, walking, hygiene, reach, grip, activities (errands, chores) (Bruce and Fries, 2003). These categories cover most domains of physical function that have been identified in the International Classification of Functioning, Disability and Health (ICF) core-set for RA (WHO, 2001; Oude Voshaar et al., 2011). About 10% of patients with RA score 0 on the HAQ-DI meaning they have no difficulties in performing activities in all eight categories of daily living (Krishnan et al., 2004). Patients with short disease duration often have relatively high restrictions in physical function mainly related to high levels of disease activity. After start of medical treatment, physical function initially improves followed by a slow decline in physical function over the years, mainly related to increased joint damage (Verstappen, 2013). Recent advances in pharmacological treatment of RA have shown large improvements in physical function after treatment, especially in patients who achieve clinical remission (Kekow et al., 2011). Restrictions in physical function tend to increase with age and are slightly higher in women than men. Also comorbidities (such as cardiovascular or respiratory diseases), low socioeconomic status, and low educational levels are associated with worse physical function (Verstappen, 2013).

The decline of physical function over time is also affected by psychosocial factors (Hazes, 2003). Depressive symptoms were related to physical disability, but they were not predictors of changes in disability over 2 years in patients with RA (Rupp et al., 2006). Kekow et al. (2011) reported that patients with severe symptoms of depression or anxiety (HADS scores ≥ 8)

before starting pharmacological treatment of RA had significantly worse physical function and achieved considerably less improvement in physical function after 2 years of treatment compared to patients without symptoms of depression or anxiety. Evers et al. (2003) showed that the use of more passive pain coping and having less social support is related to increased physical disability after 3 and 5 years in patients with recently diagnosed RA. Worse physical function is related to higher work disability (Young et al., 2002) and more psychological distress (Norton et al., 2011) over time, and is a strong predictor of mortality (Sokka et al., 2008).

There is some evidence that exercise programs and occupational therapy improves physical function in RA (Vliet Vlieland and van den Ende, 2011). A systematic review of the effectiveness of patient education programs for patients with RA shows small beneficial short-term effects on physical function, especially for programs that included behavioral instructions. However, no long-term effects on physical function were found (Riemsma et al., 2004).

Psychological (Dys)functioning: Depression and Anxiety

In the previous paragraphs we have described that pain, fatigue, and limitations in physical functioning are associated with depressive symptoms. Therefore, we now take a closer look at depression. Depression is a frequent comorbidity in RA. Prevalence of depression depends on studied population, used measures, observation time, and culture/country, but in general the levels of depression are clearly elevated compared to rates of depression in populations without RA. According to a recent systematic review and meta-analyses of 72 studies on the prevalence of depression in adult RA patients (Matcham et al., 2013), prevalence rates vary between 17 (major depressive disorder, using clinical interviews) and 39% (significant depressive symptoms, measured with screening instruments).

Merely considering clinical levels of depression fails to document subclinical levels of depressive mood, which may constitute a transitional phase to clinical levels and also mean suffering for patients with RA. In addition, both clinical and subclinical levels of depression have serious health-related consequences. Poorer adherence, maladaptive health behavior, increase of RA symptoms, poorer functional outcomes, and lower satisfaction with social relationships (network size and quality of relationships) and social participation (work- and nonwork related) and finally more health care utilization are documented as related to depression. Few studies investigated the relationship between depression and mortality in people with RA. They found clearly higher risks for all-course mortality, increased risk of having a heart attack and other comorbidities (Scherrer et al., 2009; Ang et al., 2005).

Diagnosing and interpreting depression is complex due to the overlap of other RA symptoms (fatigue, sleep problems, reduced levels of interest, and pleasure in activities) and also because of the inflammation processes that might be associated with depression through elevated levels of cytokines (Geenen et al., 2012). With regard to interpretation, it is debatable whether short-term depressive feelings reflect a normal (adaptive) reaction to the diagnosis and consequences of RA for daily functioning. Patients with stable elevated levels of mood problems, however, should definitely be screened for depression and treated accordingly. Van't Land et al. (2010) investigated whether depression precedes RA. They found that people with arthritis are more likely to develop depressive symptoms compared to people without arthritis, whereas having any mental disorder did not predict the new-onset of arthritis. In younger age groups (<45 years of age) this relationship was even more pronounced. These results suggest that younger patients with RA are even more at risk to develop clinical forms of depression. Also the meta-analyses by Matcham et al. (2013) found that younger age was associated with increased depression prevalence.

Whereas depression has been investigated widely, anxiety seems to be an overlooked aspect of problematic psychological functioning in patients with RA. Although depression and anxiety have a substantial overlap (VanDyke et al., 2004, reported 69% shared variance) there are at least three reasons to investigate and monitor anxiety and depression separately. First, the prevalence of anxiety might be higher than the prevalence of depression (20–30% elevated levels of anxiety, Geenen, et al., 2012) and anxiety might have stronger effects on current and next week pain (twice as large according to Smith and Zautra, 2008). Second, different mechanisms of anxiety and depression on pain in arthritis patients have been postulated: while anxiety is assumed to affect pain through anxiety sensitivity, pain-relevant anxiety, and fear avoidance, depression and pain might be linked through catastrophizing less positive reinforcement and helplessness (Smith and Zautra, 2008). These assumptions, however, need more substantiation but they potentially imply different treatment strategies for anxiety and depression. The third reason is the possible temporal sequence with first anxiety followed by depression that would suggest more focus on anxiety to prevent depressive episodes (VanDyke et al., 2004).

In summary, a range of factors have been identified to be associated with distress for patients with RA, which might manifest in subclinical or clinical levels of depression and anxiety. Prominent candidates for developing mood problems are somatic symptoms, functional limitations, proinflammatory cytokines, helplessness, and fear for the future due to the unpredictable and progressive character of the disease.

Social Consequences: Social Participation and Work Participation

RA can lead to restrictions in participation in work and other social roles (Geuskens et al., 2007). The ICF Model of the World Health Organization defines social participation restrictions as problems an individual may have in involvement in life situations (e.g., looking after family members or doing paid work); they are distinguished from physical limitations a person may have in executing basic tasks (e.g., walking or washing/personal hygiene) (WHO, 2001). Geuskens et al. (2007) have shown that patients with RA are more restricted in work participation and socializing compared to the general population. In cross-sectional studies it was found that role and social functioning, as measured with the Short Form 36 (a self-report questionnaire that measures physical, mental, and social

aspects of health-related quality of life with 36 items (Ware and Sherbourne, 1992)) were associated with disease activity, pain, physical limitations, vitality, and mental health (Fransen et al., 2002; Kuhlow et al., 2010). In a recent study involving patients with polyarthritis, restrictions in social and work participation were cross-sectionally related to functional limitations and fatigue (Arends et al., 2013). This study revealed that also psychological factors were significantly associated to restrictions in social and work participation: self-efficacy in coping with arthritis and the use of goal management strategies (i.e., the way people generally react to threatened life goals) played an important role in the satisfaction with social participation. Katz et al. (2006) studied loss of participation in valued life activities (VLA), a concept that partly overlaps with social participation restrictions as defined by the ICF model. VLA are grouped into obligatory activities (e.g., hygiene and self-care, walking, using transportation), committed activities (e.g., paid work, household, and family care responsibilities), and discretionary activities (e.g., socializing, recreation, hobbies). Only restrictions that are deemed important by a patient are incorporated in individual total VLA scores. In a cross-sectional study among RA patients more restrictions in valued discretionary and committed activities were found than in obligatory activities. Difficulties in performing VLA could be predicted by higher age, being female, longer disease duration, and symptom measures such as fatigue, pain, morning stiffness, and feet deformities. In this study, relationships between symptom measures and difficulty in performing VLA were mediated by functional limitations (Katz et al., 2006).

Longitudinal studies are needed to gain better insight in the causal relationships between disease activity, symptoms as pain and fatigue, physical limitations and restrictions in social and work participation, and the role of psychosocial factors in this context.

Living a Good Life with Arthritis: Supported by Regular Monitoring and Timely Support

In the last decennium, modern pharmacological treatment strategies have contributed to meaningful progress in the treatment of RA. About 50% of remission in early RA patients has been reported, resulting in lower disability rates, less radiological damage, and also fewer patient-reported symptoms. Nevertheless, findings also indicate that the presence of elevated levels of depression or anxiety at start of treatment delays clinical remission (Kekow et al., 2011). This points to the fact that monitoring of psychological functioning and early intervention can even increase the potential of successful pharmacological therapies, which are directed at controlling disease activity.

While these advances in pharmacological treatment mean less impact of RA on their lives for a number of patients, for others it remains a challenge to deal with symptoms of pain, fatigue, decreased physical functioning, and fewer possibilities for social participation and use of their personal potential. Developing skills to manage RA symptoms and finding a new satisfying life balance can be supported in various ways. Patient education and self-management programs have been developed with the aim to enhance living well with RA. The results of these programs, however, are mixed. Many short-term effects cannot be shown to last over longer periods (>12 months). Unfortunately, many effect studies have focused on (medical) outcomes that were not targeted in the interventions, disadvantaging potential benefits of these programs. Iversen et al. (2010) have reviewed self-management interventions and have described common characteristics of effective programs: duration of at least 6 weeks, explicit use of social cognitive theory or cognitive behavioral therapy approaches, individualized action plans with progress review, highly protocolized with participant handbooks, and led by the same trained leaders.

Currently, frequent monitoring of patient-reported symptoms in combination with stepped-care approaches seems to be the best way to support and empower patients in dealing with arthritis. Also, shifting the emphasis from only treating psychosocial problems to enhancing patients in dealing proactively with their disease might prevent the negative downward spiral of physical, psychological, and social limitations and opens opportunities for a new meaningful life with RA (Arends et al., 2013). Savings in health care systems and focus on medical treatment rather than holistic approaches ask for creative and innovative solutions. E-health applications for monitoring and proactive self-management of patients have the potential to deliver a significant contribution. Whether this way of health communication and treatment can reach all groups of patients and can be supported by all health professionals is questionable. Tailoring the choice of communication and treatment channels to patients' capacities and also to opportunities of care providers might be necessary to prevent new inequalities in health care and diminish negative psychosocial consequences of RA as much as possible.

See also: Chronic Pain: Models and Treatment Approaches; Co-morbidity of Mental and Physical Conditions; Control Beliefs: Health Perspectives; Illness Behavior and Care-Seeking; Pain, Health Psychology of; Self-Efficacy and Health.

Bibliography

Ang, D.C., Choi, H., Kroenke, K., Wolfe, F., 2005. Comorbid depression is an independent risk factor for mortality in patients with rheumatoid arthritis. Journal of Rheumatology 32 (6), 1013–1019.

Arends, R.Y., Bode, C., Taal, E., Van de Laar, M.A.F.J., 2013. The role of goal management for successful adaptation to arthritis. Patient Education and Counseling 93 (1), 130–138.

Buitinga, L., Braakman-Jansen, L.M.A., Taal, E., Van de Laar, M.A.F.J., 2012. Worst-case future scenarios of patients with rheumatoid arthritis: a cross-sectional study. Rheumatology 51, 2027–2033.

Bruce, B., Fries, J., 2003. The Stanford health assessment questionnaire: a review of its history, issues, progress, and documentation. Journal of Rheumatology 30 (1), 167–178.

Chauffier, K., Salliot, C., Berenbaum, F., Sellam, J., 2012. Effect of biotherapies on fatigue in rheumatoid arthritis: a systematic review of the literature and meta-analysis. Rheumatology 51, 60–68.

Cramp, F., Hewlett, S., Almeida, C., Kirwan, J.R., Choy, E.H.S., Chandler, T., Pollock, J., Christensen, R., 2013. Non-pharmacological interventions for fatigue in rheumatoid arthritis. Cochrane Database of Systematic Reviews (8): CD008322.

Covic, T., Adamson, B., Spencer, D., Howe, G., 2003. A biopsychosocial model of pain and depression in rheumatoid arthritis: a 12-month longitudinal study. Rheumatology 42, 1287–1294.

De Croon, E.M., Sluiter, J.K., Nijssen, T.F., Kammeijer, M., et al., 2005. Work ability of Dutch employees with rheumatoid arthritis. Scandinavian Journal of Rheumatology 34 (4), 277–283.

Evers, A.W.M., Kraaimaat, F.W., Geenen, R., Jacobs, J.W.G., Bijlsma, J.W.J., 2003. Pain coping and social support as predictors of long-term functional disability and pain in early rheumatoid arthritis. Behaviour Research and Therapy 41 (11), 1295–1310.

Fransen, J., Uebelhart, D., Stucki, G., Langenegger, T., Seitz, M., Michel, B.A., 2002. The ICIDH-2 as a framework for the assessment of functioning and disability in rheumatoid arthritis. Annals of the Rheumatic Diseases 61, 225–231.

Geenen, R., Newman, S., Bossema, E.R., Vriezekolk, J.E., Boelen, P.A., 2012. Psychological interventions for patients with rheumatic diseases and anxiety or depression. Best Practice & Research Clinical Rheumatology 26, 305–319.

Geuskens, G.A., Burdorf, A., Hazes, J.M.W., 2007. Consequences of rheumatoid arthritis for performance of social roles – a literature review. Journal of Rheumatology 34, 1248–1260.

Hazes, J.M.W., 2003. Determinants of physical function in rheumatoid arthritis: association with the disease process. Rheumatology 42 (Suppl. 2), Sii17–S25.

Hewlett, S., Cockshott, Z., Byron, M., Kitchen, K., Tipler, S., Pope, D., et al., 2005. Patients' perceptions of fatigue in rheumatoid arthritis: overwhelming, uncontrollable, ignored. Arthritis Care and Research 53, 697–702.

Hewlett, S., Chalder, T., Choy, E., Cramp, F., Davis, B., Dures, E., et al., 2011. Fatigue in rheumatoid arthritis: time for a conceptual model. Rheumatology 50, 1004–1006.

Iversen, M., Hammond, A., Betteridge, N., 2010. Self-management of rheumatic diseases: state of the art and future perspectives. Annals of the Rheumatic Diseases 69, 955–963.

Katz, P.P., Morris, A., Yelin, E.H., 2006. Prevalence and predictors of disability in valued life activities among individuals with rheumatoid arthritis. Annals of the Rheumatic Diseases 65, 763–769.

Keefe, J., Somers, T.J., Martire, L.M., 2008. Psychologic interventions and lifestyle modifications for arthritis pain management. Rheumatic Disease Clinics of North America 34, 351–368.

Kekow, J., Moots, R., Khandker, R., Melin, J., Freundlich, B., Singh, A., 2011. Improvements in patient-reported outcomes, symptoms of depression and anxiety, and their association with clinical remission among patients with moderate-to-severe active early rheumatoid arthritis. Rheumatology 50, 401–409.

Krishnan, E., Tugwell, P., Fries, J.F., 2004. Percentile benchmarks in patients with rheumatoid arthritis: health assessment questionnaire as a quality indicator (QI). Arthritis Research and Therapy 6 (6), R505–513.

Kuhlow, H., Fransen, J., Ewert, T., Stucki, G., Forster, A., Langenegger, T., et al., 2010. Factors explaining limitations in activities and restrictions in participation in rheumatoid arthritis. European Journal of Physical and Rehabilitation Medicine 46, 169–177.

Kvien, T.K., Scherer, H.U., Burmester, G.-R., 2009. Rheumatoid arthritis. In: Bijlsma, J.W.J., et al. (Eds.), EULAR Compendium on Rheumatic Diseases. BMJ Publishing Group and European League Against Rheumatism, London, pp. 61–80.

Matcham, F., Rayner, L., Steer, S., Hotopf, M., 2013. The prevalence of depression in rheumatoid arthritis: a systematic review and meta-analysis. Rheumatology 52, 2136–2148.

Nicassio, P.M., Ormseth, S.R., Custodio, M., Irwin, M.R., Olmstaed, R., Weisman, M.H., 2012. A multidimensional of fatigue in patients with rheumatoid arthritis. Journal of Rheumatology 39, 1807–1813.

Nikolaus, S., Bode, C., Taal, E., Van de Laar, M.A.F.J., 2013. Fatigue and factors related to fatigue in rheumatoid arthritis: a systematic review. Arthritis Care and Research 65 (7), 1128–1146.

Nikolaus, S., Bode, C., Taal, E., Van de Laar, M.A.F.J., 2010. New insights into the experience of fatigue among patients with rheumatoid arthritis: a qualitative study. Annals of the Rheumatic Diseases 69, 895–897.

Norton, S., Sacker, A., Young, A., Done, J., 2011. Distinct psychological distress trajectories in rheumatoid arthritis: findings from an inception cohort. Journal of Psychosomatic Research 71 (5), 290–295.

Odegard, S., Finset, A., Mowinckel, P., Kvien, T.K., Uhlig, T., 2007. Pain and psychological health status over a 10 year-period in patients with recent onset rheumatoid arthritis. Annals of the Rheumatic Diseases 66, 1195–1201.

Oude Voshaar, M.A., ten Klooster, P.M., Taal, E., van de Laar, M.A., 2011. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health and Quality of Life Outcomes 9, 99.

Pollard, L.C., Choy, E.H., Gonzalez, J., Koshaba, B., Scott, D.L., 2006. Fatigue in rheumatoid arthritis reflects pain, not disease activity. Rheumatology 45, 885–889.

Repping-Wuts, H., Uitterhoeve, R., van Riel, P., van Achterberg, T., 2008. Fatigue as experienced by patients with rheumatoid arthritis (RA): a qualitative study. International Journal of Nursing Studies 45, 995–1002.

Riemsma, R.P., Taal, E., Kirwan, J.R., Rasker, J.J., 2004. A systematic review of rheumatoid arthritis patient education. Arthritis Care and Research 51, 1045–1059.

Rupp, I., Boshuizen, H.C., Dinant, H.J., Jacobi, C.E., Van Den Bos, G.A.M., 2006. Disability and health-related quality of life among patients with rheumatoid arthritis: association with radiographic joint damage, disease activity, pain, and depressive symptoms. Scandinavian Journal of Rheumatology 35 (3), 175–181.

Scherrer, J.F., Virgo, K.S., Zerinque, A.M.S., Bucholz, K.K., Jacob, T., Johnson, R.G., et al., 2009. Depression increases risk of incident myocardial infarction among veteran administration patients with rheumatoid arthritis. General Hospital Psychiatry 31, 353–359.

Schieir, O., Thombs, B.D., Hudson, M., Taillefer, S., Steele, R., Berksonm, L., et al., 2009. Symptoms of depression predict the trajectory of pain among patients with early inflammatory arthritis: a path analysis approach to assessing change. Journal of Rheumatology 36, 231–239.

Scott, D., Wolfe, F., Huizinga, T.W., 2010. Rheumatoid arthritis. Lancet 376, 1094–1108.

Smith, B.W., Zautra, A.J., 2008. The effects of anxiety and depression on weekly pain in women with arthritis. Pain 138, 354–361.

Smolen, J.S., Landewe, R., Breedveld, F.C., Dougados, M., et al., 2010. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs. Annals of the Rheumatic Diseases 69, 964–975.

Sokka, T., Abelson, B., Pincus, T., 2008. Mortality in rheumatoid arthritis: 2008 update. Clinical and Experimental Rheumatology 26, S35–S61.

Taal, E., Rasker, J.J., Seydel, E.R., Wiegman, O., 1993. Health status, adherence with health recommendations, self-efficacy and social support in patients with rheumatoid arthritis. Patient Education and Counseling 20, 63–76.

Ten Klooster, P.M., Taal, E., Veehof, M.M., Riel, P.L.C.M., Van de Laar, M.A.F.J., 2007. Changes in priorities for improvement in patients with rheumatoid arthritis during one year of anti-TNF Treatment. Annals of Rheumatic Diseases 66 (11), 1485–1490.

VanDyke, M., Parker, J.C., Smarr, K.L., Hewett, J.E., Johnson, G.E., Slaughter, J.R., Walker, S.E., 2004. Anxiety in rheumatoid arthritis. Arthritis Care and Research 51 (3), 408–412.

Van't Land, H., Verdurmen, J., Ten Have, M., et al., 2010. The association between arthritis and psychiatric disorders; results from a longitudinal population-based study. Journal of Psychosomatic Research 68 (2), 187–193.

Verstappen, S.M., 2013. Outcomes of early rheumatoid arthritis – the WHO ICF framework. Best Practice & Research Clinical Rheumatology 27 (4), 555–570.

Vliet Vlieland, T.P., van den Ende, C.H., 2011. Nonpharmacological treatment of rheumatoid arthritis. Current Opinion in Rheumatology 23 (3), 259–264.

Walsh, D.A., McWilliams, D.F., 2012. Pain in rheumatoid arthritis. Current Pain and Headache Reports 16, 509–517.

Ware Jr., J.E., Sherbourne, C.D., 1992. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care 30 (6), 473–483.

World Health Organization, 2001. The International Classification of Functioning, Disability and Health – ICF. World Health Organization, Geneva (Switzerland).

Wolfe, F., Hawley, D.J., Wilson, K., 1996. The prevalence and meaning of fatigue in rheumatic disease. Journal of Rheumatology 23 (8), 1407–1417.

Young, A., Dixey, J., Kulinskaya, E., Cox, N., Davies, P., Devlin, J., Emery, P., Gough, A., James, D., Prouse, P., Williams, P., Winfield, J., 2002. Which patients stop working because of rheumatoid arthritis? Results of five years' follow up in 732 patients from the early RA study (ERAS). Annals of the Rheumatic Diseases 61, 335–340.

Ricardo, David (1772–1823)

Paul A Samuelson, Massachusetts Institute of Technology, Cambridge, MA, USA
Bryan S Turner, The Graduate Center, The City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P.A. Samuelson, volume 20, pp. 13330–13334, © 2001, Elsevier Ltd.

Abstract

David Ricardo is widely regarded as a founding figure in both classical economics and political economy alongside such figures as Adam Smith and Thomas Malthus. As a practicing financier on the London stock market, he brought his practical understanding of economics to basic problems in economic theory. Although his contributions to economics have often been criticized, his position as a founding father of the discipline is secure. A full-scale biography was published in 2013 and his collected works were edited by Piero Sraffa in 11 volumes (1951–73). Ricardo made important contributions to political debates in his time such as the corn trade. He made lasting contributions to both microeconomics and macroeconomics in the law of diminishing returns and to the labor theory of value; consequently, his work was much admired by Karl Marx. By recognizing that technological improvements could reduce the need for labor if not displace it, he anticipated what is now referred to as the de-skilling of labor. However, according to Paul Samuelson, Ricardo's major contribution was to the theory of comparative advantage, which says that, given international competition, an economy should specialize. Finally, Ricardo was a champion of free trade and the so-called *laissez-faire* economics. While Ricardo is highly regarded in economics, his contribution to the theory of social classes in his analysis of rent, profit, and wages has been neglected in sociology.

One of the most influential of the classical economists, David Ricardo was born in London in 1772 to a Sephardic Jewish family of Portuguese origin, who had emigrated from the Netherlands. At 14 years of age Ricardo joined his father at the London Stock Exchange, where he acquired a practical understanding of the actual mechanics of both domestic and international finance. These early experiences laid the foundation for his subsequent success in amassing a personal fortune. Rejecting the orthodox beliefs of his Jewish family, he eloped at 21 years of age with Priscilla Anne Wilkinson in 1793, a Quaker by whom he had eight children. On subsequently becoming a Unitarian, he was rejected by his family. The elopement and separation from his parents resulted in a period of personal hardship, forcing him to seek new means to support himself.

With his practical experience of finance, he became a highly successful stockbroker by unscrupulously trading on the bond market during the Battle of Waterloo in 1815. He was able to post an observer at the famous battle conducted by the Duke of Wellington to convey the results of the military engagement. By creating the impression that the battle had been won by the French, he bought government bonds at a sharp discount in the ensuing market panic to which he had contributed. This early success formed the basis of his personal fortune, which was estimated on his death to be worth $100 million in present day dollar values. His private wealth allowed Ricardo to retire early and to engage in his scientific interests. He became attracted to the study of economics as a result of reading Adam Smith's *The Wealth of Nations* (1776) while on vacation in the English resort of Bath in 1799. Ricardo began to write influential articles about the pressing economic matters of the day such as the corn trade and became acquainted in 1811, with such influential figures as Jeremy Bentham and James Mill. He engaged a lengthy correspondence with Thomas Malthus over the role of the landowning class in British society. He achieved prominence as a member of several influential London clubs such as the Political Economy Club, the King's Club, and the Geological Society. These social connections were important in his entry into Parliament in 1818 through purchasing the borough of Lord Portarlington for 4000 sterling. Ricardo died at Gatcombe Park, Gloucestershire, at 51 years of age from an ear infection that resulted in septicemia. Buried at St Nicholas Churchyard Chippenham, Wiltshire, he never lived to see the 1836 Repeal of the Corn Laws, which was a vindication of his pamphleteering for free trade, and he did not live to witness Ireland's great famine of the 1840s with its brutal confirmation of the laws of the free market.

Among the educated public only a few names stand out in the history of economic thought. These are Adam Smith, the spokesman for individualism and competitive capitalism; Robert Malthus, who earned for economics the title of 'the dismal science' from Thomas Carlyle in 1849, by virtue of his belief that any temporary gain in well-being would be undermined as a consequence of inducing a population explosion that would entail diminishing returns and a retreat back to a subsistence ration of real wages; Karl Marx, the prophet of capitalist crises and the exponent of allegedly scientific socialism; John Maynard Keynes, proponent of the mixed economy, who rescued capitalism from the post-1929 Great Depression and designed Britain's postwar recovery by interventionist governmental macroeconomic policies and public regulation. The classical political economy of Malthus and Ricardo is often associated with the idea of negative unintended consequences of action. Both men were criticized by Albert O. Hirschman for their 'rhetoric of reaction' or the 'perversity thesis,' namely "any purposive action to improve some feature of the political, social, or economic order only serves to exacerbate the condition one wishes to remedy" (Hirschman, 1991: 7). For example, raising the wages of the working class would be self-defeating, because it would increase the fertility rate of the poor resulting eventually in more poverty.

Within the economics profession, Ricardo has now achieved a prominent status, but in the past he was not always respected and recognized. In his *History of Economic Analysis*, Joseph A. Schumpeter (1954) coined the phrase 'the Ricardian vice' to describe how Ricardo piled up arguments to hide his confusion of facts and values. He also complained that Ricardo had failed to make any real progress in the development of a general equilibrium model. Among twentieth-century economists perhaps the most important tribute to Ricardo came from Paul A. Samuelson (1969), who was once challenged by Stanislaw Ulam the Polish mathematician, to name any nontrivial universal law in the social sciences. After some reflection, Samuelson claimed that it was Ricardo's analysis of 'comparative advantage,' which stated that to achieve successful development a society should not attempt to focus on a range of diverse productive activities, but only on those economic activities that gave them a comparative advantage in international trade.

In retrospect we can now see him as the archetypical classical economist, the economist's economist who supposedly improved upon Smith and spelled out with Euclidean logic the eternal behavior of the market system. In an amiable exchange of letters with Malthus, Ricardo was undoubtedly the outright winner, because Malthusian fears of recurrent depression stemming from underconsumption spending were argued to be fallacies. The self-made millionaire David Ricardo captured the respect of the Victorian devotees to *laissez-faire* as well as socialistic critics of Marxian or romantic stripe. He has recently received an excellent if somewhat overdue full-scale biography (King, 2013).

Ricardo was also fortunate in having Piero Sraffa edit for the Royal Economic Society, over a period of 25 years, a magnificent and complete edition in 11 volumes of *The Works and Correspondence of David Ricardo* (1951–73). Although this collection revived interest in Ricardo's economic thought, oddly it seems to have contributed to a revisionist erosion of his analytic reputation. Whereas the late-Victorian Alfred Marshall, pillar of post-1870 neoclassicism, had idiosyncratically argued that the good modern stuff was *already* in Ricardo, and whereas Keynes had written in the 1920s that Ricardo had 'the finest mind that had come to economics' – neither Ricardo's late twentieth-century admirers nor detractors could agree among themselves as to what his different virtues and vices actually were. A scholar who is much misunderstood cannot be acquitted of all responsibility for ambiguity.

Reversing the usual order of intellectual progress, Ricardo first made his mark in what has come to be called *macroeconomics*. His unqualified verdict typifies his noneclectic style. To finance a very long war against Napoleon, Britain expanded its money supply of paper currency in order to procure goods at home and abroad for its armies. Price-level inflation was unsurprisingly part of the economic history in that period. The eclectic theory of Adam Smith would no doubt have predicted as well a worsening of terms of trade as Britain and other belligerents bid up the relative prices of needed imports. Had gold and silver been the sole money used then in Britain and elsewhere, the 1750 theories of David Hume would have predicted a drain of specie from London to pay for the war-induced trade deficit of exports compared to imports.

However, even in those distant times much of the money employed for purchase and sale transactions involved the use of paper money currency. As is customary in times of war, whatever the supply of domestic gold or silver, there ensued a considerable rotation of the printing press to finance government needs and accommodate the desire for the enhanced cash necessary to conduct the enlarged volume of goods transactions and their higher average prices. Under these circumstances, whenever a government chooses *not* to redeem paper currency notes into specie at their previous face value, those notes will depreciate in price relative to ounces of specie. The high price of gold is what the person in the street calls this phenomenon and there was considerable controversy as to its causes. Ricardo attributed *all* of the rise in gold prices to overissue of paper notes.

Today one can entertain counterfactual history. Somehow the British government might have taxed the populace more heavily to finance the enhanced war effort and maintain a balanced budget and the inescapable rise in relative import prices. As a general strategy the government could have imposed tightness in bank lending, enforceable only at higher interest rates, and conceivably could have kept the stock of paper currency unchanged. The outcomes would almost certainly have included higher unemployment and extra loss of consumption on the civilian populace. With heroic tightness, trade balances might even have been forced to cancel out to zero. It is hard to envisage how normal peacetime unemployment levels could have somehow been made to prevail from say 1795 to 1810 in this counterfactual scenario. An inflation-free war can thus be painfully conjured up in counterfactual history and with no wartime increment of paper currency issue.

In actual history, paper notes depreciated and price levels soared. As happened a century later, during the 1920–23, German hyperinflation, economists held opposing views of appropriate economic policies. Ricardo argued that overissue in currency M explained virtually *all* of the rise in the price of bullion. Henry Thornton, practical banker and Bank of England authority (at first), argued that the microeconomic phenomena of adverse trade balances and war-induced buoyant business activity explained a significant fraction of the observed appreciation of bullion. Ricardo's 1809–10 strong assertion of what was called the 'bullionist view' won him instant celebrity, particularly with the dogmatic James Mill. But this economic debate was not really a zero-sum game in which Ricardo's gain had to equal Thornton's loss. Both men agreed on the likely longest-run outlook. The price of gold in terms of paper would be the same in 1850 if, in the many preceding decades, the supply of paper currency was limited enough to permit free specie–currency convertibility at the old 1790 ratio. In this instance, the so-called 'Ricardian vice' was his use of *long-run* true relations to characterize actual short-run patterns in economic history.

After the first decade of the nineteenth century, Ricardo concentrated mostly on *microeconomics*. In 1815 (and *not* prior to then), along with Malthus and Edward West he helped to articulate the 'law of diminishing returns' and its consequences for land rent determination. Society's product, for example, corn, is produced by labor and by land, both working together. When you increase labor in two *equal* increments, working on fixed land, you acquire two successive increments of corn. But

the second gain is presumptively *less* than the first – a technological law of nature. Thus the wage that the competitive market can pay to labor goes down when labor/land density increases, and the rent that the market will pay for land will go up. This is not a new idea – it is in Benjamin Franklin (1755) early in the eighteenth century and already in Smith (1776) – but it is an important insight that needed to be made explicit.

Already Ricardo had perceived what he considered faults in Adam Smith. Posterity can be grateful that this perception tempted Ricardo into economics, even though it can be argued convincingly that Smith was mostly correct in the cases where Ricardo criticized him. For example, in the capital-and-land-and-produced-capital models that both writers worked with, Smith was right to jettison the 'labor theory of value' as soon as either the interest rate and/or the land rent rate becomes significantly positive.

As between the two tasks of political economy – (1) understanding how relative scarcities of factors of production constrain the alternative availabilities of different goods and their valuation relations; and (2) how society's harvest gets *distributed* among and between the different social classes of workers and owners of property – Ricardo arbitrarily declared it was distribution that was all important. Yet nowhere in the 11 volumes of his *Works and Correspondence* (1951–73) will the reader find a *complete* and satisfactory theory of distribution that went beyond what was already in Smith, Malthus, Mill, or modern mainstream writers.

Perhaps the lack of a university education contributed to Ricardo's methodological style. Frequently, he asserted general 'truths' that admit of important exceptions; repeatedly he failed to distinguish between *necessary* and *sufficient* conditions for an argument. He was a superb miniaturist but with an absence of coherence, particularly in his post-1815 textbook phase. Ricardo's theory of comparative advantage in international trade, similar to that of the contemporary Colonel Robert Torrens, is a brilliant example of his miniaturist virtuosity.

Perhaps with hindsight, this is a somewhat harsh judgment on 1810–23 David Ricardo, but one can be even harsher on Ricardo's twentieth-century commentators. The brightest and the best of them – Piero Sraffa, George Stigler, Samuel Hollander, Mark Blaug, and scores of others – while understanding that the positive interest and time-phasing of production generically invalidate Ricardo's championing of the *labor theory of value* (as, indeed, Ricardo himself intermittently admitted), at the same time they appear blind to the parallel fact that positive rent(s) on scarce land(s) must similarly invalidate the labor theory of value. Edwin Cannan, Knut Wicksell, Jacob Viner, and Lionel Robbins are exceptions to this general criticism about how labor/land intensities generically invalidate any labor theory of value in the same logical way that time intensities are known to do so.

Around 1815, between his macroeconomic works on gold and his *Principles of Political Economy and Taxation* (1817, 1819, 1821), Ricardo puzzled over a simple scenario involving a farm good (corn) and one or more manufactured goods (cloth and other products), all produced by, say, homogeneous labor, homogeneous land, and by one or more *produced* inputs (corn seed, raw materials from farming used for manufactures, and durable tools). He may even have worked out, in an unpublished manuscript that has not survived, a one-sector farm-only model. A determinate model of distribution could have cogently defined a long-run stationary-state equilibrium for (real corn wage rate, interest rate, real corn rent rate) = ($W/P_{\text{corn}} = \bar{w}, \bar{r}$, Rent$/P_{\text{corn}} = R^*$). How would it have gone?

Like Malthus and other predecessors, Ricardo would have specified *ex*ogenously a classical 'subsistence wage' rate, $\bar{w}$, needed to keep population just reproducing itself. Like his classical successors Ricardo would have specified *ex*ogenously a minimal interest rate, $\bar{r}$, needed to keep capital(s) just reproducing themselves. Land being a classical constant, $\bar{A}$, its stationary equilibrium rent rate he would *residually* determine once variable labor-and-capital stocks, L_t and K_t, dynamically converged to their long-run, asymptotic stationary equilibrium levels: $L_t \to L^*$; $K_t \to K^*$, when K is a single scalar. Finally, short-run determination of w_t and r_t would be market-clearing rates dependent on the relative transitional abundances of L_t and K_t.

Here is how in the spirit of Prokofiev's 'Classical' Symphony, an ambitious neophyte in a 1932 Hicks–Wicksell *neo*classical workshop might synthesize these classical insights into a coherent simplistic dynamic equation system, using the notational convention $dY_t/dt = \dot{Y}_t$:

$$\begin{aligned} \text{Factor supplies}: \dot{L}_t/L_t &= |\bar{\alpha}|\varepsilon(w_t - \bar{w}) \\ \dot{K}_t/K_t &= |\bar{\beta}|(r_t - \bar{r}), A_t \equiv \bar{A} = 1 \end{aligned} \qquad [1]$$

$$\begin{aligned} \text{Factor demands}: w_t &= \bar{\sigma}K_t^{1/3}/L^{2/3} \\ r_t &= \bar{\sigma}L_t^{1/3}/K_t^{2/3} \end{aligned} \qquad [2]$$

where $\bar{\sigma}$ is an exogenous parameter denoting level of technical productivity.

When the modern crank of deduction gets turned, Ricardo's [1] and [2] do lead convergently to a determinate classical stationary state. From any *initial* $(L_0\ K_0\ \bar{A})_0$, as $t \to \infty$,

$$\begin{aligned} &(w_t, r_t; L_t/\bar{A}, K_t/\bar{A}; R_t) \\ &\to \left(\bar{w}, \bar{r}; \frac{1}{3}\bar{\sigma}\Big/\left[\bar{w}^2\bar{r}\right], \frac{1}{3}\bar{\sigma}\Big/\left[\bar{r}^2\bar{w}\right]; \frac{1}{3}\bar{\sigma}[rw]\right). \end{aligned} \qquad [3]$$

One must commend the classical tradition for so subtle an accomplishment. The version worked out here differs from Paul Samuelson's 'canonical' model, which had more faithfully adhered to the conventional labor-capital 'dose' version (Samuelson, 1978).

This above supply–demand equilibrium is definitely *not* independent of the problem of how consumers wish to allocate their earned incomes among the many goods, when there are many noncorn goods in the market. Then there are no 'natural prices' definable independently of how consumers choose to spend their incomes.

Definitely goods' relative price ratios do not remain close to proportionality with their respective embodied labor contents, $P^i{}_t/P^j{}_t \neq (L^i{}_t/q^i{}_t)/(L^j/q^j{}_t)$. Likewise embodied land contents $(A^i{}_t/q^i{}_t)/(A^j{}_t/q^j{}_t)$ and embodied capital contents $(K^i{}_t/q^i{}_t)/(K^j{}_t/q^j{}_t)$ similarly fail as approximations to empirical actual P^i/P^j ratios.

Since Ricardo did not coherently understand the intricacies of his own classical scenario, he never realized how

unattainable was his hankering to have all relative prices determinable purely *by objective technology and cost data*. Subjective variabilities (1) of consumers' utilities and demand choices, and (2) of intertemporal saving-and-consuming choices – which became central preoccupations of post-1870 neoclassical thinkers to the classical school – were thus *already* unavoidable in the capital, land, and labor models of classical theory and in the competitive scenarios that Marx grappled with all his working life. This inevitability will come with pain for the small sect of Sraffian neo-Ricardians who proliferated after 1925 and especially 1960.

Ricardo was low keyed and civil in argumentation, albeit firm in holding to his strong positions. The letters between Ricardo and Malthus, notwithstanding their obvious differences, were a model of courtesy and affection. Despite Ricardo's typical unqualified dicta, he did not hesitate to change his mind and admit to previous mistakes. Marx, who abhorred classical authorities and for whom 'Parson' Malthus was a particular *bête noir*, was uncharacteristically soft toward Ricardo. Here is an illustrative incident. After the first edition of *Principles*, but prior to its third edition, Ricardo had declared that technical invention *must always* help *all* factors of production, including the working man. He encouraged disciples, like J.R. McCulloch, to affirm this strong view. However, on reconsideration, Ricardo came to realize that some inventions can assuredly *replace labor and reduce the demand for labor in the short and the long run*. Therefore, despite warnings from his friends that this position would weaken the case for *laissez-faire*, Ricardo added a new chapter on Machinery to the third (1820) edition of the *Principles*, presenting exposition and numerical examples designed to show how labor's share of national income could suffer from certain possible inventions and how long-run equilibrium populations and gross income levels could be permanently lowered. With virtually no exceptions, ancient- and twentieth-century authorities agreed that for once Homer had nodded. For once Ricardo must have invoked rigidity of wage rates with its resulting unemployment, thus for once denying J.-B. Say's sacred dogma that overproduction is impossible in a capitalist system. Karl Marx praised the scholar for his honesty.

It is a story that casts discredit on virtually all. The Ricardo–McCulloch original sweeping position was gratuitously wrong. A new wind that raises an economy's potential for production and consumption most certainly can do harm to *some* competitors while helping others. Ricardo did not claim, as J.S. Mill and Vilfredo Pareto were to argue subsequently, that gainers could always bribe losers so that *with intervention all could gain*. Nor did Ricardo's new understanding motivate him to abandon a dogmatic defense of *laissez-faire*. His critics on Machinery, both then and in our time – including Knut Wicksell, Nicholas Kaldor, Joseph Schumpeter, and George Stigler – failed to see the absence in Ricardo's exposition of any denial of dogmatic Say's Law and its ruling out theoretical harm from underconsumption. Instead his usual classical scenario of population decline dictated by the induced drop in the wage below an alleged equilibrium *subsistence* rate would, as the Machinery chapter alleged, call for an ultimate drop in population and gross product as a result of labor-saving invention(s). Just as the Industrial Revolution was beginning to raise wages by a genuine filter-down process, the armchair philosophers were getting their first intimations of the complexity of how a market system hands out its distributive awards.

In summary, David Ricardo, like the twentieth-century economists Friedrich Hayek and Milton Friedman, importantly pushed voters and public opinion toward libertarian *laissez-faire*. Victorian England's Whig society and Manchester School owed much to him. In the modern debate about how a modern mixed economy can optimally compromise between market mechanisms and democratic rules of conduct, Adam Smith and John Stuart Mill appear to be more subtle classical thinkers. From these two, autocratic rulers like Lenin, Stalin, and Mao could have had much to learn. But among the worldly philosophers, David Ricardo must be counted as an important influence on the formation of the twenty-first-century mind.

While Ricardo's status among classical economists is secure, he has not received sufficient attention from sociologists. Ricardo developed the basis of a theory of social classes before Karl Marx's class conflict theory and before Max Weber (1978) had developed his famous model of 'class, status, and power.' There were important social classes in Ricardo's work – the working class who depend on wages, the landowning class who extract rent, and the bourgeoisie who exploit the labor of the workers to create profit. Ricardo had been especially critical of 'Old Corruption,' namely the ability of the rent-seeking aristocracy to dominate Parliament in their own narrow interests.

In classical sociology, Marx argued that with the maturation of capitalist society there would be a class polarization and two dominant classes would appear – the proletariat and the bourgeoisie – as the driving force of history. Barrington Moore (1966) in *The Social Origins of Dictatorship and Democracy* provided a more complex comparative analysis more consist with Ricardo in which democracy required a thriving bourgeois class, while authoritarian regimes are the product of a weak bourgeoisie and an alliance between the peasantry and aristocratic landowning class such as the Junkers in Germany. He advanced the slogan – 'No bourgeoisie, no democracy.' Advanced capitalist societies are no longer so obviously stratified into social classes that exert collective political influence, and some sociologists argue that, following Ricardo, the technological revolutions of the twentieth century have brought about some de-skilling of the workforce leading to an erosion of the political clout of the working class (Braverman, 1974). Modern societies are instead characterized less by class conflict and more by growing income inequality and the hollowing out of the middle class (Krugman, 2003). These sociological debates indicate that Ricardo should be regarded as a founder of modern sociology as well as of classical political economy.

See also: Economic Growth: Measurement; Economics, History of; Innovation and Technological Change, Economics of; Libertarianism; Malthus, Thomas Robert (1766–1834); Market Structure and Performance; Marx, Karl (1818–83); Monetary Policy Since the 2007–2009 Financial Crisis; Political Economy, History of; Regulation, Economic Theory of; Smith, Adam (1723–90); Social Stratification; Sociological Theory; Stock Market Predictability.

Bibliography

Braverman, H., 1974. Labor and Monopoly Capitalism. The Degradation of Work in the Twentieth Century. Monthly Review Press, New York.

Franklin, B., 1755. Observations Concerning the Increase of Mankind and the Peopling of Countries.

Hirschman, A.O., 1991. The Rhetoric of Reaction. Perversity, Futility, Jeopardy. The Belknap Press of Harvard University Press, Cambridge, MA.

Keynes, J.M., 1936. The General Theory of Employment, Interest and Money. Reprinted in: 1977–81 the Collected Writings of John Maynard Keynes. Macmillan, London.

King, J.E., 2013. David Ricardo. PalgraveMacmillan, London.

Krugman, P., 2003. The Great Unravelling. Losing Our Way in the New Century. W.W. Norton, New York.

Malthus, T.R., 1798. An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet and Other Writers. J. Johnson, London [Reprinted by Macmillan, London, 1926].

Marshall, A., 1890–1920. Principles of Economics. Macmillan, London.

Marx, K., 1867, 1885, 1894. In: Capital Volumes I, II, and III. Verlag von Otto Meissner, Hamburg [Penguin Books, Harmondsworth, 1976, 1978, 1981].

Moore, B., 1966. The Social Origins of Dictatorship and Democracy. Lord and Peasant in the Making of the Modern World. Beacon Press, Boston.

Niehans, J., 1990. A History of Economic Theory. Classic Contributions, 1720–1980. The Johns Hopkins University Press, Baltimore.

Ricardo, D., 1817. Principles of Political Economy and Taxation. John Murray, London.

Ricardo, D., 1951–73. In: Sraffa, P., Dobb, M.H. (Eds.), The Works and Correspondence of David Ricardo, 9 vols. Cambridge University Press, Cambridge, UK.

Samuelson, P.A., 1969. The way of an economist. In: Samuelson, P.A. (Ed.), International Economic Relations: Proceedings of the International Economic Association. Macmillan, London, pp. 1–11.

Samuelson, P.A., 1978. The canonical classical model of political economy. Journal of Economic Literature 16, 1415–1434.

Samuelson, P.A., 1989. Ricardo was right! Scandinavian Journal of Economics 91, 47–62.

Schumpeter, J., 1954. History of Economic Analysis. Oxford University Press, New York.

Smith, A., 1776/1937. In: Cannan, E. (Ed.), An Inquiry into the Nature and Causes of the Wealth of Nations. The Modern Library, New York.

Sraffa, P., 1926. The laws of returns under competitive conditions. The Economic Journal 36, 535–550.

Sraffa, P., 1960. Production of Commodities by Means of Commodities. Cambridge University Press, Cambridge, UK.

Thornton, H., 1802/1962. An Enquiry into the Nature and Effects of the Paper Credit of Great Britain. Augustus Kelley, New York. Reprinted, Frank Cass, London.

Weber, M., 1928/1978. Economy and Society. University of California Press, Berkeley.

Ricoeur, Paul (1913–2005)

Jonathan Roberge, Institut National de la Recherche Scientifique, Québec, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Paul Ricoeur (1913–2005) was an eminent French philosopher working at the crossroads of many schools of thought (hermeneutics, phenomenology, analytical philosophy, etc.) and in constant dialogue with the human sciences. He taught mainly in France and the United States, and gained international recognition after the publication of *Oneself as Another* in the early 1990s. This book is emblematic of Ricoeur's broader work in the way it binds existence and interpretation through language, culture, and interactions with others. Ricoeur proposed a practical approach to what he called 'meaningful action,' which is open to different and often conflicting readings. For him, meanings are not ascribed once and for all, but must be constantly negotiated and experienced as such. This also holds true for politics and ethics, as Ricoeur reworked the ancient notion of *phronesis*.

Life and Career

Paul Ricoeur was born on 27 February 1913 in Valence, France. While the World War I was a difficult ordeal for everyone across Europe, it was a particularly troubling experience for the young Paul: his mother died few months after he was born, and his father died during the Battle of the Marne in September 1914. Orphaned and considered wards of the state, he and his sister were brought to Rennes to be raised by their paternal grandparents and an unmarried aunt. It was during the 1930s that Ricoeur discovered philosophy, first under the influence of Roland Dalbiez and then in a more robust study program in Rennes and Paris, where he encountered many important philosophical figures of that time, including Gabriel Marcel. In 1935, he married Simone Lejas, and they went on to have five children together. (Biographical details of Ricoeur's life and career are taken from the following sources in English and French: Dosse (2008), Greich (2001), Hahn (1994), and Kaplan (2008).)

Ricoeur was drafted into the French army in 1940; he went to the battlefield for a very short time before being captured. He then spent 5 long years in captivity during which his only consolation were his fellow prisoners, including a diverse array of philosophers: Mikel Dufrenne (with whom he later published a book on Karl Jaspers), Roger Ikor, Jacques Desbiez, etc. It was also in these rather uncomfortable circumstances that Ricoeur worked on the French translation of Husserl's *Ideen I*, which was published in 1950.

After the war, Ricoeur finished his dissertation, published some papers, and began his academic career, first at Strasbourg (1948–57), then the Sorbonne (1957–64), and later at Nanterre (1965–70). A respected professor, he was keen to engage, sometimes critically, with fellow scholars from other disciplines, especially the human sciences in general, and structuralism in particular. His debate with Claude Lévi-Strauss, for example, in the journal *Esprit* (1963) is symptomatic of this, even though Ricoeur had little chance of challenging the then extremely fashionable intellectual trend on *la rive gauche*; indeed, his criticism of structuralism garnered much of his own criticism, some of which more personal than theoretically grounded.

The late 1960s was yet another time of turmoil in the Western world, and in Ricoeur's life. The student protests of May 1968 were particularly intense in Nanterre, where Ricoeur acted briefly as dean and was incapable of restoring order on campus. He finally resigned from that position in March 1970, which also marked a period of relative withdrawal from French intellectual life.

Ricoeur's international career took off in the 1970s through different conferences and teaching assignments (at Louvain, Columbia, and Yale, for instance). But it was in the American Midwest that Ricoeur would primarily make his home, as he succeeded Paul Tillich as the John Nuveen professor of philosophical theology at the Divinity School of the University of Chicago, with a joint appointment in the philosophy department (1970–92). Such an environment proved to be extremely fruitful: both Hannah Arendt and Mircea Eliade also worked there at that time, and Ricoeur managed to write his highly influential books *The Rule of Metaphor* (1975) and the three-volume *Time and Narrative* (1983–85).

The next decade or so signaled a return to the French intellectual scene for Ricoeur, as well as his inclusion among the most preeminent philosophers of the century. While the publication in 1986 of *From Text to Action* can be considered an important step in this process, the most exposure came with *Oneself as Another* in 1990. At that point, Ricoeur had been working for more than 50 years, and this book is generally considered to be emblematic of his mature thought.

Reflecting his profound and personal dedication to philosophy, Ricoeur continued working into the 2000s, publishing several books including *Memory, History, Forgetting* (2000). He passed away on 20 May 2005. In his prolific career, he received over 30 honorary degrees, many of which from highly distinguished institutions such as Northwestern (1977), Göttingen (1987), and McGill (1992). He also received multiple awards such as the French Academy's Grand Prize for Philosophy (1991), the Balzan Prize (1999), and the John W. Kluge Prize in the Human Sciences (2004).

Ricoeur's Work as Dialogue and Translation

Such a prolific career as Ricoeur's inevitably raises the question of what constitutes its unity, or, similarly, what could be its *modus operandi*? As mentioned above, Ricoeur often engaged over the years with the human sciences and indeed, this 'step outside of philosophy' could be considered a guiding thread of his work (see Ricoeur, 1977; for instance; for commentary, see McGuire, 1981; and Moore, 1990). His is a thought rooted in dialogue, open to dialogue, striving for dialogue. By turns, Ricoeur referenced world events and engaged with other disciplines, intellectual currents, and traditions – be they from French, German, or Anglo-American universities. In each of his books, he developed what, following Jaspers, he called 'loving struggles' with different authors, aiming at better self-understanding.

In 2004, for instance, he published a small yet dense book called *On Translation*, which not only tackles important questions about the current state of globalization, but also retrospectively applies to his own academic journey (see Jervolino, 2008). In it, he asks fundamental questions: What does it mean to translate? What kind of possibilities does it open up? According to Ricoeur, the act of translating is one by which the translator accepts the uniqueness of the original text while simultaneously affirming that it can be otherwise communicated. The goal is to stay true to the original meaning; but even though the perfect translation does not exist, this does not mean that all efforts are made in vain. To translate is to engage in an act of meaning-making informed by a close and patient reading of something that precedes it. It is to accept both the capacities and the limitations inherent in the art of interpretation. Indeed, Ricoeur's hermeneutics never ceases to deal with this issue; in that sense, it is concurrently a method, a philosophical project, and an overall way to read the world.

This particular alliance between the individual, meaning, and interpretation runs exceptionally deep throughout Ricoeur's work. It is a conceptual articulation that informs his early philosophical anthropology and also proves central in his later ontological developments. Books such as *The Voluntary and the Involuntary* (1950) and *Fallible Man* (1960), for instance, explore these themes, especially the constitutive disproportion of the human experience and the opacity of the cogito. Our embodied lives and the fact that the world always precedes us generate a finite situation with an infinite task: the desire to express and understand ourselves. "The ultimate root of our problem," notes Ricoeur, "lies in this primitive connection between the act of existing and the signs we deploy in our work; reflection must become interpretation because I cannot grasp the act of existing except in the signs scattered in the world" (1965/1970: 46). But the problem is the solution, and vice versa: meaning and culture are both a long and complicated detour around being, as well as the only resource for a cogito that cannot have direct access to itself, *à la* Descartes.

The question of what constitutes a symbol arises from this delicate connection between existence and interpretation, which Ricoeur explores in depth in the *Symbolism of Evil* (1960). In this work, he looks at different backgrounds and cultures, all of which have tried to come to terms with evil by devising myths and stories to make sense of both evil's origin and its end. Symbols – Eve, the snake, and the apple in the Bible, for instance – operate at the junction between two types of meaning: the literal and mundane on one side, and the figurative and spiritual on the other. There is thus a double sense to all symbols; they are in a constant state of polysemy. But for Ricoeur, they 'give rise to thought' precisely for this reason, as they propose a 'surplus of meaning' that has to be decoded and appropriated by individuals.

Gradually, however, it is the notion of text that would become the central feature of Ricoeur's hermeneutical approach. The two main forms of language, speech and text, are fundamentally different, as the latter introduces a significant distance into human communication. A text has an internal logic that creates an inner structure, which in turn allows for a certain closure or autonomy. This can also be thought of in terms of semiotics, and Ricoeur would use this concept to engage with the structuralists of his time, Lévi-Strauss most prominent among them. But a text is more than just its own architecture of signs; it wants to express something, which makes it a semantic entity unto itself. It refers to a projected world and thus to "understand a text is to follow its movement from sense to reference, […] from what it says, to what it talks about" (1976: 87–88). The world of the text is our world, i.e., it deals with human experience in such a fashion that it creates something like a *Lebenswelt* – a living world that the reader is able to relate to.

Ricoeur's pivotal notion of text thus connects semiotics and semantics, explanation and understanding, more than it separates them. For him, as well, textual exegesis must be rigorous; there is a validation procedure that should be applied, such that "it is not true that all interpretations are equal [… as the text] is a limited field of possible constructions" (1986/1991: 160). This holds true for the text's reference, if one accepts that it deals with the social field of human action. In a seminal essay entitled "The Model of the Text: Meaningful Action Considered as a Text," Ricoeur develops four analogies that militates for the objectification of the practical field. First, while meaning surpasses its occurrence, what constitutes the significance of an act goes beyond the simple fact that it happened. Second, the propositional content of an act or a text, as much as its illocutionary force, points toward a "logical status as having such-and such identifiable meaning or 'sense-content'" (1981: 205) – saying 'I do' at a wedding, for example. Third, and because of what precedes, the meaning of an act can detach itself from the will of its agents and produce its own outcomes. And last but not least, actions or texts can be interpreted by a wide range of readers, present and future, who propose different interpretations and argue for their relative merits, like the social sciences would do, for instance.

Over the years, and especially in the 1980s and 1990s, this hermeneutics of text and action produced a remnant, a questioning that for Ricoeur had to be transformed or translated. He was also growing unsatisfied with his treatment of the self and its capacities. In books like *Time and Narrative* and *Oneself as Another*, he explored the fragile human condition by focusing on two themes: identity and temporality. What is it to maintain the self through constant change and evolution, or even the simple fact that people age? In asking how to answer the question 'who?' Ricoeur's response is to build on a distinction

in Latin between *idem* and *ipse*. To stay the course, to persist unchanged in one character is to promote a sort of tautology, an *idem*-same, or self-same. On the contrary, through *ipseity* one would accept and welcome changes, understanding the unity of his life in analogous terms and thus the self as *comprised* of its different actions. For Ricoeur, in other words, what is fundamental for the self is to be *like* the author of his own life and to be recognizable as such.

The development of such a hermeneutics of the self necessarily involved the use of language, and Ricoeur insisted on the narrative construction of the self and thus developed the concept of 'narrative identity.' Meaning-making goes hand in hand with storytelling, and through this process a person becomes a persona, while his life is both a work in progress and an emplotment. "It is the identity of the story," notes Ricoeur, "that makes the identity of the character" (1990/1992: 148). And this, in turn, also implicates other people and other personae: the language of the self is a form of dialogue insofar as action is always a form of *interaction*. Ricoeur's interpretation is rather innovative as it allows for connections with both the social sciences and ethics. The key here is 'recognition,' which becomes more obvious in later books such as *The Course of Recognition* (2005), and emerges as mutuality, social esteem or even as a promise. For example, "I promise that in the future I will respect what I say now; I will be true to my words, to myself, and—especially—to you."

It is by way of analogy and translation, once again, that this relationship between alter and ego allowed Ricoeur to speak more generally about history, society, culture, and politics. From two individuals, to three and beyond, up to the creation of institutions, there is both continuity and rupture. The advent of institutions – what Husserl called 'personality of the higher order' and Weber 'groups of domination' – signals in effect a certain crystallization of human action, the creation of an order or a structure. These exist to achieve goals and make difficult collective decisions, but precisely because of this function, they can also introduce forms of coercion. Therein lies their fragility, or what Ricoeur in 1957 named the 'paradox of politics,' in reference to the events that had recently occurred in Prague. Simply put, rulers demand more authority and control than the ruled are willing to give them.

If the use of force has proven to be constant throughout history, the same can be said about the many justifications made by political entrepreneurs, in terms of rhetoric and ideology. Broadly speaking, this is where politics meets interpretation and culture: "Ideology is a function of the distance that separates the social meaning from the inaugural event which must nevertheless be repeated" (1981: 225). In other words, no group can reproduce its identity without narrating its origins, and yet this ongoing narrative process hides as much as it reveals, as ideology inherently introduces ambivalence and distortion.

For Ricoeur, the fundamental question with regards to politics is not truth per se, but rather what constitutes legitimacy. Politics is rooted in language and triggered by conflict: conflicts of interpretation, conflicts of power and interests, etc. The question, time and again, is what *should* people do? What is acceptable? What kind of agreement can be reached? What can be considered fair treatment? Ricoeur held that democracy is fundamentally informed by these equivocal questions. Ambiguity, in other words, is the necessary condition for deliberation and participation. To engage and to argue is an endless task to be sure, but it is this sort of collective work that shapes and maintains healthy citizenship. Ricoeur's analysis, through this latest question, begins to demonstrate some of its ethical importance. For instance, in response to the ultimate question of what constitutes justice, he would plead for less inequality and more overall social inclusion. Justice, then, is the horizon. In *Oneself as Another*, he offers a succinct expression that summarizes his thought, stating the goal is to be "with and for others in just institutions" (1990/1992: 262).

Justice is to politics what the 'good life' is to ethics. For Ricoeur, these two *telos* or aims are intimately linked. They share a common undecidability and, because of this, must be experienced in a practical way, i.e., through action and language. Following Aristotle's notion of *phronesis*, Ricoeur insists on the many differences between such practical wisdom and what could be considered a more robust understanding of morality built from the reinforcement of norms and obligations. From this, it follows that the road toward the good life has flexible guidelines, among which three are of greater interest. First, according to Ricoeur, the idea that the other shall be respected both as capable and fragile – the theme of the 'Fallible Man' is present here – informs a solicitude for which the other is an end in itself and never a means. Second, as I recognize the uniqueness of other individuals, I also recognize my own uniqueness. Consequently, I myself am an agent and shall be held accountable for my actions. Third, as the self tries to be respectful and responsible, he finds himself immersed in a never-ending process of reflection, such that the good life is and remains only an interpretation. For Ricoeur, however, this is not a flaw, but an extraordinary resource, something that helps us make sense of the most difficult questions. With the help of reflection, doubts are introduced that both allow us to reassess and reaffirm our convictions.

Ricoeur's Legacy

2013 would have marked Ricoeur's 100th birthday, and in his honor several conferences were organized and commemorative publications were produced. Together, these efforts have reinforced Ricoeur's position as a leading figure in contemporary philosophy, a continuation of the accolades and respect he received throughout his career. Even more significant in this regard, however, is the creation in 2010 of the *Fonds Ricoeur* in Paris, a project that acts as a focal point, a *passage obligé*, for all scholars who want to study his work in depth. All of Ricoeur's archives can be found there, along with his personal library. The devoted group who maintains the *Fonds* makes sure that everything published on the philosopher is properly acknowledged; they are also aiming to build an international scholarly network that includes the Society for Ricoeur Studies and a Web journal entitled *Etudes Ricoeuriennes/Ricoeur Studies*. Thus, it is quite possible today to say that the work has not only survived the death of its author, but that it has acquired a life of its own. Its finitude continues to open plural interpretations and, in that sense, the work still fundamentally concerns dialogue and translation.

Another and perhaps broader way to assess Ricoeur's legacy is by examining the scope of his work. Is it vast enough to

be more than just a collection of disparate moments and intellectual topics? Does it point toward a form of universal hermeneutics or is it more local and specific? To what extent, in short, is it systematic? These questions are perhaps best addressed by measuring the distance separating Ricoeur's work from Hegel's dialectic, which strives toward synthesis through Reason. 'Renouncing Hegel,' in fact, is the title of an important chapter of *Time and Narrative*, in which Ricoeur is insistent in his critique of Idealism; while meanings are complicated and plural, they are more ambiguous than logical and rational, as they are for Hegel. This also implies that it is always difficult to resolve practical or social conflict and, ultimately, that political powers, especially states, cannot be understood as a source of more than justification. The links between social and political conflicts and conflicts of interpretation are, to be sure, both constant and robust in Ricoeur's work; the dialogue among them is essentially what replaces Hegel's dialectic. Furthermore, this allows for a connection between the universal and specific aspects of hermeneutics. It is not mandatory to choose one over the other, for both aspects cohabitate and find a rare and interesting balance in Ricoeur's work.

Similarly, such a balance can be seen in Ricoeur's negotiation between philosophy and the human sciences. This conversation is at the same time peripheral and central; and again, there is no need to choose one over the other. Ricoeur engaged with the human sciences beginning in the 1960s and continued to do so up until the very end of his career, receiving many insights and offering many of his own, as well. Indeed, it is possible to argue that his work is a philosophy *of* the human sciences, or *in* the human sciences; his work can vitally inform their epistemological foundations, the realities they face, and their ethical horizons. And this explains why Ricoeur's dialogue with the human sciences is so important. Who makes this kind of effort nowadays? Who is trying to make such important connections among these disciplines? The present era is more geared toward specialization, even hyper-specialization, and because of funding requirements and institutional retrenchment, philosophy and the human sciences have tended to drift further apart. Ricoeur's most important legacy, then, is the crucial reminder that it is still possible to do otherwise.

See also: Hermeneutics; Identity and Identification; Justice: Philosophical Aspects; Phenomenology: Philosophical Aspects; Semiotics; Social Science, The Idea of.

Bibliography

Dosse, François, 2008. Paul Ricoeur. Le sens d'une vie (1913–2005). La Découverte, Paris.
Greich, Jean, 2001. Paul Ricoeur. L'itinérance du sens. Jerome Millon, Paris.
Hahn, Lewis E. (Ed.), 1994. The Philosophy of Paul Ricoeur. Open Court, Chicago/La Salle.
Jervolino, Dominico, 2008. Rethinking Ricoeur: the unity of his work and the paradigm of translation. In: Kaplan, David M. (Ed.), Reading Ricoeur. State University of New York, Albany, pp. 225–236.
Kaplan, David M. (Ed.), 2008. Reading Ricoeur. State University of New York, Albany.
McGuire, Steven J., 1981. Interpretative sociology and Paul Ricoeur. Human Studies 4 (2), 179–200.
Moore, Henrietta, 1990. Paul Ricoeur: action, meaning and text. In: Tilley, C. (Ed.), Reading Material Culture. Structuralism, Hermeneutics and Post-Structuralism. Basil Blackwell, Oxford, pp. 85–120.
Ricoeur, Paul, 1950(1966). The Voluntary and the Involuntary. Northwestern University Press, Evanston.
Ricoeur, Paul, 1957. Le paradoxe politique. Esprit 260–285.
Ricoeur, Paul, 1960(1965). Fallible Man. Henry Regnery, Chicago.
Ricoeur, Paul, 1960(1967). The Symbol of Evil. Harper Row, New York.
Ricoeur, Paul, 1965(1970). Freud and Philosophy: An Essay on Interpretation. Yale University Press, New Haven.
Ricoeur, Paul, 1973. The model of text: meaningful action considered as a text. New Literary History 5 (1), 91–117.
Ricoeur, Paul, 1975(1978). The Rule of Metaphor: Multi-disciplinary Studies in the Creation of Meaning in Language. Routledge & Kegan Paul, London.
Ricoeur, Paul, 1976. Interpretation Theory: Discourse and the Surplus of Meaning. Christian University Press, Forth Worth.
Ricoeur, Paul, 1977. Phenomenology and the social science. The Annals of Phenomenological Sociology 2, 145–159.
Ricoeur, Paul, 1981. Hermeneutics and the Human Science. Cambridge University Press, Cambridge.
Ricoeur, Paul, 1983–85(1984–88). Time and Narrative, 3 vols. Chicago University Press, Chicago.
Ricoeur, Paul, 1986(1991). From Text to Action: Essays in Hermeneutics II. Northwestern University Press, Evanston.
Ricoeur, Paul, 1990(1992). Oneself as Another. University of Chicago Press, Chicago.
Ricoeur, Paul, 2000(2009). Memory, History, Forgetting. University of Chicago Press, Chicago.
Ricoeur, Paul, 2004(2006). On Translation. Routledge, London.
Ricoeur, Paul, 2004(2005). The Course of Recognition. Harvard University Press, Cambridge.

Riesman, David (1909–2002)

Christian Fleck, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

David Riesman became famous because of his 1950 bestseller *The Lonely Crowd*. Before that he received his education at Harvard and was a professor of law at Buffalo and of social sciences in the College of the University of Chicago. Later, he returned to Harvard. Besides his devotion to teaching, he acted as a public intellectual during his lifetime. In the *Lonely Crowd* study, he proposed the change of social character from tradition-directed, to inner-directed, and then to other-directed. This heavily value-laden concept helped make the book one of the rare examples of a social science bestseller. Whereas an inner-directed person follows the instruction one picked up early in one's life, the other-directed person orients his behavior according to the cues he gleans from his peers and the prevailing mass culture. Riesman later on abandoned his own characterology but he continued to be concerned about changes in society and politics. His less-known publications on higher education demonstrate his familiarity with the world of undergraduate education in the United States and show his incorruptible personality vis-à-vis cultural pressures.

Like many others from his generation, David Riesman chose sociology because he thought the knowledge developed in this discipline might help not only to understand the present time but also to change it. More than many others, Riesman enjoyed real influence in his time and much beyond his own sociocultural environment. Looking at him and his *oeuvre* from a present viewpoint, even an unprejudiced observer would not file him and his writings under 'sociology'; at least they do not fit into the contours of that discipline as it is today. The sociology Riesman discovered in the aftermath of the World War II was quite different. Riesman questioned whether this obvious change was to the better of the discipline. After Riesman's death, one of his numerous friends, collaborators, and admirers published an op-ed essay on him and his legacy in the *New York Times*. He titled it 'The last sociologist' (Patterson, 2002). True or not, Riesman was a sociologist of a particular shape.

Apprenticeship Years

Riesman was born in Philadelphia in 1909 into a wealthy family of German-Jewish origin. His father was a professor of medicine at the University of Pennsylvania and his mother a college graduate with a broad spectrum of interests. Their son, David Jr., showed similarly broad interests as early as in high school. He thought first to follow his father's path and become a doctor too. At college, he majored therefore in biochemistry, but the college was Harvard with its wide spectrum of digressions, several of which Riesman took. He became an editor of the student newspaper, *The Harvard Crimson*, where he collaborated with, besides many others, the later Marxist scholar and economist Paul Sweezy (Riesman, 1990: 29, 45). Riesman later wrote that the *Crimson* was his 'major educative enterprise' and that the curriculum bored him (Riesman, 1953). In one of his autobiographical essays, he admitted that at this time, he accepted his mother's definition of him as an 'uncreative,' so his next decision might have been seen by his mother as a proof of her opinion (Riesman, 1988: 3). He graduated in 1931 and at this time his interests had changed to a complete different field, law. He chose the Law School 'aimless' but with a clear agenda to remain near his new found mentor, Carl J. Friedrich, whose friendship lasted for the rest of their lives. Friedrich, 9 years older than Riesman, was a quite recent immigrant from Germany, and held a position in the Department of Government. Riesman admired him both for his erudition and encouragement and later claimed that he "owe[d his] introduction to the social sciences to Friedrich" (Riesman, 1988). Riesman must have possessed a particular brightness that his mother might have overlooked because in his second year as a graduate student he was recruited to the group of editors of the *Harvard Law Review*. For more than a decade, Riesman pursued law. After graduation, the eminent member of the Law School and future Supreme Court Justice, Felix Frankfurter, tried to channel Riesman into academic work, something the young man resisted for a while. Instead, he took up the private practice of law. In between Riesman served 1 year as a law clerk to U.S. Supreme Court Justice Louis D. Brandeis. The young man started an argument during the hiring conversations by calling Zionism a Jewish fascism, probably because Brandeis's sternness reminded him of his father's one (Horowitz, 2010: 1008). Riesman did not like the year as a helping hand on the Supreme Court nor did he accept offers he received to join one of the New Deal institutions. In 1937, he got an offer from the University of Buffalo to teach at its Law School. Looking at Riesman's list of publications, the job offer must have been based on other achievements than written ones.

Riesman had recently married Evelyn Thompson and the young couple moved to upstate New York where they remained for 4 years. For the academic year 1941–42, Riesman moved down the Hudson to spend time as a visiting fellow at Columbia University's Law School. Even before this occupationally related move, Riesman joined another broad movement, the one that brought educated and well-to-do people to the couches of psychoanalysts. In Riesman's case, it came at the 'urging' of his mother, who was a patient of Karen Horney whereas he had been commuting since 1939 to New York to see Erich Fromm (Riesman, 1988: 7). Whatever positive functions psychoanalysis might have had for the young professor, it seems that one shortcoming never burdened him: shyness

(Riesman laments his juvenile awkwardness and slow reading as hurdles, in Riesman, 1990: 24, Riesman, 1988: passim). During all the years as a young aspiring man he made the acquaintance of many people, famous and ordinary. In Manhattan, he rubbed shoulders with what became then known as the New York intellectuals and this climate was much more to his liking than the 'abstract and unempirical' world both of Friedrich's political science and the lawyers' environments at Harvard, Buffalo or Washington, DC. During his early stays in the metropolis he got in touch with Robert and Helen Lynd, Margaret Mead, and Ruth Benedict and joined the American Association for Public Opinion Research. The entry of the United States into the World War II forced the University of Buffalo, then a private institution, to close its law school and Professor Riesman had to see where he should continue his trade. He first served for 1 year as an assistant district attorney in New York and worked for the next 3 years as an executive at the Sperry Gyroscope Company on Long Island, a middle-sized war contractor firm (Anon., 2002). An article he wrote during his fellowship year at Columbia and came out in print in 1942 (Riesman, 1942) attracted the attention of Edward Shils who managed to invite Riesman for a job interview to Chicago. There Riesman started his sociological journey in 1946.

For a long period of time, Riesman was affiliated with the undergraduate College of the University of Chicago, not with any of the university's graduate departments. He started as a visiting assistant professor and became a tenured full professor the following year (Riesman, 1988: 8). The College consisted of regular faculty and Riesman belonged to the social science staff, hired primarily to teach an extensive course on social sciences for undergraduate students (Riesman, 1990). Riesman obviously enjoyed teaching and his contribution to the series devoted to autobiographies in the *Annual Review of Sociology* is dedicated mostly to his teaching experiences. A great many students came under Riesman's spell during his decades teaching social science, but his sociological writings made him even more well-known (Gans et al., 1979). One book in particular stands out from all the rest: *The Lonely Crowd*.

A Sociological Bestseller

In 1947, the interdisciplinary Committee on National Policy at Yale University, where Harold D. Lasswell pulled the strings, invited Riesman for a 1-year stay there to do some research on anything, which fell into the frame of the inviting committee. Riesman happily accepted the offer. He was able to persuade his superiors at Chicago that he would come back and found a perfect substitute to teach 'Soc2' there: C. Wright Mills. At Yale Riesman had enough funds to hire a research assistant and managed to persuade Nathan Glazer to fill this post part-time by commuting between New York and New Haven.

The book appeared in print in 1950, after less than 3 years, and has been reprinted several times, and translated into all major languages (Riesman et al., 1950, [1950]1969). The story of how *Lonely Crowd* came into being has been told by Riesman and Glazer independently and needs not to be repeated here in detail (Glazer, 2012). Instead, a kind of analytic appreciation of the book should help to grasp its nearly unique place in the postwar-social science scenery.

Looking purely at the text, one might have troubles to identify the discipline to which it belongs. The then usual form of presenting sociological research results was not observed by its author. If one would put it into the sociology basket, it would fit better with older examples, probably best be located near the two *Middletown* volumes by Robert and Helen Lynd. One is not very surprised to read in one of Riesman's memoirs that he would have been most happy, if the Lynds had invited him for another community study endeavor.

The Lonely Crowd most closely resembles books written by anthropologists in the course of World War II effort to understand the social and cultural side of the enemy nations, Japan and Germany. People with whom Riesman had made professional acquaintance the years before he started at Chicago, such as Margaret Mead and Ruth Benedict, authored books to whom Riesman paid tribute with *The Lonely Crowd's* subtitle: *A study of the changing American character*. Character was then a widely used concept, in most cases defined more closely by prefacing the term with the word 'social.' 'Social character' meant typical patterns of behavior and belief that particular societies force upon or elicit from their members. Moreover, not only outside observers but even particularly well-prepared members of a given society were thought to be able to perceive the social character of that society with some exactitude. Riesman obviously did not have any reservations with regard to any of his own prepossessions. Nowhere in the book's first edition, do readers get any hint of relativism or 'perspectivism,' as it was called by Karl Mannheim, an author Riesman obviously did not hold in high esteem, probably because of Mannheim's relativism.

The author's self-assurance is demonstrated in a then unusual usage of the first person singular. Very often Riesman speaks of himself, the author, by saying 'I,' starting as early as at the second page: "as I do throughout this book," "I do not plan to delay …," "most of these writer assume as I do – […] most of them agree – as I do," and finally "my collaborators and I base ourselves on this broad platform of agreement, and do not plan to discuss in what way these writers differ from each other and we from them" (1950: 4). Some 160 pages later, Riesman comments on the usage of 'I' by people classified as inner-directed ones. Summarizing a field exploration by some of his friends, conducted in rural Vermont, he has this to say about first person singular speech: There the old people "express feelings of responsibility for politics […] In referring to events they use the pronoun 'I': 'I' think, 'I' want, 'I' hate, and so on. They talk as if it were up to them to judge what happens in politics and, to the limits of their gifts and available energies, to guide it" (1950: 169). Very much the same could be said about the author David Riesman, who felt himself entitled to speak out about the failures of his society and politics in his country.

Besides the two classics from anthropology already mentioned, Riesman names Erich Fromm and Karen Horney, his and his mother's psychotherapists, Abram Kardiner, and Geoffrey Gorer as belonging to the orbit *Lonely Crowd* wanted to join. The social character all these authors tried to unravel was seen as one formed by parents, a relatively stable set of behaviors and beliefs within a particular society. Whereas Fromm and Horney, quite recent immigrants into the United

States, proposed to analyze modern societies as such, Riesman wanted to be one degree more down to earth and concentrated his analysis much more on his own society, that is, America. This alone put *Lonely Crowd* on a pedestal of its own and most probably contributed to its wide recognition and bestseller status (Gans, 1997).

In the 1940s and 1950s, personality and culture, personality and society, national character, and many more similar sounding publications and study programs flooded bookstores and classrooms. So, Riesman could expect that his book would be recognized in academia at least but due to his mastery of language and probably with the help of both his wife Evelyn Thompson Riesman, an experienced editor, and in particular Reuel Denney, a poet, he must have had ambitions going beyond the restricted readership of colleagues. And he succeeded. Within just a couple of month, David Riesman became a nationally recognized person. *Time Magazine* put him on its cover on 27 September 1954, announcing that 'social scientist David Riesman' will answer the question, "What is the American character?" The painting consists of a portrait of Riesman and two elderly men floating above his shoulders wearing satellite dishes as backpacks, probably an illustration of the 'other-directed character,' which occupied the center of the book.

The 300-pages-long study consists of three parts: First, Riesman outlines what he called his 'theory of character.' Second, he embeds the character types into American politics, and third, he ends with a plea for overcoming the restrictions of the character types. Crucial for the reception and the success of *Lonely Crowd* was the wording Riesman applied to the three character types outlined at length in the first part of the book: 'Tradition-directed,' 'inner-directed,' and 'other-directed.' One could be sure that the book would have got a different reception, if the author would have called his three types: T, I, and O, or 1, 2, 3. Only because of the surplus meaning in particular of 'inner' and 'other' Riesman has been able to gain the status of a diagnostician of the present.

The three types coincide with the trajectory of all societies: During the long period of static demographics – high birth and death rates balance each other – individuals are directed by tradition. Only when dynamic forces put a society on an expansive path, what later on W.W. Rostow would call the take-off, a new type of personality has to be formed: the inner-directed. Riesman locates this 'transition' to the period of Renaissance and Reformation, which means that the inner-oriented type's first appearance must have happened somewhere in Europe (more than once Riesman parallels his inner-directed type with those guided by Max Weber's Protestant Ethic). Putting all differences between Catholics and Protestants, Renaissance and Reformation, and so on, to the side, the core of the inner-directed character has to be seen in the fact that "the source of direction for the individual is 'inner' in the sense that it is implanted early in life by the elders and directed toward generalized but nonetheless inescapably destined goals" (Riesman et al., 1950: 15).

Whereas all Western civilizations were driven by inner-directed inhabitants, only 'in very recent years,' in particular parts of American society and in a handful of areas, did there emerge a new character type: "in the upper middle class of our larger cities: more prominently in New York than in Boston, in Los Angeles than in Spokane, in Cincinnati than in Chillicothe" characters were formed who, astonishingly, look 'strikingly similar' to the ones seen by the French traveler to America, Alexis Tocqueville, a century earlier. Nevertheless Riesman argues that the other-directed type 'does find itself most at home in America' and adds as the defining streaks: "What is common to all the other-directed people is that their contemporaries are the source of direction for the individual – either those known to him or those with whom he is indirectly acquainted, through friends and through the mass media" (Riesman et al., 1950: 20–21).

The contrast between a person who follows his father's instruction and the person who looks and listens primarily to his peers fills at least half of *Lonely Crowd's* pages. But only because the first was named inner-directed, and the second other-directed did readers conclude, probably correctly, that the author prefers the older over the newer type. By focusing on the then relatively new gadgets of modern life – movies, radio, yellow press media, abundance, consumerism, and personal improvement techniques – Riesman insinuated to his audience that the present might be worse than the recent past but never argued unequivocally against the present.

Lonely Crowd functioned like a Rorschach test: Every reader could come across what she wanted to read to better understand her rapidly changing environment. And indeed the US society and economy of the late 1940s was on the tip of another take-off, complete with rising earnings, increasing mobility, expanding suburbia, college education for more people, and, of course, the great baby boom. In a word, what later sociologists called Fordism, an era of full employment, mass society, mass media, and all the rest that cried out for analysis, indeed for understanding in the midst of bewildering change. Riesman, who wanted to understand himself and his contemporaries, offered Americans a portrait of themselves, very similar to the portrait *Middletown* drew two decades earlier or the portrait drawn by the huge and constant flood of diagnostics that became commonplace after *Lonely Crowd* smoothed the way.

It seems to be appropriate to file *Lonely Crowd* in the 'diagnosis of our time' genre. This type of writing assembles between two covers whatever its author has been able to collect as proof for his thesis – and this is the one big problem with this kind of reasoning. Methodologically, they are failures from the very beginning because it is always possible to find confirming instances for any thesis one wants to propose.

An Academic Life after the Bestseller

Very much different to other diagnostic writers, Riesman was open to criticism and dismissed his typology within a short period of time. In prefaces to reprints of *Lonely Crowd* and other replies to critics (see for a collection of critical comments: Lipset and Lowenthal, 1961), Riesman was willing to remove the backbone of his diagnosis. What he did not countermand was his interest in understanding the forces behind recent social changes. Since some of the criticism was directed toward the opaque methodology of *Lonely Crowd*, he and Glazer edited a second volume, *Faces in the Crowd: Individual Studies in*

Character and Politics (Riesman and Glazer, 1952) where one can find some of the empirical material corroborating the first book. Over the next decades, Riesman continued to write about his country and society, favored individualism and remained a scholar with his own style, or habitus. By bringing two concepts together one could call him a *citoyen* social researcher, which means that Riesman embodied quintessentially the concerned citizen and a scholar who preferred empiricism over abstract philosophical reasoning. What makes Riesman different to nearly all the others who want to contribute something grand intellectually seems to be his prioritizing of key questions to be raised and answered empirically, instead of clinging to abstract concepts one imposes on the world. Or, in Riesman's own words: "thinking countercyclically" (1988: 23).

Within a short period of time Riesman found another métier for his curiosity: higher education. In 1955, Paul Lazarsfeld asked Riesman whether he would be interested to do a kind of follow-up study of a quite recently finished survey, which became *The Academic Mind* (Lazarsfeld and Thielens, [1958] 1977). In this study, Lazarsfeld examined the side effects of the hysterical mood caused by the anti-Communist crusade, usually called by the name of its fiercest spokesman, Senator Joseph McCarthy. The former law professor and now best-known public intellectual David Riesman, with several publications on civil liberties on his *curriculum vitae* seemed to be the best choice to contribute a separate investigation of the validity of the questionnaire as an instrument. Riesman accepted and made himself familiar with the diversity of higher education institutions in the United States by visiting more than four dozen of them personally (Riesman, 1979).

When his close friend Everett C. Hughes lost the battle over the University of Chicago's sociology department's chairmanship, Riesman accepted an invitation by his old university and returned to Harvard, where he continued to give undergraduate courses under the title 'American character and social structure.' During the next couple of years Riesman published several books on higher education, mostly coauthored with junior faculty (best known is probably Jencks and Riesman *The Academic Revolution* from 1969), and continued to write magazine articles and commentaries for a larger audience (see Riesman, 1954, 1964). The student movement of the 1960s did not meet with his approval and he did not refrain from confronting the rebels with his arguments. Since Riesman did not scorn popular culture – in *Lonely Crowd* he even quotes children rhymes – he might have gotten some satisfaction when Bob Dylan paid tribute to the title of his bestseller: "Standing next to me in this lonely crowd/Is a man who swears he's not to blame."

In the social sciences Riesman's name can still be remembered by insiders but from his huge *oeuvre* only one title survived. Orlando Patterson was not right to call him the last sociologist (Patterson, 2002), but he was right insofar as scholars like Riesman are now available only in the history books of the social and behavioral sciences – most probably not to their advantage.

See also: Benedict, Ruth (1887–1948); Communication Research and Media Studies, History of; Conflict: Education and Youth; Conformity: Sociological Aspects; Consumption, History of; Fromm, Erich (1900–80); Higher Education Management; Higher Education Market; Higher Education Research; Lazarsfeld, Paul Felix (1901–76); Leisure and Cultural Consumption: US Perspective; Mass Society: History of the Concept; Mead, Margaret (1901–78); Mills, Charles Wright (1916–62); National Character, History of; Professions, Sociology of; Psychoanalysis in Sociology; Psychoanalysis, History of; Social Structure: History of the Concept; Sociology, History of; Tocqueville, Alexis de (1805–59).

Bibliography

Anon., 13 May 2002. David Riesman obituary. The Telegraph. http://www.telegraph.co.uk/news/obituaries/1394001/David-Riesman.html.

Gans, H.J., 1997. Best sellers by sociologists: an exploratory study. Contemporary Sociology 26 (2), 131–135.

Gans, H.J., Glazer, N., Gusfield, J.R., Jencks, C. (Eds.), 1979. On the Making of Americans: Essays in Honor of David Riesman. University of Pennsylvania Press, Philadelphia.

Glazer, N., 2012. My life in sociology. Annual Review of Sociology 38, 1–16.

Horowitz, D., 2010. David Riesman: from law to social criticism. Buffalo Law Review 58, 1005–1029.

Lazarsfeld, P.F., Thielens Jr., W., [1958]1977. The Academic Mind: Social Scientists in a Time of Crisis. Arno Press, New York.

Lipset, S.M., Lowenthal, L. (Eds.), 1961. Culture and Social Character: The Work of David Riesman Reviewed. Free Press of Glencoe, New York.

Patterson, O., 2002. The last sociologist. New York Times 19. May 2002, correction: 24 May.

Riesman, D., 1942. Civil liberties in a period of transition. In: Friedrich, C.J., Mason, E.S. (Eds.), Public Policy: A Yearbook of the Graduate School of Public Administration, Harvard University 1941. Wiley, New York, pp. 33–96.

Riesman, D., Glazer, N., Denney, R., 1950. The Lonely Crowd: A Study of the Changing American Character. Yale University Press, New Haven.

Riesman, D., Glazer, N., Denney, R., [1950]1969. The Lonely Crowd: A Study of the Changing American Character. Abridged Edition with a 1969 Preface. Yale University Press, New Haven.

Riesman, D., 1954. Individualism Reconsidered and Other Essays. Free Press, Glencoe.

Riesman, D., 1964. Abundance for What? And Other Essays. Doubleday, Garden City.

Riesman, D., 1979. Ethical and practical dilemmas of fieldwork in academic settings. A personal memoir. In: Merton, R.K., Coleman, J.S., Rossi, P.H. (Eds.), Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld. Free Press, New York, pp. 210–231.

Riesman, D., 1988. On discovering and teaching sociology: a memoir. Annual Review of Sociology 14, 1–24.

Riesman, D., 1990. Becoming an academic man. In: Berger, B.M. (Ed.), Authors of Their Own Lives: Intellectual Autobiographies by Twenty American Sociologists. University of California Press, Berkeley, pp. 22–71.

Riesman, D., Glazer, N., 1952. Faces in the Crowd: Individual Studies in Character and Politics. Yale University Press, New Haven.

Riesman Jr., D., 8 January 1953. Paper, Not Class Taught Sociologist David Riesman. The Harvard Crimson.

Right to the City: A Liberal-Democratic Perspective

Shenjing He, Sun Yat-sen University, Guangzhou, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the last four decades, the concept 'the right to the city,' first developed by French philosopher Henry Lefebvre, has traveled from Europe to the rest of the world. This article discusses the key questions relating to the right to the city, and provides an overview of the evolution of this concept and how this abstract concept has inspired and empowered the struggles to fight against capitalist globalization worldwide, in particular the global south. In the global north, the right to the city has been widely exploited as an important tool to safeguard citizen's equal access to scarce social resources and opportunities, whereas in the global south, its enlightening power is being actively excavated to foster social equity, civil society, and democracy. The creative exploitation of this concept has been put in place in a variety of locales and leads to the enactment of a number of extended agenda locally and internationally. The enduring global economic recession provides opportunities for more expansive and interconnected exercises of the right to the city through allying the North and the South.

Originally formulated by French philosopher Henri Lefebvre, the idea of 'right to the city' was first inscribed as the right to 'renewed centrality, to places of encounter and exchange, to life rhythms and time uses' (Lefebvre, 1996[1968]: p. 179). Essentially, this thesis advocates for a vision of a life fully lived for urban inhabitants. The crux of the thesis emphasizes use values as the organizing principles of urban spaces, though the use values have been radically encroached by exchange values and colonized by the dominant classes in a world prevailed by capitalist globalization. Written just before the 1968 student riots in Paris, the right to the city thesis was rising from crisis and continues to face different crises today, from the crises of Fordism to the crises of Neoliberalism. The currently enduring global economic recession further exposes the inherent crisis and conflicts invariably piled up under the imperatives of neoliberalism over the last three decades, and leads to the outbreaks of political and legitimation crisis simultaneously. Nonetheless, these crises have offered challenges and opportunities for the dissemination and implementation of the right to the city worldwide.

Since the date of formulation, the right to the city has been widely (re)interpreted and excavated in academia, and extensively implemented in urban social movements and policy making. Building upon Lefebvre's original writings, Harvey (1973) made a focused and thoughtful attempt at interpreting the right to the city and social justice, by defining the right to the city as a collective right to the democratic management of urban resources based on a threefold criteria – need, contribution to common goods, and merit. The idea of the right to the city soon traveled from Anglo-North America to the rest of the world with its resonance found among activists and policy makers. Castells (1977) further developed this theory through offering a normative definition for urban social movements, which are considered to be the highest level of citizen action that brings about radical changes in political power or specific outcomes in the political arena, or redefines the urban system and its meanings. More recently, the idea of the right to the city has been commonly incorporated into the discussions on gentrification, immigration, housing, citizenship, urban public space, and social exclusion (see Amin and Thrift, 2002; Dikeç, 2001; Harvey, 2008; Mitchell, 2003). This can be seen as a restoration of critical urban theory and an endeavor to seek for a more democratic urban politics.

Apart from being a 'political ideal,' the right to the city has literately become 'a working slogan' inspiring and empowering numerous urban social movements worldwide. The idea is also articulated in a number of charters promoting justice, sustainability, and inclusion in cities, such as the World Charter for the Right to the City, the European Charter for Human Rights in the City, the Montreal Charter of Rights and Responsibilities, and the Mexico City Charter for the Right to the City, just to name a few. In translating the theory into practice, controversies and incongruences occurred among academics, activists, and international initiatives. Yet, efforts have been made to bridge the theoretical and practical explorations of the powerful idea of the right to the city.

Evolution of the Concept

Lefebvre's original writings on the right to the city challenge the social and political urban structures of the twentieth century, by proposing a new type of 'citizenship' at the urban level, distinct from the formal citizenship of the nation-state, to embrace heterogeneity, shared space, and societal ethics. This citizenship is to be achieved through public participation and democratic practices (Harvey, 2008). In the past four decades, the concept of the right to the city has evolved through the works of social activists and academics, the growing social and environmental awareness in the 1960s–90s, and recent innovations in social practice. It has been elaborated and practiced at different scales, from individual, household, and neighborhood level to the city level and beyond. The right to the city now encapsulates expansive conceptual and practical meanings, including the right to political space, the right to public space and housing space, the right to autonomy, and the right against state domination and policy brutality. It might be

associated with a variety of meanings and elaborations, fundamentally it is an outcry for social justice, a potent quest for the rights of various social groups in the city, and a compelling antithesis of the widely implemented neoliberal agenda. Most importantly, apart from informing and inspiring the daily struggles for specific city rights, it can also be turned into a more powerful and inclusive concept to encompass people's 'freedom to make and remake our cities and ourselves' (Harvey, 2008: p. 23). Apparently, the right to the city needs to be understood as a multiscalar and multilayer concept bearing millions of possibilities for seeking and creating a brighter urban future.

As one of the most influential urban theories, the right to the city has invoked heated debates and constant contemplations, and thus gives rise to the important questions of 'whose right is it about, what right is it and to what city?' (Marcuse, 2009: p. 185). Answering these questions would be crucial to fully comprehend the rich conation of the concept both as a critical urban theory and a working slogan in social practices. These questions also carve the story line of the evolving concept, and shape future opportunities for the development and application of the idea.

Whose Rights

Among academics and activisms, a strong belief is held that the right to the city entails the possibilities for enfranchisement and social inclusion, which eventually benefit all urban inhabitants. In this sense, a broader concept of citizenship is embedded in the right to the city thesis. It has been widely recognized that one's everyday experience of inhabiting the city rather than one's nation-state citizenship constitutes the substantive and meaningful citizenship. In practice, numerous attempts have been made to bring about a more inclusive citizenship. For instance, the Montreal charter has made possible a joint declaration by the state and citizens to ensure that citizens enjoy all sorts of rights to the city, and redefine "citizens" as "any person living within the city territory" (Brown and Kristiansen, 2008). Meanwhile, some scholars have made clear distinction between this new type of citizenship based on inhabitance and the conventional notion of national citizenship (Brown and Kristiansen, 2008). Specifically, some scholars propose disposable urban citizenships or multiple citizenships based on residency and place of work, which could be taken up on arrival but abandoned on departure (Bauböck, 2003). These endeavors have widened the definitions of citizenship and who has a right to the city based on residency. However, they still exclude the majority of the urban poor, such as migrants, workers in the informal economy, refugees and immigrants without legal documents, and commuters who go back and forth between different places on a daily basis. Apparently, framing citizenship in formal and territorial terms has failed to take urban politics and the social relations of power into account (Dikeç and Gilbert, 2002).

Therefore, the 'rights to the city' need to be defined with cautions to safeguard the livelihood and interests for the urban marginalized people. It should be the right for inhabitants of cities, especially those without citizenship or 'proper' documentations. Without bearing this in mind, the so-called 'slum-free' movements in support of the free-market ethos would be turned into a nightmare for a vast volume of dwellers in the substandard settlements in the developing world. Attending to these potential problems, Mark Purcell (2005: p. 14) contended that the right to the city must be ascribed to those who 'live in the city, who contribute to the body of lived experience and lived space.'

In addition, the rights to the city also include the right to difference, i.e., different race, ethnicity, class, and culture. This is of great significance in an age of globalization when flows of population have reached unprecedented speed and scale. More specifically, Peter Marcuse (2009) assigns the right to the city to those treated unjustly, those in want, oppressed, and dispossessed: the homeless, the disabled, the hungry, impoverished, and unemployed, and those disrespected and treated unequally: immigrants, racial minorities, sexual minorities, political activists, and so forth. For these people, the right to the city can be turned into a powerful weapon to fight against any forms of unjust hegemonic power and policies.

What Rights

Although there is little controversies on whose right the right to the city thesis entails, it has not been clearly stated in Lefebvre's original writings what kind of rights they should be. It has been widely recognized that the right to the city should be promoted as a collective right rather than specific and individual rights (Harvey, 2008). It might be more accurate to put it this way: the collective right to the city advocated by Harvey and others may include, but is not confined by specific rights – such as the right to housing, right to public space, or right to employment. In social practices, the right to the city has been rendered certain flexibility and openness. In many instances, it is a socioeconomic right, e.g., a right to housing, to transportation, and to natural resources like water. In other instances, the right to the city is more of a right to a communal good, like community, tradition, and culture. Such openness and fuzziness is not necessarily bad. Mitchell and Heynen (2009: p. 616) rightly point out the value of the concept of the right to the city 'is precisely its capaciousness,' 'because it allows for solidarity across political struggles, while at the same time focusing attention on the most basic condition of survivability.'

In general, the right to the city incorporates two principal rights: the right to participation and the right of appropriation. The right of appropriation refers to the rights of city residents to access and to make use of city space. The right to participation means that urban residents should have the opportunity and power to redefine and transform the urban space. In a similar vein, Harvey (2003: p. 939) proposes two sets of rights "the right to the city is not merely a right of access to what already exists, but a right to change it after our heart's desire," i.e., the right of now and the right of future. In addition, Douglass and Friedmann (1998) identify three claims for rights – the rights to voice, to difference, and to human flourishing. These claims, in particular the last point, have resonance with the vision of 'a life fully lived' advocated by Lefebvre. In sum, the right to the city is a set of tool kits for radical urban change, which generates opportunities and space for citizens to address and access their needs. To reach that end and to achieve substantive

citizenship, citizens must claim rights of participation, in addition to gaining access to rights of appropriation.

Overall, the right to the city encompasses rights of many kinds, be it a socioeconomic right or a liberty right, a legal right or a moral right, a prima facie right or an absolute right (Attoh, 2011). Hence, the right to the city constitutes not a singular right, but a bundle of connected rights. That said, these different rights claims might be in conflict with one another. As any other types of rights, the right to the city is not free of conflict and trade-off. For instance, David Harvey (2008) sees the right to the city as a collective right to the democratic management of urban resources, yet Don Mitchell (2003) illustrates the conflicts between the right to the city of the homeless and the 'democratic' management of public space. The question of what kind of right is a right to the city could serve to reveal the level of incoherence within the current debate and therefore to sharpen the right to the city's political edge. And one possible way to square such conflicts is to integrate the right to the city into a general theory of social justice or substantive democracy (Attoh, 2011).

To What City?

Despite some inherent incongruence, there are some common values shared by different adaptation of 'right to the city' envisioned by the academic, policy, and activist circles. Essentially, the right to the city looks beyond the physical space of the city, by interrogating all power structures and social relations determining the lives and experiences of urban inhabitants. In this sense, arduous tasks remain to be done to reach the city embraced by all its inhabitants, through combating against various influences and structures dictated by capitalist values.

To best address the political ideal remaining unreachable and to contrast the reality and ideal, Lefebvre had deliberately distinguished the 'city' from the 'urban.' The 'city' might be torn by corporate culture and dominant neoliberal agenda, while 'the urban' is a society beyond capitalism, a connected space shared by inhabitants and supported by meaningful social interactions (Purcell, 2013). The right to the city is therefore a struggle to take back urban space from capitalism and return it to its inhabitants. The struggles for the right to the city are therefore essentially the battles for the inhabitants who make the city a meaningful space to repossess the use value of urban space from the property rights owners chasing for exchange values. Claiming the right to the city is to restore the contaminated and comprised city to rediscover the urban and make it flourish (Purcell, 2013). For Lefebvre, 'the urban' foretells a revolution consisted of millions of everyday struggles and innovations, and can only be achieved through sustained political struggles.

As an important concept integrated to the right to the city thesis, *autogestion* – 'self-management' or 'worker control' – suggests replacing the powerful state-centered organizations with democratic management by citizen themselves through radical, democratic political mobilization and decentralization of political power (Lefebvre, 2001). Based on this idea, the vision of an ideal city would be "a city of centrality, access to information, truly broad participation and enfranchisement, of unalienated labor, and offering the opportunity to live life to the fullest" (McCann, 2002: p. 78).

Bridging 'Political Ideal' and 'Working Slogan'

Under the overarching framework of the right to the city, activists and international initiatives have led the challenge of translating the right to the city into practice worldwide, through urban social movements and institutionalization/legalization of the right to the city. In the global north, urban social movements have evolved to fight against dominant development paradigms over the past four decades, from urban renewal of the 1970s, to the sweeping neoliberalization of the 1980s, to the frenzy for local economic development of the 1990s, to the post-2000 highly unjust financial and property markets (Mayer, 2009). In the South, urban social movements have long been interwoven with the fight for basic human rights and political freedoms, e.g., claims for access to clean water, shelter, and dependable livelihoods. Meanwhile, other movements are also actively fighting against globalization and global justice to challenge the dominant economic system, and to contest the urban system shaped by the logic of capital. For instance, the Right to the City Alliance, a network of community organizations in the United States, was launched in 2007, bringing together community-based organizations to deal with various urban problems, such as gentrification, environmental justice, homelessness, cultural preservation, juvenile justice, and the well-being of gay, lesbian, bisexual, and transgender youth (Purcell, 2013). A similar effort began in Europe in 2010 to organize the networks of activist groups under the right to the city as an umbrella idea.

In the past four decades, the right to the city has not only informed activists in various social movements, but has also been incorporated into formal legislation and institutions, and officially employed as a vehicle for social inclusion and well-being for various excluded and alienated urban groups by a number of municipal, regional, national, and international organizations. Examples include Habitat International Coalition, the Centre on Housing Rights and Evictions, Action Aid, and the National Forum for Urban Reform, the World Charter, the Montréal Charter of Rights and Responsibilities, the Aberdeen Agenda of the Commonwealth Local Government Forum, the Brazil City Statute of 2001, and the Urban Policies and the Right to the City research project launched by UNESCO with UN-HABITAT (Brown and Kristiansen, 2008). Among them, World Charter is probably the most influential one, with its principles being adopted widely while not becoming an official legal document itself.

However, translating the critical theory of the right to the city into practice is not free of controversies. Some contemporary initiatives launched by international institutions are accused of limiting their claims to granting or institutionalizing the rights for certain disadvantaged groups through state legislation, while others are blamed for remaining as a tamed negotiation with the capital machinery ruling the world. These initiatives are considered drifting too far away from Lefebvre's original exhortation for going beyond the state and capitalism through radical struggles (Mayer, 2009). For these reasons, the

extensive use of the right to the city by certain global NGOs is considered losing its political edge (Mayer, 2009; Purcell, 2013). Their goals are often stated as 'inclusion' aiming to integrate currently excluded groups into the structures of government and the formal economy. Purcell (2013) further points out another drawback of contemporary initiatives which tend to confine the right to the city within a liberal-democratic framework, by granting people a relatively weak measure of democratic control through a set of state apparatus such as elections, parties, laws, and stable state institutions. Under this framework, the right to the city is at best proposed as a modest addition to the list of existing liberal-democratic rights, without challenging its legal foundation and inherent problems. For this sake, Purcell (2013) advocates returning to the pristine meaning of the right to the city to radicalize and inform political actions. By so doing, it is believed that a radical democratic society beyond state and capitalism can be formed by urban inhabitants themselves collectively. In this sense, the right to the city should be treated as political claims made by the mobilized groups. For Lefebvre, even though the horizon of '*autogestion*' will never be reached, it is of great importance to arouse the awareness and determination of struggles toward the horizon through the contractual citizenship binding the citizen together based on collective action and self-management.

On the other hand, for movements fighting against the capitalist urban system which are often equipped with a bottom-up approach and radical quests, such as the Right to the City Alliance in the USA and similar urban movements across Europe, they are deemed to be a direct response to Lefebvre's exhortation to create and claim rights to the city through social and political actions. In this sense, the right to the city is less a legal right, but rather a radical claim to subvert the extant social structure and system ruled by the rich and powerful. This revolutionary form of appropriation to the city space is considered to have a closer link to the Lefebvre's original idea.

Given the fact that the right to the city is so widely practiced in social movements and international initiatives, how the claim is perceived and interpreted would be rather different. While radical scholars tend to interpret the ultimate goal of the right to the city as to eventually disintegrate the relationship between citizens and the state (Purcell, 2013), others are accentuating the important role of government, particularly local government, in implementing the right to the city, as an intermediary between the urban poor and the global economy (Brown, 2013), or employing the concept to ameliorate the intensified confrontation between the 'have' and the 'have not,' by emphasizing the right to the city as a collective right. In many cases, the right to the city is also exploited to better achieve social and economic justice for poor people in urban areas, by prioritizing the social function of property and the city, and incorporating these principles within constitutional and statutory provisions and employing these principles at the local level through ambitious and innovative property programs (Pindell, 2006). Contrasting views on these social practices are also found among academics. On the one hand, Margit Mayer (2009) tends to contrasts those who seek to create rights through social and political actions with those international NGOs and UN programs seeking institutionalization of rights through legislation and city charters, to advocate for radical politics. On the other, Brown (2013) points out the limitation of 'radical politics' advocated by Mayer (2009) and others to confront the immediate problems suffered by citizens in the global south, and the danger to wipe out the potential for fundamental, legal, and institutional reform to make a fairer contract between state and citizen in the long run. She further contends that radical revolution and confrontation should be taken as a last resort to challenge oppression of rights to the city. In particular in the context of global south, radical confrontation and violent struggles could cause enormous aftermath for individuals and the country as a whole that might take decades to recover. Alternatively, Brown (2013) is in support of Fernandes' (2007) accentuation on the important role of legal and political order to change the exclusionary nature of urban development and to do justice to informal development in developing countries. Yet, this is not to suggest that the right to the city is reduced to the right to be remained in the city, or to be housed and serviced. The right to the city in fact holds real potential for liberal democracy, for enfranchisement, and social inclusion (Purcell, 2013).

To bridge the two ends of the right to the city both as 'political ideal' and 'working slogan,' Harvey (2001) gave a more pragmatic formula to exercise the right to the city, by contending that insurgent politics and grassroots activism may provide a route for realizing social change, as "transformative politics are constantly intermingling with local mobilizations," but needs to go beyond local concerns. In a similar vein, Brown (2013) points out that the most powerful agendas and exciting ideas will emerge from the intersection between creative theory and grounded practice of those addressing critical problems faced by the deprived and the alienated, especially those in the global south.

In addition, a fuller understanding of the right to the city and its practices also calls for a multiscalar reading. It is quite clear that the right to the city is to be understood and implemented at the city scale. Yet the exercise of the right to the city can be played out and be reached out by stakeholders at different scales: regional, national, international, and intraurban. Adding to the alteration in national legislation, changes in local institutions, including city governments and grassroots organizations, are also crucial to translate the right to the city into practice. At the intraurban scale, public space has long been the focal point of rights struggles in both the North and South. It has become the arena of claiming for democratic rights, and the shelter for those excluded from the commodified private domain, but is increasingly policed and controlled (Brown, 2013). Meanwhile, public space functions as a key forum for debates on rights and citizenship, particularly in societies dominated by corporate culture and 'big capital,' where the idea of citizen's rights to public urban space are rapidly obscured as corporate space has quietly penetrated into social space (Cuthbert and McKinnell, 1997). Having said that the exercises of the right to the city need to be examined at various scales, it also relies on a careful examination on the relationship between state and citizen, so as to reframe the configuration of 'the bundle of rights,' and redefine the connotation of

citizenship, through constant social struggles and intellectual contemplations.

The Right to the City in Global South and Beyond

Cities of the global south have been moving steadily from the margin to the center, demographically, economically, socially, and politically. The surging research interests in 'Southern urbanism' are certainly a significant progress going beyond understanding cities in the South based on 'Northern/Western theories.' If the 'right to the city' in the global north is often associated with democracy and social equality, in the global south, the concept deals more specifically with citizenship, housing rights, the right against forced eviction and the right of squatting (Davis, 2006; Fernandes and Varley, 1998), and the rights to livelihood that are often claimed through informal economic activities in public space. In the emerging economies and newly industrialized zones, the right to the city is of particular significance for the large volume of rural-turned urban population, who are often turned into the urban underclass. Similarly, in the context of transitional economies, a systemic transition from planned economy to market system ignites waves of rights-claiming movements, in which the rights to the city can be seen as a broad set of collective and individual rights concerning both the poor and the middle class against the party state and its neoliberal policies (Gainsborough, 2002; He, 2007; Sykora, 2005). City spaces have increasingly become a highly contested ground where diverse interest groups seek rights to survive, inhabit, and thrive. Some of these claims are modest, e.g., acquiring a livelihood and a living space of one's own, whereas others are more ambitious in its scope, e.g., challenging the predatory capitalist system through revolutionary movements. Regardless of the patterns and strengths of the struggles for the right to the city, the efforts of the fighters, in particular the less privileged urban social groups, will not only bring about a better life for themselves, but also have the potential to reshape the urbanization trajectories of their cities where elite groups have steered the courses of urban change under the imperatives of neoliberal accumulation regime (He and Chen, 2012).

Representing two rising powers in the world economy and political stage, Brazil and China exemplify two different stages and strategies of right to the city. In Brazil, the installation of City Statue gave priority to the social function of the city so as to safeguard city residents' rights of participation and occupation. In general, it has broadened the opportunities for urban inhabitants to actively participate in land use planning processes and has opened up the opportunities for them to acquire legally recognized property rights. The most remarkable change in Brazilian legislative system is granting legal property rights to the 'squatters.' Its major achievement should not be understood simply as sheltering the squatters alone, but should be interpreted as a bold attempt to entitle the poor and the marginalized with the opportunities to effectively access housing and shelter. It also sows the seeds for hope to fully access to a bundle of rights to the city beyond the right to housing, eventually to reshape the city in a way through which all inhabitants could have a fully lived life. Besides the installation of City Statue in Brazil, the right to the city has also been adopted elsewhere in Latin America, e.g., in Colombia's law no. 388/1997, Article 31 of Ecuador's 2008 Constitution, in housing rights in Chile, and in the Charter for the Right to the City in Mexico City (Brown, 2013). Certainly, these initiatives will not solve all the problems commonly faced by these countries, nonetheless the way how they are executed and stretch out to other spheres and other contexts will be critical to progress the fundamental reform that the right to the city entails. At the local level, empowered by the right to the city, struggles either in the form of resistance or participation are prevalent across Latin America, e.g., appropriating public space through Mexican 'public cultural spectacle' and 'symbolic spatial practices' (Villa, 2000), protests, and demonstrations following the tradition of 'taking to the streets' (Friedmann, 1992). This tradition not only transforms the urban social landscapes in Latin America, but is also taken up as a weapon fighting against various crises persisted in the neoliberal urban system in other locales.

In China, it was quite unusual to see urbanites seeking for any kind of right to the city under socialism. With different forms of dispossession revolving around rights to urban spaces becoming ubiquitous, Chinese cities have seen a rise of the '*weiquan*' (meaning defending rights) movements. Yet, so far urban social movements in China are still limited to a very small proportion of the urban population, mainly seen in two forms: labor protests led by laid-off workers and migrant workers, and property rights defending by homeowners (see Cai, 2002; Friedman, 2009; Lee, 2007). As the country and its dominating class have voluntarily subscribed to the neoliberal urban ideology with intense commodification of urban spaces, several tiers of citizenship are taking shape in the Chinese city, especially after marketization and neoliberalization taking up speed since the 1990s. Class stratification is formed among rural migrants and formal urban citizens of different statues. Among the latter, former socialist workers who became the new urban proletarian and the emerging *nouveau riche* who gains a windfall fortune during the economic reform are two contrasting social groups. For both rural migrants and other urban underclass alike, their rights to the use values of city spaces are vulnerable to be traded off by the exchange values of city spaces, which are often dominated by the emerging affluent class in order to maintain or redefine their class status in a rapidly changing society. Subjected to the dual forces of institutional inheritance from the socialist period and the capitalist imperatives introduced by the market reforms, the development path of the present-day Chinese city is featuring suppressed rights to the city, especially for those deprived by both forces. In this sense, the right to the city could potentially serve as a powerful weapon to fight against the emerging urbanism dominated by intensified inequality and marginality. Although over a long period of time, most Chinese urbanites chose to remain as the 'silent majority' in face of social inequalities, the installation of the Law of Property Rights in 2007 and the constantly enlarging volume of work done by NGOs and activists have profoundly wakened their consciousness of the right to the city. However, confined by the existing legislative framework and rigid political control, most urban social movements in China remain as individual

struggles, most commonly exemplified in the 'nail household' phenomenon which largely aims to bargain for better compensation packages, or limit to claiming economic rights rather than political rights and citizenship. Whereas the magnitude and amplitude of urban social movements under the rubric of the right to the city in China are not comparable to those in Latin America, the rising consciousness of rights to the city and diverse struggle tactics fighting against social inequality are encouragingly progressing. For instance, amidst the surging waves of gentrification in Guangzhou, the third largest city in China, local residents have started to progress from claiming the rights to appropriation of housing space to claiming the rights to participation in urban redevelopment decision making (He, 2012). In the meantime, under the conceptual framework of the right to the city, scholarly work has been done to interrogate how the remaking of the Chinese city is being shaped by the imperative of capital accumulation and by the subordination of social interests and marginal groups to the neoliberal forces emanating from different spatial scales, also to understand those bottom-up struggles to contest the unequal rights (see He and Chen, 2012; Qian and He, 2012).

In the global south, social inequality and poverty are pandemic, be it in Latin America or China. Several themes are at the frontier of struggles for the right to the city, such as contested public space, informal settlement/slum redevelopment, fast building of new towns/cities or the so-called 'instant urbanism,' various forms of enclaves associated with 'splintering urbanism,' and rigid urban governance as neoliberal institutional apparatus (Samara et al., 2013). Of course, these themes are not unique to the global south, which are recurring in different locales and becoming the focal point of rights struggles. It is worth noting that the breath and intensity of right struggles in global south are definitely unprecedented in the north, and are thus of great potential to cultivate revolutionary transformations to achieve the ultimate goal of the right to the city. Nowadays, in North American suburbs where the aftermath of the subprime crisis is still pandemic, people are picking up the strategies of land rights claiming from the global south, where a tradition of taking rights struggles to the street and legislative endeavors has long existed.

Along with the sizable and influential urban social movements, the achievements of international initiatives such as UN, and legislative endeavors at national level, e.g., the City Statue in Brazil, academic work on the global south has now acquired a distinct body of scholarship with its own questions, approaches, and contributions to the right to the city literature (Samara et al., 2013). This scholarship has started to bridge with and inform the social practices and academics in the north, as the two poles of the urban world increasingly bend toward each other in some important ways (Smith, 2002), accompanying the constantly unfolding global economic recession starting from 2007. As a harbinger bridging the South and North scholarship, Castells (1977) offers a comparative perspective between the cities of South America and of Europe, which are sensitive to cultural and political particularities of the South. In addition, Mayer (2009) explicitly foretells the possible opportunities for the right to the city in an age of globalization: a variety of social urban movements has become intricately linked alongside the world economy. In certain ways, the global economic recession justifies urban social movements' claims and arguments about the unsustainability and destructiveness of the neoliberal growth model. Such a situation also defends the ultimate declaration of the right to the city and expands the opportunities to make it work, and advocates for democratizing cities and their decision-making processes. Through exposing and politicizing the implications of the current global crisis, in particular the outrageous environmental, social, and political problems, social movements for the right to the city could make good use of these opportunities (Mayer, 2009). Evidently, the global south is becoming the epicenter both for social struggles for the right to the city and cutting-edge urban theories development deriving from the so-called 'Southern Urbanism.'

See also: Citizenship, Anthropology of; Environmental Justice in the United States; Human Rights and Social Work; Social Justice and Social Inequalities.

Bibliography

Amin, A., Thrift, N., 2002. Cities: Reimagining the Urban. Polity Press, Cambridge.
Attoh, K.A., 2011. What kind of right is the right to the city? Progress in Human Geography 35 (5), 669–685.
Bauböck, R., 2003. Reinventing urban citizenship. Citizenship Studies 7 (2), 139–145.
Brown, A., Kristiansen, A., 2008. Urban Policies and the Right to the City: Rights, Responsibilities and Citizenship. UNESCO, MOST (Management of Social Transformations) Programme, Paris.
Brown, A., 2013. The right to the city: road to Rio 2010. International Journal of Urban and Regional Research 37 (3), 957–971.
Cai, Y., 2002. The resistance of Chinese laid-off workers in the reform period. The China Quarterly 170, 327–344.
Castells, M., 1977. The Urban Question: A Marxist Approach. MIT Press, Cambridge, MA.
Cuthbert, A.R., McKinnell, K.G., 1997. Ambiguous space, ambiguous rights – corporate power and social control in Hong Kong. Cities 14 (5), 295–311.
Davis, M., 2006. Planet of Slums. Verso, London and New York.
Dikeç, M., 2001. Justice and the spatial imagination. Environment and Planning A 33, 1785–1805.
Dikeç, M., Gilbert, L., 2002. Right to the city: homage or a new social ethics? Capitalism Nature Socialism 14 (2), 58–74.
Douglass, M., Friedmann, J. (Eds.), 1998. Cities for Citizens: Planning and the Rise of Civil Society in a Global Age. Wiley, New York.
Fernandes, E., 2007. Constructing the right to the city in Brazil. Social Legal Studies 16 (2), 201–219.
Fernandes, E., Varley, A., 1998. Illegal Cities: Law and Urban Change in Developing Countries. Zed Books, London.
Friedman, E., 2009. External pressure and local mobilization: transnational activism and the emergence of the Chinese labor movement. Mobilization: An International Journal 14 (2), 199–218.
Friedmann, J., 1992. The right to the city. In: Morse, Richard M., Hardoy, Jorge E. (Eds.), Rethinking the Latin American City. Woodrow Wilson Center Press, Washington, DC, pp. 135–151.
Gainsborough, M., 2002. Understanding communist transition: property rights in Ho Chi Minh City in the late 1990s. Post-Communist Economies 14 (2), 227–243.
Harvey, D., 1973. Social Justice and the City. Johns Hopkins University Press, Baltimore.
Harvey, D., 2001. Spaces of Capital: Towards a Critical Geography. Edinburgh University Press, Edinburgh.
Harvey, D., 2003. The right to the city. International Journal of Urban and Regional Research 27, 939–941.
Harvey, D., 2008. The right to the city. New Left Review 53, 23–40.
He, S.J., 2007. State-sponsored gentrification under market transition – the case of Shanghai. Urban Affairs Review 43 (2), 171–198.

He, S.J., 2012. Two waves of gentrification and emerging rights issues in Guangzhou, China. Environment and Planning A 44 (12), 2817–2833.
He, S.J., Chen, G., 2012. Interrogating unequal rights to the Chinese city. Environment and Planning A 44 (12), 2794–2800.
Lee, C.K., 2007. Against the Law: Labor Protests in China's Rustbelt and Sunbelt. University of California Press, Berkeley and Los Angeles.
Lefebvre, H., 1996[1968]. Writings on Cities. Blackwell, Oxford.
Lefebvre, H., 2001. Comments on a new state form. Antipode 33 (5), 769–782.
Marcuse, P., 2009. From critical urban theory to the right to the city. City 13 (2), 185–197.
Mayer, M., 2009. The 'right to the city' in the context of shifting mottos of urban social movements. City 13 (2), 362–374.
McCann, E., 2002. Space, citizenship, and the right to the city: a brief overview. GeoJournal 58, 77–79.
Mitchell, D., 2003. The Right to the City: Social Justice and the Fight for Public Space. The Guilford Press, New York.
Mitchell, D., Heynen, N., 2009. The geography of survival and the right to the city: speculations on surveillance, legal innovation, and the criminalization of intervention. Urban Geography 30 (6), 611–632.
Pindell, N., 2006. Finding a right to the city: exploring property and community in Brazil and in the United States. Vanderbilt Journal of Transnational Law 39, 435.
Purcell, M., 2005. Globalization, urban enfranchisement, and the right to the city: towards an urban politics of the inhabitant. In: Wastl-Walter, D., Staeheli, L., Dowler, L. (Eds.), Rights to the City, International Geographical Union, Home of Geography Publication Series, vol. III. Societa Geografica Italiana, Rome, pp. 11–25.
Purcell, M., 2013. Possible worlds: Henri Lefebvre and the right to the city. Journal of Urban Affairs (online early view).
Qian, J.X., He, S.J., 2012. Rethinking social power and the right to the city midst China's emerging urbanism. Environment and Planning A 44 (12), 2801–2816.
Samara, T.R., He, S.J., Chen, G., 2013. Locating Right to the City in the Global South. Routledge, New York.
Smith, N., 2002. New globalism, new urbanism: gentrification as global urban strategy. Antipode 34 (3), 427–450.
Sykora, L., 2005. Gentrification in post-communist cities. In: Atkinson, R., Bridge, G. (Eds.), Gentrification in a Global Context: The New Urban Colonialism. Routledge, London, pp. 90–105.
Villa, R.H., 2000. The right to the city in Los Angeles: discourse and practice of a Chicano alternative public sphere. In: Hill, Mike, Montag, Warren (Eds.), Masses, Classes, and the Public Sphere. Verso, London and New York, pp. 41–61.

Rights: Legal Aspects

Michael McCann and Stuart Scheingold[†], University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The language of legal rights generally connotes potentially enforceable claims of entitlement to certain freedoms, goods, or conditions. Rights traditionally have been identified with Western liberal constitutional regimes, but non-Western foundations and international human rights conventions have become increasingly salient in contemporary world politics. This article reviews a wide range of perspectives about how rights are conceptualized and critically assessed by legal philosophers, jurisprudential analysts, and empirical social scientists.

Definition and Background

The Basic Logic of Rights

The language of rights generally connotes potentially enforceable claims of entitlement to certain freedoms, goods, or conditions. To assert that 'I have a right …' is to articulate a normative claim that certain actions should or should not be performed by other individuals, agencies of the state, or intermediate social institutions. As such, rights are related to, but different from, preferences, interests, and needs. Although rights are often invoked in the service of preferences and interests, only those rights that confer legal and/or moral authority that obligate others are enforceable by official authority. Whereas the assertion of a preference connotes merely that 'I want…,' to assert a right is to declare that 'I am entitled to….' Rights thus assume priority in the normative schemes of social ordering; they take priority over – or 'trump', as Ronald Dworkin (1984) memorably said – preferences and interests. Rights similarly take precedence over needs in much the same way, but with one major difference: needs, unlike preferences, convey an objective status that can be validated by parties other than the claimants themselves. This feature is underlined by the fact that rights are invested with general authority by the constitutional and statutory texts in which they are enumerated, by judicial constructions within various governmental orders, and by the potential for state enforcement. Although claims to rights are subject to third-party validation and enforcement, they are typically initiated by individual agents or groups acting on their own behalf (Waldron, 1996).

Because rights constitute relationships between the bearers of rights and other persons or institutions, each person's legal rights impose correlative duties on others. These duties may be *negative* in that they mandate noninterference or inaction by others, as in one's claim of a right to liberty, privacy, or autonomy. Conversely, the duties imposed by rights may be *positive* in that other parties are called upon for action and assistance – as, for example, to ensure the fulfillment of contractual terms or to provide accessible facilities for voting in political elections. Rights may be primarily negative or primarily positive, but they may also involve elements of both.

Similarly, the exercise of rights typically confers responsibilities on the rights-bearer. For example, a person may have a basic right to own property, but the right imposes enforceable responsibilities that limit activity of the property owner that causes harms to other property owners or citizens. Finally, violations of rights typically authorize legal remedies to rights claimants. Thus, for example, not only does the right to equal treatment oblige others to refrain from invidious racial discrimination, but should wrongful discrimination occur, the right to equal treatment requires a legal remedy to compensate for and/or to correct the denial of a basic entitlement.

The Origin and Evolution of Rights

The most widely recognized origin of liberal rights is in European social contract theory dating back to the seventeenth century, and especially to John Locke's classic *Two Treatises on Government*. From this work, rights emerged as a moral imperative that is found at the very core of several of the most important political documents in the modern age, including the United States Declaration of Independence as well as the Bill of Rights, and the French Declaration of the Rights of Man and Citizen. Although the concept of rights thus dates back several hundred years, and has provided a founding principle for several long-standing liberal constitutional orders, it is in the second half of the twentieth century that the moral ethos of rights took on widespread significance around the world. The persuasive power of rights developed in part in reaction to horrors of authoritarian states – Hitler's genocide, Stalin's purges, colonial repression, and the like – and in part in recognition of deprivations suffered by ethnic and racial minorities, women, sexual minorities, and other marginalized groups in nations throughout the world.

The Expanding Variety of Rights Claims

The evolution and diffusion of rights through the centuries and around the globe has been boiled down by analysts to three separate forms, or generations, of rights claims.

First-Generation Rights

These are closely linked to the rule of law and to liberal social contract theory; they are regularly found in the constitutions of liberal democratic regimes throughout the world, as well as in

[†] Deceased.

such transnational legal orders as the European Convention on Human Rights. First-generation, or classic, rights are embedded in the logic of both property ownership and the rule of law itself. On one hand, such rights tend toward the more negative pole, demarcate limits on interference by the state or other social actors, and turn on conceptions of liberty, privacy, due process, and the like. On the other hand, classic first-generation rights also include more positive rights to political participation – to free speech and association, to vote, to run for office, and so forth – and to fair treatment by others according to the principles of due process and equal protection. These rights claims are the most widely recognized and the least contested in general, although specific constructions vary greatly across time and jurisdictional space.

Second-Generation Rights

These can be traced to the social democratic traditions of Europe and Latin America but have during the last century gained currency elsewhere – although not so much in the United States. Second-generation rights have to do primarily with entitlements to economic and social well-being of the sort formulated in Articles 22–27 of the Universal Declaration of Human Rights (1948) – including, for example, basic education, health care, minimum income, secure employment, workplace organization, food, housing, social security, and the like. While first-generation rights have been celebrated as resources for challenging excessive state violence and making governments more accountable, many analysts find that second-generation socioeconomic rights, which usually require wealth redistribution, have generally proved less potent sociolegal conventions.

Third-Generation Rights

These are newer in conception, different in kind, and particularly controversial. Advocates of third-generation rights seek to confer entitlements to nonindividualistic, or communal, goods such as peace, security, a healthy environment, natural resource preservation, group self-determination or sovereignty, and animal welfare. The challenge posed by third-generation rights, which compounds the limited realization in practice of second-generation rights, will be addressed below.

Rights beyond the Liberal Western Tradition

A widely varied and robust area of academic inquiry in recent decades, largely responding to the palpable expansion of rights talk around the world, has addressed the development of rights-like conventions in many different religious and ethnic traditions. Scholars differ about whether, or how much, these traditions developed independent of Western influence or from historical encounter with Western cultural power and especially colonial influence. For example, Feldman (2000) argues that, contrary to long-held views about the strong cultural traditions that emphasize social harmony, Japanese citizens have long enacted general conceptions of rights and engaged in both individual and collective rights-based disputing. Many scholars likewise have identified the strong traditions of rights – including social and economic rights as well as private and political rights – that developed historically in Latin America prior to the twentieth century and contributed in unique ways to the development of human rights (see Wright-Carozza, 2003). Other scholars have demonstrated how rights traditions have developed in Muslim societies, sometimes independently and sometimes as hybrid constructions of sharia practices and Western secular rights traditions (Osanloo, 2009). Indeed, many scholars predict that such hybrid constructions will continue to proliferate around the world as international conventions on human rights are 'translated' and adapted to fit the wide variety of local traditions (Merry, 2006).

Contemporary Philosophical Debates about Liberal Legal Rights

The original conceptions of so-called liberal Western thinkers in the social contract tradition grounded rights in the theocratic logic of God's divine plan and natural law. Later versions downplayed the divine hand but continued the emphasis on rights as self-evident and inalienable features sewn into the rational order of the natural universe or human nature. By the twentieth century, the philosophical foundations for rights became thoroughly secular – whether justified by various forms of moral rationality or appeals to proven historical efficacy. The tensions among these diverse legitimating logics have defined the terrain for many of the most important philosophical controversies about rights. We identify just a few of such persistent debates, which mostly refer to Western liberal traditions in the United States and Europe.

'Nonsense Built on Stilts'

This critique, which can be traced back at least to Jeremy Bentham and continues today among critical 'neorealist' philosophers, turns on the recognition that rights are indeterminate (see Tushnet, 1984). Its core contention is that if rights are indeterminate and contingent, if they always are changing, and vary with context, then they really mean little at all and convey no real social power. Rights, it is argued, are simply moral or legal fictions that possess none of the objectivity, coherence, independence, and universality that philosophers, state officials, and popular culture assign to them. At best, rights are epiphenomenal reflections and rationalizations for relations grounded in other types of causal forces or power structures. Respondents to this position acknowledge that rights are social constructions, but the very fact that people believe that rights matter and act accordingly in interpersonal relations, often in the shadow of official enforcement, demonstrates the power of rights. It is further argued that all moral discourses are indeterminate and that it is, therefore, inappropriate to dismiss rights discourses as meaningless because they are unstable and contestable. Indeed, according to this argument, the indeterminacy of rights discourse, as with moral and legal discourse more generally, is desirable in providing a flexible, dynamic, and pragmatic tradition for ordering social relations in authoritative, just ways.

Who Has Rights?

The debate over the qualifications for rights-bearing is as old as the concept itself and has expanded among moral philosophers

and jurisprudes as specific rights claims themselves have proliferated. From the start, rights typically have been understood to be properties belonging to all individual human agents. Legal rights are thus often understood as inclusive and universalistic, belonging to all capable humans. Because rights traditionally have been authorized and enforced by nation states, though, legal rights have often been the prerogative of recognized citizens in specific sovereign domains. Indeed, it is commonplace to recognize that national citizenship confers the legal 'right to have rights.'

Somewhat ironically, though, this designation of rights-bearers as qualified human agents has limited the reach of, and some might say betrayed, the promise of rights, including both general citizenship rights and specific rights entitlements. Indeed, the claim that rights apply only to rational, disciplined, educated, human agents capable of self-government has been invoked to deny legal rights to children, women, ethnic and racial minorities, religious adherents, immigrants, and others in many societies that generally recognize individual rights (see Williams, 1991). Similarly, the principle that rights inherently presume qualified individual human claimants has been central to contemporary arguments that third-generation entitlements of animals and other nonhuman forms of life (e.g., trees), as well as nondivisible communal goods such as peace and a safe environment, are not suitable for rights claims. The philosophical skeptics argue that such claims lack theoretical coherence and might even undermine the social power of rights discourse itself. Such debates have generated extensive reconsideration about the foundation of rights as a social practice.

One implication of this inquiry into who can effectively claim rights is to raise questions about the politics of who grants legal rights. Although sovereign nation states traditionally have been the primary sources of authorized rights construction, administration, and enforcement, historical investigation have made clear that, whether principled or not, dominant groups control who gets rights – when, where, and how much – in practice within national and subnational contexts. Despite the pretenses to general rights entitlement that attend most rule of law systems, ascriptive and tribal standards routinely motivate the powerful to exclude many persons from full rights status. In short, all persons may be said to 'have' equal rights, but how persons are perceived by dominant groups shapes the rights they can in fact enjoy in practice; all rights politics arguably is status politics, to some degree. Moreover, if full rights status entitles disciplined subjects to treatment by liberal legal norms and protocols – including rights to due process, equal protection, property use, political participation, humane and noncoercive treatment, and the like – then those subjects marked as undisciplined, undeserving, or dangerous are doomed to less than full standing as rights-bearers. The latter thus remain vulnerable to more arbitrary, coercive, paternalistic, or even brutal forms of social control. Hence the familiar exercise of legally permissible violence against, or exclusion of, noncitizens, second-class citizens, 'criminals,' and the like for those marked as unqualified for or undeserving of rights in many times and places. And this historical politics of variable rights allocation by dominant groups, both through and beyond nation states, is one of the primary catalysts to the universalistic appeals of human rights that might trump national or local hierarchical power structures.

Do Rights Encourage Atomism, Egoism, and Selfishness?

The contention that rights practices encourage excessive individualism, egoism, and selfishness *among* rights-bearers was perhaps most forcefully articulated by Karl Marx and is commonly asserted by moralists across the political spectrum today. Critics emphasize in particular the tendency of rights to celebrate the individual and thus to undermine citizens' sense of responsibility for their actions, duties to others, and shared commitment or obligations to 'community' (McCann and Scheingold, 2012).

Theorists more favorably disposed to liberal conceptions of rights acknowledge that rights discourses presume, encourage, and protect individuals. Indeed, securing and advancing the exercise of individual citizen capacities for choice, action, self-development, and diversity against the harmful designs of other individuals, social groups, popular majorities, and the state has been the *raison d'être* of rights as a moral concept from the start. At the same time, however, most theorists understand the legal rights of individuals in relational rather than insular terms. Rights practices sustain particular obligations that are reciprocal among citizens as well as a general tolerance and even respect for the human dignity of others. In Kant's (1948: pp. 95–96) famous words, "Act in such a way that you always treat humanity, whether in your own person or in the person of another, always as an end and never simply as a means." Defenders thus argue that, far from undermining a sense of duty, rights support and enforce dutiful interaction among citizens. According to this way of thinking, rights connote a particular vision of community where individual agency and dignity are respected but balanced with conceptions of fairness, justice, and equal treatment for all. It follows that it is precisely the disregard for basic rights that encourages or permits unjust or exclusionary acts that are corrosive of community.

Critical Left theorists have offered insights to this debate. For example, Wendy Brown (1996) and others draw on Karl Marx and Michel Foucault to recognize that the disciplined, normalized subjectivity that is required for legal rights eligibility privileges private pursuits while impeding political dissent that challenges the hierarchical status quo. In short, rights reinforce commitments to private freedom and public deference or apathy. Other theorists (Williams, 1991) recognize that some rights are more self-regarding than others. Marx's powerful indictment, after all, was that it is, specifically, property rights that encourage selfishness, materialism, and competition for dominance through capital ownership. As a consequence, he argued, the positive communal promise of political rights is overwhelmed, thus reducing the promise of political solidarity and equality to an illusion. The underlying point here is that critiques of legal rights that fail to distinguish among different types of rights, within the first generation as well as among the different generations of rights concepts, are unconvincing to many moral and legal theorists. Nevertheless, the claim that rights claiming is selfish and detrimental to the common good or sovereign majority remains an effective political tactic by some

parties seeking to undermine the rights of other persons or groups.

Jurisprudence and Legal Theory

Contemporary discourse on rights within jurisprudence and legal theory is driven by an effort to identify a distinctively legal conception of rights. This discourse is organized around two foundational debates – pitting rights against democracy on the one hand, and natural law against legal positivism on the other. Historically, there have been profound and heated conflicts among the parties to these debates. These disputes are often entangled in other debates over specific policy preferences, over sovereignty when international standards conflict with national authority, and over who is the final arbiter of constitutional, common law, and/or statutory rights – specifically over the role of courts vis-à-vis other state or social actors – within constitutional regimes.

But more recently a rapprochement has developed around the recognition that rights and law are mutually constitutive – that rights without the law are virtually unthinkable, as is law without rights. This convergence not only identifies the core common ground of legal theory, it also reveals the distance that separates the legal theory of rights from the understanding of rights that is found in moral and political philosophy. Whereas philosophical understandings of rights reach well beyond the law in search of universal moral meaning, rights in legal theory are much more constrained – principally by way of the intrinsic connection in legal thought between rights and the institutional processes in which rights under law are embedded. In particular, legal theories of rights seek to specify and to justify the responsibility of courts and judges for resolving disputes over rights, and for policing the boundaries of legitimate rights claiming and recognition.

Legal Rights vs Democracy

The dilemma of legitimating rights claims is especially troublesome to legal officials in democracies and republics, where majority will is understood to carry considerable authority. We noted earlier that legal rights, by their very nature, are intended to protect and advance individual entitlements that restrict the actions of others, including those of popular majorities. As such, rights, both in general and in specific instances, are often condemned as counterdemocratic forces. Utilitarian theorists raise a parallel objection that begins with the assertion that rational public policy should provide the greatest good for the greatest number, thus giving considerable weight to the preferences of the majority. Rights are thus often challenged as costly obstacles to the development of democratic and efficient (see below) social policy.

Other legal theorists respond that rights do not thwart so much as give structure to utilitarian policy calculation, by assigning higher values to certain basic entitlements – especially those concerning protection for individual freedom and dignity – than to other competing considerations. Consequently, securing rights may impose costs on particular social goals, but these costs are arguably incurred in order to sustain larger, enduring social commitments. Likewise, it is argued, rights claims do not absolutely throttle democratic majorities or elected officials; rather, rights claims tend only to place specific, limited restrictions on particular policy actions that defy core social values and protection for individuals. Moreover, it is important to underline that many types of rights advance – indeed, are actually essential to – the democratic process. Such entitlements include, at the least, protections for dissenting speech, political association, free elections, due process, and basic public education. There is no simple way out of this debate. Although it may well be true that rights in general provide sustenance to democratic polities, particular rights constructions certainly do, and indeed are intended to, defy the will of elected officials and popular majorities. Hence the enduring challenge for judicial authorities is to justify by legal argument their specific rights constructions.

Rights and Justice

Historically, a divisive debate in legal theory between positivism and natural law has pitted a theory of law that recognizes little in common between morality and law against a conception of law as being inseparable from morality. The latter is virtually indistinguishable from the moral philosophy of rights discussed earlier, and represented by the work of Locke and Kant, among others. The former is rooted in analytic philosophy and in a logical positivism that eschewed the search for universal truths. During the late twentieth century, the gap between legal positivists and natural law theorists narrowed (but was by no means closed) by the emergence of a widespread agreement on a least-common-denominator conception of justice as the 'right to have rights.' According to this way of thinking, the guarantees associated with rights are confined to assuring predictability, reciprocity, and impartiality in the application of the law. What remains unresolved in the debate between theorists of natural law and legal positivism is whether justice, as thus defined, is the differentiating element of legal right, or only a quality that tends to be closely associated with legal rights.

Although many jurists have not reached out to the broader realms of political and moral philosophy, Ronald Dworkin (1977), one of the most prominent twentieth-century legal theorists of rights, has done so. At the core of his analysis is the distinction between policy and principle. Whereas policy is driven by utility, principle – the defining element of law – should be observed not because it will advance or secure an economic, political, or other situation deemed desirable, but because it is a requirement of justice, fairness, or some other dimension of morality. In thus linking law to principle, and principle to justice, Dworkin identifies himself with the relatively noncontroversial common ground between positivism and (neo)natural law. On the other hand, by introducing some other dimension of morality he clearly broke new ground. For Dworkin, constitutions and constitutional rights "fuse legal and moral issues" (Dworkin, 1977: p. 185). Constitutions thus provide not only the right to have rights, but particular substantive rights as well.

For Dworkin, then, rights are inextricably linked not only to the foundational elements of positive law, but also, at least in their political and constitutional incarnation, to the foundational elements of the state. In this sense Dworkin can be seen

as a kind of transitional figure among legal theorists, reaching out as he does to moral and political philosophy and incorporating them into his analysis of rights constructions by judges and other state officials. But unlike the theories of moral and political philosophers, Dworkin's theory of rights is embedded in, and inextricably linked via principle to, what is distinctively legal. Moral philosophy thus becomes an adjunct to the law and is, in effect, put at the disposal of the law, as a source of principle. In this sense, Dworkin's theory of rights returns to the starting point and indeed to the cornerstone of jurisprudence and legal theory; that is, it is a search for what is uniquely legal, as opposed to moral, ethical, or political.

Empirical Social Science Studies of Rights Practice

Whereas philosophic and jurisprudential inquiries into rights tend to be normative and abstract, social science research regarding rights practices is explanatory and empirical, although it is often informed by questions from critical social theory. Generally speaking, in the social sciences, legal rights are treated less as predictable, state-authorized entitlements of historical persons than as contingent resources that are mobilized both by individuals in everyday life and by organized interests in political advocacy. Rights are not viewed as 'above politics' and power so much as constitutive of political activity and power relations. The primary objective of the social sciences thus is to assess the value of legal rights as resources, and to explain variations in how rights work and to what effect in different sociopolitical contexts. To do so, social scientists shift the locus of rights inquiry from formal systems of logic and idealized institutions to inductively organized research on legal rights as practices in actual institutional settings, everyday life, and political advocacy.

Rights and Everyday Disputing

There has been a rapidly expanding scholarly literature on the ordinary disputing practices of citizens, especially in the United States but increasingly around the world. Many studies of 'ordinary' (non-elite) individual citizens in the United States suggest that rights matter a great deal, and that these individuals are not hesitant to assert them. Indeed, some scholars demonstrate that lay people, more so than judges, clerks, police, and other legal authorities, act on the premise and the expectation that the legal system exists to protect rights. One of the most interesting findings of Ewick and Silbey (1998) is that ordinary, non-elite citizens in the United States can relate to rights in flexible, creative, and complex ways, entertaining several conflicting notions at once. Interviewees in their study could view rights instrumentally as malleable rules of a game, authoritatively as an objective source of valid law, and skeptically as a burden that must be endured, evaded, or resisted.

Other studies qualify or challenge assumptions about the salience of legal rights. Research on the disputes of individuals suggests, for example, that Americans and others often refrain from asserting their rights; many studies document a widespread reluctance to seek formal redress from the state or other authoritative institutions when rights are violated and entitlements denied. This is especially the case among lower-income and minority citizens. Moreover, some studies offer evidence of considerable cultural biases and popular pressure against rights claiming, at least through formal adversarial channels, even in allegedly 'rights-obsessed cultures such as the United States. Moreover, much sociolegal studies question whether most citizens are as beguiled by what Scheingold famously labeled the "myth of rights" – the promise that legal institutions protect and enforce inherited rights for all persons – as once was believed (Lovell, 2012). Patterns of individual rights practices thus tend to be complex, contradictory, and context specific.

Studies into everyday rights construction and practices beyond the US context have proliferated in recent years. On the one hand, there is much evidence that legal rights talk and engagement is growing in many places around the globe, including in contexts of Islamic faith (Osanloo, 2009) and much of the global South (Wright-Carozza, 2003). On the other hand, much evidence suggests that citizens in many contexts with well-established constitutional mechanisms are reluctant to claim rights even when they have legitimate legal grievances about harm or neglect (Atuahene, unpublished). Especially relevant in this regard is Engel's admonition, grounded in his rich empirical study of provincial communities in Thailand, that increasing diffusion of rights discourse among cosmopolitan elites is often not shared by many ordinary people (2012). As Engel encourages, we need much more study of ordinary disputing and relationships in varying contexts to understand how much (or how little) the diffusion of human rights has mattered in practical affairs.

The Politics of Rights: Mobilizing Rights in Political Advocacy

The value of legal rights as a political resource for challenging hierarchy and injustice is heavily debated within the social science literature. The most stringent critics of rights strategies claim that to deploy rights is at best ineffectual, and at worst counterproductive. Some studies suggest that rights litigation has no more than negligible effect on policy outcomes. According to realist critics, only political power built around electoral coalitions and influential interests can bring about changes in well-established legal policies and practices. Other scholars demonstrate that rights-based practices often do intrude into policy processes, but with the unfortunate consequence of making these processes rigid and inefficient. And some empirical studies show that human rights discourse promotes illusionary promises of freedom that divert attention from fundamental transformations in hierarchical power relations (Englund, 2006).

There is, however, a substantial body of research that suggests rights strategies can make meaningful contributions to a transformative politics, albeit in complex, contingent, and unpredictable ways. This research does not so much invalidate the findings of critics as shift the focus and the standards by which rights strategies are to be judged. Instead of concentrating exclusively on the direct effects of court-articulated rights, the indirect and constitutive effects of rights-based action are explored and assessed. Moreover, in this research, the criterion for judging rights strategies shifts from policy efficiency to social equity and empowerment, with the salience of this latter criterion most compellingly articulated in the work of critical race and critical gender theorists. Especially in the

United States, where much of the available research has been conducted, the indirect and transformative potential of a 'politics of rights' (Scheingold, 1974) has been persuasively demonstrated. Just as rights have a constitutive impact on individual disputants, so too have the connotations of entitlement associated with legal rights had a mobilizing impact on social movements for minority inclusion, women's rights, pay equity, disability rights, and animal rights. Insofar as a need or a grievance can be articulated as a legal or constitutional right, expectations can be altered, quiescence overcome, and a politics of rights generated. Note that this research focuses very much on the discursive power of legal rights to constitute social and political life. Moreover, despite their liberal connotations and associations, rights can also be deployed on behalf of substantive agendas that span the political spectrum (Santos and Rodriguez-Garavito, 2005).

Despite all of their persuasive power, however, most scholars agree that rights claiming alone ensures very little transformative power. Nor is rights mobilization a matter of spontaneous combustion. The constitutive power of legal rights depends on deploying rights shrewdly, in conjunction with other political resources, including shrewd leadership and legal 'support networks' (Epp, 1998) working in combination with established organizations like churches and other religious associations, labor unions, NGO advocacy teams, and other kinds of social-action organizations (McCann, 1994). Even then a transformative politics does not necessarily follow. Political mobilization may, for example, generate effective countermobilization by rival or dominant groups. But as the struggles over abortion in the United States and gender violence in many parts of the world demonstrate, rights can matter, and a well-executed strategy of rights can at the very least destabilize the political arena, thus creating opportunities that were previously foreclosed. In sum, the increasing body of empirical research on rights advocacy in a variety of settings underscores the variable, contingent, and often-contradictory implications of a politics of rights (Santos and Rodriguez-Garavito, 2005).

See also: Civil Liberties and Human Rights; Civil Rights; Cultural Rights and Culture Defense: Cultural Concerns; Human Rights, Anthropology of; Human Rights, History of.

Bibliography

Atuahene, B., unpublished. The Limits of the Law: The Story of the Popela Community.
Brown, W., 1996. States of Injury: Power and Freedom in Late Modernity. Princeton University Press, Princeton, NJ.
Bumiller, K., 1988. The Civil Rights Society: The Social Construction of Victims. The Johns Hopkins University Press, Baltimore, MD.
Crenshaw, K.W., 1988. Race, reform and entrenchment: transformation and legitimation in antidiscrimination law. Harvard Law Review 101, 1331–1387.
Dworkin, R., 1977. Taking Rights Seriously. Harvard University Press, Cambridge, MA.
Englund, H., 2006. Prisoners of Freedom: Human Rights and the African Poor. University of California Press, Berkeley, CA.
Epp, C., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. University of Chicago Press, Chicago, IL.
Ewick, P., Silbey, S., 1998. The Common Place of Law: Stories from Everyday Life. University of Chicago Press, Chicago.
Feldman, E., 2000. The Ritual of Rights in Japan: Law, Society, and Health Policy. Cambridge University Press, Cambridge, UK.
Fuller, L., 1964. The Morality of Law. Yale University Press, New Haven, CT.
Glendon, M.A., 1987. Rights Talk: The Impoverishment of Political Discourse. Free Press, New York.
Hart, H.L.A., 1961. The Concept of Law. Clarendon Press, Oxford, UK.
Kant, I., 1948. The Moral Law (H.J. Paton, Trans.). Hutchinson, London.
Lovell, G., 2012. The myth of the myth of rights. In: Sarat, A. (Ed.), Studies in Law, Politics, and Society, vol. 56. Emerald Insight, Bingley, UK, pp. 1–30.
McCann, M.W., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago, IL.
McCann, M.W., Scheingold, S.A., 2012. Should we take seriously the conservative communitarian critique of rights?. In: Sarat, A. (Ed.), Studies in Law, Politics, and Society, vol. 56. Emerald Insight, Bingley, UK, pp. 199–231.
Melden, A.I., 1977. Rights and Persons. University of California Press, Berkeley, CA.
Merry, S.E., 1990. Getting Justice, Getting Even: Legal Consciousness among Working Class Americans. University of Chicago Press, Chicago.
Merry, S.E., 2006. Human Rights and Gender Violence: Translating International Law into Local Justice. University of Chicago Press, Chicago.
Minow, M., 1990. Making All the Difference: Inclusion, Exclusion, and American Law. Cornell University Press, Ithaca, NY.
Osanloo, A., 2009. The Politics of Women's Rights in Iran. Princeton University Press, Princeton, NJ.
Santos, B., Rodriguez-Garavito, C. (Eds.), 2005. Law and Globalization from Below: Towards a Cosmopolitan Legality. Cambridge University Press, Cambridge, UK.
Scheingold, S., 1974. The Politics of Rights: Lawyers, Public Policy, and Political Change. Yale University Press, New Haven, CT.
Tushnet, M., 1984. An essay on rights. Texas Law Review 62, 1363.
Waldron, J., 1996. Rights and needs: the myth of disjunction. In: Sarat, A., Hearns, T. (Eds.), Legal Rights: Historical and Philosophical Perspectives. University of Michigan Press, Ann Arbor, MI.
Williams, P., 1991. The Alchemy of Race and Rights. Harvard University Press, Cambridge, MA.
Wright-Carozza, P., 2003. From conquest to constitutions: retrieving a Latin American tradition of the idea of human rights. Human Rights Quarterly 21, 281–313.

Relevant Websites

www.amnestyusa.org/ – Amnesty International (organization).
http://www.ohchr.org/Documents/Publications/FactSheet2Rev.1en.pdf – Fact Sheet, The International Bill of Human Rights.
www.hrw.org/ – Human Rights Watch (organization).
http://plato.stanford.edu/entries/rights/ – Stanford Encyclopedia of Philosophy on Rights.
http://ccnmtl.columbia.edu/projects/mmt/udhr/index.html – The Universal Declaration of Human Rights.
http://libraryresources.unog.ch/udhr – United Nations Resource Guide on the Universal Declaration of Human Rights.
http://www.archives.gov/exhibits/charters/bill_of_rights.html – United States Bill of Rights.

Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice

Tim Blackman, The Open University, Milton Keynes, UK
Brid Featherstone, The Open University, Milton Keynes, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Uncertainty is one of the few certainties about policy and professional practice in social work, where knowledge is often imperfect and contested. Virtually every decision takes place in a context of some degree of uncertainty. This article explores differing sociological approaches to uncertainty and risk, and considers what these mean for evidence-based practice. It offers applied examples from the areas of child abuse, violence, and dementia to illustrate contemporary approaches to managing uncertainty and risk.

Sociological Approaches to Uncertainty and Risk

Uncertainty, whether due to 'known unknowns' or 'unknown unknowns,' is one of the few certainties about policy making and professional practice. Virtually every decision takes place in a context of some degree of uncertainty and therefore entails risk.

Risk is often described as having two dimensions: probability and impact (Fischhoff and Kadvany, 2011; Zinn, 2008). In risk management practice, the representation of risk along these two dimensions often uses scales such as unlikely/likely/certain for probability and negligible/moderate/critical for impact. Responses are then formulated and estimates made of how far these reduce the probability and impact of each risk. Representing risk in this way is common practice (Cabinet Office, 2010; Hopkin, 2012). However, it can create the impression of having more certainty than justified. Descriptors such as 'likely' and 'unlikely' have been shown to elicit different meanings for different people. Scores rather than descriptors are sometimes used but can give an illusory precision. Even if there is a sound basis for quantifying risk, how this is represented can alter its perception (O'Hagan et al., 2006). Lacking personal control over a risk or being unfamiliar with it generally heightens sensitivity to risk (Taylor-Gooby and Zinn, 2006). As we consider later, this is one reason why governments may overreact to risk by imposing audit and control regimes on the practice of care professionals.

In professional practice, risk is about managing uncertainty. If a risk of danger or harm materializes – nonaccidental injury to a child, for example – then professionals involved in safeguarding children are expected to account for their actions. However, by definition, in a world of probabilities the improbable does happen. Even if we can calculate how unlikely it is, doing the 'right' thing is no guarantee of a successful outcome. Similarly, a successful outcome may have nothing to do with doing the 'right' thing. Uncertainty and risk are far more than just knowledge problems in a positivist sense that the facts are 'there' to guide action, if only we can know what they are. Even psychological knowledge of how interpretations of facts are affected by preconceptions suggests there is a single reality about which we could be certain if we fully understood it, and that may not be the case.

Sociological approaches to risk and uncertainty question the notion of a single reality: that knowledge is a given with which everyone will agree as long as they have the facts, even probabilistic facts (statistical predictions from prior facts). Incomplete knowledge is often blamed for failures in care services, such as key information about a vulnerable person not being passed between agencies. This has seen responses to improve information sharing and communication across agencies and services, but these can easily stall if there are different interpretations of 'facts.' There may, for example, be very different constructions of 'harm' and 'abuse' across professional fields of practice and different cultures (Bowes and Daniel, 2010).

This is not just a question of different constructions but also, fundamentally, of discourse. Discourses are ways of thinking, communicating, and acting that impute meanings and serve particular interests. A discourse approach treats phenomena as imbued with subjective notions and shaped by social relations, especially relations of unequal power (Tulloch and Lupton, 2003). Who, for example, decides an issue is a risk? To what extent are decisions made by those affected, or made for them? How constrained are the options, who bears the risk, who decides what to do, and with what accountability?

These issues can be investigated using social theory. Macro-theory is about societies as a whole, and at this level theories of uncertainty and risk can give a perspective beyond everyday understandings and draw attention to how risk and uncertainty are manufactured by the kind of society in which people live. Middle-range theory, on the other hand, is more concerned with explaining outcomes at local or organizational level, where it can assist practical decision making. This theory can help to explain why an outcome is to be expected, or why it does not happen, such as the 'if … then' reasoning that is used to construct what are known as 'theories of change' for the purpose of evaluating interventions (Pawson and Tilley, 1997).

We will first explore some macro-theory approaches.

Risk Society: Risk as a New External Condition

German sociologist Ulrich Beck looked to new social movements in the 1990s as a way to understand 'late modernity'

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28097-5

(Beck, 1992). Beck argues that Western societies have transitioned from a dominance of tradition shaping daily lives and routines to a dominant conception of autonomous individuals freed from tradition. During much of modernity there was an orientation to progress, associated with the growth of welfare states and consumer societies, but in late modernity public concerns have turned to risk as anxieties have grown with the loss of traditional reference points and the rise of new risks borne out of modernization itself.

Rather than being empowered by what might be seen as a new freedom to decide for yourself in modern societies, Beck's argument is that people in the West have become ever more conscious of risk. This is due to the certainties of traditional beliefs having been lost and the unavoidability of having to make decisions. Decisions involve risks and, increasingly Beck argues, these risks appear harder to calculate and manage. This is because there is a new type of risk in late modern societies: nebulous, large scale, and very dangerous, created by societies themselves as individualization drives economic growth, new technologies, and social change, from global warming to family breakdown (Stalker, 2003).

In choosing social movements as his main object of study, Beck's theory took shape in a particular way, emphasizing global environmental risks. His theory is also realist: he considers that he is dealing with an objective external reality, albeit potentially interpreted in different ways. The opposite perspective, constructivism, argues that reality is entirely socially constructed and self-referential rather than externally referenced, with no single theory or explanation possible without recognizing this as relative and contestable from a different standpoint (Dybicz, 2011; Garrett, 2013; Houston, 2001).

Teenage pregnancy, for example, is a construction in these terms. It is often constructed as a 'risk' to be avoided, needing intervention with social programmes because of evidence of relatively poor outcomes for both mothers and babies, but may be constructed quite differently by mothers themselves buoyed by the experience of having a baby to love and care for (Duncan, 2007; Graham and McDermott, 2005). Social scientists have also argued that risk in teenage pregnancy is a consequence of social inequality rather than motherhood, although with a range of views from there being nothing intrinsically 'wrong' with teenage motherhood to it being a 'choice' made in the face of no positive alternatives (Harden et al., 2009).

Constructivism is sometimes parodied as an 'anything goes' theory that opens the door to all kinds of positions and beliefs with no means to decide between them, but its main contribution has been to oppose all-embracing grand theory. This is a theory that fits every social phenomenon to explanations of one kind, such as that all people are rational and narrowly self-interested actors (as in rational choice theory) or society's relationships and ideas are determined by economic forces (such as in structural Marxism) (Ritzer, 2003).

Beck's work is certainly macro-theory and has been criticized for its degree of generalization, but his interest in social movements was substantially because they opposed the 'rationality' that drives new, environmentally damaging technologies. These movements are cultural phenomena not explainable as class struggle in orthodox Marxist terms but, nevertheless, a force in society.

Beck, though, was most definitely not part of the cultural turn that saw much of the social sciences in the 1980s and 1990s shift to a relativist paradigm of constructivism. His analysis is offered as 'right' because it is argued as the best possible knowledge of what has been happening in society. However, terms such as 'right,' 'wrong,' and 'choice' are loaded with overtones and connotations. Who is to say what knowledge is right or wrong, or that a choice with its attendant risk is a real or forced choice? Who is responsible for a baby failing to thrive: the teenage mother, the preventative services that failed to intervene successfully, or the society that allows the obvious deprivation that makes it so much more likely that the teenage mother is poor? (Stalker, 2003).

This brings us to the second well-known theory of risk and uncertainty: governmentality.

Governmentality: Risk As a New Internal Condition

Governmentality is a concept from the French social theorist Michel Foucault's work on power and discourse (Foucault, 1991). While Beck emphasizes the mass protest and alternative movements that have mobilized against the risks of late modernity, governmentality sees attitudes to risk as embedded in the norms and practices of society, so that conformity is encouraged and nonconformity stigmatized and even demonized.

Governmentality is about how people become governed by internalizing the instrumental purposes of governing elites into their own subjective rationalities and self-conduct. The 'right' decisions are reinforced and rewarded psychologically by affirming norms, such as young people delaying childbirth until economically self-supporting. Deviant behaviors are constructed as 'other' and dangerous because they infringe the social rules on which social order depends, labeled by pejorative terms such as 'welfare dependency' (Spicker, 2008).

This is most obvious in power strategies that aim to change people's behavior directly, however well intentioned, such as compulsory education and mandatory pension schemes, or outlawing smoking and drinking alcohol in public areas. Risk is a fundamental justification for these strategies. But it is not always governments that pursue them: recently, there has been a rise of 'technologies of the self.' An example is Google Now, a type of virtual personal assistant that anticipates personal information needs based on what Google already knows about you. The latest version reports each month how many miles you have walked or cycled, a type of ethical surveillance ('walking and cycling are good for you') that would probably see protests on the street if it was government surveillance. Yet, "with Google at the helm … resistance is minimal. We don't mind our phones spying on us …" (Morozov, 2012).

Of course, delivering and sharing information via mobile phones, increasingly with easy-to-use applications, may be empowering. Harvesting data from these interactions can also be used to predict, prompt, and recommend in useful ways. But an important feature of technologies of the self is the role of expertise in allaying anxiety about risk: "… expertise operated not through social planning, paternalism and bureaucracy, but in terms of a logic of choice … through inculcating desires for self-development that expertise itself can guide" (Rose, 1999: p. 88).

It is possible to see 'expertise' at work in a further aspect of modern society, that of calculative practice, or the identification of groups in the population with risk characteristics, increasingly termed 'predictive analytics' (Siegel, 2013). Examples range from setting insurance premiums to preventative social programs. While these approaches can be helpful, a potential problem is the uncertainty that surrounds calculations often based on population averages and relative likelihoods. This means there is a risk of false positives and false negatives: the models may either label an individual or family as needing intervention when this is not the case, or fail to identify someone needing help. Care professionals are often wary of these approaches since it is the individual case they happen to be working with that matters – whether that is a person, family, or community – and not managing people as statistical categories.

Predictive analytics works with variables, modeling relationships in a statistical space where the variables are disembodied from actual cases (so an 'average' case, for example, may be rare or nonexistent when nonaverageness is typical). To the extent that predictions work, it is because they capture a combination of circumstances in people's lives. What is important is not a statistical calculation but being able to recognize a qualitative state or set of conditions that, by combining, cause an emergent outcome. Understanding the 'why' of this causation needs tacit, intuitive, and formal knowledge of each case and context, reaching shared understandings that aim to recognize the conditions that matter and intervene to change them (Blackman, 2013).

Evidence, however, is just one of multiple streams into policy action and is rarely, if ever, unfiltered by bureaucratic, business, and political interests and considerations. We now turn to illustrate this with examples from child and domestic abuse.

Managing Abuse and Violence

A desire by the public and politicians for care professionals to have certainty and 'omnicompetence' has been evident in a range of domains but one of the most high profile areas has been social work with children who are considered to be suffering from abuse. Gilbert et al. (2011) conducted a review of developments in 10 countries (an exercise first conducted over a decade before) and noted that there had been a broadening in definitions of risk. This had been accompanied by an expansion in the numbers of children and families referred to statutory services and in receipt of services, and of children placed away from home. Developments in all 10 countries demonstrated an increase in the formal procedures regulating child protection work. A key development was that across the countries surveyed, a range of different tools and technologies had been introduced to make front-line professionals in particular more formally accountable for what they did and how they did it. This increase was evident in countries with very different welfare regimes (e.g., the United States, Germany, and Norway) and differing histories of operating along the continuum of supporting-controlling families.

Smith (2001) has explored the literature on confidence and trust in order to understand the implications of the growth in accountability measures. She notes an important distinction between confidence and trust. Confidence refers to the general sense of safety and reliability that we invest in systems, which includes having certain expectations in relation to professional roles and the regulatory frameworks governing these systems. Thus, in relation to child care services, confidence would refer to the qualifications of those working in the system, expectations about their role, and expectations about the frameworks concerning the regulation and inspection of what they do.

Trust serves as a guide to interpersonal relationships where the outcome cannot be guaranteed and, indeed, where the possibilities of disappointment and regret are always present. If activities such as social work are to be really meaningful in the sense of bringing about positive outcomes in people's lives, then trust is crucial for the following reasons. First, many of those who need services will have experienced situations where their trust was betrayed very profoundly. This might lead to the conclusion that it is better, therefore, to concentrate on developing systems based upon rights and entitlements. While rights are a vital underpinning for services, they are exercised in interpersonal encounters: services, including those based upon rights, are mediated by people. Research evidence suggests that how a service is delivered really matters in terms of whether people continue to access it. For example, young people constantly give feedback on the importance of how they are talked to by workers and whether they feel such workers are genuine or not (Featherstone and Evans, 2004). Secondly, developing a trusting relationship with another can help repair the damage that has been done by earlier disappointing relationships. Furthermore, social workers can demonstrate that to trust is not always foolish and that change is possible.

Trust and confidence are related but not the same and, moreover, systems that are only focused upon confidence building can destroy the possibilities for developing the kind of trusting relationships we have described above. Under many of the audit arrangements set up to make child protection more accountable, such as those introduced by governments in England over the decade from the late 1990s, strict guidelines were set down on how long it should take to do an assessment and how the assessment should be carried out, using prescriptive computerized programs. However, ethnographic research found a key casualty of this approach that was the amount of time spent in face-to face contact with service users. Researchers demonstrated that the audit tail was well and truly wagging the practice dog as social workers prioritized inputting data into computers over visiting families and spending time with them in order to explore uncertain and fluid situations (White et al., 2010). Moreover, it became apparent that too much attention was paid to measuring the time it took to do certain tasks but not to exploring how the task was done.

How the task is done is crucial. It is important to be able to measure how many children attend the meetings that are held to discuss their care, but it is just as important, if not more so, to devise meaningful measures that assess their level of participation and how they feel about the quality of those meetings. It may be in these circumstances that simply designing more and more individualized performance measures is not the answer, but rather developing good quality, well-resourced consultation mechanisms with children and young people.

The implications for practitioners of an agenda with its focus on standardization, e-technology and performance management was to receive sustained critical scrutiny in the wake of the death of baby Peter Connelly in Haringey, London in 2007, as research evidence emerged of social workers tethered to their computer stations with direct contact with families a precarious, time limited project. A government commissioned review by Professor Eileen Munro urged a paradigm shift from a risk averse to a risk sensible approach and, in that context, the promotion of practices that involve the exercise of professional judgment rather than compliance with prescription (Munro, 2011). She noted the importance of recognizing that risk can never be eliminated, that uncertainty lies at the heart of human existence and that, when deaths happen, it is not always the case that someone is to blame. It is not possible to know for sure what is going on in the privacy of family life, nor can we predict with certainty what will happen. A fundamental change needed was realistic expectations of how well professionals can protect children and young people.

Munro (2011) concluded that the demands of bureaucracy had reduced practitioner capacity to work directly with children, young people, and families. Services had become so standardized that they did not provide the required range of responses to the variety of need presented. She urged the importance of moving from a compliance culture to a learning culture. She also recommended that services be redesigned using 'sociotechnical' principles, with due attention to the role of professional judgment. Such established principles have been used to good effect in health care systems, exemplified in the patient safety movement (Wachter, 2010).

The review built upon and promoted an interest in thinking about the role of systems design in supporting practitioners in areas where human judgments have to be made, where certainty is absent and the stakes are high. An examination of the literature on human factors in decision making shows that the intrinsic characteristics of information processing by human beings operate as both friend and foe in social work decision making (White, 2009; White et al., 2010). At an individual level, human beings are equipped with an innate apparatus to assess our fellow human beings on an intuitive and emotional level, and alongside this we have particular cognitive biases. The generation of hypotheses about situations is affected by our cognitive capacities in two principal ways: it is limited by what is available in memory and by 'psychological commitment' to the first hypothesis. This is confounded by the related tendency to seek out evidence that confirms a hypothesis, rather than searching for disconfirming evidence. Thus, once we have settled on an interpretation of events we tend to deviate little from our initial hypothesis. When we add to the equation the social, psychological, and sociological dimensions, which generate powerfully normative cultural practices, we have a heady cocktail indeed (White, 2009). Practitioners need to have the time, space, argumentative flexibility, analytic ability, and trusting relationships to debate and make sense of what is seen and recorded (Featherstone et al., 2012b).

Current developments in policy and practice underscore the difficulties of embracing uncertainty and the ubiquity of the search for certainty (Featherstone et al., 2012b). Child welfare policies in relation to early intervention are increasingly underpinned by a neuroscientific argument, which has been widely critiqued from within neuroscience itself but is often unchallenged in current policies (Wastell and White, 2012). Read carefully, the original neuroscience literature shows that the infant brain has quite remarkable resilience and plasticity when exposed to ordinary patterns of 'chaotic' neglect usually seen in the population referred to children's social care services. If changes to the brain were the criterion for removal from parents, very few children would be removed. Yet, the rhetorical potency of the 'now or never' argument is so great that it is supporting a drive toward early removal of children from their parents and has become a powerful professional mantra in England.

Many, if not most, of the measures introduced in child protection over recent decades are central to more 'top-down' attempts to ensure certainty in an area subject to public scrutiny. However, it is instructive to look at another area, that of working with men who are domestically violent, to see 'bottom up' demands for certainty. Here the imperative appears linked to historic concerns about women's and children's safety and the best way of ensuring such safety.

There is a considerable history of developing programs that work with men who are violent to women in intimate relationships (Rivett, 2010). These have usually been called perpetrator programs and have their roots in both the therapeutic, antisexist men's movement and the women's refuge movement in the United States (Featherstone et al., 2007). Historically, it is the latter that emerged to set standards for treatment and safety as a result of concerns that those which had a more therapeutic focus were in danger of excusing men's behavior. Moreover, they were considered to be too isolated from mainstream services and, therefore, unable to ensure the safety of women and children. The Duluth program emerged over time as the foremost program reflecting a feminist perspective on the causes of violence as rooted in men's control and power over women in a patriarchal society and masculine socialization practices. It had its origins in community reaction to the murder of a woman in a specific area in the United States.

Currently, programs consist of a set format where power, control, and equality issues are systematically addressed and where cognitive-behavioral therapies are used. This is the model that has been supported by UK governmental guidance and is the regulated program for criminal justice settings. Practitioners within the field have created a charity called Respect to lay down standards of good practice and to accreditation.

Over the years, Duluth programs have been subject to a number of criticisms. Their set format is considered to be too prescriptive and insufficiently sensitive to the differing needs of men attending them (Rivett, 2010). An allied critique has contested the underlying theoretical approach as it assumes singular explanations for why men are violent and relies on cognitive-behavioral approaches. Some have argued for psychosocial approaches that engage with unresolved childhood pain and trauma. A range of writers have suggested the importance of recognizing that not all violent men are the same and that not all violence is the same. Moreover, it is argued that the role played by factors such as mental health difficulties and

substance misuse need more consideration than that found in Duluth. However, contemporary debates reveal considerable anxieties about moving away from a set format (Featherstone and Fraser, 2012). The tensions and anxieties here are linked to long-standing debates within feminist and profeminist scholarship and practice, and the continued difficulties with moving away from dichotomizing and splitting, especially in relation to understanding and working with men's violence. Singular explanations and approaches seem to offer certainty in a field where engaging with diversity and complexity has been equated with excusing men's violence and therefore jeopardizing women's safety. The evidence base remains underdeveloped, and there is not a straightforward recourse to evidence to settle arguments.

Evidence-Based Practice

The 'objectivity' of predictive models is attractive when it comes to legitimizing a professional decision for which an individual practitioner may be held to account and asked for more justification than intuition or tacit know-how. Calculative technology brings risk into being and creates types of risk discourse, notably that of 'evidence-based practice.' The promise of evidence-based practice has been to create and act upon explicit knowledge for doing the right things right, based on robust findings from rigorous research, which has usually meant quantitative research because of its claims to reliability and generalizability (Blackman, 2013). The cultural approach of many valid types of knowledge and rationality means that every decision is a risk because an optimal decision is impossible without objective reference points. Evidence-based practice holds the promise of a single rationality where risk is at worst a calculated risk.

The complexity of the social world and human behavior often confound this approach. There is simply not enough knowledge to understand what will happen when intervening in complex, context-specific situations that research designs such as randomized controlled trials struggle to simulate. We have noted this complexity already with our examples of child protection and domestic violence. Another example is dementia care: what can an aspiring evidence-based practitioner make of the following account of what we understand about dementia wandering?

> (T)he same behaviour or type of wandering might occur for different reasons in different individuals; any theoretical formulation is going to have to allow for different triggers for the behaviour and so to get a "one size fits all" kind of explanation is unlikely. (Hermans et al., 2007)

The danger with this kind of uncertainty is that we resort to risk minimization, such as keeping people with dementia indoors and 'safe.' Dementia is still often constructed as an illness rather than a disability (Loveday, 2013). If framed as a disability, the focus is commonly on the impaired individual rather than the risks created by their environment or the behavior of others around them. Knowing the person and how they interact with the world around them is the route to understanding what is 'right' in a given set of circumstances, discussed and agreed between all parties, and in fact few decisions can be based on the generalizations and predictions of quantitative research. At best, all that may be available is trial and error as different approaches are tried out.

This chimes with the views of many advocates of qualitative methods, but while the case-based approaches of qualitative research with their rich descriptions can often give insight into *how* something happens, they are much more limited in answering *why* questions because these need a systematic comparison of cases. Methods are available for doing this, such as qualitative comparative analysis, which groups cases into sets of common attributes associated with particular outcomes (Ragin, 2000). The method handles situations such as the same outcome resulting from different combinations of conditions, or a condition having different effects depending on its combination with other conditions. This recognizes interaction, contingency, and complexity, which conventional quantitative research often fails to do (Blackman, 2013).

Conclusion

While the cultural turn in the social sciences has been important in emphasizing subjectivity and the different rationalities by which people live their lives, its relativism is a problem for policy makers and practitioners who need to make decisions. The cultural approach encourages empathy and self-reflexivity, raising awareness among practitioners about their own position in relationships of power, and their influence on how 'reality' is constructed through the influence of their own experiences and involvement. But practitioners and those they work with also need evidence on which to base whether to do (a) or (b), and these are often binary decisions. However rich the narrative of a qualitative study or specific a statistical probability or correlation, the evidence needed is middle-range theory that can isolate, as well as possible, a few explanatory factors that explain important but bounded aspects of why an outcome is to be expected or risk managed.

This is the approach we see with realistic evaluation, based on developing theories of change (Pawson and Tilley, 1997). These set out what is thought needs to happen to achieve the outcome aimed at. If the theory is right, the outcome and intermediate signs on the way to it will be as expected. If it is wrong, the learning is recycled back to improve the theory. The approach does not solve the dilemmas and challenges posed by macro-theories such as risk society and governmentality. These remain important for the critical and reflexive thinking about practice they engender. But it does support logical 'if … then' reasoning about outcomes from intervention that makes expectations explicit. Policy and practice are not technical exercises but complex blends of facts, interpretations, opinions, and evaluations (Majone, 1989). Yet, assessments and decisions need to be articulated explicitly if learning is to be possible.

See also: Green Social Work; Marxist Approaches to Social Work; Radical Social Work; Social Work Theory.

Bibliography

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Blackman, T., 2013. Rethinking policy-related research: charting a path using qualitative comparative analysis and complexity theory. Contemporary Social Science. Journal of the Academy of Social Sciences 8 (3), 333–345.

Bowes, A., Daniel, B., 2010. Introduction: interrogating harm and abuse: a lifespan approach. Social Policy & Society 9, 221–229.

Cabinet Office, 2010. Management of Risk – Guidance for Practitioners. The Stationery Office, London.

Duncan, S., 2007. What's the problem with teenage parents? And what's the problem with policy? Critical Social Policy 27 (3), 307–334.

Dybicz, P., 2011. Anything goes? Science and social constructions in competing discourses. Journal of Sociology & Social Welfare 38 (3), 101–122.

Featherstone, B., Evans, H., 2004. Children Experiencing Maltreatment: Who Do They Turn to? NSPCC, London.

Featherstone, B., Fraser, C., 2012. Working with fathers and domestic violence: contemporary debates. Child Abuse Review 21, 255–263.

Featherstone, B., Rivett, M., Scourfield, J., 2007. Working with Men in Health and Social Care. Sage, London.

Featherstone, B., Broadhurst, K., Holt, K., 2012a. Thinking systemically – thinking politically, building strong relationships with children and families in a context of rising inequality. British Journal of Social Work 42, 618–633.

Featherstone, B., White, S., Wastell, D., 2012b. Ireland's opportunity to learn from England's difficulties? Auditing uncertainty in child protection. Irish Journal of Applied Social Studies 12 (1), 28–42.

Fischhoff, B., Kadvany, J., 2011. Risk: A Very Short Introduction. Oxford University Press, Oxford.

Foucault, M., 1991. Governmentality. In: Burchell, G., Gordon, C., Miller, P. (Eds.), The Foucault Effect: Studies in Governmentality. Harvester Wheatsheaf, London, pp. 87–104.

Garrett, P.M., 2013. Social Work and Social Theory. Policy Press, Bristol.

Gilbert, N., Parton, N., Skivenes, M., 2011. Child Protection Systems: International Trends and Orientations. Oxford University Press, Oxford.

Graham, H., McDermott, E., 2005. Qualitative research and the evidence base of policy: insights from studies of teenage mothers in the UK. Journal of Social Policy 35, 21–37.

Harden, A., Brunton, G., Fletcher, A., Oakley, A., 2009. Teenage pregnancy and social disadvantage: systematic review integrating controlled trials and qualitative studies. British Medical Journal 339, b4254.

Hermans, D.G., Htay, U.H., McShane, R., 2007. Non-pharmacological interventions for wandering of people with dementia in the domestic setting. Cochrane Database of Systematic Reviews (1).

Hopkin, P., 2012. Fundamentals of Risk Management: Understanding, Evaluating and Implementing Effective Risk Management. Kogan Page, London.

Houston, S., 2001. Beyond social constructionism: critical realism and social work. British Journal of Social Work 31 (6), 845.

Loveday, B., 2013. Leadership for Person-centred Dementia Care. Jessica Kingsley, London.

Majone, G., 1989. Evidence, Argument and Persuasion in the Policy Process. Yale University Press, New Haven.

Morozov, E., 2012. Google Should Not Choose Right and Wrong. Financial Times. http://www.ft.com/cms/s/0/ed17b556–49ee-11e2-a7b1-00144feab49a.html#axzz2GRhphOhS.

Munro, E., 2011. The Munro Review of Child Protection: Final Report. A Child Centred System. Department for Education.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, Chichester.

Pawson, R., Tilley, N., 1997. Realistic Evaluation. Sage, London.

Ragin, C., 2000. Fuzzy-set Social Science. University of Chicago Press, Chicago.

Ritzer, G., 2003. Contemporary Sociological Theory and Its Classical Roots: The Basics. McGraw-Hill, London.

Rivett, M., 2010. Working with violent male carers (fathers and step-fathers). In: Featherstone, B., Hooper, C.-A., Scourfield, J., Taylor, J. (Eds.), Gender and Child Welfare in Society. Wiley, Chichester, pp. 195–223.

Rose, N., 1999. Powers of Freedom: Reframing Political Thought. Cambridge University Press, Cambridge.

Siegel, E., 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die. John Wiley & Sons, London.

Smith, C., 2001. Trust and confidence: possibilities for social work in 'high modernity'. British Journal of Social Work 31 (27), 287–305.

Spicker, P., 2008. Social Policy: Themes and Approaches. Policy Press, Bristol.

Stalker, K., 2003. Managing risk and uncertainty in social work: a literature review. Journal of Social Work 3 (2), 211–233.

Taylor-Gooby, P., Zinn, J.O., 2006. Current directions in risk research: new developments in psychology and sociology. Risk Analysis 26 (2), 397–411.

Tulloch, J., Lupton, D., 2003. Risk and Everyday Life. Sage, London.

Wachter, R.M., 2010. Patient safety at ten: unmistakable progress, troubling gaps. Health Affairs 29 (1), 165–173.

Wastell, D., White, S., 2012. Blinded by neuroscience: social policy and the myth of the infant brain. Families, Relationships and Societies 1, 397–414.

White, S., 2009. Fabled uncertainty in social work: a coda to Spafford et al. Journal of Social Work 9, 222–235.

White, S., Wastell, D., Broadhurst, K., Hall, C., 2010. When policy o'erleaps itself: the tragic tale of the integrated children's system. Critical Social Policy 30, 405–429.

Zen, J.O., 2008. Introduction: the contribution of sociology to the discourse on risk and uncertainty. In: Zinn, J.O. (Ed.), Social Theories of Risk and Uncertainty: An Introduction. Blackwell, Oxford, pp. 1–17.

Zinn, J.O., 2008. A comparison of sociological theorizing on risk and uncertainty. In: Zinn, J.O. (Ed.), Social Theories of Risk and Uncertainty: An Introduction. Blackwell, Oxford.

Risk Assessment in Social Work

Marilyn Callahan, School of Social Work, University of Victoria, Victoria, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Risk is a ubiquitous concept used in social work and other professions, most frequently meaning the chance of reducing negative occurrences in the future. Other definitions of risk and interrogations about the impact of its use in social work are infrequent, in particular its connections with neoliberal reformation of welfare state policies and programs. Instead, assessing and managing risk has become a central activity in social work, in particular the fields of child welfare, corrections, and community care. The unintended consequences of adopting risk as a core concept in social work are explored and alternative ways of thinking based on social work principles are provided.

Introduction

The promise of risk is a compelling one. If social workers can identify individuals who are 'at risk' of having bad things happen to them or of doing harm to others then they can focus their helping efforts where they can do the most good. They can justify intruding on the lives of certain individuals even if they have not sought help because they are preventing problems from occurring.

These two dilemmas, how to be trusted helpers for individuals in need and yet at the same time serve as agents of social control, and how to move from being professionals who mop up problems after they have occurred rather than working to prevent them from happening in the first place, have bedeviled social work since its inception. It is no wonder that when risk assessment and risk management imbued the thinking of business, government, the environmental movement, and the professions during the last two decades, social work was quick to join forces.

The intent of this article is to scrutinize this promise. After years of hard work spent refining risk assessment tools and risk management approaches, have risk and the attendant practices associated with it lived up to its early expectations?

The Complexities of Risk

From the outset, risk is a tricky concept to define. Although a taken-for-granted term used in common parlance, the various and sometimes contradictory meanings of risk and theories about its function in modern society have inspired protracted debates. In many ways, its present usage, as a calculation of what could happen in the future, can be dated at least from the Renaissance and the discovery of probability laws. These laws represented a challenge to superstitious and religious explanations for seemingly random events and offered a more scientific understanding (Bernstein, 1996; Covella and Mumpower, 1995). Risk is thought to be one of the key concepts defining modernity, not just in the hope that it can bring rational understanding to apparently irrational events but its promise to predict future disasters largely based on past experience and mitigate them by controlling present behavior. Hurricane Sandy, the devastating storm that struck Manhattan and the eastern seaboard of the US in 2012, demonstrates the distance that still must be traveled between predicting disasters and managing them, if indeed this is ever possible.

Risk has always contained positive connotations. In some cases, it is similar to bravery, as in the shrewd investor who takes a calculated risk or chance to maximize his profits or the gold medal alpine skier who attains highest speeds and therefore takes the greatest risks. In this sense, taking risks is necessary to succeed. Most commonly, however, risk continues to be associated with danger and disaster, as even for those who take risks, tragedy is a distinct possibility. And eschewing risk is most often recommended. For instance, physicians tell us to reduce our weight to avoid the risk of heart disease, investment counselors tell us to put aside money for retirement to avoid the risk of poverty in old age, and weather forecasters tell us to leave our homes in the face of a storm. In fact the deluge of risk warnings, often contradictory, has sometimes weakened the credibility of the scientific knowledge and professional credibility. Yet, the power of risk remains in spite of these setbacks. Prudence in personal and organizational behavior is applauded.

Social work has similarly embraced a risk avoidance approach particularly in those fields where it is charged with protecting vulnerable individuals and those around them. The profession may be becoming so 'risk-adverse' that it is losing its autonomy and commitment to change (Changing Lives, 2006). However, among the many factors reinforcing this commitment to risk avoidance is 'the duty of care' ethos (Granich, 2012), whereby social workers make ethical commitments "to avoid acts or omissions which one can reasonably foresee would be likely to injure another; also, the duty of people in particular circumstances and occupations to protect and control others" (Australian Association of Social Workers Code of Ethics, 1999: 26).

The opening paragraph in a recent article is typical of the casual way risk language permeates social work discourses.

> Research suggests that when fathers are positively involved in the care of their children (e.g., by providing emotional support, appropriate discipline, supervision), the risk of negative youth behaviours such as delinquency are reduced. Conversely, when fathers are negatively involved with family members (i.e., substance abuse,

> mental health problems, intimate partner violence), there is increased risk of harm to the children in the home.
> Nutton, 2012, http://www.mcgill.ca/crcf/projects/researchwatch

The choice to examine fathers' behavior, its variability across temporal periods, classes and cultures, the quality of the science suggesting a link to future behavior of children, and the impact of this research on fathers and social workers, are left unexplored in these few sentences, as if such questions are unnecessary. Yet, at the core of the commitment to risk thinking is the longing to have a more scientific basis to social work practice. It is assumed that risk factors connecting one occurrence to another have been identified through rigorous research yet the existence and quality of that research may be unexamined. The ideology that rests behind these assertions is also unexplored.

The dominance of risk thinking in social work and other professions as a reflection of ideological imperatives is the subject of much theorizing. Kelly (2001) traces the connections between globalizing tendencies toward economic growth and exploitation of new global markets and the restructuring of liberal welfare governments. Neoliberalism, the political approach taken to support this direction, focuses upon facilitating the global movement of capital and producing wealth, at least for some. These goals are given priority over support and assistance to people 'in need,' and require that these individuals decrease their demands on the state and simultaneously increase their activities as producers and consumers in service to economic growth.

Rose (1996) has theorized 'risk' as one of the organizing concepts of this shift in governing. Reacting against welfare state policies, which they claim, drives budget deficits and high taxes, neoliberal regimes typically focus on shifting attention away from notions of individual need and social problems. Instead, the organizing concept of risk is used, focusing on possible future harms that individuals themselves must address, while the new task of governments is to create smaller 'risk classes' requiring state support (Rosanvallon, 2000). This fundamental shift in thinking replaces 'need' with 'risk' as the organizing concept for services.

The primary administrative strategy for accomplishing the shift from welfare state to the neoliberal agenda has been referred to as 'new managerialism,' an approach well summarized by Tsui and Cheung (2004). They list eight features of new managerial strategies, including a strong focus on the market rather than society or community; efficiency rather than effectiveness as the main criterion of success; money and contracts rather than care and concern as the 'foundation' of relationships; and standardization as the measure of quality in service provision. In this scenario, professional knowledge is subordinate to managerial knowledge, and professionals themselves are less important, since managers can carry out required tasks through delegation and standardized controls.

While neoliberal ideas have faced some challenges in recent years, their basic beliefs and legacy of managerialism remain influential in shaping human services. Good organizations are deemed to be those that adhere to sound business practices and risk management, even as the vast differences between helping individuals and making money continue uncontested.

The environmental movement has picked up this focus on risk and used it to great advantage in advancing its aims. Since 1972 when a Greenpeace mission successfully stopped nuclear testing on Amchitka Island in Alaska, environmental activists have successfully drawn attention to significant risks to the environment. Using the power of an increasingly sophisticated science and using measuring tools such as risk assessments, they have changed the way much of development now occurs and, in the process, brought credibility to the concept of risk and its importance for policy-making. Many social workers are sympathetic to the aims of the environmental movement and admire its methods.

This brief discussion of risk, a topic that has generated a great deal of theorizing, nonetheless underscores several contradictions inherent in the concept that are central to social work. The issue of risk and choice is important. Accompanying risk, thinking is the idea of choice; good citizens choose appropriate actions to manage well-known risks. Yet clients frequently have few resources to address risks and may appear careless and blame worthy as a result.

Underlying risk is fear: of the future, of personal failure, and of the repercussions of that failure. Fear of failure and the accompanying shame often further demoralizes those who already feel inadequate (Stanford, 2010). Workers also realize that they too will be blamed and shamed if they fail to assess risk thoroughly using agency provided tools. In the name of risk, a great many actions that might be questionable are deemed acceptable.

> True fear is a gift. It is a survival signal that sounds only in the presence of danger. Yet unwarranted fear has assumed a power over us that it holds over no other creature on Earth.
> de Becker, 1997

Risks are defined outside the milieu of clients and social workers as they sit down to begin work together. Instead, they have been named by others in positions of considerably more power. The cornerstone of practice, how clients define their needs, is not recognized in this discourse. Working from a risk avoidance stance, social workers often fail to recognize the risks they are asking their clients to take. Leaving a spouse or neighborhood can seem like enormous risks calling upon a great deal of fortitude to those undertaking such drastic changes but are rarely seen in that light.

Social work is founded upon a commitment to growth and change. Yet, risk calculations are often made on the basis that the past determines the future, a contradiction that can prove difficult to resolve. Overall, the possibility that risk is used to transform economic and social changes into individual problems is a process that social work must explore and resist.

In a study of social workers involved in risk assessment in child welfare, these contradictions were rarely identified (Swift and Callahan, 2009). The prospect of preventing harm to children in the messy, busy work of child protection encouraged social workers to accept risk as an uncontested concept supposedly buttressed by science. Further, most of the research on risk in social work has taken place in first world countries, primarily Europe, Canada, the United States, and Australia and remains largely accepting of current conceptions of risk and its

importance for improving practice. The small body of research that has raised critical questions about the possible impact of risk practices on increasing inequality and on blurring the importance of culture and environment is frequently unheralded. Thus notions of the importance of risk assessment instruments and risk management approaches in improving social work practice drift over to other countries where social work is an emerging profession. They may be adopted unchallenged as evidence that the social work profession is gaining ground.

This discussion of risk and its implications for social work is told as a cautionary tale. Nonetheless, it is undeniable that many social workers work in present day organizations where assessing risk and responding to these assessments is a requirement of their practice. And their clients do face risks that, if left unattended, may cause harm. Thus, all of these contradictions inherent in risk thinking may seem tolerable for social workers if assessing and managing risk proved effective in reducing incidence of harm.

Risk Assessment

> Risk assessment is inherently subjective and represents a blending of science and judgment with important psychological, social, cultural, and political factors …. Whoever controls the definition of risk controls the rational solution to the problem at hand. If risk is defined one way, then one option will rise to the top as the most cost-effective or the safest or the best. If it is defined another way, perhaps incorporating qualitative characteristics and other contextual factors, one will likely get a different ordering of action solutions. Defining risk is thus an exercise in power.
>
> Slavic, 1999: 689

Translating risk into risk assessment instruments designed to measure risk levels and thus triaging often scarce resources is a logical step and indeed these instruments have proliferated in social work, as well as other professions. A review of the literature across health and human service professions indicates that risk assessment has flourished in teaching, policing, nursing, medicine, criminology, gerontology, psychology, and psychiatry as well as social work, although studies that compare these developments across more than one profession are rare (some exceptions are Swift and Callahan, 2009; Hannah-Moffat and Maaurutto, 2003; Parsloe, 1999). Indeed, risk assessment is considered a means by which professionals coming from different orientations and organizations can work together to standardize their assessments of complex situations and 'sing from the same hymnbook.' The field of violence against women and children is a clear example where common risk assessments across professions are encouraged (Family Violence, Risk Assessment and Risk Management, 2007).

Examining the risk assessments of various professions reveals that these instruments have a great deal in common. Risk factors are identified from research and practice with a scale for judging the levels of risk accompanying each factor. Criteria for assigning a numerical score for each risk factor are also provided along with opportunities for those completing the assessment to offer comments and compute an overall score. Some of the instruments developed more recently provide a list of mitigating factors or strengths, again to be scored. Overall risk assessments appear like something of a checklist with factors and categories of factors, scores, and identified tasks for professionals to complete the assessment.

However, some significant differences exist. The requirement to complete a risk assessment for each client varies among professions and their particular field of practice. Not surprisingly, in those fields where the central task of professionals is to determine individual risks to others and the community, risk assessments appear as mandatory features of practice with more rigorous procedures attached to them. In social work, risk assessments have taken a central place in practice in corrections, community care, and child welfare (Barry, 2007). Over time, risk assessments that began as voluntary procedures may become mandatory as they become incorporated into the overall risk management strategy of an organization. In situations where risk assessments are mandatory, the process of completing these are often embedded in a computer-guided work process, where each step is required for the completion of the next and where failure to follow procedures can result in sanctions.

Most risk assessments focus upon individual characteristics, experiences, and behavior. For instance, a typical risk assessment in child welfare includes characteristics of the caregiver such as age, gender, past experiences of abuse as a child, history of mental health problems, substance use, behavior of partner in the home, and so forth. Some of these characteristics are unchangeable: age, gender, and past experiences, for instance, and some can be the focus of potential work.

Less frequently risk assessment includes environmental, economic, and social factors as these factors are seen as beyond the mandate of personal social services to address. But there are exceptions. Public health and social welfare researchers may examine the incidence of a particular phenomenon among a large population and make public policy recommendations. For instance, the PRAMS program in the United States explored the diminishing reduction in infant mortality using large population databases and identified those factors leading to increased risks of death for infants with accompanying program and policy solutions: a social risk assessment rather than an individual one. In the field of community care, environmental risk assessments may occur, such as safety audits of seniors' facilities to mitigate the occurrence of falls.

The degree to which clients participate in risk assessments also varies. For some, particularly those that are voluntary, participation may be an intrinsic part of the engagement with clients, helping them to assess and manage their own risks. However, many times clients are unaware of risk assessments. Data may be taken from their file and worker observations without their knowing, particularly if their behavior is seen to be a potential risk to others. The degree of professional preparation and training for those administering risk assessments also varies. Some worry that the prescriptive nature of risk assessment work endangers professional autonomy and may lead to the replacement of professionals with less prepared workers.

The construction of risk assessments across disciplines has taken two courses. One path, based upon actuarial thinking, uses rigorous research procedures to determine causal links between individuals, situations, and behavior. Another model,

termed consensual, focuses largely upon theories and professional practice wisdom to identify risk factors. It remains challenging to develop actuarial models in social work, although many attempts exist. The most egregious behaviors to be predicted and prevented such as suicide, serious injury, and homicide occur rarely and thus the sample size is small and includes a myriad of intervening variables (Powell et al., 2000; Caldwell, 2002; Munro and Rumgay, 2000). However actuarial models are demonstrated to be more accurate than consensual ones and there is an ongoing search to develop and refine new ones (Gambrill and Shlonsky, 2000; Baird and Wagner, 2000).

Risk assessments in social work are usually a combination of actuarial and consensual models with an emphasis on professional theories and practice wisdom (Shlonsky and Wagner, 2005). For instance, an analysis of risk assessments commonly used in child welfare in Canada and the United States indicates that three particular theories guide the selection of variables: cycle of violence, child developmental stages, and attachment theories (Swift and Callahan, 2009). While all of these theories and the research supporting them are important to social work they have nonetheless faced critical analysis for many reasons including concerns about their ethnocentric foundations.

More problematic is the accuracy of these instruments in predicting future harm to children or other individuals. Exhaustive reviews of the literature on risk assessment in various fields of practice in social work have led to very cautionary conclusions about their capacity to forecast danger (Rycus and Hughes, 2003; Barry, 2007). One recent review covering international research and literature on risk assessment in corrections, community care, and child welfare concludes:

> There is little confidence in the predictability of risk assessment tools and yet they are becoming the priority and the focus of much worker-client contact; tools thus tend to replace rather than inform professional judgement; … the relationship between worker and client is paramount to effective working and yet is being eroded by the language and politics of risk.
>
> Barry, 2007: 8

The Social Work of Risk Assessment

Social work has always been problem focussed, whereby problems bring or compel people to seek help from social workers. From its inception, social workers have focussed upon individual and social assessments as a central task in analyzing problems and commencing strategies to work with their clients. Mary Richmond, the acknowledged pioneer of social work, framed assessments as "the attempt to arrive at as exact a definition of the social situation and personality of a given client" (p. 50) through the gathering of real, testimonial, and circumstantial evidence.

Risk assessment continues this history of making assessment on the basis of evidence but with one distinctive difference: previously identified and specifically named risks of future calamities have largely replaced an assessment of the client's present problems. These risks have been arrived at outside the realm of the social worker and individual client yet significantly shape their work together. The client's definition of the problem is not a focus of the assessment and in fact clients learn to be careful in identifying their problems lest it cause their risk scores to rise. Also they must take care challenging the selected risks on the risk assessment as to do so might appear that they were questioning science and not just the worker. Contesting their scores is also difficult as they may not have access to the information upon which the judgment was made. It is a test given by an expert which they pass or fail.

One of the major gaps in the literature and research on risk assessment is the actual work involved in completing such assessments (some exceptions are Parada, 2004; Broadhurst et al., 2010; Swift and Callahan, 2009; Munro, 2010). An intensive study of this process in child welfare where risk assessments have been mandatory features of practice and heavily governed by procedures reveals that although assessing risk is the expressed aim, social workers struggled to believe in this goal in the face of challenges such as the increased time required to gather and record information, the immediate needs of the clients, the limited resources of the agency, and the rigidity of the risk assessment instrument in conveying clients situations. Although the instrument appeared as scientifically based and objective, social workers were fully aware of the subjective nature of information gathering and the assigning of scores. In particular, social workers grappled with the problems of seeing their clients in terms of the risks they present and not their strengths. Some found ways to skirt the standardized system of assessment and bring individual needs and their own assessment to the situation at hand. Managing their own cognitive dissonance was a daily feature of the work.

It is also true that social workers place value on risk assessments for a variety of reasons. Working in highly volatile and dangerous situations where their professional expertise may be ignored by others, social workers can use the power of risk assessments to gain creditability for their opinions from other professionals as well as recalcitrant clients. They can use them as an 'aide memoire' covering areas in their assessments that they may not have considered without the risk assessment at hand. Even if calamities occur, as they will, social workers can protect themselves from blame by following all risk assessment protocols (Munro, 2004).

However in the face of these significant problems with risk assessment, changes have occurred. Researchers have intensified attempts to nail down risk factors supported by quantitative methods. Risk assessments themselves have been modified to include strengths as well as risks, family assessments and immediate safety assessments. These latter developments beg the question whether 'pure' risk assessments designed to anticipate future events are fading from view and are being replaced gradually by Richmond's original vision, gaining 'a picture of the social situation and personality of the client.'

Risk Management

While a great deal of research and practice has focused upon perfecting risk assessments, very little attention has been paid to addressing risks, once identified (some exceptions are Brown, 2004; Swift and Callahan, 2009; Munro, 2010). Overall the essential value dilemma in risk management is illustrated in public debate between the Canadian Paediatric Association

and their American counterparts regarding increased screening for autism. The Canadian association members argued that it was unethical to spend more resources screening for autism when in fact treatment was uncertain and unavailable for most children already diagnosed. Further, they argued, resources would continue to be directed to assessment from treatment, an imbalance that already existed.

An in-depth study of risk management in child welfare revealed how this dilemma plays out in practice (Swift and Callahan, 2009). Social workers charged with risk reduction must address the risks identified on the assessment with the same tools they have always had at their disposal: closing the case before service, offering programs for parents and children, removing parents from the home, and/or removing children. The time-consuming process of risk assessment continued to limit the resources available for any of these options except case closure and child removal, a legislated response. Not surprisingly the selection of these two options rose markedly in Canada during the time when risk assessment was the guiding policy in child welfare.

However, upon further examination, it was also apparent that an increasing percentage of parents were 'repeaters,' those whose cases were closed after they had received service to lower their risk scores, only to be reported again. The risk assessment and management processes produced this circle. Factors on the risk assessment based upon past events score highly, for instance parents raised in foster care or parents about whom previous complaints had been made. The programs they were required to take were not necessarily related to the risks that had been identified but were simply programs usually available in child welfare. As there were limited spaces in these programs, staff members whose agencies depended upon government contracts were encouraged to demonstrate risk reduction and social workers similarly were pressured to move the case along. Hence, parents were discharged, only to be reported again. Risk scores were higher the next time and so the cycle was repeated.

Not surprisingly, parents were never sure where they were in the process but felt governed nonetheless by the specter of the social worker appearing again. They were often right.

Given the confluence of factors: the heightened attention of risk to individuals and others, the uncertainty of risk assessment instruments, the modest resources available for risk reduction and the public demand for political, organizational, and professional accountability when things go wrong, taking the safest risk management approach often appears to be the best course. This is true for child welfare, corrections, and community care, where the safest course in all these fields may be institutionalization. It is however the most expensive. This fact alone may be most significant in rethinking risk assessment.

The Future of Risk

This article began with the question about whether risk assessment has lived up to its original promise: to help social workers predict future tragedies and prevent problems and at the same time to justify intrusion into the lives of individuals on the basis of prevention. The evidence from several different jurisdictions and fields of practice indicates otherwise, at least as risk assessments are applied to individuals. Instead they have increasingly consumed resources while taking away from efforts to deal with present problems. Clients have been significantly disenfranchised in the process, sometimes having little input into their risk assessments and certainly minimal power to contest their scores. At the same time social workers are expected to reduce risks with no additional resources or technologies and face risks themselves if they fail to do so. The contradictions, always apparent in work with unwilling clients, have been exacerbated and the tried and true staples of social work: relationship and client involvement are more difficult to incorporate into practice.

The capacity of financial prediction formulas to adequately predict and respond to risk was found greatly wanting during the economic meltdown of 2008 and current promises are more modest. A recent advertisement for a prominent investment company states:

> No one can fully predict tomorrow's risks. But in today's fundamentally changed investment world, the risks are less foreseeable and more misunderstood than ever…
>
> *The New Yorker*, 3 December 2012, 53

Rather than accept assumptions that increasingly sophisticated risk assessments applied to individuals can prevent tragedies from happening, social work would be wise to reframe the message, as indeed financial institutions have done. Social workers and their clients also work in a world fraught with hazards, where resources have diminished and inequalities have been exacerbated. Predicting individual catastrophes is a far more complex task than forecasting financial disasters. Parton (1998) has long urged the need for rehabilitating concepts of ambiguity and uncertainty in social work.

While social workers may be less capable of predicting what one individual might do in the future, they have solid evidence that conditions such as poverty, poor health, limited education and opportunities, stigma and homelessness increase the likelihood that 'bad things' will happen to individuals overall. Social risk assessments can continue to illuminate these messages.

Social risk remains a useful term to mobilize responses for particular groups of individuals. Programs for such groups as those at risk of contracting diseases or dropping out of school, or giving birth to underweight babies, are easier to mount with evidence of social risk as a central rationale. There are limitations. Social risks do not necessarily translate across boundaries both in substance and the priority that any one culture will give them.

But what about individual risk assessments? The experience over the past several decades leads to clear cautions. Those instruments most reliable and valid are those with the fewest risk factors and contain items that social workers know already to be associated with possible misfortunes. Moreover, the question whether responses and resources are available to deal with risk, once uncovered, is crucial. The portability of individual risk assessments across geographical boundaries is highly questionable without intense scrutiny of the different economic, social, and historical contexts which frame them.

At the core of these cautions is the fundamental concern that the frail science of prediction is used to paper over ideological agendas that work against individual well-being, professional practice, and social justice.

It might be helpful to look beyond risk assessments at other paradigms that address clients whose destructive behavior affects themselves and others and who resist the intervention of social workers. One option is harm reduction, an approach that has gained a great deal of traction in the field of addictions, not without controversy. It refocuses the gaze from risk in the future to harm in the present. Harm includes damage to individuals as well as to their family and the community at large. The aim of intervention is to help individuals stabilize their situation by reducing their destructive behaviors. Interventions include learning about safer injection practices and general health care, offering free substances and medications, and a wide range of supports, including peer helping. It does not begin with the goal of abstinence but indeed abstinence may occur at some point. It is not necessarily time limited as stabilization of individuals may require ongoing support, although over time those being helped may become involved in harm reduction work on behalf of others. One of the major challenges of harm reduction is using strategies to reduce the stigma attached to those who misuse drugs and alcohol. All of these actions reduce costs to families and communities. Adapting harm reduction thinking to corrections, child welfare, and community care could address many of the shortcomings of risk assessment.

We began this article with the aim of examining the promise of risk assessment after two decades of ongoing research and practice. Social work took risk assessment on board energetically and without much scrutiny because it seemed to address intractable contradictions concerning helping and threatening, treating and preventing. There is a lesson to be learned from this experience concerning the necessity of critical analysis. As Theodore Parker (1810–60) once said: "Magnificent promises are always to be suspected."

See also: Environmental Risk; HIV Risk Interventions; Health Insurance: Economic and Risk Aspects; Health Risk Appraisal and Optimistic Bias; Health Risk Perception; Law and Risk; Personality and Risk Taking; Probation, Offenders and Risk; Rheumatoid Arthritis: Psychosocial Aspects; Risk Screening, Testing, and Diagnosis: Ethical Aspects; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice; Risk, Sociology and Politics of; Risk: Social Psychological Perspectives.

Bibliography

Australian Association of Social Workers Code of Ethics., 1999. http://www.aasw.asn.au.

Baird, C., Wagner, D., 2000. The relative validity of actuarial and consensus-based risk assessment systems. Children and Youth Services Review 22 (11–12), 839–871.

Barry, M., 2007. Effective Approaches to Risk Assessment in Social Work: An International Literature Review. University of Stirling, Social Work Research Centre, Stirling.

Broadhurst, K., Hall, C., Wastell, D., White, S., Pithouse, A., 2010. Risk, instrumentalism and the humane project in social work: identifying the informal logics of risk management in children's statutory services. British Journal of Social Work 40 (4), 1046–1064.

de Becker, G., 1997. The Gift of Fear. Dell, New York.

Brown, D., 2004. Working the System: Rethinking the Role of Parents and the Reduction of 'Risk' in Child Protection Work (Masters thesis). University of Victoria, Victoria, BC.

Caldwell, M., 2002. What we do not know about juvenile sexual reoffense risk. Child Maltreatment 7 (4), 291–302.

Changing Lives Report of the Recommendations Made by the 21st Century Social Work Review Group for the Future of Social Services in Scotland, 2006. Government of Scotland, Glasgow.

Covello, V., Mumpower, J., May 1985. Risk analysis and risk management: an historical perspective. Risk Analysis 29, 103–120.

Family Violence, Risk Assessment and Risk Management, 2007. Family Violence Coordination Unit, Melbourne.

Gambrill, E., Shlonsky, A., 2000. Risk assessment in context. Children and Youth Services Review 22 (11–12), 813–837.

Granich, S., 2012. Duty to Warn, Duty to Protect. The New Social Worker Online. http://www.socialworker.com.

Hannah-Moffat, K., Maaurutto, P., 2003. Youth Risk/Need Assessment: An Overview of Issues and Practices. Department of Justice, Ottawa.

Lee, S.J., Kim, J., Yaylor, C., Perron, B., 6 November 2011. Profiles of disciplinary behaviors among biological fathers. Child Maltreatment 16 (1), 51–62. http://www.mcgill.ca/crcf/projects/researchwatch.

Munro, E., 2004. A simpler way to understand the results of risk assessment instruments. Children and Youth Services Review 26 (9), 873–883.

Munro, E., 2010. Learning to reduce risk in child protection. British Journal of Social Work 404, 1135–1151.

Munro, E., Rumgay, J., 2000. Role of risk assessment in reducing homicides by people with mental illness. British Journal of Psychiatry 176, 116–120.

Nutton, J., 2012. Article Review of: A Person-Centered Approach to Understanding the Parenting Behaviours of Biological Fathers.

Parada, H., 2004. Social work practices within the restructured child welfare system in Ontario. Canadian Social Work Review 21 (1), 67–85.

Parsloe, P., 1999. Risk Assessment in Social Care and Social Work. Jessica Kingsley, London.

Parton, N., 1998. Risk, advanced liberalism and child welfare: the need to rediscover uncertainty and ambiguity. British Journal of Social Work 28, 5–27.

Powell, J., Geddes, J., Hawton, K., Deeks, J., Goldacre, M., 2000. Suicide in psychiatric hospital in-patients: risk factors and their predictive power. British Journal of Psychiatry 176, 266–272.

Rosanvallon, P., 2000. The New Social Question: Rethinking the Welfare State. Princeton University Press, Princeton.

Rose, N., 1996. Governing 'advanced' liberal democracies. In: Barry, A., Osborne, T., Rose, N. (Eds.), Foucault and Political Reason: Liberalism, Neoliberalism and Rationalities of Government. University of Chicago Press, Chicago, pp. 37–64.

Rycus, J., Hughes, R., 2003. Issues in Risk Assessment in Child Protective Services. North American Resource Centre for Child Welfare, Columbus, OH.

Shlonsky, A., Wagner, D., 2005. The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. Children and Youth Services Review 27 (4), 409–427. http://www.sciencedirect.com/science/article/pii/S0190740904002403-aff2.

Slavic, P., 1999. Trust, emotion, sex, politics, and science: surveying the risk-assessment battlefield. Risk Analysis 19 (4), 689–701.

Stanford, S., 2010. 'Speaking back' to fear: responding to the moral dilemmas of risk in social work practice. British of Journal Social Work 40 (4), 1065–1080.

Swift, K., Callahan, M., 2009. At Risk Social Justice in Child Welfare and Other Human Services. University of Toronto Press, Toronto.

Toumbourou, J.W., Stockwell, T., Neighbors, C., Marlatt, G.A., 2007. Interventions to Reduce Harm Associated with Adolescent Substance Use. The Lancet, Elsevier.

Tsui, N., Cheung, F., 2004. Gone with the wind: the impacts of managerialism on human services. British Journal of Social Work 34, 437–442.

Risk Screening, Testing, and Diagnosis: Ethical Aspects

Göran Hermerén, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After an initial clarification of the notion of the risk, a brief discussion of the distinction between testing and screening, and of the difficulties of defining 'genetic information' in a precise way follows an overview of public concerns raised by genetic testing and screening. Next section reviews various types of ethical issues in this context. The challenges are not the same for all types and purposes of such methods. The following sections therefore discuss issues raised by preimplantation diagnosis, prenatal diagnosis, prenatal and neonatal screening, presymptomatic testing, testing at the workplace and forensic uses of genetic tests. Finally some future concerns are outlined briefly.

The Human Genome Organization project has increased enormously the possibilities to predict disability or disease either for an individual or for his or her offspring. Information obtained from genetic testing raises ethical and legal issues that have been debated for several decades. New issues are emerging due to rapid advances in molecular genetics.

From an ethical point of view, risk screening, genetic testing, and diagnosis belong to the most controversial and contested areas in contemporary medicine. The new genetics will have a profound impact on our self-images and on our understanding of the relations between environment and heritage. Conceptual, normative, and empirical issues are here intertwined in a complex way. To begin with, a preliminary clarification of the key concepts will be necessary.

The expression 'risk' will here be taken to refer to an adverse or negative future event that is probable but not certain, for example, given a certain exposure. Any risk can then be analyzed in two components: the probability of occurrence of an event and the magnitude of its negativity (including kind, degree, duration, and timing) according to certain more or less explicitly stated norms and values. Risk assessment accordingly analyzes these two components, while risk management proposes strategies for handling the risk in ways that minimize the negativity and probability.

However, risks involving probabilities are not the only the sort of risks that are relevant in this context. The expression 'epistemic risk' has been coined to refer to risks due to ignorance, and in particular due to ignorance of what we are ignorant about (e.g., Sahlin and Persson, 1994). This notion is particularly relevant to many clinical applications of molecular genetics, involving unknown factors. The ways in which these risks are communicated will then be important from an ethical point of view.

Testing should be distinguished from screening.

Harper has proposed the following definition: "Genetic testing is the analysis of a specific gene, its product or function, or other DNA and chromosome analysis, to detect or exclude an alteration likely to be associated with a genetic disorder" (Harper and Clarke, 1997). Testing refers to procedures performed at the request of individuals or families, and should always be voluntary. This holds also for genetic testing of members of families known to be at high risk, such as the siblings of persons with cystic fibrosis (Wertz and Fletcher, 1989).

'Screening,' however, implies application to large population groups, to entire populations or subsets of populations (e.g., pregnant women, newborns, or job applicants). Screening is not performed at the request of individuals or families but rather based on policy or public health decisions. It may be mandatory or voluntary.

Both the expressions 'genetic information' and 'genetic disorder' are vague and ambiguous concepts (Zimmern, 1999). The same holds for the key concepts of health, disease, and illness. The advances in molecular and clinical genetics may profoundly change our understanding of the latter concepts. They may have an impact on health care policy, on the patient–physician relationship as well as the way in which access to care is allocated (Murray et al., 1996).

Genetic information differs from other private information in that it reveals information not only about a particular individual but also about that individual's blood relatives. Genetic test results may also provide information about population groups. Moreover, genetic information can be obtained from any cell in a person's body, not just by examining, for example, particular malfunctioning organs or tissues.

The concept of genetic information is difficult to define and delimit in a clear and uncontroversial way. Just by asking for the age of a person's parents and grandparents, genetic information in a certain sense can be obtained. The necessary and sufficient conditions for a disease or disorder to be 'genetic' or 'hereditary' are far from clear. If the expressions 'genetic information' and 'genetic disorder' are to appear in legislative texts, they have to be made precise.

There are many kinds of genetic disorders, including (1) chromosome changes, like trisomy 21 or Down's syndrome, (2) monogenetic disorders which depend on a mutation in one gene, like Huntington's chorea, and (3) polygenetic disorders like diabetes which are due to mutations in several genes in combination with environmental factors. They are complex and multifactorial.

These distinctions are important for several reasons. What holds for monogenetic diseases concerning genetic determinism cannot be generalized to genetic disorders of other kinds. Thus, presymptomatic testing can be done for disorders in group (2) but rarely and with much less certainty, if at all, for those in group (3).

Public Concerns

Genetic testing may be beneficial in those cases where early detection makes a difference, for example, makes it possible to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31032-7

prevent or cure the disease. The validity and the reliability of the test methods will then be of crucial importance. The background of many national screening programs, like the one launched in Ontario, Canada, in 1999, to detect risk of hereditary breast, ovarian, and colon cancers is that early detection can lead to medical interventions and surveillance. This can significantly reduce the risk of developing the disease and hence be cost-effective.

The concern expressed over genetic testing and screening, however, is that it may threaten privacy and civil rights and be the basis of genetic discrimination or stigmatization of individuals as well as of groups. Moreover, it has been feared that widespread use of genetic testing will help to foster intolerance of people with genetic variations. It may be difficult, too, if not impossible, for people with certain inherited disorders to obtain a life insurance. Finally, people may feel a pressure to be tested, as a result of cost–benefit analysis. In general, of course, the negative consequences of testing people and disclosing information about their test results to blood relatives will have to be weighed against the benefits of this, if any, for those concerned. However, this weighing operation is far from simple.

Direct-to-consumer testing (DTC) has emerged as an increasingly controversial issue. Under what conditions can it be justified? What could and should be done about different forms of DTC? Consumer protection is obviously a relevant concern. But internet sales are difficult to regulate. Hogarth et al., 2008 provide an overview of the current landscape for the developing market for DTC genetic testing, and reviews the legal, ethical, and policy issues raised by this testing.

The need to be cautious when considering and interpreting such testing has been stressed (Dandara et al., 2013). The reason is that the clinical validity and utility of genetic tests for many complex multifactorial disorders is questionable. Recommendations to protect consumers and healthcare providers are proposed by these authors. Patrinos et al. (2013) discuss a variant of the DTC business model, where private molecular genetic testing services are offered for sale over the counter by pharmacies, and the authors stress the lack of awareness on the part of both the patients and the general public with respect to the potential benefits and risks of the tests offered.

Guidelines for genetic testing and screening have been proposed in the attempt to maximize the beneficial consequences and minimize the adverse ones (by e.g., the Nuffield Council on Bioethics (1993); the Danish Council of Ethics (1993); UNESCO (1997); the Council of Europe (1996, 2008)). These guidelines include recommendations concerning genetic counseling, informed consent, disclosure to individuals and family members, confidentiality, employment, insurance, as well as public policy. The commercial availability of genetic test kits raises both operational and ethical issues, dealt within the principles and recommendations suggested by National Institutes of Health (1997). This report discusses the validity of genetic testing, deals with genetic counseling and the informed consent process, as well as recommendations on quality assurance measures for laboratories performing genetic tests and the need for a national body with the authority to review genetic testing procedures.

The future development, indicated by the increased transition from chance to choice (Buchanan et al., 2000), includes a growing multitude of choices, possibilities to test earlier, and without procedure-related risks. This raises many ethical issues, not only about who will get access to the new technologies.

Ethical Issues

The ethical issues include the risk for potential harm by false positives and false negative, and by invasive procedures, the risks for genetic discrimination, how to promote and protect important values like autonomy (respect for persons), privacy, confidentiality, justice and fair access, and more generally how to estimate, weight and balance benefits and harms. There has to be a reasonable proportion between risks and benefits, but how this is to be understood more precisely is not clear (Hermerén, 2012).

Harms and Benefits

Invasive procedures can be harmful, e.g., by causing miscarriage. But harm can also be created in other ways. The accuracy of the tests – their validity and reliability – and the predictive value of the test or the screening method is clearly an important issue in this context. To estimate the predictive value, one has to be able to compute the number of false positives (who are worried unnecessarily) and false negatives (who are lulled into false security) as well as to know about the prevalence of the genetic disorder.

Autonomy

It has often been recommended that genetic information in health care should be acquired and used "in a manner that respects the autonomy of individuals" (Institute of Medicine, 1994), but it has proved hard to explain the precise meaning of this in a noncontroversial way. The autonomy of what individuals should be respected? Sometimes respecting the autonomy of one individual may conflict with respecting the autonomy of another. The principle of reproductive autonomy is problematic in the context of prenatal testing, because of possible interference with the supposed rights of the unborn child (Haker, 2008).

Traditional ways to promote and protect autonomy include nondirective counseling and informed consent. A general problem raised by all forms of genetic testing and screening is, first, how to give nondirective information, that is, information that makes autonomous decisions possible by individuals and couples, and then to obtain consent in a way that avoids exercise of pressure.

Why is nondirectiveness an important goal? In brief, the reasons are the value of autonomy and the importance for geneticists to dissociate themselves from the dark history of genetics, from abuse of genetics for eugenic purposes. But nondirectiveness in practice is not uncomplicated, and the goal is also controversial. Hardly surprising, there is an ongoing debate on the process of genetic counseling beyond nondirectiveness (Clarke, 1994; Harper and Clarke, 1997).

Particular problems are raised by the testing of individuals or groups with reduced autonomy, like children, demented people, or immigrants who do not understand the language.

The testing of a healthy child could arise in at least three different contexts (Chadwick, 1998): predictive testing, to see if the child will develop a specific disease that runs in the family; carrier testing, to see if the child may be at risk of having an affected child; and a screening test unrelated to family history to identify children with increased susceptibility to some common disease such as diabetes or ischemic heart disease. Children are vulnerable and their parents therefore have to be involved in the decision procedure.

But how? The nature of parental consent and the degree to which children are to be involved according to their maturity need close examination. If relatives are to be informed, who should do this? The geneticist, the hospital, or those already tested? And in what way? How many relatives should be informed? How closely related to the person asking to be tested (the index person) should the relatives be for the clinical geneticist to be responsible for informing them?

Different consent models are available, each with their own pros and cons. One main alternative is called 'opt in,' which means that if you have not said yes, you have said no; another is called 'opt out,' which means that if you have not said no, you have said yes. The conditions under which any of these models may or must be used is an important ethical issue.

Informed consent has traditionally been one of the basic principles in practical medical ethics. But the ambiguity of the term 'informed consent' has been stressed by Manson and O'Neill 2007. Moreover, new developments in genetic testing challenge this principle in its traditional form. It has, for instance, been suggested that opportunistic screening presents the most serious challenge to patient autonomy in this century, particularly in prostate-specific antigen screening, newborn screening and prenatal diagnosis of maternal blood tests for fetal anomalies (Davis, 2013).

Recent developments in genomics technologies have paved the way for broad genome-wide testing for dozens of diseases at the same time. Bunnik et al. (2013a) have proposed a tiered-layered-staged model for informed consent in personal genome testing. It is suggested that this model for informed consent can help to overcome the ethical problems of information provision and informed consent in direct to consumer personal genome testing.

Whole-genome sequencing and microarray based analyses also challenge the feasibility of providing adequate pretest information and achieving autonomous decision making. It has been argued that informed consent is a prerequisite but requires a new approach (Bunnik et al., 2013b). Preliminary and general directions for the testing offer and for the informed consent process are presented in their paper.

Privacy and The Right to Know

The challenge to genetic privacy concerns the control of personal genetic information. Privacy has two dimensions, according to the Canadian Human Rights Commission (1992): protection from the intrusion of others and protection from one's own secrets. Does this entail a right to genetic ignorance, a right not to know about one's own or one's relatives' genetic risk profile? Is there a duty for geneticists or tested persons to disclose such information to those who might benefit from it?

If so, what is the basis of such rights and duties, and what are the consequences for the various people concerned of accepting or rejecting them? Opposite positions have been argued for by Rhodes (1998) and Takala and Hayry (2000). Several compromise positions have been put forward. Tested persons should be encouraged to share what they know about their genetic risk profile with their blood relatives, when this information would be of vital health interest to their relatives (Chadwick et al., 1997; Hermerén, 2000).

The right to know your genetic parents has been a lively discussed issue in recent years. A central role in the debate has been the comparison between different situations. Some have argued that the right to know the identity of the gamete donor should be extended to naturally conceived children with misattributed paternity. Routine paternity testing at birth should therefore be carried out (Ravelingien and Pennings, 2013). Others have questioned the similarities between gamete donation and misattributed paternity and argued that they should be treated differently. Underlying this debate is the contested value of genetic relatedness, on which there are many different views.

Confidentiality of Genetic Information

The confidentiality of the information needs to be protected against third party interests, including other family members, employers, and insurers. It has been suggested that genetic information should not be given to unrelated third parties, without the explicit and informed consent of the tested persons. In many countries, access to genetic information is regulated by law. The conditions or criteria under which medical confidentiality may or must be breached are bound to be controversial (Harper and Clarke, 1997).

The practical problems raised by genetic testing also have to do with how to estimate and evaluate risks, communicate them in an understandable way, and decide which risks are acceptable and which are not. This includes the risk that a person will have an increased risk for a genetic disorder and the risk that the test result is not accurate. Communicating risks is obviously an important part of the genetic counseling process.

Justice and Fair Access

Finally, there are issues of justice and fair access. The health care systems in most countries are under increasing economic pressure. Different ways of cutting costs, manage the care, and require copayment by patients have been introduced. If testing and screening is to be done on an equitable basis, it should not be available on the basis of such accidents as in which geographical region a person happens to be born.

The ethical problems outlined above arise in somewhat different ways, according to the type and purpose of testing or screening. The challenges are not the same in all cases. In particular, certain forms of screening and presymptomatic testing for late onset genetic disorders present problems from an ethical point of view. Some of the most common types of genetic testing and screening will therefore be described below along with a brief indication of the particular ethical problems they raise.

Types of Testing and Screening

Taking the chronological development of a person's life as a point of departure, rather than the historical development of the various technologies, we may distinguish between methods used before implantation of the fertilized egg in uterus (preimplantation), before the individual is born (prenatal), when the individual is newborn (neonatal), after birth but before symptoms are visible (presymptomatic) or later to find out if the person or his or her future children are at risk to develop some hereditary disease (carrier). In addition to them, there are genetic tests and screening methods for specific applications, for instance, in workplaces (susceptibility) or in potentially legal cases (forensic).

Preimplantation Diagnosis

Preimplantation diagnosis means that the diagnosis is carried out before (in connection with *in vitro fertilization*) the pre-embryo is implanted in the woman's uterus. If there were a history of sex-linked diseases in a family, it would then be possible to prevent a child with such a disease being born. Thus the resulting pregnancy could be normal and the woman would not have to abort the fetus at a later stage or give birth to a child with a very serious or lethal disease. Preimplantation diagnosis is becoming an established medical practice in many countries.

The first serious ethical problem that arises is to whom these tests should be offered. Usually, they are limited to severe illnesses and to persons in specific 'high risk' groups. But both these criteria are value laden, and the indications may be narrowed down or widened. There are also considerable variations not only between lay people but also between geneticists over the severity of different genetic conditions (Chadwick, 1998).

Preimplantation diagnosis at the beginning of the twenty-first century is thus limited to sorting out fertilized eggs carrying genes associated with severe diseases for which there is no cure at present, like Huntington's chorea and certain X-linked diseases. Concern has been expressed that the practice may be extended to diseases that are less serious (Shenfield, 1997). Obviously, there is a risk of a slippery slope in this area.

Preimplantation diagnosis also, at least theoretically, opens up possibilities for a shift of power from the woman to the geneticist, as well as for cloning and manipulation of human pre-embryos that many people find too risky or simply ethically unacceptable. In many countries, manipulation of human pre-embryos is illegal.

The technological development in this area is rapid. For instance, Harper and Harton 2010 describe the current use of array technology and make suggestions for future use of arrays in preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS). They also discuss the steps that need to be taken if array-based testing is going to prove useful.

Comprehensive techniques like microarray technology and whole genome sequencing have the potential to change the practice of PGD and PGS. Hens et al., 2013 describe an expert panel study of how to deal with the extra information these procedures yield, including which conditions to test for and who should have a final say on which embryo to select.

Prenatal Diagnosis

Prenatal diagnosis can be regarded as a special case of presymptomatic testing. By prenatal diagnosis it is possible to discern whether a fetus is at risk for various disabilities or diseases.

There are several methods for prenatal diagnosis. Some are visual – either noninvasive, like ultrasonography, or invasive, like embryoscopy or fetoscopy, which uses a camera on a needle inserted in the uterus to view the fetus. Other methods are based on the analysis of fetal tissues (like amniocentesis or chorionic villus sampling). By ultrasound tomography it is possible to create fetal images on a screen.

The distinction between invasive and noninvasive methods is important. The former can do harm during the test procedure, e.g. by causing miscarriage. From that point of view various noninvasive techniques have emerged as very promising, provided that the problem of avoiding false positive and false negative test results can be dealt with adequately. The analysis of fetal DNA in the blood of the mother is one such method that has been gaining ground in recent years.

There are several overviews and discussions of current applications of circulating fetal nucleic acids (cell-free fetal DNA) in genetic testing for different kinds of hereditary diseases, particularly in diagnosis of monogenic disorders (Ma et al., 2013). Genetic skin diseases are of varying severity and this raises issues of the responsibility of doctors and parents in the prenatal genetic testing. The complexity of the issues calls for guidelines about the best practice in this kind of testing.

Allyse et al., 2013 have proposed an ethical framework for clinicians and companies providing noninvasive prenatal testing using cell-free fetal DNA or whole fetal cells as a basis for a set of best practices for the provision of noninvasive prenatal genetic testing. The recommendations include the amendment of current informed consent procedures. The authors strongly recommend that these tests should only be available through licensed medical providers and not directly to consumers.

The indications of when prenatal diagnosis should be offered include advanced maternal age, a genetic history of abnormalities in the family, repeated miscarriages, or previous infants with birth defects. Then a prenatal diagnosis may be offered to provide information to couples about what they can expect. A normal result is found in most cases. If not, prenatal diagnosis may make it possible for them to adjust and may also influence the way in which the baby is delivered.

There has been great concern about the risk of spontaneous abortion induced by the use of invasive methods like amniocentesis or chorionic villus sampling. This is a problem one does not have with ultrasonography and other noninvasive methods. On the other hand, ultrasonography may give information that is unexpected or not wanted, for example, that some organs or limbs are malformed. With other methods, too, one may make unexpected or unwanted discoveries, for example, that the social father is not the biological father.

A well-known issue that continues to be controversial is how to deal with incidental findings in genetic testing. The new genetic technologies have made this challenge even more urgent. Incidental findings can range from identifying

increased risk for certain inherited diseases to revealing misattributed paternity. Should the tested person always be told? If not, what criteria could and should be used to single out the exceptions?

Should incidental findings discovered with whole-genome sequencing or testing be sought and reported to ordering clinicians and to patients or their surrogates? Contradictory answers to this question have been argued for in a discussion of a statement by the American College of Medical Genetics and Genomics (Green et al., 2013; Friedman Ross, 2013).

McGuire et al., 2013 also discuss the recommendations of the American College of Medical Genetics and Genomics to seek and report pathogenic mutations for a short list of carefully chosen genes and conditions. The authors focus on the ethical framework for these recommendations. They argue that standards are needed, that patient autonomy – including the right not to know – should be respected, and that children and adults should be treated equally.

All this serves to underline the necessity of careful counseling before the diagnosis, so that the woman or the couple knows what the test may show and how certain the result is. Informed consent is crucial, but the difficulties of nondirective counseling should not be underestimated.

Other ethical issues raised by this type of diagnosis are concerned with how to preserve the confidentiality of the information obtained. Since prenatal diagnostics sometimes leads to abortion, it may be controversial in the eyes of some people for the same reasons that abortion is controversial. The impact of prenatal diagnosis on our views of people with handicap has also been a matter of concern.

Prenatal and Neonatal Screening

Newborn screening involves the analysis of blood or tissue samples taken in early infancy in order to detect inherited diseases where early intervention can minimize or eliminate the risk of later serious health problems or premature death. In prenatal screening, conventional karyotyping is being replaced by genome-wide array-based techniques (A de Jong et al., 2011).

In screening, a group of individuals is examined directly with DNA, RNA, or biochemical analysis to find those who will later in life have an increased risk of developing a particular genetic disorder. A good screening method is reliable and valid, with a minimum of false positives and false negatives.

A well-known example of uncontroversial neonatal screening is screening for phenylketonuria (PKU), a genetic disease. The reason is that PKU, if undetected, can lead to serious mental retardation. Moreover, it can be discovered easily and prevented through early intervention. For many other conditions, such as cystic fibrosis and galactosemia, the benefits of early detection are more controversial. Screening programs for newborns also raise the issues of freedom of choice and informed consent by parents (Hermerén, 1999).

Newborn screening raises several ethical issues, among them the nature of informed parental consent and how it is obtained, confidentiality and privacy, and the interests of third parties, as well as issues of justice such as equal access to testing and treatment. Moreover, it is clear that more research needs to be carried out on the psychological responses to newborn screening, particularly when the health benefits are disputed (Harper and Clarke, 1997).

Recent technological developments have made it possible to collect a huge amount of information about the genetic status of the fetus. However, this information is to some extent uncertain, its meaning is variable, and its relevance to the woman or the couple contested. It has been suggested that the ethical debate should focus on the clinical context and on the ends of prenatal screening and diagnostics (Schmitz, 2013). If that is done, the ethical debate would be able to provide a more comprehensive and useful analysis of the ethical challenges presented by in particular the new technologies.

People with certain genetic disorders have had difficulties in getting a life insurance. In some countries the right of insurance companies to ask for genetic information has been regulated by law. In Norway, it is forbidden. In other countries – The Netherlands and Sweden, for instance – there has been a temporary voluntary moratorium during which the insurance companies will not ask for genetic information, provided that the life insurance asked for will not exceed a certain amount. This is still a very controversial area (Sandberg, 1996).

Presymptomatic Testing

Individuals may be tested for monogenetic disorders (of which there are many though each is fairly rare) by DNA analysis before any symptoms are shown. If an early diagnosis is of value, because the disease can be prevented, presymptomatic testing is fairly uncontroversial. If the disorder cannot be prevented, the value of presymptomatic testing is more difficult to judge. This holds also in cases where there is a method of prevention, but it is nevertheless impossible to guarantee that the person tested positively will not get the disease (e.g., breast cancer).

In addition to the obvious difficulty of informing people in such a way that they can make autonomous decisions, presymptomatic testing raises problems about how the information obtained from the tests should be handled, and who should have access to it. Likewise, it is controversial whether blood relatives should be contacted and who should contact them in that case (the person tested, the hospital, the clinic, the geneticist, etc.). Moreover, in many countries there is great concern about how to prevent certain external third parties, like employers or insurers, from getting access to information about the test result.

Carrier Screening

Carrier screening identifies individuals with a gene or chromosome abnormality that may cause problems either for offspring or for the person screened. The testing of blood or tissue samples can indicate the existence of particular genetic trait changes in chromosomes or changes in DNA that are associated with inherited diseases in asymptomatic individuals. Carrier screening exists for sickle cell anemia, for Tay Sachs disease, as well as for cystic fibrosis, Duchenne muscular dystrophy, hemophilia, and Huntington's chorea.

Carriers have one normal copy of a gene and one copy that varies from the normal gene. Since one gene is normal, the carriers ordinarily do not exhibit symptoms of a genetic

disorder. A woman who is a carrier of an X-linked disease such as Duchenne muscular dystrophy or fragile X syndrome, has a variant gene on one of their two X (sex-determining) chromosomes. Her risk of transmitting the variant gene to each child can be estimated accurately.

The purpose of carrier testing is to inform potential parents of their genetic risk profile so that they can make informed reproductive decisions. National screening programs are usually justified by cost–benefit analysis. The question is then which costs and whose costs are taken into account, and how the benefits are defined and evaluated. The importance of cultural differences as to what constitutes benefits, and how benefits are ranked, should not be underestimated.

The right to autonomous choices is an important issue also in this area. Whether the choices open to people found to carry a disease gene are acceptable or not, may depend on their life plans, their ethical views, the laws of their country, as well as their culture and social customs. Before a carrier screening program is decided on and implemented, these questions need to be considered. In practice there may be subtle coercion, even if a carrier screening program is formally voluntary.

Screening the subset of a population at risk rather than the entire population makes sense from a medical and economic point of view, but selecting target groups for such screening programs also creates ethical problems. For example, there have been in the past controversies over whether some screening programs in effect had a eugenic purpose, particularly when the diseases screened for are overrepresented in certain ethnic groups.

Testing in the Work Place

Susceptibility screening can be used as a form of presymptomatic testing, but is also used to identify workers who may be susceptible to toxic or carcinogenic substances such as benzene in their workplace (Van Damme et al., 1997; Surralles et al., 1997; Jarvholm et al., 1999). This susceptibility may lead to occupational health problems and future severe disabilities.

Van Damme et al. (1995) have proposed a conceptual model of the complex interactions between exposure, acquired and inherited susceptibility, and risk for disease. The validity of tests for determining genotype and phenotype and their relevance for the disease must be evaluated critically to provide an objective basis for ethical discussions. The acceptability of such tests is related to a number of issues which Van Damme et al. identify and discuss.

Genetic testing of employees and job applicants is controversial. The first basic question concerning testing in the work place is simple – for whose sake is the test performed? Is the test performed in order to protect persons with increased (inherited) susceptibility to hazards at the workplace? Or is the test in the interest of the employer, so that the employer can avoid improving the working conditions and avoid hiring people who would be damaged by the work they do? Here it is essential to separate the situation of the job applicant from the situation of the person already with a job. The latter could be monitored for exposures and early effects of, for example, toxic substances on the genes.

There are methods of detecting inherited predisposition to cardiovascular diseases as well as to adverse effects from otherwise well-tolerated exposures related to the job. Clearly, with widespread use of genetic testing in the workplace there is the obvious danger of discrimination and stigmatization of individuals or groups. It may be tempting for employers to use tests to select only workers who will be able to work in the present working conditions, instead of improving the working conditions. The result could be that large groups of people would be barred from the job market.

The discussion of the validity and predictive value of the methods used in genetic testing in the workplace including the ethical and societal issues raised by these methods has continued with many contributions. For instance, MacDonald and Williams-Jones 2002 discuss arguments for genetic testing the workplace, the objections to such testing, and propose pragmatic criteria, which – they argue – if met, "would make it permissible for employers to offer but not to require workplace genetic testing."

Particular attention has been focused on the risks for genetic discrimination (Krumm, 2002). This has been the central issue in the debate. Holtzman discusses the opinion by the European Group on Ethics on genetic testing in the workplace, in particular the implications of false positives and false negatives. Issues of employment in relation to genetic testing and Huntington's disease (HD) have also been discussed against the background of a widely publicized case in Germany where a woman was refused employment as a teacher because of a family history of HD (Harper et al., 2004).

Forensic Testing

A comparatively new use of genetic testing is to examine a possible genetic linkage between, for example, alleged fathers and their children, or between suspects and evidence discovered in criminal investigations, like hair, blood, sperm, saliva, or skin. The idea is to use genetic methods of analysis to clear the innocent and identify the guilty.

However, concern has been expressed that mistakes can be made. Critics argue that it is necessary to improve the quality control of the test laboratories. Another problem is the confidentiality of the DNA profiles obtained from criminal investigations and stored in national police databanks. For how long time will they be stored? In some countries there are explicit limits to how long time they can be kept. Moreover, who will have access to them while they are stored, and for what purpose?

In some countries, for instance Sweden, apparently no complete genetic analysis is carried out, and the samples are not saved. The information collected in criminal contexts is limited to what is required for a profile that is unique enough (with the exception for monozygotic twins and very close relatives). The intention is to minimize the possibility to extract general genetic information from the forensic data bases.

Concluding Remarks

Since genomic technologies are able to detect genetic variations in patients with high accuracy and at reduced cost, these technologies may change the practice of medicine in the future. But genome-wide data raise many challenges concerning how they are to be interpreted and handled.

Genetic testing and screening should be done for 'health purposes.' This is sometimes stated as a necessary condition for genetic testing (Council of Europe, 1996, 2008). Health purposes could include diagnostic and therapeutic purposes as well as medical research and health care planning. Moreover, 'health' can be defined in several ways. It is important to state more precisely what this means.

Careful counseling remains crucial, as already mentioned. The uncertainties and gaps in our knowledge referred to earlier have important implications for the information provided and the choices actually made. This applies both to monogenic and multifactorial inherited conditions. One of the present concerns is that do-it-yourself test kits for genetic disorders may be bought via the Internet or over the counter without any professional counseling to explain what the findings mean, how certain the test is, etc. There is an urgent need to find ways of regulating the economic interest of commercial biotechnological companies in order to minimize the harm of premature use and overuse of genetic testing.

A matter of controversy is the extent to which attempts to individualize risks by pharmacogenomics (personalized medicine) will save money – and if so, how much – for the health care system. This depends on how the costs are estimated. Which costs, and whose costs, are included, and how are they estimated? Clearly it will have far-reaching effects on the education of medical doctors and the training of general practitioners – costs, which are not easy to estimate. Moreover, and perhaps more important, it will also have effects on the health care system, health care insurance, and priority setting in health care (Sahlin and Hermerén, 2011).

Genetics has a somewhat dark history linked to eugenics. Eugenics can be organized by a state or a political party in power. But eugenic purposes can be also achieved by testing or screening methods in combination with social pressure to abort – and be facilitated by new technologies. Fear has sometimes been expressed that advances in molecular biology might speed up *in vitro* eugenics (Sparrow, 2013). Others (Da Fonseca, 2013) have argued that this is far away, due to the practical limitations to the technique of creating human gametes from stem cells.

See also: Eugenics as a Basis of Population Policy; Eugenics as an International Movement; Genetic Counseling: Historical, Ethical, and Practical Aspects; Genetic Screening for Disease.

Bibliography

Allyse, M.A., Sayres, L.C., Havard, M., King, J.S., Greely, H.T., Hudgins, L., Taylor, J., Norton, M.E., Cho, M.K., Magnus, D., Ormond, K.E., July 2013. Best ethical practices for clinicians and laboratories in the provision of noninvasive prenatal testing. Prenatal Diagnosis 33 (7), 656–661. http://dx.doi.org/10.1002/pd.4144.

Buchanan, A., Brock, W., Daniels, N., Wikler, D., 2000. From Chance to Choice. Genetics and Justice. Cambridge University Press, Cambridge, New York.

Bunnik, E.M., Janssens, A.C., Schermer, M.H., June 2013a. A tiered-layered-staged model for informed consent in personal genome testing. European Journal of Human Genetics 21 (6), 596–601.

Bunnik, E.M., de Jong, A., Nijsingh, N., de Wert, G.M., July 2013b. The new genetics and informed consent: differentiating choice to preserve autonomy. Bioethics 27 (6), 348–355.

Canadian Human Rights Commission. Privacy Commissioner, 1992. Genetic Testing and Privacy. Privacy Commissioner of Canada, Ottawa, ON.

Chadwick, R., 1998. The Ethics of Genetic Screening. Kluwer, Amsterdam, The Netherlands.

Chadwick, R., Levitt, M., Shickle, D., 1997. The Right to Know and the Right Not to Know. Avebury, Aldershot, UK.

Chadwick, R., ten Have, H., Husted, J., Levitt, M., McGleenan, D., Shickle, D., Wiesing, U., 1998. Genetic testing and ethics: European perspectives. Journal of Medicine and Philosophy 23, 255–273.

Clarke, A. (Ed.), 1994. Genetic Counselling. Practice and Principles. Routledge, London.

Clarke, A., Flinter, F., 1996. The genetic testing of children: a clinical perspective. In: Marteau, T.M., Richards, M.P. (Eds.), The Troubled Helix: Social and Psychological Implications of the New Human Genetics. Cambridge University Press, Cambridge, pp. 164–167.

Council of Europe, 1996. Convention on Human Rights and Biomedicine. Council of Europe, Strasbourg, France.

Council of Europe, 2008. Additional Protocol to the Convention on Human Rights and Biomedicine, Concerning Genetic Testing for Health Purposes. CETS, Strasbourg. No. 203.

Dandara, C., Greenberg, J., Lambie, L., Lombard, Z., Naicker, T., Ramesar, R., Ramsay, M., Roberts, L., Theron, M., Venter, P., Bardien-Kruger, S., 4 July 2013. Direct-to-consumer genetic testing: to test or not to test, that is the question. South African Journal of Medicine 103 (8), 510–512.

Danish Council of Ethics, 1993. Ethics and the Mapping of the Human Genome. Danish Council of Ethics, Copenhagen, Denmark.

Davis, D.S., Spring 2013. Opportunistic testing: the death of informed consent? Health Matrix Clevel 23 (1), 35–54.

Friedman Ross, L., 24/31 July 2013. Mandatory extended searches in all genome sequencing. "Incidental findings," patient autonomy, and shared decision making. Journal of American Medical Association 310 (4), 367.

da Fonseca, F.G., Ribeiro, D.M., Carvalho, N.P., Stancioli, B., 13 August 2013. Human in vitro eugenics: close, yet far away. Journal of Medical Ethics. http://dx.doi.org/10.1136/medethics-2013-101674.

Green, R.C., Lupski, J.R., Biesecker, L.G., 24/31 July 2013. Reporting genomic sequencing results to ordering clinicians. Incidental, but not exceptional. Journal of American Medical Association 310 (4), 365.

Haker, H., 2008. Etische Dimensionen in der pränatalen Diagnostik. In: Hürlimann, D.C., Baumann-Hölzle, R., Müller, H. (Eds.), Der Beratungsprozess in der Pränatalen Diagnostik. Peter Lang, Bern, pp. 43–68.

Harper, J.C., Harton, G., 2010. The use of arrays in preimplantation genetic diagnosis and screening. Fertility and Sterility 94 (4), 1173–1177.

Harper, P.S., Clarke, A.J., 1997. Genetics, Society and Clinical Practice. Bios, Oxford, UK.

Harper, P.S., Gevers, S., de Wert, G., Creighton, S., Bombard, S., Hayden, M.R., April 2004. Genetic Testing and Huntington's disease: issues of employment. Lancet Neurology 3, 249–252.

Hens, K., Dondorp, W., de Wert, G., 2013. Embryos without secrets. An expert panel study on comprehensive embryo testing and the responsibility of the clinician. European of Journal Medical Genetics 56, 67–71.

Hepburn, E., 1996. Genetic testing and early diagnosis and intervention: boon or burden? Journal of Medical Ethics 22 (2), 105–110.

Hermerén, G., 1999. Neonatal screening: ethical aspects. Acta Paediatrica (Suppl. 432), 99–103.

Hermerén, G., 2000. The right to know and the right not to know. In: Who Owns Our Genes? Danish Council of Ethics, Copenhagen, Denmark, pp. 135–157.

Hermerén, G., November 2012. The principle of proportionality revisited: interpretations and applications. Medicine, Healthcare and Philosophy 15 (4), 373–382. http://dx.doi.org/10.1007/s11019-011-9360x.

Hogarth, S., Javitt, G., Melzer, D., 2008. The current landscape for direct-to-consumer genetic testing: legal, ethical, and policy issues. Annual Review of Genomics and Human Genetics 9, 161–182.

Institute of Medicine, 1994. In: Andrew, L.B. (Ed.), Assessing Genetic Risks: Implications for Health and Social Policy. National Academy Press, Washington, DC.

Jarvholm, B., Nordstrom, G., Hogstedt, B., Levin, J.O., Wahlstrom, J., Ostman, C., Bergendahl, C., 1999. Exposure to polycyclic aromatic hydrocarbons and genotoxic effects on nonsmoking Swedish road pavement workers. Scandinavian Journal of Work Environment & Health 25 (2), 131–136.

de Jong, A., Dondorp, W.J., Timmermans, D.R., van Lith, J.M., de Wert, G.M., October 2011. Rapid aneuploidy detection or karyotyping? Ethical reflection. European Journal of Human Genetics 19 (10), 1020–1025.

Krumm, J., 2002. Genetic discrimination. Why Congress must ban genetic testing in the workplace. Journal of Legal Medicine 23, 491–521.

Ma, Y., Gong, H., Wen, Y., June 2013. Nucleic acid-based non-invasive prenatal diagnosis of genetic skin diseases: are we ready? Experimental Dermatology 22 (6), 392–395. http://dx.doi.org/10.111/exd.12156.

MacDonald, C., Williams-Jones, B., February 2002. Ethics and genetics: susceptibility testing in the workplace. Journal of Business Ethics 35 (3), 235–241.

McGuire, A.L., Joffe, S., Koenig, B.A., Biesecker, B.B., McCullough, L.B., Blumenthal-Barby, J.S., Caulfield, T., Terry, S.F., Green, R.C., May 31, 2013. Point-counterpoint. Ethics and genomic incidental findings. Science 340 (6136), 1047–1048.

Manson, N.C., O'Neill, O., 2007. Re-thinking Informed Consent in Bioethics. Cambridge University Press, Cambridge.

Murray, T.H., 1983. Warning: screening workers for genetic risk. Hastings Center Report 13 (1), 5–8.

Murray, T.H., Rothstein, M.A., Murray, R.F. (Eds.), 1996. The Human Genome Project and the Future of Health Care. Indiana University Press, Bloomington, IN.

National Institutes of Health (National Human Genome Research Institute), 1997. Promoting Safe and Effective Genetic Testing in the United States: Principles and Recommendations. The report is available on the Internet at: http://www.nhgri.nih.gov/ELSI/TFGT_final.

Nuffield Council on Bioethics, 1993. Genetic Screening: Ethical Issues. Nuffield, London.

Patrinos, G.P., Baker, D.J., Al-Mulla, F., Vasiliou, V., Cooper, D.N., 2013. Genetic tests obtainable through pharmacies: the good, the bad, and the ugly. Human Genomics 8 (7), 17. http://dx.doi.org/10.1186/1479-7364-7-17.

Ravelingien, A., Pennings, G., 2013. The right to know your genetic parents: from open-identity gamete donation to routine paternity testing. American Journal of Bioethics 13 (5), 33–41.

Rose, N.C., Dolan, S.M., October 2012. Newborn screening and the obstetrician. Obstetrics & Gynecology 120 (4), 908–917.

Rhodes, R., 1998. Genetic links, family ties, and social bonds: rights and responsibilities in the face of genetic knowledge. Journal of Medicine and Philosophy 23 (1), 10–30.

Sahlin, N.-E., Persson, J., 1994. Epistemic risk: the significance of knowing what one does not know. In: Brehmer, B., Sahlin, N.-E. (Eds.), Future Risks and Risk Management. Kluwer, Dordrecht, The Netherlands, pp. 37–62.

Sahlin, N., Hermerén, G., 2011. Personalised, predictive and preventive medicine: a decision-theoretic perspective. Journal of Risk Research. http://dx.doi.org/10.1080/13669877.2011.634524.

Sandberg, P., 1996. Genetic Information and Life Insurance: A Proposal for an Ethical European Policy. Norwegian University of Science and Technology, Department of Biotechnology, Trondheim, Norway.

Schmitz, D., 2013. A new era in prenatal testing: are we prepared? Medical Healthcare and Philosophy 16, 357–364.

Shenfield, F., 1997. Current ethical dilemmas in assisted reproduction. International Journal of Andrology 20 (Suppl. 3), 74–78.

Shenfield, F., Sureau, C., 1997a. Ethics of embryo research. In: Shenfield, F., Sureau, C. (Eds.), Ethical Dilemmas in Assisted Reproduction. Parthenon, New York, pp. 15–21.

Shenfield, F., Sureau, C. (Eds.), 1997b. Ethical Dilemmas in Assisted Reproduction. Parthenon, New York.

Sparrow, R., 2013. In vitro eugenics. Journal of Medical Ethics. http://dx.doi.org/10.1136/medethics-2012-101200. Published online First 4 April 2013.

Surralles, J., Autio, K., Nylund, L., Jarventus, H., Norppa, H., Veidebaum, T., Sorsa, M., Peltonen, K., 1997. Molecular cytogenetic analysis of buccal cells and lymphocytes from benzene-exposed workers. Carcinogenesis 18 (4), 817–823.

Takala, T., Hayry, M., 2000. Genetic ignorance, moral obligations and social duties. Journal of Medicine and Philosophy 25 (1), 107–113. Discussion 114–20.

Tamir, S., March 2013. "Obligated aliens": recognizing sperm donor's ethical obligation to disclose genetic information. Kennedy Institute of Ethics Journal 23 (1), 19–52.

UNESCO, 1997. Universal Declaration on the Human Genome and Human Rights. UNESCO, Paris.

Van Damme, K., Casteleyn, L., Heseltine, E., Huici, A., Sorsa, M., van Larebeke, N., Vineis, P., 1995. Individual susceptibility and prevention of occupational diseases: scientific and ethical issues. Journal of Occupational and Environmental Medicine 37 (1), 91–99.

Van Damme, K., Vineis, P., Sorsa, M., Casteleyn, L., 1997. Ethical issues in genetic screening and genetic monitoring of employees. Annals of the New York Academy of Sciences 837, 554–565.

Van Damme, K., Casteleyn, L., 1998. Ethical, social and scientific problems related to the application of genetic screening and genetic monitoring for workers in the context of a European approach to health and safety at work. La Medicina del Lavoro 89 (Suppl. 1).

Wahlstrom, J., 1989. Inherited mental disorders. Acta Psychiatrica Scandinavica 80 (2), 111–117.

Wertz, D., Fletcher, J., 1989. Ethics and Human Genetics: A Cross- Cultural Perspective. Springer, Berlin.

Wright, C., 2009. Cell-free Fetal Nucleic Acids for Non-invasive Prenatal Diagnosis. Report of the UK expert working group. PHG Foundation.

Zimmern, P., 1999. Genetic testing: a conceptual exploration. Journal of Medical Ethics 25 (2), 151–156.

Risk, Sociology and Politics of

Steve Yearley, University of Edinburgh, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of risk plays a central role in many contemporary policy debates, but social scientists have observed a paradoxical quality to the social treatment of risk in recent decades. On the one hand, there appears to have been a trend toward the taming of risk; understanding of probability and chance has grown, and certain forms of risk exposure (to some illnesses and natural hazards, for example) have declined. But modernist confidence about the prospects for the control of risk has been moderated and to some extent undermined by related developments. As an example, twentieth-century technologies were associated with new forms of risk, including risks to the whole ecosystem. It seemed that the forces of progress could lead to novel hazards as well as to enhanced security. Furthermore, the legal frameworks and intellectual tools developed for the regulation of risk made it possible to elaborate arguments about risk in ever more sophisticated ways, often challenging 'establishment' interpretations of safety, and thus engendering a more vivid appreciation of risk. The ubiquity of contentious claims about risk and safety has even led some scientists to characterize high modern cultures as 'risk societies.'

Risk and chance have long been understood as key elements of the human predicament, but risk awareness besets whole cultures as well as individuals. Social scientists have typically been interested in risk both from the social-psychological stance of particular actors and in terms of cultural responses to collective exposures. At one level, sociologists have discerned a twofold historical trend toward the taming of risk. Risks have been controlled both in the sense that certain forms of societal and individual exposure to hazards have been reduced, and in the sense that risk and probability have themselves become better understood and made more tractable. Nineteenth- and early twentieth-century confidence in progress and control over nature suggested both that risks were diminishing – previously fearsome diseases could be controlled, dangers to food supply countered through scientific farm management, and so on – and that better understanding of risk and chance was possible, so that one's exposure to risk could be managed and reduced. But this modernist confidence about the prospects for the control of risk has been moderated and to some extent undermined by subsequent developments. For one thing, pivotal twentieth-century technologies including the internal combustion engine and nuclear energy were associated with new forms of risk, including (through climate change) risks to the entire world ecosystem. It seemed apparent that the forces of progress could lead to novel hazards as well as to enhanced security. Furthermore, the legal frameworks and intellectual tools developed for the regulation of risk made it possible to elaborate arguments about risk in ever more sophisticated ways, ironically engendering a more vivid appreciation of risk and eroding confidence in calculations of specific risks. Finally, the complexity of social development fostered distinctively social risks, to do with crime and the increasing uncertainty of social life and the fear of surveillance. This discussion will begin by focusing on the regulatory treatment of risk before moving on to consider the broader social experience of risk.

Risk Assessment: Regulators and Risks

One systematic and extensive body of literature on risk grew out of concerns to make policy and to legislate for risk; this literature can be categorized as concerned with 'risk assessment' and with attempts to set risk assessment on an objective footing. Given the ubiquity of risks, the question for regulators was customarily thought of as: 'how safe is safe enough?' Governments and official agencies recognize that, regrettably, train crashes occur, that drivers are daily involved in automobile collisions, that leaks will sometimes happen even from well-managed chemical plants, and that farm chemicals may impact the environment, agricultural workers, or even occasional consumers. Given that no complex system can be guaranteed to be perfectly safe, the leading approach was to ask about the price of additional increments in safety. From within this consequentialist world view, risk is thought of as the mathematical product of the likelihood of a hazard occurring and the costs to which that hazard gives rise. The policy maker's art is then to minimize risk within the budget that society is willing to spend on safety.

In practical terms, this task was always much more difficult than might be supposed, and certainly more intractable than economists routinely implied. In part, difficulties arose because of the inherent problems of trying to harmonize risk data across various fields. It may be possible to compare the costs and risks of various railway signaling systems, but (in the absence of a superministry of risk reduction) transport, industrial, medical, and agricultural risks cannot be brought within the same calculus. Worse still, both the likelihood of the problem and its consequences commonly defy exact specification. While there are good data on the medical and related risks of typical motor accidents on US and European freeways, the probability and the consequences of nuclear power station incidents can only be calculated in hypothetical ways. Similarly, the risks and costs of 'mad-cow disease' (bovine spongiform encephalopathy, or BSE), an unprecedented form of infection that spread through the food chain near the close of the twentieth century, could not be subjected to the risk calculus in the standard way because we still have no agreed way of knowing who has been infected (until postmortem examination) nor how long it may take for symptoms to appear. Further, the only way to obtain quantitative measures of many risks to humans is to extrapolate from indirect observations, such as rodent bioassays, thereby introducing

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85019-9

additional sources of uncertainty into risk assessment. Without good data, all calculations will be 'rough and ready' and even relatively minor changes in key parameters can radically alter the projected figures for risk. Nor is it even clear that there is a single currency into which all sorts of harms can be converted for the purposes of cost–benefit calculations, not least because some harms may be suffered by natural environments, landscapes or more abstract entities such as biodiversity. Thus in the recent, largely (but not exclusively) European, controversy over the possible link between declining bee numbers and the agricultural use of neonicotinoid insecticides, policymakers and analysts who wished to address the issue in terms of risks faced numerous problems. It was uncertain how much of the harm to bee health was attributable to the insecticides. It was unclear what chemicals farmers would turn to if neonicotinoids were withdrawn. And it was impossible to put an agreed figure on key costs: the decline in agricultural productivity that might result from not using the insecticides, the value of the pollination services carried out by the bees that might be protected by banning neonicotinoids, and the worth of the lives of the bees themselves that might be saved.

Despite these persistent difficulties, one line of literature about risk has been devoted to exploring what are thought to be shortcomings in the public's understanding of risk. Lay people appear to make irrational and statistically unsupported assessments of the relative risks of different hazards, and to demand high levels of risk reduction without being aware, or taking due account, of the costs. In the extreme case, the public appears unwilling even to accept the cost–benefit approach, but has no systematic alternative with which to replace it. As an added complication, despite frequent public protestations of anxieties about risk, market mechanisms seem to imply that people are not necessarily very risk averse: in the automobile market, vehicles have tended to sell on aesthetic and performance criteria rather than on safety; and consumers continue to eat fatty foods and to avoid exercise, despite well publicized links to heart disease.

The apparent discrepancies between public and expert versions of risk have given rise to a series of psychometric studies (Slovic, 1992). For example, research has shown that a tendency by people to undervalue systematically the risks to which they are routinely exposed (because of their lifestyle or occupation), and to overvalue the seriousness of risks that are novel or could be catastrophic. Moreover, there appear to be significant differences between people's assessment of the risks to which they expose themselves and those to which they believe they are subjected by others. These perceptions translate into corresponding pressures in policy. People seem more likely to demand that automobiles be engineered to lower traffic fatalities than they would be to force themselves to drive those automobiles more cautiously.

The resulting tension between public and expert interpretations of risk has caused a problem for policy makers. If they base regulations on expert judgments – that is, keyed only to the statistical probability of harm – policies may be unpopular or even subverted, whereas basing policies on the public's apparent preferences threatens to make regulations arbitrary, unscientific, or too costly (Breyer, 1993). If the citizenry really is more tolerant of self-imposed risks than of risks visited on them by others, then there is an argument for reflecting this in public policy, whatever the 'actual' exposure to risk. Similarly, if, as Slovic (1992: 121) reports, the public is more concerned about certain 'dread' risks than other risks that experts hold to be of equal hazard, then maybe the cost–benefit equation has to be widened to take into account people's manifest preferences. Regulators find themselves confronting a tension familiar to liberal democracies, that between people's 'revealed' preferences and the recommendations supported by expert opinion.

Risk Expertise: The Reflexivity of Risk

These problems have become amplified in particular institutional contexts in ways that social science research has sought to explicate. Risk assessments have been developed, particularly by official regulatory agencies, so as to apply to statistically representative, or, in some other sense, 'average' people (Jasanoff, 1990). Calculations of how long it takes to evacuate an aircraft, and thus of the requisite number of emergency exits, the width of aisles and so on, are supported by evacuation trials, but these in turn depend on assumptions about the make-up of a typical passenger cohort. Similarly, arguments about exposure to pollutants have had to construct some notion of the average person as the unit for measuring the at-risk population (with, in many cases, the need to substitute the average risk animal for the average human). Yet this ideal type hardly exists in reality and may obscure threats to specific subpopulations. Women may be different from men; pregnant women are more clearly different. The young may differ from the elderly, the housebound from the active, and so on. These differences may prove to be far from 'academic' in particular contexts of exposure. For example, people who live near one of the UK's most controversial nuclear power plant sites, Sellafield, and who happen to favor a shellfish-rich diet may be exposed to a greater nuclear hazard than the average person, because of the way in which shellfish filter material from sea water. In this case, a seemingly personal behavioral choice interacts unexpectedly with other risk factors. In other cases, however, numerous background factors may act in synergy for the poor or socially disadvantaged, as in the case of American inner-city minority populations, who may be exposed to toxic pollutants from multiple sources (Bullard, 1994). In neither the UK nor the US, do standard risk-assessment methodologies, based on average characteristics, compute adequately the risks to especially vulnerable populations in particular cases. Indeed, technical risk assessment has tended to obscure the distributive implications of risk production. In this sense, risk is a social technology that reinforces existing structures of power (Winner, 1986).

Ironically, the great increase in formal assessments of risk has not so much diminished as contributed to the exacerbation of risk concerns (Beck, 1992[1986]). Given the complexities of risk assessment and the high stakes involved, it is understandable that risk assessment methodologies have been subjected to legal and other formal challenges (see Jasanoff, 1990). These proceedings have subjected risk assessments to critical deconstruction, questioning the precise basis for the choice of methodologies. Where these methods have had consequences injurious to some but not all sections of the population, the

procedure has been made to appear discriminatory. Environmental justice movements have formed to press the point that risk assessments of individual chemicals are not equivalent to a demonstration of the impact of a cocktail of pollutants. Communities experiencing extensive hazardous exposures – each element of which may be deemed not to be unduly risky on its own – argue that the overall impact of multiple sources is not well gauged by standard methodologies.

Furthermore, as Jasanoff (1990) has demonstrated, particularly in the US, the combination of adversarial cross-examination with the separation of powers and scope for judicial review of executive agencies has meant that extensive financial and intellectual resources have been directed at deconstructing risk assessments. Given all the complexities outlined above (the hypothetical nature of many risks, the difficulty of identifying an average case, the impossibility of most human experimentation, the need for surrogate measures, and so on), there has been no prospect of finding an incontestable scientific basis for defending particular risk assessments. The crowning irony has been that the language of risk has been developed in the hope of measuring and comparing risks objectively, but in practice risk assessments have repeatedly (though not, of course, always) failed to deliver a robust and compelling objectivity.

Sociological research has focused also on the strategic constructions of certainty and uncertainty associated with risk assessments. Thus, Wynne (1992) has argued that risks and probabilities are made up of many kinds of not-knowing, only a fraction of which (for example, uncertainties of extrapolation from rodent bioassays) may be acknowledged openly in formal risk assessment processes. In some cases, it is possible to establish a hierarchy of uncertainty, as between 'risks' and 'uncertainties': with risks one knows the odds, but with uncertainties only the general parameters of the problem. Most practical questions that science-in-public has to face, however, involve an additional kind of noncertainty. This Wynne terms ignorance. Ignorance refers to aspects of a problem that are consciously or unconsciously bracketed off, and commonly not further investigated. These may be issues that lie outside the disciplinary paradigms of the sciences in question and are thus in a sense necessarily, rather than perniciously, excluded from day-to-day research. Nonetheless, ignorance in this sense is a different form of not-knowing than mere uncertainty; it is not adequately captured by being treated simply as extreme uncertainty. Wynne's distinction was famously reaffirmed in a quite different context when the former US Defense Secretary Rumsfeld talked of 'known unknowns' (uncertainties) and 'unknown unknowns' (ignorance).

In principle at least, more knowledge might assist in handling these kinds of not-knowing. Uncertainties might be turned into risks. New understandings might clarify specific areas of former ignorance, though, of course, there is no prospect of ignorance being overcome in general. But, in addition, Wynne argues that there is a fourth consideration, which he styles 'indeterminacy,' resulting from 'real open-endedness in the sense that outcomes depend on how intermediate actors will behave' (Wynne, 1992: 117). In other words, the safety or reliability of systems with an organizational or human component is deeply dependent on how the systems are operated. By treating most systems as determinate, conventional risk-assessment practices typically adopt tacit, unexamined and untested sociological hypotheses about those social practices that are central to the risk-producing activities. Thus, evaluations of risks from potentially hazardous technologies do not depend only on the behavior of physical and biological systems (difficult to model though they may be), but also every bit as much on the behavior of plant managers, operators, and even those exposed.

Several insights into the sociology of expertise follow from these claims. First Wynne (1992) observes that experts, when faced with the need or chance to regulate in an area of noncertainty, are tempted to treat all forms of not-knowing as statistically tractable uncertainty, even though things of which they are ignorant cannot, by definition, be quantified. This process has been visible recently in relation to new forms of natural gas (methane) extraction, known as hydraulic fracturing or 'fracking.' Those in favor of fracking acknowledge that there are risks but argue that these risks are acceptable in the light of the benefits. Opponents of fracking tend to argue that we lack good understanding of the implications of fracking – in terms of geological settlement and contamination of ground water in the fracking process – so, the statements about risks are really claims characterized by uncertainty and ignorance (Wilber, 2012). Second, Wynne argues that the public may be significantly more expert than the scientists in relation to some of the matters covered by indeterminacy. His view, as expressed with Irwin, is that "science offers a framework which is unavoidably social as well as technical since in public domains scientific knowledge embodies implicit models or assumptions about the social world" (Irwin and Wynne, 1996: 2–3). In other words, insofar as expert assessments depend on assumptions about particular social, cultural, or occupational practices of lay groups, it is likely that these publics will be more expert in these matters than technical experts more distant from the relevant experiential insights. A clear example relates to recent attempts to manage urban air quality by using computer simulations of air pollutants. These simulations evidently depend on knowledge about the way gases move and persist in the air, but they also depend on assumptions about how potentially polluting factories and power plants, or even buses, are operated. Models routinely assume that plant managers do not exceed legal emissions-limits, but this view is not always shared by local respondents who put forward their own evidence about emissions and about the way plants are operated (Yearley, 2006).

A related body of research ties this manifestation of hazard within complex sociotechnical systems to features of organizational practice and culture. Thus, Perrow's discussion of 'normal accidents' (1984) makes the case convincingly that many hazardous occurrences are a routine (if often unpredictable) outcome of the interaction of the manifold technical and organizational components of modern power stations and production plants. Accidents are to be expected even if no particular accident is predictable. Subsequently, Vaughan applied associated insights to understanding the decision to launch the ill-fated Challenger mission (Vaughan, 1996). Of course, questions still remain about the extent to which organizational arrangements give priority to concerns for production, or for safety (Clarke, 1989), and many recent case studies show corporations placing the most value on production and profitability.

In sum, though risk discourse has been promoted by scientists and by scientific attitudes to the valuation of the natural world, it is clear that public experience of risk assessment has encouraged successful challenges to the official orthodoxy, as in the emergence of an environmental justice agenda, and that 'scientific' risk assessment has been undermined as it has been exposed (in courts and enquiries) to legal-rational reflection on the validity of the assumptions it necessarily makes. In response to such difficulties, official agencies commonly are left with no alternative but to demand 'more and better' science; yet there are few grounds for thinking that further research or rationalization will resolve the problems outlined above, particularly in innovative areas such as nanotechnology and synthetic biology where novel entities are intended to display wholly new attributes (Yearley, 2010). Ironically, social movement campaigners find themselves confronting a similar impasse. In opposing the establishment, they commonly produce counterclaims about health or environmental risks, yet find that related difficulties underpin their own risk claims. Such experiences often generate an ambivalent attitude toward scientific expertise (Yearley, 1992). These phenomena serve as a good example of what Ulrich Beck (1992[1986]) has termed "reflexive modernization." The institutions of modernity, notably the traditions of scientific analysis and legal fact-finding, have been turned upon themselves with deconstructive consequences. For Beck, this reflexive modernization is just one facet of the 'risk society' thesis, which portrays late-millennial industrialized societies as being unusually and overwhelmingly concerned with the distribution of risks and other 'bads.' In this view, social scientific interest in risk consists not so much in the study of societal responses to particular hazards; rather it is the key to characterizing the dynamics of contemporary society as a whole.

Risks and Cultures

Beck's thesis in its widest form asserts that the most modern risks (termed by others 'late' or 'high' modern (Giddens, 1991)) are reflexive in the sense that they are self-induced. Risks in the early modern period were external to the self-conscious control of social actors. Diseases would spread, bad weather would damage harvests, fires would consume urban areas as though under the influence of external, natural, or supernatural forces. Even if some of these risks were exacerbated by human interventions, the contemporary perception was that they were uncontrollable. In high-modernity, by contrast, risks such as the threat of disastrous nuclear reactor incidents are plainly the consequences of human activities (Beck, 1992[1986]). In this view, Victorian and early twentieth-century confidence about the progressive diminution of risk marks, in an ironic fashion, the transition from apparently external to societally induced risks. Scientific and technological developments are caught up in this reflexive pattern because modern risks are typically the result of technological ventures (nuclear power, ozone-depleting chemicals in the Earth's atmosphere); science and technology are involved in the cause, the diagnosis, and, with luck, eventual rectification of the problem.

Contemporary societies do still face external risks, for example, the possibility of a collision with an asteroid or other extraterrestrial body as monitored by National Aeronautics and Space Administration (NASA)'s 'Near Earth Object Program.' But the clear preponderance of contemporary risks, in the global north at least, has a strong – often dominant – human or societal component. One clear example is offered by smallpox. Previous smallpox infections had an 'external' quality. But after the disease was eradicated in 1980, the external risk effectively disappeared and inoculation programs ceased. However, the virus was maintained in at least two military research facilities, one in the USA and other in the then Union of Soviet Socialist Republics (USSR), purportedly so that research could be done on the virus to protect against 'weaponized' versions. Now, therefore, smallpox is no longer an external, but a societally mediated, risk since the threat could only come from accidental or illegal releases from these two facilities, from biological weapons produced with this virus material from the facilities, or from viruses synthesized in other laboratories. The risk is now focused on those who control the smallpox virus and much less on the viruses themselves.

Beck's overall analytical claim has been greeted with widespread enthusiasm by many social scientists, and the term 'risk society' has popularly been adopted. However, the details of his analysis have not met with the same welcome. For one thing, it is unclear how modern all present-day risks are. The incidence of mad-cow disease – eventually transmitted to humans – is thought to have arisen from the low-technology business of producing cattle feed from animal protein, specifically in the context of energy-saving, low-temperature process innovations. More significantly, Beck's favored examples – such as the risk of fall-out from the Chernobyl reactor explosion – have a primitively 'democratic' quality. On the face of it, the fall-out may descend upon the poor and wealthy alike. In that sense, the risk society is everyone's problem (Beck, 1992 [1986]). But as the environmental justice movement, particularly in the USA, has made clear, environmental 'bads' are still often distributed quite unequally along ethnic, gender, and class lines (Bullard, 1994). The risks of the risk society may not be shared as evenly as Beck implies. Finally, Beck has rather little to say about the reasons for persistent cultural differences in the salience of different kinds of risk, although it is clear, for example, that the framing of, and policy responses to, risks vary widely even between socially and economically comparable regions, such as Europe and North America (Jasanoff, 1986).

The relationship of risk and culture has been explored by the anthropologist Mary Douglas and several of her colleagues. Developing Durkheim's celebrated suggestion that religious cosmologies reflect social structures, Douglas proposed that cultural views of the characteristics of nature indicate as much about the reflected character of society as about the underlying features of nature itself. Carefree, individualistic, dynamic cultures tend to view nature as resilient and able to look after itself, while cultures which are precarious or worry about protecting their boundaries tend to view nature as fragile and in need of protection. Douglas subsequently developed this approach for the examination of risk in advanced industrial societies. A society's risk anxieties, in her view, relate as much to the cultural 'insecurities' of that society as to the actual extent of hazards (Douglas and Wildavsky, 1982). Douglas's cultural theory has proven difficult to test (but see Marris et al., 1998) and it raises a host of theoretical problems, such as the unit of cultural analysis and the sources of cultural change. However,

her approach does serve to underline the extent to which a culture's awareness of risks may be an imputation onto nature of more general cultural anxieties.

Finally, it is clear that not all modern risks arise from human interventions in nature. Though concerns about medical and environmental risks have increased conspicuously since the mid-1970s, worries about crime, immigration, internet security, financial markets and global social dislocation have also characterized the networked and industrialized world (see also Science and Social Movements). Risk anxieties can thus be inscribed not only to nature and to the human body, but also to the self-understanding of modern societies. As social life is increasingly conducted though on-line platforms, novel and unanticipated forms of risk (such as exposure to 'cyber-bullying') are created.

See also: Citizen Science; Environmental Sciences; HIV Risk Interventions; Law and Risk; Personality and Risk Taking; Probability and Chance: Philosophical Aspects; Risk: Social Psychological Perspectives; Science and Law; Science and Technology Studies: Experts and Expertise; Sexual Risk Behaviors.

Bibliography

Beck, U., 1992[1986]. Risk Society: Towards a New Modernity. Sage, London [Originally, Risikogesellschaft: Auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt am Main, 1986].

Breyer, S.G., 1993. Breaking the Vicious Circle: Toward Effective Risk Regulation. Harvard University Press, Cambridge, MA.

Bullard, R.D., 1994. Dumping in Dixie: Race, Class, and Environmental Quality. Westview Press, Boulder, CO.

Clarke, L., 1989. Acceptable Risk? Making Decisions in a Toxic Environment. University of California Press, Berkeley.

Douglas, M., Wildavsky, A., 1982. Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers. University of California Press, Berkeley.

Giddens, A., 1991. Modernity and Self-identity. Polity Press, Cambridge.

Irwin, A., Wynne, B., 1996. Introduction. In: Irwin, A., Wynne, B. (Eds.), Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge, UK, pp. 1–18.

Jasanoff, S.S., 1986. Risk Management and Political Culture. Russell Sage, New York.

Jasanoff, S.S., 1990. The Fifth Branch: Science Advisers as Policymakers. Harvard University Press, Cambridge, MA.

Marris, C., Langford, I., O'Riordan, T., 1998. A quantitative test of the cultural theory of risk perception: comparisons with the psychometric paradigm. Risk Analysis 18, 635–648.

Perrow, C., 1984. Normal Accidents: Living with High Risk Technologies. Basic Books, New York.

Slovic, P., 1992. Perception of risk: reflections on the psychometric paradigm. In: Krimsky, S., Golding, D. (Eds.), Social Theories of Risk. Praeger, London.

Vaughan, D., 1996. The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA. University of Chicago Press, Chicago.

Wilber, T., 2012. Under the Surface: Fracking, Fortunes, and the Fate of the Marcellus Shale. Cornell University Press, Ithaca, NY.

Winner, L., 1986. The Whale and the Reactor: A Search for Limits in an Age of High Technology. University of Chicago Press, Chicago.

Wynne, B., 1992. Uncertainty and environmental learning. Global Environmental Change 2, 111–127.

Yearley, S., 1992. Green ambivalence about science: legal–rational authority and the scientific legitimation of a social movement. British Journal of Sociology 43, 511–532.

Yearley, S., 2006. Bridging the science-policy divide in urban air-quality management: evaluating ways to make models more robust through public engagement. Environment and Planning C 24, 701–714.

Yearley, S., 2010. Understanding responses to the environmental and ethical aspects of innovative technologies: the case of synthetic biology. In: Groß, M., Heinrichs, H. (Eds.), Environmental Sociology: European Perspectives and Interdisciplinary Challenges. Springer, Berlin, pp. 97–108.

Risk: Social Psychological Perspectives

Dame GM Breakwell, University of Bath, Bath, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article presents current social psychological models that analyze how risk is perceived, represented, and communicated. This article focuses upon the recent interpretations of risk responses offered by social representations theory and by identity process theory. It also describes the mental models approach to risk reactions in individuals and the social amplification of risk framework for examining societal effects of actual or anticipated hazardous events.

Introduction

It is necessary to differentiate between hazard and risk even though they are often used interchangeably in ordinary conversation. Technically, a hazard is something that can precipitate undesired effects. Risk is the measure of the probability that a hazard will actually produce defined undesired effects. Risk essentially indexes the likelihood of a specified adverse occurrence. However, risk also has a second meaning in common parlance. It is used to refer to the magnitude of the detriment associated with an adverse event. Risk is used both to refer to the likelihood and the severity of an adverse effect resulting from a hazard. Take an example: a terrorist attack might be a hazard identified by a police force and they would seek to estimate the risk of the attack in terms of the chances that the attack will take place and/or the extent of the damage it might produce. The example is useful in highlighting a number of things about hazards and risk. What constitutes a hazard will depend upon your viewpoint concerning what is undesirable – the terrorists would not consider the attack itself a hazard, though the prospect of its failure would be a hazard for them. Debates about the existence (or otherwise) of a hazard are often in reality disagreements about what is or is not undesirable. The example also points to a common problem with risk – there are normally uncertainties associated with any estimates of probability of occurrence or magnitude of effect. Debates about risk levels (i.e., about the probability or magnitude of undesired outcomes) are often based upon missing, flawed, or contested information.

The ambiguities, conflicts, and complexity that surround hazards and risk make them ideal objects of social psychological examination. Both the conceptualization of the hazard and the estimation of the risk are products of social psychological processes in addition to any technical assessment. So, too, are the reactions to risk. This article describes some of the models that explain these social psychological processes. No one model alone will explain everything about representations and actions in relation to risk. Risk as an object of study requires consideration of how models interrelate and can be used in conjunction (Breakwell, 2007).

A risk representation is the product of a social psychological process in which a hazard is (1) recognized, (2) its characteristics identified, (3) the probability of its negative impacts occurring estimated, and (4) the magnitude of its effects hypothesized. A complex system of interacting elements brings about a risk representation. It is necessary first to consider the characteristics of a hazard, then to examine how its risk is conceptualized (the mental model), to go on to consider how emotions influence risk responses, and to assess the role of personal differences (particularly in cognitive style). Beyond that, it is important to explore how social influence processes inform the social representation of the risk and through this, mediated by identity processes, construct individuals' risk representations and behavioral choices. This article examines each of these elements in turn.

Characterizing the Hazard: 'The Psychometric Paradigm'

The 'psychometric paradigm' of risk perception provides one way of systematically examining the perceived characteristics of a hazard. The approach uses a variety of psychometric scaling methods (including magnitude estimation techniques, numerical rating scales, and attitude indices) to produce quantitative measures of perceived risk and perceived benefit/detriment. Typically, a large number of individuals are asked to rate a variety of hazards on a series of characteristics. Slovic (2000) and coworkers, over many studies, have shown that two orthogonal factors account for much of the variability in the correlations between ratings of hazard characteristics. The first is 'dread', which covers risk severity, uncontrollability, catastrophic impact, difficulty of prevention, plus how far its consequences are fatal, inequitable, threatening to future generations, not easily reduced, growing greater, involuntary, and personally threatening. The second is 'familiarity', which covers how far the hazard is observable, is known (to those exposed and to science), has immediate consequences and lacks novelty. This two-dimensional grid has been used extensively to compare the relative perceptions of different hazards.

Personal Variability in Risk Representations

The psychometric paradigm has offered a method for exploring how individuals perceive hazards. It is not a model that explains why the individual perceives the hazard in a particular way or why one individual differs from another in their perception of the hazard. Nor does it explain why a hazard may

be perceived differently across its life cycle. However, there are models of social psychological processes that can explain these personal (and social category) variations in hazard and risk perception.

There are significant differences between individuals and between different groups or social categories in the specific beliefs held about particular hazards. Some studies show women perceive greater risk levels than men (Slovic, 1997). Some studies indicate systematic ethnic or socioeconomic differences (e.g., Flynn et al., 1994), with people in more disadvantaged social positions typically perceiving risk to be greater. But more recent studies have suggested that these differences in risk estimates are explained better by differences in self-assessed vulnerability than merely by social category membership (Breakwell, 2007). This is true even though social category membership is frequently linked to both self-assessed and objective vulnerability.

Vulnerabilities can be psychological and/or material. Psychological vulnerabilities would include personal characteristics, such as enduring anxiety levels, cognitive style biases, and proclivity to avoid novelty. Material vulnerability would include limitations in control over exposure to the hazard, for instance, as a consequence of socioeconomic position. Vulnerability does not in itself explain all risk perception variation. For instance, differentials in knowledge will explain some differences in risk perception. But actual vulnerability, and, perhaps more importantly, subjective perception of vulnerability, can be a vital ingredient in determining the representation of a hazard that an individual or group would accept as appropriate or, indeed, help to create. For instance, the more vulnerable may generate a higher estimate of the risk of a hazard. They may do this because they hold a representation of the hazard that posits that it cannot and will not be controllable, perhaps because their own perceived vulnerability and inability to protect themselves undermines any argument that they will be protected from the hazard (Conti et al., 2011).

Perceived vulnerability, obviously, is not completely independent of the level of perceived risk. Even those who habitually regard themselves as relatively invulnerable and immune will recognize hazards that raise their vulnerability levels. So, in examining the relationship between vulnerability and risk representation one is always looking at a system with complex feedback routes.

This explanation for differences in risk perception assumes that people are actively engaged in the interpretation of hazards but they are constrained in how they do this by the psychological, interpersonal, societal, and material resources that they have available. Two psychological characteristics seem particularly relevant in shaping risk representations. The first, and most important, is self-efficacy and the second, a related but distinct concept, is locus of control. Both are related to psychological vulnerability.

Perceived self-efficacy reflects a person's belief in their capacity to perform in ways that give them control over events that affect their lives (Bandura, 1997). High efficacy is associated with greater cognitive resourcefulness, strategic flexibility, and effectiveness in managing one's social and physical environment. Belief in one's efficacy affects how much stress, anxiety, and depression is experienced in threatening or demanding situations. Those who think they can manage threats find them less aversive. Beliefs in efficacy also facilitate control over perturbing and dejecting thoughts. When faced with difficult tasks or problems, those low in efficacy focus on the difficulties they face not on the possible solutions to them. Low efficacy is associated with lower motivation to succeed and greater willingness to attribute failure to personal inadequacies.

Studies of self-efficacy in relation to risk estimation show consistent, if complex, patterns. The relationship between self-efficacy and risk taking depends on whether the risk is taken voluntarily or is imposed. In relation to voluntary risk taking, the link to self-efficacy appears to depend upon the desirability of the outcome of taking the risk. If taking a risk satisfies some desire or need, those high in self-efficacy are more likely to take the risk. For instance, with regard to high-risk sports, Schumacher and Roth (2004) found that those higher in self-efficacy took greater risks. However, if avoiding a risk serves some objective, then the highly efficacious are less likely to take the risk. For example, Siero et al. (2004) found those higher in self-efficacy were more likely to be able to quit smoking. In relation to imposed risk, the effect of higher self-efficacy is to motivate attempts to escape, avoid, neutralize, or eradicate the hazard. Floyd et al. (2000) showed that faced with a range of natural hazards, the self-efficacious would attempt adaptive self-protection.

Self-efficacy should also be linked to risk estimation. Since lower self-efficacy suggests a perceived inability to take action to protect oneself, it would be expected to be associated with higher perceived personal risk. This is supported. For example, Macintyre et al. (2004) found that males who feel self-efficacious with regard to the use of condoms perceive their personal risk of HIV infection to be lower. Other evidence suggests that efficacy may also be important in determining estimates of risk that are not personalized. Gifford and Heath (2006) found lower self-efficacy was associated with rejection of the existence of global warming and denial that it is artificial.

Locus of control (Rotter, 1966), a cognitive style bias, is also related to risk reactions. Internal locus of control (believing things happen as a result of one's own action) is associated with lower levels of risk taking in contexts where danger is salient (Miller and Mulligan, 2002) and with more adaptive precautions when the individual is at risk (McMath and Prentice-Dunn, 2005). There is also some evidence that where the hazard is subject to some degree of personal control, those with greater internal locus of control appear to be more accurate in their personal risk estimates.

Mental Models of Risk

Risk representations can be examined in a more holistic way than is offered within the psychometric paradigm. Fischhoff (1994) introduced the 'mental models' approach. A mental model usually has both an image of the hazard (its characteristics and their relationship to each other) and a story about the origin of the hazard. Sometimes it will include a narrative about the relationship between the hazard and the individual using the mental model (e.g., the extent to which the individual is at risk). The person's mental model of the hazard is elicited using a variety of techniques: think aloud protocols, recall measures, problem-solving tasks, and knowledge tests.

The methods entail significant engagement and active participation and thought from the respondents.

The mental models approach seeks to identify for a particular hazard both accurate and inaccurate beliefs that are held by a target population. These are then used as the basis for developing risk communication material that will correct misunderstandings. In this approach, the object is to bridge the gap between lay and expert models of the risk by adding missing concepts, correcting mistakes, strengthening correct beliefs, and minimizing peripheral ones. It is argued that people have an 'intuitive understanding' of risks and that they can be helped to a better appreciation of them and consequently be placed in a position to make more informed decisions if they are given new information in a format that is consistent with their initial belief system (Morgan et al., 2002). It is therefore crucial to map the initial belief system about the hazard that is the target for risk communication that is designed to change the risk representation and the behavioral decisions it services.

The mental model approach is founded upon the assumption that specific experiences and information will shape the mental model but the precise mechanisms whereby this occurs still need to be established. Experience of a hazard sometimes seems to be linked to increased complexity of the mental model. Wagner (2007), for instance, showed personal experience was associated with more complex mental models of flash floods and landslides. However, experience does not always enrich the mental model. Whitmarsh (2008) found personal experiences of environmental hazards do not affect perceptions of and behavioral responses to climate change. The role of mere experience of the hazard in shaping mental models is not understood. It seems likely that experience must be actively interpreted – given meaning – by other elements (e.g., social representational processes or identity processes) in the complex system that creates risk representations before it will be reflected in mental models.

Risk and Affect

The mental models approach is valuable because it requires the researcher to think about the complex system of interacting beliefs that underpins risk representation and decision making. It is also valuable because it highlights the significance of affect (emotions) in risk representation and response. Fischhoff (2007) suggested that risk communication can be unpersuasive when it fails to take into account the role of emotions.

The role of affect in risk representation is important (Slovic et al., 2004). This includes not just the impact of fear or anger as a response to danger (though these can be very important – see Fischhoff et al., 2005) but the subtler layers of affect that are attached to all our perceptions. Peters et al. (2006) summarize the four functions of affect in decision making (affect as information, as a spotlight, as a motivator, and as common currency). Finucane et al. (2000) argued that people use an 'affect heuristic' when making judgments. Representations of objects and events are imbued with varying degrees and forms of emotion, people allow these emotions to serve as a cue to decision making alongside any cognitive process that is involved (Slovic et al., 2005). Lowenstein and Hsee et al. (2001) found that, in a wide range of contexts, the immediate emotional reaction is more influential in determining choice than the rational evaluation of the options that may have been conducted beforehand. Peters et al. (2004) also found emotional reaction, and the strong affect-laden imagery associated with it, determines perceived risk rather than the reverse in their examination of reactions to nuclear power plants, radioactive waste, and radiation from nuclear weapons.

Predictions of some emotional reaction in the future can also influence risk representations and decisions. Regret is feeling sorry about an event or for a loss. Regret occurs after something has happened. However, risk researchers have observed that regret can work prospectively. They talk about 'anticipated regret'. People are motivated to avoid certain risks because they anticipate that they will feel regret later if they do take them. Zeelenberg (1999) showed it operates in choices concerning financial decisions, gambling, interpersonal relationships, and consumption. This work suggested that anticipated regret motivates risk avoidance but this is not mediated by heighten risk perception.

The role of affect can also be heightened or dampened. Gasper and Clore (1998) found that when people were made to pay attention to their emotional state their judgments were influenced by it, otherwise they were not, and, simply telling people to pay attention to their feelings – rather than to the facts – shifted the focus of judgments.

Social Amplification of Risk

The social psychological perspectives outlined above largely focus upon the individual in relation to the risk representation. The social amplification of risk framework (SARF) was developed specifically to explain how societal reactions to hazards occur. It describes both the social and individual factors that act to amplify or attenuate perceptions of risk and then generate secondary effects, such as regulatory changes, economic losses, or stigmatization of technologies. The SARF was proposed by Kasperson et al. (1988) and has been subsequently elaborated (Kasperson et al., 2003). The SARF proposes that hazards or risk events will have an impact not only through their primary physical effects but also, and often more importantly, through the way people and institutions (called 'stations' in SARF) communicate them to others. While 'stations' are often thought to be embodied in the media or official agencies, anyone can be one. Furthermore, transmission of signals is now multiplex; the Internet and social media sites taking an increasing role.

The act of communication requires that the risk is translated into various 'risk signals' (images, symbols, and signs) that will interact with a variety of psychological, social, institutional, or cultural processes and this will result in the intensification (amplification) or dampening (attenuation) of the perceptions of the risk and its manageability. Amplification refers to qualitative restructuring of the information about the risk – a change in its meaning or significance, not just a change of degree.

SARF is a social influence model of risk representation construction. While it has been criticized and a more differentiated taxonomy of amplification processes has been proposed (Breakwell and Barnett, 2001), it nevertheless serves

to emphasize that a hazard becomes imbued with meaning as a consequence of the system of social influence processes within which it is embedded. Moreover, SARF suggests that the consequences of the hazard are themselves in part determined by the reactions of those engaged in communicating about it. Yet, as originally proposed, SARF is descriptive rather than predictive. For instance, it will not predict under what conditions intensification or attenuation will occur.

Following studies of a series of risk events using the SARF approach, Breakwell and Barnett (2003) suggested that there were discernable regularities in the way in which amplification processes worked that could provide the basis for developing it into a more predictive model. They highlighted the roles of what they labeled 'hazard sequences', 'hazard templates', and 'hazard negotiations'.

Hazard sequences refer to a simple notion: any one risk event is often represented or conceptualized as part of a sequence of risk events that share common features. So, for instance, a train crash will be viewed within the frame of a series of other train crashes. The existence of the hazard sequence allows the new risk event to be located in terms of its severity and meaning because it offers the backcloth of knowledge about similar events in the past. Reference to other past risk events can also provide the emotional connotations necessary to achieve either intensification or attenuation.

Hazard templates are an extension of the notion of hazard sequences. When a new risk event occurs the way it will be represented will be determined by the 'templates' that already exist for the representation of that type of risk. The template will provide ready-made phrases to be used about it. It will provide common explanations for the origin of the risk event, the individuals, groups, or institutions who should be held responsible, the forms of remedial or punishment action that should be taken, etc. The form of the template is roughly common across hazards but the substantive content will differ widely. So, for instance, the content of the template for a multicar smash on the highway in fog (e.g., reckless drivers, driving too close and too fast; poor signage; overcrowded highways causing increasing occurrence; need for improved driver training and highway control; efficient response of the emergency services who worked tirelessly; etc.) is different from the content of the template for an outbreak of food poisoning in a hotel (e.g., poor hygiene, poor management, inadequate regulatory controls, increasing incidence of such cases, etc.). Nevertheless, some underlying similarities emerge: the cause, the significance, the responsibility, and the redress. The hazard template is the basis for the specific mental model of the risk. These templates are commonly used by the mass media to structure their stories but they are also the frameworks that lay people use when they come to talk about a new risk event. They appear to have a risk vocabulary and rhetoric that is built upon understandings and expectations of the way different hazards work. These understandings and expectations are shared to varying degrees.

Hazard negotiation is the concept introduced to acknowledge that understandings will not always be totally shared. When 'stations' find that they are broadcasting messages about a risk event that are at variance with each other, there is scope for them to attempt to negotiate a common interpretation or, if this proves impossible, at least to seek a compatible set of interpretations. Hazard negotiations can range from the formally structured exchanges with clear rules of engagement to the informal free for all. They are rarely, irrespective of format, conflict free. The objectives of parties in the hazard negotiation will be driven by their self-interest. To reiterate all aspects of the processes of amplification (including hazard sequences, templates, and negotiation) will play a part in determining the mental models that individuals hold about a hazard.

SARF itself is silent about the nature of the processes that determine substance, form, or acceptance of risk representations. The work on hazard sequences, templates, and negotiations goes some way to rectifying this. However, there are two other approaches that have been used: social representations theory (SRT) and identity process theory (IPT).

Social Representations of Risk

SRT (Moscovici and Farr, 1984), one of the major social psychological theories of social influence processes, can provide the basis for explaining how an individual may acquire a representation (or mental model) of a hazard. According to Moscovici, social representation operates with two prime processes: objectification and anchoring.

Objectification entails translating something that is abstract into something which is almost concrete, gaining a density of meaning that ultimately makes it a common and 'natural' part of thinking about the object. Anchoring entails categorizing a new object into preexisting cognitive frameworks in order to render them familiar – reducing the strange and unfamiliar object to the level of an ordinary object set in a familiar context.

SRT may be particularly relevant to explaining why particular mental models of hazards evolve because it was initially developed to explain what happens when people are faced with having to make sense of, or give meaning to, ideas or data that are novel to them, under conditions of uncertainty and where claims are contested. Plainly these are quintessentially characteristics of the initial interface between people and information about hazards. Indeed, it might be argued that encountering a newly identified hazard inevitably triggers social representation processes. Castro and Gomes (2005) illustrated this through an analysis of the way the media discussion of genetically modified organisms developed in the Portuguese newspapers between 1999 and 2001. They showed anchoring in prior ideological systems concerning agriculture and scientific innovation.

The processes of objectification and anchoring could account for the structure often evident in mental models of hazards held by lay members of the public. Frequently, systematic empirical investigations expose lay representations of a novel hazard that are characterized by objectification (the abstract properties of the hazard being translated into commonly understood metaphors or represented through tangible, if tangential, exemplars) and by anchoring (given meaning through association or through claiming similarity with other types of explanation that are more familiar). SRT would suggest that the representation that is constructed of any hazard will be determined by the proclivities of the individual for particular forms of

objectification and anchoring. However, SRT would emphasize that the representation produced of any hazard is not idiosyncratic or purely individualized. SRT states that objectification and anchoring are not individual processes. They are processes that normally involve social interaction and the establishment of shared meaning and consensus through communication among people.

If mental models of hazards are generated through processes of social representation, they will be substantially shared by members of discrete subcultures. This does not mean that all members of the subculture would hold absolutely identical mental models of a hazard but rather that their mental model would share certain common core elements. SRT suggests that subcultures (and social categories and groups) shape the social representations that they develop to serve their self-interests. Thus the construction of a mental model of a hazard by a subculture will be purposeful, motivated to achieve particular objectives. Willis and DeKay (2007), for instance, showed how differences in stakeholder group membership predicted ecological risk representations. Support for climate change policy is particularly affected by socioeconomic factors (Dietz et al., 2007). Of course, the media are particularly active in creating social representations of climate change and offer subcultures exposure to competing risk representations.

There is overwhelming evidence that risk in general is fertile ground for contested social representation and competing mental models of hazards are common. This SRT interpretation of the genesis of the representation of hazards has significant implications for explanations of how intensification or attenuation in SARF might work. It suggests that the subcultural base for any individual's representation of a hazard will influence their susceptibility to any reframing attempts, including official interventions aimed at changing risk perceptions.

Identity Processes and Risk

Although social representations have been defined at one level as being a widely shared set of beliefs, it is not the case that beliefs are accepted and used by individuals in their entirety. Among alternative competing social representations, the hazard representations of which an individual is aware, will accept, and may use will firstly depend upon the significance they have for identity. That is, the content and usage of a social representation can be predicted in relation to its likely impact upon the demands of individual identity requirements.

IPT (Breakwell, 1986), which links social representation and identity models to predict risk perceptions, proposes that the extent to which any message about a hazard will be received and incorporated into an individual's belief systems is affected by the ways in which this may threaten the individual identity. IPT argues that identity requirements are important determinants of awareness, understanding, acceptance, assimilation, and salience of social representations (and, indeed, of mental models of hazards). IPT postulates that individuals are motivated to achieve certain characteristics in their identity structure: continuity, distinctiveness, self-efficacy, and self-esteem. These four motives vary in their relative and absolute salience over time and across situations. A threat to identity occurs when these motives cannot be satisfied. If they find that esteem, continuity, distinctiveness, or efficacy are threatened, individuals will use a variety of strategies to protect or regain them.

IPT has now generated an extensive series of studies, which illustrate that individuals in the same social category will accept and use (i.e., reproduce or act in accordance with) a particular social representation to differing degrees depending upon its potential impact upon their identity esteem, continuity, distinctiveness, and efficacy. Many of these studies have concerned social representations of hazards. For example, individuals will reject social representations of their local environment as being polluted (despite objective evidence that it is) if attachment to that place features as an important aspect of their personal sense of distinctiveness (Bonaiuto et al., 1996). Such studies would suggest that use and acceptance of social representations generated in the course of social conflict or controversy are most significantly mediated by identity processes. Many environmental hazards generate such 'polemical' social representations of hazards, created by agencies or groups in conflict (Masuda and Garvin, 2006). Under these circumstances the social representations of hazards that the individual chooses will be highly influenced by identity requirements. Essentially, individuals will resist acceptance or use of social representations of hazards if they threaten identity continuity, distinctiveness, self-esteem, or self-efficacy. Of course, self-efficacy has been highlighted earlier as a predictor of risk representations.

Conclusion

Some of the social psychological approaches that describe how individuals and societies construct risk representations have been presented. The psychometric paradigm and the mental models approach have highlighted methods for exploring the structure of hazard characterization and identifying personal variation in risk representation. The work on vulnerability, particularly psychological vulnerability associated with cognitive style and affect, shows that individual differences in risk representation are in part explicable in terms of personal characteristics. Used in conjunction, SARF, SRT, and IPT offer a way of understanding the social processes involved in shaping the way certain risk representations evolve and who will accept, use, and act upon them. Together the three approaches offer a route into an integrated social psychological model of risk. The models encompass consideration of the cognitive and affective processes that shape the individual's risk representation. They also examine the social processes that create the dynamic milieu within which the individual thinks and feels about the hazard. The important thing is that taken together these models of risk construction are capable of describing individual cognitive, affective, and oretic factors as they interact with societal structures and ideologies to shape risk representations and behavior.

See also: Identity and Identification, Social Psychology of; Shared Belief: Philosophical Aspects; Social Cognition; Social Identity in Social Psychology; Social Psychology.

Bibliography

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, New York.

Bonaiuto, M., Breakwell, G.M., Cano, I., 1996. Identity processes and environmental threat: the effects of nationalism and local identity upon perception of beach pollution. Journal of Community & Applied Social Psychology 6, 157–175.

Breakwell, G.M., 1986. Coping with Threatened Identities. Methuen, London.

Breakwell, G.M., 2007. The Psychology of Risk. Cambridge University Press, Cambridge.

Breakwell, G.M., Barnett, J., 2001. The Impact of Social Amplification on Risk Communication. Research Report 322/2001. HSE Books, Sudbury.

Breakwell, G.M., Barnett, J., 2003. Social amplification of risk and the layering method. In: Pidgeon, N., Kasperson, R., Slovic, P. (Eds.), The Social Amplification of Risk. Cambridge University Press, Cambridge, pp. 80–101.

Castro, P., Gomes, I., 2005. Genetically modified organisms in the Portuguese press: thematization and anchoring. Journal for the Theory of Social Behaviour 35 (1), 1–17.

Conti, J., Terre, S., Harthorn, B., 2011. Vulnerability and social justice as factors in emergent US nanotechnology risk perceptions. Risk Analysis 31 (11), 1734–1748.

Dietz, T., Dan, A., Shwom, R., 2007. Support for climate change policy: social psychological and social structural influences. Rural Sociology 72 (2), 185–214.

Finucane, M.L., Alhakami, A.S., Slovic, P., Johnson, S.M., 2000. The affect heuristic in judgments of risk and benefits. Journal of Behavioural Decision Making 13 (1), 1–17.

Fischhoff, B., 1994. The psychology of risk characterization. In: Sahlin, N.E., Brehmer, B. (Eds.), Future Risks and Risk Management. Kluwer Academic/Plenum Publishers, New York, pp. 125–139.

Fischhoff, B., 2007. Nonpersuasive communication about matters of greatest urgency: climate change. Environmental Science & Technology 41 (21), 7204–7208.

Fischhoff, B., Gonzalez, R.M., Lerner, J.S.S.D.A., 2005. Evolving judgments of terror risks: foresight, hindsight, and emotion. Journal of Experimental Psychology: Applied 11 (2), 124–139.

Floyd, D.L., Prentice-Dunn, S., Rogers, R.W., 2000. A meta-analysis of research on protection motivation theory. Journal of Applied Social Psychology 30 (2), 407–429.

Flynn, J., Slovic, P., Mertz, C.K., 1994. Gender, race and perception of environmental health risks. Risk Analysis 14, 1101–1108.

Gasper, K., Clore, G.L., 1998. The persistent use of negative affect by anxious individuals to estimate risk. Journal of Personality and Social Psychology 74 (5), 1350–1363.

Gifford, R., Heath, Y., 2006. Free-market ideology and environmental degradation the case of belief in global climate change. Environment and Behaviour 38 (1), 48–71.

Kasperson, J., Kasperson, R.E., Pidgeon, N., Slovic, P. (Eds.), 2003. The Social Amplification of Risk: Assessing 15 Years of Research and Theory. Cambridge University Press, Cambridge.

Kasperson, R.E., Renn, O., Slovic, P., Brown, H., Emel, J., Goble, R., Kasperson, J., Ratick, S., 1988. The social amplification of risk: a conceptual framework. Risk Analysis 8, 177–187.

Kuttschreuter, M., Gutteling, J.M., 2004. Experience based processing or risk information: the case of the millennium bug. Journal of Risk Research 7 (1), 3–16.

Lowenstein, G.F., Weber, E.U., Hsee, C.K., Wells, G.L., 2001. Risk as feelings. Psychological Bulletin 127, 267–286.

Macintyre, K., Rutenberg, N., Brown, L., Karim, A., 2004. Understanding perceptions of HIV risk among adolescents in KwaZulu Natal. AIDS and Behaviour 8 (3), 237–250.

Masuda, J.R., Garvin, T., 2006. Place, culture and the social amplication of risk. Risk Analysis 26, 437.

McMath, B.F., Prentice-Dunn, S., 2005. Protection motivation theory and skin cancer risk: the role of individual differences in responses to persuasive appeals. Journal of Applied Social Psychology 35 (3), 631–643.

Miller, R.L., Mulligan, R.D., 2002. Terror Management: the effects of mortality salience and locus on control on risk-taking behaviours. Personality and Individual Differences 33, 1203–1214.

Morgan, M.G.I., Fischoff, B., Bostram, A., Atman, C.J., 2002. Risk Communication: A Mental Models Approach. CUP, New York.

Moscovici, S., Farr, R. (Eds.), 1984. Social Representations. CUP, Cambridge.

Peters, E.M., Burraston, B., Mertz, C.K., 2004. An emotion based model of risk perception and stigma susceptibility: cognitive appraisals of emotion, affective reactivity, worldviews, and risk perceptions in the generation of technological stigma. Risk Analysis 24 (5), 1349–1367.

Peters, E.M., Slovic, P., Hibbard, J.H., Tusler, M., 2006. Why worry? Worry, risk perceptions, and willingness to act to reduce medical errors. Health Psychology 25 (2), 144–152.

Rotter, J., 1966. Generalised expectancies for internal versus external control of reinforcement. Psychological Monographs 80 (1). Whole issue.

Schumacher, J., Roth, M., 2004. Sensation seeking, health-related cognitions and participation in high-risk sports/sensation seeking, gesundheitsbezogene Kognitionen und Partizipation am Risikosport. Zeitschrift fur Gesundheitspsychologie 12 (4), 148–158.

Siero, F.W., van Diem, M.T., Voorrips, R., Willemsen, M.C., 2004. Pre-conceptional smoking: an exploratory study of determinants of change in smoking behaviour among women in the fertile age range. Health Education Research 19 (4), 418–429.

Slovic, P., 1997. Trust, emotion, sex, politics, and science: surveying the risk-assessment battlefield. In: Bazerman, M.H., Messick, D.M., Tenbrunsel, A.E., et al. (Eds.), Environment, Ethics and Behaviour. New Lexington, San Francisco, CA, pp. 277–313.

Slovic, P., 2000. The Perception of Risk. Earthscan, London.

Slovic, P., 2001. Cigarette Smokers: Rational Actors or Rational Fools? Sage Publications, Inc., Thousand Oaks, CA, pp. 97–124.

Slovic, P., 2010. The Feeling of Risk: New Perspectives on Risk Perception. Earthscan, London.

Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2004. Risk as analysis and risk as feelings: some thoughts about affect, reason, risk and rationality. Risk Analysis 24 (2), 311–322.

Slovic, P., Peters, E., Finucane, M.L., MacGregor, D.G., 2005. Affect, risk and decision making. Health Psychology 24 (Suppl. 4), S35–S40.

Wagner, K., 2007. Mental models of flash floods and landslides. Risk Analysis 27 (3), 671–682.

Whitmarsh, L., 2008. Are flood victims more concerned about climate change than other people? The role of direct experience in risk perception and behavioural response. Journal of Risk Research 11 (3), 351–374.

Willis, H.H., DeKay, M.L., 2007. The roles of group membership, beliefs, and norms in ecological risk perception. Risk Analysis 27 (5), 1365–1380.

Zeelenberg, M., 1999. Anticipated regret, expected feedback and behavioural decision making. Journal of Behavioural Decision Making 12 (2), 93–106.

Ritual

David J Parkin, All Souls College, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ritual is rule-governed action, which is repeated on special occasions. It may sometimes be regarded as sacred and part of religion, but at other times may be seen as secular. It may also be accompanied by prayer or other forms of regularized verbal expression. It is commonly associated with specific places, such as temples or shrines, and with particular bodily movements or such collective events as processions. Although knowledge of ritual is handed down from one generation to the next, it may also undergo change in its form and effects. It can be studied as part of biocultural evolution.

Sacrifice could well be the historical starting point for anthropological definitions of ritual. It is the practice that is central in early anthropological studies of religion by Robertson Smith, Frazer, and Tylor. More abstract notions of ritual came later, sometimes dissociated from religion and expressing other social relationships. For Durkheim (1964), ritual is how religion makes society. It is social activity falling within the domain of the sacred. The sacred is opposed to the profane, which does not contain ritual action. Leach (1954), however, rejects Durkheim's hard and fast distinction between sacred and profane, preferring to see these as aspects of social actions lying along a continuum. At one end are those actions that are largely technical and profane, while at the other are those, which have no technical functions and are largely symbolic, aesthetic, and sacred. Most actions in between these two poles have aspects, which are both sacred and profane.

Ritual in Leach's sense is, then, by no means confined to ideas of the sacred. By removing ritual as exclusively to do with the sacred, Leach is able to broaden the definition of ritual. He thinks of it as a message-making kind of activity, and so as an aspect of any kind of behavior in which repetitive or customary acts signal messages. Handshakes or wearing a tie in the office are as much forms of ritual as a priest praying to ancestors at a shrine. So-called magical and secular acts are, therefore, both communicative and both are examples of ritual action, Ritual for Leach is, then, a way in which humans communicate with each other in different aspects of their social structure.

Thus, we may be perfectly aware of social class or status differences and even be able to talk about them discursively. But we still signal or symbolize these differences in a number of short-cut ways, through distinctive accents, differences of dress and lifestyle, and even bodily movement and gesture.

Sometimes we are not so conscious of the differences our ritual behavior communicates, and sometimes the differences exist simply through lack of knowledge of how to behave on, say, a formal occasion. In other words, the sociostructural approach argues that ritual action does not occur in a political and economic vacuum: politicoeconomic interests are always involved to some degree, even if sometimes minimally. Leach also claimed that separating ritual from myth was nonsensical for myth 'says' in words what ritual 'says' in action, a view which relates to but is somewhat different from Malinowski's view of the two as separable but sometimes linked: with rituals being accounted for or explained in myths. In accepting the idea that ritual is communication, Bloch (1974) proposes that its repetitive, formalized nature precludes argument and so reinforces the authority of those who control and advocate the ritual. Other scholars counterargue that onlookers and participants in a ritual may have different ideas as to how it should be conducted, and so change the interpretation, meaning, and social effect of the ritual.

A second approach to the study of ritual emphasizes its logical qualities. Here, ritual communicates fundamental ideas that may be specific to a particular culture, or that may even be universal. It is an approach that owes much to structuralism in anthropology, which is predominantly associated with the work of Lévi-Strauss. It turns on the idea of culture as organized through binary oppositions.

Thus, Needham (1961) has suggested that, when a male diviner or shaman uses his left hand to throw the cowrie shells, which he or his possessory spirit will interpret in order to understand the cause and course of an illness or misfortune, he is reversing a kind of natural order. For men are normally associated with the right hand and women with the left. But since diviners fall outside the realm of normal human capacities, being able to see what other humans cannot, it makes logical sense to reverse the equation of maleness with the right hand. There is no functional explanation for these kinds of ritual reversal for they are governed by proclivities of the human mind.

That said, local political and economic circumstances can sometimes take up these proclivities and harness them to a cause. Thus, prophets leading religious movements may communicate their power and distinctiveness through rituals, which reverse normal expectations. Many other kinds of ritual behavior, like some beliefs and myths, achieve their communicative effectiveness by being considered natural.

Now, while it is true that ritual may be thought of as communicating messages about status or about the way the human mind works or how the world is, it varies enormously in the degree to which it is elaborated. After all, a handshake can hardly compare with a full-scale ritual investiture of a monarch spread over days and involving thousands of people and lavish expenditure. To call both these kinds of activity 'ritual' clearly leaves out much of what is special to the latter. Similarly, some hunting and gathering peoples such as the African Mbuti forest dwellers do not perform elaborate rituals, while other African groups noted for their masquerades do.

Peoples do, then, seem to differ in the relative amounts of ritual, which they perform; although, in making this claim, there is a dangerous Eurocentric qualitative assumption that the significance of ritual is measured by the amount of time, materials, and people involved. Who is to say that an apparently slight activity, taking only a moment and few resources, may not in fact 'say' much more than a drawn-out ceremony of great proportions whose influence on people is in fact limited? Nevertheless, it is a useful if provisional starting point of analysis to try and understand how variations in the complexity and lavishness of ritual reflect those of social organization and cultural logic.

Arising from this approach is another, which recognizes that as well as communicating messages about the current social structure or about fixed properties of the human mind, rituals also change things or enable changes to take place. We may call this third, the transformational approach to ritual. For Van Gennep (1965), the rites of passage as marking changes of status in the human life cycle – for example, birth, puberty or adolescence, marriage, first child, first grandchild, and death – could be broken down into three phases: separation, liminality, and incorporation. In other words, the individual undergoing the ritual must first be separated from the community of which they are a member, perhaps by removal of the head hair or some other bodily alteration. They then enter the liminal phase in which they are regarded by members of the community as outside and beyond normal human society. This is the phase in which there is transition from one status to another, for example, from child to adult, bachelor to husband, etc. Van Gennep saw this as a period of death, i.e., of the child, and of rebirth, i.e., of the adult. It is a theme elaborated in many later studies that showed the ritualization of death as the condition of new life or regeneration. The third and final phase is when the individual is reincorporated in the community, now no longer a child, but an adult with the rights and duties expected of this new status.

Van Gennep saw these features of rites of passage as also typifying other kinds of transition in addition to those of the life cycle, such as the seasonal or other movements of groups, fertility rites, and entering and leaving dwellings, special days or months, and cult societies. Victor Turner (1967) made the notion of liminality central to his work on ritual, extending to the idea of antistructure, by emphasizing the rejection of society on the part of people undergoing transition or in a marginal state, and also communitas, by which he meant the way in which people on ritual occasions temporarily transcended their normal, everyday social conflicts and cleavages and emphasized their common humanity. In addition to the life crisis rites of Van Gennep, Turner further distinguishes rites of affliction, which are held to reverse the effects of, for example, women's infertility, crop failure, sickness, and other misfortunes adversely affecting the lives of people living together in a single village or other dwelling. They also enable people to move from one state to another, although not as part of the human life cycle, but more as part of the unpredictability of the human and social condition in which conflicts and contradictions are not so much represented in ritual as presented, addressed, and then redressed, an approach which thus builds on and advances that of Durkheim.

A subset of this approach might conveniently be called Marxist in the sense that social inequality and exploitation of working by a ruling class is presupposed.

A Marxist approach would, for instance, claim that the official, media, and public ritual attention devoted to the British royal family evokes ordinary people's unquestioning compliance and loyalty, to the extent that they cannot see that the institution of royalty in fact underpins the system of landed privilege that defines an aristocracy and provides an imitative model for other ruling and bourgeois classes. The 1990s in the UK saw an extraordinary rise in the degree to which the media and many of these same ordinary citizens began to question the worthwhileness of royalty. Questions were raised as to whether this was an example of royal rituals losing their capacity to retain subjects' loyalty, a kind of demystification of ritual. Or perhaps it was a case of the different interpretations of ritual enabling social change to occur. In any event, by the 2010s, public attitudes to royalty warmed again, as instanced especially at royal weddings such as that of the young heir to the throne, the Duke of Cambridge and his popular 'commoner' bride, the former Kate Middleton.

Royal rituals have been a common focus of anthropological analysis, in which their transformative power has been highlighted. Gluckman (1963) was dissatisfied with studies of royal and other rituals, which saw them as simply affirming the sovereign's authority and the solidarity of the society under his rule. Using the example of the Swazi *Incwala* royal ritual in southern Africa, Gluckman claimed instead that its socially cohesive function was only possible after it had revealed the society's internal conflicts: such conflicts had to be addressed before they could be accepted as part of the normal social process.

Beidelman (1966) criticizes Gluckman's overreliance on social conflict as needing to be expunged through royal rituals, and argues for a view of them as marking out the divine or mystical properties of kingship, so providing the king with legitimate recognition to reign over his subjects.

This example moves us on to a fourth approach to ritual, which is to see them not necessarily as achieving a particular social or logical function, but as expressive in and of themselves. Beattie (1966) attacked attempts to explain rituals as practical and intellectual attempts to solve problems. Rituals draw instead on people's human predilection to engage in creative thinking, sometimes through culturally enjoined associations of ideas, and were thus closer to art and drama than to technology or science. Rituals in this sense do not 'do' things so much as express human dilemmas, symbolically and often under some kind of duress. They cannot be reduced to questions of whether they are rational or irrational, or true or false propositions, any more than aesthetics can be explained in these terms. In many respects this explanation of ritual as expressive rather than functional reproduces Leach's distinction between technical and aesthetic acts as occupying two extreme poles of a continuum between which they are both aspects of any activity. Nevertheless, expressed in the way that Beattie did, it freed anthropologists from having to seek practical explanations for a particular form of ritual behavior. If a ritual appeared to perform a particular function in society, this could better be regarded as an unintended consequence of an activity whose participants did not have such consequences

in mind. Thus, it may well be that the British royal rituals for a long time had the effect of persuading the public to accept the stratified status quo as rightfully headed and legitimized by the monarch and the monarch's family. It may well be, moreover, that politicians and members of the ruling elite recognized, seized on, and developed this effect, so turning consequence into purpose. The expressive here becomes instrumentally effective.

However, not all that which we understand as ritual behavior can be interpreted in this way. People throughout time do not normally calculate in advance what might pass as effective ritual, nor indeed separate ritual as a social category from nonritual; nor may we assume that what we define as rituals always have significant social and intellectual consequences.

This leads on to a fifth approach that derives from the widely held idea that ritual is principally performative and not necessarily aimed at any other than satisfying its own rules of procedure. Geertz's analysis of the Balinese state as theater is an example (1980). Another example is a study of the Indian Jain rite of worship by Humphrey and Laidlaw (1994). Their intriguing starting point is the fact that in some societies, people may distrust ritual practices as being empty or false, and yet continue with them. Similarly, noting that for many people, religion and ritual are seen as going together, they also refer to the way in which practitioners of some religions turn on their rituals as unworthy and even frightening. The European Reformation attempted to strip off from Christian religious belief what were regarded as some unacceptable ritual practices, as Wahabi reformers in Islam are doing today.

By contrast, some argue that ritual is good for religion and humanity. Mary Douglas (1973) drew some of her own ideas from trying to understand the rejection of ritual in the Catholic Church by many of its senior adherents who also, for good measure, dispensed with the Latin language as if transparent clarity was all that mattered, over and above the expressive power of faith.

The disjunction between religion and ritual led Humphrey and Laidlaw (1994) to suggest another disjunction, which is that between the intention of actors in ritual and the ritual actions themselves. Sometimes intention and act fit, and sometimes they diverge irretrievably, but sometimes they are aligned.

Such lack of fit is not to be understood as the result of mistakes in intention and performance, but as evidence of the way in which ordinary actions may become endowed with the quality of ritual. For these authors, ritual is a quality that actions take on: they become ritualized. While everyday actions are intrinsically intentional, ritual acts enter a kind of phenomenological space freed of such everyday intentions, and, paradoxically, are 'intended' to do so. There is here a 'ritual commitment,' a notion close to that expressed in a classical study of ritual by Gilbert Lewis (1980). Humphrey and Laidlaw (1994), moreover, oppose the idea that ritual is necessarily communicative and expressive of messages and beliefs.

One is here reminded of Sperber's (1975) view that symbols do not intrinsically mean anything. Rather it is people who interpret a symbol as such, and who thereby give it meaning. The concept of communication here has to be established. It may mean one person making an assertion and expecting a response from the hearer. Or it may simply mean making a gesture or uttering a conventional phrase ('It is hot today'), for which no response is expected and which is therefore really a form of what has been called phatic communion, a device to secure solidarity with another person. Rituals may sometimes communicate but, as Rappaport (1979) points out, so do many other kinds of phenomena, and by privileging its expressive functions we miss what is otherwise distinctive of ritual.

The view that ritualization is a process by which actions become endowed with a ritual quality, shares something with Gerholm's (1988) notion that ritual is, so to speak, partly made up as it proceeds. That is to say, there are at least some kinds of ritual for which precise textual rules do not exist or which have been forgotten, with the consequence that participants have little more than a sense of ritual occasion or commitment and endeavor to convey that sense in actions and words, inventing and borrowing as much as they can. This ushers in a query as to how much ritual depends on precise wording as well as actions. A Christian prayer depends heavily on personally composed as well as prescribed verbal forms, and is firmly regarded as ritual. But rituals are also organized through spatial orientation and bodily positioning. The many different kinds of Islamic prayer are also personally or formulaically verbal, but crucially require orientation to Mecca, while a particular class of prayer or devotion called *salat* is made up of a complex sequence of prostrations as well as recitations (Bowen, 1989). In the event of the mosque imam taking on the recitations on behalf of the congregation, the prescribed bodily actions become the individual worshipper's ritual responsibility and must be repeated if they are performed incorrectly. Parkin (1992) has argued that it is not always the case that the verbal is important in ritual, while set spatial directions and orientations, and bodily gestures, movement, and positions are what mark off and so fundamentally characterize ritual. Bell (1997) also points to the way in which the disciplined body in rituals has to obey formal rules regarding its posture and positioning.

Finally, running throughout the various perspectives on ritual from Tylor to present day is its association with sociocultural change and evolution. Most often this is implicit, as in ethnographic demonstrations of how ritual can reflect, bring about, or deny social change while itself undergoing transformation. Sometimes ritual is presented explicitly as the template and instrument of biocultural and cognitive development as in the work of Barrett (2004), Boyer (2002), Cohen (2007), and Whitehouse (2004). Here religious beliefs and their ritualization are based on innate or intuitive cognitive dispositions, which are developed along with increasing sociocultural complexity.

A challenge is to avoid essentialist definitions of ritual, which seek to set up criteria that must be met before this label can be applied to an activity.

Phenomenologically, ritual is to do with the paradox of being both the agent of an action, but also being taken along by this same action. It presupposes being able to move within an action: from citing its purpose to having no need, wish, or clear idea of what it is about, and yet valuing it. Nevertheless, ritual must have some kind of structure, the most elemental of which is that individuals alone or with others take up positions with regard to each other or with regard to objects, experiencing such positioning as unlike any other.

See also: Control, Social; Durkheim, Emile (1858–1917); Magic, Anthropology of; Myth in Religion; Religion, Sociology of; Ritual and Symbolism, Archaeology of; Taboo; Totemism.

Bibliography

Barrett, J.L., 2004. Why Would Anyone Believe in God? AltaMira Press, Walnut Creek, CA.

Beattie, J., 1966. Ritual and social change. Man: The Journal of the Royal Anthropological Institute 1, 60–74.

Beidelman, T.O., 1966. Swazi royal ritual. Africa 36, 373–405.

Bell, Catherine, 1997. Ritual: Perspectives and Dimensions. Oxford University Press, New York.

Bloch, M., 1974. Symbols, song, dance and features of articulation: is religion an extreme form of traditional authority? Archives European de Sociologie 15 (1), 55–84.

Bowen, J.R., 1989. Salat in Indonesia: the social meanings on an Islamic ritual. Man: The Journal of the Royal Anthropological Institute 24, 600–619.

Boyer, P., 2002. Religion Explained: The Evolutionary Origins of Religious Thought. Basic Books, New York.

Cohen, E., 2007. The Mind Possessed. The Cognition of Spirit Possession in the Afro-Brazilian Religious Tradition. Oxford University Press, Oxford.

Douglas, M., 1973. Natural Symbols: Explorations in Cosmology. Barrie and Jenkins, London.

Durkheim, E., 1964. The Elementary Forms of the Religious Life. Allen and Unwin, London.

Geertz, C., 2006(1980). Negara: The Theatre State in Nineteenth Century Bali. Princeton University Press, Princeton.

Gerholm, T., 1988. On ritual: a postmodernist view. Ethnos 53, 190–203.

Gluckman, M., 1963. Rituals of rebellion in south east Africa. In: Gluckman, M. (Ed.), Order and Rebellion in Tribal Africa. Cohen and West, London.

Humphrey, C., Laidlaw, J., 1994. The Archetypal Actions of Ritual. Oxford University Press, Oxford, UK.

Leach, E., 1954. Political Systems of Highland Burma: A Study of Kachin Social Structure. G. Bell, London.

Lewis, G., 1980. Day on Shining Red: An Essay in Understanding Ritual. Cambridge University Press, Cambridge, UK.

Needham, R. (Ed.), 1961. The left hand of the Mugwe: an analytic note on the structure of Meru symbolism. Africa 31, 28–33.

Parkin, D., 1992. Ritual as spatial direction and bodily division. In: De Coppet, D. (Ed.), Understanding Rituals. Routledge, London, pp. 11–25.

Rappaport, R.A., 1979. Ecology, Meaning, and Religion. North Atlantic Books, Berkeley, CA.

Sperber, D., 1975. Rethinking Symbolism. Cambridge University Press, Cambridge, UK.

Turner, V.W., 1967. The Forest of Symbols: Aspects of Ndembu Ritual. Cornell University Press, Ithaca, NY.

Van Gennep, A., 1965. The Rites of Passage. Routledge and Kegan Paul, London.

Whitehouse, H., 2004. Modes of Religiosity: A Cognitive Theory of Religious Transmission. AltaMira Press, Walnut Creek, CA.

Ritual and Symbolism, Archaeology of

Rosemary A Joyce, University of California, Berkeley, CA, USA

Published by Elsevier Ltd.

Abstract

Archaeologists approach ritual and symbolism from a material basis: through the places, things, and trace residues left behind as a result of human action in the past. Where historically, archaeologists adopted definitions of ritual, religion, and symbolism that were rooted in structuralism, today they increasingly draw on the semiotics of Charles Sanders Peirce and practice theory, broadly rooted in the work of Michel de Certeau and Pierre Bourdieu, to interpret materials in terms of action and meaning-making in the past.

Archaeology of Symbolism and Ritual

Archaeologists derive their understandings of the aspects of human social life through material traces: the spaces, objects, and residues created when people carried out activities at some point in the past. As a result of this material mediation of knowledge production, archaeologists approach topics involving ideas or cognition in distinctive ways. In recent decades, this has involved a move away from the identification of structures or categories reflected in things or features of things, to an examination of how material media shaped and were shaped in actions that we can understand as meaning-making. Archaeologists interested in meaning critique reliance on structural linguistic models rooted in the work of Ferdinand de Saussure, and advocate broader semiotic models derived from the work of Charles Sanders Peirce. A major focus of archaeological research on symbolism has been the linked domains of ritual and religion. Religion, understood as belief, was initially thought to be difficult if not impossible for archaeologists to investigate. Ritual, understood as a behavioral aspect of religion, in contrast was defined in ways that potentially allowed identification of material evidence, such as distinctive spaces where rituals took place, and objects used in ritual practices. Increasingly, archaeologists debate the dichotomy between religion and ritual, the nature of any relationship between them, and the ease of distinguishing ritual action from everyday action. While many archaeologists continue to explore ritualized practice, others have turned to investigation of the experiential dimensions of religion, or examine the material settings associated with known world religions, where it is imperative to relate material spaces and objects to concepts recorded in texts.

Archaeology of Symbolism

Symbolism requires representation of one thing, often more abstract, by reference to another, often more concrete. Relations of substitution are thus fundamental to the definition of symbols. This implies that symbolizing requires a conscious act of reference to associate meaning and the thing taken to carry meaning. Most archaeologists have, until relatively recently, based their understanding of the relations of reference involved on a structuralist model of symbolism rooted in the work of Ferdinand de Saussure (Preucel, 2010). This approach saw symbolizing as the making of an arbitrary relation between a concrete form (a vehicle) and a usually more abstract meaning intended by the producer of the arbitrary symbolic relation. The task of analysis thus became to decode the intended meaning, making archaeology of symbolism dependent on a claim to reproduce ancient thought.

Drawing on information theory, archaeologists treated symbols conceived in this way as messages intended to communicate clear meanings (Wobst, 1977). If symbols were media to convey messages, not just arbitrary meanings, then there would have been strong social pressures, and even adaptive advantages, to having shared interpretations of their meanings. The application of communication models to the archaeological study of symbols facilitated the development of an archaeological theory of style, the definition and interpretation of sociopolitical status badges (items that convey rank or role, such as a crown), and the interpretation of differences in material culture as markers of ethnic boundaries and cultural identities. However, ethnoarchaeological research demonstrated that the degree to which the use of common material forms conveys either a conscious or unconscious intent to associate a form and meaning is highly variable (Hodder, 1982a). Arguably, only those instances where a conscious connection was forged between vehicle and meaning could properly be considered symbolic. An example would be artifacts interpreted as status badges created as deliberate signs of rank, such as a scepter or crown. Where artifact style (the manner of making something that produces a distinctive appearance of similarity in a series of things) had been viewed, using communication theory, as deliberately intended to convey a sense of identity, ethnoarchaeological critics found little evidence of the necessary intent to signify belonging that would have justified this under the model of the arbitrary symbol.

Critics of the model of archaeologically recovered objects as vehicles carrying preexisting meanings developed starting at least as early as the 1980s (Robb, 1998). Critics argued that symbolic meaning, rather than being installed in a symbol by the intent of the original maker or user of a symbolic object, was negotiated, shaped, created, and recreated through the use of cultural forms, including the durable ones recovered by archaeologists. This view of symbolism requires attention to contextual relations in order to understand how symbolic meaning was produced. Archaeologists developing these

critiques expressed uncertainty about their ability to understand the specific meanings symbols had for past actors, yet they routinely made arguments about the deployment of symbolism as part of strategies of actors and factions within past societies.

Growing from this understanding of meaning as emergent from the act of interpretation, contemporary archaeologists have explored the three-part model of the sign offered by Charles Sanders Peirce (Bauer, 2002; Joyce, 2007, 2008; Lele, 2006; Preucel, 2010; Preucel and Bauer, 2001). For Preucel and Bauer (2001: pp. 88–89) the critical shift is from the two-part relation of signifier and signified in the Saussurean sign, to a three-part relation of Sign, Object, and Interpretant in the Peircean sign. The third part of the Peircean sign, the Interpretant, places meaning-making in historical and contextual circumstances, as an active process of creating relations among people and things. An interpretant actively links the representamen (sign, concrete object that represents) to an object. Consequently, "sign relations are neither linguistic nor completely arbitrary … the representamen's relation with an object represented for and through an interpretant is not necessarily arbitrary" (Lele, 2006: p. 50).

This latter point has become especially important in the contemporary exploration of symbolism in archaeology. While an arbitrary relation of representamen and object in a symbol is one mode of meaning-making, two other modes can be defined where the relationship is not arbitrary, and thus potentially open to analysis without access to the intentions of the original maker of a symbol. These two additional forms of semiosis (meaning-making) explored by Peirce are iconicity and indexicality. Indexicality has been particularly productive in archaeological analyses (Joyce, 2008; Lele, 2006; Preucel and Bauer, 2001). The representamen of an indexical sign is existentially connected to the object of the sign, rather than being simply conventional: the classic example is a footprint interpreted as a sign of the person or animal that produced it. A rock painting in the form of a handprint, even if painted as a design rather than directly printed by dipping a hand in paint, can be understood as indexing, or pointing toward, the presence of a human being, as a trace of a person who had been present at this place. From this indexical relationship, archaeologists can propose that the hand print might have been interpreted as a sign of a human presence, rather than as an entirely arbitrary symbol.

Using the concept of indexicality, archaeologists always can work from some form of connection between a material trace and a cause to begin to suggest that meaning-making took place, and even to begin to suggest a range of meaning that might have been in play in the past. Preucel and Bauer (2001: p. 90) cite Richard Parmentier's discussion of archaeological analysis of pottery style as an example. Every pot recognizable as in a specific style indexes a set of repeated traits that make up the style, which can through interpretation point to an existential identification among the makers of the different pots, developing their skills and manner of making pots together. This use of indexicality avoids the problematic assumption of self-conscious, intentional identification through style critiqued by earlier analysts (Hodder, 1982b). It also makes clear that archaeologists identifying a shared style as evidence of a shared identity are creating a sign in the present using things made in the past. Potentially, the same kind of interpretation could have occurred among the makers and users of these pots – although this cannot be certain. Parmentier's model of potential but not automatic identification through shared style better matches the evidence gathered through ethnoarchaeological studies.

Iconicity is also important in archaeological analyses of symbolism and meaning. Iconic signs create meaning by the recognition of resemblance between the representamen and object. What constitutes resemblance is culturally specific. Preucel and Bauer (2001: p. 90) note that for Parmentier, legibility of different pots as being in the same style is a product of their mutual resemblance, an iconic sign. Archaeological analysis of graphic images routinely relies on iconicity, usually without explicit consideration of how meaning is being inferred. An example would be the analysis of small-scale objects made of clay, figurines, routinely identified as representing human or animal bodies even when only very general symmetry provides the ability to define a head and body, and when most bodily features are not represented (Joyce, 2007). This example shows that the identification of resemblance that underpins iconicity also requires an act of interpretation. Again, the use of a Peircean model clarifies that both the modern analyst and the original makers and users of any object carry out acts of interpretation, which may or may not result in congruent meanings.

Archaeology of Ritual

Archaeologists have always identified some things as evidence of ritual. Beginning in the 1950s explicit concern was addressed to the methodological status of archaeological inferences concerning belief and meaning. Interpretation of ideational systems, including religion, was viewed as more problematic than other topics, such as economics and subsistence. As early as the 1960s, some archaeologists argued that ritual and symbolism were no more difficult to attempt to analyze than other topics, questioning the analytical separation of material and ideational aspects of society. Since the 1980s, archaeologists influenced by practice theory have attempted to demonstrate that the archaeological analysis of ritual benefits from explicit attention to human agency. This includes discussions of the experiential dimension of religion, whether conceived of structurally, cognitively, or in terms of biocultural effects. Most recently, archaeologists studying world religions have emphasized the need to create an archaeology of religion, independent of an archaeology of ritual.

The 'Ladder of Inference' and Processual Approaches

In 1954, Christopher Hawkes (1954) proposed that in the absence of texts, archaeologists were constrained by a ladder of inference which governed the probable reliability of their explanations. According to the ladder of inference, some topics were more tightly circumscribed by pragmatic considerations. As a result, archaeologists were likely to recover a greater proportion of the evidence necessary for understanding these topics, and their explanations were likely to suffer less from crucial missing data or from variability in the nature of the evidence. Areas such as religion and belief were understood to

be less tightly tied to materiality, and thus archaeologists were likely to lack crucial pieces of evidence to support explanation, or they might recover material that could be explained in multiple ways and have no basis to choose between explanations. Hawkes explicitly argued that understanding areas like religion and symbolism would be most feasible where written texts were available.

In contrast, in initial programmatic statements for the processual archaeology of the early 1960s, Lewis Binford (1962) argued that all aspects of culture, including ritual, were open to scientific archaeological analysis if questions were properly framed. Binford argued that objects could be identified whose primary use was to convey symbolic meanings as part of the reproduction of society. These Binford labeled sociotechnic and ideotechnic. An example would be an axe or knife made of soft metal like copper. Not useful as an actual weapon, this object would not qualify to be identified as technomic (used for pragmatic purposes, as a knife made out of hard material capable of being used to cut something). As a sociotechnic object, it might have served social purposes, such as marking rank or status. As an ideotechnic object, it might have been a pure symbol of mental concepts, such as those that formed part of religions.

Employing a processual framework, Robert Drennan (1983) developed an explicit set of expectations for how archaeologists could identify the material evidence of ritual. Drennan based his understanding of religion on the work of social anthropologist Roy Rappaport (1979). For Rappaport (1979: pp. 117–121, 207–211, 231–234), religion was a set of widely shared ideas about the nature of the world and the place of people within it. Defined in this way, religion would, as Hawkes (1954) argued, be beyond archaeological investigation, unless the archaeologist had access to written texts. What made Rappaport's approach useful for archaeologists was his argument that religion was enacted in ritual. Ritual contrasted with belief, which was interior, private, and inaccessible (Rappaport, 1979: pp. 194–195). The effect of ritual was to combine "the most abstract and distant of conceptions" with "the most immediate and substantial of experiences," transforming "the dubious, the arbitrary, and the conventional into the correct, the necessary, and the natural" (Rappaport, 1979: p. 217). Rituals were "stylized, stereotyped and repetitive"; they required "the proper persons, occasion, place, and time" (Rappaport, 1979: pp. 175–176). As a consequence, rituals left material traces that archaeologists might recover.

Following Rappaport, Drennan defined rituals as sequences of repeated action directed at intervening in relations between humans and supernatural forces or beings. Among the key features he and other processual archaeologists used to attempt to objectively identify remains stemming from ritual were associations between otherwise rare items, and the recovery of such rare items from public spaces. For example, unusual elements of animal skeletons found near open public spaces could be interpreted as part of costumes worn during ritual performances. The approach Drennan employed assumed a separation between an everyday sphere of activities and a separate sphere, the domain of ritual, detectable by distinctive location and unusual artifacts.

Processual approaches required external bodies of information for the formulation of models of ritual and symbolic systems, but unlike the limitation to historical situations where beliefs were recorded in texts, proposed by Hawkes (1954), processual archaeologists drew on analogies among different societies to construct models for less well-documented societies. Often this involved using societies observed by ethnographers to propose models for archaeological societies. Specific ethnographic analogy, commonly known as the direct historic approach, was based on the study of living societies believed to be descended from the archaeological population that created a specific site. For specific analogies to work, there had to be an assumption that little or no change had occurred over time. General analogies were based on comparative ethnographic samples such as those contained in the Human Relations Area Files (http://www.yale.edu/hraf/). The power of explanation rested on the strength of arguments linking the analogies to the test case, requiring extremely broad uniformitarian generalization. The level of generalization required made explanations of ritual and symbolism extremely vague, and tended to substitute the operations of superorganic entities for individual actions (Wylie, 1985).

The most systematic processual definitions of the conditions needed to identify spaces and objects as evidence for ritual have been provided by British archaeologist Colin Renfrew (1985, 1994). In his view, religion is unobservable, because it concerns belief in metaphysical beings. Religion could in theory affect any part of human activity, and thus could structure any part of the material world. Ritual, in contrast, was the practice of specific actions motivated by religious belief. Renfrew (1985) provided what he characterized as a framework to infer religion, which he later modified (Renfrew, 1994), that has remained influential. In his original formulation, intended to help him argue that certain sites were cult centers, he called for attention to those features of the archaeological record that allowed a 'focusing of attention' by ritual participants. Understanding ritual as situated at a boundary zone between the everyday world and a supernatural world, Renfrew's model directed archaeological attention to certain kinds of spaces, of which hill-top shrines were an example. Religious rituals involved making deities present, through actions such as offerings, which would produce identifiable material residues. These linking arguments allowed Renfrew to propose that certain locations, features, and objects might be expected in places where ritual was carried out. These included the practice of ritual in special geographic settings or special built environments; the use of objects to assist in focusing, including altars, ritual vessels, lamps, or the like; the production of images of deities or of animals symbolically related to supernatural beings; and ritual actions such as music, dance, and other bodily practices that might leave material traces.

Renfrew (1994) revisited his original formulation in a more general discussion of the possibility of a cognitive processual approach to religion as part of an archaeology of mind. This involved defining religion as "the role of symbols in coping with the unknown and the supernatural" (Renfrew, 1994: 47). He argued that archaeologists were most likely to identify religious activities when they took place in specialized locations, even though there might be religious practices

'embedded' in everyday life. Renfrew suggested that iconographic representation would likely be a useful route to pursue religious ideas, such as the separation of a soul from a body after death. Some of these iconographically represented ideas, he suggested, might be common if not universal aspects of human religions.

While Renfrew's framework has been influential, it is not without critics. Timothy Insoll (2004: pp. 92–96) criticizes Renfrew's approach for leaving out the numinous, an aspect of individual experience, and for projecting a universal view of religion that implicitly is based on Christian practices that assume separation of the sacred and religious from the everyday and secular. These concerns have become more important in contemporary archaeology of religion, including the study of world religions and the neuropsychology of individual religious experience.

Ritual as Practice and Structured Deposition

While the processual approaches reviewed above are influential, they were ultimately critiqued for failure to provide satisfactory accounts of human agency. Following such postprocessual critiques (Hodder, 1982b), archaeologists working on questions of ritual rapidly adopted new models of religion. Especially influential was the perspective developed by Catherine Bell (1992) who proposed that in place of ritual, analysts should examine ritualization: "a way of acting that is designed and orchestrated to distinguish and privilege that which is being done in comparison to other, usually more quotidian, activities" (Bell, 1992: p. 74). Rather than being a question of identifying sites used for ritual, or objects used in ritual, this definition offered the potential to examine what people were doing that marked some actions as different from others.

In this vein, archaeologist Richard Bradley (2003) argued that treating ritual as action rather than as a series of symbolic statements "makes it possible to consider the contexts in which particular rituals are created and performed, and the consequences of such actions, whether they have been intended or not" (Bradley, 2003: p. 12). Bradley based his analyses on the archaeological concept of 'structured deposition' (Richards and Thomas, 1984): a means to understand archaeological deposits as products of more or less intentional activity, including the kind of stylized practices Bell would call 'ritualized.' Bradley (2003: p. 8) argued that the association of otherwise unremarkable objects in structured deposits in some British Neolithic sites was "governed by a number of conventions" suggesting "a degree of formality" that is "a defining feature of ritual."

Independently, archaeologist William Walker (1998) developed a concept parallel to that of structured deposition from a behavioral approach, rather than one grounded in practice theory. Walker rejected the idea that archaeologists could identify specific objects as evidence of ritual. Instead, he emphasized contextual clues indicating that objects were used in ritualized ways. Walker emphasized extreme destruction or fragmentation as a key sign of the ritualized disposal of things used in ritualized practices.

The convergence of these two analyses, one rooted in consideration of intentional human agency, the other explicitly behavioral and thus avoiding any presumptions about intentions, has added weight to the use of contextual patterns in deposition as evidence of ritualization by archaeologists from a wide spectrum of approaches.

The Material Basis for an Archaeology of Ritual and Symbolism

Despite the shifts in emphasis over time, the specific material bases employed in archaeology of ritual have remained fairly stable. Offerings or caches, burials, particular spatial locales, and visual images have consistently provided the subject matter for analyses of ritual.

Caches and Offerings

Caches, hoards, or offering deposits figure as primary evidence of ritual in several regional traditions in archaeology, particularly North America, Central America, the eastern Mediterranean, and Neolithic Europe (Bradley, 1991). Interpretations of caches vary from purely materialistic to ideational. Some analysts stress the economic effect of the withdrawal from circulation of a large quantity of specialized goods at one time. Others propose structural-symbolic interpretations in which materials employed worked to create specific associations between ideas about the supernatural world and the locale of the offering, consecrated by the deposit. Explanations of both kinds have been offered for the same deposits, and indeed, they are not mutually exclusive, since a ritual action can have both pragmatic and symbolic effects.

Burials

Mortuary analysis, the study of sites and artifacts related to human burial, has occupied a central place in the archaeology of ritual. Studies conducted under normative assumptions described burial customs, including the laying out of the body according to symbolic beliefs (such as the eastward orientation of the head, toward the rising sun). The removal of skeletal elements from burials and the practice of secondary burial have been seen as unambiguous evidence of ritual practice. The deliberate burial of animals has generally also been regarded as evidence of ritual activity.

Processual archaeology directed attention to the generation of covering laws or assumptions that could link the forms of burials, and particularly, diversity among them, to social differences. Burials were seen as reflecting the social identity of the deceased. They operated as mechanisms to assert claims for territory. Postprocessual archaeologists viewed burial assemblages as the material media of ancient rituals, and demonstrated that mortuary remains can mask social relations as easily as they can reflect them (Hodder, 1982b).

Temples and 'Sacred Landscapes'

The identification of special sites for ritual, through spatial segregation, specific construction, or a combination thereof, is also a consistent aspect of archaeology of ritual. While buildings called temples today may sometimes be unambiguously identified through written documents, such as ancient Egyptian and Maya inscriptions, in other cases the identification of temples is solely made based on inferences of the analyst. In keeping with other approaches that examine the active role of

humans in ritual, some archaeologists have explored how different spatial settings structure ritual action (Moore, 1996).

Spaces that can be viewed as nonordinary by nature are sometimes treated as effectively unproblematic locales for an archaeology of ritual. Caves and natural springs are prominent foci for some definitions of such sacred geographies. In contrast, the investigation of entire landscapes, such as those of megalithic Europe, emphasizes the social marking of space through ritual action, denying an attempt to separate categorically sacred landscape features from the pragmatic experiences of human agents (Richards, 1996).

Visual Images

Visual images recovered archaeologically have been a staple in reconstructing ritual. They have been approached in two complementary ways. On one hand, researchers have argued that images or scenes represented mythologies or beliefs about the supernatural. Some have been identified as models for ritual enactment. For example, elaborately dressed individuals buried at the coastal Peruvian Sipan site have been identified with the clearly nonhuman figures depicted in presumably mythological scenes on pots made and used within the Moche kingdom (Chapdelaine, 2011). The implication that such acts as ritual decapitation and mountaintop sacrifice actually took place within ancient Peruvian society has been supported by the recovery of caches of skulls and of frozen bodies on the high Andes. Here, the pottery scenes are viewed as imagery of mythology, and the burials and sacrificial deposits as re-enactments of these mythological scenes.

In other cases, representations have been interpreted as recording ritual practices, but not their mythological charters, usually because of the embedding of ritual in other aspects of social life. For example, Classic Maya monuments have been identified as showing members of the ruling class engaged in ritual bloodletting and the sacrificial burning of paper and resin, ritual actions that simultaneously served to validate sociopolitical hierarchies (Schele and Miller, 1986).

Current Issues

The archaeology of ritual began with an emphasis on the study of societies where texts governing practice were either not in widespread use, or were unavailable for analysis. This led effectively to an archaeology of ritual unrelated to questions about religious belief. Current debate in the field, based especially on new attention to religious practice in world religions, traditions with broadly shared bodies of theological texts, has criticized the primacy of studies of ritual, and the ways that ritual and religion have been linked by archaeologists.

Thus Fogelin (2007: p. 66) suggests that "archaeologists often assume that ritual is a form of human action that leaves material traces, whereas religion is a more abstract symbolic system consisting of beliefs, myths, and doctrines." He suggests that archaeologists differ on whether they see religion or ritual as primary. Taking religion (belief) as primary leads to viewing the field of actions through which people engage with the world as a play of symbols. Approaching the same body of activities as pragmatic action leads to an emphasis on tracing the sequence of actions and their repetition as the historical reproduction of practices, and the sequential process of semiosis.

Many contemporary archaeologists continue to pursue the study of ritual, understood as the more materially evident aspect of belief-inspired action (e.g., Kyriakidis, 2007). But at the same time, a large body of archaeological work on world religions has developed, for the most part avoiding definitional questions, and stressing particularities over generalization (Insoll, 2001). Examples include studies of the religiosity of Christian nuns (Gilchrist, 1994); the effects of Quaker belief on the material world in which Quakers lived (Chenoweth, 2009); and investigations of religions like Buddhism and Islam where, unlike in Christian practice, a separation of everyday and spiritual realms should not be expected and cannot be used as a methodological approach (Insoll, 2001).

Insoll (2004: p. 6) argues that archaeologists using ritual as a methodological focus are trapped by structures specific to one religion, Christianity, with its division of experience between religious and mundane. He insists that the object of study for archaeologists must be religion, not ritual. However, he declines to give a precise definition of 'religion,' arguing that "in many respects it is indefinable, being concerned with thoughts, beliefs, actions and material … religion also includes the intangible, the irrational, and the indefinable" (Insoll, 2004: p. 7). Insoll identifies the visibility of religion, the role of myth, ritual, temporality, its relation to identity, 'belief/emotion/experience,' and the status of 'the numinous/the holy' as among the questions that any archaeological analysis of religion needs to address (Insoll, 2004: pp. 148–150).

One way other archaeologists have directly addressed the experiential dimension of religion is by relying on universal neurobiological responses to stimuli that are materially evident, or can be assumed to have been present. The most developed examples of this approach concern experiences of altered states (such as trances), especially associated with models of shamanic ritual practice (Lewis-Williams, 2002). Trance, induced through a variety of means (including ingestion of drugs, music and dancing, and even specific bodily stances) has been found to produce the same sequence of effects on the human body. These effects include visualization of geometric shapes such as wavy lines, called entoptic phenomena, which are interpreted by the brain as intelligible forms, often combining human and animal features. The person ultimately experiences sensations of physical movement, often described as flying. Lewis-Williams (2002) and others argue that various forms of rock art, including spirals and animal-human hybrids, are related to these neurobiological experiences. The argument that these experiences are grounded in universal biological capacities provides a means for these archaeologists to argue that ancient subjects were experiencing particular sensations that might be considered numinous.

As archaeologists have turned their attention to religious experience, some have begun to engage with specialists in religious studies. Recent interdisciplinary collaborations have directly addressed such questions as belief in spiritual domains and the emergence of human representations (Renfrew and Morley, 2007). One outcome of these interdisciplinary engagements is the identification of a concern with 'modes of

religiosity' in place of either religion itself, or ritual practice. In a joint discussion of the evidence from Neolithic Çatalhöyük, for example, Ian Hodder and Harvey Whitehouse (2010) discuss how the particular evidence at the site for a historical transition from highly marked, but less frequent, ritual settings, to more frequent, structured, and perhaps narrative religious practices might be related to Whitehouse's concepts of imagistic and doctrinal 'modes of religiosity.' Imagistic modes are understood to be more idiosyncratic, and emotionally charged, while doctrinal modes are abstracted and generalized. Whitehouse's discussion makes it possible to recognize that most archaeology of ritual involves doctrinal modes of religiosity, and leaves out the more personal but perhaps sometimes more compelling experiences he describes as imagistic.

Symbolism, ritual, and religion are clearly subject to study through material traces. Semiotic analyses may play a stronger role in the future in encouraging understanding of how religious meaning is produced (Schoenfelder, 2011). Personal experience, religiosity, and the ways objects and deposits are enlisted in religious practices will likely continue to be focal as archaeologists continue to explore ritual and religion from the perspective of the individual human agent and institutional religion.

See also: Belief, Anthropology of; Cognitive Archaeology; Death, Anthropology of; Practice: An Anthropological Approach; Religion, Anthropology of; Ritual; Sacrifice; Shamanism; Symbolism in Anthropology.

Bibliography

Bauer, A., 2002. Is what you see what you get? Recognizing meaning in archaeology. Journal of Social Archaeology 2, 37–52.
Bell, C., 1992. Ritual Theory, Ritual Practice. Oxford University Press, Oxford.
Binford, L.R., 1962. Archaeology as anthropology. American Antiquity 28, 217–225.
Bradley, R.J., 1991. The Passage of Arms: An Archaeological Analysis of Prehistoric Hoards and Votive Deposits. Cambridge University Press, Cambridge.
Bradley, R.J., 2003. A life less ordinary: the ritualization of the domestic sphere in later prehistoric Europe. Cambridge Archaeological Journal 13, 5–23.
Chapdelaine, C., 2011. Recent advances in Moche archaeology. Journal of Archaeological Research 19, 191–231.
Chenoweth, J., 2009. Social identity, material culture, and the archaeology of religion: quaker practices in context. Journal of Social Archaeology 9, 319–340.
Drennan, R.D., 1983. Ritual and ceremonial development at the hunter gatherer level, and ritual and ceremonial development at the early village level. In: Flannery, K.V., Marcus, J. (Eds.), The Cloud People. Academic Press, New York, pp. 30–32, 46–50.
Fogelin, L., 2007. The archaeology of religious ritual. Annual Review of Anthropology 36, 55–71.
Gilchrist, R., 1994. Gender and Material Culture: The Archaeology of Religious Women. Routledge, London.
Hawkes, C., 1954. Archaeological theory and method: some suggestions from the old world. American Anthropologist 56, 155–168.
Hodder, I. (Ed.), 1982a. Symbols in Action: Ethnoarchaeological Studies of Material Culture. Cambridge University Press, Cambridge.
Hodder, I. (Ed.), 1982b. Symbolic and Structural Archaeology. Cambridge University Press, Cambridge.
Insoll, T., 2001. Archaeology and World Religion. Routledge, London.
Insoll, T., 2004. Archaeology, Ritual, Religion. Routledge, London.
Joyce, R.A., 2007. Figurines, meaning, and meaning-making in early Mesoamerica. In: Renfrew, C., Morley, I. (Eds.), Material Beginnings: A Global Prehistory of Figurative Representation. McDonald Institute for Archaeological Research, Cambridge, pp. 107–116.
Joyce, R.A., 2008. Practice in and as deposition. In: Mills, B., Walker, W. (Eds.), Memory Work. School of American Research Press, Santa Fe, NM, pp. 25–40.
Kyriakidis, E. (Ed.), 2007. The Archaeology of Ritual. Cotsen Institute of Archaeology. University of California, Los Angeles, CA.
Lele, V.P., 2006. Material habits, identity, semeiotic. Journal of Social Archaeology 6, 48–70.
Lewis-Williams, J.D., 2002. A Cosmos in Stone: Interpreting Religion and Society through Rock Art. Altamira, Walnut Creek, CA.
Moore, J.D., 1996. Archaeology of plazas and the proxemics of ritual: three Andean traditions. American Anthropologist 98, 789–802.
Preucel, R.W., 2010. Archaeological Semiotics. Willey-Blackwell, Malden, MA.
Preucel, R.W., Bauer, A., 2001. Archaeological pragmatics. Norwegian Archaeological Review 34, 85–96.
Rappaport, R., 1979. Ecology, Meaning and Religion. North Atlantic Books, Richmond, CA.
Renfrew, C., 1985. The Archaeology of Cult: The Sanctuary at Phylakopi. Thames and Hudson, London.
Renfrew, C., 1994. The archaeology of religion. In: Renfrew, C., Zubrow, E. (Eds.), The Ancient Mind: Elements of Cognitive Archaeology. Cambridge University Press, Cambridge, pp. 47–54.
Renfrew, C., Morley, I. (Eds.), 2007. Material Beginnings: A Global Prehistory of Figurative Representation. McDonald Institute for Archaeological Research, Cambridge.
Richards, C., 1996. Monuments as landscape: creating the center of the world in late Neolithic Orkney. World Archaeology 28, 190–208.
Richards, C., Thomas, J., 1984. Ritual activity and structured deposition in later Neolithic Wessex. In: Bradley, R., Gardiner, J. (Eds.), Neolithic Studies: A Review of Some Current Research, B.A.R. British Series, vol. 133. British Archaeological Reports, Oxford, pp. 189–218.
Robb, J.E., 1998. The archaeology of symbols. Annual Review of Anthropology 27, 329–346.
Schele, L., Miller, M.E., 1986. The Blood of Kings: Dynasty and Ritual in Maya Art. Kimbell Art Museum, Ft. Worth, TX.
Schoenfelder, J., 2011. The semiotics of Gunung Kawi: materializing affiliation at ancient and modern Balinese temples. In: Rowan, Y. (Ed.), Beyond Belief: The Archaeology of Religion and Ritual. Archeological Papers of the American Anthropological Association, vol. 21. American Anthropological Association, Arlington, VA, pp. 147–165.
Walker, W.H., 1998. Where are the witches of prehistory? Journal of Archaeological Method and Theory 5, 245–308.
Whitehouse, H., Hodder, I., 2010. Modes of religiosity at Çatalhöyük. In: Hodder, I. (Ed.), Religion in the Emergence of Civilization: Çatalhöyük as a Case Study. Cambridge University Press, Cambridge, pp. 122–145.
Wobst, M.H., 1977. Stylistic behavior and information exchange. In: Cleland, C.L. (Ed.), For the Director: Research Essays in Honor of James B. Griffin. University of Michigan Museum of Anthropology Papers in Anthropology, No. 61. University of Michigan Museum of Anthropology, Ann Arbor, MI, pp. 317–342.
Wylie, A., 1985. The reaction against analogy. Advances in Archaeological Method and Theory 8, 63–111.

Robinson, Joan (1903–83)

Geoffrey C Harcourt, University of Cambridge, Cambridge, UK
PM Kerr, Nedlands, WA, Australia

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13375–13379, © 2001, Elsevier Ltd.

Abstract

Joan Robinson was born into an upper middle class English family with a tradition of dissent. She was educated at St. Paul's Girls' School and Girton College Cambridge where she read economics in the 1920s. After 2 years in India she made her home in Cambridge as a member of the Faculty of Economics and Politics from 1934 on. She made major contributions in virtually all branches of economic theory, becoming progressively more critical of mainstream neoclassical economics as the years wore on. Her approach increasingly derived from the Classicals, Marx, Kalecki, and Keynes of *The General Theory*. She was an inspired teacher. She died in August 1983.

Joan Robinson was born on 31 October 1903, in Camberley, Surrey, into an upper middle-class English family characterized by vigorous dissent and independence of mind. Her great-grandfather was F. D. Maurice, the Christian Socialist who lost his Chair of Theology at King's College, London, on a point of principle; her father was Major General Sir Frederick Maurice, the victim of the infamous Maurice debate in 1918 and subsequently Principal of what is now Queen Mary College in the University of London. Her mother was Helen Margaret Marsh, the daughter of Frederick Howard Marsh, Professor of Surgery and Master of Downing at Cambridge. Joan Robinson was born into a milieu of controversy. She was educated at St. Paul's Girls' School in London, where she read history, and from 1922 to 1925 at Girton College, Cambridge, where she was Gilchrist Scholar. There she read for the Economies Tripos, searching for rational argument and hoping to understand poverty and its cures. She graduated in 1925 with second-class honours ('a great disappointment' – her words) although it was not until 1948, when women were given full membership of the university, that she was admitted to the degree. In 1926, she married Austin Robinson, the Cambridge economist, and the couple sailed for India for what was to become the first of many visits. They stayed for over 2 years and returned to Cambridge in 1929, making it their intellectual and domestic base for the rest of their lives. They had two daughters and five grandchildren. Joan joined the Cambridge Faculty of Economics and Politics as a Faculty Assistant Lecturer in 1934, becoming part of an outstanding group of scholars in Maynard Keynes, A. C. Pigou, Dennis Robertson, Austin Robinson, Maurice Dobb, Richard Kahn, Gerald Shove, and Piero Sraffa. She became a University Lecturer in 1937, Reader in 1949, and Professor in 1965. She was elected to a Professorial Fellowship at Newnham and made Honorary Fellow of Girton in 1965, and Honorary Fellow of Newnham in 1971; she was the first woman to be made an Honorary Fellow of King's (in 1979). From 1958 to 1971 she was an FBA. She 'retired' from her Chair in 1971, remaining as active as ever in Cambridge and abroad, until her death in Cambridge on 3 August 1983.

Early in her career Joan Robinson set out her views on methodology for economics, (Robinson 1932a, Harcourt 1990a). Her procedure was to identify and make explicit her assumptions and to develop a 'box of tools' for understanding and exploring the analytical object, to construct a logically sound analysis of the processes concerned: these successive stages she saw as independent of each other. Extending these views, she suggested (1937b, pp. 99–102) that controversy could result from lack of understanding, lack of knowledge, errors of logic, from working from different assumptions but, most commonly in economics, from differences in (political) objectives: the goals of a theory could direct its analysis as well as its policy conclusions. Throughout her life she argued that economics is not neutral, it is always ideologically driven. Nevertheless she believed that it was possible to develop pure analysis independently of dispute.

Her early contributions to economic theory were mainly in the realm of value theory (microeconomics, as it is called now) and *The Economics of Imperfect Competition* (1933), her first book, generously acknowledged her debts to Marshall (via Pigou), Kahn, Austin Robinson, Shove, and Sraffa. It was a controversial and much celebrated success, revealing her great clarity of thought and exceptional analytical powers. It was her interpretation, within a Marshallian context, of several contemporary theoretical issues. In particular, it was a response to the heretical claims about decreasing costs and increasing returns being made in the *Economic Journal*, especially Sraffa (1926), Symposium (see Robertson et al., 1930), and including Robinson (1932b). Robinson replaced an analytical basis in Marshall's free competition with one in imperfect competition, where free competition became a special case. She also moved the focus from Marshall's long period to a short period. Within that context the book also presented new concerns. On the basis of the downward sloping demand curve for each firm, Robinson introduced the marginal revenue curve as an integral analytical construct. She showed how the response of relative factor shares to a change in demand depended on whether you used her partial analysis of the elasticity of substitution or Hicks', aggregate. She reconceived Marshall's 'representative' firm as an 'equilibrium' firm where each such firm faced a less than infinitely elastic demand curve for its product in a monopolistic market. What then became 'the industry' in such a market? And was the simple profit-maximizing rationale too reductionist to depict the firm? How was a monopolist to be identified theoretically?

Or a duopolist? How indeed was imperfect competition definable?

The Economics of Imperfect Competition provoked; each stage of the theoretical process, even her geometrical method, incited reaction. More controversial than her analysis were its implications for policy. She represented perfect competition theory as demonstrating that the economy left alone would generate the optimal outcomes: it suggested a political philosophy of *laissez faire*. Imperfect competition showed that left alone the economy could become stuck at a position of unemployed capacity and labour and indeed, at that time, this was the case. The debates on the elasticity of substitution further illustrated what the emerging critique of perfect competition theory could mean for the status quo. The issues raised in this book were to fill the journals for decades to come. She herself came to reject its analyses (postwar) as a 'shameless fudge.'

Robinson's interests were not confined to microeconomics. During this time she was a member of a group of economists who met to discuss John Maynard Keynes's *Treatise on Money* (1930) and his emerging new theory. His ideas appealed to her, they represented a break from the orthodox and they were focused on the pressing problem of the persistence of high levels of unemployment. As a member of the Circus, with Meade, Kahn, Austin Robinson, and Sraffa, Robinson contributed to the critical scrutiny of drafts of Keynes's *The General Theory* (1936). A formidable group, but they were persuaded by the dissenting view of the behaviour of investment and of saving in Keynes's new theory of output and employment. Immediately following the book's publication, she wrote to develop its themes and analysis. *Essays in the Theory of Employment* (1937a) and the popular *Introduction to the Theory of Employment* (1937b) were both published in the year following *The General Theory* (and in the same year that her second daughter was born). She sent Keynes drafts of the two books as well as the proofs, and the related correspondence reveals his respect for Robinson in her role as his interpreter. *Introduction* is still one of the best guides to *The General Theory*.

During the interwar years, Michal Kalecki had a major impact on Robinson's thinking. His preemptive version of the theory of effective demand, which interwove imperfect competition with a macro-economic theory of distribution, was based on Karl Marx's schema of reproduction and his analysis of the realization crisis. She integrated his double-sided relation between investment and profits and his short-period theory of income distribution to develop her theory of employment and output. Kalecki provided a structural base to Robinson's theoretical explorations and extensions of Keynesianism and later of growth. During the war Robinson wrote *An Essay on Marxian Economics* (1942). Her interest was partly attributable to Kalecki's and Sraffa's Marxist-based work, and partly to the structure of society portrayed by, and fundamental to, such classical and Marxist analysis; her continued dissatisfaction with orthodox theory was also reflected in her essay which compared Marxian with mainstream ideas.

Robinson's interpretation of Marx again illustrates her approach to constructing theory. Responses to her book were to her methodology, as well as to her critical remarks about both Marxian and neoclassical economics. She extracted from Marx's writing the circuits of deductive logic with their attendant assumptions and political implications. The methodology of her censorship was not so inclusive as to embrace Marx's theory of history and of the processes of change. Yet it was these elements – the centrality of historical analysis to Marx's method and of class to understanding effective demand – that attracted Robinson to Marx's writings. The pervasive implications of time and history underlay her vision of capitalism and informed all of her subsequent economic research even though she ultimately was to abandon the classical method. Her book was a challenge to her colleagues to read and learn from Marx, if selectively and from a Keynesian perspective. It remains a useful introduction to Marxist economics.

In the early postwar years Robinson became committed to developing a dynamic analysis of capitalism. Inspired by Harrod's (1939, 1948) studies of long-period growth, *The Rate of Interest and Other Essays* (1952) was a collection in preparation for this project, in particular. The generalization of *The General Theory* attempted to extend Keynes into the long period. This essay was an interpretation of Keynes in terms of the Marshallian short period, characterized by uncertainty, incompatible decisions and unrealizable expectations, by a new definition of equilibrium. The elements of Robinson's growth theory were coming together in a non-neoclassical and long-period framework.

Integral with Robinson's exploration of long-period accumulation was an increasing dissatisfaction with the conceptual and logical foundations of neoclassical theory and its method, for which she became most well known, if not best understood. She prefaced her constructive contributions to dynamic theory with the critique of the neoclassical production function, Robinson (1953–54). This article fired a debate that continued for more than two decades. The implications of these exchanges were for the legitimacy of the very concept of capital and hence of the aggregate production function, fundamental to neoclassical theory. In particular, Robinson targeted its distribution analysis and its corollary, value theory. Ultimately, she criticized its method, that "the neo-neoclassics still refuse to understand the difference between a comparison of timeless equilibrium positions and the effects to be expected from a change taking place at a particular moment" (1978a, p. xix). That particular moment was characterized by its own inherited structure of decisions, expectations, and uncertainties. She rejected the neoclassical method of long-period equilibrium as a point toward which the actual economy is moving; the implied stationarity of that theory meant that the movement toward it, and the rest point were logical, not historical. Her own account would not be in equilibrium terms but would take into account the actual features of the economy, the existence of uncertainty, and disappointed expectations. Her understanding of capitalism was based on classes and relations of conflict, while the neoclassicals begin with technical factors and harmonious distributive shares. What she wanted to put in the place of neoclassical theory was not a rival complete theory, which she considered would be only another 'box of tricks,' but an approach that started from institutions, history, and the 'rules of the game' (Harcourt, 1990b). Robinson's optimism about the power of economic debate was continually frustrated, yet she determinedly defended what seemed to her to be right and

just arguments. Her conservative colleagues, she felt, treated new ideas 'like heresies.'

The Accumulation of Capital (1956) continued the 'generalization' within a classical framework. The questions she set out to explore were: what are the conditions for cumulative long-term growth of income and capital (Golden Age) and what is the outcome of this process in terms of growth of gross and net output and the distribution of income between wages and profits, given a certain evolution through time of the labor force and technology? Inspired by Harrod's 'fruitful provocation,' *The Accumulation of Capital* brought together Marx's historicism, Harrod's economic dynamics, and Wicksell's capital theory. True to her methodology, she started out by presenting her most basic model and then devoted the rest of this remarkable book to relaxing the 'most severely simple' assumptions and tracing out the implications. 'The whole argument,' she proposed, "is set out ... as an analytical construction, with the minimum of controversy" (1956, p. x). In contrast to the mathematical style endemic to American economics, Robinson wrote her entire work of pure theory in literary form. It was a difficult book to assimilate and several years later she followed it up with a hands-on guide. *Exercises in Economic Analysis* (1960) fills in the gaps, clarifies obscurities, and extends the arguments from *The Accumulation of Capital*. Its object, Robinson stated, was 'to sort out questions' (1960, p. xx). It offers readers an opportunity to formulate for themselves the analytical examples she derives in *The Accumulation of Capital*. This book (1960) was immediately followed by *Essays in the Theory of Economic Growth* (1962) which presents a retrospective introduction for the 'too terse' earlier (1956) book. Although there was a hint of re-switching in Robinson (1953–54) as well as in *The Accumulation of Capital*, Robinson dismissed the phenomenon at the time and saw it as a distraction from the main methodological point of her critique of neoclassical capital.

Dissatisfied with her quest for a long-period theory of growth, Robinson looked more to Kalecki's theory of cycles as a depiction of capitalism. She was a pioneer in the development of what was to become post-Keynesian theory, its bases in Sraffa's interpretation of Ricardo and in Kalecki's interpretation of the theory of employment. To this she brought 'a general framework of long- and short-period analysis which will enable us to bring the insights of Marx, Keynes, and Kalecki into coherent form and apply them to the contemporary scene' (1978b, pp. 1, 18). Within this framework, post-Keynesian theory offered a 'method of analysis which takes account of the difference between the future and the past' (1978b, p. 12). Here was a broad conceptual structure that brought together Robinson's major theoretical concerns. From this analysis, post-Keynesian theory is evolving in several, not always compatible, directions, establishing its authority in the body of economic theory.

Robinson was a charismatic teacher; she inspired successive generations of students to develop their own ideas from her methods of analysis and brilliant insights, to think critically and for themselves. She spent many years demonstrating that 'mainstream teaching has been inculcating defective methodology'; she wanted to provide a logically sound, historically based dynamic approach to first-year economic analysis. *An Introduction to Modern Economics* (1973) with John Eatwell was a response to the 'fashionable textbooks [of] mainstream' economics. It presented a history of thought, a classical-cum-Kaleckian/post-Keynesian analysis. It was an ambitious, presumptuous book, at once a critique of neoclassical theory, and a terse introduction to the alternative. Robinson overestimated the intuitive powers of her audience and it proved too difficult for beginners; it remains a challenge for more advanced students.

Throughout her life Robinson pursued an active interest in the Third World, in India and China in particular, and in the Soviet Union. Her initial writing on development was based on the growth models developed in *The Accumulation of Capital*. Kalecki influenced this book. Development, she argued, should proceed by appropriate land reform and industrialization in both agriculture and manufacture, mindful of the balance of payments constraint and a minimum real wage or its equivalent. Socialist development, she continued, could mobilize savings and enable the state to organize the investment of the surplus, so achieving faster if not more efficient accumulation and growth than under capitalism; socialism, furthermore, led to a more equitable distribution of income (although the standard of living she saw as a political decision). She was less enthusiastic about the Soviet-style socialism of Eastern Europe than she was about China. Her mission to eradicate poverty and to meet basic needs always guided her judgements. She leaves a body of work; some reprinted in her *Collected Economic Papers* as well as Robinson (1979), on development planning and socialism that is relevant and provocative.

Joan Robinson depicted her work as an agenda for discussion as well as formulating the questions. Yet she has left an important legacy in her substantial literary output. She wrote for colleagues, students, and for the general, interested reader. As well as her many books, there is a stream of challenging evaluative essays written through the 50 years of her career. Her writing includes a theoretically rigorous critique of neoclassical economics that as yet goes unanswered; the questions she has raised present a formidable program for further research. She leaves an exacting methodology, a theoretical structure, and significant analytical details of a new, unified system of political economy.

See also: Capitalism: Global; Capitalism; East Asian Studies: Economics; Economic Growth: Measurement; Economic Growth; Economic History; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Employment and Labor, Regulation of; Labor Supply; Market Structure and Performance; Political Economy, History of; Unemployment: Structural.

Bibliography

Harcourt, G.C., 1990a. Joan Robinson's early views on method. History of Political Economy 22, 411–427.

Harcourt, G.C., 1990b. Robinson, Joan Violet. In: Blake, Lord, Nichols, C.J. (Eds.), The Dictionary of National Biography 1981–1985. Oxford University Press, Oxford, UK, pp. 346–347.

Harcourt, G.C., 1996. Some reflections on Joan Robinson's changes of mind and their relationship to post-Keynesianism and the economics profession. In:

Marcuzzo, M.C., Pasinetti, L., Roncaglia, A. (Eds.), The Economics of Joan Robinson. Routledge, London, pp. 317–329.
Harrod, R.F., 1939. An essay in dynamic theory. Economic Journal 49, 14–33.
Harrod, R.F., 1948. Towards a Dynamic Economics. Macmillan, London.
Keynes, J.M., 1930. A treatise on money, vol. 2. In: Collected Writings, vols. V, VI. Macmillan, London.
Keynes, J.M., 1936. The general theory of employment, interest and money. In: Collected Writings, vol. VII. Macmillan, London, 1973.
Robertson, D.H., Shove, G.F., Sraffa, P., 1930. Increasing returns, and the representative firm: a symposium. Economic Journal 40, 79–116.
Robinson, J., 1932a. Economics is a Serious Subject: The Apologia of an Economist to the Mathematician, the Scientist and the Plain Man. Heffer, Cambridge, UK.
Robinson, J., 1932b. Imperfect competition and falling supply price. Economic Journal 42, 544–554.
Robinson, J., 1933. The Economics of Imperfect Competition. Macmillan, London.
Robinson, J., 1937a. Essays in the Theory of Employment. Macmillan, London.
Robinson, J., 1937b. Introduction to the Theory of Employment. Macmillan, London.
Robinson, J., 1942. An Essay on Marxian Economics. Macmillan, London.
Robinson, J., 1952. The Rate of Interest and Other Essays. Macmillan, London.
Robinson, J., 1952–1980. Collected Economic Papers, vols. I–VI. Blackwell, Oxford, UK.
Robinson, J., 1953–1954. The production function and the theory of capital. Review of Economic Studies 21, 81–106.
Robinson, J., 1956. The Accumulation of Capital. Macmillan, London.
Robinson, J., 1960. Exercises in Economic Analysis. Macmillan, London.
Robinson, J., 1962. Essays in the Theory of Economic Growth. Macmillan, London.
Robinson, J., 1978a. Contributions to Modern Economics. Blackwell, Oxford, UK.
Robinson, J., 1978b. Keynes and Ricardo. Journal of Post Keynesian Economics 1, 12–18.
Robinson, J., 1979. Aspects of Development and Underdevelopment. Cambridge University Press, Cambridge, UK.
Robinson, J., Eatwell, J., 1973. An Introduction to Modern Economics. McGraw-Hill, Maidenhead, UK.
Sraffa, P., 1926. The laws of returns under competitive conditions. Economic Journal 36, 535–550.

Robustness

Stephan Morgenthaler, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Published by Elsevier Ltd.

Abstract

Robustness is a crucial quality indicator of any procedure used in a data analysis. The most basic attribute of robustness of such a procedure is its sensitivity (or influence function). The sensitivity assesses the effect of an additional observation on the results of the data analysis, if that observation takes an arbitrary fictitious value u. If we estimate something with an estimator that has bounded sensitivity, the results are more trustworthy. If the estimator has unbounded sensitivity, we cannot be confident because a single observation may have determined some of the conclusions.

The breakdown point of a statistical method is the minimal fraction of the observations that, when manipulated, have the potential to render a data analysis meaningless. This is an example of a global attribute of robustness. It goes beyond sensitivity in an important respect, because it assesses the behavior of a procedure if a sizable proportion of the database were arbitrary and fictitious, not only one observation.

Contaminated distributions such as $(1-\varepsilon)\Phi(\gamma)+\varepsilon H(\gamma)$ are useful in modeling such fictitious and arbitrary values. With a (large) probability of $(1-\varepsilon)$, this mixture distribution results in an observation drawn from a normal distribution Φ, whereas with a (small) probability ε the contaminating distribution H generates the observation. Such models can provide useful checks for choosing estimators. While the arithmetic mean is the correct choice of estimator and if the observations have random errors with $\varepsilon=0$, it turns out that even for tiny positive ε this is no longer true. The arithmetic mean's place as the best choice is then taken by the Huber estimator.

We all routinely use the arithmetic mean for estimating the expectation of a random variable based on a sequence of observations. It is indeed the best way to summarize observations, if they are distributed according to the bell-shaped normal density. Tiny deviations of the actual distribution from the normal ideal, however, can turn the arithmetic mean into a horribly bad estimator even though the observations come from a density that is very nearly a normal density. This illustrates what is meant by robustness of an estimator. The estimator's properties must be stable (or robust) under deviations from the postulated stochastic model. What I had in mind when modeling the data is different from the actual stochastic properties of the observed data, even if the model is a very good one. These differences should not be allowed to affect the performance of the procedure used in the data analysis.

The robustness of estimators or other procedure may also be defined without reference to a model. The arithmetic mean of the three observations (14.0%, 13.7%, 26.3%), for example, is equal to 18.0%. The lack of stability of this estimated average percentage is visible through the large effect of the single observation 26.3% on the estimated value.

Implicit and explicit assumptions of a data analysis based on a stochastic model are of various types. In a generalized linear model, for example, they include the link function, which determines the scale on which the relationship between the response and the explanatory variables (or covariates) is of the form of a linear combination of the covariates. Another assumption concerns the cumulative distribution for the response, which determines certain aspects of the inherent randomness of the data, such as the form of the relationship between the standard deviation and the expectation of the response variable or the rate of growth of the extreme $(1-\alpha)$ quantile of the response variable as α tends to zero. The best-known phenomenon observed in data concerns the so-called outliers. This is a term used in describing various incidents in a database. It may refer to gross errors in observations due to faulty transcription, mistakes in the manipulation of a measurement instrument, errors in the preparation of a particular experiment, confusion as to which experiment should have been performed, or some other such mistake. The frequency of such gross errors is much higher than one would expect at first thought, even in data obtained through a carefully designed experiment and of course probably even more so in observational studies. It is also possible that an atypical observation is an actual measurement in situations where the range of the errors of the instrument is simply much larger than we expected. The term outlier may also refer to a situation, in which a sizable proportion of the data seems to behave differently from the rest. In a linear regression, for example, it is sometimes observed that any one regression surface fits only part of the data. This occurs with observational studies or designed experiments, if important covariates are missing and the data cannot be separated into more homogeneous pieces based on the available information.

Of course, the data can be used to check assumptions and it is certainly a good statistical practice to do so. To check the details of a model, however, requires huge amounts of data and it is misleading to suggest that diagnostic checking of model assumptions can solve the robustness problem. Good practice in statistics requires the use of methods that do not suffer the drawbacks of the arithmetic mean of loosing performance under tiny violations of the idealistic assumptions. This is especially true in routine and less thoroughly scrutinized applications of statistical methodology. Robust methods retain very good performance, while being less sensitive to assumptions.

In the future, data analysis will increasingly be applied to very large databases composed of data from multiple sources of uneven quality. In such applications, the call for robust procedures will grow louder.

Robustness seems similar to nonparametric or, more precisely, distribution-free statistics. This is a superficial and even wrong view. Converting data to ranks, for example, in order to liberate the statistical analysis from distributional assumptions, does not give particularly good protection from the deleterious effects of gross errors.

Indicators of Robustness

Sensitivity Functions

Let $T(y_1, \ldots, y_n)$ be a statistic based on the observations $(y_1, \ldots, y_n)$ and suppose that T is estimating some unknown quantity. The sensitivity function is a simple idea leading to a better understanding of the behavior of the statistic. The question being asked is: how would the estimator have come out, if the data had not been exactly the ones I observed? To answer it, one considers the sensitivity function,

$$\mathrm{SC}(u) = (n+1)(T(u, y_1, \ldots, y_n) - T(y_1, \ldots, y_n))$$

which is the change in the estimate induced by an additional fictitious observation u. The multiplier $(n+1)$ in this formula provides a natural scaling of the size of the change. Inversely, it follows from this definition that

$$T(u, y_1, \ldots, y_n) = T(y_1, \ldots, y_n) + \frac{1}{n+1}\mathrm{SC}(u)$$

which is a formula for updating an estimator is a new observation that becomes available. Examples, for the arithmetic mean $T(y_1, \ldots, y_n) = \bar{y} = (y_1 + \cdots + y_n)/n$ one finds

$$\mathrm{SC}(u) = (n+1)\left(\frac{u + y_1 + \cdots + y_n}{n+1} - \frac{y_1 + \cdots + y_n}{n}\right) = u - \bar{y}$$

This sensitivity function shows that the sensitivity of the arithmetic mean equals the distance of the additional observation u from the average of the observed data. If all values of u are considered, the sensitivity is unbounded.

For the trimmed mean, $T(y_1, \ldots, y_n) = (y_{(2)} + \cdots + y_{(n-1)})/(n-2)$, obtained by deleting the largest $(y_{(n)})$ and the smallest $(y_{(1)})$ observations before taking the arithmetic mean, the sensitivity depends on the position of the additional observation u with regard to the initial data. If u is between $y_{(1)}$ and $y_{(n)}$, it will not be trimmed from the sample and $\mathrm{SC}(u) = (n+1)((y_{(2)} + \cdots + y_{(n-1)} + u)/(n-1) - (y_{(2)} + \cdots + y_{(n-1)})/(n-2)) = (n+1)(u - T(y_1, \ldots, y_n))/(n-1)$. If on the other hand, u is larger than $y_{(n)}$, then u will be trimmed and the sensitivity no longer depends on u and is equal to $\mathrm{SC}(y) = (n+1)((y_{(2)} + \cdots + y_{(n)})/(n-1) - (y_{(2)} + \cdots + y_{(n-1)})/(n-2)) = (n+1)(y_{(n)} - T(y_1, \ldots, y_n))/(n-1)$. An analogous formula holds when u is smaller than $y_{(1)}$. Merely by trimming the extremes from the observations before computing the arithmetic mean, we can thus bound the sensitivity, and in view of the desirability of using robust estimators, the trimmed mean is much better than the arithmetic mean. But we cannot stop here. The trimmed mean is much better, but is it good enough?

As a third example, we now consider the least-squares estimator of a slope (or a regression coefficient) in the model $y = \alpha + \beta x +$ random error. In this case, we observe the couples of a covariate and a response, $(x_1, y_1), \ldots, (x_n, y_n)$, and we add a fictitious observation at (u, v). Recall that the least-squares estimate of the slope β is $\widehat{\beta} = T(y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i(x_i - \bar{x})/\sum_{i=1}^{n}(x_i - \bar{x})^2$, which can be simplified by centering the covariate $(\bar{x} = 0)$. Assuming a centered covariate, the estimate of α is $\widehat{\alpha} = \bar{y}$. Computing the sensitivity $\mathrm{SC}(u, v)$ caused by an additional fictitious observation is complicated and we leave it to the interested readers as an exercise. The result is approximately proportional to a product of two terms, the residual $(v - \widehat{\alpha} - \widehat{\beta}u)$ (residual with respect to the regression line based on the initial data) and a term that is essentially linear in u. As in the other two cases, this is a revealing formula. It makes it clear that the least-squares regression coefficients are sensitive to the outlyingness of the additional observation with regard to the regression line of the initial data as measured by the residual. But, this effect is multiplied by the position u, the further u is placed from the mean of the initial covariates, the more the effect of outlyingness is leveraged. Adding an outlying observation at or beyond the edge of the initial covariate values is much worse than adding it centrally.

In practice, the sensitivity function can easily be computed numerically for any given estimator and underlying data set and one can thus construct the plot of $\mathrm{SC}(u)$ versus u, which in many cases will serve as a good basis for judging the sensitivity robustness of the estimator. From such a plot, one can deduce several indicators, the simplest one being the maximal sensitivity

$$g^* = \sup_y |\mathrm{SC}(y)|$$

where $|\mathrm{SC}|$ denotes the absolute value of SC. In the above examples, the maximal sensitivity is equal to ∞ for the mean and the least-squares estimate of the regression coefficient – the first and the third example. It is finite for the trimmed mean. An estimator with infinite maximal sensitivity is necessarily nonresistant in the sense that a single wild observation that behaves differently from the rest can destroy the estimate and lead to arbitrary conclusions based on a single observation, conclusions not supported by the remainder of the data.

Influence Functions

The sensitivity function has a more formal counterpart in the influence function, which is defined as a derivative of a statistical functional. In many cases it is simply the limit as the sample size n tends to infinity of the sensitivity function computed for a sample generated from a fixed model.

Examples

Let $y_1, \ldots, y_n$ be a sample from some cumulative distribution F with expectation μ_F. As n increases, the sensitivity function of the arithmetic mean converges to the influence function, which is $\mathrm{IF}_F(u) = u - \mu_F$. The influence function of the trimmed mean obtained by trimming a proportion α of the largest and the smallest observations is

$$\begin{aligned}\mathrm{IF}_F(u) &= u - qF\left(\frac{1}{2}\right)/(1-2\alpha), \quad \text{for } qF(\alpha) \le u \le qF(1-\alpha)\\ &= \mathrm{sign}\left(u - qF\left(\frac{1}{2}\right)\right)\left(qF(1-\alpha) - qF\left(\frac{1}{2}\right)\right)/(1-2\alpha),\\ &\quad \text{otherwise}\end{aligned}$$

This formula is correct for symmetric densities f. The symbol $qF(\alpha) = F^{-1}(\alpha)$ denotes the α-quantile of the distribution function.

The influence function is very similar to the sensitivity function and can thus be interpreted similarly. In particular, the gross error sensitivity is

$$\gamma_F^* = \sup_y |\mathrm{IF}_F(y)|$$

It has other uses that can be extremely helpful. It turns out that the squared influence function of an estimator is directly linked to its asymptotic variance. If our stochastic model stipulates that the observations $Y_1, \ldots, Y_n$ are independent and identically distributed with distribution F_θ for some real parameter θ, one can show that most estimators T of θ satisfy the following approximate relationship:

$$T(Y_1, \ldots, Y_n) \approx \theta + \mathrm{ave}(\mathrm{IF}_{F_\theta}(Y_i))$$

If this approximation is sufficiently precise, it implies that the asymptotic variance of the estimate T is

$$\mathrm{Var}_{\mathrm{asy}}(T) = \int (\mathrm{IF}_{F_\theta}(r))^2 f_\theta(r)\,\mathrm{d}r$$

which is, the expected value of the squared influence function. In this case, one can also apply the central limit theorem to conclude that

$$P(T \le t) \approx \Phi\left(\frac{\sqrt{n}(t-\theta)}{\sqrt{\mathrm{Var}_{\mathrm{asy}}(T)}}\right)$$

where Φ denotes the normal distribution function with density $\varphi(t) = \exp(-t^2/2)/\sqrt{2\pi}$. This in turn shows that

$$\left[T \pm 1.96\sqrt{\mathrm{Var}_{\mathrm{asy}}(T)}/\sqrt{n}\right]$$

is an approximated 95% confidence interval for θ. And finally, the arithmetic average of the squared influence evaluated at the observations is an estimate of the asymptotic variance. Equation [5] gives an example.

Breakdown and High Resistance

A limitation of the influence function is its local nature. It is based on the change of the value of the estimator from one sample to another one that differs by adding a single 'contamination.' The breakdown point is the simplest global measure of robustness. It is the largest percentage of wild values in a data set that an estimator can tolerate without giving wild estimates. This notion is inspired by the trimmed means, computed as the mean of a sample after deleting $100\alpha\%$ of the largest and of the smallest data points. This estimator has a breakdown point of $100\alpha\%$, since any contamination by wild data below this fraction is being trimmed. A special case of the trimmed means is the median, which for an odd-sized sample $(n = 2k+1)$ has a breakdown point of $k/n \approx 50\%$. This is so, because the median is computed as the $k+1$st value in the sample and anything less than k wild values will not lead to a wild estimate. More generally, global indicators describe the properties of an estimator under contamination of a sizable portion of the data.

A simple stochastic models that embraces outlying values among random errors are the so-called contaminated distributions of the form $(1-\varepsilon)F + \varepsilon H$. This is a mixture of a 'good' part F and a 'bad' part H. The contaminated normal distribution $(1-\varepsilon)$ Normal $(0, 1) + \varepsilon$ Normal $(0, \sigma^2)$, for example, describes random errors that can either result from a standard normal distribution, or from a normal distribution having a larger variance $\sigma > 1$. As ε grows from 0 to 50%, a larger and larger fraction of the data is expected to exhibit a distinct behavior. A global robustness indicator for estimators is based on behavior under positive fractions ε, whereas local properties, such as the influence function, merely consider limits as ε tends to zero.

An estimator with a breakdown point of 50% is called highly resistant. Such estimators can be very revealing when applied to data containing a large proportion of outliers, because they lead to a clear separation of the 'good' and the 'bad' data. The median is an example of a highly resistant location estimator. In a previous example, we computed the arithmetic mean of the three observations (14.0%, 13.7%, 26.3%) as being 18%. The corresponding residuals are equal to (14.0% − 18.0% = −4.0%, −4.3%, 8.3%). The median of the three observations is equal to 14%, with residuals (14.0% − 14.0% = 0.0%, −0.3%, 12.3%). The highly resistant median produces residuals that make it immediately evident that the third observation is a wild value. The ability of estimators to reveal hidden outliers is explained by the fact that they have bounded sensitivity with regard to single observations, but also, and more importantly, that even a largish fraction of outliers does not exert an undue influence on the estimate.

Another example of a highly resistant estimators is the median absolute deviation (MAD), defined as.

$$\mathrm{MAD}(y_1, \ldots, y_n) = \mathrm{median}(|y_1 - m|, \ldots, |y_n - m|)$$

where m denotes the median of the observations, $m = \mathrm{median}(y_1, \ldots, y_n)$. The MAD can be used instead of the standard deviation as an estimate of σ. (If we want consistency at the normal model, it has to be multiplied by 1.48.) In the example above of the three percentages, the MAD is equal to the median of the absolute values of the residuals, that is, MAD = 0.3%.

Robust Estimators

Least-Squares, Maximum Likelihood, and M-estimators

Least-squares estimators minimize the sum of squared residuals

$$T(y_1, \ldots, y_n) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \mu_i(\boldsymbol{\theta}))^2$$

that is, the estimated value is equal to the value of the parameter that minimizes the sum of squares. The parameters may form a vector, which is the reason for setting it boldface. As a general rule, least-squares estimators should be avoided because they have unbounded influence and a breakdown point of 0%. Maximum likelihood estimators (m.l.e.) for independent data are based on a stochastic model with density $f_i(x, \boldsymbol{\theta})$ for the ith observation and with $\boldsymbol{\theta}$ as the unknown vector of parameters. The m.l.e. is defined as

$$T(y_1, \ldots, y_n) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(f_i(y_i, \boldsymbol{\theta})\right)$$
$$= \text{solution of} \left(\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(f_i(y_i, \boldsymbol{\theta})\right) = 0 \right)$$

that is, the estimator maximizes the logarithm of the likelihood. These estimators can be sensitivity robust, depending on the properties of the densities f_i. But if we use normal, poisson, or binomial models, as is mostly the case, they do not improve on the least-squares estimates.

Generalizing either of these two concepts is the M-estimator, defined by

$$T(y_1, \ldots, y_n) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho_i(y_i, \boldsymbol{\theta})$$

that is, the estimator minimizes an arbitrary sum of functions. Usually, and for computability, the function $\rho_i(y_i, \boldsymbol{\theta})$ is convex in the parameter vector, but otherwise quite free. The corresponding estimating equation for computing the estimator $\boldsymbol{T}$ is

$$\sum_{i=1}^{n} \psi_i(y_i, \boldsymbol{T}(y_1, \ldots, y_n)) = 0$$

where $\psi_i(y, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \rho_i(y_i, \boldsymbol{\theta})$.

Example

Consider the linear regression case in which $\mu_i(\boldsymbol{\theta}) = \boldsymbol{x}_i^T \boldsymbol{\theta}$ is a linear combination of the covariates (which must include a constant value 1, if a constant is part of the regression equation). The superscript T will be used from now on to indicate the scalar product between two vectors. The model-based m.l.e. then has densities $f_i(y_i, \boldsymbol{\theta}) = f(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta})$, where $f(r)$ is the density of the random errors and the function $\rho(r, \boldsymbol{\theta})$ finally is of the form $\rho_i(y_i, \boldsymbol{\theta}) = \rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta})$. It follows that the derivative with regard to the parameters is $\psi_i(y_i, \boldsymbol{\theta}) = \psi(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}) \boldsymbol{x}_i$ with $\psi(r) = \rho'(r)$. An M-estimator $\boldsymbol{T}$ of a vector of regression coefficients $\boldsymbol{\theta}$ thus satisfies the system of equations (one for each regression coefficient)

$$\sum_{i=1}^{n} \psi(y_i - \boldsymbol{x}_i^T \boldsymbol{\theta}) \boldsymbol{x}_i = 0 \qquad [1]$$

The M-estimator with $\rho(r) = r^2/2$ (rho-function) and $\psi(r) = r$ (psi-function) is equal to the least-squares estimator, which is also equal to the m.l.e. for random errors with a normal distribution. Any m.l.e. can be thought of as an M-estimator with $\rho(t) = -\log(f(t))$, but thinking in terms of M-estimators is an advantage and an important conceptual step since it frees us from any artificial constraints. Recall that the variance of the least-squares estimator with random errors with constant variance σ^2 is equal to

$$\sigma^2 \left(\mathrm{X}^T \mathrm{X}\right)^{-1}$$

where X denotes the matrix of the values of the covariates. For general M-estimators, the same type of expression holds for the asymptotic variance, except that σ^2 is replaced by

$$\frac{\int_{-\infty}^{\infty} \psi^2(r) f(r) \mathrm{d}r}{\left(\int_{-\infty}^{\infty} \psi'(r) f(r) \mathrm{d}r\right)^2} \qquad [2]$$

where $f(r)$ is the density of the random errors. The fact that the covariate part of the asymptotic variance remains unchanged shows that M-estimators of this type do nothing to control the leverage exerted by outlying values of the covariates. For this reason, these estimators have bounded influence and good breakdown behavior only if such outlying values are excluded. If arbitrary fictitious covariate values are considered, then the influence is also unbounded and the breakdown point is 0.

The M-estimators are defined by a rho-function and an associated psi-function. Boundedness of the psi-function is necessary for robustness. Careful tuning of the shape of the psi-function permits both robustness and efficiency. It is well-known that the least-squares estimator, which has psi-function $\psi(r) = r$, is optimally efficient when the error distribution is exactly Gaussian (i.e., normal). To preserve a high Gaussian efficiency, the psi-function should remain unaltered near zero, that is, $\psi(r) \approx r$ for small r. Three standard choices are given in Table 1. The M-equation [1] must be adapted to the scale of the random errors. This requires an estimate S of the residual scale, such as the MAD of the residuals and the M-equation becomes

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{T}}{k \times S}\right) \boldsymbol{x}_i = 0 \qquad [3]$$

The multiplier k of the scale S can be any positive number and determines the loss of Gaussian efficiency, which is biggest for small k. The Huber and Tukey estimators give values very close to the least-squares estimator when k is large. This is not true for the least absolute deviations estimator which is unaffected by S and k. A reasonable compromise choice for the denominator $k \times S$ is $1.3 \times$ MAD (residuals) for the Huber estimator and $7 \times$ MAD (residuals) for the Tukey estimator. Some statistical packages offer the choice of robust regression estimators instead of the least-squares solution. An example is the reg function in the R package (reference).

Theoretical Justification of Huber's Estimator

A concise discussion of many of the topics discussed in this article can be found in the two books by Huber (1981,

Table 1 The Huber function satisfies $\psi_H(r) = r$ for r between -1 and 1 and outside of this interval, the function is constant equal to the sign of r. This gives an S-shape to this function. The Tukey function also satisfies $\psi_T(r) \approx r$ for small r and it also satisfies $\psi_T(-r) = \psi_T(r)$ as does the Huber function. However, the Tukey function is equal to zero for r bigger than 1 or smaller than -1 and thus decreases for some intermediate values of r. The Huber function is nondecreasing. Minimizing the absolute deviation leads to a ψ-function with a jump at the origin where the values go abruptly from -1 to 1. It does not satisfy $\psi_A(r) \approx r$ from small r and it does indeed suffer from fairly poor efficiency when the random errors follow a normal distribution

Name	*Psi-function*
Huber-estimator	$\psi_H(r) = \min(1, \max(r, -1))$
Tukey-estimator	$\psi_T(r) = r(1 - \min(r^2, 1))^2$
Least absolute deviations	$\psi_A(r) = \mathrm{sign}(r)$

1996). Huber's location estimators T_k in the case of known scale ($S = 1$) are solutions of

$$\sum_{i=1}^{n} \psi_H\left(\frac{y_i - T_k(y_1, \ldots, y_n)}{k}\right) = \sum_{i=1}^{n} \psi_k(y_i - T_k(y_1, \ldots, y_n))/k = 0 \quad [4]$$

where

$$\psi_k(r) = \min(k, \max(-k, r)) = \begin{cases} -k, & \text{if } t \le -k \\ t, & \text{if } -k < t \le k \\ k, & \text{if } t > k \end{cases}$$

This is an m.l.e. for random errors with density $f_{k(t)}$ which is normal-like in the center and has tails that decrease exponentially. These models for random errors were discovered by Huber and have been discussed in his seminal paper (Huber, 1964). Huber showed that f_k is in some mathematical sense the worst among all contaminated densities with a contaminating fraction of ε_k. As a consequence, the corresponding Huber estimator T_k is the estimator that best protects against a contamination of $100\varepsilon_k\%$ of the data. The Huber estimator is therefore the best one in a global sense. Even if a sizable portion of the data contains no useable information, this estimator will perform as well as be possible under these circumstances.

Indicators of robustness such as the breakdown point and the gross error sensitivity generate similar optimality questions as the one solved by Huber (1964). One is interested, for example, in finding an estimator that satisfies the two conditions of consistency and bounded gross error sensitivity and, given these conditions, is most precise. This problem is of a local nature, because we do not consider sizable fractions of outliers, but rather limits for very small fractions. The problem is again solved by the Huber estimators, which shows – not very surprisingly – that a globally best estimator is also locally best. The estimators T_k form a family of least variable estimators with gross error sensitivities that increase as k increases. In other words, they provide a bridge between the median ($k = 0$) and the arithmetic mean ($k = \infty$) and between a minimal value of the gross error sensitivity and an infinite gross error sensitivity. Considering only local properties defined by the influence function has the advantage of being more easily generalized. Thus, it is possible to solve a variety of problems of this type. The book by Hampel et al. (1986) provides a detailed discussion of related results.

Redescending M-Estimators and Highly Resistant Regression Estimators

The form of the Huber estimator comes as a surprise because it never entirely rejects gross errors among the observations. In order to completely reject such outliers, the psi-function $\psi(r)$ would have to be equal to zero for sufficiently large $|r|$. Such an estimator is called redescending, because the psi-function remains roughly linear near the origin, reaches a bound, and subsequently redescends toward zero. Clearly, the Tukey estimator is inspired by this insight. Because ψ_T is not an increasing function, the estimating eqn [1] of the biweight estimator may have multiple solutions. The estimate is thus not well defined. However, among all the solutions of the estimating equations, an appropriate one can usually be picked out by looking at the resulting residuals. Also, when multiple solutions do exist, several of them may be of interest.

Redescending M-estimators are also of interest in the search for highly resistant regression estimators, because they have the potential to identify and eliminate outliers with high leverage. The so-called S-estimators are an implementation of this idea (reference).

Rousseeuw's least-median-of-squares (LMS) estimator is another highly resistant regression estimator. It is defined by the criterion.

$$\text{criterion}(\boldsymbol{T}) = \text{median}\left\{\left(y_1 - \boldsymbol{x}_1^T\boldsymbol{T}\right)^2, \ldots, \left(y_n - \boldsymbol{x}_n^T\boldsymbol{T}\right)^2\right\}$$

which can be seen as a modification of the least-squares criterion, since it replaces the sum by the median. The LMS estimator minimizes this criterion function. The LMS is a useful, often revealing, estimate. It is, however, quite unstable, not with regard to outliers, but with regard to other features of the data.

Finite Samples

A robust data analysis must be valid and efficient beyond a narrow set of circumstances. The classical regression estimator, for example, is chosen to be the best possible for normally distributed observations.

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$$

Robustness demands joint optimality for a broader set of error distributions. One way in which this can be achieved, is by choosing a finite number of challenges, in the simplest case a couple such as {Normal, Student's t with 2 degrees of freedom}, the idea being that the estimator would have to perform well, whether the errors were generated by one or the other of the distributions. For such a set of challenges, the performance of an estimator is determined by a set of efficiencies, one for each challenge. Typically, an estimator that is highly efficient for one of the challenges will have a low efficiency for all the others. Robust estimators on the other hand must have as high an efficiency as possible for all challenges. How to compute estimators that maximize jointly the whole set of efficiencies is discussed in Morgenthaler and Tukey (1991).

Iteratively Reweighted Least-Squares

The computation of the least-squares estimator is straightforward, because the estimating equation is linear. A program for computing the weighted least-squares solution is all that is needed to find the values of M-estimators. To illustrate the idea, we consider the solution of [2]. Rewriting the estimating equation as,

$$\sum_{i=1}^{n}\left[\frac{\psi_k(y_i - T_k(y_1, \ldots, y_n))}{y_i - T_k(y_1, \ldots, y_n)}(y_i - T_k \le (y_1, \ldots, y_n))\right] = 0$$

an algorithm immediately suggests itself. We could start with a reasonable guess T_k^0 for the value of T_k, such as the median of the observations. Next, we compute weights

$$w_i^0 = \frac{\psi_k(y_i - T_k^0)}{y_i - T_k^0}$$

and then solve the estimating equation, $\sum_{i=1}^n w_i^0(y_i - T_k(y_1, \ldots, y_n)) = 0$. From this, we obtain the next guess for T_k, viz.

$$T_k^1 = \sum_{i=1}^n w_i^0 y_i / \sum_{i=1}^n w_i^0$$

With this new value for T_k we are back at the start of the loop, ready to recompute the weights, and then the next value for T_k. Repeating the loop a fixed number of times or until convergence leads to practical values for M-estimates. This idea of iteratively reweighting reveals robust estimators as being weighted least-squares estimators with weights that are not predetermined by the user (as is the case in the weighted least-squares setting), but are automatically picked out by the algorithm. Inspecting the final weights reveals the observations that have been downweighted and allows the user to inspect them and investigate further the potential reasons for the low weight.

Other Estimation Problems

In order to explain the fundamental principles of robust statistics, I have focused mainly on the linear regression model. It is clear, however, that many other procedures much used in the social sciences are in need of robust alternatives.

Robust Inference

The sampling distributions of robust estimators can often be approximated sufficiently well by their asymptotic normal distributions. For the regression M-estimator the following formula gives an estimated variance matrix of the estimated regression coefficients.

$$\frac{n\sum_{i=1}^n (\psi(r_i))^2}{\left(\sum_{i=1}^n \psi'(r_i)\right)^2}\left(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \qquad [5]$$

where $r_1, \ldots, r_n$ are the residuals and $\boldsymbol{x}_i$ is the vector of covariate values for the ith observation. Using the approximate normality of the estimate, this equation allows one to compute robust confidence intervals for individual parameters. The robustness in the context of confidence intervals means that the coverage of the parameter is equal to the nominal coverage rate even under deviations from the model. Furthermore, the interval should not achieve this feat by being overly long.

Distribution-free tests are very resistant in one respect. Their size, that is the probability of falsely rejecting the null hypothesis, is stable. The power, however, depends of course on the distribution of the observations and need not be stable under changes to this distribution. Robust tests must have both, stable sizes and good power under a variety of circumstances. This point illustrates once more the fact that robustness and distribution-freeness are in general two distinct aspects of statistical inference.

Acknowledgment

The writing of this article was supported by a grant from the Swiss National Science Foundation.

See also: Errors in Statistical Assumptions, Effects of; Estimation: Point and Interval; Hypothesis Testing in Statistics; Significance, Tests of; Statistical Distribution Models: Goodness of Fit Tests.

Bibliography

Box, G.E.P., Leonard, T., Wu, C.F. (Eds.), 1983. Scientific Inference, Data Analysis, and Robustness. Academic Press, New York.
Gnanadesikan, R., 1977. Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.
Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.
Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), 1985. Exploring Data Tables, Trends, and Shapes. Wiley, New York.
Huber, P.J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics 35, 73–101.
Huber, P.J., 1981. Robust Statistics. Wiley, New York.
Huber, P.J., 1996. Robust Statistical Procedures In: CBMS–NSF Regional Conference Series in Applied Mathematics, second ed., vol. 68. SIAM, Philadelphia.
Jurečová, J., Sen, P.K., 1996. Robust Statistical Procedures: Asymptotics and Interrelations. Wiley, New York.
Kadane, J.B. (Ed.), 1984. Robustness of Bayesian Analyses. North-Holland/Elsevier, Amsterdam.
Launer, R.L., Wilkinson, G.N. (Eds.), 1979. Robustness in Statistics. Academic Press, New York.
Morgenthaler, S., Ronchetti, E., Stahel, W.A. (Eds.), 1993. New Directions in Statistical Data Analysis and Robustness. Birkhäuser, Basel.
Morgenthaler, S., Tukey, J.W. (Eds.), 1991. Configural Polysampling: A Route to Practical Robustness. Wiley, New York.
Rieder, H., 1994. Robust Asymptotic Statistics. Springer-Verlag, Berlin.
Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.
Sprent, P., 1993. Applied Nonparametric Statistical Methods, second ed. Chapman and Hall, London.
Stahel, W., Weisberg, S. (Eds.), 1991. Directions in Robust Statistics and Diagnostics, Part I and II. Springer-Verlag, Berlin.
Staudte, R., Sheather, S.J., 1990. Robust Estimation and Testing. Wiley, New York.
Wilcox, R., 1997. Introduction to Robust Estimation and Hypothesis Testing. Academic Press, San Diego.

Rogers, Carl Ransom (1902–87)

Howard Kirschenbaum, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Carl R. Rogers (1902–87) was America's most influential psychotherapist and one of its most prominent psychologists. He developed an approach to psychotherapy, known successively as the 'nondirective,' 'client-centered,' and 'person-centered' approach. He pioneered in the recording and publishing of psychotherapy sessions and cases. He sponsored more scientific research on counseling and psychotherapy than had ever been undertaken. He was largely responsible for extending professional counseling and psychotherapy beyond psychiatry and psychoanalysis to many helping professions – psychology, social work, education, pastoral counseling, and others. Rogers was a leader in the humanistic psychology movement, encounter groups, and using effective interpersonal communication to resolve intergroup conflict.

Carl R. Rogers was America's most influential psychotherapist and one of its most prominent psychologists. His therapeutic approach, known most often as 'client-centered therapy' or 'person-centered therapy,' was validated through voluminous scientific research and taught widely, profoundly affecting the fields of counseling and psychotherapy throughout the world and permeating many other helping professions. A leader in the humanistic psychology movement of the later twentieth century, Rogers also extended his theory and practice to include person-centered approaches to education, group leadership, intergroup conflict, cross-cultural communication, and peace-keeping. His writings and personal example stimulated professional, social, and political change in many countries.

Client-Centered Therapy

Rogers' approach to counseling and psychotherapy was based on a core hypothesis about human growth and personality change, which he summarized in 1950:

> Client-centered therapy operates primarily upon one central and basic hypothesis which has undergone relatively little change with the years. This hypothesis is that the client has within himself the capacity, latent if not evident, to understand those aspects of his life and of himself which are causing him pain, and the capacity and the tendency to reorganize himself and his relationship to life in the direction of self-actualization and maturity in such a way as to bring a greater degree of internal comfort. The function of the therapist is to create such a psychological atmosphere as will permit this capacity and this strength to become effective rather than latent or potential (p. 443).

While other therapies might profess similar beliefs, Rogers' method of creating the therapeutic psychological atmosphere was radically different from other approaches commonly employed. Avoiding questions, interpretation, or other directive techniques, Rogers' initial 'nondirective' method relied exclusively on careful listening and skillful 'reflection of feelings,' leading to client insight and positive action. Although he always remained primarily nondirective in his practice, Rogers soon recognized that the counselor's attitudes of acceptance and understanding were more important than his particular techniques. Still later he clarified that it was the relationship in therapy, which the attitudes helped create, that was growth-producing. Rogers called this therapeutic relationship 'client-centered' and described three 'core conditions' in the relationship which bring about positive change in clients.

The first is to accept the client as he is, as a person of inherent worth possessing both positive and negative feeling and impulses. Adopting a term coined by his student Stanley Standal, Rogers called this acceptance and prizing of the person 'unconditional positive regard.' The second is "the therapist's willingness and sensitive ability to understand the client's thoughts, feelings and struggles form the client's point of view…to adopt his frame of reference" (1949: 84). Rogers used the term 'empathy' to describe this condition, popularizing a concept already in use in the field. The third is, be "genuine, or whole, or congruent in the relationship…It is only as [the therapist] is, in that relationship, a unified person, with his experienced feeling, his awareness of his feelings, and his expression of those feelings all congruent or similar, that he is most able to facilitate therapy" (pp. 199–200). He most often used 'congruence' to describe this third ingredient of the therapeutic relationship.

When a counselor is in psychological contact with a client and is able to communicate these attitudes so that the client can perceive them, the 'necessary and sufficient conditions for therapeutic personality change' are present. Rogers argued and demonstrated that the client has within himself the ability and tendency to understand his needs and problems, to gain insight, to reorganize his personality, and to take constructive action. What clients need is not the judgment, interpretation, advice, or direction of experts, but supportive counselors and therapists to help them rediscover and trust their 'inner experiencing' (a concept borrowed from Gendlin), achieve their own insights, and set their own direction.

Until Rogers, the fields of counseling and psychotherapy had been dominated by the guidance movement and psychoanalysis, respectively. While Rogers was not the first to argue for a 'newer approach' to counseling and psychotherapy – an approach that relied less on the professional's expertise and direction and more on the client's direction and resources – Rogers' system was arguably the most clearly described,

comprehensive, and extreme, and the newer direction in mental health work became associated with him.

Recording Therapy

One reason Rogers was able to demonstrate his propositions in a cogent and convincing manner was that he was the first person in history to record and publish complete cases of psychotherapy. Before the invention of tape recorders this was a remarkable achievement, requiring a microphone in the counseling room connected to alternating phonograph machines in an adjoining room, which cut grooves in blank record disks that had to be changed every 3 minutes. With his student Bernard Covner, Rogers and his team recorded thousands of disks, involving scores of clients. These recordings were pivotal in the clinical training of psychotherapists which, in the 1940s, Rogers may have been the first to offer in an American university setting.

In addition to the audio recording of therapy sessions, Rogers was among the first to make cinematic recordings of counseling and psychotherapy. The American Academy of Psychotherapists became a leading distributor of training tapes and movies, with Rogers the most frequent therapist portrayed. A still widely distributed set of training films showed Rogers, Gestalt therapist Frederick (Fritz) Perls, and rationale-emotive therapist Albert Ellis each demonstrating his method with the same client, 'Gloria.'

Scientific Research and Theory

The recording of actual therapy sessions also provided the data on which Rogers and his colleagues carried out more scientific research on one therapeutic approach than had ever been undertaken before. Rogers and his students and colleagues devised and used numerous instruments for measuring the variables of client-centered therapy and its outcomes: therapist acceptance, empathic understanding, and congruence; client expression of feelings, insight, self-concept, self-acceptance, self-ideal, and level of experiencing; and clients' positive actions, social adjustment, and numerous other variables. In 1959 the American Psychological Association awarded Rogers its first Distinguished Scientific Contribution Award for "developing an original method to objectify the description and analysis of the psychotherapeutic process, for formulating a testable theory of psychotherapy and its effects on personality and behavior, and for extensive systematic research to exhibit the value of the method and explore and test the implications of the theory" (p. 128).

As the scientific award citation suggests, Rogers was interested in psychological theory and in the effects of therapy on personality as well as behavior. Building upon the Gestalt and phenomenological movements in psychology, and the work of his students Victor Raimy and Arthur Combs, he developed a 'self-theory' of personality which is still included in many textbooks (Rogers, 1959). The theory describes how an individual's concept of self emerges, how the process of socialization causes individuals to distrust their feelings and sense of self, how experiences that are inconsistent with the concept of self become denied and distorted, causing personal distress and psychological problems, and how the therapeutic relationship can help the individual restructure the sense of self, allowing previously denied and distorted experience into awareness, leading to reduction in stress and openness to new experiencing.

Initial Impact on the Professions

More than any individual Rogers was responsible for the spread of professional counseling and psychotherapy beyond psychiatry and psychoanalysis to other helping professions. While he was not the first to use the term 'client' for the recipient of therapy, he popularized it, emphasizing that a person seeking help should not be treated as a dependent patient but as a responsible client, and that those in psychological distress were not necessarily sick requiring treatment by medical specialists. Rather all people could be helped by the growth-producing conditions of empathic understanding, unconditional positive regard, and congruence, and professionals from many walks of life could be trained to provide these conditions. Thus counselors, psychologists, social workers, clergy, medical workers, youth and family workers, and other helping professionals could employ counseling methods and client-centered attitudes in their work. Rogers' books *Counseling and Psychotherapy* (1942) and *Client-Centered Therapy* (1951) described the principles of effective therapy and presented ample case studies from recorded sessions to illustrate his points.

Rogers also impacted the helping professions through leadership in many professional associations. Earlier in his career he served on the executive committee of American Association of Social Workers and was vice-president of American Association of Orthopsychiatry (social workers). As he became more prominent in the psychology profession, he served as president of American Association of Applied Psychology, American Psychological Association, and American Academy of Psychotherapists, among other distinguished positions.

Applications to Other Fields

In his own counseling practice, teaching, administration, and writing, Rogers continued to apply his theory and method to other fields – education, parenting, group leadership, and the health professions, to name a few. In each instance he demonstrated how the facilitative conditions of empathy, positive regard, and congruence could unleash growth, creativity, learning, and healing in children, students, group members, patients, and others. Applied to education, his views on 'student-centered learning' coincided with and contributed to the 'open education' movement in the United States, Great Britain, and elsewhere. His book *Freedom to Learn* (1969) went through two new editions over the next 25 years.

He expanded his theory and practice into the areas of group leadership and group facilitation over several decades. In the 1940s he, Thomas Gordon, and colleagues at the University of Chicago had experimented with 'group-centered leadership,' whereby the leader's acceptance, understanding, genuineness,

and willingness to let the group set its own directions stimulated great energy, creativity, and productivity among group members. In the late 1950s and 1960s, Gordon, Richard Farson, Rogers, and associates extended this approach to what Rogers called the 'basic encounter group' – an unstructured group experience in which 'normal' group members came to greater self-understanding, spontaneity, improved communication, and genuineness in relationships. Rogers led scores of encounter groups in professional, business, religious, medical, academic, personal growth, and other settings. *Look* magazine called Rogers an 'elder statesmen of encounter groups.' *Carl Rogers on Encounter Groups* (1970) was a major seller, and his and Farson's film of an encounter group they facilitated, 'Journey Into Self,' won an Academy Award (an 'Oscar') for the best feature-length documentary in 1968.

He applied his approach to couples' relationships in *Becoming Partners: Marriage and Its Alternatives* (1972). Through case studies of couples in varying types of relationships – from traditional to openly experimental – he demonstrated how open communication, genuineness, mutual support for the partner's growth, and acceptance of change were ingredients for healthy relationships. Similarly, in earlier articles on 'The Nature of Man' and 'A Therapist's View of the Good Life: The Fully Functioning Person,' and later essays like those collected in *A Way of Being* (1980), he posited a philosophical–psychological model of the 'person of tomorrow' or 'the emerging person' – an individual who is open to the outer world but also deeply connected to one's own inner experience, who trusts one's own organismic valuing process, who is authentic, who is skeptical of authority, science, and technology, who values wholeness, and who understands that life is an ongoing process of change, among other qualities.

Political Implications

Beyond the psychological and philosophical realms, Rogers eventually recognized the political implications of his theories and methods – not in the sense of partisan politics, but in how power and influence are distributed in all human relationships. He explored these in *Carl Rogers on Personal Power* (1977). He and others had always recognized that, in the counseling or therapy relationships, the client-centered approach reversed the power relationship implicit in the traditional medical model. The expert who prescribed treatment was now the facilitator of growth who followed the client's lead. Gradually it became apparent that Rogers was arguing for a similar shift in power relations in other professional and human relationships. Recognizing the ever-widening applicability of the client-centered, student-centered, and group-centered approach, Rogers increasingly used a broader term, 'person-centered,' to describe his work.

The shift was not just semantic; it took the work in new directions. In the 1970s and 1980s he experimented with applications of the person-centered approach to enhancing cross-cultural communication and resolving intergroup and international conflict. Through workshops and filmed encounter groups involving different races, generations, and political persuasions – including Catholic and Irish protagonists from Northern Ireland – Rogers demonstrated how the same growth-promoting conditions useful in all helping relationships can lead to improved communication and greater understanding among previously antagonistic parties. He and his colleagues led person-centered workshops for groups from 100 to 800 participants in Brazil, Mexico, Spain, Hungary, Soviet Union, and many other countries. In South Africa, for example, Rogers facilitated demonstration encounter groups before large audiences in which, for the first time in their lives, white and black South Africans could actually communicate with one another in an atmosphere of equality and respect.

A 4-day gathering of international leaders in Austria on resolving tensions in Central America was a laboratory demonstration of the potential of the person-centered approach for resolving real international conflict. At the invitation of the Carl Rogers Peace Project, over 40 participants – including three former presidents of Central American countries, a current vice-president, seven ambassadors, current and former US senators and representatives, representatives of the Contra and Sandanista warring factions, and others – sat together as 83-year-old Rogers and his colleagues facilitated an unprecedented level of open communication among the participants. Former Costa Rican President Rodrigo Carazo later credited Rogers and his group for breaking the stalemate and allowing the peace process in Central America to move forward and eventually succeed. Other testimonials suggested that Rogers' efforts in professional development and citizen diplomacy helped influence future developments in several countries. In acknowledgment of his work, Rogers was nominated posthumously for the Nobel Prize for Peace. At Rogers' memorial service, Richard Farson eulogized him as 'a quiet revolutionary.'

Humanistic Psychology

Rogers was a leader in what Abraham Maslow described as a 'third force' in psychology, which emerged as a visible and influential movement in the latter half of the twentieth century. 'Humanistic psychology,' as it came to be known, differed from psychoanalysis and behaviorism in at least three ways. First, this psychology gave much more emphasis and credence to the individual's 'phenomenal field' – e.g., the client-centered therapists trying to understand the client's frame of reference rather than evaluating or diagnosing from the outside, or the existential psychotherapists helping the patient find meaning in his life. Second, this psychology focused not just on the remediation of psychological problems, but on psychological health, creativity, self-actualization, or what Rogers described as 'the fully functioning person.' It was concerned not only with helping people adjust, but helping them experience their positive human potential to the fullest. Third, it was a psychology interested in those factors that distinguish human beings from other species. Choice, will, freedom, values, feelings, goals, and other humanistic concerns were all central subjects of study.

As Rogers' career and that of leading behavioral psychologist B.F. Skinner paralleled one another in their timing, productivity, and influence, their views inevitably were contrasted. Meeting on several occasions, including a 6-hour debate-dialogue in 1962, one of their exchanges on 'Some

Issues Concerning the Control of Human Behavior' (1956) became one of the most frequently reprinted articles in the behavioral sciences. Rogers became a leading spokesperson for the humanistic psychology movement. In that unofficial capacity, he was often asked to engage in dialogues with some of the leading theologians, philosophers, and intellectuals of the later twentieth century, including Martin Buber, Paul Tillich, Reinhold Niebuhr, Michael Polanyi, Gregory Bateson, Rollo May, Skinner, R.D. Laing, and others (Kirschenbaum and Henderson, 1989a). Humanistic psychology continues to exert a significant influence on society and the helping professions, including serving as precursor and foundation of the contemporary positive psychology movement.

Enduring Influence

While Rogers was still alive, he presented a vivid role model of the person-centered approach for almost six decades, demonstrating his theories through teaching, lecturing, live demonstrations and workshops, and audio–visual recordings. By virtually all accounts, he embodied his theories by being an exceptional listener and communicator and a decent, honorable person. Near the end of his career, 1982 surveys in the *Journal of Counseling Psychology* and *American Psychologist* still ranked Carl Rogers as the most influential author and counselor/psychotherapist.

Since his death in 1987, Rogers' influence has remained strong. In the United States, although cognitive–behavioral therapy has supplanted the person-centered approach as the leading therapeutic approach, Rogers still remains an influential figure. A survey of 2600 therapists in the *Psychotherapy Networker* in 2006 easily ranked Rogers as the respondents' most often cited influence.

Part of this is due to the continuing impact of his writings. Rogers was one of the helping professions' most prolific writers, authoring 17 books and more than 200 professional articles and research studies. Millions of copies of his books have been printed, including over 80 foreign language editions of his works. His most popular book *On Becoming a Person* (1961), written for both a professional and general audience, *The Carl Rogers Reader* (Kirschenbaum and Henderson, 1989b), and other works still in print continue to spread his ideas.

Around the world, person-centered psychotherapy continues to be a leading approach in many countries. One explanation may be that, of all the major counseling theories, the person-centered approach arguably is unique in having clear political implications – reversing traditional, hierarchal professional roles with more egalitarian relationships between helpers and clients, teachers and students, leaders and group members, and empowering counselors and therapists to work for conflict resolution and social change in their own communities and wider world. This has appealed to the segment of the helping professions, particularly in Europe, that is concerned about the wider social context in which counseling and therapy are practiced.

After Rogers' death some of the greatest new interest in his work has been in Eastern Europe, Russia, Latin America, and other emerging democracies. As a Japanese counselor explained in the 1960s, Rogers helped "teach me … to be democratic and not authoritative." The totality of his life's work was devoted to demonstrating how supportive, growth-producing conditions can unleash healing, responsible self-direction, and creativity in individuals and groups in all walks of life. As many countries strive to resolve old, intergroup tensions and put self-government and self-determination into practice, they have recognized in Rogers' work not only particular methods of use to counselors, teachers, helping professionals, and group leaders, but a positive, person-centered, democratic philosophy consistent with their national aspirations.

Meanwhile, in North America and elsewhere, Rogers' influence endures not only in the traditional client-centered approach to counseling and psychotherapy that Rogers promulgated and practiced, but in a number of new paradigms. For many practitioners and scholars around the world, the person-centered approach as described by Carl Rogers remains the only person-centered approach. In counseling and psychotherapy, this approach may be characterized by:

- A belief in the client's 'self-actualizing tendency'; that is, an innate motivation to grow and mature and realize his or her self-interest, especially when provided with a supportive environment.
- A reliance on the therapeutic relationship, characterized by the core conditions of congruence, empathy, and unconditional positive regard, for therapeutic progress.
- A continuing focus on the client's inner experience, hence ….
- An absence of directive techniques or perspectives introduced by the therapist, such as questions, interpretation, advice, coaching, and the like (except for relatively rare expressions of counselor congruence).
- An avoidance of diagnosis, treatment plans, and other therapist-centered methods that reflect the medical model of mental illness.
- A view of the client as a whole person in the process of 'becoming'; that is, becoming a more fully functioning person; therefore, counseling focuses not simply on a presenting problem but on more holistic change, so the client can continue to grow and exercise self-direction beyond the therapeutic relationship.

One might say these characteristics describe what has been called 'traditional,' 'classic,' or 'orthodox' client-centered/person-centered counseling and psychotherapy. Some simply prefer the term 'client-centered.' However, for many other practitioners and scholars, there is more than a single person-centered approach. They believe it is possible to introduce certain techniques to further a client's self-exploration while still remaining essentially person centered. Thus, Natalie Rogers, Carl Rogers' daughter, developed 'person-centered expressive therapy' which engages clients and group members in the creative arts and employs empathic listening to help them explore the meaning of their creative expression and its implication for their lives. Eugene Gendlin developed 'focusing' and 'focusing-oriented, experiential psychotherapy' as a method for helping clients tune into their 'inner experiencing' in greater depth, while continuing to use empathic listening as a primary method for furthering that exploration. Leslie Greenberg, Robert Elliott, and others developed 'process-experiential,' then 'emotion-focused therapy' which combined

Gestalt and other techniques to guide the therapeutic process. All three of these approaches – and others – have identified themselves as person centered and have significant followings around the world. Their proponents would argue that they remain essentially person centered, with a major emphasis on the core conditions, including a predominant use of empathic listening.

Thus, there is some controversy in the person-centered world today as to whether 'person centered' should be used exclusively to describe Rogers' traditional or classic approach to counseling and psychotherapy, or whether there is room for 'many tribes' in the person-centered nation. In some countries, such as the United Kingdom, classic client-centered counseling has been the predominant mode; in others, like Germany, other suborientations of the person-centered movement have commanded greater allegiance. Whatever the orientation, the person-centered approach is institutionalized in a number of European countries, with universities and training institutes turning out person-centered practitioners, robust research programs, government recognition for licensure, and reimbursement by insurance companies, although political turf battles around recognition and reimbursement are ongoing. There are scores of person-centered, professional organizations around the world, numbering in the hundreds to the thousands. One 2005 study showed that in the 15 years after Rogers' death, there were more publications on Rogers and the person-centered approach than in the 40 years before. This trend has continued.

Criticism

In 1972, Rogers was awarded the American Psychological Association's Distinguished Professional Contribution Award, becoming the first psychologist to receive both the organization's highest scientific and highest professional honor. He was recognized for "devising, practicing, evaluating and teaching a method of psychotherapy and counseling which reaches to the very roots of human potentiality and individuality … [H]e has caused all psychotherapists to reexamine their procedures in a new light." The committee also noted how Rogers' "commitment to the whole person has been an example which has guided the practice of psychology in schools, industry, and throughout the community. Innovator in personality research, pioneer in the encounter movement, and respected gadfly of organized psychology, he has made a lasting impression on the profession of psychology" (1973: 71).

Not everyone agrees that Rogers' lasting impression is a positive one. As critic Christopher Lasch once wrote, "As a founding father of humanistic psychology, the human potential movement and the encounter group, Carl Rogers has a lot to answer for." Critics of Rogers' work have argued that client-centered therapy is superficial, unworkable with some populations, and unmindful of multicultural and feminist issues, the social context, and recent advances in behavioral, drug, and alternative therapies; that Rogers' views on human nature are unrealistically optimistic and underestimate human evil; that encounter groups and humanistic psychology have fostered widespread selfishness, narcissism, and moral permissiveness; and that Rogers' experiments with organizational change were naive and counterproductive. Such criticisms have sometimes been fair; for Rogers, like any individual, was a product of his times, with personal and historical limitations. Just as often, criticisms of Rogers and his work have been wanting, because the critic was unfamiliar with the full scope of Rogers' theories, research, and ever-widening practice, or biased, because the critic's own politics were antagonistic to that of the explicitly egalitarian, progressive politics of the person-centered approach.

Critics notwithstanding, Rogers more than anyone helped spread professional counseling and psychotherapy beyond psychiatry and psychoanalysis to psychology and other helping professions. Interestingly, as meta-analyses of psychotherapy research continue to emerge, the data increasingly suggest that the success of counseling and therapy is not due to any particular method, whether cognitive behavioral, psychodynamic, client-centered, or other. Rather the research demonstrates that there are a number of 'common factors' in the context of the therapy relationship which account for successful outcomes. Many of these common factors point back to the therapist's support, empathic understanding, and ability to form a therapeutic alliance with the client. Ironically, Rogers' core conditions for therapeutic change, decades later, are being validated by the latest generation of scientific research. While this research suggests that positive regard, empathy, and congruence may not be absolutely necessary in every case nor sufficient for all counseling relationships, what the research does affirm are: first, Rogers' initial insights about the importance of the therapeutic relationship; second, the usefulness and practicality of the core conditions for forming the essential therapeutic alliance; and third, the efficacy of empathy and positive regard/acceptance, and, possibly, congruence for achieving positive counseling outcomes. Consequently, in recent years virtually all therapeutic approaches have come to acknowledge the importance of the therapeutic relationship and more intentionally incorporate it into their practice.

Education and Influences

Rogers attended the University of Wisconsin in Madison, first majoring in agriculture and then switching to history after experiencing the call to religious work. As one of 10 U.S. delegates to the 1922 World Student Christian Federation Conference in Peking (Beijing), he subsequently participated in a 3-month lecture tour to youth groups around China and Japan, all of which helped to liberalize his religious and social views. Back home, motivated by the 'social gospel' as much as religious conviction, he eschewed his parents' dogmatism and attended liberal Union Theological Seminary in New York City, which allowed him to take courses at adjoining Teachers College, Columbia University. There religious doubt combined with fascination with psychology. Influenced by instructors Leta Hollingworth, Goodwin Watson, William Heard Kilpatrick, the leading proponent of John Dewey's progressive education philosophy, and others, Rogers transferred to Teachers College to pursue a doctorate in clinical psychology. His exposure there to the testing and measurement movement of E.L. Thorndike was balanced by his experiences in clinical work and a fellowship at the Institute for

Child Guidance, where he encountered Freudian thought, a lecture by Alfred Adler, Rorschach testing, and other psychoanalytic and psychiatric approaches.

Seeking to integrate these influences, Rogers came to appreciate the importance of attending to and honoring clients' inner experience while also striving to objectively assess any changes that might result from treatment. His doctoral dissertation, *Measuring Personality Adjustment in Children Nine to Thirteen Years of Age* (1931), combined both subjective and objective measures. Rogers' 'Personality Adjustment Inventory' subsequently was published by YMCA Press and used for over 50 years. His next major learning experience was the 12 years of working with thousands of troubled children and adults in Rochester, New York, where he gradually developed his nondirective approach to counseling and psychotherapy. During this period, he was influenced by students of Otto Rank, especially Jessie Taft whose 'relationship therapy' shifted emphasis from past content to a focus on the patient's self-insight and self-acceptance within the therapeutic relationship. Throughout his life, Rogers frequently cited his students and younger colleagues as being significant influences on his work – challenging his assumptions, introducing him to new thinking and practice, and helping him remain on the cutting edge.

Positions Held

After leaving Columbia, Rogers' first position was at Rochester (NY) Society for the Prevention of Cruelty to Children (1928–38) where he was director of the Child Study Department, then director of the Rochester Guidance Center (1939–40). Based on his practical experience and first book, *Clinical Treatment of the Problem Child* (1939), he was offered a full-professorship at Ohio State University, where he taught from 1940 to 1944. After an interim year of training United Service Organization workers in counseling techniques with returned servicemen, he moved to University of Chicago, where he developed and ran the internationally renowned Counseling Center and taught in the Psychology Department from 1945 to 1957. He served a joint appointment in the Departments of Psychology and Psychiatry at the University of Wisconsin from 1957 to 1963. Somewhat burned out with academia and attracted by the warm weather and freedom of the 1960s human potential movement in California, he moved to the Western Behavioral Sciences Institute (WBSI) in La Jolla, California in 1963, where he was able to pursue his own interests and spend more time traveling for professional engagements around the world. In 1972, after a painful schism at WBSI he and other colleagues split off to form their own organization, Center for Studies of the Person, where Rogers remained until his death in 1987.

Personal Life

Rogers grew up in suburban Oak Park, then rural Glen Ellen, Illinois, with three brothers, a sister, and conservative, Protestant parents. Spending teenage years on the family farm, he developed a love of nature and serious working knowledge of scientific method, conducting agricultural experiments on a plot he and his brothers were responsible for. Upon college graduation he married childhood friend Helen Elliott (1901–79), an aspiring commercial artist, commencing a 55-year union. Their two children, David (1926–94) and Natalie (b. 1928) distinguished themselves professionally, David in medicine (e.g., Dean of Medicine at Johns Hopkins and president of the Robert Wood Johnson Foundation) and Natalie as an author and pioneer of 'client-centered expressive therapy' in the field of art therapy. For many years Rogers spent the academic winter quarter in the Caribbean and Mexico, where he read widely and developed a number of his theories, Helen painted, and both snorkeled. They regularly traveled to Rogers' professional engagements around the world and to be with their children and six grandchildren. Carl pursued lifelong hobbies of photography, making mobiles, and gardening. When Helen became ill in her seventies, Carl cared for her for several years until her death in 1979. Thereafter he remained involved in work, cultivated old and new friendships with both men and women, and traveled widely. He was active until his death at age 85 from a heart attack stemming from hip surgery following a fall.

While Rogers was by and large an excellent exemplar of his own theories about growth, change, and the fully functioning person, and while he led a rich and rewarding life in many respects, he was not without personal problems. He and Helen struggled with relationship issues in their later years, periodically causing unhappiness for both. Privately he suffered from alcohol dependency for several decades, although remarkably it appeared to have little or no effect on his work or relationships.

After Rogers left the church in his early twenties, he showed little or no interest in religion for most of his life. Yet in his seventies and eighties, surrounded by younger colleagues and family involved in various spiritual movements and practices in California, Rogers and Helen experimented with spiritualism and the paranormal, and he thought and wrote seriously on the nature of reality. Although he never had a significant religious or mystical experience himself in his later years, his spiritual journey did lead him to extend his concept of an actualizing tendency in individuals to a 'formative tendency in the universe,' his last theoretical formulation.

See also: Allport, Gordon W (1897–1967); Counseling Psychology; Dewey, John (1859–1952); Diagnostic Processes in Clinical Psychology and Psychotherapy; Freud, Sigmund (1856–1939); Group Psychotherapy, Clinical Psychology of; Learning Theories and Educational Paradigms; Lewin, Kurt (1890–1947); Motivation, Learning, and Instruction; Skinner, Burrhus Frederick (1904–90); Therapist–Patient Relationship.

Bibliography

American Psychological Association, March 1957. Distinguished scientific contribution awards for 1956 – Carl R. Rogers. American Psychologist 12, 125–133.

American Psychological Association, 1973. Distinguished professional contribution award for 1972. American Psychologist 28 (1), 71.

Kirschenbaum, H., 2007, 2009. The Life and Work of Carl Rogers. PCCS Books, Ross-on-Wye, England. American Counseling Association, Alexandria, VA.

Kirschenbaum, H., Henderson, V. (Eds.), 1989a. Carl Rogers: Dialogues. Houghton Mifflin, Boston.

Kirschenbaum, H., Henderson, V. (Eds.), 1989b. Carl Rogers Reader. Houghton Mifflin, Boston.

Rogers, C.R., 1939. Clinical Treatment of the Problem Child. Houghton Mifflin, Boston.
Rogers, C.R., 1942. Counseling and Psychotherapy: Newer Concepts in Practice. Houghton Mifflin, Boston.
Rogers, C.R., 1949. The attitude and orientation of the counselor in client-centered therapy. Journal of Consulting Psychology 13, 82–94.
Rogers, C.R., 1950. A current formulation of client-centered therapy. Social Service Review 24, 442–450.
Rogers, C.R., 1951. Client-centered Therapy. Houghton Mifflin, Boston.
Rogers, C.R., 1956. Client-centered therapy: a current view. In: Fromm-Reichmann, F., Moreno, J.L. (Eds.), Progress in Psychotherapy. Grune and Stratton, New York, pp. 199–209.
Rogers, C.R., 1959. A theory of therapy, personality, and interpersonal relationships as developed in the client-centered framework. In: Koch, S. (Ed.), Psychology: A Study of a Science, vol. III. McGraw-Hill, New York, pp. 184–256.
Rogers, C.R., 1961. On Becoming a Person. Houghton Mifflin, Boston.
Rogers, C.R., 1969. Freedom to Learn: A View of What Education Might Become. Charles Merrill, Columbus, OH.
Rogers, C.R., 1970. Carl Rogers on Encounter Groups. Harper and Row, New York.
Rogers, C.R., 1972. Marriage and Its Alternatives. Delacorte Press, New York.
Rogers, C.R., 1977. Carl Rogers on Personal Power. Delacorte Press, New York.
Rogers, C.R., 1980. A Way of Being. Houghton Mifflin, Boston.
Rogers, C.R., Skinner, B.F., 1956. Some issues concerning the control of human behavior. Science 134 (3231), 1057–1066.

Rokkan, Stein (1921–79)

Peter Flora, University of Mannheim, Mannheim, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13390–13394, © 2001, Elsevier Ltd.

Abstract

Stein Rokkan (1921–79), Norwegian, was one of the world's leading social scientists after World War II. He played a unique role in organizing international social science cooperation in the 1960s and 1970s, and was the driving force in the development of infrastructure services for comparative research. But he was also a prolific writer, and it was the combination of all his contributions to comparative research, international cooperation, and infrastructure development that made him the comparativist of his time.

Stein Rokkan was one of the world's leading social scientists after World War II. He played a unique role in organizing international social science cooperation in the 1960s and 1970s, and was the driving force in the development of infrastructure services for comparative research. But he was also a prolific writer, and it was the combination of all his contributions to comparative research, international cooperation, and infrastructure development that made him the comparativist of his time.

Stein Rokkan, the Norwegian

Rokkan was born 4 July 1921 in Vågan on the Lofoten Islands, on the northern periphery of Norway, and grew up in Narvik. Between 1938 and 1948 he attended the University of Oslo, where he first studied philology and then political philosophy, in which he received his doctorate with a dissertation on David Hume. From 1948 to 1951 Rokkan continued his studies in the USA and England: at Columbia University, where his work with Paul Lazersfeld acquainted him with the methods of modern social research, and at the London School of Economics, where he met T. H. Marshall, whose historical-sociological research profoundly influenced his own.

Rokkan's educational career already contained the seeds of a rare combination of highly diverse talents: above all his great openness towards the 'new' social sciences while maintaining a firm grounding in the 'old' cultural sciences. His extensive linguistic, philosophical, and historical studies gave him a keen eye for the significance of cultural differences in international comparison, and thus for the often misleading simplicity of purely quantitative approaches. At the same time he recognized the potential of such approaches in comparative research and became one of their most important supporters. Rokkan's thought was interdisciplinary from early on. His scholarly interests included the entire spectrum of the social sciences, and this characterized his core scientific work: a political sociology in the tradition of classic macrosociology, the historical-comparative analysis of whole societies and polities.

From the very beginning, Rokkan combined a genuine international orientation with a deep devotion to his homeland that shaped his comparative approach. From his intimate knowledge of Norway he developed concepts (e.g., the center–periphery dichotomy) and questions (e.g., on the interaction of mass movements and institutional change) that he then applied to comparative research. This made him aware of the complexity of each single 'case' and led him to assess realistically the possibilities and limits of international comparisons.

Rokkan developed his system of thought, which was rooted in his origin and education, systematically throughout his entire scholarly career until his early death on 22 July 1979 in Bergen. From 1951 to 1958 he worked at the Institute of Social Research in Oslo, where he soon became engaged in various comparative research projects. In the same period he initiated, together with Henry Valen, the Norwegian Program of Electoral Research which was oriented on the work of the Survey Research Center in Ann Arbor, Michigan, but already showed Rokkan's signature in its emphasis on context variables.

In 1958 Rokkan became research professor at the Christian Michelsen Institute in Bergen, where in 1966 he was awarded a university chair and founded a new department of comparative politics. During this time he was also engaged in efforts to increase cooperation between the Scandinavian countries, which resulted in the publication of *Scandinavian Political Studies* since 1966, and in the founding of the Nordic Political Science Association in 1975.

Rokkan and the Internationalization of the Social Sciences

Rokkan was active at the international level from the very beginning of his professional career. Developing international research programs, organizing international projects and conferences, and building institutions and infrastructures for comparative research became the focus of his activities from the early 1960s. Three contexts of these activities can be distinguished: UNESCO and the International Social Science Council, the International Committee on Political Sociology, and the European Consortium for Political Research.

Rokkan, then Secretary-General of the International Sociological Association, participated in the founding of the International Social Science Council (ISSC) by UNESCO in 1952. In 1961, he proposed to the ISSC a broad program in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61114-5

cross-national and cross-cultural research which he then directed for over 15 years. He became a member of the Executive Committee of the ISSC in 1965 and served as president of the ISSC from 1973 to 1977.

The 1961 program led to a broad spectrum of activities to develop infrastructure services for comparative research and to promote concerted interdisciplinary research. Whole series of conferences were organized on comparative survey research and aggregate statistical comparisons of nations in the first half of the 1960s. Some of the further work was channeled into the European Coordination Centre for Research and Documentation in Social Sciences, set up in Vienna in 1963; other activities were headed up by two ISSC standing committees for data archives and comparative research established in 1966–67. Rokkan remained the driving force behind most of the subsequent activities, which in the 1970s also included summer schools and workbooks for training on comparative analysis and the promotion of an International Federation of Data Organization.

In the late 1960s, with the support of UNESCO Rokkan launched another series of conferences on national developments, asking what conclusions could be drawn from the early European experiences of state formation and nation-building for developmental processes outside Europe. Rokkan here developed his basic ideas about a macromodel of European political development. This work was also linked to the International Committee on Political Sociology that Rokkan, together with Shmuel Eisenstadt, Morris Janowitz, and Martin Lipset, had founded in 1960 and for which he served as Secretary from 1960 to 1970. Originally affiliated only with the International Sociological Association, in which Rokkan was vice-president from 1966 to 1970, it became a joint ISA/IPSA committee in 1970 when Rokkan was elected president of the International Political Science Association. During the 1960s the committee was an exclusive group of international scholars with a common interest in studying major historical changes. It was in this second context that Rokkan developed his ideas on democratization, cleavage structures, and party systems.

The third context was the European Consortium for Political Research. Rokkan was one of its founders and its first Chair from 1970 to 1976. Here too he pursued his central goals of developing joint research programs and improving infrastructure in the broadest sense: from bibliographic tools, research documentation, and data archives to computer applications (including cartography), and the training of the next generation of social scientists.

Rokkan's decisive role in the internationalization of the social sciences and the promotion of comparative research resulted from a congruence of personality and historical situation. In the 1960s the social sciences were expanding rapidly in Western Europe, but they suffered from national fragmentation. International cooperation was needed, and succeeded first via the USA. Only later did a genuine 'Europeanization' begin. Rokkan's broad education and international orientation proved invaluable in building bridges between the USA and Europe and reconciling different European traditions. His enormous talent for combining personal networking and institution-building was a crucial asset in this start-up phase and the close linkage he saw between programmatic conceptualization, cooperative organization, and infrastructure-building gave this endeavor a cohesiveness that seems to be lacking today.

Rokkan's historical role has somewhat overshadowed his scientific contributions, apart from a few famous single hypotheses and typologies. His lasting contribution will be his attempt to develop a macro-model of the political development of Europe. In particular, his comparative studies on the structuring of European mass democracies, and his analyses of the processes of state formation and nation-building in Europe, but also his general approach to comparative research.

The Structuring of Mass Politics in Western Europe

Rokkan's work on the democratization of the European nation-states concentrated on the institutional and organizational variations in the structuring of mass politics up to World War II. Here he developed his well-known ideas of four institutional thresholds of democratization and of historical cleavage structures which in the process of democratization are transformed into party systems via the formation of alliances and oppositions.

In institutional terms, democratization for Rokkan meant the successive lowering or removal of four thresholds: the threshold to political opposition (legitimation), to political participation (incorporation), to access to parliament (representation), and to participation in government (executive power).

The lowering of the first two thresholds, through the recognition of civic rights and the extension of suffrage, ushered in the age of mass politics with mass parties and mass election campaigns. For Rokkan, an unbroken tradition of representation, a history of continuous center building, and a Protestant nationalization of territorial culture were the main factors favoring an earlier, more gradual, and stable democratization.

The lowering of the third threshold implied a transition from majoritarian to proportional representation, easing the access of minorities to representation in the legislature. In his view, the pressures for proportional representation were higher in the ethnically and/or religiously heterogeneous countries, and in general in the smaller nations with their greater international dependence.

The introduction of universal (male) suffrage in all Western European countries at the end of World War I and the transition to proportional representation in most of them tended to stabilize the party systems of the 1920s for a long time, in most cases until the 1960s (Rokkan's famous 'freezing' thesis). In Rokkan's view, the party systems of the 1920s reflected older cleavages that had emerged in the earlier processes of state formation and nation-building and became politicized and organized in the process of democratization.

Cleavages are understood as strong and enduring conflicts that tend to polarize a political system. Rokkan distinguished between four or five basic cleavages that emerged at different critical junctures in European history. The *National Revolution* (1789 and after) produced two cleavages, with even older

roots in the early state formation processes and the Reformation:

1. the *center–periphery cleavage:* conflicts between the central nation-building culture and ethnically linguistically distinct subject populations in the peripheries;
2. the *state–church cleavage:* conflicts between the aspirations of the centralizing, standardizing, and mobilizing nation-state and the corporate claims of the church.

Two further cleavages were produced by the *Industrial Revolution* in the nineteenth century:

3. the *land–industry* cleavage: conflicts between the landed interests and the rising class of commercial/industrial entrepreneurs; and
4. the *worker–employer* cleavage: conflicts between owners/employers and tenants/workers.

The class cleavage was overlaid by a fifth cleavage between proletarian internationalism and nation-accepting socialism as a result of the *International Revolution* (1917 and after).

According to Rokkan, the differences between the European party systems largely were produced by the first three cleavages, reflected in the presence or absence of regional, Christian, or agrarian parties. Whereas Christian parties emerged mainly in the religiously mixed and some of the Catholic countries, agrarian parties are found only in Protestant areas. The class cleavage, by contrast, produced working-class parties in every country, but their strength and unity differed greatly due to the socialist–communist split.

State Formation and Nation-Building in Europe

In analyzing the historical development of the different cleavage structures, Rokkan took a decisive step beyond traditional comparative research: he no longer regarded the units of comparison, the nation-states, as given, but rather thematized their development itself. In these later works, Rokkan was influenced greatly by Hirschman's concepts of exit, voice, and loyalty, which he used in his analyses of the development of the nation-state in terms of boundary-building, internal structuring, and system-building.

Rokkan distinguished between four basic processes of loyalty-producing system-building: state formation and nation-building; two processes mainly of boundary-building and two processes of mainly restructuring, the establishing of political citizenship by equalizing the rights to participation and the establishing of social citizenship by redistributing resources and benefits. In the older countries these processes tended to follow each other in time, whereas in the newer countries the problems associated with them often cumulated.

In the European context, according to Rokkan, these processes were influenced by six givens: first, the heritage of the Roman Empire (law and citizenship); second, the tradition of the Germanic kingdoms (legislative/judicial assemblies); third, the crucial role of the supraterritorial Catholic Church; fourth, the growth of a central city belt; fifth, the development of feudal and manorial agrarian structures; and sixth, the emergence of literatures in vernacular languages.

The European world of states and nations emerged from the disintegration of the Western Roman Empire, after centuries of territorial fragmentation and retrenchment, followed by processes of expansion and reorganization. For Rokkan, 'the alphabet and the city decided the fate of Europe' (Rokkan, 1999) in these processes. As a result of various waves of migration, a new 'ethnic-linguistic infrastructure' arose in the Middle Ages. The alphabet made it possible to turn spoken into written standards of communication, and this vernacularization provided a basis for the later nation-building.

The city belt in turn was a decisive factor in center formation and early state building. The revival of trade between the Orient, the Mediterranean, and the North Sea had produced a belt of independent cities from Italy to Flanders. The polycephality of the citystate Europe either delayed state formation (Germany, Italy) or led to confederations (The Netherlands, Switzerland). By contrast, monocephalic city structures conducive to centralized state development formed at the western and later also at the eastern end of the city belt. The rise of Atlantic capitalism, however, brought with it a European division of labor that strengthened the Western states, while the Eastern states largely became incorporated into empires.

The variations in city structure and thus the tripartite division of Europe into a Western Europe of unitary, centralized nation-states, a federalist Central Europe, and an Eastern Europe of multi-ethnic empires constitute the 'west–east axis' in Rokkan's well-known typological–topological model or conceptual map of Europe. It helps to explain the great paradox of European development: that the strongest and most durable systems emerged at the periphery of the city belt and the old Empire.

While the 'west–east axis' defines differences in economic conditions of state formation, the 'south–north axis' denotes differences in cultural conditions of nation building, depending on the geopolitical distance northward from Rome. Rokkan interpreted the Reformation as the first major step in defining territorial nations in Northern Europe, whereas in Catholic Europe the Church remained supraterritorial.

The conditions for successful center-building and territorial consolidation as well as for cultural standardization and identity-building, therefore, varied across Europe in a typical and topological way. The easier these processes of territorial and cultural boundary-building, Rokkan generalized, the more gradual and less violent were the internal restructuring of the nation-states through the establishment of citizenship rights.

Towards the end of his life Rokkan attempted to integrate his comparative studies into a comprehensive model of Europe spanning the entire history of state formation, nation-building, and mass politics in Europe and identifying the crucial variables in the long and complex process that led up to the current constellations of territories, economies, and systems of political alignment. This macro-model remained a sketch exemplifying the nature of most of Rokkan's studies, which were essentially research programs.

Rokkan's Comparative Approach

Rokkan defined the decisive challenge of cross-cultural, cross-societal, and cross-national research as follows: 'We propose to direct our attention to one single, central set of issues: the

possibilities of translating "grand theory" into empirically workable "typologies of macro-settings" for variations in human behaviour, and the consequences of such typologies for decisions on the cultural and the geographical range of comparisons at the level of communities, households, and individuals' (Rokkan, 1970b).

From this fundamental orientation, Rokkan developed a series of basic elements that characterize his comparative approach: the need for context-specific comparisons, the multidimensionality of contexts, the twin concepts of structuring and boundary-building, the dynamics of freezing and unfreezing at critical junctures, and finally the method of retrospective diachronics.

For Rokkan model-building always had to be context-specific: 'No single explanatory or intervening variable can be linked up with a dependent variable in isolation from the context, whether across systems or across stages' (Rokkan, 1999). Because such macro-contexts are comparable only to a certain degree, Rokkan was skeptical of universalizing comparisons and instead preferred region- and period-specific models, particularly in his cleavage model and his conceptual map of Europe.

The macrocontexts are in principle multidimensional, since for Rokkan they always resulted from interactions between three different systems, the political, economic, and cultural systems. Due to their relative autonomy, he started in his macromodel from configurations, rather than hierarchies of factors: 'There is no economic determinism in the model, nor a geopolitical, nor a cultural: in this sense it seeks to combine the traditions of Karl Marx with those of Max Weber and Emile Durkheim.'

While all three systems have a territorial dimension, only the political system is ultimately capable of controlling territory. Territory is a crucial concept in Rokkan's comparative analysis of political systems. It is a starting point for his most basic twin concepts of structuring and boundary-building in the formation of territorial systems. Such systems essentially are structured through the formation of political, economic, and cultural centers and the incorporation of peripheries. Their consolidation presupposes a process of political, economic, and cultural boundary-building which in turn contributes to the emergence of territory-wide cleavages. Center-periphery relations and cleavage structures are the two dimensions that for Rokkan underlay the institutional and organizational structuring of political systems.

With his twin concepts, Rokkan went beyond traditional comparative research that was limited to comparing structures and focused on the formation of systems as a process of structuring and boundary-building. In this historical process critical junctures are crucial. This concept of Rokkan's refers to political, economic, and cultural 'revolutions,' historical events and processes which he used to distinguish between periods: between 'critical junctures' when basic new decisions are taken, structures are 'unfrozen' and boundaries are 'opened,' and periods in which the institutional and organizational structures remain 'frozen' and the boundaries 'closed.' This attempt at periodization underlies Rokkan's basic method of retrospective diachronic analysis which addresses the question: 'given an observed contrast in the values of variables at time t_i, what combinations of variables for earlier phases t_{i-t}, t_{i-2}, and so on, can best account for these differences?' (Rokkan, 1999).

In this way, Rokkan attempted to explain systematically current structural differences between the European societies and polities as arising from their long-term development. In many cases, his efforts to do so took the form of research programs. With his contributions to developing a social science infrastructure, he wished to create a basis for carrying out these programs. In both respects, Rokkan has left us an important legacy.

See also: Comparative History; Comparative Studies: Method and Design; International Communication: History.

Bibliography

Lipset, S., Rokkan, S. (Eds.), 1967. Party Systems and Voter Alignments: Cross-national Perspectives. Free Press, New York.

Rokkan, S., 1970a. Citizens, Elections, Parties. Approaches to the Comparative Study of the Processes of Development. Universitetsforlaget, Oslo, Norway.

Rokkan, S., 1970b. Cross-cultural, cross-societal and cross-national research. In: UNESCO, Main Trends of Research in the Social and Human Sciences. Mouton, Paris, p. 1.

Rokkan, S., 1999. State Formation, Nation-building, and Mass Politics in Europe. The Theory of Stein Rokkan. Oxford University Press, Oxford, UK.

Rokkan, S., et al. (Eds.), 1987. Centre–Periphery Structures in Europe: An ISSC Workbook in Comparative Analysis. Campus, Frankfurt, Germany.

Rokkan, S., Dogan, M. (Eds.), 1969. Quantitative Ecological Analysis in the Social Sciences. MIT Press, Cambridge, MA.

Rokkan, S., Eisenstadt, S. (Eds.), 1973. Building States and Nations, vol. 2. Sage, Beverly Hills, CA.

Rokkan, S., Merritt, R. (Eds.), 1966. Comparing Nations. The Use of Quantitative Data in Cross-national Research. Yale University Press, New Haven, CT.

Saelen, K., 1981. Stein Rokkan: A bibliography. In: Torsvik, P. (Ed.), Mobilization, Center–Periphery Structures and Nation-building. Universitetsforlaget, Bergen, Norway.

Roma People

Michael Stewart, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Roma people – formerly often referred to as 'Gypsies' – have a unique history, and a precarious current existence. They are estimated to be about 10 million in Europe, and some 2–3 million elsewhere. Their lives have been characterized by mobility within and between countries, and social marginality. The analysis of Romani language shows evidence of an Indian origin. In Europe the Roma have historically been strongly linked to a variety of occupations of trade and craftsmanship, but unemployment has tended to be high. While twentieth-century socialist regimes in Europe tended to integrate Roma into national economies by varied means, later economic regimes have again left them more often excluded.

In the year 2013, when this article was written, an ugly new tale of abuse, harassment, or persecution of Roma in different countries of Europe seemed to emerge daily. The claim from the Socialist French Interior minister that the Romany way of life was incompatible with French values stood out as did the witch-hunt style kidnapping of (blonde) Romany children from their families in Greece, Poland, and then Ireland; but one story seemed to evoke particularly powerfully the plight of Europe's largest minority.

Three years earlier, in 2010, a 40-year-old Romani beggar – who was a notorious vagrant in the town of Kosice, Slovakia – had been stabbed to death, in a savage 40 s attack, by a Slovak man half his age. Since then the killer had not spent 1 day in custody – originally, and incomprehensibly, charged with battery he had remained free when the charge was increased to murder. He then refused to testify during the preliminary proceedings and maintained his silence during the main trial.

In December 2013, the District Court in the Slovak town of Košice acquitted this non-Romani youth and set him free. The judges based their decision on psychiatric reports that for a fraction of a second, the killer had found himself under the influence of a delusional, psychotic, but temporary, disorder. He had been in a state close to unconsciousness in which his ability to tell right from wrong and his capacity for self-control disappeared and believed that the beggar intended to kill him. The psychiatrists added that they had no reason to think this man suffered from any enduring chronic mental disease or disorder – he should be set free at once.

In Europe one would have to go back to 1930s Germany or in the United States to the precivil war period to find court proceedings so racist and unjust, in which the life of a citizen is held in such contempt. But in 2013, so common are these kinds of events that they barely excite comment outside the online circles of engaged academics.

Who Are 'The Roma'?

The 'Roma' represent a more or less unique case in the history of the modern world. The label 'Roma'– which has largely replaced the former 'Gypsy' is now used to refer to a large globally dispersed 'community of communities' estimated at 10 million in Europe and maybe 2–3 million elsewhere, that, in anthropological terms, defies simple definition. Indeed, despite the fact that when 'Roma' travel from one country to another they have little difficulty recognizing other 'Roma' as 'fellow Roma,' but invariably determine that these are 'not our sort of Roma,' it makes no sense politically, sociologically, or anthropologically to talk of 'The Roma' as a single entity. In the absence of a Romany state or state-like institutions anywhere in the world, the homogenizing forces that produce national (or ethnic) type identities, 'The Roma' remain a remarkable exception to the iron laws of modern development that predicate cultural persistence on robust political bases. And yet Romany communities reproduce themselves while living both among a politically dominant majority and dispersed among such. How this persistence of Romany social forms is achieved remains a true puzzle for modern social science.

Scholarly debate on Romany identity and community can be broadly divided into two main positions. On the one hand we find scholars who adopt an 'ethnic origin of nations' approach (as in the work of Anthony Smith) and see 'The Roma' today as an ethnic diaspora united by the current or historical use of dialects of the unitary Romany language. For these scholars if Roma see themselves as a less well-defined group than, say, Jews or Armenians, this is more to do with aleatory conditions of history than their capacity to form a nation in the modern sense of the term. Such scholars will talk of the Indian ancestry not just of the language, Romani, but also of the people themselves. On the other hand we find scholars influenced by the institutional constructivism of Benedict Anderson as well as by anthropological approaches to ethnicity and nationalism. They see nations and ethnicities as the product of historically contingent, modern, political arrangements and refuse to project ethnic type identities back into the premodern world. These scholars, while being more or less interested in the historical connections among Romany populations of the world today, reject talk of diaspora, are more interested in diverse ancestral populations (in the plural) than in finding one root-stock and focus on the socio and economic niche that has enabled Romany persistence in Europe in particular since the first fourteenth-century reports of speakers of Romany were published. As a consequence, the former group of scholars tend (at least when being consistent) to exclude from their definition of Roma, all those populations like the Irish Travelers, the German Jenische, and others who

appear to have no historical connection with speakers of Romany. There remains, however, plenty of room for ambiguity in such tidying up. Do the 'Boyash (aka Baiesi)' who speak neither Romani nor use Romani terms in their dialects of Romanian or English, but whose way of life is distinctively 'Gypsy-like,' count as 'Roma'? For what it is worth, for the most part Baiesi deny their historical connection with 'Roma.'

From a political point of view, Romany diversity poses a huge challenge both for activists who would seek to build the European equivalent of a civil rights movement and for governments or transnational institutions like the EU or Council of Europe that seek 'legitimate partners' for dialogue and community leadership. While activists often seek to impose a phantasmagoric nineteenth-century vision of a unitary Romany culture forged in the 'Indic cradle of civilization,' the political institutions have created an equally phantasmagoric 'community of Roma' with self-appointed representatives of which they can then deal 'democratically and transparently.' From the point of view of social science, however, the diversity of Romany social forms alongside the undoubted familial resemblance of different Romany populations offers a unique chance to study the sources of social and cultural diversity in a nonnational context. It might also be noted that whatever the source of Romany adaptiveness, the spread of these communities over the world and their persistence through waves of persecution and assimilation suggests a resilient social form.

Language Histories – Romani

Romani is the name commonly used in scholarly literature for the language of the Rom. It is the (feminine singular) adjective derived from the collective noun that self-designates speakers of Romany, rom (the now pc Roma is a late interpellation from some western dialects (NB with a hacek on the r)), as in the phrase *romani chib (hacek on c too)*. Colloquially, speakers of Romani will refer to their language as *romanes* (literally 'in the Romany way') but in the academic literature the term Romani has stuck. It is the only Indo-Aryan language that is spoken exclusively in Europe (and among migrant populations in the New World). As an Indic language it is also unique in having a European ('Balkan') syntax.

The earliest source on Romani is a short list of 13 sentences published as *Egipt speche*, with English translation, in 1542 by Andrew Borde, but it was not until the late eighteenth century that an international circle of scholars engaged in a correspondence that led to Professor Johann Rüdigger of Halle publishing, in 1777, the first systematic comparison of Romany spoken in Europe with 'Hindustani' that established the Indic origin of this language. Because of the absence of written sources the history of the language can only be constructed speculatively using the principles and procedures of historical linguistics. Although the eighteenth-century scholars were pioneering autodidacts, modern historical linguistics has confirmed in great detail their initial insights. In the most substantial examination of the topic, Yaron Matras of Manchester University, has concluded that proto-Romani – the language that was spoken in India – before any contact with middle eastern languages like Persian – was something of an Indian hybrid language, a central Indian dialect that at some time in the first millennium CE underwent a convergence with northern Indic languages. As Matras (2002) stresses throughout his work, Romani has always been a language in contact, and the shared traces of contact languages found in all modern version of Romani (in particular Persian and even more importantly Greek loan words) provide compelling evidence of the habitat of ancestral speakers of the language. Extending our understanding of the linguistic data to population history is, however, fraught with difficulty. Does the presence of many Greek loan words in Romani reflect the 'relatively recent' contact or a more powerful (possibly more enduring?) social contact between speakers of Greek (possibly in Asia Minor) and Romani speakers?

Though Romani is a language in contact it is not – as some have sometimes speculated – a contact language in the sense that Yiddish arose as an ethnolect of Jews living in the Rhine area which then spread in particular to eastern Europe and was subsequently spoken from Hamburg through Prague and Krakow to Alexandria in Egypt. In accordance with its origin, Yiddish shares most of its lexicon and morpho-Syntax with (Southwestern Franconian varieties of) German. Romani, with its significant Indic lexicon and Indic inflections, is quite a different case.

However, extrapolating social history from linguistic history is no simple matter. Speaking rather generally, the existence of a number of dialect groups among Romani speakers in Europe demonstrates a fairly clear pattern of demographic division. The linguistic map today demonstrates a distinction between Balkan (e.g., Arli or Xoraxane), Vlax (originating from Transylvania and adjoining regions, e.g., Lovari, Kelderas), Central (of the former Hapsburg lands, e.g., the 'Romungro' dialect spoken today in Slovakia and western Romania), Northeast (Baltic–North Russian, e.g., Polska and Latvian Romany), and Northwest (German–Scandinavian, e.g., Sinti and Finnish Romany) dialects. To a large extent this map reflects early modern political boundaries. But the more outlandish claims of an earlier generation of amateur linguists and philologists, like the Hungarian Jozsef Vekerdi – that for instance Romany social history (nomadism, supposed aversion to agriculture, etc.) could be traced in the lexical retentions from Indo-Aryan – are now wholeheartedly rejected by modern linguistics.

Because the linguistic evidence is all we have to go on in reconstructing Romany population histories, considerable effort has been put in by a small group of population geneticists to try and trace 'origins' of some of today's Romany speakers. At considerable public expense they have proven nothing more than that Romany has been passed on within families over the past 60 or 70 generations, and that, therefore, there is clear evidence of an Indic origin of some of the ancestral speakers of Romany. This work – complete with its overblown attempts to date demographic events and frequent downplaying of profound founder effects in the data – represents a weak tail-end of the popular field of population genetics. Far more interesting and significant has been work in comparative linguistics that has shown the historical links of Romani and other Indic languages spoken in the middle east – Domari in particular. Thanks to this work – most of it carried out under Yaron Matras' leadership – it is now clear that Romany is but one of a group of Indic languages that have been

spoken outside of South Asia in the past 1000 years. Of perhaps greatest interest is the possibility that, linguistically, Romani and Domari separated long before speakers of either language migrated west, but that speakers of both languages may have occupied parallel occupational and social niches in the host societies of the near and middle east producing various parallel evolutions in the languages.

Of today's estimated 9–10 million 'Roma' in Europe, some 3.5 million or more speak Romany. The younger generation of these speakers is currently in the process of creating a loose international dialect and standard written form based on electronic communication. Though there have been self-serving attempts by some self-appointed 'experts' to impose a uniform standardized Romany it appears that the sheer mass creativity unleashed by the Internet will overwhelm these academic replays of nineteenth century 'language renewal programs.'

In some regions of Europe, especially the western margins (Britain, the Iberian peninsula, Scandinavia), Romani speaking communities have given up their language in favor of the majority language, but have retained some Romani-derived vocabulary as an in-group code. Such codes, for instance Angloromani (Britain), Caló (Spain), or Rommani (Scandinavia) are usually referred to as *Para*-Romani varieties. Probably a majority of those identified as Gypsies in Europe today are more or less monolingual speakers of their national tongue, like the large Gitano population in Spain and the 'Hungarian Gypsies' within modern Hungarian borders. The extent to which they identify with Romany speakers is particularly labile and open to entrepreneurial innovation.

Occupational Histories – The Legacies of Infamy

Today we find individuals who self-identify as Rom in all walks of life from Presidential counselors, through the legal and medical professions to political activists and academics, actors, classical and jazz musicians, writers, poets and journalists, as well as the whole range of manual and technically skilled professions. There is also a strata of Rom who are actively involved in various entrepreneurial activities from animal markets through second hand cars to antiques and real estate. However, the vast majority of people who either self-identify as Roma or are identified as 'Gypsy' by the majority society are in occasional, manual, or other unskilled employment. Across eastern Europe where rates of employment of unskilled manual labor are very low (compared with western Europe) and especially where economic reform has stagnated (Hungary, Slovakia, Bosnia above all) massive unemployment characterizes Romany communities. In more dynamic economies, like that of Romania, migration to the labor hungry markets of Italy and Spain has provided new demand for low-paid labor.

Historically, it is most unlikely that Romany communities ever adopted a single economic niche. But the range of their economic activities does seem to have been generally constrained by a combination of: preadaptation to the provision of services to the agricultural or urban economy; legal and practical restrictions on residence/land-ownership (no Romany communities have benefited from the various land reforms of the past century in eastern Europe, not even the most recent post-1990 one); and, finally a political culture that promoted independence over security. Almost certainly, however, in the distant as in the recent past Romany persistence relied upon enormous flexibility in economic strategy – a willingness to combine day-laboring with begging with trade with artistic performance, as opportunities presented themselves. Through history such flexibility would seem to be essential for survival on the social margins.

In terms of historical evidence the German speaking states of the eighteenth century provide the best documentary evidence for premodern Romany economies. All those people officially labeled *Zigeuner* were part of a vast swathe of population below the three official Estates, which 'wandered about the country,' as the distinctly pejorative German verb *herumziehen* had it. The *Zigeuner* were thus just one part of a broader but inherently suspicious group that was more or less socially alien to local society. Boundaries between the *Zigeuner* and other 'masterless men' were weak. Among these footloose people were all those who dealt with carcasses and the polluting processes of bodily decay as well those who put their persons on display in public, like actors, puppeteers, musicians and prostitutes, people whose very touch could pollute honorable persons. This *unterschicht* ranked in its numbers all those who supported themselves precariously through the peddling of trade and services, the makers of shingle, sulfur-matches, spindles, pick-hammers, blow-pipes and tinder, spoons, baskets, sieves, boxes, and powder-horns, repairers of coffee-grinders, book-binders, knife- and scissor-grinders, pitch-burners, itinerant traders in Frankincense, coopers, hawkers of lace, pewter, heels, rope, southern fruits, hats, pots, leather, vegetables, and tobacco. Certain specialisms recur in the records in conjunction with the label *Zigeuner*, some of which involve metal work, like the pot-binders (who repaired shattered pots with wire), buckle-makers, producers of dough-trays, grills, combs, wool-combs, drills and book-clasp makers; hosiery, broom and brush-makers, as well as the occasional 'stone-breaker' also make an appearance linked to the label *Zigeuner*.

In a fashion that is now quite hard to fathom, there was for eighteenth and even nineteenth century society something inherently suspicious and dubious about the providers of such a chaotic array of services and goods. They were a 'rabble of masterless men,' (*herrenloses Gesindel)* on edge of the *Gaunertum*, 'the world of rogues and vagrants.' They were involved an unsettled and wandering (*nichtssesshafte*) existence, and this, in and of itself, rendered a person infamous or dishonorable to a greater or lesser extent because it placed them outside of the hierarchy that was the basis of social life. Moreover, while the actor or musician was explicitly infamous for putting himself on display, the destitute peddler-beggar was, in the words of a French historian, Laurence Fontaine, "ultimately selling himself and all that his words – spoken or sung – could stir in the imagination of his listeners" and, as such, existed only a step away from his infamous self-exposing cousin.

In the poorer region of eastern Europe, where the great mass of European Rom have lived we can presume that income from such unsettled sources played only a minor role in most families' survival. Almost certainly then as in the twentieth century enduring ties to local farming families would have provided sufficient day laboring to get through the year. But here too this placed the Gypsies at the bottom of the social

order since the day laborer was the lowest of the low – even more worthless than the serf who at least had a permanent master.

In the context of this quasi-feudal social order that endured into the nineteenth century, one occupation is particularly worthy of mention – musicianship. In the Hapsburg lands in particular but also in premodern Spain, the provision of music provided Romany families with one means to transcend their inherited social status. Then, as today, the ability to conjure magic from singing or playing an instrument was the most promising route to transcending outcaste status for the socially marginal. In some cases – Hungary after the defeat of the 1848 revolution in particular, but twentieth century Spain provides another example – Romany musicians came to be associated with 'the national cause' and acquired an enduring elevation in social status. In the communist period, for example, Hungarian Gypsy musicians who traveled abroad in 'Gypsy Orchestras' saw themselves as outstanding ambassadors for their country. In post-cold war Europe a few bands of Romany musicians have also been able to plug into the world music scene with similarly transformative impact.

The greatest single transformation in the economic niche of Romany communities in Europe has, however, been the enforced employment programs of the communist regimes from 1947 to 1989. Across eastern Europe and the former Soviet Union the vast majority of those identified by the state as 'Gypsies' were both enabled and compelled to take up full time employment in 'socialist wage labor.' By the 1980s over 90% of Romany men and two thirds of the women were in full-time employment across the region. The economic policy drive to full employment was also the central plank in a social campaign to assimilate the Gypsy minority. On this front, all the evidence suggests, the regimes were less effective. Even under communist rule, Gypsies were given the worst paid, the dirtiest, the least attractive, and the least secure jobs and the poverty of their material conditions was reproduced despite all the egalitarian rhetoric of the 'workers' states. The more robust Romany communities also sought out areas of employment where whole families could work together and not under the controlling eye of a foreman – for instance in town sweeping (carried out in groups unlike the individualized rota of a hospital cleaner) or in nautical scrap iron yards. In other areas of employment, Roma benefitted from socialist planned economics. Over 20 000 Romany musicians were officially registered in Hungary in the late 1980s with several 1000 playing fixed slots in restaurants in the capital city. All these artificially created employment opportunities disappeared with the system change. Factories closed or were replaced by machines, for instance, several (like 10) unskilled laborers with one machine operated by a single semiskilled mechanic. And even in music fashions changed. By 2003, with all employment restrictions lifted, there were barely few score 'Gypsy musicians' making their living in Budapest's restaurants.

Although a significant part of eastern European Romany populations have benefited from the system change a majority have found themselves increasingly economically and socially isolated. The mixed society of 'masterless men' continued in one form or another through to 1989. But with the radical change in agriculture and the ending of subsidies to live in cities many Roma have found themselves driven out to villages from which the majority population has fled. For the first time in Europe we are seeing the emergence of *de facto* ethnic ghettos entirely cut off from the outside world. The demographic, social, and economic consequences of these developments have yet to be felt.

Political Histories

Narratives of unending victimhood have come to dominate activist-inspired 'histories' of the Romany people. In part inspired by the paucity of sources on Romany history (for much of the past 600 years we have only the occasional judicial record), the early history of Romany communities is often written as a progression from mistrustful welcome, through antagonism to repression and expulsion. In such accounts, centuries of discrimination reached their climax in the Second World War when the Nazis and their allies murdered well over 100 000 Roma (half a million according to activist histories). Communism then imposed a new version of repression in the form of assimilationist policies. According to such accounts it is only now in the twenty-first century that Roma may be able to stand up for their rights and demand the recognition that they deserve.

This kind of simplistic narrative is, of course, not wholly false but there is now substantial historical work showing how one-sided and misleading the miserablist account is. Most broadly speaking, the very paucity of data about Romany populations and even the repeated expulsion orders (repeated because unenforced) may be taken as evidence for a peaceful and, within the frame discussed above, tolerant coexistence, indeed symbiosis, of Romany groups with their host populations for most of the past 600 years. The work of Thomas Fricke and Ulrich Opfermann demonstrates a far more complex picture of mutual dependence, intermarriage, and flexible identities at least within the '*unterschicht*' of premodern society. Because the administrative term '*Zigeuner*' did not truly refer to an ethnic group in the modern sense of the term (there were no such things in eighteenth-century Europe) but rather to a (mostly inherited) life-style and because even Romany speaking *Zigeuner* were always bilingual in German, it was perfectly possible for people branded as *Zigeuner* to claim as did one woman interrogated in Stuttgart in 1781 that "now we are no longer Gypsies [since] we have been given *schutz* (protection)" that is to say entered a formal relationship of subordination with a master. She freely admitted that she spoke 'the Gypsy language' since her parents had used this when she was a child, but since becoming a resident of a community she and her family were now allowed to hawk their goods (porcelain from the Duke's factories) throughout the country and so no longer counted as *Zigeuner*. Or, as another Romany speaking man claimed some generations earlier, "I should not be seen as a Gypsy, rather a Trumpeter of the Kaiser" (Opfermann) – whatever these people's familial identifications as speakers of Romany, their labeling by the state was far less rigid than a twentieth century 'ethnic' framework would suggest.

More broadly, historical work has demonstrated that there has been no single 'European' history of Romany populations. A Foucaultian approach would suggest that a deviant population,

like itinerant Romany speakers, would have universally become a prime target of the modern surveillance state. In fact, as Leo Lucassen has demonstrated variations in the institutional configurations of modern states has profoundly affected the way Romany populations are treated by officialdom. There was a profound difference between the experience of Romany speakers and Gypsies in Great Britain, where after 1660 issues of vagrancy gradually became municipal and not national issues and Germany where from the eighteenth through the twentieth century there was a gradual criminalization of 'vagrancy.' Later, the way the police in Germany became significantly involved in administering welfare and public health issues also put them on the frontline in dealing with problems of poverty in the new cities of the Reich. As a result by the early twentieth century the German police had in place a whole system of surveillance of the 'hard to identify' populations – a system that the Nazified Criminal Police, or *Kripo*, put to extensive use in the 1940s when determining who was deported from Germany.

Communist policies toward the Romany populations of eastern Europe displayed a particularly radical form of what we might call the enlightenment approach first pioneered by the Hapsburg ruler, Maria Tereza, in the late eighteenth century. Being a 'Gypsy' was treated as a lifestyle and it was assumed that Romany cultural and linguistic practices were superficial traits that would be abandoned once full social integration was on offer. Postcommunist, democratic polities have, by and large, taken the opposite stance – stressing the distinct cultural and linguistic practices of Romany communities while condemning a large part of the Romany population to ever greater economic and social marginalization. Under the name of respect for multicultural difference, radical new systems of segregation are being invented in parts of Europe today.

Cultural Histories

The persistence of Romany sociocultural forms represents a truly fascinating sociological puzzle. Because Roma have not historically formed territorial settlements, all 'Roma' communities live in the middle of and dispersed among majority populations; they are always more or less familiar, indeed intimate, with the cultural world around them – always speaking the language/s of the dominant community; and yet they reproduce their 'communities' with apparent ease, retaining a sense of themselves as 'Rom,' and they do so without shared religion, without any form of ritual or political leadership, without overarching or underpinning political organization and without recourse to identity strategies rooted in transnational or diasporic rhetoric. In other words, they challenge that casual assumption promoted by the experience of growing up in well-homogenized nation-states that cultural and social continuity is predicated on the ability of states politically and institutionally to ground cultural forms. How Roma communities 'get by' has not yet been properly explained by any of the scholarly disciplines that deal with Romany sociality.

The diversity of Romany socialities – understandings of what constitutes a social order – has driven those scholars familiar with the inner life-world of Romany communities to reject the whole idea of there being a single 'Romany culture.' Patrick Williams and I recently coedited a text entitled *Des Tsiganes en Europe* (Gypsies in Europe – i.e., *not* The Gypsies of Europe) in order to stress this point. And yet there are continuities. Living in communities that exist within another culture and do so under conditions of dispersal, Roma use strategies of visibility and invisibility vis-à-vis the 'host' society, with a strong preference – as Patrick Williams (2003) has long stressed – for rendering the socially valuable invisible to the non-Rom world. It is perhaps for this reason that Rom are often keen to conceal their kinship ties in official papers – as if aware that this is the one area of social life where the authorities might take hold of them.

Another common feature of Romany strategy is that in the absence of any overarching social order, the atom of Romany sociality becomes the reproducing family – that is to say the network of kin within whom it is possible to reproduce over time. It is striking that words that are used by outsiders as if ethnonyms are, in a colloquial use, nothing of the sort. When German 'Sinti' talk of their community as 'amare sinte' the correct translation into English is not 'our kind of sinte' but 'our relatives.' Of course, in the national political arena German Romany activists are perfectly capable of adopting (in the German language) the malapropism of 'Die Sinti' as if the term named a people but, in a curious way, this very gesture also undermines their local credence as indigenous leaders.

Such strategies aside, how Roma accomplish the core business of social reproduction varies significantly from place to place and time to time, at least if the ethnographers are to be believed. Among Spanish Gitanos described by Paloma Gay y Blasco (1999, 2011), for instance, Romany sociality is constructed in the performance of a gendered identity by each individual Gitano and Gitana. Rigidly enforced by excommunication, the 'laws' of the community are not codified other than in the embodied gendered performance of each Gitano individual. Among the 'Manouches' of France – and probably 'the Sinte' of Piedmont and beyond the social is constructed around each individual's respect for his or her own dead and the knowledge that others have of this. In Patrick Williams' account, Manouches transcend the uncivilized, almost barbaric lack of restraint they observe among the Gaźos (non-Gypsies) by systematic but silent acts of renunciation. These unspoken acts of respect for the dead create, invisibly, an extraordinary moral edifice untouchable by the outside world. Among the Vlach Rom of Hungary, in the socialist period at least, sociality centered on the creation of a world of brotherhood. In particular in alcohol fuelled musical celebrations a short lived, single sexed community of Romani-ness provided an image of a social order that would transcend all the importune demands of a civilizing social order. Among the Gabor and Cortorari Rom of Romania the social order seems constructed around the transmission of valuable kula-type wealth in the form of putatively ancient beakers which contain familial potency. I could go on – in each of these Romany communities social value is constructed and its transmission through time takes place in significantly divergent ways. And yet, when these Rom come to meet up – in a French hospital Accident and Emergency department or a Spanish agricultural labor market, let us say – they recognize each other as 'Rom' but 'Rom of another kind' (*streina rom*). How such a sense of commonality may yet ground new *trans*-communal identity strategies – through

intermarriage, through social networking fora, through joint political activity – remains to be seen.

Certainly, ethnographic studies of the postcommunist period demonstrate how fluid apparently fixed arrangements can be. Former musician families in Hungary who had lost their old secure employment in the late 1980s were rapidly able to convert their social capital, working as waiters for shorter or longer periods in their former restaurants and then moving into commercial work in neighboring Austria. They were also able to generate new political capital in the postcommunist system of ethnic representation. As Romany cultural identity was being revived as a political value capital by the ethnicizing post-communist regime so these musicians, whose grandparents had begun an apparently irreversible trend toward cultural assimilation into the majority population, suddenly revived their cultural 'romaniness.' This itself led to new forms of social interaction among 'groups' of Roma who had, in the communist period, tended to keep to themselves.

Such shifts in the social value and significance of Romany ethnic identity can be found across eastern Europe in the past 20 years as two young Hungarian ethnographers, Cecilia Kovai and Kata Horvath have demonstrated. In the communist past although the Gypsy/non-Gypsy distinction provided the basis of all social interaction between members of these two categories in villages and small towns, because of the rhetoric of assimilation this fact was *never named* and so no one was ever called 'a Gypsy' to their face. This meant that the Gypsies lived under a constant pressure, that they should not behave in such a way as would force someone to call them 'a Gypsy.' They lived with the false promise that if they became educated, acquired wealth, lived like 'normal people' (read 'peasants' or 'workers') then they would be treated as such. But the promise was illusory. Now, however, thanks to economic, demographic, and, above all, political changes, the promise of assimilation has been taken back and 'the Gypsy' is constantly named and the Gypsies more or less happily bring their 'Gypsiness' to the public space. In consequence, the majority feels (quite unreasonably in many respects) that they are engaged in a battle for territory with this newly named presence, 'the Gypsies,' who are taking over their institutions (the school, the bars, economic life). They feel 'their' space is constantly encroached upon by this previously unnameable but ever present social force.

This shift, and many others in the stagnating polities of Hungary, Bulgaria, Slovakia, and the Czech Republic in particular have contributed to the rise of threatening and very innovative political forms that – for the first time in Europe – target 'The Gypsies' as the main social menace and the dominant social problem of our time. Political parties like the Hungarian *Jobbik* or Bulgarian *Attaka* or Slovak '*People's Party – our Slovakia*' represent the most innovative and dangerous forms of this new populist 'anti-Gypsyism.'

See also: Diaspora; Ethnicity and Migration in Europe; Europe: Sociocultural Overview; Immigration, Citizenship, and Integration: Social Work Connections; Roma People: Discrimination and Social Work Practice; Xenophobia and Nativism.

Bibliography

van Baar, H., 2011. The European Roma: Minority Representation, Memory, and the Limits of Transnational Governmentality (Ph.D. thesis). University of Amsterdam.

Gay y Blasco, P., 1999. Gypsies in Madrid: Sex, Gender and the Performance of Identity. Berg, Oxford, UK.

Gay y Blasco, P., 2011. Agata's story: singular lives and the reach of the "Gitano law". Journal of the Royal Anthropological Institute 17 (3), 445–461.

Gropper, R., 1975. Gypsies in the City: Culture Patterns and Survival. Darwin, Princeton, NJ.

Lucassen, L., Willems, W., Cottaar, A., 1998. Gypsies and Other Itinerant Groups: A Socio-Historical Approach. MacMillan/St Martin's, London.

Matras, Y., 2002. Romani: A Linguistic Introduction. Cambridge University Press, Cambridge, UK.

Stewart, M. (Ed.), 2012. The Gypsy "Menace": Populism and the New Anti-Gypsy Politics. Hurst, London.

Weiss-Wendt, A. (Ed.), 2013. The Nazi Genocide of the Roma: Reassessment and Commemoration (War and Genocide). Berghahn, London.

Williams, P., 2003. Gypsy World: The Silence of the Living and the Voices of the Dead. University of Chicago Press, Chicago.

Roma People: Discrimination and Social Work Practice

Špela Humljan Urh, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents the history, origin, common stereotypes, and current ethnic reality of the Roma people. The two most explicit myths (nomadism and homogeneity) are challenged. The Roma are recognized to be one of the most marginalized ethnic groups worldwide. Throughout their history, they have faced negative social attitudes among the dominant society, which is one of the strongest factors contributing to their marginality. Due to their ethnicity, Roma people face discrimination more often and in more spheres of everyday life than non-Roma. Roma children and Roma women, especially, face multiple discriminations. Antiracist social work theory provides social workers with knowledge on how to work with marginalized communities.

Introduction

This article presents the history, origins, common stereotypes, and current ethnic reality of the Roma people. The two most explicit myths (nomadism and homogeneity) are challenged. Roma people, numbering from 10 to 15 million in Europe, are recognized to be one of the most marginalized ethnic groups worldwide. Throughout their history, they have faced negative social attitudes from the dominant society. This is one of the strongest factors contributing to their marginality. Schooling, employment, housing, political representation, and negative media representation are the most criticized areas by those advocating on behalf of Roma people and their issues. In this article, special concern is placed on the two most oppressed groups of Roma ethnicity – Roma children and Roma women. One-half of all Roma in Europe are school children with only a small proportion of them being included in formal education. However, those in school often face different forms of discrimination like minimalization of their abilities and ethnic segregation in 'special' schools. Roma women, on the other hand, often experience double discrimination: sexual discrimination in their own community as well as ethnic discrimination as members of the Roma minority in the wider community. Special emphasis is placed to the antiracist social work when it comes to reacting to violence, aggression, and conflict behavior in Roma families. In such cases, ethnic pathologizing occurs when social workers attribute such behavior to their 'Romani nature.' Many examples of the social workers' reactions reveal racism in their practice. At the end, a clear stance to integration along antiracist principles (Dominelli, 1988; Jones, 2002; Thompson, 2001) in social work education is made. The author also encourages future social workers, especially members of Roma ethnic groups, to follow an emancipatory research approach.

Roma, Gypsies, Travellers – Who Are They?

The Origin of the Roma

Some Gypsies prefer to be called Roma (meaning 'men' in the Romani language, singular 'Rom' and adjective 'Romani'). Others prefer to be called 'Gypsies' in the official language of the country (e.g., Tzigany, Cigan, and Zigeuner). However, in some countries, the term 'Gypsy' has a pejorative connotation. It usually serves as a synonym for a thief, dirt, and untidiness; also the verb ciganiti in Slovenian means to rip somebody off or to cheat (Štrukelj, 2004: p. 66). However, the word Roma was accepted at the first Roma World Congress on 8 April 1971, in London, as a polite expression for members of the Roma minority. The term 'Gypsy' also refers to Irish Travellers (also called tinkers, itinerants). However, they do not share the same origins as Roma – the Roma originate from India, but Irish Travellers are recognized as an ethnic minority who separated from the settled Irish community. Despite their different origins, both minority groups share many common experiences of prejudice, exclusion, and marginalization (Hayes and Acton, 2007).

Soon after the Roma arrived in Europe, many authors began to write about them, speculating about their origin, and describing their cultural characteristics. Many myths were then constructed about their origin and migratory routes until the linguist A. Frederich Pott scientifically studied the Romani language and explained that the homeland of the Romani people was India, more precisely from the Punjab region in northwestern India. According to Hancock (1988: p. 183), "it has been the Gypsy language, or rather the study of its fragmented dialects, more than any other single factor which has led to the general acceptance of an Indian origin for its speakers."

They emigrated probably because of "repeated incursions by Islamic warriors" (Barany, 2002: p. 9) in the sixth century and reached Europe in the fourteenth century (Barany, 2002; Hancock, 1988, 1991; Matras, 2000, 2002). As noted by Matras (2000: p. 34), "migration forms a repetitive pattern throughout Romani history." Today, many have settled and a nomadic life for Roma people is rare. However, many migrate due to poverty and persecution and seek asylum in other countries. In the past, Romani people were subjected to persecution and expulsion, leading them to leave their places of residence. Their migratory movements continue nowadays as a result of political and economic circumstances.

Ethnic Statistics of Roma

There are no unified statistics on how many Roma live around the world. The existing statistics are most often being based on informal estimates of nongovernmental organizations in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28080-X

individual countries. European Commission estimates there are 10–12 million Roma in Europe. A variety of ethnic groups are covered by the label 'Gypsies: Cigany (Hungarian), Cyganie (Polish), Zigeuner (German), Zigenare (Swedish), Čigonai (Lithuanian), Tigani (Romanian), Zingari (Italian), Tsigane (French), Gitano (Spanish), and Cigani (Slovenian, Croatian, and Serbian).' They recognize divisions among themselves based on territorial, cultural, and dialectal differences. There are five main Roma subgroups; however, there are also internal differentiations based on the various crafts that they practice (or have practiced in the past) (Hancock, 2001; Matras, 2002).

Romani Language

Matras (2002: p. 238) states that the Romani language is primarily an oral language. The use of the Romani dialects (around 600) is generally limited to members of extended Romani families and it is estimated that the majority of adult Roma are bilingual (Matras, 2002: p. 238). Also, the Romani language serves as a symbol for the separation of Romani and Gadže (a specific term regularly used by the Roma to refer to those who do not belong to their group, also non-Romani) and as central criterion for group membership (Matras, 2002). A Romani language can be characterized as an ethnic language, maintaining a social distance between Romani and non-Romani people. Also Para-Romani languages (Matras, 2002: pp. 242–249) have gained ground – a mixture of Romani words and words belonging to the dominant majority. Thus, rather than speaking about one, unified Romani language (romani čhib), one speaks about the Spanish Romani, Hungarian Romani, Czech Romani languages, and so on.

Common Myths about the Roma

One of the most common myths about Roma is that they are nomadic people. Yet, a nomadic life was not always their free choice, but a mechanism of survival in a social environment made hostile by the dominant society in a certain local area (after the first emigration from India). Constant threats and not owning land themselves forced Romani people to move. In Western countries, Roma have also practiced commercial nomadism, but this has not been only a Romani way of life. Additionally, in Eastern Europe, Roma were pressurized into accepting a sedentary way of life. Forced settlement was (and still is) recognized as a main political measure applied to Gypsies. As a consequence, a majority of Roma in these parts of Europe is sedentary and nomadism is rare.

Yaron Matras (2000) highlights the distinction between Roma, Gypsies, and Travellers, which are often used as synonyms, assuming an inherent link between these groups. Associating Romani minorities with traveling groups has its roots partly in the traditional romantic stereotype of 'the traveling Gypsies.' But in the context of European migrations, the perception of their migration as traditional nomadism fails to acknowledge that they occurred as a consequence of political circumstances. Nowadays, migrations should not be confused with nomadism. In contrast to the general perception and image of 'nomadic Gypsies,' Matras (2000: p. 33) states that very few Roma are genuinely nomadic in the sense of lacking any permanent dwelling. There are Roma who have a permanent residence and are engaged in itinerant occupations, many travel seasonally to join their extended families, or due to international open-market trade activities.

Another common myth about Roma is they are a homogeneous people, constituting one people with a number of fixed characteristics. Traditionally, it was believed that Roma people have a specific 'Gypsy nature,' which makes them more aggressive or criminally active (Barany, 2002; Belton, 2005; Macura-Milovanović, 2006; Okely, 1994; Schuringa, 2005; Stewart, 1997). These ideas, which imply racist views, have played a prominent role in defining Gypsy identity since the end of the nineteenth century among many scholars. Belton (2005), for example, opposes these ideas and seeks answers in research that show that Roma people are composed of diverse populations that have both nomadic and sedentary ways of life, different dialects, religious practices, occupations, and nationalities. The diversity of Gypsy people is also one of the contributing reasons for an unclear answer to the question of what the Romani identity really is (Barany, 2002; Belton, 2005). What is more, Szuhay (1995: p. 111) believes that who falls into which category is more a political decision than a scientific one.

Marginalized Ethnic Community

The Roma are exposed to multidimensional marginality in everyday life. They are a "transnational, non-territorially based people who do not have a 'home state' to provide a haven or extend protection to them" (Barany, 2002: p. 2). In comparison to other ethnic groups, who share a similar position and experiences of marginality (Kurds and Jews), the situation of the Roma is exceptional and far more comprehensive (Barany, 2002). They came to Europe about seven centuries ago but, despite many changes in the continent's political systems, their position remains politically, socially, culturally, and economically marginalized.

The Roma community is historically recognized as one of the most oppressed and excluded ethnic groups in Europe. Throughout their history, they have faced negative social attitudes among the dominant society, which is one of the strongest factors contributing to their marginality. Roma peoples' development was systematically prohibited or constrained by the societies in which they had settled through the pursuit of assimilation strategies that oppressed them and inhibited their development. Such an attitude to Roma people was especially typical in the territory of the former Austro-Hungarian monarchy during the period of the rule of Maria Theresa and Joseph II, when the Romani language, culture, and history were denigrated and excluded (Barany, 2002; Guy, 1998; Klopčič, 2007; Štrukelj, 2004; Zimmermann, 1999).

In countries under communist regimes (Eastern European countries), hostility toward Roma people was covert. Today, in many countries they experience the denial of their formal and legal rights, as well as overt and covert, intentional and unintentional racism. Several international reports by the European Roma Rights Centre (ERRC), Open Society Reports

(OSI, 2001, 2002), and many smaller research studies from the countries where Roma people live, show that their social position is still worse than that of the majority population. This is evident in cases of mass structural unemployment of Roma people (in all European countries the level exceeds 50% of unemployment of Roma people), minimal educational standards for Roma children in schools, their absence (and silence) in the political sphere, negative media representation of Roma minorities, dehumanizing attitudes of social and health personnel, and social avoidance and ignorance among the majority population (for more, see ERRC web page; also Barany, 2002; Fonseca, 1995; Varsa, 2005).

Two of the most excluded social groups are Roma children and women. Being determined by their age and their gender, respectively, alongside their ethnicity, they represent two of the most vulnerable groups in society. The Roma ethnic minority group is usually discussed in general terms without a mention of gender and age differences. Roma women are a particularly invisible part of the Roma community. Their specific experiences in the public and private spheres remain hidden, with the resulting gender-insensitive practices including in social work contributing to the further discrimination of Roma women.

Roma Children

The Roma are a young population regardless of which society they live within. One-half of all Roma in Europe are school children with only a small proportion of them being included in formal education (Macura-Milovanović, 2006; OSI, 2001, 2002; Peček and Lesar, 2006). Typically, the percentage of the Roma children enrolled in primary schools tends to be low, and the rates of repeating a grade and dropping out are high. Additionally, the participation rate in secondary school programs is low.

One form of the most blatant discrimination against the Roma in education is the segregation of Roma children in special schools for intellectually impaired children (OSI, 2001, 2002; Zaviršek, 2001; Zaviršek and Urh, 2005). There is an evident disproportion, with the share of Roma children attending special schools being even higher in materially more underprivileged Roma settlements. Such disproportions can be found in all European countries where the Roma live. In some of these places, e.g., Germany and Hungary, these special schools are called 'Gypsy schools' (OSI, 2001, 2002). The children's poverty and the ethnicity rather than their potentially low individual intellectual abilities are being the two most common arguments for systematically directing them into special schools and justify 'special' treatment, segregation, and institutionalization. Roma culture is often pathologized and functions as a projection screen for everything deviant, undesired, and unvalued in the dominant culture. Ignorance, misunderstanding, and disrespect of cultural realities often result in the exclusion of Roma children from the system of regular education. Similar cases have also been known in the United States in relation to children of African-American origin (Harry and Klinger, 2006), indigenous people (Haig-Brown, 1988), and other ethnic groups (Devore and Schlesinger, 1999; Eriksen, 2002; Laird, 2008).

According to Thomas H. Eriksen (2002: p. 25), prejudices created and maintained by society in relation to members of ethnic minorities include several mechanisms that maintain the dominant position of the majority society. One of these is the stereotypes that function as 'self-fulfilling prophecies' that constrain, inhibit, or stunt the intellectual development of members of an ethnic minority group by systematically telling them that they are inferior. This is illustrated by the following examples gained through the authors' research:

"The Roma children have low abilities in mathematics" (statement by an expert worker when sending a Roma child to a special school; Zaviršek and Urh, 2005).

"Roma children obstruct classes and impede normal work" (stated by a teacher when justifying a separated school section for Roma children in a regular primary school; Krek, 2005).

"It would be better if the Roma learn to draw as they will never learn to write anyway" (stated by a teacher in the first grade of a primary school while teaching writing; personal communication, March 2005).

Also, a pediatrician tried to persuade a Roma mother during her daughter's systemic medical checkup conducted before starting school to enroll her daughter in a special school where her experience of schooling would supposedly be easier (personal conversation with young Roma girl, March 2012). Today, this young girl is completing her doctoral studies and will be the first Roma woman in Slovenia to have obtained a doctorate. It is common knowledge that children from a poor environment are more willing to accept the justifications of experts who direct them to special schools. However, some research has proven the opposite, i.e., children who were materially and socially supported became more successful with better learning outcomes despite belonging to an ethnic minority (Krek, 2005; Macura-Milovanović, 2006; Peček and Lesar, 2006; Rorke and Wilkens, 2006). This constitutes important knowledge for social workers because it shows that positive encouragements and enabling positive experiences of underprivileged children can bring better results, and improve the development of their sense of being included and appreciated.

Roma Women

Roma women often experience double discrimination: first as women in their own community (sexual discrimination), and, second, as members of the Roma minority in the wider community (ethnic discrimination). Both forms of discrimination frequently coexist. Social workers often believe that violence, aggression, and conflict behavior are part of Roma culture, and part of their 'Romani nature' (Urh, 2009). This behavior among Roma people is considered more acceptable by mainstream society. It is also why experts often do not intervene or intervene much later than they would in the case of violence or child neglect in a non-Roma family, or in marriages between minors among the Roma. In the latter instance, social workers who addressed one such case attributed the problem to Roma culture rather than defining it as sexual abuse. This exemplifies the denial of sexual abuse. When a Roma woman becomes a victim of violence and looks for expert support, she is mainly seen as a Roma, or a member of the Roma community, rather than a victim of violence. Therefore, ethnicity becomes closely interrelated with pathologization as a "process where a certain individual's

or group's ethnic/cultural characteristic[s] [are] constructed as natural and innate, and described with medical and psychological concepts determining individuals or groups as being prone to certain physical and mental illnesses, pathological violence and lower intelligence or even mental underdevelopment" (Zaviršek, 2010: p. 91). Expert workers who do not reflect on their prejudices, and do not acquire knowledge about ethnically sensitive work principles (Devore and Schlesinger, 1999), allow their assumed stereotypes unintentionally and sometimes unconsciously to influence their professional work (Urh, 2008, 2011). This means that their practice maintains existing forms of discrimination or creates new ones.

In many Romani traditions, Roma women are expected to assume the role of a woman at a very early age, to get married young, and have children early. All these roles prevent Roma women from accessing the labor market and make it difficult for them or even impede their having a family and working life at the same time. Particularly vulnerable are underage Roma girls who take on the caring role (for younger brothers and sisters or their own children), thereby interrupting their schooling before they have completed it. This is why a low education rate and a high illiteracy rate are characteristic of Roma communities throughout Europe (Barany, 2002; Corsi et al., 2008; Minority protection, 2001, 2002). The European report on the position of Roma women (Corsi et al., 2008: pp. 107–108) reveals a high rate of poor education among the Roma living in the countries included in the report.

The low education rate is one reason for the low employment rate of Roma people, but it is not the only one. Cases of prejudice and discrimination against Roma workers at the workplace by their employers and coworkers are evidenced across Europe. Apart from these, Roma women encounter additional obstacles, namely, having to take on their traditional role in the family, experiencing frequent pregnancies, and lacking qualifications and knowledge, which results in their having an even higher unemployment rate than Roma men (Corsi et al., 2008: p. 119).

One of the most debated areas regarding the rights of Roma women concerns their reproductive rights. Special rooms are known to exist for Roma women in maternity hospitals, e.g., 'Gypsy rooms' in Hungary (Izsák, 2004), along with worse health care treatment in general (OSI, 2001, 2002). The rooms for Roma women also reflect their stigmatization and their ascribed uncleanliness. The most frequent cause of their being separated lies in the implicit fear of non-Roma that coming in contact with Roma people leads to becoming infected by the stigma that they are supposedly transmitting. In a way, the rooms for Roma women represent a social distance creating and maintaining differences between 'clean' and 'unclean,' and maintaining the discourse of purity (Douglas, 2005) and stigmatization (Goffman, 1991).

Social Workers' Engagement with Roma Service Users

Racism in Social Work

Many reports recognize the vulnerability and experiences of exclusion and poverty of Roma children and Roma women are much higher and much more common than that among Roma community in general (see Web page of ERRC; also OSI 2001 and 2002 country reports). This recognition constitutes important knowledge for social work, which should trigger responses from social services. Social work seems to be the 'ideal' profession to contribute to the abolition of discrimination because of its insights into microlevel dynamics, knowing peoples' environments, social networks, and their everyday experiences; and having actionable knowledge about methods of working with marginalized social groups, professional ethics, and an understanding of sociological phenomena. Nonetheless, relations between social workers, who belong to the majority population, and Roma people, who face discrimination, are often not satisfactory (on either side) and inefficient. This can be seen in legalistic approaches or traditional social work relationships where the social worker is an expert and Roma service users are passive recipients of the services, and many unintentional, implicit racist responses. They are often treated through racist practices reflected in stricter control over their use of social benefits, the limitation or abolition of social benefits due to nonfulfillment of certain conditions only applying to Roma people, indifference, the passive attitudes of social workers who conduct 'ping-ponging' strategies by sending Roma service users from door to door, and placing responsibility for improvement on the Roma themselves (victim blaming) or other institutions. Every real-life case shows that expert services and public discourses refer to the cultural relativism and pathologization of Romani culture, especially in relation to deviant behavior such as violence against women and children, alcoholism, child neglect, and sexual abuse. In this way, social work contributes to racist practices or maintains the status quo.

Antiracist Responses

Social services are not immune from racist influences in their personal, cultural, and institutional contexts, making the integration of antiracist principles (Dalrymple and Burke, 2003; Dominelli, 1988; Thompson, 2001) into social work practice important. In Great Britain, social work has a rich history of antiracist social work development in the educational system as well as in practice from the 1980s onward and has influenced the rise of these perspectives in the social work education in other Western European countries. However, in Eastern Europe, antiracism is in its infancy. Until recently, ethnically sensitive and antiracist theory had attracted little attention in Eastern Europe compared to other social work topics (Urh, 2008; Zaviršek, 2007). One influential reason was political. The official ideology of the communist/socialist leadership in the previous political system until the end of 1989 focused on class equality, which was to overcome all other differences including gender, ethnicity, age, and religion. The fact that social inequalities were an undesirable and politically conflicting subject of discourse hindered the development of antiracist social work.

There is a strong need to develop a culturally sensitive learning model for social workers (Devore and Schlesinger, 1999; Laird, 2008; Thompson, 2001) that is based on ethnic/cultural sensitivity and antiracist principles. The model is based on the idea of recognizing the cultural needs and differences of local ethnic groups, as well as antiracist principles that recognize discrimination and oppression (Dominelli, 1988). One crucial component is an understanding of historical

experiences of the discriminated groups. This would enable social workers to raise their awareness about the discriminated groups and their collective experiences of not being accepted as belonging in the societies that they live within. It also offers a better understanding of certain types of their behavior (fatalism, conformism, passivity, and distrust), which is often the result of the prolonged experience of exclusion and being underprivileged (Urh, 2009). Understanding the dynamics of racism, which have been changing historically and taking on different forms (Dominelli, 2007), also enables social workers to understand its complexity – its extensiveness, multidimensionality, and how the many levels of racism are intertwined and influence each other (Dominelli, 1988; Jones, 2002; Thompson, 2001).

How should social workers respond to the ethnic reality of Roma people or Travellers whose origins may be different but have marginalization in common? In the United Kingdom, special emphasis is placed to education of Traveller children's families (Cudworth, 2008). Local education authorities ensure the availability of education for all school-age children in the area, whether permanently or temporarily that is appropriate to their age, ability, and aptitude. This is of particular relevance to Traveller children.

One of the recognized examples of good practice can be seen in community social work from Haringey Council (UK). Social workers follow the community social work model to provide access to education for Travellers, Gypsies, and Roma families, thereby raising the achievements of Traveller, Gypsy, and Roma children and young people, promoting Traveller and Gypsy culture in schools, and helping Traveller, Gypsy, and Roma pupils and young people to achieve their full potential (Davis, 2010). The community social work model is also recognized as one the most successful models working with marginalized Gypsy communities in Eastern Europe (Schuringa, 2005; Rorke and Wilkens, 2006). The main focus of the community model is shifting from 'the culture of poverty' toward 'the culture of development' (Schuringa, 2005: p. 20). The community development approach refers to the establishment of mechanisms for mobilizing community resources, which lead to community problem solving through a grassroots approach. The Roma are seen as subjects and participating fully in all stages of the work and are not treated as passive objects of support provided by the others (top-down approach). Its main characteristics are building on a community's strengths, partnership, and emancipatory approaches, which also allow for positive experiences (e.g., feeling their own capabilities and engendering responsibility).

Many principles of good practice are derived from the Open Society Institute's (OSI) many years of working with Roma people. Developing different integrative projects for Roma people over the years, OSI has learned not to fund projects that in their design, budgeting, and implementation neglect the key issues of structured or substantive Romani involvement and partnership; are devoid of strategic value; have no prospect of sustainability; leave no worthwhile legacy or concrete lasting results; projects that fail to make connections between inputs and outcomes or confuse means and ends; offer no clear outcome that justifies activities, training sessions, and seminars; and neglect to include Roma in the processes of monitoring and evaluation (Rorke and Wilkens, 2006).

Cultural advocacy is one of the competences of culturally competent social workers. Being an advocate for Gypsies/Travellers means to defend, support, and argue for them on individual and/or collective levels with the intention of improving their quality of life and ensuring their citizenship and civil rights. The goal of the cultural advocacy is to strengthen marginalized ethnic communities and encourage their emancipation. Cultural advocates understand and respect cultural differences and experiences of certain communities, and serve as mediators between cultures. In a way, cultural advocacy in social work is the voice of marginalized and often unheard members of ethnic minorities.

Recommendations for Future Practice and Research

Despite the rise of antiracist and ethnically sensitive social work education in the United Kingdom, Canada, and the United States from 1980 onward, East European countries still remain mostly without adequate literature regarding this field. In these environments, there is a strong need to develop culturally sensitive learning models for social workers that are based on ethnic/cultural sensitivity and antiracist principles.

In the last decade, there have been several publications about Roma peoples, in which the authors focus mainly on theoretical perspective and discuss their situations. These are valuable contributions to understanding their ethnic reality and have triggered several discussions about discrimination of Roma people. A disadvantage of these publications is that they rest at the level of analyzing the current situation and problematizing existing systems without focusing on the question of how to tackle Roma peoples' exclusion.

Another problem with many publications is the absence of Roma peoples' voices, thus disadvantaging them. Also, social workers are usually members of ethnic majorities. Therefore, one future role of the profession is to recruit ethnic minority social workers. This is not an easy task since many members of ethnic minorities are subjected to discrimination also through the field of education – which offers social work another challenge, namely, to create circumstances that would enable inclusion. This cannot be done overnight, but needs long-term social action:

- Provide Roma employment opportunities that will reduce their dependence on social benefits and provide positive role models for young generations.
- Provide support to include Roma people in sociopolitical discourses that empower the whole community.
- Be critical of negative media representations of Roma people that generate conflictual attitudes to the whole community, proposing instead positive ones.

From the perspective of research, understanding the position of researchers is important in avoiding their contributing to the creation of 'otherness' (Fine, 1994; Dominelli, 2007). This makes adopting an emancipatory research approach (Humphries, 2000) and following 'the natives' point of view' crucial (Geertz, 1974). Respecting and following this perspective makes 'the objects' of the research into 'subjects,' gives them voice, and encourages them to write their own biographies. Emancipatory research enables the researchers to

avoid one-dimensional 'expert' positions, the manipulation of ethnic minority peoples' realities, and misinterpretations of their feelings.

Social work should more strongly and explicitly accept the role of promoting social change and acting as a catalyst between institutions and the community over being an actor promoting bureaucratic demands (Urh, 2009). Social workers have to constantly rethink their position of being supportive on one side and controlling on the other. Neil Thompson (2001: p. 11) challenges social workers to think about whether they are a part of the solution or a part of the problem because "… a social work practice which does not take into account oppression and discrimination cannot be seen as a good practice …" (Thompson, 2001: p. 11).

See also: Discrimination and the Law; Minorities; Multiculturalism.

Bibliography

Babič Ivaniš, N., Urh, Š., Klopčič, V., Adamič, M., 2006. Researching educational and vocational aspirations of Roma. In: Žagar, N., Klopčič, V. (Eds.), Poklicno Informiranje in Svetovanje Za Rome (PISR). Zavod za izobraževanje in kulturo, Črnomelj, pp. 197–241.

Barany, Z.D., 2002. The East European Gypsies: Regime Change, Marginality, and Ethnopolitics. Cambridge University Press, Cambridge.

Belton, B., 2005. Questioning Gypsy Identity: Ethnic Narratives in Britain and America. AltaMira Press, Walnut Creek.

Corsi, M., Crepaldi, C., Samek Lodovici, M., Boccagni, P., Vasilescu, C., 2008. Ethnic Minority and Roma Women in Europe: A Case for Gender Equality?. http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=492&type=2&furtherPubs=no (accessed 23.11.12.).

Cudworth, D., 2008. 'There is a little more than just delivering the stuff': policy, pedagogy and the education of Gypsy/Traveller children. Critical Social Policy 28 (3), 361–377.

Dalrymple, J., Burke, B., 2003. Anti-oppressive Practice. Social Care and the Law. Open University Press, Maidenhead.

Davis, R., 2010. Specialist team earns trust of travellers. Community Care, 20–21 (10.06.10.).

Devore, W., Schlesinger, E.G., 1999. Ethnic-Sensitive Social Work Practice, fifth ed. Allyn & Bacon, Boston.

Dominelli, L., 1988. Anti-racist Social Work: A Challenge for White Practitioners and Educators, first ed. Macmillan Education, Basingstoke.

Dominelli, L., 2007. Multi-ethnic Europe: diversity and the challenges of 'race', racism, ethnicity and nationalism. In: Zaviršek, D., Zorn, J., Rihter, L., Žnidarec Demšar, S. (Eds.), Ethnicity in Eastern Europe: A Challenge for Social Work Education. Faculty of Social Work, University of Ljubljana, Ljubljana, pp. 19–38.

Douglas, M., 2005. Purity and Danger. Routledge, London.

Eriksen, T.H., 2002. Ethnicity and Nationalism: Anthropological Perspectives. Pluto Press, London.

Fine, M., 1994. Working the hyphens. Reinventing self and other in qualitative research. In: Denzin, N.K., Lincoln, Y.S. (Eds.), Handbook of Qualitative Research. Sage, London, pp. 70–82.

Fonseca, I., 1995. Bury Me Standing: The Gypsies and Their Journey. Vintage Books, New York.

Geertz, C., 1974. "From the native's point of view": on the nature of anthropological understanding. Bulletin of the American Academy of Arts and Sciences 28 (1), 26–45.

Goffman, E., 1991. Stigma. Časopis Za Kritiko Znanosti 19, 137–154.

Guy, W., 1998. Ways of looking at Roma: the case of Czechoslovakia 1975. In: Tong, D. (Ed.), Gypsies. Garland Publishing, Inc, New York, London.

Haig-Brown, C., 1988. Resistance and Renewal: Surviving the Indian Residential School. Arsenal Pulp Press, Vancouver.

Hancock, I., 1988. The development of Romani linguistics. In: Jazayery, M.A., Winter, W. (Eds.), Languages and Culture – Studies in Honor of Edgar C. Plome. Mouton de Gruyter, Berlin, pp. 183–223.

Hancock, I., 1991. The Eastern European roots of Romani nationalism. In: Crowe, D., Kolsti (Eds.), The Gypsies of Eastern Europe. M.E. Sharpe, Armonk, pp. 133–150.

Hancock, I., 2001. Ame sam e rromane džene/We are the Romani People. The Open Society Institute, New York.

Harry, B., Klinger, J., 2006. Why are So Many Minority Students in Special Education? Understanding Race & Disability in Schools. Teachers College Press, Columbia University, New York.

Hayes, M., Acton, T. (Eds.), 2007. Travellers, Gypsies, Roma: The Demonisation of Difference. Cambridge Scholars Publishing, Newcastle.

Humphries, B., 2000. The researcher-participant relationship: arguments for an 'emancipatory' research paradigm. In: Truman, C.M., Humphries, B. (Eds.), Research and Inequality. UCL Press, London, pp. 191–205.

Humljan Urh, S., 2014. Anti-Roma racism in Europe: past and recent perspectives. In: Lavalette, M., Penketh, L. (Eds.), Race, Racism and Social Work: Contemporary Issues and Debates. Policy Press, Bristol, pp. 115–130.

International Helsinki Federation for Human Rights (IHF), June 8–9, 2005. The Situation of Roma in Selected Western European Countries. Report to the OSCE Conference on Anti-semitism and on Other Forms of Intolerance. Spain, Cordoba. http://www.refworld.org/docid/46963b005e.html (accessed 03.07.13.).

Izsák, R., 2004. "Gypsy Rooms" and Other Discriminatory Treatment against Romani Women in Hungarian Hospitals. http://www.errc.org/cikk.php?cikk=2063 (accessed 25.11.12.).

Janko Spreizer, A., 2006. Migration Roms: exemple de deux familles en Slovénie. Etudes Tsiganes, nouvelle série 27 (28), 56–69.

Janko Spreizer, A., 2009a. Engagement for informed anthropological knowledge on Gypsies. Antropologija 9, 55–72.

Janko Spreizer, A., 2009b. Situation of the Roma in Slovenia: politics for a regulation of the Roma ethnic community protection. Social Sciences Eastern Europe 2, 37–40.

Jones, C.P., 2002. Levels of racism. A theoretical framework and a gardner's tale. In: La Veist, T.A., Isacc, L. (Eds.), Race, Ethnicity and Health: A Public Health Reader. Jossey-Bass, San Francisco, pp. 311–318.

Klopčič, V., 2001. The legal status of the Roma in Eastern and Central Europe. Donauraum 40 (1–2), 71–83.

Klopčič, V., 2004. Roma in the process of European integration. Etniškumo Studijos, 102–113.

Klopčič, V., 2007. The Situation of Roma in Slovenia: Roma and Gadže. Inštitut za narodnostna vprašanja, Ljubljana.

Krek, J., 2005. Action model of educating Roma children in primary school Bršljin. In: Autor, S., Kuhar, R. (Eds.), Poročilo Za Spremljanje Nestrpnosti 04. Mirovni inštitut, Ljubljana, pp. 6–14.

Laird, S.E., 2008. Anti-oppressive Social Work: A Guide for Developing Cultural Competence. Sage, Los Angeles.

Macura-Milovanović, S., 2006. Children from the Romani Settlement Deponija. Pedagogical Perspectives of Inclusion of Roma Children in Education System – an Analysis of Action Research. Pedagoška fakulteta, Ljubljana.

Matras, Y., 2000. Romani migrations in the post-communist era: their historical and political significance. Cambridge Review of International Affairs 12 (2), 32–50.

Matras, Y., 2002. Romani: A Linguistic Introduction. Cambridge University Press, Cambridge.

Okely, J., 1994. Constructing difference: Gypsies as "other". Anthropological Yearbook of European Cultures 2, 55–73.

OSI - Open Society Institute, 2001. EU Accession Monitoring Program 2001: Monitoring the EU Accession Process: Minority Protection. Open Society Institute, Budapest.

OSI - Open Society Institute, 2002. EU Accession Monitoring Program 2002: Monitoring the EU Accession Process: Minority Protection. Open Society Institute, Budapest.

Pajnik, M., Bajt, V., 2012. Migrant women's transnationalism: family patterns and policies. International Migration 50 (5), 153–168.

Peček, M., Lesar, I., 2006. Pravičnost slovenske šole: mit ali realnost? Založba Sophia, Ljubljana.

Rorke, B., Wilkens, A. (Eds.), 2006. Roma Inclusion. Lessons Learned from OSI's Roma Programming. Open Society Institute, Budapest. http://www.soros.org/initiatives/roma/articles_publications/publications/inclusion_20060605 (accessed 31.01.13.).

Schuringa, L., 2005. Community Work and Roma Inclusion. Spolu international foundation, Utrecht.

Stewart, M., 1997. The Time of the Gypsies. Westview Press, Boulder.

Štrukelj, P., 2004. Tisočletne Podobe Nemirnih Nomadov: Zgodovina in Kultura Romov V Sloveniji. Družina, Ljubljana.

Szuhay, P., 1995. Constructing a Gypsy national culture. Budapest Review of Books 5 (3), 111–120.

Thompson, N., 2001. Anti-discriminatory Practice, third ed. Palgrave Macmillan, London.

Urh, Š.H., 2008. The development of an ethnically sensitive approach in social work in Slovenia. European Journal of Social Work 11 (2), 117–129.

Urh, Š.H., 2009. Etnično Občutljivo Socialno Delo Z Romi. Doctoral dissertation. Fakulteta za socialno delo, Ljubljana.

Urh, Š.H., 2011. Ethnic sensitivity: a challenge for social work. International Social Work 54 (4), 471–484.

Urh, Š.H., 2014. Anti-Roma racism in Europe: past and recent perspectives. In: Lavalette, M., Penketh, L. (Eds.), Race, Racism and Social Work: Contemporary Issues and Debates. Policy Press, Bristol, pp. 115–130.

Varsa, E., 2005. Class, ethnicity and gender – structures of differentiation in state socialist employment and welfare politics, 1960–1980. In: Shilde, K., Schulte, S.D. (Eds.), Need and Care – Glimpses into the Beginnings of Eastern Europe's Professional Welfare. Barbara Budrich Publishers, Bloomfield Hills, pp. 197–220.

Weckman, S., 1998. Researching Finnish Gypsies: advice from a Gypsy. In: Tong, D. (Ed.), Gypsies. An Interdisciplinary Reader. Garland Publishing, New York, pp. 3–10.

Zaviršek, D., 2001. Lost in public care: the ethnic rights of the ethnic minority children. In: Dominelli, L., Soydan, H., Lorenz, W. (Eds.), Beyond Racial Divides: Ethnicities in Social Work Practice. Ashgate, Aldershot, pp. 171–188.

Zaviršek, D., 2007. Pathologized ethnicities and meaningful internationalism. In: Zaviršek, D., Zorn, J., Rihter, L., Žnidarec Demšar, S. (Eds.), Ethnicity in Eastern Europe: A Challenge for Social Work Education. Faculty of Social Work, University of Ljubljana, Ljubljana, pp. 7–18.

Zaviršek, D., 2010. Ethnization and patologization of Roma people and Roma communities. Socialno Delo 49 (2–3), 85–98.

Zaviršek, D., Urh, Š., 2005. Rights of People with Intellectual Disabilities. Access to Education and Employment. Slovenia, Monitoring Report. Open Society Institute, Budapest.

Zaviršek, D., Zorn, J., Rihter, L., Žnidarec Demšar, S. (Eds.), 2007. Ethnicity in Eastern Europe: A Challenge for Social Work Education. Faculty of Social Work, University of Ljubljana, Ljubljana.

Zimmermann, M., 1999. The national socialist "solution of the Gypsy question". In: Herbert, U. (Ed.), National Socialist Extermination Policies. Contemporary German Perspectives and Controversies. Berghahn Books, New York, pp. 186–209.

Relevant Websites

http://www.errc.org/ – ERRC Reports (accessed 31.01.13.).

http://ec.europa.eu/justice/discrimination/roma/index_en.htm – EU and Roma (accessed 31.01.13).

http://www.opensocietyfoundations.org/reports/monitoring-eu-accession-process-minority-protection – OSI Reports (accessed 31.01.13.).

Rorty, Richard (1931–2007)

Maria Baghramian and Fergal McHugh, School of Philosophy, University College, Dublin, Ireland

Published by Elsevier Ltd.

Abstract

Richard Rorty (4 October 1931–8 June 2007) is a highly influential American philosopher who is as divisive as he is popular. Rorty's many writings touch on major themes within both analytic and continental philosophy. Issues in epistemology, philosophy of mind, philosophy of science, and naturalism featured centrally in his early work. His later writings placed increasing stress on questions of democracy and culture. However, Rorty's 'great theme' is philosophy itself, and his relentless interrogation of philosophy's place and purpose both inside and outside the academy is unmatched in contemporary philosophy. The present article provides an overview of central features of his thought under the headings of 'Anti-representationalism,' 'Pragmatism,' and 'The Scope and Role of Philosophy.' The examination is loosely chronological; however, since Rorty's thought demonstrates deep continuities, each heading presents an aspect of his thought rather than, exclusively, a period of his development.

Life

Richard Rorty was born in New York in 1931 to James Rorty and Winifried Rauschenbusch. Rorty was raised in New York City and Flatbrookville, New Jersey in a predominantly left wing but anti-Stalinist political environment. He entered the University of Chicago in 1946, starting at the age of 15 years where he studied with greatest of the logical positivists Rudolf Carnap and also Leo Straus. After completing his MA in 1952, he moved to Yale where he received a PhD in philosophy in 1956 with a dissertation on the *Concept of Potentiality* under the supervision of the metaphysician Paul Weiss. After a brief stint in the army signal corps, he taught at Wellesley College (1958–60) and Princeton University (1961–82). At Princeton, Rorty was immersed in the problems, texts, and method of analytic philosophy, which dominated the philosophical culture there. The fruit of that immersion was the groundbreaking *Philosophy and the Mirror of Nature* (1979), a work that was strongly critical of that tradition while attempting to supply impetus for its renewal. Rorty's next move, in 1982, was to the University of Virginia where his primary teaching role shifted from 'Professor of Philosophy' to 'Professor of Humanities.' The move to the University of Virginia both enabled and complemented his increasing interest in 'continental' philosophers such as Derrida, Foucault, and Habermas. *Contingency, Irony and Solidarity* (1989), a somewhat polemical work ranging across topics from epistemology to modernist literature, is an important product of Rorty's broadened horizons. Rorty remained at the University of Virginia until 1998 when he took up a chair in the Department of 'Comparative Literature' at Stanford University where he saw out his teaching career. He retired in 2005. Rorty's fourth volume of philosophical papers, *Philosophy as Cultural Politics* was published in 2007. He died in June of the same year from pancreatic cancer. In a short essay published posthumously in *Poetry* (November 2007), Rorty wrote of his regret at not having spent more of his life with poetry:

> Cultures with richer vocabularies are more fully human – farther removed from the beasts – than those with poorer ones, individual men and women are more fully human when their memories are amply stocked with verse.
>
> Rorty, 2010a: p. 521

This is a signature Rorty remark, wide-ranging, polemical, and yet precise, both heartfelt and ironic.

Work

The breadth of Rorty's writing is remarkable. The content of his numerous books and essays spans themes in epistemology and metaphysics, through a variety of literary forms, to ethics and politics. Rorty represents something of a puzzle for contemporary philosophy; he is an analytic philosopher by training as well as style of writing whose crucial interventions with respect to that tradition take the form of an attempt to close it down. His writings on pragmatism are central to a resurgence of interest in the classical pragmatists (William James and John Dewey) and the establishment of what has become known as neopragmatism; however, his interpretations are hotly disputed and criticized as selective and self-serving (e.g., Susan Haack, 1993: pp. 182–202). Rorty's interest in philosophy itself has resulted in him being called a 'philosopher's philosopher,' a surprising title given his popularity outside of philosophy, and indeed outside the academy (Hall, 1994: p. 3; Kraugerud and Ramberg, 2010: p. 49); he has also been called an 'anti-philosopher's philosopher' (Ramberg, 2009). It is tempting to regard Rorty as something of a contradiction; however, to do so, neglects the remarkable consistency and continuity of his thought. Rorty's writing, from his landmark introduction to the collection *The Linguistic Turn* (1967) to late essays such as 'Philosophy as Transitional Genre' (2004) enacts a deep preoccupation with metaphilosophical questions concerning the methods, the value, and ultimately the fate of philosophy.

Despite this 'inward' turn, Rorty's metaphilosophy rarely tended toward abstraction; his overriding metaphilosophical concern was the relevance of philosophy to our lives. It is, in fact, not possible to separate Rorty's metaphilosophical views from his substantive positions. Thus, the following chronological survey of his key positions in philosophy is an account of both the substantive and the metaphilosophical points at stake.

Antirepresentationalism

In *Philosophy and the Mirror of Nature* (1979), Rorty opposes himself to a conception of philosophy that thought of the "discipline as one which discussed perennial, eternal problems – problems which arise as soon as one reflects" (Rorty, 1980: p. 3). By contrast, *Mirror* is tasked with recounting the historical and social narratives both of how such problems were specified and why they have come to demand philosophy's attention. Certain 'ways of talking,' he argues, have become remote from the purposes for which they were developed, they have become obsolete, in certain circumstances perniciously so. Most notably, a specific 'picture' has come to dominate the way philosophers think about our relation to reality, a picture that is neither obligatory nor useful.

> The picture which holds traditional philosophy captive is that of the mind as a great mirror, containing various representations – some accurate, some not – and capable of being studied by pure, non-empirical methods.
>
> Rorty, 1980: p. 12

This is how Rorty begins a career-long attack on what he characterized as the prevailing 'representational paradigm' both illustrated and in part produced by the 'ocular metaphor' of 'mirroring.' The figure of the mirror articulates an understanding of knowledge as a relationship between a mind and the world in terms of representation. Philosophy builds on this understanding, claiming as its own a specialized domain of knowledge about representation, specifically about what counts as truly representing the world. Rorty's response is to contextualize the view, to show that the view has a history and to reject the idea that it is compulsory or natural. In a language influenced by Michel Foucault, Rorty seeks to show "epistemology-centered philosophy as an episode in the history of European culture" (Rorty, 1980: p. 390). In the account Rorty provides, 'representationalism,' a paradigm routed primarily in the Cartesian epistemic turn in philosophy, achieves its institutional apotheosis in the work of Kant. Kant combined the best insights of rationalist 'mirror talk' (encapsulated by Cartesian clear and distinct ideas) and empiricist 'mirror talk' (Lockean and Humean conceptions of ideas and impressions) into 'representations' that are actively synthesized through specialized conceptual capacities only describable at a 'transcendental level.' In Rorty's account, Kant's putatively stable blend of empirical realism and transcendental idealism carved out a specific domain of reflective activity that supported the requirements of the empirical sciences while simultaneously positing a transcendental domain immune to the methods of those sciences and requiring specialized investigation with specialized tools. The mind (or in Rorty's parlance the 'Mind') is a key invention of Kant's revolution in philosophical approach, securing a place for the study of 'mirroring' at a level inaccessible to empirical means. At the outset of his career, Rorty had advocated the disappearance theory of the mind (Rorty, 1965), a precursor of the influential 'eliminative materialism,' where mind talk was to be eliminated from a naturalist description of the world to be replaced by a scientific vocabulary. This approach was subsequently softened somewhat; however, the 'capital M' mind remained a key site of attack. Crucially, for Rorty the relationship between mind and world is not one of representing but 'coping': "[a]s long as we think of knowledge as representing reality rather than coping with it, mind or language will continue to seem numinous" (Rorty, 1982: p. 202). More recently, he countered the traditional privileging of mind with the deprivileged social account: "the mind is not a representational apparatus, but rather a set of norm-governed social practices" (Rorty, 2007: p. 158).

Rorty's criticisms are not merely leveled at the Kantian and post-Kantian traditions but also at contemporary analytic philosophy, particularly the 'linguistic turn,' which he sees as an "attempt to find 'successor subjects' to epistemology" (Rorty, 1980: p. 10). While for many, a gulf separates Kant's 'idealism' from Russell and Frege's 'descriptivism,' from Rorty's perspective all these views are underwritten by unquestioned allegiance to the representational paradigm.

> Without the notion of the mind as mirror, the notion of knowledge as accuracy of representation would not have suggested itself. Without this latter notion, the strategy common to Descartes and Kant – getting more accurate representations by inspecting, repairing, and polishing the mirror, so to speak – would not have made sense. Without this strategy in mind, recent claims that philosophy could consist of "conceptual analysis" or "explications of meanings" or examination of "the logic of our language" or of "the structure of the constituting activity of consciousness would not have made sense."
>
> Rorty, 1980: p. 12

Rorty divides the history of philosophy into two distinct camps of systematic versus edifying philosophers. The divide cuts across the more conventional division between analytic and continental philosophy. Hegel and Husserl are system builders as are a great majority of the empiricists, rationalist, and analytic philosophers. While, the later Wittgenstein, Heidegger, and Dewey are concerned with 'edification,' with creating a space within philosophy that enables us to grow and develop in direct opposition to the approach where philosophy seeks to provide "'grounding' for the intuitions and customs of the present" (Rorty, 1980: p. 12). Wittgenstein's therapeutic mode shows how philosophical problems force themselves on us through captivating pictures, but his approach 'needs to be supplemented by historical awareness'; Heidegger's historical approach enables us to "distance' ourselves from the tradition' to tell a historical story about how the 'captivating' picture emerges. However, neither Wittgenstein nor Heidegger pays enough attention to the 'social perspective' of the picture's emergence (Rorty, 1980: p. 12). Dewey provides that crucial 'social perspective' but also, as part of that perspective the idea of an aesthetic mode of being that fills out the idea of 'coping.' In Rorty's words, Dewey seeks a "culture no longer dominated by the idea of objective cognition but by that of aesthetic enhancement" (Rorty, 1980: p. 13). All three contribute to Rorty's call for a change in our overall understanding of philosophy.

The cultural-political themes and motivations that dominate Rorty's later writings are already a feature of *Mirror*. Rorty

claims: "If we see knowledge as a matter of conversation and of social practice, rather than as an attempt to mirror nature, we will not be likely to envisage a meta-practice which will be the critique of all possible forms of social practice" (Rorty, 1980: p. 171). The edifying philosophers help us to recast our relation with reality in terms of 'coping' rather than 'representations.' Coping is this practical activity that releases us from the constraining chimera of 'mirroring': "[w]e should not look for skyhooks, but only for toeholds" (Rorty, 1991b: pp. 13–14).

Pragmatism

Rorty, more than any other contemporary American philosopher, is responsible for the resurgence of what has become known as neopragmatism in contemporary American philosophy. Pragmatism, for Rorty, first and foremost, is a kind of antiessentialism with respect to philosophical categories such as truth, knowledge, and morality (Rorty, 1982: p. 162), and a counterforce to the foundationalist traditions of modern philosophy since Descartes. A second hallmark of pragmatism is its resistance to unhelpful dichotomies such as the

> epistemological difference between truth about what ought to be and truth abut what is, nor any metaphysical difference between facts and values, nor any methodological difference between morality and science.
> Rorty, 1982: p. 163

In this Rorty shares common grounds with his long-time philosophical interlocutor Hilary Putnam. A third feature of pragmatism is one that we now most closely associate with Rorty and is at the heart of Rorty's disagreement with not only many of the philosophers in the analytic tradition but also with Hilary Putnam:

> [Pragmatism is] the doctrine that there are no constraints on inquiry save conversational ones – no wholesale constraints derived from the nature of the objects, or of the mind, or of language, but only those retail constraints provided by the remarks of our fellow inquirers… The pragmatist tells us that it is useless to hope that objects will constrain us to believe the truth about them, if only they are approached with an unclouded mental eye, or a rigorous method, or a perspicuous language… The only sense in which we are constrained to truth is that, as Peirce suggested, we can make no sense of the notion that the view which can survive all objections might be false. But objections – conversational constraints – cannot be anticipated. There is no method for knowing when one has reached the truth, or when one is closer to it than before.
> Rorty, 1982: p. 165–166

In parallel to the picture of 'the mirror' pursued in *Philosophy and the Mirror of Nature*, Rorty's pragmatism attempts to release us from the grip of the picture of 'the world.' In a recent essay Rorty writes:

> Despite my veneration for Wilfrid Sellars, who originated this talk of manifest and scientific images, I would like to jettison these visual metaphors. We should not be held captive by the world-picture picture. We do not need a synoptic view of something called "the world." At most, we need a synoptic narrative of how we came to talk as we do. We should stop trying for a unified picture, and for a master vocabulary. We should confine ourselves to making sure that we are not burdened with obsolete ways of speaking, and then ensuring that those vocabularies that are still useful stay out of each other's way.
> Rorty, 2010b: p. 58

Pragmatism, as Rorty presents it, helps to free us from yet another pernicious dichotomy, that of a division between the world and the conceptual schemes that supposedly categorize it and render it intelligible. Following Donald Davidson (Davidson, 1974), Rorty finds the very division between a scheme and the world as something unfitted falling under it, not only incoherent but also pernicious both for those who accord them metaphysical status and for those who locate them in history (and any position in between).

> The notion of "the world" as used in a phrase like 'different conceptual schemes carve up the world differently' must be the notion of something completely unspecified and unspecifiable-the thing-in-itself, in fact. As soon as we start thinking of "the world" as atoms and the void, or sense data and awareness of them, or "stimuli" of a certain sort brought to bear upon organs of a certain sort, we have changed the name of the game. For we are now well within some particular theory about how the world is.
> Rorty, 1982: pp. 14–15

In his highly influential book *Contingency, Irony and Solidarity* (1989) and a number of papers to follow, Rorty is at pains to distinguish between the goal of giving meaning to our lives and thoughts with reference to a 'nonhuman reality' versus the alternative of envisaging their role in terms of contributions to a community. Realist philosophers, since Plato, have followed the first route and relaying on the correspondence theory of truth, a key component of the metaphor of the mind as a mirror, to achieve the ideal of objectivity with no reference to our historical situatedness. The alternative route is to seek solidarity with a community of enquirers, to reject the dualism of objective and subjective, knowledge versus mere opinion, and to eschew the talk of truth in other than in the sense of complimenting or commending the best beliefs and practices available to us at any given time (Rorty, 1985/1991) where the 'us' is a community of thoughtful, well-informed, and reasonable but historically located thinkers. The approach is best captured in the writings of the pragmatist philosophers' attempt to 'reduce objectivity to solidarity' (Rorty, 1991b: p. 21).

Rorty calls this approach 'ethnocentrism' and resists the charge, leveled at him by Hilary Putnam among many, that the abandonment of the ideal of mind-independent objectivity amounts to a form of relativism (Baghramian, 2004: pp. 144–151). He thinks relativism is no more than the claim that all beliefs are equally good and he finds the idea incoherent. Ethnocentrism, on the other hand, is the "purely negative point that we should drop the traditional distinction between knowledge and opinion, construed as the distinction between truth as correspondence to reality and truth as a commendatory term for well-justified beliefs" (Rorty, 1991b: pp. 23–24).

From the perspective of his pragmatism, dispensing with talk of 'truth,' 'knowledge,' and even 'the world' as

metaphysically inflated categories provides release from a series of endless and fruitless debates that are only pertinent if one's thinking is constrained by the metaphor of mirroring. Knowledge is belief justified within a social context, while truth is a compliment paid to the best-justified beliefs we have at any given moment, beliefs that 'we can defend against all comers.' We do not need to refer to a mind-independent world or objective accuracy of representations in our understanding of these and other cognate epistemic concepts. For Rorty, the very idea of 'the world' produces the illusion of a vantage from which distinctions between true by our lights and true simpliciter, but such world is 'well lost' (Rorty, 1982: p. 3), as we cannot make sense of the very idea of its existence outside of our linguistic and social practices. The pragmatist alternative allows us to construct a narrative of how we created and lost the world. The mode is ironic, since from the pragmatist vantage the world was always a story and the story is told for the purpose of edification. 'Solidarity,' however, is not a purely abstract philosophical notion, but plays a major role in Rorty's political outlook. As an unapologetic liberal, Rorty sees solidarity as the basis for obtaining incremental improvements in the scope of our freedoms and rights, and to help, by Rorty's lights, in the core liberal aims of preventing and reducing cruelty and decreasing intolerance. In politics, as in philosophy, the task is to be carried out through replacing confrontation by conversation within an ever-growing circle of peers. Philosophy, in its pragmatist form, is an aid to democratic politics, in so far as it encourages such conversation.

The approach results in one of the important intellectual advances of our century a 'steady decline in interest in this quarrel between Plato and Nietzsche about what we are really like' (Rorty, 1998: p. 169); there is now, Rorty thinks,

> a growing willingness to neglect the questions "What is our nature" and to substitute the question "What can we make of ourselves." We are much less inclined than our ancestors to take "theories of human nature" seriously, much less inclined to take ontology or history or ethology as a guide to life. We are much less inclined to pose the ontological question "What are we?" because we have come to see the main lesson of both history and anthropology is our extraordinarily malleability. We are coming to think of ourselves as the flexible, protean, self-shaping animal rather than as the rational animal or the cruel animal.
>
> Rorty, 1998: pp. 169–170

Here, too, Rorty remains the eliminativist philosopher *par excellence*. We not only should eliminate the reified mind talk and Truth talk from philosophy but also should stop thinking of ourselves in essentialist terms, as creatures endowed with a specific nature.

Rorty's version and interpretation of pragmatism has come under sustained attack. Susan Haack, for instance, an unflinching critic, thinks Rortyism is no more than vulgar pragmatism and accuses him in have misrepresented the work of classical pragmatists, James, Dewey, and Peirce (Haack, 1993: pp. 182–202). Defenders, on the other hand, have hailed him as crucial revivalist, a philosopher who has delivered previously unfashionable pragmatist thinking to new audiences, reinvigorating the pragmatist ideas in the process.

The Scope and Role of Philosophy

Much of Rorty's work, as indicated above, has a metaphilosophical tone. This was clear even in his very early publications, but it is not always an easy task to distill a unified account of philosophy from Rorty's writings on the topic. A few strands of thinking, however, are prominent. The *Linguistic Turn* (1992) signaled Rorty's nascent interest in a historic and contextualized understanding of philosophy. *Philosophy and the Mirror of Nature*, where Rorty puts the current "predicament of philosophy" (Rorty, 1980), as he sees it, in the historic context of a turn to epistemology, expresses this 'genealogical' understanding of philosophy even more completely.

> In this book I have offered a sort of prolegomenon to a history of epistemology-centred philosophy as an episode in the history of European culture… A proper historical treatment would require both learning and skills I do not possess. But I would hope the prolegomenon has been sufficient to let one see the contemporary issues in philosophy as events in a certain stage or a conversation – a conversation which once knew nothing of these issues and may know nothing of them again.
>
> Rorty, 1980: pp. 390–391

He also notes, "[a]lmost as soon as I began to study philosophy, I was impressed by the way in which philosophical problems appeared, disappeared, or changed shape as a result of new assumptions or vocabularies" (Rorty, 1980: p. xiii). The historical perspective makes us sensitive to a lack of fundamentals and the apparent mutability of philosophical problems themselves, themes that permeate Rorty's work.

A second prominent tendency is an invitation to modesty and even quietism in philosophy. Rorty had questioned the 'cash value' of philosophy even in his earliest writings. In an early essay 'The Philosopher as Expert' (1958–61/2009), unpublished in his lifetime but recently included in the 30th anniversary edition of *Mirror*, Rorty already demonstrates both the deep metaphilosophical tendencies that will inhabit his mature thought combined with a discomfort with the idea of philosophy as distinct set of jobs for which only the philosopher need apply. Thus, Rorty resists the assumption that "there is a body of special 'philosophical' truths, or the assumption that there is a special 'philosophical' technique that somehow works equally well on all problems, and is distinct from the ordinary, vulgar methods used to solve them" (Rorty, 2009: p. 421). In a new introduction written for the reissue of *The Linguistic Turn* (1992), Rorty remarked that in direct contrast to other types of inquiry such as science "what counts as philosophical knowledge seems itself to be a matter of opinion" (Rorty, 1992: p. 2). It is for Rorty an important irony that this lack of consensus has not prevented philosophers from believing that their discipline should critique and regulate what counts as knowledge within specific empirical domains. Philosophy, Rorty tells us, is not in a position to sort out its own affairs, never mind the affairs of other activities. There was, Rorty argues, a particular historical juncture at which Philosophy (with a capital P) served a 'useful purpose,' for example, "by suggesting ways of dealing with the triumph of mechanistic materialism" (Rorty, 2010b: p. 56) without recoil into some form of supernaturalism. That purpose is now achieved and the

attendant self-image of the discipline only possessed meaning within that task; that self-image should no longer be honored. Instead, Rorty encourages us to replace the vocabulary of discovery with that of creation; our task (or more precisely our tasks) *qua* philosophers is to create, to innovate, and to work for the better of the human conversation (and the humans who contribute to it) through inventiveness, playfulness, and brio. In *Contingency, Irony and Solidarity*, Rorty named this 'activity' irony. In its broader, institutional practiced form it is 'cultural politics'; 'Philosophy as Cultural Politics' is the title and theme of his last collection of papers.

Thirdly, Rorty's conception of traditional philosophy crucially exceeds institutional boundaries. Even in *Philosophy and the Mirror of Nature*, the term 'philosophy' is not restricted to analytic philosophy. "[T]he difference between 'analytic' and other sorts of philosophy is relatively unimportant – a matter of style and tradition rather than a difference of 'method' or of first principles" (Rorty, 1980: p. 8). However almost from the beginning, Rorty seems to express a preference for 'continental philosophy,' specifically those types of continental philosophy that combine social and cultural critique with a deep suspicion vis-à-vis the prior Western philosophical tradition. A feature of the kind of continental philosophy Rorty likes is its embrace (at least in his reading) of contingency as well as the edifying form of philosophizing. But his writing, as well as his academic career, also crosses the boundaries between philosophy and literature. He also wishes to transgress the 'purity of academic disciplines' by inviting philosophers to become participants in debates on moral and political topics that shape our lives and our times. As Rorty remarks in the introduction to his fourth volume of papers

> The more philosophy interacts with other human activities – not just natural science, but art, literature, religion and politics as well – the more relevant to cultural politics it becomes, and thus the more useful. The more it strives for autonomy, the less attention it deserves.
>
> Rorty, 2007: p. ix–x

Intellectuals, Rorty thinks, have a duty "to function as citizens of a democracy," "to use the mechanisms of democratic government to help prevent the rich from ripping off the poor, the strong from trampling on the weak" (Rorty, 1991a: p. 490), and philosophy should not turn its back on these tasks, and yet he points out that in "political deliberation, philosophy is a good servant but a bad master. If one knows what one wants and has some hope of getting it, philosophy can be useful in formulating redescriptions of social phenomena" (Rorty, 1999: p. 232).

Legacy

The impact of Rorty's work can be felt far beyond the boundaries of what is typically considered academic philosophy; however, any appreciation of this influence must begin with an assessment of his philosophical legacy. Rorty's legacy can be divided with respect to what commentators take to be his most significant work, *Philosophy and the Mirror of Nature* or *Contingency, Irony and Solidarity*. Those philosophers who take themselves to inherit from the Rorty of *Mirror*, broadly speaking analytic philosophers, are at once critical and admiring. In many instances, their aims are to rehabilitate a Rorty who in numerous ways and on a host of issues simply went too far; their task is to retain the best of his insights while calming or abandoning his rhetoric. For example, John McDowell shares Rorty's belief that the philosophical problems posed within the Cartesian and British empiricist traditions are to be addressed therapeutically rather than to be straightforwardly answered or solved, but he thinks Rorty goes too far in denying the idea of an answerability to the way things are that goes beyond our communal say so (McDowell, 2000: p. 110). Other contemporary analytic philosophers who inherit from Rorty include Robert Brandon, Michael Williams, and Huw Price. Those who read *Contingency, Irony and Solidarity* as Rorty's most important work, broadly speaking continental philosophers, tend to be more focused on the possibilities his work supplies for the pursuit of a postphilosophical mode of philosophy. They include Nancy Fraser who has further developed Rorty's concept of 'private irony' (Fraser, 1988, 1991) as part of her brand of feminist critical theory. Equally, Rorty's impact on the social sciences both extends from that work and is at its strongest where the social sciences and philosophy are at their most enmeshed; those social scientists who take Louis Althusser, Michel Foucault, Jacques Derrida, Jean Baudrillard, and Alain Badiou seriously are likely to be interested in what Rorty has to say concerning the active creation of vocabularies, the resistance to systematization, and the embrace of contingency in a mode of critical historicism. In an early essay 'Method, Social Science and Hope' (1981/1982), Rorty argued that the social sciences should not occupy themselves with scientific description at the expense of the normative, nor should they attempt to artificially separate these strands. Rather, they should concern themselves with being useful. Rorty recommends a "Deweyan approach to both social science and morality, one which emphasizes the utility of narratives and vocabularies rather than the objectivity of laws and theories" (Rorty, 1982: p. 195). The social scientist, thinks Rorty, should be in the business of telling good stories that make our lives better.

Appropriately, for a polemical philosopher such as Rorty his relation to key thinkers from the continental tradition is both complex and contested. For example, his reading of Derrida as a poet rather than a philosopher is, by some lights, typical of Rorty's shallow 'way' with philosophical texts (e.g., Bouveresse, 2000: pp. 133–134). Also, despite his leftward leaning, Rorty's ethnocentrism sits uncomfortably with those who aim to undo what they see as Enlightenment-inspired structures of domination masked behind a rhetoric of emancipation (e.g., Haber, 1994: pp. 43–72). For such critics Rorty is the effective steward of old-Europe enlightenment wolves in 'theory' laden sheep's clothing.

Finally, we might say that Rorty himself has, appropriately, attracted attention as an intellectual phenomenon that requires a sociohistorical account. In an innovative work *Richard Rorty: The Making of an American Philosopher* (2008), Neil Gross, a sociologist, has attempted to turn Rorty's ceaseless demand for context back on Rorty himself. In many ways, Gross's engagement with Rorty is exemplar of Rorty's influence on the social sciences and his lessons for the social sciences that there

is no such thing as a value-free, ahistorical mode of inquiry, that the question of context must always be addressed, and that the story of how we reached our 'present' must be told.

See also: Continental Philosophy of Language; Dewey, John (1859–1952); Kant, Immanuel (1724–1804); Law and Morality: A Continental European Perspective; Law and Morality: An Analytical Perspective; Liberalism: Political Doctrine and Impact on Social Science; Naturalism; Pragmatism: Philosophical Aspects.

Bibliography

Baghramian, M., 2004. Relativism. Routledge, London/New York.

Bouveresse, J., 2000. Reading Rorty: pragmatism and its consequences. In: Brandom, R. (Ed.), Rorty and His Critics. Blackwell, Oxford.

Davidson, D., 1974. On the very idea of a conceptual scheme. Proceedings and Addresses of the American Philosophical Association 47, 5–20.

Fraser, N., October 1988. Solidarity or singularity? Richard Rorty between romanticism and technocracy. Praxis 8, 257–272.

Fraser, N., Spring 1991. From irony to prophecy to politics: a response to Richard Rorty. Michigan Quarterly Review 30, 259–266.

Gross, N., 2008. Richard Rorty: The Making of an American Philosopher. University of Chicago Press, Chicago.

Haack, S., 1993. Evidence and Inquiry. Blackwell, Oxford.

Haber, H.F., 1994. Beyond Postmodern Politics: Lyotard, Rorty, Foucault. Routledge, New York.

Hall, D., 1994. Richard Rorty: Prophet and Poet of the New Pragmatism. SUNY Press, Albany.

Kraugerud, H.A., Ramberg, B.T., 2010. The new loud: Richard Rorty, quietist? Common Knowledge 16 (1), 48–65. http://dx.doi.org/10.1215/0961754X-2009-060.

McDowell, J., 2000. Towards rehabilitating objectivity. In: Brandom, R. (Ed.), Rorty and His Critics. Blackwell, Oxford.

Ramberg, B., 2009. Richard Rorty (retrieved 11.04.14.). http://plato.stanford.edu/archives/spr2009/entries/rorty/%3E.

Rorty, R., 1965. Mind-Body Identity, Privacy and Categories. Springer Verlag, New York.

Rorty, R., 1980. Philosophy and the Mirror of Nature. Blackwell, Oxford.

Rorty, R., 1982. Consequences of Pragmatism. University of Minnesota Press, Minneapolis.

Rorty, R., 1991a. Intellectuals in politics. Dissent 38 (Fall), 483–490.

Rorty, R., 1991b. Objectivity, Relativism, and Truth: Philosophical Papers, vol. 1. Cambridge University Press, Cambridge.

Rorty, R., 1992. Introduction to the linguistic turn. In: Rorty, R. (Ed.), The Linguistic Turn, second ed. University of Chicago Press, Chicago.

Rorty, R., 1998. Truth and Progress: Philosophical Papers, vol. 3. Cambridge University Press, Cambridge.

Rorty, R., 1999. Philosophy and Social Hope. Penguin, London.

Rorty, R., 2007. Philosophy as Cultural Politics: Philosophical Papers, vol. 4. Cambridge University Press, Cambridge.

Rorty, R., 2009. The Philosopher as Expert Philosophy and the Mirror of Nature. Princeton University Press, Princeton, NJ.

Rorty, R., 2010a. The fire of life. In: Bernstein, R.J., Voparil, C.J. (Eds.), The Rorty Reader. Wiley-Blackwell, Malden, MA.

Rorty, R., 2010b. Naturalism and quietism. In: De Caro, M., Macarthur, D. (Eds.), Naturalism and Normativity. Columbia University Press, New York.

Rousseau, Jean-Jacques (1712–78)

Jürgen Oelkers, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Jean-Jacques Rousseau, 'citoyen de Genève,' lived a life of constant unrest. Having been part of the Parisian circles of Enlightenment for a short time, he became the successful outsider who wrote for a European public. In 1761–62, he published the three essential works that have continued to influence literature, philosophy, economy, sociology, and education into the twenty-first century. Rousseau delivered provocative ideas, which contradicted the mainstream. His thinking is paradoxical in a very attractive way. In education the basic concept of 'natural' and 'negative education' was considered to be the theory putting not religion or society but the child and its development in the center of respect. Rousseau influenced almost every reform movement in education, as well as eminent writers such as Jean Piaget and Maria Montessori. The progressive education of the twentieth century confirmed Rousseau's child centeredness and made it popular, although Rousseau, in his time, had no real impact on professional education and the development of schooling. His ideas were used by those interested in alternative education and society, and this was made possible by the strict dualism in Rousseau's thinking. Nature stands against society, and natural education was the opposite of artificial teaching, which used books instead of experience. Rousseau's language is at the same time elegant and sentimental, and together with his ideas explains his lasting influence.

Life and Work

Rousseau was born on 28 June 1712 and raised in Geneva, the center of Calvinism. Jean Calvin, not a native Genevan, published his magistral *Christinae Religionis Institutio*, the central work on education, in 1536. Its final version, dated 1559, justified general priesthood by way of the doctrine of predestination. Calvin dissolves the 'two empires' of laymen and ecclesiastical office in favor of a parish without differences, one which saw itself as an elected body (Calvin, 1957: III/21). Clergymen are chosen teachers whose sermon is intended to *educate* (Calvin, 1958: IV/1, 5) without the possibility of it being associated with primacy or hierarchy (Calvin, 1958: IV/74). Rousseau's life and work is, in many ways, a manner of dealing with Calvinism. The concept of *equality* determines the Social Contract, the theory of education and Rousseau's natural theology. Calvin's doctrine of grace is translated into the predestination of nature, along with a rigorous inwardness that has the only true form of education on its side. Virtue becomes a model for society given the education of man's nature. Calvinistic rigor fascinated Rousseau, following the banning and public burning of his books, he dispensed with the citizenship of the Republic of Geneva in 1763, but not, however, his membership in Calvin's church that he had renewed in 1754.

Rousseau's dramatic life (Cranston, 1991–97) (documents in *Le CD-Rom Jean-Jacques Rousseau*, 1999; Damrosch, 2005) is characterized by ascent outside the elite and by being an outsider. The young Rousseau completed an apprenticeship as an engraver, in view of education he was largely an autodidact. His wavering career run between Geneva and Paris, later included Switzerland and England, independently and restlessly all the time. Rousseau himself had attempted to gain control over his biography with the writing of his *Confessions* (Rousseau, 1961: pp. 1–656) that was published posthumously in 1782 and indeed determined the image of Rousseau since then. After various wanderings and low-level activities, Rousseau took up a position as a private tutor in Lyon in 1740 and wrote his first text on education (*Mémoire à Mr de Mably*: Rousseau, 1969: pp. 1–32). In 1742 Rousseau went to Paris, publishing his *Dissertation sur la musique moderne* in 1743. In 1749 Rousseau wrote the articles on music for Diderot's *Encyclopédie*, his first major literary success came one year later with the *Discours sur les sciences et les art*, which was awarded a prize by the Academy of Dijon. From 1756 on Rousseau, living now in the *hermitage* of Montmorency, concentrated on his three major works: the *Nouvelle Héloise* (1761), throughout Europe a successful, widely read novel; the *Contrat social* (1762), Rousseau's theory of society; and *Emile ou de l'éducation* (1762), the theory of 'new education.' Both books, the one about the social contract and that on education, were banned directly after publication and publicly burned in Paris and Geneva.

Since then Rousseau was on the run. He was granted asylum by the Prussian governor in Neuchâtel soon after his fleeing from Paris, at his place of refuge in Môtiers he was subjected to attacks by such differering authors as the general procurator of Geneva, Trochin, and his arch enemy Voltaire ('Sentiment des citoyens,' Voltaire, 1961: pp. 715–718). At the same time Rousseau became the idol of the generation of 'Sturm und Drang,' young *literati* from throughout Europe visited Rousseau in Môtiers and made him famous. In January 1766, after having passed more places on the run, Rousseau made for England following an invitation by David Hume. Both had wrong hopes and soon their relationship broke into pieces. Rousseau returned to France and worked, with increasing paranoia, on his autobiography which can be seen as an attempt at self-therapy. His precarious financial situation was revived at short notice in May 1778 when Rousseau and Therèse Levasseur, whom he had married in 1768 following a long-standing concubinage, moved to Ermenonville into an estate owned by Marquis de Girardin, Rousseaus's last benefactor. The sudden death on 2 July 1778 ended a very unlikely literary and

philosophical career, which was intended to be *à rebour* – contrary to convention.

Education and Society

The archbishop of Paris, Christophe de Beaumont, justified the banning and burning of Rousseau's *Emile* above all with the denial of the original sin. In fact Rousseau assumes that the nature of man is not tainted. "There is no original sin in the heart of man" (Rousseau, 1969: S. 322). In his state of defense against Beaumont, printed at the beginning of 1763, Rousseau formulated a far-reaching dualism, which since then has dominated political and educational conceptions. The two sides of dualism are *nature* and *society*; both are treated to be contradictious. The present social order contradicts human nature in every respect, society is the alienation of nature, and this explains the vices of men and the evils of life. The assumption of the original sin then is superfluous, man could live without sin if nature and society correspond and harmonize (Rousseau, 1969: pp. 966 et seq.). In this respect Rousseau had a post-Augustinian concept of education in mind which is developed in *Emile*.

Essentially, the theory has three axes: the political difference of *homme civil* and *homme naturel*, the assumption of phases of 'natural development,' and the anthropological difference of *amour de soi* and *amour propre*. Rousseau's concept of education is constructed using these axes. This concept is peculiar *not* because it stresses 'natural education,' a term that was established in educational discourses long before Rousseau (Mercier, 1961); the peculiarity arises, moreover, from the paradoxical attempt to solve the contradiction between nature and society *by way of education*. To this end Rousseau, already in the draft of *Emile*, the *Manuscrit Favre*, draws a distinction between two contrary types of education, that of *nature* and that of *society* (Rousseau, 1969: p. 58). They refer to two ways of life, the life of natural and of civil man both described contrary to Hobbes, who, in his *Leviathan* (1651), distinguished between *natural* and *social condition*. Society, for Hobbes, tames and cultivates nature, thus repression of nature is unavoidable. Rousseau reversed this argument, for him the social condition is that of ongoing civil war (bellum omnium contra omnes) while the natural condition is considered to be of presocial peaceableness. The education for the 'publique business' by Hobbes (1968: S. 289 et seq.) is completely ignored.

Rousseau justified his fundamental thesis in the second *Discours* dated 1765, which was subjected to the evolution of inequality (Rousseau, 1964: S. 109–194). Rousseau explains *inegalité* in the form of society's corruption of human nature, which originally was in a position to live harmoniously with man's own needs. What is called *homme sauvage* (savage man) is considered to exist on the basis of his own strengths, while the *homme civilisé* (civil man) develops social needs, which render him dependent and weaken his nature (Rousseau, 1964: pp. 135 et seq.). The division of labor, the development of knowledge, and the connected social differentiation (Rousseau, 1964: pp. 143 et seq.) forcefully brought about a progressive inequality among men. Leave is taken of the condition of innocence and natural freedom (Rousseau, 1964: p. 152), the invention of the *first* society irresistibly causes all those following (Rousseau, 1964: p. 178), for the prize of alienation of mankind from its own nature. In this respect *society* is the fall of men, not nature.

The theory of education in *Emile* picks up the evolution thesis and processes it with regard to the development of the child. This forces a reversal: the social condition does not follow the natural condition; moreover, the social condition is led back to the natural condition given that education can develop nature on the basis of its own potentials. Rousseau's risky experiment of thought goes as follows: If children are seen as the *homme sauvage*, education would have to apply to this condition entirely, provided that the social condition can be excluded for the duration of education. This assumption leads to *Emile's* basic scenario: an 'élêve imaginaire,' described in greater detail merely as an 'orphan,' is exclusively educated by a 'governor' outside any society. The scene is an anonymous countryside far away from corrupt cities and raw villages, i.e., independent from the social condition as Rousseau viewed it (Rousseau, 1969: S. 264, 267, 279). The place in which the story is told is described simply as 'au milieu des champs' (in the middle of fields) (Rousseau, 1969: p. 277) without any details regarding origin and personal history of the two protagonists. Insofar, the tale is not a novel of education but a *treatise* that is intended to describe the paradigm of true education. The *name* 'Emile' is presumably reminiscent of Plutarch (Shanks, 1927), it has no biographical meaning. Accordingly the governor is not given a name and is thus not distinguishable. Both are paradigms, not persons.

'Natural' is *negative* education. Rousseau largely draws on the phases of human development from the *Histoire Naturelle* by Buffon (1989: pp. 1–114). Again not the schema of 'development,' formulated at the beginning of the *Manuscrit Favre* (Rousseau, 1969: p. 60), is original but the linkage with the axes of educational theory. Rousseau understands the entire childhood as the 'age of nature,' which must take place outside the social condition. Thus the entire first education is *negative*. It consists not at all in teaching virtue or truth, but in avoiding vices and errors to grant the innocence of heart and the integrity of mind (Rousseau, 1969: p. 323). The concept of negative education is aimed against Locke (Rousseau, 1969: pp. 317 et seq.) whose *Thoughts Concerning Education* had been available in French in a translation by Pierre Coste since 1695 and had considerable influence. Locke's (1989: p. 142) central idea, "Children are to be treated as rational creatures," aroused a passionate contradiction in Rousseau. For him "*Reasoning* with children" (Locke, 1989: p. 142) was nothing other than the reversal of nature's plan. Reason (raison) is the aim of education, but the end cannot determine the beginning. "If children could understand reason there would be no need to educate them" (Rousseau, 1969: p. 317). Childhood must have its *own* manner (manière) of seeing, thinking, and feeling, which is not that of the adults (Rousseau, 1969: p. 319). Locke's concept of reason is that of adults outside the world of children.

It is above all this idea, the own world of childhood, that had justified Rousseau's fame in education, along with the paradisical metaphor of pure education outside society and the spontaneity of children that are considered to be led by immediate interests (intérests présents) (Rousseau, 1969: p. 358). Rousseau seems to be the creator of modern education,

which stresses the 'new' image of child and childhood. In this respect it is often overlooked that *Emile* clearly has a theological center, that the treatise ends by giving preference to republic, not to nature, and that the whole theory presents *two* concepts of education with differing gender forms. The center of *Emile* is the concept of 'natural religion,' justified in the *Profession de foi du Vicaire Savoyard*, a treatise within the treatise (Rousseau, 1969: pp. 565–635). Rousseau takes the role of the vicar who against the materialism of the Parisian philosophy announces a supreme being or an active creator who is assumed to be "King of Earth" (Rousseau, 1969: p. 382). Being in unison with creation cannot mean anything sinister: "Where everything is good nothing can be injust" (Rousseau, 1969: p. 588). The good can be seen in creation, justice eminates from the good, only a just person can live happily (Rousseau, 1969: p. 589), but the enthusiasm for virtue (Rousseau, 1969: p. 596) is an *inner* principle (Rousseau, 1969: p. 598), that of the heart and not of reason (Rousseau, 1969: p. 627). Consequently, the key to the profession of faith is the following: "We can be men without being scholars" (Rousseau, 1969: p. 601).

What Rousseau called "natural religion" (Rousseau, 1969: p. 607) is a *theology of the heart* (Osterwalder, 1995; cf Chevallier, 1994), it does not require any positive dogma, moreover inner enlightenment, that needs no teachers (pasteurs) (Rousseau, 1969: p. 609) but the inspiration of belief in seeing God's creation. Above all, Rousseau radicalizes the Calvinistic doctrine here, the metaphor of 'heart' signals independence even from the community of belief. 'Equality' is therefore not a social construction, which would require comparison with others but a sovereignty in solitude. 'Sovereignty' or *inner strength* presupposes identity *and* distance. Rousseau justified this difference with the relation between two instances of the soul called *amour de soi* and *amour propre*. The fact that *two* instances are mentioned violates again a convention of literature. In French virtue literature like La Rochefoucauld's *Maximes* (1665) or Blaise Pascal's *Pensées* (1670) 'amour propre' is the formula for egoism and self-love that has to be restricted by education. Rousseau used the formula in a new relation: He took over the negative meaning of *amour propre* but added a second positive instance, that of *amour de soi*. The positive instance grants that man looks only for himself and is content with the satisfaction of his own true needs. The negative instance, *amour propre*, leads to comparison with others and is never satisfied because nobody will give preference to others more than to himself (Rousseau, 1969: p. 493).

Loving and tender passions occur in *amour de soi*, hate and temper occur in *amour propre*, i.e., from comparison with others. Thus, man is all the more happier with the fewer needs he has and the more he can avoid comparisons with others; in contrast, the things which make man suffer ills and evils (méchants) are *too many* needs and *too ma*ny opinions, i.e., the dependency on others. Consequently, Rousseau develops the concept of 'éducation solitaire' (Rousseau, 1969: p. 341), which isolates the fictious Emile throughout his childhood, totally controls learning (la liberté bien réglée: Rousseau, 1969: p. 321) and stipulates the didactic scenario on the theory of natural requirements and needs. Therefore Emile does not receive lessons, is not provided with any written works and is excluded from all forms of cultural education. Only the *age de raison* – youth – requires formal lessons, the entire childhood should apply exclusively to nature which, of course, is only experienced in a *didactic* manner, i.e., extremely artificial. The governor (and he alone) arranges the entire experience, and it is not by chance that the island of Robinson Crusoe is the only literary evidence that Emile is given to read (Rousseau, 1969: pp. 455 et seq.).

The concept of *éducation solitaire* refers to the education of man (*faire un homme*: Rousseau, 1969: p. 248). However, in the beginning of the fifth book Rousseau introduces 'Sophie ou la femme,' i.e., he is forced to react to the difference in sex. The education of 'the woman' is given an equally generalized treatment as was previously the case with 'the child.' But now it is recognised as a *male* child and not as a sexless 'person' in general. Rousseau's *homme sauvage* is stated as male, namely 'strong,' 'sovereign,' and 'independent.' Sophie, the generalized women, is educated in a manner that she can see herself as being in complementary dependence to males. The education of sexes is therefore not equal, it is dissimilar (Rousseau, 1969: pp. 700 et seq.). This is subject to the precondition of sexual difference that cannot be dissolved or harmonized (Rousseau, 1969: p. 697). Because men want to seduce, they are dependent in the crucial moment; women have to agree and thus can say 'yes' or 'no.' This female strength can only be compensated by binding the woman to the house and forcing her publicly to virtuousness. The 'mutual dependency' is therefore not symmetric: men are dependent on women because of their desires (désirs), women are dependent on men as a result of their wishes *and* needs (besoins). Accordingly, "by the law of nature itself women are at the mercy of men" (Rousseau, 1969: p. 702), not vice versa.

Rousseau's treatise about true education ends with a grand tour (Rousseau, 1969: pp. 826–855), which is intended to introduce Emile to basics of government theory (Rousseau, 1969: p. 833), i.e., the doctrines of Rousseus's own *Contrat social*. Endeavors are made regarding "le dernière perfection" (Rousseau, 1969: p. 799), it is envisaged that the well-educated *man* now is to become a respectable *citizen*. The citizen's know-how is taught with the theory of the "social contract" (Rousseau, 1969: pp. 839 et seq.). Emile has to learn what defines the status of citizens (citoyens) and thus the constitution of society (Rousseau, 1969: p. 840). The small *republic* is given preference, one which can create an ideal relationship between population and government (Rousseau, 1969: p. 840). Obviously the Republic of Geneva, idealized by Rousseau during his lifetime, is the paragon for this thesis. The ideal Republic is seldom and thus hard to find. So the governor concludes his teachings on government as follows: The obligations of the citizen presuppose a true country (patrie), a country is true insofar it can be considered as a model of political virtue. Those who do not find their 'patries' will not face political abysm, "car qui n'a pas une patrie a du moins un pays" (Rousseau, 1969: p. 858).

The connection of education to society fails. Rousseau had two views on society, first the ideal of the 'contrat social' prevailing a model for social order that is mainly granted by the concept of *volonté génerale*, the general will behind all singular wills (Rousseau, 1964: pp. 361 et seq.), second the polemical description of general decadence within the existing society. In no place does society comply with the ideal, therefore the grand tour ends without result. The couple Emile and Sophie,

educated differently, are not released into society *as citizens* but led to marriage and family (Rousseau, 1969: p. 867). More is not possible, the ideal society inspired by the volonté générale and thus safe from the corruption of power is not realized anywhere. The social ideal has no social place. The decadent society, on the other hand, is morally not acceptable, the real society has its place but not a legitimate form. Thus, loneliness (solitude) is in the end stronger than sociality (Starobinski, 1971), because education will not change society but can only develop man's nature. The great project of 'humanization' through natural education (as stated by Ravier, 1941) fails due to its own ambitions and contradictions. 'Nature' in the framework of Rousseau's *Emile* is an artificial experience, learning is bound by didactic parameters, control is total, sentiment for the children's *own* world arises from the fiction of a good and equal nature that does not mirror any child's individuality (Oelkers, 2008).

Reception and Enduring Influence

The adoption of Rousseau's work demonstrates the often dramatic and always radical conflict between convinced supporters and equally convinced adversaries. Rousseau divides his readers until today, not the least because he wrote against reality. His ideals of antiquity, especially those of Sparta and Athens, suggest an *âge d'or* (Terrasse, 1970; Leduc-Fayette, 1974), which can be understood as an anticipation of the future, the renovation of the 'true society' that has been lost in history. Rousseau is therefore not by chance the hero of the Jacobins, his cult was established in the French Revolution (Barny, 1986) against all conservative theories who negate revolution in favor of long-term development of society without the sentiment of decadence. Rousseau's Social Contract represents the *new* society that conservatives can only deny. The tension between freedom and equality attributed to, and paradoxically and provocatively described by Rousseau, characterizes one major part of political theory up to John Rawls' reformulation of the *Contrat Social*. The same applies in similar fashion to the theory of 'natural education': Rousseau's *Emile* is a key source for Tolstoy and the reform movements of the nineteenth century, a central inspiration for Piaget (Oelkers, 1996) and the development of child psychology, and a milestone for *progressive education*.

Of central importance are the dualisms and thus Rousseau's language. He stresses contradictions and paradoxes between nature and society, men and citizens, children and adults, and not at least male and female. Even John Dewey, who refused dualisms, had to claim that Rousseau – and only he – was the founder of the theory of 'natural development' (Dewey, 1985: pp. 211 et seq.) as opposed to artificial schooling and thus read in a dualistic manner. Educational concepts are irritated by Rousseau, who is regarded as a 'classical writer' of education (Oksenberg Rorty, 1998), *because* of the unsolved paradoxes. The concept of education makes no attempt of solving the paradoxes and dilemmas, but left them open thus puzzling readers with keen contradictions, which are right and wrong at the same time. What counts is not logic but the heart, and this seems to be part of the success. Radical conclusions seem to be obvious, politics within a new society and to make this possible the education of 'new man' according to his own true nature. The famous ascription 'return to nature,' being part of nineteenth century's readings of Rousseau and nowhere stated by him directly, was regarded as a emancipation from alienation, thus a project of the left. But Rousseau was at the same time radical *and* conservative, society should return to the golden age, education should leave schooling for true nature, man should be first man and then citizen, it is this rigor without reality that fascinated Rousseau's readers, followers as well as opponents.

The theory of education is provocatively puzzling: The 'negative education' has no objective, nature develops itself. However, the learning process is subject to extreme regulations (Oksenberg Rorty, 1998: p. 248). The present should not be made a victim of the future (Rousseau, 1969: p. 309), but every education is a deal with future, and this is true for Rousseau, too. The ages of childhood of youth are clearly defined, so are the phases of education, nature is not chance but necessity, aiming at the future. Children should be educated solely dependent on *things* (Rousseau, 1969: p. 311), nevertheless the governor dominates the whole education. Nature should lead the way (Rousseau, 1969: p. 259), but everything possible must be taken to avoid the first wrong step (Rousseau, 1969: p. 317). Children stand like savages outside the law and are completely natural (Rousseau, 1969: p. 320), but education must undertake everything so as to *exclude* chance (Rousseau, 1969: p. 342) without being able to act *solely* to the necessity of nature. The first education should be that of the *senses* alone (Rousseau, 1969: p. 380), but that requires a rational plan of education that cannot simply be drawn from nature. And in consequence, Rousseau's educational scene is one of extreme regulations: Emile does not play, he does not develop any form of his own, and is not allowed to listen to music, his learning differs in every way to 'amusement' (Rousseau, 1969: p. 407), education should be 'realistic' (Rousseau, 1969: p. 418), and this is possible only in an extremely artificial manner. Because society is excluded educational rigorism can rule. The basis for this is *nature* in the sense of *self-love* (amour de soi). Self-love is always good and always in compliance with the natural order (Rousseau, 1969: p. 491), which begs the question why education does not concentrate fully on this *amour de soi*.

Rousseau wrote the counterproject to sensualistic educational theory, which dominated the pedagogy of the eighteenth century. Children, according to Rousseau, should not be viewed as empty vessels or *tabula rasa*, but instead as part of nature, which develops of its own accord. It cannot simply be influenced, education is not merely the establishment of habits and customs, moreover it is the child's nature that limits all education. This fundamental concept is weakened by the *implicit* sensualism, the education of the senses (Rousseau, 1969: pp. 415 et seq.), that is necessary, because education is inconceivable *without any* influence. But the provocation remains, and it defines Rousseau's standing as an educational author. Education is limited by nature, nature has nothing to do with sin, the child is innocent because of nature's original goodness, thus education can take place without any burdens of history and society. It is a renewal of mankind with every new child. This is the peak or the basic point of Rousseau's theory, which is read up to now because it provocates *and* stimulates educational thought, without any attempt to solve

the puzzles. Rousseau is read because he defines the problems, not the solutions.

See also: Democracy, History of; Education, History of; Educational Philosophy; Enlightenment; Evolution, Natural and Social: Philosophical Aspects; French Revolution, The; Inequality, Social; Nature-Nurture Controversy, History of; Religion: Morality and Social Control.

Bibliography

Barny, R., 1986. Rousseau dans la Révolution: le personnage de Jean-Jacques et les débuts du culte révolutionnaire (1787–1791). The Voltaire Foundation at the Taylor Institution, Oxford (Studies on Voltaire and the Eighteenth Century, Mason, H.T. (Ed.), vol. 246).

Buffon, G.L., Leclerc, Comte de, 1989. Histoire naturelle de l'homme et des animaux. Jean de Bonnot, Paris.

Calvin, J., 1957. Institution de la Religion chrétienne. Livre troisième. Édition nouvelle publiée par La Société Calviniste de France. Labor et Fides, Genève.

Calvin, J., 1958. Institution de la Religion chrétienne. Livre quatrième. Édition nouvelle publiée par La Société Calviniste de France. Labor et Fides, Genève.

Chevallier, M., 1994. Pierre Poiret (1646–1719). Du protestantisme à la mystique. Labor et Fides, Genève.

Cranston, M., 1991. Jean-Jacques. The Early Life and Work of Jean-Jacques Rousseau 1712–1754. University of Chicago Press, Chicago.

Damrosch, L., 2005. Jean-Jacques Rousseau: Restless Genius. Houghton, Mifflin, Harcourt, New York.

Dewey, J., 1985. The middle works, 1899–1924. In: Baydston, J.A., introduction by Hook, S. (Eds.), Essays on Education and Politics 1915, vol. 8. Southern Illinois University Press, Carbondale/Edwardsville.

Hobbes, Th, 1968. In: Introduction by MacPherson, C.B. (Ed.), Leviathan. Penguin Books, Hamondsworth/Middlesex.

Hobbes, Th, 1991. The Noble Savage. Jean-Jacques Rousseau 1754–1762. The University of Chicago Press, Chicago.

Hobbes, Th, 1997. The Solitary Self. Jean-Jacques Rousseau in Exile and Adversity. With a Foreword by S. Lakoff. The University of Chicago Press, Chicago.

Le CD-Rom Jean-Jacques Rousseau, 1999. Prod. par INCOPROM SA (Genève), L'INSTITUT de FRANCE (Paris) et CYBELE PRODUCTIONS (La Muraz). Concept général par C. Richardet.

Leduc-Fayette, D., 1974. Jean-Jacques Rousseau et le mythe de l'antiquité. J. Vrin, Paris.

Locke, J., 1989. In: Yolton, J.W., Yolton, J.S. (Eds.), Some Thoughts Concerning Education. Clarendon Press, Oxford.

Mercier, R., 1961. L'enfant dans la société du XVIIIe siècle (before Emile). Thèse complémentaire pour le Doctorat ès lettres présentée à la Faculté des Lettres de l'Université de Paris. Université de Paris. Faculté des Lettres et Sciences Humaines, Paris.

Oelkers, J., 1996. In: Barraletta, J.-M., Perret-Clermont, A.-N. (Eds.), Piaget et l'éducation nouvelle. Jean Piaget et Neuchâtel. L'apprenti et le savant. Éditions Payot, Lausanne, pp. 165–176.

Oelkers, J., 2008. Jean-Jacques Rousseau. Continuum International Publishing Group, London.

Oksenberg Rorty, A., 1998. In: Rorty, A. Oksenberg (Ed.), Rousseau's Educational Experiments. Philosophers on Education. Historical Perspectives. Routledge, London/New York, pp. 238–254.

Osterwalder, F., 1995. Die pädagogischen Konzepte des Jansenismus im ausgehenden 17. Jahrhundert und ihre Begründung. In: Theologische Ursprünge des modernen pädagogischen Paradigmas. Jahrbuch für Historische Bildungsforschung, Band 2. Juventa Verlag, Weinheim/München, pp. 59–83.

Ravier, A., 1941. L'éducation de l'homme nouveau. Essai Historique et Critique sur le Livre de l'Émile de J.-J. Rousseau. t. I/II. Éditions Spes, Issoudun.

Rousseau, J.-J., 1961. Oeuvres Complètes. In: Gagnebin, B., Raymond, M. (Eds.), T. I: Les Confessions. Autres Textes Autobiographiques. Gallimard, Paris.

Rousseau, J.-J., 1964. Oeuvres Complètes. In: Gagnebin, B., Raymond, M. (Eds.), T. III: Du Contrat Social. Écrits Politiques. Gallimard, Paris.

Rousseau, J.-J., 1969. Oeuvres Complètes. In: Gagnebin, B., Raymond, M. (Eds.), T. IV: Émile. Education – Morale – Botanique. Gallimard, Paris.

Shanks, L.P., 1927. A possible source for Rousseau's name Emile. Modern Language Notes XVII, 243–244.

Starobinski, J., 1971. Jean-Jacques Rousseau. La transparence et l'obstacle. Éditions Gallimard, Paris.

Terrasse, J., 1970. Jean-Jacques Rousseau et la quête de l'âge d'or. Palais des Académies, Bruxelles.

Voltaire, 1961. Mélanges. Préface par E. Berl. Text établi et annoté par J. van der Heuvel. Gallimard, Paris.

Routine Activities, Criminal Opportunities, Crime and Crime Prevention

Pamela Wilcox, University of Cincinnati, Cincinnati, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A theoretical framework focused on crime as a function of situation-specific opportunity emerged in the field of criminology in the latter part of the twentieth century. This framework consists of several overlapping yet distinct theories, including routine activities theory. Though routine activities theory is perhaps most notable, a number of other perspectives emphasizing situational opportunity were developed around the same time as the initial explication of routine activities theory. Today, the overlap in these distinct theoretical statements is widely recognized, and they are thus increasingly treated in an integrative fashion, so as to create a more general crime opportunity perspective.

Criminology is largely focused on understanding why it is that some people – whether as individuals or as aggregates – are more likely to break the law than other people. In other words, criminologists have long sought to explain differences in propensity for law-violating behavior, often referred to as criminality. However, a subset of criminologists have reoriented scholarship away from trying to understand criminality and toward an understanding of crime events instead. They note that for a crime event to take place, offenders must have more than criminality; they must also have the opportunity to act on their desires.

This 'opportunity perspective' of crime events emerged in the field of criminology around 1980, and it now consists of several overlapping yet distinct theories. Though initially viewed as outside the mainstream of criminological theory, the opportunity perspective has since gained many followers, and it is now considered a major paradigm in the field. This article elaborates on the development of the crime opportunity perspective. This description begins with an explanation of routine activities theory, arguably the most important theory comprising the opportunity perspective.

Routine Activities and Crime

Though crime long has been a social concern in the United States, public and scholarly concern over the crime problem was clearly justified in the 1970s and 1980s. Historical trends of national crime data show that there were unprecedented increases between 1960 and 1970, in particular, with rates still climbing, albeit less dramatically, up to 1980. But what were the social changes that were behind these trends? Accounting for the unprecedented increase, nationally, was the starting point for a series of articles published by University of Illinois colleagues Lawrence Cohen, Marcus Felson, and Kenneth Land during the late 1970s and through the 1980s. (Marcus Felson left Illinois for University of Southern California in 1984. He was there until 1995, when he joined the faculty in the School of Criminal Justice at Rutgers University. Most recently, he joined the faculty at Texas State University. Kenneth Land departed University of Illinois for University of Texas's Department of Sociology, where he was on faculty from 1981 to 1985. In 1986, he then moved to the Department of Sociology at Duke University, where he remains today. Lawrence Cohen also left University of Illinois for University of Texas. He subsequently joined the sociology faculty at University of California–Davis.)

This team of researchers rejected the prevailing theories of the day as possible explanations for the historical, national-level trends, they were observing in the time series spanning from the late 1940s to the mid-1970s. For instance, they eschewed the notion that the dramatic increases could be attributed to economic deprivation, which was often depicted as the culprit. In fact, they identified a 'sociological paradox' in the examination of trends in economic conditions in the United States vis-à-vis trends in crime. During the 1960s, household income increased, and the disparity between white and black household incomes declined; the percentage of urban residents living in poverty decreased; unemployment decreased; and median education level increased, especially among blacks (Cohen and Felson, 1979). Yet, between 1960 and 1975, robbery rates increased by 263%; aggravated assault, rape, and homicide increased between 164 and 188%; and burglary increased by 200% (Cohen and Felson, 1979).

In 1979, Cohen and Felson published the seminal article in the American Sociological Review, *Social Change and Crime Rate Trends: A Routine Activity Approach*. It was the first in a series of articles to be published by the team of Cohen, Felson, and Land. In that article, Cohen and Felson addressed the paradox as follows:

> Why, we must ask, have urban violent crime rates increased substantially during the past decade when the conditions that are supposed to cause violent crime have not worsened – have, indeed, generally improved?…we consider these paradoxical trends in crime rates in terms of the changes in the "routine activities" of everyday life. We believe the structure of such activities affects criminal opportunity….
>
> (Cohen and Felson, 1979: pp. 588–589)

The Minimal Elements of Crime

As a starting point for explaining how routine activities produce criminal activity, Cohen and Felson enumerated three minimal elements for a crime event (see also Cohen et al., 1980, 1981;

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45080-4

Felson and Cohen, 1980; Felson, 1987, 1994). Though inclination to offend (the focus of traditional criminological approaches) is one of these elements, the mere presence of an offender is not enough for a crime event. If a motivated offender does not encounter a suitable victim or target, a crime event cannot occur. Or, if a potentially suitable victim or target do meet, but there is someone or something present that is capable of intervening or preventing the crime, then a crime event cannot occur. Thus, at minimum, a crime event needs the following:

1. a motivated offender – someone inclined, willing, and able to commit the crime;
2. a suitable victim or target – a person or thing that allows the offender to fulfill criminal objectives; and
3. absence of capable guardianship – any persons or things that can discourage/prevent the crime.

Cohen, Felson, Land, and colleagues stressed that these three minimal elements must converge in a place/time context or spatiotemporal setting. There is not criminal opportunity without a motivated offender converging in place and time with a suitable victim or target and absence of capable guardianship.

The Role of Routine Activities

Cohen, Felson, Land, and colleagues presumed that the convergence in time and space of the three ingredients for crime was affected by the patterns of activity comprising everyday life. Activities outside the home were seen as particularly opportunistic. Going to school, to work, to shop, to dine out, or to be entertained – these are all legitimate, public activities comprising everyday life. Yet, these legal, routine activities were presumed to have consequences for illegal activities, nonetheless, because they affect criminal opportunity. More specifically, these activities affect both the supply of suitable victims or targets available to motivated offenders and the level of guardianship provided to those victims/targets. As such, they affect the extent to which there will be spatiotemporal convergence of motivated offenders and suitable targets in the absence of capable guardianship. In short, criminal and noncriminal activities are presumed to be interdependent.

As an example of how legitimate activities feed illegitimate activities, Cohen, Felson, Land, and colleagues posited that the dramatic rise in crime witnessed in the United States between 1960 and 1970 was a result of profound changes in the routine, daily activities of its residents. The change in routine activities increased the overall supply of suitable targets with ineffective guardianship, thereby increasing criminal opportunity. But how, specifically, were American activities changing, beginning in 1960, so as to affect criminal opportunity? The short answer to the question was that Americans were engaging in many more nonhousehold activities. Simply put, with more Americans engaging in activities outside the home, they were more accessible to motivated offenders, and they and their property were less effectively guarded. Substantial increases in college attendance and labor force participation among women between 1960 and 1970 was responsible for a large share of the overall shift in Americans' activities. Women's pursuit of post-secondary education and work outside the home increased their public exposure, thereby also enhancing the pool of available, suitable targets for predatory crimes like robbery and sexual assault. At the same time, less guardianship was being provided to American households during the daytime, thereby increasing opportunity for crimes such as burglary. For instance, in 1960, about 30% of households were unoccupied by someone 14 years or older between the hours of 8 a.m. and 3 p.m.; by 1971, that figure exceeded 40% (e.g., Cohen and Felson, 1979).

The time period between 1960 and 1970 saw shifts in activities beyond the realms of school and work. Shifts in education- and employment-related routines of women had the effect of generating more disposable income for American households – income which was channeled toward more leisure pursuits outside the home (i.e., attending sporting events, movies, restaurants, or vacation destinations) and the purchase of more household durable goods (i.e., televisions, stereos, cars, and bicycles). This proliferation of public leisure activities and household items (especially valuable but lightweight items) served to increase opportunities for crime. Thus, the overall crux of the explanation offered by routine activities theory for the large uptick in crime after 1960 (especially between 1960 and 1970) was that motivated offenders were increasingly likely to encounter people with money and other valuables on their person and/or encounter an empty house with plentiful goods for the taking.

Cohen and Felson (1979) first tested routine activities theory in their 1979 article by examining whether a 'household activity ratio' was related to the changing rates of predatory crime in the United States between the years 1947 and 1974. They measured the household activity ratio as the proportion of US households that were either 'non-husband-wife' households or households that could be classified as "married, husband-present, with a female labor-force participant" (Cohen and Felson, 1979: p. 600). They surmised that these household structures were likely to engage in public activities that would increase exposure to motivated offenders, increase target suitability, and decrease guardianship. They found that the household activity ratio was, in fact, positively and significantly associated with five different types of predatory crime rates. In a follow-up article, 1 year later, Cohen et al. (1980) also found that consumer expenditures on durable goods (nonautomobile) were positively related to 25-year trends in US burglary rates.

In short, Cohen, Felson, and Land put forth a theory that challenged conventional way of thinking about crime as a pathology that emerged from social ills. Their challenge was particularly interesting given the social and political context of the 1970s. This was a time of unprecedented change. It was not just that crime was rising; there were riots in the inner cities and protests were abundant. In response, presidential crime commissions were formed, and there were conservative calls for law and order (e.g., Wilson, 1975). It was natural that everyone was worried about the 'bad things' happening in society. But Cohen, Felson, and Land took a step back and boldly presented an alternative view that 'good things' were causing fluctuations in crime.

Lifestyle, Exposure, and Criminal Victimization

The year before Cohen, Felson, Land, and colleagues began publishing numerous works on the influence of routine activities on temporal trends in US crime, a similar idea had been put forth as a way to understand variation in victimization risk across demographic subgroups by a group of criminal justice scholars at State University of New York at Albany. In the 1970s and 1980s, the Criminal Justice Research Center at Albany, under the founding directorship of Michael Hindelang, was engaged in analyses of (then-new) national victimization survey data (from the National Crime Survey, now called the National Crime and Victimization Survey). The analyses they provided of victimization data were transformative for a field that had historically relied upon official police reports for the measurement of crime. Yet, the National Crime Survey data did more than provide an alternative aggregate measure of crime in the United States. It also provided a mechanism for understanding the victim's role in crime occurrences.

One particularly influential product of the Albany-based analysis of the National Crime Survey was the 1978 book, *Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization* by Michael Hindelang, Michael Gottfredson, and James Garofalo. The crux of this work was the suggestion that victimization risk was partly a function of lifestyle. Their argument was that demographic groups experiencing relatively higher rates of victimization – men, younger adults, and African-Americans, in particular – tended to have lifestyles that exposed them to victimization opportunities. In particular, Hindelang et al. (1978) posited that the lifestyles of these high-risk demographic groups, relative to low-risk demographic groups (i.e., women, the elderly, whites), involved more time in public (especially at night), more time away from family or household members, and greater proximity to and/or association with high-offending groups. This idea was termed the 'lifestyle–exposure theory' of victimization.

The theoretical overlap between lifestyle–exposure theory and routine activities theory is evident. Lifestyle–exposure theory was originally developed as a way to understand variation in victimization risk across different demographic groups in the United States, whereas routine activities theory was posited as a way to understand temporal patterns in national-level crime rates. While they developed independently, and within the context of very different academic departments and research programs, both stressed that lifestyles and daily activities affected opportunities for crime/victimization events. Taken together, these perspectives suggested that lifestyles and activities, and the opportunities for crime that they created, varied across demographic groups and spatiotemporal context. Given the similarity and compatibility among these perspectives, they were eventually merged into a more general 'lifestyle–routine activities theory' (L-RAT) (e.g., Cohen et al., 1981; Miethe et al., 1987). By the 1990s, L-RAT had become a popular approach for understanding individual variation in victimization risk, in particular (e.g., Miethe and Meier, 1994; Miethe and McDowall, 1993; Fisher et al., 1998; Mustaine and Tewksbury, 1998; Schwartz and Pitts, 1995; Wilcox Rountree et al., 1994).

Routine Activities and Offender Searches: Crime at Specific Nodes, Paths, and Edges

Around the same time that Cohen, Felson, and Land were developing routine activities theory and Hindelang, Gottfreson, and Garofalo were explicating a lifestyle–exposure theory of victimization, a pair of criminologists working in Canada were also thinking about crime in terms of opportunity as opposed to motivation. Paul and Patricia Brantingham edited their seminal book *Environmental Criminology* in 1981. It offered a view of how offenders searched for targets, with the result of their searches being discernible, nonrandom patterning of crime (creating hot spots) within communities. Their work suggested that targets were chosen through purposeful search, in an attempt to balance issues of effort, risk, and profit. As such, targets that were easily accessible, vulnerable, and valuable were most often sought by offenders. However, the Brantinghams recognized that the target selection decision was also hierarchical or multistaged. In other words, before certain targets are selected, a general area (i.e., community) within the offender's activity space was chosen (Brantingham and Brantingham, 1981, 1993; see also Taylor and Gottfredson, 1986). In this way, offenders have been described as foragers, who "must find a good hunting ground before starting to chase prey" (Bernasco and Block, 2009: p. 96; see also Bernasco and Nieuwbeerta, 2005; Coupe and Blake, 2006; Felson, 2006). Thus, environments in which specific targets are located – not just the specific targets themselves – are key in providing criminal opportunity to offenders. Simply, places vary in terms of their suitability for offenders' target searches.

What sorts of places are perceived by offenders as advantageous hunting grounds? The Brantinghams suggest that offenders will find their targets within the geographic space where they carry out their routine activities; they refer to this area as an offender's 'activity space' (e.g., see Brantingham and Brantingham, 1981, 1993, 1995). Specifically, offenders are likely to choose attractive targets at or near the places where they conduct their routine activity. These activity nodes include schools, workplaces, shopping or entertainment areas, and more generally, areas near offenders' residences. Motivated offenders can readily come into contact with suitable targets at such nodes. Hence, activity nodes generate plentiful opportunities for crime. Major routes (streets, walkways, etc.) that surround and connect various nodes also provide numerous targets for offenders searching for crime. These major routes, which the Brantinghams refer to as 'paths,' are therefore likely to host higher than average crime events. Some central nodes and pathways will generate so much opportunity that they might actually become 'crime attractors' – places that offenders travel to specifically for crime purposes.

The Brantinghams also suggested that offenders often find targets in their search for crime, not only at major nodes and paths within their activity spaces but also at 'edge space' (Brantingham and Brantingham, 1981, 1993, 1995). According to the Brantinghams, edge spaces are boundary areas, which form the transition from one distinct physical space to another. Given their transitional nature, there are sometimes multiple users of edge spaces; those who use each

space often converge at the edges, creating a 'crossover' of users. This crossover provides an abundance of targets for offenders. At the same time, there is often ambiguity in terms of who 'owns' and thus controls edge space, thus diminishing guardianship.

Crime and Place Research

Several scientific lines of inquiry have supported the idea, emerging from the Brantinghams' work, that crime is particularly plentiful at major nodes, in particular. For instance, beginning in the 1980s, Dennis Roncek and his colleagues published a series of studies highlighting how streets with nonresidential land uses had heightened crime risk (Roncek and Bell, 1981; Roncek and Faggiani, 1985; Roncek and LoBosco, 1983; Roncek and Maier, 1991). Specifically, their work showed that presence of on-premise alcohol outlets and high schools were positively related to crime on street blocks. These establishments were presumed to generate crime on the blocks containing them by increasing population density, even if the density was ephemeral (i.e., due to nonresidents visiting at particular times of day). Such high-density land uses – which attracted teenage and young adult users and/or users under the influence of alcohol – were particularly conducive to a convergence of motivated offenders and suitable, unguarded targets. Contemporary research since Roncek's pathbreaking work continues to support the idea that street blocks or block groups with more nonresidential land uses experience higher rates of crime (Kurtz et al., 1998; LaGrange, 1999; Rice and Smith, 2002; Smith et al., 2000).

The classic study on crime 'hot spots' by Sherman et al. (1989) also supports the Brantinghams' theoretical ideas. Sherman et al. analyzed all calls for service for 1 year in Minneapolis. These crimes could have potentially occurred at 115 000 different 'places' (i.e., addresses or intersections) within the city. However, their analysis of the calls revealed that most places experienced no crime. Instead, crime tended to cluster nonrandomly in just a few 'hot spots' with repeat crime problems. In fact, fully 50% of all crimes in the city occurred at just 3% of the places throughout the city. Among the identified 'hot spots' were nodes such as department stores, bars, and convenience stores. Subsequent work by David Weisburd and his colleagues has shown similar patterns of hot spot concentration in Seattle, with crime clustering nonrandomly on certain street segments and largely remaining stable in this patterns of clustering over time (Groff et al., 2010; Weisburd et al., 2004, 2012). This group's work has emphasized the clustering of crime at nodes and along major paths; streets segments which are arterial roads (paths), bus routes (paths), and near public facilities (nodes) are among those with heightened likelihood of being chronic crime hot spots as opposed to crime-free street segments. Recent work has shown similar crime clustering in other cities (e.g., Bernasco and Block, 2011; Braga et al., 2011).

In sum, routine activities theory was initially tested at the national level with time series data on US crime rates. The theory was subsequently integrated with lifestyle–exposure theory and applied at the individual level in order to account for variation in victimization risk. Most recently, it has been integrated with theory on offender search patterns and applied to microspatial units such as places and street segments in order to help understand crime hot spots.

Environmental Design and Crime

Alongside the development of lifestyle–routine activities theory in the 1970s and 1980s was a theoretical and empirical line of inquiry stressing the idea that a community's physical design – its 'built environment' – played a major role in the opportunities for crime therein. The origins of the focus on the relationship between environmental design and crime can be traced to influential, overlapping works by Jane Jacobs (1961), Elizabeth Wood (1961), C. Ray Jeffrey (1971), and Oscar Newman (1972). Oscar Newman's book, *Defensible Space: Crime Prevention through Urban Design*, was particularly useful in articulating an explicit theory of crime (and the prevention thereof).

His famous work, *Defensible Space*, was a study of urban residential housing communities, the various physical forms they take, and the implications of those forms for crime rates within. In contrast to the high-rise public housing developments, Newman noted that low-income residential communities that consisted of low-rise 'walk-up' style housing (i.e., townhouse style) tended to have few problems. Since social characteristics of residents were similar across many of the high- and low-rise public housing communities he observed, he claimed that differences in physical designs were behind differences in crime and other social ills across the developments. According to Newman, the physical layout of the low-rise walk-ups allowed the space to be more easily defended or controlled by residents. In short, the low-rise style housing communities had defensible space. Newman elaborated that defensible space was related to four principles: territoriality, natural surveillance, image, and milieu.

Newman defined territoriality as "the capacity of the physical environment to create perceived zones of territorial influences" (Newman, 1972: p. 51). In other words, territoriality is the extent to which a space conveys a sense of being 'owned' or 'private' and as having clearly designated purposes (i.e., people have a clear sense of how and when to use the space) through use of boundary markers, signs, and other visual cues. Natural surveillance referred to "the capacity of physical design to provide surveillance opportunities for residents and their agents" (Newman, 1972: p. 78). The height of buildings, the placement of buildings vis-à-vis the street and other buildings, the number of access routes in and out of an area, the height and placement of shrubbery or trees, and the presence or absence of alcoves, niches, or alleyways are examples of physical design features that affect the potential for resident-based surveillance. "Image" was defined by Newman as "the capacity of design to influence the perception of an area as unique, well-maintained, and non-isolated" (Newman, 1972: p. 102). The use of unique designs and distinctive materials in the building of a space, as well as the upkeep and maintenance of the space, conveys the message that the area is being cared for, thus deterring incivility and making it

more easily defended. Poorly maintained spaces, on the other hand, generate a negative type of distinctiveness – a negative stigma that can attract offenders. Also related to image is 'milieu,' or the location of the space within the broader urban locale. Spaces adjacent to risky places can experience 'spillover'.

Overall, then, Newman suggested that spaces which provided territoriality, natural surveillance, positive image, and 'safe' adjacency would be more easily defended by residents. His theory, therefore, assumed that the physical environment was important because of how it affected the social control exerted within. Newman, in fact, defined defensible space as "a model for residential environments which inhibits crime by creating the physical expression of a social fabric that defends itself" (Newman, 1972: p. 3). He saw territorial behavior – inhabitants taking primary responsibility for controlling behavior – as a latent characteristic of small communities, like public housing projects. This latent territorial behavior only needed to be fostered through smart physical design.

Crime Prevention through Environmental Design

Newman's idea that physical design affected crime through prompting resident-based territorial attitudes and behavior was influential yet unpopular among some. Critics provided evidence that 'defensible' physical design did not necessarily affect the willingness of residents of an area to engage in social control (e.g., Donnelly and Kimble, 1997; Merry, 1981). However, many aspects of Newman's work were lasting. In particular, it fostered a movement that focused on the idea that physical design of spaces can affect the opportunity for crime, regardless of whether resident-based territorial attitudes or behaviors are affected and therefore can be manipulated for prevention purposes. This alternative conceptualization of the role of physical design is referred to as Crime prevention through environmental design (CPTED) and it is popular in a variety of settings today (e.g., see Crowe, 2000).

CPTED posits that aspects of physical space – such as an area's accessibility, surveillance potential, management or ownership, and level of support offered for legitimate activities – should affect crime by impacting the costs versus benefits of crime available to offenders. A number of studies over the past three decades have shown support for the importance of physical design features that affect surveillance, in particular. For instance, number of access routes, traffic volume, presence of escape routes (i.e., proximity to highway), and obstacles to surveillance within a neighborhood are positively related to rates of crime (Bernasco and Luykx, 2003; Greenberg et al., 1982; Taylor and Harrell, 1996; White, 1990).

Offender Decision Making and Crime Patterns

Around the same time that routine activities theory, lifestyle–exposure theory, offender search theory, and environmental design theory were being developed, Ronald Clarke, in collaboration with Derek Cornish, was defining a contemporary rational choice theory of offending. This theory, like the aforementioned theories, highlighted the importance of criminal opportunity.

A seminal 1985 book chapter by Clarke and Cornish, *Modeling Offenders' Decisions: A Framework for Research and Policy*, was particularly key in defining their perspective. In this piece, they provided a framework countering positivist theories of crime which posited that criminal behavior was caused or determined by distinct biological, psychological, or social characteristics. Rather, Clarke and Cornish suggested that offenders make active, utilitarian decisions about alternative lines of action, including crime, through a process of 'bounded rationality'. In other words, offenders were thought to actively weigh the costs and benefits of alternative lines of action (i.e., crime versus legitimate alternatives), though it was recognized that their perceptions of costs versus benefits were likely shaped by biological, psychological, and social 'background factors, previous experiences, immediate needs or pressures, and so on. Thus, criminal choices may not be optimal, but to Clarke and Cornish, they represented offenders' best efforts at optimizing outcomes (Clarke and Cornish, 1985: pp. 163–164; see also Shover, 1996). Regardless of the many factors shaping an offender's perceptions of cost versus benefit associated with criminal action, such action is still ultimately presumed to be based on the utilitarian pursuit of pleasure over pain according to Clarke and Cornish's perspective.

Clarke and Cornish delineate two major types of decisions made by offenders: crime involvement decisions and crime event decisions. A crime involvement decision refers to the offender's choice that crime is an acceptable line of action. Event decisions consist of the choices that are made in actually carrying out an offense. For instance, decisions about where to offend (area selection), against whom or what to offend (target selection), and how to offend (modus operandi) are event decisions. Event decision making, in particular, is presumed to involve offenders making reasoned choices – choices about the particular ecological settings within which they will offend and the particular targets they will select. Choices about settings and targets are thought to be based on environmental cues regarding risk, reward, and effort (e.g., Clarke, 1980; Clarke and Cornish, 1985). In other words, criminal opportunity weighs heavily in choices during crime events according to Clarke and Cornish's rational choice perspective.

Integrated Perspectives

All perspectives discussed up to this point in the article emphasize 'criminal opportunity' as key in understanding offender decisions and the resulting spatiotemporal clustering of crime. The various perspectives assume that offenders make conscious decisions, albeit sometimes severely constrained, about their crime events. Collectively, they implicate as key various features of the crime setting but are united in the underlying principle that crime is ecologically rooted. Though overlapping and conceptually compatible, these various perspectives have largely developed independent of one another and have traditionally been treated by most scholars

as rather disparate. However, that trend has been changing with the development of integrative approaches.

The 25 Techniques of Situational Crime Prevention

Perhaps the most widely known integrative approach to understanding the myriad environmentally based sources of criminal opportunity comes from Ronald Clarke and various colleagues who have, over the last few decades, developed an explicit approach to reducing crime at particular spatial locations. Their approach – termed situational crime prevention (SCP) – builds directly on Clarke and Cornish's Rational Choice Theory in that it assumes that offenders make choices about committing crime at particular places using utilitarian principles. Specifically, SCP suggests that the concepts of effort, risk, reward, provocation, and excuses provide a framework for understanding the decisions of offenders. Offenders are likely to choose to commit a particular crime at a particular location if the crime is easy to commit, if it is low risk and high reward, if there are situational provocations, and if there are available excuses provided by the environmental setting. Thus, in order to prevent future crime at hot spots, steps can be taken to change the environment in which the crime is repeatedly occurring such that the effort required on the part of offenders is enhanced, risk is increased, rewards are lessened, provocations are reduced and excuses are removed (see, e.g., Cornish and Clarke, 2003; Smith and Clarke, 2012).

Specific techniques offered by Cornish and Clarke for crime reduction are based on their elaboration of the situational factors affecting offender's perceptions of effort, risk, reward, provocation, and excuses. The situational factors they identify appear to draw upon the theories discussed above, including routine activities theory and environmental design theory. For instance, perceived effort is influenced by whether a potential target is 'hardened' (i.e., protected), by whether the location in which the target is situated can be accessed by the offender, by whether the location in which the target is located can be exited (by the offender) without difficulty or obstacle, and by whether any tools or weaponry necessary for committing a crime are available to the offender. An offender's perception of risk is likely influenced by such things as the amount of guardianship provided to specific targets, the amount of natural surveillance available in the environment in which the target and offender are located, and the amount of both informal and formal surveillance and control exerted within the environment in which the target and offender are located. Perception of reward is influenced by factors such as the level of visibility and portability of targets, the extent to which targets can be traced, and the availability of markets that allow offenders to move stolen merchandise. Provocations can be influenced by environmental conditions that incite anxiety, foster disputes, or encourage boundary testing. Such conditions include extreme noise, overcrowding, rivals within close proximity (i.e., rival sports fans watching 'the big game' at the same bar), disorder, and peer pressure. Finally, offenders may perceive excuses for crime in settings in which, for example, rules are ambiguous, behavioral instructions are not evident, or where compliance is not facilitated.

Thus, conceptual integration is inherent in theory behind the techniques of SCP; overlapping constructs from previously disparate perspectives are woven together and used to understand how offenders perceive opportunity and, in turn, how we might reduce opportunity. While conceptually integrative, those practicing SCP have really shied away from integration across multiple contexts and have instead emphasized opportunity and its reduction at one primary unit of analysis – the immediate, microspatial setting in which repeated crime events are occurring. However, as discussed more fully below, recent theoretical work has suggested that situational prevention is perhaps optimally applied simultaneously to multiple embedded contexts.

Multilevel Opportunity Theory

A general, multilevel opportunity perspective emphasizes that opportunity stems from multiple, overlapping theoretical processes and opportunity is structured at multiple, embedded levels of analysis (Wilcox et al., 2003; Wilcox et al., 2013). For example, a multilevel perspective on opportunity stresses the importance of recognizing that crime is patterned nonrandomly across multiple levels of analysis, including individuals, places, streets (or street segments), neighborhoods, and the like. These various levels represent unique types of behavior settings and, as such, they have distinct structures of opportunity (Taylor and Gottfredson, 1986; Taylor, 1998). The various levels/settings and their respective opportunity structures – additively and in interaction with one another – can affect the spatial patterning of crime that we see at various levels of analysis.

As an example of a general, multilevel opportunity theory, Wilcox et al. (2003) conceptualized and operationalized opportunity at multiple levels. They drew upon routine activities theory most directly, suggesting the opportunity was a function of three major factors: (1) exposure to motivated offenders, (2) target suitability, and (3) absence of capable guardianship. However, they suggested that these three factors not only emerged from routine activities but also were affected by environmental design and various facets of social control (including social bonds/handles, social disorganization, and deterrence). In addition to conceptualizing opportunity by integrating concepts from various opportunity theories, Wilcox et al. (2003) suggested that opportunity existed at both individual and environmental levels, with 'environmental' potentially referring to streets, schools, neighborhoods, or broader areas. Finally, they suggested that individual-level and environmental-level indicators of opportunity could interact.

There is mounting empirical evidence that the tenets of a general, multilevel opportunity theory – such as that posited by Wilcox et al. (2003) – have merit (for a recent review, see Wilcox et al., 2013). For instance, there are now a number of multilevel studies of victimization which, collectively, promote the idea that characteristics associated with routine activities, environmental design, and social control at individual/household and neighborhood levels of analysis impact victimization risk (e.g., Kennedy and Forde, 1990; Sampson and Wooldredge, 1987; Smith and Jarjoura,

1989). Furthermore, a number of the multilevel studies show that victimization risk is affected by the interaction between individual- and neighborhood-level indicators of opportunity (Miethe and McDowall, 1993; Wilcox et al., 2007; Wilcox Rountree et al., 1994).

Conclusion

Routine activities theory is perhaps the most noted criminological perspective emphasizing crime opportunity as a key theoretical construct. However, a number of other perspectives emphasizing the importance of opportunity were being developed by various groups of scholars in different parts of the world around the same time as routine activities theory. Over the course of the past several decades, these overlapping perspectives have coalesced to form a more overarching, general opportunity perspective on crime.

See also: Crime Trends and Debates; Criminology: Development and Current Status of the Discipline; Economics and Crime; Property Crime: Victimization and Trends; Spatial Dynamics and Crime.

Bibliography

Bernasco, W., Block, R., 2009. Where offenders choose to attack: a discrete choice model of robberies in Chicago. Criminology 47, 93–130.

Bernasco, W., Block, R., 2011. Robberies in Chicago: a block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. Journal of Research in Crime and Delinquency 48, 33–57.

Bernasco, W., Luykx, F., 2003. Effects of attractiveness, opportunity and accessibility to burglars on residential burglary rates of urban neighborhoods. Criminology 41, 981–1002.

Bernasco, W., Nieuwbeerta, P., 2005. How do residential burglars select target areas? British Journal of Criminology 45, 296–315.

Braga, A.A., Hureau, D.M., Papachristos, A.V., 2011. The relevance of micro places to citywide robbery trends: a longitudinal analysis of robbery incidents at street corners and block faces in Boston. Journal of Research in Crime and Delinquency 48, 7–32.

Brantingham, P.J., Brantingham, P.L. (Eds.), 1981. Environmental Criminology. Sage, Beverly Hills, CA.

Brantingham, P.L., Brantingham, P.J., 1993. Nodes, paths, and edges: considerations on the complexity of crime and the physical environment. Journal of Environmental Psychology 13, 3–28.

Brantingham, P.L., Brantingham, P.J., 1995. Crime generators and crime attractors. European Journal of Crime Policy & Research 3, 5–26.

Clarke, R.V., 1980. 'Situational' crime prevention: theory and practice. British Journal of Criminology 20, 136–147.

Clarke, R.V., Cornish, D., 1985. Modeling offender's decisions: a framework for research and policy. Crime & Justice 6, 147–185.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Cohen, L.E., Felson, M., Land, K.C., 1980. Property crime rates in the United States: a macrodynamic analysis, 1947–1977; with ex ante forecasts for the mid-1980s. American Journal of Sociology 86, 90–118.

Cohen, L.E., Kluegel, J.R., Land, K.C., 1981. Social inequality and predatory victimization: an exposition and test of a formal theory. American Sociological Review 46, 505–524.

Cornish, D.B., Clarke, R.V., 2003. Opportunities, precipitators and criminal decisions: a reply to Wortley's critique of situational crime prevention. In: Smith, M.J., Cornish, D.B. (Eds.), Theory for Practice in Situational Crime Prevention, Crime Prevention Studies, vol. 16. Criminal Justice Press, Monsey, NY, pp. 41–96.

Coupe, T., Blake, L., 2006. Daylight and darkness targeting strategies and the risks of being seen at residential burglaries. Criminology 44, 431–464.

Crowe, Timothy D., 2000. Crime Prevention through Environmental Design: Applications of Architectural Design and Space Management Concepts, second ed. National Crime Prevention Institute, Louisville, KY.

Donnelly, P.G., Kimble, C.E., 1997. Community organizing, environmental change, and neighborhood crime. Crime and Delinquency 43, 493–511.

Felson, M., 1987. Routine activities and crime prevention in the developing metropolis. Criminology 25, 911–932.

Felson, M., 1994. Crime and Everyday Life: Insight and Implications for Society. Pine Forge Press, Thousand Oaks, CA.

Felson, M., 2006. Crime and Nature. Sage, Thousand Oaks, CA.

Felson, M., Cohen, L.E., 1980. Human ecology and crime – a routine activity approach. Human Ecology 8, 389–406.

Fisher, B.S., Sloan, J.J., Cullen, F.T., Lu, C., 1998. Crime in the ivory tower: the level and sources of student victimization. Criminology 36, 671–710.

Greenberg, S.W., Rohe, W.M., Williams, J.R., 1982. Safety in urban neighborhoods: a comparison of physical characteristics and informal territorial control in high and low crime neighborhoods. Population and Environment 5, 141–165.

Groff, E.R., Weisburd, D., Yang, S., 2010. Is it important to examine crime trends at a local "micro" level? A longitudinal analysis of street to street variability in crime trajectories. Journal of Quantitative Criminology 26, 7–32.

Hindelang, M.J., Gottfredson, M.R., Garafolo, J., 1978. Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization. Ballinger, Cambridge, MA.

Jacobs, J., 1961. The Death and Life of Great American Cities. Vintage, New York.

Jeffery, C.R., 1971. Crime Prevention through Environmental Design. Sage Publications, Beverly Hills, CA.

Kennedy, L.W., Forde, D.R., 1990. Routine activities and crime: an analysis of victimization in Canada. Criminology 28, 137–152.

Kurtz, E., Koons, B., Taylor, R.B., 1998. Land use, physical deterioration, resident-based control, and calls for service on urban streetblocks. Justice Quarterly 15, 121–149.

LaGrange, T.C., 1999. The impact of neighborhoods, schools, and malls on the spatial distribution of property damage. Journal of Research in Crime & Delinquency 36, 393–422.

Merry, S.E., 1981. Defensible space undefended: social factors in crime control through environmental design. Urban Affairs Quarterly 16, 397–422.

Miethe, T.D., McDowall, D., 1993. Contextual effects in models of criminal victimization. Social Forces 71, 741–759.

Miethe, T.D., Meier, R.F., 1994. Crime and Its Social Context: Toward an Integrated Theory of Offenders, Victims, and Situations. State University of New York Press, Albany.

Miethe, T.D., Stafford, M.C., Long, J.S., 1987. Social differentiation in criminal victimization: a test of routine activities/lifestyle theories. American Sociological Review 52, 184–194.

Mustaine, E.E., Tewksbury, R., 1998. Predicting risks of larceny theft victimization: a routine activity analysis using refined lifestyle measures. Criminology 36, 829–857.

Newman, O., 1972. Defensible Space: Crime Prevention through Urban Design. Macmillan, New York.

Rice, K.J., Smith, W.R., 2002. Socioecological models of automotive theft: integrating routine activity and social disorganization approaches. Journal of Research in Crime & Delinquency 39, 304–336.

Roncek, D., Bell, R., 1981. Bars, blocks, and crimes. Journal of Environmental Systems 11, 35–47.

Roncek, D., Faggiani, D., 1985. High schools and crime: a replication. Sociological Quarterly 26, 49–505.

Roncek, D., LoBosco, A., 1983. The effect of high schools on crime in their neighborhood. Social Science Quarterly 64, 598–613.

Roncek, D.W., Maier, P., 1991. Bars, blocks, and crimes revisited: linking the theory of routine activities to the empiricism of 'Hot Spots'. Criminology 29, 725–753.

Sampson, R.J., Wooldredge, J., 1987. Linking the micro- and macro-level dimensions of lifestyle-routine activity and opportunity models of predatory victimization. Journal of Quantitative Criminology 3, 371–393.

Schwartz, M.D., Pitts, V.L., 1995. Exploring a feminist routine activities approach to explaining sexual assault. Justice Quarterly 12, 9–31.

Sherman, L., Gartin, P., Buerger, M., 1989. Hot spots of predatory crime: routine activities and the criminology of place. Criminology 27, 27–55.

Shover, N., 1996. Great Pretenders: Pursuits and Careers of Persistent Thieves. Westview, Boulder, CO.

Smith, M.J., Clarke, R.V., 2012. Situational crime prevention: classifying techniques using "good enough" theory. In: Welsh, B.C., Farrington, D.P. (Eds.), The Oxford Handbook of Crime Prevention. Oxford University Press, New York, pp. 291–315.

Smith, W.R., Frazee, S.G., Davison, E.L., 2000. Furthering the integration of routine activity and social disorganization theories: small units of analysis and the study of street robbery as a diffusion process. Criminology 38, 489–523.

Smith, D.A., Jarjoura, G.R., 1989. Household characteristics, neighborhood composition, and victimization risk. Social Forces 68, 621–640.

Taylor, R.B., 1998. Crime and small-scale place: what we know, what we can prevent, and what else we need to know. In: Crime and Place: Plenary Papers of the 1997 Conference on Criminal Justice Research and Evaluation. NIJ, Washington, DC.

Taylor, R.B., Gottfredson, S.D., 1986. Environmental design, crime, and prevention: an examination of community dynamics. In: Reiss Jr., A.J., Tonry, M. (Eds.), Communities and Crime, Crime and Justice: A Review of Research, vol. 8. University of Chicago Press, Chicago.

Taylor, R.B., Harrell, A., 1996. Physical Environment and Crime. US Department of Justice, Office of Justice Programs. National Institute of Justice. NCJ 157311.

Weisburd, D., Bushway, S., Lum, C., Yang, S.-M., 2004. Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. Criminology 42, 283–321.

Weisburd, D., Groff, E.R., Yang, S.-M., 2012. The Criminology of Place: Street Segments and Our Understanding of the Crime Problem. Oxford University Press, New York.

White, G.F., 1990. Neighborhood permeability and burglary rates. Justice Quarterly 7, 57–67.

Wilcox, P., Gialopsos, B.M., Land, K.C., 2013. Multilevel criminal opportunity. In: Cullen, F.T., Wilcox, P. (Eds.), Handbook of Criminological Theory. Oxford University Press, New York, pp. 579–601.

Wilcox, P., Land, K.C., Hunt, S.A., 2003. Criminal Circumstance: A Dynamic, Multicontextual Criminal Opportunity Theory. deGruyter, Aldine, New York.

Wilcox, P., Madensen, T.D., Tillyer, M.S., 2007. Guardianship in context: implications for burglary risk and prevention. Criminology 45, 771–804.

Wilcox, R.P., Land, K.C., Miethe, T.D., 1994. Macro- micro integration in the study of victimization: a hierarchical logistic model analysis across Seattle neighborhoods. Criminology 32, 387–414.

Wilson, J.Q., 1975. Thinking about Crime. Basic Books, New York.

Wood, E., 1961. Housing Design: A Social Theory. Citizens' Housing and Planning Council of New York.

Relevant Website

http://www.popcenter.org – Center for Problem-Oriented Policing.

Rule of Law (and Rechtsstaat)

Martin Krygier, University of New South Wales, Sydney, NSW, Australia; and Regulatory Institutions Network, Australian National University, Acton, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In all its many versions, the rule of law has to do with the relationship between law and the exercise of power, particularly public power. As an ideal, it signals that law can and does well to contribute to articulating, channeling, constraining, and informing – rather than merely serving – such exercise. Beyond that, what it rules out, what it allows, what it depends on, and indeed what it is, are all matters of disagreements that stem from differences among political and legal histories and traditions, and reflect dilemmas and choices that recur, in different forms and weights, in many such histories and traditions. This article is concerned with these enduring themes, dilemmas, and choices, as they occur within particular traditions, especially the common law 'rule of law' tradition, on the one hand, and the Continental Rechtsstaat tradition, on the other.

From the last quarter of the twentieth century, the rule of law has come to occupy an increasing amount of discursive space, not only among lawyers, for whom it had been an old theme, but also among promoters of economic development, human rights, democratization, state building, and political and legal reform. Increasingly, it is alleged, the rule of law is a key ingredient in the attainment of all these good things and others. As one author has observed,

> Among a plethora of development and security agencies, a new "rule of law consensus" has emerged. This consensus consists of two elements: (1) the belief that the rule of law is essential to virtually every Western liberal foreign policy goal – human rights, democracy, economic and political stability, international security from terrorist and other transnational threats, and transnational free trade and investment; and (2) the belief that international interventions, be they through money, people, or ideas, must include a rule-of-law component. (Call, 2007: p. 4)

In this transformation, the rule of law has gained a great deal in modishness but less, actually nothing, in clarity. But clarity was never the concept's strong suit. Like many other important moral, political, and legal ideals, among them democracy, justice, and liberty, its meaning, scope, conditions, and significance are all highly, perhaps essentially, contested (Waldron, 2002). And like those ideals, not only are there enduring common themes but also common axes of argument and disputation that pervade discourse on the rule of law. As suggested elsewhere (Krygier, 2011: p. 69), these contests do not render such concepts meaningless or useless. On the contrary, some of them are the most important we have. We will not resolve those contests, here or anywhere, but it might be possible to clarify a few of them and suggest why they – and the rule of law – are important.

The concept of the rule of law embodies ideals that have figured in political and constitutional discourse at least since Aristotle, who contrasted 'the rule of the law' with 'that of any individual.' Those ideals have varied, so too the strategies to achieve them and the verbal formulation of them. They revolve around enduring themes and concerns; however, the character of which is nicely captured by Otto Kirchheimer's laconic observation that:

> for all the differences in historical roots and particular legal traditions their common denominator lies in the simple thought that the security of the individual is better served when specific claims can be addressed to institutions counting rules and permanency among their stock-in-trade than by reliance on transitory personal relations and situations. Beyond that, a good part of their common success probably lies in the mixture of implied promise and convenient vagueness. (Kirchheimer, 1992; first published 1967: p. 244)

In all versions, the rule of law has to do with the relationship between law and the exercise of power, particularly public power. As an ideal, it signals that law can and does well to contribute to articulating, channeling, constraining, and informing – rather than merely serving – such exercise. That takes us some distance from those who see law simply as one of the means by which power is exercised, neither better nor worse than any other. For there are lots of ways to exercise power; partisans of the rule of law insist that it helps us block some of them, including many of the worst of them (see Rundle, 2009), and to channel others in salutary directions and ways. But what it rules out, what it allows, what it depends on and indeed what it is, are all matters of disagreement. This is so for several reasons, often stemming from differences among political and legal histories and traditions, and reflecting dilemmas and choices that recur, in different forms and weights, in many such histories and traditions.

Law and State

These days, and in contemporary language, the words 'law' and 'state' are rarely far apart. However, it was not always and everywhere so, and even now it makes a difference whether the connection is seen as necessary or contingent, even more so whether it is seen as a conceptual rather than an historical connection. A clue to this comes from the terms used in various languages. In particular, there is an obvious semantic difference between 'rule of law,' the term used in English, and those found

in many European languages to cover some, but not all, of the same terrain. Each of these has a context and a history that cannot be ignored or simply elided, but in a host of European languages there is one thing commonly built into the concept, which is missing from the English phrase: the State. Whether it is *Rechtsstaat* (German: state of law; law-governed state), *état de droit* (French), *statto diritto* (Italian), *estado de derecho* (Spanish), *państwo prawa* (Polish), or *pravovoe gosudarstvo* (Russian), law is inextricably connected to the state. It is the subject – grammatically and ontologically – of each rendition. However, unlike all these terms, the rule of law does not mention the state. This is not an accident.

The concept of the state is not part of English constitutional jurisprudence, while in Australia and the United States it refers to what Germans would call *Länder* (see MacCormick, 1984: p. 65). More deeply, the English tradition was long pluralistic in its conceptions of the sources of law (Rosenfeld, 2001), with multiple cumulative and competing authoritative sources, among them custom, court decisions and statutes. Indeed, while the common law courts were long agents of the Crown, some of the mythologically most powerful contests in the English rule of law tradition, particularly the constitutional struggles of the seventeenth century, pitted them against successive wearers of that Crown, even at the cost of the head of one of those (Charles I), what might in other countries be called the head of State.

According to the common law tradition, popular custom, an "ancient collection of unwritten maxims and customs" (Blackstone [1765–69] 1979, vol. 1: p. 17) was long seen as a primary source of law, and "[t]he only method of proving, that this or that maxim is a rule of the common law, is by showing that it hath been always the custom to observe it" (Blackstone [1765–69] 1979, vol. 1: p. 68). That custom was evidenced (rather than made, it was claimed) by another nonlegislative source, the judgments of courts in particular cases brought to them. This was the 'common law' (Krygier, 1998), which seventeenth-century partisans of the rule of law preferred to the commands of the King; "Law established by customary practice, law that was not the creation of will, command, or sovereignty, was a restraint on government – a restraint on discretionary power" (Reid, 2004: p. 12). True, the courts were the King's courts, but the law they adjudicated was not, in the main, considered to be the King's creation. It was not just an instrument with which the Crown and the state could direct activities and control public policy. For the King, like his subjects, was subject to 'the law of the land.'

Deliberate, secular, purpose-guided, prosaic (not sacred) legislation, as Max Weber observed, is a central and distinctive characteristic of modernity in law (see Weber, 1968: pp. 760–768). Today, of course, in its exponential rush since the eighteenth century (see Lieberman, 1989), legislation has swamped custom and even judicial decisions as a quantitatively primary and increasingly imperious source of law, in the common law world as elsewhere. However, the notion that the rule of law draws upon sources other than legislative fiat (i.e., an order for an act to be carried out), that the judiciary is a fundamental guardian of it, and that all, even the most powerful, are and should be subject to it, goes deep in the common law tradition and has not lost resonance. It was expressed in many ways over centuries, but the canonical connection between the term 'rule of law' and those thoughts came to displace other descriptors, primarily as a result of the hugely influential late nineteenth century work by A.V. Dicey (Dicey, 1961 (first published 1885)). Dicey's formulations distilled (and in some respects distorted) a very old English legal tradition.

No parallel existed in the nineteenth and early twentieth century *Rechtsstaat*. The term was coined only at the very end of the eighteenth century (Heuschling, 2002: p. 29), to capture a new phenomenon, the modern State with its monopoly of force. That state was the subject of this concept, and also the legal source of law. The *Rechtsstaat* ruled *by or through* law, whereas other states, such as the *Machtstaat* (power-governed state) or *Polizeistaat* (police-governed state) (see Raeff, 1983) might dispense with it and exercise power in other ways. What was distinctive of a *Rechtsstaat* was not that the state was subject to law that had other sources and independent guardians, but that it acts in a *rechtlich* (lawful, legal) way; "according to some nineteenth century (and early twentieth century) constructions, there is a relation of near-identity between the state and its law … within the system of rule the law is the state's standard mode of expression, its very language, the essential medium of its activity" (Poggi, 1978: p. 102). That is how we recognize it as a *Rechtsstaat*, as distinct from any other type of state. There was no conceptual space to say to the State what Sir Henry Bracton already said in the thirteenth century:

> The king has a superior, namely God. Also the law by which he is made king. Also his *curia*, namely the earls and barons, because if he is without bridle, that is without law, they ought to put the bridle upon him. (quoted in Reid, 2004: p. 11, and see Palombella, 2010; *passim*)

That is an important rule-of-law claim.

Arbitrary Power: Uncontrolled or Unruly

That the law should rule even over the most powerful people and institutions is a very old theme in the English legal tradition. The rule of law is commonly contrasted with *arbitrary* exercise of power; that, above all, is the evil that the rule of law is supposed to curb. This leads to another difference, this time masked rather than revealed by semantics. For arbitrariness is itself an ambiguous concept. Is it, for example, 'uncontrolled interference' or 'interference that is not subject to established rules' (Pettit, 2012: p. 58)? These are two of several (see Richardson, 2002; Chapter 3) conceptions of the concept. They have particular relevance to law; the former commonly being referred to as 'government under law,' the latter as 'government *by* law.' If one had to choose, there are strong arguments to favor the former over the latter (see Pettit, 2012), but ideally one would encounter the first always, and in exercises of public power the second as well. Not everyone opposed to 'arbitrary power' has had both these senses in mind, however.

As we saw, the common law tradition, from at least the thirteenth century until well into the eighteenth, maintained that the king was subject to a law that he had not made, indeed

that made him king. For the king, for anyone, to ignore or override that law was to act arbitrarily (see Reid, 2004; Palombella, 2010). Liberties, and procedures to protect them such as habeas corpus and due process, were enshrined in that law, and encroachment on such liberties was barred, even to the monarch, by the law. That that law was often *not* expressed in clear, prospective, general rules (see Maitland, 1965; first edition 1908: p. 383), today regarded as the essence of the rule of law, was not to the point (Reid, 2004). Indeed, given the customary, dynamic, fluxful, and evolutionary character of the common law as theorized by its adepts, it was beside the point (see Postema, 1986; Chapter 1). The issue was that it was superior.

From the eighteenth century, however, law came to be viewed increasingly as the direct or indirect product of the political legislator, the 'sovereign,' and in English law there was, at least arguably until very recently as a result of EU membership, no legal superior to the sovereign legislator. The conception of the rule of law gradually became more preoccupied with the character of the rules that the sovereign enacted: they should be clear, prospective, consistent, etc.

Shocked by this downgrading of the notion of a law superior to the sovereign, and by what they regarded as their own 'arbitrary' treatment by the British sovereign Parliament (Reid, 2004), the American colonists first staged a revolution, and then a pathbreaking innovation: a written constitution binding on the legislator, and in due course routinely overseen by an independent Supreme Court, whose decisions also came to be seen as binding on the legislator. This was a novel way of vindicating a very old principle. In England, old conceptions persevered, but in increasing tension with the legislative bias of modernity.

On the Continent, things were different. A *Rechtsstaat* was not just any sort of state, as we have seen, but one which operated on the basis of legal rules configured in particular ways. "[S]ituated at the heart of the theory of the *Rechtsstaat* is the question of the arbitrariness of power, of the potential violence inscribed in all relations of domination, whether private or public" (Heuschling, 2002: p. 42). However, partisans of the *Rechtsstaat* did not envisage a law superior to the state, a basis for appeal to some higher notion or other source of law. Law was a characteristic of the *Rechtsstaat*, but it was also its product. The nonarbitrary *rechtlich* quality of the state was a matter of the degree to which its edicts took the form of general rules that conformed to specific formal criteria and were supposed, in particular, to guarantee certainty and predictability. In this understanding, "[t]he *Rechtsstaat* means that the law is the structure of the State, not an external limitation to it. … Liberty is a consequence not truly a premise of the law. The authority vested in this conservative aristocratic state protected civil liberties as a service offered through the State" (Palombella, 2010: pp. 11–12).

The notion that state agencies must comply with a law above the State, only came with the development of written and legally binding constitutions, particularly in reaction to the Nazi calamity in the middle of the last century. Until then, although "it was in the state's interest to promote its self-limitation through self-binding to legal norms" (Loughlin, 2010: p. 320), and the people's interest too, it was up to the state to bind itself (*Selbstbeschränkung*). The common law tradition frequently, and in the seventeenth century vociferously, conceived of individual rights as protected by the courts *against* the Crown, whereas no such opposition existed in the German conception of the *Rechtsstaat* (see Rosenfeld, 2001: p. 1319), which, as Leondard Krieger shows, projected "older national assumptions which made the idea of liberty not the polar antithesis but the historical associate of princely authority" (Krieger, 1957: p. 5). The contrast is deep. As Gozzi observes:

> In Germany … the doctrine of the *Rechtsstaat* precludes the possibility of the primacy of law over the state. Indeed, it is precisely in the relationship between law and state – which in the German case is settled with the primacy of the state – that the most significant feature of the doctrine of the German *Rechtsstaat* emerges. Conversely, the English doctrine of the government of law is most clearly distinguished by grounding the rule of law on the superiority of law as proclaimed by the courts of justice. (Gozzi, 2007: p. 238)

From the point of view of those subject to the exercise of power, both its control and its manner of exercise, government under and by law, are important. But they are not the same. A state could be controlled but act under decrees with quite particular targets, kept secret from citizens, or inconsistent with each other, or retrospective, or without any decrees, let alone laws, at all. It could, conversely, be uncontrolled but act through promulgated, clear, consistent etc. laws. In either case, something significant would be lacking. For where arbitrariness in either of these senses is linked with significant power, it at the very least raises the reasonable apprehension that it will tend to: threaten the *liberty* of anyone subject to it; generate reasonable and enduring *fear* among them; and deprive citizens of sources of reliable sources of expectations of, and *coordination* with, each other and with the state. And as Lon Fuller (1969) and Jeremy Waldron (2011a) have emphasized, it threatens the *dignity* of all who find themselves mere objects of power exercisable at the whim or caprice of another.

These are four good reasons to value reduction of the possibility of arbitrary exercise of power (see Krygier, 2011: pp. 79–81). There may well be many others, such as those that commend themselves to many economists, having to do with the alleged contribution of the rule of law to economic development (Dam, 2006). To the extent that the rule of law can help deliver such reductions, this is reason to value it. This is not, of course, merely a negative matter of removing evils, but can be expressed positively. A society in which law contributes to securing freedom, confidence, coordination and dignity, is some great and positive distance from many available alternatives. There are other things we want from law, and many more things we might want in a good society, but ways of serving these values are goods immeasurably harder to attain without institutionalizing constraints on arbitrariness, in both these senses, in the exercise of power.

For some thinkers, speculating in these ways about what good might flow from reducing arbitrariness in the exercise of power, what it might be *for*, takes us beyond the analytical task of understanding what the rule of law *is*; for others it does not get us close to what matters. The former favor 'thin' accounts;

the latter lard their accounts of the rule of law with more ingredients and of different kinds. These are often known in the literature as 'thick' accounts of the rule of law.

Thin or Thick

Apart from questions of control versus character of law, writers on the rule of law often distinguish between 'thin' or 'formal,' on the one hand, and 'thick,' 'substantive,' or 'material' conceptions of it, on the other. The former limit themselves to formal properties of laws and legal institutions, which are purported to constitute the rule of law. The latter require substantive elements from a larger vision of a good society and polity – democratic, free-market, human rights respecting, or some such – to be present.

The first track is favored by modern analytical jurists. They have often adopted (Hart, 1967: pp. 273–274) or extended (Raz, 1979; Walker, 1988) Lon Fuller's eight principles of what he called 'the morality of law' as defining characteristics of the rule of law, even when they disagreed with him over whether they deserved to be called moral principles. According to Fuller (1969), these characteristics were that there must be rules, these rules must be publicly available, prospective, understandable, consistent, possible to perform, sufficiently stable for citizens to orient their actions by them, and administered in ways congruent with their terms. There is controversy over whether there is any reliable connection between such thin principles and substantive values beyond them (see Krygier, 2010: pp. 114–120) but whatever the view on that, on a thin conception those further values are something other than the rule of law.

Again, the *Rechtsstaat* has oscillated between thick and thin through its 200 years of evolution. It was first theorized by German liberal constitutional and administrative theorists, prominent among them Karl Rotteck, Karl Theodor Welcker and Robert von Mohl, seeking to characterize a legal order in terms of values it served (those values in their turn to be realized *in* and not *against* the state). The post-Nazi *Rechtsstaat* returned to, and richly amplified, a normative characterization based on the fundamental value, inscribed in the first article of the German Basic Law of 1949, of human dignity. However, in between times, after the failure of the 1848 revolutions and particularly under and after Bismarck, its late nineteenth-century and early twentieth-century versions were pared of normative adhesions, and strictly devoted to elaborating the formal components of a legal order that might properly be called a *Rechtsstaat*.

Not only legal philosophers, but also legal comparativists (see Peerenboom, 2004) tend to favor 'thin' versions of the concept, what might be called rule-of-law-lite: easier to identify and able to travel further, because it carries less baggage. Many governments, too, particularly authoritarian ones, prefer to be assessed against thin formal criteria, easier to satisfy than thick morally demanding ones. Today international businessmen, unwilling to buy into controversial questions about democracy, human rights and other large values in, say, Singapore and China (with both of which they might want to do business), often prefer a formal, thin, conception too.

Many, however, find thin conceptions quite inadequate. Ronald Dworkin, for example, was skeptical of conventional 'rule book' conceptions of the rule of law, which insist that "so far as is possible, the power of the state should never be exercised against individual citizens except in accordance with rules explicitly set out in a public rule book available to all. ... Those who have this conception of the rule of law do care about the content of the rules in the rule book, but they say that this is a matter of substantive justice, and that substantive justice is an independent ideal, in no sense part of the ideal of the rule of law" (Dworkin, 1985: p. 11). He, by contrast, regarded the rule of law as incorporating an ideal and an eminently positive and substantive one, "the ideal of the rule by an accurate public conception of individual rights" (Dworkin, 1985: pp. 11–12).

The sociologist Philip Selznick had a more complex combination of thin and thick. He agreed with those political realists who stressed the importance of strict legality as a restraint on, and saw the rule of law as a precious protection against abuse of, power (Selznick, 1992: p. 174). On the other hand, he insisted that there was a 'larger promise of the rule of law,' and this "thicker, more positive vision speaks to more than abuse of power. It responds to values that can be *realized*, not merely protected, within a legal process. These include respect for the dignity, integrity, and moral equality of persons and groups. Thus understood, the rule of law enlarges horizons even as it conveys a message of restraint" (Selznick, 1999: p. 26).

In Germany, the circumstances that moved prevailing conceptions of the *Rechtsstaat* from thin to thick were more dramatic than those that preoccupied Dworkin and Selznick. Indeed they were tragic. Already in the Weimar Republic, Hermann Heller rejected the legal positivist, formalistic, conception of the *Rechtsstaat* as crystallized by his contemporary, Hans Kelsen, which could accommodate any state. He argued for one that insisted that only a democratic state that depended upon and then institutionalized fundamental ethical principles was a *Rechtsstaat* (for the debates between Carl Schmitt, Hans Kelsen and Hermann Heller, which centered on the nature of the *Rechtsstaat*, see Dyzenhaus, 1997). Although Heller died in 1933, he already saw fascism as the great threat to such a state. His nightmare became real in the ensuing years.

In the perspective of German post-Nazi retrospection and introspection, thin conceptions came to seem not merely inadequate, but on their own positively dangerous. The *Rechtsstaat* embodied in post-war German jurisprudence thus embodies a strong commitment to fundamental rights and to the dignitarian premise of its 1949 *Grundgesetz* or Basic Law (see Grote, 1999), grounded in its unamendable Article I. This proclaims that "Human dignity shall be inviolable. To respect and protect it shall be the duty of all state authority." Particularly through the interpretations of the Federal Constitutional Court, this has spawned a rich jurisprudence of fundamental rights that characterizes the modern German understanding of the *Rechtsstaat*, or as it is frequently expanded (and at times complicated) to join the social welfare state (*Sozialstaat*), the *sozialer Rechtsstaat*.

There are problems at the extremes. What is gained by defining down a concept that bears so much normative

resonance, in terms that ignore any interrogation of what its point might be, and simply focus on the characteristics of institutions and practices? Particularly when it is not clear whether the characteristics chosen by 'thin' theorists relate as much to what law does and should do in the world, as they do to lawyers' unevenly informed intuitions and guesses about these matters. Again, what of the exercise of power by extralegal forces, social networks? If they are free to act arbitrarily, capriciously, whatever the law says, does it make sense to insist that nevertheless the rule of law exists because certain formal elements of a legal order are present? Excessively thin conceptions often seem urgently in need of a feed.

Yet, conversely, accounts that purport to be thin as a rake are often rather plumper than intended, particularly where – as is frequently the case in well-intentioned first world interventions in benighted countries – they embody parochial suggestions as to what features of familiar legal orders generate rule-of-law friendly results. When packages of legal *bric-à-brac* are asked to travel, it often turns out that they work very differently or not at all where they land (see Krygier, 2011). It might also turn out that institutions and practices of sorts not known in the homes of confident rule of law exporters perform adequately in their own homes, even if they look quite strange to visitors. Whether they do or not should be a matter of investigation, not overbearing legalistic assumption. Too often, however, imported assumptions about the working of legal institutions, based on distant histories, traditions, institutions, and practices, have been smuggled in and then re-sold as though of universal applicability. When they fail to 'take' is it because the rule of law is a false ideal, or because what has been exported is not the rule of law itself but parochial institutions taken to be necessary for values that might yet be reached, and need to be reached, in other ways?

Partisans of thin versions, on the other hand, often associate the thickness insisted on by moralists with a combination of parochialism and imperialism about values and institutions. Why should we assume, either as a matter of fact or of value, that all cultures value the same things from law? Metaethical disputes this raises are too large to be resolved here, but there is another worry about too thick an account of the rule of law. As Joseph Raz has argued, 'thick' conceptions have a tendency to wash away all distinction between the rule of law and anything else we might want. That lays them open to the criticism that "[i]f the rule of law is the rule of the good law then to explain its nature is to propound a complete social philosophy. But if so the term lacks any useful function. We have no need to be converted to the rule of law just in order to discover that to believe in it is to believe that good should triumph" (Raz, 1979: p. 211).

Such criticism points up another inadequacy of the choice on offer: the dichotomy between a spare and formalistic thinness, on the one hand, and a pudgy confection of everything we would like to find in a good society, does not exhaust the field. There is space for values particularly associated with the exercise of power, what might be called distinctively *legal* values were it not for the fact that the differences between them and other values are unlikely to be categorical, but rather matters of focus, shades, and degree. Many legal orders bear and transmit long histories of observation, experience, and reflection upon the pathologies of unconstrained and/or capricious exercise of power and on what might be done to avoid or moderate such pathologies. Legal orders typically embody and generate values related to what they do, both in their animating principles and ideals, and in the complaints they provoke when their practices flout the sources of their legitimacy (see Selznick, 1999; Waldron, 2011a). These have included such values as due process or natural justice, ideals of fair treatment and notice, and in particular legal traditions much more. They have to do with treating a subject of power with the respect due to a person, rather than a 'rabid animal or a dilapidated house' (Waldron, 2011a: p. 16). The rule of law might well be argued to be incomplete to the extent that some such power-related values, that have to do with how to arrange and transact potentially harmful interactions between it and its subjects (too often treated as objects), are dishonored. Attempts to vindicate such values, often implicit in many legal principles and traditions if not all legal rules, might be considered a service to the rule of law, even if they go beyond purely formal aspects of laws and notwithstanding that they might fall short of justice more holistically conceived. On one view, the special disease, to which the rule of law is part of a remedy, is the propensity of power unconstrained to be exercised in arbitrary ways. It is a widespread disease.

Anatomy or Teleology

More fundamental than contests over the fleshiness of the rule of law is one over whether we should begin with focus on its anatomy, the institutional features one should expect to find in such a creature – or its point – the reasons one is concerned with this rather than something else. The long common law tradition was not fixated on specific institutions, even less the precise character of legal rules. As Reid demonstrates, it was all too murky: "Ironically, the medieval constitutional law out of which today's rule of law developed would not have met the requirements of clarity or precision. There was always an air of indefiniteness, a smoky vagueness surrounding this all-embracing restraining 'law' of English constitutionalism. Even its authority as law was shrouded in immeasurability" (Reid, 2004: p. 16). But the common law tradition was clear on one thing: "It was, Viscount Bolinbroke said in the eighteenth century, a matter of curbing power and not of the type and structure of government. Whether power was vested in a single monarch, in 'the principle Persons of the Community, or in the whole Body of the People' was immaterial. What mattered was whether power was without control. 'Such Governments are Governments of *arbitrary Will*,' he contended" (Reid, 2004: p. 42). Bolinbroke would likely be puzzled by the rule of law toolkits carried by UN and World Bank rule of law promoters throughout the world today; uniform in character, diverse in application, apparently universal in application. Why are *those* particular institutions sacrosanct? What is the *point*? How has the point influenced the kit? Even for those of us who have left the mythologies and hagiographies of common law theory long behind, these are not bad questions.

Similar choices are also found in the *Rechtsstaat* tradition. The early propagators of the concept were not legal anatomists but proto-liberals in the main, many influenced by Kant, seeking to establish an order of citizens equal before the law, whose personal autonomy and property were protected by the law. A state was a *Rechtsstaat* to the extent that it achieved these tasks, not because it had this or that particular form. As von Mohl put it, "the objective of the *Rechtsstaat* is not logically entwined with a particular form of government; on the contrary, every arrangement of public power which guarantees the right and the development of all human activities, is admissible" (quoted in Heuschling, 2002: p. 59; my translation from Heuschling's French). These writers were explicit that what mattered was achieving what they conceived to be the point of the rule of law, not whether it exemplified a particular form of institutional architecture. That understanding did not continue through the nineteenth century, but was supplanted by a formalistic, anatomizing conception, stressing the positivistically characterized features of a state, that qualified it to be declared a *Rechtsstaat*.

On the teleological account, however, the rule of law cannot adequately be explicated by a list of features of legal institutions, rules or practices. For the rule of law occurs when and to the extent that there is a social *achievement* to which law contributes. If we say, for example, that there were lots of laws under Stalin and a lot of rule, but there was not much rule of law, we are not saying something controversial, and you would not have to know much about Dicey or Fuller to agree. So at least among the legally and philosophically unwashed, the rule of law has something to do with what the law does, rather than simply with what it has been somewhere declared to be.

Moreover, if the law is enlisted to do things we associate with the rule of law but the mission fails, we might say that there was an attempt to achieve the rule of law, but it was unsuccessful: laws were of the sorts we associate with the rule of law, everyone was trying, but they were overborne, for whatever reason. To say the rule of law exists in a society is to imply an accomplishment; among its partisans a valued accomplishment: an ideal to which law is taken to contribute has been approached.

On this view, the rule of law is not a natural entity like a tree, simply awaiting scientific description, or even a man-made contrivance like *a* rule of law in a statute book, which might be identified by pointing to it. It exists to the extent that a certain state of affairs, one in which power is exercised in relatively nonarbitrary ways, exists in the world. Law is supposed to contribute, though it will never do so on its own. The aspiration or ideal is satisfied only *insofar as* some purpose or goal for law is realized. While such an achievement could in principle be thought value neutral or even valueless, and has been, the rule of law also has partisans – today perhaps, even too many – who think it valuable, an ideal for law. If we value that ideal we should of course seek to identify what might be necessary to generate it. But that is a second step. Without some principle of selection even if only tacit, we won't find a bunch of legal bits and pieces waiting 'out there' and recognizable as the rule of law.

The teleological contention is, then, that to understand what the rule of law requires we need to start by reflecting first on its *point* rather than, as is more common, with an enumeration of purportedly defining legal-institutional *features*, whether they be particular institutions such as common law courts (Dicey, 1961), particular formal qualities of rules, such as prospectivity, clarity, etc., or even traditions and procedures, such as defences, *habeas corpus*, and so on (Waldron, 2011a), though the last is getting closer to explicit concern with the specific point of the rule of law.

At first blush, this looks like a repetition of the distinction between thin and thick, and it is true that anatomical accounts of the rule of law are often 'thin,' since they focus on delineating the characters of legal institutions. But there are two differences. First, particularly among rule of law promoters, it is rare that anyone has thin ambitions, more commonly they just have confused ideas of what the rule of law is about. Rule of law promotion, after all, is ostensibly an attempt to enlist the rule of law to do good in the world, not just to build replicas of institutions from home. However, rule of law promoters are often restricted by the conventional identification of the rule of law, or *Rechtsstaat*, with a particular box of tricks, and proceed to try to vindicate some purpose with an a priori catalog of what is needed to achieve it, rather than an openness to the possibility that they might need to learn some new tricks. Awareness that one should start with the end, as it were, rather than purported means, might avoid a lot of grief over transplants that fail to do what is expected of them: promote the rule of law.

Moreover, the distinctions between thin versus thick, on the one hand, and anatomical versus teleological, on the other, do not occupy the same plane. A teleological account is not necessarily normatively thick; it might occupy itself with a small point, say predictability in the legal environment. How normatively enriched the point of the rule of law might be is a legitimate matter of debate, but it is a debate on the teleological plane. On that plane, the question is not, first of all, how much normative weight the concept carries but where one should start to think about the rule of law – by enumerating a set of purported (and typically universal) features or by asking what it is might be good for. Since it is hard to know what features matter unless one has sorted out what they are for, the suggestion here is to start with the end.

Legal or Socio-legal

If one is concerned with underlying values that inspire commitment to the rule of law, this has significant and somewhat paradoxical implications for where one should look to vindicate whatever one decides such values to be. For the search to redeem them is likely to decenter law itself. After all, it is in principle an open, and likely variable, matter what in the world best minimizes arbitrariness in the exercise of power, and the same might be said of any other values that we imagine law helps vindicate. Yet if ends matter, then it is not clear that one should *assume* that law is always key to achievement of the animating values of the rule of law, even less the state. This is so, whatever the values one has in mind, and it will be all the more so as one ramps up the values one associates with the rule of law. A good society is quite an achievement, and law only a small, if precious, contributor.

This applies at virtually every level. If arbitrary power is to be feared, then wherever power is powerful enough to be fearful, rule of law concerns are relevant. Thus, preoccupation with the state is not always appropriate, in circumstances where many of the sources of restraint on arbitrary power, many dangers flowing from it, and many of the goods accomplished by its curtailment, lie outside the state, and many of the means of achieving those goods lie outside the law as well (see Krygier, 2011: pp. 85–91).

If that is the case, it is not obvious that the familiar institutions associated by legal anatomists with the rule of law will always be the best to wield. Law will never accomplish much in the world on its own, and a key accompaniment of investigation of the rule of law should be, but rarely has been, study of what else is needed, beside and beyond, law, to attain its ends.

On this view, law should be viewed, not as the always-necessary centerpiece of power-taming policy to which other measures are subordinate or supplementary addenda, but as one implement among several, of varying significance, in some respects and circumstances of potentially unique importance, but dependent for its success on many other things, and perhaps not more important for the achievement of its own goal than they. Similar reflections apply, but all the more, to the State of the *Rechtsstaat*.

There is something to be said for the legal pluralism embodied in the old common law tradition, and squeezed out by the rise of the contemporary state: in principle in the *Rechtsstaat*, and by the statophilic tendencies of modernity more generally. Even with its dominance, and especially where it is 'failed,' 'fragile,' or 'transitional,' the state is never the only game in town. That is a sociological platitude, but it should have more bearing on legal platitudes than it has. Lawyers will naturally, habitually, focus their attention on state and legal agencies, but those interested in promoting the values that underpin the rule of law and make it worthwhile will need to look further afield.

If the foregoing considerations are plausible *within* existing nation states, they must be all the more compelling for anyone who wants, as many today do, to speak about an emerging 'international rule of law' (Palombella, 2009; Waldron, 2011b). For whatever that might mean, there is no international *Staat* to be its lawgiving and enforcing source. It might be a matter of argument whether it is a good idea to seek to extend the rule of law to the international sphere, but it would seem a strange truncation of the argument simply to rule out a nonstate-centered rule of law by *definition*.

This suggests the need for a sociological awareness and sensibility not especially common among lawyers, whether rule-of-lawyers, or *rechtsstaatlich* ones. There will be other sectors of a society altogether that influence the extent to which the values at the heart of the rule of law will be attained. Paradoxically, in order to reach those values we will have to look far beyond the institutions we have most conventionally associated with them.

That does not make either the law or the state unimportant. However, it might enable us to see their importance in perspective, give due weight to other phenomena that might need enlisting to serve such goals, and release us from the hold of a mantra, whether 'rule of law' or *Rechtsstaat*, which in their modish ubiquity threaten to obscure the valuable, indeed precious purposes for which they were pushed into the fray, instead promiscuously to serve virtually any purpose you want to name.

See also: Constitutionalism, Comparative; Justice and Law; Law and Democracy; Law and Development; Law: The Socio-Legal Perspective; Legal Positivism.

Bibliography

Blackstone, William, 1979. Commentaries on the Laws of England. University of Chicago Press, Chicago.

Call, Charles T., 2007. 'Introduction' to call. In: Constructing Justice and Security after War. United States Institute of Peace Press, Washington, D.C.

Dam, Kenneth W., 2006. The Law-growth Nexus: The Rule of Law and Economic Development. Brookings Institution Press, Washington.

Dicey, Albert V., 1961. Introduction to the Law of the Constitution, tenth ed. Macmillan, London.

Dworkin, Ronald, 1985. Political judges and the rule of law. In: A Matter of Principle. Harvard University Press, Cambridge, MA.

Dyzenhaus, David, 1997. Legality and Legitimacy. Carl Schmitt, Hans Kelsen and Hermann Heller in Weimar. Oxford University Press, Oxford.

Fuller, Lon L., 1969. The Morality of Law, revised ed. Yale University Press, New Haven.

Gozzi, Gustavo, 2007. *Rechtsstaat* and individual rights in German constitutional history. In: Costa, Pietro, Zolo, Danilo (Eds.), The Rule of Law. History, Theory and Criticism. Springer Verlag, Dordrecht, pp. 237–259.

Grote, Rainer, 1999. Rule of law, Rechtsstaat and "Etat de droit". In: Starck, Christian (Ed.), Constitutionalism, Universalism and Democracy: A Comparative Analysis. Nomos Verlagsgesellschaft, Baden-Baden, pp. 269–306.

Hart, H.L.A., 1967. Problems of philosophy of law. In: The Encyclopedia of Philosophy, vol. 6, 264–276.

Heuschling, Luc, 2002. État de droit, Rechtsstaat, Rule of Law. In: Nouvelle Bibliothèque de Thèses. Éditions Dalloz, Paris.

Kirchheimer, Otto, 1967. The *Rechtsstaat* as magic wall. In: Scheuerman, William E. (Ed.), The Rule of Law under Siege. Selected Essays of Franz L. Neumann and Otto Kirchheimer. University of California Press, California, pp. 243–263.

Krieger, Leonard, 1957. The German Idea of Freedom. University of Chicago Press, Chicago.

Krygier, Martin, 1998. Common law. vol. 10. In: Craig, Edward (Ed.), Routledge Encyclopedia of Philosophy, vol. 1. Routledge, London, pp. 440–446.

Krygier, Martin, 2010. The Hart–Fuller debate, transitional societies and the rule of law. In: Cane, Peter (Ed.), The Hart-fuller Debate in the Twenty-first Century. Hart Publishers, Oxford, pp. 107–134.

Krygier, Martin, 2011. Four puzzles about the rule of law: why, what, where? And who cares? In: Fleming, James E. (Ed.), Getting to the Rule of Law. New York University Press, New York, pp. 64–104.

Lieberman, David, 1989. The Province of Legislation Determined. Cambridge University Press, Cambridge.

Loughlin, Martin, 2010. Foundations of Public Law. Oxford University Press, Oxford.

MacCormick, D. Neil, 1984. Der Rechtsstaat und die rule of law. Juristenzeitung 50 (50), 65–70.

Maitland, F.W., 1965. The Constitutional History of England. Cambridge University Press, Cambridge.

Palombella, Gianluigi, 2009. The rule of law beyond the state: failures, promises, and theory. International Journal of Constitutional Law 7 (3), 442–467.

Palombella, Gianluigi, 2010. The rule of law as institutional ideal. In: Morlino, L., Palombella, G. (Eds.), Rule of Law and Democracy. Brill, Leiden.

Peerenboom, Randall (Ed.), 2004. Asian Discourses of the Rule of Law. Routledge Curzon, London.

Pettit, Philip, 2012. On the People's Terms. A Republican Theory and Model of Democracy. Cambridge University Press, Cambridge.

Postema, Gerald, 1986. Bentham and the Common Law Tradition. Clarendon Press, Oxford.

Raeff, Marc, 1983. The Well-ordered Police State: Social and Institutional Change through Law in the Germanies and Russia, 1600–1800. Yale University Press, New Haven.

Raz, Joseph, 1979. The rule of law and its virtue. In: The Authority of Law. Clarendon Press, Oxford.
Reid, John Philip, 2004. The Rule of Law. Northern Illinois University Press, De Kalb.
Richardson, Henry S., 2002. Democratic Autonomy. Public Reasoning about the Ends of Policy. Oxford University Press, Oxford.
Rosenfeld, Michel, 2001. The rule of law and the legitimacy of constitutional democracy. Southern California Law Review 74, 1307–1352.
Rundle, Kristen, 2009. The impossibility of an exterminatory legality: law and the holocaust. University of Toronto Law Journal 59, 65–125.
Selznick, Philip, 1992. The Moral Commonwealth. University of California Press, Berkeley, CA.
Selznick, Philip, 1999. Legal cultures and the rule of law. In: Krygier, Martin, Czarnota, Adam (Eds.), The Rule of Law after Communism. Ashgate/Dartmouth, Aldershot, pp. 21–38.
Waldron, Jeremy, 2002. Is the rule of law an essentially contested concept (in Florida)? Law and Philosophy 21 (2), 137–164.
Waldron, Jeremy, 2011a. The rule of law and the importance of procedure. In: Fleming, James E. (Ed.), Getting to the Rule of Law. New York University Press, New York, pp. 3–31.
Waldron, Jeremy, 2011b. Are sovereigns entitled to the benefit of the international rule of law? The European Journal of International Law 22 (2), 315–343.
Walker, Geoffrey de Q., 1988. The Rule of Law. Foundation of Constitutional Democracy. Melbourne University Press, Carlton.
Weber, Max, 1968. Economy and Society, vol. II. Bedminster Press, New York.

Rumors and Urban Legends

Pamela Donovan, Bloomsburg University of Pennsylvania, Bloomsburg, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by J.B. Renard, volume 20, pp. 13413–13416, © 2001, Elsevier Ltd.

Abstract

Rumors are claims about events or situations that are passed primarily outside official channels and are not based upon substantiated facts. Urban legends are a subgenre of rumor with a more complete narrative. Rumor research in social science is limited and sporadic, yet plays a vital role in understanding beliefs, attitudes, emotional states, and worldviews that cannot, or will not, be openly articulated. Both social and technological shifts have influenced the analysis of rumor and urban legends.

Introduction

Rumors are claims about reality – particularly events, people, and situations – that are circulated primarily outside official sources of knowledge, such as mass media sources, organizational authorities, or government agencies. There is nothing in this definition, widely agreed upon in rumor studies, which requires a rumor to be false. It is hard not to associate it with falsehood, since false rumors fascinate and frustrate us, and in some cases, they pose grave dangers. Nonetheless, rumor is marked by its process and not the truth or falsity of its content.

Indeed, the troublesome image of rumor that we have inherited stems from its capacity for wreaking havoc – whether the rumor is false or true. Social and behavioral scientists have had a tendency to drift toward this 'havoc' view, despite sporadically acknowledging that rumor is an integral part of social life, and that often it is a side-effect of conflict rather than its instigator. Dramatic events are associated with it, we fear its power, and yet our disciplines have had difficulty over the years in sustaining interest in the topic. New scholarship on the subject is often spawned by necessity as a real world set of problems emerges, followed by years of relative inactivity in published research. As such, many social scientists interested in the topic have relied on the work of humanities scholars, as well, especially the interdisciplinary field of folklore. In this field, rumor is regarded as a timeless genre, and as such continuities across history and cultures have been emphasized. Folklore has been particularly good about cross-cultural research, the historical lineage of particular rumors, and symbolism. Social science approaches, in contrast, demonstrate strength in relating rumors to social consensus and conflict, to the role of rumor in crisis situations, and to its relationship to psychological states, cognition, and psychodynamics. Social psychology has demonstrated more consistent effort in rumor studies than other social science fields.

Historically, rumor has been spread by informal speech between individuals, either serially one to another, or in groups. Occasionally, both rumors and urban legends have been transmitted with handbills or flyers. Recently, of course, a significant conduit has been the Internet, including e-mail, webpages, and social networking. *Urban legends* are a subgenre of rumor. They typically stand alone as full narratives whereas rumors are claims, which are only sometimes linked together. Urban legends typically have greater longevity than rumors, can change their details rapidly to recontextualize their content to time and place, and resist debunking because elements of their claims are often flexible. Urban legends are sometimes called contemporary legends by researchers, reflecting the fact that these tales are often not 'urban' in the sense of being tied to city life, but rather that they have strong themes of danger and redemption specific to modernity and mass society, in contrast to small-scale societies and the folklore generated in those contexts.

Why is social science input on rumor and urban legend important? First, rumor can have social impacts that may be subtle or bold, dire, or trivial. It can accompany collective violence (though it is often not clear whether rumor incites it, or is used by aggressors as post hoc justification for it). It is sometimes genuinely hard to tell how responsible rumor is for an outcome involving it. In some cases, we may be able to effectively intervene if a rumor is leading to destructive behavior, harmful misperceptions, or individual or collective stress. As Knapp (1944) noted in his landmark article on the psychology of rumor, rumor studies enable the study of social attitudes, beliefs, and opinions in a way that circumvents social desirability bias. Since respondents do not claim to be the authors of a rumor, we can find the tacit beliefs, fears, wishes, and attitudes 'underneath' public opinion. The implications for social science are also general: the continued importance of learning to study things that are not easily measured, the integration of key insights from studies using different methodologies, and the clear need for cross-cultural and cross-disciplinary research.

The History and Heyday of Rumor Research

The birth of modern psychology in the nineteenth century brought the first experimental and psychoanalytic approaches to rumor. From the beginning, emerging tensions within psychology about the nature of group communication and behavior shaped the study of rumor. One experimental approach was designed around the general problem of hearsay, but applied to rumor more specifically. The analogy that experimenters appeared to be using was the children's game 'Whisper Down the Lane' or the parlor game 'Telephone,' where

a message changes meaning as it is passed from person to person. Distortion was primarily an informational problem, with a series of linear, cumulative chain errors appearing as the text was passed along (Hösch-Ernst, 1915; Gorphe, 1927). Of course, the researchers conceded that the new, distorted text had to be intelligible and plausible, but it was otherwise devoid of intentional creation or reshaping – rather, it was an ordinary consequence of communication. Although this somewhat mechanical view faced competition from the psychoanalytic model, and would later be supplanted in experiments by a more complex view of human communication, it did endorse an important element of rumor – the ordinariness of it. Rumor, once it exists, will travel, and it will be distorted, regardless of its content. Strangely, the ordinariness or inevitability of rumor has often been obscured in later work. Even so, the insights from these early studies were limited. Was rumor really like any text passed along from one person to another, or was it distinctive? In what way? Rumors are also sometimes true – how would their factual nature remain preserved if distortion was inevitable?

Early psychoanalytic approaches emphasized a different model of rumor, and some of these basic insights have stood the test of time. The key role of the unconscious emerges when we realize, as Von Roretz (1915: p. 28) suggested, that rumors purport to be claims about reality or news; neither opinion nor speculation, but real. As such, the teller avoids responsibility for what they are saying, and merely claims to be passing along information. The emotional investment in rumor, Prasad (1950) suggested, stems from the way in which it can provide justification for anxiety or fear, and reflect a desire for companionship in one's feelings. Many agree with the view of Rosnow and Fine (1976) that rumor plays a role in reducing anxiety – whether collective or individual – by repackaging information into more acceptable forms. The task of the researcher, in this case, is to identify not only the source of anxiety but how the repackaging fixes the problem or resolves confusion. If the anxiety-reducing factor is central to the continued appeal of rumors in a multimedia, multiplatform world of information, it is probably because rumors and urban legends offer the teller and hearer a creative role in shaping the claim while disavowing their very shaping of it.

During the heyday of rumor research in the United States, during World War II and directly after, social scientists attempted to formulate a working set of observations or 'laws' of rumor. This effort has been modestly successful in the sense that some distilled wisdom from wartime rumor research has stood the test of time and new venues. The unconscious hostile impulse, in their view, partially explained rumors, but researchers did not necessarily agree what weight to give it, nor did they agree how the problem could be most expertly managed, or whether critical-mindedness could prevent it. Caplow (1947) found that critical-mindedness did not necessarily stop a rumor; rather the rumor was more likely to be more *accurate* as it passed through skeptical ears. Critical-minded participants were more likely to look for verification, and this, rather than transmission per se, explained some sharpening of detail.

Another open question that stems from this era is just how strong a rumor-monger, or a belief in a rumor, needs to be in order for the rumor to take hold. Wartime rumor studies always considered the role of intentional subversive propaganda in fostering rumor, even finding instances of Axis radio messages that seemed related to the content of domestic rumors in Ally countries. Yet, by and large, rumor researchers have been more interested in how and why rumor intrigues the general public, rather than who started it. It is a broadly shared axiom among rumor scholars that without some key feeling of relevance to individuals or groups, a rumor will not be successful. Intentionality or malice in the spread of rumor continues to be emphasized in some communications and political science research (cf Sunstein, 2009). Most social scientists believe that the origins of rumors are often not singular or specifiable, and are often sincerely believed, even when clearly vicious, ridiculous, or destructive.

Knapp's work in a Massachusetts public safety agency led to his widely adopted typology of rumor (1944). There are *wedge-drivers*, *bogies*, and *pipe-dreams*. Wedge-drivers typically accuse a group within society of collective bad deeds against the interest of the whole. For instance, Knapp found war rumors, collected in September 1942 in the United States, that accused Catholics or Jews of evading the draft, or accused Ally countries of manipulating the war effort to American detriment. It is noteworthy in thinking about the distinctiveness of rumor that wedge-drivers were rarely directed at the Axis countries; this makes sense if we understand rumor to be a thing that cannot be easily voiced or endorsed – there was really no social risk in saying negative things about Germany, Japan, or Italy during the war, since that animosity was quite overt. Two-thirds of the rumors Knapp collected were wedge-drivers. Bogies, or fear-driven rumors, were the second most common, accounting for about one-quarter of the war rumors. The major theme for these rumors was that things were worse than they seemed, and these rumors were often graphic in their claims of hidden mass casualties (bodies washing ashore that were not reported; or the entire fleet at Pearl Harbor destroyed) or contamination of everyday products. Finally, a small number of war rumors were pipe-dreams or wishful thinking rumors, such as an early end to the war because of depleted Axis supplies. Knapp also speculated that the preponderance of different kinds of crisis rumor might shift as the crisis unfolded – perhaps more wishfulness in the beginning and more scapegoating, wedge-driving ones as the burdens of the crisis impinged on regular life. The preponderance of wedge-drivers during war is not necessarily universal: Nkpa's study (1975) of rumor during the Nigerian civil war, 1967–70, found many more pipe-dream rumors and lower levels of wedge-driving ones.

In all cases, war rumors generally presume that the truth is being at least obscured if not completely hidden from the public, and in some cases official secrecy makes this plausible. As such, Knapp was the first to propose a still widely endorsed view of rumor management: intensity of rumors varies inversely with the availability and trustworthiness of official information. The greater candor and clarity possible from government, civic, and media organizations, the easier it would be to quash or at least limit the appeal of rumor.

Common Denominators in Rumor Episodes

Allport and Postman (1947) first identified three characteristics of rumor as they travel. First, rumors are *leveled*: this is the

tendency of the rumor to become shorter and simpler. Second, they are *sharpened*: some details drop away and others are retained, and the total quantity of detail is reduced. Finally, the rumors are *assimilated*: new hearers, if motivated to reshare it, will make small changes, and sometimes big ones, that are consistent with their own interests and assumptions. Allport and Postman's experimental methodology reproduced the serial transmission model of the early psychology studies; many researchers have since questioned whether this is really a suitable analogy for the spread of rumor in real life (Buckner, 1965; Shibutani, 1966). Prasad (1935), who studied rumor and natural disaster in India, found relatively little distortion of rumor; the same rumor would emerge among different groups. Peterson and Gist (1951) found a great deal of *snowballing* or creative elaboration, rather than leveling or sharpening, in the circulation of rumors. Other researchers have found that rumors can invert their claims depending on their social location, for instance, when traveling across a racial or ethnic divide (Turner, 1993; Knopf, 1975). A rumor of an atrocity in a neighboring town can accuse a group of black assailants of attacking a white victim, and nearly simultaneously the same rumor will appear with a group of white assailants and a black victim. Certainly the common observation that rumors tend to beget other rumors complicates the matter a bit, even if the content of specific ones becomes simplified over time.

Assimilation, which has been perhaps the most agreed upon of Allport and Postman's three characteristics, also implies that a rumor's strength may be tied to how easily it can be adapted to an individual, group, or society's interests and assumptions. "Rumors are credible because they reference the worldviews of those that share them" (Fine and Turner, 2001). While Allport and Postman emphasized the individual's readiness to integrate rumor material into their worldview, subsequent researchers (Bordia and DiFonzo, 2002) emphasize the need to add cultural, social, and historical layers in order to understand the plausibility and appeal of a rumor. DiFonzo and Bordia's analysis of corporate rumors (2002) further suggests that levels of anxiety may explain much of what we think of as interest or importance. Assimilation helps us understand why rumor spread patterns look the way that they do; for instance, why one rumor might be a small sector rumor and another more widespread and traveling across more varied social territory.

The work of pioneering researchers during World War II was prompted by the desire to control rumors. Government agencies, mainstream media sources, and civic organizations seemed to agree that rumors about the war and its effects on the home front threatened subversion, in 'loose lips sink ships' form (Rosnow and Fine, 1976) or that rumors could erode morale, particularly ones that alleged that certain imposed hardships at home, like rationing, were either unnecessary or being evaded by elites. Allport and Lepkin's study (1945) of rationing rumors concludes that the primary factor in their promulgation was the lack of access to accurate information. About the research participants that represented the stressed public, they asserted that "the more information he has, and the more carefully he thinks, the less chance that he will be led into believing by his inner, and often unrecognized, impulses." The alternative, rejected hypothesis was that the privation of rationing itself generated enough stress to cause greater vulnerability to such rumors. This assumption set the stage for the 'fight claims with counter-claims' model of rumor intervention that would later prove less than effective, since it tended to sideline the emotional fuel of rumor despite its recognition of 'impulse.'

Allport and Postman identified *salience* and *ambiguity* as the key elements of rumor's vitality. By salience the authors meant that the topic must be of some importance to the group of people who hear it and perhaps retell it. Without this basic element, a mere word-of-mouth fiction would not be transmitted much, or it will only be transmitted in groups that find it salient. For instance, rumors in the 1980s about large corporations being fronts for satanic organizations were of little interest beyond people who were a part of conservative religious movements (Koenig, 1985). Nonetheless, the rumor enjoyed an intense period of transmission and belief within that subculture. Salience does not require fervent factual belief in a topic, or at least about its central claims. Victor (1993) describes a 1988 rumor panic in rural southwestern New York state, alleging ritual sacrifice and other kinds of satanic cult activity. The panic affected everyday life and cast suspicion on nonconformists. Its salience was not limited to religious conservatives, but extended to those who 'half-believed' the rumors enough to think that something must be going on that threatened the safety of their children – and that threat was enough to shape behavior. "A rumor persists," Victor notes, "when it offers a plausible explanation for people's shared anxieties." The region was suffering from a rapid postindustrial economic decline and an alarming rise in family strain indicators, such as divorce filings and child welfare caseloads. The anxiety around the future was real enough, and the primal need to protect children made the rumors salient even to those suspicious that it all might be nothing. Anxiety may be one sort of emotional engagement, which seems necessary to turn something merely relevant to one's situation into something salient enough to be worth believing or passing on. Rosnow argues that anxiety may well be intertwined with heightened perceptions of ambiguity as well as a motivational ingredient in salience (Rosnow, 1980).

By ambiguity, Allport and Postman meant that some larger situation, of which a rumor is a specific part or attempted explanation, is characterized by uncertainty or a lack of trusted information from a reliable source. This latter factor, ambiguity, is the reason why rumor seems to intensify during crisis situations. Reliable sources may be absent, or simply unable to communicate what is known or unknown about the crisis. Official sources might also be distrusted, either because the source has failed to be candid and forthcoming in the past, or because there is generalized distrust between the source and the public. A central assumption emerged that ambiguous situations create a vacuum, which rumor fills, and thus a reduction in ambiguity will reduce rumor intensity. Yet this is easier said than done, since in some cases ambiguity is not objective, but asserted and elaborated by rumor hearers and tellers themselves (Donovan, 2007).

The anthropologist Raymond Firth (2011) found that among the Tikopia of the Solomon Islands in the early twentieth century – as is certainly the case for many parts of the world still today – the distinction between 'news' and 'talk' was a bit artificial. "The news is received as conveyed by individuals, but often treated as if it stemmed from a generic social source."

Yet this did not preclude scrutiny and doubt, and when a rumor proved false, children were often blamed. Firth notes that the Tikopia often suspended judgment of the truthfulness of accounts until a later date, and passed it on conditionally – but nonetheless continually.

Modern Mediated Societies and Rumor

In Western industrialized societies, there is an expectation that because continuous and accessible government, civic, and media sources are available, they will be utilized by members of society to clarify what is true and what is not, or at least what is verified versus what has been just heard by word of mouth. Research like Firth's should give us pause about this expectation – does it even uniformly apply, or describe people's real world behavior, in a media-saturated environment like our own? Ralph Turner's study in the 1980s of earthquake rumors in the Western United States suggested that increased scientific and official communication did not quell rumor – it simply shifted its focus. About half of the people in Turner's study believed that authorities were withholding something; demand for information increased alongside increased official communication. In contrast to Allport and Postman's model, Turner suggests (1994) that "the reduction of ambiguity forces people to pay attention to unpalatable news and instigates intensified individual and collective information seeking." The official reduction of ambiguity, in essence, can backfire as it enhances overall salience, and thus spawn even more speculative rumor. More recently, the work of Weeks and Southwell (2010) seems to confirm this problem. Using Google Trends to track its search engine activity related to the rumor that US President Obama is a 'secret Muslim' the researchers find evidence that media reports about the rumor – even when, as in most cases, the reports debunk the rumor – decisively influence Google searches the next day. Online searches then lead the curious into more speculative and unsubstantiated territory.

A further problem emerges when authorities themselves promulgate rumors, and these are treated respectfully by the media. Consider tales of kidney theft, which circulate worldwide. Even those given to skepticism about domestic kidney theft (the businessman on holiday who loses his kidney in Vegas) will give more credence to the idea, once the environs shift to the peripheries of Eastern Europe, Asia, or South America, and when the victim shifts to poor or homeless children. In her book *Organ Theft Legends*, Campion-Vincent (2005) shows the relative success of this version of the tale, weaving together local rumor, news accounts, and unsupported claims by politicians and activists, and providing a global chronology. This approach avoids the common mistake of overlocalizing the meaning of rumors based on a current, site-specific set of strains, and often well-earned suspicions. An imagined black market in stolen organs would have to overcome formidable technological and social barriers, but would also be heavily dependent on impeccable, permanent silence of the highly skilled thieves and brokers. Opinion leaders were found in Campion-Vincent's study to repeat each others' claims and long-debunked allegations 'as if they had never been denied' even though they were all in a position to help dispel rather than encourage the rumors. Campion-Vincent shows that not only do organ theft rumors occasionally become pretexts for mob violence, but that trust between doctors and poor patients, and between the government and citizens is eroded. Of course, corrupt, brutal, and neglectful governments may help foment the conditions that cause belief in such rumors in the first place, but the recognition that such rumors also flourish in developed countries should give pause to those who make a simplistic equation between structural exploitation and the plausibility of rumors.

Allport and Postman's model treated ambiguity as an objective characteristic of the circumstances in question, though the authors acknowledge that individuals could subjectively find it were it not there. For wartime researchers and those that have followed, it is a matter of emphasis. Knopf, who wrote about race, rumor, and riots in the 1960s, asks whether ambiguity really stands alone as a measurable variable. In her account, the underlying conflict shapes what is known and unknown. "Take away the adversative aspects of an event and you take away much of its ambiguity" (Knopf, 1975). Thus while the ongoing rumor process may be built upon ambiguity, this ambiguity in turn is inherently built upon existing conflict. People decide how clear things already are – or are not.

Shibutani also studied rumor during wartime, but from a quite different angle. Though his landmark book *Improvised News* (1966) incorporated previous work on World War II rumors, the heart of the study was his own on-the-scene account of rumor among Japanese-Americans in internment camps in California. It was methodologically distinct, and inspired others later to look at rumor in social context, rather than only through experimental methods. Shibutani's major thesis was that rumor was a normal and not a pathological part of social life. He saw it as a problem-solving activity by groups of people who were deprived of enough trustworthy information. Shibutani's work also shifted focus away from individual tendencies toward gullibility or rumor mongering to a model where people with a variety of critical abilities and social intentions participated in the shaping of rumor. *Improvised News* inaugurated an era of careful, qualitative, situated studies – about rumors concerning everything from kidnapping and sex slavery in Orleans, France (Morin, 1970), to the covered-up 'death' of Beatle Paul McCartney (Suczek, 1972; Rosnow and Fine, 1976). Followers in this method, though, later wondered whether Shibutani had overemphasized the cognitive and rational needs of internees for clarity and closure. Was information seeking their only motive for collective rumor building? What about the respondents in many on-the-ground studies who have no desire to seek clarification, confirmation, or the ultimate truth-status of a rumor? Goode and Ben-Yehuda (1994) argue instead that "the last thing people who tell rumors want is verified information."

Despite a decline in interest in the topic of rumor after the late 1960s, a handful of significant empirical and theoretical challenges emerge from scattered studies. Folklorists began to take more interest in contemporary rumor in the 1970s and in English-language journals, interest shifts to the 'urban legend.' Jan Harold Brunvand has collected urban legends for longer than 30 years and has published several popular-audience books, with an emphasis on those legends circulating in the United States. Tying a number of widely believed tall tales together, Brunvand (1984) senses subterranean regret about

the loss of small-world charm and social familiarity; a strong undercurrent of general suspicion of others animates these types of rumor, as opposed to the specific suspicions of more provincial times.

Urban legends, as a subgenre, underscore the often idle nature of rumor. Not all negative, accusatory rumor results in overt conflict or violence. Sometimes, there is a simple meeting of minds over rumor where the interlocutors commiserate that 'that's how it is nowadays.' Though they may be destructive in the long run, they often shore up short-term solidarity and a feeling of apprehending the same greater reality as others through the symbolism of the claim or tale. Little action is provoked. Miller's study (1992) 'Snakes in the Greens' examines the emergence of this already-circulated-elsewhere urban legend in a low-income, mostly African-American neighborhood, alleging that an elderly woman was bitten by a snake in a grocery store's produce section, hiding in bunches of greens. Miller shows how the rumor was used – as opposed to being believed or disbelieved, per se – among neighbors in casual conversation. It illustrated how little regard this corporate-owned grocery store had for its patrons, who did not have many options in choosing a place to buy food. Many who heard or shared the tale knew the rumor was 'silly' and likely a tall tale, but it provided a common object to foster a larger discussion of corporate malfeasance and community neglect. There was no desire for official confirmation (or denial), nor any call for a boycott.

Patricia Turner's study (1993) of modern rumor in African-American culture examines how racial oppression and conflict throughout American history have been accompanied by rumor cycles among both blacks and whites, alleging planned and emerging group violence. But consumer legends also are customized to community themes. The rumor that a soda brand popular in minority areas causes sterility in black men, or ones that allege Ku Klux Klan ownership of certain fast-food franchises or expensive sneakers, act as conduits of more long-standing alienation.

Another question about rumor as a whole is raised in these examples: Are rumors generally negative in content? While there is no central database of rumor worldwide that would allow a definitive answer to this question, it does appear that most rumors and legends have a negative core. French researchers are comfortable suggesting that the ratio is likely 9-to-1 in favor of the negative (Renard, 2005). Rumors tend to warn of such things as a new, incipient threat or a secret reality that we have been hoodwinked about. Even positive rumors such as consumer reward rumors, where a rumor suggests that a certain behavior (saving soda can lids, forwarding a certain number of e-mails) will result in the consumer or a charity being rewarded, have a subtle element of negativity to them: either the implication that the finance of ordinary activities is dependent on these sporadic and minor contributions, or that the wealthier entity is 'holding out on' or exploiting the ordinary consumer and the rumor cleverly exposes the secret. There is nothing superficially negative about a chocolate chip cookie recipe that has been shared informally for the last 20 years – but it is wrapped in an interesting rumor. A certain premier department store, it has been said, was serving the cookies and offering to sell its customers the recipe. Assuming the cost would be nominal, the shoppers agreed, only to find on their charge card statements that they had been billed $200 or more for the recipe. The rumor, which unleashed the recipe for free, purported to be a consumer's revenge. The store denied that it had ever served the cookies or sold any sort of proprietary recipe. Everyone agreed that the recipe makes a rather delicious cookie, so it is still interesting that a negative story comes with it (Brunvand, 2012).

Conversely, even some very negative rumors and urban legends contain positive elements, usually intertwined with the act of dissemination itself. In the actual performance of warning rumors and legends, the promulgating of the rumor itself claims to be an act of protection toward others. In Donovan's study of crime legends (2004) the promulgation of the rumor is tied to the desire to keep others safe, even though safety information could simply be hoarded for a personal gain in perceived safety, or a sense of superiority over more foolhardy nonbelievers. This helping stance is apparent both from the text and the talk around the text that went on in online forums. Incompetent criminals and others engaged in antisocial behavior are quite often getting their comeuppance in urban legends, such as the robber identified by a unique injury obtained during a crime in 'The Choking Doberman' (Brunvand, 1984).

Heath et al.'s (2001) study of the content of contemporary urban legends found that there was a significant bias toward the 'disgusting' as a common element of urban legends. They also suggest, from their experimental-method studies of reactions to urban legends, that they can inflict, rather than just reflect, things like disgust, which are broadly shared 'across individuals.' This quality essentially circumvents critical evaluation. Koenig (1985) noted about fast-food contamination rumors that, "Just the thought in the back of one's mind of worms in hamburgers was enough to steer one to the pizza parlor." The fast-food hamburger joint could suffer in sales, even if hearers of the rumor scoffed at its likelihood. These insights may explain the staying power of some everyday rumors – ones that are not particularly bound to crisis or conflict situations, at least not overtly.

So far we have assumed a stability of elements of rumor telling, even if we struggle to characterize it. But are rumor behavior and our relationship to information the same in a society over time? Even before the rise of the Internet, Klapp (1972) suggested that rumor serves each historical circumstance differently. One of the modern appeals of rumor may be offering an overmediated but less trusting society with more raw material for constructing and elaborating rumors. In the case of online networks, we are drawn back to the original question of whether an information vacuum really fosters rumor, or whether in fact information overload or the suspicion of traditional, official sources might spawn or accelerate rumor, as well. This problem may be modern, but appears to have preceded the Internet. Rosnow and Fine (1976) suggested that "in a society with a free and divided press, different sources may contradict one another, which must inevitably add to the confusion and increase feelings of anxiety and fear."

Changing Information Ecology and New Media

There are other historical shifts since the heyday of rumor research that are important in thinking about the spread of

rumors besides the ease and opportunity provided by online communication. Renard (2005) asks us to consider that the ways in which truth and falsehood are currently differentiated, and the means by which we evaluate evidence in public discussions. Renard points to an increased acceptance of *cognitive relativism* where knowledge is 'neither objective nor final.' Reality and fiction are increasingly mixed together in many forms of popular culture, and this includes online interactions.

The Internet has also emboldened skeptics and debunkers of rumor and urban legends, and theoretically could facilitate antirumor activity on a broader scale. The computing revolution has also produced rumor modeling. Based on epidemiological research, modeling can produce practical insights about rumor intervention, and has shed some light on why some approaches work better than others, based on the formal properties of diffusion. Recognizing that attempts to defuse a rumor will include a time delay, Tripathy et al. (2013) simulated three antirumor strategies: the *delayed start* model, the *beacon* model, and the *neighborhood* model. Each of the models has 10 out of 22 'infected nodes' within a network that can spread to adjacent nodes, in the way that a well-circulated rumor can be repeated to a close associate. As the spread of the rumor in these formal models can be countered by the presence of antirumor nodes (representing effective debunkers or at least nonbelievers), the simulation can demonstrate upper and lower limits for saturation, and determine the most formally effective antirumor strategy in a network. The delayed start model launches an antirumor from a single node that travels to all neighboring nodes, after a time lag. This simulates the official response of a single or united authority to a rumor. In the beacon model, an antirumor campaign begins when an infected node in proximity is noticed by the preselected beacon. This model parallels the idea, already in practice in some places such as Barcelona, which trains 'sentinels' in communities that can be activated as needed to quell a rumor. The neighborhood model shows an antirumor campaign launched by random nodes. The delayed start model fared worst: even a rapid incline in antirumor activity allowed 70% of the nodes to be infected at a given time. The beacon model has a shorter time lag, and combated the rumor somewhat more effectively, but only reduced the saturation to 40%, even when the number of beacons was increased. The neighborhood model launches an antirumor process from random sentinels that are next to an infected node, rather than prechosen nodes. This model was the most effective; slightly better than the beacon model in its speed of the total episode. Tripathy et al. suggest that since beacon sentinels are chosen by authorities, there is too much network overlap in the pathways, and random selection avoids this problem somewhat. The analogy here is quite interesting: beacon (sentinel) social networks are perhaps more overlapping than we think. And yet it would be difficult to imagine harnessing an effective antirumor campaign by trusting random debunking. The authors suggest, as other rumor researchers have in the past, that a general social orientation to skeptical-mindedness in the face of unverified rumor might be the most effective approach – though the most ambitious. "The only way that rumors can be quelled," they note "is by using the power of social relationships."

See also: Collective Beliefs, Sociological Explanation of; Conversation Analysis: Sociological; Diffusion in Sociological Analysis; Emotions, Sociology of; Knowledge, Sociology of; Social Influence, Sociology of; Social Networks; Social Psychology, Sociological; Urban Sociology.

Bibliography

Allport, F., Lepkin, M., 1945. Wartime rumors of waste and special privilege: why some people believe them. Journal of Abnormal and Social Psychology 40, 3–36.
Allport, G., Postman, L., 1947. The Psychology of Rumor. Henry Holt, New York.
Bordia, P., DiFonzo, N., 2002. When social psychology became less social: Prasad and the history of rumor research. Asian Journal of Social Psychology 5, 49–61.
Brunvand, J.H., 1984. The Choking Doberman and Other 'New' Urban Legends. W.W. Norton, New York.
Brunvand, J.H., 2012. Encyclopedia of Urban Legends, Updated and Expanded. ABC-CLIO, Santa Barbara.
Buckner, H.T., 1965. A theory of rumor transmission. Public Opinion Quarterly 29, 54–70.
Campion-Vincent, V., 2005(1997). Organ Theft Legends (J. Simpson, Trans.). University of Mississippi Press, Jackson.
Caplow, T., 1947. Rumors in war. Social Forces 25, 298–302.
DiFonzo, N., Bordia, P., 2002. Corporate rumor activity, belief and accuracy. Public Relations Review 28, 1–19.
Donovan, P., 2004. No Way of Knowing: Crime, Urban Legends, and the Internet. Routledge, New York.
Donovan, P., 2007. How idle is idle talk? One hundred years of rumor research. Diogenes 54, 59–82.
Fine, G.A., Campion-Vincent, V., Heath, C. (Eds.), 2005. Rumor Mills: The Social Impact of Rumor and Legend. Aldine Transaction, New Brunswick, NJ.
Fine, G.A., Turner, P.A., 2001. Whispers on the Color Line: Rumor and Race in America. University of California Press, Berkeley, CA.
Firth, R., 2011. Tikopia Ritual and Belief. Routledge, New York.
Goode, E., Ben-Yehuda, N., 1994. Moral Panics: The Social Construction of Deviance. Wiley-Blackwell, Hoboken, NJ.
Gorphe, F., 1927. La Critique du Témoignage [Critique of Testimony]. Librairie Dalloz, Paris.
Heath, C., Bell, C., Sternberg, E., 2001. Emotional selection in memes: the case of urban legends. Journal of Personality and Social Psychology 81, 1028–1041.
Hösch-Ernst, L., 1915. Die psychologie der aussage [psychology of statements]. Internationale Rundschau 1, 15–33.
Klapp, O., 1972. Currents of Unrest: An Introduction to Collective Behavior. Aldine de Gruyter, New York.
Knapp, R., 1944. A psychology of rumor. Public Opinion Quarterly 8, 22–37.
Knopf, T.A., 1975. Rumors, Race, and Riots. Transaction Books, New Brunswick, NJ.
Koenig, F., 1985. Rumor in the Marketplace: The Social Psychology of Commercial Hearsay. Auburn House, Dover, MA.
Miller, D.E., 1992. 'Snakes in the greens' and rumor in the inner city. Social Science Journal 29, 381–393.
Morin, E., 1970. Rumor in Orleans. Heinemann, London.
Nkpa, N.K., 1975. Rumor mongering in war time. Journal of Social Psychology 96, 27–35.
Peterson, W., Gist, N.P., 1951. Rumor and public opinion. American Journal of Sociology 57, 159–167.
Prasad, J., 1935. The psychology of rumor: a study relating to the great Indian earthquake of 1934. British Journal of Psychology 26, 1–15.
Prasad, J., 1950. A comparative study of rumors and reports in earthquakes. British Journal of Psychology 41, 129–144.
Renard, J.B., 2005. Negatory rumors: from the denial of reality to conspiracy theory. In: Fine, G.A., Campion-Vincent, V., Heath, C. (Eds.), Rumor Mills: The Social Impact of Rumor and Legend. Aldine Transaction, New Brunswick, NJ.
Rosnow, R.L., 1980. Psychology of rumor reconsidered. Psychological Bulletin 87, 578–591.
Rosnow, R.L., Fine, G.A., 1976. Rumor and Gossip: The Social Psychology of Hearsay. Elsevier, New York.
Shibutani, T., 1966. Improvised News: A Sociological Study of Rumor. Bobbs-Merrill, Indianapolis, IN.
Suczek, B., 1972. The curious case of the "Death" of Paul McCartney. Urban Life and Culture 1, 61–76.
Sunstein, C., 2009. On Rumors. Farrar Straus and Giroux, New York.

Tripathy, R.M., Bagchi, A., Mehta, S., 2013. Toward combating rumors in social networks: models and metrics. Intelligent Data Analysis 17, 149–175.

Turner, P.A., 1993. I Heard It through the Grapevine: Rumor in African-American Culture. University of California Press, Berkeley, CA.

Turner, R., 1994. Rumor as intensified information seeking: earthquake rumors in China and the United States. In: Dynes, R.R., Tierney, K.J. (Eds.), Disasters, Collective Behavior, and Social, Organization. University of Delaware Press, Newark, DE.

Victor, J.S., 1993. Satanic Panic: The Creation of a Contemporary Legend. Open Court Publishing, Chicago, IL.

Von Roretz, K., 1915. Zur Psychologie des Gerüchtes. [On the Psychology of Rumor]. Österreichische Rundschau 44, 205–212.

Weeks, B., Southwell, B., 2010. The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics. Mass Communication and Society 13, 341–360.

Relevant Websites

http://citiesofmigration.ca/good_idea/fighting-fiction-with-facts-the-bcn-anti-rumour-campaign/ – Fighting Fiction with Facts: The BCN Anti-Rumour Campaign. Retrieved December 1, 2013.

http://www.snopes.com – The Urban Legend Reference Pages. Retrieved December 1, 2013.

Rural Geography

Jo Little, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the development of rural geography, focusing in particular on more recent work on the adoption of cultural perspectives to the study of rural societies and places. During the 1990s a significant body of research and writing in rural geography started to examine the social construction of rurality. This highlighted the importance of idyllic associations of the countryside and of the implications of dominant representations and meanings of rurality to the experience and well-being of rural people. It opened up a major direction in rural geography on the marginalization of 'excluded others' from rural life and on the nature and extent of inequality within rural communities. The article then goes on to illustrate this focus with particular reference to gender and the relationship between rurality and dominant constructions of masculinity and femininity, showing how idyllic rurality has been associated with particularly conservative and traditional notions of sexuality. Finally, the article highlights three current foci of research; food and security, well-being and animals, and explores their significance as central to the current development of rural geographies.

Introduction

Reviews of the history of rural geography as a subdiscipline have generally charted the development of research over time from regional land use studies through political economic approaches to the introduction of more 'cultural' perspectives. In so doing, they note the shift from highly descriptive, empirical research that tended to focus on rural areas in isolation, to more theoretically engaged studies, which connected with and drew on approaches from geography more generally (Cloke, 2006a; Woods, 2005). While this shift was (and remains) highly significant to the emergence of rural geography as a more legitimate subdiscipline with a valuable contribution to make to geographical knowledge and understanding, its importance is now well recognized. The purpose of this review is therefore to concentrate on more recent developments in rural geography and to discuss, in particular, work that has taken place in the past 10–15 years, and that has built on the recent cultural perspectives.

The article demonstrates through the discussion of work on the social and cultural construction of rurality and on the performance of identity, how rural geography has become more theoretically informed and confident. Perhaps somewhat ironically, the theoretical developments that have ensured rural geography's position within mainstream geographical thought, have also reopened questions about the robustness and relevance of the category 'rural,' which warn against an automatic association of particular characteristics and qualities with rural places and people (Cloke, 1997). This is somewhat reminiscent of former criticisms of rural geography when 'rurality' was frequently used as an explanatory variable. As the discussion here shows, however, more recent critiques of the notion of rurality have led, not as in the past to an abandoning of the concept, but to a more theoretically informed examination of the notion of space itself and of the coconstruction of space and space as relational concepts. The article begins by briefly examining the adoption of cultural geography perspectives in the study of the rural. It then goes on to consider in greater depth the subsequent research on rural identities and on the different and multiple experiences of rurality. The article then explores the growing interest in 'alternative' conceptions of rurality in which traditional constructions of rural places and people are challenged by new attitudes and ways of life within the rural.

The Legacy of Political Economy

During the 1980s the application of political economic approaches to the study of rural geography within Western economics had encouraged a focus on structural processes and inequality (Bradley and Lowe, 1984; Cloke, 1980). The development of rural regions and places was attributed to the uneven penetration of capital resulting from global patterns of investment and growth (Marsden et al., 1993; Urry, 1984). Agriculture in particular was seen to be driven by world markets and by political decision-making at an international level. Research in rural geography focused on the economic restructuring that resulted from the global movement of capital, identifying the implications for agricultural businesses at a range of different scales (see, in particular, Buttel and Newby, 1980; Goodman et al., 1987; Marsden et al., 1986). Research also paid attention to the effects of economic restructuring and uneven development on the lives of rural people, and particularly on the well-being or farming families. The major structural inequalities were shown to have very real effects, not only on the distribution of wealth among those working in rural industries but also on the broader access to and control of resources in rural communities. Thus the uneven development of capital was also seen to drive population migration and fuel a meaningful shift in the class composition of some rural areas.

During the 1980s and 1990s there developed a significant body of research and writing in geography and sociology on the nature and implications of an expanding 'new' rural middle class in developed, industrialized economies. Counterurbanization was identified as both a threat to and savior of a traditional rural way of life (Little, 1987; Newby, 1979). While appreciated for their commitment to rural living and for the new capital they brought with them, the middle-class 'incomers' were also blamed for contributing to the poverty of

'locals' by competing for scarce housing resources, neglecting the declining local services in favor of the more distant urban-based provision and resisting any new housing development that might help to increase the housing pool for the less well off. Such research provided important detail on the class changes taking place in rural areas, initiated by national and international patterns of economic change and an increasingly centralized system of decision-making (see Cloke and Little, 1990; Goodwin, 1998). Space (and rurality) was recognized as important (especially in relation to the unevenness of processes of development and capital accumulation) but there was a strong assertion within this research of the economic basis of social inequality. Research moved away from the idea that 'rurality' itself was the cause of the gulf between rich and poor, which had been identified as a significant characteristic of rural life but rather stressed the role of structural processes of class inequality in shaping the lives of those living in rural communities.

Counterurbanization, it was acknowledged, resulted in the gentrification of many rural communities, especially in more desirable commuter zones. Research documented the impact of gentrification on the class composition of rural areas and on their prosperity (showing, particularly in the UK, how wealthy migrants were creating middle-class ghettoes in many parts of lowland rural England). While initial research did recognize the impact of gentrification on the lives of the existing villagers, a more sustained focus on the differential experience of rural counterurbanization and associated issues of welfare and inequality came with the adoption of cultural approaches (Cloke and Little, 1997; Milbourne, 1997).

Cultural Geography and Neglected Rural 'Others'

A widely quoted springboard for new cultural approaches in rural geography came in the form of Chris Philo's 1992 article on 'neglected rural geographies.' In this short piece, Philo challenged the implicit and explicit preoccupation of rural geographers with the experiences of 'Mr Average,' arguing that both the content and the practice of rural geography were dominated by middle-aged, middle-class, white, males. He suggested that this emphasis be redressed in the study of 'other' elements of the rural population and that in so doing, researchers take on board the variety that exists in the composition and experience of the rural population. The fact that such variety has been 'hidden' from both academic and lay discourses of rural spaces reflects the dominance of particular constructions of rurality.

Following Philo's article, there was a concerted attempt by rural geographers to focus on the experiences of previously neglected, and frequently marginalized groups. Studies of, for example, young people (Jones, 1999; Valentine, 1997), black and ethnic minority people (Snipp, 1996), gays and lesbians (Smith and Holt, 2005), the homeless (Cloke et al., 2002) have helped to increase our understanding of the diversity of rural lifestyles. Moreover, many studies of 'other' groups within rural society have attempted to locate their findings within a theorization of difference and in so doing to draw on critical debates in social and cultural theory. These attempts to theorize difference and diversity have demonstrated the existence of multiple ways of experiencing rural spaces. They have also recognized that everyday experiences of those living in rural areas are shaped by the interaction of identity and place (Woods, 2011).

The development of work on 'neglected others' in rural geography was important in its focus on forms of disadvantage that were not entirely class-based. Previously political economy approaches had concentrated almost exclusively on poverty and access to economic resources in identifying marginalized groups and individuals (Lowe et al., 1987). At times poverty was used as a short hand to the consideration of groups such as the elderly and the unemployed but even so the class-based nature of the analysis still suggested that the main lines of difference within rural communities were economic. As the focus shifted beyond poverty to social exclusion, it was clear that the forces of division within rural society were much more varied than simply income, and that issues around inclusion and exclusion were as central to the welfare of rural society and individual rural residents as those of wealth and poverty.

The concept of social exclusion and the attention to marginalized or neglected 'others' provided rural researchers with far greater scope in identifying and understanding the divisions within rural society (Sibley, 2006). It encouraged attention to the detail of rural lifestyles and to the day to day and mundane experiences of those living in rural communities. Such attention allowed researchers to acknowledge that people's individual encounters with rural spaces and societies, their ability to 'fit in,' to enjoy and to flourish in rural areas were potentially determined by aspects of their identity that went beyond their economic prosperity and by the norms and attitudes that accompanied particular lifestyles.

Following the initial focus on neglected others inspired by Philo's article, rural geographers began to explore the broader ideas of identity in general. They moved from being solely concerned with a particular group experience to an interest in the ways different identities were performed and how they 'connected' with rurality. This led to a more dynamic understanding of the relationship between subjectivity and place and of the shifting nature of rural cultures and of the exclusion/acceptance of particular markers of identity (Cloke, 2006b, Woods, 2011). It also included concerns about 'intersectionality' in rural areas, that is, the ways in which experiences of some characteristics of identity were shaped and changed through an interaction with others (see Crenshaw, 1993). Increasingly, rural geographers became concerned about the sense in which the rural gave rise to particular performances of identity. Key questions about whether rural spaces and places created or reinforced certain ways of being, for example, gay or elderly or male. In other words, geographers recognized that at the center of understandings of the particular experiences of different groups was the role of rurality in those experiences. Such questions opened up debates on the *social and cultural construction* of rurality, a recognition that the rural is socially created and shaped, and that rurality is not simply a place but a set of cultural characteristics.

Rurality as a Social Construct

In showing that the experiences of different identities were profoundly affected by space and place, geographers began to

draw attention to the ways in which we understand rurality arguing that it was this cultural construction of the rural that lay at the heart of social exclusion as experienced by some groups of rural residents. The case was made strongly, for example, that the racism experienced by some black people who chose to live or visit rural areas of the UK derived either wholly or partially from the notion of the rural as a white space. The centrality of historical and traditional constructions of (white) rural communities and of the place of the countryside in notions of Englishness serve, it was argued, to exclude those who are believed not to belong in such a space (Kinsman, 1995). Similarly constructions of the rural as a place for families and close-knit social relations (as discussed further below) were seen to reinforce the strong presumption of heterosexuality as the only legitimate form of sexuality in rural areas and the consequent marginalization of gay and lesbian residents.

These dominant and, it was argued, highly powerful constructions of the rural have long featured in rural geographical research, collectively known as the 'rural idyll' (see Short, 1991; Newby, 1979) and comprising what in the UK in particular are very familiar characteristics associated with pastoral landscapes and communities. The characteristics of the rural idyll, while acknowledged to vary in some degree from place to place, have been shown to be not only well known but also very stable and enduring. The significance of the rural idyll became more prominent when research linked both the concept as a whole but also, on occasion, its constituent parts to experiences of social exclusion and marginalization (Bell, 2006). Constructions of rural places and societies were revealed as potentially responsible, in Western economies, not only for the general attitudes toward the countryside – in the UK in particular for the overwhelmingly positive representation of rural lifestyles – but also for more specific responses to social groups and events. It should also be remembered, of course, that in recognizing the power of idyllic constructions of rurality geographers also noted the ways in which they have been regarded as positive by many, particularly in terms of the endurance of the rural community. Representations of the rural community as supportive, friendly, and close knit have encouraged neighborliness and preserved expectations of social contact in a way that has shaped the behavior of, in particular, incoming residents (Little and Austin, 1996), ensuring the continued circulation of symbolic meaning of rurality.

The adoption of cultural perspectives in rural geography has had a major impact on the direction and content of research within the subdiscipline. It has also had a far-reaching influence on the methods employed by rural researchers; on the way data is collected, analyzed, and used. In common with other areas of geography, there has been a significant growth in the adoption of ethnographic methods in rural research and in the use of 'grounded theory' and while such approaches may have characterized some of the early rural 'community studies,' it is only recently that they have been more widely applied to a range of economic, environmental, and policy-based issues in rural studies. Those working on rural lifestyles and marginalization have argued, in particular, that a real understanding of the lives of those living in the countryside needs to emerge from the detailed analysis of their behavior, attitudes, and beliefs and that this can only be obtained through ethnographic methods. In addition, rural researchers have started to make use of a variety of other sources of data, drawing on art, literature, diaries, and other cultural artifacts, both historical and contemporary, in their work. With a shift in the methods used by (some) rural researchers has come something of a change in attitudes toward research. The sensitivity of work on marginalization and deprivation has opened rural geographers' eyes to the potential dangers of research tourism and encouraged them to think more carefully about the way their work may be used to the advantage or disadvantage of particular groups at particular times.

Gender and Sexuality in Rural Geography

One area of research and writing that deserves attention as emerging very strongly from cultural approaches in rural geography is that of gender and sexuality. The experiences of women in the countryside were referred to fleetingly in the early development of rural geography, in, for example, the discussion of labor relations within family farm businesses (Sachs, 1983; Symes and Marsden, 1983) and the sociology of the rural community (see Davidoff et al., 1976). The adoption of political economic perspectives in 1980s was accompanied by (limited) research on gender equality in rural areas and feminist perspectives started to debate the relative importance of class and patriarchy as systems of rural power relations (Little, 1997; Whatmore, 1991). Such work made relatively little headway in identifying women's inequality, however, since (outside agriculture) the plight of middle-class women living in rural areas rarely stimulated any great desire for change or recognition of marginalization. Hence, unlike in other areas of feminist geography, the distinct lack of political will for change among the majority of rural women (and men), acted to detract academic attention away from developing understanding of either the detail of women's lives or wider webs of gender inequality. It was not really until the introduction of cultural perspectives and the examination of the relationship between identity and the sociocultural construction of rurality that much attention was given to the gendered experiences of rural life.

In the early adoption of cultural approaches it was perhaps difficult to see rural women per se as a 'neglected other' in the same way as, for example, homeless people, since, as noted, the majority of contemporary women living in rural areas in developed countries are middle class, many are employed and exercise a significant level of choice in relation to their way of life (Cloke, 2003). It was, however, possible to identify a strong gender dimension in the social and cultural construction of rurality. Idyllic rurality held very firm ideas about the different roles of men and women and it was soon argued by geographers that, underpinning the dominant themes of community and family and the values of friendship, support and self-sufficiency, so powerful within the rural idyll, were particular constructions of gender (Little and Austin, 1996). Conventional rural life, it was argued, rested on the willingness of women in particular to fulfill particular roles, managing the continued existence and well-being of the rural community and ensuring the reproduction of the nuclear heterosexual family. It was recognized that rural women may feel marginalized not through lack of employment or income but rather

through an inability to comply with or buy into the dominant models of womanhood associated with rural life (see Hughes, 1997).

More recently, the debate on gender and rurality has extended to focus specifically on notions of femininity and masculinity in recognition of the importance attached to the dominant ways in which gender is 'done' rather than to fixed structures and relations of power. Research has drawn significantly on the work of social and gender theorists such as Judith Butler (1990) and Elizabeth Grosz (2005) in exploring how masculinity and femininity are not only constructed and represented but also performed in a rural context. Thus emphasis has been placed on the actual enacting of gender and of the values and expectations associated with sexual identities. Initial work in this area concentrated on key characteristics of rural masculinity among farmers and identified how qualities such as physical fitness and strength have come to define the identity of (male) farmers. The work of Brandth (1995), Saugeres (2002), and Liepins (2000) showed how the valorization of certain forms of masculinity within the performance of day to day agricultural tasks has shaped gender relations more widely within farming underpinning the designation of certain activities as 'men's work' and others as women's. Research has also shown how the performance of masculinity and femininity within agriculture is dynamic, changing over time despite their ubiquity. For example, Brandth and Haugen (2000) have noted how expertise in land-based professions in Norway (including agriculture) is no longer exclusively associated solely with men and that women are now acknowledged as having equal capacity for understanding and using the technical knowledge required to perform such expertise. In a different way, Trauger (2004) has shown that performances of masculinity and femininity within recent sustainable agriculture enterprises reflect far greater degree of equality than in more conventional forms of farming. She argues that the increasing number of such enterprises is potentially changing performances of masculinity and femininity in agriculture more generally.

Research on the performance of masculinity and femininity has also shown the ways in which gender identities continue to be shaped by traditional forms of heterosexuality in rural areas, to the exclusion of gay and lesbian identities. Little's (2003, 2007) work on rural dating and relationship formation among farming couples has again identified a range of characteristics and behaviors associated with dominant constructions of masculinity and femininity. For example, in seeking ideal partners, male farmers require women to be practical, homely, and 'low maintenance' and suggest that those who do not conform to more traditional forms of femininity will struggle to 'fit in' to rural communities and lifestyles. Women's expectations of their ideal rural partner also demonstrate highly conventional forms of masculinity in valorizing physical fitness, reliability, and loyalty. This, and similar research, demonstrates the continuing salience of tradition gender identities in the rural and that, even with some recent shifts to accommodate contemporary gender expectations (or, for example, women's involvement in paid employment), conventional ideals of masculinity and femininity continue to hold power in shaping rural social relations.

Recent Directions in Rural Geography

In the final main section of the article, I turn to emerging rural research and consider some of the topics that have recently come to the forefront of rural geography. Of course some of these areas of research are not entirely new in themselves but perhaps have gained a greater prominence in recent years and started to influence the shape and direction of rural research more broadly. In the remaining section, I briefly touch on three related issues of research; food and security, well-being, and animals.

Clearly, the centrality of the study of agriculture to rural geography means that food has long been a well-established topic of research. During the 1990s geographers became particularly interested in 'alternative' food production. They explored the increasingly high-profile concerns around food quality and environmental damage in the context of the use of new production methods inspired by the continued drive for increased productivity. Food safety became a major issue with producers and consumers responding to fears about the growing uncertainties associated with what were seen as 'unnatural' methods of both plant and animal reproduction. For rural geographers, this led to increased research specifically on food consumption (Goodman and DuPuis, 2002; Ilbery et al., 2005; Lockie, 2002) and on the choices being made in search of 'healthy' food produced through 'environmentally friendly' farming methods (Lockie et al., 2002). Research explored the implications of shifts in consumption toward organic and/or local food, fueled not only by the worry about food safety but also wider concerns among both the public and the farming industry about the ethics of food (especially animal) production and implications of new patterns of shopping for food and eating within the family (Jackson, 2010).

The interest in food safety has recently developed in more specific research in rural geography on biosecurity. This research has started to look at a variety of risks and threats to health of humans and animals connected to contemporary food production. Such research has opened up debates on the 'borderlands' (see Hinchliffe et al., 2013) of different global concerns in the production of food, and also questions about the ways in which 'life' itself is valued (Davies, 2013). Understanding the transference of disease together with broader questions about animal lives and welfare have led geographers to focus on what is meant by 'natural' and on how safety is ensured in a context in which nature itself is at risk and uncertain. The development of work from food safety to the broader notion of food security has acknowledged the ways in which questions of consumer preference and concern about food quality cannot be divorced from the more global issue of access to food – whether this be local and national concerns about the price of food and the quality of cheap food to international debates about the possible influence of climate change and limits to food production. Research here has only just started to engage with these questions but this is clearly an area where significant growth is set to take place in the next few years.

An interest in nature and its changing place in our lives also lies at the heart of the two other areas of rural geography research included in this section. Building on the work on food safety and drawing on more established and traditional

research in agricultural and rural geography, is a growing body of research on 'animal geography' (Buller, 2013, Philo and Wilbert, 2000; Matthews, 2012). As with the work on food security, the study of animals by rural geographers is not completely new (geographers have long been interested, for example, in livestock farming and in the contribution of animals to the rural landscape) but the current attention to animals is much more theoretically informed than in the past and has opened up new ways of conceptualizing animal–human relations, challenging normative and supposedly natural dualisms in the distinction of humans and animals. Animal geographies are also concerned with the materialities of how "exclusions, inclusions and physical distance between people and animals are managed and imagined" (Matthews, 2012: p. 127).

Work on animal welfare has drawn attention to ethical dimensions of new animal economies not only in relation to farming and food production but also in wildlife trade and biotechnical research. Parallels are drawn within animal and postcolonial geographies between the treatment of animals and that of colonized, indigenous people and researchers have talked about ways in which their research can give voice as well as speak about the 'other.' Such wok has raised methodological questions about the anthropocentric nature of research on animals and the power relations surrounding human–animal relations. As with work on food security and biotechnology, geographers have pointed to the rich potential of research currently being carried out and that still to be done, stressing that such research is highly important to future ways of living with animals.

A final area of recent research in rural geography that should be mentioned here is the relationship between health and well-being and rurality. The notion of therapeutic landscapes has received attention in the past but, as with the other areas mentioned specifically in this section, recent work has developed new ways of thinking about the topic. Thus recent geographical research has gone beyond the idea that certain environments produce or enhance well-being (by inducing calm and relieving stress, for example) to consider the coconstruction of place and health and look in more detail at the specific connections formed between people and landscape. Such work has drawn attention to the emotional and spiritual dimensions of being in landscapes and focused on the embodied connections made with the landscape/environment in the course of activities such as retreats (Conradson, 2005), yoga (Lea, 2008), and exercise (Little, 2012). Once again 'nature' has been shown to be at the heart of the therapeutic landscapes concept, in both a material and a spiritual sense, encapsulating the restorative qualities of the environment and the body. Therapeutic landscapes are thus regarded as those in which a connection can be made with nature in a return to a premodern engagement with 'pure' nature.

Conclusion

This short article has highlighted some key research in past and contemporary rural geography. It has resisted the temptation to provide a chronological review of the development of the subdiscipline in favor of drawing out more specific research topics. Rural Geography as a subdiscipline has gone through a number of cycles in recent years in which it's influence has waxed and waned within mainstream critical geographical thought. Following a period of relative stagnation, work on political economic approaches and then, more recently, cultural perspectives ensured rural geography acquired a central role in geography's development. More recently, perhaps, rural geography has lost something of its coherence and currency – a reflection perhaps of the changing place of 'the rural' in the public consciousness of those living in Western economies, in particular. Importantly, however, this relative decline in the coherence of rural geography is not reflected in academic research on 'the rural.' This continues to grow and develop in the focus on new and innovative topics and the search for original theoretical and conceptual debates to which rural geography and rural geographers can contribute. Such development has signaled a greater willingness to engage with cross-disciplinary research and to apply ideas and debates from beyond the narrow confines of rural studies.

See also: Geographies of Food: Politics, Place, Power Reprised; Landscape; Rural Sociology; Urban Geography.

Bibliography

Bell, D., 2006. Variations on the rural idyll. In: Cloke, P., Marsden, T., Mooney, P. (Eds.), Handbook of Rural Studies. Sage, London.

Bell, D., Valentine, G., 1995. Queer country: rural lesbian and gay lives. Journal of Rural Studies 11, 113–122.

Bradley, T., Lowe, P. (Eds.), 1984. Locality and Rurality. Geo Books, Norwich, UK.

Brandth, B., 1995. Rural masculinity in transition: gender images in tractor advertisements. Journal of Rural Studies 11, 123–133.

Brandth, B., Haugen, M., 2000. From lumberjack to business manager: masculinity in the Norwegian forestry press. Journal of Rural Studies 16, 343–356.

Buller, H., 2013. Individuation, the mass and farm animals. Theory, Culture and Society 30, 155–175.

Buttel, F., Newby, H. (Eds.), 1980. The Rural Sociology of the Advanced Societies: Critical Perspective. Allenheld Osmun, Montclair, NJ.

Butler, J., 1990. Gender Trouble. Routledge, London.

Cloke, P.J., 1980. New emphases for applied rural geography. Progress in Human Geography 4, 181–217.

Cloke, P., 1997. Country backwater to virtual village? Rural studies and 'the cultural turn.'. Journal of Rural Studies 13, 367–375.

Cloke, P., Little, J. (Eds.), 1997. Contested Countryside Cultures: Otherness, Marginalisation and Rurality. Routledge, London.

Cloke, P. (Ed.), 2003. Country Visions. Pearson, London.

Cloke, P., 2006a. Conceptualising rurality. In: Cloke, P., Marsden, T., Mooney, P. (Eds.), Handbook of Rural Studies. Sage, London, pp. 18–29.

Cloke, P., 2006b. Rurality and otherness. In: Cloke, P., Marsden, T., Mooney, P. (Eds.), Handbook of Rural Studies. Sage, London, pp. 447–456.

Cloke, P., Little, J., 1990. The Rural State: Limits to Planning in Rural Society. Oxford University Press, Oxford, UK.

Cloke, P., Milbourne, P., Widdowfield, R., 2002. Rural Homelessness: Issues, Experiences and Policy Responses. Policy Press, Bristol.

Conradson, D., 2005. Landscape, care and the relational self: therapeutic encounters in rural England. Health and Place 11, 337–348.

Crenshaw, K., 1993. mapping the margins. Intersectionality, identity politics and violence against women of colour. In: Albertson Fineman, M., Mykituk, R. (Eds.), The Public Nature of Private Violence. Routledge, New York.

Davies, G., 2013. Mobilizing experimental life: spaces of becoming with mutant mice. Theory and Society 30, 129–153.

Davidoff, L., L'Esperance, J., Newby, H., 1976. Landscape with figures. In: Mitchell, J., Oakley, A. (Eds.), The Rights and Wrongs of Women. Penguin, Harmondsworth, UK, pp. 139–175.

Goodman, D., Sorj, B., Wilkinson, J., 1987. From Farming to Bio-technology: A Theory of Agro-Industrial Development. Blackwell, Oxford, UK.
Goodman, D., DuPuis, M., 2002. Knowing food and growing food: beyond the production-consumption debate in the sociology of agriculture. Sociologia Ruralis 42, 5–22.
Goodwin, M., 1998. The governance of rural areas: some emerging research issues and agendas. Journal of Rural Studies 14, 5–12.
Grosz, E., 2005. Time Travels: Feminism, Nature, Power. Duke University Press, Durham, UK.
Hinchliffe, S., Allen, J., Levau, S., Bingham, N., Carter, S., 2013. Biosecurity and the topologies of infected life: from borderlines to borderlands. Transactions of the Institute of British Geographers 38, 531–543.
Hughes, A., 1997. Rurality and cultures of 'womanhood.'. In: Cloke, P., Little, J. (Eds.), Contested Countryside Cultures: Otherness, Marginalisation and Rurality. Routledge, London, pp. 123–137.
Ilbery, B., Morris, C., Buller, H., Maye, D., Kneafsey, M., 2005. Product, process and place: an examination of food marketing and labelling schemes in Europe and North America. European Urban and Regional Studies 12, 116–132.
Jackson, P., 2010. Food stories: consumption in an age of anxiety. Cultural Geographies 17, 147–165.
Jones, O., 1999. Tomboy tales: the rural, nature and the gender of childhood. Gender, Place and Culture 6, 117–136.
Kinsman, P., 1995. Landscape, race and national identity: the photography of Ingrid Pollard. Area 27, 300–310.
Lea, J., 2008. Retreating to nature: rethinking 'therapeutic landscapes'. Area 40, 90–98.
Liepins, R., 2000. Making men: the construction and representation of agriculture-based masculinities in Australia and New Zealand. Rural Sociology 65 (4), 605–620.
Little, J., 1987. Gentrification and the influence of local level planning. In: Cloke, P. (Ed.), Rural Planning: Policy into Action? Harper and Row, London.
Little, J., 1997. Employment, marginality and women's self-identity. In: Cloke, P., Little, J. (Eds.), Contested Countryside Cultures: Otherness, Marginalisation and Rurality. Routledge, London, pp. 137–157.
Little, J., 2002. Gender and Rural Geography. Pearson, London.
Little, J., 2003. 'Riding the rural love train': heterosexuality and the rural community. Sociologia Ruralis 43, 401–417.
Little, J., 2007. Constructing nature in the performance of rural heterosexualities. Society and Space 25, 851–866.
Little, J., 2012. Fit farms and fat camps: nature, the body and therapeutic landscapes. Sociologia Ruralis 3, 257–271.
Little, J., Austin, P., 1996. Women and the rural idyll. Journal of Rural Studies 12, 101–111.
Lockie, S., 2002. The invisible mouth: mobilising the consumer in food production-consumption networks. Sociologia Ruralis 42, 278–294.
Lockie, S., Lyons, K., Lawrence, G., Mummery, K., 2002. Eating 'Green': motivations behind organic food consumption in Australia. Sociologia Ruralis 42, 32–40.
Lowe, P., Bradley, T., Wright, S. (Eds.), 1987. Deprivation and Welfare in Rural Areas. GeoBooks, Norwich, UK.
Marsden, T., Munton, R., Whatmore, S., Little, J., 1986. Towards a political economy of capitalist agriculture: a British perspective. International Journal of Urban and Regional Research 10, 498–521.
Marsden, T., Murdoch, J., Lowe, P., Munton, R., Flynn, A., 1993. Constructing the Countryside. UCL Press, London.
Matthews, J., 2012. Compassion, geography and the question of the animal. Environmental Values 21, 125–142.
Milbourne, P. (Ed.), 1997. Revealing Rural 'Others', vol. 15. Pinter, London, pp. 731–756.
Newby, H., 1979. Green and Pleasant Land? Social Change in Rural England. Hutchinson, London.
Philo, C., 1992. Neglected rural geographies: a review. Journal of Rural Studies 8, 192–207.
Philo, C., Wilbert, C. (Eds.), 2000. Animal Spaces, Beastly Places: New Geographies of Human Animal Relations. Routledge, London.
Sachs, C., 1983. Invisible Farmers. Women's Work in Agricultural Production. Rhinehart Allenheld, Totowa, NJ.
Saugeres, L., 2002. The cultural representation of the farming landscape: masculinity, power and nature. Journal of Rural Studies 18, 373–404.
Short, J., 1991. Imagined Country: Society, Culture and Environment. Routledge, London.
Sibley, D., 2006. Inclusions/exclusions in rural space. In: Cloke, P., Marsden, T., Mooney, P. (Eds.), Handbook of Rural Studies. Sage, London, pp. 402–411.
Smith, D., Holt, L., 2005. lesbian migrants in the gentrified valley and other geographies of rural gentrification. Journal of Rural Studies 21, 313–322.
Snipp, C.M., 1996. Understanding race and ethnicity in rural America. Rural Sociology 61, 125–142.
Symes, D.G., Marsden, T.K., 1983. Complementary roles and asymmetrical lives: farmer's wives in a large farm environment. Sociologia Ruralis 23, 229–241.
Trauger, A., 2004. 'Because they can do the work': women farmers in sustainable agriculture in Pennsylvania, USA. Gender, Place and Culture 11, 289–307.
Urry, J., 1984. Capitalist restructuring, recomposition and the regions. In: Bradley, T., Lowe, P. (Eds.), Locality and Rurality. Geo Books, Norwich, UK, pp. 45–64.
Valentine, G., 1997. A safe place to grow up? Parenting, perceptions of children's safety and the rural idyll. Journal of Rural Studies 13, 137–148.
Whatmore, S., 1991. Life cycle or patriarchy? Gender divisions in family farming. Journal of Rural Studies 7, 71–76.
Whatmore, S., Thorne, L., 1998. Wild(er)ness: reconfiguring the geographies of wildlife. Transactions of the Institute of British Geographers 23, 435–454.
Woods, M., 2005. Rural Geography: Processes, Responses and Experiences in Rural Restructuring. Sage, London.
Woods, M., 2011. Rural. Key Ideas in Geography Series. Sage, London.

Rural Planning: General

Hemalata C Dandekar, California Polytechnic State University, San Luis Obispo, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rural planning is about developing and protecting physical and human capital and addressing the at times conflicting goals of economic development and resource conservation. It encompasses both the development of agriculture and natural resources – minerals, forests, fisheries, and scenery – and the conservation of these along with improving the access of rural people to infrastructure, education, housing, and amenities. Rural planning aims to bring about economic development, yet maintain the sustainability of renewable resources and reduce income disparities. It has been variously defined, including that it is place-based in focus and pertains to places that are nonurban, small town, and open spaces. The goals of rural planning post–World War II varied depending on the political ideology of nation states that formed the first, second, and third worlds. Forces of globalization during recent decades have erased many of the international differences in rural planning. The current global challenge of rural planning is to build up and diversify the rural economy, reduce inequity, and exploit the land and natural resources in a way that allows renewal and is sustainable.

Introduction

Elements of what is now termed rural planning can be found historically in the customary traditions and mores that organized agrarian societies, in all their diversity, throughout the world. Rules pertaining to land ownership and its inheritance, customs regarding cultivation practices including crops and tilling cycles, protective practices and rights in forest areas and common grazing lands, division of labor and gender responsibilities, water sharing rights and irrigation practices among others evolved to complement the specificity of social and political organization and climatic and natural resources of particular societies. They were fine-tuned to the microcontext of their constituent communities. Given civilization's historically rural roots, traditions were established which enabled populations to cultivate land and improve the quality of human existence but do so in balance with, and so as to sustain, the area's natural resources and land quality, the base of a traditional village economy.

Development Decades

Following the end of World War II, countries newly liberated from colonialism embarked on centralized, national development planning. Many of these countries were predominantly rural and so planning for rural development was often an integral sector of the national plan. Nations self-identified, or were identified as, first, second, or third world – a categorization some based on 'stages of development' and others on political ideology – to the left, right, or centrist. In the five or so development decades which followed World War II, these economic and ideological differences shaped individual countries' attitudes and approaches to planning in general and rural development in particular. Understandably, there was little consensus across nations on the goals of rural planning. For rural development, considered one of many interrelated aspects of planning, and having a decidedly spatial, locational (nonurban, noncity) connotation, the practices and objectives were many and varied. A general model for rural planning with a widely shared set of objectives and strategies to attain them was elusive. But the governmental concern about the rural reality, regardless of whether or not it was labeled rural planning, was predominantly about the agricultural enterprise, its efficiency, diversity, and ability to not only sustain the rural population but produce surpluses to support urban residents and for trade in world markets. In some locations, the focus was on the exploitation of natural resources more broadly (fisheries, timber, minerals). Their contribution to economic growth was a key indicator of success in the development planning model.

Impact of Globalization

Since the 1990s, the forces of globalization have impacted nations around the world in quite similar ways. They have also resulted, to a large extent, in creating similarities in thinking about, and visioning, rural planning efforts around the world. These globalization forces include an international acceptance of the idea of free global trade and capital flows; large-scale migration and immigration; a strengthening of connectivity and flow of information by way of technology such as the Internet which has served to reduce the 'friction of distance'; the increase in larger scale, often corporate agriculture; the increase in the scale and corporatization of natural resource extraction; a growing recognition of climate change; and the need to adapt and develop strategies which are sustainable, thus creating change in communities to make them more resilient.

There is an emerging 'one world' approach to planning one in which a first-, second-, third-world configuration no longer seems relevant. It is characterized by enhanced cross-fertilization of ideas and approaches, an increase in the flows of people, ideas, and culture around the world; and the evolution of national economies which privilege those with knowledge and technical expertise. In this globally integrated system, the characteristic attributes of first, second, and third worlds are juxtaposed within a country. They are to be found in

the location of people with resources and those without, within regions, within urban areas, and increasingly within rural areas too. Throughout the world, the notions of rural development have been transformed, and the issues of disparity and unequal distribution of assets and income are now emerging as priorities in the conditions of the rural. The industrialization of agricultural cultivation, processing, and natural resource extraction has disconnected the historically immediate and physical connection of rural communities from their economies. The emerging call in rural planning has been for developing sustainable practices to keep production in balance with replacement so as to maintain the resource base, strive for equity, and allow an adaptation to climate change.

Defining Rural Planning

Scanning the many variations of rural planning that are practiced internationally, it is possible to find some themes and an emerging congruence in aspirations. Rural planning encompasses policy and programs that explicitly or implicitly affect economic, social, and physical aspects of rural conditions. Rural planning has been defined as a set of two complementary concerns related to how people will earn a living, and how they will do it so that the natural, environmental, and social resources of a place will be conserved. Rural planning is about physical and human capital in regions that do not have the characteristics of urban places such as density, municipal governance, and a critical clustering of services and amenities. Rural planning addresses issues related to agriculture and also, more broadly, natural resources (fisheries, timber, and minerals). Traditionally, the concerns of rural planning were distinct from those of the urban, but increasingly, as the world has become more interconnected, there appears to be convergence in the identification of the most salient issues, and strategies and plans to address them. Increasingly in rural planning, in addition to the concern about equity and opportunity in urban versus rural contexts, there is a concern about the increasing economic gap between the affluent and those in jobs that pay minimum wages. Often these are related to, in countries of high immigration such as the United States and northern Europe, the legality and entitlement of residents who are documented and legal versus those who are not. Thus issues that were in the past thought to be concerns in urban contexts have increased in significance in rural contexts too and served to broaden the concerns of rural planning beyond agriculture to human resource development and infrastructure improvement.

Rural Planning in the Development Decades (1940s–80s)

During the development decades following World War II, approaches to rural planning in the United States, India, and China were based on quite differing ideologies about the role of rural areas in national development. The modern farm, which evolved in the United States with assistance from research at land grant universities and extension services funded by the government, was emulated worldwide. It involved the use of high-yield hybrids, fertilizers, pest controls, and extensive mechanization on a large scale to serve national and international markets. The approach of modernizing agriculture to attain economic development and growth in rural areas was tempered in Third World countries with a concern for attaining social justice through a redistribution of productive rural assets, and thus the benefits, accruing from development. Consequently, rural planning in second and third world countries explicitly embraced notions of equity and redistributive justice.

The heading 'rural planning' as a sector of government involvement is sometimes absent from plan documents in many countries. But policy and legislation related to sectors such as agriculture, transportation and communications, education, health and human services, and social welfare, which were addressed in national or state plans, have had far-reaching and profound impacts on rural areas. They have impacted the physical conditions of rural areas and shaped the availability of infrastructure and jobs. They have also determined the range and quality of services accessible to rural populations. When headings such as rural planning, rural development, and integrated rural development are present in plan documents, as discrete sectors requiring attention and having related funding lines in national and state budgets, these sectors are often defined in terms of engagement with social welfare agendas. The goals of such sectoral policies have often included statements about the need to redress inequalities in access to public services and redistributive justice. A broad definition of rural planning as encompassing policy that has both explicit and implicit intention to impact rural conditions, economic, social, and physical is adopted here.

During the formative period of the United States in the nineteenth century, agriculture production was a central element in its robust growth and development. Similarly, an agriculture-based economy and a predominantly agrarian population was the legacy of heavily populated, newly independent, countries like China and India, which faced the task of nation building in the second half of the twentieth century. But over the years, these countries have differed significantly in their approach to rural planning. The interventions of their governments in the rural sector define a spectrum of strategies and approaches, from total centralized control to indirect, market-driven change. Furthermore, these strategies, which took divergent paths at the start of nation formation, have over time sharply changed direction.

The United States, India, and China provide examples of countries that have developed rural planning-related policies based on quite differing ideologies about planning and the role of rural areas in national development. Their positions have ranged from the free market, laissez faire orientation typical of 'First World' countries like the United States to policies of 'Second World' countries like China that initially adopted a centralized, rural-cooperatives based model of rural planning from the former Soviet Union. China later embraced a market-based, global trade and industrialization focused approach, the success of which has made it a key competitor for dominance in the global economy. It has adopted policies that aim at radical reform and acquiesced to the notion of individual land ownership and private property.

Seeking a more intermediate path between these two extremes were countries like India, a founding member of the bloc of so-called 'Third World' countries that sought to chart

a distinct path, other than the United States or former Soviet bloc, first- and second-world approach to development. In the last few decades, it too has adapted and adopted policies that allow a greater play of market mechanisms in the rural, agricultural, and resource sector. Examining the development of the rural planning concept in countries like the United States and India provides a glimpse of the differing interpretations of rural planning that have been evolving around the world in countries where the government's ability to control, plan, and implement has been limited by, and has been cognizant of, the significant role of the private sector in the farming endeavor. Study of the China model provides an insight into the strengths and limitations of sustained, centralized efforts to control developments in rural economy and society.

The US Model of Rural Planning

The United States is acknowledged to be one of the most open and laissez faire economies in the world. Within the United States, the idea that attention needs to be turned specifically to rural planning is not a widely embraced concept. But models of economic growth that guide policy are accepted, and in these the agricultural sector is factored as an integral component of the national economy. Historically, US agricultural policy has defined and shaped rural conditions and had far-reaching consequences on the development of rural areas, but it has not been recognized as an instrument for rural planning.

This American ambivalence toward acknowledging the existence of rural planning is understandable given the historical antecedents. After the American independence, the feudalistic land law that prevailed under the British rule was replaced with a freehold land tenure system under which public lands were sold to raise cash for the new government. To facilitate sales, the US Congress ordered a land survey in 1785 that established a grid system of land division as the dominant pattern for the area west of Pittsburgh. Its indelible stamp ordered the transportation and communication systems and helped give shape to rural administrative areas. Subsequent land laws and purchase allowed "Such a mass distribution of land to those who tilled the soil was unique in the history of the world up to that time" (Cochrane, 1993, p. 57). The philosophy of land ownership as vesting freehold in an individual, and the establishment of a unique American tradition of independent and self-reliant farm families scattered in a regular pattern, each in isolation, on their own individual pieces of land, has perhaps contributed to a national antipathy to identifying government interventions as responsible for rural development. Nonetheless, these early interventions shaped the rural context, with plans that were not always openly agreed to or publicly articulated and embedded it with desired outcomes that might have been contested if formally acknowledged in planning documents.

The movement of the US agricultural enterprise from the subsistence cultivation of the pioneer to the high-tech, mechanized, computerized, agro industry of today's corporate farm is not the product of free-market forces but of significant government actions. The role of the government in shaping US farming is noted and generally accepted in the literature pertaining to American agriculture (Rasmussen, 1960; Cochrane, 1993; McConnell, 1953). They delineate a broad and inclusive set of categories in which government has been active, establishing that it has been involved in almost every aspect of agricultural development. With the passage of the Agricultural Adjustment Acts of 1933 and 1938, the Department of Agriculture has supported and managed agricultural production pursuing policies affecting both the supply and distribution systems of agricultural production (Cochrane, 1993; pp. 140–143).

Modern Farming in the United States

Government activities in the transport, research, and education sectors have transformed American agriculture from subsistence to cash cropping for regional and national markets and later to capital-intensive agribusiness. Significant government support of educational and research institutions assisting in agricultural development through scientific and technological innovations has been sustained for more than a century. The Morrill Act of 1862 enabled government to fund land grant colleges in each state and the Hatch Act of 1887 allowed annual allocations for state agricultural experimentation stations devoted to basic agriculture research (Rasmussen, 1960). The impact of this activity on shaping the scientific base of modern American farming and increasing productivity and efficiency was significant and to date has had resonance and wielded influence around the world.

In response to the worldwide Depression of 1929, New Deal agricultural economists created the Agricultural Adjustment Acts of 1933 and 1938 to set the price and allowable level of farm production. The central theme in these legislative efforts was to stabilize agricultural prices and income by offering subsidy payments to farmers to reduce production, price support measures, and commodity storage. In the postwar period from 1945 to 1975, the modern farm evolved in the context of rapid advances in national/international transport, the emergence of the United States as a major world power, and the corresponding change in its domestic and foreign policy, and tremendous advances in scientific/technological farming in the United States including development of high-yield hybrids, fertilizers, pest controls, and extensive mechanization. The modern farm featured large-scale production for national and international markets. Cultivation for sale rather than family subsistence became the central production objective of the farm enterprise. However, the ideological underpinnings of the earlier time of minimum government control in the farm enterprise persisted. There was support for government to be an agent responsible for presenting new options and alternatives which might voluntarily be subscribed to, or not, by the American farmer.

Corporate Farming in the United States

A shift from modern to corporate farming after 1975 was triggered by changes in the role the United States played in world affairs and the tack it took with respect to its foreign policy following World War II. The United States emerged as a leading proponent of global free trade. This posture meant that the highly efficient farms in the United States began to function as the 'world's breadbasket.' Farmers produced and

exported crops to India, Russia, and other countries when they faced domestic shortfalls in food production. Advances in national and international transport aided this expansion of the market. 'Green revolution' technology packages of high-yielding hybrids, superfertilizers, and pest controls were developed leading to unexpectedly large crop yields. Specialization and monocrop farms become common. Although energy intensive in their use of new products and machinery, they were viable as energy costs were kept relatively low by the government. A process of consolidation of family and modern farms into larger and larger vertically and horizontally integrated corporate entities occurred. The resulting corporate superfarms have changed the rural landscape tremendously as fewer and fewer farms account for more and more of the total farm sales and total farm profits in the United States. This concentration is apparent in the fact that 2% of all farms in the United States produced 50% of total farm sales in 1997 and by 2007 some 187,816 of the total 2.2 million farms in the United States accounted for 63% of sales of agricultural products and 125,000 farms produced 75% of sales.

Services-Oriented Rural Planning

In the 1980s and 1990s it has been increasingly obvious in the United States that although the link between agriculture policy and rural planning has traditionally been very strong it may be less important in the emerging 'global' economy of the twenty-first century. Today the link that needs to be forged for rural planning is with economic sectors such as manufacturing and services. The emerging 'new economy' is based on the growth of the service sector and geared to the information industry that has centralized population in metropolitan areas. The impact of this centralization on populations that remain rural and dispersed is significant. For example, in rural areas, public services in education, health, and social welfare provision are of lower quality than in the metropolitan areas. Currently, US rural planning efforts have focused on providing quality services despite the delivery difficulties posed by lower population densities that mean that greater distances must be covered to reach targeted rural populations. A second strand of rural planning pertains to national settlement patterns and the belief that it is important to maintain numbers of people in rural areas. Planning efforts creating rural enterprise and empowerment zones have involved job training programs as an economic development policy to attract entrepreneurs. The second element of this strategy aims to reduce costs for entrepreneurship in rural areas by programs that make capital available for business development. The scale of such efforts is, however, quite modest. Finally, the current paradigm of rural planning focuses on equalizing rural communication and infrastructure with those of the urban. Supporting the expansion of Internet and high-speed communications links, wireless, and satellite connections all aimed at linking dispersed populations are an integral part of this agenda.

The Third World Approach to Rural Planning

Numerous countries attained independence from European colonial rule at the end of World War II. Some, like the Peoples' Republic of China, became Second World countries that adopted planning models emulating Soviet-style control by a strong central government. Another bloc of countries attempted to forge a separate third path to development differing from United States-style capitalism and Soviet-style socialism. This Third World forged constitutions endorsing centralized planning by government to attain modernization by facilitating rapid industrialization of both the private and public sectors. In these models, rural planning tended to be a key element in national economic development. Rural planning measures aimed not only to increase production but also to redistribute wealth on the grounds of social justice and human welfare. The latter provided legitimacy for governmental planning which was essential in countries recently released from colonization. Furthermore, the majority of these nations' populations were rural. In approaching the development of rural areas, many of these countries emulated the US model. They aimed to increase and improve agricultural production through the application of scientific research that was to be developed, disseminated, and implemented in farming practice through agricultural extension services. This was seen as the modern way for rural planning to support national economic development.

Third World countries aimed to use agricultural production to raise capital for industrialization and the development of related infrastructure, as well as for the provision of urban- as well as rural-based populations. However, they tempered their desire to increase agriculture production with the broader objective of bringing about an increased living standard and quality of life for economically disadvantaged populations in the countryside. Their emphasis on rural planning for human development reflected their recognition that the majority of their populations were rural, asset poor, involved in subsistence production, and members of lower-status social groups or social structures that were feudal in character. Over time, these countries evolved a planning model for 'integrated rural development' that recognized gender and class differences and attempted to address social inequalities in access to rural resources, a process differentiated from earlier production-oriented agricultural development models. This approach to rural planning attempted to combine and balance between agricultural and economic development and social welfare and redistributive justice.

The Indian Model

At the point of independence from British rule, India embraced rural planning directly. Initial Indian national planning efforts to achieve rapid socioeconomic change concentrated on industrialization at chosen metropolitan centers, on the assumption that the benefits would 'trickle down' through the economy. Planners believed that although initially benefits would be concentrated in the core cities, they would spread to rural areas. However, they also recognized the need to promote complementary rural planning to reduce existing and anticipated urban–rural disparities in wealth.

Rural planning measures in India conspicuously emulated the US model of agricultural development. They included irrigation projects; the promotion of scientifically developed

cash crops through the provision of improved seeds, equipment and fertilizer (the so-called Green Revolution approach to making agriculture productive); providing credit for agricultural investments to stimulate agricultural production; and the establishment of a system of extension services to introduce scientific farming techniques at the village level. They also regulated markets to stimulate trade in agricultural commodities; invested in roads and communications; and founded cooperatives to buy and sell agricultural products and facilitate the marketing of goods in rural areas.

The national policy of investing in industry and commercial agriculture increased the gap between those involved in the corporate formal sectors of the economy, urban and rural, and those in the informal, traditional, subsistence sectors. Rural landless and land poor, squeezed out of traditional positions in village society, migrated to the city in unprecedented numbers. Cities were unable to accommodate the massive influx of displaced, often unskilled, rural laborers and their families and urban environments deteriorated as slum settlements of the poor proliferated. Responding to this phenomenon inspired a new set of rural planning policies. Stemming migration to the city by improving rural life and diversifying the rural economy emerged as prime planning objectives. Making rural living more attractive included provision of postal service, healthcare facilities, clean drinking water, and electricity in the countryside. As this policy evolved, its objectives also expanded to include increasing equity and distributing the benefits of development to the poor more explicitly. This change inspired such planning measures as putting a ceiling on individual land holdings, reforming tenancy to enhance the rights to the land of those who cultivated it, initiating programs to support traditional artisans with better tools, establishing credit and marketing networks to allow them to compete with industry or at least sustain them until alternative jobs were created. Finally, the Indian government emphasized free, mandatory elementary school education throughout rural areas to improve skills the poor needed to compete for development benefits.

In 1976, A.T. Mosher an international agricultural development expert summarized the various phases of Indian rural development planning in *Thinking about Rural Development*. He described some common characteristics:

- recognition that although economic development is important it is not an essential precondition to the social transformation that is considered the central objective of a rural development strategy;
- attention to increasing efficiency with the technical revolution of fertilizers and high-yield varieties of cereals; the attempt was to boost both food grain production and commercial crops and monoculture production of commodities for which there was an international market;
- development of programs that addressed equity and asset redistribution; and
- enhancing rural citizen's participation in the evolution of locally appropriate planning programs.

Current Priorities and Approaches in Rural Planning

Mosher's insightful observations made several decades ago remain salient internationally in the current discourse on rural development and rural planning. Elements of the phases of rural planning policy and objectives in India – moral and social transformation, economic efficiency, redistribution and social justice, and popular participation in decision making – are to be found in the current rural planning approaches of many countries around the world although there is of course a great deal of variation in policy, in emphasis, and in the capacity to implement. There has emerged one key and significant addition to the above list of characteristics or rural planning discourse, that of the notion of sustainability – social, resource-based, and economic – of adopted approaches to, and practices of, rural planning.

Rural planning oriented to improving physical infrastructure and the quality of life for rural populations was an important policy agenda in developing countries in the earlier periods of their postindependence histories. More recently, as rates of urbanization have remained high, these countries and cities have experienced hypergrowth (despite efforts to stem rural migration and population increases); rural planning in many countries has become a lower priority endeavor. In the worldwide liberalization of economies since the 1980s and 1990s, attention has been focused on world cities of over 10 million people, and, as a result, notions of intervention in rural areas to improve rural conditions have been given far less significance.

A questioning of 'top-down' science-driven Green Revolution strategies and of intensified agricultural cultivation has emerged as the negative impacts of this approach on the natural environment have become increasingly apparent. A call for 'sustainable agriculture' is heard and policy to achieve it is proposed (Cochrane, 2003, pp. 65–82, 98–101). The notion of sustainability in agriculture, forestry, mining, and other interrelated sectors is also apparent. The policies and plans which are suggested to achieve this goal of sustainability incorporate some practices and values of earlier traditional modes of cultivation.

The concerns about the long-term negative impacts of the Green Revolution and its consequences have surfaced in many parts of the world. They are related to degradation of soil quality; depletion of groundwater tables; the salinization of land; and high fertilizer use and pollution of water systems from agricultural run-off. The long-term sustainability of modern agricultural cultivation on prime land is questioned, and as supply of such land is exhausted, concern about the viability of applications of this technology package on poorer quality, more fragile lands is raised. There is a rising concern about genetically modified crops entering the food system of human populations. These concerns have led to calls for a modified, mixed approach to rural development that recognizes traditional, sustainable practices. In contexts of risk-prone environments, development planners are describing a need for a rural planning strategy in which 'top-down' science-driven Green Revolution strategies of intensifying agricultural cultivation are tempered with the understanding that can be attained by tapping the knowledge of rural cultivators who are following traditional, self-sustaining forms of agricultural production (Pichón et al., 1999). This position, labeled 'sustainable agriculture,' has stimulated rural planning efforts based on more traditional modes of agriculture (National Research Council, 1989). The task of helping rural places has

been defined as a set of two complementary concerns related to how people will earn a living (Lapping et al., 1989) and how they will do it so that natural, environmental resources will be conserved (Sargent et al., 1991).

In the First World rural planning is coming a full circle. Rural areas are serving as an escape and retreat from the pressures and stress of cities for the affluent and as a source of raw materials and natural resources for urban areas and industry. The corporatization and large-scale production of agricultural enterprises has required less labor and resulted in fewer jobs in rural areas. From a 'universalistic' modern science-driven notion of development and progress, there is a renewed orientation to the specificities of the local, rural, asset base of both physical and human capital. Rural planning is construed to involve development and support of agriculture and a broader set of natural resources including fisheries, timber, and minerals and the conservation and protection of these resources with a view to sustainability and conservation. The challenge for rural planning has been to answer the question of how to rebuild a local economic base and sustain rural communities. Some high-amenity places – coastal communities, beach-front towns, and sites of historic and scenic natural resources such as lakes, mountains, and forests – have capitalized on tourism and amenity migration of higher-income residents.

The literature on planning amenity-led development is substantial (Marcouiller et al., 2002). In many communities, along with amenity migrants, in parallel, has been the in-migration of low-wage workers filling jobs to provide the services that amenity migrants need and demand. High housing costs in these amenity-rich locations compel these service workers to commute and find affordable housing and make community in settlements that are away from the amenity core in less desirable locations (Nelson and Nelson, 2010). They bear the cost of travel time and expense of their commute for work. This has raised equity issues.

Other options for rural development have included efforts to establish rural manufacturing and back office support such as call centers and agricultural processing such as meat packing plants for poultry and beef, all of which also tend to create a majority of jobs for low-wage workers. These trends were recognized in the 1994 creation of USDA Rural Development to administer nonfarm financial programs for housing, community facilities, water and waste disposal, rural businesses, and rural electric (Rural Cooperatives, 2004). Rural policy in the twenty-first century also acknowledges the need to support human resource development, improve education levels and skills, enhance telecommunications infrastructure to improve connectivity and opportunity for rural residents (Marshall, 2001).

Today, the first/second/third world configuration that shaped mid-twentieth century rural planning strategies seems less relevant. First and third world-like economies coexist in cities around the world, not only between nations, and, increasingly, that duality is emerging in rural areas too. This trend is global. A 'one world' approach to planning in general is emerging with the forces of globalization that are creating systemic similarities around the world. The cross-fertilization of ideas and approaches has been ever facilitated by the increased flow of communication that technological innovation has enabled. In much of the developing world, facing the waves of change engulfing countries in the throes of economic liberalization, rural planning has been placed much lower in priority to programs that address issues of rapid urbanization and globalization. Thus programs related to social justice, participatory decision making, sustainability, and the environment have taken second place to economic transformation of rural production. But there is a concurrent literature and policy emphasis in rural planning that is place-based and enjoying resurgence due to concerns about sustainability and resilience of settlements and communities (Olfert and Partridge, 2010). Thus, the inherent tension between growth and sustainability that emerged toward the latter part of the twentieth century in the thinking about national development planning is increasingly apparent in the ways that rural planning is currently being defined and framed around the world.

See also: Planning During Post-socialism; Planning and Development; Urban Informality: The Production and Regulation of Space.

Bibliography

Cochrane, W.W., 1993. The Development of American Agriculture: A Historical Analysis, second ed. University of Minnesota Press, Minneapolis, MN.

Cochrane, W.W., 2003. The Curse of American Agricultural Abundance: A Sustainable Solution. University of Nebraska Press, Lincoln and London.

Lapping, M.B., Daniels, T.L., Keller, J.W., 1989. Rural Planning and Development in the United States. Guilford Press, New York.

McConnell, G., 1953. The Decline of Agrarian Democracy. Atheneum/University of California Press, New York.

Marcouiller, D., Clendenning, J., Kedzior, R., 2002. Natural amenity-led development and rural planning. Journal of Planning Literature 16, 515. http://jpl.sagepub.com/content/16/4/515.

Marshall, R., January 2001. Rural policy in the new century. International Regional Science Review 24, 59–83.

Mosher, A.T., 1976. Thinking about Rural Development. Agricultural Development Council, New York.

National Research Council, 1989. Alternative Agriculture. National Academy Press, Washington, DC.

Nelson, L., Nelson, P.B., 2010. The global rural: gentrification and linked migration in the rural USA. Progress in Human Geography 35 (4), 441–459.

Olfert, M. Rose, Partridge, M., 2010. Best practices in twenty-first-century rural development and policy. Growth and Change 41 (2), 147–164.

Pichón, J.F., Uquillas, J.E., Frechione, J., 1999. Traditional and Modern Natural Resource Management in Latin America. University of Pittsburgh Press, Pittsburgh, PA.

Rural Cooperatives, May/June 2004. USDA Rural Development: Committed to the Future of America's Rural Communities. pp. 42–44.

Rasmussen, W.D., 1960. Readings in the History of American Agriculture. University of Illinois Press, Urbana, IL.

Sargent, F.O., Lusk, P., Rivera, J.A., Varela, M., 1991. Rural Environmental Planning for Sustainable Communities. Island Press, Washington, DC.

Rural Social Work

Margaret Alston, Monash University, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rural practice is a critical emerging social work field. While social workers have always worked in rural areas, climate change, environmental disasters, and the ongoing impacts of globalization are placing significant stress on rural people and communities, and leading to a new evaluation of the environment and its significance for social work. This article examines rural policies and practices that challenge and disrupt effective social work services into rural communities. It also highlights that rural social workers adopt innovative and creative practice models to deal with isolation and other barriers, thus creating the space for a reimagined and radical social work.

Introduction

Rural social work is a critical, and increasingly prominent, field of practice. While social workers have always worked in rural areas undertaking a number of roles in a variety of areas of practice, two key factors are shaping a new form of practice interface – the link between environmental crises and disasters and globalization. The first relates to ongoing climate and environmental disasters and includes both slow onset events such as droughts and desertification, and cataclysmic events such as earthquakes, fires, floods, river erosion, and mud slides all of which have had major impacts on landscape and significantly changed the social relationships and industry base of many rural communities. Add to this globalization – a process that is drawing industry, people, and wealth away from rural areas for which the result is environmental, economic, and social decline overlaid with climate uncertainty leading to stressed and highly vulnerable communities.

In this article I present the view that rural social work is a complex field of practice requiring a critical understanding of the link between people and their environment shaped by the social, economic, environmental, political, and cultural factors that identify rural spaces. Across the world rural areas are sites of ongoing change, typified by growing uncertainty, food and water insecurity, and evidenced by increased poverty, an aging profile, and inefficient or nonexistent services. In this context, social workers must be equipped with a rich understanding of the link between people and place (Zapf, 2009) and the structural inequalities that reduce the political power of rural people to address their disadvantage.

Defining the Rural

Since 2008, and for the first time in history, a majority of the world's population now resides in urban areas (UNFPA, 2007). Politically, this growing urbanization reduces attention to, and the resourcing of, rural areas. Nonetheless, rural places are as diverse as city suburbs and this makes an adequate definition of 'rural' very complex – a conundrum that the UN largely leaves to nation-states. The way rurality is understood and experienced in different countries is shaped by social, cultural, and political factors (Pugh and Cheers, 2010). Most see the 'rural' as places 'out there' away from the cities, a view that leads countries such as Australia to define rural spaces in terms of what they are not – that is, anything that is 'nonmetropolitan.'

Attempts at conceptualizing rurality more carefully usually center on one of three bases: a spatial element that draws on population numbers and size of communities as a unique indicator (see, for example, Hugo, 2005); an economic definition that is based on the industry base of the area (typically agriculture or mining) (Alston, 2009); or a reliance on sociocultural factors that denote somewhat controversially that rural communities are typified by shared relationships and values – a rural idyll that has often been portrayed as morally superior (Leipins, 2000). None of these definitions are entirely satisfactory either together or in isolation. Perhaps it is best to adhere to Rios's (1988: 1) comment that *people know when they are rural!*

Setting and Context

Nonetheless, the inherent complexity of rural areas should not be diminished. A common reminder in rural literature for practitioners and policymakers is that "if you've seen one rural community, you've seen one rural community!" (Raitano, 2010). In other words, rural communities are never the same, varying as they do in size and scale, in industry base, in population composition, in landscape detail, and in the degrees of socioeconomic complexity. Nonetheless, the further a community is from a capital city or regional base, the less likely it is to be economically viable, the less likely it is to have efficient, available services, the more likely it is to have an aging and masculinized population profile, the more likely it is that working aged people will have at best a precarious attachment to the labor force, the more likely it is that mortality and morbidity rates are higher than in urban areas, and the more likely it is that there are large numbers of socially excluded people including higher numbers of indigenous people (Cheers and Taylor, 2005; Foster, 2007; AIHW, 2008; Alston, 2009). The reasons for these trends include rising levels of poverty, the largely agricultural base in more remote communities, patrilineal inheritance practices in agriculture, the out-migration of young people for work and education, the greater out-migration of young women, the decline in services as these

are shifted to regional centers and cities, the loss of jobs and businesses, the higher rate of risk-taking behaviors, and the greater exposure to environmental hazards (Alston, 2010a, 2009; Alston and Kent, 2006; Gray and Lawrence, 2001).

Where these trends may vary, if only temporarily, might be when a mining industry is established and a quick development of workforce occurs. These rapidly expanding mining communities can be stressed for different reasons including a fly-in fly-out, largely male workforce, low levels of services, high housing prices, a skewed labor force where lower paid jobs outside the mining industry can be harder to fill, (Petkova et al., 2009) and high levels of environmental degradation (Dominelli, 2012).

Social workers will note that the challenging dynamics in rural communities described above provide fertile ground for a range of social outcomes and emerging social problems. These include complex multifaceted social problems, involuntary separation of families if a family member is forced to out-migrate for work and/or education, the precarious nature of employment options, health issues, aging, and poor access to services. When we add climate uncertainty and globalization, these problems are magnified (Alston, 2009; Alston and Kent, 2006; Gray and Lawrence, 2001).

Critical Factors – Climate Uncertainty and Globalization

Global warming is having an unprecedented impact on our environment, landscape, and social spaces and cuts deeply into rural areas. Slow onset events such as drought and rising temperatures affect cropping cycles, the types of crops that can be grown, the capacity of farmers to produce an income, and both food and water security (Garnaut, 2008, 2011). Cataclysmic events such as cyclones, fires, and floods cause deaths and injuries, decimate communities, affect agricultural and fishing industries, destroy infrastructure and income generating activity, and change the landscape, often irrevocably (Opdam et al., 2009; Campbell, 2008; Proudley, 2008).

Across the world, the increasing frequency and intensity of climate and environmental disasters are deeply affecting rural people and places. A range of recent disasters causing major social trauma include flooding in the Philippines (Tatlonghari and Paris, 2012), bushfires in Australia (Hargreaves, 2012; Proudley, 2008), drought and cyclones in Bangladesh (Alston, 2013), hurricanes in the US (Enarson, 2009), rising sea levels in the Pacific (Alston and Vize, 2010), and earthquakes in Japan (Barusch, 2011) and New Zealand (Lynch, 2011). Many of these events are located in rural areas where whole communities are affected by trauma and tragedy. Social workers should note that climate change is not only causing environmental disasters, but also immense social distress and thus is a human rights issue (Tester, 2012).

Climate change events are but one of a suite of social, economic, and environmental factors that will impact on food producing areas. With the world population expected to rise from 7 thousand to 9 thousand million by 2040 (The United Nations Secretary General's High-level Panel on Global Security, 2012), creating significant consequences for food and water security; a growing middle class in Asia and the subcontinent placing more pressure on food production (Gillard, 2012); predicted fuel shortages (Newman et al., 2009); conflict and political instability in many areas; and massive people movements – resulting from these factors (McAdam, 2012), the stresses on the environment and on people in rural areas are magnified. By 2030, the world will need at least 50% more food, 45% more energy, and 30% more water despite the fact that food production cycles are interrupted by climate events (The United Nations Secretary General's High-level Panel on Global Security, 2012). Rural spaces will become the locus of major transformations, instability, and ongoing environmental pressures at the same time as the world requires significantly more food and water resources.

Globalization, or the instigation of a global marketplace creating global interdependencies, is drawing wealth and capacity from the local to the global arena (Gray and Lawrence, 2001). In the process, it is providing a compelling influence for people to move to cities to work, often in jobs based on the knowledge economy. Globalization has resulted in significant changes in rural areas as agricultural producers are now dependent on a world market, their income subject to global circumstances and their capacity to respond to higher productivity demands diminished (Lawrence, 2005). Without adequate policies and programs, communities plagued by poverty, climate challenges, and out-migration pressures are ill-equipped to address future uncertainties.

Part of the problem is that there are limited attempts to create complex interlinked rural policies that cross-portfolio areas. Thus, it would seem evident that enhanced information technology will assist health delivery in isolated areas with low levels of services. Yet, in Australia at least, many areas are without mobile phone coverage and significant portions of rural Australia are not covered by technology that will assist service delivery. Additionally, with the centralization of services away from rural areas it would seem obvious that cheap, readily available transport is necessary to ensure that access is not corrupted by distance. Again, this is not the case, leaving the impression that, as the political influence of these areas decline, rural communities are abandoned and the citizenship rights of rural people are overlooked under the banner of efficiency and neoliberal economic philosophy.

Rural Policy

The decline of the welfare state and the rapid rise of neoliberalism across the world have resulted in a downgrading of social supports at the same time as principles of individualism and self-sufficiency are championed by governments. Neoliberalism is very much focused on market values, leaving those outside the labor force in an increasingly precarious position. Thus the aged, disabled, carers, and many rural people with few job opportunities are at risk. While economic neoliberalism dominates in Australia, the development of supportive rural policies and programs varies across the world. Many nations like Australia are committed to an economic rationalist orthodoxy that blinds attention to the human rights of people most affected by global challenges.

Climate variability and policies designed to address these challenges are a good example. Defining climate change as one

of the greatest moral dilemmas of our time, Molnar (2010: 12) reminds us that rural people will be most adversely affected by climate events and by policies aimed at climate change mitigation:

> the distribution of rewards from climate change mitigation is broad and diffuse; the distribution of costs, compromised livelihoods, and community disruption often is focused. The opportunity for institutional recreancy through benign neglect or active exploitation of less-populated locales in the conduct of climate change mitigation are many.

Neoliberal- and efficiency-based policy drivers are evident in many countries through processes of privatization of services, the removal of government from the front line of service delivery, regionalization and centralization of services such as health and hospitals, and the introduction of user-pays principles to basic service structures (Alston, 2009). These policy principles, reliant as they are on a critical mass of service users and a profit motive, inherently disadvantage rural people. Centralization of health services, for example, results in rural people needing to travel long distances for basic services such as birthing, cancer treatment, and other specialist attention. This has significant gendered impacts. In the case of birthing, women are left in highly vulnerable situations often needing to travel long distances in labor (Dietsch et al., 2008), and rural women also may opt for radical cancer treatments such as mastectomies rather than less invasive but more time intensive treatments (Alston and Kent, 2006). Mental health services for men that address issues associated with rural decline are also lacking (Alston, 2010c). Major cutbacks in public spending and a profit rationale for service delivery in countries such as Australia, where vast distances separate people from services, result in a significant impost on the human rights and citizenship rights of rural people. Inevitably this results in a reduced number of welfare services, declining infrastructure and transport, restricted employment opportunities, and growing poverty – problems that are largely invisible to the general population (Pugh and Cheers 2010). The loss of services also means a loss of a professional class of health and welfare workers, and those that remain, experience significantly increased workloads.

Nonetheless, rural policy across the world varies from an overreliance on the marketplace to determine outcomes and futures, to one where policy incorporates inclusive practices and equity goals. While the United States and Australia favor a market-based approach, many European Union countries and the United Kingdom have introduced a greater focus on social equity (Shortall, 2008). The distinguishing feature of this differentiation is that countries such as Australia and the United States tend to conflate rural policy with agricultural policy, limiting their vision to productivity gains and technological advances in this industry (Alston, 2010b). The inevitable consequences of these policies are that agricultural producers are on a constant treadmill, needing to work harder and faster and to continually expand their operations. This philosophy has dire consequences for rural areas. It results in larger farms and a declining number of small farms or the gentrification of rural areas, loss of rural families and thus a rise in commuting families in areas close to urban centers, lower numbers of school children and hence a loss of teachers and services such as school buses, lower numbers of people available for sports teams and social activities, a spiraling loss of social capital (Lawrence, 2005), higher levels of racism in rural areas, and a lack of human rights in countries such as China where registration under the hukou system reduces the rights of rural people to education and employment relief (Caixiong, 2013).

By contrast, EU and UK policies have given a greater focus to people and place, adopting a *territorial* rather than a *sectoral* approach, putting emphasis on the inherent complexities of rural areas and the diverse nature of the whole community rather than on the agricultural base as the leading driver of policy (Warner and Shortall, 2008; Shucksmith, 2008). Policies in these areas tend to rest on a notion of sustainable development with sustainability recognized as incorporating social, economic, and environmental aspects.

The principles underpinning sustainable development draw from the landmark Brundtland Report (1987: 9) that defined sustainability in terms of intergenerational equity – meeting the needs of the present generation without compromising the ability of future generations to meet their own needs. Subsequently there has been a significant focus on social sustainability and socially sustainable communities (Cocklin and Alston, 2003). McKenzie (2004: 18), for example, defines socially sustainable communities as *equitable, diverse, connected and democratic and providing a good quality of life*.

The introduction of sustainability as a basis for rural policy has led to what Marsden (2009) terms place-making and Shucksmith (2009) place-shaping – a recognition of the needs of the whole community rather than just the agricultural sector. Place-shaping policies acknowledge that a socially inclusive community requires services and supports that aid integration and resilience. However, while globalizing forces and environmental disasters continue to drive rural change, territorially focused policies shine a light on social complexities, the need for alternative industry jobs and service infrastructure, and recognition of the inherent rights of citizens.

Social workers should understand the principles underpinning policies that affect their rural areas and be prepared to advocate for social implications to be considered, for cost-effective services to be accessible and available, for equity, and for the rights of rural people to be addressed in policy determinations. Reflecting on the nature of social work practice, bringing the voices of the marginalized to the policy table, and reflecting on the rights-based focus of services and service structures are a critical feature of the work of rural social workers.

Rural Practice

The complexities of people and place and the impact of globalization, increasing poverty, aging, and service decline are reshaping rural social work as a unique and evolving field of practice – one that Dylan (2012: 79) terms *radical, sustainable social work practice*. Rural social work practice is distinguished by its setting and context (Pugh and Cheers, 2010; Maidment,

2012), and, as with all fields of practice, it is useful to approach rural social work with an understanding of professional aims. Payne (2000) suggests a contemplation of whether their purpose is to ensure that the continuance of the preexisting (potentially inequitable) social order, to assist and empower people to take control of their own lives (with limited resources), or to facilitate transformative social change (see for example Pelling, 2011). Picking up on this theme Dominelli (2012) notes that environmental crises may provide the space to challenge structural inequalities through a new approach to rural spaces. Further she reminds us (2012: 3):

> Contemporary social work has a vested interest in attending to environmental issues as an integral part of its daily remit if it is to retain its currency in contemporary societies [and] emphasise the social issues that people have to resolve in the twenty-first century…

These social issues are exposing the rural context to increased scrutiny from governments and NGOs. Social workers are implicated in crisis interventions and reconstruction activities following sudden onset environmental disasters and conflict situations, and in rural places subject to ongoing but slowly evolving changes and can make a significant contribution to policy and practice advice.

Australian social workers, for example, have been deployed in large numbers to areas within country affected by floods and bushfires, and have been prominent in postdisaster work in areas across the Asia Pacific region, most recently including in Christchurch and Japan following major earthquakes, in Bali following a terrorist attack, and across several Pacific regions following the 2004 Tsunami (Hargreaves, 2012). It is not unusual for large numbers of social workers to be sent by the Australian government within 24 h of a crisis situation. This work usually short term – involves long, harrowing hours of work – is focused on crisis intervention and reestablishing communities, and requires collaborative work with other professionals and groups. Social workers are also engaged in medium and long-term reestablishment phases of postdisaster situations, but it is the immediate aftermath where social work action is most visible (Hargreaves, 2012).

Rural social work is not just about postdisaster situations. Social workers in rural areas work in a number of fields including health, mental health, child protection, community development, as well as a range of other areas in government, nongovernment, and private organizations. Their work incorporates all types of practice including casework, group work, advocacy, community development, research, and social activism. Most rural social workers are generalists, working across a number of areas with a strong focus on community development (Locke and Winship, 2005) in practices that Maidment (2012) describes as generalist and generic.

Pugh and Cheers (2010) suggest the importance of informing rural practice with an understanding of the history of the community, its informal politics, local networks, and local knowledge – a position supported by Locke and Winship (2005), who also note the importance of cultural factors. This adds significant strength to rural practice regardless of the type of work being done. Nonetheless, statutory obligations in some legislative areas such as child protection and mental health blur the lines between social care and social control (Harns and Connolly, 2008) and in rural areas, the high visibility of workers makes their work in these areas very complex.

Moreover, in small communities, professionals can be stigmatized for their roles in certain social interventions particularly related to family relationships, may feel a little 'on guard' and constantly in professional mode and thus continually check their social relations. Certainly, there is a degree of difference in this respect between those who are resident in a community and those who are visiting professionals, those more able to maintain strict boundaries.

This personal boundary and privacy issue is not limited to workers. Clients wishing to visit a social worker can be challenged if a service is too open. Gossip is a powerful form of information sharing in rural communities. For this reason many social workers have offices away from central areas, in shared locations with other health professionals or at the local hospital or health service and many have off-street parking to further protect clients.

A further critical personal barrier is the high visibility and lack of anonymity workers have in rural communities (Maidment, 2012). It is not unusual to see clients in the street, the supermarket, at social events and to feel that the lines between the personal and professional are constantly blurred. Many workers report their fear of violence, particularly in situations where they are dealing with violent perpetrators or child protection issues. Pugh (2007) terms these associations 'dual relationships' – the out of work hours connections with clients that he argues challenge our ethical understandings of good practice and confidentiality and reveal the difficulty of being a professional worker in a rural community, a position also recognized by First Nation and indigenous workers in Australia and Canada.

Recruiting and retaining social workers in rural areas is one of the main constraints on this field of practice. Identified reasons for this include the perceived professional and social isolation of rural work, a lack of professional opportunities, high workloads, expectations of on-call work, the temporary nature of some workplace contracts, poor employment supports, burnout, travel demands, and high visibility (Alston, 2009; Cheers et al., 2007). Professional isolation appears to be growing as areas decline, supervision arrangements are becoming more tenuous, and workers are working longer hours (Bowles, 2012; AIHW, 2009; Cheers et al., 2007). Despite the complexity and growing social issues in this field, rural health practice has been termed a 'professional nursery' because it attracts new graduates who stay for a couple of years before moving back to the city (Healy, 2004). The effects of the temporary and highly mobile workforce are that rural practice lacks not only a solid base of research but also long-term workers in rural practice can feel professionally isolated.

Rural Practice Opportunities

Nonetheless, there continues to be a rich vein of experienced social workers who are attracted to rural communities and

stay for lengthy periods. These long-term highly experienced workers provide a major, and often unheralded, contribution to the development of innovative social work practice. They have the capacity and creativity to try new programs, extend existing knowledge, and expand beyond traditional practice boundaries. It is here that rural social work practice finds its exciting and unique edge or what Maidment (2012: 17) describes as its "rich source of professional enlivenment." Perceived barriers can be turned into outstanding opportunities by these workers who know the strengths of the community. They can also challenge evident government neglect, expand community development opportunities, and drive innovative and critical change.

Where rural workers are sole practitioners or in small agencies, it is not unusual to find they have established extensive interagency groups to work collaboratively across agencies including the local police, ambulance and other health workers, high school representatives, community service organizations, and local neighborhood groups. Beddoe and Burley (2012) note that these multidisciplinary formal or informal networks provide professional enrichment as well as a way of addressing personal safety and isolation. These rich and diverse networks allow extensive interagency cooperation, empower the community, mobilize diverse groups, and often enable significant funding to be sourced for community development and innovative developments. Workers in these communities are not afraid to try new technologies, to develop professional opportunities, to expand the horizons of their community members, to activate governments and NGOs, and to bring their community's issues to the attention of policymakers. A good rural social worker can have a significant impact on the lives of those who live in the communities they serve and the opportunities available to them. These social workers have a deep commitment to their work, a strong desire to live in a rural area, extensive friendship and professional networks, high levels of autonomy, a delight in introducing innovative practice ideas, and a creative ability to balance work and family life (Alston, 2009; Cheers et al., 2007). There is no doubt these workers are among the leaders of our profession.

Improved Model of Rural Practice

While innovative practice is evident across rural areas, there are significant challenges that require an informed practice focus and framework. In this instance, the work of Maidment and Bay (2012: 222) in developing a set of principles that can shape a new understanding of rural practice is particularly useful. These include acknowledging the significance of place, governmentality, and the consequent way 'rural' is positioned in policy; positioning rural social work practice in a social justice framework; accepting interdisciplinarity; and developing comprehensive rural practice research to inform and develop this critical area of practice.

In previous work (Alston, 2004, 2005, 2007, 2009), the writer has argued for the identification of rural social work strategies and further development of a rural social work practice model. These can be summed up as follows. Rural social workers should:

- locate their work in the space between the community and external bodies;
- use a full range of social work skills;
- be strategic in using these skills;
- be inclusive;
- work across boundaries within and beyond the community;
- use existing formal and informal networks;
- work as a community animator (Cheers et al., 2007) mobilizing community strengths and shaping goals and activities;
- work with and support other human service workers;
- advocate for rural communities and people;
- advise governments and others on rural disadvantage and policy; and
- enable a voice for rural people (Alston, 2009).

Developing a more informed model of human service delivery is essential to improving the social supports available to rural communities. The withdrawal of governments from service delivery has left a hodge-podge, of disjointed, opportunistic, and often overlapping services with significant gaps in other areas. What is required is a new model of service infrastructure for rural areas that gives greater security to communities and to the workers who serve them.

This model should include a number of strategic elements including the following:

- rural policies based on social, as well as economic and environmental sustainability;
- regional social work hubs of skilled social workers available to provide supervision, professional development, and support for local social workers;
- the development of this hub-and-spoke model to provide a regional office base for local social workers to periodically work from, in order to gain collegial support and discuss strategies and cases;
- the employment of social workers at local government level to undertake community development and to oversee service infrastructure and development;
- the establishment of Social Inclusion Committees at local government level to assist and advise social workers;
- the establishment of a competitive rural fund similar to the European LEADER program that allows local government areas to bid for funding to develop socially inclusive practices and structures;
- extension of the use of existing infrastructure such as health services, high schools, and local government offices to maximize existing buildings for extended community capacity building activities such as education, IT support, and community activities; and
- advocacy and rural policy advice.

Conclusion

Rural social work practice is emerging as a leading field of innovative practice not least because of the factors shaping twenty-first century uncertainties. Climate variability,

globalization, growing uncertainties, and increasingly vulnerable populations are shaping rural areas as sites of increasing poverty and social insecurities. This article has outlined the factors that act to reinforce these insecurities – the nature of policies and practices that ignore and discount human rights and social impacts. This is fertile ground for social workers and there is increasing evidence of innovative social work practice in sites of slow rural decline or cataclysmic climate events.

Environmental incidents, globalization, and rural restructuring increasingly impact rural areas. Neoliberal policies and a focus on economic growth tend to blind policymakers to the social injustices inherent in a market-based approach. Social workers have a critical understanding of factors that shape social sustainability and are key professionals in rural areas.

Yet, rural social work is not without its challenges. Nonetheless, the opportunities presented by rural social work to be innovative and autonomous have enabled many social workers to find their home in rural areas, investing their professional energies in developing this critical area of practice. With more attention from policymakers and the profession, rural social work can be understood as one of the most critical areas of practice and an area of major significance to the advancement of human rights.

See also: Climate Change; Ecological Social Work: Shifting Paradigms in Environmental Practice; Green Social Work; Sustainable Development and Social Work.

Bibliography

Alston, Margaret, 2004. 'You don't want to be a check-out chick all your life': The out-migration of young people from Australia's small rural towns. Australian Journal of Social Issues 39 (3), 299–315.

Alston, Margaret, 2005. Forging a new paradigm: Australian rural social work practice. Rural Society 15 (3), 276–283.

Alston, Margaret, 2007. 'It's really not easy to get help': services to drought affected families. Australian Social Work 60 (4), 421–435.

Alston, Margaret, 2009. Innovative Human Services Practice: Australia's Changing Landscapes. Palgrave macmillan, Melbourne.

Alston, Margaret, 2010a. Australia's rural welfare policy: overlooked and demoralised. In: Milbourne, P. (Ed.), Welfare Reform in Rural Places: Comparative Perspectives. Emerald Group Publishing Ltd, Bingley, pp. 199–217.

Alston, Margaret, 2010b. Gender and climate change in Australia. Journal of Sociology 47 (1), 53–70.

Alston, Margaret, 2010c. Rural male suicide in Australia. Special Edition of Social Science and Medicine 70.

Alston, Margaret, 2012. Introduction: gender and climate change. In: Alston, Margaret, Whittenbury, Kerri (Eds.), Gender and Climate Change: Research, Policy and Action. Springer, Netherlands.

Alston, Margaret, 2013. Environmental social work: accounting for gender in climate disasters. Special Edition of Australian Social Work 66 (2), 218–233.

Alston, Margaret, Kent, Jenny, 2004. Social Impacts of Drought: Report to NSW Agriculture. Centre for Rural Social Research, Charles Sturt University, Wagga Wagga.

Alston, Margaret, Kent, Jenny, 2006. Impact of Drought on Rural and Remote Education Access: A Report to DEST and Rural Education Fund of FRRR. Centre for Rural Social Research, Charles Sturt University, Wagga Wagga.

Alston, Margaret, Kent, Jenny, 2008. The big dry: exacerbating the link between rural masculinities and poor health outcomes for rural men. Australian Journal of Sociology 44 (2), 133–147.

Alston, Margaret, Kent, Jenny, 2009. Generation X-pendable: declining human capital investment in rural and remote young people. Australian Journal of Sociology 45 (1), 89–107.

Alston, Margaret, Vize, S., 2010. Gender and Climate Change in the Pacific. Gender, Leadership and Social Sustainability (GLASS) Research Unit, Monash University, Melbourne.

Australian Institute of Health and Welfare, 2008. Report Profile: Rural, Regional and Remote Health: Indicators of Health System Performance. http://www.aihw.gov.au/publication-detail/?id=6442468150 (date accessed 28.10.12.).

Barusch, Amanda S., 2011. Disaster, vulnerability, and older adults: toward a social work response. Journal of Gerontological Social Work 54 (4), 347–350.

Beddoe, Liz, Burley, Molly, 2012. Collaborating within and across interprofessional teams. In: Maidment, Jane, Bay, Uschi (Eds.), Social Work in Rural Australia: Enabling Practice. Allen & Unwin, Melbourne, pp. 40–60.

Bowles, Wendy, 2012. Caregiving in small rural and regional towns. In: Maidment, Jane, Bay, Uschi (Eds.), Social Work in Rural Australia: Enabling Practice. Allen & Unwin, Melbourne, pp. 119–132.

Caixiong, Zheng, 10 May 2013. Guangzhou moves to abolish rural hukou. China Daily. http://www.chinadaily.com.cn/china/2013-05/10/content_16488955.htm (date accessed 09.08.13.).

Campbell, A., 2008. Managing Australian Landscapes in a Changing Climate. Department of Climate Change and Energy Efficiency, Commonwealth of Australia, Canberra.

Cheers, Brian, 1998. Welfare Bushed: Social Care in Rural Australia. Ashgate Publishing Company, Brookfield, Vic.

Cheers, B., 1999. Community-embedded social care practice. In: Briskman, L., La Nauze, H., Lynn, M. (Eds.), Challenging Rural Practice: Human Services in Australia. Deakin University Press, Melbourne, VIC.

Cheers, Brian, Taylor, Judith, 2005. Social work in rural and remote Australia. In: Alston, M., McKinnon, J. (Eds.), Social Work Fields of Practice, second ed. Oxford University Press, Melbourne.

Cheers, B., Darracott, R., Lonne, B., 2007. Social Care Practice in Rural Communities. Federation Press, Leichardt, NSW.

Cocklin, Chris, Alston, Margaret, 2003. Introduction. In: Cocklin, Chris, Alston, Margaret (Eds.), Community Sustainability in Rural Australia: A Question of Capital. Centre for Rural Social Research, Wagga Wagga, NSW.

Dietsch, E., Davies, C., Shackleton, P., Alston, M., McLeod, M., 2008. Luckily We Had a Torch: Birthing in Rural and Remote NSW. A Report to the Nurses and Midwives Board of NSW. Institute for Land, Water and Society, Charles Sturt University, Wagga Wagga.

Dominelli, Lena, 2012. Green social Work: From Environmental Crisis to Environmental Justice. Polity Press, Cambridge.

Dylan, Arielle, 2012. Environmental sustainability, sustainable development and social work. In: Gray, Mel, Coates, John, Hetherington, Tiani (Eds.), Environmental Social Work. Routledge, Abingdon, pp. 62–87.

Enarson, Elaine, 2009. Women, gender and Disaster. Sage Publications, US.

Foster, M., 2007. Fields of health service provision. In: Taylor, S., Foster, M., Fleming, J. (Eds.), Health Care Practice in Australia. Oxford University Press, South Melbourne.

Garnaut, Ross, 2008. The Garnaut Climate Change Review Final Report. Cambridge University Press, Melbourne.

Garnaut, Ross, 2011. Garnaut Climate Change Review – Update: Australia in the Global Response to Climate Change. http://www.garnautreview.org.au/update-2011/about-review.html (date accessed 05.07.11.).

Gillard, Julia, 2012. Gillard 2012 Speech to the National Farmers' Federation National Congress, Tuesday 23 October. Canberra. http://www.pm.gov.au/press-office/speech-national-farmers-federation-national-congress (date accessed 27.10.12.).

Gray, Ian, Lawrence, Geoffrey, 2001. A Future for Regional Australia: Escaping Global Misfortune. Cambridge University Press, Cambridge.

Hargreaves, Desley, 2012. Gender and climate change: implications for responding to the needs of those affected by natural disasters and other severe weather events. Change. In: Alston, Margaret, Whittenbury, Kerri (Eds.), Gender and Climate Change: Research, Policy and Action. Springer, Netherlands, pp. 277–281.

Harns, Louise, Connolly, Marie, 2008. The art and science of social work. In: Connolly, Marie, Harns, Louise (Eds.), Social Work Context and Practice, second ed. Oxford University Press, South Melbourne, VIC.

Healy, Guy, 14–20 July 2004. Regional Unis Play the role of Nursery. Campus Review, 5.

Hugo, Graeme, 2005. The state of rural populations. In: Cocklin, Chris, Dibden, Jacqui (Eds.), Sustainability and Change in Rural Australia. UNSW Press, Sydney, pp. 56–79.

Lawrence, Geoffrey, 1987. Capitalism in the Countryside. Pluto Press, Sydney.

Lawrence, Geoffrey, 2005. Globalisation, agricultural production systems and rural restructuring. In: Cocklin, Chris, Dibden, Jacquie (Eds.), Sustainability and Change in Rural Australia. UNSW Press, Sydney, pp. 104–120.

Leipins, Ruth, 2000. Making Men: the construction and representation of agriculture based masculinities in Aust and NZ. Rural Sociology 65 (4), 605–620.

Locke, Barry L., Winship, Jim, 2005. Social work in rural America: lessons from the past and trends for the future. In: Lohmann, N., Lohmann, R.A. (Eds.), Rural Social Work Practice. Columbia University, New York, pp. 3–24.

Lynch, K., 2011. Spike in Domestic Violence after Christchurch Earthquake. National. http://www.stuff.co.nz/national/christchurch-earthquake/4745720/Spike-in-domestic-violence-after-Christchurch-earthquake (date accessed 14.03.11.).

Maidment, Jane, 2012. Understanding rurality: a conceptual framework. In: Maidment, Jane, Bay, Uschi (Eds.), Social Work in Rural Australia: Enabling Practice. Allen & Unwin, Melbourne, pp. 3–20.

Maidment, Jane, Bay, Uschi, 2012. Rural practice: an agenda for the future. In: Maidment, Jane, Bay, Uschi (Eds.), Social Work in Rural Australia: Enabling Practice. Allen & Unwin, Melbourne, pp. 221–234.

Marsden, Terry, 2009. Mobilities, vulnerabilities and sustainabilities: exploring pathways from denial to sustainable rural development. Sociologia Ruralis 49 (2), 113–131.

McAdam, Jane, 2012. Climate Change, Forced Migration, and International Law. Oxford University Press, New York.

McKenzie, Stephen, 2004. Social Sustainability: Towards Some Definitions: Hawke Research Institute Working Paper Series No 27. Hawke Research Institute University of South Australia Magill, South Australia. http://w3.unisa.edu.au/hawkeinstitute/publications/downloads/wp27.pdf (date accessed 27.10.12.).

Molnar, Joseph J., 2010. Climate change and social response: livelihoods, communities and the environment. Rural Sociology 75 (1), 1–16.

Newman, Peter, Beatley, Timothy, Boyer, Heather, 2009. Resilient Cities: Responding to Peak Oil and Climate Change. Island Press, Washington.

Opdam, Paul, Luque, Sandra, Bruce Jones, K., 2009. Changing landscapes to accommodate for climate change impacts: a call for landscape ecology. Landscape Ecol 24, 715–721.

Payne, Malcolm, 2000. Teamwork in Multiprofesional Practice. Macmillan, London.

Pelling, Mark, 2011. Adaptation to Climate Change: From Resilience to Transformation. Routledge, London.

Petkova, Vanessa, Lockie, Stewart, Rolfe, John, Ivanova, Galina, 2009. Mining developments and social impacts on communities: bowen basin case studies. Rural Society 19 (3), 211–228.

Proudley, Mae A., 2008. Fire, Families and Decisions (unpublished master of applied science thesis). RMIT, Melbourne.

Pugh, Richard, 2007. Dual relationships: personal and professional boundaries in rural social work. British Journal of Social Work 37 (8), 1405–1423.

Pugh, Richard, Cheers, Brian, 2010. Rural Social Work: An International Perspective. Policy Press, Bristol.

Raitano, Florine, 2010. Rural Policy Matters: Rural Issues, Concerns, Challenges and Recommendations, Opinion Piece for the 2010 Democratic Rural Policy Forum. Washington. http://www.rcac.org/assets/white-papers/RuralPolicyMatters.pdf (date accessed 28.10.12).

Rios, Betty Rose D., 1988. "Rural"-A Concept beyond Definition? ERIC Clearinghouse on Rural Education and Small Schools Las Cruces NM. http://www.ericdigests.org/pre-929/concept.htm.

Shortall, Sally, 2008. Are rural development programmes socially inclusive? Social inclusion, civic engagement, participation and social capital: exploring the differences. Journal of Rural Studies 24, 450–457.

Shucksmith, Mark, 2008. New labour's countryside in international perspective. In: Woods, Michael (Ed.), New Labour's Countryside: Rural Policy in Britain since 1997. Policy Press, Bristol, pp. 59–77.

Shucksmith, Mark, 2009. Disintegrated rural development? Neo-endogenous rural development, planning and place-shaping in diffused power contexts. Sociologia Ruralis 50 (1), 1–14.

Tatlonghari, Gerlie, Paris, Thelma R., 2012. Gendered adaptations to climate change: a case study from the Philippines. In change. In: Alston, Margaret, Whittenbury, Kerri (Eds.), Gender and Climate Change: Research, Policy and Action. Springer, Netherlands, pp. 237–250.

Tester, Frank, 2012. Climate change as a human rights issue. In: Gray, Mel, Coates, John, Hetherington, Tiani (Eds.), Environmental Social Work. Routledge, Abingdon, pp. 102–118.

United Nations Population Fund (UNFPA), 2007. State of the World Population 2007: Unleashing the Potential of Urban Growth. UNFPA, United Nations.

United Nations Secretary-General's High-level Panel on Global Sustainability, 2012. Resilient People, Resilient Planet: A Future Worth Choosing. United Nations, New York. http://www.un.org/gsp/sites/default/files/attachments/GSPReport_unformatted_30Jan.pd (date accessed 05.02.12.).

Warner, Mildred, Shortall, Sally, 2008. Growth coalitions and rural development policy in the EU and the us *EuroChoices* special issue: focus on policy reform: the CAP health check and the US. Farm Bill 7 (3), 34–37.

World Commission on Environment and Development (WCED), 1987. Our Common Future. Oxford University Press, Oxford.

Zapf, M.K., 2009. Social Work and the Environment: Understanding People and Place. Canadian Scholars Press, Toronto, ON.

Rural Sociology

Daniel T Lichter, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by F.H. Buttel, volume 20, pp. 13429–13433, © 2001, Elsevier Ltd.

Abstract

Rural sociology is a field of applied sociological research and training that has historically focused on rural people and places. It has its intellectual origins and administrative home in the US Land-Grant System of colleges. Theoretical concerns historically centered on the implications of urbanization, modernization, and capitalist economic development. While the sociology of agriculture and environmental society represent major foci of investigation, rural sociology is much less parochial today than in the past, embracing substantive research themes, alternative methodologies, and policy issues that blur the boundaries with sociology. A new rural sociology emphasizes an emerging synthesis of research at the rural–urban interface.

Rural sociology is alternatively viewed as a subfield or specialty area in the discipline of sociology or as a discipline in its own right. This dualism reflects rural sociology's intellectual origins and history, its administratively ambiguous but changing place within the academy, and the recurring antipathy between rural sociologists and mainstream sociology that sometimes pivots on questions of the continuing relevance of rural issues, especially as urbanization continues apace. This intellectual and disciplinary separation also is rooted in different sources of funding (under the US Land-Grant System – or not), different emphases on theory and applied research and public engagement, and the physical separation and isolation of rural sociologists within America's public universities. But these historically distinguishing features of rural sociology are now rapidly breaking down as traditional sources of university funding (especially state and federal sources that in the past supported separate rural sociology departments, faculty appointments, and graduate student training) have dried up under the Land-Grant System. Today, rural and agricultural-related research (in the Agricultural Experiment Station) and public engagement and outreach (through the US Department of Agriculture's (USDA) Cooperative Extension Service) have focused on technical rather than on behavioral aspects of rural development and community change.

Rural sociology has traditionally been defined as the study of social organization and change that is characteristic of regions and places where populations are small, spatially dispersed, and economically underdeveloped (i.e., in rural areas, however defined). It has its origins in nineteenth-century intellectual thought, which often centered on debates over the implications for traditional rural social organization and farm life (e.g., cohesiveness or moral superiority) in the face of capitalist industrialization and modernization (Buttel, 2001). Rural sociology is rooted in the early work of Tonnies (*Gemeinschaft* and *Gesellschaft*) and the so-called rural–urban continuum (and the replacement of mechanical solidarity with organic solidarity). And, even today, research in rural sociology typically focuses on rural–urban differences, especially economic and social disparities in well-being.

The intellectual home of most self-identified rural sociologists is the Rural Sociological Society (RSS). Its flagship journal, *Rural Sociology*, was first published in 1937. It provides a forum for research on 'rural people, places, and problems.' The journal includes mostly empirical-based research studies related to rural economic development, environmental and demographic impacts, the structure of food and agricultural production, and rural–urban linkages, among other topics. But, increasingly, rural areas or communities simply serve as spatial containers for the comparative study of institutional change (e.g., families, health care, education, religion, and public policy). Some observers, in fact, claim that there is no 'rural' left in rural sociology (Friedland, 2002), that is, rural and urban people – at least in the United States and some Western industrial societies – are largely indistinguishable.

Indeed, the intellectual boundaries that sometimes separate rural sociology from sociology are rapidly changing. This reflects shifts in university priorities and funding. It also reflects demographic processes – aging and cohort replacement of its practitioners – within rural sociology itself. The heyday of rural sociology, if measured in resources and hiring, occurred in the 1970s. Over the past decade or two, rural sociology departments have shrunk in size (as has membership in the RSS) with the retirement of faculty and programmatic staff. In several cases (e.g., Washington State University or Texas A&M University), departments have been eliminated altogether and faculty have been dispersed or consolidated within other university programs (e.g., rural education and leadership). Other departments have redefined their research and teaching niche within the university and rural sociology. Some leading departments of rural sociology, for example, have renamed themselves, becoming Departments of Development Sociology (e.g., Cornell University), Community and Environmental Sociology (University of Wisconsin–Madison), or Community and Leadership Development (University of Kentucky) to better reflect their changing missions. In some Land-Grant Universities, sociology departments have regularly hired and trained rural sociologists, providing separate faculty lines both in colleges of agricultural science and other colleges, such as arts and sciences (e.g., North Carolina State University, Michigan State University, and Iowa State University). But the marriage of sociology and rural sociology has sometimes been an unhappy one, reflecting differences in theoretical perspectives, resources (e.g., 9- versus 12-month appointments), teaching loads, and missions (e.g., basic versus applied), which

inevitably led to differences in departmental expectations and rewards (e.g., tenure and promotion, and salary disparities). Other university sociology departments (such as Brigham Young University or the University of New Hampshire) have melded faculty with strong interests in rural sociology with those of other faculty who do not share their interests, but who nevertheless appreciate the value of rural-oriented applied work.

These changing institutional arrangements have fundamentally altered the practice of rural sociology over the past decade or two, blurring the boundaries that have historically separated its research, training, and outreach missions from mainstream sociology departments. For example, newly mined PhDs trained in rural sociology departments have increasingly found faculty appointments or research positions outside the Land-Grant System, in sociology departments and other programs in major research and teaching universities, in smaller state universities, and in private liberal arts colleges. The traditional topics of rural research and teaching – sociology of agriculture, food systems, environmental sociology, population studies, and the sociology of development – have been as a result not only dispersed more widely across higher education but also diluted by the research and training missions of their new homes outside the Land-Grant System.

Rural sociology, whether viewed as a sociological specialty or as a separate discipline (because of its institutional location within the University), has its origins within the US Land-Grant agricultural college system but owes its historical development (and the waxing and waning of scholarly interest and resources) to trends in agrarian politics and changing demography (e.g., urbanization and the decline in the farm population). To illustrate these points, this article is divided into three major sections. First, the author provides a short history of the origins of rural sociology, while highlighting its past and currents connections to mainstream sociological theory and research. Second, the author highlights recent changes in the major substantive research interests and empirical approaches in rural sociology since the boom period of the 1970s. And, third, the author describes a 'new' rural sociology that acknowledges the increasing interconnections between rural and urban scholarship at the rural–urban interface in a rapidly globalizing and interconnected world (Lichter and Brown, 2011). Rural sociology arguably is being mainstreamed within the broader discipline of Sociology and expanding outside the United States, especially in European countries. Rural sociology's institutional and intellectual home is currently in flux, but reemerging as a significant subfield or specialty areas within sociology itself, a place where it started nearly 100 years ago.

Institutional and Intellectual Origins of Rural Sociology

Rural sociology is a product of the US Land-Grant System, as established by the Morrill Land-Grant College Act of 1862. In 1890, this legislation was extended to include traditionally black institutions in southern states (e.g., Alcorn State University, Florida A&M University, Tuskegee University, and Prairie View University), which are still sometimes called '1890s schools.' Each state established agricultural experiment stations, usually in a designated Land-Grant University, that worked collaboratively with the USDA in Washington, DC, to provide empirically based, scientific, and practical agricultural information (e.g., hybrid seeding, plant rotation, irrigation systems, and social erosion) for rural farmers and ranchers – to make their farm operations more cost-effective and productive. In 1887, the Hatch Act forged a new federal–state partnership for agricultural research. This was followed in 1914 by the Smith–Lever Act, which established the cooperative extension service. The idea, which started at the University of Wisconsin–Madison (the 'Wisconsin idea'), was to extend scientific knowledge (and in some cases, 'culture') to the countryside, and rural sociologists participated through research supported by agricultural experiment stations. To promote economic development, extension rural sociologists cooperated in outreach activities with state and local organizations, public officials, and policy makers, and with their counterparts in the USDA.

In 1908, President Theodore Roosevelt appointed a Presidential Commission on Country Life. The County Life Commission Report, which was published in 1909, made the case that economic development and modernization were required to improve the conditions of rural people and places. According to Buttel (2001: p. 13430), the Commission recommended "the harnessing of the social sciences, particularly agricultural economics and rural sociology, to support the technological modernizationist vision for rural America." The Commission's report was a major boon to rural-related social sciences in the Land-Grant system.

Indeed, some of the early and most influential work by rural sociologists focused on the adoption-diffusion process of innovative farming practices, as pioneered by Everett Rodgers and colleagues and students at Iowa State University (Fliegel, 1993). And, even today, many rural sociologists in the Land-Grant System hold part-time experiment station and/or extension appointments rather than teaching appointments. Over the last decade or two, however, appointments of this type have declined significantly. Faculty members in rural sociology departments today nevertheless continue to be heavily engaged in applied sociological research – in public sociology – that serves state and community stakeholders. But unlike the past, their work has become less parochial, focusing less directly on state or community issues and more on national and international concerns (e.g., economic development, environmental degradation, or rural depopulation). The days of writing descriptive experiment station bulletins that mostly focused on state or local issues is over, replaced by a new emphasis on basic and applied research, teaching, and service that meets the usual standards of scholarly productivity used by their academic counterparts in sociology departments on the other side of campus or at other institutions.

This represents a return to the past, at least in some ways. In the beginning, the field's early practitioners typically were identified as rural sociology section members under the organizational umbrella of the American Sociological Society (now the American Sociological Association (ASA)). But the rural sociology section broke off from ASA in 1937 to form the RSS. Since then, membership in the RSS reached its peak of 1101 in 1997 (Krannich, 2008). Today, RSS membership is down to

about 600 members, but holding steady over the past several years. Membership declines have understandably been cause for alarm and some self-deprecation (Beaulieu, 2005; Krannich, 2008; Singelmann, 1996), but such pessimism is warranted only if recent institutional shifts are considered in isolation of other measures of vitality and impact. For example, declines in RSS membership partly reflect healthy organizational shifts in mainstream sociology, such as the establishment of highly specialized ASA sections in development sociology or environmental sociology, as well as new professional associations that have attracted rural sociologists but sometimes siphoned membership from RSS. Many rural sociologists, including RSS members, are heavily involved today with the Community Development Society (which publishes *Journal of the Community Development Society*), International Association for Society and Natural Resources (*Society & Natural Resources: An International Journal*), and the Agriculture, Food, and Human Values Society (*Agriculture and Human Values*).

Equally important to rural sociology, if viewed as a subfield of sociology, has been the growth of interest in rural sociology and rural studies in Europe (Cloke et al., 2006) and throughout the international scholarly community of sociologists. The European Society for Rural Sociology was formed in 1957, and began publishing *Sociologia Ruralis* in 1960. The International Rural Sociological Association was established in 1976, which merged the activities of three regional rural sociology societies. A 'World Congress of Rural Sociology' has been held every 4 years since 1964 (when it was first held in Dijon, France). The XIV World Congress of Rural Sociology is scheduled for Toronto, Canada, in 2016. Buttressing international interest of rural sociology and rural social science, the multidisciplinary *Journal of Rural Studies* was first published in 1985. The paradox today is that growing interest in the subject matter, theoretical perspectives, and empirical approaches of rural sociologists has perhaps never been greater, yet the institutional support for rural sociology in the academy has undoubtedly waned, especially in the United States.

The point here is straightforward: The intellectual health and vitality of rural sociology cannot be measured simply by the number and size of rural sociology departments, by membership in the RSS, or by changes in the number of North American practitioners. Today, the growing scholarly integration between rural and (urban) sociology harkens back to an earlier period of history of sociological thought when some of sociology's most distinguished practitioners had their intellectual roots in rural sociology. Just a few examples suffice to make the point. Rural sociological research was spearheaded in the 1910s by Charles J. Galpin through his rural life studies at the University of Wisconsin. The program was housed in the Department of the Agricultural Economics, and a separate Rural Sociology department was first established in 1930 by J.H. Kolb. Franklin Giddings helped establish a major course of study in rural sociology at Columbia University in the 1920s. Pitirim Sorokin and Charles Zimmerman published *Principles of Rural-Urban Sociology* in 1929. More recently, Otis Dudley Duncan started his career in rural sociology in the 1950s at Pennsylvania State University. Another iconic figure in social stratification and mobility, William Sewell, began his career in the Department of Rural Sociology at the University of Wisconsin–Madison. Even more recently, a prominent Princeton sociologist, Robert Wuthnow, has published *Small-Town America: Finding Community, Shaping the Future* (2013). The point here is that rural sociology has not only a proud history, but also an optimistic future if the spotlight is placed less on organizational change and more on the growing interest in the subject matter itself and on newly formed connections to mainstream sociology and to the broader social scientific community.

Research Themes in Rural Sociology

To understand rural sociology today requires separating its historical origins and current institutional arrangements from its scholarly content (i.e., what rural sociologists actually do). A recent review (Buttel, 2001) claims that rural sociology is divided into six major branches: rural population, rural community, natural resources and environment, agriculture, sociology of international rural development, and sociology of agricultural science and technology. But this assessment arguably gives too much internal coherence to a field that is in flux. It may also narrow the scope of the field, while creating artificial and sometimes unnecessary boundaries between substantive areas in rural sociology. Moreover, most of the usual fields of endeavor that typically define mainstream sociology are also represented in rural sociology. And some newly emerging research themes in rural sociology, such as immigration, health, and agrifood systems, reflect ongoing social, economic, and political changes that now characterize the rural landscape, both in the United States and in other developed and developing societies.

One way to define rural sociology is by the way rural sociologists define themselves – by what they study and write about. Every 10 years or so, the RSS takes stock of the field, providing its "readers a forward-looking, policy-relevant state of the art understanding of rural people and places and the challenges they face" (Bailey et al., 2014). The latest volume, published in 2014, is the fourth in the series, beginning with Donald Dillman and Darryl Hobbs' (1982) *Rural Society in the U.S.: Issues for the 1980s*. This inaugural volume was followed by state-of-field volumes in 1991 by Cornelia Flora and James Christenson (*Rural Policies for the 1990s*) and again in 2003 by David Brown and Louis Swanson (*Challenges for Rural America in the Twenty-First Century*). The most recent volume, coedited by Connor Bailey, Leif Jensen, and Elizabeth Ransom, is entitled *Rural America in a Globalizing World: Problems and Prospects for the 2010s*. The title itself is clearly indicative of rural sociology's expanding research concerns about the impact of globalization and rural America's growing interconnections with urban America and the rest of the world. But it is also distinctive in other ways that reflect changes in rural sociology, including for the first time a coeditor located outside the Land-Grant System and a woman (reflecting the gender transformation of the field over the past decade or two).

As a document of record, *Rural America in a Globalizing World* is divided into five major sections that fairly define the rural sociology: (1) changing structure of agriculture, (2) natural resources and environment, (3) population change in rural North America, (4) diversity in rural America, and (5) rural economies, community, and quality of life. The chapters on agriculture focus primarily on the agrifood system, on the bifurcation of the America (i.e., corporate and hobby/organic/specialty farming), land ownership and consolidation,

agricultural workers and changing technology, food safety, animal welfare (e.g., humane slaughtering practices), and agrifood social movements (e.g., organic farming, farmer's markets, and hormone-free meat). The guiding perspective today focuses on the changing structure of agriculture and food systems, the movement from a system based on family farms to one that is highly industrialized and highly interconnected to many aspects of the global economy (Lobao and Meyer, 2001). North American agriculture has moved from a system based on family farms to global commodity systems dominated by multinational corporations. Neoliberalism, the deregulation of financial markets, free trade and global markets, and the consolidation of corporate power in the food industry have had profound impacts on rural workers and communities (Bonanno, 2014). Today, the sociology of agriculture may be the largest research area in rural sociology, and it has expanded in scope and influence over the past two decades.

Agriculture arguably also is an area unique to the field; there are few sociologists in mainstream sociology departments who study the changing structure of agriculture. But this is likely to change as the implications of food systems for traditional areas of sociological and economic study are made clear in the future. At the same time, rural agricultural issues – farming and food systems – is related to many significant dimensions of the United States and global economy that have attracted broad sociological and demographic interest. For example, changes in meat production, processing, and distribution have impacted America's small towns, especially in the agricultural Midwest and Southeast. The demand for cheap labor has attracted new immigrant workers and their families; Hispanics accounted for more than half of all growth in nonmetropolitan counties over the period 2000–10. New immigration has dramatically changed the ethnic mix of rural communities, patterns of economic and political incorporation, and many rural institutions, including education, religion, and family (Lichter, 2012). In some cases, the in-migration of Hispanic immigrants has been a lifeline to dying small towns. Figuratively, many mainstream sociologists have followed new immigrants to the rural countryside (Massey, 2008) and blurred the disciplinary and intellectual boundaries that have sometimes separated sociology and rural sociology.

With its historical emphasis on natural resources (e.g., wildlife, forests, minerals, water, and soil), environmental sociology has been a fundamental topic of intellectual pursuit among rural sociologists (Fortmann et al., 2014). Interest has now expanded with growing concerns about the societal implications of climate change, environmental degradation, and sustainability, both in the United States and elsewhere (Dunlap, 2010; Freudenburg, 2006). This is a big topic today of immense importance for sociological theory and research (e.g., social movements, inequality, and social disorganization) and public policy. It also merges interests with the natural and biological sciences. Rural America, for example, comprises over 70% of America's land area, a fact that requires good stewardship. New demands for cheap supplies of renewal energy also have spurred rural-oriented sociologists to become increasingly interested in 'fracking' for natural gas – its impacts on water quality, community disorganization and crime, traffic congestion and housing, and population and employment growth. Innovations in biofuels (e.g., ethanol from corn or other plants) and electricity from wind turbines also have potentially large economic and social implications for sparsely settled rural areas. Financialization, especially in the form of commodification (e.g., of nature), marketization (e.g., international 'land grabs'), and privatization (e.g., of government land) is often expressed in conservation policy (e.g., cap and trade) (Fortmann et al., 2014). In some cases, rural areas also have become 'dumping grounds' for urban refuse (e.g., garbage) and hazardous wastes that raise concerns about the health and well-being of rural people and communities.

This work often follows in the tradition of sociological work in resource-dependent communities and the economic and environmental exploitation of outside corporate interests, which is sometimes typified in poor coal-producing areas in central Appalachia and timber-dependent communities in the Pacific Northwest and elsewhere. The work also is rooted theoretically in The Chicago School of Sociology, which used Chicago neighborhoods in the early 1900s as naturally occurring laboratories for studying the consequences of urbanization and growing ethnic diversity for community life (e.g., social disorganization and anomie). Rapidly changing rural communities arguably now occupy this place in America's social history.

The continuing consolidation of agricultural production, rural industrial restructuring, and economic competition from overseas markets have contributed to slow growth and even population decline in many parts of rural America. Indeed, for the first time ever, nonmetropolitan counties – in the aggregate – experienced depopulation (over 2010–12). The long-term declines in rural population have provided a continuing source of urban growth, usually in the form of new in-migrants of reproductive ages. Rural out-migration and population decline has always been a major topic of research interest among rural sociologists, a topic usually framed and informed by Amos Hawley's Human Ecology perspective (Fuguitt et al., 1989). More recent research has given greater agency to rural communities (by embracing a 'political economy of places'). Population trends are linked to economic development, community fiscal viability, and health and economic well-being. In many cases, chronic out-migration – the loss of rural America's 'best and brightest' – has sapped the lifeblood from rural agricultural and resource-dependent areas. Moreover, a large second-order effect of chronic rural out-migration has been natural decrease – the excess of deaths over births. Rural populations have aged rapidly in combination with below-replacement fertility, both in the United States and in Europe. Most nonmetropolitan counties today are experiencing natural decrease over the 2000–10 period. Many parts of rural Europe (e.g., southwest Germany and the northern parts of The Netherlands and Northern Ireland) face similar problems of population aging, out-migration, and depopulation.

These population shifts have occurred unevenly across regions of the United States and Europe, and are experienced differently among different demographic segments (e.g., children, the elderly, and minorities) of the population (Kulcsar and Curtis, 2012). Uneven demographic change has raised the specter of increasing spatial inequality (Lobao et al., 2007), as some rural areas have emerged as economic winners (e.g., rural and recreational amenity areas), while others have languished (e.g., rural parts of Central Appalachia). Growing

income and wealth inequality – from one place or region to another – reflects, at least in part, growing economic inequality more generally in America. A new kind of geographic economic balkanization has attracted the scholarly interest of sociologists (who tend to focus on concentrated neighborhood poverty and economic segregation) and rural-oriented sociologists (whose work typically has centered on 'places left behind' by economic restructuring and globalism, such as Appalachia, the Delta, and isolated areas in the rural South).

Rural communities are not homogenous – economically, politically, or culturally. Increasingly, rural communities provide convenient, accessible, and relatively inexpensive venues for cutting-edge comparative research, using quantitative (based on secondary or primary survey data), qualitative, or mixed-methods approaches. What is sometimes sacrificed in representativeness is more than made up with depth and new sociological insights. A few examples suffice to illustrate the point.

For example, in *Hollowing Out the Middle*, Patrick Carr and Maria Kefalas (2009) spent several months living in a small town in Iowa, where they sought to better understand why some high school graduates left the community (labeled 'achievers') and others stayed, and the role that community and school investments played in reinforcing out-migration and contributing to their own demise. This community-based ethnography primarily focused on the rural 'brain drain' but provided more general sociological and policy lessons for all communities, not just small towns. A central point was that class divisions cause some communities to invest most heavily in young people who they prepare to leave, rather than in those students who are most likely to stay and contribute positively to civic life. Another recent book, by David Brown and Nina Glasgow (2008), used a mixed-method approach to examine retirement migration. *Rural Retirement Migration* is based on national level secondary data analysis, survey data of older persons (in-migrants and longer term residents) in 14 counties with high rates of elderly in-migration, and in-depth studies of four retirement destinations. The authors demonstrated the large and diverse economic and civic contributions of older in-migrants – who often come from big cities – and how they lead productive lives in older age, but also represent 'gray gold' in some communities. Again, this is a volume of broad sociological interest; rural communities provide a useful geographic container for studying successful aging in America. Finally, Jennifer Sherman's *Those Who Work, Those Who Don't: Poverty, Morality, and Family in Rural America* (2009) provided a cultural sociology of small town life (in the America Northwest). She focused not only on families experiencing economic hardship, but also on how choices were affected by economic constraints and shaped by cultural and moral understandings that organize social life in cohesive communities. Here, rural communities illustrate fundamental sociological lessons about how low-income mothers adapt to economic exigencies and make ends meet.

Rural America today arguably is on the frontline of social changes that have broad sociological and demographic significance. Its diversity also has advantages for comparative studies of social organization and processes – race relations, politics, social stratification, institutional change (e.g., schools), cultural clashes, and development. For example, rural areas are sometimes viewed as safe havens for traditional families and social ties. But the evidence suggests otherwise (Smith and Tickamyer, 2011). Rural families today look remarkably similar to urban families – in patterns of nonmarital fertility, cohabitation, divorce, single-parent families, and children exposed to diverse living arrangements. Family changes have been unusually rapid in rural communities. Population aging has been rapid in many rural areas, owing in large part to chronic out-migration, low fertility, and 'aging in place.' Rural America has become a new destination for Hispanics – both natives and foreign born. Issues of diversity and racial relations are rapidly being transformed in many new 'majority–minority' rural communities (Lichter, 2012). At the same time, black poverty and racial inequality remain extreme in the rural South, a legacy of slavery, the plantation economy, and its racially oppressive aftermath (Green, 2014). America's Indian reservations also highlight issues of spatial and racial inequality, as well as illustrate America's incorrigible geography of exclusion. Diversity is also illustrated by a new geography of sexuality, where some small towns have become home for sexual minorities (especially lesbians) or have emerged as places of recreation (e.g., Wellfleet, MA, or Key West, FL).

A New Rural Sociology?

In his 1983 paper in the *Annual Review of Sociology*, Howard Newby (1983) called for a 'new' rural sociology that emphasized agriculture as the intellectual lynchpin, and also provided a critique of rural sociology and, by extension, its practitioners in the Land-Grant System, who are sometimes painted as apologists for 'big agriculture.' Indeed, the aforementioned rural realities of the last half century (and more) ushered in various critical theories that sometimes framed the new rural sociology. They drew heavily from neo-Marxism and also from neo-Weberianism perspectives, the peasant studies literature, and critical theories drawn from hermeneutics and related traditions (Buttel, 2001). They represented a critique of modernization (as exemplified in diffusion studies and rural demography) within rural sociology, of quantitative research and training, and of the Land-Grant missions as defined by experiment station research and cooperative extension.

This 'new' rural sociology – its critical perspective on corporate agriculture and its ties to the Land-Grant System – has been highly influential in the past, and it arguably also has contributed to rural sociology's marginalization within the university (as measured by the withdrawal of faculty lines and resources). The work of rural sociologists has sometimes been viewed as critical rather than constructive, as ideological rather than empirical. Of course, any 'new sociology' inevitability becomes 'old sociology,' and that is arguably the case with the new rural sociology. Recent changes in the theoretical, substantive, and organizational content of rural sociology arguably reflect the march of neoliberalism, the waning of Marxist and other critical perspectives in sociology, and the usual demographic turnover of university faculty.

Rural sociology today cannot be reduced to sociology of agriculture (located within the RSS) or even be viewed as a "heuristic device with which to begin reconstructing rural sociology" (Newby, 1983: p. 78). A new rural sociology for the

twenty-first century is perhaps best represented by a new synthesis of research and training between rural sociology and mainstream sociology. This should be viewed constructively as symmetrical rather than asymmetrical process of disciplinary, substantive, and intellectual influence, where rural sociology is not lost to sociology or otherwise extinguished intellectually. It is represented by not only a new organizational or administrative synthesis (described previously), but also the new realities of societal and global change and spatial interconnections. This can be observed by innovative new research – both by urban-oriented and rural-oriented sociologists – at the rural–urban interface. Rural and urban issues are increasingly viewed as flipsides of the same coin. The 'new' rural America is marked by growing spatial interdependence – a blurring of traditional rural–urban spatial and symbolic boundaries (Lichter and Brown, 2011).

Rural-related theory and research arguably has become increasingly integrated with urban and global perspectives, and vice versa. Indeed, a new rural sociology is also evident in its global spread, due in part to the missionary zeal of US rural sociologists – both today and in the past – who have sowed the intellectual and organizational seeds in a large number of developed and developing countries (Buttel, 2001). Today's sociology of agriculture, for example, has clear implications for labor mobility, hunger and obesity, poverty and spatial inequality, power and cooperation, urbanization and development, social movements and change, and many other substantive areas that almost always link spatial and social processes (e.g., commuting, leisure, cultural change, and media) to the broader sociological enterprise. Simple binary views of urban and rural represent conceptual and empirical roadblocks to understanding. In fact, the rural–urban interface is increasingly a zone of interdependence and social, economic, and political interaction; they arguably should be treated as such in sociological research and training (see Lichter and Brown, 2011). The new rural sociology is a place for disciplinary and substantive growth and innovation.

See also: Agriculture, Economics of; Agroecology and Agricultural Change; Community Sociology; Crime in the Rural Context; Global Migration; Human Ecology; Immigration: Political Aspects; Internal Migration in Industrialized Countries; Migration and Development; Migration: Anthropological Perspectives; Rural Geography; Rural Planning: General; Rural Social Work; Social Ecology: The Chicago School; Urbanization.

Bibliography

Bailey, C., Jensen, L., Ransom, E. (Eds.), 2014. Rural America in a Globalizing World. University of West Virginia Press, Morgantown, WV.

Beaulieu, L.J., 2005. Breaking walls, building bridges: expanding the presence and relevance of rural sociology. Rural Sociology 70, 1–27.

Bonanno, A., 2014. Changing structure of agriculture in the 2010s. In: Bailey, C., Jensen, L., Ransom, E. (Eds.), Rural America in a Globalizing World. University of West Virginia Press, Morgantown,WV.

Brown, D.L., Glasgow, N., 2008. Rural Retirement Migration. Springer, Dordrecht NL.

Brown, D.L., Schafft, K.A., 2011. Rural People and Communities in the 21st Century: Resilience & Transformation. Polity Press, Malden, MA.

Brown, D.L., Swanson, L.E. (Eds.), 2003. Challenges for Rural America in the Twenty-first Century. Penn State University Press, University Park, PA.

Busch, L., Lacy, W.B., 1983. Science, Agriculture, and the Politics of Research. Westview Press, Boulder, CO.

Buttel, F.H., 2001. Rural sociology. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences. Elsevier Ltd, Amsterdam, pp. 13429–13433.

Buttel, F.H., Larson, O.F., Gillespie Jr., G.W., 1990. The Sociology of Agriculture. Greenwood Press, Westport, CT.

Carr, P.J., Kefalas, M.J., 2009. Hollowing Out the Middle: The Rural Brain Drain and What It Means for America. Beacon Press, Boston.

Cloke, P., Marsden, T., Mooney, P.H. (Eds.), 2006. Handbook of Rural Studies. Sage, London.

Dillman, D., Hobbs, D.J. (Eds.), 1982. Rural Society in the US: Issues for the 1980s. Westview Press, Boulder, CO.

Dunlap, R.E., 2010. Climate change and rural sociology: broadening the research agenda. Rural Sociology 75, 17–27.

Flora, C.B., Christenson, J.A. (Eds.), 1991. Rural Policies for the 1990s. Westview Press, Boulder, CO.

Fliegel, F.C., 1993. Diffusion Research in Rural Sociology. Greenwood Press, Westport, CT.

Fortmann, L., Baker-Médard, M., Kelly, A., 2014. Connections: the next decade of rural sociological research on natural resources and the environment. In: Bailey, C., Jensen, L., Ransom, E. (Eds.), Rural America in a Globalizing World. University of West Virginia Press, Morgantown, WV.

Freudenburg, W.R., 2006. Environmental degradation, disproportionality, and the double diversion: reaching out, reaching ahead, and reaching beyond. Rural Sociology 71, 3–32.

Friedland, W.H., 2002. Agriculture and rurality: beginning the "final separation?". Rural Sociology 67, 350–371.

Fuguitt, G.V., Brown, D.L., Beale, C.L., 1989. Rural and Small Town America. Russell Sage Foundation, New York.

Green, J.J., 2014. The status of African Americans in the rural United States. In: Bailey, C., Jensen, L., Ransom, E. (Eds.), Rural America in a Globalizing World. University of West Virginia Press, Morgantown, WV.

Krannich, R.S., 2008. Rural sociology at the crossroads. Rural Sociology 73, 1–21.

Kulcsar, L.J., Curtis, K. (Eds.), 2012. International Handbook of Rural Demography. Springer, New York.

Lichter, D.T., 2012. Immigration and the new racial diversity in rural America. Rural Sociology 77, 3–35.

Lichter, D.T., Brown, D.L., 2011. Rural America in an urban society: changing spatial and social boundaries. Annual Review of Sociology 37, 565–592.

Lobao, L., Meyer, K., 2001. The great agricultural transition: crisis, change, and social consequences of twentieth century US farming. Annual Review of Sociology 27, 103–124.

Lobao, L.M., Hooks, G.M., Tickamyer, A.R., 2007. The Sociology of Spatial Inequality. SUNY Press, Albany, NY.

Massey, D.S. (Ed.), 2008. New Faces in New Places: The Changing Geography of American Immigration. Russell Sage Foundation, New York.

Newby, H., 1983. The sociology of agriculture: toward a new rural sociology. Annual Review of Sociology 9, 67–81.

Sherman, J., 2009. Those Who Work, Those Who Don't: Poverty, Morality, and Family in Rural America. University of Minnesota Press, Minnesota, MN.

Singelmann, J., 1996. Will rural areas still matter in the 21st century? (or) can rural sociology remain relevant? Rural Sociology 61, 143–158.

Smith, K., Tickamyer, A., 2011. Economic Restructuring and Family Well-Being in Rural America. Penn State University Press, University Park, PA.

Sorokin, P.A., Zimmerman, C.C., 1929. Principles of Rural–Urban Sociology. Henry Holt, New York.

Tonnies, F., 1963[1887]. Community and Society (Gemeinschaft and Gesellschaft). Harper and Row, New York.

Wuthnow, R., 2013. Small-Town America: Finding Community, Shaping the Future. Princeton University Press, Princeton, NJ.

Russian and Post-Soviet Studies: Culture

Alexander Etkind, European University Institute, Florence, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by O. Sezneva, volume 22, pp. 14710–14715, © 2001, Elsevier Ltd.

Abstract

Using its custom-made version of Marxism as a road map for the transformation of economy and society, the Bolshevik state acknowledged a role of culture. The early 1920s saw the unbridled creativity of futuristic projects in many domains of high culture. The next decade saw a decisive turn of the Soviet culture toward conservative ideas, with mass violence becoming the routine mechanism of political control. Paradoxically, oppression made culture politically relevant and power culturally productive. The Gulag became the central institution of Soviet culture, and mourning for its victims defined its later developments till the collapse of 1991.

Studies of Soviet culture (1922–91) – possibly, the most ideologically driven culture in modern history – were traditionally polarized by the ideological beliefs of its historians. Now that many of these controversies have passed away along with the Soviet Union, a new post-Soviet view of Soviet culture focuses on its symbolic center and the source of its unique energy: violence. Turning the experience of Stalinism and its crucial manifestation, the gulag system, into the central subject of Soviet cultural studies, this essay looks at the futuristic origins of Soviet culture, its culmination in the state terror of Stalinism, pubic mourning for its victims, and the final decay.

Origins

Collapse of the Russian Empire in the course of World War I opened up the space of Northern Eurasia for utopian ideas and radical practices. One of many competing Russian versions of Marxism, Bolshevism, emphasized the violent break with the past, which in its view could not be achieved by democratic methods. More than other varieties of Marxism, Bolshevism was preoccupied with the state, which it perceived as the agent of radical change while objectifying other spheres – private business, civil society, personal life – as the targets of the state and its transformative efforts. Bolshevism proclaimed the internationalist nature of the new state, which would pursue the global revolution rather than any national interest. Bolshevism saw its base in the union between proletariat and intellectuals, which was supposed to overcome the resistance of the old order. Finally, Bolshevism promoted collectivist values, which would develop the traditions of the organized proletariat rather than of the peasant commune and therefore, cement the institutes of the state rather than undermine them (Stites, 1989; Malia, 1994; Kharkhordin, 1999). Led by Leon Trotsky, the first generation of Bolshevik leaders (many of them former émigrés with years of cosmopolitan experience) understood the revolution in Russia as the first stage of the forthcoming 'world revolution,' which they imagined as a combination of class warfare with decolonization of global empires. In the devastated Europe, this militant utopianism gained a broad acclaim. When World War I ended, Bolshevism was internationally perceived as a menace to the old order. Records of the Paris Peace Conference (1919) document fears of the French, British, and even American statesmen that Bolshevism would spread from the East to the West of Europe, presenting a major alternative to the liberal, anti-imperial but pro-capitalist worldview of Woodrow Wilson's era.

Using 'the science of Marxism' as a road map for the transformation of economy and society, the Bolshevik state inadvertently acknowledged a major role of culture in this endeavor. The issue was not trivial: the Marxist political economy attributed to culture – a part of the 'superstructure' rather than of the 'basis' of social production – a secondary role. However, the radical transformational measures, which the early Soviet leadership implemented in education, labor control, family legislation, hygiene, social work, and other sectors of 'superstructure,' needed a theoretical justification. Trotsky and other leaders sponsored debates about the role of culture in the societal transformation, usually arriving at formulations of 'partial autonomy' of culture in social change. The early 1920s saw the unbridled creativity of futuristic projects in many domains of high culture – architecture, music, theater, visual arts, etc. Despite the brutality of Russia's Civil War, it was still an era of social optimism, which connected cultural innovation with the proclaimed purposes of the Soviet government. Created in 1922, the Soviet Union of four 'socialist republics' embodied the internationalist ideas of early Bolshevism. But in the mid-1920s, power slipped into the hands of Joseph Stalin and the second generation of Bolshevik leaders, who centralized political decisions and economic planning in Moscow, proclaimed the end of the 'world revolution,' and launched some of the most violent operations of the Soviet period, such as 'collectivization' of 1928. The beginning of the 1930s saw a decisive turn of the official Soviet culture toward conservative ideas in a broad range of human and social issues, from neoclassical trends in architecture and urbanism to 'pedological' experiments with testing in high schools, to the attempts at enforcing the 'new Soviet moral' in sex, family, and procreation, to confronting Soviet artists, writers, and musicians with the mandatory, ideologically charged style of 'Socialist Realism.' Gradually and without much reflection, culture was promoted into the rank of the highest value of the state and supervised personally by the

supreme leader, Stalin. Undermining the operative value of traditional values such as salvation, reason, or profit, Stalinism supplanted them by a broad, undifferentiated concept of culture. Now the Soviet state understood 'culture' as a controllable and controlled, inherently resistant but eventually complicit, system of symbols and institutions that promote the major goals of 'socialism.' On an individual level, Soviet culture defined its behavioral manifestation, 'culturedness,' which operated as a measure of personal appropriation of the officially approved culture and a skill of manifesting this culture in everyday life (Volkov, 2000). The ideal of 'culturedness' provided the agents of the state with a coherent, unilinear dimension of social control, which underlined various practices of discipline, assessment, and promotion. This period saw a proliferation of various methods, sometimes pioneering and sometimes imported from the West, which aimed at transforming everyday life by means of high culture: culinary books, methods of training in birth-giving, nursing, swaddling and other aspects of parenting, and mock trials that focused on sexual and marital life of the members of the 'working collective.'

Due to a paradoxical mechanism that Michel Foucault helps to elucidate in his "repressive hypothesis" (Rothberg, 2009), oppression made culture politically relevant and power culturally productive. Like its predecessor, the Russian empire, the Soviet authoritarian regime ambivalently perceived 'culture' as its favorite instrument but also an unknown threat – a weapon of resistance, a source of plurality and violation of the rules, and even a foreshadowing of the revolution. Culture became a screen on which the endangered society saw itself – a unique organ of self-awareness, critical feedback, mourning, and warning (Etkind, 2011).

Nationalities Question

Created as an open, voluntary coalition of formally independent states, the Soviet Union underwent a relentless centralization in political, economic, and military matters. But its leadership tolerated and even promoted a variety of cultural practices in its colonial domains, which spanned from Ukraine to the Caucasus to Central Asia to the Pacific coast to the enormous Arctic region. Trying to find the usable distinction between the cultural and the political, Stalinism endorsed a limited extent of cultural autonomy of the officially recognized 'republics.' Dozens of other peoples of the Soviet Union, who happened to dwell within the Russian or other 'socialist republics,' were deprived of their group rights. Some scholars see in these cultural policies an analog of 'affirmative action' toward republic-making ethnicities (Martin, 2001), but this analogy is limited. While specific agencies of the Soviet government promoted the local variety of folk dance, songs, and poetry in the republics such as Ukraine, Georgia, or Uzbekistan, other agencies caused enormous damage to their peoples, economies, and cultures by centralization, strategic miscalculations, and sometimes (e.g., in the case of the Ukrainian famine of 1932–33) purposeful acts of mass murder. Supporting the folkish manifestations of popular culture in these republics, the Soviet government in Moscow and its agents in the republics invariably suppressed nationalist movements and their cultural manifestations. While republican Academies of Sciences and Unions of Writers used the state support to cultivate the less political aspects of national traditions, public spheres in these republics were deteriorating. Before the collapse of the Union in 1991, no republic was able to create the modern institutions of free press, local self-government, or independent education. Even the central and most powerful of the Soviet republics, Russia, featured the same processes. While Russian folk ensembles flourished under the Soviet rule and Russian classical literature was promoted in schools and universities all over the Soviet Union, Russian nationalist movements were suppressed. Many directions of literature, music, visual arts, and other genres of public culture were effectively frozen. Like the poet Osip Mandelstam, many thousands of writers, artists, musicians, journalists, and architects ended their lives in the labor camps of the Gulag system. Like the writer Mikhail Bulgakov, many authors did not see their texts published in their lifetime. Like the filmmaker Sergei Eisenstein, some authors served the regime with loyalty and talent but still saw their best creations brutally destroyed. Like the composer Dmitri Shostakovich, some spent years in an unfair struggle with the system, complying with its direct orders though engaging in various forms of Aesopian writing wherever they felt they could deceive censorship. Like the writer Vsevolod Ivanov, some worked as the agents of the overcontrolling state while writing their experimental, subversive writings with no hope for publication.

Soviet cultural figures were not only objects of ideological manipulations and violent coercion on the part of the state. Its institutions were also created and run by people who shared the same characteristics – class, ethnic, and human – as those who were censored, tortured, or murdered by them. The Soviet Union's cultural elite was connected to its political elite by a tight network of family connections, friendship, and ideological loyalty. Among the prominent Soviet authors were open enthusiasts of the regime such as Maxim Gorky or Vladimir Mayakovsky, but also its quite dissidents such as Daniil Kharms, stubborn rebels such as Osip Mandelstam, and bewildered victims such as Isaac Babel. Moreover, as biographies of many of these people show, support for the regime could turn into active dissidence, or the other way around. Reflecting this complexity, the historiography of Stalinism has featured a long-lived debate between the so-called Totalitarian School and the Revisionists. It was the Revisionists who carried the day in Western universities in the 1990s though much to their chagrin, it was the Totalitarian model that would largely be embraced in post-Soviet Russia. While the Totalitarianists revealed the top-down control and coercion in Soviet society, the Revisionists emphasized the intrinsic agency of the common people and of the low-level structures of the state. After the demise of the Soviet Union, a new generation of historians has shifted the focus of these debates toward the self-disciplining practices employed in the hearts, minds, and diaries of Soviet citizens, practices which are viewed as the construction of a particular kind of subjectivity (Halfin, 2002; Hellbeck, 2006). One consequence of this body of research has been to displace to the margins the massive violence that was the definitive feature of Stalinism. While Hannah Arendt (1966) and Tzvetan Todorov (1996) see terror as the essence of totalitarian government, the new wave of historiography has

focused on those aspects of Soviet culture and subjectivity that preserve some features of freedom even under terror, such as diaries, autobiographies, and dreams. These approaches, one that focuses on violence and another that focuses on discourse, should be creatively combined. It was the state-sponsored violence of the high Soviet period (1928–53), always massive and ever increasing, which imposed self-disciplinary practices on the Soviet subjects. These practices would not have developed without the omnipresent backdrop of the gulag, where discipline was external, diaries forbidden, and human life returned to its animal origins. During this period, any member of the communist elite could be sacked, arrested, tortured, and killed at any moment and for many reasons, one of which could be insufficient loyalty. An understanding of the environment of coercion, violence, and angst is crucial for the analysis of the ideology, culture, and discursive practices of the Soviet period. The resulting complex of affects – fear, resentment, compassion, and mournfulness – was a permanent feature of the Soviet life.

Violence

While the transformation of human life had been a proclaimed purpose of the Bolsheviks from the start to the end of the Soviet project, methods of this transformation changed considerably during the long period of the Soviet rule. Stalinism relied on mass violence as the major and increasingly, the only method of the desired transformation. Focusing on these links between the enlightening efforts of the state and its murderous practices, Soviet cultural studies have overcome the traditional (particularly for the Totalitarian School) separation of politics from culture. Experience of the Soviet state, which organized life on the enormous space of Eurasia by applying mass violence to its subjects, is crucial for understanding the circular connections between cultural propaganda and political violence. The scholars of Soviet period have produced sophisticated studies of literature, art, music, theater, and other manifestations of high culture of the period. Less is known about the culture of Soviet-era torture rooms, prisons, labor camps, and special settlements. What was the cultural substance of myriads of communications that organized life and death in the huge 'zones' of violence, which were defined politically rather than geographically? How did it happen that mass violence, which we tend to imagine as a vicious circle of uncontrollable and destabilizing events – 'Terror,' as these events are known in both French and Russian – succeeded in establishing a long-living regime of Soviet socialism? As Hannah Arendt (1970) declared, violence cannot create power but actually destroys it. Symbolized and largely realized by the gulag, violence was both catalyzer and corroder of the Soviet culture.

The available estimates of the number of victims of Stalinism range between 5 and 30 millions, and even number of the official 'rehabilitations' ranges, by different accounts, from 1.2 to 4.5 million. Until now, scholars do not have anything like a full list of victims or a full list of executioners. And in the space of Northern Eurasia there are no adequate memorials, museums, and monuments, which could stabilize the understanding of these events for generations to come. However, it was the gulag camp that was the highest – the most direct, complete, and original – manifestation of the Soviet culture. Analyzing the early reports of the survivors of the Nazi and Soviet camps, Arendt (1994, pp. 233, 240, 241) identified in these writings a common tone, which she described as a 'curious air of unreality.' She argued that the incomprehensibility of the camps for their inmates, such that they frequently described their experience there as a nightmare, resulted from the anti-utilitarian character of the camps. From the common sense point of view, "neither the institution itself … nor its political role makes any sense whatsoever." Nothing offends common sense more, Arendt said, than "the complete senselessness" of the camps where the innocent suffer more than the criminal, labor bears no fruit, and crimes do not benefit their authors. This reign of pure violence needs an explanation, but the task of providing one presents an 'extraordinary difficulty.' Responding to this puzzle, Arendt presented the camps as 'laboratories in the experiment of total domination,' a goal that can be achieved only in a 'human-made hell.'

It was not the logic of production that organized the life and work in the gulag, but the logic of torture. Investigative torture has become one of the most memorable features of the Stalinist terror. However, investigative torture was applied mostly in prisons rather than in camps, where millions were kept under a different kind of torture. There in the camps, the effects of hunger, hard labor, untreated illnesses, freezing temperatures, separation from families, isolation from the world, and vulnerability to violent attacks by fellow inmates combined in an overwhelming pain that, being purposefully inflicted by the regime, should also be considered as torture. As Elaine Scarry (1985, p. 56) revealed in her classical analysis, "torture consists of acts that magnify the way in which pain destroys a person's world." Pain, Scarry argues, has a unique ability to destroy language by degrading speech to the sounds anterior to language. If pain destroys the sufferer's world, torture has an additional element: it expands the torturer's self and its self-perceived power. The pain of torture results in the dissolution of the boundary between inside and outside, conflation of private and public, and substitution of the victim's agency by that of the torturers. "The torturer's growing sense of self is carried outward on the prisoner's swelling pain."

Soviet concentration camps were torture camps, not extermination camps, and if people died in high numbers it was the result of negligence rather than purposeful intent. The gulag's primary function was to destroy the convicts' individual 'worlds,' and to substitute it by the official culture. If the need for information, as Scarry shows, is a false justification for investigative torture, the economic needs of the Soviet state were a similarly false justification for the torture endemic to the camps. Economic productivity in the labor camps was about 50% of the average level achieved by free labor in the same industries (Applebaum, 2003; Khlevniuk, 2004; Barnes, 2011). Many of the gulag's projects had no rational justification, and many have never been completed. Those that were finished, such as Belomor Canal, proved to be economically useless. The gulag managers' permanent problem was not a deficit but a surplus of labor. They did not arrest people in order to realize their projects; they invented projects in order to occupy the prisoners. The regime partially recognized the nonutilitarian nature of the camps, when it talked about their ideological, cultural, and psychological – in a word,

transformative – functions. Used universally in the prisons and camps of the gulag era, transformative torture – "re-molding and re-forging human nature," to use the language of Stalinism – efficiently turned those who were generously endowed with language, creativity, and what Arendt called 'the world,' into goners who were indifferent to everything but their next morsel of bread, or their hated neighbor.

Memory and Mourning

After Stalin's death, the enormous system of the gulag melted away with amazing ease, foreshadowing the collapse of the Soviet Union a few decades later. In 1953, there were 166 labor camps in the Soviet Union, plus a huge number of other penitentiary institutions, from prisons to special settlements. About 10 million people served time in these institutions. Almost all of them were released in 1954–56. During this and later periods, cultural memory of terror shaped into an enormous cultural formation, which encompassed different media and genres, incompatible versions of history, and various rituals of mourning. In contrast to the defeated, occupied Germany of the late 1940s, in the Soviet Union of the 1950s it was the very same institutions and personalities that had organized the mass crimes that later revealed them voluntarily. But under their control, this process of working through could not achieve closure. Ironically, Stalin's major achievement, the military strength of the Soviet Union, created a unique situation in which his regime had to analyze, exculpate, and dismantle itself. There was nobody else to do it.

In 1956, the head of the Soviet state, Nikita Khrushchev, started the de-Stalinization process. He was personally implicated in the 'repressions' that he oversaw through the decades of his service as the party leader of Ukraine and Moscow. There was nothing coercing Khrushchev to confess other than his own guilty memory of the terror and his fear of the terror's reenactment. This autonomous, self-imposed character of Khrushchev's revelations makes them unique, even unprecedented in the history of the twentieth-century violence. 'Never again' resounded as a leitmotif both in his famous report to the Twentieth Party Congress (1956) and later in his memoirs. Mourning and warning were tightly linked in his rhetoric, with the latter providing political arguments for the former. You may feel no remorse, he kept warning his peers, and you are not required to confess your own crimes; but if you do not mourn your victims, your fate will be the same as theirs. Khrushchev succeeded in transforming his milieu to such an extent that he escaped the worst: deposed in 1964; he was left in peace to live for many more years, dictated his memoirs, and was buried by his family.

Desperately in need of a conceptual apparatus that neither Marxist tradition nor the Stalinist legacy could provide, Khrushchev coined two operational concepts: 'unjustified repressions,' to embrace mass arrests, tortures, and deportations, and the 'cult of personality,' to describe the accompanying ideological practices. Both concepts served to scapegoat Stalin. 'Unjustified repressions' blamed the preceding ruler for the meaningless, unmotivated man-made catastrophe. The quasi-religious concept of the 'cult of personality' accused Stalinism of violating two fundamental tenets of Marxism – rationality and egalitarianism. These conceptual contributions were substantial, if euphemistic. If the Nazi Holocaust was ended and exposed by others, the Soviet terror was ended and exposed by its former perpetrators, who were also its potential victims. This self-victimization makes it difficult to comprehend, represent, and remember the Soviet terror. In retrospect, its agents and targets seem blurred, agency dispersed, purpose uncertain, causality cyclical, and renunciation and reconciliation incomplete.

After the collapse of the gulag, the victims of repressions were coming back from their camps as uncanny returnees. While perpetrators saw them returning to take a revenge, others imagined the fate of the repressed with horrifying details which mixed oral histories, personal fears, and literary paradigms. The whole country turned into a contact zone between the innocent victims of the gulag and the guilty survivors who managed to escape it. In 1956, Anna Akhmatova said: "Now two Russias are eyeball to eyeball. Those who were in prison and those who put them there." But these two Russias were also aware of the third, which died in the process. Shaping a zone of indistinction between bare life and state law, between repression and rehabilitation, and finally, between mourning and oblivion, this triangular relation determined the cultural dynamics of the late Soviet period.

Mourning and haunting have etched their imprints on many products of late Soviet and post-Soviet culture. A popular writer who outlived Stalinism, Ilia Ehrenburg, named this period 'the Thaw.' Tragically incomplete, the Thaw was the most successful of Russia's de-Stalinization projects. Led by poets and writers, Russia's mourners had their moment of recognition that eventuated in their unprecedented popularity during the Thaw and in the later period. Less than 10 years after Stalin's death, his heirs staged a performative act of major importance: the removal of Stalin's corpse from the Mausoleum on Red Square (1961). Defrosting three decades of Stalinism (1924–53), this remarkable decade ended in 1964, when a vigorous protest from the Communist party against Khrushchev's capricious reforms led to his dismissal. During the long, dreadful stagnation that followed (1964–85), the authorities resumed their attempts to escape from the memory of Stalinism. However, those inconsistent moves revealed as much as they obscured (Etkind, 2013; Jones, 2013). Banished from politics and transferred to culture, the work of mourning became the most sensitive ideological issue in the late Soviet period. Remembering its losses, a postcatastrophic culture lives on through the subsequent generations, as the survivors who struggle with their traumas give way to the descendants who mourn the victims of the catastrophe.

Later in the mid-1980s, Mikhail Gorbachev launched *glasnost*, a sweeping project of public truth-seeking that revealed more about the Soviet past than about Russia's present. Newly published memoirs, archival findings, and popular histories documented the processes, institutions, and personalities of the Soviet terror in unprecedented detail. The great texts of the Thaw that remained unpublished at the time or published abroad – Alexander Solzhenitsyn, Varlam Shalamov, Vasily Grossman – were all published in Russia's large print runs (Toker, 2000; Merridale, 2000; Paperno, 2009). Hundreds of memoirs and autobiographies published in the 1980s dealt with the sufferings of their authors or their parents under the

Soviet regime. In 1993, the literary historian Marietta Chudakova, then an adviser to President Yeltsin, wrote that such writing would fulfill the function of a 'Russian Nuremberg,' staged not in the courtroom as in Germany, but on the pages of memoirs. This hope failed to materialize. In fact, historians and journalists often lament the cynicism of the post-Soviet public, which instead of pluralism has learned a sad lesson about the dependency of historical truth on the changing interests of those in power. Artistic developments of the late Soviet culture, such as 'conceptualism' of Ilia Kabakov or 'necrorealism' of Evgenii Yufit, recycled Soviet symbols and ironically distanced them, contrasting these symbols with the new ambiguous contexts of consumption, death, and self-orientalization.

For the Soviet regime, an insistence on its own definition of truth was crucial; as soon as it faltered in this, it collapsed. In the twenty-first century, Russia's authoritarian 'transition' demonstrates the tortured complexity of its postcatastrophic situation, in which the past haunts the citizenry, divides the society, and limits political choice. If the second and third generations live on the same territory where the catastrophe happened; if the political regime on this territory, despite having gone through multiple transformations since the catastrophe, remains ambiguous in its treatment of the catastrophic past; if the perpetrators are not condemned, the victims are not compensated, the criminal institutions are not banned, the monuments are not built – the memory of the catastrophe acquires intense and peculiar forms. Transforming time in a way that is typical for the posttraumatic condition, the mourned and dreadful loss, the subject of remembrance, continues into the present and shapes the future. The memory of the past becomes indistinguishable from the obsessive fear of its repetition, and the dread of the future takes the shape of compulsory repetitions or creative remembrances of the past. Mourning merges with warning, shaping a temporal zone of indistinction, which combines the past and the future in a joint effort to obscure the present.

See also: Commemoration; Communism, History of; Cultural History; Russian Revolution, The; Socialism: Historical Aspects; Totalitarianism.

Bibliography

Applebaum, Anne, 2003. Gulag: A History of the Soviet Camps. Allen Lane, London.
Arendt, Hannah, 1966. Origins of Totalitarianism. Harcourt, New York.
Arendt, Hannah, 1970. On Violence. Harcourt, Orlando, FL.
Arendt, Hannah, 1994. "Social Science Techniques and the Study of Concentration Camps" [1950], in Her Essays in Understanding. Schocken, New York.
Barnes, Steven A., 2011. Death and Redemption: The GULAG and the Shaping of Soviet Society. Princeton University Press, Princeton, NJ.
Etkind, Alexander, 2011. Internal Colonization. Russia's Imperial Experience. Polity, Cambridge.
Etkind, Alexander, 2013. Warped Mourning. Stories of the Undead in the Land of the Unburied. Stanford University Press, Palo Alto.
Fitzpatrick, Sheila, 1992. The Cultural Front. Power and Culture in Revolutionary Russia. Cornell University Press, Ithaca.
Halfin, Igal (Ed.), 2002. Language and Revolution: Making Modern Political Identities. Frank Cass, London.
Hellbeck, Jochen, 2006. Revolution on My Mind: Writing a Diary under Stalin. Harvard University Press, Cambridge, MA.
Jones, Polly, 2013. Myth, Memory, Trauma: Rethinking the Stalinist Past in the Soviet Union, 1953–70. Yale University Press, New Haven.
Kharkhordin, Oleg, 1999. The Collective and the Individual in Russia: A Study of Practices. University of California Press, Berkeley.
Khlevniuk, Oleg, 2004. The History of the Gulag: From Collectivization to the Great Terror. Yale University Press, New Haven.
Malia, Martin, 1994. Soviet Tragedy: A History of Socialism in Russia. Free Press, New York.
Merridale, Catherine, 2000. Night of Stone: Death and Memory in Twentieth-Century Russia. Granta, London.
Martin, Terry, 2001. The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923–1939. Cornell University Press, Ithaca.
Paperno, Irina, 2009. Stories of the Soviet Experience: Memoirs, Diaries, Dreams. Cornell University Press, Ithaca, NY.
Rothberg, Michael, 2009. Multidirectional Memory. Stanford University Press, Stanford, CA.
Stites, Richard, 1989. Revolutionary Dreams: Utopian Vision and Experimental Life in the Russian Revolution. Oxford University Press, New York.
Scarry, Elaine, 1985. The Body in Pain. The Making and Unmaking of the World. Oxford University Press, New York.
Todorov, Tzvetan, 1996. Facing the Extreme: Moral Life in Concentration Camps. Holt, New York.
Toker, Leona, 2000. Return from the Archipelago: Narratives of GULAG Survivors. Indiana University Press, Bloomington.
Volkov, Vadim, 2000. The concept of Kul'turnost': notes on the stalinist civilizing process. In: Fitzpatrick, Sheila (Ed.), Stalinism: New Directions. Routledge, London, pp. 210–231.

Russian and Post-Soviet Studies: Education

Mark S Johnson, University of Wisconsin–Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Soviet education both shaped and constrained the building of Soviet socialism and of modern Russian and other societies in what is now referred to as the 'post-Soviet space.' While Soviet and international studies of Soviet education diverged sharply on some points of interpretation, almost all agreed that the Soviet educational system as launched in the 1920s, expanded in the 1930s, and consolidated in the 1940s and 1950s was fundamentally functional and conducive to modernization. Despite post-Stalinist and postwar reforms, many of the fundamental features of this system endured until the more radical attempted reforms of the late 1980s, and in many ways even after the dissolution of the Soviet Union itself in 1991. While later Russian and international authors have expressed greater skepticism about the viability and functionality of some aspects of the Soviet educational system, it did undeniably contribute, if in complex and inefficient ways, to the development of modern and literate mass societies in Russia and Eurasia.

Introduction

The vast majority of Soviet and international studies of Soviet education since the 1930s have focused in positive ways on the curricular and instructional rigor of the educational system, and its contribution to Soviet social and economic development. Almost all of this work was framed conceptually within an unproblematic model of 'modernization,' whether seen from the inside by Soviet policymakers or from the outside by sympathetic Western socialists or international experts. The official line that dominated Stalinist and postwar Soviet research argued that the history of Soviet education was a straightforward unfolding of state planning and modern development, with the Communist party-state leading the people upward into modernity and enlightenment through mass literacy, rigorous science and technology programs, scientific Marxism, and militant materialism. Perhaps ironically, this ethos of modernization was echoed in much of the contemporary Western literature, which also considered the contribution of the Soviet educational system to urbanization, industrialization, and social differentiation to be relatively direct and unproblematic. The only real point of contention was over the ultimate effects of rising literacy and the growth of an increasingly educated mass society. While Soviet scholars and educators argued that these developments would lead to the realization of true communism, and thereby establish a model for all of mankind to follow, Western scholars tended to argue that this modernization and rapid educational development would result in cultural and political liberalization on the Western model, or perhaps in the 'convergence' of East and West around a shared vision of the modern welfare state. With the exception of some émigré and dissident scholars, very few analysts seemed to consider the possibility that the Soviet educational system could have been maladaptive or dysfunctional in its ability to contribute to social and economic development, or that official accounts masked potentially serious or systemic educational problems.

This unanimity in almost all published studies of Soviet education since the Russian Revolution of 1917, and especially in studies of Soviet education since the late 1920s when rapid development began in earnest, has also obscured the distinct historical phases and sharp policy conflicts that actually characterized much of that history. In fact, the history of Soviet education is more significant for comparative and international developments than official Soviet accounts would suggest. As detailed below, the years after the Russian revolutions of 1917 were characterized by an official policy of radical progressive or child-centered education that was very much influenced by, and directly contributed to, broader international trends in educational policy. This was followed by increasingly conservative educational policies in the Stalinist 1930s and after, when the enduring foundations of the modern Soviet educational system were established. In fact, while obscured by Soviet social and behavioral science, which always claimed that Soviet developments were utterly unique (and by ideological definition more advanced than anywhere else in the world), this traditionalist restoration was actually also part of a broader international turn away from Deweyan progressive education and experimental schooling in many nations during the middle decades of the twentieth century. Many of these historical phases in educational policy, and the political and professional struggles that shaped them, have long been obscured by official Soviet historiography. While there were undeniably remarkable achievements and rapid educational expansion, the very real limits and failures of Soviet education have also long been obscured by opaque Soviet statistics and official accounts. More recent post-Soviet research, both Western and Russian, has begun to detail how, throughout its history, Soviet education was successful in many ways, and yet also heavily politicized, narrowly vocational, bureaucratically inflexible, and subordinated to the Soviet state and the Communist Party of the Soviet Union (CPSU). Ultimately, it could be argued that the inflexibility of the system, and its difficulty in adapting to the needs of postwar social and economic development, caused Soviet education to become a serious 'brake' on the viability of Soviet socialism. During the post-Stalinist period, from the 1950s through the late 1980s, various reforms were attempted in an effort to 'humanize' and differentiate the rigid and intellectually stifling system. While partially successful, the limitations of these reforms then led, in the post-Soviet years after 1991, to

attempted radical reforms, questioning of the entire Soviet legacy, and to deep system-wide educational changes in the 1990s. While the Russian state has attempted to restore funding and academic standards since 2000–01, and has continued to pursue closer links with the global educational community, serious educational problems remain in many rural and post-industrial regions, and educational conditions remain uneven in many of the other post-Soviet successor states.

Soviet Educational Experiments and Radical Progressive Education

The models for what became Soviet or socialist education were sketched out in the decades before the Russian revolutions of 1917 in response to the need for rapid social and economic development, and especially in response to the poverty and dislocations caused by World War I. The fundamental principles of Soviet progressive education included expanding the conception of education and culture to include a broad array of programs for illiterate adults, working people, peasants, and children, including early childhood care. These policies also entailed the provision of extensive public health and welfare services through educational institutions, and a radical rethinking of the nature of curriculum and instruction to bring learning 'closer to life' and industrial work. Finally, educational access and promotion were to be radically expanded to make education open to all regardless of age, ethnicity, class, or gender (Medynskii, 1925; Pinkevitch, 1929).

These early Soviet educational policies were inspired in part by European and American progressive or child-centered pedagogy, and were hailed as revolutionary by Western radicals and some educators. However, Soviet education then went far beyond Western influences in its emphasis on including previously disadvantaged populations, as well as in its increasingly dogmatic emphasis on political indoctrination (Harper, 1929). Soviet educational experiments with adult and community education were, in many ways, far ahead of their time (for an encyclopedic summary, see Kalashnikov and Epshtein, 1927–29; also Fradkin, 1990). Progressive Soviet educators sought to synthesize general and vocational education for all students in a common curriculum that was known as 'polytechnicalism,' although some vocational–technical tracking was later reintroduced. The educational experiments of the 1920s also included interdisciplinary and thematic curricula (to tie together studies of 'man,' 'nature,' and 'labor,' for example); sophisticated developmental instruction that was carefully targeted to stages of students' psychological and intellectual development (as pioneered by Lev Vygotskii and others); and efforts to foster local studies (*kraevedenie*) and culturally relevant pedagogies for national minority populations. Other innovations included student and faculty self-government, a focus on experiential learning and community service, and the curtailing of most formal assessments and criteria for promotion in the name of social equality and universal access to education. Early Soviet education was also characterized by the extensive use of Western-style child study and psychological testing, known in Soviet practice as 'pedology,' or the 'Marxist science of the child.' Resources were also invested in the mass provision of new media, libraries, and extracurricular activities for all ages, albeit with a strong political or ideological content. Finally, the 1920s also witnessed the development of modernized languages, new alphabets, primers, and comprehensive educational services for many indigenous peoples and the smaller ethnic minorities of the Soviet Union. These efforts also included ambitious programs to foster new 'national' cultures for the larger nationalities of the USSR (such as Ukrainians, Kazakhs, Uzbeks, Belarusians, Georgians, Armenians, and others; for a Western account that emphasizes modernization, see Pennar et al., 1971), yet again all within rigid ideological constraints.

Almost all accounts of these developments in the 1920s, both Soviet and Western, were extremely positive about the early progressive experiments, although some Western observers such as John Dewey (1929) and Samuel Harper (1929) were also critical of the Soviet emphasis on ideological indoctrination and their philosophical dogmatism. For all of the early Soviet educators' idealism, however, there were clearly also many internal contradictions and problems with the progressive experiments of the 1920s (Holmes, 1991). For example, while expanding access to previously disadvantaged populations, the communist regime also sought to deny or limit educational opportunity to members of the prerevolutionary elites or to real or imagined political and class enemies. This created serious social tensions, and drove many of those who could have become leading researchers, teachers, and students away from Soviet educational institutions. Public or state education was also often funded on a 'residual' basis, which meant that it was financed only after other budgetary obligations were met, even as the Soviet regime categorically prohibited any private funding from compensating for the inadequacies of state provision. Nonetheless, early Soviet progressive education, for all its utopianism and politicization, was undeniably an unprecedented and truly revolutionary experiment in the provision of mass education and culture for all. In fact, despite its internal complications and mixed successes, the image of Soviet 'education for all,' which was disseminated globally in effective propaganda campaigns, helped to inspire drives for gender equity and mass access in many Western nations as well as to inspire anticolonial educational movements and antiracism education in the non-Western world.

Stalinist Traditionalism and System Building, 1930s–50s

However, as detailed in subsequent Soviet and international research, the many problems with these early experiments also helped to discredit them, and provoked a sharp reaction in the 1930s. Soviet accounts during the period of the Stalinist dictatorship tended to blame the progressive educators for their allegedly 'anti-Leninist' errors and 'petty-bourgeois' utopianism, and even went so far as to accuse them of deliberate sabotage and counter-revolutionary 'wrecking' (Svadkovskii, 1938). As a result of these increasingly virulent attacks, many Soviet progressive educators and teachers were purged in the 1930s and their lives and work erased from the historical record, only to be partially rediscovered when Soviet archives began to open up in the late 1980s (Fradkin, 1990). Later

official Soviet accounts glossed over the distinctive progressive and experimental policies of the 1920s, obscured the sharp factional conflicts of the time, and blamed the progressive educators for their alleged impracticality and adherence to inappropriately 'bourgeois' pedagogical theory (Kuzin et al., 1980). Many international analysts simply argued that the progressive policies were not conducive to rapid social and economic modernization, and that it was thus virtually inevitable that more 'stable' and traditional curricula and instructional methods would be restored to fit the Soviet educational system to the needs of modern industrial society (Anweiler, 1964; also Pennar et al., 1971). Other Western analyses of the failure of the progressive experiments focused on issues such as the acute lack of funding that crippled the attempted reforms; popular resistance by many parents and teachers (Holmes, 1991); and the broad social support for the upward mobility opened up by the Stalinist educational restoration (Fitzpatrick, 1979). Almost all accounts since the 1930s agree that the progressive experiments became increasingly unpopular as they were accused of undermining the authority of teachers and parents, weakening curricular integrity and instructional rigor, and failing to produce the highly trained technical specialists and skilled workers needed to 'build socialism' according to Stalinist tenets. From a broader and more critical perspective, with the consolidation of the Stalinist dictatorship in the 1930s, the task became less to mobilize and empower all citizens through popular enlightenment and more to train specialists (Bailes, 1978) and to indoctrinate subjects for state-directed industrial and military development.

Struggles over educational and social policy raged on in the 1930s, and resulted in the thorough repudiation of progressive education, and the forging of a new and intensely conservative model of Soviet education. This also entailed the rejection of child-centered or developmental schooling in the name of the struggle against 'pedological deviations' and a return to a school- and teacher-centered system (Ewing, 2002). This traditionalist restoration also led to the imposition of increasingly elaborate systems of formal 'didactics,' as codified in a series of mandatory mass-produced lesson plans, textbooks, and assessments (Kairov, 1939). This Stalinist educational model was intensely traditional in curriculum and instructional methods; thoroughly politicized, especially in history and the social sciences; heavily dominated by the physical sciences and engineering; and rigidly bureaucratized and hierarchical in its administration. In fact, the CPSU prioritized control in educational administration and the social sciences, but Stalinist ideology saturated all academic disciplines, often with debilitating effects. From a more critical perspective, it could be argued that the tight system of party-state control over all aspects of Soviet intellectual and professional life that resulted from these struggles became, over the following decades, debilitating for the Soviet educational system as a whole. This was perhaps especially true in regard to the educational system's global connections and competitiveness, with near total isolation reinforced by ruthless censorship and limits on foreign travel for scholars and students.

The fundamental principles of this Stalinist or traditionalist restoration are well known through the intense international attention that Soviet education received in the postwar era, especially after the launch of the first earth-orbiting satellite (Sputnik) in 1957, which was followed by other Soviet 'triumphs' in the Cold War technology race.

The Stalinist educational system was organized through content-heavy and rigorous discipline-based curricula, and embraced traditional instructional methods such as lectures that compelled students to master and reproduce that content knowledge. The system was built upon a canon of standardized textbooks and teaching guides, and emphasized lecturing, memorization, and rigorous oral and written examinations. More positively, this system was also buttressed by a comprehensive array of teaching guides and specialized publications for teachers, school libraries, and professional development opportunities (Long and Long, 1999). An essential goal throughout was to reinforce the authority of the party-state over the entire system, of school directors and administrators over their staff, and of teachers and professors over their students. Formal assessments and rigid criteria for promotion were elaborated, with harsh sanctions for failure, lack of discipline (real or imagined), or dissidence. Many of the programs that sought to cultivate the languages and cultures of indigenous peoples and national minorities were curtailed, and many ethnic groups subjected to involuntary linguistic Russification and cultural Sovietization (Shorish, 1984), often enforced through police terror. Overall, Soviet education in the 1930s and after became increasingly rigorous, hierarchical, and formal (Konstantinov and Medynskii, 1948). Despite the damage done by Stalinist terror and World War II (Dunstan, 1997), Soviet education in this period was permeated with a spirit of discipline and order, and was closely coordinated with the Soviet planned economy and the needs of the Stalinist military–industrial complex.

There were, undeniably, real successes for Soviet education in the early Soviet and Stalinist periods, especially in raising literacy rates and building a comprehensive mass or public education system. It also seems that the traditionalist restoration was quite popular with the new Soviet elites, many professional educators and teachers, and many culturally conservative parents and officials (Holmes, 1999; Hoffman, 2003), even if some objected to its more extreme measures such as attempts to end coeducation during World War II and after (Ewing, 2010). The Soviet education system did produce unprecedented numbers of technical specialists and skilled and semiskilled workers, albeit in increasingly narrow vocational specializations. The system also proved successful at some forms of research and development, especially in mathematics, the physical sciences, and military technologies, yet at the same time built stultifying ideological filters around many disciplines and isolated researchers in 'closed' cities and secret military research institutes. As the Stalinist regime embraced conservative school practices and cultural values, it arguably also fostered compensatory legitimacy for its otherwise intensely repressive and austere version of state socialism, as demonstrated by both Soviet and émigré memoirs. Yet for all its claims to universality and unanimity, unofficial cultural and educational traditions did survive, especially among religious and national minorities, as well as among the Soviet intelligentsia in the form of unsanctioned teaching practices (Kline, 1957) and private and family tutoring.

Official Soviet research in this period was, by definition, triumphal about the onward and upward march of literacy and

enlightenment, and about the indispensable role of the leadership of the Soviet party-state (Medynskii, 1951; Petrovskii, 1953). There was also a distinct and growing tone of xenophobia, with Soviet developments depicted as both entirely self-contained and superior to all comparable international trends and technologies. The emphasis on the singular role of state power tended to obscure the more complex relations between the party-state and admittedly dependent interest groups within the teaching and academic professions, as well as the role and interests of the Soviet public, however, muted or constrained the latter (Johnson and Kortunov, 2011). In fact, one could argue that the primary mechanism to meet public 'demand' throughout these decades was the continued rapid expansion of the system at all levels, even if public opinion per se was often irrelevant in policymaking, and even as selective repression and involuntary educational tracking and career assignment continued unabated. Judgments of this system in the 1950s ranged from the triumphal official Soviet accounts to more nuanced international analyses. For example, the disillusioned Western progressive George Counts (1957) was quite harsh in his judgment of the repressive and dogmatic aspects of Soviet education, yet he also acknowledged its rigor and utility for a forced march version of modernization (see also Korol, 1958; Pennar et al., 1971). Almost all other international and émigré accounts were positive in their analyses of the system's curricular and instructional principles and its effectiveness, even if they sometimes also acknowledged its rigidity and intellectual limitations.

However, alongside the successes and popularity of the traditionalist restoration, there were severe and chronic problems with Soviet education as well. The human cost of the Stalinist terror in education and culture was devastating, and resulted in the repression of many talented researchers, educational leaders, and teachers, especially among the national minorities. While adhering to the basic principles of modernization theory, some Western authors did draw attention to the more ambiguous effects of Soviet standardized education for the development of the non-Russian peoples and languages of the Soviet Union. The system became extremely top heavy and driven by administrative interests, with little space for professional autonomy or intellectual innovation. Research in the highly specialized institutes of the Academy of Sciences was often separated from teaching in the universities and other specialized postsecondary institutions. The entire system was permeated by the belief that theory – as shaped by political ideology – should drive educational policy and pedagogical practice. Furthermore, educational institutions and professional networks were often isolated from one another through bureaucratic parallelism, as each economic sector and industry in the planned economy created its own educational and research hierarchies, which often competed for budget resources. Educational institutions and even individual teachers were closely supervised, even as some contrived elaborate and creative ways to work around or subvert party-state controls. Most significantly, the entire system was structured into narrow vocational–technical specializations that limited the development of the job skills, professional initiative, and creativity that were needed for sustainable postwar social and economic growth. These more critical perspectives only began to coalesce in the 1980s and after (Dunstan, 1992), as some late Soviet and international authors began to focus on the more maladaptive or dysfunctional aspects of the official Soviet model. In conclusion, and taken as a whole, the Stalinist or traditionalist reforms in Soviet education and research successfully and rapidly built a mass education system, but arguably did so at the price of bureaucratic rigidity, intellectual stagnation, and vocational overspecialization.

The Potential and Limits of Postwar Reforms in Soviet Education

Ironically, even as international researchers were popularizing the curricular and instructional rigor of Soviet education in the post-Stalinist period after 1953 and especially after 1957, Soviet educators began to quietly debate the inflexibility and limits of their own system. Attempted postwar reforms encompassed both more 'progressive' as well as some more 'traditional' ideas, and not all were successful. For example, an attempted reform associated with Communist Party leader Nikita Khrushchev in 1958 sought to restore the 'polytechnical' principles of the system, only this time as more rigorous and involuntary vocational tracking. This attempted reform sought to force down the educational aspirations of the Soviet public through a requirement that all students entering higher education have at least 2 years of work or production experience (Goncharov and Korolev, 1960). Significantly, this reform was essentially blocked by the combined resistance of enterprise managers, parents concerned about their children's upward mobility, and more traditionalist educators resistant to such experiments. The regime also attempted, throughout the post-Stalinist period, to strengthen 'civic-social' indoctrination in response to what was seen as growing juvenile delinquency, alienation from official ideological values, and subversive cultural influences from the West.

Other reforms in the post-Soviet period reflected the more progressive currents of the 1920s, especially the tradition of social-cultural psychology pioneered by Lev Vygotskii (1896–1934, see Vygotskii, 1991). Vygotskean cultural psychology and developmental instruction, which also received significant attention from Western researchers in the 1980s and 1990s (Wertsch, 1985), may well prove to be the most internationally significant and enduring contribution of the Soviet educational system. In a very real sense, Vygotskii's revolutionary conviction that, with a proper educational environment and sophisticated teaching, children of all backgrounds and even the disabled could be drawn up successfully through their 'zones of proximal development,' contributed to a fundamental conviction that all children could be successfully educated. Other attempted reforms in the postwar period included greater curricular differentiation; the introduction of electives for some students (Matthews, 1982); and the development of programs for talented and gifted education in the arts, sciences, and sports, as well as increased attention to disabled students (Riordan, 1988). There was also a relaxation of linguistic Russification, with relatively more flexibility offered to educational institutions in the national minority regions, and more latitude for the use of national languages. Other reforms in this period focused on shifting students out of low-quality part-time and 'correspondence' (*zaochnyi*) instruction and

into full-time academic programs. Yet it must be noted that, for all the potential of these and other ambitious reforms, they were often limited by budget constraints, or thwarted by the combined power of party-state administrators and the conservative educational establishment, perhaps especially in the various Republic-level Ministries of Education and the Academy of Pedagogical Sciences of the USSR, which later became criticized as a 'citadel of dogmatism.'

Taken as a whole, the reform movements in Soviet education from the 1950s through the early 1980s did have the effect of consolidating and extending the system, while also diversifying and enriching curricular and instructional practices. Perhaps most importantly, the Soviet educational system by the 1970s had trained a vast new body of professional teachers and researchers, and had achieved essentially universal literacy and nearly universal secondary education. Not surprisingly, official Soviet accounts focus on these achievements with little or no acknowledgment of any persistent problems or struggles over policy or practice (Kolmakova et al., 1986). Yet these achievements were, to a significant degree, gained at the expense of individual and professional autonomy, creativity, and systemic flexibility. It must also be noted that, for all of its rhetoric of equality and universality, severe inequities persisted, especially between urban and rural areas, and that literacy and educational provision lagged badly behind in many of the national minority regions. The Soviet educational system was also slow to introduce new instructional and informational technologies, and became gradually weaker and more sporadic in its provision of nonpolitical continuing education and vocational retraining.

It seems that little international attention was paid to these systemic problems, perhaps in part because sustained Western attention to Soviet education, which grew dramatically after 1957, had narrowed considerably by the early 1980s. It also seems that sustained international cooperation with Soviet educators was constrained not only by Cold War limits on travel and censorship, but also by the conservatism of many of the neo-Stalinist educators and officials who continued to dominate the system well into the 1980s.

The Crisis of Soviet Education and Its Ambiguous Contemporary Legacies

The final phase of Soviet education in the 1980s and early 1990s witnessed both radical reforms and severe systemic crises. Several attempted reforms in the early and mid-1980s sought to tighten vocational tracking and increase party-state ideological controls, but in a very real sense Soviet educated society as a whole, as well as some elements within the educational profession who were no longer willing to tolerate such reactionary policies. There was a growing sense that Soviet public education was sliding into crisis, a systemic crisis shaped by too much uniformity and rigidity and too little creativity and autonomy. This led to the emergence of a growing 'innovation movement' among educators and teachers, which emphasized Vygotskean developmental instruction, greater curricular freedom, and more child-centered pedagogy (Bazhenova, 1987; Kerr, 1990). This movement sought to humanize and democratize Soviet education, to loosen party-state controls, and foster greater professional autonomy, to recover the spirit of the progressive experiments of the 1920s (Fradkin, 1990), and especially to expand contacts with international researchers and practitioners (Muckle, 1991; also Eklof and Dneprov, 1993).

After the collapse of the Soviet Union in late 1991, the 15 newly independent republics embraced an array of educational policies. Some tried to essentially preserve the centralized Soviet system, as in Belarus and parts of Central Asia, with very uneven success. Others struggled to radically reform the Soviet system through attempted privatization and decentralization, as in the Baltic nations and the Russian Federation (Dneprov et al., 1991). Other analysts argued that the system's institutional conservatism and pedagogical inertia actually helped preserve some of its stronger features amid economic crisis and social upheaval (Webber, 2000). The reforms of the 1990s have, undeniably, unleashed a veritable educational revolution of innovation in curriculum and instruction, and finally joined together the long isolated Russian educational community with its international counterparts. Much of this reintegration into the world community of educational research and practice was facilitated by an unprecedented level of international assistance provided to post-Soviet educators and students in the 1990s, in the form of research grants and vastly expanded new educational exchanges. Yet, it must also be noted, all of the Soviet successor states experienced severe if not catastrophic crises in the funding and provision of education, and that as a result both excellence and equity suffered severely in the early and mid-1990s. It could be argued that all the Soviet republics, including the long and still-dominant Russian Federation, underestimated how entwined and interdependent both the Soviet economy and its related system of education and research really were, and how much damage would be done by its abrupt dissolution. Tragically, once post-Soviet educators and teachers were able to finally rejoin the world community from which they had been isolated since the 1920s, many lacked the financial resources to remain in the profession, or to take full advantage of the new opportunities for international cooperation or professional development. Furthermore, the economic and demographic crises that gripped the entire region in the 1990s directly threatened the health and readiness to learn of an entire generation of children and young adults. While the collapse of the Soviet Union opened up unprecedented new personal freedoms and economic opportunities for some, it also destroyed much of the socialist safety net of public education and guaranteed employment, and plunged many millions into poverty. To summarize, while the legacies of the Soviet educational system were mixed, the ultimate effects of post-Soviet reforms and educational policies remain unconsolidated, and there is a clear nostalgia in some social circles and regions for what has been lost.

Conclusions

In conclusion, Soviet education both shaped and constrained the building of Soviet socialism and of modern Russian and other societies in what is now referred to as the 'post-Soviet space.' While Soviet and international studies of Soviet education diverged sharply on some points of interpretation,

almost all agreed that the Soviet educational system as launched in the 1920s, expanded in the 1930s, and consolidated in the 1940s and 1950s was fundamentally functional and conducive to modernization. Despite post-Stalinist and postwar reforms, many of the fundamental features of this system endured until the more radical attempted reforms of the late 1980s, and in many ways even after the dissolution of the Soviet Union itself in 1991. While later Russian and international authors, perhaps especially émigré and dissident observers, expressed greater skepticism about the viability and functionality of some aspects of the Soviet educational system, it did undeniably contribute, if in complex ways, to the development of modern and literate mass societies in Russia and Eurasia. Finally, the most rigid and repressive aspects of the system were repudiated in the late 1980s and especially during the 1990s, even if the urgency of recent economic crises have obscured the urgency of such historical analysis.

See also: Russian and Post-Soviet Studies: Culture; Russian and Post-Soviet Studies: Gender; Russian and Post-Soviet Studies: Politics; Russian and Post-Soviet Studies: Society.

Bibliography

Anweiler, O., 1964. Geschichte der Schule und Padagogik in Russland, vom Ende des Zarenreiches bis zum Beginn der Stalin-Ara. Quelle & Meyer, Heidelberg, Germany.

Bailes, K., 1978. Technology and Society under Lenin and Stalin: Origins of the Soviet Technical Intelligentsia, 1917–1941. Princeton University Press, Princeton.

Bazhenova, I.N. (Ed.), 1987. Pedagogicheskii poisk [Pedagogical search]. Pedagogika, Moscow.

Counts, G.S., 1957. The Challenge of Soviet Education. McGraw Hill, New York.

Dewey, J., 1929. Impressions of Soviet Russia and the Revolutionary World, Mexico-China-Turkey. New Republic, New York.

Dneprov, E.D., Lazarev, V.S., Sobkin, V.S. (Eds.), 1991. Rossiiskoe obrazovanie v perekhodnoi period: programma stabilizatsii I razvitiia [Russian Education in the Transition Period: Program of Stabilization and Development]. Ministry of Education of the Russian Soviet Federated Socialist Republic, Moscow.

Dunstan, J. (Ed.), 1992. Soviet Education under Perestroika. Routledge, New York.

Dunstan, J., 1997. Soviet Schooling in the Second World War. St. Martin's Press, New York.

Eklof, A.B., Dneprov, E.D., 1993. Democracy in the Russian School: The Reform Movement in Education since 1984. Westview, Boulder, CO.

Ewing, E.T., 2002. The Teachers of Stalinism: Policy, Practice and Power in Soviet Schools in the 1930s. Peter Lang, New York.

Ewing, E.T., 2010. Separate Schools: Gender, Policy and Practice in Postwar Soviet Education. Northern Illinois University Press, De Kalb.

Fitzpatrick, S., 1979. Education and Social Mobility in the Soviet Union, 1921–1934. Cambridge University Press, New York.

Fradkin, F.A., 1990. A Search in Pedagogics: Discussions of the 1920s and 1930s. Progress, Moscow.

Goncharov, N.K., Korolev, F.F. (Eds.), 1960. Novaia systema narodnogo obrazovaniia v SSSR: Sbornik dokumentov I statei [The New System of Public Education in the USSR: Collection of Documents and Articles]. Academy of Pedagogical Sciences of the RSFSR, Moscow.

Harper, S.N., 1929. Civic Training in Soviet Russia. University of Chicago Press, Chicago.

Hoffman, D., 2003. Stalinist Values: The Cultural Norms of Soviet Modernity, 1917–1941. Cornell University Press, Ithaca.

Holmes, L., 1991. The Kremlin and the Schoolhouse: Reforming Education in Soviet Russia, 1917–1931. Indiana University Press, Bloomington.

Holmes, L., 1999. Stalin's School: Moscow's Model School No. 25, 1931–1937. University of Pittsburgh Press, Pittsburgh.

Johnson, M.S., Kortunov, A.V., 2011. The state, the university and society in Soviet and Russian higher education: the search for a new public mission. In: Calhoun, C., Rhoten, D. (Eds.), Knowledge Matters: The Public Mission of the Research University. Columbia University Press and Social Science Research Council, New York.

Kairov, I.A., 1939. Pedagogika [Pedagogy]. Uchpedgiz, Moscow.

Kalashnikov, A.G., Epshtein, M.S. (Eds.), 1927–1929. Pedagogicheskaia entsiklopediia [Pedagogical encyclopedia], 3 vols. Rabotnik prosveshcheniia, Moscow.

Kerr, S.T., 1990. Will glasnost' lead to perestroika? Directions of educational reform in the USSR. Educational Researcher 18 (7), 26–31.

Kline, G. (Ed.), 1957. Soviet Education. Columbia University Press, New York.

Kolmakova, M.N., Gellershtein, L.S., Ravkin, Z.I. (Eds.), 1986. Ocherki istorii pedagogicheskoi nauki v SSSR (1917–1980 gg) [Notes on the History of Pedagogical Science in the USSR 1917–1980]. Pedagogika, Moscow.

Konstantinov, N.A., Medynskii, E.N., 1948. Ocherki po istorii sovetskoi shkoly RSFSR za 30 let [Notes on the History of the Soviet School in the Russian Soviet Federated Socialist Republic at 30 Years]. Gosuchpedgiz, Moscow.

Korol, A.G., 1958. Soviet Education for Science and Technology. Technological Press of the Massachusetts Institute of Technology, Cambridge.

Kuzin, N.P., Kolmakova, M.N., Ravkin, Z.I. (Eds.), 1980. Ocherki po istorii shkoly I pedagogicheskoi mysli narodov SSSR, 1917–1941 gg. [Notes on the History of the School and Pedagogical Thought of the Peoples of the USSR, 1917–1941]. Pedagogika, Moscow.

Long, D., Long, R., 1999. Education of Teachers in Russia. Greenwood Press, Westport, CT.

Matthews, M., 1982. Education in the Soviet Union: Policies and Institutions since Stalin. Allen & Unwin, London.

Medynskii, E.N., 1925. Entsiklopediia vneshkol'nogo obrazovaniia, second ed. Gosizdat, Moscow.

Medynskii, E.N., 1951. Public Education in the USSR, second ed. Foreign Languages Publishing House, Moscow.

Muckle, J., 1991. Portrait of a Soviet School under Glasnost. Palgrave Macmillan, New York.

Pennar, J., Bakalo, I.I., Bereday, G.Z.F., 1971. Modernization and Diversity in Soviet Education: With Special References to Nationality Groups. Praeger, New York.

Petrovskii, I.G., 1953. Higher Education in the U.S.S.R. Soviet News, Moscow.

Pinkevitch, A.P., 1929. The New Education in the Soviet Republic. John Day, New York.

Riordan, J., 1988. Soviet Education: The Gifted and Handicapped. Routledge, New York.

Shorish, M., 1984. Education of ethnic minorities in the Soviet Union. In: Altbach, P.G., Kelly, G. (Eds.), Education and the Colonial Experience, Transaction. New Brunswick, New Jersey.

Svadkovskii, I.F. (Ed.), 1938. Protiv pedologicheskikh izvrashchenii. Sbornik statei [Against Pedological Deviations. Collection of Articles]. Gosuchpedgiz, Moscow-Leningrad.

Vygotskii, L.V., 1991. In: Davydov, V.V. (Ed.), Pedagogicheskaia psikhologiia [Pedagogical Psychology]. Pedagogika, Moscow.

Webber, S.L., 2000. School, Reform and Society in the New Russia. St. Martin's Press, New York.

Wertsch, J.V., 1985. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge.

Russian and Post-Soviet Studies: Environment

Charles E Ziegler, University of Louisville, Louisville, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Throughout the period of Soviet socialism (1917–91) environmental considerations were subordinate to state imperatives of rapid industrial development. Although there existed small enclaves of scientists and specialists committed to conservation and ecology issues, centralized authoritarian rule and state repression made environmental activism virtually impossible, while censorship obscured the scope and severity of ecological problems. Environmental groups mushroomed in the more liberal late-Soviet and early post-Soviet period, but the unrestrained capitalism and natural resource-based economy of the Putin era (1999–) relegated ecology protection to the margins of Russian politics.

Soviet Socialism and the Environment

Since the Soviet Union was the largest country in the world, the development and environmental policies enacted by the communist state impacted approximately one-sixth of the world's total landmass. Throughout Soviet history the main goal of the ruling Communist Party was economic expansion, in largely quantitative terms, with little regard for the environmental consequences of its industrial and agricultural policies. Ecological awareness did emerge in the 1960s and 1970s, roughly parallel with the development of environmental movements in Western countries, but the regime's response to environmental degradation was tepid. When censorship and political repression were relaxed in the late 1980s under General Secretary Mikhail Gorbachev (1931–), environmental movements mushroomed across the USSR. In the non-Russian republics these ecology groups often assumed nationalist overtones, accentuating demands for separation from Moscow and ultimately contributing to the dissolution of the USSR.

Under Vladimir Lenin (1870–1924), Joseph Stalin (1879–1953), and Nikita Khrushchev (1894–1971), the fixation on producing ever greater quantities of industrial and agricultural goods through the mechanism of the centrally planned economy left little room for environmental considerations. The state controlled virtually the entire economy, and at most times opposition to development goals was not tolerated. In the more relaxed political environments of the 1920s, the 1960s, and the 1980s, concerned scientists and influential cultural figures pressed the regime to be more environmentally responsible. For example, Douglas Weiner (1988) showed that conservationists worked assiduously in the 1920s and early 1930s to preserve and expand the system of nature reserves (zapovedniki) in the face of hostile Soviet developers. Ultimately the Stalinist industrialization regime, with its utilitarian view of nature, prevailed. A modest conservation movement was resurrected in the late 1950s, but it was not until the early to mid-1960s that a modern environmental consciousness was sparked by the debate over threats over Lake Baikal.

This geologically ancient lake, situated in East Siberia and holding one-fifth of the world's supply of fresh water, was threatened by pollution from a large cellulose factory, ostensibly designed to produce cord for military aircraft tires. Outraged by the danger to this natural treasure – Baikal was designated a World Heritage Site by the United Nations in 1996 – scientists and writers were joined by a few key officials in a campaign to save the lake. The Communist Party and government responded by passing resolutions creating a protected zone around Baikal and mandating the installation of water purification facilities. While in many respects a unique incident, the campaign to save Baikal was one example of the regime responding to demands from influential sectors of the public.

By the late 1960s policy makers were increasingly aware of the costs to the national economy and the dangers to public health associated with pollution, particularly of the country's surface waters. The problems were severe. Heavily polluted bodies of water included the Volga and Don rivers, Lake Ladoga, the Caspian, Black, Azov, and Baltic seas, Lake Sevan, and many others. A study prepared in the 1970s by the Institute of Geography for top government officials, and declassified in 1989, mapped out 4 million square kilometers of Soviet territory as ecological disaster areas, nearly one-fifth of the entire country. Half of the most polluted territory was located within the ecological fragile Far North and Siberia.

Egregious examples of environmental destruction were found throughout the USSR. In Central Asia irrigation for cotton production reduced the flow of the Amu Darya and Syr Darya rivers by 90%. These rivers fed the Aral Sea, a body of water the size of West Virginia. In less than three decades, the sea's level had dropped by 40 feet (13 m), its area decreased by 40%, and its volume decreased by two-thirds. The fishing industry collapsed, and dust storms blew toxic salts and agricultural chemicals across neighboring regions, elevating rates of intestinal illness, anemia, cancer, and birth defects (Pryde, 1991). Uzbekistan and Kazakhstan experienced the worst effects of the sea's desiccation; reportedly, three-fourths of the population in the Karakalpak autonomous republic in northern Uzbekistan was ill in the late 1980s, and the infant mortality rate had soared to 90 per 1000 live births. Kazakhstan also suffered from nuclear testing at the Semipalatinsk range and from pollution associated with the Baikonur cosmodrome, home to the Soviet space program. Oil-producing regions in Siberia and Azerbaijan were heavily polluted by spills and

leaks, and large areas of western Siberia were contaminated by nuclear waste from defense plants.

There were few bright spots in the Soviet environmental record. Some analysts pointed to the central thermal stations that provided heat to apartment buildings and workplaces as one advantage of the centralized economy although the system was clearly inefficient. The Soviet zapovednik system, which consisted of approximately 160 nature reserves encompassing over 54 million acres (22 million hectares), protected unique species of flora and fauna. In studying the history of the zapovedniki, Weiner (1999) found that even during the most repressive years Soviet scientists asserted some autonomy and preserved a distinct social identity separate from the party-state in their efforts to defend and promote these pristine nature reserves. Though Soviet scientists, writers, and students were neither dissidents nor civil society activists in the Western sense, they did manage to carve out a little 'archipelago of freedom' in a highly repressive state.

Similarly, in the early years of the Bolshevik regime forestry specialists continued progressive and scientific practices that had emerged in the later decades of Tsarist Russia. Although by the 1930s the Communist Party had subordinated these specialists to the political goals of the Soviet regime, through the 1920s Russian foresters resisted the peasant outlook on nature as an inexhaustible resource from God (Bonhomme, 2005).

Some progress was made in preserving endangered wildlife through inclusion in the Soviet Red Book, two volumes detailing threatened flora and fauna, which were based on the Red List established by the International Union for Conservation of Nature and Natural Resources. With the publication of the first official Soviet Red Book in 1978 lists of endangered species were compiled by ecologists and protected by law. Official industrial policy favored the use of low-waste or waste-free technologies in production, but closed-cycle processes were seldom implemented in practice. There was a modest effort at recycling paper – just under a third of all paper used was recycled by the mid-1980s. One of the most ecologically benign aspects of the Soviet economy was the absence of modern packaging, which greatly reduced the amount of solid waste generated by consumers.

Legal and Organizational Issues

Faced with a growing litany of environmental problems, the Soviet central government enacted laws on rational land use (1968), water quality (1970), mineral use (1975), forest resources (1977), air quality (1981), and protection of animal species (1981). In each case the legislation was vague and declaratory, and legal mechanisms for enforcement ineffective or nonexistent. All Soviet laws incorporated the principle of state ownership of natural resources, rejecting the concept of property rights as a motive to use resources more efficiently. While the goal of this legislation was generally the more rational use of natural resources, there were few, if any, economic incentives for conservation or economic penalties for polluting. Fees and other price mechanisms – for water usage, mineral exploitation rights (royalties), stumpage fees – were either nonexistent or inconsequential and did little to reduce massive waste and inefficiency. Polluters did not face the threat of legal action to force them to comply with environmental legislation, as did firms in the West. In any case Soviet legislation tended to be overly stringent and therefore not readily enforceable. It was more symbolic than substantial, designed to illustrate publicly the regime's commitment to environmental protection without jeopardizing production goals (Ziegler, 1987).

Environmentalists and foreign observers frequently cited bureaucratic obstructionism as a major obstacle to effective Soviet environmental policy. The incentive structure was such that each ministry and department pursued its own narrow interests, often at the expense of the common good. Moreover, there was no single agency tasked with environmental protection. In 1978, a joint resolution of the Communist Party Central Committee and the Council of Ministers restructured the State Hydrometeorological Service into a State Committee on Hydrometeorology and the Environment, with expanded powers in environmental protection, particularly in monitoring air pollution. However, this new agency had neither the budget nor the administrative authority to control pollution generated by the more powerful defense and heavy industries. Nor did the USSR enact a law mandating environmental impact studies of new development projects. Bureaucratic responsibility for environmental protection remained divided among a large number of ministries and agencies. Capital investments for air and water pollution abatement increased substantially through the 1970s, but these larger allocations seem to have had a negligible effect since modern pollution abatement equipment was often in short supply.

One of the major problems in developing an effective environmental policy was the practice of tasking the producing ministries in agriculture and industry with self-policing responsibilities in environmental protection. A problematic arrangement under the best of circumstances, the situation was made worse in the USSR by the clear bias of the planned economy toward production. Officials and managers could lose their jobs or forfeit bonuses if they failed to fulfill the plan, while the modest fines for pollution violations could easily be covered out of the plant's general fund.

The authoritarian nature of the Soviet political system contributed to its overall environmental inefficiency. Strict censorship of the media and classification of pollution data guaranteed that the public, and much of the elite, remain ignorant of the full extent of environmental degradation. The lack of a civil society led to a deep divide between the goals of the state and those of Soviet citizens (Josephson, 2013). Soviet newspapers in the Leonid Brezhnev (1906–82) era frequently carried stories of pollution violations, waste, and inefficiency, or letters from citizens concerned about threats to nature, but such reportage was carefully monitored, and often solicited, by the party and government. Critical coverage was permitted if it served state interests or if it was useful in discrediting officials who had fallen out of favor, but the full extent of environmental degradation was not made public until late in the Gorbachev era.

Environmental Activism

Autonomous ecology groups were not permitted to organize until Mikhail Gorbachev's liberalizing reforms; only mobilized

and structured forms of participation were allowed under Soviet rule. For example, nature protection societies with limited authority operated in each of the 15 republics. The All-Russian Nature Protection Society organized beautification campaigns, sponsored radio and television broadcasts, and helped organize educational programs and propaganda. Supposedly the society was empowered to comment on draft legislation, conduct factory inspections, and provide input on siting new facilities. Another mass organization with some input into the implementation side of environmental policy was the People's Control Committee, whose members could, under party and government supervision, investigate environmental violations. As with press coverage of environmental issues, these organizations only functioned with the guidance of party officials, and had no independent authority.

Formal representative institutions in the Soviet Union – the Supreme Soviet at the national level, and the union republic, region, area, district, and local soviets – were tightly controlled by the Communist Party, as were elections to these offices. The Supreme Soviet's Conservation Commission could, with party approval, criticize bureaucrats for not following party directives on environmental protection, but it was far from a representative body. Union republic governments could and did create their own environmental or nature protection agencies, but in the absence of legal or budgetary authority these structures could not effect significant change in policy.

Environmental concerns were articulated more effectively by scientists, whose professional expertise and positions gave their opinions greater authority. The most active advocates for the environment included geographers, economists, biologists, and specialists within the departments and agencies tasked with environmental protection. Scientific expertise was highly valued in the Brezhnev period, probably as part of the elite's reaction to Khrushchev's impulsive and ill-considered experiments with agriculture, and his uncritical support for the quack geneticist Trofim Lysenko (1898–1976). Environmental activism was tolerated if the principals were specialists who could speak with authority in their respective fields.

In the final two decades of the Soviet Union, specialists were able to exert modest influence on environmental policies as long as their suggestions and criticism did not challenge party leadership or the system's shibboleths, such as socialist ownership and industrial growth. The environmental efforts of Soviet scientists were augmented by support from another influential group, writers and other cultural icons. Soviet novelists such as Mikhail Sholokov (1905–84) had the stature to advocate passionately for preserving Lake Baikal. A similar coalition of scientists and writers – most prominently, the Russian writer Valentin Rasputin (1937–) and poet Evgeny Evtushenko (1933–) – played a vital role in defeating a proposal to divert Siberia's great rivers southward, to replenish Central Asian waters depleted through irrigation. This monumental project, which was shelved in 1986 after 13 years of planning and debate, ignited tensions between Central Asian advocates of diversion and Russian nationalists determined to protect Siberia's water resources.

Soviet environmental policies also suffered from the cult of secrecy surrounding the huge military establishment. Environmental problems linked to defense industries, Soviet military operations, and the nuclear industry were generally off limits to discussion, even for the most prominent scientists and writers. Defense expenditures in the late Soviet period were estimated to consume between 25 and 30% of gross domestic product, and defense-related production dominated much of the most heavily polluting industries. A dissident ecologist writing under the pseudonym of Boris Komarov (Zeev Wolfson, 1944–) sent a manuscript to the West in 1978, in which he detailed a litany of environmental disasters. A constant theme throughout the book is the Soviet government's secretive approach to ecology issues and the practice of invoking national security as an excuse for covering up the scope of pollution.

Glasnost and the Environment

The Chernobyl nuclear accident in April 1986 gave a major impetus to Mikhail Gorbachev's announced policy of greater openness, or glasnost. The Kremlin initially sought to minimize the extent of the disaster but reversed its secretive approach when European nations measured and publicized radiation levels drifting in their direction and pressured Moscow to be more forthcoming. Chernobyl heightened public awareness of the dangers to health from various forms of pollution, and in the relaxed political atmosphere of the late 1980s a large number of informal ecology groups were organized. Women concerned with the health effects on their children frequently held leadership positions in these new organizations. These late-Soviet era ecology groups were generally local in their focus, although a few umbrella organizations – most notably the Socio-Ecological Union – tried to coordinate environmental activities at the national level.

Environmental activism peaked toward the end of the Soviet period as Chernobyl and revelations about the extent of environmental degradation exposed the government's incompetence and disregard for the people's well-being. Environmental concerns in the ethnic union republics meshed with nationalism, as the Lithuanians, Latvians, Estonians, Ukrainians, Armenians, and others accused Moscow of ecological colonialism. Kazakhstan's population bitterly resented the hundreds of nuclear weapons tests that had been carried out over four decades at Semipalatinsk in eastern Kazakhstan. The Nevada-Semipalatinsk movement organized demonstrations and protests across Kazakhstan calling for an end to nuclear weapons' testing, which reportedly contributed to high rates of cancer, skin disorders, miscarriages, genetic mutation, and depression. Led by the respected Kazakh poet Olzhas Suleimenov (1936–), the movement quickly collected over 2 million signatures in favor of closing the test site. Nevada-Semipalatinsk lobbied Alma Ata and Moscow and in 1990 secured an indefinite moratorium on nuclear testing.

Beyond the immediate impact of such environmental victories, revelations of callous indifference to human life and well-being dramatically affected public opinion. The Communist Party's assertions, that Soviet socialism was morally superior to capitalism and that national minorities benefited by living under Russian tutelage, were shown to be lies. Soviet central government officials and party leaders who were responsible for locating heavily polluting industries, unsafe nuclear power plants, and nuclear testing ranges in

national homelands lost what little legitimacy remained. This erosion of public confidence accelerated the process of political disintegration.

Popular concerns with the environment found expression in the new political institutions of the reform era. In the March 1989 elections to the Congress of People's Deputies many candidates campaigned on environmental issues, and environmental complaints were voiced by numerous deputies during the opening session of the Congress in May. The smaller working body elected by the Congress, the new Supreme Soviet, formed a large and active Committee on Questions of Ecology and the Rational Use of Natural Resources, chaired by a deputy from the Karakalpak autonomous republic. The Supreme Soviet also passed an urgent resolution to promote ecological recovery, but as with previous legislation it was overly vague and largely declarative. In union republic elections in 1990 environmental issues proved equally important to voters and candidates. In the Soviet context political measures proved inadequate; a thorough overhaul of the Soviet administrative structure was needed before an effective environmental policy could be realized.

Gorbachev was clearly sympathetic to environmental concerns, as evidenced by his formation of Green Cross International following his resignation from the Soviet presidency. In January 1988 the Soviet leadership created a State Committee on Nature Protection (Goskompriroda), the first all-union government agency tasked exclusively with monitoring and regulating environmental protection. The approach to environmental protection was to be systematic rather than discretely technical, in keeping with more recent holistic approaches to environmental protection. Market-based incentives such as the polluter pays principle were to replace purely administrative measures. In addition to setting and monitoring pollution norms, Goskompriroda was empowered to conduct environmental impact studies and to develop a comprehensive program for cleaning up the environment and planning a more environmentally benign economic strategy. Although headed by committed individuals – Nikolai Vorontsov (1934–2000), a talented biologist and committed environmentalist, replaced Feodor Morgun (1924–) in summer 1989 – most of Goskompriroda's staff lacked training or experience in environmental protection. The new agency faced the monumental challenge of formulating and implementing policy in the increasingly chaotic political climate. In any event Goskompriroda simply did not have the clout to overcome embedded bureaucratic resistance.

Gorbachev's reform program shifted initiative away from Moscow, resulting in a groundswell of environmental regulations among previously disempowered republics, regions, and cities. Many local governments established permit fees and fines to raise revenues so they would not be dependent on the central government in Moscow. The result was a patchwork of fees and regulations imposed without clear reference to costs and benefits and implemented in the context of an acute economic crisis. Environmental protection in the final years of the Soviet Union reflected the general devolution of authority and chaos that characterized the newly independent post-communist states of the 1990s.

Although there were many factors contributing to the collapse of the Soviet Union, the long list of environmental problems and the failure of the Soviet state to enact and implement an effective environmental policy must be considered among the more important. The willingness of the regime to countenance the destruction of huge tracts of land, the poisoning of the air and water, and the waste of valuable natural resources, while covering up the effects of environmental degradation on the population, exposed the fundamental inhumanity of the communist system. Environmental activism laid the foundation for nationalist activism, which in turn fragmented the country into 15 separate states. In this sense an autopsy on the Soviet Union may indeed rule it death by ecocide (Feshbach and Friendly, 1992).

Soviet Legacies

Soviet legacies continued to shape environmental politics and policies for years following the collapse of the USSR. Optimism that democratization and the creation of a market economy would improve Russia's environmental record proved unfounded (Josephson, 2013). The environmental activism that flourished during Gorbachev's perestroika reforms waned in the 1990s as inflation and unemployment soared and simple survival took precedence over ecology. Russia's economic recovery after 2000 was paralleled by Vladimir Putin's (1952–) construction of an authoritarian regime that promoted the interests of the newly rich oligarchs, who had little regard for the natural environment, and favored centralized state solutions to social, economic, and ecological problems.

In this context environmental organizations functioned but did not flourish. Private sources of funding for environmental organizations were limited, and the transnational funding that had supported ecological activism in the 1990s became politically suspect in the 2000s. Greater central control and state cooptation limited civil society organizations' freedom of maneuver. Major media outlets were either under state control or exercised self-censorship – few reported regularly on ecological problems, and few Russian citizens were overly concerned about ecology issues. Environmental activists who challenged the power structure were either ignored or intimidated; antinuclear activists were charged with treason and imprisoned for publicizing 'state secrets' when they revealed instances of radioactive pollution by the Russian military. Activists opposed to the destruction of the Khimki forest outside Moscow to build a highway to St. Petersburg were harassed and severely beaten by thugs (Evans, 2012).

Much as in Soviet times, impressive environmental laws were frequently not enforced, and projects favored by powerful state and private interests were implemented with scant regard for environmental consequences. Russia's domestic and international resurgence in the 2000s was due largely to exploitation of natural resources, primarily oil and gas. As an energy superpower Russia supported national champions such as Gazprom and Rosneft, and these corporations in turn promoted Putin's goal of a strong corporatist state, generally with minimal attention to the environmental consequences of hydrocarbon production. Most of the new oil and gas fields were located in remote parts of Siberia, and their development often came at the expense of the marginalized indigenous peoples (Agyeman and Ogneva–Himmelberger, 2009).

International cooperation on environmental issues was greatly overshadowed by Russia's role as an oil and gas exporter. Russia signed the Kyoto Protocol on climate change (in 2004), but less from a genuine environmental commitment or domestic pressure than from international considerations. Moscow anticipated economic benefits from its proposed membership in the World Trade Organization, which was from the Kremlin's perspective linked to Kyoto, and was concerned with its international image. In late 2012 following the Doha meeting, Russia's government declared Kyoto had lost its effectiveness and declined to sign on to the second commitment period of the agreement.

See also: Cities Under Postsocialism; Economic Transformation: From Central Planning to Market Economy; Environment and Development; Environmental History; Environmental Movements; Environmental Sciences; Planning During Post-Socialism.

Bibliography

Agyeman, J., Ogneva-Himmelberger, Y. (Eds.), 2009. Environmental Justice and Sustainability in the Former Soviet Union. MIT Press, Cambridge, MA.

Bonhomme, B., 2005. Forests, Peasants, and Revolutionaries: Forest Conservation and Organization in Russia. East European Monographs, Boulder.

Evans, A.B., 2012. Protests and civil society in Russia: the struggle for the Khimki forest. Communist & Post-Communist Studies 45, 233–242.

Feshbach, M., Friendly, A., 1992. Ecocide in the USSR. Health and Nature under Siege. BasicBooks, New York.

Henry, L.A., 2010. Red to Green: Environmental Activism in Post-Soviet Russia. Cornell University Press, Ithaca, NY.

Josephson, P., et al., 2013. An Environmental History of Russia. Cambridge University Press, Cambridge, UK.

Komarov, B., 1978. Unichtozhenie Prirody. (Frankfurt, Possev, Trans.) As the Destruction of Nature in the Soviet Union. M.E. Sharpe, White Plains, NY.

Oldfield, J.D., 2005. Russian Nature: Exploring the Environmental Consequences of Societal Change. Ashgate, Aldershot, UK.

Peterson, D.J., 1993. Troubled Lands. The Legacy of Soviet Environmental Destruction. Westview Press, Boulder, CO.

Pryde, P.R., 1991. Environmental Management in the Soviet Union. Cambridge University Press, Cambridge, UK.

Stewart, J.M. (Ed.), 1992. The Soviet Environment: Problems, Policies, and Politics. Cambridge University Press, Cambridge, UK.

Weiner, D.R., 1988. Models of Nature. Ecology, Conservation, and Cultural Revolution in Soviet Russia. Indiana University Press, Bloomington, IN.

Weiner, D.R., 1999. A Little Corner of Freedom: Russian Nature Protection from Stalin to Gorbachev. University of California Press, Berkeley, CA.

Ziegler, C.E., 1987. Environmental Policy in the USSR. University of Massachusetts Press, Amherst, MA.

Relevant Websites

http://baikalwave.blogspot.com/ – Baikal Wave.
http://www.bellona.org/ – Bellona-Russian Legal Issues.
http://www.wild-russia.org/html/crnc.htm – Center for Russian Nature Conservation.
http://ecodefense.ru/ – Eco Defense (in Russian).
http://www.greenpeace.org/russia/en/ – Greenpeace Russia.
http://www.greenparty.ru/main_en.php – Russian Ecology Party (Greens).
http://www.seu.ru/index.en.htm – Socio-Ecological Union.
http://www.wwf.ru/eng – World Wildlife Fund-Russia.

Russian and Post-Soviet Studies: Gender

Lynne Attwood, School of Arts, Languages and Cultures, University of Manchester, Manchester, UK; and Derbyshire, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the changes in the Russian understanding of male and female roles from the time of the socialist revolution in 1917 to the present day. It outlines the original revolutionary plans for altering the relationship between men and women to ensure gender equality, and then shows how these were eroded during the course of Soviet history with 'equality' periodically redefined not in accordance with the needs of Soviet men and women, but of the Soviet state. The article concludes by exploring the impact that the transition to capitalism has had on gender relations.

Introduction

Like so much else in the new capitalist Russia, gender issues have undergone enormous change in the past two decades since the collapse of the Soviet Union. While the supposed commitment to gender equality in the Soviet era does not stand up to much scrutiny, the predominant image of men and women, as paraded in the visual arts and the media of those times, was of comrades who shared the same hopes and aspirations for the communist future, who worked alongside each other in factories and farms, and, indeed, who fought alongside each other on the battlefield when this was required. While the understanding of equality underwent a number of changes during the course of Soviet history, it was always portrayed as an essential part of socialism and a defining feature of Soviet ideology and practice. The rapid transition to capitalism, however, did away with the need even to pay lip-service to gender equality. It ushered in an essentialist approach to gender difference reminiscent of that which prevailed in the West in the wake of the Second World War, when 'Rosie the Riveter' was encouraged to put down her factory tools and tie on her apron.

This article explores the changes in the Russian understanding of male and female roles from the time of the Revolution in 1917, throughout the 70 years of the Soviet era, and up to the present day. It begins by establishing the meaning of the term 'gender,' then outlines the original revolutionary plans for altering the relationship between men and women to ensure their equality, and reviews the ways in which equality was periodically redefined, not in accordance with the needs of Soviet men and women, but with those of the state. Finally, the article discusses the impact of transition to capitalism on gender relations.

Description of Terms

'Gender' refers to expected practices and behavior associated with either men or women that are not rooted in biology but are determined by culture. Western feminists have been careful to distinguish between the terms 'gender' and 'sex,' the latter referring to the biological differences between men and women. However, Soviet writers and thinkers only acknowledged this distinction shortly before the disintegration of the Soviet Union, when, in 1990, the Center for Gender Studies was founded in Moscow. The Soviet authorities always insisted, throughout the 70 years of the Soviet Union's existence, that they were committed to the establishment of complete equality for men and women. However, no attempt was made to analyze or understand the extent to which differences between men and women were the result of biology or culture. Alexandra Kollontai, the country's principal theoretician on the 'woman question' at the time of the Revolution, did come close to addressing the concept of gender (although without actually using the term), but dismissed it as an irrelevancy – indeed, as a bourgeois preoccupation. What was important, she argued, was that all people, male and female, had a real opportunity to develop their personal potential. In her words:

> Leaving it to the bourgeois scholars to absorb themselves in discussion of the question of the superiority of one sex over the other, or the weighing of brains and the comparing of the psychological structure of men and women, the followers of historical materialism fully accept the natural specificities of each sex and demand only that each person, whether man or woman, has a real opportunity for the fullest and freest self-determination, and the widest scope for the development and application of all natural inclinations.
>
> Kollontai, 1977, p. 581

Yet, despite their refusal to address the issue of gender openly, Soviet politicians were in reality very much concerned with it, and made determined attempts at different times either to eradicate or strengthen the cultural differences between men and women.

The Soviet Era

In its first decade of Soviet power, the policy of the Communist Party was to challenge traditional understandings of what were appropriate behavior and social roles of male and female. At a time when the West was promoting and preserving separate male and female spheres, Soviet propagandists were urging women to leave that female bastion, the home, and enter the public world of work and politics, until then a largely male

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10141-2

environment. Women were assured that state institutions would take over their domestic tasks. In any case, the family would become less important as people got used to living communally. Some theorists, most notably Kollontai, went so far as to argue that the family would serve no purpose at all in a communist society and in due course would simply 'wither away.'

Kollontai championed the emergence of a 'new woman' under socialism, who would embody the following qualities:

> Self-discipline instead of emotionality; the ability to value her freedom and independence, instead of being submissive and lacking in personality; the affirmation of her individuality, instead of a naïve attempt to absorb and reflect someone else's cast of mind, that of the "beloved" … no mere wife, no shadow of a man, [but] an individual, a woman who is above all a person.
>
> Kollontai, 1919, p. 29

Yet, there was no attempt to rethink the gender roles of both male and female. Women were to be incorporated into what was previously seen as the male sphere of public service, but there was no question of men taking on what were traditionally seen as women's responsibilities. In other words, women could enter the workforce, but they still had to do the housework when they came home. Articles on the organization of communes made it clear that women would still be responsible for the cooking, laundry, child care, and so on, but they would now take on this work collectively, for the commune as a whole, rather than individually, for their own families. There was also no question of them choosing not to have children. Rearing the next generation of workers was a patriotic duty, not a choice.

Widespread unemployment in the 1920s meant that few women were actually drawn into social production. This began to change in the late 1920s, when Stalin came to power. In 1929 he launched a huge industrialization drive. All workers were now needed, women as well as men, and by 1932 the number of women workers was almost double what it had been in 1928. At the same time, the pledge to free women from their traditional domestic duties was quietly abandoned. All resources were now required for industry, and if women were still willing to perform those traditional functions for free, the state would hardly fight for the right to fund them. Predictions about the demise of the family also ceased. On the contrary, the family was now held up as the 'basic cell' of Soviet society.

There are a number of reasons for this change in approach. There had already been concern about population decline, resulting largely from industrialization and modernization, before the introduction of the Plan, but the Soviet regime had insisted that the benefits introduced by socialism would encourage people to have more children. The demographic nose-dive was, then, perceived as a major crisis. It was hoped that rehabilitating the family might help reverse it. The familiarity of traditional family life might also provide some stability in people's lives in the face of such radical change. Furthermore, the family now took on a symbolic function; Soviet society as a whole was depicted as one large family, with Stalin the ultimate patriarch. Far from being an outdated institution that had no place in a communist society, the family had assumed a new significance.

Women now found themselves laboring under a 'double burden.' They were required to go out to work: there were simply not enough men to bring about industrialization at the speed or to the extent required by Stalin. However, they also had to perform their traditional domestic and nurturing functions. They were to do this with no assistance from men, nor from any services or products that could help reduce the domestic workload – given the concentration on heavy rather than consumer industry. Yet, this combination of functions was ironically hailed as evidence of women's equality, and the basis of their self-fulfillment. The 'woman question' was declared as closed. This was certainly true to some extent, since women did not have quite the same pressures as men to commit enormous amounts of time to the Party rituals. Women were much less likely to be Party members or to aim for a career in politics, for which active participation, for example, in Komsomol rituals was an essential step.

When Khrushchev came to power and launched his 'destalinization' program, the 'woman question' was partially reopened. The press on the whole continued to applaud the country's successful establishment of women's equality, but some authors now acknowledged that the real situation was more complex, and that women fell behind men in both professional and political hierarchies. At the same time, sociologists began conducting 'time budget studies' which made it clear that women devoted an enormous amount of time to domestic work, and suggested that this was a significant factor in their inferior standing in public life. There were now some suggestions, albeit tentative, that husbands might be persuaded to ease their wives' burden by 'helping' them in the home.

From the l970s, with Brezhnev in charge, there was another shift in policy. The notion that women's sense of self-fulfillment rested on their combining the traditionally female roles in the family with traditionally male roles in politics and social production began to be questioned, largely because of the seemingly intractable demographic implications: the birthrate, at least in the European republics of the Soviet Union, had remained stubbornly low. At the same time, women workers were essential in the Soviet Union's labor-intensive economy, and it was simply not possible to remove them from the workforce. Instead, there was an attempt to alter the balance between work and family in their lives and to persuade them to see the family as paramount and work as secondary. This meant that they would continue to work, but would devote less time and energy to improving their qualifications and climbing the professional ladder. Instead, they would see raising a family as their main source of satisfaction. Accordingly, a media campaign and a new school course were launched to raise the status of the family, persuade women to have more children, and encourage boys and girls to develop more traditional gender-specific traits. Not having children had never been a choice for Soviet women, but the attempted manipulation of their reproductive function now reached a new height.

Gender, although it was still not named as such, had become a major issue. Theorists now insisted that "the study of psychological differences between members of the male and female sex is an important scientific and practical question, no less important than the continuing study of their biological and physical differences" (Khripkova and Kolesov,

1981, p. 74). In the West there was also an interest in studying 'psychological differences between members of the male and female sex,' but in order to understand how those differences came about, the extent to which they challenged equality, and what could be done to overcome them. In the Soviet Union, the primary concern was to reinforce them.

When Gorbachev came to power in 1985, his program of economic reform required the 'rationalization' of the labor-intensive Soviet economy. This included the introduction of *khozraschet*, or 'self-accounting.' What this meant was that unproductive enterprises would no longer be bailed out by the government but would have to ensure their own profitability. The policy of full-employment had resulted in overstaffing enterprises, and this surplus labor had to be shed.

Inevitably, the least productive and/or the least cost-effective members of the workforce would be the first to go, and these were generally women: not only did they have to take time off work to have children, but they were also the ones who stayed at home when the children were sick. Many family benefits were paid to women through the workplace, which meant they were more expensive to employ than men (Bridger et al., 1996, p. 5). They were, as the economist Zoya Khotkina put it, 'excessively burdened by social privileges' (Khotkina, 1992, interview with the author).

Removing women from the workforce would also help disguise the widespread unemployment that was now seen as a serious likelihood. In a country that had consistently proclaimed work to be both a human right and a duty, unemployment was likely to have a considerable psychological effect on the population – and could also result in negative repercussions for the politicians who were seen as responsible for bringing it about. This could be mitigated by convincing women with small children that for them, not working outside the home did not constitute unemployment. It might also encourage them to have more children. Accordingly, staying at home was presented to women as a positive development: at last they had the 'luxury' of spending more time with their children. Sociologist Igor Bestuzhev-Lada, one of the 'experts' who regularly appeared in the press endorsing the state's policies, insisted that women only wanted to go out to work because of decades of misguided upbringing, and that given the choice, most would stay at home. They could not be considered unemployed because "the workplace for a woman is with her children. She is not without work – children are her work" (Bestuzhev-Lada, 1988, p. 21). Bestuzhev-Lada held that women should receive a state grant for this 'maternal work' that would be sufficient to ensure their financial independence. However, after the Soviet Union's collapse this clearly sounded too much like socialism, and, as we shall see, he went on to advocate self-help rather than government handouts (Levina, 1992, pp. 10, 11).

Yet, the new openness that began in the Gorbachev era meant that criticism of government policy had become possible, and articles defending women's right to work began to appear in the press. This extended even to the Communist Party's own theoretical journal, *Kommunist*. In a groundbreaking article appearing in 1989, three female scholars from the Institute of Socio-Economic Problems challenged the policy of sacrificing female workers in order to streamline the economy and improve its efficiency. This, they argued, was the motive behind the introduction of several supposed 'benefits' for women:

> In order to facilitate this female "exodus" from social production, it has been suggested that they be granted a series of privileges which will result in the reduction of their workload: the work day will be shortened, the period of paid leave to look after a new-born baby lengthened, and so on.
>
> Zakharova et al., 1989, p. 58

Within two years of this article appearing its authors had founded the Center for Gender Studies in Moscow, which, as the first feminist organization in the Soviet Union, attracted considerable attention from both Western scholars and feminist activists. It set about publishing material previously unimaginable in the Soviet Union, mounting university courses, and, with the help of grants from sympathetic Western organizations, hosting international conferences on gender issues in Soviet society.

The Post-Soviet Situation

The Soviet Union disintegrated in December 1991. In post-Soviet Russia, the tentative moves toward capitalism, which had begun in the late Gorbachev era, now proceeded at a gallop under the leadership of Boris Yeltsin. The new market economy was portrayed as the antithesis of socialism, not only in literal but also in symbolic terms, and gender was a key feature of this symbolization. Socialism, with its extensive welfare provision, was portrayed as an overprotective mother, who had prevented the development of personal initiative and entrepreneurship. The market, in contrast, emphasized the go-getting qualities traditionally associated with men, such as aggression, competition, and a willingness to take risks. The old Soviet image of men and women as partners in the building of socialism was now replaced by a much more traditional view of male and female roles: the man would go out to work and provide financial support for his family, while the woman would stay at home and provide emotional sustenance.

The post-Soviet woman was not only depicted as mother and housewife, but was also cast as sexual plaything. Pornography was ubiquitous in the new Russia. It not only spawned a mass of new men's magazines, but also invaded the pages of the mainstream press. At the same time, prostitution was presented as an attractive and lucrative career choice for women. There was also a clear expectation that women employed in the burgeoning private sector would accept the provision of sexual favors to clients as part of their job description.

The sex industry was one of the few growth areas for women in the new market economy. Unemployment was not as widespread as had been feared, partly because of the injection of government credits into the ailing and problematic mechanisms of market exit that would have created unemployment in the first place (Bridger et al., 1996, p. 40). However, there was much 'hidden' unemployment. Many people working in state enterprises were forced to work part time, or accept lengthy unpaid periods of leave; and even those who remained

in full-time employment often had to put up with long delays in the payment of wages.

Statistics are unreliable, but in the early post-Soviet period women were said to constitute more than 70% of the newly unemployed (Bridger et al., 1996, p. 51; Zoya Khotkina, 1992, interview with author). They were encouraged to see self-employment as the solution to unemployment. Bestuzhev-Lada, who had previously suggested that women be provided with grants to enable them to give up work and devote themselves to their children, now embraced the market and insisted that women should contribute to their own support, by, for example, selling home-made pies or craft items (Levina, 1992, pp. 10, 11). A similar assumption, that supposedly innate female skills and aptitudes present the most appropriate possibilities for women's self-employment, was expressed by a host of other commentators, despite the fact that most women caught up in the first wave of unemployment were engineers, technical specialists, and other professionals who were not doing traditional 'female' jobs (Bridger et al., 1996, p. 122).

While women are still not seen as employees of equal worth, they have compensated for this bias by showing greater flexibility than men at the work they are willing to do and the wages they are willing to accept. As Sarah Ashwin points out, "however biased they may be against women, employers offering low wages know that they are only likely to get female applicants" (Ashwin, 2006, p. 214). Accordingly, the fear that "unemployment [would have] a female face" has not been realized (Ashwin, 2006, p. 2).

At the same time, the resurgence of essentialist ideas about male and female roles has meant that there is a tendency to see being the breadwinner as a defining feature of being male. Indeed, some Russian commentators insist that women should not have to make any contribution at all to the family's material support (Kay, 2006, p. 166). Yet, this is only possible for the families of men who have risen to the top of the new capitalist hierarchy. Most men are in less secure employment than they were in the Soviet era, and families also have more financial obligations because the services that were provided free in the Soviet Union now have to be paid for. In her study of men in post-Soviet Russia, Rebecca Kay found that the majority embraced traditional ideas about male roles and duties and suffered considerable psychological discomfort if they were unable to fill them. There is a high level of male violence against women, other men, and themselves; around 14 000 Russian women are reportedly killed by their husbands or boyfriends every year, while 100 000 men die from suicide or murder (Kay, 2006, p. 8). Male life expectancy has also reached an astonishing low; while women live an average of 72 years, men barely reach 59 years (Ashwin, 2006, p. 1).

Male alcohol consumption, which is reportedly 10 times higher than that of women (Ashwin, 2006, p. 3), is one of the main factors behind these figures. A link has also been found between unemployment and alcohol abuse on the part of men, but not women. This might be explained by the difference in male and female domestic roles. Although women's extra family duties resulted in a 'double burden' in Soviet times and have made them less attractive employees in post-Soviet Russia, the family also provides them with a sense of meaning and purpose outside of the workplace. As Ashwin notes, this has made it "easier for them to cope with the psychological effects of unemployment or downward mobility," while men, who see their primary domestic function as providing for their families, are less likely to derive this compensatory satisfaction from family life if they lose their jobs (Ashwin, 2006, p. 216).

Yet, the sharp decline in the birthrate following the collapse of the Soviet Union would suggest that the family is not as important for women as it once was. While the low birthrate was a cause for concern throughout Soviet history, it is now seen as catastrophic – and, combined with a severe drop in the life expectancy of Russian men, has resulted in a considerable decrease in the Russian population. This has been partly offset by inward migration, mainly of ethnic Russians from the other republics of the former Soviet Union, but as Timothy Heleniak has pointed out, "Net migration…has compensated for less than half of the natural decrease" (Heleniak, 2002). According to demographer Anatoly Vishnevsky, the population shrank by 12.3 million between 1992 and 2008, and is predicted to shrink by more than another 11 million in the 19-year period between 2008 and 2025 (Vishnevsky, 2009, pp. 18, 20). Putin has described population decline as the biggest problem currently confronting Russia (BBC News, 19 May 2006).

One reason for the drop in the birthrate is that people now have more choice as to how to live their lives, including whether or not to have children. In the Soviet era the availability of contraception was severely controlled (although abortion, legal from 1955, was famously used as a form of birth control); and with the state in charge of the distribution of almost all housing, there was no guarantee that the unmarried and childless would be assigned apartments. Now contraception is readily available, and the housing market has been privatized. Accordingly, people have the means to opt out of state-promoted family life. Women's magazine *Domashnii ochag* (*Family Hearth*) explained in 2007 that a range of relationships and lifestyles is now possible, and for a variety of reasons some women prefer not to have children: they might have a strong career orientation, feel their material conditions are inadequate, be concerned about the fragility of marriage and not wish to become single mothers, or wish to spend their free time as they choose (*Sem'ya segodya*, 2007, pp. 151–156).

Some women are simply not interested in having children, as indicated by the arrival of the 'Childfree' movement in Russia. This began in the 1970s in North America to support people who chose to remain childless. In Russia it is primarily an Internet community; there is no sure way of knowing how many 'members' it has, but according to one estimate there are several thousands (Russia Today, 2011).

The Russian government is clearly concerned about the Childfree movement's potential contribution to the demographic crisis. In June 2011 a Demographic Summit was held in Moscow under the patronage of Svetlana Medvedeva, wife of the then-president Dmitrii Medvedev. Organized by the Illinois-based World Congress of Families, its publicity material expressed opposition to the Childfree movement, as well as to sexual minorities and 'radical feminists.' One of its participants, the American social scientist Steven Mosher, who is well known for rejecting overpopulation and the 'anti-natal' lobby, gave a talk on "How to create a baby-boom in Russia." His suggestions were a combination of practical measures and – reminiscent of

the old Soviet approach – propaganda aimed at schoolchildren. His three principal proposals were for the Russian Constitution to be amended so that it protected life from the moment of conception (hence, challenging women's right to abortion); for school textbooks to be carefully checked to eliminate any anti-natal sentiments and, instead, to promote the idea that "human beings are [a country's] most valuable resource," and for young couples with children to be 'sheltered' – or, in the case of families with three or more children – to be completely exempted from taxation (Mosher, 2011).

A number of financial benefits have already been introduced to encourage women to have more children. Most notably, in January 2007 Putin introduced 'maternity capital,' a one-off payment of 250 000 rubles (around € 7200), to be increased annually in line with inflation. This is paid to the mother on the birth or adoption of a second child, or a third or subsequent child if she already has more than one child and has not yet received this award. It can only be spent once the child reaches the age of three, and only for three designated purposes: the child's education, the family's housing, or to improve the mother's pension (Vishnevsky and Aakharov, 2009, pp. 32, 38, 50).

How have Russian women responded to these attempts to encourage a more traditional family orientation? As we have seen, women had very limited choice as to how they lived their lives throughout the Soviet era. Then, in the early post-Soviet period, there was an enormous outburst of feminist activity, much of it concerned with increasing this choice. A plethora of women's NGOs were set up, many with the help of funding from Western charitable organizations such as the Soros Foundation, the Ford Foundation, and USAID. However, this funding eventually dried up – as Janet Elise Johnson explains, after the terrorist attacks on New York in 9/11, "many donors, who had already grown weary of funding post-communist civil society, redirected funds towards new hotspots" (Johnson, 2009, p. 60). A large proportion of women's groups did not survive. The Center for Gender Studies that helped set up many of the now-defunct groups, however, does still exist and continues to engage in a range of academic and practical projects. It proclaims on its website that one of its aims is to raise awareness of gender issues among people directly engaged in social policy, such as members of parliament, journalists, teachers, and youth workers.

A more flamboyant defense of free choice has been staged by the self-proclaimed feminist punk band Pussy Riot, who became internationally famous after three of its members were arrested while performing a 'Punk Prayer' in Moscow's Cathedral of Christ the Savior in February 2012. One was released on appeal, but two remain in prison. Pussy Riot's 'prayer' called on the Virgin Mary to become a feminist and drive Putin out, and listed the crimes committed against the Russian people by the authorities, aided and abetted by the Orthodox church. These included homophobia ("Gay pride sent to Siberia in chains") and the promotion of traditional female roles ("So as not to offend His Holiness, women must give birth and love"). The appearance of the band members is itself a challenge to gender norms: their garish, ill-fitting clothes and balaclavas defy both traditional notions of female beauty and of the sexualized images of women which still abound in the new Russia. Pussy Riot have not eschewed family life, but also do not want it to be imposed on women. In fact, both of the imprisoned women are mothers of small children. One appealed for her sentence to be deferred until her five-year-old son reaches the age of 14 on the grounds that he needs his mother; despite the official concern about family life in Russia, the appeal was refused (*The Guardian*, 2013).

Concern about traditional family life is also a factor in the homophobic attitudes of the Russian authorities. In the Soviet Union, homosexuality clearly could not be accommodated within the country's demand for strict adherence to an officially sanctioned lifestyle. Male homosexuality was illegal and punishable by imprisonment, while lesbianism was treated as a mental illness. The more liberal atmosphere of the Gorbachev era resulted in the creation of a number of gay rights groups, many of which, like the women's NGOs, received financial support from Western organizations. Homosexuality was decriminalized by Yeltsin in 1993, although historian Dan Healey sees this as a response to pressure from the Council for Europe rather than from the gay lobbyist groups in Russia (Healey, 2004). Nonetheless, this was a major breakthrough for the movement, culminating in 1996 in the creation of a gay advice and cultural center in Moscow and the staging of a national conference attended by more than 125 delegates. However, as with the women's groups, these activities were largely dependent on Western funding and could not survive without it. Triangle, the major national lobbying organization, disappeared in 1997 when its Western funding ceased (Healey, 2004). For several years gay activism was almost nonexistent, until in 2005 Project Gay Russia was launched. It has carried out various campaigns to end discrimination and promote the rights of gays in Russia and, since 2009, in Belarus; these have included the right to freedom of assembly, and an end to the ban on gay men donating blood. Gay Russia also began organizing Gay Pride marches in Moscow in 2006. In 2006 the LGBT Network was founded, consisting of a range of regional gay support and lobbying groups.

These organizations face an uphill battle. While homophobia has always been an issue in Russia, official attitudes have been hardening. There have been repeated calls by members of the Russian parliament for homosexuality to be recriminalized and for lesbianism to be made illegal for the first time, although these have so far been rejected. Moscow Pride marches were banned from the outset by the then-mayor of Moscow, Yuri Luzhkov; and although they went ahead regardless, they have been beset by violent clashes with anti-gay protesters. In 2012 Moscow City Council banned Gay Pride marches for the next 100 years. That same year, St. Petersburg passed a law banning the 'propagandizing' of male homosexuality, lesbianism, bisexuality, and transgenderism to minors, a move which has since been adopted by other cities and regions throughout the Russian Federation. As journalist Masha Gessen explains, this outlaws anything which could be deemed to "harm […] the process of moral and spiritual development of minors, including forming among them the false perception that traditional and nontraditional relationships are socially equal." Herself a lesbian and a mother, she is concerned that for gay parents, "simply living with their children might be a breach" (Gessen, 2012).

While demographic concerns are clearly not the only reason for this resurgence in homophobia, they have played

a contributory role. As Healey explains, anti-gay legislation has been supported by religious leaders, Soviet-minded doctors, and other conservative forces who view the population decline as the gravest issue facing Russia, and blame a 'women's birth strike' and sexual perversion for 'the desecration of the national gene' (Healey, 2004). Putin himself, when asked for his view on the ban on the Moscow Gay Pride march at a press conference in 2007, explained that his attitude toward gay parades and sexual minorities was "connected with my official duties and the fact that one of the country's main problems is demographic" (Putin, 2007).

Conclusion

The 'new woman' championed by Aleksandra Kollontai turned out to be more myth than reality throughout Soviet history. She did not earn equal pay, did not have the same opportunities for professional advancement, was absent from the important decision-making political bodies, and had almost all of the domestic work to deal with. She was not an independent being but a servant of the state, serving it both in terms of production and reproduction. However, she did have a life outside the home, relative economic security, and (according to sociological studies) a sense of her own worth. Indeed, some commentators have argued that men suffered more than women in the Soviet Union, since the latter were able to use their family responsibilities as an excuse to withdraw as far as possible from the public realm, where state control was at its greatest (Funk and Mueller, 1993, p. 87). As the ultimate nanny state, the Soviet Union was also said to have deprived men of traditional masculine attributes; this led to concern that they had lost their sense of self-value and become 'feminized,' indolent, and alcoholic (Attwood, 1990).

The mad dash to the market in post-Soviet Russia turned this situation on its head. There was a revival of essentialist ideas about gender roles, and a celebration of traditional masculine characteristics that were thought to be particularly appropriate in a market economy. In the domestic realm, the role of 'breadwinner' was, in theory, restored to men. In reality, however, the new economic environment did not provide the conditions in which the majority of men could actually fulfill this role. This has contributed to a range of problems including alcohol abuse, violence, and a huge drop in male life expectancy.

While women have not been treated as equally valuable workers in the new Russia, their greater flexibility has ensured that they have retained an important role in the workplace. The family is still considered a primarily female sphere, resulting in a greater workload for women; but on the positive side, it provides them with psychological protection if they experience instability and deterioration in their work conditions.

People can make personal choices about how they live now that would have been unimaginable in Soviet days. However, some of these choices have not been welcomed by the authorities, and are, indeed, under threat. The most notable of these is the choice to reject family life in its traditional sense. With the demographic crisis seen as one of the key problems confronting Russia today, not having children is tantamount to an act of rebellion. Putin has claimed to "respect personal freedom in all its forms, in all its manifestations" (Putin, 2007), but it is hard to see how this is consistent with state pressure on childless women to procreate – and on the gay community to disappear.

See also: East Asian Studies: Gender; Gender History; Gender and Feminist Studies in Geography; Gender: Anthropological Aspects; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Near Middle East/North Africa Studies: Gender; Russian Revolution, The; Sexual Minorities in Modern Russia; Social Work in Russia.

Bibliography

Ashwin, S. (Ed.), 2006. Adapting to Russia's New Labour Market: Gender and Employment Behaviour. Routledge, Abingdon; New York.

Atkinson, D., Dallin, A., Lapidus, G.W. (Eds.), 1978. Women in Russia. Harvester, Hassocks, Sussex, UK.

Attwood, L., 1990. The New Soviet Man and Woman: Sex Role Socialization in the USSR. Macmillan, London.

Attwood, L., 1996. The post-Soviet woman in the move to the market: a return to domesticity and dependence? In: Marsh, R. (Ed.), Women in Russia and Ukraine. Cambridge University Press, Cambridge, UK, pp. 255–266.

Bestuzhev-Lada, I.V., 1988. Net detei – net i budushchego i naroda. Nedel'ya 15–21 August, 21.

Bridger, S., Kay, R., Pinnick, K., 1996. No More Heroines. Routledge, London.

Buckley, M., 1989. Women and Ideology in the Soviet Union. University of Michigan Press, Ann Arbor, MI.

Edmondson, L. (Ed.), 1992. Women and Society in Russia and the Soviet Union. Cambridge University Press, Cambridge, UK.

Farnsworth, B., Kollontai, A., 1980. Socialism, Feminism, and the Bolshevik Revolution. Stanford University Press, Stanford, CA.

Field, M.G., 2000. The health and demographic crisis in post-Soviet Russia: a two-phase development. In: Field, M.G., Twigg, J.L. (Eds.), Russia's Torn Safety Nets. Macmillan, Basingstoke and London, pp. 11–42.

Funk, N., Mueller, M. (Eds.), 1993. Gender Politics and Post-communism: Reflections from Eastern Europe and the Former Soviet Union. Routledge, London.

Goldman, W.Z., 1993. Women the State and Revolution: Soviet Family Policy and Social Life 1917–1936. Cambridge University Press, Cambridge, UK.

Healey, D., 2004. Russia, GLBTQ: An Encyclopedia of Gay, Lesbian, Bisexual, Transgender, and Queer Culture. www.glbtq.com/social-sciences/russia.htmlRussia (accessed 29.07.12.).

Heleniak, T., 2002. Migration dilemmas haunt post-Soviet Russia. In: Migration Information. http://www.migrationinformation.org/feature/display.cfm?ID=62.

Ilic, M., 1999. Women Workers in the Soviet Inter-War Economy. Macmillan, London, UK.

Johnson, J.E., 2009. Gender Violence in Russia: The Politics of Feminist Intervention. Indiana University Press, Bloomington, IN.

Kay, R., 2006. Men in Contemporary Russia: The Fallen Heroes of Post-Soviet Change? Ashgate Publishing, Burlington, VA.

Khripkova, A.G., Kolesov, D.V., 1981. Devochka-podrostok-devuska (Girl-Teenager-Young Woman). Prosveshchenie, Moscow.

Kollontai, A., 1919. Novaya Moral' i rabochii klass (The New Morality and the Working Class). Vserossiiskii Tsentral'nyi Ispolnitel'nyi Komitet Sovetov R., K. i K Deputatov, Moscow.

Kollontai, A., 1977. Selected Writings (Translated, and with commentary, by Alix Holt). Allison & Busby, London, UK.

Kon, I., Riordan, J. (Eds.), 1993. Sex and Russian Society. Indiana University Press, Bloomington, IN.

Lapidus, G.W., 1979. Women in Soviet Society: Equality, Development and Social Change. University of California Press, Berkeley, CA.

Levina, A., 1992. Yarmarka vakansii, ili 5 sposobov bor'by s zhenskoi bezrabotitsei (The fair of vacancies, or five ways of fighting women's unemployment). Interview with I.V. Bestuzhev-Lada Rabotnitsa 2, 10–11.

Liljestrom, M., Muntysaari, B., Rosenham, A. (Eds.), 1993. Gender Restructuring in Russian Studies. Tampere University Press, Tampere, Finland.

Mosher, S., 2011. Speech to the Moscow Demographic Summit. http://www.youtube.com/watch?v=CRWLkKZQ0AQ (accessed 11.03.13.).

Moskovskii tsentr genderrnykh issledovanii (Moscow Centre for Gender Studies), http://www.gender.ru/ (accessed 10.03.13.).

Pussy Riot Member Denied Appeal, January 17, 2013. The Guardian. http://www.guardian.co.uk/music/2013/jan/17/pussy-riot-member-prison-sentence (accessed 11.03.11.).

Putin, V.V., February 1, 2007. Press Conference for Russian and Foreign Media. http://www.usrccne.org/news2.phtml?m=305 (accessed 11.03.12.).

Russia Today, 2011. http://rt.com/news/principled-childless-gain-russia/ (accessed 20.01.13.).

Sem'ya segodya. Domashnii ochag, April 2007 151–156 (author not named).

Stites, R., 1978. The Women's Liberation Movement in Russia. Princeton University Press, Princeton, NJ.

Suchland, Jennifer, 2012. Contextualising Pussy Riot in Russia and beyond. e-International Relations. http://www.e-ir.info/2012/08/28/contextualizing-pussy-riot-in-russia-and-beyond (accessed 09.03.13.).

Vishnevsky, A.G., 2009. A new stage of Russian demographic development. In: Vishenevsky, A.G., Boblylev, N. (Eds.), Russia Facing Demographic Challenges, National Human Development Report Russian Federation 2008, Moscow. Also available at: www.undp.ru/documents/NHDR_2008_Eng.pdf (accessed 24.05.12.).

Vishnevsky, A.G., Aakharov, S.V., 2009. Growth of fertility: the start of a road with distant horizons. In: Vishenevsky, A.G., Boblylev, N. (Eds.), Russia Facing Demographic Challenges. National Human Development Report Russian Federation 2008. Prepress, Moscow.

Zakharova, N., Posadskaya, A., Rimashevskaya, N., 1989. Kak my reshaem zhenskii vopros (How we are resolving the woman question). Kommunist (Communist) 4, 56–65.

Russian and Post-Soviet Studies: Politics

Jeffrey Hass, University of Richmond, Richmond, VA, USA; and St. Petersburg State University, St. Petersburg, Russia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L. Shevtsova, volume 22, pp. 14733–14735, © 2001, Elsevier Ltd.

Abstract

This article examines politics – the sources, structure, and operation of power – in the Soviet Union and post-Soviet Russia. The Soviet political system was a grand experiment in an alternative political and social system of applying Enlightenment principles and socialist ideology. Yet, this system generated contradictions that would be its undoing: an economic centralization that bred stagnation, a social contract trading political loyalty for accountability, and institutionalization of ethnic identities. Boris Yeltsin's political and economic reforms unraveled much institutional structure, providing the impetus and opportunity for the rise of a new elite (oligarchs) and then the reimposition of the state power under the leadership of Vladimir Putin and a new elite, the *siloviki*. Yet, the new model of state-centered corporatism faces challenges and pressure from economic strains, possible political opposition, and growing tensions with the West.

Soviet Politics: A Brief Historical Overview

To make sense of contemporary Russian politics, we have to understand what came before: the Soviet era provided institutions, ideologies, and language that Russians have not entirely escaped. To an extent Vladimir Putin's state-centered political economy does appropriate some facets of the Soviet system; for example, mobilization of popular opinion via propaganda and political or youth organizations, the use of informal patronage, a central role for state elite in the polity and economy, and a distrust of autonomous social forces. These echo Soviet political institutions and practices: an elite seeking to create a version of Western modernity with unique differences, to hold power out of distrust for autonomous social forces, and seeking a place in the global order (cf Hosking, 1997 for a comprehensive overview, also Bunce, 1999).

A combination of the strains of World War I and social tensions borne of socioeconomic change and the growth of new classes led to revolutions in February and October 1917 (March and November by the western Gregorian calendar). The first brought the end of the tsarist regime, but the unstable situation of 'dual power' – political authority split contentiously between the Provisional Government and the Petrograd Soviet of Workers' and Soldiers' Deputies – did not hold, and in October 1917 the Bolshevik Party, led by Vladimir Lenin and Lev Trotsky, took power in a *coup d'état*. In January 1918 the Constituent Assembly met in Petrograd, but Lenin quickly disbanded that body. This led to a bloody civil war from 1918 until 1922, during which the Bolsheviks reconquered most territory of the former tsarist empire (except for Poland, the Baltic countries, and parts of Ukraine and Belarus that were reabsorbed in 1939) and set up institutions that would be of supreme importance, the Red Army and the secret police (originally the Chekha, later renamed the GPU, NKVD, and then KGB). In the aftermath of the social dislocation brought on by World War I and the Civil War, the new Soviet regime instituted the New Economic Policy (NEP), a broad set of economic policies that slowed down the desired pace of economic growth and allowed some autonomy to peasants and artisans. Lenin died in 1924 after a series of strokes, and a leadership struggle ensued that would take up much of the 1920s. Eventually, Joseph Stalin emerged victorious, in part because as General Secretary of the Communist Party he could appoint his supporters to important positions in the Party. By the end of the 1920s, Stalin was the supreme leader of the Party and of the state.

Part of the struggle for power involved articulating 'proper' policy according to political needs and the tenets of Marxism–Leninism, and two crucial policies would emerge from these struggles that would shape the fate of the USSR: industrialization and collectivization. In the late 1920s, while he was allied with Bukharin and superficially supported NEP, Stalin faced three challenges. First, industrial growth was sluggish, and an independent peasantry was slowly accumulating wealth at the seeming expense of the cities and Soviet state, from which the peasants could withhold their produce. Second, the rise of fascism in Italy and its possible appearance in Germany, and Chiang Kai-shek's 1927 massacre of Chinese communists in Shanghai, raised the specter of the fledgling socialist state being dragged into a war for which its economic foundations were not yet ready. Third, to solidify his hold on power, Stalin needed to distance himself from NEP. Thus, he unleashed two radical policies designed to expand the economy, provide an alternative to Bukharin's NEP, and hopefully prepare the country for future war: rapid industrialization and collectivization of agriculture. Industrialization was not without problems and inefficiencies, given the breakneck tempo of building factories and producing increasing quantities of output. Collectivization was a disaster. The idea behind this policy was to bring peasants into collective farms run by the state or peasant collectives, and resources would be pooled and output collected for distribution throughout the country. However, peasants did not always want to join collective or state farms (*kolkhoz* and *sovkhoz*), and this resistance was met by force: the Red Army was at times used to round up peasants or to surround villages that refused collectivization. The result was a decrease in agricultural output, and to feed the cities the regime squeezed food from the peasantry – resulting in peasant hardship and even famine in parts of Ukraine in the early 1930s.

With the consolidation of Party supremacy and a centralized, planned economy bereft of private property, Stalin and his leadership gained risks as well as privileges: any negative policy outcomes could reflect very badly on the leadership. Stalin's own paranoia, and the realization that his policies and his victories in the power struggle of the 1920s gave his opponents ammunition to remove him, helped encourage Stalin to unleash 'the Terror,' waves of arrests sanctioned by Stalin and carried out by the NKVD (People's Commissariat of Internal Affairs) against elites, average Party members, some average members of society, and eventually Red Army officers and members of the NKVD itself. By the outbreak of World War II in 1939, Stalin's personal rule and authority of the regime rested on a combination of coercion, propaganda, and promises of welfare in return for political loyalty and quiescence. World War II tested the mettle of the regime and population; ultimately, the Soviet Union prevailed against Nazi Germany, although at an enormous cost of lives and materiel. After World War II, the Soviet population expected a relaxation of coercive politics as a reward for sacrifices and victory. Instead, Stalin began a new wave of coercion that began with the Leningrad Affair (repression of Leningrad's wartime leaders) and ended only when Stalin died of a stroke in 1953.

For the next 3 years, Nikita Khrushchev and Georgii Malenkov courted other elites for support, with Khrushchev ultimately prevailing. A member of Stalin's Politburo who once had been head of the Ukrainian Communist Party and who had participated in the Terror, Khrushchev initiated 'de-Stalinization' with his 'Secret Speech' in 1956, in which he blamed Stalin for the tragedies of the previous decades but highlighted the sanctity of the Communist Party of the Soviet Union (CPSU). Khrushchev also set out to improve the Soviet economy so as to compete better with the United States in the new Cold War: in particular, Khrushchev decentralized economic decision making, but his reforms brought on institutional confusion and were abandoned once he was removed from leadership in 1964. Despite this seeming political defeat, Khrushchev would leave the important legacy of legitimating reforms, especially for a generation of younger communists that included future leader and reformist Mikhail Gorbachev.

When Leonid Brezhnev assumed the post of General Secretary, he distanced himself from Khrushchev's style of leadership by proclaiming 'trust in cadres,' an implicit social contract exchanging institutional stability and modest but constant improvements in standards of living, in return for political loyalty (cf Bunce, 1983; Millar, 1985). In practice, this meant weakened accountability of Party and state officials. Most Soviets welcomed such stability after decades of turbulence under Stalin and Khrushchev, and in the 1960s and into the 1970s it seemed that Brezhnev had a formula for success: the USSR continued to hold its own in completion with the United States (which was in the throes of social convulsions of Civil Rights and protests against the Vietnam War). However, into the 1970s stagnation began to take root due to the combination of a centralized socialist system without private property, and weak accountability central to Brezhnev's social contract. Fulfillment of the Five-Year Plan and its constituent monthly plan targets was the basis for reward, and because innovation naturally entails some risk of failure, Soviet industrial managers embraced conservative economic practices. Economic troubles increased after 1979: the Soviet invasion of Afghanistan led to increased oil output by Saudi Arabia, and the ensuing oil glut reduced the petrodollar income that the Soviet Union badly needed to purchase grain and industrial machinery from abroad. In the first half of the 1980s and after a decade of détente, Ronald Reagan increased American military expenditures, including investment in technological advances such as the 'star wars' missile defense system.

In the twilight of the Brezhnev era, Soviet institutions, practices, and political culture had crystallized, even if there were growing economic and social problems that would eventually trigger reforms and the collapse of the system but that would also provide templates for post-Soviet politics and political economy (Hass, 2011). While these principles reflected elite struggles and historical contingencies (e.g., the Civil War and World War II), ideological principles were not unimportant. Previous decades of struggle and policies had been carried out in part to realize an alternative modernity constructed according to the ideas of Marx and Lenin: even if their principles were not perfectly followed, Soviet leaders still based their legitimacy on at least some manner of social security and provision and ending inequality based on ownership of capital.

The first institutional logic was a parallel dual structure of power, in which the Soviet state and the CPSU coexisted with sometimes complementary and sometimes overlapping functions. The Soviet state encompassed far more of economy and society than other modern states. In its turn, the CPSU was not merely a political party in the European mold, that is, a collection of political elites and activists sharing ideology and policy goals. Rather, it had a monopoly on political power and acted as an ideological watchdog or functional equivalent of a secular church maintaining ideological purity (Kotkin, 1995). People filling important positions in the state (such as factory managers and heads of security forces) were required to be Party members; this was the infamous *nomenklatura*, literally a 'list' of positions only filled by Party members. As the Party leadership could demand discipline and obedience in return for membership, and membership was required for important state positions, the CPSU inevitably was the true ruling body of the USSR – more specifically, the Central Committee of the CPSU, then the Politburo (the ruling body that met in the absence of the Central Committee), and at the apex the General Secretary.

A second important facet of Soviet politics was the fusion of polity and economy, as the latter was subordinated to the former through centralized state control. While this provided sources of power for the Soviet elite, it also made them vulnerable to criticism and even possible delegitimation (formal or informal), should that economy stagnate for a long period of time. Further, that centralized economy, oriented to the Five-Year Plan, grew out of attempts to control autonomous social forces that could possibly threaten Soviet power. While the 'command economy' could mobilize manpower and resources (e.g., to fight in World War II), the lack of competition and accountability bred inefficiencies that eventually led to stagnation. Yet the command economy remained the basis for a future model, with the role of the state perhaps more modest: something akin to *dirigisme*, active state participation in the economy.

Another important facet of Soviet institutions and structures was ethnic foundations to the Soviet political structure. In the 1920s, after the Red Army had regained most of the territory lost in the Revolution and Civil War, Stalin was responsible for the regime's ethnic policy, including designing new internal territorial boundaries for the USSR. Ethnic identities and conflicts had been a problem in the multiethnic Russian empire, and the Bolsheviks faced not only the possibility such tensions would resume, but also possible contradictions between ethnic and class identity. The Bolsheviks' solution was to allow (relatively) free expression of ethic identities that would still be subordinate both to an all-encompassing (if elastic) 'Soviet' identity and identities and politics of class. Further, under Stalin's direction, the Soviet government constructed internal boundaries corresponding to distribution of ethnic groups as understood in Moscow. Boundaries of 12 'republics' (with the three Baltic countries incorporated by the Red Army in 1939), and additional lower-level administrative boundaries within the RSFSR (the Russian Republic), did not correspond perfectly to the distribution of people claiming those identities. So long as the Communist Party and Soviet state could control the content, expression, and presentation of ethnic identities and relations, this would not be a problem – for the moment (cf Martin, 2001).

From USSR to Russian Federation, and 14 Other Republics

By the early 1980s it was a badly-kept secret that the Soviet economy was stagnating (aging industrial machinery and continuously disappointing harvests) and social problems were on the rise (e.g., alcoholism was increasing the male life expectancy had reached a plateau and was beginning to fall). While Brezhnev's social contract made it difficult to address these issues without threatening his rule and that of his closest allies, after his death in 1982 the time for reform was necessary – but not yet ripe. Paradoxically, the Communist Party's monopoly on political authority ultimately generated both conservative and reform forces: those who gained from the existing system or were ideologically invested in it made up conservatives, but those who desired reform could also find patrons in the Party, so long as they were quiet about desires for reform and bided their time (Cohen, 1979). And the Party did generate its reformers. Iurii Andropov, former head of the KBG (and thus privy to information about the real state of social problems), succeeded Brezhnev as General Secretary (1982–84) and implemented reforms to increase labor and social discipline. While he suffered from kidney failure and was unable to push through much change, he did sow the seeds for reform by bringing in Mikhail Gorbachev as future Party leader. After a year with Konstantin Chernenko as General Secretary (1984–85) – a sop to the Brezhnev generation – Gorbachev was elected to take the late Chernenko's position.

As he consolidated his authority, Gorbachev continued Andropov's policies of tightening discipline; for example, by having the police raid movie theaters to check for workers who should be on the job, or reducing the production of wine in an attempt to combat alcoholism. Needless to say, such tepid reforms did not work. Further, the explosion of the American space shuttle *Challenger* and the disaster at the Chernobyl nuclear power plant in 1985 drove home to Gorbachev and allies that the status quo could not continue. Gorbachev's first bold reform was thus 'New Political Thinking,' an attempt at a new détente with the United States and NATO. Reducing geopolitical tensions would not only reduce the threat of nuclear catastrophe; it would also reduce the need to invest as heavily in the military–industrial complex.

Reducing tensions was a start but would not fundamentally improve. Gorbachev faced a major obstacle to reform: ministries and officials in the command economy would block any attempts at fundamental reform. (In contrast, Deng Xiaoping could institute market reforms because the state bureaucracies had been weakened in the 10-year Cultural Revolution.) Thus, Gorbachev instituted two sets of reforms. The first was *glasnost* (openness), political reform aimed at reducing censorship and freeing up public discourse, in hopes that public criticisms would force state bureaucrats and conservative elites to reduce their resistance to change. The second reform was *perestroika* (restructuring), in which enterprises were given more autonomy from ministries in Moscow and Plan targets. Between 1987 and 1989, enterprise managers could open independent cooperatives or allow shopfloors to produce independently, after they had fulfilled production quotas for state purchases, and sell that output (goods or services) at independently set prices.

In fact, both reforms worked too well. Once it was clear that *glasnost* was a serious reform, critical public discourse expanded, but the target went beyond conservative bureaucrats to the Soviet past and possible legitimacy of communism. As well, managers and workers took the new-found autonomy of *perestroika* and expanded private production, but increasingly at the expense of production for the state and in the name of speculative profit – rather than adding value, managers would obtain deficit goods (whether raw materials or other goods) at cheap rates from the state and resell them at higher prices to other managers who needed these goods. Essentially, Gorbachev's reforms increased managers' autonomy without creating rules and institutions that would guarantee accountability and provide incentives for rational use of resources rather than speculation alone (e.g., private property, bankruptcy law, and complex financial regulations).

As economic performance continued to worsen under *perestroika*, and as *glasnost* unleashed increasing criticism of the status quo, Gorbachev introduced a third reform: *demokratizatsiia* (democratization). While Gorbachev did not intend full-blown democracy and free elections, he did allow for increased participation by independent individuals (but not independent parties) in elections to the Supreme Soviet in 1990. The combination of electoral liberalization and *glasnost* ultimately encouraged the force that would tear the USSR apart: ethnic sentiments and conflict. Despite Moscow's attempts at control, the peoples of various groups did maintain some adherence to ethnic identities, and sometimes those identities led to conflict. The first such moment was in Nagorno–Karabakh, an enclave of ethnic Armenians inside the Azerbaijan Soviet Socialist Republic. In 1988, Armenians in Karabakh began demonstrating for unification with the Armenian Republic, which the Azerbaijani leadership refused. Following this conflict, ethnic movements arose elsewhere. While many of these were

initially organized clandestinely by the KGB to co-opt and control ethnic sentiments, eventually local leaders began to act autonomously: Gorbachev's reforms were weakening Moscow's grip on the country, and officials and elites at all levels now saw the opportunity to make bids for power.

As the increasingly open media reported ethnic mobilization, a demonstration effect ensued, and a 'wave' of ethnic mobilization picked up that would bring down the USSR (Beissinger, 2002). By 1991, the three Baltic republics were calling for independence, and there were public calls for renegotiating the relations between the republics and the basis of the entire USSR. Importantly, democratization and ethnic mobilization had come to Russia: in 1991 Boris Yeltsin was elected the President of the Russian Federation (RSFSR) and could compete directly with Gorbachev, the *Soviet* leader. With the threads of Soviet polity and economy unraveling, a group of Politburo elites decided to restore order with an attempted putsch in August 1991. While the leaders of this coup were able to take Gorbachev into custody, Yeltsin remained free in Moscow to publicly mobilize the crowds gathering to defend the reformers. After several days the putsch failed, but the damage was done. Weary of further possible attempts to reverse political liberalization – and in the process to reduce their newfound privileges – the leaders of the Russian, Ukrainian, and Belarusian republics (Boris Yeltsin, Leonid Kravchuk, and Stanislav Shushkevich, respectively) – met and on December 8, 1991 signed the Belovezh Accords, which dissolved the USSR and replaced it with a weak confederation, the Commonwealth of Independent States.

Ultimately, the USSR suddenly and unexpectedly collapsed because of a perfect storm: a conjuncture of economic stagnation and reforms that created autonomy without accountability led to deterioration in living standards and created the impression that Moscow was losing control of the reform process; this image was only enhanced once liberalization of the media allowed decades of pent-up frustrations and injustices to surface and drive and increasingly critical discourse; and ethnic identities that could use ethnically-grounded institutions of state and Party (i.e., the ethnically defined Soviet republics) as the basis for autonomy vis-à-vis Moscow. Order and control in polity, economy, and society had come apart. Putting it all back together was the supreme challenge that the successors to the USSR would face.

The Yeltsin Years and a Confused Transition to Democracy and Capitalism

Rather than reform Soviet politics and political economy, Yeltsin set out to destroy the old system and build something new. As in the Soviet era (and arguably most other countries), polity and economy were closely intertwined: economic reforms reshaped the landscape of political power and the topics of discourse, and political institutions and policies engendered the reshaping of the economy in the first place.

Russia in 1992 was a country in turmoil. Price liberalization led to spikes in inflation, and privatization of state-owned enterprises was a real project rather than abstract discussion. For 1992, the executive (President Boris Yeltsin) and the legislature (the Supreme Soviet, chaired by Ruslan Khasbulatov) agreed to an uneasy truce in sharing power, with Yeltsin enjoying a year of governing via presidential decree to speed up necessary economic reforms. However, constant inflation, falling output, wage arrears, and rising inter-enterprise debt threatened to bring the economy to its knees (Gaddy and Ickes, 2002). In 1993, the truce between the two branches of government fell apart, and Yeltsin and Khasbulatov engaged in constant sparring over policies and who had ultimate political authority. The tension came to a head when a crowd marched on the Ostankino television station in September 1993; Yeltsin called in the military to quell the protests in Moscow, during which the White House (in which Khasbulatov and other members of the Supreme Soviet were taking shelter) was shelled. Yeltsin called for new elections in December 1993 to approve a new constitution and to select representatives for the Duma, the lower house of a new parliament. (At that time half of the Duma's 450 members were selected in individual mandates – direct election of the representative – while half were selected via proportional representation from party lists.) While the constitution passed, the parliamentary election had a few surprises. Yeltsin's nominal pro-reform party Russia's Choice ended up with the most seats, but the Liberal Democratic Party, headed by flamboyant Vladimir Zhirinovskii, and the Communist Party came in second and third, respectively. This resulted in a Duma that could challenge Yeltsin until the next elections in 1995. While the constitution increased the powers of the president – for example, he could dissolve the Duma and call early elections if they refused his proposal for prime minster three times, and the presidential administration had more power over economic reforms – a standoff persisted between president and legislature. The December 1995 election gave the Communist Party even more seats, and more importantly suggested that their leader Gennady Zyuganov might actually win the 1996 presidential race. While polls in early 1996 suggested Yeltsin would lose badly, by summer 1996 he had managed to improve his ratings, such that Yeltsin and Zyuganov faced off in a run-off election that Yeltsin won on claims that the Communists would turn back the clock on hard-earned reforms for the future.

Yeltsin's second term would be no less contentious than his first, and arguably less successful. The president's health continued to deteriorate, and he was now surrounded not by reformers as in the past, but by his daughter, his chief of staff, and oligarch Boris Berezovskii. Increasingly, Yeltsin appeared as distant and out of touch with reality. Even more important was the August 1998 crash of the ruble, sparked by the inability of the government to pay state debts on time. As the ruble depreciated by five times again the dollar, Russia's economy was hit badly. By summer 1999, a combination of economic pain and a sense of humiliation over NATO attacks on Serbia (long held to be a natural Russian ally) created public anger that would ultimately turn against Yeltsin and reformers. Facing increasing challenges from the Communists in the Duma, in 1999 Yeltsin recruited a seemingly competent and loyal functionary who would defend Yeltsin and his allies. This new prime minister, and eventual president once Yeltsin made the decision to step down before the end of his term, was the head of the FSB (Federal Security Services, the successor to the KGB), Vladimir Putin.

Economic reforms and Yeltsin's own political practices sowed the seeds for the eventual downfall of Russian democracy and the transformation of nascent Russian market capital into state-centered corporatism. First, Yeltsin's strength – his image as independent political actor beholden to no one – was also his weakness: Yeltsin never did create a political party to mobilize support of the public and the legislature for his reforms. His policies of economic liberalization made it difficult to create patronage structures: rewards were a one-time deal (e.g., selling off state-owned firms on the cheap), rather than a constant stream of rewards (e.g., offering positions in the state that had access to rewards, such as bribes or status). In fact, Yeltsin's basic political *modus operandi* was to employ the *lack* of structure: rather than bind allies or opponents or coerce them (except in exceptional circumstances), Yeltsin preferred to take his chances in open competition. Further, Yeltsin's economic reforms created both economic pain and a sense of economic injustice. Price liberalization led to inflation that wiped out average Russians' savings and hindered the capacity of many larger Russian firms to invest in new production or to pay wages in full and on time. Further, to speed up investment in the economy, to gain support for his 1996 presidential campaign, and to create a new bourgeoisie as a bulwark against any return to a Soviet-style economy, Yeltsin used privatization in a manner perceived as less than just. In particular, between 1995 and 1997 Yeltsin sold off potentially lucrative firms, mostly in natural resource extraction, at low prices to economic elites whom Yeltsin courted as political allies. This elite, dubbed 'oligarchs,' had made massive profits in speculation or similar economic activities under Gorbachev or the first years of Yeltsin's administration, and they looked to create their own economic empires: for example, Mikhail Khodorkovskii's oil-centered Menatep empire, or Vladimir Potanin's Interros (centered on Oneximbank and Norilsk Nickel), or Boris Berezovskii's empire in oil, aluminum, and media. While some oligarchs did invest in improving performance and profitability, in public opinion privatization was rigged in favor of a new elite that replaced the old *nomenklatura*, and democracy was undercut by a new capitalism as unequal, or worse, than that under Soviet socialism. While Yeltsin and his reform team were criticized for implementing reforms that were too radical in too rapid a manner – creating more pain and hindering the consolidation of market capitalism and democracy – those within Yeltsin's inner circle claimed that they felt they had a very short period to 'destatize' the economy, that is, to move economic resources into private hands, so that when (not if) communists or other non-liberal politicians came to power, it would be near impossible to rebuild a Soviet-style economy and polity (Hoffman, 2002).

Post-2000: The Politics of Vladimir Putin and the *Siloviki*

At the end of 1999 Boris Yeltsin stepped down from the presidency, handing it over to Putin. Using new political party Unity (Edinstvo) and slowly increasing state control of the media and key firms and sectors in the economy – for example, maintaining state control over Gazprom and Rosneft, taking control of independent television station NTV, and increasing direct management of ORT – Putin maintained the trappings of democratic procedure (i.e., elections) while undercutting the foundations for real political and electoral competition (Fish, 2005; Shevtsova, 2007). Putin was not alone in his quest for power: his regime has drawn personnel and support from the *siloviki*, security forces (literally a 'power elite') who were losers in the redistribution of property and power in the 1990s but who were trained in political manipulation, socialized to have a corporate identity and political discipline, and have few compunctions about using coercion (cf Hass, 2011, Chapter 5 for more data and sources on the *siloviki*; also Shevtsova, 2005, 2007).

If Gorbachev was a child of Khrushchev's era of reform, Putin was a child of Brezhnev's era of order, and while he has been willing to use market and democratic procedures to his own ends, he seems distrustful of both as foundations of political and economic practice (Sakwa, 2007, 2009). While Yeltsin preferred to dismantle much state structure and power, Putin and *siloviki* preferred to rebuild institutional and organizational foundations of power, including parties and the state. In short, Putin took a more corporatist approach to politics and political economy, and his strategy for consolidating power – to the extent there was a coherent, conscious 'strategy' – reveals an appreciation for building institutional hierarchy and centralizing politics around the *siloviki* in the Kremlin. In the 1999 Duma elections Putin made use of the new party Unity to gain a foothold in the Duma – rather than threatening, cajoling, or negotiating from the outside, Putin would have a tool inside the legislative branch. While Unity did not have a majority between 1999 and 2003, Putin could also use a second tool, *kompromat*, to gain political acquiescence. *Kompromat* means legally compromising materials; for example, evidence about legal violations. Given the incoherent nature of Russian laws in the 1990s, it was near impossible to make money or status without breaking some laws, especially vis-à-vis privatization or tax payments. While there were hints Putin made use of *kompromat* to gain Duma support for his policies, Putin used this tool more forcefully in his next step for consolidating political power. Before long he began his attack on the oligarchs, beginning with Boris Berezovskii and Vladimir Gusinskii. Both had extensive media holdings that Putin valued: controlling television meant controlling messages and using once again those tools of propaganda that Soviet leaders had relied upon (although until the 2010s not as bizarre as much Soviet propaganda). Putin threatened Gusinskii and Berezovskii with prosecution for supposed violations of various tax and other laws to force them to sell their media holdings to state-run corporations, such as Gazprom.

With control of television (NTV and ORT) and some newspapers, Putin could reduce criticism of his regime (e.g., of his handling of the second war in Chechnya, begun in 1999). His next step was to silence oil oligarch Mikhail Khodorkovskii, who had been critical of Putin and hinted at political ambitions. Khodorkovskii was accused of violating tax laws, and eventually he was sent to jail and his empire cannibalized, with Iuganskneftegaz, the prime jewel of Khodorkovskii's oil giant Iukos, transferred to state-owned (thus Putin-controlled) oil conglomerate Rosneft. In one swoop Putin had increased the resources and economic reach of the state, now adding

greater control of oil resources as well as natural gas (via gas monopoly Gazporm) and defeating yet another potential political opponent.

United Russia (as Unity was now named) increased its status in the Duma, gaining almost half of all seats in the 2003 Duma elections. By 2004 Putin seemed supreme: the Duma was essentially under his control, and the state was increasing its reach into the economy (especially after the creation of Rossiiskaia Tekhnologiia, a state-owned holding company that owned shares in various industrial and defense enterprises across the country, cf Hass, 2011). Rising oil prices meant increasing money for the state budget, so that Putin could buy political support by increasing pensions and spending money on public works, including repairing buildings, schools, parks, and the like that had been ignored in the cash-strapped 1990s. Corruption increased as well, possibly part of a new social compact between the political elite and the rest of the state apparatus. Russians were living better and believed they could thank Putin. However, new challenges were on the horizon. First, in late 2004 and 2005 Ukrainian citizens took to the streets of Kiev to protest perceived attempts by Putin's ally Viktor Yanukovich (and by extension, of Putin) to steal the presidential election from more Western-oriented Viktor Yushchenko. These protests, dubbed the Orange Revolution, eventually pressured Yanukovich to concede the election, and Putin faced defeat in the one neighbor the Kremlin hoped to keep in its orbit. While Yushchenko was unable to reform Ukraine's polity and economy, and Yanukovich won the next presidential election, the Orange Revolution had spooked the Kremlin elite: if organized dissent could bring down Yanukovich, were Putin and *siloviki* safe? The Kremlin stepped up corporatist control of the polity, including covert (and later overt) coercion of possible opponents and then meddling in electoral procedures to ensure not only victory but victory by a wide margin.

For the 2007 Duma elections, electoral laws were changed that made it more difficult for smaller parties to participate, and United Russia won 70% of the vote, suggesting that Putin had a clear mandate to rule. But Putin made it clear that he intended to follow the constitution and stand down, as he could only serve two consecutive terms. His handpicked successor, Dmitrii Medvedev, won the 2008 presidential election. Medvedev was perceived in some quarters as a potential reformer, but he ultimately did little in his 4 years. First, as part of the deal for stepping down, Putin remained in power as Prime Minister; second, Medvedev did not have a political vehicle like United Russia, as did Putin; and Putin, not Medvedev, was the central player among the *siloviki* networks in the Kremlin. One change that did occur under Putin was a change in the constitution to lengthen the presidential term from 4 to 6 years.

Putin had stepped aside when Russia's economy was still strong from petrodollar wealth, but by 2011 those economic gains were at risk from the global recession and decline in petrodollar income, as well as corruption and problematic economic competitiveness. In the 2011 Duma elections United Russia gained only 49% of votes and 53% of seats, a clear decline from the previous election. Putin and Medvedev decided that Putin should return to the presidency. Accusations of manipulations of voting procedures and other violations of democratic procedure were widespread in the 2011 Duma elections, and Putin's announcement of this return provoked demonstrations in Moscow that were met with police repression against demonstrators and leaders of the opposition, in particular Aleksandr Navalny.

Putin and his elite weathered the storm, although in response new laws were passed that increased the power of the state to scrutinize and control private actors: for example, foreign NGOs were subjected to greater restrictions, as a result of public claims and suspicions that many such organizations were fronts for foreign interference in Russian politics. (Given that Western NGOs had helped Yushchenko in Ukraine's 2004 presidential election, such suspicions could seem valid, even if the reality might be different.) The FSB lobbied for more control over Internet services, which was eventually granted in 2014. While a minority continued to be critical of Putin's regime for its authoritarian practices and lackadaisical approach to corruption, Putin still enjoyed popular support and the lack of an organized, legitimate opposition.

Variation in the Rest of the Former USSR

An analysis of variation in political trajectories for all former Soviet republics is a book in its own right, and this is merely a brief overview (cf McFaul, 2002; Hale, 2005; Ioffe, 2004). Briefly put, post-Soviet political change has taken three forms. The first is basic consolidation of democratic procedures and civil rights: this is the case of the three Baltic countries. The five Central Asian countries (Kazakhstan, Turkmenistan, Tajikistan, Kyrgyzstan, and Uzbekistan) have weak civil freedoms and are basically authoritarian regimes, although some are more managed and relatively peaceful (Kazakhstan) or flirted with relatively more democratic politics (Kyrgyzstan with its 'tulip revolution' in 2005). While Azerbaijan has formal elections, the country is essentially run by one family (Heydar Aliyev, and his son Ilhem Aliyev). Armenia and Moldova also have formal democratic procedures, although suspicions of corrupt electoral procedures persist, such that Freedom House and other groups are skeptical or the real degree of democratization in these two countries. Georgia's transition to democracy has been on firmer ground, relative to other post-Soviet countries, although the balance of power between the executive and legislature has been adjusted several times.

The Slavic countries of Russia, Belarus, and Ukraine are interesting not only in their own rights but also in comparison with these other post-Soviet countries. Russia, as we have seen, began with political liberalization under Yeltsin, but under Putin Russia became increasingly authoritarian. This mirrored the political trajectory of Belarus, which began with democratic reforms after 1991 but, by 1996, made a rapid slide into dictatorship once Aleksandr Lukashenka won his first presidential term. Upon taking office, Lukashenka began repressing political opposition and subjecting localities to direct control from Minsk (the capital). Economic reforms, which were slow and tepid, were essentially frozen, and Lukashenka's state maintained ownership of vast swaths of industry. Ukraine, however, has had an almost random trajectory between democracy and dictatorship (Way, 2005b).

Unlike Russia, Ukraine has had competitive presidential elections in which losers have ceded authority to winners: for example, in 1994 President Leonid Kravchuk lost and ceded authority to Leonid Kuchma, and in 2010 President Yushchenko failed to secure enough votes for the run-off election (eventually between Iulii Tymoshenko and Viktor Yanukovich) and stepped aside. In contrast, a post-Soviet Russian president has yet to step down after a loss: Yeltsin (2000), Putin (2008), and Medvedev (2012) stepped aside for hand-picked successors, and Lukashenka has not allowed real opposition in elections. Unlike in Russia or Belarus, the balance of power between the Rada (the parliament) and the presidency makes it difficult for one branch to have hegemony. At the same time, charges of corruption in Ukrainian politics abound (and have provided the impetus for waves of protest in 2004 and 2014), and Ukrainian parties have shallow roots. Thus, Ukraine still lacks organizational and institutional configurations for consolidating authoritarian rule; but the Ukrainian state and political parties are still too weak to maintain the integrity of politics and create stability in political parties and ideologies between elections.

One explanation for this variation invokes political structures at the end of the Soviet era. Much as in East European countries, the Baltic countries had existing organizations (popular front movements) that could step into the power vacuum after the collapse of the USSR. However, these organizations were not elite tools, but rather were means for political mobilization and articulation of different interests. That is, the foundations for civil society were present in the Baltics. Central Asian republics in the Soviet era were essentially run by a small elite supported by Moscow, and popular organizations were absent. When the Soviet Union dissolved, elites in Central Asia could simply take that political power for themselves without answering to broader movements. Belarus had little real popular organization, leaving a vacuum after 1991: this vacuum led to continuing economic crisis until the parliament created a presidency, and when Lukashenka won that first election, he filled the power vacuum with his patronage structure. Like the Baltics, Ukraine had an organized civic movement, Rukh, whose early participation in post-Soviet politics helped strengthen the role of the Rada and contributed to that fragile balance between executive and legislature. As we have seen, political institutions remained unstable under Yeltsin: this presented Putin with opportunities to create vehicles for elite power. However, this comparison should be viewed as preliminary; more rigorous comparative research into post-socialist political trajectories remains to be done (but see Way, 2005a).

In Place of a Conclusion: Russian Politics and the Crisis in Ukraine

As of this writing, Russian politics is certainly unsettled. Putin's regime remains firmly in power due to a combination of public support, control over media and much of the economy, and overt use of repression against perceived opposition. While Putin has enjoyed constant popular support, it is not entirely clear how much of that support is genuine and how much is manufactured through fear and propaganda, both of which might evaporate were coercion to decrease (as happened to Gorbachev in the late 1980s). The reliance on petrodollars aided Putin's consolidation of power and popularity in the first half of the 2000s, but once the global crisis of 2008 set in and the price of oil dropped, the Russian economy stagnated and one important foundation for elite control weakened. Putin's return to power in 2012 did not go as smoothly as hoped, and he has faced opposition from a segment of the public that fears a real turn to dictatorship and no longer trusts the president (Pavlovsky, 2014).

Putin's newest term became even more contentious in February 2014, when demonstrations in Kiev – the 'Maidan' protests – took a turn toward the violent. Initial attempts to end the months-long demonstration with coercion backfired, and Ukrainian president and Putin ally Viktor Yanukovich fled to Moscow. In the wake of this popular revolution against the corrupt Yanukovich regime, Kremlin-led media framed the new regime as 'fascist,' and suddenly armed military forces in ambiguous uniforms appeared in Crimea – the control of which had been a point of contention between Ukraine and Russia after 1991, in part because of the Russian naval base in the peninsula, and in part because of the sizable ethnic Russian population and popular beliefs in Russian sovereignty over the region. With the approval of the Kremlin, a referendum of dubious validity was held, in which support for unification with Russia received overwhelming support. Not long afterward, pro-Russia separatists in Donetsk and further east organized a movement for local independence and took up arms; the government was cautious at the start but escalated the military response, until full-blown engagement erupted in the summer of 2014. Circumstantial evidence points to Russian participation – for example, the provision of arms and 'volunteers' for the separatist forces – although the degree of participation is not clear. Some groups inside Russia, such as the NGO Soldiers' Mothers of St. Petersburg, claimed that Russian soldiers were wounded in Ukraine and that perhaps 1000 or more were participating in the conflict. (In response, Putin's regime branded this home-grown NGO a 'foreign agent.') In response to Putin's actions, the United States and European Union applied increasingly harsh sanctions to restrict the ability of important Russian firms to raise capital on Western markets and to restrict some Russian elites' rights to travel abroad. While the ultimate duration and effect of sanctions is difficult to predict at present, the immediate effects have been the weakening of the ruble vis-à-vis the euro and dollar and marginal reduction in GDP; combined with existing economic problems, Russia's economy is stagnating even more as a result of these new geopolitical actions. Perhaps just as important, this Western stance has placed Putin in a corner: he has relied on increasingly aggressive rhetoric and positions and ultimately has fewer resources at his command than Western leaders (although whether Western leaders have the collective will to press their advantages remains to be seen): for example, Ukraine and Europe need Russian natural gas, but Russia needs that income more desperately. Further, his increasing intransigence will make it more difficult for Putin to come to agreement with the West without appearing weak.

The consequences of the conflict in Ukraine and responses of the United States, NATO, and European Union are cloudy. The Kremlin has been ramping up anti-Western rhetoric in formal pronouncements and media propaganda, but there is as yet no apparent increase in coercion against critics or potential opponents – and the author did not notice much apparent tension when he visited Russia in summer 2014. Yet, this does not deny the possibility that Putin and his regime are feeling increasing international isolation and are fearful of increasing social tensions at home, especially if international sanctions and expenses of the Crimean annexation and support for the Ukrainian separatists continue to burden the economy. Putin's use of some Soviet traditions – propaganda, active state participation in the economy to subordinate much of that economy to the Kremlin's needs and power, mobilizing parties – served him well in consolidating *siloviki* positions, but 'Putinism' seems incapable of facilitating the fundamental transformation necessary to give Russia a dynamic twenty-first century economy, polity, and society. While economic stagnation by itself did not bring down the USSR, it was a contributor; the Soviet economy, and thus polity, was dependent on petrodollars; and authoritarianism ultimately bred resistance, especially when its leadership appeared increasingly out of touch and irrational. Whether Putin and *siloviki* will strategically alter course or fall into similar traps remains to be seen – the confusion and contention of post-Soviet politics is far from over.

Bibliography

Beissinger, M., 2002. Nationalist Mobilization and the Collapse of the Soviet State. Cambridge University Press, New York.

Bunce, V., 1983. The political economy of the Brezhnev era: the rise and fall of corporatism. British Journal of Political Science 13, 129–158.

Bunce, V., 1999. Subversive Institutions. The Design and Destruction of Socialism and the State. Cambridge University Press, Cambridge, UK.

Cohen, S., 1979. The friends and foes of change: reformism and conservatism in the Soviet Union. Slavic Review 38, 187–202.

Fish, M. Steven, 2005. Democracy Derailed in Russia: The Failure of Open Politics. Cambridge University Press, New York.

Gaddy, C., Ickes, B.W., 2002. Russia's Virtual Economy. Brookings, Washington, DC.

Hale, H., 2005. Regime cycles. Democracy, autocracy, and revolution in post-Soviet Eurasia. World Politics 58, 133–165.

Hass, J.K., 2011. Power, Culture, and Economic Change in Russia 1988–2008. Routledge, New York and Abingdon, UK.

Hoffman, D., 2002. The Oligarchs: Wealth and Power in the New Russia. PublicAffairs, New York.

Hosking, G., 1997. The First Socialist Society. Harvard University Press, Cambridge, MA.

Ioffe, G., January 2004. Understanding Belarus: economy and political landscape. Europe-Asia Studies 56 (1), 85–118.

Kotkin, S., 1995. Magnetic Mountain. Stalinism as a Civilization. University of California Press, Berkeley, CA.

Martin, T., 2001. The Affirmative Action Empire. Cornell University Press, Ithaca.

McFaul, M., 2002. The fourth wave of democracy and dictatorship: noncooperative transitions in the postcommunist world. World Politics 54, 212–244.

Millar, J.R., 1985. The little deal: Brezhnev's contribution to acquisitive socialism. Slavic Review 44, 694–706.

Pavlovsky, G., July/August 2014. Putin's world outlook. New Left Review 8, 54–66.

Sakwa, R., 2007. Putin: Russia's Choice. Routledge, New York and Abingdon, UK.

Sakwa, R., 2009. The Quality of Freedom: Khodorkovsky, Putin, and and the Yukos Affair. Oxford University Press, New York.

Shevtsova, L., 2005. Putin's Russia. Carnegie Endowment for International Peace, Washington, DC.

Shevtsova, L., 2007. Russia: Lost in Transition. Carnegie Endowment for International Peace, Washington, DC.

Way, L., 2005a. Authoritarian state building and the sources of regime competitiveness in the fourth wave. World Politics 57, 231–261.

Way, L., April 2005b. Ukraine's Orange Revolution: Kuchma's failed authoritarianism. Journal of Democracy 16 (2), 131–145.

Russian and Post-Soviet Studies: Society

Andrew Buck, University of Southern Indiana, Evansville, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the distinctive characteristics of Soviet society. It addresses the main social cleavages in Soviet society: status, class, ethnicity, and gender. The article also presents key features of Soviet culture, trust networks, and deviance. Finally, it focuses on issues related to post-Soviet society: inequality, market reform, organized crime, and democratization.

Soviet society is part of a broader and more generic category of society called 'state socialism.' Comparative historical research contrasts state socialist societies with feudal, capitalist, and postindustrial ones. Soviet society most precisely expresses the characteristics of state socialism because the Soviet Union initially adopted this type of society decades before other countries and then exported it throughout the globe. Soviet society and, later, its related state socialist satellites in the Eastern Bloc held a key role in the world system during the twentieth century, but finally collapsed in 1991.

Main Characteristics

Soviet society was influenced by the writings of Karl Marx and his followers. Marx claimed inequality in capitalist societies comes from the social relations of production and, specifically, whether persons own the means for production or not. Therefore, unlike capitalist societies, Soviet society had no private property aside from household and personal items, such as combs, automobiles, etc.; productive resources, such as housing, factories, land, etc. were owned collectively. According to Marxist thinkers, a dictatorship of the proletariat over productive resources and other societal organizations would be necessary as a transitional phase before reaching the ideals of communism. When the Bolshevik Revolution occurred in 1917, the new Soviet leaders did not have blueprints for what this communist future would look like after the transitional phase because very little had been written about it.

In the Soviet Union, the Communist Party and the government bureaucracy, the 'party-state,' coordinated the control of all aspects of society, including productive resources, the political structure, and societal organizations. Far from Lenin's predicted withering away of the state as the transition progressed, the official bureaucracy eventually overshadowed the more ideological focus of the Communist Party and elaborated its logic throughout society by the 1970s. The disjoint between the ideals of communism and the bureaucratic reality led scholars to call the reality 'actually existing' state socialism (Bahro, 1978) to distinguish it from the ideology of communist leaders. As an example of this heavy-handed bureaucracy, Soviet citizens even had limited freedom of movement within the Soviet Union.

Research on Soviet Society

The social scientific study of Soviet society did not reach its full potential during the Soviet Union. Social science in capitalist countries was heavily influenced by the politics of the Cold War. Capitalist governments had an interest in staying abreast of events and trends in the Soviet Union and granted funding for research on it. But, scholars had a hard time accessing objective and systematic data for their studies. Soviet authorities separated foreign academics from ordinary Soviet citizens, making it difficult to undertake their studies. This isolation encouraged academics to use publicly available information from officials and other elites as an alternative to conduct their research. Not surprisingly, sociology and anthropology were underrepresented in the field of Soviet studies, while research about elite politics (most precisely expressed in 'Kremlinology') and economics had a greater presence.

Not only was social science in capitalist countries unable to develop fully, but Soviet scholarship on Soviet society was limited too. It was not until the 1960s that sociology was even recognized as a legitimate discipline in the Soviet Union. After sociology's professionalization, academics found it hard to conduct research projects on current trends in Soviet society. Non-Marxist sociological perspectives of Soviet society were forbidden and even Marxist interpretations of current trends were potentially suspected because they could contradict official ideology and policies. Moreover, one of the main tools of social science, large-scale survey research, did not develop until quite late. Part of this delay stems from the political tradition of letter writing in the Soviet Union. Whereas the ruling classes in some capitalist societies rely on monitoring public opinion through large-scale surveys, Soviet authorities encouraged citizens to write personal letters to newspapers and government bodies about their concerns and opinions. However, these limitations in the social sciences of capitalist countries and the Soviet Union did not stop Soviet studies from generating considerable insights into Soviet society.

Soviet Stratification

Despite its official ideology, the Soviet Union was not a classless society; Soviet society was deeply stratified throughout its existence. But, stratification in the Soviet Union was comparatively less unequal than in many of its contemporary capitalist

societies. The party-state redistributed resources among various societal sectors, organizations, and groups according to its ideology. Official ideology, among other things, promised full employment and welfare, including vacations, health care, and education, to all citizens. The party-state virtually stamped out illiteracy and social privilege based on intergenerational transfers of wealth in a country that had been rural with a ruling aristocracy. While state socialism diminished absolute levels of poverty in the Soviet Union over time, many groups by the 1980s lived modestly compared to persons in structurally equivalent positions in contemporary capitalist societies.

There was also upward mobility for members of the working and agrarian classes during the Soviet Union. Climbing the social ladder partly occurred through improved educational opportunities for citizens during Soviet-style modernization (Fitzpatrick, 1979). The promise of education, health care, and other social services to citizens created position vacancies for qualified persons too. Another factor in the mobility of persons with modest family origins came from the downward mobility of other classes, like the old intelligentsia and rich peasants, through brutal purges and emigration. Other upward streams happened because the regime valorized agrarian and industrial work. The value placed on occupations, such as dairy worker and lathe operator, was not just talk and ideology. These occupations received salaries relatively commensurate with higher skilled occupations, such as teacher and doctor.

Status

While the regime did level these kinds of inequalities to some degree through 'social homogenization' policies, status and power still deeply stratified Soviet society. Elaborated bureaucratic control over society created fine-grained occupational niches and encouraged professional regression to 'pure' kinds of work. Occupational prestige had similarities to capitalist societies, with white-collar jobs enjoying greater status than blue-collar ones (Babushkina and Shubkin, 1986). But, even lower-status work, such as metallurgy and carpentry, that had previously revolved around various jobs formalized into elaborate professions that granted credentials, advancement, and seniority to persons. The privileges associated with certain kinds of white-collar occupations also persisted. So, for example, doctors', teachers', and journalists' children continued to have access to better education, health care, and services through their social networks and socialization. Persons with these backgrounds translated their social and cultural capital into better opportunities for their careers and, ultimately, occupational status in Soviet society.

Another important factor in Soviet stratification – especially at the top of the hierarchy – was power. The Communist Party held the prerogative to promote, appoint, and dismiss personnel for a list of top positions called the 'nomenklatura' (around 750 000 jobs in the Soviet Union). With only about 7% of society as members of the Communist Party, Soviet citizens often joined the party to further their careers because it had such a determining role for advancement. The centralization of control in Soviet society gave the nomenklatura considerable privileges as well. The nomenklatura and their families had access to higher quality products and services than the rest of society. The separation of the Soviet ruling class from the masses was present from the beginning of the Bolshevik Revolution, but became more conspicuous over time. Ordinary Soviet citizens tried to attach themselves to members of the nomenklatura in patron–client chains to improve their situation. By the Brezhnev era, Soviet society was riven by clientelistic relations that fragmented the stratification system into personal tyrannies of elites and their followers (Voslensky, 1984; Rigby and Harasymiw, 1983).

Class

Even though the party-state abolished most forms of private property, there was evidence of a class system operating in the stratification hierarchy. Class analyses often came from abroad or the near abroad in the state socialist societies of the Eastern Bloc because scholars had greater freedom and an audience for their critiques. According to Konrád and Szelényi (1979), the new dominant class of state socialism did not base its power on ownership of the means production, but on its political position to redistribute resources. This class of managers extracted surplus value from the production of workers. As state socialism was the dictatorship of the proletariat, it legitimated its dominant position by claiming centralized control over resources was for the good of all citizens. Expressing this ideological concern for workers, bureaucrats from government and enterprises had strong relationships with the leaders of labor unions and very rarely did ordinary Soviet workers organize in contentious activity against their political masters. Thus, the new dominant class of state socialism was partly able to extract surplus value from labor by co-opting the leaders of labor unions into its ranks.

At the same time, workers had considerable control over their labor on the shop floor because management was focused more on resource flows external to the firm than supervising work inside (Burawoy and Lukács, 1992). The lack of supervision over workers led to chronic problems of foot-dragging, absenteeism, and drunkenness at Soviet workplaces. In less despotic times and places, workers would also use their places of work during regular working and after hours for their own entrepreneurial activities in 'the second economy' (see Stark, 1986). The use of official resources to produce goods and services that could be sold for profit in an unofficial marketplace became widespread and almost tolerated by the 1980s. This contradictory set of class relations between managerial and working classes is encapsulated in a saying of Soviet workers: 'Management pretends to pay us and we pretend to work for management.'

Soviet Ethnic and Gender Relations

Along with status and class, Soviet society had other important social cleavages. The Soviet Union was ethnically heterogeneous with 140 different nationalities officially recognized. The party-state recognized some of these differences with the formal institutions of federalism and other territorial designations. Even within the 15 republic nationalities, there were numerous unique ethnic groups that existed without formal territorial recognition. Soviet authorities tried to diminish

ethnic differences by creating a new Soviet person, 'homo Sovieticus.' The party-state exported Russian language and culture to different parts of the Soviet Union as a 'civilizing' force to create this new Soviet future because Russia was the political center of Soviet nations and there were still legacies of Russian chauvinism from the Russian Empire. Authorities promoted, with varying levels of vigor, Russian language education as a way to achieve the merging of nations. Leaders of different Soviet national republics often conducted official activities in Russian and, in the case of Central Asian nations, even had their native languages adopt Cyrillic-based alphabets.

Despite efforts to merge Soviet nations, the party-state was sensitive not to force an ethnically diverse society into a Russian-only language policy and promoted bilingualism for different ethnic groups. Ethnic minorities had considerable rights to preserve cultural practices that did not contradict Soviet ideology (see Mandel, 1985). Moreover, ethnic minorities were prominent in top positions of their own republics and regions. The achievements in promoting diversity and granting self-determination did not mean that Soviet society had eradicated discrimination and prejudice. Conflict between different ethnic groups in the Caucasus and Central Asia erupted into violence in the 1980s. In the European part of the Soviet Union, many persons from Slavic and Baltic ethnicities continued to hold prejudices and to engage in everyday forms of discrimination against non-European minorities.

The status of women in Soviet society noticeably improved from the prerevolutionary era. Marxist thinkers saw women subordinated in capitalist societies because many women engaged in unpaid labor as wives and mothers in the household economy. Soviet authorities encouraged women to participate in the labor force partly for this reason and partly because there were shortages of males. Losses in World Wars I and II created unbalanced sex ratios with significant deficits of males. Soviet women achieved high labor force participation rates decades before women in capitalist countries. Still, sex segregation of work was considerable with men hoarding more high status positions. Sex segregation was particularly noticeable because women achieved higher levels of education without analogous rates of change in better quality jobs (Gruzdeva and Chertikhina, 1986). Moreover, Soviet women continued to work more than men in the household economy. This situation is similar to the position of working women in contemporary capitalist societies, who often work the 'second shift' at home while their male partners have more time for leisure.

Soviet Culture, Networks, and Deviance

Soviet society had a distinctive culture that accompanied its social structures. There were very few social scientific studies of daily life in the Soviet Union because the barriers were great for international researchers to immerse themselves in Soviet society and ethnography was not a prominent research tradition in Soviet social science. A notable exception was Humphrey's (1983) study of a collective farm that documented the daily rituals and struggles of rural workers. Recent historical work has been reconstructing the distinctiveness of Soviet culture through archival research. Kotkin's (1995) study of Magnitogorsk during Stalin's reign uncovered how ordinary Soviet citizens developed a unique way of 'speaking Bolshevik' to make claims on authorities and others. Further research on Soviet culture has established other everyday forms of resistance against authority. The findings of cultural research underscore that Soviet society had more avenues for resistance than research on formal organizations and social structures claimed.

Much like capitalist countries, the Soviet Union had informal social networks that interpenetrated the formal organizations dominating their society. One type of these relations was called 'blat' in Russian and helped persons overcome the shortages and formalism associated with Soviet bureaucracy. Persons used trust networks to acquire resources that were difficult to obtain, such as household items and good job opportunities. But, these interpersonal networks were not 'one-off,' asymmetric relations that scholars have described in capitalist societies. Instead, blat required more emotional investment and reciprocity to be effective (Ledeneva, 1998). Also, in contrast with some capitalist societies, these trust networks had an instrumental quality rather than one of fostering a civil society. Given the deficit economy and single-party rule, blat relations had an economic rationality in them and lacked the demand to depoliticize interpersonal ties.

Even though blat relations can be used for corrupt purposes, they were distinct from structures surrounding crime and deviance in the Soviet Union. As bureaucratic control expanded throughout Soviet society, so too did the activities deemed to be illegal and deviant. Economic and political activities, such as speculation and political opposition, had harsh penalties that sent millions of perpetrators to the Gulag (see Barnes, 2011). During Stalin's reign, the mere suspicion or even appearance of such wrongdoing was enough for brutal punishment. State repression against its citizens lessened after Stalin's death, but fear of the Soviet penal system drove criminal activity deeper underground. Criminal groups developed elaborate structures and codes to stop detection from Soviet authorities. Among other deviant activities, criminal groups played an important role as enforcers and participants in the second economy and other market-based ventures because profit-making was illegal until the late 1980s. As we shall see, the position of criminal groups in these market structures put them at an advantage during the transition to a market-democratic society.

Post-Soviet Society

Paradoxically, one of the strengths of Soviet society, its ethnic diversity, played a large role in its demise. Popular mobilization of different ethnic groups called for greater national self-determination and grew precipitously after the breakup of the Eastern Bloc in 1989. Starting in the Baltics and then diffusing to other republics, ethnic leaders harnessed massive popular action for independence from Soviet rule. At the end of this wave of protest, the dominant Slavic republics finally declared independence and the Soviet Union collapsed in 1991 (Beissinger, 2002). The rise in contentious gatherings may be related to Soviet citizens recognizing that the standard of living in Western European and North American capitalist societies was higher than what they were experiencing. This diminished

the legitimacy of the party-state and its weakened position debilitated members from stopping popular discontent.

One line of inquiry in post-Soviet studies looks at what happened to the nomenklatura and other elites since the collapse. Have they been able to stay in power or have new elites emerged? Large-scale studies in Eastern Europe and the former Soviet Union address whether the nomenklatura has been able to stay on the top of the postcommunist stratification system (Eyal et al., 1998). Part of the interest in this question is related to the presumed expansion of the polity that should come with adopting democratic institutions. There is significant variation among the 15 former Soviet republics, but overall the evidence suggests that the nomenklatura has been able to maintain its elite positions in a new market-democratic society. Where there has been a changing of the guard, like in Russia, the circulation of elites has seen shallow recruiting from the 'second rung' of the Soviet elite (Hanley et al., 1995).

Assessing other 'winners' and 'losers' of the transition to capitalism has been the focus of further studies. The rapid, 'shock therapy' policies that most of the former Soviet Union adopted to create market-based economies had a deleterious effect on economic growth and inequality during 1990s. The returns in the stratification system for women, entrepreneurs, and the educated were still less than those for persons with political power in 1990s' Russia (Gerber and Hout, 1998). Critics of this rapid approach to reform often point to another state socialist society, China, and its economic prosperity as evidence that gradual policies are a better route to capitalism (Burawoy, 1996). Such comparisons imply that state socialist societies would be better off creating capitalist economies without reforming their political structures because the party-state is still intact in China.

Another line of inquiry in post-Soviet studies has been assessing economic reforms and the effects of mass privatization and rapid marketization on society. The very foundations of the social sciences tried to grapple with an analogous large-scale economic change. Marx, Weber, Simmel, and Durkheim were primarily concerned with understanding a similar question posed by the former Soviet Union's transition to capitalism: What happens when a market economy emerges? Burawoy and Krotov (1992) argue Russia's transition to capitalism has similarities to previous ones and find its economy in a 'merchant capitalism' phase – a developmental state prior to a capitalist mode of production. Other scholars point to distinctive features in the emerging post-Soviet market economy. Hass (2011), for example, argues that theories imported to guide and to analyze market reform in Russia ignore the historical development of various institutions that ordinary Russians have adopted. Practices and structures developed in the post-Soviet economy, such as alternative monies and financial–industrial groups, reflect specific historical circumstances and not just what previous theory predicts.

Another distinctive feature of post-Soviet society has been the expansion of violent entrepreneurship (Volkov, 1999). Criminal groups were in a good position to take advantage of market opportunities that emerged in the 1990s because some of them were already working in the second economy during the late Soviet period. Additionally, the contraction of the military after the Cold War made many specialists in violence and would-be soldiers open to using their knowledge for entrepreneurial endeavors. The demand for protection was also high because the government did not have the capacity to protect the rights of a growing population of new property owners (Varese, 1994). The success of criminal groups in providing protection and extracting tribute from clients eventually helped them get a foothold in various levels of government because these activities have many similarities with the main functions of a modern state.

The economic decline and growing chaos of most of the former Soviet Union during 1990s created conditions for greater control over the economy and polity. With the notable exception of the Baltic countries, much of the former Soviet Union has rolled back freedoms in the economy and politics that some citizens enjoyed during the 1990s (see Levitsky and Way, 2010). Authorities in Russia, for example, deprivatized valuable assets, such as gas and oil companies, owned by foreigners and business leaders supporting political opponents to ensure corporate ownership would cooperate with them. In reasserting control over the economy, authorities have become adept at using industry to further their national and geopolitical interests in the near and far abroad. At the same time, many authorities in the former Soviet Union have quashed internal opposition to ruling interests. The expansion of violent entrepreneurs into business and politics during 1990s has made it easy for members of ruling coalitions to find compromising material on challengers even if they themselves had connections to criminal groups as well.

Recent de-democratization in the former Soviet Union has the potential to compromise some real advances made in social science on the region. In Western Europe and North America, the study of post-Soviet societies has now became part of the mainstream of the academy as another site for social scientific inquiry. In the former Soviet Union, domestic social scientists conduct large-scale, empirical research projects using a variety of methods standards in social science. But, post-Soviet authorities have recently denounced some critical findings of social science research as designed by foreign interests to be biased against them. Emerging official suspicion of and barriers to conducting independent research projects could hurt cooperation between international and domestic social scientists and stall the progress already made.

See also: Area Studies, History of; Capitalism: The History of the Concept; Class: History of the Concept; Development and the State; Eastern European Studies: Society; Elites: History of the Concept; Organizations and Culture; Russian and Post-Soviet Studies: Gender; Russian and Post-Soviet Studies: Politics; State and Society: History of the Concept.

Bibliography

Babushkina, T.A., Shubkin, V.N., 1986. The statics and dynamics of occupational prestige. In: Yanowitch, M. (Ed.), The Social Structure of the USSR. M.E. Sharpe, Inc., Armonk, New York, pp. 126–149.

Bahro, R., 1978. The Alternative in Eastern Europe. New Left Books, London.

Barnes, S., 2011. Death and Redemption: The Gulag and the Shaping of Soviet Society. Princeton University Press, Princeton, NJ.

Beissinger, M., 2002. Nationalist Mobilization and the Collapse of the Soviet State. Cambridge University Press, New York.

Burawoy, M., 1996. The state and economic involution: Russia through a China lens. World Development 24, 1105–1117.
Burawoy, M., Krotov, P., 1992. The soviet transition from socialism to capitalism: worker control and economic bargaining in the wood industry. American Sociological Review 57, 34–52.
Burawoy, M., Lukács, J., 1992. The Radiant Past: Ideology and Reality in Hungary's Road to Capitalism. University of Chicago Press, Chicago.
Eyal, G., Szelényi, I., Townsley, E., 1998. Making Capitalism without Capitalists: Class Formation and Elite Struggles in Post-Communist Central Europe. Verso, London.
Fitzpatrick, S., 1979. Education and Social Mobility in the Soviet Union, 1921–1934. Cambridge University Press, New York.
Gerber, T., Hout, M., 1998. More shock than therapy: market transition, employment and income in Russia, 1991–1995. American Journal of Sociology 104, 1–50.
Gruzdeva, E.V., Chertikhina, E.S., 1986. Soviet women: problems of work and daily life. In: Yanowitch, M. (Ed.), The Social Structure of the USSR. M.E. Sharpe, Inc., Armonk, New York, pp. 150–169.
Hanley, E., Yershova, N., Anderson, R., 1995. Russia – old wine in a new bottle? The circulation and reproduction of Russian elites 1983–1993. Theory and Society 24, 639–668.
Hass, J., 2011. Power, Culture, and Practices in Russia. To the Undiscovered Country of Post-Socialism, 1988–2008. Routledge, New York.
Humphrey, C., 1983. Karl Marx Collective: Economy, Society and Religion in Siberian Collective Farm. Cambridge University Press, New York.
Konrád, G., Szelényi, I., 1979. The Intellectuals on the Road to Class Power. Harcourt Brace Jovanovich, New York.
Kotkin, S., 1995. Magnetic Mountain: Stalinism as a Civilization. University of California Press, Berkeley.
Ledeneva, A.V., 1998. Russia's Economy of Favours. Blat, Networking and Informal Exchange. Cambridge University Press, New York.
Levitsky, S., Way, L., 2010. Competitive Authoritarianism: Hybrid Regimes after the Cold War. Cambridge University Press, Cambridge.
Mandel, W., 1985. Soviet but Not Russia. The 'Other' Peoples of the Soviet Union. Ramparts Press, Palo Alto, CA.
Rigby, T.H., Harasymiw, B., 1983. Leadership Selection and Patron-Client Relations in the USSR and Yugoslavia. Allen Unwin, London.
Stark, D., 1986. Rethinking internal labor markets: new insights from a comparative perspective. American Sociological Review 51, 492–504.
Varese, F., 1994. Is Sicily the future of Russia? Private Protection and the rise of the Russian Mafia. Archives Europeennes de Sociologie 35, 224–258.
Volkov, V., 1999. Violent entrepreneurship in post-communist Russia. Europe-Asia Studies 51, 741–754.
Voslensky, M., 1984. Nomenklatura: The Soviet Ruling Class. Doubleday, Garden City, NY.

Russian Revolution, The

Manfred Hildermeier, Goettingen University, Goettingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following brief comments on the term and time frame, the article outlines the principle aspects and effects of the Revolution. In the foreground are the political constitution and system of government (parliamentary vs soviet government), the social upheaval (rise of the lower classes, new functionary elites), the disintegration of the empire, and the rise of national independence movements like the cultural revolutionary 'Iconoclasm'.

The term, Russian Revolution, is used ambiguously. It refers to varying events and time frames. Normally, it means the major upheaval of 1917 that replaced the tsarist empire with Bolshevik rule. In this case, the term can refer to either the demise of the monarchy in the so-called *February Revolution* (24 February/8 March NS to 2/15 March) or to the *overthrow of 25–26 October* (7–8 November) when the Bolsheviks seized power in the Petrograd Soviet of Workers' and Soldiers' Deputies. But 'Russian Revolution' is often extended to denote longer processes and obviously related sequences of events of the same type. In this case, the term encompasses the 1918–20 Civil War, which immediately followed the October Revolution as well as the subsequent early Soviet period up to the *Stalin Revolution* of 1929–30 or – the widest definition – the entire transformation of Russia from the turn of the century and so-called *First Revolution* of 1905–06 all the way down to the Stalin Era. Looking at it this way, the 'Russian Revolution' covers Russia's peculiar political and socioeconomic modernization crisis, which ended in the establishment of the Soviet Communist dictatorship (1900–30).

Substantive reasons and the question of research feasibility suggest narrowing this range of definitions somewhat. 'Russian Revolution' should be restricted to the convulsive transition from the Russian *ancien régime* to the Bolshevik form of government. The term properly denotes the two core events of the transfer of power: the overthrows of February and October and their consequences. However, there are valid reasons to include the Civil War because it was by nature the inevitable, belatedly waged war of revolution. Both time frames (1917 and 1917–21) lend themselves to differentiation of processes into *politics and power struggle*, *socioeconomic* change, and *cultural and nationalist* developments.

Political Constitution and System of Government

Political change in 1917 consisted of the succession from the monarchy to the 'emergency democracy' (D. Geyer), the Provisional Government (2/15 March to 26 October/8 November), to the soviet-based rule of the Bolsheviks with the 'Left Socialist Revolutionary Party' as a junior partner. Neither the direction of change nor the options in the year of revolution can be understood without a look at preceding developments. By the same token, the consolidation and transformation of the outcome cannot be explained without reference to the Civil War. The impetus for criticism of the autocracy definitely came from circles just outside the seat of power. Liberal forces in the aristocracy, the industrial–commercial upper class, which had developed ample political self-confidence, and the recently established academic intelligentsia rallied to the cause of constitutional restraint of the absolute monarch. But already during the grave state crisis of 1905–06 widely differing views were held on the extent of a say the parliament (Duma) should get in government. Following the restabilization of tsarist rule only moderate demands were still voiced by the groups that were allowed to operate legally. Nonetheless the 'revolution', a term, which cannot withstand any close scrutiny, clearly engendered a new political system and apparently a new distribution of power, although opinion is divided on the latter claim. Its legacy included a parliament, a party spectrum (albeit an incomplete one) and a politically outspoken press. Moreover, regional self-government (*zemstva*) had appeared alongside the central government and gained more and more powers. The First World War accelerated this shift. Ultimately the *zemstva* guaranteed the last remnants of public administration, which otherwise almost had ceased to function. Food supply had broken down, fuel was scarce, and the black market flourishing.

Only this background explains why a protest by workers' wives against hunger and cold swelled to a mass uprising, which toppled the monarchy. The last week of the 300 year Romanov rule began with an International Women's Day demonstration, which only differed from numerous past rallies in the number of participants. A repeat performance represented a new quality. In the next few days the movement escalated into a general strike of workers in the capital, and local military units joined the protest. The revolution, in the strict legal sense as a new edition of sovereignty, took place on 28 February when a Duma committee without authorization of the tsar appointed itself as the Provisional Committee for the Restoration of Public Order, thereby seizing power. At the same time the Workers' Soviet, which had convened the day before on the lines of the fall 1905 council, elected an analogous body. This established the *two principal political institutions*, which constituted the so-called *February Regime* ('dual power') of the following year. They represented two different social classes, requiring mediation of interests by the Provisional Government, first formed by the Liberals (embodied by the *Constitutional Democratic* Party), after 5/8 May together with the moderate wings of the *Socialist Revolutionary* and *Menshevik*

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62076-7

Social Democratic parties. The general correspondence between social affiliation and party orientation including institutional representation is indicative first of the marked polarization in the society and second of the absolutely crucial requisite for survival of the February order: building a bridge between the two fronts.

This became harder and harder to do from early summer on. Several factors contributed. With the government and the ruling parties it was a growing induration of personal animosities and of positions on issues. But a more serious aggravation was the result of omissions and a bad decision. The *Kerensky* Offensive of June/July, named after the premier, was a fatal error. It ended in a fiasco, which not only caused foreign policy damage. It also discredited the Provisional Government at home. Economic depression worsened simultaneously. Unemployment spread as more and more companies failed. Strikes and demonstrations proliferated, bringing the state to the brink of anarchy. The formation of new coalitions and two 'Reconciliation Congresses' (with different makeups) were unable to check the rapid erosion of authority of the February Regime.

This benefited the *Bolshevik* Social Democratic Party, which declared an irreconcilable struggle with the 'capitalist government' when their undisputed leader, V.I. Lenin, returned to Petrograd from Swiss exile on 3/16 April. They only had to channel the workers' and soldiers' misery and disenchantment over the failure of the new system. With all this grist for their mills, the Bolsheviks easily won great popularity. The spring splinter party was the strongest political force by the time the capital city councils were reelected in September. What had happened and was about to happen was clearly evidenced by the dramatic change of leadership in the Petrograd Workers Soviet. The Bolshevik Trotsky replaced the Menshevik Chkheidze as chairman.

The 'Red October' (R.V. Daniels) has produced much controversy. Soviet historiography, adhering to ideology, saw it as the investing of the first socialist state in the history of the world by the power of the masses. Others insisted on interpreting it as a military coup. There is no question that the seizure of power in the city garrisons from 22 October to 4 November was executed before everyone's eyes by the Bolshevik-dominated Military Revolution Committee in defense of the Soviet. It is also undisputed that the Bolsheviks were able to obtain the support of a majority of delegates to the Second All-Russian Congress of Soviets, which convened on 25 October. With this backing they managed to elect a presidium, more than half of whose members were their loyal supporters, while another third consisted of their Left Socialist Revolutionary sympathizers. To top it all off, the 'Right' Social Revolutionaries and the Mensheviks did them the favor of resigning their seats. Finally, the unchallenged occupation of strategic buildings and points in the city suggests that (unlike in Moscow a little later) there was no resistance worth mentioning. However, it is equally obvious that armed force was used against a legitimate government that had become unpopular but had been elected by parliament. The capture of the Winter Palace and the arrest of ministers pointedly attest to the nature of the overthrow. The maneuver accomplished by the Bolsheviks with the help of their Red Guard was 'technically' a military rebellion and politically a *coup d'état*. But it was only possible because it was undergirded by the majority of the Soviet deputies and the tolerance, if not the sympathy, of the urban lower classes, and locally stationed soldiers in particular. Thus, the 'art of insurrection' preached and wholeheartedly practiced by Lenin more or less boiled down to picking up the power that was lying in the street (N. Sukhanov).

The only question remaining was whether the Bolsheviks could maintain control. The wishes of the rebellious and potentially rebellious masses were quickly satisfied. Early on 26 October, the Soviet Congress passed a Peace Decree and a Land Decree. The intention of obtaining an immediate cease-fire secured the support of the soldiers for the overthrow. Distribution of large estates among village communes sanctioned transfers that had taken place much earlier and neutralized the peasantry. Overcoming the other parties and adherents of the *ancien régime* entailed a real struggle. It was inextricably tied to the independence movements of non-Russian nations and lasted until March 1921. Beginning in spring 1918 it escalated into the bloodiest war fought on Russian soil since the early seventeenth century. Actually, it started, still without recourse to weapons, right after the October coup. In the end the Bolsheviks maintained the upper hand so that they could begin the peaceful construction of 'their' state. However, at this juncture the fundamental decisions on its structure had already been made, and they were only superficially and temporarily amended, if at all, by the proclamation of the New Economic Policy.

The most important trend reflected Lenin's practical 'interpretation' of the slogan, 'All Power to the Soviets', as 'All Power to the Bolsheviks', which is what he had in mind in the first place. This usurpation was gradual. The cooperation between the Bolsheviks and the Left Socialist Revolutionaries in the October night session became a formal coalition in mid-November. Even this collaboration was based on the exclusion of not only the Liberals, but also the Mensheviks and 'Right' Socialist Revolutionaries. This confrontation manifested itself in the *alternative* between the *soviets* and the *Constituent Assembly*. When the latter was forcibly dispersed on 6/19 January 1918, it was not only the final curtain on the attempt to erect a democracy in posttsarist Russia. It also strained the common ground between the coalition partners. Though the Left Socialist Revolutionaries' Central Committee agreed to the dissolution, tensions within the bipartisan government grew. The collapse came after intense conflicts over the separate peace treaty of Brest-Litovsk in March. Now one-party rule and the fusion of party and state began in earnest. The system that lasted until 1991 was established.

Moreover, at least three other features of the new government and political system had their roots in the post-October clashes. First, the life-and-death struggle fostered the emergence of extraordinary organs and extreme centralization of decision making. Both measures were only partially rescinded. The most important institutional legacy was the Cheka, which was soon an army of its own, operating in the hinterland and authorized to administer martial law. As a secret police, the Cheka later became a synonym for totalitarian rule because it was empowered to execute arbitrary force beyond and above the law and was only subordinated to the top party leaders. A second fundamental decision was subjecting the economy to the control of the party and the government. Starting in

December 1917, the regime nationalized first the banks and big industry, then during the Civil War commerce and small business as well. Officially control was assumed by a pyramid of economic councils; *de facto*, however, since the unions had been forced to march in step, by members of the only legal party. In addition, an increasing intolerance of intraparty criticism reinforced the concentration of power and authority. The relative laxness, which the Civil War encouraged in 1919 yielded at the conclusion of hostilities to a 'prohibition of factions' (March 1921), which Stalin used as a decisive weapon in the power struggle for Lenin's succession. Finally, the Red Army was created in the fight for the spoils of the October Revolution. Even though its structure and personnel underwent several reshuffles, one birthmark was never effaced – its orientation on the regime and the party.

Social Upheaval

The revolutionary movement united political and social objectives from the start. It laid its claims in the name of the lower classes, which it always viewed as exploited. The Socialist Revolutionaries acted on behalf of all the 'enslaved', but primarily the peasants, while the Social Democrats marched in the name of the urban 'proletariat'. The Liberals, on the other hand, considered themselves to be 'class neutral'. But actually this camp had a social stratum of its own, ranging from the reform-oriented aristocracy and the intelligentsia to the commercial–industrial bourgeoisie. In the first revolution, these various tendencies underwent symbiosis. The nearly identical timing of the general strike in the capital and the peasants' storming of the agricultural estates in fall 1905 shocked the state to such an extent that it agreed to concessions to the democratic demands of the Liberals.

In February 1917 similar convergence took place, with one fundamental difference. Instead of the pincer attack of workers and peasants, an alliance of workers and garrisoned (not front line) soldiers led the assault. This time the *peasants* remained aloof. The agrarian revolution took place later, but more violently than ever before, *after* the collapse of the tsarist state and its rule by force in summer 1917. Later the rural population divorced itself from the urban revolutionary movement and went its own separate way until fall 1918. The core significance of the upheaval can be summed up with *one* result: the wildcat (i.e., spontaneous) redistribution of the property of the landed aristocracy. The peasants seized that which they had always believed to be their rightful possession. The predominant form of the operations was characteristic of the peasants' ties to tradition. Most villages acted collectively and allotted the land to commune members according to the dictates of custom to be used for a certain length of time. Ownership remained in the hands of the *obshchina*, and private property came into existence *de facto* only because hardly any redistribution occurred thereafter. Following this, their revolution, the peasants were sated. Events and problems beyond their village left them cold. Since the New Economic Policy legalized the reform, they could perceive themselves as the winners of the revolution despite the horrors of the Civil War. However, this redistribution wreaked havoc on the economy. It abolished the large cultivated holdings which yielded a large part of the farm produce sold elsewhere and reduced the average acreage to a level which hardly sufficed for efficient and stable production for the market. The outcome was a crisis-proneness that provided a convincing argument for a centrally planned economy.

The consequences of the February 1917 general strike of the Petrograd *workers* were more complicated. From the standpoint of the protagonists, the medium-term ramifications were negative. The rebels only managed to obtain their demands in the short run. Wages increased substantially, and working hours were reduced, even though the principal demand, a law stipulating an 8 h workday, was not achieved. Personal and social security were markedly improved. But the galloping inflation and unemployment following the economic collapse offset these benefits from early summer on. This setback made it possible for the syndical factory committee movement to extend the say in management it had gained in Petrograd in March to full control more than ever before. Its victory was hollow because the victors inherited the administration of the bankruptcy. To add insult to injury, the expectations it cherished after the October overthrow were completely dashed. If the Bolsheviks had a typical way of dealing with 'their' constituency, it was revealed in their relations with the factory committees. Before the overthrow they adopted the demands of the committees; after the Bolsheviks seized power, they also succeeded in gaining control of these organs. At the beginning of the year the 'anarchistic-decentralized' factory committees were muzzled by the unions, which themselves were on the way to becoming a 'transmission belt' (Lenin) of the party. All that remained was the disappointment over unkept promises, which triggered spreading strikes, the rebellion of Kronstadt sailors and intraparty criticism ('Workers' Opposition'). All this dissidence was suppressed by spring 1921, in many cases with extreme brutality.

Disintegration of the Empire and National Independence Movements

The liberal political and sociological views of the Russian Revolution have overlooked too much evidence that it was also a *national* movement. Incipient emancipation campaigns were exhibited in the first 'crisis of authority' in 1905–06. Social unrest in the Baltic Provinces and in Poland were particularly violent. There and in the Caucasus it expressed a desire for autonomy. The first authentically freely elected State Duma had several national factions and a total of about 45% non-Russian deputies. Following the 'prewar' and war period, which gave these movements an enormous boost, and the collapse of the tsarist empire, the national outbursts flared up as attempts to secede by force. Favored by the 'an-archy' in the literal sense created by the power vacuum, not only the 'old' nations like Poland and Finland and non-Slavic Baltic and Caucasian regions declared independence from the Petrograd central government. Nuclei of the old empire such as Ukraine joined the breakaway trend. It was not without cause that the June 1917 resolution of the Ukrainian *Rada* (council) declaring it would act as a formally independent body plunged the Provisional Government into an extremely severe crisis. The Bolsheviks supported this national emancipation only verbally

and for tactical reasons. Prior to their power seizure and during the Civil War in the peripheral areas of the empire, they espoused the doctrine that nation states had to reclaim their historical rights in the 'bourgeois phase' of the revolution. However, the stronger the *socialist* self-awareness of the new regime became, the more it again embraced the necessity for a unitary state. In fact, the Red Army acted from the beginning as an arm and an instrument of Great Russian centralism, which regained control of the old empire – with the exception of Finland, Poland, and the new Baltic republics – by the end of the Civil War. If one looks very closely, this resulted inevitably from the political foundations of the new state. Since only one political force was legal, it had to be centralist. Therefore the Soviet Union established in 1922 was only formally a federal system. In truth it was an anachronistic multinational empire in a postwar world organized into nation states and dominated by nationalisms.

Iconoclasm

Last but not least, scholars have recently remembered to highlight the extremely radical cultural upheaval the Russian Revolution entailed. In the first place, it had a salient 'ideocratic' dimension. The Leninist version of radical Marxism was an ideology with the political power to stake a claim to shaping all sectors of public life, ultimately also of private life of the people under its sway, from work to education, according to fundamentally new, theoretically preconceived principles. The October overthrow ushered in the first regime in modern history to make a specified interpretation of world history and destiny, including a programmatic, militant atheism, obligatory. The radical Bolshevik Revolution unleashed an unprecedented outburst of all kinds of 'antiestablishment' movements, almost entirely borrowed from Western European modernism. They were of a sociopolitical (women's liberation), intellectual aesthetic (abstract art, '*Proletkult*') and pedagogical nature (polytechnic education, public health). Even though this experimental exuberance was a source of aggravation to the party and government leadership and its wings were gradually clipped after the conclusion of the Civil War, it survived in isolated pockets until definitively forced into submission by Stalin. In the aesthetic sector, the 'revolutionary culture' is universally credited with significant achievements.

See also: Bureaucratization and Bureaucracy, History of; Capitalism; Communism, History of; Dictatorship; Enlightenment; First World War, The; Imperialism; Industrialization; Inequality, Social; Labor History; Modernization and Modernity in History; National Liberation Movements; Political Thought, History of; Social Movements: A Social Psychological Perspective.

Bibliography

Altrichter, H., 1997. Rußland 1917. Ein Land auf der Suche nach sich selbst. Schöningh, Paderborn.
Brovkin, V.N., 1987. The Mensheviks after October. Socialist Opposition and the Rise of Bolshevik Dictatorship. Stanford University Press, Stanford, CA.
Brovkin, V.N., 1994. Behind the Front Lines of the Civil War. Political Parties and Social Movements in Russia, 1918–1922. Princeton University Press, Princeton, NJ.
Figes, O., 1989. Peasant Russia, Civil War. The Volga Countryside in Revolution (1917–1921). Clarendon, Oxford.
Figes, O., 1996. A People's Tragedy. The Russian Revolution 1891–1925. Cape, London.
Galili, Z., 1989. The Menshevik Leaders in the Russian Revolution: Social Realities and Political Strategies. Princeton University Press, Princeton, NJ.
Gill, G., 1979. Peasants and Governments in the Russian Revolution. Macmillan, London.
Hasegawa, T., 1981. The February Revolution: Petrograd 1917. University of Washington Press, Seattle, WA.
Koenker, D.P., Rosenberg, W.G., 1989. Strikes and Revolution in Russia, 1917. Princeton University Press, Princeton, NJ.
Leggett, G., 1981. The Cheka: Lenin's Political Police. Clarendon, Oxford.
Pipes, R., 1990. The Russian Revolution 1899–1919. Knopf, New York.
Rabinowitch, A., 1978. The Bolsheviks Come to Power. Norton, New York.
Rabinowitch, A., 2007. The Bolsheviks in Power. The First Year of Soviet Rule in Petrograd. Indiana University Press, Bloomington, IN.
Rosenberg, W.G., 1974. Liberals in the Russian Revolution. The Constitutional Democratic Party, 1917–1921. Princeton University Press, Princeton, NJ.
Steinberg, M.D., 2001. Voices of Revolution, 1917. Yale University Press, New Haven, CT.
Stites, R., 1989. Revolutionary Dreams. Utopian Vision and Experimental Life in the Russian Revolution. Oxford University Press, Oxford.
Wade, R.A., 2008. The Russian Revolution, 1917, 3 pr. Cambridge University Press, New Haven, CT.
Wildman, A.K., 1979, 1987. The End of the Russian Imperial Army, 2 vols. Princeton University Press, Princeton, NJ.

S

Sacrifice

David Hicks, State University of New York, Stony Brook, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13439–13441, © 2001, Elsevier Ltd.

Abstract

There is no scholarly consensus regarding the meaning of the term 'sacrifice,' and the term is best regarded as a polythetic category that includes heterogeneous modes of behavior motivated by human intentions of variable kinds. These include commemoration, initiation, expiation, propitiation, establishing a covenant, or affecting a transition in social status. In more elaborate forms than straightforward ritualistic gift-giving, sacrifice typically involves a set of four terms: (1) sacrificer, the executor of the ritual; (2) sacrifier, the beneficiary of the sacrifice; (3) the object sacrificed; (4) the deity who receives the sacrifice. Various 'approaches' to the study of sacrifice have been proposed, and among the most useful are: (1) sacrifice as gift; (2) sacrifice as communion; (3) sacrifice as causality; (4) sacrifice as symbol system; and (5) sacrifice as catharsis. Sacrificial rituals have an ancient pedigree that extends into prehistory. So pervasive is it in time and space that a case could be made that sacrificial behavior is one of the natural dispositions common to the species *Homo sapiens*.

There is no scholarly consensus regarding the meaning of the term 'sacrifice.' In its widest usage the term, which derives from the Latin *sacrificium*, is applied to virtually any form of gift-giving to a deity. A more restricted application limits its use to situations in which the gift is mutilated or destroyed. An even more restrictive usage is suggested by the Oxford English Dictionary (Simpson and Weiner, 1989/1933), whose primary definition is 'the slaughter of an animal (often including the subsequent consumption of it by fire) as an offering to God or a deity.'

The term sacrifice is, therefore, best regarded as a polythetic category which includes heterogeneous modes of behavior motivated by intentions of variable kinds. These include commemoration, initiation, expiation, propitiation, establishing a covenant, or affecting a transition in social status. Regarding sacrifice not as constituting a class of phenomena having a fixed set of features in common but as a multivocal term applying to phenomena that merely have 'family resemblances' in common has the advantage of mitigating somewhat the Euro-centric bias inherent in the work of many of the most influential thinkers on the subject, among them Robertson Smith (1894/1927), Hubert and Mauss (1898), and Durkheim (1912). 'The Sacrifice' for this generation of scholars was the central feature of religion, a Judaic Christian bias still evident as late as 1956, the year in which E. E. Evans-Pritchard published his classic, *Nuer Religion*. Tellingly, Evans-Pritchard did not identify any word in the Nuer language that unambiguously corresponded to the English 'sacrifice.' This lack of correspondence appears widespread. Benveniste (1969: p. 223), for example, could discover no common word for sacrifice in the Indo-European languages.

The Terminology of Sacrifice

In more elaborate forms than straightforward ritualized gift-giving, sacrifice typically involves a set of four terms. First, there is the sacrificer, the executor of the ritual. Second, there is the sacrifier, the beneficiary of the sacrifice. This may be the executor of the ritual, individual members of the community, or the community *in toto*. Third, there is the object sacrificed. Although this gift may be anything, ethnographic descriptions usually emphasize the bloody sacrifice of a living being, in which case blood typically is described as connoting such ideas as fertility, renewal, replenishment, vitality, life force, and empowerment.

Access by the sacrifier to these desirables is acquired when this substance is shed in ritual by the act of forcing life out of the victim. Substitution, or the use of a surrogate in cases when the offering is that of a human being, is a common occurrence, so frequently resorted to that some writers, including two who have made important contributions to the topic – de Heusch (1986) and Bloch (1992: p. 32) – consider substitution to be a key feature of sacrificial rituals. Fourth, there is the deity to whom the sacrifice is made.

Approaches to the Sacrifice

A number of distinguished thinkers have advanced what might be termed 'approaches' to the study of sacrifice, which afford insightful, though overlapping, perspectives. Among the most useful are the following five.

Sacrifice as Gift

For Tylor (1871/1913), following Plato, the sacrifice was a gift made to an anthropomorphic divinity, much as a gift might be given to a chief, whose acceptance made the deity indebted to humanity. This approach was popular with nineteenth century scholars, who attributed to the act of sacrificing such motives as expiating sins, obtaining benefits, nourishing the deity, and

establishing good relations between human beings and spirits. This theory lays stress on the beliefs of the givers rather than on the ritual performance itself, which in this interpretation is relegated to logically subservient status. This approach is at the same time too broad and too narrow. It is overly broad in that these motives undoubtedly are widespread in many rituals of sacrifice all over the world, yet excessively narrow since sacrificial rituals are employed to do much more than simply offer gifts.

Sacrifice as Communion

Robertson Smith (1894/1927), in contrast to Tylor, granted priority to ritual, a ranking modern ethnographic fieldwork tends to validate. Robertson Smith's main point was that early Semitic sacrifice, and by implication sacrifice among nonliterate peoples, was a feast human beings communally shared with their deity. He understood the immolated victim to be the clan totem, an animal that shared in the blood of its human kin. It was also the clan ancestor, which was the clan deity, so in consuming their animal victim members of the community were eating their deity. This act of consumption conferred spiritual power.

Later research in the field undermined much of Robertson Smith's argument, but unquestionably the sacrifice does offer a means by which human beings may conjoin with their deity and establish a relationship, and here Robertson Smith did have insight. He strongly influenced Hubert and Mauss, whose *Essai sur la nature et la fonction du sacrifice*, which appeared in 1898, remains the basic work on the topic. Like all the studies of the time, it depended entirely upon secondary and tertiary sources rather than original ethnography, and the authors further restricted their sources to Sanskrit and Hebrew texts. To them the most significant feature about the sacrificial feast was that it provided the mechanism by which communion, in the sense of communication between human beings and deity, could be brought about. It was made possible by the fact that the victim represented both sacrifier and divinity, thus breaking down the barrier between the sacred and the profane (a moral contrast that played a major role in their approach). Yet, although the sacrifical act conjoined the sacred and profane, it could also disjoin them since while there were sacrifices that brought the spirit into the human world and made the profane sacred, there were also sacrifices that removed the spirit from the human world and made what had become sacred, profane. The former Hubert and Mauss called rituals of 'sacralization;' the latter, rituals of 'desacralization.' Even though Evans-Pritchard – who was very much influenced by their ideas – found this distinction among the Nuer (1956: pp. 197–230), it is by no means as universal as they supposed.

Hubert and Mauss were intrigued by those sacrifices in which it is the deity that offers his life for the benefit of his worshippers. As they saw it, the exemplary sacrifice was that of a god, who, through unqualified self-abnegation, offers himself to humankind, in this way bestowing life on lesser beings. The most celebrated example of this is the central ritual of the Catholic religion, the Mass, where Christ's sacrifice on the Cross is re-enacted. In a sacrifice of this nature the relationship between the deity and worshippers raises the question of relative status. From Hubert and Mauss' perspective the relationship is one of grave imbalance since the deity gives more than it gets. This may appear so, but Valeri (1985: pp. 66–67) has advanced a sophisticated argument designed to prove that sacrifice actually creates a bond of 'mutual' indebtedness between the human and the divine, which makes each party dependent upon the other. This proposition may not be true universally, but support for Valeri's argument comes from the bear sacrifices carried out by the Ainu of Japan. In these, the immolation of bears brings fertility to the community and nurture to the spirits (Irimoto, 1996).

Durkheim's (1912) distinctive contribution, which otherwise relied heavily upon Hubert and Mauss, and Robertson Smith, was to promote the importance of sacrificial activities in bringing about social cohesion and the values that contribute to this cohesion. In Timor, a South Pacific island, a ritual is performed in which a local king is figuratively slain by members of two ethnically different communities (Hicks, 1996). The king, who comes from one of these groups, symbolizes their unity as a single social entity at the same time that he symbolizes their local god. In addition, by periodically immolating their king the two groups regenerate their notions of 'society' and 'god' as epistemological categories.

Sacrifice as Causality

For Hocart (1970: p. 217) the most compelling reason for performing collective rituals was the 'communal pursuit of fertility.' A sacrifice is offered to the gods in order that life and the fertility that engenders and sustains it are acquired by the community, a view that focuses upon sacrifice as a practical device to affect empirical transformation. Catholic rituals before the Reformation were thought of in this way; afterwards they were reinterpreted as a symbolic process (Muir, 1997). For Catholics the Mass materially transformed the bread into Christ's body; for Protestants the bread merely represented Christ's sacrifice.

Sacrifice as Symbol System

More contemporary scholars, such as Valeri (1985), have emphasized the symbolic character of sacrificial behavior, analyzing its structure and isolating its motifs within the wider context of a society's ideology. In this approach, the array of symbols isolated is then apprehended as a semantic system whose meaning requires decoding. Valeri's interpretation of Hawaiian sacrifice, for example, requires that sacrificial rituals be understood as a symbolic action 'that effects transformations of the relationships of sacrifier, god, and group by representing them in public' (1985: pp. 70–71). Although vulnerable to the criticism that it relies too heavily on the subjective bias of the analyst, this approach has the advantage of enabling sacrificial practices to be understood within a wider social context, at the same time as it allows an analysis of sacrifice to open up a distinctive perspective on society.

Sacrifice as Catharsis

Another approach to sacrifice is to focus on violence. Girard (1972), most notably, has suggested that sacrifice offers

a socially controlled, and therefore, socially acceptable, outlet for the aggressive urges of human beings. The typical manner by which this is accomplished is to use a sacrificial victim as a scapegoat. A more recent scholar who has stressed the importance of violence in sacrifice is Bloch (1992), but his approach differs from that of Girard in that for Bloch sacrifice requires violence because human beings need 'to create the transcendental in religion and politics,' and that this is attainable through violence done to the victim (Bloch, 1992: p. 7). One difficulty with this approach is that behavior that one might arguably class as sacrificial may involve no violence at all.

Origins of Sacrifice

Sacrificial rituals have an extensive pedigree. One of the earliest for which a reasonably detailed record is available dates from about 3000 years ago. The Vedic Aryans held sacrifice to be their central religious activity (Flood, 1996: pp. 40–41), and it consisted of putting milk, grains of rice and barley, and domesticated animals into a sacred fire that would transport the offerings to supernatural entities. As the Agnicayana, this ritual is still carried out in Kerala, southwest India.

Evidence for sacrificial behavior extends further back than Vedic times, to the civilizations of Sumer 5000 years ago and China 4000 years ago, and its presence seems accounted for even further in the past. This long history, taken with the widespread practice of sacrifice all over the world, suggests that a case could be made that sacrificial behavior is one of the natural dispositions common to the species *Homo sapiens*. As Needham (1985: p. 177) has averred for ritual in general, it may be that sacrifice, 'Considered in its most characteristic features … is a kind of activity – like speech or dancing – that man as a ceremonial animal happens naturally to perform.'

Bibliography

Benveniste, E., 1969. Le Vocabulaire des institutions indoeuropéennes, I. Economie, parenté, société; II. Pouvoir, droit, religion. Les Editions de Minuit, Paris.

Bloch, M., 1992. Prey into Hunter: The Politics of Religious Experience. Cambridge University Press, Cambridge, UK.

Durkheim, E., 1912/1968. Les Formes élémentaires de la vie religieuse: le système totémique en Australie. Presse Universitaires de France, Paris.

Evans-Pritchard, E.E., 1956. Nuer Religion. Clarendon Press, Oxford, UK.

Flood, G.O., 1996. An Introduction to Hinduism. Cambridge University Press, Cambridge, UK.

Frazer, J.G., 1922. The Golden Bough: A Study in Magic and Religion. vol. 1. Macmillan, London, UK.

Girard, R., 1972. La Violence et le sacré. Bernard Grasset, Paris.

de Heusch, L., 1986. Le Sacrifice dans les religions africaines. Editions Gallimard, Paris.

Hicks, D., 1996. Making the king divine: a case study in ritual regicide from Timor. Journal of the Royal Anthropological Institute 2, 611–624.

Hocart, A.M., 1970. Kings and Councillors: An Essay in the Comparative Anatomy of Human Society. University of Chicago Press, Chicago.

Hubert, H., Mauss, H., 1898. Essai sur la nature et la fonction du sacrifice. Année sociologique 2, 29–138.

Irimoto, T., 1996. Ainu worldview and bear hunting strategies. In: Pentikäinen, J. (Ed.), Shamanism and Northern Ecology (Religion and Society 36). Mouton de Gruyter, Berlin, pp. 293–303.

Muir, E., 1997. Ritual in Early Modern Europe. Cambridge University Press, Cambridge, UK.

Needham, R., 1985. Exemplars. University of California Press, Berkeley, CA.

Simpson, J.A., Weiner, E.S.C., 1933/1989. Oxford English Dictionary, second ed. Clarendon Press, Oxford, UK.

Smith, W.R., 1894/1927. Lectures on the Religion of the Semites, third ed. Macmillan, New York.

Tylor, E.B., 1871/1913. Primitive Culture, fourth ed. vol. 2. Murray, London.

Valeri, V., 1985. Kinship and Sacrifice: Ritual and Society in Ancient Hawaii (Wissing, P., Trans.). University of Chicago Press, Chicago.

Safety, Economics of

John F Burton, Jr., Cornell University, Ithaca, NY, USA; and Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The economics of safety examines workplace injuries. In neoclassical theory, employers pay a risk premium for hazardous work and invest in safety to reduce the premium. 'Modified neoclassical theory' argues that limited types of government regulation improve the operation of the labor market. The 'old institutional economists' emphasize the unequal bargaining power of individual workers. One strain of 'new institutional economics' asserts that, when transaction costs are significant, changing the legal rules about liability can affect the allocation of resources. The government mandate theory argues that government standards improve safety. Evidence supporting the various theories is evaluated.

Introduction

The economics of safety is concerned with the causes and prevention of accidents resulting in economic losses. The primary focus is on accidents causing workplace injuries and diseases. The topic is examined using several variants of economic theory. In 'pure' neoclassical economics, employers pay a risk premium to workers for hazardous jobs and invest in safety to reduce the premiums. There is evidence of risk premiums, but the amount may not fully compensate workers for their risks. Modified neoclassical economics recognizes that some types of government intervention are appropriate in order to overcome the attributes of the labor market that do not correspond to the assumptions of pure neoclassical economics. Old institutional economics (OIE) emphasize the unequal bargaining power of individual workers as another limitation of the labor market as envisioned by pure neoclassical economics. Workers' compensation is a no-fault program providing benefits to injured workers that is endorsed by critics of neoclassical economics because premiums are experienced rated – they vary among industries and firms based on prior benefit payments – which in theory provides economic incentives to employers to improve safety. However, recent studies have cast doubt on whether experience rating improves workplace safety. Consistent with OIE, evidence suggests that collective action by workers improves workplace safety.

New institutional economics (NIE) includes devotees of the Coase theory and transaction cost economics, who argue that when transactions costs (including the expenses of drafting and enforcing agreements) are high, changes in liability rules can affect the allocation of resources, including those devoted to safety. Evidence supporting this theory includes the improvement in workplace safety when the workers' compensation program replaced tort remedies in many states in the early twentieth century.

Law and economics (L&E) theory draws on neoclassical economics and transaction cost economics in the examination of legal institutions. One application considers the stimulus of tort law to safety: while theory postulates a positive effect of tort suits on safety, the evidence suggests these suits do not provide much assistance in deterring accidents.

Government mandate theory argues that government promulgation and enforcement of standards will improve workplace safety and health. However, several studies suggest that safety standards in the United States have done little to improve workplace safety and that, rather than overcoming deficiencies of an unregulated labor market, have often been excessively stringent.

'Pure' Neoclassical Economics

The neoclassical theory of work injuries is examined in Thomason and Burton (1993). Workplace accidents are undesirable by-products of processes that produce goods for consumption (Oi, 1974). Several assumptions are used in the neoclassical analysis of these processes. Employers and workers have an ongoing relationship. Employers maximize profits. Workers maximize utility, not just pecuniary income. Workers at the margin are mobile and possess accurate information concerning the risks and consequences of work injuries. One implication of these assumptions is that employers pay higher wages for hazardous work than for nonhazardous work. On the assumption that accident insurance is not provided by employers or the government, the equilibrium wage for the hazardous work will include a risk premium equal to the expected costs of the injuries, which includes lost wages, medical care, and the disutility caused by the injury. If workers are risk averse, the premium will also include a payment for the uncertainty about who will be injured.

The employer has an incentive to invest in safety in order to reduce accidents and thus the risk premium. The firm will make safety investments until the marginal expenditure on safety is equal to the marginal reduction in the risk premium. Since there is a rising marginal cost to investments in safety, equilibrium will occur with a positive value for the risk premium, which means that in equilibrium there will be some work injuries.

Evidence Consistent with the Neoclassical Model

There is an extensive literature on compensating wage differentials for the risks of workplace death and injury. Ehrenberg and Smith (2012) report that recent estimates suggest wages are about 1.0% higher for workers facing a fatality risk twice the average risk of job fatalities (about 1 per 25 000 workers

per year). The evidence on risk premiums can be used to estimate a value of statistical life (VSL), which represents evidence from the market about the willingness of people to pay for a reduction in the risk that they will be killed. Shapiro (2009) provides an example: if the average annual wage premium for the risk of work-related death is $400 and the annual fatality risk is 1 out of 10 000 workers, then the VSL is $4.0 million ($400 divided by 0.000 1).

Viscusi and Aldy (2003) reviewed more than 60 studies of mortality risk premiums from 10 countries and approximately 40 studies estimating injury risk premiums. Estimates using US labor market data typically show a VSL in the range of $4–$9 million (in 2000 dollars). Their review included 20 studies of non-US labor markets, including those of the United Kingdom (with significantly higher VSLs than in the United States), Canada (with the most VSL estimates in the range of US$3–US$6 million), and several developing countries, including India and Korea, in which VSL estimates are generally much lower than in developed countries. Across countries, the VSL generally increased with levels of income (income elasticities of about 0.5–0.6). Other general results include higher VSLs for union workers and decreasing VSLs for older workers. Leigh and Ruser (2003) found that compensating wage differentials for fatal and nonfatal injury risks vary according to gender and race.

Qualifications Concerning the Neoclassical Theory

The model that postulates workplace safety results from risk premiums, which provide financial incentives to employers to invest in safety, has been challenged. Some critics assert that risk premiums are inadequate because workers lack sufficient information about the risks of work injuries and/or have limited mobility to move to less hazardous jobs. However, Viscusi (1993) refers to a sizable literature in psychology and economics documenting that individuals tend to overestimate low probability events, such as workplace injuries, which could result in inappropriately large risk premiums. Ehrenberg and Smith (2012) conclude that workers have enough knowledge to form judgments about the relative risks of various jobs 'with some accuracy.' Both Ehrenberg and Smith (2012) and Viscusi (1993) argue there is sufficient worker mobility to generate risk premiums. Another challenger to the neoclassical model is Rose-Ackerman (1992), who argues that the labor market does not generate proper risk premiums because certain costs resulting from work injuries are not borne by workers but are externalized.

The empirical evidence on risk premiums has also been challenged or qualified. Ehrenberg and Smith (2012), for example, conclude that the studies of compensating wage differentials for the risks of injury or death on the job "generally, but not always, support the prediction that wages will be higher whenever risks on the job are higher." Viscusi (1993) concludes: "The wage-risk relationship is not as robust as is, for example, the effect of education on wages."

The most telling attack on the compensating wage differential evidence is by Dorman and Hagstrom (1998), who argue that most studies have been specified improperly. They argue that, after controlling for industry-level factors, the only evidence for positive compensating wage differential pertains to unionized workers. For nonunion workers, they argue that properly specified regressions suggest workers in dangerous jobs are likely to be paid less than equivalent workers in safer jobs.

The Dorman and Hagstrom attack on the previously generally accepted view – dangerous work results in risks premiums for most workers – was challenged by Viscusi and Aldy (2003). However, even if every empirical study finds risk premiums for workplace fatalities and injuries, the evidence would not conclusively validate the neoclassical economics approach. Ehrenberg (1988) provides the necessary qualification: if there are any market imperfections (such as lack of information or mobility), the "mere existence of *some* wage differential does *not* imply that it is a fully compensating one." And if the risk premium is not fully compensating (or is more than fully compensating), then inter alia, the market does not provide the proper incentive to employers to invest in safety.

Modified Neoclassical Economics and the Old Institutional Economics

Some economists rely on a 'modified' version of neoclassical economics theory that recognizes some types of government intervention are appropriate in order to overcome the attributes of the labor market that do not correspond to the assumptions of pure neoclassical economics. These attributes include the lack of sufficient knowledge and mobility by employees, and the possible lack of sufficient knowledge or motivation of employers about the relationship between expenditures on safety and the reduction in risk premiums. Proponents of the OIE would agree with these critiques of lack of knowledge and mobility, and would also emphasize factors such as the unequal bargaining power of individual workers as another limitation of the labor market as envisioned by neoclassical economists.

An example of government promulgation of information in order to overcome the lack of knowledge in the labor market is the Occupational Safety and Health Act (OSHAct) Hazard Communication standard, which requires labeling of hazardous substances and notification to workers and customers and which is estimated to save 200 lives per year (Viscusi, 1996). (Other aspects of the OSHAct are discussed below.)

Workers' Compensation

Workers' compensation programs represent another type of government intervention in the labor market in order to prevent workplace injuries and diseases, as well as to improve compensation for injured workers. Until late in the nineteenth century, the principal approaches used to achieve the goals of prevention and compensation were reliance on the risk premiums generated in the labor market and on negligence (tort) suits by injured workers against their employers (discussed below). Because of the dissatisfaction with these approaches, many countries enacted workers' compensation programs, beginning with Germany in 1884. Workers' compensation programs are a form of social insurance

providing cash benefits and medical care to disabled workers on a no-fault basis.

The enactment of workers' compensation programs result in moral hazard, which Harrington and Niehaus (1998) define as "the effect of insurance on the insured's incentives to reduce expected losses." Three types of moral hazard for workers resulting from the introduction of a workers' compensation program or from increases in workers' compensation benefits have been identified (Guo and Burton, 2010; Butler et al., 2013): (1) the true injury effect – workers may be less concerned about job safety and more willing to accept risk, which results in an increase in injury frequency and injury severity; (2) the reporting effect – workers will be encouraged to submit more claims for losses they otherwise would not have bothered to report; and (3) the duration effect – workers will extend the period for which they claim benefits.

Workers' compensation programs rely on several strategies to reduce moral hazard for workers, including (1) deductibles – workers do not receive benefits in the initial period after the workplace injuries occur because of waiting periods; (2) coinsurance – benefits only replace a portion of lost wages (generally 66 2/3% of wages in US state programs); (3) limits on eligibility for benefits – benefits are limited to those workers who meet legal tests included in workers' compensation statutes; and (4) limits on duration of benefits – some programs limit the duration of benefits, especially for long-term disabilities.

Moral hazard for employers also results from the introduction of a workers' compensation program or from increases in workers' compensation benefits, with the nature of the moral hazard depending on the method used to finance the program. Some workers' compensation programs are financed by premiums (or assessments) that are uniform for all employers. In contrast, the workers' compensation programs in the United States and to some degree in other countries, including Canada, France, and Germany, are financed by experience-rated premiums. The US variant of experience rating has two levels: industry-level experience rating, which largely bases the insurance rate for each industry on prior benefits in the industry, and firm-level experience rating, which determines the premium for each firm above a minimum size by comparing the firm's prior benefit payments to those of other firms in the industry. Analysts have identified three types of moral hazard for workers' compensation programs that relay on experience rating as opposed to programs financed by premiums that do not vary among employers (Guo and Burton, 2010; Butler et al., 2013): (1) the safety effect – employers will have more incentive to keep the workplace safe; (2) the underreporting or monitoring effect – employers will have more incentive to deny claims, which may be fraudulent, questionable, or legitimate; and (3) the rehabilitation or return-to-work effect – employers will have more incentive to adopt claims management practices that reduce the duration of benefit payments.

Two studies of US workers' compensation programs examined whether the moral hazard (or incentive) effects for workers or the incentive effects for employers dominate. Bronchetti and McInerney (2012) used a national data set on individuals in the United States and found that higher benefits resulted in fewer workers' compensation claims (the benefits–claims elasticities were less than 1.0). Guo and Burton (2010) examined the determinants of interstate differences in the Bureau of Labor Statistics (BLS) injury rate for two periods (1975–89 and 1990–99). The coefficients on expected benefits were statistically insignificant in both periods. The authors indicated that "One interpretation of these results is that true injury effect is offset by the safety effect." Guo and Burton also examined the determinants of interstate differences in workers' compensation cash benefits per 100 000 workers. The authors found that "the benefit elasticity (the association between expected benefits and actual benefit payments) was significantly less than 1.0 in both our study periods (1975–89 and 1990–99)." One interpretation of this result is that the monitoring and rehabilitation effects for employers were stronger than the reporting and duration effects for workers.

The Effects of Experience Rating according to Different Economic Theories

Experience rating has received considerable attention in recent decades. In the neoclassical economics model, the introduction of workers' compensation with experience rating should make no difference in the safety incentives for employers compared to the incentives provided by the labor market without workers' compensation. Under assumptions explicated by Burton and Chelius (1997), such as perfect experience rating, risk neutrality by workers, and actuarially fair workers' compensation premiums, the risk premium portion of the wage paid by the employer will be reduced by an amount exactly equal to the amount of the workers' compensation premium. Also, under these assumptions, the employer has the same economic incentives to invest in safety both before and after the introduction of the workers' compensation program. Under an alternative variant of the pure neoclassical economics approach, in which the assumption of perfect experience rating is dropped, the introduction of workers' compensation will result in reduced incentives for employers to reduce accidents.

In contrast, the OIE approach argues that the introduction of workers' compensation with experience rating should improve safety because the limitations of knowledge and mobility and the unequal bargaining power for employees mean that the risk premiums generated in the labor market are inadequate to provide employers the safety incentives postulated by the pure neoclassical economics approach. Commons (1934), a leading figure in the OIE approach, claimed that unemployment is the leading cause of labor market problems, including injuries and fatalities, because slack labor markets undercut the mechanism that generates compensating wage differentials. Commons asserted that experience rating provides employers economic incentives to get the 'safety spirit' that would otherwise be lacking. Many modified neoclassical economists also accept the idea that experience rating should help improve safety by providing stronger incentives to employers to avoid accidents, although they place less emphasis on the role of unemployment in undercutting compensating wage differentials and more emphasis on the failure of employers to recognize the cost savings possible from improved safety without the clear signals provided by experience-rated premiums.

Evidence on the Effects of Experience Rating

A number of studies of experience rating provide evidence that should help assess the virtues of the various economic theories. Almost without exception, the studies find that experience rating reduces the number of workers' compensation claims, which supports the argument that 'pure' neoclassical economics does not provide a sufficient explanation of the operation of the labor market. What is unclear, however, is whether the reduction in claims is a result of the safety effect – with fewer actual injuries – or is due to the monitoring effect – with employers denying more claims. The evidence on this issue is inconclusive. The survey of experience-rating studies by Boden (1995) concluded that "research on the safety impacts has not provided a clear answer to whether workers' compensation improves workplace safety." In contrast, Thomason (2005) asserted that most (11 of 14) studies he surveyed found that experience rating improves safety and health and concluded: "Taken as a whole, the evidence is quite compelling: experience rating works." A markedly different conclusion was reached by Mansfield et al. (2012), who concluded that:

> Although experience rating is intended to stimulate safer workplaces, a growing body of literature reveals that it has not achieved that effect . . . The absence of a safety effect may arise because employers focus on managing reported claims rather than prevention.

Mansfield et al. (2012) is contained in a collection of articles that provide largely negative assessments of experience rating. For example, Tompa et al. (2012) conclude their study of Ontario "suggests that experience rating provides an incentive for secondary prevention, but less so for primary prevention." There are also largely skeptical assessments of experience rating in Australia (Clayton, 2012; Purse, 2012) and in New Zealand (Armstrong and Bunn, 2012). The only 'good news' for experience rating was provided by Seabury et al. (2012), who found that workers injured at self-insuring employers (who are perfectly experience rated) have significantly improved return-to-work experience for up to 5 years after the date of injury.

The recent evidence that experience rating does not improve safety is inconsistent with the positive impact on safety postulated by the OIE approach and is also contrary to the views of some proponents of modified neoclassical economics. Additional studies of experience rating can be anticipated, as can disagreements about the wisdom of relying on this approach to promoting safety, in part because of recent studies demonstrating that many workers with disabilities caused by work do not receive workers' compensation benefits (Spieler and Burton, 2012).

Collective Action

While some modified neoclassical economists and the OIE largely agree on the desirability of several prevention approaches, such as the use of experience rating in workers' compensation, the OIE endorsed another approach that many modified neoclassical economists would not support, namely, the use of collective bargaining and other policies that empower workers. Collective bargaining agreements can minimize unsafe activities or at least explicitly require employers to pay a wage premium for unsafe work. Many agreements also establish safety committees that assume responsibility for certain activities, such as participation in inspections conducted by government officials. If workers are injured, unions can help them obtain workers' compensation benefits, thereby increasing the financial incentives for employers to improve workplace safety.

The beneficial effects predicted by the OIE approach for these collective efforts appear to be achieved. Several studies, including Weil (1999) concluded that Occupational Safety and Health Administration (OSHA) enforcement activity was greater in unionized firms than in nonunionized firms. Moore and Viscusi (1990) and other researchers found that unionized workers receive larger compensating wage differentials for job risks than unorganized workers. Hirsch et al. (1997) found that "unionized workers were substantially more likely to receive workers' compensation benefits than were similar nonunion workers."

There are also safety committees in some jurisdictions that are utilized in nonunionized firms. Burton and Chelius (1997) reviewed several studies involving such committees in the United States and Canada, and found limited empirical support for their beneficial effects. Reilly et al. (1995) provided a more positive assessment for their accomplishments in the United Kingdom. These studies provide some additional support for the beneficial role of collective action postulated by the OIE.

The New Institutional Economics

NIE authors generally argue that market forces encourage efficient forms of economic organization without government assistance and that opportunities for efficiency-improving public interventions are rare. This section considers only the Coase theory/transaction costs economics strain of NIE.

The Coase Theory and Transaction Costs Economics

In the absence of costs involved in carrying out market transactions, changing the legal rules about who is liable for damages resulting from an accident will not affect decisions involving expenditures of resources that increase the combined wealth of the parties. The classic example offered by Coase (1988) involves the case of straying cattle that can destroy crops growing on neighboring land. The parties will negotiate the best solution to the size of the herd, the construction of a fence, and the amount of crop loss due to the cattle whether or not the cattle rancher is assigned liability for the crop damages.

Coase recognized that the assumption that there were no costs involved in carrying out transactions was 'very unrealistic.' According to Coase (1988):

> In order to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with … and on what terms, to conduct negotiations …, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed, and so on. These operations are often … sufficiently costly … to prevent many transactions that would be carried out in a world in which the pricing system worked without cost.

The goal of transactions costs economics is to examine these costs and to determine their effect on the operation of the economy. As scholars of transaction costs have demonstrated, when transactions costs are significant, changing the legal rules about initial liability can affect the allocation of resources.

Evidence Concerning Changes in Liability Rules

Workplace safety regulation provides a good example of a change in liability rules, since in a relatively short period (1910–20), most US states replaced tort suits (the employer was only responsible for damages if negligent and employers had extraordinary defenses even if negligent) with employer liability acts (the employer was required to be negligent but some or all of the extraordinary defenses were abrogated), which in turn were replaced with workers' compensation (the employer is required to provide benefits under a no-fault rule) as the basic remedy for workplace injuries. Fishback (1998) and Butler and Worrall (2008) provide excellent reviews of the empirical studies of these changes in liability rules, which are summarized in Burton (2009).

Studies differ on whether the changes in liability rules increased or decreased workplace accidents, depending on the workers' occupation and industry. The overall conclusion of Butler and Worrall (2008) was that "those who argue that the tort system generated better safety incentives than workers' compensation are wrong." This conclusion suggests that high transaction costs associated with the determination of fault in negligence suits were an obstacle to achieving the proper incentives to workplace safety, and that the institutional features of workers' compensation, including the no-fault principle, provided a relatively more efficient approach to the prevention of work injuries than tort suits for most workers. The transaction economics component of the NIE theory thus appears to provide a useful supplement to neoclassical economics, since institutional features, such as liability rules, can have a major impact on the economic incentives for accident prevention.

Law and Economics

L&E theory draws on neoclassical economics and transaction cost economics, but is distinctive in the extent to which it examines legal institutions and legal problems. This section examines the tort law branch of L&E theory, while the next section examines the employment law branch.

Theoretical Stimulus of Tort Law to Safety

Tort law typically is used when one party harms another party, and the parties do not have an ongoing relationship. However, even though workers and employers have a continuing relationship, tort suits were generally used as a remedy for workplace accidents in the United States until workers' compensation programs were established. When negligence is the legal standard used for tort suits, if the employer has not taken proper measures to prevent accidents and thus is at fault, the employer will be liable for all of the consequences of the injury. The standard for the proper prevention measure was developed by Judge Learned Hand and restated by Posner (1972) as:

> The judge (or jury) should attempt to measure three things in ascertaining negligence: the magnitude of the loss if an accident occurs; the probability of the accident's occurring; and the burden (cost) of taking precautions to prevent it. If the product of the first two terms, the expected benefits, exceeds the burden of precautions, the failure to take those precautions is negligence.

Posner argued that proper application of this standard would result in economically efficient incentives to avoid accidents. Burton (2009) examines some qualifications to this conclusion.

Evidence on the Tort Law Stimulus to Safety

Two types of empirical evidence suggest that skepticism is warranted about the stimulus to workplace safety from tort suits. First, tort suits were used as the remedy for workplace injuries in the late 1800s and early 1900s. As previously discussed, most studies find that replacement of the negligence remedy with workers' compensation led to a general reduction in workplace fatalities.

Second, in other areas of tort law, there is a major controversy among legal scholars about whether the theoretical incentives for safety resulting from tort suits actually work. One school of thought is exemplified by Landes and Posner (1987), who state that "although there has been little systematic study of the deterrent effect of tort law, what empirical evidence there is indicates that tort law … deters." An opposing view of the deterrent effects of tort law is provided by Priest (1991), who finds almost no relationship between actual liability payouts and the accident rate for general aviation and states that "this relationship between liability payouts and accidents appears typical of other areas of modern tort law as well, such as medical malpractice and product liability.'

A study of tort law by Schwartz (1994) distinguished a strong form of deterrence (as postulated by Landes and Posner) from a moderate form of deterrence, in which "tort law provides a significant amount of deterrence, yet considerably less that the economists' formulae tend to predict." Schwartz surveyed a variety of areas where tort law is used, including motorist liability, medical malpractice, and product liability, and concluded that the evidence undermines the strong form of deterrence but provides adequate support for the moderate form of deterrence. Dewees et al. (1996), after reviewing the evidence from several fields involving personal injury, including accidents involving automobiles, medical practice, and products, concluded that "the tort system performs unevenly in deterring the causes of personal injury, so its scope should be restricted to situations where its effect seems likely to justify its high cost." They recommended expanding the no-fault approach exemplified by workers' compensation to the motor vehicle and medical malpractice fields.

Burton and Chelius (1997) concluded and Burton (2009) affirmed that based on both the historical experience of the impact of tort suits on workplace safety and the current controversy over the deterrence effect in other areas of tort law, "the law and economics theory concerning tort law does not

provide much assistance in designing an optimal policy for workplace safety and health."

Government Mandate Theory vs Law and Economics Theory

The Government Mandate Theory

The government mandate theory argues that the government promulgation of health and safety standards and enforcement of the standards by inspections and fines will improve workplace safety and health. This is basically a legal theory, although many of the supporting arguments involve reinterpretation or rejection of studies conducted by economists. For example, proponents of the theory object to the evidence on compensating wage differentials and object in principle to economists' reliance on cost-benefit analysis. Supporters of the theory, such as McGarity and Shapiro (1996), also provide a positive case why OSHA is necessary:

> OSHA's capacity to write safety and health regulations is not bounded by any individual worker's limited financial resources. Likewise, OSHA's capacity to stimulate an employer to action does not depend upon the employees' knowledge of occupation risks or bargaining power.

The government mandate theory would not be endorsed by many economists, including the OIE. Commons and Andrews (1936), for example, criticized at length the punitive approaches that used factory inspectors in the form of policemen, since this turned employers into adversaries with the law. The minimum standards supported by the OIE were those developed by a tripartite commission, involving employers, employees, and the public, rather than standards promulgated by the government. While the OIE theory is, thus, unsympathetic to the government mandate theory, the sharpest attack is derived from the L&E theory.

The L&E Theory Concerning Government Regulations

L&E scholars make a distinction between mandatory, minimum terms (standards) and those terms that are merely default provisions (or guidelines) that employers and employees can agree to override. Most employment laws, including workplace safety laws, create standards and are thus objectionable to the L&E scholars.

Willborn (1988) articulated the standard economic objections to mandatory terms. Employers will treat newly imposed standards like exogenous wage increases and in the short run will respond by laying off workers. In the long run, employers will try to respond to mandates by lowering the wage. The final wage–benefits–standards employment package will make workers worse off than they were before the imposition of standards – otherwise the employers and workers would have bargaining for the package without legal compulsion.

Evidence on the Effects of OSHA Standards

Shapiro (2009) concluded that the studies of OSHA's activities to improve workplace safety produced inconsistent results. For example, Viscusi (1992) found that OSHA inspection activity reduced the lost workday rate by 1.5–3.6% at the industry level, but he indicated that other studies at this level of aggregation found no such effect. There were also conflicting results from studies at the individual-plant level. For example, Gray and Scholz (1991) found that inspections imposing a penalty reduced injuries, while Ruser and Smith (1991) found no such effect. On particular interest, Gray and Mendeloff (2005) found a decline over time in the effect of OHSA inspections on injury rates, with insignificant reductions in the most recent period in their study (1987–91).

Overall, the evidence suggests that the Occupational Safety and Health Act (OSHAct) has done little to improve workplace safety, thus lending more support to the L&E theory than to the government mandate theory. OSHA's ineffectiveness in part may be due to the lack of inspection activity, since Weil and Pyles (2005) note that the annual probability a US workplace will be inspected is well below 0.001. But the evidence also suggests that allocating additional resources to plant inspections may be imprudent, given the mixed results of inspections on safety.

In addition to the questionable effectiveness of OSHA inspections, some standards promulgated by OSHA have been criticized as excessively stringent. Viscusi (1996) examined OSHA standards using an implicit value of life of $5 million (derived from the compensating wage differential studies) as the standard for an efficient regulation. Four of the five OSHA safety regulations, but only one of the five OSHA health regulations adopted as final rules, had costs per life saved of less than $5 million. This evidence caused Burton and Chelius (1997) to provide a strong critique of the government mandate theory:

> To be sure, cost-benefit analysis of health standards issued under the OSHAct is not legal, and so those standards that fail the cost-benefit test (considering both the lives saved plus injuries and illnesses avoided) do not violate the letter and presumably the purpose of the law. But to the extent that the rationale offered by the government mandate theorists for regulation of health is that workers lack enough information to make correct decisions and therefore the government is in a better position to make decisions about how to improve workplace health, the evidence on the variability of the cost/benefit ratios for OSHA health standards is disquieting. Rather than OSHA standards reflecting interventions in the marketplace that overcome deficiencies of the marketplace, the explanation of why the stringency of regulation varies so much among industries would appear at best to be a result of technology-based decisions that could well aggravate the alleged misallocation of resources resulting from operation of the market and at worst could reflect relative political power of the workers and employers in various industries.

Conclusions

This article suggests that understanding the economics of workplace safety involves a rather eclectic mix of theories. Burton and Chelius (1997) were least impressed with the arguments and evidence pertaining to the pure neoclassical economics and the government mandate theories. They concluded that, among the other economic theories pertaining to safety, no single theory provides an adequate understanding of the causes and prevention of workplace accidents. Rather,

a combination of the theories, though untidy, is needed. Their conclusion has stood the test of time.

See also: Environmental Health and Safety: Social Aspects; Institutionalism and Neo-institutionalism: History of the Concepts; Occupational Health and Safety, Regulation of; Old and New Institutionalism in Economics; Regulation, Political Economy of; Value of Life, Economics of; Workplace Safety and Health.

Bibliography

Armstrong, Hazel, Bunn, Kristen, 2012. Experience rating and occupational diseases: a New Zealand case study. Policy and Practice in Health and Safety 1, 63–75.

Boden, Leslie I., 1995. Creating economic incentives: lessons from workers' compensation systems. In: Voos, Paula (Ed.), Proceedings of the 47th Annual Meeting. Industrial Relations Research Association, Madison, WI, pp. 282–292.

Bronchetti, Erin T., McInerney, Melissa, 2012. Revisiting incentive effects in workers' compensation: do higher benefits really induce more claims? Industrial and Labor Relations Review 65, 286–315.

Burton Jr., John F., Chelius, James R., 1997. Workplace safety and health regulations: rationale and results. In: Kaufman, Bruce (Ed.), Government Regulation of the Employment Relationship. Industrial Relations Research Association, Madison, WI, pp. 253–293.

Burton Jr., John F., 2009. Workers' compensation. In: Dau-Schmidt Kenneth, G., Harris, Seth D., Lobel, Orley (Eds.), Labor and Employment Law and Economics. Edward Elgar, Cheltenham, UK, pp. 235–274.

Butler, Richard J., Worrall, John D., 2008. Wage and injury response to shifts in workplace liability. Industrial and Labor Relations Review 61, 181–200.

Butler, Richard J., Gardner, Harold H., Kleinman, Nathan L., 2013. Workers' compensation: occupational injury insurance's influence on the workplace. In: Dionne, Georges (Ed.), Handbook of Insurance, second ed. Springer, New York.

Clayton, Alan, 2012. Economic incentives in the prevention and compensation of work injury and illness. Policy and Practice in Health and Safety 1, 27–43.

Coase, Ronald H., 1988. The Firm, the Market, and the Law. University of Chicago Press, Chicago, IL.

Commons, John R., 1934. Institutional Economics: Its Place in Political Economy. MacMillan, New York.

Commons, John R., Andrews, John B., 1936. Principles of Labor Legislation, fourth ed. MacMillan, New York.

Dewees, Don, Duff, David, Trebilcock, Michael, 1996. Exploring the Domain of Accident Law: Taking the Facts Seriously. Oxford University Press, Oxford.

Dorman, Peter, Hagstrom, Paul, 1998. Wage compensation for dangerous work revisited. Industrial and Labor Relations Review 52, 116–135.

Ehrenberg, Ronald G., 1988. Workers' compensation, wages, and the risk of injury. In: Burton Jr., John F. (Ed.), New Perspectives on Workers' Compensation. ILR Press, Ithaca, NY, pp. 71–96.

Ehrenberg, Ronald G., Smith, Robert S., 2012. Modern Labor Economics: Theory and Public Policy, twelfth ed. Addison-Wesley, Reading, MA.

Fishback, Price V., 1998. Operations of "unfettered" labor markets: exit and voice in American labor markets at the turn of the century. Journal of Economic Literature 36, 722–765.

Gray, Wayne B., Scholz, John T., 1991. Does regulatory enforcement work: a panel analysis of OSHA enforcement. Law and Society Review 27, 177–213.

Gray, Wayne B., Mendeloff, John, 2005. The declining effects of OSHA's inspections on manufacturing industries 1979–1998. Industrial and Labor Relations Review 58, 571–587.

Guo, Xuguang, Burton Jr., John F., 2010. Workers' compensation: recent developments in moral hazard and benefits payments. Industrial and Labor Relations Review 63, 340–355.

Harrington, Scott E., Niehaus, Gregory R., 1998. Risk Management and Insurance. McGraw-Hill, Columbus, OH.

Hirsch, Barry T., Macpherson, David A., Dumond, J. Michael , 1997. Workers' compensation recipiency in union and nonunion workplaces. Industrial and Labor Relations Review 50, 213–236.

Landes, William M., Posner, Richard A., 1987. The Economic Structure of Tort Law. Harvard University Press, Cambridge, MA.

Leigh, J. Paul, Ruser, John, 2003. Compensating wage differentials for fatal and nonfatal injury risk by gender and race. Journal of Risk and Insurance 27, 257–277.

Mansfield, Liz, MacEachen, Ellen, Tompa, Emile, Kalcevich, Christina, Endicott, Marion, Mi, Natalie Y., 2012. A critical review of literature on experience rating in workers' compensation systems. Policy and Practice in Health and Safety 1, 3–25.

McGarity, Thomas O., Shapiro, Sidney A., 1996. OSHA's critics and regulatory reform. Wake Forest Legal Review 31, 587–646.

Moore, Michael J., Viscusi, W. Kip, 1990. Compensation Mechanisms for Job Risks: Wages, Workers' Compensation, and Product Liability. Princeton University Press, Princeton, NJ.

Oi, Walter Y., 1974. On the economics of industrial safety. Law and Contemporary Problems 38, 669–699.

Posner, Richard, 1972. A theory of negligence. Journal of Legal Studies 1, 29–66.

Priest, George L., 1991. The modern expansion of tort liability: its sources, its effects, and its reform. Journal of Economic Perspectives 5, 31–50.

Purse, Kevin, 2012. Experience rating: an Australian post mortem. Policy and Practice in Health and Safety 1, 46–61.

Reilly, B., Paci, P., Holl, P., 1995. Unions, safety committees and workplace injuries. British Journal of Industrial Relations 33, 275–288.

Rose-Ackerman, Susan, 1992. Rethinking the Progressive Agenda: The Reform of the American Regulatory State. Free Press, New York.

Ruser, John, Smith, Robert S., 1991. Reestimating OSHA's effects: have the data changed? Journal of Human Resources 26, 212–236.

Schwartz, Gary T., 1994. Reality in the economic analysis of tort law: does tort law really deter? UCLA Law Review 42, 377–444.

Seabury, Seth A., McLaren, Christopher F., Reville, Robert, Neuhauser, Frank, Mendeloff, John, 2012. Workers' compensation experience rating and return to work. Policy and Practice in Health and Safety 1, 97–115.

Shapiro, Sidney A., 2009. Occupational health and safety regulation. In: Dau-Schmidt, Kenneth G., Harris, Seth D., Lobel, Orley (Eds.), Labor and Employment Law and Economics. Edward Elgar, Cheltenham, UK, pp. 273–295.

Spieler, Emily A., Burton Jr., John F., 2012. The lack of correspondence between work-related disability and receipt of workers' compensation benefits. American Journal of Industrial Medicine 55, 487–505.

Thomason, Terry, Burton Jr., John F., 1993. Economic effects of workers' compensation in the United States: private insurance and the administration of compensation claims. Journal of Labor Economics 11 (Part 2), S1–S37.

Thomason, Terry, 2005. Economic incentives and workplace safety. In: Roberts, Karen, Burton Jr., F. John, Bodah, Matthew (Eds.), Workplace Injuries and Diseases: Prevention and Compensation. W.E. Upjohn Institute for Employment Research, Kalamazoo, MI, pp. 9–35.

Tompa, Emile, Hogg-Johnson, Shelia, Amick, Benjamin C., Wang, Ying, Shen, Enquig, Mustard, Cam, Robson, Lynda, 2012. Financial incentives in workers' compensation: an analysis of the experience-rating programme in Ontario, Canada. Policy and Practice in Health and Safety 1, 117–137.

Viscusi, W. Kip, 1992. Fatal Trade-offs: Public and Private Responsibilities for Risk. Oxford University Press, New York.

Viscusi, W. Kip, 1993. The value of risks to life and health. Journal of Economic Literature 31, 1912–1946.

Viscusi, W. Kip, 1996. Economic foundations of the current regulatory reform efforts. Journal of Economic Perspectives 10, 119–134.

Viscusi, W. Kip, Aldy, Joseph E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. Journal of Risk and Uncertainty 27, 5–76.

Weil, David, 1999. Are mandated health and safety committees substitutes for or supplements to labor unions? Industrial and Labor Relations Review 52, 339–360.

Weil, David, Pyles, Amanda, 2005. Why complain? complaints, compliance and the problem of enforcement in the U.S. workforce. Comparative Labor Law and Policy Journal 27, 59–92.

Willborn, Steven L., 1988. Individual employment rights and the standard economic objections: theory and empiricism. Nebraska Legal Review 67, 101–139.

Same-Sex Sexualities in Turkey

Cenk Özbay, Bogazici Universitesi, Bebek, Istanbul, Turkey

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the background and the current situation of same-sex sexualities in Turkey. First, a short recent history of homosexuality in the nation and the ways it has been represented are explained. Then, the most significant factors that reshape contemporary Turkish same-sex sexualities, namely, Islam, social class, and the role of the state are expounded. Finally, some of the emergent problems and issues that gay, lesbian, bisexual, transgender, and queer people in Turkey have to deal with are contrasted to their persistence at seeking respect.

A Short History of Same-Sex Sexualities in Turkey

Following a social constructionist perspective on same-sex sexualities and homosexual identities throughout modern history (D'Emilio, 1998; Nardi and Schneider, 1998; Plummer, 1992; Seidman, 2009; Weeks, 1990), one can discern an ambiguous time in Turkish history when there were no sexual self-identities or medical diagnosis of sexual 'abnormalities,' but same-sex sexual acts took place and virtually everybody had knowledge about them. This undocumented and thus ambiguous period starts from the late Ottoman Empire and continues for 100 years or more into the temporal realm of modern Turkish Republic, which was founded in 1923. The new republic radically transfigured almost all spheres of social, cultural, and political public life via various developmentalist projects including Europeanization, secularization, westernization, and modernization. Not only the empire was abolished and a new nation with the nation-state was created but also a new, modern subject-citizen was formed under the influence of Enlightenment ideals by state discourses, policies, and practices. In spite of this grand transformation, same-sex sexual acts, or the type of 'the homosexual,' were not explicitly monitored, regulated, named, or criminalized by the new regime. Although it worked meticulously on the structure of the nuclear family, the role of women in public and private spheres, and the gendered politics of representation (Arat, 1997; Sirman, 2005), the modern revolution in Turkey was seemingly too busy to deal with, cure, ban, or intervene into homosexuality. Same-sex sexual activities became a significant part of the abject, invisible yet connived urban underground culture. Contemporary scholars can trace this hazy time period through local and foreign historians' work as well as novels and other literary forms. Among others, the great Turkish author Kemal Tahir's numerous novels and stories and the social historian Resat Ekrem Kocu's unfinished yet marvelous *Encyclopaedia of Istanbul* (1946) and his other books can reveal this unnamed yet fully experienced sexual riptide.

After a relatively insipid period in terms of homosexuality in the 1950s, the famous singer Zeki Muren became the first queer public figure in the history of the country in 1960s. In the beginning of his career, he was enacting a kind of alternative masculinity that contradicted with the hegemonic displays of manhood of its time. Later he started to destabilize gender and sexual norms more manifestly, and continued that until his death in 1991. People loved him as if he was a national hero and called him 'pasha,' a title used for the very masculine military leaders, despite the fact that he was not related emotionally or sexually to women and performed at stage in women's clothes with heavy makeup on his face. For most of his lifetime, he was an open transvestite and clandestine homosexual, a well-respected queer king (Özbay, 2009).

Despite his exceptional queer public display, Muren was not alone in enjoying a homosexual lifestyle in Turkey. The Beyoglu district of Istanbul has long been the focal point of queer lives in Turkey (Oz, 2009; Yuzgun, 1986, 1993). The country's first popular 'gay bar' Vat-69 (opening year 1975) and almost all its successors have been located in Beyoglu and vicinity. By attending to these bars as well as specific public beaches, parks, and Turkish baths at certain times, a new, distinct, modern homosexual person, who is located within a web of same-sex sexual relations instead of conventional family ties and collegial sociabilities, started to emerge in Istanbul after 1960s.

Another crucial figure in this particular history is also a singer, Bulent Ersoy. After becoming immensely popular in the late 1970s as an apparently homosexual man, in 1981 she became transsexual through a sex reassignment surgery. Although Muren's subverting queer performance was never oppressed by the state, Ersoy was legally banned to perform at stage by the postcoup military government. She was exiled to Germany until 1988 when the new, liberal government put an end to the prohibition. In contrast to Muren's highly adaptable, almost officially sanctioned transvestism, her radical transgenderism with the narratives about the surgeries and the bodily transformation, and the political conflict with the military power, made Ersoy a trenchant symbol of gender and sexual struggle that is still effective in the 'structure of feeling' about homosexuality or same-sex intimacies in Turkey (Yuzgun, 1986; Hocaoglu, 2002). For homosexuals, the 1980s was also the decade full of brutal police patrols, torture in police stations, harassment in public spaces and bars, and the new stigma that came with the first human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS) cases.

By the late 1980s, the incipient neoliberalization of the economy, further integration to the Western world, and the strong tides of globalization paralleled public Turkish culture becoming gradually demilitarized. Along with these changes,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10219-3

modern gay men emerged in major Turkish cities. They tended to imitate a Euro-American style of gayness, embodied body-oriented gay masculinity, and declared a symbolic war against the feminine public image of homosexuality, which more or less stemmed from the popular transgender singers (Hocaoglu, 2002; Tapinç, 1992). A more surreptitious but better organized urban gay culture flourished in this period. Men who have sex with men started to identify themselves as 'gay' as in English, having sex with other gay men, having emotional and long-term relationships, desiring and emulating a straight-looking bodily demeanor, and coming together not only for purposes of sex and joy but also for political commitment and sexual activism against homophobia. The first commercial movies regarding the issues of same-sex sexuality were produced in the Turkish film industry in the 1980s (Oztek, 2007) and homosexual stories started to reappear in Turkish novels, after a long period of silence. Representations of homosexuality in film and novel increased steadily in the 1990s and notable examples of queer art became well known such as gay authors Murathan Mungan's and Selim Ileri's books and Ferzan Ozpetek's critically acclaimed first film Steam-The Turkish Bath (1997).

Gay life in the recent history of modern Turkey reached its peak in the early 2000s, followed by its eventual decline in terms of its visibility and diversity in social and physical spaces in the metropolitan areas. Numerous gay (and to a certain extent lesbian) cafes, bars, and clubs opened during those years in Istanbul, the businesses were full with hundreds of avid homosexual people, the popular press and the news media had quite positive coverage about the glittering gay bars and their clientele, academic publications and research accounts were beginning to be published, sexual activists became more discernable and respected, and the confusion that masses had about the meaning of the concepts gayness and transgenderism vanished a little bit (Hocaoglu, 2002; Bereket and Adam, 2006). In addition to the self-assured gay men, lesbians became for the first time visible and came together as a separate social group to discuss their own problems that might be different than gay men's (Basdas, 2007; Özbay and Soydan, 2003; Savci, 2011).

In this very formation of modern gay and lesbian social identity in Turkey, three different developments played important roles. The first one was the gradual decrease and the eventual end of police raids and violence toward homosexuals and transgenders (Oz, 2009; Yuzgun, 1993). In this sense, it was deemed a little bit more acceptable and feasible to open and manage gay commercial venues, especially in metropolitan areas. Visitors in these queer businesses did not have fear of being exposed to the police when they were socializing. The second point was about the change in the attitude of the newspapers and the news media. Broadly speaking, the marginalizing, otherizing, and even dehumanizing language was dropped off and gays and lesbians started to appear in the pages through their own agendas and words within a framework of tolerance and esteem (Savci, 2011). A third cornerstone in this process was the diffusion of the Internet. It enabled gay men and lesbians to become connected to each other, social organizations and groups, and the world, without being revealed in front of the heterosexual society or having to come out of the closet unwillingly and uncontrollably (Hocaoglu, 2002; Özbay and Soydan, 2003). In the beginning, European dating Web sites brought large numbers of people together online and then Turkish Web sites were founded and replaced the European ones, became even more popular in urban zones and in the provincial towns. Maybe for the first time in the modern history, being gay and engaging into same-sex sexual activities had no social costs and consequences in Turkey.

Current Situation

Today, Turkey is one of the few countries in which homosexuality or counternormative sexualities are legal yet they are subjugated by state institutions, including the restrictive actions of police forces and public prosecutors (Oz, 2009; Savci, 2011). A simultaneous double life is experienced in terms of same-sex sexualities. On the one hand, the Turkish state and society are becoming more conservative, religious, and oppressive, in the sense that they forbid not only certain types of sexualities but sexuality itself and all its public manifestations are being constructed as taboo. On the other hand, same-sex sexualities are performed as they are tolerated within the zones of exception, especially in certain neighborhoods of the major metropolitan areas. These urban areas, under strict surveillance, provide social and physical spaces in which queer citizens engage same-sex sexualities and experience gay sociabilities while the state authorities are able to watch and govern the flow of homosexuals outside those areas. In this double configuration of same-sex sexual cultures, tolerance and intolerance, respect and intervention, freedom and restriction, grassroots diversity and superimposed uniformity, amalgamate by the participation of the state, the heterosexual public and gay men and lesbians as condoned exceptions (Basdas, 2007; Bereket and Adam, 2006, 2008; Özbay, 2010; Özyegin, 2012).

Subjects who have same-sex inclinations promulgate their own characteristics of flexibility, versatility, multifariousness, adaptability, discretion, self-centeredness, and even a precarious insincerity, in order to navigate within this compelling binary social structure that oscillates between a globalizing tendency for multiculturalism and an imagined insular isle of normality and undemocratic homogeneity (Özbay, 2010). The first step in this process is the formation of modern, Western, Euro-American gay and lesbian identities. These people construct a self-image that does not strive to hide their sexual identities and actions; on the contrary, as the Western role models suggest, coming out strategies and narratives play a crucial role in constituting who they really are and to what extent they internalize gay identities (Bereket and Adam, 2006; Hocaoglu, 2002; Özbay and Soydan, 2003; Özyegin, 2012). In this Westernized view of sexuality, there is a neat border between heterosexual and homosexual affect, culture, eroticism, and identity. A person is either gay or straight and the definitions are exact and stabilized.

There are factors complicating this supposedly neat, yet never neutral, division between what is homosexuality and what is heterosexuality. One of the most significant of them is the capacities of Islam. Islamic doctrines clearly prohibit same-sex sexual acts. Some of the most popular religious discourses and interpretations, as well as the religious public, openly deny

and ostracize gay men and lesbians as sinful people. This otherizing, robust antigay position is visible at least for the Sunni majority, while for the Alevi minority of Islam the attitude toward homosexuality is rather vague but supposedly more democratic and flexible. As Islam is not singular and unified at almost any subject, it has also multiple perspectives and principles about homosexuality. Despite the lack of elaborate discussions about religion and homosexuality as in the case of Christianity, homosexual citizens in Turkey have to deal with religion in their everyday lives and, to a certain extent, they have to find a satisfying answer to the inescapable question of whether they had faith (Bereket and Adam, 2008). Mostly, gay men and lesbians reconstruct themselves as entirely outside of the religious domain, believing in God but disconnecting from religious practices and developing a tacit silence, while some of them unmistakably reject religious and belief systems (Hocaoglu, 2002; Özbay and Soydan, 2003).

Another critical factor that shapes the current homosexual mise-en-scène in Turkey is class, which is not entirely independent from the social organization of religion in Turkish society. As it signifies a modern, global, refined, European, elite person whose cultural capital is sufficiently high, the term gay has a certain middle and upper class connotation in the Turkish vernacular. In other words, being gay in the Turkish context is not only about erotic subjectivity and sexual acts, but always a matter of social class (Armbrecht, 2001; Hocaoglu, 2002; Özbay, 2010). According to this classed understanding of sexuality, a person's erotic and intimate escapades cannot mark or stabilize the person's sexual identity; defining oneself as gay (or lesbian), sharing intimacy with self-identified homosexuals, or being a part of queer social environments are not enough alone to become gay. One needs to have a certain set of class criteria to achieve the proper, inherently modern gay identity. The most recurrent of these class signifiers are having or sharing an apartment in one of the decent middle–class areas of the city, being a college graduate or student, speaking foreign languages, adopting a secular lifestyle, following global cultural flows and fashions, traveling abroad, embodying and performing a specific style that is imbued with consumerism, self-care, and masculinity (at least for men). Being gay in this very setting is translated into becoming gay through utilizing cultural capital and mastering symbolic codes and then buoyantly sustaining it.

What happens when one has same-sex sexual affinities but for class reasons cannot follow the right codes of the middle upper–class gay notion? If he is more masculine and has a straight-acting aura then he is tagged as 'varos.' Varos in popular Turkish means both the destitute neighborhoods of informal housing and the poor, working class people who live in these areas. The word varos has obviously negative overtones and it is generally used in a derogatory sense. However, in the Turkish gay slang, varos is transformed into a word that signifies poverty and the lack of middle-class values, while it also highlights robust virility and an authentic, uncontaminated masculinity (Özbay, 2010). In this sense, the Turkish case bears a resemblance to the contexts in other countries in which the power of working-class masculinity defines or rejects sexual identities, labels, and communities (Aggleton, 1999; Barrett and Pollack, 2005; Parker, 1999). If the person who lacks the higher class qualities is more feminine, embodying an effeminate style, then this person is more easily deemed as a 'lubunya.' Lubunya is a term that was borrowed from transgender culture and it simply means sissy or unmanly (Berghan, 2007). Lubunya people are easier to match with the famous transgender singers' behavior by the public, and in some contexts they can be seen as candidates for future transvestism and transsexuality. Class for homosexuals in Turkey is deeply intertwined with desires that govern one's bodily presentations, gendered acts, and the ways of interpretation, which effect how they are related to other people.

Another significant aspect of contemporary homosexuality is the state institutions' intervention into the domain of morality, righteousness, and value through policy implementation at different levels. The most important aspect of this state intervention takes place in the compulsory military service (Biricik, 2009; Tarhan, 2008). According to law, all male Turkish citizens have to spend a certain amount of time in the army, serving for military purposes in the barracks. Only citizens with a predefined, extraordinary medical condition can avoid the service. Among the medical conditions that disqualify one from the service, in a very outdated psychological terminology, homosexuality is framed as an 'advanced psychosexual disorder' and thus people who are able to prove that they have this particular condition can waive the military service. After experiencing the awkward and torturous examinations held by the military officials, self-claimed homosexuals are entitled to receive a medical document called 'the rotten report,' which marks them as officially recognized homosexuals and tacitly excludes them from public life (Biricik, 2009). People, both gays and heterosexuals, believe that a person with the rotten report cannot work in state institutions, including public schools, and when they apply to jobs elsewhere their homosexuality will be known by the employers because of the circulation of the document between state and nonstate organizations. In other words, gay men in Turkey must find their way between two difficult strategies: perform the long compulsory military service as a closeted gay who travails to pass as straight, or get the rotten report, which has the potential to destroy their coming out strategies and their control over their personal and professional lives.

Despite the fact that the state, especially through the army, acknowledges the existence of homosexual citizens and labels them whenever it can, law does not recognize them otherwise and does not guarantee any sort of social rights that heterosexual citizens virtually have including but not limited to marriage, civil union contracts, and partnership benefits such as retirement, heritage, insurance, social security, and access to the corpse in case of death. No Turkish law specifically forbids discrimination on the basis of sexual orientation or sexuality in social institutions, government offices, or corporations. In fact, Turkey encourages discrimination through the distinct Ministry for Family and Social Policies (previously The State Ministry for Women), which reiterates the priority to protect families instead of individuals who opt for staying out of families; precludes alternative definitions of the family with a focus on definitions based on the conventional, traditional heterosexual family; and proselytizes reproductive policies. The previous head of the Ministry, Mrs Kavaf, even stated in 2010 that homosexuality was a disease that needed to be cured. Although she lost her chair

right afterward, this iteration is believed by many to reflect the official state attitude toward citizens with same-sex sexual inclinations. Supporting this assumption of state negativity, HIV/AIDS is still largely seen as a homosexual illness and the nongovernmental organizations that aim to prevent infection and to help HIV-positive people maintaining their lives are not supported by the state most probably because the sufferers as well as the charities about this illness are labeled as gay.

Issues, Visibilities, and Absences

Today, sexual minorities in urban Turkey are in great diversity consisting of gay men, lesbians, bisexuals, varos and lubunya individuals, transgenders, bears, and queers, who are mostly from the younger generations and defend the fluidity not fixity of sexual identities in the original sense of the term in English (Cakirlar and Delice, 2012; Özbay, 2010; Özyegin, 2012; Savci, 2011). In addition to the matters that emerge at the intersection of sexuality, Islam, social class, military policy, and the state, homosexuals have other issues to deal with in the course of everyday life. Among these issues are the fierce or symbolic violence they face; the continuing (although decreasing) confusion between gay and transgender identities; murders of both gays and transgenders; the homophobic approach that popular press and even politicians do not hesitate to use out of blue; the limited availability and despicably low physical standards of queer spaces; the huge secrecy, concealment, and insecurity in their lives; the rigid top–bottom sexual duality in terms of physical penetration; the scarce academic and scholarly research about queer lives and sexualities; and the intense marginalization and exclusion experienced by those intellectuals who think or write about sexualities. Some of these issues are evidently common in multiple international cultures, while some are strictly local and unique to the Turkish context and history.

Despite these limits, the emergence and the rise of same-sex sexual or queer activism also deserved to be indicated. Lambda and Kaos, in Istanbul and Ankara, respectively, have been active since early 1990s and they have improved a lot in terms of visibility and respectability. Queer college students at different campuses started to come together in the 2000s and they are still the most powerful group regarding the mobilization of young people against homophobia and heteronormativity. Turkey is a conservative and religious country and its politicians, whether from left or right wings, have traditionally stayed away from any sort of sexuality issues, starting with homosexual politics. Nevertheless, in the last couple of years, observers began to see a minor shift, especially in the two leftist political parties, the secular-modernist Republican People's Party (CHP) and the Kurdish Peace and Democracy Party (BDP). One BDP parliament member stated in the commission for rewriting the constitution that the less particular expressions such as everybody is equal in front of law should instead be more explicit and specifically declare that lesbians, gay men, bisexuals, and transgender individuals' situation need to be clearly mentioned in the new constitution. A group of CHP parliamentary members, among them former academicians, held a press conference with families of gay and transgender people and advocated for equal rights to sexual minorities. Another CHP congress member visited the gay and transgender section in a prison and talked to them about their demands. These are small but symbolically remarkable actions, exceptional and meaningful steps in the long process of having equal sexual citizenship and a democratic sense of identification and representation.

Recently, there has been a boom in queer artistic production and representation. Major novelists, among them Perihan Magden (2006, 2012), Duygu Asena (2006), and Elif Shafak (2007), as well as more independent literary figures, such as Niyazi Zorlu (2003) and Mehmet Murat Somer (2008, 2009), published thriving books while the queer filmmaker Kutlug Ataman's movie *Two Girls* (2007), which was adapted from Magden's novel, was noted as the first modern lesbian film that became mainstreamed in the Turkish popular culture. The respected director and university professor Can Candan produced a documentary about parents of queer children, *My Child* (2013), which became a national sensation and discussed widely in heterosexual public as well as among queer circles.

In terms of queer visibility, the greatest event that happens in Turkey is the Gay Pride celebrations that have taken place each June on the Istiklal Street in the Beyoglu district of Istanbul since 2001. Thousands of people, queer and straight, men and women and transgender, young and old, in all their diversity, gather and parade in attires that are quite counternormative for the standard Turkish taste, and act out, kiss each other, and dance together, which in other settings across the city in a manner that would typically create physically abusive, homophobic reaction. Onlookers can easily detect the carnivalesque collective spirit that inspires many otherwise closeted gays and lesbians, tired from oppression and concealment, protesting the heterosexist social values by their conspicuous presence in the heart of the city. The police, while normally harsh toward the protesters of any sort, do not attack the participants of the Pride, let the queer citizens walk, and maintain security through surveillance from distance.

At the end of the carnival, the burdensome conditions for queer people in Turkey recommence. There are serious taboos about same-sex sexualities in many spheres of social life. Sports, for example, especially football is one of them. The first and only out-of-closet gay football referee is not appointed to games any more by the national federation regardless of the support by many fans' associations. The theological circles, legislation, and the bureaucratic elites are rather passive and silent if not entirely against having queer members. Higher education is also dubious and not necessarily inclusionary – if not explicitly discriminating existing or prospective queer scholars. There is no out queer politician at either local or national level in Turkey. Except for a couple of elder actors and writers, there is no single esteemed, well-achieved, inspiring queer role model in the country.

Conclusion

The governing logic of the conservative, pious, neoliberal, corporate collective consciousness in Turkey avoids, forbids, condemns, marginalizes, and blacklists queer people and same-sex sexual acts. Although disempowered by much of culture and the state, these people strive to survive, make peace with the idea

of whom they are, congregate, socialize, and attempt to change the hostile social order in a bold and electrifying manner. Dynamics of globalization and transnational flows of ideas and expressions brought certain novelties, not only in Turkey but also in the Middle East as a whole, in terms of sexualities and intimacies (Özbay, 2012). It is difficult to say whether 50 years ago same-sex sexualities were freer in Turkey or nowadays they are more oppressed because it is almost impossible to fully grasp the clandestine nature of queer acts in the past. However, it is most certain that today one talks, thinks, and knows more about sexualities and in this discursive frame same-sex sexualities and queer identities occupy a great place. The Turkish case presents a twisted example simply because globalizing, more flexible, liberating currents are juxtaposed here with a more conservative, authoritarian, subjugating undertow. Thus, it is not entirely possible to exactly foresee which direction Turkish same-sex sexualities will take in the future, but it is obvious from the recent developments that the symbolic, social, and political struggle of queers against homophobia in Turkey has been ignited.

See also: Clinical Psychology, Psychiatry and Homosexuality; Gay and Lesbian Movements; Gay and Lesbian Tourism; Heterosexism and Homophobia; Islam: Middle East; Mass Media, Representations In; Sexual Minorities and Sexual Citizenship; Sexual Orientation: Historical and Social Construction; Social Protest.

Bibliography

Aggleton, P. (Ed.), 1999. Men Who Sell Sex: International Perspectives on Male Prostitution and HIV/AIDS. Temple University Press, Philadelphia.
Arat, Y., 1997. The project of modernity and women in Turkey. In: Bozdogan, S. (Ed.), Rethinking Modernity and National Identity in Turkey. University of Washington Press, Seattle.
Armbrecht, T., 2001. The cucumber seller: homosexuality and class in Turkey. Gay Men's Fiction Quarterly 3, 51–61.
Asena, D., 2006. Paramparca. Dogan, Istanbul.
Barrett, D.C., Pollack, L.M., 2005. Whose gay community? Social class, sexual self-expression, and gay community involvement. Sociological Quarterly 46, 437–456.
Basdas, B., 2007. Cosmopolitanism in the City: Contested Claims to Bodies and Sexualities in Beyoglu, Istanbul (unpublished Ph.D. dissertation). Department of Geography, University of California, Los Angeles.
Bereket, T., Adam, B.D., 2006. The emergence of gay identities in contemporary Turkey. Sexualities 9, 131–151.
Bereket, T., Adam, B.D., 2008. Navigating Islam and same-sex liaisons among men in Turkey. Journal of Homosexuality 55, 204–222.
Berghan, S., 2007. Lubunya: Transeksuel Beden ve Kimlik. Metis, Istanbul.
Biricik, A., 2009. Rotten report and reconstructing hegemonic masculinity in Turkey. In: Cinar, O.H., Usterci, C. (Eds.), Conscientious Objection: Resisting Militarized Society. Zed, London.
Cakirlar, C., Delice, S. (Eds.), 2012. Cinsellik Muammasi: Turkiye'de Queer Kultur ve Muhalefet. Metis, Istanbul.
D'Emilio, J., 1998. Sexual Politics, Sexual Communities. University of Chicago Press, Chicago.
Hocaoglu, M., 2002. Escinsel Erkekler: Yirmi Bes Taniklik. Metis, Istanbul.
Kocu, R.E., 1946. Istanbul Ansiklopedisi. Istanbul Yayinevi, Istanbul.
Magden, P., 2006. Two Girls. Serpent's Tail, London.
Magden, P., 2012. Ali and Ramazan. AmazonCrossing, Las Vegas.
Nardi, P.M., Schneider, B.E. (Eds.), 1998. Social Perspectives in Lesbian and Gay Studies: A Reader. Routledge, New York.
Oz, Y., 2009. Ahlaksizlarin Merkansal Dislanmasi. In: Alkan, A. (Ed.), Cins Cins Mekan. Varlik, Istanbul.
Özbay, C., 27 September 2009. Queer Kral. Radikal Iki.
Özbay, C., 2010. Nocturnal queers: rent boys' masculinity in Turkey. Sexualities 13, 645–663.
Özbay, C., 2012. Sexuality. In: Stanton, A.L. (Ed.), Cultural Sociology of the Middle East, Asia, and Africa. Sage, London.
Özbay, C., Soydan, S., 2003. Escinsel Kadinlar: Yirmi Dort Taniklik. Metis, Istanbul.
Oztek, A., 2007. Turk Sinemasinda Escinsellik. Kaos GL 52.
Özyegin, G., 2012. Reading the closet through connectivity. Social Identities 18, 201–222.
Parker, R., 1999. Beneath the Equator: Cultures of Desire, Male Homosexuality and Emerging Gay Communities in Brazil. Routledge, New York.
Plummer, K. (Ed.), 1992. Modern Homosexualities: Fragments of Lesbian and Gay Experiences. Routledge, London.
Savci, E., 2011. Queer in Translation: Paradoxes of Westernization and Sexual Others in Turkish Nation (unpublished Ph.D. dissertation). Department of Sociology, University of Southern California, Los Angeles.
Seidman, S., 2009. The Social Construction of Sexuality. Norton, New York.
Shafak, E., 2007. Pinhan. Metis, Istanbul.
Sirman, N., 2005. The making of familial citizenship in Turkey. In: Icduygu, A. (Ed.), Citizenship in a Global World – European Questions and Turkish Experiences. Routledge, London.
Somer, M.M., 2008. The Prophet Murders: A Hop-Ciki-Yaya Thriller. Serpent's Tail, London.
Somer, M.M., 2009. The Jigolo Murder: A Turkish Delight Mystery. Penguin, London.
Tapinç, H., 1992. Masculinity, femininity and Turkish male homosexuality. In: Plummer, K. (Ed.), Modern Homosexualities. Routledge, London.
Tarhan, M., 2008. Zorunlu Askerlik ve Sivil Alternatif Hizmete Direnis Olarak Vicdani Red. In: Muluer, N. (Ed.), Cinsiyet Halleri. Varlik, Istanbul.
Weeks, J., 1990. Sexuality and Its Discontents: Meanings, Myths and Modern Sexualities. Routledge, London.
Yuzgun, A., 1986. Turkiye'de Escinsellik (Homosexuality in Turkey). Huryuz, Istanbul.
Yuzgun, A., 1993. Homosexuality and police terror in Turkey. Journal of Homosexuality 24, 159–169.
Zorlu, N., 2003. Hergele Asiklar. Metis, Istanbul.

Sample Survey Methodology, History of

Stephen E Fienberg, Carnegie Mellon University, Pittsburgh, PA, USA
Judith M Tanur, Stony Brook University, Montauk, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sampling methods have their roots in attempts to measure the characteristics of a nation's population by using only a part instead of the whole population, i.e., a census. The movement from the exclusive use of censuses to the at least occasional use of samples was slow and laborious. A watershed was the debate over representative methods and the rise of probability sampling, including the seminal 1934 paper of Jerzy Neyman. This article describes these developments as well as the impact in the United States of the results of this work on sample surveys, the contributions of statisticians working in government statistical agencies, and the subsequent establishment of random sampling and survey methodology more broadly.

Introduction

Sampling methods have their roots in attempts to measure the characteristics of a nation's population by using only a part instead of the whole population. The movement from the exclusive use of censuses to the at least occasional use of samples was slow and laborious. Two intertwined intellectual puzzles had to be solved before the move was complete. The first such puzzle was whether it is possible to derive valid information about a population by examining only a portion thereof (the 'representative' method). And the second puzzle concerned the method for choosing that portion. Issues of choice themselves fall into two categories: the structuring of the population itself in order to improve the accuracy of the sample, and whether to select the units for inclusion by purposive or random methods. This article considers the development of methods for the sampling of human populations. We begin by describing two antecedents of modern survey methods – the development of government statistics and censuses and the general development of statistical methodology, especially for ratio estimation. Then we deal with the movement from censuses to sample surveys, examining the debate over representative methods and the rise of probability sampling, including the seminal 1934 paper of Jerzy Neyman. We describe the impact in the United States of the results obtained by Neyman on sample surveys and the contributions of statisticians working in US government statistical agencies. We end with a brief description of the subsequent establishment of random sampling and survey methodology more broadly and examine some history of the movement from the study of the design of sampling schemes to the study of other issues arising in the surveying of human populations.

Antecedents of Modern Sample Survey Methods

In order to determine facts about their subjects or citizens, and more often about their number, governments have long conducted censuses. Thus, one set of roots of sample survey methodology is intertwined with the history of methods for census taking. The origins of the modern census are found in biblical censuses described in the Old Testament (e.g., see the discussions in Duncan, 1984 and Madansky, 1986) as well as in censuses carried out by the ancient Egyptians, Greeks, Japanese, Persians, and Romans (Taeuber, 1978). For most practical purposes we can skip from biblical times to the end of the eighteenth century and the initiation of census activities in the United States of America, even though there is some debate as to whether Canada, Sweden, or the United States should be credited with originating the modern census (Willcox, 1930).

The word statistics comes to us via the German word *Statistik*, introduced by Gottfried Achenwall in 1749, which derived from the Latin status is related to the word state (John, 1883). He used it to designate the analysis of data about the state, which in English was referred to as political arithmetic. In the eighteenth century, statistics was typically used in the context of data about society, often in what we have come to know as official statistics, but it also had other more specific applications in economics, demography, and politics. These uses of statistics lacked formal methodological techniques for gathering and analyzing data.

Another antecedent of sampling lies in the problem of estimating the size of a population when it is difficult or impossible to conduct a census or complete enumeration. Attempts to solve this problem even in the absence of formal sampling methods were the inspiration for what we now know as ratio estimation. These ideas emerged as early as the seventeenth century. For example, John Graunt (1620–74) used the technique to estimate the population of England from the number of births and some assumptions about the birth rate and the size of families. Graunt's research, later dubbed political arithmetic, was based on administrative records (especially parish records) and personal observation, as in his seminal work, *Natural and Political Observations Made Upon the Bills of Mortality*, published in 1662. To accomplish what Graunt did by assumption, others looked to hard data from subsets of a population. Thus, Sir Frederick Morton Eden, for example, found the average number of people per house in selected districts of Great Britain in 1800. Using the total number of households in Great Britain from tax rolls (and allowing for those houses missing from the rolls), Eden estimated the population of Great Britain as 9 million, a figure confirmed by the first British census in 1801 (Stephan, 1948).

Even earlier, in 1765 and 1778, civil servants published population estimates for France using an enumeration of the population in selected districts and counts of births, deaths, and marriages, for the country as a whole (see Stephan, 1948).

It was Pierre Simon de Laplace (1749–1827), however, who first formally described the method of ratio estimation during the 1780s, and he then employed it in connection with a sample survey that he initiated to estimate the population of France as of 22 September 1802. He arranged for the government to take a sample of administrative units (communes), in which the total population, y, and the number of registered births in the preceding year, x, were measured. Laplace then estimated the total population of France by $\hat{Y} = Xy/x$, where X was the total number of registered births. Laplace was the first to estimate the asymptotic bias and variance of $\hat{Y}$, through the use of a superpopulation model for the ratio $p = y/x$ (see Laplace, 1814). William Cochran (1978) describes this derivation and links it to results that followed 150 years later. Quetelet later applied the ratio estimation approach in Belgium in the 1820s but abandoned it following an attack by others on some of the underlying assumptions (see Stigler, 1986: pp. 162–166).

From Censuses to Surveys

The move from censuses to sample surveys to measure the characteristics of a nation's population was slow and laborious. Bellhouse (1988), Kruskal and Mosteller (1980), and Seng (1951) trace some of this movement, especially as it was reflected in the discussions regarding surveys that took place at the congresses of the International Statistical Institute (ISI). The move succeeded when investigators combined random selection methods with the structuring of the population and developed a theory for relating estimates from samples, however structured, to the population itself.

As early as the 1895 ISI meeting, Kiaer (1896) argued for a 'representative method' or 'partial investigation,' in which the investigator would first choose districts, cities, etc., and then units (individuals) within those primary choices. The choosing at each level was to be done purposively, with an eye to the inclusion of all types of units. That coverage tenet, together with the large sample sizes recommended at all levels of sampling, was what was judged to make the selection representative. Thus the sample was, approximately, a miniature of the population. Kiaer used systematic sampling in his 1895 survey of workers as a means of facilitating special tabulations from census schedules for a study of family data in the 1900 Norwegian census. Kiaer actually introduced the notion of random selection in his description of this work by noting that at the lowest level of the sample structure "the sample should be selected in a haphazard and random way, so that a sample selected in this manner would turn out in the same way as would have been the case had the sample been selected through the drawing of lots …" (Kiaer, 1897: p. 39).

The idea of less than a complete enumeration was opposed widely, and Kiaer presented arguments for sampling at ISI meetings in 1897, 1901, and 1903. Lucien March, in a discussion of Kiaer's paper at the 1903 meeting, formally introduced the concepts of simple random sampling without replacement and simple cluster sampling, although not using these names (Bellhouse, 1988).

Arthur Lyon Bowley did a number of empirical studies on the validity of random sampling, motivated at least in part by a 1912 paper by Francis Ysidro Edgeworth. Bowley used large sample normal approximations (but did not conceptualize what he was doing as sampling from a finite population) to actually test the notion of representativeness. Then, he carried out a 1912 sample survey on poverty in Reading, in which he drew the respondents at random (see Bowley, 1913), although he appears to have equated simple random sampling with systematic sampling (Bellhouse, 1988).

At about the same time, in what appears to be an independent line of discovery, Tchouproff and others were overseeing the implementation of sample survey methods in Russia, especially during World War I and the period immediately following it. This work was partially documented after the Russian Revolution in *Vestnik Statistiki*, the official publication of the Central Statistical Administration (Zarkovic, 1956, 1962). Even though there is clear evidence that Tchouproff had developed a good deal of the statistical theory for random sampling from finite populations during this period (see Seneta, 1985), whether he actually implemented forms of random selection in these early surveys is unclear. He later published the formulae for the behavior of sample estimates under simple random sampling and stratified random sampling from finite populations in Tchouproff (1918a,b, 1923a,b).

In a seemingly independent development of these basic ideas for sampling in the context of his work on agricultural experiments, Neyman in Splawa-Neyman (1923) described them in the form of the drawing of balls without replacement from an urn. In the resulting 1925 paper (Splawa-Neyman, 1925), he gave the basic elements of the theory for sampling from finite populations and its relationship with sampling from infinite populations. These results clearly overlapped those of Tchouproff, and 2 years later following Tchouproff's death, his partisans took Neyman to task for his lack of citation to the Russian work as well as to earlier work of others (see Fienberg and Tanur, 1995, 1996, for a discussion of the controversy).

By 1925, the record of the ISI suggests that the representative method was taken for granted, and the discussions centered around how to accomplish representativeness and how to measure the precision of sample-based estimates, with the key presentations being made by Bowley (1926) and in the same year by the Danish statistician Adolph Jensen. Notions of clustering and stratification were put forward, and Bowley presented a theory of proportionate stratified sampling as well as the concept of a frame, but purposive sampling was still the method of choice. It was not until Gini and Galvani made a purposive choice of which returns of an Italian census to preserve, and found that districts chosen to represent the country's average on seven variables were, in that sense, unrepresentative on other variables, that purposive sampling was definitively discredited among professional statisticians (Gini, 1928; Gini and Galvani, 1929). The notion of representative sampling lives on, however, especially in the modern era of Internet polling and sampling.

Neyman's (1934) Paper on Sampling

In their work on the Italian census, Gini and Galvani seemed to call into question the accuracy of sampling. Neyman took up their challenge in his classic 1934 paper presented before the

Royal Statistical Society, "On the two different aspects of the 'representative method'." In it, he compared purposive and random sampling, and concluded that it was not sampling that was problematic but rather Gini and Galvani's purposive selection.

Elements of synthesis were prominent in the paper as well. Neyman explicitly uncoupled clustering and purposive sampling, saying, "In fact the circumstance that the elements of sampling are not human individuals, but groups of these individuals, does not necessarily involve a negation of the randomness of the sampling" (1952: p. 571). He called this procedure, 'random sampling by groups' and pointed out that, although Bowley did not consider it theoretically, he used it in practice in London, as did O. Anderson in Bulgaria. Neyman also combined stratification with clustering to form 'random stratified sampling by groups,' and he provided a method for deciding how best to allocate samples across strata (optimal allocation). The immediate effect of Neyman's paper was to establish the primacy of the method of stratified random sampling over the method of purposive selection, something that was left in doubt by the 1925 ISI presentations by Jensen and Bowley. But the paper's longer-term importance for sampling was the consequence of Neyman's wisdom in rescuing clustering from the clutches of those who were the advocates of purposive sampling and integrating it with stratification in a synthesis that laid the groundwork for modern-day multistage probability sampling.

Surprisingly, for many statisticians the memorable parts of Neyman's paper were not these innovations in sampling methodology but Neyman's introduction of general statistical theory for point and interval estimation, especially the method of confidence intervals.

As pathbreaking as Neyman's paper was, a number of its results had appeared in an earlier work by Tschuprow (1923a, 1923b), in particular the result on optimal allocation in stratified sampling (see Fienberg and Tanur, 1995, 1996, for discussions of this point). The Danish mathematician Gram (1883) derived the method much earlier in a paper dealing with calculations for the cover of a forest based on a sample of trees. Gram's work has been rediscovered only in the twenty-first century and seems not to have been accessible to either Tchouproff or Neyman (see Merton, 1961/1973, for a discussion of instances of such 'multiples' in science).

Neyman provided the recipe for others to follow and he continued to explain its use in convincing detail to those who were eager to make random sampling a standard diet for practical consumption (e.g., see Neyman, 1952 for a description based on his 1937 lectures on the topic at the US Department of Agriculture Graduate School, as discussed below).

The Development of Random Sampling in the United States

One might have thought that the resolution of the controversy over the representative method and the articulation of the basic elements of a theory of sample surveys by Neyman would have triggered extensive application of random sampling throughout the world. Surprisingly, this was not the case. With some notable exceptions, for example, in England (see Cochran, 1939; Yates, 1946) and India, the primary application occurred in the United States and this led to a spate of new and important methodological developments.

As late as 1932, however, there were few examples of probability sampling anywhere in the US federal government (Duncan and Shelton, 1978) and the federal statistical agencies had difficulty responding to the demand for statistics to monitor the effects of the programs of President Franklin Roosevelt's New Deal. In 1933, the American Statistical Association (ASA) set up an advisory committee that grew into the Committee on Government Statistics and Information Services (COGSIS), sponsored jointly by ASA and the Social Science Research Council. COGSIS helped to stimulate the use of probability sampling methods in various parts of the Federal government, and it encouraged employees of statistical agencies to carry out research on sampling theory. For example, to establish a technical basis for unemployment estimates, COGSIS, and the Central Statistical Board which it helped to establish, organized an experimental Trial Census of Unemployment as a Civil Works Administration project in three cities, using probability sampling, and carried out in late 1933 and early 1934. The positive results from this study led in 1940 to the establishment of the first large-scale, ongoing sample survey on employment and unemployment using probability sampling methods. This survey later became known as the Current Population Survey and it continues to the present day.

Another somewhat indirect outcome of the COGSIS emphasis on probability sampling took place at the Department of Agriculture Graduate School where W. Edwards Deming, who recognized the importance of Neyman's (1934) paper, invited Neyman to present a series of lectures in 1937 on sampling and other statistical methods (Neyman, 1952). These lectures had a profound impact on the further development of sampling theory not simply in agriculture, but across the government as well as in universities.

Among those who worked on the probability-sampling-based trial Census of Unemployment at the Bureau of the Census was Morris Hansen, who was then assigned with a few others to explore the field of sampling for other possible uses at the Bureau, and went on to work on the 1937 sample Unemployment Census. After working on the sample component of the 1940 decennial census (under the direction of Deming), Hansen worked with others to redesign the unemployment survey based on new ideas on multistage probability samples and cluster sampling (Hansen and Hurwitz, 1942, 1943). They expanded and applied their approach in various Bureau surveys, often in collaboration and interaction with others, and this effort culminated in 1953 with the publication of a two-volume compendium of theory and methodology (Hansen et al., 1953a,b).

Cochran's (1939) paper, written in England and independently of the US developments, is especially notable because of its use of the analysis of variance in sampling settings and the introduction of superpopulation and modeling approaches to the analysis of survey data. In the 1940s, as results from these two separate schools appeared in various statistical journals, we see some convergence of ideas and results. The theory of estimation in samples with unequal probabilities of selection also emerged around this time (see Horvitz and Thompson, 1952; Hansen et al., 1985).

Statisticians have continued to develop the theoretical basis of alternative methods of probability sampling and statistical inference from sampling data since the middle of the twentieth century (see, e.g., Rao and Bellhouse, 1990; Sarndal et al., 1992). The issue of statistical inference for models from survey data remains controversial, at least for those trained from a traditional finite sampling perspective.

Market Research and Polling

The 1930s, and especially the period after World War II, however, saw a flowering of survey methodology in market research and polling, as well as in the social sciences more broadly. Initial stimulation came from a number of committees at the Social Science Research Council and social scientists such as Hadley Cantril, Paul Lazarsfeld, Rensis Likert, William Ogburn, and Samuel Stouffer (e.g., see Stephan, 1948; Converse, 1987 for a discussion). Stouffer actually spent several months in England in 1931–32, and learned about sampling and other statistical ideas from Bowley, Karl and Egon Pearson, and R.A. Fisher. He then participated in the Census of Unemployment at the Bureau of the Census.

Market research and polling trace their own prehistory to election straw votes collected by newspapers, dating back at least to the beginning of the nineteenth century. Converse (1987) points out, however, a more serious journalistic base; election polls were taken and published by such reputable magazines as the *Literary Digest* (which had gained a reputation for accuracy before the 1936 fiasco). Then, as now, election forecasting was taken as the acid test of survey validity. A reputation for accuracy in 'calling' elections was thought to spill over to a presumption of accuracy in other, less verifiable areas.

There was a parallel tradition in market research, dating back to just before the turn of the twentieth century, attempting to measure consumers' product preferences and the effectiveness of advertising. It was seen as only a short step from measuring the opinions of potential consumers about products to measuring the opinions of the general public about other objects, either material or conceptual. By the mid-1930s there were several well-established market research firms. Many of them conducted election polls in 1936 and achieved much greater accuracy than did the *Literary Digest*. It was the principals of these firms (e.g., Archibald Crossley, George Gallup, and Elmo Roper) who put polling – election, public opinion, and consumer – on the map in the immediate pre-World War II period.

The polling and market research surveys of Crossley, Gallup, Roper, and others were based on a sampling method involving 'quota controls,' but did not involve random sampling. Stephan (1948) observed the close link between their work and the method of purposive sampling that had gained currency much earlier in government and academic research circles, but which by the 1930s had been supplanted by random sampling techniques.

From Sampling Theory to the Study of Nonsampling Error

The 1940s saw a rapid spread of probability sampling methods to a broad array of government agencies. It was, however, only after the fiasco of the 1948 presidential preelection poll predictions (Mosteller et al., 1949) that market research firms and others shifted toward probability sampling. Even today many organizations use a version of probability sampling with quotas (Sudman, 1967).

Amidst the flurry of activity on the theory and practice of probability sampling during the 1940s, attention was being focused on issues of nonresponse and other forms of nonsampling error (e.g., see Deming, 1944) such as difficulty in understanding questions, or remembering answers, etc. A milestone in this effort to understand and model nonresponse errors was the development of an integrated model for sampling and nonsampling error in censuses and surveys, in connection with planning for and evaluation of the 1950 census (Hansen et al., 1951). This analysis-of-variance-like model, or variants of it, has served as the basis of much of the work on nonsampling error since.

Developments in sampling since the mid-twentieth century have to do less with the design of samples and more to do with the structuring of the survey instrument and interview. Thus, there has been a move from face-to-face interviewing to telephone interviewing (with the attendant problems connected with random-digit dialing), and then to the use of computers to assist in interviewing. Nonsampling errors continue to be studied broadly, often under the rubric of cognitive aspects of survey design.

In the early twenty-first century several challenges loomed for the survey effort. As the volume of surveys increased, response rates continued their decline, even for survey conducted for branches of the Federal Government (Curtin et al., 2005; Groves, 2006, 2011). A good deal of work was done to try to increase response rates, and more was done to evaluate the possible error introduced by nonresponse. In addition, as more of the population in the developed world moved from landlines to mobile phones, methods of sampling cell phones began to be explored. Such methods had to take into account two broad sets of issues. First, a mobile phone usually belongs to an individual, in contrast to a landline that typically belongs to a household. Hence the *ultimate sampling unit* is typically an individual, not a household, and many surveys need to be adjusted to take this difference into account. Second, because cell phone users, at least in the US, normally pay for incoming as well as outgoing calls, new issues of respondent cooperation arose, as did government regulations that forbade the use of automatic dialing devices to reach cell phones.

As telephone surveys became harder to carry out and developed greater nonresponse problems, researchers began exploring the possibility of using the Internet to carry out sample surveys. It became commonplace, even for government-sponsored surveys (for example, the NSF Survey of Doctoral Recipients), to give respondents the choice of responding by mail, phone, or the Internet. In addition, a new Internet survey mode was developed, where respondents were recruited via e-mail (sometimes based on a random sample of phone contacts) and formed a panel from which members were repeatedly recruited for surveys. At the time of preparation of this article, there is no clear consensus on the accuracy of these Internet panels. For an excellent set of articles on web surveys, see Couper and Miller (2008) and the special issue of the journal *Public Opinion Quarterly* to which that article is an introduction.

With the WWW and advances in computing and storage, we have entered the era of *big data*. Many believe that more data are better and that there no longer is the need for sampling and the statistical theory associated with it. 'Found data' such as those amassed on individuals from the WWW have many attractions, but there remains great value in examining carefully selected samples of Web pages instead of looking at the entire WWW. How to do this efficiently is one of the new challenges for sampling theory. Moreover, 'found data' are not a substitute for careful survey measurement on data gathered from a randomly selected subset of a well-defined population.

See also: Censuses: History and Methods; Estimation: Point and Interval; Government Statistics; Linear Hypothesis; Sample Surveys: Cognitive Aspects of Survey Design.

Bibliography

Bellhouse, D., 1988. A brief history of random sampling methods. In: Krishnaiah, P., Rao, C.R. (Eds.), Handbook of Statistics, vol. 6. North Holland, Amsterdam, pp. 1–14.

Bowley, A., 1913. Working-class households in Reading. Journal of the Royal Statistical Society 76, 672–701.

Bowley, A., 1926. Measurement of the precision attained in sampling. Bulletin of the International Statistical Institute 22 (1), 6–62.

Cochran, W.G., 1939. The use of the analysis of variance in enumeration by sampling. Journal of the American Statistical Association 128, 124–135.

Cochran, W.G., 1978. Laplace's ratio estimator. In: David, H. (Ed.), Contributions to Survey Sampling and Applied Statistics: Papers in Honor of H. O. Hartley. Academic Press, New York, pp. 3–10.

Converse, J., 1987. Survey Research in the United States: Roots and Emergence 1890–1960. University of California Press, Berkeley, CA.

Couper, M., Miller, P.V., 2008. Web survey methods. Public Opinion Quarterly 72, 831–835.

Curtin, R., Presser, S., Singer, E., 2005. Changes in telephone survey nonresponse over the past quarter century. Public Opinion Quarterly 69 (1), 87–98.

Deming, W., 1944. On errors in surveys. American Sociological Review 19, 359–369.

Duncan, J., Shelton, W., 1978. Revolution in United States Government Statistics, 1926–1976. US Government Printing Office, Washington, DC.

Duncan, O.D., 1984. Notes on Social Measurement. Sage, New York.

Edgeworth, F.Y., 1912. On the use of the theory of probabilities in statistics relating to society. Journal of the Royal Statistical Society 76, 165–193.

Fienberg, S.E., Tanur, J.M., 1995. Reconsidering Neyman on experimentation and sampling: controversies and fundamental contributions. Probability and Mathematical Statistics 15, 47–60.

Fienberg, S.E., Tanur, J.M., 1996. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. International Statistical Review 64, 237–253.

Gini, C., 1928. Une application de la methode representative aux materiaux du dernier recensement de la population italienne (1er decembre 1921). Bulletin of the International Statistical Institute 23 (2), 198–215.

Gini, C., Galvani, L., 1929. Di una applicazione del metodo rappresentative all'ultimo censimento italiano della popolazione (1 Decembri 1921). Annali di Statistica 6 (4), 1–107.

Gram, J., 1883. About calculation of the mass of a forest cover by means of test trees (in Danish). Tidsskrft for Skovbrug 6, 137–198.

Groves, R.M., 2006. Nonresponse bias in household surveys. Public Opinion Quarterly 70 (5), 646–675.

Groves, R.M., 2011. Three eras of survey research. Public Opinion Quarterly 75 (5), 861–871.

Hansen, M., Dalenius, T., Tepping, B., 1985. The development of sample surveys of finite populations. In: Atkinson, A.C., Fienberg, S.E. (Eds.), A Celebration of Statistics: The ISI Centenary Volume. Springer-Verlag, New York, pp. 327–354.

Hansen, M., Hurwitz, W., 1942. Relative efficiencies of various sampling units in population inquiries. Journal of the American Statistical Association 37, 89–94.

Hansen, M., Hurwitz, W., 1943. On the theory of sampling from finite populations. Annals of Mathematical Statistics 14, 333–362.

Hansen, M., Hurwitz, W., Madow, W., 1953a. Sample Survey Methods and Theory, vol. 1. Wiley, New York.

Hansen, M., Hurwitz, W., Madow, W., 1953b. Sample Survey Methods and Theory, vol. 2. Wiley, New York.

Hansen, M., Hurwitz, W., Marks, E., Mauldin, W., 1951. Response errors in surveys. Journal of the American Statistical Association 46, 147–190.

Horvitz, D., Thompson, D., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Jenson, A., 1926. Report on the representative method in statistics. Bulletin of the International Statistical Institute 22, 359–380.

John, V., 1883. The term 'statistics.'. Journal of the Statistical Society, London 46 (4), 656–679.

Kiaer, A., 1895/1896. Observation et experiences concernant des denombrements representatifs. Bulletin of the International Statistical Institute 9 (2), 176–183.

Kiaer, A., 1897. The representative method of statistical surveys (in Norwegian). Christiana. Videnskabsselskabets Skrifter. II Hisoriskfilosofiske 4 [Reprinted with an English translation by Central Bureau of Statistics of Norway, Oslo, 1976].

Kruskal, W., Mosteller, F., 1980. Representative sampling, iv: the history of the concept in statistics, 1895–1939. International Statistical Review 48, 169–195.

Laplace, P., 1814. Essai philosophique sur les probabilites. Dover, New York [Translation of 1840 sixth ed. appeared in 1951 as A Philosophical Essay on Probabilities, F.W. Truscott, F.L. Emory, Trans.].

Madansky, A., 1986. On biblical censuses. Journal of Official Statistics 2, 561–569.

Mahalanobis, P., 1944. On large-scale sample surveys. Philosophical Transactions of the Royal Society of London 231 (B), 329–451.

Mahalanobis, P., 1946. Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society 109, 325–378.

Merton, R.K., 1961. Singletons and multiples in scientific discovery: a chapter in the sociology of science. Proceedings of the American Philosophical Society 105.5, 470–486 [Reprinted in Merton, R.K., 1973. The Sciology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago, IL, pp. 343–370].

Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S., Truman, D.B., 1949. The Pre-Election Polls of 1948. Social Science Research Council, New York.

Neyman, J., 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97, 558–625.

Neyman, J., 1952. Lectures and Conferences on Mathematical Statistics and Probability. US Department of Agriculture, Washington, DC [The 1952 edition is an expanded and revised version of the original 1938 mimeographed edition].

Rao, J., Bellhouse, D., 1990. The history and development of the theoretical foundations of survey based estimation and statistical analysis. Survey Methodology 16, 3–29.

Sarndal, C., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Seneta, E., 1985. A sketch of the history of survey sampling in Russia. Journal of the Royal Statistical Society Series a 148, 118–125.

Seng, Y., 1951. Historical survey of the development of sampling theories and practice. Journal of the Royal Statistical Society, Series a 114, 214–231.

Splawa-Neyman, J., 1923. On the application of probability theory to agricultural experiments. Essay on principles (in Polish). Roczniki Nauk Rolniczych Tom (Annals of Agricultural Sciences) X, 1–51 [Sect. 9, translated and edited by D.M. Dabrowska and T.P. Speed, appeared with discussion in *Statistical Science* (1990) 5: 463–480].

Splawa-Neyman, J., 1925. Contributions of the theory of small samples drawn from a finite population. Biometrika 17, 472–479 [The note on this republication reads 'These results with others were originally published in La Re ue Mensuelle de Statistique, publ. Par l' Office Central de Statistique de la Republique Polonaise 1923 tom. vi, 1–29].

Stephan, F., 1948. History of the use of modern sampling procedures. Journal of the American Statistical Association 43, 12–39.

Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, MA.

Sudman, S., 1967. Reducing the Costs of Surveys. Aldine, Chicago.

Taeuber, C., 1978. Census. In: Kruskal, W.H., Tanur, J.M. (Eds.), International Encyclopedia of Statistics. Macmillan and the Free Press, New York, pp. 42–46.

Tchouproff, A.A., 1918a. On the mathematical expectation of the moments of frequency distributions (Chapter 1 and 2). Biometrika 12, 140–169.
Tchouproff, A.A., 1918b. On the mathematical expectation of the moments of frequency distributions (Chapter 3 and 4). Biometrika 12, 185–210.
Tschuprow, A.A., 1923a. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chapters i–iii). Metron 2, 461–493.
Tschuprow, A.A., 1923b. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations (Chapters iv–vi). Metron 2, 646–680.
Willcox, W., 1930. Census. In: Seligman, E.R.A., Johnson, A. (Eds.), Encyclopedia of Social Sciences, vol. 3. Macmillan, New York, pp. 295–300.
Yates, F., 1946. A review of recent developments in sampling and sample surveys (with discussion). Journal of the Royal Statistical Society 109, 12–42.
Zarkovic, S.S., 1956. Note on the history of sampling methods in Russia. Journal of the Royal Statistical Society, Series A 119, 336–338.
Zarkovich, S.S., 1962. A supplement to 'note on the history of sampling methods in Russia.'. Journal of the Royal Statistical Society, Series A 125, 580–582.

Sample Surveys: Cognitive Aspects of Survey Design

Judith M Tanur, Stony Brook University, Stony Brook, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cognitive aspects of survey design refer to a research orientation that began in the early 1980s to integrate methods, techniques, and insights from the cognitive sciences into the continuing effort to render data from sample surveys of human populations more valid and reliable, and to understand systematically the threats to such reliability and validity. It was also intended to offer to cognitive scientists, especially cognitive psychologists, opportunities to use large-scale surveys to generalize and further test the results of their laboratory experiments.

Background and History

While a sample survey of human beings can be of good and measurable accuracy only if it is accomplished through probability sampling (*Sample Surveys: The Field* and *Sample Surveys: Methods*), probability sampling is only the first step in assuring that a survey is successful in producing results accurate enough for use in predicting elections, for gauging public opinion, for measuring the incidence of a disease or the amount of acreage planted in a crop, or for any of the other myriad of important uses that depend on survey data. Questions must be worded not only in ways that are unbiased, but also understandable to respondents and convey the same meaning to respondents as was intended by the survey's author. If the questions refer to facts about the past, they must be presented in ways that help respondents to remember the facts accurately and report truthfully. And interviewers must be able to understand respondents' answers correctly to record and categorize them appropriately.

These and other nonsampling issues have been of concern to survey researchers for many decades (e.g., Payne, 1951). Researchers were puzzled by the fact that changing the wording slightly in a question, sometimes resulted in a different distribution of answers and sometimes did not; worried that sometimes the context of other questions affected answers to a particular question, and sometimes did not; concerned that sometimes respondents were able to remember things accurately and sometimes were not. In the late 1970s and early 1980s, as survey data came to be used ever more extensively for public policy purposes, concerns for the validity of those data became especially widespread.

These concerns led to a cross-disciplinary research endeavor that would bring the insights and methods of the cognitive sciences, especially cognitive psychology, to bear on the problems raised by surveys, and that would encourage researchers in the cognitive sciences to use surveys as a means of testing, broadening, and generalizing their laboratory-based conclusions. Since then, work in the field has incorporated theories and viewpoints from disciplines not usually classified as the cognitive sciences, notably linguistics, conversational analysis, anthropology, and ethnography.

Described below are some of the practical and theoretical achievements of the movement.

These include methodological transfers from the cognitive sciences to survey research, especially so-called cognitive interviewing, the broad establishment of cognitive laboratories in governmental and nongovernmental survey research centers, theoretical formulations that account for many of the mysteries generated by survey responses, and practical applications to the development and administration of ongoing surveys. Some references appear at the close describing these achievements more fully and speculating on the future course of the field. Perhaps surprisingly, there is little to report on the hoped-for use of surveys to generalize the findings of the cognitive sciences, despite the encouragement of such authors as Wright and Loftus (1998).

Methodological Transfers

Much work in the field is rooted in a classic information processing model applied to the cognitive tasks in a survey interview. The model suggests that the tasks the respondent must accomplish, in rough order in which they must be tackled (although respondents go back and forth among them), are comprehension/interpretation, retrieval, judgment, and communication. These models have been intertwined with the earlier models of the survey interview as a social interaction (Sudman and Bradburn, 1974) through the efforts of such researchers as Clark and Schober (1992) and Sudman et al. (1996) to take into account both the social communication and individual cognitive processes that play a part in the survey interview.

Guided by this model, one of the earliest achievements of the movement to study cognitive aspects of survey design was the adoption of some of the methods of the cognitive psychology laboratory for the pretesting of survey questionnaires. In particular, these methods are aimed at insuring that questions are comprehensible to respondents and that the meanings transmitted are those intended by the investigator. They also can point to problems in retrieval and communication. Such cognitive pretesting includes 'think-aloud protocols,' in which specially recruited respondents answer the proposed questionnaire aloud to a researcher, and describe, either concurrently with each question or retrospectively at the end of the questionnaire, their thought processes as they attempt to understand the question and to retrieve or construct an answer. This is an

adaptation of protocol analysis, a technique developed by Ericsson and Simon (1980, 1993). Beatty and Willis (2007) argue that this cognitive interviewing is perhaps the most widespread and important outcome of the CASM movement. (CASM is the original acronym chosen for this movement, standing for cognitive aspects of survey methodology. The acronym was chosen purposefully to pun on the broad chasm researchers in the diverse fields would have to bridge in order to work together.) They offer a review of the range of different types of cognitive interviewing used in practice.

Also used is behavioral coding, based on procedures originally developed by Cannell and coworkers (e.g., Cannell and Oskenberg, 1988), which note questions for which respondents ask for clarification or for which their responses are inadequate. Other methods include reviews of questionnaires by cognitively trained experts, the use of focus groups, and for questionnaires that are automated, measures of response latency. Detailed descriptions of these methods and their applicability to detect problems at various stages of the question answering process appear in Sudman et al. (1996), Schwarz and Sudman (1996), and Forsyth and Lessler (1991).

Interviewer Behavior

Although survey interviewing began with investigators themselves merely going out and having conversations with respondents, as the survey enterprise aspired to greater scientific respectability and as it became sufficiently large scale so that many interviewers were needed for any project, standardized interviewing became the norm. Printed questionnaires were to be followed to the letter; interviewers were instructed to answer respondents' requests for clarification by merely rereading the question or by saying that the meaning of any term is 'whatever it means to you.' While it is probably true that skilled interviewers have always deviated from these rigid standards in the interest of maintaining respondent cooperation and getting valid data, the robotlike interviewer delivering the same 'stimulus' to every respondent was considered the ideal. Working with videotapes of interviews for the National Health Interview Survey and for the General Social Survey and using techniques of conversational analysis, Suchman and Jordan (1990) showed that misunderstandings frequently arose because of the exclusion of normal conversational resources from the survey interview, and that these misunderstandings resulted not only in discomfort on the part of respondents, but also resulted in data that did not meet the needs of the survey designer. Suchman and Jordan recommended a more collaborative interviewing style, in which interviewer and respondent work together to elicit the respondent's story and then fill in the questionnaire. Note that this recommendation applies almost exclusively when the information sought is about the respondent's activities or experiences rather than about his/her attitudes or opinions. Both theoretical and practical advances have followed on these insights.

On a theoretical level, several researchers have used variants of the 'cooperativeness' principle advanced by Grice (1975) to help explain some of the anomalies that have puzzled survey researchers over the decades. Many are understandable as respondents bringing to the survey interview the communicative assumptions and skills that serve well in everyday life. The maxims of the cooperativeness principle require a participant in a conversation to be truthful, relevant, informative, and clear (Sudman et al., 1996: p. 63). Thus, for example, it has long been known that respondents are willing to report on their attitudes toward fictitious issues or nonexistent groups, and investigators have taken this willingness as evidence of the gullibility of respondents and perhaps of the futility of trying to measure attitudes at all. But to 'catch on' to the 'trick' nature of such a question, a respondent would have to assume that the interviewer violated all the maxims of the cooperativeness principle. Parsimony suggests, instead, that the respondent would assume that the interviewer knew what he or she was talking about, and answer accordingly. Several puzzles about context effects, discussed below, also seem understandable in light of Grice's maxims.

On a practical level, freeing interviewers to be more conversational with respondents is not costless; when interviewers explain more, interviews take longer, and are thus more costly in interviewers' wages and respondents' patience. Schober and Conrad (e.g., 1997) have embarked on a program of research that has demonstrated that conversational interviewing generates better data than does standardized interviewing when the mappings of the respondent's experiences onto the survey's concepts are complicated, but indeed at the cost of longer interviews. Perhaps a compromise might be to seek a point between fully standardized and fully conversational interviewing that balances the benefits in accuracy to be gained by a more conversational style with the costs in interview length. Conrad and Schober (e.g., 2008) also explore how endowing an animated figure in a Web-based interview with conversational abilities to prompt the respondent with definitions affects respondent accuracy.

Factual Questions versus Questions about Attitudes or Opinions

The distinction between questions that ask respondents to report on facts, usually autobiographical, and those that ask for an attitude or an opinion has been an important one in survey research. Indeed, the explicit origins of the movement to study cognitive aspects of survey design lie in a desire to improve the accuracy of factual reports from respondents; the extension of concern to attitudinal questions came somewhat later, although such a concern was certainly prefigured by the work reported in Turner and Martin (1984). The distinction remains important for interviewer behavior, in the sense that the kinds of conversational interviewing being investigated are designed to help respondents better comprehend the intent of the question and to recall more accurately their experiences. Interviewer interventions designed for these purposes, useful when the question is aimed at getting a factual report, might well bias the respondent's answer when the aim of the question is to elicit an attitude or opinion. The distinction is also valid in the sense that the cognitive theories about autobiographical memory that have been used to understand problems of the recall of factual information are less applicable to attitudes and opinions, and the cognitive techniques that have been

developed to aid respondent recall for autobiographical questions are not appropriate for the recall of attitudes or opinions. But in important ways the distinction obscures similarities between the two types of questions, as will be discussed in the section on context effects, below.

Improving the Accuracy of Factual Reports

Very often, survey questions ask for the frequency with which a respondent has accomplished a behavior (e.g., visited a doctor) or experienced an event (e.g., had an illness) during a specified time period (usually starting sometime in the past and ending with the date of the interview) called the reference period. Thus, the respondent, having understood what events or experiences should be included in the report, has a twofold task; he or she must recall the events or experiences, and also determine whether those events or experiences fell inside the reference period. Theories from the cognitive sciences about how experiences are stored in autobiographical memory and retrieved therefrom have been marshaled to understand both phases of the respondent's recall task and to improve their accuracy.

In a phenomenon called 'telescoping,' respondents are known often to move events or experiences forward in time, reporting them as falling within the reference period, although in fact they occurred earlier. One of the earliest contributions to the literature on cognitive aspects of survey design was the proposal of a technique dubbed 'landmarking' to control telescoping (Loftus and Marburger, 1983). Here, rather than inquiring about the last 6 months or the last year, a respondent is supplied with (or is asked to supply) a memorable date, and then is asked about events or experiences since that date. (For example, the respondent might be asked to report events since a public event, such as a holiday, or since a private event, such as a birthday.) The related concept of bounding is also useful in controlling telescoping. In a panel survey, in which respondents are interviewed repeatedly at regular intervals, informing them of the events or experiences reported at the last interview will prevent such events or experiences from being placed in the time period between the previous interview and the current one. An analogous technique useable in a single interview is the two-time frame procedure in which the respondent is first asked the report events or experiences during a long reference period (e.g., the last 6 months) and then for a shorter period that is really of interest to the investigator (e.g., the last 2 months). This technique seems both to relieve the pressure on respondents to report socially desirable events or experiences and to stress the interviewer's interest in correct dating, encouraging respondents to strive for greater accuracy. A theoretical explanation of telescoping, taking into account the workings of memory for the storage of timing and the effects of rounding to culturally stereotypical values (e.g., 7, 10, or 30 days), is provided by Hutenlocher et al. (1990).

The task of counting or estimating the number of events or experiences in the reference period can also provide a challenge to the respondent. Issues of frequency, regularity, and similarity influence whether the respondent is likely to actually count the instances or to employ an estimation strategy, as well as influence whether counting or estimation is likely to be more accurate. In general, respondents are more likely to retrieve from memory and count events or experiences if their number is few and to estimate if their number is large. But regular events (e.g., going to the dentist every 6 months) and similar events (e.g., trips to the grocery store) are more likely to be estimated. And investigators have found that estimation (perhaps by retrieving a rate and applying that rate to the length of the reference period) is usually more accurate, than the attempt to retrieve and count regular, similar events or experiences. For a full discussion of these issues, see Sudman et al. (1996).

Context Effects

Context effects occur when the content of the questions preceding a particular question (or even following that question in a self-administered questionnaire in which the respondent is free to return to earlier questions after reading later ones) influence how some respondents answer that question. The understanding of the term has been broadened to include also the effects on the distribution of responses of the response categories offered in a closed-ended question. Investigators have used the maxims of Grice's cooperativeness principle to approach a systematic understanding of these effects, which have seemed mysterious for many years. Full treatments of these issues may be found in Sudman et al. (1996) and Tourangeau (1999); by necessity only some selected findings can be presented here.

A pair of puzzling context effects that have fascinated researchers for years are assimilation effects and contrast effects. An assimilation effect occurs when the content of the preceding items moves the responses to the target item in the direction of the preceding items; a contrast effect occurs when the movement is in the opposite direction. These effects can now be understood through the inclusion/exclusion model of the judgment process in survey responding, proposed by Schwarz and Bless (1992). The model holds that when a respondent has to make a judgment about a target stimulus, he or she must construct a cognitive representation, not only of the target stimulus, but also of a standard with which to compare the target stimulus. To construct these representations, respondents must retrieve information from memory; since not all information in memory can be retrieved, respondents retrieve only that which is most easily accessible. And information can be accessible either because it is 'chronically' accessible or because it is temporarily accessible; it is the temporary accessibility of information that is affected by the context of the question. Positive temporarily accessible information supplied by the context and included in the representation of the target will render the judgment of the target more positive: negative context information thus included will render the judgment more negative. Both these processes produce assimilation effects. Contrast effects are created when the context suggests that information be excluded from the representation of the target, or included in the representation of the standard to which the target is to be compared. For example, Schwarz and Bless (1992) were able to manipulate the frequency of positive evaluations of a political party by using a reference to a popular politician in a preceding question. They were able to create an assimilation effect, resulting in more positive evaluations, by encouraging respondents to include the politician in their representations of the party (by having them

recall that the politician had been a member of the party). They were also able to create a contrast effect, resulting in less positive evaluations of the political party, by encouraging respondents to exclude the politician from their representation of the party (by having them recall that his present position took him out of party politics). The model predicts inclusion of more information when the target is a general question, thus encouraging assimilation effects; it also predicts contrast effects when the target is narrow. Thus, a preceding question about marital happiness seems to result in higher reports of general happiness for those who are happily married, but lower reports of general happiness for those who are unhappy in their marriages. This influence of the preceding question can be eliminated by having respondents see the questions about marital happiness and general happiness as part of a single unit (either by an introduction that links the two questions or by a typographical link on a printed questionnaire); then, in obedience to the conversational maxim that speakers should present new information, respondents exclude their marital happiness from their evaluation of their general happiness.

Similar recourse to conversational maxims helps explain the influence of the response alternatives presented in closed-answer questions on the distribution of responses. For example, Schwarz et al. (1985) found that 37.5% of respondents reported watching television (TV) for more than two and a half hours a day when the response scale ranged from 'up to $2^1/_2$ h' to 'more than $4^1/_2$ h,' but only 16.2% reported watching more than two and a half hours a day when the response scale ranged from 'up to $^1/_2$ hour' to 'more than $2^1/_2$ hours.' Respondents take the response scales to be conversationally informative about what typical TV watching habits are and calibrate their report accordingly.

Some Achievements of the Movement to Study Cognitive Aspects of Surveys

The Impact of Theories

Perhaps the most profound effect of the movement to study cognitive aspects of survey design is the introduction of theory into the research area. There has been space above to present only a few of the many such theoretical advances, but their utility should be clear. For example, consulting the maxims of cooperativeness in conversation gives us a framework for predicting and preventing response effects. In the same sense, the inclusion/exclusion model, much more fully developed than is presented here, makes predictions for what to expect from various manipulations of question ordering and content. These predictions offer guidance for the experimental advance of the research domain, and their testing systematically advances the knowledge, both of the phenomena and of the applicability of the theory.

The Cognitive Laboratories and Their Impact

Cognitive interviews in laboratory settings have become standard practice in several US government agencies (including the National Center for Health Statistics, the Bureau of Labor Statistics, and the Census Bureau), as well as in government agencies around the world, and at many academic and commercial survey organizations. Work in these laboratories supplements the usual large-scale field pretests of surveys. For illustrative purposes, it is worth describing two major efforts at US government agencies.

The Current Population Survey (CPS) is a large-scale government survey, sponsored by the Bureau of Labor Statistics and carried out by the Census Bureau. From interviews with some 60 000 households monthly are derived major government statistical series, including the unemployment rate. In the early 1990s, the CPS underwent its decennial revision, this time using the facilities of the government cognitive laboratories as well as traditional field testing of proposed innovations. In particular, this research found that the ordering and wording of the key questions regarding labor force participation caused some respondents to report themselves as out of the labor force when in fact they were either working part time or looking for work. It also found that the concept of layoff, which the CPS takes to mean temporary furlough from work with a specific date set for return, was not interpreted that way by respondents. Respondents seemed to understand 'on layoff' as a polite term for 'fired.' Appropriate revision of these questions, followed by careful field testing and running the new and old versions in parallel for some months resulted in estimates of the unemployment rate in which policy makers and administrators can have greater confidence. More details can be found in Norwood and Tanur (1994).

The US Decennial Census has always asked for a racial characterization of each resident. Originally designed to distinguish whites (mostly free) from blacks (mostly slaves) for purposes of counting for apportionment (the US Constitution originally mandated that a slave be counted as 3/5 of a man), the uses to which answers to the race identification question was put evolved over time. Civil Rights legislation starting in the 1960s designated protected groups, and it thus became important to know the population of such groups. At the same time, individuals were becoming both more conscious of their racial self-identification and less willing to be pigeonholed into a small set of discrete categories, especially as more individuals were self-identifying with more than one racial group. The 1990 Census asked two questions in this general area – one on racial identity followed by one on Hispanic ethnic identification. Experimental work (Martin et al., 1990) indicated that the ordering of these questions mattered to respondents. For example, when the Hispanic ethnicity question appeared first, fewer people chose 'other' as their racial category. This and other research made a convincing case that the complexity of Americans' racial and ethnic self-identification needed a more complex set of choices on the Census form. A 4-year program of research resulted in the conclusion that allowing respondents to choose as many racial categories as they believe apply to them is a solution superior to the provision of a multiracial category (see Tucker et al., 1996). (For more details on this and other census research see *Censuses: History and Methods*.) The 2000 Decennial Census (and that done in 2010) allowed respondents to choose as many racial categories as they wished; in 2000, approximately 2.4% of the population chose two or more races to identify with.

Conclusions

The movement to study cognitive aspects of surveys has thus given us some theoretical insights and practical approaches. Ideally, the hope of the movement to study cognitive aspects of survey design might well be to develop an overarching theory of the survey designing and responding process, but the field is currently far from that goal and may perhaps never be able to attain it. But the frameworks so far provided offer clear guidance of how to proceed in an effort to reach systematic understanding. Far fuller accounts of the results of the movement to study cognitive aspects of survey design and ideas of future directions can be found in Sirken et al. (1999a,b), Sudman et al. (1996), Tanur (1992), and Tourangeau et al. (2000).

See also: Sample Survey Methodology, History of; Sample Surveys: Model-Based Approaches.

Bibliography

Beatty, P.C., Willis, G.B., 2007. Research synthesis: the practice of cognitive interviewing. Public Opinion Quarterly 71, 287–311.

Cannel, C.F., Oskenberg, L., 1988. Observation of behavior in telephone interviews. In: Groves, R.M., Biemer, P.B., Lyberg, L.E., Massey, J.T., Nichols II, W.L., Waksberg, J. (Eds.), Telephone; Survey Methodology. Wiley, New York.

Clark, H.H., Schober, M.F., 1992. Asking questions and influencing answers. In: Tanur, J. (Ed.), Questions about Questions: Inquiries into the Cognitive Bases of Surveys. Sage, New York.

Conrad, F.G., Schober, M.F., 2008. New frontiers in standardized survey interviewing. In: Hesse-Biber, S.N., Leavy, P. (Eds.), Handbook of Emergent Methods. Guilford, New York.

Ericsson, K.A., Simon, H.A., 1980. Verbal reports as data. Psychological Review 87, 215–251.

Ericsson, K.A., Simon, H.A., 1993. Protocol Analysis: Verbal Reports as Data. MIT, Cambridge, MA.

Forsyth, B.H., Lessler, J., 1991. Cognitive laboratory methods: a taxonomy. In: Biemer, P., Groves, R.M., Lyberg, L.E., Mathiowetz, N., Sudman, S. (Eds.), Measurement Errors in Surveys. Wiley, New York.

Grice, H.P., 1975. Logic and conversation. In: Cole, P., Morgan, J.L. (Eds.), Syntax and Semantics, Speech Acts, vol. 3. Academic Press, New York.

Hippler, H.J., Schwarz, N., Sudman, S. (Eds.), 1987. Social Information Processing and Survey Methodology. Springer-Verlag, New York.

Huttenlocher, J., Hedges, L.V., Bradburn, N.M., 1990. Reports of elapsed time: bounding and rounding processes in estimation. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 196–213.

Loftus, E.F., Marburger, W., 1983. Since the eruption of Mt. St. Helens, did anyone beat you up? Improving the accuracy of retrospective reports with landmark events. Memory and Cognition 11, 114–120.

Martin, E., Demaio, T.J., Campanelli, P.C., 1990. Context effects for census measures of race and Hispanic origin. Public Opinion Quarterly 54, 551–566.

Norwood, J.L., Tanur, J.M., 1994. Measuring unemployment in the nineties. Public Opinion Quarterly 58, 277–294.

Payne, S.L., 1951. The Art of Asking Questions. Princeton University Press, Princeton, NJ.

Schober, M.F., Conrad, F.G., 1997. Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly 61, 576–602.

Schwarz, N., Bless, H., 1992. Constructing reality and its alternatives: assimilation and contrast effects in social judgment. In: Martin, L.L., Tesser, A. (Eds.), The Construction of Social; Judgment. Erlbaum, Hillsdale, NJ.

Schwarz, N., Hippler, H.J., Deutsch, B., Strack, F., 1985. Response categories: effects on behavioral reports and comparative judgments. Public Opinion Quarterly 49, 388–395.

Schwarz, N., Sudman, S. (Eds.), 1996. Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. Jossey-Bass, San Francisco.

Sinaiko, H., Broedling, L.A. (Eds.), 1976. Perspectives on Attitude Assessment Surveys and Their Alternatives. Pendleton, Champaign, IL.

Sirken, M.G., Herrmann, D.J., Schechter, S., Schwarz, N., Tanur, J.M., Tourangeau, R. (Eds.), 1999a. Cognition and Survey Research. Wiley, New York.

Sirken, M.G., Jabine, T., Willis, G., Martin, E., Tucker, C. (Eds.), 1999b. A New Agenda for Interdisciplinary Survey Research Methods: Proceedings of the CASM II Seminar. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

Suchman, L., Jordan, B., 1990. Interactional troubles in face-to-face survey interviews. Journal of the American Statistical Association 85, 232–253. With discussion.

Sudman, S., Bradburn, N.M., 1974. Response Effects in Surveys: A Review and Synthesis. Aldine, Chicago.

Sudman, S., Bradburn, N.M., Schwarz, N., 1996. Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. Jossey-Bass, San Francisco.

Tanur, J.M. (Ed.), 1992. Questions about Questions: Inquiries into the Cognitive Bases of Surveys. Sage, New York.

Tourangeau, R., 1999. Context effects on answers to attitude questions. In: Sirken, M.G., Hermann, D.J., Schwarz, N., Tanur, J.M., Tourangeau, R. (Eds.), Cognition and Survey Research. Wiley, New York.

Tourangeau, R., Rips, L.J., Rasinski, K., 2000. The Psychology of Survey Response. Cambridge University Press, Cambridge.

Tucker, C., McKay, R., Kojetin, B., Harrison, R., de la Puente, M., Stinson, L., Robison, E., 1996. Testing Methods of Collecting Racial and Ethnic Information: Results of the Current Population Survey Supplement on Race and Ethnicity. Bureau of Labor Statistics. Statistical Notes, No. 40.

Turner, C.F., Martin, E. (Eds.), 1984. Surveying Subjective Phenomena. Sage, New York.

Wright, D.B., Loftus, E.F., 1998. How memory research can benefit from CASM. Memory 6, 467–474.

Sample Surveys: Model-Based Approaches

Sharon L Lohr, Arizona State University, Tempe, Arizona, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 20, pp. 13462–13467, © 2001, Elsevier Ltd., with an updated Bibliography setion supplied by the Editor.

Abstract

The theory for sample surveys has been developed using two theoretical frameworks: design-based and model-based. The design-based approach uses the probabilities with which units are selected for the sample for inference; in the model-based approach, the investigator hypothesizes a joint probability distribution for elements in the finite population, and uses that probability distribution for inference. In this article, the two approaches are described and compared, and guidelines are given for when, and how, one should perform a model-based analysis of survey data. The use of models for survey design is discussed briefly.

The theory for sample surveys has been developed using two theoretical frameworks: design-based and model-based. The design-based approach uses the probabilities with which units are selected for the sample for inference; in the model-based approach, the investigator hypothesizes a joint probability distribution for elements in the finite population, and uses that probability distribution for inference. In this article, the two approaches are described and compared, and guidelines are given for when, and how, one should perform a model-based analysis of survey data. The use of models for survey design is discussed briefly.

Inference in Sample Surveys

How does one generalize from individuals in a sample to those not observed? The problem of induction from a sample was much debated by philosophers, social scientists, and mathematicians of the eighteenth and nineteenth centuries, including Immanuel Kant, Charles Peirce, John Venn, and Adolphe Quetelet. In the early years of the twentieth century, many investigators resisted the idea of using survey samples rather than censuses for 'serious statistics' because of inference issues (*see* Sample Survey Methodology, History of). The debates involving official uses of sample surveys in these years resulted in the development of two philosophical frameworks for inference from a sample: design-based inference and model-based inference.

In design-based inference, first expounded systematically by Neyman (1934), the sample design provides the mechanism for inferences about the population. Suppose that a without-replacement probability sample (*see* Sample Survey Methodology, History of) of n units is to be taken from a population of N units. A random variable Z_i is associated with the ith unit in the population; $Z_i = 1$ if the unit is selected for inclusion in the sample, and $Z_i = 0$ if the unit is not selected. The joint probability distribution of $\{Z_1,\ldots, Z_N\}$ is used for inference statements such as confidence intervals (*see* Estimation: Point and Interval).

The quantity being measured on unit i, y_i, is irrelevant for inference in the design-based approach. Whether y_i is household income, years of piano lessons, or number of cockroaches in the kitchen, properties of estimators depend exclusively on properties of the random variables $\{Z_1, \ldots, Z_N\}$ that describe the probability sampling design. The Horvitz–Thompson (1952) estimator,

$$\sum_{i=1}^{N} Z_i y_i / \pi_i \qquad [1]$$

where $\pi_i = P(Z_i = 1)$, is an unbiased estimator of the population total $\sum_{i=1}^{N} y_i$; the variance of the Horvitz–Thompson estimator,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \left(\pi_i \pi_j\right)^{-1} \mathrm{Cov}\left(Z_i, Z_j\right)$$

depends on the covariance structure of $\{Z_1, \ldots, Z_N\}$.

The design-based approach differs from the inferential framework used in most other areas of statistics. There, y_i is the observed value of a random variable Y_i; the joint probability distribution of the Y_i's and proposed stochastic models allow inferential statements to be made. Following questions raised by Godambe (1955) about optimal estimation and survey inference, Brewer (1963) and Royall (1970) suggested that the same model-based frameworks used in other areas of statistics also be used in finite population sampling. Thompson (1997) summarized various approaches to prediction using models.

In the model-based approach to inference, the finite population values are assumed to be generated from a stochastic model. Regression models (*see* Linear Hypothesis: Regression (Basics)) are often adopted for this purpose: if covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$ are known for every unit in the population, a possible model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad [2]$$

where the ε_i's are random variables with mean zero and specified covariance structure. The parameters β_j for $j = 0, \ldots, p$ may be estimated by standard techniques such as generalized least squares, and the regression equation used to predict the value of y for units not in the sample. Then the finite population total is estimated by summing the observed values of the y_i's for units in the sample and the predicted values of the response for units not in the sample.

To illustrate the two approaches, consider a hypothetical survey taken to study mental health status in a population of 10,000 women over age 65. A stratified random sample of 100 urban women and 100 rural women is drawn from a population of 6,000 urban women and 4,000 rural women over age 65. One response of interest is the score on a depression inventory. Let $\bar{y}_U$ denote the sample mean score for the urban women and let $\bar{y}_R$ denote the sample mean score for the rural women.

Under the design-based approach as exposited in Cochran (1977), $\pi_i = 1/60$ if person i is in an urban area and $\pi_i = 1/40$ if person i is in a rural area. Every urban woman in the sample represents herself and 59 other urban women who are in the population but not in the sample; every sampled rural woman represents herself plus 39 other rural women. The Horvitz–Thompson estimator for the mean depression score for the population is $N^{-1}\sum_{i=1}^{N} Z_i y_i/\pi_i$; here, the estimated population mean score under the stratified random sampling design is $0.6\bar{y}_U + 0.4\bar{y}_R$. A 95% confidence interval (CI) for the mean refers to the finite population: if a CI were calculated for each possible sample that could be generated using the sampling design, the selection probabilities of the samples whose CIs include the true value of the population mean depression score sum to 0.95. Inference refers to repeated sampling from the population, not to the particular sample drawn.

Any number of stochastic models might be considered in a model-based approach. Consider first a special case of the regression model in Eqn. [2], and assume that the ε_i's are independent and normally distributed with constant variance. Model 1 has the form $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $x_i = 1$ if person i is an urban resident and $x_i = 0$ if person i is a rural resident. The stochastic model provides a link between sampled and unsampled units: women who are not in the sample are assumed to have the same mean depression score as women with the same urban/rural status who are in the sample. Under Model 1, $\widehat{\beta}_0 = \bar{y}_R$ and $\widehat{\beta}_1 = \bar{y}_U - \bar{y}_R$ are the least squares estimates of β_0 and β_1. Thus the predicted value of y for each urban woman is $\bar{y}_U$, and the predicted value of y for each rural woman is $\bar{y}_R$. The model-based estimate of mean depression in this population is thus

$$\frac{1}{10,000}[100\bar{y}_U + 100\bar{y}_R + 5,900\bar{y}_U + 3,900\bar{y}_R] = 0.6\bar{y}_U + 0.4\bar{y}_R$$

With Model 1, the point and interval estimates of the finite population mean are the same as for the design-based approach. The 95% confidence interval is interpreted differently, however; 95% of CIs from samples with the same values of x_i that could be generated from the model are expected to contain the true value of $\beta_0 + 0.6\beta_1$.

In this example, model-based inference with Model 1 accords with the results from design-based inference. But suppose that the model adopted is Model 2: $Y_i = \mu + \varepsilon_i$. Then the predicted value of the response for all women in the finite population is $\bar{y} = (\bar{y}_U + \bar{y}_R)/2$, and the mean depression score for the finite population is also estimated by $\bar{y}$. If depression is higher among rural women than among urban women, the estimate $\bar{y}$ from Model 2 will likely overestimate the true mean depression score in the population of 10,000 women because rural women are not proportionately represented in the sample.

In the model-based approach, inference is not limited to the 10,000 persons from whom the sample is drawn but applies to any set of persons for whom the model is appropriate. Random selection of units is not required for inference as long as the model assumptions hold. If the model is comprehensive enough to include all important information, as in Model 1 above, then the sampling design is irrelevant and can be disregarded for inference.

The two types of inference differ in the conception of randomness. Design-based inference relies on the randomness involved in sample selection; it relates to the actual finite population, and additional assumptions are needed to extend the inferences to other possible populations. Model-based inference is conditional on the selected sample; the randomness is built into the model, and the model assumptions are used to make inferences about units not in the sample. Design-based inference depends on other possible samples that could have been selected from the finite population but were not, while model-based inference depends on other possible populations that could have been generated under the model but were not.

The approaches are not completely separated in practice. Rao (1997) summarized a conditional design-based approach to inference, in which inference is restricted to a subset of possible samples. Särndal et al. (1992) advocated a *model-assisted* approach, in which a population model inspires the choice of estimator, but inference is based on the sampling design. In the depression example, a model-assisted estimator could incorporate auxiliary information such as race, ethnicity, and marital status through a model of the form in Eqn. [2]; however, the stratified random sampling design is used to calculate estimates and standard errors.

Models in Descriptive and Analytic Uses of Surveys

Estimating a finite population mean or total is an example of a descriptive use of a survey: the characteristics of a particular finite population are of interest. In much social science research, survey data are used for analytic purposes: investigating relationships between factors and testing sociological theories. Data from the US National Crime Victimization Survey may be used to estimate the robbery rate in 1999 (descriptive), or they may be used to investigate a hypothesized relationship between routine activities and likelihood of victimization (analytic). In the former case, the population of inference is definite and conceivably measurable through a census. In the latter, the population of inference is conceptual; the investigator may well be interested in predicting the likelihood of victimization of a future person with given demographic and routine activity variables.

Smith (1994) argued that design-based inference is the appropriate paradigm for official descriptive statistics based on probability samples. Part of his justification for this position was the work of Hansen et al. (1983), who provided an example in which small deviations from an assumed model led to large biases in inference. Brewer (1999) summarized work on design-based and model-based estimation for estimating

population totals and concluded that a model-assisted generalized regression estimator (see Särndal et al. 1992), used with design-based inference, captures the best features of both approaches. Models must of course always be used for inference in nonprobability samples; they may also be desirable in probability samples that are too small to allow the central limit theorem to be applied for inference.

Lohr (1999, Chap. 11) distinguished between obtaining descriptive official statistics and uncovering a 'universal truth' in an analytic use of a survey. Returning to the depression example, the investigator might be interested in the relationship between depression score (y) and variables such as marital status, financial resources, number of chronic health problems, and ability to care for oneself. In this case, the investigator would be interested in testing a theory that would be assumed to hold not just for the particular population of 10,000 women but for other populations as well, and should be making inferential statements about the βs in model (2). The quantity for inference in the design-based setting is b_p, the least squares estimate of β that would be obtained if the x_is and y_i were known for all 10,000 persons in the finite population. The quantity b_p would rarely be of primary interest to the investigator, though, since it is merely a summary statistic for this particular finite population. In social research, models are generally motivated by theories, and a model-based analysis allows these theories to be tested empirically.

The generalized least squares estimator of β, $\widehat{\beta}_{LS}$, would be the estimator of choice under a pure model-based approach because of its optimality properties under the proposed model. This estimator is, however, sensitive to model misspecification. An alternative, which achieves a degree of robustness to the model at the expense of a possibly higher variance, is to use the design-based estimator of b_p. If the proposed stochastic model is indeed generating the finite population and if certain regularity conditions are met, an estimator that is consistent for estimating b_p will also be consistent for estimating β. Under this scenario, a design-based estimate for b_p also estimates the quantity of primary interest β and has the advantage of being less sensitive to model misspecification.

Regardless of philosophical differences on other matters of inference, it is generally agreed that two aspects of descriptive statistics require the use of models. All methods currently used to adjust for nonresponse (*see* Nonsampling Errors) employ models to relate nonrespondents to respondents, although the models are not necessarily testable. In small area estimation, sample sizes in some subpopulations of interest are too small to allow estimates of sufficient precision; models are used to relate such subpopulations to similar subpopulations and to useful covariates.

Models for Small Area Estimation

In small area estimation, a model is used to estimate the response in subpopulations with few or no sample observations. As an example, the US Current Population Survey (CPS) provides accurate statistics about income and poverty for the nation as a whole. It was not designed, though, to provide accurate estimates in domains such as states, counties, or school districts – the sample would have to be prohibitively large in order to provide precise estimates of poverty for every county in the USA. These domains are called small areas – the term 'small' does not refer to the size of the area or the population, but to the fact that the sample size in the domain is small or may even be zero.

Consider the states to be the small areas, and let y_k be the proportion of school-age children who are poor in state k. The direct estimate $\bar{y}_k$ of y_k is calculated using data exclusively from the CPS, and $\widehat{V}(\bar{y}_k)$ is an estimate of the variance of $\bar{y}_k$. Since in some states $\widehat{V}(\bar{y}_k)$ is unacceptably large, the current practice for estimating poverty at the state level (see National Research Council, 2000, p. 49) uses auxiliary information from tax returns, food stamp programs, and the decennial census to supplement the data from the CPS. A regression model for predicting y_k using auxiliary information gives predicted values

$$\widehat{y}_k = \widehat{\beta}_0 + \sum_{j=1}^{4} \widehat{\beta}_j x_{jk}$$

where the x_{jk} s represent covariates for state k (e.g., x_{1k} is the proportion of child exemptions reported by families in poverty in state k, and x_{2k} is the proportion of people receiving food stamps in state k). The predicted value $\hat{y}_k$ from the regression equation is combined with the direct estimate $\bar{y}_k$ from the CPS according to the relative amounts of information present in each: the small area estimate for state k is

$$\tilde{y}_k = \gamma_k \bar{y}_k + (1 - \gamma_k)\widehat{y}_k$$

where γ_k is determined by the relative precision of $\bar{y}_k$ and $\hat{y}_k$. If the direct estimate is precise for a state, i.e., $\widehat{V}(\bar{y}_k)$ is small, then γ_k is close to one and the small area estimate $\widetilde{y_k}$ relies mostly on the direct estimate. Conversely, if the CPS contains little information about state k's poverty rate, then γ_k is close to zero and $\widetilde{y_k}$ relies mostly on the predicted value from the regression model.

The small area model allows the estimator for area k to 'borrow strength' from other areas and incorporate auxiliary information from administrative data or other sources. Ghosh and Rao (1994) and Rao (1999) review properties of this model and other models used in small area estimation.

Performing a Model-Based Analysis

The first step in a model-based analysis for either descriptive or analytical use is to propose and fit a model to the data. Dependence among units, such as dependence among children in the same school, can be treated using hierarchical linear models or other methods discussed in Skinner et al. (1989).

The biggest concern in a model-based analysis, as pointed out by Hansen et al. (1983), is that the model may be misspecified. Many, but not all, of the assumptions implicit in a model can be checked using the sample data. Appropriate plots of the data provide some graphical checks of model adequacy and correctness of the assumed variance structure, as described in Lohr (1999). These assumptions can also be partially checked by performing hypothesis tests of nested models, and by fitting alternative models to the data. In the depression example, plotting the data separately for rural and

urban residents would reveal inadequacy of Model 2 relative to Model 1.

Another method that can sometimes detect model inadequacy is comparison of design-based and model-based estimates of model parameters. As mentioned in Sect. 2, if the model is correct for units in the finite population, then the design-based estimates and the model-based estimates should both be consistent for the model parameters. A substantial difference in the estimates could indicate that the sample design contains information not captured in the model, and that perhaps more covariates are needed in the model.

One crucial assumption that cannot be checked using sample data is that the model describes units not n the sample. This assumption is especially important in nonprobability samples and in use of models for nonresponse adjustment or small area estimation.

Models in Survey Design

Kalton (1983) distinguished between the use of models in survey analysis and in survey design, stating that 'the use of models to guide the choice of sample design is well-established and noncontroversial.' In good survey practice, a stratified sampling design is often chosen because it is thought that there are differences among stratum means. An unequal probability design may be employed because of prior belief that large counties have more variability in total number of crime victimizations than small counties; models provide a mechanism for formalizing some of the knowledge about population structure and exploring results of alternative assumptions. Cochran (1977) illustrated the use of models for designing systematic samples. Särndal et al. (1992, Chap. 12) summarized research on optimal survey design, in which auxiliary information about the population is used to select a design that minimizes the anticipated variance of an estimator under the model and design.

Models used for design purposes do not affect the validity of estimates in design-based inference. A poor model adopted while designing a probability sample may lead to larger variance of design-based estimates, but the estimates will retain properties such as unbiasedness under repeated sampling. A good model at the design stage often leads to a design with greatly increased efficiency.

A model-based analysis can be conducted on data from any sample, probability or non-probability. Probability sampling is in theory unnecessary from a pure model-based perspective, and Thompson (1997) and Brewer (1999) concluded that certain forms of purposive non-probability sampling can be superior to probability sampling when a model-based analysis is to be conducted and the model is correct. In practice, however, there is always concern that the assumed model may miss salient features of the population, and probability sampling provides some protection against this concern. For a ratio model, with $Y_i = \beta x_i + \varepsilon_i$ and $V[\varepsilon_i] = \sigma^2 x_i$, the model-based optimal design specifies a purposive sample of the population units with the largest x values. Such a design does not allow an investigator to check whether the model is appropriate for small xs; an unequal probability sample with π_i proportional to x_i does allow such model checking, and allows inferences under either design- or model-based frameworks. As Brewer (1999) pointed out, there is widespread public perception that 'randomized sampling is fair' and that perception provides a powerful argument for using probability sampling for official statistics.

The following sources are useful for further exploration of modes of inference in sample surveys. Lohr (1999, Chap. 11) provides a more detailed heuristic discussion of the role of models in survey sampling; Thompson (1997) gives a more mathematical treatment. The articles by Smith (1994), Rao (1997), and Brewer (1999) contrast inferential philosophies, discuss appropriate use of models in analysis of survey data, and provide additional references.

See also: Sample Survey Methodology, History of; Sample Surveys: Cognitive Aspects of Survey Design.

Bibliography

Brewer, K.R.W., 1963. Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. Australian Journal of Statistics 5, 93–105.

Brewer, K.R.W., 1999. Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. International Statistical Review 67, 35–47.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley, New York.

Ghosh, M., Rao, J.N.K., 1994. Small area estimation: An appraisal. Statistical Science 9, 55–76.

Godambe, V.P., 1955. A unified theory of sampling from finite populations. Journal of the Royal Statistical Society B 17, 269–278.

Hansen, M.H., Madow, W.G., Tepping, B.J., 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association 78, 776–793.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

Kalton, G., 1983. Models in the practice of survey sampling. International Statistical Review 51, 175–188.

Lohr, S.L., 1999. Sampling: Design and Analysis. Duxbury Press, Pacific Grove, CA.

National Research Council, 2000. Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. In: Panel on Estimates of Poverty for Small Geographic Areas. Committee on National Statistics. National Academy Press, Washington, DC.

Neyman, J., 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97, 558–606.

Rao, J.N.K., 1997. Developments in sample survey theory: an appraisal. Canadian Journal of Statistics 25, 1–21.

Rao, J.N.K., 1999. Some recent advances in model-based small area estimation. Survey Methology 25, 175–186.

Royall, R.M., 1970. On finite population sampling theory under certain linear regression models. Biometrika 57, 377–387.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Skinner, C.J., Holt, D., Smith, T.M.F., 1989. Analysis of Complex Surveys. Wiley, New York.

Smith, T.M.F., 1994. Sample surveys 1975–1990; an age of reconciliation? International Statistical Review 62, 5–34.

Thompson, M.E., 1997. Theory of Sample Surveys. Chapman & Hall, London.

Särndal, C.E., Swensson, B., Wretman, J., 2003. Model assisted survey sampling. Springer, New York.

Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R.B., Marques, T.A., Burnham, K.P., 2010. Distance Software: Design and Analysis of Distance Sampling Surveys for Estimating Population Size. Journal of Applied Ecology 47 (1), 5–14.

Lumley, T., 2004. Analysis of complex survey samples. Journal of Statistical Software 9 (1), 1–19.

Sanctions in Political Science

David A Baldwin, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A sanction is an action by one actor (A) intended to affect the behavior of another actor (B) by enhancing or reducing the values available to B. Influence attempts by A using actual or threatened punishments of B are instances of negative sanctions. Influence attempts by A using actual or promised rewards to B are instances of positive sanctions. Until the mid-twentieth century, sanctions were viewed primarily as mechanisms for enforcing societal norms, including those embedded in laws. Although this normative view of sanctions continues in legal theory, ethics, political theory, and sociology, a broader usage has emerged in political science during the latter part of the twentieth century. This broader usage depicts sanctions as potentially relevant to any type of influence attempt. This usage is typically found in discussions of influence and power. The subfield of political science that has devoted the most attention to sanctions in recent years is international relations and foreign policy. Most of this research has focused on the use of economic sanctions as foreign policy tools. Here, as elsewhere in political science, more attention has been devoted to the negative sanctions than to the positive ones.

A sanction is an action by one actor (A) intended to affect the behavior of another actor (B) by enhancing or reducing the values available to B. Influence attempts by A using actual or threatened punishments of B are instances of negative sanctions. Influence attempts by A using actual or promised rewards to B are instances of positive sanctions. Not all influence attempts involve sanctions. Actor A may influence actor B by reason, for example, or the provision of information without the use of sanctions.

Concepts

Although the definitions of positive and negative sanctions may appear simple, there are both conceptual and empirical difficulties in distinguishing between the two. Some things take the form of positive sanctions, but actually are not; for example, giving a bonus of $100 to an employee who expected a bonus of $500, or promising not to kill a person who never expected to be killed in the first place. Likewise, some things take the form of negative sanctions, but actually are not; for example, a threat to reduce the salary of a person who expected to be fired or the beating of a masochist.

Is withholding a reward ever a punishment? Always a punishment? Is withholding a punishment ever a reward? Always a reward? The answers depend on Actor B's perception of the situation. In order to distinguish rewards from punishments one must establish B's baseline of expectations at the moment when A's influence attempt begins (Blau, 1986). This baseline is defined in terms of B's expected future value position; that is, expectations about B's future position relative to the things B values. Positive sanctions, then, are actual or promised improvements in B's value position relative to B's baseline of expectations, and negative sanctions are actual or threatened deprivations relative to the same baseline.

The determination of B's baseline of expectations is complicated by "prospect theory" (Kahneman and Tversky, 1979). The prospect theory suggests that B is likely to prefer to avoid losses of $100 compared to failure to gain $100 even when the objective probabilities are the same. This psychological tendency must be taken into consideration in estimating B's baseline of expectations.

Historical Context

Although references to both positive and negative sanctions can be traced to the seventeenth century, the most common usage of the term 'sanctions' refers to negative sanctions. Until the mid-twentieth century, sanctions were viewed primarily as mechanisms for enforcing societal norms, including those embodied in laws. Although this normative view of sanctions continues in legal theory, ethics, political theory, and sociology (Barry, 1995; Coleman, 1990; Waldron, 1994), a broader usage has emerged in political science during the latter part of the twentieth century. This broader usage depicts sanctions as potentially relevant to any type of influence attempt, regardless of whether it is aimed at enforcing social norms or not. This usage is typically found in discussions of influence and power (Baldwin, 1971, 1989; Blau, 1986; Oppenheim, 1981; Lasswell and Kaplan, 1950).

Kinds of Sanctions

Because sanctions are defined in terms of the values of the target of an influence attempt (actor B), they may take a variety of forms. Lasswell and Kaplan (1950) identified eight exemplary values that could serve as the basis for positive or negative sanctions: respect, power, wealth, rectitude, physical well-being, affection, skill, and enlightenment.

Although it is sometimes suggested that positive and negative sanctions are opposites in the sense that generalizations about one type are equally applicable to the other, *mutatis mutandis*, this is often not true. Listed below are some of the hypothesized differences between positive and negative sanctions.

A's Burden of Response

When A's influence attempt is based on a promise, B's compliance obligates A to respond with a reward; whereas B's failure to comply calls for no further response from A.

Role of Costs

One important consequence of the asymmetry between positive and negative sanctions is that promises tend to cost more

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93126-X

when they succeed, while threats tend to cost more when they fail (Boulding, 1989; Parsons, 1963; Schelling, 1960). The difference can be summarized as follows: The bigger the threat, the higher the probability of success; the higher the probability of success, the less the probability of having to implement the threat; the less the probability of having to implement the threat, the cheaper it is to make big threats. The bigger the promise, the higher the probability of success; the higher the probability of success, the higher the probability of having to implement the promise; the higher the probability of having to implement the promise, the more expensive it is to make big promises (*ceteris paribus*).

Indicators of Success

While a successful threat requires no action by A, a successful promise obligates A to implement that sanction. In a well-integrated social system, where the probability that B will comply with A's wishes is relatively high, promises will be more visible than threats. Indeed, it is precisely because threats are so successful in domestic politics that they are so difficult to detect. Since most citizens obey most laws most of the time, threats to punish lawbreakers have to be carried out with respect to only a small minority of citizens in most polities.

Efficacy

The question of the relative efficacy of positive versus negative sanctions in exercising influence has been much debated. Parsons (1963) has argued that negative sanctions have more intrinsic effectiveness than positive sanctions in deterrence situations. And ever since Machiavelli and Hobbes, there have been those who argue that force is the ultimate influence technique. Identifying the conditions under which one type of sanction is more effective than another is likely to continue as a focus of research in political science as well as in other social sciences.

Legitimation

It is usually easier to legitimize demands based on positive sanctions than demands based on negative ones. An example is provided by one of the most important social institutions in the world – the institution of private property. Most societies have one set of rules specifying the conditions under which a person may be deprived of property and quite a different set of rules specifying conditions under which one person may augment another person's property.

Foreign Policy Sanctions

Political scientists interested in international relations and foreign policy have devoted a great deal of research to sanctions during the last two decades of the twentieth century. Most of this research has focused on economic sanctions, with comparatively little attention devoted to military, diplomatic, or other noneconomic forms of sanctions. Although progress has occurred with respect to systematic gathering of empirical data (Hufbauer et al., 2007), refinement of concepts (Baldwin, 1985, 2000), and development of theories, social scientific research on the role of sanctions in international relations is still in its infancy.

Methodological Problems

Five methodological problems provide useful foci for future research:

1. *Lack of agreement on terms and concepts*. While some scholars conceive of economic sanctions in terms of mechanisms for making any type of influence attempt (Baldwin, 1985), others define economic sanctions in terms of particular policy goals (Pape, 1997). This lack of agreement on the basic concepts not only impedes debate within the field of foreign policy studies, but is also an obstacle to cross-fertilization between the study of economic sanctions in foreign policy and discussions of other kinds of sanctions in other areas of social science.
2. *Preoccupation with economic sanctions*. Research on economic sanctions has been largely insulated from research on other techniques of statecraft. This insulation sometimes gives the impression that economic sanctions are *sui generis*, rather than treating them as a particular type of statecraft. Although actual or threatened military force is a military sanction, it is also usually viewed as a unique focus of inquiry. In order to understand when and why policy makers use economic sanctions, it is necessary to understand the alternatives from which the policy makers choose. These alternatives include diplomatic, military, and/or verbal sanctions (propaganda). Future research on sanctions might usefully focus on comparative studies of alternative types of sanctions and the conditions under which each type is most likely to be cost-effective.
3. *Neglect of positive sanctions*. Research on sanctions is heavily skewed toward negative sanctions. This is true not only of research by international relations scholars, but also of political science in general. During the last two decades, research on positive sanctions has increased (Davis, 2000; Kahler and Kastner, 2006; Newnham, 2000; Nincic, 2010; Solingen, 2012), but the emphasis is still disproportionately on negative sanctions.
4. *Lack of agreement on criteria of success*. Scholars disagree as to which criteria should be used in measuring the success of economic sanctions. Some consider only the effectiveness of the influence attempt in achieving its goals (Morgan and Schwebach, 1997; Pape, 1997). Others argue that estimates of success should take into consideration the costs of the sanctions to the user, the costs for noncompliance inflicted on the target, the difficulty of the undertaking, and the comparative utility of alternative policy options (Baldwin, 1985, 2000). This lack of agreement on how to conceive of success is not peculiar to the international relations literature. Unlike economics, political science lacks a standardized measure of value in terms of which success can be measured. There is much less agreement among political scientists as to what constitutes political success than there is among economists as to what constitutes economic success (Baldwin, 1989). Developing a set of

agreed-upon criteria of success applicable to noneconomic, as well as economic, sanctions would be a valuable step in sanctions research (Johnson and Tierney, 2006).

5. *Premature generalization*. Research on economic sanctions by international relations scholars frequently leads to premature attempts to specify policy implications. Some scholars, for example, imply that only foolish policy makers would use such policy instruments (Morgan and Schwebach, 1997; Tsebelis, 1990). In order to make judgments as to the wisdom of using economic sanctions in a given situation, however, one would need to know how effective such sanctions were likely to be, with respect to which goals and targets, at what cost, and in comparison with which policy alternatives (Baldwin, 2000). Current research on economic sanctions rarely asks such questions.

See also: Consequentialism (Including Utilitarianism); Control, Social; Deterrence; Efficacy, Political; Game Theory; Law: Economics of Its Public Enforcement; Legal Culture and Legal Consciousness; National Security Studies and War Potential of Nations; Norms; Power; Prospect Theory; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects.

Bibliography

Baldwin, D.A., 1971. The power of positive sanctions. World Politics 24, 19–38.
Baldwin, D.A., 1985. Economic Statecraft. Princeton University Press, Princeton, NJ.
Baldwin, D.A., 1989. Paradoxes of Power. Basil Blackwell, New York.
Baldwin, D.A., 2000. The sanctions debate and the logic of choice. International Security 24, 80–107.
Barry, B., 1995. Justice as Impartiality. Clarendon Press, Oxford, UK.
Blau, P.M., 1986. Exchange and Power in Social Life. Transaction, New Brunswick, NJ.
Boulding, K.E., 1989. Three Faces of Power. Sage, Newbury, CA.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap, Cambridge, MA.
Davis, J.W., 2000. Threats and Promises: The Pursuit of International Influence. The Johns Hopkins University Press, Baltimore, MD.
Hufbauer, G.C., Schott, J.J., Elliott, K.A., Oegg, B., 2007. Economic Sanctions Reconsidered, third ed. Institute for International Economics, Washington, DC.
Johnson, D.D.P., Tierney, D., 2006. Failing to Win: Perceptions of Victory and Defeat in International Politics. Harvard University Press, Cambridge, MA.
Kahler, M., Kastner, S.L., 2006. Strategic uses of economic interdependence. Journal of Peace Research 43, 523–541.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291.
Lasswell, H.D., Kaplan, A., 1950. Power and Society: A Framework for Political Inquiry. Yale University Press, New Haven, CT.
Morgan, T.C., Schwebach, V.L., 1997. Fools suffer gladly: the use of economic sanctions in international crises. International Studies Quarterly 41, 27–50.
Newnham, R.E., 2000. More flies with honey: positive economic linkage in German Ostpolitik from Bismarck to Kohl. International Studies Quarterly 44, 73–96.
Nincic, M., 2010. Getting what you want: positive inducements in international relations. International Security 35, 138–183.
Oppenheim, F.E., 1981. Political Concepts: A Reconstruction. University of Chicago Press, Chicago.
Pape, R.A., 1997. Why economic sanctions do not work. International Security 22, 90–136.
Parsons, T., 1963. On the concept of political power. Proceedings of the American Philosophical Society 107, 232–262.
Solingen, E. (Ed.), 2012. Sanctions, Statecraft, and Nuclear Proliferation. Cambridge University Press, New York.
Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.
Tsebelis, G., 1990. Are sanctions effective? A game-theoretic analysis. Journal of Conflict Resolution 34, 3–28.
Waldron, J., 1994. Kagan on requirements: Mill on sanctions. Ethics 104, 310–324.

Sandel, Michael (1953–)

Andrew Tsz Wan Hung, Hong Kong Community College, The Hong Kong Polytechnic University, Kowloon, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Michael Sandel, a prominent communitarian philosopher, is famous in his criticism of John Rawls' *A Theory of Justice* and his lively teaching skill demonstrated in the Harvard course 'Justice.' He criticizes Rawls' liberalism for assuming a notion of an unencumbered self, which is not only in tension with his principles of justice, but also denying the human capability of deep evaluation on moral good thus discouraging the public deliberation of morality. By his historical retrieval, Sandel shows how the tradition of civic republicanism has been gradually replaced by liberalism in the American political history. The triumph of liberalism might be due to mass migration, the globalizing economy and the culture of consumerism, in which commodification of nearly everything and the unrestraint use of bioengineering has finally crowded out morality and corrupting human values that are important to republican self-government. In the face of modernity, Sandel calls for reviving civic republicanism with sovereignty diffused into multiplicity of political communities.

Biography

Michael J. Sandel is the son of a record salesman. He grew up at Minneapolis, Minnesota. At the age of 13, he moved with his family to Los Angeles. He studied at Palisades High School and became the president of his senior class in high school in 1971.

He was brought up in a Jewish family, and he also raises his children in a Jewish tradition. For Sandel, religion offers a youth framework for moral investigation and leads them to think about the life beyond materialism, even though his religious viewpoints are not palpable in his writings.

In 1975, he was graduated Phi Beta Kappa from Brandeis University in Massachusetts. While still a university student, he was an intern at the *Houston Chronicle* in Washington, DC and as there was a problem in staffing, he got the chance to cover the Watergate hearings. The experience of being a reporter is so important to him. He almost became a political journalist. Nevertheless, the proximity to those events finally motivated him to move the public debate by doing philosophy.

In 1981, he received a doctorate as a Rhodes Scholar from Balliol College, Oxford. At first, he intended to study welfare economics in Oxford. In the mid-1970s, he went to southern Spain with an economist. They were going to try to develop in a paper an economic model, which could integrate with the concern of equality. Meanwhile, one of his tutors at Oxford, Alan Montefiore, advised him to investigate in theoretical grounding of the political economy, and suggested him to take some philosophical books with him to Spain. He ended up taking four: John Rawls's *A Theory of Justice* (1971, *TJ*), Kant's *Critique of Pure Reason* (1997), Robert Nozick's *Anarchy, State and Utopia* (1974), and Hannah Arendt's *The Human Condition* (1958). He spent the days reading the books and worked with his economist friend on the paper in the evenings. Finally, those philosophical books he read have wrecked the economic project. Sandel thought that the contemporary science of economics has inherited too much from utilitarianism. It was too pessimistic about human nature, conceiving humans as simply bundles of preferences and desires. His reading from Rawls and Kant has offered him a convincing critique of utilitarianism. Nevertheless, Sandel also found an important fault in Rawls' conception of society and human nature. At the time, Sandel met Charles Taylor, a professor in political philosophy in Oxford, who finally became a friend and mentor to Sandel. Taylor encouraged him to develop his critique of Rawls.

In 1982, Sandel's first book, *Liberalism and the Limits of Justice* (*LLJ*), based on his Oxford thesis, was published. This book was a substantial criticism on Rawls' model of liberalism set out in *TJ*. His second book, *Democracy's Discontent: America in Search of a Public Philosophy*, was published in 1996.

In 1980, he has gotten a teaching post at Harvard University as an assistant professor, later as an associate professor in 1983, and, in 1999, was named a Harvard College Professor in a special recognition of his teaching skills. He is currently the Anne T. and Robert M. Bass Professor of Government at Harvard University. He teaches ethics, market and morality, and contemporary political philosophy. His most famous undergraduate course is 'Justice,' which also comes as a book, *Justice: What's the Right Thing to Do?* (2009), and the sourcebook of readings *Justice: A Reader* (2007).

In 1998, he presented the Tanner Lectures on Human Values at Oxford University. In these lectures, he gave an initial formulation of the book *What Money Can't Buy*. In 2009, he delivered the BBC Reith Lectures on 'A New Citizenship' that discussed the vision for what Sandel calls 'A New Politics of the Common Good,' two of the four still dealt with markets and morals.

From 2002 to 2005, Sandel worked for the Bush administration's President's Council on Bioethics, which focused on the issue of cloning at that time. The research formed the basis of another book, *The Case against Perfection: Ethics in the Age of Genetic Engineering*. Sandel's other works include *Public Philosophy: Essays on Morality in Politics*, and articles in scholarly journals, law reviews, and other publications. He has also edited *Liberalism and Its Critics*.

Main Writings

Liberalism and the Limits of Justice

In this book, Sandel argues that Rawls' *TJ*, as a kind of reformulation of deontological liberalism, has failed. Inherited from Kant's philosophy, deontological liberalism upholds the

notion of priority of right over good, which is based on the notion of the autonomous subject. The foundation of deontology is Kant's transcendental argument of a notion of subject, in which the self is antecedent to any particular experience and human conditions. Nevertheless, for Rawls, Kant's disembodied subject is inappropriate to be the determinate principles of justice, or else it cannot be applied to actual human beings in the legislation based on such a subject. Therefore, Rawls attempts to reformulate and preserve Kant's deontological teaching by replacing a "hypothetical choice situation (the original position), characterized by conditions meant to yield a determinate outcome fit for actual human beings" (p. 14).

Sandel's analysis shows that Rawls, in setting the foundation of the primacy of justice, endorses his idea of the original position to satisfy two conflicting requirements. On the one hand, standard of evaluating justice must be prior to the current desires to avoid being contingent; on the other, the foundation comes from a thin theory of the good related to actual human desires rather than totally detached from the real world. However, on the one hand, Sandel criticizes Rawls thin theory of the good for being too thick to satisfy Kantian conception of autonomy. On the other, Rawls' assumption of the subject as mutually disinterested has already provided an assumption of human motivation, which is not universal. For Rawls, the assumption is supposed to offer an empirical precondition to affirm the priority of justice unconditionally. However, Sandel argues, we can imagine cases of more intimate, solidaristic social institution, such as family or religious cultural communities, in which the significance of justice is much less prevailing.

According to Sandel, Rawls' theory includes two conceptions of the self. The first is the self as an unencumbered subject with possession, in which the self is antecedently individuated, and is related to its end by choice but without being detached altogether. The second is the priority of plurality over unity in its relation with others. However, Rawls' two conceptions of the self, Sandel argues, are in conflict with his demands of justice.

First, regarding to his difference principle, Rawls has assumed individuals' natural endowments as common assets, which has to be shared and arranged so that the least advantaged may share in the resources of the fortunate. However, Rawls' such notion of endowment is criticized by Nozick as contradicting to deontological liberalism, which emphasizes the inviolability of individual and distinction between persons; it has neglected the difference among persons. With the assumption of the unencumbered self, Rawls may defend by separating their attributes from the selves, so that even the redistribution of attributes, not the person, does not violate the difference among persons. However, such separation leaves us with a subject taken away all empirically identifiable characteristic that resembles Kantian transcendental self, which Rawls tries to avoid. Indeed, Sandel does not reject Rawls' egalitarian tendency of his theory. Instead, he argues that Rawls can defend his notion of endowment, if he appeals to an intersubjective conception of the self, arguing that those who share in my assets are not understood as 'others,' rather as a 'we,' which implies the existence of community in the constitutive sense. One has to see that he/she is brought up by the community, is part of the community, and completed by the community. In this case, we come to see ourselves, "less as individuated subjects with certain things in common, and more as members of a wider subjectivity, less as 'others' and more as participants in a common identity, be it a family or community or class or people or nation" (p. 143). However, such intersubjective conception of the self is rejected by Rawls' antecedently individuated self assumed in his deontological ethics. Indeed, without the constitutive sense of community, Sandel argues, we also cannot explain why the community as a whole deserves the assets of the successful persons in its province and the benefits that flow from them.

The second conflict is related to Rawls' contractarian argument in which the principles of justice are based on the agreement in the original position. There are two important elements in the original position. The first is choice. The second is the need to resolve conflicting claims among plurality of human subjects. Nevertheless, in the real contract, the resolution of confliction inevitably involves the problem of contingency and conventionalism, so the agreement achieved is never unconditional. For Rawls, the original position can overcome the contingency and conventionalism, because behind the veil of ignorance, different parties are assumed to be similarly situated in all relevant respects, differences of power and knowledge disappear; no one is able to choose on the basis of contingently given attributes, so ideal of autonomy is fulfilled. Any agreement reached in the original position becomes a case of pure procedural justice; and the outcome would therefore be just. However, Sandel questions the validity of these provisions in the original position for the choosers. With the veil of ignorance, everyone is similarly situated. People are guaranteed to reason in the same way. There is no difference in the interests or knowledge and so no basis for bargaining and discussion. In such situation, people are not really free to choose, but to choose under certain constrains with an assumed single correct conception of justice. The assumption of plurality and voluntary is called into question.

Furthermore, in Rawls' theory, the conception of good is determined by the consequence for realization of ends and assessing the relative intensity of agent's needs and desires. Sandel criticizes Rawls' subjectivity for not allowing individuals to have deep evaluation of desires and reflection of their identity because Rawls' subject is antecedently given and barren of constituent traits; there is nothing in the self for reflection, and there is no ground to assess the desirability. So the good is not determined by values, but by one's circumstance and arbitrary desires. Instead, Sandel asserts that our agency is capable of a deeper introspection than merely knowing our needs and desires. Nevertheless, if we want to be able to have a more thoroughgoing reflection, we cannot be wholly unencumbered subjects of possession. Rather, we must be subjects constituted, at least partly, by our central attachments and aspirations, always open to progress and renovation in the light of revised self-understandings. As self-understandings are partly constituted by our belonging community, the deepening of the community will diminished the priority of justice. For Sandel, if utilitarianism fails to take seriously our distinctness, Rawls' justice as fairness fails to take seriously our commonality.

In face of communitarian challenge, Rawls' later proposed the idea of political liberalism, which is political, not metaphysical, in nature, to rescue the priority of the right from

criticisms about the nature of the self. However, Sandel finds such rescue as rendering it vulnerable to other criticisms.

First, when grave moral questions are concerned, it is not always appropriate to bracket those values for political purposes. Sandel gives the example of controversy of abortion, whether fetus is human in the case of abortion is critical to the issues. If Catholic's doctrine is right and the fetus is really human, then abortion is a kind of murder. It is not reasonable to bracket the morality in order to attain political agreement.

Second, the case for the priority of the right over the good in his political liberalism is based on the situation that there exists reasonable pluralism about the good in modern democratic societies. However, the fact of reasonable pluralism also applies to the controversies of justice in democratic societies. If Rawls believes that we can reason about justice, it is unreasonable to reject the deliberation of good in the same way.

Third, political liberalism discourages the kind of public deliberation important to examine the plausibility of competing comprehensive moralities. For Rawls, public deliberation has to be limited by the use of public reason, in which political discourse should not refer to comprehensive moral and religious conceptions. However, Sandel reminds us that the argument for the immediate emancipation of the slaves in the abolitionist movement was based on religious reasons. Insisting on liberal public reason might have great moral cost, if political conception of justice approves toleration of grave moral problems, such as slavery or abortion. It might also have political cost. As we can see, political discourse nowadays is lacking sufficient moral discussion, and becomes increasingly preoccupied with the scandalous, sensational media talk shows. It creates a moral vacuum that opens the way for the fanatical, the prejudiced, and other misguided moralism.

Democracy's Discontent: America in Search of a Public Philosophy

In this book, Sandel contrasts the tradition of civic republicanism with that of procedural liberalism in the US political history. The presentation is organized as the intertwining of philosophical and mostly historical analyses. Philosophically, based on *LLJ*, Sandel continuous his criticism of liberalism and argues for the idea of civic republicanism with the sense of multiply situated selves. Historically, Sandel shows, while both procedural liberalism and civic republicanism used to be present throughout American politics, American political discourse, in the recent decades, has become dominated by procedural liberalism, and has steadily crowded out the republican understandings of citizenship, which is important for self-government.

Sandel's historical retrieval is tracing mainly through development of political economy and constitutional law. In part one, Sandel has discussed the controversies and development of American constitutionalism in the area of religious liberty, free speech, and the regulation of pornography, abortion, homosexuality, and no-fault divorce laws. Throughout the discussion, Sandel shows that while the Constitution is a framework of rights with neutrality among ends, "[w]hat counts as neutrality is often controversial and admits competing interpretations" (p. 50). Underlying these interpretations are different conceptions of what free moral agency consists in. These judgments are therefore not neutral after all. Nevertheless, the republican vision has been steadily replaced by procedural liberalism, which stresses basic individual rights to constrain the majority will. For Sandel, procedural liberalism has implicitly assumed the idea of unencumbered self, which is free to choose their own ends. It fails to respect persons, as member of particular communities to which they belong (p. 89). Such persons are bound by duties derived from religious or traditional belief, that is true for them, but not chosen by themselves (p. 67).

Sandel reminds us that the American Revolution was originally aspiring to generate a new community of common good. By separating from England, Americans attempt to stave off corruption and to realize republican ideals, to "renew the moral spirit that suited Americans to republican government" (p. 127). Unfortunately, in the years following independence, leading politicians and writers started to worry the corruption of the public spirit by the rampant pursuit of luxury and self-interest. Nowadays, most of American practices and institutions have thoroughly embodied the philosophy of procedural liberalism. Despite its philosophical problem, it has offered the public philosophy by which Americans live.

In part two, Sandel has laid out the American history of political economy. Regarding to the debates of large-scale manufacturing in the early republic, it shows the eclipse of Jefferson's vision of the virtue and independence of yeoman farmers. The emergency of factory life raises the controversy of working for a wage and its relation to self-government. The alteration of labor movement and the acceptance of wage labor as a permanent condition lead to the failure of the artisanal republican tradition and the rise of voluntarist conception of freedom. During the Progressive era, there is a contest between the decentralist vision of Louis Brandeis and Woodrow Wilson and the nationalization of big business of Theodore Roosevelt and Herbert Croly on how to preserve democratic self-government in the face of concentrated economic power. However, in the twentieth century, the culture has shifted to the political economy based on consumer identity. Finally, the 'political economy of citizenship' was replaced by the 'political economy of growth and distributive justice' in the Keynesian revolution of Franklin Roosevelt. Throughout these debates, the underlying concern of both sides initially appealed to the consideration of civic character formation, which is most hospitable to self-government (p. 124). Nevertheless, the consequences of these debates have finally excluded the civic consideration from the scope of contemporary American politics. In the last few decades, in particular in the civil rights movement, the procedural republic has won a total victory and the republican tradition has been almost repressed. The predomination of procedural liberalism nevertheless leads to anxiety for "the loss of self-government and the erosion of community" (p. 294).

According to Sandel, the triumph of procedural liberalism could be due to the great wave of migration and the growing diversity of nation. These factors have made the republican character formation of citizens difficult to achieve and strengthened the appeal of moral neutrality. Furthermore, the rise of a national economy and globalization has complicated the republican ideal of subjecting economy to democratic

account. In particular, the culture of consumerism in the capitalist society has offered an individualistic, privatized vision of freedom, which has diminished the aspiration of republican freedom as self-government.

With the loss of community and shared public life necessary to self-government under the aura of procedural liberalism, Sandel calls for an American revival of the republican ideal. Instead of being cosmopolitanism, Sandel suggests to restore politics of self-government, which plays itself out in a multiplicity of political communities in which sovereignty is diffused (p. 345). For Sandel, civic life is best nourished in a particular community, which is intermediate between the individual and the state. Only when sovereignty is scattered and citizenship formed across multiple sites of civic participation can the regime be integrated to rival global market forces, while public life remains to be differentiated, which inspire citizens' reflective allegiance to the whole nation.

What Money Cannot Buy: The Moral Limits of Markets

As the unfettered market and consumerism is one of the great impediments to a politics of civic virtue, Sandel, in this book, explores the relations between market and morality with examples of how we are degrading public life as part of an unrestrained marketization and commodification of almost everything.

From the time of Margaret Thatcher and Ronald Reagan to the financial crisis of 2008, it was an era of what Sandel calls 'market triumphalism' (p. 6). Market was held to be the key to prosperity and freedom. Nearly everything can be priced. Our everyday life has been dominated by market value and thinking. However, Sandel argues throughout this book that market would lead to two problems: inequality and corruption.

The book explores five area of problematic commercialization of our life: (1) paying for the queue; (2) paying for motivating better health, education and environmental protection; (3) commodification of honorary degree, admission of prestigious colleges or universities; (4) commercialization of one's death; and (5) the ubiquity of commercial advertising.

First, for Sandel, allowing payment for queue for certain public service would be unfair to the poor. To turn free public service into market commodity would corrupt the nature of service. Charging admission to congressional hearings, or to allow scalpers to profit from free public performance and religious rituals would violate their original aims, thus degrading their values.

Second, although market power is efficient to motivate certain moral activities, such as reading, health care and environment protection, it indeed induces, Sandel argues, people to do the right things with pervasive incentives. Paying people to be healthy can rebound by failing to cultivate the attitudes of maintaining good health. The monetary incentives of reading books may replace the intrinsic incentives of learning. Economists usually assume that the operation of markets is value-neutral. However, Sandel refutes this assumption, "Markets leave their mark on social norms. Often, market incentives erode or crowd out nonmarket incentives" (p. 64).

Third, if honorary degree or admission of prestigious universities can be sold or bought, it is not only unfair to the poor children, perpetuating social inequality, but also corrupts the institutional integrity. This is because "higher education not only equips students for remunerative jobs, but that it also embodies certain ideals – the pursuit of truth, the promotion of scholarly and scientific excellence, the advancement of humane teaching and learning, the cultivation of civic virtue" (p. 110). Allowing the need of fund-raising to be the most important agenda of college would distort the very reason for existence of university.

Fourth, with the janitors insurance, Sandel shows how life insurance, which originally attempts to diminish the financial impact of death on survivors and dependents, was finally corrupted into a kind of gambling against other people's lives.

Fifth, the commercial advertising nowadays has reached places we would not have ever imagined, such as public transports, schools, classrooms, public area, police cars, fire hydrants, academic books, and human bodies. The culture of commercialization of everything is pernicious to the character building of the youth. The commercial advertisement in the school would further obscure the boundary between places, and makes everywhere a site of selling. While public schools teach students to be discipline and to evaluate their own desire, markets motivate students to pursuit the satisfaction of desires. It violates the ideal of education (p. 200).

All these problems of marketization of almost everything cannot be simply explained in terms of freedom or equality. Sandel argues, in order to identify what trouble us about the current condition, we need the moral terminology of corruption and degradation, which appeals to conceptions of good life (p. 187). The corruption argument also brought to us an important insight that "markets are not mere mechanisms; they embody certain values. And sometimes, market values crowd out non-market norms worth caring about" (p. 113). Our society tends to avoid the deliberation of the conception of good life because of its controversial nature. Such worries impede us to bring our moral conviction in the public area. However, as Sandel states, "shrinking from these questions does not leave them undecided. It simply means that markets will decide them for us. This is the lesson of the last three decades" (p. 202).

The problem of inequality also leads to what Sandel calls "the skyboxification of American life" (p. 203). People from different classes are separated in their living, working, studying, and playing. There is a lack of common good, common life, and sharing of common experience among citizens. It is harmful to social solidarity and democracy. "Democracy does not require perfect equality, but it does require that citizens share a common life. What matters is that people of different backgrounds and social positions encounter one another, and bump up against one another, in the course of ordinary life" (p. 203).

The Case against Perfection: Ethics in the Age of Genetic Engineering

In discussing issues about the idea of eugenics in the era of genetic engineering, such as muscle-enhancement, memory-enhancement, height-enhancement, and reproductive technologies, Sandel shows us the deficiency of fairness argument and autonomy argument. First, even without genetic enhancement, some people are always better endowed than others in competitions, but we never consider it as violating

principle of fairness. Second, even if coercion is absent, eugenics is not therefore vindicated. The autonomy argument cannot elucidate our moral perplex about people who use genetic technology not to cure illnesses, but to extend beyond health, to enhance their abilities so that we are uplifted above the norm. Sandel argues, the way of dealing with the ethics of genetic enhancement is to "questions about the moral status of nature, and about the proper stance of human beings toward the given world" (p. 9). For Sandel, the crucial problem of eugenics and genetic enhancement is that they exemplify "the one-sided triumph of willfulness over giftedness, of dominion over reverence, of molding over beholding" (p. 85). From the religious perspective, to perceive our talents as wholly our own doing is to play God, to confuse our role and the idea of creation. From the aspect of secular philosophy, the idea of the 'self-made man,' represented by bioengineering, would transform our understanding of three values: humility, responsibility, and solidarity (p. 86). To see ourselves as a creature of God or nature is to remind us that we are not fully responsible for our shortcomings. Today, if a basketball player misses a rebound, he/she may be accused of being out of position by his coach. In a world we can master of our genetic endowments, she/he may be accused for being too short. We need to bear a much greater responsibility on our talents and performance in such a manipulating society. If the world is used to specifying genetic traits of their children, the world would also be unfriendly to the unbidden. Therefore, to see children as gift is to learn humility. To accept them as they come rather than products of design teaches parents "to be open to the unbidden" (p. 86). The notion of giftedness also helps to justify redistribution and enhance solidarity. The reason the successful have an obligation to the least-advantaged members of community is that they are not fully deserved for their achievement. Their achievements partly come from the outcome of the genetic lottery. The successful persons are therefore responsible to share their rewards with those who lack those endowments.

His Influence

Sandel's writings are easy to read. His books are well-known globally. His works have been translated into 23 languages. In particular, his *LLJ* has been widely regarded as one of the most important critique of Rawls' *TJ*. Many scholars perceive that Rawls attempts to come to terms with Sandel's criticisms in his *Political Liberalism*, a case for liberalism that is political not metaphysical. Sandel's criticisms have also been extended to the critique of other liberal equalitarian, such as Ronald Dworkin, as well as Robert Nozick, the libertarian. This book has secured his academic reputation in the United States, conceiving him as one of the major proponents of communitarianism, on a par with other communitarians, such as Alasdair MacIntyre, Charles Taylor, and Michael Walzer. This book helped shape the liberal-communitarian debates in the 1980s and 1990s. It has set out the issues of the self and its relation to the community as the clear lines of debate for the theoretical discussion. However, Sandel is resistant to the label 'communitarianism,' because it is usually misunderstood as a kind of majoritarianism.

His second book *Democracy's Discontent* has also provoked lots of responses from both liberals and communitarians. A few years later, *Debating Democracy's Discontent*, the collected responding essays made by political and legal theorists is published. Sandel's thesis of civic republicanism is regarded as helpful guidance to reinvigorate the older philosophical tradition in the modern context (Hall, 2003: p. 113). However, it is also regarded by liberals as risky of being paternalistic and oppressive to individuals. Sandel admits such risk, but he reminds, there are also risks of being moral vacuum by the procedural liberalism, which has "left liberals ill-equipped to engage moral and religious argument in political debate" (Sandel, 1998b: p. 329).

Sandel's lectures are legendary. Up to now, more than 15 000 students have enrolled on his undergraduate course 'Justice,' an introduction to moral and political philosophy. His classes have been filmed by Harvard and WGBH (Boston's PBS station) and produced a public television series that aired across the country in 2009. It is also the first Harvard course to be freely available online and on television. It becomes the hallmark of Harvard philosophy course. In 1985, Sandel was the recipient of the Harvard-Radcliffe Phi Beta Kappa Teaching Prize and honored for excellence in teaching by the American Political Science Association in 2008. His lecturing style is in the form of Socratic dialogue with students debating about moral goods related to different social issues. To certain extent, his lecture can be considered as exemplifying his republican ideal about public moral deliberation.

Sandel has a huge audience not only in the West, but also in Asia. For instance, he has given lectures in mainland China in 2007, Taiwan and Tokyo in 2010, and Seoul in 2012. The Japan Times describes Sandel's visit, "Few philosophers are compared to rock stars or TV celebrities, but that is the kind of popularity Michael Sandel enjoys in Japan." In 2010, China Newsweek named him the 'most influential foreign figure of the year' in China.

See also: Liberalism: Political Doctrine and Impact on Social Science; Rawls, John (1921–2002); Republicanism: Philosophical Aspects; Self: Philosophical Aspects; Taylor, Charles (1931–).

Bibliography

Allen, A.L., Regan, M.C. (Eds.), 1998. Debating Democracy's Discontent. Oxford University Press, New York.
Arendt, H., 1958. The Human Condition. University of Chicago Press, IL.
Hall, T., 2003. Beyond the procedural republic: the communitarian liberalism of Michael Sandel. In: Wolfe, C. (Ed.), Liberalism at the Crossroads: An Introduction to Contemporary Liberal Political Theory and Its Critics. Rowman & Littlefield, Lanham, pp. 97–119.
Kant, I., 1997. Critique of Pure Reason (P. Guyer, A. Wood, Trans. and Eds.). Cambridge University Press, Cambridge.
Mulhall, S., Swift, A., 1996. Liberals and Communitarians, second ed. Blackwell, MA, pp. 40–69.
Nozick, R., 1974. Anarchy, State and Utopia. Basic Books, New York.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, MA, (rev. ed., 1999).
Rawls, J., 1993. Political Liberalism. Columbia University Press, New York, (second ed., 2005).
Sandel, M., 1996. Democracy's Discontent: America in Search of a Public Philosophy. Belknap Press.

Sandel, M., 1998a. Liberalism and the Limits of Justice, second ed. Cambridge University Press, Cambridge, (first ed., 1982).
Sandel, M., 1998b. Reply to critics. In: Allen, A.L., Regan, M.C. (Eds.), Debating Democracy's Discontent. Oxford University Press, New York, pp. 319–335.
Sandel, M., 2005. Public Philosophy: Essays on Morality in Politics. Harvard University Press, MA.
Sandel, M., 2007a. The Case against Perfection: Ethics in the Age of Genetic Engineering. Belknap Press of Harvard University Press, MA.
Sandel, M. (Ed.), 2007b. Justice: A Reader. Oxford University Press, New York.
Sandel, M., 2009. Justice: What's the Right Thing to Do? Farrar, Straus and Giroux, New York.

Relevant Websites

http://nigelwarburton.typepad.com/michael_sandel_links/ – BBC, 2009. Reith Lecture Series: A New Citizenship.
http://www.bbc.co.uk/programmes/b00xyzg7 – BBC, 2011. Justice: Fairness and the Big Society Debate.
www.JusticeHarvard.org – Harvard University, Justice: What's The Right Thing To Do?.
http://nigelwarburton.typepad.com/philosophy_bites/2008/05/michael-sandel.html – Philosophy Bites, 2008. Michael Sandel on Genetic Enhancement in Sport.
http://nigelwarburton.typepad.com/philosophy_bites/2009/05/michael-sandel-on-what-shouldnt-be-sold.html – Philosophy Bites, 2009. Michael Sandel on What Shouldn't Be Sold.

Sapir, Edward (1884–1939)

Victor Golla, Humboldt State University, Arcata, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13482–13486, © 2001, Elsevier Ltd.

Abstract

Edward Sapir studied anthropology under Franz Boas, and became the leading exponent of linguistic anthropology in the first half of the twentieth century. He carried out important work on the descriptive linguistics of North American Indian languages, including definitive studies of Takelma, Southern Paiute, Nootka, and the Athabaskan languages. He was also deeply concerned with exploring historical links among American Indian languages, and made significant contributions to the comparative linguistics of the Uto-Aztecan, Athabaskan, Hokan, and Penutian families. His career until 1925 was almost entirely devoted to this research, but from then until his death he taught at the University of Chicago and at Yale, where he had considerable influence in establishing the paradigm for twentieth-century linguistic research and in training the first generation of American structural linguists. He was also involved deeply in the development of a psychologically sophisticated model of social research, both within anthropology (where he helped create the Personality and Culture school) and in interdisciplinary collaboration, in particular with the psychiatrist Harry Stack Sullivan. A gifted teacher and lucid writer, Sapir was a major intellectual figure in the 1920s and 1930s. His general book, *Language* (1921), was read widely and is still in print.

General Career

Edward Sapir was born in Lauenburg, Germany (now Lebork, Poland) on 26 January 1884, but his family emigrated to the United States when he was 5 years old and eventually settled in New York City. His brilliance earned him a full scholarship to the prestigious Horace Mann School and subsequently a Pulitzer fellowship to Columbia College, where he received his BA in 1904. He continued with graduate work in Germanic philology, but was soon drawn into Franz Boas's orbit and took up an anthropological career, for which he proved extraordinarily well fitted. In 1905 he began a series of field visits to the West that resulted in the detailed documentation of several American Indian languages and cultures, including Wishram Chinook, Takelma, Yana, Southern Paiute, and Nootka. The speed and accuracy with which Sapir collected linguistic and ethnographic data has probably never been surpassed, and his notes, even from his earliest field trips, are among the most valuable manuscripts in American Indian studies.

After receiving his doctorate from Columbia in 1909 with a dissertation on the Takelma language, Sapir taught briefly at the University of Pennsylvania, and in 1910 became Chief of the Division of Anthropology in the Geological Survey of Canada. He remained in this position until 1925, making the Nootka (Nuuchahnulth) language the focus of his research from 1910 to 1914. Between 1915 and 1920, when World War I and its aftermath brought field work to a halt, he devoted a considerable amount of time to the comparative linguistics of North American languages, establishing the Na-Dene relationship between Athabaskan, Tlingit, and Haida, and proposing expansions of the Penutian and Hokan stocks to include a large number of languages in North and Central America. During this period he also began making a name for himself as a literary and social commentator, and began publishing his experimental poetry.

Sapir married Florence Delson in 1911, and they had three children. Shortly after the birth of their third child in 1918, Florence Sapir began to manifest signs of serious illness, partly psychological in character. Although his wife's deteriorating health became a great concern to Sapir, and her death in 1924 was emotionally devastating, he remained committed to the intensive field documentation of American Indian languages, and in 1922 commenced an Athabaskan research program that eventually encompassed fullscale descriptive studies of Sarsi, Kutchin (Gwich'in), Hupa, and Navajo. This work was motivated partly by Sapir's conviction that a historical relationship could be demonstrated between the Na-Dene languages of North America and the Sino-Tibetan languages of East Asia.

After the publication of his highly influential book *Language* (1921) Sapir was regarded widely as one of the leading linguists of his generation. In 1925 he accepted Fay-Cooper Cole's offer of an academic position in the newly reorganized Department of Anthropology and Sociology at the University of Chicago. The move from Ottawa, and his second marriage to Jean McClenaghan, with whom he was to have two more children, was an emotional and intellectual watershed. He found the interdisciplinary atmosphere at Chicago stimulating and increasingly he addressed general and theoretical topics in his professional writing. He also began contributing sprightly and provocative essays on general social and cultural topics to such publications as Mencken's *American Mercury*. While much of his teaching was in linguistics and he took a leading role in establishing that discipline as an autonomous field of study, Sapir also became involved in developing a general model for social science. After he moved to Yale in 1931 as Sterling Professor of Anthropology and Linguistics he became interested particularly in a psychologically realistic paradigm for social research (see Irvine, 1999) and led a seminar on personality and culture. Meanwhile linguistic work on Athabaskan, specifically Navajo, continued to absorb him, and the prospect of finding remote connections of the Na-Dene languages in Asia had led him to the study of Tibetan and Chinese.

This extraordinarily diverse agenda was abruptly suspended in the summer of 1937 when Sapir suffered a serious heart attack, from which he never fully recovered. After a year and

a half of precarious health and restricted activity he died in New Haven, Connecticut on 4 February 1939, a few days after his 55th birthday.

Scientific Contribution

Sapir's scientific work can be divided into three distinct parts. First, there is his substantive work in descriptive and comparative linguistics, almost entirely devoted to North American Indian languages. Second, there is his role in establishing the paradigm for twentieth century linguistic research. Finally, there are the flashes of insight – seldom elaborated into formal hypotheses – with which Sapir from time to time illuminated the landscape of linguistic and social theory.

Substantive Work on Languages

The face that American Indian linguistics presents to the world at the beginning of the twenty-first century probably owes more to Sapir than to any other scholar. When Sapir took up the anthropological study of American Indian languages in 1905, the field was dominated by the classificatory concerns of John Wesley Powell's Bureau of American Ethnology, which saw as its principal task the identification of language and dialect boundaries, and the grouping of languages into families whose historical relationship was undoubted. Only a few earlier scholars like Pickering and Humboldt had been interested in the general philological study of these languages. It was Franz Boas, Sapir's mentor at Columbia, who first proposed making comparative investigation of the grammatical structures of American Indian languages (and of non-European languages generally) a topic for sustained scientific research. Drawing his model from a German tradition of associating language and social behavior that led back through Steinthal to Humboldt, Boas impressed upon his anthropological students the necessity of understanding how linguistic 'morphology' (by which he meant grammatical structure) channeled ideas into expressive forms. Unfortunately for Boas, few of his students were equipped either by training or by intellectual inclination to carry out linguistic research of a more than superficial kind.

The only significant exception was Sapir, whom one may imagine was attracted to Boas's anthropology for precisely this reason. From his earliest fieldwork on Wishram Chinook in 1905, Sapir made grammatical analysis the centerpiece of his research, and from his first publications portrayed American Indian languages with a descriptive clarity they had seldom before enjoyed. His accomplishments are legendary to the scholars who today study these languages, and must rank in the first tier of the grammarian's art. The most renowned of his published studies include a full grammar of Takelma, a now-extinct language of southern Oregon (1922); a full grammar and dictionary of Southern Paiute (1930–31/1992); and an outline grammar of Nootka, with texts and vocabulary, prepared with the assistance of Swadesh (1939).

Sapir's grammatical descriptions are couched in a metalanguage derived from European comparative philology, tempered by Boas's insistence that the structure of every language is *sui generis*. In discussions of the evolution of grammatical theory, Sapir's grammars are sometimes portrayed as 'processual' or as early examples of 'generative' descriptions, but such labels imply a theoretical deliberation that was uncharacteristic of Sapir's work. His concern was to explicate the patterns of the language under consideration as lucidly and unambiguously as possible, not to test a general theory of linguistic structure.

The aptness with which Sapir's descriptions captured the spirit of the languages he analyzed is no better illustrated than by his model of Athabaskan grammar. Although in this case it was not embodied in a full grammatical treatment, he laid out the basic features of his descriptive system for Athabaskan in a number of shorter works, and passed it on to his students and successors through his teaching and in his files. Sixty years after his death it remains the standard descriptive model for all work in Athabaskan linguistics, regardless of the theoretical stance of the analyst.

Throughout his career Sapir maintained a deep interest in historical relationships among languages. He took pride in extending the rigor of the reconstructive method to American Indian language families, and laid the foundations for the comparative study of both Uto-Aztecan and Athabaskan. In the 1930s he returned to Indo-European linguistics and made major contributions to the Laryngeal Hypothesis (the proposal, originating with Ferdinand de Saussure, that the phonology of Proto-Indo-European included one or more laryngeal or pharyngeal consonants not attested in the extant Indo-European languages).

The Professionalization of American Linguistics

By 1921, Sapir was able to draw on the analytic details of the American Indian languages on which he had worked to illustrate, in his book *Language*, the wide variety of grammatical structures represented in human speech. One of the most significant impacts of this highly successful book was to provide a model for the professionalization of academic linguistics in the US during the 1920s and early 1930s.

Under Sapir's guidance, a distinctive American School of linguistics arose, focused on the empirical documentation of language, primarily in field situations. Although most of the students Sapir himself trained at Chicago and Yale worked largely if not exclusively on American Indian languages, the methods that were developed were transferable to other languages. In the mid-1930s Sapir directed a project to analyze English, and during World War II many of Sapir's former students were recruited to use linguistic methods to develop teaching materials for such strategically important languages as Thai, Burmese, Mandarin, and Russian.

A distinction is sometimes drawn between the prewar generation of American linguists – dominated by Sapir and his students and emphasizing holistic descriptions of American Indian languages – and the immediate postwar generation, whose more rigid and focused formal methods were codified by Leonard Bloomfield and others. Since many of the major figures of 'Bloomfieldian' linguistics were Sapir's students, this distinction is somewhat artificial, and a single American Structuralist tradition can be identified extending from the late 1920s through 1960 (Hymes and Fought, 1981). There is little doubt that Sapir's influence on this tradition was decisive.

Theoretical Insights

From his earliest work under Boas, Sapir's independent intellectual style often carried him well beyond the bounds of the academic paradigm he was ostensibly working within. He was not, however, a programmatic thinker, and his groundbreaking work, while deeply admired by his students and colleagues, seldom resulted in significant institutional changes, at least in the short term. The cliché 'ahead of his time' is especially apt in Sapir's case, and the influence of some of his views continues to be felt.

This is most striking in structural linguistics, where Sapir must certainly be accounted one of the most influential figures of the twentieth century. As early as 1910 Sapir was commenting on the importance of formal patterning in phonology, and in the 1920s he was among the first to enunciate the 'phonemic principle' which later figured importantly in his teaching at Chicago and Yale. Characteristically, he left it to his students and such colleagues as Leonard Bloomfield to formalize an analytic methodology, while he himself pursued the psychological implications of formal patterning.

Sapir's (1917) trenchant critique of the culture concept, particularly as defined by A. L. Kroeber, was largely ignored at the time. Sapir argued that the attribution of cultural patterning to an emergent 'superorganic' collective consciousness was an intellectual dead-end, and that research would be better directed at understanding the individual psychology of collective patterned behavior. While Kroeber's view was undoubtedly the dominant one in anthropology and general social science for much of the twentieth century, Sapir's analysis is much more consistent with recent models of human sociocultural behavior that have been developed by evolutionary psychologists. The fact that Sapir's views came from someone with extraordinary insight into the elaborate self-referential patterns of language – traditionally, the most formalized and objectified of social behaviors – can hardly be accidental, and again is consistent with recent developments in cognitive science.

In the late 1920s Sapir found an important intellectual ally in Harry Stack Sullivan, a psychiatrist whose interpersonal theory of the genesis of schizophrenia resonated with Sapir's views (Perry, 1982). The two became close friends, and together with Harold Lasswell they planned and organized the William Alanson White Psychiatric Foundation, a research and teaching institution that ultimately was located in Washington, DC. In the year before he died Sapir gave serious consideration to leaving Yale and working with Sullivan in a research position at the Foundation.

Sapir's views on the history of language were linked to his view of the abstraction of patterns, and were equally controversial or misunderstood. A distinction must be drawn between Sapir's work, noted earlier, as a historical linguist within a family of languages whose relationship was secure (Athabaskan, Uto-Aztecan, Indo-European), and his explorations of much less certain relationships (Hokan, Penutian, Na-Dene) and possible interhemispheric connections (Sino-Dene). In the former, Sapir worked – with characteristic creativity and insight – with tools and models derived from a long tradition of comparative linguistics. In the latter, it was (or seemed to his contemporaries) often a matter of brilliant intuition. In fact, in this work he usually relied on an assessment of similarities in structural pattern that distinguished features susceptible to unconscious change from generation to generation (e.g., regular inflectional patterns, words) from those that are largely inaccessible to individual cognition. Although he never offered a theoretical explication, his idea of what constituted 'deep' linguistic patterns was exemplified in his classification of North and Central American Indian languages (1929). This classification, which remains influential, is still untested in its own terms.

Impact and Current Importance

A charismatic teacher, Sapir had a succession of highly motivated students both at Chicago and at Yale, the most prominent among them Morris Swadesh (who collaborated with him on Nootka research), Harry Hoijer (who codified Sapir's analysis of Athabaskan grammar), Mary R. Haas, Stanley Newman, C. F. Voegelin, George L. Trager, Zellig Harris, David G. Mandelbaum, and Benjamin L. Whorf. Through these students Sapir exercised a considerable posthumous influence on intellectual and institutional developments in both linguistics and in anthropology through the 1960s. A postwar collection of Sapir's most important general papers, edited by Mandelbaum (1949), was widely read and is still consulted. Harris' (1951) extended review of this book provides a comprehensive summary of Sapir's *oeuvre* as it was understood by his immediate circle.

Sapir is cited most frequently today for the 'Sapir-Whorf Hypothesis' of linguistic relativity, a name that inaccurately implies an intellectual collaboration between Sapir and his Yale student, Benjamin Whorf, who himself died in 1941. The more correctly designated 'Whorf theory complex' (Lee, 1996) was a retrospective construct that derived largely from writings of Whorf's that were unpublished at the time of Sapir's death. Although undoubtedly stimulated by Sapir's writing and teaching, Whorf's proposal that the structure of a language to some extent determines the cognitive and behavioral habits of its speakers cannot be connected directly with Sapir's mature thought on the psychology of language and culture.

Sapir's most enduring achievement is his own descriptive linguistic work. Long after the writings of most of his contemporaries have been forgotten, Sapir's grammatical studies continue to be held in the highest esteem. In recent years his holistic analytic technique has been emulated by a number of linguists seeking an alternative to narrow formalism, particularly when working with American Indian or other indigenous languages.

Sapir's life and work was the subject of a 1984 conference (Cowan et al., 1986), from which emerged a plan to publish a standard edition of all of Sapir's work, including edited versions of unfinished manuscripts; by year 2000 seven volumes had appeared. A biography by Darnell (1990) is useful for the externals of Sapir's career, but her reluctance to give an intellectual account of Sapir's work, particularly in linguistics, leaves some important issues still to be addressed (Silverstein, 1991).

See also: Americas, Sociocultural Overviews: North America, Native Americans; Human Language, Evolution of; Linguistic Anthropology; Linguistic Fieldwork; Linguistic Typology; Phonology; Population Composition by Race and Ethnicity: North America; Psycholinguistics: Overview; Sapir-Whorf Hypothesis; Sociolinguistics.

Bibliography

Cowan, W., Foster, M.K., Koerner, K. (Eds.), 1986. New Perspectives in Language, Culture and Personality. Benjamins, Amsterdam and Philadelphia.

Darnell, R., 1990. Edward Sapir: Linguist, Anthropologist, Humanist. University of California Press, Berkeley and Los Angeles, CA.

Harris, Z.S., 1951. Review of D.G., Mandelbaum (Ed.). Selected writings of Edward Sapir in language, culture, and personality. Language 27, 288–333.

Hymes, D., Fought, J., 1981. American Structuralism. Mouton, The Hague, The Netherlands.

The psychology of culture: A course of lectures by Edward Sapir, 1927–1937. In: Irvine, J.T. (Ed.), The Collected Works of Edward Sapir, vol. 3. Culture. Mouton de Gruyter, Berlin and New York, pp. 385–686 [also published as a separate volume, The Psychology of Culture, Mouton de Gruyter 1993].

Lee, P., 1996. The Whorf Theory Complex: A Critical Reconstruction. Benjamins, Amsterdam and Philadelphia.

Mandelbaum, D.M. (Ed.), 1949. Selected Writings of Edward Sapir in Language, Culture, and Personality. University of California Press, Berkeley and Los Angeles, CA.

Perry, H.S., 1982. Psychiatrist of America: The Life of Harry Stack Sullivan. Belknap Press, Cambridge, UK and London.

Sapir, E., 1917. Do we need a 'superorganic'? American Anthropologist 19, 441–447.

Sapir, E., 1921. Language. Harcourt Brace, New York.

Sapir, E., 1922. The Takelma language of Southwestern Oregon. In: Boas, F. (Ed.), Handbook of American Indian Languages, Part 2. Bureau of American Ethnology, Washington, DC, pp. 1–296.

Sapir, E., 1929. Central and North American Indian languages. Encyclopaedia Britannica, fourteenth ed. vol. 5. pp. 138–141.

Sapir, E., 1930–31/1992. The Southern Paiute Language. The Collected Works of Edward Sapir, vol. 10. Mouton de Gruyter, Berlin and New York.

Sapir, E., Swadesh, M., 1939. Nootka Texts: Tales and Ethnological Narratives with Grammatical Notes and Lexical Materials. Linguistic Society of America, Philadelphia.

Silverstein, M., 1991. Problems of Sapir historiography. His-toriographia Linguistica 18, 181–204.

Sapir-Whorf Hypothesis

John A Lucy, University of Chicago, Chicago, IL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13486–13490, © 2001, Elsevier Ltd.

Abstract

The Sapir–Whorf hypothesis, also known as the linguistic relativity hypothesis, refers to the proposal that the particular language one speaks influences the way one thinks about reality. Linguistic relativity stands in close relation to semiotic-level concerns with the general relation of language and thought, and to discourse-level concerns with how patterns of language use in cultural context can affect thought. Linguistic relativity is distinguished both from simple linguistic diversity and from strict linguistic determinism. The long history of the hypothesis is sketched with an emphasis on the hierarchical formulations characteristic of most early efforts. This is followed by a description of the work of Sapir and Whorf which departs markedly from this earlier tradition and has been influential in the contemporary period, hence the association of their names with the issue. Whorf's basic argument about analogical influences is outlined in some detail. Despite widespread interest, quality empirical research has been in short supply. Recent efforts to remedy this are described. The research is divided into structure-centered, domain-centered, and behavior-centered types, depending on their manner of approaching the problem. The contemporary period has seen a rapid improvement in the quality of some of these efforts. Current trends likely to characterize future research are briefly characterized.

Nature and Scope of the Hypothesis

The Sapir–Whorf hypothesis, also known as the linguistic relativity hypothesis, refers to the proposal that the particular language one speaks influences the way one thinks about reality. Although proposals concerning linguistic relativity have long been debated, American linguists Edward Sapir (1884–1939) and Benjamin Lee Whorf (1897–1941) advanced particularly influential formulations during the second quarter of the twentieth century, and the topic has since become associated with their names. The linguistic relativity hypothesis focuses on structural differences among natural languages such as Hopi, Chinese, and English, and asks whether the classifications of reality implicit in such structures affect our thinking about reality more generally.

Analytically, linguistic relativity as an issue stands between two others: a semiotic-level concern with how speaking any natural language whatsoever might influence the general potential for human thinking (i.e., the general role of natural language in the evolution or development of human intellectual functioning), and a functional- or discourse-level concern with how using any given language code in a particular way might influence thinking (i.e., the impact of special discursive practices such as schooling and literacy on formal thought). Although analytically distinct, the three issues are intimately related in both theory and practice. For example, claims about linguistic relativity depend on understanding the general psychological mechanisms linking language to thinking, and on understanding the diverse uses of speech in discourse to accomplish acts of descriptive reference. Hence, the relation of particular linguistic structures to patterns of thinking forms only one part of the broader array of questions about the significance of language for thought.

Proposals of linguistic relativity necessarily develop two linked claims among the key terms of the hypothesis (i.e., language, thought, and reality). First, languages differ significantly in their *interpretations* of experienced reality – both what they select for representation and how they arrange it. Second, language interpretations have *influences* on thought about reality more generally – whether at the individual or cultural level. Claims for linguistic relativity thus require both articulating the contrasting interpretations of reality latent in the structures of different languages, and assessing their broader influences on, or relationships to, the cognitive interpretation of reality.

Simple demonstrations of *linguistic diversity* are sometimes mistakenly regarded as sufficient in themselves to prove linguistic relativity, but they cannot in themselves show that the language differences affect thought more generally. (Much confusion arises in this regard because of the practice in linguistics of describing the meaningful significance of individual elements in a language as 'relative to' the grammatical system as a whole. But this latter relativity of the meaning of linguistic elements to the encompassing linguistic structure should be distinguished from broader claims for a relativity of thought more generally to the form of the speaker's language.) A variety of other arguments to the effect that distinctive perceptual or cognitive skills are required to produce and comprehend different languages likewise usually fail to establish any general effects on thought (see Niemeier and Dirven, 2000).

Linguistic relativity proposals are sometimes characterized as equivalent to *linguistic determinism*, that is the view that all thought is strictly determined by language. Such characterizations of the language–thought linkage bear little resemblance to the proposals of Sapir or Whorf, who spoke in more general terms about language influencing habitual patterns of thought, especially at the conceptual level. Indeed, no serious scholar working on the linguistic relativity problem as such has subscribed to a strict determinism. (There are, of course, some who simply equate language and thought, but under this assumption of identity, the question of influence or determinism is no longer relevant.) Between the patent linguistic diversity that nearly everyone agrees exists and a claim of

linguistic determinism that no one actually espouses, lies the proposal of linguistic relativity, that is the proposal that our thought may in some way be taken as relative to the language spoken.

Historical Development of the Hypothesis

Interest in the intellectual significance of the diversity of language categories has deep roots in the European tradition (Aarsleff, 1988; Werlen, 1989; Koerner, 1992). Formulations related to contemporary ones appear during the Enlightenment period in the UK (Locke), France (Condillac, Diderot), and Germany (Hamman, Herder). They are stimulated variously by opposition to the universal grammarians, by concerns about the reliability of language-based knowledge, and by practical efforts to consolidate national identities and cope with colonial expansion. Most of this work construes the differences among languages in terms of a hierarchical scheme of adequacy with respect to reality, to reason, or to both. Later, nineteenth-century work in Germany by Humboldt and in France/Switzerland by Saussure drew heavily on this earlier tradition and set the stage for the approaches of Sapir and Whorf. Humboldt's arguments, in particular, are often regarded as anticipating the Sapir–Whorf approach. He argued for a linguistic relativity according to the formal processes used by a language (e.g., inflection, agglutination, etc.). Ultimately this remains a hierarchical relativity in which certain language types (i.e., European inflectional ones) are viewed as more adequate vehicles of thought and civilization – a view distinctly at odds with what is to follow.

Working within the US anthropological tradition of Franz Boas and stimulated by the diversity and complexity of Native American languages, Edward Sapir (1949) and Benjamin Lee Whorf (1956) reinvigorated and reoriented investigation of linguistic relativity in several ways (Lucy, 1992a; Lee, 1997). First, they advocated intensive first-hand scientific investigation of exotic languages; second, they focused on structures of meaning, rather than on formal grammatical process such as inflection; and third, they approached these languages within a framework of egalitarian regard. Although not always well understood, theirs is the tradition of linguistic relativity most widely known and debated today. Whorf's writings, in particular, form the canonical starting point for all subsequent discussion.

Whorf proposed a specific mechanism for how language influences thought, sought empirical evidence for language effects, and articulated the reflexive implications of linguistic relativity for scholarly thought itself. In his view, each language refers to an infinite variety of experiences with a finite array of formal categories (both lexical and grammatical) by grouping experiences together as analogically 'the same' for the purposes of speech. The categories in a language also interrelate in a coherent way, reinforcing and complementing one another, so as to constitute an overall interpretation of experience. These linguistic classifications vary considerably across languages not only in the basic distinctions they recognize but also in the assemblage of these categories into a coherent system of reference. Thus the system of categories which each language provides to its speakers is not a common, universal system, but a particular 'fashion of speaking.'

Whorf argued that these linguistic structures influence habitual thought by serving as a guide to the interpretation of experience. Speakers tend to assume that the categories and distinctions of their language are entirely natural and given by external reality, and thus can be used as a guide to it. When speakers attempt to interpret an experience in terms of a category available in their language, they unwittingly involve other language-specific meanings implicit in that particular category and in the overall configuration of categories in which it is embedded. In Whorf's view language does not blind speakers to some obvious reality, but rather it suggests associations which are not necessarily entailed by experience. Because language is such a pervasive and transparent aspect of behavior, speakers do not understand that the associations they 'see' are from language, but rather assume that they are 'in' the external situation and patently obvious to all. In the absence of another language (natural or artificial) with which to talk about experience, speakers will not be able to recognize the conventional nature of their linguistically based understandings. Whorf argues that by influencing everyday habitual thought in this way, language can come to influence cultural institutions generally, including philosophical and scientific activity.

In his empirical research Whorf showed that the Hopi and English languages treat 'time' differently, and that this difference corresponds to distinct cultural orientations toward temporal notions. Specifically, Whorf argued that speakers of English treat cyclic experiences of various sorts (e.g., the passage of a day or a year) in the same grammatical frame used for ordinary object nouns. Thus, English speakers are led to treat these cycles as object-like in that they can be measured and counted just like tangible objects. English also treats objects as if they each have a form and a substance. Since the cyclic words get put into this object frame, English speakers are led to ask what is the substance associated with the forms *a day*, *a year*, and so forth. Whorf argues that our global, abstract notion of 'time' as a continuous, homogeneous, formless something can be seen to arise to fill in the blank in this linguistic analogy. The Hopi, by contrast, do not treat these cycles as objects but as recurrent events. Thus, although they have, as Whorf acknowledged, words for what English speakers would recognize as temporal cycles (e.g., days, years, etc.), the formal analogical structuration of these terms in their grammar does not give rise to the abstract notion of 'time' that English speakers have. (Ironically, critics of Whorf's Hopi data often miss his point about structural analogy and focus narrowly on individual lexical items.) Finally, grouping referents and concepts as formally 'the same' for the purposes of speech has led speakers to group those referents and concepts as substantively 'the same' for action generally, as evidenced by related cultural patterns of belief and behavior he describes.

Empirical Research on the Hypothesis

Although the Sapir–Whorf proposal has had wide impact on thinking in the humanities and social sciences, it has not been extensively investigated empirically. Indeed, some believe it is too difficult, if not impossible in principle, to investigate.

Further, a good deal of the empirical work that was first developed was quite narrowly confined to attacking Whorf's analyses, documenting particular cases of language diversity, or exploring the implications in domains such as color terms that represent somewhat marginal aspects of language structure. In large part, therefore, acceptance or rejection of the proposal for many years depended more on personal and professional predilections than on solid evidence. Nonetheless, a variety of modern initiatives have stimulated renewed interest in mounting empirical assessments of the hypothesis.

Contemporary empirical efforts can be classed into three broad types, depending on which of the three key terms in the hypothesis they take as their point of departure: language, reality, or thought (Lucy, 1997).

A *structure-centered approach* begins with an observed difference between languages, elaborates the interpretations of reality implicit in them, and then seeks evidence for their influence on thought. The approach remains open to unexpected interpretations of reality but often has difficulty establishing a neutral basis for comparison. The classic example of a language-centered approach is Whorf's pioneering comparison of Hopi and English described above. The most extensive contemporary effort to extend and improve the comparative fundamentals in a structure-centered approach has sought to establish a relation between variations in grammatical number marking and attentiveness to number and shape (Lucy, 1992b). This research remedies some of the traditional difficulties of structure-centered approaches by framing the linguistic analysis typologically so as to enhance comparison, and by supplementing ethnographic observation with a rigorous assessment of individual thought. This then makes possible the realization of the benefits of the structure-centered approach: placing the languages at issue on an equal footing, exploring semantically significant lexical and grammatical patterns, and developing connections to interrelated semantic patterns in each language.

A *domain-centered approach* begins with a domain of experienced reality, typically characterized independently of language(s), and asks how various languages select from, encode, and organize it. Typically, speakers of different languages are asked to refer to 'the same' materials or situations so the different linguistic construals become clear. The approach facilitates controlled comparison, but often at the expense of regimenting the linguistic data rather narrowly. The classic example of this approach, developed by Roger Brown and Eric Lenneberg in the 1950s, showed that some colors are more lexically encodable than others, and that more codable colors are remembered better. This line of research was later extended by Brent Berlin, Paul Kay, and their colleagues, but to argue instead that there are cross-linguistic *universals* in the encoding of the color domain such that a small number of 'basic' color terms emerge in languages as a function of biological constraints. Although this research has been widely accepted as evidence against the validity of linguistic relativity hypothesis, it actually deals largely with constraints on linguistic diversity rather than with relativity as such. Subsequent research has challenged Berlin and Kay's universal semantic claim, and shown that different color-term systems do in fact influence color categorization and memory. (For discussions and references, see Lucy, 1992a; Hardin and Maffi, 1997; Roberson et al., 2000.) The most successful effort to improve the quality of the linguistic comparison in a domain-centered approach has sought to show cognitive differences in the spatial domain between languages favoring the use of body coordinates to describe arrangements of objects (e.g., 'the man is left of the tree') and those favoring systems anchored in cardinal direction terms or topographic features (e.g., 'the man is east/uphill of the tree') (Pederson et al., 1998; Levinson, 2003). This research on space remedies some of the traditional difficulties of domain-centered approaches by developing a more rigorous and substantive linguistic analysis to complement the ready comparisons facilitated by this approach.

A *behavior-centered approach* begins with a marked difference in behavior which the researcher comes to believe has its roots in and provides evidence for a pattern of thought arising from language practices. The behavior at issue typically has clear practical consequences (either for theory or for native speakers), but since the research does not begin with an intent to address the linguistic relativity question, the theoretical and empirical analyses of language and reality are often weakly developed. The most famous example of a behavior-centered approach is the effort to account for differences in Chinese and English speakers' facility with counterfactual or hypothetical reasoning by reference to the marking of counterfactuals in the two languages (Bloom, 1981). The interpretation of these results remains controversial (Lucy, 1992a).

Future Prospects

Two research trends are unfolding at the present time. First, in the cognitive and psychological sciences awareness is increasing of the nature and scope of language differences. This has led to a greater number of studies focused on the possible cognitive consequences of such differences (e.g., Levinson, 2003; Niemeier and Dirven, 2000). Second, an increasing integration is emerging among the three levels of language and thought problem (i.e., the semiotic, structural, and functional levels). On the semiotic side, for example, research on the relationship between language and thought in development is increasingly informing and informed by work on linguistic relativity (Bowerman and Levinson, 2001). On the functional side, research on the relationship of cultural and discursive patterns of use is increasingly being brought into dialogue with Whorfian issues (Silverstein, 1979; Friedrich, 1986; Wierzbicka, 1992; Hill and Mannheim, 1992; Gumperz and Levinson, 1996).

The continued relevance of the linguistic relativity issue seems assured by the same impulses found historically: the patent relevance of language to human sociality and intellect, the reflexive concern with the role of language in scholarly practice, and the practical encounter with linguistic diversity. To this we must add the increasing concern with the unknown implications for human thought of the impending loss of many if not most of the world's languages (Fishman, 1982).

See also: Cognitive Psychology: History; Human Cognition, Evolution of; Language and Thought: The Neo-Whorfian

Hypothesis; Linguistic Anthropology; Linguistics: Overview; Sapir, Edward (1884–1939); Semiotics; Wittgenstein, Ludwig (1889–1951).

Bibliography

Aarsleff, H., 1988. Introduction (P. Heath, Trans.). In: Von Humboldt, W. (Ed.), On Language: The Diversity of Human Language-structure and its Influence on the Mental Development of Mankind. Cambridge University Press, Cambridge, UK, pp. vii–vxv.

Bloom, A.H., 1981. The Linguistic Shaping of Thought. Lawrence Erlbaum, Hillsdale, NJ.

Bowerman, M., Levinson, S.C., 2001. Language Acquisition and Conceptual Development. Cambridge University Press, Cambridge, UK.

Fishman, J., 1982. Whorfianism of the third kind: Ethnolinguistic diversity as a worldwide societal asset (The Whorfian Hypothesis: Varieties of validation, confirmation, and disconfirmation II). Language in Society 11, 1–14.

Friedrich, P., 1986. The Language Parallax: Linguistic Relativism and Poetic Indeterminacy. University of Texas, Austin, TX.

Gumperz, J.J., Levinson, S.C. (Eds.), 1996. Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, UK.

Hardin, C.L., Maffi, L. (Eds.), 1997. Color Categories in Thought and Language. Cambridge University Press, Cambridge, UK.

Hill, J.H., Mannheim, B., 1992. Language and world view. Annual Review of Anthropology 21, 381–406.

Koerner, E.F.K., 1992. The Sapir–Whorf hypothesis: A preliminary history and a bibliographic essay. Journal of Linguistic Anthropology 2, 173–178.

Lee, P., 1997. The Whorf Theory Complex: A Critical Reconstruction. John Benjamins, Amsterdam, The Netherlands.

Levinson, S.C., 2003. Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge, UK.

Lucy, J.A., 1992a. Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis. Cambridge University Press, Cambridge, UK.

Lucy, J.A., 1992b. Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis. Cambridge University Press, Cambridge, UK.

Lucy, J.A., 1997. Linguistic relativity. Annual Review of Anthropology 26, 291–312.

Niemeier, S., Dirven, R. (Eds.), 2000. Evidence for Linguistic Relativity. John Benjamins, Amsterdam, The Netherlands.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., Senft, G., 1998. Semantic typology and spatial conceptualization. Language 74, 508–556.

Roberson, D., Davies, I., Davidoff, J., 2000. Color categories are not universal: Replications and new evidence from a stone-age culture. Journal of Experimental Psychology – General 129, 369–398.

Sapir, E., 1949. In: Mandelbaum, D.G. (Ed.), The Selected Writings of Edward Sapir in Language, Culture, and Personality. University California Press, Berkeley, CA.

Silverstein, M., 1979. Language structure and linguistic ideology. In: Clyne, P., Hanks, W., Hofbauer, C. (Eds.), The Elements: A Parasession on Linguistic Units and Levels. Chicago Linguistic Society, Chicago, pp. 193–247.

Werlen, I., 1989. Sprache, Mensch und Welt: Geschichte und Bedeutung des Prinzips der sprachlichen Relativität. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

Whorf, B.L., 1956. In: Carroll, J.B. (Ed.), Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. MIT Press, Cambridge, MA.

Wierzbicka, A., 1992. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-specific Configurations. Oxford University Press, Oxford, UK.

Sarmiento, Domingo Faustino (1811–88)

Carlos Altamirano, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Writer, thinker, educator, and politician, Domingo Faustino Sarmiento (1811–88) was one of the great intellectual figures of Argentinian culture in the nineteenth century. He was born in a poor family, but Sarmiento, who attended just a few years of elementary school, became a career educator and man of letters. He wrote his best books in exile, where he had been thrown by the civil war that divided the country. His thesis on the South American caudillo (military leaders) particularly his famous reflection on Argentinian political life in *Facundo* had great influence in Latin America. Sarmiento ruled Argentina between 1868 and 1874.

Biography

Domingo Faustino Sarmiento was one of the great figures of the Argentinian intellectual elite of the nineteenth century. Sarmiento was a liberal writer and thinker and his work reflects the development of social and political concerns that were partly his own but also shared with the generation of which he was a part (Bunkley, 1952).

After the War of Independence against Spain and the ensuing political and social dislocation that produced the 1810 revolution, Argentina, like the majority of Latin American countries, sank into a long crisis, with various stretches of bloody turf war (war). Finding a way out of this crisis was precisely the central concern of the so-called 'generation of 1837,' which introduced literary and social Romanticism to the country (Weinberg, 1958). The program of this small, young intelligentsia was based on the premise that the study of local reality and customs was required to establish an adequate political order reflecting national aspirations (i.e., a legitimately Argentinian political order). According to this 'Romantic generation,' understanding the nature of South American history was necessary to understand the hegemony of the *caudillos* (military leaders that ruled most of Latin America in this era) and to comprehend the failure of even enlightened governments to evolve political institutions suitable for the national progress. The circle, composed mostly of university students, first met in a club, the Literary Salon, established in 1837 in Buenos Aires, and soon after, they met in a secret lodge: the Association of the Young Generation of Argentina. The lodge was short-lived, but from the founders and supporters of the ideological project of 1837, a liberal constitutional doctrine would proceed, along with the emergence of a national historiography and guidance for research on the literary past that would impose itself in the second half of the nineteenth century (Romero, 2001). One of the names associated with this intellectual program is that of Domingo Faustino Sarmiento.

Sarmiento was born in 1811 in a secluded corner of Argentina, in the city of San Juan, the capital of San Juan Province. His parents were *decente* (poor but respectable and considered white – i.e., with no Indian or African blood). From infancy, Sarmiento knew the tribulations of a life of destitution in a home that depended on maternal work. The nine years he attended the *Escuelas de la Patria* (Schools of the Homeland) provided his only formal education. Sarmiento would recall later that he also received lessons from his priest uncles, whose teachings of religion, Latin, geography, and grammar extended far beyond his initial childhood training. From these familial lectures, he obtained a love of reading, a passion for books, a yearning for knowledge, and the passion for educational action. Given the outstanding skills that the child showed in his studies, his parents imagined that his education and knowledge would provide him a role and a reputation in the public life of the country (Anderson Imbert, 1960).

But this was not an easy road in a country torn with civil war (Halperin Donghi, 1972). In effect, nineteenth-century Argentina was divided into two irreconcilable factions: the *unitarios*, who had their moment in the spotlight under the enlightened rule of Bernardino Rivadavia; and the *federales*, who had defeated the *unitarios* and aligned themselves with the provincial *caudillos* and their popular rural and urban base. Since 1835, the most powerful of these *caudillos*, Juan Manuel de Rosas, ruled the province of Buenos Aires with an iron fist and, with the most important of the provinces as his base, also ruled the Argentinian Confederacy almost continuously from 1829 to 1852. (The Argentine Confederacy was the official name of a political network of agreements and commitments between Rosas and the leaders of the other provinces.)

Although the Sarmiento family had been on the federal side, Sarmiento himself took up arms and joined the *unitario* cause, the faction supported by the majority of the country's intellectual elite. Sarmiento was still an adolescent when he enrolled with the *unitarios* and participated in some of the skirmishes of the civil war that dominated political life in Argentina at that time. The troops he had joined were defeated and that led to his imprisonment and subsequent exile in neighboring Chile. He returned to his province the following year, but because of military setbacks, he had to go into exile in Chile again, where he remained from 1831 to 1836 (Gálvez, 1945).

In 1836 Sarmiento returned to San Juan, where a federal leader governed with a more benign and tolerant style. There he would found a school for women and a newspaper, *El Zonda*. He also reorganized a theatrical association and began to write. More importantly for his ideological evolution is the discovery he made between 1838 and 1840 of a new set of doctrines derived from the European thinking of that time. A young

lawyer from San Juan, who had been a member of the Romantic ideological in Buenos Aires, carried to San Juan not only the ideas of that circle, but also his personal library. Through his friend's library, the voracious self-taught Sarmiento read books and magazines that spread the doctrine of the Romantic movement in France. This literature mixed the philosophy of history with historical studies, philosophical eclecticism and the critique of eclecticism, social humanitarianism and liberalism, literary theories with theories of the law. The authors and titles that would evoke quotes years later from this time of discovery were Guizot, Chateaubriand, Lammenais, Madame de Stäel, Pierre Leroux, Quinet, Lerminier, together with the French journals *La Revue des Deux Mondes* y *La Revue Enciclopédique*. Significantly for Sarmiento, de Tocqueville's *Democracy in America* became an intellectual guide, as it was to the rest of his generation. When in 1840 an argument with the federal *caudillo* that governed San Juan forced him to take exile again, he took with him to Chile this new wealth of readings and ideas (Botana, 1984).

Sarmiento soon became a prominent personage in Chile, where he counted on the protection of Manuel Montt, a leading figure in the Chilean government's conservative party (Galvez, 1945). Access to the press was the first step in his remarkable public career. Through journalism, he soon revealed his talent for prose about combat, the literature of ideas, and the chronicling of customs. He threw everything he had read into his articles, sometimes in a rather jumbled way. He also began to test the biographical genre, and in a short text evocative of an Argentinian *caudillo*, Sarmiento showed his narrative ability through the story of Félix Aldao. His second biography appeared as a series in the newspaper he directed and led to his fame as a writer and man of ideas: Civilization and barbarism, the life of Facundo Quiroga; and physical aspects, customs and habits of the Republic of Argentina, published in 1845 and known simply as *Facundo*. This same year, Minister Montt charged him with the task of studying *in situ* the operation of primary education in Europe and the United States. The commission was not unjustified, since education was another area where Sarmiento excelled in Chile; indeed, in 1842 he had been appointed Director of the Normal School of Primary Instruction and, 2 years later, wrote a *Method of Gradual Reading* that would be adopted in all primary schools in the country. This background, along with other instructional works, testified to the passionate commitment of the Argentinian writer to public education, which he saw as an instrument of modernization and democratization for all the societies of South America (Verdevoye, 1963).

During his long second exile in Chile, and at the behest of the Chilean government, Sarmiento traveled widely throughout South America, Europe, and the United States, the fruits of which were his books *Popular Education* and *Travels*, both published in Chile in 1849. When the end of the Rosas government seemed near, Sarmiento produced a proposed reorganization of his country, *Argirópolis*, and also wrote his celebrated autobiography, *Memories of Provincial Life*, a book that his rivals saw as the opening of a campaign for the Presidency of Argentina. He left exile in 1851 to participate in the struggle against Rosas but soon returned to Chile because, in his view, the *caudillos* continued to rule even though Rosas was no longer in power. Finally, in 1855 he returned to Argentina and settled in Buenos Aires, then a separate state of the national political order. In Buenos Aires, Sarmiento began his career in public office: municipal adviser, repeatedly a senator, government minister, governor of San Juan for 2 years, and the head of numerous diplomatic missions. While serving as ambassador to the United States, many leaders and officials of the Army offered him the candidacy for president of the republic in the elections to be held in 1868. Sarmiento agreed, was elected and maintained executive power for 6 years – years marked by a war with neighboring Paraguay and various episodes of struggle between internal factions.

Once he completed his work as president, he returned to his two passions, public education and preaching through media. Now that he was a marginal player in politics, a continuing controversy surrounding the law of common education offered Sarmiento, the chance to return to battling for what he deemed essential for the civilization and progress of his country – universal education (Weinberg, 1984). The time of writing great books was over; his most significant works were produced when he was exiled in Chile. His most ambitious post-Presidential intellectual project, *Conflict and Harmonies of the Races in America* (1883), paid tribute to the racism of positivist sociology in that era, but it was an unsuccessful work. At the time of his death in 1888, Argentina had changed dramatically because of the enormous drive that Sarmiento had for the education of his country (among other reasons, of course).

The Literary Legacy of Sarmiento

Sarmiento was the premier Argentinian author of the nineteenth century. Almost all of his written work is compiled in the collection organized by Luis Montt and Augusto Belin, *Works of D.F. Sarmiento*. Much of the 52 volumes of work reflect Sarmiento's journalistic labors, much of which comprise some of the best writings that his pen produced. The Argentinian writer Eduardo Mallea collected some of the literary articles and documented social habits from Sarmiento under the title *Prose of Seeing and Thinking* (Sarmiento, 1943). These writings showcase an almost wild and vivid talent for uniting with his subjects, an animated and personal prose style, an alert and thoughtful vision, acute observation of the curious and the quaint and ensuing generalization about the habits of men, and most of all the values Sarmiento preached to achieve progress throughout South America. There are also complaints and quarrels in these pages, reflections of an author who felt at home amid controversy. In many of the debates waged in the press, as in the discussion of Romanticism or about the orthography of Spanish, Sarmiento revealed himself to be a skilled intellectual fighter.

There is scholarly consensus that Sarmiento's most successful works are those composed during the years he was exiled in Chile. There are three works that cement his fame as a writer and have confirmed their place as national literary classics: *Facundo*, *Travels*, and *Memories of Provincial Life*. None of these shows a master plan or unity of structure. It is, rather, the mix of styles and subjects that distinguishes and inspires his discourse. If we put aside *Facundo*, to which we will return, the second of these works, whose complete title is *Travels through Europe, Africa and America* (1845–47), shows almost naturally,

through the very genre chosen to express his thoughts, the functioning of Sarmiento's intelligence and imagination. *Travels* compiles long letters to friends in Argentina and Chile, to whom Sarmiento wrote about the places he visited, the personalities he encountered, political institutions, the customs of dress and entertainment; or in short, the wide array of circumstances and events that he came across during his long journey that began in Valparaíso and ended in the United States (Sarmiento, 1993).

The epistolary form and the topical material that is characteristic of travel literature allow Sarmiento to jump from one topic to another, from a descriptive story to portraits of individuals that left impressions on his spirit as he roamed the world. He wanted to see and feel those cities that he had heard about but only knew through reading – from Montevideo to Rio de Janeiro in South America, to Paris, Madrid, Rome, Munich, Berlin, and Oran (an African colony), then to New York, Baltimore, Philadelphia, and Washington. Of course, Sarmiento was obsessed with the fate of the South American people, who he deemed backward, and thus wanted to write about useful things, about lessons that could help his people. He did not neglect, on the other hand, the fulfillment of the mission that motivated the trip: the study of public education in the countries he would visit, work that he would also consider in *Popular Education*. But at the same time, he did not hesitate to stick his nose everywhere he wanted. The Paris that he discovered was much like the Paris of his readings, i.e., the cultural center of the civilized world. But upon a closer look, French society appeared much more unequal than he had imagined. The months he was in Spain only confirmed what he already knew: the country was stuck in an earlier time – the spirit of the Inquisition and Phillip II still reigned. Spain was a country that did not know itself – he observed that every new idea entered Spain just as it would enter South America: as a rendition, a copy, not as the original thing. Sarmiento dedicated many pages to the bullfights, an entertainment that he judged as cruel and barbaric but strangely seductive. The ancient metropolis, the seat of the former Spanish Empire, in short, could not be the example of progress for its former colonies.

What dazzled Sarmiento the most about his travels was the sociopolitical experiment that was taking place in North America, as he confessed in a letter about the United States. Not everything about the US seemed admirable to Sarmiento, but he admired everything he judged as essential to the social experiment and the behaviors upon which it rested: industriousness, simplicity, and the intelligence to appropriate, generalize, and perfect all the methods and instruments of progress, all of which were shown to be within the reach of man. He paid close attention to the most varied aspects of his experience, and especially to those aspects for which he had no prior model: immigration from Europe, the electoral system, even the homogeneous style of the houses. What he saw in the US, as he mentions to a friend, was not the republic he had dreamed of, but rather, without doubt, what accompanies democracy.

Some argue that from a purely literary viewpoint, Sarmiento's the best book is his autobiography, *Memories of Provincial Life*, published in 1850, when he envisioned that the Rosas dictatorship was coming to an end. Rivals dismissed the book as a self-serving self-portrait of an aspiring presidential candidate. There is, of course, some truth in this depiction, but it would be wrong to reduce the text to a simple political ploy. *Memories* was not the first time Sarmiento wrote in defense of his merits and honor; nor was it a secret that he had a high opinion of himself. *Memories of Provincial Life* also reads as a defense of Sarmiento's public career. Nevertheless, a good part of the book is a reminiscence of ancestors and life in the province before and after independence. What mattered to Sarmiento, in other words, was not only the history of himself, but the local portion of a larger past from whence he came. In this evocative text, the discussions of customs take turns with physical and moral portraits, political and historical judgments, family memories and the vicissitudes of one's biographical path. In short, *Memories of Provincial Life* weaves autobiography with history as he praises his Creole lineage and subsequent career. Sarmiento understood, first and foremost, that he was the progeny of his works (Altamirano and Sarlo, 1997).

Sarmiento and His Interpretation of South American *Caudillismo*

The *caudillo* and *caudillismo*, the century-long domination of continental politics by military leaders after liberation, have been important issues in the historiography, social thinking, and the Hispanic narrative throughout South America. The book that began this tradition is Sarmiento's *Facundo*, making this a celebrated work of intellectual history, not only for Argentina, but also for all of Spanish America (Sarmiento, 2012). This famous work originally appeared as a series of installments in the Chilean newspaper *El Progreso* that he directed. Announcing the publication of this work, Sarmiento alleged that what had driven him to write the series was the desire to counter the Rosas propaganda against Argentinian emigrants in Chile. But even in the introduction, he realizes that his ambition was more than simply adding another chapter to the abundant anti-Rosas literature. His ambition was to explain the Argentinian civil wars and the nature of *caudillismo* in South America; to tell a biographical tale, full of strange and interesting events, within an environment that was still uncultivated; to disseminate an outline of a political and social agenda.

Why a biography? Sarmiento loved the genre, which can be attributed to his enormous cognitive and pedagogic power. He thought the synoptic capacity of the biography (the life of an eminent individual, a representative man) would allow for an evocation of the customs and spirit of an age and simultaneously make historic events accessible to the people. *Facundo* is thus not a work of historical scholarship, but it is a work of historical and sociological imagination (Salomón, 1984). Sarmiento had cultivated this imagination in his diverse readings – the historical novels of Walter Scott; then later, with the literature of the Romantic movement in France. Later, in Chile, there would be the discovery of a new history, a new way to conceive and practice the study of the past, that of historic Romanticism. The authors he praised were Francois Guizot, Augustin Thierry, and Jules Michelet. As Sarmiento saw it, France was the center of a historic vision that fascinated him,

namely, the view that philosophical history, inherited from the Enlightenment, could be expanded into a social history that included in its scope not only the customs, belief, and institutions of civil and governmental life, but also the action and confrontation of collective social and cultural forces (Orgaz, 1940).

The eminent individual, the representative man, of the book is the Argentinian *caudillo* Facundo Quiroga, a real person and not a fictional creation. In Sarmiento's eyes, Facundo was not just a *caudillo*, but the incarnation of a social force, an organic expression of Argentinian life as it had been created by Spanish colonization and the peculiarities of the physical space, of the vast plains of the Pampa. The task of the book is to explain this character and his social projection by the data of the physical and historical environment in which he lived. Hence the two parts that divide the history of Facundo: the first, which fills the first four chapters, involves the landscape of the nearly uninhabited Pampa, the stage on which he is going to represent the historical drama; then secondly, which covers the final nine chapters, the exegesis of the protagonist with his fierce appearance, his ideas and his way of working. The author believed that in this he was adopting the principles of the Romantic conception of history, where a harmony exists between the character and his environment, that each mutually reflects the other (Halperin Donghi, 1996).

The stage in which the figure of the *caudillo* will make its appearance, in his most authentic emanation, is the Pampa. In this vast and sparsely populated plain, long populated by Indians and Spaniards and their joint offspring the *mestizos*, Sarmiento observed that the colonial years entailed a way of life already forged that was distinct from that in the urban centers. Primitive, hard, and exposed to the immediate pressures of the environment, and distant from the law and doctrines of the city, the way of life of the countryside cattle rancher had created its own customs and social types, among which the exceptional character was the *gaucho*, the Pampa's cowboy. The knowledge and skills required to survive in this environment were rigorous – skill with the horse and the knife, the skills of the *baqueano* (trail guide), and the *rastreador* (tracker); and also the values of the habitants of this elemental world. The essential requirements for the rudimentary tasks of the cattle rancher and *gaucho*, their lives permanently surrounded by danger, are the stuff of epics. Life was rough, unsocialized; reputation meant nothing. What produced fame were the abilities of the gauchos and their performance in constant tests of physical courage. This was the environment of 'barbarism,' a term that in the ideological language of the day was a diatribe. And the antithesis of this barbaric space is the city. The city is the realm of law, commerce and ideas, the nucleus of 'civilization,' or at least a European-style civilization but surrounded by South American nature, the Pampa, the desert. The city is the place for cultivated, scholarly men.

Sarmiento observed that until 1810 – when the revolution began – two forms of human establishment, two societies, one next to the other, coexisted in the future territory of Argentina. Both resulted from the combined forces of South American nature and Spanish colonization and each unfolded in its own context. Each of these two societies sheltered its own spirit and principles. One was the city, the seed of European civilization or civilization itself; the other was the countryside, the cradle of barbarism, the antagonist of civilization. Both remained indifferent to each other until the revolution of 1810 put them into active contact.

The revolution, Sarmiento observed, was in fact two revolutions that largely ignored one another until external forces drove them into conflict. One was the revolution against Spanish domination, and it was driven by European ideas of freedom, progress, and modern institutions and largely led by the intellectual elite. The other was the revolution against the city, led by the gauchos, cattle ranchers, and military men who embraced 'barbarism.' Sarmiento formulated an interpretation of the era that, to him, shed light on the Argentinian convulsions. The revolutionary movement, in short, activated a double struggle: the war in the cities that fought against the Spanish by those seeking advancement of a more refined European culture; and the war against the cities, fought by the *caudillos*, representing the spirit of the countryside. The object of this second war led by the *caudillos* was not to put an end to Spanish authority, but to all authority and civil order. For the countryside, the revolution was only an opportunity to deploy, in a larger arena, all the tendencies that were hostile to the civilized spirit of the city.

Within this interpretive formula, which Sarmiento thought would clarify the issues that were tearing apart Argentinian political life, the story of the life of Facundo Quiroga begins. If this interpretative scheme explained the conditions and general trends that set the stage for the emerging *caudillo* in the province of La Rioja, the biography proposed a further link to a singular and representative destination, namely the discontinuous and scattered elements of collective history. In the exposition, Sarmiento develops a 'language of knowledge' – an historical science of the 'new way of being,' the American spirit, which in his view had not yet received adequate intellectual representation.

Sarmiento was not unaware of the 'poetic side' of barbaric life and at times his text delivers an admiring description of the nature and personalities of the crude and elemental rural world. However, Sarmiento (like Romantic Argentinians in general) accompanied his teachers, the European Romantics, only to a certain point: he did not look for nor did he discover the pastoral Eden, the culture of the original people, a culture that had been lost in the cities through the works of a cosmopolitan civilization. The aesthetic value of the countryside, the admiration of the wisdom of the cattle rancher, the empirical wisdom of the *baqueano*, or the inspiration of the singing gaucho do not imply a critique of progress nor a corrective to civilization. The reality of the barbarians, in Sarmiento's mind, was foreign and exotic, not the wellspring of an autonomous or alternative society.

What Sarmiento valued in the idea of civilization was not only civilized habits and institutions – manners, refinement of customs, the schools, the courts, business, the arts of industry, the cultivation of readings, etc. – but also something even more basic, something that can be captured in what the countryside shepherd does not provide. And that is the regular contexts that promote interactions among men, interactions that can moderate the impulses of 'natural' (unsocialized) men and generate a sense of and interest in the public good. The city, in contrast, multiplies the sites for these interactions. While the pastoral countryside disperses its habitants and their energies,

the city concentrates them and inserts them in the settings of civilization. Finally, the city engenders public space – space for animated deliberation, space where enlightened citizens can express themselves. Thus, the countryside destroys itself when its representatives seize the city. From that moment forward, opinion can no longer limit power. For Sarmiento the *caudillo* regime is a kind of despotism, a type of government in which one commands without submitting to any law.

Sarmiento did not invent every element of this interpretation of the social world that existed before the movement that began in 1810. But his book was never intended or described as the fruit of a learned and serene study. Rather, it is a book of combat and confrontation. There is a fierceness of the pamphlet that often seizes Sarmiento's text. This fierceness tirelessly curses the *caudillos* but then prophecies a bright future when the country would open itself to European immigration and allow for the navigation of its interior rivers.

Posterity has judged that this magnificent book was not just wrapped in the political passions of that time. Sarmiento and his *Facundo* do not live just in the libraries, scholarly studies, and university teaching: they also live in the ideological debates that occurred in Argentina in the twentieth century, especially after 1930, and that continue to the present day. Indeed, the *Facundo* still inspires concern about the two Argentinas, about the antagonism between scholars and *caudillos* that some believe summarizes Argentine national history. Still, neither the figure of Sarmiento, the intellectual legislator, nor that of *Facundo*, the barbaric *caudillo*, resonate solely within Argentinian culture. They resonate throughout contemporary Latin America (Brading, 1992).

See also: Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Civil Wars; Civilization, Concept and History of; Elites: History of the Concept; Intellectuals: History of the Concept; Liberalism, Anthropology of; Progress: History of the Concept; Revolution; South America, Archaeology of.

Bibliography

Altamirano, C., Sarlo, B., 1997. Una vida ejemplar: la estrategia de *Recuerdos de provincia*. In: Ensayos argentinos: de Sarmiento a la vanguardia. Ariel, Buenos Aires.

Anderson Imbert, Enrique, 1960. Genio y figura de Sarmiento. Eudeba, Buenos Aires.

Botana, Natalio, 1984. La tradición republicana. Alberdi, Sarmiento y las ideas políticas de su tiempo. Sudamericana, Buenos Aires.

Brading, David, 1992. Civilización y barbarie. In: Orbe indiano. De la monarquía católica a la república criolla. FCE, México.

Bunkley, Allison Williams, 1952. The life of Sarmiento. Princeton University Press, New Jersey.

Gálvez, Manuel, 1945. Vida de Sarmiento, el hombre de autoridad. Emecé, Buenos Aires.

Halperin Donghi, Tulio, 1972. Revolución y guerra. Siglo XXI, Buenos Aires.

Halperin Donghi, Tulio, 1995. Una nación para el desierto argentino. In: Proyecto y construcción de una nación (Argentina 1846–1880). Ariel, Buenos Aires.

Halperin Donghi, Tulio, 1996. Facundo y el historicismo romántico. In: Ensayos de historiografía. El cielo por asalto, Buenos Aires.

Orgaz, Raul, 1940. Sarmiento y el naturalismo histórico. Imprenta, Argentina, Córdoba.

Romero, José Luis, 2001. Situaciones e ideologías en América Latina. Universidad de Antioquia, Medellín, Colombia.

Salomón, Noël, 1984. Realidad, ideología y literatura en el "Facundo" de Sarmiento. Rodopi, Amsterdam.

Sarmiento, D.F., 1943. Prosa de ver y pensar. Emecé, Buenos Aires.

Sarmiento, D.F., 1993. Viajes por Europa, Africa i América. Colection Archivo, critical ed. UNESCO, Paris.

Sarmiento, D.F., 2012. Facundo. Eudeba, Buenos Aires.

Verdevoye, Paul, 1963. Sarmiento, éducateur et publiciste, entre 1839 et 1852. Institut de Hautes Etudes de l'Amerique Latine, París.

Weinberg, Félix, 1958. El Salón Literario de 1837. Hachette, Buenos Aires.

Weinberg, Gregorio, 1984. Los modelos educativos en la historia de América Latina. Kapeluz, Buenos Aires.

Satisfaction and Commitment in Long-Term Heterosexual Relationships

Aaron Ben-Ze'ev, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Satisfaction and commitment are crucial for maintaining long-term heterosexual romantic relationships. In light of the relatively brief duration of emotional experiences and the abundance of alluring romantic alternatives, maintaining a long-term romantic relationship is seen to be 'against all odds.' The presence of many tempting alternatives to the present relationship generates romantic compromises and makes it extremely difficult to be happy with one's own relationship. In order to ensure their long-term survival, there is a need to revise the nature of heterosexual relationships. A nurturing approach is suggested for this purpose.

Profound Satisfaction and Superficial Pleasure

Satisfaction involves pleasurable feelings as well as an element of fulfillment associated with contentment. When you are satisfied, you feel good about what you have and what you have achieved; you have had enough and you are happy with it.

Profound activities have a lingering impact upon our life and are essential for our development and flourishing. Superficial activities have a more limited impact: they affect only the surface of our lives; their impact becomes significant only if we frequently engage in them. Profound satisfaction, which is associated, for example, with profound love, profound involvement in work, or profound intellectual activities, is essential to human flourishing since it involves optimal functioning that, over time, develops the agent's essential capacities and attitudes. Part of profound satisfaction is the ability to overcome problems and feeling a sense of self-fulfillment while doing so. Superficial pleasure, such as that derived from, for instance, casual sex, watching television or gossiping, is an immediately rewarding, relatively short-lived experience requiring few or no profound human capacities.

Fulfilling ideal romantic love which endures for a long time is hard to achieve and hence low satisfaction and feeling of being romantically compromised are quite common. Unlike the total and uncompromising nature of ideal love, in reality, love comes in degrees and entails various types of behaviors. Preventing dissatisfaction typically requires idealization and positive illusions. While extreme romantic beliefs generally tend to engender unrealistic standards, a certain degree of idealization is usually beneficial.

Life is complex, and we cannot have it all. Knowing our limitations is one of our great virtues. Perhaps one secret of profound satisfaction in heterosexual relationship is not striving for the best of all worlds where passionate romantic love is always in the air in an intense manner, but rather knowing how to be satisfied with one's smaller lot, which might not always be filled with such love. Love is important, but love is not all you need.

Being satisfied with your own lot does not necessarily mean lowering your standards, but rather the ability to see the positive aspects of your situation. Ideals are necessary in human life, but it is useful to be aware of the limitations that reality poses upon the fulfillment of our ideals. Accordingly, compromises are inevitable. Any substantial disregard of reality is likely to lead to other problems.

Commitment in Long-Term Heterosexual Relationships

Commitment involves a pledge to do something in the future. Robert Sternberg (1988) suggested that romantic relationships characteristically have three major components: intimacy, passion, and commitment. Intimacy expresses closeness, bonding, and connection. Passion is largely the expression of desires and needs and is expressed in physical and mental arousal. Commitment is the extent to which a person is likely to stick with someone and see her or him through to the end.

Commitment is essential for long-term romantic relationships; thus, the lack of it has been found to be a significant predictor of relationships breaking up. Commitment theory distinguishes between factors that motivate connection versus factors that increase the costs of leaving. Loving someone motivates us to establish a romantic connection with that person, while being married discourages us to leave the relationship because of its high cost. Commitment theory adds another important factor: the alternatives to the present situation. Commitment can be seen then as the degree to which attracting forces overwhelm repelling forces, despite the given alternatives. Commitment is strengthened by the amount of satisfaction and the extent of the cost, and weakened by possible alternatives to that relationship. Satisfaction level was found to be significantly more predictive of commitment than were the quality of alternatives or the amount of cost of separation. This suggests that the quality of the relationship has the greatest impact upon its continuation, much more than external factors such as the cost of switching or the available alternatives. However, when satisfaction is not high, the extent of the cost and the attractiveness of the alternatives can be quite significant (Benjamin and Agnew, 2003; Jones and Adams, 1999; Rusbult, 1983).

At the basis of the psychological objections to marital commitment is the fact that while marriage is a committed, stable mode of living, intense emotions are essentially transient and unstable states. Taking love to be the essence of marriage runs the risk of ruining either marriage or love. In response to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.35035-8

this objection, it can be claimed that a commitment to love is not aimed at creating emotions from scratch; it is aimed at sustaining an already-present dynamic. A commitment to love does not imply manipulation of emotions; it implies the intention to ignore the inevitable periods in a relationship when the emotion of love is not experienced. Marriage is an act of placing trust in the relationship, rather than merely hoping or expecting it to succeed. Such a trust is essential, since it tends to be self-fulfilling. Indeed, many contemporary marriages still last for life. Nevertheless, commitment to a relationship remains conditional on the partnership retaining its value.

Difficulties in Maintaining a Long-Term Romantic Relationship

Two major factors increase the difficulties of maintaining satisfaction and commitment in long-term heterosexual romantic relationships in present-day society: (1) the great role that changes play in generating emotions, and (2) the greater plausibility of leaving a current relationship and starting a new one. These two factors significantly reduce commitment to and satisfaction from current relationships.

The first difficulty, which is related to the nature of the emotional system, refers to the fact that intense emotions are typically transient and unstable states, whereas stability is needed to generate the satisfaction we experience in long-term committed relationships. Emotions are usually experienced when we perceive a significant change in our situation (Ben-Ze'ev, 2000). Whereas change intensifies emotions, commitment entails stability. Although long-term relationships involve some changes even as they remain intact, the most significant change, that is, changing the partner, is absent. The sexual excitement and satisfaction experienced initially with a new partner is often difficult to sustain in long-term relationships, as routineness takes over. However, the importance of sexual excitement remains high, and marital satisfaction is associated with sexual desire (Brezsnyak and Whisman, 2004).

The second major difficulty for maintaining satisfaction in long-term romantic relationships refers to certain developments in current society that have increased the availability of romantic partners. In current society, we are witnessing the lifting of most of the constraints that once prevented long-term committed relationships from dissolving; another factor in this regard is the apparent presence of many attractive alternatives.

Nowadays, being satisfied and staying within a committed relationship has become a more difficult option – one that requires the partners to constantly reexamine the value of their relationship in light of, among other issues, the presence of love. Lovers might face not only constant doubts about which road to take, but also constant regrets about the many roads not taken. The abundance of alternatives and the perpetual possibility of achieving something 'better' can undermine commitment and satisfaction.

The notion of the availability of alternatives may explain many seemingly puzzling situations, such as people who remain in unsatisfied marriages or jobs. Although their satisfaction from the present situation is low, these people perceive other possible alternatives to be even worse. The other direction is found as well: people leave reasonably satisfied situations for unknown risky ones; this is because the evaluation of possible alternatives often supersedes those of existing positions. Since the realm of infinite romantic possibilities so overwhelms us with tempting alternatives, we are unable to reasonably consider the present situation as romantically satisfactory. We have become slaves of the possible and spoilers of the present.

Having the freedom to choose between many alternatives has its own cost; excessive freedom can increase people's uncertainty, insecurity, dissatisfaction, and depression. Therefore, the availability of a greater choice can be a mixed blessing. Indeed, sometimes adding options makes the task of choosing less attractive and more cumbersome; consequently, there are people who (occasionally) prefer others to make choices for them. For example, many people in loveless marriages would be happy if their spouse initiates a separation, rather than having to raise the issue themselves. As Barry Schwartz (2000) shows, too much freedom from constraints is a bad thing, as unconstrained freedom leads to paralysis and becomes a kind of self-defeating tyranny. He argues that due to the multiplicity of choices available at all times and on all fronts, people no longer know how to be satisfied with 'just good enough.' They always seek perfection. Such freedom also undermines the notion of deep commitment and social belonging to groups and institutions, bonds that are vital to our mental health. Freedom constrained by ideals and boundaries may in fact be more satisfying and less dangerous than unconstrained freedom.

The above difficulties make romantic dissatisfaction a real threat to marriages and other committed relationships. The feeling of being romantically compromised becomes more dominant.

Satisfaction and Romantic Compromises

The need to make a commitment indicates that there is a potential conflict between long-term satisfaction and the commitment. However, both commitments and satisfaction are important parts of human social life.

In romantic compromises, we give up a romantic value, such as romantic freedom and intense passionate love, in exchange for a nonromantic value, like the wish to live comfortably without financial concerns. Nevertheless, in our hearts we keep yearning for the road not taken – the one with greater romantic freedom and with a more attractive partner (Ben-Ze'ev, 2011).

We may discern two major types of romantic compromises: (1) compromises on romantic freedom that are made when entering a marriage (or another committed relationship), and (2) compromises on the choice of partner. Romantic compromises involve two major concerns: (1) giving up alluring possible alternatives while still yearning for them, and (2) accepting negative aspects of the partner. Both concerns are present in the two types of compromises, but with different weightages. The concern regarding giving up alluring alternatives is high at both types of compromises but higher in the first. The concern regarding accepting negativity is mainly present in the second type.

In the type of compromises referring to the choice of the partner we may discern two major subtypes: (1) compromise on the overall value of the person, and (2) compromise on the value of the person as a romantic partner. In the first subtype, we compromise on something very essential to intimate relationships: having a partner who is up to our standards. Although this is strictly speaking not a romantic compromise, it typically leads to such a compromise as we feel that we deserve more. The second subtype expresses the very core of romantic compromise – there is no profound romantic love in this relationship – as something essential for love is missing. Whereas giving up romantic freedom underlies the inherent romantic compromise in marriage, the dissonance between the ideal, perfect partner and the actual partner is at the center of the compromise in choosing the partner. Due to the perceived nature of the ideal beloved, it is almost impossible to find such a person. However, when we take into account our own flaws, nonperfect people may be seen as perfect enough.

Both types of romantic compromise – giving up romantic freedom and accepting a partner with obvious flaws – indicate the necessity of compromise in committed romantic relationships. Many couples in less-than-perfect marriages would stay together if they understood that compromise is an essential feature of married life as well as of any other committed relationship. These people take an unnecessarily harsh view of their marriage because they fail to perceive the value of compromise in a less-than-perfect world. On the other hand, many people stay in a rotten relationship because they do not understand the difference between 'good' compromise and 'bad' compromise.

Factors Influencing Satisfaction and Commitment

So far, I have discussed some general considerations concerning satisfaction and commitment. There are other types of factors influencing satisfaction and commitment, for example, age, gender, education, religion, socioeconomic level, and personal characteristics and attitudes. I will now discuss a few of those factors that are most relevant to the above considerations. I focus on inequality, respect, sacrifice, age, and personal space.

Inequality

Two major types of inequality in heterosexual relationships are (1) inequality in a comparative status and (2) inequality in emotional involvement. These types of inequality correspond to the two subtypes of romantic compromises: (1) compromising on the overall value of the person and (2) compromising on the value of the person as a romantic partner.

Concerning the first kind of inequality, the equity theory postulates that those involved in an inequitable romantic relationship consider themselves to be undeserving of the situation and hence their satisfaction and commitment are considerably reduced. This is the case both for those perceived to be inferior, who feel guilty for receiving more from the relationship than they feel their partner does, as well as for those perceived to be superior, who feel that they deserve more. Involvement in extramarital relationships, which expresses both low commitment and low satisfaction, may serve as a way of compensating for such inequity. The 'superior' people may perceive extramarital relationships as something they deserve because their spouse gets more from the marriage than they do. The 'inferior' people tend to be involved in extramarital relationships in order to escape the unpleasant state of inequity and to prove to themselves and to their partner that they actually are deserving and attractive to the opposite sex (Prins et al., 1993). Indeed, feeling inferior to one's partner is associated with less commitment, less satisfaction, and less love for the partner (Murray et al., 2005).

In a similar vein, people who stand to lose the most are apt to be the least likely to risk ending the relationship by having another sexual partner. People who feel that they are compromising will be more likely to take romantic risks. Thus, if the woman has a higher level of education than her partner, she is more likely to risk the relationship by having a secondary sexual partner than if both members of the couple have equal levels of education (Forste and Tanfer, 1996).

In the short term, the inequities might give rise to great admiration and hence may increase love and sexual desire. Thus, people who can provide us with social status, such as the rich, the famous, and the powerful, will generate more intense sexual desire and sexual satisfaction. In these cases, we are less likely to feel that we are compromising. A similar impact can be found in the other direction; a low level of an admirable characteristic, such as intellect, reduces sexual satisfaction. In the long term, significant inequalities become problematic for both sides, whereupon superficial short-term goals (such as being associating with a famous person) become less important. For example, the 'higher status' person may begin to show a lack of reciprocity, which will eventually damage the 'lower status' person's love and could even generate negative emotions such as envy, jealousy, and anger.

The situation is made more complex by the fact that the extent of the gap and the overall comparative value of each partner also play a part. Inequality in certain domain, such as attractiveness, can persist if the overall comparative value is similar. In these circumstances, the partner's inferiority in one domain may be compensated by superiority in another. Thus, when people are certain of their worth, they may prefer a partner who is a bit superior to them in one domain and hence will be more beneficial for them. In this case, admiration may be the relevant emotion. Thus, in one study, 89% of high-achieving men report that they would like to marry or have already married a woman who is as intelligent as they are, or who is more so (Whelan, 2006). These men believe that in marrying such woman they have made the better deal.

We may also speak about unequal emotional involvement between partners. This inequality occurs in most romantic relationships and has a considerably negative impact upon marital satisfaction and commitment, therefore increasing the likelihood of a breakup. The consequences of this inequality have been described as an example of the 'principle of least interest.' The least interested partner is less committed and has more control over the continuation of the relationship. Accordingly, this partner is often (but not always) the one who terminates the relationship (Sprecher et al., 2006). A relationship involving unequal romantic involvement is a compromise for both partners, as the low level of satisfaction is common to both of them; it is less clear who is making the more profound compromise.

Respect

Mutual respect, which is central to marital satisfaction and commitment, is related to the issue of equality. Respect is less likely to present in relationships characterized by inequality in terms of the overall value given to the partnership or inequality in the partners' emotional involvement.

Respect, which is an attitude and not an emotion, is directed toward a partner who is trustworthy, considerate, and accepting. Respect has been found to be a stronger predictor of relationship satisfaction than measures of attachment, love, and liking (Frei and Shaver, 2002). Respect, which involves both caring and equality, enables the partner to enjoy reciprocity, autonomy, flourishing, and personal growth.

A certain measure of distance exists in respect that allows partners to live as they wish, regardless of how good they are to their partner. This respect gives them a personal space, which has its own intrinsic merits. Sexual desire is not about respect, but about breaking distance and boundaries. Romantic love includes both an appropriate distance that allows respect and a close emotional involvement, which is involved in intimate sexual relationship.

John Gottman (1998), who is known for his work on marital stability and relationship analysis, noted that most couples want two main things from their marriage – love and respect. Indeed contempt, which is the opposite of respect, is one of the major predicators of divorce. Another study listed respect as one of four core relationship values, the other three being commitment, intimacy, and forgiveness. Yet another study found that people often mention respect spontaneously when talking about why their relationships function well or poorly and that respect is one of the five most frequently mentioned and valued features in the romantic partner – the others are being accepting, trustworthy, honest, and loyal. Respect was also found to be closely connected with marital satisfaction. Disrespect is a powerful expression of lovelessness in marriages. Respect implies our right to be treated in a way that enhances our positive self-esteem. In intimate relationships, respect implies and asserts equality. The importance of equality in successful marriages and intimate relationships is seen in the crucial role of respect in romantic relationships and the potential damage that lack of respect inflicts on such relationships (Hendrick and Hendrick, 2006).

Sacrifices

In close relationships, the willingness to sacrifice is associated with strong commitment to and high satisfaction in the relationship. Sacrifice is a potent longitudinal feature of marital adjustment. Accordingly, sacrifice is also associated with good relationship functioning and higher quality relationships. People identify sacrifice with caring, trust, respect, and loyalty and hence with what they consider as love. The tendency to sacrifice expresses more profound love and enhances a sense of security, which is assumed to be essential for marital success. The benefits of sacrifice are especially pronounced when people sacrifice in order to promote the other's well-being or to enhance intimacy in their relationship (Impett and Gordon, 2008; Van Lange et al., 1997).

Sacrifices can be harmful when they are extensive and are not reciprocal, such as when one partner (usually the woman) carries the full burden of sacrifice in the relationship. Extensive sacrifices underlie the traditional other-validated model of love in which the other is the ultimate preoccupation of a person's meaningful world. According to this view, sacrificing one's life for the beloved is a reasonable option. It is often the case that people who lack power in their relationships (once again, often women) are more likely to engage in sacrifice. This is made even worse when those making the sacrifice are 'silencing' their own opinions and desires in the relationships.

Reciprocity is important in generating the will to sacrifice, as it indicates that the other is equal to you and will make sacrifices for you if needed. Willingness to sacrifice is associated with greater relationship satisfaction and stability. Despite the fact that sacrifices are intended to promote the others' well-being and not that of the sacrificers, the latter may also benefit from the sacrifice, as it can increase their self-image, enhance others' evaluation of them, and raise the chances that their partner will make sacrifices for them in return. Accordingly, they too might gain from their sacrifice (Impett and Gordon, 2008).

Age

The common assumption that happiness declines with age is mistaken. On the contrary, research indicates that older people are actually happier and more satisfied with their lives than younger people are. One possible explanation is that when we realize that our years are numbered, we change our perspective and tend to focus on meaningful and positive present experiences. Our emotional experiences are more likely to consist of peacefulness and serenity rather than excitement and joy. Sonja Lyubomirsky summarizes these findings by noting that for most people, the best years are in the second half of life (Carstensen, 2009; Lyubomirsky, 2013).

In accordance with the above claims, it was found that older individuals perceive their spouse as warm during both disagreement and collaborative tasks and report high marital satisfaction. Older married couples have fewer marital conflicts than their younger counterparts although they report that erotic bonds are less central in their lives. Companionate love, which is based upon friendship, appears to be the cardinal feature of their lives. Intimate relationships in old age are, by and large, harmonious and satisfying (Berscheid, 2010; Charles and Carstensen, 2009).

Despite the above findings, there are still disagreements concerning the precise nature of the relation between marital satisfaction and age, one reason being that the notion of 'satisfaction' is perceived differently at various ages.

Many studies claim that overall marital satisfaction follows a U-shaped association between marital duration and marital happiness, with satisfaction high in the early years of marriage, relatively low in the middle years, and returning to higher levels in old age (Charles and Carstensen, 2009). This claim is disputable and some studies suggest that the U-shaped results are artifact from the selection process of the married couples. Their studies suggest that marital satisfaction does not increase in the later years of marriage (VanLaningham et al., 2001).

It appears that the ambiguous findings concerning the association between age and marital satisfaction (and with happiness in general) are due to the shifting meaning of happiness and satisfaction at different ages. The main difference is that the meaning of happiness moves gradually from excitement to peacefulness (or contentment) as one gets older (Mogilner et al., 2011).

I will discuss the implications of this claim further in the concluding section. In any case, it seems that the problem of romantic compromises becomes less acute as people grow older. The major factors of such compromises, that is, the presence of negative features in the partner and yearning for the possible, are of lesser weight in old age. Over time, people get accustomed to their spouse's negative aspects. They learn to live with them while minimizing their negative impact. In older age, when people realize that their time is running out and that their available alternatives are decreasing, they are more likely to accept their limitations and not feel they are compromising when they do not pursue an attractive option. Elderly couples more readily take the attitude of being happy with their lot.

Personal Space

The issue of personal space is directly connected to the essence of romantic compromises: giving up a seemingly better alternative and hence yearning for the possible. In a flourishing relationship, the importance of a significant personal space cannot be exaggerated; its absence expresses a profound romantic compromise. Personal space does not necessarily involve sexual freedom. Thus, in commuter marriages, where personal space is significantly higher than in typical marriages, commitment is high and the percentage of extramarital affairs is similar to that of standard marriages; sexual satisfaction is also similar (Gerstel and Gross, 1984; Kelmer et al., 2013; Stafford, 2005).

Online relationships are similar to commuter marriages in having a significant personal space, and in assuming that coresidence is not necessary for loving relationships. However, unlike commuter marriages, commitment in online relationship is very low.

It seems that some kind of distance, providing a greater personal space, is important for close relationships. Distance may focus the partners' attention on the profound aspects of their relationships and help them to disregard the superficial ones. Significant and temporally extended physical distance may harm relationship, but a more limited distance may be beneficial. As the saying goes, "Absence makes the heart grow fonder."

Hedonic Adaptation

Emotions typically occur when we perceive positive or negative significant changes in our personal situation, or in the situation of those related to us. A major positive or negative change significantly improves or interrupts a stable situation relevant to our concerns. Like burglar alarms going off when an intruder appears, emotions signal that something needs attention. When no attention is needed, the signaling system can be switched off. We respond to the unusual by paying attention to it.

In light of the crucial role of change in our emotional life, humans have a built-in tendency to return to a relatively stable level of happiness after major positive or negative events. This phenomenon, which is called 'hedonic adaptation,' suggests that the joys of new loves and new triumphs and the sorrows for new losses will fade with time. Hedonic adaptation constitutes an obstacle for maintaining a high level of excitement and satisfaction in long-term loving relationships. Such adaptation also indicates that long-term profound love is not mainly based on features such as sexual desire, which greatly depend on change and hence are more under the influence of hedonic adaptation; rather, it depends on features such as friendship, caring, and reciprocity, whose value remains constant (and even increases) as time goes by.

Another implication of hedonic adaptation is that romantic compromises are part and parcel of long-term committed relationships. This is so since romantic compromises involve the limitation of one's romantic freedom and the need to be satisfied with a familiar partner, who is less exciting than a new one. However, hedonic adaptation can help us accommodate to a situation that at first is seen as compromise. Moreover, intervention in order to increase one's happiness in marriage is possible if the agent knows about, endorses, and is committed to the intervention (Lyubomirsky, 2013).

Hedonic adaptation works within certain limits. Extreme experiences, such as intense love or intense domestic violence, are typically outside these limits. In some cases, intense love can endure for many years without returning to the basic level of happiness that prevailed before this love. And some people adapt to domestic violence, although it is highly risky for them. As in many other aspects discussed here, hedonic adaptation applies to most average circumstances, but not to extreme situations.

Maintaining Long-Term Satisfaction

Satisfaction in marriage depends upon the couple's ability to be satisfied with their romantic lot. When people are satisfied with their lot, they do not feel that they are compromising. In this concluding section, I sum up a few major related factors that enable couples to maintain long-term satisfaction in marriage: personal flourishing, being integrated in a larger positive framework, and adapting to new criteria of satisfaction.

In seeking to cope with marital dissatisfaction and romantic compromises, the agent's own flourishing and development are of considerable importance. Satisfaction in marriage is not merely about making the beloved happy, but also about working to ensure that the lover is profoundly satisfied with the way in which the relationship is nurturing her own development and flourishing (Ben-Ze'ev and Goussinsky, 2008). The emphasis is shifting here from a model based upon other-validation to one of self-validation (Mitchell, 2002). Only a deeply satisfied person can provide love and happiness to another. The self-validated model is by no means a matter of egocentrism. It is not egocentric to attempt to nurture your capacities and genuine needs, while at the same time developing a loving equal relationship with another person.

The emphasis upon self-worth is also expressed in the central role played by promoting, rather than preventing, types of behavior. Accordingly, caring is more significant in love than prohibiting various types of sexual practices. Likewise, focusing on the partner's uniqueness is of greater importance than focusing on the exclusivity of the lover-beloved relationship.

In profound loving relationships, the flourishing and autonomy of the agent is central; this leaves more options open and reduces the feeling of captivity. The agent's general flourishing, which is expressed in having more options open to her, does not negatively affect her partner. On the contrary, the impact upon the partner is usually positive as well; the flourishing of one partner is likely to improve the relationship as a whole and benefit the other partner as well. This is true in those areas of the agent's life that do not involve the romantic relationship itself; however, in the romantic area itself, too much romantic freedom and the availability of more romantic options often reduce the satisfaction from and commitment to the partner.

The importance of the agent's personal flourishing is related to the importance of establishing marital satisfaction within the larger context of a happy life. In the same way, living healthily is not just eating healthily. It is a complex, dynamic framework that is related to many factors in our physical and psychological environment that determine healthy living; for instance, whether we engage in outdoor activities, are stressed, are socially active, and have balanced meals. The notion of a healthy life is complex and refers to various means taken in different ways and to different extents. Like healthy life in general, a healthy romantic relationship is also a complex framework involving various factors, such as friendship, sexual desire, caring, and reciprocity. The larger framework makes the relationship more solid, as it is based not on one narrow leg but upon a broader foundation.

An important feature in maintaining long-term marital satisfaction is the typical shift that occurs over time in the way that we rate such satisfaction. Cassie Mogilner et al. (2011) claim that whereas in their youth, people are more likely to associate happiness with excitement, as they get older they become more likely to associate happiness with peacefulness. This change involves a redirection of attention from the future to the present as people age. When people's attention is focused on the present, they are more likely to be happy with their own lot.

To sum up, maintaining satisfaction in a long-term marriage (and in other committed relationships) depends significantly on the framework in which this relationship exists. In some aspects, this larger framework extends the satisfaction to other aspects in the people's lives, such as personal flourishing, independent of the quality of the relationship. When people are highly satisfied with their personal activities, it is more likely that they will be satisfied with their relationship with the partner. From another aspect, the focus of the framework becomes more limited with time; in older age, the focus is more on the given present rather than on the larger unknown future. Both such extensions and limitations of the context in which the committed relationship exists can increase people's ability to be satisfied with their lot.

See also: Heterosexual Expression at Older Ages; Heterosexuality; Love and Intimacy, Psychology of; Sexual Motivation; Sexuality Over the Life Course.

Bibliography

Benjamin, L., Agnew, C., 2003. Commitment and its theorized determinants: a meta-analysis of the investment model. Personal Relationships 10, 37–57.

Ben-Ze'ev, A., 2000. The Subtlety of Emotions. MIT Press, Cambridge, MA.

Ben-Ze'ev, A., 2011. The nature and morality of romantic compromises. In: Bagnoli, C. (Ed.), Morality and the Emotions. Oxford University Press, Oxford, pp. 95–114.

Ben-Ze'ev, A., Goussinsky, R., 2008. In the Name of Love: Romantic Ideology and Its Victims. Oxford University Press, Oxford.

Berscheid, E., 2010. Love in the fourth dimension. Annual Review of Psychology 61, 1–25.

Brezsnyak, M., Whisman, M.A., 2004. Sexual desire and relationship functioning: the effects of marital satisfaction and power. Journal of Sex and Marital Therapy 30, 199–217.

Carstensen, L.L., 2009. A Long Bright Future. Broadway, New York.

Charles, S.T., Carstensen, L.L., 2009. Social and emotional aging. Annual Review of Psychology 61, 383–409.

Forste, R., Tanfer, K., 1996. Sexual exclusivity among dating, cohabiting, and married women. Journal of Marriage and the Family 56, 33–47.

Frei, J.R., Shaver, P.R., 2002. Respect in close relationships: prototype definition, self-report assessment, and initial correlates. Personal Relationships 9, 121–139.

Gerstel, N., Gross, H., 1984. Commuter Marriage: A Study of Work and Family. Guilford Press, New York.

Gottman, J., 1998. Why Marriages Succeed or Fail: And How You Can Make Yours Last. Bloomsbury, London.

Hendrick, S.S., Hendrick, C., 2006. Measuring respect in close relationships. Journal of Social and Personal Relationships 23, 881–899.

Impett, E.A., Gordon, A., 2008. For the good of others: toward a positive psychology of sacrifice. In: Lopez, S.J. (Ed.), Positive Psychology: Exploring the Best in People. Greenwood, Westport, CT, pp. 79–100.

Jones, W., Adams, J., 1999. Handbook of Interpersonal Commitment and Relationship Stability. Plenum, New York.

Kelmer, G., Rhoades, G.K., Stanley, S.M., Markman, H.J., 2013. Relationship quality, commitment, and stability in long-distance relationships. Family Process 52, 257–270.

Lyubomirsky, S., 2013. The Myths of Happiness. Penguin, New York.

Mitchell, S.A., 2002. Can Love Last? The Fate of Romance over Time. Norton, New York.

Mogilner, C., Kamvar, S.D., Aaker, J., 2011. The shifting meaning of happiness. Social Psychological and Personality Science 2, 395–402.

Murray, S.L., Rose, P., Holmes, J.G., Derrick, J., Podchaski, E.J., Bellavia, G., Griffin, D.W., 2005. Putting the partner within reach: a dyadic perspective on felt security in close relationships. Journal of Personality and Social Psychology 88, 327–347.

Prins, K.S., Buunk, B.P., VanYperen, N.W., 1993. Equity, normative disapproval and extramarital relationships. Journal of Social and Personal Relationships 10, 39–53.

Rusbult, C.E., 1983. A longitudinal test of the investment model: the development (and deterioration) of satisfaction and commitment in heterosexual involvements. Journal of Personality and Social Psychology 45, 101–117.

Schwartz, B., 2000. Self-determination: the tyranny of freedom. American Psychologist 55, 79–88.

Sprecher, S., Schmeeckle, M., Felmlee, D., 2006. The principle of least interest: inequality in emotional involvement. Journal of Family Issues 27, 1255–1280.

Stafford, L., 2005. Maintaining Long-distance and Cross-residential Relationships. Lawrence Erlbaum, Mahwah, NJ.

Sternberg, R.J., 1988. The Triangle of Love. Basic Books, New York.

Van Lange, P.A.M., Rusbult, C.E., Drigotas, S.M.A., Arriaga, X.B., Witcher, B.S., Cox, C.L., 1997. Willingness to sacrifice in close relationships. Journal of Personality and Social Psychology 72, 1373–1395.

VanLaningham, J., Johnson, D.R., Amato, P.R., 2001. Marital happiness, marital duration, and the U-shaped curve: evidence from a five-wave panel study. Social Forces 79, 1313–1341.

Whelan, C.B., 2006. Why Smart Men Marry Smart Women. Simon & Schuster, New York.

Sauer, Carl Ortwin (1889–1975)

Michael Williams, University of Oxford, Oxford, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13490–13492, © 2001, Elsevier Ltd.

Abstract

Although formally a geographer writing on early humans and agriculture in Mexico and the southwestern USA, Carl Sauer's writings and concerns straddled a wide range of topics, including plant domestication, colonial destructive impacts, modern mass production society, academic freedom, the role of the social sciences, cultural diversity, and environmental quality. His early rebellion against environmental determinism in geography led to collaboration with anthropology and the adoption of culture as the ultimate shaping force in geography. His pithy style and simple, straightforward language meant that latterly he had a following among avant-garde garde poets, literati, and environmentalists in California and elsewhere.

Carl Sauer was one of the towering intellectual figures of the twentieth century, not only in geography but also in a wider intellectual sphere. However, because of his wide-ranging thought, speculative sweep, and world perspective over a number of subjects, it is difficult to define his contribution to the social and behavioral sciences precisely and easily. But one can say that some of his underlying concerns were about scholarship, independence of thought, opposition to academic bureaucracy, a sympathy and identification with rural folk, concern for cultural diversity and environmental quality, and a distaste for the technological and scientific 'fix,' particularly the solutions offered by the emerging social sciences after 1945.

He was born in Warrenton, Missouri on 24 December 1889 of German parents, and because of his background and 3 years of schooling in Calw (near Stuttgart) he was influenced by German culture and literature. He completed his graduate studies at Chicago University in 1915 under the geographer Ellen Semple Churchill, the geologist Rollin D. Salisbury, and plant ecologist Henry C. Cowles, the latter two of whom made a lasting intellectual impression on him. He then taught at the University of Michigan until 1923, when he moved to geography at Berkeley where he taught for 34 years (32 as Chair) and established one of the most distinctive graduate schools of American Geography that would always be associated with 'cultural geography.' After he retired in 1955 he enjoyed 20 remarkably productive years that saw the publication of four books and a score of influential papers, all distinguished by big and speculative ideas, that were the fruits of a lifetime's reflection and unhurried reading. His reputation soared so that an 'aura of sage, philosopher-king, and even oracle surrounded him' (Hooson, 1981, p. 166). He died in Berkeley on 18 July 1975.

Cultural Geography

Sauer rebelled against the sterile environmental determinism of contemporary geography with its emphasis on humans as response mechanisms to physical factors. If nothing else, his experience in the Economic Land Survey in the Michigan Cutovers had shown him that humans radically transformed the earth, often for the worse, and in the process created cultural landscapes. In his search for a new, humane geography, cultural anthropology seemed to offer a means of dealing with the diversity of humankind and its cultural landscapes through time. On arriving in Berkeley he found natural soul mates in the anthropologists Alfred L. Kroeber and Robert H. Lowie. The concept of 'culture' subsequently pervaded all his teaching and writing. In *The Morphology of Landscape* he distilled an almost wholly German geographical literature, established the primacy of human agency in the formation of cultural landscapes 'fashioned out of the natural landscape by a cultural group,' and the importance of a time-based approach. In addition, he placed great importance on observation and contemplation in the field – a *Verstehen* or empathetic understanding and intuitive insight into behavior or object in order to achieve 'a quality of reasoning at a higher plane' than the tangible facts (see Sauer, 1925; Williams, 1983).

Sauer wrote 'Morphology' in order to 'emancipate' himself from determinist thinking and, with a few more papers on the cultural landscape of the midwest frontier, he put the epistemological game behind him and started substantive field work on early settlement and society in Lower California, Arizona, and Mexico. During the 1930s he produced many works, including *Aztatlan* and *The Road to Cibola* both published in a new monograph series *Ibero-Americana* that he founded in the same year with Alfred Kroeber and the historian H. E. Bolton. (For these and many other of Sauer's publications see Leighly, 1963.) These investigations drew him into the controversy about New World plant domestication and plant origins, and into collaboration with botanists, archeologists and ethnologists, whom he found congenial intellectual company.

Widening Horizons

All the time Sauer's horizons were getting wider, his ideas more speculative, and his ethical values more refined. Toward the end of the 1930s he wrote a slight, but ultimately influential, paper which was a sustained and biting critique of the destructive social and environmental impact that resulted from the predatory outreach of Europe, which had few counterparts at that time except, perhaps, in the writing of Karl Marx (Sauer, 1938). He drew inspiration from the work of George Perkins Marsh on the human transformation of the earth and

Ernst Friedrich's concept of *Raub-wirtschaft*, or destructive exploitation (see Marsh, 1865, 1965; Friedrich, 1904). His experience and knowledge of Latin and Central America and their history suggested to him that the Spanish conquest had led to a devastating and permanent impoverishment of the land and of its cultures and societies. Disease, warfare, and enslavement had disrupted traditional value systems. Thus, the diffusion of technologically superior societies could affect humans and their culture just as much as it could physical resources.

But two works more than any others established his world reputation and heralded a remarkable decade of multifaceted yet interrelated speculative understanding of the place of humans on earth. First was *Agricultural Origins and Dispersals* (Sauer, 1952a) that flowered later into a string of publications into the human uses of the organic world, and early humans in the Americas from the Ice Age onward. Unfortunately, radiocarbon dating came too late to inform Sauer's writing, but although he may not have provided the answers, he defined the questions brilliantly.

Second, in 1956, with the collaboration of Marston Bates and Lewis Mumford, he masterminded the Princetown symposium on 'Man's Role in Changing the Face of the Earth,' the theme of which thereafter became his overriding interest. (See Mumford, Lewis (1895–1990).) All his learning and concerns culminated in this volume, and in his chapter "The Agency of Man on Earth" (Sauer, 1956). The capacity of humans to alter the natural environment – the 'deformation of the pristine' – the cult of progress and waste that stemmed from mass production ('commodity fetishism'), and the alien intrusion of humans into world ecology, were included. In contemporary terms, the theme was the degradation of the environment, and it was an early and influential statement. It also had another dimension: globally, the 'imperialism of production' was as bad as the old, colonial imperialism, and might ultimately be no better than Marxist totalitarianism; mass culture was eliminating not only biological diversity but also cultural diversity, and older and less robust societies. Somehow, humans had to rise above this mindless, short-term exploitative mode. 'The high moments of history have come not when man was concerned with the comforts and displays of the flesh but when his spirit was moved to grow in grace.' Therefore, what was needed was "an ethic and aesthetic under which man, practising the qualities of prudence and moderation, may indeed pass on to posterity a good Earth" (Sauer, 1956, p. 68).

His simply articulated ideas had a resonance with many activists and intellectuals, as well as Californian avant-garde poets and literati, who extolled his work as an example of cultural and ecological sensitivity and respect, tinged with deep historical insight and scholarship, and made attractive by his simple and pithy language. He also tapped a deep spring of feeling during the 1960s and 1970s at the time of Vietnam and student unrest with their concerns at the limits of the earth and technological/political power. Yet Sauer was a complex mix. He was congenitally nonconformist but deeply conservative, and although profoundly concerned with conservation was never formally an 'environmentalist,' and indeed, he thought the movement was little more than an 'ecological binge.'

Distrust of the Social Sciences

In many ways it is ironical that Sauer has a place in this *Encyclopedia* because he had such a deep distrust and distaste for the behavioral and social sciences. Although he eschewed methodological and epistemological discussion, his writing, and particularly his personal correspondence, reveal that he was "a philosopher in spite of himself" (Entrikin, 1984). Culture history became his model of social science, and it was drawn from the natural sciences, not the social sciences. He was basically a pragmatist who was influenced by the writings of the German cultural geographers Friedrich Ratzel and Eduard Hahn, and by the methods of geology and anthropology and their emphasis on the provisional character of working hypotheses which were no more than a means to an end. His later association in Berkeley with people who worked on 'tangible things,' like plant ecologists, agricultural scientists, botanists (e.g., Ernest Babcock), experts in the evolution of population genetics (see Wright, Sewall (1889–1988)), and geneticists reinforced this.

He argued for a theoretical and methodological pluralism that stemmed naturally from the inherent diversity of nature and culture. The natural science idea of a dynamic balance and diversity that arose from organic evolution was embedded deeply in his thought, and he felt that the modern world had disrupted these processes so that it was out of balance. Hence his deep distrust of US capitalism and all bureaucratic systems which would destroy diversity and local community. Liberal social scientists who designed, planned, and directed community life were more likely to destroy than enhance it by imposing universalizing concepts of social organization, and by ignoring the inherent pluralism and diversity of nature and culture that stimulated the naive curiosity about the world which was the essence of geography. Theoretical and normative social science as practiced by economists, sociologists, and political scientists, grated on him with its exaggerated confidence in the statistical and the inductive, and its 'dialectic atmosphere.' Their focus on the present precluded a better insight into the origins and evolution of any topic, and gave 'an exaggerated accent on contemporaneity.' His social science was based firmly in history and geography – this was culture history.

During the late 1930s he talked jokingly of the two patron saints of social scientists – St. Bureaucraticus, who represented rationalism and professionalism in US academic life, and St. Scholasticus, who represented social theorists who sought normative generalizations about humankind and society. Consistently, these 'Sons of Daedalus' were the target of his criticism because their well-funded procedures and programs emasculated the independence of the impressionable younger scholars (Williams, 1983).

His address, 'Folkways of the Social Sciences' was an eloquent, if not audacious, plea to social scientists to 'give back the search for truth and beauty to the individual scholar to grow in grace as best he can' and to reinstate time and place into the study of the USA (Sauer, 1952b).

Sauer's other *bête noire* – mass production – promoted the homogenization of society, not only in the USA but globally, as the USA firmly assumed the position of superpower during the 1950s, with what he saw as a disdain for the 'lesser breeds.'

Sauer was aware that he was 'out of step' with his colleagues and intellectuals, and spoke of himself as an 'unofficial outsider' and even as a 'peasant.' This has been typified as part of the antimodernism that characterized early twentieth century US intellectual life. But it went much deeper than that; his ideas anticipated society's fears and disenchantment with progress, science, the elimination of diversity, and the degradation of the environment. Perhaps the ultimate relevance of Sauer's work was in his groping toward a *rapprochement* between the social sciences, the humanities, and the biological life sciences. By being behind he was far ahead.

See also: Agricultural Sciences and Technology; Environmental Determinism; Place in Geography.

Bibliography

Entrikin, J.N., 1984. Carl O. Sauer: philosopher in spite of himself. Geographical Review 74, 387–408.

Friedrich, E., 1904. Wesen und geographische Verbreitung der 'Raubwirtschaft.'. Petermanns Mitteilungen 50, 68–79, 92–95.

Hooson, D., 1981. Carl Ortwin Sauer. In: Blouet, B.W. (Ed.), The Origins of Academic Geography in the United States. Archon, Hamden, CT, pp. 165–174.

Leighly, J. (Ed.), 1963. Land and Life: A Selection from the Writings of Carl Ortwin Sauer. University of California Press, Berkeley, CA.

Marsh, G.P., 1965. Man and nature: physical geography as modified by human action. In: Lowenthal, D. (Ed.). Harvard University Press, Cambridge, MA.

Sauer, C.O., 1925. The Morphology of Landscape, vol. 2. University of California Publications in Geography, Berkeley, CA, 19–53.

Sauer, C.O., 1938. Destructive exploitation in modern colonial expansion. Comptes Rendus du Congrès International de Géographie, Amsterdam 2 (Sect. 3c), 494–499.

Sauer, C.O., 1952a. Agricultural Origins and Dispersals. Bowman Memorial Lectures, Series 2. American Geographical Society, New York.

Sauer, C.O., 1952b. Folkways of social science. In: The Social Sciences at Mid-Century: Papers Delivered at the Dedication of Ford Hall, April 19–21, 1951. University of Minnesota Press, Minneapolis, MN.

Sauer, C.O., 1956. The agency of man on earth. In: Thomas, W.L. (Ed.), Man's Role in Changing the Face of the Earth. Chicago University Press, Chicago, IL, pp. 49–69.

Williams, M., 1983. The apple of my eye: Carl Sauer and historical geography. Journal of Historical Geography 9, 1–28.

Williams, M., 1987. Carl Sauer and man's role in changing the face of the earth. Geographical Review 77, 218–231.

Saussure, Ferdinand de (1857–1913)

Andrea Cossu, University of Trento, Italy

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.F.K. Koerner, volume 20, pp. 13492–13495, © 2001, Elsevier Ltd.

Abstract

Few scholars have left on their discipline the mark Ferdinand de Saussure left on linguistics. After his death, the publication of the Course in General Linguistics was perceived as a turning point in the discipline and a move toward a more systemic, structural analysis of language. Saussure's ideas – as conveyed in the Course – greatly influenced the rise of French Structuralism and, consequently, cultural studies in the whole realm of the humanities and the social sciences.

Saussure's Interdisciplinary Legacy

Ferdinand de Saussure is known today as one of the founding fathers of modern semiotics, as a forerunner of structuralism, and as a thinker who has influenced several generations of scholars not only in linguistics, but in all the disciplines that were touched by the 'cultural turn' that started in the 1960s and 1970s. Ironically enough, this huge intellectual status is posthumous and apocryphal. Saussure's fame derives mainly – one might say almost exclusively – from a work, the *Cours de Linguistique Générale* (Saussure, 1916), which was edited from the notes of the three courses he taught in the period 1907–11. The *Cours* was published in 1916, three years after Saussure's death, and in a few years' time it gained the reputation of being a turning point in modern linguistics. From then on, it acquired a life of its own and became a 'totemic' text.

During his lifetime, Saussure was seen under a quite different light: a precocious enfant prodige, able to write a landmark work in Indo-European linguistics (Saussure, 1879) at age 21; a man whose perfectionism prevented him from being a prolific author; a scholar who sometimes struggled to follow his great intuitions with systematic presentations of his findings; a prominent expert in historical and comparative linguistics.

There are therefore many Saussures, depending on the angle from which we consider his figure. For the majority of his readers, Saussure is today the 'author' of the *Cours*, and especially the author of the part of the *Cours* that deals with the general principles of linguistics. But time has led also to a philological reconstruction of his ideas, and to the discovery of surprisingly modern positions on semantics and pragmatics that could not always be reconciled with the legacy of his masterwork.

Education and Career

Saussure was born in Geneva on 26 November 1857 from a distinguished Protestant family. From a very young age, he developed an interest for poetry and the study of language, following the tracks of his maternal grandfather Alexandre-Joseph, himself an amateur linguist. The family was more known for its contribution to the natural sciences, starting with Saussure's great-grandfather Horace-Bénédict (1740–99) and his great-uncle Nicholas-Théodore (1767–1845). His father Henri, on his part, was an amateur entomologist and an agronomist. Following the steps of these renowned relatives, Saussure enrolled at the University of Geneva in 1875 with the prospect of becoming a scientist himself. There, he took courses in chemistry and physics, but also in philosophy, history of art, and classical languages. Perhaps the burden was too heavy, since Saussure missed his chemistry examination and, the following year, his father agreed to his transfer to Leipzig, where he would study linguistics full time.

Leipzig was at the time a very important center for linguistics, where the legacy of Romantic philology was being challenged by the group of scholars that were half-mockingly described as the *Junggrammatiker* (the Neogrammarians). In Leipzig, Saussure was able to form a network – not always characterized by easy relationships – with some of the most prominent scholars of his time. Saussure studied with Georg Curtius (1820–85), the doyen of linguistics at Leipzig, and with younger scholars including Hermann Osthoff (1847–1909), August Leskien (1840–1916), and Karl Bruggman (1849–1919) and others, well-versed in many of the languages of the Indo-European family.

By that time, Saussure had also established fruitful contacts with the other important center of linguistics in Europe, Paris. It was a move that would have important consequences on his career. In 1876, Saussure applied for membership in the Societé Linguistique de Paris, and accompanied the application with a paper on the suffix *-t-* that was quickly accepted for publication (Saussure, 1877). For years to come, Saussure played a central role in the activities of the Societé, on whose board he served for many years as Adjunct secretary. The *Mémoires de la Société de Linguistique de Paris* became the journal where the majority of the articles he published during his lifetime appeared. Thanks to his affiliation to the Société, Saussure became acquainted with the work of French linguists, like Michel Bréal (1832–1915), Gaston Paris (1839–1903), Louis Havet (1849–1925), Arséne Darmesteter (1848–88), as well as Antoine Meillet (1866–1936), his most gifted student and later a prominent linguist and a member of Émile Durkheim's interdisciplinary circle.

Despite the early acceptance of his work by Parisian linguists, Saussure was still a student in Leipzig. After the series of short papers that appeared on the *Mémoire* in 1877, Saussure began work on a much more ambitious project, one that in his

vision would advance comparative linguistics to a higher level, in terms of methodological approach and conceptual rigor. For the better part of 1878, Saussure rushed to finish a manuscript that became a breakthrough work in Indo-European linguistics. The *Mémoire sur le systéme primitif des voyelles dans les langues Indo-Européennes* (Saussure, 1879) was a 300-page book and the work for which Saussure gained acclaim and recognition during his lifetime.

In this book, which he self-published, Saussure hypothesized the existence of a sound in Proto-Indo-European (which he identified as A, although he was careful to claim that it had an unknown phonetic value). Saussure's A, in combination with other 'sonant coefficients,' would develop into various phonemes. With this ingenious internal reconstruction and with a move toward abstraction, Saussure was able to build a rigorous hypothesis on the dynamics of *ablaut* (i.e., vowel variations) in the Indo-European linguistic system, an hypothesis that was going to be proved only in 1927 by Polish linguist Jerzy Kuryowicz. More than the technicalities of Saussure's ingenious solution, or the advancement in the knowledge of Indo-European phonology, the *Mémoire* is most notable today for it contains the kernel of some assumptions that Saussure would develop in his unpublished work on linguistics and that would eventually lead to the lectures of the *Cours*. Saussure was among the first to use a modern concept of 'phoneme,' and his view of language as a system that works through the combination of interrelated parts that acquire meaning only as part of that system finds is advanced for the first time in the *Mémoire*.

Soon after, Saussure defended – *summa cum laude* – his doctoral thesis on the genitive absolute in Sanskrit, an ambitious work on the origins of the form, which he was not able to complete to his full satisfaction and which was published as *De l'emploi* (On the Use) *du génitif absolu en sanscrit* (Saussure, 1881). He then spent a period at his family home in Geneva and made an unproductive field trip to Lithuania (the interest for Lithuanian returned in later years on his work on accentuation). In September 1880, he moved to Paris, where he spent the next decade in the nontenured position of maitre des conferences at the Ecole des Hautes Etudes.

While in Paris, Saussure carried on his duties as a preeminent member of the Societé Linguistique. He also began to write in a more systematic matter on topics related to general linguistics, although it was with frustration that he realized that he was not able to bring any of these projects to its completion. Indeed, many of the themes that appear in the *Cours* were already being sketched by the beginning of the 1880s in unfinished manuscripts like his notebooks on phonology of 1881 (where he discussed the linearity of language and used the term 'semiotic'), the relationship between sound and idea, and the necessity to complement historical analysis with the study of language as a 'state' when dealing with 'semiological facts' (Joseph, 2010; Joseph, 2012: 306; Saussure, 2006).

Saussure was unable to secure a chair in Paris, where politics played a major role and his Swiss nationality acted as an obstacle in the context of the Third Republic. That was probably the reason why he moved back to Geneva in 1891, where he was appointed as extraordinary professor of Sanskrit and Comparative Grammar, a position that was made permanent with the institution of a chair of Comparative Philology in 1896.

The first years in Geneva were, as it had happened in Paris, years of methodological and theoretical reflection that form the bulk of the unpublished manuscript concerning the 'double essence' of language (Saussure, 2006: 3–60) in which he pointed out the difference between vocal sounds and meaningful signs. Only when ideas are attached to vocal sounds can language exists. Nor this was the only idea that anticipated the contents of the *Cours*. The notes contained in "On the Double Essence" also deal with issues of synchrony and diachrony, and with the proposal of Saussure's theory of *value*, which is determined by contrast, opposition, and difference with other signs.

It would take more than a decade before Saussure started the cycle of lectures that made him posthumously famous and in which these ideas were made public. Following the retirement of one of his colleagues, he was assigned the task of teaching a course in General Linguistics that took place every other year. Saussure taught three courses: in 1907, 1908–09, and 1910–11. The students' notes from these courses, and some of Saussure's own notes, constitute the bulk of what became the *Cours*. It was the second landmark work in his intellectual trajectory, although he did not live enough to see this success. Saussure died on 22 February 1913.

The Course in General Linguistics

The decision to publish a book out of Saussure's notes was made soon after his death. It was a daunting task because he had not planned to write a book on general linguistics and he had left very few systematic notes. What became the *Cours* was thus put together using some of Saussure's manuscripts, and mostly students' notes from the three courses.

The *Cours* was published in 1916, after three years of gestation and editing by two of Saussure's colleagues, Charles Bally and Albert Sechehaye, with the contribution of Albert Riedlinger (a student who had attended the first course in 1917). Bally and Sechehaye, who were among the first bearers of Saussure's flag, have sometimes been accused of assembling a text that reflected more their own visions than Saussure's focus on historical and comparative linguistics.

Today, the picture is much more accurate. Not only Saussure's writings on general linguistics have been found and published, but the material on which the *Cours* as published was drawn has been made available, including critical editions of the *Cours*, annotated editions, and the students' notebooks for all the three courses.

While these sources offer us a clearer idea of how the three courses were organized, and follow the development of Saussure's arguments, for the broader audience it was the *Cours* as originally published that had considerable influence, especially the long introduction and the section on the general principles for the study of language. In these sections, that represent the starting point for any foray in the work of Saussure, a powerful framework is proposed, with concepts that have shaped not only contemporary linguistics but many other disciplines: the distinction between *langue* and *parole*, and between synchronic and diachronic; the analysis of the linguistic sign and its

arbitrariness; the theory of linguistic value; and the analysis of the types of relations among signs, at the syntagmatic and the associative level. These are the main tenets of the *Cours*, and have represented for many a milestone in the development of structuralism and semiology, the general science that Saussure promised and of which his linguistics was a particular application.

The Main Tenets of the Course

The *Cours de linguistique générale* is an exploration that seeks to find the object of linguistics, its relation to other facts connected to language (mainly, its concrete use by speakers), the nature of linguistic units, and how they contribute to the stability or change of the linguistic system. The *Cours* is not – in Saussure's vision and in the one of his editors' – a speculative work on language. It is, on the contrary, the ambitious attempt to construct linguistics as a scientific and rigorous enterprise by defining its object and its position *vis-à-vis* other social sciences.

Saussure's argument for the definition of the object of linguistics proceeds by subtraction and abstraction. His key idea is that language needs to be studied as a system (he rarely used the word 'structure,' and never in the technical sense that was popularized decades later), which is composed by interrelated linguistic units whose function can be described only with reference to other elements of the system. Only in this way the systematic character of language can emerge, against the background of the heterogeneous mass of linguistic phenomena.

The strategy of isolating the system in analytic terms leads Saussure to the formulation of the first major dichotomy that is present in the *Cours*: that between the use of language in acts of speaking (*la parole*) and language as a socially shared repository of signs and rules (*la langue*). This distinction separates what is individual in the execution of language from what is social, the latter being the proper object of linguistic analysis. While the *parole* requires an individual *act* in a communicative situation, the role of the individual with regards to the *langue* is less as an agent and more as a repository of representations that are fundamentally transindividual.

This distinction is clear in Saussure's description of the speaking circuit, which involves a 'psychological phenomenon' (the association of a sound-image to a concept), a 'physiological process' (the emission of sound), and a 'physical process' (the transfer of sound waves through the air to the ear of the listener). The analysis of the *langue* is interested only in the first of these aspects, while the others belong to the *parole*. Linguistics – and semiology as a broader, more encompassing science of signs as part of social life – is therefore a part of social psychology, whose object is the nature of the association – a largely mental phenomenon, however socially shaped – between sound-images (signifiers) and concepts or ideas (signifieds).

This is Saussure's second major dichotomy. The vision that a sign entails the relation between an expression and a conceptual counterpart is not a radical invention on Saussure's part, and has many antecedents in the history of philosophy of language. Quite radical, with the hindsight of a hundred years of reception of the *Cours*, is the combination of antireferentialism and arbitrariness that Saussure sees as key elements of linguistic signs.

Saussure's antireferentialism is one important consequence of his distinction between langue and parole. Language (as *la langue*) is not a naming process that relates words to things in the world. Rather, the linguistic sign is the association of a 'sound-image' and a 'concept,' not the one between a word and a thing. Like words and things, however, signifiers and signifieds are in a special relation with peculiar characteristics. They are both required to signify and they cannot be separated from the concrete point of view, as 'each recalls the other' like the two sides of a sheet of paper that cannot – fronte and verso – be cut. The linguistic sign is therefore an intimate unity between two sides that could not be conceivable without each other.

The relation between the signifier and the signified is a particular one, characterized – according to Saussure – by two fundamental principles: the arbitrariness of the sign and the linearity of the signifier. Arbitrariness is one of the fundamental concepts of Saussure's *Cours*. Given the internal solution that he favors in the definition of the object of linguistics, the arbitrary nature of the sign does not mean the arbitrariness of its relationship with an external object. It is an *internal* arbitrariness that takes place between the signifier and the signified. By this, Saussure means that there is no motivated relation within the sign, that is., there is nothing natural in the association of a given signifier to a given signified; it is, by all measures, the result of social activity that gets crystallized in a particular form, the sign.

The concept of linearity is often overlooked, but it is important in order to understand the types of relations that exist *among* signs, and not only *within* signs. Being of auditory nature, a signifier can be conceived only as it unfolds in time: its representation is a line. This formal constraint governs the way signifiers are organized: at least in language, two different signifiers cannot occupy the same position at the same time. They can appear either as distinct elements in a succession, or one can substitute the other, creating thus a different value or meaning.

To understand this fundamental proposition, one need not to consider signs in isolation, but the types of association that exist among signs. Saussure – in his usual dichotomic fashion – argues that relations among signs can be either syntagmatic or associative (later, Jakobson would use the term 'paradigmatic'). These relations work in different ways: while syntagmatic relations are characterized by the succession of syntagms in a linguistic chain (and thus are characterized by presence or copresence of signs in a syntagmatic chain), associative relations are part of a 'mnemonic series' that brings together signs that have something in common (they work in absentia).

The three dichotomies that Saussure had proposed in the *Cours* are of fundamental importance to describe the internal mechanisms of the langue and the types of relations that exist within the langue. The historical dimension does not play an important role, even though there is an internal dimension of temporality that is crucial to understand how the langue works. The last dichotomy addressed the issue of linguistic change and historical comparison. A langue, in principle, can and must be approached in a static way, by producing a description of the configuration of the system at a particular moment where

structural changes do not occur. This is what Saussure means by synchronic analysis. A diachronic analysis, on the contrary, considers at least two different states of language that are not only temporally distinct, but in which the use of language has produced socially accepted changes that have invariably altered the system.

Langue/parole, signifier/signified, synchrony/diachrony, and syntagmatic/associative are the four major oppositions that Saussure proposes in his strategy to build linguistics on scientific grounds. In the reception of Saussure's work, the synchronic approach to the langue has taken a central position outside the boundaries of linguistics. While this runs contrary to many of Saussure's interests and to the core of intellectual orientation (he was, and always had been, mainly a comparativist) it was through the adaptation, rereading and to some extent even misunderstanding of the *Cours* that he was able to secure a central position as a revolutionary thinker for generations of scholars.

Saussure after the *Cours*: Reception and Adaptation

Saussure's legacy as one of the most prominent linguists of his generation was secured, ironically, by words he had not written and by positions that were made probably more radical than he had originally thought. Like many classic works (Max Weber's *Economy and Society*, Antonio Gramsci's *Prison Notebooks*, or Ludwig Wittgenstein's *Philosophical Investigations*), it mattered less what he had actually said, and more the use one could make of his words. The decades that passed after the publication of the *Cours* have thus been characterized by the attempt to reconcile different – sometimes radically different – visions of Saussure: the comparativist, the founding father of structuralism, the institutionalist, the scholar who promised – but did not realize – the foundation of a new science like semiology (Harris, 2003; Sanders, 2004).

To all the parties involved in this retrospective reconstruction, the *Cours* is still the door to Saussure's vision of linguistics. One could work, however, with the *Cours* or despite the shadow it casts on his 'author.' This tension often meant the effort to go beyond the published work and reconstruct a more authentic Saussure with the aid of his original writings.

The reception of the *Cours* came in waves. Even in its 1916 incarnation and other editions (1922, 1931) the *Cours* took some time before it was perceived as a foundational work in linguistics, and again it happened thanks to the mediation of scholars who stretched some of the main positions it contained in order to fit their own goals. Early reviews were mixed, and the reception of the *Cours* took place in a milieu still characterized by different research paradigms (historical linguistics in France and descriptivism in the United States) that were at odds with Saussure's emphasis on synchronic analysis, although the *Cours* found some champions in both countries, like Meillet and Leonard Bloomfield (1887–1949). On their part, Soviet linguistics criticized the *Cours* for its lack of attention to extralinguistic forces, despite Saussure's own caveats and his insistence on the existence of an 'external linguistics' (a sociological critique advanced by philosopher Valentin Voloshinov, 1929).

The trajectory of the *Cours*' appreciation turned for the better at the end of the 1920s, when linguistic structuralism was about to become a recognized tendency. Louis Hjelmslev (1899–1965) in Copenhagen and Nikolai Trubetzkoy (1890–1938) and Roman Jakobson (1896–1982) in Prague acted as the main agents in this new appreciation of the *Cours*, one that – despite criticism and substantial changes – stayed true to the book's perceived program of static, synchronic, and formal linguistics. However, they also called for a renewed attention to problems that Saussure had dismissed or neglected, like the relation between form and substance, the type of function established between a signifier and a signified, the importance of the phonological level, and issues of semantics and pragmatics that are not present in the *Cours* (although some of these problems are hinted at in the *Writings on General Linguistics*).

Hjelmslev's and Jakobson's work (Hjelmslev, 1928, 1953; Jakobson, 1995) was at the same time more formal than Saussure's and more explicit in the definition of the conceptual boundaries of semiotics as a discipline. Saussure had defined linguistics as a discipline that could analyze the role of signs in the context of social life, but had devoted just a few suggestive pages to the role of semiology and its position *vis-à-vis* the other psychological and sociological disciplines. To him, general linguistics was part of semiology insofar as it could describe the systemic character of *la langue* and its units (signs). But, when it came to semiology, he characterized it as a science that did not yet exist.

Hjelmslev started mainly from these premises, but criticized three crucial aspects of Saussure's approach to the l*angue*: his cognitivism (the langue is embedded in the human mind), his attention to the orality of language, and the lack of analysis of the semantic level. From Saussure, Hjelmslev derived a hierarchical vision and an inclination for dichotomic classification, in some cases specifying Saussure's ideas, and in others creating a new conceptual toolbox that complemented Saussure's: form and substance, expression and content in lieu of signifier and signified, system and process as internal patterns of relations.

Jakobson's work proved even more influential in the popularization of Saussure, even though he always remained very critical with regards to the rigor of the categories that Saussure had adopted. Starting from work on phonology in the 1920s, Jakobson became the true herald of structuralism in linguistics. However, he also developed a line of inquiry that was broader than Saussure, even when he seemed to change little from Saussure's ideas: Jakobson's notions of code and message are note a mere translation of Saussure's *langue* and *parole*, but the attempt to integrate linguistics with information theory; his idea of syntagmatic and paradigmatic relations is more formal than Saussure's, without the focus on the open-endedness of chains of signs that assigned to associative relations; his model of communication describes functions that apply to any kind of manifestation of language, including poetic texts.

While Saussure was commonly perceived as the originator of structuralism, it was however through the work of these scholars that structuralism became a true intellectual force starting from the 1940s (Dosse, 1998a,b). Saussure's ideas were incorporated in this attempt to shape structuralism (and, within structuralism, semiotics) as a theory and a method. From this point of view, this meant a reversal of the *Cours*' insistence on linguistics as a province of semiology. Rather,

linguistics became the paradigm for the development of semiotics, insofar as all semiotic phenomena were organized as a language or could be analyzed as a language (Barthes, 1968).

The rise, and the almost sudden wane, of structuralism meant another turn in Saussure's reception. The *Cours*' editors had imagined and portrayed Saussure as a generalist interested in synchronic analysis. Barthes and those associated with structuralism (among the most prominent ones, one could mention Maurice Merleau-Ponty, Jacques Lacan, Roman Jakobson, Claude Levi-Strauss, and Algirdas Greimas) offered another mediation, that focused on linguistics as a discipline that could offer a paradigm, as well as a vocabulary modeled after Saussure but highly different from Saussure's intent and scope: a toolbox made of binary oppositions, codes and messages, deep structures, generative models, that could be stretched in the effort to be connected to Saussure's ideas, but which represented also a creative (and fruitful) betrayal of Saussure as much as they were a betrayal of the *Cours*.

Meanwhile, Saussure's figure and work were being subject to a more careful scrutiny in order to come to a genealogical reconstruction of his ideas (Godel, 1957; Koerner, 1973). Saussurean studies were less interested in extending the application of Saussure's ideas than they were in trying to figure out how the *Cours* represented the synthesis of reflections that Saussure had made during his entire life.

This genealogical work took several directions, and the rediscovery of Saussure as a general linguist was not without surprises. It was possible to trace Saussure's interest for theoretical issues back to his Paris years. The appearance – and then publication – of Saussure's *Writings in General Linguistics* (2006) allowed for more careful consideration of the path that had led him to formulate his vision of the linguistic sign, the role of semiology and the tasks of the discipline. While new scholarship on Saussure did not radically change the picture, some of the assumptions that scholars had made about the course had to be tuned according to the availability of handwritten notes that included – crucially – his own lecture notes for the *Cours*, which complemented the students' notebooks critically edited by Eisuke Komatsu (Saussure, 1993, 1996, 1997).

Thus, we can argue that today there are many Saussures, depending on the prevalence of either a philological approach or a creative reconstruction of his image. Both camps, however, pay homage to a highly original thinker, regardless of the many mediations that have made his work available. While linguistics, semiotics, and cultural studies have fruitfully moved 'beyond Saussure,' he is still a relevant turning point, the originator of lines of inquiry, even when theoretical reflection goes against his positions. The very fact that we think of all these approaches as developments that took place 'after Saussure' and thanks to this work, is perhaps the best testimony of his intellectual legacy and of the conceptual power of his proposals.

See also: Comparative Method in Linguistics; Jakobson, Roman (1896–1982); Linguistic Typology; Linguistics: Overview; Semiotics; Structuralism.

Bibliography

Barthes, Roland, 1968[1964]. Elements of Semiology. Wang and Hill, New York.

Dosse, Francois, 1998a. History of Structuralism Volume One: The Rising Sign 1945–1966. University of Minnesota Press, Minneapolis.

Dosse, Francois, 1998b. History of Structuralism Volume Two: The Sign Sets 1967–Present. University of Minnesota Press, Minneapolis.

Gehr, Johannes, 2000. Saussure entre linguistique et sémiologie. Presses Universitaires de France, Paris.

Godel, Robert, 1957. Les sources manuscrites du Cours de Linguistique générale de F. de Saussure. Droz, Genève.

Harris, Roy, 2003. Saussure and His Interpreters. Edinburgh University Press, Edinburgh.

Hjelmslev, Louis, 1928. Principes de grammaire générale. Bianco Lundo, Copenhagen.

Hjelmslev, Louis, 1953[1943]. Prolegomena to a Theory of Language. Indiana University Publications in Anthropology and Linguistics, Bloomington.

Jakobson, Roman, 1995. On Language. Harvard University Press, Cambridge.

Joseph, John E., 2010. Saussure's notes of 1881–1885 on inner speech, linguistic signs and language change. Historiographia Linguistica 37, 105–132.

Joseph, John E., 2012. Saussure. Oxford University Press, Oxford.

Koerner, Ernst F.K., 1973. Ferdinand de Saussure: Origin and Development of His Linguistic Thought in Western Studies of Language: A Contribution to the History and Theory of Linguistics. Vieweg, Braunschweig.

Sanders, Carol (Ed.), 2004. The Cambridge Companion to Saussure. Cambridge University Press, Cambridge.

Saussure, Ferdinand de, 1879. Mémoire sur le système primitif des voyelles dans les langues Indo-Européennes. B.G. Teubner, Leipzig (Reprinted Cambridge University Press, Cambridge, 2009).

Saussure, Ferdinand de, 1877. Le suffixe -t-. Memoires de la societé linguistique parisienne 3, 197–209.

Saussure, Ferdinand de, 1881. De l'emploi du génitif absolu en sanscrit. Imprimerie Jules-Guillaume Fick, Genève.

Saussure, Ferdinand de, 1916. Cours de linguistique générale. In: Charles Bally, Albert Sechehaye (Eds.), with the collaboration of Albert Riedlinger. Payot, Lausanne and Paris. (Second ed., Payot, Paris, 1922; third and last corrected ed., 1931; fourth ed., 1949; fifth ed., 1960, etc.–English Trans.: (1) By Wade Baskin, Course in General Linguistics, Philosophical Library, London and New York, 1959 (reprinted, McGraw-Hill, New York, 1966; rev. ed., Collins, Fontana, London, 1974), and (2) by Roy Harris, Course in General Linguistics, Duckworth, London, 1983).

Saussure, Ferdinand de, 1922. Recueil des publications scientifiques de Ferdinand de Saussure. Sonor, Genève.

Saussure, Ferdinand de, 1993. Troisiéme cours de linguistique générale (1910–1911), d'après les cahiers d'Émile Constantin/Saussure's Third Course of Lectures on General Linguistics (1910–1911), from the Notebooks of Émile Constantin. Pergamon, Oxford and New York.

Saussure, Ferdinand de, 1996. Premier cours de linguistique générale (1907), d'après les cahiers d'Albert Riedlinger/Saussure's First Course of Lectures on General Linguistics (1907), from the Notebooks of Albert Riedlinger. Pergamon, Oxford and New York.

Saussure, Ferdinand de, 1997. Deuxième cours del linguistique générale (1908–1909), d'après les cahiers d'Albert Riedlinger et Charles Patois/Saussure's Second Course on General Linguistics (1908–1909), from the Notebooks of Albert Riedlinger and Charles Patois. Pergamon, Oxford and New York.

Saussure, Ferdinand de, 2006. Writings in General Linguistics. Oxford University Press, Oxford.

Voloshinov, Valentin, 1986[1929]. Marxism and the Philosophy of Language. Harvard University Press, Cambridge.

Savage, Leonard J (1917–71)

Dennis V Lindley[†], Minehead, Somerset, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13495–13499, © 2001, Elsevier Ltd.

Abstract

L. J. Savage was born in Detroit in 1917 and obtained a mathematics Ph.D at the University of Michigan. After a series of posts, he went, as a statistician, to the University of Chicago in 1946, staying until 1960. He then returned to Michigan, before going in 1964 to Yale, where he died in 1971. He is today principally known for his 1954 book *The Foundations of Statistics*, the first seven chapters of which develop an axiomatic approach to decision-making under uncertainty. The main result is that the uncertainties should be described by probabilities, the worth of the outcomes by utilities, and that the best decision is that obtained by maximizing expected utility. This approach is today usually described as Bayesian. Although Savage developed it as a justification for accepted statistical practice, it effectively destroyed that practice, instead producing a system based on probability as a degree of belief. The remainder of his life was devoted to exploring the consequences of his approach. In this he worked with the great Italian probabilist, de Finetti and, with Hewitt, proved an important extension of a major result of the Italian. He also wrote a book on gambling with Dubins. For social scientists, the main importance of his work lies in the construction of a belief system about the world as appreciated by a single individual.

L. J. Savage, always addressed by friends as Jimmie, was born in Detroit, Michigan on 20 November 1917. His father was in the real-estate business, his mother a nurse. Throughout his life he suffered from poor eyesight, a combination of nystagmus and extreme myopia, which was to affect the development of his career in many ways. Most of us accept reading as an easy, even casual, activity, whereas to him it was a serious business with the text held close to the eye. The result was that he appeared to regard the material as important and would absorb it with an intensity that often escapes those with normal vision. To have him read something you had written was, at first, an uncomfortable business, for he would politely question much of the material, but later to realize that he was being constructive and the final result benefited enormously from his study of it. He was a superb lecturer, once remarking that he could hold himself spell-bound for an hour. When, towards the end of his life, he gave the Fisher lecture, he held the audience spell-bound for considerably more than the allotted hour.

The poor eyesight at first affected him adversely, his high-school teacher not recommending him for further education. His parents were insistent and he went initially to Wayne University and then to the University of Michigan. His attempts to do biology failed because he could not draw, and chemistry was ruled out when he dropped the beaker in the laboratory. Eventually he met a fine mathematics teacher who introduced him to a field in which the reading had to be intense, yet minimal, where his brilliance was recognized and he gained a Ph.D in the application of vectorial methods to metric geometry. After a year each at Princeton, Cornell, and Brown, 1944 found him a member of the Statistical Research Group at Columbia University, a group he described as 'one of the greatest hotbeds statistics has ever had,' a judgement that is surely sound, as almost all those who were to become the leading statisticians in the US of the 1950s were in the group or associated with it.

So he became a statistician, and although his greatest work was in the theory, as such he was able to indulge his interest in science, so that throughout his career he enjoyed, and was extremely good at, consulting with his scientific colleagues. It was not just science that interested him, for he had inherited from his father a respect for business, and had acquired from Milton Friedman, a member of the group, an appreciation of economics, that made his advice useful and generously available to all serious thinkers.

In 1946, Savage went to the University of Chicago where, from 1949, he was in the Statistics department founded by Allen Wallis, becoming chairman from 1957 until he left the university in 1960. His departure was a turning point in his life. A reason for leaving was personal difficulties in his marriage to Jane Kretschmer that they hoped could be mended by a move to Michigan with their two sons, Sam and Frank. It did not work out and they were divorced in 1964. That year he married Jean Pearce and moved to Yale. It was in New Haven on 1 November 1971 that he tragically died at the early age of 53, and statistics lost a great scholar and many of us a dear friend. The move from Chicago was partly influenced by his perception of his relations with colleagues there. He felt, as we will see below, that he had produced a sound axiomatization of statistics showing that many standard statistical procedures were unsatisfactory, so that his colleagues should either explain where the axioms were wrong, or else abandon the faulty procedures. They did neither. They were not alone in this; most statisticians acted as the Chicago faculty, refusing to contest the axioms, yet refusing to use the results (many still do so today) and Savage found it difficult to get a suitable post until Yale obliged. The monograph by Savage with others (1962), although now rather dated, indicates the sort of antagonism that existed, though, in that book, all the protagonists are on their best behavior. Savage (1981) contains most of his more important papers, tributes from others and a technical summary of his work. Many of his papers, contained therein, deal, not with technicalities, but with his growing appreciation of statistical ideas as they developed throughout his life.

Savage is justifiably famous for one tremendous contribution to statistics and the scientific method, contained in his 1954 book *The Foundations of Statistics*. Or, to be more correct,

[†] Deceased.

the first seven chapters of that book. Anyone contemplating study of this glorious achievement should first read his preface to the second edition in 1972, for this is one of the most honest pieces of scientific writing that I know. To appreciate the importance of the book, it is necessary to cast one's mind back to the state of statistics around 1950. Fisher made enormous advances from 1920 onwards. Neyman and Pearson had introduced new ideas and Wald was developing decision analysis. But the statistical procedures available, though undoubtedly valuable and much appreciated by scientists, especially biologists, were a disparate collection of ideas lacking coherence. Savage had the idea that statistics should be like other branches of mathematics and be based on a set of concepts, termed axioms, from which the procedures would follow as theorems. He wanted to provide a firm basis for Fisher's results. Another way of understanding is to recognize that Fisher had suggested procedures whose properties had been studied by him and others, whereas Savage turned the method around and asked what properties we wanted and discovered what procedures provided them. His expectation was that they would prove to be just what statisticians were using. To his great surprise, they were not. Interestingly, his last paper, Savage et al. (1976), is a brilliant, sympathetic review of Fisher's work.

Savage had worked with von Neumann in Princeton and had much appreciated his work on games with Morgenstern (1944), that introduced axioms leading to the concept of a utility function for the outcomes of a game. However, their treatment had used probability without justification, so Savage conceived the idea of developing an axiomatic system for decision-making under uncertainty which would not use probability in the axioms but derive it in theorems, so uniting the ideas of probability and utility. Also influential was the development by Wald (1950) of an attempt to unite statistical ideas under the concept of decision-making, using von Neumann's idea of minimax, but lacking any axiomatic foundation. Savage's 1954 book presents axioms for a single decision-maker, now often referred to as 'you,' faced with selecting a course of action in the face of uncertainty. He then proves three theorems: first, that your various uncertainties present must combine according to the rules of probability; second, that your perceptions of the merits of the possible outcomes that result from your actions must be described by a real-valued utility function (which is itself based on probability); third, that your optimum decision is that which maximizes your expected utility (MEU), the expectation being evaluated according to the probabilities developed in the first part of this trilogy. One of the axioms he used encapsulated a simple notion that has since been seen to be of some importance. If you prefer A to B when C is true and, at the same time, prefer A to B when C is false; then you prefer A to B when you do not know whether or not C is true. It is called the 'sure-thing' principle because the preference of A over B is sure, in American parlance, at least as far as C is concerned.

The introduction of MEU was not new, it having been used by Daniel Bernoulli in the eighteenth century, but the development extended its use considerably, in particular by showing that it was the only sensible method for a single decision-maker to use. Utility had recently been explained, as we have seen, by von Neumann. The original and dramatically important result was the first, saying that probability was the only sensible description for uncertainty, so that the rules of probability were not arbitrary but dictated by the sensible and modest requirements expressed in the axioms. Others, from Laplace onwards, had used probability without justification: Savage showed that it was inevitable. So it was the extensive and unique use of probability that was Savage's main contribution in the book, and there are two ways in which his analysis was influential.

First was the fact that the probability was personal, expressing the uncertainty of the decision-maker, 'you.' It was not *the* probability, but rather *your* probability. An immediate reaction is that this contradicts the very nature of the scientific method which is supposed to yield objective results. This apparent contradiction was resolved in a later paper with Edwards and Lindeman (Savage et al. 1963), which showed in the principle of stable estimation, that under general conditions, people with different uncertainties would reach agreement on receipt of enough data. This result agrees with the observation that scientists typically disagree with one another in the early stages of an investigation but are eventually brought together as the evidence accumulates. Many now think that the reduction of uncertainty, through data, incorporated in the laws of probability, expresses the nature of induction and is part of the scientific method.

A second, influential factor was the result that probability was the *only* sensible description of uncertainty, for scientists, encouraged by statisticians, had been using other methods, for example in the use of tail-area significance tests of a null hypothesis, H, where the relevant quantity is the probability in the tail of a distribution, given that H is true. Savage's argument was that the proper evaluation is the probability of H, given the data, not of an aspect of the data, given H. Similar contradictions arise with confidence intervals for a parameter which involve probability statements about the interval, not about the parameter. These are instances where his conclusions clashed with standard practice, leading to disagreements like that at Chicago mentioned above.

There was a third way in which Savage's ideas contradicted the current practice, though this was not fully realized until Birnbaum's (1962) paper. The axiomatic development showed that if you had a statistical model with probability $p(x/\theta)$ for data x given parameter θ, then the only way the data could affect your uncertainty of θ, given x, was through the likelihood function $p(x/\theta)$ expressing how that probability varied with θ for the fixed, observed x. This is the likelihood principle. Significance tests clearly violate the principle since they use a tail-area which involves integration of $p(x/\theta)$ over x-values, that is, over data that you do not have. Indeed, many statistical procedures, even today, violate the principle. A related example of conflict with current practice is found in optional stopping. It had long been known that, under commonly occurring circumstances, it is possible to continue sampling until a null hypothesis is rejected at any preassigned significance level, so that you can opt to stop at your convenience. Savage showed that it was the significance test that was at fault, not the stopping procedure.

It is an astonishing fact, that he explains in the preface to the second edition, that he and others who worked with him, including myself, in the mid-1950s, did not appreciate these ideas and did not comprehend that, rather than his work

justifying statistical practice, it altered it. The latter part of his book in which the justification is attempted, largely using the tool of minimax that von Neumann had employed with great effect in the theory of games, is today of little interest, in contrast to the brilliance of the early chapters. Savage was a true scholar, a man who studied the work of others and accepted any valid criticism of his work, to incorporate their ideas with due acknowledgment. And so he understood the ideas of Ramsey (1926) who had, in a less rigorous presentation, developed similar ideas, ideas which no one had understood until Savage. But more importantly, he appreciated the work of the Italian, de Finetti, who, from an entirely different standpoint, and one which is perhaps more forceful than Savage's, developed the concept and uniqueness of personal probability. (For English readers, the best references are de Finetti (1972, 1974, 1975), though his ideas originate in the 1930s.) Savage acquired skill in Italian and worked with him. One of the problems de Finetti had studied was your assessment of probability. You are uncertain whether it will rain tomorrow in your city; so according to their thesis, you have a probability for rain. How are you to assess its value? De Finetti suggested a scoring rule: if you choose p as your value, you will receive a penalty score $(I - p)^2$ if it rains and p^2 if not. Scores for different events will be added. To minimize your total score the values that you give must be probabilities. The basic paper was written by Savage (1972) with due acknowledgment to the originator. With Hewitt (1955), Savage also extended significantly another result of de Finetti on exchangeability.

Savage conducted a major study with Dubins that resulted in their book (Savage and Dubins, 1965). The problem here is that you are in an unfavorable gaming situation at a casino and your object is to maximize your chance of winning a fixed sum of money; how should you play? They showed that boldness pays; bet all you have or the maximum amount the casino will allow. This led to a thorough study of betting strategies and many interesting and useful results were obtained.

For reasons that are not entirely clear, the school of statistics that rests on Savage's results is called Bayesian, after the eighteenth century cleric, Thomas Bayes partly because much use is made of Bayes's theorem. Today it is a flourishing school of statistics with a large and increasing number of adherents, that co-exists alongside the earlier school, often called frequentist because probability therein refers to frequency, whereas to a Bayesian it expresses your uncertainty of an event. Bayesians believe that their approach to the handling of data, and the making of decisions that use the data, is important for social and behavioral sciences for the following reasons.

If these subjects are to be truly scientific, their arguments must depend heavily on data which, properly interpreted, can be used to defend theories in the interplay of theory and practice that is the scientific method. In the physical sciences, the data ordinarily take the form of results of planned experiments in the field or laboratory, experiments which can, and should be, repeated by others, in accord with stable estimation, their repetition connecting naturally with frequency. In the social sciences, such planning and repetition is rarely possible, so that reliance has to be placed on observational data, and the behavioral sciences are scarcely better off. It is now widely appreciated that observational data have to be analyzed with considerable care, so that any conclusions are soundly based and not confounded with other factors that the lack of planning has been unable to control. Statistics therefore becomes a more important ingredient in the social, than the physical, sciences. Furthermore, social decisions must often be made in the face of uncertainty, whereas more time and effort by an experimental scientist can substantially reduce, if not eliminate, the uncertainty. The Bayesian approach recognizes and incorporates all of these factors.

There is another sense in which the Bayesian approach may be valuable and that lies in its ability to express uncertainty about any aspect of a study, not confining probability to the repetitious, frequency aspects. Often the conclusions in the social and behavioral sciences are tentative and the personalistic perception of probability, as a measure of your belief in a hypothesis, can provide an expression of uncertainty that can be conveyed to others. We still have a long way to go in employing these ideas, when even a weather-forecaster is reluctant to say there is a probability of 0.8 that it will rain tomorrow, or, if he does, sometimes does not appreciate the meaning of what is being said.

Savage had been described as 'the Euclid of statistics' in that he did for statistics what Euclid did for geometry, laying down a set of axioms from which, as theorems, the only sensible statistical procedures emerge. As with Euclid, other geometries will emerge, so other statistical systems may arise, but, for the moment, Savage's Bayesian approach is the only systematic one available and appears to work. He was, in the sense of Kuhn, a true revolutionary, who overturned one paradigm, replacing it by another without, at first, realizing what he had done.

See also: Bayesian Statistics; Bayesian Theory, History, Applications, and Contemporary Directions; Fisher, Ronald A (1890–1962); Hotelling, Harold (1895–1973); Probability Theory: Interpretations; Statistics, History of; Tversky, Amos (1937–96).

Bibliography

Birnbaum, A., 1962. On the foundations of statistical inference. Journal of the American Statistical Association 57, 269–306.
de Finetti, B., 1972. Probability, Induction and Statistics: The Art of Guessing. Wiley, London.
de Finetti, B., 1974/5. Theory of Probability, vol. 2. Wiley, London.
Dubins, L.E., Savage, L.J., 1965. How to Gamble if You Must: Inequalities for Stochastic Processes. McGraw-Hill, New York.
Ramsey, F.P., 1926. Truth and probability. In: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Logical Essays. Routledge and Kegan Paul, London.
Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.
Savage, L.J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783–801.
Savage, L.J., 1972. The Foundations of Statistics, second ed. Dover, New York.
Savage, L.J., et al., 1976. On re-reading R A Fisher. Annals of Statistics 4, 441–500.
Savage, L.J., 1981. The Writings of Leonard Jimmie Savage—A Memorial Selection. American Statistical Association and Institute of Mathematics, Washington.
Savage, L.J., Edwards, W., Lindman, H., 1963. Bayesian statistical inference for psychological research. Psychological Reviews 70, 193–242.
Savage, L.J., Hewitt, E., 1955. Symmetric measures on Cartesian products. Transactions of the American Mathematical Society 80, 470–501.
Savage, L.J., et al., 1962. The Foundations of Statistical Inference: A Discussion. Methuen, London.
von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.
Wald, A., 1950. Statistical Decision Functions. Wiley, New York.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 21

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 21

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 21 *xlv*

VOLUME 21

Scale in Geography 1
Daniel R Montello

Scaling and Classification in Social Measurement 4
David R Heise

Schizophrenia 9
Jürgen Zielasek and Wolfgang Gaebel

Schizophrenia, Treatment of 16
Thomas Pollmächer

Schizophrenia: Metabolic Consequences 21
Antonio Ventriglio, Antonello Bellomo, and Dinesh Bhugra

Schmitt, Carl (1888–1985) 28
Lars Vinx

Scholars, Private Lives of 34
Barbara Louis and Thilo Neidhöfer

School (Alternative Models): Ideas and Institutions 39
Achim Leschinsky

School Achievement as Moderated by Sociodemographic Status: The US Case 42
Selcuk R Sirin and Taveeshi Gupta

School Achievement: Motivational Determinants and Processes 48
Friedrich-Wilhelm Schrader and Andreas Helmke

School as a Social System 55
Peter W Cookson, Jr

School Burnout and Engagement: Lessons from a Longitudinal Study in Finland 59
Katariina Salmela-Aro

School Business Management 65
Paul Armstrong

School Choice 71
Mark Berends

School Effectiveness Research 80
Jaap Scheerens

School Environment for LGBTQ/Sexual Minority Youth 86
Stephen T Russell, Amanda M Pollitt, and Jennifer M Hoenig

School Improvement 91
Bert PM Creemers and Leonidas Kyriakides

School Learning for Transfer 96
Gavriel Salomon and David N Perkins

School Management 101
Brian J Caldwell

School Safety 107
Thomas J Mowen

School to School Collaboration 112
Marilyn Friend and Tammy Barron

Schooling: Impact on Cognitive and Motivational Development 119
Ulrich Trautwein, Hanna Dumont, and Anna–Lena Dicke

Schooling: Total Impact of 125
Hans Luyten

Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance 128
Christian Dayé

Schools, Micropolitics of 134
Herbert Altrichter and Robert Moosbrugger

Schumpeter, Joseph A. (1883–1950) 141
Richard Swedberg

Schütz, Alfred (1899–1959) 146
Ilja Srubar

Science and Development 150
Marcus A Ynalvez and Wesley M Shrum

Science and Law 156
Sheila Jasanoff

Science and Media 163
Bruce V Lewenstein

Science and Politics: Value Neutrality 167
Sven Eliaeson

Science and Religion 172
Ryan T Cragun

Science and Social Movements 176
Andrew Jamison

Science and Technology, Anthropology of 182
Michael MJ Fischer

Science and Technology, Social Study of: Computers and Information Technology 186
Geoffrey C Bowker and Susan Leigh Star

Science and Technology Studies, Ethnomethodology of 192
Michael Lynch

Science and Technology Studies, History of 200
Harald Rohracher

Science and Technology Studies: Experts and Expertise 206
Brian Wynne and Michael Lynch

Science and The State 212
Yaron Ezrahi

Science Communication 219
Jane Gregory

Science: Constructivist Perspectives, History of 225
Iwan R Morus

Science Education 229
John K Gilbert

Science, Geography of 236
Diarmid A Finnegan

Science, History of 241
Lorraine Daston

Science, New Forms of 248
Jerome R Ravetz and Silvio O Funtowicz

Science, Public Engagement with 255
Alan Irwin

Science, Sociology of 261
Thomas F Gieryn and Kathleen C Oberlin

Scientific Concepts: Development in Children 268
Jamie Jirout and Amy Pace

Scientific Concepts during Childhood, Development of 275
Mesut Saçkes, Kathy C Trundle, and Mandy McCormick Smith

Scientific Controversies 281
Trevor Pinch

Scientific Disciplines, History of 287
Rudolf Stichweh

Scientific Discovery, Computational Models of 291
Ronald R Grau and Peter C-H Cheng

Scientific Evidence: Legal Aspects 297
Joseph Sanders

Scientific Explanation 302
C Mantzavinos

Scientific Knowledge, Sociology of 308
HM Collins

Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct 313
Ragnvald Kalleberg

Scientific Revolution, History and Sociology of 318
Jürgen Renn and Peter Damerow

Scientometrics 322
Loet Leydesdorff and Staša Milojević

Screening and Selection 328
Joseph L Gastwirth

Search, Economics of 333
Christopher A Pissarides

Searle, John (1932–) 340
Alessandro Salice

Second Demographic Transition 346
James M Raymo

Second Language Acquisition 349
Gerrit Jan Kootstra, Ton Dijkstra, and Marianne Starren

Second Language Learning 360
Ruiqin Miao

Second World War, The 368
Richard JB Bosworth

Secrecy, Anthropology of 374
Lamont Lindstrom

Secular Religions 380
Arthur L Greil

Secularization 383
Jose Casanova

Secularization: History of the Concept 388
Stijn Latré and Guido Vanheeswijck

Securitization 395
Philippe Bourbeau

Security: Commercial Transactions 400
Frank Decker and Sheelagh McCracken

Segregation, Demographic Effects of 407
Michael J White

Segregation Indices 411
Douglas S Massey

Selection Bias, Statistics of 414
Richard Breen

Self and Brain 419
Manos Tsakiris

Self and Emotional Development in Adulthood and Later Life 424
Gisela Labouvie-Vief

Self and Identity Development during Adolescence across Cultures 432
Juan José Zacarés and Alejandro Iborra

Self in Culture: Early Development 439
Vasudevi Reddy

Self: History of the Concept 445
Peter Wagner

Self: Philosophical Aspects 449
Carol Rovane

Self-Categorization Theory 455
S Alexander Haslam and Stephen D Reicher

Self-Concept: From Unidimensional to Multidimensional and Beyond 460
Brooke Van Zanden, Herbert W Marsh, Marjorie Seaton, and Philip Parker

Self-Concepts: Educational Aspects 469
Reinhard Pekrun and Elizabeth J Stephens

Selfconscious Emotions, Psychology of 475
June P Tangney

Self-Control Theory and Crime 481
George E Higgins and Dustin Osborne

Self-Determination Theory 486
Edward L Deci and Richard M Ryan

Self-Development in Childhood and Adolescence 492
Susan Harter

Self-Directed Care and Individual Budgets for Older People: Lessons from the UK 498
Jill Manthorpe

Self-Efficacy 504
August Flammer

Self-Efficacy and Health 509
David P French

Self-Efficacy: Education Aspects 515
Dale H Schunk and Maria K DiBenedetto

Self-Esteem 522
Christian H Jordan, Virgil Zeigler-Hill, and Jessica J Cameron

Self-Organizing Dynamical Systems 529
Nigel Gilbert, David Anzola, Peter Johnson, Corinna Elsenbroich, Tina Balke, and Ozge Dilaver

Self-Regulated Learning 535
Philip H Winne

Self-Regulated Learning: Theories, Measures, and Outcomes 541
Barry J Zimmerman

Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations 547
Edmond P Bowers, G John Geldhof, Paul A Chase, Richard M Lerner, Steinunn Gestsdóttir, and Jennifer Brown Urban

Self-Regulation During Early Childhood Across Cultures, Development of 553
Daina Crafa

Self-Regulation in Adulthood 557
Alexandra M Freund and Marie Hennecke

Semantic Differential 563
Andrea Ploder and Anja Eder

Semantic Memory 572
Julie S Snowden

Semantic Similarity, Cognitive Psychology of 579
Ulrike Hahn and Evan Heit

Semantics 585
John I Saeed

Semiotics 592
Klaus B Jensen

Semiparametric Models 598
Joel L Horowitz

Sensation and Perception: Direct Scaling 604
Martha Teghtsoonian

Sensation Seeking: Behavioral Expressions and Biosocial Bases 607
Marvin Zuckerman

Senses, Anthropology of the 615
David Howes

Sentence Comprehension, Psychology of 621
Charles Clifton, Jr

Sentencing, Trends, and Disparities: Crime 627
Mindy S Bradley and Jeffery T Ulmer

Sequential Statistical Methods 634
Donald A Berry

Serial Order, Representation of 639
Eric Heuer

Serial Verb Constructions 645
Mark Sebba

Service Economy, Geography of 648
Brita Hermelin and Grete Rusten

Settlement and Landscape Archaeology 654
Gary M Feinman

Sex Crimes 659
Mary A Finn and Sebastian Czech

Sex Life Challenges: The Finnish Case 665
Osmo Kontula

Sex Offenders 672
Malcolm Cowburn and Steve Myers

Sex Role Development and Education 678
Angel Nga-Man Leung and Henry KS Ng

Sex Selective Abortion 686
Daniel Goodkind

Sex Therapy, Clinical Psychology of 689
Walter Everaerd

Sexism 693
Fabio Lorenzi-Cioldi and Clara Kulich

Sexual Behavior and Social Networks 700
Benjamin Cornwell and Rachel Behler

Sexual Debut 705
Anthony Paik

Sexual Dysfunction, Epidemiology of 711
David L Rowland and Beth A Adamski

Sexual Harassment: United States and Beyond 721
Anna-Maria Marshall

Sexual Minorities and Sexual Citizenship 728
Carol Johnson

Sexual Minorities in China 734
Day Wong

Sexual Minorities in Modern Russia 740
Brian J Baer

Sexual Minorities in South Africa 744
Zethu Matebeni

Sexual Minorities in the Maghreb 750
Jarrod Hayes

Sexual Minority Families of Choice and Family Networks: We Are Family 755
Emily E Wheeler and Sharon G Horne

Sexual Motivation 759
Martha K McClintock

Sexual Orientation and the Law 768
Stu Marvel and Martha M Ertman

Sexual Orientation: Biological Influences 773
Malvina N Skorska and Anthony F Bogaert

Sexual Orientation: Genetic Aspects 779
Khytam Dawood

Sexual Orientation: Historical and Social Construction 785
Jeffrey Weeks

Sexual Politics in the Arab World 791
Shereen El Feki

Sexual Risk Behaviors 797
Leickness C Simbayi

Sexuality and Social Work 802
Mark Henrickson

Sexuality and Space 808
Lynda Johnston

Sexuality and the Internet 813
Aaron Ben-Ze'ev

Sexuality: Cultural Aspects 819
Daniel Jordan Smith

Sexuality Education 825
Sara S Marques, Norman A Constantine, Eva S Goldfarb, and Jane Mauldon

Sexuality in Australia 833
Juliet Richters

Sexuality Over the Life Course 840
Linda Waite and Cameron Charme

Sexuality, Theories of 846
Elke D Reissing and Heather VanZuylen

Sexually Transmitted Infections: Social Network Analysis 853
Yoosik Youm

Shadow Work: Measurement 862
Friedrich Schneider

Shamanism 868
Donald Joralemon

Shame and the Social Bond 871
Thomas Scheff

Shared Belief: Philosophical Aspects 877
Raimo Tuomela

Sherrington, Sir Charles Scott (1857–1952) 881
Eike D Schomburg

Shils, Edward (1910–95) 884
Stephen Turner and Steven Grosby

Shklar, Judith N (1928–92) 889
Andreas Hess

Short- and Long-term Memory during Infancy and Early Childhood across Cultures 896
Patricia J Bauer

Short Term Memories, Theories of 901
Evie Vergauwe and Nelson Cowan

Short-Term Memory: Psychological and Neural Aspects 909
Giuseppe Vallar

Shrinking Cities 917
Robert Beauregard

Sibling-Order Effects 923
Frank J Sulloway

Sign Language 928
Ursula Bellugi and Edward S Klima

Sign Language, Disorders of 934
David P Corina

Sign Language: Psychological and Neural Aspects 942
David P Corina

Signal Detection Theory 946
Michael Hautus

Signal Detection Theory, History of 952
Carleton Douglas Creelman

Significance, Tests of 957
Nancy Reid

Simmel, Georg (1858–1918) 963
O Rammstedt and Natàlia Cantó-Milà

Simon, Herbert (1916–2001) 970
John-Christopher Spender

Simultaneous Equation Estimates (Exact and Approximate), Distribution of 975
Kimio Morimune

Simultaneous Equation Estimation: Overview 980
CF Christ

Single-Case Designs: Methodology 984
Christopher R Berghoff and John P Forsyth

This page intentionally left blank

CONTRIBUTORS TO VOLUME 21

Beth A. Adamski
Valparaiso University, Valparaiso, IN, USA

Herbert Altrichter
Johannes Kepler University, Linz, Austria

David Anzola
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Paul Armstrong
University of Manchester, Manchester, UK

Brian J. Baer
Kent State University, Kent, OH, USA

Tina Balke
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Tammy Barron
The University of North Carolina at Greensboro, Cooleemee, NC, USA

Patricia J. Bauer
Emory University, Atlanta, GA, USA

Robert Beauregard
Columbia University, New York, NY, USA

Rachel Behler
Cornell University, Ithaca, NY, USA

Antonello Bellomo
University of Foggia, Foggia, Italy

Ursula Bellugi
Salk Institute, La Jolla, CA, USA

Aaron Ben-Ze'ev
University of Haifa, Haifa, Israel

Mark Berends
Center for Research on Educational Opportunity (CREO), Research Center on Educational Policy (RCEP), University of Notre Dame, Notre Dame, IN, USA

Christopher R. Berghoff
University at Albany, Albany, NY, USA

Donald A. Berry
University of Texas, Houston, TX, USA

Dinesh Bhugra
Institute of Psychiatry, King's College, London, UK

Anthony F. Bogaert
Brock University, St. Catharines, ON, Canada

Richard J.B. Bosworth
Jesus College, Oxford, UK; The University of Western Australia (UWA), Crawley, WA, Australia; and Reading University, Reading, UK

Philippe Bourbeau
University of Cambridge, Cambridge, UK

Edmond P. Bowers
Tufts University, Medford, MA, USA

Geoffrey C. Bowker
Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, CA, USA

Mindy S. Bradley
University of Arkansas, Fayetteville, AR, USA

Richard Breen
Yale University, New Haven, CT, USA

Brian J. Caldwell
University of Melbourne, Melbourne, VIC, Australia

Jessica J. Cameron
University of Manitoba, Winnipeg, MB, Canada

Natàlia Cantó-Milà
Universitat Oberta de Catalunya (UOC), Barcelona, Spain

Jose Casanova
New School of Social Research, New York, NY, USA

Cameron Charme
NORC, University of Chicago, Chicago, IL, USA

Paul A. Chase
Tufts University, Medford, MA, USA

Peter C.-H. Cheng
University of Sussex, Falmer, Brighton, UK

C.F. Christ
The Johns Hopkins University, Baltimore, MD, USA

Charles Clifton
University of Massachusetts, Amherst, MA, USA

H.M. Collins
SOCSI, Cardiff University, Cardiff, UK

Norman A. Constantine
University of California, Berkeley, CA, USA; and Public Health Institute, Oakland, CA, USA

Peter W. Cookson
American Institutes for Research, Washington, DC, USA

David P. Corina
Cognitive Neurolinguistics Laboratory, Center for Mind and Brain, University of California, Davis, CA, USA

David P. Corina
University of Washington, Seattle, WA, USA

Benjamin Cornwell
Cornell University, Ithaca, NY, USA

Nelson Cowan
University of Missouri-Columbia, Columbia, MO, USA

Malcolm Cowburn
Sheffield Hallam University, Sheffield, UK

Daina Crafa
Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

Ryan T. Cragun
University of Tampa, Tampa, FL, USA

Carleton Douglas Creelman
University of Toronto, Toronto, ON, Canada

Bert P.M. Creemers
Faculty of Behavioural and Social Sciences, GION Institute for Educational Research, University of Groningen, The Netherlands

Sebastian Czech
Georgia State University, Atlanta, GA, USA

Peter Damerow[†]
Max Planck Institute for the History of Science, Berlin, Germany

Lorraine Daston
Max Planck Institute for the History of Science, Berlin, Germany

Khytam Dawood
Pennsylvania State University, University Park, PA, USA

Christian Dayé
University of Graz, Graz, Austria

Edward L. Deci
University of Rochester, Rochester, NY, USA

Frank Decker
NSW, Australia

Maria K. DiBenedetto
The City University of New York, New York, NY, USA

Anna–Lena Dicke
University of Tübingen, Tübingen, Germany

Ton Dijkstra
Donders Institute for Brain, Cognition, and Behaviour; and Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

Ozge Dilaver
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Hanna Dumont
German Institute of International Educational Research, Berlin, Germany

Anja Eder
University of Graz, Graz, Austria

Shereen El Feki
London, UK

Sven Eliaeson
Uppsala centre for Russian and Eurasian at Uppsala University, Uppsala and Centre for Social Studies, Instiute for Philosophy and Sociology at the Polish Academy of Science, Warsaw

Corinna Elsenbroich
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Martha M. Ertman
University of Maryland, Francis King Carey School of Law, Baltimore, MD, USA

Walter Everaerd
University of Amsterdam, Amsterdam, The Netherlands

Yaron Ezrahi
The Hebrew University of Jerusalem, Jerusalem, Israel

[†] Deceased.

Gary M. Feinman
The Field Museum, Chicago, IL, USA

Mary A. Finn
Georgia State University, Atlanta, GA, USA

Diarmid A. Finnegan
Queen's University, Belfast, UK

Michael M.J. Fischer
MIT, Cambridge, MA, USA; Harvard Medical School, Boston, MA, USA; and National University of Singapore, Singapore

August Flammer
University of Berne, Bolligen, Switzerland

John P. Forsyth
University at Albany, Albany, NY, USA

David P. French
School of Psychological Sciences, University of Manchester, Manchester, UK

Alexandra M. Freund
University of Zurich, Zurich, Switzerland

Marilyn Friend
The University of North Carolina at Greensboro, Greensboro, NC, USA

Silvio O. Funtowicz
Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Bergen, Norway

Wolfgang Gaebel
Heinrich Heine University, Düsseldorf, Germany

Joseph L. Gastwirth
George Washington University, Washington, DC, USA

G. John Geldhof
Tufts University, Medford, MA, USA

Steinunn Gestsdóttir
University of Iceland, Stakkahlíð, Reykjavík, Iceland

Thomas F. Gieryn
Indiana University Bloomington, Bloomington, IN, USA

John K. Gilbert
King's College London, London, UK; and The University of Reading, Reading, UK

Nigel Gilbert
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Eva S. Goldfarb
Montclair State University, Montclair, NJ, USA

Daniel Goodkind
Independent Researcher, Arlington, VA, USA

Ronald R. Grau
University of Sussex, Falmer, Brighton, UK

Jane Gregory
CHSTM, University of Manchester, Manchester, UK

Arthur L. Greil
Alfred University, Alfred, NY, USA

Steven Grosby
Clemson University, Clemson, SC, USA

Taveeshi Gupta
New York University, New York, NY, USA

Ulrike Hahn
Birkbeck, University of London, London, UK

Susan Harter
University of Denver, Denver, CO, USA

S. Alexander Haslam
School of Psychology, University of Queensland, St Lucia, QLD, Australia

Michael Hautus
The University of Auckland, Auckland, New Zealand

Jarrod Hayes
University of Michigan, Ann Arbor, MI, USA

David R. Heise
Indiana University, Bloomington, IN, USA

Evan Heit
University of California, Merced, CA, USA

Andreas Helmke
University Koblenz-Landau, Landau, Germany

Marie Hennecke
University of Zurich, Zurich, Switzerland

Mark Henrickson
Massey University, Auckland, New Zealand

Brita Hermelin
Linköping University, Linköping, Sweden

Andreas Hess
School of Sociology, University College Dublin, Dublin, Ireland

Eric Heuer
University of Hawaii at Hilo, Hilo, HI, USA

George E. Higgins
University of Louisville, Louisville, KY, USA

Jennifer M. Hoenig
University of Arizona, Tucson, AZ, USA

Sharon G. Horne
University of Massachusetts Boston, Boston, MA, USA

Joel L. Horowitz
Northwestern University, Evanston, IL, USA

David Howes
Concordia University, Montreal, QC, Canada

Alejandro Iborra
Faculty of Education, University of Alcalá, Alcalá de Henares, Spain

Alan Irwin
Copenhagen Business School, Frederiksberg, Denmark

Andrew Jamison
Aalborg University, Aalborg, Denmark

Sheila Jasanoff
John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA

Klaus B. Jensen
University of Copenhagen, Copenhagen, Denmark

Jamie Jirout
Temple University, Ambler, PA, USA

Carol Johnson
University of Adelaide, Adelaide, SA, Australia

Peter Johnson
Centre for Research in Social Simulation, University of Surrey, Guildford, UK

Lynda Johnston
University of Waikato, Hamilton, Aotearoa, New Zealand

Donald Joralemon
Smith College, Northampton, MA, USA

Christian H. Jordan
Wilfrid Laurier University, Waterloo, ON, Canada

Ragnvald Kalleberg
University of Oslo, Oslo, Norway

Edward S. Klima
Salk Institute, La Jolla, CA, USA

Osmo Kontula
Population Research Institute, Family Federation of Finland, Helsinki, Finland

Gerrit Jan Kootstra
Center for Language Science, Pennsylvania State University, University Park, PA, USA

Henry KS Ng
The University of Hong Kong, Hong Kong

Clara Kulich
Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

Leonidas Kyriakides
Faculty of Social Sciences and Education, University of Cyprus, Cyprus

Gisela Labouvie-Vief
University of Geneva, Geneva, Switzerland

Stijn Latré
Center Pieter Gillis and Center for Philosophy of Culture, University of Antwerp, Antwerp, Belgium

Susan Leigh Star[†]
University of California, La Jolla, CA, USA

Richard M. Lerner
Tufts University, Medford, MA, USA

Achim Leschinsky
Humboldt University, Berlin, Germany

Angel Nga-Man Leung
Center for Psychosocial Health and Aging, The Hong Kong Institute of Education, New Territories, Hong Kong

Bruce V. Lewenstein
Cornell University, Ithaca, NY, USA

Loet Leydesdorff
University of Amsterdam, Amsterdam, The Netherlands

Lamont Lindstrom
University of Tulsa, Tulsa, OK, USA

Fabio Lorenzi-Cioldi
Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

Barbara Louis
University of Minnesota, Minneapolis, MN, USA

Hans Luyten
Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

Michael Lynch
Cornell University, Ithaca, NY, USA

Jill Manthorpe
King's College London, London, UK

[†] Deceased.

C. Mantzavinos
University of Athens, Athens, Greece

Sara S. Marques
University of California, Berkeley, CA, USA; and Public Health Institute, Oakland, CA, USA

Herbert W. Marsh
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield NSW, Australia

Anna-Maria Marshall
University of Illinois, Urbana-Champaign, Urbana, IL, USA

Stu Marvel
Emory University School of Law, Atlanta, GA, USA

Douglas S. Massey
Office of Population Research, Princeton University, Princeton, NJ, USA

Zethu Matebeni
University of Cape Town, Rondebosch, Cape Town, South Africa

Jane Mauldon
University of California, Berkeley, CA, USA

Martha K. McClintock
The Institute for Mind and Biology, The University of Chicago, Chicago, IL, USA

Sheelagh McCracken
The University of Sydney, Sydney, NSW, Australia

Ruiqin Miao
School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, People's Republic of China

Staša Milojević
Indiana University, Bloomington, IN, USA

Daniel R. Montello
University of California, Santa Barbara, Santa Barbara, CA, USA

Robert Moosbrugger
Johannes Kepler University, Linz, Austria

Kimio Morimune
Kyoto University, Kyoto, Japan

Iwan R. Morus
Aberystwyth University, Aberystwyth, UK

Thomas J. Mowen
University of Delaware, Newark, DE, USA

Steve Myers
University of Salford, Salford, UK

Thilo Neidhöfer
Johannes Kepler University of Linz, Linz, Austria

Kathleen C. Oberlin
Indiana University Bloomington, Bloomington, IN, USA

Dustin Osborne
University of Louisville, Louisville, KY, USA

Amy Pace
Temple University, Ambler, PA, USA

Anthony Paik
University of Iowa, Iowa City, IA, USA

Philip Parker
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield NSW, Australia

Reinhard Pekrun
University of Munich, Munich, Germany

David N. Perkins
Harvard University, Cambridge, MA, USA

Trevor Pinch
Cornell University, Ithaca, NY, USA

Christopher A. Pissarides
London School of Economics, London, UK

Andrea Ploder
University of Graz, Graz, Austria

Amanda M. Pollitt
University of Arizona, Tucson, AZ, USA

Thomas Pollmächer
Max Planck Institute, Munich, Germany

O. Rammstedt
University of Bielefeld, Germany

Jerome R. Ravetz
Institute for Science, Innovation and Society, University of Oxford, Oxford, UK

James M. Raymo
University of Wisconsin-Madison, Madison, WI, USA

Vasudevi Reddy
University of Portsmouth, Portsmouth, UK

Stephen D. Reicher
School of Psychology, University of St. Andrews, St Andrews, UK

Nancy Reid
University of Toronto, Toronto, ON, Canada

Elke D. Reissing
School of Psychology, University of Ottawa, Ottawa, ON, Canada

Jürgen Renn
Max Planck Institute for the History of Science, Berlin, Germany

Juliet Richters
University of New South Wales, Sydney, NSW, Australia

Harald Rohracher
Linköping University, Linköping, Sweden

Carol Rovane
Columbia University, New York, NY, USA

David L. Rowland
Valparaiso University, Valparaiso, IN, USA

Stephen T. Russell
University of Arizona, Tucson, AZ, USA

Grete Rusten
University of Bergen, Bergen, Norway

Richard M. Ryan
University of Rochester, Rochester, NY, USA

Mesut Saçkes
Balõkesir University, Balõkesir, Turkey

John I. Saeed
Centre for Language and Communication Studies, Trinity College Dublin, Dublin, Ireland

Alessandro Salice
Center for Subjectivity Research, University of Copenhagen, Copenhagen S, Denmark

Katariina Salmela-Aro
Cicero Learning, University of Helsinki, Finland

Gavriel Salomon
Haifa University, Haifa, Israel

Joseph Sanders
University of Houston Law Center, Houston, TX, USA

Jaap Scheerens
Faculty of Behavioral Sciences, University of Twente, Enschede, The Netherlands

Thomas Scheff
UCSB, Santa Barbara, CA, USA

Friedrich Schneider
Johannes Kepler University of Linz, Linz, Austria

Eike D. Schomburg
Georg August University, Gottengen, Germany

Friedrich-Wilhelm Schrader
University Koblenz-Landau, Landau, Germany

Dale H. Schunk
The University of North Carolina at Greensboro, Greensboro, NC, USA

Marjorie Seaton
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield NSW, Australia

Mark Sebba
Lancaster University, Lancaster, UK

Wesley M. Shrum
Louisiana State University, Baton Rouge, LA, USA

Leickness C. Simbayi
Human Sciences Research Council, Cape Town, South Africa; and Psychiatry and Mental Health, The University of Cape Town, South Africa

Selcuk R. Sirin
New York University, New York, NY, USA

Malvina N. Skorska
Brock University, St. Catharines, ON, Canada

Daniel Jordan Smith
Brown University, Providence, RI, USA

Mandy McCormick Smith
The Ohio State University, Columbus, OH, USA

Julie S. Snowden
Manchester Academic Health Sciences Centre, Institute of Brain, Behaviour and Mental Health, University of Manchester, UK; and Greater Manchester Neuroscience Centre, Salford Royal NHS Foundation Trust, Salford, UK

John-Christopher Spender
ESADE, Universitat Ramon Lull, Barcelona, Spain; and School of Economics and Management, Lund University, Lund, Sweden

Ilja Srubar
Institute of Sociology, University of Erlangen, Konstanz, Germany

Marianne Starren
Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

Elizabeth J. Stephens
University of Munich, Munich, Germany

Rudolf Stichweh
Universität Bielefeld, Bielefeld, Germany

Frank J. Sulloway
University of California Berkeley, Berkeley, CA, USA

Richard Swedberg
Stockholm University, Stockholm, Sweden

June P. Tangney
George Mason University, Fairfax, VA, USA

Martha Teghtsoonian
Smith College, Northampton, MA, USA

Ulrich Trautwein
University of Tübingen, Tübingen, Germany

Kathy C. Trundle
The Ohio State University, Columbus, OH, USA

Manos Tsakiris
Royal Holloway, University of London, Egham, UK

Raimo Tuomela
University of Helsinki, Helsinki, Finland; and University of Munich, Munich, Germany

Stephen Turner
University of South Florida, Tampa, FL, USA

Jeffery T. Ulmer
The Pennsylvania State University, University Park, PA, USA

Jennifer Brown Urban
Montclair State University, Montclair, NJ, USA

Giuseppe Vallar
University of Milano-Bicocca, Milan, Italy; and IRCCS Istituto Auxologico Italiano, Milan, Italy

Guido Vanheeswijck
Center Pieter Gillis and Center for Philosophy of Culture, University of Antwerp, Antwerp, Belgium

Brooke Van Zanden
Institute for Positive Psychology and Education, Australian Catholic University, Strathfield NSW, Australia

Heather VanZuylen
School of Psychology, University of Ottawa, Ottawa, ON, Canada

Antonio Ventriglio
University of Foggia, Foggia, Italy

Evie Vergauwe
University of Missouri-Columbia, Columbia, MO, USA

Lars Vinx
Bilkent University, Ankara, Turkey

Peter Wagner
University of Warwick, Coventry, UK; and European University Institute, San Domenico di Fiesole, Italy

Linda Waite
University of Chicago, Chicago, IL, USA

Jeffrey Weeks
London South Bank University, London, UK

Emily E. Wheeler
University of Massachusetts Boston, Boston, MA, USA

Michael J. White
Population Studies and Training Center, Brown University, Providence, RI, USA

Philip H. Winne
Simon Fraser University, Burnaby, BC, Canada

Day Wong
Hong Kong Baptist University, Hong Kong

Brian Wynne
University of Lancaster, Lancaster, UK

Marcus A. Ynalvez
Texas A&M International University, Laredo, TX, USA

Yoosik Youm
Yonsei University, Seo-dae-moon-ku, Seoul, South Korea

Juan José Zacarés
Faculty of Psychology, University of Valencia, Valencia, Spain

Virgil Zeigler-Hill
Oakland University, Rochester, MI, USA

Jürgen Zielasek
Heinrich Heine University, Düsseldorf, Germany

Barry J. Zimmerman
City University of New York Graduate Center, NY, USA

Marvin Zuckerman
University of Delaware, Philadelphia, PA, USA

This page intentionally left blank

Scale in Geography

Daniel R Montello, University of California, Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scale is about size, and is one of the most fundamental issues in geography and other disciplines that study phenomena distributed over the earth's surface. The concept of scale has at least three referents: cartographic scale, analysis scale, and phenomenon scale. It is widely recognized that in order to observe and study geographic phenomena most accurately, the scale at which they are mapped and analyzed must match the actual scale of the phenomena. Scale is also important in part because it is closely allied with the problem of the generalization of geographic information.

Scale is about size, either relative or absolute, and involves a fundamental set of issues in geography (Howitt, 2003; Hudson, 1992). Scale primarily concerns spatiality in geography, and this article focuses on spatial scale. However, temporal scale and thematic scale are also important. Temporal scale deals with the size of temporal units; an example is a process that changes on a daily versus annual basis. Thematic scale concerns the grouping of entities or attributes into classes; an example is studying lakes versus water bodies versus landscape features. Whether spatial, temporal, or thematic scale in fact has several meanings.

Three Meanings of Scale

The concept of scale can be confusing, insofar as it has multiple referents (Sheppard and McMaster, 2004). Cartographic scale refers to the depicted size of a feature on a map relative to its actual size in the world. Analysis scale refers to the size of the unit at which some phenomenon is empirically measured and analyzed, such as at the county or state level. Phenomenon scale refers to the size at which human or physical earth structures or processes exist in reality, regardless of how they are studied or represented. Although discussions of scale frequently treat the three referents independently, they are in fact interrelated in important ways that are relevant to all geographers, and the focus of research for some. For example, choices concerning the scale at which a map should be made depend in part on the scale at which measurements of earth features are made and the scale at which a phenomenon of interest actually exists.

Cartographic Scale

Maps are smaller than the part of the earth's surface they depict. Cartographic scale expresses this relationship, traditionally in one of three ways (Muehrcke et al., 2001). A verbal scale statement expresses the amount of distance on the map that represents a particular distance on the earth's surface in words; for example, '1 cm equals 10 km.' The representative fraction (RF) expresses scale as a numerical ratio of map distance to earth distance; for example, 1:500 000. As compared to a verbal scale statement, the RF has the advantage of being a unitless measure. Finally, a graphic scale bar uses a line of particular length drawn on the map and annotated to show how much earth distance it represents. A graphic scale bar has the advantage that it changes size appropriately when the map is enlarged or reduced. Alternatively, all three expressions of scale may refer to areal measurements rather than linear measurements; for example, a 1-cm square may represent 1 ha on the earth. This areal expression is nontraditional.

Given a map image of fixed size, as the size of the earth surface it represents gets larger, the RF gets smaller (i.e., the denominator of the RF becomes a larger number). Hence, a 'large-scale map' shows a relatively small area of the earth, such as a county or city, and a 'small-scale map' shows a relatively large area, such as a continent or a hemisphere of the earth. This cartographic scale terminology is frequently felt to be counterintuitive when applied to analysis or phenomenon scales, where small scale and large scale usually refer to small and large entities, respectively.

An important complexity about cartographic scale is that flat maps invariably distort spatial relations on the earth's surface, due to the intractable problem of displaying a round earth surface on a flat image. Of course, spatial continuity is always violated, as flat images must have discrete edges even though the near-spherical earth surface has no such edges. Distances, directions, and shapes can never be shown accurately everywhere on a map surface, although particular maps can show subsets of such properties accurately. In fact, only the areal sizes of global features can be shown accurately everywhere across the map image, and very often they are not shown this way. How maps distort spatial relations is part of the topic of map projections. In many projections, especially small-scale maps that show large parts of the earth, this distortion is extreme so that linear or areal scale on one part of the map is very different from that on other parts. Variable scale is sometimes shown on a map by the use of a special symbol or multiple symbols at different locations.

Analysis Scale

Analysis scale refers to the size of the units in which phenomena are measured and the size of the units into which measurements are aggregated for data analysis and mapping. It is essentially the scale at which geographic phenomena are

understood. Terms such as 'resolution' or 'granularity' are often used as synonyms for the scale of analysis, particularly when geographers work with digital representations of the earth's surface in a computer by means of a regular grid of small cells in a satellite image (rasters) or on a computer screen (pixels). Analysis scale here refers to the area of earth surface represented by a single cell.

It has long been recognized that in order to observe and study a phenomenon most accurately, the scale of analysis must match the actual scale of the phenomenon. This is true for all three domains of scales – spatial, temporal, and thematic. Identifying the correct scale of phenomena is thus a central problem for geographers. Particularly when talking about thematic scale, using data at one scale to make inferences about phenomena at other scales is known as the cross-level fallacy (the narrower case of using aggregated data to make inferences about disaggregated data is well known as the ecological fallacy).

Geographers often analyze phenomena at what might be called 'available scale,' the units that are present in available data. Many problems of analysis scale arise from this practice, but it is unavoidable given the difficulty and expense involved in collecting many types of data over large parts of the earth's surface. Geographers have little choice in some cases but to analyze phenomena with secondary data, data collected by others not specifically for the purposes of a particular research project. For example, census bureaus in many countries provide a wealth of data on many social, demographic, and economic characteristics of their populace. Frequently, the phenomenon of interest does not operate according to the boundaries of existing administrative or political units in the data, which after all were not created primarily to serve the needs of geographic analysis. The resolution of image scanners on remote-sensing satellites provides another important example. Landsat imagery is derived from thematic mapper sensors, producing earth measurements at a resolution of about 30 by 30 m. However, many phenomena occur at finer resolutions than these data can provide.

It would be very useful to have a theory about the scale of a phenomenon's existence. Frequently lacking this, but realizing that the available scale may not be suitable, geographers use empirical 'trial-and-error' approaches to try to identify the appropriate scale at which a phenomenon should be analyzed. Given spatial units of a particular size, one can readily aggregate or combine them into larger units; it is not possible without additional information or theory to disaggregate them into smaller units. Even given observations measured at very small units, however, there is still the problem of deciding in what way the units should be aggregated. This is known as the modifiable areal unit problem (MAUP, or MTUP in the case of temporal scale), and concerns not only the appropriate size at which to aggregate units but also the shape and location of the aggregated units. Various techniques have been developed to study the implications of MAUP (Openshaw, 1983), and it is an ongoing issue for researchers.

Phenomenon Scale

Phenomenon scale refers to the size at which geographic structures exist and over which geographic processes operate in the world. It is the 'true' scale of geographic phenomena. Determining the scale of phenomena is clearly a major research goal in geography. It is a common geographic dictum that scale matters. Numerous concepts in geography reflect the idea that phenomena are scale dependent or are defined in part by their scale (Meentemeyer, 1989). Vegetation stands are smaller than vegetation regions, and linguistic dialects are distributed over smaller areas than languages. The possibility that some geographic phenomena are scale independent is important, however. Patterns seen at one scale may often be observed at other scales; whether this is a matter of analogy or of the same processes operating at multiple scales is theoretically important. The mathematics of fractals has been applied in geography as a way of understanding and formalizing phenomena such as coastlines that are self-similar at different scales (Lam and Quattrochi, 1992).

The belief has often been expressed that the discipline of geography, as the study of the earth as the home of humanity, can be defined partially by its focus on phenomena at certain scales, such as cities or continents, and not other scales. The range of scales of interest to geographers is often summarized by the use of terminological continua such as 'local–global' or 'micro-, meso-, and macroscales' (Meentemeyer, 1989; Meyer et al., 1992). The view that geographers must restrict their focus to particular ranges of scales is not shared universally, however, and advances have and will continue to occur when geographers stretch the boundaries of their subject matter. Nonetheless, few would argue that subatomic or interplanetary scales are properly of concern for geography.

It is widely recognized that various scales of geographic phenomena interact, or that phenomena at one scale emerge from smaller or larger scale phenomena. This is captured by the notion of a hierarchy of scales, in which smaller phenomena are nested within larger phenomena (although the term 'nested' should not be taken to indicate that causal relations operate in only one direction, from the superordinate to the subordinate, for instance). Local economies are nested within regional economies, rivers are nested within larger hydrologic systems. Conceptualizing, measuring, and modeling such scale hierarchies can be quite difficult, and the traditional practice within geography of focusing on a single scale largely continues (Herod, 2011).

Generalization

The world can never be studied, modeled, or represented in all of its full detail and complexity. Scale is important in part because of its consequences for the degree to which geographic information is generalized. Generalization refers to the amount of detail included in information; it is essentially an issue of simplification, but also includes aspects of selection and enhancement of features of particular interest (Buttenfield and McMaster, 1991). As one studies or represents smaller pieces of the earth, one tends strongly to deal with more detailed or more fine-grained aspects of geographic features. For example, large-scale maps almost always show features on the earth's surface in greater detail than do small-scale maps; rivers appear to

meander more when they are shown on large-scale maps, for instance. Studied most extensively by cartographers, generalization is in fact relevant to all three meanings of scale, and to all three domains of spatial, temporal, and thematic scales.

Conclusion

Issues of scale have always been central to geographic theory and research. Advances in the understanding of scale and the ability to investigate scale-related problems will continue, particularly with the increasingly common representation of geographic phenomena through the medium of digital geographic information (Quattrochi and Goodchild, 1997). Cartographic scale is becoming 'visualization' scale. How is scale – whether spatial, temporal, or thematic – communicated in dynamic, multidimensional, and multimodal representations, including when visualized in virtual environments? Great progress has been made on the problem of automated generalization, programing artificial intelligence to make generalization changes in geographic data as scale changes. The ability to perform multiscale and hierarchical analysis will be developed further. More profound than these advances, however, the widespread emergence of the digital world will foster new conceptions of scale in geography, for both laypersons and specialists (Goodchild and Proctor, 1997).

See also: Cartography; Censuses: Current Approaches and Methods; Ecological Fallacy, Statistics of; Geographic Information Systems; Geospatial Applications in Archaeology; Location, Absolute and Relative; Place in Geography; Region: Geographical; Remote Sensing; Spatial Analysis in Geography; Spatial Data; Spatial Pattern, Analysis of; Spatial Sampling; Time-Space in Geography.

Bibliography

Buttenfield, Barbara P., McMaster, Robert B. (Eds.), 1991. Map Generalization: Making Rules for Knowledge Representation. Wiley, New York.

Goodchild, Michael F., Proctor, James, 1997. Scale in a digital geographic world. Geographical and Environmental Modeling 1, 5–23.

Herod, Andrew, 2011. Scale. Routledge, New York.

Howitt, Richard, 2003. Scale. In: Agnew, John, Mitchell, Katheryne, Toal, Gerard (Eds.), A Companion to Political Geography. Blackwell, Oxford, pp. 138–157.

Hudson, John C., 1992. Scale in space and time. In: F Abler, Ronald, G Marcus, Melvin, Olson, Judy M. (Eds.), Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography. Rutgers University Press, New Brunswick, NJ, pp. 280–297.

Lam, Nina S-N., Quattrochi, Dale A., 1992. On the issues of scale, resolution, and fractal analysis in the mapping sciences. The Professional Geographer 44, 88–98.

Meentemeyer, Vernon, 1989. Geographical perspectives of space, time, and scale. Landscape Ecology 3, 163–173.

Meyer, William B., Gregory, Derek, Turner II, B.L., McDowell, Patricia F., 1992. The local-global continuum. In: F Abler, Ronald, G Marcus, Melvin, Olson, Judy M. (Eds.), Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography. Rutgers University Press, New Brunswick, NJ, pp. 255–279.

Muehrcke, Philip C., Muehrcke, Juliana O., Kimerling, A. Jon, 2001. Map Use: Reading, Analysis, and Interpretation, third ed. JP Publications, Madison, WI.

Openshaw, Stanley, 1983. The Modifiable Areal Unit Problem. Geo Books, Norfolk, UK.

Quattrochi, Dale A., Goodchild, Michael F., 1997. Scale in Remote Sensing and GIS. Lewis Publishers, Boca Raton, FL.

Sheppard, Eric, McMaster, Robert B. (Eds.), 2004. Scale and Geographic Inquiry: Nature, Society, and Method. Blackwell, Oxford.

Scaling and Classification in Social Measurement

David R Heise, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13504–13508, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Every classification category is located within a taxonomy. A typology differentiates entities at a particular level of a taxonomy in terms of one or more properties. Quantitative measurements also differentiate entities at a given taxonomic level, serving like typological classifications, but obtaining greater logical and mathematical power by ordering the classification categories. Classifications ordered along a dimension of variation to provide quantitative measurements constitute a scale. Measurements and theories are interdependent for several reasons. First, taxonomies and typologies are theoretical constructions and are part of all measurements. Second, the notion of measurement error arises from the idea that theories apply to and define true values as opposed to observed values. Third, the development of meaningful measurements requires theoretical conceptualization of relations between variables, while testing a theory's validity requires adequate measurements.

Social measurements translate observed characteristics of individuals, events, relationships, organizations, societies, etc. into symbolic classifications that enable reasoning of a verbal, logical, or mathematical nature. Qualitative research and censuses together define one realm of measurement, concerned with assignment of entities to classification categories embedded within taxonomies and typologies. Another topic in measurement involves scaling discrete items of information such as answers to questions so as to produce quantitative measurements for mathematical analyses. A third issue is the linkage between social measurements and social theories.

Classifications

Classification assimilates perceived phenomena into symbolically labeled categories. Anthropological studies of folk classification systems (D'Andrade, 1995) have advanced the understanding of scientific classification systems, though scientific usages involve criteria that folk systems may not meet entirely.

Two areas of social science employ classification systems centrally. Qualitative analyses such as ethnographies, histories, case studies, etc. offer classifications – sometimes newly invented – for translating experiences in unfamiliar cultures or minds into familiar terms. Censuses of individuals, of occurrences, or of aggregate social units apply classifications – usually traditional – in order to count entities and their variations. Both types of work depend on theoretical constructions that link classification categories.

Taxonomies

Every classification category is located within a taxonomy. Some more general categorization, Y, determines which entities are in the domain for the focal categorization, X; so an X must always be a kind of Y. 'X is a kind of Y' is the linguistic frame for specifying taxonomies. Concepts constituting a taxonomy form a logic tree, with subordinate elements implying superordinate items.

Taxonomic enclosure of a classification category is a social construction that may have both theoretical and practical consequences. For example, if only violent crimes are subject to classification as homicides, then 'homicide' is a kind of 'violent crime,' and deaths caused by executive directives to release deadly pollution could not be homicides.

Typologies

A typology differentiates entities at a particular level of a taxonomy in terms of one or more of their properties. The differentiating property (sometimes called a feature or attribute) essentially acts as a modifier of entities at that taxonomic level. For example, in the United States kinship system siblings are distinguished in terms of whether they are male or female; in Japan by comparison, siblings are schematized in terms of whether they are older as well as whether they are male or female.

A scientific typology differentiates entities into types that are exclusive and exhaustive: every entity at the relevant taxonomic level is of one defined type only, and every entity is of some defined type. A division into two types is a dichotomy, into three types a trichotomy, and into more than three types a polytomy.

Polytomous typologies are often constructed by crossing multiple properties, forming a table in which each cell is a theoretical type. (The crossed properties might be referred to as variables, dimensions, or factors in the typology.) For example, members of a multiplex society have been characterized according to whether they do or do not accept the society's goals on the one hand, and whether they do or do not accept the society's means of achieving goals on the other hand; then crossing acceptance of goals and means produces a fourfold table defining conformists and three types of deviants.

Etic–emic analyses involve defining a typology with properties of scientific interest (the etic system) and then discovering ethnographically which types and combinations of types are recognized in folk meanings (the emic system). Latent structure analysis statistically processes observed properties of

a sample of entities in order to confirm the existence of hypothesized types and to define the types operationally.

Aggregate Entities

Aggregate social entities such as organizations, communities, and cultures may be studied as unique cases, where measurements identify and order internal characteristics of the entity rather than relate one aggregate entity to another.

A seeming enigma in social measurement is how aggregate social entities can be described satisfactorily on the basis of the reports of relatively few informants, even though statistical theory calls for substantial samples of respondents to survey populations. The key is that informants all report on the same thing – a single culture, community, or organization – whereas respondents in a social survey typically report on diverse things – their own personal characteristics, beliefs, or experiences. Thus, reports from informants serve as multiple indicators of a single state, and the number needed depends on how many observations are needed to define a point reliably, rather than how many respondents are needed to describe a population's diversity reliably. As few as seven expert informants can yield reliable descriptions of aggregate social entities, though more are needed as informants' expertise declines (Romney et al., 1986). Informant expertise correlates with greater intelligence and experience (D'Andrade, 1995) and with having a high level of social integration (Thomas and Heise, 1995).

Relations

Case grammar in linguistics defines events and relationships in terms of an actor, action, object, and perhaps instrumentation, setting, products, and other factors as well. Mapping sentences (Shye et al., 1994) apply the case grammar idea with relatively small lists of entities in order to classify relational phenomena within social aggregates. For example, interpersonal relations in a group can be specified by pairing group members with actions such as loves, admires, annoys, befriends, and angers. Mapping sentences defining the relations between individuals or among social organizations constitute the measurement model for social network research (Watts, 2004).

Scaling

Quantitative measurements differentiate entities at a given taxonomic level – serving like typological classifications, but obtaining greater logical and mathematical power by ordering the classification categories.

An influential conceptualization (Stevens, 1951) posited four levels of quantification in terms of how numbers relate to classification categories. *Nominal* numeration involves assigning numbers arbitrarily simply to give categories unique names, such as 'batch 243.' An *ordinal* scale's categories are ordered monotonically in terms of greater-than and less-than, and numbering corresponds to the rank of each category. Numerical ranking of an individual's preferences for different foods is an example of ordinal measurement. Differences between categories can be compared in an *interval* scale, and numbers applied to categories reflect degrees of differences. Calendar dates are an example of interval measurements – we know from their birth years that William Shakespeare was closer in time to Geoffrey Chaucer than Albert Einstein was to Isaac Newton. In a *ratio* scale, categories have magnitudes that are whole or fractional multiples of one another, and numbers assigned to the categories represent these magnitudes. Population sizes are an example of ratio measurements. Knowing the populations of both nations, we can say that Japan is at least 35 times bigger than Jamaica.

A key methodological concern in psychometrics (Hopkins, 1998) has been: how do you measure entities on an interval scale given merely nominal or ordinal information?

Scaling Dichotomous Items

Nominal data are often dichotomous yes–no answers to questions, a judge's presence–absence judgments about the features of entities, an expert's claims about truth–falsity of propositions, etc. Answers of *yes, present, true*, etc. typically are coded as 'one' and *no, absent, false*, etc. as 'zero.' The goal, then, is to translate zero–one answers for each case in a sample of entities into a number representing the case's position on an interval scale of measurement.

The first step requires identifying how items relate to the interval scale of measurement in terms of a graph of the items' characteristic curves. The horizontal axis of such a graph is the interval scale of measurement, confined to the practical range of variation of entities actually being observed. The vertical axis indicates probability that a specific dichotomous item has the value one for an entity with a given position on the interval scale of measurement. An item's characteristic curve traces the changing probability of the item having the value one as an entity moves from having a minimal value on the interval scale to having the maximal value on the interval scale.

Item characteristic curves have essentially three different shapes corresponding to three different formulations about how items combine into a scale.

Spanning items have characteristic curves that essentially are straight lines stretching across the range of entity variation. A spanning item's line may start as a low probability value and rise to a high probability value, or fall from a high value to a low value. A rising line means that the item is unlikely to have a value of one with entities having a low score on the interval scale; the item is likely to have a score of one for entities having a high score on the interval scale; and the probability of the item being valued at one increases regularly for entities between the low and high positions on the scale.

Knowing an entity's value on any one spanning item does not permit assessing the entity's position along the interval scale. However, knowing the entity's values on multiple spanning items does allow an estimate of positioning to be made. Suppose heuristically that we are working with a large number of equivalent spanning items, each having an item characteristic curve that starts at probability 0.00 at the minimal point of the interval scale, and rises in a straight line to probability 1.00 at the maximal point of the interval scale. The probability of an item being valued at one can be estimated from the observed proportion of all these items that are valued at one, which is simply the mean item score when items are scored zero–one.

Then we can use the characteristic curve for the items to find the point on the interval scale where the entity must be positioned in order to have the estimated item probability. This is the basic scheme involved in construction of *composite scales*, where an averaged or summated score on multiple items is used to estimate an entity's interval-scale value on a dimension of interest (Lord and Novick, 1968; Martinez-Martin, 2010).

The more items that are averaged, the better the estimate of an entity's position on the interval scale. The upper bound on number of items is pragmatic, determined by how much precision is needed and how much it costs to collect data with more items. The quality of the estimate also depends on how ideal the items are in terms of having straight-line characteristic curves terminating at the extremes of probability. Irrelevant items with a flat characteristic curve would not yield an estimate of scale position no matter how many of them are averaged, because a flat curve means that the probability of the item having a value of one is uncorrelated with the entity's position on the interval scale. Inferences are possible with scales that include relevant but imperfect items, but more items are required to achieve a given level of precision, and greater weight needs to be given to the more perfect items.

Declivitous items have characteristic curves that rise sharply at a particular point on the horizontal axis. Idealized, the probability of the item having a value of one increases from 0.00 to the left of the inflection point to 1.00 to the right of the inflection point; or alternatively the probability declines from 1.00 to 0.00 in passing the inflection point. Realistically, the characteristic curve of a declivitous item is S-shaped with a steep rise in the middle and graduated approaches to 0.00 at the bottom and to 1.00 at the top.

The value of a single declivitous item tells little about an entity's position along the interval scale. However, an inference about an entity's scale position can be made from a set of declivitous items with different inflection points, or difficulties that form a *cumulative scale*. Suppose heuristically that each item increases stepwise at its inflection point. Then for an entity midway along the horizontal axis, items at the left end of the scale will all have the value of one, items at the right end of the scale will all have the value zero, and the entity's value on the interval scale is between the items with a score of one and the items with a score of zero.

If the items' inflection points are evenly distributed along the interval scale, then the sum of items' zero–one scores for an entity constitutes an estimate of where the entity is positioned along the interval scale. That is, few of the items have a value of one if the entity is on the lower end of the interval scale, and many of the items are valued at one if the entity is at the upper end of the interval scale. This is the basic scheme involved in Guttman scalogram analysis (e.g., see Shye, 1978: Part 5). On the other hand, we might use empirical data to estimate the position of each item's inflection point on the interval scale, while simultaneously estimating entity scores that take account of the item difficulties. This is the basic scheme involved in scaling with Rasch models (Andrich, 1988; Engelhard, 2008).

Entities' positions on the interval scale can be pinned down as closely as desired through the use of more declivitous items with inflection points spaced closer and closer together. However, adding items to achieve more measurement precision at the low end of the interval scale does not help at the middle or the high end of the interval scale. Thus, obtaining high precision over the entire range of variation requires a large number of items, and it could be costly to obtain so much data. Alternatively, one can seek items whose characteristic curves rise gradually over a range of the interval scale such that sequential items on the scale have overlapping characteristic curves, whereby an entity's position along the interval scale is indicated by several items.

Regional items have characteristic curves that rise and fall within a limited range of the interval scale. That is, moving an entity up the interval scale increases the probability of a particular item having a value of one for a while, and then decreases the probability after the entity passes the characteristic curve's maximum value. For example, in a scale measuring prejudice toward a particular ethnic group, the probability of agreeing with the item 'they require equal but separate facilities' increases as a person moves away from an apartheid position, and then decreases as the person moves further up the scale toward a nondiscriminatory position. A regional item's characteristic curve is approximately bell-shaped if its maximum is at the middle of the interval scale, but characteristic curves at the ends of the scale are subject to floor and ceiling clipping, making them look like declivitous items.

If an entity has a value of one on a regional item, then the entity's position along the interval scale is known approximately, since the entity must be positioned in the part of the scale where that item has a nonzero probability of being valued at one. However, a value of zero on the same item can result from a variety of positions along the interval scale and reveals little about the entity's position. Thus, multiple regional items have to be used to assess positions along the whole range of the scale. The items have to be sequenced relatively closely along the scale with overlapping characteristic curves so that no entity will end up in the noninformative state of having a zero value on all items.

One could ask judges to rate the scale position of each item, and average across judges to get an item score; then, later, respondents can be scored with the average of the items they accept. This simplistic approach to regional items was employed in some early attempts to measure social attitudes. Another approach is statistical unfolding of respondents' choices of items on either side of their own positions on the interval scale in order to estimate scale values for items and respondents simultaneously (Coombs, 1964).

Item analysis is a routine aspect of scale construction with spanning items, declivitous items, or regional items. One typically starts with a notion of what one wants to measure, assembles items that should relate to the dimension, and tests the items in order to select the items that work best. Since a criterion measurement that can be used for assessing item quality typically is lacking, the items as a group are assumed to measure what they are supposed to measure, and scores based on this assumption are used to evaluate individual items.

Items in a scale are presumed to measure a single dimension rather than multiple dimensions. Examining the dimensionality assumption brings in additional technology, such as component analysis or factor analysis in the case of spanning items, multidimensional scalogram analysis in the case of declivitous items, and nonmetric multidimensional scaling in the case of regional items. These statistical methods help in

refining the conception of the focal dimension and in selecting the best items for measuring that dimension.

Ordered Assessments

Ordinal data, the starting point for a three-volume mathematical treatise on measurement theory (Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990), may arise from individuals' preferences, gradings of agreement with statements, reckonings of similarity between stimuli, etc. Conjoint analysis (Luce and Tukey, 1964; Michell, 1990; Green et al., 2001) offers a general mathematical model for analyzing such data. According to the conjoint theory of measurement, positions on any viable quantitative dimension are predictable from positions on two other quantitative dimensions, and this assumption leads to tests of a dimension's usefulness given just information of an ordinal nature. For example, societies might be ranked in terms of their socioeconomic development and also arrayed in terms of the extents of their patrifocal technologies (like herding) and matrifocal technologies (like horticulture), each of which contributes additively to socioeconomic development. Conjoint analyses could be conducted to test the meaningfulness of these dimensions, preliminarily to developing interval scales of socioeconomic development and patrifocal and matrifocal technologies. Specific scaling methodologies, like Rasch scaling and nonmetric multidimensional scaling, can be interpreted within the conjoint analysis framework.

Magnitude estimations involve comparisons to an anchor, for example: "Here is a reference sound. … How loud is this next sound relative to the reference sound." Trained judges using such a procedure can assess intensities of sensations and of a variety of social opinions on ratio scales (Stevens, 1951; Lodge, 1981). Comparing magnitude estimation in social surveys to the more common procedure of obtaining ratings on category scales with a fixed number of options, Lodge (1981) found that magnitude estimations are more costly but more accurate, especially in registering extreme positions.

Rating scales with bipolar adjective anchors like good–bad are often used to assess affective meanings of perceptions, individuals, events, etc. Such scales traditionally provided seven or nine answer positions between the opposing poles with the middle position defined as neutral. Computerized presentations of such scales with hundreds of rating positions along a graphic line yield greater precision by incorporating some aspects of magnitude estimation. Cross-cultural and cross-linguistic research in dozens of societies has demonstrated that bipolar rating scales align with three dimensions – evaluation, potency, and activity (Osgood et al., 1975). An implication is that research employing bipolar rating scales should include scales measuring the standard three dimensions in order to identify contributions of these dimensions to rating variance on other bipolar scales.

Measurements and Theory

Theories and measurements are bound inextricably. In the first place, taxonomies and typologies, which are theoretical constructions, even when rooted in folk classification systems, are entailed in defining which entities are to be measured, so are part of all measurements.

Second, scientists routinely assume that any particular measurement is wrong to some degree, even a measurement based on a scale, and that combining multiple measurements improves measurement precision. The underlying premise is that a true value exists for that which is being measured, as opposed to observed values, and that theories apply to true values, not ephemeral observed values. In essence, theories set expectations about what should be observed, in contrast to what is observed, and deviation from theoretical expectations is interpreted as measurement error (Kyburg, 1992). A notion of measurement error as deviation from theoretical expectations is widely applicable, even in qualitative research (McCullagh and Behan, 1984; Heise, 1989).

Third, the conjoint theory of measurement underlying many current measurement technologies requires theoretical specification of relations between variables before the variables' viability can be tested or meaningful scales constructed. This approach is in creative tension with traditional deductive science, wherein variable measurements are gathered in order to determine if relations among variables exist and theories are correct. In fact, a frequent theme in social science during the late twentieth century was that measurement technology had to improve in order to foster the growth of more powerful theories. It is clear now that the dependence between measurements and theories is more bidirectional than was supposed before the development of conjoint theory.

See also: Best Worst Scaling: Theory and Practice; Factor Analysis and Latent Structure Analysis: Overview; Factor Analysis and Latent Structure: IRT and Rasch Models; Measurement Error Models; Measurement Theory: Conjoint; Measurement Theory: History and Philosophy; Measurement, Representational Theory of; Multidimensional Scaling I; Multidimensional Scaling II; Reliability: Measurement; Sensation and Perception: Direct Scaling.

Bibliography

Andrich, D., 1988. Rasch Models for Measurement. Sage, Newbury Park, CA.
Coombs, C.H., 1964. A Theory of Data. John Wiley & Sons, New York.
D'Andrade, R., 1995. The Development of Cognitive Anthropology. Cambridge University Press, New York.
Engelhard Jr., G., 2008. Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. Measurement 6, 155–189.
Green, P.E., Krieger, A.M., Wind, Y., 2001. Thirty years of conjoint analysis: reflections and prospects. Interfaces 31, S56–S73.
Heise, D.R., 1989. Modeling event structures. Journal of Mathematical Sociology 14, 139–169.
Hopkins, K.D., 1998. Educational and Psychological Measurement and Evaluation, eighth ed. Allyn and Bacon, Boston.
Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A., 1971. Foundations of measurement. In: Additive and Polynomial Representations, vol. 1. Academic Press, New York.
Kyburg Jr., H.E., 1992. Measuring errors of measurement. In: Savage, C.W., Ehrlich, P. (Eds.), Philosophical and Foundational Issues in Measurement. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 75–91.
Lodge, M., 1981. Magnitude Scaling: Quantitative Measurement of Opinions. Sage, Beverly Hills, CA.
Lord, F.N., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Luce, R.D., Krantz, D.H., Suppes, P., Tversky, A., 1990. Foundations of measurement. In: Representation, Axiomatization, and Invariance, vol. 3. Academic Press, New York.

Luce, R.D., Tukey, J.W., 1964. Simultaneous conjoint measurement: a new type of fundamental measurement. Journal of Mathematical Psychology 1, 1–27.

Martinez-Martin, P., 2010. Composite rating scales. Journal of the Neurological Sciences 289, 7–11.

McCullagh, B.C., Behan, C., 1984. Justifying Historical Descriptions. Cambridge University Press, New York.

Michell, J., 1990. An Introduction to the Logic of Psychological Measurement. Lawrence Erlbaum Associates, Hillsdale, NJ.

Osgood, C., May, W.H., Miron, M.S., 1975. Cross-cultural Universals of Affective Meaning. University of Illinois Press, Urbana, IL.

Romney, A.K., Weller, S.C., Batchelder, W.H., 1986. Culture as consensus: a theory of culture and informant accuracy. American Anthropologist 88, 313–338.

Shye, S. (Ed.), 1978. Theory Construction and Data Analysis in the Behavioral Sciences, first ed. Jossey-Bass, San Francisco, CA.

Shye, S., Elizur, D., Hoffman, M., 1994. Introduction to Facet Theory: Content Design and Intrinsic Data Analysis in Behavioral Research. Sage, Thousand Oaks, CA.

Stevens, S.S., 1951. Mathematics, measurement, and psychophysics. In: Stevens, S.S. (Ed.), Handbook of Experimental Psychology. John Wiley and Sons, New York, pp. 1–49.

Suppes, P., Krantz, D.H., Luce, R.D., Tversky, A., 1989. Foundations of measurement. In: Geometrical, Threshold, and Probabilistic Representations, vol. 2. Academic Press, New York.

Thomas, L., Heise, D.R., 1995. Mining error variance and hitting pay-dirt: discovering systematic variation in social sentiments. The Sociological Quarterly 36, 425–439.

Watts, D.J., 2004. The "new" science of networks. Annual Review of Sociology 30, 243–270.

Schizophrenia

Jürgen Zielasek and Wolfgang Gaebel, Heinrich Heine University, Düsseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Warner, volume 20, pp. 13530–13537, © 2001, Elsevier Ltd.

Abstract

Schizophrenia is a severe mental disorder characterized by a combination of symptoms including delusions, hallucinations, and disorganized thinking with disease onset in early adulthood. Diagnostic criteria include the typical clinical symptoms, a minimum duration of symptoms, and the exclusion of other causes of psychotic disorders like brain inflammation or the use of illicit drugs. While some persons with schizophrenia have a mild disease course with single episodes, the majority will have a chronic disease course with repeated episodes and/or increasing functional impairments. Treatment is mainly with antipsychotic medication. Manualized psychotherapeutic approaches are becoming available.

Definitions

Schizophrenia is defined as a mental disorder characterized by psychotic symptoms (delusions, hallucinations, thought disorder) of a certain duration occurring in the absence of other recognizable causes of psychotic disorders like several general somatic disorders, substance use, or withdrawal of substances of addiction. As schizophrenia occurs in the absence of other known general or mental disorders that may lead to psychotic symptoms, schizophrenia is considered a 'primary' psychotic disorder. The psychotic disorders occurring due to brain or general somatic disorders are 'secondary' psychotic disorders. An example of the latter would be the occurrence of psychotic symptoms in a patient with epilepsy or in a state of delirium, in which the psychotic symptoms remit once the underlying general medical condition is normalized. Note that the definition used here to denote 'psychosis' (the occurrence of delusions, hallucinations, and/or a thought disorder) is a rather narrow definition. The concept of 'psychosis' has undergone considerable changes over time, but its core symptoms are the above-mentioned symptoms regardless of their causes (Beer, 1996).

Intellectual Backgrounds, Including Schools of Thought and Leading Advocates

The concept of a dichotomy between primary psychotic disorders having a poor prognosis leading to rapid clinical deterioration ('dementia praecox,' where 'dementia' does not indicate the dementia as outlined in the modern classification systems of brain disorders, but rather a mental disorder with severe clinical symptoms and a poor prognosis) and those with a good prognosis (manic-depressive illnesses) had been introduced by Emil Kraepelin (Palm and Möller, 2011). The Swiss psychiatrist E. Bleuler reformulated Kraepelin's concept of 'dementia praecox' and introduced the term 'schizophrenia' (Bleuler, 1908). The clinical picture of schizophrenia in Bleuler's sense was characterized by four basic symptoms, i.e., formal thought disorder, affect disorder, a lack of drive, and autism. Bleuler also added accessory symptoms, i.e., delusions, hallucinations, and catatonia, the last of these being a state of severe psychomotor disturbances characterized, among other clinical features, by an increased muscle tone. Kurt Schneider in 1938 revised the hierarchical order of these symptoms by differentiating first-rank symptoms and second-rank symptoms (Table 1).

These three schools of thinking (Kraepelin, Bleuler, Schneider) were the major influences when the modern classification criteria for mental disorders were developed: for the International Classification of Mental Disorders (ICD) published by the World Health Organization (WHO) (now in its 10th revision, hence ICD-10; World Health Organization, 1992) and the Diagnostic and Statistical Manual (DSM) published by the American Psychiatric Association (now in its fifth revision, hence DSM-5; American Psychiatric Association, 2013). These classification systems retained the dichotomy of schizophrenia-like and manic-depressive psychotic illnesses.

Regarding nomenclature, the term 'schizophrenia' has been substituted in several countries due to its stigmatizing connotations (e.g., now 'integration disorder' in Japan; Sato, 2006). In Japan, this renaming was accompanied by a paradigm shift in the concept of the disorder toward a manageable – as opposed to the previous concept of an intractable – mental disorder. There are two major issues concerning the concept of schizophrenia, which have been debated since the initial concept was introduced by Bleuler: (1), is schizophrenia

Table 1 First-rank symptoms of schizophrenia according to K. Schneider (English translation: 1959)

- Auditory hallucinations
  - Voices heard arguing or giving instructions (second person)
  - Voices heard commenting on one's actions (running commentary, usually describing the patient in the third person)
- Somatic/thought passivity experiences (delusions of control/of being controlled)
- Thought echo (thoughts being heard aloud)
- Thought withdrawal
- Thought insertion (thoughts are ascribed to other people who are intruding into the patient's mind)
- Thought broadcasting (also called thought diffusion)
- Delusional perception (linking a normal sensory perception to a bizarre conclusion, e.g., seeing an airplane means the patient is the president)

a nosologically distinct mental disorder or rather a collection of different mental disorders with similar clinical and course characteristics (the latter defining a 'group of schizophrenias') and (2) what are the causes of schizophrenia. Since the latter is the question underlying the first issue, the authors attend to the question in some detail in the following article.

Current Knowledge

Currently, the pathophysiology of schizophrenia is unknown. A diagnosis of schizophrenia can be made if the appropriate psychotic symptoms are present, the necessary duration criterion is fulfilled, and other medical reasons for the occurrence of psychotic symptoms have been excluded. Thus, schizophrenia is a clinical diagnosis arrived at by careful clinical investigation, history taking, and additional diagnostic procedures like neuroimaging if necessary to exclude other medical conditions.

Schizophrenia may take different courses. About one-third of all cases are self-limiting, single episodes. In another one-third of the cases, recurrent episodes are characteristic with intervening complete or incomplete symptom remission. A final third of cases has a chronic progressive disease course with increasing and enduring symptoms and disability, with occasional acute psychotic exacerbations (**Figure 1**). Thus, first-episode schizophrenia patients in one-third of the cases have a comparatively good prognosis, while two-thirds will experience a chronic disease course (Ram et al., 1992). Functional impairments are especially prominent during acute psychotic episodes in all three disease course types due to 'positive' symptoms like delusions or hallucinations. In the chronic progressive course type, 'negative' symptoms like lack of drive or attentional deficits, and cognitive symptoms like memory impairments, become additional causes of functional impairment.

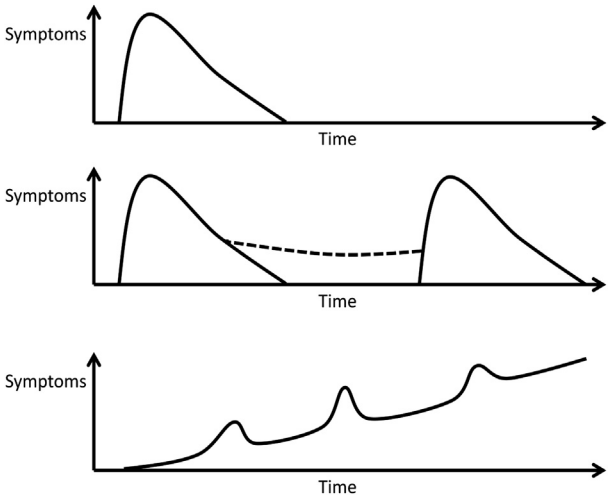

Figure 1 Disease course characteristics of schizophrenia. In schizophrenia, three major disease types can be distinguished. Upper panel: single lifetime episode. Middle panel: repeated episodes with complete symptom remission between episodes. In some cases, there is only incomplete symptomatic remission between episodes (dotted line). Lower panel: chronic progressive schizophrenia with superimposed episodes.

Epidemiology, Illness-Associated Factors and Burden of Disease

Schizophrenia usually begins with unspecific prodromal symptoms in late adolescence, and such prodromal symptoms like a lack of drive or concentration problems may lead to problems in school or in social relationships. Men seem to be at a slightly higher risk for first-episode psychosis and tend to be younger at onset. The diagnosis of schizophrenia is only made once a full clinical psychotic picture has developed. The lifetime prevalence of schizophrenia worldwide is around 1% with little regional variation or change over time (Jablensky, 1997), with prevalence rates using DSM-III and later criteria being somewhat lower due to more restrictive criteria compared to earlier classification criteria (after DSM-III around 0.7%; Bromet and Fennig, 1999). The annual incidence rate is about 0.2/1000. There is a strong aggregation of schizophrenia within affected families, with the highest rates reported for monozygotic twins (around 50%), followed by children with two parents with schizophrenia (46%), dizygotic twins (15%), and children with one parent having schizophrenia (12.8%) (Bromet and Fennig, 1999). There is an epidemiological association between urbanicity and schizophrenia, but it is unknown whether this is a causal association. Other, weaker associations were described for obstetric, birth, and early childhood complications. The disease burden caused by schizophrenia is considerable since approximately two-thirds of patients with schizophrenia have chronically progressive or chronically relapsing-remitting disease course types with ensuing chronic functional impairments. The 1-year functional outcome is favorable in about half of all schizophrenia patients and is associated with a short duration of untreated psychosis, rapid response toward treatment, low total symptom load, younger age, female gender, good premorbid social adjustment, and employment (Spellmann et al., 2012). In persons with chronic schizophrenia, high degrees of mental health care and other social services utilization are observed leading to high societal costs (Ekman et al., 2013; Mangalore and Knapp, 2007). In a European survey of costs for mental and neurological disorders, psychotic disorders (including schizophrenia) were ranked third following dementia and mood disorders (Gustavsson et al., 2011). While hospital costs are a major contributor to the societal costs of schizophrenia, analyses also show that replacing inpatient treatment by outpatient treatment will not reduce total costs (Mangalore and Knapp, 2007) and that non-healthcare-related factors like loss of productivity come into play. A special issue is related to the epidemiology and mental health care of schizophrenia in low-income countries. While the prevalence appears to be similar worldwide, there are indications that the functional outcome is better in low-income countries, which may be due to better family and workplace-related support compared to that in high-income countries (Naheed et al., 2012). However, this viewpoint is currently debated and there are also indication

of undertreatment, neglect, and human rights violations in developing countries warranting further investigations (Patel et al., 2006).

Diagnosis and Classification

Both major international classification systems for mental disorders, DSM-5 and ICD-10, include a category 'schizophrenia.' The diagnostic criteria are summarized in **Table 2** and include lists of symptoms that may be present, exclusion criteria, and duration criteria. Due to the differences in details of the diagnostic criteria in both systems, a person diagnosed with schizophrenia following the guidelines of one classification system may not fulfill the criteria of the other system, or vice versa. Such distinguishing differences are, for example, the different duration criteria and the different symptom lists. Also, ICD-10 does not require social or occupational dysfunction, as DSM-5 does, for establishing a diagnosis of schizophrenia.

The ICD-10 is currently under revision and proposals for revised schizophrenia criteria have been put forward by the WHO working group on psychotic disorders (Gaebel et al., 2013). These would lead to a harmonization of the course criteria with DSM-5, but would retain the different duration criteria. Similar to DSM-5, ICD-11 would deemphasize

Table 2 Comparison of the diagnostic criteria of schizophrenia in DSM-5 (American Psychiatric Association, 2013) and ICD-10 (World Health Organization, 1992)

Criterion	*DSM-5*	*ICD-10*
Clinical symptoms	Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated). At least one of these must be (1), (2), or (3): 1. Delusions 2. Hallucinations 3. Disorganized speech (e.g., frequent derailment or incoherence) 4. Grossly disorganized or catatonic behavior 5. Negative symptoms (i.e., diminished emotional expression or avolition)	One very clear symptom (and usually two or more if less clear-cut) belonging to any one of the following groups listed as (1) to (2), or symptoms from at least two of the groups (5) to (8): 1. Thought echo, thought insertion or withdrawal, and thought broadcasting 2. Delusions of control, influence, or passivity, clearly referred to body or limb movements or specific thoughts, actions, or sensations; delusional perception 3. Hallucinatory voices giving a running commentary on the patient's behavior, or discussing the patient among themselves, or other types of hallucinatory voices coming from some part of the body 4. Persistent delusions of other kinds that are culturally inappropriate and completely impossible, such as religious or political identity, or superhuman powers and abilities (e.g., being able to control the weather, or being in communication with aliens from another world) 5. Persistent hallucinations in any modality, when accompanied either by fleeting or half-formed delusions without clear affective content, or by persistent overvalued ideas, or when occurring every day for weeks or months on end 6. Breaks or interpolations in the train of thought, resulting in incoherence or irrelevant speech, or neologisms 7. Catatonic behavior, such as excitement, posturing, or waxy flexibility, negativism, mutism, and stupor 8. Negative' symptoms such as marked apathy, paucity of speech, and blunting or incongruity of emotional responses, usually resulting in social withdrawal and lowering of social performance; it must be clear that these are not due to depression or neuroleptic medication 9. A significant and consistent change in the overall quality of some aspects of personal behavior, manifest as loss of interest, aimlessness, idleness, a self-absorbed attitude, and social withdrawal
Functional impairments	Social/occupational dysfunction	None required
Duration	6 months	1 month
Exclusion	Schizoaffective and mood disorder exclusion Substance/general medical condition exclusion No relationship to a pervasive developmental disorder	

Schneiderian first-rank symptoms and would introduce severity/symptom specifiers.

Treatment

Treatment is based on the three pillars: pharmacotherapy, psychotherapy, and social therapy. Guidelines for schizophrenia treatment have been developed worldwide, but differ in the details of their therapeutic recommendations reflecting different interpretation of clinical study results, different therapeutic traditions, and different approaches toward mental health care in general (Gaebel et al., 2011). Yet, some treatment principles are common ground: during the acute psychotic phase, antipsychotic medication is usually indicated to shorten the psychotic episode (Hasan et al., 2012). Antipsychotic medication is also effective in preventing future psychotic episodes, so that a certain period of continuing antipsychotic drug therapy after an initial psychotic episode is recommended (Hasan et al., 2013a). This may need to be continued for extended periods of time if repeated episodes occur. Antipsychotic medication is especially effective in the treatment of the 'positive' symptoms of schizophrenia, i.e., hallucinations and delusions, and 90% of patients experience full or partial symptomatic remission of positive symptoms (Schennach et al., 2012). Antipsychotic pharmacotherapy is less effective in treating the 'negative' symptoms, i.e., anhedonia and lack of drive (Gaebel et al., 2014). Blunted affect is the most frequently persisting negative symptom in patients with first-episode schizophrenia and its occurrence is associated with a long duration of untreated psychosis and with treatment cessation (Galderisi et al., 2013). Alternatively or in combination with antipsychotic pharmacotherapy, effective modes of cognitive psychotherapy are now available for schizophrenia patients. Cognitive psychotherapy was shown to be effective for the treatment of positive and negative symptoms, but different psychotherapeutic programs are needed. Cognitive remediation therapy was shown to be effective in the treatment of cognitive dysfunctions in schizophrenia.

One of the controversial issues in schizophrenia treatment is the differential indication for 'first'- or 'second'-generation antipsychotic drugs. While both are equally effective in treating positive symptoms, the 'first-generation antipsychotic drugs' are considered problematic because of their frequent side effects on motor functions like gait. Second-generation antipsychotic drugs are viewed by some as better able to treat negative symptoms, but this opinion is debated. It is currently also debated whether second-generation antipsychotic drugs produce less motor side effects compared to first-generation antipsychotic drugs. These potential advantages of second-generation antipsychotic drugs need to be viewed in the light of increased rates of metabolic side effects in second-generation antipsychotic drugs, e.g., on glucose homeostasis. Guideline recommendations as to a preference for first- or second-generation drugs are divergent and it is therefore recommended to decide on which antipsychotic drug to use in individual persons depending on the availability of each drug in the respective country, patient preferences, side effect profiles, and practitioner's familiarity and experiences with each drug.

Different stages of the disease may demand different treatment strategies. In the first episode of schizophrenia, initial antipsychotic treatment accompanied by psychotherapy and sociotherapy, the latter involving rehabilitation to gain or regain education and/or employment, will usually be indicated. Methods of art therapy, sports therapy, and specially structured rehabilitation programs may also be added. It is recommended to continue antipsychotic therapy for at least 2 years following the initial psychotic episode. If new episodes occur, a relapsing disease course must be assumed and longer periods of prophylactic antipsychotic medications are warranted, e.g., for 5 years or even lifelong. In such cases, depot drug formulations may be warranted. However, the question of maintenance therapy is still controversial in the guidelines regarding the necessary duration and strategies for changing antipsychotic therapy in treatment-refractory cases (Takeuchi et al., 2012). If negative symptoms become dominant, especially in the chronic progressive forms of schizophrenia, psychotherapy and sociotherapy take centerstage because the negative symptoms would not be expected to respond well to antipsychotic pharmacotherapy. However, the latter should be continued with a view to prevent acute psychotic exacerbations. Little is known about schizophrenia of persons with greater ages. In the approximately 30% of cases with chronic symptoms, negative symptoms usually prevail and may be difficult to distinguish from cognitive impairments, which may be the consequence of higher age or incipient dementing illnesses. Another open question is whether treatment guidelines developed for early-onset patients are also valid for late-onset patients, i.e., disease onset after 40 years of age, who may constitute up to approximately 25% of all schizophrenia cases. Few studies have addressed this issue and currently, no recommendations based on large-scale studies can be made (Essali and Ali, 2012). Another important issue is the availability of medication in low-income countries (Naheed et al., 2012).

Prevention

Currently, there are no known effective preventive measures against the development of schizophrenia in individual persons. The large number of involved genes leading to genetic heterogeneity, pleiotropy, and incomplete penetrance are the major factors explaining why studies with all available genetic methods (genetic polymorphisms, genome-wide association studies, copy number variation studies, studies of epigenetic factors) have so far not been able to identify genes allowing individual predictions (Addington and Rapoport, 2012). However, there is a prodromal phase of months to years with attenuated psychotic symptoms or unspecific symptoms like concentration or affective problems before the full psychotic picture emerges, and such a prodromal period is reported by 80–90% of persons with schizophrenia (Yung and McGorry, 1996). The prodromal phase offers the potential for future preventive measures like pharmacotherapy or psychotherapy. There is, however, no consensus yet about the typical clinical prodromal presentation that indicates the highest risk of incipient development of schizophrenia, since less than half of all persons who experience

prodromal symptoms continue to develop schizophrenia in the future. Research is ongoing to better define the high-risk prodromal clinical picture and to better predict the imminent development of schizophrenia in individual persons. This knowledge would be a prerequisite for employing preventive therapeutic measures, which may have severe side effects. Current pharmacological research aims at establishing safe ways of indicated or selective prevention once a high risk for the development of schizophrenia is detected in individual persons (Addington and Heinssen, 2013). Given the uncertainties surrounding the prodromal phase of schizophrenia, DSM-5 has only defined an 'attenuated psychosis syndrome' as a mental state requiring further research, but not as a full psychotic disorder entity (Tsuang et al., 2013).

Changes in Focus or Emphasis over Time

With the introduction of rigorous clinical research methods in psychiatry and the discovery of the central role of dopamine neurotransmitter alterations in the pathophysiology of schizophrenia, clinical schizophrenia research initially (i.e., in the second half of the past century) was focused on the development of antidopaminergic drugs for the treatment of schizophrenia. This led to the development of the antipsychotic drugs still in use today. From the 1960s to the 1970s, research into social and environmental causes of schizophrenia took centerstage showing associations with prebirth viral infections and urbanicity. Starting in the late twentieth century, research into the role of cognitive dysfunctions in schizophrenia gained considerable momentum following the findings that social cognition and neurocognition play an important role in determining everyday functioning in persons with schizophrenia (Mehta et al., 2013), that cognitive dysfunctions are amenable to treatment (Saperstein and Kurtz, 2013), and that cognitive therapy improves the outcome of work rehabilitation in schizophrenia (Kurzban et al., 2010). The turn of the century was characterized by increased research into the genetic associations of schizophrenia, leading not only to the discovery of a plethora of genetic alterations in persons with schizophrenia but also to the result that these were not able to fully explain the clinical picture. However, many of the genes shown to be associated with schizophrenia were identified as genes of neural structures or functions, providing new impetus to cranial neuroimaging and neurophysiological investigations in schizophrenia research. Reevaluating the available evidence, the DSM-5 work group on psychotic disorders reasoned to abolish the traditional schizophrenia subtypes (paranoid, hebephrenic, and catatonic) and to deemphasize Schneiderian first-rank symptoms by substituting them with symptom specifiers including symptoms for positive and negative symptoms and by introducing a catatonia specifier (Tandon et al., 2013a,b).

Current and Emerging Directions in Theory and Research

The interindividual phenotypic variability of the psychopathological picture of schizophrenia, the multitude of genes with only minor contributions from single genes or genetic polymorphisms to this variability, and the complexity of the clinical picture are major challenges to current schizophrenia research. These conceptual challenges also lead to considerable problems in developing appropriate animal models of psychotic disorders, beyond the general conceptual problem of ascertaining delusional or hallucinatory states in experimental animals (O'Tuathaigh et al., 2013). There is an increasing research emphasis on the biological mechanisms of the pathophysiology of schizophrenia addressing several key questions: (1) which are the neural pathways involved in the pathophysiology of schizophrenia, (2) how can environmental factors exert an influence on disease development, and (3) which different biopsychosocial mechanisms interact in the pathophysiology of schizophrenia. As regards the identification of neural pathways involved in the pathophysiology of schizophrenia, several lines of investigation have shown that there are changes in the complex structure of brain neural circuits, which may be due to dysmyelinating changes of brain fiber tracts. Considering the role of environmental influences, current research focuses on identifying epigenetic effects on the genome of affected individuals and showing how early life adverse events may carry over neurobiological effects into adolescence and adulthood. Regarding the manifold biological mechanisms, it is currently being investigated whether symptom- and not diagnosis-centered research may help to identify common pathophysiologic factors for similar symptoms occurring in different mental disorders, with a view to identify unifying pathophysiological principles.

It is becoming increasingly appreciated that the psychotic symptoms occurring in schizophrenia are not readily explained by neurobiological (neurophysiological or neurogenetic and other) mechanisms. For example, in hallucination research, novel interest focuses on the psychopathological components of a hallucinatory experience (e.g., location of hallucinations, loudness, linguistic complexity), how each of these components is determined by different neural functions, and how these functions or their interplay are disturbed in schizophrenia (Stephane, 2013). Another approach is to link genomic aberrations in at-risk persons with nongenomic (environmental) influences to demonstrate ensuing aberrant gene expression and neurobiological alterations, as, for example, excess glutamate production leading to overproduction of (psychotogenic) brain dopamine and brain structural alterations in the hippocampus of at-risk persons becoming psychotic if the situation is not ameliorated (Moghaddam, 2013). Importantly, initial research shows that the pathophysiological factors determining the transition from 'at risk' to 'schizophrenia' may be different from those sustaining chronic positive or negative symptoms (Moghaddam, 2013), leading to the concept that stage-specific pathophysiologic mechanisms may make stage-specific interventions necessary. For example, based on the findings that antidopaminergic drugs may increase glutamate release, these drugs may not be the best choice in the prodromal stages, in which research indicates that excess glutamate is a key driver of the pathophysiologic process driving the progression from prodromal to psychotic states. Also, antiinflammatory drugs like omega-3-fatty acids may be more effective in the prodromal than in the later psychotic disease stages

(Mossaheb et al., 2013). Other approaches aim at correcting disturbed neural circuitry with electric or magnetic brain stimulation techniques, but these have so far only had limited success (Hasan et al., 2013b).

Research from neuroimaging and neurophysiology using electroencephalography and highly sophisticated modeling approaches shows that there are complex disturbances of the hierarchical and temporal order of neural functions once schizophrenia has fully developed. Therefore, a new research initiative by the National Institutes of Mental Health is paradigmatic for the new neural-systems-centered, integrative approach toward explanations of the pathophysiology of mental disorders in general and of schizophrenia in particular: The Research Domain Criteria (RDoC) initiative aims to identify the disturbed neural pathways, establish assessment methods, and explain the relationship of neural dysfunctions with psychopathology (Cuthbert and Insel, 2013). Based on the many neurobiological findings in schizophrenia research, novel therapeutic and diagnostic strategies will become available, but probably not in the near future. Integrative models of schizophrenia pathophysiology are warranted to reflect and guide such future advances.

Based on such insights into the pathophysiology of schizophrenia, novel drugs are being developed, for example, drugs targeting nicotinic acetylcholine receptors (Wallace and Porter, 2011) or the neurotransmitter glutamate (Kinon and Gómez, 2013). However, drug development is slow in schizophrenia given the complex longitudinal dynamics of individual disease course patterns. This also pertains to invasive therapeutic approaches like psychosurgery or deep brain stimulation, which have been successful in other mental disorders, and which may offer benefits for the future of schizophrenia treatment (Soares et al., 2013).

Another important focus for future research will be the role of stigmatization in schizophrenia. A recent systematic review showed that more than half of all persons with schizophrenia perceive or experience stigmatization (Gerlinger et al., 2013). Stigmatization also affects psychiatry and psychiatrists (Sartorius et al., 2010). A major challenge for the future will be to define the best ways to reduce and possibly eliminate the stigmatization of persons with schizophrenia and the persons and institutions caring for persons with schizophrenia, since finding a cure for schizophrenia will only become possible once the pathophysiology of this complex and severe mental disorder has been elucidated.

See also: Bipolar (Manic-Depressive) Disorder; Differential Diagnosis in Psychiatry; Mental Disorders, Epidemiology of; Mental Illness, Etiology of; Mental Illness, Genetics of; Mentally Ill, Public Attitudes towards; Nosology in Psychiatry; Psychiatric Assessment: Negative Symptoms; Schizophrenia, Treatment of; Schizophrenia: Metabolic Consequences; Social Intervention in Psychiatric Disorders; Stigma, Social Psychology of.

Bibliography

Addington, A.M., Rapoport, J.L., May 2012. Annual research review: impact of advances in genetics in understanding developmental psychopathology. Journal of Child Psychology and Psychiatry 53 (5), 510–518. http://dx.doi.org/10.1111/j.1469-7610.2011.02478.x [Epub 8 November 2011].

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders. DSM-5, fifth ed. American Psychiatric Publishing, Washington, London.

Beer, M.D., March 1996. The endogenous psychoses: a conceptual history. History of Psychiatry 7 (25), 1–29.

Bleuler, E., 1908. Die Prognose der Dementia praecox (Schizophreniegruppe). Allg Zeitschr Psychiatrie psychischgerichtl Med 65, 436–464.

Bromet, E.J., Fennig, S., October 1999. Epidemiology and natural history of schizophrenia. Biological Psychiatry 46 (7), 871–881.

Cuthbert, B.N., Insel, T.R., 14 May 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Medicine 11, 126. http://dx.doi.org/10.1186/1741-7015-11-126.

Ekman, M., Granstrom, O., Omerov, S., Jacob, J., Landen, M., March 2013. The societal cost of schizophrenia in Sweden. Journal of Mental Health Policy and Economics 16 (1), 13–25.

Essali, A., Ali, G., 15 February 2012. Antipsychotic drug treatment for elderly people with late-onset schizophrenia. Cochrane Database Systematic Review 2, CD004162. http://dx.doi.org/10.1002/14651858.CD004162.pub2.

Gaebel, W., Riesbeck, M., Wobrock, T., August 2011. Schizophrenia guidelines across the world: a selective review and comparison. International Review of Psychiatry 23 (4), 379–387. http://dx.doi.org/10.3109/09540261.2011.606801.

Gaebel, W., Riesbeck, M., Wölwer, W., Klimke, A., Eickhoff, M., von Wilmsdorff, M., Heuser, I., Maier, W., Klosterkötter, J., Falkai, P., Schlösser, R., Schmitt, A., Riedel, M., Klingberg, S., Köpcke, W., Ohmann, C., Möller, H.J., February 2014. Rates and predictors of remission in first-episode schizophrenia within 1 year of antipsychotic maintenance treatment. Results of a randomized controlled trial within the German Research Network on Schizophrenia. Schizophrenia Research 152 (2–3), 478–486.

Gaebel, W., Zielasek, J., Cleveland, H.R., June 2013. Psychotic disorders in ICD-11. Asian Journal of Psychiatry 6 (3), 263–265.

Galderisi, S., Mucci, A., Bitter, I., Libiger, J., Bucci, P., Fleischhacker, W.W., Kahn, R.S., March 2013. Eufest Study Group. Persistent negative symptoms in first episode patients with schizophrenia: results from the European First Episode Schizophrenia Trial. European Neuropsychopharmacology 23 (3), 196–204.

Gerlinger, G., Hauser, M., De Hert, M., Lacluyse, K., Wampers, M., Correll, C.U., June 2013. Personal stigma in schizophrenia spectrum disorders: a systematic review of prevalence rates, correlates, impact and interventions. World Psychiatry 12 (2), 155–164. http://dx.doi.org/10.1002/wps.20040.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D.H., Jennum, P., Jordanova, A., Jönsson, L., Karampampa, K., Knapp, M., Kobelt, G., Kurth, T., Lieb, R., Linde, M., Ljungcrantz, C., Maercker, A., Melin, B., Moscarelli, M., Musayev, A., Norwood, F., Preisig, M., Pugliatti, M., Rehm, J., Salvador-Carulla, L., Schlehofer, B., Simon, R., Steinhausen, H.C., Stovner, L.J., Vallat, J.M., Van den Bergh, P., van Os, J., Vos, P., Xu, W., Wittchen, H.U., Jönsson, B., Olesen, J., October 2011. CDBE2010 Study Group. Cost of disorders of the brain in Europe 2010. European Neuropsychopharmacology 21 (10), 718–779.

Hasan, A., Falkai, P., Wobrock, T., Lieberman, J., Glenthoj, B., Gattaz, W.F., Thibaut, F., Möller, H.J., July 2012. World Federation of Societies of Biological Psychiatry (WFSBP) Task Force on Treatment Guidelines for Schizophrenia. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Schizophrenia, part 1: update 2012 on the acute treatment of schizophrenia and the management of treatment resistance. World Journal of Biological Psychiatry 13 (5), 318–378.

Hasan, A., Falkai, P., Wobrock, T., Lieberman, J., Glenthoj, B., Gattaz, W.F., Thibaut, F., Möller, H.J., February 2013a. WFSBP Task force on Treatment Guidelines for Schizophrenia. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, part 2: update 2012 on the long-term treatment of schizophrenia and management of antipsychotic-induced side effects. World Journal of Biological Psychiatry 14 (1), 2–44.

Hasan, A., Falkai, P., Wobrock, T., 2013b. Transcranial brain stimulation in schizophrenia: targeting cortical excitability, connectivity and plasticity. Current Medicinal Chemistry 20 (3), 405–413.

Jablensky, A., December 19, 1997. The 100-year epidemiology of schizophrenia. Schizophrenia Research 28 (2–3), 111–125.

Kinon, B.J., Gómez, J.C., March 2013. Clinical development of pomaglumetad methionil: a non-dopaminergic treatment for schizophrenia. Neuropharmacology 66, 82–86.

Klosterkötter, J., Heuser, I., Maier, W., Lemke, M.R., Rüther, E., Klingberg, S., Gastpar, M., Möller, H.J., August 15, 2012. One-year functional outcomes of naturalistically treated patients with schizophrenia. Psychiatry Research 198 (3), 378–385.

Kurzban, S., Davis, L., Brekke, J.S., August 2010. Vocational, social, and cognitive rehabilitation for individuals diagnosed with schizophrenia: a review of recent research and trends. Current Psychiatry Reports 12 (4), 345–355.

Mangalore, R., Knapp, M., March 2007. Cost of schizophrenia in England. Journal of Mental Health Policy and Economics 10 (1), 23–41.

Mehta, U.M., Thirthalli, J., Subbakrishna, D.K., Gangadhar, B.N., Eack, S.M., Keshavan, M.S., August 2013. Social and neuro-cognition as distinct cognitive factors in schizophrenia: a systematic review. Schizophrenia Research 148 (1–3), 3–11.

Moghaddam, B., April 10, 2013. A mechanistic approach to preventing schizophrenia in at-risk individuals. Neuron 78 (1), 1–3.

Mossaheb, N., Schäfer, M.R., Schlögelhofer, M., Klier, C.M., Cotton, S.M., McGorry, P.D., Amminger, G.P., August 2013. Effect of omega-3 fatty acids for indicated prevention of young patients at risk for psychosis: when do they begin to be effective? Schizophrenia Resea 148 (1–3), 163–167. pii: S0920-9964(13)00282-X. doi:10.1016/j.schres.2013.05.027. [Epub ahead of print].

O'Tuathaigh, C.M., Moran, P.M., Waddington, J.L., October 2013. Genetic models of schizophrenia and related psychotic disorders: progress and pitfalls across the methodological "minefield". Cell & Tissue Research 354 (1), 247–257.

Palm, U., Möller, H.J., June 2011. Reception of Kraepelin's ideas 1900-1960. Psychiatry and Clinical Neuroscience 65 (4), 318–325.

Patel, V., Cohen, A., Thara, R., Gureje, O., June 2006. Is the outcome of schizophrenia really better in developing countries? Revista Brasileira de Psiquiatria 28 (2), 149–152.

Ram, R., Bromet, E.J., Eaton, W.W., Pato, C., Schwartz, J.E., 1992. The natural course of schizophrenia: a review of first-admission studies. Schizophrenia Bulletin 18 (2), 185–207.

Saperstein, A.M., Kurtz, M.M., June 2013. Current trends in the empirical study of cognitive remediation for schizophrenia. Canadian Journal of Psychiatry 58 (6), 311–318.

Sartorius, N., Gaebel, W., Cleveland, H.R., Stuart, H., Akiyama, T., Arboleda-Flórez, J., Baumann, A.E., Gureje, O., Jorge, M.R., Kastrup, M., Suzuki, Y., Tasman, A., October 2010. WPA guidance on how to combat stigmatization of psychiatry and psychiatrists. World Psychiatry 9 (3), 131–144.

Sato, M., February 2006. Renaming schizophrenia: a Japanese perspective. World Psychiatry 5 (1), 53–55.

Schennach, R., Meyer, S., Seemüller, F., Jäger, M., Schmauss, M., Laux, G., Pfeiffer, H., Naber, D., Schmidt, L.G., Gaebel, W., Klosterkötter, J., Heuser, I., Maier, W., Lemke, M.R., Rüther, E., Klingberg, S., Gastpar, M., Musil, R., Möller, H.J., Riedel, M., August 2012. Response trajectories in "real-world" naturalistically treated schizophrenia patients. Schizophrenia Research 139 (1–3), 218–224.

Schneider, K., 1959. Clinical Psychopathology. Grune and Stratton, New York.

Soares, M.S., Paiva, W.S., Guertzenstein, E.Z., Amorim, R.L., Bernardo, L.S., Pereira, J.F., Fonoff, E.T., Teixeira, M.J., 2013. Psychosurgery for schizophrenia: history and perspectives. Neuropsychiatric Disease and Treatment 9, 509–515.

Spellmann, I., Riedel, M., Schennach, R., Seemüller, F., Obermeier, M., Musil, R., Jäger, M., Schmauß, M., Laux, G., Pfeiffer, H., Naber, D., Schmidt, L.G., Gaebel, W., Klosterkötter, J., Heuser, I., Maier, W., Lemke, M.R., Rüther, E., Klingberg, S., Gastpar, M., Möller, H.J., August 2012. One-year functional outcomes of naturalistically treated patients with schizophrenia. Psychiatry Research 198 (3), 378–385.

Stephane, M., May 2013. Auditory verbal hallucinations result from combinatoric associations of multiple neural events. Frontier in Human Neuroscience 7, 239. http://dx.doi.org/10.3389/fnhum.2013.00239.

Takeuchi, H., Suzuki, T., Uchida, H., Watanabe, K., Mimura, M., February 2012. Antipsychotic treatment for schizophrenia in the maintenance phase: a systematic review of the guidelines and algorithms. Schizophrenia Research 134 (2–3), 219–225.

Tandon, R., Heckers, S., Bustillo, J., Barch, D.M., Gaebel, W., Gur, R.E., Malaspina, D., Owen, M.J., Schultz, S., Tsuang, M., van Os, J., Carpenter, W., June 24, 2013a. Catatonia in DSM-5. Schizophrenia Research pii:S0920-9964(13)00234-X. doi:10.1016/j.schres.2013.04.034.

Tandon, R., Gaebel, W., Barch, D.M., Bustillo, J., Gur, R.E., Heckers, S., Malaspina, D., Owen, M.J., Schultz, S., Tsuang, M., van Os, J., Carpenter, W., June 22, 2013b. Definition and description of schizophrenia in the DSM-5. Schizophrenia Research pii:S0920-9964(13)00283-1. doi:10.1016/j.schres.2013.05.028.

Tsuang, M.T., Van Os, J., Tandon, R., Barch, D.M., Bustillo, J., Gaebel, W., Gur, R.E., Heckers, S., Malaspina, D., Owen, M.J., Schultz, S., Carpenter, W., October 2013. Attenuated psychosis syndrome in DSM-5. Schizophrenia Research 150 (1), 31–35. pii:S0920-9964(13) 00259-4. doi.

Wallace, T.L., Porter, R.H., October 15, 2011. Targeting the nicotinic alpha7 acetylcholine receptor to enhance cognition in disease. Biochemical Pharmacology 82 (8), 891–903.

World Health Organization, 1992. The ICD-10 Classification of Mental and Behavioural Disorders. Clinical Descriptions and Diagnostic Guidelines. World Health Organization, Geneva.

Yung, A.R., McGorry, P.D., 1996. The prodromal phase of first-episode psychosis: past and current conceptualizations. Schizophrenia Bulletin 22 (2), 353–370.

Relevant Websites

http://www.schizophreniaforum.org/ – Schizophrenia Research Forum is a website dedicated to convey information about clinical and research aspects of schizophrenia and related psychotic disorders. It hosts a lively forum where opinion leaders in the field of schizophrenia care and research discuss current topics.

http://www.europsy.net/de/ – This homepage of the European Psychiatric Association contains guidelines on mental healthcare pertinent also to the care of persons with schizophrenia ("EPA guidance"). An additional feature is a position statement on the diagnosis and treatment of cardiovascular disorders in patients with severe mental disorders including schizophrenia.

http://www.nice.org.uk/nicemedia/live/11786/43608/43608.pdf – This is the link to the UK National Institute of Clinical Excellence guideline on schizophrenia treatment.

http://www.guideline.gov/ – This homepage of the U.S. governmental Agency for Healthcare Research and Quality is a good starting point to retrieve international treatment guidelines for schizophrenia using the "search" function and "schizophrenia" as the search term.

http://www.wfsbp.org/ – This homepage of the World Federation of Societies of Biological Psychiatry features two recent guidelines for acute and long-term schizophrenia treatment and a consensus paper on the role of biological markers in schizophrenia research and treatment.

Schizophrenia, Treatment of

Thomas Pollmächer, Max Planck Institute, Munich, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13546–13552, © 2001, Elsevier Ltd.

Abstract

Schizophrenia is a frequent brain disorder (live time prevalence about 1.5%) of unknown origin which tends to have a chronic course and deteriorates numerous complex brain functions including thought, emotion, perception, cognition, and behavior. There is no cure for schizophrenia, but the combined administration of pharmacological and psychosocial interventions yields considerable benefit for the patients affected. The targets of treatment are positive symptoms (e.g., hallucinations and delusions) and negative ones (e.g., affective flattening, anhedonia). Antipsychotic drugs (also called neuroleptic drugs) are particularly effective against negative symptoms, probably by influencing neurotransmitter systems within the central nervous system, such as dopaminergic and serotoninergic systems. Psychosocial interventions, which among others include educational programs, social-skill and problem solving training, and supervised residential living arrangements, improve to some extent negative symptoms and social integration. Recent developments and future directions of treatment include the successful pharmacological targeting of negative symptoms, the use of pharmacogenetic strategies to improve treatment response, and the development of drugs directed toward various physiological systems which are supposed to play a pathophysiological role, such as glutamatergic neurotransmission, the immune system, and phospholipid metabolism.

Introduction

Schizophrenia is a brain disorder of unknown origin which severely deteriorates numerous complex functions of the central nervous system including thought, emotion, perception, cognition, and behavior. Schizophrenia is one of the most debilitating psychiatric disorders. Due to the high lifetime prevalence (about 1.5%), the typical onset in early adulthood, and the strong tendency toward a chronic course, the disorder requires a very high degree of health care provisions. About one-quarter of hospital beds are occupied by schizophrenic patients, and the total costs of treatment are enormous (e.g., US$50 billion per year in the United States). Although there is no cure for schizophrenia, the combined administration of pharmacological and psychosocial interventions considerably improves outcome, enhances quality of life in the patients affected, and enables social integration to a large extent.

Causes and Pathophysiology of Schizophrenia

The causes of schizophrenia are unknown so far. Therefore, the term 'schizophrenia' refers to an empirically defined syndrome characterized by a combination of certain symptoms which occur in a particular temporal pattern. The idea that schizophrenia is a distinct brain disorder is rooted in Emil Kraepelin's concept of dementia praecox (Kraepelin, 1910). This concept emphasized one particular aspect of the disorder: the onset of persistent cognitive disturbances early in life. The term 'schizophrenia' was coined by Eugen Bleuler (1916), who wanted to emphasize the loss of coherence between thought, emotion, and behavior which represents another important feature of the disorder. Bleuler actually spoke about 'schizophrenias,' implying a group of diseases rather than one distinct disease entity. The present diagnostic systems compiled by the World Health Organization (*ICD-10* 1992) and the American Psychiatric Association (*DSM-IV* 1994) distinguish various subtypes of schizophrenia (see **Table 1**) which are classified according to particular symptom combinations, and according to certain aspects of the course and prognosis of the disease. However, these subtypes are defined in a phenomenological manner which neither implies distinct causes for any of the subtypes nor any particular treatment.

Although we do not know individual causes of schizophrenia, we know that the basis for the disorder is an interaction between genetic susceptibility factors and environmental components. According to numerous genetic studies the estimates of hereditability converge to about 80% (Owen and Cardno, 1999). However, although linkage studies have yielded evidence for a number of susceptibility loci on various chromosomes, neither a single gene nor a combination of genes have been discovered so far which are definitively involved. Similarly, a number of environmental factors (e.g., birth and pregnancy complications, viral infections) are likely to play a role, but details of this role remain to be established.

Although the symptoms of schizophrenia clearly demonstrate a severe disturbance of brain functions, only subtle alterations of brain morphology have been detected. The most

Table 1 Subtypes of schizophrenia according to the major classification systems

ICD-10	*DSM-IV*
Paranoid schizophrenia	Schizophrenia, paranoid type
Catatonic schizophrenia	Schizophrenia, catatonic type
Undifferentiated schizophrenia	Schizophrenia, undifferentiated type
Residual schizophrenia	Schizophrenia, residual type
Hebephrenic schizophrenia	—
—	Schizophrenia, disorganized type
Simple schizophrenia	—

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27064-5

stable finding is a slight enlargement of brain ventricles. Recent sophisticated studies combining morphological and functional brain imaging have been suggested to indicate that subtle structural or functional lesions particularly in prefrontal and limbic neural circuits impair the integrity of these circuits (Andreasen et al., 1998). Although it is still a matter of debate whether these lesions are due to a neurodevelopmental or a neurodegenerative process (Lieberman, 1999), it is clear that disturbed functioning of neuronal circuits goes along with altered neurotransmission. Dopamine, serotonin, and glutamate are the neurotransmitters most frequently implicated in the pathophysiology of schizophrenia. However, it is still not clear whether disturbed neurotransmission is the cause or the consequence of the major disease process.

Due to this scantiness of etiological and pathophysiological knowledge, the treatment of patients suffering from schizophrenia is based exclusively on empirical clinical knowledge. As outlined in detail below, the treatment approach is multimodal and based on the symptom patterns present in any individual patient (*see* Mental Illness, Etiology of).

Symptoms of Schizophrenia – the Targets for Treatment

Schizophrenia affects almost all areas of complex brain functions, and thought, perception, emotion, cognition, and behavior in particular. There have been numerous approaches to systematize the plethora of symptoms according to a variety of theoretical concepts. One widely accepted concept is the distinction between positive and negative symptoms (Andreasen et al., 1995, **Table 2**) (*see* Psychiatric Assessment: Negative Symptoms). This approach groups symptoms according to whether they represent a loss of or a deficiency in a normal brain function (negative symptoms) or the appearance of abnormal phenomena (positive symptoms). From the perspective of treatment the positive–negative dichotomy is of importance because these two domains react somewhat differently to treatment. In general, positive symptoms respond fairly well to treatment with antipsychotic drugs, whereas negative symptoms are rather difficult to influence. The latter is particularly problematic, because negative symptoms are usually the limiting factor for personal and social rehabilitation. In addition to the typical symptoms, patients suffering from schizophrenia often present numerous other psychiatric problems. Among these, anxiety, sleep disturbances, obsessions and compulsions, depression and drug or substance abuse are particularly frequent. Although these additional symptoms and problems may in many cases be the consequence of the typical symptoms rather than signs of independent additional psychiatric disorders, they complicate the course of the disease and often necessitate specific treatment approaches.

Table 2 Positive and negative symptoms of schizophrenia. Areas of symptoms are listed and selected examples appear in parentheses

Positive symptoms
Hallucinations (*auditory, somato-tactile, visual, olfactory*)
Delusions (*persecutory, religious, thought broadcasting, thought insertion*)
Bizarre behavior (*unusual clothing or appearance, repetitive, stereotyped behavior*)
Positive formal thought disorders (*incoherence, illogicality, tangentiality*)
Negative symptoms
Affective flattening (*unchanging facial expression, paucity of expressive gestures*)
Alogia (*poverty of speech, increased response latency*)
Avolition-apathy (*physical anergia, problems with personal hygiene*)
Anhedonia-asociality (*reduction in social activities, sexual interest, closeness*)
Attention (*social inattentiveness, inattentiveness during testing*)

The course of schizophrenia varies considerably between patients. Typically, the disease begins during early adulthood with the appearance of rather unspecific negative symptoms resulting in social withdrawal and impaired social and scholastic adjustment. Months to years later positive symptoms such as delusions and hallucinations appear either gradually or abruptly. Across time, positive symptoms tend to show an episodic course, whereas negative symptoms either remain quite stable or even progress. The associated psychiatric problems mentioned above do not show any systematic temporal relationship to the course of schizophrenia itself. Since the syndrome of schizophrenia has been defined, data on the final outcome have varied considerably, depending mainly on the diagnostic concepts applied. There is no doubt that full remission occurs, but probably in less than 20% of patients. In the majority of patients schizophrenia takes a chronic course (Schultz and Andreasen, 1999). The major benefit from modern treatment strategies is probably not a substantial increase in the rate of full remissions, but a significant reduction in the number of extremely ill patients, and in the severity and number of episodes characterized by prominent positive symptoms.

The Principles of Treatment

Treatment of schizophrenia is usually multimodal and comprises approaches from two major areas, which are drug treatment and psychosocial interventions. In general, appropriate drug treatment is a prerequisite for the ability of the patients to comply with and actively take part in psychosocial treatments. The more effective drug treatment is, the more specialized and sophisticated psychosocial interventions can be successfully applied. Vice versa, appropriate psychosocial treatment considerably improves the compliance with drug treatment, because it enhances insight into the disease process, which initially is poor in many patients suffering from schizophrenia.

Antipsychotic drugs (also called neuroleptics) are the most important and most effective therapeutic weapon. The major targets of these drugs are positive symptoms, although newer substances might also reduce negative symptoms to some extent (see below). Treatment with antipsychotic drugs is usually a long-term treatment, whereas drugs to control accessory symptoms (anxiolytics, hypnotics, and antidepressants) are prescribed intermittently when needed. The use of electroconvulsive therapy was widespread before antipsychotic

drugs were available, but today is very limited, although this treatment is effective in certain conditions (Fink and Sackeim, 1996).

Among the psychosocial interventions, supportive and psychoeducative approaches are feasible for the majority of patients, whereas structured social skill training, family therapy, or complex programs including cognitive-behavioral therapies require a considerable degree of insight and compliance. Although recent psychosocial treatment approaches incorporate also psychodynamic aspects, classical psychodynamic psychotherapy, in general, is not effective and sometimes even counterproductive.

Prior to the discovery of antipsychotic drugs, most schizophrenic patients had been hospitalized for decades or even for their entire life. Today, very differentiated treatment facilities are available, which to a large extent are community based. These include, besides classical hospitals and outpatient departments, short-term crisis intervention facilities, day-time clinics, and supported living facilities. This network of facilities considerably increases the chance for social integration and reduces the time spent overall in hospitals. However, recent aggressive approaches to shorten the duration of hospital stays dramatically might lead to a high frequency of rehospitalization, poor long-term outcome, and an overall increase in the costs of treatment.

An important limiting factor for all treatment approaches is the patients' compliance. In particular in severely ill patients compliance is often poor. In schizophrenia noncompliance is particularly due to a lack of insight into the fact of being ill, reduced ability to actively take part in the treatment process due to negative symptoms, or active avoidance of treatment due to positive symptoms such as delusions or imperative voices. To enable stable treatment compliance in patients suffering from schizophrenia, a confidential and empathetic attitude of the people involved is an important prerequisite.

Particular Aspects of Treatment

Pharmacotherapy

The introduction of chlorpromazine into clinical practice in 1952 is the hallmark of the pharmacological treatment of schizophrenia (Delay and Deniker, 1952). Chemically, chlorpromazine is a phenothiazine and other substances of this group also possess antipsychotic properties. In the 1950s and 1960s, numerous antipsychotic drugs with different structures (in particular butyrophenones (e.g., haloperidol) and thioxanthenes (e.g., flupentixol)) were developed. Despite this diversity in chemical structure, these first-generation drugs do not differ qualitatively with respect to their profile of beneficial and undesired effects. They target mainly the positive symptoms of schizophrenia mentioned in **Table 2** without preferentially affecting one or the other of these symptoms. Their effectiveness does not depend on the subtype of schizophrenia (see **Table 1**).

The common mode of action of the first-generation antipsychotics is believed to be blockade of the D2 subtype of dopamine receptors (Pickar, 1995). Blockade of these receptors in the mesolimbic dopamine system is thought to cause a reduction in positive symptoms, and blockade in the nigrostriatal dopamine system is believed to be responsible for the typical side effects of these drugs. These are severe disturbances of motor behavior caused by a drug-induced dysfunction of the dopaminergic extrapyramidal system, which plays a pivotal role in the control of movements. **Table 3** summarizes these extrapyramidal syndromes. Those which occur acutely are often very disturbing, but respond fairly well to treatment (e.g., acute dystonia to anticholinergic drugs), or at least cease upon drug discontinuation or dose reduction. Tardive extrapyramidal syndromes, however, are frequent (5–10% on long-term treatment with first-generation antipsychotics), often resistant to treatment approaches, and they usually persist if the antipsychotic drug is stopped. Another severe side effect of unknown cause, which might be related to the extrapyramidal system, is the neuroleptic malignant syndrome (Pelonero et al., 1998; see also Section Emergency Treatment). Extrapyramidal side effects are common to all first-generation drugs, whereas other, less specific side effects (e.g., sedation, weight gain, postural hypotension) occur with some but not all substances (see Psychopharmacotherapy: Side Effects).

Due to the central role attributed to D2 dopamine receptor blockade, it has been thought for a long time that antipsychotic effectiveness and extrapyramidal syndromes are inevitably linked to each other. However, in the late 1960s it turned out that the dibenzodiazepine clozapine is a potent antipsychotic drug that almost never induces extrapyramidal symptoms and is effective in a proportion of otherwise treatment-resistant patients (Angst et al., 1971). Moreover, it was shown that clozapine not only improves positive but to some extent also negative symptoms. Therefore, the introduction of clozapine into clinical practice was the first qualitative advance after the discovery of chlorpromazine, and it is justified to call this substance a second-generation antipsychotic drug. Unfortunately it turned out quickly that clozapine has a very particular and important drawback: it induces life-threatening agranulocytosis in about 1% of patients treated (Baldessarini and Frankenburg, 1991). Therefore, the use of this drug is only possible when white blood cell counts are monitored in short intervals.

It is still not known why clozapine is a potent antipsychotic drug, but lacks extrapyramidal side effects. Among the possible reasons are that clozapine affects mainly the mesolimbic

Table 3 Extrapyramidal syndromes induced by first-generation antipsychotic drugs

Acute (within days)	*(Late or tardive (within months to years))*
Drug-induced Parkinsonism (*bradykinesia, increased muscular tone, tremor*)	Tardive dyskinesia or dystonia (*irregular choreiform, or dystonic movements in any voluntary muscle group*)
Dystonia (*sudden-onset, sustained muscular contractions preferably in the cephalic musculature*)	Tardive tics (*repetitive, short-lasting, stereotyped movements or other kind of behavior (e.g., vocalization)*)
Akathisia (*internal restlessness and urge to move, often coupled with pacing behavior*)	Tardive myoclonus (*brief asynchronous movements which are not stereotyped*)

dopamine system, that it blocks D1 receptors to the same extent as D2 receptors, that it has a high affinity to D4 receptors, and that it also has high affinity to serotonin receptors. This theoretical framework was the basis for the development of various second-generation antipsychotics (see Table 4). Most of them have high affinities for serotonin receptors, although only one substance (olanzapine) is similar to clozapine in the equivalent binding to D1 and D2 receptors. So far there is no indication that any of these newer drugs carries a substantial risk of agranulocytosis, the most dangerous side effect of clozapine. Moreover, these substances have shown to be effective when compared to first-generation antipsychotic drugs and indeed to induce fewer extrapyramidal side effects. For two of them (olanzapine and risperidone) there are data suggesting that they also might ameliorate negative symptoms. However, it still remains to be seen how the effectiveness of these newer drugs compares to clozapine.

Treatment with antipsychotic drugs in schizophrenic patients is usually a long-term treatment continuing for many years. Intermittently, patients may need additional other psychotropic medications. Most frequently anxiolytics, hypnotics, and antidepressants are administered. The recognition and treatment of intervening depressive episodes is particularly important because depressive and negative symptoms are sometimes difficult to differentiate, but need different treatment. When administering antidepressant drugs to schizophrenic patients, one should be aware that during treatment with antidepressants positive symptoms sometimes exacerbate.

Psychosocial Treatment

It is now generally accepted that psychosocial treatment of patients with schizophrenia is not an alternative to pharmacological treatment, but rather an important complementary approach (Bustillo et al., 2000). Psychosocial measures yield little benefit for positive symptoms; in contrast, intensive psychodynamic psychotherapy might induce exacerbations. The most important and successful interventions are psychoeducative teaching and training in social skills and problem solving. Although it is still debated whether these treatments are directly effective against negative symptoms, there is no doubt that they considerably improve coping with these symptoms.

The major goal of educational approaches is to increase the knowledge about and the insight into the disease process and, as a consequence, to improve treatment compliance. Educational programs are focused on knowledge about the disease process and it's course, the mode of action and side effects of medication, early signs of exacerbation such as restlessness and insomnia, and detailed information about the various community based or hospital facilities available. These programs are particularly effective when important family members are educated as well. Moreover, approaches to improve the communication styles within the patient's network of social relationships are often useful, although the concept that particular forms of communication ('high expressed emotions') play a causative role in schizophrenia has been seriously challenged.

Schizophrenic patients' social skills are often poor. This is one major reason for the often dramatically reduced amount of social contacts. Training of social skills in patients suffering from schizophrenia is performed in a group setting. Such groups are focused on the repetitive training of behavior in various social situations of importance for everyday life. In addition, some rather basic teaching of the theoretical background of successful social interaction might be included.

Impaired problem solving in schizophrenic patients is due to a combination of disturbances in divided attention, planning, and motivation. Problem solving training includes the analysis of problems relevant to daily life and, particularly, dividing a problem into subproblems that can be successively solved. Similar to the training of social skills, the steps of the solution are practiced repeatedly.

In recent years, more complex psychotherapeutic approaches have been developed, combining cognitive, behavioral, or psychodynamic approaches, which, however, are suitable for the minority of patients only.

In particular for chronic patients who in earlier times often were hospitalized for decades, providing supervised residential living arrangements has proven to be of considerable benefit because in this setting the degree of self-sustainment can be maximized in parallel with the permanent provision of professional help.

Emergency Treatment

Schizophrenia is not only a debilitating but also a life-threatening disorder. If severe positive symptoms are present, such as verbal imperative hallucinations, delusions of persecution, or severe formal thought disturbances, the patient's life is threatened by a misinterpretation of his or her environment

Table 4 Relative receptor affinities of selected newer antipsychotic compounds compared to the first-generation drug haloperidol

	(Dopamine)		*(Serotonin)*		*(Noradrenaline)*		*Histamine*	*Acetylcholine*
	D_1	D_2	$5HT_1$	$5HT_{2\alpha}$	α_1	A_2	H_1	M_1
Haloperidol	+++	++++	..	+	++	..	..	..
Clozapine	++	++	+	+++	+++	+++	++++	+++++
Risperidone	++	++++	++	+++++	+++	+++	++	..
Olanzapine	+++	+++	..	+++	+++	..	++++	+++++
Seroquel	+	++	..	+	++++	+	++++	+++
Sertindole	+++	+++++	++	++++	+++	..	+	+
Ziprasidone	+	++++	*	++++	++	..	+	..

Adapted from Pickar, 1995a. Prototype first generation antipsychotic drug.

(e.g., of dangerous objects or situations) and by suicidal intentions. Suicidality can also emerge when negative symptoms are prominent. Acute suicidality and an acute confusional psychotic state both necessitate immediate treatment, including the administration of anxiolytics, antipsychotics, or both, preferably in a hospital setting. Although every effort should be undertaken to get the informed consent of the patient, it is often necessary to initiate treatment against his or her explicit will. This has to be done in accordance with the legal regulations in the respective country.

Less frequently occurring life-threatening situations in patients suffering from schizophrenia are febrile catatonia and the neuroleptic malignant syndrome, a rare side effect of antipsychotic drugs (Pelonero et al., 1998). Both conditions show similarities, including fever, disturbances of motor behavior, and reduced responsiveness to external stimuli. The differential diagnosis is difficult, but very important because in febrile catatonia treatment with antipsychotic drugs is necessary, whereas such treatment has to be immediately stopped when a neuroleptic malignant syndrome is suspected. In both of these conditions admission to an intensive care unit is necessary.

Future Perspectives on the Treatment of Schizophrenia

Despite the rapid development of the knowledge about human genetics, it is unlikely that we will see a major advance in our understanding of the genetic causes of schizophrenia in the near future. The reason is that the available epidemiological data render a major contribution of single genes quite unlikely and suggest that multiple genes contribute to the susceptibility to schizophrenia. However, it is likely that genetic investigations will improve the management of schizophrenia by enabling prediction of the treatment response to antipsychotic drugs. Very recently, for example, it has been shown that a combination of six polymorphisms in neurotransmitter-receptor-related genes enable a significant prediction of the treatment response to clozapine (Arranz et al., 2000). Although these preliminary studies must be replicated and extended to other drugs, it is probable that this approach will significantly enhance the success of drug treatment in schizophrenia. It is also likely that the near future will bring new drugs which are the result of a 'fine-tuning' of the chemical structure and the binding profile to dopamine and serotonin receptors of those substances which are presently available. Another approach which is currently being pursued is to target glutamate receptors, although first clinical trials are only in part encouraging (Farber et al., 1999). Based on a theory implicating the immune system in the pathophysiology of schizophrenia, studies are being started that use immunomodulation as a treatment. This is an interesting and promising approach, which also might help to understand the mode of action of presently available antipsychotic drugs because some of them influence the production of cytokines which are important immune mediators (Pollmächer et al., 2000). Another approach based on pathophysiological theories is to target phospholipids. These are an important cell-membrane component and there is some evidence suggesting that either their uptake is deficient or that they are excessively broken down in schizophrenic patients (Horrobin, 1999).

See also: Mental Illness, Etiology of; Nosology in Psychiatry; Psychiatric Assessment: Negative Symptoms.

Bibliography

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, Washington, DC.

Andreasen, N.C., Arndt, S., Alliger, R., Miller, D., Flaum, M., 1995. Symptoms of schizophrenia. Methods, meanings, and mechanisms. Archives of General Psychiatry 52, 341–351.

Andreasen, N.C., Paradiso, S., O'Leary, D.S., 1998. "Cognitive dysmetria" as an integrative theory of schizophrenia: a dysfunction in cortical-subcortical-cerebellar circuitry? Schizophrenia Bulletin 24, 203–218.

Angst, J., Bente, D., Berner, P., Heimann, H., Helmchen, H., Hippius, H., 1971. Das klinische Wirkungsbild von Clozapin. Pharmakopsychiatrie 4, 201–211.

Arranz, M.J., Munro, J., Birkett, J., Bolonna, A., Mancama, D., Sodhi, M., Lesch, K.P., Meyer, J.F., Sham, P., Collier, D.A., Murray, R.M., Kerwin, R.W., 2000. Pharmacogenetic prediction of clozapine response. Lancet 355, 1615–1616.

Baldessarini, R.J., Frankenburg, F.R., 1991. Drug therapy – Clozapine. A novel antipsychotic agent. New England Journal of Medicine 324, 746–754.

Bustillo, J., Keith, S.J., Lauriello, J., 2000. Schizophrenia: psychosocial treatment. In: Sadock, B.J., Sadock, V.A. (Eds.), Kaplan & Sadock's Comprehensive Textbook of Psychiatry, seventh ed. Lippincott Williams and Wilkins, Baltimore, MD, pp. 1210–1217.

Delay, J., Deniker, P., 1952. 38 cas de psychoses traitées par la cure prolongée et continue de 4560 R.P. Annales médico-psychologique 110, 364–741.

Farber, N.B., Newcomer, J.W., Olney, J.W., 1999. Glycine agonists: what can they teach us about schizophrenia? Archives of General Psychiatry 56, 13–17.

Fink, M., Sackeim, H.A., 1996. Convulsive therapy in schizophrenia? Schizophrenia Bulletin 22, 27–39.

Horrobin, D.F., 1999. Lipid metabolism, human evolution and schizophrenia. Prostaglandins Leukotrienes and Essential Fatty Acids 60, 431–437.

Kraepelin, E., 1910. Psychiatrie. Ein Lehrbuch fār Studierende und Ärzte. 8 Auflage. Barth, Leipzig.

Lieberman, J.A., 1999. Is schizophrenia a neurodegenerative disorder? A clinical and neurobiological perspective. Biological Psychiatry 46, 729–739.

Owen, M.J., Cardno, A.G., 1999. Psychiatric genetics: progress, problems, and potential. Lancet 354 (Suppl. 1), SI11–SI14.

Pelonero, A.L., Levenson, J.L., Pandurangi, A.K., 1998. Neuroleptic malignant syndrome: a review. Psychiatric Services 49, 1163–1172.

Pickar, D., 1995. Prospects for pharmacotherapy of schizophrenia. Lancet 345, 557–562.

Pollmächer, T., Haack, M., Schuld, A., Kraus, T., Hinze-Selch, D., 2000. Effects of antipsychotic drugs on cytokine networks. Journal of Psychiatric Research 34, 369–382.

Schultz, S.K., Andreasen, N.C., 1999. Schizophrenia. Lancet 353, 1425–1430.

World Health Organization, 1996. International Statistical Classification of Diseases. World Health Organization, Geneva, Switzerland.

Schizophrenia: Metabolic Consequences

Antonio Ventriglio and Antonello Bellomo, University of Foggia, Foggia, Italy
Dinesh Bhugra, Institute of Psychiatry, King's College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

People affected by schizophrenia and other severe mental illnesses show a reduced expectancy of life as compared with the general population because of a high rate of comorbidities (obesity, diabetes, hypertension, dyslipidemias, metabolic syndrome) and unhealthy lifestyles. In fact, the cardiometabolic risk is considered to be higher for people affected by schizophrenia with a two- to threefold increased risk of death mostly from cardiovascular and metabolic disease. Antipsychotic agents are an effective and necessary component of the management of such major severe mental illness. Although there may be an independent link between schizophrenia and metabolic conditions, there is a considerable literature regarding the metabolic adverse effects of antipsychotic agents, particularly for the second-generation antipsychotics. Over recent years, international organizations have developed screening and monitoring guidelines in order to reduce modifiable risk factors, comorbidity, and mortality rates among patients affected by schizophrenia and improve the outcome of illness, including patients' functioning and quality of life.

Introduction

People affected by schizophrenia and other severe mental illness show a reduced expectancy of life as compared with the general population because of a high rate of comorbidities and unhealthy lifestyles (Saha et al., 2007; Fleischacker et al., 2008).

In fact, they report a two- to threefold increased risk of death mostly from cardiovascular and metabolic disease (Saha et al., 2007; Fleischacker, 2008). In 2004 the American Diabetes Association (ADA, www.diabetes.org) promoted an initiative regarding the cardiometabolic risk (CMR) in the general population and identified some risk factors considered to be modifiable (overweight/obesity, high blood glucose, hypertension, abnormal lipid metabolism, physical inactivity, smoking) or nonmodifiable (age, race and ethnicity, gender, family history). The CMR is considered to be increased for people affected by schizophrenia and major mental illness since prevalence and relative risk of modifiable factors are higher in the psychiatric population as compared with the general one (Correll, 2007). Also, many authors (Fleischhacker et al., 2008; Newcomer, 2005; Cradock-O′ Leary et al., 2002; Druss et al., 2001) describe that schizophrenia patients may have poorer access to health care and an unhealthy lifestyle (poor diet and physical exercise).

In the last decades, many authors have demonstrated that antipsychotic agents (APs) may have a great impact on the modifiable risk factors leading to weight gain and significant metabolic changes among patients treated (Nasrallah, 2008; Newcomer, 2007; Papanastasiou, 2013). Also, metabolic syndrome that is a constellation of clinical conditions, recognized as independent risk factors for cardiovascular disease (including abdominal obesity, insulin resistance, dyslipidemias, hypertension), is highly prevalent in individuals with schizophrenia and conveys significant cardiovascular risk and mortality (Papanastasiou, 2013).

Figure 1 summarizes all factors involved in the increased risk for cardiometabolic diseases among schizophrenia patients.

Mortality in Schizophrenia

Correll in 2007 estimated that the prevalence of CMR modifiable factors and relative risk (RR) are higher among patients affected by schizophrenia as compared with the general population. In fact those patients are more likely to be overweight/obese (45–55%, RR: 1.5–2), smokers (50–80%, RR: 2–3), to have diabetes (10–15%, RR: 2), hypertension (19–58%, RR: 2–3), dyslipidemias (25–69%, RR $\leq$ 5), and metabolic syndrome (37–63%, RR: 2–3). The median

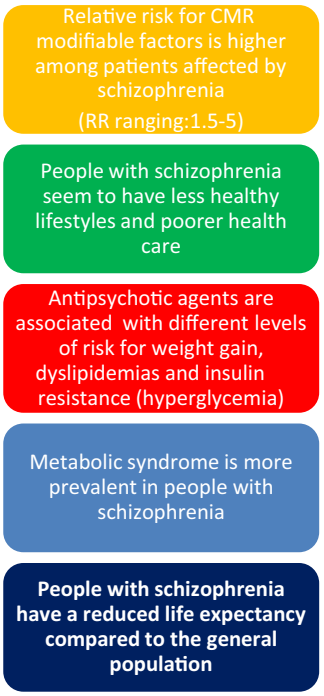

Figure 1 Cardiometabolic risk (CMR) among patients affected by schizophrenia.

all-causes standardized mortality ratio (SMR) for people affected by schizophrenia was 1.84 in the 1970s, 2.98 in the 1980s, and 3.20 in the 1990s, with an increasing mortality over time (De Hert et al., 2009a). In particular, cardiovascular mortality in schizophrenia increased in 1976–95 with the greatest increase in SMRs in men from 1991 to 1995 (Osby et al., 2000). Moreover, medical comorbidities, in particular cardiovascular disease, in patients with schizophrenia may lead to a 20% shorter lifespan (10–15 years) (Harris and Barraclough, 1998; McLean et al., 2014; Castillo Sánchez et al., 2014). In fact, it has been widely described that schizophrenia is associated with premature mortality and a high rate of sudden, unexpected deaths mostly caused by structural cardiovascular, respiratory, and neurological abnormalities, with most cases due to acute myocardial infarction (Ifteni et al., 2014). In conclusion, early recognition and treatment of cardiovascular and metabolic conditions must become a clinical priority for all adults with schizophrenia as the standard risk assessment for such comorbidities underestimates the risk of death in schizophrenia (McLean et al., 2014; Ifteni et al., 2014).

Vulnerability

The link between metabolic abnormalities and schizophrenia has been described before the era of modern antipsychotic drugs. Genetic links between diabetes and schizophrenia have been recognized by Sir Henry Maudsley in his 1897 text, *The Pathology of Mind*, who noted: "Diabetes often shows itself in families in which insanity prevails." Also, weight gain and abnormal eating behaviors were commented by Kraepelin and Bleuler in the beginning of 1900.

Evidence from latest decades confirmed that there may be a link between major mental illness and metabolic conditions without any influence by APs. In fact, metabolic disturbances have been found in drug-naïve patients with first-episode illness and may be an inherent part of the illness.

First-episode, drug-naïve patients with schizophrenia have impaired fasting glucose tolerance, are more insulin resistant, and have higher levels of glucose, insulin, and cortisol than healthy comparison subjects (Ryan et al., 2003; Thakore, 2004). These patients have higher than expected rates of visceral obesity which may be related to a subtle disturbance of the hypothalamic–pituitary–adrenal axis (Thakore, 2004). It has been recently described that molecular abnormalities in metabolic, hormonal, and immune pathways are present in peripheral body fluids of schizophrenia patients. These findings suggest the presence of a molecular endophenotype involving disruption of insulin and growth factor signaling pathways as an increased risk factor for schizophrenia (van Beveren et al., 2014). In addition, drug-naïve, first-episode schizophrenia patients, even if with normal weight, present an upregulated inflammatory status (increased levels of IL-1β, IL-6, and TNF-α) and higher adiponectin levels that may play a proinflammatory role leading to later metabolic abnormalities in this patient population (Song et al., 2013).

Mitchell et al., in 2013, described that metabolic factors (including weight, triglycerides, impaired fasting glucose, cholesterol levels, diabetes, smoking) are lower among patients at first episode of psychosis or unmedicated patients as compared with treated schizophrenia patients, but factors are altered in such early-stage patients and metabolic syndrome may be present in the unmedicated group more than in the first-episode one.

Although the cardiovascular risk is not significantly higher at this stage of illness, the first-episode patients should be screened for the primary prevention of cardiovascular morbidity and mortality (Phutane et al., 2011).

Antipsychotic Agents

Psychopharmacological treatment of schizophrenia is an effective and necessary component for the management of such severe major mental disorder. It has been clearly established that unmedicated patients report much higher mortality and suicide rates, more hospitalizations, and worst outcomes of illness with greater cognitive and functional impairment (De Hert et al., 2009a; Tiihonen et al., 2006).

However, there is a considerable literature regarding the metabolic adverse effects of APs. Growing evidence demonstrates that newer agents, second-generation antipsychotics (SGAs), induce more weight gain and metabolic abnormalities as compared with first-generation antipsychotics (FGAs) (De Hert et al., 2009a; Nasrallah, 2008; Newcomer, 2007; Papanastasiou, 2013).

It has been a shift of risk perception for patients and clinicians employing newer agents, i.e., SGAs, with a major concern regarding metabolic side effects more than extrapyramidal symptoms and tardive dyskinesia. In fact, SGAs are responsible for new cases of diabetes. Although diabetes is considered to be related to other metabolic dysfunctions (hypertriglyceridemia, insulin resistance, and weight gain) observed during the treatment with low-potency FGAs, many types of evidence show that there may be specific diabetogenic actions of some SGAs, in particular for clozapine and olanzapine (Reynolds and Kirk, 2010; Stahl, 2013). The diabetogenic mechanism may involve a muscarinic M3-receptor antagonism, which is remarkable for olanzapine and clozapine. M3 receptors are localized on beta-cells of the pancreas and regulate the insulin release and glucose level homeostasis. M3-receptor blockage might alter glucose metabolism with clinical conditions of diabetes, diabetic ketoacidosis, and hyperosmolar syndrome, all observed in patients treated with olanzapine and clozapine (Reynolds and Kirk, 2010; Stahl, 2013). Weight gain and dyslipidemia (hypertriglyceridemia) are triggering factors for diabetes and are frequently induced by antipsychotics like olanzapine and clozapine whereas aripiprazole and ziprasidone show safer tolerability profiles to be considered in case of switching from other antipsychotics for metabolic side effects (Reynolds and Kirk, 2010; Stahl, 2013). Weight gain might be induced by 5HT2C- and H1-receptor affinity of newer antipsychotics; in particular H1-receptors are localized in the hypothalamus (hunger and satiety centers) and may be involved in inducing hyperphagia. Chlorpromazine, as an FGA, is also associated with weight gain for its high H1-receptor affinity (Reynolds and Kirk, 2010; Stahl, 2013). Other receptor mechanisms may have additive or synergistic

effects; dopamine D2 receptor antagonism can enhance 5-HT2C-mediated effects on food intake, as well as influencing lipid and glucose metabolism via disinhibition of prolactin secretion (Reynolds and Kirk, 2010; Stahl, 2013). SGAs might also have effects on the hypothalamic regions involved in control of food intake: olanzapine enhances neuropeptide Y (NPY) expression in the ARC (arcuate nucleus) (Reynolds and Kirk, 2010).

Table 1 shows the atypical SGAs and their tolerability profiles according to the literature evidence (Leucht et al., 2009a; American Diabetes Association/American Psychiatric Association, 2004; Tschoner et al., 2007; Cohn and Sernyak, 2006; De Hert et al., 2006; De Nayer et al., 2005).

Weight Gain

Weight gain during the acute and maintenance treatment of schizophrenia involves 15–72% of patients (De Hert, 2009). A *metaanalysis* of clinical trials showed significant weight gain with a 10-week antipsychotic treatment ranging as: clozapine (+4.45 kg) > olanzapine (+4.15 kg) > quetiapine and risperidone (+2.1 kg) > aripiprazole and ziprasidone (<1 kg) (Allison et al., 1999; De Hert et al., 2009). This hierarchy of induced weight gain was confirmed by following studies and *metaanalyses* including Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) and European First Episode Schizophrenia Trial (EUFEST) Studies (De Hert et al., 2009a; Daumit et al., 2008; Leucht et al., 2009b; Goff et al., 2005; Lieberman et al., 2005).

There are some demographic and clinical features associated with a greater propensity for weight gain: younger age, lower initial BMI, personal and family history of obesity, nonwhite ethnic background, tendency to overeat in times of stress, cannabis use, and first-episode psychosis (De Hert et al., 2009a). A comprehensive literature review, however, did not find evidence for a dose–response at the doses used to treat psychotic illnesses for most antipsychotics, except clozapine and olanzapine (Simon et al., 2009).

Although there are significant findings in many other genes involved in the antipsychotic-induced weight gain, the most consistently replicated findings are in the melanocortin 4 receptor (MC4R), the serotonin 2C receptor (HTR2C), the leptin, the neuropeptide Y (NPY) and the cannabinoid receptor 1 (CNR1) genes (Shams and Müller, 2014).

Diabetes

There is a higher risk of developing diabetes mellitus type 2 for people receiving antipsychotic medications. The sources of information range from case histories, through pharmacoepidemiological studies to randomized controlled studies (De Hert et al., 2009a).

Causality between antipsychotic use and the onset of diabetes can be difficult. People with schizophrenia show increased risk for diabetes independent of drug therapy.

In fact, approximately 15% of people with schizophrenia develop diabetes with a prevalence that is approximately two times greater than the general population (Bushe and Holt, 2004). In 2008 schizophrenia was recognized by the Canadian Diabetes Association (CDA) as an independent risk factor for diabetes. The CDA also lists SGAs as a risk factor for diabetes.

In contrast to some studied antipsychotics, clozapine and olanzapine are associated with a concentration-dependent increase in circulating insulin levels thought to be secondary to a direct stimulating effect of these agents on pancreatic insulin secretion (Melkersson et al., 2004). In the CATIE study, there was a significantly greater increase in HbA1c with olanzapine (0.4%) > quetiapine (0.04%) > risperidone (0.07%) > perphenazine (0.09%) > ziprasidone (0.11%) (Lieberman et al., 2005).

A baseline assessment of glucose regulation is important to screen new cases of diabetes due to the antipsychotic treatment. Clinicians managing patients treated in the long term need to make efforts to prevent diabetes and respond to it early when it does occur.

Dyslipidemias

Risk factors for coronary artery disease include increased low-density lipoprotein cholesterol (LDL-C) and/or decreased high-density lipoprotein cholesterol (HDL-C).

Many studies show that the use of antipsychotics is associated with an increase of LDL-C and a decrease of HDL-C (Daumit et al., 2008; Goff et al., 2005).

It has been indicated that the effect on total and LDL cholesterol differs between antipsychotics: drugs associated with the most weight gain, such as clozapine and olanzapine, are associated with the largest increase in serum LDL and triglycerides

Table 1 Side effects of second-generation antipsychotics (SGAs)

	Weight gain	*Dyslipidemia*	*Diabetes*	*EPS*	*Prolactin increase*	*Sedation*	*QT-interval prolongation*
Amisulpride	+	–	+	+	+++	–	–
Aripiprazole	–	–	–	+/–	–	+	–
Asenapine	+/–	+/–	–	–	–	++	–
Clozapine	+++	++	+++	–	–	+++	+
Olanzapine	+++	++	+++	–	–	+++	–
Quetiapine	++	+	++	–	–	++	–
Risperidone	++	+/–	++	+	+++	+	++
Ziprasidone	+/–	–	–	–	+	–	++

+++ = high incidence, ++ = moderate incidence, + = low incidence, – = very low incidence, EPS: extrapyramidal symptoms.
Modified by Leucht, S., Komossa, K., Rummel-Kluge, C., Corves, C., Hunger, H., Schmid, F., Asenjo Lobos, C., Schwarz, S., Davis, J.M., 2009a. A *metaanalysis* of head-to-head comparisons of second-generation antipsychotics in the treatment of schizophrenia. American Journal of Psychiatry 166 (2), 152–163.

(Daumit et al., 2008; Goff et al., 2005). The most marked increase is observed in the early phase of treatment when weight gain occurs most rapidly (Daumit et al., 2008; Goff et al., 2005). Most studies suggest lipid measurements after short periods of treatment (2–3 months) and there is a need to assess what the longer-term effects on triglyceride concentrations are once weight has stabilized (De Hert et al., 2009a). Lipid concentration may decrease rapidly after a psychoeducational and monitoring program (Ventriglio et al., 2014).

Blood Pressure

Increases in blood pressure may occur sporadically, mostly involving clozapine (Gardner and Theean, 2011). Reduction of blood pressure with associated symptoms of feeling weak is a common occurrence with many antipsychotics. Syncopal episodes related to orthostasis may happen, mostly due to low-potency FGAs (α1-adrenergic and calcium blockade are generally thought to account for this adverse effect). Posture-related drops in blood pressure are especially common with FGAs. A slower dosage titration was recommended when it was observed. With most drugs, tolerance develops with continued use, and the effects recede (Gardner and Theean, 2011).

Metabolic Syndrome

The metabolic syndrome (MetS; also called syndrome X or Reaven's syndrome) is a constellation of different conditions, including abdominal obesity, insulin resistance, dyslipidemia, and elevated blood pressure. All components of the MetS have been recognized as independent risk factors for cardiovascular disease and are associated with other conditions such as the prothrombotic state, the proinflammatory state, nonalcoholic fatty liver disease, and reproductive disorders (Papanastasiou, 2013).

Many definitions have been given to MetS: World Health Organization (WHO) in 1985, European Group for the Study of Insulin Resistance in 1999, National Cholesterol Education Program's Adult Treatment Panel III (NCEP-ATP III) in 2001 (modified in 2003), American Association of Clinical Endocrinologists in 2003, and finally by the International Diabetes Federation (IDF) in 2006 (Papanastasiou, 2013).

MetS may be diagnosed if three out of the following five conditions are present in the same patient:

1. Elevated waist circumference, M > 102 cm F > 88 cm (IDF: M $\geq$ 94 cm F $\geq$ 80 cm).
2. Elevated triglycerides ($\geq$150 mg dl^{-1}) or drug treatment for dyslipidemia.
3. Reduced high-density lipoprotein (HDL) (<40 mg dl^{-1}) or drug treatment for hypercholesterolemia.
4. Elevated blood pressure (systolic $\geq$130 mmHg, diastolic $\geq$85 mmHg) or a history of hypertension or drug treatment for hypertension.
5. Elevated fasting glucose ($\geq$100 mg dl^{-1}) or drug treatment for hyperglycemia.

In a large review of literature Papanastasiou (2013) described that the prevalence rates may vary largely across studies (reflecting the groups of patients studied that may be different for age, sex, ethnicity, medication status, smoking, duration of illness, and outcomes). The lowest prevalence rate reported was 3.9%, and the highest prevalence rate reported was 68%, derived from patients treated with FGAs and/or SGAs. MetS shows a peak of prevalence in the fourth/fifth decade of life among patients studied. For patients treated with antipsychotics, clozapine and olanzapine appeared to be related to higher MetS rates than other APs. Also, the reviews show an increase from 24.7% to 49.6% in the comparison between previously unmedicated and currently treated patients. Patients treated with SGAs showed higher rates of MetS as compared with those treated with FGAs (27.8% versus 9.8%) (De Hert et al., 2009b; Papanastasiou, 2013).

Recently it has been described that MetS might aggravate injury of cognitive function in chronic schizophrenia, especially in immediate memory, delayed memory, and attention (Li et al., 2014).

Prolactin Levels

Antipsychotic-induced hyperprolactinemia results from dopaminergic blockade in the tuberoinfundibular dopaminergic system of the hypothalamus. The magnitude of hyperprolactinemia correlates strongly with dopamine D2 receptor blockade potency. Prolactin level may be associated with clinical signs and consequences: galactorrhea, gynecomastia, oligomenorrhea, amenorrhea, galactorrhea, reduced libido, dyspareunia, vaginal dryness for women; gynecomastia, reduced libido, erectile dysfunction, ejaculatory dysfunction for men (Halbreich et al., 2003; Gardner and Theean, 2011). Osteoporosis, hypogonadism, and breast cancer have been associated with prolactinomas, in which prolactin levels can exceed 5000 ng ml^{-1} (Halbreich et al., 2003; Gardner and Theean, 2011).

Hyperprolactinemia occurs more frequently in females than males, despite lower doses used in women. Return to normal prolactin levels occurs rapidly upon discontinuation of treatment (Halbreich et al., 2003; Gardner and Theean, 2011).

According to literature evidence, antipsychotics are associated with increase of prolactin ranging as: risperidone = paliperidone = amisulpride = haloperidol (74–91%)> olanzapine (2–20%) > quetiapine (no significant increase) > aripiprazole = ziprasidone (no significant increase) (Halbreich et al., 2003; Gardner and Theean, 2011).

Prolactin should be measured at baseline and over the follow-up period (monitoring hyperprolactinemia-related signs and symptoms). Prolactin levels range: nonpregnant women: 0–20 ng ml^{-1} (0–424 mIU l^{-1}); pregnant women: $\leq$200 ng ml^{-1} ($\leq$4240 mIU l^{-1}); men: 0–15 ng ml^{-1} (0–318 mIU l^{-1}) (Kratz et al., 2004).

When prolactin levels are high but lower than 150 ng ml^{-1}, clinicians may consider a switch from the current antipsychotic to another, which is not associated with prolactin increase or, taking into account the risk of psychotic rebounds, may introduce a dopamine agonist like cabergoline, bromocriptine, or pergolide. If prolactin levels are higher than 150 ng ml^{-1}, an endocrinology consult is needed and the application of nuclear magnetic resonance imaging to exclude prolactin-producing pituitary and juxtasellar adenomas is required (De Hert et al., 2014).

QT-Interval Prolongation

Recently, many studies have been reporting data on QT prolongation during the antipsychotic treatment.

QT interval is a measure of the time between the start of the Q wave and the end of the T wave in the heart's electrical cycle. In general, the QT interval represents electrical depolarization (Na^+ and Ca^{2+} intake in the cells) and repolarization (K^+ output) of the ventricles.

QT is corrected for heart rate (QTc) derived from the heart rate (HR) as 60 beats per minute according to Bazett's formula.

Although QT may be variable depending on age, sex, and circadian rhythms, a lengthened QT interval is a marker for the potential of ventricular tachyarrhythmias like *torsades de pointes* and a risk factor for sudden death.

A QTc value above 500 ms or an increase of 60 ms (20% from baseline) is strongly associated with arrhythmias and *torsade de pointes* (TdP) (**Table 2**).

APs may prolong QTc interval depending on their binding potency to IKr channels of ventricle repolarization (hERG channels).

SGAs mostly involved with QTc prolongation are ziprasidone, risperidone, paliperidone, clozapine, and quetiapine. Aripiprazole, olanzapine, and asenapine are associated with lower variations in QTc interval (Hasnain and Vieweg, 2014; Gardner and Theean, 2011).

FGAs are responsible for higher QTc prolongations, in particular haloperidol. Thioridazine and sertindole have been withdrawn worldwide since they may lead to arrhythmias and sudden death.

Antipsychotics should be prescribed with caution for patients affected by cardiovascular diseases (congenital long QT syndrome, myocardial infarction, cardiac failure, arrhythmias) or other risky conditions (previous TdP or hypokalemia) (Hasnain and Vieweg, 2014; Gardner and Theean, 2011).

Finally, it is of interest that myocarditis may occur during the treatment with clozapine, in particular in the first 2 months of it (85% of cases). Also, patients treated with chlorpromazine or fluphenazine may present with myocarditis. If clinical signs of myocarditis are suspected (e.g., tachycardia), the drug should be stopped (Gardner and Theean, 2011).

Screening and Monitoring for Metabolic Consequences

Over recent years, national and international groups have developed screening and monitoring guidelines. The aim of such procedures is to reduce modifiable factors associated with the risk of cardiometabolic conditions among patients affected by schizophrenia (De Hert et al., 2009a). We will consider guidelines from the European Psychiatric Association (EPA), supported by the European Association for the Study of Diabetes (EASD) and the European Society of Cardiology (ESC), also known as EPA/EASD/ESC Position Statement (De Hert et al., 2009a).

These recommendations are poorly employed in the clinical settings even if, when applied, they seem to improve the metabolic outcome of schizophrenia patients (De Hert et al., 2009a; De Hert et al., 2011; Ventriglio et al., 2014).

The baseline medical history and examination should include (De Hert et al., 2009a) history of previous cardiovascular diseases, diabetes, or other related disease; family history of premature CVD, diabetes, or other related disease; smoking habit, food selection, physical activity; weight and height in order to calculate body mass index (BMI) and waist circumference; fasting blood glucose; fasting blood lipids: total cholesterol, triglycerides, LDL-C (by calculation), and HDL-C; blood pressure (measured twice and average taken), heart rate, heart and lung auscultation, foot pulses; ECG.

If metabolic parameters are in normal ranges in patients on antipsychotic treatment for more than 12 months, clinicians have to repeat previous factors annually. If values are in normal ranges at baseline in drug-naïve patients, monitoring has to be repeated at week 6, week 12, and annually (paying particular attention to weight gain within the 6th week of treatment) (De Hert et al., 2009a).

In the presence of one or more risk factors, the antipsychotic treatment has to be selected carefully, by a psychiatric and metabolic prospective. Treatment may be required to reduce the cardiovascular risk. Clinicians have to involve general practitioners in order to collaborate with them in managing the risk factors: BMI should be ≤ 25 kg m^{-2}, waist circumference <102 cm (men)/<88 cm (women), blood pressure $\leq$140/90 mmHg – triglycerides <190 mg dL^{-1} – LDL <115 mg dL^{-1} in nondiabetic patients, glycosylated hemoglobin <7% – blood pressure $\leq$130/80 mmHg – triglycerides <175 mg dL^{-1} – LDL <100 mg dL^{-1} in diabetic patients. Statins treatment is encouraged to reach the recommended values of serum lipids (De Hert et al., 2009a). Diabetic patients should repeat glycosylated hemoglobin three-monthly, microalbuminuria, renal

Table 2 QTc-interval values

QTc (ms)	*Males*	*Females*
Normal	<430	<450
Borderline	431–450	451–470
Prolonged	>450	>470

Table 3 Monitoring and screening flow-chart for metabolic consequences (according to EPA/EASD/ESC Position Statement, De Hert et al., 2009a)

Parameter/time	*Baseline*	*APs prescription*	*Week 6*	*Week 12*	*Yearly*
Weight	✓	✓	✓	✓	✓
Height	✓	✓			✓
BMI	✓	✓	✓	✓	✓
Waist circumference	✓	✓	✓	✓	✓
Fasting blood glucose	✓	✓	✓	✓	✓
Serum lipids	✓	✓	✓	✓	✓
Gly-HB	✓	✓	✓	✓	✓
Blood pressure	✓	✓	✓	✓	✓
Heart rate	✓				✓
Heart and lung auscultation	✓				✓
Foot pulses	✓				✓
ECG	✓				✓

Note: Diabetic patients should repeat glycosylated hemoglobin three-monthly, microalbuminuria, renal functioning, ophthalmological and feet examination yearly.

functioning, ophthalmological and feet examination yearly (De Hert et al., 2009a; **Table 3**).

Conclusion

Metabolic consequences of schizophrenia and antipsychotic treatments may increase the risk of mortality among patients affected. In fact, they show higher levels of cardiometabolic comorbidities as compared with the general population and a two- to threefold increased risk of death. Screening and monitoring recommendations from international organizations aim to reduce the comorbidity and mortality rates among patients affected by schizophrenia and might improve the outcome of illness, including patients' functioning and quality of life.

See also: Antidepressant Drugs; Endocrinology and Psychiatry; Mania; Psychopharmacotherapy, Costs of; Psychopharmacotherapy: Side Effects; Psychotherapy and Pharmacotherapy; Schizophrenia, Treatment of; Schizophrenia.

Bibliography

Allison, D.B., Mentore, J.L., Heo, M., Chandler, L.P., Cappelleri, J.C., Infante, M.C., et al., 1999. Antipsychotic-induced weight gain: a comprehensive research synthesis. American Journal of Psychiatry 156 (11), 1686–1696.

American Diabetes Association/American Psychiatric Association, 2004. Consensus development conference on antipsychotic drugs and obesity and diabetes. Diabetes Care 27, 596–601.

Bushe, C., Holt, R., 2004. Prevalence of diabetes and impaired glucose tolerance in patients with schizophrenia. British Journal of Psychiatry-Supplement 47, S67–S71.

van Beveren, N.J., Schwarz, E., Noll, R., Guest, P.C., Meijer, C., de Haan, L., Bahn, S., 2014. Evidence for disturbed insulin and growth hormone signaling as potential risk factors in the development of schizophrenia. Translational Psychiatry 4, e430.

Canadian Diabetes Association Clinical Practice Guidelines Expert Committee, 2008. Canadian diabetes association 2008 clinical practice guidelines for the prevention and management of diabetes in Canada. Canadian Journal of Diabetes 32 (Suppl. 1), S1–S201.

Castillo Sánchez, M., Fàbregas Escurriola, M., Bergè Baquero, D., Goday Arno, A., Vallès Callol, J.A., 2014. Psychosis, cardiovascular risk and associated mortality: are we on the right track? Clinica e Investigacion Arteriosclerosis 26 (1), 23–32.

Cohn, T.A., Sernyak, M.J., 2006. Metabolic monitoring for patients treated with antipsychotic medications. Canadian Journal of Psychiatry 51, 492–501.

Correll, C.U., 2007. Balancing efficacy and safety in treatment with antipsychotics. CNS Spectrums 12 (Suppl. 17), 12–20.

Cradock-O'Leary, J., Young, A.S., Yano, E.M., et al., 2002. Use of general medical services by VA patients with psychiatric disorders. Psychiatric Services 53, 874–878.

Daumit, G.L., Goff DC,Meyer, J.M., Davis, V.G., Nasrallah HA,McEvoy, J.P., et al., 2008. Antipsychotic effects on estimated 10-year coronary heart disease risk in the CATIE schizophrenia study. Schizophrenia Research 105, 175–187.

De Hert, M., Van Eyck, D., De Nayer, A., 2006. Metabolic abnormalities associated with second generation antipsychotics: fact or fiction? Development of guidelines for screening and monitoring. International Clinical Psychopharmacology 21 (Suppl. 2), 11–15.

De Hert, M., Dekker, J.M., Wood, D., Kahl, K.G., Holt, R.I., Möller, H.J., 2009a. Cardiovascular disease and diabetes in people with severe mental illness position statement from the European Psychiatric Association (EPA), supported by the European Association for the study of diabetes (EASD) and the European Society of Cardiology (ESC). European Psychiatry 24 (6), 412–424.

De Hert, M., Schreurs, V., Vancampfort, D., VAN Winkel, R., 2009b. Metabolic syndrome in people with schizophrenia: a review. World Psychiatry 8 (1), 15–22.

De Hert, M., Vancampfort, D., Correll, C.U., Mercken, V., Peuskens, J., Sweers, K., van Winkel, R., Mitchell, A.J., 2011. Guidelines for screening and monitoring of cardiometabolic risk in schizophrenia: systematic evaluation. British Journal of Psychiatry 199 (2), 99–105.

De Hert, M., Detraux, J., Peuskens, J., 2014. Second-generation and newly approved antipsychotics, serum prolactin levels and sexual dysfunctions: a critical literature review. Expert Opinion on Drug Safety 13 (5), 605–624.

De Nayer, A., De Hert, M., Scheen, A., et al., 2005. Conference report: Belgian consensus on metabolic problems associated with second-generation antipsychotics. International Journal of Psychiatry in Clinical Practice 9, 130–137.

Druss, B.G., Bradford, W.D., Rosenheck, R.A., et al., 2001. Quality of medical care and excess mortality in older patients with mental disorders. Archives of General Psychiatry 58, 565–572.

Fleischhacker, W., Cetkovich-Bakmas, M., De Hert, M., et al., 2008. Comorbid somatic illnesses in patients with severe mental disorders: clinical, policy, and research challenges. Journal of Clinical Psychiatry 69, 514–519.

Gardner, D.M., Teehan, M.D., 2011. Antipsychotics and Their Side Effects. Cambridge University Press.

Goff, D.C., Sullivan, L.M., McEvoy, J.P., Meyer, J., Nasrallah, H., Daumit, G., et al., 2005. A comparison of ten-year cardiac risk estimates in schizophrenia patients from the CATIE study and matched controls. Schizophrenia Research 80 (1), 45–53.

Halbreich, U., Kinon, B.J., Gilmore, J.A., Kahn, L.S., 2003. Elevated prolactin levels in patients with schizophrenia: mechanisms and related adverse effects. Psychoneuroendocrinology 28 (Suppl. 1), 53–67.

Harris, E.C., Barraclough, B., 1998. Excess mortality of mental disorder. British Journal of Psychiatry 173, 11–53.

Hasnain, M., Vieweg, W.V., 2014. QTc interval prolongation and torsade de pointes associated with second-generation antipsychotics and antidepressants: a comprehensive review. CNS Drugs.

Henry, Mausdley, 1897. The Pathology of Mind. D. Appleton and Company, New York.

Ifteni, P., Correll, C.U., Burtea, V., Kane, J.M., Manu, P., 2014. Sudden unexpected death in schizophrenia: autopsy findings in psychiatric inpatients. Schizophrenia Research 155 (1–3), 72–76.

Kratz, A., Ferraro, M., Sluss, P.M., Lewandrowski, K.B., 2004. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Laboratory reference values. New England Journal of Medicine 351 (15), 1548–1563.

Leucht, S., Komossa, K., Rummel-Kluge, C., Corves, C., Hunger, H., Schmid, F., Asenjo Lobos, C., Schwarz, S., Davis, J.M., 2009a. A *metaanalysis* of head-to-head comparisons of second-generation antipsychotics in the treatment of schizophrenia. American Journal of Psychiatry 166 (2), 152–163.

Leucht, S., Corves, C., Arbter, D., Engel, R.R., Li, C., Davis, J.M., 2009b. Second generation versus first-generation antipsychotic drugs for schizophrenia: a *meta-analysis*. Lancet 373 (9657), 31–41.

Li, C., Zhan, G., Rao, S., Zhang, H., 2014. Metabolic syndrome and its factors affect cognitive function in chronic schizophrenia complicated by metabolic syndrome. Journal of Nervous and Mental Disease 202 (4), 313–318.

Lieberman, J.A., Stroup, T.S., McEvoy, J.P., Swartz, M.S., Rosenheck, R.A., Perkins, D.O., et al., 2005. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New England Journal of Medicine 353 (12), 1209–1223.

McLean, G., Martin, J.L., Martin, D.J., Guthrie, B., Mercer, S.W., Smith, D.J., 2014. Standard cardiovascular disease risk algorithms underestimate the risk of cardiovascular disease in schizophrenia: evidence from a national primary care database. Schizophrenia Research.

Melkersson, K.I., Dahl, M.L., Hulting, A.L., 2004. Guidelines for prevention and treatment of adverse effects of antipsychotic drugs on glucose-insulin homeostasis and lipid metabolism. Psychopharmacology (Berl) 175 (1), 1–6.

Mitchell, A.J., Vancampfort, D., Sweers, K., van Winkel, R., Yu, W., De Hert, M., 2013. Prevalence of metabolic syndrome and metabolic abnormalities in schizophrenia and related disorders–a systematic review and *metaanalysis*. Schizophrenia Bulletin 39 (2), 306–318.

Nasrallah, H.A., 2008. Atypical antipsychotic-induced metabolic side effects: insights from receptor-binding profiles. Molecular Psychiatry 13, 27–35.

Newcomer, J.W., 2005. Second-generation (atypical) antipsychotics and metabolic effects: a comprehensive literature review. CNS Drugs 19, 1–93.

Newcomer, J.W., 2007. Antipsychotic medications: metabolic and cardiovascular risk. Journal of Clinical Psychiatry 68 (Suppl. 4), 8–13.

Osby, U., Correia, N., Brandt, L., Ekbom, A., Sparén, P., 2000. Time trends in schizophrenia mortality in Stockholm county, Sweden: cohort study. BMJ 321 (7259), 483–484.

Papanastasiou, E., 2013. The prevalence and mechanisms of metabolic syndrome in schizophrenia: a review. Therapeutic Advances in Psychopharmacology 3 (1), 33–51.

Phutane, V.H., Tek, C., Chwastiak, L., Ratliff, J.C., Ozyuksel, B., Woods, S.W., Srihari, V.H., 2011. Cardiovascular risk in a first-episode psychosis sample: a 'critical period' for prevention? Schizophrenia Research 127 (1–3), 257–261.

Reynolds, G.P., Kirk, S.L., 2010. Metabolic side effects of antipsychotic drug treatment–pharmacological mechanisms. Pharmacology & Therapeutics 125 (1), 169–179.

Ryan, M.C., Collins, P., Thakore, J.H., 2003. Impaired fasting glucose tolerance in first-episode, drug-naive patients with schizophrenia. American Journal of Psychiatry 160 (2), 284–289.

Saha, S., Chant, D., Mcgrath, J., 2007. A systematic review of mortality in schizophrenia: is the differential mortality gap worsening overtime? Archives of General Psychiatry 64, 1123–1131.

Shams, T.A., Müller, D.J., 2014. Antipsychotic induced weight gain: genetics, epigenetics, and biomarkers reviewed. Current Psychiatry Reports 16 (10), 473.

Simon, V., vanWinkel, R., De Hert, M., 2009. Are weight gain and metabolic side effects of atypical antipsychotics dose-dependent? A literature review. Journal of Clinical Psychiatry 70 (7), 1041–1050.

Song, X., Fan, X., Song, X., Zhang, J., Zhang, W., Li, X., Gao, J., Harrington, A., Ziedonis, D., Lv, L., 2013. Elevated levels of adiponectin and other cytokines in drug naïve, first episode schizophrenia patients with normal weight. Schizophrenia Research 150 (1), 269–273.

Stahl, S.M., 2013. Essential Psychopharmacology. Cambridge University Press.

Thakore, J.H., 2004. Metabolic disturbance in first-episode schizophrenia. British Journal of Psychiatry-Supplement 47, S76–S79.

Tiihonen, J., Wahlbeck, K., Lönnqvist, J., Klaukka, T., Ioannidis, J.P., Volavka, J., et al., 2006. Effectiveness of antipsychotic treatments in a nationwide cohort of patients in community care after first hospitalisation due to schizophrenia and schizoaffective disorder: observational follow-up study. British Medical Journal 333 (7561), 224.

Tschoner, A., Engl, J., Laimer, M., et al., 2007. Metabolic side effects of antipsychotic medication. International Journal of Clinical Practice 61, 1356–1370.

Ventriglio, A., Gentile, A., Baldessarini, R.J., Martone, S., Vitrani, G., La Marca, A., Bellomo, A., 2014. Improvements in metabolic abnormalities among overweight schizophrenia and bipolar disorder patients. European Psychiatry.

Relevant Website

http://professional.diabetes.org/resourcesforprofessionals.aspx?cid=60379 – ADA Cardiometabolic Risk Initiative – (accessed 08.31.14.).

Schmitt, Carl (1888–1985)

Lars Vinx, Bilkent University, Ankara, Turkey

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Carl Schmitt (1888–1985) was a conservative German public lawyer and political theorist. Schmitt is best known for his trenchant critique of liberal constitutionalism, parliamentary democracy, and legal cosmopolitanism. His contributions to debate on these issues continue to attract scholarly attention today, but they are considered to be highly controversial, due to Schmitt's association with National Socialism. This article presents an overview of the key themes and arguments in Schmitt's major writings, and it offers an assessment of Schmitt's relevance for contemporary political, legal, and constitutional theory.

Biographical Sketch

Schmitt published his first academic works in the field of legal theory in the last years of the Wilhelmine Empire (see for Schmitt's biography: Bendersky, 1983; Balakrishnan, 2000; Mehring, 2009). The most productive period of Schmitt's career, however, was the era of the Weimar Republic (1919–33), which saw the publication of most of Schmitt's major political-theoretical works. In these works, Schmitt often presented himself as a defender of the Weimar Constitution, though on the basis of a rather authoritarian and highly contestable interpretation of that constitution that was rejected by the majority of German public lawyers at the time. In particular, Schmitt championed the claim that the popularly elected president of the Weimar Republic possessed an all but unlimited power of dictatorship, under article 48 of the Weimar Constitution, which would allow him to act as a 'guardian of the constitution' and to legislate and govern, in a situation of constitutional crisis, without any parliamentary support (Dyzenhaus, 1997: 70–85; Kennedy, 2004: 154–178).

While Schmitt seems to have been opposed to the prospect of a National Socialist government before 1933, he quickly sided with the Nazis after Hitler's accession to power and came to be perceived as the 'Crown Jurist' of Nazi Germany (Mehring, 2009: 304–436). While some scholars have argued that Schmitt's willingness to support the new regime was born of his ambition and opportunism, and thus did not indicate an ideological affinity to Nazism (Bendersky, 1983: 195–242), there can be little doubt that Schmitt held strongly anti-Semitic views and that his constitutional ideas of the Weimar period were open to be adapted to the new system (Dyzenhaus, 1997: 85–101; Scheuerman, 1999). Schmitt's active involvement with the National Socialist system, however, was cut short in 1936, when he was accused by academic competitors with better National Socialist credentials of being a turncoat who had sided with the new government only to advance his career.

During the later years of the Nazi period, Schmitt therefore turned his attention increasingly to questions of international law. While Schmitt did attempt to provide a justification for Nazi Germany's claim to hegemony on the European continent, he also developed a general account of the conditions of a stable international legal order.

After the end of the war, Schmitt was briefly detained and interrogated by the allied powers, who considered the possibility of putting Schmitt to trial in Nuremberg. Though nothing came of these plans, and Schmitt was set free, he was not allowed, due to his involvement with the National Socialist regime, to return to his academic position after 1945 (Mehring, 2009: 438–463). Schmitt, nevertheless, remained a highly influential figure in West Germany's conservative intellectual scene until his death in 1985 (Müller, 2003). He continued to publish occasionally and to informally supervise young academics.

Schmitt on Sovereignty and Dictatorship

In his early *Political Theology* (Schmitt, 2005), Schmitt famously defends the continuing relevance of sovereignty, both to legal theory and to political practice.

Liberal constitutionalism, according to Schmitt, aims to banish sovereign authority from political life, while it declares the concept of sovereignty to be theoretically dispensable. Its goal is to subject all exercises of political power to the rule of law. But in Schmitt's view, liberal constitutionalism overlooks that the applicability of general legal norms that are to control the exercise of political power presupposes a situation of social normality. No law, Schmitt argues, can be applied to a situation of utter chaos or disorder. Rather, such application requires a 'homogeneous medium,' a situation of social normality, so that the application of the norm will lead to predictable results. A state of emergency or of exception, on the other hand, cannot be governed by legal norms, because a legislator cannot foresee what measures may turn out to be necessary to pacify and normalize an abnormal situation of grave crisis (Kahn, 2011; Croce and Salvatore, 2012). Schmitt concludes that a political community, or rather someone who can act on its behalf, must be endowed with the power to completely suspend the application of the law, and then to use dictatorial measures to restore a state of normality. This conclusion underpins Schmitt's famous claim that 'sovereign is he who decides on the exception' (Schmitt, 2005: 5).

Schmitt is not referring here to an established constitutional competence. What he argues, rather, is that sovereign authority is manifested in the successful exercise of the power to take a decision on the exception so as to create the conditions of the applicability of law. Schmitt holds that any legal order, even one where the constitution does not endow anyone with an

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61279-5

authority to declare a total exception, is based on a decision (to suspend the legal order or not to suspend it), and not on a norm. The fact that the constitutional façade of a modern, liberal-democratic political system typically does not show a sovereign authority, Schmitt thinks, is irrelevant to this point. The true sovereign will appear in a situation of crisis.

Schmitt concedes that a defense of sovereignty for the contemporary world cannot return to the age of absolutism, and portray sovereign authority as vested, by divine right, in a particular person or family (Schmitt, 1985: 22–32, 2008a: 101–111). The doctrine of sovereignty must, therefore, be translated into a democratic ideological context. It is hard to see, though, how such a translation would work. In a democratic context, after all, sovereignty ultimately rests with the people. But the people, it seems, will be able to take collective decisions only once constitutional rules have been put in place that allow them to speak with a united voice. The people, it would appear, are not a suitable bearer of Schmittian sovereignty who takes a decision on the exception.

Schmitt attempted to solve this difficulty through his conception of sovereign dictatorship, which he put forward in his seminal study on the history of the institution of dictatorship (Schmitt, 2014; McCormick, 1997: 121–156; Cristi, 1998: 108–125).

Though originally distinct, dictatorship and sovereignty, Schmitt argues, nevertheless, entered into an intimate relationship in the course of the French Revolution. The French Revolutionary governments employed extralegal, dictatorial force not merely – as the Romans had done – to restore the condition of social normality presupposed by an existing constitution, but so as to create a new situation of normality that would allow a new constitution to come into force. However, the French Revolutionary governments did not claim to be the sovereign. Rather, they took themselves to exercise sovereign authority in the name of the people.

Schmitt's systematic claim here is that democratic constituent power, the power of the people to give itself a new constitution, can only be actualized in the form of sovereign dictatorship: The sovereign people by itself, as an unorganized mass, cannot purposefully act outside the law to create a new situation of normality that will allow for the establishment of a new constitution. It needs to be led by a person or group of persons exercising dictatorial power in the name of the people. But this person or group is dependent on the trust of the people, on the people's willingness to be led and to accept the decisions taken and measures implemented as its own. Sovereign authority, then, still exists in a democratic state: it is the power to take the decision on the exception, so as to establish a new constitutional condition, insofar as it is exercised, and successfully, by appeal to the will of the people (Schmitt, 2008a: 109–110, 265–266).

The Concept of the Political and the Critique of Liberalism

Schmitt's conception of democratic sovereignty assumes that a people or political community can exist prior to or outside of an established framework of constitutional rules. Schmitt's famous claim that politics is based on a distinction between friend and enemy (Schmitt, 2007: 26; Kennedy, 2004: 92–118; Slomp, 2009: 21–37) may usefully be understood as an attempt to explicate the way in which a political community can exist outside of any legal or constitutional framework.

Schmitt emphasizes that the distinction between friend and enemy is essentially public, not private. Politics involves social groups that face off as mutual enemies. Two groups are in a situation of mutual enmity, if and only if, there is a possibility of violent conflict between them. The distinction between friend and enemy marks the 'utmost degree of intensity … of an association or dissociation' (Schmitt, 2007: 26). The utmost degree of association, in other words, is the willingness to fight and to die for a group one identifies with, while the ultimate degree of disassociation consists in the readiness to kill others for the reason that they are members of a hostile group. A political community, then, differs from other kinds of community in that it disposes off the life of its members. Hence, wherever people are members of several kinds of community at the same time, political community will remain the decisive community, and it will typically manage to keep all disagreements between its members below the threshold of violent conflict.

Every political community, in Schmitt's view, requires a substantive collective identity. People who unite into political community must agree to rely on some commonality between themselves as the mark of inclusion in (or exclusion from) the community. But Schmitt does not think that there is any particular kind of difference of identity that is necessarily political. Neither does he hold, consequently, that the distinction between friend and enemy must always be based, in whole or in part, on some particular kind of difference. All kinds of differences of identity – be they ethnic, religious, linguistic, economic – can turn into the substance of a political difference if they turn out, in particular historical context, to have the power to sort people into mutually hostile groups (Schmitt, 2007: 25–27). It would therefore be naïve, Schmitt thinks, to hold that the political will disappear once the conflicts arising from a particular difference of identity have been solved or become irrelevant. History suggests, rather, that political conflict will simply latch on to another difference of identity (Schmitt, 2007: 80–96).

Finally, Schmitt argues in the *Concept of the Political* that a group can be said to exist as a political community only as long as it is self-determining (Schmitt, 2007: 45–53). The members of a group that is to qualify as a political community must take the friend–enemy decision that differentiates their community from other polities by themselves and for themselves. If the friend–enemy decision is taken for a group, but not by the group, but rather by some external power, the group no longer exists as a political community. It follows from this that to deny a group the right to draw the friend–enemy distinction as it sees fit, for example by subjecting it to some binding international regime for the solution of conflict, is to deny the political existence of the group. As long as politics as such is considered to be legitimate, Schmitt concludes, any political community must be entitled to reject such an assault on its political existence.

Schmitt's analysis of the political is at least partly descriptive. The claim is that human beings, throughout history, have shown a tendency to form political communities based on a friend–enemy distinction. A group's ability to take that

distinction, Schmitt holds, does not presuppose that the group is already formally organized, that it already possesses a system of explicit constitutional rules for the creation of a common will. A political community, rather, can have a common existence prior to all legal form as long as its members identify with the substance of the group's identity strongly enough to be willing to fight and die for it. As long as a political community in this sense exists, it will be capable to support a sovereign dictatorship exercised in its name (Schmitt, 2008a: 126–135).

Though partly descriptive, Schmitt's conception of politics has clear normative implications. These are made explicit in Schmitt's critique of liberalism.

In Schmitt's view, liberalism's main characteristic is that it attempts to avoid genuinely political decisions (McCormick, 1997; Dyzenhaus, 1997: 58–70). According to liberals, it is not necessary or desirable for individuals to form groups constituted by friend–enemy distinctions. Liberals hold, rather, that all conflicts among human beings can, in principle, be solved through amicable compromise, as well as through the improvement of civilization, technology, and social organization. Consequently, liberalism is unable to provide markers of identity that can become the basis of political community. While liberalism aims to domesticate political power, in the name of individual freedom, it lacks the power to constitute political community (Schmitt, 2007: 69–79, 1985: 33–50).

This problem, Schmitt thinks, manifests itself in the inability of a liberal state to draw a distinction between friend and enemy. A liberal state will tend to recognize as citizens, and to offer legal protection to, all those who live law-abidingly under the rules laid down in some formal constitution. In doing so, however, liberal states are likely to extend membership to individuals who do not truly identify with the political community that underpins the state. Hence, there is a danger that, in a liberal system, the aims of the political nation will come to be frustrated by internal enemies who abuse the protection afforded by constitutional legality. Eventually, Schmitt, fears, the political nation will slowly disappear, as a result of a process of depoliticization encouraged by liberal ideology (Schmitt, 2008b: 65–77).

It is the job of the sovereign dictator, in Schmitt's view, to prevent such political disintegration, by means of a dictatorial suppression or elimination of internal enemies in a state of exception. In thus protecting the substantive homogeneity of the people, the sovereign dictator creates the condition for the legitimate applicability of the law. The application of constitutional law in nonhomogeneous society, Schmitt argues, can only lead to the legalized suppression of one social group by another (Schmitt, 2008b). Liberal protections of individual freedom and of rights of political association are to be welcomed, therefore, only once all citizens subscribe to a homogeneous political identity.

Though Schmitt expresses confidence that the political will not disappear any time soon, he appears to admit that a completely depoliticized global liberal utopia is possible. The claim that liberalism intends a depoliticization of the world, even if true, thus raises the further question as to why one should judge that the realization of a liberal utopia would be undesirable.

Schmitt offers a number of different responses to this challenge. A first answer emphasizes the Hobbesian claim that the authority of a state is dependent on its power to protect citizens. A state that has suffered a subversion of the political, due to the spread of liberal ideology, may turn out to be incapable to mobilize the coercive force that is needed to protect its subjects, either from an indirect rule of pluralist interest groups who have colonized the state or from external enemies seeking to subdue the political community (Schmitt, 2007: 37–45, 51–53). Schmitt also argues, secondly, that a life without politics would be shallow and meaningless. A completely depoliticized world would offer human beings no other identity than that of a producer and/or consumer. In such a world, there would no longer be any values that would license risking one's own life, and thus give a meaning to one's existence that transcends the satisfaction of private desires (Schmitt, 2007: 35, 57–58). Finally, Schmitt intimates that the willingness to distinguish between friend and enemy is a theological duty (Schmitt, 2005; Meier, 1998). Liberalism, in this reading, is to be rejected because it purports to offer a secular redemption from man's fallen and corrupted nature. But such a promise, as well as the attempt to realize it, is to be rejected, Schmitt appears to suggest, as a rebellion against God, since it is God alone, but only at the end of history, who will deliver humanity from political enmity.

At any rate, Schmitt argues that the political is unlikely to disappear as long as there are at least some who are willing to stake their existence on the attachment to a political identity. Wherever there is one political community, it will force other communities to choose a political existence or to accept subjection. To avoid the latter fate, Schmitt warns, Germans must remain willing to live politically. His analysis of the political at times shades over into a plea for living politically. Schmitt is not just a theorist, but also a partisan of the political.

Schmitt's Theory of Democracy and Constitutional Theory

Democracy, literally, is the self-rule of the people. Or as Schmitt prefers to put the point: a democratic political system must be characterized by the fact that the ruling will is identical to the will of the ruled (Schmitt, 1985: 25–30).

In political practice, of course, such democratic identity is merely presumed, for instance, when the decision of a majority (or of a majority of representatives) is regarded as the decision of the people as a whole. One must therefore ask for the conditions under which the identification of the will of the majority with the will of the whole is justified. Schmitt's answer, unsurprisingly, is based on his analysis of the political. Schmitt claims, as we have seen, that every political community is based on a constitutive distinction between friends and enemies. The equality of democratic citizens, he concludes, must not be confused with the general equality of all human beings. It is an expression, rather, of a shared substantive identity (Schmitt, 2008a: 255–267).

A legitimate identification of the will of the majority with the will of all, then, presupposes that all those who participate in a democracy share one and the same substantive identity. It is only under this condition that one can demand of the members of an outvoted minority that they accept the majority's choice as their own. Schmitt goes on to argue that the ideal of democracy

is not as intimately tied to majoritarian decision-taking as we are inclined to believe (Schmitt, 1985: 28–32). Forms of collective decision-taking other than majority voting may well turn out to be expressive of an identity of ruler and ruled. A populist dictatorship claiming to exercise the sovereignty of the people, Schmitt argues, need not be any less democratic than a system of majority rule, if the dictator and his followers indeed share a political identity. A system of majority voting, on the other hand, whether direct or representative, may turn out not to be democratically legitimate if it is not expressive of an underlying identity. In that case, the rule of the majority will degenerate, Schmitt holds, into an illegitimate and potentially oppressive rule of one social faction over another.

Above all, Schmitt is emphatic that democracy must not be identified with parliamentary government (Schmitt, 1985). The principle of legitimacy of parliamentary government, Schmitt believes, is not democracy, i.e., the identity of ruler and ruled, but rather the view that parliamentary deliberation, in front of an interested and educated bourgeois public, is most likely to lead to correct or sound legislative decisions that cater to the common interest. Schmitt is convinced that this assumption does not apply to modern mass-democracies, where parliaments are controlled by tightly organized parties that represent firmly entrenched social interests. Since parliament has become the scene of a pluralist disintegration of the political community, its decisions are not likely to be expressive of a shared interest or political identity of the whole.

The general effect of Schmitt's reflections on democracy is to dissociate the ideal of democracy from the institutional forms – majority rule and parliamentary government – with which it is most commonly associated, and to link it up with exercises of sovereignty in a state of exception. This revisionist understanding democracy informs Schmitt's constitutional theory and his interpretation of the Weimar Constitution (Dyzenhaus, 1997: 38–101; Scheuerman, 1999: 61–84; Kennedy, 2004: 119–153).

Schmitt's *Constitutional Theory* argues that a democratic constitution is not to be regarded as a social contract, but rather as a unilateral determination, on the part of an already existing political community, of the concrete institutional form of its political existence. The constituent power of the people, i.e., its power to choose for itself any constitution it sees fit, is inalienable, since the people, as long as it exists as a political community, always retains the ability to engage in renewed exercises of constitution-making (Schmitt, 2008: 75–77, 125–130, 140–146; Kalyvas, 2008: 79–186).

Of course, it would make little sense to claim that any written constitution is ever a choice of the people down to its last detail. A people, Schmitt thinks, can do no more than to express its preference for a certain basic constitutional structure, which will then have to be filled out and concretized by the sovereign dictator acting in its name. But it would be wrong, Schmitt holds, to think that the particular norms of a written constitution possess the same normative force as the people's original choice for a certain kind of political system. Even where a written constitution that implements the people's choice for a certain kind of political regime does not impose any material limits on the constituted power of constitutional amendment, it would still be unconstitutional, Schmitt argues, to use that power to transform the very identity of the polity. Such an action, even if it was supported by a supermajority of members of parliament, would amount to usurpation, on the part of constituted powers, of the constituent power of the people (Schmitt, 2004: 85–94, 2008: 77–82, 147–158).

Before 1933, Schmitt relied on this argument to oppose a formally legal Nazi seizure of power. However, Schmitt's constitutional theory did not provide a principled defense of the liberal and democratic constitutional system of the Weimar Republic. While Schmitt denied the constitutionality of changing the fundamental nature of the constitution from within, he, nevertheless, endorsed the possibility of fundamental constitutional change through sovereign dictatorship. Hence, his constitutional theory deliberately left open the possibility of a renewed exercise of constituent power that would sweep away the liberal and parliamentary democracy of the Weimar Republic.

The ambiguity of Schmitt's position toward the Weimar Constitution is equally visible in his interpretation of the powers of the president under the Weimar Constitution (Schmitt, 1931, 2014: 180–226). By assimilating the powers of the president under article 48 of the constitution to those of a sovereign dictator, Schmitt defended an unusually extensive interpretation of presidential authority that subjected most constitutional rights to the discretionary interference of the executive headed by the president. At the same time, Schmitt vehemently opposed the establishment of a constitutional court for the protection of the Weimar Constitution. Such a court, Schmitt argued, would have to limit itself to uncontroversial cases or it would have to act as a constitutional legislator that illicitly claims the power to determine the political identity of the people.

Schmitt seems to have hoped, during the terminal crisis of the Weimar Republic, that presidential dictatorship in defense of the existing constitution would be able to recreate a situation of normality that would allow the constitution to come to function properly again (Kennedy, 2004: 154–183). But Schmitt's theory of presidential guardianship of the constitution is at times difficult to distinguish from a call for an authoritarian transformation of the constitution. The project, at any rate, failed when the Nazis managed to take power through the kind of abuse of constitutional procedure, Schmitt had warned against.

It did not take Schmitt very long to throw in his lot with the new regime, and to offer his services as a legal advisor. Schmitt's readiness to support the Nazis has often been explained as the result of mere personal ambition and opportunism. However, Schmitt's constitutional theory clearly did allow for an interpretation of the Nazi *Machtergreifung* as a valid exercise of constituent power. It thus had the potential to portray Hitler's rule as perfectly legitimate. Hence, it seems unnecessary to postulate a deep discontinuity between Schmitt's views before and after 1933 (Dyzenhaus, 1997: 82–101; Scheuerman, 1999: 113–139).

Liberal Cosmopolitanism and the Foundations of International Legal Order

Schmitt's conception of the political implies two conditions of the legitimacy of international order (Hooker, 2009; Slomp,

2009). Every true political community must claim an unrestricted *ius ad bellum* or right to decide whether to go to war to protect what it sees as its vital interests. If a group no longer takes the distinction between friend and enemy by itself – if that distinction is taken for the group, for instance, by a hegemon or an international organization – the group in question no longer exists as a political community. Hence, an international order that acknowledges the legitimacy of politics must recognize the *ius as bellum* of all groups that have successfully constituted themselves as political communities. Second, Schmitt's conception of the political implies that a group can only exist as a political community if there is at least one other political community that is its enemy. Since the friend–enemy distinction must be based on a difference in substantive identity, politics depends on the availability of several identity-defining forms of social and political life. As a result, an international order that makes room for the political must acknowledge that there are several legitimate forms of social and political life.

These two demands may appear to be a recipe for anarchy, as they seem to rule out an international legal order that successfully regulates the use of force between political communities and that imposes minimal human rights standards on the several members of international society. Schmitt's reply to this objection is to argue that attempts to suppress the *ius ad bellum* and to impose standards of moral conformity on the internal governance of members of international society will only lead to greater disorder and violence than we can expect to see in an international order that is built on a recognition of the political.

Schmitt's argument for this claim proceeds by way of historical example (Schmitt, 2003). Traditional European public international law, the *ius publicum Europaeum*, did not distinguish, according to Schmitt, between just and unjust war. Rather, it acknowledged that every European state had the right to decide to go to war on the basis of its own judgment of justice and necessity. In interstatal wars, both parties were by default recognized as legitimate belligerents. This practice, Schmitt argues, had a number of beneficial consequences that restrained the destructive force of war. The refusal to apportion blame for war made it easier for states to make peace. States not directly involved in a conflict, moreover, were free to adopt a posture of neutrality or to side with either belligerent. This freedom allowed states to contain conflicts by balancing or simply by staying out of the fight. Finally, the mutual recognition of legitimate belligerency permitted states to agree, on the level of the *ius in bello*, on stringent constraints on the permissible means of warfare.

This containment of war, Schmitt argues, was contingent on the willingness to bracket the question of whether wars had a just cause or not (Schmitt, 2011: 30–74). If one adopts the view that a war is usually legitimate or just on one side, but illegitimate or unjust on the other, one is forced to conclude, Schmitt argues, that it must be morally wrong to grant the status of legitimate belligerency to both parties. But in that case, it will no longer be possible to hold on to the view that third parties have a right to side with either belligerent or to remain neutral. One will have to hold, rather, that they are obliged to side with the belligerent that fights justly. Moreover, it will become more difficult to make lasting peace after a war if one side is to be punished for having waged an unjust war. Most importantly, the refusal to accept mutually legitimate belligerency will put pressure on the protections afforded by the *ius in bello*. If one party to a violent conflict is perceived to fight unjustly, it will seem legitimate to use more destructive methods against it than would otherwise appear permissible. Any attempt to create an international order based on a 'discriminatory concept of war' that will subject a political community's decision to use force to substantive criteria of legal or moral legitimacy, Schmitt concludes, is only going to undercut the containment of war achieved by the *ius publicum Europaeum*.

In the *Nomos of the Earth* (Schmitt, 2003), Schmitt tried to answer the question how a regime of the containment of war might come to be constructed on a global level. Groups that are related by mutual enmity can come to coexist in a shared framework that limits the destructive consequences of war, Schmitt proposes, if their enmity does not reach the level of absolute enmity. Enmity, in other words, must take a form that does not require the complete destruction of the enemy's political or even physical existence. Absolute enmity can be prevented, Schmitt holds, if friend–enemy distinctions are aligned with territorial boundaries. If the forms of life of two opposing groups are each tied to and expressed in a particular territory, the friend–enemy distinction between them will have become spatialized. As a result, the defense of a group's political existence can take the form of a defense of its own territory. It will not require the group to interfere with the internal organization of some other polity or to eliminate the latter's political existence.

To spatialize the friend–enemy distinction it is necessary, however, that all and only the people who share one and the same political identity live in the same territory. Some political identities resist such a spatialization of the political. A political community, for instance, whose identity is based on the promotion of liberal-humanitarian values that it takes to be universal, is unlikely to accept a reduction of political conflict to territorial conflict. It will want to interfere where another polity's internal order fails to live up to those values. A stable international order that limits the destructive consequences of political conflict will thus remain out of reach, Schmitt believes, as long as the world's foremost powers are committed to ideologies that imply a rejection of the spatialization of conflict. The way forward to the construction of a durable international order, Schmitt concludes, must consist in division of the globe into several hegemonial spheres that are characterized by internal ideological homogeneity and that recognize each other as mutually legitimate belligerents despite their ideological differences (Schmitt, 2011: 75–124).

Schmitt's theory of the basis of international order is closely related to his account of the conditions under which a domestic constitution can function well. The spatialization of conflict requires political communities capable of enforcing internal political homogeneity. Political communities are likely to be unable to enforce internal homogeneity, however, if they have to live in an international environment that lacks a clear spatial order, because it is controlled by powers committed to universalistic ideologies. Legitimate domestic and legitimate international order, then, are two sides of the same coin for Schmitt (Vinx, 2013). Both require a defense of the political.

It should not occasion surprise that the significance and value of Schmitt's work are hotly contested. But there is a consensus in the growing scholarly literature on Schmitt that Schmitt's work put forward an important challenge to liberal approaches in political and constitutional theory. This challenge deserves to be taken seriously, even while Schmitt's own counterproposals are often unattractive. Schmitt's work clearly offers provocative and sophisticated discussion of a number of issues in legal and political theory – such as the nature and limits of emergency powers, the nature of constituent power, the limits of constitutional amendment, just war theory, the legitimacy of international criminal law and of humanitarian intervention – that are still the subject of extensive debate. In all these areas, Schmitt's arguments have the potential to enrich contemporary debate, precisely by going against the grain of received opinion.

See also: Collective Identity; Cosmopolitanism; Democracy: Normative Theory; Democratic Theory; Dictatorship; Geopolitics; Liberalism: Political Doctrine and Impact on Social Science; Parliamentary Government; Pluralism; Rule of Law (and Rechtsstaat); World Republic.

Bibliography

Balakrishnan, G., 2000. The Enemy. An Intellectual Portrait of Carl Schmitt. Verso, London.

Bendersky, J.W., 1983. Carl Schmitt. Theorist for the Reich. Princeton University Press, Princeton, NJ.

Cristi, R., 1998. Carl Schmitt and Authoritarian Liberalism. Strong State, Free Economy. University of Wales Press, Cardiff.

Croce, M., Salvatore, A., 2012. The Legal Theory of Carl Schmitt. Routledge, Abingdon.

Dyzenhaus, D., 1997. Legality and Legitimacy. Carl Schmitt, Hans Kelsen and Hermann Heller in Weimar. Oxford University Press, Oxford.

Hooker, W., 2009. Carl Schmitt's International Thought. Order and Orientation. Cambridge University Press, Cambridge.

Kahn, P.W., 2011. Political Theology: Four New Chapters on the Concept of Sovereignty. Columbia University Press, New York.

Kalyvas, A., 2008. Democracy and the Politics of the Extraordinary. Max Weber, Carl Schmitt and Hannah Arendt. Cambridge University Press, New York.

Kennedy, E., 2004. Constitutional Failure. Carl Schmitt in Weimar. Duke University Press, Durham, NC and London.

McCormick, J.P., 1997. Carl Schmitt's Critique of Liberalism. Against Politics as Technology. Cambridge University Press, Cambridge.

Mehring, R., 2009. Carl Schmitt: Aufstieg und Fall. Eine Biographie. C.H. Beck, Munich.

Meier, H., 1998. The Lesson of Carl Schmitt. Four Chapters on the Distinction between Political Theology and Political Philosophy. University of Chicago Press, Chicago.

Müller, J.W., 2003. A Dangerous Mind. Carl Schmitt in Post-War European Thought. Yale University Press, New Haven.

Scheuerman, W., 1999. Carl Schmitt. The End of Law. Rowman&Littlefield, Lanham, MD.

Schmitt, C., 1931. Der Hüter der Verfassung. Duncker & Humblot, Berlin.

Schmitt, C., 1985. The Crisis of Parliamentary Democracy (E. Kennedy, Trans.). MIT Press, Cambridge, MA.

Schmitt, C., 2003. The Nomos of the Earth in the International Law of the Jus Publicum Europaeum (G. Ulmen, Trans.). Telos Press, New York.

Schmitt, C., 2004. Legality and Legitimacy (J. Seitzer, Trans.). Duke University Press, Durham, NC and London.

Schmitt, C., 2005. Political Theology. Four Chapters on the Concept of Sovereignty (G. Schwab, Trans.). University of Chicago Press, Chicago.

Schmitt, C., 2007. The Concept of the Political. Expanded Edition (G. Schwab, Trans.). University of Chicago Press, Chicago.

Schmitt, C., 2008a. Constitutional Theory (J. Seitzer, Trans.). Duke University Press, Durham, NC and London.

Schmitt, C., 2008b. The Leviathan in the State Theory of Thomas Hobbes (G. Schwab and E. Hilfstein, Trans.). University of Chicago Press, Chicago.

Schmitt, C., 2011. Writings on War (T. Nunan, Trans.). Polity Press, Cambridge.

Schmitt, C., 2014. Dictatorship. From the Origin of the Modern Concept of Sovereignty to Proletarian Class Struggle (M. Hoelzl and G. Ward, Trans.). Polity Press, Cambridge.

Slomp, G., 2009. Carl Schmitt and the Politics of Hostilty, Violence and Terror. Palgrave Macmillan, Basingstoke.

Vinx, L., 2013. Carl Schmitt and the analogy between constitutional and international law: are constitutional and international law inherently political? Global Constitutionalism 2, 91–124.

Relevant Website

http://plato.stanford.edu/entries/schmitt/ – Vinx, L., 2010. Carl Schmitt. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy.

Scholars, Private Lives of

Barbara Louis, University of Minnesota, Minneapolis, MN, USA
Thilo Neidhöfer, Johannes Kepler University of Linz, Linz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Private lives of scholars have received some attention by researchers, which is reflected in a still modest but growing body of literature. Traditionally biographies of great men have provided insights into the personal lives of protagonists predominantly as background information for their professional enterprises. Feminist historiography pointed out the interrelatedness of private and professional spheres as well as the productive roles of women in scientific collaboration. This article will discuss collaborative couples and antinepotism rules in academia as two rather well-researched areas and point out avenues for further research, also taking cues from the history of the natural sciences.

Private lives of scholars in the social and behavioral sciences is an understudied topic and remains wide open for research. Until recently authors typically focused on one person, usually a man, and concentrated their accounts on the protagonist's work. Biographies of great men frequently portrayed the female partners, if mentioned at all, as secondary, for instance in their roles of caring and supporting wives or assistants. Although literature on collaborative couples in the sciences is widely available, many of these books can be categorized as popular science accounts. While such works often pose claims of historical accuracy, they use the private lives, especially the marriages and partnerships, as a literary tool to thrill readers, and in doing so, often reproduce outdated views on gender relations. For instance, they point out the asymmetry between the man as creative and active and the woman as supportive and assisting, or they emphasize distinct spheres of action for men in the public and women in the private realm.

Historians of science, however, have shown that private and professional spheres are inseparable, and furthermore that "domestic arrangements are part of the culture of science" (Schiebinger, 1999, p. 188). Historians have also come to understand science both as a social practice and a collaborative effort. As a consequence, research has expanded from focusing on the lone genius or the big man of science to include women, assistants, family members, and lay persons. It sought to identify the actors involved and to appraise their contributions to the production of knowledge and to the scientific enterprise more generally. As historians have conceptualized gender as a social construct and historical category and originally focused mostly on women, feminist historiography was initially dedicated to including women in historical accounts and restoring them to their deserved historical place. Taken to the extreme, however, one-sided approaches risked to create accounts of great women (Eckstein, 1996; Chiu, 2008). Thus, gender has been increasingly conceptualized as an analytical tool to explore the relational character of the concept and the ways in which meanings of gender are constructed, which results in attempts to consider both genders equally (Scott, 1986, 2010).

Much of the work on private lives of scholars refers to natural scientists. While traditional biographies of big men of science used details of private events and everyday life of scientists to round out narratives of their professional lives, historians have recently cast a more analytical and productive glance at the intersection of private and professional lives. As Donald Opitz pointed out, explorations of the private realm as a site of knowledge production informing scientific endeavors have led to new interpretations of the work of scientists such as Galileo Galilei, Isaac Newton, Charles Darwin, and Louis Pasteur by revealing the psychological, personal, and social influences on their scientific work (Opitz in Lykknes et al., 2012, p. 249). This is linked to an understanding of science as a social and cultural activity and an interest in the dynamics of scientific practice rather than a focus on the successful outcomes only. This article surveys work on two aspects of private lives of scholars that have received attention by historians: first, familial arrangements and collaborative couples and, second, antinepotism rules in academia and their consequences. While the focus is on the social and behavioral sciences, examples from other disciplines serve as models for future research.

Familial Arrangements and Collaborative Couples

The roles of family members, particularly wives, in scientific activities were at the center of historians' attention, as they started to investigate the roles and activities of various actors in scientific collaboration. A long-standing perception based predominantly on published records and a historiography focused on big men of science held that, if wives were involved in their husbands' work at all, they served as secretaries or assistants, or provided a comfortable and undisturbed atmosphere for his scientific work. As historians of science such as Sally Gregory Kohlstedt and Margaret Rossiter explored the ways in which women were involved in scientific work, they revealed that women played more manifold and substantial roles in the production, communication, and application of scientific knowledge than previous literature had suggested (Kohlstedt, 1999; Rossiter, 1982, 1995, 2012). The range of capacities in which women were involved in scientific pursuits include, but are not limited to, researcher, research assistant, coauthor, editor, discussion partner, critic, typewriter, secretary, manager, fund-raiser, public relations organizer, and all the tasks

connected to running a household, taking care of the family, and being a friend and romantic partner. While this research focused on the natural sciences, some social and behavioral scientists were included. About the same time, social scientists started to uncover women's spheres of action in and contributions to their own disciplines (Deegan, 1991; Laslett and Thorne, 1997; Silverberg, 1998).

In order to interrogate systematically the intersection of private and professional lives, scientific couples became a productive unit of analysis that allowed the exploration of scientific collaboration by taking seriously the contributions of both partners in their joint endeavors. The seminal publication *Creative Couples in the Sciences*, edited by Pycior et al. (1996), probed cross-gender collaboration for different variants of collaboration and aimed at restoring women to the historical record, while neglecting, as Opitz et al. have argued, to explore "how collaboration – or, in its own case, degrees of collaboration – might be historicized and explained as products of partners' agency within particular, local contexts" (Opitz et al. in Lykknes et al., 2012, p. 3).

The recent follow-up book *For Better or for Worse?* (Lykknes et al., 2012) expanded the questions and analytical angles of *Creative Couples*. As it can serve as a model for how to approach familial collaboration in the social and behavioral sciences and for private lives of scholars more generally, it offers strategies for further research by covering a wide range of disciplines, geographical locations, time periods, and actors, and by exploring a variety of modes of collaboration between couples. The contributions in this book ask how the meanings of gender roles were negotiated in relation to each other and how these meanings change over time and vary according to historical context. Such inquiry focuses less on the outcomes and assessments of contributions, and more on processes by exploring the dynamics, negotiations, and representations of collaboration.

In *For Better or for Worse?* Per Wisselgren explored three social science couples in Sweden at the turn of the twentieth century within the context of academization of the social sciences and found different types of collaboration that he situated within 'action spaces' at the intersection of social reform and social sciences (Wisselgren in Lykknes et al., 2012, p. 196). With the husbands holding academic posts, Wisselgren found that the wives contributed to their work in varying capacities and intensity, but also worked independently on projects of their own. As opposed to the limited wives-as-assistant model, he identified three ideal types of spousal collaboration. First, the 'two roles model' describes a mode of collaboration in which family and other kinds of work can be combined, with the family having priority. Second, in the 'in the shadow of' type, the wives' interests are subordinate to the ones of their husbands. While wives made significant contributions to the husbands' work, they remained largely unacknowledged. Finally, the 'separate worlds' model describes an arrangement in which the spouses pursue their own interests independently from each other (Wisselgren in Lykknes et al., 2012, p. 209). Wisselgren's analysis broadens the scope of understanding spousal collaboration and the context in which modern social sciences unfolded. It can serve as an example of the manifold avenues open for future research to shed light on the understudied but rewarding topic of private lives of scholars. His research also cautions not to anachronistically accept the boundaries of present-day disciplines or academia more generally as a frame for analysis. Showing how particularly the women were active in social movements outside the academic realm and collaborated on what is now understood as the husbands' work suggests that a wider angle on scholarly activity has the potential to provide more comprehensive and nuanced insights not only into personal lives, but into the creation and application of knowledge.

Eileen Yeo's analysis of the Webbs, the Bosanquets, and the Branfords, three social science couples in Great Britain at the turn of the twentieth century also delineates different models of intellectual collaboration and married life (Yeo in Lykknes et al., 2012). This examination of the gendered division of labor in these marriages exemplifies diverse arrangements ranging from a traditional constellation with the man doing intellectual social science work and the woman being more practically and social-action oriented to reversed gender roles. The husbands' involvement in household issues also varied. These examples highlight not only the different possible arrangements of spousal collaboration and private life, which call for more research. They also accentuate the importance of careful attention to the historical context that conditions the options available to the couples, for example, regarding contemporary gender roles and expectations but also regarding institutional structures of social science and practice. Furthermore, as the professionalization of the social sciences in Great Britain with more extra-academic opportunities followed a different path than in the US where social science strove for academization and rigid objectivity in the early twentieth century, this context needs to be taken into account in order to detect possible spheres or action and configurations of professional and private lives. Moreover, these examples point to cross-country comparisons and transnational approaches as fruitful modes of analysis.

By and large the heterosexual married couple is the unit of analysis of scientific couples' collaboration. In his study of the British socialist, philosopher, and activist Edward Carpenter, Opitz transcended this common unit to explore the modes of collaboration of a homosexual couple. This analysis of an alternative to the normative heterosexual married couple and their social science and activism reveals that the practice and meaning of collaboration are historically contingent and need to be carefully historicized within their social, cultural, and political contexts. In his example, therefore, Opitz suggested the term 'cooperative comradeship' to describe the mode of Carpenter's domestic partnership and collaboration with his partner George Merrill (Opitz in Lykknes et al., 2012, p. 246). Following Opitz' lead to include same-sex collaboration in various domestic arrangements suggests a variety of new research that not only expands the core unit of analysis but through intersections with other historiographical themes could add new perspectives on the lives and work of social and behavioral scientists. Even though same-sex partnerships are often mentioned in passing, systematic explorations of the partners' living and collaboration arrangements are pending. For example, the

sociologist Jessie Taft and her partner of 50 years, Virginia Robinson, substantially shaped the social work program at the University of Pennsylvania, but their lives and work have not been systematically studied from a couple's perspective. Furthermore, relationships do not necessarily have to be lifelong commitments in order to be significant. They can be temporary or change character over time, as exemplified by Margaret Mead and Ruth Benedict, whose personal and work relationship also invites further research.

While collaborative couples in the same discipline or closely related fields are prevalent in literature, cross-disciplinary couples may reveal different dynamics of living arrangements and collaboration. The focus on same-disciplinary collaboration is plausible and reflects both reality and scholarly interest. On the one hand, many partners met as students, or students and teachers, or colleagues in the same department or field. On the other hand, researchers who are interested in a particular discipline are perhaps more prone to select couples from this area (Handler, 2004).

Looking across disciplinary boundaries, however, will very likely prove productive, since research at the intersection of disciplines can open a vantage point for studying cross-fertilization of different fields, of different disciplinary cultures, as well as their interactions. As historians of science have been interested in ideas and methods circulating and passing through various disciplines, being appropriated, modified, and put to use in new ways, a look at cross-disciplinary collaboration may also help to shed light on such dynamics. The lives and work of Louis Guttman, an American mathematician and sociologist, and his wife, Austrian-born geneticist Ruth Halpern, serve as a case in point (Wurzinger, 2002). After World War II, the Guttmans moved to Israel, where Louis established the Israel Institute of Applied Social Research. While Ruth continued her work in genetics, cancer research, and psychology, she also served as the senior scientist at her husband's institute. Both applied Facet Theory to their own research specialties, a concept whose development is commonly attributed to Louis. Their marriage and collaboration provides an opportunity to explore further their personal and professional roles as they pursued their own separate careers, but also collaborated, and shared a family life including three children. Coming from the US and Europe, respectively, and building their careers in Israel, this example promises an exceptionally rich and complex constellation to study, not only because of their different disciplinary, social and cultural backgrounds, but also because they did so in unfamiliar terrain, which potentially requires more intense and explicit reflexion and negotiations than the native society where role expectations and social and professional arrangements are clearer.

Modifying the units of analysis can lead to innovative new research. After the solitary individual of the past, after looking at couples, schools, and laboratory groups, more recently new modes of communication, telecommuting, and increased mobility of scholars has transformed collaboration. As Nancy G. Slack pointed out, these changes are altering both private and professional lives and are leading to more flexibility in collaboration of scholars in research networks without being at the same place physically. Current-day technology and mobility also open up new ways of organizing the balance between family and professional lives. Moreover, gender roles in society have been changing, leading to more egalitarian division of household responsibilities among couples than in the past and potentially resulting in new configurations of lives and careers for both men and women (Slack in Lykknes et al., 2012).

But even further back in history, a wider angle on collaboration yields new and productive insights. Deborah Coen (2007) has shown in her study of the extended Exner family in fin-de-siècle Vienna how private life, politics, and science intersected and, in fact, exerted strong mutual influence. Coen detailed how the activities in the Exners' semiprivate annual summer retreat in the countryside substantially influenced their scientific thinking, which manifested itself in their works in fields such as physics, meteorology, and ethology. While an extensive scientific dynasty like the Exners is rare, scientific families, even on a smaller scale, do invite further analysis. The Odum family comes to mind, with Howard W. Odum, a distinguished sociologist, and his sons Howard T. and Eugene Odum, who became two of the most eminent American ecologists of the twentieth century. The sons were lifelong collaborators and were at least motivated by their father's sociological framework and social activism, which begs the question of intergenerational transmission of (social) scientific thought, culture, and collaboration.

In a recent study of the Swedish social scientists Alva and Gunnar Myrdal, Thomas Etzemüller (2010) explored the profound entanglements and interdependencies of the Myrdals' private and professional lives. In their capacity as social engineers in Sweden they understood their own personal lives as a role model for the modern, social democratic society in whose reform they were involved. In their private lives, which they publicly displayed, they attempted to demonstrate rationality, equality, and control of emotions, which they envisioned as the main pillars of the reformed Swedish society. The negotiations and conflicts between the spouses about equality, collaboration, and the tension between individual liberty and dedication to the shared lives and work reflected the ambivalence of the larger process of modernization in society. This study of the Myrdals stands out as a model for exploring the intersection of the private and professional lives and inspires future research along similar lines.

As in the past most scholars came from educated upper and middle-class backgrounds, access to universities for larger parts of the population and policies like affirmative action in the United States, as well as efforts to increase the number of women and members of minorities in higher education have started to change the demographics of the social and behavioral sciences. This development opens up an entire new universe of experiences from different social and cultural vantage points to explore how family life is organized and how personal and professional lives are negotiated, and to study further how scholarly work is negotiated, practiced, and presented. In the mid-twentieth century, however, as large numbers of women graduated from universities and sought employment at colleges and universities, hiring policies guided by antinepotism rules profoundly affected their lives and careers as the next section explores.

Antinepotism Rules and Their Consequences

Initially implemented to base academic appointments and advancement on merit rather than on personal relationships, antinepotism regulations led to large-scale discrimination of women in particular. Instead of holding full-fledged faculty positions as appropriate for their training and experience, many highly educated women found themselves relegated to low-rank, often unstable positions with little prestige and small salaries. These disadvantages predominantly affected the women themselves, but also bore consequences for their families. A lower total family income is the obvious consequence, but the women's positions were often temporary, geographically scattered, time-consuming, and scheduled at inconvenient times. In addition to exploring the effects of antinepotism rules on women's careers, more research is needed on the consequences for the couples' relationships, for family life including children, for these families' social lives and networks. How did inferior and time-consuming jobs affect the fathers' involvement in family life? What were the reactions of the husbands, who succeeded professionally at the cost of their wives' careers while, at the same time, they valued their wives' academic skills?

Cases of stalled or diverted careers because of antinepotism rules in the United States are increasingly well documented, even though this documentation and analysis remains far from exhaustive. One prominent example is the psychologist Else Frenkel-Brunswik, who is best known for her coauthorship of *The Authoriatian Personality*. Since her husband Egon Brunswik was a faculty member at the Department of Psychology at Berkeley, she could not obtain a tenured position at the same institution. Instead, the University's Institute of Child Welfare became her institutional home, where she was hired as a research associate. In addition to maintaining a productive research agenda, she taught courses at the Department of Psychology, but without being accepted into the faculty. At the University of Kansas, the renowned psychologist Beatrice Wright was barred from teaching, because her husband was a faculty member. Her textbook on disability, however, was used in the classes she was not allowed to teach. Not least because of Wright's case, the University of Kansas reversed its policy and Wright could obtain a faculty position eventually (Rossiter, 1995, p. 141). The sociologist Caroline Baer Rose is another prominent example of the effects of antinepotism rules on women in academia. She was married to Arnold Rose, a sociologist at the University of Minnesota. Before they moved to Minnesota in 1952, both had worked on the research and the manuscript of Myrdal's *An American Dilemma*. Even though the Roses collaborated throughout their marriage and coauthored several books and articles, Arnold's position prevented Caroline from becoming a faculty member at the Department of Sociology. Thus, she held a temporary position in a small college, worked on high school curricula, taught correspondence courses, and found other ways of employment at the University. She was only able to join the faculty and become full professor after Arnold died in 1968 (Howery in Deegan, 1991).

These antinepotism regulations, even though they were applied in varying severity at different institutions, locked out highly qualified women particularly from the most prestigious universities. A study by John Parrish, a professor of labor economics at the University of Illinois, analyzing faculty numbers by field and sex at the 20 leading universities in the United States in 1960 illustrates the gender-segregation at the top ranks of academia. In psychology 5 out of 184 full professors (or 2.72%) were women (for the total including lower ranks, i.e., assistant professors and above, 30 out of 416 faculty members were women, which is 7.21%). In sociology, all of 2 women could be found among a total of 108 full professors, which equals 1.85% (the total faculty was 259, out of whom 12 were women, which translates into 4.63%). The situation in economics was similar: The total of 205 full professors included 4 women, or 1.95% (including lower ranks: 16 out of 435, i.e., 3.68%). One woman held a full professorship in anthropology vis-à-vis 46 men (2.17%), whereas the total faculty of 94 included 3 women, i.e., 3.19%. Political science had 2 full female professors out of 161 (1.24%) as opposed to 6 out of 336 (1.79%) including assistant professors and above. Finally, there was no woman among 44 full professors in geography, whereas one woman could be found in the total faculty sample of 97 (1.03%). (Parrish cited in Rossiter, 1995, p. 128ff.)

The women affected by antinepotism regulations turned to various strategies to deal with this discrimination. As one popular strategy, they took on assignments in different university departments (if allowed) or lower-rank colleges in the region. In doing so, they were at least able to work in the field in which they were interested and trained. Some tried to accumulate so many outside accolades that pressured university administrations to hire or advance them, as it was eventually the case with Beatrice Wright at the University of Kansas. Yet another strategy was not to marry at all during their active time as professionals, as did the archeologists Madeline Kneberg and Thomas Lewis (Rossiter, 1995, p. 125).

Antinepotism rules did not entirely seal off women from positions in higher education, however. Particularly lower-rank and public colleges and universities, as well as newly founded institutions tended to interpret the antinepotism rules more generously than long-standing top-tier universities. For example, both the sociologists Lewis and Rose Coser were offered professorships by the State University of New York Stony Brook, while Luther and Jessie Bernard, also sociologists, were hired by Pennsylvania State University.

By the early 1970s, more and more American universities started to loosen or reverse their antinepotism rules, not least because these rules conflicted with affirmative action policies. While there are indications that nepotism restrictions continue to exist at some colleges and universities, albeit in a more covert manner, it is the term 'dual career couples' that currently garners attention. Academic couples still often fail in their attempt to find appropriate positions for both partners in the same institution or even geographical region, and it is still mostly the women who cannot attain their professional goals. A recent project in Germany explored the internal and external conditions of the lives and work of academics that continue to adversely affect women's employment (Rusconi and Solga, 2011). While this project focused on the current situation, it can serve as a repository of themes to pursue relevant questions

in historical research on private lives of scholars in the social and behavioral sciences.

More systematic study of the intended and unintended consequences of university hiring policies and practices for the women and their spouses and families has the potential to illuminate the intersection of the personal and the professional. As research has focused on North America and more recently Germany, an international comparative perspective also seems worthwhile, as does a transnational one particularly looking at the international migrants in academia. Furthermore, research has focused on academics who, despite having faced discrimination and setbacks, were determined and tenacious enough to stay in or distinguish themselves outside academia. This viewpoint is a variant of the winner's perspective which restricts the subjects under investigation to those individuals who were eventually successful. If sources permit, in order to gain a better understanding of the processes and dynamics involved, future projects should include the lives and careers of persons who got discouraged or had no other opportunity than leave academia for different career paths.

Conclusion

As the body of literature on the private lives of scholars is modest but growing, a variety of emerging avenues for research hold the promise for multifaceted new insights into the history of the social and behavioral sciences and the experiences of scholars. Historical research depends on sources, however, and the lack of material particularly covering the personal realm is often an inhibiting factor for research. In many cases the extant papers equal the published record, which often went through a selective process that weeded out material that was not published or did not lead directly to publications. For the big men of science, records about their private lives, their wives and their shared lives and work are usually scant, and the same is true for the few famous female social and behavioral scientists. Margaret Mead stands out as an exception, since her papers also contain material on her husbands. However, researchers can take their cue from women's history, which has demonstrated how historical narratives can be revisited and revised both by reinterpreting archival material and by consulting previously neglected sources. Additionally, unconventional sources may provide starting points for further research. For instance, the protagonist in Philip Roth's campus novel *The Human Stain*, is modeled after Melvin Tumin, professor of sociology at Princeton University (Flood, 2012).

While finding and presenting new material and case studies is certainly important, it is desirable for further analysis to move toward comparative, transnational, longitudinal, and comprehensive analyses. Investigating various perspectives of private lives of scholars in the context of different cultures, disciplines, historical contextualization, and discourses will significantly enrich the historiography of the social and behavioral sciences and their practitioners.

Bibliography

Biagioli, M., 1993. Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago, IL.

Browne, J., 1996. Charles Darwin: Voyaging. Princeton University Press, Princeton, NJ.

Browne, J., 2002. Charles Darwin: The Power of Place. Princeton, NJ.

Chiu, C.S., 2008. Women in the Shadows. Lang, New York, NY.

Coen, D.R., 2007. Vienna in the Age of Uncertainty. Univ. of Chicago Press, Chicago, IL.

Deegan, M.J. (Ed.), 1991. Women in Sociology: A Bio-Bibliographical Sourcebook. Greenwood Press, New York, NY.

Eckstein, V. (Ed.), 1996. Kultfrauen, second ed. Bollmann, Mannheim.

Etzemüller, T., 2010. Die Romantik der Rationalität. Transcript-Verl, Bielefeld.

Flood, A., 2012. Philip Roth's Complaint to Wikipedia. The Guardian. http://www.theguardian.com/books/2012/sep/11/philip-roth-wikipedia.

Fölsing, U., 1999. Geniale Beziehungen. Beck, München.

Geison, G.L., 1995. The Private Science of Louis Pasteur. Princeton University Press, Princeton, NJ.

Handler, R. (Ed.), 2004. Significant Others: Interpersonal and Professional Commitments in Anthropology. University of Wisconsin Press, Madison.

Kohlstedt, S.G. (Ed.), 1999. History of Women in the Sciences: Readings from Isis. University of Chicago Press, Chicago.

Laslett, B., Thorne, B. (Eds.), 1997. Feminist Sociology: Life Histories of a Movement. Rutgers University Press, New Brunswick, NJ.

Lykknes, A., Opitz, D., Tiggelen, B.van (Eds.), 2012. For Better or for Worse? Collaborative Couples in the Sciences. Birkhäuser, Basel.

Pycior, H.M., Helena, M., Slack, N.G., Abir-Am, P.G. (Eds.), 1996. Creative Couples in the Sciences. Rutgers University Press, New Brunswick, NJ.

Rossiter, M.W., 1982. Women Scientists in America: Struggles and Strategies to 1940. The Johns Hopkins University Press, Baltimore.

Rossiter, M.W., 1995. Women Scientists in America: Before Affirmative Action 1940–1972. The Johns Hopkins University Press, Baltimore.

Rossiter, M.W., 2012. Women Scientists in America: Forging a New World since 1972. The Johns Hopkins University Press, Baltimore.

Rusconi, A., Solga, H. (Eds.), 2011. Gemeinsam Karriere machen: Die Verflechtung von Berufskarrieren und Familie in Akademikerpartnerschaften. Budrich, Opladen.

Schiebinger, L.L., 1999. Has Feminism Changed Science? Harvard University Press, Cambridge, MA.

Scott, J.W., 1986. Gender: a useful category of historical analysis. The American Historical Review 91, 1053–1075.

Scott, J.W., 2010. Gender: still a useful category of analysis? Diogenes 57, 7–14.

Shortland, M., Yeo, R. (Eds.), 1996. Telling Lives in Science: Essays on Scientific Biography. Cambridge University Press, Cambridge.

Silverberg, H. (Ed.), 1998. Gender and American Social Science: The Formative Years. Princeton University Press, Princeton, NJ.

Westfall, R.S., 1980. Never at Rest: A Biography of Isaac Newton. Cambridge University Press, Cambridge.

Wurzinger, C., 2002. Ruth Guttman, geb. Halpern. In: Keintzel, B., Korotin, I. (Eds.), Wissenschafterinnen in und aus Österreich: Leben – Werk – Wirken. Böhlau, Wien, pp. 270–273.

School (Alternative Models): Ideas and Institutions

Achim Leschinsky, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13560–13563, © 2001, Elsevier Ltd.

Abstract

Ever since its beginnings at the turn of the eighteenth to nineteenth century, the modern educational system has been accompanied by alternative schools, the numbers and broad effects of which have varied over the course of the years. This coexistence is by no means based on independence, but is the manifestation of complex reciprocal relations which exist despite fundamental differences in the ideologies of the two approaches. Phases of intensive development in the domain of traditional schools have been mirrored by similar phases of intense activity in the alternative school movement; alternative schools fulfill important functions within the general spectrum of educational possibilities.

Background and Conception

Because proponents of alternative schooling are critical of contemporary manifestations of the traditional school system, the focus of their critique has shifted over time. In general, however, their criticism is based on objections to the structures and functions of the modern school. They do not restrict their demands to a return to earlier historical epochs and structures in which schools still played a minor role. Their objectives are for duress (compulsory education and instruction) to be replaced by freedom; teacher-directed schooling and instruction by self-directed action on the part of the student; the curriculum by the individual needs of the student; competition and rivalry by community; compartmentalization by wholeness and integral methods in instruction; the separation of life and school by an interweaving of the two spheres (e.g., Holzman, 1997). The fundamental criticism of traditional schooling is based on the conviction that this form of education is burdened by a basic contradiction: Although adults – and particularly adults living in modern democracies – are granted dignity, majority, and self-determination, children and adolescents at school are denied these basic rights, albeit on the pretext that they first have to learn how to exercise them. Because it is not possible to recreate the historical conditions which preceded institutionalized education and its inherent ambivalences, however, the critics of traditional schooling have to resort to so-called 'counterschool' initiatives (German: *Gegenschulen*). Although some of these are embedded in a variety of alternative organizations (e.g., flat-sharing communes, communal child care, and even dietetic measures), in general it is not society that is deschooled, but school itself. The extent to which this radical rearrangement leads to the complete disintegration of institutions is dependent on its (partly unintended) consequences.

Structural and Organizational Characteristics

Pedagogical aims are reflected in the specific organizational characteristics of any given school (e.g., the general structure of the school, the relationships between teachers, parents, and students, the organization of classes, the methods implemented). In turn, these characteristics constitute the decisive prerequisites allowing many alternative institutions to exceed the norm with regard to content and in both the pedagogical and the social fields. Indeed, the typical approach of the alternative schools can be reduced to a simple formula: in contrast to the universal institutions of the traditional system, which are characterized by high levels of internal differentiation and complicated organizational structures, alternative institutions are distinguished by social clarity and immediacy owing to the fact that they are, from the outset, geared to the particularities of their respective clients. Accordingly, the student body in an alternative school on average numbers less than 200 (cf Mintz et al., 1994). Thus, irrespective of content-related differences, the alternative school represents a special type of school or, more specifically, a type of school offering a special kind of education. While this offer may formally be an open one, it in fact only reaches the few who know how to take it up.

Due to these invisible processes of selection, alternative schools can be expected to be more or less immune to many of the problems arising in schools which really are open to all (as is made clear by the reports and evaluations of alternative schools; cf as a very vivid example Duke and Perry, 1978). In accordance with the very definition of the alternative school, however, these assets cannot readily be transferred to the mass school system as a whole. Instead of having to simultaneously meet the needs of students from different backgrounds and with various levels of qualification, alternative schools are able to respond to individual differences and developments emanating from a more homogeneous context of attitudes and lifestyles. The additional potential for motivation and identification (in both students and teachers) which is inherent in a free choice of school or in the decision to embark upon a rather unusual experiment constitutes another factor which cannot easily be transferred to the traditional school. As a rule, 'counterschools' are spared the arduous coordination of pedagogical subfunctions on the basis of a division of labor: because of the small size of these institutions, instruction and education, counseling and care all rest in the hands of a few. Instead of having to strictly adhere to general and at least partly formulated rules in order to ensure the smooth running of the

school and its classes, confidence in the closer personal ties and connections of the school community makes it possible to act more effortlessly, easily, and flexibly. Finally, the danger of merely running out of inspiration within the confines of internal school processes is reduced, as the high levels of input from the local environment mean that the system's contours remain rather diffuse. The close links between the home and the school, which are not limited to financial support, but also involve the parents' practical cooperation, can be seen as another important asset.

Developments Since the 1960s

The concepts behind alternative schooling and its implementation reach far back in time, and in some respects they refer to the founders of modern educational theory (e.g., Rousseau or Herbart). Progressive education in the United States and so-called *Reformpädagogik* in Germany, both of which emerged after the end of the nineteenth century, may certainly be considered to be phases of activity to which a number of existing institutions can be traced back (e.g., Semel and Sadovnik, 1999). However, the main wave of alternative school foundation resulted from political and social conditions including the failure or petering out of large-scale attempts at educational reform (e.g., Carnie et al., 1996); this holds especially for the Anglo-American countries since the 1960s. Most of the alternative institutions founded during this phase were set up in the United States, and the majority are still to be found there (cf Mintz et al., 1994). Although less than 5 percent of students at elementary and secondary school actually attend alternative schools, the institutions now number several thousands in the United States, most of them elementary schools. The institutions are highly divergent in terms of their targets, methods, integration of parents and students, links with the traditional system, locus of control, and circles of addressees. It would thus seem justified to focus this overview of alternative schooling since the 1960s on the situation in the United States, all the more so because both the fundamental principle of alternative schools ('as many institutions as there are needs and groups') and current trends in development (adopting the tasks of the traditional system) are to be observed there.

One important form of alternative education is home schooling (cf Evans, 1973), where parents instruct their children at home with the permission of the state. If – as is quite often the case – these children meet up with other home students in (elementary) schools on single days of the week, there is a sliding transition to school cooperatives aiming at unconstrained, playful, spontaneous learning in open instruction. Other institutions specifically target problematic cases, so-called at-risk students, including not only chronic truants and low-performing students, but also talented nonconformers (in the broadest sense). A well-balanced program of drill and freedom is intended to render school attractive again, thus helping these students to enhance their performance.

However, this development has also led to a reverse trend, a sort of back-to-basics shift to formal 'three Rs' schools, which specifically emphasize the basic skills, discipline, and respect for the authority of parents and teachers (cf, e.g., Deal and Nolan, 1978). Similar trends are to be observed at the secondary level. The large number of schools-within-schools – smaller, less complex units running different experimental programs, and affiliated to larger state schools – is particularly notable. The spectrum also includes schools attempting to set themselves up as 'just community schools' based on the principles of Kohlberg's developmental theory (cf Kohlberg, 1985). The so-called 'schools without walls' represent another form of alternative education. They not only offer a 'normal' curriculum (differing from that of the traditional school in terms of content and time), but give students the opportunity to learn in the adult world. This is intended to provide students with extrascholastic experience and to enable them to prove themselves in the world of work. Although this approach initially seemed very promising, experience has shown that it is in fact very hard to put into practice.

Recent Trends and Perspectives

Over the past few decades, there have been signs of a kind of division of function and labor between the traditional and the alternative schools in the United States. This has been expressed in an increase in the state-controlled administration of alternative institutions. The fact that such large numbers of US alternative schools were founded at the beginning of the 1980s is probably related to this development. The radical free schools (e.g., Dennison, 1969), which signaled the upheavals of the 1960s and certainly provided considerable impetus for subsequent developments, play a lesser role in today's broader stream of initiatives. The basic conceptions and expectations of alternative schooling have also been affected, resulting in the extremely short lifespan of some of these institutions. There is still a great deal of fluctuation, making it hard to make valid evaluations of these institutions (cf the data in Mintz et al., 1994: p. 10ff.). Even the early approaches and experiments demonstrated the great weaknesses and dangers of such institutions – the cost of their structure-related advantages (cf also Swidler, 1979). Critics have reduced this to the polemic formula that precisely those principles and premises which allowed the pure 'counterschool' initiatives of the early period to emerge later threatened their ability to survive. Wherever the concept of freedom as a counterpole to the compulsory character of the traditional institutions went hand in hand with systematic shortcomings in the specification of learning targets, there was the risk that later conflicts be all the more severe due to the differing hopes and expectations of those involved (cf Barth, 1972: especially p. 108ff). When every school-related action was dependent on the consensus of every single member of the school community, the school's fate was put in the hands of changing minorities with the power of veto, all the more so when the small size of these private institutions kindled fears of sudden financial ruin. Allowing almost all of those concerned to take an active role in the decision-making process did not in fact ensure the equality of all participants – instead of increasing the capacity to solve problems, there was a growing risk of disappointment, and the necessary basic consensus was threatened by frequent conflicts and failures in minor matters. The firm and wide-ranging commitment of parents, extending well into issues of classroom practice, often

resulted not so much in the enhancement as in the disturbance of school activities. Teachers were consequently made to feel insecure and were put on the defensive instead being provided with support and control. Attempts to simultaneously cope with the factual demands, organizational affairs, and personal standards inherent in these institutions – preferably without referring to preset guidelines – led to the systematic overexertion of those concerned, and the end of many 'counter-schools' was marked by the burnout of the teaching staff (Dennison described this phenomenon as early as 1969). The unconditional openness with respect to the other members of the school community and the intimacy of the small group responsible for the school meant that factual disputes often escalated into personal trials. Moreover, school affairs were often overshadowed by secondary motives such as parental self-realization or the search for security.

By integrating and promoting alternative schools, the state aims to increase the diversification of the educational spectrum, to enhance the overall efficiency of schools (which has been subject to criticism for decades; cf, e.g., Goodlad, 1984), and to better meet the specific needs of certain problem groups in both school and society. The US programs of community schools, magnet schools and, more recently, charter schools provide evidence for this trend, and are looked upon with growing interest by other countries as possible models for the resolution of system-related problems. However, more detailed analyses of these models point to significant limitations in their problem-solving capacity, limitations which are ultimately rooted in the basic concepts of market and choice (cf Lauder et al., 1999). The community schools program has already been thwarted by the US authorities' desegregation measures: although these schools were directed at the advancement of members of the underprivileged minorities, most of whom are colored, they undermined the desired objective of racial integration. Magnet schools were supposed to be a free choice of school with a special educational program which acted like a magnet, attracting students from beyond the racial boundaries marked out in residential areas and school districts. They were intended to promote racial integration, which can scarcely be achieved through the compulsory measure of busing. A number of objections have been voiced against this school form. For example, it is argued that the enlargement of catchment areas weakens the link between the school and the parental home. Furthermore, although magnet schools may well establish a balanced relation between whites and blacks in their own student bodies, there are too few magnet schools to ensure social compensation on a large scale.

Moreover, there is an imperceptible shift in their circle of addressees: instead of reaching the underprivileged, out-of-school colored population with their specific educational provision, magnet schools in fact seem to appeal to white parents and their children who would otherwise switch to private schools or the more privileged suburbs (cf Smrekar and Goldring, 1999).

The charter schools, on the other hand, which in a way transfer the principles of the private school (i.e., the autonomy of schools and a free choice of school) to the state school system, are faced with comparable problems on a different level inasmuch as unforeseen side effects have been revealed in the social process. Despite the fact that the numbers of charter schools have increased considerably over the past few years, they are still far from ever being able to constitute the majority of state schools, belonging by definition to the optional domain of educational provision. The half-hearted legislative reservations perceptible in most of the federal states against according too many privileges to these schools threaten to render the entire charter program futile (cf Hassel, 1999).

See also: School Business Management; School Choice; School Management; School as a Social System; School to School Collaboration; Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance.

Bibliography

Barth, R.S., 1972. Open Education and the American School. Agathon Press, New York.

Carnie, F., Tasker, M., Large, M. (Eds.), 1996. Freeing Education. Steps Towards Real Choice and Diversity in Schools. Hawthorn Press, Stroud, UK.

Deal, T.E., Nolan, R.R. (Eds.), 1978. Alternative Schools. Ideologies, Realities, Guidelines. Nelson-Hall, Chicago, IL.

Dennison, G., 1969. The Lives of Children. The Story of the First Street School. Random House, New York.

Duke, D.L., Perry, C., 1978. Can alternative schools succeed where Benjamin Spock, Spiro Agnew, and B. F. Skinner have failed? Adolescence 13, 375–392.

Evans, T., 1973. The School in the Home. Harper & Row, New York.

Goodlad, J.I.A., 1984. Place Called School: Prospects for the Future. McGraw Hill, New York.

Hassel, B.C., 1999. The Charter School Challenge – Avoiding the Pitfalls, Fulfilling the Promise. Brookings Institution Press, Washington, DC.

Holzman, L., 1997. Schools for Growth. Radical Alternatives to Current Educational Models. Lawrence Erlbaum Associates, Mahwah, NJ.

Kohlberg, L., 1985. The just community approach to moral education in theory and practice. In: Berkowitz, M.W., Oser, F. (Eds.), Moral Education: Theory and Application. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 27–87.

Lauder, H., Hughes, D., Watson, S., Waslander, S., Thrupp, M., Strathdee, R., Simiyu, I., Dupuis, A., McGlinn, J., Hamlin, J., 1999. Trading in Futures. Why Markets in Education Don't Work. Open University Press, Buckingham, UK.

Mintz, J., Solomon, R., Solomon, S., Muscat, A. (Eds.), 1994. The Handbook of Alternative Education. A Solomon Press Book. Macmillan, New York.

Ramseger, J., 1975. Gegenschulen. Radikale Reformschulen in der Praxis [Radical Reform Schools in Practice]. Klinkhardt, Bad Heilbrunn, Germany.

Semel, S.F., Sadovnik, A.R. (Eds.), 1999. Schools of Tomorrow,' Schools of Today. What Happened to Progressive Education History of School and Schooling, vol. 8. Peter Lang Verlag, New York.

Smrekar, C., Goldring, E., 1999. School Choice in Urban America. Magnet Schools and the Pursuit of Equity. Teachers' College, Columbia University, New York.

Swidler, A., 1979. Organization Without Authority. Dilemmas of Social Control in Free Schools. Harvard University Press, Cambridge, MA.

School Achievement as Moderated by Sociodemographic Status: The US Case

Selcuk R Sirin and Taveeshi Gupta, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

School achievement is greatly moderated by sociodemographic factors, such as socioeconomic status, minority status, students' gender and grade level as well as the location of schools. Based on a comprehensive review of the literature, the effect of socioeconomic status on school achievement is illustrated with empirical evidence.

School achievement has been defined as the degree of students' academic learning. Assessment of the existence, extent, or depth of such learning is conducted by using a number of measurement tools across research, policy, and interventions; the most commonly used indicators are grades in math, science, and reading, standardized achievement tests, and IQ tests. Although high correlations are reported among these various measures of academic performance, they also capture a unique aspect of individual student learning. For instance, achievement tests are specifically designed to measure efficiently the amount of knowledge and/or skill a person has acquired at school using a comparison with a standard or norm that is usually based on an expected level of mastery in a content area. Often these tests are used to measure and compare students, schools, districts, and states, based on norm-referenced scales of achievement. Similarly, IQ tests are intended to assess general cognitive competence, rather than measuring learning in a particular area of schooling.

While these are the most widely used ways to assess school achievement, many problems have been found with the use of these tools. A comprehensive review of literature using the meta-analysis technique compared the average correlations between socioeconomic status (SES) and school achievement as measured by different types of assessment tools (Sirin, 2005). The measures of assessment include verbal achievement, math achievement, composite achievement, grade point average (GPA) for all subjects, and IQ scores. The findings from this meta-analysis show that most studies used verbal achievement measures, followed by math achievement, IQ, composite achievement, and GPA scores. Although IQ tests are sometimes included as a measure of academic competence, there has been an ongoing dispute about the use of IQ tests in educational research. Opponents of using IQ tests have identified evidence of a bias against various disadvantaged groups, such as minorities and women. Using grades as the measure of achievement can also be quite problematic, particularly when making comparisons across studies, because grades awarded by teachers are usually considered more subjective than standardized achievement tests.

School Achievement and SES

The relation between SES and school achievement is one of the most widely replicated findings in educational research. SES refers to one's position in society as it is determined by a variety of factors including income, education, occupation, and accumulated wealth (Bollen et al., 2001). Conceptually, the social and economic conditions of children provide the physical, social, and intellectual background for effective learning (Coleman et al., 1966). It is important to note that while social class refers to a more unidimensional construct, SES refers to a more multidimensional conceptualization of social and economical characteristics (Bollen et al., 2001). SES plays two roles in the educational process both of which may exert influences on school achievement, both directly and through associated behaviors and experiences. The first aspect of SES includes resources, such as parental education, income, and wealth and the second aspect of SES includes 'social capital,' i.e., social status as a function of relative positions in a hierarchy, such as social class. While the first aspect to a great extent dictates life conditions for students by determining the share and quality of life to which they have access, the second aspect provides membership in a certain social class that creates its own personal socialization processes, through which the schooling behaviors are acquired (Coleman, 1988). The family's level of educational values, expectations, and school-related activities in the home is partly determined by their SES. In other words, families from different SES groups create different learning environments that affect the child's academic achievement. Hence, the impact of SES on school achievement is not solely due to the amount of parental income or parental education, but also due to social capital that is composed of ideas, behaviors, attitudes, values, and language.

Overall, educational research has been consistent in its finding that SES has an impact on students' school achievement (Coleman et al., 1966; Sirin, 2005). Coleman et al.'s (1966) report ('The Equality of Educational Opportunity'), is widely considered the most significant education study of the past century. James Coleman, the head of a large team of scholars from various fields was commissioned to write this report by the US Office of Education in accordance with the Civil Rights Act of 1964. Using data from over 575 000 students and their teachers at 4000 schools across the United States, the team examined how differences in various types of physical, human, and social resources across schools related to average school achievement levels.

In their final report (which is usually referred as *the Coleman Report*), Coleman and his colleagues presented four major findings: (1) for all ethnic groups, the strongest predictor of achievement was the family's SES (parental education, parental

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23102-4

income, etc.); (2) for students who were not White, the SES of the school as assessed by its percentage of White students and the average economic background of all students, was the second best predictor of school achievement; (3) teacher characteristics assessed by educational degree and teaching experience was a modest but significant predictor of school achievement only for the school performance of African-American students living in the South; and, (4) once all the above characteristics were controlled, school-based economic characteristics, such as per-student spending and instructional resources, had little or no effect on school achievement (Coleman et al., 1966).

Although the Coleman report made it clear that the social composition of the school, the student's sense of control of his or her environment, and the verbal skills of the teachers were relatively related to school achievement, the biggest contribution of this report is its first finding. The report's overall message that family SES is the most important predictor of academic performance has challenged educational researchers and policy makers ever since its publication. At a time when civil rights were on top of the agenda for many, the report's findings that academic achievement was less related to the quality of a student's school and more related to the student's family background was received with controversy, as it suggested that inequities in economical resources allocated to schools had little or no influence on school achievement.

Many scholars criticized the report and pointed out a number of methodological shortcomings of the study (see Buchmann and Hannum, 2001; Murnane, 1975). A major area of dispute was that the study failed to portray accurately the way schools operate, by conceiving of schools as a kind of a closed production facility with school and family resources at one end and student achievement at the other. Therefore, the study did not sufficiently assess the possibility that school achievement might vary between schools as a function of student characteristics, such as race, gender, or urbanity. Students within the same school may not share equal access to the school resources such as tutoring, library, guidance services, or access to the most experienced teachers; and they usually experience difficult educational experiences due to certain groupings based on their characteristics. Further, studies showed that in poorer countries, the impact of the school and the teacher on achievement was greater than family characteristics (Buchmann and Hannum, 2001). Despite these shortcomings, the Coleman Report nevertheless boosted research on SES and schooling processes (Wong and Nicotera, 2004).

Sirin (2005) conducted the last and most comprehensive meta-analytic study that reviewed the literature on SES and academic achievement, most of which was published after the Coleman Report. The results of this meta-analysis clearly show that the significant impact of SES on student's academic achievement is as strong today as it was when investigated by Coleman et al. (1966). Sirin found that based on 78 independent samples, the overall correlations effect size (ES) was 0.29, which reflects a medium level of impact. This general finding in the Coleman and Sirin studies has been confirmed by many other large-scale surveys. For example, the 1985 National Assessment of Educational Progress (NAEP) report on reading found that 17-year-olds from low-income backgrounds could read only at an elementary school level (a level achieved by advantaged students at age 13). The national results in 2009 from the NAEP show that there were no significant changes from 1992 to 2009 in the reading score gaps between White and Black students or between White and Hispanic students. Likewise, demographic data show that retained students tend to come from lower SES backgrounds than nonretained students. It has been shown that approximately 40% of repeaters come from the lowest SES quartile, whereas approximately 8.5% come from the highest SES quartile. Research also shows that adolescents who come from lower SES families are more likely to drop out of high school and less likely to attend college (Gándara et al., 2004).

Recent research shows that SES is not only 'directly' linked to school performance, but also that its impact is multifaceted. For example, students from lower SES backgrounds are more likely to lack basic academic skills even as early as kindergarten, which puts them at risk for failure in later years of schooling (Levin and Belfield, 2002). They are also more likely to be at risk for poorer health and nutrition, which could affect their ability to concentrate in the classroom (McLoyd, 1988). Students coming from low SES families also tend to have less access to both social and cultural capital that helps them navigate systems crucial to educational success (Wadpole, 2003). Moreover, families from low SES backgrounds are more likely to live in poor neighborhoods, with larger number of family members in a household, with fewer educational and social resources available to both parents and student for help toward academic success (Duncan and Magnuson, 2005).

Minority Status, SES, and School Achievement

Minority status continues to be a critical demographic factor in educational achievement. Recent surveys conducted by the National Center for Education Statistics (NCES) indicate that the achievement gap between minority students and their White peers, measured by various subject tests and the Scholastic Aptitude test (SAT), has remained the same over the last two decades. Research with Latino/a students, for example, have found that Latino/a youth earn lower grades than other ethnic groups and are at increased risk for dropping out of high school. There is also a persistent achievement gap between African-American and White students, and African-American students are under-represented in classes for the gifted, and over-represented in special education classes (Lleras, 2008). Many factors have been suggested to explain the lower academic achievement of minority students, but the major explanation has been the fact that minorities or disadvantaged families usually live in harsher social and economical circumstances. On average, minority students are more likely to live in low-income households or in single-parent families, more likely to attend under-funded schools, and their parents are likely to have less education – all of which are components of SES and linked to school achievement (Brooks-Gunn and Duncan, 1997). Moreover, research shows that minority students may face discrimination and prejudice at school and the value systems of school may be in conflict with some of their cultural values. In other words, race and ethnicity intersect with SES so

that much of the effect that race and ethnicity have on school achievement may be actually due to the influence of SES.

The moderating role of race in the relation between SES and race in relation to school achievement is still being debated in educational research. There is mixed evidence when studies attempt to untangle the effect of ethnicity versus SES on achievement. On the one hand, some studies show that race/ethnicity, and not the family SES of students, has a stronger influence on school performance for minority students. Catterall (1998), using national longitudinal data, found that within African-American cohorts, SES does not have an influence on their school performance. Similar findings were also found in ethnographic investigations which show race as the most consistent and strongest predictor of school achievement above and beyond SES. Moreover, despite recent national surveys indicating that although the gap in the parental educational level of minorities, particularly between Blacks and Whites, narrowed substantially during the last three decades, there is still a continuing achievement gap between minority groups and Whites.

On the other hand, a number of studies have found the opposite – that it is SES and not race/ethnicity that carries most of the variance in school performance among minority students. For example, Blair et al. (1999) investigated achievement differences across ethnic groups using two waves of a national data set and found that SES-based characteristics were the best overall predictors of academic performance above and beyond minority status. Similarly, other studies have found that regardless of their race/ethnicity, middle class students had higher school achievement than White students who came from working and underclass backgrounds. Likewise, Howell and Frese (1982) illustrated that dropout rates for White students are actually higher than those for minorities, after controlling for SES. There is no agreement at this point as to how to interpret these contradictory findings, partly because for minority students in the United States, race/ethnicity and SES are deeply linked.

Sirin's review shows two separate findings that provided empirical evidence for the role of race/ethnicity and SES on school achievement in the United States. First, a comparison of White only samples with minority only samples showed significant differences between the groups. More specifically, family SES was a stronger predictor of school achievement for White students than for minority students. Additional evidence that minority status acts as a moderating factor came from the significant association between the percentage of minority students and the magnitude of the correlations between SES and school achievement. It was seen that as the proportion of minority students increased in a sample, the magnitude of SES-school achievement correlation decreased significantly. Put differently, the more minority students in a sample, the lesser the impact of SES on school achievement. Based on these findings, it is fair to suggest that minority students do not benefit from their family background as much as White students do.

The finding that family background variables such as parental education, income, and occupation are less predictive for minorities should be of concern not only for reasons of future research methodology, but also for social policy implications. Research suggests that neighborhood and school SES may exert a more powerful effect over and above family SES on academic achievement within minority communities, particularly within African-American communities. For example, Gonzales et al. (1996), using a 1-year prospective design, examined the combined effect of family and neighborhood influences on the school performance of African-American high school students. As expected, they found that family SES variables were not as predictive of school achievement, but neighborhood risk factors were related to lower grades and also moderated several other factors such as parenting support and control. In a large scaled examination of the same issue, Dornbusch and Ritter (1991) also compare the differential effects of parental SES and neighborhood SES in relation to school achievement. They showed that the ability of family SES to predict school achievement is minimized when students, regardless of their ethnic background, live in a census tract with a substantial proportion of minority residents. In other words, the reduced influence of family SES on school achievement among African-Americans is not because of their ethnic background, but because most African-Americans and fewer Whites live in neighborhoods with various higher risk factors for school success. Therefore, neighborhoods with disadvantaged communities may lack the resources necessary for school achievement and neighborhood SES is an important factor to be considered.

SES and Gender Differences in Achievement

Sex and gender (analogous to race) have been cited as factors that influence the relation between SES and school achievement. Unlike sex (that is understood to be the biological and physical characteristics of an individual), gender is socially constructed via processes that can vary with SES. In recent years, the recognition of gender differences in school achievement has led investigators to incorporate gender into research on the impact of social and economic factors on the academic performance of students.

In 1992, the report of the American Association of University Women (AAUW) titled 'How Schools Shortchange Girls' pointed out that it has only been in the past two decades that scholarly work on the effect of gender differences in school achievement has become popular. Despite progress in closing the gender gap in many domains, current research shows a number of significant gender differences in school achievement. The majority of the research has focused on differential math performance between boys and girls. It is reported that girls score higher, on average, than boys on tests of verbal abilities; while boys consistently score higher, on average, than girls on tests of science and spatial abilities. One limitation of the research examining gender gaps is that it usually focuses on early adolescence, because studies show that boys and girls have similar achievements in mathematics in the early grades but that boys tend to improve their math scores, relative to girls, as they move through school.

By finding that boys consistently achieve higher scores in math and science and that girls score higher in reading and writing, the research seems to suggest that external social factors are operating beyond biological correlates to explain this difference. The question of whether or not these differences are

changing due to SES received very limited attention in the literature. For example, in line with the observation of the AAUW report that gender was not a salient variable, White's (1982) meta-analysis did not include any variable to account for gender differences. Similarly, in the 2012 Recommendation to the NCES report on 'Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress,' Cowan et al. do not mention gender as a differentiating variable in the measurement of SES.

Therefore, based on separate studies on the impact of gender and SES differences on school achievement, it is plausible to argue that students' gender may moderate the relation between SES and school achievement. In other words, one could argue that the social and psychological pressures experienced by a girl from a lower SES family might be very different from those experienced by a boy who comes from a lower SES family, stemming from the differing cultural expectations operating on a girl or a boy.

Sirin's analysis of gender as a moderator of the relation between SES and school achievement was not significant. Based on independent samples (i.e., all-girls samples and all-boys samples), the impact of SES on school achievement was not significantly different for the two gender groups. Likewise, a second analysis found that the proportion of girls in the sample as a possible moderator of the correlation between SES and school achievement was not significant, indicating that the proportion of girls in a sample was not significantly associated with the magnitude of the correlation between SES and school achievement.

Students' Age, SES, and School Achievement

Yet another important factor affecting the social and economic circumstances of school achievement is the developmental stage of the students. Since the impact of contextual factors is time variant, they lead to questions about the impact of SES on school performance over time. One of the major developmental questions is whether the relation between SES and school achievement varies over the developmental stages, from early childhood to late adolescence. Questions of a developmental nature ask whether the impact of SES that occurs in early schooling is different from the impact of SES in later years at school. Furthermore, scholars examine whether various components of SES (e.g., income, education, and occupation) take on different weights in their relation to school achievement at different ages.

Longitudinal studies that measure SES from early childhood to adolescence have only recently begun to be investigated, so it is difficult to know the answers to these questions from the existing literature (McLoyd, 1988). In one of the few existing studies, Duncan et al. (1994) look at how family SES impacts school achievement over time and found that the preschool years are periods of increased risk for school achievement in poor children. Duncan et al. investigated the Panel Study of Income Dynamics and showed that poverty during the first 2 years of life has a much more negative impact on educational attainment than poverty during middle childhood and adolescence. Similar trends can be seen in the school transition literature which shows that students from lower SES backgrounds present an especially sharp drop in school performance after the transition to junior high school. Likewise, the impact of being economically disadvantaged is seen even within developmental stages. In their study of over 21 000 children from kindergarten to third grade using the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99, Aikens and Barbarin (2008) find that higher SES relates to higher initial achievement in the fall of kindergarten and more rapid reading growth per month across elementary years.

In Sirin's meta-analysis, students' grade level was found to be a significant moderator of correlations between SES and academic achievement with an ES of 0.19 for kindergarten students, 0.27 for elementary school students, 0.31 for middle school students, and 0.26 for high school students. Thus, with the exception of the high school students, there seems to be a trend of increasing ES from kindergarten to middle school. Thus, it can be concluded that the developmental stages moderate the effect of SES on school achievement, at least until middle school.

Single Parenthood, SES, and School Achievement

Family structure has long been considered one of the key aspects of family background that may affect various developmental outcomes of children, including academic performance and achievement. There are two important reasons for considering family structure in the measurement of SES. Firstly, family structure is usually considered as an indicator of social and economic circumstances because single-parent families generally have lower incomes than two-parent families. Secondly, in addition to economical decline in family income, single-parent families also have less resources for parental involvement, which creates another possible link to lower achievement.

Over the past three decades, large-scale national surveys indicate that students who live in single-parent families are more likely to be placed in special education classes, to have lower grades, are less likely to be in school at age 17, and less likely to graduate from high school than students who live in two-parent households. A similar trend was observed in achievement tests; younger students who come from single-parent households score significantly lower on achievement tests than those who come from two-parent families.

Despite the evidence that links single parenthood with low academic achievement, several investigators have moved away from this deficit model and pointed to a shift in the impact of single parenthood on school performance. For example, the national trends between 1970 and 1990 showed that on comparing single-parent households and two-parent households in conflict, some of the characteristics of single-parent households are indeed conducive to school success. Sustained conflict within a family can significantly affect children's development, and children who live in single-parent families resulting from divorce are probably much more likely to have been exposed to this detrimental environment in their original family. Thus, for these children, the transition to a single-parent family may not have direct negative consequences, and, in some cases, might even create a better developmental environment.

Furthermore, the drop in SES due to parental breakup probably does have a negative effect on academic achievement; but because mothers are more educated in recent years as compared with the 1970s, the impact of single parenthood on children is not as detrimental as once assumed. Therefore, once other family differences are accounted for, the lack of a direct effect on achievement from being in a single-parent family cannot be extrapolated to imply that children who live in a nonconflictual or positive two-parent environment would do as well in a single-family environment.

Sirin's meta-analysis showed that the proportion of single parents in a sample does not significantly moderate the magnitude of the relation between SES and school achievement. The results of his study, therefore, do not support the notion that family structure moderates the impact of SES on school achievement. Despite the fact that these results were based on a small number of studies, the insignificant finding is still worth noting because the current literature on the impact of single parenthood is inconclusive. On the one hand, many studies in the past reported that students reared by single parents tended to do worse in school, while on the other hand, analysis of national data and other studies found no negative impact of the single-parent background, particularly after controlling for family SES. In fact, Duncan and Magnuson (2005) suggest that interventions that are designed to increase parents' SES (by improving single-parent education levels) will provide perhaps a modest increase in achievement for children. Instead, they call for policies that directly target children's aptitude or mental and physical health. Thus, the conclusion that families headed by single parents were a source of disadvantage is brought into question by findings from the meta-analysis.

School Location, SES, and School Achievement

One of the moderators of the effect of SES on school achievement that is tested by researchers is that of school location. Data from the NAEP, for example, indicated that the achievement of children in affluent suburban schools was significantly and consistently higher than that of children in 'disadvantaged' urban schools (U.S. Department of Education, 2000). Clearly then, the location of a school is closely related to the social and economic conditions of its students. Therefore, there is some indication that rural–urban–suburban difference may actually be another indication of differences in students' SES background.

'Location, Poverty and Schools,' a study by the U.S. Department of Education (1996) showed that SES is an important factor in accounting for differences in the achievement gap across urban, rural, and suburban schools districts (e.g., the higher the levels of poverty, the lower the levels of achievement). Urban children were more likely to be living in poverty than those in suburban locations – 20% compared with 11% in 2011. Therefore, unless SES was held constant, no findings of rural–urban difference are meaningful on their own.

Interestingly, even after accounting for SES, there seems to be a number of significant differences between urban, rural, and suburban schools. Urban schools on average have larger enrollments but less funding for their schools than suburban schools. Moreover, research suggests that students at urban schools are more likely to present higher rates of student behavior problems such as absenteeism, classroom discipline, and student pregnancy than nonurban schools. Since urban schools are also more likely to cater to immigrant populations, students in urban schools are also more likely than those in other locations to have difficulty in speaking English and are subject to numerous health and safety risks that present greater challenges to their school performance. Beyond student level, teacher absenteeism was more of a problem in urban schools than in rural or suburban schools. Finally, urban schools had fewer resources available for teachers and schools and teachers had less control over their curriculum than suburban or rural schools, at both the elementary and secondary levels.

The variable of geographical location of schools in the studies reported in Sirin's meta-analysis was found to be a significant moderating factor. Students who go to urban schools seem to have higher disadvantage in terms of the impact of family background on school achievement than students in suburban schools. The difference between the correlations across the groups was significant and large, indicating that the geographical location of the school is a strong moderator on the impact of family SES on school achievement. One of the reasons why geographical location seems to have such a strong effect on the impact of SES on school achievement is due to the fact that the location of the school largely determines the amount of financial resources available for students. In a nationwide study of over 17 000 school districts, investigators have found that higher neighborhood SES, as measured by the value of owner-occupied housing or by residents' educational attainment, is significantly related to greater school expenditure per student. In addition to this nationwide analysis, several statewide surveys showed the same phenomenon, in which differential resources were available for schools in different locations. For example, an examination of data from Virginia shows that resources are associated with a school district's SES characteristics and that resources were related to students' academic achievement. Such discrepancies in funding allocation continue to plague the educational system even today. These findings further illustrate that the relation between SES and school achievement is moderated by geographical location.

The finding that the impact of family SES on school achievement was the weakest for urban schools compared with nonurban schools should be interpreted in the context of the economical desegregation of people. According to U.S. Department of Education's (1996) report, many school outcome measures were largely the same across urban, rural, and suburban schools, once the percentage of students in poverty was controlled. This is an important factor to consider when interventions are aimed at improving child outcomes nationally.

Conclusions

The quest to assess the relationship between SES and school achievement produced a wealth of research, from which it is clear that SES continues to influence students' performance in

school. The observed magnitude of this influence, however, depends on several sociodemographic factors. Minority status, age, and family structure of the student, all affect the correlation between SES and school achievement. Finally, contextual factors as determined by the geographic location of the school and the percentage of minorities in the school, change the magnitude of the relation between SES and school achievement.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Data Bases and Statistical Systems: Education, General; Defiant Behavior During Adolescence across Cultures; Ethnicity and Educational Achievement; Family and Schooling; Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments; School Achievement: Motivational Determinants and Processes; Social Inequality and Schooling; Socioeconomic Status and Academic Achievement.

Bibliography

Aikens, N.L., Barbarin, O., 2008. Socioeconomic differences in reading trajectories: the contribution of family, neighborhood, and school contexts. Journal of Educational Psychology 100 (2), 235–251 doi:10.1037/0022-0663.100.2.235.

American Association of University Women, 1992. How Schools Shortchange Girls: A Study of Major Findings on Girls and Education. AAUW Educational Foundation, The Wellesley College Center for Research on Women, Washington, DC.

Blair, S.L., Blair, M.C., Madamba, A.B., 1999. Racial/ethnic differences in high school students' academic performance: understanding the interweave of social class and ethnicity in the family context. Journal of Comparative Family Studies 30 (3), 539–555.

Bollen, K., Glanville, J.A., Stecklov, G., 2001. Socioeconomic status and class in studies of fertility and health in developing countries. Annual Review of Sociology 27, 153–185.

Brooks-Gunn, J., Duncan, G.J., 1997. The effects of poverty on children. The Future of Children 7 (2), 55–71.

Buchmann, C., Hannum, E., 2001. Education and stratification in developing countries: a review of theories and research. Annual Review of Sociology 27, 77–102. doi: 0360-0572/01/0811-0077.

Catterall, J.S., 1998. Risk and resilience in student transitions to high school. American Journal of Education 106, 302–333.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. Government Printing Office, Washington, DC.

Dornbusch, S.M., Ritter, P.L., April 18, 1991. Family Decision-making and Authoritative Parenting. Oral presentation. Society for Research on Child Development, Seattle, Washington, DC.

Duncan, G., Brooks-Gunn, J., Klebanov, P., 1994. Economic deprivation and early childhood development. Child Development 65 (2), 296–318.

Duncan, G.J., Magnuson, K., 2005. Can family socioeconomic resources account for racial and ethnic test score gaps? The Future of Children 15, 35–54.

Gándara, P., O'Hara, S., Gutiérrez, D., 2004. The changing shape of aspirations: peer influence on achievement behavior. In: Gibson, M.A., Gándara, P., Peterson Koyama, J. (Eds.), School Connections: U.S. Mexican Youth, Peers, and School Achievement. Teachers College Press, New York, pp. 39–62.

Gonzales, N.A., Cauce, A.M., Friedman, R.J., Mason, C.A., 1996. Family, peer, and neighborhood influences on academic achievement among African-American adolescents: one-year prospective effects. American Journal of Community Psychology 24 (3), 365–387.

Howell, F.M., Frese, W., 1982. Early transition into adult roles: some antecedents and outcomes. American Educational Research Journal 19, 51–73.

Kao, G., Thompson, J., 2003. Racial and ethnic stratification in educational achievement and attainment. Annual Review of Sociology 29, 417–442.

Levin, H.M., Belfield, C.R., 2002. Families as Contractual Partners in Education. UCLA Law Review 49, 1799–1830.

Lleras, C., 2008. Race, racial concentration, and the dynamics of educational inequality across urban and suburban schools. American Educational Research Journal 45 (4), 886–912.

McLoyd, V., 1988. Socioeconomic disadvantage and child development. American Psychologist 53, 185–204.

Murnane, R.J., 1975. The Impact of School Resources on the Learning of Inner-City Children. Ballinger, Cambridge, MA.

National Center for Educational Progress, 2009. Grade 12 Reading and Mathematics 2009. National and Pilot State Results. Institute of Educational Sciences.

National Center for Educational Statistics, 2012. Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress: A Theoretical Foundation. Institute of Educational Sciences. Review 49, pp. 1799–1824.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. Review of Educational Research 75 (3), 417–453.

U.S. Department of Education, National Center for Education Statistics (U.S.), 1996. Location, Poverty and Schools. U.S. Government Printing Office, Washington, DC.

U.S. Department of Education, National Center for Education Statistics, 2000. The Condition of Education 2000, NCES 2000-602. U.S. Government Printing Office, Washington, DC.

Wadpole, M., 2003. Socioeconomic status and college: how SES affects college experiences and outcomes. Review of Higher Education 27, 45–73.

White, K., 1982. The relation between socioeconomic status and academic achievement. Psychological Bulletin 91, 461–481.

Wong, K.K., Nicotera, A.C., 2004. Brown v. Board of Education and the Coleman Report: social science research and the debate on educational equality. Peabody Journal of Education 79 (2), 122–135.

School Achievement: Motivational Determinants and Processes

Friedrich-Wilhelm Schrader and Andreas Helmke, University Koblenz-Landau, Landau, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Among the many factors that influence school achievement, motivation plays a major role. Motivational approaches are classified using a framework based on expectancies and values. The role of some important motivational theories (self-related concepts, attribution, interest, goals and goal orientations, avoidance) to achievement is outlined, followed by two integrative models (Eccles' expectancy-value model, process models of learning, and self-regulation). The focus is on motivational constructs that aim at explaining learning and achievement in schools. For most motivational constructs, effect sizes from meta-analyses are reported. Finally, complex relations among motivation and achievement are discussed.

School achievement (SA) – its structure, determinants, correlates, and consequences – has always been a major issue of educational psychology, especially effectiveness of teaching and learning. Interest in achievement as a central outcome of schooling has considerably increased during the last two decades. In particular, large-scale projects like Progress in International Reading Literacy Study, Trends in International Mathematics and Science Study, and Program for International Student Assessment had and still do have a strong and enduring impact on educational policy, as well as on educational research, and have also raised the public interest in scholastic achievement. Among the many determinants of SA, motivation is presumably the one that is given the most attention by laypersons, teachers, and parents. In educational psychology, much research has been done on motivation and many motivational constructs that are important for understanding learning processes were explored. But few studies, especially longitudinal studies that include different student characteristics and indicators of classroom context as predictors of achievement, have examined the direct links of motivation to SA. Recently, syntheses of meta-analyses of factors contributing to achievement were published by Hattie (2009, 2012). This work includes motivational factors and so can give an overview of the empirical relations between motivation and achievement. The aim of this article is to address the role of motivational variables with respect to scholastic achievement. We refer to achievement as a result of individual learning processes in school classrooms, and consequently, focus on research that is especially useful for classroom learning.

SA and Its Determinants

SA refers to cognitive learning outcomes of students that are products of instruction or aimed at by instruction within a school context. Cognitive outcomes mainly comprise not only procedural and declarative knowledge but also problem-solving skills and strategies. Student achievement is measured by tests or by teacher judgment, i.e., grades. Research on SA is typically faced with a multilevel problem, that is, for conceptual and statistical reasons, the individual student's level has to be separated from the classroom level, school level, and even higher levels (Helmke and Weinert, 1997). In this contribution, we focus on achievement at an individual level.

Achievement as a result of classroom learning is typically a product of cumulative learning, i.e., an accumulation of specific learning processes and opportunities during extended periods of time – for instance, a school year (Helmke and Weinert, 1997). Research on determinants of achievement, then, deals with the question if and to what degree long-term achievement can be predicted by variables that are assumed to influence achievement, with context, instructional processes, and individual student characteristics such as intelligence, prior knowledge, or motivation as predictors. One line of inquiry is based on large-scale classroom studies and longitudinal designs and is correlational. Motivation is often measured by questionnaires that have been developed for different constructs, and improved measurement has led to many advances in this field (Graham and Weiner, 2012). Training studies, which are experimental or quasi-experimental, constitute another type of research. While this kind of research often takes place within classroom, basic research on motivation that seeks to clarify motivational processes and mechanisms as well as mediational processes relating motivation to learning processes has used laboratory studies. Results summarized in meta-analyses mostly deal with the individual difference perspective, whereas models and theories of motivation often focus on processes and mechanisms.

Achievement is multidetermined, which means that it is influenced by many different factors simultaneously (Helmke and Weinert, 1997). **Figure 1** shows a theoretical model that represents important determinants of SA. Individual motivational determinants include not only task-specific factors such as expectancies but also broader goals and values. Motivational factors are embedded in a complex system of individual, parental, and school-related determinants and depend on the given social, classroom, and cultural context. As shown by this model, individual characteristics have a direct impact on learning activities and achievement, whereas the effect of the other determinants, such as teachers' or parents' influences on learning and achievement, is only indirect and mediated by processes at the level of the individual learner.

Cognitive determinants such as intelligence and prior knowledge account for high portions of individual differences in SA, with an overall effect size of 0.67 (Hattie, 2009). Intelligence and prior knowledge represent the individual determinants that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26055-8

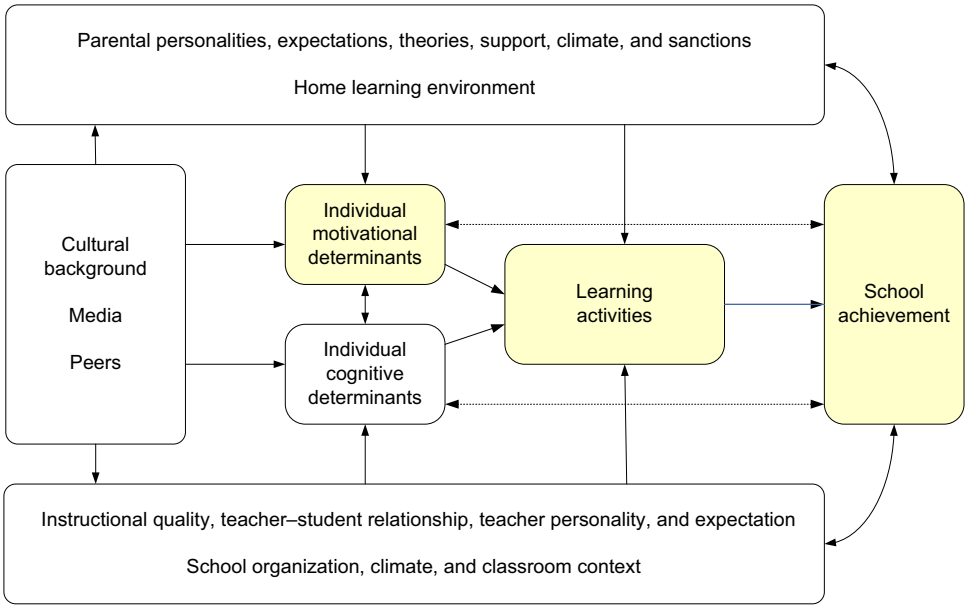

Figure 1 Interplay of individual motivational and other determinants of school achievement.

show the strongest associations with achievement, and so may serve as a reference. With respect to educational outcomes, an effect size of $d = 0.2$ can be considered as small, $d = 0.4$ as medium, and $d = 0.6$ as large (Hattie, 2009: p. 9). Cognitive factors are strongly intertwined with motivational factors, such as subjective competencies and beliefs, and a combination of cognitive and motivational factors characterizes complex processes such as self-regulation and learning strategies.

Motivation and Scholastic Achievement: A General Framework

Motivation deals with why a behavior occurs, and, therefore, deals with the factors that initiate, direct, energize, maintain, and terminate a behavior (e.g., Graham and Weiner, 2012). Expectancies and values are constituent elements of many kinds of human motivation. A person's propensity to choose a behavior depends on the value the consequences of this behavior have for the person, and the expectation or subjective probability that these consequences are met. Different approaches can be subsumed under two basic questions a person may ask before deciding to act: (1) Can I do the task? This question refers to the expectancy component that covers self-concept and related constructs (such as self-efficacy, subjective competence). (2) Do I want to do the task? This question deals with the value of a task (task values) and addresses the importance or personal significance of consequences. Following Eccles and Wigfield (2002) and others, we use expectancies and values as a basic framework to give an overview over motivational constructs that are relevant for achievement. In addition, we present two integrative approaches: (1) expectancy-value models that combine expectancies and values, and include antecedent conditions and achievement; (2) process models of learning activities and self-regulation.

There is considerable conceptual overlap among different theoretical approaches of motivation. Graham and Weiner (2012) characterize motivational approaches and constructs in more depth by using additional questions. A basic distinction is between extrinsic and intrinsic motivation ('Why am I engaging in this activity?'). Whereas many theories of motivation focus on behaviors that are instrumental for certain incentives or goals, the behavior in itself may have intrinsic value. Attribution theory covers causal explanations ('Why did I succeed or fail?'), which are often raised after finishing an action. Emotions ('How do I feel?') are important constituents of motivation and represent motivational factors on their own (see Pekrun et al., 2007, for an overview over emotions related to achievement). Because emotions are broad constructs that require a more detailed presentation, we do not include them in this article.

Motivation refers to factors that influence decision-making and forming of intentions. But in order to achieve a goal, motivation is not sufficient. Decisions and intentions have to be transformed into actions, which often are complex behaviors extending over long periods of time as is characteristic for learning activities. In this case, some kind of external or internal regulation (planning, monitoring, feedback) is required. From the standpoint of an individual learner, self-regulatory or volitional processes ('How am I doing?') should take place. Theories that combine motivation and cognition and include self-regulation and volition as factors supporting effective goal attainment represent a more complete framework (Eccles and Wigfield, 2002).

Constructs Relating to Expectancy

Self-Concept

Self-related cognitions characterize the interplay between cognitive abilities and beliefs (see **Figure 1**) and, therefore,

play an important role in the domain of education. Self-concept is a general and comprehensive construct that comprises cognitions and beliefs about one's abilities and academic achievements (academic self-concept, self-concept of ability), but also social and physical characteristics. Self-concept has an affective component, commonly termed as self-esteem, and a behavioral component, which corresponds to self-efficacy. According to Marsh et al. (2012), a distinction between the cognitive and the affective component does not seem to be fruitful in empirical research. In present conceptualizations, self-concept is often conceived as hierarchically ordered, with general self-concept or self-esteem at the top, self-perceptions in different domains at an intermediate level, and self-perception of specific task behaviors within a domain at lower levels. Whereas social and physical self-concept is important for psychological health and well-being, self-concept of ability (largely equivalent with academic self-concept, subjective competence, achievement-related self-confidence, expectation of success) is relevant to learning and achievement.

The relation between self-concept and achievement is well established. In meta-analyses, effect sizes ranging from 0.32 to 0.41 are reported (Hattie, 2009, 2012). Correlations between domain-specific self-related cognitions and achievement are substantial whereas achievement is unrelated to general self-esteem and nonacademic components of self-concept (Marsh et al., 2012). Correlations are especially high, when self-concept is conceptualized as domain-specific, and when pupils are older (Helmke and Weinert, 1997). What is, yet, controversial is the causal ordering between self-concept and achievement. Research based on longitudinal assessment and structural equation modeling of both self-concept and achievement has demonstrated that self-concept depends on former achievement (skill-development model) but also predicts subsequent achievement (self-enhancement model) (Helmke and van Aken, 1995; Marsh et al., 2012). That is, self-concept is a motivating factor, but also reflects the impact of prior achievement on beliefs and cognition. Therefore, achievement and self-concept should be considered as mutual reinforcing influences. Furthermore, both variables are important success criteria of schooling, so, from a practical view, ameliorating both of them is an important goal (Marsh et al., 2012).

Self-Efficacy

Self-efficacy, which is theoretically and empirically related to self-concept, has a clear motivational focus and, though also influenced by former achievement, predicts subsequent achievement. Self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments, i.e., to solve a problem or accomplish a task. Self-efficacy is a task-specific construct, relating to a specific situation, and not a general expectation. How much it may generalize across different tasks and situations is an open question (Bong and Skaalvik, 2003). With respect to achievement, meta-analyses report effect sizes of 0.37 (Hattie, 2009), and correlations with academic achievement in the range of 0.49–0.70 reported when self-efficacy corresponds closely to the criterion task (Schunk and Pajares, 2009).

Self-efficacy is assumed to influence choice of activities, effort, and persistence. Self-related cognitions may influence scholastic performance by means of various mechanisms: Students with a high self-concept of ability initiate learning activities more easily and more quickly, show more persistence and less procrastination behavior, are more apt to continue learning when a task is unexpectedly difficult, and are better protected against interfering cognitions such as self-doubt and other worries (Helmke and Weinert, 1997).

But the self-concept–achievement relation may depend on classroom context. Weinert and Helmke (1987) found that self-concept predicted achievement of 5th graders when teacher support was low, but lost its predictive power when teacher support was high. This is interpreted as a kind of functional compensation between self-concept and instruction.

Implicit Theories of Intelligence

The impact self-related cognitions have on learning and achievement depends on the view persons have about the nature of their capabilities, especially their general intelligence but also their domain-specific abilities (Dweck and Master, 2009). Implicit theories either constitute an entity theory or an incremental theory, both having different motivational consequences. A person with an entity theory believes that intelligence or ability is a fixed attribute that cannot be changed and task performance reflects this ability. As a consequence these students are concerned about their intelligence. When involved in tasks, they are motivated to demonstrate their intelligence by successful performance (performance goals). Failure suggests a lack of intelligence, and students wish to avoid such a situation, which may threaten their self-esteem. Effort, which is not a salient causal factor for these students, is considered as useless, or even worse, as damaging because failure in spite of effort suggests low ability. An incremental view of intelligence means that intelligence is seen as changeable and open to improvement by effort. Students with an incremental theory of intelligence consider tasks as an opportunity to learn and improve their abilities (learning goals). Dweck and Master (2009) discuss several consequences of these different motivational tendencies with respect to students' goals, learning opportunities, effort, reactions to failure, self-esteem, and performance. Implicit theories of intelligence seem to have an effect at critical points in one's life when new demands arise and increased effort is required. Blackwell et al. (2007) demonstrate achievement gains for students with an incremental theory of intelligence over 2 years of junior high school whereas students with an entity theory do not gain.

Attribution

Attribution theory that can be traced back to the classical model of Weiner (Graham and Weiner, 2012) deals with causal factors

to which success and failure are attributed. These are internal (ability, effort) and external (task difficulty, luck) causes, which may be more or less stable or controllable. As a consequence, it is psychologically favorable when failure is attributed to controllable factors such as lack of effort, and unfavorable when failure is attributed to stable and uncontrollable factors such as ability. Attributions of success and failure influence expectations for success and failure in future. In control theory, a basic distinction is between internal and external locus of control. Meta-analyses show effect sizes of 0.23 and 0.36 for internal locus of control and of 0.56 for success versus failure attributions with respect to achievement (Hattie, 2009).

Constructs Relating to Values

Motivation to Learn and Attitude toward Learning

Motivation to learn is a general construct that comprises subjective expectations as well as incentive values (anticipated consequences like pride, sorrow, shame, or reactions of significant others, the incentive of learning action itself, i.e., interest in activity). For the impact of motivation to learn on achievement, a moderate effect size of 0.34 has been reported. Attitude toward learning means the affective (negative or positive) aspect of the orientation toward learning that corresponds to the value component of learning motivation. Attitude can refer to school or school subjects. For attitude to school subjects, meta-analyses have demonstrated effect sizes of 0.32 for science and of 0.29 and 0.47 for mathematics (Hattie, 2009). There are several more or less interrelated concepts that are associated with the subjective value of the domain or subject under consideration, the respective activities, or on a more generalized level, teachers and school.

Intrinsic Motivation and Interest

Intrinsic motivation is defined as motivation to engage in a task or activity for its own sake whereas with extrinsic motivation engagement occurs for instrumental reasons, such as reward (Linnenbrink and Pintrich, 2002). Intrinsic motivation theories comprise self-determination theory, flow theory, and interests (Eccles and Wigfield, 2002). Self-determinations theory is a theoretical approach that specifies conditions leading to intrinsic motivation. Intrinsic motivation is promoted when people feel competent, self-determined (autonomy), and experience a sense of relatedness (Graham and Weiner, 2012). Flow is the immediate subjective experience of fluent and mentally absorbing processes that occurs when people are engaged in an activity that can be mastered (Eccles and Wigfield, 2002). Studies that explore that actual role of intrinsic motivation for SA are rare. Gottfried (1985) reported moderate correlations between academic intrinsic motivation measured by a self-report inventory and test achievement and teacher grades for reading, language, and mathematics in grades 4–8.

Interest can be considered as a special kind of intrinsic motivation that is related to specific objects, subject areas, or activities (Schiefele, 2009). A basic distinction is between individual, or personal, and situational interest, which corresponds to the well-known distinction between trait and state. Individual interest is a general and relatively enduring affective-evaluative orientation toward certain subject areas or objects, situational interest is a temporary state aroused by specific features of an object. Individual interest comprises feeling-related and value-related valences. The former relates to the feelings that are associated with an object or activity, the latter to the personal significance or importance of an object or activity. As summarized by Linnenbrink and Pintrich (2002), personal interest is positively associated with achievement, the use of deeper cognitive strategies, and increased attention and persistence. For the relation between interest and achievement, an effect size of 0.65 has been reported (Hattie, 2009). Köller et al. (2001) found in a longitudinal study that interest had a significant effect on the development of mathematical achievement from grade 10 to 12 but not from grade 7 to 10. That is, motivation played a role in higher grades when instruction is less structured than in earlier grades. This result illustrates the role of context for motivation.

Goal Orientations and Goals

Achievement goals refer to the purpose why persons engage in achievement behavior. Mastery, learning, or task-involvement goals are distinguished from performance, relative ability, or ego-involvement goals (Maehr and Zusho, 2009). The former are characterized by a tendency to develop one's own competencies and skills, the latter by a tendency to compare one's achievement with that of others and outperform other persons. Mastery goals often are seen as favorable to learning, and performance goals as detrimental (Maehr and Zusho, 2009). But the role of achievement goals is more complex, depending on achievement indicators (grades, test scores), age, and other moderators. In many studies performance goals were related to higher grades, especially for older students (Linnenbrink-Garcia et al., 2008). Furthermore, many researchers considered it necessary to distinguish between performance approach and performance avoidance goals (three-factor model), or to combine mastery and performance orientations with approach and avoidance tendencies (four-factor model). As was demonstrated in a recent meta-analysis based on 151 studies, the overall relation between goals and achievement is rather weak: ranging from $r = 0.13$ for mastery goals in the traditional two-factor model to $r = -0.13$ for performance avoidance goals in the four-factor model (Huang, 2012).

Beyond general orientations like mastery and performance goals, individuals strive for a diversity of goals (Boekaerts, 2009; Eccles and Wigfield, 2002) that compete for limited resources. Little is known about goal coordination and the conditions under which alternative goals are supportive or debilitating. Goal conflicts can arise, for example, between academic goals and leisure time goals. Different cultures differ in the degree goal conflicts play, which are especially high in individualized Western countries (Hofer et al., 2010).

Avoidance Motivation

For many students, achievement is aversive or, at least, ambiguous. Achievement motivation comprises two general

motives: Hope for success and fear of failure. Self-related cognitions play an important role: Students expect failure, because their subjective competence is low (or because they are uncertain about their competence); and they fear failure, because failure has negative consequences for their self-esteem and the worth they have for persons such as parents, teachers, or peers. The self-worth theory of achievement motivation (Covington, 1984) views ability and effort as causal factors for achievement. Ability is a highly valued attribute and is inferred from successful performance that is achieved with low effort. On the other hand, effort is a risk because failures, in spite of high effort, lead to the inference that ability is low. To avoid these results behaviors like procrastination or self-handicapping may result.

Integrative Approaches

Eccles' Expectancy-Value Model

One of the most elaborate models is Eccles' expectancy-value theory (Eccles, 2005). In this model, expectation of success and subjective task value which directly predict achievement-related choice and performance are linked to anteceding factors. Proximal antecedents include goals, self-schemata, and self-concept. Distal factors are stable person characteristic such as gender and aptitudes or cultural milieu. Four components of task value are distinguished: attainment value (i.e., the perceived importance of doing well), intrinsic value (i.e., how much enjoyment the individual derives from the task), utility value (i.e., how the task relates to future goals), and costs (i.e., what the individual has to give up to engage in a task) (Eccles and Wigfield, 2002). Eccles's model has been tested intensively with a focus on achievement related choice (career choice, course plans, enrollment decisions). Whereas choice is mainly predicted by the value component, achievement seems to be directly related to performance expectancies (Wigfield et al., 2009). A further important result is that expectancy and value judgments are positively associated (Eccles and Wigfield, 2002).

Contrary to the multiplicative model originally used in motivation research, Eccles' expectancy-value model is an additive model. In a recent classroom study using a large sample of 2.508 students and complex statistical techniques Trautwein et al., (2012) not only confirmed that expectancy beliefs were a stronger predictor for achievement than value beliefs, but also found significant interactions between expectancies and values. These results suggest that the full model including a multiplicative termin could be more appropriate than the additive model although a proof of interactions is difficult in most cases.

Process Models of Learning and Self-Regulation

The influence of motivation on achievement is mediated by learning activities with engagement and concentration as indicators of effective learning. Meta-analyses for engagement in science and concentration on achievement with effect sizes of 1.09 and 0.61, respectively (Hattie, 2009) demonstrate the significance of these factors.

Managing one's own learning activities is indispensable for higher learning processes and the acquisition of a rich and elaborated knowledge base. Self-regulation comprises different approaches focusing on cognitive, metacognitive, and motivational processes, and is characterized by a cyclical structure that is based on a goal, self-initiated behaviors toward that goal, feedback, and reactions depending on feedback.

Zimmerman and Labuhn (2012) distinguish three phases: a forethought phase (task analysis, goal setting, strategic planning), a performance phase (task strategies, self-instruction), and a self-reflection phase (self-judgment, monitoring). These processes are influenced by specific forms of motivation. During the foresight phase, key motives are goal orientation, self-efficacy, interest, and task values; during self-reflection, causal attributions and self-evaluative judgments such as self-satisfaction and affect come into play. Effect sizes for training programs related to self-regulation such as teaching metacognitive strategies (0.69), study skill (0.59), self-verbalization, and self-questioning (0.64) are at least moderate (Hattie, 2009). But in some cases large effect sizes could be demonstrated. For self-regulated interventions for math, reading and writing with primary and secondary school students effect sizes that are in a range from 0.87 to 1.44 are reported by Zimmerman and Labuhn (2012). But in school learning, learning activities always represent some mixture of external regulation and self-regulation, especially for younger and less able students (Corno and Snow, 1986). As Figure 1 shows, teachers and instructional processes as well as parents' expectations and behaviors influence learning. Furthermore, classroom teaching, and, to a minor degree, parental support constitute contexts that differ in structure and so offer different opportunities for self-regulated learning (Helmke and Weinert, 1997).

According to Boekaerts and Corno (2005), volitional aspects are emphasized in some models of self-regulation. To understand why some people – in spite of sufficient motivation – fail to transform their learning intentions into correspondent learning behavior, approaches that include volition based on theoretical work by Heckhausen, Kuhl, Gollwitzer and others (see Heckhausen and Heckhausen, 2008) have proven to be helpful. Action control refers to the ability to protect intentions against competitive tendencies, and action versus state orientation is a disposition that relates to individual differences to regulate cognitions, emotions, and behaviors in order to accomplish intended actions. While direct links between action-state orientation and achievement have rarely been explored, some studies could show that implementation intentions were associated with academic achievement (Hattie, 2009). But relations between volitional constructs and achievement are complex and mediated by many intervening processes so that more complex modeling is required.

Conclusions

Meta-analyses show that associations between motivation and achievement are moderate, sometimes rather weak, contrary to popular views of the role of motivation. But these moderate

associations may underestimate the significance of motivation for several reasons as listed below.

- Effect sizes reported in meta-analyses do not provide a complete picture if variability is not taken into account. Exploring this variability and mediating factors in more detail can suggest conditions under which stronger associations result and can lead to theoretical clarification and elaboration.
- As Figure 1 shows, motivation is only one of several important determinants of achievement. Motivational research has identified different approaches. Although most of them are conceptually related they focus on different factors. Which of them come into play, may differ and depend on student personality as well as on context.
- Motivation, in a strict sense, deals with decision-making and forming of intentions. But in order to achieve a goal, actions have to be performed, often over extended periods of time. With respect to achievement, learning activities have to be carried out. All these actions and activities require some kind of regulation and volitional processes.
- The impact of motivation is related to students' age for, at least, two reasons. First, processes such as forming an ability concept, understanding the compensatory relation between ability and effort, the use of strategies and metacognitive competences depend on cognitive development. Second, the impact of motivation varies with respect to context, and contexts differ for younger and older students.
- The role of context for motivation is as follows: Younger students are given fewer opportunities to choose and perform tasks on their own, and so the role of motivation may be more limited. The potential influence of motivation will be higher when the contexts allow or demand self-regulation. Otherwise, with strong external control such as classrooms where learning is completely structured and supported, motivational factors may lose their impact.
- Finally, research of motivation in educational contexts often uses questionnaire data that assess explicit motives. But some theorists regard implicit motives as equally important, and explicit and implicit motives may not be associated (see Heckhausen and Heckhausen, 2008).

In spite of these limitations, motivational factors are crucial for understanding learning and achievement. From a theoretical perspective, motivational and volitional processes are essential steps in a chain of processes leading from tasks to achievement but their significance depends on many other factors. In many cases, one cannot expect direct, linear relations and simple correlations between motivational and volitional characteristics and SA, but complex interactions and manifold possibilities of mutual compensation, e.g., of inefficient learning activities by increased effort. For example, excellent performance requires high degrees of both ability and effort, whereas with normal or average achievement, lack of ability can be compensated (as long as it does not drop below a critical threshold value) by increased effort (and vice versa). There is a dynamic interplay between SA and its individual motivational determinants: SA is affected by motivation, e.g., self-concept, and affects motivation itself. From this perspective, SA and its determinants change their states as independent and dependent variables.

During the last decades, a number of promising approaches and constructs were developed that make important contributions to understand the role of motivation for school achievement. As the relations between motivation and school achievement are complex, more research is needed to explore the conditions under which motivational variables have an effect upon school achievement in more detail. Models that link motivational constructs with antecedent and subsequent factors such as Eccles' model (Eccles, 2005) or the model depicted in Figure 1 could represent an appropriate framework. In future research, a broader range of stable personal characteristics such as aptitudes should be explored. Furthermore, school achievement as the main dependent variable is not a uniform construct but comprises various kinds of knowledge ranging from factual knowledge and simple skills to strategic knowledge. The effect of motivation on school achievement is mediated by learning activities and volitional processes which also should be explored in more detail. Context plays an important role to specify the conditions under which motivation is effective. One context is the learning environment which can offer different degrees of support. A highly structured supportive learning environment puts low demands on self-regulation and motivation (Corno and Snow, 1986). To understand the impact of motivation for learning in schools in more depth, the role of instruction and the interplay between instructional support and learners' activities should be considered. To examine these relations, longitudinal studies as well as experimental research are needed.

Motivation is important for life-long learning, knowledge acquisition, and self-regulation, which are basic requirements in modern societies. Self-regulation and motivational orientations that promote achievement are highly valued educational goals. This is especially true for older students but the acquisition of these skills and orientations begins much earlier (Wigfield, 1994). So, for younger children, developing motivational orientations, learning strategies and self-regulatory skills which are favorable to learning is of great importance, even if the full impact of these achievements may show up only later in their development.

See also: Academic Achievement Motivation, Development of; Avoidance and Approach Motivation: A Brief History; Emotion and Expression; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self-Regulated Learning: Theories, Measures, and Outcomes; Temperament and Motivation; Test Anxiety and Academic Achievement.

Bibliography

Blackwell, Lisa S., Trzesniewski, Kali H., Dweck, Carol S., 2007. Implicit theories of intelligence predict achievement across an adolescent tradition: a longitudinal study and an intervention. Child Development 78, 246–263.

Boekaerts, Monique, 2009. Goal-directed behavior in the classroom. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 105–122.

Boekaerts, Monique, Corno, Lyn, 2005. Self-regulation in the classroom: a perspective on assessment and intervention. Applied Psychology: An International Review 54, 199–231.

Bong, Mimi, Skaalvik, Einar M., 2003. Academic self-concept and self-efficacy: how different are they really? Educational Psychology Review 15, 1–40.

Corno, Lyn, Snow, Richard E., 1986. Adapting teaching to individual differences among learners. In: Wittrock, Merlin C. (Ed.), Handbook of Research on Teaching, third ed. Macmillan, New York, pp. 605–629.

Covington, Martin V., 1984. The Self-worth theory of achievement motivation: findings and implications. The Elementary School Journal 85, 4–20.

Dweck, Carol S., Master, A., 2009. Self-theories and motivation: students' beliefs about intelligence. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 123–140.

Eccles, Jacquelynne S., 2005. Subjective task value and the Eccles et al. model of achievement related choices. In: Elliot, Andrew J., Dweck, Carol S. (Eds.), Handbook of Competence and Motivation. Guilford Press, New York, pp. 105–121.

Eccles, Jacquelynne S., Wigfield, Allan, 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Gottfried, Adele E., 1985. Academic intrinsic motivation in elementary and junior high school students. Journal of Educational Psychology 77, 631–645.

Graham, Sandra, Weiner, Bernhard, 2012. Motivation: past, present, and future. In: Harris, Karen R., Graham, Sandra, Urdan, Tim (Eds.), APA Educational Psychology Handbook, Theories, Constructs, and Critical Issues, vol. 1. American Psychological Association, Washington, DC, pp. 367–397.

Hattie, John A.C., 2009. Visible Learning. A Synthesis of over 800 Metaanalyses Relating to Achievement. Routledge, London.

Hattie, John A.C., 2012. Visible Learning for Teachers. Maximizing Impact on Learning. Routledge, London.

Heckhausen, Jutta, Heckhausen, Heinz (Eds.), 2008. Motivation and Action, second ed. Cambridge University Press, Cambridge.

Helmke, Andreas, van Aken, Marcel A.G., 1995. The causal ordering of academic achievement and self-concept of ability during elementary school: a longitudinal study. Journal of Educational Psychology 87, 624–637.

Helmke, Andreas, Weinert, Franz E., 1997. Bedingungsfaktoren schulischer Leistungen. In: Weinert Franz, E. (Ed.), Enzyklopädie der Psychologie. Pädagogische Psychologie. Psychologie des Unterrichts und der Schule. Hogrefe, Göttingen, pp. 71–176 [Determinants of school achievement. In: Weinert Franz, E (Ed.), Encyclopedia of psychology. Educational Psychology. Psychology of Instruction and Schooling. Hogrefe, Göttingen].

Hofer, Manfred, Fries, Stefan, Helmke, Andreas, et al., 2010. Value orientations and motivational interference in school-leisure conflicts: the case of Vietnam. Learning and Instruction 20, 239–249.

Huang, Chiungjung, 2012. Discriminant and criterion-related validity of achievement goals in predicting academic achievement: a meta-analysis. Journal of Educational Psychology 104, 48–73.

Köller, Olaf, Baumert, Jürgen, Schnabel, Kai, 2001. Does interest matter? The relationship between academic interest and achievement in mathematics. Journal for Research in Mathematics Education 32, 448–470.

Linnenbrink, Elisabeth A., Pintrich, Paul R., 2002. Motivation as an enabler for academic success. School Psychology Review 31, 313–327.

Linnenbrink-Garcia, Lisa, Tyson, Diana F., Patall, Erika A., 2008. When are achievement goal orientations beneficial for academic achievement? A closer look at main effects and moderating factors. International Review of Social Psychology 21, 19–70.

Maehr, Martin L., Zusho, Akane, 2009. Achievement goal theory. The past, present, and future. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 78–104.

Marsh, Herbert W., Xu, Man, Martin, Andrew J., 2012. Self-concept: a synergy of theory, method, and application. In: Harris, Karen R., Graham, Sandra, Urdan, Tim (Eds.), APA Educational Psychology Handbook, Theories, Constructs, and Critical Issues, vol. 1. American Psychological Association, Washington, DC, pp. 427–458.

Pekrun, Reinhard, Frenzel, Anne C., Goetz, Thomas, Perry, Raymond P., 2007. The control-value theory of achievement emotions: an integrative approach to emotions in education. In: Schultz, Paul A., Pekrun, Reinhard (Eds.), Emotions in Education. Elsevier, Amsterdam, The Netherlands, pp. 13–36.

Schiefele, Ulrich, 2009. Situational and individual interest. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 197–222.

Schunk, Dale H., Pajares, Frank, 2009. Self-efficacy theory. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 35–53.

Trautwein, Ulrich, Marsh, Herbert W., Nagengast, Benjamin, Lüdtke, Oliver, Nagy, Gabriel, Jonkmann, Kathrin, 2012. Probing for the multiplicative term in modern expectancy-value theory: a latent interaction modeling study. Journal of Educational Psychology 104, 763–777.

Weinert, Franz E., Helmke, Andreas, 1987. Compensatory effects of student self-concept and instructional quality on academic achievement. In: Halisch, Frank, Kuhl, Julius (Eds.), Motivation, Intention, and Volition. Springer, Berlin, pp. 233–247.

Wigfield, Allan, 1994. Expectancy-value theory of achievement motivation: a developmental perspective. Educational Psychology Review 6, 49–77.

Wigfield, Allan, Tonks, Stephen, Klauda, Susan Lutz, 2009. Expectancy-value theory. In: Wentzel, Kathryn R., Wigfield, Allan (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 55–75.

Zimmerman, Barry J., Labuhn, Andju S., 2012. Self-regulation of learning: process approaches to personal development. In: Harris, Karen R., Graham, Sandra, Urdan, Tim (Eds.), APA Educational Psychology Handbook, Theories, Constructs, and Critical Issues, vol. 1. American Psychological Association, Washington, DC, pp. 399–425.

School as a Social System

Peter W Cookson, Jr., American Institutes for Research, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Hallinan, volume 20, pp. 13563–13567, © 2001, Elsevier Ltd.

Abstract

School as a social system describes schools as institutions and suggests a model of analysis that addresses the complex and often contradictory purposes of schooling. In particular, this article examines the socialization effects of schools in terms of norms, sanctions, and rites of passage and analyzes how allocation is structured within schools and between schools. The conclusion suggests that the social systems of schools may change dramatically in the future because of technological and social changes outside of the school.

School as a Social System: Introduction

A social system is characterized by an interdependent network of subunits, a defined population, an identifiable configuration of social relationships, and a distinctive culture (Parsons, 1951). While we generally think of social systems as encompassing whole populations, institutions can also be thought of as social systems because they are structured by many of the same characteristics associated with larger social systems. Seen in this light, schools can be considered as social systems complete with their own established configuration of social relations, distinctive cultures, norms, and legal obligations (Gordon, 1957; Parsons, 1959; Smelser, 1988).

To understand how the social systems of schools are created, developed, and influenced larger social systems requires a nuanced understanding of the complexity of schools. They are designed to promote learning, but they also reproduce existing social relations with regularity. They are sites of cooperation, but also of conflict. Schools are also embedded in larger social systems; society precedes schools but schools help shape society. This reflexive and interactive relationship means that schools act as conserving institutions as well as agents of change.

The missions of schools are multiple and not always aligned with the publically stated purposes of schooling: The manifest purpose of a school's social system is student learning, socialization, and allocation according to merit. In reality, the manifest purposes of schools are accompanied by a host of latent purposes such as the transmission of privilege and creation of shared norms that may have little to do with the formal curriculum or may be in conflict with the manifest purposes of schooling.

Viewed in this light, we can see that the social systems of schools are capable of producing apparently contradictory outcomes; the process of maximizing student achievement may run counter to the process of ensuring the reproduction of privilege, the elevation of the intellect may be subverted by the evaluation of material success. Schools are complex societies that when viewed analytically reveal a great deal about the larger social system in which they are embedded.

In schools there is a well-defined authority structure and a system of rewards and punishments. Schools are not bucolic societies where authority is exercised solely for the purpose of student learning. The sociologist of education Willard Waller (1932: p. 11) described the social system of schools as "despotisms in a state of perilous equilibrium." The social systems of schools are dynamic environments where conflict creation and conflict resolution are arrayed on a continuous line of human interaction.

The social system of schools includes more groups than students. Parents; teachers; administrators; custodial and security staff; local, regional, and national educational authorities; public servants; the media; and the public-at-large all participate either directly or indirectly in the social system of schools. Recently, the integration of the new communications technologies into teaching and learning has caused the social milieu of schools to expand to include people and organizations far from the physical location of the school.

The Purposes of Schooling

Schools are purposeful societies; they are social systems that are goal directed with observable and measurable outcomes. In general, there are four major purposes of schooling: intellectual, political, social, and economic (Bennett and LeCompte, 1990: pp. 5–21).

The intellectual purposes of schooling are to teach basic cognitive skills such as reading, writing, and mathematics; to transmit specific knowledge (e.g., in literature, history and the sciences, etc.); and to help students acquire higher-order thinking skills such as analysis, evaluation, and synthesis.

The political purposes of schooling are to inculcate allegiance to the existing political order; to prepare citizens who will participate in this political order; to help assimilate diverse cultural groups into a common political order; and to teach children the basic laws of society.

The social purposes of schooling are to help solve social problems; to work as one of many institutions, such as the family and religion to ensure social cohesion; and to socialize children into the various roles, behaviors, and values of the society. This process is referred to as socialization.

The economic purposes of schooling are to prepare students for their later occupational roles and to select, train, and allocate individuals into the division of labor. The degree to which schools directly prepare students for work varies from society to

society, but most schools play at least an indirect role in the process.

While these purposes are in many ways aligned in structuring schools as social systems, they may also come into conflict. This inherent tension was expressed clearly by the educational historian Laurence Cremin (1977: p. 37):

> Schooling – like education in general – never liberates without at the same time limiting. It never empowers without at the same time constraining. It never frees without at the same time socializing. The question is not whether one or the other is occurring in isolation but what the balance is, and to what end, and in the light of what alternatives.

Schools are complex societies. Intellectual creativity is not necessarily aligned with the political purposes of schooling, and social purposes of schooling are not always meant to serve the economic purposes of schooling. Understanding how these complex configurations make meaningful patterns requires a theoretical perspective.

Two Theoretical Perspectives

Not all scholars would describe schools as social systems in the same way. Structural functionalists see schools as socializing students into mainstream societal values, providing students with a common culture and language, enabling and encouraging completion of an education cycle, and preparing students for the labor market. Central to the structural functional perspective is that schools, as social systems, seek equilibrium. When the social system of the school is disrupted it accommodates to change by finding a new stable state. There is a tendency in this theory to emphasize the reproduction of existing social relations as a smooth process according to agreed-upon norms (Durkheim, 1956; Parsons, 1951).

A conflict theoretical view presents a different viewpoint. Conflict theory argues that schools are theaters of competition for status, wealth, and power. Social change is inevitable in the conflict perspective because unequal access to power, prestige, and resources inherently leads to destabilizing conflict. In this perspective, schools are social sites where conflict is continuous, power relations are legitimized or delegitimized, and the reproduction of inequality is disguised by the creation of sorting and selecting mechanisms that appear to be meritocratic, but in fact are based on race, class, and gender (Bowles and Gintis, 1976; Collins, 1971; Cookson, 2013).

These two perspectives provide alternative narratives for understanding schools as social systems. Both theories consider the major purposes of schooling to be socialization and allocation but draw different conclusions about the relationship of school to society and about how schools operate as social systems.

Socialization

Socialization in schools has been studied extensively (Coleman, 1961; Dreeben, 1968, Kerckhoff, 1972; Kohn, 1976; Willis, 1977; Lareau, 2003). Socialization is the process by which students internalize the values of the school and by extension the larger social system. These internalized values become norms that structure behavior.

Norms

Dreeben (1968) identified four norms of particular significance in industrialized societies: independence, achievement, universalism, and specificity. He explains (Sadovnik et al., 2012: pp. 131–132):

> In speaking of independence, achievement, universalism and specificity as norms, I mean that individuals accept the obligations, respectively: To act by themselves (unless collaborative effort is called for) and accept personal responsibility and accountability for their conduct and its consequences (independence); to perform tasks actively and master the environment according to standards of excellence (achievement); and to acknowledge the right of others to treat them as members of categories often based on a few discrete characteristics rather than on the full constellation of them representing the whole person (universalism and specificity).

Sanctions

The internalization of norms does not happen automatically or without resistance on the part of some students. Sanctions are applied to ensure that the norms of the school are internalized. School-based socialization is enforced by a system of rewards and punishments. Systematic regimes of rewards and punishments are generally codified in a written form and enforced by a school's academic personnel.

In theory, sanctions range in severity in proportion to the severity of the offense. Minor offenses, such as being occasionally late for class, might warrant a warning by a teacher or principal whereas cheating, drug use, or carrying a concealed weapon might warrant expulsion and in some cases criminal prosecution.

Sanctions are often applied unevenly despite their codification. Because schools educate the young and are dynamic human environments, if all failures to follow rules automatically resulted in sanctions there would be little time for instruction as students resist authority in many small and sometimes large ways that are overlooked by school authorities.

In addition, sanctions can vary by the social class and race of students. In a school educating poor and minority students, a minor drug offense might result in mandatory expulsion due to 'zero-tolerance' discipline policies, whereas the same offense in a middle-class school is more likely to result in a call to the student's parents.

The enforcement of norms, however, is not solely a function of formal rule enforcement. For the school to function as a stable social system, the majority of students (and adults) must internalize the norms of the school as their own value system. This occurs through systematic socialization in which school members internalize the dominant values of the school. This process of internalization is not left to chance, but encoded into the rituals, routines, and rationales of daily life. Students pass through an elaborate socialization experience which can be thought of as a rite of passage.

Rites of Passage

A rite of passage is a ceremony marking a significant transition in an individual's or a group's life. Rites of passage may be short or extended; for socialization purposes, extended rites of passage are most effective because they allow for repetition and permanent internalization. Rites of passage are transformative and enduring. In this sense schools are organized to produce certain types of graduates. According to Kamens (1977: pp. 217–218):

> Schools symbolically redefine people and make them eligible for membership in social categories to which specific sets of rights are assigned, e.g., income. The social organization of schools is major symbolic index of the kind of socialization that has occurred the thus legitimates the conferral of specific status rights.

By undergoing an organized set of socially sanctioned and enforced collective social experiences, students develop collective memories, shared prejudices, and social attitudes that remain with them into adulthood.

Two Interpretations

From the structural functional perspective, socialization allows for the intergenerational transmission of the shared norms required for the stability and continuity of society; from a conflict perspective, the deep internalization of school norms can lead to powerlessness and alienation. For structural functionalists, the internalization of norms is the foundation of a cohesive society; for conflict theorists the deep socialization that occurs in schools primarily serves the powerful who benefit from schools that reproduce people with differing status rights.

Allocation

Schools act as one of society's most efficient sorting and selecting mechanisms. These allocation mechanisms have consequences for students in terms of securing future educational opportunities and employment. Schools do this in a number of ways.

Within-School Allocation

Schools sort and select students according to measured abilities or skills. In theory, this type of allocation matches students' courses of study commensurate with their abilities and most likely educational and career destinations (Gamoran et al., 2000; Hallinan, 1994; Oakes, 1994; Lucas, 1999). Within-school allocation is commonly known as tracking.

While tracking can be found at the elementary level, most tracking occurs at the high school level. In the United States, students typically are assigned to academic, vocational, and general tracks. Students assigned to an academic track take college-preparatory courses, those assigned to the vocational track take skills- and job-related courses, and students assigned to general courses are offered low-level courses and skills courses. Research indicates that there is a strong correlation between track assignment and class and race (Oakes, 1985).

By allocating students to tracks, schools effectively channel them into prescribed life chances. By and large, working-class students end up in vocational tracks and middle-class students in academic tracks. Studies show that students placed in 'high-ability' tracks spend more time on actual teaching and learning activities; are able to use more interesting materials; and consistently receive better teachers, better laboratory facilities, and more interesting extracurricular activities than do their lower-track peers (Oakes, 1985; Goodlad, 1984). It has been found that track placement directly affects cognitive development (Rosenbaum, 1976).

Between-School Allocation

Schools vary in their academic and social power. Much of this differentiation is based on the social class and racial composition of families who send their children to a particular school (Coleman et al., 1981; Finn, 2009; Cookson, 2013). It is common place in educational research to ignore the social power of high schools under the assumption that years of schooling is the major allocation outcome variable of importance.

However, in the United States, the barriers to academic promotion are low in most cases and in some cases nonexistent. Because of this, the amount of schooling a student secures may not be directly related to upward mobility, but where a student goes to school, or his or her educational route, may be highly correlated to future success. Educational route determines with whom a student goes to school, the nature of his or her socialization experiences, and the social power of the credential with which he or she graduates in the transition to higher education and career (Cookson and Persell, 1985; Cookson, 2013).

Two Interpretations

For structural functionalists, within-school tracking is a mechanism for matching the talents of students to their most likely educational and career futures. In this view, tracking is not discriminatory; differentiations based on differing capacities are one of the key purposes of schooling (Parsons, 1951). Discussions of the legitimacy of allocation based on the social status of schools are generally not featured in structural functionalist literature.

For conflict theorists, schooling tends to reproduce social systems (Apple, 1995; Golden, 2007; Persell, 1977; Rothstein, 2004; Sacks, 2007). In this view, schools play a leading role in legitimating inequalities by conflating merit with ascription.

The efficacy of these rivaling perspectives is not just of academic interest; they bear directly on the meaning of equality of educational opportunity. If in fact students are allocated by factors that have little to do with achievement and merit, then the legitimating manifest purposes of education are open to question.

Conclusion: A Social System in Flux

In the last two decades, science and technology have undergone a major transformation. We are living in an age of invention

and discovery. To what degree will schools change and to what degree will their social systems change? Will technology alter the structure of education and in doing so alter socialization and allocation practices? Will schools evolve into institutions without walls? Will the socialization function of schooling include interactions with peers across the world? Will the power to allocate students be weakened by increased access to the World Wide Web?

There are those who predict that schooling as we know it today will disappear and a new era of freedom of information and global communication will undo traditional notions of schooling. Individuals will be able to chart their own educational courses without having to submit to formal schooling. In this scenario, individual agency will undermine the institutions of schooling. Schools will cease to operate as social systems.

This remains to be seen. Social structure is much like steel; it is marginally flexible but keeps its shape under a great deal of strain. Will schools as social systems change dramatically in the foreseeable future? This will remain an unanswered question for some time to come as a great deal depends on what happens in the larger society. But uncertainty is no barrier to inquiry; the subject of school as a social system is likely to remain as important tomorrow as it is today.

See also: Schooling: Total Impact of; Social Inequality and Schooling; Social System; Socialization in Infancy and Childhood.

Bibliography

Apple, M., 1995. Education and Power. Routledge, New York.
Bennett, K.P., LeCompte, M.D., 1990. How Schools Work. Longman, New York.
Bowles, S., Gintis, H., 1976. Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life. Basic Books, New York.
Coleman, J.S., 1961. The Adolescent Society: The Social Life of the Teenager and Its Impact on Education. Free Press, New York.
Coleman, J.S., Hoffer, T., Kilgore, S., 1981. Public and Private Schools. National Opinion Research Center, Chicago.
Collins, R., 1971. Functional and conflict theories of educational stratification. American Sociological Review 36, 1002–1019.
Cookson Jr., P.W., 2013. Class Rules: Exposing Inequality in American High Schools. Teachers College Press, New York.
Cookson Jr., P.W., Persell, C.H., 1985. Preparing for Power: America's Elite Boarding Schools. Basic Books, New York.
Cremin, L.A., 1977. Traditions of American Education. Basic Books, New York.
Dreeben, R., 1968. On What Is Learned in Schools. Addison-Wesley, Reading, MA.
Durkheim, E., 1956. Education and Sociology. Free Press, New York.
Finn, P.J., 2009. Literacy with an Attitude: Educating Working-Class Children in Their Self-Interest, second ed. State University of New York Press, Albany, NY.
Gamoran, A., Secada, W., Marret, C., 2000. The organizational context of teaching and learning: changing theoretical perspectives. In: Hallinan, M.T. (Ed.), Handbook of the Sociology of Education. Kluwer/Academic/Plenum, New York, pp. 37–64. Chapter 2.
Golden, D., 2007. The Price of Admissions: How America's Ruling Class Buys Its Way into Elite Colleges and Who Gets Left Outside the Gates. Three Rivers Press, New York.
Goodlad, J.I., 1984. A Place Called School: Prospects for the Future. McGraw-Hill, New York.
Gordon, C., 1957. The Social System of the High School. Free Press, Glencoe, IL.
Hallinan, M., 1994. Tracking: from theory to practice. Sociology of Education 67 (2), 79–84, 89–91.
Kamens, D., 1977. Legitimating myths and educational organization: the relationship between organizational ideology and formal structure. American Sociological Review 42, 208–219.
Kerckhoff, A.C., 1972. Socialization and Social Class. Prentice-Hall, Englewood Cliffs, NJ.
Kohn, M.L., 1976. Class and parental values: another confirmation of the relationship. American Sociological Review 41, 538–545.
Lareau, A., 2003. Unequal Childhoods: Class, Race and Family Life. University of California Press, Berkeley, CA.
Lucas, S.R., 1999. Tracking Inequality: Stratification and Mobility in American High Schools. Teachers College Press, New York.
Oakes, J., 1985. Keeping Track: How Schools Structure Inequality. Yale University Press, New Haven, CT.
Oakes, J., 1994. More than misapplied technology: a normative and political response to Hallinan on tracking. Sociology of Education 67 (20), 84–89, 91.
Parsons, T., 1951. The Social System. Free Press, New York.
Parsons, T., 1959. The school class as a social system: some of its functions in American society. Harvard Educational Review 29, 297–318.
Persell, C.H., 1977. Education and Inequality. Free Press, New York.
Rosenbaum, J.E., 1976. Making Inequality: The Hidden Curriculum of High Schools. Wiley, New York.
Rothstein, R., 2004. Class and Schools: Using Social, Economic and Educational Reform to Close the Black-White Achievement Gap. Teachers College Press, New York.
Sacks, P., 2007. Tearing Down the Gates: Confronting the Class Divide in American Education. University of California Press, Berkeley, CA.
Sadovnik, A.R., Cookson Jr., P.W., Semel, S.F., 2012. Exploring Education. Routledge, New York.
Smelser, N., 1988. Social structure. In: Smelser, N. (Ed.), Handbook of Sociology. Sage, Newbury Park, CA, pp. 103–129.
Waller, W., 1932. The Sociology of Teaching. Wiley, New York.
Willis, P., 1977. Learning to Labour: How Working Class Kids Get Working Class Jobs. Columbia University Press, New York.

School Burnout and Engagement: Lessons from a Longitudinal Study in Finland

Katariina Salmela-Aro, Cicero Learning, University of Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents recent findings on student school burnout and engagement in light of the ongoing FinEdu longitudinal data. Schoolwork engagement is conceptualized as study-related vigor, absorption, and dedication, whereas school burnout is defined as exhaustion, cynicism, and inadequacy as a student. First, the role of school burnout and engagement are presented in the context of the demands–resources model. Accordingly, high school demands lead to burnout, while high personal and school resources lead to engagement. Burnout in turn is connected to later depressive symptoms, and engagement to later overall satisfaction with life. Second, taking a person-oriented approach to burnout and engagement, different burnout–engagement profiles in the high school context are presented. Third, the role of educational transitions in gendered school burnout trajectories is examined. Next, school burnout and engagement are considered in relation to different social contexts. Finally, an intervention program aimed at increasing preparedness for possible setbacks in the educational career and promoting later engagement and decreasing burnout is described. These recent findings are then discussed in the context of the 4-C (channeling, choice, coregulation, compensation) life-span model of motivation.

Introduction

School burnout is a new concept (Salmela-Aro, 2012), particularly when compared to research on school engagement (Newmann, 1991; Skinner et al., 1990), which has increased dramatically over the past few years (Appleton et al., 2008; Fredricks, 2011; Fredricks et al., 2004). A recent search of PsychInfo yielded 1100 hits for studies on school engagement (Upadyaya and Salmela-Aro, 2013). Engagement has also been compared to flow (Shernoff et al., 2003), a state of mind in which the student is so intensively involved in an activity that nothing else seems to matter (Csikszentmihalyi, 1990). The burgeoning interest in engagement reflects the increasing attention shown in positive psychology, where the focus is on human strengths and optimal functioning (Reschly et al., 2008) rather than on weaknesses and malfunctioning (Schaufeli et al., 2002a; Seligman and Csikszentmihalyi, 2000). However, school burnout is also a very important influence on school achievement and mental health.

In this article, first, the role of school burnout and engagement are presented in the context of a demands–resources model. According to the model, high school demands lead to burnout, whereas high personal and school resources lead to engagement. The latter is connected to later overall satisfaction with life, whereas burnout leads to later depressive symptoms. Second, taking a person-oriented approach to burnout and engagement, different burnout–engagement profiles within the high school context are described. Third, the role of educational transitions with regard to gendered school burnout trajectories is summarized. Next, school burnout and engagement are considered in relation to different social contexts. Finally, the results of an intervention program with the aim of increasing preparedness for possible setbacks in the educational career and promoting later on engagement and decreasing burnout are described. These recent findings are then discussed in the context of the 4-C (channeling, choice, coregulation, compensation) life-span model of motivation (Salmela-Aro, 2009).

The results presented in this article are based on longitudinal data drawn from the ongoing FinEdu study in which data have been collected on all the young people resident in a medium-sized town in central Finland. They have been followed intensively for 10 years, starting at age 15 (year 2003), when they were at the end of comprehensive school, to age 25 (year 2013).

School Burnout and Engagement

School burnout is a school-related stress disorder that can be described along three dimensions: exhaustion, cynicism toward school, and a sense of inadequacy as a student (Salmela-Aro et al., 2009). Cynicism is manifested in an indifferent or distal attitude toward schoolwork in general, a loss of interest in schoolwork, and not seeing it as meaningful. Inadequacy refers to a diminished sense of competence, achievement, and accomplishment as a student. Earlier FinEdu results showed that all three dimensions were stable from ages 15 to 19 (Parker and Salmela-Aro, 2011). Furthermore, both exhaustion due to school demands and cynicism toward school predicted feelings of inadequacy as a student later on (Parker and Salmela-Aro, 2011).

Schoolwork engagement, like work engagement (Schaufeli et al., 2006), can be defined along three dimensions: energy, dedication, and absorption (Salmela-Aro and Upadyaya, 2012, under review; Schaufeli et al., 2002b; Upadyaya and Salmela-Aro, 2013). Schoolwork engagement resembles flow, which describes an experience that is so enjoyable in itself that people desire it, even at great cost, purely for its own sake (Csikszentmihalyi, 1990). The main difference between the concepts of school engagement and flow is that flow refers to a short-term peak experience, whereas engagement is a more persistent state of mind. Of the three dimensions of engagement, energy refers to high mental resilience while studying, a willingness to invest effort in one's schoolwork, and a positive approach to school (Salmela-Aro and Upadyaya, 2012;

Schaufeli et al., 2002b). Dedication is characterized by a sense of significance, enthusiasm, pride, identification, and inspiration regarding school, as well as perceiving schoolwork as meaningful (Salmela-Aro and Upadyaya, 2012; Schaufeli et al., 2002b). Absorption is characterized by behavioral accomplishments and flowlike experiences, such as being fully concentrated and happily engrossed in one's studying so that time passes quickly (Salmela-Aro and Upadyaya, 2012; Schaufeli et al., 2002b). These three dimensions of school engagement are separate but highly correlated constructs (Salmela-Aro and Upadyaya, 2012; Schaufeli et al., 2002b). Developmental changes seem to co-occur in the composition of these engagement dimensions (Salmela-Aro and Upadyaya, 2012). At the beginning of postcomprehensive education, when students have experienced the transition to high school (i.e., academic track) or vocational school (i.e., vocational track), their engagement is perhaps better described as overall engagement with school. However, in their postcomprehensive education and especially in university-level studies, engagement is characterized by the separate dimensions of energy, dedication, and absorption (Salmela-Aro and Upadyaya, 2012; Schaufeli et al., 2002a,b). These changes may be due to the closeness of the transition to working life, a time when students become more oriented toward work and the new demands of working life. It is at this point when students' schoolwork engagement becomes more differentiated and similar to work engagement (Salmela-Aro and Upadyaya, 2012).

The Demands–Resources Model in the School Context

One new approach to examining school burnout and engagement in the school context is based on the demands–resources model (Salmela-Aro and Upadyaya, 2014). The demands–resources model, typically applied in studies on work engagement, suggests that high engagement in the educational setting promotes students' subsequent well-being, life satisfaction, personal growth, and learning (Demerouti et al., 2001; Salmela-Aro and Upadyaya, 2014), whereas burnout predicts problems later on. The demands–resources model assumes two processes (Demerouti et al., 2001): an effort-driven energetic process of overtaxing and wearing out in which severe study demands exhaust the student's energy and lead first to burnout and then to diminished mental health, for example in the form of depression. The other is a motivational process in which the availability of resources leads to engagement and fosters later life satisfaction. According to the demands–resources model, several resources such as high self-efficacy, and demands such as study challenges and overload, promote or hinder one's engagement over the school years (Demerouti et al., 2001; Salmela-Aro and Upadyaya, 2014).

We documented pathways from study demands and study resources to later schoolwork-related engagement and burnout and then to later life adjustment on the basis of the FinEdu data (Salmela-Aro and Upadyaya, 2014). In the model, study demands are seen as presenting a challenge to study-related goals, while resources are seen as functional in achieving study-related goals. For example, study-related stress can be interpreted as a demand felt by students that they adjust to school, whereas high self-efficacy may function as a resource for their engagement with school (Vasalampi et al., 2010). In line with the demands–resources model, study-related resources constituted more robust determinants of engagement with schoolwork, while study demands constituted more robust determinants of school burnout. Our results showed that the model fit the data well and can thus be used as a theoretical background for burnout and engagement in the school context. Supporting the model (see **Figure 1**), study demands in the last year of comprehensive school were positively related to school burnout in upper secondary education 1 year later, while study resources were positively related to engagement with schoolwork 1 year later. In addition, personal resources, such as self-efficacy in the last year of comprehensive school, were positively related to engagement with schoolwork and negatively to school burnout 1 year later.

In line with the model, engagement with schoolwork and school burnout were both stable during upper secondary education and negatively associated (see **Figure 1**). Furthermore, the cross-lagged paths between engagement with schoolwork and school burnout revealed that school burnout negatively predicted engagement with schoolwork 1 year later, and engagement with schoolwork was positively related to life satisfaction 2 years later, while school burnout was positively related to later depressive symptoms (see also Salmela-Aro et al., 2008c). School burnout was also negatively related to life satisfaction 2 years later. Our results thus indicate that the demands–resources model is a suitable framework for

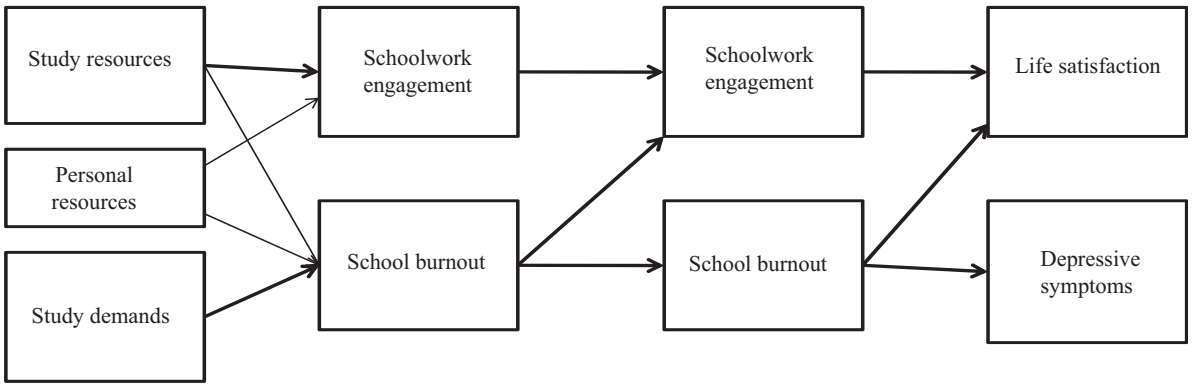

Figure 1 The demands–resources model in the school context.

explaining student pathways to adjustment and maladjustment in the school context and general adjustment later on. The model includes both an effort-driven energetic process of overtaxing and wearing out in which great demands exhaust energy and lead to burnout and diminished mental health, and a motivational process in which the availability of resources leads to engagement. School burnout mediated the relationship between study demands and mental health outcomes.

Person-Oriented Approach to Engagement and Burnout

Using a person-oriented approach, we have recently examined what kinds of schoolwork engagement and burnout profiles can be identified among high school students (Tuominen-Soini et al., 2011; Tuominen-Sioni and Salmela-Aro, 2014). The results, based on the FinEdu data, showed that students' engagement with school does not necessarily decrease across the school years, but rather that different subgroups of students with varying levels and developmental trends of school engagement can be identified (Upadyaya and Salmela-Aro, in press). In fact, the four-group solution fit the data well. We labeled these groups engaged, engaged–exhausted, cynical, and burnout. The first two were engaged. Those in the engaged group (the largest group with 44% of the students) scored high on schoolwork engagement and relatively low on all the burnout dimensions; in contrast, those in the engaged–exhausted group (28%) scored high on schoolwork engagement but also reported rather high levels of exhaustion and inadequacy. Thus, students can feel engagement and disengagement simultaneously. The other two groups were disengaged: those in the cynical group (14%) scored high, particularly in cynicism, and low in engagement, whereas those in the burned-out group (14%) were characterized by relatively high levels of all three burnout dimensions – exhaustion, cynicism, and inadequacy as a student.

These results show that most students are engaged. As a whole, students who are highly engaged with school typically follow an overall positive youth development trajectory. They produce high academic outcomes, possess a mastery-oriented approach to studying, seldom drop out of school, adapt better to challenging situations, spend their time out of school in constructive activities, and experience overall better well-being (Tuominen-Soini et al., 2011). However, one-third of students evidence clear signs of burnout. We examined these students further during their transition to working life and found that membership in the cynical group increased and that burnout symptoms tended to increase in severity over time in the burnout group (Tuominen-Sioni and Salmela-Aro, 2014).

Developmental Changes in Burnout and Engagement during Educational Transitions

We next analyzed changes in the three components of school burnout – exhaustion, cynicism, feelings of inadequacy – during the transition from comprehensive school to upper secondary education (Salmela-Aro et al., 2008b; Salmela-Aro and Tynkkynen, 2012). The developmental patterns varied by gender and educational track. Although exhaustion increased among boys on the academic track, girls on the academic track reported higher levels of exhaustion than their male peers at all time points. Cynicism decreased among girls on the vocational track but increased among boys on the academic track. Among girls on the academic track and boys on the vocational track, cynicism remained stable. At the end of upper secondary education, cynicism was highest among boys on the academic track. Finally, the level of feelings of inadequacy remained stable among both boys and girls on the vocational track, whereas it increased among both girls and boys on the academic track.

These patterns are consistent with the idea that changes in the fit between students' developmental needs and school settings may change across school transitions differently for various groups of students (Eccles and Roeser, 2009, 2011). During educational transitions, the fit between the student and the school is a crucial factor affecting school engagement and burnout. When inconsistent changes occur in both the adolescent and the school context (Eccles and Roeser, 2009), the stage–environment fit (Eccles and Midgley, 1989) can become impaired. According to the stage–environment fit theory (Eccles and Midgley, 1989), many of the changes associated with educational transitions, in particular in the academic track, are at odds with the changing developmental needs of adolescents (e.g., increased emphasis on grades and competition, decline in adolescents' perception of emotional support from teachers, and sense of belonging in their classrooms). According to the stage–environment fit theory, if schools do not provide developmentally appropriate educational environments for adolescents, they are not offering the kind of social context that will continue to motivate students' interest and, consequently, may result in increasing burnout and decreasing engagement. A negative developmental fit can lead to alienation from school and give rise to cynicism. In contrast when the school context fits well with students' interests, goals, and psychological needs, the result should be continuing high engagement and well-being and low burnout (Eccles and Midgley, 1989; Salmela-Aro et al., 2008b; Salmela-Aro and Tynkkynen, 2012). This scenario may also occur after the transition to higher education or work (Upadyaya and Salmela-Aro, in press). The results from tFinEdu study suggest that, while the level of overall engagement might be low during postcomprehensive education, it may increase after the transition to higher education or work (Upadyaya and Salmela-Aro, in press) because the new study or work environment may provide a better person–environment fit for some young adults (Eccles and Midgley, 1989; Eccles and Roeser, 2009).

Girls often exhibit higher levels of behavioral, emotional, and overall school engagement than boys (Salmela-Aro and Upadyaya, 2012; Wang and Eccles, 2012), which may be related to girls' tendency to perform better at school (Pomerantz et al., 2002) and to attribute greater importance to academic achievement than boys, as well as a better fit between the students' needs and the school culture for girls than for boys (Berndt and Miller, 1990). Consistent with these hypotheses, boys are more likely than girls to experience rapid

decreases in engagement and to follow unstable school engagement trajectories, often leading to school dropout (Li and Lerner, 2011).

For both boys and girls, high levels of secondary school engagement predict higher education and better job possibilities (Upadyaya and Salmela-Aro, in press; Vasalampi et al., 2009), as well as life satisfaction (Salmela-Aro and Tuominen-Soini) and well-being (Salmela-Aro and Upadyaya, 2014) later in life. In contrast, burnout predicts more gap years and difficulties in negotiating the transition to working life and later education (Vasalampi et al., 2009). Recent studies have also revealed that during involuntary gap years after high school, burnout decreased and engagement increased (Parker et al., submitted for publication).

School Burnout and Engagement in Varying Social Contexts

Students' social contexts (parents, peers, teachers, and the school environment) also predict school-related engagement and burnout (see reviews by Upadyaya and Salmela-Aro, 2013; see also Finn and Rock, 1997; Finn and Voelkl, 1993; Eccles and Roeser, 2009, 2011). Supportive relationships with parents promote students' school engagement: parental involvement and support positively predict students' school engagement (Upadyaya and Salmela-Aro, in press; Wang and Eccles, 2012). When students have a good relationship with their parents, they also seem more likely to complete a successful transition to higher education or working life (Upadyaya and Salmela-Aro, in press; Vasalampi et al., 2009). A good, supportive relationship with parents is an important resource for school engagement and continued educational and career success (Upadyaya and Salmela-Aro, in press; Wang and Eccles, 2012). However, the FinEdu results also revealed that burnout is shared in the family (Salmela-Aro et al., 2011). In particular, levels of mothers' work burnout and daughters' school burnout are correlated as are fathers' work burnout and sons' school burnout.

School engagement and burnout can also be shared among students in the same peer group (Kiuru et al., 2008, 2009). In the FinEdu data, both the level of and changes in school burnout were shared in a peer group (Kiuru et al., 2008). Moreover, a selection influence was found for school engagement. In addition, high levels of school engagement were shared especially among tight peer groups of girls, probably because girls in highly cohesive peer groups experience a higher sense of belonging, which is related to school engagement (Kiuru et al., 2009).

But who influences whom? Using FinEdu data, we found that maternal affection moderates the influence that higher-burnout friends exert over the subsequent schoolwork engagement of their lower-burnout friends (Marion et al., 2014) such that the schoolwork engagement of the friend reporting higher levels of school burnout predicts a decline in the schoolwork engagement of the friend reporting lower school burnout only if the latter reported lower than average maternal affection. These results suggest that warm family relationships can buffer against negative peer influences.

In addition, teachers can play a role in school burnout and engagement (Pyhältö et al., 2011; Salmela-Aro et al., 2008a). Although the effects were small, students' burnout is predicted by higher levels of haste and lower levels of general atmosphere. In contrast, the fairer and more encouraging the teachers in high school are perceived to be, and the more supporting adults are available, the less burnout is reported by the students (Salmela-Aro et al., 2008a). Thus, during adolescence, teacher support and fairness can help support students' sense of engagement (Salmela-Aro et al., 2008a), as can perceived support for student autonomy and feelings of relatedness (Deci et al., 1991).

Intervention to Promote Engagement and Prevent Burnout

Although school transitions include several risk factors for student engagement (Roeser et al., 1999), they also offer opportunities for change and possible interventions. One possibility is to boost students' transition efficacy and related resources and provide them with information on the opportunities and challenges of further education and career opportunities prior to their transition to vocational studies, higher education, or work. Such preparedness (Salmela-Aro et al., 2012) would enhance students' subsequent engagement and the value they attach to educational goals, and help them in adjusting to the upcoming educational transitions (Vuori et al., 2008).

Recently, an intervention titled Towards Working Life was implemented at the end of comprehensive school in Finland with the aim of promoting student preparedness for the impending educational transition (Vuori et al., 2008). It was assumed that transition preparedness would subsequently increase school engagement and decrease burnout. Transition preparedness has two components: promoting transition-related resources, such as self-efficacy, and increasing preparedness for possible setbacks and challenges during the transition. The results of the randomized trial documented the effectiveness of this intervention for increasing students' transition preparedness. In addition, increased transition preparedness predicted decreasing levels of students' school burnout later in upper secondary school (Salmela-Aro et al., 2010). Key educational transition periods seem to be windows of opportunity for changes in burnout and engagement. These are also ideal places to carry out interventions to boost transition preparedness and school engagement.

Conclusion

In this article, processes associated with both schoolwork engagement and school burnout were presented in the context of the school demands–resources model. According to this demands–resources model, high school demands lead to burnout, whereas high personal and school resources lead to engagement. In turn, engagement can lead later on to overall satisfaction with life, whereas burnout can lead to later depressive symptoms. Moreover, school burnout predicts decreases in engagement later on rather than vice versa. Second, taking a person-oriented approach, four burnout–engagement profiles were presented in the high school context. Although most students were engaged, one-third

suffered from burnout. In addition, about one-third experienced high engagement and exhaustion at the same time. This combination can be demanding later on during transition from high school to further studies. Third, attention was paid to the role of educational transitions in gendered school burnout trajectories. Girls seem to suffer more from burnout, but both girls and boys on the academic track reported an increase in feelings of inadequacy. School burnout and engagement were also considered in the social context. The role of peers, parents, and teachers all seem to be crucial. Finally, the results of a recent intervention program aiming at increased transition preparedness for possible setbacks related to educational career were described. It was found that transition preparedness promoted school engagement and decreased burnout in students' subsequent educational careers.

These findings can be approached in the context of the 4C (channeling, choice, coregulation, compensation) life-span model of motivation (Salmela-Aro, 2009). In line with this model (Salmela-Aro, 2009; Dietrich et al., 2012), various environmental characteristics, such as the transition to school and to work, channel students' engagement and burnout. However, students coregulate their engagement and burnout in reciprocal relationships with parents, peers, and teachers. The changing fit between students' developmental needs and school settings (Eccles and Roeser, 2009) may be reflected in changes in students' engagement. However, a crucial role in engagement is played by young people's own choices, agency, and related personal resources. Finally, young people can be helped to reduce problems of engagement and burnout (compensate): boosting their transition preparedness played a key role during these major educational transitions by promoting engagement and preventing burnout later on. Perseverance in the face of obstacles seems to lead to advantageous outcomes such as engagement.

Future Directions

There are several issues that should be taken into account in future research on school burnout and engagement. First, students' burnout and engagement need to be examined in different academic domains, such as STEM (science, technology, engineering, and math). Some students might feel highly engaged with math and sciences but lack engagement with the arts and languages. This would help to understand the processes through which engagement and burnout develop in school. Second, more research using experience sampling methods is needed to reveal burnout and engagement processes during the school day. What really engages students and when are students most burned out? Do students feel more engagement when doing sports than academic subjects? Daily and weekly fluctuations in burnout and engagement, and situational and contextual factors affecting burnout and engagement might be then revealed. Moreover, this article is based on data collected in Finland, and as school systems and contexts may vary greatly between countries more research is needed in other countries and school contexts. In addition, this article focused on adolescents, and thus engagement and burnout processes during elementary and middle school should also be examined. Finally, new challenges for schools, such as segregation in neighborhoods and increasing immigration, need to be addressed in the context of burnout and engagement.

See also: Academic Engagement; Academic Motivation and Performance: Task Value Interventions; Burnout, Psychology of; Coping across the Lifespan; Flow in Motivational Psychology; Gender and Academic Motivation; Grit; Interest, Psychology of; Mastery Learning; Motivation, Learning, and Instruction; Passion and Motivation; Personal Projects; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Test Anxiety and Academic Achievement.

Bibliography

Appleton, J.J., Christenson, S.L., Furlong, M.J., 2008. Student engagement with school: critical conceptual and methodological issues of the construct. Psychology in the Schools 45, 369–386. http://dx.doi.org/10.1016/j.jsp.2006.04.002.

Bergman, L.R., Andersson, H., 2010. The person and the variable in developmental psychology. Journal of Psychology 218 (3), 155–165. http://dx.doi.org/10.1027/0044-3409/a000025.

Berndt, T.J., Miller, K.E., 1990. Expectancies, values, and achievement in junior high school. Journal of Educational Psychology 82, 319–326. http://dx.doi.org/10.1037/0022-0663.82.4.664.

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. Harper and Row, New York.

Deci, E.L., Vallerand, R.J., Pelletier, L.G., Ryan, R.M., 1991. Motivation and education: the self-determination perspective. Educational Psychologist 26, 325–346. http://dx.doi.org/10.1080/00461520.1991.9653137.

Demerouti, E., Bakker, A., Nachreiner, F., Schaufeli, W., 2001. The job demands–resources model of burnout. Journal of Applied Psychology 86, 499–512. http://dx.doi.org/10.1037/0021-9010.86.3.499.

Dietrich, J., Parker, P., Salmela-Aro, K., 2012. Phase-adequate engagement at the post-school transition. Developmental Psychology 48 (6), 1575–1593. http://dx.doi.org/10.1037/a0030188.

Eccles, J., Midgley, C., 1989. Stage/environment fit: developmentally appropriate classrooms for young adolescents. In: Ames, R., Ames, C. (Eds.), Research on Motivation and Education: Goals and Cognitions, vol. 3. Academic Press, New York: NY, pp. 139–186.

Eccles, J.S., Roeser, R.W., 2011. Schools as developmental contexts during adolescence. Journal of Research on Adolescence 21, 225–241. http://dx.doi.org/10.1111/j.1532-7795.2010.00725.x.

Eccles, J.S., Roeser, R.W., 2009. Schools, academic motivation, and stage-environment fit. In: Lerner, R.M., Steinber, L. (Eds.), Handbook of Adolescent Psychology, third ed. John Wiley & Sons, Hoboken, N.J, pp. 404–434.

Finn, J.D., 1989. Withdrawing from school. Review of Educational Research 59, 117–142. http://dx.doi.org/10.3102/00346543059002117.

Finn, J.D., Rock, D.A., 1997. Academic success among students at risk for school failure. Journal of Applied Psychology 82, 221–234. http://dx.doi.org/10.1037/0021-9010.82.2.221.

Finn, J.D., Voelkl, K.E., 1993. School characteristics related to student engagement. Journal of Negro Education 62, 249–268.

Fredricks, J.A., 2011. Engagement in school and out-of-school contexts: a multidimensional view of engagement. Theory Into Practice 50, 327–335. http://dx.doi.org/10.1080/00405841.2011.607401.

Fredricks, J.A., Blumenfeld, P.C., Paris, A.H., 2004. School engagement: potential of the concept, state of the evidence. Review of Educational Research 74, 59–109. http://dx.doi.org/10.3102/00346543074001059.

Kiuru, N., Aunola, K., Nurmi, J.-E., Leskinen, E., Salmela-Aro, K., 2008. Peer group influence and selection in adolescents' school burnout: a longitudinal study. Merrill-Palmer Quarterly 54 (1), 23–55. http://dx.doi.org/10.1353/mpq.2008.0008.

Kiuru, N., Nurmi, J.-E., Aunola, K., Salmela-Aro, K., 2009. Peer group homogeneity in adolescents' school adjustment varies according to peer group type and gender. International Journal of Behavioral Development 33, 65–76. http://dx.doi.org/10.1177/0165025408098014.

Li, Y., Lerner, R.M., 2011. Trajectories of school engagement during adolescence: implications for grades, depression, delinquency, and substance use. Developmental Psychology 47, 233–247. http://dx.doi.org/10.1037/a0021307.

Marion, D., Laursen, B., Kiuru, N., Nurmi, J., Salmela-Aro, K., 2014. Maternal affection moderates friend influence on schoolwork engagement. Developmental Psychology 50 (3), 766–771. http://dx.doi.org/10.1037/a0034295.

Newmann, F., 1991. Student engagement in academic work: expanding the perspective on secondary school effectiveness. In: Bliss, J.R., Firestone, W.A. (Eds.), Rethinking Effective Schools: Research and Practice. Prentice-Hall, Englewood Cliffs, NJ, pp. 58–76.

Parker, P.D., Salmela-Aro, K., 2011. Developmental processes in school burnout: a comparison of major developmental models. Learning and Individual Differences 21 (2), 244–248. http://dx.doi.org/10.1016/j.lindif.2011.01.005.

Parker, P. Felix, Salmela-Aro, K. Burnout and engagement during the gap years, submitted for publication.

Pomerantz, E.M., Altermatt, E.R., Saxon, J.L., 2002. Making the grade but feeling distressed: gender differences in academic performance and internal distress. Journal of Educational Psychology 94, 396–404. http://dx.doi.org/10.1037/0022-0663.94.2.396.

Pyhältö, K., Pietarinen, J., Salmela-Aro, K., 2011. Teacher–working-environment fit as a framework for burnout experienced by Finnish teachers. Teaching and Teacher Education 27 (7), 1101–1110. http://dx.doi.org/10.1016/j.tate.2011.05.006.

Reschly, A.L., Huebner, E.S., Appleton, J.J., Antaramian, S., 2008. Engagement as flourishing: the contribution of positive emotions and coping to adolescents' engagement at school and with learning. Psychology in the Schools 45, 419–431. http://dx.doi.org/10.1002/pits.20306.

Roeser, R.W., Eccles, J.S., Freedman-Doan, C., 1999. Academic functioning and mental health in adolescence – patterns, progressions, and routes from childhood. Journal of Adolescent Research 14, 135–174. http://dx.doi.org/10.1177/0743558499142002.

Salmela-Aro, K., 2009. Personal goals and well-being during critical life transitions: the four C's – channelling, choice, co-agency and compensation. Advances in Life Course Research 14 (1–2), 63–73. http://dx.doi.org/10.1016/j.alcr.2009.03.003.

Salmela-Aro, K., 2012. Motivation, burnout and engagement during critical transitions from school to work. In: Karabenick, S., Urdan, T. (Eds.), Advances in Motivation and Achievement, vol. 17, pp. 79–108.

Salmela-Aro, K., Kiuru, N., Leskinen, E., Nurmi, J.-E., 2009. School burnout inventory: reliability and validity. European Journal of Psychological Assessment 25, 48–57. http://dx.doi.org/10.1027/1015–5759.25.1.48.

Salmela-Aro, K., Kiuru, N., Pietikäinen, M., Jokela, J., 2008a. Does school matter? European Psychologist 13 (1), 12–23. http://dx.doi.org/10.1027/1016-9040.13.1.12.

Salmela-Aro, K., Kiuru, N., Nurmi, J.-E., 2008b. The role of educational track in adolescents' burn out: a longitudinal study. British Journal of Educational Psychology 78, 663–689. http://dx.doi.org/10.1348/000709908X281628.

Salmela-Aro, K., Mutanen, P., Koivisto, P., Vuori, J., 2010. Adolescents' future education-related personal goals, concerns, and internal motivation during the "towards working life" group intervention. European Journal of Developmental Psychology 7 (4), 445–462. http://dx.doi.org/10.1080/17405620802591628.

Salmela-Aro, K., Mutanen, P., Vuori, J., 2012. Promoting career preparedness and intrinsic work-goal motivation: RCT intervention. Journal of Vocational Behavior 80 (1), 67–75. http://dx.doi.org/10.1016/j.jvb.2011.07.001.

Salmela-Aro, K., Savolainen, H., Holopainen, L., 2008c. Depressive symptoms and school burnout during adolescence: evidence from two cross-lagged longitudinal studies. Journal of Youth and Adolescence 38 (10), 1316–1327. http://dx.doi.org/10.1007/s10964-008-9334-3.

Salmela-Aro, K., Tuominen-Soini, H., 2010. Adolescents' life satisfaction during the transition to post-comprehensive education: antecedents and consequences. Journal of Happiness Studies 11 (6), 683–701. http://dx.doi.org/10.1007/s10902-009-9156-3.

Salmela-Aro, K., Tynkkynen, L., 2012. Gendered pathways in school burnout among adolescents. Journal of Adolescence 35 (4), 929–939. http://dx.doi.org/10.1016/j.adolescence.2012.01.001.

Salmela-Aro, K., Tynkkynen, L., Vuori, J., 2011. Parents' work burnout and adolescents' school burnout: are they shared? European Journal of Developmental Psychology 8 (2), 215–227. http://dx.doi.org/10.1080/17405620903578060.

Salmela-Aro, K., Upadyaya, K., 2012. The schoolwork engagement inventory: energy, dedication and absorption (EDA). European Journal of Psychological Assessment 28, 60–67. http://dx.doi.org/10.1027/1015-5759/a000091.

Salmela-Aro, K., Upadyaya, K., 2014. School burnout and engagement in the context of demands–resources model. British Journal of Educational Psychology 84 (1), 137–151. http://dx.doi.org/10.1111/bjep.12018.

Schaufeli, W.B., Bakker, A.B., Salanova, M., 2006. The measurement of work engagement with a short questionnaire: a cross-national study. Educational and Psychological Measurement 66, 701–716. http://dx.doi.org/10.1177/0013164405282471.

Schaufeli, W.B., Martinez, I., Pinto, A.M., Salanova, M., Bakker, A., 2002a. Burnout and engagement in university students: a cross-national study. Journal of Cross-Cultural Psychology 33 (5), 464–481. http://dx.doi.org/10.1177/0022022102033005003.

Schaufeli, W.B., Salanova, M., Gonzalez-Roma, V., Bakker, A.B., 2002b. The measurement of engagement and burnout: a two sample confirmatory factor analytic approach. Journal of Happiness Studies 3, 71–92. http://dx.doi.org/10.1023/A:1015630930326.

Seligman, M.E.P., Csikszentmihalyi, M., 2000. Positive psychology: an introduction. American Psychologist 55, 5–14. http://dx.doi.org/10.1037/0003-066X.55.1.5.

Shernoff, D.J., Csikszentmihalyi, M., Scheiner, B., Shernoff, E.S., 2003. Student engagement in high school classrooms from the perspective of flow theory. School Psychology Quarterly 18, 158–176. http://dx.doi.org/10.1521/scpq.18.2.158.21860.

Skinner, E.A., Wellborn, J.G., Connell, J.P., 1990. What it takes to do well in school and whether I've got it: the role of perceived control in children's engagement and school achievement. Journal of Educational Psychology 82, 22–32. http://dx.doi.org/10.1037/0022-0663.82.1.22.

Tuominen-Sioni, H., Salmela-Aro, K., 2014. Schoolwork engagement and burnout among Finnish high school students and young adults: profiles, progressions, and educational outcomes. Developmental Psychology 50 (3), 649–662. http://dx.doi.org/10.1037/a0033898.

Tuominen-Soini, H., Salmela-Aro, K., Niemivirta, M., 2011. Stability and change in achievement goal orientations: a person-centered approach. Contemporary Educational Psychology 36, 82–100. http://dx.doi.org/10.1016/j.cedpsych.2010.08.002.

Upadyaya, K., Salmela-Aro, K., 2013. Development of studies engagement in association with academic success and well-being in varying social contexts: a review of empirical research. European Psychologist 18 (2), 136–147. http://dx.doi.org/10.1027/1016-9040/a000143.

Upadyaya, K. Salmela-Aro, K. (in press). Engagement with studies and work: trajectories from post-comprehensive education to higher education and work, Journal of Emerging Adulthood

Vasalampi, K., Salmela-Aro, K., Nurmi, J.-E., 2009. Adolescents' self-concordance, school engagement, and burnout predict their educational trajectories. European Psychologist 14 (4), 332–341. http://dx.doi.org/10.1027/1016-9040.14.4.332.

Vasalampi, K., Salmela-Aro, K., Nurmi, J.-E., 2010. Education-related goal appraisals and self-esteem during the transition to secondary education: a longitudinal study. International Journal of Behavioral Development 34 (6), 481–490. http://dx.doi.org/10.1177/0165025409359888.

Vuori, J., Koivisto, P., Mutanen, P., Jokisaari, M., Salmela-Aro, K., 2008. Towards working life: effects of an intervention on mental health and transition to post-basic education. Journal of Vocational Behavior 72 (1), 67–80. http://dx.doi.org/10.1016/j.jvb.2007.10.003.

Wang, M.-T., Willet, J.B., Eccles, J.S., 2011. The assessment of school engagement: examining dimensionality and measurement invariance by gender and race/ethnicity. Journal of School Psychology 49, 465–480. http://dx.doi.org/10.1016/j.jsp.2011.04.001.

Wang, M.-T., Eccles, J.S., 2012. Social support matters: longitudinal effects of social support on three dimensions of school engagement from middle to high school. Child Development 83, 877–895. http://dx.doi.org/10.1111/j.1467-8624.2012.01745.

Wigfield, A., Eccles, J.E., 2000. Expectancy-value theory of achievement motivation. Contemporary Educational Psychology 25, 68–81. http://dx.doi.org/10.1006/ceps.1999.1015.

Wigfield, A., Eccles, J.S., Schiefele, U., Roeser, R., Davis-Kean, P., 2006. Development of achievement motivation. In: Damon, W. (Series Ed.), Eisenberg, N. (Eds.), Handbook of Child Psychology, Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York, pp. 933–1002.

School Business Management

Paul Armstrong, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The role of school business managers (SBMs) in the English state school system has grown exponentially over the past decade. The roots of the SBM can be traced to its forerunner, the school bursar, and private schools where the role had been long established before a period of large-scale educational reform provided the catalyst for its growth within the state school sector. In 2002, a national government initiative launched the inaugural program of training and certification for school bursars in a bid to dramatically increase their numbers in schools in England. Since then the role has evolved into the SBM, a post that encompasses a range of responsibilities far greater in scope and diversity than those of their bursarial predecessors. Research evidence suggests that while SBMs are making a positive impact in English schools there remain cultural barriers from some educational stakeholders who have yet to fully embrace the profession and its potential. Looking forward, indications suggest that SBMs will have a central role to play in the modern day educational setting.

Introduction

This article outlines the growth of school business management in the English state school system. It begins with the origins of the bursar within the private school sector before discussing the gradual introduction of the role into state schools, a process fueled significantly by the educational reforms of the late 1980s. The chapter then considers how a large-scale government strategy in 2002 initiated the evolution of the school business manager (SBM) role, a role that would become increasingly professionalized over the following decade. International comparisons are drawn followed by some of the recent research evidence on the impact being made by SBMs in England. The chapter concludes with some thoughts on the future of the profession.

School Business Management: A Brief History

The roots of school business management in the English state school system can be traced back to bursarship, a profession traditionally associated with the independent school sector. The bursar would have been predominantly responsible for the financial management of a school, college, or university with typical candidates being exmilitary personnel who often found their personal management and organizational skills suitable for the role (Kerry, 2001). One of the more illustrious incumbents of this profession was the economist John Maynard Keynes (1883–1946) who was bursar at King's College, Cambridge between 1924 and 1944, a time during which he 'amassed a considerable personal fortune from the financial markets' and 'greatly improved the college's financial position' (BBC, 2012). In a more informative account of the specifics of the traditional bursar, O'Sullivan et al. (2000) cite the biography of Robert Somerville, both master and bursar at the Harrow School for 33 years between 1887 and 1920, whose responsibilities included relieving teachers of their nonteaching duties, advising on school development plans, and dealing with contractors and estates.

Such insights illustrate the long and distinguished history of bursarship (the roll of bursars at King's College is traceable to the 1400s) and what is also interesting about this account are the parallels between Robert Somerville's role and those of his modern day successor, the SBM. Perhaps because of their predisposition to employment within the independent school sector, such references to bursars are rare within historical research literature.

Within the state school sector, it was not until the early 1980s that scholarly attention turned toward those members of the school workforce not directly involved with teaching and learning. Colin Riches conducted some of the earliest academic research studies on what he termed 'nonteaching staff,' an umbrella term he used to categorize all school staff who were not qualified teachers such as administrative and clerical assistants, caretakers and cleaners, grounds staff and gardeners, and school meals staff (Riches, 1981). Within his research he alludes to the fact that nonteaching staff are often equivalent in number to their teaching counterparts while indicating the implications of this evolving group of school support staff for the head teacher tasked with:

> managing a workforce which is to some extent not his own, and yet these staff are often/sometime in positions of influence in the school, e.g. the controlling caretaker and the "gatekeeper" secretary.
> Riches, 1982, p. 154

It is notable that in modern day schools, the management and coordination of this 'workforce' is often the responsibility of the SBM whereby Riches' observations were perhaps early signs of the growing nonteaching obligations of the head teacher. Around this time, the Local Education Authority (LEA) for London undertook some research exploring the implications of increasing workloads for teachers with a view to improving achievement in schools. While nonteaching staff were not the primary focus of this research, one of the recommendations from the final report suggested an increase in provision for administrative assistants in schools to relieve deputy head teachers of their clerical duties (thus allowing them more time in the classroom) (The Hargreaves Report, ILEA, 1984). Though sporadic, these glimpses of nonteaching staff highlight their developing

profile, which, by the end of the decade, was set to increase significantly.

Bursars in State Schools

The educational reforms of the late 1980s, in particular the Education Act (HMSO, 1988), ushered in a new era of autonomy for schools handing them greater control over their finances and strategic planning. One of the key measures of this act was a system of site-based management known as Local Management of Schools (LMS) which gave head teachers and governing bodies more autonomy over their school's budget (the assumption being that increased decision making at school management level would bring about improved educational outcomes for pupils; Levačić, 1998). Beforehand, schools were only responsible for the area of the budget relating to teaching materials (e.g., books and stationery) with the LEA held accountable for staff employment and buildings/premises maintenance. LMS gave schools responsibility for the majority of their budgets, which, in turn had implications for the role of the head teacher:

> LMS dramatically changed the role of the head teacher from educationalist to institutional manager. S/he now had to learn about recruitment and selection procedures, employment law, buildings maintenance etc.
>
> Gillard, 2007

This facilitated an emergence of specified bursarial roles, previously not seen or required, within the state school sector to accommodate this new level of responsibility. One head teacher at this time reports on the recruitment of a bursar in his secondary school following the onset of LMS and the new financial management responsibilities it would bring:

> the need for informed and inspired leadership is greater than ever but so is the threat that time for leadership will be swamped by the need for good management. Hence the need for a bursar who can remove the clutter of LMS from the academic and pastoral leadership in the school.
>
> Gittins, 1989, p. 169

From an early stage there was an awareness of the advantages of employing a bursar. Mortimore et al. (1994) underline the benefits of the role such as the streamlining of administrative and financial operations within a school and the relieving of heads and deputies of nonteaching duties. However, they also point to the "the disparity between the new post and the traditional secretary" and "the considerable workload, post-LMS, of the new post-holder" (p. 28). Knight (1993) goes further, suggesting possible tensions should nonteaching staff such as bursars become too involved in decisions concerning pupil education and that a newly appointed bursar "needs a thorough induction into this strange new world" (p. 49).

As the profile of school bursars was slowly developing so was their role evolving to meet the needs of a school system in flux. For instance, Szemeremyi (1991) discusses the possibility of schools generating their own income via fund raising and grant applications. He advises that "appeals to firms will have to be couched in more business-like terms" (p. 113) and composed as "methodologically prepared business plans" (p. 114) before concluding that the appointment of a "facilities development manager whose salary is paid from the extra income generated" (p. 114) would perhaps be the most sensible option. Without specifically mentioning bursars, the remit he describes is typical of a modern day SBM who is expected to have the requisite business acumen to raise money and to justify at least a portion of their salary from the income they generate.

The majority of these early references to bursars in state education focused on the secondary school sector which is unsurprising, given that they tend to be larger schools with wider bursarial needs (with regard to premises, buildings, and financial management) and bigger budgets to justify the salary of a bursar. Yet, there is evidence from this time that primary school head teachers were beginning to feel the strain of an ever increasing administrative burden coupled with their curriculum and leadership duties (Webb and Vulliamy, 1996) and that this was dampening the aspirations for headship for those leaders further down the career ladder such as those of deputy heads (Webb and Vulliamy, 1995). However, for most primary schools a bursar remained an unaffordable luxury (Knight, 1993) although there were suggestions that clusters of smaller schools might pool their finances to share a bursar as a way around this problem (Harrison and Gill, 1992).

Throughout the 1990s the role of the bursar continued to develop in tandem with the leadership and management requirements in schools. The following extract is taken from an interview with a head teacher speculating how the role of the bursar in her school might need to evolve to release her deputy head from an increasingly onerous management burden:

> We may need to change the role of the bursar. We might need someone to take on those activities and also to take away the detailed financial responsibilities carried by one of the deputies who, after all, should have an educational function … We have got to be aware, because of the new management functions we have had to take on, that senior staff do not lose sight of the fact that their main managerial functions should focus on the education of the pupils.
>
> Head teacher in Ribbins and Marland, 1994, p. 48

This gives an insight into the way schools were beginning to adapt and reorganize their staffing structures, to accommodate this new level of independence from the LEA, and the opportunities this would provide for bursars. However, it is also an indicator of the potential pitfalls of these changes in terms of the senior school leaders' workload priorities:

> If I think back, for example, to the start of LMS none of us was trained in managing budgets. None of us had come into teaching expecting that would be part of any role we would have to play as senior managers.
>
> Ribbins and Marland, 1994, p. 48

There were suggestions from academics that the shifting educational landscape was changing the culture of headship considerably so that the role was becoming complex and variable than ever before (Hellawell, 1991; Evetts, 1991; Weindling, 1992) and that the head teacher was being transformed into a kind of corporate manager (Evetts, 1994).

The stage seemed set for bursars to enter and relinquish head teachers and senior leaders of their nonteaching responsibilities though their steady influx into schools was not an entirely smooth transition. In a study of 18 secondary schools, exploring their use of resourcing post LMS, Thomas and Martin (1996) found that the majority had created a new bursarial position to accommodate the changes resulting from LMS although none were appointed at a grade comparable to the deputy head teachers and perceived, as they were, 'to be in a service and support role' (p. 57). They also discovered tension in one school whose governing body was concerned about the level of influence a bursar might develop given the budgetary information they had at their disposal. In the first book to focus entirely on bursarship, O'Sullivan et al. (2000) draw on their research with bursars in state schools to highlight the growing profile of the role and its increasing significance in the operation of the school. They describe the bursar "as a 'fixer' in a turbulent sea of problems and issues" from which they emerge "as a well-known, respected, and approachable colleague with considerable technical and managerial expertise" (p. 43). However, they also point to an "underlying tension and a sense of inferiority" whereby "teachers were regarded as finding it very difficult to come to terms with the idea of bursars as anything other than office clerks" (p. 68).

While it is not entirely straightforward trying to coherently piece together these rather patchy references to bursars there are some recurrent and consistent themes. Firstly, that the educational reforms of the late 1980s were having a significant impact on the amount and type of work that head teachers and senior school leaders were responsible for with schools adapting their staffing structures to cope with this additional workload. Secondly, although these new structures appeared to be alleviating the financial and administrative burden from teachers, allowing them to spend more time on the educational agenda, there were also inherent cultural tensions arising from these organizational changes. These themes would remain ever more relevant heading into the new millennium.

The Evolution of the SBM

In 2001, responding to a perceived crisis in the recruitment and retention of teachers due to heavy workloads, the Department for Education and Skills (DfES, as was), appointed PricewaterhouseCoopers to conduct a large-scale research study into teachers' and head teachers' workloads. The findings from the *Teacher Workload Study* (PricewaterhouseCoopers, 2001) were largely confirmatory highlighting increasingly onerous workloads for teachers and head teachers (compared to managers and professionals in other vocations) specifically the amount of administrative work and the lack of bursarial support, unrelated to teaching and learning, which could be carried out by support staff. Smithers and Robinson (2003) supported these findings in questioning over 1000 departing teachers across all phases of schools to reveal that almost 45% blamed a heavy workload as their main reason for leaving. According to Gunter and Butt (2005) an aging teaching population coupled with significant numbers of newly qualified teachers leaving the profession within just 3 years had led to "a generation of teachers in school who will retire in the next decade" and whose "replacements are not staying the pace" (p. 1). The government's response was an upheaval in the structure and design of the school workforce to facilitate support staff (such as secretaries, technicians, teaching assistants, and bursars) taking on more of the work traditionally done by teachers. *Raising standards and tackling workload: a national agreement* (DfES, 2003b) was the beginning of a period of what would come to be known as *workforce remodeling* in which support staff roles in schools were increased and developed to accommodate more responsibility.

Bursars had a significant role to play in this restructuring process, and in 2001 Estelle Morris, the then Secretary of State for Education, announced that the government would fund training for 1000 new school bursars within 5 years. The previous year the New Labor government had signaled their intent to raise the profile of school leadership and leadership preparation by establishing the National College for School Leadership (NCSL) with an aim of putting in place the world's largest leadership development program by 2004 (NCSL, 2001). It was the NCSL who the government commissioned to develop a national program of training for bursars in a bid to facilitate recruitment and professionalize the role while offering them a clear and structured career path. In 2003, following some pilot work, the Bursar Development Program (BDP) was rolled out nationally with an entry level Certificate of School Business Management (CSBM) covering the basic skills for school finance and resource management. This following year the more advanced Diploma of School Business Management (DSBM) was initiated and it was also around this time that the title of School Business Manager (SBM) was introduced to more accurately reflect the expanding nature of a role which was no longer predominantly concerned with school finances, as indicated by the more traditional title of 'bursar' (Southworth, 2010). The renamed School Business Manager Development Programme (SBMDP) would continue to expand adding an Advanced Diploma of School Business Management (ADSBM) in 2008 and a School Business Director (SBD) program in 2010 along with an SBM competency framework, developed with the National Association of School Business Managers (NASBM) illustrating the six main areas of this continuously evolving role.

School Business Management: A Growing Profession

With the establishment of a national training program and accompanying competency framework, SBMs (as they were now more commonly referred to) were becoming increasingly qualified and professionalized (Coulbeck, 2006). Furthermore, the nature of school leadership was changing to become more dispersed among senior staff members (Hartley, 2007; Gronn, 2003) and, as such, links were beginning to be forged between SBMs and the role they might play in the leadership and management of schools.

For example, drawing on participant evaluation and case study data from early cohorts of the CSBM, Moorcroft and Summerson (2006) found instances where head teachers had relinquished elements of operational control of the school to their SBMs and recruited them onto their Senior Management/Leadership Teams. In such cases the status of the SBM in their school was markedly improved as were the teaching staff attitudes toward the role. Conversely, they also found examples of SBMs facing barriers to change particularly where senior teaching colleagues were apprehensive about nonteaching staff taking on leadership roles. The authors suggest that this may be due to them having little understanding of the merits of the SBM role coupled with a lack of trust in their nonteaching counterparts. Similarly, Coulbeck (2006), working with SBMs on the C/DSBM courses, reports on firsthand accounts from participants to describe the positive impact they are having on the leadership, management, and administration of schools. Further, she argues that the growing number of professional qualifications for SBMs is a clear sign that school business management is emerging as a new profession. However, she also highlights a number of challenges faced by SBMs in establishing their professional identities including a lack of acknowledgment from teaching staff and governors; skepticism about the value of their professional training and; a reluctance of senior members of staff to recognize the candidacy of the SBM for membership of the SLT. These findings are akin to those from Aldridge (2008) who conducted cross-sectional surveys with SBMs on the CSBM to explore their roles in the evolving nature of school leadership; while his results indicated "distributed leadership teams, incorporating SBMs/Bursars, across many schools, their uptake is patchy with head teachers, governors and senior leadership members not always being aware of all the potential benefits" (p. 1).

The evidence presented here indicates the potential of SBMs to play an increasingly active role in school leadership set against the wider cultural barriers that accompany skepticism toward staff from nonteaching backgrounds taking on senior leadership roles in school.

Such notions are supported by Woods (2009) who, based on her own interviews with SBMs, draws attention to the growth of school business management, the link between the SBM role and distributed school leadership, and the tensions inherent in this relationship. She suggests that ongoing educational reforms have significantly altered the organization and purpose of schools leading to schools offering child welfare services far beyond the capacity of the traditional educational model (e.g., *Every Child Matters*, DfES, 2003a). Furthermore, schools are increasingly working in partnership with other schools (i.e., as federations and trusts) and morphing into larger, more complex structures (i.e., academy chains, learning communities) which "imply an increased role for SBMs in assisting their teaching colleagues in managing their schools' complex agendas in future" (p. 89). While acknowledging the progress and development of school business management Woods believes there remains a tension between teachers feeling their profession threatened by nonteaching staff and SBMs frustrated at being unable to realize their professional leadership training and potential within the school. She proposes that a clearer understanding from both groups of their mutual objective – of improving the educational and welfare outcomes of pupils – may help to harmonize the relationship in the future:

> Through fostering a shared understanding of how their contributions are different, but nonetheless essential, given the way schools now operate, potential for greater interdependence between teaching and administrative staff may result over the longer term.
>
> Woods, 2009, p. 90

Despite the cultural barriers faced by SBMs, the evidence presented here illustrates how far the role has progressed and developed since its earlier incarnation, the bursar, which first emerged in the state school system back in the late 1980s.

SBMs: An International Perspective

Internationally, there are signs of a similar push to raise the profile and develop the role of nonteaching school staff members with business, finance, administration, and human resource acumen as the requirement for roles incorporating these skills increases in educational settings across the globe.

In South Africa, a pilot group of 107 graduates became the first SBMs in their country to qualify for the inaugural Certificate for School Business Administrators (CSBA). This was the result of over 5 years of development work between the NCSL, Manchester Metropolitan University (MMU), and the Western Cape Education Department in which the NCSL provided course materials from its own CSBM to inform the design of this new accreditation. Covering areas such as financial management, human resources, risk assessment, facilities management, ICT systems, and the wider school environment the CSBA has, for the first time, given South African SBMs the opportunity to become qualified professionals (NCSL, 2010; Education Update, 2012).

As a profession, school business management is more established in Australia which has a number of professional organizations supporting SBMs (e.g., Association of Business Managers in Victorian State Schools (www.education.vic.gov.au), Schools Business Managers' Association Queensland Inc. (http://www.sbmaq.com.au)) although provision varies across the country with some regions yet to embrace the role. For example, primary and secondary school principals in New South Wales recently proposed the introduction of business managers into their schools to help alleviate their nonteaching responsibilities thus allowing them greater focus on the teaching and learning agenda. However, differences between states in regulations surrounding school staffing budget allocations mean that schools in Victoria for example have more flexibility in their workforce recruitment. As such, SBMs are a more familiar sight in Victorian schools (Newman, 2009). Research has also highlighted the range of responsibilities encompassed by SBMs across Australia including finance, facilities, and resource management; governance and strategic planning; human resources; accountability and legal oversight. Additionally, they often manage support staff, supervise extra-curricular provision (e.g., breakfast clubs, school trips), and are even

involved in making time tables and in teacher/classroom allocation (Starr, 2012).

In the United States, school business management is a well-established profession (the Association of School Business Officials International (www.asbointl.org) was formed in 1910) and forms an integral part of their education system. Unlike the United Kingdom and other countries discussed, School Business Officials (SBOs) operate at the district level in an overarching role often working with several schools and principals simultaneously although many of their areas of responsibility would be familiar to SBMs in other parts of the world. These include budgeting, purchasing, financial management, maintenance, human resources, ICT, pupil transportation and security, catering, and health care (ASBO, 2012). As business management is a district responsibility, school principals tend to be much more focused on classroom instruction and are able to spend a lot of their time working with students and teachers.

SBMs: Current Research

Recent large-scale studies have offered further illustrations as to the impact SBMs are having in schools. For instance, the NCSL funded a national program of interschool partnerships where clusters of schools were funded to develop bespoke models of shared, effective business management expertise. An evaluation of this program from researchers at the University of Manchester evidenced a variety of ways in which SBMs were positively impacting on the schools in which they were working. For instance, their experience in site and project management was relieving senior teaching colleagues of these duties thus saving them time and effort; they were taking advantage of economies of scale to secure best value from suppliers for the schools in their cluster; taking on responsibility for school policies and procedures (e.g., health and safety) thus relinquishing teaching staff from these obligations; utilizing their knowledge of human resources knowledge to aid the professional development of , often, large numbers of support staff employed across their clusters; and networking with other SBMs, outside agencies and the wider community to benefit their school(s). Their impact had also led to improvements in both the head teacher job satisfaction and pupil outcomes (University of Manchester, 2010). Moreover, evidence also emerged from this large-scale program of significant financial benefits associated with increased savings and higher income generation as a result of schools having access to a suitably qualified SBM (PricewaterhouseCoopers, 2010; Oakleigh, 2010).

Drawing on the outcomes of this research, Woods et al. (2013) highlight the benefits of school to school business management partnerships, with a suitable qualified SBM (or a range of staff members performing elements of the SBM function) as the fulcrum of such arrangements, as a viable alternative to school collaboration based on educational expertise. Similarly, Mertkan (2011) refers to the SBM role as a potential vehicle for closer collaboration and partnerships between schools and businesses while Southworth (2010) underlines how SBMs have moved on from dealing purely with school finances and now have a "portfolio of responsibilities and skills that supplement and complement those of other school leaders" (p. 3). He believes that this upsurge in the training and deployment of the thousands of new SBMs across schools in England over the past decade is a "revolution in education" (p. 3) to rival the rise in classroom teaching assistants or the escalation of school leaders operating beyond their own sites to support and lead other settings. Woods et al. (2012) explored the impact of SBMs on primary head teacher succession. Their findings indicate a number of ways in which having a suitably qualified SBM at that post could impact on school leaders at different stages of their careers and also aid leadership succession. These included making the head teacher role more attractive for aspiring leaders; easing the transition period from new or acting to full-fledged head teacher; assisting the head teacher in a challenging or crisis situation (e.g., a poor inspection); and helping retain experienced head teachers who were thinking of early retirement. Other research has uncovered evidence of SBMs rising to influential strategic leadership positions in their schools or being appointed as part of an improvement plan to turn around a failing school (Armstrong, 2012).

Looking Forward

Following over two decades of educational reform, instigated largely by the LMS initiative, increased demands on schools and school leaders have coincided with a restructuring of education to adopt more business like approaches. This has generated significant challenges to educational leaders whose workload has expanded in both volume and diversity with a range of pressures, in addition to the teaching and learning agenda (e.g., financial, legal, and technical) which have become impossible to deal with independently (PricewaterhouseCoopers, 2007). As the system has been rearranged, this has created new spaces that have been populated by nonteaching professionals with different skill sets to support and compliment the work of educational leaders. The election of a new coalition government in the United Kingdom in 2010 has intensified further the pace and intensity of educational reform. However, while previous governments had tinkered with deregulation and market forces, the new government has freed up the system to a previously unseen extent (Hatcher and Jones, 2011). Initiatives such as the academies programs have removed significant numbers of schools from the services of their Local Authorities (LA; essentially rendering this middle tier of government obsolete) and the introduction of Free Schools into the system has added a further dimension to this shifting educational landscape (Gunter, 2011).

For SBMs such changes have meant an increased workload as schools converting to academy status must find alternatives to the services their LA once supplied. Furthermore, the regulations surrounding the requirement for academies to employ a qualified accountant were recently relaxed thus creating opportunities for SBMs to take on the financial management of schools without the need for accountancy qualifications. As schools and educational settings moved forward and are increasingly reliant upon themselves for funds, services, and resources in a competitive business-like environment there is

an expectation that the SBM role will become a necessary requirement (DfE, 2010). It would, therefore, seem that SBMs are set to play an increasingly vital role in the modern state school system in England.

Bibliography

Aldridge, M., 2008. School Business Managers: Their Role in Distributed Leadership. How; Can SBMs/Bursars Compliment and Support Distributed Leadership? NCSL, Nottingham.

Armstrong, P., 3–6 January 2012. Managing the business of schools: perspectives on the expansion and 'professionalisation' of the school business manager role in England. In: Paper presented at the International Congress for School Effectiveness and Improvement (ICSEI). Sweden, Malmo.

Association of School Business Officials International, 2012. About ASBO International. From: http://www.asbointl.org/AboutASBO.htm (accessed on 21.08.12.).

British Broadcasting Cooperation, 2012. History: John Maynard Keynes. From: http://www.bbc.co.uk/history/historic_figures/keynes_john_maynard.shtml (accessed on 13.08.12.).

Coulbeck, J., 2006. School business managers: the emergence of a new profession? Professional Development Today 9 (3), 14–19.

Department for Education and Skills, 2003a. Every Child Matters. HMSO, London.

Department for Education and Skills, 2003b. Raising Standards and Tackling Workload: A National Agreement. DfES, London.

Department for Education, 2010. The Schools White Paper 2010: The Importance of Teaching. From: http://www.official-documents.gov.uk/ (accessed on 23.08.12.).

Education Update, March 2012. Pilot Group in School Business Management Course Graduate, ed. 12. From: http://www.designinfestation.com/WCEDNEWS/index.php?option=com_content&view=article&id=84%3Apilot-group-in-school-business-management-course-graduate&Itemid=2&lang=en. (accessed on 21.08.12.).

Evetts, J., 1991. The experience of secondary headship selection: continuity and change. Educational Studies 17 (3), 285–294.

Evetts, J., 1994. Becoming a Secondary Headteacher. Cassell, London.

Gillard, D., 2007. Education in England: A Brief History. From: http://www.educationengland.org.uk/history/ (accessed on 13.08.12.).

Gittins, C., 1989. A bursar in a secondary school. In: Fidler, B., Bowles, G. (Eds.), Effective Local Management of Schools: A Strategic Approach. Longman, Harlow.

Gronn, P., 2003. The New Work of Educational Leaders. Paul Chapman Press, London.

Gunter, H.M., Butt, G., 2005. Challenging modernization: remodelling the education workforce. Educational Review 57 (2), 131–137.

Gunter, H.M. (Ed.), 2011. The State and Education Policy: The Academies Programme. Continuum, London.

Harrison, M., Gill, S., 1992. Primary School Management. Heinemann, Oxford.

Hartley, D., 2007. The emergence of distributed leadership in education: why now? British Journal of Educational Studies 55 (2), 202–214.

Hatcher, R., Jones, K. (Eds.), 2011. No Country for the Young. Tufnell Press, London.

Hellawell, D., 1991. The changing role of the head in the primary school in England. School Organisation 11 (3), 321–337.

HMSO, 1988. Education Reform Act. HMSO, London.

ILEA, 1984. Improving Secondary School (The Hargreaves Report), Report of the Committee on the Curriculum and Organisation of Secondary Schools. ILEA, London.

Kerry, T., 2001. Working with Support Staff: Their Roles and Effective Management in Schools. Pearson Education, Harlow.

Knight, B.A.A., 1993. Financial Management for Schools: The Thinking Manager's Guide. Heinemann Educational, Oxford.

Levačić, R., 1998. Local management of schools: results after six years. Journal of Education Policy 13 (3), 331–350.

Mertkan, S., 2011. Leadership support through public-private partnerships: views of school leaders. Educational Management Administration & Leadership 39 (2), 156–171.

Moorcroft, R., Summerson, T., 2006. Leaders backing leaders: a programme of school business management. Journal of In-service Education 32 (2), 255–274.

Mortimore, P., Mortimore, J., Thomas, H., 1994. Managing Associate Staff: Innovation in Primary and Secondary Schools. Paul Chapman, London.

National College for School Leadership, 2001. First Corporate Plan: Launch Year 2001–02. NCSL, Nottingham.

National College for Leadership of Schools and Children's Services, 2010. School Business Management - a First for South Africa. From: http://www.nationalcollege.org.uk/index/about-us/news/news-item.htm?id=133280 (accessed on 01.08.12.).

Newman, A., 2009. The Impact of School Bursars (Business Managers) on the Role of the Principal in Leading and Managing the School Educationally. Leadership Fellowship 2007–08, Jim Harkin Award. NSW Department of Education and Training, New South Wales.

Oakleigh Consulting, 2010. Financial Impact Assessment of the National College's School Business Manager Demonstration Projects. National College for Leadership of Schools and Children's Services (NCSL), Nottingham.

O'Sullivan, F., Thody, A., Wood, E., 2000. From Bursar to School Business Manager: Reengineering Leadership for Resource Management. Pearson Education, Harlow.

PricewaterhouseCoopers, 2001. Teacher Workload Study. DfES, London.

PricewaterhouseCoopers, 2007. Independent Study into School Leadership. DfES, London.

PricewaterhouseCoopers, 2010. Cost Benefit Analysis of the School Business Management Programme. National College for Leadership of Schools and Children's Services (NCLSCS), Nottingham.

Ribbins, P., Marland, M., 1994. Headship Matters: Conversations with Seven Secondary School Headteachers. Longman, Harlow.

Riches, C.R., 1981. Non-teaching Staff in Primary and Secondary Schools, E323 Management and the School, Block 6, the Management of Staff. Open University Press.

Riches, C.R., 1982. The micropolitics of the management of non-teaching staff. Educational Management Administration & Leadership 10, 151–156.

Smithers, A., Robinson, P., 2003. Factors Affecting Teachers' Decisions to Leave the Profession. DfES, Nottingham.

Southworth, G., 2010. School Business Management: A Quiet Revolution. National College for School Leadership (NCSL), Nottingham.

Starr, K., 2012. Worth their weight in gold: the boon of having a school business manager. School Business Affairs 78 (6), 29–30.

Szemeremyi, S., 1991. Money matters. In: Stevens, D.B. (Ed.), Under New Management: Strategies for Secondary Schools in the 1990s. Harlow, London.

Thomas, H., Martin, J., 1996. Managing Resources for School Improvement: Creating a Cost Effective School. Routledge, London.

University of Manchester, 2010. School Business Manager Demonstration Project Evaluation Study – Final Report. National College for Leadership of Schools and Children's Services, Nottingham.

Webb, R., Vulliamy, G., 1996. The changing role of the primary-school headteacher. Educational Management Administration & Leadership 24 (3), 301–315.

Webb, R., Vulliamy, G., 1995. The changing role of the primary school deputy headteacher. School Organisation 15, 53–63.

Weindling, D., 1992. Marathon running on a Sand Dune: the changing role of the headteacher in England and Wales. Journal of Educational Administration 30 (3), 63–76.

Woods, C., 2009. Remodelling and distributed leadership: the case of the school business manager. In: Chapman, C., Gunter, H.M. (Eds.), Radical Reforms: Perspectives on an Era of Educational Change. Routledge, London, pp. 80–90.

Woods, C., Armstrong, P., Pearson, D., 2012. Facilitating primary head teacher succession in England: the role of the school business manager. School Leadership & Management 32 (2), 141–157.

Woods, C., Armstrong, P., Pearson, D., 2013. Perfect Partners or Uneasy Bedfellows? Competing understandings of the place of business management within contemporary educational partnerships. Educational Leadership and Management 41 (6), 751–766.

School Choice

Mark Berends, Center for Research on Educational Opportunity (CREO), Research Center on Educational Policy (RCEP), University of Notre Dame, Notre Dame, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

School choice is one of the several educational reforms being implemented in the United States to address persistent inequalities among students and the communities in which they live. Yet, school choice takes on a variety of forms, whether that be private or religious schools, charter public schools, choice scholarships or voucher programs, educational tax credits, or homeschooling. This article examines the research on these various aspects of choice in the United States, placing the research in broader theoretical frameworks, and directing attention to future research to inform the conditions under which school choice is effective or not.

Introduction to School Choice

Although there are a myriad of school reforms that have been proposed and implemented over the past several decades, one that receives a great deal of attention is school choice, which refers to a variety of programs providing families the option to choose the school their children attend. School choice options may include neighborhood public schools, private schools, magnet schools, charter public schools, vouchers or tuition tax credits, homeschooling, inter- and intradistrict choice plans, and supplemental educational services (Berends et al., 2009; Berends, 2014). The public may not understand the meaning of the many forms of school choice, but the idea that parents should have some choice in the education of their children is deeply ingrained in the U.S. culture as well as many other countries (Berends et al., 2009, 2011).

In this article, I will place the research on school choice within a theoretical framework, drawing on market theory and institutional theory. I then will discuss the research and findings on the various aspects of school choice in the United States and other countries, focusing on Catholic schools, charter public schools, choice scholarships or voucher programs, education tax credits, and homeschooling. After highlighting the findings from the more recent rigorous studies on school choice effects, I will point the direction toward research that draws on the social organization of schooling as a way to further research in this area. In this perspective, researchers focus their attention inside schools – examining the relationships between teachers and students as well as the other social relationships among parents, peers, and administrators – to understand how the organization of schools influences the learning opportunities and schooling processes within schools (Bidwell and Kasarda, 1980). Summarizing the importance of this perspective when examining school choice, Schneider (2003: 212) writes

> Studies of school choice have also not yet overcome the inadequacies we find in studies of the social organization of public schools. One reason may be that the pressure to link school choice with gains in student achievement has placed issues of selection bias and test score differences at the forefront, thus overshadowing serious efforts to examine qualitative differences in the learning opportunities experienced in choice versus other types of school settings.

Evolving Aspects of School Choice in the United States

Federal policies have provided support for school choice over the past decade. The No Child Left Behind Act (NCLB) promoted choice for families if schools failed to make adequate progress over time. As part of the more recent $4.0 billion Race to the Top program, the Obama administration has called for an expansion of charter schools, awarding funding to states for school reforms related to adopting standards and assessments for college and workplace readiness, recruiting effective teachers and principals, building data systems to measure student progress and inform educators, turning around low-achieving schools, and expanding the number of charter schools.

Some choice options that have grown significantly in the past two decades is parents sending their children to charter schools – schools that are publicly funded but run under a charter by parents, educators, community groups, universities, or private organizations to encourage school autonomy and innovation. As the fastest growing area of school choice, there are currently over 6000 charter schools serving more than 2.2 million children across the United States or 4.5% of the K-12 student population (National Alliance for Public Charter Schools, 2014; U.S. Department of Education, 2012).

Another growing area of school choice is scholarships or vouchers and tuition tax credits. *Vouchers* or *choice scholarships* provide parents the option of sending their children to the school of their choice, whether public or private, religious or nonreligious schools. With public funds usually spent by the district, vouchers are allocated to families for tuition payments, in part or perhaps in full. *Education tax credits* are provided to either corporations, individual families, or both. At the *corporate* level, businesses receive a state tax credit for making donations to nonprofit organizations called Scholarship Granting Organizations (SGOs), which use the donated money to fund private school scholarships for students. Anyone can start an SGO, and some establish certain income criteria for students to be eligible for scholarships. At the *individual* level, parents can receive a tax credit or tax deduction from state income taxes for approved educational expenses, which typically include private school tuition, books, supplies,

computers, tutors, and transportation. Some programs have income restrictions for family eligibility or set the amount families can claim on their taxes.

The number of students participating in different voucher programs across the nation has increased significantly in the last decade – even though the total number of students receiving vouchers is a small fraction of the total number of students in the United States (see Berends, 2014). (For a description of these voucher programs and the urban locales in which they are being implemented, see http://www.edchoice.org/School-Choice/School-Choice-rograms.aspx.)

For example, one of the more well-known voucher programs is the Milwaukee Parental Choice Program, which started in 1990; the program provides vouchers (capped at $6442) to families earning up to 300% of the federal poverty level (about $69 000 for a family of four in 2012). In 1991, 337 Milwaukee students received a voucher; in 2013, 24 027 students received one. Another voucher program of note is Florida's John M. McKay Scholarships for Students with Disabilities, providing vouchers to public school students with disabilities to attend private or other public schools. Starting in 2000, the program served 23 011 students in 2013. A more recent voucher program is Indiana's Choice Scholarship Program, which allows student in low- and middle-income families to receive vouchers to attend private schools. Beginning in 2011 with 3919 students, the number of participating students increased to 20 047 students in 2013–14, making the program one of the largest in the United States.

Another area of growth in school choice is the home-schooling – the education of students at home, typically by a parent. The number of homeschooled students in the United States has increased from 850 000 in 1999 to over 1.5 million in 2007. In 2007, about 2.9% of the K-12 student population was homeschooled (U.S. Department of Education, 2012).

Despite the growth in charter schools, voucher programs, education tax credits, and homeschooling, other areas of school choice are declining. Nationally representative data reveal that between 1995–96 and 2009–10, there has been a general decline of students in pre-K-12 private schools in the United States (Aud et al., 2013). This overall trend masks some shifts in the private sector between 1996 and 2010. For example, there has been a decrease in the number of students attending Catholic schools and increases in the number of students attending unaffiliated private schools – schools that have a more general religious orientation or purpose, not identifying with a specific religion or faith (Aud et al., 2013).

The Catholic school sector in particular is undergoing significant changes. One important change is due to severe financial crises, resulting in many schools either closing or consolidating with other Catholic schools. At the same time, increasing operational expenses in conjunction with dire economic conditions make it difficult for many schools to survive. Between 2000 and 2011, 1755 Catholic schools either closed or consolidated, and there has been a Catholic school enrollment decline of over 22% (McDonald and Schultz, 2011; see also Uecker, 2008). Because of the changing demographics surrounding the Catholic sector in the United States, continued research on Catholic school effects is worthwhile to examine changes over time.

Theoretical Perspectives on School Choice

As school choice programs continue to expand, additional research will likely increase our understanding of the impact of choice options on families and students. As this research unfolds, it is important to keep in mind different theoretical frameworks that guide the research. Specifically, there are two competing theories about the possible impact of choice schools on teaching and learning and in school organizational conditions – market theory and institutional theory (Berends et al., 2010). Many reformers maintain that market style mechanisms of consumer choice and competition between autonomous schools will encourage diverse and innovative approaches to school organization, teaching, and learning (e.g., Chubb and Moe, 1990; Walberg, 2011; Walberg and Bast, 2003). The assumption is that as school choice undercuts bureaucratic political control of public education, it provides educators in schools of choice the opportunity and motivation to experiment with new organizational and instructional strategies to improve student achievement.

Proponents of choice argue that providing this freedom not only diversifies educational opportunities but also creates incentives for the improvement of traditional public schooling through increased market competition for services (Chubb and Moe, 1990; Friedman, 1962). In large part, this argument is about how market competition decreases the amount and influence of historical bureaucratic structures to increase the opportunities for parents and school staff to better relate to address parents' demands.

Some hold that privatization and school choice can bring about 'creative destruction,' borrowing from Joseph Schumpeter, who in the 1940s argued that entrepreneurs relied on radically new technologies – whether more effective, more efficient, or both – to promote economic progress and replace older technologies (Walberg, 2011; Schumpeter, 1942). As the theory goes, such privatization and increased choice will lead to better outcomes, lower costs, and greater satisfaction of employees, parents, and students. According to Walberg (2011: 73), "market-based consumer-driven school choice seems the best hope for creative destruction of new technologies, the expansion of choice, competition, and diversity for substantial, sustained achievement improvement." Chubb and Moe (1990: 217) hold a similar position, calling choice a "revolutionary reform that introduces a new system of public education."

Critics of the market model, however, raise questions about the empirical validity of its key assumptions about parent-consumers (demand-side), schools (supply-side), and the products that a market in education would generate (Finnegan, 2007; Henig, 1999). From such criticism comes an alternative theory about the consequences of school choice: institutional theory. Stemming from broader organizational analysis, this *new institutionalism*, developed by John Meyer and colleagues over several decades (Meyer, 1977; Meyer and Rowan, 1977, 1978; Powell and DiMaggio, 1991; Scott and Meyer, 1994; Scott and Davis, 2007), characterizes schools as

institutions with persistent patterns of social action that individuals take for granted.

Agreeing with market theorists that the bureaucratic form of schooling dominates the public school sector in the United States (and many other countries), institutional theorists take a different tack in their analysis of the education environment. For instance, the increase in bureaucratization of schools has led to an increase in rational coordination among the nested layers of the school – from the federal government to the state, districts, schools, and classrooms. According to institutional theorists such as Meyer and Rowan (1977, 1978), this bureaucratic, rational network has resulted in a system of categories or rules, called 'ritual classifications,' that define the actions of schools, teachers, and students. Over time these ritual classifications become institutionalized and accepted as the norm for what constitutes a legitimate school and its activities (Bidwell and Kasarda, 1980). Institutional theorists refer to this as *isomorphism* and have documented its diffusion both in the United States and throughout the world (Meyer and Ramirez, 2000; Bidwell and Dreeben, 2006).

School compliance to ritual classifications is important for legitimacy – more important, according to institutional theorists, than maximizing efficiency and innovations of school operations (Meyer and Rowan, 1977; Scott and Meyer, 1994). In other words, schools adapt to their environments by adopting accepted rules and structures, leaving actual classroom instruction and learning relatively unexamined and unmonitored. Such loose coupling helps schools maintain their validity (Weick, 1976) and is further promoted by schools' logic of confidence that delegates instruction to teacher professionals who ultimately control what goes on inside their classrooms. It is the teachers who typically have made the choices about what to teach and how to teach in their classrooms – a set of activities that until more recently has been beyond the purview of principals, parents, and policymakers at the district and state levels (Weick, 1976; Elmore, 2007). With the implementation of NCLB in the United States over the past decade, states and districts have attempted to guide instruction in schools through high-states accountability, but even with specific instructional materials, it is the teachers in the classrooms and their relationships with students who are making choices about what to teach, how to teach, and how to learn (see Stecher and Vernez, 2010).

DiMaggio and Powell (1983) argue that there are different types of environmental pressures on organizations to make them more similar than different. These pressures include 'coercive isomorphism,' which stems from formal and informal pressures by organizations and groups on which the school depends (e.g., federal and state mandates under NCLB), 'mimetic isomorphism,' which stems from the adoption of similar structures and practices when facing uncertain tasks; and 'normative isomorphism,' which stems primarily from professionalization of educators and professional networks.

When applied to school choice, institutional theory emphasizes that *all* schools operate within highly institutionalized environments, which define what counts as legitimate education. All types of schools, no matter the sector or organizational form, adopt rituals, norms, and myths to support their validity (Meyer and Rowan, 1977, 1978; Scott and Davis, 2007). Thus, even schools of choice pay attention to institutional rules such as teacher certification, curricular subject matter, instructional time, reasonable class size, and mostly age-based grade organization.

In short, institutional theorists argue that the institutional environment of education is so strong that significant changes in instruction are likely to be rare or short lived (see Elmore, 2007), and market theorists claim that increased choice will result in widespread autonomy promoting innovation, competition, and increased satisfaction and outcomes. Despite a couple of decades of school choice reform, researchers have not definitively supported market theory vis-à-vis institutional theory. The limited empirical research is mixed on improved and differentiated instruction and in-school organizational conditions, curriculum content, and pedagogy in schools of choice, supporting neither market nor institutional theories (see Berends et al., 2010; Lubienski, 2003; Preston et al., 2012).

Private and Catholic Schools

Many people argue that public schools are outperformed by private schools, particularly Catholic schools (Chubb and Moe, 1990; Coleman and Hoffer, 1987; Coleman et al., 1982). However, the magnitude of these effects and their implications are often the center of heated debate (see Lee and Bryk, 1993). Much of this debate focuses on whether students attending Catholic schools score higher on achievement tests than students in public schools. When comparing sector differences, there is a continuing concern about selection bias – that is, there is something different about students who attend Catholic schools vis-à-vis public schools, such as social background, motivation, values and beliefs, among other characteristics, some measureable and some difficult to measure (Goldberger and Cain, 1982).

Within these comparative analyses of school sector effects on student outcomes, researchers often focus on the issues of the *Catholic school advantage* (CSA) and the common school ideal (see also Hallinan and Kubitschek, 2010, 2012; Marks, 2009). That is, the CSA refers to the higher educational outcomes of Catholic school students compared with their public school peers, whether that be higher academic achievement or college attendance. The common school ideal, according to Hallinan and Kubitschek (2012), "signifies the goal of providing equal access to learning opportunities for all students."

In terms of academic achievement, the research to date suggests that there is variability in the effects of Catholic schools by level (e.g., high school vs elementary). For example, when considering effects on mathematics achievement, there is some evidence that the effect of attending a Catholic high school on student achievement has been consistently positive in nationally representative cohorts of high school students from the early 1980s to the present. For example, in his review of the average achievement differences among high school seniors, Hoffer (2009: p. 432) states that "the Catholic-public differences in mathematics were 0.48 SD units in the 1982 HSB survey (Hoffer et al., 1985). 0.37 SD units in the NELS:88 sample (Morgan, 2001), and 0.50 SD units in the ELS:2002 sample (Carbonaro and Covay, 2010)."

Controlling for important measures in nationally representative data from the 1980s, Coleman and Hoffer (1987) found that between grades 10 and 12, students in Catholic schools

outperformed public school students by about one grade level equivalent in both mathematics and reading. With more sophisticated multivariate models in more recent nationally representative data (2002–04), Carbonaro and Covay (2010) found that Catholic school students outperformed their public school counterparts in mathematics.

At the elementary school level, Catholic school effects seem less robust. For instance, analyzing the Early Childhood Longitudinal Study (ECLS-K), Carbonaro (2006) found that kindergarten students in public and Catholic schools experienced similar achievement gains in mathematics and general knowledge, net of other characteristics. Analyzing the ECLS-K data and the gains of students between the 3rd and 5th grades, Reardon et al. (2009) found that public school students outperformed their Catholic school peers in math but experience similar gains in reading.

Further research comparing Catholic schools at the pre-K through 8th grade levels would help shed light on the sector differences, with an emphasis on theoretical development and educational policy implications. In particular, further examination of the organizational and instructional conditions may explain sector differences at earlier grade levels and the variability between as well as within sector (Bryk et al., 1993; Berends et al., 2008). For example, at the high school level, a consistent finding in explaining the CSA has been the academic focus and more rigorous course taking patterns of Catholic school students compared with their public school peers. Yet, at the elementary grade levels, these organizational and instructional conditions have not been fully analyzed. In addition, there are several other factors – such as communities of learning and trust – that prior research and theory have pointed to but so far not fully examined as important for explaining sector differences (Bryk and Driscoll, 1988; Bryk and Schneider, 2002), social capital (Berends and Zottola, 2009; Hoffer, 2009; Marks, 2009), peer effects (Marks, 2002), or market vis-à-vis institutional hypotheses (Berends et al., 2010).

As mentioned above, to date most scholars – particularly sociologists – have focused on student outcomes in terms of standardized test scores. Others – including economists – have concentrated on college attainment, since this is an outcome of interest that has more direct economic implications for labor market participation and earnings (see Marks, 2009). Several studies analyzing nationally representative data from the 1980s to the present reveal that students attending Catholic schools are more likely to attend college (Altonji et al., 2005; Evans and Schwab, 1995; Grogger and Neal, 2000; Nguyen et al., 2006).

Internationally, there is a great deal of variation in the size of the private school sector, ranging from about 10% in the United States to 6% in the United Kingdom, 38% in Australia, 56% in Korea, 61% in Ireland, and 77% in the Netherlands (see Woessman, 2009). In their analyses of international data for 29 countries in the Programme for International Student Assessment 2003 data, Woessman et al. (2009) find that students' achievement scores are higher in countries that have a greater percentage of privately managed schools. They go on to argue important mechanisms explaining these results, including private schools' freedom to make autonomous decisions about school organization and processes and accountability systems that inform parents' choices. Although informative, it is important to know that these findings are correlational and may be subject to significant selection bias (for a review of international assessment data and private schools, see Woessman (2009)).

Charter Public Schools

Charter schools are public schools funded by the government, but they have a different governance structure compared with traditional public schools in that they are established under a charter by parents, educators, community groups, or private organizations to encourage school autonomy and innovation. In exchange for such autonomy and flexibility, charter schools are held accountable to current state and federal accountability standards, such as NCLB, which requires testing in certain grades and sets performance targets over time (overall and various subgroups). When a charter school has more students applying than there are seats available (i.e., oversubscription), the school is required to hold a lottery to select students for open seats at random.

Although the charter school sector is the fastest growing area of school choice, the evidence base is weak for scaling up charter school reform via federal policy and programs. For instance, some studies using randomized designs show modest effects on academic achievement gains for students in charter schools compared with those students who are not so enrolled (Abdulkadiroglu et al., 2009; Angrist et al., 2011; Dobbie and Fryer, 2011; Hoxby and Murarka, 2008; Hoxby et al., 2009). Other studies relying on broader samples of schools within a randomized design (Gleason et al., 2010) and those using quasi-experimental methods show mixed results for charter school effects on achievement (for a review see Betts and Tang, 2011; CREDO, 2013; Teasley, 2009).

Although it is important to understand that overall charter school studies show that effects on achievement are mixed (some positive, some negative, and some neutral), it is important to note that some studies have found significant and substantial positive effects of charter schools, particularly in urban areas where it has been difficult to implement meaning educational reforms. For example, comparing students who won and lost charter school lotteries, Hoxby et al. (2009) found that charter students outscored the comparison group in both mathematics and English. The authors noted that students who attended charter schools in New York City over a longer period of time (e.g., kindergarten through 8th grade) matched the mathematics performance gains of their peers in affluent suburban schools – what they called the closing of the 'Harlem–Scarsdale' achievement gap. Dobbie and Fryer (2011), who examined students who won and lost the charter school lotteries in the Harlem Children's Zone, found that the effects of charter elementary schools were large enough to close the racial achievement gap across subjects – i.e., students gained about 0.20 of a standard deviation a year in both mathematics and English/language arts. Large effects of charter schools have also been found in Boston (Angrist et al., 2011).

Although studies of school choice shed some light on the main effects in different locales, they provide limited information about the schools as organizations and the conditions within them that may promote student achievement,

particularly the curriculum and instruction that is most likely to affect student learning. Many researchers and policymakers advocate looking inside schools to better understand the conditions under which schools of choice have (or do not have) positive effects on achievement, pointing to the importance of detailed information about curriculum, instruction, organizational conditions that promote achievement, and teacher characteristics and qualifications (Berends et al., 2008, 2010; Betts and Loveless, 2005; Gill et al., 2007; Zimmer and Buddin, 2007, 2008; Zimmer et al., 2003). To date, however, these calls to look at the organization of schooling within the charter sector have not been heeded, by and large – suggesting an important avenue for future research.

This is an important research agenda for cross-national comparisons as well, for we know that many other countries have 'independent schools' that are publicly funded, but privately managed. For instance, in the United Kingdom, there are several models that represent a basic departure from the "'pure model' of sole public funding and sole public provision" of educational services (Brewer and Hentschke, 2009: p. 228). These models of departure include the foundation schools, grant-maintained schools, city technology schools, and academies, and school founders in a many of these cases can be for-profit businesses or faith-based organizations. School founders can play a major role in determining the instructional design, curriculum, and appointing members to the governing body if the founders put in an investment of 10% of the academies' capital costs (or about £2 million, whichever is less).

In the Netherlands, independent schools are the norm; the government provides financial support to private schools (including religiously affiliated institutions). Schools have control over curriculum and the hiring of teaching staff, but they cannot exceed the national standard of student–teacher ratio. Another important constraint is the national curriculum standards and national exit exams at the end of primary and secondary schooling (Brewer and Hentschke, 2009).

Likewise, more recently in Qatar there is a significant growth in publicly funded, but privately operated independent schools (or charters). First introduced in 2002, about half of the nation's students were enrolled in independent schools by 2008. Brewer et al. (2007) find that the founders varied from those from the business sector to former school administrators and educational ministry officials.

As these examples reveal, there is much to learn about how independent (or something very similar to charter schools) are operated and how they affect students' academic and social outcomes.

Choice Scholarships and Voucher Programs

Besides the changes in the charter school sector, another choice option is vouchers, or school choice scholarships, that provide parents with public funding to send their children to the school of their choice, which includes public or private schools (religious or nonreligious). These public vouchers are often means-tested and are allocated to families for tuition payments to pay for part or all of the tuition. Vouchers have also been highly contested and debated in educational policy, the media, and research communities. For instance, studies in the 1990s of the Milwaukee voucher program, researchers disagreed about the effects of the program on student outcomes (see Berends et al., 2011). Although policymakers suspected that evaluation of the program would settle disputes about the effects of vouchers on students, the research findings instead were nuanced and mixed (for review, see Figlio, 2009; Zimmer and Bettinger, 2008).

With the expansion in the number of voucher programs, the research addressing the effects of these programs has increased as well. However, the research portrays a mixed view of voucher impacts (Figlio, 2009; Zimmer and Bettinger, 2008). The first study of a school voucher program in the United States was the official evaluation of the pioneering Milwaukee Parental Choice Program, 1990–95 (see Witte, 2000). Subsequent work analyzed a portion of the data generated by that evaluation, and was reported in Greene et al. (1999b) and Rouse (1998). These studies were different from Witte in their methodological approaches and, therefore, the findings reported differed. Based on regression models strengthened with Heckman selection corrections, the Witte work found generally no systematic differences on student achievement outcomes between voucher and public school students. The Greene et al. analysis was based on a subset of voucher participants who had won their voucher via a lottery system to a small number of oversubscribed private schools, finding positive achievement impacts associated with participation. Rouse employed a series of quasi-experimental approaches from student fixed effects and instrumental variable designs, finding no effect in reading but positive achievement impacts in math.

These differences between findings of positive and no achievement effects have since been reflected in other studies of voucher programs. For example, Greene (2001) found positive achievement impacts from an experimental analysis of the privately funded voucher program in Charlotte, a result generally supported by Cowen (2008) using more sophisticated statistical models. Howell et al.'s (2002) primary analysis of lottery-based privately funded programs in New York City, Washington, DC, and Dayton, Ohio revealed positive student achievement outcomes for African-American students, but not for the overall sample, a finding both confirmed (Barnard et al., 2003) and questioned (Krueger and Zhu, 2003) by subsequent analysts of the New York data. Studies of the Cleveland Scholarship Program similarly found both positive impacts for voucher students (Greene et al., 1999b) and no significant differences (Metcalf et al., 2003), findings that differed due to study design and sample. Using a regression discontinuity approach on the statewide voucher system in Florida, Figlio and Hart (2011) found that the program generated statistically significant positive impacts on student reading achievement, at least for students near the income-eligibility ceiling. Witte et al. (2012) reported that Milwaukee voucher students on average gained more than a matched sample of public school students in reading but not in math in the final year of that 4-year evaluation. They noted that a new high-stakes testing policy may have been partly or wholly responsible for the voucher gains.

The evaluation of the first federally funded voucher program in Washington, DC relied on an experimental design based on scholarship lotteries (Wolf et al., 2010, 2011). The series of

reports described significant achievement gains in reading in the third year of the evaluation but no significant reading impacts in other years, including the fourth and final year, or in math in any year. This same study revealed significant gains in voucher students' high school graduation rates. Cowen et al. (2013) reported gains in educational attainment from the Milwaukee voucher program, both in terms of rates of high school graduation and 4-year college enrollment. Recently, Chingos and Peterson (2012) found that African-Americans who participated in the New York City privately funded voucher experiment enrolled in college at higher rates than the experimental control group, though there were no significant attainments effects of the program on the entire sample of students.

Internationally, the research on voucher programs is more positive in terms of effects on student achievement scores. Reviewing studies from Chile, Columbia, Denmark, South Korea, Sweden, the Netherlands, and Denmark, Barrera-Osorio and Patrinos (2009) summarize the international research on vouchers as follows: "The most rigorous studies available show that voucher programs lead to significant improvements in access to secondary schooling for relatively poorer students; these programs also lead to significant increases in test scores. Additionally, studies show that voucher recipients are more likely to complete secondary school, enter university, postpone marriage, and increase their earnings" (p. 339).

Similar to the research on charter and independent schools, there have been very few studies that examined the specific learning conditions that students experienced in their schools of choice vis-à-vis comparable students in traditional public schools (Figlio et al., 2013; Zimmer and Bettinger, 2008). Although such studies are difficult to design and implement, more research is needed on school and classroom experiences to understand the conditions under which voucher programs provide more meaningful and substantive learning opportunities (or not).

Education Tax Credits

Another school choice option that has increased dramatically over the past decade is providing education tax credits to businesses and individuals. Parents can receive a tax credit or deduction for educational expenses (tuition, books, supplies, transportation), and businesses can receive a tax credit for making donations to nonprofits called SGOs, which use the funds for private school student scholarships. Often education tax credits are seen as equivalent to vouchers in that they provide students with funds to attend private schools; it is the transfer mechanisms that differ. In addition, education tax credits have been seen as a more palatable way to get legislation passed in different states because of the controversies and debates around vouchers and the direct provision of public funds to attend private schools. Education tax credits are a more indirect transfer of public funds.

Perhaps because education tax credits seem more acceptable in state legislatures and are not under scrutiny like voucher programs, there has been a great deal of policy activity at the state level since the 1980s, and certainly over the past few years (Huerta and d'Entremont, 2007; Schaeffer, 2009). For example, there were a total of 26 tax credit scholarship bills in 2008, and over a 3-year period this number more than doubled to 53 in 2011 (Berends, 2014).

The research on education tax credits is very limited (as mentioned by Belfield, 2005; Huerta and d'Entremont, 2007; Schaeffer, 2009; Zimmer and Bettinger, 2008), probably because education tax credits are often equated with vouchers. Similar to voucher programs, which in the end result in students transferring from the public to private sector, more information is needed about the instructional practices and school organizational conditions that students using tax credits experience compared with their counterparts. Some questions that might further our understanding of tax credit effects include: Who chooses to apply and use a tax credit? What are the characteristics of students and their families receiving tax credits? To what schools do such students apply and attend? What are the learning opportunities in these schools vis-à-vis the schools that students attended before? What are the differences in the organizational and instructional conditions in the schools of destination (attending with vouchers) compared with schools of origin (where students attended previously)?

Unfortunately, the research appears to be less than conclusive – 'unfortunate' because this provides fodder for continued debate at different levels of policy about whether or not to scale up these choice options. What is needed is more systematic research that looks at the educational trajectories of students in choice and nonchoice schools with data not only on test score gains and graduation rates, but other noncognitive measures of student outcomes (e.g., behavior, engagement, motivation, educational and occupational expectations). In addition, more systematic information needs to be gathered from the choice and nonchoice learning environments, including not only instructional conditions but also differences in the social organization of schools (see Berends et al., 2010).

Homeschooling

Homeschooling is one of several parental choice options, and often, it is overshadowed by other aspects of choice, such as private or Catholic schools, charter schools, voucher programs, and tuition tax credits (see Berends et al., 2009). Nearly 3% of the K-12 student population in the United States is homeschooled – making the homeschooling sector nearly as large as the sectors of Catholic schools and charter schools. Yet, homeschooling remains one of the most understudied areas of school choice. Unlike research on other areas of school choice – such as the research on charter schools and voucher programs that has produced a growing evidence base for understanding impacts on students – the research on homeschooling is sparse and disappointingly weak.

We do know a few descriptive aspects of homeschoolers. For instance, the families that tend to homeschool their children tend to be better educated and solidly middle class compared with parents who do not homeschool, and this status is achieved with only one wage earner, typically the father (Murphy, 2012). Moreover, homeschool families also tend to have cohesive marriages and are overwhelmingly Protestant Christians, often fundamentalist (Murphy, 2012).

In terms of academic achievement differences between students who are and are not homeschooled, the data are extremely limited, but research suggests that homeschooled students compare well when examining national norms for test scores or when looking at college entrance scores, such as the ACT or SAT (Murphy, 2012). For example, Ray (2010) found that homeschool students performed above the 80th percentile in reading, language, mathematics, science, and social studies. Belfield (2004) found that homeschooled students outperformed public school students on the SAT; for instance, homeschoolers scored 527 on the mathematics portion and 567 on the verbal portion of the SAT compared with public school students' scores of 510 in math and 503 in verbal. However, Belfield was careful to point out the serious issues of self-selection of SAT test-takers and the lack of controls for relevant measures when looking at these differences.

We need to know a great deal more about homeschooling on students' social, behavioral, and academic outcomes and its variation in the United States and other countries. It would be helpful to gathering longitudinal data about households, families, and children across different homeschooling contexts to inform this significant gap in the empirical base along with comparisons to other parental choice options, including traditional public schools. Researchers would make contributions by also examining 'dosage.' By that I mean that it is not clear how long students typically stay in homeschooling situations and whether that matters for important youth outcomes. In future research, gathering information about how long students are homeschooled and why parents make those decisions would provide important new insights into this understudied area.

Future Research Possibilities

As mentioned throughout, additional research is critical for understanding the various effects of school choice reforms, particularly as the landscape of choice reforms continues to expand in a dramatic fashion for students around the world. Such research can make significant contributions to educational policy related to private or Catholic schools, charter/independent schools, vouchers, education tax credits, and homeschooling. An important component of this future research should address questions that go beyond horse races between sectors (e.g., charter and non-charter students, voucher and non-voucher students, students benefiting from tax credits vis-à-vis those who do not, and homeschooler outcomes vis-à-vis public school student outcomes). Rather, understanding the conditions under which these choice options are effective or not will help push policy debates forward. Further cross-national research is also important to understand the school choice policy conditions related to student outcomes in different contexts. Such research will hopefully cumulate to further systematic knowledge-building and revision of policies and theories within research and policy communities, so that there is systematic growth in our understanding of the policies and research (see Cohen, 2003). In the future, research is needed not only to inform the policy debates about school choice, but to examine the effects on students who have historically not had access to the resources and educational opportunities that lead to future educational and adult attainments.

Bibliography

Abdulkadiroglu, A., Angrist, J., Cohodes, S.R., Dynarski, S., Fullerton, J., Kane, T.J., Pathak, P., 2009. Informing the Debate: Comparing Boston's Charter, Pilot, and Traditional Schools. The Boston Foundation, Boston, MA.

Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic Schools. Journal of Political Economy 113 (1), 151–184.

Angrist, J.D., Cohodes, S.R., Dynarski, S.M., Fullerton, J.B., Kane, T.J., Pathak, P.A., Walters, C.R., 2011. Student Achievement in Charter Schools. Center for Educational Policy Research, Harvard University, Boston, MA.

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., Zhang, J., 2013. The Condition of Education 2013 (NCES 2013-037). U.S. Department of Education, National Center for Education Statistics, Washington, DC. http://nces.ed.gov/pubsearch (retrieved from 30.12.13.).

Barrera-Osorio, F., Patrinos, H.A., 2009. An international perspective on school vouchers. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 339–358.

Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B., 2003. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City. Journal of the American Statistical Association 98 (42), 299–323.

Belfield, C., 2004. Homeschooling in the US. (Occasional Paper No. 88). Columbia University, Teachers College, National Center for the Study of Privatization in Education, New York.

Belfield, C., 2005. Homeschoolers: how well do they perform on the SAT for college admissions? In: Cooper, B.S. (Ed.), Homeschooling in Full View. Information Age, Greenwich, CT, pp. 167–177.

Berends, M., 2014. The evolving landscape of school choice in the United States. In: Richard Milner IV, H., Lomotey, K. (Eds.), Handbook of Urban Education. Routledge, New York, pp. 451–473.

Berends, M., Zottola, G.C., 2009. Social perspectives on school choice. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 35–53.

Berends, M., Cannata, M., Goldring, E.B. (Eds.), 2011. School Choice and School Improvement. Harvard Education Press, Cambridge, MA.

Berends, M., Goldring, E., Stein, M., Cravens, X., 2010. Instructional conditions in charter schools and students' mathematics achievement gains. American Journal of Education 116 (3), 303–335.

Berends, M., Springer, M.G., Walberg, H.J. (Eds.), 2008. Charter School Outcomes. Lawrence Erlbaum Associates/Taylor & Francis Group, Mahweh, NJ.

Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), 2009. Handbook of Research on School Choice. Routledge, New York.

Betts, J.R., Loveless, T. (Eds.), 2005. Getting Choice Right: Ensuring Equity and Efficiency in Education Policy. Brookings Institution Press, Washington, DC.

Betts, J.R., Tang, E., 2011. The Effect of Charter Schools on Student Achievement: A Meta-analysis of the Literature. National Charter School Research Project, Center on Reinventing Public Education, University of Washington, Seattle, WA.

Bidwell, C.E., Dreeben, R., 2006. Public and private education: conceptualizing the distinction. In: Hallinan, M.T. (Ed.), School Sector and Student Outcomes. University of Notre Dame Press, Notre Dame, IN, pp. 9–37.

Bidwell, C.E., Kasarda, J.D., 1980. Conceptualizing and measuring the effects of school and schooling. American Journal of Education 88, 401–430.

Brewer, D.J., Hentschke, G.C., 2009. An international perspective on publicly-financed, privately-operated schools. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 227–246.

Brewer, D.J., Augustine, C., Zellman, G., Ryan, G., Goldman, C.A., Staz, C., Constant, L., 2007. Education for a New Era: Design and Implementation of K-12 Education Reform in Qatar. RAND, Santa Monica, CA.

Bryk, A.S., Driscoll, M.E., 1988. The High School as Community: Contextual Influences and Consequences for Students and Teachers. National Center on Effective Secondary Schools, University of Wisconsin, Madison WI.

Bryk, A.S., Schneider, B., 2002. Trust in Schools: A Core Resource for Improvement. Russell Sage Foundation, New York.

Bryk, A.S., Lee, V., Holland, P., 1993. Catholic Schools and the Common Good. Harvard University Press, Cambridge, MA.

Carbonaro, W., 2006. Public-private difference in achievement among kindergarten students: differences in learning opportunities and student outcomes. American Journal of Education 113 (1), 31–65.a.

Carbonaro, W., Covay, E., 2010. School sector and student achievement in the era of standards based reforms. Sociology of Education 83 (2), 160–182.

Center for Research on Education Outcomes (CREDO), 2013. National Charter School Study. Author, Stanford, CA.

Chingos, M.M., Peterson, P.E., 2012. The Effects of School Vouchers on College Enrollment: Experimental Evidence from New York City. Harvard Kennedy School Program on Education Policy and Governance, Cambridge, MA.

Chubb, J.E., Moe, T.E., 1990. Politics, Markets, and America's Schools. The Brookings Institution Press, Washington, DC.

Cohen, D., 2003. Workshop on Understanding and Promoting Knowledge Accumulation in Education: Tools and Strategies for Education Research. National Academy of Sciences, Washington, DC.

Coleman, J.S., Hoffer, T., Kilgore, S., 1982. High School Achievement: Public, Catholic, and Private Schools Compared. Basic Books, New York.

Coleman, J.S., Hoffer, T., 1987. Public and Private High Schools: The Impact of Communities. Basic Books, New York.

Cowen, J.M., 2008. School choice as a latent variable: estimating the 'complier average causal effect' of vouchers in Charlotte. Policy Studies Journal 36 (2), 301–315.

Cowen, J.M., Fleming, D.J., Witte, J.F., Wolf, P.J., Kisida, B., 2013. School vouchers and student attainment: evidence from a state-mandated study of Milwaukee's Parental Choice Program. Policy Studies Journal 41 (1), 147–167.

DiMaggio, P.J., Powell, W.W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48, 147–160.

Dobbie, W., Fryer, R.G., 2011. Are high quality schools enough to close the achievement gap? Evidence from a social experiment in Harlem. American Economic Journal: Applied Economics 3, 158–187.

Elmore, R., 2007. School Reform from the inside Out: Policy, Practice, and Performance. Harvard Education Press, Boston, MA.

Evans, W.N., Schwab, R.M., November 1995. Finishing high school and starting college: do catholic schools make a difference? Quarterly Journal of Economics 110, 941–974.

Figlio, D.N., 2009. Voucher outcomes. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, NY, pp. 321–337.

Figlio, D.N., Hart, C.M.D., 2011. Competitive Effects of Means-tested School Vouchers. NBER Working Paper No. 16056. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w16056.

Figlio, D.N., Goldhaber, D., Hannaway, J., Rouse, C., 2013. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. American Economic Journal: Economic Policy 5 (2), 251–281.

Finnegan, K.S., 2007. Charter school autonomy: the mismatch between theory and practice. Educational Policy 21 (3), 503–526.

Friedman, M., 1962. Capitalism and Freedom. Chicago University Press, Chicago.

Gill, B.P., Timpane, P.M., Ross, K.E., Brewer, D.J., 2007. Rhetoric versus Reality: What We Know and what We Need to Know about Vouchers and Charter Schools. RAND, Santa Monica, CA.

Gleason, P., Clark, M., Tuttle, C.C., Dwoyer, E., 2010. The Evaluation of Charter School Impacts: Final Report. NCEE 2010-4029. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC. http://ies.ed.gov/ncee/pubs/20104029/pdf/20104029.pdf.

Goldberger, A.S., Cain, G.C., 1982. The causal analysis of cognitive outcomes in the coleman, hoffer, and Kilgore report. Sociology of Education 55 (2/3), 103–122.

Greene, J.P., 2001. Vouchers in Charlotte. Education Matters 1 (2), 55–60.

Greene, J.P., Howell, W.G., Peterson, P.E., 1999a. Lessons from the Cleveland scholarship program. In: Peterson, P.E., Hassel, B.C. (Eds.), Learning from School Choice. Brookings Institution Press, Washington, DC.

Greene, J.P., Peterson, P.E., Du, J., 1999b. Effectiveness of school choice: the Milwaukee experiment. Education and Urban Society 31 (1), 190–213.

Grogger, J., Neal, D., 2000. Further evidence on the effects of catholic secondary schooling. Brookings-Wharton Papers on Urban Affairs 151–193.

Hallinan, M.T., Kubitschek, W.N., 2010. School sector, school poverty, and the catholic school advantage. Catholic Education: A Journal of Inquiry and Practice 14 (2), 143–172.

Hallinan, M.T., Kubitschek, W.N., 2012. A comparison of academic achievement adherence to the common school idea in public and catholic schools. Sociology of Education 85 (1), 1–22.

Henig, J., 1999. School choice outcomes. In: Sugarman, S., Kemerer, F. (Eds.), School Choice and Social Controversy. Brookings Institute Press, Washington, DC.

Heyneman, S.P., 2009. International perspectives on school choice. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, NY, pp. 79–95.

Hoffer, T.B., 2009. Perspectives on private schools. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 429–446.

Hoffer, T.B., Greeley, A.M., Coleman, J.S., 1985. Achievement growth in public and catholic schools. Sociology of Education 58 (2), 74–97.

Howell, W.G., Wolf, P.J., Campbell, D.E., Peterson, P.E., 2002. School vouchers and academic performance: results from three randomized field trials. Journal of Policy Analysis and Management 21 (2), 191–217.

Hoxby, C.M., Murarka, S., 2008. Methods of assessing achievement of students in charter schools. In: Berends, M., Springer, M.G., Walberg, & H.J. (Eds.), Charter School Outcomes. Taylor & Francis Group, Mahweh, NJ, pp. 7–37.

Hoxby, C.M., Murarka, S., Kang, J., 2009. How New York City's Charter Schools Affect Achievement. New York City Charter Schools Evaluation Project, Cambridge, MA. http://www.nber.org/~schools/charterschoolseval/how_NYC_charter_schools_affect_achievement_sept2009.pdf.

Huerta, L.A., d'Entremont, C., 2007. Education tax credits in a post-Zelman era: legal, political, and policy alternatives to vouchers? Educational Policy 21, 73–109.

Krueger, A.B., Zhu, P., 2003. Comment on "principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City." Journal of the American Statistical Association 98 (42), 314–318.

Ladd, H.F., Fiske, E.B., Ruijs, N., 2011. Does parental choice foster segregated schools? Insights from the Netherlands. In: Berends, M., Cannata, M., Goldring, E.B. (Eds.), School Choice and School Improvement. Harvard Education Press, Cambridge, MA, pp. 233–253.

Lee, Valerie E., Bryk, Anthony S., 1993. Science or policy argument? A review of the quantitative evidence in chubb and Moe's politics, markets, and America's schools. In: Rasell, E., Rothstein, R. (Eds.), School Choice: Examining the Evidence. Economic Policy Institute, Washington, DC, pp. 185–208.

Lubienski, C., 2003. Innovation in education markets: theory and evidence on the impact of competition and choice in charter schools. American Educational Research Journal 402, 395–443.

Marks, H.M., 2002. School composition and peer effects in distinctive organizational settings. International Journal of Educational Research 37 (5), 505–519.

Marks, H.M., 2009. Perspectives on catholic schools. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 479–499.

McDonald, D., Schultz, M.M., 2011. United States Catholic Elementary and Secondary Schools, 2010-2011. National Catholic Educational Association, Washington, DC.

Metcalf, K.K., West, S.D., Legan, N.A., Paul, K.M., Boone, W.J., 2003. Evaluation of the Cleveland Scholarship and Tutoring Program, Summary Report 1998–2001. Indiana Center for Evaluation, Bloomington, IN.

Meyer, J.W., 1977. The effects of education as an institution. American Journal of Sociology 83 (1), 55–77.

Meyer, J.W., Ramirez, F., 2000. The world institutionalization of education. In: Schriewer, J. (Ed.), Discourse Formation in Comparative Education. Peter Lang, Frankfurt, pp. 111–132.

Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Meyer, J.W., Rowan, B., 1978. The structure of educational organizations. In: Meyer, M.W., Associates (Eds.), Environments and Organizations. Jossey-Bass, San Francisco, pp. 78–109.

Morgan, S.L., 2001. Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. Sociology of Education 74, 341–374.

Murphy, Joseph, 2012. Homeschooling in America: Capturing and Assessing the Movement. Corwin Press, Thousand Oaks, CA.

National Alliance for Public Charter Schools. 2014. http://dashboard.publiccharters.org/dashboard/schools/page/overview/year/2012/.

Nguyen, A.N., Taylor, J., Bradley, S., 2006. The estimated effect of catholic schooling on educational outcomes using propensity score matching. Bulletin of Economic Research 58 (4), 285–307.

Powell, W.W., DiMaggio, P.J., 1991. The New Institutionalism in Organizational Analysis. The University of Chicago Press, Chicago, IL.

Preston, C., Goldring, E., Berends, M., Cannata, M., 2012. School innovation in district context: comparing traditional public schools and charter schools. Economics of Education Review 31, 318–330.

Ray, B., 2010. Academic achievement and demographic traits of homeschool students: a nationwide study. Academic Leadership Live 8 (1). Retrived from: http://www.academicleadership.org/.

Reardon, Sean, Cheadle, Jacob, Robinson, Joseph, 2009. The effect of catholic schooling on math and Reading development in kindergarten through Fifth Grade. Journal of Research on Educational Effectiveness 2 (1), 45–87.

Rouse, C.E., 1998. Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program. The Quarterly Journal of Economics 113 (2), 553–602.

Schaeffer, A.B., 2009. Education tax credits. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 593–613.

Schneider, B., 2003. Sociology of education: an overview of the field at the turn of the twenty-first century. In: Hallinan, Maureen T., Gamoran, Adam, Kubitsheck, Warren, Loveless, Tom (Eds.), Stability and Change in American Education: Structure, Process, and Outcomes. Eliot Werner Publications, Clinton Corners, NY, pp. 193–226.

Schumpeter, J.A., 1942. Capitalism, Socialism, and Democracy. Harper, New York.

Scott, W.R., Davis, G.F., 2007. Organizations & Organizing: Rational, Natural and Open System Perspectives. Prentice Hall, Englewood Cliffs.

Scott, W.R., Meyer, J.W., 1994. Institutional Environments and Organizations: Structural Complexity and Individualism. Sage, Thousand Oaks, CA.

Stecher, B.M., Vernez, G., 2010. Reauthorizing No Child Left behind: Facts and Recommendations. RAND, Santa Monica, CA.

Teasley, B., 2009. Charter school outcomes. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 209–225.

Uecker, J.E., 2008. Alternative schooling strategies and the religious lives of american adolescents. Journal of the Scientific Study of Religion 47 (4), 563–584.

U.S. Department of Education, 2012. Digest of Education Statistics 2012. National Center for Education Statistics, Washington, DC. http://nces.ed.gov/programs/digest/2012menu_tables.asp.

Walberg, H.J., 2011. Tests, Testing, and Genuine School Reform. Hoover Institution Press, Stanford, CA.

Walberg, H.J., Bast, J.L., 2003. Education and Capitalism: How Overcoming Our Fear of Markets and Economics Can Improve America's Schools. Hoover Institution Press, Stanford, CA.

Weick, K.E., 1976. Educational organizations as loosely coupled systems. Administrative Science Quarterly 21 (1), 1–19.

Witte, J.F., 2000. The Market Approach to Education: An Analysis of America's First Voucher Program. Princeton University Press, Princeton, NJ.

Witte, J.F., Cowen, J.M., Wolf, P.J., Fleming, D.J., 2012. Milwaukee Parental Choice Program Longitudinal Educational Growth Study: Fifth Year Report. University of Arkansas, Fayetteville, AR.

Woessman, L., 2009. An international perspective on private schools. In: Berends, M., Springer, M.G., Ballou, D., Walberg, H.J. (Eds.), Handbook of Research on School Choice. Routledge, New York, pp. 501–518.

Woessman, L., Ludemann, E., Schutz, G., West, M.R., 2009. School Accountability, Autonomy, and Choice Around the World. Edward Elgar, Cheltenham.

Wolf, P.J., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., Eissa, N., Carr, M., 2010. Evaluation of the DC Opportunity Scholarship Program: Final Report (NCEE 2010–4018). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Wolf, P.J., Kisida, B., Gutmann, G., Puma, M., Rizzo, L., Eissa, N., 2011. School vouchers in the nation's capital. In: Berends, M., Cannata, M., Goldring, E.B. (Eds.), School Choice and School Improvement. Harvard Education Press, Boston, MA, pp. 17–33.

Zimmer, R., Bettinger, E.P., 2008. Beyond the rhetoric: surveying the evidence on vouchers and tax credits. In: Ladd, H.F., Fiske, E.B. (Eds.), Handbook of Research in Education Finance and Policy. Routledge, New York, pp. 447–466.

Zimmer, R., Buddin, R., 2007. Getting inside the black box: examining how the operations of charter schools affect performance. Peabody Journal 82 (2–3), 231–273.

Zimmer, R., Buddin, R., 2008. Charter schools in California. In: Berends, M., Springer, M.G., Walberg, H.J. (Eds.), Charter School Outcomes. Taylor & Francis, New York, pp. 163–194.

Zimmer, R., Buddin, R., Chau, D., Gill, B., Guarino, C., Hamilton, L., Krop, C., McCaffrey, D., Sandler, M., Brewer, D., 2003. Charter School Operations and Performance: Evidence from California. RAND, Santa Monica, CA.

School Effectiveness Research

Jaap Scheerens, Faculty of Behavioral Sciences, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

School effectiveness research is described as the scientific approach to determine the causal influence of malleable conditions of schooling. The article describes how different strands of school effectiveness have developed and are now increasingly combined into more integrative approaches. The knowledge base of school effectiveness research is discussed on the basis of the results from meta-analyses and international comparative assessment studies. The size and educational relevance of the effects of malleable school conditions is still contested. Theory-based interpretation and theory-driven research are a relatively neglected area in this field. Promising developments are found in conceptual models and application of more established organizational and economic theories. Methodologically the field is likely to benefit from more rigorous research design, such as quasi-experimental evaluations of school improvement projects and randomized field trials.

School Effectiveness and School Effectiveness Research

In the most general sense 'school effectiveness' refers to the level of goal attainment of a school. Although average achievement scores in core subjects, established at the end of a fixed program, are the most probable 'school effects,' alternative criteria like the responsiveness of the school to the community and the satisfaction of the teachers may also be considered.

Assessment of school effects occurs in various types of applied contexts like the evaluation of school improvement programs or comparing schools for accountability purposes, by governments, municipalities, or individual schools.

School effectiveness *research* attempts to deal with the causal aspects inherent in the effectiveness concept by means of scientific methods. Not only is assessment of school effects considered, but particularly the attribution of differences in school effects to malleable conditions. Usually, school effects are assessed in a comparative way, e.g., by comparing average achievement scores between schools. In order to determine the 'net' effect of malleable conditions, like the use of different teaching methods or a particular form of school management, achievement measures have to be adjusted for intake differences between schools. For this purpose student background characteristics like socioeconomic status, general scholastic aptitudes, or initial achievement in a subject are used as control variables. This type of statistical adjustment in research studies has an applied parallel in the strive for 'fair comparisons' between schools, known under the label of '*value-added.*'

School effectiveness research is often compared with the related field of school improvement. Seen from a research perspective the differences are gradual, with the additional focus on strategies for educational change in school improvement studies being the major difference (Scheerens, 2013).

Strands of Educational Effectiveness Research

School effectiveness research has developed as a gradual integration of several research traditions. The roots of current 'state-of-the-art' school effectiveness research are sketched by briefly referring to each of these research traditions. Integration implies that school effectiveness research, specifically focused at the effects of school level factors, is now often embedded in multilevel *educational* effectiveness models. This issue is further addressed in the Section Strands of Educational Effectiveness Research, but it implies that certain looseness in terminology will be used as far as the terms school effectiveness and educational effectiveness are concerned. When speaking of school effectiveness studies that have addressed, in addition to school factors, teaching conditions, either at classroom level are aggregated up to the school level, will be included under this label.

The elementary design of school effectiveness research is the association of hypothetical effectiveness enhancing conditions of schooling and output measures, mostly student achievement. A basic model is from systems theory, where the school is seen as a black box, within which processes or 'throughput' take place to transform inputs into outputs. The inclusion of an environmental or context dimension completes this model (see **Figure 1**). The major task of school effectiveness research is to reveal the impact of relevant input characteristics on output and to 'break open' the black box in order to show which process or throughput factors 'work,' next to the impact of contextual conditions. Within the school it is helpful to distinguish a school and a classroom level and, accordingly, school organizational and instructional processes.

Research tradition in educational effectiveness varies according to the emphasis that is put on the various antecedent conditions of educational outputs. These traditions also have a disciplinary basis. The common denominator of the six areas of effectiveness research that will be distinguished is that in each case the elementary design of associating outputs or

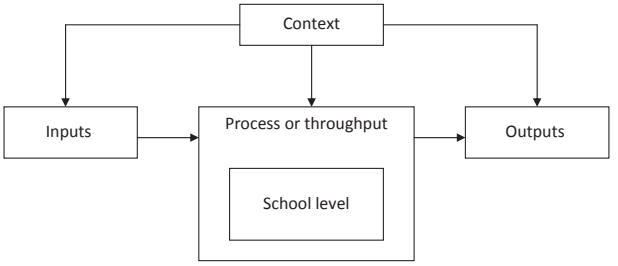

Figure 1 A basic systems model of school functioning.

Table 1 General characteristics of types of school effectiveness research

	Independent variable type	*Dependent variable type*	*Discipline*	*Main study type*
1. (Un)equal opportunities	Socioeconomic status and IQ of pupil, material school characteristics	Attainment	Sociology	Survey
2. Production functions	Material school characteristics	Achievement level	Economics	Survey
3. Evaluation compensatory programs	Specific curricula	Achievement level	Interdisciplinary pedagogy	Quasi-experiment
4. Effective schools	'Process' characteristics of schools	Achievement level	Interdisciplinary pedagogy	Case study
5. Effective instruction	Characteristics of teachers, instruction, class organization	Achievement level	Educational psychology	Experiment observation
6. System level effectiveness	System level policies and institutional arrangements	Achievement and attainment	Economics	Background studies based on international assessment programs

outcomes of schooling with antecedent conditions (inputs, processes, or contextual) applies.

The following research areas or research traditions can be distinguished:

1. Research on equality of opportunities in education and the significance of the school in this.
2. Economic studies on education production functions.
3. The evaluation of compensatory programs and school improvement programs.
4. Studies of unusually effective schools.
5. Studies on the effectiveness of teachers, classes, and instructional procedures.
6. Studies on the effectiveness of system level policies and institutional arrangements.

For a further discussion of the first five of these research traditions the reader is referred to Scheerens et al. (2003: Chapter 11). A schematic characterization of research orientation and disciplinary background is given in **Table 1**. The sixth research area is an emerging field, which is very much stimulated by international assessment programs, such as TIMSS and PISA. System level variables that have been addressed in these kinds of studies are decentralization, choice, and accountability arrangements in national educational systems (cf Woessmann et al., 2009). Scheerens et al. (2012), in a study that used PISA 2009 data, investigated more complex models in which indirect effects of system level factors through intermediary school conditions on student performance were computed.

A very interesting methodological development is the new interest in the use of randomized field trials in school effectiveness research (cf Bosker, 2011). Many relevant examples are presented on the website attached to the biannual SREE (Society for Research on Educational Effectiveness) conferences (https://www.sree.org/).

Integrated School Effectiveness Research

In recent school effectiveness studies these various approaches to educational effectiveness have become integrated. Integration was manifested in the conceptual modeling and the choice of variables. At the technical level multilevel analysis has contributed significantly to this development. In contributions to the conceptual modeling of school effectiveness, schools became depicted as a set of 'nested layers' (Purkey and Smith, 1983), where the central assumption was that higher organizational levels facilitated effectiveness enhancing conditions at lower levels (Scheerens and Creemers, 1989). In this way a synthesis between production functions, instructional effectiveness, and school effectiveness became possible, by including the key variables from each tradition, each at the appropriate 'layer' or level of school functioning [the school environment, the level of school organization and management, the classroom level, and the level of the individual student]. Conceptual models that were developed according to this integrative perspective are those by Scheerens (1990), Creemers (1994), Stringfield and Slavin (1992), and Creemers and Kyriakides (2008). By way of illustration the Scheerens model is shown in **Figure 2**.

Exemplary cases of integrative, multilevel school effectiveness studies are those by Mortimore et al. (1988), Brandsma (1993), Hill et al. (1995), Sammons et al. (1995), and Grisay (1996). The study by Borman et al. (2003) provides a review and meta-analysis of evaluations of Comprehensive School Reform (CSR) projects. CSR projects form a very interesting blending of school effectiveness and school improvement, as well as an application of integrating effectiveness enhancing conditions at school context, school, and classroom level.

The Knowledge Base on School Effectiveness

Review and meta-analyses on school effectiveness, e.g., Wang et al. (1993) and Scheerens and Bosker (1997), indicate that resource-input factors on average have a negligible effect, school factors have a small effect, while instructional variables have an average to large effect. Another feature, noted by Scheerens and Bosker (1997), is a remarkable difference between the relatively small effect size for the school level variables reported in meta-analyses and the degree of certainty

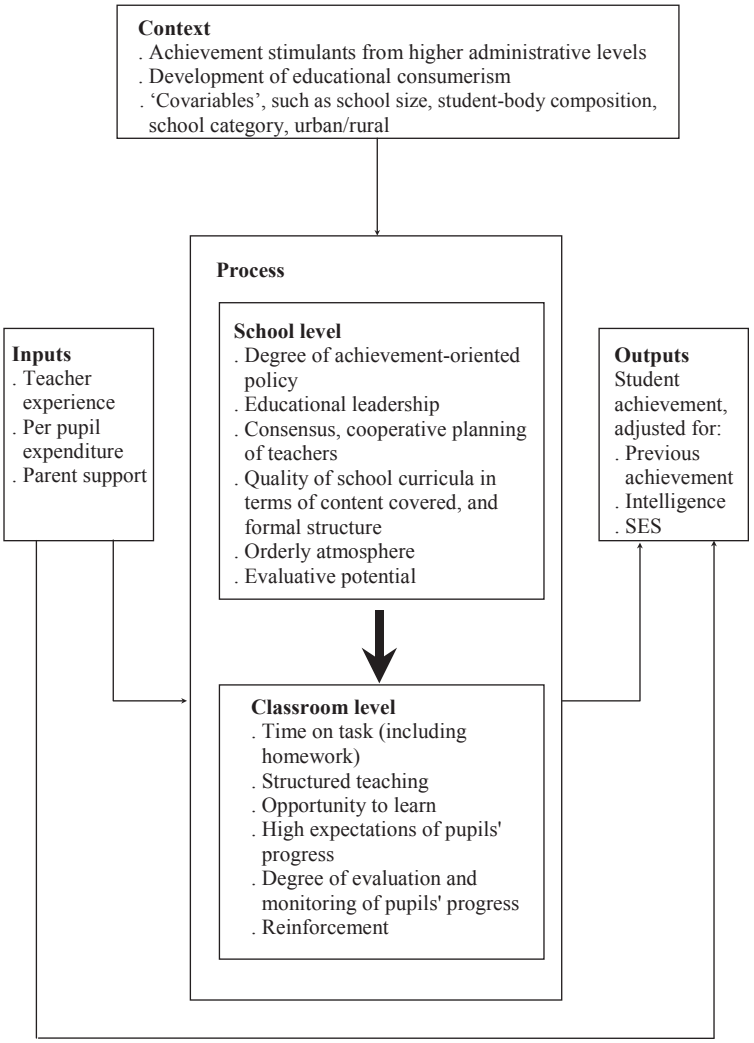

Figure 2 An integrated model of school effectiveness. From Scheerens, J., 1990. School effectiveness and the development of process indicators of school functioning. School Effectiveness and School Improvement, 61–80.

and consensus on the relevance of these factors in the more qualitative research reviews.

The same kind of relative ordering of effect sizes is evident from more recent meta-analyses, as those by Hattie (2009) and Scheerens et al. (2007). Still as it comes to effect sizes, important differences are reported in different meta-analyses. Hattie (2009) provides massive quantitative evidence on the association of numerous school, teacher, and teaching variables with student achievement. Average effect sizes for school, curriculum, teacher, and teaching factors in terms of the *d* coefficient (standardized difference between means) reported by Hattie are 0.23, 0.45, 0.49, and 0.43, respectively (Hattie, 2009: pp. 74, 109, 130, 162, 201). According to Cohen (1988), effect sizes of 0.2 are considered small, 0.5 medium, and 0.8 large. According to these standards the average effect sizes should be considered as slightly below medium. Still, meta-analyses that are carried out by European authors show effect sizes that are even lower; see for example Witziers et al. (2003), Scheerens et al. (2007), Seidel and Shavelson (2007), and Creemers and Kyriakides (2008). By way of illustration some of the results on key variables like educational leadership, evaluation and monitoring, learning time, structured teaching, and quantity of teaching are compared. For educational leadership Hattie (2009: p. 82) reports a coefficient *d* of 0.36, which expressed as a correlation coefficient is 0.18; Scheerens et al. (2007) and Creemers and Kyriakides (2008: p. 202) report effects (correlations) of 0.05 and 0.07, respectively. For feedback and monitoring at classroom level from Hattie's results, an average effect of 0.33 (correlation) can be computed. (This is the average of effect sizes found for 'feedback' (p. 173), 'frequent testing' (p. 178), and 'formative evaluation' (p. 181), expressed as a correlation.) Scheerens et al. (2007) and Seidel and Shavelson found effects of 0.07 and 0.01, respectively. Using structured teaching as an overarching label, Hattie reports an effect size comparable to a correlation of 0.30 for direct teaching. Creemers and Kyriakides found an effect size of 0.17 for 'quality of teaching,' while Scheerens et al. (2007) and Seidel and Shavelson (2007) report effects of 0.09 and 0.02, respectively. Variables associated with quantity of teaching, time on task, and opportunity to learn show effect sizes of 0.17 (Hattie, 2009: p. 184; Creemers and

Kyriakides, 2008: p. 201), but 0.08 and 0.03 in the studies by Scheerens et al. (2007) and Seidel and Shavelson (2007). A more extensive table with comparisons of results from meta-analyses is provided in Scheerens et al. (2011: p. 144).

According to established scientific standards the effect sizes for the key school and teaching variables are medium when one considers the results by Hattie and small when one considers the other meta-analyses. One of the explanations Hattie (2009: p. 202) offers for the differences in effect sizes between his results and those by Seidel and Shavelson is that these latter authors have used only studies that controlled for student prerequisites. This could be seen as a sign that the more Europe-based studies used stricter quality controls in selecting studies, and might therefore have more credible results. The other explanation might be that effect sizes in the USA, Great Britain, and Australia are higher, perhaps due to greater variability in processes and outcomes.

Results from international studies indicate that the generalizability of educational effectiveness outcomes is rather poor. Results based on PISA, for example, show that the school variable that did best in PISA 2000, 'orderly climate,' was significantly associated with reading literacy achievement in just 11 out of 37 countries (Luyten et al., 2005).

Interesting debates can be started about further interpretation of the effect sizes. It could be argued that even 'small' effects are educationally relevant when compared, for example, to the magnitude of achievement gain in 1 year of schooling (cf Baumert et al., 2006). This is beyond the scope of this article. Still the excursion into the world of quantitative meta-analyses gives reason to be prudent in the interpretation of the strength of the educational effectiveness knowledge base as far as positive process/outcome associations are concerned.

Conceptual Models and Theory in School Effectiveness Research

Usually the field of educational effectiveness research is considered as rather a-theoretical. This impression is enforced in a 'research mapping' study by Nordenbo et al. (2009). According to this study, out of 111 empirical school effectiveness studies only 23 made explicit reference to theory or a conceptual model.

According to Snow (1973) theory development is to be seen as a continuum, ranging from relatively simple generalizations to axiomatic theory. Conceptual models have a place on this continuum. The 'Dynamic Model' by Creemers and Kyriakides (2008) illustrates various applications of theory. Its micro level core is based on the Caroll model, which in turn has a connection with learning theory. Next the model is an example of conceptual mapping, using a multilevel framework and various metatheoretical concepts, such as consistency across levels and control. A major challenge of this and other multilevel conceptual models is to understand the way factors at different levels interact and mediate achievement outcomes. Finally, the Dynamic Model represents an elaborate measurement model of effectiveness enhancing processes. The pragmatic road ahead for educational effectiveness would seem to be the continued connection of conceptual models, such as the Dynamic Model, and advanced statistical analysis techniques, such as multilevel structural equation modeling. The knowledge about 'indirect effects' provides more insight into the mechanisms by which effectiveness enhancing conditions work. Interesting examples are provided in indirect effect models of school leadership, where instructional conditions gradually become better integrated as important intermediary factors; cf Heck and Moriyama (2010), Louis et al. (2010), and Scheerens (2013).

Two other conceptual models that have been used as a basis for school effectiveness research are the *Quinn and Rohrbaugh Competing Values Framework* as applied by Cameron and Whetten (1983) and Quinn and Rohrbaugh (1983), and the concept of *High Reliability Organizations*, as applied to schools by Stringfield et al. (2008).

The Competing Values Framework is based on alternative models of organizational effectiveness. Each model is based on long-standing schools of thought in organizational and management science: the idea of economic rationality and rational planning, the idea of organizations as open systems, the human relations orientation, and the idea of formalized structure, the bureaucracy. Quinn and Rohrbaugh derive four distinct models from these basic orientations, the Rational Goal (RG) Model, the Open Systems (OS) Model, the Human Relations (HR) Model, and the Internal Process Model (IP). Each is oriented toward a specific effectiveness criterion: RG toward primary production, OS toward adaptability and responsiveness toward the environment, HR toward staff job satisfaction, and IP toward formal structures and procedures. Griffith (2003) uses this framework to map malleable variables that have received empirical support in school effectiveness research according to each of these four models. For example, optimizing learning time is seen as a measure that fits RG, stimulating parental involvement as belonging to OS, participative leadership is subsumed under the HR model, and creating an orderly atmosphere is seen as a measure fitting the IP approach. By means of path analysis Griffith models the effects of each of the four organizational models on student achievement. The Quinn and Rohrbaugh framework has also been used, with varying success, to model direct and indirect effects of school leadership (e.g., Ten Bruggencate, 2009; for an overview of studies see Scheerens, 2012). By subsuming specific effectiveness enhancing variables under four broader concepts, this approach succeeds in providing a more parsimonious conceptualization of educational effectiveness. The four orientations to organizational effectiveness can easily be interpreted as different strategies to school improvement.

The second theoretical approach that was represented in the set of studies covered by Nordenbo et al. (2009) and which can be associated with the Quinn and Rohrbaugh framework is the work on schools as High Reliability Organizations (Stringfield et al., 2008). High Reliability Organizations operate in a context where failure to attain the key goals would be disastrous. Specific characteristics are clear and finite goals, alertness to surprises and lapses, the imperative to identify flaws, use of powerful databases, standard operating procedures, rigorous evaluation, and hierarchical structure, through collective decision making. High Reliability Organizations are very much in line with the IP Model, of the Competing Values Framework, and are aligned to the RG Model as well. Bureaucratic structuring and formalization of procedures are at the heart of this approach to school improvement.

Despite these interesting examples, a more fundamental theory-oriented basis for school effectiveness research is still an unaccomplished ideal. Crucial ingredients for bringing this about would be a comprehensive reflection on the knowledge base, stricter formalization of key associations in multilevel conceptual models, and interpretation of the key findings against the background of more established theory and constructs, e.g., loose coupling versus alignment. An interesting complementary perspective in theory-oriented work could be guided by conceptualization and empirical data on school failure and mechanisms that would create 'ineffectiveness': e.g., Stringfield (1998) and Scheerens (2014).

The Future of School Effectiveness Research

From this article it could be concluded that school effectiveness research can be defined in a broad and a narrow sense. In the broadest sense one could refer to all types of studies which relate school and classroom conditions to achievement outcomes, preferably after taking into account the effects of relevant student background variables. In a narrower sense state-of-the-art integrative school effectiveness studies and fundamental studies could be seen as the core.

Following the broader definition the future of school effectiveness studies is guaranteed, particularly in the sense of 'applied' studies, like cohort studies, large-scale effect studies carried out for accountability purposes, monitoring studies, and (international) assessment studies.

State-of-the-art, fundamental, and foundational school effectiveness studies are a much more vulnerable species, one of the major difficulties being the organizational complexity and costs of the 'state-of-the-art' types of study. Given the shortage of these kinds of studies the more fundamental and foundational types of studies are likely to be dependent on the quality of data sets that have been acquired for 'applied purposes.'

Therefore the best guarantee for continued fundamental school effectiveness research lies in the enhanced research technical quality of applied studies. Examples are quasi-experimental evaluation of carefully designed school improvement projects and randomized field trials on specific effectiveness enhancing factors, such as longer school days, alignment between content taught and content tested, result-oriented leadership, and focused teaching.

The coherence of the field is served by regular comprehensive reviews, such as the three state-of-the-art reviews, published in the journal *School Effectiveness and School Improvement*: Reynolds et al. (2012), Muijs et al. (2012), and Hopkins et al. (2012).

See also: Educational Effectiveness Theory and Research: Recent Advances; Educational Effectiveness, The Field of; Equity and Education; School Improvement; School Management.

Bibliography

Baumert, J., Luedtke, O., Trautwein, U., 2006. Interpreting Effect Sizes in Large-Scale Educational Assessments. Internal paper. Max Planck Institute for Human Development, Berlin.

Borman, G.D., Hewes, G.M., Overman, L.T., Brown, S., 2003. Comprehensive school reform and achievement: a meta-analysis. Review of Educational Research 73, 125–230.

Bosker, R.J., 2011. From Educational Effectiveness to Evidence Based Education. Cyprus. Key Note Address, ICSEI.

Brandsma, H.P., 1993. Basisschoolkenmerken en de kwaliteit van het onderwijs (Characteristics of Primary Schools and the Quality of Education). RION, Groningen.

Ten Bruggencate, G.C., 2009. Maken Schoolleiders Het Verschil? (Do School Leaders Make a Difference?). University of Twente, Enschede.

Cameron, K.S., Whetten, D.A. (Eds.), 1983. Organizational Effectiveness. A Comparison of Multiple Models. Academic Press, New York.

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences. Lawrence Erlbaum and Associates, Hillsdale, NJ.

Creemers, B.P.M., 1994. The Effective Classroom. Cassell, London.

Creemers, B.P.M., Kyriakides, L., 2008. The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools. Routledge, London.

Griffith, J., 2003. Schools as organizational models: implications for examining school effectiveness. The Elementary School Journal 104, 29–47.

Grisay, A., 1996. Evolution des acquis cognitifs et socio-affectifs des eleves au cours des annees de college. Université de Liège, Liège.

Hattie, J.A.C., 2009. Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge, London.

Heck, R.H., Moriyama, K., 2010. Examining relationships among elementary schools' contexts, leadership, instructional practices, and added-year outcomes: a regression discontinuity approach. School Effectiveness and School Improvement 21, 377–408.

Hill, P.W., Rowe, K.J., Jones, T., October 1995. SIIS: School Improvement Information Service. Version 1.1. The University of Melbourne, Centre for Applied Educational Research, Melbourne.

Hopkins, D., Stringfield, S., Harries, A., Stoll, L., Mackay, T., 2012. School and system improvement: a state of the art review. School Effectiveness and School Improvement 23.

Louis, K.S., Dretzke, B., Wahlstrom, K., 2010. How does leadership affect student achievement? Results from a national US survey. School Effectiveness and School Improvement 21, 315–336.

Luyten, H., Scheerens, J., Visscher, A., Maslowski, R., Witziers, B., Steen, R., 2005. School Factors Related to Quality and Equity. Results from PISA 2000. OECD, Paris.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., Ecob, R., 1988. School Matters: The Junior Years. Open Books, Somerset.

Muijs, D., Creemers, B., Kyriakides, L., Van der Werf, G., Timperley, H., Earl, L., 2012. Teaching effectiveness. A state of the art review. School Effectiveness and School Improvement 23.

Nordenbo, S.E., Holm, A., Elstad, E., Scheerens, J., Soegaard Larsen, M., Uljens, M., Fibaek Laursen, P., Hauge, T.E., 2009. Research Mapping of Input, Process and Learning in Primary and Lower Secondary Schools. Technical report. Danmarks Pædagogiske Universitetsskole, Aarhus Universitet, Copenhagen.

Purkey, S.C., Smith, M.S., 1983. Effective schools: a review. The Elementary School Journal 83, 427–452.

Quinn, R.E., Rohrbaugh, J., 1983. Spatial models of effectiveness criteria: towards a competing values approach to organizational analysis. Management Science 29, 363–377.

Reynolds, D., Sammons, P., De Fraine, B., Townsend, T., Van Damme, J., Teddlie, C., Stringfield, S., 2012. Educational effectiveness research (EER): a state of the art review. School Effectiveness and School Improvement 23.

Sammons, P., Hillman, J., Mortimore, P., 1995. Key Characteristics of Effective Schools: A Review of School Effectiveness Research. OFSTED, London.

Scheerens, J., 1990. School effectiveness and the development of process indicators of school functioning. School Effectiveness and School Improvement 61–80.

Scheerens, J., 1999. School Effectiveness in Developed and Developing Countries; A Review of the Research Evidence. World Bank Paper, Washington.

Scheerens, J., 2013. The use of theory in school effectiveness research revisited. School Effectiveness and School Improvement. 24, 1–138.

Scheerens, J., 2014. School, teaching, and system educational effectiveness: some comments on three state of the art reviews. School Effectiveness Research and School Improvement 25, 282–291.

Scheerens, J. (Ed.), 2012. School Leadership Effects Revisited. Review and Meta-analysis of Empirical Studies. Springer, Dordrecht, Heidelberg, London, New York.

Scheerens, J., Bosker, R.J., 1997. The Foundations of Educational Effectiveness. Pergamon, Oxford.

Scheerens, J., Creemers, B.P.M. (Eds.), 1989. Developments in school effectiveness research. Special Themed Issue of the International Journal of Educational Research 13.

Scheerens, J., Glas, C.W., Luyten, H., Jehangir, Kh, Steen, R., 2012. System Level Correlates of Educational Performance. Analyses Based on PISA 2009. University of Twente, Enschede.

Scheerens, J., Glas, C.W., Thomas, S.M., 2003. Educational Evaluation, Assessment and Monitoring. A Systemic Approach. Swets and Zeitlinger, Lisse.

Scheerens, J., Luyten, H., Steen, R., Luyten-de Thouars, Y., 2007. Review and Meta-analyses of School and Teaching Effectiveness. University of Twente, Department of Education, Enschede.

Scheerens, J., Luyten, H., Van Ravens, J., 2011. Perspectives on Educational Quality. Illustrative Outcomes on Primary and Secondary Education in the Netherlands. Research Briefs, Springer, p. 148.

Seidel, T., Shavelson, R.J., 2007. Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. Review of Educational Research 77, 454–499.

Stringfield, S., 1998. An anatomy of ineffectiveness. In: Stoll, L., Myers, K. (Eds.), No Quick Fixes: Perspectives on Schools in Difficulties. Falmer, London, pp. 209–221.

Stringfield, S.C., Slavin, R.E., 1992. A hierarchical longitudinal model for elementary school effects. ICO-Publication 2. In: Creemers, B.P.M., Reezigt, G.J. (Eds.), Evaluation of Effectiveness. GION, Groningen.

Stringfield, S., Reynolds, D., Schaffer, E., 2008. Improving secondary students' academic achievement through a focus on reform reliability: 4- and 9-year findings from the high 19 reliability schools project. School Effectiveness and School Improvement 19, 409–428.

Stringfield, S.C., Reynolds, D., Schaffer, E.C., 2011. Toward high reliability, high-quality. In: Noteworthy Perspectives. McREL, Denver, CO.

Snow, R.E., 1973. Theory construction for research on teaching. In: Travers, R.M.W. (Ed.), Handbook of Research on Teaching. Rand McNally, Chicago, pp. 77–112.

Wang, M.C., Haertel, G.D., Walberg, H.J., 1993. Towards a knowledge base for school learning. Review of Educational Research 63, 249–294.

Witziers, B., Bosker, R.J., Krüger, M.L., 2003. Educational leadership and student achievement: the elusive search for an association. Educational Administrative Quarterly 39, 398–425.

Woessmann, L., Luedemann, E., Schuetz, G., West, M.R., 2009. School Accountability, Autonomy and Choice Around the World. Edward Elgar, Cheltenham, UK and Northampton, MA, USA.

School Environment for LGBTQ/Sexual Minority Youth

Stephen T Russell, Amanda M Pollitt, and Jennifer M Hoenig, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There has been growing attention in the last decade to the school experiences of lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ) or sexual minority youth. Unlike youth of prior generations, contemporary LGBTQ youth have opportunities to 'come out' in their families, communities, and schools, a factor that has changed the climate at school for all students. This article provides an overview of homophobia and bullying at school. Consistent with recent research in this area, it considers both individual student experiences as well as broader questions about school climates of safety and support for LGBTQ students and their heterosexual allies. A number of strategies to promote positive school climates for LGBTQ and all students have been proposed in recent years, and are reviewed: nondiscrimination and antibullying laws and policies, curricular inclusion, and school-based clubs such as gay-straight alliances. In closing, implications are considered for future research, policies, and programs to support the health and development of LGBTQ youth in schools.

In recent years there has been growing public awareness of lesbian, gay, bisexual, transgender, queer, and questioning (LGBTQ) or sexual minority youth and their experiences at school. News reports of the deaths by suicide of gay boys following chronic harassment at school brought unprecedented attention to the issue of homophobic school bullying, as well as to the broader issue of the degree to which school climates are safe and supportive of LGBTQ students. In the same period there have been controversies in many communities, also played out in news media, about challenges to long-standing heteronormative practices of schools: whether students can take same-sex dates to the prom; whether a boy can be a homecoming queen; and whether students can choose their gender presentation in senior class portraits. This public attention is in stark contrast to the silence that many LGBTQ students in prior cohorts experienced at school: For many generations they were simply invisible.

This growing public awareness and understanding of LGBTQ students and their experiences at school has emerged in part because LGBTQ youth are 'coming out' (acknowledging and disclosing their sexual and gender identities to others) at younger ages. Broad social acceptance has shifted during this period, as evidenced by multiple factors, most notably the growing public affirmation of marriage for same-sex couples. Yet for youth there is a tension between growing social acceptance on one hand, and the persistent heteronormativity of schools on the other. Thus, even as acceptance grows for LGBTQ individuals, schools remain places where homophobia, heterosexism, and discrimination continue to exist. The norms, beliefs, and attitudes held by peers, teachers, and administrators shape a school's climate. A growing body of literature has demonstrated that positive school climate supports positive development (Birkett et al., 2009). However, numerous studies have indicated that the climate of schools in the United States is hostile and unsafe for many LGBTQ youth.

We consider the lives of contemporary LGBTQ youth: their coming out in today's context, and their experiences at school. After providing an overview of homophobia and bullying at school, we shift from a focus on individual student experiences to consider broader questions about school climates of safety and support for LGBTQ students and their heterosexual allies. A number of strategies to promote positive school climates for LGBTQ and all students have been proposed in recent years: nondiscrimination and antibullying laws and policies, curricular inclusion, and school-based clubs such as gay-straight alliances (GSA). Finally, we consider implications for future research, policies, and programs to support the health and development of LGBTQ youth in schools.

Contemporary LGBTQ Youth

There have been dramatic social changes in attitudes toward homosexuality during recent decades, propelled in part by a growing number of positive LGBTQ role models, including positive portrayals of LGBTQ individuals and families in television, film, and literature. At the same time there has been growing advocacy and legislation related to LGBTQ people and their lives, most notably related to marriage, which is now legal for same-sex couples in multiple states of the United States and countries around the world (Loftus, 2001). Parallel to increasing social acceptance, LGBTQ adolescents come out at earlier ages compared to youth in prior decades, and many disclose their sexual identities in the context of supportive environments (rather than in the context of alienation and shame, as was often the case for prior cohorts). Recent young adult cohorts report coming out between ages 16 and 17 compared to older adult cohorts who report having come out between the ages of 24 and 27 (Grov et al., 2006); notably, patterns of sexual identity and behavior, as well as development of sexual identity milestones such as coming out, show few differences across racial/ethnic groups (Rosario et al., 2004). Some studies show that youth are coming out at even younger ages, including during middle school (Poteat and Anderson, 2012).

At the same time, youth generally report stronger prejudice and more engagement in social exclusion at younger ages. These trends in prejudice and social exclusion may have multiple origins. On one hand there are significant cognitive and developmental changes during adolescence that enable teens to evaluate human rights, fairness, and prejudice with more sophistication as they age. Other explanations emphasize self-consciousness and peer regulation that are heightened during the early adolescent years. Whether explained by cognitive, interpersonal, or social developmental factors,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10243-0

compared to their older counterparts, younger adolescents are less adept at navigating social differences that may involve stigma or prejudice. Thus, as youth have begun to come out at younger ages across recent years, this downward trend in age at coming out collides with age-related trends in homophobia. Considering these two competing patterns – the historical trend that LGBTQ youth are coming out at younger ages and the developmental course of prejudicial attitudes and behaviors – it is likely that contemporary youth's school experiences may be more characterized by homophobia and prejudice than was true for prior cohorts, despite the broad trends in social acceptance of LGBTQ people and issues (Poteat and Russell, 2013).

Homophobia and Bullying

There is a broad discourse regarding homophobia and bullying; in the research literature studies have considered multiple relevant dimensions: homophobic epithets (name-calling; use of the phrase 'that's so gay'), perceptions of safety and belonging, physical and verbal harassment, and bullying and cyberbullying in schools. Homophobic epithets are ubiquitous in US high schools. According to a report by the Gay, Lesbian and Straight Education Network (GLSEN; Kosciw et al., 2012), "84.9% of students heard 'gay' used in a negative way (e.g., 'that's so gay') frequently or often, and 91.4% reported that they felt distressed because of this language" (p. xiv). In addition, over two-thirds heard other homophobic remarks (e.g., 'dyke' or 'faggot') frequently or often while 61.4% heard negative remarks about gender nonconformity. Students report homophobic language in school not only from peers, but from educators as well: Over half of students reported hearing negative remarks about sexual orientation and gender nonconformity from teachers or other school staff.

From the perspective of many youth, the phrase 'that's so gay' is not discriminatory but is understood as a general put-down; this perspective has been acknowledged by scholars (e.g., McCormack, 2012). Yet perspective and perception matters; there is clear evidence that for LGBTQ young people, 'that's so gay' is interpreted as personally undermining or distressing (Kosciw et al., 2012). The GLSEN report documents a slow but steady decline in the frequency of expressions such as 'that's so gay' since 2001 (Kosciw et al., 2012). This result comes from a large nationwide study of LGBTQ students, and is the first empirical indication that this form of homophobia in schools may be on the decline. Although the use of this phrase seems to be declining, the same study showed that negative remarks about gender nonconformity have changed very little. Thus, although there may be increasing understanding that the use of the word 'gay' as a put-down is problematic, a trend which follows the growing social acceptance of LGBTQ people and issues, there appears to be little change in adolescents' sensitivity to negative remarks regarding conformity to the boundaries of masculinity and femininity.

Homophobic language at school contributes to school climates that may be unsafe for LGBTQ students. In fact, feeling unsafe at school is a critical issue for many sexual and gender minority students: 63.5% of LGBTQ students felt unsafe because of their sexual orientation, and 43.9% because of their gender expression. More than one-third (38.3%) were physically harassed in the past year because of their sexual orientation, and 27.1% because of their gender expression (Kosciw et al., 2012). This leads to multiple negative outcomes for sexual and gender minority students: Sexual minority students who experience high levels of victimization are more likely to skip school because they fear for their safety, smoke cigarettes, drink and use substances more often, and engage in suicidal behavior compared to nonsexual minority students (Bontempo and D'Augelli, 2002). Further, LGBTQ youth who are gender nonconforming are at risk for depression and life dissatisfaction due to LGBT-specific harassment (Toomey et al., 2010).

LGBTQ youth are frequently victims of harassment, discrimination, and violence within schools. This harassment takes on many forms in addition to bias-based language, including verbal teasing, bullying, and physical aggression (Russell et al., 2010). GLSEN reported that over 50% of LGBTQ students were verbally harassed and over 25% were physically harassed (Kosciw et al., 2012). However, there are group differences between sexual and racial/ethnic minorities: Bisexuals experience higher levels of victimization than other sexual minority groups and white and Latino adolescents experience higher levels of victimization than Black and Asian adolescents (Russell et al., in press). In a study of California schools, white students reported more sexual identity-based victimization than students of color (Russell et al., 2009). Similar patterns of victimization have been found in school samples in other countries. For example, homophobic behaviors are reported to be widespread among Italian students, with verbal and physical abuse being commonplace within schools (Prati et al., 2011).

Constant exposure to these conditions has been shown to have negative consequences, as discrimination and stigmatization within schools can have a negative impact on the mental health of youth. A number of studies have documented associations between school bullying, victimization, and abuse and compromised mental health among LGBTQ youth. In particular, D'Augelli and colleagues (2002) found that youth who were verbally abused in school were likely to experience symptoms of depression, anxiety, and sleep disturbances. Similarly, reports by GLSEN show associations between severe victimization and lower self-esteem as well as higher levels of depression (Kosciw et al., 2012, 2013). Other research shows that depression and suicidal feelings were low for all students, regardless of sexual orientation, when they attended a school with a positive climate (Birkett et al., 2009).

Although more attention has focused on the link between homophobia and bullying at school and compromised mental and behavioral health, some studies have examined links with education attainment and academic outcomes. There is evidence to suggest an association between bias-based harassment and academic achievement or grade-point average (Kosciw et al., 2012, 2013; Russell et al., 2001). Victimization and feeling unsafe and unwelcome in schools can also lead to missing school and negative attitudes toward school (Espelage et al., 2008; Kosciw et al., 2012; Birkett et al., 2009).

Finally, recent studies have begun to document the impact of discriminatory bullying in comparison to bullying that is not due to discrimination or prejudice. This growing body of

evidence shows that bias-based or discriminatory bullying is particularly undermining of adolescent health. Boys who are bullied based on perceived sexual orientation report higher anxiety and depression and more frequent verbal and physical bullying than boys who are bullied for other reasons (Swearer et al., 2008). In a Brazilian sample, victimization in schools was related to lower commitment to school, and this association was stronger for homophobic compared to nonhomophobic victimization (Alexander et al., 2011). Additionally, Sinclair and colleagues (2012) demonstrated that compared to general harassment, adolescents who were victims of bias-based harassment were at significantly higher risk for negative mental health outcomes, such as depression, physical self-harm, and suicidal ideation and attempts. The same study showed that the link between discriminatory bullying and negative health and behavior outcomes was magnified for youth who also reported cyberbullying. Given the increasing use of mobile technologies and social media by youth, these results raise concern about the intersection of cyberbullying and discriminatory attitudes and prejudice.

School Climate

Although school climate has been found to be overwhelmingly negative for many LGBTQ students, school cultures and educational contexts vary: Some schools are safer and more supportive of LGBTQ students than others. There is a growing body of work that focuses on the structural conditions of schools and how policies and programs can change the experiences of LGBTQ students (Russell et al., 2010). With this focus, an increasing number of policies, programs, and practices are being introduced to promote a positive school environment for all students. In addition there is growing attention (including policy attention) to curriculum that is inclusive of LGBTQ people and issues. Finally, school clubs such as GSAs have emerged as an important setting for positive development and support for LGBTQ students, and for creating positive school climates.

National and State-Level Laws

At this time there are no nationwide protections in place for LGBTQ students in the United States. Both the Student Non-Discrimination Act, introduced in the 111th session of Congress, and the Safe Schools Improvement Act, introduced in the 112th session of Congress, did not pass beyond the committee level. The proposed legislation aimed to protect LGBT students against bullying and discrimination in schools. Although there are no national protections for LGBT students, many states (as well as the District of Columbia) have laws prohibiting discrimination in schools based on sexual orientation (GLSEN, 2013). Of those states, more than a dozen have fully enumerated antibullying laws which ban harassment of students based on sexual orientation and gender identity (although nearly all of such laws to date are regarded as 'unfunded mandates' that provide no specific provision for capacity for enforcement; Russell et al., 2010).

In addition to enacting protective policies, a central issue in promoting a positive school climate is making students, teachers, and school officials fully aware of existing inclusive school policies: Students must be aware of and understand policies and how these policies affect them (Russell et al., 2010). Sandfort and colleagues (2010) found that when schools had consistent and clear rules and expectations, Dutch students with and without same-sex attraction did not differ in mental health outcomes. Further, teachers and administrators must fully understand existing policies and implement them when necessary; studies have shown that teacher support and intervention in harassment is strongly linked to students' perceptions of school safety for LGBTQ students (Russell et al., 2010).

Inclusive Curriculum

In addition to the support they provide, educators can also shape school climate with the material that they teach. Including material on LGBTQ history or on LGBTQ individuals who have influenced the fields of math, science, and literature can contribute to an inclusive environment in which students can feel less isolated and more empowered. Students identified an LGBTQ-inclusive curriculum as a strategy to promote positive school climate (Burdge et al., 2013): The *Implementing Lessons That Matter* report notes that school climate surveys conducted in schools that included an LGBTQ-inclusive curriculum showed greater feelings of safety and perceived support for LGBTQ individuals. Additionally, inclusive curriculum may help promote academic engagement in LGBTQ youth, as it is associated with higher student achievement (Kosciw et al., 2013). This study also found that it may be a moderating factor for the association between harassment and self-esteem, particularly for students in poorer school climates with worse harassment. Although positive associations were found from this inclusive curriculum, students did note that it was rare; most schools that did offer this curriculum did not include it across the school curriculum.

Gay-Straight Alliances

Along with inclusive school-level policies and curricula, GSAs may also positively impact a school's climate. These student-led clubs had their roots in early efforts to provide counseling and support to LGBT students in schools, and emerged as youth-led clubs in the northeast and west coast during the 1990s; by the mid-2000s it was clear that GSAs represented a national movement (Herdt et al., 2007). GSAs create a sense of community for LGBT students and their allies. Often, these clubs advocate for education of issues facing LGBT students among the school community. GSAs have been shown to create a safe space for students in schools; schools that include a GSA have been found to have decreased rates of victimization and greater feelings of safety and school connectedness (Goodenow et al., 2006; Kosciw et al., 2013). In a population-based sample of high school students, GSAs were associated with lower truancy, smoking, drinking, and suicidal ideation and attempts for both LGBTQ and heterosexual youth, though these associations were stronger for LGBTQ youth (Poteat et al., 2012). Important to consider, however, are the different ways that GSAs are associated with youth outcomes. A study by Toomey and colleagues (2011) found that

although the presence of a GSA in a high school was associated with higher well-being and educational attainment, GSA participation was not associated with lower negative outcomes or higher positive outcomes. However, higher perceived GSA effectiveness was associated with higher educational attainment and lower depression and substance abuse problems. This implies that it is the salience of GSAs and their perceived effectiveness in changing school environments that promotes the most positive outcomes for LGBTQ youth, regardless of individual student participation.

Conclusion

Implications for Research

Research on GSAs, policies for youth, and inclusive curriculum have contributed to a growing body of evidence on how best to create positive school climates for LGBTQ youth, yet there is still much more to be done. There is a dearth of interventions in the school environment that address the issues discussed, and even less research testing the programs, strategies, and interventions that do exist. It is crucial that interventions addressing bullying, mental and physical health, sexual education, and other important LGBTQ concerns are devised, implemented, and evaluated within the school environment. Additionally, our knowledge of the experiences of ethnic and racial minority LGBTQ adolescents is limited and much research in this area is necessary to understanding the school environment for all LGBTQ adolescents.

Implications for Schools, Parents, and Students

There are a number of ways in which to promote a positive school climate for all students. Advocating for national policy to protect students from discrimination and harassment based on sexual orientation and gender expression is paramount. Until such a policy is enacted, state laws continue to be crucial, and the nature of those laws must be disseminated to students and teachers.

Teachers are an especially important support for LGBTQ students. Professional development programs like *The Reduction of Stigma in Schools* program (Payne and Smith, 2011) can provide teachers with the information they need to support students and intervene as necessary when harassment occurs. Additionally, training teachers to implement inclusive curriculum can lead to safer schools for all students. At this time, however, these strategies are not typically requirements in schools of education and other teacher training programs. Many new and current teachers are unaware of LGBTQ issues and ways to support their students. It is imperative that classroom strategies to promote respect for all students are integrated into teacher and school administrator training and professional development programs.

See also: Gay and Lesbian Movements; Gay, Lesbian, and Bisexual Youth; Heterosexism and Homophobia; Sexual Orientation: Historical and Social Construction.

Bibliography

Alexander, M.M., Santo, J.B., Da Cunha, J., Weber, L., Russell, S.T., 2011. Effects of homophobic versus nonhomophobic victimization on school commitment and the moderating effect of teacher attitudes in Brazilian public schools. Journal of LGBT Youth 8, 289–308.

Birkett, M., Espelage, D.L., Koenig, B., 2009. LGB and questioning students in schools: the moderating effects of homophobic bullying and school climate on negative outcomes. Journal of Youth and Adolescence 38, 989–1000.

Bontempo, D.E., d'Augelli, A.R., 2002. Effects of at-school victimization and sexual orientation on lesbian, gay, or bisexual youths' health risk behavior. Journal of Adolescent Health 30, 364–374.

Burdge, H., Snapp, S., Laub, C., Russell, S.T., Moody, R., 2013. Implementing Lessons that Matter: The Impact of LGBTQ-Inclusive Curriculum on Student Safety, Well-being, and Achievement. Gay-Straight Alliance Network and Tucson, AZ: Frances McClelland Institute for Children, Youth, and Families at the University of Arizona, San Francisco, CA.

D'Augelli, A.R., Pilkington, N.W., Hershberger, S.L., 2002. Incidence and mental health impact of sexual orientation victimization of lesbian, gay, and bisexual youths in high school. School Psychology Quarterly 17, 148–167.

Espelage, D.L., Aragon, S.R., Birkett, M., Koenig, B.W., 2008. Homophobic teasing, psychological outcomes, and sexual orientation among high school students: what influence do parents and schools have? School Psychology Review 37 (2), 202–216.

Floyd, F.J., Bakeman, R., 2006. Coming-out across the life course: implications of age and historical context. Archives of Sexual Behavior 35, 287–296.

Gay, Lesbian, & Straight Education Network (GLSEN), 2013. Retrieved March 8, 2013 from http://glsen.org/article/state-maps.

Goodenow, C., Szalacha, L., Westheimer, K., 2006. School support groups, other school factors, and the safety of sexual minority adolescents. Psychology in the Schools 43, 573–589.

Grov, C., Bimbi, D.S., Nanín, J.E., Parsons, J.T., 2006. Race, ethnicity, gender, and generational factors associated with the coming-out process among gay, lesbian, and bisexual individuals. Journal of Sex Research 43 (2), 115–121.

Herdt, G., Russell, S.T., Sweat, J., Marzullo, M., 2007. Sexual inequality, youth empowerment, and the GSA: a community study in California. In: Teunis, N., Herdt, G. (Eds.), Sexual Inequalities. University of California Press, Berkeley, pp. 233–252.

Kosciw, J.G., Greytak, E.A., Diaz, E.M., 2009. Who, what, where, when, and why: demographic and ecological factors contributing to hostile school climate for lesbian, gay, bisexual, and transgender youth. Journal of Youth and Adolescence 38, 976–988.

Kosciw, J.G., Greytak, E.A., Bartkiewicz, M.J., Boesen, M.J., Palmer, N.A., 2012. The 2011 National School Climate Survey: The Experiences of Lesbian, Gay, Bisexual and Transgender Youth in Our Nation's Schools. GLSEN, New York.

Kosciw, J.G., Palmer, N.A., Kull, R.M., Greytak, E.A., 2013. The effect of negative school climate on academic outcomes for LGBT youth and the role of in-school supports. Journal of School Violence 12, 45–63.

Loftus, J., 2001. America's liberalization in attitudes toward homosexuality, 1973 to 1998. American Sociological Review 66, 762–782.

McCormack, M., 2012. The Declining Significance of Homophobia: How Teenage Boys are Redefining Masculinity and Heterosexuality. Oxford University Press, Oxford; New York.

Payne, E.C., Smith, M., 2011. The reduction of stigma in schools: a new professional development model for empowering educators to support LGBTQ students. Journal of LGBT Youth 8, 174–200.

Poteat, V.P., Anderson, C.J., 2012. Developmental changes in sexual prejudice from early to late adolescence: the effects of gender, race, and ideology on different patterns of change. Developmental Psychology 48, 1403–1415.

Poteat, V.P., Russell, S.T., 2013. Understanding homophobic behavior and its implications for policy and practice. Theory into Practice 52 (4), 264–271.

Poteat, V.P., Sinclair, K.O., DiGiovanni, C.D., Koenig, B.W., Russell, S.T., 2012. Gay-straight alliances are associated with student health: a multischool comparison of LGBTQ and heterosexual youth. Journal of Research on Adolescence, 1–12.

Prati, G., Pietrantoni, L., D'Augelli, A.R., 2011. Aspects of homophobia in Italian high schools: students' attitudes and perceptions of school climate. Journal of Applied Social Psychology 41, 2600–2620.

Rosario, M., Schrimshaw, E.W., Hunter, J., 2004. Ethnic/racial differences in the coming-out process of lesbian, gay, and bisexual youths: a comparison of sexual identity over time. Cultural Diversity and Ethnic Minority Psychology 10, 215–228.

Russell, S.T., Clarke, T.J., Laub, C., 2009. Multiple Forms of Biased-Based Harassment at School (California Safe Schools Coalition Research Brief No. 8). California Safe Schools Coalition, San Francisco, CA.

Russell, S.T., Everett, B.G., Rosario, M., Birkett, M.A. Indicators of victimization and sexual orientation among adolescents: analyses from youth risk behavior surveys. American Journal of Public Health, in press.

Russell, S.T., Seif, H., Truong, N.L., 2001. School outcomes of sexual minority youth in the United States: evidence from a national study. Journal of Adolescence 24, 111–127.

Russell, S.T., Kosciw, J., Horn, S.S., Saewyc, E., 2010. Safe schools policy for LGBTQ students. Social Policy Report 24, 1–24.

Sandfort, T.G.M., Bos, H.M.W., Collier, K.L., Metselaar, M., 2010. School environment and the mental health of sexual minority youths: a study among Dutch young adolescents. American Journal of Public Health 100, 1696–1700.

Sinclair, K.O., Bauman, S., Poteat, V.P., Koenig, B., Russell, S.T., 2012. Cyber and bias-based harassment: associations with academic, substance use, and mental health problems. Journal of Adolescent Health 50, 521–523.

Swearer, S.M., Turner, R.K., Givens, J.E., Pollack, W.S., 2008. "You're so gay!": Do different forms of bullying matter for adolescent males? School Psychology Review 37, 160–173.

Toomey, R.B., Ryan, C., Diaz, R.M., Card, N.A., Russell, S.T., 2010. Gender-nonconforming lesbian, gay, bisexual, and transgender youth: school victimization and young adult psychosocial adjustment. Developmental Psychology 46, 1580–1589.

Toomey, R.B., Ryan, C., Diaz, R.M., Russell, S.T., 2011. High school gay-straight alliances (GSAs) and young adult well-being: an examination of GSA presence, participation, and perceived effectiveness. Applied Developmental Science 15, 175–185.

School Improvement

Bert PM Creemers, Faculty of Behavioural and Social Sciences, GION Institute for Educational Research, University of Groningen, The Netherlands
Leonidas Kyriakides, Faculty of Social Sciences and Education, University of Cyprus, Cyprus

© 2015 Elsevier Ltd. All rights reserved.

Abstract

School improvement is an innovation or planned change with specific means and specific goals. There are many and diverse improvements in the paper projects but the relation between the strategies and the outcomes of student learning is difficult to answer. The merging of educational effectiveness research, which concentrates on the relation between causes and effects in education, and school improvement, which emphasizes how this knowledge can be used, can result in theory-driven and evidence-based approaches to school improvement. The article presents and discusses successful examples of this recent development in school improvement.

School improvement is an innovation or planned change with specific means (a change in learning and other internal conditions of the school) and specific goals (to accomplish educational goals more effectively). There is a wide variety of improvement efforts, which more or less meet the elements of the definition, some enhancing student outcomes, others more directed toward the professional development of teachers, toward restructuring the organization of the school, or toward improving the relationship between schools and their clients (Creemers and Kyriakides, 2012). Because of systematic evaluation of most school improvement efforts, questions about the relationship between improvement strategies, the technicalities and outcomes of improvement – ultimately students' learning outcomes – cannot be answered. A major contribution to establishing cause–effect and/or means–ends relationships in school improvement is the merging of school improvement and school effectiveness. School effectiveness can provide a knowledge base about 'what works' and 'why' in education; school improvement provides understanding of 'how' that knowledge can be introduced in educational practice. This merging of school effectiveness and school improvement has resulted in some successful projects, and the formulation of principles and conditions for effective school improvement.

School Improvement

Different terms are used to describe change processes in schools and educational systems as a whole. They include 'educational change,' 'innovation,' 'reform,' and 'development.' These terms are quite often used in an interchangeable manner, and their meaning in the literature is frequently ambiguous. Many authors do not define them, but instead tend to trust everyday usage to convey the meaning of the words. Other authors define the concept by including in the definition important conditions and crucial characteristics of the processes.

In the definition of these different concepts generally, and to distinguish school improvement from other concepts, at least three points should be taken into consideration.

1. The level in the educational system where the change takes place: the school, the district, or the state. Ultimately, every change process has to reach the classroom and the level of the individual student. In order to make a distinction, educational change at the national level is often called 'systemic reform.' The term 'school improvement' makes it clear that the change takes place at the school level.
2. In all definitions, distinctions should be made between the process of educational change (i.e., the strategies used and the characteristics of the change process) and the outcomes of the change process. In everyday use, the term can refer to one of these aspects or even to a combination of them, for example, to strategy and the product of change. Fullan clarifies what he calls "the confusion between the terms *change* and *progress*" (Fullan, 1991, p. 4). Not every change is progress: the actual outcome of the change processes in terms of progress is not as important as the intended outcomes in the desired direction.
3. When it comes to the outcomes of change, there is often no well-defined concept of what the change is aimed at: the context of education, the inputs and processes, or, ultimately, the outcomes in terms of student achievement.

The International School Improvement Project defines school improvement as a systematic sustained effort aimed at change in learning conditions and other related internal conditions in one or more schools, with the ultimate aim of accomplishing educational goals more effectively (Van Velzen et al., 1985). This definition specifies improvement as an innovation or planned change with specific means (change in learning and other internal conditions of the school) and specific goals (ultimately to accomplish educational goals more effectively).

Practice and Theory in School Improvement

School improvement as a process has been demarcated into three global phases: the initiation, the implementation, and the continuation processes. Stakeholders of school improvement are more narrowly circumscribed and consist of students, teachers, principals, and parents at the school level, and of local educational authorities, consultants, and the local community at the local level. In exceptional cases, where it is a regional or national strategy to improve all schools, the stakeholders at these levels are also relevant.

School improvement is a very widespread phenomenon and a wide variety of improvement efforts can be found. For example, there is a lot of improvement going on which does not aim at enhancing student outcomes at all. These types of improvement focus, for instance, on the career development of teachers, on restructuring the organization of the school, on the way decisions are made, or on the relationships between schools and their clients. Sometimes restructuring takes place at the classroom level. Changes in the writing practices of elementary schoolteachers, for example, have been described in great detail. Nevertheless, the actual impact of the changes on students and on student achievement is usually left out.

Some schools practice improvement on their own and try to find their own solutions to their problems. Other schools have chosen to implement as accurately as possible improvement programs that have been developed elsewhere. This is called fidelity implementation. Often they do not consider alternative educational options. For example, they may opt for a specific program primarily because another school was satisfied with it (Stringfield, 1995). Some schools are involved in improvement only because their government expects them to be (Hopkins et al., 1994). Depending on the extent to which educational policies are translated into clear outlines and contents, fidelity implementation is a more or less appropriate concept. When the educational policy is rather prescriptive with respect to curricular content and outcome levels, as it is in the UK (Hopkins et al., 1994), it is theoretically possible to check whether or not schools are implementing this policy. When the educational policy is rather open-ended, as is the Dutch policy on educational priorities (Van der Werf, 1995) or the Dutch inclusion policy, it is virtually impossible to apply the notion of fidelity implementation.

Only a small part of school improvement is based on research (Stringfield, 1995). Innovations are hardly ever tested before they are implemented in educational practice, and an adequate *evaluation* of their impact is rare. The same applies to the occurrence of experimental or quasiexperimental designs in improvement projects. Some school improvers have preferred forms of action research instead of research-based experiments. Sometimes projects based on changing only a couple of factors report great successes. It may be that changes in just a few important areas can alter a whole system, but the question is whether these changes – often cases of educational mass hysteria – will stand the test of time. In education, the margins for change are generally small, because of the impact of noneducational factors such as the individual characteristics of students.

Evaluations are often not carried out satisfactorily or are performed after a time span that is too short. Because of inadequate evaluations, questions about the causation of effects, extremely important for the school effectiveness knowledge base, cannot be answered. School improvement, therefore, should consistently pursue assessment of its results, pay attention to its failures and try to learn from these, and limit its goals to prevent confusion of cause and effect (Hopkins et al., 1999).

In school improvement, the orientation toward educational practice and policy-making is emphasized. Although in educational practice schools and classrooms can be found that succeed much better than others, theories cannot be based on exemplary practice only. Insofar as theoretical notions have been developed, they have not yet been empirically and systematically tested. The typologies of school cultures (Hargreaves, 1995), for example, have not been studied in educational practice and their effects on the success of school improvement are as yet unknown. Moreover, their relationships to the first criterion for school improvement (enhancing student outcomes) are not always very clear. The same holds for the factors that are supposed to be important in different stages of educational change, outlined by Fullan (1991) and for his ideas about the essential elements in educational change at the classroom level (beliefs, curriculum materials, and teaching approaches). Even though these ideas are derived from school improvement practice, their importance and their potential effects are not accounted for in detail, and have not yet been studied in research.

A contribution to theory development is the generic framework for school improvement provided by Hopkins (1996). In this framework, three major components are depicted: educational givens, a strategic dimension, and a capacity-building dimension. Educational givens cannot be changed easily. Givens can be external to the school (such as an external impetus for change) and internal (such as the school's background, organization, and values). The strategic dimension refers to the competency of a school to set its priorities for further development, to define student learning goals and teacher development, and to choose a strategy to achieve these goals successfully. The capacity-building dimension refers to the need to focus on conditions for classroom practice and for school development during the various stages of improvement. Finally, the school culture has a central place in the framework. Changes in the school culture will support teaching–learning processes which will, in turn, improve student outcomes (Hopkins, 1996).

Despite the obvious gaps in theory development and testing in the field of school improvement, there are already some elements of a knowledge base. By trying to improve schools, knowledge on the implementation of classroom and school effectiveness factors in educational practice has become available. This has provided the possibility of studying, to different degrees, the influence of factors and variables on educational outcomes (Stoll and Fink, 1994).

The Link between School Improvement and School Effectiveness

School effectiveness and school improvement conjoin because of their mutual interest, although their actual relationship may be very complicated (Creemers and Reezigt, 1997; Reynolds et al., 1993). The major aim in the field of school effectiveness was always to link theory and research on the one hand, and practice and policy-making on the other. School improvement can be fostered by a knowledge base covering what works in education that can be applied in educational practice. The combination of theory, research, and development is not new in education. Almost all movements start out to make knowledge useful for educational practice and policy-making, or state their goal in terms of supplementing policy practice with a knowledge base supplied by theory and research

from a cyclical point of view. The next step is to use practical knowledge for further advances in theory and research. In this way, research and improvement can have a relationship as a surplus benefit for both.

School effectiveness has led to major shifts in educational policy in many countries by emphasizing the accountability of schools and the responsibility of educators to provide all children with opportunities for high achievement, thereby enhancing the need for school improvement (Mortimore, 1991). School effectiveness has pointed to the need for school improvement, in particular by focusing on alterable school factors. School improvement projects were necessary to find out how schools could become more effective. These projects were often supposed to implement effective school factors in educational practice (Scheerens and Bosker, 1997) and, in doing so, could yield useful feedback for school effectiveness. School improvement might point to inaccurate conceptions of effectiveness, such as the notion of linearity or one-dimensionality (Hargreaves, 1995). In addition, school improvement might give more insight into the strategies for changing schools successfully in the direction of effectiveness.

The relatively short history of school effectiveness and improvement shows some successes of this linkage. Research results are being used in educational practice, sometimes with good results. School improvement findings are sometimes used as input for new research. Renihan and Renihan state that "the effective schools research has paid off, if for no other reason than that it has been the catalyst for school improvement efforts" (Renihan and Renihan, 1989, p. 365).

Most authors, however, are more skeptical (Reynolds et al., 1993). Fullan states that school effectiveness "has mostly focused on narrow educational goals, and the research itself tells us almost nothing about how an effective school got that way and if it stayed effective" (Fullan, 1991, p. 22). Stoll and Fink (1992) think that school effectiveness should have done more to make clear how schools can become effective. According to Mortimore (1991), a lot of improvement efforts have failed because research results were not translated adequately into guidelines for educational practice. Changes were sometimes forced upon a school, and when the results were disappointing the principals and teachers were blamed. Teddlie and Roberts (1993) suggest that effectiveness and improvement representatives do not cooperate automatically, but tend to see each other as competitors. Links between school effectiveness and school improvement were stronger in some countries than in others (Reynolds, 1996). Reynolds et al. (2000) came to the conclusion based on the analyses of Dutch, British, and North American initiatives that the following principles are fundamental to a successful merger of school effectiveness and school improvement:

1. a focus on teaching, learning, and the classroom level,
2. use of data for decision-making,
3. a focus on pupil outcomes,
4. addressing schools' internal conditions,
5. enhanced consistency (through implementation of 'reliable' programs), and
6. pulling levers to affect all levels, both within and beyond the school.

Recently, these elements were taken into account in developing a project aiming to use the knowledge base of school effectiveness research to develop school improvement initiatives, namely the high reliability project, and some evidence illustrating the impact of this approach to student achievement gains has been provided (Kyriakides and Christoforou, 2011; Reynolds, 2011).

In addition, an evidence-based and theory-driven approach to school improvement has been developed (Creemers and Kyriakides, 2012). This approach makes use of the dynamic model of educational effectiveness (Creemers and Kyriakides, 2008) which was developed in order to address some of the weaknesses of integrated models of educational effectiveness developed in the 1990s especially since these models do not take into account the dynamic character of educational effectiveness. The dynamic model is multilevel in nature and refers to factors operating at four different levels: student, classroom, school, and system. Changes in the functioning of school factors were found to be associated with changes in the effectiveness status of schools (Creemers and Kyriakides, 2010). Thus, school improvement projects are expected to refer to these factors. It is also claimed that factors are interrelated and thereby comprehensive strategies of school improvement addressing grouping of factors can be developed. Finally, five dimensions are used to measure the functioning of each factor and these can be used for generating constructive feedback to schools on both quantitative and qualitative characteristics of school factors (Kyriakides and Creemers, 2008).

A distinctive feature of the dynamic model is that it does not only refer to factors that are important for explaining variation in educational effectiveness but it also attempts to explain why these factors are important by integrating different theoretical orientations to effectiveness (Sammons, 2009). In this way, teachers and other school stakeholders involved in improvement efforts are offered the opportunity to use in a flexible way the knowledge base of school effectiveness research, adapt it to their specific needs, and develop their own strategies for school improvement. Thus, a dynamic approach to school improvement (DASI) has recently been developed (Creemers and Kyriakides, 2012). The dynamic model is used as the theoretical framework of this approach and a significant role to school self-evaluation for identifying improvement priorities is given. In addition, the advisory and research team (A&RTeam) has an important role to play since it is expected to support school stakeholders in their attempt to develop and implement their strategies and action plans for school improvement.

More specifically, school stakeholders and the A&RTeam should first of all define the general aim of the school improvement intervention by taking into account that promotion of student learning must be the ultimate aim of any school improvement effort. At the second step, the dynamic model and its factors are presented to the school stakeholders. This presentation will help them understand how and why by addressing the school factors student learning can be promoted. At the third step, schools should develop their own school self-evaluation mechanisms in order to collect and analyze data about the functioning of school factors and identify their priority area(s) for improvement. The fourth step is one of the most important steps of DASI. The A&RTeam should work closely with the

school stakeholders in order to help them define their strategies and action plans for improvement. School stakeholders are expected to take into account the available knowledge base of educational effectiveness research and adopt the guidelines emerging from the literature on their school context (with the help of the A&RTeam). Then school stakeholders and the A&RTeam should develop mechanisms for monitoring the implementation of the intervention. At this point, DASI gives emphasis to the role of formative evaluation and the importance of using evaluation data to further develop the school improvement strategies and action plans. Finally, the A&RTeam and the school stakeholders should develop summative evaluation mechanisms in order to measure the impact of DASI on promoting student learning. This step will not only help school stakeholders test the theory upon which their intervention is based but may also reveal the importance of identifying a new priority area for improvement. If summative evaluation reveals that a school has managed to substantially improve the functioning of the factor(s) addressed, school stakeholders and the A&RTeam may decide to collect new evaluation data and identify a new priority improvement area. By conducting school self-evaluation a new priority area will be identified and a new improvement project will be developed. It can therefore be claimed that DASI is based on the assumption that effective schools always search for ways to improve further their effectiveness status.

In order to demonstrate the development and use of DASI in educational practice, projects showing the conditions and methods with which the proposed approach can be used for school and teacher improvement have been undertaken (Antoniou and Kyriakides, 2011; Bren et al., 2012). These projects address important challenges which many schools in different countries face, such as establishing school self-evaluation mechanisms, facing and reducing bullying, improving assessment practices, and promoting teacher professional development. It was shown in these studies that the DASI approach helped participating schools to develop their own strategies and actions for improvement, in relation to specific challenges they faced, and also assisted in the improvement of their effectiveness status as this was measured through student achievement in different learning outcomes.

During the last decade, we have different approaches to school improvement that are based on the knowledge base of educational effectiveness research (Hallinger and Heck, 2011; Houtveen and van de Grift, 2012; Mongon and Chapman, 2012; Muijs et al., 2011). Some projects investigating the extent to which these approaches can be used in practice have been undertaken. Further research is needed to compare the elements of these approaches and their impact on student learning outcomes. In this way, a more comprehensive approach to school improvement which will be based on a theory of effective education and will be able to promote quality and equity in education could be developed.

See also: Educational Effectiveness Theory and Research: Recent Advances; Educational Effectiveness, The Field of; Educational Evaluation: Overview; School Effectiveness Research; School Management.

Bibliography

Antoniou, P., Kyriakides, L., 2011. The impact of a dynamic approach to professional development on teacher instruction and student learning: results from an experimental study. School Effectiveness and School Improvement 22 (3), 291–311.

Bren, M., Hauptman, A., Cankar, G., Zupanc, D., 2012. Schools establishing school self-evaluation mechanisms for improvement purposes: case studies of two primary schools in Slovenia. In: Creemers, B.P.M., Kyriakides, L. (Eds.), Improving Quality in Education: Dynamic Approaches to School Improvement. Routledge, London, pp. 106–118.

Creemers, B.P.M., Kyriakides, L., 2008. The Dynamics of Educational Effectiveness: A Contribution to Policy, Practice and Theory in Contemporary Schools. Routledge, London.

Creemers, B.P.M., Kyriakides, L., 2010. Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. School Effectiveness and School Improvement 21 (4), 409–427.

Creemers, B.P.M., Kyriakides, L., 2012. Improving Quality in Education: Dynamic Approaches to School Improvement. Routledge, London.

Creemers, B.P.M., Reezigt, G.J., 1997. School effectiveness and school improvement: sustaining links. School Effectiveness and School Improvement 8 (4), 396–429.

Fullan, M., 1991. The New Meaning of Educational Change. Teachers College, New York.

Hallinger, P., Heck, R.H., 2011. Exploring the journey of school improvement: classifying and analyzing patterns of change in school improvement processes and learning outcomes. School Effectiveness and School Improvement 22 (1), 1–27.

Hargreaves, D.H., 1995. School culture, school effectiveness and school improvement. School Effectiveness and School Improvement 6, 23–47.

Hopkins, D., 1996. Towards a theory for school improvement. In: Gray, J., Reynolds, D., Fitz-Gibbon, C., Jesson, D. (Eds.), Merging Traditions: The Future of Research on School Effectiveness and School Improvement. Cassell, London, pp. 30–51.

Hopkins, D., Ainscow, M., West, M., 1994. School Improvement in an Era of Change. Cassell, London.

Hopkins, D., Reynolds, D., Gray, J., 1999. Moving on and moving up: confronting the complexities of school improvement in the improving schools project. Educational Research and Evaluation 5, 22–40.

Houtveen, T., van de Grift, W., 2012. Improving reading achievements of struggling learners. School Effectiveness and School Improvement 23 (1), 71–93.

Kyriakides, L., Christoforou, C., April 2011. A synthesis of studies searching for teacher factors: implications for educational effectiveness theory. In: Paper Presented at the American Educational Research Association (AERA) 2011 Conference. New Orleans.

Kyriakides, L., Creemers, B.P.M., 2008. Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. School Effectiveness and School Improvement 19 (2), 183–205.

Mongon, D., Chapman, C., 2012. High Leverage Leadership. Routledge, London; New York.

Mortimore, P., 1991. School effectiveness research: which way at the crossroads? School Effectiveness and School Improvement 2 (3), 213–229.

Muijs, D., Ainscow, M., Chapman, C., West, M., 2011. Collaboration and Networking in Education. Springer, New York.

Renihan, F.I., Renihan, P.J., 1989. School improvement: second generation issues and strategies. In: Creemers, B.P.M., Peters, T.A., Reynolds, D. (Eds.), School Effectiveness and School Improvement. Swets & Zeitlinger, Amsterdam; Lisse, pp. 365–377.

Reynolds, D., 1996. County reports from Australia, the United Kingdom, and the United States of America. Introduction and overview. School Effectiveness and School Improvement 7, 111–113.

Reynolds, D., 2011. Failure-Free Education? The Past, Present and Future of School Effectiveness and School Improvement. Routledge, London.

Reynolds, D., Hopkins, D., Stoll, L., 1993. Linking school effectiveness knowledge and school improvement practice: towards a synergy. School Effectiveness and School Improvement 4 (1), 37–58.

Reynolds, D., Teddlie, C., Hopkins, D., Stringfield, S., 2000. Linking school effectiveness and school improvement. In: Teddlie, C., Reynolds, D. (Eds.), The International Handbook of School Effectiveness Research. Falmer Press, London, pp. 206–231.

Sammons, P., 2009. The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools. School Effectiveness and School Improvement 20 (1), 123–129.

Scheerens, J., Bosker, R.J., 1997. The Foundations of Educational Effectiveness. Pergamon, Oxford.

Stoll, L., Fink, D., 1992. Effecting school change: the Halton approach. School Effectiveness and School Improvement 3 (1), 19–41.

Stoll, L., Fink, D., 1994. School effectiveness and school improvement: voices from the field. School Effectiveness and School Improvement 5 (2), 149–177.

Stringfield, S., 1995. Attempting to enhance students' learning through innovative programs: the case for schools evolving into high reliability organisations. School Effectiveness and School Improvement 6 (1), 67–96.

Teddlie, C., Roberts, S.P., April 1993. More clearly defining the field: a survey of subtopics in school effectiveness research. In: Paper Presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Van der Werf, M.P.C., 1995. The Educational Priority Policy in The Netherlands: Content, Implementation and Outcomes. SVO, Den Haag, The Netherlands.

Van Velzen, W.G., Miles, M.B., Ekholm, M., Hameyer, U., Robin, D., 1985. Making School Improvement Work: A Conceptual Guide to Practice. ACCO, Leuven, Belgium.

School Learning for Transfer

Gavriel Salomon, Haifa University, Haifa, Israel
David N Perkins, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Most of what is taught in school is expected to affect learning and performance that transcend the mastery of those subjects. Such transfer from A to B is likely to occur when more elements are common to both or when one acquires a general principle that can be transferred to new instances. However, there is a great discrepancy between the expectation for transfer from school learning to the acquisition of new subjects and the handling of daily events where transfer appears more frequently. According to one theory, transfer takes place either through automatic application of well-rehearsed knowledge ('low road') or through mindful processes of abstraction. Often, neither road is taken and thus transfer fails. Recent developments challenge the basic paradigm and conception of transfer, suggesting for example to focus on transfer as 'preparation for future learning,' or seeing learning as a matter of participation in a community of learners in an active process of *knowing-as-construction within a social context of activity*. Some suggest a *detect–elect–connect* model, while others offer a more *expansive framing* that emphasize the future uses of learned content. Transfer can be facilitated by reflective abstraction (high road), practice to near automaticity (low road), emphasis on significance and potential diverse uses (expansive versus bounded framing), and fostering seeing through the surface to deeper structure.

Most of what is taught in school is expected to affect learning and performance that transcend the mastery of those subjects: one learns addition to prepare the grounds for the study of multiplication, science to understand the surrounding physical and natural world, and history to gain a sense of identity and a deeper understanding of current events. Such expectations concern the transfer of what is learned in school in one domain to other realms of learning and activity in school and beyond, a spillover of learning that functions like a ripple effect. Thus, the study of topic A is supposed to facilitate learning of and/or performance on B, a later topic or activity often of a very different nature and in a different context. The prior study of A makes B easier, faster, or better understood relative to B without the learning of A preceding it.

Sometimes a prior A may interfere with B; this is generally called negative transfer, as when having mastered driving a car interferes with steering a semitrailer truck in reverse, a theme not pursued further here. Also, the current literature recognizes complex interactions between A and B. In facing situation B and drawing on A to deal with it, the learner may reconstruct A to some extent and thereby learn more about A, for instance broadening its meaning or identifying variants (e.g., Cook, 2003; Ross and Kennedy, 1990).

Transfer has been of special interest to researchers and educators because of many findings of failure to transfer – school learning of A not having the spillover effects hoped for to B. Some have held that the prospects for transfer markedly beyond what is most immediately and specifically learned are intrinsically poor. Educational aspirations and planning should take this into account. However, the prospects for transfer are contentious, a theme developed further below.

Two Basic Approaches

Two approaches to the study of transfer have dominated the field since the beginning of the twentieth century: transfer as a function of similar elements between A and B, and as a function of the understanding of general principles drawn from A (and perhaps other context) that are transferable to a variety of situations including B. The history of research on transfer shows how these two major principles underlie the development of the field, even when other basic assumptions about learning and thinking have recently been challenged.

The expectation for transfer has as long a history as institutional learning itself. Plato argued that the study of abstract reasoning assists the solution of daily problems. Similarly, the debates about Talmudic and Biblical texts in ancient times were argued to 'sharpen minds.' The scientific study of transfer has a long history as well; it dates back to the beginning of the twentieth century with the studies by Thorndike and Woodworth (1901) in which the idea that Latin and other taxing school subjects developed one's general 'faculties of mind' was challenged. The expected transfer was not found; mastery of Latin, Greek, or geometry did not facilitate either performance or efficiency in learning other subjects. Thorndike's School Improvement findings led to his formulation of the theory of 'identical elements': transfer from A to B is likely to occur when more elements are common to both. It followed that transfer can take place only in the relatively rare case when there is clear and apparent similarity between the constituent elements of the learned topics. Thus, one would have to learn a host of independent issues and procedures as no transfer in the absence of identical elements should be expected.

Thorndike's theory was countered by a more cognitively oriented and Gestalt-flavored alternative that emphasized the meaningful conceptual understanding of underlying principles, which then could be thoughtfully applied to new situations. The more general the learned principle, the farther could it be transferred to new instances. Judd (1908) had his subjects learn to shoot at targets submerged in water after having learned principles of light refraction in water. Having learned these principles, subjects were much better at hitting the targets than those who have only practiced the activity.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92040-3

As we shall see below, the two approaches – transfer as a matter of common elements and transfer as a matter of higher-order cognitive processes of understanding and abstraction – although having undergone profound modifications and developments over the years, continue to this day to dominate the field.

Transfer Missing in Action

The history of research on transfer as well as practitioners' expectations for transfer might be characterized as an ongoing pendulum-like oscillation between the two approaches. One of the hallmarks of the field is the great discrepancy between the expectation for transfer from school learning to the acquisition of new subjects or to the handling of daily events and the findings from controlled transfer studies. The latter often do not yield evidence of the hoped-for transfer. A typical case comes out of research on children learning programming, which some thought might develop general cognitive capabilities of various kinds. Pea and Kurland (1984) found, contrary to expectations, that children learning LOGO programming showed no advantage in their ability to plan over children who did not master LOGO.

Transfer, when at all found in carefully carried out studies, appears to be highly specific, as when a particular procedure or principle is applied in a new situation which bears great and apparent (often analogical) similarity to the learning material, and when the process is accompanied by such facilitation as coaching, specific guidance, and cueing (see Bransford and Schwartz, 1999; Cox, 1997, for detailed reviews). Such findings would seem to support Thorndike's theory and pessimism about transfer, as summarized by Detterman (1993): "In short, from studies that claim to show transfer and that don't show transfer, there is no evidence to contradict Thorndike's general conclusions: Transfer is rare, and its likelihood of occurrence is directly related to the similarity between the two situations" (p. 15).

There is a notable anomaly here, however. Whatever the pattern of findings from formal studies, it is apparent that in everyday life a great deal of transfer occurs unproblematically (Bereiter, 1995). People readily understand metaphors, jokes, and explanatory analogies, all of which are based on the transfer of ideas from elsewhere – and often construct them on the fly – even though these commonly require connecting two very different conceptual domains. People routinely learn skills and stretch them to other contexts, as in knowing how to drive a car and readily driving a rented moving van. People acquire robust skills such as reading and apply them in diverse contexts. Transfer of certain sorts in certain settings is so plentiful that plainly there is more to understand about when transfer succeeds or fails and why.

Transfer on the Rebound

While Thorndike's pessimism seemed to have won the day for a while, the research tradition of Judd (1908), with its emphasis on comprehension and general (nonspecific) transfer gradually showed its strength. As research of this tradition tended to show, when relevant principles or strategies are mindfully attended to, or better yet – either abstracted by learners and/or metacognitively monitored – transfer to a variety of instances can be obtained even in the absence of apparent common elements.

At the same time, other research on the transfer of skills, ostensibly still in a tradition more similar to that originated by Thorndike (Thorndike and Woodworth, 1901), has also yielded positive results. Thus, for example, Singley and Anderson (1989) have shown that when training to near automaticity of discrete skill components is carried out, transfer from one task (e.g., a line editor) to another (text editor) can be virtually perfect. Also more field-based research found support for impressive transfer from the study of university-taught disciplines: Nisbett et al. (1993) have shown that the study of psychology, and also to a lesser extent medicine (but not law or chemistry), improves students' abilities to apply statistical and methodological rules of inference to both scientific and daily problems.

Such positive cases of transfer of learning stand in contrast to the many seeming cases of failure to transfer. How might the discrepancy be understood? Chi (in press) notes that failure to transfer has often been explained by shallow initial learning. In such circumstances, one still might speak of failure to transfer, albeit from impoverished initial learning. However, arguably, such cases are not best viewed as failure to transfer but simply as failure to achieve reasonably robust learning as a basis for transfer. This notion is developed in the following section.

The High and the Low Roads to Transfer

The renewed success of the two lines of research, often addressing higher-order cognitions and information processing exercised by multiple activities, suggests that transfer may not be a unitary process as the two approaches differ in important ways. Observing these differences, Salomon and Perkins (1989) have developed a theory to account for the possibility that transfer takes either one of two routes (or a combination thereof), described as the *high road* and the *low road* of transfer.

The low road reflects to an extent the Thorndikian line, and more recently that of the Singley and Anderson (1989) skill-acquisition and transfer theory. It is taken when skills, behaviors, or action tendencies are repeatedly practiced in a variety of situations until they are mastered to near automaticity and are quite effortlessly applied to situations whose resemblance to the learning situations is apparent and easily perceived (e.g., reading and driving). In this respect, transfer is not based on the development of abstract mental representations, "but through the incremental refinement of knowledge resources that account for – rather than overlook – contextual variation" (p. 1). Transfer is thus constructed piece by piece (Wagner, 2006). Other candidates for low-road transfer are attitudes, cognitive styles, dispositions, and belief systems, the application of which to new situations is rarely a mindful process.

In contrast, the high road to transfer is characterized by the process of mindful abstraction of knowledge elements that afford logical abstraction: principles, rules, concepts, procedures, and the like. It is this mindfully abstracted, decontextualized idea ('ethnic oppression may lead to revolt') that

becomes a candidate for transfer from one particular instance (the Austro-Hungarian Empire) to another (Russia and Cechenia). Simon (1980) commented in this respect that "To secure substantial transfer of skill acquired in the environment of one task, learners need to be made explicitly aware of these skills, *abstracted from their specific task content*" (p. 82, emphasis added).

There is ample evidence to support the role of the high road in obtaining transfer. One of the studies by Gick and Holyoak (1983) illustrates this point. Participants who were given two stories and were asked to write a summary of how the two stories resembled each other (i.e., their common moral) showed a 91% transfer to the solution of another story, relative to 30% transfer of a nonsummary group. Bassok (1990) showed that mastering algebraic abstractions (plus examples) allowed students to view physics problems as particular instances to which the more abstract algebraic operations could be applied. Physics, on the other hand, is too particular and thus students do not expect and do not recognize any possible relationship between it and algebraic operations. Research also directs attention to the role played by self-regulation and metacognitions in the process of mindful abstraction and transfer via the high road. Following the studies and review of Campione et al. (1982), Salomon et al. (1989) have shown that students interacting with a semi-intelligent computerized Reading Partner that provides metacognitive-like guidance tend to internalize that guidance and transfer it to new reading as well as to writing situations.

The high road–low road theory sheds light on the many failures of obtaining transfer in controlled studies. Close examination of such studies suggests that in many cases neither the low road of repeated practice nor the high road of mindful abstraction was taken. Not enough time is allocated for practice for the former, and not enough attention is given for mindful abstraction for the latter. As a consequence, neither near automatic transfer on the basis of easily recognized common elements, nor farther transfer on the basis of metacognitively guided mindful abstraction can be attained.

New Approaches

Research on transfer, until recently, did not challenge the basic paradigm and conception of transfer as a process involving change of one's performance on a new task as a result of his or her prior performance on a preceding and different task. Such a conception of transfer has recently become challenged on the basis of new theories of learning, the role of activity, and the place of cognitions therein. These challenges can be arranged along a dimension ranging from the least to the more radical, relative to traditional notions of transfer.

Bransford and Schwartz (1999), coming from a mainstream tradition of cognitive science applied to instructional issues, argue that the current ways of demonstrating transfer are more appropriate for studying only full-blown expertise. According to them, "There are no opportunities for…[students] to demonstrate their abilities to learn to solve new problems by seeking help from other resources such as text or colleagues or by trying things out, receiving feedback, and getting opportunities to revise" (p. 68). They recommend replacing the typical transfer task which they call 'sequestered problem solving' (SPS) by a conception of transfer as 'preparation for future learning' (PFL). Thus, one would measure transfer not by having students apply previously acquired knowledge or skill to new situations, demonstrating either *knowledge how* or *knowledge that* something, but rather by demonstrating *knowledge with* – thinking, perceiving, and judging with whatever knowledge and tools are available even if that knowledge is not consciously recalled. In other words, transfer would be expected to become demonstrated by way of 'people's abilities to learn new information and relate their learning to previous experiences' (Bransford and Schwartz (1999: 69).

An illustration of the above can be seen in a study in which college and fifth-grade students did not differ in their utilization of previous knowledge and educational experiences for the creation of a statewide plan to recover the bald eagle. This is a typical SPS way of measuring transfer, suggesting in this case that previous educational experiences did not transfer to the handling of the new problem. However, when the students were asked to generate questions that would need to be studied in preparation for the creation of a recovery plan (applying a PFL approach), striking differences were found between the two age groups in favor of the college students. Focusing on different issues, the two groups showed how, knowingly or not, previous learning has facilitated the generation of the list of questions to be asked. The use of tools and information resources would also show how previous learning of skills and knowledge becomes actively involved in the process of solving a new problem.

While Bransford and Schwartz (1999) continue to adhere to the traditional conception of learning as knowledge acquisition and to cognitions as general and transferable tools of the mind, others have developed an alternative, sociocultural approach according to which neither the mind and its content (e.g., arithmetic algorithms) should be treated as a neutral toolbox ready for application wherever suitable, nor should knowledge-in-use and social context of activity be taken as two independent entities. Learning is a matter of participation in a community of learners; thus knowledge is not a noun denoting *possession* but rather a verb denoting the active process of *knowing-as-construction within a social context of activity*. One underlying assumption is that learner, activity, and content become united through culturally constructed tools. A second assumption is that learning is highly situated in a particular activity context of participation, and its outcomes are 'becoming better attuned to constraints and affordances of activity systems so that the learner's contribution to the interaction is more successful' (Greeno, 1997: 12). Seen in this light, the traditional *cognitivist* concern with the transferability of acquired knowledge across tasks becomes, from the newer *situative* perspective, an issue concerned with the task and the 'consistency or inconsistency of patterns of participatory processes across situations' (Greeno, 1997: 12).

An important practical implication of this approach is that one would need to take into consideration the kinds of participatory activities to which school-based learning might be expected to transfer, rather than expect decontextualized school materials to transfer all on its own to other tasks, situations, and activities. Students are likely to restructure new situations to fit their previous practices and thus make the old available

for transfer to the new. Illustrating this approach to learning and cognition, Saxe (1991) showed that Brazilian children's mathematical experience as street vendors affected their math learning when later on they became enrolled in school, and vice versa. "[I]n recurring activities like practices of candy selling and schooling – contexts in which knowledge is well learned and in which similar problems emerge on a repeated basis – transfer is a protracted process. In their repeated efforts to solve practice-linked problems, individuals attempt to structure problem contexts in coherent and more adequate ways" (p. 179).

The theme of looking at transfer realistically in context, asking what can be expected, and examining what kinds of learning experiences might prepare learners for challenging transfers has proved persistently fruitful. For instance, Lobato (2003) has urged the importance of an *actor-oriented* perspective on transfer. She warns that studies of transfer commonly neglect what's going on with the learner. They reflect an outside or expert perspective, seeking the wide and deep transfer that would mark mastery of ideas. Lobato suggests that transfer in fact is common but takes modest forms initially, gradually building toward something broader, given further learning opportunities. Teaching and learning practices should expect and build upon this.

Chi (in press) also emphasizes how commonplace useful transfer is. Transfer comes easily when surface similarities and deep structural similarities go together, which they often do. The surface similarities cue the learner to make the deeper connections. Transfer gets tricky when surface dissimilarities mask deep structural similarities, leading learners to miss the opportunity, or when surface similarities trigger mistaken efforts to match up circumstances dissimilar under the surface. Appropriate transfer in such cases requires 'seeing through' the surface to deeper relationships. Studies of topics can be organized to cultivate this.

Relatedly, Perkins and Salomon (in press) urge understanding successes and failures of transfer according to a detect–elect–connect model. They note that treatments of transfer for the most part do not differentiate how learners *detect* possible relationships in the moment, why they *elect* to pursue them, and, if so, how successfully they build a meaningful *connection*. Any of the three can lead to failure of transfer. Imagine a student who has learned some good principles of informal negotiation. In a tense real-life situation, the student might never even think of those principles (a problem of *detect*), feel too indifferent to use them (*elect*), or have difficulty adapting them to the circumstances (*connect*). Effective teaching for transfer needs to prepare the learner for all three.

Emphasizing how classroom dynamics cultivate learner mindsets for or against transfer, Engle (2006) draws a provocative distinction between *bounded framing* and *expansive framing* of learning. In bounded framing, classroom practices treat the content to be learned as for today, for the test, to master because it's in the curriculum. In expansive framing, classroom practices emphasize the future uses of content in school and beyond, how content connects across areas, how learners can be responsible for bringing forward content from past learning and carrying content forward from today's learning, and similar themes. Engle argues that bounded versus expansive framing influences considerably what sorts of transfer occur.

The new approaches to transfer clearly deviate from the traditional views. They suggest new ways of measuring transfer, and more radically – challenge traditional assumptions by treating knowledge, skill, task, activity, and context as one participatory activity. By so doing, they emphasize the uniqueness of each activity setting. An important aspect of this approach is its ability to incorporate within the idea of participatory activity motivational and dispositional elements also. These elements have so far been neglected by much research on transfer, although they may be crucial for students' choice to treat each situation as different, or as allowing students to exercise familiar patterns of participation. Still, it appears that despite these important novelties, the principles of transfer as a function of similarity between situations (or activities) and of depth of understanding general principles (e.g., of useful participation in a community of learners) have not yet been replaced but only redefined.

The Prospects of Teaching for Transfer

Early conceptions of transfer and its teaching assumed that special strategies were not required. One taught as usual for the usual student attainments and expected substantial transfer. Abundant research has shown this optimism to be misplaced. For many kinds of content and skill, transfer to hoped-for targets turned out not to be an automatic consequence of routine school learning.

A century of research on transfer of learning has yielded a much more differentiated conception of the conditions under which various kinds of transfer are likely to occur. First of all, there are conditions of initial learning that influence the prospects of transfer: emphasis on reflective abstraction (high road), practice to near automaticity (low road), emphasis on significance and potential diverse uses (expansive versus bounded framing), fostering seeing through the surface to deeper structure, exploration of ideas before formal instruction as PFL, and so on. Classroom learning certainly can be organized to attend to such conditions. Moreover, doing so is highly synergistic with other educational goals, such as learning for understanding. Indeed, as emphasized earlier, many of the cases of seeming failure to transfer in the literature can be attributed to impoverished initial learning.

Furthermore, there are conditions for transfer that figure in moments of use – strong versus weak cuing in everyday circumstances, intrinsic or extrinsic motivation in the moment to pursue a connection, and the presence or absence of rival responses such as entrenched habits or attitudes. Classroom learning can be organized to anticipate the conditions of moments of use, for instance by alerting students to everyday clues and preparing students to deal with rival responses.

In summary, research on school learning for transfer certainly tells educators to adjust their expectations. However, it is not so much that educators need to lower their aspirations for transfer. Rather, educators need to raise their standards to styles of classroom learning supportive of transfer.

See also: Learning Theories and Educational Paradigms; Learning to Learn; Metacognitive Development: Educational Implications; School Effectiveness Research; Situated Learning (Learning In Situ); Thorndike, Edward Lee (1874–1949).

Bibliography

Bassok, M., 1990. Transfer to domain-specific problem solving procedures. Journal of Experimental Psychology: Learning Memory and Cognition 6, 522–533.

Bereiter, C., 1995. A dispositional view of transfer. In: McKeough, A., Lupart, J., Marini, A. (Eds.), Teaching for Transfer: Fostering Generalization in Learning. Erlbaum, Hillsdale, NJ.

Bransford, J., Schwartz, D., 1999. Rethinking transfer: a simple proposal with multiple implications. Review of Educational Research 24, 61–100.

Campione, J.C., Brown, A.L., Ferrara, R.A., 1982. Mental retardation and intelligence: contributions from research with retarded children. Intelligence 2, 279–304.

Chi, M. Seeing deep structure through interactions of surface features: a hypothesis for transfer. Educational Psychologist, in press.

Cook, V.J., 2003. Effects of the Second Language on the First. Multilingual Matters, Clevedon.

Cox, B.D., 1997. A rediscovery of the active learner in adaptive contexts: a developmental-historical analysis of transfer of training. Educational Psychologist 32, 41–55.

Detterman, D.R., 1993. The case for prosecution: transfer as an epiphenomenon. In: Detterman, D.K., Sternberg, R.J. (Eds.), Transfer on Trial: Intelligence, Cognition and Instruction. Ablex, Norwood, NJ.

Engle, R.A., 2006. Framing interactions to foster generative learning: a situative account of transfer in a community of learners classroom. Journal of the Learning Sciences 15, 451–498.

Gick, M.L., Holyoak, K.J., 1983. Schema induction and analogical transfer. Cognitive Psychology 15, 1–38.

Greeno, J.G., 1997. Response: on claims that answer the wrong questions. Educational Researcher 26, 5–17.

Judd, C.H., 1908. The relation of special training to general intelligence. Educational Review 36, 28–42.

Lobato, J., 2003. How design experiments can inform a rethinking of transfer and vice versa. Educational Researcher 32, 17–20.

Nisbett, R.E., Fong, G.T., Lehman, D.R., Cheng, P.W., 1993. Teaching reasoning. In: Nisbett, R.E. (Ed.), Rules for Reasoning. Erlbaum, Hillsdale, NJ.

Pea, R.D., Kurland, D.M., 1984. On the cognitive effects of learning computer programming. New Ideas in Psychology 2, 37–68.

Perkins, D.N., Salomon, G. Knowledge to go: a motivational and dispositional view of transfer. Educational Psychologist, in press.

Ross, B.H., Kennedy, P.T., 1990. Generalizing from the use of earlier examples in problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition 16, 42–55.

Salomon, G., Globerson, T., Guterman, E., 1989. The computer as a zone of proximal development: internalizing reading-related metacognitions from a reading partner. Journal of Educational Psychology 81, 620–627.

Salomon, G., Perkins, D.N., 1989. Rocky roads to transfer: rethinking mechanisms of a neglected phenomenon. Educational Psychology 24, 113–142.

Saxe, G.B., 1991. Culture and Cognitive Development: Studies in Mathematical Understanding. Erlbaum, Hillsdale, NJ.

Simon, H.A., 1980. Problem solving and education. In: Tuma, D.T., Reif, R. (Eds.), Problem Solving and Education: Issues in Teaching and Research. Erlbaum, Hillsdale, NJ.

Singley, M.K., Anderson, J.R., 1989. The Transfer of Cognitive Skill. Harvard University Press, Cambridge, MA.

Thorndike, E.L., Woodworth, R.S., 1901. The influence of improvement in one mental function upon the efficacy of other functions. Psychological Review 8, 247–261.

Wagner, J.F., 2006. Transfer in pieces. Cognition and Instruction 24, 1–71.

School Management

Brian J Caldwell, University of Melbourne, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The somewhat negative view of school management at the end of the twentieth century has given way to a more upbeat and positive account, reflecting a tighter integration of management and learning. Large-scale studies or meta-analyses of developments in different countries have led to a more nuanced approach to school management especially in high-performing systems. Significant challenges lie ahead in school management for innovation, a capacity for futures thinking, and preparation and ongoing professional development.

The article for school management in the first edition of *International Encyclopedia of the Social and Behavioral Sciences* provided a somewhat negative even depressing account of the standing of the field. 'School management is in considerable disarray, if not turmoil, and is likely to remain so in the first decade of the twenty-first century.' It argued that this was a reflection of (1) the sweeping transformations in society and their impact on schools, (2) the disjunction between educational management as a field of practice and a field of study, and (3) the weak connection between school management and learning.

An appraisal in this second edition is much more upbeat to the point that the last decade may have seen the most important advances in the study and practice of educational management since the field was established more than five decades ago. Even though the aforementioned disjunction may still be evident and the link to learning is still not as strong as it should be, the outlook for the decade ahead is more optimistic. Especially noteworthy is that developments in school management have been incorporated in a narrative of what has transpired and this is a reflection of how leading scholars who have contributed to the field over time are now able to place events in perspective. Also noteworthy is the way meta-analyses of developments in different countries have yielded dependable knowledge of approaches to successful management *for* schools and *in* schools.

There are, however, significant challenges for the decade ahead. These lie in three domains: (1) the management of innovation in schools; (2) the development of a futures perspective given the rapid changes in education, reflecting continuing sweeping changes in society; and (3) the preparation and ongoing development of those engaged in school management.

Definition of School Management

In some nations, notably the USA, administration is used to describe a range of functions performed by the most senior people in an organization, with management a more limited, sometimes routine set of tasks, often performed by those in supporting roles. The work of Kotter (1990) is helpful in resolving the issue. Leadership is concerned with change and entails the three broad processes of establishing direction, aligning people, and motivating and inspiring. Management is concerned with the achievement of outcomes, and the three broad processes of planning and budgeting, organizing and staffing, and controlling and problem solving.

There is a generally held view that the work of those who hold senior positions in schools, especially principals or headteachers, entails both leadership and management. They should certainly be leaders. In some instances, managers will support them. There is concern when they act solely as managers. Kotter (1990) offered a contingent view that is as helpful for schools as it is for organizations in general. He contends that the emphasis and balance in leadership and management are contingent on two variables: the amount of change that is required and the complexity of the operation. Complex schools in a turbulent environment require high levels of leadership and management. Schools that are relatively simple in size and structure but are faced with major change require leadership more than management. Schools that are complex in stable circumstances require management more than leadership. Relatively simple schools in stable settings may require little of each.

School Management as a Field of Practice

The three developments described in the first edition are even more evident at the start of the second decade of the twenty-first century: (1) the shift of authority and responsibility to schools in systems of public education within centrally determined frameworks of curriculum, standards, and accountabilities (school autonomy, school-based management, local management, or self-management); (2) higher expectations for schools as governments realize that the knowledge and skill of their people are now the chief resource if nations are to succeed in a global economy; and (3) change in the nature of schooling itself, driven to a large extent by advances in information and communications technology so that much of the learning that in the past could only occur in the classroom can now occur at anytime and anywhere there is access to a networked computer.

School Management as a Field of Study

School management has emerged relatively recently as a subfield of study within the broader field of educational

management (administration). There were few formal programs and little research on the phenomenon until the 1950s. Prior to that, approaches to the management of schools paralleled those in the industry. Departments of educational administration made their appearance in universities in the US in the mid-twentieth century and soon proliferated. They were established in the UK soon after, with many other nations following suit, drawing on the US and, to a lesser extent, the UK for their intellectual underpinnings. The first quarter-century, from the early 1950s to about the mid-1970s, was characterized by efforts to build theory in educational management, modeling to a large degree the discipline approach of the behavioral sciences. In the USA, *Educational Administration Quarterly* was an attempt to provide for education what *Administrative Science Quarterly* had done for administration (management). Separate subdisciplines were created in fields such as economics, finance, governance, human relations, industrial relations, law, planning, policy, and politics.

The theory movement in educational management was the subject of powerful attack in the mid-1970s, notably by the Canadian scholar T.B. Greenfield (1975). The critique was largely directed at efforts to build a scientific theory of educational management; the separation of management from learning and teaching; and fundamental considerations of context, culture, ethics, meaning, and values.

On a more pragmatic level, the critique of educational management as a field of study referred to its failure to impact practice on a large scale and to assist in a timely manner the implementation of reform and the resolution of concerns. A promising approach to unifying the field was to place learning at the heart of the effort. This means that policy interest in the improvement of student outcomes should be the driver of this 'quest for a center,' as the US scholar Joseph Murphy (1999) described it.

Much momentum was gained in the 1990s with the establishment of the International Congress for School Effectiveness and Improvement, which made its mission the bringing together of policymakers, practitioners, and researchers. *School Effectiveness and School Improvement* quickly established itself as a leading international journal.

Extending the Narrative

A narrative is important to give a sense of coherence and continuity to what at first sight may be a series of relatively disconnected or fragmented events. Strategic intentions can be communicated through a narrative in ways that detailed plans and policies often miss. Commitment and support can be marshaled if there is an attractive and credible narrative.

Murphy (1999) placed developments to the end of the twentieth century in a narrative starting from 1820 and continuing to the end of the twentieth century. He suggested that the 'center of gravity' of educational administration (management) shifted from philosophy (1820–1900), through management (1901–45) and social sciences (1945–85) to school improvement (1986–99). The foundation moved during these time frames from religion through administrative functions and academic disciplines to education. The 'engine' changed from values through practical knowledge and theoretical knowledge to applied knowledge. The dialectic from 1986 continues to the present with a more robust consensus on this reading of developments than at the time of Murphy's account.

The narrative has been extended in recent years but in a way that reflects a stronger link between management and learning. This is illustrated in the identification by Hargreaves and Shirley (2009) of approaches to leading and managing change over more than four decades. Hargreaves has had a major interest in teaching, teacher professionalism, school leadership, and change over several decades and has offered a powerful critique of aspects of educational reform from time to time. They described four 'ways' or broad approaches to change (Hargreaves and Shirley, 2009, p. 44), the first three of which were (1) from the end of World War II to the mid-1970s characterized by innovation and inconsistency, with an 'interregnum' from the mid-1970s to the late 1980s to a situation that involved a high level of complexity and contradiction; (2) from the late 1980s to the mid-1990s with an emphasis on markets and standardization; and (3) from the late 1990s to 2009 that shifted the focus to performance and partnerships. Each of the three ways is analyzed in a number of domains including control, purpose, trust, community engagement, curriculum, teaching and learning, professionalism, professional learning communities, assessment and accountability, and lateral relations. For example, for teaching and learning, there have been different emphases over time, with developments being eclectic and uneven (first way), prescriptively driven by standards and testing (interregnum), direct instruction to meet standards and test requirements (second way), and autocratically data driven yet customized approaches (third way).

The 'fourth way' that is emerging in the second decade of the twenty-first century, and advocated by Hargreaves and Shirley, calls for the retention of aspects of the first three ways and abandonment of others. For example, moving from the third way to the fourth should see the retention of balance and inclusiveness, public involvement, financial reinvestment, better evidence and professional networks, and the abandonment of 'persistent autocracy,' imposed targets, obsession with data, and 'effervescent interactions.' There are implications for each of the four ways for school management on two dimensions: management *for* schools and management *in* schools.

While narratives along these lines are helpful, and do indeed give a sense of coherence and continuity in school management, there have been further developments since the turn of the century that suggest a more nuanced approach to description and advocacy in different settings, and these emerged from an unprecedented series of international studies, the findings of which have been synthesized through meta-analyses.

Meta-Analyses of School Management in Different International Settings

A major development in the first decade of the twenty-first century is the large-scale meta-analyses of test results and other indicators of high performance in many countries, the most prominent of which have been conducted by the Organization for Economic Cooperation and Development (OECD) and the

international consultancy McKinsey & Company. The findings and their implications for school management are summarized below.

Most systems of public education around the world have included decentralization of decision-making in their plans for school reform or restructuring. There has also been significant centralization. The OECD explained it in these terms in summarizing the findings in its annual report in 2004:

> Placing more decision-making authority at lower levels of the educational system has been a key aim in educational restructuring and systemic reform in many countries since the early 1980s. Yet, simultaneously, there have been frequent examples of strengthening the influence of central authorities in some areas.
>
> OECD, 2004, p. 34

The OECD report described the locus and mode of decision-making in four domains. Locus referred to which of six levels decisions were made: national, state, regional, municipal, local, or school. Mode referred to which of four ways decisions were made: full autonomy at the level concerned, consultation with other bodies at that level, independently but within a framework set by a higher authority, or other. The four domains were organization of instruction, personnel management, planning and structures, and resources. There have been important changes to school management as a result of these trends. While there are different patterns among the nations, it is evident that those immediately concerned with the management of schools have needed a heightened capacity in each of these domains.

Findings from analyses of test results in Program for International Student Assessment (PISA) conducted by OECD suggest that the most successful systems of schools secure an optimal balance of autonomy, accountability, and choice. A model was constructed to explain the joint impact of school and system resources, practices, and policies on student performance. Of the 15 factors in the model, the system average on the school autonomy index in budgeting is by far the most powerful (OECD, 2007). Particularly striking are two studies conducted for OECD by the Ifo Institute for Economic Research at the University of Munich (Department of Human Capital and Innovation), one focusing on the level of student achievement and the other on equity of student achievement. On level of student achievement, the following findings are striking:

> On average, students perform better if schools have autonomy to decide on staffing and to hire their own teachers, while student achievement is lower when schools have autonomy in areas with large scope for opportunistic behavior, such as formulating their own budget.
>
> Wößmann et al., 2007, p. 59

> Students perform substantially better in systems where private school operation creates choice and competition. At the same time, student achievement increases along with government funding of schools. A level playing field in terms of government funding for public and private schools proves significantly performance enhancing.
>
> Wößmann et al., 2007, pp. 59–60

As far as equity is concerned:

> [R]ather than harming disadvantaged students, accountability, autonomy, and choice are tides that lift all the boats … there is not a single case where a policy designed to introduce accountability, autonomy, or choice into schooling benefits high-SES [socio-economic status] students to the detriment of low-SES students.
>
> Schütz et al., 2007, pp. 34, 35

There is tension between these findings and the narrative offered by Hargreaves and Shirley as summarized earlier, especially with respect to competition. However, there are important implications for those concerned with school management, who are operating or may need to operate in a milieu of relatively high levels of autonomy, accountability, and choice.

Three reports of McKinsey & Company shed light on developments in school management in different countries with a particular focus on approaches associated with high performance. The first (Barber and Mourshed, 2007) drew on practice in several countries and school systems to explain how the world's 'best performing systems come out on top.' The report drew extensively on practice in Finland that had attracted headlines for its consistently high performance in PISA but also other high performers including Alberta (Canada), Australia, Belgium, Finland, Hong Kong, Japan, Netherlands, New Zealand, Ontario (Canada), Singapore, and South Korea. Jurisdictions with a strong improvement strategy included the following in the United States: Atlanta, Boston, Chicago, New York, and Ohio. The main finding was concerned with the quality of teaching, with the key themes being: getting the right people to become teachers, developing effective instructors, and ensuring every student performs well. Indicators of 'best practice' in the second theme illustrated the relevance of the study as far as school management is concerned. Key issues for this second theme are illustrated in the following, with indicators included in parentheses: coaching for new teachers (at least 20 weeks), time spent in professional development (10% of working time), role of school leaders in effective instruction (best coaches are selected as leaders), and school-based research on effective practice (research budget of USD50 per student). These findings suggest that school leaders have an important role in the management of instruction in high-performing school systems. The concept of 'talent management' has emerged in recent literature (for example, Davies and Davies, 2011).

The findings in another report of McKinsey & Company (Mourshed et al., 2010) highlighted the need for a nuanced approach to school management depending on the setting. In this instance, the study was concerned with school systems that were already high performing and were getting better and others that were starting from a low base and were showing noteworthy improvement. Examples of the former included Singapore, Hong Kong, and South Korea. The latter included Western Cape in South Africa, Minas Gerais in Brazil, and Madhya Pradesh in India. There were 18 countries in the sample (a total of 20 systems including three from the United States). The 20 systems were classified on a robust evidence base in four 'journeys': poor to fair ('achieving the basics of

literacy and numeracy'), fair to good ('getting the foundations in place'), good to great ('shaping the profession'), and great to excellent ('improving through peers and innovation'). There were three main findings:

1. It's a system thing, not a single thing: There is a common pattern in the interventions improving systems use to move from one performance stage to the next, irrespective of geography, time, or culture.
2. Prescribe adequacy, unleash greatness: There is a strong correlation between a school system's improvement journey stage and the tightness of central control over the individual schools' activities and performance.
3. Common but different: Our findings indicate that six interventions occur with equal frequency across all the improvement journeys, though manifesting differently in each one (Mourshed et al., 2010, pp. 33, 34).

The second finding ('prescribe adequacy, unleash greatness') is especially applicable for the journey from good to great wherein only loose, central guidelines for teaching and learning processes are provided 'in order to encourage peer-led creativity and innovation inside schools, the core driver for raising performance at this stage.' The 'intervention cluster' for the journey from great to excellent includes cultivating peer-led learning for teachers and principals, creating additional support mechanisms for professionals, and system-sponsored experimentation/innovation across schools (Mourshed et al., 2010, p. 36).

The report provides examples of interventions for each of the three findings. For the first finding ('it's a system thing, not a single thing'), interventions in the journey from great to excellent include collaborative practice among educators, decentralizing pedagogical rights to schools and teachers, creating rotation and secondment programs across schools and between center and schools, providing additional administrative staff, sharing innovation from the front line, and funding for innovation (p. 51). For the second finding ('prescribe adequacy, unleash greatness'), the report concluded that "lower-performing systems focus on raising the floor, while higher performing ones focus on opening up the ceiling" (p. 52). For the third finding ('common but different'), particular attention is given in the journey from great to excellent to 'attracting top talent,' with teachers' base salary significantly above per capita GDP (p. 62).

These findings reinforce the theme that there is no one best way as far as school management is concerned. Those schools starting from a low base but showing rapid improvement such as those in Brazil, India, and South Africa will focus on a different set of tasks and have different priorities than those that are high performing and getting better, as in Hong Kong, Singapore, and South Korea. School managers in the former are on a journey "characterised by lower skill educators [and] exercise tight, central control over teaching and learning processes in order to minimise the degree of variation between individual classes and across schools." School management in the latter is conducted in contexts that "provide only loose, central guidelines for teaching and learning processes, in order to encourage peer led creativity and innovation inside schools, the core driver for raising performance at this stage." As noted above in the context of the OECD studies, there is a tension in these more nuanced approaches with the 'fourth way' described and advocated by Hargreaves and Shirley (2009).

A third report of McKinsey & Company provided a sharper focus on leadership (Barber et al., n.d.). It contained a review of developments to 2010 in eight jurisdictions: Alberta (Canada), England, Netherlands, New York, New Zealand, Ontario (Canada), Singapore, and Victoria (Australia). The authors claim that "all perform strongly on international tests, or their performance in this respect is improving, and they demonstrate good practice in school leadership" (p. 4). The report drew on a literature review; interviews with experts, policymakers, and leaders in each jurisdiction; and a survey of 1850 leaders. While the broad findings and the specific illustrations stand up well to critical scrutiny, the limitations of meta-analyses on this scale are evident. For example, the claim of strong or improving performance may be challenged in the cases of England and New York, although examples of outstanding practice can be found in each system. Broad-based generalizations about policy and practice cannot be made across a jurisdiction, a point that was made well in the second McKinsey report cited above (Mourshed et al., 2010).

The report cited the view of an official in Singapore who stated that "one of the key revelations over the past ten years is that school leadership is not just an HR issue – it is a strategic issue" (p. 5). Reference was made to an analysis of practice in England that suggests that:

> [T]he overall performance of a school almost never exceeds the quality of its leadership and management. For every 100 schools that have good leadership and management, 93 will have good standards of student achievement. For every 100 schools that do not have good leadership and management, only one will have good standards of achievement.
>
> Barber et al., n.d., p. 5

The report lists evidence-based practices that are common across jurisdictions in the study, with "clear evidence in the literature that developing teachers makes the biggest contribution to student learning outcomes" (p. 6).

These meta-analyses continued at the start of the second decade of the twenty-first century. For example, the Grattan Institute (Jensen et al., 2012) in Australia took a similar approach to McKinsey & Company but focused on high-performing systems in East Asia: China Hong Kong, China Shanghai, Singapore, and South Korea. The authors found that stereotypes of East Asian education don't stand up to critical scrutiny in explaining the success of these countries; for example, Confucian culture and the influence of 'tiger mums.' While many may challenge this contention and refer also to differences in terms of the longer school day and special tutoring and coaching, the central theme is difficult to challenge, namely, the quality of teaching. These findings are consistent with those of McKinsey & Company (Barber and Mourshed, 2007). A focus of school management in these countries is on building the capacity of teachers through the preparation and selection of outstanding applicants and devoting substantial amounts of time to school-based professional development and research.

What may be the most comprehensive of meta-analyses was conducted by Hattie (2009), who synthesized more than 800 meta-analyses to specify the effects of various interventions on student achievement. He was able to rank order the effects of 138 interventions. He drew conclusions from the literature on management to identify practices with positive effects, the most successful being goal setting and team building and the last successful being 'techno-structural' interventions such as job redesign and job enrichment (Hattie, 2009, p. 85). Overall, the effects of school leadership and school management are indirect rather than direct, that is, they involve the creation of conditions under which these 138 interventions may be established and made effective.

Innovation and a Futures Perspective

The second McKinsey & Company study cited above (Mourshed et al., 2010) found that a capacity for innovation was a feature of management in high-performing systems that 'keep getting better.' There is therefore a need for leaders in schools and school systems to have a capacity for innovation. However, a case can be made that such a capacity should be nurtured in all or most settings, even those who are still on the journey from 'poor' to fair' since the context for schooling is changing in dramatic fashion. Notwithstanding the different stages in the evolution of approaches to school leadership (Murphy, 1999) or the four ways of change (Hargreaves and Shirley, 2009), the broader social and economic context was relatively stable for decades. That is not the case now, especially with advances in technology and the challenge of aligning education, society, and economy in an era of globalization. The prognosis for school management in the decade ahead is likely to include an expanded role for innovation and a 'futures perspective.'

The Merriam Webster Dictionary defines innovation as 'the introduction of something new' or 'a new idea, method, or device.' However, in education as in other settings, what might be an innovation in one school may be routine in another. A related issue is the connection between innovation, change, and reform. Innovation and reform are intersecting subsets in the broader domain of change, as illustrated in **Figure 1**. Expressed simply, not all reform is innovation; not all innovation is reform; reform and innovation are examples of change. Dealing with these relationships in different contexts is evident in school management.

In many systems of education, innovation is the responsibility of a central unit, with 'successful' innovations then scaled up for implementation across all schools. Increasingly in some settings, especially those on the journey from good to great or great to excellent (in the formulation of the second McKinsey & Company study reported above), a capacity for innovation is becoming embedded in schools themselves.

A major contribution was made by Christensen et al. (2008) with their concept of 'disruptive change' that has powerful implications for school management. Every public school principal in Victoria, Australia – about 1600 – was provided with a copy. Dyer et al. (2011) built on the concept to describe 'the innovator's DNA.' Their work was an outcome of research of over 8 years in which leaders of the most innovative companies were interviewed to determine their personal attributes. The questions were penetrating to the extent that these attributes are quite fundamental in a personal sense, hence the idea of 'the innovators DNA.' The companies were private profit-making enterprises but the authors made clear that the findings are more broadly applicable. There is no reason to believe that they are not pertinent to innovation in schools, with the authors identifying the following as the core skills: (1) associational thinking – innovative thinkers connect fields, problems, or ideas that others find unrelated; (2) questioning – innovators are consummate questioners who show a passion for inquiry; (3) observing – innovators are intense observers who gain insights into and ideas for new ways of doing things; (4) networking – innovators spend a lot of time and energy finding and testing ideas through a diverse network of individuals with different backgrounds and perspectives; and (5) experimenting – innovators are constantly trying out new experiences and piloting new ideas. The following are implications for school management: (1) every leader in every unit should have these five capacities and position descriptions and selection criteria should be changed accordingly; (2) particular attention is required to the appointment of principals and, within schools, to the appointment of staff; (3) mentors and coaches should have high levels of skill in these areas – coaches and mentors who help others reproduce the status quo may have limited value; (4) changing the culture of schools will take time; and (5) nothing in the findings and these implications suggest that a balance of leadership and management is unimportant – not every aspect of a school's operations involves innovation.

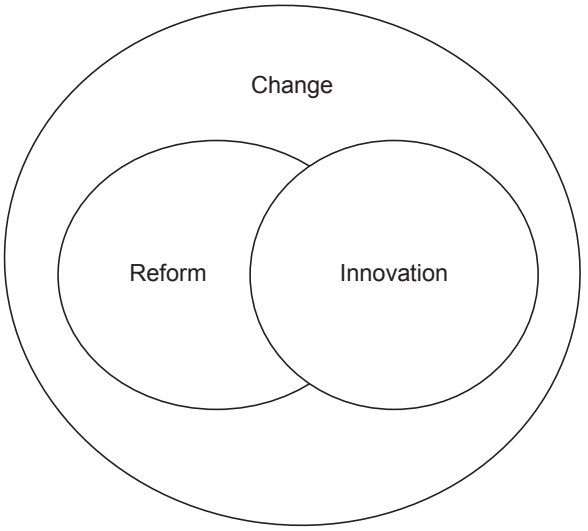

Figure 1 The relationship between change, reform, and innovation.

Leadbeater concluded that "innovation in education should take place simultaneously – at different levels and in different settings from the daily practice of teachers and learners – through organizations, systems and platforms, to the social movements and the ideologies that inspire them" (Leadbeater, 2011, p. 4).

Complementing the need for deeply embedding a capacity for innovation in school management is the importance of

gaining a 'futures perspective.' Drawing on earlier work by Mintzberg (1995); Caldwell and Loader, 2010) adopted the metaphor of 'seeing' in describing a futures-focused school. A futures-focused school 'sees ahead,' but it also 'sees behind,' honoring and extending its accomplishments in the past. It 'sees above' in the sense of understanding the policy context. It 'sees below,' demonstrating a deep understanding of the needs, interests, motivations, and aspirations of students and staff. It 'sees beside' by networking professional knowledge to take account of best practice in other schools in similar settings. It 'sees beyond' by seeking out best practice in other nations and in fields other than education. It is consistent and persistent; it 'sees it through.' The metaphor of 'sensing' is also helpful given that 'seeing' refers to what is already in place or is projected. A futures-focused school is alert to signals in its internal and external environment that may influence what may occur in the future and that may subsequently be 'seen.' These signals may be strong or weak and a high level of sensitivity is required to distinguish among them (Caldwell and Loader, 2010, p. 20; see also Caldwell et al., 2012, pp. 411, 412). The balance of leadership and management is weighted in favor of the former in the capacities required to create a futures-focused school. However, consistent with the relationships and distinctions provided by Kotter (1990), set out at the start of this article, there is a management component. Structures and processes must be established; personnel must have the required capacities; and ongoing evaluation and change are demanded.

Preparation and Ongoing Development of Those Engaged in School Management

There have been significant, if not dramatic, changes in approaches to preparing and developing those engaged in school management. These reflect the narratives provided earlier (Murphy, 1999; Hargreaves and Shirley, 2009). In summary, programs for preparation and ongoing development at the end of the twentieth century tended to be isolated from the core work of schools, namely, learning and teaching and the support of learning and teaching. As far as graduate programs in educational administration (management) were concerned, subjects like change, finance, human resource management, leadership, organizational development, planning, policy, and supervision were in vogue. Participants were expected to apply generic knowledge to areas of professional interest. Studies in curriculum and pedagogy were rarely included. There was steady change throughout the first decade of the twenty-first century, the chief feature of which has been the inclusion of studies in curriculum and pedagogy and building a capacity to analyze and utilize data on student achievement. Most school systems have established centers for leadership development and these often involve stronger partnerships between universities, school systems, and schools than existed in the past. These developments are likely to continue.

See also: Classroom Level in Educational Effectiveness Research; Educational Effectiveness Theory and Research: Recent Advances; Leadership Education: Theory and Practice; School Improvement; Teacher Leadership.

Bibliography

Barber, M., Mourshed, M., 2007. How the World's Best-Performing School Systems Come Out on Top. McKinsey & Company, London.

Barber, M., Whelan, F., Clark, M., n.d. Capturing the Leadership Premium. McKinsey & Company, London.

Caldwell, B.J., Loader, D.N., 2010. Our School Our Future. Education Services Australia, Melbourne.

Caldwell, B., Loader, D., Harris, J., Zhao, Y., 2012. Capital formation in the futures focused school: Indicators of a breakthrough in school improvement. In: Day, C. (Ed.), The Routledge International Handbook of Teacher and School Development. Routledge, London and New York, pp. 410–420 (Chapter 32).

Christensen, C., Johnson, C.W., Horn, M.B., 2008. Disrupting Class. McGraw-Hill, New York.

Caldwell, B.J., Harris, J., 2008. Why Not the Best Schools? ACER Press, Melbourne.

Davies, B., Davies, B.J., 2011. Talent Management in Education. SAGE, London.

Day, C. (Ed.), 2012. The Routledge International Handbook of Teacher and School Development. Routledge, London and New York.

Dyer, J., Gregersen, H., Christensen, C.M., 2011. The Innovator's DNA: Mastering the Five Skills of Disruptive Innovators. Harvard Business Review Press, Boston.

Greenfield, T.B., 1975. Theory about organizations: a new perspective and implications for schools. In: Hughes, M.G. (Ed.), Administering Education: International Challenge. Athlone, London, pp. 71–99.

Hargreaves, A., Shirley, D., 2009. The Fourth Way: The Inspiring Future for Educational Change. Hawker Brownlow Education, Melbourne (originally published by Corwin Press).

Hattie, J., 2009. Visible Learning. Routledge, London and New York.

Jensen, B., Hunter, A., Sonnemann, J., Burns, T., 2012. Catching Up: Learning from the Best School Systems in East Asia. Grattan Institute, Melbourne.

Kotter, J.P., 1990. A Force for Change: How Leadership Differs from Management. Free Press, New York.

Leadbeater, C., 2011. Rethinking Innovation in Education: Opening up the Debate. No. 207 in the Seminar Series. Center for Strategic Education, Melbourne.

Mourshed, M., Chijioke, C., Barber, M., 2010. How the World's Most Improved School Systems Keep Getting Better. McKinsey & Company, London.

Murphy, J., 1999. The quest for a center: notes on the state of the profession of educational leadership. In: Invited Paper at the Annual Meeting of the American Educational Research Association, Montreal, April.

Mintzberg, H., 1995. Strategic thinking as 'seeing'. In: Garratt, B. (Ed.), Developing Strategic Thought: Rediscovering the Art of Direction-Giving. McGraw-Hill, London (Chapter 5).

OECD, 2004. Education at a Glance. OECD, Paris.

OECD, 2007. PISA 2006 Science Competencies for Tomorrow's World. OECD, Paris.

Schütz, G., Wößmann, L., West, M.R., 2007. School accountability, autonomy, choice, and level of student achievement: international evidence from PISA 2003. In: Education Working Paper No. 14 (Directorate of Education). OECD, Paris.

Wößmann, L., Lüdemann, E., Schütz, G., West, M.R., 2007. School Accountability, Autonomy, Choice, and the Equity of Student Achievement: international evidence from PISA 2003. Education Working Paper No. 13 (Directorate of Education). OECD, Paris.

School Safety

Thomas J Mowen, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The safety of schools in the United States has always been a concern to policy makers, teachers, parents, and students. However, over the past three decades, there has been an increase in the concern over school safety given the highly publicized shootings of the 1990s and 2000s. As a result, schools are employing more security measures to ensure the safety of schools, and policy makers and school administrators are implementing more and more policies aimed at maintaining safe schools.

Introduction

Schools in the United States are often considered safety zones for youth and young adults. While neighborhoods, certain areas of town, or even the home environment may be potential places of violence for children, schools have traditionally been a safe haven for all youth in the minds of most Americans. The reality is that all schools must deal with issues of violence and threats against the school or members of the school. Because of this, all schools must also deal with how to maintain a safe school environment. Certainly the degree to which a school must consider threats varies due to myriad factors. For example, the way in which a suburban school secures its' hallways is likely very different than the means employed by an urban or rural school. Overall, however, the safety of schools in the United States has occupied an important position in educational policy since the introduction of the formal educational system in the United States.

School safety exists, of course, in response to perceived threats to the well-being of the educational institution, as well as the well-being of the individual members of the institution including students, staff, administration, family members, and anyone within the school or involved in school business. Because of this, school safety can take on many forms and many meanings; it changes over time and between spaces. For example, the instruments of school safety used in modern schools are very different than the instruments used in schools of the 1930s and 1940s. Similarly, school policies and practices aimed at protecting the school have also changed over time. This is due, in large part, because the concern of school safety; that is, what the safety is guarding against, has shifted drastically over the past half century.

Defining School Safety Over Time

School safety is comprised of two primary areas of concern: (1) threats to the school from external forces such as individuals who may seek to burglarize the school, for example, or threaten the safety of students and staff members; and (2) threats to the school from internal sources such as student misbehavior, teacher, and staff misconduct, or other crimes committed from someone within the school. As subsequent sections will highlight, the former area of concern has received significantly more attention than the latter. Given this duality of concern, however, it is important to consider how defining and understanding school safety is bound within time.

The goals and instruments of school safety – like most anything – have evolved over time. Documents reveal that in the 1930s and 1940s, the concern of school safety was aimed at discouraging children from speaking out of turn, or encouraging students to dispose of their trash in a proper manner. There is little evidence that school officials were considered with any school safety policy aimed at deterring threats outside of the school; the focus was aimed exclusively at reducing minor misbehavior by students (Stouffer, 1952).

Moving into the 1960s and 1970s, highly publicized events such as the Civil Rights Movement, and other national-level events such as the Vietnam War highlighted the importance of safety in the nation. As a result, schools began to target more aggressive forms of problem behavior such as student–teacher disobedience, and physical confrontations between students. In addition, school policies aimed at reducing drug use and gang activity began to emerge within the school security discourse.

Transitioning into the 1980s and early 1990s, as the United States Department of Justice intensified the war on crime, schools and policy makers became concerned with guns, gang violence, and drug violence. Because of this rise in concern, zero tolerances policies (outlined in subsequent sections) quickly gained ascendancy into the school safety conversation, and have continued to dominate educational policy.

The 2000s and 2010s have seen a number of highly sensationalized school shootings including Columbine, Virginia Tech, and more recently, the shootings at an elementary school in Newtown, Connecticut. Such events changed the way in which school safety is regarded. These recent events have resulted in a drastic rise in concern over safety in the hallways of American schools (see Casella, 2006). Because of this, researchers, policy makers, and parents have turned their attention to understanding what makes schools safe, and how can threats to the school be neutralized or avoided.

The Rise of Zero Tolerance

Prompted by the increased awareness and attention given to the threat of gangs and gun violence in the hallways of American schools, the late 1980s saw the ascendency of zero tolerance policies within American schools. Though initially a term

used by the Drug Enforcement Agency regarding illegal shipping containers in the mid-1980s, zero tolerance school policies quickly gained rapid ascension beginning in the late 1980s. By 1989, nearly half of all state school systems had adopted a zero tolerance perceptive on drugs and gangs (Phaneuf, 2009), and by 1990, nearly 80% of all public school systems in the United States employed some level of zero tolerance.

As previously highlighted, zero tolerance is the primary method of employing discipline and punishment on students within the school. Zero tolerance refers to giving no leniency for particular crimes; that is, punishments become required at a minimum for particular offenses. The goal of zero tolerance is twofold. First, it sends a clear message to students that particular offenses will not be tolerated. Second, it is aimed to ensure that all punishments given by an administrator are uniform based on the zero tolerance policy. Thus, it encompasses both a deterrent effect, while providing uniformity in discipline, creating a sense of equality among diverse student bodies. Because of this, forms of zero tolerance exist in nearly all schools across the United States.

In 1994, with the passage of the Gun-Free Schools Act, public schools in the United States became legally required to expel students from school for no less than 1 year if they were found with a gun on school grounds. The law, as it was originally drafted by the United States Congress, was narrowly focused on dangerous weapon possessions – such as guns and explosives – within the school proper (Sughrue, 2003). However, by 1995, most schools had voluntarily expanded the scope of this bill to include other weapons such as daggers, knives, brass knuckles, and lighters. As other scholars highlight, many schools also voluntarily expanded zero tolerance to include destruction of property, nonviolent repeated defiance of school authority (Insley, 2001), possession of over-the-counter pain relievers, toy guns, and classroom disruption (Schwartz and Len, 2001).

As noted above, there are some levels of zero tolerance that are mandated by law. For example, it is likely that if a student were found with a gun in any school, they would face a minimum sentence of a 1-year expulsion and also face legal prosecution. However, not all schools have zero tolerance for more minor behavior such as repeated defiance of school authority or destruction of school property. Thus, while most schools in the United States employ some level of zero tolerance, but there are important variations. Public schools are much more likely to employ a greater number of zero tolerance policies than private schools. High school and middle schools are more likely to use a greater degree of zero tolerance than elementary schools. Finally, urban schools are more likely to use zero tolerance policies than rural or suburban schools.

Columbine Aftermath

On 20 April 1999, two students at Columbine High School in Colorado committed suicide, but not before killing 12 students, one teacher, and injuring 21 other individuals during a premeditated takeover of the high school. Perhaps more so than any other single event in the history of schooling in the United States, this event reverberated throughout the nation. In turn, this prompted lawmakers, administrators, and policy makers to review the safety of schools in the United States. The outcome of this was the Safe School Initiative (SSI). The goal of the SSI was to explore past violence in schools, and to implement a series of best practices for schools to engage in order to reduce the risk of school shootings in the future.

The SSI, a collaboration between the National Threat Assessment Center, the US Department of Education, and the US Secret Service is one of the most ambitious and lengthy studies on school safety in the Unites States to date. The study examined 37 lethal shootings involving 41 attackers that occurred in schools between 1974 and 2000. The research teams examined school records, school security instruments present, demographic information, and other pertinent details that shed light on these acts of school violence. The final report highlighted 10 major findings, which are presented below along with additional comments concerning the outcome of each point.

1. Incidents of targeted violence at school rarely are sudden impulsive acts.

It is important to note that although a violent act occurring within the school is not necessarily impulsive, the time between when the attacker decides to carry out violence against the school and the execution of the plan may not be long enough for proper intervention. At the same time, however, the results of the SSI reveal that most of the incidences involved detailed planning of the act before execution. This finding suggests that if 'at risk' students can be identified, it may be possible to provide resources to the student and the student's family that could potentially result in a different outcome.

2. Prior to most incidents, other people knew about the attacker's idea and/or plan to attack. In most cases, those who knew were other kids–friends, schoolmates, siblings, and others. However, this information rarely made its way to an adult.

This key is especially important as it suggests that other students or friends may be in a position to intervene, or pass important information onto an adult. The SSI suggests that schools should ensure they have a fair way of responding to concerns of students. Thus, school safety policy should encourage communication between students and adults, and provide an avenue for the students to reach out without placing themselves at risk.

3. Most attackers did not threaten their targets directly prior to advancing the attack.

According to the SSI, this point underscores the importance of acting on any information that may indicate a problematic situation; not simply waiting for a threat to materialize. Moreover, the report concludes that school administrators should respond to any threat made by students, regardless of how small.

4. There is no accurate or useful 'profile' of students who engaged in targeted school violence.

Such findings indicate that profiling a student based on sociodemographic factors, lifestyle choices, such as clothing or musical tastes, are not useful methods in predicting whether or not a student who may become violent. This is especially

relevant as one of the outcomes of Columbine was many parents and policy makers pointed to Marylyn Manson and the 'goth' lifestyle as inspiring the violence that occurred. Likewise, the shooting in Newton, CT, has inspired great debate over the prevalence and influence of violent video games. Yet, research has yet to establish any 'profile' of students who may engage in school shootings, and the SSI underscores that this would be a fruitless endeavor.

5. Most attackers engaged in some behavior prior to the incident that caused others concern or indicated a need for help.

Overall, this recommendation suggests that a truly safe school cannot simply surround itself with defense objects such as metal detectors and police officers (discussed in subsequent sections) to prevent school tragedies and increase school safety. While security instruments play an important role in preventing school shootings, school officials must be trained to recognize calls for help from at risk students. Moreover, this finding suggests that most troubled youth do show some visible signs of distress that can be recognized. One of the tangible outcomes of this report was an increase in school counselors and school psychologists available to students in many schools.

6. Most attackers had difficulty coping with significant loses or personal failures. Moreover, many had considered or attempted suicide.

It is also important to highlight – and the SSI does – that most students who experience a loss in life will not become attackers within their school. However, the majority of cases investigated by the DoE experienced personal failures or loses. This finding suggests that school staff and administration should pay attention to students experiencing hardship, not simply because they 'may' become a danger to the school, but because the school is viewed as a safe haven (see also Dwyer et al., 1998). A closer integration between the private life of the student, and the students' school life may be an effective way to reduce school shootings and increase the safety of the school.

7. Many attackers felt bullied, persecuted, or injured by others prior to the attack.

As with the above point, not all students who are bullied are problematic, but the vast majority of cases investigated had students who experienced bullying and were often classified as social outcasts by other students, teachers, and staff. Moreover, bullying has recently received a significant attention at the national level. Recently, increased legislative attention to bullying has resulted in the government launching specific programs designed to reduce bullying within the school including the launch of a government website aimed specifically at reducing bullying within schools (United States Department of Health and Human Services, 2013).

8. Most attackers had access to and had used weapons prior to the attack.

The issue of gun control falls outside of the scope of the current article, but the implications of this finding suggests that if it were more difficult for students to gain access to guns it *may* be possible that difficulty could serve as a deterrent. While students could turn to other forms of weapons, such as knives, it is likely that within access to a gun, the number of lives lost due to an attack within the school would be much less. Such thoughts are gaining popularity as shown by the passage of the strictest gun control law in the nation in Connecticut, the home state of the school shooting in Newton.

9. In many cases, other students were involved in some capacity.

Much like point 2, this finding suggests that friends and other students may be able to provide feedback to troubled individuals in deterring them from engaging in an attack. At the same time, however, it also suggests that students may not have believed that the individual would carry out an attack. Because of this, school officials should encourage students to come forward with any concern, no matter how small.

10. Despite prompt law enforcement responses, most shooting incidents were stopped by means other than law enforcement intervention.

Again, this highlights the need for schools to focus not just on defensive technologies, but also on training of school staff to recognize potential situations, and to defuse situations and seek treatment for the troubled individuals. It also suggests that an increase in security guards may not be enough to prevent school shootings.

Overall, the conclusions of the SSI suggest that many of the attacks may have been preventable. The Department of Education concludes that schools should actively, and continually, engage in threat assessment and develop strategies to prevent potential attacks on the school from occurring. In large part, this document has resulted in schools adopting greater numbers and types of school security measures and policy.

Instruments of School Safety

As highlighted by the SSI, one of the major takeaways was the implementation and use of school security measures. As of 2010, the most currently available data, nearly 98% of all schools in the Unites States control access to their grounds through check in areas. Approximately 80% of schools maintain a closed campus during the entire school day, requiring students to stay within the school and school groups; 21% of all schools utilize metal detectors at the entrance of the school; 69% utilize random drug sweeps on the student bottom to search for contraband; nearly 75% utilize security cameras or closed-circuit televisions; and approximately 55% of all schools utilize a police or school resource officers (U.S. Department of Education, 2012). The overall goal of these measures is to increase the safety for students, teachers, parents, and other stakeholders within the school.

It's important to highlight that the types of school safety instruments used vary by region, school type, and school size. Larger schools are more likely than smaller schools to utilize any instrument of school security measure outlined above. This is due, in part, to the budget of the school, as well as the location. Similarly, public schools – which tend to be larger than private schools – tend to employ more instruments of school security. Urban schools, as opposed to rural and

Table 1 US suspension and expulsion rates in schools K-12

	1974	*1998*	*2000*	*2002*	*2004*	*2006*
Suspensions						
Number of suspensions	1 700 000	3 200 000	3 053 449	3 083 810	3 279 745	3 328 754
Percent of suspensions	3.7	6.8	6.5	6.6	6.8	6.9
Expulsion						
Number of expulsions			97 177	89 131	106 222	102 077
Percent of expulsions			0.3	0.2	0.2	0.2

suburban schools, are the most likely to utilize instruments of school security. Finally, schools in the Southern United States are more likely to use instruments of school security compared to schools in the Midwest or Northeast regions of the United States.

In addition to the use of school security instruments, school administrators and policy makers also rely on the use of detentions, suspensions, and in the most severe cases, expulsions as a means to regulate the behavior of students. As of 2006, approximately 6.9% of all students in the United States had received a suspension at some point over the course of their schooling. While data are not easily accessible, this rate has increased from 3.7% of all students in 1974. At the same time, expulsion rates have remained relatively stable since 2000 with approximately 0.2% of all students having been subject to expulsion has a form of punishment. Please see **Table 1**.

School Safety, Drugs, and Gangs

The major impact of the Columbine incident is only part of the story in explaining how policy and procedures have been shaped in order to keep schools a safe place. While school shootings have become a more pressing concern in recent years, one major concern of school safety efforts is focused on reducing the availability of drugs and illegal substances within the school. This concern gained ascendency in the early 1980s as the United States moved into the period of War on Drugs, but it is important to note that there have been many fluctuations in the amount of drugs within American schools, and there have been many changes in policy to address drug use and school safety.

A University of Michigan-based report entitled Monitoring the Future (MTF) has studied drug use within schools since 1975. That year, the study concluded that approximately 55% of all students in the United States had used an illicit substance prior to graduating high school. While this does not mean that illicit substances were used within the school, but studies such as MTF find that most youth who engage in illegal substance abuse generally procure the means to get the drug through the school. Moving forward, by 1981, this number increased to 66%, but soon after began to decline finally setting on a low of 41% in 1991 before rising to 55% in 1999. In recent years, this number has dropped to approximately 44% by 2006 where it has remained relatively stable since (MTF, 2012). MTF reports that marijuana is the most used drug by students, and most students who consume marijuana receive the drug through contacts within the school.

Such findings highlight the important role school security measures and school safety policy play in the reducing of the availability of drugs to students in school. At the same time, however, it is not possible for security measures/policy to effectively clear drugs from schools. It is interesting to note that as schools began to use more security measures (see the above section), the number of students reporting drugs dropped. Although that number has bounced between 40 and 55% over the last two decades, it is interesting to note that the number of students reporting drugs has never reached 'pre'modern security measure and zero tolerance numbers.

The second major concern of school safety – also growing out the 1980s – is that of gangs and violence within schools. While research is mixed on the extent to which gangs operate within the school, gang activity is a major concern. For example, research has found that most gang members are not interested in attending school, and gang activity is often regulated by zero tolerance. Because of this, students who are members of a gang are often faced with expulsion. Individuals that support this view often show that gangs, overall, do not operate within schools to a significant degree compared to their presence outside of the school.

At the same time, no one denies that gang activity, when it is present, is problematic within schools, and school security measures and policy must be implemented to dissuade gangs from operating within schools. As a recent report by the Bureau of Justice Assistance highlights, youth – including youth who are involved with gangs – have a legal right to an education from public schools in the United States and because of this, gang activity will continue to persist (U.S. Department of Justice, 1999). Therefore, schools continue to take action and implement school security policies to deter gang activity.

Side Effects of School Safety Measures

While it is important to assess the relative success of school safety on reducing the presence of drugs, gangs, and violence in schools, there have also been a number of studies that have assessed other outcomes of school safety, namely, the effect of school safety on students. Recent works have found a strong racial imbalance in the application of punishment (Skiba et al., 2000) with Black and Hispanic students significantly more likely to be expelled or suspended compared to White students for similar misbehaviors. In addition, Black males are

more likely to be punished for minor misbehavior than males of any other race. Not surprisingly, projects have also found that Black students are more likely than White students to perceive that discipline is applied unfairly (Kupchik and Ellis, 2008).

Others have become considered with the authoritative nature of school security and policies. For example, many researchers and scholars argue that the modern school is one that is focused more on the safety and security of the schools, and less with the true goal of education. On the other hand, some research has found that school disciplinary policy and school security measures are becoming so increasingly common that students have become desensitized to having to walk through metal detectors or encounter police officers in the hallway of their school (Bracy, 2010). In this sense, school safety police and practice has become normalized for many students in the US schools.

Clearly, these bodies of literature do not suggest that there should not be school security measures and policy within the school. Instead, these scholars show that there may be important side effects of zero tolerance and other policies that do not just affect the safety of students in the school. The question, for many, revolves around understanding where the balance between safety and autonomy lies, something that will, no doubt, be further explored in the future.

The Future of School Safety

In light of recent tragic acts of violence that have taken place in schools around the country, the safety of the hallways in the US schools remains a pressing and important issue. While no one can predict what events in the future will change how we protect the nation's schools, we do know that parents, teachers, administrators, and policy makers will continue to play a pivotal role in protecting America's youth within the institution of education. While no single national reformatory proposal has been, or probably will be, universally embraced, most groups agree maintaining appropriate safety structures within each school to safeguard the nation's youth against violence will remain a top priority for everyone involved.

See also: Delinquency, Sociology of; Deterrence Theory: Crime; Drugs: Illicit Use and Prevention; Prevention of Crime and Delinquency; Youth Gangs.

Bibliography

Bracy, N.L., 2010. Students Perceptions of high-security school environments. Youth & Society 43, 365–395.

Casella, R., 2006. Selling Us the Fortress: The Promotion of Techno-security Equipment for Schools. Routledge, Taylor & Francis, New York.

Dwyer, K., Osher, D., Wagner, C., 1998. Early Warning, Timely Response: A Guide to Safe Schools. U.S. Department of Education, Washington, DC.

Insley, A., 2001. Suspending and expelling children from educational opportunities: time to reevaluate zero tolerance policies. American University Law Review 50, 1039–1074.

Kupchik, A.C., Ellis, N.D., 2008. School discipline and security: fair for all students? Youth & Society 39, 549–574.

Monitoring the Future, 2012. Key Findings on Adolescent Drug Use. The National Institute on Drug Abuse, National Institutes of Health. http://www.monitoringthefuture.org//pubs/monographs/mtf-overview2012.pdf.

Phaneuf, S.W., 2009. Security in Schools: Its Effect on Students. LFB Scholarly Publishing LLC, El Paso: TX. ISBN: 10:1593322755.

Reddy, M., Borum, R., Berglund, J., Vossekuil, B., Fein, R., Modzeleski, W., 2001. Evaluating risk for targeted violence in schools: comparing risk assessment, threat assessment, and other approaches. Psychology in the Schools 38, 157–172.

Schwartz, R., Len, R., 2001. Zero tolerance as mandatory sentencing. In: Ayers, W., Dohrn, B., Ayers, R. (Eds.), Zero Tolerance: Resisting the Drive for Punishment in Our Schools. The New Press, New York, pp. 126–135.

Skiba, R.J., Michael, R.S., Nardo, A.C., Peterson, R., 2000. The Color of Discipline: Sources of Racial and Gender Disproportionality in School Punishment. Indiana Education Policy Center. Research Report SRS1.

Stouffer, G.A., 1952. Behavior problems of children as viewed by teachers and mental hygienist, a study of present attitudes as compared with those reported by E.K. Wickman. Mental Hygiene 36, 271–285.

Sughrue, J., 2003. Zero tolerance for children: two Wrongs do not make a right. Education Administration Quarterly 39, 238–258.

United States Department of Justice, 1999. Addressing Community Gang Problems: A Practical Guide (91-DD-CX-K058). https://www.ncjrs.gov/pdffiles/164273.pdf.

United States Secret Service and United Sates Department of Education, 2002. The Final Report and Findings of the Safe School Initiative: Implications for the Prevention of School Attacks in the United States. http://www.secretservice.gov/ntac/ssi_final_report.pdf.

United States Department of Education, National Center for Education Statistics, 2012. Indicators of School Crime and Safety: 2011 (NCES 2012-002).

United States Department of Health & Human Services, 2013. Stop Bullying. http://www.stopbullying.gov.

School to School Collaboration

Marilyn Friend, The University of North Carolina at Greensboro, Greensboro, NC, USA
Tammy Barron, The University of North Carolina at Greensboro, Cooleemee, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Contemporary education is a complex endeavor, and educators have come to value collaboration as a vehicle for accomplishing its purposes. Collaboration is the style educators use when they partner to achieve shared goals. It is based on voluntariness, parity, shared responsibility for key decisions, shared accountability for outcomes, and shared resources. Collaboration matures with ongoing implementation. Contemporary applications of collaboration include those that are intraschool (teams, professional learning communities, consultation, team teaching, co-teaching) and extra-school (school-to-school collaboration and other partnerships, including those with parents). Challenges to collaboration include clear evidence of its efficacy derived from research, the lack of supportiveness of some school cultures, professionals' preparation for collaborative roles, and availability of time for collaboration as well as other practical barriers. In spite of such challenges, collaboration's importance for education is likely to continue to grow.

Introduction

Although teaching was once considered a profession characterized by isolation, such is no longer the case. This change has been driven by many forces, including the exponentially expanding amount of information in today's world, the rapidly increasing role of technology in schooling, pressures coming from business and industry as they evolve in this second decade of the twenty-first century, the worldwide movement to include individuals with disabilities, and ongoing school reform efforts. The result is that educators have recognized that their complementary but disparate abilities and talents can be shared in order to design effective and innovative instruction, create supportive but rigorous learning environments, solve the many problems they face each day, and generally improve outcomes for all students (Friend and Cook, 1990; McCoach et al., 2010).

In some ways, though, the use of the term *collaboration* has outstripped technical understandings of it and the knowledge based on its practice. That is, collaboration at times seems to be a panacea: It is referenced frequently in the education literature but often in vague or confusing ways. It has been employed to describe any interaction among adults in schools as well as those between school professionals and parents, agency personnel, and community members (e.g., Kampwirth and Powers, 2012). Collaboration has also been assigned to the communication that occurs between adults and children or activities undertaken by groups of children (e.g., Gailbreath and Winterbottom, 2011). Clearly, this range of uses suggests incomplete understanding. Unfortunately, it also sometimes results in professionals forming a negative perception of this essential dimension of contemporary education.

The purpose of this article is to explicate the concept of collaboration as it occurs in school-related settings, also placing it in the context of a larger society, and to describe several of its most common applications in today's schools. In addition, challenges related to professional school collaboration are considered, and conclusions are offered concerning enhancing its current and future implementation.

Conceptual Underpinnings of Collaboration in Education

In everyday parlance, collaboration connotes camaraderie or partnership. However, for professionals it is important to understand its technical meaning and characteristics as well as its prominence in a larger societal context.

A Technical Definition of Collaboration

According to Friend and Cook (2013), collaboration is "an interpersonal style for direct interaction between at least two coequal parties voluntarily engaged in shared decision-making as they work toward a common goal" (p. 6). Two especially central elements of this definition are these: first, as a style, collaboration represents how individual interact rather than what they are interacting about. This is a critical concept. It suggests that any interaction between educators or educators and parents may be collaborative, if that interaction meets the other criteria of the definition. It also infers that other styles exist, and they do. For example, some interactions are directive, in which one individual directly or indirectly guides the other (e.g., 'you should …'). Second, the definition implies that collaboration cannot stand alone; it always is attached to a substantive topic or process. If a professional reports that she/he is collaborating, it only suggests a style; the subject on which the style is focused and the process being used (e.g., problem solving) remain unknown.

Foundational Characteristics

Reviewing the definition of collaboration is a mere first step in understanding its role in schools. Exploring the foundational characteristics that complete the definition of collaboration illustrates its complexity.

Collaboration Is Voluntary

School administrators and other school leaders often have the appropriate responsibility of assigning teachers and others to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92129-9

work together; they determine which individuals will work in proximity, often with the assumption that collaboration will occur. However, it is the participants who choose whether to use the style of collaboration. For example, a school team is created to problem solve about students who struggle academically. This proximity was mandated. However, on one such team members actively contribute, respect all opinions, and keep in mind the goal of improving student achievement. On another team, one person tends to avoid participation, annoyed at being assigned to work on the team. Another argues repeatedly for only his/her ideas, often without considering the value of others' recommendations. The former is collaborative; the latter, obviously, is not. And yet in both cases, membership was assigned.

Collaboration Requires Parity

The contribution of each individual who is part of collaboration must be equally valued. Further, collaborators must share equal power in critical decision making. This does not mean that participants necessarily have parity in interactions outside the collaborative endeavor. For example, a principal and a teacher must have parity if they are collaborating about adopting a new instructional program, even if typically this is not the nature of their professional relationship. It should be noted, too, that parity is often much more complex than implied here, at least in part because it can be profoundly influenced by perception and culture.

Collaboration Includes Mutual Goals

Collaboration is not possible unless participants identify goals that they share. They may have other goals that potentially are in conflict or are unrelated, but at least one mutual goal must be critical enough to be worthy of setting differences aside. For example, professionals may come together to problem solve about a student's behavior. However, if some members are seeking a means of removing the student from class while others are working to keep the student there, collaboration is unlikely, even if all participants agree they want to 'help' the student. A conversation about goals in this case would be essential.

Collaboration Requires Shared Responsibility for Key Decision Making

When school professionals collaborate they actively contribute to making key decisions, whether they concern a new data collection strategy, a new instructional program, a common assessment strategy to be used, or an intervention to assist a student's learning. In today's busy work settings, a common piece of advice is that collaborators should share key decisions related to accomplishing their goal so that they are comfortable in their knowledge of what is planned, but they should divide the labor involved so that it can be accomplished efficiently and effectively – but not necessarily shared.

Collaboration Requires Shared Resources

Limited resources often are one of the most profound reasons school professionals seek to collaborate. The pooling of resources may be the only way to reach a goal. These can be tangible resources, such as materials; skills, such as a teaching methodology or knowledge of particular academic content; a specialized understanding, such as familiarity with and experience in working with students with autism; or availability to complete a particular task that is key to the success of the group.

Collaboration Requires Shared Accountability

When collaboration leads to success, participants are eager to share in the accomplishment, but if the effort is not successful or even led to additional challenges, participants may distance themselves from the outcome. Taking ownership of the entire process with the understanding that there sometimes will be setbacks when working toward a goal is required for members of a collaborative group. This sustains the group through hardships and unexpected challenges, ensures the best possible environment to attain the shared goal, and fosters celebration of goal attainment.

Emergent Characteristics

Collaboration also has emergent characteristics. These characteristics must be present in some minimal way at the outset of collaboration, but they then develop over time as the collaborative relationship strengthens because of its success. Emergent characteristics include trust among participants, accompanied by respect for what each contributes to the endeavor. Gradually, a deep sense of community is cultivated. The developmental nature of collaboration is relevant because participants may need to persist through early, less-than-satisfactory interactions in order to eventually reach a strong and productive relationship.

Collaboration in a Broader Context

The concept of collaboration is by no means limited to the educational arena; it is integral to many aspects of business and industry as well as the medical and health fields. In fact, successful businesses have noted the impact on efficiency and the importance of having professionals who can collaborate since the 1950s (Blanton and Kenett, 2007; Ishikawa, 1990; Juran, 1962). A similar history can be found in the medical community where the interprofessional nature of health care is seen as a means of collecting more sensitive and broader patient assessments with higher reliability of decisions (Zwarenstein and Reeves, 2006).

The use of collaboration in schools as a reflection of larger societal trends can perhaps be best illustrated by current trends and priorities in the business world. It is widely acknowledged that today's business climate makes collaboration a necessity rather than a luxury. Globalization of the marketplace, the amount and complexity of data that must be managed, and the speed at which change occurs all indicate to business leaders that close and carefully engineered partnerships, formed both internally and externally, are foundational to success (Beier, 2014). These leaders are urged to seek job candidates with strong potential to work collaboratively, to encourage collaborative practices by distributing leadership among employees, and to create teams and other internal structures that foster the enriched outcomes created through collaboration (Jensen, 2013; Miller and Katz, 2014). Additionally, contemporary collaboration in business environments does not necessarily

rely on face-to-face interactions. Strong emphasis is placed on electronic collaboration and its potential to facilitate interactions among participants from around the world in effective and efficient ways (Boeri, 2012). What is clear is that collaboration in many forms has become a central element of business success and is likely to continue to grow in prominence and sophistication.

Applications of Collaboration in Educational Endeavors

Although school professionals may not yet be implementing collaborative activities in the ways being pioneered in business settings, the number of applications and their importance in achieving the goals of education have expanded over the past two decades. The following are brief descriptions of the most typical types of school collaboration.

Intraschool Examples of Collaboration

Most school collaboration is focused on the working relationships among professionals and between professionals and parents. All of these partnerships are intended to pool the knowledge and expertise of the participants in a way that leads to improvement in student academic achievement as well as in student social and behavioral well-being.

Teams

Teams are an increasingly common collaboration structure in today's schools. A *team* is an interdependent group of individuals who share their knowledge and skills in order to accomplish a mutual and clearly articulated goal, usually related to improving student outcomes. The assumptions for creating teams are that the task at hand is too complex to be reached by a single individual and that multiple perspectives will lead to more positive results.

Effective teams share several characteristics that have been well researched, both in education and in other domains, including business and health occupations (e.g., Bang et al., 2010). First, teams have clear goals. Without such goals, there is no purpose for this group of individuals to interact. Second, team members are aware of and committed to their membership in the group and the accompanying responsibilities for meaningful participation. Although this characteristic seems almost too fundamental to mention, it sometimes is an issue: Educators occasionally do not recognize the need for them to actively contribute, to work on team goals outside meetings, and to put the work of the team above individual agendas.

Third, teams formally or informally adopt shared norms. That is, they have a set of expectations related to their actions when they interact during meetings, procedures they follow, expectations for work completion outside of meetings, and so on. This leads to the fourth characteristic of teams: Team members have a strong sense of interdependence, the understanding that it is the shared effort that will lead to a positive result. Finally, effective teams are comprised of members who have different types of expertise. If participants are similar, they are less likely to generate creative options; if they have a wide range of knowledge and skills, more and higher quality solutions are likely to result.

School teams may be formed for a variety of reasons. In some locales, they exist to implement schoolwide systems intended to address students who are struggling academically or behaviorally, such as response to intervention, positive behavior supports, or multitiered systems of support. They also are integral to the special education/special educational needs process in which students who may have disabilities are assessed, their eligibility for services determined, and the amount and nature of such services decided. Finally, school teams may be formed for narrower but critical work such as redesigning systems for including students with special needs, selecting and implementing new curricula, incorporating technology into existing programs, or addressing a problem or issue confronting the school.

One currently popular example of school teaming is a professional learning community (PLC). A PLC may consist of teachers on the same grade level, a curriculum coach, possibly an administrator and also other specialized professionals who meet on a scheduled basis. This may be all high school English teachers, a curriculum coach, and a special education teacher coming together to learn about the new English curriculum. On a specified day, twice per month, teachers may meet to share information from various sources regarding the curriculum. They may monitor student outcomes through data analysis of a common unit assessment all the teachers have administered. They then may develop strategies for struggling learners. Each member of the group may take a lead role at various times, depending on the topic at hand. For example, one member may take the lead on discussion of a book about literacy grouping and facilitate the development of methods for further instruction. Another member may take the lead on facilitating analysis of student data. PLCs are intended to function for at least an entire school year and require a collective commitment. Specifically, there must be a collective commitment to the collaborative effort, to the success of each student, and to high levels of learning for PLCs to be effective (Dufour et al., 2006; Dufour, 2004).

School teams may face challenges in accomplishing their goals. For example, time to meet may be limited, especially when various professionals may have schedules that may be difficult to coordinate. In addition, each participant must fully and actively commit to the team's work; if some individuals, perhaps assigned to the team but unwillingly so, isolate themselves, the team's goal is unlikely to be met. Finally, the differences among team members that provide team strength also can make interdependence difficult: When members represent different areas of specialization, come from different cultures, or possibly have different primary interests, teaming requires extraordinary effort and discipline.

Consultation

Consultation has long been recognized as a valuable tool for supporting students with a wide variety of special needs and their teachers. Drawing from earlier applications in the medical professions, consultation began in schools in the middle of the twentieth century primarily as a role for school psychologists and other specialists. Its purpose was to educate teachers about their students who needed support because there simply were

not enough school psychologists (nor would there ever be) to directly intervene with each needy student. In more recent years, consultation has expanded to include teacher–teacher interactions as well as coaching and mentoring. Most importantly, it has shifted from directive models to those based on collaboration. In general, consultation is a process in which one professional with a particular type of expertise provides support or guidance to another professional in order to address a concern or problem related to a third party, usually a student. The result is that the third party is supported but also that the teacher acquires skills for addressing similar situations in the future.

Consultation typically follows a series of steps that can be characterized as specialized problem solving (Newell, 2010). It is premised on professionals' willing participation, and the understanding that, while the person functioning as the expert may learn from the consultee, the goal is directional, that is, the process is undertaken to assist the latter individual. Because of this arrangement, consultants and consultees have different responsibilities and accountability. Consultants are responsible for diligently striving to understand the issue at hand, for offering ideas and strategies that are ethical and feasible, and for supporting the consultee's implementation of those suggestions, interpreting results, and continuing the process until success is achieved. The consultee is responsible for accurately describing the dilemma to be addressed, carefully implementing the consultant's suggestions, gathering needed data, and being willing to persist if additional options need to be explored.

Consultation often is employed when teachers find that they are not being successful in meeting a student's needs. This could pertain to a student who has unusual behaviors, a student who is exceptionally gifted or talented, a student with a particular disorder such as autism, or a class group experiencing extraordinary problems in social interactions. In another application, early career educators may work with a mentor consultant to refine their overall instructional or classroom management skills and to have a successful induction into the profession (Ingersoll and Strong, 2011). Coaching also sometimes is a consultation strategy (Gallucci et al., 2010): For example, if a new reading program is introduced at the elementary school level, a coach may demonstrate for teachers how to complete required instructional activities, occasionally spending time in the classroom but also meeting at other times to discuss progress in implementation.

When specialists consult with teachers and collaboration is stressed, its effectiveness is well established. Far less is known about teacher–teacher consultation. The latter can be complicated by several factors, including the value placed on and the perceived appropriateness of shared peer expertise, the availability of time, and the skills of both the teacher consultant and the teacher consultee. Regardless of who serves in the consulting role, effectiveness is determined by perceived relevance and feasibility of the strategies offered, their immediate impact on the identified concern, and the quality of the relationship between the participants.

Team Teaching

Begun as a high school instructional model in Great Britain and intended to draw on teachers' particular areas of content expertise, team teaching originally consisted of one teacher delivering a lecture to a large group of students, and other teachers providing follow-up, discussing the content, responding to student questions, and monitoring student understanding (Geen, 1985; Warwick, 1971). In the US, this approach was adopted as the Trump model. Over time, it evolved into a variety of contemporary applications. Today, team teaching refers to an instructional arrangement in which two (or more) teachers, each with an assigned full-class group of students, share instructional responsibilities for all or part of the school day.

The key characteristics of this model include the following: First, the teachers involved may have similar or different content expertise. In some instances two elementary teachers use team teaching, blending their classes in order to better differentiate instruction, and so their expertise is similar. In high school, a history teacher and a literature teacher may blend their classes with the goal of teaching history through literature, each contributing their unique expertise. In addition, this application of collaboration implies that joint teaching occurs, but that is not always the case. In schools using a middle school model, for example, four teachers may share responsibility for a group of 100 or more students and plan coordinated instructional activities, but they may not teach together. Further, in most examples of team teaching, the teacher-to-student ratio is constant. In a few applications, though, two teachers may be partnered to teach a single class group, sometimes to improve overall classroom management and occasionally because there happens to be a teacher available for this purpose.

The examples already provided begin to illustrate how team teaching is occurring in today's schools. Other uses of this approach may include multigrade teaching opportunities at the elementary school level or even the option for language immersion (for example, a Spanish teacher and an elementary teacher working with a large group of students for all or part of the school day). In middle and high schools, many examples can be found embedded in the trend to integrate disciplines. For example, science and math teaching may be blended with students applying their mathematical learning in a biology or physics course, led by two teachers. In another example, two teachers lead a student group in a blended anatomy and physical education class.

Several issues can arise with team teaching. First, the professionals participating must have a collaborative relationship that fosters the blending of their discipline-specific expertise. Second, each must be committed to this type of integrated instruction and partner to contribute to its content so that students have adequate coverage of both subject areas. In addition, a physical space conducive to team teaching must be available, that is, a classroom or teaching area that can accommodate two or possibly more class groups of students.

Co-Teaching

The origins of co-teaching lie in the early school inclusion efforts of the 1980s, but its recent increased utilization is the result of the worldwide trend toward inclusive practices as well as school reform initiatives designed to ensure access and achievement for all learners. Co-teaching is the instructional model in which two or more professionals, most often

a general education teacher and a special education teacher, jointly deliver instruction to a diverse group of students in a shared classroom space, an arrangement that may occur for the entire school day or for a specified amount of time or class period.

A key characteristic of co-teaching is the contribution that each professional makes. For example, a special education teacher may work with a general education teacher in math class. The general education teacher's appropriate primary focus is on content and pacing while the special educator incorporates specialized techniques to facilitate the learning of students with special needs. Because successful co-teaching relies on trust between the teachers and strong willingness to share authority in the classroom, collaboration is essential.

Specific co-teaching approaches have been identified that maximize this instructional option (Friend, 2014). These are summarized in **Table 1**, although creative practitioners have added many variations to the basic structures.

Effective co-teachers employ the approaches based on student need and the nature of the planned instruction, fluidly moving among them across lessons or even within a single lesson. For example, in the math example already introduced and using station teaching, one teacher might work with students on vocabulary, one might introduce new concepts, and the students might independently work on a problem solving activity based on previous learning. The math teacher's responsibility is to be sure the new concepts are clearly presented; the special educator's responsibility is to embed in the instruction learning strategies to help students remember the vocabulary as well as the new concepts.

Co-teaching is a sophisticated collaboration application influenced strongly by professionals' working relationships. Patterns or expectations relate to diversity can directly affect co-teachers' shared instruction. For example, a male teacher from a culture in which women should defer to men's judgment may find co-teaching difficult with a female co-teacher. Ultimately, one of the most important elements of co-teaching is flexibility.

Table 1 Co-teaching approaches

Co-teaching approach	*Description*
One teaching, one observing	One teacher is solely responsible for the delivery of instruction while the other teacher gathers data on student performance.
Station teaching	Students are divided into three groups and rotated through three instructional activities, two of which are facilitated by teachers and one completed independently.
Parallel teaching	Each teacher is actively teaching during the lesson but may use slightly different strategies to teach the same concepts.
Alternative teaching	One teacher is responsible for large group instruction while the other teacher focuses on small-group work in which instructional goals may be different.
Teaming	Teachers work in complementary ways during a lesson, working toward the same goals. Instructional leadership is shared and integrated.
One teaching, one assisting	One teacher leads instruction of the whole class while the other teacher assists students individually.

Extra-School Examples of Collaboration

School applications of collaboration are not limited to those that occur among the professionals at a single site. School professionals have recognized the benefits of collaboration outside their educational setting to enhance instructional techniques, share resources, and improve learning opportunities for students. Such collaboration includes that which occurs between schools and agencies that which is school-to-school.

School–Agency Collaboration

Educators in many schools have reason to collaborate with professionals in other organizations or agencies. For example, specialists working with young children with developmental delays often collaborate across multiple organizations (and, of course, with parents) to provide services to these at-risk students. Special programs for at-risk youngsters, social services, and other agencies serving families all may participate in efforts to design and deliver early education for the children. School professionals also may work with community agency personnel to provide social, mental health, or medical assistance for children and families and to transition children through different school experiences. Each agency may have a focused goal for the child, but those goals are part of the larger collaborative goal of supporting parents and children.

School-To-School Collaboration

Many types of school-to-school collaboration have been reported in the professional literature. These partnerships may occur between schools at the same level of instruction. For example, fourth grade teachers in one primary school partner with teachers in another primary school in the same school district to make common assessments for shared units of curriculum. Another form of school-to-school collaboration includes schools of different levels of instruction. For example, secondary school teachers plan with teachers from a primary school to create a tutoring program that will assist the younger students and create option for service for the older students.

One compelling example of school-to-school collaboration is the Afri Twin project. In this ambitious initiative 75 South African primary and secondary schools were paired with 75 comparable schools in Britain and then 'twinned' with a third school in a rural area (Klerk and Nienaber, 2011). Each group of three schools shared one common goal and focused on forging of friendships and improving communication skills. The schools shared resources by building international links, creating a two-way knowledge transfer, and learning alternative skills and values. In addition, the project addressed parity and diversity issues by stressing pride, self-esteem, and personal growth for all participants, and it fostered racial tolerance.

The Complexity of Collaboration in Education

Given the importance of collaboration for contemporary schooling and the many collaborative applications that are

enabling educators to optimize their students' learning, it might seem that the road forward would be obvious, that is, continue to foster collaborative practices and ensure they are sustainable. However, a number of complexities ensure that the future of school collaboration is neither readily apparent nor a particularly straight path.

The Evidence Base for School Collaboration

One potential barrier to expanding collaborative applications in schools is the paucity of rigorous studies related to them. Some evidence does exist, but it often comes from evaluation studies or anecdotal reports rather than from traditional experimental research (e.g., Solis et al., 2012). In fact, collaboration occasionally has been criticized as tangential to effective education practices or a distraction from them.

Part of the reason for the dearth of research is the very nature of collaboration. That is, as noted at the beginning of this article, it never exists by itself but is always associated with a process, activity, or program. That makes it difficult to determine whether the activity itself or the collaborative approach adopted to implement the activity that was responsible for the results obtained.

School Culture Supportive of Collaboration

Another factor contributing to or detracting from the growth of collaborative applications is school culture. Although nearly every school leader would profess to embrace collaboration and foster its development, data suggest that the words are not always accompanied by actions. In some schools, professionals are rewarded for their individual efforts with little attention paid to collegial work. In yet others, leaders tend to use a directive style that causes teachers and other personnel to isolate themselves, completing assigned tasks but avoiding commitment to shared work. In the US, such cultures sometimes are an unintended consequence of the current trend toward individual educator accountability for student academic achievement as measured on standardized high stakes assessments.

Professional Preparation for Collaborative Responsibilities

Whether an individual professional is prepared at the preservice level for the collaborative activities they are likely to experience in schools depends in large part on the approach taken by the preparation program they completed. Most special education personnel are required to take coursework on collaboration and have intensive field experiences with teaming and co-teaching. A similar comment could be made for middle school teachers and team teaching. However, elementary and high school teachers may or may not have any direct preparation related to collaboration. If they do not, the consequences may be sobering. These educators may find an expectation for collaboration as foreign and potentially intrusive. They may be resistant to the notion of sharing responsibility and accountability, and most of all, they may lack the skills for effective collaborative practice, including communication and problem solving skills, which, in turn, may limit outcomes for their students.

The Logistics of Collaboration

In too many schools, even when commitment to collaboration is strong, professionals understand and welcome collaborative roles, and school leaders encourage collaborative practices, those practices occur only informally or on a significantly limited basis. The reasons too often relate to logistics. One issue may be time available for professionals to work with one another in any other than a superficial way. That is, school schedules may not incorporate time for professionals to regularly work together, and times before and after school may be filled with other responsibilities and meetings. Teachers may perceive that they do not have sufficient time to adequately discuss problems and may feel pressure to find solutions regardless of the insufficient analysis of them (Williamson and McLeskey, 2011). Further, specialists such as data or behavior experts, may not be available to attend meetings at times convenient for teachers, and the teachers have to proceed without potentially valuable input. The result can be interventions or solutions that are not effective.

A second example of logistics challenges concerns physical space, which may affect the feasibility, structure, and character of collaboration. Professionals need appropriate settings for their collaborative efforts, whether that is a classroom for co-teaching or team teaching or a safe and private conference room for consultation or professional learning communities. When reasonable space cannot be provided, problems can occur. For example, a group trying to find a solution to a problem left to work in a library where students and other staff roam may be less likely to speak freely and record their ideas on a chart due to confidentiality concerns. This may disrupt the group's brainstorming, resulting in dissatisfaction on the part of participants and planned interventions that may not be the product of the group's best thinking.

Conclusion

Collaboration, as understood in a technical professional sense, is essential in contemporary schools and is a reflection of the trends toward shared work in other disciplines, including business and medicine. Examples of collaborative applications include teaming, professional learning communities, consultation, co-teaching, team teaching, school–agency, and school-to-school collaboration. And even more examples could have been described, including the collaboration that school professionals have with parents and family members, the partnerships that exist between schools and universities, and projects that emphasize collaboration between schools and community organizations. As expectations continue to increase for student achievement and educators are faced with increasingly complex job demands, collaboration likely will continue to blossom as a vehicle through which higher student achievement and other positive student outcomes can be obtained.

See also: Communities of Practice; Cooperative Learning in Schools; Educational Planning Worldwide; Educational Productivity; Educational Sociology; Educational Systems:

North America; Environments for Education; Family and Schooling; Group Processes in the Classroom; Open Education: Learning and Teaching; Parents and Teachers as Partners in Education; School Effectiveness Research; School Improvement; School Management; School as a Social System; Schools, Micropolitics of; Socialization and Education: Theoretical Perspectives; Teacher Behaviours and Student Outcomes; Teacher Education; Teacher Expertise; Teaching as a Profession: United States.

Bibliography

Bang, H., Fuglesang, S.L., Ovesen, M.R., Eilertsen, D.E., 2010. Effectiveness in top management group meetings: the role of goal clarity, focused communication, and learning behavior. Scandinavian Journal of Psychology 51, 253–261.

Beier, Y., 2014. The collaborative advantage. Communication World 31 (1), 22–25.

Blanton, G.A., Kenett, R., 2007. Joseph M. Juran; a perspective on past contributions and future impact. Quality and Reliability Engineering International 22 (6), 653–663.

Boeri, R., 2012. Collaborative digital documents: from creation to consumption. Econtent 35 (3), 29.

De Klerk, W., Nienaber, A.W., 2011. A discussion of a unique collaboration model between schools. Psychology Research 1 (3), 170–181.

Dufour, R., 2004. What is professional learning communities? Educational Leadership 61 (8), 6–11.

Dufour, R., Dufour, R., Lopez, D., Muhammad, A., 2006. Promises kept: collective commitments to students become a catalyst for improved professional practice. Journal of Staff Development 27 (3), 53–56.

Friend, M., 2014. Co-teach: Building and Sustaining Effective Classroom Partnerships in Inclusive Schools, second ed. Marilyn Friend, Inc, Greensboro, NC.

Friend, M., Cook, L., 2013. Interactions: Collaboration Skills for School Professionals, seventh ed. Merrill/Pearson, Upper Saddle River, NJ.

Friend, M., Cook, L., 1990. Collaboration as a predictor for success in school reform. Journal of Educational and Psychological Consultation 1, 68–83.

Gailbreath, J., Winterbottom, M., 2011. Peer tutoring: what's in it for the tutor? Educational Studies 37, 321–332.

Gallucci, C., Van Lare, M., Yoon, I.H., Boatright, B., 2010. Instructional coaching: building theory about the role and organizational support for professional learning. American Educational Research Journal 47, 919–963.

Geen, A.G., 1985. Team teaching in the secondary schools of England and Wales. Educational Review 37, 29–38.

Ingersoll, R.M., Strong, M., 2011. The impact of induction and mentoring programs for beginning teachers: a critical review of the research. Review of Educational Research 81, 201–233.

Ishikawa, K., 1990. Introduction to Quality Control. 3A Corp, Tokyo.

Jensen, K.R., 2013. Cooperative efforts. Communication World 30 (7), 15–18.

Juran, J., 1962. New life for staff departments. Management of Personnel Quarterly 1 (3), 16–19.

Kampwirth, T.J., Powers, K.M., 2012. Collaborative Consultation in the Schools: Effective Practices for Students with Learning and Behavior Problems, fourth ed. Prentice-Hall, Upper Saddle River, NJ.

McCoach, D.B., Goldstein, J., Behuniak, P., Reis, S.M., Black, A.C., Sullivan, E.E., et al., 2010. Examining the unexpected: outlier analyses of factors affecting student achievement. Journal of Advanced Academics 21, 426–468.

Miller, F.A., Katz, J.H., 2014. 4 keys to accelerating collaboration. OD Practitioner 46 (1), 6–11.

Newell, M., 2010. The implementation of problem solving consultation: an analysis of problem conceptualization in a multiracial context. Journal of Educational and Psychological Consultation 20, 83–105.

Solis, M., Vaughn, S., Swanson, E., McCulley, L., 2012. Collaborative models of instruction: the empirical foundations of inclusion and co-teaching. Psychology in the Schools 4, 498–510.

Warwick, D., 1971. Team Teaching. University of London, London.

Williamson, P., McLeskey, J., 2011. An investigation into the nature of inclusion problem-solving teams. The Teacher Educator 46, 316–334.

Zwarenstein, M., Reeves, S., 2006. Knowledge translation and interprofessional collaboration: where the rubber of evidence-based care hits the road of teamwork. The Journal of Continuing Education in the Health Professions 26, 46–54.

Relevant Websites

http://cecp.air.org/ – Center for Effective Collaboration and Practice: Homepage and links.

http://www.edutopia.org/blog/school-community-collaboration-brendan-okeefe – Edutopia: Five Steps to Better School/Community Collaboration.

http://iris.peabody.vanderbilt.edu/module/esp/ – The Iris Center: Effective School Practices – Promoting Collaboration and Monitoring Student Progress.

http://www.aasa.org/content.aspx?id=12512 – School Superintendents Association (AASA): 5 Ways To Build a Culture of Collaboration with Staff, Teachers and Parents (by Sharon D. Kruse).

http://cooltoolsforschools.wikispaces.com/Collaborative+Tools – Wikispaces: Web 2.0 – Cool Tools for Schools (Electronic Collaboration Options).

Schooling: Impact on Cognitive and Motivational Development

Ulrich Trautwein, University of Tübingen, Tübingen, Germany
Hanna Dumont, German Institute of International Educational Research, Berlin, Germany
Anna–Lena Dicke, University of Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F.E. Weinert, volume 20, pp. 13589–13594, © 2001, Elsevier Ltd.

Abstract

How strongly does a student's cognitive and motivational development depend on various characteristics of schooling? This article describes the origins and developing profile of the field of research on differential effects of learning environments, the current knowledge about central characteristics of schooling, persistent methodological challenges, and implications for educational practice.

How much difference does it make for math achievement if one has the best math teacher in the best school of town rather than an average teacher in a typical school? How strongly is a student's motivation affected by the students she or he is studying with? These and similar questions are extremely important issues – for both research and practice. Millions of parents around the world think about these issues when they make decisions about their children's school biographies, and an increasing number of researchers from several different disciplines are busy in finding adequate ways of gauging the effects of schooling.

In examining "effects of schooling," it is important to distinguish between two perspectives. When school versus no school conditions (e.g., "is there an effect of two more years of schooling on cognitive development?") are compared, one studies 'general' or 'average' effects of schooling on cognitive or motivational development. When differences among classes, tracks, schools, school systems, and the like are studied (e.g., "how much difference does school quality make?"), the primary focus is on what we will call "differential effects of learning environments." Learning environments such as school classes, schools, or school types function as differential learning and developmental environments "when young people – independent of, and over and above, their individual intellectual, cultural, social, and economic resources – are afforded different developmental opportunities" depending on the class, school, or school type they attend (adapted from Baumert et al., 2006, p. 99, own translation). This article focuses on the emerging field of research on differential effects of learning environments as evidenced in large-scale school achievement studies.

Schools as Differential Learning Environments: The Emergence of an Interdisciplinary Research Field

The starting point of modern empirical research on the impact of schooling on children's development is typically seen as back in the 1960s when the Coleman report (Coleman et al., 1966) was published and the International Association for the Evaluation of Educational Achievement (IEA) conducted its first study comparing the academic achievement of 13-year-olds in 12 countries (Foshay et al., 1962). Since these pioneer studies, empirical research on the impact of schooling has increased tremendously – not only with respect to national studies, but also regarding large-scale cross-national studies: The IEA now conducts several studies with up to 60 participating education systems. The Organization for Economic Development and Cooperation (OECD) has started its own Programme for International Student Assessment (PISA) in 2000 involving more than 70 countries to date. The increasing breadth of the data base also concerns the variables under study. Whereas earlier studies on the effects of schooling (including international school achievement studies) had their main focus on students' academic achievement, in the last two decades students' motivational development and their engagement in learning have become a key focus of empirical studies on the impact of schooling.

From its beginning, research on the effects of schooling has been a field that attracted researchers with diverse disciplinary backgrounds, and this multidisciplinary background has contributed to the richness of conceptualizations of the effects of schooling. As observed by Wang et al. (1993), researchers from different disciplines typically had a focus on different variables. For instance, sociologists were usually concerned with demographic variables such as social class as well as institutional differences between schools. Political scientists studied the impact of federal, state, and district policy variables. Psychologists typically studied individual characteristics of teachers and learners and the interactions between them. Adding to that list, pedagogical research concentrated on the aims of schooling and the design and content of curriculum.

Over the last two decades, new developments can be observed with regard to the disciplinary composition of research on the effects of schooling. Additional disciplines have increased their efforts in this field, making the field more heterogeneous. Notably, economics of education has attracted much interest (e.g., Hanushek et al., 2010) and brought in a focus on efficiency and the so-called 'returns to education' (the monetary exchange value of education). Moreover, the field of didactics has become more important, specifically when explaining domain-specific outcomes of teaching. In addition, the construction of standardized achievement tests and decisions about research designs has increasingly been informed by psychometricians. Although currently more

focused on learning processes rather than learning outcomes, the emerging field of 'neuroeducation' might also contribute to research on the effects of schooling in the future.

On the one hand, these developments have added additional perspectives and methodologies to the study of the effects of schooling; on the other hand, they have helped to move toward more integrative frameworks that bring together the wisdom from various disciplines. In fact research on schooling effects (and, more broadly, effects of educational institutions) has become an interdisciplinary field in itself, to varying degrees across different countries, which brings together different streams of theorizing from disciplines such as educational research, psychology, sociology, economics, statistics, and others. In some countries, the interdisciplinary endeavor to study effects of schooling has become a separate discipline in its own right.

Integrative frameworks on the effects of schooling come in different forms. The so-called Educational Effectiveness Models (Creemers, 2009; Scheerens and Bosker, 1997; Wang et al., 1993) integrate the wisdom from various disciplines by systematically collecting variables at various different analytical levels with potential effects on academic outcomes and gauge their relative effect. For instance, in the theoretical framework by Wang et al. (1993) 228 variables are grouped into six broad theoretical constructs. The six theoretical constructs used to organize the framework are (1) State and District Governance and Organization; (2) Home and Community Educational Contexts; (3) School Demographics, Culture, Climate, Policies, and Practices; (4) Design and Delivery of Curriculum and Instruction; (5) Classroom Practices; and (6) Student Characteristics. The prominent overview study by Hattie (2009) on the most important determinants of academic achievement is also rooted in this tradition. A second form of integrative frameworks describes the process of learning and development as an interplay of individual characteristics and learning environments whose form may be different across different ages, schools, and school systems ("differential learning and developmental milieu" or "schools as multilevel social organizations"; Baumert et al., 2006; Eccles and Roeser, 2011). This second form of integrative models is typically less exhaustive in terms of variables, but rich in its theoretical foundation and more specific in explaining learning processes and the development of achievement and motivation.

Effects of Schooling: Early Findings and Current Wisdom

To the shock of many politicians, educators, and teachers, the influential Coleman Report (Coleman et al., 1966) concluded "that school brings little influence to bear on a child's achievement that is independent of his background and social context" (p. 325). Along similar lines, Jencks et al. (1972) stated a few years later "that the character of schools' output depends largely on a single input, namely, the characteristics of the entering children. Everything else – the schools' budget, its policies, the characteristics of the teachers – is either secondary or completely irrelevant" (p. 256).

The early accounts of the effects of schooling are likely to have underestimated the role of differential learning environments for several reasons. First, reading competence was the most important outcome in the Coleman report. However, as is known by now, differential learning effects in reading are smaller than in, for instance, mathematics. Second, the sample of schools used in the Coleman report was less diverse than it could have been, and this might have led to the observation of smaller effects. It should also be noted that the understanding of what is to be considered a small/meaningful effect has changed over the years.

In contrast to the earlier studies, contemporary wisdom on the effects of schooling emphasizes systematic differences in cognitive and motivational outcomes that can be traced back to differential schooling quality. At the same time, recent research clearly highlights the need to differentiate between various factors at several levels, different domains, and multiple outcomes when gauging the effects of schooling. The main findings from recent research on the impact of schooling on children's cognitive and motivational development are summarized next.

Differential Impact of Schooling on Cognitive Development

Schools are expected to teach students knowledge and skills in a number of different subjects, such as languages, mathematics, or science, which students need in later life. In recent years, both in practice as well as in research, there has been a shift in the educational objectives of schools. Whereas the focus used to be on the acquisition and reproduction of subject-specific knowledge, students are now expected to "apply knowledge and skills in key subject areas and to analyze, reason and communicate effectively as they pose, interpret and solve problems in a variety of situations" (OECD, 2010a, p. 18).

Against this backdrop, the concept of competence – similar to the concept of literacy used by the OECD and the IEA – has become prominent in studies assessing the impact of schooling on cognitive development. Competencies can be defined as context-specific cognitive dispositions for achievement which are acquired through learning and are needed to successfully cope with situations and tasks in specific domains (Klieme et al., 2008). In most studies on the impact of schooling, domain-specific competencies such as reading literacy, mathematical competence, or scientific competence are assessed. Whereas competence tests used in national studies focus more on specific school subjects and usually have higher curricular validity (therefore oftentimes called 'academic achievement tests'), the assessment of more broadly defined key competencies such as problem-solving skills or information and communication technology (ICT) literacy in addition to subject-specific competencies is becoming increasingly important in international large-scale assessment studies.

Characterized by its context specificity and its focus on 'real life' situations, the concept of competence is to be distinguished from general cognitive abilities such as intelligence. The concept of intelligence refers to the "relatively decontextualized ability to reason and to address new problems without content-specific knowledge" (Baumert et al., 2009). Whereas competencies have to be acquired through learning, intelligence can be learned only to a certain extent and is usually understood as a relatively stable characteristic of a person. In that sense, intelligence can be seen as an important determinant of the

development of domain-specific competencies (Baumert et al., 2009; Klieme et al., 2008). Even though intelligence is, consequently, oftentimes used as a predictor or a control variable in studies investigating the impact of schooling on students' competence development, there is evidence showing that intelligence is also affected by schooling (Becker et al., 2012; Ceci, 1991). However, in the following, we will focus on empirical studies investigating schooling effects on students' academic competencies.

How large are differences in students' competencies across schools and school systems? When comparing different countries in cross-national studies as those conducted by the IEA or the OECD, massive effects of the quantity and quality of schooling can be ascertained, in particular when comparing developed countries with developing countries. In PISA 2009, for instance, the gap in reading literacy between the highest and the lowest performing country was 242 score points, which is the equivalent of more than 6 years of formal schooling (OECD, 2010a). Similarly, in within-country analyses, substantial differences between mean scores in the best-achieving schools and the worst-achieving schools can be observed. In addition, even when controlling for differential intake (e.g., prior achievement and social background of students), longitudinal studies document substantial differences across schools and classes in cognitive development.

What aspects of schooling lead to these observable differences? In the classic review by Wang et al. (1993), "the theoretical construct with the greatest effect was Student Characteristics, followed by Classroom Practices, and Home and Community Educational Contexts. Having less effect were Design and Delivery of Curriculum and Instruction and School Demographics, Culture, Climate, Policies, and Practices, while State and District Governance and Organization had the least effect" (p. 270). This main conclusion is also echoed in more recent accounts. Drawing on Hattie's (2009) extensive synopsis of empirical findings on what affects students' academic achievement, which is based on over 800 *metaanalyses*, as well as other research reviews on what constitutes student learning (e.g., Eccles and Roeser, 2011), there is now a large consensus among educational researchers that proximal variables such as the instructional quality provided by the teacher exert a much stronger influence on students' competencies than distal variables such as school resources. Or put differently, general characteristics of the school environment do not necessarily impact student learning, they only do so when they reach down to the classroom level and affect teachers' instructional practices. This general result will next be illustrated with respect to three frequently discussed aspects of schooling (school resources, tracking effects, and teaching quality).

School resources: One characteristic of schools, which is often subject of many debates among parents, teachers, and policy makers, is the amount of resources available. Overall, the combined effects of resources are often overestimated (Hanushek and Lindseth, 2009). Whereas few researchers would argue that money does not matter at all, its impact clearly depends on the way it is spent. For instance, if a school uses additional resources to hire qualified teachers, the positive effect that school resources can have on student outcomes is evident. On the other hand, using more resources to reduce class size is not by itself going to have a positive impact on student outcomes; its impact depends on the instructional quality provided. Merely reducing the number of children in a learning group does not necessarily lead to different and more effective instructional strategies of teachers (for a summary, see Hattie, 2009).

Tracking/composition: Another practice in many school systems around the world is tracking, that is grouping students into different courses, study programs, or schools according to their achievement. Even though a large amount of research has been conducted on the effects of tracking on students' achievement in recent years, few strong and definitive answers have emerged. One general conclusion, which can be drawn, is that grouping students by ability has an effect insofar as it determines the curriculum and instruction students receive and exposes them to different peers. More specifically, it has been shown that students in higher tracks may develop higher competencies than students in lower tracks because they receive a different type of instruction and curriculum. Tracking also exerts an effect on students' competence development by determining the composition of the student body. There is strong evidence showing that being surrounded by high achieving students has a positive effect on students' competence development. But not only the mean achievement level of a school or a class is of importance, the social and – to a lesser degree – the ethnic composition of the student body has also been shown to impact students' learning: As the ratio of students from disadvantaged families goes up, the achievement of students goes down. These effects of the composition of the student body are again assumed to be mediated by more proximal variables, such as teachers' practices and daily interactions between students. To that effect, the composition of the class has a greater influence than the composition of the school.

Teaching quality: Generally, the instruction a teacher provides needs to be intellectually challenging and cognitively stimulating for students by paying close attention to the prior knowledge of students (Kunter et al., 2013). This may be achieved by many different practices and strategies such as cooperative learning, direct instruction, concept mapping, reciprocal teaching, or formative assessment. Whatever may be the specific strategy, close attention must be paid to the domain specificity of the competence that is to be learned. At the same time, teachers need to transport clear guidelines, communicate their expectations, and provide students with regular feedback. It is evident that this kind of instruction can be more easily given by teachers who have higher levels of teacher qualifications, more knowledge in the subject matter as well as more pedagogical content knowledge. But not only characteristics of the teachers themselves are important, there is also evidence showing that principals or school leaders may play an important role in the development of students' competencies when they show strong instructional leadership, including a learning climate free of disruption, a system of clear teaching objectives, and high expectations for teachers and students.

Impact of Schooling on Motivational Development

A positive development in motivation is seen as both a goal of schooling in itself and a strong determinant of academic

achievement and students' educational choices. Whereas student motivation was understood to be externally driven within the behaviorist tradition of the early twentieth century, current theorizing understands student motivation as a social-cognitive process. Student motivation is, thus, determined by students' subjective beliefs and attitudes toward schooling, which are developed in response to their social surroundings. Motivation has more stable and more variable, that is more situation-specific, elements. Because of space restrictions, our emphasis in this article is on interindividual differences in relatively enduring motivational predispositions to engage in academic activities. Pintrich (2003) differentiated among five central motivational constructs: competence beliefs, control beliefs, interest, value, and goals. This article exemplarily focuses on two of these constructs – competence beliefs and value – which have been prevalent in motivational education research in the last few decades.

Students' self-beliefs about their own abilities, capabilities, and competencies in certain areas, such as specific subjects and tasks, do influence the strength of their engagement with these specific tasks. The higher students' subjective expectancies to successfully engage in and master a specific situation are, the stronger their motivation will be to engage in this specific situation will be. Similarly, students' motivation to engage within the school context is seen to depend on the extent to which the specific context matches their values, needs, and goals. In addition, the subjectively perceived usefulness of a specific task in terms of accomplishing one's personal goals and the perceived enjoyment while performing that task is seen to increase the likelihood of students' engagement (Eccles and Wigfield, 2002).

Summarizing a huge number of studies, several characteristics of the learning environment are substantially associated with motivational development. As with achievement, classroom processes are closely associated with motivational development, whereas somewhat 'paradoxical' effects are found at the level of schools/tracks.

Tracking/classroom composition: The composition of the student body with regard to its academic achievement levels has been found to be one important contextual influence. In particular, students' self-beliefs about their ability are strongly affected by the achievement levels of their immediate classmates (Marsh et al., 2008). When assessing their own abilities and academic standing, students will naturally use the achievement levels of their immediate classmates as a frame of reference. Thus, equally able students have been found to report lower confidence in their academic competencies if they are attending schools with higher average levels of ability, a phenomenon that is commonly described as the big-fish-little-pond-effect. Further research suggests that students' valuing and enjoyment of school, e.g., their interests, might potentially be affected in a similar way. Selective schooling thus might produce negative motivational patterns for students of high ability. To fully depict the complexity of the influence of peers' ability levels on individual student motivation, however, the type of ability tracking needs to be taken into account as well (Chmielewski et al., 2013).

Teaching quality: The type of classroom environment that is created by teachers has been shown to be particularly influential to the development of student motivation. Teachers influence the development of student motivation in multiple ways within the classroom setting: through their choice of tasks and materials, the instructional practices they employ, and the type of relationship they foster with their students (Schunk et al., 2008). With regard to the choice of tasks and materials, certain task characteristics have been identified to be beneficial for the development of student motivation. The use of surprising task features, the use of challenging tasks in combination with appropriate instructional support as well as highlighting the relevance of the educational content for students' lives appear to positively influence students' enjoyment, their feelings of competence, and the value they attach to school. Considering teachers' instructional practices in the classroom, several instructional practices have been identified that promote student motivation in the classroom. On the one hand, the application of student-centered classroom practices that support students' feeling of autonomy such as the provision of choices has been shown to increase student motivation and engagement in the classroom. Additionally, teacher practices that support students' feelings of competence such as the provision of informative and constructive feedback that focuses on students' individual improvement over time will increase student motivation. At the same time, teachers' general classroom management influences student motivation. The use of proactive practices such as communicating clear expectations and rules with regard to classroom conduct has been shown to support students' commitment and attachment of value to school. Lastly, the type of relationship that teachers foster with the students in their classroom is connected to the development of student motivation (Wentzel, 2009). The provision of an emotionally safe and trusting environment in which teachers act as role models as well as confidantes providing social support has been found to increase student commitment and valuation of school. A supportive relationship with the teacher appears to be instrumental for student motivation over and above the support that is provided by parents and peers in some circumstances.

Interest differentiation: Schools are sometimes seen as a killer of student interest rather than a developer (Travers, 1978, p. 125). In fact, the trajectories for motivational dispositions such as self-concept and interest show a decline from school entrance to adolescence, and this decline might be particularly strong under some conditions. For instance, it has been suggested that the decline in student motivation from elementary to secondary schooling commonly found in national and cross-national studies is actually related to systematic changes in the classroom environment. It is assumed that the classroom environment in secondary schools fails to meet the specific developmental needs of students (Eccles et al., 1993). Students' increase in self-awareness and their increased need for independence during adolescence appear to be mismatched with the organization of the classroom environment that is shaped by a high level of control and promotion of social comparison and competition within secondary schools. At the same time, the overall decline in self-concept and interest may also reflect age-normative developmental trends in terms of intraindividual differentiation across different subjects: Although students partly lose interest in some or the majority of their subjects, most students have stable or even increasing interests in other subjects.

Persisting Methodological Challenges

Because of several methodological challenges, it is too early – and in fact impossible – to make final statements about the size of the effects of differential learning environments. First, it is quite obvious that the size of these effects differs across different samples; for instance, between-school differences may tend to be larger in more differentiated school systems. Moreover, the size of effects is likely to change over time as school environments and societies develop. Second, the outcome variables under study also determine the observed effects (for instance, larger effects have been found for mathematics than for reading). Third, a fundamental challenge for research on schooling effects are problems with 'differential intake' or 'selection.' In order to come to an accurate estimate of the impact of a certain aspect of schooling (such as the composition of the student body of a school) on students' competencies, there is a need to (statistically) control for other variables affecting student outcomes, most importantly students' family background and characteristics of students such as intelligence and their prior knowledge (Scheerens and Bosker, 1997). Unfortunately, not all empirical studies control for these variables when analyzing the impact of schooling on students' cognitive development. Experimental and quasiexperimental intervention and reform studies are still rare. Fourth, measurement issues are relevant because the restricted reliability and validity of some achievement tests will lead to a downwardly biased estimate of schooling effects. Fifth, there are a number of challenges for data analysis including the handling of hierarchically structured data and the almost inevitable problem of missing values that have not been overcome in a satisfactory way; missing values are particularly troubling if there are differential participation rates/differential attrition across different learning environments.

Practical Implications

How can current research findings on the effects of schooling inform educational policy and practices? First, the research findings which indicate that students' development differs to a meaningful degree depending on the learning environment they experience highlight issues of distributional justice. Differential outcomes despite similar initial student characteristics are at odds with meritocratic principles embraced in many societies. In fact, in many countries there is the explicit goal and widespread claim that all schools are of comparable quality so that, in general, there are no – or only small – differential effects of schooling, and student development is believed to be similar across different learning environments. In a less strict variation of this approach, in some countries with a rather differentiated school system, it is claimed that although schools differ in some respects, each student is assigned to a school that offers a perfect fit for his or her strengths and needs. At first sight, differential effects of schooling are less of a normative problem in school systems in which differences in cognitive and motivational development as a consequence of differential learning environments are considered to be in line with meritocratic principles because the placement in 'good' schools is believed to depend on prior achievement and/or student aptitude; however, research findings may not unanimously support these claims and thus challenge policy practices in providing each student with equal opportunities.

A second implication of the evidence for the existence of differential learning environments is that in most school systems there is considerable room for improvement. Intervening with crucial determinants of cognitive and motivational development will likely have a positive impact. For instance, if all schools within a state became as effective as the most effective schools within this state, the average achievement in this state might increase considerably, especially if their effectiveness was quite different in the beginning. Along these lines, ways to improve the quality of schooling have been discussed in many countries, and several measures have been put in place. An influential, well-known example from the recent past is the No Child Left Behind Act of 2001 in the US which is often referred to as a highly ambitious perspective on how much improvement in educational effectiveness is possible over a short timeframe. For instance, one element of this act is the Adequate Yearly Progress (AYP) in test scores on standardized achievement tests that schools need to 'produce' in order to receive federal funding. However, current research does not unanimously support the optimistic view on how much progress can be achieved over a relatively short time span as formulated in this and similar reforms. Importantly, unlike less important factors such as class size and the use of tracking, the most powerful determinants of student achievement are not easily influenced by political interventions. For instance, teaching quality cannot quickly and easily be increased, but rather requires long-term efforts to increase the quality of teacher education and to attract the best candidates to the teaching profession. Moreover, research highlights the fact that – given the huge selectivity effects in school composition – even schools with below-average student achievement can be very effective schools (and therefore may not have much room to improve).

Future Directions

Whereas current research allows for a general understanding of the sources of differential cognitive and motivational development, there is a need for ambitious research programs that lead to a better understanding of schooling effects. For instance,-more research is needed to understand the seemingly opposing effects of classroom composition on cognitive and motivational development. Moreover, the strength of the association between proficiency and motivation varies considerably between countries, which suggests potential differential influences of the school system on student motivation across different countries. Similarly, we have yet to better understand how student characteristics may moderate the effects of characteristics of the learning environment (i.e., there is a lack of systematic studies of aptitude–treatment interactions). In addition, there is a need for more intervention work. Intervention studies are particularly strong in terms of establishing causality and could also help in finding ways to positively impact students' cognitive and motivational development. More generally, rather than just 'predicting' cognitive and motivational development, there seems to be a need for longitudinal studies with a broad theoretical background encompassing the

wisdom of several disciplines in order to better understand the processes underlying cognitive and motivational development. Longitudinal studies will also help to understand the long-term impact of schooling (mediated by cognitive and motivational development) on lifelong learning, educational attainment in higher education, and active participation in society.

See also: Academic Motivation and Performance: Task Value Interventions; Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Flow in Motivational Psychology; Gender and Academic Motivation; Grit; Interest, Psychology of; Leisure Activities Choices among Adolescents; Mastery Learning; Motivation, Learning, and Instruction; Passion and Motivation; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Temperament and Motivation; Test Anxiety and Academic Achievement.

Bibliography

Baumert, J., Lüdtke, O., Trautwein, U., Brunner, M., 2009. Large-scale assessment studies measure the results of processes of knowledge acquisition: evidence in support of the distinction between intelligence and student achievement. Educational Research Review 4, 165–176.

Baumert, J., Stanat, P., Watermann, R., 2006. Schulstruktur und die Entstehung differentieller Lern- und Entwicklungsmilieus (School structure and the emergence of differential learning environments). In: Baumert, J., Stanat, P., Watermann, R. (Eds.), Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Verlag für Sozialwissenschaften, Wiesbaden, Germany, pp. 95–188.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., Baumert, J., 2012. The differential effects of school traching on psychometric intelligence: do academic-track schools make students smarter? Journal of Educational Psychology 104, 682–699.

Ceci, S.J., 1991. How much does schooling influence general intelligence and its cognitive outcomes? A reassessment of evidence. Developmental Psychology 27, 703–722.

Chmielewski, A.K., Dumont, H., Trautwein, U., 2013. Tracking effects depend on tracking type: an international comparison of students' mathematics self-concept. American Educational Research Journal 50, 925–957.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., York, R.L., 1966. Equality of Educational Opportunity. US Government Printing Office, Washington, DC.

Creemers, B.P.M., 2009. Educational Research for Theory & Practice. Routledge, London, UK.

Eccles, J.S., Midgley, C., Wigfield, A., Buchanan, C.M., Reuman, D., Flanagan, C., Mac Iver, D., 1993. Development during adolescence: the impact of stage-environment fit on young adolescents' experiences in schools and in families. American Psychologist 48, 90–101.

Eccles, J.S., Roeser, R.W., 2011. School and community influences on human development. In: Bornstein, M.H., Lamb, M.E. (Eds.), Developmental Sciences: An Advanced Textbook, sixth ed. Psychology Press, New York, pp. 571–644.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pidegon, D.A., Walker, D.A., 1962. Educational Achievements of Thirteen-Year-Olds in Twelve Countries. UNESCO Institute for Education, Hamburg.

Hanushek, E.A., Lindseth, A.A., 2009. Schoolhouses, Courthouses, and Statehouses. Princeton University Press, Princeton.

Hanushek, E.A., Machin, S.J., Wößmann, L., 2010. Handbook of the Economics of Education, vol. 3. North Holland, Amsterdam.

Hattie, J.A.C., 2009. Visible Learning. A Synthesis of over 800 MetaAnalyses Relating to Achievement. Routledge, Oxon.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., Michelson, S., 1972. Inequality. Basic Books, New York.

Klieme, E., Hartig, J., Rauch, D., 2008. The concept of competence in educational contexts. In: Hartig, J. (Ed.), Assessment of Competencies in Educational Contexts. Hogrefe, Göttingen, pp. 3–22.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., 2013. Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers. Results from the COACTIV Project. Springer, New York, NY.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.T., O'Mara, A.J., Craven, R.G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research. Educational Psychology Review 20, 319–350.

OECD, 2010a. PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science, vol. 1. OECD, Paris.

OECD, 2010b. PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices, vol. III. OECD Publishing, PISA.

Pintrich, P.R., 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. Journal of Educational Psychology 95, 667–686.

Scheerens, J., Bosker, R.J., 1997. The Foundations of Educational Effectiveness. Pergamon, Oxford.

Schunk, D.H., Pintrich, P.R., Meece, J.L., 2008. In: Motivation in Education: Theory, Research, and Applications, third ed. Pearson/Merrill Prentice Hall, Upper Saddle River, NJ.

Suter, L.E., 2000. Is student achievement immutable? Evidence from internationals studies on schooling and student achievement. Review of Educational Research 70, 529–545.

Travers, R.M.W., 1978. Children's Interest. Michigan University, College of Education, Kalamazoo, MI.

Wang, M.C., Haertel, G.D., Walberg, H.J., 1993. Toward a knowledge base for school learning. Review of Educational Research 63, 249–294.

Weinert, F.E., Helmke, A., 1995. Interclassroom differences in instructional quality and interindividual differences in cognitive development. Educational Psychologist 30, 15–20.

Wentzel, K.R., 2009. Students' relationships with teachers as motivational contexts. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Routledge, New York, NY, pp. 301–322.

Schooling: Total Impact of

Hans Luyten, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The total effect of schooling on the cognitive development of young people (its prime objective) has been addressed in a remarkably small number of studies. The available findings indicate that at least 50% of the learning gain across grades can be attributed to the effect of schooling. This figure contrasts sharply with the more frequently cited 'school effect' of only 15%. Some additional questions remain to be settled: for example, to what extent the impact of schooling varies across age ranges, content areas (e.g., language vs mathematics), background (e.g., by socioeconomic status or gender), and (last but not the least) across schools.

Introduction: The Importance of Schooling

Young people spend many hours in school. Thus it seems plausible that the impact of formal schooling on their cognitive development is considerable. However, one cannot assume that all progress made during the school years is due to formal schooling; for example, during their years at primary school children grow approximately 6 cm in length per year, but it seems hardly plausible that classroom instruction has any impact on this growth. In contrast, some cognitive skills developed during the school career are probably acquired in large part through classroom instruction (e.g., spelling and arithmetic), but other skills (e.g., vocabulary) may largely be learned outside school.

The theoretical underpinning of the presumed benefits of education (for personal income and nationwide economic growth) largely stems from the work by human capital economists (e.g., Becker, 1993). In human capital research the results of education are typically measured as qualifications attained or amount (i.e., years) of schooling received. Research in this field strongly focuses on the relationship between academic credentials and success on the labor market. The number of empirical studies that explicitly address the impact of formal schooling on the cognitive development of students (its prime objective) turns out to be surprisingly small.

One might expect to find extensive information on the impact of formal education on cognitive skills in the research literature on educational effectiveness. However, this line of research has traditionally focused on comparing schools, classes, and teachers and has yielded a considerable knowledge base on the variation between schools, classes, and teachers with regard to the outcomes of learning. More importantly, it has produced useful insights on the factors that may account for this variation and provide levers for improvement (Teddlie and Reynolds, 2000). However, assessing the impact of education on the development of children in its own right has never been an issue of major concern in this field.

A well-confirmed finding in educational effectiveness research is the percentage of variance in student achievement scores at the school level. This is usually referred to as the *school effect* and has been found to equal 15% on average over a huge number of studies. It is important to note that this school effect relates to differences in outcomes between schools. It is perfectly conceivable that all schools make a similarly large contribution to student development. This would produce a modest school effect even though the *total effect of schooling* is considerable.

Assessing the Total Effect of Schooling – Two Viable Alternatives

When trying to assess the contribution of schooling to the development of students, one is faced with the challenge that nearly everyone attends school. In order to estimate the contribution, one should compare those who attend school to an equivalent control group that receives no schooling, which is not feasible in practice. Ceci (1991) has reviewed a number of methods that may be helpful in assessing the impact of schooling on cognitive development. Most of these capitalize on exceptional circumstances, for example, (temporary) absence of schooling in remote communities, interrupted schooling during times of war, and changes in mandatory schooling. Studies based on these methods provide compelling evidence for a strong impact of schooling on the development of cognitive skills. Needless to say, such methods cannot be applied on a routine basis.

Regression Discontinuity

According to Ceci (1991), one of the best ways to document the impact of schooling is by comparing children of the same chronological age in different grades. This can be done by applying the logic of the regression-discontinuity design (Trochim, 1984). It capitalizes on the fact that assignment to grades is mainly determined by date of birth. In most education systems, students born just a few days before a certain cut-off date are much more likely to end up in a higher grade than the ones that are just a little younger. Such cut-off dates are clear examples of instrumental variables (Angrist and Pischke, 2009). They affect the score on the explanatory variable of interest (i.e., grade), but are not causally related to the outcome variable. Instrumental variables are sometimes said to produce 'natural experiments.' Thus they allow for unbiased estimates

of the effect that is the focus of the analysis, even when data on confounding variables are not available.

Application of regression discontinuity requires hardly any background information of the students. Besides measures of learning outcomes (e.g., test scores), information on each student's date of birth and actual grade are needed. Two final requirements are that the dataset should relate to students in adjacent grades and that the outcome measures should relate to a common scale, even though the students are in different grades. When students from a wide range of grades (e.g., grades 1–6) are included in the analysis, one cannot expect all students to take the same tests, but this problem can be solved by making use of equated test scores so that all scores can be expressed on a common scale (Verhelst, 2010).

It is also necessary to reflect on the meaning of the effect of education as assessed through regression discontinuity. The effects estimated actually reflect the difference in achievement due to assignment to a higher or lower grade. It should be acknowledged that this possibly includes components that may not be considered effects of education in a more narrow sense. For one thing, the grade effect will also include a peer group effect. It is also conceivable that parents make adaptations in the way they support their child when (s)he is assigned to a higher grade. It should be noted, though, that all these effects are *caused* by assignment to grades. In that sense they still are effects of schooling. However, one may consider these as side-effects of schooling rather than its core effects.

Seasonality of Learning

If one assumes that during the school year learning gain is affected by both school and nonschool factors, whereas only nonschool factors are at work during the summer vacation, the effect of schooling can be estimated by means of a comparison between the per month learning gain during the school year and the summer vacation (Heyns, 1978). Although this approach seems useful for estimating the total effect of schooling, it is also important to point to some problems.

The main problem relates to the assumption that the summer period presents a useful alternative for the control condition in a randomized trial. This is valid if we can be sure that the impact of nonschool factors (e.g., parental support) during the summer vacation provides a reliable indication of their impact in the (hypothetical) absence of schooling, but it is conceivable that certain parents (e.g., the well-educated ones) engage more in academically stimulating interactions with their children than they would otherwise during the summer vacation. It remains an open question whether the summer period presents an equally viable alternative for a control condition for all students. Another implicit assumption of the approach is that the effect of schooling stops as soon as the summer vacation starts. Most teachers would hope that their efforts have a more enduring effect and that learning still continues out of school due to their teaching efforts.

The main focus in seasonality research has been on the summer gain/loss itself and particularly on inequalities in this respect. The bulk of the studies have been conducted in the USA (probably because of the long summer vacations).

Findings

Numerous studies on the seasonality of learning have consistently shown much stronger gains during the school year than during the summer vacation. Sometimes zero learning gains and learning losses have been reported for the summer period. A consistent finding is the large variation in learning gains during the summer vacation. When school is in session, the differences in cognitive growth between students are much smaller than during periods without schooling. Especially students from disadvantaged backgrounds appear to make relatively little progress during the summer period. This suggests that education presents an equalizing force of considerable importance (Downey et al., 2004; Alexander et al., 2001).

The number of studies based on regression discontinuity is much smaller, but these studies focus explicitly on the total effect of schooling. One of the first well-known studies was conducted by Cahan and Davis (1987) and relates to mathematics and reading comprehension in grades 1 and 2 of primary education in Israel. The findings indicate that about $^{2}/_{3}$ of the difference in achievement between grade 1 and 2 students is due to schooling versus $^{1}/_{3}$ due to the 1-year difference in age. A study using the same approach to assess the impact of schooling on IQ development of Israeli children in grades 4 to 6 (Cahan and Cohen, 1989) reports larger effects for verbal than nonverbal tests. It relates to 12 different tests and for nine of these the effect of schooling outweighed the age effect.

A more recent study by Luyten (2006) applied regression discontinuity within a multilevel framework in a secondary analysis of TIMSS-95 data (http://timssandpirls.bc.edu/) for eight different countries. The effect of schooling on mathematics achievement in grades 3 and 4 of primary education was modeled as random at the school level. This approach yields both an assessment of the average effect of one year of schooling and an estimate of its variation across schools. In line with the Israeli studies, this study reports larger effects of schooling than age in most countries. It also reveals that in most countries a substantial minority of schools could be discerned that failed to show a positive effect of being in a higher grade (4 vs 3).

Most studies based on regression discontinuity assume a sharp cut-off date, which implies that students with delayed or accelerated school careers are excluded from the analyses. A study by Luyten and Veldkamp (2011) applies a methodology that does not assume a sharp cut-off date and takes into account the effect of unmeasured factors that affect both learning outcomes and assignment to grade (i.e., delay or acceleration of school careers). Their reanalysis of TIMSS-95 data includes 15 countries. They report that for mathematics and science, respectively, 54% and 47% of the differences between grades are due to schooling. When expressed as an effect size (Cohen's d), the average effect of grade level across 15 countries is 0.46 and 0.35 for mathematics and science, respectively. By themselves these are not particularly large effect sizes, but if they apply to all grades in primary education, the total effect for all six grades would fall between 2.00 and 3.00, whereas 0.80 is generally considered a very large effect.

Concluding Remarks

Although the total effect of schooling on the development of students has been the explicit focus of a limited number of studies, there can be no doubt about its impact. The effect of schooling appears to be clearly stronger than the effect of physical maturation and at least half of the cognitive growth that occurs during primary education should be considered an effect of schooling. This implies that the effect of schooling is much larger than the 'school effect' (i.e., 50% vs 15%). The latter relates to differences in outcomes between schools. Studies on the seasonality of learning suggest that formal education works as an equalizing force. Disparities in knowledge and skills tend to widen mostly during periods when school is not in session.

Nevertheless some additional questions remain to be settled. It is unclear to what extent the effect of schooling varies over different age ranges. It seems likely that the effect of age decreases as children grow older. How far this applies to the effect of schooling is still an open question. Whether the effect of schooling differs for students from various backgrounds (e.g., low vs high SES) has hardly been addressed, although the findings from seasonality of learning suggest that the effect is similar for all students. Last but not the least, little is known about the variation of the effect of schooling across schools. The only study that has addressed this issue (Luyten, 2006) suggests substantial variation between schools, even though the effect of schooling appears to be considerably larger than the school effect.

See also: Data Bases and Statistical Systems: Education, Statistical Systems; Education, Economics of; School Effectiveness Research.

Bibliography

Alexander, K.L., Entwisle, D.R., Olson, L.S., 2001. Schools, achievement and inequality. A seasonal perspective. Educational Evaluation and Policy Analysis 23, 171–191.

Angrist, J.D., Pischke, J.S., 2009. Mostly Harmless Econometrics, an Empiricists Companion. Princeton University Press, Princeton.

Becker, G.S., 1993. Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, third ed. Chicago University Press, Chicago.

Cahan, S., Cohen, N., 1989. Age versus schooling effects on intelligence development. Child Development 60, 1239–1249.

Cahan, S., Davis, D., 1987. A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. American Educational Research Journal 24, 1–12.

Ceci, S.J., 1991. How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. Developmental Psychology 27, 703–722.

Downey, D.B., Von Hippel, P.T., Broh, B.A., 2004. Are schools the great equalizer? Cognitive inequality during the summer months and the school year. American Sociological Review 69, 613–635.

Heyns, B., 1978. Summer Learning and the Effects of Schooling. Academic Press, New York.

Luyten, H., 2006. An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. Oxford Review of Education 32, 397–429.

Luyten, H., Veldkamp, B., 2011. Assessing effects of schooling with cross-sectional data: between-grades differences addressed as a selection-bias problem. Journal of Research on Educational Effectiveness 4, 264–288.

Teddlie, C., Reynolds, D., 2000. The International Handbook of School Effectiveness Research. Falmer Press, London, New York.

Trochim, W.M.K., 1984. Research Design for Program Evaluation, the Regression-Discontinuity Approach. SAGE Publications, London.

Verhelst, N., 2010. IRT models. Parameter estimation, statistical testing and application in EER. In: Creemers, B.P.M., Kyriakides, L., Sammons, P. (Eds.), Methodological Advances in Educational Effectiveness Research. Routledge, London, New York, pp. 183–218.

Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance

Christian Dayé, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'schools' in social and behavioral sciences (SBS) refers to two different types of concepts: the school of thought, which consists mainly of a coherent set of interrelated ideas and positions, and the concept of the school as institution, which takes in view not only a system of ideas, but also the social organization in which this system is developed and stabilized. The article introduces theoretical approaches to these concepts and discusses their relevance for the history of SBS. Both concepts of schools imply that SBS disciplines are not characterized by a consensus on theoretical and methodological principles, but must be conceived of as multiparadigmatic.

Introduction

A variety of ideas and concepts are associated with the notion of schools in social and behavioral sciences (SBS). Basically, the main thrust of all use of the term is to argue that a given field is not characterized by homogeneity as regards theory, methodology, or research practice. Instead, a variety of lines of reasoning, or paradigms, coexist under the umbrella of the discipline. This, in turn, implies that the intellectual development of a discipline or field at large cannot be understood as showing patterns of linear accumulation of knowledge. Rather, the intellectual development is a complex product emerging of the interplay between the schools in the field.

Four main concepts of schools in SBS can be discerned: (1) a typological, (2) a psychological, (3) an institutional, and (4) a concept of 'national schools' (cf Szacki, 1975).

The typological and the psychological concept can be seen as discernible though partly complementary dimensions of what is often called a 'school of thought'. (1) The thrust of a typological understanding is to classify scholars or scientists according to the ways they work and think, "independent of whether they maintained any contacts or not, and whether they would have themselves included in that particular 'school'" (Szacki, 1975: p. 176). (2) When, in contrast, a scholar or researcher defines himself or herself as belonging to a school of thought, we have an instance of a psychological concept of a school. The feeling of being not alone in following a certain set of ideas and convictions might have psychological benefits, even if the persons thus related never actually meet. Proponents of new approaches can draw profit from finding principles of their approach adumbrated or anticipated in the works of the founders of their disciplines. Such relations help to stabilize new approaches in the decisive early phases, mostly because they appear to justify the establishment of a new field within a discipline (or at its borders). For example, the references to Georg Simmel in the literature on social network analysis can be understood by the psychological benefits arising from defining oneself as member of an already existing school of thought, even if the person who formulated the school's basic tenets is long deceased. An interesting phenomenon is that in some cases, the psychological concept predated the typological. Bronisław Malinowski's invoking of a functionalist school in his early writings crucially contributed to functionalism becoming established typologically as a school of thought in SBS.

Opposed to schools of thought, more sociological or cultural uses of the term can be found in the historical literature. The main types here are (3) concepts based on an institutional argument and (4) concepts focusing on national schools. (3) Basically, an institutional concept of school implies that the "interests, assumptions, theses, etc., common to a given group of scholars are formed and consolidated within one institutional framework" (Szacki, 1975: p. 174). (4) To speak of national schools in SBS, finally, suggests that institutional and cultural environments within a nation fundamentally and crucially informed the basic tenets of a given discipline within this country. This concept implies that there is, to a certain extent, homogeneity within a discipline all across the country and that the links between the institutional and cultural habitat of a science within a nation-state and its theoretical and methodological peculiarities can be specified. Schools that at first sight appear as national schools, for instance the Austrian School in economics, might actually be better treated as schools of thought.

With regard to SBS disciplines and fields, the most relevant understandings of the notion of school were typological and institutional concepts. The following two sections of this article focus on these. An adjacent section focuses on the methodological implications of the discussed concepts in order to compare them to the neighboring concepts of invisible colleges (Crane, 1972), theory groups (Mullins, 1973), and academic tribes (Becher and Trowler, 2001). The concluding section attempts an outlook on the future of the various school concepts.

Schools of Thought

Characteristics

Speaking of schools of thought presupposes both an internal coherence and an external distinctiveness of their system of thought (Negru, 2013). Internal coherence refers both to the internal logic of a system of thought and to the homogeneity of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03010-5

the various contributions to this system, although the latter aspect should not be taken to say that there is no disagreement or conflict among various members of a school. Rather, homogeneity is a matter of degree. Furthermore, the system of thought must be sufficiently distinctive from other traditions. Such system of thought might comprise ontological, epistemological, methodological, ethical, and axiological, as well as ideological elements. The importance of a coherent and distinctive system of thought discerns schools of thoughts from the neighboring concept of the scientific community (Hagstrom, 1965).

Several sources might provide a school with a coherent and distinctive system of thought. One prominent source, and one close to the original meaning of the term 'school', is the teacher: an individual who developed a doctrine and is regarded as a model by his pupils. Durkheimian sociology is one example, Keynesian economics another. Another source is broad theoretical frameworks, like structuralism, functionalism, or symbolic interactionism. Further, research techniques and methods can also provide sufficient coherence and distinctiveness to become the intellectual core of a school of thought. It should be emphasized that schools of thought are not restricted to one field or discipline but can, like e.g., functionalism, spread across a variety of fields, and although the actual shape of the system of thought can show differences from discipline to discipline, the basic tenets can be broad and flexible enough to guarantee coherence. But that need not be the case: cross-disciplinary schools of thought can disintegrate into several disciplinary schools, if the differences among disciplinary proxies become too obvious.

How schools of thought are established and, conversely, why some schools of thought fail to do so are questions of vital relevance to the history of SBS. The second question has been explored by Neil McLaughlin (1998) with regard to Neo-Freudianism. Neo-Freudianism was a culture-oriented reformulation of Freudian psychoanalysis that emerged in the 1930s. In the following two decades, it became highly influential among both mental health practitioners and scientists from the SBS, but also left its imprint on the broader intellectual culture. But then, "attacks from members of the orthodox psychoanalytic establishment increasingly isolated" this line of thinking (McLaughlin, 1998: p. 113). Its influence declined, and by the 1980s, there was no Neo-Freudian school of thought. The failure of Neo-Freudianism to establish itself lastingly as a school of thought cannot be explored solely by personal differences between the main protagonists, Karen Horney, Erich Fromm, and Harry Stack Sullivan. Rather, the "goals, ideas, politics, and institutional and sociological positions of the major neo-Freudians were too different from each to make a school of thought possible" (McLaughlin, 1998: p. 128f.). The system of thought was distinctive, but not coherent enough. Its center was a sophisticated critique of Freudian thinking, which nonetheless did not lead to constructive theoretical or methodological developments that Horney, Fromm, and Sullivan could unanimously share. This was if not caused, then aggravated by the different positions the protagonists occupied – Horney being a therapist with "a proto-feminist bent", Sullivan a "brilliant and charismatic clinician […] who understood the importance of institution building and selling psychiatry to policymaking elites", and Fromm a "renegade sociologist and more of a prophetic figure than an academic expert or even a clinician" (McLaughlin, 1998: p. 129). Also, personal tensions among the proponents have played a role.

Historical Relevance

Assessing the relevance of schools of thought for the history of SBS is difficult because the answer must address both the levels of historical fact and of the historiography of a field. Some SBS disciplines show a tradition of producing textbooks that are organized along schools of thought in a typological understanding. This is most notably the case with psychology: taking E.G. Boring's *A History of Experimental Psychology* (1929) as their model, textbooks like Robert S. Woodworth's *Contemporary Schools of Psychology* (1931) and Edna Heidbredder's *Seven Psychologies* (1933) treated the field as organized into schools (cf Green et al., 2013: p. 168). Though not always with explicit use of the term, and sometimes in combination with topic-oriented chapters (e.g., Goodwin, 2012), the presentation of psychology as a multitude of mostly coexisting schools of thought – Wundtian experimental psychology, structuralism, functionalism, gestalt psychology, behaviorism, and so forth – has been an important pattern in psychology textbooks throughout the twentieth century. This also holds true with those SBS that are overtly multiparadigmatic, most prominently with economics (e.g., Dow, 1996; Mair and Miller, 1991; Snowdon et al., 1994), sociology (Martindale, 1960; Sorokin, 1928), and political science (Bleek and Lietzmann, 1999), and occasionally with subfields such as interpretive social research (Scott and Keates, 2001) or international relations (Thompson, 1996).

Schools of thought have thus been historically influential first and foremost as didactical devices that help beginners demarcate the field of study. Textbooks that are organized according to schools convey a mental map of the field, in which a small number of intellectual gravitational centers, usually not more than a dozen, are located. In theoretical and methodological debates in the classroom and beyond, schools of thought then function as a proxy for discursive positions that may help to clarify an argument. Major battles evolve around the inconsistencies and conflicts between schools and define a good part of the shape of the discipline or field.

Discussion

The criticism most often voiced against the use of the concept of schools of thought is that it is a Procrustean bed. It cuts off important differences and idiosyncrasies of authors in order to fit a banalized version of their thought into a predesigned scheme. (This charge has also been voiced against Nicholas Mullins' (1973) description of theory groups; see below.) The usefulness of such an artificial concept for historical analyses of the development of SBS has been disputed from a variety of methodological and theoretical orientations, and some authors sought to develop an institutional concept of schools to overcome the shortcomings of the concept of schools of thought.

Schools as Institutions

Characteristics

As a rule of thumb, an author refers to an institutional concept of school if the school's name includes the name of its location. To take only some examples from the history of economics, literature locates schools in Cambridge (Backhouse, 2006; Pasinetti, 2005), Chicago (Emmett, 2009; van Overtveldt, 2007), Kiev (Sušjan, 2010), and Stockholm (Erixon, 2011), among other places. Of course, the degree to which each of these conforms to the criteria of an institutional concept must be established by detailed analysis, and this requires a workable conceptualization of schools as institutions. Most conceptual work in this area comes from sociology, with other SBS disciplines contributing to it.

One of the most elaborated conceptualizations of schools as institutions has been proposed by Edward A. Tiryakian (1979). A school consists of people who are in relatively close contact to each other and form "a real group of intellectuals, a small community of persons whose origin and formative period can be localized in time and place" (Tiryakian, 1979: p. 216). Sociologically speaking, they share several characteristics with religious communities. The basic differentiation within a school is between a founder-leader and his followers, who are committed to the doctrines of the founder-leader and united in the mission of establishing them. This hierarchy discerns schools from formally more egalitarian collaborative circles (Farrell, 2001). The founder-leader has developed to a certain point a system of thought, and against the orthodoxies and established views of the field, he or she believes in the value and viability of his or her system. Usually, he or she is a charismatic person, a fact that further strengthens the commitment of the followers. Schools as institutions thus are a collective and mediating entities located between the individual researcher and the scientific community.

Among the followers, various roles can be discerned. One of these is the interpreter. "Very often, the founder-leader is one who, because he has such an intense, innovative way of conceptualizing the reality of the world, has trouble expressing himself in ordinary language" (Tiryakian, 1979: p. 219). To convey the message in a way accessible to those not so close to the founder-leader becomes the task of the interpreter, or popularizer. A second role is the pragmatic organizer, who takes care of the school's day-to-day activities. A further group of followers potentially decisive for the success of the school are the converts, scholars who received their training elsewhere (and might thus have the same age as the founder-leader) but later became convinced of the surpassing quality of the founder-leader's worldview. Their contacts to the field outside the school can prove vital in disseminating its doctrines. Another crucial group are the students, or lieutenants, some of whom might become the leaders of the school after the departure of the founder-leader and the converts, from whom they take over the task of promoting the institutionalization of the doctrine in the discipline at large. Further positions include the auxiliary, a 'loyal foot soldier' assisting in promoting the school's visibility, but without any original intellectual contribution; the patron, who does not fully participate in the intellectual debates within the school, but provides material and financial assistance; and the confidant, an individual personally close to the founder-leader who provides emotional, but not intellectual, support.

Beyond these aspects of the social structure of the school, other criteria have to be met by a school in order to be able to produce works of scientific eminence. A school needs an institutional affiliation; the higher in the rankings of academia, the better, because of the social and cultural capital related to such an affiliation. Further, the school needs some form of regular publication, usually one or two journals or book series. Also, a manifesto is required, a "document in the nature of a professional proclamation of [the schools] basic mode of perceiving and relating to the world" (Tiryakian, 1979: p. 223).

Tiryakian also offers some thoughts on the trajectories of schools, which roughly resemble Nicholas Mullins' characterization of developmental stages of what he calls 'theory groups' (Mullins, 1973). The distinctiveness of the system of thought developed by the school attracts both attention and new members. This leads to an increased visibility of the school within academia. Its ideas become popularized and are melted into the standard body of knowledge of the discipline. Also, the "charisma of the school becomes 'routinized'; […] The more the school and its achievements become integrated into the discipline, the more depersonalized becomes its paradigm" (Tiryakian, 1979: p. 218). This depersonalization, on the one hand, leads to a higher freedom in using and transforming the ideas. On the other hand, their specific sociocultural origins slowly fall into oblivion, as do those presuppositions that did not find their way into the written products of the school – basic ontological, metaphysical, existential, and perhaps ideological groundings of a paradigm. "Over time, and with an increased and perforce more heterogeneous membership, the presuppositions of the school become blunted, diluted, trivialized, or compromised" (Tiryakian, 1979: p. 218).

Apart from identifying schools and roles within them, historical analyses of institutional schools usually address one of two more general problems. One relates to the consequences of the existence of such schools for the development of knowledge. This issue is treated in more detail in the next section. The other more general set of problems is concerned with the importance of local practices in research. The separation of these two problems is mainly of analytical value. Of course, these two problems are interrelated, because they both converge on the importance of tacit forms of knowledge (Polanyi, 2005), of knowledge that does not travel easily on written documents but is acquired through model-learning or hands-on experience.

Historical Relevance

Tiryakian conceives of schools as being oriented toward the discipline at large. The mission of its members is to spread the doctrines of the founder-leader. Granted this (which is disputed; see next section), schools can be identified as important units in the development of SBS, firstly because they are key agents in the process of academic institutionalization of a new discipline. Secondly, they play a paramount role in the

cognitive development of a discipline. Because "[t]he articulation and demonstration of the efficacy of a new methodology are much more a communal affair than that of a single individual" (Tiryakian, 1979: p. 211), schools are the places where innovative paradigms are crafted. The growth of sociology "to a large measure represents, or is a function of, periodic infusions brought about by a small number of major 'schools'" (Tiryakian, 1979: p. 211).

While the justification part of the above explanation – that new paradigms usually are the product of a collective – is certainly correct, the assertion that a school is the most important social form capable of uniting scholars to collaborate on the articulation and presentation of a system of thought is dubious. All the more interesting is the striking similarity of Tiryakian's argument to the coeval discourse in Marxist philosophy of science. Referring to the notion of a scientific and technological revolution established by Radovan Richta in the 1960s, the editors of a book on scientific schools that emerged from a series of colloquia among researchers from the Academies of Science in Moscow and East Berlin wrote in the introduction that "the task to investigate the collective creativity in science and research and lead it to its highest efficiency gains crucial importance for the acceleration of the scientific and technological progress" (Mikulinskij et al., 1977b: p. 7, vol. I; see also Mikulinskij et al., 1977a). To them, scientific schools are paramount for scientific progress, because they are "a type of organization that realizes the dialectical unity of individual and collective epistemic progress" (Mikulinskij et al., 1977b: p. 9, vol. I). While it will not be possible to create schools by *fiat*, "the comprehensive, but partly widely scattered material on the past and present of scientific schools must be collected and analyzed to generate recommendations for the design of creative work under the conditions of socialized science in the developed socialist society" (Mikulinskij et al., 1977b: p. 9, vol. I).

It has been noted that the debate on scientific schools has been more widespread in formerly socialist countries (Stichweh, 1999: p. 23f.). Philosophers and sociologists of science in East Europe and Russia were indeed cognizant of the works of English-speaking scholars, while the opposite has not always been true. The criteria for defining a school developed by Eastern scholars, however, do not differ much from Tiryakian's; for instance, Eduard Michajlovič Mirskij defines the following characteristics of a school (in the natural sciences): (1) it is a coalition of researchers with a high-profile leader which locates itself within the wider scientific community; (2) it exists as an autonomous entity which is in opposition to the wider community both with regards to the content and procedure of scientific work and to scientific standards and norms; (3) the research program proposed by the school differs from those proposed elsewhere; (4) its existence is caused by a set of factors that include objective factors, like the research program, as well as subjective factors, like the personality of its key members; (5) the research program must be productive, i.e., it must be capable of being developed; (6) the scientific authority of the leader of the school must be acknowledged by the academic environment; (7) both composition and growth of a school depend on the organizational capabilities of its leader; and (8) the value of a school is not only to be seen in its research program. The existence of a school also contributes to the maintenance of a continuous "interest in the explication and critical discussion of problems of theory and methodology within a given research area […] [T]he scientific school is virtually the only type of environment which *forces each member* to consider critically the tenets common in his science and to see them as mere *conventional assumptions of a specific group of scientists*" (Mirskij, 1977: p. 188f.). Despite the apparent consensus about the core features of schools as institutions, there is a richness of interesting and nuanced differences in their treatment of the notion of school by Marxist authors (e.g., Gasilov, 1977; Laitko, 1977). Furthermore, all the cited authors agree that the importance of the formation and existence of schools with regard to the growth of knowledge cannot be underestimated, be it in the SBS or in other areas of scientific endeavor.

Discussion

The supposed existence of schools as institutions in academic life has often led to severe criticisms. Informed in part by Robert K. Merton's (1996) treatment of the ethos of science, and in striking contrast to the apparent consensus on the positive functions of schools for the development of science described above, scholars have argued that schools are closed and introverted systems characterized by "the uncritical acceptance by a group of disciples of a leader's idea system" (Krantz, 1972: p. 86). They emphasize a sectarian meaning of the notion of school, arguing that schools reject external influence and validation of their work. The creation of new publication outlets by the school is seen as a measure to "bypass the criticism of referees from other areas" (Crane, 1972: p. 87). The existence of schools invariably leads to either a field-internal segregation or a prolonged confrontation; especially in the latter case, these authors expect the growth of the field to be inhibited.

The viability of an institutional conceptualization of schools within the history of SBS has also confronted criticism, and its use has declined in the past decades. Most criticism focused on the supposed stability within schools, both with regard to the internal hierarchization and role distribution, but also with regard to the system of thought and the commitment of the individual members to it. A school is more fragile an institution and individuals more ambivalent toward their own membership in them than these concepts suggest. Also, it is not easy to demarcate a school as historical entity. A rather recent, highly sophisticated approach has been developed by Andrew Abbott (1999), who argued that the first Chicago school (A.W. Small, R.E. Park, E.W. Burgess, W.I. Thomas, L. Wirth, etc.; cf Bulmer, 1984) in sociology can be treated as a historical entity for its social structural consequences. These initiated a complex process of mutual reinforcement, which did not only intensify the structural consequences, but out of which also emerged the Chicago school as a cultural object. "The Chicago School was not a fully formed cultural object until after it had virtually completed its life as a social structure." The cultural object, thus, is to a high degree a retrospective creation of the second Chicago school period after World War II (E.C. Hughes, H. Blumer, W.L. Warner, etc.; cf Fine, 1995).

Methodological Considerations

The specificities of the two concepts of school discussed above come into view best if contrasted with neighboring concepts with a focus on the respective implications for the design of a study in the history of SBS. In the case of typological schools, the most obvious are analyses of textbooks (Platt, 2008). Recently, another research design has been suggested by Green et al. (2013). Their approach has been to quantify the level of verbal similarity between pairs of articles published in a predefined period in a selected journal. Such methodologically sophisticated analysis has become less difficult in the course of the digitalization of scientific knowledge.

Comparable studies of communication processes in science have also informed many of those concepts that can be seen as neighboring the institutional concept of school, such as 'invisible colleges' and 'theory groups'. Invisible colleges are communication networks that link the leaders of local research groups and usually comprise about 100 scientists (Price, 1963). These leaders "communicate with each other and transmit information informally across the whole field. This enables them to monitor the rapidly changing research 'front' and to keep up with new findings during a period of rapid growth" (Crane, 1972: p. 35). The concept of theory groups considers a broader variety of social relationships between individual scientists and includes a developmental perspective on the institutionalization processes of a theory group (Mullins, 1973). Mullins's conceptualization offers a fruitful empirical framework that, apart from a quantitative analysis of communication patterns, also investigates many features of the above characterization of schools (basic manifesto, intellectual leader, organizational leader, etc.); however, as his own analyses show, this framework can also be too ambitious.

Analyses of schools as institutions usually rely on historical and archival work in search for sources that document the internal organization of the school. The objective of such research is to gain insights into the cultural life of the school, and to show how this is reflected in the works of the members. Citation analyses and the like are only of secondary relevance. This also holds true for the concept of academic tribes (Becher and Trowler, 2001). By academic tribes, the authors understand the distinctive cultures within academic communities. The emphasis is on the relation of these cultures to the academic ideas they propagate – the territories, in Becher and Trowler's words. The relationship between the two "involves a mutually dependent interplay of, on the one hand, the structural force of the epistemological character of disciplines that conditions culture and, on the other, the agentic capacity of individuals and groups for autonomous action, including interpretative acts" (Becher and Trowler, 2001: p. 23f.). In both cases, thus, the analysis is more on the cultural aspects of academic life and its relations to the prevailing ideas within these social entities. The methodologies used in this line of work are predominantly of an interpretative kind.

Conclusion

A brief outlook on the future of the concepts here described might be appropriate. In general, there is reason to assume that the notion of schools of thought, due to their didactic function and their role in theoretical debate, will remain in use, though this does not necessarily apply for the term itself. The trajectory of the institutional concept of schools is dubious. A variety of factors that characterize contemporary academic life appear to counteract the establishment of schools: the increasing amount of temporary contracts makes it hard to maintain a stable core of collaborators; the standardization of academic education and the internationalization of the academic labor market transform the decision to commit to an idiosyncratic (though perhaps highly innovative) paradigm into a highly critical career decision, because of the potential lack of connectability of one's own work to the 'mainstream'; the mere population size of SBS disciplines has increased dramatically during the last century, so that the attention space available for alternative paradigms decreases; and finally, the internationalization of the SBS and the gradual erosion of national academic spaces homogenize the academic landscape, so that if institutional schools will emerge in the future, they will probably do so in the metropolitan centers of this homogenized, global landscape.

See also: Academic Research and Employment: Recent Changes in Europe and the United States; Discipline Formation in the Social Sciences; Scientific Disciplines, History of.

Bibliography

Abbott, A., 1999. Department and Discipline. Chicago Sociology at One Hundred. The University of Chicago Press, Chicago, London.

Backhouse, R.E., 2006. Sidgwick, Marshall, and the Cambridge school of economics. History of Political Economy 38, 15–44.

Becher, T., Trowler, P., 2001. Academic Tribes and Territories: Intellectual Enquiry and the Culture of Disciplines, second ed. Open University Press, Buckingham, Philadelphia, PA.

Bleek, W., Lietzmann, H.J. (Eds.), 1999. Schulen der deutschen Politikwissenschaft. Leske & Budrich, Opladen.

Boring, E.G., 1929. A History of Experimental Psychology. Century, New York.

Bulmer, M., 1984. The Chicago School of Sociology. Institutionalization, Diversity, and the Rise of Sociological Research. The University of Chicago Press, Chicago, London.

Crane, D., 1972. Invisible Colleges: Diffusion of Knowledge in Scientific Communities. The University of Chicago Press, Chicago, London.

Dow, S.C., 1996. The Methodology of Macroeconomic Thought: A Conceptual Analysis of Schools of Thought in Economics. Edward Elgar, Aldershot.

Emmett, R.B., 2009. Frank Knight and the Chicago School in American Economics. Routledge, London.

Erixon, L., 2011. A social innovation or a product of its time? The Rehn–Meidner model's relation to contemporary economics and the Stockholm school. The European Journal of the History of Economic Thought 18, 85–123.

Farrell, M.P., 2001. Collaborative Circles: Friendship Dynamics and Creative Work. The University of Chicago Press, Chicago, London.

Fine, G.A. (Ed.), 1995. A Second Chicago School? The Development of a Postwar American Sociology. The University of Chicago Press, Chicago.

Gasilov, V.B., 1977. Analyse der interpretation des terminus "wissenschaftliche Schule". In: Mikulinskij, S.R., Jaroševskij, M.G., Steiner, H., Winkler, R.-L. (Eds.), Wissenschaftliche Schulen, 2 Bde. Akademie Verlag, Berlin, pp. 291–321.

Goodwin, C.J., 2012. A History of Modern Psychology, fourth ed. John Wiley & Sons, Hoboken, NJ.

Green, C.D., Feinerer, I., Burman, J.T., 2013. Beyond the schools of psychology I: a digital analysis of psychological review, 1894–1903. Journal of the History of the Behavioral Sciences 49, 167–189.

Hagstrom, W.O., 1965. The Scientific Community. Southern Illinois University Press, Carbondale, Edwardsville.
Heidbredder, E., 1933. Seven Psychologies. Appleton-Century-Crofts, New York.
Krantz, D.L., 1972. Schools and systems: the mutual isolation of operant and non-operant psychology as a case study. Journal of the History of the Behavioral Sciences 8, 86–102.
Laitko, H., 1977. Der Begriff der wissenschaftlichen Schule – theoretische und praktische Konsequenzen seiner Bestimmung. In: Mikulinskij, S.R., Jaroševskij, M.G., Steiner, H., Winkler, R.-L. (Eds.), Wissenschaftliche Schulen, 2 Bde. Akademie Verlag, Berlin, pp. 257–290.
Mair, D., Miller, A., 1991. A Modern Guide to Economic Thought: An Introduction to Comparative Schools of Thought in Economics. Edward Elgar, Aldershot.
Martindale, D., 1960. The Nature and Types of Sociological Theory. Houghton Mifflin, Boston.
McLaughlin, N.G., 1998. Why do schools of thought fail? Neo-freudianism as a case study in the sociology of knowledge. Journal of the History of the Behavioral Sciences 34, 113–134.
Merton, R.K., 1996. The ethos of science. In: Sztompka, P. (Ed.), On Social Structure and Science. The University of Chicago Press, Chicago, pp. 267–276.
Mikulinskij, S.R., Jaroševskij, M.G., Kröber, G., Steiner, H. (Eds.), 1977a. Shkoly V Nauke. Nauka, Moskva.
Mikulinskij, S.R., Jaroševskij, M.G., Steiner, H., Winkler, R.-L. (Eds.), 1977b. Wissenschaftliche Schulen, 2 Bde. Akademie Verlag, Berlin.
Mirskij, E.M., 1977. Die naturwissenschaftliche Schule im System der wissenschaftlichen Tätigkeit. In: Mikulinskij, S.R., Jaroševskij, M.G., Steiner, H., Winkler, R.-L. (Eds.), Wissenschaftliche Schulen, 2 Bde. Akademie Verlag, Berlin, pp. 169–191.
Mullins, N.C., 1973. Theories and Theory Groups in Contemporary American Sociology. Harper & Row, New York.
Negru, I., 2013. Revisiting the concept of schools of thought in economics: the example of the Austrian school. American Journal of Economy and Sociology 72, 983–1008.
van Overtveldt, J., 2007. The Chicago School: How the University of Chicago Assembled the Thinkers Who Revolutionized Economics and Business. Agate, Chicago.
Pasinetti, L.L., 2005. The Cambridge school of Keynesian economics. Cambridge Journal of Economics 29, 837–848.
Platt, J. (Ed.), 2008. Introductions to Sociology: History, National Traditions, Paedagogies. Current Sociology Monograph. Sage, London.
Polanyi, M., 2005. Personal Knowledge. Towards a Post-Critical Philosophy. Routledge, London.
Price, D.J.S., 1963. Little Science, Big Science. Columbia University Press, New York.
Scott, J.W., Keates, D. (Eds.), 2001. Schools of Thought: Twenty-Five Years of Interpretive Social Science. Princeton University Press, Princeton.
Snowdon, B., Vane, H., Wynarczyk, P. (Eds.), 1994. A Modern Guide to Macroeconomics: An Introduction to Competing Schools of Thought. Edward Elgar, Aldershot.
Sorokin, P.A., 1928. Contemporary Sociological Theories. Harper Brothers, New York.
Stichweh, R., 1999. Zur Soziologie wissenschaftlicher Schulen. In: Bleek, W., Lietzmann, H.J. (Eds.), Schulen der deutschen Politikwissenschaft. Leske & Budrich, Opladen, pp. 19–32.
Sušjan, A., 2010. Historicism and neoclassicism in the Kiev school of economics: the case of Aleksander Bilimovich. Journal of the History of Economic Thought 32, 199–219.
Szacki, J., 1975. “Schools” in sociology. Social Science Information 14, 173–182.
Thompson, K.W., 1996. Schools of Thought in International Relations: Interpreters, Issues, and Morality. Louisiana State University Press, Baton Rouge.
Tiryakian, E.A., 1979. The significance of schools in the development of sociology. In: Snizek, W.E., Fuhrman, E.R., Miller, M.K. (Eds.), Contemporary Issues in Theory and Research. A Metasociological Perspective. Aldwych Press, London, pp. 211–233.
Woodworth, R.S., 1931. Contemporary Schools of Psychology. Ronald, New York.

Schools, Micropolitics of

Herbert Altrichter and Robert Moosbrugger, Johannes Kepler University, Linz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Micropolitics' is a specific perspective in organization theory. It focuses on activities within organizations to acquire, develop, and use power and other resources. This article explores the uses and usefulness of this concept in educational research. It starts with outlining the micropolitical view on educational organizations, then gives some examples of representative micropolitical research in order to illustrate typical topics and research strategies. The article concludes with discussion on criticisms of micropolitics and indicating areas for future development.

'Micropolitics' is a specific perspective in organization theory. It focuses on "those activities taken within organizations to acquire, develop, and use power and other resources to obtain one's preferred outcomes in a situation in which there is uncertainty or dissent" (Pfeffer, 1981: p. 7) because it considers them as most important for the constitution and the workings of organizations.

This article explores the uses and usefulness of this concept in educational research. It starts with outlining the micropolitical view on educational organizations. Then, it gives some examples of representative micropolitical research in order to illustrate typical topics and research strategies. Finally, it discusses criticisms of micropolitics and indicates areas for future development.

The Concept of Micropolitics

Main Elements

Traditional organizational theories of different origins seem to converge in describing organizations as goal-oriented and rationally planned, characterized by stable objective structures, a high degree of integration and a common purpose of its members. Conflicts between the members of an organization are considered costly and irrational 'pathologies' to be eradicated as soon as possible. This view on organizations which has also been used for studying educational establishments suggests "that it is possible within schools to arrive at a time and place where politics and conflict are not present" (Flessa, 2009: p. 340).

However, conflicts are endemic in organizational life (Ball, 1987), they are a "naturally occurring phenomenon" (Achinstein, 2002: p. 422). Different stakeholders pursue their own interests which are not necessarily identical with the formulated goals. Coalitions are formed, meetings are boycotted, and formal structures are ignored. As such phenomena frequently occur in organizations, it is not helpful to exclude them from organizational theorizing: models are needed which can cope with this "dark side of organizational life" (Hoyle, 1982: p. 87) and take account of the complexities "in managing a diversity of beliefs and practice while maintaining strong communities" (Achinstein, 2002: p. 422). Such models also offer the chance to examine and learn from conflict (see Flessa, 2009: p. 331; Achinstein, 2002: p. 422). They challenge traditional rational theories of organizations by highlighting the difference between 'policy rhetoric' and reality (Björk and Blase, 2009: p. 198); what micropolitics is interested in is "the study of how things *really* work, not how an organizational chart or a principal's action plan would like them to work" (Flessa, 2009: p. 331).

The micropolitical perspective was originally developed in the realm of profit organizations (see e.g., Bacharach and Lawler, 1980; Pfeffer, 1981). In a 1982 seminal paper, Eric Hoyle put forward arguments for exploring the concept also for education. In 1987, Stephen Ball published the book *The Micro-Politics of the School* in which he used empirical material to push the development of the concept. In its wake the term 'micropolitics' appeared more frequently in various research papers (see e.g., Blase, 1991 for a representative collection). What are the main elements of a micropolitical view of schools:

1. The micropolitical perspective is based on a specific *view of organizations*. They are seen as characterized by diverse goals, by interaction and relationships rather than structures, by diffuse borders and unclear areas of influence rather than clear-cut conditions of superordination and delegation, by continuous, unsystematic, reactive change rather than by longer phases of stable performance and limited projects of development. The main focus of micropolitical approaches is not the organization as it aspires to be in mission statements and organizational charts, but the organization-in-action and, in particular, the "space between structures" (Hoyle, 1982: p. 88), which produces enough ambiguity to allow political activities to flourish.
2. A micropolitical approach is based on a specific *image of actors*. Members of organizations are characterized by differing interests, goals, status, power, and authority and are "pursuing their own interests, resulting in fights over real and symbolic resources" (Kelchtermans and Ballet, 2002a: p. 107). Actors may do this as individuals; they may coalesce in loosely associated interest sets, or use subdivisions of the organization (e.g., departments, professionals vs administration) as power bases. In an organization everyone influences everyone, regardless of job, title, or formal position (Brosky, 2011: p. 8). In order to protect or enhance their organizational room for maneuver, actors aim to retain or obtain control of resources, such as the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92070-1

following (see Kelchtermans and Vandenberghe, 1996: p. 7; Ball, 1987: p. 16):

a. *Material resources*: such as time, funds, teaching materials, infrastructure, and time tabling.
b. *Organizational resources:* such as procedures, roles and positions which enable and legitimate specific actions and decisions. They define 'prohibited territories' and relationships of subordination or autonomy. Thereby, they have their bearing on the actors' chances to obtain other resources (e.g., through career and participation).
c. *Normative or ideological resources*: such as values, philosophical commitments, and educational preferences. A particularly important ideological resource is the 'definition of the organization' which indicates what is legitimate and important in the organizational arena.
d. *Informational resources*: The members of an organization are interested in obtaining organizationally relevant information and to have their expertise, knowledge, and experience acknowledged.
e. *Social resources*: Affiliation to, support from, and reputation with influential groups within and outside the school are assets which may be mobilized for one's organizational standing.
f. *Personal resources:* also personal characteristics – such as being respected as a person and being grounded in an unchallenged identity as a teacher – are resources in organizational interactions.

3. A micropolitical perspective pays special attention to *interaction processes in organizations*. Schools are arenas of interaction and struggle, and it is interaction which ultimately produces their specific shape. Activities and decisions of actors in these arenas may be understood as contributing to some formal organizational goals; however, at the same time and with the same activities actors are using their influence and power, formal and informal, in order to foster their own interests and goals (see Malen and Cochran, 2008: p. 149; Blase and Blase, 2002: p. 9). Although schools are reflecting the society they are located within (see Caffyn, 2007: p. 345) and are bound to and influenced by law and guidelines, their organizational structures and processes are critically shaped by power struggles within them (see Busher, 2001: p. 80; Brosky, 2011: p. 2). Micropolitics influences all types of decision making, behavior, processes, and structures. The shape of the organization, its norms and values, and its strategies for goal attainment, are the result of ongoing micropolitical interactions within and among individuals and groups related to the organization (see Blase, 2005: p. 265). This has the following implications:
   a. Organizational interaction is seen as *conflictual and competitive*, "as a continuing struggle for control, power and influence" (Eilertsen et al., 2008: p. 300). "Confronted by competition for scarce resources and with ideologies, interests and personalities at variance, bargaining becomes crucial" (Gronn, 1986: p. 45). The relationship between control and conflict (or domination and resistance) is conceived "as the fundamental and contradictory base of organizational life" (Ball, 1994: p. 3822).
   b. Organizational interaction is *power-impregnated*. It draws on the resources mentioned above for influencing the workings of an organization. However, these resources have to be mobilized in interaction in order to become 'power.'
   c. Power *is in need of relationship*. In order to satisfy his or her interests an actor needs some interaction from other actors. Power builds on dependency which makes the relationship reciprocal, but, as a rule, asymmetric.
   d. Actors use a multitude of *strategies and tactics* to interactively mobilize their resources and to prevent others from doing so, such as setting the agenda of meetings, boycotting and escalating, avoiding visibility, confronting openly or complying, and carefully orchestrating scenes behind closed doors or in public arenas.
   e. *Micropolitical literacy* is an important competence of organizational actors. Actors who are well aware of micropolitics are in the position to cope with conflicts and profit from them (see Achinstein, 2002: p. 450). In order "to be effective in political environments, individuals need to possess *political skill*. Political skill refers to the exercise of influence through persuasion, manipulation, and negotiation" (Brosky, 2011: p. 3). Actors who are micropolitically literate are able to read the power aspect of organizational interaction and to navigate through roles, power, interests, and norms within an organization (Curry et al., 2008: p. 662).
   f. Power is employed *from both ends of the organizational hierarchy*. There are advantages of position, however, "micropolitical skills may be well distributed throughout the organization and position is not always an accurate indicator of influence" (Ball, 1994: p. 3824). For example, Brosky (2011: p. 4) identified six influence tactics used by subordinates: (1) *ingratiation*: for example, getting another actor in good mood; (2) *exchange*: using promises, rewards, benefits, and so on; (3) *rationality*: relying on logical arguments and evidence; (4) *assertiveness*: requesting in a forceful manner; (5) *upward appeal*: appealing to higher instances; and (6) *coalitions*: seeking the aid of others.
7. *Organizational change is non-teleological, pervasive, and value laden* since it advances the position of certain groups at the expense of the relative status of others (see Ball, 1987: p. 32). Schools are not organizations easily changed; a 'quick-fix orientation' in innovation will not be appropriate to their cultural complexities (see Corbett et al., 1987: p. 57).

Micropolitics certainly cultivates an interest in the conflict side of organizational interaction. Conflict offers special opportunities for both theoretical understanding and practical development of organizations. However, conflict also "generates opportunities to strengthen communities, for in the conflict lies an occasion to examine differences of beliefs, solicit alternative voices, bridge across differences to find common ground, and seek opportunities for change and growth" (Achinstein, 2002: p. 449). In this view, conflicts do not necessarily have a negative impact, they also offer an opportunity to learn and grow. Nevertheless, 'embracing conflicts' may be less appealing to those who have to live with them. Frustration about repeated conflicts and unsolved problems may result in stress, burnout and, as

a consequence, in high teacher turnovers (see Achinstein, 2002: p. 449).

Conflict is endemic to organizations, although, it need not be dominant in all instances. Political systems do have conflicting and cooperative elements. Collaboration, coalition, routine, unspoken nonaggression pacts, outspoken negotiations, political alliances, and obligations may establish comparatively smooth time spans in organizational life (see Kelchtermans and Ballet, 2002a: p. 107; Blase, 2005: p. 264). However, also these are considered to be power-impregnated, representing a more or less fragile balance of power.

Research in the Micropolitics of Education

Methods and Strategies

Micropolitical processes are a constant and normal part of organizational life and they contribute to the organizational shape and to the performance of a school. Studying micropolitics may help to understand why schools operate the way they do and to explain organizational outcomes which do not seem 'rational' at first sight. It may help identifying relevant actors who are not highlighted by the official organizational charts and ways of operating and problem solving which have not been foreseen by the respective organizational guidelines.

However, if part of the micropolitical activities which are interesting for the researcher take place in privacy, if these processes are often overlain by routine activities (see Ball, 1994: p. 3822), and if experienced actors regularly use strategies such as covering up actions and deceiving about real intentions, then researching the micropolitics of schools will have to cope with some methodological problems. To come to grips with their object of study, micropolitical studies focus on critical incidents, persons, and phases that disrupt routine (e.g., the appointment of a new principal, or the introduction of new tasks or programs), instances of structural reorganization (e.g., a school merger), and so on. Micropolitical studies pay special attention to the processes of (re-)distribution of resources, rewards, and benefits (see Ball, 1994: p. 3823) or analyze instances of organizational conflict to find out how new organizational order has been established. These strategies provide a quicker access to indicative processes, but are certainly in danger of inducing an overestimation of the proportion of conflict in organizations. Micropolitical studies try to enhance their credibility through triangulation of methods, collection of multiple perspectives, long-term engagement in the field, critical discourse in a team of researchers, and feedback from participants and external researchers. Although there have been some attempts to tackle organizational micropolitics with quantitative means (see e.g., Blickle, 1995), most studies are based on qualitative methods. There have been interview and observation studies, however, the typical strategy is a case study approach mainly based on interviews, field notes from participant or non-participant observation, transcripts of meetings, analyses of documents, and cross-case comparison of different sites.

A good example of such a case study is Sparkes' (1990) 3-year study of a newly appointed departmental head's attempts at innovation: The head seeks to orientate Physical Education teaching toward a child-centered approach emphasizing mixed-ability grouping and individual self-paced activities which runs counter to the then prevalent 'sporting ideology.' In the beginning, the new head believes in rational change and democratic participation. He tries to make other department members understand his own particular teaching approach before introducing changes to the curriculum. However, not all staff can be convinced through dialogue.

In his management position the departmental head can control the agenda of department meetings, time and procedure of arriving at decisions, and so on. He secures the agreement of crucial staff members before the meetings and networks with other influential persons outside his department. His proposals for innovation are nevertheless met with resistance by some staff. However, this resistance cannot prevent the structural changes being introduced. On the other hand, these structural changes cannot guarantee teachers' conformity to the head's educational goals within the confines of individual classrooms. Thus, he goes on to consolidate his domination of the department by careful selection of new staff and exclusion of the unwilling.

Although being successful in that his proposals for curriculum change were implemented within one year, the departmental head ended up in a situation that was characterized by conflict and uncertainty and, thus, in many respects, counteracted his own aspirations of democratic leadership. Sparkes suggests that his findings have some bearing beyond the specific case. He warns that the "ideology of school-centered innovation … fails to acknowledge the presence and importance of conflict and struggle both within, and between, departments. … the commonly held view of teacher participation in the innovative process as 'good' in itself is highly questionable. The findings presented indicate that qualitative differences exist in the forms of participation available to those involved and that, to a large extent, these are intimately linked to the differential access that individuals have to a range of power resources" (Sparkes, 1990: p. 177).

Research Topics

Potential topics are manifold. Björk and Blase (2009), for instance, analyzed how strategic alliances by district middle managers hampered efforts of decentralization, Kelchtermans and Ballet (2002a) studied beginning teachers efforts in coping with the 'real-life requirements' of their profession, and Chen (2003) investigated strategies of candidates in schoolhead election at a primary school in Taiwan. The most typical topics of micropolitical studies include:

Processes of Change

When organizational 'routine games' are disrupted, for example, through educational, curricular, or organizational innovations, through the recruitment of new principals or teachers, through the introduction of quality assurance systems, then intensified processes of micropolitical relevance are to be expected. Some studies investigate how externally mandated reform is dealt with in schools: When Ball and Bowe (1991: p. 23) studied the impact of the 1988 British Educational Reform Act, they found that "the implementation of externally initiated changes is mediated by the established culture and history of the institution, and that such changes,

their acceptance, and their implementation become sites as well as the stake of internal dispute." The reform puts some possible alternative definitions about the meaning and the workings of schools on the agenda which have to be 'processed' in the local environments through disputes over topics such as "serving a community versus marketing a product" or "priority to managerial or professional views of schooling."

Leadership

Traditionally, educational ethnography had a strong classroom orientation. Micropolitics may be understood as a move to apply ethnographic methods to study the interaction of the members of the school outside the classroom. Although parents,' pupils,' and ancillary staff's voices are included in some studies, most research focuses on the interaction of teachers. Since steering and regulation of school work are at the heart of the micropolitical approach, it is no wonder that issues of leadership have been frequently investigated (see e.g., Ball, 1987: p. 80). Focusing on leadership does not mean personalizing organizational affairs, since leadership is seen as a political process, as "one of the best policy levers to pull to achieve desirable ends" (Flessa, 2009: p. 337).

School leaders play a central role in shaping a school. They significantly affect the school culture, define basic rules, and determine topics and agendas as well as responsibilities and tasks of other members (see Ehrich and Cranston, 2004: p. 2). By having the power to reward or punish they show subordinates the potential costs of actions (see Malen and Cochran, 2008: p. 161). Evidence shows that the micropolitical orientation of teachers varies according to those of their principals. 'Open principals' provoke an open and diplomatic stance in teachers while a closed or manipulative principal encourages protective actions like avoidance and ingratiation (see Blase and Blase, 2002: p. 15).

Holders of senior positions must acquire micropolitical skills for their career advancement and for managing organizational control (see Ball, 1994: p. 3824). "(…) the skill school leaders need is not the perfect strategy to engage in the impossible struggle to eliminate micropolitics but instead the skill of inquiry to look for it in every instance, to examine who gains and who loses under important school decisions, and to develop a better understanding of the concerns that motivate people to join together in support or opposition" (Flessa, 2009: p. 346). Skilled headpersons can read conflicts as sources of information about what matters most to various school stakeholders.

Managing Relationships within the Organization and with Stakeholders

Under organizational conditions the welfare of children is bound to the well-being of the adults who work with them (Nias, 2005: p. 228). Analyzing the micropolitics of schools often means analyzing adults' behavior at the workplace. Colleagues are, not only for beginning teachers, an important source for backup, assistance, new ideas, and emotional support, but also may be potential partners for discussion, for sharing knowledge, for advice, and for consolation. Overall, they may be of great help to cope with the demands of being a teacher (see Nias, 2005: p. 204), but they also may be a source of constant pressure, distress, or uncertainty. Working conditions and relationships in and around school are on the one hand significantly influencing school culture, and on the other hand influenced by school culture itself. They have an impact on how different demands and issues are dealt with, on the overall satisfaction of staff members and other stakeholders, and on the output of the organization.

Socialization and Professional Development of Teachers

The politically impregnated organizational climate in schools is the context in which individual teachers develop their professional identity (Kelchtermans and Vandenberghe, 1996). Teachers do not only have to deal with issues at the classroom level, but are also members of an organization and confronted with 'the micropolitical reality' of the organization (see Kelchtermans and Ballet, 2002a: p. 105). Thus, micropolitical literacy would be important for beginning teachers who are to develop their professional identity in a political arena. However, training programs for teachers usually fail to provide sufficient preparation for this aspect of work life; the same holds true for programs of leadership development, although in the meantime to a lesser degree.

During the first years of their career, teachers concentrate on acquiring knowledge and competencies with respect to academic and control aspects of classroom instruction. Later on they build up a 'diplomatic political perspective,' because they feel under 'constant scrutiny' by pupils, parents, fellow teachers, and administrators (see Blase, 1991: p. 189; Kelchtermans and Vandenberghe, 1996). A pervasive feeling of 'vulnerability' induces many teachers to develop a range of protective strategies such as low-risk grading, documentation of assessment and instruction, and avoidance of risky topics and extracurricular activities (see Blase, 1991: p. 193). Thus, it is no wonder that some authors claim that the acquisition of political competence is fundamental to teachers' career satisfaction (see Ball, 1994: p. 3824). "Explicitly and implicitly, most micropolitical studies of teaching have demonstrated that teachers routinely use power strategies to influence and to protect themselves from others (particularly principals, parents, and students) and that they develop cognitive and affective responses consistent with a micropolitical perspective" (Blase and Blase, 2002: p. 14).

Understanding Schools as Organizations

Since micropolitics is an approach to organization theory, all these studies aim to contribute to our understanding of schools as organizations. Schools as organizations are placed in an environment of complex relationships where they are confronted with diverse requirements from parties within and outside its boundaries (Caffyn, 2007). Many studies prove the intensity of power-impregnated activity in schools and analyze typical causes and process forms. It has been argued that political control strategies are a characteristic part of school life "precisely because the professional norms of teacher autonomy limit the use of the effectiveness of more overt forms of control" (Ball, 1994: p. 3824).

In Iannaconne's (1991: p. 468) view, the unique contribution a micropolitical approach can offer lies in exploiting the metaphor of 'society' for schools: A typical school is organized like a 'caste society' since a majority of its

members are "subject to its laws without the right to share in making them" (Iannaconne, 1991: p. 469). From the characteristics of a 'caste society' some hypotheses about schools may be derived, for example: (1) Such societies are characterized by great tensions between the castes and a fragile balance of power. (2) They tend to conceal the internal differences of a caste from others. Consequently, critical self-examination of castes and open debate over policy issues is difficult. (3) They tend to avoid fundamental conflicts and to displace them into petty value conflicts. "In schools, as in small rural towns, the etiquette of gossip characterizes teacher talk. Faculty meetings wrangle over trivial matters, avoiding philosophic and ethical issues like the plague" (Iannaconne, 1991: p. 469).

Achinstein (2002) contrasts two schools and their differences in dealing with conflict over underscoring pupils. One school is, in accordance with Iannaconne's analysis, avoiding conflict at all cost. Head and staff employ strategies in order to avoid conflict or immediately restore consensus, for example, blaming students, avoiding criticism of teaching practices, sarcasm, and choosing the 'right staff.' In contrast, the other school – a 'conflict embracing community' – has established a culture where conflict is seen as something to learn from. Norms, values, and practices are constantly contested and double loop learning takes place.

Conclusion, Criticism, and Areas for Future Development

The first wave of research and conceptual work in micropolitics in the beginning of the 1990s has not only triggered some criticism which will be discussed in order to identify possible limitations, but also areas for future development:

1. Many authors of the micropolitical approach revel in examples of illegitimate strategies: secrecy, deception, and crime add much to the thrill of micropolitical studies. Some critics argue that talking about the pervasiveness of 'illegitimate power' *produces itself a 'micropoliticization' of practice and will undermine confidence in organizations*. This claim implies that a second meaning of 'micropolitics' is introduced: It does not (only) refer to an *approach in organization theory* (which analyses both legitimate and illegitimate interactions and their effects for the structuration of an organization), but refers to those *specific activities and attitudes* in organizations which are markedly based on power politics with illegitimate means. This conceptual move, however, produces problems: How should it be possible to draw the line between 'good' and 'bad organizational politics' since it were precisely situations characterized by diverse goals and contested expertise which made a concept like 'micropolitics' necessary? (see Kelchtermans, 2005: p. 1001)
2. Although the proponents of micropolitics authors are right to criticize a presupposition of organizational consensus, it is unsatisfactory to interpret any consensus as a form of domination (see Ball, 1987: p. 278; but also Ball, 1994: p. 3822). Many early studies did not include cooperative or consensual interactions (see Blase, 1991: p. 9). It is necessary to develop conceptual devices to account for relationships of co-agency which neither dissolve any cooperation into subtle forms of domination nor gloss over the traces of power which are at the core of collaboration and support (see Altrichter and Salzgeber, 2000: p. 104).
3. In focusing on strategic aspects of action, micropolitical approaches are again in danger of *overemphasizing the 'rationality' of organizational action*. To understand the workings of organizations it may be necessary to consider actions which are mainly influenced by emotions and affect, which have not been carefully planned according to some interests, as well as unintended consequences of otherwise motivated actions. Kelchtermans and Ballet (2002b: p. 765) distinguish three aspects of micropolitical literacy: (1) knowledge – seeing, understanding, and interpreting micropolitics in a particular situation; (2) instrumental aspect –the 'repertoire' of micropolitical strategies; and (3) experiential aspect – how confrontation with micropolitical reality is experienced. In particular, when highly valued issues – such as the self-understanding as professionals and working conditions – are at stake, intense emotions are triggered (see Kelchtermans, 2005: p. 1004).
4. Although being right to put the spotlight on the potential instability of organizational processes, early micropolitical studies had some difficulty in conceptualizing *the relative stability and durability of organizations, which we can observe day in, day out in many organizations*. For this purpose, it will be necessary to bring agency and structure in a balanced conceptual relationship: micropolitics will not deal with "relationships rather than structures" (Ball, 1994: p. 3822), but put precisely with the interplay between interaction and structure.
5. The *understanding of politics displayed in many early micropolitical studies was quite unsatisfactory, not to say unpolitical*. Politics was often seen as the opposite of truth and reason; it was associated with treason, conspiracy, and accumulation of influence as if fair and democratic ways of negotiation in organizations were inconceivable.

   Many of these problems may be overcome by reference to Giddens' (1992) theory of structuration, Crozier and Friedberg's (1993) concept of organizational games, and Ortmann et al. (1990) studies of power in organizational innovation. On this conceptual basis, organizational structure may be seen as an outcome of and resource for social actions which enables and restricts actions, but never fully determines them. Organizations are webs woven from concrete interactions of (self-)interested actors. They acquire some spatial–temporal extension through the actors' use and reproduction of 'games.' The 'game' is the point of conceptual intersection in which the analytically divided elements of agency and structure, of individual and organization meet. The game works as an "indirect integration mechanism which relates the diverging and/or contradictory action of relatively autonomous actors" (Crozier and Friedberg, 1993: p. 4). In order to take action and to pursue their own interests members of organizations must at least partly use structural resources of the organization, that is, they are forced to play within the organizational rules, and thereby, they reproduce these rules. The functioning of an organization may be explained as "result of a series of games articulated

among themselves, whose formal and informal rules indirectly integrate the organization members' contradictory power strategies" (Ortmann et al., 1990: p. 56). The interlocked character of the games and the high amount of routine elements in social practices are major reasons for the relative stability of organizations (see Altrichter and Salzgeber, 2000: p. 106).

6. It was also criticized that in some micropolitical studies *the external relationships of the organization were dissolved into the interactions between individual actors*. On the other hand it must be said that many studies took care to account for the societal embedding of organizational interactions (see e.g., Ball and Bowe, 1991). In the meantime there is general agreement that this must be a feature of sound micropolitical studies (see Blase, 1991: p. 237).

This is even more important since many national schools' systems have recently faced and are still facing massive changes in their governance structure. New policies, such as standards, accountability, autonomy, and access policies, have to be implemented at district and school levels, necessarily involving micropolitical struggles over the new distribution of rights and resources (Altrichter et al., 2014) and producing ambiguous situations: Schools are required to act in the interest of its students and according to the professional norms of its staff while at the same time to "pursue broader social and economic imperatives" (Johnson, 2004: p. 268). New policies are also providing new levers for external influence on schools, for example, accountability policies contain the "potential to produce information that attentive publics … might use to press for school reforms that they view as key" (Malen and Cochran, 2008: p. 152).

Theoretically, these developments in the field call for linking analyses on macro- and micro-political level. The 'New Model of Micropolitics' by Smeed et al. (2009) begins with recognizing the impact of the macropolitical context – for example, heightened accountability requirements, testing, and pressure from the wider community – on the micropolitics at school level. Some researchers who have contributed to the literature on micropolitics have recently turned to study the *governance of school systems* (e.g., Ball, 2007). Governance studies aim to analyze how systemic coordination is achieved by the transactions of various social actors in a multilevel system (de Boer et al., 2007; Altrichter, 2010). It is particularly important to understand how new policies which have been developed on a particular system level are perceived on another level (e.g., on school level) and translated into agency and structures adequate to the norms and work traditions prevalent on this level. Fend (2006: p. 69) has captured these processes of translation and transformation by the concept of 'recontextualization' and we will not go wrong by assuming that these processes are 'micropolitical' in that they may be conceived as negotiation of interested actors in a power-impregnated field.

See also: Conflict: Organizational; Educational Institutions, History of; Educational Policy, Mechanisms for the Development of; Governance; Leadership Education: Theory and Practice; Leadership in Organizations, Sociology of; Learning: Organizational; Negotiation and Bargaining: Organizational Aspects; Organizational Decision Making; Organizations, Sociology of; Organizations: Negotiated Orders; Organizations: Unintended Consequences; People in Organizations; Professions in Organizations; School Improvement; School Management; School as a Social System; Teacher Education; Teacher Expertise.

Bibliography

Achinstein, B., 2002. Conflict amid community: the micropolitics of teacher collaboration. Teachers College Record 104, 421–455.

Altrichter, H., 2010. Theory and evidence on governance: conceptual and empirical strategies of research on governance in education. European Educational Research Journal 9 (2), 147–158.

Altrichter, H., Heinrich, M., Soukup-Altrichter, K., 2014. School decentralisation as a process of differentiation, hierarchization and selection. Journal of Education Policy 29 (5), 675–699.

Altrichter, H., Heinrich, M., Soukup-Altrichter, K., 2011. Schulentwicklung durch Schulprofilierung? VS, Wiesbaden.

Altrichter, H., Salzgeber, S., 2000. Some elements of a micro-political theory of school development. In: Altrichter, H., Elliott, J. (Eds.), Images of Educational Change. Open University Press, Buckingham, pp. 99–110.

Bacharach, S., Lawler, E., 1980. Power and Politics in Organizations. Jossey Bass, San Francisco.

Ball, S.J., 1987. The Micro-Politics of the School. Routledge, London.

Ball, S.J., 1994. Micropolitics of schools. In: Husen, T., Postlethwaite, T.N. (Eds.), The International Encyclopedia of Education, second ed. Pergamon, Oxford, pp. 3821–3826.

Ball, S.J., 2007. Education Plc. Understanding Private Sector Participation in Public Sector Education. Routledge, London.

Ball, S.J., Bowe, R., 1991. Micropolitics of radical change. Budgets, management, and control in British schools. In: Blase, J. (Ed.), The Politics of Life in Schools. Sage, Newbury Park, pp. 19–45.

Björk, L., Blase, J., 2009. The micropolitics of educational change and reform: cracking open the black box. In: Hargreaves, A., Lieberman, A., Fullan, M., Hopkins, D. (Eds.), Second International Handbook of Educational Change. Part 1, vol. 23. Springer, Dordrecht, pp. 237–258.

Blase, J., 1991. The Politics of Life in Schools. Sage, Newbury Park.

Blase, J., 2005. The micropolitics of educational change. In: Hargreaves, A. (Ed.), Extending Educational Change. International Handbook of Educational Change. Springer, Dordrecht, pp. 264–277.

Blase, J., Anderson, G.L., 1995. The Micropolitics of Educational Leadership. Cassell, London.

Blase, J., Blase, J., 2002. The micropolitics of instructional supervision: a call for research. Educational Administration Quarterly 38, 6–44.

Blickle, G., 1995. Wie beeinflussen Personen erfolgreich Vorgesetzte, KollegInnen und Untergebene? (How do persons successfully influence superiors, colleagues, and subordinates?). Diagnostica 41, 245–260.

Brosky, D., 2011. Micropolitics in the schools: teacher leaders' use of political skill and influence tactics. International Journal of Educational Leadership Preperation 6 (1), 331–340.

Busher, H., 2001. The micro-politics of change, improvement and effectiveness in schools. In: Harris, A., Bennett, N. (Eds.), School Effectiveness and School Improvement. Alternative Perspectives. Continuum, London, pp. 75–97.

de Boer, H., Enders, J., Schimank, U., 2007. On the way towards new public management? The governance of university systems in England, the Netherlands, Austria, and Germany. In: Jansen, D. (Ed.), New Forms of Governance in Research Organisations. Springer, Dordrecht, pp. 137–152.

Caffyn, R., 2007. Fragmentation in international schools: a micropolitical discourse of management, culture and postmodern society. In: Hayden, M., Levy, J., Thompson, J. (Eds.), The SAGE Handbook of Research in International Education. Sage, London, pp. 339–350.

Chen, H., September 2003. School-Based Management: The Micro-Political Interactions among the Headteacher, Directors, Teachers and Parents. Paper presented at the Britisch Educational Research Association Annual Conference, Heriot-Watt University, Edinburgh. British Education Online: www.leeds.ac.uk/educol/documents/00003234.htm, 11–13.

Corbett, H.D., Firestone, W.A., Rossman, G.B., 1987. Resistance to planned change and the sacred in school cultures. Education Administration Quarterly 23, 36–59.

Crozier, M., Friedberg, E., 1993/1980. Die Zwänge kollektiven Handelns. Hain, Frankfurt/M. Actors and Systems, University of Chicago Press, Chicago.

Curry, M., Jaxon, K., Russell, J.L., Callahan, M.A., Bicais, J., 2008. Examining the practice of beginning teachers' micropolitical literacy within professional inquiry communities. Teaching and Teacher Education 24, 660–673.

Ehrich, L.C., Cranston, N., 2004. Developing Senior Management Teams in Schools: Can Micropolitics Help? School of Learning & Professional Studies, Faculty of Education, Queensland University of Technology, Australia. eprints.qut.edu.au/1383/1/Dev_Snr_Mngmt.pdf.

Eilertsen, T.-V., Gustafson, N., Salo, P., 2008. Action research and the micropolitics in schools. Educational Action Research 16 (3), 295–308.

Fend, H., 2006. Neue Theorie der Schule. VS, Wiesbaden.

Flessa, J., 2009. Educational micropolitics and distributed leadership. Peabody Journal of Education 84, 331–349.

Giddens, A., 1992/1984. Die Konstitution der Gesellschaft. M.: Campus, Frankfurt. The Constitution of Society, Polity Press, Cambridge.

Gronn, P., 1986. Politics, power and the management of schools. In: Hoyle, E., McMahon, A. (Eds.), World Yearbook of Education. Kogan Page, London, pp. 45–54.

Hoyle, E., 1982. Micropolitics of educational organizations. Educational Management and Administration 10, 87–98.

Iannaconne, L., 1991. Micropolitics of education – what and why. Education and Urban Society 23, 465–471.

Johnson, B., 2004. Local school micropolitical agency: an antidote to new managerialism. School Leadership & Management 24, 267–286.

Kelchtermans, G., 2005. Teachers' emotions in educational reforms: self-understanding, vulnerable commitment and micropolitical literacy. Teaching and Teacher Education 21, 995–1006.

Kelchtermans, G., Ballet, K., 2002a. The micropolitics of teacher induction. A narrative-biographical study on teacher socialisation. Teaching and Teacher Education 18, 105–120.

Kelchtermans, G., Ballet, K., 2002b. Micropolitical literacy: reconstructing a neglected dimension in teacher development. International Journal of Educational Research 37, 755–767.

Kelchtermans, G., Vandenberghe, R., 1996. Becoming Political: A Dimension in Teachers' Professional Development. Paper presented at the AREA-Conference, New York (ERIC-document: ED 395–921).

Malen, B., Cochran, M.V., 2008. Beyond pluralistic patterns of power: research on the micropolitics of schools. In: Coopers, B.S., Cibulka, J.G., Fusarelli, L.D. (Eds.), Handbook of Education Politics and Policy. Routledge, New York, pp. 148–178.

Nias, J., 2005. Why teachers need their colleagues: a developmental perspective. In: Hopkins, D. (Ed.), The Practice and Theory of School Improvement. Springer, Dordrecht, pp. 223–237.

Ortmann, G., Windeler, A., Becker, A., Schulz, H.-J., 1990. Computer und Macht in Organisationen (Computers and Power in Organizations). Westdeutscher Verlag, Opladen.

Pfeffer, J., 1981. Power in Organizations. Pitman, Boston.

Sparkes, A.C., 1990. Power, domination and resistance in the process of teacher-initiated innovation. Research Papers in Education 5, 153–178.

Smeed, J., Kimber, M., Millwater, J., Ehrich, L., 2009. Power over, with and through: another look at micropolitics. Leading & Managing 15, 26–41.

Schumpeter, Joseph A. (1883–1950)

Richard Swedberg, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Joseph A. Schumpeter (1883–1950) is generally considered one of the major economists of the twentieth century. He was born in the Austro-Hungarian Empire but emigrated to the United States in the 1930s where he worked as a professor of Economics at Harvard University. Schumpeter's most important contribution is to the theory of entrepreneurship, especially as developed in *The Theory of Economic Development* (1912; trans. 1934). According to Schumpeter's well-known definition, entrepreneurship means "the carrying out of new combinations." His work *Capitalism, Socialism and Democracy* (1942) is probably his best-known work and is regarded as a classic not only in economics but also in political science. Of special interest in this work is Schumpeter's definition of democracy as the election by the people of a new elite.

Joseph A. Schumpeter (1883–1950) is generally considered one of the major economists of the twentieth century. His main effort was to develop a theory of economic change that could complement Walras' theory of general equilibrium. Schumpeter centered his theory of economic change on the entrepreneur, and in *Theorie der wirtschaftlichen Entwicklung* (1912; tr. 1934) he defined entrepreneurship as "the carrying out of new combinations." The initial innovation, according to Schumpeter, attracted other entrepreneurs and eventually set off a business cycle. Schumpeter's best known work is *Capitalism, Socialism and Democracy* (1942), which is regarded as a classic not only in economics but also in political science. Several of Schumpeter's analyses in this work are still controversial in nature, such as his suggestion that monopolies can be healthy for the economy and that modern capitalism is digging its own grave and preparing the road for socialism. The most celebrated contribution of this work, however, is to be found in Schumpeter's discussion of democracy, more precisely in his critique of the traditional way of defining democracy as an attempt to realize the will of the people and in his own preference for a definition of democracy as competition among political leaders for the votes of the citizens.

Joseph A. Schumpeter is one of the best known and most flamboyant economists of the twentieth century. His main accomplishments are to have introduced the entrepreneur into economic theory through *Theorie der wirtschaftlichen Entwicklung* (1912) and to have produced a magnificent history of economic thought, *History of Economic Analysis* (posthumously published in 1954a). Schumpeter's interests were wide ranging, and his writings also include contributions to sociology and political science. His analysis of democracy, which can be found in his most popular work, *Capitalism, Socialism and Democracy* (1942), is generally regarded as an important addition to the theory of democracy.

Early Life and Career

Joseph Alois Schumpeter was born on 8 February 1883, in the small town of Triesch in the Austro-Hungarian Empire (today Trest in Slovakia). His father was a cloth manufacturer and belonged to a family that had been prominent in the town for many generations, and his mother, who was the daughter of a medical doctor, came from a nearby village. Schumpeter's father died when he was 4 years old, an event that was to have great consequences for the future course of his life. Sometime later, his mother moved to Graz and then to Vienna, where she got married to a retired officer from a well-known, aristocratic family, Sigismund von Kéler. Schumpeter was consequently born into a small-town bourgeois family but raised in an aristocratic one, and his personality and his work were to contain a curious mixture of values from both of these worlds.

After having graduated in 1901 from the preparatory school for the children of the elite in the Empire, Theresianum, Schumpeter enrolled at the University of Vienna. His goal from the very start was to become an economist. Around the turn of the century, the University of Vienna had one of the best educations in economics in the world, thanks to Carl Menger and his two disciples Eugen von Böhm-Bawerk and Friedrich von Wieser. Schumpeter had the latter two as his teachers, and he also attended lectures in mathematics on his own. When Schumpeter received his doctorate in 1906, he was 23 years old and the youngest person to have earned this degree in the Empire.

After having traveled around for a few years and added to his life experience as well as to his knowledge of economics, Schumpeter returned in 1908 to Vienna to present his *Habilitationsschrift*. Its title was *Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie* (1908; *The Nature and Essence of Theoretical Economics*), and it can be characterized as a work in economic methodology from an analytical perspective. In 1909, Schumpeter was appointed assistant professor at the University of Czernowitz and thereby became the youngest professor in German-speaking academia. It was also at Czernowitz that Schumpeter produced his second work, the famous *Theorie der wirtschaftlichen Entwicklung* (1912; second ed. from 1926 trans. as *The Theory of Economic Development*). At the University of Graz, to which he moved as a full professor in 1911, Schumpeter produced a third major study, *Epochen der Dogmen- und Methodengeschichte* (1914; trans. in 1954 as *Economic Doctrine and Method*). Over a period of less than 10 years Schumpeter had produced three important works, and one understands why he referred to the third decade in a scholar's life as "the decade of sacred fertility."

The Works from the Early Years

Das Wesen is today best known for being the place where the term 'methodological individualism' was coined. In its time,

however, this work played an important role in introducing German students to analytical economics, which Gustav von Schmoller and other members of the Historical School had tried to ban from the universities in Germany. *Das Wesen* is well written and well argued, but Schumpeter would later regard it as a youthful error and did not allow a second edition to appear. The reason for Schumpeter's verdict is not known, but it is probably connected to his vigorous argument in the book that economics must sever all links with the other social sciences.

While the emphasis in *Das Wesen* had been exclusively on statics, in his second book, *Theorie der wirtschaflichen Entwicklung*, it was on dynamics. Schumpeter very much admired Walras' theory of general equilibrium, which he had tried to explicate in *Das Wesen*, but he was also disturbed by Walras' failure to handle economic change. What was needed in economics, Schumpeter argued, was a theory according to which economic change grew out of the economic system itself and not, as in Walras, a theoretical scheme in which economic change was simply explained as a reaction to a disturbance from outside the economic system.

Schumpeter starts the argument in *Theorie* by presenting an economy where nothing new ever happens and that therefore can be analyzed with the help of static theory, that is, all goods are promptly sold, no new goods are produced or wanted, and profits are zero. Schumpeter then contrasts this situation of a 'circular flow' with a situation in which the activities of the entrepreneur sets the whole economic system in motion, and which can only be explained with the help of a dynamic theory. The model that Schumpeter now presented in *Theorie* would later be repeated and embellished upon in many of his writings: the entrepreneur makes an innovation, which earns him a huge profit, this profit attracts a swarm of other, less innovative entrepreneurs and as a result of all these activities, a wave of economic change begins to work its way through the economic system, a business cycle, in brief, has been set in motion. The theory of business cycles that one can find in *Theorie* is fairly rudimentary, and Schumpeter would later spend much time in improving upon it.

Much of Schumpeter's analysis in *Theorie* is today forgotten and few economists are today interested in the way that Schumpeter explains interest, capital, and profit by relating these to the activities of the entrepreneur. Some parts of *Theorie*, however, are still very much alive and often referred to in studies of entrepreneurship. In one of these, Schumpeter discusses the nature of an innovation, and explains how it differs from an invention; while an invention is something radically new, an innovation consists of an attempt to turn something – for example, an invention – into a money-making enterprise. In another famous passage, Schumpeter discusses the five main forms of innovations:

> (1) The introduction of a new good … or of a new quality of a good. (2) The introduction of a new method of production … (3) The opening of a new market … (4) The conquest of a new source of supply of raw materials or half-manufactured goods … (5) The carrying out of the new organization of any industry, like the creation of a monopoly position … or the breaking up of a monopoly position.

Schumpeter's description of what drives an entrepreneur is also often referred to, as is his general definition of entrepreneurship: "the carrying out of new combinations."

Schumpeter's third work from these early years is a brief history of economics, commissioned by Max Weber (1914–20) for a handbook in political economy (*Grundriss der Sozialökonomik*). Its main thesis is that economics came into being when analytical precision, of the type that one can find in philosophy, came into contact with an interest in practical affairs, of the type that is common among businessmen; and to Schumpeter this meeting took place in the works of the Physiocrats. *Epochen der Dogmen- und Methodengeschichte* was later overshadowed by the much longer *History of Economic Analysis*, but while the latter work tends to be used exclusively as a reference work, Schumpeter's study from 1914 can be read straight through and is also easier to appreciate.

Difficult Mid Years

By the mid-1910s, Schumpeter had established himself as one of the most brilliant economists in his generation, and like other well-known economists in the Austro-Hungarian Empire, he hoped for a prominent political position, such as finance minister or economic adviser to the Emperor. During World War I, Schumpeter wrote a series of memoranda, which were circulated among the elite of the Empire, and through which he hoped to establish himself as a candidate for political office. The various measures that Schumpeter suggested in these memoranda are of little interest today, for example, that the Empire must not enter into a customs union with its powerful neighbor, Germany. What is of considerably more interest, however, is that they provide an insight into Schumpeter's political views when he was in his early 30s. Schumpeter's political profile at this time can be described in the following way: he was a royalist, deeply conservative, and an admirer of tory democracy of the British kind.

Schumpeter's attempt during World War I to gain a political position failed, but his fortune changed after the Empire had fallen apart and the state of Austria came into being. In early 1919, Schumpeter was appointed finance minister in a joint Social Democratic and Catholic Conservative government in Austria led by Karl Renner. Already by the fall of 1919, however, Schumpeter was forced to resign, mainly because the Social Democrats felt that he had betrayed them and could not be trusted. The Social Democrats in particular thought that Schumpeter had gone behind their backs and stopped their attempts to nationalize parts of Austria's industry. Schumpeter always denied this, and in retrospect it is difficult to establish the truth. What is clear, however, is that the Austrian Social Democrats advocated a democratic and peaceful form of socialism, while Schumpeter detested all forms of socialism.

Loath to go back to his position at the University of Graz, Schumpeter spent the early part of the 1920s working for a small bank in Vienna, the Biedermann Bank. Schumpeter did not take part in the everyday activities of the bank, but mainly worked as an independent investor. These activities ended badly, however, and by the mid-1920s, Schumpeter was forced to leave the Biedermann Bank and had by this time accumulated a considerable personal debt. For many years to come, Schumpeter would give speeches and write articles in order to

pay back his debts. Adding to these economic misfortunes was the death in 1926 of his beloved wife Anna ('Annie') Reisinger, a 20 years younger working class girl to whom Schumpeter had only been married for a year. From this time on, Schumpeter's friends would later say, there was a streak of pessimism and resignation in his personality.

In 1925, Schumpeter was appointed professor of public finance at the University of Bonn, and it is clear that he was both pleased and relieved to be back in academia. While his ambition, no doubt, extended to politics and business, he always felt an academic at heart. The great creativity that Schumpeter had shown in the 1910s was, however, not to return in the 1920s, and Annie's death was no doubt an important reason for this. Schumpeter's major project during the 1920s was a book in the theory of money, but this work was never completed to his satisfaction (Schumpeter, 1970). Schumpeter did, however, write several interesting articles during these years, some of which are still worth reading. One of these is *The Instability of Capitalism* (1989), in which Schumpeter suggests that capitalism is undermining itself and will eventually turn into socialism. Another worthwhile article is *Gustav vs. Schmoller und die Probleme von heute* (1926), in which Schumpeter argues that economics has to be a broad science, encompassing not only economic theory but also economic history, economic sociology, and statistics. Schumpeter's term for this type of general economics is *Sozialökonomik*.

To illustrate what economists can accomplish with the help of sociology, Schumpeter wrote an article on social classes (1991c), which is still often referred to. Schumpeter starts out by drawing a sharp line between the way that the concept of class is used in economics and in sociology. In economics, Schumpeter says, class is basically used as a category (e.g., wage earner/non-wage earner), while in sociology class is seen as a piece of living reality. An attempt was also made by Schumpeter to link his theory of the entrepreneur to the idea of class.

In mentioning Schumpeter's articles on social classes, something must also be said about two of his other famous studies in sociology: *The Fiscal Crisis of the Tax State* (1918) and *The Sociology of Imperialisms* (1991b). The former study can be described as a pioneer effort in the sociology of taxation, which not only focuses on the situation in Austria just after World War I but also attempts to formulate some general propositions about the relationship between taxation and society. Schumpeter, for example, suggests that if the citizens demand more and more subventions from the state, but are unwilling to pay for these through ever higher taxes, the state will collapse. In the article on imperialism, Schumpeter argues that imperialism grows out of a social situation that is not to be found in capitalism, and whatever imperialist forms that still exist represent an atavism and a leftover from feudal society. Schumpeter's famous definition of imperialism is as follows: "imperialism is the objectless disposition on the part of the state to unlimited forcible expansion."

The Years in the United States

By 1930, Schumpeter was still depressed about his life and quite unhappy about his career; he had found no one to replace his wife on an emotional level and he felt that he deserved a better professional position than Bonn. When Werner Sombart's chair at the University of Berlin became vacant, Schumpeter had high hopes that he would get it. When this did not happen, however, he instead accepted an offer from Harvard University, and in 1932, he moved to Cambridge, Massachusetts, for good. Schumpeter was slowly to regain his emotional composure, and also his capacity to work, in the United States. In 1937, he got married to economist Elizabeth Boody, who regarded it as her duty in life to take care of Schumpeter and manage his worldly affairs. With the support of his wife, Schumpeter produced three major works: *Business Cycles* (1939), *Capitalism, Socialism and Democracy* (1942), and *History of Economic Analysis* (published posthumously in 1954).

The Works from the American Period

Of these three works, the one for which Schumpeter had the highest hopes was definitely *Business Cycles* (1939), which appeared exactly 25 years after his last book in Europe. This work is more than 1100 pages long and it traces business cycles in Great Britain, the United States, and Germany during the years 1787–1938. Schumpeter had hoped that *Business Cycles* would get the kind of reception that Keynes' *General Theory* got, and he was deeply disappointed by the lack of interest that his colleagues and other economists showed. The impression one gets from histories of economic thought, however, is that *Business Cycles* is rarely read today, mainly because few people are persuaded by Schumpeter's theory that business cycles are set off by innovations and that their motion can be best described as a giant Kondratieff cycle, modified by Juglar and Kitchin cycles; Schumpeter's handling of statistics is also considered poor. To a large extent this critique is true, although it should also be noted that a number of themes and issues in this and other of Schumpeter's works are increasingly being referred to in evolutionary economics.

After many years of excruciatingly hard work on *Business Cycles*, Schumpeter decided that he needed an easier task to work on than the recasting of economic theory that was on his agenda since the time in Bonn. In the late 1930s, Schumpeter therefore began working on what he saw as a minor study of socialism. This minor study, however, quickly grew in size, and when it finally appeared in 1942 – under the title *Capitalism, Socialism and Democracy* – it was nearly 400 pages long. Of all of Schumpeter's works (1942), this book was to become the most popular one, especially in the United States where it is still considered a standard work in political science. *Capitalism, Socialism and Democracy* has also been translated into more languages than any other work by Schumpeter, and it is still in print.

Capitalism, Socialism and Democracy contains an interesting history of socialist parties (part V), a fine discussion of the nature of socialism (part III), and an excellent discussion of the work of Karl Marx (part I). What has made this work into a social science classic, however, are its two remaining parts with their discussion of capitalism (part II) and the nature of democracy (part IV). Capitalism, Schumpeter says, can be characterized as a process of continuous economic change, through which new enterprises and industries are continuously being created and old ones continuously being destroyed.

Schumpeter's famous term for this process is 'creative destruction.' Perfect competition of the type that economists talk about, he also says, is a myth, and monopolies, contrary to common belief and economists' dogma, are often good for the economy. Without monopolies, for example, very expensive forms of research and investments would seldom be undertaken.

Schumpeter's theory of monopolies has led to a large number of studies of innovation. By 'the Schumpeterian hypothesis' two propositions are meant: that large firms are more innovative than small firms, and that innovation tends to be more frequent in monopolistic industries than in competitive ones. Each of these hypotheses can in their turn be broken down into further hypotheses. Many decades later, the results of all of these studies are still inconclusive, and there is no consensus if monopolistic forms of enterprise further or obstruct innovation (for an overview of research on the Schumpeterian hypothesis, see Kamien and Schwartz, 1982).

Schumpeter's overall theory in *Capitalism, Socialism and Democracy* – the idea that capitalism is on its way to disappear and be replaced by socialism – has, on the other hand, led to little concrete research. One reason for this is probably that many readers of Schumpeter's book have felt that his argument is very general in nature and at times also quite idiosyncratic. Schumpeter argues, for example, that capitalism tends to breed intellectuals, and that these are hostile to capitalism; that the bourgeois class is becoming unable to appreciate property; and that the general will of the bourgeoisie is rapidly eroding. "Can capitalism survive?" Schumpeter asks rhetorically, and then gives the following answer, "No. I do not think it can." Schumpeter's conclusion about the inevitable fall of capitalism has often been sharply criticized (e.g., Heertje, 1981). His more general attempt to tackle the question of how social and economic institutions are needed for an economic system to work properly has, however, remained unexplored. There also exists a certain tendency to cite catchy phrases from Schumpeter's work without paying attention to their organic place in his arguments.

What has always been much appreciated in *Capitalism, Socialism and Democracy* is Schumpeter's theory of democracy. Schumpeter starts out by criticizing what he calls "the classical doctrine of democracy," which he characterizes in the following way: "the democratic method is that institutional arrangement for arriving at political decisions which realizes the common good by making the people itself decide issues through the election of individuals who are to assemble in order to carry out its will." What is wrong with this theory, he argues, is that there is no way of establishing what the common good is. There is also the additional fact that politicians do not have interests of their own, according to this theory; their function is exclusively to translate the desires of the people into reality.

To the classical type of democracy, Schumpeter counterposes his own theory of democracy, which he defines in the following manner: "the democratic method is that institutional arrangement for arriving at political decisions in which individuals acquire the power to decide by means of a competitive struggle for the people's vote." Some critics have argued that Schumpeter's theory is elitist in nature, since democracy is reduced to voting for a political leader at regular intervals. This may well be true, and Schumpeter's theory of democracy no doubt coexists in his work with contempt for the masses. Regardless of this critique, however, Schumpeter's theory is generally regarded as being considerably more realistic than the classical doctrine of democracy.

World War II was a very trying time for Schumpeter since he felt emotionally tied to German speaking Europe, even though he by now was a US citizen. Schumpeter did not support the Nazis but like many conservatives felt that Communism constituted a more dangerous enemy than Hitler. Schumpeter's intense hatred of Roosevelt, whom he suspected of being a socialist, also put him on a collision course with the academic community in Cambridge during World War II. In his diary, Schumpeter expressed his ambivalence toward Hitler, Jews, Roosevelt, and many other things, and in his everyday life he more and more withdrew into solitude and scholarship. The War, in brief, was a very difficult and demoralizing time for Schumpeter.

From the early 1940s, Schumpeter started to work feverishly on a giant history of economic thought that was to occupy him till his death some 10 years later. He did not succeed in completing *History of Economic Analysis*, which was instead put together by Elizabeth Boody Schumpeter after several years of work. *History of Economic Analysis* is many times longer than Schumpeter's history of economics from 1914, and it is also considerably more sophisticated in its approach. Schumpeter's ambition with this work, it should first of all be noted, was not only to write the history of economic theory, but also to say something about economic sociology, economic history, and economic statistics, the four 'fundamental fields' of economics or *Sozialökonomik*. It should also be mentioned that *History of Economic Analysis* begins with a very interesting methodological discussion of the nature of economics. This is then followed by a brilliant, though occasionally willful, discussion of economics from classical Greece to the years after World War I. The section on twentieth century economics was unfortunately never completed, but it is clear that Schumpeter was very discontent with the failure of his contemporaries to create a truly dynamic economic theory.

Schumpeter died in his sleep on 8 January 1950, and the cause was put down as cerebral hemorrhage. Some of his friends thought that the real reason was overwork, while his diary shows that he had been tired of life for a long time. Schumpeter's private notes also show that behind the image of brilliant lecturer and iconoclastic author, which Schumpeter liked to project, there was a person who felt deeply unhappy with the way that his life had evolved and with his failure to produce a theory that would revolutionize economic thought. "With a fraction of my ideas," he wrote in dismay a few years before his death, "a new economics could have been founded."

See also: Business History; Capitalism; Democracy, History of; Democratic Theory; Determinism: Social and Economic; Economic History; Economic Sociology; Economic Transformation: From Central Planning to Market Economy; Economics, History of; Entrepreneurship and Organizations; Innovation and Technological Change, Economics of; Innovation, Theory of; Marx, Karl (1818–83); Political Economy, History of; Socialism.

Bibliography

Allen, R.L., 1991. Opening Doors: The Life and Work of Joseph Schumpeter. Transaction Publishers, New Brunswick, NJ.

Augello, M., 1990. Joseph Alois Schumpeter: A Reference Guide. Springer-Verlag, New York.

Harries, S.E. (Ed.), 1951. Schumpeter: Social Scientist. Books for Libraries Press, Freeport, New York.

Heertje, A. (Ed.), 1981. Schumpeter's Vision: Capitalism, Socialism and Democracy after 40 Years. Praeger, New York.

Kamien, M.I., Schwartz, N.L., 1982. Market Structure and Innovation. Cambridge University Press, Cambridge, UK.

McCraw, T.K., 2007. Prophet of Innovation: Joseph Schumpeter and Creative Destruction. The Belknap Press, Cambridge.

Schumpeter, J.A. Werke, Briefe, Bibliografien. Ein Schumpeter-Archiv. www.schumpeter.info/.

Schumpeter, J.A., 1908. Das Wesen und der Hauptinhalt der theoretischen Nationalökonomie. Duncker and Humblot, Leipzig.

Schumpeter, J.A., 1912. Theorie der wirtschaftlichen Entwicklung. Duncker & Humblot, Leipzig (actually appeared in 1911, second ed. 1926) [trans. into English in 1934 as The Theory of Economic Development. Harvard University Press, Cambridge, MA].

Schumpeter, J.A., 1914. Epochen der Dogmen- und Metho-dengeschichte. In: Bcher, J.A., et al. (Eds.), Grundriss der Sozialökonomik. I. Abteilung, Wirtschaft und Wirtschaftswissenschaft. J.C.B. Mohr, Tbingen, Germany, pp. 19–124.

Schumpeter, 1926. Gustav vs. Schmoller und die probleme von heute. Schmollers Jahrbuch 50, 337–388.

Schumpeter, J.A., 1939. Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process. McGraw-Hill, New York.

Schumpeter, J.A., 1942. Capitalism, Socialism and Democracy. Harper and Brothers, New York.

Schumpeter, J.A., 1954a. History of Economic Analysis. Oxford University Press, New York.

Schumpeter, J.A., 1954b. Economic Doctrine and Method. Oxford University Press, New York.

Schumpeter, J.A., 1970. Das Wesen des Geldes. Vandenhoeck and Ruprecht, Göttingen, Germany.

Schumpeter, J.A., 1989. The instability of capitalism. In: Schumpeter, J.A. (Ed.), Essays. Transaction Publishers, New Brunswick, NJ.

Schumpeter, J.A., 1991a. The crisis of the tax state. In: Schumpeter, J.A. (Ed.), The Economics and Sociology of Capitalism. Princeton University Press, Princeton, NJ.

Schumpeter, J.A., 1991b. The sociology of imperialisms. In: Schumpeter, J.A. (Ed.), The Economics and Sociology of Capitalism. Princeton University Press, Princeton, NJ.

Schumpeter, J.A., 1991c. Social classes in an ethically homogenous milieu. In: Schumpeter, J.A. (Ed.), The Economics and Sociology of Capitalism. Princeton University Press, Princeton, NJ.

Schumpeter, J.A., 1991d. The Economics and Sociology of Capitalism. Princeton University Press, Princeton, NJ.

Schumpter, J.A., 2000. Briefe/Letters. J.C.B. Mohr, Tbingen, Germany.

Schumpeter, J.A., 2012. The Entrepreneur: Classic Texts. Stanford University Press, Stanford.

Stolper, W.F., 1994. Joseph Alois Schumpeter: The Private Life of a Public Man. Princeton University Press, Princeton, NJ.

Swedberg, R., 1991. Schumpeter: A Biography. Princeton University Press, Princeton, NJ.

Weber, M. (Ed.), 1914–1920. Grundriss der Sozialökonomik. J.C.B. Mohr, Tbingen, Germany.

Schütz, Alfred (1899–1959)

Ilja Srubar, Institute of Sociology, University of Erlangen, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alfred Schütz, born in Vienna, emigrated in 1938 via Paris to New York. He was founder of the phenomenological approach in sociology, which is one of the main paradigms in the *interpretative social sciences*. He formulated a theory of the life world and its structures showing how actors produce and understand social reality in everyday interaction and communication. His concepts initiated inquiry into the everyday life of societies (sociology of everyday life) and entered into the mainstream of sociological theory. His methodological suggestions sparked innovations in the field of qualitative methods of social research. Beyond sociology, his ideas were conceived in philosophy and in many fields of cultural sciences.

Vitae

Alfred Schütz was born in Vienna on 13 April 1899. Here, he studied law and social sciences from 1918 to 1921. After receiving his Doctorate of Law, he continued his studies in the social sciences until 1923. His teachers included Hans Kelsen, an eminent lawyer, Ludwig von Mises, and Friedrich von Wieser, prominent representatives of the Austrian school of economics. Equally important as his university studies was his participation in the informal academic life of Vienna, e.g., his participation in the private seminars of von Mises, where Schütz also cultivated his philosophical interests. In addition to his scholarly activities, Schütz held a full-time job as a bank attorney, a dual life that lasted until 1956. After the annexation of Austria to the Third Reich in 1938, Schütz and his family escaped to New York. On his arrival, Schütz got in touch with Talcott Parsons whose *Structure of Social Action* he intended to review. But despite their intense correspondence (Schütz and Parsons, 1978), the contact between Schütz and Parsons broke down. Instead, Schütz became affiliated with the Graduate Faculty at the New School for Social Research in New York where he started teaching as Lecturer in 1943, advancing to a full Professor of Sociology in 1952 and also of Philosophy in 1956. He died on 20 May 1959 in New York.

Major Aims of Schütz' Theory

Alfred Schütz' social theory focuses on the concept of the life world, which is understood as a social–cultural world as it is perceived and generated by the humans living in it. Schütz identifies everyday interaction and communication as being the main processes in which the constitution of the life world's social reality takes place. In this sense, his approach pursues two major goals: the development of a theory of action, which reveals the constitution of social reality with its meaning structure given in the commonly shared typical patterns of knowledge, and a description of the life world with its multiple strata of reality emerging from these constituting processes. These leitmotifs merge in a theory of life world constitution – a project which occupied Schütz during the final years of his life.

Main Intellectual Sources of the Schützian Approach

Schütz' conception originated in the intellectual discourse in the social sciences and philosophy of the early decades of this century. Schütz developed his position from within a triangular interdisciplinary field marked by Max Weber's historising interpretative sociology (*verstehende Soziologie*), the Austrian economic approach represented by Ludwig von Mises, and the philosophical theories of the stream of consciousness formulated by H. Bergson and E. Husserl. From the very onset, he adopted Max Weber's view of social reality as a meaningful social–cultural world, as well as his concept of social action as a meaning-oriented behavior, and shared Weber's search for an interpretative method in sociology. He also shared the methodological individualism advocated by Weber and by the Austrian economists who denoted individual action as the starting point of social research. At the same time, he accepted the need for a generalizing theory of action, as stressed by his teacher Ludwig von Mises, and criticized Weber for neglecting to develop the basic framework of such a theory and especially for not inquiring into the constitution of meaning attached to social action. But Schütz' criticism also addressed the 'general theory of action' based on an 'a priori' postulate of *rational choice* put forth by Ludwig von Mises. Schütz rejected this conception since it imposed an abstract framework on actors' orientation of action and ignored their actual stock of knowledge.

In order to analyze how the meaning attached to action is revealed, Schütz referred to the philosophical concepts of Henri Bergons and Edmund Husserl. Both offer insights into the stream of lived experience and into the acts of consciousness in which our world is meaningfully constituted. Schütz adopted the Bergsonian idea of the stream of consciousness in his first scholarly attempts in 1924–28 (Schütz, 1982), only to later recognize the difficulties in Bergson's intuitivism and to turn to Husserl's phenomenology, which ultimately gave the name to his approach. Husserl was concerned with the acts of consciousness, which establish the meaningful taken-for-grantedness of the world in humans' natural attitude. His term 'life world,' which Schütz later assumes, is aimed at the reality constituted in those acts. Based primarily on this philosophical approach, Schütz wrote his first monograph, *Der sinnhafte Aufbau der*

sozialen Welt, in 1932, in which he developed the basic concept of his theory (Schütz, 1932). Starting with the criticism on Max Weber mentioned above, he devoted himself to the process in which humans create the social world as a reality that is meaningful and understandable to them.

Theory of the Life World with Its Everyday Core

In order to grasp the constitution of the social world, Schütz (1932) had to transcend the realm of consciousness and perception analyzed by Husserl. He included both acts of consciousness as well as human action and interaction into the constitutive process under scrutiny. Departing from the transcendental philosophical approach, he developed his own *mundane* phenomenology, which analyzes the constitution of the meaningful reality within social relationships in the everyday world.

He proceeded in three steps dealing with three problems constitutive for any theory of action, which is searching for the construction of social reality.

1. How does a meaningful orientation of human action emerge?
2. How can we understand the Other?
3. How is socially shared, intersubjective valid knowledge generated?

Following the principle of methodological individualism, Schütz started with the analysis of single individual action. Relying on Bergson's concept of the inner stream of lived experience and on Husserl's investigations on intentionality and temporality of consciousness, Schütz first explored the constitution of meaning in subjects as a temporal process. The meaning attached to a lived experience emerges from successive acts of selective reflection aimed toward it and framing it into a context of other experiences. The schemes of experience on which the primary meaning of action, i.e., its project is based, arise from this basic temporal dynamics and plasticity of consciousness. The temporality of the meaning attached to an action manifests itself in two kinds of motivation attached to its different temporal dimensions: '*in-order-to-motives*,' which direct action toward the future, and '*becausxe-motives*' representing the roots of the action in the past. However, the meaning of the project of action changes during the time when the projected action is going on. The final meaning of an action can be found in the experience perceived by the subject when looking at the changes that have emerged in the meaning structure of the action when the action is completed. The meaning attached to an action and consequently the schemes of experience are thus not only affected by the acts of consciousness but also by the action itself.

In his second step, Schütz proceeded to show how the schemes of experience are shaped by interaction and communication. He perceived communication as a process in which two subjective streams of consciousness are coordinated within a social interaction (*Wirkensbeziehung*). Thus, communication signifies an interaction in which the meaning of ego's action consists in the intention to evoke a reaction of the alter ego. Actions in this sense have the function of signs, which are mutually indicated and interpreted. The ultimate meaning of my action is revealed in the reaction of the Other and vice versa, therefore communication generates a *chain of motives* where my *in-order-to-motives* become *because-motives* of the Other and provide a common stock of shared patterns of interpretation, which allows for mutual understanding even if each of the actors are always referring to their own schemes of experience. In this concept of understanding, based on interaction, Schütz offers his own solution to the problem of intersubjectivity posed by Husserl.

The constitution and appropriation of shared knowledge primarily takes place in long lasting face-to-face interaction (we-relations) where the mutual expectations are learned, verified, and sedimented to typical patterns that can be applied to more remote and anonymous strata of the social world. As a result, the meaning structure of the social world is characterized by *typifications* of actions, situations, and persons generated in interaction and communication.

In these three steps, Schütz laid down the main features of his theory of the life world. In his later work, Schütz (1962, 1964, 1966) determined this communicatively created social reality as the world of everyday life, in which typical patterns are taken for granted, and which represents the intersubjective common core of the reality we live in. Later on (Schütz, 1962, 1970), he disclosed further structural characteristics of this everyday core of the life world. Since its typical structure greatly depends on action, it is also the pragmatic orientation selecting the areas where typification processes take place. Both typicality and this kind of selection, which Schütz designated as *pragmatic relevance* (Schütz, 2003), represent two generative principles of order in the everyday world. They determine its formal structure and at the same time, when realized in action, they shape this structure into a concrete social–cultural world, which is thus characterized by social distribution and differentiation of knowledge. A third moment structuring the everyday world can be found in the rootedness of its constitution in individual action. Here actors and their bodies represent the central point of the *everyday world and its temporal, spatial, and social dimensions* that are arranged into spheres of past and future, of within-reach and distant and of intimacy and anonymity, respectively, to the actors' own position (Schütz and Luckmann, 1973).

Everyday reality is nevertheless not identical with the life world on the whole. By suspending their pragmatic interests, actors are able to modify their everyday experiences and perceive them as objects of a game, fantasy, art, science, or, if the conscious attention is completely absent, as a dream. There are areas – even in the realm of everyday action directed by the principal of pragmatic relevance – which are beyond the sphere of actors' everyday practice and therefore transcend it. All these modifications represent different provinces of meaning transcending the everyday world and constituting *multiple realities* (Schütz, 1962), which make up the life world. Nevertheless, among the different provinces of the life world, the everyday core denotes a *paramount reality* since it is only here that actors, driven by the *fundamental anxiety* when facing the finality of their lives, have to master their living conditions and are engaged in communicative processes producing common knowledge, which makes mutual understanding possible.

Schütz (1962) viewed communication as a substantial constitutive mechanism of social reality and stressed the role of *language* in this process. He considered language – including its syntax and semantics – as an objectivation of sedimented, social provided stock of knowledge that preserves relevances and typification inherent to cultures and social groups, and thus as crucial for the constitution of the life world (Schütz and Luckmann, 1989). Schütz saw language as delineating an essential case of social objectified systems of appresentation, which bridges the transcendence between different areas and meaning provinces of the life world. Another important integrative mechanism in the life world are symbolic structures, which are often based on language, and intermediate between the everyday world and noneveryday realities such as religion, arts, or politics (Schütz, 1962).

Schütz did not restrict his study of the structure of the life world solely to theoretical research. He also applied his theory as a general interpretative scheme to several social fields and problems. He explored courses of intercultural communication and the social distribution of knowledge, the social conditions of equality in modern societies as well as the fields of music and literature (Schütz, 1964, 1982; Embree, 1998).

Consequences of the Theory of Life World for the Methodology of the Social Sciences

Schütz considered the construction of ideal types to be the general methodological device in the social sciences since sociology, economy, political science, etc., explain social phenomena by modeling ideal actors and courses of action, which they endorse with special features as explanatory variables. However, he did not see ideal types in the Weberian sense, i.e., as a scientific model, which does not need to correspond to reality in all respects. For Schütz, the ideal-typical method is legitimized by his findings on the typicality of everyday world, which is the object of social research. Social reality can be approached by type-construction on the scientific level because its immanent structure is typified itself. Thus, the Schützian theory of the life world and its structures represents a methodological tool to bridge the gap between sociological theoretical reasoning and its object. The methodological rule, which Schütz derived from his theoretical approach consists correspondingly in the *postulate of adequacy* (Schütz, 1962, 1964) between everyday and scientific typifications. This postulate holds that ideal types featured by the social sciences have to be constructed in correspondence to the structure of the everyday typifications, so that everyday actors can take the model for granted when they act under the conditions stated in the ideal type. In this sense, social scientists as well as everyday actors have to follow the same frame of reference given by the structure of the life world but their cognitive attitude differ in several respects: scientists neither share the pragmatic interest of everyday actors nor do they share their everyday rationality restricted by their beliefs in the grantedness of typical knowledge. Opposing T. Parson's functionalism and C.G. Hempel's methodologic positivism, Schütz (1962) stressed that scientists must not impose their own theoretical concepts and rationality on their object of study – the life world. First, they must discover and then respect its immanent meaning structure.

The Significance of the Schützian Approach for the Social Sciences

Since the 1960s, Schütz' theory has drawn sociologists' attention to inquiries into everyday interaction, communication as well as to the insight that social reality has to be considered as a construction produced within these processes. Sociologists started to examine the practices of everyday action, communication, and interpretation from which social reality emerges. Garfinkel's (1967) *Ethnomethology*, which was aimed at finding the formal properties of everyday practices, led to a series of case studies covering a wide range of everyday life in society and its institutions. Continuing this line of research, examinations of everyday communication were provided by Schegloff (2007) and Sacks (1995), whose *Conversational Analysis* became a widespread method in qualitative sociological research. A. Cicourel's (1964) *Cognitive Sociology* revealed the constructed character of data in social institutions as well as in science and thus initiated a series of studies in the sociology of organization and in the sociology sciences. *Milieus* as formations of everyday interaction are the subject of the *Social Phenomenology* R. Grathoff (1986).

P.L. Berger and Th. F. Luckmann's (1966) idea of *the Social Construction of Reality* (as being a general process in which cultural worlds emerge) triggered new impulses in both the sociology of knowledge and the sociology of culture reconceived now in Schützian terms. In this context, Schütz' impact can also be seen in the sociology of language and the sociology of religion (Luckmann, 1962, 2002). Contemporary Marxian theory saw the way in which Schütz focused on the everyday practice as a possible mediation between social structure and individual consciousness. The term *phenomenological sociology* was coined in the 1970s (Psathas, 1973) to address the spectrum of approaches oriented and inspired by Schütz' theory. Under this label, Schütz' approach became one of the general paradigms in the interpretative social science and theory of action. The diffusion and empirical application of the Schützian approach enforced the search for qualitative research methods, which would reveal data pertaining to the social construction of social reality in everyday life. Aside from ethnomethodology and conversational analysis mentioned above, this quest especially led to a development of methods in the fields of the ethnography of everyday life (Honer, 1993), social hermeneutics (Soeffner, 1989; Hitzler et al., 1999) as well as to a refinement in the techniques of narrative interviews and in biographical research (Rosenthal, 2005).

Once established in the 1970s, the Schützian paradigm influenced the mainstream of sociological theory, which became sensitive not only to the social construction of life world but also to the phenomenological background of the Schützian theory. The 'life world,' in the sense of a basic social reality provided by humans in their 'natural' intercourse,

became one of the central terms in social theory (Habermas, 1981). The everyday construction of social reality was recognized as a crucial mechanism in which society emerges (Bourdieu, 1972; Giddens, 1976; Bauman, 1991). The phenomenological conception of meaning constitution reformulated as an *autopoiesis* (self-creation) of social and psychic systems influenced the development of the contemporary sociological system theory (Luhmann, 1996).

Beyond the scope of sociology, other humanities also gained innovative impulses from the Schützian approach. In philosophy, Schütz' theory led to a critical assessment of the Husserlian view of intersubjectivity and to the conceptions of a *worldly phenomenology* (Embree, 1988) or of a *philosophy of modern anonymity* (Natanson, 1986). In the literature, Schütz' constructionism inspired the *esthetics of reception*, which pointed out the beholder's participation in cocreating the autonomous reality of literary works (Jauss, 1982). Schütz' concept of the structure of life world also affected theorizing in social geography (Werlen, 1993), educational theory (Meyer-Drawe, 1984), and political science (Voegelin, 1966). New editions of his writings (Schütz, 2003–2013, 2011) as well as the recent studies on his intellectual biography (Endreß, 2006; Barber, 2004; Nasu et al., 2009), both based on his literary estate, reveal that wide scope of Schutz' scientific impact.

See also: Constructivism and Constructionism: Methodology; Culture, Sociology of; Everyday Life, Anthropology of; Husserl, Edmund (1859–1938); Interactionism, Symbolic; Interpretive Methods: Micromethods; Knowledge, Sociology of; Methodological Individualism: Philosophical Aspects; Parsons, Talcott (1902–79); Phenomenology in Sociology; Phenomenology: Philosophical Aspects; Weber, Max (1864–1920).

Bibliography

Barber, M., 2004. The Participating Citizen. A Biography of Alfred Schutz. State University of New York Press, Albany, NY.
Bauman, Z., 1991. Modernity and Ambivalence. Polity Press, Oxford, UK.
Berger, P.L., Luckmann, Th, 1966. The Social Construction of Reality. Doubleday, Garden City, New York.
Bourdieu, P., 1972. Esquisse d'une théorie de la pratique, précéd de trois études d'ethnologie kabyle. Droz S.A., Geneva, Switzerland.
Cefai, D., 1998. Phénomenologie et les sciences sociales: la naissance d'une philosophie anthropologique. Libraire Droz, Geneva, Switzerland.
Cicourel, A.V., 1964. Method and Measurement in Sociology. The Free Press of Glencoe, New York.
Emanuel A. Schegloff, E.A., 2007. Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, Cambridge (UK).
Embree, L., 1988. Wordly Phenomenology: The Continuing Influence of Alfred Schutz on North American Human Sciences. University Press of America, Washington, DC.
Embree, L. (Ed.), 1998. Alfred Schutz's "Sociological Aspects of Literature". Kluwer Academic Publishers, Dordrecht, Boston, London.
Endreß, M., 2006. Alfred Schütz. UVK Verlagsgesellschaft, Konstanz, Germany.
Garfinkel, G., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Giddens, A., 1976. New Rules of Sociological Method. Hutchinson & Co., London, UK.
Grathoff, R., 1986. Milieu und Lebenswelt. Suhrkamp, Frankfurt/M, Germany.
Habermas, J., 1981. Theorie des kommunikativen Handelns. Suhrkamp, Frankfurt/M, Germany.
Hitzler, R., Reichertz, J., Schröer, N. (Eds.), 1999. Hermeneutische Wissenssoziologie. UVK Verlagsgesellschaft, Konstanz, Germany.
Honer, A., 1993. Die lebensweltliche Ethnographie. Deutscher Universitätsverlag, Wiesbadeb, Germany.
Jauss, H.P., 1982. Ästhetische Erfahrung und literarische Hermeneutik. Suhrkamp, Frankfurt/M, Germany.
Luckmann, T., 1962. The Invisible Religion. Macmillan, New York.
Luckmann, T., 2002. Wissen und Gesellschaft. UVK Verlagsgesellschaft, Konstanz, Germany.
Luhmann, N., 1996. Die neuzeitlichen Wissenschaften und die Phänomenologie. Picus, Vienna, Austria.
Meyer-Drawe, K., 1984. Leiblichkeit und Sozialität. Fink, Munich, Germany.
Natanson, M., 1986. Anonymity: A Study in the Philosophy of Alfred Schutz. Indiana University Press, Bloomington, IN.
Nasu, H., Embree, L., Psathas, G., Srubar, I. (Eds.), 2009. Alfred Schutz and His Intellectual Partners. UVK Verlagsgesellschaft, Konstanz, Germany.
Psathas, G. (Ed.), 1973. Phenomenological Sociology. Issues and Applications. John Wiley & Sons, New York.
Rosenthal, G., 2005. Interpretative Sozialforschung. Juventa, München, Germany.
Sacks, H., 1995. Lectures on Conversation. Blackwell, Oxford, UK.
Schegloff, E.A., 2007. Sequence Organization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, Cambridge, UK.
Schütz, A., 1932. Der sinnhafte Aufbau der sozialen Welt. Springer, Vienna.
Schütz, A., 1962. Collected Papers I. Nijhoff, The Hague, The Netherlands.
Schütz, A., 1964. Collected Papers II. Nijhoff, The Hague, The Netherlands.
Schütz, A., 1966. Collected Papers III. Nijhoff, The Hague, The Netherlands.
Schütz, A., 1970. Reflections on the Problem of Relevance. Yale University Press, New Haven, CT.
Schütz, A., 1982. Life Forms and Meaning Structure. Routledge and Kegan Paul, London.
Schütz, A., 1995. Collected Papers IV. Nijhoff, The Hague, The Netherlands.
Schütz, A., 1996. Collected Papers IV. Nifhoff, The Hague, The Netherlands.
Schütz, A., Luckmann, T., 1973. Structures of Life World I. Northwestern University Press, Evanston, IL.
Schütz, A., 2003. Theorie der Lebenswelt 1: Alfred Schütz Werkausgabe, Vol V.1. UVK Verlagsgesellschaft, Konstanz, Germany.
Schütz, A., 2011. Collected Papers V. Springer, Dordrecht, Heidelberg, etc., The Netherlands, Germany, etc.
Schütz, A., 2013. Collected Papers VI. Springer, Dordrecht, Heidelberg etc., The Netherlands, Germany etc.
Schütz, A., 2003–2013. Alfred Schütz Werkausgabe, Vol. I–IX. UVK Verlagsgesellschaft, Konstanz, Germany.
Schütz, A., Luckmann, T., 1989. Structures of Life World II. Northwestern University Press, Evanston, IL.
Schütz, A., Parsons, T., 1978. Theory of Social Action: The Correspondence of Alfred Schütz and Talcott Parsons. Indiana University Press, Bloomington, IN.
Soeffner, H.G., 1989. Soeffner, Hans-Georg: Auslegung des Alltags - der Alltag der Auslegung. Suhrkamp, Frankfurt am Main, Germany.
Srubar, I., 1988. Kosmion die Genese der pragmatischen Lebensweltheorie von Alfred Schütz und ihr anthropologischer Hintergrund. Suhrkamp, Frankfurt am Main, Germany.
Srubar, I., 2007. Phänomenologie und soziologische Theorie. VS Verlag für Sozialwissenschaften, Wiesbaden, German.
Voegelin, E., 1966. Anamnesis. Zur Theorie der Geschichte und Politik. Piper, Munich.
Wagner, H.R., 1983. Alfred Schütz. An Intellectual Biography. University of Chicago Press, Chicago.
Werlen, B., 1993. Society, Action and Space. An Alternative Human Geography. Routledge, London.

Relevant Websites

http://www.waseda.jp/Schutz/AlfredEng.htm – The Alfred Schutz Archive (Department of Sociology/Waseda University).
http://phenomenologycenter.org/about/specialized-bibliographies/schutz/ – Bibliography of Secondary Sources on Alfred Schutz, Center of Advanced Studies in Phenomenology.
http://www.zetabooks.com/schutzian-research-2.html – Schutzian Research.
http://cms.uni-konstanz.de/en/soz-archiv/news/ – Sozialwissenschaftliches Archiv Konstanz, Alfred-Schütz-Memorial-Archives, University of Konstanz.

Science and Development

Marcus A Ynalvez, Texas A&M International University, Laredo, TX, USA
Wesley M Shrum, Louisiana State University, Baton Rouge, LA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W. Shrum, volume 20, pp. 13607–13610, © 2001, Elsevier Ltd.

Abstract

Aimed at a contemporary view of science and development that takes into account changes brought about by information and communication technologies, this article examines why low-income countries are in pursuit of the 'Western way of knowing' known as science. Toward this aim, several perspectives are reviewed: modernization, dependency and world systems, neoinstitutional, global systems, postcolonialism, and reagency. The intent is to highlight how each casts the nature of science and its relation to the project of development.

Introduction

Macrolevel orientations in the sociology of science and in the sociology of international development and social change have alluded to the connection between science and development because national economic growth and scientific capacity (i.e., knowledge production, technological development, technical innovation, and scientific discovery) are empirically associated with each other (Castells, 2000; Grammig, 2002; Malmgren et al., 2010; Schofer, 2004; Schofer et al., 2000). For example, scientific knowledge production and technological development are almost always monopolized by economically powerful countries, such as the Group of Seven and the Organization for Economic Cooperation and Development countries (Bradshaw et al., 1993; Castells, 2000; Gereffi and Fonda, 1992; Migdal, 1988; Sernau, 2012). Countries in Africa (with the exception of South Africa), Asia (with the exception of Japan, Singapore, South Korea, and Taiwan and more recently China and India), and Latin America (with the exception of Argentina, Brazil, and Mexico) have almost always been consumers of but never producers of knowledge, innovation, and technology; beneficiaries of technology transfer programs and initiatives but rarely benefactors. However, when knowledge, innovations, and technologies travel, or are exported to travel from their production sites in developed countries to test sites in developing countries, more often than not these packages of Western knowledge have proved largely ineffective, insensitive to, and incompatible with local needs, and have all too often resulted in a dependent and the uneven development of countries at the periphery of the global economic system (Bradshaw, 1988; Grammig, 2002).

Macrolevel data from the United Nations Development Programme indicate that countries at the high human development level have, on average, about 2335 scientists, technicians, and engineers per million people. Countries at the medium human development level have, on average, about 588 scientists, technicians, and engineers per million, while estimates are not available for countries at the low human development level (UNDP, 2003). Implicitly, these numbers indicate that science and technology are indispensable tools for national competitiveness and hegemony. Indeed, the geography of scientific and technological capacity has a major impact on the centers and networks of global society (Castells, 2000). With the goal of coming up with a contemporary view of science and development that takes into account the changes and shifts brought about by new information and communication technologies (ICTs), this article attempts to explain why developing countries (i.e., low-income countries or peripheral countries) are so enmeshed in pursuing a 'Western way of knowing' called science. Toward this goal, several theoretical perspectives are presented: modernization, dependency/world systems, institutional, global systems, postcolonial, and reagency. The intent is to show how each perspective views the nature of science. Toward the end, the authors propose a perspective that takes into account the changes and recalibrations in the logic (i.e., spatial and temporal) of social interaction resulting from the 'time–space compression' brought about by advances in ICT (Castells, 2000, 2001).

Modernization Theory

Modernization theory casts development as a uniform evolutionary route that all societies follow, from agricultural, rural, and traditional societies to postindustrial, urban, and modern forms (Bradshaw, 1987; Escobar, 1995; Chirot and Hall, 1982; Shrum, 2000). In other words, all societies, once engaged in the modernization process, follow a predetermined sequence of developmental stages: traditional economies, transition to takeoff, takeoff itself, drive to maturity, age of high consumption, and postindustrial society (Chirot and Hall, 1982: 82). Modernization theory emphasizes internal forces and sources of socioeconomic development such as formal education, market-based economy, and democratic and secular political structures. Although modernization theory does not rule out external forces and sources of social change and economic development, it focuses less on foreign influences (Jenkins and Scanlan, 2001; Shrum, 2000).

Among external influences, however, science is exceptional because it is considered beneficial to developing countries by way of 'knowledge and technology transfer' from developed countries (Shrum, 2000). In other words, societies can be

fast-tracked to modernization by 'importing' Western technical capital, forms of organization, and science and technology to developing countries (Herkenrath and Bornschier, 2003; Shrum, 2000). W.W. Rostow argues that adoption of scientific methods and scientific ways of thinking and acquisition of technoscientific skills are critical at the 'transition to takeoff' stage of development (Chirot and Hall, 1982). Essentially, proponents of modernization theory view science and technology as catalysts for development. Science and technology provide conducive environments for economic growth in developing countries through their ability to provide rational protocols in decision making for the efficient use of material and human resources (Shrum and Shenhav, 1995).

As far as modernization theory is concerned, development is simply a matter of knowledge and technology transfer that is unproblematic and straightforward, context free, and not disruptive of existing social and cultural arrangements in developing countries (Herkenrath and Bornschier, 2003). Modernization theory also seems to be unmindful of the fact that much of the knowledge and technology critical for national development and national competitiveness are within the domain of proprietary knowledge production. In a way, modernization theory implies a monolithic, one-way, and top-down development scheme that holds true for all identities, for all time, for all places, and for all contexts. The same holds true for knowledge generation, production, dissemination, and representation. In this top-down development model, the sources of knowledge are foreign to the places and identities to which knowledge is applied or exported. As a model for social change and development, modernization theory fails to consider the possibility of having an interactive and multifarious process of knowledge generation and exchange, which is made possible by recent advances in ICT.

As far as modernization theory is concerned, science is seen as exceptional and different from other institutions in Western developed countries, and is assumed to be independent of and invariant to the limitations of contexts of interpretation and use, and further is viewed as a search, by means of generating objective and rational knowledge claims, for empirical truths and universal laws. Such casting of science is reminiscent of the limited translation model that Callon (1995) critiques. In other words, modernization theory emphasizes the beneficial role of objective, rational science in national socioeconomic development. It de-emphasizes science's possible higher order and negative interactional effects upon local contexts and the identities that populate such contexts.

Dependency and World Systems Theory

Modernization theory stresses the importance of internal, local, and native factors as catalysts for socioeconomic development while making science – an external agent of change – an exception (Chase-Dunn, 1998; Shrum and Shenhav, 1995). Dependency theory (with its three main strains, classical dependency, dependent development, and mature dependency) and the more encompassing world systems theory both stress the important role of external and foreign factors in the development of underdevelopment (Bradshaw, 1987, 1988; Jenkins and Scanlan, 2001). Proponents of dependency theory contend that linkages and relations with developed countries, particularly when mediated by multinational corporations, are a means of control, domination, and exploitation that stifle and stunt economic growth and social development in developing countries (Bradshaw, 1988; Chirot and Hall, 1982; Herkenrath and Bornschier, 2003).

In contrast to the world systems approach, dependency theory focuses on individual countries or nations: their role as sources of raw materials and cheap and disempowered labor, their relegation and assignment to a primary sector type economy, and their functions as markets for high-tech and expensive products of developed countries (Chase-Dunn, 1998; Shrum, 2000). These asymmetrical ties between core (or developed) and peripheral (or developing) countries result in the stunted socioeconomic development of the latter due mainly to the unequal balance of payments and trade. Classical dependency theory states that underdeveloped nations remain dependent because foreign investment in raw material extraction results in highly specialized, low value-added, export-oriented activities that impede the internal structural differentiation necessary for sustainable development (Bradshaw, 1988; Firebaugh, 1992).

A variant of dependency theory known as dependent development theory is characterized by its focus on a triple alliance of foreign capital, domestic capital, and the domestic state (Bradshaw, 1988; Bradshaw et al., 1993; Jenkins and Scanlan, 2001). Dependent development qualifies the argument of classical dependency. It contends that structural dependency on foreign (direct and indirect) investments and foreign markets in the developed countries constrains and distorts, but is not inconsistent with development. Rather, this type of dependency simultaneously triggers rapid economic development in the domestic cores and increasing economic disparities in the domestic peripheries of developing countries.

What is sometimes called mature dependency is exhibited by developing countries that are dependent on foreign – direct and indirect – investment (Bradshaw, 1988; Bradshaw et al., 1993). The nature of this dependency is indicated by the negative relationship between foreign investment and long-term economic development of developing countries. Castells (2000) and Escobar (1994) contend that the advent of new digital technologies results in a new form of dependency of technologically lagging nations on leading nations in ICT development. For these analysts, there is a need to recalibrate the economic and the political relations between developed and developing countries in a cyber economy and society.

Wallerstein's world systems theory focuses on the supra-macronetwork of societies in the modern and industrialized core, the traditional and impoverished periphery, and a group of semiperipheral countries in which mobility and position within the world system is conditioned by the capital, labor, resources, and constraints confronted by distinct sets of nations (Chase-Dunn, 1998; Gereffi and Fonda, 1992; Wallerstein, 1974). This system can be viewed as an international division of labor with the economies of semiperipheral and peripheral countries controlled by the core countries. With their

emphasis on the production of primary, low value-added products, the unchecked economic influence of multinational corporations, and the political power wielded by local capitalist agents, peripheral economies continue to have high levels of socioeconomic inequality, coupled with low levels of democracy and economic growth (Shrum, 2000). This perspective posits that a nation's location and incorporation into the capitalist world economy are the key determinants of national development outcomes (Gereffi and Fonda, 1992).

For Shrum (2000), an interesting point in the dependency/world systems account is that it is naïve to construe science as benign or neutral. When fostered by foreign investments brought in by multinational corporations into developing countries, science translates into one of a group of institutions that contribute to the development of 'sustained underdevelopment' in developing countries (Bradshaw, 1988; Shrum, 2000). Science is seen as part of an aggressively debilitating 'social cancer' when it is generated in core countries and then introduced into peripheral countries. As a case in point, scientific activities are disproportionately concentrated in core countries, and most such activities are conducted for the benefit of those countries, with applications directed at advancing their political hegemony and economic interest. The invention, patenting and production of technical innovations and process for maximal profit by core countries (and international corporations headquartered in them) typically go with the introduction and diffusion of commodities that are often inconsistent with the needs and conditions in peripheral countries, serving to draw raw materials and scarce resources away from more pressing developmental activities (Bradshaw, 1988; Shrum, 2000). Indeed, within the theory of dependency/world systems, science is a vehicle for world economic domination and for the maintenance of geopolitical hegemony of agents in the core countries.

Dependency/world systems theory explanations of the link between science and development are derived from both neoclassical and Marxist economics. On the one hand, neoclassical economics contends that knowledge producers are very much like capitalists who are driven by rationality and are oriented toward maximizing personal profits and extrinsic rewards. On the other hand, Marxist economics contends that knowledge producers are not only interested to produce knowledge for its own sake, but also to produce knowledge in research areas where profits and rewards are high. Granted, it is then logical to argue that knowledge producers play an important background role in the maintenance of dependency between core and peripheral countries, or between developed and developing countries. Knowledge producers comprise the needed technoscientific workforce that would facilitate the conversion from basic to applied knowledge. This conversion serves as the critical precursor for the creation and production of proprietary science and technology, which ultimately make up the essential components of foreign direct and indirect investment made by multinational corporations in developing countries. These investments facilitate the conversion of resources – using cheap, disempowered, and typically feminized labor – to commodities. Hence, it is reasonable to assert that the dependency/world systems theory's view of science is in sync with Callon's (1995) competition view in which the knowledge production process is driven by rationality, profit generation, and the accumulation of extrinsic rewards.

Neoinstitutional Theory

In its attempts to explain why developing countries are so much involved in the adoption and expansion of science – despite its generally unpredictable consequences – neoinstitutional theory calls attention to the role of international economic and financial organizations such as the International Bank for Reconstruction and Development (World Bank), International Monetary Fund, United Nations Education, Science, and Cultural Education, and other multilateral organizations (Schofer, 2004; Shrum, 2000). According to Schofer (2004), international organizations espouse science and encourage its dissemination in developing countries by deploying rhetoric to the effect that science is a core instrument for development (Drori et al., 2003). Indeed, much of the involvement of developing countries is founded on a belief in the universality of science as means to national socioeconomic development.

This belief, which inspired the institutionalization of science, has been introduced by Western development experts, propagated through technical assistance programs and monetary lending policies of international economic and financial institutions, affirmed in development plans and discourse, supported by international policy makers, and transmitted through Western education systems (Schofer et al., 2000; Schofer, 2004). Two closely related processes that give real force to the faith in science as an instrument for national socioeconomic development are those of professionalization and of institutionalization (Escobar, 1995; Schofer, 2004). The first refers to a process that draws developing countries into the politics of expert knowledge and Western science; the second refers to the creation of institutions (e.g., development banks such as the Asian Development Bank and the African Development Bank) from which discourses about development are formulated, generated, and disseminated to developing member countries (Escobar, 1995).

In essence, neoinstitutional theory is an account of the emergence, morphology, and expansion of academic and government research sectors as they become successful organizations in developed countries (Shrum, 2000). These organizations serve as models for developing countries to adopt in order for them to achieve the high levels of social and economic development characteristic of advanced Western societies. Neoinstitutional theory's construction of the nature of science is a hybrid of an objective and rational view with Callon's (1995) perspective on science as competition. It argues that scientists, experts, and scientific communities are principal exponents of universalism, espousing the idea of a global, universal, and context-independent science system (Shrum, 2000). Although neoinstitutional theory does not stress conflict, or the power and interests of actors, it describes processes fueled by competition.

As an explanation of why countries are so much dedicated to and involved in scientific pursuits, neoinstitutional theory is

enlightening because it identifies yet another factor external to a country – not multinational corporations in this case, but international economic and financial institutions – that triggers the spread and expansion of science in developing countries while legitimizing and reinforcing the idea that science is tantamount to development.

Global Systems Theory

Modernization theory emphasizes intranational (domestic and internal) forces, processes, structures, and organizations as influences on socioeconomic development. In contrast, dependency or world systems and institutional theory emphasize international (foreign and external) forces, processes, structures, and organizations. Despite their differences, these perspectives explain development through a nation-centric approach. They focus on the authority and power of the nation-state in effecting social change, and utilize nation-states as the units of analysis (Sernau, 2012). Furthermore, these perspectives assume the dominance of 'linear time' over 'space of place' – a dominance that contemporary social theorists claim is typical of social perspectives prior to the advent of ICTs (Castells, 2000; Sklair, 2000).

In contrast, global systems theory emphasizes transnational forces, processes, linkages, and organizations – which cross national borders but do not derive power and authority from actors and organizations associated with the nation-state – as forces configuring socioeconomic development (Bradshaw et al., 1993; Pinnington and Gray, 2007; Sernau, 2012). Linkages refer to transnational economic linkages, examples of which are foreign aid, foreign trade, foreign direct investment, and foreign loans (Bradshaw et al., 1993). Social theorists such as Castells (2000), Sassen (2000), and Sklair (2000) describe contemporary global society as transnational in nature, with a spatial logic that follows the 'space of flows' instead of the 'space of place,' and with a temporal logic that follows 'random time' instead of 'linear time.' According to these theorists, such shifts and recalibrations in the fundamental logic of social interaction are consequences of ICTs' capacity for high-speed communication and high-quality information exchange among globally dispersed actors, groups, and organizations.

Global system theory posits that contemporary economic and social inequalities are moving away from the traditional divisions among nations – core, semiperiphery, and periphery – and societal sectors – urban, suburban, and rural (Bergesen and Bata, 2002; Sassen, 2000; Sklair, 2000). Instead, inequalities are shifting toward a divide between membership and nonmembership in what is now called the transnational capitalist class (TCC), or the 'globally coordinated elite' (Sklair, 2000). The TCC is situated in and identified with no particular nation, state, or country; instead, it derives and exercises power and authority through transnational corporations (TNCs), and pursues resources – material and human – all over the world in its insatiable desire for profit and capital accumulation on a global scale (Evans, 2005). As a class, TCC comprises corporate executives, globalizing bureaucrats and politicians, globalizing professionals, and consumerist elites (Castells, 2000; Sklair, 2000). In the global economy, TNCs increasingly erode or circumvent the power of the state. This means that such transnational actors stand in an increasingly objective global position with respect to other global classes as opposed to the social classes within nation and states (Bergesen and Bata, 2002).

In other words, socioeconomic development is very much a function of whether an actor (in this case, an individual, a group, or an organization) is tied to the network structure of the TCC. This network structure comprises actors who possess capital (i.e., cultural, organizational, political, and technoscientific capital) that is critical for advancing the interests of the TCC. It is important to note that the possession of such capital is a necessary condition for an actor to be connected to the network structure. The network structure also consists of connections that serve as global channels for the movement of data, information, knowledge, and resources necessary for the further accumulation of wealth by the TCC. Hence, actors outside this network structure constitute an emerging domain called the global periphery, while actors connected to this network structure comprise the global core.

This account views science as a global network that relies on communication among scientists and other key actors around the world, wherein ICT expansion contributes to the formation of a global scientific community (Castells, 2000). However, despite the globality of science, its orientation and practice are geared toward the concerns and interests of the TCC (Castells, 2000; Sklair, 2000). In other words, global systems theory frames the nature of science as globally networked and coordinated, generated and produced by actors within science, influenced and configured by actors outside science, and generally oriented toward and distributed according to the interests and goals of the TCC. In essence, global system theory assumes the nature and dynamics of science to be uniform across network sites (i.e., localities and levels) and resistant to the social and cultural influences of localities.

As for ICTs' role in the formation of a global scientific community, this account assumes that such technologies have a uniformly enhancing effect on the content and intensity of communication, collaboration, productivity, and visibility of scientists in different network sites. However, recent studies on scientific research systems at the global periphery criticize these characterizations of contemporary science for being too one-dimensional, shallow, and devoid of empirical substantiation. Such criticism deserves serious attention in light of recent findings that transnational scientific collaboration and productivity do not immediately follow or result from ICT expansion and adoption at the global periphery. Despite the rapid diffusion of ICTs in peripheral areas, scientists' communicative behavior, collaboration patterns, professional networks, and research productivity are largely shaped by cultural, organizational, social, and sectoral contexts (Duque et al., 2005; Ynalvez et al., 2005, 2010).

Postcolonialism and Reagency

Postcolonial theory challenges the entrenched assumption that Western knowledge – as manifested by Western science and

technologies (broadly construed as idea systems, initiatives and projects, and technical artifacts) – is authoritative, objective and unbiased, reliable, universally applicable, and value-free (Anderson, 2002; Anderson and Adams, 2008). Like proponents of global systems theory, postcolonial theorists shy away from analyzing, critiquing, and framing science and development from a macrolevel, state-centered vantage point. Unlike global systems theory, proponents of postcolonial theory are hesitant to focus heavily on transnational forces, institutions, linkages, and processes, such as on the role of transnational networks. Instead, they emphasize new ways to examine the politicoeconomic dimensions of science and social change; the mutual and reciprocal reorganization between the global and the local; the increasingly global flow of cultures, information, people, practices, and technologies; and debates about intellectual property (Anderson, 2002).

An example of postcolonial theory is the work of Helen Verran (2002), who juxtaposes and describes two different Australian knowledge bases – among indigenous people and Western scientists – about firing strategies – practices for restoring land with controlled burns – with the intent of enhancing opportunities and increasing possibilities of cooperation and understanding between these two distinct groups with their very different ways of knowing. Indeed, postcolonial theory attempts to do so, without demarcating geopolitical boundaries between nation-states and empires – demarcations that have occupied central roles for casting and understanding social change – and pushes vigorously for a focus on 'contact zones' (Pratt, 1992). Contact zones are not geographic places with stable and symbolic meanings. Instead, they represent attempts at hegemony, but are – at the same time – sites of multivocality: of negotiation, borrowing, exchange, and interaction; and of redeployment, redefinition, reversal, and even reagency (Gilbert Joseph, quoted in Anderson, 2002; Shrum, 2005; Ynalvez et al., 2010). In other words, as a perspective for understanding science and development, postcolonial theory has the potential to explain how scientific endeavors become sites for generating and linking local and global identities, as well as sites for disrupting and challenging the distinction between the global and the local (Anderson, 2002; Anderson and Adams, 2008).

Most recently, the field of science and technology studies has seen the emergence of a reagency perspective from a long-term collaboration by a group of investigators in Africa and Asia. Unapologetically contrary to the very notion of development (Ynalvez et al., 2010), the original idea of reagency was a sensitizing concept based on the analogy with a chemical reagent – initiating certain kinds of reactions, often unpredictable. In the context of low-income areas, development on the ground is generally a resource intervention that stimulates chains of encounters, but without direction or progress in any determinate sense (Shrum, 2005). Reagency studies tend to focus on scientific and technological networks, donors (multinational and bilateral), and nongovernmental organizations that initiate programs and projects with explicit stories of making a difference, sustainability, and local participation.

The reagency perspective views development as a network of actors and encounters where knowledge, agendas, initiatives, and programs move across geographical boundaries (generally from higher to lower income areas) generating resource flows, bits of microorganization, and stories Shrum (2005); Ynalvez et al. (2010), Shrum et al. (2010). Illustrating this approach is a study by Harsh et al. (2010) that examines the implications of resource chains for nongovernmental organizations with projects in Kenya. Using the logic of reagency, they examine the difficult strategic choice faced by agencies and programs between targeting particular aid recipients (with massive inefficiencies) or efficient distribution of resources (with wildly divergent flows). Reagency shares the critical focus on the combined strength of human and nonhuman actors made famous by actor network theory (Latour, 2005). Rejecting such essentializing concepts as hegemony, center, and periphery, as well as the idea of an organized TCC, reagency understands social processes as local clusters of encounters and knowledge as an unproblematized set of stories. It provides science and technology studies with a concept and framework that emphasizes the unexpected results that come about from the interaction of initiatives and programs with local contexts, identities, and meanings.

Conclusion

Based on the foregoing review, the authors propose an approach to science and technology studies that takes a more 'situated' investigation of the connections between science and development. This emergent approach is essentially a hybrid between global systems theory and reagency, with some qualifications and alterations. Although this hybridized approach focuses on concrete transnational movements of capital, culture, knowledge, and technological innovation, it expresses caution in assuming global uniformity in the dynamics, nature, orientation, style, and practice of science. It rejects the assumption of uniformly positive impacts of ICTs on scientific communication, collaboration, networking, and productivity across localities with varied and distinct cultural, economic, historical, political, and social configurations.

In contrast to modernization, dependency or world systems, and neoinstitutional theories, this hybrid approach questions the following assumptions about development, knowledge production, social networks, and ICT diffusion at the global scientific periphery: (1) that development is a unidirectional, monotonic process of socioeconomic change that faithfully traverses certain trajectories; (2) that the diffusion of ICTs has globalized science, through networking and collaboration between knowledge producers from the core and periphery (Escobar, 1995; Shrum, 2000). Instead, what occurs as knowledge and technologies travel from one place to another is better described by the concept of reagency than by the ideologically loaded concept of development.

Whatever the approach to science and development, an important, unanswered question does center on the role of new ICTs in the process of research. The conventional view is that the diffusion of ICTs will generate the kinds of network shifts that have been characterized as globalization. But empirical findings on global research networks are mixed (Shrum, 2005: Ynalvez and Shrum, 2011; Ynalvez et al., 2010) and one lacks

convincing generalizations to characterize changes in cross-national networks brought about by new media. In time, some homogenization might occur, but shifting patterns in the global research network characterize the past 20 years. Actors in the outer space of science (Shrum et al., 2010) are more heterogeneous and disinterested than one may wish to contemplate.

Acknowledgment

The section starting from the introduction to global systems theory has been adapted with modifications from Ynalvez (2006) with permission from ProQuest Information and Learning Company.

See also: Actor-Network Theory; Development Studies; Indigenous Knowledge: Science and Technology Studies; Patenting in the Biological Sciences; Postcoloniality; Science and The State; Science, New Forms of; Science, Sociology of; Technology, Social Construction of.

Bibliography

Anderson, W., 2002. Postcolonial technoscience. Social Studies of Science 32, 643–658.

Anderson, W., Adams, V., 2008. Pramoedya's chickens: postcolonial studies of techoscience. In: Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge, MA, pp. 181–204.

Bergesen, A.J., Bata, M., 2002. Global and national inequality: are they connected? Journal of World Systems Research 8, 130–144.

Bradshaw, Y., 1987. Urbanization and development: a global study of modernization, urban bias, and economic dependency. American Sociological Review 52, 224–239.

Bradshaw, Y., 1988. Reassessing economic dependency and uneven development: the Kenyan experience. American Sociological Review 53, 693–708.

Bradshaw, Y., Kim, Y.J., London, B., 1993. Transnational economic linkages, the state, and dependent development in South Korea, 1966–1988: a time series analysis. Social Forces 72, 315–345.

Callon, M., 1995. Four models for the dynamics of science. In: Jasanoff, S., Markle, G.E., Petersen, J.C., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage Publication, Inc., Thousand Oaks, CA, pp. 29–63.

Castells, M., 2000. The Information Age: Economy, Society and Culture: The Rise of the Network Society. Blackwell Publishers, Inc., Malden, MA.

Castells, M., 2001. The Internet Galaxy. Reflections on the Internet, Business, and Society. Oxford University Press, Inc., New York.

Chase-Dunn, C., 1998. Global Formation: Structures of the World Economy. Basil Blackwell, New York.

Chirot, D., Hall, T., 1982. World-system theory. Annual Review of Sociology 8, 81–106.

Drori, G.S., Meyer, J.W., Ramirez, F.O., Schofer, E., 2003. Science in the Modern World Polity: Institutionalization and Globalization. Stanford University Press, Stanford, CA.

Duque, R.B., Ynalvez, M., Sooryamoorthy, R., Mbatia, P., Dzorgbo, D.B.S., Shrum, W., 2005. Collaboration paradox: scientific productivity, the internet, and problems of research in developing areas. Social Studies of Science 35, 755–785.

Escobar, A., 1994. Welcome to Cyberia: notes on the anthropology of cyberculture. Current Anthropology 35, 211–231.

Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.

Evans, P., 2005. Counter-hegemonic globalization: transnational social movements in the contemporary global political economy. In: Janoski, T., Alford, R.R., Hicks, A.M., Schwartz, M. (Eds.), The Handbook of Political Sociology. Cambridge University Press, Cambridge, UK, pp. 655–670.

Firebaugh, G., 1992. Growth effects of foreign and domestic investment. American Journal of Sociology 98, 105–130.

Gereffi, G., Fonda, S., 1992. Regional paths of development. Annual Review of Sociology 18, 419–448.

Grammig, T., 2002. Technical Knowledge and Development: Observing Aid Projects and Processes. Routledge, New York.

Harsh, M., Mbatia, P., Shrum, W., 2010. Accountability and inaction: NGOs and resource lodging in development. Development and Change 41, 253–278.

Herkenrath, M., Bornschier, V., 2003. Transnational corporations in world development: still harmful? Journal of World Systems Research IX, 105–139.

Jenkins, J.C., Scanlan, S.J., 2001. Food security in less developed countries, 1970 to 1990. American Sociological Review 66, 718–744.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor Network Analysis. Oxford University Press, New York.

Malmgren, R.D., Ottino, J.M., Amaral, L.A.N., 2010. The role of mentorship in protégé performance. Nature 465, 622–626.

Migdal, J.S., 1988. Strong Societies and Weak States: State-Society Relations and State Capabilities in the Third World. Princeton University Press, Princeton, NJ.

Pratt, M.L., 1992. Imperial Eyes: Travel Writing and Transculturation. Routledge, New York.

Sassen, S., 2000. Cities in a World Economy. Pine Forge Press, London.

Schofer, E., 2004. Cross-national differences in the expansion of science, 1970–1990. Social Forces 83, 215–248.

Schofer, E., Ramirez, F.O., Meyer, J.W., 2000. The effects of science on national economic development, 1970–1990. American Sociological Review 65, 866–887.

Sernau, S., 2012. Global Problems: The Search for Equity, Peace and Sustainability. Pearson/Allyn & Bacon, Boston.

Shrum, W., 2000. Science and story in development: the emergence of non-governmental organizations in agricultural research. Social Studies of Science 30, 95–124.

Shrum, W., 2005. Reagency of the internet, or, how i became a guest for science? Social Studies of Science 35, 723–754.

Shrum, W.M., Duque, R.B., Ynalvez, M.A., 2010. Outer space of science: a video ethnography of reagency in Ghana. In: Meusburger, P., Livingstone, D.N., Jons, H., Stiftung, K.T. (Eds.), Geographies of Science. Springer, Dordrecht, pp. 151–163.

Shrum Jr., W., Shenhav, Y., 1995. Science and technology in less developed countries. In: Jasanoff, S., Markle, G.E., Petersen, J.C., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage Publications, Inc., Thousand Oaks, CA, pp. 621–651.

Sklair, L., 2000. The transnational capitalist class and the discourse of globalization. Cambridge Review of International Affairs 14, 67–85.

UNDP, 2003. Human Development Report 2003, United Nations Development Programme. Oxford University Press, New York.

Verran, H., 2002. A postcolonial moment in science studies: alternative firing regimes of environmental scientists and aboriginal landowners. Social Studies of Science 32, 729–762.

Wallerstein, I., 1974. The Modern World System. Academic Press, New York.

Ynalvez, M., Duque, R.B., Mbatia, P., Sooryamoorthy, R., Palackal, A., Shrum, W., 2005. When do scientists "adopt" the internet? Dimensions of connectivity in developing areas. Scientometrics 63, 39–67.

Ynalvez, M.A., Duque, R.B., Shrum, W.M., 2010. Shaping research in developing areas. In: Dutton, W.H., Jeffreys, P.W. (Eds.), Worldwide Research: Reshaping the Sciences and Humanities. MIT Press, Cambridge, MA, pp. 325–342.

Ynalvez, M.A., Shrum, W.M., 2011. Professional networks, scientific collaboration, and publication productivity in resource-constrained research institutions in a developing country. Research Policy 40, 204–216.

Science and Law

Sheila Jasanoff, John F. Kennedy School of Government, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews scholarship on the nature and implications of the law–science relationship. Four major themes are identified: first, how the two institutions are thought to shape the authority structures and practices of modernity; second, how law affects the emergence and professionalization of scientific and technical disciplines; third, the role of expert witnesses and rules governing the admissibility of expert testimony; and fourth, how law functions as a site for generating new scientific knowledge and technical expertise.

The relationship between law and science has occupied scientists and philosophers since the early modern period, but in the twentieth century it also became a topic of social analysis and policymaking. Law–science interactions are significant for the social and behavioral sciences in two respects. First, because of their foundational role in modern social order, law–science interactions have emerged as a distinct branch of social science inquiry, though not one that can be encompassed within a single disciplinary envelope. Studies on this topic derive in the first instance from scholars and practitioners of law, but also from neighboring fields such as anthropology, sociology, and increasingly science and technology studies. Second, the intertwining of these two institutions affects developments in science as well as society, including the sciences of human behavior and action. These developments in turn have prompted investigation and critique by analysts of scientific knowledge in the making.

Three long-held presuppositions had to be set aside in order to position law–science relations within the scope of social science inquiry: first, that science represents 'truth to nature' and hence has no significant sociality; second, that law itself is, or should be, a science; and, third, that science and law operate independently of one another. Science was seen by its early practitioners as the effort to discern and reveal laws inherent in nature – in effect as god's laws, hence universal, immutable, and true. In time, law came to be seen as an institution in need of rationalization, to be made more like a science if possible in order to avoid arbitrary outcomes. For much of the twentieth century, too, legal theorists and practitioners regarded law and science as independent institutions, each with its own objectives, ethos, organization, and practices. Similarities and differences between the two fields preoccupied many observers, with frequent commentary on misalignments between law's demand for justice and science's commitment to truth.

By the end of the century, however, the role of human agency in making both science and law became more apparent, as did the interdependence of the two institutions. A new focus on the law's instrumental uses of science, and more generally expertise, led to divergent schools of thought about the epistemological status of scientific evidence. Those views in turn influenced thinking about how legal institutions should interact with scientific and technical experts. An onslaught of technically complex litigation caused judicial systems to reassess the rules of evidence. Analysts of the legal process grappled with wider questions about the role of the jury (in common law countries), the proper qualification of experts, the law's capacity to act in the face of scientific and social uncertainty, and the knowledge-producing and processing functions of legal proceedings.

This article reviews the resulting scholarship on the nature and implications of the law–science relationship. Four major strands are identified. One strand concerns the ways in which social analysts have conceptualized the interplay of these two institutions in shaping the authority structures and practices of modernity, particularly those legitimizing exercises of power in democratic societies. A second area relates to the law's impact on the emergence and professionalization of scientific and technical disciplines, especially in the social and behavioral sciences. A third focuses on responses by courts and their critics to the expanding influence of expert witnesses and associated shifts in rules governing the admissibility of expert testimony. Fourth and finally, a fast-growing stream of research looks at the law as a site for generating new scientific knowledge and technical expertise and inquires into the implications of this mode of knowledge-making for individual and social justice.

Law, Science, and the Transition to Modernity

Since the origins of modern scientific thought, the term 'law' has been used to describe both regularities discernible in nature and rules laid down by religious or secular authorities to guide human conduct. Both kinds of laws, it was once assumed, embody unchanging principles that work always and everywhere in the same fashion, even though Robert Boyle, a founder of experimental science, cautioned against a literal assimilation of the behavior of matter under natural law to the behavior of human agents under civil law (Shapin, 1994: pp. 330–331). For H.L.A. Hart, the renowned British legal theorist, certain natural facts (such as the approximate equality of humans, their vulnerability, and the limits of their resources and altruism) underwrote a 'natural law' without which human societies would not be able to retain their salient moral characteristics (Hart, 1961). Taking a more sociological tack reminiscent of Robert Merton's well-known 1942 essay on the norms of science, the American legal philosopher Lon Fuller argued that law and science resemble each other in possessing an internal morality that results from the distinctive

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86107-3

arrangements, practices, and fiduciary demands of each institution (Fuller, 1964). Later legal scholarship contrasted the 'cultures' of law and science and likewise underscored their normative and procedural particularities (Goldberg, 1994; Schuck, 1993).

Yet, historical research shows that science and law have continually drawn on each other's institutional strengths and attributes. Thus, early experimental science derived its legitimacy from enlisting the testimony of witnesses, both real and 'virtual' (Shapin and Schaffer, 1985), as if in a court of law. In reverse, some argued that science's self-regulating, nonhierarchical modes of governance offered a model or prototype for the democratic operation of society (Polanyi, 1962). The Enlightenment of the eighteenth century and the rise of liberal democracies brought science and government together in pragmatic projects to ameliorate people's lives (Price, 1985). In the United States, Thomas Jefferson drafted the first Patent Act of the fledgling American republic in 1793, giving operational effect to the constitutional provision that Congress has power "to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." More recently, the partnership of scientific and political authority has furthered instrumental rather than idealistic ends. Modern governments turned to science and technology to discipline citizens and to secure their assent to state rule (Foucault, 1979). State-sponsored technology programs to some degree bypassed consensus-building through deliberation, the staple mode of democratic lawmaking. Instead, states sought to establish their effectiveness through public demonstrations. In projects ranging from vaccine development to space exploration to military exploits, publics were called on to attest to the state's performance, much as scientists bore witness to each other's observations in early experimental laboratories (Ezrahi, 1990). Unlike legislation, however, the design and deployment of technologies by the state often resisted democratic supervision and control (MacKenzie, 1990; Winner, 1986).

The law's dependence on science was perhaps most transparently visible in the US, a society in which lawsuits abound and routinely pave the way to significant policy change. By the beginning of the twentieth century, ideological and political dogma favored the belief that state social policies could and should be responsive to empirical evidence from the sciences. An icon of the progressive era, U.S. Supreme Court Justice Louis D. Brandeis, pushed this development through his historic 'Brandeis brief' in *Muller v. Oregon* (1908). Written to defend Oregon's 10-hour working-day restriction on women's employment, the 100-page brief is widely regarded as the first attempt to bring social science arguments into the courts. Brandeis, then a public interest lawyer in Boston, successfully argued that women were sufficiently different from men to justify special legislative protection, thereby distinguishing the Oregon case from the Court's earlier, infamous decision in *Lochner v. New York* (1905), which struck down a similar state law curtailing working hours for male bakers.

Social science evidence received a further boost when, in overturning the 'separate but equal' doctrine that supported racial segregation in public schools, the U.S. Supreme Court in *Brown v. Board of Education* (1954) relied on psychological studies of segregation's 'detrimental effect upon the colored children.' In later years, students with disabilities benefited from similar sorts of evidence in claiming rights to special education (Minow, 1990). Employment discrimination, racial segregation, gay rights, antitrust, capital punishment, and environmental liability were among the many litigated subjects that brought floods of social science testimony, often involving new methods of counting and quantification, into the US courts through the turn of the twenty-first century. Scientific professional societies, such as the American Psychological Association and the American Sociological Association, filed *amicus curiae* (friends of the court) briefs in important appellate cases, to both assert the social value of their branches of science and their own right to declare what counts as authoritative knowledge with respect to the issues at stake. These developments kept alive the view that objective scientific information about society could be invoked in legal proceedings to combat recognized forms of oppression and social injustice.

The relationship between law and science in modern Europe was no less pervasive, though in some ways more covert than in the United States, where litigation regularly serves as politics by other means. The crises of industrialization and the rise of the welfare state created a demand for credible public action supported by science. States responded with rapid increases in their institutional capacity to order and manage society through newly accredited social science disciplines (Wagner et al., 1991). In turn, the classifications and measurement technologies offered by the human sciences provided the basis for new forms of self-understanding, normalization, and social control. Whether in medicine, psychology, criminology, or studies of intelligence (Carson, 2007), science offered criteria for dividing people into categories of normal or pathological, mainstream or deviant. The scientific study of human characteristics thus enabled members of society to see and discipline each other without the need for constant, formal regulatory oversight by the state (Foucault, 1990, 1979). Science, on this account, not only underwrote the state's formal legal authority to govern its subjects, but also replaced some of the ordering functions of positive law.

The demand for science-based decision-making created a widening sphere for expert advice to governments around the world. Studies of 'regulatory science' (Jasanoff, 1990) called attention to policy-relevant knowledge as a distinctive domain of knowledge production, different from research science in its contexts of production and application, and presenting its own problems of credibility, accountability, and quality control (Lentsch and Weingart, 2011). More, however, was at stake in the shift from science, anchored by the ideal of truth to nature, to expertise, rooted in the exercise of judgment. Nature seems to speak for itself through science; experts, in contrast, are mere humans, and therefore fallible, and they are called upon in just those situations where certainty about facts is most elusive. What, then, accounts for the law's faith in experts? At issue in investigations of regulatory science were the foundations of trust in experts' cognitive authority and the state's performances of public reason. The need for comparative, cross-cultural research on such issues became evident (Jasanoff, 2005), though research capacity, including relevant linguistic and cultural knowledge, remained in short supply.

Disciplines, Professions, and the Law

The interactivity of law and science spurred new disciplinary formations and helped change professional norms, especially in the human sciences, which serve both as putatively neutral allies in regulatory and law enforcement programs and as uneasy subjects of the law's skeptical gaze. This dual relationship between law and the social sciences – at once necessary and suspect – prompted new lines of research and theorizing, along with periodic encroachment by legal authorities upon science's professional autonomy. Developments in the fields of risk studies and psychology clearly illustrate these patterns.

The Social Sciences of Risk

The scientific study of risk grew out of state efforts to manage the hazards of industrial production, complementing modern governments' reliance on older formal decision-making tools such as cost–benefit analysis. Uncertainty about technology's risks and benefits increased through the twentieth century, along with the size and complexity of networks of production and distribution. From village-based economies, in which injured parties knew with fair certainty the causes of the harms they had suffered, the global economy pulled consumers into webs of relationships that diffused both information about risks and responsibility for regulation and compensation. Epitomizing the plight of industrialization's innocent victims, millions of asbestos workers worldwide were unknowingly exposed to severe health risks, and hundreds of thousands died of asbestos-induced illnesses. As the magnitude of such disasters became apparent, legislatures sought to enact preventive rules, ensuring that proof no longer needed to arrive in the form of 'body counts.' Risk, not harm, became the basis for action under new laws safeguarding health, safety, and the environment; and state agencies struggled to develop scientific techniques for measuring, assessing, and mitigating risks before injuries occurred on mass scales.

From relatively modest beginnings rooted largely in the engineering and physical sciences (as in calculating the risk of accidents in nuclear power plants), risk assessment to support regulatory standards grew into an important, and contested, branch of the social sciences by the end of the twentieth century. Risk studies acquired markers of professionalization, with dedicated journals, societies, curricula, and research centers at major universities. An important focus of work in this field, particularly in the United States, was the construction of reliable models for measuring the probability of harm, ranging from the use of animals as stand-ins for humans to sophisticated methods of representing uncertainty in risk estimates (National Research Council, 1994). Risk assessment presented not only methodological but also institutional challenges for governments needing to persuade industry and the public that regulatory decisions were politically balanced and scientifically sound (Jasanoff, 1990). Lawsuits questioning the risk assessment practices of US federal agencies were commonplace and led to significant judicial decisions, including Supreme Court rulings on the occupational standard for benzene in 1980, on air quality standard for particulates in 2001, and on the Environmental Protection Agency's failure to regulate greenhouse gases in 2007.

Other researchers meanwhile approached risk from the standpoint of social psychology, asking why lay observers frequently diverge from experts in judgments concerning the relative magnitude of different kinds of risk. This question produced a sizable and influential body of experimental and survey work on risk perception (Slovic, 2000), as well as theoretical critiques of such work (Irwin and Wynne, 1996). While some social scientists focused on identifying and measuring 'real' risks, and public perceptions of them, others treated *risk* itself as a new organizing principle in social life, noting that vulnerability to risk no longer maps neatly onto older divisions of society according to race, class, and economic status (Beck, 1992).

A new line of research at the intersection of economics and cognitive psychology rose in prominence and policy relevance by the turn of the century. At the base were studies showing that individuals make – from an economic standpoint – systematically irrational choices when confronted with certain kinds of judgments, including decisions about risk (Kahneman, 2011; Kahneman et al., 1982). Originally designed to contest core assumptions about human rationality in economics, these studies demonstrated that the capacity to choose rationally is often compromised by people's prior assumptions and experiences. These biases and heuristics frequently lead to overly fearful or overly optimistic assessments of loss and gain that do not correspond well to actual measured probabilities of events. Findings from the new decision sciences quickly migrated into law and policy. In particular, for government regulators seeking a politically neutral basis for governance, these findings supplied welcome grounds for the 'libertarian paternalism' of 'nudging,' that is, structuring choice frameworks so as to steer people toward making what policy specialists deemed to be the right or most rational choices (Thaler and Sunstein, 2008).

In sharp contrast to these approaches grounded in individual cognitive psychology, the US environmental justice movement generated politically persuasive, if also methodologically contested, data suggesting that social inequality still matters profoundly in the distribution of industrial risks, which fall disproportionately on poor and minority populations (Bullard, 1990). This research led in 1994 to a presidential executive order demanding that US federal agencies should consider the equity implications of their actions. More generally, social impact analysis emerged as an offshoot of existing methodologies for assessing risk, and statistically grounded claims of environmental injustice offered new grounds for demanding regulatory action.

Psychology and Law

Unlike formal risk analysis, which is largely a product of twentieth century concerns, law and the mental health disciplines have existed in close reciprocity since at least 1843, when the British House of Lords laid down the so-called M'Naghten rule for determining the responsibility of persons who try to excuse their unlawful behavior on grounds of insanity. The human sciences' claimed ability to sort people's behavior and capacities into seemingly objective, 'natural' categories, such

as insanity, served the needs of both democratic and totalitarian states in a variety of legal contexts. In US law, for example, psychiatric evidence was used to determine whether an accused criminal could stand trial, whether conditions for sentencing leniency or parole were met, and whether a criminal convicted of a heinous crime posed a sufficient threat of long-term dangerousness to deserve the death penalty. These intersections provided new professional opportunities for psychiatry and psychology, spawning a thriving pedagogic specialty of law and mental health (Gutheil and Appelbaum, 2000). Indeed, the state's need to determine people's mental states in civil as well as criminal proceedings was an important driver of professionalization in mental health fields.

Entanglements with the law, however, also revealed disagreements among psychiatric experts and undermined their disciplinary authority in other ways. A notable example occurred in the 2012 trial in Norway of the notorious mass killer Anders Breivik who had fatally bombed and shot 77 people the year before. A team of two court-appointed psychiatrists initially found him to be a paranoid schizophrenic, but a second two-person team diagnosed him as sane and suffering from serious personality disorders. The latter diagnosis meant that Breivik would serve time in prison rather than in a low security psychiatric hospital. The majority of Norwegians, and Breivik himself, agreed with the finding of sanity, but the case also showed how pliant forensic psychiatric expertise can be, especially when confronted with strong public opinion.

An early but seminal challenge to psychiatrists' professional autonomy occurred in the 1974 US case of *Tarasoff v. Regents of the University of California* (redecided in 1976). Here, the California Supreme Court ruled that psychiatrists are required to warn (later changed to have a duty to protect) third parties threatened by their patients. Although the decision did not open the floodgates to liability as therapists had feared, *Tarasoff* was widely seen as influencing therapeutic practice by making practitioners more attentive to threats posed by violent patients. In another notable decision, the U.S. Supreme Court in *Barefoot v. Estelle* (1983), a death penalty case, essentially ignored an *amicus* brief by the American Psychiatric Association (APA) stating that psychiatric predictions of dangerousness are wrong in as many as two out of three cases. The Court expressed confidence that juries would be able to evaluate such evidence even if it was weak or deficient, thereby denying privileged status to the APA's professional expertise.

By the end of the twentieth century, the alliance between law and the human sciences – joined, eventually, by genomics and powerful imaging techniques from the neurosciences (Dumit, 2004) – helped to define a host of novel disorders and syndromes invoked to extenuate or penalize various types of conduct. Posttraumatic stress disorder (PTSD), for example, along with its specific manifestations such as rape trauma syndrome, entered the DSM-IV, the official diagnostic handbook for psychologists, in spite of scholarly doubts that PTSD is a scientifically discoverable, universal phenomenon (Young, 1995). Codified in DSM-IV, PTSD became not only an identifiable and presumably treatable medical condition, but also a basis for asserting new legal claims (as in lawsuits for stress-related psychological injury in the workplace), or defending oneself against charges (as in murder trials of sexually abused defendants).

A notorious synergy between law and psychology occurred in a rash of so-called recovered memory and child abuse cases, which gained prominence in the United States in the 1980s (Hacking, 1995). Psychiatrists and psychologists testified as experts, initially to the reliability of recovered childhood memories, but these cases caused scandal as witnesses later recanted and it became apparent that some experts had more likely instilled than elicited the shocking 'memories.' The profession eventually organized to deny the scientific validity of the claims and the expertise of the claimants (Loftus and Ketcham, 1994).

Research on eyewitness identification, another branch of science dealing with memory and recall, also received its chief impetus from the legal process and remained in constant dialogue with the law (Wells and Quinlivan, 2009). Troubled by high error rates on the part of eyewitnesses, psychologists tested people's ability to recognize persons they had encountered in stressful situations and discovered that many factors, such as race, inhibit accurate identification under these circumstances (Wells, 1988). Hundreds of studies with similar findings began to have an impact on police procedure by the late 1990s, as awareness dawned that relatively simple changes, such as sequential rather than simultaneous presentation of suspects in line-ups, could substantially improve the reliability of eyewitness testimony. Nonetheless, in 2012 the U.S. Supreme Court ruled that, in the absence of police misconduct, the introduction of eyewitness testimony without prior judicial scrutiny was not a violation of due process.

Scientific Evidence and Expert Witnesses

The civil law system relegated the production of scientific and technical evidence largely to court-appointed experts, thereby short-circuiting public conflicts between experts. Common law proceedings by contrast traditionally left it to the parties themselves to produce evidentiary support for their claims, with disputes resolved through questioning in pretrial and courtroom contexts. The adversary process, with its ritual of cross-examination, was assumed to be equal to the task of testing the evidence, no matter how arcane or technical, thereby permitting the fact-finder – the judge or the jury – to discern the truth. Although commentators often deplored the practice of treating experts as hired guns, the parties' right to present their own evidence was not in doubt, and courts rarely used their acknowledged power to appoint independent experts who might offer a more disinterested view (Jasanoff, 1995).

By the turn of the century, several social transformations pressured judges into taking a more proactive stance in relation to expert witnesses. The sheer volume of cases demanding some degree of technical analysis was one key factor, especially in the US, where low entry barriers to litigation and the inadequacy of social safety nets routinely brought to the courts controversies that were resolved by other means in most common law jurisdictions. Highly public instances of expertise gone awry, as in the recovered memory cases, shook people's confidence in the power of cross-examination to expose charlatans posing as scientists. The economic consequences of successful products liability lawsuits strained the system from

another direction, especially in a growing number of mass tort actions that threatened entire industries with bankruptcy. It was tempting to blame these developments on the epistemic shortcomings of the adversary system: the passivity and low scientific training of judges, the technical illiteracy of juries, the gamesmanship of lawyers, and the bias or incompetence of experts selected by parties. These perceived deficits led to calls for reforming the process for admitting expert testimony into the courtroom.

Early critique reflected a turn away from the liberal ideals of the 1970s. In an era of governmental retrenchment and economic conservatism, the attack on courtroom science proved to be politically powerful, though it was neither methodologically rigorous nor of lasting scholarly value. One influential broadside accused courts of deciding cases on the basis of 'junk science,' a rhetorically useful label that took hold with opponents of the burgeoning tort system though it proved difficult to characterize systematically (Huber, 1991). The barely concealed purpose of this and related writings was to remove the testing of scientific evidence as much as possible from the adversary system, particularly from jury deliberations, and to decide admissibility issues either in pretrial proceedings conducted by judges or with the aid of scientists appointed directly by the courts. This literature displayed an abiding faith in science's self-corrective ability and a technocratic conviction that truth would eventually win out if only the legal system left scientific fact-finding to scientists (Foster and Huber, 1997). These attacks also turned the clock back in a sense, toward the older presumption that, on any contested issue, mainstream expert opinion exists independently of the legal process and can be validated with unambiguous tests of 'good' and 'bad' science. These positions were eventually thrown into doubt, but not until their polemical force had left a substantial mark on American scientific and legal thought.

Studies by practicing scientists, physicians and some legal academics converged on the theme of the law's misuse of science. Critics cited tort cases in which multimillion dollar jury verdicts favorable to plaintiffs conflicted with the opinions of respected scientists who denied any causal connection between the alleged harmful agent, such as a drug or workplace toxin, and the harm suffered. Particularly noteworthy was the litigation involving Bendectin, a medication prescribed to pregnant women for morning sickness and later suspected of causing birth defects in their children. Juries frequently awarded damages despite repeated assertions by epidemiologists that there was no statistically significant evidence that Bendectin was responsible for the claimed injuries. Sociological research showed suggestive differences in behavior (such as discrepant rates of repeat witnessing) between experts for plaintiffs and for defendants (Sanders, 1998), leading some to question whether offers of scientific testimony were being adequately screened. Another case that drew considerable critical commentary was litigation involving silicone gel breast implants. Close to a half-million women surgically implanted with these devices sued the manufacturer, Dow Corning, claiming injuries ranging from minor discomfort to permanent immune system damage. The company's initial settlement offer broke down under a torrent of legal action, but epidemiological studies, undertaken only after the commencement of lawsuits, showed no causal connection between the implants and immune system disorders. Publication of these results in the prominent *New England Journal of Medicine* led its executive editor to join the chorus of accusation against the legal system's apparent disregard for sound science (Angell, 1996).

Rising discontent about the quality of scientific evidence led the US Supreme Court to address this issue for the first time in 1993. In *Daubert vs Merrell Dow Pharmaceuticals*, the Court set aside the so-called *Frye* rule that had governed the admissibility of expert testimony in federal courts for the preceding 70 years. Instead of simply demanding, as *Frye* had, that scientific evidence be 'generally accepted' within the relevant peer community, the Court urged judges to screen science in accordance with standards that would be used by scientists themselves. As examples, the Court offered four criteria: (1) was the evidence based on a testable theory or technique, and had it been tested; (2) had it been peer reviewed; (3) did it have a known error rate; and, recapitulating the older admissibility rule, (4) was the underlying science generally accepted? Two more major evidence decisions of the 1990s cemented the high court's message that trial judges should play a vastly more proactive and independent gatekeeping role when confronted with scientific and technical evidence. As judicial involvement in admissibility decisions increased, new complaints appeared on the horizon: that judges were inappropriately using the *Daubert* criteria as an inflexible checklist rather than as the guidelines they were meant to be; that mechanical application of *Daubert* and its progeny was trumping deserving claims; that judges were usurping the jury's constitutional role; and that misinterpretation of *Daubert* was introducing unscientific biases and impermissibly raising the burden of proof in civil cases.

In most civil law jurisdictions, by contrast, the inquisitorial approach to testing evidence, coupled with the state's near monopoly in generating forensic science, precluded similar controversy about the legitimacy of experts or the quality of their testimony. As US courts moved to consolidate their gatekeeping function in the 1990s, constricting the entry routes for experts, a countermove of sorts could be discerned in a few European legal systems to open the courtroom door to a wider diversity of expert opinion (Van Kampen, 1998).

Law as a Site of Knowledge-Making

Investigation of the law–science relationship took a more sophisticated turn in the 1990s as it attracted the attention of a new generation of researchers in science and technology studies. From this disciplinary vantage point, the use-misuse framing that had dominated popular writing appeared superficial, especially when set against the observable richness of social interactions between the two institutions. Rather, as this body of work sought to demonstrate, it was the hybridization of law and science that held greatest interest and significance for both science and society. At this live frontier of social problem-solving, one could observe in fine-grained detail how important normative and epistemological commitments either reinforce or contradict each other in contemporary societies. For example, even within the common law's adversarial environment, the ritualistic character of legal proceedings successfully holds the line

against certain forms of radical skepticism, leaving untouched traditional hierarchies of expert authority (Wynne, 1982), belief in the ultimate accessibility of the truth (Smith and Wynne, 1989), and an abiding faith in technological progress (Jasanoff, 1995).

At the same time, the legal process came to be recognized as a distinctive site for the production of socially significant yet otherwise unavailable knowledge and expertise. Even in cases where litigation arguably produced anomalous results – such as those involving Bendectin and breast implants – resort to the courts added to society's aggregate store of information. In these and countless other cases, plaintiffs' claims spurred inquiry into social phenomena that no one had previously thought to study; litigation-driven scientific knowledge was produced under conditions of methodological instability and potential bias, but was relevant nonetheless for righting old wrongs and protecting against future risks. Corporate resources too played an increasing role in the production, and the nonproduction (Proctor and Schiebinger, 2008), of knowledge relevant to policy and social justice. In numerous cases, corporate funding sought to destabilize relatively secure, peer reviewed science supportive of regulation by commissioning studies to raise doubts about the prevalent consensus (Michaels, 2008). All these efforts can be seen as part of the broad sociological shift in knowledge production away from ivory tower science toward ends-oriented research in the later twentieth century (Gibbons et al., 1994).

Other cases of law-driven knowledge production seemed unambiguously progressive. A noteworthy example was the importation of molecular biological techniques into the domains of criminal identification and paternity testing in the form of 'DNA typing.' Legal needs for better identifiers provided the initial context for the technique's move from the laboratory into forensic settings (Aronson, 2007). Subsequent contestation over the reliability of DNA-based identification among ethnic sub-populations had impacts on the field of population genetics and spurred the formation of new testing methods and agencies, as well as the standardization of commonly used tests (Lynch et al., 2008; Lynch and Jasanoff, 1998). As the technique became black-boxed, new market niches for it emerged within the legal system, most spectacularly in the formation of the US-based Innocence Project, a legal collective dedicated to exonerating mistakenly convicted persons. The success of DNA as a forensic tool also helped to destabilize the legal system's longstanding commitment to eyewitness evidence, thereby providing for research critical of the ancient privileging of visual memory (Wells, 1988). But the massive uptake of forensic DNA analysis by the criminal justice system also opened up new areas of social concern, such as privacy and database protection (Hindmarsh and Prainsack, 2010), not to mention outright misconduct, which called for continued vigilance and ingenuity on the part of the legal system (Jasanoff, 2011).

The injection of molecular biology created unexpected ripple effects for other branches of forensic science. Here finally was a domain of law enforcement that seemed firmly grounded in knowledge of nature as opposed to the say-so of potentially biased human experts whose job was to facilitate prosecutions. Fresh skepticism challenged even that most black-boxed tool of the criminal justice system, fingerprinting, revealing weaknesses in its validation and inconsistencies in its use (Cole, 2001). Just as regulatory science had emerged as a field of contestation in the last quarter of the twentieth century, so forensic science took center stage as a new arena for controversy and demand for reforms in the 21st. In 2009, the influential U.S. National Academy of Sciences called for the formation of a new institute to oversee training, accreditation, and research in forensic science, though implementation in a time of financial austerity looked unrealistic.

The rise of DNA typing, like that of risk analysis and PTSD, reflected the dynamics of an epoch in which the establishment of any form of social order became virtually unthinkable without the mutual engagement, or coproduction, of law and science. In such a time, commentary deploring the law's alleged misuses of science appeared increasingly simplistic and out of touch with the everyday reality of deep interpenetration between the two institutions. The emergence of science and law as a special topic within the social sciences, especially in science and technology studies, responded to the inadequacies of earlier analyses premised on reductionist assumptions about science, law, and their interactions. A major contribution of this work was to discredit once for all the pared down 'clashing cultures' model of the law–science relationship and to put in its place a more nuanced picture of the interplay between epistemic and normative authority. By exploring in detail how the law not only 'uses' and 'misuses' science, but also generates it, questions it, assesses science-based technologies, and often compensates for their inadequacies, the new scholarship on science and law invited reflection on the intricate balance of truth and justice in technologically advanced societies.

See also: Biotechnology; Civil Law; Common Law; Critical Race Theory; Expert Testimony; Forensic Genetic Databases: Ethical and Social Dimensions; Forensic Psychiatry; Forensic Psychology; Law and Neuroscience; Law and Society; Law and Technology; Law, Bioethics, and Biotechnology; Norms in Science; Science and Technology Studies: Experts and Expertise; Science and The State; Scientific Evidence: Legal Aspects.

Bibliography

Angell, M., 1996. Science on Trial: The Clash of Medical Evidence and the Law in the Breast Implant Case. Norton, New York.

Aronson, J.D., 2007. Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling. Rutgers University Press, New Brunswick, NJ.

Beck, U., 1992[1984]. Risk Society: Towards a New Modernity. Sage Publications, London.

Bullard, R.D., 1990. Dumping in Dixie: Race, Class, and Environmental Quality. Westview, Boulder, CO.

Carson, J., 2007. The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750–1940. Princeton University Press, Princeton, NJ.

Cole, S., 2001. Suspect Identities: A History of Fingerprinting and Criminal Identification. Harvard University Press, Cambridge, MA.

Dumit, J., 2004. Picturing Personhood: Brain Scans and Biomedical Identity. MIT Press, Cambridge, MA.

Ezrahi, Y., 1990. The Descent of Icarus: Science and the Transformation of Contemporary Democracy. Harvard University Press, Cambridge, MA.

Foster, K.R., Huber, P.W., 1997. Judging Science: Scientific Knowledge and the Federal Courts. MIT Press, Cambridge, MA.

Foucault, M., 1990[1976]. The History of Sexuality. Vintage, New York.

Foucault, M., 1979. Discipline and Punish. Vintage, New York.

Fuller, L., 1964. The Morality of Law. Yale University Press, New Haven, CT.
Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge. Sage, London.
Goldberg, S., 1994. Culture Clash. New York University Press, New York.
Gutheil, T.G., Appelbaum, P.S., 2000. Clinical Handbook of Psychiatry and the Law, third ed. Lippincott Williams & Wilkins, Philadelphia.
Hacking, I., 1995. Rewriting the Soul. Princeton University Press, Princeton, NJ.
Hart, H.L.A., 1961. The Concept of Law. Oxford University Press, Oxford.
Hindmarsh, R., Prainsack, B. (Eds.), 2010. Genetic Suspects: Global Governance of Forensic DNA Profiling and Databasing. Cambridge University Press, Cambridge, UK.
Huber, P., 1991. Galileo's Revenge: Junk Science in the Courtroom. Basic Books, New York.
Irwin, A., Wynne, B. (Eds.), 1996. Misunderstanding Science. Cambridge University Press, Cambridge, UK.
Jasanoff, S., 1990. The Fifth Branch: Science Advisers as Policymakers. Harvard University Press, Cambridge, MA.
Jasanoff, S., 1995. Science at the Bar: Law, Science and Technology in America. Harvard University Press, Cambridge, MA.
Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.
Jasanoff, S. (Ed.), 2011. Reframing Rights: Bioconstitutionalism in the Genetic Age. MIT Press, Cambridge, MA.
Kahneman, D., 2011. Thinking Fast and Slow. Farrar, Straus and Giroux, New York.
Kahneman, D., Slovic, P., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.
Lentsch, J., Weingart, P., 2011. The Politics of Scientific Advice: Institutional Design for Quality Assurance. Cambridge University Press, Cambridge.
Loftus, E.F., Ketcham, K., 1994. The Myth of Repressed Memory: False Memories and Allegations of Sexual Abuse. St. Martin's Press, New York.
Lynch, M., Cole, S.A., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.
Lynch, M., Jasanoff, S. (Eds.), 1998. Contested Identities: Science, Law and Forensic Practice. Social Studies of Science 28 (5–6) (Special Issue).
MacKenzie, D., 1990. Inventing Accuracy: An Historical Sociology of Nuclear Missile Guidance. MIT Press, Cambridge, MA.
Michaels, D., 2008. Doubt Is Their Product: How Industry's Assault on Science Threatens Your Health. Oxford University Press, New York.
Minow, M., 1990. Making All the Difference: Inclusion, Exclusion, and American Law. Cornell University Press, Ithaca, NY.
National Research Council (NRC), 1994. Science and Judgment in Risk Assessment. National Academy Press, Washington, DC.
Polanyi, M., 1962. The republic of science. Minerva 1, 54–73.
Price, D.K., 1985. America's Unwritten Constitution: Science, Religion, and Political Responsibility. Harvard University Press, Cambridge, MA.
Proctor, R.N., Schiebinger, L. (Eds.), 2008. Agnotology: The Making and Unmaking of Ignorance. Stanford University Press, Stanford, CA.
Sanders, J., 1998. Bendectin on Trial: A Study of Mass Tort Litigation. University of Michigan Press, Ann Arbor.
Schuck, P., 1993. Multi-culturalism redux: science, law, and politics. Yale Law and Policy Review 11, 1–46.
Shapin, S., 1994. A Social History of Truth. University of Chicago Press, Chicago.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press, Princeton, NJ.
Slovic, P., 2000. The Perception of Risk. Earthscan Publications, London.
Smith, R., Wynne, B. (Eds.), 1989. Expert Evidence: Interpreting Science in the Law. Routledge, London.
Thaler, R., Sunstein, C., 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press, New Haven, CT.
Van Kampen, P.T.C., 1998. Expert Evidence Compared: Rules and Practices in the Dutch and American Criminal Justice System. Intersentia, Antwerp.
Wagner, P., Wittrock, B., Whitley, R. (Eds.), 1991. Discourses on Society: The Shaping of the Social Science Disciplines. Kluwer, Dordrecht.
Wells, G.L., 1988. Eyewitness Identification: A System Handbook. Carswell, Toronto.
Wells, G.L., Quinlivan, D.S., 2009. Suggestive eyewitness identification procedures and the Supreme Court's reliability test in light of eyewitness science: 30 years later. Law and Human Behavior 33, 1–24.
Winner, L., 1986. The Whale and the Reactor: A Search for Limits in an Age of High Technology. University of Chicago Press, Chicago.
Wynne, B., 1982. Rationality and Ritual: The Windscale Inquiry and Nuclear Decisions in Britain. British Society for the History of Science, Chalfont St. Giles, UK.
Young, A., 1995. The Harmony of Illusions: Inventing Post-Traumatic Stress Disorder. Princeton University Press, Princeton, NJ.

Science and Media

Bruce V Lewenstein, Cornell University, Ithaca, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13654–13657, © 2001, Elsevier Ltd.

Abstract

Although science and media have long been perceived as separate realms, recent scholarship has highlighted their mutual dependence. Media play a fundamental role in the creation of scientific knowledge, and at the same time provide a forum in which public debate about science shapes the social matrix in which scientific institutions operate. Scholars have identified a complex web of interaction among different scientific activities and different media with different audiences. Current research in the field explores political and economic interests of science media producers, rhetorical conventions of scientific presentations, and audience responses to science information.

The phrase 'science and the media' as a single concept encompasses two major social institutions. Science is both a system of reliable knowledge about the natural world and a complex social system for developing and maintaining that knowledge. Similarly, the media comprise a variegated system for collecting and presenting information as well as being an institution with significant economic, political, and social impact. Though long perceived as separate realms, recent scholarship has highlighted their mutual dependence. Increasing episodes of tension and conflict in the late twentieth century led scholars, scientists, media producers, and social critics to explore the interactions of science and the media. Working scientists and media producers sought guidance on how to use each others' resources more effectively, while analysts asked how such issues as political and economic interests, rhetorical conventions, and audience responses shed light on the roles of science and media in shaping each other. One effect of recent scholarship has been to elide the difference between 'media' (all forms of media) and 'the media' (particularly the major institutions of newspapers, magazines, television, and radio), an elision that will be evident throughout this article.

Science and the Media: A Brief History

The history of science and the media is the history of a growing mesh: increasing use of media by science, increasing attention to scientific ideas by media institutions, and increasing tensions caused by the rising interaction.

Through the middle of the twentieth century, science and media were largely seen as separate realms. Science had emerged in the nineteenth century from natural philosophy as a systematic approach to understanding the natural world. With the advent of research-based industries such as chemical dyes and electronic communication, science began to attract the attention of capitalists and government leaders who understood the value of controlling scientific knowledge. By the end of World War II, the emergence of 'big science' had made science a powerful institutional force as well as a system of accredited knowledge with increasingly far-reaching applications. The media had also evolved in the nineteenth century from small, local organizations into nationally and internationally circulated publications that served as tools of merchant capitalism and political control, full of advertising, business and political news, and the ideology of the industrial revolution. By the early twentieth century, new electronic media such as movies and radio provided the means for newly developed forms of 'publicity' and propaganda to help produce modern mass society, as well as to increase public access to new forms of cultural entertainment.

Since the late seventeenth century, natural philosophers had relied on print media, particularly professional journals such as the *Philosophical Transactions* and books such as Isaac Newton's *Principia*, as tools for disseminating the results of their investigations. These publications were not only records of experiments or philosophical investigations, but also served as active carriers of the rhetorical structures that enabled natural philosophers to convince people outside their immediate vicinity of the truth of their claims (Shapin and Schaffer, 1985).

Early in the nineteenth century, new forms of science media emerged, particularly a genre known variously as 'popularization,' *vulgarisation* (French), or *divulgacion* (Spanish). These books and magazines provided entrée into the knowledge of the natural world for non-scientist readers. They ranged from textbooks for young women through to new mass-circulation magazines filled with instruction on the latest achievements of a rapidly industrializing society. The so-called 'great men' of late nineteenth-century England (most notably Thomas Huxley and John Tyndall) became evangelists for science, lecturing widely to enlist public support for the enterprise of rational understanding of the natural world; they converted their lectures into articles and books to spread their reach even further. At the same time, scientists' use of media for disseminating knowledge among the scientific community grew dramatically, with the growth of new journals and the founding of publishing houses committed to scientific information.

The growth of both science and media in the twentieth century forced specialization into the fields, leading to a distinction between what some scientists have called 'communication in science' and 'communication about science.' For scientists in their daily work, broad synthetic journals such as *Science* and *Nature* (published respectively in

the USA and the UK) were joined by topical journals in chemistry, physics, biology, and their many sub-disciplines. For non-scientists, magazines and other media (such as radio and new 'wire services' for newspapers) increasingly focused on particular audiences, such as teachers, inventors, 'high culture' readers, general media users, and so on. Although scientific ideas appeared in some entertainment media, science in most non-technical media became a component of news coverage for different audiences. In the 1920s and 1930s, a small group of professional journalists began to write almost exclusively about science. (These developments were clearest in the highly developed countries of Europe and North America. The patterns in the rest of the world have been little studied and so cannot be described.)

By the second half of the twentieth century, the interaction of science and media had become a complex web, in which new scientific research appeared in abstracts, journals, press reports, and online, while other forms of science – not necessarily derivative of the research reports – appeared regularly in media outlets such as newspapers, televisions, websites, radio programs, puppet shows, traveling circuses, and museums. Moreover, as science became central to modern culture, the presence of science in media such as poetry, sculpture, fine art, entertainment films, and so on, became more evident. In the journalistic media, many reports circulated internationally, through cooperative agreements between public broadcasting systems and worldwide news services. The rise of the World Wide Web in the 1990s provided further opportunities for information about science to circulate more fully between the developed and developing world.

However, from about the mid-century onward, no matter what media were involved, many working scientists perceived presentations in non-technical media to be incorrect, oversimplified, or sensationalized. Thus, tensions developed between the institutions of science and those of the media. Those tensions led eventually to the development of an analytic field of science and media, populated largely by media sociologists, communication content scholars, and sociologists of science.

Understanding Science and Media

Analysis of science and media has led to a deeper understanding of the essential role of media in creating scientific knowledge, the ways that media presentations shape understandings of nature, and the role of public debate in constituting scientific issues.

The earliest attempts to understand the relationship of science and media came from two directions: the overwhelming growth of scientific information, leading to attempts by scientists and scientific institutions to learn how to manage scientific publications, and emerging tensions between scientists and journalists, as specialized reporting developed in the 1930s. Research and publications before and after World War II led by the 1970s to two common understandings.

First, sociologists of science had come to understand the fundamental role of communication in the production of reliable knowledge about the natural world. Scientists do not work in isolation, but must constantly present their ideas to colleagues for acknowledgement, testing, modification, and approbation. The process of communication, occurring in various media, is the 'essence of science' (Garvey, 1979), leading to science becoming 'public knowledge' (Ziman, 1968). At the same time, the apparent differences between the goals of science (methodical, tentative, but nonetheless relatively certain statements about the natural world) and the needs of public media (rapid, attention-getting narratives about issues directly related to readers and viewers, regardless of certainty) had also been explicated (Krieghbaum, 1967), leading to practical attempts to 'bridge the gap' between science and the media but also to recurring fears that the gap was unbridgeable.

Beginning in the 1970s, new developments in the sociology of scientific knowledge opened up new ways of conceiving of the relationship between science and media (Barnes, 1974; Latour and Woolgar, 1979). In particular, attention to the rhetorical goals of scientists in their use of media suggested that a distinction between communication in science and communication about science could not be maintained. Instead, some researchers began to discuss the 'expository' nature of science, showing how scientists tailored their communication to meet the needs of specific media and specific communication contexts (Bazerman, 1988; Shinn and Whitley, 1985). These researchers highlighted the interaction of the idealized intellectual goals of science (for the production of reliable knowledge about the natural world) with the social and institutional goals of scientists and their employers or patrons (for priority, status in the community, ownership of ideas and thus patents, and so on).

In this conception of science and media, scientific knowledge itself might differ in different presentations – in some cases, appearing as a 'narrative of science' (focusing on the methodological and theoretical structures leading to particular knowledge of the natural world), in other cases as a 'narrative of nature' (focusing on the relationships between organisms or entities, with emphasis on the 'story' linking different aspects of those organisms and entities) (Myers, 1990). The very boundary between 'science' and 'non-science' (or 'mere' popularization), between communication in science and communication about science, was shown to be highly mutable, itself an object of rhetorical construction used for political purposes by participants in scientific controversies to establish control over elements of a debate (Hilgartner, 1990).

A second new line of inquiry highlighted the institutional interdependence of science and the media. Pointing to the growing post-World War II need for scientific institutions to generate public and political support, and to the media's need to draw on scientific developments for constant infusions of drama and 'newness,' the new research identified 'selling science' as a major element of the political economy linking science and the media (Nelkin, 1987).

Finally, work on the images of science that appeared over time helped show the ways that social concerns interacted with scientific developments to create cultural representations of science (Weart, 1988; LaFollette, 1990; Nelkin and Lindee, 1995).

At the same time that these developments were occurring in the understanding of science communication, new approaches were developing in understanding the relationship between science and the public. Spurred by concerns in the scientific community about lack of 'scientific literacy,' the studies ultimately questioned the idea that the public has a 'deficit' of knowledge that needs to be 'improved' (Irwin and Wynne, 1996). Instead, the new approach focused on the possibility of engaging the public in discussions about scientific issues, recognizing the contingent nature of understanding of scientific issues and thus the alternative meanings that might emerge in democratic discussions instead of authoritarian pronouncements (Sclove, 1995). Much of the new work looked at issues of uncertainty and risk, highlighting the social negotiations required to make personal and policy decisions in a context of incomplete information (Friedman et al., 1999). The new work clearly implied that science and the media cannot be perceived as separate institutions with separate goals. Instead, better understanding comes from analyzing science and media across audiences, looking at the political and economic interests of media producers, at the rhetorical meanings of media conventions, and at the audience responses to media content (especially in terms of meaning-making).

Approaches to Studying Science and Media

Political and Economic Interests of Media Producers

The most detailed and well-developed understandings of science and the media point to the political and economic interests that motivate the producers of media representations of science. In many countries, the media depend on government subsidies or at least on political tolerance. Moreover, media are expensive to produce. While the presence of the media in democratic societies is often seen as fundamental to political liberty, the economic need to appeal to broad audiences (even for government-controlled media) often leads to editorial decisions that emphasize broad coverage and sensationalism over depth and sobriety. Science media are no different, and even the most prestigious scientific journals such as *Science* and *Nature* have been criticized for highlighting the most 'newsworthy' research, for hyping scientific reports beyond their value, for engaging in 'selling science.' Newspaper and television coverage of science is routinely criticized for its sensationalism, particularly in the coverage of controversial issues such as food contamination (e.g., the BSE or 'mad cow' epidemic in Britain in the 1980s and 1990s), genetic modification of food, and global climate change. Journalists (and their employers) routinely defend their practices on the basis of 'what readers/viewers want,' which is often judged by sales success. Entertainment media are criticized even more heavily for creating movie plots that involve impossible science or focus on highly unlikely scenarios; producers disclaim any responsibility for accuracy, claiming only to be providing (profitable) escapism. These recurring disputes between the scientific community and media producers highlight the irreconcilable tension between the goals of science and the goals of the media.

Rhetorical Meanings of Media Conventions

An emerging area of understanding is the way that media conventions shape the rhetorical meaning of science communication. Scientific papers reporting original experimental research, for example, have developed a standardized format (called 'IMRAD,' for introduction, methods, results, analysis, discussion). The IMRAD format, conventionally written in an impersonal tone, with passive grammatical constructions, hides the presence of the researcher in the research, highlighting the objective character of scientific knowledge that is claimed to exist independently of its observer. In journalistic reports of scientific controversies, the conventions of objective reporting require 'balance' – the presentation of both sides of a controversy (Bazerman, 1988). Though 99.9% of all researchers may hold that a new claim (such as the eventually discredited 1989 announcement of a new form of 'cold nuclear fusion') is scientifically untenable, the journalistic norm of balance may lead to approximately equal attention to all claims. Similarly, journalistic goals of story telling often lead to emphasis on the human, narrative dimensions of an issue over the theoretical or mathematical components of the science. New research is exploring the meaning of visual representations of science, at every level from the technical outputs of genome analyzers to mass media manipulations of false-color images produced by orbiting astronomical telescopes (Lynch and Woolgar, 1990).

Audience Responses to Media Content

One of the most difficult areas to investigate is the audience response to particular media presentations. Virtually no work, for example, has been done on how scientists respond to specialized science media, despite anecdotal stories suggesting that changes in information presentation (such as World Wide Web access to scientific databases) may have dramatic impacts on how scientists conceive of the natural world and frame their hypotheses and theories. Some work has been done at broader public levels, such as responses to media presentations of risk information. But, although the psychology of risk perceptions, for example, has been extensively investigated, the factors affecting individual responses to media coverage of specific risky incidents are so varied as to defy measurement. Though the complexity leads some researchers to reject the possibility of quantitative analysis, others insist that social scientists can learn to isolate appropriate variables. Certainly some areas of audience response could be investigated more carefully; audience reaction to images of science in entertainment films, for example, has never been systematically tested, despite frequent statements that 'the image of scientists in the movies is bad.'

The shift away from separate analysis of 'science' and 'media' to an integrated understanding of 'science and media' has led in recent years to the possibility of cross-cutting analyses, such as those of producers, rhetorical structures, and audience reception. These analyses will ultimately yield a much richer understanding of the interaction of science and media. Less clear is how such improved understanding will contribute to the practical concerns of those in the scientific community and elsewhere, who worry about possible linkages among science literacy, the image of science, and public support for science.

See also: Educational Media: Potentials for Learning; Genetics and the Media; Media Ethics; Media Events; Media and Child Development; Media and History; Public Broadcasting; Research Publication: Ethical Aspects.

Bibliography

Barnes, B., 1974. Scientific Knowledge and Sociological Theory. Routledge and Kegan Paul, London.

Bazerman, C., 1988. Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. University of Wisconsin Press, Madison, WI.

Friedman, S.M., Dunwoody, S., Rogers, C.L. (Eds.), 1999. Communicating Uncertainty: Media Coverage of New and Controversial Science. Erlbaum Associates, Mahway, NJ.

Garvey, W.D., 1979. Communication: The Essence of Science-Facilitating Information Exchange Among Librarians, Scientists, Engineers and Students. Pergamon Press, New York.

Hilgartner, S., 1990. The dominant view of popularization: conceptual problems, political uses. Social Studies of Science 20 (3), 519–539.

Irwin, A., Wynne, B. (Eds.), 1996. Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge, UK.

Krieghbaum, H., 1967. Science and the Mass Media. New York University Press, New York.

LaFollette, M., 1990. Making Science Our Own: Public Images of Science, 1910–1955. University of Chicago Press, Chicago.

Latour, B., Woolgar, S., 1979. Laboratory Life. Sage, Beverly Hills, CA.

Lynch, M., Woolgar, S., 1990. Representation in Scientific Practice. MIT Press, Cambridge, MA.

Myers, G., 1990. Writing Biology: Texts in the Social Construction of Scientific Knowledge. University of Wisconsin Press, Madison, WI.

Nelkin, D., 1987. Selling Science: How the Press Covers Science and Technology. Freeman, New York.

Nelkin, D., Lindee, M.S., 1995. The DNA Mystique: The Gene as a Cultural Icon. Freeman, New York.

Sclove, R., 1995. Democracy and Technology. Guilford, New York.

Shapin, S., Schaffer, S., 1985. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press, Princeton, NJ.

Shinn, T., Whitley, R. (Eds.), 1985, Expository Science: Forms and Functions of Popularisation D Reidel, vol. 9. Dordrecht, The Netherlands.

Weart, S., 1988. Nuclear Fear: A History of Images. Harvard University Press, Cambridge, MA.

Ziman, J.M., 1968. Public Knowledge: An Essay Concerning the Social Dimension of Science. Cambridge University Press, Cambridge, UK.

Science and Politics: Value Neutrality

Sven Eliaeson, Uppsala centre for Russian and Eurasian at Uppsala University, Uppsala; and Centre for Social Studies, Instiute for Philosophy and Sociology at the Polish Academy of Science, Warsaw

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Science and politics is a theme where past and present meet and is hard to ignore, in particular in social science, where the research frontier is split and there are intervening vested interests; uncontrolled value intrusion is a recurrent phenomenon. Max Weber's views on value neutrality and the relation between value neutral knowledge production and political passion are popularized in his 'twin-lectures'. Myrdal, under the influence of Hägerström's value-nihilism provides a pragmatic way out of the value-incommensurability predicament, by the use of significant and relevant explicit value premises, as points of departure for social engineering. Merton codifies the ethos of unbiased science. The autonomy of research is under siege from increasing commodification, bureaucratization, and ideologization. Even if recurrently challenged, Weber's decisionism remains the standard position.

Economic notions of various origins are challenging the traditional, academic value-orientation. The ideas of unconditional truth-seeking and an institutional method of securing scientific quality are being contested. The institutionalized norm system for the pursuit of scholarship nurtures countercultures.

Science and Politics in Social Science

The topic of science and politics confronts us with a chicken-and-egg problem. Machiavellian and Hobbesian statecraft are cases in point. Civil society discourse in Scottish Enlightenment, and Samuel Pufendorf (1632–94) as an 'early bird', could be included as well. Nineteenth-century scholars such as Herbert Spencer (1820–1903) and Karl Marx (1818–83) did not make a difference between scholarship and politics. Throughout the nineteenth century, history writing served state interest and nation building, and not only in Germany. In the German case 'system Althoff' (Friedrich Althoff (1839–1908) was *Universitetsreferent* in the Prussian department of culture and managed to launch more than 140 new institutes, seminars, and clinics.) was very successful a century ago and its aim was not only to promote science but also to strengthen Prussia as a great power. Germany was at the cutting edge and Germans received many early Nobel prizes, not only in medicine and physics; Theodor Mommsen got the literature prize in 1902.

However, it was in the *Werturteilsstreit* within the *Verein für Sozialpolitik* (Social Policy Association) that the issue became imminent and Max Weber (1864–1920) left the room slamming the door in anger (see Nau, 1996).

It is reasonable to focus on the second wave of post-Enlightenment social science and take Weber and his 'twin-lectures' as points of departure. *Politik als Beruf* (The Profession and Vocation of Politics) is not a scholarly text but a popular lecture delivered in front of students at a tumultuous time in Munich's history. *Wissenschaft als Beruf* (Science as a Calling) was delivered earlier, in November 1917, when an end to World War I was still in the future.

Bavaria for a while became a Soviet republic *(Räterepublik)* with escalating violence.

With momentous events expected outside the window, the students wanted political guidance rather than the dry, taxonomical typologization which Weber had to offer.

The Weberian paradox extracted out of these lectures means that we, in our capacity as scholars, have to adhere to value-rationality and the ethics of conviction *(Gesinnungsethik)* in order to promote instrumental means–end rationality *(Zweckrationalität)* and the ethics of responsibility/consequences, which we expect from political decision makers. The scholar buys his freedom at the price of his influence in political decision-making. Alternatives could be, for instance, a technocratic order, with rule by experts.

A politician who is totally value-rational and unwilling to compromise is like a motorist driving against the traffic on the autobahn. Value-rational *Café literati* are anathema for Weber.

The Christian does what is right and places the outcome in God's hands, Weber wrote (Weber, 1994 [1919], p. 359), while the democratically elected politician is accountable and has to answer for the consequences. This division of labor between science and politics is sometimes labeled 'decisionism', a term also associated with Carl Schmitt (about Schmitt's significance in the junction of scholarship and politics, see Turner, 1999).

The background to Weber's methodological essays is a couple of intertwined controversies, *Methodenstreit* between positivism and antipositivism being the central one, but also a debate on the antipositivist side about the best way to protect the human sciences from being invaded by unified science ideals.

Despite actually siding with Wilhelm Dilthey (1833–1911) on the issue of demarcation between the cultural and natural sciences – the object and not the method being the divide – Weber makes use of the neo-Kantian philosopher Heinrich Rickert (1863–1936) in order to accomplish objectivity and intersubjectivity. There are problems of interpretation here. That Weber refers to and depends on Rickert – 'the father of historical relativism' – is explicit. The difference between Rickert and Weber remains a matter of dispute.

Rickert is a case of incoherent 'backdoor-normativism', since he claims the validity of some supra-historical, eternal generally, and unconditionally valid 'true' values. Weber does not agree, but his combined stance of distancing himself from Rickert while also defending Rickert, is not crystal clear either.

Historicism promotes historical relativism and contributes to the erosion and demise of natural law doctrines. Weber's

neighbor and friend, the theologian Ernst Troeltsch (1865–1923) discusses this in terms of polytheism in post-Enlightenment, as the problem of historicism.

Weber's 'Objectivity' essay from 1904 has a strategic place in intellectual history and was explicitly written in the aftermath of the huge *Methodenstreit* between Gustav von Schmoller (1838–1917) and Carl Menger (1840–1921), representing history versus theory. Schmoller was the dominant figure in the younger historical school and also in the founding generation of *Verein*. Historicism was hegemonic and almost all German scholars, including Weber, were 'historicists'. Carl Menger from Vienna brought British-style 'Marginalist' political economy into the German cultural sphere:

> That this problem (of objectivity, my remark, SE) is genuine and not just the fruit of empty speculation on our part, will be apparent to anyone who observes the conflict about methods, 'basic concepts', and presuppositions; the constant shift in 'viewpoints'; and the continual redefinition of the 'concepts' employed, and recognizes that the theoretical and the historical approaches are still separated from each other by an apparently unbridgeable gulf: "*two* sciences of economics!" as a despairing examinee in Vienna once sorrowfully complained. What does 'objectivity' mean in this context? *That* is the only question that will be addressed in the following discussion.
> (Weber, 2012 [1904], p. 108).

Most of the 'Objectivity' essay is dedicated to responding to the problem of the despairing student in Vienna. 'The Meaning of Ethical Neutrality', an essay from 1917 is, however, even more illustrative to the crisis of historicism as the backdrop of Weber's methodological endeavor.

Values are assigned the pivotal role in a selection process from the vast reality in which concept formation is transparent and models testable (*nachvollziehbar*), precisely what the younger historical school (von Schmoller) had difficulties to accomplish.

In the *Methodenstreit*, the political dimension was not exactly hidden. The confrontation was between the exact science and real science, abstractions versus *Wirklichkeitswissenschaft* – "as we would like to pursue it", Weber adds, in order to both mark his belonging to, and ambition to renew historicism. Persuasive and derogatory definitions were in use. British 'Robinsonaden' (rational actor model, individual rationality) were anemic and lacking in *Gestalt*. Germany was still in a process of delayed nation building in which historicism had a defining role – identity provided by the historic past, origin – and the Germans had a big brother complex against the British.

In the controversy over method, the members of the younger historical school had lowered their guard by debating how to establish laws. This made them very vulnerable since they had methodological feet of clay, where the Austrian marginalists had clear models allowing for deductions; in contrast to the mysterious nontransparent 'virgin birth' of *Realfaktoren* contributing to events; the historicist inclination was to collect data and give *Gestalt*, rather than search for *Gesetz;* they were idiographic rather than nomothetic, to use Wilhelm Windelband's terminology.

Historicism has a dual role. It is part of a German, Romantic, anti-Enlightenment trend. Yet, despite its idealist and romanticist background, as part of *deutscher Sonderweg* (A problematic concept, since it seems to assume a nomothetic notion, of a normal route; possibly *Sonderbedingungen* would have been better terminology.) and late nation-building in the search for roots, historicism also contributed to the advancement of scientific techniques and brought about both a new technical awareness, a 'feeling' for the actual material that historical research works with, and a more developed source-criticism. However, an important trademark of historicism was its animosity toward abstract concepts. Historicists felt that reality was either revealed through empirical research or grasped by some sort of empathetic or intuitive act of understanding: either way one could embrace the essence of the phenomenon under study.

Opaque practices in concept formation give leeway to uncontrolled value intrusion. This is notorious in the pre-neo-Kantian historicist tradition, as in the case of Hegel and his independent conceptual reality; while neo-Kantian nominalism enables concept-formation in a lucid way (Lepsius, 1977). Concepts are artifacts, but for historicists such as Hegel, the conceptual world is independent of us and the real one, although we might catch a glimpse of it only occasionally, e.g., Napoleon as an incarnation of the *Weltgeist*.

Weber had "two souls dwelling in his bosom" and politics was his "secret love". He was pioneering policy studies, trying to define the national interest in studies of agrarian economic trends East of the Elbe, in about 800 pages during the 1890s (sponsored by *Verein)*, with a condensed version in Weber's inaugural speech *(Der Nationalstaat und die Volkswirtschaftspolitik*, 'Freiburger Antrittsrede', 1895); (see Agevall, 2004 and Aldenhoff, 1991). Nation – more specifically the German nation as carrier of culture – was an ultimate top value to Weber.

On the crucial topic of neo-Kantian nominalism as a remedy against Hegelian essentialism and conceptual realism, see, for instance, Ralf Dahrendorf (1987, p. 574) with his strong statement that Weber's methodology is far from one-dimensional, nevertheless "above all, non-Hegelian". The same point is made by Bjarne Jacobsen (1998). It is also documented in Weber's famous statement that there are two ways to proceed: "Hegel's way or *our* way". (Documented in Bruun (2007) [1972]: 39n, letter to von Eulenburg on 11 May 1909).

What objective science can accomplish is the rationalization of value-hierarchies, at the service of means-end-rational policy decisions. The problem of norm-sender and incommensurability of ultimate 'top-values' is not solved by Weber. Weber's 'Wahlrecht und Demokratie in Deutschland' (Suffrage and Democracy in Germany) from 1917 is a telling text, possibly the best one to search for a key to Weber's own value hierarchy. He is a functional democrat or a *Vernunftsdemokrat*, just like Alexis de Tocqueville (1805–59). The soldiers returning from the trenches of World War 1 deserve a stake in the running of the government, after the collapse of the Prussian Empire. Weber is free from any illusions about the "true will of the people."

The predicament of anxiety of choice is still with us and even accentuated in post-Modernity. Evaluative judgments cannot be verified or refuted by the means afforded by any science, as Weber writes, soon after endorsing absolute polytheism as the only appropriate metaphysics and characterizing the alternatives between values as an irreconcilable

death-struggle, like that between 'God' and the 'Devil'. Weber in a suggestive manner describes the existentialist dilemma of the post-Enlightenment 'polytheist' predicament of value incommensurability. The last paragraph in 'Zwischen Zwei Gesetzen' (Between Two Laws) is very telling, where with reference to J.S. Mill, Weber characterizes the struggle between multiple and contradictory sets of values from which the scholar "has to choose which of these gods he will and should serve, or when he should serve the one and when the other" (Weber, 1994 [1916], p. 78f).

Robert K. Merton's (1910–2003) pioneering efforts (1949 [1938]) in the sociology of knowledge is a Weberian extension. Merton sees an affinity between Puritan virtues, liberal democracy and experimental research, and innovations.

Merton's CUDOS (an acronym for communalism, universalism, disinterestedness, originality and skepticism) first formulated in 1942, defines the ethos of uncompromising scholarship and is still ever-fresh, although for natural reasons questioned in the recent era of exaggerated commodification. Merton's norms apply to basic research, while applied research has a different logic. Policy research and public choice are naturally problem-orientated, just like medicine and engineering. It is no surprise that peace research promotes peace, and social science today deals with matters of environmentalism, climate change, and risk evaluation.

The Nazi government violated the 'Mertonian' ethos of science in a self-punishing way, which resulted in a mass *diaspora* of scholars from, in particular, Frankfurt am Main, with an energizing effect in American social science, where migrants play a significant role.

Edward Shils (1911–95) was engaged in the Cold Culture Wars of the Congress for Cultural Freedom (Berghahn, 2001) and monitored external threats to the autonomy of scholarship in his journal 'Minerva: A Review of Science, Learning and Policy', which he founded in the early 1960s and himself edited until his death. The 'End of Ideology' theme, engaging also S.M. Lipset (1922–2006) and Daniel Bell (1919–2011), is a good example of duality, academic freedom, and politics intertwined, 'separate but equal'.

Both Merton and Shils are – and see themselves – as extensions of Max Weber. We could here of course add Talcott Parsons (1902–79) with his focus on value-orientation combined with political commitments, for instance, concerning democratic political education in post-World War II Germany (Gerhardt, 2011).

Hans Kelsen (1881–1973) carries straws to the same stack, in his contribution to the American Political Science Review in the early 1950s (Kelsen, 1951).

In general, the narrative of scholars as politicians is a short one and not exactly a success story, except, arguably, for Thomas Masaryk (1850–1937). Historians (e.g., Adolphe Thiers (1797–1877, President of France 1871–73) were more prominent as statesmen in the nineteenth-century era of nation-building. We, perhaps, are wise to apply historical relativism, since in Europe, East of the Elbe, after the implosion of the Soviet Empire, intellectuals and politicians are closer and with some personal overlaps (e.g., Bronislaw Geremek in Poland) as in Swedish social engineering in the 1930s. While sociology is a shrinking discipline in the developed world, it has a growing market in developing countries.

The ambiguity or interplay between politics and science is characteristic for both Weber and Gunnar Myrdal (1898–1987), who offers an 'operationalization' of Weber/Rickert. Myrdal's research in the mid-1930s is dedicated to educational social engineering; and in Myrdal & Myrdal, Kris i Befolkningsfrågan (*Crisis in the Population Question*, 1934) he deals with demographic aspects of a sustainable welfare state. Housing, education, unemployment, nutrition, health care, etc., are dealt with from a national perspective. The book from 1934 is not yet translated into English, but is partly available in Alva Myrdal's *Nation and Family* (1941).

Gunnar Myrdal is both a scientist and a politician, pursuing double careers. The problem of science and values – to bridge David Hume's gulf between 'is' and 'ought' – found a pragmatic solution in Myrdal, with his argument that the chosen explicit value premises providing ultimate top-values in rationalizing value hierarchies for instrumental policy formation, should be 'significant' and 'relevant'. One way to become efficient and 'make a difference' is to connect with contemporary social movements, such as the labor movement. Weber was more like 'dry swimming' compared to Myrdal's praxis orientation, promoting intersubjectivity and testability.

Myrdal was obsessed with how to avoid value bias throughout his life, a recurrent theme in his production from the late 1920s to the early 1970s.

The main source of inspiration was the Swedish 'nihilist' philosopher Axel Hägerström (1868–1939). In his inaugural lecture ('On the Truth of Moral Propositions', in Hägerström 1964) as professor in practical philosophy at Uppsala University in 1911, he claimed that there is no science in morals, only about morals. This lecture is the first coherent formulation of a negative-value ontology, although it had high affinity with similar strands in Cambridge (G.E. Moore) and Vienna (or Graz: Alexius Meinong).

Myrdal, who also was a do-gooder and compassionate world improver and active social engineer and public intellectual, illustrates that there is no necessary contradiction in combining this with a negative-value ontology and non-cognitivist value sentence theory. Value judgments might be scientifically meaningless but there is no harm in promoting them, as long as it is not in the form of 'hidden agendas'.

The difference with Weber is that there is an element of infinite regress in Weber's neo-Kantian approach: we cannot know the borders of rational scientific knowledge formations in the future; we can only not today by scientific means judge between values. To Hägerström and his pupils, values are cognitively meaningless, just as coughs or sneezes. The idea that law is morally binding appears to Hägerström as superstition and magic.

His most dedicated pupils are within the school of Scandinavian Legal Realism (Karl Olivecrona, Vilhelm Lundstedt, and Alf Ross), a school to which also Myrdal belonged. The orthodox adherents would not write about rights but 'rights' with quotation marks, or instead adding 'the realities this term implies'. Rights and duties are artifacts, an instrumental human invention for the pursuit of happiness in a *Rechtsstaat*.

'Scientific value relativism' is a common label for both Weber and Myrdal. Arnold Brecht (1959) is the best source for its provenance and characteristics. Brecht's *Political Theory: The Foundations of Twentieth-Century Political Thought* (1959) is one

of those books without a 'best before' date. Brecht is an excellent example of what he called the 'gulf-doctrine', codifying the scientific method and its limits, yet motivated by consolidating peace.

The predicament in post-Enlightenment polytheism with its new anxiety of choice is common to Weber/Rickert and Myrdal/Hägerström. In a secularized world we cannot unite upon any true values. We might, at best, establish some working 'community of assumption' in order to enjoy a civilized order and perhaps, in the spirit of Bentham and Kant, discuss various tentative or postulated highest values, such as 'the pursuit of happiness'. 'Public welfare' was also popular in post-Jacobine Enlightenment. The dignity of the human individual as a cultural being is an alternative close to both Kant and Weber. But the 'Archimedean point' has to be postulated.

Myrdal emphasizes and applies explicit value premises in his main work *An American Dilemma* (1944), in ambitious appendices.

Myrdal (1944) is a pivotal text for understanding science versus politics as interrelated activities, communicating bowls. It is a 'commodified' investigation financed by the Carnegie Foundation, with the explicit goal of promoting human and minority rights, which is the reason Myrdal is so very ambitious providing detailed exposure of the value premises at work.

Weber was also a public intellectual *avant la lettre*, with many critical contributions to the liberal Frankfurter Zeitung during World War 1. He was a loyal patriotic dissident on the verge of *lese-majesty*.

The 'enemies' of objectivity are numerous, for instance so-called Action research, in which trained nurses could turn themselves to activists in social policy matters and scholars declare solidarity with their research objects. Especially in anthropology, it is a temptation for researchers to identify with their object of research and 'go native'.

Feminist-inspired research deals with gender issues in a way that does not prioritize the criteria of Karl Popper's (1902–94) critical rationalism and the ideal of organized skepticism. We also encounter various forms of dogmatism. Karl Marx was an eminent scholar but it is hard to ignore that his theory about the contradictions in ideal-typical capitalism cannot meet testability criteria and will not easily match Popper's fallibilism, something in common to most macro-theories, due to the option to 'anticipate the consequences'.

Science is only one way to knowledge, aside from revelation. Even Weber in a letter late in life, plays with the idea of being a 'privately' religious person. His problem horizon is defined by Troeltsch's work on 'The Problems of Historicism'. Troeltsch's theology deals with the loss of meaning (the theodicy problem) which is captured by Weber in the metaphor of the 'iron cage' and which calls for values but throws us into polytheism or 'perspectivism'. In a letter to Ferdinand Tönnies (1855–1936), Weber combines a searching attitude in matters of personal conviction and the scientific and anti-metaphysical creed (Weber, 19 February 1909 in MWG II/6: 63 *et passim)*. Weber here makes a very clear demarcation between faith and science, yet indicates that on the personal level he is neither anti-religious, nor a religious, even if he is religiously unmusical (MWG II/6: 65).

Before the objectivity problem becomes urgent with antagonistic world-views in struggle using different 'language games', nascent social science was busy mapping its territory by use of refined survey techniques, which was in the United States combined with a Lutheran reform creed. Being a priest is the most popular second career choice for US sociologists, and many started as social workers. In the United Kingdom, sociologists have a different *habitus*, playing snooker and watching soccer games.

Philanthropy played a significant role in early social science, with Carnegie financing the huge educational project of which Myrdal's *An American Dilemma* was a part, and the Laura Spelman Rockefeller foundation financed the first chair in social policy in Sweden and contributed also to the budget of some Kaiser Wilhelm Institutes (today Max Planck) in Germany. The link to fast urbanization and industrialization in a society of mass production and mass consumption is a backdrop and surveys in demand. 'Fordism' is a concept grasping the interest of big enterprise to be aware of the predicament of the human being. Henry Ford's company had its own sociological department.

Political economy offers applications for political decisions, based on improved cognition. The US anti-Trust legislation could be seen as an application of the Marxian analysis of the logic of the capitalist mode of production. After the great depression, in 1929, anti-cyclical 'Keynesian' policy was successfully implemented by various regimes, in Sweden, Germany, and the United States. Unthinking the political dimension of Milton Friedman (1912–2006) would be unnatural. There is – for better or worse – an invasion by rational choice in the humanities affecting policy making also.

Today the research councils are assigned the task to canalize tax payers' money into planned research, leaving little room for serendipity. The freedom of research is limited to methods and results, but in reality there is no longer the choice of topics. The NPM (new public management) ideal now operates also on the European level. Conformism erodes autonomy; scholarship being transformed into 'scholarshop'.

The universities are hosting 'rebels in search of a cause', some of them 'Trojan horses' not sharing the value assumption of uncompromising search for truth. Extramural interests are on the board of higher institutions of research and learning. In Sweden the Chair on the Board is appointed by the government.

After *die Wende* there was a *Stunde Null* with an open future and many suggestions about future designs. The university system in BRD was reorganized. What the case of post-World War II Germany illustrates is that science can co-exist with various regimes, when its internal norm system is not damaged. This was, however, the case in social science and the humanities, resulting in a liturgic 'new speech', although here also scholarship in the academy remained competitive, for instance in history writing and linguistics.

In present-day Europe, politicians tend to see the universities as a means for social goals and economic growth, rather than intellectual development. The era of 'Bologna' invites moral confusion within the universities and the financial system, with science councils where researchers have to apply invites to self-corrupting tactics, according to critics. Syllabuses are also increasingly streamlined in a way that puts the academic teachers at disadvantage and the evaluation industry flourishes. MOOC (massive open on-line course) further

erodes the old 'Humboldtian' ideals, but increases dissemination of knowledge.

Events in Berkeley, Paris, and more places in the 1960s indicated the new era of the 'megaversity', to use a neologism by the Californian educator Clark Kerr (1911–2003), with ensuing 'polytechnic' orientation of the universities. Kerr's dismissal as President of the University of California by Ronald Reagan, moreover, being an example of state power intervening in the world of research and teaching.

Max Weber's doctrine with its demarcation between scholarship and politics remains the standard position, but is recurrently under siege. His pupil Georg Lukács (1885–1971) is one prominent example, in Die Zerstörung der Vernunft (The Destruction of Reason, 1953).

See also: Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography; Eastern European Studies: Society; Enlightenment: Impact on the Social Sciences; Intellectuals: History of the Concept; Knowledge Society, History of; Objectivity, History of; Political Theory, History of; Science and Technology Studies, History of; Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct; State and Society: History of the Concept; Weberian Social Thought, History of.

Bibliography

Adorno, Th. W., et al., 1972. Der Positivismusstreit in der deutschen Soziologie. Luchterhand, Darmstadt.

Agevall, Ola, 2004. Science, values, and the empirical argument in Max Weber's inaugural address. In: Max Weber Studies, vol. 4(2), July 2004, 157–177. Thematic Issue on Max Weber's Relevance as a Theorist of Politics, Eliaeson, Sven; Palonen, Kari (Eds.).

Aldenhoff, Rita, 1991. Nationalökonomie, Nationalstaat und Werturteile. Wissenschaftskritik in Max Webers Freiburger Antrittsrede im Kontext der Wissenschaftsdebatten in den 1890er Jahren. Archiv für Rechts- und Sozialphilosophie, Beiheft 43, 79–90.

Beck, Ulrich, 1974. Objektivität und Normativität/Die Theore-Praxis-Debatte in der modernen deutschen und amerikanischen Soziologie. Rowohlt, Reibek bei Hamburg.

Berghahn, Volker R., 2001. America and the Intellectual Cold Wars in Europe. Princeton University Press, Princeton.

Blegvad, Mogens, 1991. 'Value' in turn-of-the-century philosophy and sociology. Danish Yearbook of Philosophy 26, 51–96.

Brecht, Arnold, 1959. Political Theory: The Foundations of Twentieth-Century Political Thought. Princeton Univ. Press, Princeton.

Brecht, Arnold, 1970. The Political Education of Arnold Brecht. An Autobiography 1884-1970. Princeton University Press, Princeton.

Bruun, Hans Henrik, 2007 [1972]. Science, Values and Politics in Max Weber's Methodology. Ashgate, Aldershot.

Dahrendorf, Ralf, 1987. Max Weber and Modern Social Science. In: Mommsen, Osterhammel (Eds.), Max Weber and His Contemporaries. Unwin Hyman, London, pp. 574–580.

Eliaeson, Sven, 2002. Max Weber's Methodologies. Interpretation and Critique. Polity Press, Cambridge, UK.

Eliaeson, Sven, 2006. Gunnar Myrdal as a Weberian Public Intellectual. In: Ay, Karl-Ludwig, Borchardt, Knut (Eds.), Das Faszinosum Max Weber. Die Geschichte seiner Geltung. UVK [Universitätsverlag Konstanz], Konstanz.

Eliaeson, Sven, 2012. Value orientation and the secularization of post-enlightenment social science. History of the Human Sciences 25 (3), 3–31.

Eliaeson, Sven, 2014. Myrdal and the diffusion of Scandinavian legal realism. In: Eliaeson, Sven, Mindus, Patricia, Turner, Stephen P. (Eds.), Axel Hägerström and Modern Social Thought. Bardwell Press, Oxford, pp. 363–374.

Eyerman, Ron, 1985. Rationalizing intellectuals. Theory and Society 14, 777–808.

Gerhardt, Uta, 2011. The Social Thought of Talcott Parsons. Methodology and American Ethos. Ashgate, Farnham.

Hägerström, Axel, 1964. Philosophy and Religion (Robert T. Sandin, Trans.). Allen & Unwin, London.

Hägerström, Axel, 1953. Inquires into the nature of law and morals. In: Olivecrona (Ed.), Sthlm: Almqvist & Wiksell (Trans. C.D. Broad).

Iggers, Georg(e), 1973. Historicism. In: Dictionary of the History of Ideas. Studies of selected pivotal ideas, vol. II. Charles Scribner's sons, New York, pp. 456–464.

Iggers, Georg(e), 1983 [1968]. The German Conception of History. The National Tradition of Historical Thought from Herder to the Present. Wesleyan University Press, Middletown, Connecticut.

Iggers, Georg(e), 2005 [1997]. Historiography in the Twentieth Century. From Scientific Objectivity to the Post-Modern Challenge. Wesleyan University Press, Middletown, Connecticut.

Jacobsen, Bjarne, 1998. Max Weber und Friedrich Albert Lange. Rezeption und Innovation. DUV, Wiesbaden.

Kelsen, Hans, 1951. Science and politics. American Political Science Review 45, 641–661.

Lepsius, Rainer M., Januar 1977. Max Weber in München. Rede anlässlich der Enthüllung einer Gedenktafel. Zeitschrift für Soziologie 6 (1).

Lukács, Georg, 1953. Die Zerstörung der Vernuft. Aufbau Verlag, Berlin.

Max Weber Studies, 2004. vol. 4(2), July 2004, 157–177. Thematic Issue on Max Weber's Relevance as a Theorist of Politics, Eliaeson, Sven; Palonen, Kari (Eds.).

Merton, Robert K., 1949 [1938]. Science, Technology and Society in Seventeenth-Century England, in *Osiris*, 1949, orig. published in 1938, also New York 1970.

Mindus, Patricia, 2009. A Real Mind. The Life and Work of Axel Hägerström. Springer, Dordrecht; Heidelberg.

Mouffe, Chantal (Ed.), 1999. The Challenge of Carl Schmitt. Verso, London & New York.

Myrdal, Alva, 1941. Nation and Family: The Swedish Experiment in Democratic Family and Population Policy. Harper, New York.

Myrdal, Gunnar, 1958. Value in Social Theory. Harper & Brothers, New York.

Myrdal, Gunnar, 1990 [1953, 1929]. The Political Element in the Development of Economic Theory. Transaction, New Brunswick. Introduction by Richard Swedberg.

Myrdal, Gunnar, 1944. An American Dilemma. The Negro Problem and Modern Democracy, with the Assistance of Richard Sterner and Arnold Rose. Harper & Brothers, New York & London.

Myrdal, Gunnar, 1933. Das Zweck-Mittel-Denken in der Nationalökonomie. Zeitschrift für Nationalökonomie 305–329.

Myrdal, Gunnar, 1970. Objectivity in Social Research. Duckworth, London. Orig. in Swedish 1968.

Nau, H.H. (Ed.), 1996. Der Werturteilsstreit. Die Äuβerungen zur Werturteilsdiskussion im Ausschuβ des Vereins für Sozialpolitik.

Olivecrona, Karl, 1971 [1939]. Law as Fact. Stevens & sons, London.

Rickert, Heinrich, 1913 [1902]. Die Grenzen der naturwissenschaftlichen Begriffsbildung. Mohr, Tübingen.

Ryder, Sharon, Hasselberg, Ylva, Waluszewski, Alexandra (Eds.), 2013. Transformations in Research, Higher Education and the Academic Market: The Breakdown of Scientific Thought. Springer, Dordrecht; London.

Shils, Edward, 2006. A fragment of a sociological autobiography. In: Grosby, Steven (Ed.), The History of My Pursuit of a Few Ideas. Transaction, New Brunswick.

Shils, Edward, Fall 1970. Tradition, ecology, and institution in the history of sociology. Daedalus 99 (4), 760–825.

Stevenson, C.L., 1963 [1944]. Facts and Values. Yale University Press, New Haven and London.

Troeltsch, Ernst, 1922. Der Historismus und seine Probleme. Mohr, Tübingen.

Turner, Stephen P., 1999. The significance of Shils. Sociological Theory 17, 125–145.

Weber, Max, 1949. Methodology of the Social Sciences (Edward Shils and Henry Finch, Trans.). Macmillan & Free Press, London & New York.

Weber, Max, 2012. In: Bruun, H.H., Whimster, S. (Eds.), Max Weber. Collected Methodological Writings. Routledge, London (Trans. Bruun).

Weber, Max, 1994. In: Lassman, Peter; Speirs, Ronald (Eds.), Political Writings. CUP, Cambridge.

Weber, Max, 2008. In: Dreijmanis, Johan (Ed.), Max Weber's Complete Writings on Academic and Political Vocations. Algora, New York (Trans. Gordon C. Wells).

Weber, Max, 2012 [1904]. The 'Objectivity' of Knowledge in Social Science and Social Policy. MWG = Max Weber Gesamtausgabe, 100–138, org. as Die 'Objektivität' sozialwissenschaftlicher und sozialpolitischer Erkenntnis. Originally published in Archiv für Sozialwissenschaft und Sozialpolitik, 19 Bd, 1904).

Science and Religion

Ryan T Cragun, University of Tampa, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.A. Russell, volume 12, pp. 13621–13625, © 2001, Elsevier Ltd.

Abstract

The relationship between science and religion is contingent upon at least three factors: the period of time in question, the religion of interest, and the views of specific individuals. This article illustrates that an accurate understanding of the relationship between science and religion can only be discerned in the light of knowledge of these three factors. The article also notes the growing interest in the scientific study of religion.

The relationship between religion and science is complicated (Brooke, 1991). It is complicated by the fact that both religion and science have changed over time, and by the fact that neither may be said to constitute singular, uniform entities (*see* Science, Sociology of). While the boundaries of science are somewhat porous and it can be difficult to demarcate what should and should not be included as science (Evans and Evans, 2008), 'religion' is a broad term that refers to thousands of diverse organizations (Cho and Squier, 2008). Additionally, individuals can and do think about both science and religion in multifaceted ways (Baker, 2012; Sherkat, 2011). Thus, the relationship between religion and science is complex because it is contingent upon at least three factors: the time period in question, the religion of interest, and the individual. Each of these will be discussed in turn, as well as the increasing interest in studying religion scientifically.

Time Period

The specific methodology and concomitant knowledge for understanding the natural world referred to nowadays as 'science' are relatively recent human developments. Humans have, for thousands of years, attempted to understand the natural world, but not in the exacting way employed nowadays by scientists. Evidence for interest in understanding the natural world can be seen in famous monuments built in prehistoric times (i.e., *c.*3000 BCE), like Stonehenge and Newgrange; these monuments appear to be oriented relative to astronomical phenomena, reflecting attempts to understand the movement of celestial bodies.

Some of the earliest and most well-known efforts to systematize knowledge about the natural world come from the ancient Greeks (Roger, 1986). While they did use some experimentation not unlike that employed by scientists nowadays (Geymonat, 2010), most of their ideas were derived from speculation, contemplation, and introspection (Roochnik, 2004), which are approaches to understanding the world more commonly referred to nowadays as philosophy, not science. However, it was speculation about the natural world that ultimately became the science of the modern age. Speculation about the natural world was originally called 'natural philosophy' (Brooke, 1991). Natural philosophy evolved over thousands of years – from the ancient Greeks to the Renaissance and Enlightenment thinkers – until it eventually became what we call as science. During that evolution, the way that natural philosophy was conducted changed (Lindberg, 1986). Early natural philosophy would have included both empirical investigation, like observing the paths of celestial bodies, but also introspective contemplation or speculation, like contemplating similarities between humans and other living things. Over time, stricter criteria for observation and speculation were developed, leading to the Renaissance, the project of modernity, and ultimately to the more rigorous empiricism of modern scientific endeavor (*see* History of Science).

This brief history of science and its evolution is noted here in order to illustrate that what people typically call science has not always been the exacting and precise discipline it is today. This becomes even more apparent when the relationship between religion and science is examined.

Religion, from a historical standpoint, predates natural philosophy (Lindberg, 2007; *see* Religion, History of). Yet, the two are not entirely dissimilar, particularly in antiquity (Brooke, 1991). Religion, broadly, is collective beliefs relative to the supernatural – collective because religions are social institutions and supernatural because religions incorporate beliefs and ideas that fall outside of nature. However, the distinction between natural and supernatural is not fixed; it is a dynamic distinction based on the accumulation of knowledge. The use of religion vis-à-vis natural philosophy to explain the natural world is likewise dynamic. When very little was understood about the natural world, as was the case in prehistoric times and antiquity, religion was often invoked to offer such explanations. For instance, the occurrence of a flood or tsunami, while fully understood by modern science, would not have been explainable by natural philosophers thousands of years ago (Dundes, 1988). Without a naturalistic (i.e., nonsupernatural) explanation, a supernatural one might be suggested, like any of the dozens of documented flood myths that are widespread among religions around the world, e.g., the epic of Gilgamesh or the Noachian flood of the Bible/Torah, that invoke the wrath of a deity to explain the sudden, devastating influx of water (Dundes, 1988). Thus, when very little knowledge about the natural world existed, religion was invoked to offer explanations. In the ancient world, things like earthquakes, rain, stars, floods, storms, and even life and death were all explained through religion, because so little was known about the natural

world (Campbell and Moyers, 1991). However, as natural philosophy expanded and eventually evolved into modern science, naturalistic explanations for natural phenomena increased, shrinking the domain of religious explanations (Grant, 1986). While many modern theologians argue against the idea that religion now offers a 'god of the gaps' – i.e., god and religious explanations are inserted into the gaps of scientific knowledge (Dixon, 2008; Evans and Evans, 2008), most religions no longer suggest supernatural explanations for natural phenomena that do not have alternative explanations in scripture (e.g., lightning strikes or volcanic eruptions).

The depiction of the relationship between religion and science in the previous paragraph, while necessary to illustrate how science and religion are different, is overly simplistic. It suggests that religion and science are distinct and incompatible. While they are distinct epistemologies and in many ways incompatible (see below), this has not always been the case. Early natural philosophers would often rely in part on supernatural phenomena as explanations for natural phenomena when naturalistic explanations were not sufficient. Many natural philosophers were motivated by their religious beliefs to pursue an understanding of natural phenomena, often with the aim of supporting their religious beliefs. This was the case for one of the most famous scientists, Isaac Newton. While unorthodox in his religious views, Newton was driven by his belief in the supernatural to try to discern order in the universe (Popkin et al., 1999). Upon finding such order, Newton did not stop believing in the supernatural but rather felt that he had helped to illustrate that there was a divine order in the movement of planets and stars (Westfall, 1986; Jacob, 1986). Many natural philosophers/scientists were motivated by their religious worldview to investigate nature, and many of these individuals retained their religious beliefs – though often in a modified form – after advancing science through discovery (Westfall, 1986; Brooke, 1991).

Another famous scientist, Galileo, is often invoked to argue that the relationship between religion and natural philosophy was hostile, given his imprisonment and forced recanting of the heliocentric model of the solar system he proposed, by the Roman Catholic Church. Yet, even that famous case was complicated by the context in which it occurred – the Protestant Reformation had just begun and the Catholic Church became less tolerant of alternative views as a result (Shea, 1986; *see* Reformation and Confessionalization). The Roman Catholic Church did ban specific books, and the threat of punishment from the Church was an ever-present concern for many natural philosophers during the Late Middle Ages and through the Renaissance (Shea, 1986; *see* Renaissance). Yet, there were also many natural philosophers who were also religious functionaries – e.g., priests and monks – during the late middle ages and early Renaissance (Ashworth, 1986). In short, the historical relationship between religion and science was never perfectly distinct nor was it entirely oppositional.

The complicated relationship between religion and science is also apparent in other religious traditions, like Islam. Islam had a period of significant scientific advancement, often referred to as the Islamic Golden Age (*see* Historical Thought and Historiography: Islamic Tradition). During this time, scholars throughout the Muslim world made significant advances in science and, like Newton, were often inspired by their religion to do so. For instance, Al-Khwarizmi developed algebra in part to help solve Islamic inheritance laws (Gandz, 1938). Likewise, advances in astronomy, geography, and trigonometry were motivated in part by the desire to be able to accurately determine the direction of the Qibla for Salah prayers (Gingerich, 1986). While support for scientific endeavors began to wane in the Muslim world in the 1500s, there was a period of time during which Islam strongly motivated individuals to pursue a better understanding of the natural world and there are many Muslims today who see no conflict between modern science and their religion (see below).

It is only since about the middle of the nineteenth century that the relationship between science and some religions has been formulated as a distinctly antagonistic one (Dupree, 1986; Gregory, 1986; Numbers, 1986). This leads to the second factor that must be considered when examining the relationship between science and religion – specific religious institutions.

Religion of Interest

'Religion' refers to the broad group of organizations that various scholars have suggested should be grouped together. What these organizations have in common is that they, in some way, relate to the supernatural. The supernatural would include gods, angels, demons, spirits, or essences that exist outside of the natural or empirically detectable world. Thus, by definition, religions include elements that lie outside the domain of science, which is concerned with the natural world. As noted above, the relationship between the natural and supernatural is dynamic, as many things that were once thought to be supernatural – e.g., lightning and the origins of species – have now been shown to be natural phenomena.

More importantly, there is a great deal of diversity within the broad category of religion. There are numerous ways to distinguish between different types of religions. One way to categorize religions is based on their relative geographic origins – distinguishing between Eastern and Western religions. Another way would be to group religions into broad families that are historically, organizationally, and theologically similar, e.g., Hinduism, Buddhism, Judaism, Christianity, Islam, and other religions. These classification schemes are useful for understanding historical similarities between religions, but there is a better way of understanding religions when it comes to the relationship between science and religion today.

Originating in the US in the late nineteenth and early twentieth centuries was a split among Protestant religions over the acceptance of science (Antoun, 2008; *see* Christianity: Evangelical, Revivalist, and Pentecostal). Some Protestant religions embraced the findings of the increasingly rigorous scientific community. As a result of embracing these findings, like Darwin's theory of evolution, these religions began to 'liberalize,' adjusting their supernatural beliefs according to the findings of modern science (Chaves, 1997; *see* Liberal Christianity). A literal interpretation of the Bible morphed into a metaphorical interpretation, as stories like Adam and Eve being the first humans on the planet were no longer tenable in

light of the findings of geology, biology, and other scientific disciplines (Yandell, 1986). The Protestant religions that accommodated science came to be viewed as 'liberal' religions. But not all religions were willing to accommodate science. Some rejected what they saw as capitulation to modernity and instead advocated a return to the 'fundamentals' of religion. These religions came to be viewed as 'fundamentalists.' While originally a term used to describe just this group of Protestants, the term is now more widely used to describe any religious group that exhibits specific characteristics, namely scriptural literalism and black and white thinking (Antoun, 2008). Holding a literalistic view of scripture results in conflict with science, as science has shown that various aspects of the Bible (but also other scriptures, like the Quran, the Bhagavad Gita, the Book of Mormon, etc.) are inaccurate when interpreted literally.

There has been, then, a split among religions over acceptance of science. While this split is not as clean as the previous paragraph suggests – some religions lie in between, accepting some aspects of science but rejecting others (Cragun, 2013) – the split illustrates that the relationship between religion and science is religion-specific. Some religious groups, like Reformed Judaism or the United Church of Christ, largely accept the findings of modern science. Other religions, like Southern Baptists and many Pentecostal congregations, reject the findings of modern science that conflict with their literalistic interpretation of scripture. Likewise, within Islam there are individuals who are more receptive to science and individuals who reject tenets of modern science that contradict their interpretation of their religion. Discerning which religions have embraced modern science and which have rejected it is somewhat easier in the Christian West as some of the divisions between religions have been over these very issues (e.g., Lutheran divisions in the US are based on this distinction). However, this type of distinction, between fundamentalist, moderate, and liberal interpretations of religion can be observed in every major religious tradition, including Judaism, Islam, Buddhism, Christianity, and Hinduism. In short, without knowing which specific religion is of interest, and perhaps even which religious congregation (assuming a congregational structure), it is difficult to know exactly how religions view modern science. Some religions embrace it; others reject it; and still others embrace some elements of science but not all of it.

Individuality

Yet, even recognizing the distinctions that exist between religions, it has increasingly been shown that individual members within religions are more or less receptive to science, regardless of the formal teachings or doctrine of the religion (Baker, 2012; Sherkat, 2011). Recent research has shown that denominational distinctions are breaking down and that there is greater diversity within religions than between religions. **Table 1**, using data from a dozen countries around the world, illustrates the diversity that exists within religious traditions. It shows that even within the same religion, attitudes toward science vary considerably.

As the table illustrates, many Muslims, Buddhists, Hindus, Protestants, and Catholics around the world are more and less receptive to science. It is the case nowadays that fundamentalist and liberal Catholics sit side-by-side during Mass. Likewise, there are Muslims in the Middle East who are scientists and accept the major tenets of modern science, yet attend Friday prayers with fundamentalist Muslims who reject evolution. Knowing an individual's religion will increase the odds of accurately predicting their views of science (Sherkat, 2011), but such knowledge will not allow for perfect predictions.

The Scientific Study of Religion

Of final note, despite the complex relationship between science and religion, in the last two centuries, religion itself has increasingly become an object of scientific investigation. Such investigation can focus on the origins of religion (Barrett, 2004; Atran, 2002), the evolution of religions and religious concepts over time (Atran, 2002; Armstrong, 1994; *see* Religion: Evolution and Development), the present state of religions

Table 1 View of science and faith of members of most popular religion in various countries

	We depend too much on science and not enough on faith		
Most popular religion/country	*Disagree (1–4) (%)*	*Neutral (5) (%)*	*Agree (6–10) (%)*
Buddhists in Japan	24	22	54
Catholics in Brazil	56	20	24
Catholics in Poland	22	21	57
Evangelical Church in Germany	35	20	45
Evangelical Lutheran Church of Finland	30	14	56
Hindus in India	30	na	70
Muslims in Indonesia	40	9	51
Muslims in Jordan	53	6	41
Muslims in Turkey	24	11	64
Protestants in Ghana	28	7	65
Protestants in the United States	24	15	61
Protestants in Zambia	37	15	48

Note: Most popular religion does not mean majority religion. In some of these countries, the nonreligious outnumber the religious.
Source: World Value Survey, 2004–2007 wave.

(Chaves, 2011), and the psychology and biology of religious people (Molendijk and Pels, 1998; Ashbrook and Albright, 1997; *see* Psychology of Religion). The scientific study of religion has led to the understanding that religion is a social construction (Berger, 1990), that religion evolves in light of both internal and external pressures (Chaves, 1997; Lawson and Cragun, 2012), that human evolution and psychology contribute in intriguing ways to religious beliefs (Atran, 2002), and has even begun to predict the potential future of religion (Bruce, 2002).

Conclusion

Historically, science and religion were different from what they are today. Science was far more speculative and less rigorous in its approach to understanding the natural world than it is today, and it was also often engaged in by religious people who were pursuing it in the interest of justifying their religious beliefs. While religions did, at times, oppress scientific inquiry and scientists, it is not accurate to state that religion and science were always opposed. Today, the same relationship holds. Some religions – fundamentalist religions – reject elements of science they believe contradict their religious beliefs. But many other religions embrace modern science, modifying their understanding of or jettisoning outright elements of their religion that contradict scientific findings. And within any religion, individuals may be more or less receptive to science than the religion's own doctrine. Science has also begun to turn its unflinching gaze toward religion, revealing that even belief in the supernatural is likely a natural phenomenon.

See also: Evangelical, Revivalist, and Pentecostal Religions; Historical Thought and Historiography: Islamic Tradition; Liberal Christianity; Reformation and Confessionalization; Religion, History of; Religion, Psychology of; Religion: Evolution and Development; Renaissance; Science, History of; Science, Sociology of.

Bibliography

Antoun, R.T., 2008. Understanding Fundamentalism: Christian, Islamic, and Jewish Movements, second ed. Rowman & Littlefield Publishers.

Armstrong, K., 1994. A History of God: The 4000-year Quest of Judaism, Christianity and Islam. Ballantine Books, New York.

Ashbrook, J.B., Albright, C.R., 1997. The Humanizing Brain: Where Religion and Neuroscience Meet, first ed. Pilgrim Press.

Ashworth Jr., W.B., 1986. Catholicism and early modern science. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 136–166.

Atran, S., 2002. In Gods We Trust: The Evolutionary Landscape of Religion. Oxford University Press, USA.

Baker, J.O., 2012. Public perceptions of incompatibility between "science and religion". Public Understanding of Science 21 (3), 340–353.

Barrett, J.L., 2004. Why Would Anyone Believe in God? AltaMira Press.

Berger, P.L., 1990. The Sacred Canopy: Elements of a Sociological Theory of Religion. Anchor.

Brooke, J.H., 1991. Science and Religion: Some Historical Perspectives. Cambridge University Press.

Bruce, S., 2002. God Is Dead: Secularization in the West. Blackwell Publishers, London.

Campbell, J., Moyers, B., 1991. The Power of Myth. Anchor.

Chaves, M., 1997. Ordaining Women: Culture and Conflict in Religious Organizations. Harvard University Press, Cambridge.

Chaves, M., 2011. American Religion: Contemporary Trends. Princeton University Press.

Cho, F., Squier, R.K., 2008. "He blinded me with science": science chauvinism in the study of religion. Journal of the American Academy of Religion 76 (2), 420–448.

Cragun, R.T., 2013. What You Don't Know about Religion (But Should). Pitchstone Publishing, Charlottesville, Virginia.

Dixon, T., 2008. Science and Religion: A Very Short Introduction. Oxford University Press, USA.

Dundes, A., 1988. The Flood Myth. University of California Press, Berkeley.

Dupree, A.H., 1986. Christianity and the scientific community in the age of Darwin. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 351–368.

Evans, J.H., Evans, M.S., 2008. Religion and science: beyond the epistemological conflict narrative. Annual Review of Sociology 34, 87–105.

Gandz, S., 1938. The algebra of inheritance: a rehabilitation of Al-Khwarizmi. Osiris 5, 319–391.

Geymonat, M., 2010. The Great Archimedes. Baylor University Press, Waco, Tex.

Gingerich, O., 1986. Islamic astronomy. Scientific American 254 (4), 74–83.

Grant, E., 1986. Science and theology in the middle ages. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 49–75.

Jacob, M.C., 1986. Christianity and the Newtonian worldview. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 238–255.

Lawson, R., Cragun, R.T., 2012. Comparing the geographic distributions and growth of Mormons, Adventists, and witnesses. Journal for the Scientific Study of Religion 51 (2), 220–240.

Lindberg, D.C., 1986. Science and the early church. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 19–48.

Lindberg, D.C., 2007. The Beginnings of Western Science: The European Scientific Tradition in Philosophical, Religious, and Institutional Context, Prehistory to A.D. 1450. University of Chicago Press, Chicago.

Molendijk, A.L., Pels, P., 1998. Religion in the Making: The Emergence of the Sciences of Religion. Brill Academic Publishers.

Numbers, R.L., 1986. The creationists. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 391–423.

Popkin, R.H., Force, J.E., William Andrews Clark Memorial Library, 1999. Newton and Religion: Context, Nature, and Influence. Kluwer Academic.

Roger, J., 1986. The mechanistic conception of life. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 277–295.

Roochnik, D., 2004. Retrieving the Ancients: An Introduction to Greek Philosophy. Blackwell Publishing, Malden, MA.

Shea, W.R., 1986. Galileo and the church. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 114–135.

Sherkat, D.E., 2011. Religion and scientific literacy in the United States. Social Science Quarterly 92 (5), 1134–1150.

Westfall, R.S., 1986. The rise of science and the decline of orthodox christianity: a study of Kepler, Descartes, and Newton. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 218–237.

Yandell, K.E., 1986. Protestant theology and natural science in the twentieth century. In: Lindberg, D.C., Numbers, R.L. (Eds.), God and Nature: Historical Essays on the Encounter between Christianity and Science. University of California Press, Berkeley, CA, pp. 448–472.

Science and Social Movements

Andrew Jamison, Aalborg University, Aalborg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The relations between science and social movements are multifaceted. New frameworks for scientific activity have often emerged as part of broader political and social struggles. Social movements have also provided audiences, or new publics, for the spreading and popularization of scientific findings and ideas, as well as critical perspectives on both the form and content of science and of the uses to which science has been put. Finally, it can be suggested that many social movements are themselves producers of science, in that they provide alternative modes of knowledge production that contend with more established or hegemonic modes.

Introduction

The social movements that are discussed in this article refer to major political processes of popular mobilization that have significant societal significance, and include organized forms of protest as well as broader and more diffuse manifestations of public opinion and debate. In relation to science, social movements conceived in this way can be seen to have served as seedbeds for new modes of practicing science and organizing knowledge more generally, and also as sites for critically challenging and reconstituting the established forms of scientific activity.

Since they fall between different academic specializations and disciplinary concerns, the relations between science and social movements have tended to be a neglected subject in the social sciences. For students of social movements in sociology and political science, the relations to science have been generally of marginal interest and received little systematic attention, while in science and technology studies, social movements usually have been regarded more as contextual background than as topics of investigation in their own right. Considering their potential significance as 'missing links,' both in the social production of knowledge and in broader processes of political and social change, however, the relations between science and social movements deserve closer scrutiny. For one thing, new ideas about nature and society have often first emerged outside the world of formal scientific activity, within broader social and political movements. Social movements have also often provided audiences, or new publics, for the spreading and popularization of scientific findings and results. Movements of religious dissent were among the most significant disseminators of the new experimental philosophy in the mid-seventeenth century, for example, while, in the nineteenth and twentieth centuries, new approaches to medicine, social relations, gender roles, and the environment found receptive audiences within social and political movements. Most recently, in relation to such issues as climate change, the different attitudes to science in contending political and social movements have played a key role in the international public debate (Jamison, 2010).

What Is a Social Movement?

There Is little agreement among those who study social movements about what a social movement 'really' is. Definitions depend on which movements are seen as typical or most important, what kind of terminology or conceptual framework is applied, the particular research questions being addressed and, not least, the situation, or standpoint of the researcher (Della Porta and Diani, 2006). Since Neil Smelser (1962) in his classic work, *Theory of Collective Behavior*, divided movements into 'results-oriented' and 'values-oriented' types, there has been a bifurcation among the students of social movements, namely, between those who focus on what might be termed movements in the streets, that is, explicit protest movements, versus those who focus on movements in the mind, that is, cultural and intellectual movements. However, the various definitions that have been most actively applied in the scholarly literature can be said to share certain common features or elements that will be used to provide a 'working definition' of social movements for the purposes of this article.

On the one hand, a social movement is a collective form of social behavior that is explicitly organized for political action. A social movement is the process by which human and material resources are mobilized to try to affect political change (an influential recent discussion is McAdam et al., 2001). They make use of what Herbert Kitschelt (1986), in his analysis of antinuclear movements in the 1970s, termed 'political opportunity structures' in pursuit of one or another political cause. Social movements tend to manifest themselves through publicly recognized forms of protest, or direct action, but those acts do not themselves make a social movement. They need to be linked or connected to one another in some way, organized and coordinated by means of a common platform or program. This provides what has been termed a 'collective identity,' a set of values, precepts, or beliefs that empower those who share them and identify with them. These matters of collective identity formation have been most influentially theorized and studied by Alain Touraine and Alberto Melucci and their students and colleagues (cf Touraine, 1981, 1988; Melucci, 1996).

On the other hand, a social movement is something distinct from more formalized political parties, social institutions, or other established kinds of politics, in that it has a more informal character than established political and social activities; in this regard, social movements are related to what Ulrich Beck (1992) has termed 'subpolitics.' Many

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85022-9

theorists have attempted to distinguish movements from institutions, organizations, and political parties (perhaps most famously Alberoni, 1984), and a recurrent effort in social movement studies has been to analyze processes of institutionalization and professionalization. In recent years, social movements have come to be ever more likened to, and conceptualized as, networks, and attention has been given to distinguishing between the different kinds of networking, brokerage and (inter)mediation that goes on in social movements (e.g., McAdam and Diani, 2003). In many contemporary social movements, not least those related to climate change, much of this networking activity is now conducted 'virtually' via electronic communication and the Internet (Garrett, 2006).

A third element of almost all theories of social movements is the understanding that social movements are more than 'merely' political phenomena, and that they involve some form of what Jürgen Habermas has influentially termed communicative action. As in other fields of social science, there has been an increasing interest in these more communicative, or cultural, aspects of social movements (cf Goodwin and Jasper, 2004). Many social movement analysts use the concept of framing to discuss these matters (Snow and Benford, 1988; Benford and Snow, 2000), while for others the role of passions and emotions has come to be emphasized (cf Goodwin et al., 2001). Other theorists, usually within anthropology, have focused explicit attention on the uses of knowledge by social movements, especially the uses of so-called indigenous knowledge by social movements in non-Western countries (e.g., Aparacio and Blaser, 2008; Escobar, 2008).

The cognitive approach to social movements (Eyerman and Jamison, 1991) makes use of the terms 'cognitive praxis' and 'movement intellectuals' to emphasize the role of knowledge making in social movements, and to characterize the people who are most actively involved. Cognitive praxis is defined as the linking, or integration of ideas, ideologies, and/or world view assumptions (a cosmological dimension) to particular activities or forms of action, including technical development, information dissemination, and practical demonstration of both protest and constructive alternative (a technological dimension). The experimentation with renewable energy that took place in the environmental movements of the 1970s is only the most recent example of the kind of alternative technology development that takes place in social movements (Jamison, 2013). The movement is seen as providing an organizational dimension, a public space, for integrating the cosmology and the technology in processes of collective learning, and it is their cognitive praxis that makes social movements particularly important in the constitution and reconstitution of science and technology (Jamison, 2006). Eyerman and Jamison (1998) broaden the cognitive approach to encompass cultural practices and, by so doing, focus attention on the role that the 'mobilization of tradition' plays in the collective activities of many social movements. It is the mobilization and (re)invention of different traditions of ideas, beliefs, and ideologies that often plays an important role in attracting active participation and involvement in social movements.

Historical Perspectives

A fundamental insight of the sociology of knowledge, as it developed in the 1920s and 1930s, was to suggest that modern science emerged in the seventeenth century out of a more all-encompassing struggle for political freedom and religious reform (Scheler, 1980 [1924]; Merton, 1970 [1938]). What eventually came to be characterized as modern science represented an institutionalized form of knowledge production that, at the outset, had been inspired by more sweeping social and political transformations. The historical project of modernity did not begin as a new scientific method, or a new mechanical worldview, or, for that matter, as a new kind of state support for experimental philosophy in the form of scientific academies. As for the Reformation, it arose as a more deep-seated challenge to traditional ways of thought in social and religious life. It was a protest against the Church (which is why the members of the movement were called protestants), and it was an encompassing social and cultural movement that articulated and practiced alternative, or oppositional, forms of religion, politics, and learning as part of its political struggle (Mandrou, 1978). The teachings of Paracelsus, Giordano Bruno, and Tomaso Campanella, to name some of the better known figures, combined a questioning of established religious and political authority with an interest in scientific observation, mathematics, mechanics, and technical improvement (Merchant, 1980). But, as the broader movements were replaced by more formalized institutions in the course of the seventeenth century, the political and social experiments came to be transformed into scientific experiments; the political and religious reformation, tinged with mysticism and filled with mistrust of authority, was redefined and reconstituted, at least in part, as a scientific revolution (see **Table 1**).

In the late seventeenth and early eighteenth centuries, the scientific 'aristocracy' that had emerged in London and Paris at the Royal Society and the *Academie des Sciences* was challenged by dissenting groups and representatives of the emerging middle classes, some of whom fled from Europe to the colonies in North America, and some of whom established scientific societies, often in provincial areas in opposition to the established science of the capital cities.

Table 1 A schematic representation of the scientific revolution from movement to institution

From movement	*To institution*
Social reform	Reform of philosophy
Millenarian vision	Experimental program
Connection to radical politics	Rejection of politics
Decentralized structure	Central academy
Democratic/open to all	Elitist/professional
Spiritual (absolute) knowledge	Instrumental (probabilistic) knowledge
Technical-economic improvement	Scientific development
Informal communication	Formal communication
Pamphlets, correspondence	Journals, papers

Many of them shared with the academicians and their royal patrons a belief in what Max Weber termed the protestant ethic, that is, an interest in the value of hard work and the virtue of making money, and most had an interest in what Francis Bacon had termed 'useful knowledge.' The social movements of the enlightenment objected, however, to the limited ways in which the Royal Society and the Parisian Academy had organized the scientific spirit and institutionalized the new methods and theories of the experimental philosophy. The various attempts to democratize scientific education in the wake of the French Revolution and to apply the mechanical philosophy to social processes – that is, to view society itself as a topic for scientific research and analysis – indicate how critique and opposition helped bring about new forms of scientific practice. Adam Smith's new science of political economy developed in the Scottish hinterland, and many of the first industrial applications of experimentation and mechanical philosophy took place in the provinces, rather than in the capital cities, where the scientific academies were located (Russell, 1983).

Many of the political and cultural trends of the nineteenth century – from romanticism and utopianism to socialism and populism – also began as critical movements that at a later stage, and in different empirical manifestations, came to participate in reconstituting the scientific enterprise (cf Jamison et al., 2011). Romanticism, for example, first emerged as a challenge to the promethean ambitions of science, in the guise, for example, of Mary Shelley's Doctor Frankenstein, while socialism, in the utopian form promulgated by Robert Owen, was a reaction, among other things, to the problematic ways in which the technological applications of scientific activity were being spread into modern societies. Romantic writers and artists, like William Blake, Johann Goethe, and later Henry David Thoreau, were critical of the reductionism and cultural blindness of science and of the 'dark satanic mills' in which the mechanical worldview was being applied. They sought to mobilize the senses and the resources of myth and history in order to envision and create other ways of knowing (Roszak, 1973). Some of their protest was transformed into constructive artistic creation, and later in the century, the romantic revolt of the senses inspired both the first waves of environmentalism in the form of conservation societies and the emergence of a new kind of holistic scientific discipline, to which was given the name ecology: household knowledge.

In the late nineteenth century, the labor movement also sought a deep-going, fundamental political, and cultural transformation of society, but it would see its critical message translated, in the twentieth century, into packages of reforms and a more welfare-oriented capitalism, on the one hand, and into the 'scientific socialism' of Lenin and Stalin, on the other (Gouldner, 1980). Once again, however, science benefited from this institutionalization of the challenge of social movements; the knowledge interests of the labor movement, for example, entered into the new social science disciplines of economics and sociology. A political challenge was once again transformed into programs of scientific research and state policy, but while new forms of scientific expertise were developed, little remained of the broader democratization of knowledge production that the labor movement, in its more radical days, had articulated.

In the early twentieth century, the challenge to established science was mobilized most actively in the colonies of European imperialism, as well as in the defeated imperialist powers of Germany and Italy; the critique was primarily of scientific civilization writ large, and of what Mohandas Gandhi in India called its 'propagation of immorality.' In the name of modernism, science stood for the future and legitimated, in the colonies as well as in Europe, a wholesale destruction of the past and of the 'traditional' knowledges – the other ways of knowing – that had been developed in other civilizations (Tambiah, 1990). These social movements had an impact on the development of political ideology on both the right and left, but also inspired new sciences of ethnography and anthropology, as well as the sociology of knowledge. Even more important perhaps were the various attempts to combine the artisanal knowledges of the past and of other peoples with the modern science and technology of the present in new forms of architecture, design, and industrial production. Many of the regional development programs of the 1930s and 1940s, in Europe and North America, can trace their roots back to the cultural critique of modernism that was inspired by the Indian independence movement and by such 'movement intellectuals' as William Morris and Patrick Geddes in Britain. Both Roosevelt's New Deal and the Swedish model welfare state can be said to have mobilized critical perspectives in their projects of social reconstruction.

Science and the 'New Social Movements'

In recent decades, social movements have also served to challenge and reorient the scientific enterprise. Of the so-called new social movements of the 1960s and 1970s (anti-imperialist, environmentalist, feminist, and identity movements) have emerged a range of alternative ideas about science, in form, content, and meaning, that have given rise to new scientific theories, academic fields, and technological programs (Schiebinger, 1993; Harding, 1998). Out of critique have grown the seeds of new, and often more participatory, ways of sciencing – from technology and environmental impact assessment to women's studies, queer theory, and postcolonial discourses. What were protest movements of radical opposition in the 1960s and 1970s largely have been emptied of their political content, but they have given rise to new branches of, and approaches to, science and technology.

The new social movements rose to prominence in the downturn following a period of institutional expansion and economic growth. They emerged in opposition to the dominant social order and to its hegemonic scientific technological regime, which had been largely established during and immediately after World War II (Touraine, 1988). The war led to a fundamental transformation in the world of science and technology, and to the emergence of a new relation, or contract, between science and politics. The war effort had been based on an unprecedented mobilization of scientists to create new weapons, from radar to the atomic bomb, and

to gather and conduct intelligence operations. In the process, 'little' science was transformed into big science, or industrialized science.

Especially important for the new social movements that were to develop in the 1960s and 1970s was the fact that scientific research was placed at the center of postwar economic development. Many of the economically significant new products – nylon and other synthetic textiles, plastics, home chemicals and appliances, and television – were based directly on scientific research, and the new techniques of production were also of a different type: it was the era of chemical fertilizers and insecticides, of artificial petrochemical-based process industries, and food additives. The forms of big science also differed from the ways in which science had been organized in the past. The big science laboratories – both in the public and in the private sectors – were like industrial factories, and scientific-technical innovation came to be seen as an important concern for business managers and industrial organizers. The use of science in society had become systematized, and, as the consequences of the new order became more visible, new forms of mistrust and criticism developed (Jamison and Eyerman, 1994).

One wing of the public reacted to the destruction of the natural environment, what an early postwar writer, Fairfield Osborn, termed 'man's war with nature.' The exploitation of natural resources was increasing and, in the 1940s and 1950s, it began to be recognized that the new science-based products were more dangerous for the natural environment than those that had come before. But it would only be with Rachel Carson, and her book *Silent Spring* in 1962, that an environmental movement began to find its voice and its characteristic style of expression. It was by critically evaluating specific instances of scientific technology, particular cases of what Osborn had called the 'flattery of science' that the environmentalist critique would reach a broader public. Carson singled out the chemical insecticides for detailed scrutiny and assessment, but her point was more general. Carson's achievement was not only to direct the methods of science against science itself, but also to point to another way of doing things, the biological or ecological way – what she called in her book the road not taken.

Another source of inspiration for the new social movements came from philosophers and social historians who questioned the more general impact of science on the human spirit. It was one-dimensional thinking that critical theorists like Herbert Marcuse reacted against, the dominance of an instrumental rationality over all other forms of knowing. For Lewis Mumford, another major source of inspiration, it was the homogenization of the landscape that was most serious, the destruction of the organic rhythms and flows of life that had followed in the wake of postwar economic growth, as well as the dominance of what he termed the 'megamachine,' the use of technology for authoritarian purposes.

In the 1970s, a range of new social movements, building on these and other sources of inspiration, came to articulate an oppositional, or alternative approach to science and technology (Dickson, 1974). The so-called new social movements represented an integrated set of knowledge interests, which combined the critique of Rachel Carson with the liberation dialectics of Marcuse and the direct democracy of the student movement. The new movements involved a fundamental critique of modern science's exploitative attitude to nature, as well as an alternative organizational ideal – a democratic, or participatory ideal – for the development of knowledge. There were also distinct forms of collective learning in the new social movements of environmentalism and feminism, as well as grassroots engineering activities that went under the name of appropriate or alternative technology. The social movements provided the setting for innovative forms of cognitive praxis, combining alternative worldviews, or cosmological assumptions with alternative organizational and practical–technical criteria.

In the case of environmentalism, the cosmology was, to a large extent, the translation of a scientific paradigm into a socioeconomic paradigm; in the 1970s, the holistic concepts of systems ecology were transformed into political programs of social ecology, and an ecological worldview was to govern social and political interactions. Technology was to be developed under the general perspective that 'small is beautiful' (in the influential phrase of E.F. Schumacher), and according to the assumption that large-scale, environmentally destructive projects should be opposed and stopped. At the same time, new contexts for education and experimentation and the diffusion of research were created in the form of movement workshops and, in the Netherlands, for example, in the form of science shops, which allowed activist groups to gain access to the scientific expertise at the universities (Irwin, 1995).

Since the 1980s, the cognitive praxis of the new social movements has been decomposed into a disparate cluster of organizations and individuals, through processes of professionalization and fragmentation. The movements were transformed into various kinds of professional expertise, which made it possible to incorporate parts of a particular movement into the established political culture, and to shift at least some of the members of the movement from outsider to insider status. The political struggles for civil rights, environmental protection, women's and sexual liberation, and ethnic and national identity have inspired new approaches to knowledge that have since been institutionalized and transformed into established scientific fields, such as environmental studies, women's studies, gay and lesbian studies, African-American studies, as well as in new areas of medicine and technology.

At the same time, science itself has been reconstituted, partly as a result of the critical perspectives and cognitive challenges posed by the new social movements of the 1960s and 1970s. Many of the postmodern theories in the cultural and human sciences have been inspired by the experiences of social and political movements. Of the alternative public spaces that have been created by social and political movements has emerged a new kind of scientific pluralism, in terms of organization, worldview, and technical application.

Science and Contemporary Social Movements

In recent years, there has developed a growing literature on science, technology, and social movements that has perhaps

been especially focused on local movements in such areas as food, agriculture, energy, and housing (Hess, 2007; Hess et al., 2008). There has also been significant attention devoted to health movements, particularly in relation to so-called patient groups (Epstein, 1996, 2008). There has also been interest directed to the movements associated with so-called climate skepticism, and their connections to neoconservative groups and movements (Austin, 2002; McCright and Dunlap, 2003; Jacques et al., 2008).

The relations between science and the movement for global justice that emerged in the wake of the antiglobalization protests of the late 1990s have also received attention by a variety of scholars. The global justice movement has been called 'a movement of movements' (Klein, 2000) and is composed of a variety of groups and individuals who have begun to take political action in order to protest the quite different kinds of negative consequences that they attribute to globalization.

There are at least three different kinds of submovements, or networks, concerned with global justice, which take part in various international gatherings that are sometimes said to represent the global justice movement (Jamison, 2010). These three strands, which often find themselves in conflict when it comes to particular topics, tend to mobilize different sorts of people with different sorts of grievances, and they tend to tap into different traditions of ideas. On the one hand, there are the parties, organizations, federations, and other institutionalized legacies of the so-called old social movements of the nineteenth and early twentieth centuries, the various outgrowths of populist and labor movements that have become integral parts of the political landscape.

A second important component of the emerging movement is based on the concerns of the new social movements of the 1970s, especially the movements for environmental protection and women's liberation. These movements have tended to become established fixtures in the contemporary world, primarily in the form of nongovernmental organizations, or NGOs. There is usually in these organizations a more businesslike, professional approach to politics, and, as is well known, these organizations, much like university science and engineering, have become dependent on external funding for most of their activity, which usually takes the form of specific projects related to what are sometimes termed 'single issues.'

In recent years, these 'old' and 'new' social movement-based organizations have been complemented by a newer wave, or generation, of activists, and by groups and organizations, which are often more confrontational than the older movements. Beginning in the 1980s, sometimes in the name of 'environmental justice,' these groups have often taken shape in direct opposition to particular manifestations of global injustice. In relation to science, the protests against biotechnology and genetically modified organisms (GMOs) in food and agriculture have been particularly active (Tokar, 2001; Schurman and Munro, 2010; Kinchy, 2012). In recent years, there have been attempts to arrange gatherings and manifestations, where the different component parts of the global justice movement can meet and discuss their concerns. These various 'social forums,' as they have come to be called, have taken place both at an international level (at world social forums, which have been held each year since 2000) and at more regional, national, and local levels, particularly in Europe (Fisher and Ponniah, 2003).

There are a number of tensions, not least generational and geographical, among the various components of the emerging movement, and, as might be expected, there are also some significant differences in regard to science. The remnants or residues of the older social movements, in the parties and organizations of the 'left,' tend to be favorable toward science and technology, while the institutional legacies of the new social movements of the 1970s, usually referred to as NGOs, tend to focus on particular issues or areas of concern. It is only a relatively small group of newer organizations and activists who have emerged to alert the public to the wide range of other challenges that science and technology raise in relation to global justice and economic inequality. For these people, science and technology are, for the most part, seen as connected with corporate expansionism. A representative of this kind of criticism is the Indian physicist-turned-activist Vandana Shiva, who has been one of the leading spokespersons for international protests against biotechnology and genetically modified foods (Shiva, 2000, 2005). Shiva's criticism is not about the science itself, but rather about the cultural contexts in which science is developed:

> The conflict over genetically engineered crops and food is not a conflict between 'culture' and 'science'. It is between two cultures of science: one based on transparency, public accountability and responsibility toward the environment and people and another based on profits and the lack of transparency, accountability and responsibility. (Shiva, 2000: p. 109)

This emphasis on the cultural dimension of science on the part of contemporary social movements is central to two recent studies. In *Fighting for the Future of Food*, Rachel Schurman and William A Munro (2010) characterize the struggle over biotechnology as a conflict between opposing 'lifeworlds,' and provide a detailed investigation of the values and ways of life of the anti-GMO activists, on the one hand, and 'agribusiness' companies and networks, on the other. In *Seeds, Science, and Struggle*, Abby Kinchy (2012) presents an ethnography of local anti-GMO movements in Mexico and Canada, focusing in particular on the attitudes and approaches to scientific knowledge within them. Both in terms of research methods and general approach, these books represent what will most likely be a growing interest in studying the cultural aspects of social movements in their relations to science in the years to come.

See also: Agricultural Sciences and Technology; Civil Rights Movements; Collective Behavior, Sociology of; Environmental Movements; Health Politics and Policy; Movements, Social; Nongovernmental Organizations: Anthropological and Historical Aspects; Populism; Risk, Sociology and Politics of; Science and Technology Studies, History of; Science, Public Engagement With; Social Movements, History of: General; Technology, Social Construction of.

Bibliography

Alberoni, F., 1984. Movement and Institution. Columbia University Press, New York.

Aparacio, J., Blaser, M., 2008. The "lettered city" and the insurrection of subjugated knowledges in Latin America. Anthropological Quarterly 81 (1), 59–94.

Austin, A., 2002. Advancing accumulation and managing its discontents: the U.S. antienvironmental countermovement. Sociological Spectrum 22 (1), 71–106.

Beck, U., 1992. Risk Society. Towards a New Modernity. Sage, London.

Benford, R., Snow, D., 2000. Framing processes and social movements: an overview and an assessment. Annual Review of Sociology 26, 611–639.

Della Porta, D., Diani, M., 2006. Social Movements. An Introduction. Blackwell, Oxford.

Dickson, D., 1974. Alternative Technology and the Politics of Technical Change. Fontana, London.

Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley, CA.

Epstein, S., 2008. Patient groups and health movements. In: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies. MIT Press, Cambridge, MA, pp. 499–539.

Escobar, A., 2008. Territories of Difference. Duke University Press, Durham, NC.

Eyerman, R., Jamison, A., 1991. Social Movements. A Cognitive Approach. Penn State University Press, University Park, PA.

Eyerman, R., Jamison, A., 1998. Music and Social Movements. Cambridge University Press, Cambridge.

Fisher, W., Ponniah, A. (Eds.), 2003. Another World Is Possible: Popular Alternatives to Globalization at the World Social Forum. Zed Books, London.

Garrett, R.K., 2006. Protest in an information society: a review of literature on social movements and new ICTs. Information. Communication and Society 9 (2), 202–224.

Goodwin, J., Jasper, J., Polletta, F. (Eds.), 2001. Passionate Politics: Emotions and Social Movements. University of Chicago Press, Chicago, IL.

Goodwin, J., Jasper, J. (Eds.), 2004. Rethinking Social Movements: Structure, Meaning and Emotion. Rowman and Littlefield, Lanham, MD.

Gouldner, A., 1980. The Two Marxisms. Contradictions and Anomalies in the Development of Theory. Seabury University Press, New York.

Harding, S., 1998. Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies. Indiana University Press, Bloomington, IN.

Hess, D., 2007. Alternative Pathways in Science and Industry. Activism, Innovation and the Environment in an Era of Globalization. MIT Press, Cambridge, MA.

Hess, D., Breyman, S., Campbell, N., Martin, B., 2008. Science, technology and social movements. In: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies. MIT Press, Cambridge, MA.

Irwin, A., 1995. Citizen Science: A Study of People, Expertise and Sustainable Development. Routledge, London.

Jacques, P., Dunlap, R., Freeman, M., 2008. The organisation of denial: conservative think tanks and environmental skepticism. Environmental Politics 17 (3), 349–385.

Jamison, A., Eyerman, R., 1994. Seeds of the Sixties. University of California Press, Berkeley, CA.

Jamison, A., 2006. Social movements and science: cultural appropriations of cognitive praxis. Science As Culture 15 (1), 45–59.

Jamison, A., 2010. Climate change knowledge and social movement theory. Wiley Interdisciplinary Reviews: Climate Change 1 (6), 811–823.

Jamison, A., Christensen, S., Botin, L., 2011. A Hybrid Imagination: Science and Technology in Cultural Perspective. Morgan & Claypool, San Rafael, CA.

Jamison, A., 2013. The Making of Green Engineers. Sustainable Development and the Hybrid Imagination. Morgan & Claypool, San Rafael, CA.

Kinchy, A., 2012. Seeds, Science and Struggle: The Global Politics of Transgenic Crops. MIT Press, Cambridge, MA.

Kitschelt, H.P., 1986. Political opportunity structures and political protest: anti-nuclear movements in four democracies. British Journal of Political Science 16 (1), 57–85.

Klein, N., 2000. No Logo: No Space, No Choice, No Jobs. Flamingo.

Mandrou, R., 1978. From Humanism to Science, 1480–1700. Penguin Books, Harmondsworth, UK.

McAdam, D., Diani, M. (Eds.), 2003. Social Movements and Networks. Oxford University Press, Oxford.

McCright, A., Dunlap, R., 2003. Defeating Kyoto: the conservative movement's impact on climate change policy. Social Problems 50 (3), 348–373.

McAdam, D., Tarrow, S., Tilly, C., 2001. Dynamics of Contention. Cambridge University Press, Cambridge.

Melucci, A., 1996. Challenging Codes: Collective Action in the Information Age. Cambridge University Press, Cambridge.

Merchant, C., 1980. The Death of Nature: Women, Ecology, and the Scientific Revolution. Harper and Row, San Francisco, CA.

Merton, R.K., 1970 [1938]. Science, Technology, and Society in Seventeenth-century England. Harper and Row, New York.

Roszak, T., 1973. Where the Wasteland Ends: Politics and Transcendence in Post-industrial Society. Anchor, New York.

Russell, C., 1983. Science and Social Change, 1700–1900. Macmillan, London.

Scheler, M., 1980 [1924]. Problems of a Sociology of Knowledge. Routledge & Kegan Paul, London.

Schiebinger, L., 1993. Nature's Body: Gender in the Making of Modern Science. Beacon, Boston, MA.

Schurman, R., Munro, W., 2010. Fighting for the Future of Food: Activists versus Agribusiness in the Struggle over Biotechnology. University of Minnesota Press, Minneapolis, MN.

Shiva, V., 2000. Stolen Harvest: The Hijacking of the Global Food Supply. South End Press, Cambridge, MA.

Shiva, V., 2005. Earth Democracy: Justice, Sustainability and Peace. South End Press, Cambridge, MA.

Smelser, N., 1962. Theory of Collective Behavior. Routledge and Kegan Paul, London.

Snow, D., Benford, R., 1988. Ideology, frame resonance, and participant mobilization. International Social Movement Research 1 (1), 197–217.

Tambiah, S., 1990. Magic, Science, Religion and the Scope of Rationality. Cambridge University Press, Cambridge.

Tokar, B. (Ed.), 2001. Redesigning Life? The Worldwide Challenge to Genetic Engineering. Zed Books, London.

Touraine, A., 1981. The Voice and the Eye: An Analysis of Social Movements. Cambridge University Press, Cambridge.

Touraine, A., 1988. The Return of the Actor: Social Theory in Postindustrial Society. University of Minnesota Press, Minneapolis, MN.

Science and Technology, Anthropology of

Michael MJ Fischer, MIT, Cambridge, MA, USA; Harvard Medical School, Boston, MA, USA; and National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of science, technology, and society (STS) has emerged alongside a number of closely neighboring fields of scholarship. Four features distinguish the anthropology of STS: (1) a detailed interest in the sciences and technologies themselves; (2) a global perspective, not just an account from Western Europe and North America; (3) multilocale or multisited ethnographic access to complex distributed processes such as the global chemical industry or global clinical trials; and (4) a concern with the powerful aesthetics of imaginaries. The task of translating legacy knowledges into public futures draws upon four kinds of genealogies: test drives and libidinal drives, protocols and networks, landscapes or ethical plateaus, and reknitting global moieties split by the cold war.

Mapping the Field with Nearest Neighbors

The anthropology of science, technology, and society (STS) is in dialogue with real-world socio-political and cultural epistemic contests over ecological change, environmental damage, revenue seeking, sovereignty, and sustainability versus political–economic transformation, as well as with the creation of new science cities and globally distributed, locally modulated communities of scientists, engineers, and media developers.

The anthropology of STS emerged in the 1970s and 1980s, with work on physicists in the United States and Japan (Traweek, 1988), epistemologies of engineers, users, and communities of practice (Suchman, 1987; Cefkin, 2009), moral approaches to evolving medical testing (Fox and Swazey, 1974, 1992; Rapp, 1999; Martin, 1987, 1994, 2007; Clarke, 1998; Fujimura, 1996; Hogle, 1999), regulatory science and the law (Jasanoff, 1990, 1995, 2005), globally distributed industrial processes with unintended negative consequences (Reich, 1991; Fortun, 2001), and a 'cyborg anthropology' of prosthetic and experimental interfaces between silicon and flesh (Haraway, 1991, 1997, 2003; Downey and Dumit, 1997; Downey et al., 1995; Davis-Floyd and Dumit, 1998).

The anthropology of STS has emerged alongside, and broadened the purview of, a range of neighboring fields populated by sociologists, historians, philosophers, and bioethicists: (1) social studies of knowledge addressed itself to idealist accounts of scientific method, noting that tacit knowledge (Polanyi, 1974) and 'good hands' or skills in experiments are not easily transferrable through written protocols until stabilized in mature, repeatable, robust technologies (Collins, 1974; see also Fleck, 1935, on the troubles of stabilizing the Wasserman test for syphilis). (2) Actor–network theory (ANT) (e.g., Latour, 1988; McNeil, 1976; Mitchell, 2002; Spinage, 2012; Mahvunga, 2013) emphasizes that networks have obligatory points of passage that can reverse power ratios (e.g., from farmers' empirical knowledge to laboratory bacteriological science) and depend upon 'immutable mobiles' (the transfer of abstracted information distanced from locally modulated contexts). ANT promotes the methodological principle of including nonhuman 'actors' or actants within network analyses. (3) Social histories of science and technology (e.g., Galison, 1997; Fortun, 1993, 2008; Biagioli, 1993, 2006) have provided important coproducers of the anthropology of science and technology. (4) The social construction of technology approach (e.g., Bijker, 2002; Bijker et al., 1987) recognizes that technologies are not devoid of politics, that they are shaped by social histories, and that they can create new infrastructures for inequality.

What distinguish the anthropology of STS are four features:

1. A commitment to opening the 'black box' of the sciences and technologies, that is, to exploring their internal sociabilities, competitions, presuppositions, and articulations with managerial and political economy pressures. This involves unpacking enough ethnographically to be able to distinguish between aspirations and pragmatic pipelines of experimentation, and between what is a breakthrough and what is just normal science. It involves understanding the 'peopling' of the sciences and technologies: the importance of interpersonal relations and trust across disciplines, across institutions, and across national competitions on which the building and maintaining of these pipelines and platforms either work or do not. Opening the black box of science and technology is crucial to the anthropology of science and technology. It raises the bar of precision and refocuses what we think we know about how the world works in terms of both civic and technological epistemologies.
2. A global perspective that does not limit itself to Western Europe and the United States, thus continuing anthropology's longstanding charge of being a comparative social anthropology of the world and not just of privileged presumptive leaders or holders of political advantage. Here the rise of journals based in Asia (e.g., *EASTS: East Asian Science Technology and Society* and *Science, Technology and Society*) as well as the National University of Singapore's Biopoleis Project are harbingers of a more polycentric approach to global developments, many of which are happening in Asia.
3. A set of methods that can deal ethnographically (at the required close-up scale) with globally distributed technoscientific, political–economic, and socio-cultural

processes (at the required overview mapping scales). Doing ethnography at 'multilocale' or 'multisited' strategic points of access to multiscaled and distributed processes such as the global chemical industry, the global biotechnology industry, or climate change is an invaluable way of ground truthing, reality testing, and challenging models that use only aggregate data, superficial survey instruments, or results limited to laboratories divorced from shop floors, tinkerers, hackers, makers, and other users.

4. An artistic and literary aesthetics not only for writing the cultures of the sciences and technologies, but also for the imaginaries that inform new experimental discoveries, validating tests, as well as psychic and financial investments in new socio-technical infrastructures, conventions, and trade shows. Such liminal technoscientific spaces have long been terrains, for instance, for NASA exploration of extreme environments, cyborg possibilities, and space medicine. Among such artistic aesthetics that are increasingly pervasive are bioarts for democratizing scientific knowledges and making more visible and enjoyable the various sorts of care that maintaining living organisms require and the limits that such care and such knowledge of habitat interactions impose, while at the same time allowing the limits of possibility to be expanded.

A broader venue can be seen through the topics of recent anthropology of science and technology books and dissertations that include (nonexhaustively) the following: evidence law and the incorporation or exclusion of visual media from handwriting and finger printing to positron emission tomography scans and video animation, and the related topic of e-documents in a culture of cryptography; interpretive debates over the meaning of visualizing, amplifying, or proliferation technologies such as medical scans, tissue culture technologies, and clinical diagnoses; open software as socio-political as well as technical infrastructure, e-governance in Latin America, and e-kiosks in India; rationalized office design, e-tools for collaborative work at a distance, and its failures; Taiwanese biostatisticians as providers of bridging tools for global clinical trials, in a case of a nonrecognized country using expertise to be a player; biological citizenship after disaster, bioethnic conscription in disease studies, and bioavailability as biopolitics; surgical training, pharmaceutical marketing, and objective self-fashioning through drugs, and the turning of selves into clinical subjects; forms of biocapitalism under different national agendas; organ transplantation, its psychoses for patients, identity contradictions for physicians (in Turkey), moral rejection of organ transplantation (in Egypt), and transmission of structural biology through bodily performance; race and the marketing of medicine, genomics, and diversity; product design and safety law; executive coaching; psychiatry, depression, and negotiating different class and religion-linked public health discourses in Iran; epidemic and transgenerational addiction as a techno-culture of affect; computational a-life and American folk kinship idioms, deep-water marine biology and the logics of ancestry, and the microbiopolitics of bacteria and cheese; climate change vernaculars in Canadian First Nations, US evangelicals, corporate green audits, climate scientists in the public sphere, and climate science under different national agendas; biological crafts from do-it-yourself biologies and geometric modeling with yarns and plastics to molecular gastronomy and synthetic biology; citizen epidemiology, gas fracking, and endocrine disruptors; 'smart' electricity market making by engineers and traders in contrast to economists, and digital map making in contests between First Nations, mining geologists, and forest ecologists.

Translating Legacy Knowledges into Public Futures

Four genealogical cables, each made up of many wires or threads, feed into the anthropology of science and technology's translation of legacy knowledges into public futures, buffering future shock, and tracking emergent forms of life. I will call these four cables 'test drives and libidinal drives,' 'protocols and networks,' 'landscapes or ethical plateaus,' and 'reknitting global moieties split by the cold war.'

1. Test drives and libidinal drives: You beta test a new technology, you test a hunch or hypothesis, you test constitutional issues in court, and you have libidinal drives. Biology, arguably, has replaced physics as the lead in upsetting traditional understandings of how the world works. In biomedicine, we are arguably undergoing epistemic revolutions, all too slow for the desperately ill but stunningly fast in terms of the pedagogies in which we have been trained. Two indices for these epistemic changes – the molecular biology revolution and bioecological imaginaries – may be sufficient to direct attention and to indicate the cascades, ramifications, and implications of paradigmatic shifts. There are glimmerings not only that we might be able to biologically repair our organs, but also that the self–other distinction is far more complicated once we recognize that our bodies are composed of more microorganisms and metabolic processes than human cells. Not only might genomics help stop or mitigate pandemics such as SARS or avian flu, but also comparative genomics might help identify lively processes of defense and repair across species, a new bioecological imaginary. 'Test drives' is a name for the nexus of genealogies of getting to contemporary understandings, say, from the old idea of vitalism (i.e., that life could not be reduced to the laws of chemistry and physics) to the more contemporary recognition that indeed one can synthesize living tissue. Creating life from chemicals (again) in the lab is one meaning often given to the term 'synthetic biology,' but another is something already happening: the engineering of bacteria to produce fuel, detect arsenic in drinking water, or produce medicinal drugs. Test drives also evoke libidinal drives, the entangled emotional fears (dystopian implications) and desires (utopian ones) emplotted in popular genres and metaphors of speaking and writing about science and technology, such as the impossibilities of complete control in treating epidemics; the gaps between triumphalist development schemes and their down-geared implementations; and emotional investments in promissory venture capital and their socialities of discipline, coding, and incentivizing.
2. Protocols and networks: The genealogical thinking here involves the great debates over rationality and rationalities,

over pragmatism's understanding of the public sphere as called into existence through the unintended consequences of decisions made by experts, politicians, and others; open and closed societies; the autonomy of science versus directed science for social ends; the coproduction of political and scientific authority of science and through regulation; engineering models of social action of human and nonhuman actor–actant networks, cyborg anthropologies, parasites and viruses, companion species, and other biological fellow travelers; and open and closed cybernetic systems that can tend toward equilibria, spin out of control into disequilibria or positive feedback, or expand fractally, or autopoeitically, via self-correcting or self-destroying modes of self-assembly. I call these 'protocols' or 'writing in programmed object-oriented languages,' through which 'ontology' becomes mutated from philosophy's metaphysical (untestable) and singular 'ground of being' into malleable, plural, and utilitarian units (ontologies in the plural) that software engineers encode. The accent here is on experimentation, and on the productive disagreement or continuous dialectic between science and engineering.

3. Landscapes or ethical plateaus: Ethnography is not merely descriptive. Once heard or read, the travails of patients and doctors in Botswana's only cancer ward (Livingston, 2012) will be hard to forget. Such is the force of many ethnographic case studies: descriptive, ethnographic, comparative, and acknowledging global network perfusion throughout local worlds. Ethnography, moreover, as Livingston argues, can be an early warning system of problems that are just emerging (e.g., a cancer epidemic following efforts to control HIV/AIDS). The genealogical thinking here is ethnographic: finding out what scientists and engineers actually do, not what they say they do; following their production of social organizations, power plays, public demonstrations, infrastructures, and resources for the future; and following them through the multiple sites and scales of interaction with other disciplines, institutions, economies, polities, publics, and generations. While some social historians and autobiographies or memoirs of scientists have provided precursors, the anthropology of science and technology is relatively young, beginning in the 1970s and 1980s and picking up speed and range of topics in the 1990s. The accent here is on how scientists and engineers pursue their crafts; how those crafts have changed with new instruments, new communities of practice, and new research programs or national projects; and, in a larger sense, how general epistemes and knowledge ecologies have changed (e.g., how large data sets, computational strategies, molecular or nano-scale experimentation, and so on have changed how we think, perceive, and evaluate). Physics, biology, bacteriology and medicine, computer science and new media infrastructures, and ecological and environmental issues have all been brought under the ethnographic gaze, often with more than merely descriptive effects.
4. Reknitting global moieties split by the cold war: The calls of the twenty-first century are routed through switches, transducers, translations, and satellites from Mumbai and Shanghai, Bandung and Tehran, Chongqing and Bangalore, Dubai and Singapore, and elsewhere. Do these portals open into new epistemes, armatures, mythologies, and experimental regimes? The genealogical thinking here involves the rapid globalization of scientific and engineering competitions, such that STS or history and philosophy of science departments focused only on the United States and Europe (as has long been the case) are simply archaic and out of touch. The largest genomic-sequencing center in the world in 2013 is in China, over one-third of the postdocs in US biology labs are Chinese nationals, and a large number of Silicon Valley information technology entrepreneurs are part of the Indian diaspora, many of whom are returning to invest in a newly booming Indian and transnational economy. Acquisitions and mergers by once third and second world companies of first world enterprises are changing even the landscape of multinationals. Similarly archaic and out of touch are histories of science in India, China, or the Arab world that are simply nostalgic romanticizations of long-past glories, usually ending in the early modern period. Suddenly, the world is as if two moieties, or two halves, of a whole were allowed back into intimate exchange and interaction. The accent here is on competitive cooperation, translation, and getting along with partial understandings that have always been features of scientific and technological worlds, but that now take on increased salience, breadth, and bite across national boundaries of distrust and intense advantage seeking. The leading edges in this competitive cooperation are biosecurity and pandemic control; food, water, energy, minerals, and rare earths resource security; media, information, data, and tourist circuits; and trade and intellectual property regimes. Important harmonics in this competitive cooperation are also in the areas of holistic medicine to repair the limitations of biomedicine, experimental green city expertise, logistics and finance that are sustainable and less subject to unexpected boom-and-bust cycles, and conflict resolution over the long term involving new international and global regulatory regimes. All of these have become new areas of ethnographic-anthropological work, transforming the ways in which medical anthropology, economic anthropology, political anthropology, and studies of 'material culture' and technologies have previously been framed and conducted.

Open Futures

It is hard to keep the generations straight. Legacies channel and constrain futures, only later to be rediscovered or reconceived in new contexts, generating new and sometimes different directions of thought and action. Switching points among multifocal and multinational agendas and capacities, 'Gestalt' or 'paradigm' switches in new technoscientific ways of living; new ethical plateaus in arbitrating decision making across intersecting knowledges and imperatives; and coping with alterities, cyborgs and companion species, and rhizomes and parasites – these are some of the philosophical, semiotic, mechanical, and biological indexical terms of reference. They also gesture to knots or quiltings, striations and inequalities, of emergent forms of life. Li Shizhen (1518–93) and Garcia da Orta (1500–68), who were at different ends of the Silk Road (Yangtze and Goa, respectively) and were collectors of

materia medica, founders of botanical gardens, and compilers of the medical empirical knowledges of their times, are parts of the varied moieties pulled apart by historical polities that now are reknitting into new fabrics of exchange and competition.

See also: Cognitive Anthropology; Energy, Anthropology of; Health: Anthropological Aspects; Infrastructures of Society, Anthropology of; Knowledge, Anthropology of; Multisited Field Studies.

Bibliography

Biagioli, Mario, 1993. Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago.

Biagioli, Mario, 2006. Galileo's Instruments of Credit: Telescopes, Images, Secrecy. University of Chicago Press, Chicago.

Bijker, Wiebe, 2002. The Oosterschelde Storm Surge Barrier: a test case for Dutch water technology management and politics. Technology and Culture 43, 569–584.

Bijker, Wiebe, Hughes, Thomas, Pinch, Trevor (Eds.), 1987. The Social Construction of Technological Systems. MIT Press, Cambridge, MA.

Cefkin, Melissa (Ed.), 2009. Ethnography and the Corporate Encounter: Reflections on Research in and of Corporations. Berghahn, New York.

Clarke, Adele, 1998. Disciplining Reproduction: Modernity, American Life Sciences and 'the Problems of Sex'. University of California Press, Berkeley.

Collins, Harry, 1974. The TEA set: tacit knowledge and scientific networks. Science Studies 4, 165–186.

Davis-Floyd, Robbie, Dumit, Joseph (Eds.), 1998. Cyborg Babies: From Techno-Sex to Techno-Tots. Routledge, New York.

Downey, Gary, Dumit, Joseph (Eds.), 1997. Cyborgs and Citadels: Anthropological Inventions in Emerging Sciences and Technologies. School of American Research Press, Santa Fe.

Downey, Gary, Dumit, Joseph, Williams, Sarah, 1995. Cyborg anthropology. Cultural Anthropology 10 (2), 264–269.

Fleck, Ludwik, 1935. Genesis and Development of a Scientific Fact (English Translation 1979). University of Chicago Press, Chicago.

Fortun, Kim, 2001. Advocacy after Bhopal: Environmentalism, Disaster, New Global Orders. University of Chicago Press, Chicago.

Fortun, Michael, 1993. Mapping and Making Genes and Histories: The Genomics Project in the United States, 1980–1990 (Ph.D. dissertation). Harvard University.

Fortun, Michael, 2008. Promising Genomics: Iceland and deCODE Genetics in a World of Speculation. University of California Press, Berkeley.

Fox, Rene, Swazey, Judith, 1974. The Courage to Fail: A Social View of Organ Transplants and Dialysis. University of Chicago Press, Chicago.

Fox, Rene, Swazey, Judith, 1992. Spare Parts: Organ Replacement in American Society. Oxford University Press, New York.

Fujimura, Joan, 1996. Crafting Science: A Sociohistory of the Quest for the Genetics of Cancer. Harvard University Press, Cambridge, MA.

Galison, Peter, 1997. Image and Logic. University of Chicago Press, Chicago.

Haraway, Donna, 1991. Simians, Cyborgs, and Woman: The Reinvention of Nature. Routledge, New York.

Haraway, Donna, 1997. Modest_Witness@Second_Millenium: FemaleMan© Meets OncoMouse™. Routledge, New York.

Haraway, Donna, 2003. The Companion Species Manifesto: Dogs, People, and Significant Otherness. Prickly Paradigm, Chicago.

Hogle, Linda, 1999. Recovering the Nation's Body: Cultural Memory, Medicine, and the Politics of Redemption. Rutgers University Press.

Jasanoff, Sheila, 1990. The Fifth Branch: Science Advisors as Policy Makers. Harvard University Press, Cambridge, MA.

Jasanoff, Sheila, 1995. Science at the Bar: Law, Science, and Technology in America. Harvard University Press, Cambridge, MA.

Jasanoff, Sheila, 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton.

Latour, Bruno, 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA.

Livingston, Julie, 2012. Improvising Medicine in an African Oncology Ward. Duke University Press, Durham.

Mahvunga, Clapperton, 2013. The Mobile Workshop. MIT (forthcoming).

Martin, Emily, 1987. The Woman in the Body. Beacon, Boston.

Martin, Emily, 1994. Flexible Bodies: Tracking Immunity in American Culture from the Days of Polio to the Age of AIDS. Beacon, Boston.

Martin, Emily, 2007. Bipolar Expeditions: Mania and Depression in American Culture. Princeton University Press, Princeton.

McNeil, William H., 1976. Plagues and People. Anchor/Doubleday, New York.

Mitchell, Timothy, 2002. "Can the Mosquito Speak?" In Rule of Experts: Egypt, Techno-politics, Modernity. University of California Press, Berkeley.

Polanyi, Michael, 1974. Personal Knowledge. University of Chicago Press, Chicago.

Rapp, Rayna, 1999. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, New York.

Reich, Michael, 1991. Toxic Politics: Responding to Chemical Disasters. Cornell University Press, Ithaca.

Spinage, Clive A., 2012. African Ecology: Benchmarks and Historical Perspectives. Springer, New York.

Suchman, Lucy, 1987. Plans and Situated Actions: The Problem of Human-Machine Communication, second ed. Cambridge University Press, New York.

Traweek, Sharon, 1988. Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA.

Science and Technology, Social Study of: Computers and Information Technology

Geoffrey C Bowker, Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, CA, USA
Susan Leigh Star[†]**,** University of California, La Jolla, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S.L. Star, volume 20, pp. 13638–13644, © 2001, Elsevier Ltd.

Abstract

Computers and information technology have over the past 50 years moved from being a social and engineering form which 'impacted' society to being core to our social fabric – much as we cannot study an ecosystem without considering water supply, so we cannot study any part of society without considering the flow of bits and bytes. The richly textured fabric of this fundamental social change has been studied at various levels, without there being as yet a single integrating vision. This article identifies some key levels and points to possibilities for an integrated understanding.

Introduction

Taking 1992 as the date of origin for the World Wide Web, then every child and adolescent, every young adult, has grown up in a world as the web's meshes were becoming ever finer. Access is not easy or universal, and there have been continuing stalwart efforts on the parts of industry and governments to ensure that this remains the case across the developing and developed worlds. It is not that everyone is on the web; rather that the businesses of business, government, war craft, education, and entertainment have become increasingly dependent on web use – so the nonuser is still caught in this inexorable flow. You simply cannot do a sociology of the social without addressing the pervasiveness of computing. Therfeore, I could fight my designated title and assert that really all contemporary sociology is implicated in the social study of computers and information technology.

However tempting the *mise en abyme* of this whole encyclopedia might be, I will follow a more modest path. In a computer-free context, Jacques Revel once wrote: "the change in the scale of observation revealed not just familiar objects in miniature but different configurations of the social" (Revel, 1995; cited in Feierman, 1999, p. 206). I take his tack to analyzing social studies of computers and information technology; in so doing, I explore social and political work done at the various layers of computing and information technology before finishing with some integrative visions.

From the early nineteenth century on, we have seen the growth of vast amounts of data collection about our citizens – the rise of national censuses leading to a flourishing of the arts of statistics (etymologically information about the state) and classification. Information and communication technologies such as the steamship, the railway, and the telegraph vastly increased the number of connections to be made socially, and the extent and potential sway of empires pulsing with a single beat. A large, interconnected population must be able to be managed; the great national censuses provided the key move of taking us from dealing with people in contexts to dealing with classes of people and reasoning our policies on the basis of these classes. In the same way, the great natural history surveys over the same period allowed us to conceive of nature as something to be managed, within human temporality, rather than an infinite resource to be exploited. Information infrastructures accompanied these developments – whether with the origin of computing in the organization of the massive insurance agencies that blossomed in the nineteenth century (and whose actuarial tables, later to serve as a basis for epidemiology, relied on census techniques) or the origin of computing in the development of the Hollerith punch-card machine, which enabled the processing of vast datasets in the 1890 American census (Yates, 1989; Campbell-Kelly, 1989). In the mid- to late-nineteenth centuries, railway and insurance industries were the two great behemoths of private enterprise. The railway industry was crucial, as Chandler (1977) notes, for the development of new forms of organizational management, as was the insurance industry. The big thing here was the problem of how to sell insurance to those who traditionally could not afford and would not consume it (in the words of a modern entrepreneur, they were competing with nonconsumption). To make it affordable, you needed really precise actuarial tables (of which more anon) and very low-cost procedures for both gathering those tables and processing claims – the margins were very thin, so every efficiency counted. Thus, as Martin Campbell-Kelly (1989) argues, new forms of office organization were developed around the drive for flows of data/information. Punch-card technologies, such as the Hollerith machines, were developed in the context of this preprepared niche. In a sense, the niche created the technology which in turn developed the niche in wildly new directions – a form of circular causal chain that is familiar to biogeographers (it is interminably circular to ask whether the niche or the terraforming organism came first). (Ian Hacking (1998) in his work on the traveler's fugue makes a similar argument about niches and railway technology in mental disorders.) Hence, fairly directly, office organization gave us a second computer precursor.

And then came the Second World War. In a stunning investment in an untried concept, the US government plowed the then enormous amount of $US 3 billion into a massively

[†]Deceased.

interdisciplinary project involving chemists, mathematicians, physicists, and engineers in order to create a singular weapon of mass destruction – the atomic bomb. A major issue that developed along the way was that you needed very complicated, dense, and precise mathematical calculations in order to get conventional explosives to create just the right symmetric implosive force to bring the core of the weapon to critical mass and unleash the furies. One of the first electronic stored-program computers, the ENIAC, was born. Hence, fairly directly, a scientific need gave us a third computer processor. (There are two other stories one could tell here – about Norbert Wiener, cybernetics, and antiaircraft guns; or about Alan Turing and code-breaking – which would complicate the picture in interesting ways.)

The lesson to be drawn from this is that if one looks at 'computers' and 'society' as separate and separable things, then one immediately gets drawn into some kind of determinism. With the fundamental cognitive tools of our time, we are dealing simultaneously with industry, organization, and science – and while we might not have a good word for that which is developing and subtends all three, we need the concept if we are to make sense of the question of digital technology, sociality, and epistemology. Further, we need to understand that occurrences of new digital epistemology (and ontology) can be traced back to earlier branches of the Tree of Knowledge – they certainly predate the Apple.

Automation and the Impact of Computerization

Many early studies of computers and society or computers and organizations concerned computing as an automation of human work. These included two major analytical concerns: the replacement of humans by machines in the workplace and the 'deskilling' of existing functional jobs (Braverman, 1998). Later, the analytic picture became much richer, as researchers realized that automation and deskilling were just two dimensions of the picture. An early breakthrough article by two social-computer scientists defined this territory as 'the web of computing' (Kling and Scacchi, 1982), a term that became a touchstone for the next generation. It refers to the co-construction of human work and machine work, in the context of complex organizations and their politics. Much of this research took place in business schools; some in departments of sociology or industrial psychology. In Scandinavia and other regions with strong labor unions, partnerships between the unions and researchers were important, at first focused on job replacement and deskilling, and later on computer systems design created through partnerships with users, social scientists, and computer designers.

The social impact of computing remains a strong research strand worldwide, despite the anthropological fact that finding 'untouched tribes' is increasingly difficult. Thus, the problematics of interest to science and technology studies (STS) have shifted from understanding the initial impact on a group or institution, to understanding ongoing dynamics such as usage, local tailoring of systems, shifts in design approaches, understanding the role of infrastructure in helping to shape impacts, and the ecology of computers, paper, telephones, fax machines, social media, face-to-face communication, and so forth – dynamics that are commonly found in most offices, and increasingly in homes and schools.

The Emulation of Human Cognition and Action (Artificial Intelligence and Robotics)

The early program of artificial intelligence (AI) work, beginning in an organized way during World War II, was to create a machine that would emulate human thinking (classic AI) and action (robotics). Some of the emulation work came in the form of intelligent tutoring systems and expert systems (Wenger, 1987) meant to replace or supplement human decision-making. The decision-making side of the research continues to be strong in management schools and in government agencies; it is often critiqued by STS researchers, although some do use these tools (Gilbert and Heath, 1985; Byrd et al., 1992). The discussion about whether emulation of human thinking would be possible dates far back; it came to the attention of many STS researchers with the work of philosopher Hubert Dreyfus (1979, 1992), who argued for the irreducibility of human thought, perception, and action to computerized mechanisms. Later STS researchers (including Diana Forsythe, Susan Leigh Star, and Lucy Suchman) formed direct collaborations with AI researchers. These partnerships took both critical and system-building forms; in both cases, the social science researchers acted as informants about the nature of the 'real world' as opposed to that of the emulated world of AI. AI changed during the late 1980s and early 1990s – some computer scientists spoke of the 'AI winter,' referring to funding problems unknown in the 1970–88 period. As personal computing and e-mail spread, and later, the Web (1992→), and the early promises of AI seemed not to be paying off, AI began to lose its prestige within computer science. Many AI researchers changed to problems in information science, software engineering, or cognitive science. Researchers in a branch of AI, distributed AI, continued to interact with the STS community (Huhns, 1987; Star, 1988). Their interests were in modeling and supporting spatially and temporally distributed work and decision practices, often in applied settings. This reflected and bridged to STS concerns with community problem-solving, communication and translation issues, and the division of labor in large scientific projects.

The Enterprise of Computing, Its Military Roots and the Role of Activism

In 1983 US President Ronald Reagan introduced his (in) famous 'Star Wars' (officially known as the Strategic Defense Initiative or SDI) military defense program, a massively expensive attempt to create a distributed laser and missile-based system in space that would protect the USA from foreign attack. The proposal put computer scientists, and especially software engineers, at the center of the design effort.

An immediate outcry was raised by the computing community as well as by alarmed STS scholars close to it. There were concerns that a system of that magnitude was untestable; tools were lacking even to emulate the testing. There were concerns about its technical feasibility. There also were concerns about

its ecological impact on the field of computer science – concerns akin to those raised by organismal biologists about the Human Genome Project – that the lion's share of research funds would be funneled into this project, orphaning others.

A new group, Computer Scientists for Social Responsibility (CPSR – http://cpsr.org), was formed in Silicon Valley as a focus for these concerns. The group quickly found common ground in many areas, including activist, ethical, policy, and intellectual questions. It has flourished and now sponsors or cosponsors three conferences a year in which STS scholars often participate: Computers, Freedom and Privacy (policy); Directions in Advanced Computing (DIAC) (intellectual and design directions, and their political and ethical bases and outcomes); and the Participatory Design Conference (PDC) (which brings together users, community organizers, computer and social scientists who work on issues of codesign and appropriate technology) (Kyng and Mathiassen, 1997). CPSR, as with its counterparts in the UK (Computers and Social Responsibility) and elsewhere, became an important meeting ground between STS and computer/information science. This was especially true of those critical of military sponsorship of computing agendas (at one point in the US some 98% of all funding of computer science came from the military, especially ARPA – the Advanced Research Projects of the Army (formerly DARPA), arguably the critical actor in the creation of the Internet (Abbate, 1999).

Beginning in the 1980s, a grassroots movement sometimes called 'community computing' arose, and linked with CPSR and other similar organizations. It attempted to increase access to computing for poor people and others disenfranchised from computing for a variety of reasons (Bishop, 2000). This often involved the establishment of 'freenets,' or publicly available free computing. Terminals were placed in venues such as public libraries, churches, and after-school clubs; some were distributed directly to people in need. This general problematic came to be called the 'digital divide,' and formed an active nexus of research across many areas of interest to STS.

An important part of the attempt to enfranchise everyone arose in the world of feminist activism and scholarship, as early computer hackers were nearly all male, and there were sexist and structural barriers to women as both users and professionals. Another meeting place between STS and computer/information science was formed by numerous conferences on women in computer science, gender and computing, and feminist analyses of the problem choice and ethics of computer science. An excellent series of conferences, whose proceedings are published in book form every 4 years, is sponsored by IFIPS (the International Federation of Information Processing Societies, Working Group 9.1, entitled Women, Work and Computing). See, for example, Grundy et al. (1997).

The 'Scandinavian School'

In the 1950s the (powerful) trade unions across Scandinavia helped to pass the 'co-determination legislation.' These laws stated that unions must be involved in technological design – originally motivated by concerns about deskilling and job loss through automation. A form of sociotechnical systems analysis evolved into a set of techniques for studying work places and processes. As computers arrived in the workplace, this came to include progressive computer scientists and social scientists, many of whom now participate in STS publishing and conferences, as well as Computer Supported Cooperative Work, Human Computer Interaction, and Participatory Design conferences (Greenbaum and Kyng, 1991; Bjerknes et al., 1987; Bødker, 1991; Neumann and Star, 1996). Gregory (2003) discusses the democratic bases of the multilevel Scandinavian participatory design approach – in particular their comfort with working through conflict and contradiction.

Computing, Standards, and Design

Although we rarely look under the hoods of our own computers, their formal structure and the protocols that allow them to interoperate are core sites of social and political activity. As Laura DeNardis (2009) so beautifully demonstrates, the whole issue of Internet governance is played out in standards-setting bodies who make decisions with vast sway. While standards in general remain somewhat understudied, they are the key to understanding the imbrication of social policy, computing, and design.

This theme has been explored beautifully by Steven Jackson (2014) in an article on the 'design policy knot.' They argue that when dealing with issues such as privacy and trust (core to our computing environment), we cannot look at design as an activity precursor to or separate from policy deliberations: our policies become part of the very information infrastructures we create (Braman, 2007). This lock-in occurs at the deepest level – as Helen Nissenbaum and colleagues have uncovered while exploring the Future of Internet Architecture projects funded by the National Science Foundation. It is possible, for example, to 'hardwire' the Internet for digital rights management; it is a political issue whether to do so.

Nissenbaum developed with Geoffrey C. Bowker, Susan Leigh Star, and Carl di Salvo, among others, the Values in Design movement, which through extended workshops held in the United States and Europe has trained graduate students in the methods of social studies of science and in design principles so that they can recognize and inflect the values that get built into all information systems. The databases that we use to store our information are a case in point. Attempts to fit 'indigenous knowledge' (the term is multiply problematic) into our own database structures is a perspicuous site for these issues. As Cori Hayden (2003) has shown, one method used by big Pharma has been to slot this knowledge (which they are mandated to collect) into unstructured text fields – making it extremely difficult to access, since databases take on their power through operations performed over multiple fields. The Indian Traditional Knowledge Digital Library makes the knowledge searchable, but at the price that it can only really be used to protect intellectual property claims – a mission itself alien to the knowledge producers themselves (Rajan, 2011). Helen Verran, Michael Christie, David Turnbull, and Ramesh Srinivasan have worked with aboriginal and Native American groups to represent their knowledge in its own terms – a drive that fits into the broad rubric of 'postcolonial computing' (Philip et al., 2012).

Knowledge Infrastructures

Recent decades have seen the rise and dissolution of categories such as 'digital libraries,' or 'scientific cyber-infrastructure' (when we were in a transition from laboratory-based to computer-based scientific work). We have now entered the age of the computer-based retronym where these qualifiers have become unnecessary. However, there has been a body of pioneering work within the field of STS which has tracked and theorized this transition.

Star and Ruhleder (1996) wrote a classic paper trying to untangle just what an (information) infrastructure was. Their categories – embedded in **Figure 1** – were notably not centered on the technological base alone, but on a set of sociotechnical observations. These have proven fertile in providing a definition that led to research developments of a range of concepts on old and new themes, such as invisible work, legitimate peripheral participation, and ontic occlusion (Cory Knobel). Millerand, Baker, and Bowker subsequently developed the Cartesian grid in **Figure 1** as a means to explore design decisions taken as cyber-infrastructures are built – the argument here being that willy-nilly designers chose to situate their decisions along local/global and social/technical axes. This concern with design reflected the changing role of researchers in this field. Where early science studies practitioners argued that the social scientist should not directly participate in the field they were studying, those studying the establishment of distributed, computer-mediated patterns of work increasingly found themselves drawn into working with developers (Ribes and Baker, 2007). This move represented an integration of science studies and information science (as practiced by the iSchools, themselves inheritors of the tradition of Library and Information Science). An influential report on Understanding Infrastructure has since been widely used in the policy and scientific literature – this has been complemented by a report on Knowledge Infrastructures, where the focus is centrally on the configurations of humans and nonhumans needed (see web references below).

In order to understand an emergent information infrastructure, one needs to be able to understand codes and conventions as well as observe practice. Paul Edwards (2010) demonstrated this brilliantly with respect to climate change – showing, for example, that modeling assumptions often got built into code and passed on through generations of programs without the scientists concerned being aware of the transmission. Lev Manovich has laid the framework for the emergent field of software studies, which is developing techniques for the critical analysis of computer code.

Mobile Technology

The most significant trend in computing over the past 20 years has been the rise of mobile devices. In the field's early days, a computer was a large machine in a cold room plugged into a wall – today most people would have difficulty enumerating all the computing devices distributed throughout their homes and about their persons. They have moved into being part of the very fabric of social life. The field of Ubicomp (ubiquitous computing) has grown up to track this and design for it.

A core question here has been whether what it means to be human and to be social is changing. There is a line of argument that we are just doing the old things in new ways. Thus, in a classic paper entitled 'A Rape in Cyberspace,' Julian Dibbell (1993) demonstrated that an online libertarian community soon grew rules very similar to those operating in 'normal' society when encountering deviant behavior. The myth of the early Internet was that it was open to all, free and democratic – today a handful of media companies holds vast sway over our collective attention economy and governments and

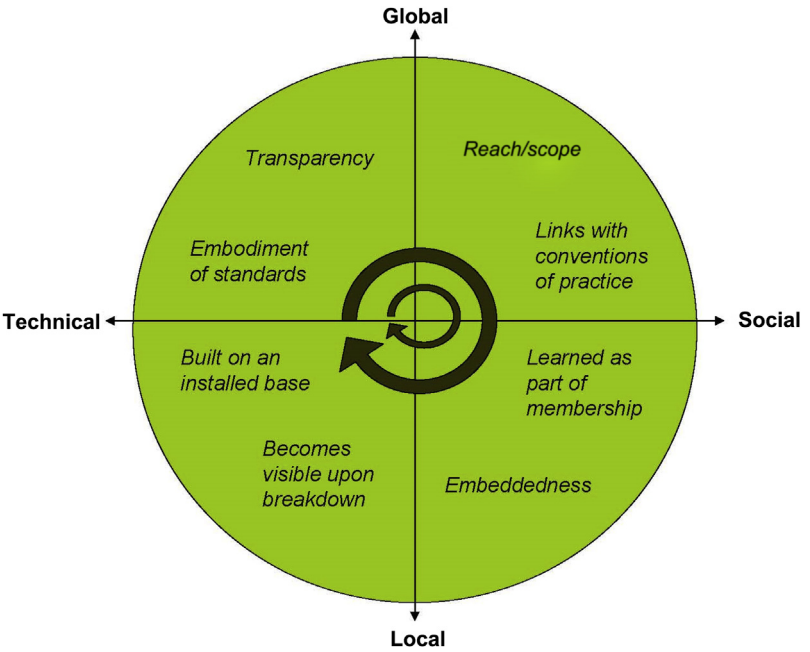

Figure 1 Cyber-infrastructure as *distributions* along technical/social and global/local axes.

corporations are attempting to build what are called 'walled gardens' – isolating the innocent user from tempestuous anarchy. The other end of this equation is the questionable assertion that new social and political movements have been enabled by mobile technology – the misnamed Arab Spring and the Occupy movement for example.

While this business-as-usual case can be somewhat persuasive, it does fall down when one recognizes that the very stuff of sociality is changing. A good analogy here is with the development of long-distance telephone networks in the last century. In the 1980s, there were myriad advertisements trumpeting the virtues of the telephone for keeping families together; yet they also helped drive families apart by allowing constant contact at a distance. Being a family member for many in the developed world came to involve trains, planes, automobiles, and telephones. Today it also involves Facebook and Twitter and their like. This interposition of (expensive) technology between people exercising social and political functions is business as normal for us – it does, although, speak to the need for radically rethinking our categories. Perhaps the most interesting theoretical development here has been Dominique Boullier's (2011) work on 'habitele' – or living at a distance. He is exploring how all aspects of our lives (reading, watching films, keeping calendars, transferring money, and so on) are converging onto our mobile devices, to the point where it is a misnomer to call them phones.

Big Data

Along the way, we are gathering huge amounts of data about ourselves and each other. Part of this is the traditional dream of social engineering – most immediately inherited from cybernetics: the controversy in 2014 around Facebook carrying out social experiments on their unwitting users is a case in point.

There are three dimensions of interest here. First is the development of new analytic tools by social scientists to study their objects. Thus, there have been a number of attempts to produce 'distant readings' (Moretti, 2013) of large text corpera and to do 'trace ethnography' (Geiger and Ribes, 2011). Distant reading is a technique akin to (but developed independently of) coword analysis – one seeks to identify patterns within huge text databases, such as the Ecco database of all published English language eighteenth century texts (de Bolla, 2014). Second is use of big data for surveillance and control by governments and corporations. This work draws in part on Foucault's concept of governmentality – what does it take to be able to govern one's citizens? Third is the study of ways in which we are wearing personal technologies which monitor our activities and provide us constant feedback. This Quantified Self social movement is currently being extended to monitoring bracelets, shoes, cars, and clothes.

Integrative Visions

The field of social studies of science had one origin in epistemological (David Bloor) and ontological questions (T.S. Kuhn) that constituted a discourse among historians, philosophers, and social scientists. The ontological status of 'nonhumans' has been a core question for Actor Network Theory (ANT). It is arguable that this interest itself derives in part from the history of computing – notably the field of cybernetics by way of Michel Serres. However one comes down on that issue, it is notable that technical talk today of the Internet of Things is wading in these ontological waters (Bogost, 2013).

The most interesting front in today's research is exploring our new social ontology. Dourish and Mazmanian (2013) are developing a corrective to the 'mind/body' split in much analysis of computing – bits and bytes do not exist in a 'cloud' but typically in vast distributed server farms; they have weight. The question of what it means to be human is changing fundamentally, and our field is primed to track this change. Jed Brubaker (Brubaker et al., 2013), for example, is exploring the changing nature of death by looking at death on Facebook – his findings range from the status of a dead person's account (Is it part of their identity, can it be 'passed on'?) to a body of users convinced that the heavenly host track their own Facebook accounts. The nature of memory is changing – just as our cognition is distributed among people and tools, so is our 'memory' increasingly not stored in the head or located elsewhere (a reference library).

A fundamental change in means of holding, storing, and sharing information is a rare historical event – one thinks of the origin of writing and the technology of the list; the invention of the printing press, and the technologies of absolutist bureaucracy together with the legacy of humanism (Clanchy, 1993). It constitutes a Kuhnian paradigm shift in the sense that fundamental social and political concepts no longer mean what they used to under previous regimes – indeed, it makes no sense to seek a continuity in the terms over the chasm. For a government to be 'democratic' means something different before and after the invention of a national press. This is being most closely felt today in the institutions of 'privacy' – the latter is just not a useful term in the modern state. For a person to be a member of a social group means something different with the invention of mobile technology. The so-called natural world is a different thing when you have terabytes of data from streaming sensors used in its monitoring and control. We as a field are realizing a return to our origins through increasing recognition that sociology, history, and philosophy can only advance hand in hand as we begin to understand the epochal transformations we are witnessing.

Acknowledgment

This article is based on work by Susan Leigh Star; updated by Geoffrey C. Bowker after her death. I have left Leigh's work, which describes the development of the field until 2000 largely untouched. I am indebted to Judith Gregory for her very helpful comments and changes – performed while on holiday and with broken toes.

See also: Citizen Science; Computers: Impact on the Social Sciences; Gender and Technology: From Exclusion to Inclusion?; Indigenous Knowledge: Science and Technology Studies; Information Society; Media and Social Movements; Mobile Communication; New Media and the Digital Divide; Science Communication; Social Media.

Bibliography

Abbate, Janet, 1999. Inventing the Internet. MIT Press, Cambridge, MA.

Bishop, A.P., 2000. Communities for the new century. Re/mediating adolescent literacies. Journal of Adolescent and Adult Literacy 43, 472–478.

Bjerknes, G., Ehn, P., Kyng, M. (Eds.), 1987. Computers and Democracy: A Scandinavian Challenge. Avebury, Aldershot, UK.

Bødker, S., 1991. Through the Interface: A Human Activity Approach to User Interface Design. Erlbaum, Hillsdale, NJ.

Bogost, Ian, 2013. Alien Phenomenology; or What It's Like to Be a Thing. University of Minnesota Press, St Paul.

Boullier, D., January 2011. Habitèle virtuelle. Revue Urbanisme 2011, 42–44.

Braman, Sandra, 2007. Change of State: Information, Policy, and Power. MIT Press, Cambridge, MA.

Braverman, Harry, 1998. Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century, 25th Anniversary Edition. Monthly Review Press, NY.

Brubaker, J., Hayes, G., Dourish, P., 2013. Beyond the grave: Facebook as a site for the expansion of death and mourning. The Information Society 29, 3.

Byrd, T.A., Cossick, K.L., Zmud, R.W., 1992. A synthesis of research on requirements analysis and knowledge acquisition techniques. MIS Quarterly 16, 117–139.

Campbell-Kelly, Martin, 1989. ICL: A Business and Technological History. Clarendon Press, Oxford.

Chandler, Alfred Dupont, 1977. The Visible Hand: The Managerial Revolution in American Business. Belknap Press, Cambridge, MA.

Clanchy, Michael T., 1993. From Memory to Written Record. Cambridge and Oxford, Blackwell, England, 1066–1307.

De Bolla, Peter, 2014. The Architecture of Concepts. Cambridge University Press, Cambridge.

DeNardis, Laura, 2009. Protocol Politics: The Globalization of Internet Governance. MIT Press, Cambridge, MA.

Dibbell, Julian, December 21, 1993. A Rape in Cyberspace. The Village Voice.

Dourish, Paul, Mazmanian, Melissa, 2013. Media as material: information representations as material foundations for organizational practice. In: Carlile, P.R., Nicolini, D., Langley, A., Tsoukas, H. (Eds.), How Matter Matters: Objects, Artifacts, and Materiality in Organization Studies. Oxford University Press, Oxford and New York, pp. 92–118.

Dreyfus, Hubert L., 1979. In: What Computers Can't Do: The Limits of Artificial Intelligence. Harper & Row, New York.

Dreyfus, Hubert L., 1992. What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press, Cambridge, MA.

Edwards, Paul, 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press, Cambridge, MA.

Feierman, Steven, 1999. Colonizers, scholars and the creation of invisible histories. In: Bonnell, V., Hunt, L. (Eds.), Beyond the Cultural Turn: New Directions in the Study of Society and Culture. University of California Press, Berkeley and Los Angeles, pp. 200–209.

Geiger, R. Stuart, Ribes, David, 2011. Trace ethnography: following coordination through documentary practices. In: Proceedings of, 44th Hawaii International Conference on System Sciences (HICSS), 4–7 January, pp. 1–10.

Gilbert, G. Nigel, Heath, Christian (Eds.), 1985. Social Action and Artificial Intelligence. Gower, Aldershot. IEEE Digital Library, UK.

Greenbaum, Joan, Kyng, Morton, 1991. Design at Work: Cooperative Design of Computer Systems. Erlbaum, Hillsdale, NJ.

Gregory, Judith, 2003. The Scandinavian approach to participatory design. The International Journal of Engineering Education 19, 63–74.

Grundy, A.F., et al. (Eds.), 1997. Women, Work, and Computerization: Spinning a Web from Past to Future. Proceedings of the 6th International IFIP Conference. Germany: Bonn, 24–27 May. Springer, Berlin.

Hacking, Ian, 1998. Mad Travelers: Reflections on the Reality of Transient Mental Illnesses. University of Virginia Press, Charlottesville.

Hayden, Cori, 2003. When Nature Goes Public: The Making and Unmaking of Bioprospecting in Mexico. Princeton University Press, Princeton.

Huhns, M. (Ed.), 1987. Distributed Artificial Intelligence. Morgan Kaufmann, Los Altos, CA.

Jackson, Steven J., Gillespie, Tarleton, Payette, Sandy, 2014. The policy knot: re-integrating policy, practice and design in cscw studies of social computing. In: CSCW '14 Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 588–602.

Kling, Rob, Scacchi, Walt, 1982. The web of computing: computing technology as social organization. Advances in Computers 21, 3–78.

Kyng, Morten, Mathiassen, Lars (Eds.), 1997. Computers and Design in Context. MIT Press, Cambridge, MA.

Moretti, Franco, 2013. Distant Reading. Verso, London.

Neumann, L., Star, S.L., 1996. Making infrastructure: the dream of a common language. In: Blomberg, J., Kensing, F., Dykstra-Erickson, E. (Eds.), Proceedings of PDC '96 (Participatory Design Conference). Computer Professionals for Social Responsibility, Palo Alto, CA, pp. 231–240.

Philip, Kavita, Dourish, Paul, Irani, Lilly, 2012. Postcolonial computing a tactical survey. Science Technology, & Human Values 37 (1), 3–29.

Rajan, Kaushik Sunder, 2011. Property, rights, and the constitution of Indian biomedicine: notes from the Gleevec case. Social Research 78 (3), 975–998.

Revel, Jacques, 1995. Microanalysis and the construction of the social. In: Revel, J., Hunt, L. (Eds.), Histories: French Constructions of the Past. New Press, New York, pp. 492–502.

Ribes, David, Baker, Karen S., 2007. Modes of social science engagement in community infrastructure design. In: Steinfield, C., Pentland, B.T., Ackerman, M., Contractor, M. (Eds.), Proceedings of Third International Conference on Communities and Technology. Springer, London, pp. 107–130.

Star, Susan L., 1988. The structure of ill-structured solutions: heterogeneous problem-solving, boundary objects and distributed artificial intelligence. In: Huhns, M., Gasser, L. (Eds.), Distributed Artificial Intelligence, vol. 2. Morgan Kauffmann, Menlo Park, CA, pp. 37–54.

Star, Susan L., Ruhleder, Karen, 1996. Steps toward an ecology of infrastructure: design and access for large information spaces. Information Systems Research 7, 111–134.

Wenger, Etienne, 1987. Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge. Morgan Kaufmann, Los Altos, CA.

Yates, JoAnne, 1989. Control through Communication: The Rise of System American Management. Johns Hopkins University Press, Baltimore, MD.

Relevant Websites

http://knowledgeinfrastructures.org/ – Knowledge Infrastructures.

http://lab.softwarestudies.com/ – Software Studies.

http://deepblue.lib.umich.edu/handle/2027.42/49353 – Understanding Infrastructure Report.

http://www.nyu.edu/projects/nissenbaum/vid/ – Values in Design.

Science and Technology Studies, Ethnomethodology of

Michael Lynch, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Ethnomethodology is a research program that studies the situated production and coordination of social practices. Starting in the 1960s, ethnomethodologists studied practices in a broad range of ordinary and professional settings. Ethnomethodology's orientation to local practices and situated knowledge influenced constructionist, actor-network, and discourse analytic approaches in science and technology studies (STS). Ethnomethodologists treat social science research practices as having no special epistemological status in relation to other scientific and nonscientific practices they study. This research policy is known as ethnomethodological indifference, and it is similar in some respects to the symmetry postulate in the sociology of scientific knowledge. However, unlike many STS scholars, ethnomethodologists do not treat professional sociology as a basis for authoritative explanations of other practices. Ethnomethodology has been criticized for its apparent lack of epistemological foundation and normative commitment, but proponents of the approach argue that their understandings and judgments have an ordinary basis in communal life rather than an epistemological foundation furnished by an academic school, theory, or method.

Ethnomethodology is a sociological approach to the study of practical actions, which has influenced the development of constructionist, actor-network, and discourse analytic approaches in science and technology studies (STS). This influence is explicitly acknowledged as an important part of the 'practice turn' in philosophy, history, and social study of science (Schatzki et al., 2001; Soler et al., 2014). Ethnomethodology's legacy also remains implicit in the now-commonplace talk of local practices, situated knowledge, discourse, and reflexivity in studies of science, technology, and medicine. In addition to the having pervasive influence throughout STS, ethnomethodology continues to develop distinctive lines of research on scientific and technical practices.

In the 1960s, Harold Garfinkel (1917–2011) coined the term *ethnomethodology* as a name for a unique sociological approach to practical actions and practical reasoning (Garfinkel, 1967, 1974; Lynch, 1993: 3–10; vom Lehn, 2013: Chapter 5). Early in his career, Garfinkel developed an interest in sociological theories of action (especially that of Talcott Parsons, 1937), but was dissatisfied with the way such theories glossed over the practical reasoning and interactional exchanges that characterize ordinary situations of action. He found inspiration in existential phenomenology, and particularly Edmund Husserl's and Alfred Schutz's phenomenological conceptions of the lifeworld. Others, and to some extent Garfinkel himself, also draw upon Ludwig Wittgenstein's later philosophy with its emphasis on language as practical action (Coulter, 1989). Literally defined, ethnomethodology is the investigation of 'folk methodology' for producing the innumerable practical and communicative actions that constitute recurrent social activities. However, the term 'methodology' should not be understood to suggest that 'folk' practices are designed along the lines of canonical scientific (or social scientific) methods. Instead, as ethnomethodological studies of routine scientific work have demonstrated, the practices in question – whether taking the form of family dinner table conversations or routine procedures in scientific laboratories – involve orders of discursive and embodied actions that rarely register in formal accounts of method.

Ethnomethodological Indifference, the Symmetry Postulate, and Reflexivity

Unlike methodologists in the philosophy of science who attribute special epistemic status to scientific methods, ethnomethodologists examine methods for composing and coordinating ordinary as well as scientific activities. Garfinkel used the term 'ethnomethodological indifference' to emphasize that any method whatsoever is worthy of study, regardless of its professional status, relative importance, adequacy, credibility, value, and necessity (Garfinkel, 2002: 170; Garfinkel and Sacks, 1970: 345–346). This does not mean that ethnomethodologists do not care about questions of value and validity; instead, it means that they believe that prior assumptions about the importance or validity of particular practices should not limit the choice of research topics. Not only might further investigation reveal that such preconceptions are misleading, but such investigation might also demonstrate that apparently trivial practices are organizationally important and resilient features of social life. And, in many cases, the investigator is in no position to make an ultimate judgment about the validity and value of the practices studied.

Ethnomethodological indifference is similar in some respects to the more familiar 'symmetry postulate' of the Strong Program in the sociology of scientific knowledge (Bloor, 1991[1976]: 4–5). The symmetry postulate is that the same general form of (social) explanation should apply to theories or doctrines of all kinds, regardless of their current standing as scientific or not, and regardless of whether the analyst initially holds them to be true, false, successful, or unsuccessful. However, there is a significant difference between ethnomethodology and the Strong Program, which is that the former delves into the practical and interactional production of embodied and discursive activity, while the latter focuses on more abstract contents of knowledge, and proposes to use established sociological concepts and methods to explain what passes for knowledge in a given time or place. Despite embracing a version of relativism, and contrary to accusations that his program is antiscientific, Bloor (1991[1976]: 160) explicitly embraces scientism for his program, not as a fallacious effort

to boost credibility but as an 'optimistic' orientation to the possibility of gaining objective (social) scientific knowledge. In part, the difference between Garfinkel's and Bloor's programs has to do with a less-obvious implication of the term ethnomethodology. In addition to being a name for an academic field that studies folk methods, the term refers to methodologies – systematic analyses of investigative procedures and evaluations of their efficacy – which are integral to many practices studied as well as the social scientific methods used for studying them. Many social scientists (including quite a few in STS) assert that scientists conduct their research unreflexively, unaware of the extent to which they implicitly construct the qualities they assign to nature, and that the intervention of an outside analyst is necessary for unmasking, explaining, and criticizing the tacit epistemologies and ontologies. In contrast, ethnomethodological indifference extends to the privileges ascribed to social scientific analysis itself:

> Persons doing ethnomethodological studies can 'care' no more or less about professional sociological reasoning than they can 'care' about the practices of legal reasoning, conversational reasoning, divinational reasoning, psychiatric reasoning, and the rest.
>
> Garfinkel and Sacks, 1970: 142

Not only does such indifference apply to topics of investigation, it also applies to methods. Perhaps most obviously in the natural sciences, systematic practical activities often include extensive methodological inquiry and debate. Natural science methodologies are often held up as exemplary, but scientists are not alone in the way they devote sustained attention to their own methodic practices. In modern (and also many ancient) societies, large bodies of literature, and countless conversations, instruct, articulate, discuss, and debate the practical, epistemic, and ethical character of a broad array of practices, including religious practices, legal procedures, etiquette, food cultivation and preparation, sexuality, child rearing, and many recreational activities. Any effort to explain such practices sociologically must first come to terms with the fact that explanations of many kinds (including social explanations) are reflexively embedded in the history, production, and teaching of the practices themselves.

Ethnomethodological Reflexivity

Garfinkel (1967: 1) spoke of the 'reflexive,' or 'incarnate,' character of accounting practices and accounts. By this he meant that social activities include endogenous practices for displaying, observing, recording, and certifying their own regular, normative, and rational properties. In other words, social agents do not just behave in orderly ways; they examine, record, and reflexively produce their actions. This particular sense of reflexivity emphasizes collective, and often highly organized, accounting practices, and differs from the Enlightenment ideal of critical self-reflection. Rather than reserving reflexivity for the savant, ethnomethodology begins with the observation that the persons and organized groups that social scientists study already observe, describe, and make sense of their own methodic practices, and that when confronted with social science findings that conflict with their own understandings, they are likely to remain unmoved (for different views on ethnomethodology's stance on reflexivity, compare Lynch (2000) and Pollner (2012)).

Bruno Latour, who occasionally expresses a qualified affiliation to Garfinkel's ideas (see Latour, 2005), also makes a deflationary move to locate reflexivity in a flat landscape in which the social or cultural analyst dwells alongside the practitioner, rather than viewing practices abstractly from an imagined height above that world. He argues that "journalists, poets and novelists are not naive make-believe constructionists. They are much more subtle, devious and clever than self-conscious methodologists. They did not have to wait for post-modern writing to tell stories; they are as self-conscious as those who naively believe they are *more* self-conscious" (Latour, 1988: 170).

Ethnomethodology's orientation to 'endogenous' or 'incarnate' accounting practices does not require an uncritical acceptance of official doctrines. For example, scientific reports and scientists' reminiscences sometimes describe scientific practices in ways reminiscent of Mertonian norms and empiricist ideals, but rather than accept such accounts at face value a historian or ethnographer needs to pay attention to the more pervasive way in which methodic understandings, and understandings of method, play a constitutive (and sometimes subversive) role in a science under study. For ethnomethodologists, the research question is not just "How do a society's (or scientific field's) members understand their social world?," but "How do members perform actions so as to make them accountable – that is, intelligible, recognizable, and acceptable in the relevant field of practice?" Accordingly, reflexive accounts are not limited to thoughtful ruminations about past actions and present circumstances; they are themselves prospective and retrospective actions performed in unfolding public interactional situations, and thus they help constitute the very contents of the fields of practice in which they occur.

Topic and Resource

The idea that explanations, explanatory concepts, and, more generally, natural language resources are common to professional sociology and the activities sociologists study, is a source of long-standing consternation and confusion in the social sciences (Winch, 1990[1958]; Hutchinson et al., 2008). As Garfinkel and Sacks (1970: 337) noted:

> The fact that natural language serves persons doing sociology, laymen or professionals, as circumstances, as topics, and as resources of their inquiries furnishes to the technology of their inquiries and to their practical sociological reasoning its circumstances, its topics, and its resources.

An injunction developed by ethnomethodologists (Zimmerman and Pollner, 1970), and adopted by proponents of discourse analysis (Gilbert and Mulkay, 1984) is that social analysts should not confuse topic and resource. That is, they should not confuse the common sense explanations given by participants in the social activities studied with the analytic resources for studying those same activities. A variant of this injunction in social studies of science warns the analyst not to adopt, for polemical purposes, the rhetoric of scientific authority that their own studies make problematic (Woolgar,

1981; Ashmore, 1989). However, if it also is the case, as Garfinkel and Sacks suggest in the above quotation that both topics and resources are furnished by the social world in which the analyst is embedded, then it remains unclear where the professional analyst can turn to find a standpoint that is not already occupied by the actor – an actor, as conversation analysis reminds us, engages in 'analysis' as a means of interactional production. This question is taken up in the following section.

Ethnomethodology and the Problem of Description

In recent decades, STS has undergone a 'normative turn' (Lynch, 2014), and many participants in the field no longer are satisfied (if they ever were) to pursue the descriptive, nonevaluative orientation that characterized both the sociology of scientific knowledge and ethnomethodology. However, if it is the case that pragmatic methodologies, analytical concepts, and discursive repertoires are reflexive to social settings, independently of any social science methods that would describe them, this certainly should be the case as well for ethical and moral evaluations. The normative turn thus continues to face the problem of how any analytical account (and certainly any evaluation or recommendation) stands with respect to those accounts that are already generated in the settings investigated. Social scientists who have cut their teeth by refuting analytical strategies that invoke the superior validity and perspicacity provided by a (social) scientific perspective – a perspective is presumptively more objective than the common sense myths, superstitions, and ideologies of the benighted natives – no longer can fall back upon such strategies. This is why the problem of description is so acute for social and cultural analysts who conspicuously eschew the epistemic privileges associated with the natural and social sciences.

Ethnomethodologists do not agree upon a single solution to the problem of description, and contradictory ways of handling it are evident in the field, but some ethnomethodologists dissolve the problem by arguing that sociological descriptions do not require special theoretical or analytical auspices in order to be intelligible and cogent. As Sharrock and Anderson (1991) argue, under most circumstances in which descriptions are made and accepted as adequate, there is no need (and, in any case, no prospect) to overcome epistemological skepticism by attaining an extraordinary degree of certainty. Consequently, the unavoidably reflexive relationship between epistemic contents and existential conditions is not a 'methodological horror' (Woolgar, 1988: 32) that makes description or explanation impossible or essentially problematic, because an ethnomethodologist is in no worse (or better) shape than anyone else who would write intelligible, cogent, and insightful descriptions of particular states of affairs. Even so, there still remains the question of the scientific, or other, grounds for ethnomethodological descriptions.

The Possibility of 'Scientific' Descriptions of Human Behavior

A possible basis for ethnomethodological descriptions is presented in an early argument written by Harvey Sacks (1992: 802–805), the founder of a field that later came to be called conversation analysis. This was a brief, but intriguing, argument about the possibility of developing a science that produces stable, reproducible, naturalistic accounts of human behavior. Sacks observed that natural scientists already produce accounts of human behavior when they write natural language descriptions of methods for reproducing observations and experiments. Social scientists who attempt to describe methods in a broad range of professional and nonprofessional activities are in no worse of a position than natural scientists who attempt to describe their own particular methods. In other words, Sacks conceived of the replication of experiments as a particular (but not necessarily special) case of the reproduction of social structures. While social scientists generally do not aim to produce manuals on how to perform the activities they describe, they can write accounts of social actions, which are adequate in a praxeological sense: adequate to the production and reproduction of the relevant practices.

Instructed Actions

It is well-established in the sociology of science that replication is problematic. Instead of being a methodological foundation for establishing natural facts, particular instances of experimental replication often beg further questions about the detailed conditions and competencies they involve (Collins, 1985). To say that replication is problematic does not imply that it is impossible, but it does raise the question of how scientists manage to secure assent to the adequacy of their experiments. The phenomenon of just how scientists conduct experiments, and how they establish particular results in the relevant communities, has become a major topic for social-historians and ethnographers of science. Consequently, in the more general case of the reproduction of social structure, it seems reasonable to conclude that Sacks identifies a phenomenon for further ethnomethodological investigation, but not a grounding for a possible sociological research program (Lynch and Bogen, 1994). Garfinkel (2002: Chapter 6) names this phenomenon 'instructed actions.' Such actions involve two parts: first, a formal (written, formulated, or otherwise transsituational) set of instructions, such as a recipe, a description of how to do a laboratory method, a procedure described in a manual, etc. and, second, a singular effort to follow or enact the instructed action. The pairing of the two is the phenomenon. That is, the question to be investigated in any case is how formal instructions relate to instructed actions. This relation might be said to involve a variant of the hermeneutic circle, but instructed action is not an interpretation of a text, it is instead a course of worldly action in particular environments in which textual resources are used in an improvised way. Variations on the theme of instructed actions are endless, depending on the occasion, the media in which the instructions are conveyed, the competency of the reader, and so on.

Normative and Ethical Considerations

Ethnomethodology has been criticized for treating normative aspects of the activities studied as 'mere phenomena'

(Habermas, 1984: 106). Such criticisms may seem to have a point, for example, when Charles Goodwin (1994) in a much-cited article draws a parallel between discursive, gestural, and inscriptive practices used in an archeological dig to instruct novices on how to see a key structure in layers of soil, and those used by the defense in a famous trial to persuade a jury on how to see a videotape that appeared to show a group of police officers relentlessly and violently beating a prone suspect. The defense attorney and witnesses used various visual props and testimonies successfully to persuade the (notoriously credulous) jurors to see the police actions as instances of normal and legitimate professional practice. Goodwin does not overtly condemn or endorse the rhetorical use of such practices, but instead describes them in detail from a detached point of view. However, the lack of any explicit ethical evaluation in Goodwin's article does not reduce normative considerations to mere phenomena or arbitrary constructs; instead, his article effectively provokes a reader's response without need for explicit commentaries, principles, or rules instructing readers on how to respond. His descriptions of the trial testimony present the adversaries' displays and specifications of the police actions; contrastive accounts and demonstrations that incorporate normative judgments and ethical claims. Instead of recasting such judgments and claims in terms one or another transcendent ethical framework, Goodwin displays and explicates the way normative judgments were produced, addressed, and fought over by the parties in the trial court. He thus allows readers to draw their own judgments from a rich body of materials, rather than preempting such judgments with a theory or set of general principles. Accordingly, when ethnomethodologists investigate highly charged uses of, and appeals to, normative judgments, they enable readers to examine specific configurations of action and judgment that do not neatly fall under recipe versions of norms and values. Rather than invoking or developing a general normative theory, ethnomethodologists invite their readers to consider intricate situations of practice and contestation that no single general framework can possibly forecast or resolve.

Practical and ethical dilemmas are no less salient for ethnomethodologists than for cultural anthropologists, lawyers, and other students of, and participants in, the activities that give rise to such dilemmas. The policy of ethnomethodological indifference does not relieve the investigator of ethical choices and responsibilities. Like other investigators, ethnomethodologists should in some circumstances find it advisable to respect norms of privacy and propriety, while in others they may be compelled to expose wrongdoing. However, as a body of concepts, research policies, and exemplary studies, ethnomethodology does not supply a set of rules or ethical guidelines for making such difficult choices. This does not mean that ethnomethodologists are unethical, but that their ethical judgments have a broader existential basis than the specific doctrines and lessons furnished by participation in an academic field; they have what Hutchinson et al. (2008: 36) call a life-long 'extramural' basis in communal living, and are not derived from a particular academic school, theory, or method.

Laboratory Studies, Constructivism, and Actor-Network Theory

Ethnomethodology has been explicitly acknowledged for its contributions to two major developments: one in science studies, in connection with of series of ethnographic studies of scientific laboratories (commonly known as laboratory studies) starting in the late 1970s; and the other in connection with on-site investigations of technical activities in high-tech workplaces starting in the mid-1980s. Important initiatives for both of these developments came from a research program that Garfinkel called 'studies of work' (1986). Such work included practices associated with specific occupations and professions in the arts, sciences, and many other fields, but it also included commonplace activities of many kinds, such as driving in traffic, recreational pursuits, and household routines. The idea of conducting ethnographic studies of work was far from novel at the time, but what Garfinkel proposed was distinctive: to study the production of practical actions that make up the technical contents of work (playing musical instruments, conducting experiments, proving mathematical theorems, navigating vehicles in traffic, etc.) (for recent collections see Rouncefield and Tolmie, 2011; Tolmie and Rouncefield, 2013).

This aim to engage with technical contents contrasted with the deliberate neglect of such contents in most previous ethnographies of work, in favor of investigations of the social backgrounds and lifestyles of the actors, the systematic social functions of the action, the cultural supports of the activity, and so forth – virtually everything 'social' about the work except for the situated interactional production of the work itself. This agenda happened to coincide with the question Bloor raised at the very start of his *Knowledge and Social Imagery* (1991 [1976]: 3) to set up the Strong Program's agenda: "Can the sociology of knowledge investigate and explain the very content and nature of scientific knowledge?"

Ethnomethodology and Constructivism

By the end of the 1970s, several sociologists of science had conducted extended on-site observations of laboratory activities and they answered Bloor's question with an enthusiastic 'Yes' (Latour and Woolgar, 1979; Knorr Cetina, 1981, 1983). Not only did they answer affirmatively, but they also presented arguments and observations to support the idea that specific natural scientific facts were constructed in and through laboratory practices and interactions.

What exactly was meant by construction or social construction, and whether or not it undermined trust in the reliability of scientific practices and the solidity of scientific results, remained (and remains) ambiguous and confusing. Often, but not always, it meant that scientific results were fraught with uncertainty, that the process of scientific investigation was subject to unanticipated contingencies, and that highly artificial conditions that were set up to control such contingencies – conditions so thoroughly artificial that the laboratory was the last place in the world that one could hope to find raw 'nature.'

The early laboratory studies by Latour and Woolgar and Knorr Cetina assimilated selected themes from

ethnomethodology into constructionist arguments, but other ethnomethodologists preferred then, as now, to speak of the production rather than the construction of natural and social orders (Lynch, 1993; Button and Sharrock, 1993). While it may be so that no fact is ever free from human intervention – or construction in the broadest sense of the word – local uses of the term in empirical science (though not in mathematics and certain branches of theory) refer to research artifacts that are distinguished from natural objects (Lynch, 1985). Ethnomethodologists prefer not to speak indiscriminately of the construction or manufacture of knowledge, and prefer instead to maintain an indifferent orientation to the way distinctions between constructed and unconstructed realities are employed in practical action and situated argument. One could say that, instead of resolving all scientific facts and findings on the side of artifacts, ethnomethodologists take a symmetrical view of scientists' working distinctions between constructed and unconstructed phenomena.

The boldest claims made in laboratory studies licensed a tendency that has remained widespread in STS, which is to view any and all naturalistic claims with skepticism, and to treat references to science, objectivity, facts, and consensus as deliberate or disguised rhetoric that lends authority to ontological claims. Such claims and tendencies have drawn considerable attention (both negative and positive) to STS in the past three decades, though central figures in the field, for different reasons and in different ways, have backed away from the more florid constructivist claims (Collins and Evans, 2002; Latour, 2004). Although sociologists and anthropologists of science, technology, and medicine continue to conduct ethnographies, for the most part constructivism is now 'black boxed' (to use a common STS idiom), so much so that the word no longer provokes much interest, and the ideas that scientific facts have no special epistemic status, and that scientific work is messy and fraught with contingency, are more often assumed than critically examined. Nevertheless, exactly how these studies lend empirical support for the philosophical and political arguments that so often cite them remains very much in doubt (Doing, 2008).

Ethnomethodological studies of mathematics (Livingston, 1986; Greiffenhagen and Sharrock, 2011), and scientific observation and laboratory work (Garfinkel et al., 1981; Lynch, 1985), are often mentioned together with constructivist studies, but there is a clear difference in the overall point of the studies and the claims they make on their own behalf. The aim is not to answer Bloor's question with documented arguments to the effect that natural science and mathematics are constructed, but to show how specific proofs, observations, experimental protocols, and so forth, are produced, assessed for their adequacy, reproduced, disseminated, disputed, and agreed upon in specific laboratories and other communicative fields. Whether or not such work supports a constructivist philosophy, or suggests that science is inherently messy and political, is a matter of indifference. Consequently, though occasionally recognized and appreciated in STS, ethnomethodological studies tend not to provide clear support for general theoretical or normative programs, whether constructivist or not. This lack of clear resolution may be disappointing to some readers, but it motivates continued investigation of work in the sciences and mathematics, but not in order to document a constructivist (or other) philosophy of science, but to address basic questions that earlier studies claimed to resolve about the nature and content of science while leaving many stones unturned (Sormani, 2014).

Ethnomethodology and Actor-Network Theory

As mentioned earlier, there are points of affinity between ethnomethodology and actor-network theory (ANT) – a theory, or rather an ontology that for at least the past two decades has been strongly established in STS. Latour (2005: 54n.) goes so far as to say that it "would be fairly accurate to describe ANT as being half Garfinkel and half Greimas" (the latter being semiotician A.J. Greimas). As mentioned earlier, one point of agreement has to do with the conception of reflexivity, not as the scholar's or metaphysician's rare ability to reflect insightfully upon the motivations and social circumstances that the unreflexive actor fails to recognize or acknowledge, but as an intrinsic aspect in the field of action itself. More generally, there is agreement that, as Latour puts it, "it's not the sociologist's duty to decide in advance and in the member's stead what the social world is made of …" (2005: 29n.). Instead, the idea is to follow how the members assemble the social world, and not as a discrete domain separate from nature to be studied by social scientists. Moreover, like Garfinkel, Latour acknowledges that the conceptual and methodological instrumentarium offered by sociology simply cannot keep pace with the varied and nuanced, not to speak of handsomely resourced, practical and discursive maneuvers that it would hope to comprehend and explain.

There also are some clear differences. Whereas ethnomethodology, at least in principle, relentlessly eschews transcendental analysis, Latour turns to semiotics to gain leverage on how locally situated actions manage to transcend their circumstances. The transcendence is achieved via texts and related literary technologies and infrastructures that afford the collection, stabilization, movement, analysis, criticism, and reconstruction of communicative actions. Latour's grander (albeit often deflationary) philosophical writings on modernity show a willingness to ride the literary mobile to span vast territories. In contrast, Garfinkel was highly distrustful of textually conveyed journeys, and he expressed this distrust through a repeated insistence that practical action and practical reasoning take place only deceptively in texts, and that sociology is doomed to be a 'literary enterprise' by virtue of its almost exclusive reliance on documents, textual instruments (surveys, transcripts, and the like), coding schemes, and stories, both as data and as exhibits of analytical work. He even distrusted the literary technologies of tape recording and transcription that so many ethnomethodologists use for their own analyses of 'naturally organized' activities. This distrust also may have contributed to his convoluted prose, and his reluctance to complete and publish many of the writings he had drafted and outlined.

At times, such denigration of texts and literary enterprises may seem to express a quixotic desire to reach beyond the literary form that gives expression to that very hope to grasp the things themselves, just as the eschewal of theorizing may seem itself to be theoretically framed. However, the deceptively simple theme of instructed actions suggests a different picture. Far from ignoring or denigrating texts and formal analysis, the phenomenon of instructed actions pairs and blends formal

analysis with a counterpart: courses of action that recurrently make use of, supplement, and selectively consult and disregard instructions. Examining instructed actions by performing and describing specific efforts to follow a given set of instructions, directions, or plans does not aim to reveal that formal instructions always turn out to be inadequate. Instead, the aim is to show whether or not, and how, they become adequate in singular courses of action, and to identify what else such action involves. Just how they are used in action is the thing itself of interest to ethnomethodology.

The puzzling issue raised by Latour's effort to combine Garfinkel with Griemas is that the latter's story grammar is just the sort of general abstract schema that studies of situated actions tend to dissolve (see Lenior (1994) for a critique of Latour's use of Greimas; and Alac (2011) and Liberman (2013) for different ways to integrate semiotics with ethnomethodology). One aspect of Garfinkel's suspicion of textual form is that the analysis of texts (such as with a coding scheme or program) draws equivalencies between apparently similar linguistic figures and syntactic positions in diverse narratives, with very thin regard for contexts of use that can easily dissociate what appears to be equivalent on paper. The much debated move that Latour makes when he creates a second axis of symmetry seems, on the one hand, to be a radical move to make all 'actants' (agents of action described in a narrative) equivalent, regardless of their status as human (a person), collective (a government agency), or nonhuman (a disease agent). This radical move of inviting all manner of entities into the 'actor' category may seem to avoid the problem created by the oversocialization of nature and technology that Latour attributes to the sociology of scientific knowledge. On the other hand, the apparent equivalence between these actants is a formal equivalence that recalls the classic Saussurean distinction between stable elements of language and fluid combinations of those elements in speech. For Latour, it seems, Garfinkel articulates the fluidity of syntagma (an elementary constituent element of a text), while Griemas gives him a vocabulary and system for stabilizing narrative elements in a network. The problem is that the former threatens the stability and seeming transcendence of the latter, so that the one axis deconstructs the other. Latour has shown an admirable willingness and ability to think outside the box (or, rather, the semiotic square), including boxes of his own creation, but semiotics has provided an abstract anchorage (if that is not an oxymoron) for ANT. Perhaps ethnomethodologists should envy such anchorage, in preference to what Garfinkel facetiously called, during lectures and unpublished presentations, his "boat without a bottom" – inviting his critics to go ahead and "throw your rocks, they'll fall right through." However, rather than anchoring ethnomethodology in an abstract system, Garfinkel suggested instead that ethnomethodological studies of work could hybridize with an open variety of other fields. Some success on that front was achieved in studies of information technology.

Ethnomethodological Studies of Technical Practice

A variant of the theme of instructed actions – plans and situated actions – proved to be immensely important for launching an offshoot of ethnomethodology and STS that became established both in academic research on information technology and in corporate research centers where such technology is designed. Lucy Suchman (2007[1987]) creatively worked with the theme of instructed actions to observe and critique the investment in formal plans in cognitive science and artificial intelligence. Her research took place at, and was supported by, the Xerox Corporation's Palo Alto Research Center, and involved detailed recording and analysis of users' efforts to comprehend instructions conveyed through an interactive display panel on an (at the time) advanced photocopier. Suchman described the all too common confusion experienced by users (some of whom struggled hilariously to make photocopies with the smart machines), and such observations led her to develop a critique of programmed instructions that took little account of the situated actions of users who, typically, did not notice, attend to, or fully comprehend what the programmed instructions were telling them to do. Part of the problem, according to Suchman, was that machine instructions were designed along the lines of computer programs in which following instructions was viewed as a matter of mechanically executing plans, rather than engaging with a machine in an embodied way, in the midst of the distractions and variations in technical skill found in particular organizational environments. Drawing from ethnomethodology, STS, and cultural anthropology, Suchman pointed to the contingent, circumstantially relative way that formal plans, rules, and instructions relate to situated actions.

Xerox, Microsoft, and other corporations that maintain research centers were convinced by Suchman's and other ethnographers' research to support in-house groups of ethnographers to advise and collaborate with design teams to better take into account the potential fate of their programmed innovations when placed in the hands of users in offices and homes. This research not only found a home in industrial research centers, it also became highly important for interdisciplinary academic fields in communication and information studies, such as computer-supported cooperative work, human–computer interaction, and workplace studies (Luff et al., 2000). The hybrid was dubbed 'technomethodology' by Dourish and Button (1998), and while it has been a success in many respects, a point of contention among some of the key participants in the emergent field has been the tendency to use hybrid studies as a basis for reverting to the familiar social science theories, arguments, and disciplinary ambitions that such studies were designed to bypass (Crabtree et al., 2009).

One particularly difficult challenge for ethnomethodological studies of highly technical activities is to contend with the 'unique adequacy' requirement (Garfinkel, 2002: 169). This is an extension and radicalization of the classical sociological requirement that an adequate description of human actions must take into account (must exhibit understanding of) the agent's purposive orientation. In studies of technical practices, the unique adequacy requirement assigns priority to the cultivated technical understandings of skilled practitioners, rather than to the descriptive language, conceptual frames, and metaphors creatively devised by social analysts who have not mastered the practice in question. To paraphrase J.L. Austin, the practitioner's account may not be the last word

on the subject, but it is the first word. However, just as technical language and skill presents an initial barrier to the novice who would master a practice, it also presents a barrier for getting access to, observing, understanding, and writing about practices – especially when the proverbial 'average intelligent reader' is unlikely to know much of anything about the practice. One way to contend with the problem of conveying technical contents to nonspecialist readers is to devise analyses and demonstrations of relatively undemanding cases – analogous to the way Wittgenstein uses examples from elementary arithmetic and geometry in his *Remarks on the Foundations of Mathematics*, and justifies these examples as no less authentically mathematical than more esoteric cases. One relatively small domain of ethnomethodological studies makes use of historical cases such as Galileo's pendulum and inclined plane demonstrations, elementary school science lessons, and devised exercises using geometric patterns and easily purchased prisms (Garfinkel, 2002: Chapter 9; Bjelic, 2003; Livingston, 2008; Lynch and Macbeth, 1998). Such studies certainly do not grant access to cutting edge science and mathematics, but they do provide a way to contend with (both in writing and in embodied exercises with students) the contingencies of handling and improvising with equipment. They also enable readers to depart from textual accounts and interpretations, to gain first-hand access to struggles with how to see phenomena: to examine their own work of following instructions in novel situations, and to experience the frustrations and delights associated with embodied practice.

See also: Ethnography; Ethnomethodology, General; Fromm, Erich (1900–80); Laboratory Studies: Historical Perspectives; Reflexivity in Science and Technology Studies; Science and Technology, Social Study of: Computers and Information Technology; Science: Constructivist Perspectives, History of; Scientific Knowledge, Sociology of; Strong Program, Sociology of.

Bibliography

Alac, M., 2011. Handling Digital Brains: A Laboratory Study of Multimodal Semiotic Interaction in the Age of Computers. MIT Press, Cambridge, MA.
Ashmore, M., 1989. The Reflexive Thesis: Wrighting the Sociology of Scientific Knowledge. University of Chicago Press, Chicago.
Bjelic, D., 2003. Galileo's Pendulum: Science, Sexuality, and the Body-Instrument Link. SUNY Press, Albany, NY.
Bloor, D., 1991[1976]. Knowledge and Social Imagery, second ed. University of Chicago Press [first ed., Routledge and Kegan Paul, London].
Button, G., Sharrock, W., 1993. A disagreement over agreement and consensus in constructionist sociology. Journal for the Theory Social Behaviour 23, 1–25.
Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.
Collins, H.M., Evans, R., 2002. The third wave of science studies: studies of expertise and experience. Social Studies of Science 32, 235–296.
Coulter, J., 1989. Mind in Action. Polity Press, Oxford.
Crabtree, A., Rodden, T., Tolmie, P., Button, G., 2009. Ethnography considered harmful. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 879–888.
Doing, P., 2008. Give me a laboratory and I will raise a discipline: the past, present, and future politics of laboratory studies. In: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge, MA, pp. 279–296.
Dourish, P., Button, G., 1998. On 'technomethodology': foundational relationships between ethnomethodology and system design. Human-Computer Interaction 13 (4), 395–432.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice Hall, Englewood Cliffs, NJ.
Garfinkel, H., 1974. On the origins of the term 'ethnomethodology'. In: Turner, R. (Ed.), Ethnomethodology. Penguin, Harmondsworth, UK, pp. 15–18.
Garfinkel, H. (Ed.), 1986. Ethnomethodological Studies of Work. Routledge & Kegan Paul.
Garfinkel, H., 2002. Ethnomethodology's Program: Working Out Durkheim's Aphorism. Rowman and Littlefield, Lanham, MD.
Garfinkel, H., Lynch, M., Livingston, E., 1981. The work of a discovering science construed with materials from the optically discovered pulsar. Philosophy of Social Science 11, 131–158.
Garfinkel, H., Sacks, H., 1970. On formal structures of practical actions. In: McKinney, J.C., Tiryakian, E.A. (Eds.), Theoretical Sociology: Perspectives and Development. Appleton-Century-Crofts, New York, pp. 337–366.
Gilbert, G.N., Mulkay, M., 1984. Opening Pandora's Box: An Analysis of Scientists' Discourse. Cambridge University Press, Cambridge.
Goodwin, C., 1994. Professional vision. American Anthropologist 96 (3), 606–633.
Greiffenhagen, C., Sharrock, W., 2011. Does mathematics look certain in the front, but fallible in the back? Social Studies of Science December 41 (6), 839–866.
Habermas, J., 1984. The Theory of Communicative Action. In: Reason and the Rationalization of Society, vol. I. Beacon Press, Boston.
Hutchinson, P., Read, R., Sharrock, W., 2008. There is no such thing as a social science. In: Defence of Peter Winch. Ashgate, Aldershot, UK.
Knorr Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon, Oxford.
Knorr Cetina, K., 1983. The ethnographic study of scientific work: towards a constructivist sociology of science. In: Knorr Cetina, K., Mulkay, M. (Eds.), Science Observed. Sage, London, pp. 115–140.
Latour, B., 1988. The politics of explanation: an alternative. In: Woolgar, S. (Ed.), Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge. Sage, London, pp. 155–177.
Latour, B., 2004. Why has critique run out of steam? From matters of fact to matters of concern. Critical Inquiry 30 (2), 225–248.
Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, New York.
Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Facts. Sage, London.
vom Lehn, D., 2013. Harold Garfinkel: The Creation and Development of Ethnomethodology. Left Coast Press, Walnut Creek, CA.
Lenoir, T., 1994. Was the last turn a right turn? The semiotic turn and A.J. Greimas. Configurations 2, 119–136.
Liberman, K., 2013. More Studies of Ethnomethodology. SUNY Press, Albany, NY.
Livingston, E., 1986. The Ethnomethodological Foundations of Mathematics. Routledge and Kegan Paul, London.
Livingston, E., 2008. Ethnographies of Reason. Ashgate, Aldershot, UK.
Luff, P., Hindmarsh, J., Heath, C., 2000. Workplace Studies: Recovering Work Practice and Informing System Design. Cambridge University Press, Cambridge and New York.
Lynch, M., 1985. Art and Artifact in Laboratory Science. Routledge and Kegan Paul, London.
Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge University Press, New York.
Lynch, M., 2000. Against reflexivity as an academic virtue and source of privileged knowledge. Theory, Culture & Society 17 (3), 26–54.
Lynch, M., 2014. From normative to descriptive and back: science and technology studies and the practice turn. In: Soler, L., Zwart, S., Lynch, M., Israel-Jost, V. (Eds.), Science after the Practice Turn in Philosophy, History, and the Social Studies of Science. Routledge, London & New York, pp. 93–113.
Lynch, M., Bogen, D., 1994. Harvey Sacks's primitive natural science. Theory, Culture & Society 11, 65–104.
Lynch, M., Macbeth, D., 1998. Demonstrating physics lessons. In: Greeno, J., Goldman, S. (Eds.), Thinking Practices in Mathematics and Science Learning. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 269–297.
Pollner, M., 2012. The end(s) of ethnomethodology. American Sociologist 43, 7–20.
Rouncefield, M., Tolmie, P., 2011. Ethnomethodology at Work. Ashgate, Farnam, UK.
Sacks, H., 1992. Lectures on Conversation, vol. one. Blackwell, Oxford.
Schatzki, T.R., Knorr-Cetina, K., von Savigny, E. (Eds.), 2001. The Practice Turn in Contemporary Theory. Routledge, London.
Sharrock, W., Anderson, B., 1991. Epistemology: professional skepticism. In: Button, G. (Ed.), Ethnomethodology and the Human Sciences. Cambridge University Press, Cambridge, pp. 51–76.

Soler, L., Zwart, S., Lynch, M., Israel-Jost, V. (Eds.), 2014. Science after the Practice Turn in Philosophy, History and the Social Studies of Science. Routledge, New York and London.

Sormani, P., 2014. Respecifying Lab Ethnography: An Ethnomethodological Study of Experimental Physics. Ashgate, Hampshire, UK.

Suchman, L., 2007[1987]. Human-Machine Reconfigurations: Plans and Situated Actions, second ed. rev. Cambridge University Press, Cambridge [first ed. titled, Plans and Situated Actions. Cambridge University Press, Cambridge].

Tolmie, P., Rouncefield, M., 2013. Ethnomethodology at Play. Ashgate, Farnam, UK.

Winch, P., 1990[1958]. The Idea of a Social Science and its Relation to Philosophy, second ed. Humanities Press, Atlantic Highlands, NJ [first ed., Routledge & Kegan Paul].

Woolgar, S., 1981. Interests and explanations in the social study of science. Social Studies of Science 11, 365–394.

Woolgar, S., 1988. Science: The Very Idea. Tavistock, London.

Zimmerman, D.H., Pollner, M., 1970. The everyday world as a phenomenon. In: Douglas, J.D. (Ed.), Understanding Everyday Life: Toward the Reconstruction of Sociological Knowledge. Aldine, Chicago, pp. 80–103.

Science and Technology Studies, History of

Harald Rohracher, Linköping University, Linköping, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The interdisciplinary field of science and technology studies (STS) has only emerged in the late 1970s but has been influenced by various lines of thought in sociology, philosophy, and economics. At its basis is the claim that not only the social institution of science or the impact of technological change on society is open to social analysis but also the very content of science and engineering, i.e., scientific facts and technological objects. This article sketches out the development of STS from its roots, among others in the sociology of knowledge and history of technology, to contemporary issues such as the problem of expertise and transformations in the production of knowledge.

Introduction

Science and technology studies (STS) have emerged in the 1970s and 1980s as a distinct interdisciplinary field – or intradisciplinary as Jasanoff (2010) puts it to stress its focus on questions not raised within existing disciplines – which deals with the inseparability of science and technology from social structures and practices. The constitutive idea in contrast to its precursors in the sociology, history, or economics of science and technology is an understanding that also the content of science and engineering (scientific facts, technologies, objects) is open to social analysis and not the result of a privileged form of knowledge about nature. Facts and artifacts are but temporarily stable outcomes of heterogeneous activities of scientists and engineers and their entanglement in wider social and political relations. From a STS perspective sociotechnical assemblages are the smallest units of analysis and cannot be simply split into preexisting social and material dimensions, which are then analyzed separately. Society and technology/materiality coconstitute each other and mutually shape each other through processes of translation, inscription, and enactment. At its best, STS thus not only contributes to a more appropriate understanding of science and technology but also of society.

There are no strict demarcations of what counts as STS and what is already outside its realm. One may take a narrower view on the formation of core concepts and emerging disciplinary structures, or a broader view on STS as a multifaceted and heterogeneous field overlapping with, e.g., economic innovation studies, anthropology or cultural studies, and also comprising of a broad range of problem- and application-oriented STS research on health, climate change, or the democratization of science and technology. This article takes the broader view and complements the formation of an academic canon with the development of an 'engaged program' of STS (Sismondo, 2008). A number of introductory textbooks (e.g., Fuller, 2007; Hess, 1997; Sismondo, 2010; Yearley, 2005) as well as articles on the history and development of the field (e.g., Jasanoff, 2010; Latour, 2004; Pestre, 2004; Woolgar, 2004) along with the STS-related entries in earlier versions of this encyclopedia have already contributed to an ex-post consolidation and ordering of the field (and at the same time self-critical reflection on such attempts). This article builds on these contributions and – given the heterogeneity and ongoing dynamic of the field – presents a somewhat eclectic selection of further lines of thought and reasoning.

Precursors of STS

Science and technology have been an object of social sciences and humanities for a long time and various approaches and schools of thought have led up to the current concepts used in STS. Science has been regarded a crucial element in processes of modernization and the rise of a capitalist economy, as, e.g., expressed in the writings of Karl Marx or Max Weber. Only later, science and knowledge generation became direct objects of social scientific research, as in Karl Mannheim's sociology of knowledge and in the investigation of science as a social system in Robert K. Merton's work. Technology also has played a long-standing role in the social sciences and humanities, whether in social history (Mumford), political economy (Marx), or philosophy (Heidegger).

Science as a Social Institution

The development of a sociology of science is closely interwoven with the scholarly work of American sociologist Robert K. Merton. Merton analyzes science as a social phenomenon in at least three distinct ways (Barnes, 2007): (1) as a social unit functioning in a larger social system and being affected by the development of its social surroundings – as, e.g., analyzed in Merton's early work on the emergence of science in puritan seventeenth-century England; (2) as a set of peers, a moral community and a subculture with a distinctive normative order – Merton's work on the ethos of science expressed by four basic norms stands for this dimension of his work; and (3) the question of how the normative order in science is sustained by a specific reward system encouraging original research. Merton's article on 'The normative structure of science' (Merton, 1973[1942]) is probably his most popular and well-known contribution to the sociology of science and reflects the political context of the 1940s by emphasizing the importance of a democratic social structure for science to thrive and to let scientists live up to their ethos expressed by a set of four norms: communism (later termed communalism), universalism,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03064-6

disinterestedness, and organized skepticism. The institutional goal of science depending on these norms is 'the extension of certified knowledge' pointing clearly to the assumption that 'science as a social phenomenon' only extends to the social behavior and practices of scientists but not to the content of science itself. This strand of a sociology of science has remained influential throughout the postwar period with often controversial discussions about the need for additional norms, their empirical validity, and relevance for the social analysis of the community, culture, and institutional makeup of science (see e.g., Zuckerman, 1988).

The Social Constitution of Knowledge

Another strand of sociological thinking preparing the ground for STS is the sociology of knowledge, best represented in the work of Max Scheler and Karl Mannheim – particularly in his book *Ideology and Utopia* (Mannheim, 1929). The key argument is that knowledge is conditioned by the social position of the person or group who formed these thoughts. Everyone's belief and possibility to know depends on the social context he/she is immersed in, be it social class, religion, profession, or age cohort. Obviously, such a position has to deal with both (Heintz, 1993), the problem of relativism (no overarching criteria for the evaluation of knowledge claims) and reflexivity (as the positionality of knowledge also applies to the social scientist). Mannheim sought an answer to relativism in the 'multiperspectivity' of different knowledge claims and the aim of sociology of knowledge to understand why and how these different perspectives and knowledge claims differed. His strategy to deal with reflexivity, however, sets Mannheim apart from the later STS, as he chooses to exclude science and mathematics from the program of a sociology of knowledge and assign them a privileged epistemological status. His contemporary Ludwik Fleck (see next section) developed precisely such a program of a sociology of scientific facts.

History and Philosophy of Science and the 'Structure of Scientific Revolutions'

Thinking about the constitution and development of scientific knowledge has long been a prerogative of epistemology and the philosophy of science, which upheld the special status of scientific knowledge and its exclusion from sociological or historic analysis. However, not least developments within the theory of science paved the way to historic analyses of the development of scientific knowledge by Thomas Kuhn and others. Two core theses set the scene for the so-called anti-positivist or historic turn in the philosophy of science: the underdetermination of scientific theory as argued by Pierre Duhem and W.V.O. Quine and the theory-ladenness of observation. If isolated theoretical hypotheses cannot be simply falsified by contradictory observations and if theory and observation cannot be strictly separated, historic or sociological arguments need to stand in to explain the development and constitution of scientific knowledge. This program has been taken up in the book *The Structure of Scientific Revolutions* by Thomas S. Kuhn (1962) who (despite being criticized on various grounds) created an intellectual space for STS analyses of the social constitution of scientific knowledge. Kuhn essentially made an argument against the linear and cumulative growth of scientific knowledge and illustrated with various cases from the history of science, how scientific development is characterized by (rarely occurring) shifts in scientific paradigms ('scientific revolutions'), which are shared within a scientific community and incommensurable with each other. By far the largest part of scientific activity ('normal science') is engaged in problem solving within the framework of an existing paradigm. Seen from an STS perspective, one can distinguish between a more radical and a more conservative interpretation of Kuhn's work (Pinch, 1997). This is particularly visible in the use of the concept of paradigm either as a description of a shared social activity or social network in science (similar to an 'invisible college') distinct from the cognitive dimension of science, or as a term emphasizing the combined sociocognitive nature of science, i.e., science understood as doing and practice. Kuhn's work also laid the ground for a rediscovery of Ludwik Fleck's (1979[1935]) book on *The Genesis and Development of a Scientific Fact* where Polish medical doctor Fleck developed his concept of 'thought collectives' and 'thought style' and a concept of scientific truth relative to these social collectives, which anticipated much of STS and social constructivism in a more contemporary way than Kuhn's writings, which took most of the credit.

The Social Impact of Technology

While the sections above sketched out different lines of thought about science as a social phenomenon, sociological and philosophical engagement with technology and artifacts took place rather separate from these approaches. The main thrust of theses approaches was to understand the impact of modern technology on society and – going beyond a one-sided relationship – the interdependence of changes in social structure (or: civilization) and technological development. Lewis Mumford and Jacques Ellul can exemplarily be mentioned in this context. In *The Myth of the Machine* Mumford (1967) develops a universal history of civilization with the development of order as the pervasive formative principle. As he points out with examples of megamachines such as pyramids, but also polytechnic traditions of medieval times, the 'idea' of the machine is deeply entrenched in the history and development of civilization, in its organization but also its ethics and aesthetics. For Jacques Ellul too, human history and condition is intricately interwoven with technology. However, where Mumford puts some hope in alternative (democratic) technological traditions, Ellul draws a more pessimistic picture of technology taking control of modern society. And similar to Mumford, 'la technique' is more than just machine technology – it is any complex of standardized means to attain predetermined results. "Technique integrates the machine into society. It constructs the kind of world the machine needs and introduces order (…), clarifies, arranges, and rationalizes. (…) It is efficient and brings efficiency to everything" (Ellul, 1964: 5).

Undoubtedly, various other lines of inquiry can be identified, which prepared the ground for understanding artifacts as being deeply intertwined with social, cultural, and political phenomena. Important inputs have, e.g., come from the sociology of industrial organizations (e.g., studies of the industrial

workplace and its interrelations with technological change); the economics of innovation and technological change (such as the work of Joseph Schumpeter and its later uptake in theories of evolutionary economics); or early sociological analyses of the social and cultural factors that shape technical invention (e.g., Gilfillan, 1935).

The Emergence of a New Field

Only in the late 1970s and early 1980s, a new field of scientific inquiry took shape, which later came to be known as 'STS' – though terms such as 'Social Studies of Science and Technology' or 'Science, Technology, and Society' existed in parallel and are sometimes still used. The common focus of early STS was to gain an understanding of the 'social fabrication' of scientific facts – be it as a product of discourses, social interests, or the heterogeneous social practices involved in producing scientific knowledge. Much of the ground for this new line of inquiry was prepared by Kuhn's writings and a 'linguistic turn' inspired by Ludwig Wittgenstein's philosophy. Two sets of approaches stood at the center of early STS: the sociology of scientific knowledge (SSK) and ethnographic studies of scientific knowledge production in laboratories. In both cases, the focus was on natural sciences, which represented the 'hardest possible case' to show how scientific knowledge was imbued with social and cultural influences and did not constitute a separate, asocial, and ahistoric realm of positive knowledge.

The Sociology of Scientific Knowledge

Two UK 'schools' stood at the center of early debates about the social construction of scientific knowledge: a group of researches at the University of Edinburgh (David Bloor, Barry Barnes, Donald MacKenzie, Steven Shapin, and others) developing the 'strong program' as well as interest theory of scientific knowledge, and a group at the University of Bath (Harry Collins, Trevor Pinch, and others) with their 'empirical programme of relativism' (EPOR) and a focus on the analysis of discourses and scientific controversies. The 'strong program' was based on several methodological claims for a more symmetric analysis of scientific knowledge generation – meaning that empirical analysis should be ignorant to whether a scientific claim was true or false (contrary to the philosophy of science approach, which is just about ways of evaluating the truth or falsity of such beliefs). The aim thus was not to methodologically decide on scientific truth claims, but to understand how they emerged, were accepted or rejected in social interactions. Historical case studies, moreover, elucidated how decisions about scientific beliefs and knowledge claims were influenced by wider social interests ('interest theory'), e.g., how statistical concepts originating in early twentieth-century England were ultimately grounded in debates about eugenics (MacKenzie, 1978). Instead of searching for such 'external social factors' the Bath school with their 'EPOR' along with scholars such as Michael Mulkay rather focused on processes of inner-scientific discourses and controversies to understand how interpretations of scientific facts were developed, rejected, or stabilized.

Laboratory Studies

Along with the SSK, another more ethnographically inspired field of research developed around the analysis of knowledge production in scientific laboratories. Key representatives of this approach were Bruno Latour, Steve Woolgar, Karin Knorr-Cetina, and Michael Lynch. Not 'science as knowledge' was at the center of these studies, but 'science as a localized practice,' which was analyzed through participant observation and the analysis of documents and discourses at the locus of knowledge production, the scientific laboratory. As Knorr-Cetina (1995) points out, these studies showed how scientific objects were not only 'technically' manufactured in laboratories but also symbolically and politically construed. Scientific knowledge thus is not only embedded in controversies and wider social processes, but also socially fabricated in concrete locales with their specific knowledge cultures and idiosyncrasies. Particularly, Bruno Latour and Steve Woolgar in their study *Laboratory Life* (Latour and Woolgar, 1979) do not focus on concepts of intentionality, but rather on the detailed description of the gestures and action observed in scientific work (Pestre, 2004). Scientific facts are not correspondents of observation, textual representation, and things, but gain 'reality' only through series or networks of translations that can each be contested and renegotiated. Particularly, Latour's study on Louis Pasteur, 'Give me a laboratory and I will raise the world' (Latour, 1983), makes clear, how laboratory practices transcend the boundaries of the laboratory and modify social order and eventually transform the world.

The Technological Turn in Science Studies

During the 1980s, the program of understanding the social fabrication of scientific facts was extended to the study of technology. Central to these concepts is that there are potential choices (not necessarily conscious and manifest ones) inherent in the design of artifacts and technical systems. There is no inherent and compelling logic of technical development; the choices made depend on organizational, political, or economic factors and on actor-strategies.

The Social Construction of Technology

In their article 'The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other,' Trevor Pinch and Wiebe Bijker (1984) apply the methodological principles and strategies of the 'EPOR' to the study of technological innovation. Their aim is to understand technology 'from inside,' as a body of knowledge and as a social system. This may be achieved (following EPOR) by a symmetric and impartial treatment of technical artifacts without an external reference in deciding why one technology works and another one does not. At the basis is the concept of interpretative flexibility, meaning that technologies are open to more than one interpretation by different relevant social groups; the existence of social mechanisms that lead to a closure of the controversy at a second stage, and finally the relation of such closure mechanisms to 'the wider context,' which they call the technological frame to contrast the

contingency and to an extent arbitrariness of interpretative flexibility. Along with the historic analyses of the messy and nonlinear ways of developing artifacts inspired by the Social Construction of Technology approach, another historian of technology gained an important influence on the social understanding of technology. In his seminal work *Networks of Power* Thomas P. Hughes (1983) demonstrated the fundamental inseparability of the social and the technical in the analysis of artifacts in his historic analysis of the development of the electricity system in the US, the UK, and Germany as an example of other 'large technical systems' and infrastructure technologies. Efforts of system building as a heterogeneous social, economic, and technical ability stood at the center of this analysis, but also the momentum and obduracy gained as such large sociotechnical ensembles mature and become increasingly entrenched in society and material structures.

Actor-Network Theory

A second direct line from science to technology studies is established by actor-network theory (ANT) – closely linked to the names of Bruno Latour, Michel Callon, and John Law – who already studied science as a practice of creating heterogeneous networks of actors, inscriptions, theories as well as scientific instruments and other artifacts and accordingly put forward the notion of 'technoscience' to stress the increasing convergence of producing scientific knowledge and technical objects. ANT follows the consequences of the 'new' social studies of science approach, especially its transgression of the separation of an internalist and externalist analysis of science to its radical end. If the social realm of subjects and the realm of scientific facts or things is not separable and reducible to each other any more, if neither nature nor society are objective arbiters, social theory has to be rebuild and conceptualized in a way that simultaneously includes humans and things. This results in a generalized principle of symmetry between human and nonhuman actants (as a term for anything that has force to change things), famously put forward in Michel Callon's (1986) article on the scallops of St. Brieuc bay, which are as much granted agency as the scientists or fishermen struggling with a solution to cultivate such mussels. The stability of human social assemblage is always shared with the nonhumans mobilized to this end. The same holds true the other way round: what we call things or scientific facts are also highly impregnated with sociality and need constant work to stabilize – as Latour once called it: technology is society made durable. Central terms employed to explain the eventual constitution of these heterogeneous assemblages are 'translation' signifying an active element in the network connections as an enrollment into an actor-network always involves displacements and changes in the identity of actors; or 'heterogeneous engineering' as the tactics of network-building, which depends on the interrelation of a range of disparate elements of varying degrees of malleability. Though frequently at the center of controversies within science studies or criticized from outside (e.g., in the so-called 'science wars,' pitching so-called 'postmodernists' against 'scientific realists' or constructivism against objectivity in science), ANT still is at the center of much of the research carried out in STS and has not lost its provocative power.

Toward an Engaged Program

Building on and extending these core approaches of STS and crossing over with other theories and research fields such as cultural studies, history or gender studies, STS has developed into a versatile and heterogeneous research field. Many of these subfields and specialties, however, have in common that they share an engagement with critically analyzing and deconstructing dominant social discourses and power structures enacted in sociomaterial relationships and reflect on the possibility of strategies and policies toward a more democratic, sustainable, or gender-equal society informed by such insights. The following section highlights a selection of these engaged approaches in STS.

Feminist STS

STS have had a productive relationship with feminist theory and perspectives since its beginning. Uncovering gender relations in science and technology and analyzing the role of science in legitimizing social forms of domination largely parallels the STS program of understanding the social construction of the content of science and technology. Feminist standpoint theory (e.g., Harding, 1991) claims that all attempts to know are socially situated and that the particular social position of women and their distinctive experiences of discrimination is not reflected in a male-dominated science. The social location of women and other oppressed can be an important source to increase objectivity in science. Donna Haraway takes this perspective further, particularly, in her seminal essay 'Situated knowledges: the science question in feminism and the privilege of partial perspective' (1991), where she develops a more dynamic and hybrid epistemology granting privilege to partial perspectives, which takes the position of the knower seriously. In her notion of situated knowledges, Haraway problematizes both subject and object and aims at thinking the world in terms of the 'apparatus of bodily production,' which brings her work in close interrelation with ANT and other STS approaches. Feminist positions have also been influential for the discussion of the social shaping of technology. As Wajcman (1991) points out, technology itself is imbued with male power and male interests, and technologies reveal a lot about the societies within which they are invented and used, about dominant images of social status and distributive justice, they currently reflect a male-dominated world.

From Epistemology to Ontology?

A recent issue of the journal Social Studies of Science asks whether one can speak of an 'ontological turn' in STS (Woolgar and Lezaun, 2013). In fact, important lines of discussion within feminist STS such as Haraway's above-mentioned 'situated knowledges' can be understood as a shift in analysis from different modes of knowing to different modes of being. Within this perspective, objects do not acquire a particular meaning through a given context, as Woolgar and Lezaun point out, but they are 'enacted in practices,' they are brought into being, or realized, in the course of a certain practical activity. What we perceive as things or social phenomena are not just

intertwined but constantly 'in the making' and the (temporarily stable) outcome of social and material processes. STS research does not just observe disembedded things, but performances, 'realities enacted into being,' as John Law puts it. Objects become ambivalent and fluid and various modes of enactment may bring about a multiplicity of objects, as, e.g., Annemarie Mol (2003) argues in *The Body Multiple* with respect to medical practices. In a similar vein, Karen Barad is concerned with the constitution of material objects in her concept of 'agential realism,' where materiality is understood as a practical achievement, a contingent upshot of practices and produced through intra-action in an apparatus consisting of bodies, norms, technologies, and other sociomaterial elements. If however, as Mol (2013) points out, realities are adaptive and multiple, if they take different shapes as they engage, and are engaged, in different relations, then questions of ontological politics become important and STS may create new normative concerns with the making-up of alternative worlds (see also Latour, 2004).

New Modes of Knowledge Production

During the past decades, STS have also increasingly turned to analyzing the changing role of science in economy, policy, and society. During the 1990s, a new phase of science-related policy has taken shape, where "basic research has become intimately intertwined with production of goods and technological development of relevance for all realms of society" (Elzinga, 1997). At the same time parts of science have undergone substantial transformations – addressed, e.g., with the notion 'mode 2' science (Gibbons et al., 1994) – characterized by a diversification of the locations of scientific knowledge production, an orientation toward problems that cannot be 'solved' by scientific research alone, require the participation of actors outside academia and are situated in a context of high uncertainty, and by demands for more social accountability and reflexivity.

STS interest has shifted accordingly to issues such as credibility (whose knowledge-claims are accepted as trustworthy?), expertise, the relation of science and policy, and the role of the wider public and civil society in science and technology. Science has become increasingly contested and at the same time plays an ever more important role in policy, court rooms the media or public controversies about various types of problems. STS analyses various sites of such contestation, such as the role of science in regulation and judiciary processes (e.g., Jasanoff, 1990), the influence AIDS activists got on the design of clinical trials for new AIDS drugs (Epstein, 1996), or patient groups on pharmacological research (Callon and Rabeharisoa, 2003).

This potential conflict between democracy and expertise is also discussed in terms of a 'problem of extension' by Collins and Evans (2002), i.e., the question who should legitimately participate in technical decision-making (see also Sismondo, 2010). Legitimate expertise extends beyond the provision of scientific knowledge and encompasses other forms such as contributory or interactional expertise, which facilitates meaningful interaction between experts and nonscientific groups. As Sheila Jasanoff highlights with her concept of 'civic epistemologies' (Jasanoff, 2005) such science–policy interactions crucially depend on, and interact with, political culture and must conform to established ways of 'public knowing' in order to gain broad-based support.

Moreover, STS approaches have increasingly become involved in studying and understanding various other contemporary social and political issues and challenges, such as climate change and the transition of systems of mobility, energy, or agriculture toward greater sustainability, the functioning of financial markets and wider processes of marketization and valuation, or the governance of innovation and technological change to name but a few of the pathways an engaged program of STS is currently taking.

See also: Actor-Network Theory; Culture and Actor Network Theory; Gender and Technology: From Exclusion to Inclusion?; Reflexivity in Science and Technology Studies; Science and Technology Studies, Ethnomethodology of; Science, Sociology of; Scientific Knowledge, Sociology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Social Constructionism; Technology, Social Construction of.

Bibliography

Barnes, B., 2007. Catching up with Robert Merton: scientific collectives as status groups. Journal of Classical Sociology 7 (2), 179–192.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Law, J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge? Routledge and Kegan Paul, London, pp. 196–233.

Callon, M., Rabeharisoa, V., 2003. Research "in the wild" and the shaping of new social identities. Technology in Society 25, 193–204.

Collins, H.M., Evans, R., 2002. The third wave of science studies: studies of expertise and experience. Social Studies of Science 32 (2), 235–296.

Ellul, J., 1964. The Technological Society. Vintage Books, New York.

Elzinga, A., 1997. The science-society contract in historical transformation: with special reference to "epistemic drift". Social Science Information 36 (3), 411–445.

Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley.

Fleck, L., 1979[1935]. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.

Fuller, S., 2007. New Frontiers in Science and Technology Studies. Polity, Cambridge.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies. Sage, London.

Gilfillan, S.C., 1935. The Sociology of Invention; an Essay in the Social Causes of Technic Invention and Some of its Social Results. Follett Publishing Company, Chicago.

Haraway, D., 1991. Situated knowledges: the science question in feminism and the privilege of partial perspective. In: Haraway, D. (Ed.), Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York, pp. 183–276.

Harding, S.G., 1991. Whose Science? Whose Knowledge?: Thinking from Women's Lives. Cornell University Press, Ithaca, NY.

Heintz, B., 1993. Wissenschaft im Kontext. Neuere Entwicklungstendenzen der Wissenschaftssoziologie. Kölner Zeitschrift für Soziologie und Sozialpsychologie 45, 528–552.

Hess, D.J., 1997. Science Studies: An Advanced Introduction. NYU Press, New York; London.

Hughes, T.P., 1983. Networks of Power. Electrification in Western Societies 1880–1930. Johns Hopkins University Press, Baltimore.

Jasanoff, S., 1990. The Fifth Branch: Science Advisers as Policymakers. Harvard University Press, Cambridge, MA.

Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton.

Jasanoff, S., 2010. A field of its own: the emergence of science and technology studies. In: Frodeman, R., Thompson Klein, J., Mitcham, C. (Eds.), The Oxford Handbook of Interdisciplinarity. Oxford University Press, Oxford, pp. 191–205.

Knorr-Cetina, K., 1995. Laboratory studies. The cultural approach to the study of science. In: Jasanoff, S., Markle, G.E., Petersen, J.C., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage Publications Inc, USA, pp. 140–166.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago; London.

Latour, B., 1983. Give me a laboratory and I will raise the world. In: Knorr-Cetina, K.D., Mulkay, M. (Eds.), Science Observed: Perspectives on the Social Study of Science. Sage, London, pp. 141–170.

Latour, B., 2004. Why has critique run out of steam? From matters of fact to matters of concern. Critical Inquiry 30, 225–248.

Latour, B., Woolgar, S., 1979. Laboratory Life. The Construction of Scientific Facts. Princeton University Press, Princeton.

MacKenzie, D., 1978. Statistical theory and social interests: a case-study. Social Studies of Science 8, 35–83.

Mannheim, K., 1929. Ideologie und Utopie. F. Cohen, (Bonn).

Merton, R.K., 1973[1942]. The Normative structure of science. In: Merton, R.K., Storer, N.W. (Eds.), The Sociology of Science: Theoretical and Empirical Investigations. The University of Chicago Press, Chicago, IL, pp. 267–278.

Mol, A., 2003. The Body Multiple: Ontology in Medical Practice. Duke University Press, Durham, NC; London.

Mol, A., 2013. Mind your plate! the ontonorms of Dutch dieting. Social Studies of Science 43 (3), 379–396.

Mumford, L., 1967. The Myth of the Machine. Technics and Human Development. Secker & Warburg, London.

Pestre, D., 2004. Thirty years of science studies: knowledge, society and the political. History & Technology 20 (4), 351–369.

Pinch, T.J., 1997. Kuhn – the conservative and radical interpretations: are some mertonians 'Kuhnians' and some Kuhnians 'Mertonians'? Social Studies of Science 27 (3), 465–482.

Pinch, T.J., Bijker, W.E., 1984. The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science 14 (3), 399–441.

Sismondo, S., 2008. Science and technology studies and an engaged program. In: Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies. MIT Press, London, pp. 13–31.

Sismondo, S., 2010. An Introduction to Science and Technology Studies. Wiley-Blackwell, Chichester.

Wajcman, J., 1991. Feminism Confronts Technology. Polity Press, Cambridge, UK.

Woolgar, S., 2004. What happened to provocation in science and technology studies? History and Technology: An International Journal 20 (4), 339–349.

Woolgar, S., Lezaun, J., 2013. The wrong bin bag: a turn to ontology in science and technology studies? Social Studies of Science 43 (3), 321–340.

Yearley, S., 2005. Making Sense of Science: Understanding the Social Study of Science. SAGE, London.

Zuckerman, H., 1988. The sociology of science. In: Smelser, N. (Ed.), Handbook of Sociology. Sage, Newbury Park, CA, pp. 511–574.

Science and Technology Studies: Experts and Expertise

Brian Wynne, University of Lancaster, Lancaster, UK
Michael Lynch, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by O. Renn, volume 20, pp. 13647–13654, © 2001, Elsevier Ltd.

Abstract

The problem of how to integrate expert knowledge with democratic politics has preoccupied scholars for more than a century but has become highly salient for Science and Technology Studies (STS) scholarship in recent decades. Political controversies over climate change, genetically modified organisms, and many other topics involve factual claims and counterclaims that invoke expert – and often scientific – authority. Much of the research in recent years on public engagement with science emphasizes the fallibility of expert knowledge and the hazards for democracy when scientific authority is overextended. At the same time, some scholars see a need for STS scholarship to develop a basis for distinguishing authentic expertise from non- and pseudo-expertise. Lively debates over these matters call attention to the extent to which public deliberation is adequate to the task of contending with, seemingly, technical matters. A key issue remains – how to define a purely technical, expert question, in public issues?

Introduction

Expert and expertise are commonplace words, but who counts as an expert and what counts as expertise (along with the often-neglected question of which issues require expert input) can be highly contested topics in many technical, political, legal, and ordinary situations. Expert claims themselves are frequently, perhaps even typically, contested, which itself generates questions over who is a legitimate expert. Some key institutional arenas, such as courts in the USA and UK though not continental Europe, use legal proceedings that depend upon formal adversary mechanisms for expert testimony. In many other social and political arenas, expert disagreement is often lamented as a cause of paralysis or uncertainty, where decisive commitment is sought. In some national systems, expert advice for policy is organized so as to be consensual, thus disallowing expert disagreements or at least, relegating divergent expert views to unofficial, 'critical science' status. Designated experts have important roles in law and governance, and the issue of how scientific expertise is defined and used in legal and governmental settings is a lively area of current Science and Technology Studies (STS) research, one with 40 or so years of history (Nelkin, 1979; Wynne, 2010 [1982]). This large and dynamic field has also studied, in depth, the varying reasons for expert controversy in countless substantive fields from climate change to contraceptives or genetically modified (GM) crops and foods. Review-based studies from STS (Cambrosio and Limoges, 1991) have noted that expert *disagreement* is natural and that expert consensus is a social achievement that needs explanation. This contradicts the conventional wisdom that disagreement signifies pathology requiring the assumed mistaken or hidden-interest 'side' to be 'exposed,' which would then reveal the true basis of 'natural' expert-revealed consensus.

As is the case with the word 'scientist,' the word 'expert' in many social contexts is used both as a descriptive category and as a term of praise. Once attained, expert credentials can launch lucrative careers and confer respect and influence. In important areas of law, politics, and everyday life, experts are granted special authority. Especially when paired with the descriptor 'scientific,' expertise in one field can sometimes be an advantage for being granted presumptive influence and authority in others. Indeed, as a general matter, the relevant question of 'Expertise *for what*?' (and further: 'Who has defined this question?') is all-too-rarely deliberated. In some contexts in which experts are called upon to give advice, their credibility is allowed to extend broadly: a scientist competent in a given area – especially, a Nobel laureate – may be treated as an authority on a wider range of basic scientific questions, including questions that hold social, ethical, political, and legal implications. In the case of a Pauling, an Einstein, or a Hawking, the person's reputation may transcend mere expertise and take on the larger-than-life, indiscriminate proportions of a genius (Mialet, 2012).

Expertise, nevertheless, is always circumscribed and more so than is normally recognized and examined. As a result, authority – for example, the authority to issue permits to release materials into the environment and thus, society – may unquestioningly be assigned where it is contested. Few, if any, recognized experts would profess to be expert in every subject and there are many cases of notable scientists who fail to transfer their authority from one field or topic to another. In some cases they have drawn criticism and ridicule for overstepping the limits of their expertise, such as when William Shockley, a physicist of materials who received a Nobel Prize for his part in the development of the transistor, made conspicuous efforts to promote eugenic programs and became notorious for helping to set up a Nobel laureate sperm bank.

This Shockley example of what we might call 'expert overextension' is typical of many others in contemporary public affairs. In general, there is probably too little open and inclusive debate as to what kinds of questions a given policy issue needs to have specialist answers to, and what kinds of experts – practical or formally certified – are most appropriate for addressing them. Such questions also require clarity about whether scientists or any other experts may be in a position to answer them (Stirling, 2010). In many domains of public policy, such as in risk assessment for regulatory decisions about new technologies, the disciplinary composition of officially

designated expert advisory committees to government has been much narrower than the public policy issue in question merits. This happens, for example, when expert advisory committees for chemical pesticides are composed of laboratory toxicologists, with little or no scientific expertise in ecology and no practical farming expertise at all. Irwin (1994) examined one such case of risk assessment and the UK policy decision on the herbicide 2,4,5-T in the 1980s. In that case, a farmers' representative described the scientific expertise to which the government had granted authority, as "cloud-cuckoo land, behind the laboratory bench" (see also Wynne, 1989). The European Environmental Agency (Horremous et al., 2002) published a collection of similar cases, with the formal proposal that such expert advice be made more systematically inclusive in order to avoid neglect of key specialist knowledge, including relevant practical nonscientific expertise such as in actual uses of pesticides.

In policy and media circles as well as in science, the representation of expertise is often rather wider than it is accurate in its attributions of authority. By contrast, as discussed below, legal procedures are often ruthlessly aggressive and quite indiscriminate in their deconstruction of the grounds of expert credibility (Oteri et al., 1982; Jasanoff, 1995).

Professions and Expertise

To acquire a reputation for being an expert in some areas of conduct, a person need not acquire professional status or acquire formal credentials, but many professions, recreational activities, and other pursuits institutionalize tests and formal certifications to mark degrees of expertise (Grand Master in chess, Black Belt in karate, professional law degrees and bar exams, etc.). While professional status does not necessarily mark a person as an expert, expertise tends to come with the territory in some professions, such as those of the physician, scientist, certified public accountant, fingerprint examiner, and so on. Certified practitioners can lose their credentials for incompetence and malpractice or for failing to maintain and renew them. Credentials themselves are subject to scrutiny, since a person can be defrocked as an imposter when discovered to have faked the necessary credentials, even if they apparently performed competently in the profession. In some circumstances such as cross-examination in adversary trials, challenging questions are raised as to whether a person who professes expertise really is an expert; but in many circumstances such expertise is taken for granted.

In order for membership in a profession to confer expertise, the activity itself must have credibility. Some credentialed activities (such as driving a motor vehicle after passing a test and obtaining a license) are deemed too ordinary to count as domains of expertise in societies with high automobile ownership. A driver cannot claim to be an expert, except in specialized or exceptional cases, such as professional racing, driving highly specialized vehicles, or training others to drive. Some entire fields of activity lack credibility in particular historical, social, or cultural contexts, regardless of a member's attainments within such fields. A person with a reputation as an experienced and highly capable astrologer is unlikely to be regarded publicly as an expert, except in social and historical contexts in which an astrologer's knowledge is regarded as credible. There also can be occasions in which an experienced astrologer's knowledge is sought even by those who do not believe in astrology, for example, in an investigation of common practices among astrologers. The same can be said for 'alternative' health practitioners such as homeopaths, in most contexts.

Expert status contrasts, on the one hand, with the inexpert standing in the respective domain of activity attributed to amateurs, novices, or others who may have credentials but are deemed not to measure up. On the other hand, expert status contrasts with the 'lay' status of those who simply lack knowledge of, and training in, the field – often because they have no interest in pursuing it. Whether or not they have had any prior experience with, or interest in, the relevant fields, laypersons frequently encounter problems in deciding whether to accept expert advice – advice from a physician about a surgical operation; advice from an investment broker about whether to invest in a particular stock; advice from an auto mechanic about an expensive repair; advice from experts delivered through the media about the health effects of various foods and medicines; the potential consequences of (not) entering into one or another military engagement; and the reality of, or the best responses to, human-induced global climate change.

In popular usage, the term 'technocrat' often refers to a leader or style of leadership akin to that of a bureaucrat: a managerial style that avoids bold, visionary, and risky decisions in favor of competent, tedious, and rational attention to precision and detail. In political theory, technocracy has a more systemic denotation as a name for a pervasive system of governance in which experts rule indirectly by feeding technical advice and information to titular heads of agencies and other managerial authorities who are not in any position to question the technicalities (Galbraith, 1967; Winner, 1978), even when these authorities also harbor political commitments (see Jasanoff, 2004). For example, government officials rely upon numerous forms of expert assessment: environmental impact statements; health and environmental risk assessments; cost-benefit analyses of proposed legislation or regulation; assessments of the effectiveness of a proposed weapons system; estimates of increments of atmospheric carbon contributed by existing coal-burning power plants. Although no single expert or group of experts controls government decisions, in a technocracy, expert advice carries increasing weight in all significant decisions; and political meanings and issues may be framed in ways that, by default, give scientific expertise more power than that of its purely technical warrant alone.

In the past two decades, another political concern about expertise has become prevalent in the United States and also in various other national and international arenas. This is a concern that, far from having too much influence over governance, well-informed technical experts have too little. This is particularly evident in contentions over anthropogenic climate change (Hulme, 2009) but similar concerns have been raised about many other public issues, such as school biology curricula, especially concerning the theory of evolution (Latour, 2004).

The Expert as a Legal Category

Experts in many different fields are frequently called upon to advise courts and governments (Smith and Wynne, 1989; Jasanoff, 1990, 1995; Risinger, 2001). Although many experts have scientific credentials, others are deemed to be expert by virtue of professional or experiential knowledge that can inform a court about relevant matters that are not part of common experience. In the United States and many other nations, an expert witness is permitted to testify about relevant knowledge acquired through education, reading, scholarship, and other sources of training and experience – rather than directly witnessed 'factual' matters. Another way of putting this is that an expert witness is permitted to testify about inferences as well as directly observed facts (though legal scholars recognize that it is difficult to draw the distinction in practice) and to testify to collective, professionally negotiated, and agreed-upon disciplinary expert 'truth.' A witness with the relevant professional credentials is permitted to give a trained opinion about such matters as the psychological or neurological competency of the defendant or the cause of a vehicular or other accident. An expert also can testify about specific items of evidence, for example, to declare that a particular suspect's fingerprints match prints collected from a crime scene. This has, more recently, also encompassed DNA evidence (Lynch et al., 2008). An expert also may be called upon to give an opinion about a profession, either from personal experience or academic research; to describe accepted procedures and methods of certification; or to cite statistical error rates for particular procedures. The normal state of knowledge or practice for a given scientific field at a given time, is also a matter of required expert testimony, where for example, negligence may be at issue and could be disproven if it were shown that the actions of (for example) a civil engineering designer-contractor at a given time, were in accordance with prevailing standard practice or scientific knowledge.

Representatives of a particular occupation or trade may also be asked to testify about matters that are collateral to their technical training: psychiatrists typically are asked to assess the future 'dangerousness' of a defendant (a legal category that is not specifically included in psychiatric nosology); a surgeon may be asked to comment on whether or not a victim's wounds would have been the cause of death; a dentist may be asked to advise a court about whether a bite mark on a victim matches the teeth of a suspect. In a classic law review article, Justice Learned Hand (1901: 46) mentions a historic case in which battle-tested sailors as well as physicians were asked for their opinions on whether a deceased person could possibly have drowned if the body had no water in the lungs. Scholars, including historians and STS scholars, occasionally testify about such matters as the history of fingerprint expertise, the scientific status of creation science and intelligent design, and whether or not tobacco smoking was generally known to be hazardous to health prior to recent decades when tobacco companies were promoting tobacco-smoking.

The problem with how to handle expert testimony in jury trials has preoccupied trial courts in the United States and elsewhere for the past century. This has included fraud trials, where the required financial expertise is extraordinarily esoteric. Although rules of evidence allow judges to summon experts to testify, usually they appear on behalf of one or the other adversary party, though in some European legal systems such adversary mechanisms are not used and experts are summoned by the court. The alignment of expert authority with partisan arguments creates a set of problems in jury trials (and also in bench trials, given the fact that judges generally do not have expertise in all or most of the specialties that feature in expert testimony). One commonly mentioned problem is that of a battle between experts, where juries are faced with choosing between contradictory testimonies but without the technical basis for making an independent choice. Another problem is that one side in an adversary trial often has an advantage over the other by having more resources for hiring expert witnesses with impressive credentials.

Uncertainty about the criteria for admitting expert witnesses in jury trials led to a series of landmark decisions by the US Supreme Court in the 1990s (*Daubert v. Merrell Dow Pharmaceuticals* [1993]; *General Electric v. Joiner* [1997]; and *Kumho Tire v. Carmichael*, [1999]). Admissibility decisions are rulings by trial judges about whether specific types of evidence – in this case expert evidence – should be presented to the jury. In a case in which expert evidence is critical, such as a class action lawsuit against a company for an allegedly unsafe product, a decision not to admit the plaintiffs' evidence can effectively shut down the case with a summary judgment before it goes to a jury. The Supreme Court's rulings in the 1990s drew a massive amount of discussion in law review articles, and also attracted the attention of STS scholars (Solomon and Hackett, 1996; Jasanoff, 1995). In the first and most famous of the cases, the Supreme Court in the *Daubert* decision addressed a question familiar to the philosophy of science: how to distinguish genuine science from pseudoscience or so-called 'junk science' (Huber, 1991). The Court famously listed four 'factors' or 'considerations' (they hesitated to use the word criteria) to guide judicial decisions about admissibility: testability, peer review, error rate, and general acceptance in the relevant field(s). The Court also qualified its recommendations with caveats about the need for judicial discretion. Nevertheless, the justices were criticized for imposing a one-size-fits-all conception of expertise on the immense range of specialties that can find their way into the courtroom (Risinger, 2001; Haack, 2005). The Supreme Court also was criticized for privileging stereotypical features of science, for assessing the admissibility of all forms of expertise, including technical and experiential evidence that makes no claim to scientific status and may be authorized by different means (Edmond, 2002).

Judicial rulings that declare plaintiffs' evidence inadmissible in class-action lawsuits often result in summary judgments that preclude a jury trial. The rationale for such rulings is that a single judge is more competent to decide the merits of such evidence than is a jury – a collective representative of the lay populace. The question of jury competence or jury understanding has been of concern in the legal domain for at least the past century. Edmond and Mercer (1997) usefully connect that issue to more general debates about the public understanding of science. A predominant viewpoint expressed by Learned Hand (1901) has it that juries are ill-equipped to contend with technical evidence and that panels of independent experts should be entrusted to advise juries

about relevant technical facts, so that their evaluations of the 'ultimate issue' (guilt, innocence, liability) will be well-informed. One counter-argument to such suggestions is that judgments about technical facts often are intertwined with more ordinary evaluations. A notorious jury verdict provides a case in point. William Thompson, who both participated as a defense lawyer in the notorious double-murder trial of *California v. Orenthal James Simpson* in 1995 and has written extensively about police uses and abuses of DNA evidence in criminal trials, contested a common view of the not-guilty verdict returned by the jury in the face of a massive amount of (seemingly) incriminating DNA evidence. The common view was that the jury either willfully ignored or was incapable of comprehending the technical evidence adduced by the prosecution. After elaborating upon a series of anomalous features of the blood evidence collected and analyzed by the prosecution which, the defense argued, pointed to evidence of tampering and incompetence, Thompson asserted:

> An examination of the competing theories also belies the claim that the DNA evidence was too complex or arcane for the jury to understand. Whether the jury believed or disbelieved the tampering theory, for example, probably had relatively little to do with the jurors' understanding of scientific arcana and a great deal to do with the jurors' assessment of facts that they were uniquely well positioned to assess …. There were a number of scientific issues in the case, but they had little to do with the fundamental science underlying DNA testing or the details of laboratory procedures for typing DNA, and everything to do with the potential for cross-contamination of samples before they reached the DNA laboratories.
> Thompson, 1996: 841

Thompson referred to a series of material indications that blood samples had been mishandled, and possibly created through fraudulent planting of blood on Simpson's car and gloves. The resulting DNA matches and probability estimates of coincidence matches in the relevant population, were thus placed in mundane contexts of police malpractice instead of the murderous actions attributed to Simpson. The basis for such judgments of the significance of technical evidence thus turned on nontechnical attributions of motive and narrative scenarios. Thompson (ibid.) concludes that it "seems fatuous to claim, without further proof, that the jury was confused about the DNA evidence." The jury's basis for judgment involved a variant of what is sometimes called 'lay expertise.'

Lay Expertise

An important general insight from the analysis of the scientific and policy field of risk assessment by STS scholarship, has been the observation that such risk assessments, claimed to be realistic and valid knowledge for governance of risk situations, are invariably based on incomplete knowledge, with relevant knowledge being excluded and scientific ignorance in play, yet denied (Wynne, 1992; Irwin and Wynne, 1996; Proctor and Shiebinger, 2008). Frequently it has been found that accredited risk experts who are given policy authority are unaware of these inadequacies in their own knowledge and unable (or unwilling) to recognize other bodies of salient knowledge, which, had they been included, might have improved the full knowledge-base (see Horremous et al., 2002; EEA, 2013). Since risks are delivered and suffered in diverse real practical situations, such salient knowledge includes not only scientific knowledge but also the specialist knowledge practised by others involved with the technologies and networks concerned. In a major controversy over HIV-AIDS treatment and research on therapeutic drugs in the United States in the 1980s, Stephen Epstein documented how gay community activists and medical practitioners joined forces to make the case for new testing and research standards and regimes, which reflected their real conditions of already being at-risk or infected and so in need of therapies even when these were not proven to be safe by normal standards. The unique practical expertise of HIV/AIDS sufferers was salient to the development of more effective policies and therapies. It was not that such experience and community-based knowledge needed to be in competition with salient medical research expertise but that the latter needed to recognize and respect – and also include – the specialist knowledge of the former, nonscientific experts. In many cases, gay community activists acquired selective medical knowledge that enabled them to more effectively collaborate with medical researchers.

In his study of the responses of the highly specialist, practical-knowledge-based, hill sheep farmers of Cumbria, Northern England, to the scientific expertise invoked to try to manage the radioactive fallout from the 1986 Chernobyl nuclear accident, Wynne (1996) described this practice-based specialist farming and related knowledge-culture as an example of 'lay expertise.' Collins and Evans (2002, 2007) took exception to this term and proposed an elaborate classification of several different types of expertise, including their notion of 'contributory expertise' as distinct from 'interactional expertise,' where contributory expertise could include scientific or practical, experience-based expertise in the form of contributing to technical knowledge of the domain at issue. Interactional expertise in any given domain is a form of competence in the specialist language but not in the specialist practices, of knowledge-development: someone who can 'talk the talk,' as it were, but not 'walk the walk.'

The differences between Wynne's approach to what he called lay expertise of the sheep-farmers who interacted with the radioactivity scientists and Collins and Evans's (2002) partly critical interpretation of his case study, have little to do with how expertise is defined, and more to do with Collins and Evans' assumption that the different knowledge-cultures of the farmers and scientists were both devoted to solving the same issue or problem, and that this is a general state of affairs in such interactions. As ethnographic field-researcher in this case, Wynne disagreed, asserting that the primary meanings given to the issue and thus the problems they saw themselves as managing, were different between farmers and scientists; thus such analytical cross-cultural unification does not make sense as an account of reality as it occurred. This question as to whether experts of different kinds are or should be providing only expertise in answering specific, shared, propositional questions ('Is it safe?' for example) or whether they are also defining the meanings of such situations ('What is the issue in question?') is often an

unstated and subtle one (see Wynne, 2003; also Jasanoff, 2003). Many other examples of expertise based on noncertified but impressively durable experience-based specialist knowledge in diverse domains, abound in the literature.

Interactional versus Contributory Expertise

As alluded to in the foregoing section, Collins and Evans (2007) made a strong point of distinguishing between what they called 'contributory' and 'interactional' expertise. While contributory expertise (as the words suggest) contributes to the salient body of specialist expert knowledge, and so is based in authentic practical and cognitive skills exclusive to that expert collective, interactional expertise is qualitatively different. Thus, while the latter form of expert is capable of communicating in the specialist language with contributory experts and interacting as if one of them, such expertise is not capable of actually contributing to the salient corpus of knowledge, for example by designing, conducting, and analyzing productive experimental tests and publishing findings. For Collins and Evans this was a key distinction. Others found it basically valid in principle but of questionable analytical worth, unless one wanted, as Collins and Evans evidently did, to use such identifications to distinguish between participants in a public issue, to decide who is making statements based in real expertise as opposed to interactional expertise or virtually no expertise at all. This corresponds with the concern that Collins and Evans have also expressed over 'the problem of extension,' where in their view, anyone who participates in a public issue tends to claim, or be allocated, equal expertise with designated experts, thus confusing public debate and decision-making to the point of paralysis. Others have argued that a far more significant problem than such over-extension, if it does occur, is a danger of confusing differences in factual knowledge with different orientations to 'matters of concern' between social groups (Latour, 2004). An exclusive attention to expertise with propositional (or factual) issues to the exclusion of differences of meaning and concern about legitimately political issues, threatens to bring us back to questions of technocracy, as discussed before.

Conclusion

The politics of expertise is by no means a new topic but it has become especially salient in STS. In the literature, there are different views on the problem of expertise. Many in STS, as part of the legacy of constructivism, take a skeptical view of expert claims in favor of more widespread public engagement with (seemingly) technical decisions about public health, safety, security, and environmental matters. Increasingly, STS scholars also recognize the need to discriminate between different domains of expertise and different political uses of both expert claims and skepticism about such claims (Turner, 2001). Debates about expertise in the field continue and are tied in with more widespread public controversies about climate change, genetically modified organisms, and many other public concerns.

See also: Biocitizenship; Environmental Sciences; Expertise, Anthropological Approaches to; Indigenous Knowledge: Science and Technology Studies; Reflexivity in Science and Technology Studies; Science and Law; Science and Social Movements; Science, Public Engagement with; Science, Sociology of; Science: Constructivist Perspectives, History of; Scientific Controversies; Scientific Evidence: Legal Aspects; Scientific Knowledge, Sociology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Technological Determinism; Truth and Credibility in Science.

Bibliography

Cambrosio, A., Limoges, C., 1991. Controversies as governing processes in technology assessment. Technology Assessment and Strategic Management 3 (4), 377–396.
Collins, H., Evans, R., 2002. The third wave of science studies: studies of expertise and experience. Social Studies of Science 32 (2), 235–296.
Collins, H., Evans, R., 2007. Rethinking Expertise. University of Chicago Press, Chicago.
EEA, 2013. Air Quality in Europe – 2013 Report. European Environment Agency, Denmark.
Edmond, G., 2002. Legal engineering: contested representations of law, science (and non-science) and society. Social Studies of Science 32 (3), 371–412.
Edmond, G., Mercer, D., 1997. Scientific literacy and the jury: reconsidering jury 'competence'. Public Understanding of Science 6, 329–357.
Epstein, S., 1998. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley.
Galbraith, J.K., 1967. The New Industrial State. Princeton University Press, Princeton, NJ.
Haack, S., 2005. Trial and error: the Supreme Court's philosophy of science. American Journal of Public Health 95 (S1), S66–S73.
Hand, B.L., 1901. Considerations regarding expert testimony. Harvard Law Review 15, 40–58.
Horremous, P., Gee, D., MacGarvin, M., Stirling, A., Keys, J., Wynne, B., Vaz, S., 2002. Late Lessons from Early Warnings: The Precautionary Principle in the 20th Century. European Environment Agency, Copenhagen. Also published (2003) by Earthscan, London and New York.
Huber, P., 1991. Galileo's Revenge: Junk Science in the Courtroom. Basic Books, New York.
Hulme, M., 2009. Why We Disagree about Climate Change. Cambridge University Press, London and New York.
Irwin, A., 1994. Citizen Science. Routledge, London.
Irwin, A., Wynne, B. (Eds.), 1996. Misunderstanding Science. Cambridge University Press, London and New York.
Jasanoff, S., 1990. The Fifth Branch: Science Advisors as Policymakers. Harvard University Press, Cambridge.
Jasanoff, S., 1995. Science at the Bar. Harvard University Press, Cambridge, MA.
Jasanoff, S., 2003. Breaking the waves in science studies: comment on H.M. Collins and Robert Evans, 'The third wave of science studies'. Social Studies of Science 33 (3), 389–400.
Jasanoff, S. (Ed.), 2004. States of Knowledge. Routledge, London and New York.
Latour, B., 2004. Why has critique run out of steam? from matters of fact to matters of concern. Critical Inquiry 30 (2), 225–248.
Lynch, M., Cole, S., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.
Mialet, H., 2012. Hawking Incorporated. University of Chicago Press, Chicago.
Nelkin, D. (Ed.), 1979. Controversy: The Politics of Technical Decisions. Sage, Beverley Hills, CA.
Oteri, J.S., Weinberg, M.G., Pinales, M.S., 1982. Cross-examination of chemists in drugs cases. In: Barnes, S.B., Edge, D.O. (Eds.), Science in Context. The Open University Press, Milton Keynes, UK, pp. 250–259.
Procter, R., Shiebinger, L. (Eds.), 2008. Agnotology: The Making and Unmaking of Ignorance. Stanford University Press, Stanford.
Risinger, D.M., 2001. Preliminary thoughts on a functional taxonomy of expertise in a post-Kumho world. Seton Hall Law Review 31, 508–536.
Smith, R., Wynne, B. (Eds.), 1989. Expert Evidence: Interpreting Science in the Law. Routledge, Chapman & Hall, London.

Solomon, S., Hackett, E., 1996. Setting boundaries between science and law: Lessons from Daubert v. Merrell Dow Pharmaceuticals, Inc. Science, Technology, & Human Values 21, 131–156.

Stirling, A., 2010. Keep it complex. Nature 468 (23/30th December), 1029–1032.

Thompson, W.C., 1996. DNA evidence in the O.J. Simpson trial. University of Colorado Law Review 67, 827–857.

Turner, S., 2001. What is the problem with experts? Social Studies of Science 31 (1), 123–149.

Winner, L., 1978. Autonomous Technology: Technics-out-of-control as a Theme in Political Thought. MIT Press, Cambridge, MA.

Wynne, B., 1989. Frameworks of rationality in risk management: towards the testing of naïve sociology. In: Brown, Jennifer (Ed.), Environmental Threats: Analysis, Perception, Management. Frances Pinter, London, pp. 33–47.

Wynne, B., 1992. Misunderstood misunderstandings: social identities and public; uptake of science. Public Understanding of Science 1 (3), 281–304.

Wynne, B., 1996. May the sheep safely graze? A reflexive view of the expert-lay knowledge divide. In: Lash, S., Szerszynski, B., Wynne, B. (Eds.), Risk, Environment, and Modernity. Towards a New Ecology. Sage, London, pp. 44–83.

Wynne, B., 2003. Seasick on the third wave? Subverting the hegemony of propositionalism: response to Collins & Evans (2002). Social Studies of Science 33 (3), 401–417.

Wynne, B., 2010 [1982]. Rationality and Ritual: The Windscale Inquiry and Nuclear Decisions in Britain. British Society for History of Science, Chalfont St Giles, Bucks, UK (Republished with new Introduction, as: Rationality and Ritual: Participation and Exclusion in Nuclear Decision Making. Earthscan, London and New York).

Science and The State

Yaron Ezrahi, The Hebrew University of Jerusalem, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As a body of knowledge, a host of techniques of investigation, a sociocultural configuration, and an outlook, science was a major factor in the emergence of modernity and the shaping of the modern state. Science and technology constituted a principal resource for the military empowerment of the modern state. In addition, they provided technical instruments, organizational tools, and powerful justifications for the interventions of the modern state in such diverse areas as public health, industry, city planning, agriculture, transportation, and more recently environmental protection. In the course of the nineteenth century, the faith in the capacity of science to objectively describe, explain, and control nature was gradually extended to human affairs as sociology, economics, psychology, and other social sciences appeared to certify scientific accounts of social facts and scientific techniques of intervention. While the natural and the social sciences could thus play an important role in the development and the modes of operation of the modern public bureaucracy, they could at the same time be enlisted in free societies as powerful means of public criticism of the state and its actions. Nevertheless, the increasing role of science and technology in modern military conflicts and economic or industrial competition has undermined the long-standing apolitical status of science and technology. Together with these developments, new challenges to the powers and authority of the modern state posed by the globalization of the economy and mass communications, the transformation of relatively homogeneous into ethnically heterogeneous societies, and the spread of multiculturalism, diminished the capacity of the state to function as an agent of large-scale science-based projects. On the other hand, the increasing role of the market and private agents in the applications of science and technology enhances the responsibilities of the state as a regulator charged with protecting the private citizen and the public against the misuses of sciences and technology and their unintended effects.

Introduction

Since the sixteenth century, the 'scientific revolution' has generated a host of new, intellectual, rhetorical, and institutional strategies for undermining traditional political authorities and constructing alternative ones. Science became a vital resource in the modern attempt to discredit traditional political hierarchies and spiritual transcendental sources of authority in the context of public affairs. By providing new rationales for order compatible with the novel modern commitments to the values of individualism, voluntarism, and egalitarianism, science and technology paradoxically also provided new grounds for novel modern forms of hierarchy and authority committed to the use of knowledge in the reconstruction of society. The widely shared notion that science went along with progress appeared to suggest that knowledge, when applied to public affairs, could in fact depoliticize public discourse and action. This faith produced wide mandates for large-scale social and political engineering and monumental state-sponsored technological projects in both authoritarian and democratic states.

While such ideals and projects were from the very beginning criticized by observers such as Edmund Burke (1729–97), throughout the twentieth century the record of relations between science and politics turned out to be sufficiently mixed to discard earlier hopes. The controversial role of scientists in the production and military deployment of means of mass destruction, the role of scientists in the Nazi experiments on human beings and the endorsement of racism, and their involvement in the industrial pollution of the environment, diminished the public trust in science fostered by such developments as medical breakthroughs and spectacular space flights.

Since the closing decades of the twentieth century, an increasing number of social scientists and historians have been pointing out that new uncertainties in the sphere of politics have been converging with newly recognized uncertainties of science and risks of technology to form novel postmodern configurations of state, science, and society (Eisenstadt, 1999; Beck, 1992). It has become increasingly recognized that, in ethnically and religiously heterogeneous and multicultural societies, science and technology can no longer be attached to uncontroversial comprehensive values that privilege claims of neutrality, objectivity, and rationality; nor can the earlier belief that knowledge can depoliticize public discourse and action be sustained any longer.

The Authorities of Science and The State

The belief that science can provide a secular version of the synoptic divine view of human affairs was politically most significant. Until its erosion toward the end of the twentieth century, the modernist configuration of science and the state was formed over the course of at least four centuries. The secularization of political power in the modern state has put the dilemma of arbitrary rule at the center of modern political theory and practice. Whereas legal and constitutional constraints appeared to be the most appropriate remedy to this problem, the rise of modern science and technology encouraged faith in the possibility of restraint and moderation through public enlightenment and by means of scientific and technical advice. Not surprisingly, the redemptive drive behind the integration of science and politics had its origins in premodern religious visions and values (Manuel, 1974).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85051-5

In both its authoritarian and democratic versions, this modernist program rested on the belief that the modern state can remedy the defects of the present sociopolitical order and realize an ideal order. Science contributed to this modernist outlook a faith in the power of secular knowledge to control nature and reconstruct society. In his dedication to Lorenzo di Medici of Florence in The Prince (1964[1513]), Niccolo Machiavelli implies that the holder of the God's eye view of the entire state is neither the King, who looks down at the people from the top of the hierarchy, nor the people, who see the King at the top from their lower state. It is a person like himself, a man of knowledge. The King, to be sure, has knowledge of the people, and the people have knowledge of the King, but, because of his intellectual perspective, 'a man of humble and obscure condition' like Machiavelli can claim to see both the King and his subjects and understand the nature of their interactions.

Political theory and political science thus claimed to have inherited the God's eye view of the whole polity and evolved a secular vision of the state as an object of knowledge. In turn, the nation-state has often found it useful to adopt the frame, if not always the content, of the inclusive scientific outlook in order to legitimize its interventions in the name of the general common good and its expressions in public goals such as security, health, and economic welfare. If, from the inclusive perspective of the state, all citizens, social groups, or institutions appeared as parts that the supreme power had the capacity to fit together or harmonize, science in its various forms often provided the technical means to rationalize, and the rhetorical strategies to depoliticize, such applications of power.

The potential of science, not only as a source of knowledge but also as a politically useable authority, was already recognized by Thomas Hobbes. This founder of modern political theory was in contact with Francis Bacon, admired William Harvey's findings about the circulation of the blood, and criticized Robert Boyle's position on the vacuum (Shapin and Schaffer, 1985). In his influential works on the state, Hobbes tried to buttress his claims by appealing to the authority of scientific, especially mathematical, certainties, hoping to enlist the special force of a language that claims to generate proofs as distinct from mere opinions (Skinner, 1996). Although he was suspicious of the experimental science of Boyle, Hobbes followed Machiavelli in describing politics in the language of causes rather than motives. Thomas Sprat (1958[1667]), the historian of the Royal Society, claimed that when compared with the contentious languages of religions, the temperate discourse of science could advance consensus across diverse social groups. In both the experimental and mathematical scientific traditions, however, knowledge, whether based on inferences or observations (or on their combination), was expected to end disputes.

The association between science and consensus was regarded as deeply significant in a society that had begun to view authority and order as constructed from the bottom up rather than coming from above (Dumont, 1986). Reinforced by the growing image of science as a cooperative, nonhierarchical, and international enterprise in which free individuals come to uncoerced agreements on the nature of the universe, scientific modes of reasoning and thinking appeared to suggest a model for evolving discipline and order among equals (Polanyi, 1958, 1962).

Deeply compatible with the notion that legitimate power and authority are constituted by social contracts, the rise of modern scientific knowledge appeared to accompany and reinforce the rise of the modern individual and his or her ability to challenge hierarchical forms of knowledge and politics. From Hobbes and Descartes, through Kant, J.S. Mill, and late-modern democratic thinkers such as Popper, Dewey, and Habermas, the commitment to the centrality of individual agency becomes inseparable from a growing faith in the possibility of expanding public enlightenment backed up by the demonstrated success of scientific modes of reasoning and acting.

Power and the Social Sciences

In succeeding centuries, one of the most persistent ambitions of Western society has been to make conclusions touching the 'things' of society, law, and politics appear as compelling and impervious to the fluctuations of mere opinion as conclusions touching the operation of heaven and earth. The faith that natural scientists are bound by the 'plain' language of numbers to speak with an authority that cannot be corrupted by fragile human judgment was gradually extended to the fields of engineering and the social sciences (Porter, 1986). The notion that experts who are disciplined by respect for objective facts are a symbol of integrity and can therefore serve as guardians of public virtues against the villains of politics and business was widely interpreted to include, beyond natural scientists, other categories of experts who speak the language of numbers. Social science disciplines like economics, social statistics, sociology, political science, and psychology adopted modes of observing, inferring, and arguing which appeared to deploy in the sphere of social experience notions of objective facts and discernible laws similar to those that the natural sciences had developed in relation to physical nature. The social sciences discovered that the language of quantification is not only a powerful tool in the production of knowledge of society but also a valuable political and bureaucratic resource for depersonalizing, and thus legitimizing, the exercise of power (Porter, 1995). In the context of social and political life the separation, facilitated by expert languages, between facts and values, causal chains and motives, appeared profoundly consequential. Until effectively challenged, mostly since the 1960s, such scientific and technical orientations to human affairs appeared to produce the belief that economic, social, political, and even moral 'facts' can be used to distinguish subjective or partisan from professional and apolitical arguments or actions. The apparent authority of science in the social and political context motivated the principal founders of modern ideologies like socialism, fascism, and liberalism to enlist science to their views of history, society, and the future.

While theoretical and mathematized scientific knowledge remained esoteric, as religious knowledge had been for centuries in the premodern state, the ethos of general enlightenment and the conception of scientific knowledge as essentially publicly made scientific claims appear agreeable to democratic publics even when those claims remained, in fact, elusive and

removed from their understanding. In fields such as physics, chemistry, and medicine, machines, instruments, and drugs could often validate, in the eyes of the lay public, knowledge claims that theories alone could not substantiate. The belief that science depoliticizes the grounds of state policies and actions, and subordinates them to objective and, at least in the eyes of some, transparent, professional standards, has been highly consequential for the uses of science in the modern state.

Political leaders and public servants discovered that the appearance of transparency and objectivity could make even centralized political power seem publicly accountable (Price, 1965; Ezrahi, 1990). Moreover, the realization that the authority of science could in various contexts be detached from substantive scientific knowledge and deployed without the latter's constraint often enhanced the political usefulness of both natural and social scientists, independently of the value actually accorded to their professional judgments. Still, the uses of expert scientific authority in the legitimization of state politics and programs have often empowered scientists to actually exercise influence on the substance of government actions or to effectively criticize such actions from institutional bases outside the government (Jasanoff, 1990). The actual uses of scientific expertise could often be just the consequence of a political demand for expert authority.

Science in Democratic and Authoritarian States

In democratic societies, the integration of scientific authority, and sometimes also of scientific knowledge and technology, into the operations of the modern state encouraged the development of a scientifically informed public criticism of the government. This process was greatly augmented by the rise of the modern mass media and the ability of scientists to call public attention to governmental failures traceable to nonuse or misuse of scientific expertise. The role of scientists in empowering public criticism of the uses of nuclear weapons and nuclear energy (Balough, 1991), in grounding public criticism of the operation of such agencies as the National Aeronautics and Space Administration (NASA) and the Food and Drug Administration (FDA), and in the criticism of public and private agencies for polluting the environment, illustrates the point. In authoritarian states like the Soviet Union, such cooperation between scientists, the mass media, and the public in criticizing government policies and actually holding the government accountable was, of course, usually repressed. In the authoritarian modernist version of science and the state, science was used mostly to justify, and partly to direct, comprehensive planning and control. Here, the force of scientific knowledge and authority rarely constrained the rationalization of centralization or warranted public skepticism toward the government or criticism of its actions (Scott, 1998).

Even in democratic states, however, science and technology were massively used to promote the goals of reconstruction, coordination, mass production, and uniformity. The redemptive role assumed by the political leadership in the name of equality, public welfare, and uniformity was no less effective in rationalizing state interventions than the totalistic utopianism enlisted to justify the direction and manipulation of society in authoritarian regimes. In all the variants of the modern state, bodies of expert knowledge such as statistics, demography, geography, macroeconomics, city planning, public medicine, mental health, applied physics, psychology, geology, and others were used to rationalize the creation of government organizations and semipublic regulatory bodies which had the mandate to shift at least part of the discretion held formerly by political, legal, and bureaucratic agents to experts (Price, 1965; Wagner et al., 1991).

In many democratic states, the receptiveness to experts within the state bureaucracy was manifest in the introduction of merit-based, selective, personnel recruitment procedures and a widening use of examinations. By contrast, the cases of the former USSR and Communist China suggest the ease with which experts and their expertise could be controlled by political loyalists. Even where it was facilitated, however, the welding of bureaucrats and professionals was bound to be conflict-ridden. The hierarchical authority structure of bureaucratic organizations has inevitably come to both challenge and be challenged by the more horizontal structure of professional peer controls (Wilson, 1989; Larson, 1977). Nevertheless, the professionalization of fields of public policy and regulation boosted the deployment of more strictly instrumental frames of public action in large areas. The synoptic gazes of power and science could thus converge in supporting such projects as holistic social engineering, city planning, and industrial scientific agriculture (Scott, 1998).

Science, War, and Politics

The most dramatic and consequential collaboration between science and the modern state took place during wars. Such collaboration was usually facilitated by the atmosphere of general mobilization, defining the war effort as the top goal of the nation. In a state of emergency, even democratic states have suspended, or at least limited, the competitive political process. Under such conditions, democracies have temporarily, albeit voluntarily, experienced the usual conditions of the authoritarian state: centralized command and control and a nationally coordinated effort. In the aftermath of such wars, politicians and experts often tried to persist in perpetuating the optimal conditions of their collaboration, regarding the resumption of usual political processes as disruptive of orderly rational procedures. In authoritarian nationalist or socialist states, open competitive politics was repressed as a matter of routine, along with the freedom of scientific research and the liberty of scientists to diffuse their findings and interpret them to students and the public at large (Mosse, 1966). In such countries, the political elite usually faced the dilemma of how to exploit the resources of science for the war effort and for advancing domestic goals without becoming vulnerable to the revolutionary potential of science as the expression of free reason, criticism, and what Robert K. Merton (1973) called "organized skepticism." In the democratic state, the special affinities between the participatory ethos of citizens, free to judge and evaluate their government, and the values of scientific knowledge and criticism did not usually allow the coherences and the clarities of the war period to survive for long, and the relations between science and the state had to readjust to

conditions of open political contests and the tensions between scientific advice to, or scientific criticism of, the state. Limitations imposed by open political contests were mitigated, however, in political cultures that encouraged citizens to evaluate and judge the government with reference to the adequacy of its performance in promoting public goals. The links between instrumental success and political legitimation preserved in such contexts the value of the interplay between expert advice to the government and to its public critics in the dynamic making of public policy and the construction of political authority. Political motives for appealing to the authority of science could of course be compatible with a readiness to ignore knowledge and adequate performance, although it did not need to contradict the willingness to actually use scientific knowledge to improve performance. Discrepancies between the uses of scientific authority and scientific knowledge became widespread, however, because of the growing gaps between the timeframes of science and politics in democratic societies. In fields such as public health, economic development, security, and general welfare, the instrumental effectiveness of policies and programs must usually be measured over a period of years, and sometimes even decades. But the politicians whose status has come to be mediated by the modern mass media need to demonstrate their achievements within a timeframe of months or even days or weeks. In contemporary politics, 1 month is a long time, and the life expectancy of issues on the public agenda is usually much shorter. Such a state of affairs means that political actors often lack the political incentives to invest in the costly resources that are required for instrumental success; investments whose payoffs are likely to appear only during the incumbency of their political successors.

Science and the Transformation of the Modern Democratic State

These conditions have encouraged leaders in democratic states increasingly to replace policy decisions and elaborate programs with grand political gestures (Ezrahi, 1990, 2012). Where instant rewards can be obtained by well-advertised commitment to a particular goal or policy, the political motive to invest in substantive moves to change reality and improve governmental performance over the longer run may easily decline.

Besides the tendency of democratic politics to privilege the present over the future in the allocation of public resources, the position of science in substantiating a long-term instrumental approach to public policy was further destabilized by the fragmentation of the normative mandates of state policies. Roughly since the 1960s, the publics of most Western democracies have become increasingly aware of the inherent constraints on the determination and ordering of values for the purpose of guiding public choices. Besides the influence of ethnic, religious, linguistic, or cultural differences on this process, such developments as the feminist movement reflect even deeper pressures to reorder basic values. Feminist spokespersons demanded, for example, that financial and scientific resources directed to control diseases be redistributed to redress gender discrimination against the treatment of specifically female diseases. In part, the feminist critique has been but one aspect of wider processes of individuation, which, toward the later part of the twentieth century, have increased the diversity of identities, tastes, lifestyles, and patterns of association in modern societies. This new pluralism implied the necessity of continually renegotiating the uses of science and technology in the social context, and locally adapting it to the diverse balances of values and interests in different communities. In the course of the twentieth century, leading scientific voices like J.B.S. Haldane, J.B. Bernal, and Jacques Monod tended to treat almost any resistance to the application of advanced scientific knowledge in human affairs as an 'abuse of science' resulting from irrationalism, ignorance, or prejudice. These and many other scientists failed to anticipate changes in social and political values that complicated the relations of science and politics during the twentieth century, and they also failed to anticipate the constraints imposed by the inherently competitive value environment of science and policy making in all modern states.

The constant, often unpredictable, changes in the value orientations of democratic publics, together with other related developments, have clearly undermined confidence in the earlier separation between facts and values, science and politics, and technological and political structures of action. Such a central element of policy making as risk assessment, for example, which for a long time was regarded as a matter best left to scientists, was gradually understood as actually a hybrid process combining science and policy judgments (Jasanoff, 1990; Beck, 1992). In contemporary society, in which the respective authorities of science and the state are demystified, knowledge has come to be regarded as too complex to directly check power, and power as too diffused to direct or repress the production and diffusion of knowledge.

Historically, one of the most important latent functions of science in the sociocultural construction of the democratic political universe was to warrant the faith in an objective world of publicly certifiable facts that can function as standards for the resolution or assessment of conflicts in opinions. To the extent that scientists and technologists could be regarded as having the authority to state the valid relations between causes and effects, they were believed to be vital to the attribution of responsibility for the desirable or undesirable consequences of public actions. Policies were regarded within this model of science and politics as hypotheses, which were subject to tests of experience before a witnessing public (Ezrahi, 1990). Elements of this conception have corresponded to the modern political experience. Although the public gaze has always been mediated, and to some extent manipulated, by hegemonic elites, publics could usually see when a war effort failed to achieve the stated objectives, when a major technology like a nuclear reactor failed and endangered millions of citizens, or when economic policy succeeded in producing affluence and stability. The expansion of this conception of government responsibility and accountability to regions outside the West even made it possible to redescribe fatal hunger in wide areas in India, for instance, as the consequence of policy rather than natural disaster (Sen, 1981). Still, in large areas, the cultural and normative presuppositions of the neutral public realm, as a shared frame for the nonpartisan or nonideological description and evaluation of reality, have not survived the proliferation of

the modern mass media and the spread of political and cultural pluralism. In contemporary democracies, the pervasive mediating role of largely commercialized electronic communication systems has diminished the power of the state to influence public perceptions of politics and accentuated the weight of emotional and esthetic factors relative to knowledge or information in the representation and construction of political reality. Influenced by diverse specialized and individualized electronic media, the proliferation of partly insulated micro-sociocultural universes within the larger society has generated a corresponding proliferation of incommensurable notions of reality, causality, and factuality.

Transformations of the Institutional Structure and Intellectual Orientations of Science

Changes in the political and institutional environment of science in the modern state have been accompanied by changes in the internal institutional and intellectual life of science itself. In the modern democratic state, the status of science as a distinct source of certified knowledge and authority was related to its institutional autonomy vis-à-vis the structure of the state and its independence from private economic institutions. Freedom of research and academic freedom were regarded as necessary conditions for the production and diffusion of scientific knowledge and, therefore, also for its powers to ground apolitical advice and criticism in the context of public affairs. Such institutional autonomy was, of course, never complete. The very conditions of independence and freedom had to be negotiated in each society and reflected at least a tacit balance between the needs of science and the state as perceived by the government. But the international universalistic ethos of science tended to ignore or underplay such local variations (Solingen, 1994).

The institutional arrangements that secured a degree of autonomy and freedom to scientists seemed to warrant the distinction between basic and applied research, between pure research directed to the advancement of knowledge and research and development directed to advancing specific industrial, medical, or other practical goals. An important function of this distinction was to balance the adaptation of science to the needs of the modern state and the preservation of the internal intellectual traditions and practices of scientific research and academic institutions. In practice, the relations between science and the state were more symbiotic than the public ethos of science would have suggested. Scientific institutions, in order to function, almost always required public and political support and often also financial assistance. On the other hand, the state could not remain indifferent to the potential uses of scientific knowledge and authority to both secure adequate responses to problems and facilitate the legitimation of its actions by reference to apolitical expert authority. Thus, while the separation between truth and power and the respect for the boundaries between pure and applied sciences, as well as between politics and administration, were considered for a long time as the appropriate way to think about the relation between science and politics (Price, 1965), the relations between science and the state, as well as between scientists and politicians, turned out to be much more interactive and conflict-ridden. Since scientists and politicians respectively controlled assets useful for each other's work, they were bound to be more active in trying to pressure each other to cooperate and engage in mutually useful exchanges (Gibbons et al., 1994).

Not all these exchanges were useful in the long run, either to science or the state. The mobilization of scientists to assist the war efforts of the modern nation-state, while it boosted the status of scientists domestically in the short run, had deleterious effects on the international network of scientific cooperation, as well as on the independence of science in the long run. Especially in cases of internally controversial wars, such as the American involvement in Vietnam, the mobilization of science inevitably split the scientific community and politicized the status of science. Nevertheless, the unprecedented outlays of public money made available to science in the name of national defense allowed many research institutions to acquire expansive advanced facilities and instruments that permitted the boosting of pure science in addition to militarily related research and development. But the gains in size and in potential scientific advance were obtained at the cost of eroding the glorious insulation of scientific research and exposing its delicate internal navigational mechanisms to the impact of external political values and institutions (Leslie, 1993). Yielding to the pressures of the modern nation-state to make science more relevant and useful to more immediate social goals inevitably reduced the influence of internal scientific considerations and the priorities of scientific research (Guston, 2000). Following such developments, the university, the laboratory, and the scientific community at large could often appear less elitist and more patriotic, but also more detached from their earlier affinities to humanistic culture and liberal values.

Like the partnership between science and the state, the partnership between science and the market, often facilitated by their collaboration during the war effort and postwar privatization of projects and services, has undermined the autonomy of science and the ethos of basic research in many late-modern states. But, whereas the interpenetration between science and the state subjected the internal intellectual values of science to the pressures of public goals and pressing political needs, the links between science and the private sector subordinated scientific norms to the private values of profit making. The conversion of scientific expertise into capital opened the way for substantial private support for the intellectual pursuits of the sciences. But the linkages between science, industry, and capital only reinforced the decline of science and its institutions as a bastion of enlightenment culture, progress, and universal intellectual values. Thus, while scientific and technical knowledge and skills were increasingly integrated into a wider spectrum of industrial productions and commercial services, science as a whole became more amorphous, less recognizable or representable as a set of distinct institutions, and a community with a shared ethos (Gibbons et al., 1994).

An instructive illustration of these processes is the pressure exerted by both the state and private business firms on scientists to compromise the cherished professional norm of publicity (making findings publicly available). In the context of science, the transparency of methodology and the publicity of research results have for a long time been reinforced by both

ethical commitment and practical needs. The methodologies and findings of any research effort are the building blocks, the raw materials, with which scientists in discrete places produce more knowledge. Transparency and publicity are necessary for the orderly flow of the research process and the operation of science as a cooperative enterprise (Merton, 1973). In addition, publicity has been a necessary element of the special status of science as public knowledge. From the very beginning, pioneering scientists such as Bacon, Boyle, and Lavoisier distinguished scientific claims from the claims of magic, alchemy, cabala, and other esoteric practices by the commitment to transparency. Transparency also was what rendered claims of scientific knowledge to appear publicly acceptable by democratic citizens (Tocqueville, 1945). Turning scientific research into a state secret in order to gain an advantage in war or into an economic (commercial) secret in order to gain an advantage in the market was not congenial to sustaining the cooperative system of science or the early luster of science as an embodiment of noble knowledge and virtues which transcended national loyalties and sectoral interests. In the aftermath of such developments, the virtues of objectivity, disinterestedness, universality, and rationality which were associated with earlier configurations of science (Polanyi, 1958, 1962; Merton, 1973) appeared unsustainable. Moreover, the fact that the resources of science and technology could be enlisted, not only by liberal and democratic but also by fascist, communist, and other authoritarian regimes, accentuated the image of science as an instrument insufficiently constrained by internal norms and flexible enough to serve contradictory, unprogressive, and extremely controversial causes.

The Postmodern Condition and the Reconfiguration of Science and The State

These developments had a profoundly paradoxical impact on the status of science in the late-modern state. Especially since the closing decades of the twentieth century, the authority of science in society and in the context of public affairs has suffered a sharp decline, even while such projects as the decoding of the human genome and the hunting for the Higgs Boson particle at the Large Hardon Collider in Geneva indicate that scientific knowledge has advanced beyond even the most optimistic predictions. In order to consider this state of affairs as reversible, one needs to believe that the insular autonomy of science can be restored and that the political and economic environments of science can become more congenial for sustaining this condition. Such expectations seem unwarranted. The long-standing assumption that science can be sharply separable from politics, or even that such separation is always desirable, has lost much of its earlier force. The attempts of science to embody the values of neutrality and objectivity in the political context have declined, due to developments not only within science but in politics as well (Ezrahi, 1996; Jasanoff, 2005).

It is important to stress that the apparent decline in the distinct social and political value of scientific authority, and in the perceived autonomy of science, may not necessarily undermine the impact of scientific knowledge on the ways states and governments act. What appears against the past as the decline of scientific authority in the larger social and political context may be redescribed as a reconfiguration and dispersion of the authority of expertise in the postmodern state. The boundaries perceived earlier between science, politics, law, ethics, economy, and the like may reflect an anachronistic ethos of science which no is longer relevant to contemporary, more sophisticated, views of the multiple decentralized interactions between various bodies of knowledge and their sociopolitical environments. Such interactions require finer, more intricate and continual adjustments and readjustments between science and power, and other perspectives on their relations.

In the absence of hegemonic ideological or national frames, the macropolitical sphere of the state divides into a multitude of – often just temporary and amorphous – subgroups, each organized around a particular order of values and interests. In each of these particular universes, scientific knowledge and expertise can occupy a privileged authority as a means to advance shared goals. When these groups clash or compete in the wider political arena, their experts tend to take sides and function as advocates. The state and its regulatory institutions enlist their own experts to back up their particular perspectives. Within this new, more open-ended and dynamic system, science is more explicitly and reflexively integrated into social, economic, and political values. In the context of the new pluralist polity, the authority and knowledge of science may therefore be less distinctly visible and less relevant to the symbolic, ideological, or cultural aspects of politics. At the same time, it is no less present in the organization, management, and the normative constitution of social order.

Several contemporary theorists have attempted to reconceptualize the relations between science and the state in light of new understandings of the diverse modes of generating and diffusing scientific knowledge and novel configurations of science, technology, politics and the economy in the contemporary state. A few scholars have advanced the idea that knowledge and social order are coproduced by complex processes that ensure a degree of compatibility between the character and authority of the knowledge produced, the framing of political power, and modes of government that become legitimate. This approach has been reinforced by studies that demonstrate the ways in which historically embodied variability in the collective imaginaries of legitimate order like monarchy and democracy evolve respectively matching forms of culture and knowledge (Ezrahi, 2012). This theory has been supported by the finding that diverse sociocultural contexts tend to generate variable perspectives on science, its symbolic contextual implications and applications (Gibbons et al., 1994; Jasanoff, 2004).

Another approach led by Bruno Latour and Michel Callon, and inspired by Gabriel Tarde, Gilles Deleuze, Donna Haraway, and other theorists of interaction and hybrid agents, has aimed at replacing the hegemonic metaphysical dichotomy of subjects and objects with an alternative theory according to which the boundaries between human and nonhuman agents, which they regard as equivalent semiotic "actants," are blurred or canceled (Latour, 2005). The material-semiotic perspective they advance focuses on an imaginary consisting of networks in which human and nonhuman agents (including microbes, inanimate matter, machines, etc.) form dynamic assemblages that generate social actions and processes which compose the social-material whole. This approach challenges conventional

hierarchies composed of persons, objects, and animals by reframing our understanding of the place of science and technology in our social and political world-making.

Whereas the above theories seem to open new ways to understand the relations between science and the state, they also elicit some instructive criticisms. The theory of the coproduction of knowledge and the social order may underestimate the unpredictability and the accelerated revolutionary potential of the advancement of knowledge to undermine established norms and institutions without evolving new ones, and the actor-network theory (ANT) has difficulties dealing with politics and political power and with its problematic diminishment of voluntarism, ethics, and responsibility as central components of social-scientific accounts of the human dimensions of society and politics.

See also: Development and the State; Funding of Social Science, History of; Norms in Science; Research Funding: Ethical Aspects; Science and Technology Studies: Experts and Expertise; Science, History of; Science, New Forms of; Science, Sociology of; Science: Constructivist Perspectives, History of; Universities and Science and Technology: United States.

Bibliography

Balough, B., 1991. Chain Reaction: Expert Debate and Public Participation in American Commercial Nuclear Power, 1945–1975. Cambridge University Press, New York.
Beck, U., 1992. Risk Society: Towards a New Modernity. Sage Publications, London.
Dumont, L., 1986. Essays on Individualism: Modern Ideology in Anthropological Perspective. University of Chicago Press, Chicago.
Eisenstadt, S.N., 1999. Paradoxes of Democracy: Fragility, Continuity and Change. The Woodrow Wilson Center Press, Washington, DC.
Ezrahi, Y., 1990. The Descent of Icarus: Science and the Transformation of Contemporary Democracy. Harvard University Press, Cambridge, MA.
Ezrahi, Y., 1996. Modes of reasoning and the politics of authority in the modern state. In: Olson, D.R., Torrance, N. (Eds.), Modes of Thought: Explorations in Culture and Cognition. Cambridge University Press, Cambridge, pp. 72–90.
Ezrahi, Y., 2012. Imagined Democracies: Necessary Political Fictions. Cambridge University Press, Cambridge.
Gibbons, M., Nowotny, H., Limoges, C., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, London.
Guston, D.H., 2000. Between Politics and Science: Assuring the Integrity of Research. Cambridge University Press, New York.
Jasanoff, S., 1990. The Fifth Branch: Science Advisers as Policymakers. Harvard University Press, Cambridge, MA.
Jasanoff, S., 2004. The idiom of coproduction. In: Jasanoff, S. (Ed.), States of Knowledge: The Coproduction of Science and the Social Order. Routledge, London, pp. 1–12.
Jasanoff, S., 2005. Designs of Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton.
Larson, M.S., 1977. The Rise of Professionalism. University of California Press, Berkeley.
Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.
Leslie, S.W., 1993. The Cold War and American Science. Columbia University Press, New York.
Machiavelli, N., 1964. The Prince. St. Martin's Press, New York.
Manuel, F.E., 1974. The Religion of Isaac Newton. Clarendon Press, Oxford.
Merton, R.K., 1973. The Normative Structure of Science. In: Merton, R., Storer, N.W. (Eds.), The Sociology of Science. University of Chicago Press, Chicago, pp. 267–278.
Mosse, G.L., 1966. Nazi Culture. Schocken Books, New York.
Polanyi, M., 1958. Personal Knowledge: Towards a Post Critical Philosophy. University of Chicago Press, Chicago.
Polanyi, M., 1962. The republic of science: its political and economic theory. Minerva 1, 54–73.
Porter, T.M., 1986. The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton.
Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton.
Price, D.K., 1965. The Scientific Estate. Belknap Press, Cambridge, MA.
Scott, J.C., 1998. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven, CT.
Sen, A.K., 1981. Poverty and Famines: An Essay on Entitlements and Deprivation. Clarendon Press, Oxford.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton, NJ [second ed., 2011].
Skinner, Q., 1996. Reason and Rhetoric in the Philosophy of Hobbes. Cambridge University Press, Cambridge.
Solingen, E. (Ed.), 1994. Scientists and the State: Domestic Structures and the International Context. University of Michigan Press, Ann Arbor.
Sprat, T., 1958[1667]. History of the Royal Society. Washington University Press, St Louis, MO.
Tocqueville, A., 1945. In: Bradley, P. (Ed.), Democracy in America. Knopf, New York.
Wagner, P., Weiss, C.H., Wittrock, B., Wollman, H., 1991. Social Sciences and Modern States. Cambridge University Press, Cambridge, UK.
Wilson, J.Q., 1989. Bureaucracy: What Government Agencies Do and Why They Do It. Basic Books, New York.

Science Communication

Jane Gregory, CHSTM, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Science communication is the communication of scientific knowledge and ideas to people who are not scientists. Communication among scientists is fundamental to the generation and affirmation of scientific knowledge; communication with people who are not scientists serves both to inform them and to identify scientists as a particular form of authority. The products of science communication, which are known as popular science, have distinctive rhetorical features, but are also subject to the trends and values of areas such as journalism and public relations. Theories of science communication have evolved from simple transmission models to ideas more congruent with ideas about the 'network society.' Science communication is used to encourage development in emerging economies, and in post-industrial societies it becomes part of the immaterial production of technological futures, bringing innovations into social life.

'Science' has many meanings. It is a form of enquiry: for some, this includes, very broadly, any systematic enquiry that involves gathering data, developing hypotheses, and testing them; and for others, and also colloquially in the English-speaking world, science means data and hypotheses about the natural world in particular, rather than the social or the metaphysical. Science is also a profession, the members of which are scientists; but in common parlance scientists as a group are usually referred to not as a profession but as a community. Unusually among professions, there is also a common word for people who are not members of this community: they are nonscientists. It is characteristic of science that it is concerned with this kind of classifying, both in its approach to problem-solving and in its assertions of its own identity, from which it draws authority to speak on some topics but absolves itself from responsibility for others. This concern among scientists themselves with demarcating science and scientists from other forms of enquiry, other problems, and other people is reflected in the fact that the communication of science with nonscientists has its own name: science communication. The products of science communication are often called 'popular science,' to distinguish them from communications between scientists.

Communication in Science

The essential role of communication in the formation of modern science is clear in the principles and practices of its founders, who came together and developed a common identity in the seventeenth century in the capitals of Western Europe (Shapin and Schaffer, 1985). The philosophy of empiricism quickly came to dominate: that is, that nature should be encountered not through the imagination or through acts of mental logic, but through direct observation and manipulation of the stuff of the natural world. These experiments would, if competently conducted, yield the same insights wherever, and by whomsoever, they were performed. It was therefore necessary that the experiments should be witnessed, reported in detail, and scrutinized by others, so that the outcomes could be understood collectively. This collective aspect of science was advocated not only as a means to truth about nature but also as having greater social value than the potentially destructive alternative of disputatious factions.

Early scientists sought to extend their community by disseminating accounts of their activities in print: scientific societies published journals that included reports of meetings, scripts of lectures, and also accounts of experiments, to allow distant readers to share in the experience. In Europe, Latin was routinely used to share scientific knowledge abroad, and, later, foreign languages became an important part of a scientist's toolkit. The social transition that resulted in the institution of modern science therefore has communication at its core.

This communication among scientists is considered a normal part of scientific work and fundamental to the social identity and epistemological integrity of science. Communicating with nonscientists, however, has often been considered by scientists as a peripheral and low-status activity. However, these evaluations have varied historically and from place to place and have recently been problematized in science communication studies.

The Origins of Popular Science

While the emerging scientific community sought ways to identify and gather its members, exponents of natural knowledge were also active contributors to the cultural turn known as the Enlightenment. In this preprofessional era for science, writing and speaking about the natural world were among the few ways in which scientists could earn money from it, and the burgeoning public sphere, new media, and new forms of commerce provided new spaces in which to do so. The idea of natural knowledge as a universal good fueled a rich and diverse scene.

The label 'science' is an anachronism at this point and does not imply the strict categorization of later centuries, of authors, subject matter, genre, style, or audience. The literature we would now identify as science was then enmeshed in the general discourse of politics, religion, and society. The proliferation of publishers, media, and readerships brought with it a wide diversity of styles and genres, and, as it framed itself for these new cultural niches, the identity of science was

problematic: one of the most prolific 'science communicators' of the eighteenth century, Buffon, author of the monumental *Natural History*, was embroiled in grave discussions with peers on the relationship between literary style and the facts of nature, including topics such as the extent to which natural knowledge can be conveyed by metaphor. The recurring question was how one can faithfully report the single voice of universal nature, while at the same time meeting the diverse demands of the broader, and ever broadening, popular culture.

The emphasis on consensus as the goal of scientific enquiry, and the wide geographical spread of correspondents and journal readers, made scientists ideal ambassadors. In times of international conflict and expanding empires, scientists were often in the colonial vanguard: ostensibly above politics, they traveled freely, spreading universal knowledge and identifying natural resources to send to the imperial center.

By the nineteenth century, science across Europe was organizing as a profession and seeking an institutional infrastructure and funding. The application of natural knowledge was marketed as the agent of progress, and from industrialization could come wealth. The label 'scientist' was adopted, identifying a new profession. One of the services offered by the new profession in the industrializing countries of Europe was to prepare a workforce for industrial labor and to support the urbanized life that industry entailed. Programs of lectures were orientated around particular industries such as mining or metal-working and on topics such as technical drawing and mechanics.

To gain the authority to advise and the social mandate to practice, scientists identified their knowledge as special and themselves as specialists. This involved displays of the power represented by the communication of their special knowledge to nonscientists, who included both social authorities and the common people. When a monarch undertakes publicity with displays of regalia, the material of power returns to the palace at the end of the tour; knowledge, however, is a nonrival good: the lecturer still has it, but the audience also take it home with them. Fears about the revolutionary potential of evolving democracies and their newly literate citizens focused the attention of the educated elites on the question of how to enable the people for this changing world but at the same time contain any ambition or potential to disrupt the social status quo. Popular science, they argued, risked empowering the masses, but its proper handling could confer discipline, reason, and compliance.

So while public communication of science has often been explicitly committed to Enlightenment ideals, it has usually also happened in the contexts of programs to manage and direct public responses to a changing world more generally and to science and technology in particular. By the end of the nineteenth century in Europe, there was a publishing industry, lecture series, and forms of education about science, directed by scientists and nonscientists. Science communication for nonscientists became distinct from science communication among professionals, which became increasingly esoteric in expensive journals. Scientific journals have been a lucrative and persistent medium: their number has grown by factor of 10 every 50 years since 1650, and they remain a significant medium in the internet era. Journals select for publication only the most significant submissions, so that journal publications are considered the 'gold standard' in scientific productivity and achievement, and are used as a measure of scientists' performance, both by policymakers looking to distribute resources and by academic analysts looking to map the structure and workings of the scientific community.

Charles Darwin (1809–82) is sometimes identified as the last of the amateurs who dominated science before it professionalized, and his theory of evolution by natural selection, published in 1859, is identified as the last significant event in the history of science to be communicated to both scientists and laypeople in one and the same language and form: in the book *On the Origin of Species*. Thereafter, a scientist was someone who worked in science for a salary, and if they wrote for the public, it was in a different forum, medium, and style from their communication with their colleagues.

Science Communication as Popular Science

By the turn of the twentieth century, there were clear, formal routes, controlled by scientific institutions, into science as a profession, and an established market for popular science alongside the flourishing professional science market. Science communicators who were not themselves scientists drew on the long tradition of public talk about natural knowledge and wrote books and magazine articles and embraced new media including film and radio. In Europe, the sudden expansion of the mass market newspaper industry came just in time not only for the highly technological World War I but also for the scientific revolution of the 'new physics,' when relativity theory and quantum mechanics challenged even the best scientists and most proficient communicators. The culturally widespread idea that science is very complicated (and scientists very clever) may have arisen at this time, underlining the value of the specialist practitioners of science communication (Bucchi, 2004).

Science communication has a clear rhetorical identity by this time. Where scientists are understood to be very clever, nonscientists are rhetorically constructed as less so. Science communication often makes explicit assertions about what it assumes its audience does not know or misunderstands: phrases such as 'Many people mistakenly believe that …' will commonly precede a startling revelation of scientific knowledge, often identified as startling by an exclamation mark. Where scientists tend to express their professional claims in tentative, conservative, and incremental terms, science communication deploys a rhetoric of achievement and change, often with confident portrayals of the future. Science communication has traditionally privileged the visual, deploying the evidence of the senses, in the same way as professional science privileges the mathematical, calling on the capacities of intellect and education. Science communication uses literary devices such as narrative, metaphor, and anthropomorphism in contrast to the largely forensic language of professional communication.

These rhetorical conventions are symptoms of the long-standing tension in science communication about how to exhibit knowledge that is esoteric enough to affirm the special status of scientists, while at the same time revealing the value of science that earns it esteem, public support, and, eventually,

taxpayers' money. This tension is also reflected in the classic popular science scenario in which social boundaries are maintained while knowledge is displayed. Difference and distance between the scientist and the nonscientist can be emphasized by the architecture of the lecture hall (often with the physical barrier of the lab bench between lecturer and audience), styles of dress (the white coat), or linguistic conventions (such as jargon). This holds the public at arm's length so as to emphasize difference but at the same time benefit from the affirmation provided by their proximity (Gieryn, 1999).

Science communication has many purposes for scientists, not least of which is the need to secure funding and other resources such as positive public opinion (Dornan, 1990). It tends to flourish in times of crisis in the political fortunes of science. It is also a way for scientists to reach each other via the public sphere, untrammelled by the selectivity and strict conventions of journal publications, and so scientists use science communication to fight their corner both in the early days of new fields and when their status is receding. A confident, well-resourced, and productive discipline or field is usually quieter in the public sphere.

Science communication typically claims the political neutrality often attributed to scientists and does not ask its audience to behave in particular ways: it is enlightening, rather than educational. Unlike education, popular science can be merely surprising; it does not have to be meaningful or comprehensible. This distinction between education and enlightenment is strict enough to have produced explicitly hybrid institutions, such as science centers, which deploy the techniques of science communication in an explicitly educational setting, and are often linked to national educational programs.

Jurdant (1993) characterizes science communication as the autobiography of science: it is the story science tells about itself. It is idealized and tidied up and excused from the contingencies of practice. It tends to retrospectively construct direct paths from experiment to discovery, when scientists may instead talk among themselves about difficulties, diversions, and set-backs; it focuses on the emerging consensus rather than controversy from which it emerged. Its neutrality is presented as objectivity. Latour (1987) characterizes this as the 'Janus face' of science, contrasting the relatively private 'inward' talk among scientists with the 'outward' public stories being told to nonscientists.

Science Communication as Politics

Professional science journalists established a recognized beat in the West during the 1930s, and science and politics came together on the global platform of World War II, most visibly in response to the atomic bomb, which established a form of journalism, delivered by specialist reporters, that was enthusiastic about the technical wizardry of science and committed to its role in shaping the post-war reconstruction. The economic and social dimensions of science gave mainstream journalists their entrée into this new beat, and their message was one of faith in scientific rationality and technological progress.

Science was also brought into mainstream politics during the 1960s in the West via restructuring of public funding for science and became an issue of domestic politics. At this time too, science journalists were rediscovering the critical spirit of mainstream journalism. In a society tending toward a critical approach to authority more generally, the public sphere was becoming an increasingly combative space for science, involving new actors and voices from outside the scientific community, such as new social movements focused on issues such as human rights, the environment, health, and war. Scientists became less confident in this pluralized discourse and to a significant extent withdrew from the public sphere, leaving space for challenges to scientific authority and for alternative worldviews. Another reason to turn back toward their own community came in 1969, when the editor of the highly prestigious *New England Journal of Medicine*, Franz Ingelfinger, declared that no research could be published in his journal if it had already been exposed in public. For individual scientists, who needed publications to secure their careers, the choice was clear. The Ingelfinger rule spread quickly among journals and remains a condition of journal publication today – one that journals exploit through their own public relations activity by turning publication, rather than the research itself, into a news event.

The turn away from public communication was problematized during the 1980s by scientists themselves. With the advent of neoliberal economics and pressure on publicly funded scientists to contribute to wealth creation, scientists became conscious of the extent to which their market value depended on public opinion. In an era considered by historians and sociologists to be a turning point in the cultural status of science, and at a time of ambivalence in relation to authorities of all kinds, scientists took a renewed interest in their public relations (Bauer and Bucchi, 2007).

As a consequence, funding was made available for science communication activity of many kinds. This also provided a commercial opportunity for publishers and other media organizations to develop science content, at a time when the global media industry was developing into a multichannel, transnational business. Through the 1990s, scientific institutions became alert to the benefits of media coverage and found ways of collaborating with journalists. Now very few research institutes in Europe are without their press office. The universities work closely with the journals that publish their scientists' research, coordinating press releases and respecting embargoes. Journals publish not only press releases but also news digests to help journalists access their esoteric contents. Thus publication in an esoteric journal can automatically mean exposure in the public sphere for an author who thus brings their work, their institution, and the journal to the attention of opinion leaders, policy-makers, and the public, without ever having to step out of the laboratory. Some observers claim that journals' prestige is now measured, at least in part, by the amount of media coverage they can attract and that media potential may be a factor in journal editors' decisions about whether or not to publish a research paper.

The internet-based journalism that arises from the press releases of high-prestige international journals has come to dominate science reporting. In the United Kingdom, science news is dominated by research press-released from journals, and a story in one newspaper will be repeated in several others, with the same photograph and quotes, often uninflected by the

newspaper's political stance. This uniformity among media outlets that traditionally made a virtue of their political differences is a peculiar feature of science reporting and one that worries some critics, who see a loss of democratic potential in the journalists' close relationships with institutions' press officers. At the same time, with scientific research in the West being undertaken increasingly in private institutions such as businesses, science news is not press-released to protect it from competition; so a proportion of scientific research is already withheld from media scrutiny.

Beyond the mass media, the tendency toward centrist politics set the context for novel approaches to many aspects of social life, including the relations between science and society (Gregory and Lock, 2008). Instead of the adversarial politics of the traditional Left and Right, a consensus-seeking approach to the middle ground deployed consultative and dialogic interaction between the public and policymakers, including in areas of scientific research such as genetically modified food and nanotechnology. This approach made science a matter for negotiation among interest groups, and science communication activity was reorientated toward events that encouraged dialogue. These events were often called 'public engagement with science' to distinguish them from traditional science communication. Although engagement as identified as a way of enhancing democratic interaction, its agenda is nevertheless, like traditional science communication, invariably set by the more powerful actors; and participants may experience a process that echoes the management of the workers in the nineteenth century. Critics of public engagement note that what appears to be an open negotiation may be very closely prescribed, with a restricted range of topics and potential outcomes, and that such events are often staged to provide a moderate alternative to street protests or activist campaigns (Lezuan and Soneryd, 2007). This prescriptive aspect takes engagement toward the principles of professional public relations. This situates engagement within the broader trend in science communication, which increasingly adopts the rhetoric and values of public relations.

Theoretical Models of Science Communication

Research into science communication has tended to focus on the efficiency of the transfer of information and has often been conflated with studies of 'scientific literacy' or 'public understanding of science,' which refer to what nonscientists know about science. This knowledge tends to be described as deficient in quality and quantity in comparison to the knowledge of scientists. In this 'deficit model' of public understanding of science there is a normative element that the deficit should be remedied; and the remedy is to transmit knowledge from scientists to laypeople (Gregory and Miller, 1998). There is often a tacit assumption that increasing the public's knowledge will also lead to more positive attitudes toward science – an assumption that is itself based on the common but often unfounded assumption of widespread negative public attitudes to science. However, research on these questions consistently indicates that greater knowledge does not necessarily generate more positive attitudes; and more science communication does not necessarily result in higher levels of knowledge. Critics have also asked 'why' the public's 'deficit' should be remedied, when their ignorance may be socially functional, and the purpose of scientists is that they should be the ones who know about science. Most people know very little of many areas of knowledge, and societies seem to function nevertheless. The deficit model shows, then, an ideology that is not grounded in empirical evidence.

Much of the criticism of the media coverage of science, and the ambitions for science communication, can be conceptualized using what Hilgartner (1990) identified as the dominant model of science communication. Hilgartner noted that much science communication by scientists is framed by the idea that science is emitted by scientists and transmitted by journalists to a receptive public. In this regard, the model is similar to the sender-transmitter-receiver model of classical communication studies. However, the model also embodies social relations, and the transmission process it expresses is vertical, with the scientists at the top. Scientists have the highest-quality knowledge, which they allow to trickle down to journalists whose job is to pass on the information as faithfully as they can. The final destination of this pristine knowledge is a passive, uncritical public, who listen but do not speak. The classic expression of this model is that science communication is about 'bringing it down to their level.' The position of the journalist as mediator maintains a separation between scientists and the public and allows the journalist to stand as scapegoat for communication failures, such as public misunderstandings or lack of political interest.

Communication scholars argue that the public communication of science does not work in this way. Instead, science communication is a complex network of institutions, ideas, and actors who collectively participate in the production of popular science. The public, journalists, and scientists are active participants from a range of backgrounds and playing multiple roles, interacting in many directions and with a variety of media, picking up, passing on, and inflecting information and comment about science (Lewenstein, 1995). Public engagement is an expression of this more fluid, participatory approach to science communication.

The idea of science communication as a complex network brings it into the mainstream of communication theory, alongside many other areas of communication. By the end of the twentieth century, it was adapting to a world transformed by new technologies; and traditional hierarchical social relations, such as those expressed in the dominant model, were sitting uncomfortably among spontaneous and ever-changing horizontal networks. In these networks, there is no clear divide between who can speak and who must listen, with many playing both parts; and the markers of expertise and authority are often difficult to identify. Science communication in early twenty-first century should therefore be understood in the context of theories about this 'network society.'

Science Communication in the 'Network Society'

According to its leading proponent Castells (2000), the 'network society' is a result of three conditions: first, the social changes wrought by and expressed in the cultural and social

movements of 1960s and 1970s, such as feminism, environmentalism, and the antiwar movements, which brought an unprecedented diversity of issues, opinions, and expertises into the public sphere; second, the turn to digital in information technology in the 1970s and the spread of the personal computer from the mid-1980s; and third, the dominance of neo-liberal economics by the end of century. Networks are a longstanding form of social organization: the innovation in the network society is that the networks process and manage information. Information management is an activity that is characteristic of a post-industrial society, where much of economic production is about immaterial goods such as knowledge and experiences. Science is increasingly organizing its relationships in mediated networks, gradually replacing or complementing social networks of face-to-face communication. For scientists, their primary social group is no longer their immediate neighbors and their national context, but instead the individuals with whom they connect from across the global network of the internet. The internet is an open network, and scientists and nonscientists share access to many of the same tools and knowledge, and information can travel not only in expensive journals but also freely, and instantaneously, through electronic networks, facilitating distributed collaborations and diverse critiques.

Scientists benefit from being in a network by the access it grants them to specific and rare resources, and also from the broader network of collaborators, and the global exchange of free-floating ideas and information that may change the course of research (Wagner, 2009). Researchers are free to join forces to tackle common problems regardless of where they are geographically, and their outputs are accessible to a geographically distributed audience of readers, commentators, and critics. The internet as a medium supports interpersonal communication, organizational/business communication, and mass communication all at the same time, which blurs the boundaries between the traditionally distinct roles and audiences in science communication: nonscientists can easily access scientists' professional communications, and anyone can publish scientific information on the Internet without it being subject to the gate-keeping performed by scientists for their professional journals. Advocates of alternative or 'fringe' sciences readily find a platform there, mobilizing the tools, rhetorics, and audiences of mainstream science. The reach, scale, and spontaneity of the network empower its users, providing scope for insurgent politics, particularly on issues of global concern such as the environment and infectious disease.

Networks also play a part in the acceptance or rejection of innovations in society. In a post-industrial, late-capitalist economy, a competitive nation will secure its place in the global market primarily through the promotion of innovation, and of the social and economic conditions for innovation, and by making ready a public for novel goods. Networks enable a small group of supporters of the innovation to share their enthusiasm not just by the traditional 'word of mouth' in interpersonal networks but also through the instantaneous global networks of the internet where science communication serves to display and elucidate the benefits of new products (Rogers, 2003). Markets quickly become saturated, and producers must adapt, making further innovations that can create new markets.

Networks can also restrain innovation by mobilizing critiques of new products and creating representations of their negative attributes or consequences. Markets are vulnerable to public opinion, and in the network society there is space for many points of view to compete for attention. Studies of responses to new technologies show how complex they are: for example, responses to biotechnology across Europe showed similarities and differences between cultures, nations, and regions that appear to relate variously to the religious context, history, media norms, linguistic resources, values, gender, individual wealth, and degree of industrialization, among other factors (Bauer and Gaskell, 2002). Such data potentially allow science communicators to intervene in attitude formation: they could, for example, attempt to mitigate moral concerns about a new product by emphasizing its usefulness.

Less industrialized countries are often more open to new technologies than their more industrialized neighbors. Science communication is increasingly promoted as a tool for development in less industrialized parts of the world, where, for example, cheap wireless technologies are not only an innovation in themselves but also bring the potential for social innovation, by providing an instant infrastructure for technologically, economically, and politically disenfranchised groups. In these as well as in other ways, science communication becomes a kind of activism, with a diverse membership and across a spectrum of loyalties, and outwith the control of any formal agency. In the network society, the traditional roles and controls no longer apply.

Science Communication as Immaterial Labor

Much of economic production in the post-industrial context involves knowledge and services. Anyone who puts information on the Internet is a producer, and consumers of this information also contribute to producing the network itself, which does not exist until and unless people use it. Thus production and consumption are conflated in a single act. Matters of public good, such as education, healthcare, and knowledge, become commodified. Science communication is readily commodified in such an environment (Thorpe and Gregory, 2010).

Technologies have social meanings, and these meanings are an important aspect of their value, which extends beyond the ostensible purpose of the technology itself. These meanings are constructed and shared through intellectual and affective work that is typical of post-industrial capitalist societies. This work, which can be described as a form of immaterial labor, brings immaterial attributes to the commodity, such as hope, freedom, or a sense of belonging. Developing these attributes means bringing the emotions and communicative capabilites of the worker into the production process, so that the communicative activities we think of as being about enlightenment and the values of the public sphere are brought into the sphere of production. Not only are the workers mobilized in this way, but so too are the consumers of the information they produce: the public is productive in that it confers as sense of reality on innovation – in effect, making it – and accepts it (or not), by bringing it into the social world. While science and

technology have long been represented as contributing goods and services to the economy, now, in the post-industrial world, the public communication of science contributes to the economy too and becomes a good in itself.

Traditional science communication has typically appealed to its readers' reason, and activists who have deployed an emotional discourse have been criticized as unscientific: 'sensationalist' is a common disparagement. But in the post-industrial society, the emotions are mobilized in support of new technologies, and the promises about these technologies are not only about their efficiency for a prescribed task, but also about aesthetics, experiences, and better futures. Hopes and expectations about future applications are essential for attracting venture capital, so the construction and management of expectations becomes integral to scientific development. The practice of 'upstream engagement' contemplates future technologies, which may have little basis in present capabilities or knowledge, but which allow public opinion a preview that minimizes the risk of future market failure. In such cases, there is nothing to the issue but talk: the only commodity is the public talk, which brings the future technology into the present. Here, the boundary is blurred between science communication and science fiction: the imaginings of new technologies are already immaterial products with value in the mass media; there they attract both government interest and private investment. Thus, science communication enrolls participants into the immaterial production of imagined futures, integrating them with social values and turning innovations into everyday objects through public talk.

Conclusion

The plurality of the network society supports science communication in a wide range of forms and media and speaks with multiple voices in a variety of roles. Science communicators work in a variety of roles in museums, businesses, government and universities, as well as media institutions. Science communication is subject both to the changing context for science and to the changing frame of the public sphere: it reflects the turn from unquestioning respect for authority to the challenge of activism; from control over the messages about science in the public sphere to a more open and fluid interaction; the emergence of public relations as a dominating mode of information management; and the shift from industrial to post-industrial production. At the same time, science communication retains its connotations of enlightenment, and is experienced as such, by a variety of audiences in many cultural niches as well as on a global scale.

See also: Capitalism; Castells, Manuel (1942–); Citizenship, Anthropology of; Darwin, Charles Robert (1809–82); Deliberation and Democracy; Empiricism, History of; Enlightenment; Information Society; Innovation, Theory of; Internet and Social Media: Anthropological Aspects; Journalism; Knowledge, Sociology of; Liberalism, Anthropology of; Literacy and Illiteracy, History of; Modernization and Modernity in History; Modernization, Sociological Theories of; Networks and Cultural Consumption; Postindustrial Societies; Prison Life, Sociology of; Professions and Professionalization, History of; Public Opinion: Social Attitudes; Public Sphere and the Media; Public Sphere: Eighteenth-Century History; Public Sphere: Nineteenth and Twentieth-Century History; Rhetorical Analysis; Science and Development; Science and Media; Science and Politics: Value Neutrality; Science, History of; Science, Public Engagement with; Scientometrics.

Bibliography

Bauer, M.W., Bucchi, M. (Eds.), 2007. Journalism, Science and Society: Science Communication between News and Public Relations. Routledge, London.

Bauer, M.W., Gaskell, G., 2002. Biotechnology: The Making of a Global Controversy. Cambridge University Press, Cambridge.

Bucchi, M., 2004. Science in Society: An Introduction to Social Studies of Science. Routledge, London.

Castells, M., 2000. The Rise of the Network Society. Wiley, Blackwell, New York.

Dornan, C., 1990. Some problems in conceptualizing the issue of 'science and the media.' Critical Studies in Mass Communication 7, 48–71.

Gieryn, T., 1999. Cultural Boundaries of Science: Credibility on the Line. Chicago University Press, Chicago, IL.

Gregory, J., Lock, S.J., 2008. The evolution of 'public understanding of science' in the UK. Sociology Compass 2 (4), 1252–1265.

Gregory, J., Miller, S., 1998. Science in Public: Communication, Culture and Credibility. Plenum, New York.

Hilgartner, S., 1990. The dominant view of popularization: conceptual problems, political uses. Social Studies of Science 20, 519–539.

Holliman, R., Thomas, J., Smidt, S., Scanlon, E., Whitelegg, E. (Eds.), 2009. Investigating Science Communication in the Information Age: Implications for Public Engagement and Popular Media. Oxford University Press, Oxford.

Hook, D., Franks, B., Bauer, M. (Eds.), 2011. The Social Psychology of Communication. Routledge, London.

Jurdant, B., 1993. Popularisation as the autobiography of science. Public Understanding of Science 2, 365–373.

Latour, B., 1987. Science in Action. Harvard University Press, Cambridge, MA.

Lewenstein, B.V., 1995. From fax to facts: communication in the cold fusion saga. Social Studies of Science 25, 403–436.

Lezuan, J., Soneryd, L., 2007. Consulting citizens: technologies of elicitation and the mobility of publics. Public Understanding of Science 16, 279–297.

Rogers, E.M., 2003. Diffusion of Innovations. Simon & Schuster, New York.

Shapin, S., Schaffer, S., 1985. Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton, NJ.

Thorpe, C., Gregory, J., 2010. Producing the post-fordist public: the political economy of public engagement with science. Science as Culture 19 (3), 273–301.

Wagner, C.S., 2009. The New Invisible College: Science for Development. Brookings Institution Press, Washington, DC.

Science: Constructivist Perspectives, History of

Iwan R Morus, Aberystwyth University, Aberystwyth, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a summary of developments in the history of science informed by social constructivist perspectives since the 1970s. It discusses key influential texts and highlights historical themes of particular importance such as the understanding of science as local culture, the material culture of experiment, science and performance, and science and labor. The article also emphasizes the constructivist understanding of science as a species of skilled practice and the ways in which such accounts draw attention to the importance of understanding the circulation of skills in the development and consolidation of new knowledge.

Historians of science inspired by the strong program in the sociology of scientific knowledge have developed new approaches to understand science's past. Rather than considering science as an abstract, disembodied body of ideas they have focused on science as a localized practice, embedded in material, national, and political cultures. They have looked at the work involved in making experimental phenomena robust and repeatable. They have looked at the skills involved: their origins, their translation from site to site and individual to individual, their instrumental embodiment, and the changing social status of their possessors. This has focused attention on the ways in which scientists have represented themselves, the knowledge they produce, and on the ways in which they and their products have been represented. Economic, social, and political interests were at stake in such processes. How knowledge was historically constructed was highly consequential for the organization of society. Such histories have also focused on how audiences have received, actively appropriated, and reformulated science in different contexts. Science, from this perspective, is a constant process of production, reproduction, and consumption.

The last quarter of the twentieth century saw the emergence of a challenging new perspective on the sciences and their history. Drawing on developments in the social studies of science, as well as new trends in labor and social history, historians began to look at science as a local practice, rather than as an abstract, disembodied, and universal set of ideas. Such historians started looking at science as part of culture, locally constituted in particular material, social, and even political practices. Constructivist historians are interested in the detailed processes of scientific knowledge production, dissemination, and reception. They differ from more traditional social historians of science in their insistence that the content as well as the context of science is a proper focus of cultural historical analysis. Constructivist historians see science as a species of skilled practice and therefore inquire into the cultural origins of those skills; the means and the spaces through which they are produced, validated, and disseminated; and the ways in which particular forms of inquiry – and particular kinds of knowledge – have historically been characterized as scientific.

Local Cultures

In their exemplary study of the early modern culture of experiment, Shapin and Schaffer lay out some of the central concerns of the constructivist history of science (Shapin and Schaffer, 2011[1985]). Focusing on the dispute between Robert Boyle and Thomas Hobbes concerning the validity of Boyle's famous air-pump experiments, they use Hobbes' critique of Boyle's experiments to lay bare the interests underlying them and the strategies that Boyle and his allies adopted to convince others of the experiments' reliability. They demonstrate how even such an apparently universal cornerstone of the scientific method as experiment has its origins in particular local and contingent circumstances. Making science local and examining the interests that underlie the articulation and resolution of scientific debates have been the preoccupations of constructivist history. Other pioneering studies include accounts of the ways that disillusionment with rational philosophy and mechanistic notions of causality in postwar Weimar Germany informed interpretations of quantum theory (Forman, 1971), and of the role social interests played in debates concerning the anatomical structure of the brain in the early nineteenth-century Edinburgh (Shapin, 1979). Such studies emphasize that such interests are constitutive rather than contingent features of even the most apparently abstract and technical forms of knowledge.

One important outcome of constructivist history is the recognition that even the issue of what kind of activity might count as science is often the product of locally articulated debates and controversies. Not only did Boyle and Hobbes disagree as to what might reliably be inferred from the outcome of air-pump experiments, but they also disagreed as to whether or not doing such experiments should be counted as a reliable way of generating scientific knowledge at all. Similarly, members of the scientific and the medical establishment who opposed the introduction of continental ideas concerning evolution and transmutation into early Victorian England did so both because they did not recognize the practices or the practitioners involved as being properly scientific and because they disagreed with the theories themselves (Desmond, 1989). As the labels 'science' and 'scientific' gained in cultural value, so too did different groups of knowledge producers disagree more vociferously as to what kinds of people and what kinds of activity should be accorded those labels (Gieryn, 1999). Constructivist historians have identified the business of policing the constantly changing boundaries of science as a crucial part of science's history. Through institutions, regimes of training, control of resources, and access to networks

of patronage, natural philosophers and scientists have struggled to circumscribe and discipline their own and others' knowledge-producing activities.

This localist historical perspective leads to recognition of the extent to which the production and validation of scientific knowledge are contingent upon more or less informal networks of trust. In order for particular groups or individuals to be considered reliable spokespersons for nature, they have to be trustworthy. They have to be the kinds of people who can be relied upon to tell the truth. This opens up the possibility of producing a cultural history of truth and objectivity. The marks of being trustworthy vary depending on time and place. In early modern England, for example, trustworthiness was considered the hallmark of the gentleman (Shapin, 1994). As a result, many of the informal (and more formal) rules of early modern gentlemanly behavior became enshrined as codes of conduct for the nascent scientific community as well. Some historians have argued that the increased scientific reliance on precision measurement and self-registering instrumentation since the beginning of the nineteenth century may be regarded as efforts to circumvent the need for trusting individuals with knowledge production (Schaffer, 1988). Others have pinpointed the increasing quantification of science as a means of overcoming the problem of trust in the production and particularly in the transfer of knowledge (Porter, 1995). The attention to local culture also emphasizes the historical importance of space and spatiality in the production and consumption of scientific knowledge (Livingstone, 2003).

The Material Culture of Experiment

Constructivist history of science pays particular attention to the process of experiment. Experimentation is taken to be a matter of human agents actively engaged in constructing the natural world around them (Gooding et al., 1989). The material culture of experiment is a crucial focus of attention. Historians have worked to recover the kinds of material resources required for successful experimentation. They have traced the emergence of the laboratory as a specialized space within which experiments take place. They have emphasized the ways in which successful manipulation of nature requires the experimenter to manipulate material and social resources successfully (Latour, 1988). Michael Faraday's discovery of electromagnetic induction, for example, involved as much his marshalling of the Royal Institution's considerable resources as it did his artful manipulations of materials in the laboratory (Morus, 1998). Constructivist historians have identified as well the important ways in which individual scientific reputations hinge upon recognition of the property relationship between experiment and experimenter. Priority disputes in science are frequent and virulent precisely because of the ways in which experiments embody the skills and reputations of their makers (Iliffe, 1992).

Seeing experiment as a skilled activity, constructivist historians have focused on understanding the historical origins of the kinds of skills required in order to carry out experiments. They have investigated how particular skills have been transferred into the laboratory from other settings, such as workshops or factories. Historical experiments have been replicated in an effort to better understand the tacit, unarticulated skills required for their successful performance. Such studies have shown how iconic experiments, such as James Prescott Joule's famous paddle-wheel experiments establishing the mechanical equivalent of heat during the 1840s, were contingent upon the technical know-how that Joule and his workers derived from their experiences in the family brewery (Sibum, 1995). Historians have also looked at the politics of skill in the laboratory and elsewhere, investigating the division and distribution of hierarchies of skilled activity within science at different times and places. Much attention has been devoted to the historical transmission of material resources and skilled personnel that accompanies the standardization of experimental practices. Work in this vein has shown how experiments move from being instances of highly skilled, difficult to replicate, and delicate procedures performed in highly circumscribed settings to being robust, standardized, and readily reproducible embodiments of group knowledge (Galison, 1987).

Apparently self-evident features of experimentation, such as precision measurement, are shown by constructivist historians to have a history. Measurement can be taken to be an expression of particular cultural values (Wise, 1995). The rise of metrology, for example, was an expression of science's increasing embeddedness in and contribution to a rapidly bureaucratizing and industrializing Western culture. Measurement could be regarded not only as a means of disciplining nature, rendering it more amenable to standardized routines of inquiry, but also as a means of disciplining the scientific workforce. Nineteenth-century scientific managers, such as George Biddell Airy, the English Astronomer Royal, regarded their laboratories and observatories as industrial workplaces and used regimes of routinized measurement and calculation to impose an industrial discipline on their workforces (Ashworth, 1998). In the late nineteenth-century laboratories at Cambridge or Glasgow, natural philosophers such as James Clerk Maxwell or Lord Kelvin saw in regimes of precision measurement in teaching and research, an expression of their affinity with industrializing society (Schaffer, 1992). Constructivist historians have investigated how adherence to cultural values like precision came to be taken as a hallmark of what it meant to be properly scientific. Such historians have also investigated the relationships between the material cultures of the laboratory and of fieldwork, showing how workers in the field, such as twentieth-century primatologists, have developed their own distinctive material and methodological cultures and disciplines (Rees, 2009). Constructivist historians have interrogated theory, using the analytical tools developed to understand experiment to look at theory as a form of cultural practice as well. From this perspective, processes of theory construction and dissemination are seen in terms of the pedagogical training and practical skills required to participate in highly specialized theoretical cultures (Warwick, 2003).

Science as Performance

One notable outcome of constructivist historical work has been to make the identity of the scientific practitioner itself problematic. Being a scientist is not a self-evident activity: the

word itself is of comparatively recent, early nineteenth-century origin. Scientific practitioners in the past occupied a number of different cultural niches. Constructivist historians have shown how natural philosophers at different times and places have been engaged actively in carving out social spaces for themselves and their work. Galileo's position at the Medici court in Renaissance Florence, for example, was not ready-made. Galileo had to work hard to promote the idea of a philosopher or mathematician courtier that he could then fulfill (Biagioli, 1993). He had to fashion himself into the kind of person who could easily inhabit that cultural niche and then to orient his work so that it would fit into the milieu of Italian Renaissance court life. The same could be said of Humphry Davy and his efforts to forge a place for the fashionable natural philosopher in Georgian London (Golinski, 1992). Constructivist approaches share with the 'new history' an emphasis on the important role of individual agents' artful engagement in constructing both themselves and their worlds.

Recent work has paid particular attention to the body of the scientific practitioner (Lawrence and Shapin, 1998). This is an outgrowth of the insistence on viewing scientific practices and the knowledge they produce as being embodied. Scientific practitioners at different times and places have needed to be particular kinds of human beings with specific bodily attributes. The ways in which those attributes were held to be best produced – whether by birth, self-discipline, or rigorous training – have been a focus of increasing historical attention. Historians have examined, for example, the relationship between bodily, mental, and moral discipline in Victorian electrical metrology (Gooday, 2004). The bodily interaction of practitioners with their experimental apparatus and the spatial organization of their laboratories have come to be recognized as playing an important role in the production of knowledge. The organization of such spaces and the kinds of material interactions between practitioners, and between them and their instruments, that have been taken to produce reliable knowledge are now seen as contingent and open to historical investigation.

Not only the identity and activities of the practitioner producing knowledge, but also the role of the audience for such productions is now recognized as constitutive. Audiences for scientific knowledge do not simply act as passive receivers of settled and unproblematic matters of fact. They have an important and active role to play in the articulation and resolution of controversies. During the early modern period and well into the nineteenth century, experiments were often performed before an audience. This was not simply a matter of demonstration. Members of the audience – prestigious and often named individually – played an important role in validating the knowledge produced. Their reputations could be used to underwrite the matters of fact they had witnessed. Audiences have participated in and responded to the processes of knowledge production through a variety of media, from direct witnessing to the virtual witnessing of reading a scientific paper at first hand, to the reception of second- or third-hand reports in museums, newspapers, reviews, or popular handbooks. Public scientific performances are therefore an important part of the dynamic process of knowledge production and validation. This draws attention to the visual culture of science too, and the historical construction of seeing as the paradigm of knowing (Morus, 2006).

Science at Work

This kind of history also looks at the many ways in which nature has been used historically as a model for social organization. The constitution of the natural economy has frequently been used as a powerful argument in favor of particular ways of ordering society (Roberts et al., 2008). Natural theology in the late eighteenth and early nineteenth centuries, for example, was widely used as an argument in favor of existing social hierarchies. Evolutionary theories were similarly cited to support new, meritocratic visions of society. The proper social economy was often held to mirror the natural order. Constructivist historians have shown how models of nature and society in the past were frequently mutually sustaining. They were produced together. Discussions in mid-nineteenth-century thermodynamics surrounding issues of work and waste in natural economies were coextensive with arguments concerning the proper organization of work and the minimization of waste in political economy (Smith and Wise, 1989). Proponents of new forms of industrial organization looked to nature for their models, while scientists looked to factory organization as a model for the best means of ordering the search for truth.

Different models for the organization of scientific work often expressed different understandings not only of how nature was ordered but also of the proper politics of knowledge production. Nineteenth-century scientific gentlemen, for example, regarded scientific work as a highly disciplined, hierarchical, and increasingly specialized activity. In principle, it was open to all. In practice, however, it could only be conducted properly either by, or under the supervision of, properly trained and vocationally committed 'gentlemen of science' (Rudwick, 1985). Other social groups had different notions of the politics of knowledge. The gentlemen's vision excluded women and artisans, for example, from anything other than a supportive role in making knowledge. Working-class men, on the other hand, could construct their own very different accounts of how knowledge making ought to be organized (Secord, 1994).

Through constructivist historical analysis, many of the traditional analytical categories employed by historians and sociologists of science have become recognized as being themselves in need of careful historical contextualization. Pronouncements by historical actors concerning matters such as the scientific method, experimentation, objectivity, or even more novel analytical categories, such as skill, are not to be taken at face value. Such pronouncements usually carry considerable freight. Deciding what kind of activity might properly be considered as an experiment, for example, is not straightforward. At different times, experimentation has been regarded as both a public and a private process. It has been characterized as both a fundamentally communal and an essentially individual activity. Again, what it means to be a skilled worker has changed a great deal since the early modern period. Skill, like method, or even science itself, is a historical construction. By looking at science as a preeminent cultural activity, constructivist histories of science have made it possible

to raise a whole host of new questions concerning the ways in which human beings, in the past, have tried to make sense of the world around them. Rather than seeing science as a static, monolithic edifice, constructivist history recognizes it as a constantly changing process of production, reproduction, and consumption; its boundaries continually shifting and its meaning always open to renegotiation.

See also: Chemical Science: History and Sociology; Citizen Science; Constructivism and Constructionism: Methodology; Experiment in Science and Technology Studies; History and the Social Sciences; History: Overview; Physical Sciences: History and Sociology; Science, History of; Scientific Knowledge, Sociology of; Social Constructivism.

Bibliography

Ashworth, W., 1998. John Herschel, George Airy, and the roaming eye of the state. History of Science 36, 151–178.

Biagioli, M., 1993. Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago.

Desmond, A., 1989. The Politics of Evolution: Morphology, Medicine, and Reform in Radical London. University of Chicago Press, Chicago.

Forman, P., 1971. Weimar culture, causality, and quantum theory, 1918–1927: adaptation by German physicists and mathematicians to a hostile intellectual environment. Historical Studies in the Physical Sciences 3, 1–115.

Galison, P., 1987. How Experiments End. University of Chicago Press, Chicago.

Gieryn, T., 1999. Cultural Boundaries of Science: Credibility on the Line. University of Chicago Press, Chicago.

Golinski, J., 1992. Science as Public Culture: Chemistry and Enlightenment Britain, 1760–1820. Cambridge University Press, Cambridge.

Gooday, G., 2004. The Morals of Measurement: Accuracy, Irony and Trust in Late Victorian Electrical Practice. Cambridge University Press, Cambridge.

Gooding, D., Pinch, T., Schaffer, S. (Eds.), 1989. The Uses of Experiment: Studies in the Natural Sciences. Cambridge University Press, Cambridge.

Iliffe, R., 1992. In the warehouse: privacy, property and priority in the early Royal Society. History of Science 30, 30–68.

Latour, B., 1988. The Pasteurization of France. Harvard University Press, Cambridge, MA.

Lawrence, C., Shapin, S. (Eds.), 1998. Science Incarnate: Historical Embodiments of Natural Knowledge. University of Chicago Press, Chicago.

Livingstone, D., 2003. Putting Science in Its Place: Geographies of Scientific Knowledge. University of Chicago Press, Chicago.

Morus, I.R., 1998. Frankenstein's Children: Electricity, Exhibition and Experiment in Early Nineteenth-Century London. Princeton University Press, Princeton, NJ.

Morus, I.R., 2006. Seeing and believing science. Isis 97, 101–110.

Porter, T., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.

Rees, A., 2009. The Infanticide Controversy: Primatology and the Art of Field Science. University of Chicago Press, Chicago.

Roberts, L., Schaffer, S., Dear, P. (Eds.), 2008. The Mindful Hand: Inquiry and Invention from the Late Renaissance to Early Industrialization. University of Chicago Press, Chicago.

Rudwick, M., 1985. The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists. University of Chicago Press, Chicago.

Schaffer, S., 1988. Astronomers mark time: discipline and the personal equation. Science in Context 2, 115–145.

Schaffer, S., 1992. Late Victorian metrology and its instrumentation: a manufactory of ohms. In: Bud, R., Cozzens, S. (Eds.), Invisible Connections: Instruments, Institutions and Science. SPIE Optical Engineering Press, Bellingham, WA, pp. 23–56.

Secord, A., 1994. Science in the pub: artisan botanists in early nineteenth century Lancashire. History of Science 32, 269–315.

Shapin, S., 1979. The politics of observation: cerebral anatomy and social interests in the Edinburgh phrenology disputes. In: Wallis, R. (Ed.), On the Margins of Science: The Social Construction of Rejected Knowledge. University of Keele Press, Keele, UK.

Shapin, S., 1994. A Social History of Truth: Civility and Science in Seventeenth-Century England. University of Chicago Press, Chicago.

Shapin, S., Schaffer, S., 2011[1985]. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life (first ed., 1985), second ed. Princeton University Press, Princeton, NJ, pp. 139–178.

Sibum, H.O., 1995. Reworking the mechanical value of heat: instruments of precision and gestures of accuracy in early Victorian England. Studies in History and Philosophy of Science 26, 73–106.

Smith, C., Wise, N., 1989. Energy & Empire: A Biographical Study of Lord Kelvin. Cambridge University Press, Cambridge.

Warwick, A., 2003. Masters of Theory: Cambridge and the Rise of Mathematical Physics. University of Chicago Press, Chicago IL.

Wise, N. (Ed.), 1995. The Values of Precision. Princeton University Press, Princeton, NJ.

Science Education

John K Gilbert, King's College London, London, UK; The University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.C. Linn, volume 23, pp. 13668–13673, © 2001, Elsevier Ltd.

Abstract

The evolution of formal school science education, in response to changing attitudes toward science, is summarized. The content and structure of the curriculum, including the relationships to technological applications and to the nature of science, are presented. Assumptions about learning, assessment, and science teacher education are given. The evolution, the nature and provision, of informal science education is summarized. The nature and use of research into both formal and informal science education are briefly discussed.

Science education refers to the teaching and learning of the ideas produced by the sciences, to the ways in which these are produced, ratified, and accepted, by the community of scientists, and to the implications of the uses to which this knowledge are or could be put. It is provided both formally, that is within a framework of a prescribed curriculum and a specialized institution (the school), and informally, that is extracurricularly and in a variety of physical situations. The development of formal science education has been one of the most significant events in school education over the last 170 years, whilst informal provision has only recently come into prominence. Some features of both can be illustrated largely with respect to England.

The Provision of Formal Science Education

The Development of the Provision of Formal Science Education

The early inclusion of science education as an option in the curriculum of some schools took place in the mid-nineteenth century in Western Europe and North America. For example, in England, the subject was being taught in some fee-paying private secondary (high) schools in the 1850s, perhaps in recognition of the growing importance of science in industrial production (Layton, 1973). By 1870, there was discussion at national level about what should be included in the science curriculum and efforts were begun to systematically train science teachers. From then on, the provision of science education expanded progressively but unevenly until it included all primary (elementary) and secondary (high) schools and all pupils, together with the systematic training and certification of science teachers (Jenkins, 1979). The increased involvement of the state led to a national school science curriculum becoming mandatory in 1988, during the subsequent evolution of which the influence of the academic science and science education communities decreased considerably (Fensham, 2013). This sequence of events has been approximately replicated throughout the world, albeit to different degrees and at different speeds.

The Evolution of Formal Science Education

Changing Purposes

At its inception, there was tension with respect to the aims of school science education. Was it to provide 'citizen science,' that is an education that would deal with the contribution of science to the everyday personal, social, economic, and cultural concerns of the general public? Or was it to provide 'prevocational' science, a basic knowledge of the most important facts, concepts, and processes of science as a preparation for students to go on to advanced study of a science at university or technical institute? The prevocational emphasis came to dominate the debate, but was continuously questioned for a number of reasons. First, as the provision of science education expanded and became compulsory to below the minimum school-leaving age, an ever-lower proportion of students in a given school class were interested in higher education in a science. Second, the rapid expansion of output from scientific research meant that the prevocational school science curriculum gradually became overloaded with content, such that students were introduced to isolated facts and concepts, being increasingly unable to see the relation between them and their significance. Third, the growing impact of science on the everyday lives of students meant that the traditional presentation of ideas as largely devoid of their applications and implications become ever less interesting to many students. Fourth, the pedagogy most commonly adopted was based on the assumption that the duty of the teacher was to present information in an efficient manner and that of the students was to memorize a mental 'copy' of it that was to be reproduced in examinations. This approach proved unsuccessful, in that only a small proportion of students were ever successful in fully meeting its demands. Taken together, these factors led to many students developing negative attitudes not only to science education, which meant that they did not continue to study it when given the option not to, but also to science itself, which was a matter of concern for the community of scientists and industrialists.

Attitudes to Science

Attitudes toward science and science education in schools are manifest in the active engagement of students in science lessons and in their interest in seeking a career in science or technology. Enquiries into such attitudes have shown them to be influenced by many factors, including their personal responses to classroom activities; self-assessment of actual and potential success in science-related activities; perceptions of the quality of the science teacher(s) encountered; and reaction to the attitudes of their peers and parents toward both science and science

education. Evaluating the research that has been done into these factors is not easy. Are the identified attitudes toward science/science education in general or are they focused on individual science subjects, e.g., physics, chemistry, biology, and their presentation in the curriculum?

Students' attitudes toward school science seem to decrease from the start of formal schooling (Pell and Jarvis, 2001) and to continue throughout the secondary (high) school phase (Yager and Penick, 1986), although it does seem that biology is rated more positively than the physical sciences (Osborne and Collins, 2001). A worldwide study (Sjoberg and Schreiner, 2010) showed that such attitudes declined throughout the span of schooling in industrialized counties (but not in developing countries), with that of girls being particularly steep. Other studies have thrown some light on the influence of specific factors, e.g., the family (Archer et al., 2012) and ethnicity (DeWitt et al., 2010), on these trends.

Content and Organization

Although there had long been calls for curricular change (e.g., Schools Council, 1970), which had been proceeding unevenly over the years, it was the report 'Beyond 2000' (Millar and Osborne, 1998) that called, amongst many things, for an emphasis on key 'explanatory stories'; the nature of scientific inquiry; the relationships between science and technology and their implications for everyday life. The influential Rocard Report (Rocard, 2007) set these suggestions in a Europe-wide context and added emphases on the relationship between the formal and informal provision of science education and, particularly, on that for girls.

The progressive implementation of these changes has been underpinned by three developments. The first of these was the clarification of the distinction between the two types of curriculum: that based on 'Vision 1,' drawing heavily on orthodox science and intended to provide prevocational science education; that based on 'Vision 2,' drawing heavily on those aspects of the everyday lives of students for which science had relevance (Roberts, 2007). The second of these has been the development of 'context-based' courses, in which the learning of key concepts is framed within the study of situations, problems, events, that have meaning for the students (Gilbert, 2006). The third of these was the focus on 'socio-scientific issues' in which the emphasis was on the social aspects of the nature of scientific inquiry and on the implications of science for the lives of individuals and societies (Sadler, 2009). Taken together, these three developments are claimed to have successfully addressed the challenges faced by science education although they have been rarely implemented in full and therefore cannot have been thoroughly researched.

Little, if anything, has been written about the practical organization of formal science education since the introduction of the national curriculum. While the provision of 'separate subjects' (i.e., physics, chemistry, biology) still takes place, particularly toward the end of the high school period, there is little evidence of the provision of 'integrated science' courses, the majority being 'modular,' with themes from the separate sciences being interspersed. Whilst it is common for all children of primary (elementary) school age to experience the same curriculum, the continuation of that format in the high (secondary) school soon gives way to 'core' courses, supported by 'optional' courses, enabling students following 'Vision 1' courses' to be progressively distinguished from the majority taking 'Vision 2' courses.

Relationship to Technology

Science can be seen as the search for understanding of the world as experienced. Technology can be seen as attempts to meet the practical needs of individuals that arise in a wide range of different physical, social, and cultural milieus. The meaning of the latter has been obscured somewhat by the North American misuse of the word 'technology' as equating with 'computers.' Science and technology as activities are related, albeit in a number of different ways (Gardner, 1994), all of which have implications for how the two activities are related in the school curriculum.

Historically, most school science teachers have viewed technology as applied science. This has led technology education to be seen as the addition of examples of technologies as appendices to topics in the science curriculum. Although this view does have a very limited validity, there have been calls to take a broader perspective on the relationship (Cajas, 2001). One solution has been to introduce the separate subject of technology into the curriculum, called 'Design and Technology' in England and Wales. However, this approach seems unlikely to succeed in the long run because of the time constraints arising from the complex relation between the concepts used in the two subjects, from the length of the school day (technology activities require sustained periods of time), from the lack of clear recruitment lines for suitable teachers from higher education, and because of the large investment in physical plant required, all of which were anticipated by Layton (1993). An approach which seems more likely to succeed is through the provision of modules, prepared by working groups, in which science and technology are brought to bear on themes about contexts of specific socio-scientific interest. An example is the 'Nature, Life and Technology' course available to both Vision 1 and Vision 2 audiences in the Netherlands in the form of 66 modules, prepared by groups of school and higher education teachers, and adopted by 44% of all senior high schools (NLT, 2008).

Nature of Science

Calls for school science education to place a much greater emphasis on how scientific knowledge is generated and validated are of long-standing duration (e.g., Schools Council, 1970). The failure to do so that has resulted reflects both the tenacity of 'folk knowledge' about science by both teachers and students and the lack of agreement about what the phrase actually encompasses and how it might be taught. Folk knowledge about science stems from two sources: first, from the persistence of 'indigenous knowledge,' i.e., those explanations about nature that stem from tenacious belief systems other than those of science (Cobern and Loving, 2000); and second, from the misrepresentations of nature of science that persist in textbooks, e.g., the ideas that scientific knowledge is produced by a process of induction from the results of

empirical enquiry, that scientific theories are truthful and unalterable, and that scientists work alone. The persistence of such ideas stems from the failure of philosophers, historians, and sociologists to come to any sort of agreement on the meaning of the phrase 'nature of science,' both amongst and between themselves. This uncertainty goes unchallenged as few future science teachers are required to take any courses on nature of science.

The currently available consensus view, which avoids major disagreements and which is used in much science education research, is that science knowledge is subject to change; has to account for empirical observations; is influenced by scientists' personal views; is proposed by acts of imagination; and in validated by discussion between scientists (AAAS, 1990). Two broad approaches to teaching nature of science have emerged over the years: providing students with opportunities to conduct scientific enquiries and of then reflecting on what they had done; and direct instruction in the main ideas of the consensus model. A review of many studies came to the conclusion that it is a blended amalgam of the two approaches that is the most effective approach (Abd-El-Khalick and Lederman, 2000).

Models and modeling is one aspect of nature of science that has made good educational progress. A scientific model is a representation of an idea, an object, an event, or a process, initially produced for a specific purpose in research, the ontology of 'model' having been explored (Gilbert and Boulter, 2000). Early work showed that the 'nature of model' was not well understood by either teachers (e.g., Crawford and Cullin, 2005) or students (e.g., Grosslight et al., 1991). Systematic interventions to remediate this situation proved very satisfactory in respect of teachers (e.g., Justi and Van Driel, 2005) and students (e.g., Raghavan and Glaser, 1995). Approaches to the development of modeling skills have been proposed (e.g., Justi and Gilbert, 2002) and successfully implemented (e.g., Menonca and Justi, 2013).

A second aspect of nature of science that is making progress – it is not clear how rapidly and successfully – concerns practical work in the context of inquiry-based teaching and learning. Since the outset of formal science education, laboratory work has been defended as a major component of the subject. It has been included in the curriculum for a number of reasons: to provide opportunities to acquire conceptual knowledge, to develop the skills of empirical inquiry, to introduce students to some aspects of how scientists actually go about their work and, perhaps most importantly, to engage the interest of students in the subject. It was perhaps because this array of objectives was indiscriminately addressed that the outcome was a high degree of uncertainty over the educational value of practical work as such (Hofstein and Lunetta, 1982). This is unfortunate given the high cost of providing laboratories and equipment. In the last decade or so, 'inquiry-based teaching and learning,' which implies a central role for the laboratory, has become a major element in the national curricula of all countries.

Implementing inquiry-based learning – attempting to replicate the activities of scientists in schools – is only slowly and hesitantly being introduced, for a number of reasons. There is no clear agreement on what that replication entails, although Bell et al. (2010) have suggested that it consists of nine elements in varying order: posing a question; generating a hypothesis; planning the inquiry; carrying out the investigation; analyzing and interpreting the data obtained; producing a model of the phenomenon; drawing conclusions as to the answer to the question posed; communicating that answer to others; and making predictions about the phenomenon based on that answer. In addition to the major problems of identifying phenomena that are suitable for inquiry within the restricted resources available in schools, the behavioral norms and social roles in many schools militate against such an approach to learning. For example, teachers will have to avoid taking the verbal initiative in classes too often, an approach that many will find difficult (Van Zee, 2000), while students will have to acquire the ability to discuss scientific issues meaningfully amongst themselves (Roychoudhury and Roth, 1996). The more authoritarian educational systems will find the introduction of inquiry-based teaching, as many are attempting to do, particularly challenging. In all cases, it may well be desirable to change existing procedures gradually.

Assumptions about Learning and Their Implications for Teaching

The traditional view about how learning takes place, in science as elsewhere in the school curriculum, has been based on the 'behaviorist' psychology of Skinner (Skinner, 1971). This assumed that a student was a '*tabula rasa*' ('blank slate') with respect to a specific idea until it had been systematically taught. That teaching would involve the careful and structured presentation of the new ideas, requiring their recitation by students until a perfect 'copy' was obtained, this process being supported by the measured use of rewards and punishments. Despite its rigorous and widespread application, attempts at learning by many science students have been unsuccessful when they have followed its precepts. In the late 1970s it gradually emerged that, in respect of many ideas that formed part of the science curriculum, students already had some understandings, albeit somewhat idiosyncratic in nature from the point of established science, before they were taught a specific idea. These came to be called 'misconceptions' or 'alternative conceptions' and were widespread, even universal, in occurrence (Gilbert and Watts, 1983) and they served to 'interfere' with the teaching being provided.

The three main alternative approaches to psychology came into prominence to account for these 'alternative conceptions.' They share the assumption that people learn by being mental active, not passively as behaviorism had assumed, and all provide accounts for how new knowledge is related to what is already known. They differ both in the relative importance of the individual in that process and that of the surrounding society.

The long-established ideas of Piaget (1929), which underpin all manifestations of what came to be called the 'constructivist movement,' place considerable emphasis on the role of the individual in learning, but do recognize the great educational value of interaction with the surrounding environment, of which other people form an important segment. However, much of his extensive empirical work was framed so as to provide support for his basic tenet: that the quality of understanding achieved was constrained by an individual's general 'stage' (as he put it) of mental capability at any one time, this

evolving roughly stepwise over the years of childhood and adolescence. The implication for teaching was that the demands made of the students could not exceed the capability of their current 'stage.' It was not clear what could be done to enable students to reach the next 'stage.' It was the rigidity of the 'stage' caveat, shown to be of doubtful validity, e.g., by Driver (1978), that led to its gradual replacement as the basis for research, initially by the ideas of Kelly and subsequently by those of Vygotsky.

Kelly (1970) focused on the individual as the major agent in learning; hence, his view is called 'personal construct psychology.' As he put it:

> Constructive alternativism holds that man (sic) understands himself, his surroundings and his potentialities by devising constructions to place upon them and then testing the tentative utility of these constructions against such *ad interim* criteria as the successful prediction and control of events.
>
> p. 1

He thus rejected any notion of 'stages' in an individual's development, seeing the pragmatic inclination to engage in prediction plus the testing of ideas as determining what can be learnt. He did accept the significance of other people in the learning process, with the caveat that:

> To the extent that one person construes the construction process of another he may play a role in his social process involving the other person.
>
> p. 2

With the 'stage' idea seemingly discounted, the social dimension to learning moved center stage in the 'social constructivist' approach of Vygotsky (1986). This saw learning as a process of social enculturation of a learner by interaction with an authority, e.g., a teacher. Only that knowledge which lies within the 'zone of proximal development' of the learner, a phrase the meaning of which is much debated, can be learned. One major consequence for teaching of the Kellian and Vygotskian approaches has been the acceptance that, for effective learning to take place, the teacher must identify what a student already knows about a subject and relate to that when introducing new knowledge. This has led to the development of clear procedures for identifying and remediating dysfunctional 'alternative conceptions' (Vosniadou et al., 2007). Meanwhile, efforts to locate students' zones of proximal development have been one of the major purposes for assessment.

Assessment

'Formative assessment' is the identification of students' knowledge about a given topic, the intention being to subsequently take action to improve that understanding. Teachers carry it out continuously during lessons, mainly through verbal questioning, and amend their subsequent teaching in the light of what they find out (Black, 1998). For formative assessment to be valid, the students must trust that the information they reveal – deliberately or accidently – will be used for their benefit. Alas, there is a paradox here, for such information can also be used to comparatively evaluate the performance of the students, which is 'summative assessment' and which is the lynchpin of institutional accountability (Bell and Cowie, 2001). Maintaining a reputation for integrity with students – a condition for retaining their trust – poses a professional dilemma for science teachers.

Whilst all educational systems include public examinations at specific intervals in the educational process, there has been a recent growth of very large-scale international systems of student assessment. The 'Trends in International Mathematics and Science Study' (TIMSS) of students in the fourth and eighth Grade, was first administered in 1995 and at 4-year intervals thereafter. It is focused on what students know of the concepts of a common core curriculum in the two subjects: an entry into the archive is provided by TIMSS (2013). The 'Programme for International Student Assessment' (PISA) study, organized by the Organisation for Economic Co-operation and Development (OECD), involves eleventh Grade students and is conducted every 3 years, starting in 2000. PISA examines the ability of students to relate what they have learnt at school to real-life situations: see OECD (2013). Both systems publish their results country by country. Students of a few countries/districts, e.g., Singapore, Hong Kong, South Korea, always do very well in these surveys. All the countries taking part in these schemes view their 'league position' as a matter of national pride and introduce reforms intended to emulate the schools of the most successful countries, i.e., in reality the increased use of behavioral psychology. It is a matter of interest that, whilst this is going on, the most 'successful' countries are moving away from such approaches.

Science Teacher Education

University graduates who become science teachers do so from a wide variety of academic backgrounds. Given that national curricula are intended to provide a fairly uniform provision of education, intending and actual science teachers normally go through a process of professional development which may be based either in a university or in a school. Whilst that education is often divided into two sections, a compulsory 'preservice' phase and a less rigidly managed 'in-service' phase, a strong case can be made for a more coordinated provision.

Such a provision can be provided with respect to three components: first, the 'professional' element, directly concerned with teaching and learning; second, the 'personal' element, concerned with the individual's perception of him-/herself as a teacher; and third, the 'social' element, concerned with the individual's perception of him-/herself as a member of a community of teachers. An orchestrated program of 'teacher professional development' can provide an evolution of professional identity across the pre- and in-service phases: in well-funded environments, this does take place (Bell and Gilbert, 1996).

That provision should be concerned with four themes. First, the teacher's knowledge of the subject(s) they are to teach. Many teachers have to work in schools outside their specialized field of undergraduate study. Moreover, they may not have a suitable depth of understanding of key concepts, for these

may not have been revisited since their own school days. Second, they will need to be introduced to the wide range of teaching techniques that are now in use. Third, they will need help in adapting concepts into a form that students can understand: they will have to deploy their so-called 'pedagogic content knowledge' (Shulman, 1987). Fourth, it will be necessary to ensure that their personal beliefs about science education are compatible with the assumptions of the curriculum they will have to teach.

Successful science teachers – those who remain in the profession for many years and enjoy the experience – will have had such an experience (Gilbert, 2010).

Research in Formal Science Education

The established pattern of research in science education – as in the rest of education – makes use of two contrasting approaches. Qualitative research, small scale, often case study, is intended to establish the nature of educational events and to identify concepts that can be used to represent them. Quantitative research, large scale, is intended to establish the incidence of educational events and the relationships between the concepts that represent those events. 'Mixed method' research uses a combination of these two types. For a wide variety of reasons, neither policy makers nor practicing teachers make much use of published science education research (Levin, 2013). Recently, researchers have begun to use 'design-based research' as the organizational framework for their activities (Cobb et al., 2003). Design-based research rests on five guiding principles. First, it seeks to solve real-world problems by making interventions into educational systems and evaluating their outcomes. Second, it is theory driven, intending that theory will be advanced. Third, it is interactive, in that the work is carried in a partnership between researchers and practitioners. Fourth, it uses a variety of existing research methodologies. Fifth, the outcomes of such work are context dependent, such that reports of what was done and why it was done must include full accounts of the educational situations in which it took place (Wang and Hannafin, 2005). It will be some time before enough examples of such work will have been published and conclusions drawn about whether it overcomes the established problems of the use of research.

The Provision of Informal Science Education

The Rise of Informal Science Education

Informal science education was first provided for adults in the nineteenth century (Layton, 1973). The development of the notion of 'museum' and the availability of cheap paper-based publications were major factors in extension of the genre, first into family-based education, and then directly into the education of school-age students.

Provided for a wide variety of purposes and in an ever-increasing range of ways, informal science education, in its purist form, has a number of characteristics that differentiate it from formal provision. It is voluntary, in that the learner decides what to learn, when, and how; based on experiences and themes that are presented without structure or sequence, so that the learner has to decide what to do; open-ended, in that the learner decides when to start and when to stop; and not certificated, so that any assessments made are only for the benefit of the learner (Wellington, 1990).

With respect to school-age learners, the increased incidence of informal science education can be attributed to the interplay of a range of factors. Science and technology advance at an ever-faster rate, producing ideas and artifacts that are of interest to young people. The development of the formal science curriculum cannot keep up, not least because the new ideas require elaborate and expensive illustrations and equipment, which national educational systems are unable, or unwilling, to afford. Private agencies will afford to do so, often in anticipation of being able to charge fees for access. The element of voluntary choice and decision by a young person leads to their engagement in learning, thus counteracting declining attitudes to science education. Perhaps most importantly of all, a young person can choose to actively learn with somebody else – family member or friend – whilst the formal education system largely prohibits this. Whilst some informal resources are based on behaviorist principles, for example when the direct transmission of factual knowledge is involved, the great majority make use of social constructivist principles, requiring the active engagement of the learner – alone or with others – in order for the experience to advance. This flexibility in terms of psychology of learning has led to a wide range of types of provision. In the next two sections, a sketch of the main forms will be given.

Provision through Conventional Media

Provision of opportunities for science education by museums, botanical gardens, and zoos can be grouped together because they all contain original objects, inanimate or animate, about which the visitor is invited to construct a narrative with the aid of a brief explanatory text. In the case of museums, an innovation in the last 20 years has been the addition of the 'science center,' in which exhibits require visitors to take particular actions and then providing feedback on the consequences of the actions taken (Rennie and McClafferty, 1996). Members of this grouping provide the maximum opportunity for visitors to learn together in self-organized groups.

Provision through 'popular books,' newspapers, and magazines may be grouped together because they are all paper based and accessible through a wide variety of outlets, e.g., newsagents, supermarkets, and train stations (Afonso and Gilbert, 2013). Their great attraction is that their timing of use is entirely in the hands of the learner.

Provision through Electronic Media

This large and very important grouping includes radio, television, the Internet, and computer games, all provided through electronic media. Of the long-standing media, radio is of particular value in providing access to learning for very low-density populations, e.g., Australia, or for very large and culturally complex countries, e.g., India. The advent of television so soon after the introduction of national radio systems has so far blighted the development of this medium for the provision of science education (Mazzonetto et al., 2005). Film and television may be grouped together because their content

and range of program types overlap, e.g., documentaries, news broadcasts, and movies. Television, in particular, has been found to be a valuable source of information and understanding (Dhingra, 2006). Of all the electronic-based media, the newest, the Internet (Hsi, 2007) and computer games (Connelly et al., 2012) are probably having the greatest impact and developing at the fastest rate.

Historically, there has been relatively little research done into the use and impact of informal learning resources, both in general and in science education in particular. The book *Learning Science in Informal Environments* is the most comprehensive account to date (Bell et al., 2009).

The Relationship between Formal and Informal Science Education

Because of the general underfunding of development for the formal science curriculum, an increasing use is being made of informal resources by co-opting them into the formal system. Where 'school visits' are made, for example to a science center, the greatest impact on learning is achieved by careful timing of the visit in relation to the coverage of the curriculum. However, it seems that most use of such recourses is made by students at their own initiative, thus retaining many of the elements of genuine informal learning (Stocklmayer et al., 2010).

In educational terms, the relative significance of the formal and informal systems will swing toward the latter, given the entrepreneurial initiative that drives the latter and the bureaucratic inertia that infests the former.

See also: Gender and Education; Health Education and Health Promotion; Mathematics Education; Science and Technology, Anthropology of.

Bibliography

AAAS, 1990. Science for All Americans. Oxford University Press, New York.

Abd-El-Khalick, F., Lederman, N., 2000. Improving science teachers' conceptions of the nature of science: a critical review of the literature. International Journal of Science Education 22, 665–701.

Afonso, A., Gilbert, J.K., 2013. The role of 'popular' books in informal chemical education. International Journal of Science Education (B) 3 (1), 77–99.

Archer, L., De Witt, J., et al., 2012. Science aspirations and family habitus: how families shape children's engagement and identification with science. American Educational Research Journal 49 (5), 881–908.

Bell, B., Cowie, B., 2001. Formative Assessment in Science Education. Kluwer, Dordrecht, The Netherlands.

Bell, B., Gilbert, J., 1996. Teacher Development: A Model from Science Education. Falmer, London.

Bell, B., Lewenstein, B., et al., 2009. Learning Science in Informal Environments: People, Places, and Pursuits. The National Academies Press, Washington, DC.

Bell, T., Urhahne, D., et al., 2010. Collaborative inquiry learning in schools: models, tools, and challenges. International Journal of Science Education 32 (3), 349–377.

Black, P., 1998. Testing: Friend or Foe? Routledge, London.

Cajas, F., 2001. The science/technology interaction: implications for science literacy. Journal of Research in Science Teaching 38, 715–729.

Cobb, P., Confrey, J., et al., 2003. Design experiments in educational research. Educational Researcher 32 (1), 9–13.

Cobern, W., Loving, C., 2000. Defining 'science' in a multicultural world: implications for science education. Science Education 80, 579–610.

Connelly, M., Boyle, E., et al., 2012. A systematic literature of empirical evidence on computer games and serious games. Computers & Education 39, 661–686.

Schools Council, 1970. Changes in School Science Teaching. Evans/Methuen, London.

Crawford, B., Cullin, M., 2005. Supporting prospective teachers' conceptions of modelling in science. International Journal of Science Education 26 (11), 1379–1401.

DeWitt, J., Osborne, J., et al., 2010. High aspirations and low progression: the science aspirations-career paradox amongst minority ethnic students. International Journal of Science and Mathematics Education 9 (2), 243–271.

Dhingra, K., 2006. Science on television: storytelling, learning and citizenship. Studies in Science Education 42, 89–123.

Driver, R., 1978. When is a stage not a stage?: A critique of Piaget's theory of cognitive development and its application to science education. Educational Research 21 (21), 54–61.

Fensham, P., 2013. The science curriculum: the decline of expertise and the rise of bureaucratise. Journal of Curriculum Studies 45 (2), 152–168.

Gardner, P., 1994. Representations of the relationships between science and technology in the curriculum. Studies in Science Education 24, 1–28.

Gilbert, J.K., 2006. On the nature of 'context' in chemical education. International Journal of Science Education 28 (9), 957–976.

Gilbert, J.K., 2010. Supporting the development of effective science teachers. In: Osborne, J., Dillon, J. (Eds.), Good Practice in Science Teaching. Open University Press, Maidenhead, UK, pp. 274–300.

Gilbert, J.K., Boulter, C.J., 2000. Developing Models in Science Education. Kluwer, Dordrecht.

Gilbert, J.K., Watts, D.M., 1983. Conceptions, misconceptions, and alternative conceptions: changing perspectives in science education. Studies in Science Education 10, 61–98.

Grosslight, L., Unger, C., et al., 1991. Understanding models and their use in science: conceptions of middle and high school students and experts. Journal of Research in Science Teaching 28 (9), 799–822.

Hofstein, A., Lunetta, V., 1982. The role of the laboratory in science teaching: neglected aspects of research. Review of Educational Research 52, 201–217.

Hsi, S., 2007. Conceptualising learning from the everyday activities of digital kids. International Journal of Science Education 29 (12), 1509–1529.

Jenkins, E.W., 1979. From Armstrong to Nuffield. Murray, London.

Justi, R., Gilbert, J.K., 2002. Modelling, teachers' views on the nature of modelling and implications for the education of modellers. International Journal of Science Education 24 (4), 369–387.

Justi, R., Van Driel, J., 2005. The level of science teachers' knowledge about models and modelling: promoting, characterising, and understanding the process. International Journal of Science Education 27 (5), 549–573.

Kelly, G., 1970. A Brief Introduction to Personal Construct Theory. Perspectives on Personal Construct Theory. Wiley, New York, pp. 1–14.

Layton, D., 1973. Science for the People. George Allen & Unwin, London.

Layton, D., 1993. Technology's Challenge to Science Education. Open University Press, Buckingham, UK.

Levin, B., 2013. To know is not enough: research knowledge and its use. Review of Education 1 (1), 2–31.

Mazzonetto, M., Merzagora, M., et al., 2005. Science in Radio Broadcasting. Polimetrica, Milan, Italy.

Menonca, P., Justi, R., 2013. The relationship between modelling and argumentation from the perspective of the model of modelling diagram. International Journal of Science Education 35 (14), 2407–2434.

Millar, R., Osborne, J., 1998. Beyond 2000. King's College London, London.

NLT, S.C., 2008. Outline of a New Course in the Sciences. From: http://www.betavak-nlt.nl/English.

OECD, 2013. Programme of International Student Assessment. From: http://www.oecd.org/pisa/pisafaq (retrieved 01.11.13.).

Osborne, J., Collins, S., 2001. Pupils' views of the role and value of the science curriculum: a focus group study. International Journal of Science Education 23 (5), 441–468.

Pell, T., Jarvis, T., 2001. Developing attitude to science scales for use with children of ages from five to eleven years. International Journal of Science Education 23, 847–862.

Piaget, J., 1929. The Child's Conception of the World. Routledge and Kegan Paul, London.

Raghavan, K., Glaser, R., 1995. Model-based analysis and reasoning in science: the MARS curriculum. Science Education 79 (1), 37–62.

Rennie, L., McClafferty, T., 1996. Science centres and science learning. Studies in Science Education 27, 53–98.

Roberts, D., 2007. Scientific literacy/science literacy. In: Abell, S., Lederman, N. (Eds.), Handbook of Research on Science Education. Erlbaum, Mahwah, NJ, pp. 729–780.
Rocard, M., 2007. Science Education NOW: A Renewed Pedagogy for the Future of Europe. European Commission, Brussels.
Roychoudhury, A., Roth, W.-M., 1996. Interactions in an open-enquiry physics laboratory. International Journal of Science Education 18 (4), 423–445.
Sadler, T., 2009. Situated learning in science education: socioscientific issues as contexts for practice. Studies in Science Education 45, 1–42.
Shulman, L., 1987. Knowledge and teaching: foundations of the new reform. Harvard Educational Review 57 (1), 1–22.
Sjoberg, S., Schreiner, C., 2010. The ROSE Project: An Overview and Key Findings. University of Oslo, Oslo, Norway.
Skinner, B., 1971. Beyond Freedom and Dignity. Jonathan Cape, London.
Stocklmayer, S., Rennie, L., et al., 2010. The roles of the formal and informal sectors in the provision of effective science education. Studies in Science Education 46 (1), 1–44.
TIMSS, 2013. TIMSS & PIRLS. From: http://timss.bc.edu (retrieved 01.11.13.).
Van Zee, E., 2000. Analysis of student-generated inquiry discussion. International Journal of Science Education 22 (2), 115–142.
Vosniadou, S., Baltas, A., et al., 2007. Reframing the Conceptual Change Approach to Learning and Instruction. Elsevier, Oxford, UK.
Vygotsky, L., 1986. Thought and Language. MIT Press, Cambridge, MA.
Wang, F., Hannafin, M., 2005. Design-based research and technology-enhanced learning environments. Educational Technology Research and Development 53 (4), 5–23.
Wellington, J., 1990. Formal and informal learning in science: the role of interactive science centres. Physics Education 25, 247–252.
Yager, R., Penick, J., 1986. Perceptions of four age groups towards science classes, teachers, and the value of science. Science Education 70 (4), 355–363.

Science, Geography of

Diarmid A Finnegan, Queen's University, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Interest in the geography of science has increased significantly over the past 20 years to create a highly interdisciplinary field of inquiry. Rejecting the spatial determinism of earlier work, geographers and many others have developed insights from the sociology and philosophy of science and from social and cultural theory to examine the spatiality of science in all of its diversity and complexity. Particular attention has been paid to the sites at which scientific knowledge has been produced, assessed, and consumed and to the transmission of science over physical, cultural, and cognitive terrain.

Introduction

Work on the geography of science over the past 20 years or so has staked out a now flourishing area of academic research (Livingstone, 2003). Scholarship on the historical geography of science has been among the most vigorous in this burgeoning field. More recently, geographers and others interested in the development of technoscience in the recent past have added to our understanding of the spatiality of scientific knowledge and praxis. This diverse body of work has been guided by a shared assumption that determining where scientific knowledge is acquired, assessed, and applied carries significance beyond a simple mapping exercise. This assumption appears, on first encounter, rather puzzling. The truth of scientific knowledge is typically understood to be independent of local circumstances or geographical contingencies.

This view, however, is at best incomplete. Scientific knowledge is always a hard-won accomplishment. To make a claim that is robustly scientific requires the mobilization of material resources, technology, institutions, and personnel. Science thus relies upon distinct and intricate spatial arrangements and, in turn, significantly transforms the places where it is made, modified, or marketed. Investigating the geography of science involves, then, unearthing the spatial conditions necessary for the relative success of particular scientific claims or projects. Before offering examples of work that takes that task seriously, filling out the historical background and conceptual foundations of recent research on the geography of science will help highlight some important theoretical underpinnings.

Histories, Theories, and Concepts

While most academic work on the geography of science has occurred in the last two decades, it is possible to identify a longer, if discontinuous, history of interest in the geography of science reaching back to the late nineteenth century. In a study published in 1873, the Swiss phytogeographer Alphonse de Candolle concluded that Europe's leading men of science had been largely confined to a triangle with a base running from Sweden to Scotland and an apex terminating in central Italy. For Candolle, the particular social and natural environment of Western Europe was the leading cause of this golden triangle of scientific excellence. Such an approach and general argument had advocates in the twentieth century including the social geographer Stephen Visher and the historian of technology, Harold Dorn. For Visher and Dorn, the geography of science referred to the influence of different kinds of physical or cultural environments on the level and concentration of scientific ability or on the distribution of specific types of scientific inquiry. The drift in this work toward spatial or environmental determinism made it unpalatable to others interested in understanding the development and growing influence of science in the modern world. By contrast, a concern with the geography of science born out of a sustained engagement with the sociology of scientific knowledge and its more recent offshoots have helped to nurture a more lasting research program.

The sociology of scientific knowledge, which emerged in the early 1970s, rested on the assumption that science was 'constitutively' or inescapably social (Shapin, 1995: p. 289). This was a radical departure from earlier sociological accounts of science, which accepted that scientific error could be explained by social factors but which assumed that properly validated science was by definition free from the taint of social or cultural influences. Proponents of the so-called strong program extended the argument that science, like any other human activity, was a fundamentally social enterprise (see Barnes et al., 1996). Arguing against making a priori judgments about the truth of certain bodies of knowledge, they aimed at social accounts that explained equally the success or failure of competing scientific claims. The underlying premise that science was social 'all the way down' also encouraged a 'localist' turn among sociologists of science (Shapin, 1995). If science was manifestly a social enterprise, it was also unavoidably 'local,' rooted in specific cultural circumstances and distinct sites of knowledge production.

In short, the realization that scientific knowledge was a form of situated knowledge (Haraway, 1988) highlighted the importance of geographic context for sociologists and historians of science (e.g., Smith and Agar, 1998). It also led to calls for geographers to take seriously the spatiality of scientific knowledge (Livingstone, 1995). At the same time, and not entirely independent, the nature of space and its role in social and cultural life in late modernity was being rethought by a number of influential thinkers within and outside of the discipline of geography. This so-called spatial turn involved

reconceptualizing space by overturning the view that regarded it as an empty or inert container for social action. Influenced by the theoretical interventions of Henri Lefebvre and others, space was redefined as part of the action, made and remade along with social and cultural life. The spaces of science, in this way of thinking, were no longer viewed simply as a backdrop to scientific change but were now analyzed as a key ingredient in the unfolding dynamics of scientific cultures.

A number of the architects of the spatial turn strongly influenced geographers and historians and sociologists of science interested in the spatiality of scientific knowledge. To name but two, Clifford Geertz's 'thick descriptions' of diverse social groups, which carefully excavated the local frames of reference that governed cultural behavior, proved useful for constructing ethnographies of scientific communities. And Michel Foucault's archaeologies of venues that facilitated the emergence of modern forms of power and surveillance provided inspiration for those trying to understand the constitutive role of particular spaces in the making and meaning of science.

The rethinking of space as an active part of science in the making also showed that carefully reconstructing the circumstances that shaped different sorts of scientific projects involved more than the reductively 'local.' In other words, the location of science could not be reduced to a set of numerical coordinates. Instead, it had to be approached as a product of social relations and scientific practices that were at once located in a given place and distributed across space. Sociologists of science also quickly realized that the multiplication of individual case studies detailing the situated nature of particular scientific initiatives could not offer a compelling explanation for one of the most remarkable features of modern science – its efficient and secure transmission from one place to many others (e.g., Harris, 1998). Science was, after all, widely supposed to be a global, because universal, form of knowledge. The geography of science had to include not only a detailed account of the lineaments of self-contained social spaces and circumscribed venues but also an understanding of the character and extent of wider networks of scientific exchange and communication. Bruno Latour (1987) was among the first to supply a set of theoretical resources to help examine the distributed nature of scientific research. In particular, Latour pointed to the creation of 'immutable mobiles' – fixed units of scientific information readily combined with similar units derived from other distant places – and to 'centers of calculation' where these units are collected, collated, and classified. Peter Galison's (1997) proposal that science relies on successful negotiations within a 'trading zone' offered another approach to understanding the efficient exchange of results, methods, and ideas across different scientific subcultures.

Broadly speaking, then, the geography of science can be analyzed using two relatively distinct modes of inquiry. The first is concerned with situating science in particular sites, places, and institutions. The second is concerned with tracking the movement of science across space. Some examples will illustrate what might be involved in pursuing these lines of investigation. As will become clear, however, this heuristic distinction between sites and movements, and between local and global knowledge, has its limitations and methodological blind spots.

Situating Science

A key claim underlying much recent work on the geography of science is that precisely where scientific knowledge is generated, transacted, applied, or consumed is of considerable importance. Science is always made somewhere, and opening up the 'black box' of those discrete settings can reveal a great deal about the nature and meaning of scientific research.

Among the huge range of sites where scientific knowledge has been, and is, generated, the laboratory is perhaps the most obvious. Designed to provide a carefully controlled environment for the production of accurate and precise knowledge about the world, the lab is often thought of as the archetypal space for conducting scientific research. Partly for this reason, it was among the first scientific sites to be explored by sociologists interested in the social dimensions of scientific knowledge production (e.g., Woolgar and Latour, 1979; Knorr-Cetina, 1981). If the social character of laboratories designed to filter out influences that might 'contaminate' scientific procedures could be determined, then the case for regarding science as an irreducibly social enterprise could be more strongly made.

What rapidly became clear through these studies is that 'the laboratory' does not refer to a single, stable space but rather to a diverse set of venues operating according to different codes of conduct, cultural norms, and epistemic protocols. Even the most carefully managed modern laboratory was found to have its own distinctive research culture and acted as an important social institution for determining scientific questions and scientific careers. The 'micro geography' of the lab – understood not just in terms of physical arrangement of rooms and equipment but also the social interactions of lab workers – proved to be of vital significance. This finding, of course, runs against the grain of the ideal to which those designing and running laboratories aspire, namely, to create a 'placeless' place entirely isolated from a messy and chaotic world (Kohler, 2002). Yet there is, as it turns out, an interesting geography to the pursuit of that ideal.

If laboratories are created to control the messiness of place, field sites might be thought of as a kind of antilaboratory, a space in which variability or contingency is preserved for the purpose of scientific study. Scholars interested in this contrasting 'site' of scientific inquiry have noted that fieldwork is an inherently risky intellectual venture pursued without the security of controlled conditions, carefully defined protocols, or the infrastructural support provided by a modern laboratory (Kuklick and Kohler, 1996). Historical studies have shown that the basic spatial division between laboratory spaces and field spaces emerged early in the history of modern science and was also apparent in the methodological preferences of different practitioners. The apparently fleeting, partial and idiosyncratic observations made in the field were regarded by some as untrustworthy and of limited value. Defenders of fieldwork responded by arguing that out-of-door observations provided data on 'raw nature' and brought the investigator closer to nature's realities than did the isolated researcher working under artificial conditions in the study or laboratory (Outram, 1996). However valued, the open-ended and unpredictable character of field spaces forcibly underlines the situated character of scientific inquiry, its 'craft competencies,'

tacit rules of action, and its responsiveness to changing local conditions (Livingstone, 2003: p. 45).

The contrast perennially drawn between field sites and laboratories also helped to form a certain sort of scientific investigator. Strategies to counteract charges of hasty observations and uncertain and unrepeatable findings included emphasizing the heroic exploits necessary to get to, and remain in, isolated and challenging field environments. These strategies were taken a stage further with the creation of other distinctive sites for scientific inquiry such as the field station, observatory, and mobile laboratory. Field stations, for example, combined the 'virtues' of lab and field to produce a hybrid venue marked by a kind of controlled adaptability. They created an experimental complex positioned, as Henke (2000: p. 484) puts it, between 'the standardized space of the lab and the messiness of the field.' The astronomical observatory emerged from a desire to offer a controlled and replicable site for making precise measurements and observations in remote places. For all that, while appealing to the rhetoric of precision associated with the laboratory, late-Victorian astronomers also described themselves in terms of the 'explorer scientist' and 'heroic fieldworker' (Lane, 2010). And the mobile laboratory on research vessels like the *Challenger* merged different ways of knowing and functioned within a distinctive social environment created by the uneasy relations between scientists and sailors (Rozwadowski, 2005).

Other spaces of scientific inquiry have been defined as much by the task of attracting public interest as by efforts to produce credible and robust knowledge. This has generated additional challenges, competencies, and patterns of inquiry. Museums provide a classic example of a scientific space that formed at the intersection between the quest for knowledge about nature and the desire to display that knowledge to a general public (Forgan, 2005). Emerging from early modern cabinets of curiosity, museums by the nineteenth century became key sites for nurturing the natural sciences and fashioning a certain sort of citizen. European metropolitan museums in particular became linked with nation building by indexing national progress through vast collections of carefully arranged natural and cultural artifacts. They also operated within a marketplace of ideas and institutions, competing with other public attractions such as the circus or freak show. Striking a balance between entertainment, instruction, and original research helped to shape internal spatial arrangements and influenced the direction, success, and wider meanings of museum research. At the same time, the micro geographies of display were also shaped by scientific disputes. The Natural History Museum in London is classic case in point. In the late nineteenth century, the director, Richard Owen, arranged the collections of organic specimens according to his own particular account of the relations between extant and extinct plant and animals (Yanni, 2000). The displays materialized Owen's dissent from Darwin's theory of evolution and promoted his own program of natural historical research.

Botanical and zoological gardens appeared in the same period as museums and, in a similar fashion, were fundamentally shaped by cultural as much as scientific concerns. As Johnson (2011) has shown, botanical gardens were constructed with both aesthetic and taxonomic imperatives in mind. With the emergence of glasshouses, gardens were also able to accommodate exotic flora, a fact that appealed to visitors as well as enabling research on tropical as well as temperate vegetation. Zoos, too, traded in the exotic, retaining an emphasis on unfamiliar fauna from the menageries from which they emerged. The scientific study of the living world was thus accompanied by a set of changing cultural expectations associated with a public's appetite for viewing exotic plants and charismatic animals.

A host of other less obvious sites where scientific knowledge has been assembled or assessed have also been interrogated by scholars interested in the mediating role of space in the making of modern science. The public house, the kitchen, and the courtroom have, for example, revealed a set of social scientific relations mediated by, respectively, artisanal sociability, domestic economies, and legal regimes. But situating science in its diverse contexts of production and dissemination need not be restricted to the micro scale. Cities and regions have also been studied as 'sites' in which science has been pursued or promoted. Science has, in the words of Lafuente and Saraiva (2004), played a 'fundamental role in the definition of urban spaces.' In the late-nineteenth-century Madrid, for example, scientific institutions functioned as heterotopias, 'symbols of a promised metropolis' (Lafuente and Saraiva, 2004: p. 531). The 'shock' cities of industrial Europe in the nineteenth century presented challenges that encouraged the rise of the 'civic scientist' with expertise in urban problems. Further, cities have become both field sites and laboratories for a variety of scientific research programs. The Chicago School of urban studies, for instance, was closely tied to the study of the city in which it was based and worked to create a 'truth spot' at an urban scale that would secure the school's credibility among other social scientists (Gieryn, 2006). Cities and regions have also helped to shape the varied reception of scientific knowledge. This has been demonstrated in abundant detail by Secord (2000) through careful reconstruction of the different responses to the anonymously authored evolutionary text *Vestiges of the Natural History of Creation* (1844) in the contrasting civic and religious cultures of Edinburgh, Liverpool, and Oxford. Livingstone (2005), too, has shown that responses to Darwinism in the American South, New Zealand, and Russia were determined in part by the religious and cultural milieux of those regions.

At whatever scale of investigation, it is clear that different sites of scientific inquiry come with their own distinctive suites of practices, protocols, and personnel. Situating science in the sites of its making thus requires attention to the specific arrangements that make scientific investigation possible and also demands attention to the ways in which scientific investigation helps to form and refashion the spaces in which it is carried out. But, as many have pointed out, the localist nature of such studies do not immediately provide answers to why modern science has been so successful and so influential (Shapin, 1998; Harris, 1998; Secord, 2004). More specifically, they do not address the question of why, if it is just another form of local knowledge, science has traveled with such unparalleled efficiency from one place to many others. Nor do they necessarily help us to understand how science earned its reputation as the quintessentially global form of knowledge. These patently geographical questions direct us to the demanding but essential task of tracing and interpreting the movement of science across space.

Tracking Science

As outlined above, detailed accounts of individual laboratories have helpfully troubled the notion that science, as a kind of birthright, floats free from the contexts in which it is made. But such micro studies have risked missing the ways in which laboratories are from the start 'plugged into' the wider world. Kohler (2008), for example, argues that laboratories should be understood, among other things, as an expression of late-modern capitalism. According to Kohler, "the analytic categories and practices of lab science were congruent with the new managerial hierarchies and procedures of large-scale industrial capitalism…[and] scientists proved useful allies for modernizers striving to transform the traditional business firm into the modern managerial leviathan" (p. 768). This analysis does not, however, apply straightforwardly to all laboratory cultures. Simon Reid-Henry's (2010) close reading of Cuba's biotech industry demonstrates that it functions according to the directives of a socialist state *and* the powerful norms of a global industry. Its 'strategic marginality' has allowed innovation and experimental creativity to flourish but not in ways easily translatable into the regulatory regimes of Western capitalism. Even so, the labs within Havana's biotech complex are thoroughly embedded in networks that stretch far beyond local horizons.

Arguably, work on field sciences has been alert from the start to the situated and mobile nature of scientific investigations pursued outside the laboratory walls. We have already noted that field sites have often been contrasted with laboratories in their unpredictability and 'messiness.' As Bruno Latour (1999: p. 30) puts it, in the laboratory 'one no longer sees science stammer.' In the field, the situation is different. There the 'gap' between science and the 'real world' clearly opens up. Fieldworkers face the formidable task of acquiring stable data for building robust theories about large-scale natural phenomena. The 'jump' from local to global knowledge cannot be done in a leap but involves a significant amount of hard scientific labor. In his account of field research in Brazil, Latour (1999) shows that the 'chains of translations' that connect the field site to the final published paper involve a series of small transformative steps. The field scientist moves through physical space (from field to home institution, for example) and, often more haltingly, across cognitive domains and between scientific disciplines. The creation of standard forms of measurement and transportable and comparable samples are among the many rather precarious techniques necessary to convert field observations into the printed diagrams and text of a scientific paper.

Fieldwork, involving as it does the traversal of real as well as cognitive terrain, can also quickly become entangled in matters not directly connected with the scientific task at hand. Historically, field exploration has been thoroughly embroiled not only in the development of scientific knowledge but also in the politics, ambitions, and at times brutal realities of colonialism. In this sense, doing scientific work 'in the field' cannot be separated from the much larger political and economic projects that, in ambition if not in fact, have spanned the globe. A classic case in point is the Great Trigonometrical Survey of India, begun in 1802 by the British army officer and geodesist William Lambton and continued for much of the nineteenth century. The ambitious aim of the survey, to determine the size and shape of the globe, was pursued alongside more mundane matters of colonial governance. The survey also provided material for an apologia for imperial expansion in the Indian subcontinent (Edney, 1997). The practical exigencies of geodetic measurements in the field were thus combined with a political project and political discourse that imbued the survey with significance well beyond its immediate practical and intellectual goals.

The kind of 'big science' carried out by British surveyors in India has been productively studied by other scholars interested in the ways in which scientific knowledge 'travels' well beyond the confines of the laboratory, field site, or study. In his work on early modern corporations, Steve Harris (1998: p. 271) has argued that "how science travels has as much to do with the problem of travel in the making of science as it does with the problem of making science travel." As Harris shows, the development of natural knowledge in the early modern period cannot be disentangled from the patronage and ambitions of 'long-distance corporations' such as the Dutch East India Company and the Society of Jesus, a Roman Catholic religious order. Mike Hulme's (2010) work on the geographies of climate change science is animated by a similar concern with the institutional settings from which 'global' knowledge claims emerge. Hulme contrasts the 'globalizing metrics' of officially sanctioned climate science with the local cultural values that shape people's responses to variable climatic conditions. Applying this to the Intergovernmental Panel on Climate Change, Hulme argues that insensitivity to a range of cultural perspectives creates a 'brittle' kind of global knowledge that cannot accommodate a 'plural and turbulent world.'

The politics of science, understood as a globalizing or 'totalizing' enterprise, has also attracted the attention of post-colonial theorists. Warwick Anderson (2009: p. 392), for example, has called for more attention to how "science and technology travel over the uneven terrain that colonialism has worked." By ignoring the legacies of European colonialism, scholars have neglected the continued impact of imperial forms of power on science. Historians of science have also underlined the importance of attending to connections between science and imperialism. Kapil Raj (2007), for example, has argued that in the eighteenth and nineteenth centuries scientific knowledge was a product of the unequal relations between Europe and the territories that it explored and colonized. At the same time, science was not simply a European imposition foisted on subjugated peoples. Nor was the knowledge of non-Europeans always cordoned off from, and alien to, Western science. Instead, science was made in intercultural contact zones in which the flow of ideas and scientific materials, while tipped toward European powers, moved in more than one direction.

Conclusion

Whether concerned with situating scientific knowledge or tracing its movements, the work summarized above reveals the importance of grasping the spatialities of science in efforts to uncover its profound but varied influence on late-modern societies. It has also become clear that the question of scale is a central one for work on the geography of science. Determining the most appropriate scale for investigating effectively

a particular scientific project is not at all straightforward. The site or sites at which scientific knowledge is produced may or may not have identifiable cartographic coordinates but a physical 'address' does not tell us much about the scale at which scientific inquiries are carried out. A laboratory may have its own 'local' cultures of research that are dependent upon a dense and widely distributed set of ideas, instruments, personnel, and methods. To overcome this difficultly, scale might be profitably regarded as an emergent property of the practice of science itself rather than something that exists independently from scientific work. There is a danger, however, that this constructivist account of scale may exaggerate the power of scientific practices to transform human politics and society. A related challenge concerns the applicability of accounts of the geography of certain scientific projects, practices, or debates to other cultural situations or to other scientific times and places. One criticism made of studies offering detailed reconstructions of scientific spaces in the past or present is that they tell us little about changing patterns that can only be discerned at a 'macro' level. As Daston (2009: p. 809) recently put it, for the average 'micro history' of science, 'the texture is fine grained, the metaphysics is nominalist and the aesthetic is pointillist.' Clearly, the difficult task of moving between scales, whether spatial or temporal, is faced by geographers concerned with investigating the spaces of modern science. This is not, of course, a new problem or one that is unique to science studies. Human geographers have long been concerned with the challenges of relating the particular and the general and while they clearly have no monopoly on approaches that might address this methodological quandary, they are well placed to offer some ways forward. Certainly, there is no reason to restrict the study of the geography of science to the 'pointillist' approach identified by Daston. Addressing these and other methodological issues will no doubt continue to prompt and guide further work in a scholarly field that is now a well-established area of geographical inquiry.

See also: Knowledge, Geography of; Space and Social Theory in Geography.

Bibliography

Anderson, W., 2009. From subjugated knowledge to conjugated subjects: science and globalisation, or postcolonial studies of science? Postcolonial Studies 12, 389–400.

Barnes, B., Bloor, D., Henry, J., 1996. Scientific Knowledge: A Sociological Analysis. University of Chicago Press, Chicago.

Daston, L., 2009. Science studies and the history of science. Critical Inquiry 35, 798–815.

Dorn, H., 1991. The Geography of Science. Johns Hopkins University Press, Baltimore.

Edney, M., 1997. Mapping an Empire: The Geographical Construction of British India, 1765–1843. University of Chicago Press, Chicago.

Forgan, S., 2005. Building the museum. Isis 96, 572–585.

Galison, P., 1997. Material culture, theoretical culture, and delocalization. In: Krige, J. (Ed.), Science in the Twentieth Century. Harwood, Amsterdam, pp. 669–682.

Geertz, C., 1983. Local Knowledge: Further Essays in Interpretative Anthropology. Basic Books, New York.

Gieryn, T.F., 1999. Cultural Boundaries of Science: Credibility on the Line. University of Chicago Press, Chicago.

Gieryn, T.F., 2006. City as truth spot. Laboratories and field sites in urban studies. Social Studies of Science 36, 5–38.

Haraway, D., 1988. Situated knowledges: the science question in feminism and the privilege of the partial perspective. Feminist Studies 14, 575–599.

Harris, S.J., 1998. Long distance corporations, big sciences, and the geography of knowledge. Configurations 6, 269–305.

Henke, C., 2000. Making a place for science: the field trial. Social Studies of Science 30, 483–511.

Hulme, M., 2010. Problems with making and governing global kinds of knowledge. Global Environmental Change 20, 558–564.

Johnson, N.C., 2011. Nature Displaced, Nature Displayed: Order and Beauty in Botanical Gardens. I. B. Tauris, London.

Knorr-Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Permagon Press, Oxford.

Kohler, R.E., 2002. Landscapes and Labscapes. Exploring the Lab-Field Border in Biology. University of Chicago Press, Chicago.

Kohler, R.E., 2008. Lab history: reflections. Isis 99, 761–768.

Kuklick, H., Kohler, R., 1996. Introduction: science in the field. Osiris 11, 1–14.

Lafuente, A., Saraiva, T., 2004. The urban scale of science and the enlargement of Madrid (1851–1936). Social Studies of Science 34, 531–569.

Lane, K.M.D., 2010. Geographies of Mars: Seeing and Knowing the Red Planet. University of Chicago Press, Chicago.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.

Latour, B., 1999. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.

Livingstone, D.N., 1995. The spaces of knowledge: contributions towards a historical geography of science. Environment and Planning D 13, 5–34.

Livingstone, D.N., 2003. Putting Science in its Place: Geographies of Scientific Knowledge. University of Chicago Press, Chicago.

Livingstone, D.N., 2005. Science, text and space: thoughts on the geography of reading. Transactions of the Institute of British Geographers 35, 391–401.

Outram, D., 1996. New spaces in natural history. In: Jardine, N., Secord, J.A., Spary, E.C. (Eds.), Cultures of Natural History. Cambridge University Press, Cambridge, pp. 249–265.

Raj, K., 2007. Relocating Modern Science: Circulation and the Construction of Knowledge in South Asia and Europe, 1650–1800. Palgrave Macmillan, Basingstoke.

Reid-Henry, S., 2010. The Cuban Cure: Reason and Resistance in Global Science. University of Chicago Press, Chicago.

Rozwadowski, H., 2005. Fathoming the Ocean: The Discovery and Exploration of the Deep Sea. Harvard University Press, Cambridge, MA.

Secord, J.A., 2000. Victorian Sensation: The Extraordinary Publication, Reception and Secret Authorship of Vestiges of the Natural History of Creation. University of Chicago Press, Chicago.

Secord, J.A., 2004. Knowledge in transit. Isis 95, 654–672.

Shapin, S., 1995. Here and everywhere: sociology of scientific knowledge. Annual Review of Sociology 21, 289–321.

Shapin, S., 1998. Placing the view from nowhere: historical and sociological problems in the location of science. Transactions of the Institute of British Geographers 23, 5–12.

Smith, C., Agar, J. (Eds.), 1998. Making Space for Science: Territorial Themes in the Shaping of Knowledge. Macmillan, Basingstoke.

Woolgar, S., Latour, B., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, London.

Yanni, C., 2000. Nature's Museums: Victorian Science and the Architecture of Display. Johns Hopkins University Press, Baltimore.

Science, History of

Lorraine Daston, Max Planck Institute for the History of Science, Berlin, Germany

Published by Elsevier Ltd.

Abstract

The history of science is a small discipline with a large and expanding subject matter: not only the history of the modern natural and human sciences but also premodern and non-Western traditions of knowledge and learning. Because it lies at the intersection of the humanities and the natural human sciences, it has been and remains a locus of interdisciplinary research.

History of Science

The history of science studies the emergence and development of systematic knowledge. Just what kind of knowledge that is – whether it is defined by subject matter, reliability of results, contradistinction to technical know-how, cumulative progress, disciplinary consensus, and/or method of inquiry, demonstration, or explanation – is a point of persistent debate and reflection within the history of science and its cognate fields in philosophy, sociology, and anthropology. Linguistic and historiographic traditions diverge sharply on this point. Whereas Francophone and Anglophone scholars have since the mid-nineteenth century largely reserved the honorific 'science' for the natural sciences and mathematics (influenced by Auguste Comte's hierarchy of sciences in his *Cours de philosophie positive*, 1830–42, in which only some had reached the stage of 'positive' knowledge), the German *Wissenschaft* still refers to the range of fields covered by the Latin *scientia*, from physics to philology. Moreover, recent explorations of experiment and other practices have blurred the venerable boundary between *ars* and *scientia*, technique and systematic knowledge (whether understood narrowly as 'science' or broadly as *Wissenschaft*). Close investigation of the sciences even in the restrictive English and French usage has furthermore revealed considerable variety among methods and explanations in, for example, biology and physics. Finally, studies of premodern and non-Western intellectual traditions have made the creation of categories like 'science,' as well as the classifications of knowledge that order such categories, themselves objects (rather than presuppositions) of historical research. Hence this article will not attempt to delineate sharply the boundaries of 'science,' but rather briefly survey the (1) emergence and status of the history of science as a recognized field of inquiry; (2) principal divisions of the field and disciplinary affinities; (3) periodizations and narratives that have shaped those divisions; and (4) major historiographical developments since about 1970 (for an earlier survey see Kuhn, 1972 [1968]).

Emergence and Divisions of the Discipline

The history of science is an ancient pursuit, but a relatively young discipline. From Aristotle through the early nineteenth century, practitioners of one or another branch of knowledge have variously used the history of their field to argue for its dignity and importance, introduce it to beginners, situate it within a broader cultural milieu, summarize the literature to date, position themselves in relationship to that literature, praise and blame predecessors, give evidence of progress, extrapolate a program for future research, and draw lessons concerning the nature of knowledge and the conditions for its flourishing. In works like Joseph Priestley's *The History and Present State of Electricity* (1767) or Georges Cuvier's *Rapport historique sur les progrès des sciences naturelles depuis 1789* (1808), the history was inseparable from the science. Particularly in fields dominated by empirical research, the judicious sifting and ordering of past results was a precondition for coherence and a means for achieving consensus on what was reliably known and where the major challenges for future research lay. These functions are preserved today in the scientific review article and the posing of key 'problems' (e.g., the celebrated Hilbert problems in mathematics), and practicing scientists occasionally still appeal to the history of their field for present guidance, especially in times of crisis (Graham et al., 1983). By the mid-nineteenth century, however, histories of science had become distinct from scientific publications, although they were still written primarily by scientists, including prominent figures such as William Whewell, Marcellin Berthelot, Ernst Mach, and Pierre Duhem. Their histories often criticized the current state of science by establishing the genealogy of a controversial hypothesis (e.g., atomism), analyzing the origins of a suspect concept (e.g., absolute space) for hidden flaws, or pleading the superiority of one approach to science over another (e.g., Kantian ideas over Comtean facts). By 1900, histories of science had become a genre distinct from science, but they were still motivated by, and deeply engaged with, contemporary scientific developments (Laudan, 1993).

History of science coalesced only gradually as a recognizable discipline in the twentieth century, with its own distinctive program of training, institutions (journals, professional societies, university positions), and scholarly standards. Spurred by the organizational efforts of such scholars as Paul Tannery in France and Karl Sudhoff in Germany, there was a spurt of professionalizing activity around 1900, when the first international congress on the subject was held in Paris. The German Gesellschaft für Geschichte der Medizin und der Naturwissenschaften was established in 1901; the first chairs in the history of science also appeared at European universities around this time (Kragh, 1987). The Belgian historian George Sarton (Pyenson and Verbruggen, 2009), who emigrated to the USA after World War I, brought with him the journal *Isis* (est. 1912), which became the official organ of the North

American History of Science Society after its establishment in 1924 and remains the specialist journal with the largest circulation. Other journals dedicated to the history of science (including the history of the human sciences, which now boasts several specialist journals and societies) in all its aspects have since proliferated in many languages, to which the entries in the annual installments of the *Isis Current Bibliography* provide at least a partial guide.

Although Sarton's Comtean view of the history of science as a saga of intellectual and moral progress that should evaluate the past in light of present scientific understanding (Sarton, 1948 [1927]) has left little trace in current historiography (Dear, 2009), his tireless attempts to institutionalize and organize the discipline in the form of university departments, specialized journals, and research tools like the *Isis Current* (earlier *Critical*) *Bibliography* (History of Science Society, 1967), and *Dictionary of Scientific Biography* (Gillispie, 1981) had a lasting impact on the field (Thackray and Merton, 1972; Holton, 2009). During the 1950s, graduate programs dedicated to the history of science (sometimes in conjunction with the philosophy or sociology of science) multiplied; by 1990, there existed almost 200 degree-granting programs in 25 countries.

By no means all history of science is done under the auspices of university departments or published in journals expressly devoted to the subject. The disciplinary affinities of the history of science are many and mutable, with distinctive national differences. Until the 1970s, the primary connection, both in Europe and North America, was probably with the sciences themselves. This connection was rooted deep in the origins of the field and sustained by the continuing importance of some form of the history of science in the intellectual and social initiation of young scientists, as well as by the recruitment of historians of science from the ranks of those with scientific training. In the UK and North America, history largely replaced philosophy during the 1980s and 1990s as the discipline to which historians of science felt most akin (and within which the majority of them found employment), although interactions with sociology and, more recently, anthropology have also exerted a powerful influence, especially in connection with programs in science studies. In Western Europe, philosophy has remained a significant framework for the history of science, in diverse forms ranging from neo-Kantianism to logical empiricism, though science and technology studies have also been powerful influences since the 1980s. In Eastern Europe, Russia, Taiwan, and China, the history of science, sometimes with a national focus, has been cultivated within state-sponsored academies of science; in Latin America, Africa, and India science studies, postcolonial studies, and the history of medicine have often been the loci of academic teaching and research. In addition to its disciplinary polyvalence with respect to choice of problems and methods, the history of science is closely linked by its subject matter to the histories of medicine and technology, albeit to differing degrees depending on the historical period in question. The phrase 'science and technology studies' bears witness to these criss-crossing ties to other disciplines, serving as an abbreviation for the conglomerate 'history, philosophy, sociology, and anthropology of science, medicine, and technology,' which, however cumbersome, accurately reflects the ecumenical perspective of many historians of science.

Narratives and Periodization

Because of its links to philosophy, sociology, and anthropology, as well as to the natural and human sciences and mathematics, the history of science has been highly theorized for centuries. Theories about what scientific knowledge is, how it develops over time, where and when it expands and contracts, and in which ways it interacts with extrascientific contexts have profoundly shaped the way in which the history of science has been periodized and narrated. The more recent developments (since about 1970) in the historiography of science are incomprehensible without some background concerning the dominant theoretical schemes that had molded the field previously.

All history is inevitably written to some extent from the standpoint of the present, but present science – its orthodoxies and heresies, its controversies and critics – has exerted a mighty influence on the history of science since the Enlightenment. The great bulk of the history of science has been written explicitly or implicitly to illuminate, justify, or criticize the current state of science. In the Enlightenment histories, e.g., Jean-Étienne Montucla's *Histoire des mathématiques* (1758) or Jean-Baptiste Delambre's *Rapport historique sur les progrès des sciences mathématiques depuis 1789* (1810), the aim was to demonstrate that at least some of the sciences had shown steady and irreversible progress, especially since the Renaissance, and that the history of science was therefore a more edifying genre than the chronicles of human folly recounted by political and military history. The notion of progress was originally applied to art (by Giorgio Vasari in 1550) and to technology (by Francis Bacon in 1620), not to science. Once this interpretation of scientific change took hold in the eighteenth century, however, science became the premier example of progressive advance and scientific progress; a criterion by which to judge the comparative worth of civilizations. Whatever may have been the cultural achievements of the Latin Middle Ages or the Chinese empire, the stagnation of the sciences and mathematics in these milieux, so ran the argument of the Enlightenment historians, revealed fundamental intellectual and social limitations – limitations that in the nineteenth and twentieth centuries came to be described as characteristic of 'traditional' or 'premodern' societies. Hence narratives of scientific progress eventually became part and parcel of larger narratives about modernization and European superiority.

Although the assumptions underlying the Enlightenment vision of scientific progress are seldom invoked explicitly by contemporary historians of science, the reigning periodization of the field and the definition of specialties as central or peripheral still bears the traces of the traditional–modern division. Despite their glaring differences in many substantive aspects, the study of science in antiquity, the Arabic and Latin Middle Ages, and in non-Western traditions such as those of China and India has often been amalgamated into a shared problematic, one which defined science as peculiarly modern and Western by opposition (Elshakry, 2010). Only from this Eurocentric (and often ideologically self-serving) viewpoint do questions like the following make sense, although they

dominated the history of premodern and non-Western science until very recently: to what extent did the knowledge of these highly disparate cultures anticipate/approximate the science of post-Renaissance Europe? And to the extent that it did not, what were the intellectual and social obstacles that blocked the way? Hence the pioneering works in these areas (e.g., Duhem, 1908; Maier, 1949–58; Clagett, 1959; Needham et al., 1954) foregrounded possible forgotten contributions to and missing preconditions for the Scientific Revolution. Knowledge traditions that did not lead to modern science, understood to commence in early modern (a term as saturated with hindsight as 'the Middle Ages') Europe, were generally relegated to the periphery of the discipline, whatever their intrinsic interest and however excellent the scholarship devoted to them. As the scholarship on the history of Arabic, Chinese, and other non-Western, as well as premodern Western traditions has expanded, the parochialism of this perspective has been sharply criticized (Sabra, 1996; Lloyd, 2004; Sivasundaram, 2010; Chemla, 2012).

Despite such critiques, the Scientific Revolution has remained oddly central to the historiography of science. Oddly, because recent scholarship on Europe in the sixteenth and seventeenth centuries has emphasized the heterogeneity and regional variations in the developments in anatomy, astronomy, natural history, mechanics, chemistry, mathematics, and physics that were once summarily lumped together as 'the Scientific Revolution' (Lindberg and Westman, 1990; Porter and Teich, 1992; Shapin, 1996; Osler, 2000; Park and Daston, 2006). Moreover, the steep increase of studies of twentieth-century science since about 1980 has highlighted the novelty and significance of post-1900 developments, as the steady increase in the number of entries under these rubrics in the *Isis Current Bibliography* in the period 1990–2010 bears witness: whereas the number of publications under the combined rubrics 'Renaissance' and 'Seventeenth Century' (the time period embraced by the Scientific Revolution) have fluctuated between approximately 15 and 20% (with downward trend) of all the entries listed chronologically, those for 'Twentieth Century' and thereafter have risen steadily from approximately 30 to 40% (with upward trend) of the total.

Yet the Scientific Revolution retains much of the mythical status first granted to it by the Enlightenment historians as the mold in which modernity was cast, one of the few points upon which both the apostles of progress and the critics of modernity agree. This periodization has shaped the division of subspecialties within the history of science, which tend to be chronological or geographic ('ancient,' 'Southeast Asian') for knowledge traditions prior to, or uninfluenced by, the Scientific Revolution, but disciplinary ('physical sciences,' 'biomedical sciences') for those following that original event – even for disciplines (e.g., astronomy) with histories that long antedate the sixteenth century and with geographical distribution well beyond Europe. Recent studies of science on a global scale have put narratives and periodizations explicitly or implicitly oriented around the European Scientific Revolution under increasing strain, but a widely accepted alternative has yet to emerge.

The progress of the sciences (especially mathematics and the exact sciences) set forth by the Enlightenment historians was cumulative but not revolutionary. In their accounts, knowledge steadily expanded around a stable core of permanent truths, which may have taken millennia to discover, but which were thereafter firmly established beyond challenge. The language of 'revolutions' declined among early nineteenth-century scientists, perhaps in order to distance the permanent improvements of natural knowledge from the brusque reversals of politics; instead, the language of 'law' (e.g., Comte's 'Law of Three Stages') was applied to the historical development of science (Cohen, 1985). In the middle decades of the nineteenth century, however, this view of durable, inexorable scientific progress gave way to one of the rapid successions of scientific theories, even in areas (e.g., optics) long believed to rest upon unshakable foundations. Scientific progress had become vertiginous and open-ended; today's truth might be discredited as tomorrow's error. Under these conditions, it became increasingly difficult for historians of science to portray the steady growth of knowledge as culminating in present triumphs, since the endpoint was constantly changing. Instead, the history of science turned critical and conceptual: in the hands of Mach, Ostwald, and other late nineteenth-century scientist-historians, history of science revealed the past roots of present incoherences and taught the mutability of scientific theories. Within the revised interpretation of scientific progress, the purpose of the history of science was to use past science to change, not confirm, present science. Despite the dramatic shift in the narrative of scientific progress and in the goals of the history of science, its presentist orientation remained unchanged.

Almost all the strands of early twentieth-century historiography of science can be seen as attempts to come to grips with the peculiar character of scientific knowledge, which seemed to be simultaneously mutable and cumulative, driven by an internal dynamic and yet sensitive to ambient social conditions. These attempts often tugged in opposite directions. For example, the logical positivist program that stamped the history and philosophy of science in the middle decades of the twentieth century, especially in the modified form imported from the Vienna Circle to the UK and the USA, fixed upon the empirical findings of science as its enduring core in the Heraclitean flux of hypotheses and theories. These 'protocol statements' (Carnap, 1928) were likened to simple observations, bringing science into the neighborhood of common sense. In contrast, the French historian and philosopher Gaston Bachelard, drawing upon the insights of psychoanalysis, insisted upon the necessity of overcoming everyday experience and intuitions in the formulation of successful scientific concepts (Bachelard, 1938).

A many-cornered debate was ignited by the attempts of Marxist historians to relate the rise of modern science to specific socioeconomic conditions (Hessen, 1931; Bernal, 1939), out of which emerged the opposition of 'internal' to 'external' approaches to the history of science (Kuhn, 1972 [1968]; Shapin, 1992). External approaches found a non-Marxist model in Robert K. Merton's influential study of how science came to be valued in seventeenth-century England (Merton, 1938; Shapin, 1988). Informed by German and French neo-Kantian traditions in the history of philosophy, Alexandre Koyré's studies of the development of science in the broader intellectual context of early modern Europe stamped the work of historians of science, especially in France and the USA,

trained in the 1950s and 1960s (Koyré, 1939–40, 1961; Jorland, 1981). Common to all of these otherwise highly diverse perspectives on the history of science was an engagement with contemporary science and its lineage. Whether positivist or psychoanalytic, externalist or internalist, the history of science took the form of a genealogy leading to ideas and discoveries canonized by contemporary science. However much they may have disagreed on historical methods and forms of explanation, the combatants agreed on what and who in the history of science was worthy of explanation, namely that which had turned out to be pregnant with the future. In these histories, a potent telos was at work, making present science appear to be not only the actual but also the inevitable outcome of past developments.

Developments 1970–2010

Thomas S. Kuhn's *The Structure of Scientific Revolutions* (Kuhn, 1970 [1962]) broke decisively with this historiography, and stimulated new trends (not all to Kuhn's liking) in the history, sociology, and philosophy of science that have transformed the understanding of science and its growth in these fields since 1970. Kuhn argued for a plurality of scientific revolutions, in addition to *the* Scientific Revolution, which alternate with periods of 'normal science' framed by a 'paradigm' that models what can count as an acceptable problem, method, and solution in a given science in a specified period. It was less, however, Kuhn's 'structures' or patterns of long-term scientific development that exercised a deep and enduring influence on the history of science (although this feature of the book dominated its reception in other disciplines, particularly the social sciences) than his frontal attack on progressive narratives, his sketchy but evocative notion of 'paradigm' that embraced practices as well as theories, and the thoroughgoing historicism of many of his examples. All of these lines of thought were further cultivated in the history, philosophy, and sociology of science (the three disciplines were drawn into intensive interaction in part through their differentiated reception of Kuhn's book) during the subsequent three decades, often with consequences that led in directions diverging markedly from those plotted in *The Structure of Scientific Revolutions*.

Kuhn's rejection of progressive narratives in the history of science derived from his view that successive paradigms established and later overthrown by scientific revolutions were, first, internally coherent, and second, in a strong sense incommensurable with one another. This implied that science could not be cumulative in any progressive sense, one paradigm extending and building upon its predecessors. Kuhn's claim that conflicts between rival paradigms were not necessarily resolved by paradigm-independent criteria (e.g., predictive accuracy of explanatory unity) unsettled the inevitability of scientific development still further. Sociologists of science Barry Barnes and David Bloor took these claims as the basis of an approach called 'constructivism' (Golinski, 1998), in which the development of scientific knowledge was to be explained without appeal to its truth or falsehood, on the grounds that the truth of a proposition is neither a necessary nor sufficient condition for its acceptance (Bloor, 1991 [1976]). A protracted and ultimately inconclusive debate between 'social constructivists' (mostly coming from the sociology of scientific knowledge, or SSK, camp) and 'realists' (mostly philosophers of science) raged during the 1980s and early 1990s, ending not with a bang but a whimper (Hacking, 1999). In some circles, the debate was construed as a postmodernist attack on the sciences, triggering the brief but bitter 'Science Wars' of the mid-1990s (Wise, 1996). During the 1980s and 1990s, the history of science and science studies cross-fertilized one another in fruitful ways that spanned the premodern/modern divide. Articles devoted to science studies topics appeared regularly in *Isis* and *Social Studies of Science*, the flagship journal of science studies, returned the favor by publishing articles dealing with pre-1900 topics. Since 1996, however, the latter journal introduced an 'occasional series of "Historic Papers,"' under which rubric only three items appeared 1996–2010. Conversely, the 'Focus' section of *Isis*, initiated in 2004 to spotlight topics of over-arching interest in the history of science, has featured only a few themes with strong science studies affinities. Although the history of science was the battleground upon which all sides fought, arming themselves with historical examples to argue their points, these squabbles have had little lingering influence on how the history of science (as opposed to science studies) was pursued a mere decade later.

For historians of science, other aspects of the Kuhnian program have proved more fertile in the long run. First, in conjunction with Kuhn's focus on periods of conflict between rival paradigms, it prompted detailed studies of how scientific controversies were in fact resolved. Second, it enlarged the definition of what could count as 'science' for the purposes of historical investigation by ruling out hindsight judgments about what was and was not legitimate by present standards (to the benefit of specialties such as the history of premodern science and of the human sciences). Third, it encouraged a broader exploration of the grounds for scientific conviction by enforcing a 'symmetric' treatment of reasons for embracing false as well as true beliefs. This deliberate refusal to adopt the perspective of present science in the investigation of past knowledge went under the name of 'anti-Whiggery' among historians of science (in allusion to Herbert Butterfield's critique of 'Whig' political history). Among sociologists of science, the symmetry principle led to an increasingly distanced approach even to present science, to the point of adopting the ethnographer's estranged perspective (Latour and Woolgar, 1986 [1979]).

Kuhnian paradigms drew upon the philosophical reflections of Ludwig Wittgenstein (Wittgenstein, 1953) and Michael Polanyi (Polanyi, 1958) to stress the importance of knowledge that could not be reduced to rules or algorithms or even to words in a normal scientific tradition. In the history, sociology, and philosophy of science, a rich literature on 'scientific practices' has taken the inadequacy of the formal aspects of theory or explicit methods to explain the distinctive aspects of scientific knowing as its departure point. Although scientific practices may embrace everything from a knack for integrating equations to collecting botanical specimens to visualizing data, a rash of brilliant studies on experiment in the 1980s became exemplary of the new practice-centered approach. On this topic, interactions among

historians (e.g., Shapin and Schaffer, 1985; Galison, 1987; Gooding et al., 1993), philosophers (e.g., Cartwright, 1983; Hacking, 1983), and sociologists (e.g., Collins, 1985) created a discussion that blended epistemological, historical, and social analyses and inspired the next generation of historians of science.

Scientific practices more generally have emerged as autonomous forms of knowledge, no longer subsidiary to and in lockstep with theory, and have been at the center of the history of science since the mid-1990s, coinciding with the ever closer integration of the field with cultural history (Pestre, 1995). Practices have also served as an alternative way of making sense of the history of science across periods and disciplines (Pickstone, 2000). Practices usually depend crucially on local factors (including skill, material culture, values) that travel only with difficulty. This emphasis upon the locality of scientific knowledge has spawned a new set of problems: How is local science made global? In the earlier historiography, science had been assumed to be universal because nature was so. For historians of scientific practice, in contrast, further explanation is required to show the ways in which localized achievements can be made at least partially communicable among practitioners divided by geography, history, or culture (Schaffer, 1992; Galison, 1999). Sociologists of science have attacked this problem from the standpoint of 'actor networks' (Latour, 1987). In keeping with the growing interest in the circulation and communication of knowledge on a global scale, historians of science have paid increasing attention to the history of reading and publishing (Secord, 2004), the geography of scientific knowledge (Livingstone, 2003), and cross-cultural encounters (Fan, 2004; Raj, 2006; Cook, 2007; Delbourgo and Dew, 2008).

Both the rejection of presentist perspectives and the close study of scientific practices have pushed historians of science toward a more thoroughly historicist approach. Whereas their predecessors practiced anachronism openly and with conviction, on the view that past science must be understood by the lights of present science, most historians of science now writing insist upon some attempt to reconstruct the conceptual and cultural categories of historical actors. This historicism has expressed itself in two important movements in recent historiography of science: the aim to understand 'science in context,' and attempts to write histories of entities previously assumed to be transhistorical.

'Science in context' (which lent its name to a journal) thickens the connections of science with its historical place and time, as well as blurring the boundaries between factors previously regarded as either internal or external to a discipline. Indeed, one of the consequences of the studies of science in context has been to make the emergence, persistence, and occasional disappearance of such boundaries, topics of historical investigation in their own right. One of the most important of the strands of contextual history of science, because of its potentially wide application, has been the investigation of how ideas and practices of gender have shaped the modern sciences (e.g., Haraway, 1989; Jordanova, 1989; Schiebinger, 1993) and the exploration of the household and family as a site of science (e.g., Browne, 2002; Coen, 2007).

Although some critics of this embedded view have feared that context impugns scientific autonomy and therefore validity, many studies have depicted the social, technological, and economic milieu of science as resources for scientific creativity, not merely as ideological contaminants (e.g., Wise, 1989–90). Histories of supposedly transhistorical entities such as experience (Dear, 1995), truth (Shapin, 1994), and objectivity (Porter, 1995; Daston and Galison, 2007) make use of the practices of science, such as the evaluation of testimony and the making of images, to anchor these ethereal abstractions in specific historical circumstances. They are also at least indirectly indebted to the remarkable studies of Georges Canguilhem (Canguilhem, 1966 [1943]) and, especially, Michel Foucault (Foucault, 1969) that challenged the universality and permanence of fundamental modern categories like normalcy and sexuality.

In keeping with the emphasis on scientific practices, recent history of science might be briefly described as materialized and embodied. The material culture of science, ranging from the architecture of the spaces in which science is done (Rheinberger et al., 1997; Galison and Thompson, 1999) to the stuff that science works with (Klein and Spary, 2009), has shifted from the periphery to the center of historical attention. Hands-on experience with objects is no longer understood as a means to a greater end, but as a force that shapes the culture and penetrates into the metaphysics of science (Galison, 1997). Embodiment has entered the history of science through several doors: the study of the interaction of experiment and craft traditions (Smith, 2004), the investigation of the scientific persona in historical context (Lawrence and Shapin, 1998; Shortland and Yeo, 1996), and the making visible of the role of women in science (Rossiter, 1982, 1995). All of these explorations of embodiment challenge the Cartesian image of the disembodied knower, as well as the hagiography of great individuals; they aim at the reconstruction of collective identities of scientists, as well as their consequences for who was allowed to do what kind of science.

By enlarging the geography, chronology, personnel, sites, and subject matter of the history of science to encompass forms of knowledge beyond those recognized as 'scientific' by modern, Western standards, these recent approaches, informed by cultural, gender, material, and economic history, have opened up new research vistas. In one direction, historians of science are expanding into the history of knowledge. Its practitioners have queried the anachronistic assumptions that restrict the gaze of historians to institutions, actors, and genres that seem most continuous with modern ones (e.g., professors at medieval universities who write treatises but not apothecaries or goldsmiths who trade recipes) or the ethnocentric assumptions that posit a European origin of modern science that slowly diffused to other parts of the world (e.g., via Jesuit missions to China). Historians of knowledge have paid particular attention to the points at which local, lay knowledge intersected with cosmopolitan, elite science (e.g., in the exchange of natural history specimens or weather observation). In the other direction, historians of science with premodern and non-Western specialties are making common cause with historians of learning. The history of learning has also emphasized analogies between the practices of scholars and

those of scientists (e.g., between the commonplace book of the Renaissance humanist and the observation notebook of the early modern naturalist) and, more recently, the importance of studying non-Western traditions (e.g., of philological learning bound to the canonical texts of the Sanskrit Vedas or Chinese Confucianism). In contrast to historians of knowledge, who are indifferent if not hostile to attempts to draw a bright line between exoteric knowledge and esoteric science, historians of learning are deeply interested in the history of esoteric scholarly traditions and their criteria of evaluation. But they too resist the strict segregation of learning and science as anachronistic for all but the modern period.

Some have wondered whether the history of science can survive so much pushing and pulling on its object of inquiry to include knowledge and learning alongside science, not to mention the centrifugal forces of interdisciplinary approaches: in addition to its closest neighbors, history and science studies, history of science intersects with sociology, philosophy, anthropology, economics, and literary studies. Yet in this still somewhat undisciplined discipline, the reward of permeable boundaries has been a rare dialogue, ongoing and sometimes heated, among the natural sciences, social sciences, and humanities.

See also: Comte, Auguste (1798–1857); Cultural Studies of Science; Darwin, Charles Robert (1809–82); Enlightenment; Feyerabend, Paul (1924–94); Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Intellectual History; Knowledge, Sociology of; Kuhn, Thomas S. (1922–96); Latour, Bruno (1947–); Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Medicine, History of Western; Merton, Robert K (1910–2003); Modernization and Modernity in History; Norms in Science; Objectivity: Philosophical Aspects; Physical Sciences: History and Sociology; Science and Religion; Science, Sociology of; Science: Constructivist Perspectives, History of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Social Constructivism; Technology, Social Construction of; Vienna Circle: Logical Empiricism; Wittgenstein, Ludwig (1889–1951).

Bibliography

Bachelard, G., 1938. La Formation de l'Esprit Scientifique. Vrin, Paris.
Bernal, J.D., 1939. The Social Function of Science. Macmillan, New York.
Bloor, D., 1991 [1976]. Knowledge and Social Imagery. University of Chicago Press, Chicago, IL.
Browne, J., 2002. Charles Darwin: The Power of Place. Jonathan Cape, London.
Canguilhem, G., 1966 [1943]. Le Normal et le Pathologique. Presses Universitaires de France, Paris.
Carnap, R., 1928. Der Logische Aufbau der Welt. Weltkreis-Verlag, Berlin.
Cartwright, N., 1983. How the Laws of Physics Lie. Clarendon Press, Oxford, UK.
Chemla, K., 2012. Historiography and history of mathematical proof: a research programme. In: Chemla, K. (Ed.), The History of Mathematical Proof in Ancient Traditions. Cambridge University Press, Cambridge UK.
Clagett, M., 1959. The Science of Mechanics in the Middle Ages. University of Wisconsin Press, Madison, WI.
Coen, D., 2007. Vienna in the Age of Uncertainty: Science, Liberalism, and Private Life. University of Chicago Press, Chicago, IL.
Cohen, I.B., 1985. Revolutions in Science. Harvard University Press, Cambridge, MA.
Collins, H., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, Beverly Hills, CA.
Cook, H., 2007. Matters of Exchange: Commerce, Medicine, and Science in the Dutch Golden Age. Yale University Press, New Haven, CT.
Daston, L., Galison, P., 2007. Objectivity. Zone Books, New York.
Dear, P., 1995. Discipline and Experience: The Mathematical Way in the Scientific Revolution. University of Chicago Press, Chicago, IL.
Dear, P., 2009. The history of science and the history of the sciences: George Sarton, *Isis*, and the two cultures. Isis 100, 89–93.
Delbourgo, J., Dew, N. (Eds.), 2008. Science and Empire in the Atlantic World. Routledge, New York.
Duhem, P., 1908. Sōzein ta Phainomena: Essai sur la Notion de Théorie Physique de Platon à Galilée. A. Hermann et Fils, Paris.
Elshakry, M., 2010. When science became Western: historiographical reflections. Isis 101, 98–109.
Fan, F., 2004. British Naturalists in Qing China: Science, Empire, and Cultural Encounter. Harvard University Press, Cambridge, MA.
Foucault, M., 1969. L'Archéologie du Savoir. Éditions Gallimard, Paris.
Galison, P., 1987. How Experiments End. University of Chicago Press, Chicago, IL.
Galison, P., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago, IL.
Galison, P., 1999. Trading zone: coordinating action and belief. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, New York, pp. 137–160.
Galison, P., Thompson, E., 1999. The Architecture of Science. MIT Press, Cambridge, MA.
Gillispie, C.C. (Ed.), 1981. Dictionary of Scientific Biography. Charles Scribners' Sons, New York.
Golinski, J., 1998. Making Natural Knowledge: Constructivism and the History of Science. Cambridge University Press, Cambridge, UK.
Gooding, D., Pinch, T., Schaffer, S. (Eds.), 1993. The Uses of Experiment: Studies in the Natural Sciences. Cambridge University Press, Cambridge, UK.
Graham, L., Lepenies, W., Weingart, P. (Eds.), 1983. Functions and Uses of Disciplinary History. D. Reidel, Dordrecht, The Netherlands.
Hacking, I., 1983. Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge University Press, Cambridge, UK.
Hacking, I., 1999. The Social Construction of What? Harvard University Press, Cambridge, MA.
Haraway, D.J., 1989. Primate Visions: Gender, Race, and Nature in the World of Modern Science. Routledge, New York.
Hessen, B., 1931. The social and economic roots of Newton's 'principia'. In: Bukharin, N.I. (Ed.), Science at the Crossroads. Kniga, London, pp. 146–212.
History of Science Society, 1967. Isis Current Bibliography (Originally Isis Critical Bibliography). University of Chicago Press, Chicago, IL.
Holton, G., 2009. George Sarton, his *Isis*, and the Aftermath. Isis 100, 89–93.
Jordanova, L., 1989. Sexual Visions: Images of Gender in Science and Medicine between the Eighteenth and Twentieth Centuries. University of Wisconsin Press, Madison, WI.
Jorland, G., 1981. La Science dans la Philosophie: Les Recherches Épistémologiques d'Alexandre Koyré. Éditions Gallimard, Paris.
Klein, U., Spary, E., 2009. Materials and Expertise in Early Modern Europe: Between Market and Laboratory. University of Chicago Press, Chicago, IL.
Koyré, A., 1939–40. Études Galiléennes. Hermann, Paris.
Koyré, A., 1961. La Révolution Astronomique: Copernic–Kepler– Borelli. Hermann, Paris.
Kragh, H., 1987. An Introduction to the Historiography of Science. Cambridge University Press, Cambridge, UK.
Kuhn, T.S., 1970 [1962]. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.
Kuhn, T.S., 1972 [1968]. History of science. In: Sills, D.L. (Ed.), International Encyclopedia of the Social Sciences, vol. 13. Macmillan and Free Press, London and New York, pp. 74–83.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.
Latour, B., Woolgar, S., 1986 [1979]. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton, NJ.
Laudan, R., 1993. Histories of the sciences and their uses: a review to 1913. History of Science 31, 1–34.
Lawrence, C., Shapin, S. (Eds.), 1998. Science Incarnate: Historical Embodiments of Natural Knowledge. University of Chicago Press, Chicago, IL.
Lindberg, D., Westman, R. (Eds.), 1990. Reappraisals of the Scientific Revolution. Cambridge University Press, Cambridge, UK.
Livingstone, D.N., 2003. Putting Science in Its Place: Geographies of Scientific Knowledge. University of Chicago Press, Chicago.
Lloyd, G.E.R., 2004. Ancient Worlds, Modern Reflections. Philosophical Perspectives on Greek and Chinese Science and Culture. Clarendon Press, Oxford, UK.

Maier, A., 1949–58. Studien zur Naturphilosophie der Spätscholastik. Edizioni di Storia e Letteratura, Rome.
Merton, R.K., 1938. Science, technology and society in seventeenth-century England. Osiris 4, 360–632.
Needham, J., et al., 1954. Science and Civilization in China. Cambridge University Press, Cambridge, UK.
Osler, M.J. (Ed.), 2000. Rethinking the Scientific Revolution. Cambridge University Press, Cambridge, UK.
Park, K., Daston, L. (Eds.), 2006. Early Modern Science. The Cambridge History of Modern Science, vol. 3. Cambridge University Press, New York, NY.
Pestre, D., 1995. Pour une histoire sociale et culturelle des sciences: nouvelles définitions, nouveaux objets, nouvelles pratiques. Annales. Histoire, Sciences Sociales 50, 487–522.
Pickstone, J., 2000. Ways of Knowing: A New History of Science, Technology and Medicine. University of Manchester Press, Manchester.
Polanyi, M., 1958. Personal Knowledge: Towards a Post-critical Philosophy. University of Chicago Press, Chicago, IL.
Porter, R., Teich, M. (Eds.), 1992. The Scientific Revolution in National Context. Cambridge University Press, Cambridge, UK.
Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.
Pyenson, L., Verbruggen, C., 2009. Ego and the international: the modernist circle of George Sarton. Isis 100, 60–78.
Raj, K., 2006. Relocating Modern Science: Circulation and the Construction of Knowledge in South Asia and Europe. Permanent Black, Delhi.
Rheinberger, H.J., Hagner, M., Wahrig-Schmidt, B. (Eds.), 1997. Räume des Wissens: Repräsentation, Codierung, Spur. Akademie-Verlag, Berlin.
Rossiter, M.W., 1982. Women Scientists in America: Struggles and Strategies to 1940. Johns Hopkins University Press, Baltimore, MD.
Rossiter, M.W., 1995. Women Scientists in America: Before Affirmative Action, 1940–1972. Johns Hopkins University Press, Baltimore, MD.
Sabra, A.I., 1996. Situating Arabic science. Isis 82, 654–670.
Sarton, G., 1948 [1927]. Introduction to the History of Science. Williams & Wilkins, Baltimore, MD.
Schaffer, S., 1992. Late Victorian metrology and its instrumentation: a manufactory of ohms. In: Bud, R., Cozzens, S. (Eds.), Invisible Connections: Instruments, Institutions, and Science. SPIE Optical Engineering Press, Bellingham, WA, pp. 23–56.
Schiebinger, L., 1993. Nature's Body: Gender in the Making of Modern Science. Beacon Press, Boston, MA.
Secord, J., 2004. Knowledge in transit. Isis 95, 654–672.
Shapin, S., 1988. Understanding the Merton thesis. Isis 79, 594--605.
Shapin, S., 1992. Discipline and bounding: the history and sociology of science as seen through the externalism–internalism debate. History of Science 30, 333–369.
Shapin, S., 1994. A Social History of Truth: Civility and Science in Seventeenth-Century England. University of Chicago Press, Chicago, IL.
Shapin, S., 1996. The Scientific Revolution. University of Chicago Press, Chicago.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press, Princeton, NJ.
Shortland, M., Yeo, R. (Eds.), 1996. Telling Lives in Science: Essays on Scientific Biography. Cambridge University Press, Cambridge, UK.
Sivasundaram, S., 2010. Sciences and the global: on methods, questions, and theory. Isis 101, 146–158.
Smith, P., 2004. The Body of the Artisan: Art and Experience in the Scientific Revolution. University of Chicago Press, Chicago, IL.
Thackray, A., Merton, R.K., 1972. On discipline building: the paradoxes of George Sarton. Isis 63, 473–495.
Wise, M.N., 1989–90. Work and waste: political economy and natural philosophy in nineteenth century Britain. History of Science 27, 263–301, 391–449; 28: 221–261.
Wise, M.N., 1996. The enemy within and the enemy without. Isis 87, 323–327.
Wittgenstein, L., 1953. Philosophical Investigations (G.E.M. Anscombe, Trans.). Blackwell, Oxford, UK.

Science, New Forms of

Jerome R Ravetz, Institute for Science, Innovation and Society, University of Oxford, Oxford, UK
Silvio O Funtowicz, Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

New forms of activity in science are appearing rapidly, due to the interaction of the new technologies of information with new political currents. Traditional practices, structures, and assumptions in key phases of the cycle of knowledge production, including resources, persons, procedures, property, publicity, quality control, and consequences have been challenged or transformed. Political activism within science has begun, as with the boycott of Elsevier; and the rigid ideas of 'scientist' or 'the paper' are dissolving. The future is highly challenging.

Preface

It is well known that the new technologies of information are revolutionizing practice in a great variety of fields, now even including the cheap manufacture of material objects. The powers of research science in very many fields are being enhanced and even transformed. Generally these changes in technical practice take place within social and conceptual structures that were inherited from previous epochs in science. They are important new developments, but not necessarily 'new forms' of science. Here we consider the practices and structures that cause surprise and concern when they are noticed, and that have their own momentum to continue rapid change. Generally they show the effects of the new technologies of information, combined with ideas and practices imported from the wider political sphere. Since this is all happening so recently and so rapidly, we have not been able to base our analysis on a body of reflective scholarly literature. Also, we assume a general familiarity with the phenomena and the potential of the new information technologies, taking insights from the standard contemporary sources while discounting their more speculative conclusions. Because of these constraints, this article will be more in the form of an analysis with examples, rather than a review and synthesis of literature. This is not a standard form for an encyclopedia, but the unusual subject matter requires it.

Historical Background

About a half century ago, there was awareness among scientists that things had changed. The transition was captured in the seminal book *Little Science*, *Big Science*, by Derek Price (1963), which reflected on the rapid expansion of science, particularly in the USA, and its harnessing to production and war as in the Manhattan Project. Over the following decades, this new situation became the occasion for many reflections, along with new lines of debate, which developed gradually. The current situation of science is a very rapid change. Attitudes and activities that were unthinkable only a decade ago are now well established. Who would have imagined that eminent scientists would organize a boycott of a leading scientific publisher? But that is what the leading mathematician, Timothy Gowers (2012), did to Elsevier, with the support of thousands, with the slogan 'Academic Spring.'

The changing self-consciousness of scientists during that previous period is a reminder that, in the sociotechnical system of science, the intellectual and material aspects of the system are deeply implicated in one another. As we move, perhaps headlong, into a new set of material and the conceptual relationships, it is helpful to gain perspective by going further back than 'big science.' Since it is the means of communication that are most obviously being transformed by 'digital' media, we can recall the impact on science of printing during the Renaissance. This enabled the diffusion of knowledge in obvious ways, feeding its growth. The publication of a book enabled a practitioner to advertise his talents, and thereby to obtain the best patronage available on the market. Agricola, Biringuccio, and Vesalius are great examples. Galileo exploited this technique, parlaying his *Sidereus Nuncius* first into a pay rise in Venice, and then to a coveted position as 'Chief Mathematician and Philosopher' at the Medici's court back home in Florence.

By the end of the sixteenth century, the great national industries of conquest, war, and royal display gave employment as never before to practitioners with mathematical skills. The traditional status barriers between 'liberal' knowledge and 'mechanical' practice were lowered, for a few generations at least. Simon Stevin was a genius who worked as a practitioner but showed philosophical ambitions, while Galileo was just on the other side of the divide and determined to stay there. William Gilbert was the first to unify 'philosophy' and experimentation in his *de Magnete*; there he also rigorously reported which experiments worked and which did not. Bacon's rhetoric about 'torturing nature' became a successful practice among the 'experimental philosophers,' most notably in the great English school with Boyle, Hooke, and Newton. By the eighteenth century, the union of theory and practice was fully established, and philosophers in the 'highland zone' of Britain combined ideology and improvement (with profit where obtainable) to great effect.

As the powers of technology grew through the nineteenth century, scientists (who then got the title) could study natural powers in artificially controlled situations, notably in relation to electricity and magnetism but also to both old and novel chemical substances. The cross-fertilization of heavy industry with mission-oriented scientific research was established, first in Germany (starting with Liebig) and then elsewhere, notably

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85027-8

in the USA in the great labs dominated by such iconic inventors as Edison and Steinmetz, along with the quasi-magical Tesla. During the same period, the university system was forced to graft research onto its traditional roles of teaching and socializing the young elite. One later product of this merger was a collective amnesia among scientists and scholars about the role of such organized, industrialized research and development during the 'little science' period, a misunderstanding that led to much confusion when 'big science' arrived in full force.

In the epoch that was inaugurated by the Manhattan Project and is now just ending, science became 'industrialized' in several respects: in overall size, in scale of individual enterprises, and in its ever closer relations with industry. It also became 'incorporated' (Rose and Rose, 1969), that is, involved in a variety of ways in the affairs of the state, far beyond the traditional connection with warfare. Somewhat naively, scientists believed that they were being asked to speak 'truth to power,' when their clients frequently wanted 'policy-based evidence.' Traditional stable subject-specialty research has been displaced by 'Mode 2' (Gibbons et al., 1994), in which scientists are fungible units of production, to be deployed or discarded as the needs of mission-oriented projects dictate. In its relations to the broader society, the manipulations and distortions of research that are now familiar in connection with profit-making industry (such as research on tobacco, food, and pharmaceuticals *inter alia*) are mirrored in the less notorious, but equally important practices of ministries and agencies. The 'Gemeinschaft' research communities of yesteryear survive only on the margins, although Steven Shapin has argued that they are being recreated out of necessity in the fast-moving fields of innovative entrepreneurial science (Shapin, 2008).

Structures of Scientific Activity

Having very briefly reviewed the historical background, we can now sketch the relevant structures of scientific activity. Then we will be equipped to understand the influence of the new technologies on the new forms of scientific practice. It is now commonly appreciated that the old picture of 'the scientist' making a discovery and then turning it over to 'society' is very radically oversimplified. First, the possible motivations for engaging in, or supporting, research are varied. Aside from the traditional 'curiosity,' there is also 'mission-oriented,' 'mandated,' 'regulatory,' and (we must now add) 'critical' motivation. The relations with broader society, the internal social structure, and the ruling criteria of adequacy and value, all depend on such initial motivation. By appreciating the basic variety in the production systems of scientific knowledge, we can then consider the cycle of activities through which knowledge production goes.

The initial motivation is but the first step in a multiphased cycle. We might call this first step the 'issue,' as that is where policy-relevant science gets its start. Next comes 'policy,' which expresses the hoped-for solution and defines the parameters of the design of the project. This is most easily seen in research on risks; for 'the risk' is itself always complex, and simplifying design choices must be made on the subjects to be studied and on the nature of the hazard they experience (e.g., adults, children, pregnant mothers, and chronic or acute doses, average or peak exposure, etc.). And 'problem,' giving greater specificity to the inquiry, perhaps stating a statistical hypothesis to be tested, comes soon after. Quite essential at this early stage are 'resources.' Since science does not operate in an ideal universe, this issue inevitably brings in politics, in the classic sense of 'who gets what, how, and why.' Politics also enters with 'persons.' Whenever policy is at stake, this can be quite crucial; even if scientists all have integrity, they act rather like judges, and this does not prevent them from leaning this way or that when confronted with uncertainty and value conflicts. This step has become problematic in several ways recently, and is closely related to the new politics of science. Then come 'procedures,' which are necessary because research is never straightforward, but is rather the study of the 'proxies' for the conceptual things and events that are of concern. In all such cases, there are numerous standards and conventions defining good practice, which are designed and implemented for avoiding the known pitfalls of data creation, interpretation, and inference. A high-quality scientific paper contains dense descriptions of those procedures, thereby assuring its readers that its results are robust. At the core of the investigation there is 'production,' where someone interacts with equipment (or respondents) to obtain the primary data, which will be the intended representatives of the proxies for the defined objects of inquiry. Out of this comes 'product,' realized in an 'inscription' which may function as 'paper,' report, or patent application. Beyond that there is 'property,' which also includes control over 'publicity.' It was once assumed that the public sharing of results was definitional of science, as opposed to the proprietary control over invention. But it has recently emerged that the publication of data is by no means universal in spite of being formally required by many funders and journals. Of course, when data are withheld it is impossible to operate Merton's norms of 'communalism' and 'organized skepticism,' which had been accepted as fundamental to 'pure' science. Near the end of the cycle are the traditional 'applications' and the recently recognized 'consequences.' These are now recognized as critical to the inquiry, and will include the 'bads,' or 'unintended consequences' that arise from the applications of scientific results. The relevant sciences here are almost polar opposite in character from the traditional lab disciplines. For the problems are usually 'postnormal' (Funtowicz and Ravetz, 1993), in having high uncertainties and high decision stakes, politicized debates over the assignment of burdens of proof, plus imbalances of prestige and of conventional institutional funding, with whistle-blowers sometimes being as important as researchers, and open conflict breaking out between pressure groups and vested interests.

Such a cycle helps to explain the complexity and confusion that attends debates when science is involved with policy. Without a shared awareness of the different phases of the production cycle, debates can wander inconclusively through a maze of topics. Moreover, each of those phases of the production cycle has its own 'quality control cycle.' This feature establishes the true complexity of the process. For whenever there is an action conducted within an organized system, it will be subject to 'control' for its quality. This operation will include particular agents, operating particular criteria and adopting particular procedures for controlling and evaluating quality. Whenever a regulatory body comes to public notice, it will be seen that it has these elements: agents, criteria, and procedures.

Moreover, the control cycle itself is subject to control cycles. Let this seem paradoxical, we recall the Latin epigram, "Quis custodiet ipsos custodes?" – who guards the guardians? Sheila Jasanoff first called attention to this phenomenon in science advice in her pioneering work *The Fifth Branch* (1990). Such an iteration is familiar in connection with the administration of justice, with a sequence of courts of appeal, right up to some highest court from which some political influence cannot be excluded. More to our present concerns, as our companion article in this Encyclopedia argues (*see* Peer Review and Quality Control), quality has replaced truth as the effective guiding principle for science. The 'quis custodiet' iteration is a reminder that quality cannot be maintained by enforcement alone. It requires an ethical commitment, one that must be seen to be practiced at the top of any institution. In its absence, corruption, and vacuous research are sure to follow. We have previously remarked on the paradox that the successful production of objective scientific knowledge depends critically on this subjective ethical commitment (Ravetz, 1971: Chapter 11).

Persons

We can now look at the production cycle of scientific knowledge in greater detail, identifying important points of novelty. In recent decades there has been a growing recognition that policy-relevant science is done better if the early stages, as issues and even problems, draw on broader experience than that of politicians, science advisors, and experts. Some time ago the term 'extended peer community' was coined (Funtowicz and Ravetz, 1993) and, since then, under many different names, the involvement of citizens has become increasingly accepted as desirable and useful. The obvious constitutional problems presented by these exercises in 'participation' have somehow not proved insuperable. Particularly on public health issues, broader participation in problem definition has become established, starting with the campaigns by AIDS activists (Epstein, 1996). Such developments have diluted the original force of the very idea of 'scientists,' as the sorts of persons involved in the research cycle become ever broader. In conventional scientific research, there is a long continuous tradition of amateurs making contributions, but more recently they had been relegated to the role of lower-status assistants. Such were the neglected sisters of astronomers in the past, and the numerous bird watchers in the present. That approach to broader involvement is the least threatening to established orders, and it motivated the pioneering 'Galaxy Zoo.' In that case, the groups of amateurs even suggested topics for closer investigation. But a more active involvement is developing all the time, such as in the Web sites for folding of molecules, and even for solving difficult mathematical problems. In another direction, there is 'garage biology,' still very small and marginal but with great potential. Now Web sites, such as Petridish.org (available at: www.petridish.org, accessed 29.08.13.), are being created to serve scientists who want to bypass the established channels for resources; although innocent in intent, such developments inevitably will affect the politics of resource distribution in science. More immediately serious is the parallel universe of the informally trained hackers of information technology (IT), including everything from pranksters and libertarians over to gangsters and state agents. As a result of all such developments, the status of 'scientist' is being progressively diluted; one no longer needs to have gone through the lengthy process of selection and training, culminating in a PhD at the top of the cleverness pyramid, in order to be accepted as a member of a recognized knowledge-producing community. The status of 'expert,' someone who has authority to solve particular problems on the basis of his specialized scientific training, is being eroded even more rapidly. With the growing contestation of science-policy issues (ranging from nuclear power to vaccination), along with the increasing use of crowd-sourcing for decisions, expertise is becoming a beleaguered role.

A most important recent development has been the loss of the presumption of Mertonian 'disinterestedness' in someone certified as a scientist. Authors of papers are now routinely required to declare any potential conflicts of interest, on the assumption that these are relevant to their claims. Also, research done with industrial support has its quality discounted, on the basis of presumed bias that is supported by surveys of outcomes (Lundh et al., 2012), to the point where some firms are abandoning public research. Since a very large proportion of ostensibly scientific research has been supported by industry and the military, this new awareness can destabilize long-established patterns.

Procedures

A similar process of change, variously seen as erosion or democratization, affects procedures. For a long time, there was a common belief in a scientific method, those procedures whereby science unerringly produced the correct answers to its problems. It was the job of history of science to show that science always got it right, and of the philosophy of science to show how this was so. With the rise of uncertainty, first in theoretical physics, and then of complexity in knowledge and policy, this motivating faith gradually lost plausibility. The inability to specify a method that is not logically fallacious has contributed to the difficulty. Popper's philosophy of 'falsificationism,' so valuable as a moral injunction, turned out to be vacuous for the construction of knowledge. Once scientific practice is dominated by statistics, there is no question of 'truth' as traditionally understood, for the results depend on counterintuitive and obscure reasonings about null hypotheses and confidence limits. Even the standard statistical techniques, applied unreflectively in all of 'normal' science, have been severely criticized, as by Ziliak and McCloskey (2008). As for models, which absorb ever more research effort in all fields, and squeeze out the more expensive and time-consuming traditional methodologies, the best that can be said is the classic line 'all models are wrong; some are useful.' The 'quis custodiet' iterative principle applies most strongly to procedures. The products of inquiry are subjected to direct testing, however imperfect, but the testing of procedures is of an entirely different order of complexity and difficulty. It is not surprising that one author (Ioannides, 2005) has found multiple examples of shoddy and low-grade work leading to unreliable and unworthy science. It is impossible to say how much of this depressing development is actually new;

certainly, poor quality science has always been with us, as has been noted in the past (Ravetz, 1971: Chapter 10). But there does seem to be a much-heightened awareness of these problems within scientific communities, and this could simultaneously have the opposing effects of stimulating reform while also reinforcing cynicism.

Property

In the previous epoch, the social relations of property had not been changing with great rapidity; there was a steady displacement of the traditional 'public knowledge' by what we might call 'corporate know-how.' However, there has been a recent recognition of a deeply paradoxical situation in the management of intellectual property. In the new emerging technologies, notably in nanotechnology, there is a confused mixture of systems of intellectual property. Much work is done on public funds and in the public domain, while much else is proprietary. And this division is not restricted to results, but also to the ancillary information that is essential for the research. There has been a recent warning that the progress of the field is being seriously impaired by the thickets of property rights that surround the various bits of knowledge and technique that are necessary for research. Thus, right here in science we have an example of Marx's description of the replacement of one mode of production by another in the *Communist Manifesto*: "the feudal relations of property became no longer compatible with the already developed productive forces; they became so many fetters. They had to be burst asunder; they were burst asunder." We can see a sort of parallel in science, whereby the small-scale semiindependent handicraft researcher has recently been displaced by the knowledge-worker in industrialized, or 'Mode 2,' science. Marx saw the boom-and-bust cycle of modern industrial capitalism as a sign of a similar deep systemic crisis for itself, and imagined that the revolutionary proletariat would soon step in. That was not to be, at least on his timescale. However, it is now widely recognized that in order to maintain the momentum of scientific innovation, the self-strangling hybrid system of intellectual property requires supplanting by the new social relations of production, characterized as 'open source' or 'creative commons' on the example of the Intellectual Property (IP) systems of some key Information Technology (IT) industries (Igoe and Mota, 2011; Brooks, 2012). Without indulging in rhetoric about 'bursting asunder' the fetters of the IP rights enshrined in patents, we can see that deep change is inevitable and already underway. Even Big Pharma has taken the point, as with the Gates Foundation on Tuberculosis, and GlaxoSmithKline in its Tres Cantos Open Lab Foundation (Nathan, 2012). Under these new circumstances, the 'gravediggers of (scientific) capitalism' will not be Marx's proletarians 'with nothing to lose but their chains,' but rather Clay Shirky's possessors of 'cognitive surplus, spreading creativity, and generosity in a connected age' of digital knowledge (2010).

Publicity

Innovation is even more rapid in the phase of publicity, which is after all the lifeblood of science and consisting of information. What is significant for the self-consciousness of science is the sudden discovery that there actually is an urgent and deeply problematic political economy of publicity. Previously accepted as the unproblematic norm with occasional deviations, publicity has rapidly become the focus of concern and even of grievance. This issue, even more than property, may become the lever whereby the idealistic ethical assumptions of 'little science' come to be seen rather like liberty in the political sphere, not to be assumed but requiring vigilance and struggle for their preservation. The distinguished mathematician invoking Tahrir Square against the publisher Elsevier may be iconic for the emerging contestations over public knowledge in science. The issue of secrecy and publicity emerged almost incidentally in the rather troubled evolution of the well-intentioned science of global warming. Leading scientists were affronted by demands for their data made by unqualified critical outsiders. Eventually it required freedom of information procedures against individuals and institutions, in the USA and UK, to unlock data that had been produced on public funds. The critics could not be isolated and neutralized, as in previous debates (notably over bovine spongiform encephalopathy (BSE) or Mad cow disease), as they formed an 'extended peer community' in the 'blogosphere,' the most popular site enjoying both a readership in the millions and prizes for quality (see www.wattsupwiththat.com, accessed 30.08.13). The issue of secrecy of publicly funded data became so urgent that there was an official response from the Royal Society of London, in the form of a working party chaired by Professor Geoffrey Boulton (Royal Society Policy Centre, 2012). Its report strongly recommended procedures for insuring that such data should be public, although it did not extend its recommendations to the computer codes whereby such raw data were converted to usable information. In the ensuing discussion, it emerged that many rules for data publicity already existed, but that they were generally ignored by all parties to the publication process. The awareness of the problem soon extended from research data to those used in the regulatory process. It became realized that public regulators commonly make decisions based on secret data. In the atmosphere of mutual suspicion, the principle was invoked, "if they are hiding something, they must have something to hide." Accusations of possible malpractice by leading pharmaceutical companies, and of serious incompetence in data management by European regulators, have been made in the most authoritative quarters (Goldacre et al., 2012). It is only natural that the issues of publicity and quality should become entangled in this way. The most effective way to seize the moral high ground in a science-policy debate is to call for transparency and openness, thereby invoking the traditional norms of science combined with the modern imperatives of participatory democracy.

On another front, ethical issues in fairness to access to information have suddenly become urgent. Publication is not free, nor even particularly cheap. How are the publishers to be recompensed? The problems of cost and recompense in the digital age that had already afflicted the creative industries have suddenly arrived in science. Previous cozy arrangements whereby academic libraries subsidized publications through their subscriptions (and also provided a hefty profit) no longer go unchallenged. Also, it is suddenly realized that there are many worthy people out there who need access to nominally public

information, and who cannot or even should not pay large fees for the privilege of sampling it. All that information waiting there on the Internet, or cloud, turns out to be free only for those who are subsidized by an institutional library that can afford the fees for an extensive license. Without membership of such a wealthy institution, the would-be researcher must produce significant payments for each item of essential information in the nominally 'public knowledge' sector, even for research to which they have contributed through their taxes. The publishers are now suddenly subjected to outrage, as academics accuse them of getting all their editorial services for free and then securing superprofits through their quasi-monopoly control of high-prestige outlets. The whole social system of scientific publication, which had been a model of success for generations, performing a great variety of functions (diffusion, archive, quality control, evaluation, and allocation of resources) is suddenly required to rethink itself. And this crisis occurs in conjunction with the arrival of the new technologies of publicity, which threaten the monopoly of paper publication with all its very useful structures and constraints. We should not underestimate the severity of this challenge to the social institution of science as we have known it.

In science, as in other areas of knowledge production, the new technologies dissolve previous boundaries. Previously the forms of publicity were atomized, in harmony with the production process and the conceptual objects. That is, 'the paper' was realized in a congealed text, printed in multiple identical copies on paper. It was the product of a defined original study by 'an author' (or, in industrialized conditions, a defined collection of authors). That study had its own distinct closed cycle from inception to completion, as reported in the paper. Even quality control was atomized, with anonymous reviewers reporting to a single editor, determining the fate of 'the paper.' The social conditions of industrialized knowledge production, particularly in relation to the reward system, which determined career opportunities, militated against leisurely investigations, as of the traditional natural history or even the reflective natural philosophy characteristic of 'gentleman-amateur' science. Accordingly, the atomized quick-returns project has come to dominate all aspects of the cycle. A parallel development in what can be called the practical ontology of science fitted in with this style: the effort was to study simple, or simplified, systems, using conceptual objects and tools, usually statistical and mathematical, which enabled simple judgments to be made. Thus among biologists the emphasis became to focus on the smallest scales possible, from biochemistry down to genetics; whole, complex systems were deemed incapable of scientific study, as indeed they were under this definition of science. The process can be seen in its extreme parodic form in mainstream economics, where abstract mathematical models of the decisions of mythical atomic actors served to reinforce the politically driven assumptions of the discipline. In strong contrast, thanks to digital publication, the ongoing inquiry can now be a focus of a continuous flow of communal dialogue and development, in which there may be a temporary crystallization of a 'product' as a matter of convenience. Leading journals in all fields now run blogs on their Web sites, providing alternate and informal channels of communication and dialogue. Thus, all the previous atomized social structures of quality, social control, publicity and access, and by extension the practical ontology of science, are thrown into confusion. Along with 'interdisciplinarity,' complexity is now becoming a keyword in all sciences relevant to policy. We see here how all the different aspects of the production of scientific knowledge are mutually implicated in this transformation; starting with publicity, we have just shown how research practice and even practical ontology are affected.

Quality Control

The effects of the new technologies on the quality control cycle are, if anything, even more dramatic. Traditionally, the quality assurance systems of science were run largely informally and largely confidentially. The mere fact of publication of a paper was taken to be its stamp of quality. All the processes of criticism and improvement were kept secret; the atomic unit of established knowledge embodied in the paper carried no penumbra of uncertainty or quality. Of course this system generally performed well, but it is highly sensitive to the quality of the quality-guardians and even more so of theirs. It is all too easy for peer review to degenerate into 'pal review.' For quite some time, the deficiencies of the peer-review system have been aired and considered, and the conclusion has always been that, rather like democracy, it is the worst possible system except for all the others. The technology of information contributed to this stasis: the sheer bulk of papers containing reviews, reports, and revisions for a single publication would be inconvenient and expensive to make available for a wider scrutiny. Opening the files of documents on all publications would be impossible. But with the new technologies of information, we have new devices. Perhaps, the most revolutionary of all the 'apps' recently invented is the Wikipedia. For, as Wikipedia itself has shown, with Wiki all the editorial processes can be both convenient and transparent. On Wikipedia, every article appears with a sort of pedigree, and cocreators are invited in to contribute to the public quality assurance process. Of course the process can be, and often is, abused, but then the abuses are usually spotted quickly and a lively open debate ensues. In many ways, this approach is being incorporated into research science. What had hitherto been a remarkably closed institution, with little public scrutiny of its internal politics, and hardly any investigative journalism (with Mae-Wan Ho and Jonathan Latham providing the main exceptions), is now starting to be opened up. We can hope that the transition to effective 'openness and transparency' in the governance of science, starting with quality assurance, will be smooth; but it must come, and will come.

Consequences

Near the end of the cycle is consequences. Traditionally, it had been assumed that this had nothing to do with scientific discovery. The consequences of science were assumed to be essentially good and overwhelmingly benign in practice, and so the scientist could, with a good conscience, turn over his products to society for development and control. During the twentieth century, first in war and then in the fields of safety, health, and environment, consequences obtruded with increasing salience. A natural reaction was to bring all those

issues under scientific control, with 'risk' becoming the dominant concept, displacing 'safety' and 'danger' in scientific discourse. This was defined as the product of the quantified probability of harm (always assuming that these had precise measures). There was even an ethical application: if a novel risk was no greater, quantitatively, than one already accepted, then only those who were misguided or malevolent would reject the new risk. Such analyses were applied in the early debates over civil nuclear power, and were discredited, along with the industry itself, by the disasters in the USA and USSR. By the end of the century, the inescapable presence of risk (still not 'danger'), along with its characteristic politics, was analyzed in the classic work *The Risk Society* (Beck, 1992). Now the issue of consequences is recognized as crucial in policy debates on new technologies, be they genetically manipulated crops or nanotechnology, as well as in broader policy debates, such as those on climate change resulting from industrialization. However, the science that is universally accepted as necessary finds it very hard to develop strength that is adequate to its tasks.

To understand this imbalance, we can see this new task as dealing with the bads, rather than with the goods that are the traditional goals of science. These problems will usually be 'postnormal' in having significant uncertainties and decision stakes. Even when there is little scientific uncertainty, vested interests can delay action by manipulating all relevant processes. The US tobacco industry, 'manufacturing doubt,' is the classic case; more are to be found in the seminal work *Late Lessons from Early Warnings* (European Environment Agency, 2002). Then, normally the bads are unintended consequences of mainstream science/technology done in the standard mainstream atomized way. For their study, the relevant fields will usually be marginal, and weak institutionally and politically, and perhaps even weak scientifically as well. Without any conscious intention, the rules of the mainstream scientific game are rigged against those who would pursue the science of bads. All this is changing rapidly under the new conditions of scientific practice. First, there is a structural asymmetry in the opposing sides of any debate on possible bads. The proposers are concentrated and organized, institutionally and personally. The opposition is typically a coalition of scattered aggrieved citizens joined up with underfunded NGOs. But just as the printing press and the rotary duplicating machine facilitated protest in earlier ages, the mobile phones and social media do so now. Also, since the protection of bads involves secrecy, the powers of new information technology to penetrate records and expose secrets redress imbalances of knowledge and hence of power. Further, in modern society there are many protestors with education and applicable skills, so the traditional class divides are dissolved. The technique of 'Community-Based Audit,' where the data, methods, and paradigms of proponents of developments are scientifically scrutinized by scientist-citizens and citizen-scientists working together as a preliminary to dialogue, is a valuable contribution to the science of bads (Tattersall, 2010). We will know that science has genuinely entered its new epoch, when the science of bads becomes recognized as a subject worthy of being taught as part of a liberal education in science. A good start has been made in the announcement of the 'Oxford Principles' for responsibility in the governance of geoengineering research (Oxford Geoengineering Programme, 2012). They could provide the same service for the science of bads, that the 'Mertonian norms' for scientific integrity have done for traditional research.

A Problematic Case

A recent case encapsulates all these issues in consequences, or the science of bads, at the intersection of state regulation, populist politics, and information technology. It started with a thoroughly traditional debate on bads, in this case the claimed possible harmful effects of a strain of genetically modified (GM) maize, which is intended for use as animal feed in the European Union (EU). Although it had been cleared by the European Food Safety Agency (EFSA), powerful forces in France were opposed to it, and they derived support from a scientist at Caen named Séralini, whose tests, he claimed, showed danger (Butler, 2012). He had completed a new series of lengthy feeding tests, and had ambushed the opposition in its publication. The respectable scientific journal publishing the paper gave no advance notice to journalists, and the publication was accompanied by a film! In the row that followed, the critical statement by the EFSA was published on their Web site, and nothing could prevent the scientist from stating his rebuttal, at length, in the commentary section. Among the many criticisms of his work made by reputable scientists was one that alleged that he had not released his primary data. He promptly converted this into a challenge to the EFSA to release the data on whose basis they had decided that the product was safe. Why indeed were they so secretive? The EFSA soon called his bluff, launching an initiative for publication of data that had actually been planned for some time (EFSA, 2013). As ever-more authoritative criticisms piled in, something entirely different happened. This was a proposal by a group of anti-GM scientists in Russia, to perform a 'rat reality show' (Sharoykina, 2012). This would be a year-long Internet–TV 24/7 real-time feeding experiment to test whether GM products are harmful. Feeding procedures would be appropriately blinded, and there would be supervision by a committee including all tendencies in the debate. The sponsors claimed that in this way the public could have a trustworthy experiment to witness, and could then draw their own conclusions on the safety of GM products. If the group cannot obtain resources for the show through orthodox channels, they plan to crowd-source the funding. There was no immediate response to this proposal. Is it a stunt? An opening to scientific demagoguery? Or a harbinger of a real 'science by the people'? Should we now start to speak of Zapatista science? For now, there is no way to know. But now that political power within science is diffused and perhaps also confused, the future is not foreclosed.

Citizen Science – an Emerging New Form of Science?

It is useful to draw together the threads of the previous discussions of citizens' activity in science, and to consider whether this is actually an embryonic new form of science. If so it would be a successor to the earlier phases, going back from industrialized/incorporated/academic to practitioner/gentleman-amateur. This new practice is based on a rapidly growing technology of social communication, has precedents in the 'creative commons' culture in IT, and draws support from

a great variety of institutions. Its own varieties of practice in relation to mainstream science have been described as: contribution, collaboration, and cocreation. The enormous differences in social organization between these roles are obvious. It can provide support for mainstream research, opportunities for individual creativity, and also means for opposition and protest. In this latter role, it can connect to the radical social-media movements, focused on issues of environmental and broader social justice. Is it developing a consciousness of itself and for itself, distinct from other forms of scientific practice? It is still too early to tell, and so far there is no sign of the internal conflicts that would inevitably accompany any attempt at self-definition. As it grows, it will need to establish procedures for quality-assurance, for excluding cranks and malefactors, and for relating to 'mainstream' science. Whatever happens, it is highly likely that in the next edition of this Encyclopedia, 'citizen science' will not be an afterthought.

Conclusion

We must keep a sense of perspective. All these developments in digital-age science, especially those on the populist fringe, are very much on the margin. The vast majority of scientific work is done just as it has been, in large labs with traditional structures of quality control. The immediate problems facing the scientific communities, such as deceleration of growth, shifting of activity to the East, disentangling the publications system, and shoring up the quality assurance system, continue to require attention from all concerned. Creativity flourishes, in a spate of new inventions as well as in debates on the scientific problems bordering philosophy and theology. But this article is devoted to 'new forms of science,' and so long as they are new they will by definition be on the margins, uncertain, indeed speculative in their future course. We might adopt the title of a recent prophetic book, *Science 3.0* (Miedema, 2012), as the motto for our journey into this unknown and challenging future.

See also: Chemical Science: History and Sociology; Citizen Science; Industrialization; Modernization and Modernity in History; Norms in Science; Objectivity of Research: Ethical Aspects; Patenting in the Biological Sciences; Peer Review and Quality Control; Physical Sciences: History and Sociology; Renaissance; Science Communication; Science and Law; Science and Technology Studies: Experts and Expertise; Science and The State; Science, History of; Science, Public Engagement With; Truth and Credibility in Science.

Bibliography

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.
Brooks, M., 8 December 2012. Welcome to the personal drone revolution. New Scientist 216 (2894), 42–45.
Butler, D., 2012. Hyped GM maize faces growing scrutiny. Nature 490 (158). http://www.nature.com/news/hyped-gm-maize-study-faces-growing-scrutiny-1.11566 (accessed 13.04.14.).
European Environment Agency, 2002. Late Lessons from Early Warnings: The Precautionary Principle 1896–2000. European Environment Agency, Copenhagen.
EFSA, January 14, 2013. EFSA Promotes Public Access to Data in Transparency Initiative. European Food Safety Agency Press Release.
Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley.
Funtowicz, S.O., Ravetz, J.R., 1993. Science for the post-normal age. Futures 25 (7), 739–755.
Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, London.
Gowers, T., 2012. Elsevier – My Part in Its Downfall. Gowers's Weblog. http://gowers.wordpress.com/2012/01/21/elsevier-my-part-in-its-downfall/ (accessed 13.04.14.).
Goldacre, et al., October 29, 2012. Drug Companies Must Come Clean. Letter about Roche. The Times, London.
Igoe, T., Mota, C., 2011. A strategist's guide to digital fabrication. Strategy+Business 64.
Ioannides, J.P.A., 2005. Why most published research findings are false. PLoS Medicine 2 (8), e124. http://dx.doi.org/10.1371/journal.pmed.0020124.
Jasanoff, S., 1990. The Fifth Branch: Science Advisors as Policymakers. Harvard University Press, Cambridge.
Lundh, A., Sismondo, S., Lexchin, J., Busuioc, O.A., Bero, L., December 2012. Industry sponsorship and research outcomes. Cochrane Database of Systematic Reviews 12. http://dx.doi.org/10.1002/14651858.MR000033.pub2. Article No.: MR000033.
Miedema, F., 2012. Science 3.0/Real Science, Real Knowledge. Amsterdam University Press, Amsterdam.
Nathan, C.F., 9 December 2012. Let's Gang up on Killer Bugs. New York Times.
Oxford Geoengineering Programme, 2012. The Oxford Principles. Oxford Geoengineering Programme, Oxford, UK.
Price, D.J., 1963. Little Science, Big Science. Columbia University Press, New York.
Ravetz, J.R., 1971. Scientific Knowledge and its Social Problems. Oxford University Press, Oxford.
Rose, S., Rose, H., 1969. Science and Society. Allen Lane, London.
The Royal Society Policy Centre, 2012. Science as an Open Enterprise. The Royal Society, London.
Shapin, S., 2008. The Scientific Life: A Moral History of a Late Modern Vocation. University of Chicago Press, Chicago.
Sharoykina, E., 2012. Rat Reality Show: Russian Scientists to Broadcast GMO Experiment. http://rt.com/news/gmo-experiment-online-rats-240/ (accessed 20.12.12.).
Shirky, C., 2010. Cognitive Surplus: Creativity and Generosity in a Connected Age. Penguin Group, New York and London.
Tattersall, P., 2010. What is community based auditing and how does it work? Futures 42, 466–474.
Ziliak, S.T., McCloskey, D., 2008. The Cult of Statistical Significance/How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor.

Science, Public Engagement with

Alan Irwin, Copenhagen Business School, Frederiksberg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Public engagement with science' evokes a series of long-standing issues concerning the relationship between members of the public (or citizens) and matters of technical expertise. However, each of the terms 'public,' 'engagement,' and 'science' is open to question, and to empirical investigation regarding their definition in institutional practice. Science and technology studies scholars have been especially active in challenging prevailing policy assumptions in this area and in considering how science–public relations might be reinterpreted and reconstructed. This article presents some of the key social scientific contributions to this developing area of discussion and practice.

Introduction

Considered in general terms, it might seem that there is little new in the relationship between expert knowledge and wider publics. As Dorothy Nelkin expressed the central political dilemma: "The complexity of public decisions seems to require highly specialized and esoteric knowledge, and those who control this knowledge have considerable power. Yet democratic ideology suggests that people must be able to influence policy decisions that affect their lives" (Nelkin, 1975, p. 37). Writing in the mid-1970s, Nelkin observed that controversy and public dispute were often associated with 'technologies of speed and power' (airports, power-generating facilities, highways, dams), creating significant challenges with regard to the public role of science. By the beginning of the new century, the issues of concern had shifted away (at least primarily) from 'speed and power' and toward the biological sciences (stem cell research, nanotechnology, synthetic biology, questions of food and food safety) (Bauer and Howard, 2012). However, many of the themes raised by Nelkin remain recognizable today: the ambivalent role of science in controversial policy arenas, the variable impact of expertise on political decision making, the at-times problematic relationship between expert evaluation and citizen action.

Although partly anticipated by Nelkin (1977), since that time – and especially since the 1990s – a public rhetoric of 'engagement' around science and technology has developed in many countries, a rhetoric that has been specifically embodied in a series of 'engagement' initiatives (including national debates, consultation exercises, consensus conferences, and the involvement of 'lay' members on scientific advisory committees). To offer one much-discussed example of this 'politics of talk' (Irwin, 2006), the UK in 2003 witnessed a large-scale public consultation over the commercial growing of genetically modified (GM) crops (Horlick-Jones et al., 2007). Other notable cases can be identified, both within European nations (for example, Sweden, Denmark, The Netherlands) (Hagendijk and Irwin, 2006) and more globally (including initiatives in the USA, Canada, New Zealand, and Australia) (see for example, Guston, 1999). In parallel – and at times in close relationship – with such initiatives, social scientists (including science and technology studies (STS) scholars) have developed empirical analyses of these developments, generating also a series of analytical constructs and case-based commentaries. Indeed, one characteristic of this field of practice and scholarship has been that terms entirely or partially developed by social scientists (for example, the deficit model, upstream engagement, citizen science) have influenced (and in turn been influenced by) the language of practice.

Seen from the perspective of STS, the very phrase 'public engagement with science' (PES) raises a series of demanding questions and challenges. How is the 'public' to be defined and what should be its relationship to matters of scientific and technological change? What is meant by 'engagement' in this context? What form should this take, and does the phrase imply that wider society can ever be 'disengaged' from scientific and technological development? It is also a central commitment of STS scholarship that 'science' (like 'the public') should be seen as plural, not singular, and as overlapping with other domains of social, technical, and material practice (as expressed in terms such as 'techno-science' or 'sociotechnical networks'). In line with this problematization of 'public,' 'engagement,' and 'science,' many academic studies have been sharply critical of institutional activities in this area, viewing specific initiatives very often as centralized and highly constrained attempts to 'capture' public opinions without any substantial commitment to the broader principle (or potential) of dialogue and engagement. Nevertheless, the issue of PES raises important questions about the relationship between science, governance, and political values, and especially about that between the larger publics and relevant scientific innovations.

This article does not set out to summarize practical experience within PES, although, given the nature of this field, some sense of the topic will necessarily be provided. Instead, it considers specific social scientific contributions to this area of investigation and debate, and covers some of the key concepts and areas of discussion developed especially by STS scholars.

The 'Deficit Model'

It is important to note that current social scientific discussions of PES have to a large degree evolved from an earlier (but still extant) series of debates and practical activities concerned with the 'public understanding of science' – a term which continues

to be employed among, for example, British science policy-makers. Specifically, a Royal Society report on that topic (1985) made an influential case for the 'better public understanding of science' but also became a particular focus for social scientific criticism. Thus, one of the major contributions from the 1990s onward by STS scholars has been to challenge the assumptions about technical knowledge made by key societal institutions (for example, government agencies and industry) when 'engaging' with larger publics. One major social scientific input has been to identify 'deficit' assumptions within much institutional policy and practice in this area. At the heart of the 'deficit model' (or 'theory') of science–public relations is the notion that 'the facts' relating to a controversial area of science and technology can be separated from political, social, and ethical matters.

Put very broadly, the institutional tendency has been to view public engagement (or 'understanding') as essentially a matter of getting these 'facts' across to public groups, which are viewed as uninformed, ignorant, or emotional in the face of new scientific developments. In the context of the 1990s, the critique of, for example, the nuclear industry was that it presented opponents of civil nuclear power as operating within a cognitive 'deficit' coupled with 'fear' or 'dread.' In line with this assumption, external criticism of the industry was defined primarily as a matter of misunderstanding the scientific case for nuclear power. The challenge for the industry was therefore to educate and inform these 'deficit' publics. While such an approach may have appealed to the developers and promoters of new science and technology, the consequence was (and continues to be) that such institutions failed to consider the possibly legitimate issues raised by critical social groups (in this case antinuclear organizations) as well as their own role in provoking public concerns. The deficit model functions therefore both to problematize critical publics and to insulate institutions from societal challenges, questions, and concerns.

While critiques of the deficit model were developed in the 1990s (see for example, Irwin and Wynne, 1996) – and have subsequently become part of the vocabulary of many engagement practitioners (including in government) – this remains a lingering issue within engagement activities. Certainly, the accusation that specific initiatives are about 'deficit rather than dialogue' has been recurrent since that period. On the one hand, the deficit model serves as a reminder that what is presented as dialogue may offer no more than one-way (or 'top-down') communication, with the wider publics considered as passive recipients rather than active partners in mutual deliberation. On the other, accusations about one or another 'deficit' should not obstruct the broader point that deficiencies of understanding are unavoidable on all sides, and indeed serve as a powerful motivation for dialogue and engagement. From a broad STS perspective, the issue may be less one of 'removing' deficits than of challenging assumptions about what should count as the relevant 'facts' within engagement exercises.

Citizen Science and the 'Knowledgeable Publics'

In this context of problematic science–public–institutional relations, it was perhaps inevitable that social scientists from the 1990s onward would point to domains in which the larger publics could be presented, not simply as recipients of knowledge and information (as in the 'deficit' formulation), but also as active creators of expert knowledge who possess distinctive understanding. In his classic study of nuclear fallout following the Chernobyl disaster, Brian Wynne (1996) observes how a group of sheep farmers demonstrated substantial knowledge of local environmental conditions and animal behavior, especially when compared with the putative 'experts.' The contrast here is between generalized models of (for example) ground contamination – as developed by 'laboratory science' – and the intimate knowledge of specific conditions built up by the farmers in question. Significantly also, the knowledge possessed by the sheep farmers did not fall within one epistemological or disciplinary category, but brought together shared experience, direct observation (often over a very long period), and personal judgments as to the credibility of competing claims to 'know best' (with the government scientists losing credibility as their ignorance of local farming conditions became steadily more evident). Linking with other treatments of 'indigenous' or 'local' knowledge, STS scholars such as Wynne thereby indicate that so-called experts may have as much to learn as the 'lay' interlocutors with whom they seek to 'communicate' within such contexts. However, 'citizen science' (Irwin, 1995) in this sense does not represent an alternative to 'conventional science' – nor does it claim to have a 'better' cognitive understanding of specific conditions. The point instead is that all knowledge claims are 'contextual' in character – including those claims that draw upon standardized protocols and methods of blind peer review.

Beyond STS, 'citizen science' has become a popular term to describe programs in which members of the public who lack formal scientific training engage in scientific projects. Often this takes the form of public participation in large-scale data collection: for example, surveys of garden birds and other species, astronomical observations, and studies of pollution and public health. Representing a form of 'crowd-sourcing,' citizen science in this sense offers a way in which groups of attentive citizens can augment and extend scientific activities, while also gaining a sense of 'engagement' with science. Although this form of 'citizen science' may not have a primarily political or dispute-based focus, it can certainly have policy-related implications, such as when local ornithologists observe changing patterns of bird migration that may in turn have consequences for the general understanding of climate change. Perhaps the main difference between the STS discussion of 'knowledgeable publics' and the more common idea of 'citizen science' is that the latter suggests a greater willingness to work within the epistemological boundaries of science whereas the former explicitly brings social and political judgments into the definition of 'knowledge' (as when the sheep farmers in question explicitly challenge the trustworthiness of externally appointed experts). In practice, however, there may be considerable blurring of these distinctions. Meanwhile, the concept of public knowledgeability and crowd-sourced science suggests perspectives on engagement that take us substantially beyond the restricted framework of the deficit model.

Engagement and the 'Crisis of Trust'

Although the United Kingdom was not an early starter in terms of public discussion, participation, and consultation around science and technology (with countries such as the USA, The Netherlands, and Denmark running substantially ahead during the 1970s), experience in that country through the 1990s, and the institutional lessons drawn from this experience, have been broadly influential both in terms of STS scholarship and international policy-making. Certainly, the UK in the 1990s witnessed a number of challenges to its traditionally closed and expert committee-based regulatory style. The case of mad cow disease (bovine spongiform encephalopathy (BSE)) was to prove especially important in this regard. The 2000 Phillips Report into BSE concluded not only that a government department had provided inappropriate technical advice about the link between contaminated beef and human health, but also that relations with 'the public' had been badly managed. As the Report noted, at the core of government activity in this area had been a 'consuming fear of provoking an irrational public scare' (Phillips et al., 2000, p. 264). The same Report very directly pointed to a feeling of betrayal among members of the public: 'Confidence in public pronouncements about risk was a further casualty of BSE' (Phillips et al., 2000, p. xviii). In making such statements about a generalized loss of public confidence in scientific advice, the Phillips Report represented one important statement among a chorus of other calls at that time for greater openness in science–public–governmental relations. As an influential House of Lords Select Committee report put it also in 2000: "Policy makers will find it hard to win public support on any issue with a science component, unless the public's attitudes and values are recognized, respected and weighed along with the scientific and other factors" (House of Lords, 2000, p. 6). More widely, the Lords report emphasized a "new mood for dialogue" over science and technology and advocated a change in the culture of policy-making "so that it becomes normal to bring science and the public into dialogue about developments at an early stage" (House of Lords, 2000, p. 13).

From a social scientific perspective, there is much that could be said about this increased official enthusiasm for at least making statements about the need for greater transparency, dialogue, and engagement with the wider publics. A key point, however, is to note the relationship between a sense of crisis (or at least strain) in science–public relations and the call for dialogue and openness. This is very explicit in the official report into BSE, but also in a number of governmental and inter-governmental statements (for example, the European Commission's (2002) Action Plan on Science and Society). Presented in this fashion, PES is not simply a desirable option (perhaps on ethical or democratic grounds), but rather an essential precondition for the future of science-based innovation. In line with this assumption, governmental efforts toward engagement have been especially prominent in relation to developments that are either already controversial (including GM foods) or where public controversy is anticipated (nanotechnology, synthetic biology). In this situation, STS scholars have raised questions both about the real institutional commitment to change (Wynne, 2006) and the likelihood that greater openness will actually serve to 'rebuild' public trust: not least since public 'trust' in official experts in contemporary society will always be both conditional and relational, but also since there is little empirical evidence for a generalized (rather than selective) public mistrust in science (Felt and Wynne, 2007). Certainly, the notion that public trust can be 'fixed' by what are often time-limited and centrally directed initiatives in engagement is open to question on several grounds.

Frameworks of Consultation

As PES has developed into a field of institutional practice, STS scholars and others have paid particular attention to the form and the 'framing' of engagement initiatives (Irwin, 2001). Given the breadth of the issues in question and also their sociotechnical complexity, how could one even start to 'engage' with groups of citizens around sets of questions that are often both unfamiliar and externally selected? This question is further complicated by the potential diversity of 'the publics' (embracing, for example, both activists and those who consider themselves entirely new to the engagement topic) and the fundamental challenge of formulating an engagement focus that is appropriate to a diversity of parties and interests. In terms of engagement initiatives then, a number of practical but also social scientific questions arise concerning such matters as the form and organization of consultation, the balance of information-giving and dialogue, the framing of consultation agenda, and the degree of passivity or activity accorded to citizen groups. In addition, a tension can often be identified between the perceived need of institutional actors (e.g., civil servants) to move forward with policy processes (often within a relatively restricted time-frame) and the requirement to provide sufficient time and space for the treatment of what can turn out to be rather open-ended issues. This can be especially significant when issues that are originally presented in rather narrow terms (e.g., the commercial growing of GM crops in the UK) may take on broader significance within public discussions (Why should North American companies dictate to European consumers?; Has greater consumer choice had a positive or negative effect on national diets?; What is the relationship between GM foods and national self-sufficiency?).

Empirical accounts of specific engagement exercises have often reached rather negative conclusions concerning their restricted nature. These criticisms focus on the assumption in such exercises that 'engaged publics' are just 'out there' waiting for their views to be harvested, rather than being specifically shaped and constructed by the exercises themselves. On the one hand, and despite the expressed enthusiasm for public dialogue, such exercises have generally been atypical of institutional practice and limited to a small number of 'high-profile' topics. On the other, those exercises have been criticized for their restricted budget and timescales, the limitations of their coverage, a general lack of clarity about their aims and purposes, and the unproven link between specific engagement initiatives and actual policy change. Furthermore, many exercises in public engagement have taken form as relatively standardized investigations conducted by social research organizations, complete with demographically selected

participants and a mix of qualitative and quantitative evidence (i.e., smaller-scale focus groups together with larger-scale questionnaires). Coupled with a general sense that policy makers may be 'hitting the notes but missing the music' (Wynne, 2006), experience since 2000 suggests that the rhetoric of openness, dialogue, and engagement may be running at some distance from innovation and knowledge development processes (Felt and Wynne, 2007). While engagement therefore at least offers the possibility of 'opening up' policy and decision processes to new perspectives, knowledge, and opportunities, STS scholars and others have often concluded that the reality has been one of 'closing down,' as established modes of policy-making marginalize public concerns about the form and direction of technical change (Stirling, 2008). For this reason also, there has been some support for the idea of 'upstream engagement,' i.e., moves to enable public engagement earlier in the innovation process when policy choices and technological alternatives have not yet been 'closed down' (Wilsdon and Willis, 2004).

Science, Democracy, and Civic Epistemologies

One important contextual element to take into account when considering the relationship between science, technology, and democracy relates to the political and institutional cultures within which specific decisions are identified, defined, and enacted. Jasanoff has proposed the concept of 'civic epistemologies' to characterize the manner in which different societies employ different modes of reasoning when making decisions about science and technology. In her study of the development of biotechnology in the United States, Britain, and Germany, Jasanoff (2005) presents the management of the life sciences not as a 'special case' but as a field that is highly interwoven with larger processes of political constitution and nation-building. In contrast to universalistic discourses on science and democracy, this approach is highly attentive to the heterogeneity of (in this case) Western political cultures, and to the varying way in which these different political cultures interpret the credibility of public knowledge claims. Thus, taking one very notable example, the contemporary German treatment of stem cell research is inevitably influenced by the powerful national reaction to the eugenics policies of the Third Reich. The more general point is that particular cases and controversies around science and technology can only be fully interpreted in the context of longer-established, culturally situated practices of interpretation – and indeed that significant rifts have appeared in the ways in which leading Western nations address the governance of genetics and biotechnology. Going further, Jasanoff suggests that the politics of science and technology are central to the development of democratic theory, so that decisions over technically sophisticated areas such as biotechnology both reflect prevailing assumptions about democratic governance and also provoke a deeper reconsideration of what 'democracy' might mean within modern knowledge societies.

The argument that relations between science and democracy play out in specific political and cultural settings serves as an important reminder that 'public engagement' methods and discussions cannot necessarily or unproblematically be exported across local and national borders – and indeed that what counts as 'engagement' can appear very different in different environments. Thus, the international enthusiasm for 'Danish-style' consensus conferences typically disregards the cultural interconnections, social understandings, and national aspirations that led to the development of consensus conferences in Denmark – and which have both profoundly shaped their operation in that country and constituted a contextually specific definition of 'consensus' (Horst and Irwin, 2010). More broadly, comparative approaches to the relationship between science, politics, and democracy take us directly into the discussion with which this article began. While, at a general and philosophical level, there may indeed be 'little new' in the relationship between expert knowledge and the wider publics, current discussions over public engagement can also be viewed as symptomatic of a wider unease over the form and direction of technical change. To offer one obvious example, how are established forms of democratic governance to respond when so many key discussions over the direction of science and technology are taken within private corporations (Beck, 1992)? If, as STS scholars generally argue, science and politics cannot be kept apart from each other, then what are the consequences for political institutions which were not designed to deal with the potentially pervasive but also highly uncertain challenges offered by developments in fields such as medical and agricultural biotechnology (Callon et al., 2011)?

Science Communication as Engagement

While the STS literature has often defined PES in terms closely related to institutional policy, the politics of expertise and citizen action, an expanding cross-disciplinary literature has examined the public communication involved with science and technology in a fashion that both moves beyond 'deficit' formulations and explores significant science–public interactions occurring outside the structures of government and industry. Seen from a 'communication' perspective, popular science books, science journalism and the mass media, science museums, cinematic science, advertising and public relations, health campaigns, and the internet all represent significant domains in which the publics engage with science (Bucchi and Trench, 2008). Indeed, when compared with the scale and reach of such activities, even large-scale exercises such as the UK's 'GM Nation?' consultation appear extremely limited. It also follows that, for example, environmentalist groups and nongovernmental organizations are obliged to act as science communicators since "empirical claims about the state of the natural environment are core to their message" (Yearley, 2008, pp. 168–169). Equally, the design of novel experiments in communication (such as science installations) can offer a means of 'connecting' with citizens within the contexts of everyday life. Thus, Horst (2011) discusses one 'debate installation' on stem cell research located within a suburban shopping center. As with public consultation exercises, science communication can take many forms: from straightforward 'popular science' based on 'one-way, one-time' assumptions through to two-way and iterative models emphasizing dialogue and contextual understanding.

In line with more institutionally oriented treatments of PES, there have also been calls among researchers in science communication for 'knowledge coproduction' and science–public collaboration in shaping research agendas. Thus, Trench offers a triad of science communication models: 'deficit,' 'dialogue,' and 'participation.' In the third of these, "(c)ommunication about science takes place between diverse groups on the basis that all can contribute, and that all have a stake in the outcome of the deliberations and discussions" (Trench, 2008, p. 132).

One underlying question linking PES research conducted within a 'communication' framework with research focusing on planned consultation exercises concerns the extent to which 'science' is seen to be central to public perspectives and concerns. Thus, and as already noted with regard to the GM Nation? exercise, public views and assessments of possible areas of innovation may take into consideration a plurality of issues regarding choice, need, health, and ethics within which 'science' can at best appear marginal. Horst and Michael (2011) have noted the manner in which an installation with a specifically scientific focus was imaginatively reinterpreted and reconstructed by members of the public (as when teenage girls expressed their concern about the closure of the world's shopping centers or employed the installation's cameras for the amusement of their friends). We thus encounter an apparent paradox in which the 'science' in PES may be less central from certain social perspectives than many engagement protagonists would assume. Indeed, broader values and reflections about the form and direction of technical change – and about societal form and direction more generally – may override the specificities of particular issues (whether stem cell research, GM foods, or synthetic biology). In that way, the areas of science and technology that often serve as the stimulus to engagement activities (and which function as their very core) can turn out to be rather peripheral from a public viewpoint (which may explain some of the disappointment which often follows such activities). Seen from a broad social scientific perspective, this observation may appear unsurprising. However, it is an important point when seeking either to analyze or capture public assessments of particular scientific fields. It also alerts those studying PES to the possibility that there are alternative frameworks of meaning within everyday life, not all of which prioritize 'science' as a cultural and 'sense-making' category.

Engaging Beyond Engagement

In line with this discussion of forms of engagement that find places outside conventional political institutions and the established discourse of 'democracy and citizen action,' a number of STS scholars have sought to explore alternative ways in which members of society 'engage' with science and society, including the relationship between consumption and citizenship, and the manner in which public participation can be 'materialized' through objects and devices. Thus, Michael (1998) has sought to open up what he presents as a complex reciprocal relation between science and consumption, noting for example that the engagement between 'science' and 'the public' can be a matter of pleasure and an expression of identity (as in the enthusiastic purchasing of new electronic devices). The point is not to present the figure of 'the consumer' as an alternative to that of 'the citizen,' but rather to consider the implications of the rise of consumer culture and the aestheticization (or commercialization) of everyday life on the construction of the layperson. At the same time, Michael's account serves as a warning that PES studies should not present science–public relations in unidimensional terms: as if for example the sheep farmers in Wynne's study shared one coherent culture, one sense of identity, and one form of engagement with science and technology. The point instead is that individuals have multiple engagements with science and technology, not all of which can be captured in the established language of 'citizenship.'

Work by Marres and Lezaun (2011) has similarly attempted to pluralize the treatment of engagement by addressing the material dimensions of citizenship. Presented in this way, everyday activities – heating, cooking, washing – represent significant performances of engagement as, for example, choices are made concerning low-energy light bulbs, the level and location of thermostats, and the composting of household waste. More generally, material practices represent a significant domain of interaction with science and technology (Marres, 2012). Their study also places importance on people's actions, activities, and lived experiences around science and technology rather than viewing engagement primarily as a matter of knowledge and knowing. This is therefore an approach to engagement as if 'things' matter. It is also a challenge to develop an explicitly STS-centered reconsideration of the language of 'engagement' rather than simply borrowing from the established vocabulary of political theory.

The discussion of engagement as it occurs in a plurality of public and private contexts – and in both material and discursive forms – suggests above all that 'PES' as a field of scholarly investigation is far from complete. Certainly, there is considerable further potential for both exploring how publics are called into being within institutional approaches to engagement (see, for example, Felt and Fochler, 2010) and the diverse ways in which members of the public both engage and are engaged within sociotechnical systems. While the relationship between expert knowledge and the wider publics might not itself be new, the mutual entanglements, political implications, and sociotechnical manifestations embedded within that relationship remain a crucial focus for social science.

See also: Agricultural Sciences and Technology; Biocitizenship; Environmental Movements; Environmental Sciences; Politics and Participation; Science Communication; Science and Law; Science and Social Movements; Science and Technology Studies: Experts and Expertise; Science and The State.

Bibliography

Bauer, M.W., Howard, S., 2012. Editorial: public understanding of science – a peer reviewed journal for turbulent times. Public Understanding of Science 21 (3), 258–267.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Bucchi, M., Trench, B. (Eds.), 2008. Handbook of Public Communication of Science and Technology. Routledge, Abingdon, New York.

Callon, M., Lascoumes, P., Barthe, Y., 2011. Acting in an Uncertain World: An Essay on Technical Democracy. MIT Press, Cambridge, MA.

European Commission, 2002. Science and Society Action Plan. Office for Official Publications of the European Communities, Luxembourg.

Felt, U., Fochler, M., 2010. Machineries for making publics: inscribing and de-scribing publics in public engagement. Minerva 48, 219–238.

Felt, U., Wynne, B., 2007. Taking European Knowledge Society Seriously. Report of the Expert Group on Science and Governance to the Science, Economy and Society Directorate, Directorate-General for Research, European Commission. Office for Official Publications of the European Communities, Luxembourg.

Guston, D.H., 1999. Evaluating the first US consensus conference: the impact of the 'citizen's panel on telecommunications and the future of democracy'. Science, Technology, & Human Values 24 (4), 451–482.

Hagendijk, R., Irwin, A., 2006. Public deliberation and governance: engaging with science and technology in contemporary Europe. Minerva 44, 167–184.

Horlick-Jones, T., Walls, J., Rowe, G., Pidgeon, N., Poortinga, W., Murdock, G., O'Riordan, T., 2007. The GM Debate: Risk, Politics and Public Engagement. Routledge, London.

Horst, M., 2011. Taking our own medicine: on an experiment in science communication. Science and Engineering Ethics 17 (4), 801–815.

Horst, M., Irwin, A., 2010. Nations at ease with radical knowledge: on consensus, consensusing and false consensusness. Social Studies of Science 40 (1), 105–126.

Horst, M., Michael, M., 2011. On the shoulders of idiots: re-thinking science communication as 'event'. Science as Culture 20 (3), 283–306.

House of Lords Select Committee on Science and Technology, February 2000. Third Report: Science and Society. The Stationery Office, London.

Irwin, A., 1995. Citizen Science: A Study of People, Expertise and Sustainable Development. Routledge, London.

Irwin, A., 2001. Constructing the scientific citizen: science and democracy in the biosciences. Public Understanding of Science 10 (1), 1–18.

Irwin, A., 2006. The politics of talk: coming to terms with the 'new' scientific governance. Social Studies of Science 36 (2), 299–320.

Irwin, A., Wynne, B. (Eds.), 1996. Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge.

Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ, and Woodstock, Oxfordshire.

Marres, N., 2012. Material Participation: Technology, the Environment and Everyday Publics. Palgrave Macmillan, Basingstoke.

Marres, N., Lezaun, J., 2011. Materials and devices of the public: an introduction. Economy and Society 40 (4), 489–509.

Michael, M., 1998. Between citizen and consumer: multiplying the meanings of 'public understanding of science'. Public Understanding of Science 7, 313–327.

Nelkin, D., 1975. The political impact of technical expertise. Social Studies of Science 5 (1), 35–54.

Nelkin, D., 1977. Technological Decisions and Democracy: European Experiments in Public Participation. Sage, Beverly Hills and London.

Phillips, L., Bridgeman, J., Ferguson-Smith, M., October 2000. The BSE Inquiry: The Phillips Report, vol. 1. The Stationery Office, London.

Royal Society, 1985. The Public Understanding of Science. Royal Society, London.

Stirling, A., 2008. 'Opening up' and 'closing down': power, participation, and pluralism in the social appraisal of technology. Science, Technology, & Human Values 33, 262–294.

Trench, B., 2008. Towards an analytical framework of science communication models. In: Cheng, D., Claessens, M., Gascoigne, T., Metcalfe, J., Schiele, B., Shi, S. (Eds.), Communicating Science in Social Contexts: New Models, New Practices. Springer, Dordrecht, pp. 119–135.

Wilsdon, J., Willis, R., 2004. See-through Science: Why Public Engagement Needs to Move Upstream. Demos, London.

Wynne, B., 1996. Misunderstood misunderstandings: social identities and public uptake of science. In: Irwin, A., Wynne, B. (Eds.), Misunderstanding Science? The Public Reconstruction of Science and Technology. Cambridge University Press, Cambridge, pp. 19–46.

Wynne, B., 2006. Public engagement as a means of restoring trust: hitting the notes but missing the music? Community Genetics 9, 211–220.

Yearley, S., 2008. Environmental action groups and other NGOs as communicators of science. In: Bucchi, M., Trench, B. (Eds.), Handbook of Public Communication of Science and Technology. Routledge, Abingdon and New York, pp. 159–171.

Science, Sociology of

Thomas F Gieryn and Kathleen C Oberlin, Indiana University Bloomington, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T. Gieryn, volume 20, pp. 13691–13698, © 2001, Elsevier Ltd.

Abstract

Science has attracted the attention of sociologists both as a powerful source of institutional authority and as a set of practices for making and using knowledge. Theoretical and methodological orientations for the study of science have changed historically alongside shifts in the home discipline of sociology. Studies anchored in mid-C20 structural–functional thought gave way to newer approaches grounded in constructivism and ethnomethodology. These, in turn, were followed by work inspired by neoinstitutionalist and social structural frameworks. The key throughout is a sense that science is a social and cultural phenomenon, the upshot of historically and situationally contingent practices and not the expression of a universal epistemology or methodology.

The effort to make science the object of sociological analysis directs attention to the production and consumption of scientific knowledge in diverse cultural contexts, institutional structures, local organizations, and immediate settings. The sociology of science divides into three broad lines of inquiry, each distinguished by a particular mix of theories and methods. The earliest systematic studies (mostly from the 1950s to the early 1970s) focus on the structural contexts of scientists' behavior, addressing questions such as what rules govern the pursuit of scientific knowledge, how are scientists judged and rewarded, and how is scientific research broken up into dense networks of specialists? In the 1980s, sociologists shift their attention to the practices through which scientific knowledge is constructed – at the laboratory bench or in the rhetoric of professional papers. Starting in the 1990s, science is put in more encompassing societal contexts, as sociologists examine scientists along with other stakeholders as purveyors of cognitive authority, and explore their linkages to power, politics, and the economy.

Precursors

It is remarkable how much the literature in sociology of science is bunched into the last third of the twentieth century. Perhaps only after the deployment of nuclear weapons, or only after genetic engineering raised eugenic nightmares, could sociologists begin to think about science as a social problem in itself, rather than as a source of consistent solutions; or perhaps earlier generations of sociologists were guided by epistemological assumptions that rendered true scientific knowledge immune from social causes – thus putting it outside the orbit of sociological explanation.

Classical Anticipations

'Science' is nowhere indexed in Max Weber's encyclopedic *Economy and Society*, a measure of his unwillingness or inability to see it as a consequential factor in human behavior or social change. Weber's interest in science was largely methodological and political. Could the causal models employed so effectively in the natural sciences be used as well to study social action? Does the objectivity and neutrality of the social scientist preclude involvement in political activity?

Emile Durkheim also sought to institutionalize sociology by making its methods appear scientifically precise, but at the same time, he considered scientific knowledge as an object of sociological study. Durkheim suggested that basic categories of thought and logic (such as time and space) are social in origin, in that they correspond to fundamental social categories (such as the division of a society along lines of family or gender). However, as human societies grew in size and as their institutions became functionally differentiated, a distinctively scientific pursuit of knowledge was gradually insulated from such social causes. The observable facts of modern science, Durkheim concluded, were in accord with the reality of the physical world – a position that forestalled examination of how observable facts are also shaped by the culture and communities in which they arise.

Karl Marx's materialism would seem to commit him to the idea that all beliefs arise amid historically specific conditions of production, as they are shaped by the goals and interests of a ruling class. The rise of science in seventeenth-century Europe is intimately bound with the rise of industrial capitalism and, for Marx, can be explained in terms of the utilities of science-based technologies for improving productivity and enlarging surplus value. But, although the rate of scientific growth may be explained by its congruence with the interests of the bourgeoisie, Marx seems to suggest that the content of scientific claims inside professionalized research networks is nonideological – that is, science provides an objective account of natural reality.

Science in the Sociology of Knowledge

Even more surprising is the failure of systematic sociological studies of science to emerge from a blossoming sociology of knowledge in the 1920 and 1930s. Neither Max Scheler nor Karl Mannheim, authors of foundational treatises on the social determinants of knowledge, inspired sustained inquiry into the social determinants of science – probably because both distinguished scientific knowledge from other kinds in a way

that truncated what sociology could say about it. Scheler isolated the content of scientific knowledge – and the criteria for ascertaining validity – by describing such criteria as absolute and timeless, and not shaped by social interests. The effects of social structure (specifically, the power of ruling elites) is limited to selections of problems and beliefs from that self-contained and essential realm of ideas.

Mannheim sustained the neo-Kantian distinction between formal knowledge of the exact sciences and sociohistorical knowledge of culture. Phenomena of the natural world are invariant, Mannheim suggests, and so therefore are criteria for deciding truth (i.e., impartial observations based on accurate measurements). In contrast, cultural phenomena become meaningful only as they are constructed through interest-laden judgments of significance, which are neither impartial nor invariant, and thus they are amenable to sociological explanation.

Robert K. Merton's 1938 classic *Science, Technology and Society in Seventeenth-century England* (see Merton, 1973) tackles a fundamental problem: why did modern science emerge with a flourish in seventeenth-century England? His answer has become known as the 'Merton Thesis': that a Puritan ethos provided both the motivating force and legitimating authority for the pursuit of scientific inquiry. Certain religious beliefs – e.g., God is glorified by an appreciation of his handiwork in Nature, or Blessed Reason separates human from beast – created a fertile cultural context for the rise of science. Merton also explains shifts in the foci of research attention among the early modern 'natural philosophers' by connecting empirical inquiry to the search for technological solutions to practical problems in mining, navigation, and ballistics.

Social Organization of the Scientific Community

When concerted sociological studies of science began in the late 1950 and 1960s, research centered on the institutions or social structures of science – with relatively less attention given to the routine practices involved in making knowledge or to the wider settings in which science was conducted. This work was largely inspired by theories of structural–functional analysis, which ask how the community of scientists is organized in order to satisfy modern society's need for certified, reliable knowledge. One distinctive feature of this first phase is a reliance on quantitative methods of analysis. With statistical data drawn from surveys of scientists and from the *Science Citation Index* (and other bibliometric sources), sociologists developed causal models to explain individual variations in research productivity and used topographical techniques such as multidimensional scaling to map the dense networks of scientists working at a research front (Barber and Hirsch, 1962; Hagstrom, 1965; Crane, 1972).

Norms of Science

The shift from analyzing science in society to analyzing its internal social organization was effected in Merton's 1942 paper on the normative structure of science (reprinted as a revised and retitled chapter in Merton, 1973). Writing in the shadow of Nazism, Merton argues that the success of scientists in extending certified knowledge depends, at once, on a salutary political context (namely, democracy, which allows science a measure of autonomy from political intrusions and whose values are said to be congruent with those of science – quite unlike fascism) and on an internal institutionalized ethos of values held to be binding upon the behavior of scientists. This ethos comprised the famous norms of science: scientists should evaluate claims impersonally (universalism), share all findings (communism), never sacrifice truth for personal gain (disinterestedness), and always question authority (organized skepticism). Behavior consonant with these moral expectations is functional for the growth of reliable knowledge, and for that reason, young scientists learn through precept how they are expected to behave, conformity is rewarded, and transgressions are met with outrage.

Subsequent work ignored Merton's conjectures about science and democracy, as sociologists instead pursued implications of the four social norms. Studies of behavioral departures from these norms – ethnocentrism, secrecy, fraud, plagiarism, dogmatism – precipitated debates over whether such deviance is best explained by idiosyncratic characteristics of a few bad apples or changing structural circumstances (such as commercialization of research) that might trigger increases in such behavior. Sociologists continue to debate the possibility that Merton's norms are better explained as useful ideological justifications of scientists' autonomy and cognitive authority. Other research suggests that the norms guiding scientific conduct vary historically, vary among disciplines, vary among organizational contexts (university research vs military or corporate research), and vary even in their situational interpretation, negotiation, and deployment – raising questions about whether the norms identified by Merton are functionally necessary for enlarging scientific knowledge.

Stratification and Scientific Careers

The norm of universalism in particular has elicited much empirical research, perhaps because it raises questions of generic sociological interest: how is scientific performance judged, and how are inequalities in the allocation of rewards and resources best described and explained? With effective quantitative measures of individual productivity (number of publications or citations to one's work), resources (grant dollars), and rewards (prizes, like the Nobel), sociologists have examined with considerable precision the determinants of individual career success or failure. Competition among scientists is intense, and the extent of inequality high: the distribution of resources and rewards in science is highly skewed. A small proportion of scientists publish most research papers (and those papers collect most citations), compete successfully for research grants and prestigious teaching posts, achieve international visibility and recognition, and win cherished prizes.

Debate centers on whether these observed inequalities in the reward system of science are compatible with the norm of universalism – which demands that contributions to knowledge be judged on their scientific merit, with resources and opportunities meted out in accordance with those judgments. The apparent elitism of science may result from an 'accumulation of advantage': work by relatively more eminent or

well-positioned scientists is simply noticed more and thus tends to receive disproportional credit – which (over time) enlarges the gap between the few very successful scientists and everybody else. Such a process may still be universalistic because it is functional for the institutional goal of improving science: giving greater attention to research of those with accomplished track records may be an efficient triage of the overwhelming number of new candidate theories or findings. Others suggest that particularism contributes to the stratification of scientists – old boy networks that protect turf and career reputations by rewarding sycophants. The underrepresentation of women in the higher echelons of science has called attention to sometimes subtle sexism that occurs early in the scientific career (restricted access to well-connected mentors or essential research equipment, or opportunities to collaborate and assignment to trivial problems or mind-numbing tasks) (Long and Fox, 1995; also see Lincoln et al., 2012 for recent work on the 'Matilda Effect').

Institutionalization of the Scientific Role

A separate line of sociological inquiry (exemplified in work by Joseph Ben-David (1991) and Edward Shils) seeks an explanation for how science first became a remunerable occupation – and later, a profession. How did the role of the scientist emerge from historically antecedent career patterns in which amateurs explored nature part-time and generally at their own expense? The arrival of the 'scientist' as an occupational self-identification with distinctive obligations and prerogatives is inseparable from the institutionalization of the modern university (itself dependent upon government patronage). Universities provided the organizational form in which the practice of science could become a full-time career – by fusing research with teaching, by allowing (ironically) the historic prestige of universities as centers of theology and scholasticism to valorize the new science, and by providing a bureaucratic means of paying wages and advancing careers.

The scientific role has also been institutionalized in corporate and government laboratories. The difficulties of transporting a 'pure science' ideal of university-based research into these very different organizational settings have been the object of considerable sociological attention. Scientists in industry or government face a variety of competing demands: their research is often directed to projects linked to potential profits or policy issues rather than steered by the agenda of their discipline or specialty; the need to maintain trade secrets or national security hampers the ability of scientists in these settings to publicize their work and receive recognition for it. And, as Jerome Ravetz (1971) suggests, the intrusion of 'bureaucratic rationality' into corporate and state science compromises the craft character of scientific work: largely implicit understandings and skills shared by the community of scientists and vital for the sustained accumulation of scientific wisdom have little place in accountabilities driven by the bottom line or policy relevance.

Disciplines and Specialties

Sociologists use a variety of empirical indicators to measure the social and cognitive connections among scientists: self-reports of those with whom a scientist exchanges ideas or preprints, subject classifications of publications in topical indexes or abstract journals, lineages of mentor–student relationships or collaborations, and patterns of who cites whom or is cited with whom ('cocitation'). The networks formed by such linkages show occasional dense clusters of small numbers of scientists whose informal communications are frequent, who typically cite each other's very recent papers, and whose research focuses on some new theory, innovative method, or breakthrough problem. Emergence of these clusters – for example, the birth of radio astronomy in England after Second World War (WWII), as described by David Edge and Michael Mulkay (1976) – is a signal that science has changed, both cognitively and socially: new beliefs and practices are ensconced in new centers for training or research with different leaders and rafts of graduate students. Over time, these specialties evolve in a patterned way: the number of scientists in the network becomes much larger and connections among them more diffuse; the field gets institutionalized with the creation of its own journals and professional associations; and precedent-shattering innovations become less common as scientists work more on filling in details or adding precision to the now-aging research framework. As one specialty matures, another dense cluster of scientists emerges elsewhere, as the research front or cutting edge moves on. More recent sociological research, following the theoretical lead of Pierre Bourdieu (2004), sees specialties and disciplines as occasions for competitive struggle among scientists seeking to advance careers by advancing their preferred understanding of reality.

Sociology of Scientific Knowledge

A sea-change in sociological studies of science began in the 1970s with a growing awareness that studies of the institutional and organizational contexts shaping scientists' behavior could not illuminate sufficiently the processes that make science: experimental tinkering, sifting of evidence, negotiation of claims, replacement of old beliefs about nature with new ones, and achievement of consensus over the truth. All these processes – observation, getting instruments and research materials (e.g., mice) to work, logic, criteria for justifying a finding as worthy of assent, choices among theories, putting arguments into words or pictures, persuading other scientists that one is correct – are uncompromisingly social, cultural, and historical phenomena, and so sociologists set about to explain and interpret the content of scientific knowledge by studying the routine practices of scientific work.

This research is guided by constructivist theories (and, less often, ethnomethodology), which center attention on the practically accomplished character of social life. Rather than allow a priori nature or given social structures to explain behavior or belief, constructivist sociologists examine how actors incessantly make and remake the structural conditions in which they work. Such research relies methodologically on historical case studies of scientific debate, up close ethnographic observations of scientific practices, and interpretative analysis of scientific texts (Mulkay, 1979, 1980; Knorr-Cetina and Mulkay, 1983).

Sociology of Discovery

Diverse studies of scientific discovery illustrate the range of sociological perspectives brought to bear on these consequential events. An early line of inquiry focuses on the social and cognitive contexts that cause the timing and placing of discoveries: why did these scientists achieve a breakthrough then and there? Historical evidence points to a pattern of simultaneous, multiple, and independent discoveries – that is, it is rare for a discovery to be made by a scientist (or a local team) who is the only one in the world doing research on that specific question. Because honor and recognition are greatest for solutions to the perceivably 'hottest' problems in a discipline, the best scientists are encouraged by the reward system of science to tackle similar lines of research. But these same social structures can also forestall discovery or engender resistance to novel claims (Brannigan, 1981). Cognitive commitments to a long-established way of seeing the natural world (reinforced by reputations and resources that depend upon those traditional perspectives) can make it difficult for scientists to see the worthiness of a new paradigm. Resistance to new ideas seems to be greatest among older scientists, especially in cases where the proposed discovery comes from scientists with little visibility or stature within the specialty or discipline that would be transformed.

More recent sociological research considers the very idea of 'discovery' to be a practical accomplishment of scientists. Studies inspired by ethnomethodology offer detailed descriptions of scientific work 'first time through,' taking note of how scientists at the laboratory bench decide whether a particular observation (among the myriad observations) constitutes a discovery (Garfinkel et al., 1981). Other sociologists locate the 'moment' of discovery in downstream interpretative work, as scientists narrate first time through research work with labels such as 'breakthrough.' Such discovery accounts are often sites of dissensus, as scientists dispute the timing or implications of an alleged discovery amid ongoing judgments of its significance for subsequent research initiatives or allocations of resources. These themes – interests, changing beliefs, ordinary scientific work, post hoc accountings, dissent, and persuasion – have become hallmarks of the sociology of scientific knowledge.

Interests and Knowledge Change

In the mid-1970s to the 1980s, sociology of science took root at the Science Studies Unit in Edinburgh, as philosopher David Bloor developed the 'strong programme,' while Barry Barnes, Steven Shapin, Donald MacKenzie, and Andrew Pickering developed its sociological analog – the 'interest model.' The goal is to provide causal explanations for changes in knowledge – say, the shift from one scientific understanding of nature to a different one. Scientists themselves, as well as philosophers of science, might account for such changes in terms of greater evidence, coherence, robustness, promise, parsimony, predictive power, or utility of the new framework. Sociologists, in turn, account for those judgments in terms of social interests of scientists that are either extended or compromised by a decision to shift to the new perspective. What becomes knowledge is thus contingent upon the criteria used by a particular community of inquirers to judge competing understandings of nature, and also upon the goals and interests that shape their interpretation and deployment of those criteria. Several caveats are noted: interests are not connected to social positions (class, for example, or nationality, discipline, specialty) in a rigidly deterministic way; social interests may change along with changes in knowledge; and choices among candidate knowledge claims are not merely strategic – that is, calculations of material or symbolic gains are bounded by considerable uncertainty and by a shared culture of inquiry that provides standards for logical or evidential adequacy and for the proper use of an apparatus or concept.

Drawing on historical case studies of theoretical disputes in science – nineteenth-century debates over phrenology and statistical theory and twentieth-century debates among high-energy physicists over quarks – two different kinds of interests are causally connected to knowledge change. Political or ideological commitments can shape scientists' judgments about candidate knowledge claims: the development of statistical theories of correlation and regression by Francis Galton, Karl Pearson, and R. A. Fisher depended vitally on the utility of such measures for eugenic objectives. Different social interests arise from the accumulated expertise in working with certain instruments or procedures, which incline scientists to prefer theories or models that allow them to capitalize on those skills (Barnes and Edge, 1982; Barnes et al., 1996).

Laboratory Practices and Scientific Discourse

As sociologists moved ever closer to the actual processes of 'doing science,' their research divided into two lines of inquiry: some went directly to the laboratory bench seeking ethnographic observations of scientists' practices in situ and others examined scientists' discourse in talk and texts – that is, their accounting practices. These studies together point to an inescapable conclusion: there is nothing not social about science. From the step-by-step procedures of an experiment to writing up discovered facts for journal publication, what scientists do is describable and explicable only as social action – meaningful choices contingent on technical, cognitive, cultural, and material circumstances that are immediate, transient, and largely of the scientists' own making.

Laboratory ethnographies by Karin Knorr-Cetina (1999), Michael Lynch (1993), and Bruno Latour and Steve Woolgar (1986) reveal a science whose order is not to be found in transcendent timeless rules of 'scientific method' or 'good laboratory procedures,' but in the circumstantial, pragmatic, revisable, and iterative choices and projects that constitute scientific work. These naturalistic studies emphasize the local character of scientific practice: the idea that knowledge making is a process situated in particular places with only these pieces of equipment or research materials or colleagues immediately available. Never sure about how things will turn out in the end, scientists incessantly revise the tasks at hand as they try to get machines to perform properly, control wild nature, interpret results, placate doubting collaborators, and rationalize failures. Even methodical procedures widely assumed to be responsible for the objective, definitive, and impersonal character of scientific claims – experimental replication, for instance – are

found to be shot through with negotiated, often implicit, and potentially endless judgments about the competence of other experimentalists and the fidelity of their replication attempts to the original (as Harry Collins (1992) has suggested).

Ethnographic studies of how scientists construct knowledge in laboratories compelled sociologists then to figure out how the outcomes of those mundane contextual practices (hard facts, established theories) could paradoxically appear so unconstructed – as if they were given in nature all along, and now just found (not made). Attention turned to the succession of 'inscriptions' through which observations become knowledge – from machine-output to laboratory notebook to draft manuscript to published report. Scientists' sequenced accounts of their fact making rhetorically erase the messy indeterminacy and opportunism that sociologists have observed at the laboratory bench, and substitute a story of logic, method, and inevitability in which nature is externalized as waiting to be discovered. Such studies of scientific discourse have opened up an enduring debate among constructivist sociologists of science: those seeking causal explanations for scientists' beliefs treat interests as definitively describable by the analyst, while others (Gilbert and Mulkay, 1984; Woolgar, 1988) suggest that sociologists must respect the diversity of participants' discursive accounts of their interests, actions, or beliefs – and thus treat actors' talk and text not as mediating the phenomena of study but as constituting them.

Actor Networks and Social Worlds

After sociologists worked to show how science is a thoroughly social thing, Bruno Latour (1987) and Michel Callon (1986) retrieved and reinserted the material: consequently, science is not only about facts, theories, interests, rhetoric, and power but also about nature and machines. Scientists accomplish facts and theories by building 'heterogeneous networks' consisting of experimental devices, research materials, images and descriptive statistics, abstract concepts and theories, the findings of other scientists, as well as persuasive texts – and, importantly, none of these are reducible to any one of them, nor to social interests. Things, machines, humans, and interests are, in the practices of scientists, unendingly interdefined in and through these networks. They take on meanings via their linkages to other 'actants' (a semiotic term for anything that has 'force' or consequence, regardless of substance or form). In reporting their results, scientists buttress claims by connecting them to as many different actants as they can, in hopes of defending the putative fact or theory against the assault of real or potential dissenters. From this perspective, length makes strength, that is, the more allies enrolled and aligned into a network – especially if that network is then stabilized or 'black boxed' – the less likely it is that dissenters will succeed in disentangling the actants and thereby weaken or kill the claim. Importantly, for this sociology of science, the human and the social are decentered, in an ontology that also ascribes agency to objects of nature or experimental apparatuses.

Actor-network theory moved the sociological study of science back outside the laboratory and professional journal – or, rather, reframed the very idea of inside and outside. Scientists and their allies 'change the world' in the course of making secure their claims about nature, and in the same manner. In Latourian vernacular, not only are other scientists, bits of nature, or empirical data enlisted and regimented but also political bodies, protest movements, the media, laws, and hoi polloi. When Louis Pasteur transformed French society by linking together microbes, anthrax, microscopes, laboratories, sick livestock, angry farmers, nervous Parisian milk drinkers, public health officials, lawmakers, and journalists into what becomes a 'momentous discovery,' the boundary between science and the rest of society became impossible to locate. Scientists are able to work autonomously at their benches precisely because so many others outside the laboratory are also 'doing science,' providing the life support (money, epistemic acquiescence, infrastructure) on which science depends.

The boundaries of science also emerge as theoretically interesting in related studies derived from the brand of symbolic interactionism developed by Everett Hughes, Herbert Blumer, Anselm Strauss, and Howard Becker (and extended into research on science by Adele Clarke (1990), Joan Fujimura, and Susan Leigh Star). On this score, science is work – and, instructively, not unlike work of any other kind. Scientists (like plumbers) pursue doable problems, where 'doability' involves the articulation of tasks across various levels of work organization: the experiment (disciplining research subjects), the laboratory (dividing labor among laboratory technicians, graduate students, and postdocs), and 'social worlds' (the wider discipline, funding agencies, or maybe animal rights activists) (Fujimura, 1987). Scientific problems become increasingly doable when 'boundary objects' allow for cooperative intersections of those working on discrete projects in different social worlds. For example, success in building California's Museum of Vertebrate Zoology in the early twentieth century depended upon the standardization of collection and preparation practices (here, the specimens themselves become boundary objects) that enabled biologists to align their work with trappers, farmers, and amateur naturalists in different social worlds (Star and Griesemer, 1989). As in actor-network theory, sociologists working in the 'social worlds' tradition make no assumption about where science leaves off and the rest of society begins – those boundaries get settled only provisionally, and remain open to challenge from those inside and out.

Science In/As Society

Recent work seeks an even better understanding of how science intersects with economic interests, political action, and cultural traditions. Sheila Jasanoff's (2005) development of the idiom of the 'coproduction' of science and society suggests that the construction of scientific knowledge is at the same time the resolution (or perpetuation) of political, economic, and cultural controversies. That is, science changes along with other social institutions and practices, simultaneously and interactively. Attention has centered on an increasingly expanding set of stakeholders, who engage 'science' either as producers or consumers, beneficiaries, or victims. These newer studies build on a foundation created by the sociology of scientific

knowledge, even as they increasingly draw on the institutionalist inclinations of an earlier generation.

Credibility and Expertise

Science is often examined as a cognitive or epistemic authority; scientists are said to have the legitimate power to define facts and assess claims to truth. This authority is not treated as an inevitable result of the character or virtue of those who become scientists, the institutional organization of science (norms, for example), or the 'scientific method.' It is, rather, an accomplished resource pursued strategically by a profession committed not only to extending knowledge but also to preserving and expanding its power, patronage, prestige, and autonomy.

Steven Shapin (1994), among others, has identified credibility as a constitutive problem for the sociology of science. Whose knowledge claims are accepted as believable, trustworthy, true, or reliably useful – and on what grounds? Plainly, contingent judgments of the validity of claims depend upon judgments of the credibility of the claimants. Examination of such judgments in historical and contemporary cases has focused sociological attention on how people use (as they define) qualities such as objectivity, expertise, competence, personal familiarity, propriety, and sincerity to decide which candidate in the universe becomes provisionally 'real.' A long-established line of sociological research examines those public controversies that hinge, in part, on 'technical' issues. Case studies of disputes over environmental and health risks find a profound ambivalence: the desire for public policy to be decided by appropriate legislative or judicial bodies in a way that is both understandable and accountable to the populace runs up against the need for expert skills and knowledge monopolized by scientific, medical, or engineering professionals.

The understanding of science today has benefited from those who bring to its analysis theories and methods from the sociological study of social movements and participatory politics. Steven Epstein (1998) develops the concept of 'credibility struggles' through a study of the science and politics of acquired immunodeficiency syndrome (AIDS). Initially, activists denounced scientific research on AIDS as unresponsive to the immediate needs of those who suffered. Then the activists shifted tactics, and sought a 'seat at the table' by learning enough about clinical trials and drug development to work alongside scientists in the deliberation of public policy. In other cases, scientists themselves become activists and embrace the agendas of various social movements (such as physicists and others who worked to resist the proliferation of nuclear weapons after WWII or the spread of artificial pesticides and herbicides). This line of inquiry has crystallized in a 'new political sociology of science,' which is especially sensitive to the diversity of stakeholders who variously challenge and exploit the cognitive authority of science in the pursuit of political, economic, or personal ambitions (Frickel and Moore, 2006; also see Kloppenburg, 1988).

Situating Science

Finally, sociologists have turned their attention to a different question: *Where* does science happen (Gieryn, 1999). Some explore the collaborative networks and organizational settings that house scientific inquiry – beyond the archetypal pure science laboratory at a university. Careers of scientists at work in the entrepreneurial networks of Silicon Valley vary substantially from those of academic researchers, as patents become as important for success as open publications (Bunker Whittington et al., 2009). Other scholars compare the distinctive trajectories of scientific research and application as these emerge in different national settings. For example, because of distinctive political, economic, and legal contexts that are themselves never fixed, biotechnological knowledge and products move along different developmental and regulatory trajectories in the United States, United Kingdom, and Germany (Jasanoff, 2005). Sociologists working from a neoinstitutionalist perspective examine the role of science in the global universalization of standards and categories, as the legitimating force of scientific authority is adopted locally (but eventually almost everywhere) to justify and rationalize political actions and structures (Drori et al., 2002). Other scholars have countered with an assertion that science itself is an inescapably local production, shaped thoroughly by the cultural, spatial, material, and even architectural particularities of sites where research is conducted. Studies of laboratory design and of the unique qualities of local field sites have initiated a promising geographic sociology of science.

See also: Actor-Network Theory; Bourdieu, Pierre (1930–2002); Cultural Studies of Science; Garfinkel, Harold (1917–2011); Institutions; Interests in Sociological Analysis; Knowledge, Sociology of; Laboratory Studies: Historical Perspectives; Norms in Science; Science and Social Movements; Science and Technology Studies, Ethnomethodology of; Science and Technology Studies: Experts and Expertise; Science and The State; Scientific Controversies; Standpoint Theory, in Science; Strong Program, Sociology of; Truth and Credibility in Science; Weber, Max (1864–1920); Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Barber, B., Hirsch, W. (Eds.), 1962. The Sociology of Science. Free Press, Glencoe, IL.

Barnes, B., Edge, D. (Eds.), 1982. Science in Context. MIT Press, Cambridge, MA.

Barnes, B., Bloor, D., Henry, J., 1996. Scientific Knowledge: A Sociological Analysis. University of Chicago Press, Chicago.

Ben-David, J., 1991. Scientific Growth. University of California Press, Berkeley, CA.

Bourdieu, P., 2004. Science of Science and Reflexivity. Polity Press, Cambridge.

Brannigan, A., 1981. The Social Basis of Scientific Discoveries. Cambridge University Press, Cambridge.

Bunker Whittington, K., Owen-Smith, J., Powell, W.W., 2009. Networks, propinquity, and innovation in knowledge intensive industries. Administrative Science Quarterly 54, 90–122.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Bliuex Bay. In: Law, J. (Ed.), Power, Action and Belief. Routledge & Kegan Paul, London, pp. 196–233.

Clarke, A.E., 1990. A social worlds research adventure. In: Cozzens, S.E., Gieryn, T.F. (Eds.), Theories of Science in Society. Indiana University Press, Bloomington, IN, pp. 15–43.

Collins, H.M., 1992. Changing Order: Replication and Induction in Scientific Practice, second ed. University of Chicago Press, Chicago.

Crane, D., 1972. Invisible Colleges: Diffusion of Knowledge in Scientific Communities. University of Chicago Press, Chicago, IL.
Drori, G., Meyer, J.W., Ramirez, F., Schofer, E., 2002. Science in the Modern World Polity: Institutionalization and Globalization. Stanford University Press, Stanford, CA.
Edge, D.O., Mulkay, M.J., 1976. Astronomy Transformed. Wiley, New York.
Epstein, S., 1998. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley, CA.
Frickel, S., Moore, K., 2006. The New Political Sociology of Science: Institutions, Networks, and Power. University of Wisconsin Press, Madison.
Fujimura, J., 1987. Constructing doable problems in cancer research. Social Studies of Science 17 (2), 257–293.
Garfinkel, H., Lynch, M., Livingston, E., 1981. The work of a discovering science construed with materials from the optically discovered pulsar. Philosophy of the Social Sciences 11, 131–158.
Gieryn, T.F., 1999. Cultural Boundaries of Science: Credibility on the Line. University of Chicago Press, Chicago.
Gilbert, G.N., Mulkay, M., 1984. Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse. Cambridge University Press, Cambridge, UK.
Hagstrom, W.O., 1965. The Scientific Community. Basic Books, New York.
Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.
Kloppenburg Jr., J.R., 1988. First the Seed: The Political Economy of Plant Biotechnology, 1492–2000. Cambridge University Press, Cambridge, UK.
Knorr-Cetina, K., 1999. Epistemic Cultures: How the Science Make Knowledge. Harvard University Press, Cambridge, MA.
Knorr-Cetina, K., Mulkay, M.J. (Eds.), 1983. Science Observed: Perspectives on the Sociology of Science. Sage, Beverly Hills, CA.
Latour, B., 1987. Science in Action: Following Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts, second ed. Princeton University Press, Princeton, NJ.
Lincoln, A.E., Pincus, S., Bandows Koster, J., Lebov, P.S., 2012. The matilda effect in science: awards and prizes in the US, 1990s and 2000s. Social Studies of Science 42 (2), 307–320.
Long, J.S., Fox, M.F., 1995. Scientific careers: universalism and particularism. Annual Review of Sociology 21, 45–71.
Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge University Press, Cambridge.
Merton, R.K., 1973. The Sociology of Science. University of Chicago Press, Chicago.
Mulkay, M.J., 1979. Science and the Sociology of Knowledge. George Allen & Unwin, London.
Mulkay, M.J., 1980. Sociology of science in the west. Current Sociology 28 (3), 1–184.
Ravetz, J.R., 1971. Scientific Knowledge and Its Social Problems. Oxford University Press, Oxford.
Shapin, S., 1994. A Social History of Truth. University of Chicago Press, Chicago.
Star, S.L., Griesemer, J., 1989. Institutional ecology, "translations" and boundary objects: amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. Social Studies of Science 19 (3), 387–420.
Woolgar, S., 1988. Science: The Very Idea. Tavistock, London.

Scientific Concepts: Development in Children

Jamie Jirout and Amy Pace, Temple University, Ambler, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by Z. Chen, R. Siegler, volume 20, pp. 13714–13719, © 2001, Elsevier Ltd.

Abstract

Children's understanding of scientific concepts begins in infancy, but develops considerably with experience over many years. Focusing primarily on from the past 20 years, we discuss the development of scientific concepts in three sections: understanding of physical concepts, understanding of biological concepts, and scientific learning processes – both how children learn science and how they learn the processes involved in scientific reasoning.

Early Understanding of Scientific Concepts

As Chen and Siegler discussed in 2001, the advancement of technology such as eye trackers over more action-oriented measures has provided support for much earlier acquisition of science knowledge. Using looking time paradigms, such as the violation of expectancy (see Chen and Siegler's previous version), researchers have found that infants have at least immature theories of scientific concepts, such as physics (Vosniadou, 2002) and biology (Setoh et al., 2013). These paradigms record infant visual fixations to stimuli presented on a computer or television screen and provide a more sensitive measure than previous methods by eliminating task demands. Recent evidence from looking time experiments provides support that children as young as 4-month old demonstrate basic physical knowledge, such as showing surprise when a solid object appears to move through another solid object, when an event violates the rules of gravity, or when an inanimate object moves without a force being applied (Wang et al., 2005; Baillarageon et al., 2009), and by the age of 3, children have a strong understanding of the differences between animate and inanimate objects (see below for evidence of this in younger infants as well; see Chen and Siegler's previous version). Though these studies suggest some early understanding of science concepts, children's science knowledge is quite immature relative to adults,' and develops significantly from early childhood through adulthood. Similarly, children have basic scientific process skills, though some researchers consider their learning processes to be quite sophisticated, calling infants 'rational constructivists' who can reason about evidence with previous knowledge and beliefs using probabilities (Xu and Kushnir, 2013). We discuss the development of scientific concepts in three sections: understanding of physical concepts, understanding of biological concepts, and scientific learning processes.

Developing Understanding of Physical Concepts

Though understanding physical concepts requires quite abstract thought, there is consistent support that children begin to demonstrate the basic understanding of physical phenomena from a very young age. It is suggested that this knowledge of the physical world evolved as a mechanism of survival (Pinker, 1997; Wellman and Gelman, 1998), for example, understanding the principle of consistency in solid objects. Children demonstrate understanding of the rule that objects continue to exist when not visible (i.e., object permanence), or that partially occluded objects are continuous (Piaget, 1952; Baillargéon et al., 1995; Baillargéon, 2008). Infants well under 1 year seem to make sense of the physical world, for example, demonstrating understanding that large items can be hidden by large containers but not small containers (Hespos and Baillargéon, 2006), and recognizing violations of laws of physics like gravity (Luo et al., 2009). While some researchers suggest that this early appearance of physical knowledge is innate, there is evidence supporting the relationship between children's experience and their understanding of physical relationships in the world (Cicchino et al., 2011).

A key area where infants and very young children tend to show weaker understanding of physical concepts is in mechanistic thinking. Children's understanding of the world is greatly enhanced with the ability to make causal inferences (Frye et al., 1996), and determining mechanisms is critical for developing accurate theories of causal relationships. However, infants and young children tend to represent causal relationships using relatively weak rules, such as covariation (Gopnik et al., 2001). When events occur together, they perceive them as causally connected. If one event tends to precede another in time or space, it is inferred to be the causal agent. Children eventually begin to invent mechanisms to explain these relationships – often incorrect – creating naive or folk theories. It is not until they learn the correct mechanisms that their theories are considered to be more scientific than naive, and the theories become more useful in understanding the physical world (Keil, 2010).

The development of children's ability to understand mechanisms is not clear. While some research suggests that children must be explicitly taught, it has also been suggested that preschool children can infer even abstract causal mechanisms with limited evidence (Schulz et al., 2008). For example, after seeing only four demonstrations of three causal kinds of blocks, children are able to infer the mechanistic power of a novel block with only a single demonstration. The three causal relationships were: block B makes a train noise when it comes in contact with block A, block C makes a siren noise when it comes in contact with block B, but contact between blocks A and C has no effect. When children are shown one instance of a novel block's effect on block B – seeing it causes

block B to make a train noise – they are then able to infer that if the novel block is placed behind a curtain and makes a train sound, that there must be another block, specifically a B block, also behind the curtain. Though young children demonstrate causal reasoning across several similar studies, this ability seems to be limited by their tendency to focus on a single causal mechanism (e.g., when two blocks touch, a sound plays; when a switch is flipped, a light comes on). An example of this limitation is observed in children's understanding of force dynamics.

Though children can reason about causality and make predictions of motion events in their first year of life, the weakness of their understanding is demonstrated in their failure on tasks involving multiple forces. For example, when preschool age children predict what will happen when a ball rolls off of a table, they expect the ball to fall straight down, rather than following a parabolic path. Göksun et al. (2013) investigated children's representations of multiple forces by asking children to make predictions in events with two forces: motion caused from a ball rolling out from a 16″ tall curved tube, and wind caused from a blower. The researchers asked children to estimate where the ball would roll in three conditions, determined by the types of causal forces: (1) a single force acting on the ball (cause), (2) two forces acting on the ball in the same direction (enable), and (3) two forces acting on the ball in opposing directions, either directly opposing each other or at a perpendicular angle (prevent). The researchers found that children were relatively accurate when predicting where the ball would roll if acted on by a single force (cause), but were not able to integrate two forces in their predictions until about 5.5 years of age. When integrating two forces, children were most accurate when the forces were moving in the same direction (enable) and much less accurate when forces were perpendicular than directly opposing each other (prevent).

Children's developing concepts of physical events involves an underlying and understudied aspect of the development of physical concepts: spatial reasoning. The development of children's ability to reason about space is fundamental in their ability to make predictions about causal relationships and about the laws of physics. Yet, little is known about when and how different spatial abilities develop. One major developmental milestone that children must reach in order to understand the physical world more accurately is to overcome egocentrism, or the ability to consider perspectives other than one's own, which typically occurs during the preschool years (see Newcombe, 1989). Object representation, discussed above, is also a part of spatial ability, though it goes beyond only representing objects to being able to rotate and transform them, for example, determining which way to orient puzzle pieces so they fit together, or inferring the shape of a cross section of an object. Mental rotation, the imagined movement of an object in space, is an ability seen as early as infancy (Möhring and Frick, 2013), but with a strong developmental trend through early childhood (Frick et al., 2013), and malleability, or sensitivity to intervention, even into adulthood. Spatial tools, for example, symbolic representation, gesture, and sketching, can also be used to learn about one's representation of the physical world, and understanding of spatial representations is foundational for important scientific skills (DeLoache, 1987). Children's ability to use very simple maps from 3 to 4 years of age suggests relatively accurate but simple mental representations of physical space, and research on the use of symbolic representations demonstrates considerable development from early childhood through adulthood, including the strategies used in the use of representations, for example, a switch from very perceptual to more metric or quantitative methods.

Though infants and young children do not typically hold complex representations of physical concepts and space, research suggests that they are quite capable of encoding information to form sufficient representations of physical events, allowing them to use their representations to recognize, categorize, and predict in the physical domain (Baillargéon et al., 2009). These representations and the ability to use them, however, are very limited, resulting in a 'naive' or 'folk' theory of physical concepts that develops considerably throughout adolescence and beyond.

Developing Understanding of Biological Concepts

Although there is still considerable debate regarding early understanding of biological concepts including life, aging, reproduction, illness, and death, researchers generally agree that young children demonstrate more sophisticated knowledge than originally proposed by Piaget (1929). A review of the contributions since the late 1990s highlights two critical factors. First, young children may have a more nuanced appreciation of biological concepts than previously thought, but important conceptual development can be observed in childhood, adolescence, and even adulthood. Second, much of what children understand about biological concepts is learned from others (Gelman, 2009) and thus, must consider the role of linguistic and sociocultural environment in their formation.

Recent evidence has identified early sensitivity to unique biological characteristics such as gaze alternation, face recognition, and biological motion within 1 year (Bidet-Ildei et al., 2013; Brand et al., 2007; de Haan et al., 2002). For example, infants can distinguish between animate and inanimate objects (Rakison and Poulin-Dubois, 2001; Woodward et al., 2001) and recognize that self-propelled movement is unique to animals (Kosugi et al., 2003). In addition, young infants anticipate that human hands, but not trains or blocks are associated with *causal agency*, or an entity's ability to contact other objects and change their physical location (Saxe et al., 2005). While these early sensitivities may provide a foundation for developing biological understanding, these representations are not necessarily aligned with the concepts of children and adults. In particular, early understanding of living things (e.g., animate entities) may conflate the perception of humanlike features and prototypical knowledge about intentional agents. At 6.5 months, for example, infants considered actions to be goal-directed when performed by a human or a humanoid robot (Kamewari et al., 2005). By 18 months, in contrast, children may be more likely to accurately categorize entities into inclusionary animate and inanimate categories even if the exemplars are perceptually distinct (e.g., people/horses; furniture/vehicles; Rostad et al., 2012).

Despite evidence for conceptual foundations in infancy, early understanding of biological concepts often involves substantial misconceptions that undergo significant change with age and experience (Inagaki and Hatano, 2002). In one study that evaluated parental reports of biological misconceptions in 3- to 6-year-old children, a majority involved the misunderstanding of causal processes or events involved in biological phenomena such as reproduction or illness, and the outcome or inevitability of aging and death (Nguyen and Rosengren, 2004).

One important developmental shift involves children's ability to overcome perceptual biases and appreciate that biological concepts share an essence that transcends surface characteristics. In one study that demonstrates this shift, Keleman et al. (2003) presented 3-, 4-, and 5-year-old children with information about the contrasting behavioral properties of two 'training' animals (e.g., a surgeonfish and a porcupine). Similar to the triad induction task developed by Gelman and Markman (1986), they were then asked which of these properties was true of a test animal (e.g., a porcupine box fish) that had overall similarity to one of the training animals (i.e., the surgeonfish) but was overall dissimilar to the other training animal (i.e., porcupine) with which it shared only a behaviorally relevant functional structure (i.e., spines). Results suggest that 4 and 5 year olds (but not 3 year olds) were more likely to overlook global perceptual similarity and attend to relevant functional cues (despite perceptual dissimilarity) as a basis for generalizing the biologically based property (i.e., its defensive strategy). Furthermore, by 4 years of age, children were increasingly able to reflect on and describe their attention to functional features when explaining why they believe animals do what they do.

Additional changes in children's ability to make inferences about functional properties can be observed from 5 to 10 years of age (Farrar and Boyer-Pennington, 2011). In this study, children learned familiar or unfamiliar internal, external, or functional properties of known and unknown animals. Younger children's beliefs about the essence of biological kinds were more likely to be guided by perceptual appearance rather than categorical information, and gave no selective importance to internal properties. Older children, however, made categorical inferences about internal and functional properties even if they were unfamiliar. As children become more knowledgeable about biology, they are able to more accurately weigh categorical vs perceptual information in the context of the specific task to distinguish which property types are likely to be shared. Although misconceptions become less prevalent with age and experience, traces may persist into adolescence and even adulthood. In a speeded classification task, adult undergraduates showed reduced accuracy and increased response times when determining life status for plants, relative to animals, and for natural and moving nonliving things compared with artifacts and nonmoving things (Goldberg and Tompson-Schill, 2009). Surprisingly, these results were replicated in a group of university biology professors, suggesting that feature biases within biological knowledge appear to reflect developmental foundations that are not completely erased with the acquisition of domain-specific expertise.

Given that certain biological processes are difficult to directly observe, these concepts may develop to a large extent through cultural transmission. Converging evidence suggests that culture and experience matter when reasoning about biological phenomena (Medin et al., 2010). Tarlowski (2006) tested four groups of 4 year olds differing along two dimensions of exposure to biological information: their direct experience with nature (i.e., urban or rural) and the biological expertise of their parents (i.e., experts vs laypeople). Results showed that children's exposure to biology had consequences for their patterns of induction: children of experts based inductions on the category *animal*, while children of laypeople relied on similarity. Furthermore, rural children were more restricted in their responses compared to urban children. In another study, 5- to 7-year-old Menominee and European American children were shown pairs of plants and animals and given the opportunity to describe how or why they might be associated (Unsworth et al., 2012). Menominee children were more likely than their rural European American counterparts to talk about their ecological relations and utility associated with nature. They were also more likely to mimic the species using sound or posture, suggesting a greater psychological closeness to nature. These findings suggest that biological concept formation is not just a product of development, but is highly contingent upon children's experience.

The persistent challenges observed when drawing distinction between living and nonliving natural kinds is complicated by the fact that different languages carve up and categorize these concepts in distinct ways. For example, the term 'alive' introduces challenges to English-speaking children since it can be applied to an inclusive category of all animate objects as well as the more restrictive concept that excludes humans. This is less problematic in other languages such as Indonesian in which the term 'alive' refers only to the more restrictive concept, while the inclusive concept remains unnamed (Anggoro et al., 2008). This study revealed that even at 9 years of age, English-speaking children were more likely to exclude plants from their classifications of entities that were 'alive' than their Indonesian-speaking counterparts.

Children's understanding of biological concepts reflects the acquisition and enrichment of content information as well as continuous reorganization of initial theories as children learn more about the biological world. Exploring the way in which infants progress from their initial knowledge base to more advanced understanding is another important piece of the puzzle.

Developing Scientific Processes Skills

In addition to children's understanding of scientific concepts in physical and biological sciences, children develop general science process skills from quite young ages. The popular phrase "scientist in the crib" (Gopnik et al., 2009) is used to describe infants' early science process skills, such as children's early explorations of cause and effect. More intentional process skills, the precursors to the foundational skills necessary in formal science, can be observed throughout early childhood. These include children's ability to recognize gaps

in their knowledge, to address these gaps through exploratory behavior and experimentation, and to analyze information and evidence to make conclusions (Jirout and Zimmerman, in press; Klahr et al., 2013). An important consideration in the development of these skills is *how* they develop, including methods used to this question.

The microgenetic approach provides many data points across trials during which performance changes and learning can be directly observed (Siegler and Crowley, 1991). This type of analysis can elucidate the underlying processes involved in learning a concept, which can then be targeted with instruction. The instructional methods most effective for developing these processes is a contentious area in the developmental and educational literature, with many factors to consider when determining how something should be taught. For example, Klahr and Chen's (2003) study of teaching children experimental design found that direct instruction of the content was most effective in children's learning, contrary to the push for 'discovery learning' in science education (see Chen and Siegler, earlier version of this article, for more detail). We extend the previous section on learning processes by discussing new methods of studying the development of scientific process skills, focusing on Bayesian Modeling and Cognitive Maps (Gopnik et al., 2004), and recent research addressing the issue of effectiveness of different pedagogical methods. We end with a brief discussion on the development of some specific science process skills.

Bayesian modeling uses children's prior knowledge or beliefs and the likelihood of different hypotheses to model learning (Perfors et al., 2011). It is thought to capture 'optimal inductive inference,' as well as making it possible to infer less than optimal ways that people think by comparing observed data to models (see Bayesian sections in the statistics chapters, or Perfors et al. (2011); for a nontechnical overview). Bayesian modeling has been successfully used to determine the processes behind children's understanding of causal mechanisms, and Bayes nets themselves have been used as a model of the processes by which children seem to infer causal relationships from evidence (Gopnik and Schulz, 2004). As an example, causal reasoning has been studied using a 'blicket detector,' a box that lights up and makes noise when different required actions and objects are used (Gopnik and Sobel, 2000). Gopnik and colleagues showed children a blicket detector that either did or did not light up when different configurations of objects were placed on it (Gopnik et al., 2001; Sobel et al., 2004). For example, children saw object *A* placed on the detector, and it lit up. It did not light up with object *B*, but did light up again when both *A* and *B* were placed on it. Children were told 'blickets' make the machine work, and were asked if each of the objects was a blicket. Children's causal inferences were similar to those generated with Bayes nets, considering the 'input,' or data children are given, and matched the predictive Bayesian models. Similarly, Xu and Kushnir (2013) and others suggest that infants demonstrate Bayesian-like inference during hypothesis testing, and Bayesian modeling has contributed to new knowledge about how children's scientific concepts develop in other domains as well.

In addition to using Bayesian models to study how development occurs, they have been informative for testing the effectiveness of pedagogical methods. Specifically, children seem to interpret types of instruction differently when forming theories that then influence learning behavior, with the different types of pedagogy influencing interpretation of the information provided. Bonawitz et al. (2011) observed preschool age children exploring a toy with four specific functions after being instructed to try and figure out how it works. When children were directly shown how the toy worked by demonstration of a single function, or saw another child receiving this demonstration, they explored for less time and discovered less novel functions than when they were not shown anything, or if the demonstrations they were shown were perceived as unintentional or incomplete. In a similar study, Buchsbaum et al. (2011) demonstrated several series of three actions on a toy, some of which included the same last two actions that resulted in music playing, some of which did not include these causal actions and did not result in music playing. When children perceived the person demonstrating the actions as teaching about the toy, they were less likely to discover the two-action cause of the music than children who thought the demonstrator was exploring the toy without knowing the underlying causal mechanism. Instead, they were more likely to directly imitate the actions, including the unnecessary third action in the sequence.

Consistent results have been found in other scientific domains as well. Fisher et al. (2013) found that in spatial learning, children learned shapes better with guided play than didactic instruction, though guided play also showed greater learning than free play. A possible explanation for these findings and several other similar findings in research on pedagogical effectiveness can perhaps be explained by Chi's (2009) theory that instruction is most effective when children are constructing evidence themselves, rather than participating or passively observing directed instruction, but learning is greatest when constructive activity is also *interactive*. This theory would suggest that instruction will be most effective in developing children's scientific concepts if it allows children to actively construct knowledge, but even further, to do so collaboratively or by some other interactive means. Further research is necessary to know what methods are most effective, and whether or not these results generalize across different learning domains. These studies suggest the importance of considering both learning goals and a teacher's knowledge base when planning pedagogical methods.

Another important aspect of early science learning and instruction is how to make the learning environment engaging. For children to learn in science, they must first be *curious* to learn something, and must recognize that there is something to be learned. Cognitive conflict, a strategy for inducing conceptual change, can lead to scientific curiosity (Klahr et al., 2013), or the desire to resolve uncertainty (Jirout and Klahr, 2012). In order for children to become curious, however, there must be some uncertainty or gap in their knowledge they wish to resolve. For example, in a study of children's exploration of causal relationships, Schulz and Bonawitz (2007) found that children preferred to explore a toy with an ambiguous causal mechanism over a novel toy, presumably to resolve the uncertainty presented by the ambiguity in the toy's function. Similar studies have found a consistent relationship with uncertainty or ambiguity and children's exploratory behavior, with an inverted U-shaped curve representing an 'optimal'

level of uncertainty that leads to exploratory behavior (Loewenstein, 1994; Litman et al., 2005), though this level appears to differ across individuals (Jirout and Klahr, 2012).

Question asking, a way children seek missing information (Kuhn and Dean, 2005), is both a foundational science process skill (NRC, 2012; Jirout and Zimmerman, in press) and a mechanism for learning science (Chouinard et al., 2007). Very young children are able to ask questions, with the majority of children's questions being identified as 'information seeking' rather than questions asking permission (e.g., Chouinard et al., 2007). Children's ability to ask *effective* questions, however, or those used to gain specific information necessary to resolve uncertainty, develops more slowly (King, 1991; Jirout and Klahr, under review). Kemler Nelson and colleagues have found that while 2- to 4-year-old children can ask topic-appropriate questions about artifact and animal functions, these questions often do not extend beyond seeking basic information of what something is called and what it does – lower level questions compared to more mechanistic ones necessary in science learning (Kemler Nelson et al., 2004; Greif et al., 2006). Asking questions, exploring, or experimenting provides children with information that they can then evaluate and use in terms of covariation as discussed above, in order to reason about causality and to revise or form theories.

The development of effective investigation through question asking or exploration is another area of continuous development, though, and is not very sophisticated in young children. For example, young children choose the better of two simple experiments in providing useful information, but are not very good at designing unconfounded experiments themselves until around third grade (Sodian et al., 1991; Piekny and Maehler, 2012) or even later (Bullock and Ziegler, 1999). Similar to their ability to reason about causal relationships in the studies discussed above, children can begin to use statistical information about covariance to reason and make conclusions from evidence, though this ability is limited by the complexity of the information children are given. In a simple study of a single cause, for example, children as young as 5 years can evaluate how evidence relates to a cause (Ruffman et al., 1993), and children as young as 4 years can predict how another person will change a prior belief with new covarying evidence (Koerberg et al., 2005). Preschool children were relatively good at drawing conclusions from perfect and partial covariation data (Piekny and Maehler, 2012), but even older children – 10–14 year olds – had difficulty interpreting noncovarying data (Kanari and Millar, 2004). Another source of difficulty for children is interpreting data that is inconsistent with their own beliefs, even with perfect covariation but especially with partial covariation (Koerber et al., 2005), though this finding is consistent with the performance seen in older children and even adults (Zimmerman and Croker, 2013).

Conclusion

In summary, research in recent decades has uncovered evidence that infants and young children may have considerably greater scientific knowledge at their fingertips than what was previously thought possible. Despite its early emergence, it is clear that this knowledge is far from complete. Although recent work has advanced our understanding of the complexities of scientific concepts in children, the conclusions drawn by Chen and Siegler 15 years ago still hold true today. Recent methodological advances, such as the use of computational modeling, allow new ways of exploring how different science knowledge is acquired. The growing literature on the understanding of conceptual domains across wide age ranges provides information about when children develop specific concepts, their ability to apply that knowledge to revise incorrect theories, the important role of language, and culture and experience in shaping scientific discovery. Findings on children's understanding of science concepts and how scientific reasoning develops have broad implications for education, including how different pedagogical methods might be most effective for learning in science.

See also: Bayesian Statistics; Cognitive Development in Infancy, Neural Mechanisms of; Cognitive Development: Mathematics Learning and Instruction; Concept Learning and Representation: Models; Learning Theories and Educational Paradigms; Science Education.

Bibliography

Anggoro, F.K., Waxman, S.R., Medin, D.L., 2008. Naming practices and the acquisition of key biological concepts: evidence from English and Indonesian. Psychological Science 19 (4), 314–319. http://dx.doi.org/10.1111/j.1467-9280.2008.02086.x.

Baillargéon, R., 2002. The acquisition of physical knowledge in infancy: a summary in eight lessons. Blackwell Handbook of Childhood Cognitive Development 1, 46–83.

Baillargéon, R., 2008. Innate ideas revisited: for a principle of persistence in infants' physical reasoning. Perspectives on Psychological Science 3 (1), 2–13.

Baillargéon, R., Kotovsky, L., Needham, A., 1995. The acquisition of physical knowledge in infancy. In: Sperber, D., Premack, D., Premack, A.J. (Eds.), Causal Cognition: A Multidisciplinary Debate. Clarendon Press, Oxford, pp. 79–116.

Baillargéon, R., Li, J., Ng, W., Yuan, S., 2009. An account of infants' physical reasoning. In: Woodward, A., Needham, A. (Eds.), Learning and the Infant Mind. Oxford University Press, New York, pp. 66–116.

Bidet-Ildei, C., Kitromilides, E., Orliaguet, J.-P., Pavlova, M., Gentaz, E., 2013. Preference for point-light human biological motion in newborns: contribution of translational displacement. Developmental Psychology. http://dx.doi.org/10.1037/a0032956. Advanced Online Publication.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., Schulz, L., 2011. The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. Cognition 120 (3), 322–330.

Brand, R.J., Shallcross, W.L., Sabatos, M.G., Massie, K.P., 2007. Fine-grained analysis of motionese: eye gaze, object exchanges, and action units in infant-versus adult-directed action. Infancy 11, 203–214.

Buchsbaum, D., Gopnik, A., Griffiths, T.L., Shafto, P., 2011. Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. Cognition 120 (3), 331–340.

Bullock, M., Ziegler, A., 1999. Scientific reasoning: developmental and individual differences. In: Weinert, F.E., Schneider, W. (Eds.), Individual Development from 3 to 12: Findings from the Munich Longitudinal Study. Cambridge University Press, Cambridge, pp. 38–54.

Chen, Z., Siegler, R.S., 2001. Scientific concepts: development in children. International Encyclopedia of the Social and Behavioral Sciences. In: Smelser, N.J., Baltes, P.B., Kintsch, W. (Eds.), Cognitive Psychology & Cognitive Science, vol. 21. Elsevier, Oxford, pp. 13714–13719.

Chi, M.T.H., 2009. Active-constructive-interactive: a conceptual framework for differentiating learning activities. Topics in Cognitive Science 1 (1), 73–105. http://dx.doi.org/10.1111/j.1756-8765.2008.01005.x.

Choinard, M.M., Harris, P.L., Maratsos, M.P., 2007. Children's questions: a mechanism for cognitive development. Monographs of the Society for Research in Child Development 72 (1).

Cicchino, J.B., Aslin, R.N., Rakison, D.H., 2011. Correspondences between what infants see and know about causal and self-propelled motion. Cognition 118 (2), 171–192. http://dx.doi.org/10.1016/j.cognition.2010.11.005.

Cimpian, A., Markman, E.M., 2009. Information learned from generic language becomes central to children's biological concepts: evidence from their open-ended explanations. Cognition 113 (1), 14–25. http://dx.doi.org/10.1016/j.cognition.2009.07.004.

De Haan, M., Pascalis, O., Johnson, M.H., 2002. Specialization of neural mechanisms underlying face recognition in human infants. Journal of Cognitive Neuroscience 14 (2), 199–209. http://dx.doi.org/10.1162/089892902317236849.

DeLoache, J.S., 1987. Rapid change in the symbolic functioning of very young children. Science 238 (4833), 1556–1557.

Farrar, M.J., Boyer-pennington, M., 2011. Developmental changes in children's inductive inferences for biological concepts: implications for the development of essentialist beliefs. 539, 525–539. http://dx.doi.org/10.1002/icd.

Fisher, K.R., Hirsh-Pasek, K., Newcombe, N., Golinkoff, R.M., 2013. Taking shape: supporting preschoolers' acquisition of geometric knowledge through guided play. Child Development 84 (6), 1872–1878.

Frick, A., Ferrara, K., Newcombe, N.S., 2013. Using a touch screen paradigm to assess the development of mental rotation between $3^1/_2$ and $5^1/_2$ years of age. Cognitive Processing, 1–11.

Frye, D., Zelazo, P.D., Brooks, P.J., Samuels, M.C., 1996. Inference and action in early causal reasoning. Developmental Psychology 32 (1), 120.

Gelman, S.A., 2009. Learning from others: children's construction of concepts. Annual Review of Psychology 60, 115–140.

Gelman, S.A., Markman, E.M., 1986. Categories and induction in young children. Cognition 23, 183–209. http://dx.doi.org/10.1016/0010-0277(86)90034-X.

Göksun, T., George, N.R., Hirsh-Pasek, K., Golinkoff, R.M., 2013. Forces and motion: how young children understand causal events. Child Development 84 (4), 1285–1295.

Goldberg, R.F., Thompson-Schill, S.L., 2009. Developmental "roots" in mature biological knowledge. Psychological Science 20 (4), 480–487.

Gopnik, A., Schulz, L., 2004. Mechanisms of theory formation in young children. Trends in Cognitive Sciences 8 (8), 371–377.

Gopnik, A., Sobel, D.M., 2000. Detecting blickets: how young children use information about novel causal powers in categorization and induction. Child Development 71 (5), 1205–1222.

Gopnik, A., Sobel, D.M., Schulz, L.E., Glymour, C., 2001. Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. Developmental Psychology 37 (5), 620.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D., 2004. A theory of causal learning in children: causal maps and Bayes nets. Psychological Review 111 (1), 3.

Gopnik, A., Meltzoff, A.N., Kuhl, P.K., 2009. The Scientist in the Crib: Minds, Brains, and How Children Learn. HarperCollins, New York.

Greif, M.L., Kemler Nelson, D.G., Keil, F.C., Gutierrez, F., 2006. What do children want to know about animals and artifacts? Domain-specific Requests for Information 17 (6), 455–460.

Hespos, B., 2006. De'calage in infants' knowledge about occlusion and containment events: converging evidence from action tasks. Cognition 99, B31–B41.

Hespos, S.J., Ferry, A.L., Rips, L.J., 2009. Five-month-old infants have different expectations for solids and liquids. Psychological Science 20 (5), 603–611. http://dx.doi.org/10.1111/j.1467-9280.2009.02331.x.

Inagaki, K., Hatano, G., 2002. Young Children's Naive Thinking about the Biological World. Psychology Press, New York.

Jirout, J., Klahr, D., 2012. Children's scientific curiosity: in search of an operational definition of an elusive concept. Developmental Review 32, 125–160.

Jirout, J., Zimmerman, C. Science process skills in young children. To appear in Trundle, K.C., Saçkes, M. (Eds.), Research in Early Childhood Science Education. Springer Publishing Company, Inc., New York, in press.

Kamewari, K., Kato, M., Kanda, T., Ishiguro, H., Hiraki, K., 2005. Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. Cognitive Development 20 (2), 303–320. http://dx.doi.org/10.1016/j.cogdev.2005.04.004.

Kanari, Z., Millar, R., 2004. Reasoning from data: how students collect and interpret data in science investigations. Journal of Research in Science Teaching 41, 748–769.

Keil, F.C., 2010. The feasibility of folk science. Cognitive Science 34 (5), 826–862.

Kelemen, D., Widdowson, D., Posner, T., Brown, A.L., Casler, K., 2003. Teleo-functional constraints on preschool children's reasoning about living things. Developmental Science 6 (3), 329–345. http://dx.doi.org/10.1111/1467-7687.00288.

Kiefer, M., Pulvermüller, F., 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex. A Journal Devoted to the Study of the Nervous System and Behavior 48 (7), 805–825. http://dx.doi.org/10.1016/j.cortex.2011.04.006.

King, A., 1991. Effects of training in strategic questioning on children's problem-solving performance. Journal of Educational Psychology 83, 307–317.

Klahr, D., Chen, Z., 2003. Overcoming the positive-capture strategy in young children: learning about indeterminacy. Child Development 74, 1275–1296.

Klahr, D., Matlen, B., Jirout, J., 2013. Children as scientific thinkers. In: Feist, G., Gorman, M. (Eds.), Handbook of the Psychology of Science. Springer, New York, pp. 223–248.

Koerberg, S., Sodian, B., Thoermer, C., Nett, U., 2005. Scientific reasoning in young children: preschoolers' ability to evaluate covariation evidence. Swiss Journal of Psychology 64, 141–152.

Kosugi, D., Ishida, H., Fujita, K., 2003. 10-month-old infants' inference of invisible agent: distinction in causality between object motion and human action. Japanese Psychological Research 45 (1), 15–24.

Kuhn, D., Dean, D., 2005. Is developing scientific thinking all about learning to control variables? Psychological Science 16, 866–870.

Kemler Nelson, D.G., Egan, L.C., Holt, M.B., 2004. When children ask, "What is it?" what do they want to know about artifacts? Psychological Science 15, 384–389.

Leddon, E.M., Waxman, S.R., Medin, D.L., 2008. Unmasking "alive:" children's appreciation of a concept linking all living things. Journal of Cognition and Development 9 (4), 461–473. http://dx.doi.org/10.1080/15248370802678463.

Litman, J.A., Hutchins, T.L., Russon, R.K., 2005. Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. Cognition & Emotion 19, 559–582.

Loewenstein, G., 1994. The psychology of curiosity: a review and reinterpretation. Psychological Bulletin 116, 75–98.

Luo, Y., Kaufman, L., Baillargeon, R., 2009. Young infants' reasoning about physical events involving inert and self-propelled objects. Cognitive Psychology 58 (4), 441–486.

Medin, D., Waxman, S., Woodring, J., Washinawatok, K., 2010. Human-centeredness is not a universal feature of young children's reasoning: culture and experience matter when reasoning about biological entities. Cognitive Development 25 (3), 197–207. http://dx.doi.org/10.1016/j.cogdev.2010.02.001.

Möhring, W., Frick, A., 2013. Touching up mental rotation: effects of manual experience on 6-month-old infants' mental object rotation. Child Development 84 (5), 1554–1565.

National Research Council, 2012. A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. National Academies Press, Washington, DC.

Newcombe, N., 1989. The development of spatial perspective taking. In: Reese, H.W. (Ed.), Advances in Child Development and Behavior, vol. 22. Academic Press, New York, pp. 203–247.

Nguyen, S., Rosengren, K., 2004. Parental reports of children's biological knowledge and misconceptions. International Journal of Behavioral Development 28 (5), 411–420. http://dx.doi.org/10.1080/01650250444000108.

Perfors, A., Tenenbaum, J.B., Griffiths, T.L., Xu, F., 2011. A tutorial introduction to Bayesian models of cognitive development. Cognition 120 (3), 302–321.

Piaget, J., 1929. The Child's Conception of the World. Harcourt, Brace, New York.

Piaget, J., 1952. The Origins of Intelligence in Children. Intentional Universities Press, New York.

Piekny, J., Maehler, C., 2012. Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. British Journal of Developmental Psychology. http://dx.doi.org/10.1111/j.2044-835X.2012.02082.x.

Pinker, S., 1997. How the Mind Works. Norton, New York.

Rakison, D.H., Poulin-Dubois, D., 2001. Developmental origin of the animate–inanimate distinction. Psychological Bulletin 127 (2), 209–228. http://dx.doi.org/10.1037/0033-2909.127.2.209.

Rostad, K., Yott, J., Poulin-Dubois, D., 2012. Development of categorization in infancy: advancing forward to the animate/inanimate level. Infant Behavior & Development 35 (3), 584–595. http://dx.doi.org/10.1016/j.infbeh.2012.05.005.

Ruffman, T., Perner, J., Olson, D.R., Doherty, M., 1993. Reflecting on scientific thinking: children's understanding of the hypothesis-evidence relation. Child Development 64, 1617–1636.

Saxe, R., Tenenbaum, J.B., Carey, S., 2005. Secret agents inferences about hidden causes by 10-and 12-month-old infants. Psychological Science 16 (12), 995–1001.

Schulz, L., Bonawitz, E.B., 2007. Serious fun: preschoolers play more when evidence is confounded. Developmental Psychology 43, 1045–1050.

Schulz, L.E., Goodman, N.D., Tenenbaum, J.B., Jenkins, A.C., 2008. Going beyond the evidence: abstract laws and preschoolers' responses to anomalous data. Cognition 109 (2), 211–223.

Setoh, P., Wu, D., Baillargeon, R., Gelman, R., 2013. Young infants have biological expectations about animals. Proceedings of the National Academy of Sciences of the United States of America 110 (40), 15937–15942. http://dx.doi.org/10.1073/pnas.1314075110.

Siegler, R.S., Crowley, K., 1991. The microgenetic method: a direct means for studying cognitive development. American Psychologist 46 (6), 606.

Sobel, D.M., Tenenbaum, J.B., Gopnik, A., 2004. Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. Cognitive Science 28 (3), 303–333.

Sodian, B., Zaitchik, D., Carey, S., 1991. Young children's differentiation of hypothetical beliefs from evidence. Child Development 62, 753–766.

Tarlowski, A., 2006. If it's an animal it has axons: experience and culture in preschool children's reasoning about animates. Cognitive Development 21, 249–265.

Unsworth, S.J., Levin, W., Bang, M., Washinawatok, K., Waxman, S.R., Medin, D.L., 2012. Cultural differences in children's ecological reasoning and psychological closeness to nature: evidence from Menominee and European American children. Journal of Cognition and Culture 12 (1), 17–29. http://dx.doi.org/10.1163/156853712X633901.

Vosniadou, S., 2002. On the nature of naïve physics. In: Limón, M., Mason, L. (Eds.), Reconsidering Conceptual Change. Issues in Theory and Practice. Kluwer Academic Publishers, The Netherlands, pp. 61–76.

Wang, S.H., Baillargeon, R., Patterson, S., 2005. Detecting continuity violations in infancy: a new account and new evidence from covering and tube events. Cognition 95, 129–173.

Wellman, H.M., Gelman, S.A., 1998. Knowledge acquisition in foundational domains. In: Kuhn, D., Damon, W., Siegler, R., Lerner, R.M. (Eds.), Handbook of Child Psychology, Cognition, Perception, and Language, vol. 2. Wiley, New York, pp. 523–573.

Woodward, A.L., Sommerville, J.A., Guajardo, J.J., 2001. How infants make sense of intentional action. In: Malle, B.F., Moses, L.J., Baldwin, D.A. (Eds.), Intentions and Intentionality: Foundations of Social Cognition. The MIT press, Cambridge, MA, pp. 149–169.

Xu, F., Kushnir, T., 2013. Infants are rational constructivist learners. Current Directions in Psychological Science 22 (1), 28–32.

Zimmerman, C., Croker, S., 2013. Learning science through inquiry. In: Feist, G., Gorman, M. (Eds.), Handbook of the Psychology of Science. Springer, New York, pp. 49–70.

Scientific Concepts during Childhood, Development of

Mesut Saçkes, Balıkesir University, Balıkesir, Turkey
Kathy C Trundle and Mandy McCormick Smith, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents a synthesis of descriptive and instructional studies on the development of scientific concepts during childhood. Preschoolers' and young children's understandings of earth and space science concepts, physical science concepts, and life science concepts are described. Strategies to support the development of scientific understandings in young children are provided.

Children construct explanations about how things work in the natural world around them, and their ideas, in general, differ from scientifically accepted models. Researchers have used different terms including *naive ideas*, *alternative conceptions*, *misconceptions*, *preconceptions* and *initial explanatory frameworks* to describe children's alternative understandings. A considerable number of studies have been devoted to children's understandings of the natural world and changes that occur in these ideas as a result of development and/or instructional intervention (e.g., Baxter, 1989; Carey, 1985; Saçkes et al., 2010; Trundle and Saçkes, 2012; Vosniadou, 1994; Vosniadou and Brewer, 1992). Here, we present a synthesis of descriptive and instructional studies on the development of scientific concepts during childhood. We organized the information under four headings: earth and space science concepts, physical science concepts, life science concepts, and strategies to support the development of scientific understandings.

Children's Understanding of Earth and Space Science Concepts

Beginning with Piaget's pioneering studies, children's conceptual understandings of various earth and space science concepts have been reported in research literature. These studies mainly focused on children's understandings of scientific concepts targeted in elementary school and include weather phenomena, the night and day cycle, the shape of the earth, and moon phases.

Studies that examined children's understandings of the water cycle and related phenomena demonstrated that most young children believe that rain is water, that they tend not to relate clouds with rain, and that they also believe that rain simply falls from the sky (Saçkes et al., 2010). Young children tend to elucidate the mechanism of rainfall with an outside agent at early ages and they often attribute human characteristics to clouds and rain (Miner, 1992). Although most young children tend to pay attention to differences that occur in clouds when they change as a prelude to rain, many young children continue not to associate rain with clouds (Saçkes et al., 2010). They also have difficulty understanding the water cycle, possibly due to their poor performance on conservation and reversibility tasks, which encompass the process of rainfall (Bar, 1989). Not until the age of 9–10 years do children begin to appreciate the natural origin of clouds and consider clouds as a cause of rain as well as a sign of it (Bar, 1989; Piaget, 1972).

Many researchers believe that children's inability to explain the mechanism of rainfall is related to their lack of understanding of the major concepts of the water cycle, namely evaporation and condensation (Bar, 1989; Saçkes et al., 2010). Elementary school children start using the concepts of condensation and evaporation in describing the formation of clouds and rain (Moyle, 1980). However, many children, despite instruction, have difficulty understanding and explaining the role of condensation and evaporation in the formation of rain and clouds and in the water cycle process (Saçkes et al., 2010). Although the idea that rain comes from colliding clouds continue to be entertained by some children in late elementary grades, many children, as a result of schooling, begin to believe that rain falls when clouds become cold or heavy or when raindrops become heavy and fall (Bar, 1989; Saçkes et al., 2010). Children explain rain as water vapor all gathered up in clouds, but few children refer to changes in temperature as part of their explanation of how rain occurs and of the cyclic process that makes rain (Moyle, 1980).

The geographic location in which children live seems to influence their understandings of the mechanism of rainfall. Children from mountainous locations tend to have a better understanding of the water cycle than children from desert locations. They understand precipitation and condensation concepts significantly better than the children who live in deserts (Miner, 1992). This finding could be attributed to the fact that children in mountain environments tend to have richer experiences with the water cycle than children in desert environments. Thus, children who live in the mountains tend to understand the water cycle and related phenomena better than the children who live in deserts.

In addition to rainfall, Piaget also investigated children's understanding of the day and night cycle. Piaget (1972) reported that children perceive the day and night cycle as a result of the movement of the sun in the sky. Other studies reported similar findings – most young children attribute the day and night cycle to the movement of the sun in the sky or provide religious explanations for the phenomenon (Sharp, 1996). Vosniadou (1994) found that young children are likely to explain the day and night cycle with reference to their initial mental models, which probably form from their everyday experiences. Young children tend to think that the sun goes

down behind the mountains or that clouds cover up the sun, causing night. Older children construct a synthetic mental model by combining some scientific ideas with their initial models. Children who hold synthetic models tend to think that the sun and the moon revolve around the stationary earth every 24 h, or that the earth rotates in an up/down motion relative to the fixed positions of the sun and the moon on opposite sides of the earth.

Children's conceptions of the shape of the earth have been the focus of several studies. In their seminal work, Vosniadou and Brewer (1992) argued that most children consistently use a small number of well-defined mental models of earth to explain the appearance of the earth. Younger children tend to have a 'flat earth model' and they conceive of the earth as a rectangle or disc. While older children tend to have a 'spherical-earth model,' some children, however, maintain a combination mental model of a flat-and-spherical earth. This type of blended or combined model is called a 'synthetic mental model.' Three types of synthetic models were identified: the dual-earth model, the hollow-sphere model, and the flattened-sphere model. Children who hold to the dual-earth model believe that there are two earths – a flat one where people live and a round one up in the sky. Children with the hollow-sphere conceptual model believe that people live on the flat surface of the earth, which is located at the bottom of the spherical earth. Children who hold to the flattened-sphere model believe that the top and bottom of the earth are flat, and these are the places where people live.

Children's conceptions of the moon and observable changes in its appearance have also been examined in the literature. Studies demonstrated that most children, even at early ages, are aware that the appearance of the moon changes over time. Although most young children are aware that the appearance of the moon changes, they have a limited understanding about the pattern of this change (Piaget, 1972; Za'rour, 1976). Studies suggest that young children's interpretation of the changes in the moon's appearance differs with age. At early ages, children see human activity and supernatural forces as being responsible for the changes in the shape of the moon. Examples of their concepts include the notion that the moon is born or cut by people, or that fairies change the moon's appearance. Children also explain these changes via natural forces such as wind and clouds (Piaget, 1972; Haupt, 1950).

When children realize that changes in the moon are actually changes in the moon's appearance and not changes in the moon's size, they begin to construct a model to explain moon phases, based on the idea that something must be blocking either the sun's light from reaching the moon or the moon's light from reaching to the earth. Baxter (1989) reported four explanations that are based on this idea. Children believe that clouds, a planet other than earth, the sun, or the earth itself, cause moon phases by blocking the light. The idea that the shadow of the earth causes moon phases was reported as the most common explanation among the participants in Baxter's study. Indeed, this alternative concept, which is often called the 'eclipse model,' was reported as the most common alternative concept in the literature (Hobson et al., 2010; Trundle et al., 2002).

Children's Understanding of Physical Science Concepts

From infancy, children develop their own ideas and understandings of many physical phenomena. For example, children construct knowledge of force, heat, sounds, and states of matter (e.g., solids and liquids) early in life (Spelke, 1991). The development of infants' understanding of the physical world seems to be influenced by five major constraints: continuity, solidity, no action at a distance, gravity, and inertia (Spelke, 1991). 'Continuity' refers to an understanding that objects continue to exist and move on a connected path; 'solidity' implies that two objects cannot occupy the same place in a space at the same time; 'no action at a distance' suggests that objects travel independently in space unless they meet; 'gravity,' which is a force, causes objects to move downward when they are not supported; and 'inertia,' a property of matter, is the resistance of an object to change in its motion or state of rest (inactivity).

Studies suggest that infants as young as 3 months expect objects to move on a connected and unobstructed path. By the age of 6 months, infants develop an initial appreciation of gravity and inertia and they make valid inferences about the behavior of falling objects (Spelke, 1991). However, children often make miscalculations when falling objects are acted upon by outside forces. Bascandziev and Harris (2010), who used gravity tubes to describe preschoolers' misconceptions of gravity and free fall, described children's gravitational misconceptions as 'gravity error.' As with other misconceptions, they found that children had a hard time overcoming 'gravity error,' even after they experimented with materials and gathered empirical evidence of falling objects. Their research indicates that input from others, such as peers and teachers, has more influence on children's understandings of gravity-related concepts than their own evidenced-based observations. These findings hold widespread developmental implications for young children's science engagement and conceptual development.

Piaget (1971) described conceptions and learning progression about sound in children aged 4–8 years. He found that children aged 4–5 years believe that nothing passes between the object emitting a sound and the ear of the person who receives the sound, indicating that sounds need a clear unobstructed path in order to be heard. Six-year-old children believe that sounds stay within objects even if they are not heard. They conceive that sounds try to reach a person's ears, and when they cannot, the sounds return to the place/object where they originated. Beginning at age 7 years, children visualize sound as straight lines emanating in all directions (Piaget, 1971; Lautrey and Mazens, 2004). Results from more recent studies indicate that young children tend to attribute substantiality to sound and they think sound cannot pass through solids or liquids. When they experience that sounds can be heard through some solids, they conclude that the solids must include invisible holes through which the sound passes. Moreover, children believe that for sound to pass through a solid, the sound must be harder or stronger than the solid. These presuppositions were proposed to explain children's mental models related to permanence, continuity, and solidity (Mazens and Lautrey, 2003).

Results of studies demonstrate that even very young children understand that matter can break into tiny invisible pieces (i.e., molecules or ions of the substance) by dissolving (Nakhleh and Samarapungavan, 1999). Children understand that different substances or solutes dissolve differently, depending on the properties of the solution or solvent used for dissolving. Children tend to use a self-evident type of logic in explaining the properties of matter, and they construct ideas about different types and states of matter. Children tend to think of some solid matter with no visible granularity, such as wood and copper wire, as being continuous; whereas they recognize the particulate nature of solids that exist in small grains like sugar and table salt. Children also are more likely to recognize the particulate nature of liquids and gases rather than solids (Nakhleh and Samarapungavan, 1999). Piagetian theory supports children's responses when they distinguish between objects and matter itself. As children develop, they are more able to distinguish between the extensive and intensive properties of matter, and therefore discern between the matter and the object (Krnel et al., 2005). Very young children (aged 3 years) often classify matter on an item-by-item basis. Thus, they change their classification criteria with each new object. Even at the age of 7 years, children have difficulty grouping more than two items together, and color is the predominate grouping factor. By the age of 11 years, children use two or more factors when classifying matter – e.g., by substance (i.e., powders/solids/liquids) and by actions (i.e., how the items would be used). Even at this advanced developmental age, color remains an important factor in considering matter properties and groupings (Krnel et al., 2005).

Although the concepts of heat and temperature are separate entities, they are often taught together and used interchangeably in nonscientific dialogue. Even though these are everyday experiences, older students and adults alike have difficulty understanding thermal events and thermodynamics (Lewis and Linn, 1994). When considering issues of heat and temperature, very young children have difficulty using the particulate nature of matter to envision or explain the underlying scientific processes. The first mental models children hold, associate heat with burning or melting. However, these initial models do not describe the internal process. When asked why an object is melting, children may respond "because it is smaller" (Shayer and Wylam, 1981). Additionally, younger students often speak of heat in terms of being instantly created or destroyed, much like an object with an on/off switch (Albert, 1978). Albert found that the child's view changes to a more dynamic characterization of heat between the ages of 7 and 8 years. She also noted that children of this age begin to describe heat as a process. In her research, very young children did not discern between 'hot' and 'warm.' This distinction, including degrees of temperature, begins between the ages of 8 and 10 years. During this developmental stage, children begin to understand heat as an energy transfer (Albert, 1978). Children do not generally incorporate a caloric or kinetic view of heat until later in their development (Shayer Wylam, 1981). Similarities exist between 8-year-old children's conceptual organization of heat and sound. Lautrey and Mazen's (2004) study found the categories of explanation for sound (permanence, continuity, and solidity) were also applicable to children's mental models of heat. They found that children more often associated many trajectories (or paths) with heat than with sound. They theorized this difference to be due to the presence of many receptors for heat in the human body, whereas there are only two ears as primary receptors for sound. Thus, children can more easily observe heat in other objects, whereas the effects of sounds on objects cannot be readily observed (Lautrey and Mazens, 2004).

Students experience 'speed' or velocity from infancy. For example, children race on the playground, move from the path of a rolling ball, observe cars on the road, and comment on the 'fastness' of objects. Children, however, hold many naïve ideas regarding their everyday experiences with velocity. For example, young children often attribute closer objects as having higher speeds (Cross and Pithkey, 1988). Understanding the dynamic nature of motion is difficult for even high school children and some college students to achieve. Previously, students' understandings of motion were classified as either dynamic or nondynamic/direction-only. The dynamic view of motion includes the development of a Newtonian-like perspective. Dykstra and Sweet (2009) suggested a new classification of young students' views of motion, the 'snap-shot view.' With this view, children see that velocity varies at different times and that motion changes. The researchers' evidence suggests that speed differences are important aids to deepening children's understandings of force and motion. Results reported in the current literature are inconclusive and educators are uncertain if attaining a dynamic view of motion is a proper curriculum goal for fourth grade students. However, research results indicate that sixth grade students are able to achieve a dynamic view of motion, but they require more physical manipulation and experimentation than older students (Dykstra and Sweet, 2009).

The research literature also includes studies on children's understanding of force, another physical science concept. The results of these studies indicate that young children consider force to be a property of physical objects, and they relate force to the object's weight and size. Young children understand force to be something that is exerted only on heavy and big objects or something that is an internal property of physical objects. Researchers often describe these naive ideas as an 'internal force' model. By the age of 8–12 years the internal force model usually is replaced by another model called an 'acquired force' model. In this model, children only associate a force with an object when the object moves (Vosniadou, 1994).

Children's presuppositions influence how they understand the natural world. The following presuppositions seem to restrict children's understanding of the concept of force: (1) physical objects exist in the natural world, (2) these physical objects have properties, (3) force is a property of physical objects, (4) inanimate objects do not move by themselves, (5) the motion of inanimate objects requires an explanation, and (6) the explanation for movement of inanimate objects should be in terms of a causal agent (Vosniadou, 1994). These presuppositions influence children's interpretations of their everyday experiences, and they give rise to the following specific naive theories about force: (1) some objects respond to external actions upon them while others do not, (2) inanimate objects need to be pushed or pulled to move, and (3) inanimate objects stop moving by themselves (Vosniadou, 1994).

Researchers also have examined children's understanding of sinking and floating. Density and buoyancy are higher-order scientific ideas. A person must have foundational ideas about both matter and volume in order to scientifically understand density. Further complexity is added to this concept, because density is a property of substances but not necessarily of objects. Piaget, who provided the first framework for children's development with regard to density, labeled stages from stage 1 (ages 4–5 years) to stage 4 (age 12 years). Piaget claims that children at stage 4 are able to correctly explain how buoyancy correlates with relative weight, and they are able to understand mass per unit volume or density (1972). More recent studies demonstrate that preschoolers use their previous life experiences to predict which objects will float or sink and why. Children base their predictions on everyday experiences, explanations, on information they hear from adults, or they use a self-evident type of logic. Preschoolers usually explain flotation as being due to air inside an object, or they suggest that the weight of an object causes it to float, with lighter objects floating and heavier objects sinking. Children also use size and the substance of which the object is made to explain flotation, but they often conceive the size of the object to be synonymous with its weight (Havu-Nuutinen, 2005). What is known as the 'size-weight illusion' (i.e., believing that between two objects, the smaller object will be lighter when lifted) has been broadly represented in the research literature. Overcoming the size-weight illusion can be difficult for both older students and adults (Flanagan and Beltzner, 2000). Interestingly, children do not exhibit the 'size-weight illusion' until they are aged 4–5 years (Kohn, 1993).

Children's Understanding of Life Science Concepts

As with the physical and earth and space science concepts previously discussed, young children's ideas about living things differ from scientific norms (Piaget, 1972). Young children usually develop the alternative conception that an object moving in any way, such as rolling car, is alive. This conception becomes more sophisticated as children develop, and they begin to consider that only things that move on their own initiative are alive. Finally, at around the age of 10 years, children develop an understanding that animals and plants are alive, and many children at this age limit their thinking to only plants and animals. Children's initial ideas about living things may be attributed to their animistic thinking. For example, many young children believe plants are not alive, whereas they think the sun is alive because of the way it appears to move across the sky.

One source for children's animistic responses could be their poor knowledge of biology, which prevents them from including plants and animals together in a single category (Carey, 1985). According to Carey, children's understanding of biology undergoes a radical change between the ages of 4 and 10 years, and their early understandings of plants and animals can generally be described as anthropocentric (i.e., children use human beings as a reference to understand other living things).

Preschoolers usually do not assign typical animal properties to animals. For example, children tend to attribute eating and breathing only to humans and not to other animals. Older children, however, perceive the ability to eat and breathe as a distinguishing property of all animals (Carey, 1985). Preschoolers understand that living things, such as animals and plants, grow, whereas artifacts do not. Young children tend to be aware of animals' basic physiological needs, and they usually are aware of conditions that support animal and plant growth. While preschool children generally seem to be able to categorize living things, many also have difficulty identifying them as 'alive' (Backscheider et al., 1993). In other words, preschoolers tend to understand that plants and animals, like all living things, share some general basic properties (e.g., both plants and animals reproduce or make babies and they both grow). Despite this general knowledge, they still are unable to conclude that plants are living things. Young children seem to attribute intentional actions to animals but not to plants. Therefore, young children tend not to classify plants as living things. However, when preschoolers are taught that plants, like animals, are able to move in response to stimuli (e.g., phototropism where plants grow toward light and thigmotropism where plants move in response to touch) only then do they conclude that plants are alive. The concept that plants can move in goal-directed ways plays a more crucial role in the restructuring of children's ideas about what it means to be alive, than the understanding that plants need water and that they can grow (Opfer and Siegler, 2004).

Instructional Strategies to Support the Development of Science Concepts

Results from research studies demonstrate that alternative conceptions are often pervasive and resistant to change through traditional forms of instruction. Instructional strategies are more likely to be successful if they are specifically geared toward helping children restructure their existing conceptual understandings. These types of instructional strategies are based on conceptual change theory, and they usually involve the following elements in their designs: (1) take children's prior knowledge into consideration, (2) devise a learning environment to create cognitive conflict or dissonance, (3) support inquiry activities, (4) provide models and external representations of natural phenomena, (5) support the use and development of meta-conceptual awareness and strategy use, and (6) foster task-related motivational beliefs (Trundle and Saçkes, 2013).

Science education literature emphasizes inquiry-based instruction as one of the most effective ways to promote conceptual understandings of scientific phenomena (Trundle and Saçkes, 2012). Inquiry-based instruction includes hands-on activities with sense-making discussions and requires the active participation of learners. In such a learning environment, children are more likely to construct a scientific understanding of the targeted natural phenomena through restructuring of their existing alternative understandings (Anderson, 2002, 2007; Trundle et al., 2010). Children also have opportunities to develop inquiry skills (e.g., observing, inferring, posing questions, forming hypotheses, and developing conclusions based on the analysis of data) (Anderson, 2007).

Several forms of inquiry-based instruction have been developed and described in the literature, such as the 5E learning cycle, the predict-observe-explain model, and the

preschool learning cycle. These models have been developed as both theoretical and practical answers to developmentally appropriate classroom implementations. Bybee's 5E learning cycle consists of five phases: engagement, exploration, explanation, elaboration, and evaluation (Trowbridge and Bybee, 1996). The 5E cycle emerged from Piagetian theory; through the exploration phase, children experience assimilation and/or disequilibrium. Children then make accommodations and organize in the explanation and elaboration phases respectively (Tobin et al., 1994). This 5E learning cycle was developed for the elementary classroom with young children's capabilities in mind.

The predict-observe-explain (POE) model, by White and Gunstone (1992), developed as a way to discover the individual student's predictions regarding a specific event, and the reasoning behind these predictions. Besides noting children's preconceptions, the POE model also provides teachers with useful information regarding children's thinking. The model enhances student engagement, discussion, and the generation of follow-up investigations (Hsu et al., 2011).

The preschool learning cycle (Trundle et al., 2013) integrates Bybee's effective 5E learning cycle with the developmental needs of the youngest learners in preschool. The three-part cyclical process starts with play, progresses to exploration, and is followed by discussion. This cycle includes hearts-on, hands-on, and minds-on learning. During the hearts-on phase, the teachers tap into the children's curiosities and link their prior knowledge with the new material. Children are given opportunities during the hands-on phase to interact with the materials in new and different ways. Finally, teachers provide age-appropriate labels or scientific terms to the learning process during the minds-on phase. Teachers can implement this model to connect to very young children's interests and foster their scientific explorations (Trundle et al., 2012).

One instructional strategy, a project approach, successfully incorporates inquiry into the teaching of science concepts. This approach engages children with in-depth investigations into science concepts that are of interest to them. Children ask questions or define problems, use tools to collect and analyze data, and produce artifacts to communicate and present their understandings. They answer questions they themselves posed or solve problems that they defined. The project approach incorporates three phases: phase 1, beginning a project; phase 2, developing the project; and phase 3, concluding the project. In each of the phases, the students move through five exercises: discussion, fieldwork, representation, investigation, and display (Katz and Chard, 2000). Teachers can use this approach to scaffold developmentally appropriate projects into their curricula (Chard, 1998; Helm and Katz, 2010; Katz, 1994).

Some researchers argue that inquiry-based instructional approaches place a heavy cognitive load on learners, especially younger children who have little cognitive processing capacity, and thus discourage children from processing novel information and restructuring their conceptual understandings (Hobson et al., 2010; Kirschner et al., 2006; Mayer, 2004). Therefore, inquiry-based instructional approaches must involve developmentally appropriate structuring and scaffolding for young children (Trundle and Saçkes, 2012). For example, guided inquiry, with an appropriate level of scaffolding and guidance in formulating research questions and procedures, should be used in early childhood classrooms rather than open inquiry where children, on their own, formulate research questions as well as procedures to answer them. Teachers can reduce the level of guidance as children develop expertise in the targeted domain, and as they age, develop, and move into the upper grades (Bell et al., 2005; Banchi and Bell, 2008).

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Development, Educational Theories of; Concept Formation in Childhood: Cross-Cultural Perspectives; Piaget's Theory of Cognitive Development.

Bibliography

Albert, Edna, 1978. Development of the concept of heat in children. Science Education 62, 389–399.

Anderson, Ronald D., 2002. Reforming science teaching: what research says about inquiry. Journal of Science Teacher Education 13, 1–12.

Anderson, Ronald D., 2007. Inquiry as an organizing theme for science curricula. In: Abell, Sandra, Lederman, Norm (Eds.), Handbook of Research on Science Education. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 807–830.

Backscheider, Andrea G., Schatz, Marilyn, Gelman, Susan A., 1993. Preschoolers' ability to distinguish living kinds as a function of regrowth. Child Development 64, 1242–1257.

Banchi, Heather, Bell, Randy, 2008. The many levels of inquiry. Science and Children 46, 26–29.

Bar, Varda, 1989. Children's views about the water cycle. Science Education 73, 481–500.

Bascandziev, Igor, Harris, Paul L., 2010. The role of testimony in young children's solution of a gravity-driven invisible displacement task. Cognitive Development 25, 233–246.

Baxter, John, 1989. Children's understanding of familiar astronomical events. International Journal of Science Education 11, 502–513.

Bell, Randy L., Smetana, Lara, Binns, Ian, 2005. Simplifying inquiry instruction. The Science Teacher 72, 30–33.

Carey, Susan, 1985. Conceptual Change in Childhood. MIT Press, Cambridge, MA.

Chard, S.C., 1998. The Project Approach: Making Curriculum Come Alive. Scholastic, New York.

Cross, R.T., Pithkey, A., 1988. Speed, education and children as pedestrians: a cognitive change approach to a potentially dangerous naive concept. International Journal of Science Education 10, 531–540.

Dykstra, Dewey I., Sweet, Dale R., 2009. Conceptual development about motion and force in elementary and middle school students. American Journal of Physics 77, 468–476.

Flanagan, J. Randall, Beltzner, Michael A., 2000. Independence of perceptual and sensorimotor predictions in the size-weight illusion. Nature Neuroscience 3, 737–741.

Haupt, George W., 1950. First grade concepts of the moon, part 2 by interview. Science Education 34, 224–234.

Havu-Nuutinen, Sari, 2005. Examining young children's conceptual change process in floating and sinking from a social constructivist perspective. International Journal of Science Education 27, 259–279.

Helm, J.H., Katz, L.G., 2010. Young Investigators: the Project Approach in the Early Years. Teachers College Pr.

Hobson, Sally M., Trundle, Kathy Cabe, Saçkes, Mesut, 2010. Using a planetarium software program to promote conceptual change with young children. Journal of Science Education and Technology 19, 165–176.

Hsu, C.Y., Tsai, C.C., Liang, J.C., 2011. Facilitating preschoolers' scientific knowledge construction via computer games regarding light and shadow: the effect of the prediction-observation-explanation (POE) strategy. Journal of Science Education and Technology 20, 482–493.

Katz, Lilian G., 1994. The project approach. ERIC Digest: EDO-PS-94-6 [Online] http://ericps. crc. uiuc. edU/eece/pubs/digests/l994.

Katz, Lilian G., Chard, Sylvia C., 2000. Engaging Children's Minds: The Project Approach. Praeger Pub Text.

Kirschner, Paul A., Sweller, John, Clark, Richard E., 2006. Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery,

problem-based, experiential and inquiry-based teaching. Educational Psychologist 41, 75–86.

Kohn, Amy S., 1993. Preschoolers' reasoning about density: will it float? Child Development 64, 1637–1650.

Krnel, Dusan, Glazar, Sasa S., Watson, Rod, 2005. The development of the concept of "matter": a cross-age study of how children classify materials. Science Education 87, 621–639.

Lautrey, Jaques, Mazens, Karine, 2004. Is children's naive knowledge consistent? A comparison of the concepts of sound and heat. Learning and Instruction 14, 399–423.

Lewis, Eileen L., Linn, Marcia C., 1994. Heat energy and temperature concepts of adolescents, adults, and experts: implications for curricular improvements. Journal of Research in Science Teaching 31, 657–677.

Mayer, Richard E., 2004. Should there be a three-strike rule against pure discovery learning? the case for guided methods of instruction. American Psychologist 59, 14–19.

Mazens, Karine, Lautrey, Jaques, 2003. Conceptual change in physics: children's naïve representation of sound. Cognitive Development 18, 159–176.

Miner, Jolynne T., 1992. An Early Childhood Study of the Water Cycle. Unpublished master thesis, University of Nevada, Las Vegas, NV.

Moyle, Ray, 1980. Weather. Learning in Science Project Working Paper 21. University of Waikato, Hamilton, NZ.

Nakhleh, Mary B., Samarapungavan, Ala, 1999. Elementary school children's beliefs about matter. Journal of Research in Science Teaching 36, 777–805.

Opfer, John. E., Siegler, Robert S., 2004. Revisiting preschoolers' living things concept: micro genetic analysis of conceptual change in basic biology. Cognitive Psychology 49, 301–332.

Piaget, Jean, 1971. Les Explications Causales. PUF, Paris.

Piaget, Jean, 1972. Child's Conceptions of the World (J. and A. Tomlinson, Trans.). Littlefield Adams, Lanham, MD.

Saçkes, Mesut, Flevares, Lucia, Trundle, Kathy Cabe, 2010. Four- to six-year-old children's conceptions of the mechanism of rainfall. Early Childhood Research Quarterly 25, 536–546.

Sharp, John G., 1996. Children's astronomical beliefs: a preliminary study of Year 6 children in southwest England. International Journal of Science Education 18, 685–712.

Shayer, Michael, Wylam, Hugh, 1981. The development of the concepts of heat and temperature in 10–13 year-olds. Journal of Research in Science Teaching 18, 419–434.

Spelke, Elizabeth Shilin, 1991. Physical knowledge in infancy: reflections on Piaget's theory. In: Carey, Susan, Gelman, Roachel (Eds.), Epigenesis of Mind: Studies in Biology and Cognition. Erlbaum, Hillsdale, NJ.

Tobin, Kenneth, Tippins, Deborah J., Gallard, Alejandro Jose, 1994. Research on instructional strategies for teaching science. In: Gabel, Dorothy L. (Ed.), Handbook of Research on Science Teaching and Learning. Macmillan Publishing Company, New York, pp. 45–93.

Trowbridge, Leslie W., Bybee, Rodger W., 1996. Teaching Secondary School Science: Strategies for Developing Scientific Literacy, sixth ed. Merrill, Englewood Cliffs, NJ.

Trundle, Kathy Cabe, Atwood, Ronald K., Christopher, John E., 2002. Preservice elementary teachers' conceptions of moon phases before and after instruction. Journal of Research in Science Teaching 39, 633–658.

Trundle, Kathy Cabe, Atwood, Ronald K., Christopher, John E., Sackes, Mesut, 2010. The effect of guided inquiry-based instruction on middle school students' understanding of lunar concepts. Research in Science Education 40 (3), 451–478.

Trundle, Kathy Cabe, Saçkes, Mesut, 2012. Science and early education. In: Pianta, R.C., Robert, C., Barnett, W. Steven, Justice, Laura M., Sheridan, Susan M. (Eds.), Handbook of Early Childhood Education. Guilford Press, New York.

Trundle, Kathy Cabe, Saçkes, Mesut, 2013. Science and early education. Students' conceptual understandings enhanced through technology instruction. In: Finson Kevin, D, Pederson John, E. (Eds.), Visual Data and Their Use in Science Education. Information Age Publishing, Charlotte, NC, pp. 189–210.

Trundle, Kathy Cabe, Saçkes, Mesut, Smith, Mandy M., Miller, Heather L., 2012 September. Preschoolers' ideas about day and night and objects in the sky. In: Paper Presented at the Third International Congress on Early Childhood Education, Adana, Turkey.

Trundle, Kathy Cabe, Mollohan, Katherine N., Smith, Mandy MCCormick, 2013. Alike and different: laying the foundation to help young children begin to understand inheritance of traits. Science and Children 50 (6), 52–57.

Vosniadou, Stella, Brewer, William F., 1992. Mental models of the earth: a study of conceptual change in childhood. Cognitive Psychology 24, 535–585.

Vosniadou, Stella, 1994. Capturing and modeling the process of conceptual change. Learning and Instruction 4, 45–69.

White, Richard T., Gunstone, Richard F., 1992. Probing Understanding. Falmer Press, Great Britain.

Za'rour, George I., 1976. Interpretation of natural phenomena by Lebanese school children. Science Education 60, 277–287.

Scientific Controversies

Trevor Pinch, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the research on scientific controversies, largely carried out within the field of science and technology studies (STS). It identifies different strands of research including priority disputes, controversy as a method, controversies to do with the impact of science and technology on society and controversies at the research frontiers of science. It concludes by arguing that within modern STS the boundary between studying controversies within science and how they ramify in the wider society has largely been dissolved.

There is a long history of studying scientific controversies within the sociology of science. Different sorts of controversy and different approaches to analyzing them have emerged over the years. Here, I summarize the main developments.

Priority Disputes

The first type of controversy to gain attention within the sociology of science was the 'priority dispute.' This is a dispute between one or more scientists or groups of scientists who claim 'priority' for a scientific discovery. In other words, there are claims and counterclaims as to who has first made a particular discovery. The reason scientists care so much about priority was first pointed out by Robert Merton (1957). He argued that the main reward in science is recognition, bestowed in prizes and the naming of discoveries. In other sorts of endeavors, for instance, technological invention, rewards take other forms, such as patent rights or money.

Priority disputes in science are frequent occurrences because science as an activity builds upon previous knowledge and techniques that are widely disseminated and shared within the scientific community. This means that scientists, who often work within what Thomas Kuhn (1962) has called a shared paradigm, are solving similar sorts of puzzles and this will lead to disputes as to who first arrived at a particular solution. New instruments and techniques, which enable new discoveries, are also collectively shared. Discoveries are, thus, very much in the air. Often scientists are aware that they are in a race and who their main 'competitors' are. During such races, the norms of sharing data, techniques, and knowledge may break down, and such races can be accompanied by acrimonious accusations of breaches of correct scientific behavior, such as occurred during the race to discover the structure of DNA, where Watson and Crick were accused of stealing Rosalind Franklin's X-ray crystallography results (Sayre, 1975).

Merton's interest in priority disputes stemmed from his claim that science has a particular normative structure or 'institutional ethos' with an accompanying set of rewards and sanctions. Because so much of the reward structure of science is built upon recognition for new discoveries, scientists are particularly concerned to establish priority for their findings. Priority disputes are legion, such as the famous fight between Newton and Leibnitz over who first discovered the calculus, or the more recent fight between the French, Institute Pasteur group, and the American researcher, Robert Gallo, over who first established HIV as the cause of AIDS. In this case there was not only scientific priority at stake, but also the licensing of the lucrative blood test for identifying AIDS. The controversy could only be settled by intervention at the highest political level. The Presidents of the U.S. and France, Ronald Reagan and Jacques Chirac, agreed to share the proceeds from the discovery. This case was marked by additional controversy because of allegations of scientific misconduct raised against Gallo, which led to Congressional and National Institute of Health (NIH) investigations. It was one of the first times in science that a controversy was settled by politicians. The AIDS priority dispute reminds us that the notion of a 'community of scientists' characterized by internal struggles for recognition does not fit science in the latter part of the twentieth century. Scientists in areas such as biotechnology increasingly move between academia and start-up companies, and seek both reputational and financial recognition through stock options and the like (Shapin, 2008).

Thomas Kuhn (1962) raised a fundamental problem for the analysis of priority disputes. A priority dispute is predicated upon a model of science, known as the 'point model' of scientific discovery, which can establish unambiguously *who* discovered *what* and *when*. Asking the question who discovered oxygen, Kuhn showed that the crucial issue is what counts as oxygen. If it is the dephlogisticated air first analyzed by Priestly then the discovery goes to him, but if it is oxygen as understood within the modern meaning of atomic weights, then the discovery must be granted to Lavoisier's later identification. The 'point model' requires discovery to be instantaneous, and for discoveries to be recognized and dated. A rival 'attributional model' of discovery, first developed by Augustin Brannigan (1981), draws attention to the social processes by which scientific discoveries are recognized and 'attributed.' This approach seems to make better sense of the fact that what counts as a discovery can vary over time. In short, it questions the Eureka moment of the point model. Woolgar (1976), in a pioneering analysis of the discovery of pulsars, showed that the date of the discovery varies depending on what stage in the process is taken to be the defining point of the discovery. If the discovery is the first announcement of 'scruff' by Jocelyn Bell, a graduate student in Anthony Hewish's investigative team, when examining a chart recording of signals from the radio telescope, then it will be dated earlier than when it was realized

following extensive investigations that the unambiguous source of this 'scruff' was a distant and novel kind of astrophysical source. This case was particularly controversial because it was alleged by the dissonant Cambridge radio astronomer Fred Hoyle that the Nobel Prize winners for this discovery should have included Jocelyn Bell. Priority disputes can thus touch on the social fabric of science, such as its gender relationships and hierarchical structure.

The point model of discovery is embedded in the reward system of science. For example, most significant rewards accrue to the scientist or group who is 'first past the post.' Interestingly with modern discovery claims posted to the Internet, the exact date and time when a paper is first posted becomes even more crucial in establishing priority.

Controversy as a Method

Controversies as a form of method became important in the sociology of scientific knowledge (SSK) in the 1970s and 1980s. SSK went on to replace the earlier Mertonian sociology of science with its focus upon the rewards and institutions of science. For SSK researchers and for other strands within the emerging field of Science and Technology Studies (STS), such as feminist science studies (Haraway, 1989), it has been important to study moments of contention in science. It is during such moments that the often invisible processes of the working of science become more visible and hence available to analysis.

One metaphor for understanding the methodological importance given to controversies is that of 'punching' a system. Scientists on occasion gain insight into natural systems by punching them. For example, one may learn more about the laws of momentum by bouncing one billiard ball off another than by watching a stationary billiard ball. Similarly, Rutherford famously used scattering experiments in which gold foil was bombarded with alpha particles to uncover the structure of the atom and, in particular, the presence of the nucleus. In sociology, a similar methodology was advocated by Harold Garfinkel, for investigating the taken-for-granted features of social life. His breaching experiments involved breaking or breaching some taken-for-granted social convention, say, in routine greetings. The methodological assumption underpinning the study of controversies is similar, only in this case the breach in the normal social operation of science is produced by the scientists themselves. By studying a scientific controversy, or moments of contestation, one learns something about the underlying dynamics of science and its relations with wider society. For instance, during a controversy the normally hidden social dimensions of science (including its gender biases) may become more explicit. Sites of contestation are places to facilitate the investigation of, for instance, the metaphors, assumptions and political struggles embedded within science (or for that matter, technology and medicine).

Three Strands of Controversy Research

One can identify three broad strands of research on controversies within STS. The first approach, which dates back to the 1960s, was concerned largely with controversies provoked by contemporary science and technology and what was perceived to be its increasingly deleterious impact on the wider society. The controversies often crystallized around particular technologies such as nuclear power, pesticides and agribusiness in general, or technologies used in warfare, such as Agent Orange, or the perceived threat of what was then called 'genetic engineering.' Often groups of concerned citizens and scientists organized to oppose such technologies – for example, Rachel Carson's warnings about the dire effects of pesticides in her book *Silent Spring*. The source of controversy was the perceived negative impact of science and technology on particular groups and it is the study of these political responses that forms the core of the analysis.

The second strand is the study in the 1970s and 1980s of controversies *within* science. This was briefly mentioned above and emerged as part of SSK and its attempts to study how scientific knowledge is 'socially constructed.' By focusing on controversies at the research frontier of science, scholars were able to delineate what was at stake for groups of scientists as they fought over very specific knowledge claims, such as the existence of a new room-temperature form of nuclear fusion known as 'cold fusion,' or whether particular experimental results in physics had been replicated or not (Collins, 1992 [1985]). This approach was also extended into technology by what has been called the study of technoscience (Latour, 1987), and by the Social construction of technology (SCOT) (Bijker et al., 2013[1987]).

The third strand of controversy studies can be thought of as a merger of the earlier two strands (and is discussed at greater length in the final section of this article). Within contemporary STS, many researchers follow scientific controversies wherever they ramify. In other words, they make less of a distinction between controversies within science and those that impact the wider society. Modern scientometric tools sometimes help analysts to follow or map such controversies and the many different participants who may be scattered globally.

Controversies over the Impact of Science on Society

Many of the controversies over science and society that dominate today – such as global warming, sustainable energy, environmental degradation, and biomedical interventions into life – have their roots in earlier debates in the 1960s. The Vietnam War protests, the emergence of the counterculture, and the civil rights movement, formed part of a wider backdrop against which science and technology increasingly came into question. The political protest movements of the 1960s, followed by the environmental movement in the 1970s, led to recognition that science and technology are neither neutral nor necessarily beneficial, and that many developments stemming from modern science and technology, such as nuclear power, the petrochemical industries, and genetic engineering, raise profound and controversial issues for a concerned citizenry.

Dorothy Nelkin, a pioneer in analyzing these types of dispute, identified a range of different sorts of political, economic, and ethical controversies that engaged the public (her research was almost exclusively based in the United States (Nelkin, 1995)). Early research by Nelkin and her colleagues focused mainly on the interest politics of the groups involved

(Nelkin, 1992): how and why they got involved in political action over science and technology; their underlying political values; and how effective they were. Participants' positions in controversies were taken to be consistent with their interests, although these interests were not necessarily determinate in bringing about closure. For instance, the demise of nuclear power in the United States had as much to do with economics as with political protest.

Since scientists themselves often played an active part in these disputes, part of the analysis touched upon how scientists deployed their specialized knowledge for political aims. By and large, this research tradition avoided using the entry of scientists into these disputes as an opportunity to examine the core processes of knowledge construction. In short, the attention was focused upon seeing how scientists became political rather than upon how politics might itself shape scientific knowledge. Political controversies were treated as analytically separable from epistemic controversies and as resolved by distinct processes of closure (Engelhardt and Caplan, 1987). Typically epistemic controversies were thought to be closed through the application of epistemic and methodological standards, while political controversies were closed through the intervention of 'nonscientific factors,' such as economic and political interests.

Controversies at the Research Frontiers of Science

The SSK, which emerged in the late 1970s, was concerned to examine how scientific knowledge itself was socially shaped or constructed. A key tenet of this new sociology of science, as formulated by David Bloor (1991[1976]) in his 'Strong Program,' was that of symmetry. This principle called upon sociologists to use the same explanatory resources to explain both successful and unsuccessful knowledge claims. This was a way to avoid a sociology of error whereby the social would be evoked to explain false knowledge, leaving true knowledge as simply revealed, or generated by some universal epistemological principle or rule of method. Controversies at the research frontiers over knowledge claims were good places to carry through the principle of 'symmetry.' With each side in a dispute claiming to have 'truth on its side,' and disparaging the efforts of the other, and with the very outcome of the controversy still unknown, there was less temptation to carry out a sociology of error. Full-blooded sociological explanations as to how credibility and trust were gained, maintained, and lost could be mounted. Unlike the approach to controversies on the impact of society favored by Nelkin, the very scientific claims made by the participants became subject to analysis.

Bloor and his colleagues in what was known as the 'Edinburgh School' pursued their program mainly through theoretical analysis supported by historical case studies (an exception was the work of Andy Pickering who studied contemporaneous controversies in particle physics). Collins and the 'Bath School' by contrast, developed an empirical method for studying the SSK in contemporaneous cases – a method based primarily upon the study of scientific controversies (Collins, 1992[1985]; Pinch, 1986; Collins and Pinch, 1998). Generalizing from several case studies of controversies, Collins (1981) argued that during controversies scientific findings exhibited 'interpretative flexibility,' with the facts at stake being debated and interpreted in radically different ways by the parties in the controversy. This interpretative flexibility did not last forever. By following a controversy over time researchers could delineate the process of 'closure' by which controversy vanished and consensus emerged. Collins defined the group of scientists involved in a controversy as the 'core set.' Only a very limited set of scientists actively partook in controversies, the rest of the scientific community depended upon the core set for their expert judgment as to what to believe.

Following controversies from their inception to the point of closure, allowed researchers to better understand 'science in the making' (Latour, 1987). Controversies when seen through the sociological lens turned out to be messy things and were very rarely resolved by experiments alone. Collins (1992[1985]) developed the notion of an 'experimenter's regress' to explain the lack of compulsion of experiments during controversies.

Often the losing side in a scientific controversy continues to fight for its position long after the consensus has turned against it. Those who continue to fight the good fight meet increasing disapprobation from their colleagues and may be forced to leave science altogether. 'Life after death' for discredited phenomena such as cold fusion goes on at the margins and often finally passes away only when the protagonists themselves die or retire (Simon, 2002). Sometimes scientists caught up in scientific controversies who find themselves on the losing side enter into an explicit form of scientific 'dissent,' which can involve them in attempts to form new alliances between academic institutions, interest groups, and the public (Martin, 1991, 1998; Delborne, 2008: 514). Opposition *outside* science to particular theories *in* science (e.g., Darwin's theory of evolution, or Einstein's theory of relativity) also presents a continuing source of controversy (Wazeck, 2013).

Powerful actors with vested political interests, such as the tobacco lobby in the United States, have latched onto the idea of scientific controversy as a way of fighting a scientific consensus that is moving against them. By raising issues of doubt and uncertainty over knowledge claims, thus setting the stage for a faux scientific controversy, over, say, the health hazards of smoking or the thesis that global warming is produced by human activities, such actors can resist the prevailing consensus (Oreskes and Conway, 2010).

The uncertain side of science is clearest during moments of controversy and it has been argued that teaching science through controversies is good way to give students a more realistic idea of scientific practice. Most scientists never experience controversies directly, and often it is those further from the research frontier who have a firmer sense of the robust nature of scientific method and the certainty of experimental results. This 'distance lends enchantment' phenomenon (Collins, 1992[1985]) has been expressed by MacKenzie (1990) for technology in terms of a 'certainty trough' whereby people closest to a technology and those furthest away from it both experience the most uncertainty about a technology's working than those at middle distance. Returning to science, after exposure to a controversy, scientists often become more aware of the social side of science and start to read in science studies and even employ ideas drawn from science studies to understand what has happened to them.

The sociological approach to controversies has also been taken up by historians (e.g., Rudwick, 1985). Such studies pose

particular methodological challenges because often the viewpoint of the losing side in a controversy has vanished from history (Dascal, 1998). Shapin and Schaffer's (2011[1985]) study of the dispute between Robert Boyle and Thomas Hobbes over Boyle's air-pump experiments was a landmark in research on scientific controversy, because it showed how the wider political climate, in this case that of Restoration Britain, could shape the outcome of a controversy and at the same time help institutionalize a new way of fact-making – the experiment – in the Royal Society. It also drew attention to the literary and technological dimensions of building factual assent in science. By documenting the witnesses to particular experimental performances a culture of 'virtual witnessing' was born.

The SSK approach to scientific controversy has also been influential in the study of technology. The SCOT framework uses concepts imported from the study of scientific controversy such as 'interpretative flexibility' and 'closure.' A variety of competing meanings are found in technological artifacts and scholars study how 'closure mechanisms' produce a stable meaning of a technology (Pinch and Bijker, 1984; Bijker, 1995).

Another influential SSK approach to the study of scientific and technological controversies has been that developed by Bruno Latour and Michel Callon. Actor-Network Theory, as it has become known, has taken on a life of its own and has gone through several reiterations. The seminal work in the 1980s formed an important strand to the SSK approach toward controversies. Callon's article on a controversy over a new method of harvesting scallops in France argues that the outcome of the controversy cannot be explained by reference to the social realm alone, because the analyst must also take account of the actions of nonhuman actors, such as scallops and ocean currents (Callon, 1986). Subsequently Latour's work on how 'trials of strength' are settled in science and technology has become especially influential. Such struggles, according to Latour (1987), involve aligning sociotechnical resources into so-called 'immutable mobiles' or black boxes: objects that remained fixed when transported along scientific networks and which contain embedded within them sets of social and material relationships.

One of the key ideas in all controversy research carried out within SSK is the notion that science has moments of openness or uncertainty followed by closure. Once a matter of fact is settled or black boxed, it is very hard to reopen the black box (Pinch, 1986; Latour, 1987). Often, these processes of closure are woven into the fabric of scientific practices and one can observe the closure or black boxing processes *in situ*. It is at this point that controversy research meets another very influential tradition within SSK known as Laboratory Studies (Lynch, 1993). For instance, Latour and Woolgar (1986[1979]) in their now classic study of a molecular biology laboratory observed scientific practices close up but also had access to a controversy in the lab. They were able to show how literary inscriptions played a special role as a marker in settling controversies. Controversies could be analyzed in terms of whether certain modalities are added to or subtracted from scientific statements making them more or less factlike. This enables the STS analyst to follow the opening and closure of a controversy and to examine how previously resolved controversies become embedded within taken-for-granted results, theories, and techniques.

A detailed examination of the role of discourse in scientific controversies was carried out by Gilbert and Mulkay (1984). They showed how particular discursive repertoires were used selectively by scientists in order to bolster their own claims or undermine those of their opponents. Subsequently, there has been much work on how a variety of rhetorical, visual, and textual resources operate during controversies (e.g., Lynch and Woolgar, 1990; Myers, 1990; Pinch, 1996). Sometimes the resolution of a controversy is only possible by drawing boundaries around the relevant experts who can play a role in the controversy. Sometimes particular scientific objects cross such boundaries and form a nexus around which a controversy can be resolved. Such 'boundary work' (Gieryn, 1983) and 'boundary objects' (Star and Griesemer, 1989) form another important analytical resource for understanding how controversies develop and are brought to closure.

Scientific Controversy in STS Today

The earlier distinction between controversies in science and those over the social impact of science of technology has largely been dissolved in contemporary STS. Another way of putting this is to say that in contemporary work the sites of contestation chosen for analysis have become more heterogeneous. Part of the reason for this heterogeneity is that increasingly the boundaries around expertise have become much harder to delineate (Collins and Evans, 2007). There is no longer assumed to be a one-to-one mapping between expert authority and expert credentials and this means that lay people and activists can themselves acquire sufficient expertise to engage scientists on their own terms (Wynne, 1989). In his well-known case study, Steven Epstein (1996) showed how AIDS activists, for instance, managed to reconfigure some of the science of clinically controlled trials of AIDS drugs.

In addition, the relationship between analyst and actor in controversy research has increasingly been problematized (Ashmore and Richards, 1996). Reflexivity is sometimes embraced by analysts who even become occasional participants in the controversies they study. The study of controversies in modern technoscience – with its porous boundaries between science, technology, politics, the media, and the citizenry – also calls for the analyst to adopt a broader array of analytical tools. With the growing role played by the media, there is also more attention upon the role of audiences for specific controversies (e.g., Delborne, 2011).

Although the fundamental insights provided by SSK remain influential, such insights are supplemented by an increased understanding of how macropolitical structures, such as the state and the legal system, enable and constrain the outcome of scientific controversies. Within the radical agenda of ANT and its sociology of associations, the very distinction between macro and micro is itself dissolved (Latour, 2005). The tools available to study controversies in contemporary science and technology have also expanded. There are several groups involved in a project initiated by Bruno Latour on 'mapping scientific controversies.' In this project, controversies are defined in the constructivist sense as "every bit of science and

technology which is not yet stabilized, closed or 'black boxed'" (Venturini, 2010). This project on the cartography of controversies is heavily influenced by ANT and has evolved several Web-based scientometric tools for following controversies, although face to face interactions still form an important part of the analysis (Venturini, 2012).

Contemporary research into controversies draws upon the lessons of constructivist SSK in not taking for granted that claims are settled by epistemic or methodological fiat. The movement from openness to closure involves deploying a whole range of resources. Such work also pays close attention to how actors themselves describe the situations they face. But with the move to study sociotechnical controversies in wider society, the sorts of actors dealt with are likely to be much more heterogeneous. The 'core set' idea of a bounded community of scientists who conduct the controversy still works for the study of controversies within the more esoteric districts of science, but analysts these days are more likely to embrace the key ANT idea that controversies ramify endlessly, especially when the wider sociotechnical assemblage is studied.

Controversies involving science and technology are today examined in the courtroom, in the media, in quasigovernmental policy organizations, in the state and its security apparatus, and in citizen action groups. A controversy or hot spot may be sparked by almost anything, such as the testing of a new drug, or a claim made about a new smart phone, or a natural (though possibly human-caused) calamity such as the collapse of the honey bee population (Suryanarayanan and Kleinman, 2013). There has been particular interest in controversies involving health and medicine, such as contentious new disease categories (Murphy, 2006), and in the activities of patient activist groups (Silverman, 2013). The strong feminist tradition in science studies has continued with, for instance, research on male contraceptives (Oudshoorn, 2003). Wherever there are sociotechnical disputes – such as, for instance, in mapping the state (Carroll, 2006; Leuenberger and Schnell, 2010) – one can study the movement between openness and closure as powerful social choices get embedded within material structures and procedures. Controversy research has largely been located in the US and Europe but researchers today increasingly explore controversies in the Global South. Controversy research is a taken-for-granted part of STS, and that is testament to its success as a method.

See also: Experiment in Science and Technology Studies; Laboratory Studies: Historical Perspectives; Norms in Science; Reflexivity in Science and Technology Studies; Science and Technology Studies, Ethnomethodology of; Science and Technology Studies: Experts and Expertise; Science, Sociology of; Scientific Knowledge, Sociology of; Technology, Social Construction of; Truth and Credibility in Science.

Bibliography

Ashmore, M., Richards, E. (Guest Eds.), 1996. The politics of SSK: neutrality, commitment and beyond. Special Issue of Social Studies of Science 26, 219–445.

Bijker, W.E., 1995. Of Bicycles, Bakelites, and Bulbs: Towards a Theory of Sociotechnical Change. MIT Press, Cambridge, MA.

Bijker, W.E., Hughes, T.S., Pinch, T.J. (Eds.), 2013[1987]. The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology, Anniversary ed. MIT Press, Cambridge, MA.

Bloor, D., 1991[1976]. Knowledge and Social Imagery, first ed. University of Chicago Press, Chicago (first ed., Routledge and Kegan Paul, 1976).

Brannigan, A., 1981. The Social Basis of Scientific Discoveries. Cambridge University Press, Cambridge.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieux Bay. In: Law, J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge? Sociological Review Monograph. Routledge & Kegan Paul, London, pp. 196–223.

Carroll, P., 2006. Science, Culture and Modern State Formation. University of California Press, Berkeley.

Collins, H.M. (Guest Ed.), 1981. Knowledge and controversy: studies in modern natural science. Special Issue of Social Studies of Science 11 (1), 1–158.

Collins, H.M., 1992[1985]. Changing Order, second ed. University of Chicago Press, Chicago (first ed., Sage, London, 1985).

Collins, H.M., Evans, R., 2007. Rethinking Expertise. University of Chicago Press, Chicago.

Collins, H.M., Pinch, T.J., 1998. The Golem: What You Should Know about Science. Canto, Cambridge.

Dascal, M., 1998. The study of controversies and the theory and history of science. Science in Context 11, 147–155.

Delborne, J., 2008. Transgenes and transgressions: scientific dissent as heterogeneous practice. Social Studies of Science 38, 509–541.

Delborne, J., 2011. Constructing audiences in scientific controversy. Social Epistemology 25 (1), 67–95.

Engelhardt, T.H., Caplan, A.L., 1987. Scientific Controversies: Case Studies in the Resolution and Closure of Disputes in Science and Technology. Cambridge University Press, Cambridge.

Epstein, S., 1996. Impure Science: AIDS, Activism and the Politics of Knowledge. University of California Press, Berkeley, Los Angeles and London.

Gieryn, T., 1983. Boundary work and the demarcation of science from non-science: strains and interests in professional ideologies of scientists. American Sociological Review 48, 781–795.

Gilbert, N.G., Mulkay, M.K., 1984. Opening Pandora's Box. Cambridge University Press, Cambridge.

Haraway, D., 1989. Primate Visions: Gender, Race and Nature in the World of Modern Science. Routledge, London.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago.

Latour, B., 1987. Science in Action. Harvard University Press, Cambridge, MA.

Latour, B., 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford and New York.

Latour, B., Woolgar, S., 1986[1979]. Laboratory Life, second ed. Princeton University Press, Princeton (first ed., Sage, London and Beverly Hills, 1979).

Leuenberger, C., Schnell, I., 2010. The politics of maps: constructing national territories in Israel. Social Studies of Science 40 (6), 803–842.

Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science. Cambridge University Press, New York.

Lynch, M., Woolgar, S. (Eds.), 1990. Representation in Scientific Practice. MIT Press, Cambridge, MA.

MacKenzie, D., 1990. Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance. MIT Press, Cambridge, MA.

Martin, B., 1991. Scientific Knowledge in Controversy: The Social Dynamics of the Fluoridation Debate. State University of New York Press, Albany.

Martin, B., 1998. Strategies for dissenting scientists. Journal of Scientific Exploration 12 (4), 605–616.

Merton, R.K., 1957. Priorities in scientific discoveries: a chapter in the sociology of science. American Sociological Review 22, 635–659.

Murphy, M., 2006. Sick Building Syndrome and the Problem of Uncertainty. Duke University Press, Durham, NC.

Myers, G., 1990. Writing Biology: Texts and the Social Construction of Scientific Knowledge. University of Wisconsin Press, Madison.

Nelkin, D. (Ed.), 1992. Controversies: Politics of Technical Decisions, third ed. Sage, Newbury Park, CA.

Nelkin, D., 1995. Science controversies: the dynamics of public disputes in the United States. In: Jasanoff, S., Markle, G.E., Petersen, J.C., Pinch, T. (Eds.), Handbook of Science and Technology Studies. Sage, Thousand Oaks, CA, pp. 444–457.

Oreskes, N., Conway, E.M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. Bloomsbury Press, New York, London, New Delhi, Sydney.

Oudshoorn, N., 2003. The Male Pill: A Biography of a Technology in the Making. Duke University Press, Durham, NC.
Pinch, T.J., 1986. Confronting Nature: The Sociology of Solar-Neutino Detection. Reidel, Dordrecht.
Pinch, T.J., 1996. Rhetoric and the cold fusion controversy. In: Krips, H., McGuire, J.E., Melia, T. (Eds.), Science, Reason, and Rhetoric. University of Pittsburgh Press, Pittsburgh, pp. 153–176.
Pinch, T.J., Bijker, W.E., 1984. The social construction of facts and artifacts. Or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science 14, 339–441.
Rudwick, M., 1985. The Great Devonian Controversy: The Shaping of Scientific Knowledge among Gentlemanly Specialists. University of Chicago Press, Chicago.
Sayre, A., 1975. Rosalind Franklin and DNA. Norton, New York.
Shapin, S., 2008. The Scientific Life. University of Chicago Press, Chicago.
Shapin, S., Schaffer, S., 2011[1985]. Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life, second ed. Princeton University Press, Princeton, NJ.
Silverman, C., 2013. Understanding Autism: Parents Doctors and the History of a Disorder. Princeton University Press, Princeton.
Simon, B., 2002. Undead Science: Science Studies and the Afterlife of Cold Fusion. Rutgers University Press, New Brunswick.
Star, S.L., Griesemer, J., 1989. Institutional ecology, "translations" and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939. Social Studies of Science 19, 387–420.
Suryanarayanan, S., Kleinman, D.L., 2013. Be(e)coming experts: the controversy over insecticides in the honey bee colony collapse disorder. Social Studies of Science 43 (2), 215–240.
Venturini, T., 2010. Diving in magma: how to explore controversies with actor-network theory. Public Understanding of Science 19 (3), 258–273.
Venturini, T., 2012. Building on faults: how to represent controversies with digital methods. Public Understanding of Science 21 (7), 796–812.
Wazeck, M., 2013. Marginalization processes in science: the controversy about the theory of relativity in the 1930s. Social Studies of Science 43 (2), 163–190.
Woolgar, S., 1976. Writing an intellectual history of scientific development: the use of discovery accounts. Social Studies of Science 6, 395–422.
Wynne, B., 1989. Sheepfarming after Chernobyl: a case study in communicating scientific information. Environment 31, 10–15, 33–39.

Scientific Disciplines, History of

Rudolf Stichweh, Universität Bielefeld, Bielefeld, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13727–13731, © 2001, Elsevier Ltd.

Abstract

The scientific discipline, as the primary unit of internal differentiation of science, is an invention of nineteenth century society. There exists a long semantic prehistory of *disciplina* as a term for the ordering of knowledge for purposes of instruction in schools and universities. But only the nineteenth century established real disciplinary communication systems. They are based on specialization of scientists, role differentiation in the organizations of science, the emergence of standard forms of scientific publication, and the rise of the research imperative which demands an incessant search for novelties. All these structural changes coalesce in the disciplinary community as a new type of communication system in science. After that the discipline functions as unit of structure formation in the social system of science, in systems of higher education, as a subject domain for teaching and learning in schools, and finally, as a designation of occupational and professional roles. Although the processes of differentiation in science have been going on ever since, the scientific discipline as a basic unit of structure formation is stabilized by these plural roles in different functional contexts of modern society. Finally, the individual discipline is embedded in an internal environment of other disciplines. The continuous mutual observation and interaction of these disciplines is the most important factor in the dynamics of modern science.

The scientific discipline as the primary unit of internal differentiation of science is an invention of nineteenth century society. There exists a long semantic prehistory of *disciplina* as a term for the ordering of knowledge for the purposes of instruction in schools and universities. But only the nineteenth century established real disciplinary communication systems. Since then the discipline has functioned as a unit of structure formation in the social system of science, in systems of higher education, as a subject domain for teaching and learning in schools, and finally as the designation of occupational and professional roles. Although the processes of differentiation in science are going on all the time, the scientific discipline as a basic unit of structure formation is stabilized by these plural roles in different functional contexts in modern society.

Unit Divisions of Knowledge

Disciplina is derived from the Latin *discere* (learning), and it has often been used since late Antiquity and the early Middle Ages as one side of the distinction *disciplina* vs. *doctrina* (Marrou, 1934). Both terms meant ways of ordering knowledge for purposes of teaching and learning. Often they were used synonymously. In other usages *doctrina* is more intellectual and *disciplina* more pedagogical, more focused on methods of inculcating knowledge. A somewhat later development among the church fathers adds to *disciplina* implications such as admonition, correction, and even punishment for mistakes. This concurs with recent interpretations of discipline, especially in the wake of Michel Foucault, making use of the ambiguity of *discipline* as a term always pointing to knowledge and disciplinary power at the same time (cf. Hoskin in Messer-Davidow et al., 1993). Finally, there is the role differentiation of teaching and learning and the distinction *doctrina/disciplina* is obviously correlated with it (Swoboda, 1979).

One can still find the same understandings of *doctrina* and *disciplina* in the literature of the eighteenth century. But what changed since the Renaissance is that these two terms no longer refer to very small particles of knowledge. They point instead to entire systems of knowledge (Ong, 1958). This goes along with the ever more extensive use by early modern Europe of classifications of knowledge and encyclopedic compilations of knowledge in which disciplines function as unit divisions of knowledge. The background to this is the growth of knowledge related to developments such as the invention of printing, the intensified contacts with other world regions, economic growth and its correlates such as mining and building activities. But in these early modern developments there still dominates the *archival function of disciplines*. The discipline is a place where one deposits knowledge after having found it out, but it is not an active system for the production of knowledge.

Disciplines as Communication Systems

A first premise for the rise of disciplines as production and communication systems in science is the specialization of scientists and the role differentiation attendant on it (Stichweh, 1984, 1992). Specialization is first of all an intellectual orientation. It depends on a decision to concentrate on a relatively small field of scientific activity, and, as is the case for any such decision, one needs a social context supporting it, that is, other persons taking the same decision. Such decisions are rare around 1750 when encyclopedic orientations dominated among professional and amateur scientists alike, but they gained in prominence in the last decades of the eighteenth century. Second, specialization as role differentiation points to the educational system, which is almost the only place in which such specialized roles can be institutionalized as occupational roles. From this results a close coupling of the emerging disciplinary structures in science and the role structures of

institutions of higher education. This coupling is realized for the first time in the reformed German universities of the first half of the nineteenth century and afterwards quickly spreads from there to other countries. Third, role differentiation in institutions of higher education depends on conditions of organizational growth and organizational pluralization. There has to be a sufficient number of organizations which must be big enough for having differentiated roles and these organizations must be interrelated in an ongoing continuity of interactions.

The emergence of *communities of specialists* is a further relevant circumstance. In this respect the rise of disciplines is synonymous with the emergence of *scientific communities* theorized about since Thomas Kuhn (Kuhn, 1970). Scientific communities rest on the intensification of interaction, shared expertise, a certain commonality of values, and the orientation of community members towards problem constellations constitutive of the respective discipline. Modern science is not based on the achievements of extraordinary individuals but on the epistemic force of disciplinary communities.

Scientific communities are *communication systems*. In this respect the emergence of the scientific discipline is equivalent to the invention of new communication forms specific of disciplinary communities. First of all one may think here of new forms of *scientific publications*. In the eighteenth century a wide spectrum of publication forms existed; they were not, however, specialized in any way. There were instructional handbooks at the university level, journals of a general scientific nature for a regional public interested in utility, and academy journals aiming at an international public, each covering a wide subject area but with rather limited communicative effects. It was only after 1780 that in France, in Germany, and finally, in England, nationwide journals with a specific orientation on such subjects as chemistry, physics, mineralogy, and philology appeared. In contrast to isolated precursors in previous decades, these journals were able to exist for longer periods exactly because they brought together a *community of authors*. These authors accepted the specialization chosen by the journal; but at the same time they continually modified this specialization by the cumulative effect of their published articles. Thus the status of the scientific publication changed. It now represented the only communicative form by which, at the macrolevel of the system of science – defined originally by national but later by supranational networks – communication complexes specialized along disciplinary lines could be bound together and persist in the long run (Stichweh, 1984, Chap. 6, Bazerman, 1988).

At the same time the scientific publication became a formal principle interfering in every scientific production process. Increasingly restrictive conditions were defined regarding what type of communication was acceptable for publication. These conditions included the requirement of identifying the problem tackled in the article, the sequential presentation of the argument, a description of the methods used, presentation of empirical evidence, restrictions on the complexity of the argument accepted within an individual publication, linkage with earlier communications by other scientists – using citations and other techniques – and the admissibility of presenting speculative thoughts. In a kind of feedback loop, publications, as the ultimate form of scientific communication, exercised pressure on the scientific production process (i.e., on research) and were thereby able to integrate disciplines as social systems.

This reorganization of the *scientific production process* adheres to one new imperative: the *search for novelties*. The history of early modern Europe was already characterized by a slow shift in the accompanying semantics associated with scientific truth, from an imperative to preserve the truth to an interest in the novelty of an invention. The success achieved in organizing traditional knowledge, as well as tendencies towards empirical methods and increased use of scientific instruments, worked toward this end. In this dimension, a further discontinuity can be observed in the genesis of the term *research* in the years after 1790. In early modern times the transition from the preservation to the enlargement of knowledge could only be perceived as a continual process. In contrast, *research* from about 1800 refers to a fundamental, and at any time realizable, questioning of the entire body of knowledge until then considered as true. Competent scientific communication then had to be based on research in this sense. What was communicated might be a small particle of knowledge, as long as it was a new particle of knowledge. Scientific disciplines then became research disciplines based on the incessant production of novelties.

The link between scientific disciplines and organizations of higher education is mediated by two more organizational structures. The first of these are *disciplinary careers*. Specialized scientists as members of disciplinary communities do not need only specialized occupational roles. Additionally there may be a need for careers in terms of these specialized roles. This again is a condition which sharply distinguishes eighteenth from nineteenth century universities. Around 1750 you still find, even in German universities, hierarchical career patterns which implied that there was a hierarchical succession of chairs inside of faculties and a hierarchical sequence of faculties by which a university career was defined as a progression of steps through these hierarchized chairs. One could, for example, rise from a chair in the philosophical faculty to an (intellectually unrelated) chair in the medical faculty. The reorganization of universities since early nineteenth century completely discontinued this pattern. Instead of a succession of chairs in one and the same university, a scientific career meant a progression through positions inside a discipline, which normally demands a career migration between universities. This presupposes intensified interactions and competitive relations among universities which compete for qualified personnel and quickly take up new specializations introduced elsewhere. In Germany such regularized career paths through the national university system were especially to be observed from around 1850.

This pattern is again closely related to *disciplinary curricula*, meaning that one follows one's disciplinary agenda not only in one's research practice and personal career, but furthermore that there exist institutional structures favoring teaching along lines close to current disciplinary core developments. The unity of teaching and research is one famous formula for this, but this formula does not yet prescribe disciplinary curricular structures which would demand that there should be a complete organization of academic studies close to the current intellectual problem situation and systematics of a scientific discipline. Only if this is the case does there arise a *professionalization of a scientific discipline*, which means that

a systematic organization of academic studies prepares for a non-academic occupational role which is close to the knowledge system of the discipline. Besides professionalization there is then the effect that the discipline educates its own future research practitioners in terms of the methods and theories constitutive of the discipline. A discipline doing this is not only closed on the level of the disciplinary communication processes, it is also closed on the level of socialization practices and the attendant recruitment of future practitioners (on the operational closure of modern science see Luhmann, 1990; Stichweh, 1990).

The Modern System of Scientific Disciplines

It is not sufficient to analyze disciplines as individual knowledge producing systems. One has to take into account that the invention of the scientific discipline brings about first a limited number, then many scientific disciplines which interact with one another. Therefore it makes sense to speak of a modern system of scientific disciplines (Parsons, 1977, p. 300ff., Stichweh, 1984) which is one of the truly innovative social structures of the modern world.

First of all, the modern system of scientific disciplines defines an *internal environment* (*milieu interne* in the sense of Claude Bernard) for any scientific activity whatsoever. Whatever goes on in fields such as physics, sociology, or neurophysiology, there exists an internal environment of other scientific disciplines which compete with that discipline, somehow comment on it, and offer ideas, methods, and concepts. There is *normal science* in a Kuhnian sense, always involved with problems to which solutions seem to be at hand in the disciplinary tradition itself; but normal science is always commented upon by a parallel level of interdisciplinary science which arises from the conflicts, provocations and stimulations generated by other disciplines and their intellectual careers.

In this first approximation it is already to be seen that the modern system of scientific disciplines is a very dynamic system in which the dynamism results from the intensification of the interactions between ever more disciplines. Dynamism implies, among other things, ever changing disciplinary boundaries. It is exactly the close coupling of a *cognitively defined discipline* and a *disciplinary community* which motivates this community to try an expansionary strategy in which the discipline attacks and takes over parts of the domain of other disciplines (Westman, 1980, pp. 105–106). This was wholly different in the disciplinary order of early modern Europe, in which a classificatory generation of disciplinary boundaries meant that the attribution of problem domains to disciplines was invariable. If one decided to do some work in another domain, one had to accept that a change over to another discipline would be necessary to do this.

Closely coupled to this internally generated and self-reinforcing dynamics of the modern system of scientific disciplines is the openness of this system to new disciplines. Here again arises a sharp difference to early modern circumstances. In early modern Europe there existed a closed and finite catalogue of scientific disciplines (Hoskin, 1993, p. 274) which was related to a hierarchical order of these disciplines (for example philosophy was a higher form of knowledge than history, and philosophy was in its turn subordinated to faculty studies such as law and theology). In modern society no such limit to the number of disciplines can be valid. New disciplines incessantly arise, some old ones even disappear or become inactive as communication systems. There is no center and no hierarchy to this system of the sciences. Nothing allows us to say that philosophy is more important than natural history or physics more scientific than geology. Of course, there are asymmetries in influence processes between disciplines, but no permanent or stable hierarchy can be derived from this.

The modern system of scientific disciplines is a global system. This makes a relevant difference from the situation of the early nineteenth century, in which the rise of the scientific discipline seemed to go along with a strengthening of national communities of science (Crawford et al., 1993; Stichweh, 1996). This nationalization effect, which may have had to do with a meaningful restriction of communicative space in newly constituted communities, has since proved to be only a temporary phenomenon, and the ongoing dynamics of (sub-) disciplinary differentiation in science seems to be the main reason why national communication contexts are no longer sufficient infrastructures for a rapidly growing number of disciplines and subdisciplines.

The Future of the Scientific Discipline

The preponderance of subdisciplinary differentiation in the late twentieth century is the reason most often cited for the presumed demise of scientific discipline postulated by a number of authors. But one may object to this hypothesis on the ground that a change from disciplinary to subdisciplinary differentiation processes does not at all affect the drivers of internal differentiation in modern science: the relevance of an internal environment as decisive stimulus for scientific variations, the openness of the system to disciplinary innovations, the nonhierarchical structure of the system. Even if one points to an increasing importance of interdisciplinary ventures (and to problem-driven interdisciplinary research) which one should expect as a consequence of the argument on the internal environment of science, this does not change the fact that disciplines and subdisciplines function as the form of consolidating interdisciplinary innovations. And, finally, there are the interrelations with the external environments of science (economic, political, etc.), which in twentieth and twenty-first century society are plural environments based on the principle of functional differentiation. Systems in the external environment of science are dependent on sufficiently stable addresses in science if they want to articulate their needs for inputs from science. This is true for the educational environment of science which has to organize school and higher education curricula in disciplinary or interdisciplinary terms, for role structures as occupational structures in the economic environment of science, and for many other demands for scientific expertise and research knowledge which always must be able to specify the subsystem in science from which the respective expertise may be legitimately expected. These interrelations based on structures of internal differentiation in science which have to be identifiable for outside observers are one of the core

components of modern society which, since the second half of the twentieth century, is often described as *knowledge society*.

See also: History and the Social Sciences; Human Sciences, History of; Knowledge Society, History of; Science: Constructivist Perspectives, History of; Scientific Revolution, History and Sociology of; Social Science and Universities; Universities and Science and Technology: Europe; Universities and Science and Technology: United States.

Bibliography

Bazerman, C., 1988. Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. University of Wisconsin Press, Madison, WI.

Crawford, E., Shinn, T., Sörlin, S., 1993. Denationalizing Science. The Contexts of International Scientific Practice. Kluwer, Dordrecht, The Netherlands.

Hoskin, K.W., 1993. Education and the genesis of disciplinarity: The unexpected reversal. In: Messer-Davidow, E., Shumway, D.R., Sylvan, D.J. (Eds.), Knowledges: Historical and Critical Studies in Disciplinarity. University Press of Virginia, Charlottesville, VA, pp. 271–305.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Luhmann, N., 1990. Die Wissenschaft der Gesellschaft. Suhrkamp, Frankfurt am Main, Germany.

Marrou, H.I., 1934. 'Doctrina' et 'Disciplina' dans la langue des pères de l'église. Archivus Latinitatis Medii Aevi 9, 5–25.

Messer-Davidow, E., Shumway, D.R., Sylvan, D.J. (Eds.), 1993. Knowledges: Historical and Critical Studies in Disciplinarity. University Press of Virginia, Charlottesville, VA.

Ong, W.J., 1958. Ramus, Method, and the Decay of Dialogue: From the Art of Discourse to the Art of Reason. Harvard University Press, Cambridge, MA.

Parsons, T., 1977. Social Systems and the Evolution of Action Theory. Free Press, New York.

Stichweh, R., 1984. Zur Entstehung des Modernen Systems Wissenschaftlicher Disziplinen – Physik in Deutschland 1740–1890. Suhrkamp, Frankfurt am Main, Germany.

Stichweh, R., 1990. Self-organization and autopoiesis in the development of modern science. In: Krohn, W., Küppers, G., Nowotny, H. (Eds.), Selforganization – Portrait of a Scientific Revolution, Sociology of the Sciences, Vol. XIV. Kluwer Academic Publishers, Boston, pp. 195–207.

Stichweh, R., 1992. The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science. Science in Context 5, 3–15.

Stichweh, R., 1996. Science in the system of world society. Social Science Information 35, 327–340.

Swoboda, W.W., 1979. Disciplines and interdisciplinarity: A historical perspective. In: Kockelmans, J.J. (Ed.), Interdisciplinarity and Higher Education. University Park, London, pp. 49–92.

Westman, R.S., 1980. The astronomers's role in the sixteenth century: A preliminary study. History of Science 18, 105–147.

Scientific Discovery, Computational Models of

Ronald R Grau and Peter C-H Cheng, University of Sussex, Falmer, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the fields of artificial intelligence and cognitive science, computational models of scientific discovery have been implemented in computer programs to support the making of scientific discoveries and investigate the nature of the creative processes of discovery. Over the past four decades, computational models and tools have been created to model many of the diverse theory-oriented tasks, and some experimental tasks, which scientists perform. The models span the full spectrum of scientific disciplines. They have not only given us a deeper insight into the richness of scientific activity and the psychology of creative problem solving, but also have been instrumental in making new scientific discoveries in their own right.

Introduction

Scientific discovery is the process by which novel, empirically valid, general, and rational knowledge about phenomena is created. It is, arguably, the pinnacle of human creative endeavors. Scientific discovery involves a diverse range of activities with which humans strive to search for new scientific knowledge, find explanations for phenomena, revise existing knowledge, or integrate it with new findings. Many academic and popular accounts of great discoveries surround the process with mystery, ascribing them to a combination of serendipity and the special talents of geniuses. However, computational models of scientific discovery have been developed that shed some light on the underlying creative processes. In contrast to more domain specific, applied approaches like data mining or scientific computation, computational models of scientific discovery usually operate on a wider scope. The making of discoveries in particular domains has been frequently used to illustrate the applicability of the models, yet the primary aim of their construction has been to create communicable scientific knowledge and advance human insight and understanding into how discovery processes work in general.

Models as a Scientific Instrument

Models are an important instrument of science, as they provide different ways of making a scientific problem tractable. The design, testing, comparison, and revision of models present a substantial part of scientific day-to-day business. Many scientific models for all kinds of fields, applications, and purposes have been developed to date, each supporting a particular aspect of the scientific process. Most notably, models have a representational function with regard to objects, phenomena, data, and theories. As such, they can be representations of observable behavior, for instance, through the use of idealizations or analogies (e.g., the billiard ball model of gases or the shell model of the atom). With respect to theories, scientific models usually present an interpretation of a calculus, which is based on some formal notation, such as mathematical equations.

Computational Models

One major advantage of computational models over other approaches to the study of scientific discovery lies in the requirements imposed by writing a running computer program. Any ambiguities and inconsistencies in the concepts used to describe discovery processes will become apparent when attempting to coherently encode them in a programing language. Computer programs can perform laborious tasks while maintaining high precision and thoroughness in their performance, such as searching large information spaces, carrying out automated operations on data, and performing inferences according to particular algorithms. Computer programs can also record user inputs or experiment results reliably and in a systematic fashion using computational storage repositories such as files or databases, and make these recordings available as input for further study. Given suitably constrained and well-defined tasks that can be encoded within a common programing paradigm, computers will usually execute these tasks faster than humans. Another advantage of computational modeling is the ability to investigate alternative methods or hypothetical situations. Different versions of a system may be constructed embodying, say, competing representations to investigate the difficulty of making the discovery with the alternatives. The same system can be run with different sets of data, for example, to explore whether a discovery could have been made had there been less data, or had different data been available. Many stages are involved in the development of the models, including, formulation of the problem, engineering of appropriate problem representations, selecting and organizing data, design and redesign of the algorithm, actual invocation of the algorithm, and filtering and interpretation of the findings of the system (Noy et al., 2000). Considering the nature and relative importance of these activities in the development of systems provides further insight into the nature of scientific discovery.

This article contains two main parts. In section Computational Models of Scientific Discovery, an overview of the work on computational models of scientific discovery is presented, covering the evolution of the field from early work to recent developments, the tasks addressed, as well as some comments on the scope and common limitations of the models. The section Computational Models and Tools describes the structure and workings of some representative models and tools that have been developed thus far.

Computational Models of Scientific Discovery

Since the early 1970s, work in artificial intelligence (AI) has developed computational models of scientific reasoning, which

resulted in a range of different discovery systems and tools. Some systems have modeled discoveries from the history of science or simulated the behavior of participants solving scientific problems in the psychology laboratory. Other systems attempt to harness computational power to operate on combinatorial spaces of information in search of new knowledge. Some systems have made significant discoveries in chemistry and biology, such as finding new reaction pathways or genetic networks. The success of these models provides some insights into the nature of human cognitive processes in scientific discovery and addresses some interesting issues about the nature of scientific discovery itself.

General Approaches

Computational models of scientific discovery have been broadly categorized into data-driven and theory-driven approaches (Langley et al., 1987), whereas data-driven and knowledge-driven approaches are distinguished in Valdes-Perez (1999). Many systems present a combination of the above categories, in particular if they address more than one discovery task or try to improve the results of one approach with the help of the other. Data-driven approaches focus on analytical tasks and are often used in domains where scientific knowledge is scarce. A common problem addressed by these approaches is the induction of laws from existing data. The related computational models usually implement some form of heuristic search that aims to find patterns in large spaces of information and data. Knowledge and theory-driven approaches make use of methods that can be informed and controlled by integrating generic and domain-specific knowledge. These approaches aim to synthesize new knowledge, enable reasoning, or inform the construction of explanatory models from theories and existing knowledge.

Scope

Knowledge structures and scientific activities constitute two categories for describing the many components that play a role in the theoretical and experimental processes of science (Shrager and Langley, 1990). Computational models of scientific discovery are often illustrated using a scientific problem from a particular domain to facilitate a single or several scientific activities in an operationalized fashion, such as observation, data elicitation, taxonomy formation, law formation, creation of models, theory formation, explanation, prediction, manipulation of experiments, or the evaluation of results. Depending on the nature of the discovery task addressed, the models integrate a varying range of domain-specific and domain-independent knowledge, information, and data for carrying out the related operations.

Many computational models of scientific discovery that have been developed thus far have exclusively addressed theoretical tasks. While much of the early work focused on taxonomy and law formation (Langley et al., 1987), the range of models available today cover a wider range of activities, including modeling, model revision, and explanation (e.g., Stolle and Bradley, 2007; Bridewell and Langley, 2010; see also Langley, 2000 for earlier examples). However, the modeling of experiments has not been completely neglected as models have been built that design experiments, to a limited extent, by using procedures to specify what properties should be manipulated and measured in an experiment and the range of magnitudes over which the properties should be varied (Kulkarni and Simon, 1988; Cheng, 1992). For these systems, the actual experimental results are either provided by the user of the system or generated by a simulated experiment in the software. Discovery systems have also been directly connected to a robot, which manipulates a simple experimental setup so that data collected from the instruments can be fed to the system directly, eliminating any human intervention (Huang and Zytkow, 1997). Nevertheless, few systems have simulated or supported substantial experimental activities, such as observing or creating new phenomena, designing experiments, inventing new experimental apparatus, developing new experimental paradigms, establishing the reliability of experiments, or turning raw data into evidence. This perhaps reflects a fundamental difference between the theoretical and experimental sides of science. While both clearly involve abstract conceptual entities, experimentation is also grounded in the construction and manipulation of physical apparatus, which involves a mixture of sophisticated perceptual abilities and motor skills. Developing models of discovery that include such capabilities would necessarily require other areas within AI beyond problem solving, such as image processing and robotics.

Limitations

Computational approaches to scientific discovery share many advantages and disadvantages that are characteristic for in vitro research (Dunbar, 1995). In particular, they allow a high degree of control and can be implemented at relatively low cost. However, computational models are also constrained by a bounded rationality that is implicit in their codified structure and function. Many systems perform rather simple, time-constrained tasks that are based on idealized models and usually fail to implement realistically the complex social interactions between scientists in their approach and examination of a scientific problem. Further, particular human characteristics and circumstances that may play an important role in discovery processes are currently difficult to operationalize. Computational models of scientific discovery often abstract from aspects like human creativity, shifts in the focus of attention, motivation, choice and change of problem representation, or irrational reasoning, just to name a few. Computational models usually draw on features of the technical environment in which they exist, such as large data storage capacity, as well as a high precision and thoroughness in their operation. These features render irrelevant some of the human limitations that might otherwise influence the choice of particular discovery strategies and related activities when computational models were absent, and that may play their part in leading to conceptual change and novel discoveries. Examples of these are the iterative recording and revision of results, the handling of anomalies, and activities of abstraction, simplification or analogical reasoning that are used by humans in order to cope cognitively with large amounts of information.

Recent advances in the design of collaborative software systems and the emergence of social academic networks have improved the computational means for social interaction and communication in science. A prominent example is presented by the various projects embedded within the Semantic Web initiative, which facilitate the collaborative construction, inspection, and manipulation of ontologies (e.g., Noy et al., 2000). While computational models of scientific discovery have evolved to address a wide range of scientific activities (Dzeroski et al., 2007), the exact workings of social interaction still escape the level of formalization that would be needed for their comprehensive simulation. This poses the question how else computational models of scientific discovery can successfully integrate social behavior to grow beyond their current limitations. One possible approach might involve embedding human scientific activities within computational discovery systems that are both data and knowledge driven and facilitate a wide range of scientific tasks in a collaborative, rather than fully automated fashion. Some recent work in the biological sciences, for example, has taken steps in this direction, away from pure simulation models toward programmable knowledge platforms that support multiple users (Shrager, 2007).

Tasks

The majority of early discovery systems have modeled a single theoretical task. The predominance of such systems has been taken as the basis for a general criticism of computational scientific discovery. The models were typically poor imitations of the diversity of activities in which human scientists are engaged and, perhaps, it is from this variety that scientific creativity arises. Researchers in this area countered such arguments by claiming that the success of single task systems presents a manifestation of the underlying nature of the process of discovery, that it is composed of subprocesses or tasks that are relatively autonomous. More complex activity can be modeled by assembling systems that perform one task into a larger system, with the inputs to a particular component subsystem being the outputs of other systems. Models that perform multiple tasks demonstrate the plausibility of this claim (e.g., Kulkarni and Simon, 1988; Cheng, 1992; Stolle and Bradley, 2007; and Bridewell and Langley, 2010). The organization of knowledge structures and procedures in those systems exploits the hierarchical decomposition of the overall process into tasks and subtasks. This in turn raises the general question about the number and variety of different tasks that constitute scientific discovery and the nature of their interactions. What distinct search spaces are involved and how is information shared among them? Computational models of scientific discovery provide some insight into this issue. At a general level, many models can be characterized in terms of two spaces, one for potential hypotheses and the other a space of instances or sets of data (Simon and Lea, 1974). Scientific discovery is then viewed as the search of each space mutually constrained by the search of the other. Inferring a hypothesis dictates the form of the data needed to test the hypothesis, while the data will determine whether the hypothesis is correct and suggest in what ways it should be amended (Klahr and Dunbar, 1988; Kulkarni and Simon, 1988). This is an image of scientific discovery that places equal importance on theory and experiment, portraying the overall process as a dynamic interaction between both components. This approach is applicable both to disciplines in which individual scientists do the theorizing and experimenting and to disciplines in which these activities are distributed among different individuals or research groups. The search of the theoretical and experimental spaces can be further decomposed into additional subspaces; for example, Cheng (1992) suggests three subspaces for hypotheses, models, and cases under the theory component, and spaces of experimental classes, setups, and tests under the experimental component.

Representation

Scientists usually create representations like notes and drawings in order to plan their research, reason about problems, and record results. Finding an effective representation may be fundamental to the making of discoveries. This issue has been directly addressed by computational models that contrast the efficacy of different representations for modeling the same historical episode (Cheng, 1996). Consistent with work in cognitive science, diagrammatic representations may in some cases be preferable to informationally equivalent propositional representations (e.g., Cheng and Barone, 2007; Klahr and Dunbar, 1988). Studies from the history of science have shown that particular discoveries were made as a progression through different stages of activities that involved different kinds of representations. An example is presented by Nersessian (2002), who analyzed the accounts of James Clerk Maxwell in an attempt to examine the processes involved in the discovery of electromagnetic theory. Among many interesting insights into his experiments and the role analogical reasoning may have played for facilitating conceptual change, it was also shown that Maxwell employed qualitative, diagrammatic representations for conceptualizing the scientific problem before creating the final quantitative formulation. The different stages of the discovery process and the related tasks and representations have also been the subject of other research in computational scientific discovery (e.g., Langley, 2000) and research on creativity and cognitive science (e.g., Helie and Sun, 2010). Although computational models of scientific discovery argue against any special abilities of great scientists beyond the scope of conventional theories of problem solving, the models also suggest that the ability of some scientists to modify or create new representations may be an explanation, at least in part, of why they were the ones to succeed.

Computational Models and Tools

Many models and tools have been developed since the advent of research on computational scientific discovery. Although the developed systems employ various methods and data structures that are common in Computer Science and Artificial Intelligence AI, their particular implementations and applications overlap, so that a structured overview of the different

types of systems that exist is somewhat difficult to create. Moreover, in contrast to early computational models of scientific discovery, which explicitly focused on exploring cognitive models of discovery processes, many of the later developments seem increasingly applied and domain specific in nature. As a consequence, even though some of these tools may make valuable contributions to our understanding of discovery processes, it is sometimes unclear whether they have been developed to represent models of discovery in their own right, or whether they use and extend models of discovery more implicitly to focus on practically addressing a scientific problem in a particular domain. The tools described below present but a small selection of examples from each of the previous decades to illustrate somewhat the breadth of applications for which models have been developed and show how systems have evolved to implement an increasing number of tasks.

Discovering Laws from Data – BACON (1978)

Many computational models of discovery can be conceptualized as performing a recursive search of a space of possible states, or expressions, defined by the representation of the problem. Procedures are used to search the space of legal states by manipulating the expressions and using tests of when the goals or subgoals have been met. To manage the search, which is typically subject to potential combinatorial explosion, heuristics are used to guide the selection of appropriate operators. This is essentially an application of the theory of human problem solving as heuristic search within a symbol processing system (Newell and Simon, 1972). BACON (Langley et al., 1987) finds algebraic formulas as parsimonious descriptions of quantitative data. States in the problem search space of BACON include simple algebraic formulas; such as P/D or P^2/D, where, for instance, P is the period of revolution of planets around the sun and D is their distance from the sun. Tests in BACON attempt to find how closely potential expressions match the given quantitative data. Given quantitative data for the planets of the solar system, one step in BACON's discovery path finds that neither P^2/D nor P/D is constant and that the first expression is monotonically increasing with respect to the second. Given this relation between the expressions, BACON applies its 'increasing' operator to give the product of the terms, that is P^3/D^2. This time, the test of whether the expression is constant, within a given margin of error, is true. $P^3/D^2 =$ constant is one of Kepler's planetary motion laws. For more complex cases with larger numbers of variables, BACON uses discovery heuristics based on notions of symmetry and the conservation of higher order terms to pare down the search space. The heuristics use the underlying regularities within the domain to obviate the need to explore parts of the search space that are structurally similar to previously explored states. Following such an approach, computational models have been developed to perform tasks spanning a full spectrum of theoretical activities including the formation of taxonomies, discovering qualitative and quantitative laws, creation of structural models and the development of process models (Langley et al., 1987; Shrager and Langley, 1990; Cheng, 1992). The range of scientific domains covered is also impressive, ranging from physics and astronomy, to chemistry and metallurgy, to biology, medicine, and genetics.

Simulating Historical Discoveries – KEKADA (1988)

KEKADA (Kulkarni and Simon, 1988) is a system that uses heuristics for discovering chemical pathways, based on a two-space model of problem solving and learning similar to that used in BACON (Newell and Simon, 1972; Simon and Lea, 1974). The system simulates the discovery of the urea synthesis cycle, which was made by the biochemists Hans Adolf Krebs and Kurt Henseleit in 1932. KEKADA searches both a space of rules, which contains hypotheses about the problem, and a space of instances, which represents possible experiments that can be carried out for testing the hypotheses. The search is configured by heuristics that are based on the particular expectations associated with the outcome of experiments, and a measure of confidence with respect to the rules applied. The system operates in a cyclic fashion in that experiments are chosen from the instance space, whose outcome changes the state of both spaces, after which another experiment is selected, and so on. KEKADA identifies any experimental results that do not comply with previous expectations as anomalies, and will attempt to find an explanation for such results. Although the simulation carried out by KEKADA was considered largely identical to the original account of the discovery, its success has also been questioned, for example, because the heuristics encoded in the system may have contained knowledge that was not available to Krebs and his colleagues at the time. Nevertheless, KEKADA presented one of the first computational systems to address multiple discovery tasks, including hypothesis formation and experimentation.

Using Literature as Data for New Discoveries – ARROWSMITH (1990)

ARROWSMITH (Swanson, 1990) is a data-driven discovery tool that searches medical literature for intersecting information based on the joint use of keywords or mutual citations. In general, the system attempts to find connections between diseases, their causes and possible treatments and recombines any factual knowledge found in the documents to infer new knowledge. For example, the system has been shown to combine the facts that (1) magnesium may inhibit spreading depression in the cortex and (2) migraine attacks may be a symptom of depression, to infer that magnesium could act as a relief for migraine. With respect to this example, clinical trials have subsequently found supporting evidence that a deficiency of magnesium may indeed be the cause for migraine.

ARROWSMITH uses some linguistic heuristics to streamline connections found and proposes suitable indexing of documents to improve search and prevent a combinatorial explosion of relations found. The system is still being used in the medical research community and has been extended further to improve search results. For example, Torvik and Smalheiser (2007) describe the development of a quantitative regression model that uses additional criteria

to establish and rank the probable relevance of any connections found by the tool.

Finding New Chemical Reaction Pathways – MECHEM (1992)

MECHEM (Valdes-Perez, 1992) is a theory-driven system for discovering chemical reaction pathways. Given empirical data on initial materials and knowledge about the products of chemical reactions, the system will infer possible reaction pathways for these materials, that is sequences of individual chemical reactions that explain how the reaction products come about. MECHEM applies heuristics that are based on existing knowledge about chemical reactions in order to constrain the number, plausibility, and likelihood of possible reaction pathways. For example, the system will favor those results that feature the fewest number of steps. However, the hypotheses are "not drawn from any stored catalog of common reactions; rather, they are generated from basic principles using algorithms minimally slanted toward particular solutions" (Valdes-Perez, 1999). In the context of modeling historical discoveries, this suggests that the knowledge base used by MECHEM can be adjusted at a relatively fine granularity to represent more realistically the bodies of knowledge that were available to scientists at some point in time.

Modeling Dynamic Systems – LAGRAMGE (1997), PRET (2007), SC-IPM (2010)

Since the emergence of BACON (Langley et al., 1987), a number of equation discovery systems have been developed that attempt to find mathematical models of data. In contrast to earlier systems, more recent developments increasingly use existing domain knowledge to improve the results of their inference mechanisms. LAGRAMGE (Todorovski and Dzeroski, 1997) is an example of an earlier system that has targeted dynamic systems and is able to specify models using algebraic as well as differential equations, using context-free grammars to define its search space. SC-IPM (Bridewell and Langley, 2010) is based on this approach and presents a system for inducing quantitative process models from time-series data. Given a specification of entities, processes, and observations that describe a system's dynamics, the program will generate explanations and predictions over this data. SC-IPM uses regression methods to operate on the data and domain knowledge to constrain its heuristic search for plausible models.

PRET (Stolle and Bradley, 2007) is a tool for automatic system identification that constructs models for given dynamic systems in the form of ordinary differential equations. Although originally designed for engineering tasks, the system employs a variety of AI methods to generate and test candidate models, including qualitative reasoning, geometric reasoning, backward chaining, numerical simulation, and others. PRET takes as input observations about the behavior of a given system, and can also take into account user-defined hypotheses and task-specific constraints. Based on its internal methods, PRET can make inferences about potential models on multiple abstraction levels, in that it applies those methods first that involve the lowest computational cost. For example, the system will attempt to rule out a potential model using qualitative reasoning first, before initiating a (computationally expensive) quantitative parameter estimation. Eventually, PRET will create a formal mathematical model of the dynamics of the system.

Equation discovery systems have been successfully applied to a range of theoretical and real-world problems in domains such as physics, robotics, biology, or climatology. Despite their representational limitations with respect to noncontinuous aspects of processes, these models will perform important functions in future discovery systems.

Conclusions and Future Directions

Given the extent of development work necessary to build a discovery system, it seems appropriate to attribute discoveries as much to the developer as to the system itself, although without the system many of the novel discoveries would not have been possible. This does not imply that machine discovery is impossible, but that care must be taken in delimiting the capabilities of discovery systems. The ability of systems like KEKADA (Kulkarni and Simon, 1988) to change their goals to investigate any surprising phenomenon they discover suggests that systems can be developed that would filter and interpret the output of existing systems, by constraining the search of the space defined by the outputs of those systems using metrics based on notions of novelty. Developing such a system, or other systems that find problems or that select appropriate representations, will require the system to possess substantially more extensive knowledge of the target domain. Although there have been vast advances in automated approaches such as inductive process modeling (Bridewell and Langley, 2010) and system identification (Stolle and Bradley, 2007), comprehensive knowledge-based systems are costly and time consuming to build. It therefore appears that the future of discovery systems may be in collaborative support systems for domain scientists rather than fully autonomous systems (Valdes-Perez, 1999; Shrager, 2007; Grau, 2009). These systems use multiple tools and representations to exploit the respective strengths of human scientists and the computational power of the models to compensate for each other's limitations.

Emerging fields such as Computational Biology or Astroinformatics have led to the development of domain-specific tools that implicitly build on computational models of scientific discovery, yet focus mostly on formalisms developed by Computer Science, rather than AI to make discoveries and solve scientific problems. While the handling of massive amounts of data is considered an important task in these domains, it has also been suggested that (1) increasing effort should be made to integrate more of the existing knowledge about models of discovery to improve these approaches, (2) the creation of models that generalize over a particular domain or scientific problem should not be neglected in order to re-enforce the creation of knowledge that can be communicated across disciplines, and (3) research activities concerning scientific computation should be balanced with those that examine the nature of discovery processes, in order to prevent the creation of a research bias

toward data-rich domains, and to continue work on equally important scientific problems where less data are available (Dzeroski et al., 2007).

Computational models of scientific discovery have evolved to handle multiple tasks and can support discovery systems that attempt to address increasingly complex problem domains. Some of the most recent research in AI, Discovery Informatics, and Cognitive Science addresses important topics such as the exploration of ill-structured knowledge domains, the modeling of structural and temporal aspects of processes, the development of multiscale models, or the integration and representation of heterogeneous knowledge (e.g., Grau, 2009). However, further effort is also needed to develop computational models for the support of experimentation, such as systems that design experimental setups or invent new apparatus for carrying out experiments.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Deductive Reasoning Systems; Intelligence: History of the Concept; Knowledge Representation; Problem Solving and Reasoning, Psychology of.

Bibliography

Bridewell, W., Langley, P., 2010. Two kinds of knowledge in scientific discovery. Topics in Cognitive Science 2, 36–52.

Cheng, P.C.-H., 1992. Approaches, models and issues in computational scientific discovery. In: Keane, M.T., Gilhooly, K. (Eds.), Advances in the Psychology of Thinking. Harvester-Wheatsheaf, Hemel Hempstead, pp. 203–236.

Cheng, P.C.-H., 1996. Scientific discovery with law encoding diagrams. Creativity Research Journal 9 (2 and 3), 145–162.

Cheng, P.C.-H., Barone, R., 2007. Representing complex problems: a representational epistemic approach. In: Jonassen, D.H. (Ed.), Learning to Solve Complex Scientific Problems. Lawrence Erlbaum Associates, Mahmah, NJ, pp. 97–130.

Dunbar, K., 1995. How scientists really reason: scientific reasoning in real-world laboratories. In: Sternberg, R.J., Davidson, J. (Eds.), Mechanisms of Insight. MIT press, Cambridge, MA, pp. 365–395.

Dzeroski, S., Langley, P., Todorovski, L., 2007. Computational discovery of scientific knowledge. In: Dzeroski, S., Todorovski, L. (Eds.), Computational Discovery of Scientific Knowledge: Introduction, Techniques, and Applications in Environmental and Life Sciences. Springer, Berlin, pp. 17–43. LNAI 4660.

Grau, R.R., 2009. The Acquisition and Representation of Knowledge about Complex Multi-dynamic Processes. Doctoral thesis, University of Sussex, School of Informatics.

Helie, S., Sun, R., 2010. Incubation, insight, and creative problem solving: a unified theory and a connectionist model. Psychological Review 117 (3), 994–1024.

Huang, K.-M., Zytkow, J., 1997. Discovering empirical equations from robot-collected data. In: Ras, Z., Skowron, A. (Eds.), Foundations of Intelligent Systems. Springer, Berlin.

Klahr, D., Dunbar, K., 1988. Dual space search during scientific reasoning. Cognitive Science 12, 1–48.

Kulkarni, D., Simon, H.A., 1988. The processes of scientific discovery: the strategy of experimentation. Cognitive Science 12, 139–175.

Langley, P., 2000. The computational support of scientific discovery. International Journal of Human-Computer Studies 53, 393–410.

Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M., 1987. Scientific Discovery: Computation Explorations of the Creative Processes. MIT Press, Cambridge, MA.

Nersessian, N.J., 2002. Maxwell and the method of physical analogy: model-based reasoning, generic abstraction, and conceptual change. In: Malamet, D. (Ed.), Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics. Open Court, Lasalle, IL, pp. 129–165.

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ.

Noy, N.F., Fergerson, R.W., Musen, M.A., 2000. The knowledge model of Protégé-2000: combining interoperability and flexibility. Second International Conference on Knowledge Engineering and Knowledge Management, France.

Shrager, J., 2007. The evolution of BioBike: community adaptation of a biocomputing platform. Studies in History and Philosophy of Science 38 (4), 642–656.

Shrager, J., Langley, P. (Eds.), 1990. Computational Models of Scientific Discovery and Theory Formation. Morgan Kaufmann, San Mateo, CA.

Simon, H.A., Lea, G., 1974. Problem solving and rule induction: a unified view. In: Gregg, L.W. (Ed.), Knowledge and Cognition. Lawrence Erlbaum, Potomac, MD, pp. 105–127.

Stolle, R., Bradley, E., 2007. Communicable knowledge in automated system identification. In: Dzeroski, S., Todorovski, L. (Eds.), Computational Discovery of Scientific Knowledge: Introduction, Techniques, and Applications in Environmental and Life Sciences. Springer, Berlin, pp. 17–43. LNAI 4660.

Swanson, D.R., 1990. Medical Literature as a Potential Source of New Knowledge. PubMedLibrary. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC225324/ (last accessed 09.05.2012).

Todorovski, L., Dzeroski, S., 1997. Declarative bias in equation discovery. In: Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA, pp. 376–384.

Torvik, V.I., Smalheiser, N.R., 2007. A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics 23 (13), 1658–1665.

Valdes-Perez, R.E., 1992. Theory-driven discovery of reaction pathways in the MECHEM system. Proceedings of National Conference on Artificial Intelligence.

Valdes-Perez, R.E., 1999. Discovery tools for science applications. Communications of the ACM 44 (11), 37–41.

Scientific Evidence: Legal Aspects

Joseph Sanders, University of Houston Law Center, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Legal consumption of scientific evidence has grown rapidly in the last few decades. Legislatures, administrative agencies, and courts turn to science to provide factual justification for laws, regulations, and decisions. The law–science relationship is characterized by tensions resulting from both cultural differences and the questions law puts to science. All legal systems control the way in which science is introduced into the legal process, but the way in which control is exercised is shaped by whether the legal system is adversarial or inquisitorial. In the process, law helps to define 'science' and the scientific enterprise. In the modern state, law and science have grown increasingly interdependent. By invoking science, law is able to justify difficult policy decisions. In turn, legal decisions contain an implicit scientific epistemology, which helps to maintain science's dominant position in Western society.

Introduction

Expertise, scientific and otherwise, has been part of the legal landscape for centuries (Hand, 1901; Landsman, 1995). In the last century, the role of scientific evidence in the law has expanded rapidly both in regulatory settings and in litigation. Statutes and treaties routinely require agencies to provide scientific justifications for regulatory decisions within and among nations (Peel, 2010). National and international policy disputes are increasingly fought out within the discourse of science (Cranor, 2006). Courts are another large consumer of science. In both criminal and civil cases, parties believe scientific testimony will make their case stronger. As science's role has grown, so has interest in the relationship between law and science. This article reviews the nature of the science–law relationship.

What Law Wants from Science

In both the administrative and the courtroom contexts, law is most frequently interested in acquiring scientific answers to practical questions of causation (e.g., which cancers are caused by exposure to benzene or other toxic agents, how greenhouse gases affect climate change, and how agricultural subsidies affect fair trade) and measurement (e.g., what is the alcohol level in a driver's blood and what are the emission levels at a power plant). These questions lead to a second set of questions about whether specific fields and techniques (e.g., epidemiology, a breathalyzer, and market surveys) are capable of producing reliable results. Theoretical questions per se, are generally of secondary interest, useful insofar as they help courts and regulators to choose among conflicting measurements, extrapolations, or causal assertions.

Tensions in the Law–Science Relationship

The law–science relationship is made more difficult by a number of inherent tensions (Haack, 2009). The tensions arise from two different sources: difference in scientific and legal cultures and differences in the causal questions they typically address.

Culture

The core objectives of legal systems are different from those of science. The business of science is open-ended inquiry while the business of law is to enshrine policy in rules and then to use those rules to resolve disputes.

The nature of open-ended scientific inquiry is reflected in Robert Merton's four norms of scientific inquiry: universalism, organized skepticism, communism, and disinterestedness (Merton, 1973). Two norms – communism (the collective ownership of data and knowledge), and disinterestedness (lack of desire to arrive at any particular outcome in searching for the truth) – are particularly relevant here. One does not have to believe that these norms – perhaps better described as epistemological ideals – are always followed in scientific inquiry to appreciate that they may be more or less valued in different legal settings. The scientific arena, like the legal arena, is often contested ground and complete individual disinterestedness is rare in either context. However, because legal actors do not see inquiry as their primary business, they do not share with those inculcated in scientific values a set of professional ethics supporting this epistemological posture.

It is common to divide modern legal systems into two types: inquisitorial and adversarial (Langbein, 1985; Van Kampen, 1998; Sklansky, 2009). Because all legal systems contain both adversarial and inquisitorial components, it is best to think of these as ideal types (Weber, 1949: p. 90).

Adversarial systems are especially likely to encourage a set of normative beliefs that promote party control and client advocacy over other values. The morality of adversarial legal systems is a matter of considerable debate (Luban, 2007; Markovits, 2008; Wendel, 2010). What is not debatable, however, is that adversarial systems do not view unbiased inquiry embodied in the norms of communism and disinterestedness as a cardinal legal virtue.

The impact of legal systems is not solely a matter of normative values. Much turns on how these values shape the structures and procedures legal systems put in place to affect the way law receives scientific information. Here, it is unwise to view any legal system as a single, homogenous regime. One should distinguish the use of scientific expertise in the administrative context from the use of science in the adjudication of disputes. Administrative agencies in both inquisitorial and adversarial legal systems routinely employ experts presumed to be capable of monitoring and assessing the science presented by interested parties and agencies may enjoy the luxury of postponing decisions when reliable evidence is unavailable. As such, administrative procedures come closer to approximating the organization of scientific inquiry.

Even here, however, one observes differences across legal cultures. Jasanoff (1991) contrasts the relatively closed, consensual, and nonlitigious British regulatory approach with the open, adversarial, and adjudicatory style found in the United States.

Fewer ameliorative structures are available in the context of litigation than in the administrative realm. Absent the use of their own court-appointed experts, decision makers have limited resources with which to assess the merits of expert arguments. Much, therefore, turns on how experts are chosen and how their knowledge is brought into the dispute. In inquisitorial legal systems, e.g., Belgium, Egypt, France, Germany, and Japan, experts are chosen by the judge trying the case. The judge also controls the expert's input (Jurs, 2012; Shaham, 2010). Parties may be given the opportunity to object to a particular expert, question the expert about the opinion rendered, or hire their own expert to rebut the court-appointed expert; but it is very difficult to successfully challenge the admissibility or even the correctness of an expert's report (Langbein, 1985; Jones, 1994).

The adversarial legal system that exists in the United States eschews these and other structural devices in favor of a set of arrangements Oscar Chase (2005) terms American procedural exceptionalism. Among other things, the set of structures in the American system permits the parties to choose, pay, and present experts and their testimony to the fact finder, which often is a civil or criminal jury. These arrangements produce selection biases (experts outside the mainstream of the opinion of their field are chosen because they are more willing than their colleagues to provide an opinion based on less evidence), and often produce witness bias insofar as party experts, when offering their opinion, eschew nuances that they might otherwise have expressed (Sanders, 2007). Moreover, party examination and cross-examination of witnesses frequently creates the perception that there is far less agreement on scientific questions than in fact exists (Angell, 1996; Haack, 2009).

Legal systems may try to ameliorate the effects of these structures in various ways, e.g., by establishing admissibility criteria (discussed below), but these devices are not a full substitute for changing the structures themselves in significant ways (Bernstein, 2008). The reluctance of legal systems to change these structures speaks to the fact that differences in legal cultures are related to larger cross-cultural differences. For example, open, adversarial approaches reflect a coordinate organization – to be distinguished from a more hierarchical organization – of justice (Damaska, 1986), which in turn is more pronounced in societies that are more individualistic and lower on measures of power distance (Hofstede, 1980; Kagan, 2001).

Scientific vs Legal Inquiry

Cultural differences are not the only source of tension in the law–science relationship. Two additional, and perhaps more intractable, sources relate to the nature of inquiry in scientific vs legal culture, specifically differences related to *timing* and differences between the *types of causal questions* normally addressed by science and the causal questions often at stake in legal proceedings. Again these differences play out differently in administrative rule making and courtroom settings.

A key component of much scientific inquiry is its long-term view, its lack of concern with immediate results. However, when law uses science, time is almost always relevant. In administrative settings, a policy must be set; and in litigation, a dispute must be settled. This produces pressures to ask experts to offer their 'best guess' and to draw a conclusion with less warrant than would be called for if one could simply wait for more evidence.

The lack of fully developed scientific knowledge is not of equal significance in the administrative and litigation arenas, because they do not require the same level of proof before reaching a conclusion. In administrative settings, it may be sufficient to know, for example, that a substance poses a significant potential risk. In a courtroom, in order to prevail, the plaintiff often will be held to a higher burden of persuasion. Scientific evidence that is sufficient in the regulatory context may be considered insufficient in court. For example, animal studies showing a relationship between saccharin consumption and cancer may be sufficient to cause an agency to impose regulations on human exposure to the sweetener, but would be insufficient to permit a plaintiff to prevail on a claim that saccharin at normal dose levels causes cancer in humans. Therefore, in litigation the question frequently arises as to whether there is some minimal warrant below which an expert should be prohibited from offering any opinion. One may even ask whether there are whole areas of alleged expertise, such as forensic bite-mark evidence and handwriting analysis, currently routinely admitted into evidence, that should be excluded absent better support for the claims of experts who testify in these areas (Risinger et al., 1989; Beecher-Monas, 2009; National Research Council, 2009: Chapter 5).

The type of causal question addressed in legal settings presents still another difficulty for the law–science enterprise. As numerous scholars have noted, science tends to focus on the general law rather than on the particular case. As a result, science usually seeks to uncover the effects of causes. Law, too, is frequently interested in such questions. What are the effects of exposure to asbestos? What are the chemical attributes of substances that cause human injury? What threshold should be established for daily exposure to benzene? At trial, however, courts often require testimony about the particular case. When this is the case, courts are interested not in the effects of causes, but rather in the causes of effects (Dawid, 2011).

Some causes-of-effects questions can be answered with a significant degree of certainty, e.g., the probability that DNA

left at a crime scene matches a defendant's DNA. But most questions of this type cannot be answered with anything like this degree of certainty (Faigman, 2010). Scientists may be able to state whether benzene causes various types of leukemia, but they have far less to say about whether a particular plaintiff's leukemia was caused by benzene. This problem is particularly acute with respect to many types of forensic evidence in criminal cases where there is little or no reliable empirical work underlying the opinions of forensic experts (Cole, 2001; National Research Council, 2009). Once again, in the face of this difficulty, courts typically lower the threshold of justification required for an expert to testify (Sanders, 2010).

Legal Control of Scientific Evidence

The way in which legal systems control expert inputs reflects both the legal culture and the law's view of science. In litigation, the fundamental distinction remains between systems in which judges pick experts, and those in which the parties do so. Judges have networks they can turn to in choosing an expert. In some jurisdictions, they are assisted in this task through formal registries of qualified individuals. This is the case, for example, in France, Germany, and Japan (Jurs, 2012; Taylor, 1996).

Judicial appointment of experts through a set of registries makes several assumptions about the nature of science and the proper relationship between law and science. A system that permits the parties to select experts with unfettered discretion rests on a very different set of assumptions. The history of admissibility decisions in US courts provides a useful way to explore these assumptions.

Control of scientific testimony at trial is much more difficult when parties are permitted to select experts. Precisely, because the knowledge possessed by the scientist is beyond the ken of the court, the judge (or the jury) often is not in a good position to determine whether what the expert is offering is or is not helpful. In a 1923 decision, *Frye v. the United States*, a federal court offered one solution to this problem: defer to the judgment of the community of specialists who are the guardians of each area of knowledge. Under the *Frye* rule, experts may testify only if the subject of their testimony is sufficiently established to have reached 'general acceptance' in the relevant scientific community. However, the *Frye* test was applied almost exclusively to criminal cases and even there the test was applied only to novel expert evidence that had not previously passed muster. This laissez-faire attitude prevailed until relatively recently. The most noteworthy example of change is in the United States, although a similar trend has occurred in Canada (Vidmar, 1999). In the 1980s, judicial control of scientific expert testimony at trial became a high-profile issue in the United States with the rise of the asbestos litigation and other mass torts, i.e., torts involving thousands of plaintiffs, which in turn increased the concern that experts chosen by the parties produce less reliable testimony (*see* Expert Testimony).

In 1993, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, and subsequent opinions, the US Supreme Court moved the federal courts away from the *Frye* rule, a move followed by most state jurisdictions as well. In its place, courts have moved toward directly examining the structure of the expert's argument and the empirical evidence available to support the various steps in the argument. When the 'analytical gap' between the evidence and the argument becomes too great, the expert testimony is excluded (*General Electric Co. v. Joiner*, 1997; Faigman et al., 2013: Chapter 1).

Legal Understanding of the Scientific Enterprise

Liberal admissibility rules reflect an adversarial legal system that institutionalizes mistrust of all claims to superior authority. They are consistent with the values of a low-power distance culture, i.e., a culture that affords relatively less legitimacy to elite authoritative opinion and, therefore, is skeptical that scientists are entitled to a privileged status. They are also consistent with the view that science is strongly influenced by other forces in society. These are perspectives on the scientific enterprise that are associated with those who adopt social constructionist views of science, i.e., a view that the social organization of science plays a large role in the production and shape of 'scientific knowledge' (Haack, 2003; Zammito, 2004).

The new admissibility filters in the United States and elsewhere clearly reject a social constructionist test, if by this one means the belief that scientific conclusions are solely or largely the result of social processes within the scientific community. At their core, these approaches require of judges that they become sufficiently knowledgeable about scientific methods so that they can fairly assess the validity of evidence offered at trial. This requirement that scientific testimony must pass methodological muster reflects an objectivist view of science.

This does not mean that the legal system necessarily has a utopian vision of science. Factors that courts consider in assessing admissibility include whether the theory or technique underlying the expert's testimony is falsifiable and has been tested, the error rate of any techniques employed, whether the expert's research was developed independent of litigation, whether the testifying expert exercised the same level of intellectual rigor that is routinely practiced by experts in the relevant field, whether the subject matter of the testimony had been subjected to peer review and publication in refereed journals, and, in a partial retention of the *Frye* test, whether the theory or technique had achieved general acceptance in the relevant scientific community. A number of these factors reflect the skeptical view that scientists frequently fail to live up to Merton's norms and that they are strongly influenced by other considerations that are both internal and external to their work (Caudill and LaRue, 2006). Even after *Daubert*, the legal system views expertise with considerable skepticism.

In their use of scientific evidence, both courts and administrative agencies seem to distinguish the process of science from its products. They accept the insight that the process of doing science is a social enterprise and is subject to the buffeting, often distorting winds of social, political, economic, and legal influences. Nevertheless, courts, agencies, and legislatures assume that the products of science may state a truth or an approximate truth about the world. The legal system's view of science adopts a version of what Cole (1992) calls a realist–constructivist position, i.e., science is socially constructed both in the laboratory and in the wider community, but the construction is constrained by input from the empirical world.

The focus on testing of hypotheses is a search for some assurance that the expert has given the empirical world a reasonable opportunity to influence and constrain the expert's conclusions. Ultimately, a prevailing legal view of science holds that there are a set of (social) practices – often given the shorthand name 'the scientific method' –that increase the likelihood someone will make positive contributions to knowledge; a set of practices that scientists themselves frequently point to as the sources of past scientific success (Goldman, 1999). It is worth noting that in the most recent US Supreme Court pronouncement in this area, *Kumho Tire Co. v. Carmichael* (1999) declares that *Daubert* applies to all expert testimony. Admissibility turns on 'reliability,' not on whether something is or is not 'scientific' (Haack, 2010).

There is a large dose of pragmatism in all of this, of course, and admissibility rules have been cited as an example of "the common law's genius for muddling through on the basis of experience rather than logic" (Jasanoff, 1995: p. 63). Not surprisingly, some have criticized the courts for failing to adopt a philosophically coherent admissibility rule. The court's admissibility rulings and the administrative agency use of science to assist policy making do seem to have proceeded in happy obliviousness to the 'science wars' that raged for the last half of the twentieth century between the defenders of a more traditional, positivist view of science and those critics who emphasize its historical, political, social, and rhetorical aspects (Fleck, 1979/1935; Kuhn, 1962; Leplin, 1997; Latour, 1999). However, like many wars, science wars seem to rumble along without a final resolution of the issues that gave rise to the conflict in the first instance (Labinger and Collins, 2001; Haack, 2003). In light of this standoff, legal pragmatism is, perhaps, the wisest course.

The rejection of fully relativist or fully idealist views of science does not mean all legal actors hold identical views. It would be valuable to map the beliefs of legal actors on central disputes in the science wars and how these beliefs impact their use of science. For example, if, as seems likely, plaintiff personal injury lawyers in the United States hold a more relativist view of science than do their defense counterparts, how does this affect their selection and preparation of experts and, in turn, how are these experts received in courts?

Law–Science Interdependence

Law's approach to science should not be understood in terms of the law's beliefs about science alone but rather in terms of the law–science interaction. There are several dimensions to this relationship. First, although modern legal systems may recognize that scientists are influenced by the social world around them, a modern legal system that would permit radical deconstruction which in turn would undermine science's claim to special status is difficult to imagine (Fuchs and Ward, 1994). The modern state is increasingly dependent on science as a source of legitimacy. By turning to science for solutions to complex environmental and safety issues, legislatures are able to avoid making difficult political choices while giving the appearance of placing decision making in the hands of apparently neutral experts who are held in very high esteem relative to other elites (Lawler, 1996). The advantages of this approach are so great that agencies at times engage in a 'science charade' in which they wrap their decision in the language of science even when there is very little research supporting a regulation (Wagner and Steinzor, 2006).

Second, and related to the first observation, the law–science interaction is one of the important ways in which the state helps to produce and maintain the dominant position of science in Western society. In both its administrative and judicial actions, law defines what is and what is not scientific knowledge, and thereby sometimes assists and sometimes impedes science in the important boundary maintenance work of excluding many from the community of 'scientists.' Until quite recently, for example, law single handedly helped many areas of forensic science to maintain the mantle of 'science' in the face of substantial evidence that many claims had little empirical support (Mnookin et al., 2011).

Even when they admit weakly supported testimony, however, legal decisions maintain an implicit epistemology that reinforces a realist view of scientific knowledge. When courts and other institutions of the state reject a relativist view of science that argues the empirical world has little, if any, influence on what is accepted as true by the scientific community, they help to marginalize anyone who adopts this position. There can be little doubt that the view is given little or no attention in law.

Third, as discussed above, the contours of the law–science interaction are shaped by the values underlying a society's legal system. One may hypothesize that all legal systems lean toward an epistemology of science and the scientific enterprise that fits comfortably within their dominant methods of law making and dispute settlement. One might expect, therefore, that legal actors in inquisitorial systems would be even more skeptical of a constructivist view of the scientific enterprise than those in adversarial legal systems.

Effects of Growing Interdependence

As this article attests, science's impact on legal processes grows apace. In many areas, an attorney unarmed with scientific expertise operates at a significant disadvantage. However, a growing number of treatises attempt to inform lawyers and judges of the state of scientific knowledge in various areas (Faigman et al., 2013). At the level of the case, there is some evidence that scientific concerns are altering the way courts and other legal actors address some types of supposed expertise (National Research Council, 2009).

What is the effect of law on science? If the supply of science on some issues is affected by demand, legal controversy should draw scientists to research topics they might otherwise have eschewed. There is evidence that this does occur. The subject matter of scientific research is shaped by legal controversy in a wide number of areas (Schuck, 1986; Sobol, 1991). Some have argued that law's interest has an additional effect of causing the production of 'worse' science (Huber, 1991), but it might also be argued that in areas such as DNA testing, law's interest has produced better science and better technology than might otherwise have existed.

Scientific groups often attempt to surround themselves with a social structure that keeps society at a distance. Increasingly,

they also attempt to propagate ethical standards for those who offer expert testimony in courts, reflecting the widespread belief that scientists who typically appear in legal arenas frequently disregard accepted interpretative canons. Finally, legal use of science affects the closure of scientific debates (Sanders, 1998). On the one hand, law may perpetuate controversy by supplying resources to both sides. On the other hand, it may help to bring a dispute to closure by authoritatively declaring one side to be correct.

See also: Empirical Legal Studies; Expert Testimony; Science and Law; Science and Technology Studies: Experts and Expertise.

Bibliography

Angell, Marcia, 1996. Science on Trial: The Clash of Medical Evidence and the Law in the Breast Implant Case. W.W. Norton, New York.

Beecher-Monas, Erica, 2009. Reality bites: the illusion of science in bite-mark evidence. Cardozo Law Review 30, 1369.

Bernstein, David, 2008. Expert witnesses, adversarial bias, and the (partial) failure of the Daubert revolution. Iowa Law Review 93, 451.

Caudill, David S., LaRue, Lewis H., 2006. No Magic Wand: The Idealization of Science in Law. Rowman & Littlefield, Lanham, MD.

Chase, Oscar, 2005. Law, Culture, and Ritual: Disputing Systems in Cross-Cultural Context. New York University Press, New York.

Cole, Stephen, 1992. Making Science: Between Nature and Society. Harvard University Press, Cambridge, MA.

Cole, Simon A., 2001. Suspect Identities: A History of Fingerprinting and Criminal Identification. Harvard University Press, Cambridge, MA.

Cranor, Carl F., 2006. Toxic Torts: Science, Law and the Possibility of Justice. Cambridge University Press, Cambridge.

Damaska, Mirjan R., 1986. The Faces of Justice and State Authority. Yale University Press, New Haven, CT.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Dawid, A. Philip, 2011. The role of scientific and statistical evidence in assessing causality. In: Goldberg, Richard (Ed.), Perspectives on Causation. Hart Publishing, Oxford, pp. 133–148.

Faigman, David, 2010. Evidentiary incommensurability: a preliminary exploration of the problem of reasoning from general scientific data to individualized legal decision making. Brooklyn Law Review 75, 1137.

Faigman, David, et al., 2013. Modern Scientific Evidence. Thompson/West, Eagan, MN.

Fleck, Ludwik, 1979/1935. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago, IL.

Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

Fuchs, Stephan, Ward, Steven, 1994. What is deconstruction, and where and when does it take place? Making facts in science, building cases in law. American Sociological Review 59, 481–500.

General Electric Co. v. Joiner, 522 U.S. 136 (1997).

Goldman, Alvin I., 1999. Knowledge in a Social World. Clarendon Press, Oxford.

Haack, Susan, 2003. Defending Science within Reason: Between Scientism and Cynicism. Prometheus Books, Amherst, NY.

Haack, Susan, 2009. Irreconcilable differences? The troubled marriage of science and law. Law & Contempory Problems 72 (Wtr), 1.

Haack, Susan, 2010. Federal philosophy of science: a deconstruction and a reconstruction. New York University Journal of Law and Liberty 5, 394.

Hand, Learned, 1901. Historical and practical considerations regarding expert testimony. Harvard Law Review 15, 40.

Hofstede, Geert, 1980. Culture's Consequences. Sage Publications, Beverly Hills, CA.

Huber, Peter, 1991. Galileo's Revenge: Junk Science in the Courtroom. Basic Books, New York.

Jasanoff, Sheila, 1995. Science at the Bar: Law Science and Technology in America. Harvard University Press, Cambridge, MA.

Jasanoff, Sheila, 1991. Acceptable evidence in a pluralistic society. In: Mayo, Deborah G., Hollander, Rachelle D. (Eds.), Acceptable Evidence: Science and Values in Risk Management. Oxford University Press, Oxford.

Jones, Carol A.G., 1994. Expert Witnesses: Science, Medicine, and the Practice of Law. Oxford University Press, New York.

Jurs, Andrew, 2012. Balancing legal processes with scientific expertise: expert witness methodology in five nations and suggestions for reform of post-daibert U.S. reliability determinations. Marquette Law Review 95, 1329.

Kagan, Robert A., 2001. Adversarial Legalism: The American Way of Law. Harvard University Press, Cambridge, MA.

Kuhn, Thomas, 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, IL.

Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).

Labinger, Jay, Collins, Harry (Eds.), 2001. The One Culture? A Conversation about Science. University of Chicago Press, Chicago, IL.

Landsman, Stephan, 1995. Of witches, madmen and product liability: an historical survey of the use of expert testimony. Behavioral Science and Law 13, 131.

Langbein, John, 1985. The German advantage in civil procedure. University of Chicago Law Review 52, 823.

Latour, Bruno, 1999. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.

Lawler, Andrew, 1996. Support for science stays strong. Science 272, 1256.

Leplin, Jarrett, 1997. A Novel Defense of Scientific Realism. Oxford University Press, New York.

Luban, David, 2007. Legal Ethics and Human Dignity. Cambridge University Press, Cambridge.

Markovits, Daniel, 2008. A Modern Legal Ethics: Adversary Advocacy in a Democratic Age. Princeton University Press, Princeton, NJ.

Merton, Robert K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago, IL.

Mnookin, Jennifer L., et al., 2011. The need for a research culture in the forensic sciences. UCLA Law Review 58, 725.

National Research Council, 2009. Strengthening Forensic Science in the United States: A Path Forward. U.S. G.P.O, Washington.

Peel, Jacqueline, 2010. Science and Risk Regulation in International Law. Cambridge University Press, New York.

Risinger, D. Michael, et al., 1989. Exorcism of ignorance as a proxy for rational knowledge: the lessons of handwriting identification expertise. University of Pennsylvania Law Review 137, 731.

Sanders, Joseph, 1998. Bendectin on Trial: A Study of Mass Tort Litigation. University of Michigan Press, Ann Arbor, MI.

Sanders, Joseph, 2007. Expert witness ethics. Fordham Law Review 76, 1539.

Sanders, Joseph, 2010. Applying Daubert inconsistently?: proof of individual causation in toxic tort and forensic cases. Brooklyn Law Review 75, 1367.

Schuck, Peter, 1986. Agent Orange on Trial: Mass Toxic Disaster in the Courts. Belknap Press, Cambridge, MA.

Shaham, Ron, 2010. The Expert Witness in Islamic Courts: Medicine and Crafts in the Service of Law. University of Chicago Press, Chicago, IL.

Sklansky, David, 2009. Anti-inquisitorialism. Harvard Law Review 122, 1634.

Sobol, Richard B., 1991. Bending the Law: The Story of the Dalkon Shield Bankruptcy. University of Chicago Press, Chicago, IL.

Taylor, Robert, 1996. A comparative study of expert testimony in France and the United States: philosophical underpinnings, history, practice, and procedure. Texas International Law Journal 31, 181.

Van Kampen, P.T.C., 1998. Expert Evidence Compared: Rules and Practices in the Dutch and American Criminal Justice System. Intersentia Rechtswetenschappen, Antwerp.

Vidmar, Neil, 1999. The Canadian criminal jury: searching for a middle ground. Law & Contemporary Problems 62 (Spr), 141.

Wagner, Wendy, Steinzor, Rena, 2006. Rescuing Science from Politics: Regulation and the Distortion of Scientific Research. Cambridge University Press, New York.

Weber, Max, 1949. The Methodology of the Social Sciences (Edward A. Shils, Henry A. Finch, Trans.). Free Press, Glencoe, IL.

Wendel, W. Bradley, 2010. Lawyers and Fidelity to Law. Princeton University Press, Princeton, NJ.

Zammito, John H., 2004. A Nice Derangement of Epistemes: Post-Positivism in the Study of Science from Quine to Latour. University of Chicago Press, Chicago, IL.

Scientific Explanation

C Mantzavinos, University of Athens, Athens, Greece

© 2015 Elsevier Ltd and Chrys Mantzavinos [Translation from the English language edition]: INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL AND BEHAVIORAL SCIENCES, SECOND EDITION, edited by James D. Wright. All Rights Reserved.

Abstract

There are three main approaches to scientific explanation in the philosophical literature. The unification approach claims that science explains by fitting the particular facts and events within a general theoretical framework. The mechanistic approach claims that science explains by identifying mechanisms. According to the manipulationist approach an explanation ought to be such that it can be used to answer a "what-if-things-had-been-different question." The article examines whether these three approaches are compatible or not in the case of the social sciences, and it concludes by defending explanatory pluralism.

Introduction

The three philosophical approaches to scientific explanation that are currently dominating the discussion are the unification approach, the causal/mechanistic approach, and the manipulationist approach. The unification approach claims that science explains by fitting the particular facts and events within a general theoretical framework. The mechanistic approach claims that science explains by identifying mechanisms understood as entities and activities organized such that they are productive of regular changes from start to termination conditions. Finally, the manipulationist approach claims that explanatory activity consists mainly in answering the "what-if-things-had-been-different question." I will address the issue of compatibility of these three approaches with a specific focus on the social sciences. Are those philosophical accounts of explanation compatible with each other and if yes, in what sense and how exactly?

The Three Main Current Philosophical Theories of Explanation

The philosophical discussion on the theory of scientific explanation centered for many decades around the epoch-making essay of Carl C Hempel and Paul Oppenheim, 'Studies in the Logic of Explanation.' Even when philosophers developed alternative views, the Hempel and Oppenheim (1948) paper served as the default position. At the end of many decades of debate three powerful schools of thought seem to dominate the scheme today: the unification approach, the mechanistic approach, and the manipulationist approach.

The unification thesis, whose chief proponents are Friedman (1974) and Kitcher (1981, 1989), holds that scientific understanding increases as one decreases the number of independent assumptions that are required to explain what goes on in the world. It seeks laws and principles of high generality with the aim of constructing a coherent world picture and fitting particular facts within this framework. Besides, it is not committed to the world picture being deterministic since it is perfectly compatible with the position that basic laws can be irreducibly statistic. The thrust of the central argument of this approach is nicely summarized in the following quote from the classic paper of Friedman (1974, 15): "I claim that this is the crucial property of scientific theories we are looking for; this is the essence of scientific explanation – science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given. A world with fewer independent phenomena is, other things equal, more comprehensible than with more." The unification approach being, of course, different from the received view of Hempel–Oppenheim, it still remains somehow close to it (mainly in virtue of its insistence on deductivism).

Some new and important work on the unification approach that stresses different dimensions of unification has been produced (for a review Psillos, 2002), and it is characteristic for its relevance that any new theoretical endeavor on scientific explanation feels obliged to take a position vis-à-vis this approach. For the purposes of this paper it will be Kitcher's formulation that will be preferred, which focuses on the scarcity of patterns of derivation. Unification is reached by deriving descriptions of many types of phenomena using one or a few argument patterns over and over again respecting certain constraints, stringency being the most important one.

The causal/mechanical approach to scientific explanation was born mainly as an attempt to repair the two most serious problems of the received view, i.e., the problem of causal asymmetries (associated with the famous flagpole counterexample) and the problem of relevance (associated with the famous example of the hexed table salt). Besides Railton (1978, 1981) and Humphreys (1981, 1989), it is Wesley Salmon who has most prominently argued in favor of bringing 'cause' back into 'because.' The straightforward way to remedy the main problems of the Hempel–Oppenheim model is supposed to consist in integrating a theory of causality into the theory of explanation or, in other words, in providing scientific explanations by identifying causes of events and/or processes. Since this approach tries to take account of the explanatory practices in science (mainly physics), it aims not only at derivations of low-level laws and generalizations from higher level theories but also at elucidating the mechanisms at work. To explain is, thus, to expose the internal workings, to lay bare the hidden mechanisms, to open the black boxes that nature presents to us. This view makes explanatory knowledge into knowledge of the hidden mechanisms by which nature works. Salmon (1984) has tried to specify the notion of mechanism by pointing to causal processes: According to his

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63100-8

theory those processes (and only those) are causal that are capable of transmitting a mark.

The first decade of the new millennium has seen an explosion of work in this direction. On the one hand, the seminal paper of Machamer et al. (2000) has provoked further the 'thinking about mechanisms.' On the other hand, further work on causality has been produced, and the accounts of causality have reached a higher level of technical sophistication than any of the accounts in the past (Cartwright, 2007). Mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, p. 3) should be sought in order to be able to explain how a phenomenon comes about or how some significant process works – this is the main message of the mechanistic approach to explanation. The search for mechanisms goes hand in hand with three claims: (1) explanations should provide causes (or reasons); (2) explanations should make phenomena intelligible; (3) explanations should exhibit the continuity among the explaining parts (Machamer, 2009). Salmon's causal/mechanical approach was mainly inspired by physics, and his mark-transmission (and later in Salmon (1998) the conserved quantity-transmission) theory of causal processes was tailored to physics, whereas the modified mechanistic approach has extended its reach to the life sciences, the cognitive sciences, and the social sciences. Defenders of this approach in the life sciences claim that mechanistic explanations differ from more traditional, nomological explanations because (1) they are not limited to linguistic representations and logical inference, but employ frequently diagrams to characterize mechanisms and simulations to reason about them; (2) the fact that mechanisms involve organized systems of component parts and operations provides direction to both the discovery and testing of mechanistic explanations; and (3) models of mechanisms are developed for specific exemplars and are not represented in terms of universally quantified statements (Bechtel and Abrahamsen, 2005; Bechtel, 2006; Darden, 2006; Bogen and Machamer, 2011).

In the cognitive neurosciences, the mechanistic approach points to the fact that explanations in neuroscience describing mechanisms are multilevel and integrate multiple fields (Craver, 2007; Bechtel, 2008). Finally, in the social sciences, though the problem of interpretation of meaningful actions remains important (Mantzavinos, 2005, 2012), a great number of both philosophers and practicing scientists hold the view that social scientific explanations require the discovery of the underlying causal mechanisms that give rise to the outcomes of interest (Hedström and Swedberg, 1996; Demeulenaere, 2011). The search for causal mechanisms is often combined with the position of methodological individualism (Mantzavinos, 2009), but more recently also with the position of methodological localism (Knight, 2009; Little, 2009). All in all, judging from the recent wealth of publications on the mechanistic approach and how it can be applied on many different domains of science, one would apparently not be very wrong to state that this approach is in the process of becoming the new mainstream in the theory of explanation.

The manipulationist approach of Jim Woodward is the third philosophical account of explanation that has gained prominence in recent years. This approach was designed as an alternative to the common view that explanation involves subsumption under laws. According to Woodward (2000, 2003) whether or not a generalization can be used to explain has to do with whether it is invariant rather than with whether it is lawful. A generalization is invariant if it is stable or robust in the sense that it would continue to hold under a relevant class of changes. For example, a generalization can be invariant even if it has exceptions or holds only over a limited spatiotemporal interval. A relationship among some variables (or magnitudes) X and Y is said to be causal if, were one to intervene to change the value of X appropriately, the relationship between X and Y would not change and the value of Y would change. In a nutshell, an explanation for Woodward ought to be such that it can be used to answer what he calls a "what-if-things-had-been-different question," i.e., the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways.

Summarizing, it seems that there are three viable accounts in the philosophical theory of explanation: the unification approach, the mechanistic approach, and the manipulationist approach. The question is whether some or all of them are compatible, and if yes, in which sense. The fundamental issue this article is tackling is whether one can build bridges between the different theories of explanation.

The Question of the Compatibility of the Three Philosophical Theories of Explanation

The main question, thus, is whether the unification approach, the mechanistic approach, and the manipulationist approach to explanation are compatible, and, if yes, how exactly. Kitcher (1985) has christened the unification approach as a 'top-down' approach since scientific explanations are supposed to generally unify the understanding of the world by fitting the particular facts and events within a general theoretical framework – explanations serve to systematize knowledge in the most efficient way. The mechanistic approach is, on the contrary, according to Kitcher's (1985) terminology, a 'bottom-up' approach, since it aims at identifying mechanisms and provides explanatory knowledge by virtue of showing how they work and what outcomes they tend to bring about. Are these two ways of looking at scientific explanations compatible with one another and with the manipulationist approach?

The question of the compatibility of the three approaches must be formulated more sharply. Are the three approaches compatible and in what sense? I want to distinguish between two senses:

Compatibility in the weak sense: The three approaches are compatible in that each approach applies in different domains of science.

Compatibility in the strong sense: Uses of different approaches in the same domain would lead to the same conclusions regarding explanatory relations.

I only defend the compatibility in the weak sense. Prima facie, at least, the three approaches are not necessarily mutually exclusive – each one offers a reasonable way of construing explanations. They might be taken as representing indeed three different, but compatible, aspects of scientific

explanation. Since scientific understanding is after all a very complex matter, there is every reason to suppose that it has different dimensions. In other words, there is no a priori reason to assume that a single theory of philosophical explanation must necessarily cover all dimensions of explanatory activity. However, this prima facie contention needs to be elaborated and defended.

The core issue that seems to divide the unification approach on the one hand and the mechanistic and manipulationist approaches on the other seems to be the question whether explanations must necessarily include causes, or in other words whether any genuine explanation has to take into consideration the causes involved. The causal/mechanical approach takes this for granted and goes on to specify how exactly causality is tied to explanation. The unification approach, on the other hand, contends that the concept of causal dependence is derivative from that of explanatory dependence. It is from the 'because' that one can infer the 'cause' and not vice versa. What is distinctive about the unification view is, thus, that it proposes to ground causal claims in claims about explanatory dependency rather than vice versa. This approach accounts for the intuition that appeals to shadows not explaining the heights of towers because shadow heights are causally dependent on tower heights. It suggests that the view of causality in this and similar cases rather stems from an appreciation of the explanatory ordering of beliefs. Put more simply, one can only identify something as a cause because this something provides the explanation of the phenomenon at hand – the concept of causal dependence is derivative from that of explanatory dependence. Adopting a position and making a decision in favor of the primacy of causal or explanatory dependence goes hand in hand with adopting the respective metaphysical commitments and coming to grips with Hume's legacy that causal judgments are epistemologically problematic.

This difference in the approaches seems to be really fundamental since it concerns the thorny metaphysical question about the nature of causality. Although more and more philosophers – surprisingly – hasten to jump on the train of causality in order to reach explanation, cautiousness is necessary. Not all explanation is causal explanation; the most prominent examples of noncausal explanation include mathematical, geometrical, and equilibrium explanation (Sober, 1983). If all these are accepted as genuine cases of explanation, then it is definitely wrong to insist on including causal arguments in every explanatory account. However, this core issue remains, and, in my judgment, constitutes the main dividing line between the unificationist approach on the one hand and the mechanistic and manipulationist approaches on the other, and in the end this is the main reason that a real bridge between them is impossible.

However, the question of their compatibility remains open and should be addressed (Strevens, 2008). It is necessary to take a close look at the scientific practices prevailing in different domains of science in order to inquire whether the three approaches are compatible in the sense that one of them best accommodates the scientific practices in some domains of science and the others accommodate best the scientific practices in other domains of science. Should this be the case, then the compatibility of the three approaches would not refer to the theoretical bridges in the philosophical theory of explanation intended to bring the three approaches closer together, but rather to the domains of application of the three approaches – the compatibility in the weak sense could be defended. I will devote more space to this task and mainly focus on the social sciences.

Compatibility of the Three Approaches of Explanation in the Social Sciences

The picture in the social sciences seems to be very disparate. On the one hand, the most theoretically developed field is probably neoclassical microeconomics, for which the unification approach with its insistence on deductivism seems to offer the best account. On the other hand, in sociology and political science one is confronted with a multitude of theoretical approaches, which do not aim at establishing a global theoretical framework but insist instead on highlighting specific, middle-range causal mechanisms as the means of explaining certain aspects of the social world. The mechanistic approach to explanation can clearly better accommodate these cases. Finally, it seems possible to capture a lot of empirical work carried out in the social sciences with the interventionist account of Jim Woodward. This is especially true for econometrics, which provides the example par excellence of the kind of toolbox used in trying to answer a "what-if-things-had-been-different question," the question that occupies center stage in Woodward's manipulationist account of explanation. I now will elaborate on my claim that each philosophical account of scientific explanation best captures a different domain of social science.

Neoclassical Microeconomic Theory

The standard neoclassical microeconomic theory is based on the theoretical construction of utility maximization. Since the marginalist revolution in the 1870s and the pioneering works of Leon Walras, Carl Menger, and William Stanley Jevons, a theory of price has been devised based on marginal utilities. Marshall's (1890/1920) *Principles of Economics* then provided a systematic account of the interplay between demand and supply on product and factor markets. Today, neoclassical microeconomic theory provides a standard axiomatization of the behavior of households and firms in markets. The general theoretical framework that underlies this neoclassical theory of markets is the rationality hypothesis. The hypothesis of utility maximization plays the fundamental role in driving economic research, and it claims to offer a theoretical account covering all those cases where two or more individuals exchange goods under conditions of scarcity. Neoclassical microeconomic theory offers a global theoretical framework for both a partial analysis of a single market and a total analysis of all markets in an economy, using the utility maximization hypothesis and focusing on the properties of economic equilibrium.

As Mäki (2001) and Mäki and Marchionni (2009) correctly point out, the ideal of explanatory unification is prevalent in economics, and more specifically in neoclassical economic theory. One of the great merits of the theory is supposed to consist precisely in its unifying power, i.e., its ability to subsume a great range of economic phenomena under a unique descriptive and explanatory scheme. Kitcher's (1989, 432) verdict seems to fit exactly to the case of neoclassical

microeconomics: "Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation over and over again, and, in demonstrating this, it teaches how to reduce the number of types of facts we have to accept as ultimate (or brute)." The unificationist approach to scientific explanation seems unambiguously to capture, thus, the heart of the neoclassical economic enterprise, to derive over and over again a wide range of descriptions of economic phenomena from a standard set of patterns, with the maximization under constraints being the most obvious and prominent one.

Now, neoclassical microeconomic theory is, and has always been, under serious attack. Alternative research programs have always prevailed. Institutional Economics is the most prominent case at hand (Mantzavinos, 2001; Mantzavinos et al., 2004) (with a few Nobel Prizes awarded to a series of leading figures working in this research program), and more recently Behavioral Economics. This is not the place to proceed to the evaluation of the alternative research programs in economics or to present the battery of critical arguments against neoclassical economics that has been shaped and elaborated on by a series of economists and philosophers. What is important here is the following: There is an *opinio communis* among critics and defenders of neoclassical economics alike that one, if not the main, virtue of the approach is its power to unify diverse phenomena under a single descriptive and explanatory scheme. It is due to this specific merit of the theory that it is probably the only piece of social science that is offered in a standardized way in every single introductory textbook taught in every single economics department around the globe.

The ideal of unification has been pushed even further by applying the neoclassical toolbox to the study of phenomena other than the traditional economic ones. Buchanan (Buchanan (1975); Buchanan and Tullock (1962); Brennan and Buchanan (1985)), Downes (1957), and Olson (1965) – just to name a few protagonists of introducing neoclassical economic theory into the study of politics – have given birth to a whole new discipline, Public Choice (Mueller, 2003), which is a consistent step toward a unification of economic and political phenomena. Besides, the same patterns of derivation of neoclassical economics 'are used over and over again' to explain such diverse social phenomena as crime and family (Becker, 1976; Becker and Posner, 2009), and law and sex (Posner, 2010). It is clear that the unificationist theory of explanation can provide a better account and accommodate better the practices of the scientists working within the neoclassical economic theoretical framework than the alternative philosophical theories of explanation. Taking into consideration that neoclassical economic theory avails of a unique reputation amidst the social sciences – rightly or wrongly so is not under dispute here – the unification approach to explanation seems to be tailored to capture the reasons for its popularity.

Social Mechanisms in Political Science

A great deal of political science, when not of a descriptive or purely historical nature, deals with laying bare the working properties of social mechanisms or opening up the black boxes between independent and dependent variables (e.g., Elster, 2007, ch. 2) by identifying the specific causal patterns at work. In comparative politics in political science, for example, the focus of political scientists is to highlight the mechanisms that exhibit a relative constancy across countries and epochs. A prominent piece of theoretical work that aims at discovering a specific social mechanism that I want to briefly present here is Kalyvas's (2006) *The Logic of Violence in Civil Wars*.

Kalyvas differentiates between the broad concept of civil war and the phenomenon of civil war violence. His study is innovative inter alia for not treating violence and war as the same phenomena as has previously been done in the literature, but for focusing on violence within civil wars as a dependent variable. In his model, civil war itself is an exogenous shock onto a society, which can activate invisible networks of grievances among its individuals. His theoretical approach breaks civil war violence down into two basic categories: indiscriminate violence is executed massively, whereas, in contrast, selective violence describes targeted aggression toward individuals identified as enemies with the help of specific local information about their actions. The aim is to provide an explanation of the spatial variance of the dependent variable, i.e., of violence in civil wars – characterized by its barbarism and the fact that victims and executioners used to live together peacefully before the outbreak of the civil war. The independent variable that Kalyvas highlights as the most important is territorial control exercised by opposed fractions. He explains the level of violence in a territory as a function of the degree of control that the warring fractions exert over a particular territory.

The mechanism that is at work starts with individual actors, who are assumed to be rational in their activity in the sense of trying to further their self-interest within a given geographical region. Civil wars are very often warred by means of irregular warfare, and despite the frequency and planning that are involved in indiscriminate violence, the warring fractions soon discover it to be counterproductive. Armed groups soon realize that selective violence better furthers their interest, but then their principal problem becomes the removal of the uncertainty about who the enemy is and who is neutral in an environment where competitors are able to hide behind civilian population. The solution to this problem requires armed groups to collect specific information about individuals, which in turn gives individual noncombatants incentives to cooperate. Since enemies can be hidden among apparent supporters of a community, the exertion of selective violence requires information about specific people, which is most easily collected from individual noncombatants acting as denouncers; and thus selective violence becomes "a *joint* process, created by the actions of both political actors and civilians" (Kalyvas, 2006, p. 209, emphasis in original). The probability of denunciations grows in a nonmonotonical way with the level of control that one side has in a given area since differing levels of control provide different incentives to potential civilian informers and thus leads them to different calculations with respect to the benefits of denunciation. The more control one side has in a given geographical territory, the less the denouncers will face risks of retaliation by the other side and thus the higher will be their propensity to collaborate and provide information – very often because of grievances against specific individuals dating back to the peaceful period. Selective violence is the outcome of the interplay of many individual decisions that take place in

certain social contexts defined precisely within the model. The evidence provided stems from regional data of the Greek Civil War, and it accounts for two-thirds of violence variance, and the validity of the model outside the Greek sample is tested through a great range of both historical and anecdotal accounts about civil wars around the world.

The mechanistic model of explanation can clearly best accommodate such pieces of social scientific research. The aim of this kind of scientific endeavor is not to establish a global theoretical framework and provide the unification of a great range of phenomena, but to insist instead on highlighting specific, middle-range causal mechanisms as the means of explaining certain aspects of the social world. The mechanistic approach to explanation is clearly the appropriate philosophical model in those cases.

Econometrics

Econometrics originally emerged as a field at the intersection of economics, mathematics, and statistics, aiming at the provision of numerical values for the parameters of economic relationships, for example, elasticities, mariginal values, etc. and at the verification or falsification of economic theories. The main insight offered by econometrics is that economic relationships contain a random element, which is typically ignored by economic theory that standardly postulates exact relationships between the various economic magnitudes. Econometrics offers a range of methods for dealing with the random component of economic relationships.

Since the heart of econometrics lies in regression equation models that specify a functional relationship between dependent and independent variables, including also a 'measurement error,' econometric techniques have been included over the years in the toolboxes of other disciplines in the social sciences. Regression equations take the following form:

$$Y = B_1X_1 + B_2X_2 + \cdots + B_nX_n + U$$

$X_1, X_2, \ldots, X_n$ are the independent variables, $B_1, B_2, \ldots, B_n$ are the coefficients, Y is the dependent variable, and U the error term (which makes the model a stochastic one).

Now, such equations cry for a causal interpretation, which can shed light upon the relationship between different variables in a specific setting. This is exactly what the philosophical theory of explanation proposed by Woodward offers: a natural causal interpretation of such regression equations. Each of X_1, X_2, …, X_n are (direct) causes of Y, and these causal relationships are understood as holding for each individual in the population of interest "in the sense that for each such individual (the regression equation) characterizes the response of the value of Y possessed by that individual to some range of interventions that change the values of $X_1, \ldots, X_n$ for that individual. On this interpretation, the error term U also has a causal interpretation: it represents the combined influence of all the other causes of Y besides $X_1, \ldots, X_n$ that are not explicitly represented in (the regression equation)" (Woodward, 2003, p. 321).

A great bulk of applied work in economics and the other social sciences consists in establishing the truth of those kinds of quantitative relationships. The core issue is that social scientists very often do not aspire to establish the truth of general laws, but rather to establish the truth of quantitative relationships that have been tested successfully and hold in specific populations and under specific circumstances. Judged from a God's point of view, those generalizations might be judged insufficient on many grounds; however, scientific knowledge in the social sciences *de facto* consists of such quantitative relationships that are invariant and hold only in specific settings. Neither the unificationist nor the mechanistic approach to explanation can adequately capture these scientific practices. Woodward's manipulationist account on the contrary seems to be tailored to them.

Summarizing, the discussion has shown that there definitely is compatibility between the three philosophical approaches to explanation, but this does not allow them to be synthesized into one approach; rather they are compatible in that they can be applied in different and distinct domains of social scientific knowledge. In other words, they are compatible because they refer to different explanatory enterprises that constitute parts of different domains of social sciences and that are all perfectly legitimate. Insofar the weak compatibility thesis is tenable.

Conclusion: Towards Explanatory Pluralism

Since it is commonly accepted that philosophy of science as a discipline can only be useful if it takes the scientific practices seriously, and it is clearly the case that in different domains different kinds of scientific practices prevail, trying to establish a monolithic philosophical theory of explanation that is supposedly good for everything cannot be an acceptable strategy. In a nutshell, the goal of a philosophical account of explanation should not be to capture the explanatory relation, but rather to capture the many ways in which explanations are provided in the different domains of science.

The position that emerges then and that I wish to adopt is that of an explanatory pluralism. This is a pluralistic account that allows for different ideal types of explanation, i.e., different exemplary accounts of what an explanation consists of, which are good as means of classification of different types of explanatory activities that are offered in different domains of the (social) sciences. This explanatory pluralism is different from the explanatory ecumenism that Jackson and Pettit (1992) endorse insofar as it is independent of any type of commitment to causality. In other words, I disagree with what Jackson and Pettit call 'causal fundamentalism,' i.e., the view that every explanation must provide information on the causal history of what is to be explained (a view also endorsed by Lewis (1986, p. 217) in his famous verdict that "to explain an event is to provide some information about its causal history"). The explanatory pluralism that I propose is a genuinely pluralistic position that makes as few commitments as possible to causality (and other metaphysical issues). It does not preclude that explanations in some domains of science are offered, debated, and criticized in terms of figuring out causal patterns, processes, or mechanisms; it also permits scientific explanations that provide scientific understanding by, for example, unifying phenomena without invoking causes.

The main danger for a defender of such a position is the possibility of sliding into a pure sociological theory of scientific explanation that just provides descriptions of different explanatory practices or, even worse, to make such a position serve a purely apologetic function by providing ex post legitimization

of any and/or all explanatory scientific practices. However, philosophy of science has a normative function to serve, so that a philosophical theory of explanation that intends to be pluralistic must clearly accommodate normative considerations. Working out the normative dimension of explanatory pluralism is clearly a desideratum. However, any kind of normative considerations cannot be formulated *in abstracto*, but must be founded on the existing scientific practices. The aim of this article has been the moderate one of establishing the claim that the philosophical theories of explanation currently on offer do not exclude each other but can rather capture some of those practices and thus of reorienting the discussion toward the feasibility of the position of explanatory pluralism.

See also: Causation: Physical, Mental, And social; Causes and Laws: Philosophical Aspects; Functional Explanation: Philosophical Aspects; Historical Explanation, Theories of: Philosophical Aspects; Methodological Individualism: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Social Science, The Idea of.

Bibliography

Becker, Gary, 1976. The Economic Approach to Human Behavior. The University of Chicago Press, Chicago and London.
Becker, Gary, Posner, Richard, 2009. Uncommon Sense: Economic Insights from Marriage to Terrorism. The University of Chicago Press, Chicago and London.
Bechtel, William, 2006. Discovering Cell Mechanisms. Cambridge University Press, Cambridge.
Bechtel, William, 2008. Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. Routledge, London.
Bechtel, William, Abrahamsen, Adele, 2005. Explanation: a mechanist alternative. Studies in History and Philosophy of Biological and Biomedical Sciences 36, 421–441.
Bogen, Jim, Machamer, Peter, 2011. Mechanistic information and causal continuity. In: McKay Illari, Phyllis, Russo, Federica, Williamson, Jon (Eds.), Causality in the Sciences. Oxford University Press, Oxford, pp. 845–864.
Brennan, Geoffrey, Buchanan, James, 1985. The Reason of Rules – Constitutional Political Economy. Cambridge University Press, Cambridge.
Buchanan, James, 1975. The Limits of Liberty. Between Anarchy and Leviathan. The University of Chicago Press, Chicago.
Buchanan, James, Tullock, Gordon, 1962. The Calculus of Consent – Logical Foundations of Constitutional Democracy. University of Michigan Press, Ann Arbor.
Cartwright, Nancy, 2007. Hunting Causes and Using Them. Cambridge University Press, Cambridge.
Craver, Carl, 2007. Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience. Oxford University Press, Oxford.
Darden, Lindley, 2006. Reasoning in Biological Discoveries. Essays in Mechanisms, Interfield Relations and Anomaly Resolution. Cambridge University Press, Cambridge.
Demeulenaere, Pierre (Ed.), 2011. Analytical Sociology and Social Mechanisms. Cambridge University Press, Cambridge.
Downes, Anthony, 1957. An Economic Theory of Democracy. Harper, New York.
Elster, Jon, 2007. Explaining Social Behavior. Cambridge University Press, Cambridge.
Friedman, Michael, 1974. Explanation and scientific understanding. Journal of Philosophy 71, 5–19.
Hedström, Peter, Swedberg, Richard (Eds.), 1996. Social Mechanisms. An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Hempel, Carl G., Oppenheim, Paul, 1948. Studies in the logic of explanation. Philosophy of Science 15, 135–175 (Reprinted in Hempel (1965)).
Hempel, Carl G, 1965. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press, New York.
Humphreys, Paul, 1981. Aleatory explanation. Synthese 48, 225–232.
Humphreys, Paul, 1989. The Chances of Explanation. Princeton University Press, Princeton, NJ.
Jackson, Frank, Pettit, Philip, 1992. In defense of explanatory ecumenism. Economics and Philosophy 8, 1–21.
Kalyvas, Stathis, 2006. The Logic of Violence in Civil War. Cambridge University Press, Cambridge.
Kitcher, Philip, 1981. Explanatory unification. Philosophy of Science 48, 251–281.
Kitcher, Philip, 1985. Two approaches to explanation. Journal of Philosophy 82, 632–639.
Kitcher, Philip, 1989. Explanatory unification and the causal structure of the world. In: Kitcher, Philip, Salmon, Wesley (Eds.), Scientific Explanation, Vol. 13 of Minnesota Studies in the Philosophy of Science, pp. 410–505.
Knight, Jack, 2009. Causal mechanisms and generalizations. In: Mantzavinos, C. (Ed.), Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice. Cambridge University Press, Cambridge, pp. 179–184.
Lewis, David, 1986. Philosophical Papers, vol. 2. Oxford University Press, Oxford.
Little, Daniel, 2009. The heterogeneous social: new thinking about the foundations of the social sciences. In: Mantzavinos, C. (Ed.), Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice. Cambridge University Press, Cambridge, pp. 154–178.
Machamer, Peter, 2009. Explaining Mechanisms, unpublished paper.
Machamer, Peter, Darden, Lindley, Craver, Carl, 2000. Thinking about mechanisms. Philosophy of Science 67, 1–25.
Mäki, Uskali, 2001. Explanatory unification: double and doubtful. Philosophy of the Social Sciences 31, 488–506.
Mäki, Uskali, Marchionni, Caterina, 2009. On the structure of explanatory unification: the case of geographical economics. Studies in History and Philosophy of Science 40, 185–195.
Mantzavinos, C., 2001. Individuals, Institutions and Markets. Cambridge University Press, Cambridge.
Mantzavinos, C., 2005. Naturalistic Hermeneutics. Cambridge University Press, Cambridge.
Mantzavinos, C. (Ed.), 2009. Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice. Cambridge University Press, Cambridge.
Mantzavinos, C., 2012. Explanations of *meaningful* actions. Philosophy of the Social Sciences 42, 224–238.
Mantzavinos, C., Douglass, North, Shariq, Syed, 2004. Learning, institutions and economic performance. Perspectives on Politics 2, 75–84.
Marshall, Alfred, 1890/1920. Principles of Economics, eighth ed. MacMillan, London.
Mueller, Dennis, 2003. Public Choice III. Cambridge University Press, Cambridge.
Olson, Mancur, 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Posner, Richard, 2010. Economic Analysis of Law, eighth ed. Aspen Publishers, New York.
Psillos, Stathis, 2002. Causation and Explanation. McGill-Queen's University Press, Montreal & Kingston.
Railton, Peter, 1978. A deductive-nomological model of probabilistic explanation. Philosophy of Science 45, 206–226.
Railton, Peter, 1981. Probability, explanation and information. Synthese 48, 233–256.
Salmon, Wesley, 1984. Scientific Explanation and the Causal Structure of the World. Princeton University Press, Princeton, NJ.
Salmon, Wesley, 1998. Causality and Explanation. Oxford University Press, Oxford.
Sober, Elliott, 1983. Equilibrium explanation. Philosophical Studies 43, 201–210.
Strevens, Michael, 2008. Depth. An Account of Scientific Explanation. Harvard University Press, Cambridge, MA.
Woodward, James, 2000. Explanation and invariance in the special sciences. British Journal for the Philosophy of Science 51, 197–254.
Woodward, James, 2003. Making Things Happen. Oxford University Press, Oxford.

Scientific Knowledge, Sociology of

HM Collins, SOCSI, Cardiff University, Cardiff, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The origins and growth of the sociology of scientific knowledge (SSK) are described along with some recent developments. The sociology of *scientific* knowledge is distinguished from the sociology of knowledge by its readiness to apply the latter's insights to science and technology whereas before they had been treated as immune – islands of rational thought with an 'internal' dynamic only. Different philosophical flavors of SSK are distinguished.

The sociology of knowledge originally treated the sciences as immune to social analysis, but from the 1970s, the very content of scientific knowledge has been treated as a social product – at least in part. Thinking in sociology of knowledge terms, this transformation could be seen as belonging to the revolution of ideas that took place in the West in the affluent 1960s, but the origins of the thought that informed the new movement goes back further. Sociology of scientific knowledge (SSK) introduced new concepts, such as the need to treat true and false ideas symmetrically in terms of social analysis. Once truth was no longer taken to determine belief, it became possible to explore the way scientific beliefs were actually formed. Experimental findings were shown to be 'interpretatively flexible' and less determining than had been thought. Scientific disputes reached 'closure' only through a variety of means beyond theory and experiment. Trust and personal judgment turned out to have central roles in the production of scientific knowledge. Different approaches to SSK have widely differing political and philosophical consequences, philosophical radicalism not always coinciding with political significance.

Origins

The central concern of the sociology of knowledge is the idea that what people take to be certain is an accident of the society in which they are born and brought up. It is obvious that religious or political truths are largely affected by their social settings. If the truths are moral, then such relativism is more troubling. But if the truths are scientific, then the idea of social determination is widely regarded as subversive, or self-defeating – since the science that purports to show that science is socially situated is itself socially situated. The sociology of knowledge, as formulated by Mannheim (1936), avoided these dangerous and dizzying consequences by putting settled scientific and mathematical knowledge in a special category; other kinds of knowledge had roots in society but the knowledge of the natural sciences was governed by nature or logic. Scientific method, then, when properly applied, would insulate scientists from social influences and their knowledge should be more true and more universal than other kinds of knowledge. From the early 1970s, however, groups of sociologists, philosophers, historians, and other social scientists began programs of analysis and research, which treated scientific knowledge as comparable with other kinds of knowledge. Their work broke down the barrier between ordinary knowledge and scientific knowledge.

The post-1970s SSK, which included major contributions from historians of science, shows us how to understand the impact of social influences on all knowledge. We can use the field's own approach to query its intellectual origins. A suitable model is a well-known paper by historian Paul Forman (1971), who argued that the rise of quantum theory owed much to the 'Weltanschauung' of Weimar Germany. In the same way, it could be argued that the 1960s saw the growth of postwar prosperity in Europe, which allowed the development of new markets with young people becoming powerful consumers of goods and producers of culture – much of it consisting of a rebellion against traditional forms of expression. The 1960s also saw the development of the birth control pill, making possible a (temporary) breakdown of sexual mores. During that era, experiments with perception-affecting drugs led to questioning of consensual reality, while protests against the Vietnam War threatened old lines of institutional authority. The new order had its counterpart in academe. Antipsychiatry questioned the barriers between the sane and the insane, while in Europe at least, romanticized versions of Marxism dominated debate. SSK's questioning of the traditional authority of science can plausibly be seen as a product of this social ferment.

The emblematic book of the period as far as science was concerned was Thomas Kuhn's (1962) *The Structure of Scientific Revolutions*. Kuhn's book was influential in helping to reinforce the intellectual atmosphere in which radical ideas about science could flourish. Kuhn's notions of the incommensurability of paradigms, and of paradigm revolution, provided a way of thinking about scientific knowledge as a product as much of culture as of nature. The book was not widely discussed until late in the 1960s, when the new social movements were at their height, and this adds force to the argument.

Philosophy of science in the early 1970s was also affected by this turn. The so-called 'Popper–Kuhn' debate (Lakatos and Musgrave, 1970) pitted Karl Popper's (e.g., 1959) notion that scientific theories could be falsified with near certainty, even if they could not be proved, against Thomas Kuhn's (1961, 1962) claim that what was taken to be true or false varied in response to sudden revolutions in scientific world view. The 'Duhem–Quine thesis' (Losee, 1980) showed that scientific theories were supported by networks of observations, such that no one observation could threaten any other part of the network directly. Imre Lakatos's (1976[1963]) brilliant analysis of the

history of Euler's theorem showed that falsification of a theory was only one choice when faced with apparently recalcitrant observations. Lakatos's work was particularly attractive to the social analyst of science because it dealt with the detailed history of a real case rather than an ideal-typical case designed to illustrate general principles. Lakatos was showing how mathematicians argued (or could have argued in principle). This was the world of philosophical activity into which the SSK was born.

Like any other group of knowers, sociologists of scientific knowledge prefer models of the genesis of their own ideas which reflect their self-image as 'rational creatures' or, in Mannheim's phrase, 'free floating intellectuals.' In these models Kuhn plays a much smaller role. There are two intellectual stories. One turns on Ludwik Fleck's (1979 [1935]) *Genesis and Development of a Scientific Fact*, which Kuhn cites in the preface of his 1962 book, and which already contained many of the key ideas found therein. Furthermore, Fleck's book anticipated the developments in SSK, which came in the 1970s in that, unlike *The Structure of Scientific Revolutions*, it included a case study of a contemporaneous scientific episode: Fleck's own research on the Wasserman reaction for the diagnosis of syphilis. Indeed, Fleck's book remains unique in sociology of science in that his social analysis was of a pioneering piece of scientific research, which he himself was pioneering. Thus, it is the most perfect piece of participant observation yet done in the social study of scientific knowledge.

Fleck's book, however, was not widely known until many years after Kuhn's was published, and not until long after the SSK had established its own way of doing things. Fleck, then, was the first to set out some of the crucial ideas for SSK, but he had little influence on the new movement except through his influence on Kuhn. For his part, Kuhn helped open up the social and intellectual space for what came after, rather than providing intellectual foundations or methodological principles; in other words, Kuhn provided the intellectual space but not, to use his own term, the paradigm.

The origin of SSK's intellectual and methodological *paradigm* has to be understood by looking at more direct influences. In the 1970s a group of philosophers, concerned with defending science against what they perceived as its critics, put forward the argument that sociological analysis could be applied only to knowledge that was in error, whereas true knowledge remained insulated from social forces. The intellectual source for the sociologists who inspired these initial attacks and fought against them was yet another philosopher, Ludwig Wittgenstein, whose work was central to the anthropological debate about whether there were universal standards of rationality (Wilson, 1970). Readings of Wittgenstein's later philosophy (Winch, 1958; Wittgenstein, 1953, 1956) produced the crucial theoretical resource for the work in SSK that had its origins in the mid-1970s. The 'Edinburgh School' and its 'Strong Programme' argued on philosophical grounds that 'symmetrical analysis' – analysis that treated true knowledge and false knowledge equally – was possible. Historical studies conducted in this framework showed how social influence affected scientific conclusions, in science in general, in early studies of the brain, in the development of statistics, and in high-energy physics (Bloor, 1973, 1976; Barnes, 1974; Shapin, 1979; MacKenzie, 1981; Pickering, 1984). Independently, a group at the University of Bath applied Wittgensteinian ideas to contemporary episodes of science, particularly controversies in physics and the fringe sciences, and developed the methodology of interviewing participants in such controversies at networks of competing scientific centers. Its program became known as the 'Empirical Programme of Relativism' (EPOR; Collins, 1975, 1985; Pinch, 1977, 1986; Travis, 1980, 1981). These approaches were to become quite widely adopted in the early years of SSK.

As the 1970s turned into the 1980s other groups, more strongly influenced by anthropological practice, entered and influenced the field. The Edinburgh–Bath approach was influenced by anthropology in that much of the discussion of the meaning of Wittgensteinian philosophy for the social sciences turned on questions that had to do with the knowledge of isolated 'tribes' and its relationship to Western knowledge. The newer contributors, however, aligned themselves with anthropology. Their work founded the tradition of 'laboratory studies.' In the Strong Programme, the source of material tended to be historical archives; in EPOR, the method was interviews with members of the 'core set' of scientific laboratories contributing to a scientific controversy; in laboratory studies, the method was a long sojourn in, or deep study of, a single laboratory (Latour and Woolgar, 1979; Knorr Cetina, 1981). The setting for the laboratory studies tended to be the life sciences rather than the hard sciences or fringe sciences.

Another important input was the interpretive tradition in sociology associated with phenomenology and ethnomethodology. Indeed the term 'social construction' is a widely used description of the approach: 'social construction of scientific knowledge' was taken from the title of a well-known book by Peter Berger and Thomas Luckmann (1967) – though Berger's *Invitation to Sociology* (1963) probably had more direct impact on the field. Ethnomethodology strongly influenced all those who practiced SSK and especially the detailed analyses of scientific practice such as those by Lynch (1985) (but written much earlier than its publication date), Woolgar (1976), and especially the move of some authors into what became known as 'discourse analysis' (Knorr Cetina and Mulkay, 1983).

As the 1980s turned into the 1990s, the concentration on language that began with the move to discourse analysis helped SSK to appear to be part of the much larger movement known as 'postmodernism.' The influence of French philosophers of literature and culture such as Derrida became marked. Since then 'cultural studies of science,' which share with SSK what could broadly be called a 'social constructivist' approach to science, have attracted large numbers of followers in the humanities as well as the social sciences. The respective methods and, more especially, the attitudes toward methodology make it possible to separate cultural studies from SSK. Crucially, SSK practitioners still take natural science as the touchstone where matters of method are concerned, though the model of science is very far from the narrow statistically based notion of science that informed the 'scientific sociology' of the 1950s and early 1960s, and continues to dominate much mainstream sociological practice in the United States. SSK stresses the more general values of careful empirical observation and

repeatability. Cultural studies, in contrast, take philosophical literary criticism, or semiotics, as the model to be followed.

Concepts

Among sociology's traditional topics was the study of the social factors that affected scientists' choices of research topics. The SSK, however, took it that the *outcomes* of research projects also were affected by their social setting. SSK showed that theoretical and experimental procedures did not determine, or fully determine, the conclusions of scientific research. The philosophical ideas described above provided an important starting point, but sociologists of scientific knowledge developed their own concepts, which were set in the working world of the scientist rather than the abstract world of the philosopher. One strand of research showed how difficult it was to transfer the skills of physics experimentation between settings; showing further that it was even harder to know that such skills had been transferred, and even harder still to know when an experiment had been satisfactorily completed. This meant that scientists did not know with certainty when negative results from an experiment should be taken to contradict a positive outcome. This argument became known as 'the experimenter's regress' (Collins, 1975, 1985), and it revealed the scope for 'interpretive flexibility' regarding the outcomes of passages of experimentation and theorization. It also means that even while SSK proponents stress the importance of empirical research and repeatability, they are aware that, by themselves, such procedures cannot turn disputed knowledge into certainty in either the natural or social sciences.

Even if experimental science produced widely accepted data, as Pinch (1981) showed, the 'evidential significance' of such data could vary and the same finding could count as revolutionary or insignificant depending upon how it was subsequently characterized. Sociological studies also described the processes that link the disordered activities of the laboratory with the orderly 'findings' (Latour and Woolgar, 1979; Knorr Cetina, 1981) or 'discoveries' (Brannigan, 1981) reported – or constructed – in the published literature. Latour and Woolgar demonstrated the way that 'modalities' – words and phrases that qualified a finding or referred to the particular time and place of its generation – were successively stripped from publications as the finding became established. What they called 'inversion' was the way the stripping of the modalities, and similar processes, comprised the establishment of the finding – a reversal of the accepted direction of the causal arrow. It was also shown that, with some exceptions, 'distance lends enchantment,' that is, certainty about scientific outcomes tends to increase as the interpreter's distance from the laboratory increases (Collins, 1985; MacKenzie, 1998). This concept explained why commentators' accounts often seemed far more confident than the reports of the scientists themselves, and explained much about the way scientific knowledge was diffused beyond the specialists.

Mechanisms of 'closure' were described through which potentially open-ended disputes were brought to a close around a particular interpretation. Different approaches to closure were taken, with EPOR concentrating on fringe sciences, the 'French School' of Actor–Network Theory (ANT) concentrating on the interplay of human and nonhuman 'actors' within networks (Latour, 1987), and the 'Edinburgh School' concentrating on large-scale political influences on the content of ideas (Shapin, 1979; MacKenzie, 1981; Shapin and Schaffer, 1985). More recently, work on closure has broadened to include studies in the social construction of technology, which applied the EPOR to technology and is widely known by the acronym, SCOT, the public understanding of science, and law and science.

Significance

The early 1990s were characterized by what became known as the 'science wars.' Scientists, along with some philosophers, attacked science studies for its misconceptions of science and complete misunderstanding of how science works. They were worried that SSK and related analyses would weaken science. That the criticism was more of a 'war' than a proper debate can be seen from some of the more ridiculous but oft-repeated claims, such as that the critics of science had no idea of how an experiment worked. This might have been true of some of those who came from an arts and humanities background, but even a superficial reading of the works that were being attacked as icons of the movement, such as *The Golem* series of books (particularly Collins and Pinch, 1993), would have revealed that the science being discussed was thoroughly understood. Fortunately, the more outrageous of these attacks were to die away, helped by at least a partial meeting of minds of natural and social scientists represented in books such as *The One Culture* (2001; coedited by Labinger (a chemist) and Collins). The science warriors made a blanket attack on science studies, obscuring that both the philosophical and political radicalism of the social analysis of science varies from program to program and study to study. Today, it has begun to seem as though some of the science warriors' concerns might have been justified, even if the means of attack were not. Thus science and technology studies, which is based on SSK and the related programs, can have an antiscience and antitechnology flavor, as well as a tendency to favor the views of the general public on technical matters over the views of experts.

The science warriors' attacks on SSK focused on its relativism. But even this relativism is not at all uniform. For example, 'epistemological relativism' implies that one social group's way of justifying its knowledge is as good as another's and that there is no external vantage point from which to judge between them; accordingly, all that can be known is known only from the point of view of one social group or another. Ontological relativism is the view that, in social groups such as those described above, reality itself is different. A combination of epistemological and/or ontological relativism can be referred to as 'philosophical relativism.' This attitude is nicely captured by McHugh in the following quotation: "We must accept that there are no adequate grounds for establishing criteria of truth except the grounds that are employed to grant or concede it – truth is conceivable only as a socially organized upshot of contingent courses of linguistic, conceptual and social behaviour" (1971: 329). Philosophical relativism, then, is a philosophically radical viewpoint.

A still more philosophically radical position is actor (or actant) network theory most closely associated with Michel Callon and Bruno Latour. The approach was first signaled by Latour and Woolgar when, in 1986, they changed the subtitle of *Laboratory Life* (Latour and Woolgar, 1979) from *The Social Construction of Scientific Facts* to *The Construction of Scientific Facts* (1986) so as to signify that the 'social' in their view no longer deserved special mention in the shaping of scientific knowledge. Subsequently, as Callon and Latour developed their ideas, scientific facts came to be seen as emerging from the interplay of 'actants' in the 'text,' of life – suggesting that ANT should most properly be included under cultural studies of science rather than SSK. Within Callon and Latour's 'text,' terms such as 'social' and 'natural' are themselves outcomes of the interplay of actants rather than original causes. Therefore, to draw attention to the social as a special factor is to begin the discussion at too low a level of generality. Under this approach, 'constructivism' has ceased to be *social* and among the actants nonhumans are given as much reality-forming agency as humans (Callon, 1986). The philosophical radicalness of ANT is clear in its refusal to accept even the distinction between human and nonhuman as primary. ANT has been enormously successful, in part, perhaps, because it enabled humanities scholars to produce literary critiques of science without the back-breaking precondition of participating in the scientific form of life being analyzed.

Methodological relativism, by contrast, says nothing direct about reality or the justification of knowledge. Methodological relativism is an attitude of mind recommended by some of those who practice SSK; it says that the sociologist or historian of science should act as though the beliefs about reality of any of the competing groups being investigated are not caused by the reality itself. The intention is to limit analysis to the kinds of cause of scientific beliefs that are located in the domain of the social. Methodological relativism is meant to push exploration of the social causes of belief to the limit without having to cut off the investigation with the argument that a belief was widely accepted because it was rational. Methodological relativism is, then, not at all radical as a philosophy, it is a (social) scientific method.

Methodological relativism may be considered to be time bounded if it claims only that scientific procedures are often inconclusive in the short term, whatever the deeper significance of any consensus about results that is attained in the long term. This kind of relativism, which makes no strong philosophical claim beyond that fact that relativism is supportable as a methodology, is all that is needed to support much current social analysis of contemporary science, especially controversies that have impact on the public domain (e.g., Collins and Pinch, 1993; Richards, 1991; Epstein, 1996).

The relationship between philosophical radicalism and political radicalism is, however, perverse. Time-bounded relativism has consequences for contemporary scientific controversies, such as those that engage environmentalists and the like: it has this power because it shows that, contrary to the way scientific controversies are treated in textbooks and many popular accounts, the 'short-term' period before scientific disputes are resolved is often many decades long. The empirical and analytic tools of SSK can explain why such disputes are sustained, and they suggest that we should not expect speedy science-based resolutions to our current technological dilemmas. Following this line of thinking, it can be argued that the decision-making power once monopolized by scientific experts should be more widely shared (Wynne, 1987). Methodological relativism, which is philosophically quietist, tends to be disturbing to the scientific community as it blatantly sets aside scientific consensus, however well established. To go to the other extreme, ANT, the most philosophically radical position in science and technology studies, by removing humans from the central position they hold in *social* constructivism, recreates a relationship between scientists and their nonhuman objects of study, which is similar to that which held before the revolutions of the 1960s and 1970s (Collins and Yearley, 1992; Callon and Latour, 1992; Fuller, 2000). What about philosophical relativism? Scientists and philosophers often conflate methodological and philosophical relativism, believing that the later justifies nonscientific beliefs or fringe sciences, such as parapsychology, whereas the aim is to analyze the social process by which they become rejected without assuming that the content of the science is the explanation. SSK, of course, argues that as knowledge moves from its seat of creation to a wider audience, it tends to be stripped down and simplified, so it is not surprising that the subtleties of the various relativist positions have escaped the more outspoken critics of the social analysis of scientific knowledge.

In recent years an attempt has been made to reconcile the relativism of SSK and its findings with the central role of science in society, under what has been termed a Third Wave of Science Studies (Collins and Evans, 2002, 2007). The aim has been to maintain a notion of expertise in the face of the vulnerability of truth to sociological deconstruction. To the extent that the ideas of the Third Wave are successful we can both have our cake of relativist SSK and eat it in the form of expertise; this should mean a final end to any lingering suspicion about the dangerous potential of the sociological analysis of scientific knowledge.

See also: Actor-Network Theory; Cultural Studies of Science; Discourse Studies; Experiment in Science and Technology Studies; Kuhn, Thomas S. (1922–96); Laboratory Studies: Historical Perspectives; Physical Sciences: History and Sociology; Popper, Karl Raimund (1902–94); Relativism, Cognitive; Science and Technology Studies, Ethnomethodology of; Science and Technology Studies: Experts and Expertise; Science, Sociology of; Science: Constructivist Perspectives, History of; Scientific Controversies; Social Constructionism; Strong Program, Sociology of; Truth and Credibility in Science; Wittgenstein, Ludwig (1889–1951).

Bibliography

Barnes, B., 1974. Scientific Knowledge and Sociological Theory. Routledge and Kegan Paul, London.

Barnes, B., Bloor, D., Henry, J., 1996. Scientific Knowledge: A Sociological Analysis. Athlone Press, London.

Berger, P.L., 1963. Invitation to Sociology. Anchor Books, Garden City, NY.

Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality. Allen Lane, London.
Bloor, D., 1973. Wittgenstein and Mannheim on the sociology of mathematics. Studies in the History and Philosophy of Science 4 (2), 173–191.
Bloor, D., 1976. Knowledge and Human Imagery. Routledge & Kegan Paul, London.
Brannigan, A., 1981. The Social Basis of Scientific Discoveries. Cambridge University Press, New York.
Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Law, J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge? Routledge & Kegan Paul, London, pp. 196–233.
Callon, M., Latour, B., 1992. Don't throw the baby out with the bath school! In: Pickering, A. (Ed.), Science as Practice and Culture. University of Chicago Press, Chicago, pp. 343–368.
Collins, H.M., 1975. The seven sexes: a study in the sociology of a phenomenon, or the replication of experiments in physics. Sociology 9 (2), 205–224.
Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, Beverly Hills & London [second ed., University of Chicago Press, 1992].
Collins, H., Evans, R., 2002. The third wave of science studies: studies of expertise and experience. Social Studies of Science 32 (2), 235–296.
Collins, H., Evans, R., 2007. Rethinking Expertise. University of Chicago Press, Chicago.
Collins, H.M., Pinch, T.J., 1993. The Golem: What Everyone Should Know about Science. Cambridge University Press, Cambridge & New York [subsequent editions, 1994, 1998].
Collins, H.M., Yearley, S., 1992. Epistemological chicken. In: Pickering, A. (Ed.), Science as Practice and Culture. University of Chicago Press, Chicago, pp. 301–326.
Epstein, S., 1996. Impure Science: AIDS, Activism and the Politics of Knowledge. University of California Press, Berkeley, Los Angeles and London.
Fleck, L., 1979. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago [first published in German in 1935].
Forman, P., 1971. Weimar culture, causality and quantum theory, 1918–1927: adaptation by German physicists and mathematicians to a hostile intellectual environment. Historical Studies in the Physical Sciences 3, 1–115.
Fuller, S., 2000. Thomas Kuhn: A Philosophical History for Our Times. University of Chicago Press, Chicago.
Knorr Cetina, K.D., 1981. The Manufacture of Knowledge. Pergamon, Oxford.
Knorr Cetina, K.D., Mulkay, M. (Eds.), 1983. Science Observed: Perspectives on the Social Study of Science. Sage, London & Beverley Hills.
Kuhn, T.S., 1961. The function of measurement in modern physical science. Isis 52, 162–176.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Labinger, J.A., Collins, H.M. (Eds.), 2001. The One Culture? A Conversation about Science. University of Chicago Press, Chicago.
Lakatos, I., 1976. Proofs and Refutations. Cambridge University Press, Cambridge [originally published in 1963. British Journal for the Philosophy of Science 14, 1–25, 120–39, 221–45, 296–342].
Lakatos, I., Musgrave, A. (Eds.), 1970. Criticism and the Growth of Knowledge. Cambridge University Press, Cambridge.
Latour, B., 1987. Science in Action. Open University Press, Milton Keynes, UK.
Latour, B., Woolgar, S. 1979. First ed. Sage, London & Beverly Hills.
Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton.
Losee, J., 1980. A Historical Introduction to the Philosophy of Science. Oxford University Press, Oxford.
Lynch, M., 1985. Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory. Routledge and Kegan Paul, London.
McHugh, P., 1971. On the failure of positivism. In: Douglas, J.D. (Ed.), Understanding Everyday Life. Routledge and Kegan Paul, London, pp. 320–335.
MacKenzie, D., 1981. Statistics in Britain, 1865–1930. Edinburgh University Press, Edinburgh.
MacKenzie, D., 1998. The Certainty trough. In: Williams, R., Faulkner, W., Fleck, J. (Eds.), Exploring Expertise: Issues and Perspectives. Macmillan, Basingstoke, UK.
Mannheim, K., 1936. Ideology and Utopia: An Introduction to the Sociology of Knowledge. University of Chicago Press, Chicago.
Pickering, A., 1984. Constructing Quarks: A Sociological History of Particle Physics. Edinburgh University Press, Edinburgh.
Pinch, T.J., 1977. What does a proof do if it does not prove? In: Mendelsohn, E., Weingart, P., Whitley, R. (Eds.), The Social Production of Scientific Knowledge. Reidel, Dordrecht, The Netherlands, pp. 171–215.
Pinch, T.J., 1981. The sun-set: the presentation of certainty in scientific life. Social Studies of Science 1 (11), 131–158.
Pinch, T.J., 1986. Confronting Nature: The Sociology of Solar-neutrino Detection. Reidel, Dordrecht, The Netherlands.
Popper, K.R., 1959. The Logic of Scientific Discovery. Harper & Row, New York.
Richards, E., 1991. Vitamin C and Cancer: Medicine or Politics. Macmillan, Basingstoke, UK.
Shapin, S., 1979. The politics of observation: cerebral anatomy and social interests in the Edinburgh Phrenology Disputes. In: Wallis, R. (Ed.), On the Margins of Science: The Social Construction of Rejected Knowledge, Sociological Review Monograph, 27. Keele University Press, Keele, UK, pp. 139–178.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton.
Travis, G.D.L., 1980. On the importance of being earnest. In: Knorr, K., Krohn, R., Whitley, R. (Eds.), The Social Process of Scientific Investigation: Sociology of the Sciences Yearbook, vol. 4. Reidel, Dordrecht, The Netherlands, pp. 165–193.
Travis, G.D.L., 1981. Replicating replication? Aspects of the social construction of learning in planarian worms. Social Studies of Science 11, 11–32.
Wilson, B. (Ed.), 1970. Rationality. Blackwell, Oxford.
Winch, P.G., 1958. The Idea of a Social Science and Its Relation to Philosophy. Routledge and Kegan Paul, London.
Woolgar, S., 1976. Writing an intellectual history of scientific developments: the use of discovery accounts. Social Studies of Science 6, 342–395.
Wittgenstein, L., 1953. Philosophical Investigations. Blackwell, Oxford.
Wittgenstein, L., 1956. Remarks on the Foundations of Mathematics. Blackwell, Oxford.
Wynne, B., 1987. Risk Management and Hazardous Wastes: Implementation and the Dialectics of Credibility. Springer, Berlin.

Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct

Ragnvald Kalleberg, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Scientific misconduct refers to fraud, to serious violations of the internal research ethical norms related to the truth-commitment of science. The three most common examples used to illustrate such misconduct, are fabrication of fictitious data, falsification of data and methods, and plagiarism. The next section is focused on the historical emergence of modern research ethics in early modernity, as an integrated element in the scientific revolution. Robert Merton labeled this system of values and norms "the ethos of science." As aset of institutional imperatives it has perhaps been the most efficient system of internal self-control invented in the modern world. The last section deals with developments in research ethics after World War II. Before World War II institutional control of misconduct was informal, based on the self-correcting power of individuals and groups. Today the traditional system of self-regulation is more explicitly cultivated and supervised, complemented with procedures and committees on higher levels, such as at the level of organizations. The task of new procedures and institutions is to comple guard, control, stimulate and further develop traditional self-regulation, not to replace it.

The Phenomenon of Scientific Misconduct

What kinds of malpractices are covered by the concept of 'scientific misconduct'? There is no general definition and terminology agreed on across disciplines, institutions, and countries. Definitions vary in breadth. In its policy document on research integrity Johns Hopkins University in 1988 used a relatively narrow definition: "Research fraud is the intentional falsification or fabrication of data or results, plagiarism, misconduct in the application of research procedures so as to bias results, or other deceptive research or reporting practices" (Steneck, 1994, p. 318). The European Science Foundation (ESF, 2011) also focuses on other 'forms of misconduct,' generally identified as 'failure to meet clear ethical and legal requirements.' The ESF gives examples of serious deviations such as "misrepresentation of interests, breach of confidentiality, lack of informed consent and abuse of research subjects or materials" (Section 1.3).

It can be useful to order the norms and values of research in four groups: (1) Internal norms about the production and testing of knowledge claims. These are norms regulating the inner workings of research communities in their search for valid knowledge. Such norms also include those safeguarding the autonomy of researchers and research institutions. (2) More external norms, regulating the relationships between researchers and living beings studied, for instance individual persons, groups, institutions, and animals. (3) Norms regulating the relations to those financing and using research. (4) Norms relevant for the communication of scientific knowledge to those outside of the specialized research fields – researchers in other specialized fields included. We can label this fourth subgroup of norms the cultural and democratic obligations of science (Kalleberg, 2014).

Norms and values in the different groups are related to each other and partly overlapping. They are all relevant for the social and behavioral sciences. Good referencing behavior, free and informed consent from those studied, conflict of interests, and appropriate dissemination of scientific knowledge to those outside of a research field are examples from each group of norms. Within the social sciences there has been much discussion about norms regulating the relationships to those studied, for instance in well-known social psychological experiments, such as Stanley Milgram's authority experiments and Philip Zimbardo's prison experiment. We can label serious violations of each type of norms as 'scientific misconduct.' In this article the expression 'scientific misconduct' is used in a more narrow meaning, primarily referring to serious violations of norms in the first group. When such norms are violated, words like fraud and dishonesty are often used. Norms and values in this group have to do with the main task of scientific and scholarly activity, its truth commitment.

There is general agreement about examples of major violations of the truth and truthfulness commitment of science. Three types of violations are regularly mentioned, often talked about as FFP. The acronym refers to fabrication of fictitious data, falsification of data and methods, and plagiarism. During the last quarter of a century several misconduct cases have been uncovered and have got much publicity. The best known cases have taken place at excellent research institutions in North America, Europe, and Asia (for some well-known cases, see Anderson et al., 2013, pp. 229–231). It is usual that such cases contain several types of misconduct, such as fabrication, falsification, and plagiarism, and also less serious violations of norms.

Most empirical studies of misconduct have been about the biomedical sciences, and most have been focused on researchers in the United States. There are, however, no convincing reasons to assume that the social sciences are special with regard to scientific misconduct. Serious misconduct and questionable practices have also been documented in these disciplines (see for instance Israel and Hay, 2006; Necker, 2014).

Fabrication refers to fabrication of fictitious data and results, whereas falsification refers to inappropriate manipulation of data, methods, and results (see for instance Israel and Hay, 2006, Chapter 8). Plagiarism concerns authorship. In the *Oxford English Dictionary*, plagiarism is defined as the "action or practice of taking someone else's work, idea, etc., and

passing it off as one's own." Obviously, the criticizable element is not to use the ideas and words of others, but to present them as ones own. Posner (2007, p. 6) identifies plagiarism as a type of 'intellectual fraud' and gives the following, more elaborate definition: "unauthorized copying that the copier claims (whether explicitly or implicitly, and whether deliberately or carelessly) is original with him /sic/ and the claim causes the copier's audience to behave otherwise than it would if it knew the truth." (For definitions of plagiarism and a typology of plagiarism, see Weber-Wulff, 2014, pp. 3–14.)

In Europe, some cases of plagiarism have become widely known during the past years, due to the fact that leading politicians have been forced to leave their positions because plagiarism was discovered in their doctoral dissertations. The two most spectacular cases have been that of the German Minister of Defense in 2011, and the German Minister of Education of Research in 2013 (Weber-Wulff, 2014, pp. 29–31, 57, 58).

Fabrication, falsification, and plagiarism do not represent an exhaustive typology of violations related to the truth and honesty commitment of science. Failure to report conflict of interest, fictitious authorship (so-called gift and ghost authorship), change of design or results because of pressure from institutions financing the research are examples of related violations. If we underline originality as a basic internal value in science, phenomena such as *salamisation* (to divide results up in so small bits as possible in order to increase the number of publications) and self-plagiarism would also be relevant examples.

Scientific and scholarly research is an institutional phenomenon, regulated by norms and values practiced within scientific networks. Such communities often include researchers and research groups spread out around the globe. Norms regulating research ethics consequently also refer to sets of norms and values of a transnational and institutional character, and not only to the practices and ideals of individual researchers in a certain location. There can be no private science or scientific knowledge, nor a science based on dignified sages disseminating esoteric knowledge that is not to be controlled by competent and independent peers. Peer review and replication of results from others are typical examples of this institutional character. Ziman (2000, p. 4) gives this characterization of the institution of science: "Although research *scientists* often have a great deal of freedom in what they do and how they do it, their individual thoughts and actions only have scientific meaning in these larger schemes. Like many facts of life, this is so obvious that it was for long overlooked … The sociological dimension is thus fundamental to our picture." It is still regularly overlooked by researchers and institutions when the focus is on research ethics. But misconduct, or good conduct, is also an institutional phenomenon, not only a matter of the morality of individual researchers.

Widely known cases of misconduct are often identified with only one name, the main person responsible for the misconduct, the single bad apple in a big basket of good apples. In investigations of misconduct cases, however, also several other problems in the relevant scientific community are regularly discovered. They can be identified as system problems or as questionable research practices (QRP) (see Steneck, 2006, pp. 59–61). Examples are sloppy control by coauthors, not thorough enough checking by opponents, advisors, and reviewers, and lack of relevant teaching and mentoring in institutions. Cases of misconduct regularly go hand in hand with questionable practices. There are also examples of institutions covering up misconduct instead of uncovering, handling, and preventing it. This is a kind of serious institutional misconduct. Institutions should instead establish research ethical committees, install procedures informing about where and how to report suspicion of possible misconduct, develop routines for protection of whistleblowers, procedures for handling of misconduct cases, and programs for promoting good conduct and preventing misconduct.

The Emergence of an Ethos of Science and Institutionalized Self-Regulation

The most influential analysis of the ethics of research as an institutional phenomenon where researchers correct each other has been given by Robert Merton. He labeled the phenomenon 'the ethos of science,' referring primarily to the internal norms and values connected to the truth commitment of science. He first published his analysis in two small articles (Merton, 1938, 1942), later followed by several others (as Merton, 1957). His analysis was both based on an extensive historical understanding of the emergence of modern science in the seventeenth century, and reflections based on the undermining and destruction of the ethos of science in the twentieth century, by market forces, nationalism, Nazism, and Communism (Kalleberg, 2007; Enebakk, 2007).

Merton identified the ethos of science like this: "The ethos of science is that affectively toned complex of values and norms which is held to be binding on the man /sic/ of science" (1942, pp. 268, 269; 1938, p. 258). This seminal definition was characterized by the triple perspective of sociological imagination (see Mills, 1959), comprising the interplay of *individuals* and *institutions* embedded in *historical processes* (biographies, traditions). To underline this sociological perspective Merton used the expression 'institutional imperative' as a synonym for 'norm.' The problem with several later definitions of research ethical norms, including many who criticized Merton, has been a too narrow orientation only to individuals and their ideals (or scientific conscience), not at the same time seeing the institutional framework disciplining its members, and the tensions between norms and counter-norms, including the complexity of role-sets embedded in evolving fields of research (Merton, 1963; Kalleberg, 2014).

Merton operated with a traditional, sociological typology of social norms. "The norms are expressed in the form of prescriptions, proscriptions, preferences and permissions" (1942, p.269). The norms and values of the ethos of science are in operation as a set of shared and internalized norms and values, institutional imperatives, characterizing well-functioning scientific communities. Merton did not claim that his typology was exhaustive, nor that the different types were exclusive. They are clearly internally related, partly overlapping each other.

Merton's scientistic interpretation of norms in science, basing them on emotions and not reasons, is not tenable. But that can be corrected at the same time as one preserves the superior insights and realism in his analysis (see Kalleberg, 2007, 2009). Merton's contributions illustrate

a typical phenomenon in the social and cultural sciences: classics are not only of historical interests, but also keep on being of contemporary and systematic interest (see Calhoun, 2010).

The ethos of science in Merton's conception can be presented as a set of six norms, the CUDOSH-norms: Communalism, Universalism, Disinterestedness, Originality, Skepticism and Humility (see Kalleberg, 2007, p. 141, 142). *Communalism* refers to the fact that scientific activity is a sociological phenomenon; there can be no isolated, private science. Knowledge and insight have to be published, made public, in order to be tested and recognized by the collective of peers. "Property rights in science are whittled down to a bare minimum by the rationale of the scientific ethic" (Merton, 1942, p. 273). *Universalism* concerns issues of scientific validity and refers to "the canon that truth claims, whatever their source, are to be subjected to *pre-established impersonal criteria*; consonant with observation and with previously confirmed knowledge" (Merton, 1942, p. 269, italics in original). This institutional imperative can also be characterized as the pursuit of truth, the truth-commitment of science. In a later article Merton explained the relatively few cases of scientific fraud, as compared to other fields, by the institutional "emphasis on the value of truth by whomsoever it is found, and a commitment to the disinterested pursuit of truth" (Merton, 1957, p. 321). *Disinterestedness* concerns impartiality in posing of questions, collection and interpretation of data, and impartiality in argumentation. Science is done for the sake of valid and reliable knowledge.

Originality requires that contributors in well-functioning scientific communities are obliged to present new knowledge and insight, and not only preserve the knowledge of a tradition. As a historical sociologist of the seventeenth century scientific revolution, Merton had a keen eye for identifying the historically new in the insistence on originality. *Skepticism* refers to the obligation to check the validity of documentation and argumentation, not just believing what authors claim. Skepticism became institutionalized in many ways, two of them being the scrutiny of peers reading the contributions and the mechanism of peer review. The basic precept can be formulated like this: only reliable documentation and consistent argumentation shall influence the formation of opinions in well-functioning scientific communities. *Humility* refers to the institutional obligation to understand how little each researcher and discipline understand of a complex reality, and how dependent each individual researcher is on the transnational community of scholars. This intellectual dependence does not only include contemporaries, but also earlier and coming generations. The age-old metaphor often used to convey the meaning of scientific humility in this strict meaning, is the following – here quoted as expressed by one of the greatest scientists of all time, Isaac Newton: "If I have seen farther it is by standing on the shoulders of giants" (Merton, 1942, p. 274, 275).

The ethos of science – like other institutional ethoses, for instance the protestant work ethic (Weber) – does not have to be codified in order to influence action and interaction within research groups and disciplines. The research ethic of science can be characterized as the oldest internal control system in the modern world, as old as modern science. Scientific communities are unthinkable without it; research ethics is internal to science. For Merton, there was little doubt that all in all the amount of fraud was smaller within the social system of science than in other institutions in a modern society. The relatively small amount of deviance in science is due to an institutional trait, its high degree of transparency. If one should try to understand the phenomenon strictly individualistically, without sociological imagination, one would have to rely on an unrealistic presupposition of researchers as moral superbeings. Merton (1957, p. 311) gave this general explanation for the effectiveness of this kind of interactional self-correction:

> Scientific research is typically, if not always, under the exacting scrutiny of fellow experts, involving as it usually, though not always does, the verifiability of results by others. Scientific inquiry is in effect subject to rigorous policing, to a degree perhaps unparalleled in any other field of human activity. Personal honesty is supported by the public and testable character of science.

With the historical emergence of the ethos of science came also violations of those norms and corresponding conflicts. Merton (1957) mentions several examples of fraud in the history science, some of them based on the well-known book by Babbage (1830) on fraud in science. Merton's most detailed discussions of scientific misconduct were focused on plagiarism and priority struggles. When the requirement for originality emerged as part of the scientific revolution, also the first struggles about priority and plagiarism also emerged. An early example is the bitter conflict between Newton and Leibnitz about the invention of the calculus (Merton, 1957, pp. 314, 315). Merton also noted the possible distorting effects of the institutional obligation of originality "in many academic institutions, to transform the sheer number of publications into a ritualized measure of scientific or scholarly accomplishment" (1957, p. 326).

Codification and Institutional Regulation of Research Ethics after World War II

Before World War II institutional control of research ethical norms was informal, based on the self-correcting power of scientific communities. Extensive codification of research ethics and institutional strategies to handle misconduct and promote research integrity, are developments characteristic of the period after World War II. The development came first in medical ethics, oriented to the more external norms of science, focused on the protection of individuals studied. A second wave started in the 1980s in the United States, focused on fraud and deception, concerned with violations of the first set of research ethical norms regulating the inner workings of scientific communities.

A classic sociological insight, associated with the work of Emile Durkheim, can be articulated like this: deviation from a norm can make visible and strengthen that same norm. Merton expressed it like this: "Yet this very deviation from the norm of universalism actually presupposed the legitimacy of the norm … in the process of contemning their violation, the mores are reaffirmed" (Merton, 1942, p. 271). It is worth noting that Merton's (1938, 1942) articles actually were self-exemplifying. He understood that the emerging anti-Semitism in Germany

after World War I perverted leading scientists and institutions, also before Hitler came to power in 1933. For instance, Nobel Prize winners in physics (Philipp Lenard, Johannes Stark) in the early 1920s rejected the work of Einstein because it was produced by a Jew ('Jewish' vs 'Aryan' physics) (Walker, 1995, Chapter 2). Such a violation of the norm of universalism in the world's leading science nation shocked the young American sociologist. He felt compelled to explicate the norms in order to identify, criticize, and correct the misconduct. Karl Popper once characterized his *The Open Society and Its Enemies* as his war effort, his intellectual contribution to promote freedom and democracy. Merton's contributions on the ethos of science can be characterized in the same way.

Violation of basic norms in science, such as fabrication of fictitious data and falsification of data, is dramatic and destructive, arousing strong emotions of anger and sorrow. But cases of scientific misconduct can help science institutions to better see the norms and their importance, and safeguard them institutionally. The strong emotional reactions to such violations can have a cognitive function in such socio-cognitive processes (Kalleberg, 2009, pp. 259–262).

The history of research ethics after World War II confirms the general sociological insight about violations also strengthening values violated. Widespread codification of the ethos was first developed after World War II and the atrocities of German medical doctors during the Nazi period. It started with the so-called Doctors' Trial, taking place in Nuremberg in 1946–47. The defendants were medical doctors. They were convicted because of cruel experimentation and murder of prisoners in concentration camps. The trial led to the *Nuremberg Code* in 1947. The focus was on the more external norms regulating the relationships between researchers and those studied. One central guideline explicated for medical research, was the requirement of free and informed consent from patients and other persons participating.

The World Medical Association (WMA) was established in 1947. After 10 years of preparation, the Association in 1964 adopted *The Declaration of Helsinki*. It was built on the Nuremberg code and has established itself as the best known and most influential statement for medical ethics concerning informing and protecting human beings participating in such research. This code has influenced legislation in the member countries of WMA. The US *Belmont Report* (1978) was produced by the 'National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.' It came as a response to serious cases of missing or inadequate information of research subjects. A fourth code was developed by the *Council for International Organizations of Medical Sciences*. This code is focused on guidelines for international medical research, also involving participation by less developed countries. Guideline 2 is about review committees – be they located on the institutional, local, regional, national, or international level. All proposals "to conduct research involving human subjects must be submitted for review of their scientific merit and ethical acceptability to one or more scientific review and ethical review committees." Projects must be approved before start-up (on these four codes, see Israel and Hay, 2006, Chapter 3).

A second wave of institutional control of ethical practices and procedures was started in the 1980s in United States. It was focused on fraud and deception, concerned with violations of the truth-commitment of science, the first set of research ethical norms presented above. This wave has generally consisted in measures introduced to secure and further develop the ethical self-correction capacities of individual researchers and research groups, for instance by developing codes, introducing teaching arrangements, or installing local and national research ethical committees to handle misconduct cases. Such measures on an institutional and national level are not introduced to replace the traditional self-correcting and self-regulating activities of individuals and groups, but to complement and strengthen them.

The reforms came as a consequence of a new understanding of the prevalence of scientific misconduct and questionable research practices. Up to the 1980s it was a widespread assumption that scientific fraud was extremely rare. In 1987 the editor of the prestigious journal *Science* even suggested only one case in a million (Koshland, 1987). During the 1990s it became more common in the United States to claim that misconduct was more widespread than this. Some estimates were based on reported cases that had been investigated by institutions like the Office for Research Integrity (ORI). Such studies indicated that the prevalence was one case in 100 000 researchers. Other studies were based on misleading surveys producing too high estimates, and talk about the tip of the iceberg. With reference to new empirical studies, including anonymous questionnaires where individual scientists informed about their own misconduct, Steneck suggested the "level of occurrence for serious misconduct" was rather one in 100 (2006, p. 58). Several recent studies indicate the same order of magnitude, between 1 and 3% (Fanelli, 2009; Anderson et al., 2013, pp. 232–235). Obviously it is very difficult to give a reliable estimate of this type of phenomenon. But we now have good grounds to claim that scientific misconduct today is not rare. One to three researchers among one hundred having at least once been responsible for scientific fraud is an alarmingly high level.

There are good reasons to assume that the amount of misconduct varies between the different types of fraud. The amount of fabrication of fictitious data, for instance, is in all probability much smaller than the amount of plagiarism. There are probably also considerable variation between disciplines, institutions, and countries. There is still little systematic knowledge about this. In comparison with the United States, the number of empirical studies of this phenomenon is low in Europe. And we have even fewer studies in other regions of the world.

Before World War II the relationships among basic research, politics, and commercial activities, were looser and more indirect than they are today. After World War II we have seen an enormous growth of science, and science has got an increased strategic and commercial relevance. To characterize the development Ziman talks about the emergence of 'post-academic science' (Ziman, 2000, pp. 67–82). The new structures have put the ethos of science under new pressure, and have probably increased the prevalence of scientific dishonesty during the past 30–40 years, although that is difficult or impossible to document.

The traditional ethos of science is an internal and informal system of self-regulation, a set of institutional imperatives

operating at the level of researchers, groups, and networks of individuals. In a recent contribution this institutional culture of self-correction was analyzed and its continuing importance underlined (Anderson et al., 2013, pp. 220–222, 253). And Ziman organizes his book (2000) around the Mertonian norms and insists on the increased importance of research ethics in the 'post-academic' situation (Ziman, 1998). Today this traditional system of self-regulation at the level of individuals and groups has not only to be more explicitly cultivated, taught, and supervised, but also has to be complemented with procedures, positions, and committees on higher levels, such as at the level of organizations, for instance with research integrity *ombud* and ethical committees in universities. Codes and institutions at the national (such as the Norwegian research ethical committees, see www.etikkom.no) and international level (see ESF, 2011; Singapore statement, 2010; Montreal statement, 2013) are also needed. The task of the new systems is to complement, guard, control, and further develop traditional self-regulation in the research communities, not to replace them. A legal–administrative replacement system, for instance, would pervert the inner workings of scientific communities and create more problems than those solved (Kalleberg, 2015). The ethos of science can only function optimally at the field-specific level of working scientists, as an essential and integrated element in their scholarly and scientific work.

Bibliography

Anderson, Melissa, Shaw, Marta, Steneck, Nicholas, Konkle, Erin, Kamata, Takehito, 2013. Research integrity and misconduct in the academic profession. In: Paulsen, Michael (Ed.), Higher Education: Handbook of Theory and Research. Springer, Dordrecht, pp. 217–261.

Babbage, C., 1830. The Decline of Science in England, and on Some of Its Causes. London.

Calhoun, Craig (Ed.), 2010. Robert K. Merton. Sociology of Science and Sociology as Science. Columbia University Press, New York.

Enebakk, Vidar, 2007. The three Merton theses. Journal of Classical Sociology 7 (2), 221–238.

ESF (European Science Foundation), 2011. Fostering Research Integrity in Europe. www.esf.org.

Fanelli, Daniele, May 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 5 (5). www.plosone.org.

Israel, Mark, Hay, Iain, 2006. Research Ethics for Social Scientists. Between Ethical Conduct and Regulatory Compliance. Sage, London.

Kalleberg, Ragnvald, 2007. A reconstruction of the ethos of science. Journal of Classical Sociology 7 (2), 137–160.

Kalleberg, R., 2009. Can normative disputes be settled rationally? On sociology as a normative discipline. In: Cherkaoui, M., Hamilton, P. (Eds.), Raymond Boudon. A Life in Sociology, vol. 2. The Bardwell Press, Oxford, UK, pp. 251–269.

Kalleberg, R., 2014. The role of public intellectual in the role-set of academics. In: Fleck, C., Hess, A. (Eds.), Knowledge for Whom? Public Sociology in the Making. Ashgate Publishing Company, Surrey, pp. 253–274.

Kalleberg, R., 2015. Plagiarism as violation of law in Norway. On inappropriate juridification of research ethics. In: Steneck, N.H., Anderson, M.S., Kleinert, S., Mayer, T. (Eds.), Integrity in the Global Research Arena. World Scientific Publishing Company, New Jersey, pp. 129—144.

Koshland, Daniel E., 1987. Fraud in science. Science 235, 141.

Montreal Statement on Research Integrity, 2013. http://www.icsu.org/icsu-asia/news-centre/news/montreal-statement-on-research-integrity.

Merton, Robert K., 1973. The Sociology of Science. Theoretical and Empirical Investigations. The University of Chicago Press, Chicago.

Merton, R.K., 1938. Science and the Social Order, in Merton 1973, 254–266.

Merton, R.K., 1942. The Normative Structure of Science, in Merton 1973, 267–278.

Merton, R.K., 1957. Priorities in Scientific Discovery, in Merton 1973, 286–324.

Merton, R.K., 1963. The Ambivalence of Scientists, in Merton 1973, 383–412.

Merton, R.K., 1990. STS: foreshadowings of an evolving research program in the sociology of science. In: Cohen, I.B. (Ed.), Puritanism and the Rise of Modern Science. Rutgers University Press, London, pp. 334–371.

Mills, Charles W., 1959. The Sociological Imagination. Grove Press, New York.

Necker, Sarah, 2014. Scientific misbehavior in economics.

Posner, Richard, 2007. The Little Book of Plagiarism. Pantheon Books, New York.

Singapore Statement on Research Integrity, 2010. http://www.singaporestatement.org.

Steneck, Nicholas, 1994. Research universities and scientific misconduct: history, policies, and the future. The Journal of Higher Education 65 (3), 310–330.

Steneck, N., 2006. Fostering integrity in research: definitions, current knowledge, and future directions. Science and Engineering Ethics 12, 53–74.

Walker, Mark, 1995. Nazi Science. Myth, Truth, and the German Atomic Bomb. Plenum Press, New York.

Weber-Wulff, Deborah, 2014. False Feathers. A Perspective on Academic Plagiarism. Springer Verlag, Berlin.

Ziman, John, December 1998. Why must scientists become more ethically sensitive than they used to be? Science (5395), 1813–1814.

Ziman, J., 2000. Real Science. What It Is and What It Means. Cambridge University Press, Cambridge.

Scientific Revolution, History and Sociology of

Jürgen Renn and Peter Damerow[†] Max Planck Institute for the History of Science, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This contribution analyzes the material, social, and cognitive dimensions of the Scientific Revolution as independent and irreducible dimensions. In discussing the social dimension, the contribution will delineate the impact of the great engineering projects of the Renaissance on the formation of a new social group, dealing both with the practical problems and with the theory of nature. In discussing the material dimension, it will address the intellectual challenges represented by new objects of contemporary technological practices and, in discussing the intellectual dimensions, it will show how mental models of traditional thinking merge with those of practical thinking to form the basis for what eventually will become the conceptual novelties generated by the Scientific Revolution.

The Term 'Scientific Revolution'

The term 'Scientific Revolution' was introduced to denote what has been considered since the nineteenth century as one of the most important discontinuities in the history of European science (H.F. Cohen, 1994; I.B. Cohen, 1985; Lindberg and Westman, 1990). It covered roughly the period between Copernicus and Newton, and led from Aristotelian natural philosophy (*see* Aristotle (384–322 BC)) deriving its dogmatic authority from the Church to the establishment of the classical sciences and their institutions, i.e., the period between 1500 and 1700 (*see* Scientific Disciplines, History of).

This period can be characterized as a period in which a new social group, the 'engineer-scientists,' (consisting of engineers, scientists, inventors, artists, and explorers) emerged and became institutionalized. This group confronted traditional natural philosophy with the challenges of practice and experience, but also engaged in self-contained explanations of natural phenomena, expecting that science is a means to master nature, as Francis Bacon puts it. Among the lasting achievements of the Scientific Revolution were the establishment of heliocentric astronomy, classical mechanics, as well as numerous contributions to optics, chemistry, physiology, and other areas of modern science (for overviews, see Butterfield, 1965; Daston and Park, 2006; Dear, 2001; Dijksterhuis, 1986; Hall, 1954; Rossi, 2001). Many of these achievements became, in fact, the basis for technological breakthroughs. However, these breakthroughs occurred, as a rule, much later than anticipated by the protagonists of the Scientific Revolution. Further, the intellectual breakthroughs responsible for the revolution's lasting impact on the development of science were mainly attained toward its end and only a generation or more after their proclamation by the early pioneers (Damerow et al., 2004).

The Scientific Revolution has become a paradigmatic reference for all approaches to the history and philosophy of science (*see* History of Science). At least since the Enlightenment, it has been conceived as the triumph of the scientific method over the irrationalism of religious beliefs (*see* Science and Religion). Opposition to this view, together with a growing professionalization of historical studies, has opened up the space for other accounts, including attempts – such as that by Pierre Duhem – to deny that the Scientific Revolution actually represented a radical break, claiming instead that it merely constituted an episode within a continuous accumulation of scientific knowledge since the Middle Ages (Duhem, 1996). Furthermore, the traditional emphasis on the role of outstanding protagonists of the Scientific Revolution, such as Bacon, Galileo, and Descartes, and their individual 'discoveries' (see, e.g., Koyré, 1978) has, in recent scholarship, increasingly receded in favor of an analysis of contexts tracing scientific achievements back to cultural, social, and economic conditions (Osler, 2000; Porter and Teich, 1992; Shapin, 1996). The Scientific Revolution, on the other hand, itself provided a model for historical explanations when Thomas Kuhn (*see* Kuhn, Thomas S. (1922–96)) radically challenged traditional historiography with his notion of 'scientific revolution' conceived of as a general structure of knowledge development (Kuhn, 1962).

Historical and Sociological Context

From a sociological perspective, the Scientific Revolution appears as part of a wider social process in which technical knowledge assumed a new role in the organization of European societies (*see* Renaissance; Enlightenment). This process took off in the late Middle Ages and was primarily rooted in the larger cities, which saw an ever more diversified and developed artisanal culture and a growing accumulation of merchandise capital. The cities of early modern Europe thus offered favorable conditions to the rapid growth of technical knowledge and the reflection of this growth in political, philosophical, and religious thinking. Large-scale ventures involving technical expertise, such as projects of military architecture, water regulation, or seafaring expeditions involved types of resources, social mobility, and an outlook on the world available only in urban centers such as Florence, Venice, Paris, and London, which in fact became, long after they had attained an outstanding economic role, also the nuclei of the Scientific Revolution. Historians such as Edgar Zilsel (2000) have considered the early modern engineering projects as a decisive condition for the practical orientation and the empirical knowledge base that distinguish the science of this period from its medieval

[†]Deceased.

antecedents (see Freudenthal and McLaughlin, 2009; Long, 2011; Olschki, 1965; Smith, 2004; Valleriani, 2010).

Beginning in the fifteenth century, ambitious practical ventures (such as the building of the cupola of the Florentine Cathedral, the search for a sea route to India, or the development of a new military technology) increasingly relied on expert knowledge comprising both logistic and technological competencies exceeding those of traditional practitioners and artisans. Such competencies could only be gained on the basis of a broader reflection on the relevant practical and theoretical knowledge resources available. This reflection became the specialty of the new group of engineer-scientists such as Filippo Brunelleschi, Christopher Columbus, Leonardo da Vinci, Niccolo Tartaglia, and Galileo Galilei. While the states of the time (if not actually at war) competed with each other in the pursuit of practical ventures, for example, building projects or seafaring expeditions, the knowledge acquired in such ventures was nevertheless constantly spread among the engineer-scientists employed by them. In the social fabric of the early modern period, engineer-scientists occupied a place similar to that earlier conquered by Renaissance humanists, administrators, and artists (*see* Art History). These practically oriented intellectuals were, as a rule, highly mobile, offering their services to whatever patronage was available. At the same time, they constituted, as a social group, a collective memory, accumulating and transmitting the new knowledge long before appropriate institutions of learning emerged and in spite of the frequent political and military turnovers of the time.

The characteristic features of the engineer-scientists and their work become understandable against the background of their uncertain social status and their dependence on the patronage of courts and city governments (see, e.g., Biagioli, 1993) with rapidly changing power structures. Examples are their incessant engagement with projects for potential future patrons; their usually unrealistic promises regarding the practical benefits of their theories, inventions, or projects; the secrecy with which they treated their discoveries; their frequent involvement in priority struggles; as well as the striving to ennoble their practical knowledge with the claim of creating 'new sciences.'

The social and political ambitions of the engineer-scientists were reflected in their pursuit of a literary culture of technical knowledge, largely emulating the humanist culture of the courts, including their reference to ancient Greek and Roman canons. Their contribution to the literary culture was in turn welcomed by those interested in challenging the transcendent, religious legitimization of the feudal order as an argument for the possibility of an immanent explanation of both the natural and the social world. This affinity, together with the fact that an all-encompassing explanation of the world on the basis of Aristotelian philosophy had been adopted as the official doctrine of the Catholic Church, brought the engineer-scientists almost unavoidably into conflict with its power structures (Feldhay, 1995).

Other developments contributed to turning the growth of technological and scientific knowledge into a force driving profound changes in the established structures of European society. The invention of printing offered a revolutionary new means of dissemination that challenged the exclusiveness of a literary culture based on manuscripts. The new dissemination technique contributed to overcoming the traditional separation between various branches of practical knowledge, confined by a transmission process relying exclusively on participation and oral communication as well as restricted by guild regulations. It also bridged the social gulf between such practical knowledge and the theoretical knowledge transmitted via scholarly texts at universities, monasteries, and courts. As a result, knowledge resources of different provenance were integrated and became widely available. Together with the accumulation of new knowledge in the context of the great practical ventures of the time, this process formed part of a veritable explosion of knowledge, both in the sense of its expansion and of its spreading across traditional social barriers.

In reaction to this knowledge explosion and its growing significance for the functioning of the advanced European societies, new institutions of learning such as the Accademia del Cimento (1657), the Royal Society (1660), and the Académie Royale des Sciences (1666) emerged in the middle of the seventeenth century (*see* Knowledge Society, History of). Traditional institutions, such as those of the Church, on the other hand, had to accommodate to the new situation, as may be illustrated by the prominent involvement of Jesuits in the scientific culture of the time (Wallace, 1991). Parallel to this process of institutionalization, science had, toward the end of the period here under consideration, gradually emancipated itself from the expectation of immediate practical benefits and could increasingly be pursued for its own sake. In summary, as a result of the Scientific Revolution, not only the production and transmission of technological knowledge but also its reflection by scientific theories became an essential factor in the social, economic, and cultural development of European societies. This is the context in which scientists such as Huygens, Leibniz, Hook, Newton, and Wallis traditionally identified with the completion of the Scientific Revolution by the creation of classical terrestrial and celestial mechanics, achieved their celebrated results.

Structures and Achievements

The cognitive, social, and material structures and achievements of the Scientific Revolution have been extensively studied and discussed amid much controversy. Recent scholarship in the context of a cultural history of science has emphasized the possibility of an integrative treatment of these various dimensions. From the point of view of the traditional history of ideas, the Scientific Revolution appears primarily as the renewal of a knowledge development going back to antiquity, as a renaissance of Greek science. In fact, the ancient tradition of mathematical science, and in particular the 'Elements' of Euclid, provided the protagonists of the Scientific Revolution with the canonical model for a mathematical theory, a model which they systematically applied to new areas such as ballistics (Tartaglia), (see Valleriani, 2013) and even ethics (Spinoza). But the ancient tradition also had in stock designs for a theory of nature not based on Aristotelian views which they could hence exploit in their struggle against scholasticism, in particular Platonism and atomism. The ancient tradition finally offered a substantial corpus of knowledge in such domains as geometry, mathematical astronomy, and mechanics, serving as

a point of departure for new scientific endeavors. The perception of their work as a renewal of antique science was typical of the self-image of the protagonists of the Scientific Revolution, who honored each other by such titles as that of a 'new Archimedes.' In short, the characterization of the Scientific Revolution as a renaissance of antique science by historians of ideas such as Koyré is in agreement the claims of contemporary scientists to have accomplished a radical break with Aristotelian scholasticism (Koyré, 1978).

This characterization, however, is in apparent conflict with the results of studies inaugurated by scholars in the tradition of Duhem, pointing to a conceptual continuity between early modern science and its medieval predecessors (and even with contemporary scholasticism), in spite of the anti-Aristotelian polemics (Duhem, 1996). Such studies have provided evidence that the intellectual means available to the engineer-scientists engaged in creating new sciences were essentially still rooted in traditional conceptual frameworks. For instance, the investigations of motion and mechanics by Galileo Galilei and his contemporaries were shaped by such Aristotelian notions as the distinction between natural and violent motion and the assumption that violent motion is caused by a moving force (Damerow et al., 2004; Renn and Damerow, 2012). Studies of medieval natural philosophy have revealed, on the other hand, that the advanced explanations of phenomena such as projectile motion, by which early modern scientists distinguished themselves from Aristotelian natural philosophy, make use of concepts of causation such as that of 'impetus' (or 'impressed force') and techniques for conceptualizing change such as Oresme's doctrine of the 'latitude of forms' that had already been developed by late antique or medieval commentators on Aristotle (Clagett, 1968).

Such contrasting accounts of early modern science appear less incompatible if other dimensions of the development of scientific knowledge are taken into account. For instance, epistemological considerations have suggested that one should differentiate between the claim of Renaissance engineer-scientist to have created a new science or a new scientific method, on the one hand, and the knowledge base they shared with their contemporaries, on the other. Using terminology introduced by Elkana (1988) we can say that their ambitious claim resulted from the 'image of knowledge,' which was determined by their fragile social status. This image is to be distinguished from the shared 'body of knowledge' comprising the antique and medieval heritage which determined what challenges they were able to master. The role of this shared knowledge for the Scientific Revolution has been analyzed not only from the point of view of the history of ideas, but also from the viewpoint of its actual function in the social and material contexts of this revolution. It has thus turned out that the material culture of the Scientific Revolution decisively shaped the way in which the knowledge of antique and medieval science was taken up or newly interpreted.

It has become evident, for instance, that the knowledge resources available to the engineer-scientists, largely structured by traditional conceptual frameworks, were challenged by their application to the new objects of a rapidly expanding range of experience, acquired in the context of the practical ventures in which the engineer-scientists were engaged. While investigations of 'challenging objects' such as the pendulum, the trajectories of artillery, purified metals, the dissected human body, the terrestrial globe, and the planetary system often remained less successful than their protagonists hoped and claimed, they nevertheless triggered elaborations and modifications of these traditional frameworks creating the foundations for the accomplishments of the age of classical science. Columbus searched for the sea route to India, but discovered America. Kepler tried in the Pythagorean tradition to unravel the harmonies of the world but from the harmonies he found only three laws of planetary motion that became the starting point of Newtonian cosmology. Galileo explained ballistics and pendulum motion on the basis of the impetus concept, but in fact he contributed to a new mechanics, which expelled this concept once and for all from science. Harvey defended Aristotle's claim of the primacy of the heart among the organs, but became the founder of an anti-Aristotelian theory, a mechanistic medicine. All these achievements resulted from coping with challenging new objects while relying on essentially traditional intellectual means.

Classical science, often considered an accomplishment of the Scientific Revolution, was actually established for the most part only in the course of the eighteenth century. Even classical mechanics, the pilot and model science of the Scientific Revolution, assumed the formulation familiar from today's physics textbooks only in the aftermath of Newton's pioneering work. The characteristics of classical science, relatively stable theoretical frameworks and generally accepted standards for the production of knowledge serving as the canonical reference for a scientific community, were, in any case, not yet shared by the scientific endeavors of the Scientific Revolution.

The conceptual framework of classical science comprising basic concepts such as inertia, a new methodical canon including the experimental method, and mathematical techniques such as differential calculus, new images of knowledge such as the mechanization of the worldview, and the new institutions of classical science such as the academies, were nevertheless consequences of the achievements of the Scientific Revolution, as a belated result of reflection on its accumulated experience.

See also: Evolution: Diffusion of Innovations; Innovation, Theory of; Kuhn, Thomas S. (1922–96); Physical Sciences: History and Sociology; Science, History of.

Bibliography

Biagioli, M., 1993. Galileo, Courtier: The Practice of Science in the Culture of Absolutism. University of Chicago Press, Chicago.

Butterfield, H., 1965. The Origins of Modern Science 1300–1800. The Free Press, New York.

Clagett, M.H. (Ed.), 1968. Nicole Oresme and the Medieval Geometry of Qualities and Motions. University of Wisconsin Press, Madison.

Cohen, H.F., 1994. The Scientific Revolution: A Historiographical Inquiry. University of Chicago Press, Chicago.

Cohen, I.B., 1985. Revolution in Science. Belknap Press, Cambridge.

Damerow, P., Freudenthal, G., MacLaughlin, P., Renn, J., 2004. Exploring the Limits of Preclassical Mechanics, second ed. Springer, New York.

Daston, L., Park, K., 2006. The Cambridge History of Science, vol. 3. Cambridge University Press, New York.

Dear, P., 2001. Revolutionizing the Sciences: European Knowledge and Its Ambitions, 1500–1700. Princeton University Press, Princeton, NJ.

Dijksterhuis, E.J., 1986. The Mechanization of the World Picture: Pythagoras to Newton. Princeton UniversityPress, Princeton, NJ.
Duhem, P., 1996. Essays in the History and Philosophy of Science. Hackett, Indiapolis.
Elkana, Y., 1988. Experiment as a second order concept. Science in Context 2, 177–196.
Feldhay, R., 1995. Galileo and the Church: Political Inquisition or Critical Dialogue? Cambridge University Press, Cambridge.
Freudenthal, G., McLaughlin, P. (Eds.), 2009. The Social and Economic Roots of the Scientific Revolution: Texts by Boris Hessen and Henryk Grossmann. Springer, Dordrecht.
Hall, A.R., 1954. The Scientific Revolution 1500–1800: The Formation of the Modern Scientific Attitude. Longmans, Green, London.
Koyré, A., 1978. Galileo Studies. Humanities Press, New Jersey.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Lindberg, D.C., Westman, R.S. (Eds.), 1990. Reappraisals of the Scientific Revolution. Cambridge University Press, Cambridge.
Long, P.O., 2011. Artisan/Practitioners and the Rise of the New Sciences, 1400–1600. Oregon State University Press, Corvallis, OR.
Olschki, L., 1965. Geschichte der neusprachlichen wissenschaftlichen Literatur, 3 vols. Kraus Reprint, Vaduz.
Osler, M.J. (Ed.), 2000. Rethinking the Scientific Revolution. Cambridge University Press, Cambridge.
Porter, R., Teich, M. (Eds.), 1992. The Scientific Revolution in National Context. Cambridge University Press, Cambridge, UK.
Renn, J., Damerow, P., 2012. The Equilibrium Controversy. Guidobaldo del Monte's Critical Notes on the Mechanics of Jordanus and Benedetti and their Historical and Conceptual Background. Edition Open Access, Berlin.
Rossi, P., 2001. The Birth of Modern Science. Blackwell, Oxford.
Shapin, S., 1996. The Scientific Revolution. University of Chicago Press, Chicago.
Smith, P.H., 2004. The Body of the Artisan: Art and Experience in the Scientific Revolution. University of Chicago Press, Chicago.
Valleriani, M., 2010. Galileo Engineer. Springer, Dordrecht.
Valleriani, M., 2013. Metallurgy, Ballistics and Epistemic Instruments: The Nova Scientia of Nicolò Tartaglia, new ed. Edition Open Access, Berlin.
Wallace, W.A., 1991. Galileo, the Jesuits, and the Medieval Aristotle. Variorum, Aldershot.
Zilsel, E., 2000. The Social Origins of Modern Science. Kluwer, Dordrecht.

Scientometrics

Loet Leydesdorff, University of Amsterdam, Amsterdam, The Netherlands
Staša Milojević, Indiana University, Bloomington, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the field of scientometrics, that is, the study of science, technology, and innovation from a quantitative perspective. We cover major historical milestones in the development of this specialism from the 1960s to today and discuss its relationship with the sociology of scientific knowledge, the library and information sciences, and science policy issues such as indicator development. The disciplinary organization of scientometrics is analyzed both conceptually and empirically. A state-of-the-art review of five major research threads is provided.

Introduction

Scientometrics can be defined as the "quantitative study of science, communication in science, and science policy" (Hess, 1997: p. 75). What started as Eugene Garfield's idea of an index to improve information retrieval in the 1960s and resulted in the creation of the *Science Citation Index* (SCI) (Garfield, 1979; Wouters, 1999) was soon recognized as a novel instrument in the empirical study of the sciences (e.g., de Solla Price, 1965; Cole and Cole, 1973). The availability of output indicators (such as databases of publications and patents) complemented ongoing efforts by the Organisation of Economic Cooperation and Development (OECD) in Paris to standardize input statistics of the scientific enterprise (OECD, 1963/1976). Based on these data, the National Science Board of the United States initiated the biannual series of *Science Indicators* in 1972. (The series was renamed into *Science and Engineering Indicators* in 1987 (National Science Board, 2012).)

The new journal *Scientometrics* was launched in 1978 and in that same year leading historians, philosophers of science, and social scientists – among them Robert K. Merton – published an edited volume entitled *Toward a Metric of Science: The Advent of Science Indicators*, in which they reflected on the new perspectives (Elkana et al., 1978). The historian Derek J. de Solla Price published a number of books and articles in the 1960s and 1970s which laid the foundations for the newly emerging field of quantitative science studies (e.g., de Solla Price, 1961, 1963, 1965), culminating in a full-fledged research program (de Solla Price, 1976).

The sociology of science, however, during the 1980s turned increasingly toward microanalysis focusing on the behavior of scientists in laboratories (e.g., Latour and Woolgar, 1979). From this perspective, the quantitative analysis of scientific literature at the macro (e.g., disciplinary) level was not considered a useful tool to explain scientific practices (Edge, 1979). Rather, with its focus on scientific communications – as a unit of analysis potentially different from scientists as authors – scientometrics developed at arm's length from the sociology of science and closer to the library and information sciences. At the same time, the value of scientometric indicators for informing science policies and research management became manifest (Irvine and Martin, 1984).

Under these diverging pressures, the field of science and technology studies increasingly bifurcated during the period 1985–2000 into qualitative 'sociology of scientific knowledge' on the one side, and the quantitative study of scientometrics and science indicators, on the other. Additionally, a third line of research emerged that published articles that use insights from the quantitative study of science and technology for evaluation and policy purposes. Such research appeared in journals such as *Research Policy*, *Research Evaluation*, and *Technology Analysis & Strategic Management* (Leydesdorff and Van den Besselaar, 1997).

During the 2000s, attention to evaluation and ranking was further enhanced after the publication of the first Academic Ranking of World Universities (ARWU) of the Shanghai Jiao Tong University in 2004 (Shin et al., 2011). The use of impact factors of journals for evaluative purposes has in the meantime pervaded the academic environment, even to the level of individual tenure decisions, which increasingly are being based on quantitative measures of publications and citations. Another popular indicator, the *h*-index (Hirsch, 2005), provides a simple impact metric for individual authors that can readily be used in online searching, for example, with Google Scholar, but is also incorporated in the major citation databases such as the Web-of-Science and Scopus. Computer programs (e.g., *Publish or Perish* at http://www.harzing.com/pop.htm) can freely be downloaded from the Internet and allow for measuring numbers of publications, citations, *h*-index, *g*-index (Egghe, 2006), etc., at the level of individuals, journals, and institutions without much prior knowledge of the scientometrics involved. In the meantime, the production and improvement of these indicators has become organized in university departments, spin-offs, and relevant companies such as Elsevier and Thomson-Reuters.

The increased access to large datasets through the Internet led to the development of the network sciences as part of computing and applied physics during the first decade of this century (e.g., Newman, 2010). A number of these studies used coauthorship and citation data for modeling the network dynamics. Although these efforts to model the evolution of the sciences statistically (e.g., Scharnhorst et al., 2012) often do not aim at contributing to social–scientific understanding and theorizing, the new methods (e.g., visualization techniques) developed by these researchers are partly derived

from and have also been adopted by scientometricians. Such interdisciplinary exchanges make scientometrics an active research specialty that in the 2000s has been experiencing a spectacular growth in its literature. More recently, the specialty has come to increasingly be considered as a 'research front' in terms of the turnover of the referencing patterns (Milojević and Leydesdorff, 2013; Wouters and Leydesdorff, 1994).

The Disciplinary Organization of 'Scientometrics'

Unlike the behavioral sciences and mainstream philosophy of science, scientometrics focuses on texts (documents) as empirical units of analysis. Figure 1 schematizes the relations with other disciplinary perspectives in science studies. Texts cannot be reduced to their authors – texts can, for example, be coauthored – nor can theories be reduced to the documents in which they are published. However, a measure in one dimension can be used as a proxy or indicator for the other given a research design.

As against texts or people, contents and theorizing remain latent and thus have to be theorized and hypothesized. Using factor analysis, however, one can reorganize data so that, for example, latent journal structures can be derived from the aggregation of citation linkages among journals. In Figure 2, we have applied techniques from network analysis to distinguish four communities in the citation networks among 32 journals that were cited by authors in *Scientometrics* during 2010. (All journals are included which contributed more than 1% to the total number of references in the journal *Scientometrics* during this year (2010).)

Figure 2 shows that *Scientometrics* as a journal has mainly been situated within the information sciences. The *Journal of the American Society for Information Science and Technology* (JASIST) has become a major publication outlet for scholars in this field in addition to the more specialized journals (such as *Journal of Informetrics* that was newly founded in 2007). However, the journal *Scientometrics* is the core journal in this field that interfaces with *Social Studies of Science*; a journal that mainly publishes qualitative science studies, and with general science journals (such as *Science* and *Nature*) and a number of journals that can be considered as part of technology and innovation studies (e.g., *R&D Management* and *Research Policy*). The relation with social network analysis and organization studies is more remote, but also visible.

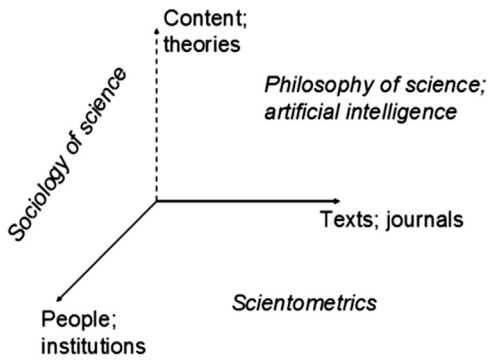

Figure 1 Three main dimensions in the dynamics of the sciences. Adapted from Leydesdorff, L., 1995. The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications. DSWO Press, Leiden University, Leiden.

At the lower-left side of Figure 2, one also can see strong links to a group of major multidisciplinary journals (*Nature*, *Science*, *Proceedings of the National Academy of Science of the USA (PNAS)*) and physics journals (*Physical Review E* and *Physica A*). These journals mostly publish the research of network science researchers – physicists and applied mathematicians – who discovered bibliographic databases as resources and study objects at around the turn of the century. These developments in physics and mathematical statistics have become part of the relevant environments for scientometric research, but the relevant journals are positioned very differently from journals such as *Social Networks* and *Administrative Science Quarterly* (at the top-right side of Figure 2). In other words, scientometric research nowadays is at the crossroads among the social sciences, information science, and advanced computing with its efforts to capture patterns in 'big data.'

Major Research Issues

The Measurement of Impact

The modern citation emerged in scientific literature – notably in chemistry – at the end of the nineteenth century as a standardized format (Bazerman, 1988; Leydesdorff and Wouters, 1999: p. 175). Citation indexing was organized during the 1960s (Garfield, 1979; Wouters, 1999; cf Cronin, 1984). With the advent of Google Scholar and Scopus as alternative citation indexes in 2004, citation analysis and rankings have become increasingly paradigmatic in research management and science policies. Discussions about the meaning of citations and the derived indicators tend to drive the scientometric enterprise.

For several decades, citation impact has been measured as the average number of citations per publication (c/p), and the impact factor (*IF*) was defined as a 2-year moving average at the level of journals. Citation distributions, however, are right-skewed because of cumulative advantages (de Solla Price, 1965, 1976) or, in other words, the 'Matthew Effect' in science (Merton, 1968) – the tendency for citation-rich authors and publications to draw further citations, in part *because* they are heavily cited. Therefore, the use of central tendency statistics (such as the arithmetic mean) is ill advised.

In addition to violating statistical assumptions, the use of the mean can affect the calculated impact when applied to different levels of aggregation by placing the number of publications in the denominator. For example, a Principal Investigator (PI) may lose average impact when his/her publications with junior staff are included in the set to be evaluated. However, a research group has more cumulative impact than its individual PI. Obviously, one needs a measure that can be aggregated, but after a normalization of the raw citation scores for differences among fields of science.

In 2010 and 2011, a scholarly debate (e.g., Gingras and Larivière, 2011) led gradually to the acceptance in measuring

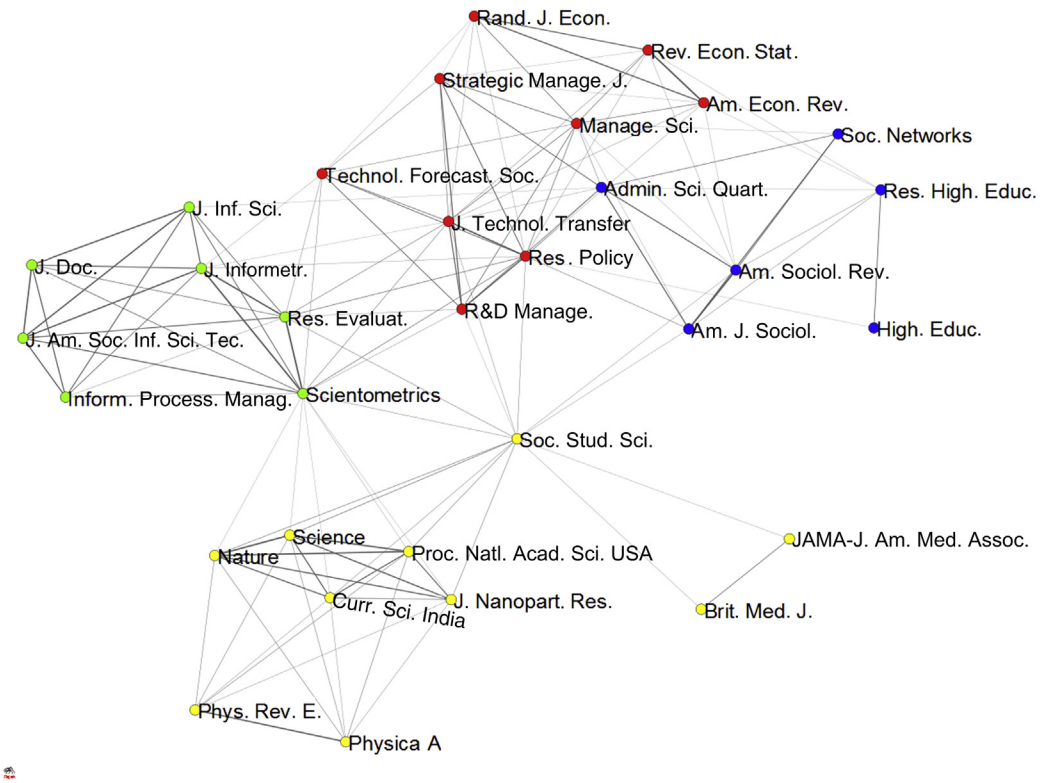

Figure 2 Map of 32 journals based on citing documents published in *Scientometrics* during 2010; Pajek used for the visualization; cosine >0.1; four communities detected with modularity $Q = 0.506$.

the impact of using percentiles and nonparametric statistics that accord with the skew in the distributions. The *Science and Engineering Indicators* (National Science Board, 2012), for example, uses six classes: the top-1%, top-5%, top-10%, top-25%, top-50%, and bottom-50%, but these measures can be further developed into appropriate statistics. Percentile ranks remain – like numbers of citations but now normalized – attributes at the article level and can be aggregated at different levels: institutions, nations, or even œ*uvres* of individual scholars. Statistical significance can be specified and error indicated, both in comparisons and with reference to expectations. The SCImago Institutions Rankings 2011 and Leiden Rankings 2011 of top universities, for example, both use the top-10% as an *excellence* indicator for which one can test for differences statistically (Bornmann et al., 2012).

Statistics based on normalization of the citation curves instead of averaging may lead to very different results. Using 2009 data, for example, Leydesdorff and Bornmann (2011) showed that PNAS could thus be attributed an impact that was significantly higher than *Science* or *Nature*, although, according to the usual measure, the latter journals had an impact factor that was three times higher. In brief, the new (nonparametric) indicators correlate more strongly with citations and publications than impact factors or other measures based on central tendency statistics.

Hirsch's (2005) *h*-index became another popular impact metric for capturing both the productivity and impact of an individual author. The index itself is crude, but there have been attempts to improve it. For example, Egghe proposed a *g*-index (2006) that increases the sensitivity of *h*-index values to highly cited papers. These indexes can be used for ranking scholars within a single discipline, because publication and citation behavior varies across disciplines. Recently, there have been proposals to create so-called 'universal' indexes, that is, indexes that can be used to evaluate and rank scientists across disciplines (e.g., Radicchi et al., 2008).

The Delineation of a Reference Set

A largely unresolved problem in scientometrics has been the delineation of reference sets for measuring the impact of journals or institutional units. Reference sets are a condition for appropriate normalization. The results of a scientometric evaluation can be highly sensitive to the attribution of papers to particular disciplines or specialties, and to the way specialties and (inter)disciplines are delineated. This makes it very difficult to delineate reference sets for universities that include multidisciplinary units (Rafols et al., 2012).

Following up on Moed (2010), Leydesdorff and Opthof (2010) proposed to use the citing papers as reference sets and to fractionate the citation credit in accordance with the length of the reference lists. This procedure takes into account that papers can address a variety of audiences with different citation practices. Using all publications in 2005 with Tsinghua University in Beijing as the institutional address, for example, Zhou and Leydesdorff (2011) showed that the department of Chinese Language and Literature would be

upgraded from the 19th to the 2nd position in the ranking among departments of this university if the citation scores were counted fractionally. Although fractionation moderates the effects of differences among disciplines, it does not enable articles that are not cited to be assessed.

In general, the problem remains that the citation matrices can be decomposed into (disciplinary) groupings to a variable extent. A single threshold therefore cannot remove the fuzziness that results from variations among fields of science. At the microlevel, a scientometrics researcher may carefully design appropriate sets of documents for representing relevant environments in considerable detail (**Figure 2** is a case in point). However, the specificity in each subset may make it difficult to aggregate sets from which more general conclusions can be drawn. The exception may be the rule in some sciences more than in others.

Theories of Citation

In line with different traditions in the sociology of science, it is possible to distinguish a normative from a constructivist theory of citations and other scientometric artifacts. Following Merton's (1942) set of norms – communalism, universalism, disinterestedness, and organized skepticism – one can argue that citations are a manifestation of influence and can therefore be considered as a currency for paying tribute in science (Zuckerman, 1987). This normative model has been countered by discourse-analytic arguments that emphasize the rhetorical construction of references in texts (Cozzens, 1989; Gilbert, 1977).

Fujigaki (1998) provided an interpretation of referencing in scientific papers as an important part of the self-organizing system in scientific fields: by providing citations, authors select those elements from the previously existing knowledge base which have to be retained for further development in a next round. Leydesdorff (1998) added that texts and authors constitute layered networks in which citations are expected to have different meanings. The system in the present can thus move forward referencing both the social and intellectual organization of science, while potentially following trajectories based on mutual adjustments between selections in these dimensions. Trajectories, however, can also be metastabilized and then globalized into regimes or paradigms. An evolutionary model of science is thus envisaged that is open for simulation and empirical research (Scharnhorst et al., 2012).

Mapping Science

Following Rip (1988: p. 254), a map of science can be defined as "the visualization of the topology of relationships between elements or aspects of science." Mapping of science can be useful for three purposes: retrieving information, understanding the dynamics of science, and informing science policy decisions about the allocation of resources (i.e., funding) and rewards. Scientometric analysis focuses on revealing the internal structure of intellectual domains, that is, mapping the components of disciplines, fields, or specialties on the basis of evidence from the literature under study. This can be achieved by mapping subject terms, documents, authors' *œuvres*, or journals. The basic data are cooccurrence counts.

Callon et al. (1986) argued that the sciences develop as heterogenous networks. In addition to heterogeneity across institutional domains (university, industry, government), organized knowledge can flexibly be codified in terms of mixtures of cited references and word usages. Words are more volatile than citations. Coauthorship relations, however, are not sufficiently informative of intellectual content for a semantic map; social network analysis has to be combined with semantic maps in order to generate rich representations. A longer-term challenge for this agenda is the animation of the informed networks over time, preferentially with specification of the statistical error involved. Advances in visualization and animation can further inform our understanding of the role of citations, title words, keywords, author names, etc., in evolving networks of scientific communication.

Policy and Management Contexts

Scientometricians also construct indicators that can be used in policy and management contexts. In these contexts, utility is sometimes more important than validity. For example, the use of impact factors to evaluate faculty members involves a combination of analyses at different levels of aggregation, and eventual applications to relatively small sets. Science maps and overlays to geographical maps can function heuristically at the aggregate level, but testing for significance requires sufficiently large sets.

The use of scientometric indicators for strategic choices at low levels of aggregation, such as 'picking the winners' (Irvine and Martin, 1984) in competitions for funding, is sensitive to the problem of relatively small sets, the operationalization of quality (e.g., the normalization of indicators across specialties and disciplines), and also the intrinsic problem of selecting the 'excellent' candidates for funding from an already preselected set of fundable proposals. Bornmann et al. (2010) showed in a number of cases that 'best rejected' proposers, when matched as pairs with grantees, scored higher on performance and impact indicators at a statistically significant level. The peer review process may introduce normative bias without intending to do so, in a process of selecting 'excellent' proposals from a pool of 'good' proposals. Scientometric metaevaluation may help with improving the selection process by enhancing awareness that it is not always possible to distinguish 'excellent' from 'good' research despite pressures for doing so for policy reasons.

Conclusions and Further Perspectives

The scientometric perspective adds a quantitative focus on texts and communication to the interdisciplinarity of science and technology studies. The philosophical distinction between 'context of discovery' and 'context of justification' (Popper, [1935]1959) was first overcome in empirical science studies that focus on interactions in which the social and cognitive organization of the sciences are continuously and actively reconstructed and recombined. However, this social and

intellectual process is textually mediated because texts can 'travel' more easily than scholars do, and thus the global dimension of paradigmatically structured horizons of meaning can be instantiated in relation to locally generated novelty. At the level of texts (e.g., manuscripts), one is able to recombine references to both local agency (author names, institutional addresses) and cognitive organizers such as title words and journal names. Furthermore, the content of manuscripts is validated (by peer review in contexts of justification) and thereupon the texts are admitted to the archive of published, and thus authenticated, scholarship on which future work can be built. These processes can be traced because they are documented.

The field of scientometrics has also expanded to different types of documents and other domains. For example, a similar process is working in the technological domain, but with a different dynamics (Dosi, 1982). Like manuscripts, patent applications contain knowledge claims that refer to 'prior art.' However, the functions and therefore the institutional incentives are different: patents are meant to protect intellectual property, whereas the public sciences are also based on the principle of gift giving (Merton, 1973). As publications, patents can be searched on the Internet, and the data can then be reorganized in terms of lines of intellectual heritage using citations, inventor names, institutional addresses, etc. Overlays on Google Maps, for example, allow for studying the geographic diffusion of new technologies based on sets and series of patented inventions.

Texts span networks of relations among authors, inventors, cognitions, institutional addresses, journals, etc. These networks contain both social (including economic) and cognitive relations. The cognitive dynamics are different from the social dynamics in that meaning is reflexively provided (that is, from the perspective of hindsight), while the historical networks develop with the time arrow. Thus, forward and backward loops are intertwined. Whereas the forward arrow necessarily generates Shannon-type information – uncertainty, variation – the backward loops can selectively reduce uncertainty by providing meanings to the events. Using the theory and computation of anticipatory systems (Dubois, 1998), the two operations can be distinguished as reconstructive incursion versus historical recursion.

In summary, the modeling of knowledge exchanges in scientific discourses cannot be reduced to the exchanges of information in coauthorship, coword, or citation relations. Models as entertained in the sciences enable researchers both to provide meaning to possible future states and to specify uncertainty. The measurement of the communication/sharing of meaning among frames of reference – for example, in university–industry–government relations – is very much on the research agenda of scientometrics (e.g., Leydesdorff and Ivanova, in press), but still a step away from how meaningful communication can further be codified in scientific discourses.

Models also serve to communicate possible future states. A construction of future states from a knowledge-based perspective can be modeled as hyperincursion. The modelers/scientists thus become carriers who are differently positioned in terms of their reflexive and communicative competencies; for example, as scientific explorers and/or appliers of engineering knowledge. The (re)constructions and their interactions update and reinforce the knowledge bases of the evolving societies and their economies. Authors in scientometrics are able to contribute to the study of science, technology, and innovation from a quantitative perspective by modeling and measuring these developments.

See also: Libraries; Peer Review and Quality Control; Science and Technology Studies: Experts and Expertise; Science and Technology, Social Study of: Computers and Information Technology; Science, History of; Social Network Algorithms and Software; Social Networks; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Bazerman, C., 1988. Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. University of Wisconsin Press, Madison, WI.

Bornmann, L., Leydesdorff, L., Van den Besselaar, P., 2010. A meta-evaluation of scientific research proposals: different ways of comparing rejected to awarded applications. Journal of Informetrics 4 (3), 211–220.

Bornmann, L., de Moya-Anegón, F., Leydesdorff, L., 2012. The new excellence indicator in the World Report of the SCImago Institutions Rankings 2011. Journal of Informetrics 6 (3), 333–335.

Callon, M., Law, J., Rip, A. (Eds.), 1986. Mapping the Dynamics of Science and Technology. Macmillan, London.

Cole, J.R., Cole, S., 1973. Social Stratification in Science. University of Chicago Press, Chicago/London.

Cozzens, S.E., 1989. What do citations count? The rhetoric-first model. Scientometrics 15 (5), 437–447.

Cronin, B., 1984. The Citation Process. The Role and Significance of Citations in Scientific Communication, vol. 1. Taylor Graham, London.

Dosi, G., 1982. Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. Research Policy 11 (3), 147–162.

Dubois, D.M., 1998. Computing anticipatory systems with incursion and hyper-incursion. In: Dubois, D.M. (Ed.), Computing Anticipatory Systems, CASYS-First International Conference, vol. 437. American Institute of Physics, Woodbury, NY, pp. 3–29.

Edge, D., 1979. Quantitative measures of communication in science: a critical overview. History of Science 17, 102–134.

Egghe, L., 2006. Theory and practise of the g-index. Scientometrics 69 (1), 131–152.

Elkana, Y., Lederberg, J., Merton, R.K., Thackray, A., Zuckerman, H., 1978. Toward a Metric of Science: The Advent of Science Indicators. Wiley, New York, etc.

Fujigaki, Y., 1998. Filling the gap between discussions on science and scientists' everyday activities: applying the autopoiesis system theory to scientific knowledge. Social Science Information 37 (1), 5–22.

Garfield, E., 1979. Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. John Wiley, New York.

Gilbert, G.N., 1977. Referencing as persuasion. Social Studies of Science 7, 113–122.

Gingras, Y., Larivière, V., 2011. There are neither "king" nor "crown" in scientometrics: comments on a supposed "alternative" method of normalization. Journal of Informetrics 5 (1), 226–227.

Hess, D.J., 1997. Science Studies: An Advanced Introduction. New York University Press, New York.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America 102 (46), 16569–16572.

Irvine, J., Martin, B.R., 1984. Foresight in Science: Picking the Winners. Frances Pinter, London.

Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, Beverly Hills, etc.

Leydesdorff, L., 1995. The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications. DSWO Press, Leiden University, Leiden.

Leydesdorff, L., 1998. Theories of citation? Scientometrics 43 (1), 5–25.
Leydesdorff, L., Bornmann, L., 2011. Integrated Impact Indicators (I3) compared with Impact Factors (IFs): an alternative design with policy implications. Journal of the American Society for Information Science and Technology 62 (11), 2133–2146. http://dx.doi.org/10.1002/asi.21609.
Leydesdorff, L., Ivanova, I.A., Mutual redundancies in inter-human communication systems: steps towards a calculus of processing meaning. Journal of the American Society for Information Science and Technology, in press.
Leydesdorff, L., Opthof, T., 2010. Scopus' source normalized impact per paper (SNIP) versus the journal impact factor based on fractional counting of citations. Journal of the American Society for Information Science and Technology 61 (11), 2365–2396.
Leydesdorff, L., Van den Besselaar, P., 1997. Scientometrics and communication theory: towards theoretically informed indicators. Scientometrics 38, 155–174.
Leydesdorff, L., Wouters, P., 1999. Between texts and contexts: advances in theories of citation? Scientometrics 44 (2), 169–182.
Merton, R.K., 1942. Science and technology in a democratic order. Journal of Legal and Political Sociology 1, 115–126.
Merton, R.K., 1968. The Matthew effect in science. Science 159, 56–63.
Merton, R.K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago/London.
Milojević, S., Leydesdorff, L., 2013. Information metrics (iMetrics): a research specialty with a socio-cognitive identity? Scientometrics 95 (1), 141–157.
Moed, H.F., 2010. Measuring contextual citation impact of scientific journals. Journal of Informetrics 4 (3), 265–277.
National Science Board, 2012. Science and Engineering Indicators. National Science Foundation, Washington, DC. http://www.nsf.gov/statistics/seind12/.
Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press, Oxford/New York.
OECD, 1963/1976. The Measurement of Scientific and Technical Activities: "Frascati Manual". OECD, Paris.
Popper, K.R., [1935]1959. The Logic of Scientific Discovery. Hutchinson, London.
Radicchi, F., Fortunato, S., Castellano, C., 2008. Universality of citation distributions: toward an objective measure of scientific impact. Proceedings of the National Academy of Sciences 105 (45), 17268–17272.
Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., Stirling, A., 2012. How journal rankings can suppress interdisciplinary research: a comparison between innovation studies and business & management. Research Policy 41 (7), 1262–1282.
Rip, A., 1988. Mapping of science: possibilities and limitations. In: Van Raan, A.F.J. (Ed.), Handbook of Quantitative Studies of Science and Technology. Elsevier, North-Holland, pp. 253–273.
Scharnhorst, A., Börner, K., van den Besselaar, P. (Eds.), 2012. Models of Science Dynamics: Encounters between Complexity Theory and Information Sciences. Springer, Heidelberg, etc.
Shin, J.C., Toutkoushian, R.K., Teichler, U., 2011. University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education. Springer, Dordrecht.
de Solla Price, D.J., 1961. Science since Babylon. Yale University Press, New Haven.
de Solla Price, D.J., 1963. Little Science, Big Science. Columbia University Press, New York.
de Solla Price, D.J., 1965. Networks of scientific papers. Science 149 (3683), 510–515.
de Solla Price, D.J., 1976. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27 (5), 292–306.
Wouters, P., 1999. The Citation Culture (unpublished Ph.D. thesis). University of Amsterdam, Amsterdam.
Wouters, P., Leydesdorff, L., 1994. Has Price's dream come true: is scientometrics a hard science? Scientometrics 31, 193–222.
Zhou, P., Leydesdorff, L., 2011. Fractional counting of citations in research evaluation: a cross- and interdisciplinary assessment of the Tsinghua University in Beijing. Journal of Informetrics 5 (3), 360–368.
Zuckerman, H., 1987. Citation analysis and the complex problem of intellectual influence. Scientometrics 12 (5), 329–338.

Screening and Selection

Joseph L Gastwirth, George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 20, pp. 13755–13760, © 2001, Elsevier Ltd, with an updated Bibliography section supplied by the Editor.

Abstract

Procedures for classifying individuals as possessing a characteristic or trait are used to make decisions concerning university admissions and employee selection, auditing tax returns, health status or the need for more extensive diagnostic procedures, and the veracity of accused criminals or job applicants for security sensitive jobs. The statistical concepts underlying these screening methods and their proper interpretation in a variety of fields are described. The role of the prevalence of the trait in the population screened is stressed. In many applications the true status of an individual cannot be determined perfectly so special studies are needed to estimate the accuracy of screening tests. In addition to selecting individuals for a position or for special treatment, results of screening tests can be used to estimate the prevalence of a trait in a population. For some purposes, e.g., screening blood donations for HIV/AIDS, a test that is not accurate enough to classify an individual can still be very useful in protecting the public.

Introduction

In order to assist in selecting individuals possessing either desirable traits such as an aptitude for higher education or skills needed for a job or undesirable ones such as having an infection, illness, or a propensity to lie, screening tests are used as a first step. A more rigorous selection device, e.g., diagnostic test or detailed interview is then used for the final classification. For some purposes, such as screening blood donations for a rare infection, the units classified as positive are not donated but further tests on donors may not be given as most will not be infected. Similarly, for estimating the prevalence of a trait in a population, the screening data may suffice provided an appropriate estimator that incorporates the error rates of the test is used (Hilden, 1979; Gastwirth, 1987; Rahme and Joseph, 1998). This article describes the measures of accuracy used to evaluate and compare screening tests and issues arising in the interpretation of the results. The importance of the prevalence of the trait on the population screened and the relative costs of the two types of misclassification are discussed. Methods for estimating the accuracy rates of screening tests are briefly described and the need to incorporate them in estimates of prevalence is illustrated.

Basic Concepts

The purpose of a screening test is to determine whether a person or object is a member of a particular class, C or its complement, $C-$. The test result indicating that the person is in C will be denoted by S and a result indicating nonmembership by $S-$. The *accuracy* of a test is described by two probabilities:

$$\eta = P[S|C]$$

being the probability that someone in C is correctly classified, or the sensitivity of the test; and

$$\theta = P[S-|C-]$$

being the probability that someone not in C is correctly classified, or the specificity of the test. Given the prevalence, $\pi = \mathrm{P}(C)$, of the trait in the population screened, from Bayes' theorem it follows that the predictive value of a positive test (PVP) is

$$P[C|S] = \pi\eta/[\pi\eta + 1 - \pi)(1 - \theta)].{}_X$$

Similarly, the predictive value of a negative test is

$$P[C-|S-] = (1-\pi)\theta/[(1-\pi)\theta + \pi(1-\eta)].$$

In the first two sections, we will assume that the accuracy rates and the prevalence are known. When they are estimated from data, appropriate sampling errors for them and the PVP are given in Gastwirth (1987).

For illustration, consider an early test for HIV, having a sensitivity of 0.98 and a specificity of 0.93, applied in two populations. The first has a very low prevalence, 1.0×10^{-3}, of the infection while the second has a prevalence of 0.25. From eqn [1], the PVP in the first population equals 0.0138, i.e., only about one-and-one-half percent of individuals classified as infected would actually be so. Nearly 99% would be false positives. Notice that if the fraction of positives, expected to be 0.0797, in the screened data were used to estimate prevalence, a severe overestimate would result. Adjusting for the error rates yields an accurate estimate. In the higher prevalence group, the PVP is 0.8235, indicating that the test could be useful in identifying individuals. Currently used tests have accuracy rates greater than 0.99, but even these still have a PVP less than 0.5 when applied to a low prevalence population. A comprehensive discussion is given in Brookmeyer and Gail (1994).

The role of the prevalence, also called the base rate, of the trait in the screened population and how well people understand its effect has been the subject of substantial research, recently reviewed by Koehler (1996). An interesting consequence is that when steps are taken to reduce the prevalence of the trait prior to screening, the PVP declines and the fraction of false-positives increases. Thus, when high-risk donors are encouraged to defer or when background checks eliminate a sizable fraction of unsuitable applicants prior to their being subjected to polygraph testing the fraction of classified positives who are truly positive is small.

In many applications screening tests yield data that are ordinal or essentially continuous, e.g., scores on psychological tests or the concentration of HIV antibodies. Any value, t, can be used as a cut-off point to delineate between individuals with the trait and 'normal.' Each t generates a corresponding sensitivity and specificity for the test and the user must then incorporate the relative costs of the two different errors and the likely prevalence of the trait in the population being screened to select the cut off. The receiver operating characteristic (ROC) curve displays the trade-off between the sensitivity and specificity defined by various choices of t and also yields a method for comparing two or more screening tests.

To define the ROC curve, assume that the distribution of the measured variable (test score or physical quantity) is $F(x)$ for the 'normal' members of the population but is $G(x)$ for those with the trait. The corresponding density functions are $f(x)$ and $g(x)$ respectively and $g(x)$ will be shifted (to the right (left) if large (small) scores indicate the trait) of $f(x)$ for a good screening test. Often one fixes the probability ($\alpha = 1 - \gamma\theta$ or one minus the specificity) of classifying a person without the characteristic as having it at a small value. Then t is determined from the equation $F(t) = 1 - \theta$. The sensitivity, η, of the test is $1 - G(t)$. The ROC curve plots η against $1 - \theta$. A perfect test would have $\eta = 1$ so the closer the ROC is to the upper left corner in **Figure 1**, the better the screening test. In **Figure 1** we assume f is a normal density with mean 0 and variance 1 while g_1 is normal with mean 1 and the same variance. For comparison we also graphed the ROC for a second test, which has a density g_2 with mean 2 and variance 1. Notice that the ROC curve for the second test is closer to the left corner (0,1) than that of the first test.

A summary measure (Campbell, 1994), which is useful in comparing two screening tests is the area, A, under the ROC.

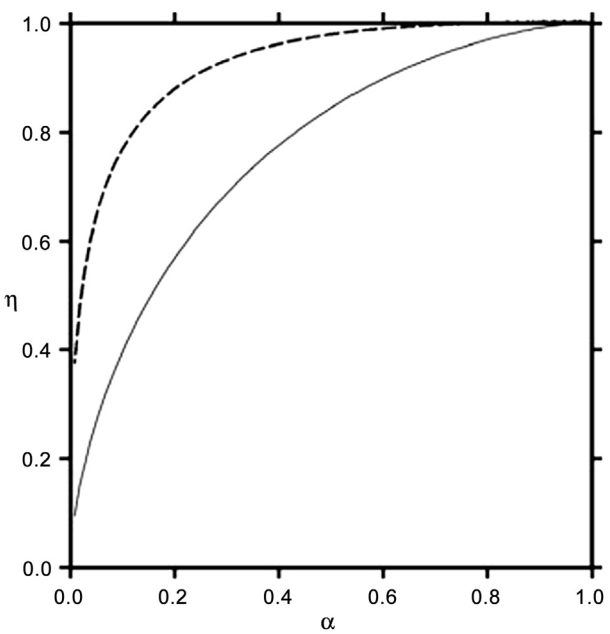

Figure 1 The ROC curves for screening tests 1 and 2. The solid line is the curve when the diseased group has mean 1 and the dashed curve is for the second group with mean 2.

The closer A is to its maximum value of 1.0, the better the test is. In **Figure 1**, the areas under the ROC curves for the two tests are 0.761 and 0.922, respectively. Thus, the areas reflect the fact that the ROC curve for the second test is closer to what a 'perfect' test would be. This area equals the probability a randomly chosen individual will have a higher score on the screening test than a normal one. This probability is the expected value of the Mann–Whitney form of the Wilcoxon test for comparing two distributions and methods for estimating it are in standard nonparametric statistics texts.

Nonparametric methods for estimating the entire ROC curve are given by Wieand et al. (1989) and Hilgers (1991) obtained distribution-free confidence bounds for it. Campbell (1994) uses the confidence bounds on F and G to construct a joint confidence interval for the sensitivity and one minus the specificity in addition to proposing alternative confidence bounds for the ROC itself. Hseih and Turnbull (1996) determine the value of t that maximizes the sum of sensitivity and specificity. Their approach can be extended to maximizing weighted average of the two accuracy rates, suggested by Gail and Green (1976). Wieand et al. (1989) also developed related statistics focusing on the portion of the ROC lying above a region, α_1 and α_2 so the analysis can be confined to values of specificity that are practically useful. Greenhouse and Mantel (1950) determine the sample sizes needed to test whether both the specificity and sensitivity of a test exceed prespecified values.

The area A under the ROC can also be estimated using parametric distributions for the densities f and g. References to this literature and an alternative approach using smoothed histograms to estimate the densities is developed in Zou et al. (1997). They also consider estimating the partial area over the important region determined by two appropriate small values of α.

The tests used to select employees need to be reliable and valid. Reliability means that replicate values are consistent while validity means that the test measures what it should, e.g., successful academic performance. Validity is often assessed by the correlation between the test score (X) and subsequent performance (Y). Often X and Y can be regarded as jointly normal random variables, especially as monotone transformations of the raw scores can be used in place of them. If a passing score on the screening or preemployment test is defined as $X \ngeq t$ and successful performance is defined as $Y \ngeq d$, then the sensitivity of the test is $P[X > t|Y > d]$, the specificity is $P[XY < d]$, and the prevalence of the trait is $P[Y > d]$ in the population of potential applicants. Hence, the aptitude and related tests can be viewed from the general screening test paradigm.

When the test and performance scores are scaled to have a standard bivariate normal distribution, both the sensitivity and specificity increase with the correlation, ρ. For example, suppose one desired to obtain employees in the upper half of the performance distribution and used a cut-off score, t, of one-standard deviation above the mean on the test (X). When $\rho = 0.3$, the sensitivity is 0.217 while the specificity is 0.899. If $\rho = 0.5$, the sensitivity is 0.255 and the specificity is 0.937. The use of a high cut-off score eliminates the less able applicants but also disqualifies a majority of applicants who are in the upper half of the performance distribution. Reducing the cut-off score to one-half a standard deviation above the average raises the

sensitivities to 0.394 and 0.454 for the two values of ρ but lowers the corresponding specificities to 0.777 and 0.837. This tradeoff is a general phenomenon as seen in the ROC curves.

The Importance of the Context in Interpreting the Results of Screening Tests

In medical and psychological applications, an individual who tests positive for a disease or condition on a screening test will be given a more accurate confirmatory test or intensive interview. The cost of a 'false positive' screening result on a medical examination is often considered very small relative to a 'false negative,' which could lead to the failure of suitable treatment to be given in a timely fashion. A false positive result presumably would be identified in a subsequent more detailed examination. Similarly, when government agencies give employees in safety or security sensitive jobs a polygraph test, the loss of potentially productive employee due to a false positive was deemed much less than the risk of hiring an employee who would might be a risk to the public or a security risk.

One can formalize the issue by including the costs of various errors and the prevalence, π, of the trait in the population being screened in determining the cut-off value of the screening test. Then the expected cost, which weights the probability of each type of error by its cost is given by

$$w\alpha(1-\pi)+(1-w)\pi G(t).$$

Here the relative costs of a false positive (negative) are w and $1-w$, respectively and as before $t=F^{-1}(1-\alpha)$. The choice of cut-off value, t_0, minimizing the expected cost satisfies:

$$\frac{g(t)}{f(t)}=\frac{w(1-\pi)}{(1-w)\pi} \qquad [1]$$

Whenever the ratio, g:f, of the density functions is a monotone function the previous equation yields an optimum cut-off point, t_0, which depends on the costs and prevalence of the trait. Note that for any value of π, the greater the cost of a false positive, the larger will be the optimum value, t_0. This reflects the fact that the specificity needs to be high in order to keep the false positive rate low.

Although the relative costs of the two types of error are not always easy to obtain and the prevalence may only be approximately known, eqn [1] may aid in choosing the critical value. In practice, one should also assess the effect slight changes in the costs and assumed prevalence have on the choice of the cut-off value.

The choice of t that satisfies condition [1] may not be optimal if one desires to estimate the prevalence of the trait in a population rather than classifying individuals. Yanagawa and Tokudome (1990) determine t when the objective is to minimize the relative absolute error of the estimator of prevalence on the basis of the screening test results.

The HIV/AIDS epidemic raised questions about the standard assumptions about the relative costs of the two types of error. A 'false positive' classification would not only mean that a well individual would worry until the results of the confirmatory test were completed, they also might have social and economic consequences if friends or their employer learned of the result.

Similar problems arise in screening blood donors and in studies concerning the association of genetic markers and serious diseases. Recall that the vast majority of donors or volunteers for genetic studies are doing a public service and are being screened to protect others or advance knowledge. If a donation tests positive, clearly it should not be used for transfusion. Should a screened-positive donor be informed of their status? Because the prevalence of infected donors is very small, the PVP is quite low so that most of the donors screened positive are 'false.' Thus, blood banks typically do not inform them and rely on approaches to encourage donors from high-risk groups to exclude themselves from the donor pool (Nusbacher et al., 1986). Similarly, in a study (Hartge et al., 1999) of the prevalence of mutations in two genes that have been linked to cancer the study participants were not notified of their results.

The screening test paradigm is useful in evaluating tests used to select employees. The utility of a test depends on the costs associated with administering the test and the costs associated with the two types of error. Traditionally, employers focused on the costs of a false positive, hiring an employee who does not perform well, such as termination costs, and the possible loss of customers. The costs of a false negative are more difficult to estimate.

The civil-rights law, which was designed to open job opportunities to minorities, emphasized the importance of using appropriate tests, i.e., tests that selected better workers. Employers need to check whether the tests or job requirements (e.g., possession of a high school diploma) have a disparate impact upon a legally protected group. When they exclude a significantly greater fraction of minority members than majority ones, the employer needs to validate it, i.e., show it is predictive of the job performance. Arvey (1979) and Paetzold and Willborn (1994) discuss these issues.

Estimating the Accuracy of the Screening Tests

So far, we have assumed that we can estimate the accuracy of the screening tests on samples from two populations where the true status of the individuals is known with certainty. In practice, this is often not the case and can lead to biased estimates of the sensitivity and specificity of a screening test, as some of the individuals believed to be normal have the trait, and vice versa.

If one has samples from only one population to which to apply both the screening and confirmatory test, then one cannot estimate the accuracy rates. The data would be organized into a 2×2 table, with four cells, only three of which are independent. There are five parameters, however, the two accuracy rates of the two tests plus the prevalence of the trait in the population. In some situations, the prevalence of the trait may vary among subpopulations. If one can find two such subpopulations and if the accuracy rates of both tests are the same in both of those sub-populations, then one has two 2×2 tables with six independent cells, with which to estimate six parameters. Then estimation can then be carried out (Hui and Walter, 1980).

This approach assumes that the two tests are conditionally independent given the true status of the individual. When this assumption is not satisfied, Vacek (1985) showed that the

estimates of sensitivity and specificity of the tests are biased. This topic is an active area of research, recently reviewed by Hui and Zhou (1998). A variety of latent-class models have been developed that relax the assumption of conditional independence (see Faraone and Tsuang, 1994; Yang and Becker, 1997 and the literature they cite).

Applications and Future Concerns

Historically screening tests were used to identify individuals with a disease or trait, e.g., as a first stage in diagnosing medical or psychological conditions or select students or employees. They are being increasingly used, often in conjunction with a second, confirmatory test, in prevalence surveys for public health planning. The techniques developed are often applicable, with suitable modifications, to social science surveys.

Some examples of prevalence surveys illustrate their utility. Katz et al. (1995) compared two instruments for determining the presence of psychiatric disorders in part to assess the needs for psychiatric care in the community and the available services. They found that increasing the original cut-off score yielded higher specificity without a substantial loss of sensitivity. Similar studies were carried out in Holland by Hodiamont et al. (1987) who found a lower prevalence (7.5%) than the 16% estimate in York. The two studies, however, used different classification systems illustrating that one needs to carefully examine the methodology underlying various surveys before making international comparisons. Gupta et al. (1997) used several criteria based on the results of an EKG and an individual's medical history to estimate the prevalence of heart disease in India.

Often one has prior knowledge of the prevalence of a trait in a population, especially if one is screening similar populations on a regular basis, as would be employers, medical plans, or blood centers. Bayesian methods incorporate this background information and can yield more accurate estimates (see Geisser, 1993; Johnson et al., 2001). A cost-effective approach is to use an inexpensive screen at a first stage and retest the positives with a more definitive test. Bayesian methodology for such studies was developed by Erkanli et al. (1997).

The problem of misclassification arises often in questionnaire surveys. Laurikka et al. (1995) estimated the sensitivity and specificity of self-reporting of varicose veins. While both measures were greater than 0.90, the specificity was lower (0.83) for individuals with a family history than those with negative histories. Sorenson (1998) observed that often self-reports are accepted as true and found that the potential misclassification could lead to noticeable (10%) errors in estimated mortality rates. The distortion that misclassification errors can have on estimates of low prevalence traits because of the high fraction of false positive classifications was illustrated in the earlier discussion of screening tests for HIV/AIDS. Hemenway (1997) applies these concepts to demonstrate that surveys typically overestimate rare events, in particular the self-defense uses of guns. Thus it is essential to incorporate the accuracy rates into the prevalence estimate (Hilden, 1979; Gastwirth, 1987; Rahme and Joseph, 1998).

Sinclair and Gastwirth (1994) utilized the Hui-Walter paradigm to assess the accuracy of both the original and reinterview (by supervisors) classifications in labor force surveys. In its evaluation, the Census Bureau assumes that the reinterview data are correct; however, those authors found that both interviews had similar accuracy rates. In situations where one can obtain three or more classifications, all the parameters are identifiable (Walter and Irwig, 1988). Gastwirth and Sinclair (1998) utilized this feature of the screening test approach to suggest an alternative design for judge–jury agreement studies that had another expert, e.g., law professor or retired judge, assess the evidence.

Conclusion

Many selection or classification problems can be viewed from the screening test paradigm. The context of the application determines the relative costs of a misclassification or erroneous identification. In criminal trials, society has decided that the cost of an erroneous conviction far outweighs the cost of an erroneous acquittal. While, in testing job applicants, the cost of not hiring a competent worker is not as serious. The two types of error vary with the threshold or cut-off value and the accuracy rates corresponding to these choices is summarized by the ROC curve.

There is a burgeoning literature in this area as researchers are incorporating relevant covariates, e.g., prior health status or educational background into the classification procedures. Recent issues of *Biometrics and Multivariate Behavioral Research*, *Psychometrika* and *Applied Psychological Measurement* as well as the medical journals cited in the article contain a variety of articles presenting new techniques and applications of them to the problems discussed.

See also: Selection Bias, Statistics of.

Bibliography

Arvey, R.D., 1979. Fairness in Selecting Employees. Addison-Wesley, Reading, MA.
Brookmeyer, R., Gail, M.H., 1994. AIDS Epidemiology: A Quantitative Approach. Oxford University Press, New York.
Campbell, G., 1994. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine 13, 499–508.
Erkanli, A., Soyer, R., Stangl, D., 1997. Bayesian inference in two-phase prevalence studies. Statistics in Medicine 16, 1121–1133.
Faraone, S.V., Tsuang, M.T., 1994. Measuring diagnostic accuracy in the absence of a gold standard. American Journal of Psychiatry 151, 650–657.
Gail, M.H., Green, S.B., 1976. A generalization of the one-sided two-sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests. Biometrics 32, 561–570.
Gastwirth, J.L., 1987. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data (with discussion). Statistical Science 2, 213–238.
Gastwirth, J.L., Sinclair, M.D., 1998. Diagnostic test methodology in the design and analysis of judge–jury agreement studies. Jurimetrics Journal 39, 59–78.
Geisser, S., 1993. Predictive Inference. Chapman and Hall, London.
Greenhouse, S.W., Mantel, N., 1950. The evaluation of diagnostic tests. Biometrics 16, 399–412.
Gupta, R., Prakash, H., Gupta, V.P., Gupta, K.D., 1997. Prevalence and determinants of coronary heart disease in a rural population of India. Journal of Clinical Epidemiology 50, 203–209.
Hartge, P., Struewing, J.P., Wacholder, S., Brody, L.C., Tucker, M.A., 1999. The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. American Journal of Human Genetics 64, 963–970.
Hemenway, D., 1997. The myth of millions of annual self-defense gun uses: a case study of survey overestimates of rare events. Chance 10, 6–10.

Hilden, J., 1979. A further comment on estimating prevalence from the results of a screening test. American Journal of Epidemiology 109, 721–722.

Hilgers, R.A., 1991. Distribution-free confidence bounds for ROC curves. Methods and Information Medicine 30, 96–101.

Hodiamont, P., Peer, N., Syben, N., 1987. Epidemiological aspects of psychiatric disorder in a Dutch health area. Psychological Medicine 17, 495–505.

Hseih, F.S., Turnbull, B.W., 1996. Nonparametric methods for evaluating diagnostic tests. Statistics Sinica 6, 47–62.

Hui, S.L., Walter, S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

Hui, S.L., Zhou, X.H., 1998. Evaluation of diagnostic tests without gold standards. Statistical Methods in Medical Research 7, 354–370.

Johnson, W.O., Gastwirth, J.L., Pearson, L.M., 2001. Screening without a gold standard: the Hui–Walter paradigm revisited. American Journal of Epidemiology 153, 921–924.

Katz, R., Stephen, J., Shaw, B.F., Matthew, A., Newman, F., Rosenbluth, M., 1995. The East York health needs study—prevalence of DSM-III-R psychiatric disorder in a sample of Canadian women. British Journal of Psychiatry 166, 100–106.

Koehler, J.J., 1996. The base rate fallacy reconsidered: descriptive, normative and methodological challenges (with discussion). Behavioral and Brain Sciences 19, 1–53.

Laurikka, J., Laara, E., Sisto, T., Tarkka, M., Auvinen, O., Hakama, M., 1995. Misclassification in a questionnaire survey of varicose veins. Journal of Clinical Epidemiology 48, 1175–1178.

Nusbacher, J., Chiavetta, J., Naiman, R., Buchner, B., Scalia, V., Horst, R., 1986. Evaluation of a confidential method of excluding blood donors exposed to human immunodeficiency virus. Transfusion 26, 539–541.

Paetzold, R., Willborn, S., 1994. Statistical Proof of Discrimination. Shepard's/McGraw Hill, Colorado Springs, CO.

Percival Zhang, Y.-H., Himmel, Michael E., Mielenz, Jonathan R., 2006. Outlook for cellulase improvement: screening and selection strategies. Biotechnology Advances 24 (5), 452–481.

Rahme, E., Joseph, L., 1998. Estimating the prevalence of a rare disease: adjusted maximum likelihood. The Statistician 47, 149–158.

Sinclair, M.D., Gastwirth, J.L., 1994. On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. Journal of American Statistical Assn 91, 961–969.

Sorenson, S.B., 1998. Identifying Hispanics in existing databases. Evaluation Review 22, 520–534.

Tammemägi, Martin C., Katki, Hormuzd A., Hocking, William G., Church, Timothy R., Caporaso, Neil, Kvale, Paul A., Chaturvedi, Anil K., Silvestri, Gerard A., Riley, Tom L., Commins, John, 2013. Selection criteria for lung-cancer screening. New England Journal of Medicine 368 (8), 728–736.

Vacek, P.M., 1985. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41, 959–968.

Vigouroux, Y., McMullen, M., Hittinger, C.T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y., Doebley, J., 2002. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proceedings of the National Academy of Sciences 99 (15), 9650–9655.

Walter, S.D., Irwig, L.M., 1988. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. Journal of Clinical Epidemiology 41, 923–937.

Wieand, S., Gail, M.H., James, B.R., James, K., 1989. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 76, 585–592.

Yanagawa, T., Tokudome, S., 1990. Use of screening tests to assess cancer risk and to estimate the risk of adult T-cell leukemia/lymphoma. Environmental Health Perspectives 87, 77–82.

Yang, I., Becker, M.P., 1997. Latent variable modeling of diagnostic accuracy. Biometrics 52, 948–958.

Zou, K.H., Hall, W.J., Shapiro, D.E., 1997. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine 16, 214–256.

Search, Economics of

Christopher A Pissarides, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The economics of search study the implications of frictions for individual behavior and market performance, due usually to imperfect information about exchange possibilities. This article reviews labor-market research in this area. Individuals search for a job offer by choosing a reservation wage and accepting jobs that pay above that wage. Firms create jobs to maximize profit. The probability of realizing a match is derived from an aggregate matching function that gives the number of jobs formed in terms of the search efforts of firms and workers. Because of monopoly rents implied by frictions, wages are determined either by bargaining or by take-it-or-leave- it offers posted by firms. Bargaining models share the surplus and give rise to a single wage for homogeneous labor that in general does not allocate jobs efficiently. Wage posting can give rise to a single wage equal to the worker's reservation wage if the worker can only observe one wage offer at a time, or to the efficient allocation if posted wages can be observed but there are job queues, or to a distribution of wage offers if employed workers can sample more than one firm at a time. Job destruction is due either to job-specific shocks or to technological obsolescence.

The economics of search study the implications of market frictions for economic behavior and market performance. 'Frictions' in this context include anything that interferes with the smooth and instantaneous exchange of goods and services. The most commonly studied problems arise from imperfect information about the location of buyers and sellers, their prices, and the quality of the goods and services that they trade. The key implication of these frictions is that individuals are prepared to spend time and other resources on exchange; they *search* before buying or selling. The labor market has attracted most theoretical and empirical interest in this area of research, because of the heterogeneities that characterize it and the existence of good data on flows of workers and jobs between activity and inactivity, which can be used to test its propositions.

Historical Background

The first formal model of individual behavior, due to Stigler (1961), was in the context of a goods market: choosing the optimal number of sellers to search before buying at the lowest price. Stigler's rule, known as a fixed-sample rule, was abandoned in favor of sequential stopping rules: choosing the optimal reservation price and buying at the first store encountered which sells at or below the reservation price (see McCall, 1970, for an early influential paper).

The first big momentum to research in the economics of search came with the publication of Phelps et al. (1970), which showed that search theory could be used to analyze the natural rate of unemployment and the inflation-unemployment trade off, the central research questions of macroeconomics at that time. Although interest in the inflation-unemployment trade-off has since waned, interest in search theory as a tool to analyze transitions in the labor market and equilibrium unemployment has increased. The momentum in this direction came in the 1980s, when contributions by Diamond (1982), Mortensen (1982), and Pissarides (1985) showed that search theory could be used to construct equilibrium models of the labor market with more accurate predictions than the traditional neoclassical model (see Pissarides, 2000; Mortensen and Pissarides, 1999a,b, for reviews). The appearance of comprehensive data on job and worker flows, which can be studied with the tools of search theory, also contributed to interest in this direction (see Leonard, 1987; Dunne et al., 1989; Davis et al., 1996; Blanchard and Diamond, 1989). This article reviews major developments since the mid-1980s, with explicit reference to labor markets.

Job Search

An individual has one unit of labor to sell to firms, which create jobs. The valuation of labor takes place under the assumptions that agents have infinite horizons and discount future income flows at the constant rate r, they know the future path of prices and wages and the stochastic processes that govern the arrival of trading partners, and they maximize the present discounted value of expected incomes. Let U_t be the expected present discounted value of a unit of labor before trade at time t (the 'value' of an unemployed worker) and W_t the expected value of an employed worker. During a short time interval δt the unemployed worker receives income $b\delta t$, and a job offer arrives with probability $a\delta t$. The frictions studied in the economics of search are summarized in the arrival process. In the absence of frictions, $a \to \infty$. With frictions and search, $a > 0$; with no search, $a = 0$. A large part of the literature is devoted to specifying the arrival process, an issue addressed in Section The Aggregate Matching Function. The choice of search intensity has also been studied, by making a an increasing function of search effort, but this issue is not addressed here (see Pissarides, 2000: Chapter 5).

If a job offer arrives, the individual has the option of taking it, for an expected return $W_{t+\delta t}$; or not taking it and keeping instead return $U_{t+\delta t}$. If no offer arrives, the individual's return is $U_{t+\delta t}$. Therefore, with discount rate r, U_t satisfies the Bellman equation

$$U_t = b\delta t + a\delta t \frac{\max(W_{t+\delta t}, U_{t+\delta t})}{1 + r\delta t} + (1 - a\delta t)\frac{U_{t+\delta t}}{1 + r\delta t} \qquad [1]$$

Rearrangement of terms yields

$$rU_t = b + a(\max(W_{t+\delta t}, U_{t+\delta t}) - U_{t+\delta t}) + \frac{U_{t+\delta t} - U_t}{\delta t} \quad [2]$$

Taking the limit of eqn [2] as $\delta t \to \infty$, and omitting subscripts for convenience, yields

$$rU = b + a\,(\max\,(W, U) - U) + \dot{U} \quad [3]$$

where $\dot{U}$ denotes the rate of change of U.

Equation [3] is a fundamental equation in the economics of search. It can be given the interpretation of an arbitrage equation for the valuation of an asset in a perfect capital market with risk-free interest rate r. This asset yields coupon payment b and at some rate a, it gives its holder the option of a discrete change in its valuation, from U to W. Optimality requires that the option is taken (and the existing valuation given up) if $W \geq U$. The last term, $\dot{U}$, shows capital gains or losses due to changes in the market valuation of the asset. In most of the economics of labor-market search, however, research concentrates on 'steady states,' namely, on situations where the discount rate, transition rates, and income flows are all constant. With infinite horizons there are then stationary solutions to the valuation equations, obtained from eqn [3] with $\dot{U} = 0$.

One simple way of solving eqn [3] is to assume that employment is an 'absorbing state,' so when a job that offers wage w is accepted, it is kept for life. Then, $W = w/r$, and if the individual is sampling from a known wage offer distribution $F(w)$, the stationary version of eqn [3] satisfies

$$rU = b + a\left(\int \max\,(w/r, U)\mathrm{d}F(w) - U\right) \quad [4]$$

The option to accept a job offer is taken if $w/r \geq U$, giving the *reservation wage* equation.

$$\xi = rU \quad [5]$$

The reservation wage is defined as the minimum acceptable wage, and it is obtained as the solution to eqns [4] and [5].

Partial models of search and empirical research in the duration of unemployment have explored generalized forms of eqn [4] to derive the properties of transitions of individuals from unemployment to employment (see Devine and Kiefer, 1991). For a known $F(w)$ with upper support A, eqn [4] specializes to

$$\xi = b + \frac{a}{r}\int_{\xi}^{A}(w - \xi)\mathrm{d}F(w) \quad [6]$$

Various forms of eqn [6] have been estimated by the empirical literature or used in the construction of partial models of the labor market. The transition from unemployment to employment (the unemployment 'hazard rate') is $a(1 - F(\xi))$ and it depends both on the arrival of offers and on the individual's reservation wage. a, r, and the parameters of the wage offer distribution can be made to depend on the individual's characteristics. The empirical literature has generally found that unemployment compensation acts as a disincentive on individual transitions through the influence of b on the reservation wage, but the effect is not strong. The offer arrival rate increases the hazard rate, despite the fact that the reservation wage increases in a. A number of personal characteristics influence reservation wages and transitions, including age, education, and race.

Two-Sided Matching for Given Wages

Recent work in the economics of search has focused mainly on the equilibrium implications of frictions and search decisions. An equilibrium model needs to specify the decisions of firms and solve for the offer arrival rate a. In addition, a mechanism is needed to ensure that search is an ongoing process. The latter is achieved by introducing a probability $\lambda\delta t$ that a negative shock will hit a job during a short time interval δt. When the negative shock arrives, the job is closed down ('destroyed'), and the worker has to search again to find another job. For the moment, λ is assumed to be a positive constant (see Section Job Destruction).

The Aggregate Matching Function

To derive the equilibrium offer arrival rate, suppose that at time t there are u unemployed workers and v vacant jobs. In a short time interval δt each unemployed worker moves to employment with probability $a\delta t$, so in a large market the total flow of workers from unemployment to employment, and the total flow of jobs from vacant state to production, are both $au\delta t$. A key assumption in the equilibrium literature is that the total flows satisfy an *aggregate matching function*. The aggregate matching function is a black box that gives the outcome of the search process in terms of the inputs into search. If the u unemployed workers are the only job seekers and they search with fixed intensity of one unit each, and firms also search with fixed intensity of one unit for each job vacancy, the matching function gives

$$m = m\,(u, v) \quad [7]$$

with m standing for the flow of matches, au. The function is usually assumed to be continuous and differentiable, with positive first partial derivatives and negative second derivatives, and to satisfy constant returns to scale (see Petrongolo and Pissarides, 2000, for a review). A commonly used matching function in the theoretical literature, derived from the assumption of uncoordinated random search, is the exponential

$$m = v\left(1 - e^{-ku/v}\right), \quad k > 0 \quad [8]$$

The empirical literature, however, estimates a log-linear (constant elasticity) form, which parallels the Cobb–Douglas production function specification, with the elasticity on unemployment estimated in the range 0.5 – 0.7.

The fact that job matching is pairwise implies that the transition rates of jobs and workers are related to Poisson processes. Given $au = m$, the rate at which workers find jobs is $a = m/u$. If q is the rate of arrival of workers to vacant jobs, then total job flows $vq = au$, and so $q = m/v$. The equilibrium literature generally ignores individual differences and treats the average rates m/u and m/v as the rates at which jobs and

workers, respectively, arrive to each searching worker and vacant job. By the properties of the matching function,

$$q = m\left(\frac{u}{v}, 1\right) \quad [9]$$

$$\equiv m\left(\theta^{-1}, 1\right) \equiv q(\theta) \quad [10]$$

with $q'(\theta) < 0$ and elasticity $-\eta(\theta) \in (-1, 0)$. Here, θ is a measure of the *tightness* of the market, of the ratio of the inputs of firms into search to the inputs of workers. Similarly, the transition rate of workers is

$$a = m\left(1, \frac{v}{u}\right) \equiv \theta q(\theta) \quad [11]$$

By the elasticity properties of $q(\theta)$, $\partial a/\partial\theta > 0$. In the steady state, the inverse of the transition rates, $1/q(\theta)$ and $1/\theta q(\theta)$, are the expected durations of a vacancy and unemployment, respectively.

The influence of tightness on the transition rates is independent of the level of wage rates. If there are more vacant jobs for each unemployed worker, the arrival rate of workers to the typical vacancy is lower and the arrival rate of job offers to the typical unemployed worker is higher, irrespective of the level of wages. When a worker and a firm meet and are considering whether or not to stay together, they are not likely to take into account the implications of their action for market tightness and the transition rates of other unmatched agents. For this reason, the influence of tightness on the transition rates is known as a *search externality*. Several papers in the economics of search have explored the efficiency properties of equilibrium given the existence of the externality (see Section Competitive Search and Diamond, 1982; Mortensen, 1982; Pissarides, 1984; Hosios, 1990).

Job Creation

A job is an asset owned by the firm and is valued in a perfect capital market characterized by the same risk-free interest rate r. Suppose that in order to recruit a worker, a firm has to bear set-up cost K to open a job vacancy and in addition has to pay a given flow cost c for the duration of the vacancy. The flow cost can be interpreted as an advertising and recruitment cost, or as the cost of having an unfilled position in what might be a complex business environment (which is not modeled).

Let V be the value of a vacant position and J the value of a filled one. Reasoning as in the case of the value of a job seeker, U, the Bellman equation satisfied by V is

$$rV = -c + q(\theta)(J - V) \quad [12]$$

The vacant job costs c and the firm is offered the option to take a worker at rate $q(\theta)$.

Since the firm has to pay set-up cost K to create a job, it will have an incentive to open a job if (and only if) $V \geq K$. The key assumption made about job creation is that the gains from job creation are always exhausted, so jobs are created up to the point where

$$V = K \quad [13]$$

Substitution of V from eqn [12] into eqn [13] yields

$$J = K + \frac{rK + c}{q(\theta)} \quad [14]$$

Competition requires that the expected present discounted value of profit when the worker arrives, the value of a filled job J, should be just sufficient to cover the initial cost K and the accumulated costs for the duration of the vacancy, interest on the initial outlay rK, and ongoing costs c for the expected duration of the vacancy, $1/q(\theta)$.

Let productivity be a constant p in all jobs and the wage rate a constant w. With break-up rate λ, the value of a job satisfies the Bellman equation

$$rJ = p - w - \lambda J \quad [15]$$

The flow of profit to the firm is $p - w$ until a negative shock arrives that reduces its value to 0. Replacing J in eqn [14] by its expression in eqn [15] yields the *job creation condition*

$$\frac{p - w}{r + \lambda} - K - \frac{rK + c}{q(\theta)} = 0 \quad [16]$$

Equation [16] determines θ for each w and parallels the conventional labor demand curve. A higher wage rate makes it more expensive for firms to open jobs, leading to lower market tightness. Frictions slow down the arrival of suitable workers to vacant jobs and so the firm incurs some additional recruitment costs. If the arrival rate is infinitely fast, as in Walrasian economics, $q(\theta)$ is infinite and the last term in eqn [16] disappears. The assumptions underlying eqn [13] ensure that at the margin, the recruitment cost is just covered.

Wage Setting

With search frictions, there is no supply of labor that can be equated with demand to give wages. The conventional supply of labor is constant here: there is a fixed number of workers in the market, which is usually normalized to unity, and each supplies a single unit of labor.

In search equilibrium, there are local monopoly rents. Firms and workers who are together can start production immediately. If they break up, they can start production only after they find another partner through an expensive process of search. The value of the search costs that they save by staying together correspond to a pure economic rent; they could be taken away from them and they would still stay together. Wages need to share those rents.

Two approaches dominate in the literature. The first and more commonly used approach employs the solution to a Nash bargaining problem (see Diamond, 1982; Pissarides, 2000). The Nash solution allocates the rents according to each side's 'threat points.' The threat points in this case are the returns from search, and the Nash solution gives to each side the same surplus over and above their expected returns from search. Generalizing this solution concept, wages are determined such that the net gain accruing to the worker from the match, $W - U$, is a fixed proportion β of the total surplus, the sum of the worker's and the firm's surplus, $J - V$. This sharing rule can be obtained as the maximization of the product

$$(W - U)^{\beta}(J - V)^{1-\beta}, \quad \beta \in (0, 1) \quad [17]$$

The coefficient β can be given the interpretation of bargaining strength, although strictly speaking, bargaining strength in the

conventional Nash solution is given by the threat points U and V (see Binmore et al., 1986, for an interpretation of β in terms of rates of time preference).

The second approach to wage determination postulates that the firm 'posts' a wage rate for the job, which the worker either takes or leaves. The posted wage may be above the worker's reservation wage because of some 'efficiency wage' type arguments, for example, in order to reduce labor turnover, encourage more effort or attract more job applicants.

Bargaining

The value of an employed worker, W, satisfies the Bellman equation.

$$rW = w - \lambda(W - U) \qquad [18]$$

The worker earns wage w and gives up the employment gain $W - U$ when the negative shock arrives. No worker has an incentive to quit into unemployment or search for another job whilst employed, for as long as $w \geq b$, which is assumed.

With eqns [15] and [17] in place, the Nash bargaining solution to eqn [18] gives the sharing rule

$$W - U = \beta(W - U + J - V) \qquad [19]$$

Making use of the value equations and sharing rule yields

$$w = rU + \beta(p - (r + \lambda)K - rU) \qquad [20]$$

$$= (1 - \beta)b + \beta[p - (r + \lambda)K + (rK + c)\theta] \qquad [21]$$

There is a premium on the reservation wage which depends on the worker's bargaining strength and the net surplus produced. Wages depend positively on unemployment income and the productivity of the job, the first because of the effect that unemployment income has on the cost of unemployment and the second because of the monopoly rents and bargaining. Wages also depend on market tightness. In more tight markets they are higher, because the expected duration of unemployment in the event of disagreement is less. Empirical evidence supporting this wage equation has been found by a number of authors (e.g., Blanchflower and Oswald, 1994), although it should be noted that similar wage equations can also be derived from other theoretical frameworks.

Equilibrium

Equation [21] replaces the conventional labor supply curve and closes the system. When combined with the job creation condition eqn [16] it gives unique solutions for wages and market tightness. A variety of intuitive properties are satisfied by this equilibrium. For example, higher labor productivity implies higher wages and tightness; higher unemployment income implies higher wages but lower tightness. It remains to obtain the employment rate in equilibrium.

The labor force size is fixed, so by appropriate normalizations, if at some time t unemployment is u_t, employment is $1 - u_t$. In a short time interval δt, $a_t u_t \delta t$ workers are matched and $\lambda(1 - u_t)\delta t$ workers lose their jobs. Given that $a_t = \theta_t q(\theta_t)$, the evolution of unemployment is given by

$$u_{t+\delta t} = u_t + \lambda\,(1 - u_t)\delta t - \theta_t q\,(\theta_t) u_t \delta t \qquad [22]$$

Dividing through by δt and taking the limit as $\delta t \to 0$ yields

$$\dot{u} = \lambda(1 - u) - \theta q\,(\theta) u \qquad [23]$$

Because the solution for θ is independent of u, this is a stable differential equation for unemployment with a unique equilibrium

$$u = \frac{\lambda}{\lambda + \theta q\,(\theta)} \qquad [24]$$

Equation [24] is often referred to as the Beveridge curve, after William Beveridge who first described such a 'frictional' equilibrium. Plotted in space with vacancies on the vertical axis and unemployment on the horizontal is a convex-to-the origin curve. In the early literature the impact of frictions on the labor market was measured by the distance of this curve from the origin (see also Pissarides, 2000; Blanchard and Diamond, 1989).

Wage Posting

Wage posting is an alternative to wage bargaining, with different implications for search equilibrium. The firm posts a wage for the job and the worker who searches it either takes it or leaves it. Three different models of wage posting are examined.

Single Offers, No Prior Information

Workers search sequentially one firm at a time, they discover the firm's offer after they have made contact and they have to accept it or reject it (with no recall) before they can sample another firm. The worker who contacts a firm that posts wage w_i has two options. Accept the wage offer of the firm and enjoy expected return W_i, obtained from eqn [18] for $w = w_i$, or reject it and continue search for return U. The firm that maximizes profit chooses w_i subject to $W_i \geq U$. Since the firm will have no incentive to offer the worker anything over and above the minimum required to make workers accept its offer, in this model wages are driven to the worker's reservation wage, $w_i = b$ (Diamond, 1971).

In terms of the bargaining solution, the 'Diamond' (see eqns [20, 21]) equilibrium requires $\beta = 0$. The model can then be solved as before, by replacing β by 0 in the job creation condition and wage equation. Tightness, and consequently unemployment, absorb all shocks other than those operating through unemployment income, which also change wages. (There is a paradox in this model, in that if wages are equal to unemployment income no worker will have an incentive to search.)

Competitive Search

The next model increases the amount of information that workers have before they contact the firm (Moen, 1997). Workers can see the wage posted by each firm but because of frictions they cannot be certain that they will get the job if they apply. If the probability of a job offer across firms is the same, all workers will apply for the highest-wage job. Queues will build up and this will reduce the probability of an offer. In equilibrium, the wage offer and queue characterizing each job have to balance each other out, so that all firms get applicants.

The length of the queue is derived from the matching process. The firm that posts wage w_i makes an offer on average after $1/q(\theta_i)$ periods and a job applicant gets an offer on average after $1/\theta_i q(\theta_i)$ periods. Implicit in this formulation is the assumption that more than one firm offers the same wage, and firms compete for the applicants at this wage. Workers apply to only one job at a time.

Suppose now there is a firm, or group of firms, such that when workers join their queue they derive expected income stream $\overline{U}$, the highest in the market. The constraint facing a firm when choosing its wage offer is that the worker who applies to it derives at least as much expected utility as $\overline{U}$. The expected profit of the firm that posts wage w_i solves

$$rV_i = -c + q(\theta_i)(J_i - V_i) \qquad [25]$$

$$rJ_i = p - w_i - \lambda J_i \qquad [26]$$

The worker's expected returns from applying to the firm posting this wage satisfy the system of equations

$$rU_i = b + \theta_i q(\theta_i)(W_i - U_i) \qquad [27]$$

$$rW_i = w_i - \lambda\left(W_i - \overline{U}\right) \qquad [28]$$

The firm chooses w_i to maximize V_i subject to $U_i \geq \overline{U}$. The first-order maximization conditions imply that all firms offer the same wage, which satisfies

$$W - U = \frac{\eta}{1-\eta}(J - V) \qquad [29]$$

Comparison of eqn [29] with eqn [19] shows that the solution is indeed similar to the Nash solution, but with the share of labor given by the (negative of the) elasticity of $q(\theta)$, (which is equal to the unemployment elasticity of the underlying matching function). The rest of the model can be solved as in the Nash case.

There is a special significance to the share of labor obtained in this formulation. In the case where wages are determined according to the Nash rules, the firm and worker choose the wage after they meet, so it is unlikely that they will internalize the search externalities; they do not take into account the effect of their choices on the transition rates of unmatched agents. It can be shown that with constant returns to scale there is a unique internalizing rule which requires $\beta = \eta$, the solution of the wage posting model considered here (Hosios, 1990; Pissarides, 1984, 2000). For this reason, this particular wage posting model is often called the *competitive search* equilibrium. The key assumption that gives efficiency is the relaxation of the informational restrictions on workers, which lets them know both the firm's wage offer and the length of the queue associated with it.

Wage Differentials

The third version of the wage posting model relaxes the assumption that workers have the choice of at most one wage offer at a time but does not allow workers knowledge of the wage offer before they apply (Burdett and Judd, 1983; Burdett and Mortensen, 1998; Montgomery, 1991). The easiest way to introduce this is to allow workers the possibility of search on the job, i.e., to let them continue looking for a better job after they accepted one. Suppose for simplicity that job offers arrive to searching workers at the same rate a, irrespective of whether they are employed or unemployed. Suppose also that job search is costless (except for the time cost) and job changing is costless. Then, if a worker is earning w now and another offer paying w' comes along, the worker accepts the new offer if (and only if) $w' > w$. The worker's reservation wage is the current wage.

Unemployment pays b, so no firm can pay below b and attract workers. Consider a firm that pays just above b. Anyone applying for a job from the state of unemployment will accept its offer, but no one else will. Employed job seekers will have no incentive to quit their jobs to work at a wage close to b. In addition, this firm's workers will be quitting to join other firms, which may be paying above b. So a firm paying a low wage will have high turnover and will be waiting long before it can fill its vacancies.

A firm paying a high wage will be attracting workers both from unemployment and from other firms paying less than itself. Moreover, it will not be losing workers to other firms. So high-wage firms will have fewer vacant positions.

Now suppose the two firms have access to the same technology. The low-wage firm enjoys a lot of profit from each position, but has a lot of vacant positions. The high-wage firm enjoys less profit from each position, but has them filled. It is possible to show under general conditions that high wage and low wage firms will co-exist in equilibrium.

Burdett and Mortensen (1998) show this by assuming that there is a distribution of wage offers for homogeneous labor, $F(w)$, and demonstrating that (1) no two firms will offer the same wage, (2) firms can choose any wage between the minimum b and a maximum $\overline{w}$, and enjoy the same profit in the steady state. The maximum is given by

$$\overline{w} = p - \left(\frac{\lambda}{a+\lambda}\right)^2 (p - b) \qquad [30]$$

where as before p is the productivity of each worker and λ is the job destruction rate.

The interesting result about wage posting in this model is that once the assumption that the worker can only consider one offer at a time is relaxed, a distribution of wage offers for homogeneous labor arises. The distribution satisfies some appealing properties. As $a \to 0$, the upper support tends to $\overline{w} = b$, the Diamond solution. At the other extreme, as $a \to \infty$, it can be shown that all wage offers converge to p, the competitive solution. Intuitively, $a = 0$ maximizes the frictions suffered by workers and $a = \infty$ eliminates them.

Job Destruction

A body of empirical literature shows that there is a lot of 'job churning,' with many jobs closing down and new ones opening to take their place. It is also found that both the job creation and job destruction rates, especially the latter, vary a lot over the cycle (see Leonard, 1987; Dunne et al., 1989; Davis et al., 1996). This is an issue addressed by search theorists. Returning to the model with a Nash wage rule, the job creation flow is the matching rate $m(u, v)$. The job creation *rate* is defined as the job creation flow divided by employment, $m(u, v)/(1 - u)$, which is

$$\frac{m(u,v)}{1-u} = \theta q(\theta)\frac{u}{1-u} \qquad [31]$$

Since both θ and u are endogenous variables of the model that respond to shocks, the model predicts a variable job creation rate that can be tested against the data (see Mortensen and Pissarides, 1994; Cole and Rogerson, 1999).

But the job destruction flow is $\lambda(1-u)$, so the job destruction rate is a constant λ, contrary to observation. Two alternative ways of making it variable are considered.

Idiosyncratic Shocks

In the discussion so far, the productivity of a job is p until a negative shock arrives that reduces it to zero. Generalizing this idea, suppose that although initially the productivity is p, when a shock arrives it changes it to some other value px. The component p is common to all jobs but x is specific to each job. It has distribution $G(x)$ in the range [0, 1]. New jobs start with specific productivity $x = 1$; over time shocks arrive at rate λ that transform this productivity to a value between 0 and 1, according to the distribution G. The firm has the choice of either continuing to produce at the new productivity, or closing the job down. The idea is that initially firms have a choice over their product type and technique and choose the combination that yields maximum productivity. Over time techniques are not reversible, so if the payoffs from a given choice change, the firm has the choice of either continuing in the new environment or destroying the job.

As in the case where workers are faced with take-it-or-leave-it choices from a given wage distribution, Mortensen and Pissarides (1994) show that when the firm has the choice of either taking a productivity or leaving it, its decision is governed by a reservation productivity, denoted R. The reservation productivity depends on all the parameters of the model. Profit maximization under the Nash solution to the wage bargain implies that both workers and firms agree about the optimal choice of R, i.e., which jobs should be closed and which should continue in operation. With knowledge of R, the flow of job closures generalizes to $\lambda G(R)(1-u)$, the fraction of jobs that get shocks below the reservation productivity. The job destruction rate then becomes $\lambda G(R)$, which responds to the parameters of the economy through the responses of R to shocks.

An interesting feature of the job creation and job destruction rates, which conforms to observation, is that because the job creation rate depends on unemployment, which is a slow-moving variable, whereas the job destruction rate does not depend on it, the job destruction rate is more volatile than the job creation rate. For example, a rise in general productivity p, associated with a positive cyclical shock, reduces job destruction immediately by reducing the reservation productivity, but increases job creation at first and then reduces it, as tightness rises at first but unemployment falls in response.

Technological Progress and Obsolescence

Another way of modeling job destruction borrows ideas from Schumpeter's theory of growth through creative destruction (see Aghion and Howitt, 1994; Caballero and Hammour, 1994). Jobs are again created at the best technology but once created, their technology cannot be updated. During technological progress, the owners of the jobs have the option of continuing in production with the initial technology or closing the job down and opening another, more advanced one. As new jobs are technologically more advanced, the wage offers that workers can get from outside improve over time. There comes a time when the worker's outside options have risen sufficiently to render the job obsolete.

Formally, the model can be set up as before, with the technology of the job fixed at the frontier technology at creation time and wages growing over time because of growth in the returns from search. The value of a job created at time 0 and becoming obsolete at T is

$$J_0 = \int_0^T e^{-(r+\lambda)t}[p(0) - w(t)]\mathrm{d}t \qquad [32]$$

As before, r is the discount rate and λ is the arrival rate of negative shocks that may lead to earlier job destruction. $p(0)$ is the initial best technology and $w(t)$ the growing wage rate. The job is destroyed when $w(t)$ reaches $p(0)$; i.e., the job life that maximizes J_0 is defined by $w(T^*) = p(0)$.

A useful restriction to have when there is growth is to assume that both unemployment income and the cost of recruitment grow at the exogenous rate of growth of the economy. As an example, consider the wage eqn [21] with the restriction $K = 0$ and $b(t) = be^{gt}$, $c(t) = ce^{gt}$, where g is the rate of growth of the economy. T^* then satisfies

$$e^{gTA*} = \frac{(1-\beta)p}{(1-\beta)b + \beta c\theta} \qquad [33]$$

Jobs are destroyed more frequently when growth is faster, when unemployment income is higher and when the tightness of the market is higher.

Job destruction in this model has two components, the jobs destroyed because of the arrival of shocks, $\lambda(1-u)$, and those destroyed because of obsolescence. The latter group was created T^* periods earlier and survived to age T^*, $\theta q(\theta)ue^{-\lambda T^*}$ (note that in the steady state the job creation flow is a constant $\theta q(\theta)u$). Therefore, the job destruction rate now is $\lambda + \theta q(\theta)ue^{-\lambda T^*}/(1-u)$, which varies in response to changes in T^* but also in response to changes in the job creation rate $\theta q(\theta_u)/(1-u)$.

See also: Behavioral Economics; Consumption, Sociology of; Information, Economics of; Labor Supply; Market Structure and Performance; Transaction Costs and Property Rights; Wage Differential and Structure; Work, Sociology of.

Bibliography

Aghion, P., Howitt, P., 1994. Growth and unemployment. Review of Economic Studies 61, 477–494.

Binmore, K.G., Rubinstein, A., Wolinsky, A., 1986. The Nash bargaining solution in economic modelling. Rand Journal of Economics 17, 176–188.

Blanchard, O.J., Diamond, P.A., 1989. The Beveridge curve. Brookings Papers on Economic Activity 1, 1–60.

Blanchard, O.J., Diamond, P.A., 1990. The cyclical behavior of the gross flows of US workers. Brookings Papers on Economic Activity 2, 85–155.

Blanchflower, D.G., Oswald, A.J., 1994. The Wage Curve. MIT Press, Cambridge, MA.

Burdett, K., Judd, K., 1983. Equilibrium price distributions. Econometrica 51, 955–970.

Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size, and unemployment. International Economic Review 39, 257–273.
Caballero, R.J., Hammour, M.L., 1994. The cleansing effect of recessions. American Economic Review 84, 1350–1368.
Cole, H.L., Rogerson, R., 1999. Can the Mortensen–Pissarides matching model match the business cycle facts? International Economic Review 40, 933–959.
Davis, S.J., Haltiwanger, J.C., Schuh, S., 1996. Job Creation and Destruction. MIT Press, Cambridge, MA.
Devine, T.J., Kiefer, N.M., 1991. Empirical Labor Economics: The Search Approach. Oxford University Press, Oxford.
Diamond, P.A., 1971. A model of price adjustment. Journal of Economic Theory 3, 156–168.
Diamond, P.A., 1982. Wage determination and efficiency in search equilibrium. Review of Economic Studies 49, 217–227.
Dunne, T., Roberts, M.J., Samuelson, L., 1989. Plant turnover and gross employment flows in the manufacturing sector. Journal of Labor Economics 7, 48–71.
Hosios, A.J., 1990. On the efficiency of matching and related models of search and unemployment. Review of Economic Studies 57, 279–298.
Leonard, J.S., 1987. In the wrong place at the wrong time: the extent of frictional and structural unemployment. In: Lang, K., Leonard, J.S. (Eds.), Unemployment and the Structure of Labor Markets. Basil Blackwell, New York.
McCall, J.J., 1970. Economics of information and job search. Quarterly Journal of Economics 84, 113–126.
Moen, E.R., 1997. Competitive search equilibrium. Journal of Political Economy 105, 385–411.
Montgomery, J., 1991. Equilibrium wage dispersion and interindustry wage differentials. Quarterly Journal of Economics 106, 163–179.
Mortensen, D.T., 1982. The matching process as a non-cooperative/bargaining game. In: McCall, J.J. (Ed.), The Economics of Information and Uncertainty. University of Chicago Press, Chicago, IL.
Mortensen, D.T., Pissarides, C.A., 1994. Job creation and job destruction in the theory of unemployment. Review of Economic Studies 61, 397–415.
Mortensen, D.T., Pissarides, C.A., 1999a. Job reallocation, employment fluctuations, and unemployment. In: Woodford, M., Taylor, J. (Eds.), Handbook of Macroeconomics. North-Holland, Amsterdam, pp. 1171–1228.
Mortensen, D.T., Pissarides, C.A., 1999b. New developments in models of search in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam.
Petrongolo, B., Pissarides, C.A., 2000. Looking into the black box: a survey of the matching function. Journal of Economic Literature 39, 390–431.
Phelps, E.S., et al., 1970. Microeconomic Foundations of Employment and Inflation Theory. Norton, New York.
Pissarides, C.A., 1984. Efficient job rejection. Economic Journal 94, 97–108.
Pissarides, C.A., 1985. Short-run equilibrium dynamics of unemployment, vacancies, and real wages. American Economic Review 75, 676–690.
Pissarides, C.A., 2000. Equilibrium Unemployment Theory, second ed. MIT Press, Cambridge, MA.
Stigler, G.J., 1961. The economics of information. Journal of Political Economy 69, 213–225.

Searle, John (1932–)

Alessandro Salice, Center for Subjectivity Research, University of Copenhagen, Copenhagen S, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

During the last 50 years, John R. Searle has developed influential views on fundamental research areas within and beyond philosophy. This article mainly focuses on Searle's contributions to three fields of research. In the first section, John Searle's concept of a speech act is tackled. Topic of the second section is Searle's theory of mind and of intentionality. Finally, the third part addresses Searle's 'Philosophy of Society' (a discipline, which broadly covers what nowadays also falls under the name of 'social ontology').

John Rogers Searle is currently Willis S. and Marion Slusser Professor of Philosophy at the University of California, Berkeley. Searle was born on 31 July 1932, in Denver, Colorado, and began his studies at the University of Wisconsin–Madison in 1949. In 1952, he was awarded a Rhodes Scholarship from the University of Oxford, where he obtained a BA in 1955. Before receiving his PhD in 1959, he served as a lecturer in philosophy at the Christ Church College, Oxford, where he came in contact with John L. Austin, Peter Strawson, and other important representatives of what has been labeled 'ordinary language philosophy.' Soon after receiving his doctorate, Searle started teaching at Berkeley. In the mid-1960s, he became the first tenured professor to join the Free Speech Movement (FSM). The FSM was a protesting group of students that, as one of its main achievements, managed to extend the student rights to a significant extent including the – previously banned – liberty to use the University campus for political activity and debate (this experience is described in his 1972 book).

Over the years, he has been granted a number of prestigious prizes and honors from numerous institutions, including the most recent Puffendorf Medal (Sweden) in 2006; the Mind and Brain Prize (Italy) in 2006; the National Humanities Medal (United States) in 2004; and the Jovellanos Prize (Spain) and Jean Nicod Prize (France) in 2000.

Philosophy of Language

In the philosophy of language, Searle developed a systematic theory built around a notion of speech acts that theoretically refines and improves the concept introduced by John L. Austin in the 1950s (see his 1962; and, before him, by Adolf Reinach in German phenomenology, cf his 1913). According to Austin (1962), every speech act is composed of a locutionary and an illocutionary act (and, in some cases, of a perlocutionary act). For instance, if Sam utters the words "I'll be there," then his act can be analyzed as follows: First, he is uttering a meaningful sentence, i.e., he is performing a locutionary act. The locutionary act is itself a complex act. Roughly, this consists in Sam producing certain sounds (*phonetic act*), in these sounds being used as *words* – i.e., in being used as belonging to the vocabulary and as conforming to the grammar of a natural language (*phatic act*) – and in these words being used with a sense and a reference (*rhetic act*). Second, in addition to the locutionary act, Sam is performing an illocutionary act, i.e., he is uttering a sentence with a certain force, e.g., he is promising, announcing, menacing, or forecasting that he will be there. Austin claims that the concepts of locutionary and illocutionary acts are mutually exclusive (i.e., it can never be the case that one act is at once a locutionary and an illocutionary act), but he also adds that every locutionary act presupposes an illocutionary one and vice versa. Third, with his utterance Sam also could have wanted to do something else, e.g., he could have wanted to warn the addressee. In this case, a third, perlocutionary act has been performed.

One of the considerations that lead Searle to dismiss key parts of Austin's analysis is the following. Suppose Sam is more explicit and says, "I promise that I will be there." In this case, the description of the locutionary act is also a description of the illocutionary act, for once one determines the meaning of the sentence, one also determines its force: "The description of the act as a happily performed locutionary act, since it involves the meaning of the sentence, is already a description of the illocutionary act, since a particular illocutionary act is determined by that meaning. They are one and the same act" (Searle, 1968: 407). This goes against Austin's thesis of a relation of mutual exclusivity between locutions and illocutions and motivates Searle in modifying Austin's analysis. In particular, he drops Austin's idea of rhetic acts and substitutes it with that of 'propositional act.' A propositional act expresses a *proposition* – which is conceived of as the truth-bearing content of the speech act.

Searle's introduction of the propositional act has three important consequences. First, the act-theoretic description finds 'utterance acts' (which encompass phonetic and phatic acts), propositional acts, and illocutionary acts as the building blocks of every speech act (Searle, 1969: 23f.). Second, the notion of proposition provides Searle with the possibility to distinguish the speaker's meaning from the conventional or sentence meaning. Contrary to what Paul Grice has argued, namely that the meaning of a sentence is to be defined solely in terms of the speaker's intention to produce some effect in the hearer by means of the recognition of this intention (cf Grice, 1957), Searle recognizes that – in addition to this meaning – the speaker's utterance expresses a proposition, which identifies the literal or sentence meaning of the utterance (Searle, 1969: 44f.; for Searle's most recent account of speaker's and sentence meaning, see his 2010: 73ff.).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61282-5

It should be noted, however, that Searle improved his theory over the years by arguing that sentence meaning is not absolutely context-free: according to his more recent view, basically all forms of representation (whether linguistic or not) go "on against a background of assumptions which are not and in most cases could not also be completely represented as part of or as presuppositions of the representation" (1979a: 131), i.e., they presuppose the Background. Searle's point is that there is no 'zero context' or 'null context' for the interpretation of sentences (Searle, 1979a: 117): for instance, "[i]f you consider the sentence 'Cut the grass!' you know that this is to be interpreted differently from 'Cut the cake!' If somebody tells me to cut the cake and I run over it with a lawn mower or they tell me to cut the grass and I rush and stab it with a knife, there is a very ordinary sense in which I did not do what I was told to do. Yet nothing in the literal meaning of those sentences blocks those wrong interpretations" (Searle, 1995: 130f.). The conclusion to be drawn from these examples is that we understand the verb of each sentence differently, because in each case our interpretation depends on the Background, i.e., it depends on a set of abilities and capacities that we acquired along the years and are now part of our mental equipment (see below for further details on this notion).

The third consequence is that the previous two points (the structure of a speech act and the role played in it by the notion of proposition) speak for the idea that one and the same proposition can be expressed in connection with different forces. For instance, the utterances of "Sam smokes habitually," "Does Sam smoke habitually?" and "Would that Sam smoked habitually" are used to perform different illocutionary acts that bear the same proposition. This insight is at the basis of Searle's notation:

$$F(p)$$

where 'F' indicates the force of an utterance and 'p' the proposition expressed. Improved with the concepts of 'illocutionary point or purpose,' 'expressed psychological state,' and 'direction of fit,' this notation is at the basis of Searle's taxonomy of speech acts (cf 1979b). The illocutionary point is what Searle also calls the 'essential condition' of a speech act and it roughly identifies the reason for which the act is performed. So, for example, the point of promises is to undertake an obligation, the point of a description is to match the world, etc. In addition to essential conditions, speech acts also have 'sincerity conditions.' They are sincere if the speaker actually has the psychological state he or she expresses. If one communicates something to someone, the speaker is taken to believe the information that he or she is conveying (and if the speaker does not have this belief, he or she is lying). Similarly, in the case of promises, sincere promises are those which the promisor also intends to honor. That is, every sincere speech act expresses a corresponding psychological state, the lack of which makes the act insincere.

The concept of 'direction of fit' (whose intuition Searle credits to Elizabeth Anscombe, cf her 1957, and whose term he took from Austin, cf his 1953) is intended to capture how the propositional content is supposed to relate to the world. There are three standard kinds of direction of fit (DoF) and one 'nonstandard' kind. The standard kinds are: the words-to-world, the world-to-words, and the double DoFs. The words-to-world DoF characterizes all that acts like descriptions, prognoses, hypotheses, etc., whose aim is to fit the world – Searle also says that in these cases, it is the responsibility of the words to fit the world (cf 2010: 28). The world-to-words DoF identifies all those acts where the responsibility of fitting lies on the world's side, so to speak – these are acts like orders, commands, questions, invitations, etc. Finally, the double DoF is a feature that qualifies the particular class of declarations: declarations have the power to generate facts by representing them as obtaining. For instance, if the officiant declares Sam and Sarah as husband and wife, Sam and Sarah *are* husband and wife. (In Section Philosophy of Society we will turn back to this point.) In addition to these kinds of DoF, there are also speech acts showing a nonstandard DoF: suppose you apologize for something, in this case the act simply *presupposes* that something has been the case (i.e., the event for which you are apologizing). Searle initially believed that these sorts of acts have no DoF at all (and hence spoke of a 'null' DoF, cf 1979a). Later he came to the idea that it is not true that they are not completely unrelated to the world; rather, this relation is *presupposed* (the new terminology is hence that of a "Presup [for: 'presupposition'] Fit," cf 2010: 29).

Another important contribution to speech act theory is the distinction Searle draws between the conventional and the rule-constituted dimension of every speech act. Austin put the notion of convention at the very core of the idea of a speech act (a speech act is a 'conventional or ritual' act, cf Austin, 1962: 14). By contrast, Searle contends that, of course, speech acts have conventional aspects, but these aspects are not the ultimate constituents of a speech act. Rather, these constitutive aspects are to be sought in the rules that define what a speech act is. So, for example, one can express one and the same promise with the French words 'je promets' and the English words 'I promise.' Expressing the promise in French or in English is a matter of convention and it is not a constitutive feature of the promise: by contrast, the fact that to promise something counts as the undertaking of an obligation is a constitutive feature of promises and, more precisely, it is a matter of rules. The rules that define speech acts are a particular kind of rules, they are 'constitutive rules' (cf Searle, 1969: 33ff.).

The easiest way to understand what constitutive rules are is to consider them in contradistinction with another kind of rules, i.e., the regulative rules. As the expression already suggests, regulative rules regulate a form of behavior that could have taken place also without these rules. Take driving as an example: the rule to drive on the right side of the street is regulative in the sense that the activity of driving does not logically depend on this rule (people can drive cars also without it). By contrast, constitutive rules make a given activity possible in the sense that without such rules one does not have even the possibility to engage in the activity at issue. Playing chess is the main example here: without the rules of chess, one could not play the game. Characteristically, constitutive rules have the form 'X counts as Y in context C' (e.g., 'to make an order counts as imposing an obligation to the addressee'), whereas regulative rules have the form of imperatives (e.g., 'officers must wear ties at dinner').

Philosophy of Mind

According to Searle, to be fully conclusive, the considerations on speech acts have to be supplemented by a theory of

intentionality. Although words and sentences refer to (or are about) objects and states of affairs, the sense of this reference or aboutness is not primitive; rather, it is secondary, or 'derived.' It is derived, Searle claims, because it presupposes that the *mind* has the power to refer to something. To put this idea differently: words and sentences refer to something because a mind *imposes* this power on them (cf Searle, 1983: 26ff.; one could also express this point by saying that the representational capacities of words and sentences are extrinsic, whereas the representational capacity of the mind is *intrinsic*, cf Searle, 1980b).

To characterize this power or this capacity of the mind, Searle employs the term 'intentionality,' which he uses in the technical sense with which it was (re)introduced by Franz Brentano in post-Kantian philosophy (cf Brentano, 1874), i.e., to denote the feature of mental states to refer to objects and states of affairs in the world. So, for instance, perceptions and beliefs are intentional because they are always about something. In this sense, 'intentionality' has no privileged link to that specific class of mental states encompassing intentions, although it can be said that intentions, too, are intentional since they are directed toward or are about something.

We said that intentionality is a property of the mind; this suggests that we first need to say something about the mind before addressing the more specific question about the nature of intentionality. According to a very influential approach in the philosophy of mind – call it the 'strong artificial intelligence' approach, cf Searle 1980a, or the 'computer functionalist' approach, cf Dennett 1978 – the mind is nothing other than a computer program implemented in brains. To function, computer programs require hardware and, in the case of the mind, the hardware is the brain: since the description of the hardware is not the description of a program, so the description of the brain is not that of the mind. Still, the mind is not an entity separated from the brain; it is merely the program running on hardware and can therefore be accounted for in terms of its inputs and outputs.

Searle rejects the 'strong artificial intelligent' approach by developing an argument known as the 'Chinese room' argument (Searle, 1980a). It is based on a thought experiment, which, in its compact form, runs as follows: imagine a person who understands no Chinese is locked in a room. From the outside, she is handed a batch of Chinese writings containing questions in Chinese (the 'inputs'). Inside the room, she finds another batch of Chinese scripts and a set of rules written in English (the 'program') telling how to correlate the symbols in the first batch with the symbols in the second (the output or 'answers'). The rules tell how to process the questions in such a way that it is not possible for an external observer to tell whether the answers have been formulated by a program or by a Chinese-speaking person. At this point, Searle asks whether the person locked in the room understands Chinese: the force of the argument is to show that, although the entire situation can be framed as if the person is processing some inputs according to a program, this is not the case: "[b]ecause the program is purely formal or syntactical and because minds have mental or semantic contents, any attempt to produce a mind purely with computer programs leaves out the essential features of the mind" (cf Searle, 1992: 45).

The Chinese room argument leaves us with a negative result: if the mind is not a computer program, what is it? Searle's strategy for answering to this question is to avoid any "obsolete seventeenth-century categories" (Searle, 2008a: 71), meaning allegedly antithetical categories such as 'mental,' 'physical,' 'mind,' 'body,' etc. Searle indeed formulates his answer in overt contrast to these dichotomies (cf Searle, 2008c: 152–160), by claiming that the mind and its features, such as intentionality and consciousness, are nothing other than biological phenomena. This position is labeled 'biological naturalism' (Searle, 1983: 264) and Searle overtly states that we still do not have a clear grasp on how exactly the brain's neurobiology causes mental states and "we are not likely to find the answer in the near future" (1998: 96).

The fulfillment of this research desideratum would not preclude the project of a science of consciousness, though. In fact, mental states are entities of the sort that could be characterized as intrinsically 'subjective.' They are subjective in two senses: according to the first, they are subjective because they are eminently first-person phenomena. That is to say, they have a qualitative shape that is there only for the subject having these mental states and which hence cannot be expressed in a third-person language (2002a: 39f.). (This first sense could be qualified as 'phenomenological,' although Searle has a rather critical stance toward phenomenology as a philosophical movement, cf Searle, 2008b, 2010: 30.) In addition, they are subjective in the ontological sense of the term, i.e., they are necessarily the mental states *of* someone (and in this sense are not ontologically objective, this property being defined as independency from a subject). Despite their ontological subjectivity, though, Searle claims that one can make *epistemically objective* statements about them. "Consider, for example, the statement, 'I now have a pain in my lower back.' That statement is completely objective in the sense that it is made true by the existence of an actual fact and is not dependent on any stance, attitudes, or opinions of observers [this dependency defining the notion of 'epistemic subjectivity,' A.S.]. However, the phenomenon itself, the actual pain itself, has a subjective mode of existence […]" (Searle, 1992: 94). A science of the mind is hence possible because objective statements about subjective entities like mental states are possible. As we will see, one of the features that most cogently characterize most mental states is their intentionality.

A useful way to map mental states is indeed by means of two properties: intentionality and consciousness. Mental states are either intentional or nonintentional (I see that it rains; I am nervous – although I do not know what I am nervous about) and they are either conscious or unconscious (I am conscious of being hungry now, but my belief that today is Thursday is not conscious). Searle hence admits the existence of unconscious mental states, but these have to be characterized as states which in principle might be or might have been conscious (in contrast to those states of the brain that are unconscious and can never became conscious; here Searle has in mind states such as the "secretion of the neurotransmitter norepinephrine into the synaptic cleft," cf 1998: 97). But what makes a given mental state intentional?

Intentional states can be analyzed into two parts: the type of state and its content. Following the notation already presented

for speech acts, one could picture the structure of intentional states in this way:

$$S(p)$$

where S points at the kind of the state (also called the 'mode' of the experience, i.e., its being a desire, a perception, an intention, etc.) and p pictures the content. (Generally the content is a proposition, but there are exceptions where the content is nominal or non-propositional – e.g., admiration for someone, see Searle, 1983: 34.) It is important to highlight that the content of a state is not that toward which the state is directed: intentional states are directed toward objects or states of affairs *through* or via their contents. (So, for instance, the presentation of Pegasus has a content, but it does not have an object.)

Like speech acts, intentional states, too, have a direction of fit: e.g., beliefs have a mind-to-world DoF, and intentions or desires have the world-to-mind DoF. Intentional states with a propositional content and a direction of fit are also said to *represent conditions of satisfaction* (on the specificity of this notion of 'representation,' see Searle, 1983: 12). Accordingly, beliefs, desires, etc. 'represent' their conditions of satisfaction in the sense that they represent those conditions, which must obtain in order for these states to be satisfied. If one believes that it is raining, the state of affairs that it is raining must obtain for the belief to be satisfied. Analogously, a desire represents its conditions of satisfaction and these are what, again, must obtain for the desire to be satisfied.

One of the most debated aspects of Searle's theory of intentionality concerns his claim that there is a primitive and irreducible form of intentionality, which is *collective*. That is to say, not only I – as an individual – can intend, accept, believe … *p*, but there is a sense in which one can say that *we* (collectively) intend, accept, believe … *p*. Searle argues that collective intentionality does not boil down to individual intentionality plus relevant cognitive relations – especially mutual beliefs – among individuals (cf 2002b; over the years, though, Searle would drop the irreducibility thesis for some forms of collective intentionality such as collective acceptance or recognition, cf 2010: 58). At the same time, though, Searle intends to avoid the ontological commitment toward group minds as bearers or owners of collective intentionality. How can both claims be squared?

Searle's suggestion is to locate collectivity in the mode or kind of the experience. For instance, the mental state of collectively intending to φ (where "φ" points to a kind of action like painting an house, playing chess or writing a paper) is such neither because of a particular *subject* of the intention nor because of a particular *content* of the intention. Rather, this mental state is such because it is a particular *kind* of intention: it is not a simple intending; it is *we* intending (or a cooperative intending; in Searle's own words: a collective-goal-by-means-of-singular-contribution-intending). Two consequences can be drawn from this idea: first, all forms of intentionality (collective intentionality included) are always of an *individual* mind: it is an individual who is *we*-intending to φ. Second, the individual can be fundamentally mistaken about his or her own mental states: e.g., if my *we*-intention brings me to act, but this *we*-intention has no counterpart (i.e., there is no other individual whose we-intention mashes with mine), then I am thinking to collectively φing, though in reality I am doing something completely different (2002b: 97f.).

All intentionality, and *a potiori* collective intentionality, is made possible by two elements: the Background and the Network. Although Searle himself admits never having been able to make a satisfactory and precise distinction between the Background and the Network (Searle, 2012: 19), these two notions can be roughly characterized as follows. The Network is intended as a network of interconnected mental states, which enable other mental states. In order to have a single perception, say, the perception of a car, one needs to have an entire series of beliefs: one needs to believe that cars are means of transportation, that generally they are driven by human beings on the earth's surface, etc. In contrast, the Background is a complex of 'capacities,' (abilities, dispositions, tendencies, and causal structures generally, 1995: 129) which is anchored in our neurophysiology. For instance, the *capacity* of speaking and understanding English or the *capacity* of playing baseball are a matter of Background (1995: 141ff.): in the beginning, the layperson first needs to think about every single step to be carried out for the practice to be successful. However, once the game is learned, i.e., once the person has acquired the corresponding capacity, the entire process goes smoothly and automatically: it is caused and guided by the Background.

Philosophy of Society

In recent contributions (most notably, in two books published in 1995 and 2010), Searle mainly focuses on a "new branch of philosophy that might be called 'The Philosophy of Society'" (Searle, 2010: 5). Such a new discipline, which broadly covers what nowadays also falls under the name 'social ontology,' has the ambition to provide an account of social reality by identifying the main principles that regulate its construction. Social reality does not exist independently of social beings and, hence, it can be said to be ontologically subjective. However, it is possible to make epistemically objective judgments about it. For instance, the fact that a given piece of paper is a €100 bill would not exist without social beings and their institutions, but the judgment that 'this piece of paper is a €100 bill' is not an arbitrary (or epistemically subjective) judgment.

Searle argues that there are three logical building blocks that are required for the construction of social reality (cf Searle, 2008d: 31). Two of them have been previously mentioned: constitutive rules (rules of the forms 'X counts as Y in context C,' that enable given practices) and collective intentionality (the power of the mind to do things together). Below, we will see how these elements enter into the picture of Searle's social ontology. The third element is the notion of function and that of assignment or imposition of function.

Searle stresses that functions are extrinsic or observer-relative features: strictly speaking, "nature knows nothing of functions" (1995: 14). To say that a given object has a given function (e.g., that the function of the heart is to pump blood) is to 'locate' the object into some system of values and to evaluate the object and its properties with respect to these values. If, for example, we consider a patient's heart in connection to the value of the patient's life, the heart is malfunctioning if it endangers the life of the patient, but "[i]f we thought the most important value in the world was to glorify God by making thumping noises, then the function of the heart would be to make a thumping noise, and the noisier heart would be the better heart" (1995: 15).

Some functions are nonagentive; others are agentive (for this taxonomy, cf 1995: 121). Generally, agentive functions indicate uses: the properties of being a screwdriver or being a paperweight indicate how a given object can be used. Nonagentive functions occur naturally (in the sense that one assigns the notion of purpose to a natural and causal process, e.g., the heart's function is to pump blood). Among the agentive functions, a further distinction can be drawn between causal and status functions. Causal functions are performed "*solely in virtue of causal and other brute features of the phenomena*" (1995: 123f.), e.g., the necessary conditions that an object has to fulfill in order to satisfy the function of being a screwdriver are its physical features (having a certain form, a certain robustness, etc.). By contrast, status functions ontologically do not rely on the physical features of their objects: they are "*performed only by way of collective acceptance*" (1995: 124). For instance, in order for something to be money, it is fairly irrelevant what features the physical object has. Money can be pieces of paper, cigarettes, squirrel pelts, etc., as long as it is collectively accepted that these objects are money. An object together with its status function is called an 'institutional fact' (e.g., this piece of paper being a $10 bill, Obama being the President of the United States, my laptop being property of the University of Copenhagen, etc.)

As we have seen, to exist or to obtain, institutional facts have to be collectively accepted and collective acceptance is a form of collective intentionality. Searle holds that all facts involving collective intentionality are social facts (1995: 37). For instance, when a group of hyenas is after a prey, their collective hunting is a social fact. So, what makes institutional facts so peculiar within the class of social facts? This question has different answers, but there are two features, which Searle seems to particularly stress in his work: the linguistic nature of institutional facts and the deontology they are able to generate.

In a broader sense, institutional facts are linguistic because they cannot exist without a linguistic system. Were there no creatures with a linguistic apparatus and linguistic competence, there would also be no institutional facts. "Consider the case of points scored in a game such as football. We say 'a touchdown counts six points.' Now, that is not a thought that anyone could have without linguistic symbols. But […] why? […] The answer, to put it simply, is that if you take away all the symbolic devices for representing points, there is nothing else there" (1995: 66). This goes not only for the game of football, but for all institutional facts – in fact, one should remember that these facts do not depend, for their existence, on the natural features of their constituting objects, and this suggests that a perfect description of the physical properties of the objects at issue would not say anything about their status functions.

This leads to a further consideration: institutional facts are linguistic – now in a narrower sense – because their very structure is linguistic. In a first version of his theory (1995), Searle argued that all institutional facts have the form 'X counts as Y in context C,' where X is a given object, generally a nonsocial one, and Y the term for the status function (note, however, that since the formula can be iterated, X could also point at a social object, as instance consider the case where a University's professor counts as the Dean of her Department). In his more recent contribution (2010), Searle – still holding that the formula captures a fundamental feature of institutional facts – restricts its validity only to a subclass of institutional facts. In addition to 'X counts as Y in context C'-structured facts, he introduces the idea of 'free-standing Y terms' to accommodate those facts where apparently no X figures in their structure (think of blindfolded chess, electronic money, or corporations as examples of such facts, 2010: 20f.). The introduction of these sorts of facts asks for further consideration: Searle needs to say something more about facts whose existence is not anchored, as it were, to the existence of a material object (of an X) and which, hence, seem to be made out of thin air.

To provide an account for these facts, Searle recurs to the notion of declaration that he previously developed in his speech-act theoretic work. Declarations were defined as speech acts with a double DoF: they create a reality by means of representing it as existing. Analogously, one can create institutional facts by performing collective declarations of the form:

> We make it the case by Declaration that the Y status function exists in context C.

Searle conceives of such a declarative mechanism as underlying the creation of *all* institutional facts. Accordingly, in the case of Ys assigned to Xs, we have a declaration of this kind:

> We make it the case by Declaration that X has the status Y […] in C.
> Searle, 2010: 99

Declarations hence acquire a central role in Searle's social ontology: "God can create light by saying 'Let there be light!' Well, we cannot create light but we have a similar remarkable capacity. We can create boundaries, kings, and corporations by saying something equivalent to 'Let this be a boundary!' 'Let the oldest son be the king!' 'Let there be a corporation!'" (Searle, 2010: 100).

A second feature that characterizes institutional facts is the deontology they generate. Searle's idea is that, whatever the metaphysical intricacies of boundaries, kings and corporations are, all these entities have a set of deontic powers attached to them, including rights, obligations, entitlements, etc. Boundaries should not be crossed without authorization, kings can declare war, corporations can enter contracts, etc. Indeed, Searle claims that *all* status functions (and, *a potiori*, *all* institutional facts, these being facts constituted by status functions) imply deontic powers (Searle, 2010: 23). That is, the status 'function itself' is to be "spelled out as a set of deontic powers" (Searle, 2010: 99) or, even more clearly, "[…] the [status] function is defined in terms of [deontic] power […]" (Searle, 1997: 451). It is hence remarkable that the complex metaphysics of such intentionality-dependent, institutional facts could also be considered as eventually boiling down to "[…] deontic phenomena" which "are not reducible to something more primitive and simple" (Searle, 1995: 70).

These deontologies provide desire-independent reasons for actions. If a person is thirsty and sees a glass of beer, the desire to quench his or her thirst might drive that person to reach for the glass and drink. Still, if someone else has ordered this glass of beer, such institutional fact has attached a deontology (an obligation) providing the desire-independent reason for

action, not to drink from others' glasses (for an example along the same lines, cf 2001: 27). According to Searle (2001), the distinction between desire-dependent and desire-independent reasons for action is essential in order to give an adequate account of (especially) practical rationality and to amend some of the weak points of what he calls the Classical Model of Rationality (a sort of more or less explicit and paradigmatic description of rationality in terms of payoff maximization, cf 2001: 5ff.).

The considerations on desire-independent reasons for action have to be understood in light of a main purpose that seems to have steered large parts of Searle's philosophical work: the attempt to make sense of concepts like language, mind, institutions, freedom of will, etc., in a world he describes as entirely composed of atoms and molecules.

Reception

During the last 50 years, John Searle has developed influential views on fundamental research areas within and beyond philosophy. Many of his publications are by now considered classics in the field of philosophy, and his ideas can be said to have profoundly impacted the scientific understanding of language, mind, and the social world.

His many works devoted to speech act theory have substantially contributed to the development of pragmatics and to its establishment as standing area of research within the philosophy of language. His book on *Intentionality* was one of the main factors that informed the process that put the mind and other mentalistic notions (like intentionality) back at the center of philosophical debates. Furthermore, the emergence of social ontology as a new field of research and the growing interest in intentionalistic explanations of collective phenomena can be seen as closely and directly linked to his more recent contributions about 'the philosophy of society.'

Acknowledgment

I would like to thank Frank Hindriks and Michael Schmitz for their helpful comments on previous draft of this article.

See also: Causation: Physical, Mental, and Social; Conventions and Norms: Philosophical Aspects; Intentionality and Rationality; Irrationality: Philosophical Aspects; Methodological Individualism: Philosophical Aspects; Objectivity: Philosophical Aspects; Rational Choice Explanation: Philosophical Aspects; Shared Belief: Philosophical Aspects; Social Ontology; Social Science, The Idea of.

Bibliography

Anscombe, G.E.M., 1957. Intention. Blackwell, Oxford.
Austin, J.L., 1953. How to talk, now. In: Urmson, J.O., Warnock, G.J. (Eds.), 1961. John L. Austin, Philosophical Papers. Oxford University Press, Oxford, pp. 134–153.
Austin, J.L., 1962. In: Urmson, J.O. (Ed.), How to Do Things with Words, the William James Lectures Delivered at Harvard in 1955. Clarendon, Oxford.
Brentano, F., 1874. Psychologie vom empirischen Standpunkt. Duncker & Humblot, Leipzig. Second ed., 1924, with introduction and notes by Kraus, O., 2 vols., Meiner, Lepzig. English translated ed. by McAlister, L.L., 1995. Psychology from an Empirical Standpoint, Routledge, London.
Dennett, D.C., 1978. Brainstorms: Philosophical Essays on Mind and Psychology. MIT Press, Cambridge.
Grice, H.P., 1957. Meaning. The Philosophical Review 66 (3), 377–388.
Reinach, A., 1913. Die apriorischen Grundlagen des bürgerlichen Rechtes. In: Schuhmann, K., Smith, B., Werke, Sämtliche (Eds.), Textkritische Ausgabe, 2 vols. Philosophia Verlag, Munich, pp. 141–278. English translated by Crosby, J., 1983. The a priori Foundations of the Civil Law. Aletheia. An International Journal of Philosophy III, 2–142.
Searle, J.R., 1968. Austin on locutionary and illocutionary acts. The Philosophical Review 77 (4), 405–424.
Searle, J.R., 1969. Speech Acts. An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.
Searle, J.R., 1972. The Campus War. Penguin Books, Harmondsworth, Middlesex.
Searle, J.R., 1975/1979a. A taxonomy of illocutionary acts. In: Expression and Meaning. Studies in the Theory of Speech Acts. Cambridge University Press, Cambridge, pp. 1–29.
Searle, J.R., 1978/1979b. Literal meaning. In: Expression and Meaning. Studies in the Theory of Speech Acts. Cambridge University Press, Cambridge, pp. 117–136.
Searle, J.R., 1980a. Minds, brains, and programs. The Behavioral and Brain Sciences 3, 417–424.
Searle, J.R., 1980b. Intrinsic intentionality: reply to criticisms of minds, brains, and programs. The Behavioral and Brain Sciences 3, 450–456.
Searle, J.R., 1983. Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press, Cambridge.
Searle, J.R., 1992. The Rediscovery of the Mind. MIT Press, Cambridge.
Searle, J.R., 1995. The Construction of Social Reality. The Free Press, New York.
Searle, J.R., 1997. Responses to critics of the construction of social reality. Philosophy and Phenomenological Research 57 (2), 449–458.
Searle, J.R., 1998. Mind, Language and Society. Basic Books, New York.
Searle, J.R., 2001. Rationality in Action. MIT Press, Cambridge.
Searle, J.R., 2000/2002a. Consciousness. In: Consciousness and Language. Cambridge University Press, Cambridge, pp. 36–60.
Searle, J.R., 1990/2002b. Collective intentions and actions. In: Consciousness and Language. Cambridge University Press, Cambridge, pp. 90–105.
Searle, J.R., 2002/2008a. Twenty-one years in the Chinese room. In: Consciousness and Language. Cambridge University Press, Cambridge, pp. 67–85.
Searle, J.R., 2005/2008b. The phenomenological illusion. In: Consciousness and Language. Cambridge University Press, Cambridge, pp. 107–136.
Searle, J.R., 2002/2008c. Why I am not a property dualist. In: Philosophy in a New Century. Cambridge University Press, Cambridge, pp. 152–160.
Searle, J.R., 2006/2008d. Social ontology: some basic principles. In: Philosophy in a New Century. Cambridge University Press, Cambridge, pp. 26–52.
Searle, J.R., 2010. Making the social world. In: The Structure of Human Civilization. Oxford University Press, Oxford, p. 208.
Searle, J.R., 2012. Preface. In: Salice, A. (Ed.), Intentionality. Historical and Systematic Perspectives. Philosophia, Munich, pp. 29–58.

Second Demographic Transition

James M Raymo, University of Wisconsin-Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.J. van de Kaa, volume 5, pp. 3486–3488, © 2001, Elsevier Ltd.

Abstract

The Second Demographic Transition (SDT) is a term used to describe dynamic interrelationships between fertility rates, a constellation of innovative demographic behaviors, and changing values in countries characterized by sustained below-replacement fertility. Key demographic changes accompanying the emergence of very low fertility include delayed marriage and childbearing, and substantial increases in cohabitation, nonmarital fertility, childlessness, maternal employment, and divorce. The core shift in values associated with the SDT involves movement away from viewing marriage and childbearing as either unquestioned or obligatory toward a life orientation in which family formation is something purposively chosen for its role in self-actualization.

The Second Demographic Transition (SDT) is a term first used in the mid-1980s by Ron Lesthaeghe and Dirk van de Kaa to describe the emergence of sustained below-replacement fertility and an accompanying 'package' of new behaviors and attitudes in several European countries (Lesthaeghe and Van de Kaa, 1986; Van de Kaa, 1987). Their reference to a 'second' transition was motivated primarily by the fact that, in contrast to a central assumption of classic demographic transition theory, period fertility rates had fallen well below replacement levels in several countries. At the core of the SDT framework is an emphasis on dynamic interrelationships between fertility rates, a constellation of 'innovative' demographic behaviors, and changes in values and life orientation. In addition to below-replacement fertility, the demographic changes commonly described as part of the SDT include delayed marriage, delayed childbearing, childlessness, increases in the proportion who never marry, and substantial increases in nonmarital cohabitation, nonmarital fertility (including fertility within cohabiting unions), maternal employment, and divorce (Lesthaeghe, 1995; Lesthaeghe and Moors, 2000; Van de Kaa, 1987).

The range of potential explanations offered for the demographic changes associated with the SDT is broad and includes shifting attitudes and values, changes in contraceptive technology, and changes in the economic opportunities and roles of men and women driven by educational expansion and economic change (e.g., Bumpass, 1990; Lesthaeghe, 1995, 1998; Lesthaeghe and Moors, 2000; McDonald, 2000). The role of shifting attitudes and values, in particular, has been central to the SDT framework from the beginning. In very general terms, the core shift in values or life orientation has been described as a move away from viewing marriage and childbearing as either an unquestioned part of life or a duty to family and society (as in the first demographic transition) toward a view that family formation is something purposively chosen for its role in self-actualization. In the SDT framework, the process by which this emerging valuation of satisfying 'higher order needs' and new family behaviors evolve is viewed as cyclical. Initially, the decline in fertility and the emergence of new family behaviors is driven by changes in values including declining familistic altruism, increasing individuation, secularization, antiauthoritarianism, and rising consumption aspirations. Innovators at the forefront of these shifting values self-select into new family behaviors including postponement or avoidance of marriage and parenthood and these experiences reinforce the underlying attitudes and values while the new patterns of demographic behavior become more widely accepted and diffuse throughout population (Lesthaeghe and Surkyn, 1988; Surkyn and Lesthaeghe, 2004).

The initial formulation of the SDT framework sought to explain observed demographic changes in a small number of countries in North and Western Europe. As the framework developed over time, increased emphasis has been placed on understanding the SDT as a broader phenomenon and on documenting its spread to other regions. Related empirical research has focused on (1) developing rankings or typologies to classify countries according to their stage of progression through the SDT (Lesthaeghe, 1995; Lesthaeghe and Moors, 2000; Sobotka, 2008), (2) describing variation across low-fertility countries in the pace and nature of changes associated with the SDT (Sobotka, 2008), and (3) evaluating whether specific countries should be considered SDT countries (Atoh et al., 2004; Lesthaeghe and Neidert, 2006). This work has shown that the levels of key demographic indicators and attitudes associated with the SDT tend to cluster in different areas, with the Scandinavian countries most 'advanced' in terms of SDT characteristics and Southern European countries less far along. It is also clear that the constellation of demographic behaviors associated with the SDT varies, with low fertility and the full package of new family behaviors emerging more or less in tandem in the original SDT countries, in piecemeal fashion in other countries (e.g., Southern Europe and East Asia), or following rapid fertility collapse after major political and economic change in Central and Eastern Europe.

Just as regional variation in the transition from high fertility to low fertility provided important insights into the roles played by language, social capital, networks, and diffusion processes, research on the transition from low to very low fertility has made use of variation in the pace and nature of

associated family changes to better understand the influence of social, historical, economic, and political context. For example, historical and cultural differences in the strength of family ties and patterns of departure from the parental home have been linked to regional differences in the spread of nonmarital cohabiting unions, especially the low prevalence in Italy and other Southern European countries (Reher, 1998). Similarly, the very low prevalence of nonmarital childbearing in low-fertility East Asian countries is thought to reflect the continued strength of social stigma and legal discrimination against 'illegitimate' children (Rindfuss et al., 2004). Social and economic factors that combine to make work–family balance more difficult for women in some societies than in others have been linked to regional differences in marriage, fertility, and married women's labor force attachment (McDonald, 2000; Rindfuss et al., 2003). Within the US, substantial regional variation in fertility and demographic behaviors associated with the SDT has been linked to regional differences in religion and attitudes (Lesthaeghe and Neidert, 2006).

Research during the past decade has increasingly emphasized socioeconomic differences in family behaviors associated with the SDT and the implications of those differences for inequality and the reproduction of disadvantage. In contrast to earlier research emphasizing the role of highly educated demographic and attitudinal innovators, this new work has focused on the growing concentration of demographic behaviors with potentially negative implications for subsequent well-being at the lower end of the socioeconomic spectrum. The impetus for this work emerged primarily from the US where it has become increasingly clear that some behaviors associated with the SDT such as nonmarital childbearing, nonmarital unions, and union dissolution are increasingly concentrated among those with the most limited socioeconomic resources (McLanahan and Percheski, 2008). In an influential paper published in 2004, Sara McLanahan portrayed this pattern of family bifurcation as a key feature of the SDT and highlighted its role in rising inequality, especially in children's resources. Building upon this work and accumulated evidence from a range of countries including those in Central and Eastern Europe, Sobotka (2008) suggests that there may be two pathways to the SDT. One is the original depiction in which highly educated innovators adopt new behaviors and associated attitudes/values which then diffuse throughout the population (as with fertility regulation in the classic demographic transition framework). In the second pathway, economic uncertainty or other structural conditions that make conventional family patterns difficult to achieve result in emergence of cohabitation, nonmarital childbearing, and other SDT behaviors among the more disadvantaged segments of the population before spreading more broadly. Careful empirical evaluation of this argument is a relatively new area of research and findings to date are mixed, with some finding consistent patterns in nonmarital childbearing in Europe (Perelli-Harris et al., 2010) while others find little evidence that educational differences in divorce are changing in a similar way Härkönen and Dronkers (2006).

The SDT framework is widely referenced but consensus regarding its interpretation is not universal. Some have questioned the appropriateness of the term SDT, arguing that recent trends, especially sustained below-replacement fertility, should be seen as an extension in the longer demographic transition process (Cliquet, 1991) or that the changes observed do not merit the term 'transition' (Coleman, 2004). Others have raised questions about the permanence of changes, suggesting that below-replacement fertility may be a temporary phenomenon generated by widespread postponement of family formation. The idea that period fertility rates will rebound when those currently postponing family formation begin to have children has inspired important efforts to disentangle the impact on the Total Fertility Rate (TFR) of changes in the timing of fertility – tempo effects and reductions in the number of children women will ever have – quantum effects (Bongaarts and Feeney, 2000). The return to near-replacement fertility in some SDT countries like France and Sweden appear consistent with a view of the SDT as a transitory phenomenon and the absence of some key features of the SDT in very low-fertility countries in Southern Europe and East Asia highlights the limited generalizability of the SDT as it was originally depicted. Despite these questions and critiques, the term SDT is now widely used by demographers as a convenient and informative means of summarizing recent demographic trends in societies characterized by very low fertility. In Sobotka's (2008: p. 213) words, "the second demographic transition seems to be particularly useful as an umbrella concept that encompasses a broad range of interrelated changes in sexuality, family, and partnership behaviours and attitudes, as well as a massive postponement of parenthood."

What is the future of the SDT? Evidence of fertility recuperation at older ages in many countries (Lesthaeghe, 2010) suggests that below-replacement fertility, or at least lowest-low fertility, may be a temporary phenomenon. However, it does appear that many of the new family forms, especially the separation of marriage and childbearing, and accompanying changes in attitudes and values are here to stay (Sobotka, 2008). Moving forward, it seems likely that efforts to evaluate the idea of an SDT or to place countries on an SDT continuum will give way to efforts to evaluate the broader and longer-term implications of the fundamental demographic changes that have come to be associated with the SDT. For example, it is clear that prolonged periods of below-replacement fertility will result in rapid population aging and population decline. Related policy responses, including pronatalist policies, work–family balance policies, and migration and guest worker policies are all of increasing interest to demographers. The same is true of the changed family landscape and new attitudes toward children and family life, including perhaps permanently altered preferences regarding the relationship between marriage and childbearing, the timing of family formation, and desired family size. Finally, because many of the family behaviors associated with the SDT have important implications for economic well-being and the intergenerational transmission of dis/advantage, efforts to understand how these changes impact poverty, stratification, and social mobility, and how these relationships differ across countries characterized by different social, economic, and policy contexts, will likely be a core focus of demographers in the years to come.

See also: Children, Value of; Cohabitation: United States; Family Size Preferences; Fertility Change: Quantum and Tempo; Fertility Trends in the Formerly Socialist Countries of Europe; Fertility of Single and Cohabiting Women.

Bibliography

Atoh, Makoto, Kandiah, Vasantha, Ivanov, Serguey, 2004. The second demographic transition in Asia? comparative analysis of the low fertility situation in east and south-east Asian countries. The Japanese Journal of Population 2, 42–75.

Bongaarts, John, Feeney, Griffith, 2000. On the quantum and tempo of fertility: reply. Population and Development Review 26, 560–564.

Bumpass, Larry L., 1990. What's happening to the family? Interactions between demographic and institutional change. Demography 27, 483–498.

Cliquet, Robert L., 1991. The Second Demographic Transition: Fact or Fiction? (No. 23). Council of Europe.

Coleman, David, 2004. Why we don't have to believe without doubting in the "Second Demographic Transition"—some agnostic comments. Vienna Yearbook of Population Research 2, 11–24.

Härkönen, Juho, Dronkers, Jaap, 2006. Stability and change in the educational gradient of divorce. A comparison of seventeen countries. European Sociological Review 22, 501–517.

Lesthaeghe, Ron, 1995. The second demographic transition in Western countries: an interpretation. In: Mason, K.O., Jensen, A.-M. (Eds.), Gender and Family Change in Industrialized Countries. Oxford University Press, New York, pp. 17–62.

Lesthaeghe, Ron, 1998. On theory development: applications to the study of family formation. Population and Development Review 24, 1–14.

Lesthaeghe, Ron, 2010. The unfolding story of the second demographic transition. Population and Development Review 36, 211–251.

Lesthaeghe, Ron, Moors, Guy, 2000. Recent trends in fertility and household formation in the industrialized world. Review of Population and Social Policy 9, 121–170.

Lesthaeghe, Ron J., Neidert, Lisa, 2006. The second demographic transition in the United States: exception or textbook example? Population and Development Review 32, 669–698.

Lesthaeghe, Ron, Surkyn, Johan, 1988. Cultural dynamics and economic theories of fertility change. Population and Development Review 14, 1–45.

Lesthaeghe, Ron, Van de Kaa, Dirk J., 1986. Twee demografische transities? (Two demographic transitions?) In: Van de Kaa, D.J., Lesthaeghe, R.L. (Eds.), Bevolking: Krimp en Groei. Van Loghum Slaterus, Deventer, pp. 9–25.

McDonald, Peter, 2000. Gender equity in theories of fertility transition. Population and Development Review 26, 427–439.

McLanahan, Sara, 2004. Diverging destinies: how children are faring under the second demographic transition. Demography 41, 607–627.

McLanahan, Sara, Percheski, Christine, 2008. Family structure and the reproduction of inequalities. Annual Review of Sociology 34, 257–276.

Perelli-Harris, Brienna, Sigle-Rushton, Wendy, Kreyenfeld, Michaela, Lappegård, Trude, Keizer, Renske, Berghammer, Caroline, 2010. The educational gradient of childbearing within cohabitation in Europe. Population and Development Review 36 (4), 775–801.

Reher, David S., 1998. Family ties in Western Europe: persistent contrasts. Population and Development Review 24, 203–234.

Rindfuss, Ronald R., Choe, Minja Kim, Bumpass, Larry L., Tsuya, Noriko O., 2004. Social networks and family change in Japan. American Sociological Review 69 (6), 838–861.

Rindfuss, Ronald R., Guzzo, Karen B., Morgan, S. Philip, 2003. The changing institutional context of low fertility. Population Research and Policy Review 22, 411–438.

Sobotka, Tomas, 2008. Overview chapter 6: the diverse faces of the second demographic transition in Europe. Demographic Research 19, 171–224.

Surkyn, Johan, Lesthaeghe, Ron, 2004. Value orientations and the second demographic transition (SDT) in Northern, Western and Southern Europe: an update. Demographic Research. Special Collection 3, 45–86.

Van de Kaa, Dirk J., 1987. Europe's second demographic transition. Population Bulletin 42 (1).

Second Language Acquisition

Gerrit Jan Kootstra, Center for Language Science, Pennsylvania State University, University Park, PA, USA
Ton Dijkstra, Donders Institute for Brain, Cognition, and Behaviour; and Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands
Marianne Starren, Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by W. Klein, volume 20, pp. 13768–13771, © 2001, Elsevier Ltd.

Abstract

This contribution provides an overview of second language acquisition research from a psycholinguistic perspective. The overview is mainly focused on cross-language interactions during word and sentence production and comprehension at different points in the second-language acquisition process. The main conclusion is that language learning is interactive and dynamic. Because of this interactivity and dynamicity, the process and outcome of second-language acquisition can be influenced by many different factors, like the similarity between the first and second language, the learning context, learner motivations, the age at which a second language is learnt, and the learner's language proficiency and background.

Introduction and Historical Context

The world counts close to 7000 known living languages spread over more than 200 countries (Lewis, 2009). It is therefore not surprising that the number of people in the world speaking multiple languages easily outperforms that of people speaking only one language (e.g., Gardner-Chloros, 2009; Grosjean, 2008), and that roughly two-thirds of all children in the world grow up in a bilingual environment (e.g., Bhatia, 2004; Crystal, 2003). In addition, because of the growing influence of international mass media, global communication through the internet, and increased mobility across countries, it is important for many people to learn to speak other languages in addition to their first language.

There are many types of second-language learners. In addition to the typical school pupil who is taught a second language for educational purposes, there are migrants, children of migrants, expats, children who are enrolled in bilingual education programs, international students, adopted children, and people who speak a dialect in addition to the national majority language. It can be concluded from this that the context in which a second language is learnt, the motivations to learn a second language, the age at which a second language is learnt, as well as the social and linguistic background of the language learner can differ widely. These are all factors that can influence the second language acquisition (SLA) process (see, e.g., Doughty and Long, 2003; Gass and Mackey, 2012; Robinson, 2013, for overviews).

It is difficult to say exactly when the study of SLA began, but an important work that fueled much early thinking was *Linguistics across Cultures* by Robert Lado (1957). Building on behaviorist theories of language learning, Lado proposed that difficulty of second-language learning is based on the distance between the native language and the target language. Elements that are similar between languages will be relatively easy to learn whereas elements that are dissimilar will be relatively difficult to learn. Later, Lado's views were questioned by researchers like Corder (1967) and Dulay and Burt (1974), who argued that the type of errors second-language learners make or the developmental sequences of SLA cannot wholly be attributed to the linguistic distance between the native language and the target language. Over the last decades, second-language research has evolved into a research field that has built on many different research disciplines, ranging from general linguistics to cognitive psychology and from sociolinguistics to educational science, in which SLA is often characterized as a complex and dynamic process that is influenced by a multitude of factors (e.g., De Bot et al., 2005; Doughty and Long, 2003; Ellis, 1994; Gass and Mackey, 2012; Robinson, 2013; Robinson and Ellis, 2008).

Much research on SLA is closely tied to research on bilingualism. The term 'bilingual,' like 'multilingual,' is often loosely used as referring to a speaker of two or more languages (see Appel and Muysken, 1987; Bhatia and Ritchie, 2013; Kroll and de Groot, 2005; Li Wei, 2007. for overviews). A central question in the field of bilingual language processing is how and to what extent a bilingual's languages interact with each other during language use. An important instigator of this research was Uriel Weinreich, who studied the extent to which a bilingual's languages compete with each other and are stored in the brain as a function of language dominance, proficiency, and use (Weinreich, 1953). This role of language dominance and proficiency is still central in current research on the bilingual speaker, for instance on issues like cross-language interactions in language production and comprehension (e.g., Dijkstra, 2009; Kroll et al., 2012; Van Hell and Tanner, 2012), cognitive control of languages in balanced and unbalanced bilinguals (e.g., Christoffels et al., 2007; Green, 1998), and the consequences of bilingualism for nonlinguistic aspects of memory and executive functions (e.g., Bialystok et al., 2012).

Considering the thematic relatedness between bilingual language processing and SLA, it is becoming more and more difficult to distinguish research on bilingualism from research on SLA. Cross-language interactions during language production and comprehension, for instance, have not only been studied from a bilingual language processing perspective, but

also from an SLA perspective, in which it is often termed *language transfer* or *crosslinguistic influence* (e.g., Jarvis and Pavlenko, 2008; Kellerman and Sharwood Smith, 1986; MacWhinney, 2008; Odlin, 1989). Likewise, more and more studies on bilingual processing focus on proficiency and learning processes and thus make the link with SLA research (see, e.g., Kroll and de Groot, 2005; de Groot, 2011, for overviews). This close connection between SLA and bilingual processing will also be the starting point of the present contribution, in which we present a perspective on SLA inspired by studies on bilingual language processing.

Core Notions in SLA and Processing

Two fundamental notions with respect to language learning are that it is both *interactive* and *dynamic*. Language learning is *interactive*, because it is affected by what the learners already know about their native language and other languages. Methods for teaching a foreign language (FL) can only be optimized if the underlying mechanisms of such interactions are clear. If learners of an FL depend heavily on the rules and representations they already have in their native language, learning will be different when it takes place at a later age (when L1 is firmly established) relative to early in life (when L1 is still under development). Psycholinguistic research in the last two decades has provided evidence that most sorts of language information, for instance, with respect to pronunciation (phonetics/phonology), words (lexicology), word formation and grammar (morphosyntax), and meaning (semantics) in bilinguals are not stored in long-term memory stores that are separate for each language but are integrated across languages (see, e.g., De Groot, 2011, for a balanced review of the data). Furthermore, during processing, information from these stores appears to be accessed in a language-independent way. As a consequence, information of different languages interacts at all linguistic levels, for instance during the processing of words and sentences. Research looking into these interactions has considered situations in which language processing converges (i.e., in the case of crosslinguistic similarities between languages) and in which it diverges (in case of crosslinguistic differences between languages).

The second fundamental notion is that language learning is *dynamic*. The language system goes through profound changes during the lifespan and it is quite sensitive to prolonged and momentary changes in the environment (Grosjean and Li, 2013, p. 10). To describe the waxing and waning of whole languages and their various representations, the *activation metaphor* is well suited. Words and grammar of a language can be more or less *active* at certain moments in time. Many researchers assume that words are characterized by a *resting-level activation* when they are not used and an *activation rate* when they are, which is dependent on how often they are used. Thus, when words are encountered during reading, listening, or speaking, they become active and available faster or slower, again depending on their frequency of usage. Because words in a second language have typically been learned later in life (i.e., they have a later age of acquisition) and are used less than words of a first language, their subjective frequency is lower, resulting in more effortful, slower, and less accurate processing (cf Grosjean and Li, 2013, p. 224). It is further assumed that when multiple languages are present, there may be simultaneous activation or *co-activation* of elements (words, sentences) of different languages, due to lexical or structural (syntactic) overlap. In the following section, we will consider the interactive activations of languages at word and sentence levels in more detail. This will be followed by a consideration of the dynamic aspects of language learning across the life span.

The Interaction of Languages

Interactions at the Word Level

At the *word* level, many studies have investigated how words are processed that have form overlap between languages. Words that are similar in their orthographic or phonological form between languages, but that differ in meaning at the same time, are called 'false friends' (also: interlingual homographs or interlingual homophones). Examples are words like COIN (English) – COIN (French, meaning 'corner') and SLIM (English) – SLIM (Dutch, meaning 'cunning'). Words that are similar in form and in meaning are called 'cognates' (such words usually have a common etymology or are borrowings in one of the languages). So-called 'identical cognates' are translation equivalents that are orthographically identical and often phonologically very similar across languages, like FILM (English) – FILM (Dutch, French, Polish, and many other languages) (Peeters et al., 2013). 'Nonidentical cognates' are translation equivalents that are crosslinguistically less similar in their orthography and phonology, like TOMATO (English) – TOMAAT (Dutch) – TOMATE (French) – ΔΟΜΑΤΑ (Greek, pronounced as /doma:ta:/). Studies with false friends have shown that their processing is typically slowed down relative to matched one-language control words (Dijkstra, 2005) and that the degree of interference depends on the relative frequency of usage of the false friends in the associated languages. The strongest interference effects have been observed for low-frequency L2 words with high-frequency L1 false friends (e.g., Dijkstra et al., 1998; Smits et al., 2006). Studies with cognates have reported that these are usually processed faster than comparable control words ('cognate facilitation') and that the degree of facilitation depends on the cognate's degree of crosslinguistic similarity. This holds for words in isolation (Dijkstra et al., 2010) and in sentences (van Assche et al., 2011). Like in false friends, the relative frequency of usage of the cognate in a language also plays a role (Peeters et al., 2013).

The difference in processing between false friends and cognates has been explained as a consequence of their representational differences. It is assumed that for both false friends and cognates, form representations in both languages are co-activated (e.g., SLIM activates both an English and a Dutch form representation). In false friends, the competition between the co-activated form representations results in slower recognition. In cognates, two form representations are also assumed to become co-activated. However, cognates are assumed to have a meaning representation that is at least partially shared between languages. This representation is co-activated by both form representations and benefits cognate

processing relative to control words (Dijkstra et al., 2010; Voga and Grainger, 2007).

The cross-language interactions described above turn out to be influenced by *relative language proficiency*. Van Hell and Dijkstra (2002) tested the size of the 'cognate facilitation' effect in trilinguals who differed in their proficiency of the third language. When Dutch (L1) university students with English as a second language (L2) and French as a third language (L3) performed a Dutch lexical decision task, a cognate facilitation effect was found for Dutch–English cognates, but not for Dutch–French cognates. Because the trilingual participants all were native speakers of Dutch (L1) with a higher proficiency in English than in French, this was to be expected. In a second experiment, trilinguals with a much higher level of French proficiency were recruited (i.e., Dutch university students of French). For this participant group, cognate facilitation effects in Dutch were obtained for both Dutch–English (L1–L2) and Dutch–French (L1–L3) cognates. Therefore, relative proficiency in another language (L2 or L3) affects the size of the cognate facilitation effect in L1, at least for this population of adult participants.

With respect to L2 word learning, cognates are prominent in many language courses, because they are easier to acquire than control words: their form and meaning overlap with L1 representations makes them easier to store and remember. When pairs of European languages are compared with respect to their lexicons, large numbers of cognates are found and much smaller numbers of false friends (Schepens et al., 2012, 2013). Thus, in spite of the misleading form overlap in the case of false friends, the relative prominence of cognate words in the initial phases of a language learning curriculum may indeed be beneficial to the learner.

A model that been proposed to understand the interactions and dynamics of language learning and language processing is the Revised Hierarchical Model (RHM; Kroll and Stewart, 1994). The RHM assumes that information about words is stored both in a lexical memory (for word forms) and in a conceptual memory (for word meanings). The conceptual memory is nonlinguistic in nature, implying that meanings of words can be shared by words from different languages. For instance, TOMATO (English) and TOMAAT (Dutch) map onto one and the same meaning. In a late L2 learner, strong connections between word forms of the native language L1 and their meanings have already been established. In the first phases of L2 learning, the L2 form representations will not be very strong, and the same holds for their connections to meaning. Because of these weak representations and connections, it might be quicker not to look up the L2 word's meaning in the lexicon directly (concept mediation), but indirectly via the stronger L1 representations and their strong connections to meaning (word association). Depending on the L2 teaching method, a learner may already have forged links between L2 and L1 form representations, for instance, between HORSE (L2) and PAARD (L1) in a Dutch learner of English (e.g., Comesaña et al., 2009). Thus, it might be faster to find the meaning of HORSE via the form representation of PAARD rather than via the form-meaning connection of HORSE itself. In later stages of learning, the L2 word form HORSE and its connection to meaning are strengthened by more frequent use. Accordingly, the direct route to meaning would become more prominent.

The RHM has inspired hundreds of studies over the last 15 years. In addition to findings of the predicted development from word association to concept mediation (e.g., Talamas et al., 1999), an increasing number of studies has found early concept mediation effects in L2 word learning (Altarriba and Mathis, 1997; Brenders et al., under revision; Comesaña et al., 2009; de Groot and Poot, 1997). To gain more insight into the relative contribution of concept mediation and word association in different stages of SLA, computational modeling of the RHM's assumptions would be useful (a first attempt to do so is Multilink, Dijkstra and Rekké, 2010).

Complementary to the RHM model, another model for language interactions at the word level that has recently been applied to L2 acquisition issues is the Bilingual Interactive Activation (Plus) Model (BIA+). The BIA+ model (Dijkstra and Van Heuven, 1998, 2002; Van Heuven et al., 1998) extends the Interactive Activation model for word recognition from McClelland and Rumelhart (1981, 1988). This symbolic (or localist) connectionist model proposes that word recognition entails the activation and de-activation of representations for letter features, letters, and words. During the recognition process, words that are similar to the input letter string compete for recognition (e.g., the input BIKE also leads to activation of LIKE, HIKE, and BAKE, so-called 'neighbors'). In addition, the task to be performed (e.g., name the word or decide whether a letter string is a word or not) codetermines the ultimate response time.

Another important determinant of the model's behavior is the lexicon that is incorporated, which can be monolingual or bilingual. Clearly, for bilingual learners, words from a second language will be included depending on their current L2 proficiency. More proficient learners will know more L2 words. In addition, the subjective frequency of usage of those words will become higher when the second language is used more often. This characterization of L2 learning makes it clear that at every stage of learning, the learners' lexicon can be considered as a combination of all L1 words with a particular frequency and a (typically smaller) set of L2 words with a particular frequency. Dijkstra et al. (2012) conducted simulations with the BIA+ model for different stages of learning, i.e., with a changing lexicon. These simulations showed that for Dutch as a native language and English as a second language, word recognition in Dutch was relatively unaffected by the increasingly stronger English, while the new English words were still recognized at reasonably normal response times.

Dijkstra et al. (2012) also report a series of simulations with a distributed connectionist model for word learning that extends an earlier model by Thomas (1997). This Revised Bilingual Single Network model assumes that learners build word-form representations for words in a new language that are linked to semantic representations during learning. Using a biologically plausible learning algorithm, called LEABRA ('Local, Error-driven and Associative, Biologically Realistic Algorithm'), it was shown that cognates were learned faster than words without a form-similar translation in another language, whereas false friends were learned as slow or somewhat slower than such control words. The simulations suggest that the inherent properties of cognates (form and

meaning overlap) and false friends (form overlap but no meaning overlap) do not only affect how these items are processed after solid learning, but that they are already different during the learning process itself.

Interactions at the Sentence Level

A number of studies at the *sentence* level indicate that, like their lexicons, the grammars of different languages are not stored independently but together. This has, for example, been investigated in sentence interpretation tasks, in which bilinguals and second-language learners are confronted with a simple sentence (e.g., 'the boy kissed the girl' or 'the man touched the stone'), where they have to indicate which of the sentence's arguments is 'doing the action.' Many studies have found that second-language learners are strongly influenced by L1 processing preferences while performing the sentence interpretation task in their L2 (e.g., Gass, 1987; Harrington, 1987; Liu et al., 1992; McDonald, 1987). However, with increasing proficiency in the L2, L2 learners have been found to gradually develop more target-language-like sentence interpretation strategies (e.g., McDonald and Heilenman, 1992). This does not mean, however, that cross-language interactions at the sentence level are absent in more proficient L2 learners. Indeed, in learners with a high proficiency in both languages, both the L1 and the L2 have been found to mutually influence both L1 and L2 sentence interpretation (e.g., Liu et al., 1992).

The sentence interpretation effects are often accounted for in terms of the Competition Model (e.g., MacWhinney and Bates, 1989; later also termed Unified Competition Model; e.g., MacWhinney, 2005, 2008; 2012). The basic tenet of this model is that languages consist of multiple mappings from linguistic form to communicative function (i.e., *cues*, such as word order, case marking, or subject–verb agreement) that interact and/or compete with each other to come to an interpretation of a sentence. These cues are not only assumed to interact with each other *within* a language, however, but also *between* languages, leading to cross-language interaction of language-specific form–function mappings. Similar to the earlier-discussed RHM, the competition model holds that these form–function mappings are stronger in the L1 than in the L2 in beginning L2 learners, such that L2 sentence processing is initially based on L1 processing strategies. In more experienced L2 learners, form–function mappings in the L2 sentence interpretation will have become stronger, making L2 sentence processing less parasitic on the L1 (and sometimes leading to mutual influences between the L1 and L2).

Cross-language interactions at the sentence level have also been observed in the form of *cross-language syntactic priming* (e.g., Bernolet et al., 2012; Loebell and Bock, 2003; Hartsuiker et al., 2004; Weber and Indefrey, 2009). Cross-language syntactic priming is the process in which a bilingual's syntactic processing of an utterance is influenced by the syntactic structure of a recently processed utterance *in a different language*. Hartsuiker et al. (2004) asked pairs of Spanish–English participants to describe cards depicting transitive events to each other in a dialogue game. One of the participants was a confederate, who produced utterances according to a prescribed format. It was found that the 'real' participants tended to use similar sentence constructions as the confederate, even though the confederate used Spanish and the real participants used English. Similar findings of cross-language syntactic priming have now been replicated across different languages, from L1 to L2 and from L2 to L1, in different syntactic structures and constituents, using different task varieties (see, e.g., Hartsuiker and Pickering, 2008, for a review).

It is interesting to note that, like in bilingual word-processing studies, cross-language interaction in sentence production is influenced by relative language proficiency. Schoonbaertet al. (2007) found, for example, that cross-language syntactic priming in Flemish learners of English occurred both from L1 to L2 and from L2 to L1, but that enhanced priming due to lexical repetition between prime and target (i.e., the lexical boost effect) was only present from L1 to L2; not from L2 to L1. Similarly, Kootstra and Doedens (submitted) found in Dutch learners of English that long-lasting effects of cross-language syntactic priming were only present from L1 to L2 and not from L2 to L1. They also found that L1 structural preferences associated with specific verbs (verb alternation bias) influenced L2 syntactic choices in picture descriptions, but L2 structural preferences influenced L1 syntactic choices to a lesser extent. Finally, Kootstraet al. (2012) found that lexical priming in the production of bilingual sentences was stronger in high-proficient bilinguals than in low-proficient bilinguals.

The cross-language syntactic priming studies are often accounted for in terms of a model in which words (lexical representations) and sentence structures (syntactic representations) are represented in a processing system that is shared for both languages (e.g., Hartsuiker et al., 2004). This shared system is exactly what makes cross-language syntactic priming possible: when a syntactic representation has just been used in Language A, it is still relatively strongly activated when a next sentence is produced in Language B, thus increasing the likelihood that this structure is used again. To account for the effects of relative language proficiency, a similar assumption is made as in the RHM and in the competition model, namely that the connections between concepts and linguistic forms are stronger in the (stronger) L1 than in the (weaker) L2. Therefore, the influence of L1 on L2 processing will be stronger than vice versa.

An interesting point with respect to SLA is that syntactic priming does not only provide a methodological tool to investigate cross-language interactions in bilingual sentence production, but may also serve as a mechanism underlying the acquisition of syntax. That is, priming of syntactic structure can be seen as reactivation and strengthening of a just used syntactic representation in memory; when this takes place frequently and consistently in the target language (e.g., the L2), priming could indeed result in the gradual acquisition of target-language form-to-meaning mappings (see, e.g., Chang et al., 2006). This potentially facilitating role of (within-language) syntactic priming for language acquisition has recently been applied and empirically confirmed in the context of L2 learning (e.g., McDonough, 2006; Trofimovich and McDonough, 2011; Shin and Christianson, 2012). Indeed, it has even been found that L2 syntactic priming can override negative effects of L1 transfer (Nitschke et al., 2010). Thus, L2 syntactic priming could serve as an important mechanism of L2 acquisition, in which L2

learners learn to associate L2 syntactic structures with their meanings on the basis of recent (and continuous) exposure to the L2. Based on this, the mechanism of priming is now also investigated as a potential technique for L2 teaching in educational settings (see Trofimovich and McDonough, 2011, for examples).

In all, the discussed studies on lexical and syntactic cross-language interactions reveal that both language comprehension and production in second-language learners and bilinguals is based on a processing system that is shared for both languages and prone to cross-language interactions. We have also shown that these cross-language effects are not only dependent on linguistic factors such as cognate status or language-specific form–function mappings, but also on learner characteristics, such as relative language proficiency. This shows that the nature of cross-language interactions in second-language learners is dynamic in itself. The dynamics of second-language learning are the topic of the next section.

The Dynamics of Language Learning across the Life Span

It is often mentioned that a child can learn a second language more quickly than an adult. Clearly, the brain of a child is still under development and relatively 'empty,' whereas that of an adult is more stabilized and filled with knowledge. There are many differences in the learning situations of both populations. Young L2 learners feel an urgent need to communicate, they are exposed more frequently to the L2, they are not afraid to make mistakes, they use concrete language about the here and now, and their memory/brain is plastic and not stowed with knowledge. However, older and later L2 learners may profit from other advantages: they already know about the world the language is referring to, they have well-developed cognitive abilities, reading abilities, learning skills and study skills, and a complete L1 language system (to which they can relate their new system). In this context, we should also consider the amount of time that someone can spend on L2 learning, whether that person already knows other languages, whether the new language is similar to the old one(s), whether the learner is a man or a woman, whether someone is talented for language or not, and what the motivation of the person is in acquiring the new language (e.g., for a living or for life).

A general observation is that the sensitivity to learn a second language becomes less when a person ages. However, it appears that talking about a strict 'critical period' for L2 learning is a bit too strong. Sensitivity does not decline catastrophically after puberty, but gradually, and learning another language at a later age is still possible.

Babies initially are sensitive to *phoneme distinctions* that exist in all different languages. Already after half a year their perception of nonnative phonemic contrasts declines and that of specific native contrasts increases (e.g., Werker et al., 1981; Kuhl, 2004). The older they get, the more difficult it becomes to learn to master the sounds of other languages. However, this development is not unitary but depends on specific acoustic/articulatory sound characteristics involved (Rivera-Gaxiola et al., 2005). Nevertheless, with motivation and effort, a very high level of pronunciation can still be reached also in later stages of life (Bongaerts et al., 1997). It appears to be somewhat easier to learn the *grammar* of an FL earlier in life, whereas the acquisition of vocabulary is less sensitive to age of acquisition (Ullman, 2001). The explanation of Ullman for this is in terms of two types of memory. The first type of memory stores facts and events (declarative memory); this is important for our 'dictionary.' The second type of memory is involved in learning and actions (procedural memory) and is important for the acquisition of sequences. Grammar is concerned with word order and the learning of grammar is therefore subject to the second type of memory. This second type of memory would be sensitive to hormonal changes, and it would be somewhat more effective before rather than after puberty. A consequence of this view is that one learns a language somewhat differently at a later age relative to an earlier age: somewhat less 'intuitively' and with a more conscious desire to derive rules. Seen from this developmental perspective, it can be understood why (late) L2 learners are betrayed as nonnative, especially in terms of their pronunciation and grammar (see, e.g., Diaz et al., 2008).

A 'real-life' event that underlines this point was described by Monica Schmid (e.g., 2012). At the brink of the Second World War, in the years 1938–39, a very large number of Jewish children were transferred from Germany to England by charity organizations to prevent their persecution. This transfer has been called the '*Kindertransport*'. The age of the children varied from very young (2 years) to post-puberty (17 years). Because the children ended up in foster parent families, they often lost all opportunities to speak German. In many cases, they were reluctant to do so anyway. However, sometimes the children escaped with other family members, and as a consequence some German communication was taking place. Since the 1990s, the testimonies of the children have been recorded in extensive archives. Using these, one can analyze the precise role played by the age at which the children acquired the FL and the moment at which they started to forget their mother tongue. Schmid showed that the age at which the children left Germany was more important for the quality of the remaining German, than whether they were transferred to England with other family members. It appeared that their language problems were especially concerned with the activation and access to their German lexicon, rather than a loss of their linguistic knowledge or the replacement of their first language by a second one.

There are many other studies on how age of acquisition and age of initial exposure (or age of arrival in a foreign country) affect the learning of a second language. In view of space limitations, we will consider only two of the most known of these. Johnson and Newport (1989) used a grammatical judgment task to test the English proficiency of 46 native Korean and Chinese speakers, who arrived in the United States between age 3 and 39, on a variety of English syntactic constructions, including past tense, plural, third person singular, present progressive, and determiners. There was a clear performance advantage for earlier arrivals relative to later arrivals that was linearly related to age of arrival up to puberty (set at age 16), but variable and unrelated to age of arrival after puberty. This effect appeared on all tested grammatical structures, although there were considerable performance differences between

constructions and individuals as well (for follow-up studies see Birdsong and Molis, 2001; Liu et al., 1992).

In an EEG study (Electroencephalography, i.e., recording of neural activity across the scalp), Weber-Fox and Neville (1996) investigated how the syntactic and semantic processing of Chinese learners of English changed depending on the age at which the learners were first exposed to English. Understandably, the learners' self-ratings for reading, listening, speaking, and writing in their new language decreased when they were older upon first exposure; in contrast, their first-language ratings increased. Similarly, their performance accuracy on sentences that had semantic or syntactic anomalies decreased with age of first exposure. Altered patterns of semantic and syntactic processing depending on age of first exposure were also visible in their EEG. For semantic processing, there were altered behavioral patterns and a latency shift in the relevant EEG markers with an age of first exposure higher than 11 years. For syntactic processing, differences in both types of measure were visible already at a much earlier delay (age of exposure 8 years or less). A conclusion is that especially syntactic processes appear to be qualitatively and quantitatively different in late L2 learners, whereas semantic processes often differ quantitatively (also see Ardal et al., 1990; Kroll and Dussias, 2004; Ojima et al., 2005).

Late Child Learners of a Second Language

In addition to the age factor, the dynamics of second-language learning have also been studied by considering the first stages of late second-language learning in children and in adults. With respect to child L2 learners, Poarch and van Hell (2012) considered the production performance of German children who had just begun to acquire English as a second language in school at ages 5–8. Three groups of young participants were tested: German learners of English, German–English bilinguals, and German–English–Language X trilinguals (where Language X was a 'dual L1'). All three groups were found to exhibit a cognate facilitation effect in picture naming in their nondominant language English. The more balanced bilinguals and trilinguals also showed a smaller facilitation effect in their dominant language, German. These findings indicate that the language processing system is not only shared for both languages in adults, but also in young children.

Two word-level studies by Brenders and colleagues considered the L2 comprehension performance of Dutch children aged 9–11 years who had just started to acquire English as a second language in primary school. The first study (Brenders et al., 2011) reported a cognate facilitation effect in L2 lexical decision for L2 learners even in the early stages. Interestingly, adding false friends to the stimulus list turned this facilitation into inhibition. The explanation given by the authors was that the presence of false friends in the stimulus list makes the young participants aware of potential problems in mapping active items to the 'yes' response (remember that false friends and cognates activate their readings in both languages of a bilingual). Thus, responses to cognates suffered because suddenly their membership of the nontarget language was considered to be relevant. In the second study, Brenders et al. (in revision) used an L2 to L1 word translation task. Combining reaction times (RTs) and event-related potentials, they found that L2 words activated their conceptual representations early in the learning process, as indicated by N400 effects already in fifth-graders with only 5 months of English instruction (the N400 is a marker in the EEG that is sensitive to lexical and semantic integration). See also Poarch et al. (2014), for a similar study with RTs. Apparently, already at an early stage Dutch, beginning L2 learners of English make use of the L2 word form to concept mappings to access the L2 word meaning without mediation of the L1 word. This is in contradiction with the predictions of the RHM (see also Altarriba and Mathis, 1997; Cheng and Leung, 1989; Kroll et al., 2002; Brenders, 2012, gives a full review).

One suggestion to resolve this issue is that differences in obtained results are due to the different teaching methods applied to L2 learning, because they pay more or less attention to meaning aspects. Comesaña et al. (2009) studied the creation of L2 words' conceptual representations in children, examining the efficacy of two L2 word-learning methods: L2–L1 word association learning and L2-picture association learning. In a translation recognition task (Experiment 2), Spanish–Basque learners, aged 10–11 years, saw semantically related pairs of words from different languages and unrelated word pairs. They had to say whether the second word in the pair was a translation of the first one. A semantic interference effect (for pairs of words where prime and target were not translations but were semantically associated, like 'table' and 'chair') was already found after only one vocabulary learning session. Because the L2 picture method produced a larger semantic interference effect than the L2–L1 method, this suggests that the speed of linking L2 word forms to semantic representations may depend on the vocabulary training method. Furthermore, it is important to point out that word translation recognition is a relatively slow task that may not be optimally suited to address the issue of fast word form to concept mappings (also see Dijkstra and Rekké, 2010).

Late Adult Learners of a Second Language

The very first stages of second-language learning have also been investigated in *adults*.

With respect to adults, studies have shown that during the first encounters with an FL, many aspects of the new language are quickly picked up. For instance, Gullberg et al. (2010, 2012) showed that already after watching 7 min of a Chinese weather forecast, the viewers had acquired a considerable amount of information about the Chinese phonology of word forms and after 15 min they had already assigned meanings to particular words.

Researchers have extended their studies on this topic by considering electrophysiological (EEG) and magnetic resonance imaging techniques (e.g., Osterhout et al., 2006; Morgan-Short et al., 2012; Weber et al., in preparation). Whereas many of these studies are concerned with grammar learning, McLaughlin et al. (2004) determined after how much FL exposure a learner's EEG responses to existing FL words and pronounceable pseudowords (nonwords) started to resemble those of a native speaker. Participants were English-speaking university students in their first year of classroom instruction in French and a control group without

any French instruction. The learners performed a lexical decision task on all materials. After only 14 h of FL instruction, the pseudowords already elicited a larger N400 EEG response than the existing words. Interestingly, the effect occurred even though learners performed at chance level in a French lexical decision task. Thus, at a neurocognitive level, the learners became quickly sensitive to FL word forms, even when there was no clear benefit at the decision level yet. After about 60 h of FL instruction, smaller N400's arose to words preceded by a related than by unrelated word, reflecting sensitivity to word meaning. Finally, after about 140 h of FL instruction, the word/nonword and relatedness differences approximated that of native English or French speakers in amplitude, even though the learners' explicit word/nonword judgments remained very poor.

In all, these EEG data suggest that adult learners rapidly integrate lexical information in their mental lexicon (about word form, and in a later stage about word meaning). These effects are not always found in standard behavioral assessments (e.g., RTs), implying that the EEG might more sensitively reflect aspects of implicit learning and continuous lexical change.

Recently, two longitudinal projects at Radboud University Nijmegen, with the acronyms VILLA and SLANG, have begun to investigate the very first stages of second-language learning in adults from a perspective that is somewhat different from that taken by McLaughlin, Osterhout, and colleagues. The project called 'Varieties of Initial Learners in Language Acquisition' (VILLA) investigates the very first phases of FL acquisition in children and adults under controlled input conditions (for its predecessor, see Gullberg et al., 2010, 2012). In VILLA, adult and child learners with different native languages (Dutch, English, French, German, Italian) were exposed to 20 h of input in a new language (Polish) with which they were completely unfamiliar. The language input across sessions was similar to those in standard FL courses, but learning was held constant under relatively natural and interactive exposure conditions. The learners' performance was examined over a period of 2 weeks with respect to a diversity of linguistic levels (phonology, lexicon, morphosyntax, and semantics/pragmatics) in order to find out more about the most elementary stages of linguistic organization resulting from the teaching. Off-line language aptitude tests and on line psycholinguistic experiments were administered to the learners before and after classroom exposure at regular time intervals during the duration of the project.

Figures 1 and 2 illustrate some of the first results of the project. In this part of the project, participants performed a Polish lexical decision task every 2 days, i.e., five times in the 2-week learning period. The stimuli in the task were Polish words and nonwords with different characteristics. The Polish words consisted of noncognates or cognates with the native language of the participants (e.g., Polish 'MUSEUM' is

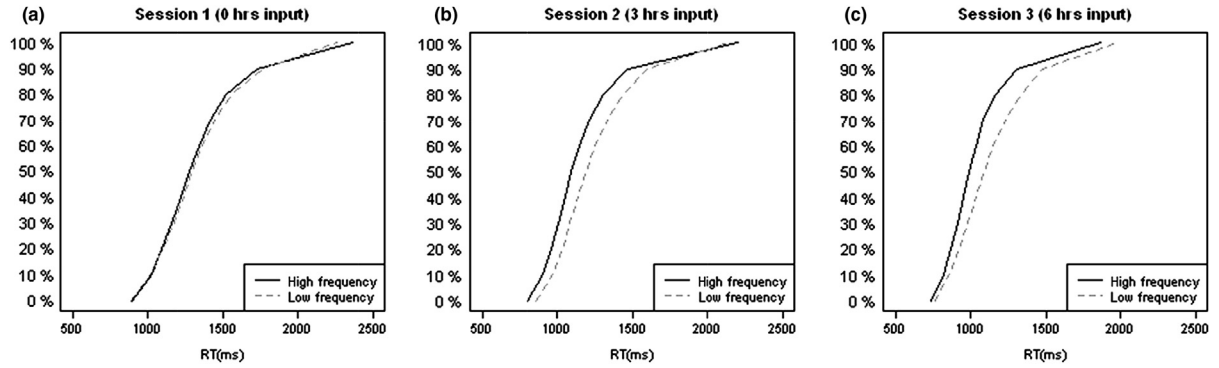

Figure 1 (a)–(c) Cumulative distribution functions for words with a high and a low frequency of presentation at the beginning of the Polish course, after 3 h, and after 6 h of input. More frequent presentation results in faster response times.

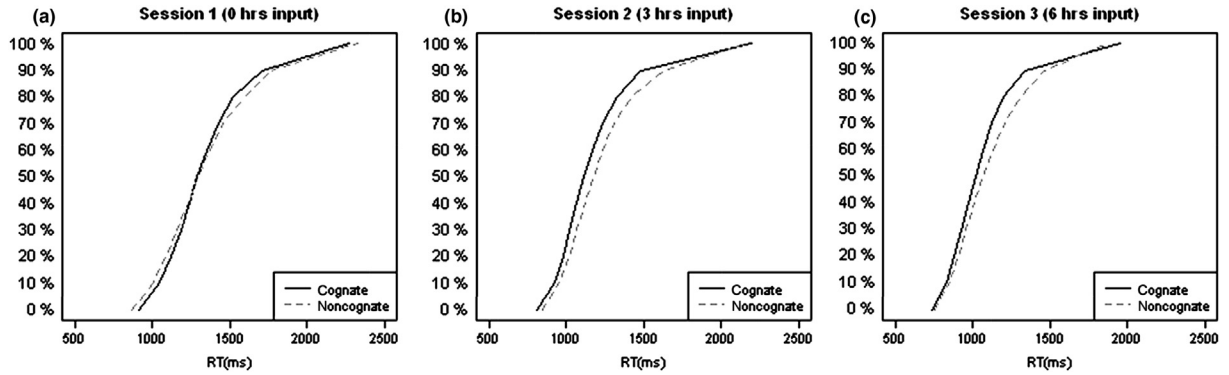

Figure 2 (a)–(c) Cumulative distribution functions for cognate and noncognate words at the beginning of the Polish course, after 3 h, and after 6 h of input. Cognates result in faster response times than noncognates.

phonologically very similar to 'MUSEUM' in Dutch) with a high or low frequency of presentation in the study (the exact number can be derived from a transcription of the teacher's speech). In Figure 1(a)–(c), it can be seen that from early on, words were processed faster and more accurately when their frequency of presentation was higher. From Figure 2(a)–(c), it is clear that participants quickly became sensitive to whether the item was a cognate between Polish and the native language (in this case, Dutch) or not. They very quickly noticed the cross-linguistic similarities in word form and meaning and used these to speed up their response times.

During the 20-h course, a morphosyntactic comprehension task was administered twice. This task was designed to investigate the development of nominal morphological marking and word order in the acquisition of the Polish morphosyntactic system. Polish is a highly inflected language with seven cases that allows relatively flexible word order. Two of the learner groups were native speakers of Dutch or French without any knowledge of Polish or another Slavic language. Dutch has much less case marking than Polish and also has fairly flexible word order. French has no case marking on the noun and has strict word order. For each source language, there were two groups of learners; one received 'meaning-based input,' in which the instructor provided no explicit information about grammar, and the other received 'form-based input.' in which the instructor highlighted aspects of morphosyntax. Results on the comprehension task, administered after 9 h and again after 13 h of Polish instruction, reveal a role of source language in that French and Dutch learners differed in their processing of case marking and word order. With regard to the type of input, no significant difference was found between the meaning-based input groups (French and Dutch) and the form-based input groups (French and Dutch) after 9 h of exposure to Polish. After 13 h, however, the French learners who received meaning-based input performed better than those who received form-based input. In contrast, the Dutch learners who received form-based input performed better than those who received meaning-based input. Analyses of individual differences tests revealed that individual variability may have played a role as well.

The VILLA project focused on behavioral aspects of FL learning during the first 2 weeks of exposure in a classroom-like setting (see Rast, 2008, for a theoretical background). However, longer lasting studies (Xiang et al., 2012) indicate that important further developments take place in the weeks and months after this initial learning stage. Furthermore, as the studies by Osterhout et al. (2006) and by Brenders et al. (in revision) suggest, EEG data might pick up early changes in the language processing system that behavioral assessments (i.e., latencies, error rates) are not yet sensitive to. For these reasons, another longitudinal project, called SLANG (for Second Language Acquisition), was started in which young adult German learners of Dutch were followed for a period of nearly half a year, with regular collection of their behavioral and electrophysiological responses to Dutch words of different types and enriched by off-line test data. Already at a very early stage of learning, words in the new language were found to be processed up to semantic and affective (emotion) levels. With respect to emotion words in the new language, the learners already responded differently after only 1–2 weeks of language learning (Middelburg et al., in preparation; Sianipar et al., in preparation). Positive words appeared to have a special status in the earliest stages of learning: RTs to these words were faster, fewer errors were made on them, and they displayed a more positive-going Late Positive Complex (LPC: an EEG marker at 600–800 ms after the stimulus event, often found in emotion word processing). Arousal of the words affected the EEG indirectly via interaction with the valence response. Valence itself played a role in L2 emotion word processing from the earliest stages of SLA onwards. In the course of half a year of L2 learning in the classroom and submerged in society, the results approached those of a native control group, but were still different at the end.

Conclusion and Future Developments

In this contribution, we discussed two core aspects of second-language acquisition, namely that it is (1) interactive and (2) dynamic. Of course, given the multifacetedness of the field of second-language acquisition, there are many more factors and notions that play a role in second-language acquisition and could have been discussed. This was not possible within the limited space of this contribution, however (see, e.g., Gass and Mackey, 2012; Robinson, 2013, for elaborate overviews of the field of second-language acquisition). Instead, we focused on the role of a number of central language-related and learner-related factors in the comprehension, production, and acquisition of a second language, and explained second-language acquisition and use from a psycholinguistic perspective on language processing.

In addition to these core issues, there are a number of developments that may lead to new insights in the future. One of these is the application of laboratory-based research techniques in schools. Brenders et al. (in revision), for instance, used EEG measurements in a study on bilingual processing in primary school children by driving a minivan equipped with EEG recording hardware to primary schools. This made it possible to test the neurocognitive mechanisms of second-language learning in primary school children, which would otherwise be much more effortful to investigate. Another development is the study of intraindividual variation in the SLA process. De Bot and Lowie (2010; see also De Bot et al., 2005), for instance, studied stability of lexical processing over time in single second-language learners and found interesting patterns of variation in these learners. This focus on variation over time can in the end lead to more insight into the dynamics of the language learning process. A final development that we name here concerns the application of insights from research on second-language acquisition in educational contexts. For instance, research has made it clear that the teaching methods used play an important role in the ways new information is integrated in the language systems that are already present in the FL learner (Comesaña et al., 2009). In addition, the insight that early learning of an FL is typically less effortful than late learning and may even lead to more general cognitive and academic advantages has led to the development that many schools in the Netherlands have now

implemented (partly) bilingual programs (mostly Dutch–English), sometimes already starting in kindergarten. Studies are now performed to investigate the circumstances in which these programs are indeed beneficial for children's linguistic, academic, and overall cognitive development.

In this article, we took a psycholinguistic perspective to show that second-language learning results from complex interactions on multiple dimensions. It is exciting to see that increased communication between researchers from different backgrounds, teachers, interpreters, language learners, and others slowly but surely results in a better understanding of this important human skill.

See also: Code Switching: Linguistic; First Language Acquisition, Linguistic Theory of; Functional Brain Imaging of Language processes; Language Acquisition; Language Contact; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexicon; Multilingual Language Development; Multilingualism; Neural Basis of Language Development; Psycholinguistics: Overview; Sentence Comprehension, Psychology of; Speech Perception; Speech Production, Psychology of; Visual Word Recognition, Neurocognitive Psychology of; Vocabulary Acquisition.

Bibliography

Altarriba, J., Mathis, K.M., 1997. Conceptual and lexical development in second language acquisition. Journal of Memory and Language 36, 550–568.

Appel, R., Muysken, P., 1987. Language Contact and Bilingualism. Edward Arnold, New York.

Ardal, S., Donald, M., Meuter, R., Muldrew, S., Luce, M., 1990. Brain responses to semantic incongruity in bilinguals. Brain and Language 39, 187–205.

Bernolet, S., Hartsuiker, R.J., Pickering, M.J., 2012. Effects of phonological feedback on the selection of syntax: evidence from between-language syntactic priming. Bilingualism: Language and Cognition 15, 503–516.

Bhatia, T.K., Ritchie, W.C., 2004. Introduction. In: Bhatia, T.K., Ritchie, W.C. (Eds.), Handbook of Bilingualism. Blackwell, Oxford, pp. 1–2.

Brenders, P., Van Hell, J.G., Dijkstra, T., 2011. Word recognition in child second language learners: Evidence from cognates and false friends. Journal of Experimental Child Psychology 109, 383–396.

Bhatia, T.K., Ritchie, W.C., 2013. The Handbook of Bilingualism and Multilingualism, second ed. Wiley-Blackwell.

Bialystok, E., Craik, F.I.M., Luk, G., 2012. Bilingualism: consequences for mind and brain. Trends in Cognitive Sciences 16, 240–250.

Birdsong, D., Molis, M., 2001. On the evidence for maturational constraints in second-language acquisition. Journal of Memory and Language 44, 235–249.

Bongaerts, T., van Summeren, C., Planken, B., Schils, E., 1997. Age and ultimate attainment in the pronunciation of a foreign language. Studies in Second Language Acquisition 19, 447–465.

Brenders, P., 2012. Cross-language Interactions in Beginning Second Language Learners (Unpublished doctoral thesis). Radboud University, Nijmegen.

Brenders, P., Van Hell, J.G., Dijkstra, T., 2011. Word recognition in child second language learners: evidence from cognates and false friends. Journal of Experimental Child Psychology 109, 383–396.

Brenders, P., van Hell, J.G., Dijkstra, T., The development of lexical and conceptual processing in second language learners. Electrophysiological and Behavioural Studies, in revision.

Chang, F., Dell, G.S., Bock, J.K., 2006. Becoming syntactic. Psychological Review 113, 234–272.

Cheng, H.-C., Leung, Y.-S., 1989. Patterns of lexical processing in a nonnative language. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 316–325.

Christoffels, I.K., Firk, C., Schiller, N.O., 2007. Bilingual language control: an event-related brain potential study. Brain Research 1147, 192–208.

Comesaña, M., Perea, M., Piñeiro, A., Fraga, I., 2009. Vocabulary teaching strategies and conceptual representations of words in L2 in children: evidence with novice learners. Journal of Experimental Child Psychology 104, 22–33.

Corder, S.P., 1967. The significance of learners' errors. International Review of Applied Linguistics 5, 160–170.

Crystal, D., 2003. English as a Global Language. Cambridge University Press.

De Bot, K., Lowie, W.M., 2010. On the stability of representations in the multilingual lexicon. In: Sicola, L., Pütz, M. (Eds.), Cognitive Processing in Second Language Acquisition. John Benjamins, Amsterdam/Philadelphia, pp. 117–134.

De Bot, K., Lowie, W.M., Verspoor, M., 2005. Second Language Acquisition: An Advanced Resource Book. Routledge, London.

De Groot, A.M.B., 2011. Language and Cognition in Bilinguals and Multilinguals: An Introduction. Psychology Press, New York, NY.

De Groot, A.M.B., Poot, R., 1997. Word translation at three levels of proficiency in a second language: the ubiquitous involvement of conceptual memory. Language Learning 47, 215–264.

Diaz, B., Baus, C., Escera, C., Costa, A., Sebastian-Galles, N., 2008. Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. Proceedings of the National Academy of Sciences 105 (42), 16083–16088.

Dijkstra, T., 2003. Lexical storage and retrieval in bilinguals. In: van Hout, R., Hulk, A., Kuiken, F., Towell, R. (Eds.), The Interface between Syntax and the Lexicon in Second Language Acquisition. John Benjamins, Amsterdam, pp. 129–150.

Dijkstra, T., 2005. Bilingual visual word recognition and lexical access. In: Kroll, J.F., De Groot, A. (Eds.), Handbook of Bilingualism: Psycholinguistic Approaches. Oxford University Press, pp. 178–201.

Dijkstra, T., 2009. The multilingual lexicon. In: Sandra, D., Östman, J.-O., Verschueren, J. (Eds.), Cognition and Language Use: Handbook of Pragmatics Highlights. Also Handbook of Pragmatics On-Line.

Dijkstra, T., Haga, F., Bijsterveld, A., Sprinkhuizen-Kuyper, I., 2012. Lexical competition in localist and distributed connectionist models of L2 acquisition. In: Altarriba, J., Isurin, L. (Eds.), Memory, Language and Bilingualism: Theoretical and Applied Approaches. Cambridge University Press, Cambridge, pp. 48–73.

Dijkstra, T., Rekké, S., 2010. Towards a localist-connectionist model for word translation. The Mental Lexicon 5 (3), 403–422. Special issue on: Jarema, G., Libben, G., Westbury, C. (Eds.), Methodological and analytic frontiers in lexical research.

Dijkstra, T., Van Heuven, W.J.B., 1998. The BIA model and bilingual word recognition. In: Grainger, J., Jacobs, A. (Eds.), Localist Connectionist Approaches to Human Cognition. Erlbaum, Hillsdale, NJ, pp. 189–225.

Dijkstra, T., van Heuven, W.J.B., 2002. The architecture of the bilingual word recognition system: from identification to decision. Bilingualism: Language and Cognition 5, 175–197.

Dijkstra, T., Van Jaarsveld, H., Ten Brinke, S., 1998. Interlingual homograph recognition: effects of task demands and language intermixing. Bilingualism: Language and Cognition 1, 51–66.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappeli, M., Baayen, R.H., 2010. How cross-language similarity affects cognate recognition. Journal of Memory and Language 62, 284–301.

Doughty, C., Long, M.H. (Eds.), 2003. Handbook of Second Language Acquisition. Blackwell, New York.

Dulay, H.S., Burt, M.K., 1974. Natural sequences in child second language acquisition. Language Learning 24, 37–53.

Ellis, R., 1994. The Study of Second Language Acquisition. Oxford University Press.

Gardner-Chloros, P., 2009. Code-Switching. Cambridge University Press.

Gass, S.M., 1987. The resolution of conflicts among competing systems: a bidirectional perspective. Applied Psycholinguistics 8, 329–350.

Gass, S.M., Mackey, A., 2012. The Routledge Handbook of Second Language Acquisition. Routledge, London/New York.

Green, D.W., 1998. Mental control of the bilingual lexico-semantic system. Bilingualism: Language and Cognition 1, 67–81.

Grosjean, F., 2008. Studying Bilinguals. Oxford University Press.

Grosjean, F., Li, P., 2013. The Psycholinguistics of Bilingualism. Wiley-Blackwell, Chichester, UK.

Gullberg, M., Roberts, L., Dimroth, C., Veroude, K., Indefrey, P., 2010. Adult language learning after minimal exposure to an unknown natural language. Language Learning 60, 5–24.

Gullberg, M., Roberts, L., Dimroth, C., 2012. What word-level knowledge can adult learners acquire after minimal exposure to a new language? International Review of Applied Linguistics 50, 239–276.

Harrington, M., 1987. Processing transfer: language-specific processing strategies as a source of interlanguage variation. Applied Psycholinguistics 8, 351–377.
Hartsuiker, R.J., Pickering, M.J., 2008. Language integration in bilingual sentence production. Acta Psychologica 128, 479–489.
Hartsuiker, R.J., Pickering, M.J., Veltkamp, E., 2004. Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. Psychological Science 15, 409–414.
Jarvis, S., Pavlenko, A., 2008. Crosslinguistic Influence in Language and Cognition. Routledge, London/New York.
Johnson, J.S., Newport, E.L., 1989. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cognitive Psychology 21, 60–99.
Kellerman, E., Sharwood Smith, M. (Eds.), 1986. Crosslinguistic Influence in Second Language Acquisition. Pergamon, Oxford.
Kootstra, G.J., Doedens, W.J. Cross-language interaction and convergence in bilingual sentence production: immediate and long-lasting effects of structural priming and verb bias between languages, in revision.
Kootstra, G.J., van Hell, J.G., Dijkstra, T., 2012. Priming of code-switches in sentences: the role of lexical repetition, cognates, and language proficiency. Bilingualism: Language and Cognition 15, 797–819.
Kroll, J.F., Bogulski, C.A., McClain, R., 2012. Psycholinguistic perspectives on second language learning and bilingualism: the course and consequence of cross-language competition. Linguistic Approaches to Bilingualism 2, 1–24.
Kroll, J.F., de Groot, A.M.B., 2005. Handbook of Bilingualism: Psycholinguistic Approaches. Oxford University Press.
Kroll, J.F., Dussias, P.E., 2004. The comprehension of words and sentences in two languages. In: Bhatia, T., Ritchie, W. (Eds.), Handbook of Bilingualism. Blackwell Publishers, Cambridge, MA, pp. 169–200.
Kroll, J.F., Michael, E., Tokowicz, N., Dufour, R., 2002. The development of lexical fluency in a second language. Second Language Research 18, 137–171.
Kroll, J.F., Stewart, E., 1994. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. Journal of Memory and Language 33, 149–174.
Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 5, 831–843.
Lado, R., 1957. Linguistics across Cultures: Applied Linguistics for Language Teachers. University of Michigan Press, Ann Arbor.
Lewis, M.P. (Ed.), 2009. Ethnologue: Languages of the World, sixteenth ed. SIL International, Dallas, TX. Retrieved from: www.ethnologue.com.
Li, W. (Ed.), 2007. The Bilingualism Reader, second ed. Routledge, London/New York.
Liu, H., Bates, E., Li, P., 1992. Sentence interpretation in bilingual speakers of English and Chinese. Applied Psycholinguistics 13, 451–484.
Loebell, H., Bock, J.K., 2003. Structural priming across languages. Linguistics 41, 791–824.
MacWhinney, B., 2005. A unified model of language acquisition. In: Kroll, J.F., de Groot, A.M.B. (Eds.), Handbook of Bilingualism: Psycholinguistic Approaches. Oxford University Press, pp. 49–67.
MacWhinney, B., 2008. A unified model. In: Robinson, P., Ellis, N.C. (Eds.), Handbook of Cognitive Linguistics and Second Language Acquisition. Routledge, New York/London, pp. 341–371.
MacWhinney, B., 2012. The logic of the unified model. In: Gass, S.M., Mackey, A. (Eds.), The Routledge Handbook of Second Language Acquisition. Routledge, London/New York, pp. 211–227.
MacWhinney, B., Bates, E., 1989. The Crosslinguistic Study of Sentence Processing. Cambridge University Press.
McClelland, J.L., Rumelhart, D.E., 1981. An interactive activation model of context effects in letter perception, part 1: an account of basic findings. Psychological Review 88, 375–405.
McClelland, J.L., Rumelhart, D.E., 1988. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: A Handbook of Models, Programs, and Exercises. Bradford Books, Cambridge, MA.
McDonald, J.L., Heilenman, L.K., 1992. Changes in sentence processing as second language proficiency increases. In: Harris, R.J. (Ed.), Cognitive Processing in Bilinguals. North Holland, Amsterdam, pp. 325–336.
McDonald, J.L., 1987. Sentence interpretation in bilingual speakers of English and Dutch. Applied Psycholinguistics 8, 379–413.
McDonough, K., 2006. Interaction and syntactic priming: English L2 speakers' production of dative constructions. Studies in Second Language Acquisition 28, 179–207.
McLaughlin, J., Osterhout, L., Kim, A., 2004. Neural correlates of second-language word learning: minimal instruction produces rapid change. Nature Neuroscience 7 (7), 703–704.
Middelburg, R., Sianipar, A., Dijkstra, T. Learning emotion words in a new language: a longitudinal behavioral and electrophysiological study, in preparation.
Morett, L.M., MacWhinney, B., 2013. Syntactic transfer in English-speaking Spanish learners. Bilingualism: Language and Cognition 16, 132–151.
Morgan-Short, K., Finger, I., Grey, S., Ullman, M.T., 2012. Second language processing shows increased native-like neural responses after months of no exposure. PLoS ONE 7 (3), e32974. http://dx.doi.org/10.1371/journal.pone.0032974.
Nitschke, S., Kidd, E., Serratrice, L., 2010. First language transfer and long-term structural priming in comprehension. Language and Cognitive Processes 25, 94–114.
Odlin, T., 1989. Language Transfer: Cross-linguistic Influence in Language Learning. Cambridge University Press.
Ojima, Sh, Nakata, H., Kakigi, R., 2005. An ERP study of second language learning after childhood: effects of proficiency. Journal of Cognitive Neuroscience 17, 1212–1228.
Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., Molinaro, N., 2006. Novice learners, longitudinal designs, and event-related potentials: a means for exploring the neurocognition of second language processing. Language Learning 56, 199–203.
Peeters, D., Dijkstra, T., Grainger, J., 2013. The representation and processing of identical cognates by late bilinguals: RT and ERP effects. Journal of Memory and Language 68, 315–332.
Poarch, G.J., Van Hell, J.G., 2012. Cross-language activation in children's speech production: Evidence from second language learners, bilinguals, and trilinguals. Journal of Experimental Child Psychology 111, 419–438.
Poarch, G.J., Van Hell, J.G., Kroll, J.F., 2014. Accessing word meaning in beginning second language learners: Lexical or conceptual mediation? Bilingualism: Language & Cognition. Advance online publication. http://dx.doi.org/10.1017/S1366728914000558.
Rast, R., 2008. Foreign Language Input: Initial Processing. Multilingual Matters, Clevedon.
Rivera-Gaxiola, M., Silva-Pereyra, J., Kuhl, P.K., 2005. Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. Developmental Science 8, 162–172.
Robinson, P., Ellis, N.C., 2008. Handbook of Cognitive Linguistics and Second Language Acquisition. Routledge, New York/London.
Robinson, P., 2013. The Routledge Encyclopedia of Second Language Acquisition. Routledge, New York.
Schepens, J., Dijkstra, T., Grootjen, F., 2012. Distribution of cognates in Europe as based on Levenshtein distance. Bilingualism: Language and Cognition 15 (1), 157–166.
Schepens, J., Dijkstra, T., van Heuven, W.J.B., Grootjen, F., 2013. Cross-language distributions of high frequency and phonetically similar cognates. PLoS ONE 8 (5), e63006.
Schmid, M.S., 2012. The impact of age and exposure on forgetting and retention of the birth language in international adoptees: a perspective from Holocaust survivors. Linguistic Approaches to Bilingualism 2 (2), 177–208.
Schoonbaert, S., Hartsuiker, R.J., Pickering, M.J., 2007. The representation of lexical and syntactic information in bilinguals: evidence from syntactic priming. Journal of Memory and Language 56, 153–171.
Shin, J.-A., Christianson, K., 2012. Structural priming and second language learning. Language Learning 62, 931–964.
Sianipar, A., Middelburg, R., Dijkstra, T. When feelings arise with meanings: how emotion and meaning of a native language affect second language processing in adult learners, in preparation.
Smits, E., Martensen, H., Dijkstra, A., Sandra, D., 2006. Naming interlingual homographs: Variable competition and the role of the decision system. Bilingualism: Language and Cognition 9, 281–297.
Talamas, A., Kroll, J.F., Dufour, R., 1999. From form to meaning: stages in the acquisition of second-language vocabulary. Bilingualism: Language and Cognition 2, 45–58.
Thomas, M.S.C., 1997. Connectionist Networks and Knowledge Representation: The Case of Bilingual Lexical Processing (Unpublished D.Phil. thesis). Oxford University.
Trofimovich, P., McDonough, K., 2011. Insights from Psycholinguistics: Applying Priming Research to L2 Learning and Teaching. John Benjamins, Amsterdam/Philadelphia.
Ullman, M.T., 2001. The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. Bilingualism: Language and Cognition 4, 105–122.
Van Assche, E., Drieghe, D., Duyck, W., Welvaert, M., Hartsuiker, R.J., 2011. The influence of semantic constraints on bilingual word recognition during sentence reading. Journal of Memory and Language 64, 88–107.

Van Hell, J., Dijkstra, A., 2002. Foreign language knowledge can influence native language performance. Psychonomic Bulletin & Review 9 (4), 780–789.

Van Hell, J.G., Tanner, D.S., 2012. Effects of second language proficiency on cross-language activation. Language Learning 62 (Suppl. 2), 148–171.

Van Heuven, W.J.B., Dijkstra, A., Grainger, J., 1998. Orthographic neighborhood effects in bilingual word recognition. Journal of Memory and Language 39, 458–483.

Voga, M., Grainger, J., 2007. Cognate status and cross-script translation priming. Memory & Cognition 35, 938–952.

Weber-Fox, C.M., Neville, H.J., 1996. Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. Journal of Cognitive Neuroscience 8, 231–256.

Weber, K., Christiansen, M.H., Indefrey, P., Hagoort, P. Syntactic and lexical priming during the first days of (second) language acquisition, in preparation.

Weber, K., Indefrey, P., 2009. Shared syntactic information in German-English bilinguals: an fMRI repetition suppression study. Neuroimage 46, 1164–1172.

Weinreich, U., 1953. Languages in Contact: Findings and Problems. Mouton, The Hague.

Werker, J.F., Gilbert, J.H.V., Humphreys, K., Tees, R.C., 1981. Developmental aspects of cross-language speech perception. Child Development 52, 349–355.

Xiang, H., Dediu, D., Roberts, L., Van Oort, E., Norris, D.G., Hagoort, P., 2012. The structural connectivity underpinning language aptitude, working memory, and IQ in the perisylvian language network. Language Learning 62 (2), 110–130.

Second Language Learning

Ruiqin Miao, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents a general introduction to second language acquisition (SLA) as an academic discipline investigating how a second language is learned. It first defines and distinguishes between some terminologies commonly used in the field. It then reviews the influential theories proposed to account for SLA and examines the major factors that may affect acquisition. In addition, it predicts the future trends of SLA research in the twenty-first century. Lastly, the chapter concludes with a summary of the discussions presented in the previous sections.

Introduction

The topic of second language learning and teaching has been intriguing scholars since ancient times, but second language acquisition (SLA) as an independent research discipline is quite new, having started in the 1960s. Unlike their predecessors who focused on pedagogy, SLA researchers are interested in the nature and process of learning. They seek answers to such questions as how a second language (L2) is acquired, and what makes individuals learn at different rates and achieve different degrees of success.

Since its very beginning, SLA has been an interdisciplinary field. To uncover and understand the mechanism of acquisition, scholars draw upon theories and findings from a variety of fields, including linguistics, psychology, education, sociology, anthropology, among others. SLA is a subject worthy of serious and scientific exploration for a number of reasons. Most importantly, of course, it can help us understand the process of second language learning. It can also benefit second language teaching, e.g., with respect to curriculum design and teaching methodology. In addition, it may shed lights on other related disciplines, especially disciplines which it has drawn knowledge from.

Definitions

The term 'second language acquisition' in a broad sense refers to the learning of a nonnative language after the first language (L1) (i.e., the native language) has been learned, either in a naturalistic setting or in a formal classroom setting. Before we go into detail, it is necessary to clarify some terms commonly used in the filed.

- Second language versus foreign language

The distinction between 'second language' and 'foreign language' is related to the function of the language in concern in the larger sociocultural setting where learning takes place. A second language is a language that plays important social and institutional functions in a country although it may not be the native language (NL) of the dominant population, for example, English in India and Singapore. In the case of language learning, a second language is a nonnative language to which the learner has natural exposure, as in the case of the learning of English by many immigrants in the United States.

A foreign language, on the other hand, is a language that is not the native language of the majority of the population, nor is it widely used as a medium of communication in the country. Instead, it is only used for speaking to foreigners or for reading written materials. It is usually learned as a subject in school. A typical example is the learning of English by a Japanese speaker in Japan.

- Second versus third language acquisition

Although SLA often refers to the learning of any language other than L1, in recent decades there has emerged an area of research focusing on the learning of the third, or even the fourth, language, which is known as multilingual acquisition. One major issue studied in this field is the relationship between the third (or fourth) language and other languages the learner has learned previously, e.g., L1 or L2 (see Gass and Selinker, 2008, pp. 21–23).

- Acquisition versus learning

SLA researchers sometimes distinguish between 'acquisition' and 'learning.' Acquisition refers to the learning of a nonnative language in a naturalistic environment (e.g., when a Russian speaker learns French in France). Learning, in contrast, refers to the learning of a nonnative language in an environment in which that language is not natively spoken (e.g., when a Russian speaker learns English in Russia). This distinction correlates with the distinction between 'second' versus 'foreign' language as discussed above, and hence the terms 'second language acquisition' and 'foreign language learning' are used with distinctions on a similar dimension.

Unless noted otherwise, this article follows the practice of most SLA researchers, without making the fine distinctions introduced above (e.g., Larsen-Freeman and Long, 1991; Ellis, 1994; Gass and Selinker, 2008). The term 'second language acquisition' is used to refer to the learning of any nonnative language, regardless of the sociocultural nature of the learning environment (second vs foreign) and the order in which the language is learned (second vs third, or fourth). Nevertheless, it should be pointed out that the learning environment exerts significant influences on SLA. For this reason, SLA in the classroom is sometimes discussed separately under various titles, e.g., 'instructed second language acquisition' (Larsen-Freeman

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92096-8

and Long, 1991), 'classroom second language acquisition' (Ellis, 1994), and 'instructed second language learning' (Gass and Selinker, 2008).

How Is a Second Language Acquired: Theories of SLA

A variety of theoretical frameworks have been proposed in the field of SLA to describe and account for the L2 learning process and "learners' underlying knowledge" (Ellis, 1994, p. 13) of the target language (TL). Some theories focus on the explanation of L2 output data, and some on formulating models of the learning process. This section will introduce the major SLA theories.

The Behaviorist Approach

The earliest approach to SLA is Contrastive Analysis (CA) developing under the influence of psychological theories, especially behaviorism prevailing in the 1950s and 1960s.

- Contrastive Analysis

During the 1940s and 1950s, researchers interested in improving language pedagogy believed that contrastive analyses of the learners' native language and the target language would reveal what structures would be easy to learn and what would be difficult. This led to the proposal of the Contrastive Analysis Hypothesis (CAH) by Lado (1957), who argues that L2 learners tend to transfer L1 forms to L2 use and that L2 errors could be predicted through systematic comparisons between L1 and L2 (cited in Larsen-Freeman and Long, 1991, pp. 52–56; Gass and Selinker, 2008, pp. 89–102). Similar structures will be easy to acquire due to positive transfer, whereas different ones will be difficult and cause negative transfer or interference (cf Flege, 1992, 1995, cited in Major, 2001, pp. 38–41, for a different view).

L1 transfer is especially powerful at the beginning stage of L2 acquisition, and it may occur at all levels of linguistic structure, from phonology, morphology, lexicon, syntax, semantics, to discourse and cultural norms. Some examples of L2 errors due to negative L1 transfer are given in **Table 1**, where L1 is Mandarin Chinese and L2 is English. For example, in the native language (Mandarin), a same word 'wen ti' can be used to mean either a question or a problem. In the target language (English), however, 'question' and 'problem' are two words of different meanings. In this case, one single form in L1 corresponds to two distinct items in L2, and hence an error occurs when the learner uses one of the two L2 words (e.g., 'question') to refer to the other (e.g., 'problem').

CAH developed at a time when the behaviorist school of psychology prevailed. According to behaviorism, learning is a process of habit formation, during which the learner applies his previously learned skills to the learning of a new one. New habits are developed through repeated imitation and intensive practice. When learning a second language, one needs to overcome L1 habits in order to acquire new habits in L2.

CAH was soon challenged by findings from empirical studies. For instance, some errors predicted by CA do not actually occur. In addition, some errors that actually occur are not predicted and hence cannot be attributed to L1 transfer. Furthermore, CA, at least in its earlier forms, cannot predict the relative difficulty of L2 structures, that is, what areas cause more difficulties than others (Major, 2001, p. 34). Despite these problems, it is undeniable that L1 transfer is an important factor influencing L2 learning. A weak version of CAH using CA to explain (rather than to predict) some L2 errors, is still commonly practiced today.

Cognitive Approaches

Some other approaches to examining the nature of the L2 learning process are based on cognitive perspectives. In such frameworks, language learning is seen as a process in which learners constantly form hypotheses by use of their own cognitive abilities. Three theories can be grouped under this category, the Error Analysis Hypothesis, the Universal Grammar (UG) Model and the Monitor Model.

- Error Analysis

In the late 1960 and 1970s, researchers started to view L2 errors from a new perspective. Corder (1967) expounds the importance of errors, arguing that errors are systematic and serve as significant evidence of the learner's language system (see reviews in Larsen-Freeman and Long, 1991, p. 58; Gass and Selinker, 2008, p. 102). Selinker (1972) introduces the term 'Interlanguage' (IL) to refer to the L2 learner's language system or grammar during the course of learning (cited in Ellis, 1994, p. 44). In his view, the learner actively constructs grammars based on the limited input he has received. These grammars constitute a continuum, with L1 at one end and L2 at the other (Ellis, 1994, pp. 350–356).

Interest in learner errors triggers a series of studies on the classification of IL errors. Major types of errors are interlingual errors – errors caused by NL transfer, and intralingual

Table 1 Examples of L1 transfer errors

	L1 (Mandarin)	*L2 (English)*	*L2 output*	*Error example*
Phonology	Lack of stop codas	Existence of stop codas	Vowel insertion	*/li:də/ for 'lead' /li:d/
Morphology	Isolating	Inflectional	Omission of inflections	*He paint the house last week. (*'Paint' for 'painted')
Syntax	Wh-in-situ	Overt wh-movement	Wh-in-situ	*You go where? (For 'where do you go?')
Semantics	One word (e.g., 'wen ti')	Two words (e.g., 'question,' 'problem')	One word (e.g., 'question')	*Water pollution is a serious question. (*'Question' for 'problem')
Paragraph development	Nonlinear	Linear	Nonlinear	e.g., Kaplan (1966)

Note: The asterisk '*' indicates an ungrammatical structure.

errors – errors independent of NL (e.g., overgeneralization and simplification errors). Other types include communication-based errors (i.e., errors resulting from a speaker's use of communicative strategies) and induced errors (i.e., errors due to training) (see discussions in Larsen-Freeman and Long, 1991, pp. 58–60; Ellis, 1994, pp. 57–63).

Error analysis (EA) replaced CA as a dominant approach to SLA research in the 1970s. As the term denotes, it focuses on the analysis of learners' errors. Unlike CA comparing L1 and L2, EA studies the learner's IL by examining the differences between the IL and L2. It was the first theory focused on serious investigations of learner language (Ellis, 1994, pp. 68–70). It broadened the scope of error sources, taking into consideration not just L1 transfer but also others (e.g., overgeneralization or training).

EA has a number of limitations. Firstly, it is not always easy to identify the type and source of an error. For instance, a deviant English form such as "She *arrive* late yesterday" produced by a Mandarin Chinese speaker learning English may be considered either as a morphological error (with lack of the past tense marker '-ed') or as a pronunciation error (with deletion of the final consonant /d/ in 'arrived'); it can be either a transfer error since Mandarin (L1) lacks inflectional morphology or a developmental error that a native English-speaking child also makes. Secondly, EA has a limited view on L2 acquisition by only looking at errors that actually occur, without considering the whole picture of the learner's behavior, e.g., the strategies to avoid using a difficult form and hence make fewer errors (see discussions in Larsen-Freeman and Long, 1991, pp. 61–62; Gass and Selinker, 2008, pp. 104–110).

- UG Model

The Universal Grammar (UG) (Chomsky, 1965, cited in White, 2003, pp. 2–4) assumes that the mental grammar an adult has of his or her native language consists of universal principles and variable settings of parameters. The principles are the same in all human languages, and parameter settings vary across languages. In language acquisition, input triggers the setting of parameters in the particular language being learned.

Since the 1980s, some researchers have started to explore the application of UG to SLA research. Major questions pursued along this line include whether UG remains accessible in SLA and how universals affect SLA. For instance, White (2003) studies the extent to which L2 acquisition is constrained by universal principles, arguing that UG constraints do function in SLA. She outlines different possibilities for the initial state of L2 learning, with different degrees of L1 transfer. For instance, in the case of full transfer (Schwartz and Sprouse, 1994, 1996, cited in White, 2003, pp. 61–68), L1 grammar constitutes the initial state. On the contrary, a full access situation (Epstein et al., 1996, 1998, cited in White, 2003, pp. 88–95) consists entirely of UG in the initial state.

However, L1 acquisition and L2 learning are different in many ways, which raises caution about the strong claims of access to UG in SLA. For instance, L2 learning starts out with the leaner having already formed a set of skills and strategies in L1, but this is not the case in L1 acquisition. L2 learners also have different motivations from children acquiring their NL. In addition, the sociocultural setting of SLA is much more complex than that of children's L1 acquisition. These differences present challenges to the arguments of UG availability in SLA.

- The Monitor Model

The Monitor Model was formulated by Krashen in the 1970s and 1980s. It seeks to account for the natural order of acquisition on the one hand and the disturbances observed in the order in specific performance tasks on the other.

Krashen (1976) distinguishes between two knowledge systems governing L2 performance (cited in Larsen-Freeman and Long, 1991, pp. 240–241). One is the acquired system, consisting of subconscious knowledge similar to what children have for native language acquisition. The other is the learned system, which results from formal instruction and serves to monitor performance (output) consciously. The two systems are completely separate, each accounting for a different aspect of performance. The natural order in which TL rules are acquired embodies the acquired system, whereas disturbances in the order result from conscious monitoring by the learned system.

The Monitor Model contains five hypotheses, two of which state crucial conditions necessary for SLA. One is the Input Hypothesis, which postulates that the learner acquires a second language through comprehensible input and that progress in learning is achieved when the input is one level above the current stage of the learner's interlanguage development (see discussions in Larsen-Freeman and Long, 1991, pp. 242–243; Gass and Selinker, 2008, pp. 309–310). The other is the Affective Filter Hypothesis, which posits that affective factors like motivation and self-confidence promote acquisition when the filter is low (e.g., when the learner is highly motivated and confident in learning) (see reviews in Larsen-Freeman and Long, 1991, pp. 243–244; Gass and Selinker, 2008, pp. 402–403).

Various components of the Monitor Model have encountered criticisms (Larsen-Freeman and Long, 1991, pp. 245–249). For instance, the model provides no measures to distinguish between acquisition and learning. Furthermore, the Input Hypothesis and the function of the affective filer cannot be easily tested. Nevertheless, this theory motivates new thoughts about the relationship between various factors in L2 acquisition and about the issues to consider in designing language instruction materials.

Interactional Approaches

Interactional approaches take into consideration the interactional nature and the communicative functions of language to account for the mechanism of second language learning.

- Conversation Analysis

Conversation Analysis, a subarea of discourse analysis, examines language learning as situated in the social interaction that the learner participates in. It recognizes that language cannot be understood in isolation from social contexts and argues that changes in the structure of the interaction discourse evidence learning (see review in Gass and Selinker, 2008, pp. 281–283). In this framework, the conversation between the L2 learner and his or her interlocutor constitutes an important source of language acquisition, serving a variety of functions, e.g., to provide feedback and corrections, to encourage participation and to negotiate meaning. For instance, Hatch

(1978) argues that a learner can acquire syntactic structures through conversations or verbal interactions with another person (cited in Larsen-Freeman and Long, 1991, pp. 70–71).

Conversation analysis focuses on the significant role of input and interaction in language learning. By including both form and function, it initiates a number of new directions to investigate SLA, e.g., foreigner talk, contrastive rhetoric, classroom discourse, and speech act analysis (see Larsen-Freeman and Long, 1991, pp. 71–73).

- The Multidimensional Model

The Multidimensional Model developed from the ZISA (Zweitsprachenwerb Italienischer und Spanischer Arbeiter) project conducted in Germany in the 1970s. It was first proposed by Meisel et al. (1981) (cited in Larsen-Freeman and Long, 1991, pp. 270–287; Ellis, 1994, pp. 382–388). Based on the development of word order rules in German as a second language, ZISA researchers propose that two independent dimensions regulate the developmental sequences of the IL. One is the developmental axis, which represents an invariant order of developmental stages. The other is the variation axis, which accounts for the differences among individual learners with respect to whether their orientation toward a learning task is a 'standard' (integrative) one or a 'simplifying' (segregative) one. By using the concept of orientation, the model incorporates the role of social and psychological factors such as motivation and interest in accounting for the learning process.

Although a powerful tool to explain IL development, the Multidimensional Model is not flawless (Larsen-Freeman and Long, 1991, pp. 283–287; Ellis, 1994, pp. 387–388). For instance, it accounts for acquisition in view of the learner's L2 production only, dismissing the interaction between comprehension and production. In addition, it is hard to identify formulaic chunks of language and variational features a priori.

- The Competition Model

The Competition Model (Bates and MacWhinney, 1982, cited in Gass and Selinker, 2008, pp. 221–226) adopts a functional approach to language development. It claims that language forms "are created, governed, constrained, acquired and used in the service of communicative functions" (MacWhinney et al., 1984, cited in Ellis, 1994, p. 374). A form fulfills one or more functions, and a function is realized in one or more forms.

According to this model, languages differ in the strengths of particular form-function mappings. The input provides four types of cues to convey meaning, namely word order, vocabulary, morphology and intonation. Since the human processing system employs only a limited number of information at a time, these different signals compete for processing space. In learning a second language, one needs to figure out which forms are mapped to what functions and what weights these mappings assume in the target language (see Ellis, 1994, pp. 373–378 for more discussions).

Sociocultural Approaches

Sociocultural approaches view language as a social phenomenon and hence language acquisition as crucially determined by sociocultural factors.

- The Acculturation Model

Based on a case of SLA in a naturalistic setting, Schumann (1978a) outlines a range of similarities between an unsuccessful learner's interlanguage and a pidgin, a simplified or reduced form of language used for communication between speakers of different native languages. For instance, both language forms display such characteristics as simple morphology, lack of certain movement rules, and reduced vocabulary (see Larsen-Freeman and Long, 1991, pp. 254–255).

Schumann proposes the Acculturation Model to account for his observations, arguing that naturalistic SLA results from acculturation, i.e., adaptation to or integration into another culture. He claims that "the degree to which a learner acculturates to the target-language group will control the degree to which he acquires the second language" (Schumann, 1978b, p. 34, cited in Larsen-Freeman and Long, 1991, p. 257; Ellis, 1994, p. 230). A high degree of acculturation leads to a greater amount of contact with native speakers of the L2, which facilitates acquisition.

Schumann develops the concepts of social distance and psychological distance to define the extent of acculturation. Each involves a number of factors regulating the relationship between the learner and the target language speaker group. A learner socially and psychologically integrated with the TL group achieves a high degree of acculturation, which further leads to successful L2 acquisition (for detailed discussions, see Larsen-Freeman and Long, 1991, pp. 251–258; Ellis, 1994, pp. 230–233).

While recognizing the significance of social and psychological factors in SLA, the Acculturation Model faces various challenges (Larsen-Freeman and Long, 1991, pp. 258–266; Ellis, 1994, pp. 233–234). For example, it does not specify how those social and psychological factors may regulate the features of the interlanguage. In addition, the factors underlying social distance between the two language groups involved in pidginization (i.e., the superstrate language group and the pidgin group) can be quite different from those involved in SLA (i.e., the individual L2 learner and the native TL speaker group). Moreover, the validity of positing that social factors have a direct effect on L2 acquisition raises doubt since they are generally considered indirect forces.

- Vygotsky's sociocultural theory

Since the 1990s, there has been a revival of interest in Vygotsky's (1978) sociocultural theory (cited in Gass and Selinker, 2008, pp. 283–285), which sees learning as social in nature and language acquisition as situated in social interaction. A central concept of this framework is the Zone of Proximal Development (ZDP), which refers to the distance between a child's actual level of development in terms of problem solving skills and the potential level of skills he is able to reach under the guidance of an adult or through collaboration with a peer who is more capable. ZDP defines a range of developmental levels where a child can learn through accomplishing tasks under guidance.

In the recent decades, a number of studies have employed this approach in SLA research. For instance, it has been shown that learners employ private speech to regulate their language in learning tasks (e.g., Lantolf and Thorne, 2006, 2007, cited in Gass and Selinker, 2008, p. 285).

In short, a great variety of models and approaches have been proposed to explain how a second language is acquired. There are some other models that are not introduced since they are not as extensively pursued as the approaches discussed above. Some examples are Givon's (1979a, 1979b) Functional-Typological theory (see Larsen-Freeman and Long, 1991, pp. 266–270), Eckman's (1977) Markedness Differential Hypothesis (see Gass and Selinker, 2008, pp. 179–182), and the Ontogeny Phylogeny Model (Major, 2001).

What Makes Individuals Learn Differently: Factors Influencing SLA

One phenomenon characteristic of second language learning is that individual learners vary greatly in the rate of learning and the level of achievement. Some learn fast, while others progress slowly. Some succeed in attaining a high proficiency, while others fail despite the enormous efforts they spend.

A plethora of factors have been investigated in the literature (e.g., Larsen-Freeman and Long, 1991; Ellis, 1994; Gass and Selinker, 2008). This section discusses a total of 12 factors, grouped into two categories, namely environmental/setting factors and individual learner factors.

Environmental/Setting Factors

The sociocultural environment and specific setting in which an L2 is learned has crucial influence on the learning process and achievement. It determines how learning is done, what resources are available to learners, what learning objectives are set, how fast learning progresses, and how motivated learners are.

- Naturalistic versus foreign environment

The naturalistic versus foreign nature of the larger sociocultural environment of L2 learning is one of the key factors determining how learning is done and how effectively it is done. Learners in a naturalistic environment acquire the TL through natural exposure in daily communication. They have access to abundant authentic input and plenty opportunities to use the TL, which may boot their motivation to learn, both the motivation to be integrated into the mainstream L2 society and that for practical reasons (e.g., to find a job). With frequent contact with native speakers of the L2, they enjoy advantages in acquiring speaking fluency and listening comprehension. In contrast, learners in a foreign environment get input mainly from classroom instruction, which is limited to interactions with teachers, peer learners and course materials. The input is usually somewhat modified to accommodate students' proficiency level. Learners have restricted opportunities to communicate with native speakers, and hence their motivation to learn is mainly instrumental. The skills they learn tend to focus on reading and writing. These differences related to the nature of the learning environment are summarized in Table 2 below.

Table 2 Differences between SLA in a naturalistic setting and a foreign setting

	Naturalistic	*Foreign*
Means of learning	Natural exposure	Instruction in classroom
Quantity of input	Abundant	Limited
Quality of input	Authentic	Modified
Learner motivation	Integrative and instrumental	Instrumental
Opportunities to use L2	Plenty	Restricted
Primary skills learned	Speaking and listening	Reading and writing

However, it needs to be pointed out that the reality of L2 learning environment is much more complex than a mere division between naturalistic and foreign settings. In a naturalistic setting, a person may attend formal classes to learn an L2 in addition to learning it through daily exposure in the native TL environment. On the other hand, a person learning an L2 in a foreign environment may sometimes create a native-like setting for himself by reading news, watching movies or play games in the TL.

In the long run, the naturalistic and foreign setting may each fare better in certain ways. For instance, learning in a natural L2 environment may lead to a higher level of fluency and communicative competency although not necessarily greater accuracy. A foreign setting, however, may benefit learners with modified input and help them develop a greater awareness of grammar accuracy, but it may lack resources facilitating acquisition of oral fluency and communicative skills.

- Technology

Technology has been an important tool for language learning ever since the mid-twentieth century. A variety of media and devices have been widely used, from conventional ones like radio, cassette recorders, television and video recorders to the current ones like mobile devices, computers and the Internet (Flowerdew and Miller, 2005, pp. 165–183; Godwin-Jones, 2008). For example, there are radio and television programs, audio and video materials, as well as computer software specially designed for language learning and teaching. In present days, it is not uncommon to find language labs and self-access language learning centers equipped with multimedia tools in schools.

Technological advances have brought about dramatic changes to the way L2 learners acquire resources (input) and approach learning tasks. For one thing, they expand the resources and opportunities for learners to get authentic L2 materials. For another, technology offers L2 learners immense freedom in planning and managing learning on their own.

The role of technology is especially significant for SLA in a foreign environment, where learners have limited natural exposure to the TL. For example, today an EFL (English as a foreign language) learner in Korea can have easy access to English news and movies on the Internet and use them to practice reading and listening comprehension anytime he or she wants. If without the aid of technology, the language resources he or she could have might come solely or mostly from the classroom – teachers, course materials and other learners.

Individual Learner Factors

A large body of SLA research looks into the influence of various learner-related factors. Major variables include the following

10: age, sex, affect, aptitude, attitudes, motivation, personality, learning styles, learning strategies, and social distance.

- Age

Studies show that children learn more successfully, but at a lower speed, than adults (see reviews in Larsen-Freeman and Long, 1991, pp. 155–163; Ellis, 1994, pp. 484–494; Major, 2001, pp. 6–12; Gass and Selinker, 2008, pp. 405–416). Young children tend to learn better than adults in the sense that they are more likely to achieve native-like proficiency. This is especially evident in phonology or pronunciation acquisition. However, the opposite tendency is revealed about the rate of acquisition. Older learners learn faster than young learners, especially at the early stage of acquiring morphology and syntax.

This observation of children's advantages over adults in attaining native-like proficiency is supported by the Critical Period Hypothesis (Lenneberg, 1967), which claims that there is a critical period of time, from age 2 to puberty, determining whether native proficiency in a second language can be acquired. For those who start to learn a language, either native or nonnative, after the critical period, it is very hard, or even impossible, for them to attain native-like levels.

- Sex

Females generally enjoy advantages over males in L2 learning (Larsen-Freeman and Long, 1991, p. 204; Ellis, 1994, pp. 202–204). The reason for this may be that women employ more standard and more prestige forms as well as more new forms than do men in using language. When learning a second language, they may be more open to the new input and more likely to eliminate errors, which contribute to their better L2 performance.

- Affect

Affect involves 'feelings or emotional reactions' about L2, the speakers as well as the culture of that language (Gass and Selinker, 2008, p. 398). It includes such specific factors as language/culture shock, stress, and anxiety. Shock and stress may cause problems to language learning. Anxiety can have either negative or positive effects, depending on how intense it is. Anxiety at a moderate or low level is helpful, but at a high level it is detrimental (see Ellis, 1994, pp. 479–483; Gass and Selinker, 2008, pp. 398–403).

- Aptitude

Aptitude is the capability of learning new skills. It is a cognitive ability closely related to, but not identical to, intelligence. Carroll (1981) defines four components of language learning aptitude, namely phonetic coding ability, grammatical sensitivity, rote learning ability for foreign language materials, and inductive language learning ability (cited in Larsen-Freeman and Long, 1991, p. 167; cf Ellis, 1994, pp. 495–496; Gass and Selinker, 2008, p. 418). A well-known instrument to measure aptitude is the Modern Language Aptitude Test (Carroll and Sapon, 1959, cited in Larsen-Freeman and Long, 1991, p. 168; Gass and Selinker, 2008, p. 419).

Aptitude is found to be a strong predictor of L2 performance. It demonstrates positive correlations to the rate of acquisition and achievement in language tests. In addition, it is linked to other factors such as motivation and social class (see discussions in Ellis, 1994, pp. 494–499; Gass and Selinker, 2008, pp. 417–426).

- Attitudes

Attitudes toward the target language and its speakers or community can have an important, although indirect, influence on L2 achievement. According to Gardner (1979), attitudes influence motivation, which in turn influences acquisition (cited in Larsen-Freeman and Long, 1991, p. 175). Generally speaking, positive attitudes toward L2 and its speakers/community facilitate learning.

- Motivation

Motivation is another key factor related to differential achievement in L2 learning. If a learner is motivated, he or she will have an intense interest and be willing to be engaged in learning (Ellis, 1994, pp. 512–514). Two kinds of motivation have been studied extensively, namely integrative and instrumental motivation proposed by Gardner and Lambert (1959, 1972) (cited in Larsen-Freeman and Long, 1991, pp. 172–174). Integrative motivation refers to the learner's desire to be integrated into or identified with the L2 community. Instrumental motivation is related to the usefulness of the language for practical goals, e.g., for career advancement. Both types of motivation serve as strong impetus to learning.

- Personality

Certain personality traits may favor L2 acquisition. Although research results are inconclusive, studies do provide evidence for the correlation between personality and L2 performance. Successful learners are believed to have high self-esteem, be extrovert and willing to take-risks, be tolerant of ambiguity, demonstrate empathy, and have low inhibition (see reviews in Larsen-Freeman and Long, 1991, pp. 184–192; Ellis, 1994, pp. 517–522).

- Learning styles

Learners differ in the styles or modes they prefer to use when approaching learning tasks. Researchers have examined such variables as field-independent/-dependent, reflective/impulsive, aural/visual, and analytic/gestalt styles (Larsen-Freeman and Long, 1991, pp. 192–197). A field–independent person, who is good at abstracting information from the context, learns better than does a field-dependent person, who pays great attention to details in the context. Reflective learners, who tend to think longer and more carefully, make fewer errors than do impulsive learners. Learners who learn by both aural and visual means perform better than do those who employ only a single modality. Furthermore, analytical learners tend to break language forms into smaller components, whereas gestalt learners take a holistic approach, assimilating language in big units and taking more risks in using it.

- Learning strategies

Learning strategies involve both cognitive and behavioral actions. O'Malley and Chamot (1990) define them as "the special thoughts or behaviors that individuals use to help them comprehend, learn, or retain new information" (p. 1) and classify them into three types, i.e., metacognitive, cognitive

and social/affective strategies (p. 43–47). Metacognitive ones are higher order skills used to plan, monitor and evaluate learning activities. Cognitive ones "operate directly on incoming information, manipulating it in ways that enhance learning" (p. 44). Examples are rehearsal, organization, summarization and detection. Finally, social/affective strategies are those learners use to interact with others and control affect (e.g., cooperation and questioning for clarification) (p. 45–46).

Good L2 learners demonstrate different patterns of strategy use from what poor learners employ. For instance, they tend to be attentive to form, monitor the speech of their own as well as others (Rubin, 1975, cited in Ellis, 1994, p. 547). They also attend to meaning, engage themselves actively in learning, and use more metacognitive strategies (see Ellis, 1994, pp. 546–550).

- Social distance

If a learner feels socially distant from the L2 community, he or she may not have a high motivation to integrate into that community and to participate in contact with the native speakers of the L2. This can lead to an inadequate amount of exposure and input, and hence hinder L2 acquisition (see Gass and Selinker, 2008, pp. 403–405). In consideration of the potential significance of such sociopsychological factors, Schumann (1978a, cited in Larsen-Freeman and Long, 1991, pp. 251–258) proposes the Acculturation Model to account for the process of SLA (see previous section).

To summarize, diverse factors related to the learning environment/setting and individual learners themselves can affect how well a second language is acquired. These variables do not function in isolation. Instead, they may be linked to each other. For instance, social distance may bear a close relationship to motivation, attitudes and affect. Aptitude may be interwoven with learning styles and learning strategies. Furthermore, learning styles may be influenced by personality, and the use of learning strategies is affected by various factors like age, motivation and the acquisition setting.

Besides those listed above, there are many other factors that more or less influence L2 acquisition too, e.g., memory, language disability, interest, birth order and prior experience with L1 (Larsen-Freeman and Long, 1991, pp. 203–206), as well as social class and ethnic identity (Ellis, 1994, pp. 204–210). But they are not as widely studied and their effects are not as significant or obvious as the aforementioned ones, thus they are not discussed here for the sake of brevity.

SLA in the Twenty-First Century

In the twenty-first century, SLA may develop toward an increasing diversity of research directions, growing attention to technology-related issues, and plurality of theoretical approaches. First, the scope of subject topics is broadening and research directions are expanding. In addition to the long-established areas such as L1 transfer, language universals and learner differences, more directions are emerging. For instance, a growing body of research in recent decades starts to investigate learner autonomy (e.g., in Dam, 2001), strategy training (e.g., Carrier, 2003) and task-based learning (e.g., in García Mayo, 2007).

Second, the development of CALL (Computer-Assisted Language Learning) triggers an array of technology-related topics worthy of investigation. Some questions waiting for exploration include the following: how to design effective CALL programs, how the mediation of technology affects the mental processes of learning, and in what way media literacy may influence the outcome of learning. For instance, Dodigovic (2005) discusses how to develop an intelligent CALL tool to assist L2 writing acquisition.

Thirdly, SLA theories are moving toward plurality. For one thing, established psycholinguistic, cognitive and sociocultural models will undergo new developments along with advances of related disciplines (e.g., cognitive science and psychology). For another, new sociocultural approaches may emerge within existent frameworks. For instance, there has been a growing interest in studying the negotiation and construction of learner identity (e.g., in Mantero, 2007) and the application of Vygotsky's sociocultural theory to SLA research.

With the increasing pace of globalization and the promotion of linguistic and cultural plurality in the modern age, it can be safely predicted that the population of second language learners and users will experience rapid growth in the new century. To account for the diversity of L2 sociocultural settings, a sound and comprehensive SLA theory in the future may well integrate the valid components from the various current approaches.

Conclusion

This article presents a general picture of SLA as an autonomous discipline of social science. It reviews the major theoretical approaches proposed in the field and examines a variety of factors that influence L2 acquisition. It also predicts some future directions of SLA research in the twenty-first century.

Acknowledgments

The writing of this article was supported by Shanghai Pujiang Program, China (Grant No. ZXDF069) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, China (Grant No. B1066W).

See also: Cognitive Styles and Learning Styles; Computer-Assisted Instruction; Cross-Cultural Study of Education; Foreign Language Teaching and Learning; Multicultural Education; Socialization and Education: Theoretical Perspectives; Tangible User Interfaces in Learning and Education; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Carroll, J.B., Sapon, S.M., 1959. Modern Language Aptitude Test "http://en.wikipedia.org/wiki/Harcourt_Assessment" "Harcourt Assessment". The Psychological Corporation, San Antonio, TX.

Carroll, J., 1981. Twenty-five years of research on foreign language aptitude. In: Diller, K. (Ed.), Individual Differences and Universals in Language Learning Aptitude. Newbury House, Rowley, MA.

Carrier, K.A., 2003. Improving high school English language learners' second language listening through strategy instruction. Bilingual Research Journal 27, 383–408.
Corder, S.P., 1967. The significance of learners' errors. International Review of Applied Linguistics 5, 161–170.
Chomsky, N., 1965. Aspects of the Theory of Syntax. The MIT Press, Cambridge.
Dam, L. (Ed.), 2001. Learner Autonomy: New Insights. The AILA Review, 15.
Dodigovic, M., 2005. Artificial Intelligence in Second Language Learning: Raising Error Awareness. Multilingual Matters, Clevedon.
Eckman, F., 1977. Markedness and the contrastive analysis hypothesis. Language Learning 27, 315–330.
Ellis, R., 1994. The Study of Second Language Acquisition. Oxford University Press, Oxford.
Flowerdew, J., Miller, L., 2005. Second Language Listening: Theory and Practice. Cambridge University Press, Cambridge.
García Mayo, M.P. (Ed.), 2007. Investigating Tasks in Formal Language Learning. Multilingual Matters, Clevedon.
Gardner, R., Lambert, W., 1959. Motivational variables in second language acquisition. Canadian Journal of Psychology 13, 266–272.
Gardner, R., Lambert, W., 1972. Attitudes and Motivation in Second language Learning. Newbury House, Rowley, MA.
Gardner, R., 1979. Social psychological aspects of second language acquisition. In: Giles, H., St. Clair, R. (Eds.), Language and Social Psychology. Basil Blackwell, Oxford, pp. 193–220.
Gass, S.M., Selinker, L., 2008. Second Language Acquisition: An Introductory Course, third ed. Routledge, New York.
Givon, T., 1979a. On Understanding Grammar. Academic Press, New York.
Givon, T., 1979b. From discourse to syntax: grammar as a processing strategy. In: Givon, T. (Ed.), Syntax and Semantics, Discourse and Syntax, vol. 12. Academic Press, New York.
Godwin-Jones, R., 2008. Emerging technologies mobile-computing trends: lighter, faster, smarter. Language Learning and Technology 12 (3), 3–9.
Hatch, E., 1978. Discourse analysis, speech acts and second language acquisition. In: Ritchie, W. (Ed.), Second Language Acquisition Research. Academic Press, New York, pp. 137–155.
Kaplan, R.B., 1966. Cultural thought patterns in inter-cultural education. Language Learning 16, 1–20.
Krashen, S., 1976. Formal and informal linguistic environments in language acquisition and language learning. TESOL Quarterly 10, 157–168.
Lado, R., 1957. Linguistics Across Cultures. University of Michigan Press, Ann Arbor, MI.
Lantolf, J.P., Thorne, S.L., 2006. Sociocultural Theory and the Genesis of Second Language Development. OUP, Oxford.
Lantolf, P., Thorne, S.L., 2007. Sociocultural theory and second language learning. In: Van Patten, B., Williams, J. (Eds.), Theories in Second Language Acquisition: An Introduction. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 693–701.
Larsen-Freeman, D., Long, M.H., 1991. An Introduction to Second Language Acquisition Research. Longman, London.
Lenneberg, E.H., 1967. Biological Foundations of Language. John Wiley & Sons, New York.
Major, R.C., 2001. Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology. Lawrence Erlbaum Associates, Mahwah, NJ.
Mantero, M. (Ed.), 2007. Identity and Second Language Learning: Culture, Inquiry, and Dialogic Activity in Educational Contexts. Information Age Publishing, Charlotte, NC.
Meisel, J., Clahsen, H., Pienemann, M., 1981. On determining developmental stages in natural second language acquisition. Studies in Second Language Acquisition 3, 109–135.
O'Malley, J.M., Chamot, A.U., 1990. Learning Strategies in Second Language Acquisition. Cambridge University Press, Cambridge.
Rubin, G., 1975. The traffic in women: notes on the 'Political Economy' of sex. In: Reiter, R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York.
Schumann, J., 1978a. The Pidginization Process: A Model for Second Language Acquisition. Newbury House, Rowley, MA.
Schumann, J., 1978b. The acculturation model for second language acquisition. In: Gingras, R. (Ed.), Second Language Acquisition and Foreign Language Teaching. Center for Applied Linguistics, Arlington, VA, pp. 27–50.
Schwartz, B., Sprouse, R., 1994. Word order and nominative case in nonnative language acquisition: a longitudinal study of (L1 Turkish) German interlanguage. In: Hoekstra, T., Schwartz, B. (Eds.), Language Acquisition Studies in Generative Grammar: Papers in Honor of Kenneth Wexler from the 1991 GLOW Workshops. John Benjamins, Philadelphia, PA, pp. 317–368.
Schwartz, B., Sprouse, R., 1996. L2 cognitive states and the Full Transfer/Full Access model. Second Language Research 12 (1), 40–72.
Selinker, L., 1972. Interlanguage. International Review of Applied Linguistics 10, 209–231.
Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.
White, L., 2003. Second Language Acquisition and Universal Grammar. Cambridge University Press, Cambridge.

Relevant Websites

http://www.aaal.org/ – American Association for Applied Linguistics.
http://www.carla.umn.edu – Center for Advanced Research on Language Acquisition (CARLA), University of Minnesota.
http://www.cal.org – Center for Applied Linguistics.
http://www.aila.info – International Association of Applied Linguistics.
http://linguistlist.org – The Linguist List.
http://homepage.ntlworld.com/vivian.c/SLA/index.htm – Second Language Acquisition Topics by Vivian Cook.
http://www.tesol.org – TESOL International Association.

Second World War, The

Richard JB Bosworth, Jesus College, Oxford, UK; The University of Western Australia (UWA), Crawley, WA, Australia; and Reading University, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

World War II was, arguably, the most significant and influential event of the twentieth century. It was also a conflict with many faces – from the absolute 'barbarization' of the eastern front, to the 'war of a kind' in the Pacific, to something different in many other places. This article reviews the causes, course, and consequences of these multiple wars, paying special attention to the historiographical debates, which, in very many societies, have kept the war alive and meaningful in the generations since 1945. In that sense, it traces the history of a 'war that never ended.'

The Second World War: A Narrative

World War II was an event of massive significance. For at least 50 years after its end in 1945 it continued to condition societies and ideas throughout the world. Much of the politics of the second half of the twentieth century can be read as occurring in an 'after-war' context. The war exacted a death toll of at least 60 million, and probably tens of millions more than that (figures for China and the rest of Asia are mere guesses and the casualty figures on the extent of USSR's sacrifice has risen from 7 to 27 or more million as time has passed, circumstances varied, and the requirements of history altered). A majority of the casualties were civilians, a drastic change from World War I when some 90% of deaths were occasioned at the fronts. Moreover, the invention of the atom bomb during the war and its deployment by the USA at Hiroshima and Nagasaki (6 and 9 August 1945) suggested that, in any future nuclear conflict, civilians would compose 90% or more of the victims. When this apparent knowledge was added to the revelations of Nazi German barbarism on the eastern front, careless half-plans to eliminate 'some 30 million' of the peoples of the USSR and the Nazis' massacre of European Jewry, either in pit killings or when deliberately transported to such death camps as Auschwitz-Birkenau, Treblinka, Sobibor, Chelmno, Belzec, and Majdanek, another casualty of the war seemed to be optimism itself. Certainly at 'Auschwitz' and perhaps at Hiroshima, 'civilization,' the modernity of the Enlightenment, the belief in the perfectibility of humankind, had led not to hope and life but instead to degradation and death.

This more or less open fearfulness, with its automatic resultant linking of a pessimism of the intellect to any optimism of the will among postwar social reformers, may be the grandest generalization that can be made about the meaning of World War II. Big history, however, should not lose sight of microhistory. Actually World War II was fought on many fronts, at different times, for different reasons, and with different effects. In this sense, there was a multiplicity of World War IIs.

In September 1939, a war broke out between Nazi Germany and authoritarian Poland. The liberal democratic leadership of Britain and France intervened saying that they would defend Poland, although in practice they made do with 'phoney war' until the Nazi forces took the military initiative, first in Denmark and Norway, and then in the Low Countries and France in April–May 1940. In June 1940, Fascist Italy entered the war as had been envisaged in the 'Pact of Steel' signed with its Nazi ally in May 1939. War now spread to the Italian empire in North and East Africa and, from October 1940, to the Balkans. Italian forces botched what they had hoped would be a successful *blitzkrieg* against Greece and the effect over the next year was to bring most of the other Balkan states into the conflict, with drastic effect on the populations. Nazi-Fascist occupation was brutal and, moreover, often these fragile states dissolved into multiple civil wars, the most multifaceted and of one of the most lasting significance began in Yugoslavia from March to April 1941.

On 22 June 1941, the Nazi Germans invaded the Soviet Union, commencing what was, in some eyes, the 'real' World War II, and certainly the one that was inspired by the most direct ideological impulse, which unleashed the most horrendous brutality and which still remains, overtly or covertly the most disputed in its meaning. In the course of the campaign in the east it is estimated that the Germans sacked 1710 towns and 70 000 villages. During the epic siege of Leningrad from 1941 to 1944, a million of the city's inhabitants starved to death. In their invasion, the Germans were joined by an assortment of anti-communist allies and friends, including military forces from authoritarian Romania and Fascist Italy. Many Lithuanians, Latvians, and Estonians, and quite a few anti-Soviet elements within the USSR (Ukrainian nationalists, people from the Caucasus, and others) acted as auxiliaries of Nazi power. The Nazis were even embarrassed by a 'Russian' army under General A.A. Vlasov, willing to fight on their side against Stalin and his system. Volunteers also came from pro-Fascist circles in France, and from Spain and Portugal, states ruled by clerical and reactionary dictators who hated communists but were not fully reconciled to the radical racist thrust of much of the Nazi-Fascist rhetoric and some of the Nazi-Fascist policies.

On 7 December 1941, the war widened again when the Japanese air force attacked the American Pacific fleet at anchor at Pearl Harbor in Hawaii. In the following weeks, the Japanese army and navy thrust south and east, dislodging the British from Singapore by 15 January 1942. They went on to seize the Philippines and Dutch East Indies (later Indonesia) and were within striking distance of Australia before being checked at the Battle of the Coral Sea in June 1942. They simultaneously

continued the terrible campaign that they had been waging in China since 1937 (or rather since 1931, when they had attacked Manchukuo or Manchuria). In their special wars, the militarist Japanese leadership tried to throw off what they called the imperialist yoke of US capitalism and the older 'white' metropolitan empires. The purity of their antiimperial motives was damaged, however, by their own commitment to empire and by their merciless killing of Asians. Moreover, as in Europe, their invasions often touched off varieties of civil war provoked by the highly complex stratification of society in the region. At the forefront of such campaigns were often local nationalists who imagined communities subservient neither to European powers nor the Japanese.

On 11 December 1941, Germany and Italy, loyal to the terms of the anti-Comintern pact, had also declared war on the USA, somewhat ironically so, given that Japan, checked by military defeats in an unofficial war against the USSR at Khalkin Gol and Nomonhan in 1938–39, had now engaged with other enemies. The Italians would find out the implications in North Africa, the Germans after the Allied invasion of France on 'D-Day' (6 June 1944), as well as in Italy from 8 September 1943 (the Fascist dictator, Benito Mussolini, overthrown on July 25 in a curiously untotalitarian cabinet crisis, was thereafter restored as a German puppet in northern Italy; allied forces moved slowly up the peninsula from the south, liberating Rome on 4 June 1944 but only reaching Milan at the end of the war in late April 1945). Of the participants in the war, the USA, once fully mobilized, possessed the biggest and most productive economy, and was therefore of crucial importance in the eventual defeat of the anti-Comintern powers. The campaign that the Americans fought with the most intense passion, and with an evident racism of their own, was the war against Japan. In another sense, the USA had a relatively soft war, not disputed on its own territory and not requiring the sort of physical or spiritual sacrifice obligatory from most other combatants. The USA's special World War II was not a visceral one, except, perhaps, in its later imagination.

If the war was fought in many different ways, it is equally true that the variety of conflicts did not all end at the same time and in the same way. The Nazi armies surrendered on 8 May 1945, the Japanese on August 15. But matters were more complicated than that. France's special World War II would commemorate 'victory' from the date of the liberation of Paris on 25 August 1944 (and General Charles De Gaulle would have the effrontery to proclaim that Paris and France had liberated themselves). In most of Nazi-/Fascist-occupied Europe and in parts of Asia, partisan movements had never altogether accepted defeat. Communists were invariably prominent in such resistance, even if quite a few still envisaged themselves as fighting as much for the Soviet revolution as for the liberty of their own nation-state. Every successive 'liberation,' in Europe most frequently coming after the military victory of the Red Army, and in the Pacific that of the USA, had its own special character. Yugoslavia and China were two especially complicated places where the resistance was very strong but where it was contested, not only by the Nazi-Fascists and the Japanese but also by local anticommunist and nationalist or particularist forces. The effects and memory of their wars were by definition to be very different from such societies as in the USA, Australia, and the UK, which did not endure foreign occupation, though the last, with its severe experience of bombing, was itself different from the other two.

In sum, World War II was not just an enormously influential event but also an extraordinarily complicated one. Its complexity has in turn stimulated many passionate and long-lasting debates about its historical meaning.

The Causes of War

For many years it was customary to argue that World War II as compared with World War I had a simple cause. This was 'Hitler's war.' Mainstream analysis continues to ascribe to the Nazi dictator great responsibility for the invasion of Poland and for the spreading of the war thereafter, and especially for the launching of Operation Barbarossa against the USSR. To be sure, from the 1960s, the course of historiography, especially as exemplified in the rise of social history, itself eventually destined to be ousted by 'cultural history' during the rise of neoliberalism to its current hegemony, for a time did not favor a 'Great Man' view of the past and tended to urge that even dictators had limits to their free will. The English historian A.J.P. Taylor (1964) argued in a book entitled *The Origins of the Second World War* that the several crises which led up to September 1939 and what he sardonically called the 'War for Danzig' needed to be understood in a variety of contexts, including the peace settlements at the end of World War I, the course of German political and social history, the institutionalization of the Russian revolution, with its victorious but feared and paranoid communist and then Stalinist regime, and the lights and shadows of democratic liberalism in Western Europe.

Taylor wrote with a flaunted stylistic brilliance and practiced a brittle historical cleverness. He was frequently misunderstood and often gloried in the misunderstanding. His book thus produced an enormous and long-lasting controversy, still the greatest of many prompted by attempts to define the meaning of World War II. At the same time, Taylor's idiosyncrasies ensured that his work could easily be dismissed by those mainstream historians who cheered the weight of their commitment to Rankean principles.

Nonetheless the issues raised by Taylor did not go away. In West Germany, the so-called Fischer controversy, sparked by the Hamburg liberal historian Fritz Fischer's Griff nach der Weltmacht, a massively documented study of German aims during World War I and also published in 1961, raged through the decade and is still an object of historiographical debate. Although Fischer, then and thereafter, wrote almost exclusively about World War I and about imperial Germany, he was read as commenting on World War II and, indeed, on Germany's divided fate in its aftermath. Four issues were prominent. Was imperial Germany an aggressive power in a way that bore comparison with the Nazi regime during the 1930s? Was the motive of the German leadership as much domestic as foreign – did they seek foreign adventure and even world war in order to divert the pressure building for social democracy? Was the appalling conflict from 1939 to 1945 caused by the 'German problem,' which began as early as 1848 and continued after 1945? And was the real story of a European civil war lasting from 1914 to 1945?

Fischer's work was more directly influential than Taylor's and the issue of the relationship between *Innenpolitik* and *Aussenpolitik* fitted neatly into the preoccupations of new social historians who, by the 1970s, were scornfully dismissing the work of 'old fashioned diplomatic historians.' For all that, specialist works on the causation of World War II continued to privilege the power of Hitler and acknowledge the ideological thrust of Nazism. The English independent Marxist, Tim Mason, may have tried to pursue a Fischerian line in asking how much the diplomatic crises of 1938–39 were prompted by the contradictions of Nazi economic and social policy, but his essays remained at the periphery of most analysis. Rather such firmly Rankean historians as Gerhard Weinberg and Donald Watt assembled evidence which, in their eyes, confirmed that the war was caused by Hitler. The most authoritative English language biographer of the *Führer*, Ian Kershaw, despite his background in social history, has not disagreed. Hitler may have been erratic as an executive. Nazi totalitarian state and society, contrary to its propaganda about militant efficiency and a people cheerfully bound into a *Volksgemeinschaft*, may in practice have often been ramshackle. But the dictator, Kershaw argued, did possess power. Indeed, so all-embracing was his will that Germans strove to 'work toward' their *Führer*, to accept his ideas and implement his policies before he had fully formulated them. For a generation in the wake of the Fischer controversy, scholarship on Nazism separated into 'intentionalists' (advocates of Great Man history) and 'functionalists' (those who preferred to emphasize the role of structures and contexts and who were especially alert to the 'institutional darwinism' of the Nazi regime). By the late 1990s and after the fall of the USSR and the cementing of neoliberal hegemony, Kershaw, often a historian of the golden mean, found a way to resolve and end that conflict. It still holds.

The only major variants on this reaffirmation that Hitler had provoked the war came from certain conservative viewpoints. Some anticommunist historians focused on the Ribbentrop–Molotov pact (23 August 1939), placing responsibility for the war on 'Stalin' and the Russian revolution in what seemed to others a highly tendentious effort to blame the victim. More common was the view expressed most succinctly by Zionist historian Lucy Dawidowicz that the whole conflict was in essence a 'war against the Jews.' Hitler (or, in the variant made notorious by Daniel Goldhagen, the Germans) wanted to kill Jews; that was the purpose of the Nazi regime; that was the aim of its wars.

In other ex-combatant societies, further examples of local focus are evident. In England, fans of appeasement still existed, the most prominent being John Charmley. For him, the problem with the war lay simply in Britain's engagement in it. As far as the British Empire was concerned, Nazism did not need to be fought and the USSR should never have been an ally. Worst was the fact that the USA had dominated the postwar world. Charmley has never quite said so, but implication of his work is that the 'real' World War II for Britain, and the one it most dramatically lost, was implicitly fought against its American 'cousins.'

The Asian-Pacific conflict has similarly been subject to historical revision. In the 1960s Gabriel Kolko and other American 'new leftists' applied a Marxian model to their nation's foreign policy, critical of the gap between its idealistic and liberal theory and its realist and capitalist practice. They were in their turn subjected to withering fire from more patriotic and traditional historians. Nonetheless a consensus grew that, at least in regard to the onset of the American–Japanese war, US policy makers carried some responsibility. By the 1980s, liberal historian John Dower was even urging that the two rivals had been 'enemies of a kind,' neither of them guiltless of racism and brutality.

In Japan, by contrast, an officially sponsored silence long hung over everything to do with the war. The Ministry of Education was particularly anxious that schoolchildren not be exposed to worrying facts about such terrible events as the massacre in Nanking in 1937, the practice of germ warfare, the exploitation of 'comfort women,' and the other examples of Japanese murder, rape, and pillage in Asia and the Pacific. Nonetheless, a stubborn undercurrent of opinion exemplified in the work of historian Ienaga Saburō continued to contest the Ministry line and, by now, the Japanese leadership has gone further than ever before in admitting some of the misdeeds of its militarist predecessors. Fully critical history may still not be especially appreciated in Tokyo. But, Japan is not the only society to behave that way in a world where neoliberalism has achieved unparalleled hegemony backed by what American democratic historian Peter Novick has called 'bumper sticker' lessons from the past.

The Course of the War

In the preceding sections, it has not always been possible to keep fully separate discussions about the coming of World War II from what happened after hostilities commenced. In any case, military history in the pure sense, just like diplomatic history, after 1945 soon lost ground professionally. To most eyes, the military history of the war can swiftly enough be told. Of the anti-Comintern states, Germany and Japan but not Italy, won rapid initial victories, exulting in their respective *blitzkriegs*. But their triumphs were always brittle. Germany and its allies may have been good at launching wars, but their ideologies made it difficult for them thereafter to contemplate any policy except the complete liquidation of the enemy. Indeed having won their present battles, they often aimed at still further wars. Historian Adam Tooze (2006) argued strongly that the Nazis never abandoned a dream, extending beyond Europe, of a final confrontation with the USA. Hitler and the Japanese imperial and military leadership thus made no attempts to offer compromise from a position of strength, and Mussolini's occasional flirtation with the idea always involved Nazi sacrifice in their wars, especially that in the east, rather than Italian loss.

Nor did the anti-Comintern states make the most of the huge territories, which they had conquered, and the immense material resources that they therefore controlled. Nazi Germany is something of a case study in this regard. In the West, where the war was always gentler, the Nazis found plenty of direct and indirect collaborators. They were, for example, able to harness a considerable proportion of the French economy to their cause. They also started to construct a new economic order that was not utterly unlike some of the developments, which would occur in Western Europe after

Nazi-Fascism had been defeated. With extraordinary contradiction for a state built on an utter commitment to racial purity, Nazi Germany, already before 1939, needed immigrants to staff its economy. Once the war began, this requirement became more pressing. One partial solution was to import workers by agreement with its ally Italy and by arrangement with the friendly Vichy regime in France. Not all such French 'guest-workers' came unwillingly and not all had especially bad wars. In Germany they joined other, more reluctant, immigrants from the east, who were often little more than slave laborers. Poles and Soviet prisoners of war constituted the majority of these; at first they were frequently worked to death. However, as the Nazi armies turned back and the war settled into one of attrition and retreat, as symbolized by the great defeat at Stalingrad (November 1942–January 1943), the Germans began to treat even laborers from the east in a way that allowed some minimum chance of their survival and which also permitted some tolerable productivity from their labor.

The exception was, of course, the Jews, who, from September to October 1941, became the objects of the 'Final Solution,' a devotion to murder confirmed by officials who attended the Wannsee conference in January 1942. In terms of fighting a war, the adoption of the policy of extermination was, of course, counterproductive in many senses, among which was the economic. In staffing and fueling the trains, which transported the Jews to the death camps, in the low productivity of these and the other camps, the Nazis wasted resources needed at the front. Saul Friedlander will write elsewhere in this volume about the debates about the meaning of the Holocaust. It is worth noting in this segment, however, that each combatant society has argued about its particular experience of being visited by the Nazi war machine and therefore of being exposed to collaboration with it. Two important examples occurred in France and the USSR.

The spring of 1940 brought disaster to the Third French Republic. Now was the time of what Marc Bloch, one of the founders of the great structuralist historical school of the Annales, a patriotic Frenchman and a Jew, called the 'strange defeat.' Military historians have demonstrated that France was not especially inferior in armament to the invading Germans. Rather, France lost for reasons to do with morale and the domestic divisions of French society. As a result, by June 1940 the French state and empire collapsed. During the next 4 years, the inheritance of the Third Republic was disputed between the Vichy regime headed by Marshal Philippe Pétain within the rump of French metropolitan territories; the 'Free French' under General Charles De Gaulle, resident in London, and there, however reluctantly and churlishly, dependent on Allied goodwill and finance; and a partisan movement which gradually became more active in the occupied zones. This last was typically divided between communists and other forces, some of whom disliked communists as much as they hated the invaders. The years of Axis occupation were thus also the time of the 'Franco-French civil war,' with killings and purge trials extending well beyond liberation.

The meaning of war, occupation, and liberation in France was long disputed after 1945. It took a film maker, Marcel Ophuls in Le Chagrin et la pitié (1971), and an American historian, Robert Paxton, to break a generation of silence about the troubling implications of this period of national history, although Socialist President François Mitterrand (1981–95), with his own equivocal experience of Vichy, was scarcely an unalloyed advocate of openness during his term in office. Historian Henry Rousso has brilliantly examined the 'Vichy syndrome' and done much to expose some of the obfuscations favored by many different leadership groups in post-1945 French politics. Perhaps his work does not go far enough, however. With its fall, and then with decolonization after 1945, France had lost its political empire, promising to become just another European state. However, this loss was curiously compensated by the rise and affirmation of French culture. In almost every area of the humanities, such French intellectuals as De Beauvoir, Braudel, Barthes, Foucault, Lévi-Strauss, Lyotard, Baudrillard, and Nora, charted the way to post-modernity. They did so sometimes invoking the pro-Nazi philosopher Martin Heidegger and almost always without much reckoning of the collapse of the French nation-state in 1940 and often with its denial. The intellectuals of France were given to forgetting their nation's fall, the better to affirm their own rights to cultural imperium.

If France provides a case study of the transmutation of defeat into victory, the USSR offers the reverse example of a victor whose peoples would eventually learn that 'actually' they had lost. The nature of the Soviet war effort is still in need of research. Sheila Fitzpatrick, the key social historian of her generation, depicted a population by 1939 brutalized and depressed by the tyranny, incompetence, and contradictions of Stalinism, if also retaining agency of their own despite the dictatorship. Her work does not really explain, however, how that same populace fought so stubbornly 'for the motherland, for Stalin.' No doubt the absurd murderousness of Nazi policies gave them little alternative. No doubt the aid that eventually flowed from the USA was of great significance. But something does remain unexplained about how 'Stalin's Russia' won its 'Great Patriotic War.'

The Consequences of the War

By now very clear, however, are the consequences of the war for the USSR. In the short term the war made the Soviet state the second superpower, the global rival to the USA, and gave Stalin, until his death in 1953, an almost deified status. It also ensured, in the West, the rise of the theory of totalitarianism, which married communism and Nazi-Fascism as the common enemies of mankind. Initially a fruitful intellectual underpinning of the Cold War, this view of Europe's history from 1914 to 1945 and beyond has never been fully abandoned by its partisans, as has been demonstrated most graphically in recent times by Timothy Snyder's powerful but controversial book, Bloodlands. More generally, as the postwar decades passed, it became clear that the USSR and its expanded sphere of influence in Eastern Europe were not recovering from the war with the speed being dramatically exemplified in Western Europe and Japan. Indeed, the history of the USSR, at least until Gorbachev's accession to the party secretaryship in 1985 and, arguably, until the fall of the Berlin Wall and the

collapse of communism (1989–91), should best be read as that of a generation who had fought and won its wars (including the terrible domestic campaigns for collectivization as well as the purges of the 1930s). Leaders and people were unwilling or unable to move beyond that visceral experience. After 1945, the USSR became the archetypal place 'where old soldiers did not die nor even fade away.' Brezhnev, Chernenko, and their many imitators further down the power structure were frozen into a past that had an ever-diminishing connection with a world facing a new technological and economic revolution. So powerful was the war myth that it transferred cheerfully to Putin's Russia (and, in different form, to the various Soviet successor states), with communism obscured, nationalism enhanced (and with Stalin by no means renounced).

Other ex-combatant societies were more open to change than the USSR but a certain sort of remembering and a certain sort of forgetting can readily enough be located in them, too. Memory proved most threatening in Yugoslavia. There the victorious partisans under Josip Broz Tito seemed for a time to have won a worthwhile victory. The special history of their campaigns against the Nazi-Fascist occupiers of their country allowed them claims to independence from the USSR, which they duly exercised after 1948. At the same time the barbarity, during the war, of the collaborationist Croat-Fascist regime under Ante Pavelić, whose savagery embarrassed even the Germans, seemed to suggest that the region was indeed best administered by a unitary state. The fall of communism, however, also brought down a communist Yugoslavia in which such Serb leaders as Slobodan Milošević, unable to believe in official ideals, increasingly recalled the nationalism of wartime Cetniks rather than the internationalist Marxism once espoused by the partisans. This memory of war and murder justified new battles and new murders, even if with the somewhat ironical result that, in the Balkan wars of the 1990s, the (Serb) winners of the last conflict had become losers and the (Slovene, Croat, Bosnian Moslem, and Kosovar) losers had become winners.

In Italy, the country with the largest communist party in the West and a polity, which had a 'border idiosyncrasy,' bearing some comparison with Yugoslavia's role in the Eastern Bloc, the inheritance of war and fascism similarly possessed peculiar features. Postwar Italy began by renouncing fascism, empire, and war, in 1946 abandoning the monarchy that had tolerated the imposition of Mussolini's dictatorship and, in 1947, adopting a constitution which made considerable claim that the new Republic would be based on labor. In practice, however, Italy took its place in the Cold War West. Its purging of ex-Fascists was soon abandoned and both the power elites and the legislative base of the new regime exhibited much continuity with their Fascist predecessors. Nonetheless, from the 1960s, an ideology of antifascism was accorded more prominence in a liberalizing society. From 1978 to 1985, Sandro Pertini, an independent socialist who had spent many years in a Fascist jail and been personally involved in the decision to execute Mussolini, became Italy's president. Widely popular, he seemed an embodiment of the national rejection of the Fascist past.

Once again, however, the process of memory was taking a turn and a different useable past was beginning to emerge. Left terrorists in the 1970s had called themselves the new Resistance and declared that they were fighting a Fascist-style state – the governing Christian Democrats were thought to be merely a mask behind which lurked the Fascist beast. The murder of Aldo Moro in 1978 drove Italians decisively away from this sort of rhetoric and, from then on, Italians sought instead a 'pacification' with the past where ex-Fascists had as much right to be heard as ex-partisans. Among historians, 'anti-anti-Fascists,' led by Renzo De Felice, the biographer of Mussolini, provided evidence and moral justification for this cause. Media magnate and conservative politician Silvio Berlusconi joined those who agreed that an anti-Fascist World War II had lost its ethical charge, stating (wrongly) that 'Mussolini never killed anyone.'

Among the Western European ex-combatant states perhaps the UK was the place where the official myth of the war survived with least challenge, despite the huge casualty total in parts of the empire during the war, for example in the Bengal famine of 1943, for which at a minimum British incompetence must share some blame. A vast span of postwar British society and behavior was influenced by Britain's war. The Welfare State, as codified by the postwar Labor government under Clement Attlee, was explained and justified as a reward for the effort of the British people in the 'people's war.' Wartime conservative Prime Minister, Winston Churchill, despite his many evident limitations, remains a national icon. A royal wedding in 2011 had to be accompanied by a flyover from World War II planes. From the alleged acuteness of their secret service activities to the alleged idealism of their rejection of Nazism, the British have constantly sought to preserve the lion's share of the positives of World War II for themselves, almost as if it is their national version of a sacred holocaust. The suspicion toward the euro and much of the European Union and the many other examples of continuing British insularity often find historical justification in the view that the British fought alone against the Nazi-Fascists from 1940 to 1941, and express the associated annoyed perplexity that somehow their wartime sacrifice has entailed a slower passage to prosperity compared with that of their continental neighbors (and especially Germany).

Memory after memory, history after history, World War IIs, in their appalling plenitude, still eddy around. In 1999, for example, another historian wrote a major book about the meaning of an aspect of the war, and about the construction of that meaning. It is a work still of major importance. In it, Peter Novick daringly wondered whether the privileging of the Nazi killing of the Jews in contemporary Jewish and even gentile American discourse is altogether a positive. Being a historical victim at one time in the past, he argued cogently, can obscure as well as explain. His caution was timely. It is probably good that World War IIs are with us still; it will be better if the interpretation of so many drastic events can still occasion democratic debate, courteous, passionate, and humble debate, and if we can therefore avoid possessing a final solution to its many problems.

See also: Cold War, The; Communism, History of; Conflict and War: Anthropological Aspects; Contemporary History; First World War, The; Genocide, Historical Aspects of; Holocaust, The; Military History; National Socialism and Fascism; War Crimes Tribunals, Crimes against Humanity and Transitional Justice; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Bartov, O., 1986. The Eastern Front 1941–5: German Troops and the Barbarisation of Warfare. St. Martin's Press, New York.

Bessel, Richard, 2009. Germany 1945 from War to Peace. Simon and Schuster, New York.

Bloxham, D., 2009. The Final Solution: A Genocide. Oxford University Press, Oxford.

Charmley, J., 1993. Churchill: The End of Glory—A Political Biography. Hodder and Stoughton, London.

Collingham, L., 2011. The Taste of War: World War Two and the Battle for Food. Allen Lane, London.

Connelly, M., 2004. We Can Take It London: Britain and the Memory or the First World War. Longman/Pearson, London.

Dawidowicz, L., 1986. The War against the Jews 1933–45, rev. ed. Penguin, Harmondsworth, UK.

Dower, J.W., 1986. War without Mercy: Race and Power in the Pacific War. Pantheon, New York.

Edele, M., 2011. Stalinist Society 1928–1953. Oxford University Press, Oxford.

Evans, R.J., 2008. The Third Reich at War: How the Nazis Led Germany from Conquest to Disaster. Allen Lane, London.

Finney, P., 2010. Remembering the Road to World War II. Routledge, London.

Fischer, F., 1967. Germany's Aims in the First World War. Chatto and Windus, London.

Fischer, F., 1986. From Kaiserreich to Third Reich: Elements of Continuity in German History, 1871–1945. Allen and Unwin, London.

Gorodetsky, G., 1999. Grand Delusion: Stalin and the German Invasion of Russia. Yale University Press, New Haven, CT.

Goto, K., Kratoska, P., 2003. Tensions of Empire: Japan and Southeast Asia in the Colonial and Postcolonial World. Ohio University Press, Athens, OH.

Hitchcock, W.I., 2009. Liberation: The Bitter Road to Freedom, Europe 1944–1945. Faber and Faber, London.

Hogan, M.J. (Ed.), 1996. Hiroshima in History and Memory. Cambridge University Press, Cambridge, UK.

Ienaga, S., 1979. Japan's Last War: World War II and the Japanese 1931–1945. Australia University Press, Canberra, Australia.

Kershaw, I., 1999. Hitler 1889–1936: Hubris. W. W. Norton, New York.

Kershaw, I., 2000. Hitler 1936–1945: Nemesis. Allen Lane, London.

Kolko, G., 1968. The Politics of War: The World and United States Foreign Policy 1943–5. Random House, New York.

Mazower, M., 2008. Hitler's Empire: How the Nazis Ruled Europe. Penguin, London.

Milward, A., 1977. War, Economy and Society 1939–1945. University of California Press, Berkeley, CA.

Novick, P., 1999. The Holocaust in American Life. Houghton Mifflin, Boston.

Reynolds, D., 2004. In Command of History: Churchill Fighting and Writing the Second World War. Random House, New York.

Rodogno, D., 2006. Fascism's European Empire: Italian Occupation during the Second World War. Cambridge University Press, Cambridge.

Rousso, H., 1991. The Vichy Syndrome: History and Memory in France since 1944. Harvard University Press, Cambridge, MA.

Snyder, T., 2010. Bloodlands: Europe between Hitler and Stalin. Basic Books, New York.

Stargardt, N., 2005. Witnesses of War – Children's Lives under the Nazis. Jonathan Cape, London.

Taylor, A.J.P., 1964. The Origins of the Second World War, rev. ed. Penguin, Harmondsworth, UK.

Tooze, A., 2006. The Wages of Destruction: The Making and Breaking of the Nazi Economy. Allen Lane, London.

Tumarkin, N., 1994. The Living and the Dead: The Rise and Fall of the Cult of World War II in Russia. Basic Books, New York.

Watt, D.C., 1989. How War Came: The Immediate Origins of the Second World War 1938–1939. Pantheon Books, New York.

Weinberg, G.L., 1994. A World at Arms: A Global History of World War II. Cambridge University Press, Cambridge, UK.

Weiner, Amir, 2001. Making Sense of War: The Second World War and the Fate of the Bolshevik Revolution. Princeton University Press, Princeton.

Secrecy, Anthropology of

Lamont Lindstrom, University of Tulsa, Tulsa, OK, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Secrecy is deliberate concealment of information. Anthropologists are primarily interested in public, or institutionalized, secrecy: those persistent practices of hiding information within kinship, religious, political, or economic groups including state and corporate secrecy. Areas of research include concealment of ritual knowledge and the social functions of secret societies. Uses of more personal, or private, secrets have also attracted attention as these relate to cultural constructions of personhood and to the dynamics of interpersonal relationships. Some theorists have argued that society itself rests on a necessary interaction of surface and secret cultural systems. The study of secrets has posed obvious methodological and ethical problems for ethnographers.

Secrecy and Culture

A simple definition of secrecy as the deliberate concealment of information has served cross-cultural comparison (e.g., Tefft, 1980; Bok, 1982). Under this definition, people in cultures everywhere have secrets. Since a secret may turn out to be 'empty' – no hidden information in fact exists – some anthropologists have suggested that attention be refocused on the practice or the 'doing' of secrets rather than on a secret's informational content. Anthropologists have been primarily interested in what could be called public, or institutionalized, secrecy: those persistent practices of hiding information within kinship, religious, political, or economic groups. Areas of research have included concealments of ritual knowledge and the social functions of secret societies. More personal, or private, secrets have also attracted attention insofar as these relate to cultural constructions of personhood, identity construction, and the dynamics of interpersonal relationships. The study of secrets and institutionalized secrecy has posed obvious methodological and ethical problems for ethnographers.

Some have argued that secrecy is a fundamental aspect of all cultural systems. If culture is more or less shared knowledge, then the less shared, less public information creates and sustains people's cultural commitments and stokes their faith and emotional attachment to what they take to be important hidden truth. Truth, as locally and collectively accepted, gains power, significance, and attraction insofar as it remains partially hidden and opaque. Cultural worlds are often mysterious and partly obscured to those of us who inhabit them. Georg Simmel (1950 [1908]), who was among the first to remark the sociological significance of secrecy, suggested that secrets are a device of cultural 'adornment' – a jewel or bauble that seduces public attention. If the observable surface of everyday existence did indeed comprise all that there is, life would be more shallow and less interesting and, bored and unattached, one might disinvest in participating in society.

Secrecy, as a practice, is more significant than whatever information itself might be concealed. Furthermore, many analysts of secrecy systems have concluded that many secrets are not actually secret. Women know men's ritual tricks: those spirit voices overheard during religious ceremony are mere bamboo flutes or bullroarers. Children pretend not to know that their parents are the real Santa Claus. Kin from one lineage are familiar with the supposedly secret names, story archives, and genealogies of their neighbors. In oral societies, the leakiness of secret knowledge, in fact, helps maintain its viability within a community. If a secret holder dies, others are around to reveal lost knowledge if need be.

Suspicion of some truer reality, which lies hidden under the surface of life and explaining that life, is widespread (Ferme, 2001:4). What one sees is not all there is. In this argument, the canny and uncanny attractions of culture itself depend on a punctuated relationship between concealment and revelation. Michael Taussig, drawing on Elias Canetti and Michel Foucault, has argued that periodic revelation – the 'unmasking' of secrets – functions to burnish and empower society as mysterious and sacred: "For are not shared secrets the basis of our social institutions, the workplace, the market, the family, and the state?" (Taussig, 1999:2). Revelation, rather than draining cultural energy by unmasking concealed information, in fact, instead may 'animate the thing defaced and the mystery revealed may become more mysterious, indicating the curious magic upon which Enlightenment, in its elimination of magic, depends' (1999:3–4). Human culture is always partly secretive – recognized but not discussed. Such 'public' secrecy is 'that which is generally known, but cannot be articulated' (Taussig, 1999:5). It is this public willingness to accept and protect communal secrets, as in the story of the Emperor's new clothes, that maintains ruling cultural systems as well as claims to power and relations of inequality.

More specifically, ethnographers have focused on ceremonial masking, wrapping, and clothing practices that symbolize, enact, and demonstrate the distinction between manifest and secret truths. Spectators may or may not know the identity of the person obscured by a mask, but either way masking depersonalizes ritual actors who may thus become avatars of spirits and gods (De Jong, 2007:128). Ordinarily hidden reality rises to the surface insofar as the mask opens a door between seen and unseen. Wrapping objects in cloth or paper, similarly, creates a distinction between outside and inside, thus reflecting a fundamental opposition between surface and secret knowledge. Tattooing and clothing the human body similarly mark the boundary between interiors and exteriors. Gell, for example, suggests that Polynesian tattooing functioned as 'the exteriorization of the interior which is simultaneously the interiorization of the exterior'

(1993:39). Masking, wrapping, clothing, and tattooing, all call attention to the existence and interdependent relations between surfaces and interiors of both the person and society at large.

In addition to public cultural secrets, postponements of revelation also lubricate everyday social relationships. In these relationships, people tack between useful, ambiguous ignorance of each other's positions and desires so that the details of relational inequalities or disagreements need not be admitted or totted up and occasions when all is revealed (Piot, 1993). Secrecy here sets 'in motion a process—of interpretation, ambiguity, and the quest for hierarchy' and keeps this process in play (1993:362). Revelation, on the other hand, creates new relational and informational equilibria and starts the game anew.

Reading secrecy as an integral and fundamental aspect of culture itself parallels developments in queer theoretical analyses of literature, notably Eve Sedgwick's *Epistemology of the Closet* (1991), which locate unarticulated public secrets about sexuality hidden beneath the surface of the Western literary canon. Secrecy, likewise, is the dark shadow of culture itself, intertwined through all human affairs. Most anthropological theories of secrecy, however, are not so grand. They instead have focused on more particular practices and institutions of secrecy including secret organizations and societies and hidden aspects of personality.

The Paradox of Secrecy

The revelation of secrets is as important as the keeping of them. Ethnographers have called this the 'paradox' of secrecy (Bellman, 1984). For a secret to persist within a social order, it must eventually be told to someone who may in turn keep the secret until it passes along again. Information kept absolutely secret, and never revealed, eventually drops out of cultural orders. The social consequence of public secrets rests on the fact that people know of their existence even though content may remain hidden. Secrets with social consequence thus must be at least partly transparent and also periodically transacted and revealed.

Systems of secrecy often rely on inequalities in rights to reveal knowledge rather than on an effective concealment of information – rights that the Kpelle summarize in the term 'you cannot talk it' (Bellman, 1984). Folk copyrights of this sort regulate who may speak about what (Lindstrom, 1990). Even though many may know the content of concealed information, these unacknowledged others may not legitimately repeat and reveal this knowledge in public without the copyright to do so. Family groups in Vanuatu, for example, own songs, myths, genealogies, and medical recipes that others may not publicly reveal without that family's permission. In private contexts, however, public secrets are often surreptitiously discussed. Knowledge is thus regulated not only by restricted transmission – by secrecy – but also by copyrights that limit who speaks publicly about this. Folk systems of copyright transform secrets into public secrets. Those supposed not to know must pretend not to know, and those supposed to know pretend to think that only they, in fact, do know.

Public Secrecy

Public awareness of the existence of secrets can afford prestige and power to those who know what others do not. Many accounts of secrecy revisit Georg Simmel's ground-breaking analysis of secrecy and secret societies (1950 [1908]), including his observation that secrets are a form of 'inner property.' Simmel's commoditization of hidden knowledge opened the door to a political economy of secrecy. If secrets are property, then they are produced, exchanged, and consumed. Secrets are public because they have exchange value within a knowledge marketplace, and because such exchange has political consequence. Systems of political hierarchy may be grounded in an unequal distribution of secrets. Those who know may dominate those who do not know so long as the ignorant grant that hidden knowledge has value.

Anthropologists have used a political economy of secrecy to account for various forms of inequality. James Mooney, in 1887, revealed the 'sacred formulas' of Cherokee shaman and healers with which they built both professional pride and practice. Twenty years later, Franz Boas published details about northwest coast American secret societies, as did Edward Evans-Pritchard in Africa who infiltrated Azande secret magical associations (González, 2012). Much subsequent ethnographic attention has focused on West African secret societies as political institutions that maintain inequalities among elders and youth and men and women. Piot discriminates four predominantly functionalist anthropological approaches to African secrecy: the structural-functionalist, which concludes that secrecy maintains political unities and initiation societies; the Marxist, which argues that secrecy serves to empower elders over youth; the Freudian, which reads secrecy partly as a projection of infantile sexual anxiety; and the semiotic, which investigates the function of secrets within communicative systems (1993:354). All four approaches presume clear connections between secrets and power, although secrecy and secret organizations have sometimes maintained ruling political institutions while at other times have undermined these.

Secrecy and Power

In the latter years of the twentieth century, anthropological attention returned notably to issues of power and inequality. Neutral definitions of culture as more or less shared knowledge gave way to new concerns with the contradictions, gaps, variation, and disparity in that knowledge – with secrets, in fact. Secret knowledge and secret societies might not only have useful social functions, but also maintain ruling political and economic regimes. The distribution of access to public secrets typically parallels that of other rights and powers. Adults hide information from children, often until the latter have been ritually initiated into adulthood. Men refuse to share ritual knowledge with women. Family and lineage groups own various sorts of genealogical, medical, or historical knowledge, sharing these only among kin. Guilds and professions jealously guard business secrets and establish apprenticeships and other credentialing barriers to their proprietary technologies and other economic knowledge.

Conversely, secrecy can be a device to resist power. Or, in places where everyone has secrets – where secrecy is democratized – enduring political inequalities may be more difficult to establish insofar as 'the entire social order remains contingent, and the possibility of power, which is partly rooted in claims of privileged access to certainty, is greatly diminished' (Petersen, 1993:334). Secrecy protects ruling authority and those who resist this alike. Where hierarchy exists, the dominated conceal much of what they know from authority. Women preserve secrets they keep from men. Slaves commune in secret languages. Children construct hidden worlds that evade adult awareness. Secrecy here is a 'weapon of the weak,' which functions to resist and deflect supervision from above. It protects the lowly and the oppressed who must keep their heads down, be these wives, slaves, children, peasants, employees, and so forth. Secrecy can also protect adherents of heterodox religions, preserving their idiosyncratic beliefs and practices from public deprecation, as once was the case, for example, with middle-class English witches joined together in Wicca covens (Luhrmann, 1989).

Cants and other sorts of restricted linguistic codes also serve to restrict information flow. These include the professional jargons and argots spoken, for example, by doctors, lawyers, and thieves, and also a variety of youthful linguistic styles that elders may fail to grasp. Some traditional Bolivian healers speak Kallawaya, a secret trade language that combines Quechua grammatical forms and an esoteric vocabulary perhaps drawn from the now extinct Pukina language. Such cants preserve trade secrets and also obscure and conceal the everyday dealings of those who use them from the uninitiated. Arcane sacerdotal languages such as Latin and Sanskrit similarly create boundaries between insiders and outsiders, and may also usefully represent the mysteriously sacred voice of supernatural gods and spirits. In Melanesia, for example, ancestral spirit communiqués are often linguistically opaque to many insofar as only the adept can understand the secret language of spirits (Lindstrom, 1990:121).

Secret Societies

Anthropologists have pursued Simmel's original concern with secret societies, documenting a variety of secret groups, cults, associations, lodges, and clubs in cultures around the world the members of which are pledged to maintain secrets, ritual, and otherwise. The common comparative ethnographic category 'secret society' is catholic enough to encompass Melanesian men's houses and grade societies, Australian totemic cults, Native American medicine lodges, West African male and female associations, Western men's social clubs, and more (Webster, 1908).

Liberal theories of democratic process and of the capitalist marketplace have been suspicious of secrecy as they are of cabals and monopolies. Similarly, classic psychological models of personality and family dynamics have deprecated secrets and promoted personal revelation and storytelling as a form of healing. One should reveal hidden personal truth in order to 'shed light on the dark folds, to open the doors to the closets of secrecy, to engage the power of story as a way to penetrate – and perhaps lift ourselves out of – the darkness of despair' (Poulos, 2009:25). Just as persons and families cleansed of dark secrets are presumed to be healthier, so does democracy requires an informed citizenry. The capitalist marketplace, too, is supposed to depend on the free flow of economic information among buyers and sellers. From this perspective, secrecy generally goes against the public good. Secret societies have frequently aroused public suspicion about their operations and goals, suspicion that has frequently led to legal and sometimes violent suppression of suspect groups. Those left in the dark have often been outraged by and mistrustful of secret organizations: "For about a century and a half large numbers of intelligent Europeans believed that much of what was happening in the world around them only happened because secret societies planned it so" (Roberts, 1972:1–2).

Although often mistaken, conspiracists on the lookout for hidden powers and forces have at least occasionally been correct in their suspicions. Some secret organizations have indeed aimed to topple established authority, just as witches, sorcerers, and shaman may actually, and secretly, attempt to kill others (Whitehead and Wright, 2004). In recent years, some West African secret societies, which once protected village political interests, have turned against the state in support of one or another rebel movement in Liberia, Sierra Leone, and elsewhere (De Jong, 2007). Earlier, European colonial authority likewise mistrusted secret society operations in West Africa as it did Melanesian 'cargo cults' in the southwest Pacific.

There are more ethnographic accounts of establishment rather than revolutionary secret societies. Structural-functionalist analysis suggests that traditional secret societies in Africa and elsewhere commonly serve important social functions, including the education of children, preservation of political authority, punishment of lawbreakers, stimulation of the economy, and the like (Little, 1949). Such 'social' secret societies address important community needs, unlike 'antisocial' secret societies whose goals are criminal or revolutionary (Wedgewood, 1930). For example, the power of Kpelle elders over youth in Liberia is grounded in their management of the secrets of the Poro society (Murphy, 1980). Elders across Sierra Leone and Liberia initiate young men into Poro and young women into a parallel female Sande secret society. Acknowledging the authority of their elders, initiates pay to join these societies and they eventually will themselves become the knowledgeable elite. In addition to accessing revealed knowledge, society initiates also learn adult practices of discretion and dissimulation (Bellman, 1984:51). In postcolonial times, these West African secret societies have also come to modulate local community relations with state authority and with global economic and other cultural forces that increasingly reach into once isolated villages (De Jong, 2007).

Beyond West Africa, anthropologists have argued that elder male keepers of secrets similarly acquired authority over women and uninitiated youth in various societies of Melanesia (Barth, 1975), Native America (particularly Southwestern Pueblo cultures), and Australia (Keen, 1994). Melanesian cult secrecy, for example, typically not only aimed to avoid colonial scrutiny but also functioned as a means 'of understanding the cultural differences between Europeans and Melanesians' insofar as cultists presumed European power also to rely on hidden knowledge (Lattas, 1998:73). Cult ideology suspected that all powerful knowledge – local

and foreign – need be secret. Male appropriation of religious ritual, access to the supernatural, technology, medicine, sacred texts, history, and other valued information cemented their authority over the ignorant. Furthermore, women, children, and others in subordinate political positions sometimes can be forced to reveal what they know, or otherwise denied rights to secrecy. Rights to keep secrets are as consequential as the right to reveal information.

Personal Secrecy

Ethnographic attention to notions and practices of personal secrecy overlaps with legal theories of the person and policy initiatives to protect rights of privacy that the US Supreme Court delineated beginning in the 1920s. Concepts of private secrets reflect cultural assumptions about personhood. Persons, too, comprise a public face or surface that overlies hidden truths. To have secrets, a person must encompass some sort of inner self or heart where private information and opinion are stored along with intentions and capacity to conceal this information. One might imagine a culture that entertains notions of persons with no inner selves that would thus obviate the possibility of private secrets. No such culture exists although Western conceptions of childhood often presume that psychologically immature children lack complete competence with secrecy – that personal aptitude to intentionally conceal information only emerges as part of the child developmental process (Bok, 1982). Literature and film, e.g., *The Invention of Lying*, have sometimes imagined humans without inner selves and private secrets, but if such persons exist today they are taken to be psychologically immature or insane.

Western historians have suggested that there have been different regimes of secrecy in the past, related to shifts in constructions of personhood. Simmel (1950 [1908]) connected the evolution of secrecy to that of individualization and urbane modernity (his argument recalls similar evolutionary accounts of personal privacy). In preurban societies, lack of individualism and everyday intimacies of interpersonal contact within small human communities made secrecy difficult. Simmel supposed that secrets increased with developing opportunities for personal reserve and discretion. More recent historians, stimulated by the work of Michel Foucault (1978), have argued instead that modernity shrinks opportunities for personal secrecy – that bureaucratic power structures increasingly have come to rely upon the monitoring of individuals, either through self-revelation or by institutional and technological devices that extract personal information. According to Foucault, the origins of such extraction trace back to the Christian practice of confession. People are obliged to reveal their secrets to ecclesiastical, juridical, and other authorities, and revealing secrets to authority confirms ones subjugation within a social order.

After the Enlightenment, sincerity and authenticity emerged as predominant modern virtues, displacing previous appreciation of personal ability to mask and to dissimulate (Snyder, 2009). Being authentic to one's own inner truths and maintaining sincerely within one's personal relationships, both can degrade onetime boundaries between public face and private secrets. Some fear such trends have now gone too far, energized and encouraged by contemporary technology. The revelation of private selves on Facebook to vastly greater numbers of spectators, corporate, and otherwise, for example, may threaten what remains of personal privacy. The modern individual, however, still possesses inner capacities to conceal information although these capacities struggle nowadays with contradictory urges and duties to reveal those secrets.

Stimulated by the work of Erving Goffman (1959) and others, students of interpersonal relationships have remarked the significance of secrets, masks, and 'face' within the everyday micropolitics of self-presentation. Such studies have noted social uses of revelation. People strengthen their relationships, creating communities of trust, by revealing secret information. This may be a secret about themselves, or a secret about another that they pass along, often in return for the promise 'not to tell.' Such secrets are a kind of gossip (Bok, 1982), the exchange of which remarks and maintains sentiments of friendship. American girls, for example, establish and strengthen friendships by sharing secrets. They also end such friendships by spilling these secrets to third parties. Personal secrets are a social currency that people invest in their relationships. Public secrets maintain political value insofar as their transmission is restricted, whereas the value of personal secrets flows from their easy everyday revelation.

Alongside preservation of regimes of power, secrecy also contributes to perceptions of individual identity. Self-understanding may transform after a person has acquired secret information. Boys, for example, come to redefine themselves as men after progressing through an initiation ritual during which they learn adult secrets (Herdt, 2003). One's acquisition of once hidden knowledge marks achievement of adulthood. Those who join secret societies may gradually access more and more esoteric knowledge, and this growing purchase on hidden reality marks their increasing status and authority within a community.

Anthropological analyzes of the personal dynamics of public secrets prefigured work on other similar institutions, including the homosexual 'closet,' a term that has now come to characterize all sorts of personal secrets that are either unknown or ignored by others (Sedgwick, 1990). *Yan Daudu*, for example, men who have sex with men in the Hausa Islamic areas of northern Nigeria and who also embrace aspects of feminine style, manage to operate within their communities insofar as others overlook the quasisecret queer spaces they create (Gaudio, 2009). This is also true of the men, typically married, with whom they enjoy sexual relationships. Such 'do not ask, do not tell' practices and policies serve to maintain multipart persons. The closet, like more general practices of clothing, masking, and wrapping, map out a layered personhood, which comprises a public face floating atop an inner, private self.

State and Corporate Secrecy

Institutional secrecy has accelerated with developing digital technologies that not only store and make available but also conceal information. States, corporations, and other contemporary institutions are often reticent to release information that does not serve their interests. At the same time,

they employ increasingly powerful data mining techniques to access and aggregate personal information. Public debate about limitations and rights to online privacy are currently underway. Although anthropologists have been encouraged to 'study up,' typical ethnographic methods including participant observation and interviewing are difficult to deploy when investigating governmental or business organizations that actively deflect outside attention. Some ethnography of such institutions and their secrecy practices has been undertaken, although this ground has been better covered by investigative journalists. Ethnographers have managed to study notoriously secret organizations including American intelligence agencies, government nuclear and other research laboratories, and the US military's Human Terrain System that was initiated during the Iraq and Afghanistan wars (González, 2012). Ethnographers have also focused on corporations and their internal organization but typically have acquired access only to the less secretive corners of these enterprises such as human resource management, product marketing campaigns, or whatever corporate culture they are willing to acknowledge.

The Study of Secrecy

Anthropology's interest in cross-cultural description as a whole has been taken to be the desire to learn and reveal other people's secrets (Taussig, 1999; De Jong, 2007:189). Anthropological authority rests on the ability to penetrate cultural screens. Methodologically, ethnographers face obvious problems of access to concealed information, but the study of secrecy raises even larger ethical puzzles. Ethnographers who infiltrate cultural systems can sometimes be accused of stealing secrets or being troublesome snitches. Codes of ethics generally require ethnographers to ensure that their research does not harm the safety, dignity, or privacy of the people they study. The 2009 Code of Ethics of the American Anthropological Association, for instance, requires that researchers must not disguise their objectives, their sources of support, or the potential impacts of their projects and, in general, should appropriately use and disseminate their research results. The code, however, recognizes that every case is different and that an ethical duty not to harm the dignity, safety, or privacy of informants may require maintaining confidentiality and their anonymity, as well as sometimes not publishing or otherwise disclosing research findings. Some ethnographers who investigate secrecy have described the structure and function of secrecy systems while avoiding publishing the details of secret knowledge content. Others have promised not to make their publications or photographs available to members of a community (women, often, or the uninitiated) who should not have access to secret information. A few have refrained from publishing at all and restrict access to their field notes for varying lengths of time.

Ethical issues are thorniest where the secrets that anthropologists probe help sustain unjust social orders. Duty to protect the interests of one's informants becomes more difficult when these are in powerful positions or organizations, and spilling the secrets of whom might in fact encourage greater social justice. Some have also noted that current ethical practices of 'informed consent,' wherein subjects must first understand and then consent to be studied, presume the existence of rational actors operating in public arenas in which information is shared and facts are transparent (Mitchell, 1993:24). Others argue that such presumptions of normative transparency and impartial contractual obligations between researchers and their subjects are not supported by more sophisticated theories of cultural secrecy (as above) that take the social world 'to be shot through with misinformation, evasion, lies, and fronts and every level, and research in kind – secret, covert, concealed, and disguised – is necessary and appropriate' (1993:30). Ethnographic attention to the secrecy practices of state and corporate bureaucracies certainly has required some dissimulation insofar as these organizations generally command the ability to evade and repel academic scrutiny. Despite the best efforts of these organizations, some information leaches out via the Internet. Sites such as wikileaks.org collect and collate formerly secret information. Researchers have also deployed Freedom of Information Laws, where these exist, to penetrate state secrets (González, 2012). Like state and corporate bureaucracies, ethnographers too can data mine at least some of the concealed information now digitally circulating around the world.

See also: Anthropological Research, Ethics of; Culture: Contemporary Views; Ethnography in Applied Social Research; Indigenous Knowledge and Intellectual Property Rights; Inequality: Comparative Aspects; Internet and Social Media: Anthropological Aspects; Mafia, Anthropology of; Organizational Culture, Anthropology of; Personhood, Anthropology of; Property: Anthropological Aspects.

Bibliography

Barth, Fredrik, 1975. Ritual and Knowledge among the Baktaman of New Guinea. Yale University Press, New Haven, CT.
Bellman, Beryl L., 1984. The Language of Secrecy: Symbols and Metaphors in Poro Ritual. Rutgers University Press, New Brunswick, NJ.
Bok, Sisella, 1982. Secrets: On the Ethics of Concealment and Revelation. Pantheon Books, New York.
De Jong, Ferdinand, 2007. Masquerades of Modernity: Power and Secrecy in Casamance, Senegal. Indiana University Press, Bloomington, IN.
Ferme, Mariane C., 2001. The Underneath of Things: Violence, History, and the Everyday in Sierra Leone. University of California Press, Berkeley, MA.
Foucault, Michel, 1978. An introduction. In: The History of Sexuality, vol. 1. Pantheon Books, New York.
Gaudio, Rudolf Pell, 2009. Allah Made Us: Sexual Outlaws in an Islamic African City. Wiley-Blackwell, Chichester, UK.
Gell, Alfred, 1993. Wrapping in Images: Tattooing in Polynesia. Clarendon Press, Oxford.
Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
González, Roberto J., 2012. Anthropology and the covert: methodological notes on researching military and intelligence programmes. Anthropology Today 28 (2), 21–25.
Herdt, Gilbert H., 2003. Secrecy and Cultural Reality: Utopian Ideologies of the New Guinea Men's House. University of Michigan Press, Ann Arbor, MI.
Keen, Ian, 1994. Knowledge and Secrecy in an Aboriginal Religion. Oxford University Press, Oxford, UK.
Lattas, Andrew, 1998. Cultures of Secrecy: Reinventing Race in Bush Kaliai Cargo Cults. University of Wisconsin Press, Madison, WI.

Lindstrom, Lamont, 1990. Knowledge and Power in a South Pacific Society. Smithsonian Institution Press, Washington, DC.
Little, K.L., 1949. The role of the secret society in cultural specialization. American Anthropologist 51 (2), 199–212.
Luhrmann, Tanya M., 1989. The magic of secrecy. Ethos 17, 131–165.
Mitchell Jr., Richard G., 1993. Secrecy and Fieldwork. Sage, Newbury Park.
Murphy, William P., 1980. Secret knowledge as property and power in Kpelle society: elders versus youth. Africa 50 (2), 193–207.
Petersen, Glenn, 1993. Kanengamah and Pohnpei politics of concealment. American Anthropologist 95 (2), 334–352.
Piot, Charles D., 1993. Secrecy, ambiguity, and the everyday in Kabre culture. American Anthropologist 95 (2), 353–370.
Poulos, Christopher N., 2009. Accidental Ethnography: An Inquiry into Family Secrecy. Left Coast Press, Walnut Creek, CA.
Roberts, J.M., 1972. The Mythology of the Secret Societies. Charles Scribner's Sons, New York.
Sedgwick, Eve Kosofsky, 1990. Epistemology of the Closet. University of California Press, Berkeley, MA.
Simmel, Georg, 1950 [1908]. Secrecy; the secret society. In: Wolff, K. (Ed.), The Sociology of Georg Simmel. Free Press, Glencoe, IL, pp. 330–376.
Snyder, Jon R., 2009. Dissimulation and the Culture of Secrecy in Early Modern Europe. University of California Press, Berkeley, MA.
Taussig, Michael, 1999. Defacement: Public Secrecy and the Labor of the Negative. Stanford University Press, Stanford, CA.
Tefft, Stanton K. (Ed.), 1980. Secrecy: A Cross-cultural Perspective. Human Sciences Press, New York.
Webster, Hutton, 1908. Primitive Secret Societies: A Study in Early Politics and Religion, second ed. rev. Macmillan, New York.
Wedgewood, Camilla H., 1930. The nature and function of secret societies. Oceania 1, 129–145.
Whitehead, Neil L., Wright, Robin (Eds.), 2004. In Darkness and Secrecy: The Anthropology of Assault Sorcery and Witchcraft in Amazonia. Duke University Press, Durham.

Relevant Websites

http://aaanet.org/issues/policy-advocacy/Code-of-Ethics.cfm – American Anthropological Association's Code of Ethics.
www.wikileaks.org – Wikileaks.

Secular Religions

Arthur L Greil, Alfred University, Alfred, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term 'secular religions' can be used to describe certain apparently secular enterprises that appear to share common features with enterprises usually thought of as 'religious.' Pursuing the analogy between such enterprises and religion raises serious theoretical questions concerning the proper definition of religion, the process of secularization, and the nature of religion in contemporary societies. Within contemporary sociology and anthropology, there exist several different research traditions that focus on the boundary between the religious and the secular, including the 'civil religion' approach, the 'implicit religion' tradition, the 'religious forms' approach, and the study of 'quasireligion.'

The term 'secular religions' can be used to describe certain apparently secular enterprises that appear to share common features with enterprises usually thought of as 'religious.' 'Secular religion' is one of several terms – including 'civil religion,' 'invisible religion,' 'parareligion,' and 'quasireligion' – which draw attention to religious and religious-like beliefs and activities that do not fit easily into the Western folk conception of religion as a distinct institutional structure focused on a transcendent being or beings.

Examples of 'Secular Religions'

At first blush, the very notion of 'secular religion' would appear to be an oxymoron. In both popular and sociological usage, the term 'religion' is typically associated with the realm of the 'sacred' and contrasted with that of the 'secular.' However, a number of scholars have pointed out striking similarities between ostensibly religious and ostensibly secular undertakings. Secular ideologies, it is argued, may, like religions, unite followers into a community of shared beliefs. They may provide adherents with a sense of meaning and ultimate purpose. They may inspire in believers a sense of transcendence usually associated with religion. Attempts to find parallels between the sacred and the secular have been especially common in studies of the political realm and in studies of therapeutic organizations and activities.

Political 'Religions'

Zuo (1991) has described the veneration of Chairman Mao during the Chinese Cultural Revolution as a political religion replete with sacred beings (Mao himself), sacred texts (the Little Red Book), and rituals (political denunciations). Gentile (1996) and others have described Fascism as secular religion. The term 'political religion' has also been employed to describe attempts made in developing societies to rally support for the concept of the nation. Crippin (1988) has gone so far as to argue that nationalism is the religion par excellence in modern society and that it is displacing more traditional forms of religion.

O'Toole (1977) has used the term 'sect' to describe certain political groups operating in Canada, including the Socialist Labor Party, followers of De Leon who wait for a communist millennium, which they regard as imminent. Social movements such as environmentalism (Brinkerhoff and Jacob, 1991), the animal rights movement, and the health food movement have been described as quasireligions insofar as they provide adherents with a coherent worldview and sense of purpose at the same time that they command a great deal of loyalty.

Therapeutic 'Religions'

Many scholars have drawn attention to the ritual aspects of Western medical practice. Others have pointed out that psychiatrists have much in common with shamans and other religious healers. A number of researchers have pointed out the similarities that exist between self-help groups and religious organizations (Galanter, 1989; Rudy and Greil, 1988). One family within the self-help movement that is perhaps more obviously 'religious' in character than most of the others includes Alcoholics Anonymous (AA) and other twelve-step or codependency groups. A few among many of the religious characteristics of AA are a conception of the sacred, ceremonies and rituals, creedal statements, experiences of transcendence, and the presence of an AA philosophy of life.

Other Examples

Other examples of attempts to point out analogies between apparently secular enterprises and religious ones abound in the social scientific literature. Within the realm of sport, attention has been paid to ritual elements and experiences of transcendence to be found in cricket, baseball, football, and the Olympic Games. It has now become commonplace among students of corporate culture to regard such activities as meetings, presentations, and retirement dinners as rituals. Ashforth and Vaidyanath (2002) have analyzed work organizations in general as secular religions. Some writers have drawn attention to the sectarian characteristics of certain types of businesses, such as home-party sales organizations and direct sales organizations. Voluntary organizations that have been interpreted as quasireligions include the Boy Scouts, groups of hobbyists and collectors, and 'fan clubs.' Jindra (1994) has described the phenomenon of Star Trek 'fandom' as religious in that it provides participants with a common worldview and inspires

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32129-8

high levels of commitment. Several writers have attempted to characterize atheism itself as a religious enterprise, characterizing Ethical Culture and other humanist groups as analogous to religions. Still others have viewed faith in science as the dominant religious perspective of the contemporary era.

'Secular Religions' and the Problem of Defining Religion

It may be argued that an interest in secular religion is a natural outgrowth of the functionalist perspective in classical sociology, which tended to view religion as serving a necessary function for the maintenance of society by uniting members of a society into a common moral universe. If, as many early sociologists thought, supernaturalistic conceptions of the universe were destined to recede in the face of industrialization and the increasing influence of science, the question arose as to what phenomena might serve as the 'social cement' of future societies. Comte's call for a 'religion of humanity' qualifies as an early sociological mention of the notion of a 'secular religion.' Durkheim's expectation that a 'cult of the individual' might play a similar role in the maintenance of society as that traditionally played by religions represents another early effort to provide intellectual justification for the idea that apparently secular ideologies may have religious characteristics.

As appealing and intuitive as the idea that secular enterprises may share important features with religions might be, the notion of a 'secular religion' must confront theoretical problems concerning the appropriate sociological definition of 'religion' and of 'the sacred.' Functional definitions of religion emphasize that the essential element in religion is the provision of an encompassing system of meaning or the ability to connect people to the ultimate conditions of their existence. Substantive definitions of religion argue that what distinguishes religion from other types of human activity is its reference to the sacred, the supernatural, or the superempirical.

The advantage of functional definitions is a breadth and inclusiveness that allows one to look at beliefs and practices not commonly referred to as religious but that may nonetheless resemble religious phenomena in important ways. One major disadvantage is that they may have the effect of so broadening the concept of religion that it becomes meaningless. While substantive definitions avoid this problem, they may result in allowing traditional Western conceptions of the nature of religion to determine what is to be included in the 'religious' category. Viewed from the perspective of the debate over the definition of religion, the theoretical problem with the concept of 'secular religion' is that it seems to rely simultaneously on both a broad functional approach to religion and a narrower substantive approach. The idea that environmentalism or nationalism could be properly called a religion relies on a functional definition, while the idea that such religions are 'secular' rather than 'sacred' necessarily depends on a substantive definition.

One might argue that both functional and substantive definitions of religion share the weakness that they privilege social scientists' conceptions of religion over folk conceptions of religion, that is to say the ways that people use the term 'religion' in everyday life. Some social scientists would therefore argue that the search for a scientifically valid definition of religion is futile and that the best that can be done is to define religion ethnographically, treating it as a 'category of discourse' with meanings that are changeable over time. The position that scholars take with regard to these definitional issues will necessarily influence the way in which they approach the study of 'secular religions.' This is perhaps one reason why there is at present no consensus within the social sciences with regard to the most appropriate method for studying phenomena residing on the border between the sacred and the secular.

Approaches to the Study of 'Secular Religions'

Within contemporary sociology and anthropology, there exist several different research traditions that focus on the boundary between the religious and the nonreligious.

The Notion of Civil Religion

While Rousseau coined the term 'civil religion,' its development as a social scientific concept is attributed to Bellah (1967). Working within the functionalist tradition, which sees religion as integrative for society, Bellah argued for the existence of an American civil religion, an 'institutionalized collection of sacred beliefs about the American nation,' which binds Americans together in spite of denominational pluralism. A key tenet of American civil religion is the conception of the United States as a nation with a divinely ordained mission. Although the civil religion concept was developed in the American context, it has been applied to the analysis of many states including Canada, Israel, and Malaysia.

The Implicit Religion Tradition

Although the term 'implicit religion' was coined and popularized by Bailey (1983), the concept owes much to the work of Luckmann (1967). Working with a broad functional definition of religion as the transcending of biological nature and the formation of a self, Luckmann argues that religion is a human universal. Luckmann maintains that traditional religions have become irrelevant in the modern world but that religion, rather than disappearing, has become personalized, privatized, and 'invisible.' Following Luckmann's lead, Bailey's implicit religion approach looks for the experience of the sacred within events of everyday life ordinarily dismissed as profane. Thus, in a study of interaction in an English public house, Bailey interprets the ethos of 'being a man,' mastering alcohol, and respecting the selves of others as implicitly religious.

The Study of 'Religious' Forms

There also exist a relatively large number of studies of general social 'forms' that are deemed to be relevant in both sacred and secular contexts. General discussions of 'secular ritual,' for example, fall into this category. Goffman's (1967) work on the functional significance of such rituals of everyday life as 'saving face' and showing 'deference' is particularly well known. Collins (1981) treats 'interaction ritual chains' as a key element

in his attempt to lay a microsociological foundation for macrosociology. Once conversations, social greetings, and athletic contests are allowed to count as ritual, ritual becomes virtually coterminous with social life. For this reason, some scholars reject 'ritual' as a meaningless category, while others argue that the ubiquity of ritual simply means that ritualizing is a fundamental human activity.

Many sociological studies of the commitment process either explicitly or implicitly present the commitment process as operating in much the same way in both sacred and secular contexts (Kanter, 1972). Some writers have developed models of the identity change process that highlight the similarity between religious conversion and other forms of identity change (Galanter, 1989). Studies that employ the term 'sect' in the analysis of apparently secular organizations have already been discussed.

The 'Quasireligion' Approach

The quasireligion approach popularized by Greil (1993) and his colleagues relies on an ethnographic approach to the term 'religion,' regarding it not as an objective category susceptible to social scientific definition but as a claim negotiated in the course of social interaction. Greil distinguishes between parareligions and quasireligions. Parareligions, which are ostensibly nonreligious entities that nonetheless deal with matters of ultimate concern, resemble the enterprises referred to in this article as 'secular religions.' The term 'quasireligion,' on the other hand, refers to groups, like AA or certain occult groups, which deal with matters of transcendence or ultimate concern, but which do not see themselves or present themselves unambiguously as religious. The concern here is not to determine whether a particular group is or is not a quasireligion but to examine the process by which particular groups attempt to claim or repudiate the religious label and by which other groups and individuals respond to these claims.

The Theoretical Significance of Secular Religions

Pursuing the analogy between ostensibly secular enterprises and 'religion' as it is usually conceived raises important questions concerning the proper definition of religion, the process of secularization, and the nature of religion in contemporary societies. The study of secular religions puts into bold relief the problematic nature of social scientific attempts to define religion. How one defines religion has important consequences for how one thinks about secular religions and for whether or not one regards the concept as useful. Attention to religious border phenomena presses one to consider the value of an ethnographic approach to definition that conceptualizes religion as a category of discourse whose precise meaning and implications are continually being negotiated in the course of social interaction. The study of secular religions also directs attention to difficulties in specifying and evaluating the secularization thesis, which claims that the influence of religion is declining in contemporary societies. It should be clear that one's approach to the question of secularization (including whether one even conceives of secularization as a theoretical possibility) depends on one's definition of religion and on one's treatment of religious border phenomena.

Several of the approaches to secular religion discussed here imply that traditional Western conceptions of religion as an institutionalized set of beliefs and practices focusing on a transcendent deity is beginning to lose its hold over many people. Likewise, many people seem to be tending to give expression to their experiences of transcendence in institutional contexts that have not typically been thought of as religious. If this is, in fact, the case, the line between religion and nonreligion may be expected to become even more blurred and the study of religious border phenomena even more central to the social scientific enterprise.

See also: Civil Religion; Religiosity; Secularization.

Bibliography

Ashforth, Blake E., Vaidyanath, Deepa, 2002. Work organizations as secular religions. Journal of Management Inquiry 11, 359–370.

Bailey, Edward, 1983. The implicit religion of contemporary society: an orientation and plea for its study. Religion 13, 69–83.

Bellah, Robert N., 1967. Civil religion in America. Daedalus 96, 1–21.

Brinkerhoff, Merlin B., Jacob, Jeffrey C., 1991. Mindfulness and quasi-religious meaning systems: an empirical exploration within the context of ecological sustainability and deep ecology. Journal for the Scientific Study of Religion 38, 524–542.

Collins, Randall, 1981. On the micro-foundations of macro-sociology. American Journal of Sociology 86, 984–1084.

Crippin, Timothy, 1988. Old and new gods in the modern world: toward a theory of religious transformation. Social Forces 67, 316–336.

Galanter, Marc, 1989. Cults: Faith, Healing, and Coercion. Oxford University Press, New York.

Gentile, Emilio, 1996. The Sacralization of Politics in Fascist Italy. Harvard University Press, Cambridge, MA.

Goffman, Erving, 1967. Interaction Ritual: Essays on Face-to-face Behavior. Doubleday, New York.

Greil, Arthur L., 1993. Explorations along the sacred frontier: notes on para-religions, quasi-religions, and other boundary phenomena. In: Bromley, David, Hadden, Jeffrey K. (Eds.), Handbook of Cults and Sects in America: Assessing Two Decades of Research and Theory Development. Religion and the Social Order, vol. 2. JAI Press, Greenwich, CT.

Jindra, Michael, 1994. Star Trek fandom as a religious phenomenon. Sociology of Religion 55, 27–51.

Kanter, Rosabeth Moss, 1972. Commitment and Community: Communes and Utopias in Sociological Perspective. Harvard University Press, Cambridge, MA.

Luckmann, Thomas, 1967. The Invisible Religion: The Problem of Religion in Modern Society. Macmillan, New York.

O'Toole, Roger, 1977. The Precipitous Path: Studies in Political Sects. Peter Martin, Toronto.

Rudy, David R., Greil, Arthur L., 1988. Is Alcoholics Anonymous a religious organization? Sociological Analysis 50, 41–51.

Zuo, Jiping, 1991. Political religion: the case of the cultural revolution in China. Sociological Analysis 52, 99–110.

Secularization

Jose Casanova, New School of Social Research, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13786–13791, © 2001, Elsevier Ltd.

Abstract

In its precise historical sense, 'secularization' refers to the transfer of persons, things, meanings, etc., from ecclesiastical or religious to civil or lay use. In its broadest sense, often postulated as a universal developmental process, secularization refers to the progressive decline of religious beliefs, practices, and institutions.

In its precise historical sense, 'secularization' refers to the transfer of persons, things, meanings, etc., from ecclesiastical or religious to civil or lay use. In its broadest sense, often postulated as a universal developmental process, secularization refers to the progressive decline of religious beliefs, practices, and institutions.

The Term 'Secularization'

Etymologically, the term secularization derives from the medieval Latin word *saeculum*, with its dual temporal-spatial connotation of secular *age* and secular *world*. Such a semantic connotation points to the fact that social reality in medieval Christendom was structured through a system of classification which divided 'this world' into two heterogeneous realms or spheres, 'the religious' and 'the secular.' This was a particular, and historically rather unique, variant of the kind of universal dualist system of classification of social reality into sacred and profane realms, postulated by Émile Durkheim.

In fact, Western European Christendom was structured through a double dualist system of classification. There was, on the one hand, the dualism between 'this world' (the City of Man) and 'the other world' (the City of God). There was, on the other hand, the dualism within 'this world' between a 'religious' and a 'secular' sphere. Both dualisms were mediated, moreover, by the sacramental nature of the church, situated in the middle, simultaneously belonging to both worlds and, therefore, able to mediate sacramentally between the two.

The differentiation between the cloistered *regular* clergy and the *secular* clergy living in the world was one of the many manifestations of this dualism. The term secularization was first used in canon law to refer to the process whereby a religious monk left the cloister to return to the world and thus become a secular priest. In reference to an actual historical process, however, the term secularization was first used to signify the lay expropriation of monasteries, landholdings, and the mortmain wealth of the church after the Protestant Reformation. Thereafter, secularization has come to designate any transfer from religious or ecclesiastical to civil or lay use.

Secularization as a Historical Process

Secularization refers to the historical process whereby the dualist system within 'this world' and the sacramental structures of mediation between this world and the other world progressively break down until the medieval system of classification disappears. Max Weber's expressive image of the breaking of the monastery walls remains perhaps the best graphic expression of this radical spatial restructuration initiated by the Protestant Reformation. This process, which Weber conceptualized as a general reorientation of religion from an other-worldly to an inner-worldly direction, is literally a process of secularization. Religious 'callings' are redirected to the *saeculum*. Salvation and religious perfection are no longer to be found in withdrawal from the world but in the midst of worldly secular activities. In fact, the symbolic wall separating the religious and the secular realms breaks down. The separation between 'this world' and 'the other world,' for the time being at least, remains. But, from now on, there will be only one single 'this world,' the secular one, within which religion will have to find its own place. To study what new systems of classification and differentiation emerge within this one secular world and what new place religion will have within it is precisely the analytical task of the theory of secularization.

Obviously, such a concept of secularization refers to a particular historical process of transformation of Western European Christian societies and might not be directly applicable to other non-Christian societies with very different modes of structuration of the sacred and profane realms. It could hardly be applicable, for instance, to such 'religions' as Confucianism or Taoism, insofar as they are not characterized by high tension with 'the world' and have no ecclesiastical organization. In a sense those religions which have always been 'worldly' and 'lay' do not need to undergo a process of secularization.

In itself such a spatial-structural concept of secularization describes only changes in the location of Christian religion from medieval to modern societies. It tells very little, however, about the extent and character of the religious beliefs, practices, and experiences of individuals and groups living in such societies. Yet the theory of secularization adopted by the modern social sciences incorporated the beliefs in progress and the critiques of religion of the Enlightenment and of Positivism and assumed that the historical process of secularization entailed the progressive decline of religion in the modern world. Thus, the theory of secularization became embedded in a philosophy of history that saw history as the progressive evolution of humanity

from superstition to reason, from belief to unbelief, from religion to science.

The Secularization Paradigm

Secularization might have been the only theory within the social sciences that was able to attain the status of a paradigm. In one form or another, with the possible exception of Alexis de Tocqueville, Vilfredo Pareto, and William James, the thesis of secularization was shared by all the founders. Paradoxically, the consensus was such that for over a century the theory of secularization remained not only uncontested but also untested. Even Durkheim's and Weber's work, while serving as the foundation for later theories of secularization, offer scant empirical analysis of modern processes of secularization, particularly of the way in which those processes affect the place, nature and role of religion in the modern world. Even after freeing themselves from some of the rationalist and positivist prejudices about religion, they still share the major intellectual assumptions of the age about the future of religion.

For Durkheim, the old gods were growing old or already dead and the dysfunctional historical religions would not be able to compete with the new functional gods and secular moralities which modern societies were bound to generate. For Weber, the process of intellectual rationalization, carried to its culmination by modern science, had ended in the complete disenchantment of the world, while the functional differentiation of the secular spheres had displaced the old integrative monotheism, replacing it with the modern polytheism of values and their unceasing and irreconcilable struggle. The old churches remain only as a refuge for those 'who cannot bear the fate of the times like a man' and are willing to make the inevitable 'intellectual sacrifice.'

Only in the 1960s one finds the first attempts to develop more systematic and empirically grounded formulations of the theory of secularization in the works of Acquaviva (1961), Berger (1967), Luckmann (1963), and Wilson (1966). It was then, at the very moment when theologians were celebrating the death of God and the secular city, that the first flaws in the theory became noticeable and the first systematic critiques were raised by Martin (1969) and Greeley (1972) in what constituted the first secularization debate. For the first time it became possible to separate the theory of secularization from its ideological origins in the Enlightenment critique of religion and to distinguish the theory of secularization, as a theory of the modern autonomous differentiation of the secular and the religious spheres, from the thesis that the end result of the process of modern differentiation would be the progressive erosion, decline and eventual disappearance of religion. Greeley (1972) already pointed out that the secularization of society, which he conceded, by no means implied the end of church religiosity, the emergence of 'secular man,' or the social irrelevance of religion in modern secular societies. Yet after three decades the secularization debate remains unabated. Defenders of the theory tend to point to the secularization of society and to the decline of church religiosity in Europe as substantiating evidence, while critics tend to emphasize the persistent religiosity in the United States and widespread signs of religious revival as damaging counterevidence that justify discarding the whole theory as a 'myth.'

The Three Subtheses of the Theory of Secularization

Although it is often viewed as a single unified theory, the paradigm of secularization is actually made up of three different and disparate propositions: secularization as differentiation of the secular spheres from religious institutions and norms, secularization as general decline of religious beliefs and practices, and secularization as privatization or marginalization of religion to a privatized sphere. Strictly speaking, the core and central thesis of the theory of secularization is the conceptualization of the historical process of societal modernization as a process of functional differentiation and emancipation of the secular spheres – primarily the state, the economy, and science – from religion and the concomitant specialized and functional differentiation of religion within its own newly found religious sphere. The other sub-theses, the decline and privatization of religion, were added as allegedly necessary structural consequences of the process of secularization. Maintaining this analytical distinction should allow the examination and testing of the validity of each of the three propositions independently of each other and thus refocus the often fruitless secularization debate into comparative historical analysis that could account for obviously different patterns of secularization.

The Differentiation and Secularization of Society

The medieval dichotomous classification of reality into religious and secular realms was to a large extent dictated by the church. In this sense, the official perspective from which medieval societies saw themselves was a religious one. Everything within the *saeculum* remained an undifferentiated whole as long as it was viewed from the outside, from the perspective of the religious. The fall of the religious walls put an end to this dichotomous way of thinking and opened up an entire new space for processes of internal differentiation of the various secular spheres. Now, for the first time, the various secular spheres – politics, economics, law, science, art, etc. – could come fully into their own, become differentiated from each other, and follow what Weber called their "internal and lawful autonomy." The religious sphere, in turn, became a less central and spatially diminished sphere within the new secular system, but also a more internally differentiated one, specializing in 'its own religious' function and losing many other 'non-religious' (clerical, educational, social welfare) functions (Luhmann, 1977). The loss in functions entailed as well a significant loss in hegemony and power.

It is unnecessary to enter into the controversial search for first causes setting the modern process of differentiation into motion. It suffices to stress the role which four related and parallel developments played in undermining the medieval religious system of classification: the Protestant Reformation; the formation of modern states; the growth of modern capitalism; and the early modern scientific revolution. Each of the four developments contributed its own dynamic to modern processes of secularization. The four of them together were certainly more than sufficient to carry the process through.

The Protestant Reformation by undermining the universalist claims of the Roman Catholic church helped to destroy the old organic system of Western Christendom and to liberate the secular spheres from religious control. Protestantism also served to legitimate the rise of bourgeois man and of the new entrepreneurial classes, the rise of the modern sovereign state against the universal Christian monarchy, and the triumph of the new science against Catholic scholasticism. Moreover, Protestantism may also be viewed as a form of internal secularization, as the vehicle through which Christian religious contents were to assume institutionalized secular forms in modern societies, thereby erasing the religious/secular divide.

If the universalist claims of the church as a salvation organization were undermined by the religious pluralism introduced by the Reformation, its monopolist compulsory character was undermined by the rise of a modern secular state which progressively was able to concentrate and monopolize the means of violence and coercion within its territory. In the early absolutist phase the alliance of throne and altar became even more accentuated. The churches attempted to reproduce the model of Christendom at the national level, but all the territorial national churches, Anglican, Lutheran, Catholic, and Orthodox, fell under the caesaro–papist control of the absolutist state. As the political costs of enforcing conformity became too high, the principle *cuius regio eius religio* turned into the principle of religious tolerance and state neutrality toward privatized religion, the liberal state's preferred form of religion. Eventually, new secular *raison d'état* principles led to the constitutional separation of church and state, even though some countries such as England and the Scandinavian Lutheran countries may have maintained formal establishment.

Capitalism, that revolutionizing force in history which according to Marx 'melts all that is solid into air and profanes all that is holy,' had already sprouted within the womb of the old Christian society in the medieval towns. No other sphere of the *saeculum* would prove more secular and more unsusceptible to religious and moral regulation than the capitalist market. Nowhere is the transvaluation of values which takes place from Medieval to Puritan Christianity as radical and as evident as in the change of attitude toward 'charity' – that most Christian of virtues – and toward poverty. Following Weber (1958), one could distinguish three phases and meanings of capitalist secularization: in the Puritan phase, 'asceticism was carried out of monastic cells into everyday life' and secular economic activities acquired the meaning and compulsion of a religious calling; in the utilitarian phase, as the religious roots dried out, the irrational compulsion turned into 'sober economic virtue' and 'utilitarian worldlines'; finally, once capitalism 'rests on mechanical foundations,' it no longer needs religious or moral support and begins to penetrate and colonize the religious sphere itself, subjecting it to the logic of commodification (Berger, 1967).

The conflict between the church and the new science, symbolized by the trial of Galileo, was not as much about the substantive truth or falsity of the new Copernican theory as it was about the validity of the claims of the new science to have discovered a new autonomous method of obtaining and verifying truth. The conflict was above all about science's claims to differentiated autonomy from religion. The Newtonian Enlightenment established a new synthesis between faith and reason, which in Anglo-Saxon countries was to last until the Darwinian crisis of the second half of the nineteenth-century. Across the Channel, however, the Enlightenment became patently radicalized and militantly anti-religious. Science was transformed into a scientific and scientistic worldview which claimed to have replaced religion the way a new scientific paradigm replaces an outmoded one.

As each of these carriers – Protestantism, the modern state, modern capitalism, and modern science – developed different dynamics in different places and at different times, the patterns and the outcomes of the historical process of secularization varied accordingly. Yet it is striking how few comparative historical studies of secularization there are which would take these four, or other, variables into account. David Martin's *A General Theory of Secularization* is perhaps the single prominent exception. Only when it comes to capitalism has it been nearly universally recognized that economic development affects the 'rates of secularization,' that is, the extent and relative distribution of religious beliefs and practices. This positive insight, however, turns into a blinder, when it is made into the sole variable accounting for different rates of secularization. As a result, those cases in which no positive correlation is found, as expected, between rates of secularization and rates of industrialization, urbanization, proletarianization, education, in short, with indicators of socio-economic development, are classified as 'exceptions' to the 'rule.'

The Decline of Religion Thesis

The assumption, often stated but mostly unstated, that religion in the modern world was declining and would likely continue to decline has been until recently a dominant postulate of the theory of secularization. It was based primarily on evidence from European societies showing that the closer people were involved in industrial production, the less religious they became, or at least, the less they took part in institutional church religiosity. The theory assumed that the European trends were universal and that non-European societies would evince similar rates of religious decline with increasing industrialization. It is this part of the theory which has proven patently wrong.

One should keep in mind the inherent difficulties in comparative studies of religion. Globally, the evidence is insufficient and very uneven. There is often no consensus as to what counts as religion and even when there is agreement on the object of study, there is likely to be disagreement on which of the dimensions of religiosity (membership affiliation, official vs. popular religion, beliefs, ritual and nonritual practices, experiences, doctrinal knowledge, and their behavioral and ethical effects) one should measure and how various dimensions should be ranked and compared.

Nevertheless, one can prudently state that since World War II, despite rapid increases in industrialization, urbanization, and education, most religious traditions in most parts of the world have either experienced some growth or maintained their vitality (Whaling, 1987). The main exceptions were: the rapid decline of primal religions mostly in exchange for more 'advanced' ones (mainly Islam and Christianity), the sudden and dramatic decline of religion in communist countries, a process which is now being reversed after the fall of

communism, and the continuous decline of religion throughout much of Western Europe (and, one could add, some of its colonial outposts such as Argentina, Uruguay, New Zealand, and Quebec).

What remains, therefore, as significant and overwhelming evidence is the progressive and apparently still continuing decline of religion in Western Europe. It is this evidence which has always served as the empirical basis for most theories of secularization. Were it not for the fact that religion shows no uniform sign of decline in Japan or the United States, two equally modern societies, one could still perhaps maintain the 'modernizing' developmentalist assumption that it is only a matter of time before the more 'backward' societies catch-up with the more 'modern' ones. But such an assumption is no longer tenable. Leaving aside the evidence from Japan, a case which should be crucial, however, for any attempt to develop a *general* theory of secularization, there is still the need to explain the obviously contrasting religious trends in Western Europe and the United States.

Until very recently, most of the comparative observations as well as attempts at explanation came from the European side. European visitors have always been struck by the vitality of American religion. The United States appeared simultaneously as the land of 'perfect disestablishment' and as 'the land of religiosity *par excellence*.' Yet, Europeans rarely felt compelled to put into question the thesis of the general decline of religion in view of the American counterevidence. Religious decline was so much taken for granted that what required an explanation was the American 'deviation' from the European 'norm.' The standard explanations have been either the expedient appeal to 'American exceptionalism' or the casuistic strategy to rule out the American evidence as irrelevant, because American religion was supposed to have become so 'secular' that it should no longer count as religion (Luckmann, 1963).

From a global–historical perspective, however, it is the dramatic decline of religion in Europe that truly demands an explanation. A plausible answer would require a search for independent variables, for those independent carriers of secularization present in Europe but absent in the United States. Looking at the four historical carriers mentioned above, neither Protestantism nor capitalism would appear as plausible candidates. The state and scientific culture, however, could serve as plausible independent variables, since church–state relations and the scientific worldviews carried by the Enlightenment were significantly different in Europe and America.

What the United States never had was an absolutist state and its ecclesiastical counterpart, a caesaro–papist state church. It was the caesaro–papist embrace of throne and altar under absolutism that perhaps more than anything else determined the decline of church religion in Europe. Consistently throughout Europe, nonestablished churches and sects in most countries have been able to withstand the secularizing trends better than the established church. It was the very attempt to preserve and prolong Christendom in every state and thus to resist modern functional differentiation that nearly destroyed the churches in Europe.

The Enlightenment and its critique of religion became themselves independent carriers of processes of secularization wherever the established churches became obstacles to the modern process of functional differentiation or resisted the emancipation of the cognitive–scientific, political–practical, or aesthetic–expressive secular spheres from religious and ecclesiastical tutelage. In such cases, the Enlightenment critique of religion was usually adopted by social movements and political parties, becoming in the process a self-fulfilling prophecy. By contrast, wherever religion itself accepted, perhaps even furthered, the functional differentiation of the secular spheres from the religious sphere, the radical Enlightenment and its critique of religion became superfluous. Simply put, the more religions resist the process of modern differentiation, that is, secularization in the first sense, the more they will tend in the long run to suffer religious decline, that is, secularization in the second sense.

The Privatization of Religion Thesis

As a corollary of the thesis of differentiation, religious disestablishment entails the privatization of religion. Religion becomes indeed 'a private affair.' Insofar as freedom of conscience, 'the first freedom' as well as the precondition of all modern freedoms, is related intrinsically to 'the right to privacy' – to the modern institutionalization of a private sphere free from governmental intrusion as well as free from ecclesiastical control – and inasmuch as the right to privacy serves as the very foundation of modern liberalism and of modern individualism, then indeed the privatization of religion is essential to modern societies.

There is, however, another more radical version of the thesis of privatization which often appears as a corollary of the decline of religion thesis. In modern secular societies, whatever residual religion remains becomes so subjective and privatized that it turns 'invisible,' that is, marginal and irrelevant from a societal point of view. Not only are traditional religious institutions becoming increasingly irrelevant but, Luckmann (1963) adds, modern religion itself is no longer to be found inside the churches. The modern quest for salvation and personal meaning has withdrawn to the private sphere of the self. Significant for the structure of modern secular societies is the fact that this quest for subjective meaning is a strictly personal affair. The primary 'public' institutions (state, economy) no longer need or are interested in maintaining a sacred cosmos or a public religious worldview.

The functionalist thesis of privatization turns problematic when instead of being a falsifiable empirical theory of dominant historical trends, it becomes a prescriptive normative theory of how religious institutions ought to behave in the modern world. Unlike secular differentiation, which remains a structural trend that serves to define the very structure of modernity, the privatization of religion is a historical option, a 'preferred option' to be sure, but an option nonetheless. Privatization is preferred internally as evinced by general pietistic trends, by processes of religious individuation, and by the reflexive nature of modern religion. Privatization is constrained externally by structural trends of differentiation which force religion into a circumscribed and differentiated religious sphere. Above all, privatization is mandated ideologically by liberal categories of thought which permeate modern political and constitutional theories.

The theory of secularization should be free from such a liberal ideological bias and admit that there may be

legitimate forms of 'public' religion in the modern world, which are not necessarily anti-modern fundamentalist reactions and which do not need to endanger either modern individual freedoms or modern differentiated structures. Many of the recent critiques and revisions of the paradigm of secularization derive from the fact that in the 1980s religions throughout the world thrust themselves unexpectedly into the public arena of moral and political contestation, demonstrating that religions in the modern secular world continue and will likely continue to have a public dimension.

See also: Religion, Phenomenology of; Religion, Sociology of; Religion: Evolution and Development; Secular Religions.

Bibliography

Acquaviva, S.S., 1961. L'ecclissi del sacro nella civiltà industrialle. Edizioni di Comunità, Milan [1979. The Decline of the Sacred in Industrial Society. Blackwell, Oxford].

Berger, P., 1967. The Sacred Canopy. Doubleday, Garden City, NY.

Casanova, J., 1994. Public Religions in the Modern World. University of Chicago Press, Chicago.

Dobbelaere, K., 1981. Secularization: A Multidimensional Concept. Sage, Beverly Hills, CA.

Greeley, A., 1972. Unsecular Man: The Persistence of Religion. Schocken, New York.

Greeley, A., 1989. Religious Change in America. Harvard University Press, Cambridge, MA.

Hadden, J.K., Shupe, A. (Eds.), 1989. Secularization and Fundamentalism Reconsidered. Paragon House, New York.

Luckmann, T., 1963. Das Problem der Religion in der modernen Gesellschaft. Rombach, Freiburg [1967. Invisible Religion. Macmillan, New York].

Luhmann, N., 1977. Funktion der Religion. Suhrkamp Verlag, Frankfurt [1984. Religious Dogmatics and the Evolution of Societies. Mellen Press, New York].

Martin, D., 1969. The Religious and the Secular. Shocken, New York.

Martin, D., 1978. A General Theory of Secularization. Blackwell, Oxford, UK.

Stark, R., Bainbridge, W.S., 1985. The Future of Religion. University of California Press, Berkeley, CA.

Weber, M., 1958. The Protestant Ethic and the Spirit of Capitalism. Scribner's Sons, New York.

Whaling, F. (Ed.), 1987. Religion in Today's World: The Religious Situation of the World from 1945 to the Present Day. T & T Clark, Edinburgh, UK.

Wilson, B., 1966. Religion in Secular Society. Watts, London.

Wilson, B., 1985. Secularization: the inherited model. In: Hammond, P. (Ed.), The Sacred in a Secular Age. University of California Press, Berkeley, CA.

Secularization: History of the Concept

Stijn Latré and Guido Vanheeswijck, Center Pieter Gillis and Center for Philosophy of Culture, University of Antwerp, Antwerp, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the historical evolution of the concept of 'secularization' in sociology and philosophy. It does not include a description of political systems and their approach to religion and secularity. The authors dwell on the classic secularization thesis and explain how this thesis was questioned in sociology and philosophy alike. The secularization debate nowadays counts many participants reflecting diverging normative positions. Despite this multitude of positions and nuances, the debate is likely to continue in the future along two particular strands: first, discussions on the use of concepts as secular/post-secular and the tenability of the religious/secular divide; second, discussions that will revolve around Western modernity/multiple modernities, or 'the West versus the rest.'

Definition, Etymology, and Classic Secularization Thesis

Secularization is a multilayered concept with at least four layers of meaning (cf Gorski and Altimordu, 2008, pp. 59–60). First, secularization etymologically derives from the Latin world *saeculum*, meaning 'century' in worldly time, as opposed to *eternitas*, eternal life where human, worldly concepts of time have no meaning or to *eschaton*, the end of times. In medieval canonical law, *saecularizatio* refers to a monk's exit from his monastery and his return to the world. The third layer dates from the period of reformation, when Protestant leaders seized church properties. As such, the concept of secularization first emerges as a French neologism in the sixteenth century, in a theological – literary essay by Pierre Crespet (1543–94) (Monod, 2002, p. 20). In 1690 the concept reappears in the posthumously published dictionary of Antoine Furetière (1619–88) with a slightly different meaning. There the entry refers to the transformation of ecclesiastic possessions, for example, the reconversion of monasteries to hospitals or universities, or to the selling of ecclesiastic goods for civil purposes. This material meaning of the process of secularization also prevails between 1648 and 1803, when church goods are rendered a worldly status as a result of the peace of Westphalia and the policy of Napoleon, respectively (Monod, 2002, p. 21). The thus far final layer of meaning emerges at the end of the nineteenth century: secularization as the political project of a secularist movement (Gorski and Altimordu, 2008, p. 61). Hence, secularization will always denote a transition from a religious to a more worldly level.

All these shades of meaning shine through in the so-called classic secularization thesis, that is, the idea that modernization would, with a sort of inner necessity, eventually lead to the disappearance of religion (De Vriese, 2010, p. 408; Gorski and Altimordu, 2008, p. 56). Adherents of the classic secularization thesis saw religion and secularization as strongly opposed to each other. Not only from a historical perspective – first there was religion, and then came secularization – but also from an ideological angle both were generally presented as antipodes. In that perspective, religion was deemed as irrational whereas secularization was considered as the product of rational progress.

Secularization Thesis under Pressure

Due to a number of historical events, the classic secularization thesis has become controversial during the last decades. The Iranian Revolution of 1979, the upsurge of the Moral Majority, the New Christian Right and Evangelicalism in the United States, the political upheaval due to fundamentalist radicalization in several countries: these and analogous developments all over the world have undermined the paradigmatic status of the secularization thesis and have given rise to a huge controversy within modern social sciences as to its objective basis (De Vriese, 2010, pp. 410–412, Habermas, 2008, p. 2)

This controversy has led to various positions among sociologists and philosophers alike. First, some scholars continue to defend the classic secularization thesis, albeit in a more moderate form. Secularity is definitive in some societies and secularization continues to progress in other societies, but the prediction that religion would entirely disappear does no longer make sense. However, the boosters of the secularization thesis remain convinced that modernity in all its ramifications has diminished the significance of religion. In this respect, the differentiation of societal functions and scientific disciplines (e.g., psychology, economy, and sociology) that formerly pertained to religion counts as one of many explanatory factors. More moderate adherents of the classic thesis argue for secularization as the best warrant for religion to flourish on a basis of freedom and equality in relation to nonreligious outlooks.

The number of adherents to the classic secularization thesis is gradually being outweighed by an increasing amount of scholars who explicitly reject its hegemony. They point to thriving religiosity on a global scale, or even see signs of desecularization within Western, allegedly, secular societies. Religion will continue to play an important role, both in the private and in the public sphere, and is entitled to do so.

Another angle from which to look at the secularization debate is that of transcendental and genealogical accounts of secularization, which have multiplied in recent years. These genealogical narratives focus on the hidden transcendental background – ideological, rhetorical, psychological, and epistemological – behind the adherence or opposition to the secularization thesis. A number of scholars (Marcel Gauchet, Jürgen Habermas, Charles Taylor, Robert Bellah, and José Casanova),

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03113-5

precisely because they are aware that theories on secularization are shaped at a much deeper, often unconscious transcendental level, aim at a clarification of the concept. Although these scholars depart from a common historical and genealogical approach, their ways part when it comes to their evaluation of secularization. In this sense, the secularization debate cannot escape meta-historical normativity.

This brief introduction into the etymology of the concept of secularization and to the current debate of secularization is elaborated in the following four sections. We begin with a selective overview of the debates in sociology; we then turn to the field of philosophy; these overviews set the stage for a reflection upon the relation between the concepts of 'post-secular,' 'secularization,' and 'secularism.' Finally, we make a (much too small) digression on the non-Western view on the concept of secularization.

Sociology of Secularization: A Selective Overview

The origin of the sociological secularization theory is strongly related to *Max Weber*'s thoughts on the rise of capitalism and its protestant origins in *The Protestant Ethic and the 'Spirit' of Capitalism*. With the rise of modern capitalism, Weber claims, money becomes a goal in itself and is no longer subservient to 'higher goals,' let alone religious ones. Weber traces this evolution back to the spirit of Protestantism. Certain strands of Protestantism, such as Calvinism, developed a theory of predestination. Man's fate is regulated by divine providence. Every person's salvation depends solely on God's grace. This paradoxically made Calvinist believers strive for worldly success. The more wealth one could accumulate within earthly existence, the firmer the proof of election by God. Of course, the economic works by man were seen as homage, an act of gratitude to the benedictions of God. But by the same token, a lot of energy was invested in worldly affairs, which were eventually cut loose from their divine purposes.

Indirectly influenced by Weber, *Peter Berger* argued in the 1960s, most notably in *The Sacred Canopy*, for a theory of secularization that considered modernization and secularization as causally connected. In other words, the general thrust of his argument was that modernity inevitably undermined the social significance of religion. As to the causes and the results of the loss of function of religion, there was a divergence of opinion. Some argued that the future of religion was complete extinction (Wallace, 1966), some attributed more weight to social or religious changes, such as the industrial revolution or the protestant reformation (Berger, 1969) than to the growth of scientific knowledge, while others thought vice versa. However divergent in the detailed elaboration of their positions, during the late 1960s and early 1970s, most sociologists of religion embraced, one way or another, the secularization thesis in its basic form.

By the late 1970s or early 1980s, the majority of sociologists of religion came to agree that the original secularization thesis was untenable in its basic form. That change was primarily inspired not by any change in theological or philosophical positions, but by the massive weight of empirical evidence, the 'revenge of facts' (De Vriese, 2010). Most conspicuously, it was Berger himself, one of the founding fathers of the classic secularization theory, who in his 1999 edited volume *The Desecularization of the World*, renounced his earlier theory in unequivocal terms, by stating that the classic secularization theory is 'essentially mistaken' (Berger, 1999, p. 2). Other scholars proclaimed that God was back (Micklethwait and Wooldridge, 2009).

One of the most extreme voices in this debate among proponents and opponents of the secularization theory was that of *Rodney Stark*. Already in the mid-1980s he and his collaborators vehemently attacked the then dominant secularization theory as propounded by Wallace and Berger. Stark and his colleagues' point of departure was a rather idiosyncratic definition of secularization as a decline in individual belief and practice, based on what they called a supply-side or religious economies model. This model allowed them to argue that neither the United States nor even Western Europe was really secularized. In the case of the United States, this is clear enough: although indisputably modern (urbanized, industrialized, democratic, and so on), it is not secular in terms of demand and supply (high rate of church attendance and membership; even higher rate of belief in God and an afterlife). But in their view, Europe is not secular either. That the levels of church attendance and church membership are lower than in the United States, is not due to a decline in religious demand but rather to structural deficiencies in the religious supply, that is, the age-old state-regulated structures of Europe's religious market. Since Stark considers religious demand as a constant and since he defines secularization as a deficiency in religious supply, he concludes that the secularization doctrine may be carried to the graveyard of failed theories (Stark, 1999, pp. 269–270; cf Gorski and Altimordu, 2008, p. 57; De Vriese, 2010, p. 409).

Admittedly, Berger's recantation of his earlier theory and Stark's outspoken dismissal of secularization as such did remain controversial. In particular, Stark, Finke, and Iannaconne's alternative supply-side model and their end-of-secularization declaration were charged as being oversimplified and much too individualistic. Mainly in reaction to the supply-side model of sociology, divergent sociologists of religion have been highlighting the supra-individual aspects that are at play in relation between religion and secularization (the power of ecclesiastical elites and institutions, cultural and symbolic structures, and so on). They claim that the process of secularization is a highly complicated multileveled process, going far beyond the demand of and supply for individual actors. Among them, there are both opponents and defenders of the secularization theory.

The most prominent defender of what is coined as the neoorthodox secularization thesis is the British sociologist *Steve Bruce*. Why neoorthodox? On the one hand, Bruce recognizes the complexity of the secularization process: he is fully aware of its cross-national varieties and historical contingencies. On the other hand, he remains embedded in the rather ahistorical and teleological background of the classic secularization theory by positing the strong link between modernization and secularization (Gorski and Altimordu, 2008, p. 59; Bruce, 2011, pp. 27, 223). Another proponent of the neoorthodox strand is *Gregory S. Paul*, who has done some effort to statistically prove correlations between democracy, secularity, and affluence (Paul, 2005, 2010).

Due to increasingly divergent definitions of secularization, the Belgian sociologist *Karel Dobbelaere* undertook an attempt at clarification. He distinguished between three levels of analysis: the individualization and privatization of religious practice and belief (micro-level); the emergence of competitive religious markets (meso-level); and the functional differentiation of societal subsystems (macro-level) (Dobbelaere, 1999). Analogous to Dobbelaere, *José Casanova* distinguishes three different connotations of the term secularization so as to overcome the impasse in the secularization debate: (1) secularization as the *decline* of religious beliefs and practices; (2) secularization as the *privatization* of religion (see also Luckmann, 1967, 1996); and (3) secularization as the *differentiation* of the secular spheres (Casanova, 1994, 2006, p. 7; Gorski and Altimordu, 2008, pp. 56–58). While the 'decline' and 'privatization' sub-definitions have been heavily criticized in the last decades, the 'differentiation' thesis remains relatively uncontested in the social sciences (Casanova, 2006, p. 9). But even then, the question remains whether it is correct to subsume the different geographical and historical patterns of differentiation into a universal process of modernization.

Each endeavor to find an answer to that latter question presupposes a genuine awareness of its complexity. Not only does one have to take into account the influence of a new concept, that is, that of 'multiple modernities,' first developed by *Shmuel N. Eisenstadt* (cf infra), but also one must realize as well that the vicissitudes within the evolution of secularization are one way or another strongly linked with internal transformations of religion. Therefore, parallel to elaborating a multilayered concept of secularization we are in need of a new concept of religion, seen in the light of their mutual relations.

Grace Davie and *Danièle Hervieu-Léger* have been working on this topic, albeit from different angles. Davie characterizes the contemporary situation in Western Europe as 'believing without belonging.' The normative link between a religious identity, belief in dogmatic propositions, and regular church practice does no longer hold for a majority of 'believers.' Hervieu-Léger offers a different, even reverse characterization of the same European situation as 'belonging without believing.' But these, at face value contradictory characterizations, illustrate the emergence of a new concept of religion that was coined by Davie as 'vicarious religion.' This concept indicates that people do not always connect explicitly to a specific religious tradition, but approve as a tacit majority of the religious activities ('belonging') or religious beliefs of an active minority (Davie, 1994, 2000; Hervieu-Léger, 1999, 2004; Casanova, 2006, p. 14; Taylor, 2007, pp. 513–514, 521–522). It exemplifies the unmistakable fact, at least in Europe, that the religious and the secular mutually condition each other (Casanova, 2006, pp. 10–14)

Is a conclusion possible from this selective overview? Hardly. It may be clear that the classic secularization thesis, once hegemonic, now finds itself in an increasingly defensive position. But that does not mean that everyone is equally prepared to dispense with the theory altogether. The debunkers are accompanied by the believers and both of them are criticized by many revisionists who opt for a redefinition of the concepts of secularization and religion alike.

Viewed over a longer period from the nineteenth to the twenty-first centuries, this evolution in the sociology of religion parallels the evolution within the domain of philosophy of religion. In post-World War II philosophy, prominent philosophers, in search for a redefinition of the concept of secularization, also favor a three-tiered scheme. Mutual references between sociologists and philosophers abound. As a consequence, main protagonists such as Weber, Gauchet, Taylor, Casanova, and so many others in the debate on the concept of secularization are figuring prominently in general reviews of sociological and philosophical journals alike (De Vriese, 2010, pp. 409–410; Taylor, 2007; Gauchet, 1997; Gorski and Altimordu, 2008).

Philosophy of Secularization: A Selective Overview

Philosophical debates reflect a profound analogy with the sociological views on secularization. First, we have the classic secularization thesis that interprets modernity as a dynamic of ongoing progress, at the cost of 'irrational' religion. In this respect, *Auguste Comte*'s positivistic view of the relation paved way for contemporary views about the irrationality or superfluity of religion in the light of scientific progress. Although the so-called 'new atheists' Richard Dawkins (2006), Daniel Dennett (2006), Sam Harris (2004), and Christopher Hitchens (2007) are, in view of sociological data, well aware that religion will not easily disappear from the globe, they nonetheless express their hope that it would wither away – and that the light of reason would further conquer the world. What distinguishes the philosophical debates about secularization from the sociological debates – despite many intersections – is the more overtly normative tone.

Max Weber (1905) also took the rationalization of culture as a key element of modernity. But unlike Comte, he deems religion capable of self-rationalization. Monotheistic religious traditions vehemently criticize magical practices of the so-called 'natural' religions. Monotheistic thinking thus breeds the possibility of modern science, eventually leading to what Weber famously called the disenchantment or 'Entzauberung' of the world. However, Weber is not unequivocally happy about disenchantment. He blames scientific rationality and its instrumental logic for creating an 'iron cage' around human society, analogous to what *Charles Taylor* (2007) has called a disembedded society, consisting of 'buffered selves.' Weber's link between capitalism and protestant ethics can be regarded as well as an example of the internal capacity of religion for self-transformation, and a warning against simplifying theories which state that 'science has substituted religion' – or will eventually do so.

In line with Weber, and also with political philosopher *Carl Schmitt*, *Karl Löwith* questioned modernity and its self-acclaimed originality. In *Meaning in History* (1949), Löwith coined a new (philosophical) secularization paradigm, that of *modernity as an illegitimate transfer* of previously religious elements to a worldly level. The transfer is illegitimate, because one cannot untie secular concepts or practices from their religious framework, without perverting the original purpose of the religious concepts and practices. For example, the Christian concept of 'hope' was deeply intertwined with a transcendent, eschatological perspective on life after death. It was by no means directed toward an inner-worldly total solution for suffering. Redemption was solely possible in the afterlife. It was only with the Franciscan movement inaugurated by Joachim of Fiore (1135–1202) that the priority of

transcendent eschatology over inner-worldly affairs was questioned. Löwith considers modernity as the heir of this Franciscan movement. When modernity transforms Christian hope into the modern idea of progress, totalitarian ideologies see their chance to promote inner-worldly goals leading to ultimate salvation *in this life*. To underline their undeniable kinship with religious schemes, *Marcel Gauchet* has called these totalitarians views – communism, fascism – in line with *Eric Voegelin* 'secular religions.'

Hans Blumenberg opposed Löwith's paradigm of secularization. He defends modernity's original claims in his classic book *The Legitimacy of the Modern Age* (1966, transl. 1983). Secularization did not come about by an illegitimate transition from Christian hope to the modern idea of progress. Medieval eschatology was not primarily based on the hope for salvation, but on the fear of condemnation. This fear was theologically deepened by the nominalist account of a voluntarist and almighty God, who could at any time alter the 'ratio' of his own creation arbitrarily and totally whimsically. Hence, man could no longer be considered as the image of a God whose almighty will and thinking infinitely transcend the human condition, nor could nature be seen as the reflection of a divine plan of creation. As a result, man was left on his own, squeezed between an almighty but capricious *deus absconditus* on the one hand, and a whimsical and irrational nature on the other. According to Blumenberg, this predicament gave rise to the modern idea of self-assertion ('Selbstbehauptung'): man had to enhance his unenviable position by becoming master and commander of nature by his own rationality, and not by understanding God's rationality. Modern science developed in this context as a perfectly legitimate endeavor.

In more recent decades, philosophers like Taylor and Gauchet, and (philosophically oriented) sociologists like *Robert Bellah* have clearly widened the scope of the Löwith – Blumenberg debate. This debate revolved around the relation between Christianity and modernity in very specific historical figures: that of Joachism in the case of Löwith, the figure of nominalism in Blumenberg's account. While these two perspectives on secularization remain true for their part, Taylor, Gauchet, and Bellah esteem that they only tell a relatively small part of the story. Ground for this suspicion lies in *Karl Jaspers' Vom Ursprung und Ziel der Geschichte* (1949; Eng. transl. 1953), where Jaspers questioned Christianity's centrality in the rise of modernity by pointing to the importance of what he coined as the era of 'axial religion,' in the period of about 800–200 BC. Bellah, Gauchet, and Taylor continue to recognize the pivotal importance of Christianity, but place that importance in a larger story about the evolution of religion.

In *The Disenchantment of the World*, Gauchet considers secularization on a political level as the irreversible transition from heteronomy to autonomy. Gauchet states that this political autonomy could only arise from deeper ontological evolutions. He claims somewhat paradoxically that the 'reign of the invisible' or 'transcendence' first needs to be articulated more deeply, in order to eventually drift away from the 'visible' or 'immanent' world. The axial age and the Judeo-Christian tradition precisely constitute historical stages that articulate ontological dualism, and thus make it possible to ultimately separate the transcendent from the immanent. Gauchet follows Weber in his description of this evolution as 'disenchantment.' Gauchet sees Christianity as the penultimate stage of disenchantment (prior to the Enlightenment, with the rise of modern science, historical consciousness, and the French Revolution), depicting Christianity as 'the religion for the departure from religion' – *la religion de la sortie de la religion*. He calls the process of disenchantment and secularization, resulting in political autonomy, irreversible.

Gauchet's view does not imply that religion has become impossible. His claim is that religion has lost its key role in the level of public society, and will never be able to reclaim that position. Nor does Gauchet assert that the ongoing development of autonomy will result in a world without heteronomy. On the contrary, since heteronomy is a structural feature of the human being, it will continue to be part of our human experience, but it need not take the form of religion.

Taylor is more prudent in his considerations about the public role of religion in the future. In his *opus magnum* on secularization, *A Secular Age*, he describes the vicissitudes of religion as a 'zigzag account,' to be opposed to linear theories about the gradual decline of religion in human history. His purpose is not to describe secularization empirically, be it as the decline of religion in the public sphere ('secularity 1') or as the decay of personal religious beliefs ('secularity 2'). Rather he wants to offer a cultural analysis on the transcendental level: which conditions made it possible for Western culture to evolve from a clearly religious age around AD 1500 to a secular age in our times?

According to Taylor, one cannot tell a linear 'subtraction story' about, for instance, science and religion, where science is held to make religion obsolete by taking over its truth function. Such subtraction story argues that once science has displaced religion as the center of knowledge about the universe, the true nature of human existence is finally revealed, and it has nothing to do with religion.

Taylor chides this simple picture for two reasons. First, the subtraction story of positivism, scientism, or new atheism ignores the essential role of religious transformations in the genesis of modern science. The history of religion proves that religion is itself capable of rationalization (the Weberian thesis). For instance, the axial period of religious history, constituting a revolution along rational and ethical lines, raises questions as: how can the Gods behave as capriciously as human beings? How can good and evil be mingled in a divine nature? Later on, the period of Protestant Reform (sixteenth century) inaugurated what *Michel Foucault* has called the 'disciplinization' of culture. This new culture of discipline certainly contributed to the rigid methods of modern science.

Second, religion and science are not inversely proportional quantities. Religion also benefited from the new culture of discipline, self-reflectivity, and austerity. In the nineteenth century and the first half of the twentieth century, Christianity competed with socialism and liberalism for the soul of the supposedly enlightened individual. Religion probably never thrived to a similar degree as in this 'age of mobilization.' It is just one example of why the story of secularization is rather a zigzag account than a linear development.

Although Taylor's and Gauchet's genealogies of modernity and secularization broadly converge on an empirical and transcendental level, their ways part when it comes to their appreciation of secularization and modernity. Gauchet's linear story

logically results in his thesis about the irreversibility of secularization in the public sphere. Taylor's zigzag account also 'logically' leads to the opposite thesis. With Bellah (and Hegel), Taylor believes that 'nothing is ever lost.' A certain historical epoch has always inherited a lot from and was always made possible by previous eras. The historic sediments of foregone ages may always be dug up again, and may provoke us to criticize our contemporary ways of living. Hence, Taylor believes that with the high degree of discipline that dominates our culture, we tend to lose sight of religious, festive, and communal forms of life. It is by no means predictable which road the future will take: will our society eventually ban all reference to religion, or will it rediscover the richness of the religious past?

Bellah's dictum that 'nothing is ever lost' also echoes, be it inadvertently, in much of contemporary deconstructionist philosophy. *Jean-Luc Nancy*, for example, develops in *La déclosion* (2005, Engl. transl. 2007) a deconstruction of Christianity that draws attention to secular elements inherent to Christianity – which ultimately is a religion of self-deconstruction. Since the secular is essentially a product of monotheistic religions, it can never be fully understood without reference to religion. Hence, it is precisely religion that withstands any full definition of 'the secular.'

A Secular or Post-Secular Age?

Given that the once hegemonic status of the classic secularization paradigm is under fire, not only outside the West or in North America, but even in the heartlands of the classic secularization paradigm, Western Europe, it is no coincidence that the German philosopher Jürgen Habermas coined the term 'post-secular.' Over the course of the past decade, that very term has gained acknowledged currency. It seems to imply that we are experiencing a new era, by leaving behind our secular existence and simultaneously rejecting, on a theoretical level, the once dominant secularization theory.

However, the concept 'post-secular' (as coined by Habermas and many others) is not neutral. Actually, it presupposes a classic interpretation of the term secular, that is, as a narrative in which secularization is synonymous to the disappearance of religion (Habermas, 2005).

As already mentioned, a significant number of sociologists and philosophers currently dispute that classic presupposition. That is, for instance, the central reason why Taylor refuses to use that very term in *A Secular Age*. Only once does he make use of the term, when referring to the situation in 'post-secular' Europe (Taylor, 2007, pp. 534–535). Taylor's alternative master narrative in *A Secular Age* argues, first, that religion has not disappeared in the secular age, but that it has become more fragile, an embattled option among other options. Second, he claims that we have become secular due to the rise of a self-reflexive concept of the subject precisely inaugurated by religion (cf Gauchet, 1997), and that (contra Gauchet, 1997), therefore, a secular society has never been completely devoid of religion but that religion has remained vital and has produced ever more diverse forms of self-experience and self-reflective attitudes, that is, vicarious forms of and approaches toward religion. In other words, "Charles Taylor can be read as arguing that only a good secularist can be a good Christian, and only a Christian can be a good secularist." (Mendieta in Gorski et al., 2012, p. 308)

This controversy around the concept 'post-secular' is strongly related to the problem of the place of religion in the public sphere and the role of religious vis-à-vis secular language (Mendieta and VanAntwerpen, 2011). Exclusionists take the position that only secular arguments and motives count in the public debate and that, consequently, religious argumentation has to be excluded (Audi, 2000). Inclusionists claim that there is no need for a proviso, even not regarding the possibility that a religious majority can take public decisions, legally legitimized, on the basis of pure religious argumentation. (Weithman, 2002; Audi and Wolterstorff, 1997). Habermas, Rawls, and Taylor take mid-positions between those of exclusionists and inclusionists, albeit from different interpretations of the concepts of the 'secular' and the 'post-secular' (Rawls, 2001; Habermas, 2008; Mendieta and VanAntwerpen, 2011).

Secularization: The West versus the Rest

Whether one designates our contemporary age as 'secular' or as 'post-secular,' the growing sensitivity for the complexity of the concept of secularization in the Western world is even intensified, when it is transferred to non-Christian religions and to other areas in the world. The case of India is especially instructive in this regard. Since its independence from Britain, secularism has been a defining feature of its political system, whereas at the same time, it has become a deeply contested concept. A whole gamut of scholars claim that Western secularism, imposed on Indian society as a hegemonic ideology indebted to Protestant Reformation, is not viable for Indian society. They argue that minority rights for the traditional believers and religious toleration offer a more reliable guarantee for ethnic and religious tolerance than secularism does. In a similar vein, analogous questions and controversies apply to the worldwide problem of migration in general and of Muslim migration in Europe in particular (Gorski and Altimordu, 2008, pp. 72–73; Bhargava, 1998).

In order to cope with the difficulty of finding a concept covering the worldwide relations between religion and contemporary societies, the classic category of secularization hardly seems to be applicable. Since religions such as Confucianism or Taoism, for instance, have, in a sense, always been 'wordly' and 'lay,' the Western concept of secularization does not make much sense as a descriptive account of their recent evolution (Casanova, 2006, pp. 12–13). Therefore, a number of scholars (Eisenstadt, 2003; Casanova, 2006; Taylor, 2007) have developed the concept of 'multiple modernities' as a more adequate tool to designate global trends in this respect compared to either secular or post-secular cosmopolitanism or a clash of civilizations between religious and secular regimes (Casanova, 2006, p. 13).

Those who support this 'multiple modernities' position reject both the position of a modern radical break with religious traditions (classic or neoorthodox secularization paradigm) as well as the position that defends an essential continuity with tradition. Traditions are forced to adjust to modern conditions, in the process of which they also help to shape the particular forms of modernity (Casanova, 2006, p. 14).

The lively discussion about topics like 'multiple modernities' and the relation between 'secular' and 'post-secular' shows that the original self-evident belief in the inevitable decline and eventual extinction of religion has made room for a versatile debate not only around a whole gamut of specific aspects of the relation between modernity and religion, but even for a renewed conceptual analysis of the term 'secularization' itself: "Which is to say that the definition of the secular and its relationship to the religious are as hotly disputed now as ever, and that the scope of the debate is probably wider than ever. Whether this period of contestation marks the beginning of a post-secular age, or merely a period of secular ebb and religious flow, only time will tell." (Gorski and Altimordu, 2008, p. 77).

See also: American and French Revolutions: Impact on the Social Sciences; Christianity, Origins of; Church and State; Civil Religion; Civil Society and the Public Sphere: History of the Concept; Comte, Auguste (1798–1857); Conceptual History: Begriffsgeschichte; Eisenstadt, Shmuel N. (1923–2010); European Religion from the Atlantic to the Black Sea; Globalization and Religion; Habermas, Jürgen (1929–); Liberal Christianity; Luhmann, Niklas (1927–98); Modernity: History of the Concept; Modernity; Modernization, Sociological Theories of; Multiculturalism; Pluralism and Tolerance; Political Religions: History of the Concept; Progress: History of the Concept; Rationality in Society; Religion and Politics in the United States; Religion, History of; Religion, Sociology of; Religion: Evolution and Development; Religion: Mobilization and Power; Schmitt, Carl (1888–1985); Science and Religion; Secular Religions; Secularization; Taylor, Charles (1931–); Translating Sociology into Japanese, Chinese and Korean Languages; Unintended Consequences: History of the Concept; Weber, Max (1864–1920); Weberian Social Thought, History of.

Bibliography

Audi, R., 2000. Religious Commitment and Secular Reason. Cambridge University Press, Cambridge; New York.

Audi, R., Wolterstorff, N., 1997. Religion in the Public Square. Rowman and Littlefield, Lanham, MD.

Baehr, P.R., Wells, G.C., 2002. The Protestant Ethic and the "Spirit" of Capitalism and Other Writings. Penguin Books, New York.

Bellah, R., 2011. Religion in Human Evolution. From the Paleolithic to the Axial Age. Harvard U.P, Cambridge, MA.

Berger, P., 1969. The Sacred Canopy: Elements of a Sociological Theory of Religion. Doubleday, Garden City, NY.

Berger, P., 1999. The desecularization of the world: a global overview. In: GWeigel, D. Martin, Sacks, J., Davie, G., Weiming, T., An-Na'im, A.A. (Eds.), The Desecularization of the World: Resurgent Religion and World Politics. Eerdmans, Grand Rapids, MI, pp. 1–18.

Bhargava, R., 1998. Secularism and Its Critics. Oxford University Press, New Delhi.

Blumenberg, H., 1983. The Legitimacy of the Modern Age. MIT Press, Cambridge, MA.

Bruce, S., 1992. Religion and Modernization: Sociologists and Historians Debate the Secularization Thesis. Oxford University Press, Oxford.

Bruce, S., 2002. God Is Dead: Secularization in the West. Blackwell, Oxford, UK.

Bruce, S., 2006. Secularization and the impotence of individualized religion. Hedgehog Review 8, 35–45.

Bruce, S., 2011. Secularization: In Defence of an Unfashionable Theory. Oxford University Press, Oxford.

Casanova, J., 1994. Public Religions in the Modern World. University of Chicago Press, Chicago.

Casanova, J., 2006. Rethinking secularization: a global comparative perspective. Hedgehog Review 8, 7–22.

Davie, G., 1994. Religion in Britain since 1945: Believing without Belonging. Blackwell, Oxford.

Davie, G., 2000. Religion in Modern Europe: A Memory Mutates. Oxford University Press, Oxford.

Dawkins, R., 2006. The God Delusion. Houghton Mifflin, Boston.

Dennett, D., 2006. Breaking the Spell: Religion as a Natural Phenomenon. Allen Lane, New York.

De Vriese, H., 2010. The charm of disenchantment: a quest for the intellectual attraction of secularization theory. Sophia 49 (3), 407–428.

Dobbelaere, K., 1999. Towards an integrated perspective of the processes related to the descriptive concept of secularization. Sociology of Religion 60 (3), 229–247.

Eisenstadt, S.N., 2003. Leiden and Boston. In: Comparative Civilizations and Multiple Modernities, 2 vols. Brill, Leiden and Boston.

Ferry, L., Gauchet, M., 2004. Le Religieux Après la Religion. Bernard Grasset, Paris.

Finke, R., Stark, R., 1988. Religious economies and sacred canopies: religious mobilization in American cities, 1906. American Sociological Review 53, 41–49.

Finke, R., Stark, R., 1989. How the upstart sects won America: 1776–1850. Journal for the Scientific Study of Religion 28, 27–44.

Finke, R., Stark, R., 1992. The Churching of America, 1776–1990: Winners and Losers in Our Religious Economy. Rutgers University Press, New Brunswick, NJ.

Finke, R., Stark, R., 1998. Religious choice and competition. American Sociological Review 63, 761–766.

Gauchet, M., 1997. The Disenchantment of the World: A Political History of Religion. Princeton University Press, Princeton, NJ.

Gauchet, M., 1998. La Religion dans la Democratie: Parcours de la Laïcité. Gallimard, Paris.

Girard, R., 1978. Des choses cachées depuis la fondation du monde. Grasset, Paris.

Girard, R., Vattimo, G., 2010 (W. McCuaig, Trans.). In: Antonello, Pierpaolo (Ed.), Christianity, Truth, and Weakening Faith: A Dialogue. Columbia University Press, New York.

Gorski, P.S., Kyuman, D., Torpey, J., VanAntwerpen, J., 2012. The Post-Secular in Question: Religion in Contemporary Society. NYUP, New York.

Gorski, P.S., Altimordu, A., 2008. After secularization? Annual Review of Sociology 34, 55–85.

Gorski, P.S., 2003. Historicizing the secularization debate: a program for research. In: Dillon, M. (Ed.), Cambridge Handbook for the Sociology of Religion. Cambridge University Press, New York; Cambridge, UK, pp. 110–122.

Gregory, B., 2012. The Unintended Reformation. How a Religious Revolution Secularized Society. Harvard University Press, Cambridge, MA.

Habermas, J., 2006. Religion in the public sphere. European Journal of Philosophy 14, 1–25.

Habermas, J., 2008. Notes on a post-secular society. New Perspectives Quarterly 25 (4), 17–29.

Habermas, 2005. Zwischen Naturalismus und Religion. Philosophische Aufsätze. Suhrkamp, Frankfurt am Main.

Habermas, J., Mendieta, E., 2002. Religion and Rationality: Essays on Reason, God, and Modernity. MIT Press, Cambridge, MA.

Habermas, J., Reemtsma, J.P., 2001. Glauben und Wissen: Friedenspreis des Deutschen Buchhandels 2001. Suhrkamp, Frankfurt am Main.

Harris, S., 2004. The End of Faith: Religion, Terror, and the Future of Reason. Norton, New York.

Hervieu-Léger, D., 1999. Religion en Mouvement: le Pelerin et le Converti. Flammarion, Paris.

Hervieu-Léger, D., 2000. Religion as a Chain of Memory. Polity, Cambridge, UK.

Hervieu-Léger, D., 2001. La Religion en Miettes ou la Question des Sectes. Calmann-Lévy, Paris.

Hervieu-Léger, D., Champion, F., 1986. Vers un Nouveau Christianisme? Introduction 'a la Sociologie du Christianisme Occidental. Cerf, Paris.

Hervieu-Léger, D., Martine Cohen, Erwan Dianteill, Isabelle Saint-Martin, dirs, 2004. La modernité rituelle. Rites politiques et religieux des sociétés modernes. Paris, L'Harmattan, coll. Religions en questions.

Hitchens, C., 2007. God Is Not Great: How Religion Poisons Everything. Twelve, New York.

Inglehart, R., Norris, P., 2004. Sacred and Secular: Religion and Politics Worldwide. Cambridge University Press, New York.

Jaspers, K., 1953. The Origin and Goal of History. Routledge & Kegan Paul, London.

Joas, H., 2007. Säkularisierung und die Weltreligionen. Fischer Taschenbuch Verlag, Frankfurt am Main.

Lehmann, H., 2004. Säkularisierung. Der Europäische Sonderweg in Sachen Religion. Wallstein, Göttingen.

Löwith, K., 1949. Meaning in History the Theological Implications of the Philosophy of History. University of Chicago Press, Chicago, IL.
Luckmann, T., 1967. The Invisible Religion: The Problem of Religion in Modern Society. MacMillan, New York; London.
Luckmann, T., 1996. The privatization of religion and morality. In: Heelas, P. (Ed.), Detraditionalization. Basil Blackwell, Oxford, pp. 72–86.
Micklethwait, J., Wooldridge, A., 2009. God Is Back. How the Global Revival of Faith Is Changing the World. The Penguin Press, New York.
Milbank, J., 2014. Beyond Secular Order: The Representation of Being and the Representation of the People. Wiley-Blackwell, Oxford.
Monod, J.-C., 2002. La Querelle de la Sécularisation: Theologie Politique et Philosophies de l'Histoire de Hegel à Blumenberg. Vrin, Paris.
Mendieta, E., VanAntwerpen, J. (Eds.), 2011. The Power of Religion in the Public Sphere. Columbia University Press, New York.
Nancy, J.-L., 2007. Dis-Enclosure: The Deconstruction of Christianity. Fordham University Press, New York.
Paul, G.S., 2005. Cross-national correlations of quantifiable societal health with popular religiosity and secularism in the prosperous democracies. A first look. Journal of Religion & Society 7, 1–17.
Paul, G.S., 2010. The evolution of popular religiosity and secularism: how first world statistics reveal why religion exists, why it has been popular, and why the most successful democracies are the most secular. In: Zuckerman, Phil (Ed.), Atheism and Secularity: Volume 1: Issues, Concepts and Definitions. Praeger, Santa Barbara, pp. 149–208.
Rawls, J., 2001. The idea of public reason revisited. In: The Law of Peoples. Harvard University Press, Cambridge, MA.
Stark, R., 1999. Secularization, RIP. Sociology of Religion 60, 249–273.
Stark, R., Bainbridge, W.S., 1985. The Future of Religion: Secularization, Revival, and Cult Formation. University of California Press, Berkeley.
Stark, R., Finke, R., 2000. Acts of Faith: Explaining the Human Side of Religion. University of California Press, Berkeley.
Stark, R., Iannaccone, L.R., 1994. A supply-side reinterpretation of the so-called secularization of Europe. Journal for the Scientific Study of Religion 33, 230–252.
Taylor, Ch, 2007. A Secular Age. Harvard University Press, Cambridge, MA.
Vattimo, G., 2002. Dopo la cristianità. Per un cristianismo non religioso. Garzanti, Milaan.
Wallace, Anthony F.C., 1966. Religion: An Anthropological View. Random House, New York.
Weithman, P., 2002. Religion and the Obligations of Citizenship. Cambridge University Press, Cambridge.
Weber, M., 1905. Die protestantische Ethik und der Geist des Kapitalismus.

Relevant Websites

http://blogs.ssrc.org/tif/.
http://www.pewforum.org/.

Securitization

Philippe Bourbeau, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Securitization is one of the most dynamic fields of research in today's security studies. This article (1) examines the various definitions developed by scholars to make sense of the phenomenon, (2) presents the two logics upon which most of the research is conducted (i.e., the logic of exception and the logic of routine), and (3) identifies avenues of research that will most likely drive future research in this field.

Securitization is one of the most dynamic fields of research in today's security studies. Several sets of arguments, factors, rationales, and explanations have been offered to better understand and explain mechanisms of securitization. Constructivists, postmodernists, political sociologists, neoliberalists, critical theorists, feminists, and normative theorists are all participating in the debate. This article is divided in three parts. First, I will examine the various definitions of securitization currently proposed by scholars. The second section will analyze the major line of cleavage among theorists of securitization, i.e., the logics of securitization. Finally, I will seek to identify future avenues of research that although by no means exhausting the possibilities will most likely drive most of the research in years to come.

Definitions of Securitization

If scholars find it hard to reach a consensus on a definition of security, they find it equally difficult to agree on a definition of securitization. A first definition was proposed by Ole Wæver (1995) in his seminal paper in which he argues that security is an intersubjective and socially constructed process and that security is about survival. Securitization seeks thus to understand the processes by which "an issue is presented as posing an existential threat to a designated referent object." In that sense, securitization studies aim to gain a better understanding of "who securitizes, on what issues (threats), for whom (referent objects), why, with what results, and, not least, under what conditions" (Buzan et al., 1998: p. 31).

This definition of securitization is contested. In the same year of the publication of Barry Buzan, Ole Wæver, and Jaap de Wilde important book, Didier Bigo proposed a different take (the paper was first published in French in 1998 and then translated into English in 2002). Focusing on the issue of migration, he argues that "the transformation of security […] is directly related to [security professionals] own immediate interests […] and to the transformation of technologies they use […]. The Europeanization and the Westernization of the logics of control and surveillance of people beyond national policies is driven by the creation of a transnational field of professionals in the management of unease. This process of securitization is now well known" (Bigo, 2002: p. 64).

Influenced by the work of Bigo, Thierry Balzacq (2005: p. 172) proposed another definition a few years later. In an important contribution to this field of research, he argues that securitization is "a strategic (pragmatic) practice that occurs within, and as part of, a configuration of circumstances, including the context, the psycho-cultural disposition of the audience, and the power that both speaker and listener bring to the interaction." Balzacq has recently slightly modified his definition of securitization in proposing a short and a long definition. For him, securitization is, "a set of interrelated practices, and the processes of their production, diffusion, and reception/translation that brings threats into being" and "an articulated assemblage of practices whereby heuristic artefacts (metaphors, policy tools, image repertoires, analogies, stereotypes, emotions, etc.) are contextually mobilized by a securitizing actor, who works to prompt an audience to build a coherent network of implications (feelings, sensations, thoughts, and intuitions), about the critical vulnerability of a referent object, that concurs with the securitizing actor's reasons for choices and actions, by investing the referent subject with such an aura of unprecedented threatening complexion that a customized policy must be undertaken immediately to block its development" (Balzacq, 2011: p. 3).

Building on these definitions, I have proposed to understand securitization as the process of integrating discursively and institutionally an issue into security frameworks that emphasize policing, control, and defense (Bourbeau, 2013b). Whereas the discursive element of this definition refers not only to the use of language but also to practices associated with language, the institutional dimension refers to notions of acceptance and the mechanisms of endurance, path dependence, and continuity in socially constructing issues as security ones.

Despite these disputes regarding definitions of securitization, the main objectives of securitization scholars remain to better understand and explain this particular social process. For that purpose, scholars have mainly proposed two logics of securitization.

Logics of Securitization

As currently organized, the literature on the securitization process rests mainly on two logics: the logic of exception and the logic of routine. These logics should not be viewed as opposing ones, as they are not competing views. In fact, it is not at all clear that they can compete, for it is not clear that they are mutually exclusive.

Building primarily on the work of Carl Schmitt, the logic of exception postulates that security is about the fight against an existential threat that necessitates exceptional measures. It focuses on speech acts legitimizing exceptional policies and practices in the face of an existential security threat. The so-called Copenhagen model (or school) best illustrates the logic of exception and has become over the years the current benchmark in securitization research (Buzan et al., 1998; Wæver, 1995). Bringing speech act theory into security studies, the Copenhagen model posits that labeling something as a security issue permeates it with a sense of importance that legitimizes the use of emergency measures outside of the usual political processes to deal with it. It is by labeling a phenomenon a security issue that it becomes one.

The model distinguishes two types of actors: a securitizing actor securitizing issues by declaring something existentially threatened and a functional actor affecting the dynamics of a sector. Securitizing agents do not navigate freely toward successful securitization. In fact, "conditions for a successful speech-act fall into two categories: (1) the internal, linguistic-grammatical – to follow the rules of the act (or, as Austin argues, accepted conventional procedures must exist, and the act has to be executed according to these procedures); and (2) the external, contextual and social – to hold a position from which the act can be made" (Buzan et al., 1998: p. 32). On top of these facilitating conditions, the Copenhagen model argues in a bold statement that "discourse that takes the form of presenting as an existential threat to a referent object does not by itself create securitization – this is a securitizing move, but the issue is securitized only if and when the audience accepts it as such … successful securitization is not decided by the securitizer but by the audience of the security speech-act" (Buzan et al., 1998: p. 25).

In the past decades, the Copenhagen model has provoked heated debates and has captured the attention of an increasing number of scholars: Articles analyzing the model's strengths and weaknesses have been published as leading article in major international relations journals, and panels have been organized at several annual conventions of the International Studies Association. While most security studies theorists disagree with at least one aspect of the Copenhagen model, they nearly all recognize that this particular take on the securitization process is one of the most creative, rich, and stimulating models in today's security studies.

Debates about the Copenhagen model have evolved through several interrelated phases, and major lines of criticisms have touched on several elements of the model. Scholars have criticized the model for fixing collective identity and have questioned the responsibility of the analysts in studying security. Others have focused their criticism on the speech act theory and its applicability to security studies arguing that the model underestimates other powerful practices such as the production of image and the use of symbols (Hansen, 2011; Williams, 2003) or that the model fails to include other categories or types of acts (Balzacq, 2005; Huysmans, 2006; Vuori, 2008). Scholars attuned to poststructural and gender perspectives have underscored that the model failed to take into account the silenced voices (Hansen, 2000). Others have argued that the model left the audience significantly under-theorized (Balzacq, 2005; Salter, 2008). Whether the model can include the issue of context in its framework has been analyzed (Bourbeau, 2011; McDonald, 2008; Stritzel, 2007). Others have focused their criticisms on the lack of universal and transcultural understanding of "the established rules of the game" (Curley and Wong, 2008). Still others have argued that the Copenhagen model, as currently organized and applied, cannot explain variation in levels of securitization and that it presents a thin understanding of the concept of power, particularly the power of contextual factors – which should be understood as the power to enable and/or constrain securitizing agents (Bourbeau, 2011).

In contrast, the logic of routine inspired by Michel Foucault's and Pierre Bourdieu's systems of thoughts sees securitization as a process of establishing and inscribing meaning through governmentality and practices. It understands the securitization process as routinized and patterned practices, particularly bureaucracies' and security professionals' practices, in which technology comes to hold a prominent place. The logic of routine comprises a broader set of security scholars than for the logic of exception. Each scholar brings to the table his own focus and influences, and although the logic is far from consistent across scholars, it cannot be ignored either.

Bigo is a central figure of the logic of routine. He argues that security is not necessarily about survival nor is it about urgency and exceptional measures. Rather, security is largely defined by mundane bureaucratic decisions and practices that create a sense of insecurity and unease. Using Bourdieu's concepts of field and habitus, Bigo contends that the field of insecurity while breaking the traditional internal–external understanding of security absorbs other fields, and in doing so produces a security continuum in which various agents come to interact under the umbrella of this field. Focusing on the case of migration, Bigo argues that the securitization of immigration emerges "from the correlation between some successful speech acts of political leaders, the mobilization they create for and against some groups of people, and the specific field of security professionals […] It also comes from a range of administrative practices […] and what may be termed a specific habitus of the 'security professional' with its ethos of secrecy and concern for the management of fear or unease" (Bigo, 2002: p. 67). Bigo's work has been very influential, and several scholars have developed and further expanded his set of arguments.

Following in these footsteps, Balzacq has recently argued that the securitization process is better understood as a strategic (pragmatic) process occurring within, and as part of, a configuration of circumstances in which context, audiences, and power relations play an important role. For him, security performatives are situated actions mediated by agents' habitus. Seeking to "reformulates the assumptions of securitization in a form appropriate to empirical studies and to the development of a comprehensive theory," Balzacq (2011: p. 1) proposes three core assumptions: the audience is central in the securitization process, the agent and context are codependent, and the *dispositif* connects different practices.

The logic of routine has so far been subjected to less critical engagements than the logic of exception (and particularly the Copenhagen model). Yet, friendly criticisms are starting to emerge. Michael C. Williams (2011), for example, argues that if

security is defined away from an explicit logic of extremity, then there is a danger that it will become difficult to distinguish any number of policy changes from ones that specifically relate to security. He further points out that "while the language of emergency can appear an obstacle to more theoretically nuanced and empirically grounded understanding of concrete practices, it may at the same time provide an anchoring device through which 'security' dynamics can be discerned and distinguished from 'normal' change with the policy process" (Williams, 2011: p. 217). Elsewhere, I have criticized the logic of routine for its understanding of who can be a securitizing actor and for offering very little guiding principles in explaining variation in levels of securitization (Bourbeau, 2014).

Future Avenues of Research

The nature, mechanisms, contours, and consequences of the securitization process are one of the most active and fast growing areas of research of global politics. Undoubtedly, there is substantial disagreement among scholars that this short article cannot fully engage with. Yet, future avenues of research (or of disputes) can be divided into four broad categories: the politicization versus the securitization process, the issue of methods in securitization research, the various strategies of contesting security, and varieties of issues under investigation.

The Politicization versus the Securitization Process

Distinguishing between the politicization process and the securitization process is an important step in our study of contemporary security politics. There is substantial disagreement as to whether the two processes are in fact distinct. Arguments on both sides of the spectrum have been put forward: from the somewhat radical position that security is in fact placing an issue beyond normal politics to the equally radical position that security is the normal politics. Yet, highlighting the importance of distinguishing the two processes does not mean that there is no politics in the securitization process. To take international migration as example, it is one thing to question the refugee determination process or the adequacy of immigrant settlement services or the over-concentration of immigrants in large cities. It is an altogether different one to argue that international migration is a national security threat to a given society or a state.

In an effort to stimulate a richer dialogue between the literature on the politicization process and the literature on the securitization process, I have recently put forward the argument that the two processes are two distinct processes but they are not mutually exclusive, for it is not clear that they are competing forces (Bourbeau, 2013a). Several hypotheses could guide future research: Can a negative politicization of an issue induce the securitization process, idiosyncratically as well as over a longer time period? Does a negative politicization necessarily lead to the securitization? Does the relationship between the two is a question of intensification: The more 'intense' the politicization is, better are the chances that it will become a security issue? If so, what is the threshold and how do we recognize it?

Williams (2012) has been one of the first to propose the intensification hypothesis. He argues that the concept of intensification may hold some promise as a means of moving beyond the division between scholar attuned to the logic of exception and advocates of a pragmatic-routine approach. Normal politics, unease, and thresholds could be seen as forms of intensification below extremity. Yet, the idea of a fundamental breaking of rules remains central to security analysis, argues Williams, precisely because it provides a much-needed limit condition.

Methods

How do we know when an issue has been securitized? This seemingly simple question constitutes one of the thorniest issues in contemporary security studies. This is best exemplified with the case of international migration. Some almost take it for granted; others strive to explain the securitization of migration without first demonstrating whether migration is securitized or not – either globally or in the country/cases under investigation. As a result, contradictory standpoints have been presented with little methodological consensus on which to build on. Indeed, whereas the majority of scholars have argued that migration has been successfully securitized in most OECD countries, others, on the contrary, have found little evidence to the effect that migration was securitized in Europe.

The literature is slowly making steps in correcting this situation, however. Two paths out of the wood can be identified. A first one is to better understand the phenomenon under study. In my comparative analysis of the securitization of migration in Canada and France, I have offered two categories of indicators to gain a better understanding of whether migration is securitized (and to what extent) or not in Canada and France. Both sets of indicators offer nominal and ordinal measurements (Bourbeau, 2011).

The first category is institutional indicator, which includes whether immigration acts mention in one way or another security provisions or not and whether official policy statements that relate to security, foreign affairs, and defense discuss immigration in security terms or not. To illustrate, whereas the previous official statement entitled *Competitiveness and Security: Direction for Canada's International Relations* published in 1985 was silent about the issue of international migration, the statement *Foreign Policy Themes and Priorities* issued in 1991 listed migration as a security threat to Canada. The inclusion of migration as a security concern was reiterated in the *White Paper on Defence* of 1994, *Canada in the World* statement of 1995, *Securing an Open Society: Canada's National Security Policy* of 2004, and *Canada's International Policy Statement* of 2005. The second category of indicators is security practices, which include interdiction policies and detention of immigrants. Using this broad range of indicators, Bourbeau demonstrates that migration is securitized in both Canada and France and that significant variations exist between the two country cases on the level of securitization, thereby opening up research possibilities to study the variation.

Indeed, the question of variation in securitization is a rather underdeveloped and undertheorized one in the literature. This might be because most of the empirical studies are case study and

not small-n or large-N comparative analysis. This fact remains that too few contributions try to make sense of the differences in levels of securitization (for example, between a thick/strong/high securitization and a thin/weak/low securitization – regardless in fact of the exact vocabulary one uses). Is it a question of number and/or strength of securitizing moves or degrees of reception by audiences (broadly defined)? If we accept Bigo's central argument that the habitus of the security professionals is the key driver of securitization, then does a variation in 'strength' of security professionals across cases explain a variation in levels of securitization? The more powerful security professionals are, the ticker/harder/stronger securitization we observe? Extending Williams' work (2012), is a variation in securitization a question of intensification? My comparative analysis offers guiding principles to account for the variation in the levels of securitization; I suggest that contextual factors, particularly when their power is understood as the capacity to enable and/or to constrain, offer a promising axiom of research into the issue of variation (Bourbeau, 2011).

A second option is to focus on methods of investigation of the securitization process. Attuned to comparative politics stress on the research design, I employ the logic of triangulation by which findings obtained with one research method are contrasted against findings obtained with another method (Bourbeau, 2011). Using a broad range of methodologies (from content analysis, to statistical analysis, to interviews, to surveys and polls) has proved successful in providing a high level of robustness to his study. Balzacq (2011) has also recently ventured into these method questions, albeit from a slightly different perspective and with a different set of questions in mind (and not necessarily in terms of international migration). He proposed to distinguish between four methodological techniques of securitization research: discourse analysis, ethnographic research, process tracing, and content analysis. Stefano Guzzini (2011) has also recently suggested a stimulating path: to use the idea of 'securitization' as a causal mechanism in process tracing and interpretivist methodology.

Despite these contributions, the literature on the securitization process remains overly theoretical with few comprehensive empirical studies within case across time and/or across cases. To fully establish itself as a dynamic and important area of research, securitization research will have to make inroads in better understanding to what extent an issue has been (or not) securitized and it will have to diversify its methodological techniques from the already well-covered content analysis and discourse analysis to small-n comparative analysis and even large-N statistical analysis.

Strategies of Contesting Securitization

What are the options for those wanting to contest a successful securitization? For a long time, scholars focused almost exclusively on one strategy: desecuritization. Initially conceptualized as somewhat the opposite of securitization, desecuritization quickly became a subject of disputes in security studies. The Copenhagen model understands desecuritization as the process of taking an issue out of the security/emergency mode and bringing it into the political sphere. Huysmans (1998) has proposed to see it as the unmaking of securitization. By the mid-2000, several scholars focused their attention on this apparently twin concept of securitization. Claudia Aradau's (2004) stimulating article has made important steps in debunking the meaning of desecuritization and has spurred an enriching dialogue on the question. She criticized the literature on desecuritization for tackling the issue analytically (and for doing so in vague terms), thereby eschewing that desecuritization is about politics and the type of politics we want. Rita Floyd (2011) has argued that desecuritization is not morally better than securitization. Paul Roe (2012) has recently revisited the normatively defined debate over securitization as a negative concept. In a superb contribution, Lene Hansen (2012) has defined desecuritization as the move of an issue out of the sphere of security and has proposed to recover the political status of desecuritization through four readings of desecuritization: (1) as a Derridarian supplement to securitization, (2) as a way of 'reworking' the friend/enemy distinction, (3) as an emphasis on responsibility and choice, and (4) as an illustration of the particular historical context of Cold-War détente.

Yet another strategy of contesting successful securitization is resilience. In recent years, a great deal has been written about the role of resilience in our social world. This scholarship has sparked vivid theoretical debates in psychology, criminology, social work, and political geography about the nature of resilience and how scholars should go about studying it. Security studies are a latecomer in the field, and while the term is sometimes employed, it is rarely debunked and theorized. Balzacq's (forthcoming) edited book takes important steps in correcting this. Resilience is one of the four strategies of contesting security he identified (the others are desecuritization, resistance, and emancipation). In addition, I suggest a theorization of resilience as applied to the securitization process in which he argues that current definition of resilience (as the positive adaptation within the context of significant adversity) is problematic because it eschews that resilience has a dark side and that it is always a matter of degree (Bourbeau, 2013b).

Going beyond Migration, Environment, and Global Health

Although most of the conversation has been done in theoretical terms, extensive empirical studies of the securitization process have been conducted mostly in the field of migration, environment, and health. Scholars have in fact conducted research on the migration–security nexus within case across time, across cases, and on almost all continents. A large strand of this literature was part of migration studies' focus on whether and to what extent the nation-state remains capable of regulating and controlling international migration. Other scholars have emphasized the effects (usually negative) of the securitization of migration on migrants themselves: some highlighting the interconnections of security practices and human rights considerations, others hypothesizing that the securitization of migration is the latest and most modern form of racism.

Environmental and energy security has also attracted the attention of several scholars. For some, the case of environment underscores the limit of the logic of exception and illustrates the need to adopt alternative approaches. Others have employed the environmental case to demonstrate the analytical strength of the Copenhagen model and to propose

a refined version of it (Floyd, 2011). Global health issues have also recently a fair amount of attention by scholars of the securitization process. Key issues include HIV/AIDS, SARS outbreak, and Asian influenza. One of the central questions that have driven most of the analysis is whether (and to what extent) these global health should be framed as security issue. Some have argued that although framing AIDS as an international security issue could raise international awareness and resources, it also runs the risk of pushing responses away from the civil society toward military and intelligence organizations (Elbe, 2006). Others have argued that domestic, political, and social context within states are the key drivers of responses to infectious disease in contrast with international discourse of health security, thereby making an important step in understanding the policy challenges of implementing the global discourse of infectious disease securitization at the domestic level (Curley and Herington, 2011).

Other nontraditional security issues need to be further studied, and some scholars have started to make inroad in this regard. Although cybersecurity, critical infrastructures, and securitization as a mechanism of the democratic peace have received insufficient attention so far (Hayes, 2012), they will most likely be at the center of several debates in securitization research for years to come.

See also: Deterrence; Insurgency; Militarism; War Propaganda; War, Sociology of.

Bibliography

Aradau, C., 2004. Security and the democratic scene: desecuritization and emancipation. Journal of International Relations and Development 7 (4), 388–413.

Balzacq, T., 2005. The three faces of securitization: political agency, audience and context. European Journal of International Relations 11 (2), 171–201.

Balzacq, T. (Ed.), 2011. Securitization Theory. How Security Problems Emerge and Dissolve. Routledge, London.

Balzacq, T. (Ed.), Security Contested: Strategies and Logics. Routledge, London, forthcoming.

Bigo, D., 2002. Security and immigration: toward a critique of the governmentality of unease. Alternatives 27, 63–92.

Bourbeau, P., 2013a. La politisation et la sécuritisation des migrations internationales: une relation à définir. Critique Internationale 61 (4), 127–145.

Bourbeau, P., 2013b. Resiliencism: premises and promises in securitization research. Resilience: International Politics, Practices and Discourses 1 (1), 3–17.

Bourbeau, P., 2014. Moving forward together: the logics of the securitization process. Millennium: Journal of International Studies 43 (1), 187–206.

Buzan, B., et al., 1998. Security: A New Framework for Analysis. Lynne Rienner Publishers, Boulder, CO.

Curley, M., Wong, S-I. (Eds.), 2008. Security and Migration in Asia: The Dynamics of Securitisation. Routledge, London.

Curley, M.G., Herington, J., 2011. The securitisation of infectious disease: international norms and domestic politics in Asia. Review of International Studies 37 (1), 141–166.

Elbe, S., 2006. Should HIV/AIDS be securitized? The ethical dilemmas of linking HIV/AIDS and security. International Studies Quarterly 50 (1), 119–144.

Floyd, R., 2011. Can securitization be used in normative analysis? Towards just securitization theory. Security Dialogue 42 (4–5), 427–439.

Guzzini, S., 2011. Securitization as causal mechanism. Security Dialogue 42 (4–5), 329–341.

Hansen, L., 2000. The little mermaid's silent security dilemma and the absence of gender in the Copenhagen school. Millennium: Journal of International Studies 29 (2), 285–306.

Hansen, L., 2011. Theorizing the image for security studies: visual securitization and the Muhammad cartoon crisis. European Journal of International Relations 17 (1), 51–74.

Hansen, L., 2012. Reconstructing desecuritisation: the normative-political in the Copenhagen school and directions for how to apply it. Review of International Studies 38 (6), 525–546.

Hayes, J., 2012. Securitization, social identity, and democratic security: Nixon, India, and the ties that bind. International Organization 66 (1), 63–93.

Huysmans, J., 1998. Revisiting Copenhagen: or, on the creative development of a security studies agenda in Europe. European Journal of International Relations 4 (4), 479–505.

Huysmans, J., 2006. The Politics of Insecurity. Fear, Migration and Asylum in the EU. Routledge, London.

McDonald, M., 2008. Securitization and the construction of security. European Journal of International Relations 14 (4), 563–587.

Roe, P., 2012. Is securitization a 'negative' concept? Revisiting the normative debate over normal versus extraordinary politics. Security Dialogue 43 (3), 249–266.

Salter, M.B., 2008. Securitization and desecuritization: dramaturgical analysis and the Canadian aviation transport security authority. Journal of International Relations and Development 11 (4), 321–349.

Stritzel, H., 2007. Towards a theory of securitization: Copenhagen and beyond. European Journal of International Relations 13 (3), 357–383.

Vuori, J.A., 2008. Illocutionary logic and strands of securitization: applying the theory of securitization to the study of non-democratic political orders. European Journal of International Relations 14 (1), 65–100.

Wæver, O., 1995. Securitization and Desecuritization. On Security. R.D. Lipschultz. Columbia University Press, New York, pp. 46–86.

Williams, M.C., 2003. Words, images, enemies: securitization and international politics. International Studies Quarterly 47 (4), 511–531.

Williams, M.C., 2011. The Continuing Evolution of Securitization Theory. Securitization Theory. How Security Problems Emerge and Dissolve. T. Balzacq. Routledge, London, pp. 212–222.

Williams, M.C., 2012. Securitization and the liberalism of fear. Security Dialogue 42 (4–5), 453–463.

Security: Commercial Transactions

Frank Decker, NSW, Australia
Sheelagh McCracken, The University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the legal framework within which debtors use their assets to secure payment or assure the performance of their obligations to their creditors. It analyzes the range of security interests recognized in major legal systems, starting with their development under Greek and Roman law. It then focuses on the modern position, drawing on the English Common Law as well as the French and German Civil Codes. It also describes a discernible trend toward a more functionalist approach, originating in Article 9 of the US Uniform Commercial Code. It concludes by placing the legal position in its economic context.

Introduction

Security Law, often also described as Secured Transactions Law, forms part of a legal jurisdiction's commercial law regime. It deals with interests in property that are created by way of security for payment or performance of an obligation. Although typically capable of securing any obligation, they are often given by borrowers to creditors in order to secure repayment of debt.

In lending transactions Security Law enables the provision of credit, in the knowledge that it may, in case of the borrower's default, be recovered through recourse to the borrower's property, which may be sold or otherwise dealt with, in order to recover the monies owed. The direct access to the particular property conferred by the security interest generally enables the lender to enforce the claim without first having to expend time and money in obtaining a court judgment. Most importantly, the lender's legal status as a secured creditor may in many legal jurisdictions confer protection in the event of the borrower becoming insolvent, enabling the secured creditor to realize the security rather than initiate insolvency proceedings. Lenders therefore frequently make the provision of security a prerequisite for the advance of a loan.

Security Law intersects with other areas of law, including contract law governing the underlying loan and security agreement; property law facilitating dealings with particular assets; and insolvency law impacting on the payment of creditors. Its operation is generally a matter of domestic law, with its scope and application in any legal jurisdiction depending on the rules of that jurisdiction. Historically, security law jurisdictions have been classified into three main groups: those of Anglo-American Common Law countries; Napoleonic civil law countries based on the French Civil Code; and Romano–Germanic countries based on the German Civil Code (Wood, 2007). Since the 1950s, however, American jurisdictions have adopted a broader functional approach to determining the scope of security interests; an approach whose influence has now begun to spread internationally and cut across some of the traditional groupings.

This article is divided into six sections. The first section outlines traditional features of security interests. The next section examines their early development under Greek and Roman law and is followed by an analysis of their current operation under the English Common Law regime and under the French and German Civil Code frameworks. The penultimate section describes the functionalist approach to Security Law and the final section places the legal discussion within its broader commercial context, emphasizing the economic importance of Security Law.

Security Interests

'Security interests' is the collective term used to describe a range of interests in property which secure the performance of an obligation. The term 'interest' refers generically to the various bundles of rights that make up particular types of interest.

Security interests may arise by operation of law. Common examples are liens where, in special circumstances, property can be detained for the satisfaction of an obligation. This article focuses, however, on consensual security interests, arising from negotiation between debtors and their creditors. Traditionally they fall into two fundamental categories identified by reference to whether the creditor must necessarily have possession of the property for the security to be enforceable. As the description implies, possessory interests require a creditor to have possession of the borrower's property. By contrast, nonpossessory interests may arise from, for example, a transfer of ownership to the creditor or the creation of some form of encumbrance over the borrower's property in favor of the creditor. In economic terms, the latter are the more important interests.

It is the creation of the actual interest in the property that distinguishes security interests from arrangements such as guarantees. Although guarantees are often commercially considered to constitute security, they are in legal terms merely contractual promises to make payment in specified circumstances. They do not of themselves confer an interest in property, although guarantors may choose additionally to give security in order to support their promise.

Typically, a security interest is created through the borrower granting the creditor a specific interest over property which the borrower owns. There are nonetheless transactions which resemble security but which legal regimes often do not recognize as part of their security law because they do not involve any grant by the borrower. The classic example is a reservation

of title (ownership) arrangement under which a vendor provides credit to a purchaser by permitting the purchaser to take possession of goods on terms that ownership to those goods will not pass until the purchaser makes full payment to the vendor. In such circumstances, there is no grant of an interest. Rather, the vendors reserve their own interest of ownership, thereby protecting themselves against nonpayment by the purchaser. Some legal jurisdictions which adopt the functional approach and focus on the economic substance of the transaction rather than its form have, however, explicitly brought reservation of title arrangements within their security law (see Functional Approach below).

Assets capable of being given as security (sometimes described as 'collateral') commonly include land as well as other property. Within this latter broad category is not only property which has a physical manifestation such as plant and equipment and motor vehicles but also property which is intangible, such as debts. Jurisdictions often develop different rules relating to security over land and security over other types of property, a consequence of the historical importance of land as a source of wealth and, more generally, its unique attributes. Jurisdictions may also provide special regimes for specific types of collateral (such as cash and financial instruments) important for the stability and efficiency of financial markets (Gullifer, 2008).

A modern security law regime is expected to provide mechanisms not only for the valid creation of a security interest in property but also for determining the priority of competing security interests and the availability of remedies to enforce the security interests in the event of default.

Graeco–Roman Origins

The development of security law has a long history and emerged as part of the creation of ownership-based societies in many parts of the ancient Mediterranean and ancient Near East (*see* Property Rights). References to security can be traced back to the earliest legislation of the Greek *polis*. When Solon, appointed as the ruler and lawgiver of Athens in 594/3 BC, canceled existing debts as part of his famous legal and economic reforms, this was signified by the removal of *horoi*. These were slabs of stone, marking the encumbrance of land as part of a loan (Finley, 1981). Approximately 250 *horoi* have been uncovered and documented. They date roughly to the period of 400–200 BC, which indicates their widespread and continuous use (Millett, 1985). One example is a marble block found within the city of Athens that reads: 'Horos of a workshop hypothecated, 750 (drachmas)' (Finley, 1951: 4).

Three distinct types of security appear on the *horoi: hypotheke, prasis epi lysei*, and *apotimema*. Of the three *prasis epi lysei*, ('sale on condition of release', Finley, 1951), was the most frequent, with 65% of *horoi* disclosing an encumbrance of this type (Millett, 1985). The property in question was 'sold' to the creditor, but on the condition that it would be 'released' or reconveyed on repayment of the loan. The debtor as a rule retained possession. It has been assumed that on default, the creditor acquired unrestricted ownership of the property and could take legal action if the debtor refused to transfer possession. *Prasis epi lysei* was regarded as the typical form of real security in ancient Athens (Harrison, 1968).

While the concept underlying *prasis epi lysei* is that of sale, the term *hypotheke* signifies the 'object put down' (Finley, 1951: 29). Only 5% of *horoi* refer to this type of security. Little evidence is available to distinguish *hypotheke* from *prasis epi lysei* transactions. It has been suggested that the *hypotheke* was a more flexible instrument. There is evidence that it was more often used in conjunction with written contracts (Finley, 1951). The absence of any reference to 'sale' suggests that ownership stayed with the debtor (Harrison, 1968). The concepts of ownership and possession were, however, less developed in Athenian law than in Roman law. Hence assumptions in relation to 'ownership' require qualification (Lipsius, 1915; Harrison, 1968). Under the *hypotheke*, the 'owner'–debtor also retained possession (Finley, 1951).

The remaining 30% of '*horoi*' identify the third type of security – *apotimema*. *Apotimema* was classified as either *pupillary apotimema* or *dotal apotimema*. The former signified a security transaction where an orphan's estate was leased out under the state's supervision and a *horos* was placed on the property of the lessee as security for the return of the lessor's property. The latter ensured the return of the dowry to the father of the bride in the case of divorce or other circumstances. It is clear from the nature of *apotimema* transactions that the party providing the security remained in possession of the property (Finley, 1951).

Two other important forms of Greek security transactions related to movables. The first was the *enechyron* (pledge), where movables were handed over to a creditor as security for loans. There are instances of Greek orators handing over gold cups, horses, slaves, and gold crowns to creditors, including bankers (Harrison, 1968). The second took the form of maritime or bottomry loans where the ship master or merchant obtained an advance of money on the security of the ship, the slaves who formed its crew, or its cargo (MacDowell, 1978).

The taking of security was also a 'universal feature of Roman life' (Crook, 1967: 243) and was critical in an economy based on debt (Harris, 2008b). It was in Rome that the law of security was developed into a system and codified. This codification, in turn, became the model for modern security law.

The Roman law *fiducia* (*fiducia cum creditore contracta*) provided that the ownership of property was formally transferred (through *mancipatio* or *in iure cessio*) for security purposes to the creditor, who agreed to reconvey the property on repayment of the loan. The transfer of possession from the debtor to the creditor was not required to make this transaction effective. Typically the debtor was left in possession. This type of security applied to *res mancipi*, i.e., property the ownership of which could formally be transferred to the creditor. It included the most important Roman asset classes such as land, slaves, cattle, horses, mules, donkeys, and rights of way and of water. However, the debtor retained a right against the creditor, the *actio fiduciae*. This precluded the creditor from selling or encumbering the assets, and obliged the creditor if in possession, to prevent any damage. If the loan was not paid back when due, the creditor could seek compensation from the property securing the loan. Contracts typically contained a clause allowing the creditor to sell the property (for example, the standard contract – *Formula Baetica* – found in a notary's office in Spain and tablets from Herculaneum, see Crook, 1967). If the sale proceeds were less than the loan obligation, a residual obligation on the part of the debtor remained. On

the other hand, if the sale proceeds exceeded the obligation arising from the loan, the excess (*superfluum*) was owed to the debtor (Kaser, 1971).

The Roman law *pignus* allowed the creation of a security interest without any formal transfer of ownership. This is believed to have originally required the delivery of the property to the creditor. However, from the third century BC a *pignus* was already capable of use as a nonpossessory security interest by agreement, without delivery of the property. *Pignus* without possession was also referred to by the Greek term *hypotheca*, although there is no direct evidence suggesting that it was introduced due to Greek influence (Kaser, 1971). If the property was delivered to the creditor, the use of the property by the creditor required permission from the debtor. Similarly, the creditor had no automatic right to take the profits and fruits of the property. This also required an agreement with the debtor, and the value of fruits and profits were typically offset against interest and loan principal ('Anticresis', Kaser, 1981).

Property that could be subject to a *pignus* included all physical things, obligations (*nomina*), and anything else saleable. A flock or a herd could be subject to a *pignus* even though the individual animals might change. Similarly, ungathered fruits or unborn animals, and children were also so subject (Roby, 1902). Moreover, a *pignus* could be created over the whole of a debtor's property, including all the debtor's current and future assets (Kaser, 1971). Similar to the *fiducia*, loan contracts regularly stipulated that the creditor would sell the asset used as security, should the loan agreement not be fulfilled when due. The debtor was entitled to any excess of the sale proceeds over and above the obligations arising from the loan. At its fullest stage of development, the *pignus* provided the creditor with the *Actio Serviana*, a special (praetorial) legal action that allowed a creditor to obtain possession of the property used as security from the debtor or any third party if a debt was not paid when due (Kaser, 1971; Jolowicz and Nicholas, 1972).

Roman creditors, who held security interests, also retained priority over unsecured creditors when execution was levied against an insolvent debtor's assets (Kaser, 1996).

English Common Law

Common Law developed in medieval England as the law common to the whole of the country and has become the generic term used to describe the English legal system as well as those legal jurisdictions which adopted a similar legal framework, whether historically through colonization or otherwise. The Common Law classification of security interests into mortgages, possessory securities, and hypothecations has been traced to Roman Law, although significant differences developed in their operation (Sykes and Walker, 1993).

Three types of consensual security interests, reflecting the above classification, are generally recognized in relation to personal property under Common Law: the pledge, the mortgage, and the charge (Beale et al., 2012). Such interests are commonly given by the person who owes payment or the performance of a particular obligation that is secured by the asset. Underlying personal liability on the part of the debtor is not, however, a necessary criterion. A third party may give any of the security interests, to support a promise made by a debtor to a creditor.

Common law classifies property and interests in property by reference to whether they are 'legal' or 'equitable'. This reflects the historical development and recognition of property and interests by courts exercising what is described as 'Common Law' or 'equitable' jurisdiction. The classification remains important in the context of security law as it can impact on the type of security that can be taken over particular property as well as on the operation of priority rules and the availability of remedies.

A pledge, a very early form of security (Goode, 2010), is created by a person (the pledgor) giving possession of an object to a creditor (the pledgee) by way of security, on the basis that the creditor will return the object on payment of the debt. In the event of a default, the pledgee may sell as agent of the pledgor. The pledgee is said to have a 'special property', meaning 'possession' as distinct from ownership. It is a bailment of tangible property (choses in possession), in which the pledgee obtains the legal interest of possession. At Common Law (unlike the position under civil law) it is not possible to grant a pledge over pure intangible property (chose in action) such as a debt, as that form of property does not have a physical manifestation and hence is incapable of possession. Where choses in action are embodied in documentation (a documentary intangible), as in the case of bills of exchange, a pledge may be taken.

More common, however, as security over personal property are the nonpossessory security interests of mortgage and charge. They can be taken over all forms of personal property and not just choses in possession.

The mortgage involves a transfer of title by the debtor who owns the property (mortgagor) to the creditor (mortgagee), on the basis that the title will be retransferred on repayment of the debt. The transfer of title may take place at law, giving rise to a legal mortgage. In theory, the transfer of legal title brings with it a transfer of possession where the property is goods, although it is common practice for the mortgagor to remain in possession. A transfer of title may alternatively take place in equity, giving rise to an equitable mortgage. Although in both cases, the interest of 'title' is transferred to the mortgagee, the mortgagor has long been recognized as having an equitable interest in that property (known as 'the equity of redemption'). A mortgagee's remedies include the remedies of sale, possession, appointment of a receiver, and foreclosure, the latter being a means whereby the mortgagee seeks court approval to extinguish the equity of redemption and take ownership of the property.

In practice in modern commercial life, it is common for companies to use the equitable charge as their preferred form of security over personal property. A charge is commonly described in terms of the borrower (chargor) appropriating property for the satisfaction of an obligation owed to the lender (chargee). It can operate as a 'fixed' security, meaning that it attaches to the property at the time of its creation as an enforceable contract. During the nineteenth century, however, courts exercising equitable jurisdiction additionally recognized the existence of a 'floating charge' in response to industrial and commercial needs (Pennington, 1960). The floating charge has been described as 'an interest not in specific assets but in

a constantly changing fund of assets' (Gullifer, 2008: 123). For as long as the charge continues to float, the chargor can deal with the assets in the ordinary course of business. Hence it is a particularly useful form of security in relation to inventory. A floating charge may crystallize, meaning that it becomes fixed and attached to specific property precluding the chargor from dealing with it. Crystallization happens on the occurrence of certain events impacting on the chargor; namely, liquidation, receivership, cessation of business, and defaults so specified in the charge agreement.

Charges can also be taken over land, as indeed can equitable mortgages. It is important, however, not to confuse a charge in the sense described above with the formal 'charge by way of legal mortgage', which has been used in England since the mid-1920s as a form of security over land. It is now the only statutory form that can be used over registered land (being land regulated by registration legislation, currently the Land Registration Act 2002). This charge by way of legal mortgage does not involve any transfer of legal title, thereby differentiating it from the traditional Common Law mortgage. The Common Law mortgage, discussed above in relation to personal property, had emerged in relation to land by the fifteenth century and survived until the mid-1920s (Harpum et al., 2012), itself differing from earlier forms of security over land which involved the mortgagee possessing the land under a lease and obtaining ownership on default under the terms of the lease (Holdsworth, 1923).

The creditor holding security under the charge by way of legal mortgage has statutory powers equivalent to that of a mortgagee at common law (Gray and Gray, 2009). Remedies capable of exercise therefore include sale, foreclosure, possession, and appointment of a receiver. Statutory charges over land, described as mortgages, can be found in other Common Law jurisdictions which since the nineteenth century adopted a conveyancing system, known colloquially as 'Torrens title'. It takes its name from Sir Robert Torrens who is credited with the introduction of initial legislation in South Australia in 1858, although it is said to be based on 'Hanseatic systems of mercantile registration' (Gray and Gray, 2009: 187). It has been adopted in other countries, including New Zealand, parts of Canada, parts of South East Asia, Africa, and the Carribean (Butt, 2010) and is described as 'the inspirational force' behind the English land registration schemes (Gray and Gray, 2009: 188).

French and German Civil Code Frameworks

The most influential civil law codes in the world are generally acknowledged to be the French (1804) and the German (1896) codes. They have each provided the basis for codification in many other countries.

The French Code Civil deals extensively with security in Title II *des sûretés réelles* (of real securities, Articles 2323–2373). Common forms of security over immovables governed by the Code Civil are the *hypothèque* (mortgage, Article 2393), the *privilège du prêteur de deniers* (lender's privilege, Article 2374, 2379) and the *gage immobilier* (Antichresis, Article 2387) (Cabrillac and Mouley, 1990). The *hypothèque* is a non-possessory security interest providing the creditor with a preferred right in the property regardless of who is owner at the time of execution. In order to be enforceable against third parties, it must be registered in the local land registry. Successive *hypothèques* rank according to their date of registration. In certain circumstances, a *hypothèque* may subsequently secure debts other than the one through which it was originally created (rechargeable *hypothèque*, Article 2422).

The *privilège du prêteur de deniers* is a nonpossessory security interest held by the lender who provided the funds to acquire the immovable. The lender obtains a preferential right over the potential proceeds from the property, subject to certain conditions and registration within 2 months. This security is less expensive and requires fewer formalities than the *hypothèque*. It is often the preferred form of security over immovables (Boughida et al., 2011).

In contrast to both *hypothèque* and *privilege*, the *gage immobilier* provides the creditor with possession of the property and the rights to the profits and revenue. However, this type of security is not used in practice (Boughida et al., 2011).

The Code Civil security over tangible movables (actual or future) such as goods, animals, furniture, and materials is the gage (pledge, *gage de meubles corporels*, Article 2333). A *gage* is a security interest in the movable that is effective against third parties if the movable is delivered to the creditor (*gage avec dépossession*) or through registration of the security in a special register held at the applicable court office (*gage sans dépossession*, Article 2338). In the case of default, the creditor may sell the property subject to the *gage* by court order or it may be agreed that the creditor becomes owner of the property. If the value of the property exceeds that of the debt, the difference is payable to the debtor or other secured creditors. The taking of security over inventory is regulated by the French Commercial Code (Articles L. 527-1 to L. 527-11, Code de commerce).

Security over intangible property such as shares, receivables, bank accounts, and debts can be created by the Code Civil *nantissement* (security, pledge, *nantissement de meubles incorporels*, Article 2355). Like the *gage*, the *nantissement* is the allocation of a set of actual or future assets as security for a loan. In the case of default, the assets subject to the *nantissement* are assigned to the creditor by court order or prior agreement. Sums in excess of the secured debt are paid to the debtor. Receivables are more commonly assigned by way of security according to the provisions of the French *Code monétaire et financier* (Articles L. 313-23). Here, ownership of the receivable is assigned to the credit institution upfront (Hubert, 2010).

The German Civil Code (BGB) provides specific regulations for the taking of security over land and buildings through the *Hypothek* (*Hypothec*, BGB §1113), the *Grundschuld* (land charge, BGB §1191) and the *Rentenschuld* (annuity land charge, BGB §1199); and over movables and rights through the *Pfand* (pledge, BGB §1204, 1273). While the provisions of the code satisfy commercial requirements in relation to the taking of security over land and buildings, the absence of nonpossessory security over movables and limitations with respect to security over claims to money and other rights, has led to the development of contractual forms of security outside the civil code (*Sicherungstreuhand* – security trust), where ownership is transferred to the creditor by agreement (Bülow, 2012).

In German property law, *Hypothek*, *Grundschuld*, *Rentenschuld*, and *Pfand* are classified as limited rights of ownership (Wolff and Raiser, 1957). The power that these rights provide is that of execution. The holder, on nonpayment of a claim, can seek execution against the property to which the security extends and be compensated from the proceeds.

Security interests over land are created by entry in the land registry (*Grundbuch*) and cannot exist without registration (Bülow, 2012). The *Hypothek* is accessory and requires an associated credit contract, while the *Grundschuld* and *Rentenschuld* are nonaccessory (abstract) and can exist without a claim or loan. This means that a land owner can raise funds or meet obligations by selling a *Grundschuld* over the desired amount without being a party to a loan contract (Wolff and Raiser, 1957). More common, however, is the creation of a *Grundschuld* alongside a credit contract (*Sicherungsgrundschuld* – security land charge). The differences between *Hypothek* and *Grundschuld* from an economic viewpoint are small, but the independence of the *Grundschuld* from the underlying claim can provide advantages to the creditor. On repayment of the loan, the *Grundschuld* remains in the hands of the creditor and the transfer to the debtor can be tied to other conditions (Bülow, 2012). The *Grundschuld* has become the most common security over real estate in Germany. The *Rentenschuld* is similar to the *Grundschuld* but provides for an annuity payment. This form of security has not been successful in practice (Bülow, 2012).

The priority of security rights over land is determined by the date and time of registration. The first highest ranking creditor is fully compensated from the execution proceeds (if these are sufficient), while the next secured creditor is compensated from what remains. The same process applies to subsequent creditors. Any residual sum falls to the debtor (Bülow, 2012).

The German civil law *Pfand* in relation to movables requires the delivery of the property to the creditor. An important application is refinancing (Lombard) loans between the German central bank and commercial banks as part of the European Central Bank system, which can be secured by bank bills and other debt securities. Another example is bank loans that are secured by shares on deposit with a bank (Bülow, 2012).

The taking of security over movables, where possession of property must stay with the debtor is implemented through contractual agreement (*Sicherungstreuhand*). It is effected by *Sicherungsuebertragung* (security transfer) which is the transfer of ownership and indirect possession (utilizing BGB §930 – constructive delivery) to the creditor for security purposes. The direct possession stays with the debtor. An important application relates to goods stored in a warehouse. In this case the security interest extends to future goods that replace those exiting the warehouse (Bülow, 2012).

Security over rights such as patents, trademarks, negotiable instruments, and claims to money are typically implemented by contractual agreement, the *Sicherungsabtretung* (security assignment), rather than the civil code *Pfand*. Here the right is assigned to the creditor for security purposes. In the case of claims to money, this avoids the requirements to make the security transaction known to the debtor (BGB §1280) and also simplifies the realization of the security (Bülow, 2012).

Efforts are underway to further harmonize security interests within the European Union. The directive of the European Parliament and the Council on credit arrangements related to residential property (also referred to as the *Euro-hypothec*) seeks to harmonize the underlying loan contracts. Principles for the taking of security over movable assets have been outlined in Book IX of the Draft Common Reference Framework for Principles, Definitions, and Model Rules of European Private Law (Von Bar et al., 2009) (see Functional Approach below).

The Functional Approach

A functional approach to security law looks beyond the actual form that a particular transaction takes, such as for example a pledge or a mortgage. Rather, it focuses on its function, considering whether the transaction gives rise to an interest in property which has the effect of securing payment or performance.

This approach to security has its origins in the US Uniform Commercial Code. Although law in the United States was initially based on English Common Law (McPherson, 2007), English and American security law diverged considerably during the nineteenth and twentieth centuries. Concern over the potential for fraud on creditors precluded judicial recognition of the floating charge (see English Common Law above).

The absence of such a flexible and economically important device resulted, at least in part, in increasing the technicality of various forms of security. A new approach was adopted from the 1950s under Article 9 of the Uniform Commercial Code (Gilmore, 1965), which has subsequently been revised on a number of occasions.

Article 9's influence has since spread to both Common Law and civil law jurisdictions, leading to significant changes in security law regimes. Prominent examples of Common Law jurisdictions are the Canadian Common Law provinces, New Zealand, and Australia, each of whom have enacted personal property securities legislation, from Ontario in 1967 to New Zealand in 1999 and Australia in 2009. The functional approach of Article 9 is reflected at an international level in the 'Convention on International Interests in Mobile Equipment' (known as the Capetown Convention), which offers a special security regime for aircraft, railway rolling stock, and space assets. It can also be found in the Legislative Guide on Secured Transactions, published by United Nations Commission on International Trade Law (UNCITRAL) in 2010 (McCormack, 2011). Building on earlier work carried out by bodies such as the European Development and Reconstruction Bank and the Asian Development Bank, the Legislative Guide provides a general model for countries wishing either to modernize their existing security law or to introduce 'efficient and effective… laws' (UNCITRAL, 2010: 1, 3–4). A similar trend is evident in Book IX of the Draft Common Reference Framework (see Section on French and German law above).

A functional approach typically extends the concept of a security interest beyond security interests granted by a borrower and encompasses other interests which are functionally equivalent, most notably the reservation of title arrangements, whether arising under conditional sale

agreements, hire purchase, or security (or finance) leases. It also may explicitly include some other types of arrangements which do not actually secure payment or performance of an obligation. Outright transfers of particular types of debts may, for example, be included. Also included in some jurisdictions are certain operating leases, generally those for more than a term of 1 year where the lessor is regularly engaged in the business of leasing goods. Under these arrangements, the lessee is commonly regarded as having sufficient rights in the property for a security interest to attach to the goods and not merely the possessory interest in those goods. The lessor thus finds itself in competition with other parties having security over the lessee's assets.

Security interests created under this type of legislation are interpreted as fixed in nature, leaving no room for the operation of a floating security (see English Common Law above). In the absence of agreement to the contrary between the secured party and the debtor, a security interest may continue in the secured property despite the debtor dealing with it by, for example, selling or leasing it to a third party. This has necessitated the express inclusion of rules whereby those third parties may be able to take the property free of security interests. The statutory regimes also provide a comprehensive framework for resolving competing priority claims, generally based initially on the secured party having perfected its security interest. For policy reasons, however, certain types of secured parties may be preferred, such as vendors and lenders providing credit for the purchase of an asset where those secured parties' interests are characterized as purchase money security interests and where they have complied with relevant procedural requirements for registration.

A striking feature of these statutory regimes is the concept of notice filing or registration of claims to a security interest, as distinct from the more traditional concept of registration of the actual security interest. Also of note is the provision of specific statutory remedies for enforcement.

The Economic Importance of Security

Security rights are critical in any monetary economy, because the creation of money relies on the ability to encumber property by way of security for loans (Decker, 2013; Heinsohn and Steiger, 2013). When central bank money is created, commercial banks enter into loan contracts with the central note-issuing bank, which must be secured by eligible central bank collateral (Alternative modes of money creation are repurchase agreements, which resemble secured loans, and outright purchases).

The critical role of security in this process is particularly visible during a liquidity crisis. When acting as a lender of last resort, a central bank must provide funds to anyone who can furnish sound loan security (Bagehot, 1873; Heinsohn and Steiger, 2013). In fact, all principal property classes of an economy must be able to be monetized in a crisis (Decker, 2010, 2012). This became apparent, for instance, during the 2008 global financial crisis, where central banks around the world had to widen the types of assets eligible for central bank loans. Another insightful example is the German Hyperinflation of 1923, where the German monetary system was completely disrupted as the central bank monetized an ever increasing amount of government debt. The situation was eventually resolved through the issue of a new currency backed by a priority land charge that was imposed by the government on all land used for agricultural, forestry, or horticultural purposes (at 4 or 5% of the assessed land value) together with a matching first ranking debenture on all industrial assets (Wolff and Raiser, 1957; Decker, 2012). With the property foundation thus reestablished, the currency was stabilized with surprising rapidity.

If central banks are reliant on security, so are commercial banks. The latter create funds in the form of bank deposits through loans to nonbanking institutions, the majority of which are secured. In their essence, banks are lenders on security and not providers of unsecured funds such as venture capital, as any entrepreneur can attest. Banks are therefore best understood as creators of funds rather than intermediaries (Hawtrey, 1923; Schumpeter, 1954). Banks 'transmute' (Crowther, 1940) property assets into money. Since the invention of property rights and money, loan security has played an integral part in this process.

When a note-issuing bank creates money for a debtor when granting a loan, the bank has to back the issued notes with its capital. At the same time the debtor must provide loan security as part of the loan. Both parties place a burden on their property in this process. While still in physical possession, both parties can no longer freely dispose of their assets. It has been argued that the economic implication is that both parties have lost 'ownership premium'. The latter is a yield that is attached to property that is unencumbered and free and explains why interest is charged as part of a loan (Heinsohn and Steiger, 2013).

It has been argued that ownership premium and loan security also play an important role during an economic crisis. As property owners try to protect themselves from asset losses, the ownership premium rises, reducing the amount of property that is made available for the purposes of money creation. As the prices of assets are deflated in this process, banks are forced to 'write down' their assets reducing their capital and their ability to make loans. At the same time near-insolvent borrowers can no longer provide security to renew their loans. Central bank action is hampered because it can neither directly recapitalize banks nor provide to businesses the missing loan security required for fresh loans (Heinsohn and Decker, 2010).

Loan security is also considered to be a critical factor in economic development. The limited access to formal security rights for large parts of the population in developing countries creates the issue of 'dead capital' (De Soto, 2000). While large amounts of assets exist, these have been created in the extra-legal sector and 'cannot readily be turned into capital' and used as security for loans (De Soto, 2000: 5). This is now regarded as one of the key reasons for lack of development and programs have been set up to facilitate the establishment of improved security law around the world (see, for example, United Nations, 2010; Alvarez de la Campa et al., 2012).

See also: Property Ownership; Property Rights, History of.

Bibliography

Alvarez de la Campa, A., Dowes, S.C., Tirelli, Hennig B., 2012. Making Securities Interests Public: Registration Mechanisms in 35 Jurisdictions. International Finance Corporation, Washington, DC.

Bagehot, W., 1873. Lombard Street. A Description of the Money Market. 1962. Richard D. Irwin Inc., Homewood, IL.

Beale, H., Bridge, M., Gullifer, L., Lomnicka, E., 2012. The Law of Security and Title-based Financing, second ed. Oxford University Press, Oxford.

Boughida, S., Levesque, G., Roux, J., 2011. Lending and Taking Security in France. http://uk.practicallaw.com/7-501-4362 (accessed 24.11.13.).

Bülow, P., 2012. Recht der Kreditsicherheiten. Sachen, Rechte und Personen, eighth ed. C. F. Müller, Heidelberg.

Butt, P., 2010. Land Law, sixth ed. Law Book Co., Australia.

Cabrillac, M., Mouley, C., 1990. Droit des suretes. litec, Paris.

Crook, J., 1967. Law and Life in Rome. Cornell University Press, New York.

Crowther, G., 1940. An Outline of Money. Thomas Nelson and Sons, London.

Decker, F., 2010. The Emergence of Money in Convict New South Wales. Metropolis, Marburg.

Decker, F., 2012. Settlement Assets: The Missing Economic Category in the Theory of Money. SSRN Working paper. http://papers.ssrn.com/abstract=2007637.

Decker, F. (Ed.), 2013. Ownership Economics. Routledge, London.

De Lacy, J., 2010. The evolution and regulation of security interests over personal property. In: De Lacy, J. (Ed.), The Reform of UK Personal Property Security Law. Routledge Cavendish, London.

Finley, M.I., 1951. Studies in Land and Credit in Ancient Athens 500-200 B. C. Rutgers University Press, New Brunswick, NJ.

Finley, M.I., 1981. In: Shaw, B.D., Slader, R.P. (Eds.), Economy and Society in Ancient Greece. Penguin Books, Harmondsworth.

Gilmore, G., 1965. Security Interests in Personal Property. Little Brown & Co, Boston. Reprinted from: Lawbook Exchange, New Jersy (2008).

Goode, R., 2010. Commercial Law (revised by McKendrick, E.), fourth ed. Penguin Books, London.

Gray, K., Gray, S.F., 2009. Elements of Land Law, fifth ed. Oxford University Press, Oxford.

Gullifer, L. (Ed.), 2008. Goode on Legal Problems of Credit and Security. Sweet & Maxwell, London.

Harpum, C., Bridge, S., Dixon, M., 2012. Megarry & Wade the Law of Real Property, eighth ed. Sweet & Maxwell, London.

Harris, W.V., 2008b. The nature of roman money. In: Harris, W.V. (Ed.), The Monetary Systems of the Greeks and Romans. The Oxford University Press, Oxford.

Harrison, A.R.W., 1968. The Law of Athens. At the Clarendon Press, Oxford.

Hawtrey, R.G., 1923. Currency and Credit. Longmans, Green and Co., London.

Heinsohn, G., Decker, F., 2010. A property economics explanation of the global financial crisis. In: Kolb, R.W. (Ed.), Lessons from the Financial Crisis: Causes, Consequences, and Our Economic Future. John Wiley & Sons, Hoboken, NJ.

Heinsohn, G., Steiger, O., 2013. Ownership economics. On the foundations of interest, money, markets, business cycles and economic development. edited and translated with comments and additions by F. Decker. In: Decker, F. (Ed.), Ownership Economics. Routledge, London.

Holdsworth, W.S., 1923. History of English Law, third ed., vol. III. Methuen, London.

Hubert, O., 2010. Enforcement of Security Interests in Banking Transactions. Multi-Jurisdictional Survey, France.

Jolowicz, H.F., Nicholas, B., 1972. Historical Introduction to the Study of Roman Law, third ed. At the University Press, Cambridge.

Kaser, M., 1971. Das Römische Privatrecht. Erster Abschnitt. Das altrömische, das vorklassische und das klassische Recht, second ed. C.H.Beck, München.

Kaser, M., 1981. Römisches Privatrecht. Juristische Kurz-lehrbücher. C.H.Beck, München.

Kaser, M., 1996. Das Römische Zivilprozessrecht, second ed. Revised and edited by K. Hackl. C.H.Beck, München.

Lawson, F.H., Rudden, B., 2002. The Law of Property, third ed. Oxford University Press, Oxford.

Lipsius, J.H., 1915. Das Attische Recht und Rechtsverfahren. O. R. Reisland, Leipzig.

MacDowell, D.M., 1978. The Law in Classical Athens. Paperback edition 1986. Cornell University Press, Ithaca, NY.

McCormack, G., 2011. Secured Credit and the Harmonisation of Law. Edward Elgar, Cheltenham.

McCracken, S., Bird, J., Stumbles, J., Tolhurst, G., 2013. Everett & McCracken's Banking and Financial Institutions Law, eighth ed. LawBook Co. (Thomson Reuters), Australia.

McPherson, B.H., 2007. The Reception of English Law Abroad. Supreme Court of Queensland Library, Queensland.

Millett, P., 1985. The Attic Horoi reconsidered in the light of recent discoveries. In: Finley, I.M. (Ed.), Studies in Land and Credit in Ancient Athens 500–200 B. C. Transactions Inc., New Brunswick, NJ.

Pennington, R., 1960. The genesis of the floating charge. Modern Law Review 23, 630.

Roby, H.J., 1902. Roman Private Law, vol. II. At the University Press, Cambridge.

Schumpeter, J.A., 1954. History of Economic Analysis, 1986 edition. Allan and Unwin, London.

De Soto, H., 2000. The Mystery of Capital. Why Capitalism Triumphs in the West and Fails Everywhere Else. Bantam Press, London.

Sykes, E.I., Walker, S., 1993. The Law of Securities, fifth ed. Law Book Co, Australia.

United Nations, 2010. UNCITRAL Legislative Guide on Secured Transactions. United Nations, New York.

Von Bar, C., Clive, E., Schulte-Nölke, H. (Eds.), 2009. Definitions and Model Rules of European Private Law. Draft Common Reference Framework for Principles. sellier, Munich.

Wolff, M., Raiser, L., 1957. Sachenrecht. In: Enneccerus, L., Kipp, T., Wolff, M. (Eds.), Lehrbuch des Bürgerlichen Rechts, Dritter Band, tenth ed. J. C. B. Mohr (Paul Siebeck), Tübingen.

Wood, P., 2007. Comparative Law of Security Interests and Title Finance, second ed. Sweet & Maxwell, London.

Segregation, Demographic Effects of

Michael J White, Population Studies and Training Center, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An abiding concern in the study of segregation is that the residential clustering would itself influence social outcomes, either by changing behavior or by changing the structure of opportunities. The premise is that residential clustering carries behavioral consequences beyond those of ethnicity, socioeconomic status, or life cycle stage. A number of models advance a causal connection between this geographic clustering of traits and social behaviors that result. Contemporary research approaches to the issue generally involve multivariate, multilevel models. Increasingly, the availability of such data allows for a more sophisticated examination of this issue than heretofore. Multivariate studies have examined the effects of segregation and neighborhood composition on a wide array of outcomes, from health to schooling to child development. Investigators continue to deal with methodological challenges associated with such analyses, but accumulated research reveals the value of adding residential context to models of social and behavioral outcomes.

Segregation is the differential distribution of social groups. Most commonly, the use of the term refers to residential segregation, the differential distribution of social groups across geography. Residential segregation has served as a key indicator of social differentiation for decades. (*See* Residential Segregation in the United States in this volume for a general review.) Segregation is usually measured by one or more summary indices, designed to reflect the differential distribution of persons or other social units across neighborhoods, occupations, or other categories (White, 1986). Nowadays, segregation is often seen as a multidimensional phenomenon tapped by a variety of indices (Massey et al., 1996; Wilkes and Iceland, 2004). While the bulk of work has centered on residential segregation among ethnic groups, virtually any trait could be considered, and research has examined segregation by income, education, occupation, nativity, age, and other demographic traits. A line of research also strives to examine the relative contribution of group income differentials in determining (confounding) observed racial and ethnic segregation (Iceland and Wilkes, 2006; Fischer, 2003). Racial segregation across schools and occupations remains an important indicator of social distance more generally.

Beyond the measurement of segregation itself, there has always been interest in the consequences of residential clustering. Theory can be invoked to suggest positive as well as negative consequences (Cutler and Glaeser, 1997; Bell et al., 2006). This ranges from the ways in which residential concentration might generate employment networks or social support institutions, to the consequences of residential isolation for educational attainment or health outcomes. From the point of view of policy, the determinants and consequences of segregation are of particular interest. There has been heightened attention to the aftermath of living in or being raised in a segregated or disadvantaged neighborhood.

Considerable empirical work has now accumulated to investigate the effects of residential segregation and related phenomena. The underlying concept is that residential clustering carries behavioral consequences beyond those of ethnicity, socioeconomic status, or lifestyle. Such notions have been present in demography and allied social sciences for some time. Indeed, the residential neighborhood may be seen as one of the several very important contextual settings helping to shape outcomes for the individual person (Brooks-Gunn et al., 1993; Massey, 2014). If anything, the concerns about the effects of local environment – from household, through immediate neighborhood, to residential community and school district – have grown with time and additional research on social inequality. At the same time, however, a number of observers have raised doubts about the depth of our understanding of spatial (segregation) mechanisms and/or the ability to alter those environments, especially for disadvantaged individuals (Burdick-Will et al., 2011; DeLuca and Dayton, 2009; Sampson, 2008).

It is important to distinguish the influence of composition from that of behavioral causation. For instance, the receipt of public assistance varies appreciably across urban neighborhoods. The degree to which this behavior can be linked causally to the concentration of social groups in neighborhoods (and the sociological patterns that arise through clustering) needs to be weighed against an explanation based just on composition. The key question turns on whether geographic clustering itself influences behavior, beyond that attributable to other background traits, such as age and education.

Since this article is dedicated to the consequences of segregation, the causal direction of interest runs *from* the neighborhood or community aggregate; one should also note that mechanisms can run from individual or household behavior *to* the neighborhood, thereby generating segregation through process of selection (or restriction), often via local geographic mobility.

Many demographic traits are segregated in urban neighborhoods, and there is a consistency to the ranking. Although empirical results on these matters derive preponderantly from North America, results are accumulating for other high-income societies (Johnston et al., 2007; Musterd, 2005). Racial and ethnic groups remain very highly segregated in most urban areas. Several decades of study, from straightforward segregation measures through factorial ecology and the like, have verified the sorting of urban space along the lines of socioeconomic status. Life cycle stage is also differentiated.

Segregation indices for sex and age are generally modest, although segregation for household size and type is more discernible. Of course, certain special concentrations of population, such as group quarters for military populations, students, and those under correctional supervision, are highly segregated and are often excluded from consideration by analysts.

Contemporary statistical estimates of neighborhood effects rely on hierarchical and multilevel models. While strategies differ, the general approach is to develop a multivariate model, in which the behavioral and demographic outcome (employment, childbearing) is taken to be first a function of fixed personal traits and a set of family and household background factors. Neighborhood and community characteristics are introduced to the model simultaneously or sequentially. With the availability of multilevel statistical techniques and associated software, it is possible to derive more sophisticated (and econometrically appropriate) estimates of neighborhood or community effects. Some of the same conceptual and methodological apparatus can be applied to patterns or demographic unevenness found in other contextual social settings, such as the workplace or schools.

Mechanisms through Which Segregation May Influence Demographic Outcomes

Beyond the simple mechanism of demographic composition itself, segregation and differential concentration might influence outcomes in several ways. The outcomes can be placed within three groupings: (1) contagion or socialization effects; (2) resource allocation; and (3) institutional structures. (Various authors differ in identifying and labeling mechanisms. Leventhal and Brooks-Gunn (2000) make reference to three types of mechanisms: institutional resources, relationships, and norms/collective efficacy.) Contagion and socialization models, also closely linked to epidemic effects and diffusion models, have a long history in social science thinking. Simply stated, individuals alter their behavior on the basis of their interpretation of those around them (Jencks and Mayer, 1990; Sampson et al., 2002). The influence of peers on child behavior is one such critical mechanism. The general reasoning (see, e.g., Wilson, 1987) is that low-income or disadvantaged neighborhoods offer fewer mainstream role models (dual parent families, stable adult workers, school completion); by contrast one is exposed to different conditions and social behaviors, so persons growing up in such areas do not acquire mainstream norms. Massey and Denton argued that the intersection of poverty and segregation creates an environment where such conditions abound (Massey and Denton, 1993). Much discussion surrounding the impact of growing up in a disadvantaged inner city neighborhood or attending an inferior school draws on this line of thinking, and the mechanism attracts increasing concern in the field of health (Diez-Roux, 2001; Diez-Roux and Mair, 2010; Kawachi and Subramanian, 2007; Robert, 1999).

A second line of thinking involves resource allocation. Residents of affluent neighborhoods (or workers in advantaged workplaces) can provide superior public goods, either through direct community provision or through the political process. More than that, these public goods are likely to confer advantages that go beyond the immediate advantages of household and family. Conversely competition over scarce resources may influence behaviors and outcomes in disadvantaged neighborhoods. And related thinking along these lines concentrates on the impact of relative deprivation, rather than absolute conditions. Other externalities of clustering include relative advantage or deprivation in local sanitation (exposure to pathogens and toxins), public safety, and the like. Housing experiments, such as those relocating low-income individuals to middle-income communities, are often premised on such mechanisms. As is clear from discussion over how to interpret such interventions, disentangling causal mechanisms is not an easy task (DeLuca and Dayton, 2009; Sampson, 2008).

The third line of thinking involves the growth and maintenance of local institutions, whether formal or informal (Duncan, 1994). The literature identifies both positive and negative consequences of segregation. In the realm of economic outcomes, the notion of the ethnic enclave suggests that concentration, building on ethnic networks, may provide benefits. Here there would seem to be a crossover to the discussion of social capital and influences it conveys. Occupational and residential segregation among co-ethnics may generate employment opportunities unavailable in a more integrated economy. A long-standing argument is that racial residential segregation enables integration by social class. In an urban neighborhood, local institutions, such as a church or neighborhood association, might foster social ties across class boundaries, although even historical accounts of ghetto life do not always portray a comfortable coexistence of disparate groups. These evolving structures can serve as social actors and then influence behavioral outcomes for other members of the community. Challenges for empirical work along these lines involve adequately identifying institutional structures and then teasing out the influences of institutions, net of individual, family, and other community traits.

Methodological Issues

The study of the effects of segregation raises several methodological issues. A fundamental issue is one of definition. A 'neighborhood' or local 'community' might have several competing geographic boundaries. Tracts, wards, block groups, and other urban statistical subdivisions become *de facto* neighborhoods (White, 1987). This has two consequences. First, such units serve as the basic geography for calculations of segregation measures. Second, these small statistical areas are the units that serve as the proxy for social conditions in multilevel models.

In recent years there has been a tremendous evolution in spatial thinking and modeling in the social sciences, along with accompanying methodological development to implement analysis suitable to that thinking (Logan, 2012). For decades, analysts concerned with neighborhood or other similar contexts were often constrained to ecological approaches, i.e., invoking models which relate aggregate values of demographic characteristics to one another. While some might argue for the focus on the neighborhood only, multilevel analysis can analyze outcomes as the result of both individual and

contextual (neighborhood, school) characteristics with more attention to proper statistical specification. The composition of the neighborhood or city, even the level of segregation characterizing the metropolitan region, could be one of these aggregate characteristics. Data are increasingly available in ways that link individual to context, but are promulgated or stored in a way so as to maintain confidentiality of the information for the respondents. Despite the considerable advances in data and statistical technology, challenges remain in model specification and even in correctly constructing contextual variables with respect to time and space.

Evolution of Empirical Investigation

From the earliest investigations of residential segregation, there has been interest – implicit or explicit – in its effects. A large social science literature grew rapidly around the measurement and analysis of racial and ethnic residential segregation. As interests expanded and as more small-area data became available, analyses of residential segregation spread to other demographic and social characteristics and to geographic settings beyond North America. Soon, interest arose in empirical testing of mechanisms linked to residential segregation, both as cause and consequence. The advent of superior data structures and estimation techniques quickly facilitated doing so. At the same time a companion literature has arisen discussing issues of measurement, inference, and other challenges to understanding contextual or neighborhood effects (DeLuca and Dayton, 2009; Logan, 2012; Sampson et al., 2002).

Additional work on the measurement of spatial segregation, per se, has also been underway. Considerable research accumulated on the dimensions of residential segregation, and since then investigators have turned to further consideration of the meaning of alternative measures and, of particular note, to more explicit concern for geographic scale of measurement (Reardon et al., 2008), to space itself, and to multiethnic patterns.

The empirically based literature on the effects of residential segregation is now voluminous. Much literature has grown up around a focus on the effects of concentrated disadvantage, such as the consequences of growing up in a poor neighborhood or attending a 'bad' school. After expressing reservations about the quality and consistency of data, a 1990 review of empirical studies of the effects of socioeconomic status and racial mix, for both neighborhoods and schools, concluded that there was no general pattern of effects across a wide range of outcomes (Jencks and Mayer, 1990). A more recent and comprehensive review focused on children and adolescents (Leventhal and Brooks-Gunn, 2000) identifies neighborhood effects for several education-related outcomes, for behavioral and emotional well-being, and for sexual behavior. Through this literature, the range of potential mechanisms and impacts across demographic groups and social characteristics is quite large. Qualitative approaches in neighborhood research, for instance, with regard to parenting strategies, might help uncover mechanisms (Jarrett, 1997; Leventhal and Brooks-Gunn, 2000). Effects of residential segregation, i.e., neighborhood composition, or equivalent have been found in many studies, but at the same time, the scale of these effects must be interpreted against the magnitude of other influences, particularly those of the immediate family. Given the argument for making investments at an early age in disadvantaged children (Heckman, 2006), there is a companion case for both understanding pathways through which segregation might influence outcomes and conducting the empirical work to measure their impact and for whom they operate.

One long-standing line of work looks at the influence of segregation itself. Empirical analyses often introduce a residential segregation index to a regression model of a behavioral outcome. The behavioral mechanism underlying the inclusion of such a covariate is not always clear, but higher levels of residential segregation may tap into underlying housing market constraints, a more hostile climate of local race relations, or differential access to public sector resources. Predictive power of such segregation indices is not always strong, but on balance they tend to show that individuals (usually minority group members) living in more segregated urban areas experience detrimental outcomes, such as inferior health outcomes (LaVeist, 1989; Subramanian et al., 2005), differential schooling outcomes (Card and Rothstein, 2007), or less mobility out of poverty (South and Crowder, 1997).

Recent developments in methods and data have pushed forward work on residential segregation, and companion analyses of neighborhood effects. We have clearer evidence on the ways in which residential segregation impacts a range of societal outcomes. While statistical findings are not always simple, they do reinforce the view, held by social scientists for decades, that neighborhood and context matter.

See also: Contextual Studies: Methodology; Demographic Measurement: Nuptiality, Mortality, Migration, and Growth; Human Ecology; Multilevel Models in Demography; Neighborhood Effects; Residential Segregation in the United States; Residential Segregation: Geographic Aspects; Segregation Indices; Spatial Analysis in Geography; Urban Internal Spatial Structure; Urban Sociology.

Bibliography

Bell, Janice F., et al., 2006. Birth outcomes among urban African-American women: a multilevel analysis of the role of racial residential segregation. Social Science & Medicine 63, 3030–3045.

Brooks-Gunn, J., Duncan, G.J., Klebanov, P.K., Sealand, N., 1993. Do neighborhoods influence child and adolescent development? American Journal of Sociology 99, 353–395.

Burdick-Will, J., Ludwig, J., Raudenbush, S.W., Sampson, R.J., Sanbonmatsu, L., Sharkey, P., 2011. Converging evidence for neighborhood effects on children's test scores: an experimental, quasi-experimental, and observational comparison. In: Duncan, G.J., Murnane, R.J. (Eds.), Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances. Sage Foundation, Russell, NY, pp. 255–276.

Card, D., Rothstein, J., 2007. Racial segregation and the black–white test score gap. Journal of Public Economics 91, 2158–2184.

Cutler, D., Glaeser, E.L., 1997. Are ghettoes good or bad? Quarterly Journal of Economics 112, 827–872.

DeLuca, S., Dayton, E., 2009. Switching social contexts: the effects of housing mobility and school choice programs on youth outcomes. Annual Review of Sociology 35, 457–491.

Diez Roux, A.V., 2001. Investigating neighborhood and area effects on health. American Journal of Public Health 91, 1783–1789.

Diez Roux, A.V., Mair, C., 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186, 125–145.

Duncan, G., 1994. Families and neighbors as sources of disadvantage in the schooling decisions of white and black adolescents. American Journal of Education 103, 20–53.
Fischer, M.J., 2003. The relative importance of income and race in determining residential outcomes in US urban areas, 1970–2000. Urban Affairs Review 38 (5), 669–696.
Heckman, James J., 2006. Skill formation and the economics of investing in disadvantaged children. Science 312, 1900–1902.
Iceland, J., Wilkes, R., 2006. Does socioeconomic status matter? Race, class, and residential segregation. Social Problems 53, 248–273.
Jarrett, R., 1997. African American family and parenting strategies in impoverished neighborhoods. Qualitative Sociology 20, 275–288.
Jencks, C., Mayer, S.E., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn, L.E., McGeary, M.G.H. (Eds.), Inner City Poverty in the United States. National Academy Press, Washington.
Johnston, R., Poulsen, M., Forrest, J., 2007. The geography of ethnic residential segregation: a comparative study of five countries. Annals of the Association of American Geographers 97 (4), 713–738.
Kawachi, I., Subramanian, S.V., 2007. Neighbourhood influences on health. Journal of Epidemiology and Community Health 61, 3–4.
LaVeist, T.A., 1989. Linking residential segregation to the infant-mortality race disparity in U.S. cities. Sociology and Social Research 73, 90–94.
Leventhal, T., Brooks-Gunn, J., 2000. The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. Psychological Bulletin 126 (2), 309.
Logan, J.R., 2012. Making a place for space: spatial thinking in social science. Annual Review of Sociology 38, 507–524.
Massey, D.S., Denton, H., 1993. American Apartheid. Harvard University Press, Cambridge, MA.
Massey, D.S., White, M.J., Phua, V.C., 1996. The dimensions of residential segregation revisited. Sociological Methods and Research 25, 172–206.
Massey, D.S., 2014. Spheres of Influence. Russell Sage, New York.
Musterd, S., 2005. Social and ethnic segregation in Europe: levels, causes, and effects. Journal of Urban Affairs 27, 331–348.
Reardon, S.F., Matthews, S.A., O'Sullivan, D., Lee, B.A., Firebaugh, G., Farrell, C.R., Bischoff, K., 2008. The geographic scale of metropolitan racial segregation. Demography 45, 489–514.
Robert, S.A., 1999. Socioeconomic position and health: the independent contribution of community socioeconomic context. Annual Review of Sociology 25, 489–516.
Sampson, R.J., 2008. Moving to inequality: neighborhood effects and experiments meet social structure. American Journal of Sociology 114, 189–231.
Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T., 2002. Assessing 'neighborhood effects': social processes and new directions in research. Annual Review of sociology 28, 443–478.
South, S., Crowder, K., 1997. Escaping distressed neighborhoods: individual, community, and metropolitan influences. American Journal of Sociology 102, 1040–1084.
Subramanian, S.V., Dolores, Acevedo-Garcia, Theresa, L. Osypuk, 2005. Racial residential segregation and geographic heterogeneity in black/white disparity in poor self-rated health in the US: a multilevel statistical analysis. Social Science & Medicine 60, 1667–1679.
White, M.J., 1987. American Neighborhoods and Residential Differentiation. Russell Sage, NY.
White, M.J., 1986. Segregation and diversity measures in population distribution. Population Index 52, 198–221.
Wilkes, R., Iceland, J., 2004. Hypersegregation in the twenty-first century. Demography 41, 23–36.
Wilson, W.J., 1987. The Truly Disadvantaged. University of Chicago Press, Chicago.

Segregation Indices

Douglas S Massey, Office of Population Research, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on residential segregation has been a topic of social science investigation since the first studies were done by sociologists in the Chicago School during the 1920s. For many years, scholars engaged in a spirited debate about how best to measure geographic segregation but a period of methodological peace was created when Duncan and Duncan in 1955 concluded that the index of dissimilarity was the best measure. The methodological debate was renewed in the 1970s but settled once again during the 1980s after Massey and Denton linked the various indices to one of five separate dimensions of segregation. Indices for each of these dimensions are defined and discussed here and current thinking on segregation and its measurement is summarized.

Residential segregation has long been a topic of interest to social scientists, dating back to the early days of the Chicago School of Sociology. In their analyses of urban spatial patterns, for example, Burgess (1925) noted the tendency of racial and ethnic groups to separate themselves into distinct residential enclaves and Park (1925) developed a theory of ecological succession to account for the formation of these enclaves. The first systematic study of segregation was published by Burgess in 1928 and since then residential segregation has maintained a constant presence in the social science literature (see Massey (1985, 2014) and Charles (2003) for reviews). For many years, social scientists searched for the ideal index of segregation and over the years many options were proposed and debated with no clear resolution.

The First Methodological Peace

In 1955, however, Duncan and Duncan undertook a systematic review of the candidate indices and concluded that, overall, the index of dissimilarity was the best measure. This index measures segregation as the degree of departure from an even residential distribution of two social groups in urban space. An even distribution occurs when each neighborhood replicates the percentage of group members in the urban area as a whole. If blacks constitute 10% of an urban area and whites 90%, for example, then in an even distribution each neighborhood would be 10% black and 90% white and the dissimilarity index would be zero. The index may be interpreted as the relative share of group members (e.g., blacks and whites) who would have to exchange neighborhoods to achieve an even residential distribution and equals 100 when two groups share no neighborhood in common.

The Duncans' analysis ushered in two decades of methodological peace in which the index of dissimilarity was routinely applied to measure ethnic and racial segregation in diverse urban areas throughout the world but especially the United States, yielding some of the field's classic studies of racial and ethnic segregation, such as Duncan and Duncan's *The Negro Population of Chicago* (1957), Liberson's *Ethnic Patterns in American Cities* (1963), Taeuber and Taeuber's *Negroes in Cities* (1965), and Grebler et al.'s *The Mexican American People* (1970).

Index Wars and the Second Methodological Peace

The period of consensus ended in 1976, however, when Cortese, Falk, and Cohen published an article in the *American Sociological Review*, the field's over-reliance on the dissimilarity index, arguing among other things that the standard of evenness was unrealistic and suggesting that a random distribution in urban space would be a better standard of comparison.

Although practitioners such as Taueber and Taueber (1976) were unimpressed with the critique, it did set off a decade of renewed debate on the best way to measure residential segregation and in the wake of their article a variety of new measures were proposed and older indices were revived (James and Taueber, 1985). One critique of the dissimilarity index was that it was aspatial: it simply measured departures from an abstract ideal of evenness but took no account of where neighborhoods were situated geographically in the urban landscape. Several new spatially sensitive measures were introduced and some scholars proposed adjusting existing indices to take account of their spatial location (Jakubs, 1981; Morgan, 1982, 1983).

After a flurry of methodological activity, the second period of debate came to an end when Massey and Denton (1988) undertook a factor analysis of all prevailing segregation measures and identified five distinct underlying dimensions of segregation and recommended an index for each one. Massey and Denton's analysis was based on 1980 census data used to compute indices of segregation for Hispanics, African Americans, and Asians in 60 metropolitan areas, but it was later replicated by Massey et al. (1996), who reached the same conclusions using 1990 census data for these groups in all metropolitan areas. The first dimension identified by Massey and Denton (1988) was unevenness, the degree to which two groups were unevenly distributed across spatial units, and they confirmed Duncan and Duncan's selection of the index of dissimilarity as the best overall indicator for this dimension of segregation. The other four dimensions were isolation, clustering, concentration, and centralization.

Isolation is the degree to which a social group lives in neighborhoods dominated by members of the same group and to measure it Massey and Denton recommended

Lieberson's (1981) P* isolation index, which indicates the percentage of own-group members in the neighborhood of the average member of that group. The black isolation index, for example, is the percentage of blacks in the neighborhood of the average African American and it ranges from the percentage of blacks in the urban area as a whole to 100. Clustering is the degree to which neighborhoods inhabited by a group are packed together in space and is best measured by White's (1983) spatial proximity index (SP) which captures the differential clustering of two groups in residential space. For blacks and whites, the index would equal zero if black and white neighborhoods were spread around an urban area like squares on a checkerboard and 100 when all blacks are clustered together in a set of contiguous black neighborhoods and all whites are clustered together in a set of contiguous white neighborhoods.

Concentration refers to the relative amount of physical space occupied by a social group within an urban area. To measure this dimension, Massey and Denton built on earlier work by Duncan et al. (1961) to define a relative concentration index (RCO). When computed for blacks and whites, the index takes a value of zero when two groups occupy the same amount of space and 100 when blacks occupy the smallest amount of physical space possible compared with whites. Centralization is the degree to which members of a social group reside near the center of any metropolitan area and Massey and Denton (1988) again drew upon the work of Duncan et al. (1961) to define an absolute centralization index (ACE) that for blacks and whites gives the proportion of blacks who would have to change neighborhoods to achieve an even residential distribution around the geographic center of the metropolitan area.

In a follow-up article, Massey and Denton (1989) drew upon 1980 census data to measure black segregation in 60 large metropolitan areas on all five dimensions and found that as of 1980 African Americans in 16 areas were highly segregated (index value >=60 on a 0–100 scale) on at least four of the five dimensions, a condition they labeled hypersegregation. In the years since the Massey–Denton work was published, social scientists have regularly computed indices for the five dimensions of segregation they identified. Analyzing data for all metropolitan areas in 1990, Denton (1994) expanded the list of black hypersegregated areas to 29, the same number found by Wilkes and Iceland (2004) in their analysis of data from the 2000 census. Working at the US Census Bureau, Iceland et al. (2002) ultimately published a battery of indices measuring segregation on all five dimensions for blacks, Hispanics, Asians, and American Indians using data from the 1980, 1990, and 2000 censuses. Descriptions, formulae, and instructions for computing the various indices are available online at https://www.census.gov/hhes/www/housing/housing_patterns/pdf/app_b.pdf.

Current Thinking on the Measurement of Segregation

The multidimensional nature has come to be widely accepted although the dissimilarity index of unevenness and P* index if usikatuib are generally computed more often than indices of the other three dimensions (see Rugh and Massey, 2013). Although the index of dissimilarity continues to be the workhorse of segregation research, another measure of evenness has come into use to make up for the limitations of the dissimilarity index. Two drawbacks of the index of dissimilarity are that it can only be computed for two groups at a time and it is not easily partitioned into components corresponding to different sources of spatial variation.

In order to measure the degree of segregation between multiple groups at the same time (say, blacks, whites, Asians, and Latinos or five income quintiles) or to decompose overall segregation into interpretable components, investigators have turned to Theil's (1972) entropy index. Reardon and Bischoff (2011) used it to assess the degree of segregation between income classes in US metropolitan areas, whereas Fischer (2003) used it to analyze the changing importance of race/ethnicity versus class in determining a group's level of segregation, and Fischer et al. (2004) employed it to decompose segregation by race, ethnicity, class, and life cycle into separate contributions by region, metropolitan area, city-suburban residence, and census tract.

When measuring segregation between two groups, the entropy index and the dissimilarity index are highly correlated, so Theil's index has generally been applied whenever investigators seek to measure segregation between multiple groups, or to decompose segregation into analytic components. In the end, however, there is no single 'correct' index of residential segregation and investigators continue to invent and try out new options. The choice of an index depends on what dimension of segregation and the kind of analysis one seeks to conduct. The fullest picture of a group's residential segregation is probably achieved by taking a multidimensional approach, though in any given analysis investigators may wish to focus on one particular dimension of interest.

See also: Community Sociology; Discrimination: Racial; Ethnicity, Sociology of; Human Ecology; Minorities; Prejudice in Society: Psychological Perspectives; Racial Relations; Racism, Sociology of; Urban Sociology.

Bibliography

Burgess, E.W., 1925. The growth of the city: an introduction to a research project. In: Park, Robert E., Burgess, Ernest W. (Eds.), The City. University of Chicago Press, Chicago, pp. 47–62.

Burgess, E.W., 1928. Residential segregation in American cities. Annals of the American Academy of Political and Social Science 140, 105–115.

Charles, C.Z., 2003. The dynamics of racial residential segregation. Annual Review of Sociology 29, 67–207.

Cortese, C.F., Falk, R.F., Cohen, J.S., 1976. Further considerations on the methodological analysis of segregation indices. American Sociological Review 41 (4), 590–592.

Denton, Nancy A., 1994. Are African Americans still hypersegregated? In: Bullard, R.D., Grigsby III, J.E., Lee, C. (Eds.), Residential Apartheid: The American Legacy. CAAS Publications, University of California, Los Angeles, CA, pp. 49–81.

Duncan, Otis D., Duncan, Beverly, 1955. A methodological analysis of segregation indices. American Sociological Review 20 (2), 210–227.

Duncan, O.D., Duncan, B., 1957. The Negro Population of Chicago: A Study of Residential Succession. University of Chicago Press, Chicago.

Duncan, O.D., Cuzzort, R.P., Duncan, B., 1961. Statistical Geography: Problems in Analyzing Areal Data. Free Press, New York.

Fischer, C.S., Stockmayer, G., Stiles, J., Hout, M., 2004. Distinguishing the geographic levels and social dimensions of U.S. metropolitan segregation, 1960–2000. Demography 41 (1), 37–59.

Fischer, M.J., 2003. The relative importance of class and race in determining residential outcomes in U.S. urban areas, 1970–2000. Urban Affairs Review 38 (5), 669–696.

Grebler, L., Moore, J.W., Guzman, R.C., 1970. The Mexican-American People: The Nation's Second Largest Minority. Free Press, New York.
Iceland, J., Weinberg, D.H., Steinmetz, E., 2002. Racial and Ethnic Residential Segregation in the United States 1980–2000. U.S. Bureau of the Census, Washington, DC.
Jakubs, J.F., 1981. A distance-based segregation index. Journal of Socioeconomic Planning Sciences 15 (3), 129–136.
James, David R., Taeuber, Karl E., 1985. Measures of segregation. In: Tuma, Nancy (Ed.), Sociological Methodology 1985. Jossey Bass, San Francisco, pp. 1–32.
Lieberson, S., 1963. Ethnic Patterns in American Cities. Free Press, New York.
Lieberson, S., 1981. An asymmetrical approach to segregation. In: Peach, C., Robinson, V., Smith, S. (Eds.), Ethnic Segregation in Cities. Croom Helm, London, pp. 61–82.
Massey, D.S., 1985. Ethnic residential segregation: a theoretical synthesis and empirical review. Sociology and Social Research 69, 315–350.
Massey, D.S., 2014. Segregation and the perpetuation of disadvantage. In: Burton, Linda, Brady, David (Eds.), The Oxford Handbook of Poverty and Society. Oxford University Press, Oxford.
Massey, D.S., Denton, N.A., 1988. The dimensions of residential segregation. Social Forces 67 (2), 281–315.
Massey, D.S., Denton, N.A., 1989. Hypersegregation in U.S. metropolitan areas: black and Hispanic segregation along five dimensions. Demography 26, 373–393.
Massey, D.S., White, M.J., Phua, V.C., 1996. The dimensions of segregation revisited. Sociological Methods and Research 25, 172–206.
Morgan, B.S., 1982. The properties of a distance-based segregation index. Journal of Socioeconomic Planning Sciences 16 (4), 167–171.
Morgan, B.S., 1983. A distance-decay interaction index to measure residential segregation. Area 15 (3), 211–216.
Park, R.E., 1925. The city: suggestions for the investigation of human behavior in the urban environment. In: Park, Robert E., Burgess, Ernest W. (Eds.), The City. University of Chicago Press, Chicago, pp. 1–46.
Reardon, S.F., Bischoff, K., 2011. Income inequality and income segregation. American Journal of Sociology 116 (4), 1092–1153.
Rugh, J.S., Massey, D.S., 2013. Segregation in post-civil rights America: stalled integration or end of the segregated century? The DuBois Review: Social Science Research on Race 12 (1), 1–28. http://dx.doi.org/10.1017/S1742058X13000180.
Taeuber, K.E., Taeuber, A.F., 1965. Negroes in Cities: Residential Segregation and Neighborhood Change. Aldine, Chicago, IL.
Taeuber, K.E., Taeuber, A.F., 1976. A practitioners perspective on the index of dissimilarity. American Sociological Review 42 (6), 884–889.
Theil, H., 1972. Statistical Decomposition Analysis. North Holland, Amsterdam.
White, M.J., 1983. The measurement of spatial segregation. American Journal of Sociology 88, 1008–1019.
Wilkes, Rima, Iceland, John, 2004. Hypersegregation in the twenty-first century: an update and analysis. Demography 41 (1), 23–36.

Selection Bias, Statistics of

Richard Breen, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 20, pp. 13795–13799, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The problem of selection bias is pervasive in the social and behavioral sciences. This article defines the problem and provides examples of where it arises. Since the 1980s, the Heckman two-step method has been the most popular method of correcting for selection bias. It is explained and the criticisms made of it are discussed. In recent years, much attention has focused on selection bias in the nonexperimental evaluation of programs. The issues in this area are briefly outlined; selection bias proves to be only one of several possible sources of bias in these circumstances. The article ends with some practical guidelines for using selection bias correction estimators in empirical research.

In almost all areas of scientific inquiry, researchers often want to infer, on the basis of sample data, the characteristics of a relationship that holds in the relevant population. Frequently, this is the relationship between a dependent variable and one or more explanatory variables. In some cases, however, although one may have complete information on the explanatory variables, information on the dependent variable is lacking for some observations or 'units.' Furthermore, whether or not this information is present may not be conditionally independent (given the model) of the value taken by the dependent variable itself. This is a case of selection bias.

Selection Bias

The following illustrates the basic problem. Suppose one hypothesizes the following relationship, in the population, between a dependent variable, Y^*, and a set of $j = 1,\ldots, J$ explanatory variables, X:

$$y_i^* = x_i'\beta + \varepsilon_i \qquad [1]$$

Here the unit-specific values of Y^* are denoted as y_i^* (where i denotes the ith observation) and each unit's values of the X variables are contained in the vector x_i. ε is a random error, and the parameter vector, β, is to be estimated. Specific assumptions about the distribution of ε will determine the choice of estimation technique.

If Y^* is considered as a latent variable, a variety of different statistical models can be generated from eqn [1] depending on the manner in which Y^* is actually observed (though still, at this point, assuming that Y^* is observed, in some fashion, for all units). To do this, write a second equation that defines the relationship between the observed variable, Y, and the latent variable, Y^*. For example, where Y^* is completely observed.

$$y_i = y_i^* \qquad [2a]$$

But one might also have

$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* > 0 \\ y_i &= 0 \quad \text{otherwise} \end{aligned} \qquad [2b]$$

In the case of eqn [2b], Y is the binary observed realization of the underlying latent variable, Y^*. More complex relationships between Y^* and Y are also possible.

To capture the idea of selection bias, suppose that there is another latent variable Z^* such that whether or not we observe a value for Y depends on the value of Z^*, which is given by

$$z_i^* = w_i'\alpha + \nu_i \qquad [3]$$

Here w is a vector of explanatory variables with coefficients α, and ν is a random error term. The observed variable Z is defined as

$$\begin{aligned} z_i &= 1 \quad \text{if } z_i^* i > 0 \\ z_i &= 0 \quad \text{otherwise} \end{aligned} \qquad [4]$$

Finally, define the observation equation for Y as depending on Z^* as follows:

$$\begin{aligned} y_i &= y_i \qquad\qquad \text{if } z_i = 1 \\ y_i &\ \text{unobserved} \quad \text{if } z_i = 0 \end{aligned} \qquad [5]$$

Equations [1], [2a], [2b], and [5], together with eqns [3] and [4], link the latent variable, Y^*, to its observed counterpart, Y, when observation of the latter depends on the value of another variable, Z^*. Selection bias occurs when there is a nonzero covariance between the error terms ε and ν. More complex selection bias models can be generated by having, for instance, more than one Z^* variable. The censored regression (or Tobit) model is a simpler case in which whether Y is observed or not depends on whether it exceeds (or, in other cases, falls below) a given threshold value.

The Problem

Selection bias is a problem because if one tries to estimate β using normal statistical methods, the resulting estimates will be poor and potentially misleading. For example, if eqn [2a] specifies the relationship between Y and Y^*, discarding cases where Y is unobserved and running an ordinary least squares (OLS) regression on the remainder will give estimates that are biased and inconsistent.

Where might this kind of selection bias arise? Suppose one draws a random sample from a population and carries out a survey. Some respondents may refuse to answer some, though not all, questions. If the response to such a question played the role of dependent variable in a regression analysis, there would be a selection bias problem if the probability of having responded to the item was not independent of the response value, given the specification of the regression model (Horowitz and Manski, 1998). This may not always be so: in the simplest case, nonresponse might be random. Alternatively,

response might be independent of the response value controlling for a set of measured covariates. In this case there is selection on observables (Heckman and Robb, 1985). But often neither of these cases holds, and a nonzero residual correlation between the selection eqn [3] and the outcome eqn [1] means that there is selection on unobservables.

In studies of the criminal justice system in which the population includes all those charged with a crime, sentence severity is observed only for those found guilty. In studies of the effectiveness of university education, the outcome (say, examination results) is only observed for those who had completed that period of education. In studies of women's earnings in paid work, the earnings of women who are not in the labor force cannot be observed. To these examples could be added very many more. Selection bias is a pervasive problem. As a consequence, a variety of methods has been proposed to deal with it.

The problem of selection bias arises because of the nonzero correlation of the errors in eqns [1] and [3], and this arises commonly in the following context. Suppose there is a nonexperimental comparison of two groups, one exposed to some policy treatment, the other not. In this case, the outcome of interest, Y, could be observed for members of both groups. Let Z now be the indicator of group membership (so that, for instance, $Z = 0$ means membership of the comparison group and $Z = 1$ means membership of the treatment group). In the case where eqn [2a] holds, write the equation for Y as

$$y_i = x_i'\beta + \gamma z_i + \varepsilon_i \qquad [6]$$

and interest would focus on the estimate of γ as the gross effect on the outcome measure of being in the treatment, rather than the comparison, group. The problem of selection bias still arises to the extent that ε and ν have a nonzero correlation. The difference between this and the type of selection bias discussed initially (namely, that now Y is observed for both groups) is more apparent than real as an alternative formulation shows. For each unit Y is observed as the response to membership of either the treatment or comparison group, but not both. Let Y_1 be the outcome given treatment and Y_0 given no treatment. Then the ith individual unit's gain from treatment is simply $\Delta_i = Y_{1i} - Y_{0i}$. But Δ cannot be measured, since one cannot simultaneously observe a unit's values of Y_1 and Y_0. Instead

$$Y_i = ZY_{1i} + (1 - Z)Y_{0i} \qquad [7]$$

is observed. Equation [7] thus specifies the incomplete observation of two latent variables, Y_0 and Y_1. This setup can easily be extended to cases with more than two groups. As might be expected, this approach is commonly encountered in the evaluation literature, but one of its earliest formulations was by Roy (1951) as the problem of self-selection. More recent works treat the problem of selection as a particular case of a broader set of challenges in the identification of causal effects (Pearl, 2009; Rubin, 2005).

The Solutions

Broadly speaking, there are two ways to address problems of selection bias: *ex ante* by research design and *ex post* by statistical adjustment. The modern literature on correcting for selection bias is itself biased toward the latter, beginning with the work of Heckman in the 1970s (Heckman, 1976, 1979), whose so-called 'two-step' method is widely used and hardwired into many econometrics programs. The argument that underlies this technique is the following: Assume the setup as defined by eqns [1]–[5] and a nonzero covariance between ε and ν. Then, we can write the 'outcome equation' (i.e., the regression equation for Y, given its observability) as

$$E(y_i|z_i = 1, x_i) = x_i'\beta + E(\varepsilon_i|z_i = 1) \qquad [8]$$

Because there are only observations of Y when $z = 1$, there is an extra term for the conditional expectation of ε. Because of the nonzero covariance between ε and ν and if, as is usually assumed, $E(\varepsilon) = 0$, this conditional expectation cannot be zero. One can write

$$E(y_i|z_i = 1, x_i) = x_i'\beta + E(\varepsilon_i|\nu_i > w_i'\alpha) \qquad [9]$$

Using a standard result for the value of a truncated bivariate normal distribution, the second term on the right-hand side of this equation is given by

$$E(\varepsilon_i|\nu_i > w_i'\alpha) = \rho\sigma_\varepsilon\sigma_\nu\frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)} \qquad [10]$$

Here the σ's are the standard deviations of the respective error terms from eqns [1] and [3]; ρ is the correlation between them; ϕ and Φ are, respectively, the density and distribution functions of the standard normal and their ratio, as it appears in eqn [10], is termed the 'inverse Mill's ratio.' Heckman's two-step method requires first running a probit regression with Z as the dependent variable, estimated using all the observations in the data. Then, for those observations for which Y is observed, the probit coefficient estimates are used to compute the value of the inverse Mill's ratio. This is then inserted as an extra variable in the OLS regression in which Y is the dependent variable to give

$$E(y_i|z_i = 1, x_i) = x_i'\beta + \theta\widehat{\lambda}_i \qquad [11]$$

where λ is the inverse Mill's ratio and the 'hat' indicates that it is estimated. Its coefficient, θ, is then itself an estimate of $\rho\sigma_\varepsilon\sigma_\nu$. This is the covariance between ε and ν (σ_ν is set to unity in the probit). This approach is extended readily to the case in which an outcome is observed for both groups ($Z = 0$ and $Z = 1$).

What are the assumptions of this approach and what are the properties of the resulting estimator when we apply this method to data from a sample of the population? First, it is assumed that all the other requirements for an OLS regression are met. Second, in the setup outlined above, the joint distribution of ε and ν should be bivariate normal (though, in general, weaker assumptions suffice, namely that ν be normally distributed; and the expectation of ε, conditional on ν, should be linear in ν (see Olsen, 1980)).

Given that the assumptions hold, the two-step estimator yields consistent estimates of the population β in eqn [1]. The standard errors are incorrect (due to heteroscedasticity and the use of an estimated value of λ), but these are corrected readily. An alternative to the two-step approach is to estimate both the selection and outcome equations simultaneously using maximum likelihood (ML). The resulting estimates are asymptotically unbiased and more efficient than those from

the two-step method (see Nelson (1984) for a comparison of OLS, the two-step method, and ML).

This two-step method is, in any case, applicable only when the relationship between Y^* and Y is as given by eqn [2a]. When this relationship is given by, for example, eqn [2b], the two-stage method is inconsistent. In these cases, ML is the most feasible option. If the outcome variable is itself binary, the joint log-likelihood of the selection and outcome equations has the same general as for a bivariate probit but one in which there are three, rather than four, possible outcomes. They are $z = 1$ and $y = 1$; $z = 1$ and $y = 0$; and $z = 0$ (in which case y is not observed).

Although the two-step method is probably the most widely used approach to correcting for selection bias, it has been subjected to much criticism. Among the main objections are the following:

1. *Sensitivity to distributional assumptions*. Practical implementation of the method renders it particularly sensitive in this respect. If the assumptions are not met, the estimator has no desirable properties (i.e., it is not even consistent).
2. *Identification and robustness*. It is common to find that the estimate of the inverse Mill's ratio is highly correlated either with the explanatory variables or with the intercept in eqn [11]. The extent to which such problems will arise depends mainly on three things. They are
   1. the specification of the selection equation;
   2. the sample correlation between the two sets of explanatory variables, X and W (call this q);
   3. the degree of sample selection in the sample (p, the proportion of cases for which $z = 1$).

For example, if X and W are identical, the two-equation system is identified only because of the nonlinearity of the probit equation. But for some ranges of p the probit function is almost linear, with the result that the estimated λ will be close to a linear function of the explanatory variables in the probit; and thus it will be highly correlated with the X variables in the outcome equation. In general, the correlation between the inverse Mill's ratio estimate and the other explanatory variables in this equation will be greater, the greater is q and the closer is p to zero or one (this issue is discussed in detail in Breen, 1996). If the selection equation does not discriminate well between the selected and unselected observations, the estimated inverse Mill's ratio will be approximately a constant, and there will therefore be a high correlation between it and the intercept of the outcome equation.

Both these objections reflect genuine difficulties in applying the model. In principle, the solution to the identification and robustness problems is simple: ensure that the probit equation discriminates well and does not rely on the nonlinearity of the probit for identification. In practice it may be rather difficult to achieve these things. On the other hand, the issue of distributional assumptions may be even less tractable, obliging recourse to semiparametric or other approaches (Manski, 1995; Lee, 1994; Cosslett, 1991).

An alternative is to try to assess the likely degree of bias that sample selection might induce in a given case. Rubin (1977) presents a Bayesian approach in which the investigator computes the likely degree of bias, conditional on a prior belief about the relationship between the parameters of the distribution of Y in the selected and nonselected samples. The mean and the variance of the latter sample might, for example, be expressed as a function of their values in the selected sample, conditioning on the values of observed covariates. It is then straightforward to express the extent of selection bias for plausible values of the parameters of this function and to place a corresponding Bayesian probability interval around any estimates. Rosenbaum (1995, 1996) suggests and illustrates other sensitivity analyses for selection bias in nonexperimental evaluations.

Program Evaluation and Selection Bias

For several years now the literature on selection bias has been dominated by discussion of how to evaluate programs (such as job training programs) when randomized assignment to treatment and control group is not possible (Imbens and Wooldridge, 2009). In a very widely cited paper, Lalonde (1986) compared the measured effectiveness of labor market programs using a randomized assignment design with their effectiveness computed using the Heckman method and showed that they led to quite different results (but see also Heckman and Hotz, 1989). While some have seen this as a fatal criticism of the method, one consequence has been the development of greater awareness of the need to ensure the suitability of different selection bias correction approaches for specific cases. Another has been a greater concern to identify other possible sources of bias in nonrandomized program evaluations.

In the statistical literature, matching methods commonly are advocated in program evaluations in the absence of randomization. They involve the direct comparison of treated and untreated units that share common, or very similar, values on a set of relevant covariates, X. The rationale for this is the assumption that, conditional on X, and assuming only selection on observables (all of which are in X), the observed outcome for nonparticipants has the same distribution as the unobserved outcome that participants would have had, had they not participated (see Heckman et al., 1997). This then allows one to estimate Δ. In passing, note that matching is also used commonly to impute missing values in item nonresponse in surveys (Little and Rubin, 1987). A central issue is how to carry out such matching. If there are K covariates, then the problem is one of matching pairs, or sets, of participants and nonparticipants in this K-dimensional space. But by a result due to Rosenbaum and Rubin (1983) (see also Rosenbaum, 1995), matching using a scalar quantity called the propensity score is equally effective. The propensity score is simply the probability of being in the treatment, rather than the comparison group, given the observed covariates. Matching on the propensity score controls bias due to all observed covariates, and even if there is selection on unobservables, the method nevertheless produces treatment and comparison groups with the same distribution of X variables. This is so under the assumption that the true propensity score is known, but Rosenbaum and Rubin (1984) show that an estimated propensity score (typically using a logit model) appears to perform at least as well.

The use of matching draws attention to the problem of the possibly different distributions of covariates (or, equally, of propensity scores) in the treatment and comparison groups. There are two important aspects: First, the support of the propensity score may be different, so some ranges of propensity score values may be present in one group and not the other. More simply, some participants may have no comparable nonparticipants. Second, the distributions of the set of common values of propensity scores (i.e., which appear in both groups) may be different (Heckman et al., 1997, 1998). Hitherto, in practice, both of these sources of bias in evaluating the program's effects commonly would have been confounded with selection bias proper. The use of propensity scores can remove the former but, if there is also selection on unobservables (i.e., selection bias proper), matching cannot be expected to solve the problem (Caliendo and Kopeinig, 2008; Imbens, 2014). Indeed, it is possible that these different forms of bias may have opposite signs, so that correcting for some but not others may make the overall bias greater. Nevertheless, the ability to draw such finer distinctions among biases is a valuable development, not least because it allows one to focus on methods for controlling selection bias, free from the contaminating effects of other biases.

Conclusion

This article has only skimmed the surface of selection bias statistics. In particular, it has looked only at models for cross-sectional data: other approaches are possible given other sorts of data. For example, given longitudinal data, with observations of Y for each unit at two (or more) points in time, one prior to the program to be evaluated and the other after it, then it may be possible to specify and test a model which assumes that unobserved causes of selection bias are unit-specific and time-invariant, and can therefore be removed by differencing (Heckman and Robb, 1985; Heckman et al., 1997, 1998). Considerations of this sort draw attention to the *ex ante* approach to dealing with selection bias that was referred to earlier. Here the emphasis falls on designing research so as to remove, or at least reduce, selection bias. Random assignment is one possibility, though not always feasible in practice. Nevertheless, quasi-experimental approaches, interrupted time series, and longitudinal designs with repeated measures on the outcome variable are among a range of possibilities that could be considered in designing one's research to minimize selection bias problems (see Rosenbaum (1999) and associated comments). Even so, there will still be many instances in which analysts use data over whose collection they have had no control and where the use of *ex post* adjustments will be unavoidable.

To conclude, much work on the selection bias problem has been undertaken since the 1980s, yet there is no widespread agreement on which statistical methods are most suitable for use in correcting for the problem. Some broad guidelines for practitioners can, however, be discerned as follows:

1. the need to ensure that methods to correct for selection bias are appropriate and that their requirements (such as distributional properties) are met;
2. the need to be aware of other, possibly confounding, sources of bias (Bareinboim and Pearl, 2012);
3. the usefulness of analyses of the sensitivity of conclusions to various possible magnitudes of selection bias, and of the sensitivity of the selection bias corrected results to the assumptions of whatever method is employed; and
4. the desirability of designing research so that selection bias problems are, as far as possible, eliminated without the need for complex and sometimes fragile *ex post* adjustment.

See also: Causal Inference: Overview; Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Mechanism-Based Causal Analysis; Probability and Chance: Philosophical Aspects; Quasi-Experimental Designs; Screening and Selection.

Bibliography

Bareinboim, E., Pearl, J., 2012. Controlling selection bias in causal inference. Journal of Machine Learning Research 22, 100–108.
Breen, R., 1996. Regression Models: Censored, Sample-Selected or Truncated Data. Sage, Thousand Oaks, CA.
Caliendo, M., Kopeinig, S., 2008. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys 22 (1), 31–72.
Cosslett, S., 1991. Semiparametric estimation of a regression model with sample selectivity. In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, UK, pp. 175–197.
Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.
Heckman, J.J., Hotz, V.J., 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. Journal of the American Statistical Association 84 (408), 862–874.
Heckman, J.J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York.
Heckman, J.J., Ichimura, H., Todd, P.E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64, 605–654.
Heckman, J.J., Ichimura, H., Smith, J., Todd, P.E., 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.
Horowitz, J.L., Manski, C.F., 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. Journal of Econometrics 84, 37–58.
Imbens, G.W., 2014. Matching Methods in Practice: Three Examples (No. w19959). National Bureau of Economic Research.
Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics program evaluation. Journal of Economic Literature 47, 5–86.
Lalonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, 604–620.
Lee, L.F., 1994. Semi-parametric two-stage estimation of sample selection models subject to Tobit-type selection rules. Journal of Econometrics 61, 305–344.
Little, R.D., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.
Manski, C.F., 1995. Identification Problems in the Social Sciences. Harvard University Press, Cambridge.
Nelson, F.D., 1984. Efficiency of the two-step estimator for models with endogenous sample selection. Journal of Econometrics 24, 181–196.
Olsen, R.J., 1980. A least squares correction for selectivity bias. Econometrica 48, 1815–1820.

Pearl, J., 2009. Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge.

Rosenbaum, P.R., 1995. Observational Studies. Springer-Verlag, New York.

Rosenbaum, P.R., 1996. Observational studies and nonrandomized experiments. In: Ghosh, S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 13. Elsevier, Amsterdam, pp. 181–197.

Rosenbaum, P.R., 1999. Choice as an alternative to control in observational studies. Statistical Science 14, 259–304.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P.R., Rubin, D.B., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.

Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3, 135–146.

Rubin, D.B., 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association 72, 538–543.

Rubin, D.B., 2005. Causal inference using potential outcomes. Journal of the American Statistical Association 100, 322–331.

Self and Brain

Manos Tsakiris, Royal Holloway, University of London, Egham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Questions that have been previously asked only in philosophical investigations, such as the nature of the self, have now become accessible to quantitative life and social sciences. Over the past decades, significant advances have enabled the cognitive neurosciences to study the neurocognitive basis of different aspects of the self, such as the affective self, the sensorimotor self, the ability to self-recognize and the distinction between self and other. The available neurocognitive data suggest that there is no single specialized neural system or a single neurocognitive function underlying selfhood. Instead, multiple unimodal and multimodal brain areas underlie distinct aspects of selfhood, such as the affective and sensorimotor self, the autobiographical and the social self. The discovery of a unifying neurocognitive basis of the self remains elusive, yet essential.

Introduction

Questions that have been previously asked only in philosophical investigations, such as the nature of the self, have now become accessible to quantitative life and social sciences. Over the past few decades, significant advances in research methods and experimental designs have enabled the cognitive neurosciences to study the neurocognitive basis of the self. However, the empirical study of the self is hindered by the multitude of selves described in psychological sciences. For example, Strawson (2009) counted up to 25 different forms of self, such as the 'cognitive,' the 'conceptual,' the 'core,' the 'embodied,' the 'empirical,' the 'interpersonal,' the 'material,' the 'narrative,' and the 'private' self to name a few. The fact that so many notions of self exist in psychology raises important methodological questions that require careful consideration when investigating how the self is represented in the brain. In addition, in similarity with other domains of information processing (e.g., language), questions of anatomical specificity (e.g., the existence or otherwise of self-specific areas or networks in the brain), functional uniqueness (e.g., whether the self is being processed in a unique way), functional independence (e.g., whether self-processing is independent from processing of other information), and species specificity (e.g., whether other species have a sense of self) are inevitably linked with current neuroscientific attempts to find the self in the brain.

A useful distinction often made in the neuroscience of self is between physical and psychological components (see Gillihan and Farah, 2005). The physical component corresponds to the affective and sensorimotor experience of one's own body, whereas the psychological component corresponds not only to the experience of one's own identity over time, but also to the experience of acting and interacting with the world. The importance of the distinction between the physical and the psychological components of selfhood lies in its applicability in neuroscientific research. Even though the physical and the psychological aspects are not experienced separately, they can be methodologically separated, enabling the empirical investigation of distinct subcomponents of selfhood, such as emotional self-awareness, awareness of one's own body and one's own actions, awareness of one's own identity over time and the awareness of the self as distinct from other people.

The Affective Self

The concept of an affective or 'core self' is based on the hypothesis that awareness of one's emotions plays a crucial adaptive role. The brain is constantly required to mediate between compelling demands arising from the internal milieu of the body, on the one hand, and the practical constraints of external reality on the other. Central to this negotiation between the inner and the outer is the neural processing of emotions that are closely linked to changes in the physiological states of the internal milieu. Feeling an emotion seems to be a simple matter, consisting of mental images arising from the neural patterns that represent the changes in the body that make up an emotion. However, having consciousness of that feeling, might have been the crucial step in the development of human self-consciousness because awareness of one's own emotions led to environmentally adaptive behaviors that went beyond simple automatic responses. The representation of internal body states is perhaps the most fundamental element of self-awareness, because the interoceptive system is used to remap via subcortical structures in the cerebral cortex the most stable aspects of the one's environment, that is, the organism's body (Damasio, 2010).

In line with this view, a primary 'core' quality of the self (Damasio, 2010) lies in the awareness of feelings that represent the affective interaction of one's own body with the environment. This type of awareness can be thought of as an inner sense directed toward one's own body, an idea going back to Locke, Brentano, and James. Interoceptive representations, that provide the content of this inner consciousness, include distinct feelings originating from the body, such as pain, temperature, itch, sensual touch, muscular and visceral sensations, hunger, and thirst among others. Recent neuroimaging studies attempted to study the neural correlates of interoceptive awareness. For example, Critchley et al. (2004) measured brain activity while participants judged the timing of their own heartbeat. Brain

activity in the right anterior insular cortex predicted the participants' accuracy at detecting their heartbeat. Converging evidence suggests that the insular cortex is responsible for representing interoceptive activity. These interoceptive representations of internal body-states seem to provide the basis for affective self-awareness. In addition to the insular cortex, other brain structures participate in the construction of first-order maps of the body, such as the hypothalamus, the brain-stem nuclei, and the somatosensory cortices. While these structures are used to represent body states, second-order maps, instantiated in the cingulate cortex and the thalamus are needed for integration awareness of the internal body with perception of external objects. These second-order maps will represent the relation between the inner and the outer world, serving an essential adaptive role.

The Sensorimotor Self

Another principal quality of the self is grounded in the experience of the body as a sensory and motor entity that belongs to 'me,' and that can be perceived not only within via interoception, but also from outside via exteroception. Two fundamental elements characterize this experience, the sense of body-ownership and the sense of agency (Gallagher, 2000).

The Sense of Body-Ownership

The feeling that the body I inhabit is mine and always with me is called body-ownership. Even though in everyday life the link between the body and its self is taken for granted, the cognitive and neural processes by which the body is experienced as linked to the self are far from being fully understood. Clinical cases of abnormal body-awareness aptly demonstrate the need for a neuroscientific account of the bodily self. Somatoparaphrenia is a neurological condition, which is usually related to anosognosia for hemiplegia, occurring after predominantly right hemispheric lesions. Patients with somatoparaphrenia believe that their limb contralateral to the side of the lesion belongs to someone else, and the disorder is often accompanied by inability to feel tactile sensations in the 'non-belonging' part of the body.

The rubber hand illusion (RHI) emerged as an experimental paradigm that allows the controlled manipulation of the experience of body-ownership. In brief, watching a rubber hand being stroked synchronously with one's own unseen hand causes the rubber hand to be attributed to one's own body, to 'feel like it is my hand' (Botvinick and Cohen, 1998). This illusion does not occur when the rubber hand is stroked asynchronously with respect to the subject's own hand. One behavioral correlate of the RHI is a change in the perceived location of the participant's own hand toward the rubber hand. Botvinick and Cohen (1998) showed that, after synchronous visuotactile stimulation of the rubber hand and the participant's hand, participants perceived the position of their hand to be closer to the rubber hand than it really was. Interestingly, the prevalence of illusion over time (Botvinick and Cohen, 1998) and the subjective intensity of the experience of body-ownership are positively correlated with changes in the felt location of the subject's own hand toward the rubber hand. Finally, the effects of multisensory integration between the rubber hand and the real hand have been shown to be specific to body parts, as synchronous multisensory stimulation between a non-corporeal object and the real hand does not lead to changes in body-ownership (for a review, see Tsakiris, 2010).

The RHI paradigm is one of the few viable ways of investigating body-ownership scientifically because it allows for an external object to be treated, rather than simply recognized, as a part of my body. The brain processes that produce the sense of body-ownership depend both on current sensory integration and also on the contribution of internal models of the body. The right temporoparietal junction (rTPJ) may underpin the assimilation of novel multisensory signals by maintaining a preexisting reference representation of one's own body. The rTPJ may underpin decisions about the incorporeability of external objects by testing the extent to which these objects fit with a reference body-model (see Tsakiris, 2010 for a review). The primary and associative somatosensory cortices seem to maintain online anatomical and postural representations of the current state of the body against which the anatomical and postural features of the stimulated object are compared. Body-related sensory integration linked to the onset of body-ownership during the RHI is related to activity in the posterior parietal cortex (PPC) and the ventral premotor cortex (Ehrsson et al., 2004). The effect of multisensory integration and recalibration of hand position, namely the experience of body-ownership of the rubber hand, has also been shown to correlate with activity in the right posterior insula (Tsakiris et al., 2007). Interestingly, a recent lesion mapping study that focused specifically on patients with somatoparaphrenic symptoms (Baier and Karnath, 2008) revealed that the right posterior insula is indeed the critical structure involved in phenomena of 'disturbed sensation of limb ownership.' Other authors have also suggested that the insula is concerned with higher order somatosensory processing of the body that is related to a subjective awareness and affective processing of bodily signals (Craig, 2009). The available imaging evidence on the RHI is consistent with the view that secondary somatosensory cortex and the insula are responsible for conscious somatosensory perception, with the right PPC contributing to spatiotemporal integration. Circuitry and connectivity analyses demonstrate both the afferent and efferent connectivity of the insula with SII, the temporoparietal area and the premotor cortex. These structures may form a network that plays a fundamental role in linking current sensory stimuli to one's own body, and thus also in self-awareness. Consistent results have been obtained more recently in studies that extended the RHI paradigm to full body illusions. As with the RHI, neuroimaging studies have implicated the premotor cortex and temporoparietal areas in the experience of illusory ownership over a full body (for a review, see Blanke, 2012).

The Sense of Agency

The studies on the sense of body-ownership highlight the importance of multisensory integration in creating a sensory sense of self. However, our bodies are first and foremost acting agents. The sense of agency is the sense of intending and

executing an action, a sense of oneself as an actor who can make things happen. One example taken from the psychiatric literature aptly illustrates how the sense of agency is linked to specific functional and neural processes. Schizophrenic patients with delusions of control demonstrate a striking failure to experience their own agency over the actions that they execute. The delusion of control is an example of a passivity experience in which a patient feels that his own actions are being created, not by himself, but by an outside force, usually an external agent. Increased activation in the right parietal cortex was related to the experience of passivity phenomena in schizophrenic patients (Spence et al., 1997). The main feature of this symptom is that the intention to act is misattributed to another agent, whereas the ownership of the body-part that executes the action is not.

At the neurocognitive level, delusions of control may arise as a result of an abnormal awareness of the predicted consequences of one's own action (Blakemore et al., 2002). An important aspect of agency involves the ability to distinguish actions and effects that are self-generated from those generated by other agents. The ability to distinguish between self-generated and externally generated events is embedded in the function of the motor system. In the case of a self-generated action, intentions and the efference copy of the motor command can be used to predict the consequent visual, auditory, and somatosensory signals produced by one's own movements. An internal 'forward model' of the motor system uses the efference copy of the motor command to predict the sensory effects of one's own actions. Once the actual sensory feedback arrives, it is compared with the prediction. If there is no mismatch between the predicted and actual feedback, the action is attributed to the self because the sensory event is recognized as being self-generated.

Several different experimental paradigms and clinical populations have been used to investigate the neural signatures of the sense of agency. Some have looked at the preaction processes, such as the formation of an intention and the preparation-selection of action, while others have focused on postaction processes that focus on the perception of action effects. A brain network that involves prefrontal and parietal areas has emerged as central to our experience of agency. In general, methods similar to the body-ownership literature, such as the systematic manipulation of visual feedback to alter the experience of one's body in action, have been used in agency research.

The supplementary motor area (SMA) is considered to be the brain structure responsible for translating motivational states originating from anterior cingulate cortex into intentional motor programs. SMA activity precedes M1 activation during the generation of voluntary movement (Deecke, 1996), and this activity is present even when a planned movement does not occur. Moreover, SMA activity seems to reflect a higher level representation of motor preparation and execution, as it is not correlated to the actual muscle activity. More direct evidence regarding the role of SMA in translating intentions into actions comes from Fried et al. (1991) who studied the functional organization of SMA in 13 patients with intractable epilepsy. Four patients who received direct electrical stimulation in SMA, below the threshold of afterdischarges, reported a subjective 'urge' to move or an anticipation that a movement was about to occur. At some of the sites where this 'urge to move' was elicited, higher current stimulation produced overt motor behavior. Overall, these observations argue in favor of Goldberg's view that SMA functions in a 'supramotor' fashion, and that it "is being engaged earlier than M1 in the translation of motive to intention to action and exerting control over M1" (Goldberg, 1985).

Given that a certain goal can be achieved in various ways, the selection of the motor program that will be executed is a central problem in the generation of intentional actions. When people are asked to perform open-ended willed actions, increased activation in the left dorsolateral prefrontal cortex (DLPFC) was found, as opposed to when people perform routine acts as response to stimuli. Such findings suggest that the DLPFC is 'sculpting the response space' from which then we can actively choose an action according to our own will (Frith, 2000).

In addition to these frontal areas, the parietal cortex is also involved in our sense of agency because of its multimodal role, whereby both efferent and multimodal afferent signals converge in order to represent accurately the current state of the organism in relation to its environment. Andersen and Buneo (2002) argued for the presence of various 'intentional maps' in the PPC, which correspond to specific goals for orienting actions. The common functional principle of the PPC is to transform multisensory signals into appropriate motor plans or 'motor intentions.' Parietal cortex is also involved in the online monitoring and perception of actions and their consequences. Fink et al. (1999) investigated the neural responses during a conflict between the intention to move and the afferent feedback. In some trials subjects performed out-of-phase bimanual movements, but they saw in-phase movements. The main effect of out-of-phase movements (i.e., a mismatch among intention, proprioception and visual feedback) was associated with increased activity in PPC and DLPFC bilaterally. It seems therefore that parietal cortex plays a pivotal role in detecting mismatches between the intended and the actual movement, especially in cases where there is ambiguous visual feedback. The parietal cortex functions as a neural workspace that facilitates the integration of efferent, reafferent and ex-afferent signals. Overall, it has been suggested that the intact function of frontoparietal neural loops is necessary for the generation, execution, and perception of intentional actions, processes that are essential for the normal experience of agency.

Self-Recognition

The face is the most distinctive feature of our appearance, the one by which we come to be known to ourselves and to others as an individual. The importance of faces to the self-concept and for social cognition leads to their physical features being processed in a specialized neural circuit. In turn, there is some evidence to suggest that representations of one's own facial appearance may be stored in a similarly specialized network of areas, which is engaged when a face that is recognized as one's own is processed (Devue and Bredart, 2011). In a typical self-face recognition experiment, participants are presented with

pictures of morphed faces that contain varying degrees of the participants' face and the face of another person with whom the participant was personally familiar (e.g., 0, 20, 40, 60, 80, or 100% of the participant's own face), and participants are asked to judge whether the face looks more like their own or not. A study by Uddin et al. (2005) reported that portions of the inferior parietal lobe (IPL) and inferior frontal gyrus (IFG) showed a parametric response, increasing with the amount of self-face in the stimuli. They also reported activity in the ventromedial portions of the frontal lobe, the precuneus, the superior frontal gyrus, the superior parietal lobe (SPL), and medial portions of a region covering the right inferior temporal gyrus (ITG) and the right middle temporal gyrus (MTG), which showed an increasingly negative response the more that stimuli contained of the self-face. A recent review of the literature highlights that these regions are those that are typically reported as activated in studies that investigate the neural basis of current self-face recognition (Devue and Bredart, 2011). More recently, Apps et al. (2012) used functional magnetic resonance imaging to investigate the neural circuits involved in the recognition of the different faces of one's self. Participants viewed images of either their own face as it currently looks morphed with the face of a familiar other or their childhood face morphed with the childhood face of the familiar other. Activity in areas that have a generalized selectivity for faces, including the inferior occipital gyrus, the superior parietal lobule, and the ITG, varied with the amount of current self-face in an image. Activity in areas involved in memory encoding and retrieval, including the hippocampus and the posterior cingulate gyrus, and areas involved in creating a sense of body-ownership, including the temporoparietal junction (TPJ) and the inferior parietal lobule, varied with the amount of childhood self in an image. Therefore, recognizing one's past and current facial appearances relies on processing in distinct neural circuits. Recognition of one's physical appearance from the past engages areas involved in body-ownership and recognition memory. This suggests that identifying a past facial appearance as one's own requires the physical features of the face to be retrieved from memory and coded for as a part of one's body. In contrast, a representation of one's current self-image is maintained and updated through plastic processes in areas of the brain that are specialized for processing faces, but not specifically one's own face. These findings pave the way for understanding the neural mechanisms underlying the plasticity of self-representations as a result of changes in appearance due to normal aging as well as due to traumatic events or reconstructive surgery.

More general, self-face, voice, and body recognition have been shown to activate unimodal regions within the core face processing network including: the inferior occipital gyrus, the fusiform gyrus, as well as face selective areas in the superior temporal sulcus (STS), portions of the auditory system in the superior temporal gyrus, and portions of extrastriate cortex that process body parts. However, it is well established that these areas process information about all faces, bodies, or voices, even if the profile of their response is different between self and other. Such findings would suggest that these regions are not processing anything that is specific to the self. Neuroimaging studies suggest that unimodal self-face, voice, and body recognition also activate a broad range of multimodal areas. The list of areas includes: the posterior cingulate gyrus, the anterior cingulate gyrus (ACG), medial portions of the superior frontal gyrus/paracingulate cortex, the TPJ, the STS, the temporal poles, the hippocampus, the anterior insula (AI), midportions of the IFG, the middle frontal gyrus (MFG), the intraparietal sulcus (IPS), and the inferior parietal lobule. Some have therefore argued that information processed in some of these regions is self-specific and it is processing in these areas that leads to a self-concept (Northoff et al., 2006). However, there is little evidence that these regions are engaged exclusively by self-recognition processes and indeed, each of these regions is not found to be activated exclusively during the processing of self-stimuli (Legrand and Ruby, 2009).

The Self-Representing Self

Our psychological sense of self-identity seems to be grounded in explicit and conceptual representations, such as one's own memories, personality traits, beliefs, and attitudes. A sense of identity over time is given by autobiographical memory. Research on autobiographical memory has not provided a definite answer as to whether autobiographical memory is functionally independent from other memory systems. Neuropsychological case studies demonstrate that it is possible to have intact autobiographical memory despite severe impairment in semantic memory system. Similarly, the evidence for specific neural correlates of autobiographical memory is inconclusive. The overall pattern across studies suggests that retrieval of autobiographical memories is associated with activations of medial frontal cortex and the left hippocampus (for a review, see Maguire, 2001). However, many of the studies have failed to control for various confounding factors between autobiographical and other memory systems (for a review, see Gillihan and Farah, 2005). Importantly, autobiographical memory provides both a sense of continuity over time and a sense of identity. The sense of continuity over time is given by the fact that autobiographical memory provides a template of the past against which new experiences can be assimilated. The sense of identity is generated by the fact that all these experiences are related to me, rather than to someone else.

Social cognitive neuroscience has focused on the neural circuits that are common to both self- and other-representations, which may be critical for social cooperation and empathy. A meta-analysis of the functional anatomy of action-execution, mental simulation of action, verbal generation of actions, and more importantly observation of other people's actions showed that for all four processes, overlapping activation was found in SMA, dorsal premotor cortex, supramarginal gyrus, and SPL (Grezes and Decety, 2001). The activation of an overlapping neural network for both the representations of self-generated and observed actions suggests that some representations are agent-neutral and shared. More evidence for common representations for self and others comes from the neural substrates of empathy. Bilateral AI, rostral anterior cingulate cortex, brainstem, and cerebellum were activated when participants received pain and also when a loved one experienced pain (Singer et al., 2004). The

existence of an overlapping neural network, or 'mirror system' that underpins both self- and other representations may support social communication and interaction. This discovery raises the interesting possibility that some aspects of the awareness of self resulted from mechanisms that initially evolved through attempts to predict the behavior of others through representing their internal states and justifies the claim that perhaps consciousness is for other people.

Future Directions

The available neurocognitive data suggest that there is no single specialized neural system or a single cognitive function underlying selfhood. Instead, multiple specialized processes underlie distinct aspects of selfhood, such as the affective and sensorimotor self, the autobiographical and the social self. These specialized processes operate at a sub-personal level and they may not be accessible to conscious awareness. The cognitive neuroscience of self is faced with the challenge of explaining how these specialized processes interact at the functional and neural level to generate a coherent and conscious sense of self.

Current empirical studies on the sense of self focus on either the physical or the psychological dimension of selfhood. Indeed, there seems to be a real difference between sensorimotor theories of self and most psychopersonal theories of self as regards the continuity of personal identity over time and autobiographical memory. Our sense of embodiment seems to be immediate and transient, while a reflective and autobiographical sense of self seems to be stable over time. Future studies need to address the neurocognitive processes that bridge the physical and the psychological dimensions of selfhood and provide a unifying neurocognitive theory of self-processing.

See also: Cognitive Neuroscience; Culture and the Self: Implications for Psychological Theory; Face Recognition Models: Computational Approaches; Face Recognition, Psychological and Neural Aspects; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Occupational Health; Parietal Lobe; Person and Self: Philosophical Aspects; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex.

Bibliography

Andersen, R.A., Buneo, C.A., 2002. Intentional maps in posterior parietal cortex. Annual Review of Neuroscience 25, 189–220.

Apps, M.A.J., Tajadura-Jiménez, A., Turley, G., Tsakiris, M., 2012. The different faces of one's self: an fMRI study into the recognition of current and past self-facial appearances. Neuroimage 63, 1720–1729.

Baier, B., Karnath, H.O., 2008. Tight link between our sense of limb ownership and self-awareness of actions. Stroke 39, 486–488.

Blakemore, S.J., Wolpert, D.M., Frith, C.D., 2002. Abnormalities in the awareness of action. Trends in Cognitive Sciences 6, 237–242.

Blanke, O., 2012. Multisensory brain mechanisms of bodily self-consciousness. Nature Reviews Neuroscience 13 (8), 556–571.

Botvinick, M., Cohen, J., 1998. Rubber hands 'feel' touch that eyes see. Nature 391, 756.

Craig, A.D., 2009. How do you feel–now? The anterior insula and human awareness. Nature Reviews Neuroscience 10, 59–70.

Critchley, H.D., Wiens, S., Rotshtein, P., Ohman, A., Dolan, R.J., 2004. Neural systems supporting interoceptive awareness. Nature Neuroscience 7, 189–195.

Damasio, A.R., 2010. Self Comes to Mind. Constructing the Conscious Brain. Heineman, London.

Deecke, L., 1996. Planning, preparation, execution, and imagery of volitional action. Brain Research: Cognitive Brain Research 3, 59–64.

Devue, C., Bredart, S., 2011. The neural correlates of visual self-recognition. Consciousness and Cognition 20, 40–51.

Ehrsson, H.H., Spence, C., Passingham, R.E., 2004. That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. Science 305, 875–877.

Fink, G.R., Marshall, J.C., Halligan, P.W., Frith, C.D., Driver, J., Frackowiak, R.S., Dolan, R.J., 1999. The neural consequences of conflict between intention and the senses. Brain 122 (Pt 3), 497–512.

Fried, I., Katz, A., McCarthy, G., Sass, K.J., Williamson, P., Spencer, S.S., Spencer, D.D., 1991. Functional organization of human supplementary motor cortex studies by electrical stimulation. Journal of Neuroscience 11, 3656–3666.

Frith, C.D., 2000. The role of dorsolateral prefrontal cortex in the selection of action, as revealed by functional imaging. In: Monsell, S., Driver, J. (Eds.), Control of Cognitive Processes: Attention and Performance, vol. XVIII. MIT Press, Cambridge, pp. 547–565.

Gallagher, S., 2000. Philosophical concepts of the self: implications for cognitive sciences. Trends in Cognitive Sciences 4, 14–21.

Gillihan, S.J., Farah, M.J., 2005. Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. Psychological Bulletin 131, 76–97.

Goldberg, G., 1985. Supplementary motor area structure and function: review and hypotheses. Behavioral and Brain Sciences 8, 567–616.

Grezes, J., Decety, J., 2001. Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. Human Brain Mapping 12, 1–19.

Legrand, D., Ruby, P., 2009. What is self-specific? Theoretical investigation and critical review of neuroimaging results. Psychological Review 116 (1), 252–282.

Maguire, E.A., 2001. Neuroimaging studies of autobiographical event memory. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 356 (1413), 1441–1451.

Northoff, G., Heinzel, A., de Greck, M., Bennpohl, F., Dobrowolny, H., Panksepp, J., 2006. Self-referential processing in our brain – a meta-analysis of imaging studies on the self. Neuroimage 31, 440–457.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., Frith, C.D., 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303, 1157–1162.

Spence, S.A., Brooks, D.J., Hirsch, S.R., Liddle, P.F., Meechan, J., Grasby, P.M., 1997. A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). Brain 120, 1997–2011.

Strawson, G., 2009. Selves: An Essay in Revisionary Metaphysics. Oxford University Press.

Tsakiris, M., 2010. My body in the brain: a neurocognitive model of body-ownership. Neuropsychologia 48 (3), 703–712.

Tsakiris, M., Hesse, M., Boy, C., Haggard, P., Fink, G.R., 2007. Neural correlates of body-ownership: a sensory network for bodily self-consciousness. Cerebral Cortex 17, 2235–2244.

Uddin, L.Q., Kaplan, J.T., Molnar-Szakacs, I., Zaidel, E., Iacoboni, M., 2005. Self-face recognition activates a frontoparietal "mirror" network in the right hemisphere: an event-related fMRI study. Neuroimage 25, 926–935.

Self and Emotional Development in Adulthood and Later Life

Gisela Labouvie-Vief, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Attention to the development of emotions and self in adulthood and aging has dramatically increased in the past half-century, and different theoretical frameworks have been presented to deal with the resulting findings. The current article reviews the major positions to have addressed these findings, and presents a view that details both strengths and vulnerabilities related to self and emotions. Thus in settings that are novel and familiar, the capacity to deal with situations of high complexity and tension is negatively affected by the process of aging; this vulnerability is offset if individuals can base themselves on skills that have been well developed and crystallized. The present article discusses these general changes over the adult life span by suggesting that while earlier development of self and emotions is characterized by an increasing capacity to deal with tension, that capacity becomes compromised and the source of vulnerability in the later part of the life span.

Introduction

Emotions are of importance throughout the total life span, from birth to late life. They play an important role in diverse areas of life. As important adaptive systems, they steer our actions in one direction or another, depending on critical circumstances. Depending on our ability to modulate and regulate them, they afford us a sense of personal well-being or lack thereof. They are of prime importance in social relations and thus constitute a basic element of our functioning within kinship and cultural systems (Magai, 2001). And finally, they constitute an important basis to our personality and sense of self.

There has been, as a consequence, much interest in the nature and development of emotions. Yet until recently, the majority of work on emotional development concerned itself with processes from infancy to childhood and adolescence. Many pioneer theories held, in fact, that development was, in essence, completed sometime around the entry into early adulthood, with little change thereafter (see Labouvie-Vief, 1994). But with increasing interest in theory and research on adult development, there has been a burst of interest in development in mature and later life as evidenced by the contributions of theorists such as Carl Gustav Jung (1933) and Erik Erikson (1978). Currently, a number of theoretical frameworks on emotional development in adulthood and later life have emerged; these frameworks offer partially convergent and partially divergent interpretations of certain aspects of mature adulthood and later life. In the present contribution, I outline several approaches to describing emotional development and discuss how they bear on processes of individual and interpersonal adaptation. I summarize, as well, a theoretical approach that is able to reconcile some of the interpretive divergences by means of a cognitive-developmental theory that provides a coherent description not just of a single period of life, but of the total span of life from birth to old age.

Theoretical Approaches to Emotional Development

Differential Emotions Theory

Emotions as Core Adaptive Survival Responses

Emotions have had a problematic status in Western intellectual history since Greek philosophy due to their inherent subjective nature. In contrast to cognitive/intellectual functions, they often were maligned as undermining our ability to behave in a well-reasoned and rational fashion (see Labouvie-Vief, 1994). While it is true that rational clarity tends to be clouded in the heat of passion, a more modern stance is to point to enormous adaptive importance of emotions in virtually all of human and animal life. This claim of the importance of emotions was introduced by Darwin, who in the wake of his theory of evolution proposed, in *The Expression of Emotion in Man and Animals* (Darwin, 1965/1882), that emotions had an extremely important function in survival and adaptation. Just like the bodily structures of primary importance in his theory of evolution, he held that emotions, too, had been established by evolution to assure that in emergency situations (for example, threat by another animal) highly automated systems of responding would ensure quick survival responses (e.g., flight or counterattack). How to cope with such highly urgent situations hence did not depend on processes of reflection, which tend to be costly for survival because they can involve too much time and conscious effort. Instead they drew on response systems that were highly automated and rapid. Darwin described different such response systems that were shared by animals and humans alike.

Darwin's theory provided an important base for scientific study of emotions. His proposal that different emotional systems were adapted for different purposes became the basis of differential emotions theory (e.g., Izard, 1971, 1977; Ekman, 1973; for review, see Magai, 2001). Following Darwin's suggestions, this theory proposed that there exist qualitatively different primary or basic emotions, such as anger, joy, sadness, fear, shame, or disgust. Each of these emotions was characterized by a specific pattern of eliciting conditions, along with distinct patterns of expressive (physiognomic and gross motor) as well as physiological response patterns.

Whereas, following Darwin, differential emotions theory has dealt with emotions on an individual level, emotion systems may be more complex and encompass the synchronization of emotional responses among individuals. An example of much recent interest is the emotion system of empathy (DeWaal, 2010). Although this system exists in a uniquely mature form, its developmental precursors are completely automatic, involving such responses as mimicking the mood of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26021-2

another through unconsciously moving our body in synchrony, as a result of which we partake in the emotional state observed.

One example particularly important from a developmental perspective is the attachment system discussed by John Bowlby (1973), a psychoanalyst who offered an ethological theory proposing that the mother–child bond had an evolutionary base because of its important survival function. To secure the survival of the young, there must be in place a repertoire of instinctual behaviors whose aim is to seek and to maintain proximity with parents, and especially to seek it in all possible situations of emergency. Ideally, such systems function in mutuality and reciprocity, but development also can bring problematic deviations from genuine mutuality. Some of such deviations are of importance in interpreting emotional responses along the total of the life span, as will be discussed more fully later.

Stability and Change of Primary Emotions

It is a general assumption of the differential emotions approach that emotional systems are in place early and remain stable throughout the life span. The importance of such stability is that it provides consistency of the adaptive aims that are subserved by the basic emotional systems. In addition, stability is crucial for a sense of self that is experienced as continuous, as opposed to one that is characterized by confusion of self and identity. In this way, emotions also become an important part of the gradual elaboration of the personality.

The core stability of emotional aspects to self and identity is not rigid, however; change is possible, as well, and with increasing development, the basic response patterns become enriched by, and interlinked with, other response systems. For example, with the growth of cognitive understanding and development, individuals become better able to differentiate the experience of self from that of others, and more adept at functioning in reflective ways that permit modulation and regulation/control of experience. Some of these changes will be discussed more fully in the section on cognitive-developmental approaches to emotional development.

Other changes in the nature of primary emotions may be due to biological changes that are apparent across the life span. For example, Sroufe (1996) studied the prototypical emotional responses of laughter and its developmental precursor, smiling, over the course of the first 2 years of life. Smiling is a low intensity response that in newborns appears as a reaction to purely internal stimulation. Over the course of the first four or so months, smiling responses become gradually quicker and more active, and more explicitly related to external events. At first, the elicitors are of low intensity, such as gentle speaking or touching, but eventually they become more dynamic, as becomes the expression of smiling, itself. After 4 months of age, in contrast, one typically observes a much more active laughter, a dynamic and vigorous response characterized by sound bursts and active movements in response to complex and highly dynamic stimuli. On the basis of his work, Sroufe proposes that early development of smiling and laughter involves an increasing capacity to tolerate and enjoy stimulation of high intensity, novelty, and complexity – stimulation that at earlier stages produces fear and distress, a result of too-high levels of tension. In fact, Labouvie-Vief proposes that the resulting lowering of tension thresholds and capacity to not even tolerate, but even embrace high levels of tension is one general characteristic of positive emotional development (see Labouvie-Vief et al., 2014).

Responses to varying levels of intensity have also been studied in adults and older individuals, and a number of studies report that the intensity of affect and the strength of impulses is lower in later life (although other studies report the opposite) (see Blanchard-Fields, 2009 and later discussion). It is not entirely clear why this should be so, but different explanations are apparent. For example, it is evident that older individuals, as a result of their accumulated experience, have become experts at regulating their emotions (Blanchard-Fields, 2009). Another explanation, to which we will return later, points, instead, to evidence that high-level stimulation becomes more difficult to tolerate and sustain for individuals in their later life; hence they have come to key in to stimulation that is less intense and more comfortable. Such changes in preference may root in basic biological changes that make regulatory efforts more difficult, as detailed later.

Hedonics of Emotions: Processes of Emotional Control and Optimization

Differential emotions theory had as its primary objective the precise, scientifically valid description of emotional processes. Another major interest has been in the question of how individuals come to regulate their emotions so as to maintain their levels of tension and arousal within acceptable levels. These approaches often take as important criteria those of well-being, such as an overall positive as opposed to negative affect tone. Here, core questions are whether individuals are able to ward off emotional stimulation that is overactivating and stressful and experienced as negative, and instead strive for, and maintain, a sense of positive well-being.

Optimization and Emotional Selectivity

Whereas aging often is construed in popular culture as a period fraught with difficulties and losses, research has tended to show that older individuals often report high levels of well-being (for review, see Fung et al., 2005). One major explanation of this apparent paradox has been proposed in Carstensen's (see Carstensen et al., 2006) social-emotional selectivity theory. This theory proposes that with increasing age, older individuals come to change their values and goals. The core reason for this change is that, unlike youth, they develop a sharpened sense of the passage and limits of time. What seems like a vast expanse of time for the young, for the aging individual comes to be experienced as time left to live. This restructuring of time, in turn, is related to a restructuring of goals and values.

According to emotional selectivity theory, time for the young is experienced as unlimited. As a result, young adults are primarily invested in goals that emphasize preparation for future success, with knowledge seeking goals being particularly evident. Older adults, in contrast, are said to experience a goal shift toward emotional and interpersonal satisfaction. They aim at optimizing pleasure and satisfaction, and they do so especially in terms of focusing on satisfying and meaningful

interactions while abandoning those that are too stressful. Thus, they appear to actively orient toward optimizing positive feelings and avoiding and decreasing negative ones. In part, they do so by trimming their social networks, restricting them more and more to the most intimate and satisfying relationships.

In support of the notion of optimization of positive emotions, emotional selectivity theory points to a plethora of findings; Kryla-Lighthall and Mather (2008), for example, cite an impressive body of research. First of all, a series of studies has demonstrated that, in general, older individuals report quite high levels of well-being, with decreased negative and increased positive affect as compared to younger adults. Nevertheless, reports are not always consistent; some studies find decrease in negative emotions such as anger over the adult life span, while others report stability of negative affect. Older adults also do appear to experience a shift in their emotional goals. For example, time spent with acquaintances and close friends declines, whereas that spent with close family members increases. In general, elders place high value on social relationships and positive contributions to society, in contrast to younger adults who are more likely to describe life goals such as emphasis on learning and taking new challenges. Elders instead direct considerable attention to optimizing positive affect and to redirect attention away from negative information. As a result, they often display superior emotion regulation strategies, especially ones that are aimed at avoiding and bypassing interpersonal conflict rather than strategies of confrontation that are likely to it.

Attachment Theory as an Explanatory Framework for Selectivity

As indicated, emotional selectivity theory has created an impressive body of research to demonstrate selectivity of emotions. Nevertheless, it is not always evident exactly how and why individuals come to optimize reports of positive emotions, because reporting positivity does not always indicate good regulation skills. Sometimes, indeed, it points to a tendency to downplay difficulty with regulation. Hence there may exist different motivational bases for reporting positivity. Attachment theory, briefly outlined previously, has addressed this ambiguity by defining different styles of attachment, each of which indicates different reasons for why individuals may seek out and/or report positive relations based on the observation of brief periods of separation from the mother (for review, see Mikulincer and Shaver, 2010): these styles have also been extended to and validated for adult populations.

In young children, a *secure style* is evident if, upon reentry of the mother, children do not avoid contact with her by turning away in anger, but approach to her and, if necessary, let themselves be consoled. In contrast, an *avoidant-dismissing* style is evident in children who avoid contact with mother upon reunion and do not reestablish contact with her; these children apparently adopt an air of apparent indifference and calm through remaining distant as a way of adapting to frequent problematic interactions with mother. *Fearfully attached* children, in contrast, are difficult to accept mothers' attempts to comfort them: they vacillate between anger and approach, indicating a relationship that is ambivalent at its core.

As noted, similar attachment styles have been established for adults by means of attachment style questionnaires. Again, research established that the motivations for displays of positivity are very different for the different styles (Mikulincer and Shaver, 2010). *Secure* adults display a positive approach toward and evaluation of self and others. *Dismissing* adults, in contrast, describe themselves in a positive fashion, but project their motives of avoidance on others whom they describe in a negative fashion. *Fearful* adults, finally, tend to describe others as rather positive, but themselves as more negative.

Mikulincer and his colleagues have conducted a series of studies that show that, especially in the case of secure and avoidant-dismissing individuals, different motivations underlie the overt display of positivity. For secure individuals, positivity appears to be a genuine, effortless expression, but for avoidant-dismissing adults, positivity is achieved through the repression of negativity by means of effortful cognitive distortion. This process is evident in experiments that require the judgment of the positivity or negativity of targets (e.g., pictures), and compare conditions in which no special effort is required to ones in which individuals also need to pay attention to other, effort-demanding tasks. Under such conditions, the positivity effect found among dismissing adults is canceled, and instead, avoidant/dismissing individuals evidence a dramatic rise in negative responses. The same reversal is not, however, evident in secure individuals. Such findings are interpreted as the presence of a powerful motivation toward *avoiding negative experiences*, rather than an *effortless orientation toward positivity*.

Interestingly, the reversal indicated has been reported by Mather and Knight (2005; see Kryla-Lighthall and Mather, 2008) who found a positivity effect in older adults' memory for pictures when individuals could direct their full attention to viewing pictures; in contrast, when individuals were distracted by means of a second effort-demanding task, this positivity effect was dramatically reversed, indicating a lowering of positive but a dramatic rise of negative responses. This finding supports the proposed interpretation of a negativity-avoidant motivation as opposed to one that is able to effortlessly seek out positivity.

Avoidance of negativity rather than genuine positivity also has been reported in Labouvie-Vief's (Labouvie-Vief, 2008) research. Among elders who reported positive affect, one subgroup also reported that they had a secure relationship style. This group also displayed high tolerance for ambiguity and affect exploration. In contrast, another subgroup placed little emphasis on openness but scored higher on conformity, denial, and repression. Of all of the elderly, this group also scored particularly low on empathy. Hence it appears that reports of high positive and low negative affect can imply very different core motivations – that of affirming positive affect, as opposed to that of avoiding negative affect. We will return to this finding in the next section on cognitive-developmental interpretations of positivity phenomena, proposing an interpretation in terms of unique vulnerabilities that emerge as part of the process of aging, and that render avoidance of negative experiences and stress especially important.

Cognitive-Developmental Theory

A third theoretical position to the study of emotions and self in aging has been developed by Labouvie-Vief (Labouvie-Vief and Marquez, 2004; Labouvie-Vief et al., 2009, 2010, 2014). This work began with the observation that approaches to the study of aging often have developed out of context of that of developmental processes of earlier life. Gerontology was established only about in the middle of the twentieth century, and emotional development in adulthood and later life began to enjoy popularity only since the later part of that period. The study of emotion began early with the work of Sigmund Freud and continued with that of Piaget (1991). Both theoreticians offered theories that focused not just on stage-specific descriptions of developmental phenomena, but were strongly interested in how mature emotions emerged gradually from precursors that were rather immature, involuntary, and reflex-like. Freud proposed that such primordial emotional responses were characteristic of 'primary process' functioning of the ID, but that with mature development, such processes were gradually encompassed, amplified, and even governed by the more conscious executive mechanisms of the ego. Both theoreticians hence proposed a view of emotional development as a series of transformations in which more unconscious and automatic processes begin to interact with more deliberate and conscious ones. This process of transformation implies a series of processes on neurobiological and cognitive processes that transform emotions from lower to higher levels of complexity at psychological and biological levels.

Cognitive and Neurobiological Mechanisms of Emotional Transformation from Childhood to Adolescence

As already pointed out, emotions and cognitive-rational functions often have been described as existing in a state of tension and polarization. One developmental psychologist who rejected such polarization was Piaget (e.g., 1991) who suggested that although the systems subserve different functions, these two are intimately linked. On that view, emotions have a dynamic and energetic function (as evident, in fact, in the Darwinian notion of affect earlier discussed). In contrast, cognition has a structural function that is oriented toward stability and equilibrium. The great Soviet psychologist Luria demonstrated in the 1930s (Luria, 1932) that the use of language acted to dampen the peripheral arousal and motor excitation that characterizes strong emotions. More recently, research has established that the use of linguistic labels function to downregulate (reduce and control) activation of brain regions involved in emotional arousal (see Labouvie-Vief et al., 2014). Thus, merely labeling emotional faces decreased activity of the amygdala, the core related to emotional arousal; at the same time, the inhibitory/control areas of the prefrontal cortex increased in activity.

Piaget also proposed such intimate interplay when stating "if our … hypotheses are correct, we shall be able to parallel, stage by stage, the intellectual structures and the levels of emotional development. Since no structure exists without dynamic and since, respectively, a new from of energizing regulation must correspond to any new structure, a certain type of cognitive structure must correspond to each new level of emotional behavior" (Piaget, 1954: 10, as quoted in Labouvie-Vief and DeVoe, 1991).

Accordingly, Piaget proposed a series of cognitive-emotional stages from birth to adolescence (see Labouvie-Vief and DeVoe, 1991). At birth, the cognitive-emotional repertoire consists of hereditary organizations such as reflexes, instinctual drives, and affective reactions present at birth. However, these soon become reorganized as a function of learning, as children link emotions with events they have experienced and thus attach to them the satisfaction or disappointments of expectations. In toddlerhood (Piaget's preoperational stage), language and representation emerge, which allow stabilizing and controlling feelings. This emerging capacity for control becomes stabilized as school-age children come to understand and internalize social rules for emotional expression and accordingly regulate their behavior along their understanding of what is socially acceptable. This capacity for internally oriented control becomes even more effective in adolescents who develop abstract thought that permits to construct ideals of how affect should be felt and regulated.

In addition to emotional controls, emotions themselves become more complex in the course of development. One example is the emotion of empathy mentioned earlier. This emotion develops out of simple precursors that reflect automatic responses of mimicking and emotional contagion to the distress of others. It is hence primarily related to the distress induced in the self when watching another. However, to come to be able to effectively offer aid to the other in distress, the development of a set of cognitively complex skills is required: the self needs to be able to differentiate the state of the other from that of the self so as to be able to imagine the other's suffering as a separate state. This, then, represents a complex cognitive operation that, in its mature form, emerges only gradually over the course of the life span, and parallels the development of complex cognitive understanding (see Labouvie-Vief et al., 2010). Only once individuals have come to master these skills, is it possible to display genuine empathic concern for others.

The emergence of such complex skills related to emotions and self-representations has been the focus of a considerable body of neurobiological research (see Labouvie-Vief et al., 2014). At the beginning, capacities for emotional regulation draw upon brainstem-related automatic homeostatic mechanisms that permit an original foundation for the automatic regulation of emotional state. Over time and with continued cortical growth, these areas become richly interconnected with the growth of higher cortical centers implicated in control and inhibition, permitting the adaptive control over states of tension and stress. At first, rather effortful and based on attention-demanding engagement, these capacities themselves become well automated or crystallized with practice, and hence are performed with decreasing effort. This process of automatization/crystallization hence implies that the modulation and tension reduction at more mature stages is accomplished by the automatization of processes that require effort at younger ages. Hence with increasingly complex cognitive skills and more complex neurological interconnections of

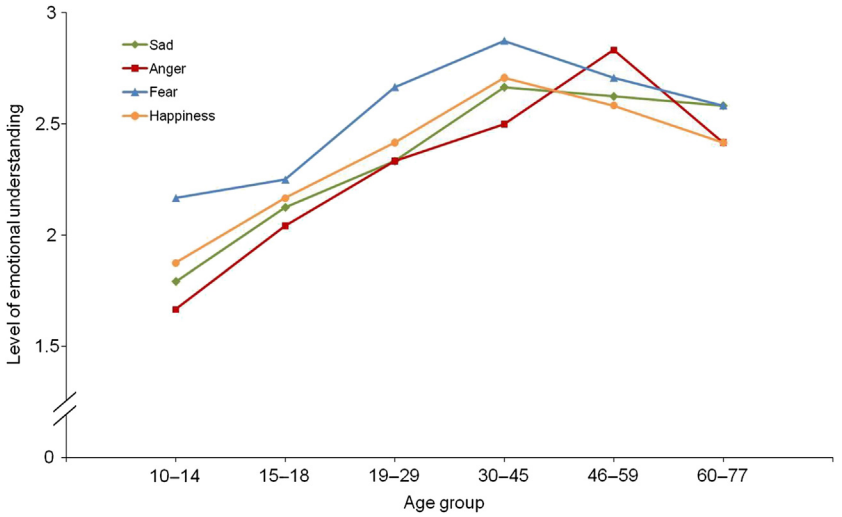

Figure 1 Levels of emotional understanding from preadolescence to late life. Reproduced by permission from Psychology and Aging, 1989, 4 (4), 425–437.

brain functions, individuals can readily master levels of activation and stress that previously were difficult and stressful. This process of lowering of tension thresholds with increasing complexity of cognitive and neurological processes is, in fact, the major result of positive development up until at least middle adulthood. How, then, does this process continue in later adulthood and aging?

Cognitive and Neurobiological Mechanisms of Emotional Transformation in Adulthood and Later Life

Piaget's general scheme of cognition–emotion relations has had a profound influence on childhood researchers of emotional development. From the perspective of adulthood, however, these capacities can have a highly conventional cast that is not ideally suited to continued development in mature and later adulthood. For example, Jung (1933; see Labouvie-Vief, 1994) proposed that the internalization of rules and norms and the constraints of institutions in the first part of life results in an overly mental and outer-directed orientation in which individuals must devalue much of their inner life as 'childish' or 'immature.' In contrast, the goal of mature and later adulthood is said to be to free oneself from these restrictions and idealizations and to explore and liberate deeper, richer, and more complex layers of the personality. A researcher who has specifically used this notion of expansion of the self and emotions was Jane Loevinger who in her theory of ego development (for a summary see Labouvie-Vief, 2002) proposed such movements from conventionality to ways of thinking and being that cherish individuality, self fulfillment, and the psychological depths of one's personality – its causes, apparent contradictions, and paradoxes.

What is the evidence for such continued developmental movements? Labouvie-Vief (see Labouvie-Vief and DeVoe, 1991) applied these notions to the development of conceptions of emotions and the self. Adolescent to adult individuals ranging in age from about 16 to the mid-80s described their experience of emotion in a first study, and of their self in a second. In both, responses were transcribed and coded into four levels, ranging concrete-presystemic to dynamic-intersubjective. At the simplest level, individuals gave global and concrete responses that simple actions, physical characteristics, but do not mention motivational processes (e.g., 'I am nice,' 'I am pretty,' etc.). At the dynamic-intersubjective level, in contrast, individuals vividly described interpersonal dynamics, and were aware of psychological processes and motivations that played a role in the biographic emergence of their personality (e.g., "I struggle with the concept of who I am and have been identified (all giving mother, self-sufficient, religious) and I think, feeling the need to be a more individualized woman with specific needs and desires").

Results showed the hypothesized growth in complexity of descriptions of emotions and (see Figure 1) and self (see Figure 2) from adolescence to late midlife; surprisingly, however, these trends did not continue past the late midlife period but rather showed an apparent reversal, with lower levels reached by older individuals in a study done quite independently. Similar results also emerged when assessing

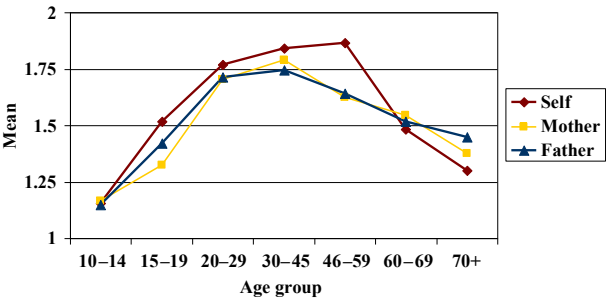

Figure 2 Levels of self-representations from preadolescence to late life. Reproduced by permission from Psychology and Aging, 1995, 10 (3), 404–415.

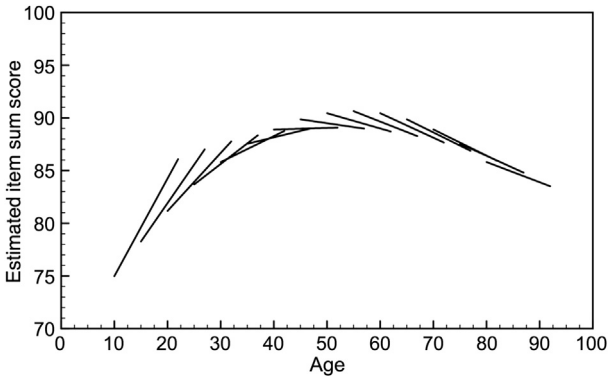

Figure 3 Twelve-year longitudinal gradients of ego development from preadolescence to old age.
Reproduced from Grühn, D., Diehl, M., Lumley, M., Labouvie-Vief, G., July 2008. Paper Presented at the 29th International Congress of Psychology, Berlin, Germany.

levels of ego development, as indicated in Figure 3. Moreover, in contrast to the first two studies, which were cross-sectional, the third reflects longitudinal data. Despite the differences in measurement and methodology, however, the results are very consistent.

These results were encouraging in the sense that, as expected, the process of cognitive-emotional integration became increasingly more complex and well integrated from adolescence to mid-adult life and that this process was evident across several different criteria of complexity. In contrast to middle adulthood, however, evidence did not suggest continuation of a similar process of continued growth and integration from that point onward. These results were disappointing in the sense that, at least by the criteria used, older adults did not adopt a similarly complex and rich approach to talking about self and personality. Since the measure of this work was strongly oriented toward aspects of cognitive complexity, the obvious question is, do such results suggest that cognitive changes often affected with aging are the culprit for such apparent declines?

The answer to this question is both yes and no. First, it is widely established that aging brings normative declines in certain cognitive functions. These functions can constrain further development as certain cognitive-emotional processes become less efficient and more effortful. On the other hand, elders bring to many tasks a rich repertoire of knowledge they have established and crystallized as a result of a lifetime of experiences. In tasks that permits them to apply such knowledge toward the solution, they often maintain superior performance and can evidence high levels of wisdom (see Labouvie-Vief et al., 2010) – a result of having available preexisting well-automated though often quite complex skills (Blanchard-Fields, 2009). However, the changes in the cognitive system are most relevant in tasks that make high demands on the processing system because they are relatively novel and not yet well integrated into an already established repertoire of cognitive-emotional skills. In contrast, novel experiences that require the formation of such complex skills *de novo* pose particular problems to the elderly; this is especially true of tasks that are of unusually high levels of difficulty, if they require that one disregard highly exercised and automated skills in contexts that are not appropriate, or if individuals are required to perform under levels of raised activation or even stress, or ones of unusually high levels of difficulty.

As a result, the picture of aging and emotion experience and regulation offers two rather distinct pathways. One indicates a pattern of vulnerabilities in situations that are novel or relatively high in stress and arousal. The second, in contrast, offers a picture of well-integrated emotion experience and regulation in contexts in which older individuals can base themselves on their rich background of well-crystallized emotion experience and knowledge.

The literature relating to these dual pathways has been extensively summarized by Labouvie-Vief and her collaborators (Labouvie-Vief and Marquez, 2004; Labouvie-Vief et al., 2009, 2010, 2014). A brief summary of the related evidence follows below.

Vulnerability of Later Life Regulation Capacities

As stated, vulnerabilities and restrictions in emotion regulation have been widely documented in situations that exceed individual's accustomed level of complexity, or that require inhibition of well-automated functioning. A number of functions are affected in such situations. *First*, in light with the notion that tension thresholds become lowered in later life, older individuals do not tolerate highly arousing situations well, and consider highly arousing material to be negative. *Second* and relatedly, this lowered tolerance for high tension tends to be disruptive in older adults' cognitive performance, whereas young adults are quite impervious to variations in arousal.

A *third* line of recent evidence suggesting that the aging system is more easily overwhelmed emerges from research using physiological measures of effort, such as systolic blood pressure and cardiovascular stress, and indicating that difficult cognitive challenge places greater physiological stress on elders' system. Such heightened stress is particularly likely in situations of competition or other added stress for the elderly, whereas younger individuals may even be stimulated by such situations.

Fourth, physiological research also indicates that older adults show more pronounced stress-induced hormonal secretion than younger adults during a cognitive challenge. They also take more time to recover from such stress-related reactions than younger adults. Indeed, studies on stress provide further evidence for the greater dysregulation of the system under high levels of arousal and/or effort with advancing age.

Fifth, older individuals are especially negatively affected by conditions of challenge and stress. For example, in a series of studies, older individuals do worse than young individuals on their performance on cognitively challenging task, particularly if these tasks are novel.

Finally in a series of studies examining the relation between attachment styles and emotional reactivity in older individuals, elders with an avoidant dismissing attachment style did fare better than other elders. However, when the situation was particularly activating, such as engaging in a discussion about a conflict, older avoidant, dismissing women had higher initial heart rate and slower recovery rates than their peers. These

results suggest that a dismissing style can serve a partially protective role for some older individuals, but not when levels of activation reach a critical level. Consistent with this proposition, several studies (see Labouvie-Vief et al., 2009; Magai, 2001; Zhang and Labouvie-Vief, 2004) have shown that the avoidant-dismissing, self-protective style of attachment just described is more prevalent in the elderly than in younger adults, suggesting that this style may reflect an attempt to use compensatory protection strategy against negative emotions and their consequences (Labouvie-Vief et al., 2009).

In summary, increasing age and reduced tolerance for high tension seems to impair the capacity to cope with high complexity emotional information. Nevertheless, this loss can be set off by the availability of strong crystallized processes that provide a protective function.

Aging and Crystallization: Strengths of Later Life Regulation Capacities

In contrast to the more problematic aging changes just discussed, a number of studies on aging suggest that automatic functioning seems little affected in later life and may even increase in importance. More specifically, preserved automaticity is likely to aid individuals' performance in situations or tasks that do not require them to inhibit their well-automated cognitive-emotional knowledge, and in tasks that are in the range of complexity for which such crystallized knowledge already is available. Indeed, effects usually associated with aging – such as reduction of cognitive control over emotions – are not always detrimental, but actually may bring positive benefits to the extent that automated knowledge systems may provide a rich base for integrating experience.

Deepened Inner Orientation

In the eyes of some theoreticians, increases in automatic functioning in later life may bring unique benefits to the aging person. Jung (1933) suggested long ago that with a relaxation of controlled functioning, mid- to late adulthood can bring a liberation of unconscious processes and with it, a turn away from preoccupations with the outer world. Instead, the resulting increases in understanding inner layers of experience may achieve their height in later life, through a process of 'centroversion' toward the inner world (see Labouvie-Vief, 1994).

One sign of this inward shift is a general way in which individuals' relationship to information becomes restructured. For the young adult, information is seen as an outer given that one attempts to reproduce in a literal way. In contrast, middle and older adults turn more to the landscape of human motivations and intentions. Hence, they may become experts at the processing of information relating to subjective processes and inner dynamics. Although this symbolic processing style can result in reduced attention to the literal level, it may imply a richer psychological texture. Indeed, a series of studies reviewed elsewhere (for review, see Jepson and Labouvie-Vief, 1992) suggests that older individuals continue to develop highly integrated and personally meaningful insights about life's regularities – the hallmark of advanced wisdom.

Somewhat relatedly, a growing body of research suggests that age differences can diminish or even disappear if experimental tasks permit the elderly to make use of the inner orientation that draws on rich knowledge of self and psychological processes. For example, recent studies on age-related differences on everyday problem solving tasks found that older adults can solve problems as effectively, or even, better than younger adults do. Especially in emotionally salient and interpersonal problems, older adults are more effective than is true of young, evidencing a greater flexibility and adaptability than do younger ones.

Finally, a large body of work suggests that apparent decrements in later life often reflect the lack of relevance for older individuals. In contrast, if tasks are constructed so as to be interesting to the elderly, their performance typically increases. Such findings suggest that age differences can be eliminated when information to-be-remembered is relevant to older adults.

Conclusion

In this article, we address the mix of positive and negative changes in emotion regulation observed with aging. It proposes that this diversity of findings can be explained and systematized by a theory that outlines the coevolution in individual development of cognitive structures and emotion regulation capacities in such a way that individuals develop more effective means of maintaining equilibrium. On this formulation, the growth of cognitive structures and their integration with emotional systems in early development is related to an expansion of the range over which individuals can maintain equilibrium, as well as a raising of the thresholds of tension individuals can manage. The resulting cognitive-emotional schemas or representations themselves can become highly automated or crystallized, providing individuals with available means of regulation that although at first dependent on the expenditure of effort, eventually come to function in a relatively effortless way.

In contrast to development from childhood to middle adulthood, later life appears to bring a cessation of this expansive developmental process as the capacity for high tension and effortful performances decreases as a result of the decline of fluid processes of cognition. As a result, aging is typically associated with a lowering of tension thresholds and a restriction of the range of equilibrium. However, these restrictions may be limiting only in cases where individuals cannot fall back on already well-developed representational systems, instead needing to expend high levels of tension, stress, and effort.

By placing cognitive-emotional processes in the context of equilibrium regulation, the proposed theory provides, we believe, a coherent framework to discuss cognitive-emotional processes from a psychological and biological perspective alike. The notion of lowered equilibrium thresholds with aging already is commonplace in the medical and biological field, such as cardiac stress tests or recommendations for physical exercise regimens. By applying a similar conception to the field of emotion regulation, the current theory provides a linkage of the cognitive functions of emotion regulation to mechanisms of arousal and activation.

See also: Coping across the Lifespan; Eudaemonism; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Blanchard-Fields, F., 2009. Flexible and adaptive socio-emotional problem solving in adult development and aging. Restorative Neurology and Neuroscience 27, 539–550.

Bowlby, J., 1973. Separation: Anxiety and Anger. Basic Books, New York.

Carstensen, L.L., Mikels, J.A., Mather, M., 2006. Aging and the intersection of cognition, motivation, and emotion. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Academic Press, San Diego, CA, pp. 343–362.

DeWaal, F., 2010. The Age of Empathy. Three Rivers Press, New York.

Darwin, C., 1965. The Expression of Emotion in Man and Animals. University of Chicago Press, Chicago (Original work published in 1882).

Ekman, P., 1973. Crosscultural studies of facial expression. In: Ekman, P. (Ed.), Darwin and Facial Expression. Academinc Press, New York, pp. 169–222.

Erikson, E., 1978. Reflections on Dr. Borg's life cycle. In: Erikson, E. (Ed.), Adulthood. W.W. Norton, New York, pp. 1–31.

Fung, H.L., Rice, C., Carstensen, L.L., 2005. Reactive and proactive motivational changes across adulthood. In: Greve, W., Rothermund, K., Wentura, D. (Eds.), The Adaptive Self: Personal Continuity and Intentional Self-Development. Hogrefe & Huber Publishers, New York, pp. 171–183.

Izard, C.E., 1979. The maximally discriminative facial action coding system (MAX). University of Delaware, Newark.

Izard, C.E., 1990. Facial expression and the regulation of emotions. Journal of Personality and Social Psychology 58, 487–489.

Jepson, K.L., Labouvie-Vief, G., 1992. Symbolic processing of youth and elders. In: West, R.L., Sinnott, J.D. (Eds.), Everyday Memory and Aging. Springer, New York, pp. 124–137.

Jung, C.G., 1933. Modern Man in Search of a Soul. Harcourt, Brace, & World, New York.

Kryla-Lighthall, N., Mather, M., 2008. The role of cognitive control in older adults' emotional well-being. In: Bengtson, V., Gans, D., Putney, N., Silverstein, M. (Eds.), Handbook of Theories of Aging, second ed. Springer Publishing, New York, pp. 323–344.

Labouvie-Vief, G., 1994. Psyche and Eros: Mind and Gender in the Life Course. Cambridge University Press, Oxford.

Labouvie-Vief, G., DeVoe, M.R., 1991. Emotional regulation in adulthood and later life: a developmental view. In: Schaie, K.W., Lawton, M.P. (Eds.), Annual Review of Gerontology and Geriatrics: Behavioral Science and Aging, vol. 11. Springer, New York, pp. 172–193.

Labouvie-Vief, G., 2002. Ego development in adulthood. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. In: Eisenberg, N. (Ed.), Development, Social, Personality, and Motivational Psychology, vol. 22. Elsevier, Oxford, pp. 4361–4365.

Labouvie-Vief, G., 2008. Dynamic integration theory: emotion, cognition, and equilibrium in later life. In: Bengtson, V., Silverstein, M., Putney, N., Gans, D. (Eds.), Handbook of Theories of Aging. Springer, New York, pp. 277–293.

Labouvie-Vief, G., Marquez, M.G., 2004. Dynamic integration: affect optimization and differentiation in development. In: Dai, D., Sternberg, R. (Eds.), Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development. Lawrence Erlbaaum, Mahwah, NJ, pp. 237–272.

Labouvie-Vief, G., Grühn, D., Mouras, H., 2009. Dynamic emotion-cognition interactions in adult development: arousal, stress, and the processing of affect. In: Bosworth, H.B., Hertzog, C. (Eds.), Aging and Cognition: Research Methodologies and Empirical Advances. American Psychological Association, Washington, DC, pp. 181–196.

Labouvie-Vief, G., Grühn, D., Studer, J., 2010. Dynamic integration of emotion and cognition: equilibrium regulation in development and aging. In: Lerner, R.M., Lamb, M.E., Freund, A.M. (Eds.), The Handbook of Life-Span Development, Social and Emotional Development, vol. 2. Wiley, Hoboken, NJ, pp. 79–115.

Labouvie-Vief, G., Gilet, A.L., Mella, N., 2014. The dynamics of cognitive-emotional integration. Complexity and hedonics in emotional development. In: Verhaegen, P., Hertzog, C. (Eds.), The Oxford Handbook of Emotion, Social Cognition, and Problem Solving in Adulthood. Oxford University Press, Oxford, pp. 83–98.

Luria, A., 1932. The Nature of Human Conflicts. Liveright, New York.

Magai, C., 2001. Emotions over the life span. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, fifth ed. Academic Press, San Diego, CA, pp. 399–426.

Mather, M., Knight, M., 2005. Goal-directed memory: the role of cognitive control in older adults' emotional memory. Psychology & Aging 20, 554–570.

Mikulincer, M., Shaver, P.R., 2010. Attachment in Adulthood: Structure, Dynamics, and Change. Guilford, New York.

Piaget, J., 1991. Intelligence and Affectivity: Their Relationship during Child Development. Annual Reviews, Palo Alto, CA.

Sroufe, L.A., 1996. Emotional Development: The Organization of Emotional Life in the Early Years. Cambridge University Press, New York, US.

Zhang, F., Labouvie-Vief, G., 2004. Stability and fluctuation in adult attachment style over a 6-year period. Attachment & Human Development 6, 419–437.

Self and Identity Development during Adolescence across Cultures

Juan José Zacarés, Faculty of Psychology, University of Valencia, Valencia, Spain
Alejandro Iborra, Faculty of Education, University of Alcalá, Alcalá de Henares, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose is to show the dynamic of change and stability in the development of identity and self during adolescence. We present some basic differences between the development of self-perceptions and the development of identity according to Erikson's psychosocial formulation. The change and stability dynamics of identity development is subjected to critique from the perspective of the status paradigm as well as more recent theoretical models. The key role of the cultural context's inevitable influence on the development of identity and self-perceptions through the production of relevant cultural variations is also analyzed.

Introduction

Adolescence is usually defined as a 'developmental transition,' because in the broader context of the life cycle it represents, a period of change, growth, and transformation from childhood immaturity to the physical, psychosocial, and sexual maturity of adulthood. All the challenges in the form of 'developmental tasks' that the teenager must face are synthesized in the establishment of a sense of identity or self at the end of this period. Today, the period of time needed for this process must be considered as lasting up to the years of so-called 'emerging adulthood,' at least for a substantial and ever larger number of youngsters in many societies around the world. While major changes in self-perception also occur during childhood, such a degree of self-awareness and capacity for self-reflection is reached for the first time in adolescence.

It is important to grasp the intrinsically psychosocial and transactional nature of identity development, as revealed by Erikson (1968). Identity is located at the major crossroads of individual and social processes: it bestows, on the one hand, individuality – the person experiences himself or herself as a unique person; and on the other, a sense of belonging linked to the social context and recognized by the subject. Adolescent identity is the result of mutual recognition between the adolescent and society: the teenager forges an identity, but at the same time society identifies the teenager. The formation of identity thus involves a continuous interaction between the person and the context. Multiple levels of analysis and different theoretical approaches to the study of the self and identity in adolescence have led to diverse modifications to this basic premise, which means that both concepts stand in need of some clarification.

Definitional Issues

The development of self and identity is first and foremost a complex, multifaceted process that combines conscious and unconscious elements. It can be considered as a series of interrelated developments covering all the changes in the way, adolescents see themselves in relation to the others and to their sociocultural context. As these changes have been addressed by both psychological and sociological research, there has been great diversity in the definitions given to the basic concepts at stake. The most persistent distinction is between 'identity' and the related term of 'self.' Although some consider it artificial, this distinction is generally assumed to be valid and to need careful clarification. Beyond this continuing lack of agreement, it is, however, possible to recognize some points of convergence in the current theoretical debate that help to frame the development of identity in adolescence.

Firstly, there is an emphasis on the careful description of the phenomena studied, whether they be specific aspects of adolescents' own beliefs and evaluations of themselves, of their particular configurations of roles, identities, and values, or of the selection and processing of self-relevant information. Likewise, it would be clearer to refer to each term as a separate construct by means of the prefix 'self' until we know how they are related. In order to perform this task of clarification, it would be helpful to differentiate between inclusive levels of self-organization combining the temporal dimension (employing different time scales) with other parameters such as the degree of internalization or reflexivity. This way, self-evaluations and self-perceptions, for instance, would be located on an intermediate time level and as part of a first-order reflexive process, while the concept of identity would be located on a longer timescale typical of long-term developmental processes and as part of a second-order reflexivity (Lichtwarck-Aschoff et al., 2008). Self-representations would thus be considered as components of the identity.

Secondly, different theoretical approaches recognize several levels at which identity and self can be defined and understood: the personal or individual level, the relational level, and the social or collective level. Personal identity refers to those aspects of self-definition at the individual level including goals, values, beliefs, and a whole set of associated self-representations and self-evaluations. Relational identity refers to the contents of identity present in 'face-to-face' interactions and that have been framed in a variety of social roles and positions (son, student, friend, etc.). It also covers the way adolescents define and interpret these roles. Collective identity refers to all the cognitive and affective aspects deriving from belonging to certain groups with which adolescents identify themselves and which place them within certain social

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23028-6

categories such as ethnicity, nationality, or gender. Alongside these three generic levels, another two levels of undeniable psychosocial significance have been identified. On the one hand, at the most basic psychological level, there is the ego identity, which remains largely unconscious and whose essential characteristic is its provision of a more subjective sense of time–space continuity than that associated with the personality. Erikson gave great importance to the function of an ego structure in creating this experience of personal continuity. On this account, the ego identity would be responsible for testing, selecting, and integrating self-images, and its chief function would be the unconscious synthesis of impressions, emotions, memories, impulses, and actions. On the other hand, cultural identity is differentiated as an aspect of social identity, which refers specifically to those cultural groups to which the person belongs and, in the case of immigrants, to their heritage and receiving communities. Cultural identity may be seen as midway between personal identity and collective identity. It includes domains such as adopted cultural practices, emotional identifications with the original cultural group and the receiving society, and value orientations.

Finally, there is currently a clear awareness of the need for theoretical and empirical integration in the study of both contents and identity processes at their several levels. As Vignoles et al. put it (2011: p. 9), "irrespective of levels of content, any given aspect of identity can be viewed as defined by individual, relational, and collective processes: as the subjective understanding or experience of individuals, as an interpersonal construction, and as a sociocultural product." Details of these tendencies are set out in the next two sections, which concentrate on the two most important approaches to the study of self and identity in adolescence.

Development of Self-Perceptions and Self-Esteem in Adolescence

Self-psychology has traditionally focused on the Jamesian distinction between the self 'as active knower' (I-self) and the self as 'known' (Me-self), in a way that is comparable to Erikson's distinction between the ego and the self. Most of the literature on self-development in adolescence has focused on the second level of analysis, i.e., the Me-self as the object of one's knowledge. In this regard, most studies have centered on the perceptions and evaluations that the teenagers have about their 'self' and its association with indicators of psychosocial adjustment and mental health.

Moreover, there is growing recognition of the difficulty of separating the evaluative dimension (self-esteem) from the purely descriptive dimension (self-concept) in self-representations because most of them are linked to a particular assessment, which is inevitably more or less positive. The current trend is to differentiate between domain-specific and global components of self-perception. On this view, the self-concept is considered to be the domain-specific assessment of capabilities (in a descriptive and an evaluative sense), while self-esteem represents the general sense of self-worth as a person. Most models further assume that self-perceptions are multidimensional, hierarchical, and progressively differentiated with age. Due to its hierarchical nature, self-esteem as a general statement of self-worth is located at a higher level, with specific domains (cognitive, physical, or social) and subdomains located in the subordinate levels of the structure of self-perceptions. Self-esteem is the most important factor in psychological well-being, but it is influenced by the self-assessments made in the subordinate domains (e.g., academic performance or physical appearance). Finally, most theoretical models agree that the self-concept is formed during childhood, when domain-specific differentiations are already under way, and undergoes faster differentiation and achieves potential consolidation during adolescence. This last point deserves more attention because of the intensity and variability of these changes.

Cognitive maturation in adolescence does not mean that the teenager quickly develops a consistent theory of the self. What happens instead is that the self-concepts undergo a series of progressive changes in their content, structure, and organization throughout this period. Therefore, the achievement of this more consistent and internally stable scheme of the self only starts to take place in early adulthood. As for the contents that teenagers often include in their self-representations, there are changes related to the vital areas, which are most important in each developmental stage. In early adolescence (12–14 years of age), attributes, interpersonal feelings, and social skills that influence interactions with others or affect the social image become the most prominent. By middle adolescence (15–16 years of age), self-descriptions contain a greater variety of psychological traits and a number of attributes reflecting increasingly personalized beliefs, values, and moral ideas. In late adolescence (17–19 years of age), these materials are typically added to concerns about 'future selves' and to the plans put in place to achieve them.

As for structure, the direction of these changes points in two directions: self-concepts become more differentiated at the same time as there is a growing tendency toward integrating these multiple self-representations into larger and more consistent units. In a developmental sense, differentiation is predominant until middle adolescence, with integration taking over in the late adolescence. The path to adulthood requires the integration of these multiple self-perceptions into a coherent, unified theory of the self. This path starts with the preparation by the teenager of some early abstractions that integrate concrete descriptions of the self (Harter, 2012). It is only in late adolescence when the capacity to build higher-order abstractions permits the integration of isolated and contradictory abstractions into a more consistent self-concept (e.g., 'depressive' and 'cheerful' do not enter into conflict with each other because they are part of being someone 'with mood changes').

Research has addressed the issue of the pattern of change in self-esteem across adolescence and whether these changes confirm the view of adolescent development as an intense period of 'storm and stress.' In general, self-esteem during adolescence tends to remain stable with age. However, there is a transitional period during early adolescence where the positive feelings about oneself decrease significantly between 12 and 13 years of age, in both boys and girls self-esteem reaches its lowest levels due to the psychosocial impact of major pubertal change. From here there is a progressive recovery until late adolescence, in a U-shaped fashion.

Some studies have shown that if stability is analyzed only in terms of average differences, qualitatively different developmental paths in self-esteem could be concealed. More flexible person-centered methodologies are required to address this important issue. In a recent study, Morin et al. (2013) used growth mixture analysis to identify four trajectories in adolescence based on global self-esteem levels and fluctuations: (1) *moderate* (56.2% of the sample, with moderate levels of self-esteem that remained stable over the course of the study), (2) *low* (19.3%, low and unstable levels of self-esteem), (3) *increasing* (11%, increasing levels of self-esteem and stability), and (4) *elevated* (13.5%, elevated levels of self-esteem and stability). As previous research has shown, each path in self-esteem differed in its pattern of correlations with other variables, in this case, those relating to individual perceptions of various aspects of school life. Their research also illustrates the existence of a group (around 20% of the sample) manifesting the most negative developmental pattern associated with the 'storm and stress' view of adolescence. The association between the levels and stability of self-esteem is another outstanding result, consistent with the self-equilibrium hypothesis. These researchers state in their conclusions, "tend to co-occur so that students with high global self-esteem show a greater level of longitudinal equilibrium in their levels of global self-esteem whereas students with low global self-esteem present an apparently highly reactive, unstable, longitudinal profile" (Morin et al., 2013: p. 19). Adolescents at greater psychosocial risk (low trajectory) are those whose self-esteem is more reactive and sensitive to changes in their context and who may thus benefit particularly from interventions designed to improve their self-perception.

Much attention has been paid to self-perceptions as a factor associated with psychological adjustment throughout the life cycle. For example, having a high self-concept in different domains and a sense of self-worth has been considered to protect against poor mental health (depression and anxiety) as well as risk activities, and antisocial or delinquent behaviors. High self-perceptions have also been shown to be connected with higher levels of academic performance and adaptation. However, somewhat more controversy attaches to giving any precise definition to the character of this widely observed association. Clearly, self-perceptions are related temporarily in a reciprocal fashion to indicators of adjustment and psychological well-being, thereby functioning as both a predictor and an outcome of healthy development. In contrast, the attribution of a true causal role to self-esteem has come in for criticism. Many of the associations detected between the level of self-esteem and some adaptation indicators may be the effects of common causal factors impacting both self-esteem and these indicators. This might have overinflated the impact of self-perceptions on adolescent functioning. So what role should be given to them? Based on current evidence, it is, on the one hand, reasonable to consider global self-esteem as a good indicator of overall psychological well-being in adolescence, or alternatively, as a risk marker for future development, and self-perception in specific areas as reflecting a component of self-efficacy in each domain. On the other hand, while the average level of self-esteem has no explanatory power per se, it has been noted that the variability of self-esteem is more likely to be the factor that impinges most on to adaptation. Several studies conclude that adolescents who show more severe fluctuations in self-esteem (compared to those with greater stability in high or low self-esteem) are more likely to behave aggressively and to suffer mental health problems. The relative independence of these two aspects has not been proven one way or the other, so it remains an open question for future research.

Development of Personal Identity in Adolescence

Erik Erikson initiated a tradition in the study of the concept of personal identity. His interest in this topic was clearly stated in his early work on World War II veterans, men who "did not know any more who they were" and therefore provided evidence of a "distinct loss of ego identity." As Erikson (1968: p. 67) observed clinically, "the sense of sameness and continuity and the belief in one's social role were gone." This was the first time in his career that he formulated the assumption of the importance of a sense of identity for a healthy person. For Erikson, adolescence was only completed when the individual could subordinate his childhood identifications to the choices and decisions he had to transform into meaningful commitments, which also needed confirmation by a significant social community. Although it was during adolescence that the task of creating a sense of identity attained its greatest ascendancy, according to Erikson every crisis showed elements of this identity formation.

The work of James Marcia marked the start of numerous critiques of the identity status paradigm, all with their origin in attempts to falsify Erikson's propositions in Ego Psychoanalytic Theory (Kroger and Marcia, 2011). Marcia's primary task was to determine what observable referents were available that might point to the presence of an identity structure. As no one can observe an identity, these observable referents were important if any empirical research was to be possible.

Marcia chose two dimensions, exploration (originally called crisis) and a commitment to empirically document the identity process. Exploration referred to some period of rethinking and trying out various roles and life plans. Commitment referred to making a choice to adhere to one's values, goals or beliefs in a given life domain and the ability to make future projections with regard to this choice. The four identity statuses based on the criteria of exploration and commitment were as follows: Identity achievement – when a person has undergone exploration and is currently committed; Moratorium – when a person is in an exploratory period and therefore bereft of commitments; Foreclosure – when the person has not explored but is committed to one or more choices; and Identity diffusion – when the person is characterized by a lack of both exploration and commitment.

Early research with identity status focused primarily on individual personality and cognitive and relational differences across the four different status positions. Although the initial method for obtaining evidence of these statuses was the Identity Status Interview, whose focus was deliberately qualitative and developmental, the most subsequent research has used questionnaires. By means of questionnaires such as the Objective Measure of Ego Identity Status, the four statuses have been correlated extensively with different personality variables

such as self-esteem, anxiety, moral reasoning, and ego development (Kroger and Marcia, 2011) with a view to finding differences between the statuses that would imply changes in the underlying identity structure.

Achievement has been associated with balanced thinking, mature interpersonal relationships, and thoughtful consideration of potential life options. Individuals with foreclosure status are inflexible and defensive. They have been associated with high self-worth but also with rigidity, closed-mindedness, and authoritarianism. Moratorium is the identity status with the lowest level of well-being and characterizes young people in an identity crisis. In many ways, they resemble achievement individuals in their cognitive complexity and higher levels of moral reasoning, yet they also demonstrate greater openness to experience. But this status has a dual trait so that an individual in moratorium may be an explorer but also suffer paralysis. Diffusion is less homogeneous as a group, but individuals with this status share an inability to make definite commitments and decline to explore options. Some diffusion individuals may go through life in a carefree, uninvolved way, while others manifest a severe psychopathology and great loneliness.

It has been recognized that the main drawback of this quantitative approach based on the use of questionnaires was the creation of a 'static' sociological conception of identity, which was more interested in classifying individuals into categories (the statuses) than in examining the complex psychodynamic that Erikson had initially described. In view of this major drawback, at the beginning of the last decade, instead of answering calls to abandon identity status theory, some models were developed in order to extend or expand its original formulation. What emerged from these extensions of the theory were such original contributions as the works of Luyckx et al. (2011) and of Crocetti et al. (2008).

Luyckx and colleagues model subdivided exploration and commitment into two different processes. Exploration included exploration in breadth and exploration in depth. In the same way, commitment differentiated between commitment making and identification with commitment. Both exploration in breadth and commitment making are equivalent to Marcia's original dimensions. Exploration in breadth refers to the gathering of information about several identity alternatives (values, roles, careers, relationships, ideologies, etc.), while commitment making refers to establishing strong choices in different identity domains as a consequence of prior exploration in breadth. What is new is the evaluation of those commitments, which were enacted previously. This commitment evaluation phase begins with exploration in depth, that is to say, a closer examination of those commitments after choosing them. Identification with commitment follows after a period of exploration in depth in order to integrate one's commitments into one's sense of self; in other words, identification with commitment is a measure of the degree of fit or agreement between those commitments and one's own wishes and personal standards. The interaction of the four processes is far from being simply linear. Not only may exploration in depth influence identification with commitment, but also a lack of identification with commitment may influence the need for a prolonged exploration in depth of those commitments or a renewed exploration in breadth of alternatives (Luyckx et al., 2011).

Crocetti et al. (2008) provide a second new extended model of identity. In their studies, identity is operationalized as a three-factor model with interplay among commitment, in-depth exploration (the extent to which individuals reflect on their commitments), and reconsideration of commitment (comparison of present commitments with alternative ones when the previous ones are no longer satisfactory enough). Individuals begin the process in early adolescence with a set of commitments in at least the ideological and interpersonal domains. These commitments are explored through reconsideration of commitments and through in-depth exploration. Reconsideration compares present commitments to alternative ones and includes decisions about whether to change them or not. In-depth exploration is a continuous monitoring of present commitments with the function of making them more consciously and maintaining them. In this model, commitment is not differentiated.

These newer distinctions of exploration and commitment offer interesting insights into the process of exploration. Luyckx et al. (2011) introduced the concept of 'ruminative exploration' to explain the paradoxical yet common association of exploration with both distress and openness to new experiences. Ruminative exploration refers to an obsessive concern about making the 'correct' choice, with the result that the person remains 'stuck' in the exploration process. This recurrent finding was in fact due to the merging of multiple exploration processes, which had not been distinguished in the original identity status theory. According to said authors, exploration in breadth and exploration in depth are associated with the positive side of exploring new alternatives, ruminative exploration with the negative side of distress.

The introduction of new distinctions in the processes of exploration and commitment has understandably provided some additional variants in the original statuses (see **Table 1**). From the combination of the three factors proposed by Crocetti et al. (2008), five identity statuses emerge: achievement, early closure, moratorium, searching moratorium, and diffusion. From the combination of exploration in breadth, exploration in depth, commitment making, and identification with commitment, cluster analysis identified the following statuses (Luyckx et al., 2011): an achievement cluster, another of foreclosure, two variants of moratorium (adaptive and arrested), and two of diffusion (carefree and diffused).

Another recent contribution of these two extended models has been to answer the age-old question of whether any developmental sequence underlies the identity statuses. Recent meta-analysis studies have tried to find out whether identity develops progressively and whether identity statuses can be ordered on a developmental continuum. Kroger and Marcia (2011) discussed a review of 124 longitudinal or cross-sectional studies performed before 2005, with adolescents (13–19 years of age) and young adults (20–36 years of age). Their conclusion was that the data generally supports a developmental process of identity formation. This means that there are more progressive than regressive developmental movements over time (D → F, D → M, D → A, F → M, F → A, M → A). Meeus (2011) concluded with respect to progression and stability that studies generally show a developmental progress in personal identity during adolescence. But, in addition to progress, many cases of stability have also been

Table 1 Identity status classification as per Crocetti–Meeus (shaded) and Luyckx models

	Processes					
Statuses	*Exploration in breadth*	*Commitment making*	*Reconsideration of commitment*	*Exploration in depth*	*Identification with commitment*	*Ruminative process*
Achievement		High	Low	High		
Achievement	High	High		High	High	Low
Early closure		Moderate	Low	Low		
Foreclosure	Low	High		Low	Moderate	Low
Searching moratorium		High	High	High		
Moratorium		Low	High	Low		
Moratorium (adaptive)	High	Low		High	Low	Low
Moratorium (arrested)	High	Low		High	Low	High
Carefree diffusion	Low	Low		Low	Low	Low
Diffused diffusion	Low	Low		Low	Low	High
Diffusion		Low	Low	Low		

reported since many adolescents and adults do not change their identity status, such stability being more common in adulthood than in adolescence. Stabilities were more frequent in the high commitment statuses than in the diffusion and moratorium statuses. It can therefore be concluded that adolescents may follow two distinct paths on the identity status continuum: D → F → A or D → M → C → A.

Recently, Meeus et al. (2012) have confirmed that it is better to conceptualize achievement, moratorium, early closure, and diffusion as trajectories rather than statuses. On this account, every status would be an example of an "over time solution of the identity puzzle" (p. 1017). By combining the three dimensions of commitment, exploration in depth and reconsideration of commitment they found the following trajectories: 15.8% of the respondents were classified within a trajectory of achievement, 39.6% within an early closure trajectory. A smaller number of respondents, mostly early-to-middle adolescents, were classified in the trajectory of searching moratorium. 20.5% of the respondents were in the moratorium trajectory. In contrast to searching moratorium, the moratorium trajectory describes participants without strong commitments, which have not been processed actively. In addition, they manifest a high level of depressive symptoms. Finally, 20.7% of respondents were in the diffusion trajectory. These trajectories are evidence of the richness and usefulness of Marcia's original work on status. But in this case, statuses are understood as long and stable trajectories instead of a path of statuses that must be crossed.

Cultural Context and the Development of Self and Identity

Although Erikson conceptualized identity as the interplay between a person and his or her social or cultural context, the focus of psychological research has traditionally been on the individual, leaving aside the influence of the cultural context. Ethnicity, as an example of the cultural influence, is one of the barriers, which can limit or facilitate the construction of a sense of identity. Although initially introduced as just another domain of identity status research, issues arising from ethnic identity led to a questioning of the assumption that the identity status model was valid for adolescents and emerging adults who were not white, Westerners, or middle-class. Phinney defined ethnic identity as a self-constructed internalization of the meaning and implications of a person's group membership based on that person's attitudes and feelings toward his or her cultural background, ethnic heritage, and racial phenotype. Compared with identity statuses, ethnic identity involves choice limitations in the sense that one cannot choose one's heritage. As Phinney and Baldelomar (2011: p. 163) put it, culture "permeates the process of identity formation not just at a macro level but at every level of analysis. Culture is not at the periphery of identity formation but at its center." From this all-encompassing idea of culture, they regard identity to be relational to the extent that one's identity serves to distinguish oneself from others and to link oneself with others. Identity is therefore situated within a cultural context, which includes the range of options available to individuals.

Phinney and Baldelomar go on to propose the cross-cultural identity status model (see **Figure 1**). This model extends the original status model by including the sort of cultural orientation, which frames the identity work. For them, Marcia's status model is valid only for an individualistic culture promoting choice, individualization, and self-expression. Consequently, an interdependent or collective culture should also be included, which promotes tradition, conformity, and replication. This culture would favor narrow exploration in order to facilitate the achievement of an available identity from authority figures. In this line, on the basis of the work of Greenfield et al. (2003), they propose two different cultural pathways of development, which frame the process of identity formation and its outcomes.

An interdependent pathway of development, which whittles down exploration could lead normatively to an interdependent identity achievement, adapted to this cultural frame; an independent pathway of development could lead to an independent identity achievement characterized by a wide-ranging process of exploration. That said, someone who explored widely in an interdependent culture (to give just one example of a countercultural process) would also end up with an achievement identity. Someone who did not explore in an independent culture (another example of a countercultural

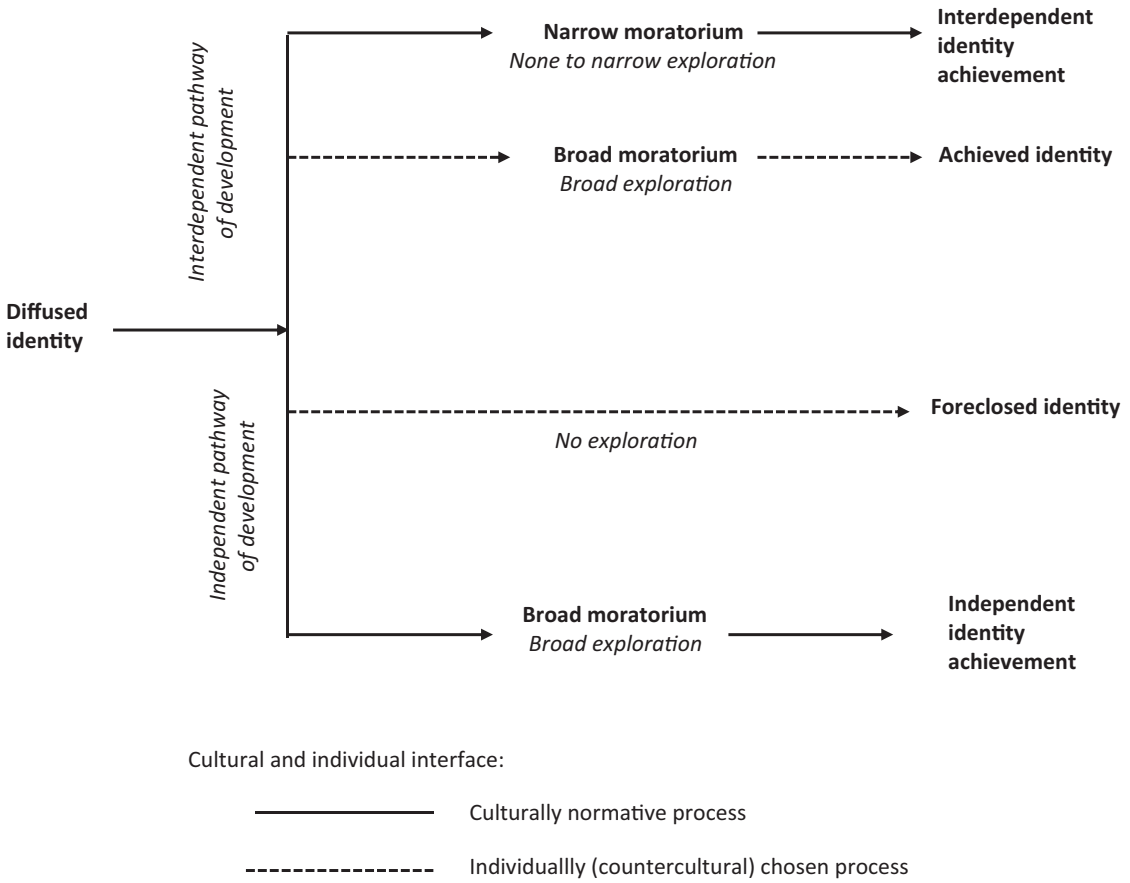

Figure 1 Sequence of identity statuses by cultural pathways of development. Phinney, J.S., Baldelomar, O.A., 2011. Identity development in multiple cultural contexts. In: Jensen, L.A. (Ed.), Bridging Cultural and Developmental Psychology: New Syntheses in Theory, Research and Policy. Oxford University Press, New York, p. 168.

process) would end up with a foreclosed identity status. It should be noted that foreclosure status would not be included in an interdependent pathway of development, in so far as adopting and internalizing the values and models proposed by the community is the ideal trajectory for successful adaptation.

Recent research conducted in different countries around the world provides interesting examples of these two cultural pathways (Schwartz, 2012). As far as independent cultural orientation is concerned, in Europe, there have been studies in Belgium, the Netherlands, Germany, and Italy. All emphasize the importance of individualized exploration and self-expression processes. Nevertheless, they present different findings with respect to the resources provided by political and economic institutions to help in supporting the younger generation in their task of exploration. The cases of Japan and Sweden show a combination of independent and interdependent cultural orientations. Research based on identity status in Japan has been revised from the optic of a culture, which stresses interdependence and its values of conformity, tradition, and replication. However, from 1990 individualistic values such as the importance of being creative and getting achievements as promoted in education and the mass media have given rise to a new configuration termed 'individualistic collectivism,' which incorporates a strain of individualism and autonomy into Japanese identity formation. For its part, and like Germany and the Netherlands, Sweden stands out for of its social structural support policies, which has generated an interesting model with a marked emphasis on tolerance, equality of outcomes, and social welfare, despite a traditionally individualistic perspective. Thanks to this, Swedish youngsters can postpone their transition to adulthood and, with backup from the social system, are able to explore openly more options.

The formal consideration of one's culture has provided interesting insights into our understanding of how our identity is formed in interplay with its immediate context. Far from considering identity as only the product of individual processes, current research demonstrates the importance of relational, social, economic, political, and historical factors. The inclusion of research from around the world fills out the conventional picture drawn by studies in North America and Europe. This is even more important when we take into account how issues of cultural diversity are becoming more common and crucial all around the world. In a globalizing world, knowing just one culture, and its implied independent or interdependent pathways of identity formation, is going to be almost impossible for today's adolescents and emerging adults. As a result of globalization and its processes of mass

immigration, all individuals, members of majority or minority groups alike, must deal with themselves as cultural beings. In today's world, developing a cultural identity entails the integration of multiple cultures.

Conclusions

The convergence between theoretical and empirical approaches to the development of self and identity in adolescence is now the trend that is shaping the future in this area of study. Research into adolescence has benefited from the awareness of the complexity of such questions as what identity means and how identity processes operate. Two broad lines of work reflect this integration. One line seeks to connect together developmentally various dimensions of personal identity and self-concept as components of a dynamic self-system. The study of the implications of these connections for the psychosocial functioning and the psychological health of the adolescents has also intensified. Longitudinal studies combining a focus on longer time periods and short-term contextual fluctuations are helping to clarify the dynamics between change and stability in the various elements and levels of identity between early adolescence and emerging adulthood. In adolescence, active wide-ranging exploration and the reduction of uncertainty through the maintenance of tentative commitments contribute to a clear self-concept, i.e., to the configuration of a consistent, stable self-concept, and high self-esteem. Self-concept clarity in turn decreases the likelihood of reconsidering commitments and amounts to protection against the internalizing problems of anxiety and depression. However, in emerging adulthood, more importance is attached to an identification with commitments when psychosocial well-being and the increase of self-esteem are at stake. Therefore, normative trends in the consolidation of self-concept in early and middle adolescence provide the foundations for social and personal identities that are ultimately expressed as the resolution of the task of identity. Thus, a form of personal agency might be postulated, which was involved in identity formation during late adolescence and emerging adulthood but did not occur in the earliest stages of the development of the self-concept and of the adoption of the first customized commitments.

On the other hand, there is a growing interest in clarifying the interrelationships between personal identity and cultural identity in the framework of the effects of globalization on adolescent development. Personal identity seems to mediate the relationship between cultural identity and psychosocial adjustment; therefore, it is essential to analyze the ways in which cultural processes affect the setting of goals, values, and beliefs characteristic of personal identity. The application of the distinction of cultural orientations (independent or interdependent) to both the configuration of a self-schema and to the way identity commitments are formed and reviewed will remain fruitful. It is clear that in some Western cultural contexts, with broad possibilities of choice in most domains, the adoption of commitments after an exploration period seems to be the most adaptive means for developing identity. However, the ways in which identity develops in culturally interdependent environments where exploration, time–space continuity of the self and personal agency are secondary are not well understood. The notions of 'wide' and 'narrow' exploration applied to different domains of self-concept and identity contained in the trans-cultural status model are suggestive. However, they are still insufficient, especially when it comes to accounting for the development of personal identity in situations of rapid cultural change as a result of migration or globalization, and all that implies for the negotiation of the multiple cultural identity options available.

See also: Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Erikson's Theory of Psychosocial Development; Ethnic Identity, Psychology of; Identity and Identification, Social Psychology of; Logical and Hypothetical Reasoning in Adolescence, Development of; Personality Changes During Adolescence Across Cultures; Self-Development in Childhood and Adolescence; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Social Identity in Social Psychology; Transition to Adulthood, Cultural Differences in; Wisdom, Psychology of.

Bibliography

Crocetti, E., Rubini, M., Meeus, W., 2008. Capturing the dynamics of identity formation in various ethnic groups: development and validation of a three-dimensional model. Journal of Adolescence 31, 207–222.

Erikson, E.H., 1968. Identity, Youth and Crisis. Norton, New York.

Greenfield, P.M., Keller, H., Fuligni, A., Maynard, A., 2003. Cultural pathways through universal development. Annual Review of Psychology 54, 461–490.

Harter, S., 2012. Emerging self-processes during childhood and adolescence. In: Leary, M.R., Tangney, J.P. (Eds.), Handbook of Self and Identity, second ed. Guilford Press, New York, pp. 680–716.

Kroger, J., Marcia, J.E., 2011. The identity statuses: origins, meanings, and interpretations. In: Schwartz, S.J., Luyckx, K., Vignoles, V.L. (Eds.), Handbook of Identity Theory and Research. Springer, New York, pp. 31–53.

Lichtwarck-Aschoff, A., van Geert, P.L.C., Bosma, H.A., Kunnen, E.S., 2008. Time and identity: a framework for research and theory formation. Developmental Review 28, 370–400.

Luyckx, K., Schwartz, S.J., Goossens, L., Beyers, W., Missotten, L., 2011. Processes of personal identity formation and evaluation. In: Schwartz, S.J., Luyckx, K., Vignoles, V.L. (Eds.), Handbook of Identity Theory and Research. Springer, New York, pp. 77–98.

Meeus, W., 2011. The study of adolescent identity formation 2000–2010: a review of longitudinal research. Journal of Research on Adolescence 21, 75–94.

Meeus, W., van de Schoot, R., Keijsers, L., Branje, S., 2012. Identity statuses as developmental trajectories: a five-wave longitudinal study in early-to-middle and middle-to-late adolescents. Journal of Youth and Adolescence 41, 1008–1021.

Morin, A.J.S., Maïano, C., Marsh, H.W., Nagengast, B., Janosz, M., 2013. School life and adolescents' self-esteem trajectories. Child Development. http://dx.doi.org/10.1111/cdev.12089.

Phinney, J.S., Baldelomar, O.A., 2011. Identity development in multiple cultural contexts. In: Jensen, L.A. (Ed.), Bridging Cultural and Developmental Psychology: New Syntheses in Theory, Research and Policy. Oxford University Press, New York, pp. 161–186.

Schwartz, S. (Ed.), 2012. Identity Around the World. New Directions for Child and Adolescent Development, vol. 138. Jossey-Bass, San Francisco.

Vignoles, V.L., Schwartz, S.J., Luyckx, K., 2011. Introduction: toward an integrative view of identity. In: Schwartz, S.J., Luyckx, K., Vignoles, V.L. (Eds.), Handbook of Identity Theory and Research. Springer, New York, pp. 1–28.

Relevant Websites

www.issiweb.org – International Society for Self and Identity (ISSI).
http://w3.fiu.edu/ydp/ – Miami Youth Development Project. Florida International University.
www.identitysrif.org– Society for Research on Identity Formation.

Self in Culture: Early Development

Vasudevi Reddy, University of Portsmouth, Portsmouth, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self and culture are both fluid processes. I focus on three aspects of self: the body, the awareness of self, and the regulation of behavior. In each of these aspects, the self can be seen to be open in relation and thus varying with cultural and situational influences. From very early in infancy, we can see cultural differences in motor skills, types of self-awareness, and difference of regulation of infant behavior. The development of selfhood is a dialogical, participatory, and thus a fundamentally cultural process.

> … the contents of our 'inner' lives are not so much 'inside' us as individuals, as 'in' our living of our lives. (Shotter, 1997)

Self and culture are arguably two of the three most slippery notions in the Social and Behavioral Sciences (the third, one could argue, is mind). The self cannot be contained inside the skin just as culture cannot be conceived of as outside it (Shweder, 1999). Both are fluid and continually developing. The self can best be understood as a relational entity, and thus, to understand the emergence of selfhood we need to understand its openness in relations of different kinds. In this article, I look at this openness within relation of three aspects of selfhood – the body, the awareness of self, and the regulation of self.

The Body

Cultures vary in the 'body techniques' they instill in their members (Mauss, 1979). These cultural habits of the body are not trivial in terms of selfhood. Our posture at a given moment reflects even trivial recent experiences; brief verbal priming such as the mention of old age, for instance, slows down the walking speed and posture of experimental subjects in laboratories (Bargh et al., 1996). And our posture can lead to different kinds of abilities and experiences; sitting with a confident posture (back erect and chest out) leads to greater confidence in our thinking than sitting with a doubtful posture (slouched forward and back curved, Brinol et al., 2009); and having our hands tied behind our backs impairs our ability to judge others' reaching ability (Ambrosini et al., 2012). Specific experience affects not only the way we move our bodies, but also the sounds we make with it in early infancy (Wermke, this volume) and the very experience of our body as our own. The rubber hand illusion, the whole body illusion, and enfacement illusions all show that if there is enough confounding of multisensory information we can be fooled into thinking that another body is our own, even if briefly and partially. This remarkable plasticity of body experience appears to be evident even in newborns (Filipetti et al., 2013).

Esther Thelen (2000) identified motor development as the foundation and the future of developmental psychology because motor development in this, the age of embodiment, is "not just motor anymore" (p. 394). Using Bernstein's revolutionary way of thinking about motor control not as something that comes only from the 'inside' but as something that lies in local factors such as joints, muscles, and the constraints and affordances of specific environment, we can see how situations and cultural invitations to move and to act require that children must learn, anew in each situation, the biodynamics of their bodies and the changing forces accompanying and producing each movement.

Motor development in infancy differs dramatically across cultures (Keller, 2002; see Formega and Kayenne, this volume). From the mid-twentieth century we know that African infants show motor precocity in a range of actions. This precocity appears to be cultural rather than genetic. Ugandan infants, for instance, are more advanced, especially in the first year of life, than American infants, regardless of race. The African advantage also seems to disappear when the children grow up in Western environments. Differences in motor skill appear to be specific to local cultural encouragement or discouragement of specific motor practice: Brazilian infants sit and grasp later than American infants, arguably because their mothers are more protective about physical risks in allowing their infants to sit and stand. Cultural variations in age of onset and proficiency in motor skills can come from different sources, from variations in practices of handling infants (which are largely logistic and pragmatic in specific environments) from ethnotheories about the stimulation of general motor skills, and from practices for training for specific skills.

Variations in maternal practices in handling infants – even newborns – vary, and these may enhance specific muscular development. Uganda mothers, for example, tend to pick up infants around the chest without head or back support, or lift by one of the shoulders or even lower newborns into cribs in this way. Holding infants out when the infant begins to urinate, or holding them in a squatting position to defecate may also help to strengthen muscles in the legs and back. Informal routines such as keeping babies lying down or sitting up also vary across cultures. French infants, for instance, spend 50–60% of their day lying down, while infants in Uganda, Botswana, and Mali spend 40–90% of their day sitting or standing, or being carried. Restricted time in a prone position in back-sleepers (the most common practice in the United Kingdom at the moment) is linked with later onset ages for crawling and related skills that require the use of upper trunk and arms. Similarly, methods of carrying and transporting

infants vary enormously between cultures and afford the infant very different opportunities to explore their environments and give them different postural challenges to cope with. When mothers carry infants on their backs rather than in their arms, the infant requires better head control to compensate for the mother's movements, has the advantage of more practice at sitting (on the back) and can reach and grasp more objects with their arms. Such practice at frequent postural adjustment and opportunities for manual exploration may result in accelerated reaching (e.g., in the !Kung) and sitting (Adolph et al., 2012).

In many cultures, there are accepted practices – and different ethnotheories – for stimulating general motor development. Formal baby massage and stretching are common in many cultures in Africa, the Caribbean, and India and have often been shown to accelerate developmental progress (Adolph et al., 2012; Keller et al., 2002). Formal training of infant motor skills is not something that is common or even encouraged in Europe and North America. And yet, the specific effects of such training are clear; infant reaching skills can be accelerated, and infant crawling can be accelerated (down to five and a half months) by mothers who teach these specific skills. Jamaican mothers in the United Kingdom who continue the massage and formal training of their native culture have infants with better head control at 1 month (also compared to English infants), who sit and walk earlier than the Jamaican infants of mothers who do not continue the massage and formal training (Hopkins and Westra, 1988, 1990). The effects of training generally seem to be specific; the frequency of formal exercise in sitting and walking predict the age of sitting and walking, but not the age of crawling. Mistakenly believing that it is better not to help with the acquisition of motor skills can lead to a delay in several motor milestones and reflects an odd assumption of the closed-ness to cultural influences of motor development.

Does advancing the age of onset or increasing the proficiency of motor skills matter for selfhood? There is considerable evidence that new locomotor skills (such as crawling or walking) transform the infant's relation to the world, opening up new and different opportunities for action and discovery and allowing different sociocognitive skills (Campos et al., 2000). Cultural variations of practice as well as expectation in terms of infant mobility thus open up new avenues for cultural differences in social cognition. What others do with our bodies in infancy affects what we do with our bodies, which in turn makes a difference to how we think about and perceive ourselves in the world.

Awareness of Self

Being is often influenced by the awareness of being. Do early experiences also influence the extent to which and more importantly, the way in which, we are aware of ourselves? The reflexive awareness of self has in the past been lauded for being the great achievement of *Homo sapiens*; although this claim to exclusivity is less strongly made nowadays, the complexity and impact of human self-consciousness on human lives is beyond doubt. This section will focus on two aspects of self-awareness that might be open to cultural variations: the way in which infants reach an awareness of the I, and the way in which they experience self-conscious affects.

Detecting the 'I'

The ability to distinguish self from nonself is an early one and very likely to emerge from the ability for agentic movement. Neonates can tell the difference between their own hand brushing their cheek and another person's hand stroking their cheek, showing more rooting responses to the latter than to the former (Rochat and Hespos, 1997). Even fetuses appear to distinguish between their own body, that of their twin, and the uterine wall; upper limb movements toward the twin are of longer duration and show longer periods of deceleration than movements toward their own body (unless toward their own face) or toward the uterine wall (Castiello et al., 2010).

However, this basic ability to detect one's own movements is only half the story in the development of an awareness of the 'I' and the ability to act and do things; the other half is how the world responds, contingently, to these movements, creating a sense of successful agency as well as the sense of a responsive world. One month-old infants detect and like to explore, perfect contingencies (the kinesthetic-perceptual match between their own moving bodies and a live video of their bodies), increasing the movements of the seen rather than unseen arm even when the arm has additional weights hung on it to make it more difficult to move (Van der Meer et al., 1995).

Contingent responses from the world are clearly pleasurable, leading to smiling and experimentation in infants (Watson, 1972) and activating the reward centers of the brain in adults (Schilbach et al., 2010). The most frequent (and possibly the most pleasurable) experiences of contingency that infants experience in the early months are of course, from other people. If these vary, as they must, between cultural situations, infant propensities for acting and for exploring actions must vary too. Violations of contingent responsiveness by others are detected very early in life (Murray and Trevarthen, 1985; Nadel et al., 1999; Markova and Legerstee, 2006) and may affect patterns of attachment and relationships as well as how infants perceive and recognize themselves (Gergely, 2004). Cultural variations in the degree and style of contingent responses, which infants receive appear to vary along the dimension of independence–interdependence, with 'independent' cultures seeing contingent responses as recognition and encouragement of infants' agency, uniqueness, and expressiveness. German mothers (from an independent cultural orientation) appear to show more contingent responsiveness, while Greek mothers (more family oriented) show more warmth to their 3-month-olds; similarly, rural Cameroonian mothers (from a more interdependent orientation) show less contingent responsiveness than do German mothers (Keller et al., 2005).

At least within a Western cultural environment, receiving more contingent responses to their smiles from their mothers tends to encourage infants to make more social bids (attempts to regain engagement) during the still face procedure (McQuaid et al., 2009). However, given that infants learn very early to prefer the degree of contingent responsiveness that they are used to (Bigelow, 1998) the generalizability of this finding to other cultures cannot be

assumed. Cultural differences in forms of contingent response appear to have knock-on effects for the way in which objective self-awareness – that is, the recognition of the self as a 'me' – develops, with infants who receive greater contingent responsiveness from their mothers showing earlier mirror self-recognition (Keller et al., 2004). These findings have to be treated with some caution, since infant mirror experience also varies between cultures and rapidly influences mirror self-recognition. Although in their main experiment Cameroonian infants performed less well than German infants in mirror self-recognition, at a follow-up test only 6 weeks after the first test, the Cameroonian Nso infant performance improved dramatically. The validity of the mirror self-recognition task has been challenged in several ways, most recently in a cross-cultural study of older toddlers, that touching the self in the mirror (taken as evidence of self-recognition) itself varies between cultures, with much fewer self-referential acts in Kenya and Fiji than in the United States, Canada, Peru, and Granada (Broesch et al., 2011).

Interestingly, it may not just be the fact of or the extent of contingent responsiveness that varies between cultures, but also its content. American mothers (more than French or Japanese mothers) tend to respond to infant overtures more by directing attention to things in the environment, and Japanese mothers tend to respond more (than American or French mothers) to social looking by the infants, and with more dyadic than extradyadic responses (Bornstein et al., 1992). Similarly, the modality of contingent responses varies between cultures. Kaertner et al. (2010) compared the proportion of visual and proximal maternal contingent responses in Nso and German families. Although at the start of the study at 4 and 6 weeks of age there were few differences between the groups, by 12 weeks of age the differences were pronounced. The German mothers significantly increased their visual contingent responses and decreased their proximal contingent responses. Thus, very different features of the self are acknowledged and affirmed in different cultural groups from extremely early in life, with the impact of the differences increasing with age.

Feeling Self-Conscious

One of the most salient ways in which we reveal self-awareness is through self-conscious affectivity – typically involving emotions such as pride, shame, embarrassment, guilt, and jealousy. Although we no longer characterize cultures as cultures of guilt and cultures of shame (Benedict, 1967), the meanings of these emotions for the self and for the action potentials from these emotions vary enormously. Shame and pride appear to carry very different connotations in different cultures. In East Asian cultures, shame appears to carry with it a recognition of the ability – and the need – to communicatively rectify the damage to interpersonal relations, whereas in Western cultures shame is more often seen as an individual failing with no option for action, as something that cannot be rectified, and thus can only be responded to by withdrawal and hiding. With such a difference in cultural framing, it is not surprising to find the daily practice of communal *hansei* or self-criticism in Japanese classrooms in sharp contrast with the American practice of encouraging children to report and focus on their worthy achievements (Miller et al., 1996). Similarly, adult Japanese respondents appear to prefer to report a situation of shame or humiliation over one of pride, a preference that would boggle most Anglo-American adults who prefer instead to hide it. Shame in East Asian cultures is "one way of showing social engagement" (p. 163), with options – or even a requirement – for public action and reparation (Mesquita and Karasawa, 2004). Consistent with its 'invisibility' in Western adulthood, the ontogenetic emergence of shame appears to be relatively late in Western cultures (Miake and Yamazaki, 1995). Pride, on the other hand, appears to be something that Anglo-American cultures are more comfortable with, with a well-known emphasis on remembering to praise children's achievements (Levine et al., 1980). This is consistent with the general encouragement at all ages to show off one's achievements rather than hide them in embarrassment or be self-effacing. That infants before the end of the first year are proficient in showing off (at least in the United Kingdom) is known (Reddy, 2005); whether there are differences between Anglo-American cultures and Asian or African cultures in terms of infant or child proclivity to show off is as yet unknown.

Shyness or bashfulness is another self-conscious affect that is generally seen as a problem in Anglo-American cultures (as juvenile, immature, and perhaps even socially problematic). In contrast, it is seen as attractive in India and in East Asia. A quick look at movie hoardings in India would yield an impression that attractiveness in young women is associated with demureness and smiling with sideways glances. Indeed, even in a work context, the term conveys different cultural values, implying a desirable quality in employment in India, as opposed to a distinctly worrying quality in the United States (Shweder, 2003). It has also been claimed that the Japanese kimono with its long sleeves particularly favors the revealing of smiling eyes without showing the smiling mouth – the classic sign of a coy smile. Infants do show positive shyness very early in life (Reddy, 2005) but we do not know as yet whether this affective expression – charming as it is in infants – is responded to differently in different cultures, nor whether its frequency differs. The suggestion that the reasons for self-conscious affects change with age, emerging in more interpersonal context early in infancy but in more triadic contexts later (Mesquita and Karasawa, 2004; Reddy, 2005), suggests an intriguing possibility that in more interdependent cultures there may be a longer continuation of the interpersonal contexts of shame, pride, and embarrassment. If this is the case, then our attempts to understand self-awareness must shift to the kinds of interactions and actions that prompt selves to feel self-conscious.

Regulating Behavior

The regulation of behavior is a problematic issue in psychology. On the one hand, the pervasive cultural value in most of Europe and America is for freedom from control by others. The early Freudian assumption of a tension between society and the individual is still a familiar one. On the other hand, given that the only way out of this tension is through some regulation of individual behavior, (Western) psychology

puts its trust in the need for self-regulation on the basis of internalized rules. The solution is culturally consistent; the individual still has the freedom to regulate himself or herself without direct pressure from others, with the internalized rules forming a depersonalized and more acceptable intrusion on freedom. The issue of self-regulation has thus become central in psychology, with other-regulation (regulation by others) seen as a developmentally early and temporary form of influence.

Within this framework, adult directives to children hold an ambivalent position. They could be seen to be limiting the child's ability to act on their own and curtailing rather than expanding the horizons of potential actions. The Piagetian–Vygotskyan debate about the role of the teacher is directly relevant here, seen either as the person who stops the child from making his or her own discoveries or as the person who opens doors through a scaffolding process to help the child make more discoveries. Studies adopt a dim (and ethnocentric) view of parents who issue 'too many' directives, simplistically equating it with intrusiveness, status orientedness, and insensitivity (Pine, 1992). Children's compliance with directives holds a similarly ambivalent position; while docility with immediate control is seen as not necessarily desirable, compliance with rules – an internalized and self-regulating compliance as it were – is seen as more mature.

However, there is considerable variation between cultures in attitudes toward both these things – direct regulation by others and compliance with such regulation (see also Kagitcibasi, this volume). Western psychology's value for the internal may have led to a relative ignorance of how adults do, in fact, engage in participatory regulation with their infants and children, with such regulation opening up richer developmental opportunities for infants and children (Rogoff, 2003; Kaertner et al., 2010). In interdependent cultures, adults appear to be more indulgent and warm, more regulating, and more 'connected' with their children in contrast with independent cultures (there is less known and less agreement about the direct effects of this on the children). Most findings about such cultural variations appear to fall within a great divide narrative and thus must be treated with some reservations (Raeff, 2006; Kagitcibasi, 2005).

Japanese teachers (in an immersion school in the United States as well as at a school in Japan) have been shown to use more direct forms of command (including the declarative command – e.g., 'you shall sit down') than their English counterparts who favor less direct forms and questions. The intimacy with the children that often accompanies the direct commands appears to make the directness inoffensive and appropriate (Falsgraf and Majors, 1995). Greater use of direct commands has also been found to characterize parents on 'welfare' in the United States as compared to middle-class parents (Hart and Risley, 1995). Iranian parents in the United States (in contrast to Anglo-American parents) use more directives to their preschool children in a goal-oriented activity, although this style is unrelated to sensitivity (Kermani and Brenner, 2000). Similarly, Chinese immigrant mothers use more attention-directing than attention-following styles with their infants (Vigil, 2002). Individual differences in collectivist attitudes within a culture (e.g., in Estonia) do not, however, appear to be linked to maternal directiveness, with greater collectivism linked instead to greater talkativeness with their children (Tulviste, 2001). Directiveness, however, is not a unitary phenomenon; it varies in degrees of supportiveness and in intrusiveness and even intrusive directives can have positive consequences, increasing infants' attention to toys and play (Masur et al., 2013).

In general, patterns of compliance in toddlers and children in more interdependent cultures also seem to imply a deeper commitment and cooperativeness in the children. In a study of Greek, Costa Rican, and Nso families (which could be seen as varying in that order from more independent to more interdependent orientations), Nso toddlers (18–20 months) showed greater rates of internal regulation (complying without reminders) than did the Costa Rican toddlers, with the Greek toddlers being the least likely to comply. The amount of close body contact at 3 months – highest in the Nso – was found to be the strongest predictor of later internal regulation (Keller et al., 2004). A similar differentiation between independent and interdependent cultures can be seen in a study showing that Chinese toddlers in Canada display more committed compliance than Canadian toddlers who show more situational compliance (Chen et al., 2003). One explanation of this is that the independence-encouraging Canadian mothers allow their infants to take risks and rely on their own monitoring and reminding functions to manage the child, while the Chinese mothers use warmth, contact, and connectedness expecting (and getting) cooperation in return from the child.

However, compliance and noncompliance may not follow the same pattern through childhood, and may not have the same meaning at all ages or in all groups. Reddy (1983), for instance, found that at 7 years children in India were less compliant than children in the United Kingdom. Although there were more frequent adult directives in India, Indian parents showed greater eventual acceptance of noncompliance than in the United Kingdom. Compliance in India was often a longer drawn out negotiation than in the UK families. It is not necessarily the case, therefore, that greater regulation by others leads to greater self-regulation. Greater efforts at regulation may also accompany greater indulgence and tolerance of noncompliance and maternal directiveness does not predict child compliance in either Japanese or American 5-year-olds (Abe and Izard, 1999). In a study of compliance at the earliest stages of its emergence, before the end of the first year, infants in India showed higher absolute frequencies of compliance with adult directives than infants in the United Kingdom, but showed no differences at all in proportions of compliance (Reddy et al., 2013).

Independence–interdependence and directiveness may be too blunt a set of instruments with which to understand selfhood and regulation in childhood. A closer look at the content of directives, especially early in infancy, might be more informative. Even between 6 and 12 months, infants in India more than in the United Kingdom, receive an impressive range of adult requests for participation in cultural activities (such as prayer and ritual actions) and for social reenactment of interpersonal gestures and expressions (such as 'show how granddad gets angry' or 'show how you cry,' Reddy et al., 2013). Infants are drawn into culture both in relation to specific skills and into the practice of participation and may benefit from this in the same

way as they benefit from being trained in motor skills. Directives can be dialogical opportunities as well as intrusions, their meaning varying across situations and cultures in a multitude of ways. Such dialogical opportunities open doors for development and posit the developing infant as a dialogical self (Bento et al., 2012) born in intersubjectivity (Trevarthen, 1993). The difference between self-regulation and other-regulation may be more apparent than real.

Conclusions

The openness of self to culture can be seen in diverse ways in which we connect with and move with the world. In this article, we have focused on three aspects of self – the body, awareness of the self, and regulation of the self – which are open to engagement with the world. The self cannot be seen as an independent entity bounded by the skin. Whether we are talking about the way we sit or walk or move, or the way we are aware of our ability to be and to act, or the way in which our behavior is regulated in everyday life, our selves are dialogical entities. They contain the cultural world, just as the cultural world contains and is itself changed by the participation of each of these selves.

Acknowledgement

This writing was supported by the Marie-Curie ITN "TESIS: toward an Embodied Science of InterSubjectivity" (FP7-PEOPLE-2010-ITN, 264828).

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Culture and the Self: Implications for Psychological Theory; Intentionality During Infancy and Early Childhood, Development of; Motor Skills: Development in Infancy and Early Childhood; Neonatal Crying Behaviors; Parenting Attitudes and Beliefs across Cultures; Personality and Self; Self-Development in Childhood and Adolescence; Self-Regulated Learning; Self: History of the Concept; Social Referencing During Infancy and Early Childhood across Cultures.

Bibliography

Abe, J.A.A., Izard, C., 1999. Compliance, non-compliance strategies and the correlates of compliance in 5-year-old Japanese and American children. Social Development 8 (1), 1–20.

Adolph, K., Karasik, L., Tamis-LeMonda, C., 2012. Moving between cultures: cross-cultural research on motor development. In: Bornstein, M. (Ed.), Handbook of Cross-cultural Developmental Science. Psychology Press, New York.

Ambrosini, E., Sinigaglia, C., Costantini, M., 2012. Tie my hands, tie my eyes. Journal of Experimental Psychology: Human Perception and Performance 38 (2), 263–266.

Bargh, J., Chen, M., Burrows, L., 1996. Automaticity of social behaviour: direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology 71 (2), 230–244.

Benedict, R., 1967. The Chrysanthemum and the Sword: Patterns of Japanese Culture. Meridien Books, Cleveland.

Bento, T., Cunha, C., Salado, J., 2012. Dialogical theory of selfhood. In: Vaalsiner, J. (Ed.), The Oxford Handbook of Culture and Psychology. Oxford University Press, New York, pp. 421–438.

Bigelow, A., 1998. Infants' sensitivity to familiar imperfect contingencies in social interaction. Infant Behavior and Development 21 (1), 149–162.

Bornstein, M., Tamis-LeMonda, C., Tal, J., Ludemann, P., Toda, S., Rahn, C., Pecheux, M.-G., Azuma, H., Vardi, D., 1992. Maternal responsiveness to infants in three societies: the United States, France and Japan. Child Development 63, 808–821.

Brinol, P., Petty, R., Wagner, B., 2009. Body posture effects on self-evaluation: a self-validation approach. European Journal of Social Psychology 39, 1053–1064.

Broesch, T., Callaghan, T., Heinrich, J., Murphy, C., Rochat, P., 2011. Cultural variations in children's mirror self-recognition. Journal of Cross-cultural Psychology 42 (6), 1018–1029.

Campos, J., Anderson, D., Barbu-Roth, M., Hubbard, E., Hertenstein, M., Witherington, D., 2000. Travel broadens the mind. Infancy 1 (2), 149–219.

Castiello, U., Becchio, C., Zoia, S., Nelini, C., Sartori, L., et al., 2010. Wired to be social: the ontogeny of human interaction. PLoS ONE 5 (10), e13199.

Chen, X., Rubin, K., Liu, M., Chen, H., Wang, L., Li, D., Gao, X., Cen, G., Gu, H., Li, B., 2003. Compliance in Chinese and Canadian toddlers: a cross-cultural study. International Journal of Behavioural Development 27 (5), 428–436.

Falsgraf, C., Majors, D., 1995. Implicit culture in Japanese immersion classroom discourse. The Journal of the Association of Teachers of Japanese 29 (2), 1–21.

Filipetti, M.L., Johnson, M.H., Lloyd-Fox, S., Dragovic, D., Farroni, T., 2013. Body perception in newborns. Current Biology 23 (23), 2413–2416.

Gergely, G., 2004. The role of contingency detection in early affect-regulative interactions and in the development of different types of attachment. Social Development 13 (3), 468–478.

Hart, B., Risley, T., 1995. Meaningful Differences in the Everyday Experiences of Young American Children. Brookes, Baltimore.

Hopkins, B., Westra, T., 1988. Maternal handling and motor development: an intracultural study. Genetic. Social and General Psychology Monographs 114 (3), 377–408.

Hopkins, B., Westra, T., 1990. Motor development, maternal expectations and the role of handling. Infant Behavior and Development 13 (1), 117–122.

Kaertner, J., Keller, H., Chaudhry, N., 2010. Cognitive and social influences on early prosocial behavior in two sociocultural contexts. Developmental Psychology 46 (4), 905–914.

Kaertner, J., Keller, H., Yovsi, R., 2010. Mother-infant interaction during the first 3 months: emergence of culture-specific contingency patterns. Child Development 81 (2), 540–554.

Kagitcibasi, C., 2005. Autonomy and relatedness in cultural context. Journal of Cross-cultural Psychology 36 (4), 403–422.

Keller, H., 2002. Development as the interface between biology and culture: a conceptualisation of early ontogenetic experiences. In: Keller, H., Poortinga, Y., Schonmerich, A. (Eds.), Between Culture and Biology: Perspectives on Ontogenetic Development. Cambridge University Press, New York.

Keller, H., Kaertner, J., Borke, J., Yovsi, R., Kleis, A., 2005. Parenting styles and the development of the categorical self: A longitudinal study on mirror self-recognition in Cameroonian Nso and German families. International Journal of Behavioral Development 29 (6), 496–504.

Keller, H., Yovsi, R., Borke, J., Kaertner, J., Jensen, H., Papaligoura, Z., 2004. Developmental consequences of early parenting experiences: self-recognition and self-regulation in three cultural communities. Child Development 75 (6), 1745–1760.

Keller, H., Yovsi, R., Voelker, S., 2002. The role of motor stimulation in parental ethnotheories: the case of Cameroonian Nso and German women. Journal of Cross-cultural Psycology 33 (4), 398–414.

Kermani, H., Brenner, M., 2000. Maternal scaffolding in the child's zone of proximal development across tasks: cross-cultural perspectives. Journal of Research in Childhood Education 15 (1), 30–52.

Levine, R., Caron, New, 1980. Anthropology and child development. New Directions for Child and Adolescent Psychology 8, 71–86.

Markova, G., Legerstee, M., 2006. Contingency, imitation and affect sharing: foundations of infants' social awareness. Developmental Psychology 42 (1), 132–141.

Markus, H.R., Kitayama, S., 1991. Culture and self: implications for cognition, emotion and motivation. Psychological Review 98 (2), 224–253.

Masur, E., Flynn, V., Lloyd, C., 2013. Infants' behaviours as antecedents and consequents of mothers' responsive and directive utterances. First Language 33 (4), 354–371.

Matsumoto, D., 1999. Culture and self: an empirical assessment of Markus and Kitayama's theory of independent and inter-dependent self-construals. Asian Journal of Social Psychology 2, 289–310.

Mauss, M., 1979. The notion of body techniques. In: Brewster, B. (Ed. and Trans.), Sociology and Psychology Essays. Routledge and Kegan Paul, London.

McQuaid, N., Bibok, M., Carpendale, J., 2009. Relation between maternal contingent responsiveness and infant social expectations. Infancy 14 (3), 390–401.

Mesquita, B., Karasawa, M., 2004. Self-conscious emotions as dynamic cultural processes. Psychological Inquiry 15 (2), 161–166.

Miake, K., Yamazaki, K., 1995. Self-conscious emotions, child rearing, and psychopathology in Japanese culture. In: P.Tangney, J., Fischer, K.W. (Eds.), Self-conscious Emotions. Guilford, New York, pp. 488–504.

Miller, J., Fung, H., Mintz, J., 1996. Self-construction through narrative practices: a Chinese and American comparison of early socialisation. Ethos 24, 237–280.

Murray, L., Trevarthen, C., 1985. Emotional regulation of interactions between two-month-olds and their mothers. In: Field, T., Fox, N. (Eds.), Social Perception in Infants. Ablex, Norwood, NJ.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., Reserbat-Plantey, D., 1999. Expetancies for social contingency in 2-month-olds. Developmental Science 2 (2), 164–173.

Pine, J., 1992. Maternal style at the early one-word stage: re-evaluating the stereotype of the directive mother. First Language 12, 169–186. Developmental Review 17 (3), 205–238.

Raeff, C., 2006. Multiple and inseparable: conceptualising the development of independence and interdependence. Human Development 49, 96–121.

Reddy, V., 1983. Responsiveness and Rules: Parent–Child Interaction in Scotland and India (Ph.D. thesis). University of Edinburgh, UK.

Reddy, V., 2005. Feeling shy and showing-off: self-conscious emotions must regulate self-awareness. In: Nadel, J., Muir, D. (Eds.), Emotional Development. Oxford University Press, London, pp. 183–204.

Reddy, V., Liebal, K., Hicks, K., Jonnalagadda, S., Chintalapuri, B., 2013. The emergent practice of infant compliance: an exploration in two cultures. Developmental Psychology 49 (9), 1754–1762.

Rochat, P., 2004. Emerging co-awareness. In: Bremner, G., Slater, A. (Eds.), Theories of Infant Development. Blackwell, Oxford.

Rochat, P., Hespos, S., 1997. Differential rooting response by neonates: evidence for an early sense of self. Early Development and Parenting 6, 105–112.

Rogoff, B., 2003. The Cultural Nature of Human Development. Oxford University Press, Inc., New York.

Schilbach, L., Wilms, M., Eickhoff, S., et al., 2010. Minds made for sharing: initiating joint attention recruits reward-related circuitry. Journal of Cognitive Neuroscience 22 (12), 2702–2715.

Shotter, J., 1997. The social construction of our inner lives. Journal of Constructivist Psychology 10 (1).

Shweder, R., 1999. Why cultural psychology? Ethos 27 (1), 62–73.

Shweder, R., 2003. Toward a deep cultural psychology of shame. Social Research: An International Quarterly 70 (4), 1109–1129.

Thelen, E., 2000. Motor development as foundation and future of developmental psychology. International Journal of Behavioral Development 24 (4), 385–397.

Trevarthen, C., 1993. The self born in intersubjectivity: the psychology of an infant communicating. In: Neisser, U. (Ed.), The Perceived Self: Ecological and Interpersonal Sources of Self-knowledge, Emory Symposia in Cognition, vol. 5. Cambridge University Press, New York, pp. 121–173.

Tulviste, T., 2001. Can variation in mother–child interaction be explained by context and collectivist attitudes? Applied Psycholinguistics 22 (4), 541–553.

Van der Meer, A., van dr Weel, F., Lee, D., 1995. The functional significance of arm movements in neonates. Science 267, 693–695.

Vigil, D., 2002. Cultural variations in attention regulation: a comparative analysis of British and Chinese-immigrant populations. International Journal of Language and Communication Disorders 37 (4), 433–458.

Watson, J., 1972. Smiling, cooing and the 'game'. Merrill-Palmer Quarterly 18 (4), 323–339.

Self: History of the Concept

Peter Wagner, University of Warwick, Coventry, UK; and European University Institute, San Domenico di Fiesole, Italy

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13833–13837, © 2001, Elsevier Ltd.

Abstract

It is often asserted that the concept of the self emerges only in early modern times in connection with the concern for subjectivity that is taken to be characteristic of modernity. While it is true that the term 'self' as a noun, describing that which in a person is really and intrinsically this person, comes into use only from the seventeenth century onwards, what can be called the question of selfhood was not unfamiliar to earlier thinkers. This is, in its core, the question whether there is – and if so, what is – some unity, or at least continuity and coherence, of a human being over the life-course beyond their bodily constitution, beyond the unity of the body. Broadly, empirical traditions tended to doubt the existence of such unity. They could observe changes, even radical transformations, in the human mind. Transcendental traditions, in contrast, tended to argue that there had to be a unity of apperception, or of consciousness. In the intellectual space between these two positions, so to say, the issue could be addressed in ways that are more specific to the social and psychological sciences. Starting out from observations by Georg Simmel and George Herbert Mead on the formation of identities in social interaction, a sociology and social psychology of selfhood and identity has developed that could conceptualize variations of self-formation in different social contexts. Over the twentieth century, research on selfhood has developed in a broad semantic space in which aspects of the self are emphasized in highly various ways, connecting the concept of self to notions of modernity, of meaning, and of difference.

It is often asserted that the concept of the self emerges only in early modern times in connection with the concern for subjectivity that is taken to be characteristic of modernity. While it is true that the term 'self' as a noun, describing that which in a person is really and intrinsically this person, comes into use only from the seventeenth century onwards, what can be called the question of selfhood was not unfamiliar to earlier thinkers. This is, in its core, the question whether there is – and if so, what is – some unity, or at least continuity and coherence, of a human being over the life-course beyond their bodily constitution, beyond the unity of the body.

In the history of human thought, this question has been answered in a great variety of different ways. Broadly, empirical traditions tended to doubt the existence of such unity. They could observe changes, even radical transformations, in the human mind and had thus no ground to postulate any *a priori* unity. Transcendental traditions, in contrast, tended to argue that there had to be a unity of apperception, or of consciousness. Otherwise, not even the question of the self could be asked.

In the intellectual space between these two positions, so to say, the issue could be addressed in ways that are more specific to the social and psychological sciences. Then, the faculty of memory, for instance, could be seen as enabling a sense of continuous self to emerge. Human subjects have the ability to narrate their lives. Or, selves could be seen as formed in the interaction with others, that is, in the mirroring of an image of oneself through the responses to one's own words and actions by others. In moral philosophy, the existence of a continuous self was seen as a precondition for holding human beings accountable for their past deeds. Selfhood was thus linked to moral and political responsibility.

Selfhood in the Social Sciences

In these forms, the question of the self was already posed at the historical moment when the social sciences, by and large in the contemporary understanding of the term, arose towards the end of the eighteenth century. It is noteworthy, then, that the emerging social sciences rather reduced the range of ways of exploring selfhood across the nineteenth century. In some of their fields, like liberal political philosophy and political economy/economics, they postulated a rational self, able to make choices and to be responsible for her, or rather: his deeds. In other fields, in particular in the sociological way of reasoning, the orientations and behaviors of human beings were seen as determined by their social position. This opposition has become known as the one between an under- and an oversocialized conception of the human being (Wrong, 1961). More cautiously, the empirical and behavioral strands of social research restricted themselves to observing human behavior and refrained from making any assumptions about selfhood at all. Regularities emerged here only through the aggregation of observations.

All these approaches had in common, though, was that they aimed at regularizing and stabilizing human orientations and behavior. Whether human beings were under- or oversocialized or just happened to behave according to patterns, the broader question of selfhood, namely whether and how a sense of continuity and coherence in a human being forms, was rather neglected or answered *a priori* by theoretical postulate.

In this light, it seems appropriate to say that the social sciences have developed a serious interest in questions of human selfhood only late, systematically only in the early twentieth century. Furthermore, they have largely ignored ways of representing the human self that were proposed in other areas, in philosophy for instance, but maybe even more importantly in literature. Thus, issues such as individuality and subjectivity, the possible idiosyncrasy of a life-course and life-project, have long remained outside the focus of the social sciences, which have tended to look at the question from the perspective of a fully developed, stable, personal identity, rather than making the continuity and coherence of the self an issue open to investigation.

This focus can hardly be understood otherwise than through the perceived need to expect a political stability of the world on the basis of a presupposed coherence of human orientations and actions.

The ground for an interest in such broader issues was prepared outside of what is conventionally recognized as social science, namely by Friedrich Nietzsche, and later by Sigmund Freud. Nietzsche radically rejected the problems of moral and political philosophy and thus liberated the self from the impositions of the rules of the collective life. Freud located the drives toward a fuller realization of one's self in the human psyche and connected the history of civilization with the repression of such drives. Against such a 'Nietzschean-Freudian' background (Rorty, 1989), Georg Simmel and George Herbert Mead could observe the ways in which identities are formed in social interaction and conceptualize variations of self-formation in different social contexts. From then on, a sociology and social psychology of selfhood and identity has developed which no longer relies on presuppositions about some essence of human nature and is able to connect its findings to both child psychology and phenomenology. It emphasizes the socially constructed nature of selfhood, but remains capable, at least in principle, to analyze the specific social contexts of self-formation thus working towards a comparative-historical sociology of selfhood. This broadened debate on selfhood has created a semantic space in which aspects of the self are emphasized in various ways (Friese, 2001). This space can be described by connecting the concept of self to notions of modernity, of meaning, and of difference.

Selfhood and Modernity

The idea of human beings as autonomous subjects is often taken to be characteristic of modernity as an era, or at least for the self-understanding of modernity as an ethos (*see* Modernity: History of the Concept). Such a view entails a conception of the self as rather continuous and coherent, since only on such a basis is the choice of a path of action and a way of life as well as the acceptance of the responsibility for one's deeds conceivable. The concept of selfhood is here conditioned by the need to maintain a notion of human autonomy and agentiality as a basic tenet of what modernity is about, namely the possibility to shape the world by conscious human action. Such modernism, however, is not necessarily tied to the atomist and rationalist individualism of some versions of economic and political thought, most notably neoclassical economics and rational choice theory.

In the light of the twentieth century developments in sociology and (social) psychology, as mentioned above, the view that connects selfhood to modernity mostly – in all its more sophisticated forms – starts out from an assumption of constitutive sociality of the human being. Not absolute autonomy, but rather the conviction that human beings have to construct their self-identities can then be seen as characteristically modern (Hollis, 1985). Unlike the concept of the rational, autonomous self, this concept is open towards important qualifications in terms of the corporeality, situatedness, and possible nonteleological character of human action (Joas, 1996). Without having to presuppose the self-sustained individual of modernity, the aim is to demonstrate how autonomous selves develop through social interactions over certain phases of the life-course (Joas, 1998; Straub, 2001).

The commitment to autonomy and agentiality, characteristic of modernity, becomes visible rather in the fact that formation of self (or, of self-identity) is here understood as the forming and determining of the durably significant orientations in a life. It is related to the formation of a consciousness of one's own existence and thus biographically predominantly to the period of adolescence. Crises of identity occur accordingly during growing up; more precisely one should speak of life crises during the formation of one's identity. Self-identity once constituted is seen as basically stable further on. No necessary connection is presupposed between self-formation and individuality; theoretically, human beings may well form highly similar identities in great numbers. Since the very concept of identity is connected to continuity and coherence of self, however, stability is turned into a conceptual assumption.

Objections against such a conceptualization go in two different directions. On the one hand, this discourse, which often has its roots in (social) psychology, stands in a basic tension to any culturalist concept of identity, which emphasizes meaning. On the other hand, doubts about the presupposition of continuity and coherence of selfhood employ notions of difference and alterity that are not reducible to the idea that selves are formed by relating to others, by intersubjectivity.

Selfhood and Meaning

Some critics of the close connection between self-formation and modernity argue that every form of selfhood is dependent on the cultural resources that are at hand to the particular human being when giving shape to their important orientations in life. Human beings give meaning to their lives by interpreting their situations with the help of moral-cultural languages that precede their own existence and surround them. Cultural determinism is the strong version of such theorizing, mostly out of use nowadays, but many current social theories adopt a weaker version of this reasoning which indeed sustains the notion of the continuity and coherence of the self but sees this self as strongly embedded in cultural contexts.

Such a view of the self has its modern source in romanticism. Philosophically, it emerged as a response to the rationalist leanings of the Enlightenment, and politically, against the conceptions of abstract freedom in individualist liberalism. Unlike cultural determinism, however, which has a strongly oversocialized conception of the human being and thus hardly any concept of self at all, romanticism emphasizes agency and creativity in the process of self-formation and self-realization. Human beings are seen in their singularity, but the relation to others is an essential and inescapable part of their understanding of their own selves.

Charles Taylor's inquiry into *The Sources of the Self* is a most recent and forceful restatement of this conception. Taylor starts out from the familiar argument that the advent of modernity

indicates that common frameworks for moral evaluation can no longer be presumed to exist. The key subsequent contention is then that the ability and inclination to question any existing such framework of meaning does not lead into a sustainable position that would hold that no such frameworks are needed at all, a view he calls the 'naturalist supposition' (Taylor, 1989, p. 30). If such frameworks of meaning are what gives human beings identity, allows them to orient themselves in social and moral space, then they are not 'things we invent' and may as well not invent. They need to be seen as 'answers to questions which inescapably pre-exist for us,' or, in other words, 'they are contestable answers to inescapable questions' (Taylor, 1989, pp. 38, 40). Taylor develops here the contours of a concept of inescapability as part of a moral-social philosophy of selfhood under conditions of modernity. Meaning-centered conceptions of selfhood have recently been underlined as a basis of communitarian positions in moral and political philosophy, such as Sandel's concept (1982) of the 'encumbered self.'

Selfhood and Otherness

Arguably, these two concepts of selfhood remain within the frame of a debate in which the under- and the oversocialized views of the human being occupy the extreme points. The introduction of interaction and intersubjectivity in self-formation has created intermediate theoretical positions, and, possibly more importantly, they have allowed different accentuations of selfhood without making positions mutually incompatible. Nevertheless, the modernity-oriented view still emphasizes a self that acts upon the world, whereas the meaning-oriented view underlines the fact that the self is provided the sense of their actions by the world of which they are a part. Since the mid-1970s, in contrast, theoretical and empirical developments have tended to break up the existing two-dimensional mode of conceptualization.

On the one hand, the notion of a 'decentering of the subject' has been proposed mainly from poststructuralist discussions. On the other hand, the observation of both multiple and changing basic orientations in human beings has led to the proliferation of the – infelicitously coined – term 'postmodern identity' as a new form of selfhood. In both cases, a strong concept of self has been abandoned, in the one case on the basis of philosophical reflection, in the other, grounded on empirical observation. Both the ideas of a 'decentring' and of a 'postmodern self' question some major implications of the more standard sociological view of selfhood, namely the existence of the human self as a unit and its persistence as the 'same' self over time.

In the former perspective, the philosophical maxim of thinking identity and difference as one double concept, rather than as one concept opposed to another, is translated into sociological thinking as the need to think self and other as a relation, rather than as a subject and a context or object. Such notions rather underline the nonidentitarian character of being by pointing to the issue of 'the other in me' (Emmanuel Lévinas) as a question of co-constitution rather than of interaction. While emphasized in recent poststructuralist thought, similar ideas can be found, for instance, in Arendt's (1978, pp. 183–187) insistence on the 'two-in-one,' on the relation to oneself as another, as the very precondition for thought. In a broad sense, they go back to the ancient view of the friend as an 'other self.' And Theunissen (1977/1965) had already conceptualized self-other relations on the basis of an understanding of 'the Other' as referring to all 'concepts through which philosophy interprets the presence and present of the fellow/co-human being (*Mitmensch*) or of the transcendental original form of this human being.'

In one particular formulation, Cavell's reflections provide an example for a thinking about selfhood that does not presuppose an idea of identity, coherence, or consistency (Cavell, 1989, 1990). He cautions against 'any fixed, metaphysical interpretation of the idea of a self' and against the idea of 'a noumenal self as one's "true self" and of this entity as having desires and requiring expression.' In contrast, Cavell suggests that the 'idea of the self must be such that it can contain, let us say, an intuition of partial compliance with its idea of itself, hence of distance from itself' or, in other words, he advocates the idea of 'the unattained but attainable self' (Cavell, 1990, pp. 31, 34). Cavell proposes here a relation of the unattained and the attainable as constitutive for the self, that is, he makes the very question of attainability a central feature of a theory of selfhood.

Selfhood and Sociohistorical Transformations

The development of this threefold semantic space has considerably enriched the conceptualization and analysis of selfhood in the social sciences during the twentieth century. If one considers the three fields of inquiry as explorations of the various dimensions of selfhood, rather than as mutually exclusive perspectives, the question of the constitution of the human self emerges as a *problématique* that concerns all human beings and for which there may be a variety of processes which can be determined generally only in their forms but not in their results. As a consequence, however, it becomes much more difficult to conclude from the prevailing character of selfhood on social structure and political order. This has often been seen as a 'weakness' of symbolic interactionism, for instance, which Mead's view of the self is said to have inaugurated, as a social theory that allegedly cannot address issues of societal constitution. But by the same move Mead allows for and recognizes a plurality of selves that has returned to the center of discussion today – after the renewal of a regressive synthesis of identity and society in Talcott Parsons's work (to which – what is often overlooked – Erik Erikson contributed as well). The question of the relation between selfhood and society and politics thus needs to be rephrased as a question of comparative-historical sociology rather than merely of social theory.

Some indications as to how this relation should be understood can be found when reading the historical sociology of twentieth century societies in this light. The social upheaval during the second half of the nineteenth century with industrialization, urbanization and the phrasing of 'the social question' is often seen as a first 'modern' uprooting of established forms of selfhood, as a first massive process of 'disembedding' (Giddens, 1990). The development towards so-called mass societies during the first half of the twentieth century lets the question of the relation between individuation and growth of

the self emerge. Totalitarianism has been analyzed in terms of an imbalance between imposed individuation and delayed self-formation, the result having been the tendency towards 'escape from freedom' and into stable collective identities (Fromm, 1941; Arendt, 1958). The first three decades after the Second World War are then considered as a form of re-embedding of selves into the institutional frames of democratic welfare societies. Most recently, the indications of dissolution and dismantling of the rather comprehensive set of social institutions of the interventionist welfare state are one of the reasons to focus sociological debate again on questions of selfhood and identity. During this period, as some contributions argue, no longer continuity and coherence but transience, instability, and inclination to change are said to be marks of the important life orientations of contemporary human beings (see Shotter and Gergen, 1989; Lash and Friedman, 1992; Kellner, 1995). However, many of these analyses are challenged on grounds of the limited representativity of the empirical sample or on conceptual grounds.

Given the theoretical insights described above as the emergence of the threefold semantic space of selfhood, any attempt to offer a full-scale reformulation of the issue of the varieties of selfhood in different sociohistorical configurations would be adventurous. Forms of self-constitution and substantive orientations of human selves are just likely to be highly variable across large populations with varieties of experiences and sociocultural positions. However, the notion of recurring crises of modernity and the identification of historically distinct processes of disembedding and re-embedding could be the basis for a socially more specific analysis of the formation and stability of social selves (Wagner, 1994).

The social existence of the 'modern' idea that human beings construct their selves is what many societies have in common throughout the 1800s and 1900s. As such, it does not give any guidance in defining different configurations. Therefore, three qualifying criteria have been introduced. First, the existence of the idea of construction of selfhood still leaves open the question whether all human beings living in a given social context share it and are affected by it. The social permeation of the idea may be limited. Second, human beings in the process of constructing their selves may consider this as a matter of choice, as a truly modernist perspective would have it. In many circumstances, however, though a knowledge and a sense of the fact of social construction prevails, it may appear to human beings as almost natural, as in some sense pre-given or ascribed, which social self they are going to have. Third, the stability of any self one has chosen may vary. Such a construction of selfhood may be considered a once-in-a-lifetime occurrence, but may also be regarded as less committing and, for instance, open to reconsideration and change at a later age. These criteria widen the scope of constructability of selfhood. All conditions of construction have existed for some individuals or groups at any time during the past two centuries in the West. But the widening of the scope of construction may mark the transitions from one to another social configuration of modernity. These transitions entail social processes of disembedding and provoke transformations of social selves, in the course of which not only other selves are acquired but the possibility of construction is also more widely perceived.

See also: Culture and the Self: Implications for Psychological Theory; Identity and Identification; Identity in Anthropology; Modernity; Person and Self: Philosophical Aspects; Self: Philosophical Aspects; Social Identity in Sociology.

Bibliography

Arendt, H., 1958. The Origins of Totalitarianism, second ed. World Publishing Company and Meridian Books, Cleveland, OH. New York.
Arendt, H., 1978. The Life of the Mind. In: Thinking, vol. 1. Harcourt, Brace, Jovanovich, New York.
Cavell, S., 1989. The New yet Unapproachable America. Lectures After Emerson After Wittgenstein. Living Batch Press, Albuquerque, NM.
Cavell, S., 1990. Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism. The University of Chicago Press, Chicago.
Friese, H., 2001. Identity: Desire, name and difference. In: Friese, H. (Ed.), Identities. Berghahn, Oxford, UK.
Fromm, E., 1941. Escape from Freedom. Holt, Rinehart, and Winston, New York.
Giddens, A., 1990. The Consequences of Modernity. Polity, Cambridge.
Hollis, M., 1985. Of masks and men. In: Carrithers, M., Collins, S., Lukes, S. (Eds.), The Category of the Person: Anthropology, Philosophy, History. Cambridge University Press, Cambridge, UK and New York.
Joas, H., 1996. The Creativity of Action. Polity, Cambridge [trans. by Gaines J, Keast P].
Joas, H., 1998. The autonomy of the self. The meadian heritage and its postmodern challenge. European Journal of Social Theory 1 (1), 7–18.
Kellner, D., 1995. Media Culture, Cultural Studies, Identity and Politics Between the Modern and the Postmodern. Routledge, London and New York.
Lash, S., Friedman, J. (Eds.), 1992. Modernity and Identity. Blackwell, Oxford, UK.
Rorty, R., 1989. Contingency, Irony, and Solidarity. Cambridge University Press, Cambridge, UK.
Sandel, M.J., 1982. Liberalism and the Limits of Justice. Cambridge University Press, Cambridge, UK and New York.
Shotter, J., Gergen, K.J. (Eds.), 1989. Texts of Identity. Sage, London.
Straub, J., 2001. Personal and collective identity. A conceptual analysis. In: Friese, H. (Ed.), Identities. Berghahn, Oxford, UK.
Taylor, C., 1989. Sources of the Self. The Making of Modern Identity. Harvard University Press, Cambridge, MA.
Theunissen, M., 1977/1965. Der Andere. de Gruyter, Berlin.
Wagner, P., 1994. A Sociology of Modernity: Liberty and Discipline. Routledge, London and New York.
Wrong, D.H., 1961. The oversocialized conception of man in modern sociology. American Sociological Review 26, 183–193.

Self: Philosophical Aspects

Carol Rovane, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first section of this article outlines the classical problem of personal identity that was originally posed by Locke and continues to generate philosophical debate. The next section discusses some of the distinctive social relations into which only selves can enter. The article then draws several analogies between such social relations and first-personal relations, and shows how these analogies point to the possibilities of group and multiple selves. It closes with some cautionary remarks about how overlooking these two possibilities can lead to confusions about methodological individualism.

This article focuses on two related ways in which philosophical reflection on the self may usefully be brought to bear in social science. One concerns the traditional problem about personal identity. The other concerns the distinctive kinds of social relations into which only selves can enter. Sustained reflection on the second issue reveals striking analogies between some of the social relations into which only selves can enter and certain first-personal relations that only a self can bear to itself. These analogies entail two possibilities that relate to the first issue concerning personal identity, namely, the possibility that there could be group selves who are composed of many human beings and the possibility that there could be multiple selves within a single human being. If the facts about group selves are not reducible to facts about their human constituents atomistically described, it might seem to follow that methodological individualism is false. But this follows only if we assume that the only individuals that exist are individual human beings, and this assumption is rendered false by the very possibility of group selves.

As an aside that there is a third way in which philosophical reflection on the self might conceivably be brought to bear in social science, which derives from a conception of the self as having a capacity for free choice and action. When the findings of social scientific investigation are presented as *laws* of human behavior, the question immediately arises whether selves, qua free agents, have the power to deviate from those laws – in other words, does it lie within our power to choose *not* to conform to certain scientific visions of what we are like? This question is closely allied to philosophical puzzles about the self that particularly intrigued Sartre, which concern whether the unity of the self – its numerical *oneness* – is necessarily undermined by any act of first-person reflection, due to the way in which the knowing *subject* that actively undertakes such reflection must consider its own self in a different, nonactive, mode as the passively known *object* of its reflection. Prima facie, it would seem that making the self into such an object is precisely what the social scientific stance invites us to do; and seconda facie, it would seem that what is thereby known may never be a complete story insofar as there is always an active self who is taking up that stance. These mysteries of the self fall under the heading 'freedom,' and although they are not entirely irrelevant to the main concerns of this article they deserve separate discussion. The focus here will be on other respects in which the essential *oneness* of the self is not a straightforward matter, and introduces complications for two fundamental concepts of the social sciences, namely, the concept of a social relation and the concept of the self as social relatum.

The Philosophical Problem about Personal Identity

To have a self is to be self-conscious in roughly the sense that Locke took to be the defining mark of the person. In his words, a person is "a thinking, intelligent being, that has reason and reflection, and can think itself as itself in different times and places" (Locke, 1979). In general, all of these capabilities – for thought, reason, reflection, and reflexive self-reference – tend to be lumped together by philosophers under the common heading of self-consciousness. Except in some generous accounts of animal intelligence, it is generally assumed that the only fully self-conscious beings are human beings.

Locke famously disagreed. This was not because he believed that there are other species of animal besides the human species whose members actually qualify as persons (though he did discuss at some length a certain parrot that was reported to have remarkable conversational abilities). It was rather because he thought that the condition of personal identity should not be understood in biological terms at all; it should be understood, rather, in phenomenological terms. As he put it, a person's identity extends exactly as far as its consciousness extends. He saw no reason why the life of a person so construed, as a persisting center of consciousness, should necessarily coincide with the biological lifespan of a given human animal. In defense of this distinction between personal and animal identity, he offered the following thought experiment: first, imagine that the consciousnesses of a prince and a cobbler are switched each into the other's body, and then consider who would be who thereafter. He thought that anyone who does consider this is bound to see that the identity of the prince is tied to his consciousness in such a way that he would remain the same person – would remain himself – even after having gained a new, cobbling body (similar remarks apply to the cobbler).

Many objections have been raised against Locke's argument and conclusion, and it remains a matter of serious dispute among philosophers whether he was right to distinguish the identity of the person from the identity of the human animal (Parfit, 1984; Perry, 1975; Rorty, 1979). Yet this philosophical

dispute has not drawn much attention from social scientists, who generally assume against Locke that the life of each individual self or person coincides with the life of a particular human being. The anti-Lockean assumption is so pervasive in the social sciences that it serves as common ground even among those who take opposing positions on the issue of methodological individualism – whether all social facts are reducible to facts about individuals or whether there are irreducibly social facts (*see* Individualism versus Collectivism: Philosophical Aspects; Methodological Individualism).

When Locke and neo-Lockeans set out to defend his distinction between the person or self and the human being, they have not generally tried to argue that there any *actual* instances of it; they have been content to argue for the *possibility* that a person's consciousness could be transferred from one human body to another – a possibility that they have tried to illustrate with various thought experiments. One drawback of Locke's original thought experiment about the prince and the cobbler is that it merely *stipulates* that two persons' consciousnesses are switched each into the other's body, without providing any supporting account of how this is supposed to happen. Subsequent thought experiments offered by neo-Lockeans have tried to do better on this score, by describing particular processes through which this might happen, such as brain transplantation or brain reprogramming. But it must be admitted that these added details belong more to the realm of science fiction than scientifically informed speculation, and so they do not give us any realistic expectation that what is being imagined will ever actually come to pass (Wilkes, 1988). It would appear to follow that social scientists may safely set aside the entire philosophical dispute about personal identity as irrelevant to their investigations, and retain their anti-Lockean assumption as the only realistic one on which to proceed, according to which the identity of an individual self is always the same as the identity of an individual human being.

Although it is indeed tempting to draw this conclusion, it would be a mistake to do so. Some new reasons have come to light for embracing a version of Locke's distinction, which are not derived from thought experiments but from a philosophical account of the nature of distinctively personal agency, and these reasons direct our attention to some *actual* instances of Locke's distinction that do potentially bear on the empirical concerns of social science. This will emerge over the course of the next several sections.

The Sociality of Selves: Rational Relations

Starting roughly with Hegel, various philosophers have mounted arguments to the effect that self-consciousness is a social achievement that simply cannot be possessed by beings who stand outside of all social relations (Wittgenstein, 1953; Davidson, 1984). The common conclusion of these arguments is overwhelmingly plausible, and it is clearly supported by a vast range of empirical work in developmental psychology. So even if none of the strictly philosophical arguments for the conclusion were wholly convincing, we could still take for granted the following well-confirmed empirical facts: self-conscious beings typically are social, and their social relations reflect their self-consciousness in interesting ways. These facts gave rise to a special field of study within micro-sociology during the last century, pioneered by Mead, called "symbolic interactionism" (Mead, 1934; Blumer, 1969).

In order to gain a full appreciation of the *philosophical* significance of the social dimension of self-consciousness, we must shift our attention away from the phenomenological dimension of it that Locke had made central in his analysis of the condition of personal identity, and shift our attention to another dimension of it that he emphasized in his initial definition of the person as a "a thinking, intelligent being, *with reason and reflection*." Obviously, one need not cite Locke as an authority in order to defend the point that exercising a capacity for reflective rationality is crucial for the achievement of selfhood, and that it is even more crucial than being the site of a single consciousness. But the point deserves careful elaboration, in order to expose some of its less obvious consequences.

Arguably, every sentient animal is the site of a single consciousness, even animals that cannot plausibly be said to either have or be a self – for example, wallabies and cockatoos. Locke's analysis of personal identity seems to suggest that what marks the difference between such animals and self-conscious beings is that the latter have a capacity for reflexive thought and reflexive reference – thoughts that are expressed in the first-person singular, with the pronoun 'I' – a pronoun whose reference must, it seems, be a self. The trouble with this suggestion is that if unself-conscious animals like wallabies and cockatoos can represent things at all (which some philosophers deny (Davidson, 1984)), it is very likely that some of their representations will function in ways that are characteristic of reflexive self-representations. Consider, for example, mental maps in which such animals represent their spatial relations to the various objects they perceive; such maps require a way to mark the distinction between their own bodies and other bodies, a distinction which is not inaptly characterized as the distinction between self and other (Evans, 1982).

Self-consciousness in the sense that goes together with selfhood involves a much more sophisticated sort of self-representation that incorporates a conception of oneself as a thinking thing and, along with it, the concept of thought itself. Armed with this concept, a self-conscious being can do more than represent itself; it can self-ascribe particular thoughts, and then also take up the critical perspective of reflective rationality toward its own thoughts.

Whenever selves in this full sense enter into social relations, they enter into relations of reciprocal recognition. Each is self-conscious; each conceives itself as an object of the other's recognition; each conceives the other as so conceiving itself (i.e., as an object of the other's recognition); each conceives both as having a common nature and, in so doing, each conceives both as so conceiving both (i.e., as having a common nature). Since this reciprocally recognized common nature is social as well as rational, it affords the possibility of the social exercise of rational capacities. Thus, corresponding to the capacity to ascribe thoughts to oneself, there is the capacity to ascribe thoughts to others; corresponding to the capacity to critically evaluate one's own thoughts there is the capacity to critically evaluate the thoughts of others; and together with all this comes another extraordinary social capacity, the capacity for specifically rational influence in which selves aim

to move one another by appealing to the normative force of reasons. Such rational influence is attempted in virtually every kind of social engagement that takes place among selves, including engaging in ordinary conversation and argument, offering bribes, lodging threats, bargaining, cooperating, and collaborating. In all of these kinds of engagement reasons are offered up in one guise or another in the hope that their normative force will move someone else.

Analogies between First-Personal and Social Relations

It is because selves have a rational nature that analogies arise between their first-personal relations to their own selves and their social relations to one another. Here are three examples.

Critical Evaluation

It is a platitude that the principles of rationality are both impersonal and universal. They lay down normative requirements that *anyone* must meet in order to be rational, and they constitute the standards by which *everyone* evaluates their own and everyone else's rational performances. It follows that when we evaluate our own thoughts by these standards, we are really considering them as if they were anyone's – which is to say, as if they were not our own. We sometimes describe this aspect of critical self-evaluation with the phrase 'critical distance.' There is a specific action by which we accomplish this distancing: we temporarily suspend our commitment to act on our thoughts until after we have determined whether keeping the commitment would be in accord with the normative requirements of rationality. If we determine that it would not be in accord with those requirements, then the temporary suspension of our commitment becomes permanent, and our critical distance on our thoughts gives way to literally disowning them. But before getting all the way to the point of disowning our thoughts we must first have gotten to a point where our relations to our thoughts are rather like our relations to the thoughts of others. For, whether we decide to disown the thoughts under critical scrutiny or to reclaim them, the normative standards on the basis of which we do so are necessarily conceived by us as impersonal and universal. In consequence, there is a straightforward sense in which we are doing the same thing when we apply the standards to our own thoughts and when we apply them to the thoughts of others. There is this difference, though: whereas self-evaluation involves distancing ourselves from our own thoughts, evaluating the thoughts of others involves a movement in the opposite direction. We must project ourselves into their points of view so that we can appeal to their own sense of how the principles of rationality apply in their specific cases. This sort of criticism that involves projection is sometimes called 'internal criticism,' in contrast with the sort of 'external criticism' that proceeds without taking into account whether its target will either grasp it or be moved by it (Williams). This difference between internal criticism of others and self-criticism – the difference between projection and distancing – does nothing to undermine the analogies between them. On the contrary, for by these opposite movements two distinct selves can bring themselves to the very same place. That is, when one self projects itself into another's point of view for the purposes of internal criticism, it can reach the very same place that the other gets to by distancing itself from its own thoughts for the purposes of self-criticism. Moreover, there is a sense in which what the two selves do from that place is not merely analogous, but the very same thing. Each of the two selves brings to bear the same critical perspective of rationality on the same particular self's thoughts. The only difference that remains is that in the one case it is a social act directed at another, whereas in the other case it is a self-directed act.

Joint Activities

Just as internal criticism of another replicates what individual selves do when they critically evaluate themselves, so also joint activities carried out by many selves may replicate the activities of an individual self. For the purposes of such activities, distinct selves need to look for areas of overlap between (or among) their points of view, such as common ends, common beliefs about their common ends, and common beliefs about how they might act together so as to realize those ends. These areas of commonly recognized overlap can then serve as the basis from which the distinct selves deliberate and act together and, as such, the areas of overlap will serve as a common point of view that is shared by those selves. Whenever distinct selves do deliberate together from such a common point of view, they will be guided by the very same principles that ordinarily govern individual deliberation, such as consistency, closure (the principle that one should accept the implications of one's attitudes), and transitivity of preferences. All such principles of rationality taken together define what it is for an individual to be fully or optimally rational. For this reason, they generally apply separately to each individual but not to groups of individuals. (Thus, different parties to a disagreement might be individually rational even though they contradict each other, but they cannot be individually rational if they each believe contradictions.) The one exception is joint activities in which such groups of individuals deliberate together from a common point of view. To take a limited example, you and I might deliberate individually about a philosophical problem or we might do so together. If we do it individually, each of us will aim to be consistent in our reasoning because that is what the principles of rationality require of us individually. But these principles do not require us to do this together; which is to say, you and I may disagree about many things without compromising our ability to pursue our individual deliberations from our separate points of view in a completely rational manner. What would be compromised, however, is our ability to work on a philosophical problem together. We cannot do that without resolving at least some of our disagreements, namely, those that pertain to the problem. This means that if we truly are committed to working on the problem together, then we should recognize a sort of conditional requirement of rationality, a requirement to achieve as much consistency between us as would be necessary to carry out our joint project. Although we need not achieve overall consistency together in order to do this – that is, we need not agree on every issue in every domain – nevertheless, we must resolve our disagreements about the relevant philosophical matters. When we resolve these disagreements we will be doing within a limited

domain the very same thing that individuals do within their whole points of view when they respond to the rational requirement of consistency. The same holds for the other principles of rationality such as closure and transitivity of preferences. In general, whenever distinct individuals pursue ends together, they temporarily cease deliberating and acting separately from their individual points of view in order to act and deliberate together from their common point of view. Correlatively, they strive to achieve by social means within a group – albeit in a limited domain – the very sorts of rational relations and outcomes that are characteristic of the individual self (Korsgaard, 1989).

Rationality over Time

Although there is some sense in which individual points of view are extended in time, the deliberations that proceed from such points of view always take place in time from the perspective of the present. The question immediately arises, why should an individual take its future attitudes into account when it deliberates from the perspective of the present? The question becomes poignant as soon as we concede that an individual's attitudes may change significantly over time. Thus, consider Parfit's Russian nobleman who believes as a young man that he should give away all of his money to the poor, but who also believes that he will become much more conservative when he is older and will then regret having given the money away. What should he do? Parfit answers as follows: if he recognizes any normative pressure toward an altruistic stance from which he is bound to take the desires of others into account, then he must extend that stance to his own future desires as well – regardless of the fact that he now disapproves of them (Parfit, 1984). Other philosophers have also insisted on parity between the normative demands of prudence and altruism. Among moral philosophers, the parity is often taken in the positive way that Parfit does, as showing that we have reason to be both prudent and altruistic (Sidgewick, 1981; Nagel, 1970). But some philosophers have taken it in a negative way, as showing that we lack reason to be either. According to them, it follows from the fact that deliberation always proceeds from the perspective of the present that rationality does not require us to take our own future attitudes into account any more than it requires us to take the attitudes of others into account. If they are right, then the rational relations that hold within an individual over time are comparable to social relations within a group (Elster, 1979; Levi, 1985). However, anyone who is struck by this last suggestion – that an individual over time is, rationally speaking, analogous to a group – would do well to bear in mind what the first two analogies brought out, which is that the social exercise of rational capacities within a group can literally replicate individual rationality.

Group Selves and Multiple Selves

It might seem implausible that the normative analogies between first-personal and social relations could ever jeopardize the metaphysical distinction between them, which is grounded in whatever facts constitute the identity and distinctness of selves. According to Locke, these facts are phenomenological; according to his opponents, the facts are biological. In both the cases, the metaphysical facts on offer appear to ground a distinction between first-personal and social relations, which are guaranteed to hold no matter how deep and pervasive the analogies between these two sorts of relations might be at the normative level.

Yet there is a way of thinking about the self that calls this into question. It is the very way of thinking that predominated in Locke's initial definition of the person as a being with "reason and reflection." When we emphasize this rational dimension of the self, we will find it natural to use a normative criterion in order to individuate selves. According to this normative criterion, we have one self wherever we find something that is subject to the normative requirements that define individual rationality. There is some disagreement among philosophers about exactly what these requirements are. But, even so, we can say this much with confidence: it is a sufficient condition for being a rational individual who is subject to the requirements of rationality that the individual have some conception of what those requirements are and is in some degree responsive to them – in the sense of at least trying to meet them when it deliberates and accepting criticism when it fails to meet them (Bilgrami, 1998). Of course, such criticism might issue from without as well as within, since a rational individual in this sense would certainly have the conceptual resources with which to comprehend efforts by others to wield rational influence over it. This ensures that there is a readily available public criterion by which such rational individuals can be identified for social purposes: there is one such individual wherever there is an interlocutor whom we can rationally engage. It must be admitted that, due to the analogies between first-personal and social relations, the boundaries between such individuals will not always be clear. Yet this does not mean that the normative conception of the self is inadequate or unworkable. It only means that the conception has unfamiliar and, indeed, revisionist consequences. For, if we press the analogies between first-personal and social relations to their logical limit, it follows that there can be group selves comprising many human beings and, also, multiple selves within a single human being (Rovane, 1998).

Take the group case first. We have seen that, for the purposes of joint activities, distinct selves must find areas of overlap between their individual points of view that can then serve as a common point of view from which they deliberate and act together. We have also seen that such joint deliberations replicate individual rationality in degree. Now, suppose that this were carried to the following extreme: a group of human beings resolved to engage only in joint activities; accordingly, they resolved to rank all of their preferences about such joint activities together and to pool all of their information; and finally, they resolved always to deliberate jointly in the light of their jointly ranked preferences and pooled information. The overlap between their points of view would then be so complete that they no longer had separate points of view at all. They would still have separate bodies and separate centers of consciousness. But, despite their biological and phenomenological separation, they would nevertheless all share a single rational point of view from which they deliberated and acted *as one*. Consequently, they could also be rationally engaged in the

ways that only individual selves can be engaged. There is good reason to regard such a group as a bona fide individual in its own right (Korsgaard, 1989; Rovane, 1998; Scruton, 1989; List and Pettit, 2011).

Now take the multiple case. Recall that it might be rational for an individual to disregard its future attitudes when it deliberates about what to do in the present, in just the way that it might be rational (though possibly immoral) to disregard the attitudes of others. If the normative demands of rationality are thus confined to one's own present attitudes, then one's future self will be no more bound by one's present deliberations than one is presently bound to take one's future attitudes into account. In other words, one's future self will be free to disregard any long-term intentions or commitments that one is thinking of undertaking now. The point is not that there are no such things as long-term intentions or commitments. The point is that when one makes them one does not thereby bind one's future self in any causal or metaphysical sense. It is always up to one's future self to comply or not as it wishes, and one could not alter this aspect of one's situation except by undermining one's future agency altogether. Although this may seem alarming, it does not follow that individual selves do not endure over time. All that follows is that the unity of an individual self over time is constituted by the same sorts of facts that constitute the unity of a group self. Just as the latter requires shared commitments to joint endeavors so too does the former, the only difference being that the latter involves sharing across bodies while the former involves sharing across time. This serves to bring out that the metaphysical facts in terms of which Locke and his opponents analyze the identity of the self over time are, in a sense, normatively impotent. The fact that there is a persisting center of consciousness, or a persisting animal body, does not suffice to bind the self together rationally over time. And, arguably (though this is more counter-intuitive), such facts do not even suffice to bind the self together rationally in the present. That is why we are able to understand the phenomenon of dissociative identity disorder, formerly known as multiple personality disorder. Human beings who suffer from this disorder exhibit more than one point of view, each of which can be rationally engaged on its own. Since none of these points of view encompasses all of the human host's attitudes, there is no possibility of rationally engaging the whole human being. At any given time one can engage only one part of it or another. Each of these parts functions more or less as a separate self, as is shown precisely by the susceptibility of each to rational engagement on its own. Although these multiple selves cannot be rationally engaged simultaneously, there is good reason to suppose that they nevertheless exist simultaneously. While one such self is talking, its companions are typically observing and thinking on their own (this generally comes out in subsequent conversation with them). Furthermore, these multiple selves who coexist simultaneously within the same body are sometimes 'co-conscious' as well – that is, they sometimes have direct phenomenological access to one another's thoughts. But their consciousness of one another's thoughts does not put them into the normative first-person relation that selves bear to their own thoughts; for they do not take one another's thoughts as a basis for their own deliberations and actions. Thus neither their shared body nor their shared consciousness suffices to bind multiple personalities together in the sense that is required by the normative conception of the self. On that conception, they qualify as individual selves in their own rights (Braude, 1995; Rovane, 1998; Wilkes, 1988).

Between these two extreme cases of group and multiple selves, the normative conception provides for a wide spectrum of cases, in which centers of rational unity are exhibited within and among human beings. The conception says that whenever there is the sort of rational unity that is characteristic of individual rationality, there is a rational being with its own point of view who can be rationally engaged as such, and who therefore qualifies as an individual self in its own right.

Implications for Methodological Individualism

It is not the case that all of the facts about the Columbia Philosophy Department can be reduced to facts about its members atomistically described. Nor is it the case that all of the facts about that department can be explained by appealing to facts about its members atomistically described. Why? Because sometimes the members of the department stop reasoning from their own separate human-size points of view and begin to reason together from the point of view of the department's distinctive projects, needs, opportunities, and so on. When the members of the department do this, they give over a portion of their human lives to the life of the department in a quite literal sense, so that the department itself can deliberate and act from a point of view that cannot be equated with any of their individual points of view. Social scientists might naturally regard these claims about the Columbia Philosophy Department as contradicting methodological individualism, but that is because they make the anti-Lockean assumption that the only individuals there are or can be are individual human beings. Given the assumption, they are bound to interpret the above claims about the Columbia Philosophy Department – that the facts about it cannot be properly described or explained by appealing exclusively to facts about its individual human members atomistically described – as pointing to irreducibly social phenomena. However, this interpretation fails to take account of the way in which the department instantiates the rational properties that are characteristic of the individual self (vs a social whole).

See also: Person and Self: Philosophical Aspects; Self: History of the Concept.

Bibliography

Bilgrami, A., 1998. Self-knowledge and resentment. In: Wright, C., et al. (Eds.), Knowing Our Own Minds. Oxford University Press, Oxford, UK.

Blumer, H., 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Braude, S.E., 1995. First Person Plural: Multiple Personality Disorder and the Philosophy of Mind. Rowman Littlefield, Lanham, MD.

Davidson, D., 1984. Inquires into Truth and Interpretation. Oxford University Press, New York.

Elster, J., 1979. Ulysses and the Sirens: Studies in Rationality and Irrationality. Cambridge University Press, New York.

Evans, G., 1982. Varieties of Reference. Oxford University Press, New York.
Gilbert, M., 1989. On Social Facts. Routledge, London.
Korsgaard, C., 1989. Personal identity and the unity of agency: a Kantian response to Parfit. Philosophy and Public Affairs 28 (2).
Levi, I., 1985. Hard Choices. Cambridge University Press, New York.
List, Christian, Pettit, Philip, 2011. Group Agency: The Possibility, Design and Status of Corporate Agents. Oxford University Press, Oxford.
Locke, J., 1979. In: Niddich, P. (Ed.), An Essay Concerning Human Understanding. Oxford University Press, Oxford, UK.
Mead, G.H., 1934. Mind, Self, and Society. University of Chicago Press, Chicago.
Nagel, T., 1970. The Possibility of Altruism. Clarendon, Oxford, UK.
Parfit, D., 1984. Reasons and Persons. Clarendon, Oxford, UK.
Perry, J. (Ed.), 1975. Personal Identity. University of California Press, Berkeley, CA.
Rorty, A.O. (Ed.), 1979. The Identities of Perons. University of California Press, Berkeley, CA.
Rovane, C., 1998. The Bounds of Agency: An Essay in Revisionary Metaphysics. Princeton University Press, Princeton, NJ.
Sartre, J.P., 1956. Being and Nothingness. Washington Square Press, New York.
Scruton, R., 1989. Corporate persons. Proceedings of the Aristotelian Society 63.
Sidgewick, H., 1981. The Methods of Ethics. Hackett, Indianapolis, IN.
Wilkes, K.V., 1988. Real People: Personal Identity without Thought Experiments. Oxford University Press, New York.
Wittgenstein, L., 1953. Philosophical Investigations. Macmillan, New York.

Self-Categorization Theory

S Alexander Haslam, School of Psychology, University of Queensland, St Lucia, QLD, Australia
Stephen D Reicher, School of Psychology, University of St. Andrews, St Andrews, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-categorization theory is a theory about the functioning of the self that provides three key insights into social behavior. First, social identity (a shared sense of 'us') is what makes group behavior possible. Second, the self system reflects a dynamic context-sensitive categorization process through which people see themselves as either sharing category membership with others or not. Third, shared self-category membership is the basis for mutual social influence. In these ways, shared social identity is understood as the basis for all forms of productive social interaction between people – in particular, those that entail cooperation, organization, and leadership.

Self-categorization theory "is a set of related assumptions and hypotheses about the functioning of the social self-concept (the concept of self based on comparison with other people and relevant social interaction)" (Turner, 1987: p. 42). The theory is a development of *social identity theory* – a theory which examines the way in which intergroup relations are structured by people's *social identity* (the sense of 'us' that reflects their self-categorization in terms of group membership; Tajfel and Turner, 1979). Together these two theories comprise the *social identity approach* (Turner, 2004; Reicher et al., 2010); however, where social identity theory is oriented to the explanation of a particular type of group behavior (intergroup discrimination), self-categorization theory is concerned with "how individuals are able to act as a group at all" (Turner, 1987: p. 42).

In this regard, one important shortcoming of social identity theory is that it offers only a limited analysis of the cognitive aspects of social identity. When do people categorize themselves in terms of group membership (as 'us' and 'we') rather than as individuals (as 'I' and 'me')? What leads to self-categorization in terms of one group membership rather than another? What is the nature of these social self-categorizations, and how does the self-categorization process transform a person's psychology and behavior? Despite the fact that the construct of social identity is obviously its core focus, social identity theory provides no real answers to questions like these. It was therefore to address such questions that self-categorization theory was developed by Turner and his colleagues in the 1980s (Turner, 1982, 1985; Turner et al., 1987, 1994). The result is a theory that has enormous and profound implications for psychology, primarily because it allows us to understand not only how individual minds are structured by social group memberships, but also how this structuring provides a basis for individuals to achieve otherwise unimaginable feats of collective organization and concerted action.

Depersonalization and Self-Stereotyping

Formative work on self-categorization theory focused on unpacking the cognitive substrates of social identity. In particular, Turner (1982) presented a more complete explanation of the process whereby people moved from seeing themselves and the world in interpersonal terms (e.g., 'me' interacting with 'you'), to seeing it in intergroup terms ('us' interacting with 'them'; as hypothesized by Tafjel's (1978), interpersonal–intergroup continuum). A key insight here is that it is a person's self-categorization in terms of social identity that actually *allows* intergroup behavior to take place. As Turner famously put it, "social identity is the cognitive mechanism that makes group behaviour possible" (p. 21; see also Haslam et al., 2003).

Turner went on to specify a *psychological process* associated with self-categorization in terms of social identity. Termed *depersonalization*, this refers to a process of *self-stereotyping* through which *the self comes to be perceived as categorically interchangeable with other in-group members*. Here individuals understand themselves and others not as differentiated individuals but as exemplars of a common in-group. This in turn plays a key role in regulating social behavior because it leads group members to act, not as unique individuals, but instead with reference to goals and norms that they share. This means, for example, that a person can act as a member of a given group (e.g., as a firefighter) only to the extent that they define themselves less as a unique individual (e.g., as someone who is married to Sam and likes woodwork) and more as someone who is categorically interchangeable with other firefighters and whose behavior is regulated by norms associated with that social category (e.g., to wear a uniform, to observe certain safety protocols, to be loyal to one's unit).

The Self-Categorization Process: Key Assumptions and Hypotheses

In suggesting that group behavior follows from an act of self-stereotyping, self-categorization theory points to the role that *categorization* – and more specifically *self*-categorization – plays in social perception and behavior. This was formalized in a number of core assumptions and related hypotheses of which five are key (Turner, 1985; Turner et al., 1987).

First, cognitive representations of the self are argued to take the form of *self-categorizations*. That is, the self is seen as a member of a particular class or category of stimuli. As such, it

is perceived to be (1) more or less equivalent to other stimuli in that category, and (2) more or less distinct from stimuli in other categories. So, for example, when a woman categorizes herself as a geographer she acknowledges her equivalence to other geographers and her difference from, say, sociologists or geologists.

Second, self- and other categories exist at different levels of abstraction with higher levels being more inclusive. Lower-level categories (e.g., geographer, geologist) can be subsumed within higher ones (e.g., earth scientist) and are defined in relation to comparisons at that higher level. For analytical purposes, the theory differentiates between three important levels of the social self-concept: self-categorization (1) at the superordinate *human* level as a human being (in contrast to other species), (2) at the intermediate *social* level as an in-group member (as distinct from out-groups), and (3) at the subordinate *personal* level as a unique individual (different from other relevant in-group members). Importantly, level of category abstraction is a relative concept – so more than one level of social self-category will typically be available for any given person. For example, our geographer may define herself in terms of social self-categories that are more or less abstract – as a scientist, as an earth scientist, or as a social geographer. Importantly, too, self-categories at all levels of abstraction are equally 'real' and just as much a reflection of a person's 'true' self. No one level of self-categorization is inherently more appropriate, useful, or valid than another and hence none is in any sense more fundamental to who or what the person is. This proposition challenges a general assumption in psychological theory which accords privileged status to personal identity – as if a person's *true* self is defined only by their individuality.

Third, the formation and salience (i.e., the cognitive activation) of any self-category is partly determined by comparisons between stimuli at a more inclusive level of abstraction. More specifically, the formation of self-categories is argued to be a function of the *metacontrast* between interclass and intraclass differences. This means that within a frame of reference comprised of various salient stimuli, the likelihood of any given collection of stimuli being perceived as a categorical entity increases to the extent that their difference from each other is perceived to be less than the difference between them and all other stimuli. So, for example, a geographer and a geologist are more likely to be seen – and see themselves – to share social identity as scientists when they are encountered in a context that includes nonscientists. This is because here the differences between them will be seen as relatively small compared to those between them and the nonscientists. In this way, metacontrast *contextualizes* categorization by tying it to in-the-moment judgments of *relative differences*. It also means that a person's categorical representation of exactly the same social stimuli will vary as a function of comparative context. Whether you and I see ourselves as 'us' will depend in large part on how different we see ourselves as being from the *other people* we are conscious of in a given setting.

Fourth, just as the metacontrast principle is a partial determinant of *which* categories perceivers use to represent a given stimulus array, so too it is a partial determinant of the *internal structure* of those categories. Following cognitive theorizing (e.g., Rosch, 1978) categories are assumed to have an internally *graded structure* such that some features of a category (i.e., particular behaviors, attributes, or individuals) define it better than others. This means that people differ in the extent to which they are perceived to be representative or *prototypical* of groups in the same way that a dog is generally more representative of the category 'pet' than a snake. More specifically, it follows from the metacontrast principle that any particular category member will be perceived to be more prototypical of a category to the extent that he or she is less different from other members of that category than from other social stimuli that are salient in a given context. Again, then, self-categorization theory predicts that the prototypicality of exactly the same exemplar for exactly the same category – and hence their capacity to represent what that category means and stands for – will vary as a function of the social context within which categorization takes place.

Finally, fifth, the salience of a categorization at a particular level of abstraction leads to the *accentuation* of perceived intraclass similarities and interclass differences between people as defined by their category membership at the same level. In this way, patterns of accentuation are reflections of people's categorical interchangeability. For example, if a woman's social self-category 'geographer' becomes salient, other geographers will be perceived to be more similar to each other (and her) and more different from other nongeographers (whose similarity to each other will also be accentuated) on dimensions that are seen to define membership of those categories (e.g., commitment to the scientific method).

Perceiver Readiness and Fit as Determinants of Social Identity Salience

To answer one of the questions we posed earlier, the above ideas allow us to understand exactly what factors dispose people to act in terms of a particular social self-categorization. When will an employee in an organization see and act in terms of the organization as a whole, or in terms of the department or team to which they belong, or as an individual? Answering such questions is extremely important, because it is apparent that people are capable of acting at *all* these levels, but that the *particular* level at which they define themselves has distinctive implications both for their own behavior and for the functioning of the organization as a whole.

To address this issue, the principles of self-categorization theory were formally applied to the analysis of social identity salience by (Oakes, 1987; Oakes et al., 1991; Turner, 1985). Following Bruner (e.g., 1957), it was argued that one crucial determinant of social category salience is *fit* – the degree to which a social categorization appears to be a sensible way of organizing and making sense of social stimuli (i.e., people and things associated with them). Fit has two components: comparative and normative.

Comparative fit is defined by the principle of metacontrast discussed above. This leads us to expect that a person will define themselves in terms of a particular self-category to the extent that the differences between members of that category on a given dimension of judgment are perceived to be

smaller than the differences between members of that category and others that are salient in a particular context. If our (female) geographer was surrounded by psychologists and other geographers, she would tend to define herself as a geographer only if the differences between the two groups appeared to be larger than the differences within them – which, for example, is more likely to be the case at a science conference than at a football match. *Normative fit* reflects the fact that in order to represent sets of people as members of a given category, the differences between them must not only appear smaller than the differences between them and others (comparative fit), but the *nature* of these differences must also be consistent with a perceiver's expectations about the categories. Our geographer at the science conference will be unlikely to classify participants as geographers and psychologists (or to act as a geographer herself) if the members of these two groups are seen to differ from each other in ways that are completely unexpected – perhaps if the geographers are concerned only with mental processes and the psychologists only with space.

Importantly too, the principles of fit also determine category salience in interaction with *perceiver readiness* (or *accessibility*; Oakes et al., 1994; Turner et al., 1994). Individuals do not engage in social encounters by mechanically processing information about similarities and differences from scratch in order to decide whether a person is one of us. Categorization also depends on the expectations, goals, and theories that perceivers *bring* to a given context – many of which derive from their previous group encounters. We thus organize and construe the world in ways that reflect the groups to which we belong and in this way *social histories* lend stability and predictability to our experience (Oakes et al., 1994; Reicher, 1996). *Identification* with a group – the extent to which the group is valued and self-involving and contributes to an enduring sense of self – is therefore one particularly important factor which affects a person's readiness to use a given social category in order to define themselves (e.g., Ellemers et al., 1997; Turner, 1999). Among other things, this means that when we identify strongly with a given group, we more readily interpret the world, and our place within it, in a manner consistent with that group's values, ideology, and culture. For this reason alone, social identification has proved to be an important construct in both social and organizational psychology and a number of researchers have developed scales that attempt to measure both its nature and strength.

Social Influence as a Determinant of Organized Social Behavior

Self-categorization theory's analysis of social identity salience is directed to the explanation of social–psychological phenomena that are often studied by looking at the responses and perceptions of noninteracting individuals. Significant as it is, such analysis therefore makes only a partial contribution to an understanding of social behavior that involves *structured social interaction*. Importantly though, the dynamic processes of self-categorization described above do not just affect the perception of individuals in the abstract. They are also assumed to have ongoing consequences for the *active coordination* of individuals' perception and behavior.

In this regard, a key assertion of self-categorization theory is that social self-categorizations serve to regulate individual cognitive activity not only by providing a shared perspective on social reality but also by providing a basis for mutual *social influence* (Turner, 1987, 1991). This is because when people perceive themselves to share category membership with another person in a given context they not only *expect to agree* with them on issues relevant to their shared identity but are also motivated to *strive actively to reach agreement* on those issues. Where only geographers and psychologists are present, a single geographer and a single psychologist may define themselves in terms of distinct social self-categories and expect to have different views, intentions, and goals. But if they meet in a context that also includes people from very different backgrounds (e.g., artists, politicians) they are more likely to redefine themselves in terms of a higher-order *shared* social self-categorization (e.g., as social scientists) that provides them with a relatively common perspective and which motivates them to coordinate that perspective further through discussion, persuasion, negotiation, and argument.

Self-categorization theory argues that social influence of this form is *necessary* because it is not possible for a person to establish the subjective validity and correctness of their beliefs simply through 'independent' activity. In this way, other members of the groups to which we see ourselves as belonging (those who contribute to our sense of 'we-ness' – i.e., in-group members) serve as essential reference points for our own perception. Moreover, it is precisely through individuals' identification of, and conformity to, norms that are perceived to be shared with others in a particular context that their potentially idiosyncratic views become socially organized and consensual. In other words, it is through this process that individual views are coordinated and transformed into *shared* values, beliefs, and behaviors. These values and beliefs also have particular force in the world (e.g., in the form of shared stereotypes) because they are no longer experienced as subjective but instead articulate a common, as-if-objective view (Haslam et al., 1998). In this way, what had been simply *personal opinion* and contingent belief now becomes *social fact* – so that, for example, '*I think* they are stupid and lazy' now becomes '*they are* stupid and lazy.'

Indeed, at the most general level, we can see that it is through these categorization-based processes of influence that low-level individual inputs are transformed into higher-order group products. These have emergent higher-order properties which ensure that the group whole is qualitatively different from ('more than') the sum of its individual parts. Moreover, it is also through this process that social identity becomes a world-making and world-changing resource.

Tests and Applications of Self-Categorization Theory

Within social psychology, early work on self-categorization theory elaborated upon the implications of depersonalization

for social perception and behavior. In particular, Turner et al.'s (1987) milestone text, *Rediscovering the Social Group*, provided novel and important insights into a broad range of social psychological topics and issues. Some of these – most obviously *group formation*, *cohesiveness*, and *stereotyping* – had been of long-standing interest to social identity theorists. However, the theory also provided a radically new analysis to a host of other topics, in particular, *group polarization* (e.g., McGarty et al., 1992) and *crowd behavior* (e.g., Reicher, 1987).

Subsequent strands of research confirmed the capacity for people's self- (and other) stereotypes to respond fluidly to changes in comparative context in the ways predicted by the metacontrast principle (e.g., Doosje et al., 1998) and pointed to the ability of these same contextual features to impact upon the internal structure of social categories. Other early work showed how patterns of stimulus accentuation serve to reflect the perceived interchangeability of category members (e.g., Haslam and Turner, 1992; Hogg and Turner, 1987) and indeed how they structure perceptions of the individual person (e.g., their *personality*; Reynolds et al., 2010).

The theory's analysis of social influence also fed into the analysis of an array of social and applied phenomena (especially *organizational behavior*; e.g., Haslam, 2004). In particular, it has been used to inform theorizing around the topics of cooperation, *communication*, *negotiation*, group decision making, and *stress*. However, probably the largest area of work in which the theory's ideas have been developed and tested relates to the study of *leadership* (e.g., Ellemers et al., 2004; Haslam et al., 2011; Hogg, 2001; Turner and Haslam, 2001). The key insight here is that leadership is not, as commonly supposed, a property that resides in a person's character *as an individual*. Rather it results from a leader's capacity to embody what a particular group means to its members in any given context. The capacity for leadership thus rests on individuals' ability to be seen as prototypical of a shared social identity and to exert influence over other group members on this basis. In other words, it is only when leaders are seen as embodying 'who we are' that their leadership stimulates followership.

As such work has developed, key elaborations of self-categorization principles have increasingly demonstrated that these processes of social categorization and influence do not simply occur within the heads of individuals, but also take place 'out there' in the social world. In this way, a great deal of research has explored the entrepreneurial and public dimensions of social identities as they are crafted by would-be leaders and contested in social interaction (e.g., Reicher and Hopkins, 2001). This work has served to elaborate and clarify the *dynamic* nature of self-categorization principles and the social identity approach more generally. It shows not only that (and how) self-categorization processes *respond* to social reality, but also that (and how) they are the means by which that reality is *transformed* – in particular, through the development and exercise of group-based *power* (Reicher, 1996; Turner, 2005). In this way, the theory speaks not only to the social dimensions of thinking and seeing but also to the political process of collective action and social change.

See also: Collective Behavior, Social Psychology of; Group Processes, Social Psychology of; Identity and Identification, Social Psychology of; Intergroup Relations; Leadership; Social Categorization; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Bruner, J.S., 1957. On perceptual readiness. Psychological Review 64, 123–152.

Doosje, B., Haslam, S.A., Spears, R., Oakes, P.J., Koomen, W., 1998. The effect of comparative context on central tendency and variability judgements and the evaluation of group characteristics. European Journal of Social Psychology 28, 173–184.

Ellemers, N., Spears, R., Doosje, B., 1997. Sticking together or falling apart: ingroup identification as a psychological determinant of group commitment versus individual mobility. Journal of Personality and Social Psychology 72, 617–626.

Ellemers, N., De Gilder, D., Haslam, S.A., 2004. Motivating individuals and groups at work: a social identity perspective on leadership and group performance. Academy of Management Review 29, 459–478.

Haslam, S.A., 2004. Psychology in Organizations: The Social Identity Approach, second ed. Sage, London.

Haslam, S.A., Postmes, T., Ellemers, N., 2003. More than a metaphor: organizational identity makes organizational life possible. British Journal of Management 14, 357–369.

Haslam, S.A., Reicher, S.D., Platow, M.J., 2011. The New Psychology of Leadership: Identity, Influence and Power. Psychology Press, London.

Haslam, S.A., Turner, J.C., 1992. Context-dependent variation in social stereotyping 2: the relationship between frame of reference, self-categorization and accentuation. European Journal of Social Psychology 22, 251–277.

Haslam, S.A., Turner, J.C., Oakes, P.J., McGarty, C., Reynolds, K.J., 1998. The group as a basis for emergent stereotype consensus. European Review of Social Psychology 8, 203–239.

Hogg, M.A., 2001. A social identity theory of leadership. Personality and Social Psychology Review 5, 184–200.

Hogg, M.A., Turner, J.C., 1987. Intergroup behaviour, self-stereotyping and the salience of social categories. British Journal of Social Psychology 26, 325–340.

McGarty, C., Turner, J.C., Hogg, M.A., David, B., Wetherell, M.S., 1992. Group polarization as conformity to the prototypical group member. British Journal of Social Psychology 31, 1–20.

Oakes, P.J., 1987. The salience of social categories. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, pp. 117–141.

Oakes, P.J., Turner, J.C., Haslam, S.A., 1991. Perceiving people as group members: the role of fit in the salience of social categorizations. British Journal of Social Psychology 30, 125–144.

Oakes, P.J., Haslam, S.A., Turner, J.C., 1994. Stereotyping and Social Reality. Blackwell, Oxford.

Reicher, S.D., 1987. Crowd behaviour as social action. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, pp. 171–202.

Reicher, S.D., 1996. Social identity and social change: rethinking the context of social psychology. In: Robinson, P. (Ed.), Social Groups and Identities: Developing the Legacy of Henri Tajfel. Butterworth–Heinemann, Oxford.

Reicher, S.D., Hopkins, N., 2001. Self and Nation: Categorization, Contestation and Mobilisation. Sage, London.

Reicher, S.D., Spears, R., Haslam, S.A., 2010. The social identity approach in social psychology. In: Wetherell, M.S., Mohanty, C.T. (Eds.), Sage Identities Handbook. Sage, London, pp. 45–62.

Reynolds, K.J., Turner, J.C., Branscombe, N.R., Mavor, K.I., Bizumic, B., Subašić, E., 2010. Interactionism in personality and social psychology: an integrated approach to understanding the mind and behaviour. European Journal of Personality 24, 458–482.

Rosch, E., 1978. Principles of categorization. In: Rosch, E., Lloyd, B.B. (Eds.), Cognition and Categorization. Erlbaum, Hillsdale, NJ, pp. 27–48.

Tafjel, H., 1978. Interindividual behaviour and intergroup behaviour. In: Tajfel, H. (Ed.), Differentiation between Social Groups: Studies in the Social Psychology of Intergroup Relations. Academic Press, London, pp. 27–60.

Tajfel, H., Turner, J.C., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

Turner, J.C., 1982. Towards a cognitive redefinition of the social group. In: Tajfel, H. (Ed.), Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, pp. 15–40.

Turner, J.C., 1985. Social categorization and the self-concept: a social cognitive theory of group behaviour. In: Lawler, E.J. (Ed.), Advances in group processes, vol. 2. JAI Press, Greenwich, CT, pp. 77–122.

Turner, J.C., 1987. A self-categorization theory. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, pp. 42–67.

Turner, J.C., 1991. Social Influence. Open University Press, Milton Keynes.

Turner, J.C., 1999. Some current issues in research on social identity and self-categorization theories. In: Ellemers, N., Spears, R., Doosje, B. (Eds.), Social Identity: Context, Commitment, Content. Blackwell, Oxford, pp. 6–34.

Turner, J.C., 2004. What the social identity approach is and why it matters. In: Foreword to Haslam, S.A. (Ed.), Psychology in Organizations: The Social Identity Approach, second ed. Sage, London, pp. xvii–xx.

Turner, J.C., 2005. Explaining the nature of power: a three-process theory. European Journal of Social Psychology 35, 1–22.

Turner, J.C., Haslam, S.A., 2001. Social identity, organizations and leadership. In: Turner, M.E. (Ed.), Groups at Work: Advances in Theory and Research. Erlbaum, Hillsdale, NJ, pp. 25–65.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford.

Turner, J.C., Oakes, P.J., 1986. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. British Journal of Social Psychology 25, 237–252.

Turner, J.C., Oakes, P.J., Haslam, S.A., McGarty, C.A., 1994. Self and collective: cognition and social context. Personality and Social Psychology Bulletin 20, 454–463.

Turner, J.C., Reynolds, K.J., 2011. Self-categorization theory. In: Van Lange, P., Kruglanski, A., Higgins, T. (Eds.), Handbook of Theories of Social Psychology. Sage, London, pp. 399–417.

Self-Concept: From Unidimensional to Multidimensional and Beyond

Brooke Van Zanden, Herbert W Marsh, Marjorie Seaton, and Philip Parker, Institute for Positive Psychology and Education, Australian Catholic University, Strathfield NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-concept is one of the oldest and most important constructs in the social sciences. It is also at the heart of the positive psychology revolution focusing on how healthy, normal, and exceptional individuals can get the most from life. We begin by distinguishing between an historical unidimensional perspective that focuses on self-esteem and a more recent multidimensional, hierarchical perspective that distinguishes between specific facets of self (e.g., academic, social, physical, and emotional). In this article we review developmental, educational, and personality perspectives of self-concept, gender differences, theoretical models, and empirical research on the reciprocal effects relating self-concept and performance, frame of reference effects based on social and dimensional comparisons that influence the formation of self-concept, and the juxtaposition between multidimensional perspectives of personality and self-concept.

Self-Concept: History and Significance

Self-concept is one of the oldest and most significant constructs in the social sciences. Highlighted by William James (1890/1963), self-concept today is viewed as a multidimensional, hierarchical construct with highly differentiated components such as emotional, physical, social, and academic self-concepts. Research has continually shown self-concept to be a critical construct in the attainment of academic achievement (Chen et al., 2013; Marsh and Craven, 2006; Pinxten et al., 2010), social and personal adjustment in children and adolescents (Harter, 1990), elite athlete success (Marsh and Perry, 2005), parent–adolescent relations (Barber et al., 2003), bullying (Marsh et al., 2004), approaches to learning (Burnett et al., 2003), and emotional adjustment and socialization (Donahue et al., 1993).

As such, enhancing self-concept is a major goal in many domains, including health, education, and the social services. According to Shavelson et al. (1976), self-concept is best thought of as a person's self-perceptions that are formed through experience and interpretations of one's environment, and incorporates feelings of self-worth, self-confidence, ability, and competence. Notably, self-concept is particularly susceptible to evaluations of significant others, reinforcements, and attributions for one's own behavior (Marsh, 2007). Additionally, not only is self-concept important as an outcome, but is also an important mediating factor that facilitates the attainment of other desirable outcomes (Shavelson et al., 1976).

The construct of self-concept has had a long and illustrious history, dating back to Socrates and Plato and to Bandura and Rogers in the present day (see Hattie, 1992). However, it can be argued that the current understanding of self-concept as a multidimensional construct can be traced back to William James (1890/1963). In his seminal work, The Principles of Psychology, James proposed the self as both multifaceted and hierarchical, "with the bodily self at the bottom, the spiritual self at the top, and the extracorporeal material selves and the various social selves between" (p. 313). This assertion, along with James' distinction between the self-as-knower (the I) and the self-as-known (the Me), played an important role in developing self-concept theory.

However, after the promising start by James, advances in self-concept theory, research, and measurement were slow and wrought with conceptual and methodological problems. Marsh (2007) proposed that the root of these problems may have been that terms such as self-concept were already so entrenched in the psychological vernacular and so widely used in everyday conversation that there was already an assumption that the paradigm was understood. Consequently, researchers may not have been motivated to explore theoretical definitions of the construct or to develop a systematic evaluation of self-concept measures (Marsh et al., 1988). This is reflected in the criticisms of pre-1980 self-concept research, which notes the dearth of sound measurement instruments, inherent methodological problems, and a pervasive inconsistency in reported results (e.g., Shavelson et al., 1976; Wylie, 1979).

Self-Concept: A Unidimensional or Multidimensional Construct?

Although James originally conceived self-concept as a multidimensional construct, there has been much debate on the value of unidimensional perspectives that emphasize a single, global domain of self-concept, often referred to as self-esteem, versus multidimensional perspectives based on multiple distinct components of self-concept (Marsh and Craven, 2006). Early self-concept research was generally dominated by a unidimensional perspective, where self-concept was represented by a single, general self-esteem score. However, this approach was criticized on the basis that a unidimensional perspective of self-concept does not allow researchers to capture the richness or specific insight that the various components of self-concept have to offer (Rosenberg et al., 1995).

It was not until Shavelson et al.'s (1976) seminal work that a multidimensional perspective again came to the forefront of self-concept research. After conducting an extensive review of

the self-concept literature, Shavelson et al. formed their own model of self-concept that was both multidimensional and hierarchical. Shavelson et al. posited a general self-concept encompassing all aspects of the self, which was then divided into academic and nonacademic self-concepts. At the next level, these self-concepts were further divided. For example, academic self-concept was further divided into self-concepts in specific subject areas (e.g., mathematics and English).

However, continued research suggested there were some weaknesses in Shavelson et al.' (1976) proposed model (see Byrne and Shavelson, 1996). This led Marsh and Shavelson (1985) to revise the model that provided a more highly differentiated view of self-concept than previous conceptions. For example, this model proposed higher-order academic factors of academic self-concept (e.g., mathematics and verbal) to explain self-concepts in specific school subjects, rather than the single dimension of academic self-concept initially proposed by Shavelson et al. Both the Shavelson et al. model and the Marsh–Shavelson revision are important, as they paved the way for the development of today's multidimensional self-concept instruments, which arguably formed the basis for considerable advances in the theoretical and methodological quality of self-concept research, as well as providing for stronger interventions (Craven et al., 1991; Marsh and Craven, 1997; Marsh and Shavelson, 1985).

The multidimensional perspective of self-concept has now been applied to a variety of settings, especially in the areas of education (Marsh, 1993) and sport (Fox and Corbin, 1989; Marsh, 2002). Importantly, the findings of all these studies have demonstrated the value of a multidimensional perspective over a global measure of self-esteem, especially in terms of predictive, convergent, and discriminant validity (Marsh, 2007). In addition, more recent research using a multidimensional perspective has proved valuable in the areas of bullying (Marsh et al., 2001b) and obesity and body image (Marsh et al., 2007).

Developmental Perspectives

Self-concept has been posited as a critical aspect of the social and emotional development of children by a number of researchers (Marsh et al., 2005a, 2002a). Thus, the issue of self-concept is a central concern to developmental psychologists. Researchers have been particularly interested in charting how self-concept develops with age. However, while it is clear that self-concept plays an important role in childhood development, there is less clarity regarding the age that children become able to report their self-concepts. Marsh et al. (2005a) argued that this has largely been due to a lack of appropriate instruments to adequately measure the self-concepts of young children.

The development of self-concept scales designed especially for young children have shed some light on how self-concept is experienced in early childhood. Using a version of the Self Description Questionnaire (SDQ) adapted for young children, Marsh et al. (2002a) demonstrated that preschool-aged children are able to distinguish between multiple dimensions of self-concept at a younger age than suggested by previous research. However, measures of verbal and mathematics self-concepts were highly correlated (0.73) in comparison to the typical small correlations between mathematics and verbal self-concepts evidenced in older children (Marsh, 1986; Marsh and Hau, 2004; Möller et al., 2009). These results suggest that very young children do not make the same clear differentiations between self-concepts that older children do.

The aforementioned findings give rise to the question of whether self-concept becomes increasingly multidimensional with age. Indeed, Shavelson et al. (1976) hypothesized that the multidimensional aspect of self-concept would increase with age. In other words, it was argued that as people get older, they become better able, and more likely, to distinguish their self-evaluations in different contexts. However, Marsh and Ayotte (2003) critiqued previous tests of this hypothesis, and found that increasing differentiation did not explain all results. Instead, Marsh and Ayotte proposed a differential distinctiveness hypothesis, which argued that with increasing age and cognitive development, there are counterbalancing processes of self-concept integration and differentiation. According to Marsh and Ayotte, integration occurs when closely related areas of self-concept (cognitive and affective components of mathematics self-concept) become amalgamated; differentiation refers to the increasing differentiation of disparate areas of self-concept (mathematics and verbal). Marsh and Ayotte supported their theoretical claims with data from young French–Canadian schoolchildren, which showed that with age dissimilar domains became more differentiated, while similar domains became more integrated, providing strong support for the differential distinctiveness hypothesis.

Researchers have also investigated how self-concept varies with age. Marsh (1989) described a curvilinear relationship between age and self-concept, whereby self-concept declines during preadolescence and early adolescence, and then levels out in middle adolescence, before increasing in late adolescence through to early adulthood. This curvilinear relationship was demonstrated by Cole et al. (2001). However, Cole et al.'s study also uncovered further complexity to the curvilinear relationship between self-concept and age. For instance, Cole and colleagues found that from middle to late childhood there were increases in academic, social, and sports competence, but an overall decrease in self-concept of behavioral conduct, and a female-only decrease in physical self-concept. At the onset of early adolescence, academic and sports self-concepts decreased significantly. Furthermore, following the transition from middle school, academic and social self-concepts recovered. Meanwhile, sports, behavior, and appearance self-concepts remained the same. Finally, the transition to high school was associated with increased self-perceived academic competence and increases in social self-concept for males and physical appearance self-concept for females. However, sports and behavioral self-concepts decreased for females at the same time. According to Cole et al., the period after the transition from high school is characterized by a plateau in self-concept scores across all five domains. This research supports Marsh's (1989) assertion that the transition from middle to high school presents the most turbulent time for self-conceptions. In addition, this time also displays more

discrepancies between male and female self-concepts than at any other time.

There are a host of theories that aim to explain the changes in self-concept that occur across the lifespan. According to Marsh and Craven (1997), children's self-concept decreases with age, as a result of increased exposure to situations that challenge high childhood self-concepts. Other researchers argue that the tumultuous transitions during puberty lead to radical decreases in self-perceptions (see Harter, 1998). Furthermore, others have posited that improved social skills, autonomy, and maturity may lead to increased self-concept (Hart et al., 1993). More research regarding the relationships between domain-specific self-concepts and age is needed to better understand the issue.

Gender Differences

Research focusing on the role of gender in self-concept has provided further support to the necessity of espousing a multidimensional model to describe self-concept. Speculation that global measures of the self are inadequate in explaining gender differences were first expressed by Wylie (1979), who argued that it was likely that gender differences in specific self-concept domains were masked by global scores of self-esteem. Instead, researchers (e.g., Dusek and Flaherty, 1981; Wylie, 1979) claimed that small differences in general self-esteem were actually reflective of counterbalancing gender differences in domain-specific self-concepts that were consistent with traditional sex stereotypes. There is a wealth of empirical evidence to support this argument. For example, in a review of gender differences in self-concept, Crain (1996) found that there was evidence of differences in domain-specific self-concepts for girls and boys that appeared to be gender stereotypic.

Other studies have further supported the notion that gender differences in self-perceptions vary across domain-specific self-concepts. Marsh's (1989) analysis of archive responses to the multidimensional SDQ instruments indicated that boys had distinctly more favorable global measures of self-concepts than their female counterparts. However, small but statistically significant, gender differences in most SDQ scales revealed that domain-specific self-concept scores typically followed sex stereotypes. For example, males had higher self-concepts in mathematics, while females fared better in measures of verbal self-concept.

This finding has been demonstrated in other studies. For example, Marsh and Ayotte (2003) replicated these findings using a French–Canadian version of the SDQ. Additionally, Lent et al. (1997) found that males exhibited higher mathematics self-concepts and higher general academic self-concepts than females. Finally, Feingold and Mazzella (1998) provided evidence for this trend in nonacademic domains of self-concepts, with the finding that male self-concepts relating to physical appearance were more positive than for their female counterparts. In summary, these studies that have analyzed gender and self-concept in a multidimensional framework are important as they provide further insight into the role gender plays in the formation of self-perceptions.

Causal Ordering of Self-Concept and Achievement: The Reciprocal Effects Model

Numerous studies have demonstrated a substantial positive correlation between academic self-concept and achievement (e.g., Marsh et al., 2006b,e). However, until recently, the causal ordering of this relation was less certain. However, this relation leaves unanswered the question of whether prior ratings of self-concept affect academic achievement, or whether academic achievement contributes to self-concept. This question is particularly important due to the theoretical and practical implications of the causal direction between these two variables. Essentially, knowing the causal ordering of self-concept and achievement is critical in determining the teaching practices that are most effective in enhancing student educational outcomes. Furthermore, answering the issue of causality is theoretically important in that much of the interest shown in self-concept stems from the belief that academic self-concept has motivational properties that contribute to achievement (Byrne, 1984).

Calysn and Kenny (1997) highlighted two opposing models relating to the causal ordering of self-concept and achievement: the self-enhancement and skill development models. The self-enhancement model posits that self-concept is a primary determinant of academic achievement and argues that interventions aiming to improve self-concept are critical in the attainment of academic success. In contrast, the skill development model argues that academic self-concept forms as a result of previous academic achievement. Thus, improving academic skills is seen as central to the enhancement of self-concept.

Until recent advances in statistical analyses, clarifying the issue of causal ordering between self-concept and academic achievement was problematic. However, the introduction of structural equation modeling into self-concept research reduced previous limitations, allowing researchers to investigate models that were theoretically and statistically more complex than the either–or arguments presented by self-enhancement and skill development models (see Marsh, 1990a; Marsh et al., 1999; Marsh and Martin, 2011). As a result of these advances, Marsh (1990a) proposed the reciprocal effects model (REM; see **Figure 1**) of academic self-concept. This model argued that the causal ordering of self-concept and achievement was not unidirectional. Instead, Marsh's model hypothesized a reciprocal relation between increases in self-concept and improvements in performance.

To test this model, Marsh (1990a) analyzed the large, nationally representative, US Youth in Transition Study. By analyzing the latent constructs of academic ability, academic self-concept, and school grades across a series of time waves, Marsh found support for a model where self-concept and achievement were reciprocally related. Marsh's study was important as it was one of the first studies that provided methodologically strong evidence for an REM, by demonstrating statistically significant pathways from prior self-concept to later achievement, and vice versa.

Cross-cultural research can inform us whether or not theories can be universally applied (Schwartz and Bilsky, 1990). Hence, the REM has been further bolstered by studies

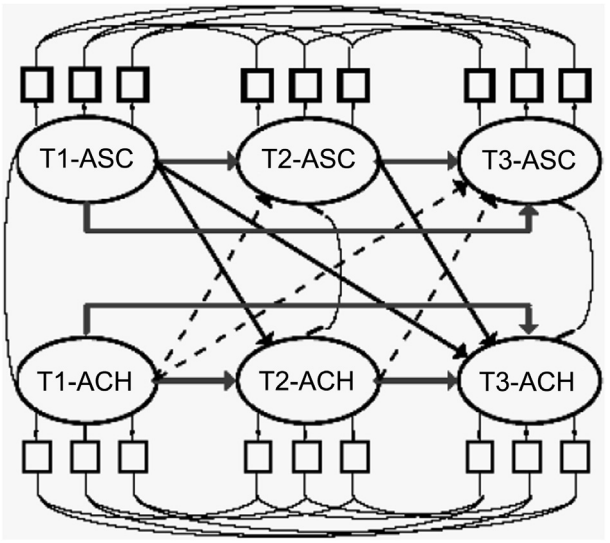

Figure 1 Prototype causal-ordering model for testing self-enhancement, skill development, and REM. In this full-forward, multi-wave, multivariable model, multiple indicators of academic self-concept (ASC) and achievement (ACH) are collected in three successive waves (T1, T2, and T3). Each latent construct (represented by oval shape) has paths leading to all latent constructs in subsequent waves. Within each wave, academic self-concept and achievement are assumed to be correlated; in the first wave, this correlation is a covariance between two latent constructs, and in subsequent waves, it is a covariance between residual factors. Curved lines at the top and bottom of the figure reflect correlated uniquenesses between responses to the same measured variable (represented by boxes) collected on different occasions. Paths connecting the same variable on multiple occasions reflect stability (the solid gray paths), but these coefficients typically differ from the corresponding test–retest correlations (which do not include the effects of other variables). Dashed lines reflect effects of prior achievement on subsequent self-concept, whereas solid black lines reflect the effects of prior self-concept on subsequent achievement. Adapted with permission from Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester, UK.

demonstrating good cross-cultural generalizability. Support for the REM in studies has come from different countries, for example, Hong Kong (Marsh et al., 2002b), Belgium (Pinxten et al., 2014), Germany (Marsh et al., 2001a, 2005b), and Canada (Guay et al., 2003).

Further strengthening the claims of the REM, evidence for reciprocal relations between self-concept and performance has been found outside the classroom. Marsh et al. (2006c) demonstrated the REM in a study investigating the causal ordering of physical self-concept and exercise behavior. Furthermore, Marsh et al. (2006a) tested the reciprocal effects, self-enhancement, and skill development models in relation to the physical self-concept and performance of gymnasts, and again found support for an REM of self-concept and performance. These findings are further supported by Marsh and Perry's (2005) study of self-concept and performance in a sample of elite swimmers in which prior self-concept was a significant positive predictor of subsequent championship performance.

Challenges to the REM

The REM provides evidence for the positive psychology movement. At the core of this movement is the belief that positive conceptions of the self are associated with enhanced life outcomes. Baumeister et al. (2003) challenged this premise in an influential review commissioned for Psychological Science in the Public Interest, arguing "that efforts to boost people's self-esteem are of little value in fostering academic achievement or preventing undesirable behavior" (p. 84).

However, in the context of recent advances in methodological and theoretical understandings of self-concept, Baumeister et al.'s (2003) conclusions are problematic. In particular, Baumeister et al. relied on a unidimensional perspective that emphasized self-esteem, largely ignoring research based on a multidimensional perspective reviewed here that focused on academic self-concept. From a multidimensional perspective it is reasonable that esteem has little or no relation with academic achievement even though academic self-concept and achievement are reciprocally related (Marsh and Craven, 2006). Marsh and O'Mara (2008) subsequently provided clear support for this theoretical claim, juxtaposing the negligible effects of self-esteem with the substantial effects of academic self-concept in a reanalysis of the classic Youth in Transition study highlighted by Baumeister et al. (in relation to self-esteem) and by Marsh and Craven (in relation to academic self-concept). This conclusion is also consistent with the meta-analytic research which found consistent support for a reciprocal relation between academic self-concept and achievement, but little to no reciprocal effect between achievement and self-esteem (Valentine and DuBois, 2005; Valentine et al., 2004).

In summary, the REM has been a critical development in self-concept theory. Firstly, the model established that positive self-concept and achievement are mutually reinforcing. Secondly, this finding supports the notion that positive self-concept is an integral part of success and achievement. Finally, these findings have important implications for educators. That is, since self-concept and achievement are mutually reinforcing and reciprocally related, interventions aimed at improving performance should not only strive to promote skill development, but also seek to simultaneously enhance self-concept to encourage achievement.

Frames of Reference Models

Since William James (1890), psychologists have stressed that self-concept cannot be understood without recognizing the role that comparative processes play in perceptions of the self. Essentially, while people may share the same objective characteristics or accomplishments, differing frames of references or standards of comparisons can lead to disparate self-concepts among individuals.

Social Comparison and the Big-Fish–Little-Pond Effect

Integrating research based on a variety of theoretical models and empirical studies (e.g., Davis, 1966; Festinger, 1954), Marsh (1984; also see Marsh et al., 2008) proposed the

big-fish–little-pond effect (BFLPE; see Figure 2) that is specific to educational settings. According to the BFLPE, students compare their own academic abilities with the abilities of their classmates, and use this social comparison as the basis of their academic self-concept. In the BFLPE, students who attend high-ability classes and schools tend to have lower academic self-concepts than equally able students who attend mixed- or low-ability classes and schools. Thus, the BFLPE explains how students with equal ability can have differing academic self-concepts as a result of the educational settings they attend.

Since the initial BFLPE study (Marsh and Parker, 1984), there has been a wealth of support for BFLPE predictions based on studies that use differing experimental and analytical approaches (Marsh, 1987; Marsh and Craven, 1994, 1997; see review by Marsh et al., 2008). Furthermore, Marsh (1991) demonstrated that students attending higher-ability high schools were likely to not only suffer from depleted academic self-concepts, but were also likely to have lower GPAs, lower educational aspirations, lower occupational aspirations, lower standardized test scores, and were also more likely to select less demanding coursework than their equally abled peers attending schools with lower average abilities. These findings are significant as they have important implications for parents, teachers, and policy makers, as they dispute the common held belief that it is advantageous to send students to schools with stellar academic reputations. Instead, Marsh argues that the findings of BFLPE indicate that many students attending high-average schools are not reaching their full academic potential.

Further support for the BFLPE was also found in the quasi-experimental setting that occurred after the fall of the Berlin Wall (Marsh et al., 2001a). Essentially, the fall provided researchers the opportunity to compare the effects of attending school systems that differed in the extent that they segregated students by ability. While West German students had previously attended academically differentiated schools, East German students had not been exposed to an academically differentiated school system. Results of the study supported the BFLPE predictions. The BFLPE was significantly larger for West German students at the start of reunification of the schooling systems. Importantly, as time passed the difference in the size of the effect between East and German students was reduced, and eventually disappeared after the East German students had been exposed to the West German schooling system for a year. According to Marsh, Köller, et al. these findings are a testament to how national educational policy differences impact upon the academic self-concepts of individual students.

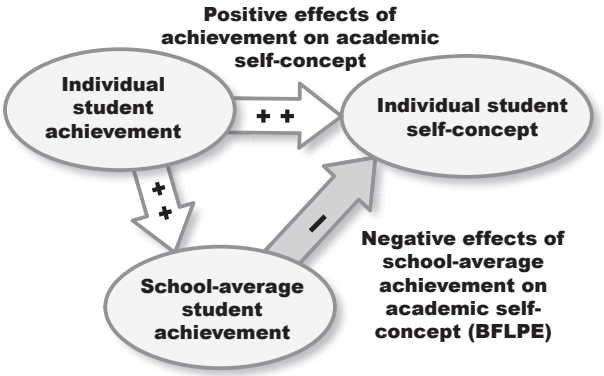

Figure 2 Big-fish–little-pond effect. Reproduced from Marsh, H.W., Hau, K.T., 2003. Big-fish–little-pond effect on academic self-concept: a cross-cultural (26 country) test of the negative effects of academically selective schools. American Psychologist, 58, 364.

In research reviewed by Marsh et al. (2008), there is consistent cross-cultural support for the BFLPE based on studies from many different countries. Demonstrating that the BFLPE is one of psychology's most cross-culturally universal phenomenon, three successive Program for International Student Assessment (PISA) data collections (Marsh and Hau, 2003; 103 558 students from 26 countries; Seaton et al., 2009, 2010; 265 180 students from 41 countries; Nagengast and Marsh, 2012; 397 500 students from 57 countries) showed that the effect of school-average ACH on ASC was negative in all but one of the 123 samples and significantly so in 114 samples.

Seaton et al. (2010) evaluated whether the BFLPE was moderated by 17 individual student characteristics (student background, learning styles, and the perceived learning environment) based on the large PISA 2003 database. However, they concluded "that the BFLPE was an extremely robust effect" (p. 390). Jonkmann et al. (2012) evaluated whether the BFLPE was moderated by personality factors (big-five traits and narcissism). They found that students high in narcissism had exaggerated academic self-concept and smaller BFLPEs. In contrast, students high in neuroticism experienced more negative BFLPEs. However, these moderation effects were modest in that they did not change the direct of the BFLPE, again supporting the robustness of the BFLPE.

While most BFLPE studies use high school student samples, the effect has been demonstrated in primary school' gifted and talented classes (Marsh et al., 1995). Therefore, the assumed benefits of gifted and talented programs are also called into question by the BFLPE. In addition to this, the BFLPE has found support among academically disadvantaged students placed in mainstream classes (Tracey et al., 2003; Marsh et al., 2006d). The findings of these studies were in contrast to labeling theory, which predicts that placement of academically disadvantaged children in special classes will lead to lower self-concepts and stigmatization. Instead, results indicated that academically disadvantaged students placed in regular classes in fact had lower self-concepts than their equally able peers who were grouped in classes with other students of similar abilities. These findings are important as they indicate that the very policies aiming to include students are at the same time having detrimental effects on students' academic self-concept.

Dimensional Comparisons and the I/E Model

Researchers have also brought attention to the role that internal comparisons play in shaping self-concept. For example, Albert (1977) proposed that individuals also undertake temporal comparisons in order to form self-evaluations. Temporal comparisons are performed within an internal frame of reference, whereby individuals compare themselves across time. Furthermore, Möller and Marsh (2013) described another internal comparative process known as dimensional

comparison, whereby individuals compare their own abilities in one area to another. For example, a student may ask, "How good am I in mathematics compared to English?"

These dimensional comparisons are the focus of the 'Classic' Internal/External Frame of Reference (I/E) model, which investigates the association between academic self-concept and achievement. In accordance with the hierarchical aspect of self-concept, correlations between mathematics and verbal self-concepts are typically weak or even negative (Marsh, 1986; see **Figure 3**). Furthermore, Marsh demonstrated that pathways between domain-specific achievement and the corresponding self-concept were substantial and positive. In contrast, paths from mathematics achievement to verbal self-concept and vice versa were significant and negative. Essentially, people tend to see themselves as either a mathematics or a verbal person, but rarely both. What makes this finding interesting is that although mathematics and verbal self-concepts are uncorrelated, objective measures of academic performance indicate that mathematics and verbal abilities are usually similar (Marsh, 1986). Therefore, there is a paradoxical relationship that exists between academic self-concepts and their corresponding measures of academic performance.

The I/E model was developed to explain the aforementioned paradoxical relationship between subject-specific self-concepts and achievement (Marsh, 1986). Marsh argues that academic self-concepts in particular school subjects are formed through, not only external references, but also internal references such as dimensional comparisons. These occur when students compare their own performance in one subject to their own performance in another school subject. According to the model, good verbal skills lead to lower mathematics self-concept as a result of these dimensional comparison processes, and vice versa.

The I/E model has been widely supported across a number of studies differing in sample nationality, age, and in the uses of self-concept and measures of achievement (see Marsh, 1990b; Marsh and Craven, 1997). Meta-analyses have demonstrated further support for the model, again showing

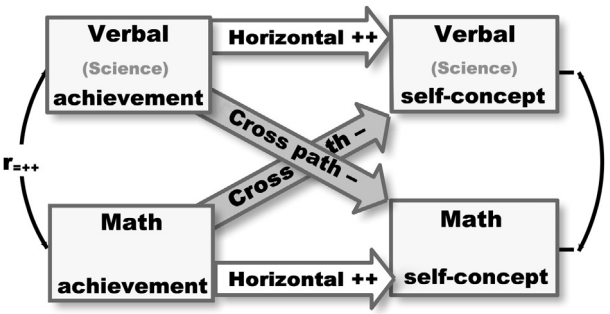

Figure 3 The 'Classic' Internal/External Frame of Reference (I/E) Model relating verbal and mathematical achievement to verbal and mathematical self-concepts. According to predictions from the I/E model, the horizontal paths from achievement to self-concept in the matching domain (content area) are predicted to be substantial and positive (++), whereas the cross paths from achievement in one domain area to self-concept in a non-matching domain (shaded in gray above) are predicted to be negative (−).

generalizability across age, gender, and nationality (Möller et al., 2009; Marsh and Hau, 2004). There has also been experimental research that has enabled confirmation of the causal hypotheses of the I/E model (Möller and Köller, 2001). In this study Möller and Köller were able to demonstrate that manipulation of feedback on achievement in one subject area had an inverse effect on self-concept in the subject at the opposite end of the verbal–mathematics continuum. Furthermore, diary studies have also confirmed that students spontaneously undertake dimensional comparisons on a day-to-day basis. Importantly, these dimensional comparisons have been shown to predict post-school education and career pathways (Parker et al., 2012; Parker et al., in pressa,b).

The I/E model has been an important development in self-concept theory as it highlighted the hierarchical aspect of self-concept, leading to the Marsh/Shavelson revision of the original Shavelson model of self-concept. This theoretical development has been important in the practice of educational psychology, as it has highlighted the fact that even academic self-concept is highly domain specific. This means that psychologists should take careful consideration of the way they choose to measure academic self-concept to ensure that the level of specificity is appropriate to research aims and intervention goals.

Personality and Self-Concept

Personality researchers make a distinction between core personality traits such as the Big Five (neuroticism, extraversion, conscientiousness, agreeableness, and openness) and self-concept, which they perceive to be a more malleable personality characteristic. Although both self-concept and personality researchers agree that their constructs should be viewed from a multidimensional perspective, Marsh et al. (2006e; see also Marsh, 2008) noted that personality researchers have relied heavily on global self-esteem measures rather than on multidimensional self-concept measures that are more commonly used by self-concept researchers. Moreover, these authors noted that there is a paucity of research combining personality and self-concept from a multidimensional perspective.

Research that has examined personality and global self-esteem has shown that the latter to be negatively related to neuroticism (Judge et al., 1998; also see Francis and James, 1996; Furr and Funder, 1998; Watson et al., 2002) and somewhat less positively related to extraversion (e.g., Francis and James, 1996; Furr and Funder, 1998; Watson et al., 2002). The few studies that have examined relations between self-esteem and the complete set of Big Five factors have found positive correlations between self-esteem and conscientiousness, openness, and agreeableness (see Watson et al., 2002).

In response to the limited amount of studies relating multidimensional personality factors with multidimensional self-concept factors, Marsh et al. (2006e) evaluated relations between 17 clearly defined self-concept factors (extended SDQIII), personality (Big Five), well-being (positive and negative affect and life satisfaction), and academic outcomes

(school grades, test scores, and coursework selection) in young adult students (17–19 years). They found that substantial amounts of variance in each of the Big Five and well-being factors could be explained by specific self-concept factors, varying between 23% and 60% (M = 39%) for the Big Five factors and between 14% and 19% (M = 17%) for the well-being factors. Quite the reverse was found for self-esteem; almost none of the variance in any of the Big Five or the well-being factors could be uniquely explained by self-esteem. That the pattern of relations between self-concept and personality factors was so highly differentiated and multivariate suggests that personality researchers should consider using multidimensional self-concept measures in place of the commonly used unidimensional ones.

Marsh et al. (2006e) found support for convergent validity (self-concept and personality factors that were logically related were substantially correlated) and divergent validity (self-concept and personality factors that were not logically related were not substantially related), most especially in seven higher-order factors encompassing both self-concept and personality factors. Moreover, personality was confirmed as a multidimensional construct, as each Big Five personality factor loaded primarily on one and only one of the seven higher-order factors. Furthermore, specific self-concept components and Big Five personality factors that one would logically associate with each other loaded on the same higher-order factor. Moreover, two higher-order factors (reflecting the quantitative-academic components of self-concept and religious self-concept) were relatively unrelated to all personality factors. Self-esteem relations confirmed the multidimensional nature of self-concept. Not only did self-esteem load on just one of the seven higher-order factors, but also was not even the highest loading of all the self-concept factors. Further demonstrating the multidimensional perspective of self-concept, the association between self-esteem and each of the nine nonacademic outcome measures was modest (rs varied from −0.03 to 0.05). Conversely, the relations between academic outcomes and the academic components of self-concept were highly differentiated and substantial. It is interesting to note that there was only a moderate association between the Big Five personality and well-being factors and the academic outcomes. Moreover, these personality constructs were unable to additionally contribute to the prediction of the academic outcomes over and above that which could be explained by the academic components of self-concept.

In summary, Marsh et al. (2006e) have provided both personality and self-concept research with a way forward by demonstrating how multidimensional personality constructs and multidimensional self-concept constructs relate to each other. In doing so, they have begun the process of integrating two very important and productive strands of social psychology.

Summary

The self-concept construct has had a long and distinguished history. Although once conceptualized as unidimensional, research over the last 35 years has demonstrated that self-concept is indeed a multidimensional construct. As Marsh and Hattie (1996) aptly declared, "there appears to be no support at all for the unidimensional perspective of self-concept or, apparently, even a unidimensional perspective of academic self-concept" (p. 44). Indeed, self-concept research has blossomed in this period, spanning issues as varied as developmental perspectives, to gender differences, effects of differing frames of reference on self-concept, self-concept's association with personality, and the reciprocal relation that it has with achievement. These advances in self-concept are exciting and augur well for the future of the discipline. Perhaps more importantly they demonstrate how crucial a positive self-concept is in many areas of human functioning and endorse Marsh and Craven's (2006) description of self-concept as a "hot variable that makes good things happen, facilitating the realization of full human potential in a range of settings" (p. 134).

See also: Academic Achievement Motivation, Development of; Concept Formation in Childhood: Cross-Cultural Perspectives; Factor Analysis and Latent Variable Models in Personality Psychology; Gender Differences in Personality and Social Behavior; Internal/External Frame of Reference Model; Motivation, Learning, and Instruction; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self and Emotional Development in Adulthood and Later Life; Self-Esteem; Self-Regulated Learning.

Bibliography

Albert, S.U., 1977. Temporal comparison theory. Psychological Review 84, 485–503.

Barber, C.N., Hall, J., Armistead, L., 2003. Parent-adolescent relationship and adolescent. psychological functioning among African-American female adolescents: self-esteem as a mediator. Journal of Child & Family Studies 12, 361–374.

Baumeister, R.F., Campbell, J.D., Krueger, J.I., Vohs, K.D., 2003. Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? Psychological Science in the Public Interest 4, 1–44.

Burnett, P.C., Pillay, H., Dart, B.C., 2003. The influences of conceptions of learning and learner self-concept on high school students' approaches to learning. School of Psychology International 24, 54–66.

Byrne, B., 1984. The general/academic self-concept nomological network: a review of construct validation research. Review of Educational Research 54, 427–456.

Byrne, B.M., Shavelson, R.J., 1996. On the structure of social self-concept for pre-, early, and late adolescents: a test of the Shavelson, Hubner, and Stanton (1976) model. Journal of Personality and Social Psychology 70, 599.

Calysn, R., Kenny, D., 1997. Self-concept of ability and perceived evaluations by others: cause or effect of academic achievement? Journal of Educational Psychology 69, 136–145.

Chen, S., Yeh, Y., Hwang, F., Lin, S.S.J., 2013. The relationship between academic self-concept and achievement: a multicohort–multioccasion study. Learning and Individual Differences 23, 172–178.

Cole, D.A., Maxwell, S.E., Martin, J.M., Peeke, L.G., Seroczynski, A.D., Tram, J.M., Maschman, T., 2001. The development of multiple domains of child and adolescent self-concept: a cohort sequential longitudinal design. Child development 72, 1723–1746.

Crain, R.M., 1996. The influence of age, race, and gender on child and adolescent multidimensional self-concept. In: Bracken, B.A. (Ed.), Handbook of Self-Concept. John Wiley & Sons, Oxford, pp. 395–420.

Craven, R.G., Marsh, H.W., Debus, R.L., 1991. Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. Journal of Educational Psychology 83, 17.

Davis, J.A., 1966. The campus as a frog pond: an application of theory or relative deprivation to career decisions for college men. American Journal of Sociology 72, 17–31.

Donahue, E.M., Robins, R.W., Roberts, B.W., John, O.P., 1993. The divided self: concurrent and longitudinal effects of psychological adjustment and social roles on self-concept differentiation. Journal of Personality and Social Psychology 64, 834–846.

Dusek, J.B., Flaherty, J.F., 1981. The development of self-concept during the adolescent years. Society for Research in Child Development 46 (Serial No. 191).

Feingold, A., Mazzella, R., 1998. Gender differences in body image are increasing. Psychological Science 9, 190–195.

Festinger, L., 1954. A theory of social comparison processes. Human relations 7, 117–140.

Fox, K.R., Corbin, C.B., 1989. The physical self-perception profile: development and preliminary validation. Journal of Sport and Exercise Psychology 11, 408–430.

Francis, L.J., James, D.J., 1996. The relationship between Rosenberg's construct of self-esteem and Eysenck's two-dimensional model of personality. Personality and Individual Differences 21, 483–488.

Furr, M.R., Funder, D.C., 1998. A multimodal analysis of personal negativity. Journal of Personality and Social Psychology 74, 1580.

Guay, F., Marsh, H.W., Boivin, M., 2003. Academic self-concept and academic achievement: developmental perspectives on their causal ordering. Journal of Educational Psychology 95, 124–136.

Hart, D., Fegley, S., Brengelman, D., 1993. Perceptions of past, present and future selves among children and adolescents. British Journal of Developmental Psychology 11, 265–282.

Harter, S., 1990. Processes underlying adolescent self-concept formation. In: Montemayor, R., Adams, G., Gulotta, T. (Eds.), From Childhood to Adolescence. Sage Publications, California.

Harter, S., 1998. The development of self-representations. In: Damon, W., Eisenberg, N. (Eds.), Handbook of Child Psychology, fifth ed. John Wiley & Sons, Hoboken, NJ.

Hattie, J., 1992. Self-Concept. Erlbaum, Hillsdale, NJ.

James, W., 1890/1963. The Principles of Psychology. Holt, Rinehart & Winston, New York.

Judge, T.A., Erez, A., Bono, J.E., 1998. The power of being positive: the relation between positive self-concept and job performance. Human Performance 11, 167–187.

Jonkmann, K., Becker, M., Marsh, H.W., Lüdtke, O., Trautwein, U., 2012. Personality traits moderate the big-fish–little-pond effect of academic self-concept. Learning and Individual Differences 22, 736–746.

Lent, R.W., Brown, S.D., Gore Jr., P.A., 1997. Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. Journal of Counseling Psychology 44, 307–315.

Marsh, H.W., 1984. Self-concept: the application of a frame of reference model to explain paradoxical results. Australian Journal of Education 28, 165–181.

Marsh, H.W., 1986. Verbal and math self-concepts: an internal/external frame of reference model. American Educational Research Journal 23, 129–149.

Marsh, H.W., 1987. The big-fish–little-pond effect on academic self-concept. Journal of Educational Psychology 79 (3), 280–295.

Marsh, H.W., 1989. Age and sex effects in multiple dimensions of self-concept: preadolescence to early adulthood. Journal of Educational Psychology 81, 417–430.

Marsh, H.W., 1990a. The causal ordering of academic self-concept and academic achievement: a multiwave, longitudinal panel analysis. Journal of Educational Psychology 82, 646–656.

Marsh, H.W., 1990b. A multidimensional, hierarchical self-concept: theoretical and empirical justification. Educational Psychology Review 2, 77–172.

Marsh, H.W., 1991. The failure of high ability schools to deliver academic benefits: the importance of academic self-concept and educational aspirations. American Educational Research Journal 28, 445–480.

Marsh, H.W., 1993. Academic self-concept: theory, measurement and research. In: Suls, J. (Ed.), Psychological Perspectives on the Self. Erlbaum, Hillsdale, NJ, pp. 59–98.

Marsh, H.W., 2002. A multidimensional physical self-concept: a construct validity approach to theory, measurement and research. Psychology: The Journal of the Hellenic Psychological Society 9, 459–493.

Marsh, H.W., 2007. Self-concept Theory, Measurement and Research into Practice: The Role of Self-concept in Educational Psychology. British Psychological Society, Leicester.

Marsh, H.W., 2008. A multidimensional, hierarchical model of self-concept: an important facet of personality. In: The Sage Handbook of Personality Theory and Assessment. Sage, Los Angeles, CA, pp. 447–469.

Marsh, H.W., Ayotte, V., 2003. Do multiple dimensions of self-concept become more differentiated with age? The differential distinctiveness hypothesis. Journal of Educational Psychology 95, 687–706.

Marsh, H.W., Byrne, B.M., Shavelson, R.J., 1988. A multifaceted academic self-concept: its hierarchical structure and its relation to academic achievement. Journal of Educational Psychology 80, 366–380.

Marsh, H.W., Byrne, B.M., Yeung, A.S., 1999. Causal ordering of academic self-concept and achievement: reanalysis of a pioneering study and revised recommendations. Educational Psychologist 34, 155–167.

Marsh, H.W., Chanal, J.P., Sarrazin, P.G., 2006a. Self-belief does make a difference: a reciprocal effects model of the causal ordering of physical self-concept and gymnastics performance. Journal of Sports Sciences 24, 101–111.

Marsh, H.W., Chessor, D., Craven, R.G., Roche, L., 1995. The effects of gifted-and-talented programmes on academic self-concept: the big fish strikes again. American Educational Research Journal 32, 285–319.

Marsh, H.W., Craven, R.G., 1994. Who benefits from selective schools? The role of academic self-concept and a call for further research. Forum of Education 49, 1.

Marsh, H.W., Craven, R.G., 1997. Academic self-concept: beyond the dustbowl. In: Phye, G. (Ed.), Handbook of Classroom Assessment: Learning, Achievement and Adjustment. Academic Press, San Diego, pp. 131–193.

Marsh, H.W., Craven, R.G., 2006. Reciprocal effects of self-concept and performance from a multidimensional perspective: beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science 1, 133–163.

Marsh, H.W., Debus, R., Bornholt, L., 2005a. Validating young children's self-concept responses: methodological ways and means to understand their responses. In: Teti, D.M. (Ed.), Handbook of Research Methods in Developmental Science. Blackwell Publishers, Oxford, pp. 138–160.

Marsh, H.W., Ellis, L.A., Craven, R.G., 2002a. How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. Developmental Psychology 38, 376–393.

Marsh, H.W., Hattie, J.A., 1996. Theoretical perspectives on the structure of self-concept. In: Bracken, B.A. (Ed.), Handbook of Self-Concept: Developmental, Social and Clinical Considerations. Wiley, Oxford, pp. 38–90.

Marsh, H.W., Hau, K.T., 2003. Big-fish–little-pond effect on academic self-concept: a cross-cultural (26 country) test of the negative effects of academically selective schools. American Psychologist 58, 364.

Marsh, H.W., Hau, K.T., 2004. Explaining paradoxical relations between academic self-concepts and achievements: cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. Journal of Educational Psychology 96, 56–67.

Marsh, H.W., Hau, K.T., Artelt, C., Baumert, J., Peschar, J.L., 2006b. OECD's brief self-report measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. International Journal of Testing 6, 311–360.

Marsh, H.W., Hau, K.T., Kong, C.K., 2002b. Multilevel causal ordering of academic self-concept and achievement: influence of language of instruction (English compared with Chinese) for Hong Kong students. American Educational Research Journal 39, 727–763.

Marsh, H.W., Hau, K.T., Sung, R.Y.T., Yu, C.W., 2007. Childhood obesity, gender, actual-ideal body image discrepancies, and physical self-concept in Hong Kong children: cultural differences in the value of moderation. Developmental Psychology 43, 647.

Marsh, H.W., Köller, O., Baumert, J., 2001a. Reunification of East and West German school systems: longitudinal multilevel modeling study of the big-fish–little-pond effect on academic self-concept. American Educational Research Journal 38 (2), 321–350.

Marsh, H.W., Martin, A.J., 2011. Academic self-concept and academic achievement: relations and causal ordering. British Journal of Educational Psychology 81, 59–77.

Marsh, H.W., O'Mara, A., 2008. Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: unidimensional and multidimensional perspectives of self-concept. Personality and Social Psychology Bulletin 34, 542–552.

Marsh, H.W., Papaionannou, A., Theodorakis, Y., 2006c. Causal ordering of physical self-concept and exercise behavior: reciprocal effects model and the influence of physical education teachers. Health Psychology 25, 316–328.

Marsh, H.W., Parada, R.H., Craven, R.G., Finger, L., 2004. In the looking glass: a reciprocal effects model elucidating the complex nature of bullying, psychological determinants and the central role of self-concept. In: Sanders, C.S., Phye, G.D. (Eds.), Bullying: Implications for the Classroom. Academic Press, Orlando.

Marsh, H.W., Parada, R.H., Yeung, A.S., Healey, J., 2001b. Aggressive school troublemakers and victims: a longitudinal model examining the pivotal role of self-concept. Journal of Educational Psychology 93, 411–419.

Marsh, H.W., Parker, J., 1984. Determinants of student self-concept: is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? Journal of Personality and Social Psychology 47, 213–231.

Marsh, H.W., Perry, C., 2005. Self-concept contributes to winning gold medals: causal ordering of self-concept and elite swimming performance. Journal of Sport & Exercise Psychology 27, 71–91.

Marsh, H.W., Seaton, M., Trautwein, U., Ludtke, O., Hau, K.T., O'Mara, A.J., Craven, R.G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research. Educational Psychology Review 20, 319–350.

Marsh, H.W., Shavelson, R., 1985. Self-concept: its multifaceted, hierarchical structure. Educational Psychologist 20, 107–125.

Marsh, H.W., Tracey, D.K., Craven, R.G., 2006d. Multidimensional self-concept structure for pre-adolescents with mild intellectual disabilities: a hybrid multigroup-mimic approach to factorial invariance and latent mean differences. Educational and Psychological Measurement 66, 795–818.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O., Baumert, J., 2006e. Integration of multidimensional self-concept and core personality constructs: construct validation and relations to well-being and achievement. Journal of Personality 74, 403–456.

Marsh, H.W., Trautwein, U., Ludtke, O., Koller, O., Baumert, J., 2005b. Academic self-concept, interest, grades and standardised test scores: reciprocal effects models of causal ordering. Child Development 76, 297–416.

Möller, J., Köller, O., 2001. Dimensional comparisons: an experimental approach to the internal/external frames of reference model. Journal of Educational Psychology 93, 826–835.

Möller, J., Marsh, H.W., 2013. Dimensional comparison theory. Psychological Review 1 no pagination specified.

Möller, J., Pohlmann, B., Köller, O., Marsh, H.W., 2009. A meta-analytic path analysis of the internal/extern frame of reference model of academic achievement and academic self-concept. Review of Educational Research 79, 1129–1167.

Nagengast, B., Marsh, H.W., 2012. Big fish in little ponds aspire more: mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. Journal of Educational Psychology 104, 1033.

Parker, P.D., Schoon, I., Tsai, Y., Nagy, G., Trautwein, U., Eccles, J., 2012. Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: a multi-context study. Developmental Psychology 48, 1629–1642.

Parker, P.D., Marsh, H.W., Ciarrochi, J., Marshall, S., Abduljabbar, A.S. Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. Educational Psychology, in press-a.

Parker, P.D., Nagy, P.D., Trautwein, U., Lüdtke, O. The internal/external frame of reference as predictors of career aspirations and university majors. In: Eccles, I., Schoon, I. (Eds.), Gender Differences in Aspirations and Attainment, in press-b.

Pinxten, M., De Fraine, B., Damme, J., D'Haenens, E., 2010. Causal ordering of academic self-concept and achievement: effects of type of achievement measure. British Journal of Educational Psychology 80, 689–709.

Pinxten, M., Marsh, H.W., De Fraine, B., Van Den Noortgate, W., Van Damme, J., 2014. Enjoying mathematics or feeling competent in mathematics? Reciprocal effects on mathematics achievement and perceived math effort expenditure. British Journal of Educational Psychology 84 (1), 152–174. http://dx.doi.org/10.1111/bjep.12028.

Rosenberg, M., Schooler, C., Schoenbach, C., Rosenberg, F., 1995. Global self-esteem and specific self-esteem: different concepts, different outcomes. American Sociological Review 60, 141–156.

Seaton, M., Marsh, H.W., Craven, R.G., 2009. Earning its place as a pan-human theory: universality of the big-fish–little-pond effect across 41 culturally and economically diverse countries. Journal of Educational Psychology 101, 403.

Seaton, M., Marsh, H.W., Craven, R.G., 2010. Big-fish–little-pond effect generalizability and moderation – two sides of the same coin. American Educational Research Journal 47, 390–433.

Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Self-concept: validation of construct interpretations. Review of Educational Research 46, 407–441.

Schwartz, S.H., Bilsky, W., 1990. Toward a theory of the universal content and structure of values: extensions and cross-cultural replications. Journal of Personality and Social Psychology 58, 878.

Tracey, D.K., Marsh, H.W., Craven, R.G., 2003. Self-concepts of preadolescent students with mild intellectual disabilities: issues of measurement and educational placement. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research. Information Age, Greenwich, pp. 203–230.

Valentine, J.C., DuBois, D.L., 2005. Effects of self-beliefs on academic achievement and vice-versa: separating the chicken from the egg. In: Marsh, H.W., Craven, R.G., McInerney, D.M. (Eds.), International Advances in Self Research. Information Age, Greenwich, pp. 203–230.

Valentine, J.C., DuBois, D.L., Cooper, H., 2004. The relations between self-beliefs and academic achievement: a systematic review. Educational Psychologist 39, 111–133.

Watson, D., Suls, J., Haig, J., 2002. Global self-esteem in relation to structural models of personality and affectivity. Journal of Personality and Social Psychology 83, 185.

Wylie, R.C., 1974. The Self-Concept: Theory and Research on Selected Topics. University of Nebraska Press, Lincoln.

Wylie, R.C., 1979. The Self-Concept, vol. 2. University of Nebraska Press, Lincoln.

Self-Concepts: Educational Aspects

Reinhard Pekrun and Elizabeth J Stephens, University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-concepts are beliefs people hold about themselves. They are stored in memory, can be organized in hierarchical ways, and can pertain both to real attributes of the self and to nonfactual (e.g., ideal) aspects of the self. Findings of educational self-concept research show that students' self-concepts are largely organized according to academic domains (e.g., verbal vs math academic self-concept). They are reciprocally linked to academic achievement, implying that achievement results underlie the formation of related self-concepts, and that self-concepts in turn influence motivation, learning, and achievement. Beyond academic self-concepts, education at home and in academic institutions also influences students' general self-esteem. Important educational factors affecting general self-esteem as well as more specific self-concepts are acceptance by significant others (e.g., parents and teachers), consistent behavioral rules, support of autonomy, academic instruction, and feedback on abilities, behavior, and achievement. Of specific importance are reference norms used for giving feedback (e.g., social comparison vs mastery-oriented norms).

The capacity to reflect on one's capabilities and actions is uniquely human. Early in life, young children begin to form beliefs about the self that help them interpret the world and themselves, and serve as frames of reference for regulating emotion, motivation, and behavior. Research on self-related beliefs can be traced back to the seminal writings of William James (1893) who distinguished the self as 'I' ('existential self') and as 'me' ('categorical self'). The latter involves cognitions about the self as an object of reflection, as represented in statements such as 'I am thinking about me.' During the first half of the twentieth century, research on self-concepts was slow to develop. However, in the 1960s and 1970s, once behaviorist approaches were gradually replaced by cognitive accounts of psychological processes, psychologists started to recognize the importance of self-related cognitions for human agency. At that time, the self-concept construct gained a central status in psychology and applied fields like educational research. Subsequently, the number of studies on self-concepts increased dramatically. Today, we have cumulative knowledge about the structures, origins, and functions of self-concepts, including their role in education (Bracken, 1996; Valentine et al., 2004; Marsh et al., 2008).

Conceptual Issues

The term 'self-concept' is used in two ways. Talking about *the* self-concept of an individual should be distinguished from referring to a number of *different* self-concepts within the individual. 'The' self-concept is the whole of an individual's cognitive representations of him- or herself that are stored in memory. Different self-concepts are subsets of such representations that relate to different attributes of the self, such as cognitive abilities, physical appearance, or social status. Both variants of the term imply that self-concepts are self-representations that are more or less enduring over time, in contrast to situational, momentary self-perceptions that can vanish when the situation is over. The definition is open to whether these representations refer to actual personal attributes, such as real abilities, or to fictitious (e.g., possible or ideal) attributes.

There are two conceptual problems with the term 'self-concept' that are not adequately solved to date. First, it remains unclear whether self-related emotions such as pride or shame should be included or not. On the one hand, it can be argued that emotions are different from cognition, and that self-concepts should be defined as (purely) cognitive constructs. From this perspective, (cognitive) self-concepts and (emotional) self-related feelings can be combined using umbrella terms such as *self-esteem*, but should be considered as distinct entities in the first place. This perspective is supported by research showing that cognitive and affective components of self-concept scales can be separated empirically (e.g., Arens et al., 2011). On the other hand, many measures of self-concepts combine cognitive and affective items reflecting self-evaluation, which implies using a more integrative perspective. An important example is H. Marsh's widely used Self-description Questionnaire (Marsh, 1988) which includes both cognitive items (e.g., 'I am good at mathematics') and affective items (e.g., 'I enjoy doing work in mathematics') to measure self-concepts.

Second, there is insufficient agreement about the conceptual overlap between self-concept and other types of self-related cognitions. Specifically, what about cognitions linking the self to one's actions and the environment – are these cognitions part of self-concept or not? Two examples are self-efficacy beliefs and beliefs about personal control (Skinner, 1996). Self-efficacy beliefs pertain to one's capabilities to successfully solve specific tasks, thus linking self-attributes (ability) to actions (task performance) and environmental demands (a domain of tasks; Pajares, 1996). Beliefs about personal control, or 'perceived control' (Perry, 2003; Skinner, 1996), pertain to expectancies that one's actions produce desired outcomes. Given that these beliefs include representations of the self (along with other components), it could be argued that they should be considered as part of an individual's self-concept. However, self-concept, self-efficacy, and perceived control are viewed as separate constructs by most researchers and have been addressed in distinct research traditions up to date.

Types and Structures of Self-Concepts

Different types of self-concepts relate to different attributes of the self, such as one's abilities versus one's physical appearance. In addition, self-concepts can be of a more descriptive or more evaluative nature, and can pertain to either the factual reality of these attributes ('real self') or to fictitious attributes (e.g., one's 'ideal self'). Moreover, self-concepts can be explicit or implicit. As yet, educational research has primarily focused on examining structures of students' explicit concepts of the real self. Most of these studies targeted students' academic self-concepts representing their cognitive abilities.

Hierarchies of Self-Concepts and Attributes

Self-concepts pertaining to different attributes can be conceptualized in hierarchical ways, similar to semantic memory networks. From such a perspective, more general self-concepts are located at higher levels of the hierarchy and specific self-concepts at lower levels. In educational research, the hierarchical self-concept model proposed by Shavelson et al. (1976) stimulated an abundant number of studies on the structure of students' self-concepts. In this model, a 'general self-concept' is located at the top of the hierarchy. The next level includes an 'academic self-concept' representing one's academic abilities, as well as social, emotional, and physical nonacademic self-concepts. Self-concepts pertaining to different academic domains and to social, emotional, and physical subdomains are located on the third level, and even more specific self-concepts involving representations of situation-specific behavior on lower levels.

Research has shown that the correlations between academic self-concepts pertaining to different academic domains, such as math versus verbal self-concept, tend to be zero, in contrast to students' academic performance which is typically correlated positively across academic domains. Specifically, zero correlations are found for self-concepts across dissimilar domains, such as mathematics and English (Marsh and Shavelson, 1985). Moreover, affective variables that are likely to depend on students' self-concepts, including students' interests, values, goals, and academic emotions, have also been found to be organized in domain-specific ways (i.e., correlations are low to zero across domains; Bong, 2001; Goetz et al., 2007).

How can we explain the domain specificity of academic self-concepts and related variables? A theoretical account is provided by Marsh's (1986) Internal/External Frame of Reference model (I/E model). This model posits that students' academic self-concepts are based on comparisons using two types of standards: an internal, within-individual standard of comparing one's performance across academic domains (e.g., 'How does my performance in math compare to my performance in English?') and an external, between-individuals standard comparing one's performance with the performance of other students (e.g., 'How does my performance in math compare to other students' performance in math?').

Use of internal standards should lead to negative correlations between different academic self-concepts (e.g., perceiving high ability in math should lead to reduced estimates for verbal ability, and vice versa). In contrast, use of external standards based on social comparison should lead to positive correlations across domains, because performance scores and one's relative achievement status within a class are positively correlated across academic domains (e.g., positive correlations between math and verbal scores). The model proposes that students use both types of standards, implying that opposing effects can cancel each other out and that the overall correlations can be zero across domains. Given the domain specificity of academic self-concepts, researchers also have proposed modifying the original Shavelson et al. (1976) hierarchical model by including math and verbal self-concepts and omitting general academic self-concept (Marsh and Shavelson, 1985).

Different Perspectives on Attributes

Self-representations can consist of descriptive and evaluative accounts of attributes (e.g., 'I am tall' vs. 'I am an attractive person'). In academic self-concepts, this distinction is blurred because the use of comparative performance standards leads to judgments that involve an amalgam of descriptive and evaluative components (e.g., the statement 'I am better in English than most of my classmates' both describes and evaluates the self).

In forming self-concepts, one can use the same types of standards that teachers use to evaluate academic achievement and students use to develop their achievement goals (Elliot et al., 2011). These standards include (1) social comparison across individuals (external frame of reference in Marsh's I/E model cited earlier); (2) intraindividual comparison of present performance with past performance (i.e., improvement vs deterioration); (3) intraindividual comparison of actual performance with one's ability (i.e., potential performance); (4) intraindividual comparison of performance across domains (internal frame of reference in Marsh's I/E model cited earlier); (5) task-oriented comparison of one's performance with task mastery standards; and (6) cooperative standards linking individual performance to group performance.

Beyond existing attributes ('real self'), self-representations can pertain to attributes of the self that do not actually exist, but might exist ('possible selves'), are desired by oneself ('ideal self'), are wanted by others ('ought self'; Higgins, 1987), etc. Concepts of nonreal selves can also be important for students' affect and motivation. For example, discrepancies between real and ideal self are thought to trigger depressive emotions, whereas discrepancies between real and ought self can induce anxiety (Higgins, 1987).

Explicit versus Implicit Self-Concepts

Similar to other variables that involve human affect and behavior (e.g., motives; Schultheiss, 2008), not only can self-concepts be explicit, reflective, and deliberate, but they can also be implicit and nonreflective. Explicit self-concepts consist of propositional structures that result from deliberate information processing. These self-concepts can be described by models capturing reflective, propositional associations between the self and its various attributes, such as the hierarchical Shavelson et al. (1976) model discussed earlier. In contrast, implicit self-concepts are thought to be based on associative structures that result from automatic information

processing involving spontaneous associations between the self and an attribute rather than any deliberate cognitive effort. Explicit self-concepts can be assessed using self-report methods such as questionnaires and interviews, whereas assessment of implicit self-concepts needs indirect methods to assess associations that are less accessible to the conscious mind, such as the Implicit Association Test (e.g., Greenwald and Farnham, 2000).

To date, educational research has focused on analyzing students' and teachers' explicit self-concepts, and has only just begun to examine the structures and functional role of implicit self-concepts. Initial findings on these self-concepts suggest that they can play a major role in students' motivation and performance. For example, recent studies found that discrepancies between students' explicit and implicit self-concept of general cognitive ability (i.e., intelligence) predicted decreased performance on intelligence tests (Dislich et al., 2011). It seems likely that implicit self-concepts are as important for students' academic agency as implicit motives have been found to be.

Self-Concepts and Academic Achievement

The link between self-related beliefs and achievement is one of the most frequently examined research questions in the social sciences. The relationship between students' self-concepts and academic performance also has been analyzed in several hundreds of studies. In the first phase of research, researchers examined this relationship by using cross-sectional designs and correlating self-concepts scores with performance scores. Later, longitudinal designs were employed to investigate predictive effects of prior self-concepts on subsequent achievement and to investigate reciprocal effects linking the two constructs over time.

Empirical Evidence

The evidence from cross-sectional studies shows that self-concepts and students' achievement are positively correlated. However, the magnitude of the correlation depends on the self-concept and achievement measures used. Academic self-concepts show higher correlations with academic performance than general self-concept or general self-esteem, in line with principles of symmetry and compatibility (Brunswick, 1955; Fishbein and Ajzen, 2010). In a meta-analysis by Hansford and Hattie (1982), the average correlation between students' general self-esteem and achievement was $r = 0.22$, whereas the correlation between students' academic self-concept and achievement was $r = 0.42$. When both self-concepts and achievement are assessed in more domain-specific ways (e.g., math self-concept and math achievement), the correlations tend to be even higher. This pattern of findings suggests that the link between self-concept and achievement is fortified when the measures for both constructs are conceptualized in domain-specific ways.

Beyond cross-sectional correlation, longitudinal studies have shown that self-concept is a positive predictor of subsequent academic achievement, even when controlling for baseline achievement scores. Valentine et al. (2004) conducted a meta-analysis of 55 research reports on the predictive achievement effects of various self-beliefs. The average Betas for the effects of self-concept, self-esteem, and self-efficacy on achievement when controlling for baseline achievement were 0.08, 0.07, and 0.11, respectively. These findings suggest that students' self-concepts do have a positive causal impact on achievement, in contrast to views positing that self-beliefs do not affect achievement or even undermine it (see Baumeister et al., 2003; Marsh and Craven, 2006).

Finally, using cross-lagged analysis and structural equation modeling of longitudinal data, researchers can simultaneously examine the effects of self-concept on subsequent achievement and of achievement on subsequent self-concept scores. Specifically, cross-lagged designs are well-suited to more fully examine the causal effects underlying self-concept/achievement correlations. Competing theoretical perspectives on these effects were proposed by the 'skill development model' maintaining that students' self-concepts result from prior experiences with academic achievement (i.e., effects of achievement on self-concept), and the 'self-enhancement approach' positing that self-concept influences achievement (Helmke and van Aken, 1995). Both perspectives are combined in models acknowledging reciprocal causation, such as the reciprocal effects model (REM) proposed by H. Marsh and his colleagues (see Marsh and Craven, 2006).

The findings from cross-lagged studies lend credence to the validity of the REM (e.g., Marsh et al., 2005). Typically, the effects of achievement scores on subsequent self-concept are somewhat stronger than the effects of self-concept on subsequent achievement. Overall, the evidence indicates that students' self-concepts positively influence their achievement, and that achievement, in turn, has positive effects on the development of students' self-concepts. Less is known about the mechanisms producing these effects, however. The next section highlights some potential mediating processes.

Effects of Self-Concept on Achievement

It seems unlikely that beliefs about one's capabilities directly affect subsequent performance - there must be mechanisms mediating these effects. Two important possible mechanisms are motivation for self-consistency and the effects of self-concepts on perceived academic control and achievement expectancies. In terms of *self-consistency*, humans seek to maintain consistency between their self-views and behaviors (Swann, 1997). By implication, students who hold positive views of their academic abilities should strive to attain success in order to be consistent with their self-view, whereas students with lower academic self-concepts may be motivated to behave in accordance with a more pessimistic self-view. Such self-consistent behavior would lead to performance scores that are consistent with students' self-concepts, thus supporting the positive correlation between self-concept and achievement.

With regard to *perceived control* and *achievement expectancies*, perceptions of one's abilities should be a major source of perceived control over achievement and related expectancies to attain success and to avoid failure. By implication, academic self-concepts should influence affective processes that depend on control appraisals and expectancies. Specifically, from the perspective of expectancy-value theories of student motivation (Eccles and Wigfield, 2002), academic self-concepts have positive effects on students' expectancies to succeed, thus

positively influencing their motivation to learn and ultimately their achievement. Similarly, from the perspective of the control-value of achievement emotions (Pekrun, 2006), academic self-concepts should have a positive impact on students' perceived control over achievement, thus increasing adaptive emotions such as enjoyment of learning and hope for success, and reducing maladaptive emotions such as anxiety, hopelessness, and boredom. Since these emotions affect students' learning, they are likely to mediate the effects of academic self-concepts on students' achievement (see e.g., Pekrun et al., 2011, for related evidence).

Effects of Achievement on Self-Concepts

Success and failure are likely to be a major source of students' academic self-concepts. Specifically, if achievement is made salient (e.g., in terms of feedback by teachers), and if successful or unsuccessful achievement experiences are repeated in cumulative, consistent ways over time, it is likely that students self-concepts of ability develop congruently with their academic performance attainment (see next section). This would explain the positive impact of achievement scores on students' academic self-concepts. Moreover, in contrast to the more complicated mechanisms mediating effects of self-concept of achievement, the effects of achievement on self-concept formation are more straightforward in terms of involving a smaller number of mediating variables. This might help to explain why the effects of achievement on self-concept are typically stronger than the effects of self-concept on achievement.

Development of Self-Concepts

Human Needs Underlying Self-Concept Development

Two categories of basic human needs are thought to govern the development of self-representations (Epstein, 1973). The first category includes needs for maximizing pleasure and minimizing pain, from which stem needs for self-enhancement. Because it is pleasurable to maintain a positive view of the self and painful to maintain a negative view, humans should generally be motivated to establish positive self-perceptions (even if they do not coincide with reality). The second category comprises needs to understand reality and foresee the future, which implies the need to maintain self-perceptions that are in and of themselves consistent as well as consistent with reality (i.e., needs for consistency). Self-enhancement and consistency converge when feedback about the self is positive; however, they can conflict when feedback is negative. Furthermore, satisfying needs for self-enhancement may lead an individual to reject negative feedback, whereas satisfying needs for consistency with reality may result in the individual endorsing such feedback if it is deemed veridical. The development of students' self-concepts can be explained by the interplay of these partially opposing needs systems.

Developmental Trajectories across the Life Span

In line with needs for self-enhancement, if information about real abilities lacks salience or is unavailable, then humans tend to overestimate their capabilities. At the same time, they search for factual information about the self and can even be motivated to endorse negative self-information if it is salient and consistent. In accord with these opposing trends, studies have shown that many children dramatically overestimate their academic abilities upon entering school, but over time lower their self-evaluations when teachers' feedback repeatedly diverges from such positive self-views (Nicholls, 1978). Young children's lack of metacognitive competencies may make it difficult for them to use feedback to develop realistic perceptions of their abilities; however, this process can be expedited later. As a general rule, the interplay between beginner's optimism and expert's realism is likely to extend beyond the elementary school years and emerge in numerous situations such as starting college, entering a new career, and getting married (Perry, 2003). Indeed it can be said that self-concept development over much of the life span is characterized by this joint action between the sometimes opposing forces of self-related optimism and realism.

Impact of Educational Environments

The influence of social environments was addressed early in the twentieth century by symbolic interactionism, which posited that interactions with significant others shape the development of self-concepts (cf. the concept of the 'looking glass self'; Cooley, 1902). The environment provides various sources of self-related information and such information can influence students' self-concepts, including attributions of ability and personal worth, opportunities to develop and demonstrate knowledge and competencies, achievement feedback, and the composition of student groups and related opportunities for social comparison of one's abilities.

First, the *attributions* teachers, parents, and peers provide for success and failure (e.g., attributions to ability vs. effort) influence students' self-concepts. These attributions can be conveyed directly, for example, in verbal messages (e.g., 'You failed the exam because you didn't prepare well'). Additionally, attributions from others can take more indirect forms, such as the implied messages communicated by others' emotional and instrumental behavior. Importantly, these oftentimes well-intentioned messages can produce contradictory effects, such as help from a teacher on an easy task or praise for successful performance on an easy task indicating low ability (paradoxical effects of praise; e.g., Rheinberg, 1988). Messages about personal worth can also be conveyed indirectly, for example, by teachers showing student support and expressing affection and positive emotions. These messages and attributions are of fundamental importance for the development of students' general self-esteem (e.g., Pekrun, 1990).

Second, learning environments provide *learning opportunities* that help build the knowledge and skills that make it possible for students to attain success and develop positive academic self-concepts. Specifically, high-quality classroom instruction can promote successful student learning and related self-concepts (Hattie, 2009). Furthermore, environments supportive of autonomy can help students acquire important self-regulatory abilities that enable them to effectively guide and manage their own learning. Finally, consistent behavioral rules can enhance the cognitive predictability of students'

environments which can also positively affect their overall sense of competence.

Third, *achievement feedback* is likely critical for self-concept development. As noted earlier, feedback that is salient, consistent across situations, and cumulative over time may be especially relevant to students' self-concepts. The content of and achievement standards used for feedback are important as well. Specifically, as noted earlier, different definitions of achievement are provided by normative, intraindividual, task-referenced, and cooperative standards (Elliot et al., 2011; Johnson and Johnson, 1974). Normative standards (i.e., social comparison standards) are competitive and define achievement in relation to one's peers (i.e., high achievement implies outperforming peers, low achievement implies performing worse than one's peers). Intraindividual standards define achievement in terms of individual improvement over time or in terms of using one's potential. In task-oriented (criterion-referenced) standards, achievement is defined in terms of attaining mastery of the learning material. Cooperative standards define achievement in terms of group performance rather than individual performance.

Because these different standards imply differential opportunities for students to experience success, they also affect students' self-concepts. Specifically, in contrast to criterion-referenced and intraindividual standards, normative standards imply that some students can attain success, whereas others must fail. The failure experiences produced by normative standards and competitive grading structures can, for some students, contribute to lowered academic self-concepts.

Finally, the *composition of student groups* influences students' within-group achievement position, thus also affecting their self-concepts of ability. Specifically, the ability levels of student groups determine students' opportunities to experience success and avoid failure, thereby impacting self-concepts. All things being equal, membership in a high-ability classroom reduces one's chances to succeed relative to others, whereas membership in a low-ability classroom increases the likelihood of attaining success. Two opposing types of social comparison processes are thought to mediate the impact of group-level ability on students' self-concepts: *Contrast effects* focus one's attention on differences between oneself and the group, implying that self-concepts of ability are reduced if individual achievement is low relative to the group, and increased if achievement is high relative to the group. *Assimilation effects* imply adapting self-concepts according to group membership, thus increasing self-concepts for members of high-ability groups and decreasing self-concepts for members of low-ability groups.

Contrast and assimilation effects on academic self-concepts have been called the *big-fish-little-pond effect* (BFLPE) and the *reflected-glory effect*, respectively (Marsh, 1987; Marsh et al., 2000). Research has shown that the BFLPE on students' academic self-concepts is usually stronger than reflected glory effects. By implication, membership in high-ability classrooms typically has negative effects on students' average academic self-concepts, whereas membership in low-ability classrooms has positive effects. The BFLPE is known to be one of the most robust effects in the social sciences and has been shown to occur in students of various age groups and in various educational institutions around the globe (Marsh et al., 2008). One of the paradoxical implications of this effect is that students' self-concepts tend to decrease after transitions to high-ability school tracks and classrooms (such as classes for gifted students), and tend to increase after transitioning to a lower-ability classroom.

Summary of Implications for Educational Practice

Fostering students' self-concepts can be beneficial for their academic performance, career trajectories, and personality development. Furthermore, since positive self-regard is a key component of psychological health and well-being, nurturing students' self-concepts is an educational goal that is in itself valuable. From the evidence cited earlier, it follows that education can contribute substantially to self-concept development. Acceptance and support in educational environments such as schools and the family are important for students' general self-esteem. In addition, providing students with positive, specific feedback about their individual strengths in terms of performance, abilities, and personality can help them develop optimistic self-concepts and life expectancies. In designing feedback, it is helpful to define achievement using criterion-referenced, individual, or cooperative standards rather than to rely on competitive standards and normative grading. Moreover, educators can promote students' self-concept development by providing high-quality instruction, establishing consistent, predictable expectations of behavior, and supporting autonomy and self-regulation in order to promote students' perceptions of control and competence.

See also: Academic Self-Concept and Achievement; Personality and Self; Self and Identity Development During Adolescence across Cultures; Self in Culture: Early Development; Self-Concept: From Unidimensional to Multidimensional and Beyond; Self-Efficacy and Health; Self-Efficacy: Education Aspects; Self-Efficacy; Self-Esteem; Self: History of the Concept.

Bibliography

Arens, A.K., Yeung, A.S., Craven, R.G., Hasselhorn, M., 2011. The twofold multidimensionality of academic self-concept: domain specificity and separation between competence and affect components. Journal of Educational Psychology 103, 970–981.

Baumeister, R.F., Campbell, J.D., Krueger, J.I., Vohs, K.D., 2003. Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? Psychological Science in the Public Interest 4 (1).

Bong, M., 2001. Between- and within-domain relations of academic motivation among middle and high school students: self-efficacy, task-value and achievement goals. Journal of Educational Psychology 93, 23–34.

Bracken, B.A. (Ed.), 1996. Handbook of Self-concept. Wiley, New York.

Brunswick, E., 1955. Representative design and probabilistic theory in a functional psychology. Psychological Review 62, 193–217.

Cooley, C.H., 1902. Human Nature and the Social Order. Scribner, New York.

Dislich, F.X.R., Imhoff, R., Banse, R., Altstötter-Gleich, C., Zinkernagel, A., Schmitt, M., 2011. Discrepancies between implicit and explicit self-concepts of intelligence predict performance on tests of intelligence. European Journal of Personality 26, 212–220.

Eccles, J.S., Wigfield, A., 2002. Motivational beliefs, values, and goals. Annual Review of Psychology 53, 109–132.

Elliot, A.J., Murayama, K., Pekrun, R., 2011. A 3 × 2 achievement goal model. Journal of Educational Psychology 103, 632–648.

Epstein, S., 1973. The self-concept revisited or a theory of a theory. American Psychologist 28, 404–416.

Fishbein, M., Ajzen, I., 2010. Predicting and Changing Behavior: The Reasoned Action Approach. Psychology Press, New York.

Goetz, T., Frenzel, A.C., Pekrun, R., Hall, N.C., Lüdtke, O., 2007. Between- and within-domain relations of students' academic emotions. Journal of Educational Psychology 99, 715–733.

Greenwald, A.G., Farnham, S.D., 2000. Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social Psychology 79, 1022–1038.

Hansford, B.C., Hattie, J.A., 1982. The relationship between self and achievement/performance measures. Review of Educational Research 52, 123–142.

Hattie, J.A.C., 2009. Visible Learning. A Synthesis of over 800 Meta-analyses Relating to Achievement. Routledge, New York.

Helmke, A., van Aken, M.A.G., 1995. The causal ordering of academic achievement and self-concept of ability during elementary school: a longitudinal study. Journal of Educational Psychology 87, 624–637.

Higgins, E.T., 1987. Self-discrepancy: a theory relating self and affect. Psychological Review 94, 319–340.

James, W., 1893. Psychology. Fawcett, New York.

Johnson, D.W., Johnson, R.T., 1974. Instructional goal structure: cooperative, competitive or individualistic. Review of Educational Research 4, 213–240.

Marsh, H.W., 1986. Verbal and math self-concepts: an internal/external frame of reference model. American Educational Research Journal 23, 129–149.

Marsh, H.W., 1987. The big-fish-little-pond effect on academic self-concept. Journal of Educational Psychology 79, 280–295.

Marsh, H.W., 1988. Self-description Questionnaire I (SDQ I): Manual and Research Monograph. Psychological Corporation, San Antonio, TX.

Marsh, H.W., Kong, C., Hau, K., 2000. Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: counterbalancing contrast and reflected glory effects in Hong Kong schools. Journal of Personality and Social Psychology 78, 337–349.

Marsh, H.W., Craven, R.G., 2006. Reciprocal effects of self-concept and performance from a multidimensional perspective: beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science 1, 133–163.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, U., Hau, K.T., O'Mara, A.J., Craven, R.G., 2008. The big-fish-little-pond-effect stands up to critical scrutiny: implications for theory, methodology, and future research. Educational Psychology Review 20, 319–350.

Marsh, H.W., Shavelson, R.J., 1985. Self-concept: its multifaceted, hierarchical structure. Educational Psychologist 20, 107–125.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O., Baumert, J., 2005. Academic self-concept, interest, grades and standardized test scores: reciprocal effects models of causal ordering. Child Development 76, 397–416.

Nicholls, J.G., 1978. The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. Child Development 49, 800–814.

Pajares, F., 1996. Self-efficacy beliefs in academic settings. Review of Educational Research 66, 543–578.

Pekrun, R., 1990. Social support, achievement evaluations, and self-concepts in adolescence. In: Oppenheimer, L. (Ed.), The Self-concept. Springer, Berlin, pp. 107–119.

Pekrun, R., 2006. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18, 315–341.

Pekrun, R., Goetz, T., Frenzel, A.C., Barchfeld, P., Perry, R.P., 2011. Measuring emotions in students' learning and performance: the Achievement Emotions Questionnaire (AEQ). Contemporary Educational Psychology 36, 36–48.

Perry, R.P., 2003. Perceived (academic) control and causal thinking in achievement settings. Canadian Psychology/Psychologie Canadienne 44, 312–331.

Rheinberg, F., 1988. Paradoxe Effekte von Lob und Tadel [Paradoxical effects of praise and blame]. Huber, Bern, Switzerland.

Schultheiss, O.C., 2008. Implicit motives. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality: Theory and Research, third ed. Guilford, New York, pp. 603–633.

Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Self-concept: validation of construct interpretations. Review of Educational Research 46, 407–441.

Skinner, E.A., 1996. A guide to constructs of control. Journal of Personality and Social Psychology 71, 549–570.

Swann, W.B., 1997. The trouble with change: self-verification and allegiance to the self. Psychological Science 8, 177–180.

Valentine, J.C., DuBois, D.L., Cooper, H., 2004. The relation between self-beliefs and academic achievement: a meta-analytic review. Educational Psychologist 39, 111–133.

Selfconscious Emotions, Psychology of

June P Tangney, George Mason University, Fairfax, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Shame, guilt, embarrassment, and pride are 'self-conscious emotions,' evoked by self-reflection and self-evaluation. A growing empirical literature indicates that shame and guilt are distinct emotions with very different implications for psychological adjustment and interpersonal behavior. In brief, shame involves a negative evaluation of the global self, whereas guilt involves a negative evaluation of a specific behavior. On balance, guilt appears to be the more adaptive emotion, benefiting relationships in a variety of ways. Shame has been associated with a variety of negative outcomes – evasion, denial, anger, and vulnerability to a range of psychological symptoms. Recent research, however, has begun to explore the conditions under which shame may be adaptive – fostering meaningful change. Embarrassment is not simply a mild form of shame, but is marked by more humor, more obvious physiological changes (e.g., blushing), and a greater sense of exposure. The diversity of situations that cause embarrassment makes a comprehensive 'account' of embarrassment difficult. There is greater consensus about its adaptive significance. Embarrassment signals appeasement to others, thus diffusing negative social evaluations and the likelihood of retaliation. In recent years, pride has received long overdue attention. Paralleling the self versus behavior distinction central to shame and guilt, there are two types of pride. The propensity to experience pride about behavior ('authentic' or 'behavioral' pride) appears to be similarly more adaptive than the propensity to experience pride in self ('hubristic' pride).

Shame, guilt, embarrassment, and pride are members of a family of 'self-conscious emotions' that are evoked by self-reflection and self-evaluation. This self-evaluation may be implicit or explicit, consciously experienced or transpiring beyond our awareness. But either way, the self is the object of self-conscious emotions. For example, when good things happen, we may feel a range of positive emotions – joy, happiness, contentment. But we feel pride in our *own* positive characteristics or actions. Similarly, when bad things happen, many negative emotions may arise – sadness, disappointment, frustration, anger. But feelings of shame and guilt typically arise from the recognition of one's *own* negative characteristics or behaviors. Even when we feel pride or shame due to another person's behavior, that person is apt to be someone with whom we are closely affiliated or identified (e.g., a family member, close friend). We experience shame because that person is part of our self-definition.

In contrast to 'basic' emotions (e.g., anger, fear, joy) which are present very early in life, self-conscious emotions have been described as 'secondary,' 'derived,' or 'complex' emotions because they emerge later and hinge on several cognitive achievements – recognition of the self separate from others and a set of standards against which the self is evaluated. For example, Lewis et al. (1989) showed that the capacity to experience embarrassment coincides with the emergence of self-recognition. Very young children first show behavioral signs of embarrassment during the same developmental phase (15–24 months) in which they show a rudimentary sense of self. Moreover, within this range, children who display self-recognition (in a 'rouge' test) are the same children who display signs of embarrassment in an unrelated task.

Also in contrast to 'basic' emotions, self-conscious emotions do not bear a one-to-one correspondence to clearly defined, facial expressions. Recent research has identified postural indicators of pride (e.g., head held high, shoulders back, Tracy and Robins, 2008), and there are physical signals of shame and embarrassment (slumped, shoulders, downcast eyes/head), but physical signs of shame and embarrassment are difficult to differentiate. These often subtle cues are not as readily apprehended as 'emotions of the face' such as anger, sadness, joy, and disgust. And there exists no known visible marker of guilt, on the face or otherwise. Thus, most researchers interested in the self-conscious emotions rely on self-report methods.

Shame and Guilt

The terms 'shame' and 'guilt' are often used interchangeably in everyday conversation. When people do distinguish the two, they typically suggest that shame arises from public exposure and disapproval of a failure or transgression, whereas guilt is more 'private,' arising from one's own conscience. Empirical research, however, has failed to support this public/private distinction. For example, Tangney et al. (1996a) found that people's real-life shame and guilt experiences are each most likely to occur in the presence of others. More important, neither the presence of others nor others' awareness of the respondents' behavior distinguished between shame and guilt. Overall, there are surprisingly few differences in the types of events that elicit shame and guilt. Shame is somewhat more likely in response to violations of social norms, but most types of events (e.g., lying, cheating, stealing, failing to help another) result in guilt for some people and shame for others.

So what is the difference between shame and guilt? Empirical research supports Helen Block Lewis's (1971) notion that shame involves a negative evaluation of the global self, whereas guilt involves a negative evaluation of a specific behavior. This differential emphasis on self ('*I* did that horrible thing') versus behavior ('I *did* that horrible *thing*') leads to different affective experiences.

Shame is an acutely painful emotion typically accompanied by a sense of shrinking, 'being small,' and by a sense of

worthlessness and powerlessness. Shamed people also feel exposed. Although shame does not necessarily involve an actual observing audience, there is often the imagery of how one's defective self would appear to others. Not surprisingly, shame often leads to a desire to escape or hide – to sink into the floor and disappear.

In contrast, guilt is typically less painful and devastating because the primary concern is with a specific behavior, not the entire self. Guilt does not affect one's core identity. (One can do a bad thing, but still be a good person.) Instead, when people feel guilt, they experience a tension, remorse, and regret over the 'bad thing done' and a nagging preoccupation with the transgression. Rather than motivating avoidance, guilt typically motivates reparative action.

Implications of Shame and Guilt for Interpersonal Adjustment

Research has consistently shown that, on balance, guilt is the more adaptive emotion, benefiting relationships in a variety of ways (Baumeister et al., 1994; Tangney, 1995). Three sets of findings illustrate the adaptive, 'relationship-enhancing functions' of guilt, in contrast to the hidden costs of shame.

First, the action tendencies associated with guilt are more apt to appease victims, repair relationships, and result in a better behavior in the future. Guilt often motivates reparative action – confessing, apologizing, undoing the harm that was done. In contrast, shame leads to attempts to deny, hide, or escape. Thus, guilt orients people in a more constructive, proactive, future-oriented direction, whereas shame orients people toward separation, distance, and defense (de Hooge et al., 2007; Lewis, 1971; Lindsay-Hartz, 1984; Sheikh and Janoff-Bulman, 2010; Tangney et al., 1996a).

Second, a special link between guilt and empathy has been observed at the levels of both emotion states and dispositions (Joireman, 2004; Leith and Baumeister, 1998; Stuewig et al., 2010; Tangney et al., 2007). For example, studies of children, college students, and adults (Tangney, 1995) show that guilt-prone individuals are generally empathic individuals. In contrast, shame-proneness is associated with an impaired capacity for other-oriented empathy and a propensity for 'self-oriented' personal distress responses. Similar findings are evident when considering feelings of shame and guilt 'in the moment.' Individual differences aside, when people describe personal guilt experiences, they convey greater empathy for others, compared to descriptions of shame experiences. It appears that by focusing on a bad behavior (as opposed to a bad self), people experiencing guilt are relatively free of the egocentric, self-involved focus of shame. Instead, their focus on a specific behavior is likely to highlight the consequences for distressed others.

Third, there is a special link between shame and anger, again observed at both the dispositional and state levels. At all ages, shame-prone individuals are also prone to feelings of anger and hostility. Moreover, once angered, they tend to manage their anger in an aggressive, unconstructive fashion. In contrast, guilt is generally associated with constructive means of handling anger (Tangney et al., 1996b). What accounts for this link between shame and anger? When feeling shame, people initially direct hostility inward ('*I'm* such a bad person'). But this hostility may be redirected outward in a defensive attempt to protect the self by shifting the blame elsewhere ('Oh what a horrible person I am, and damn it, how could you make me feel that way!'). Empirically, this relationship between shame and externalization of blame is especially strong and consistent across people from all walks of life. Moreover, evidence suggests that shame leads to aggression via its influence on externalizing blame (Stuewig et al., 2010).

In sum, studies employing diverse samples, measures and methods converge. All things equal, it is better if your friend, partner, child, or boss feels guilt than shame. Shame motivates behaviors that interfere with interpersonal relationships. Guilt helps keep people constructively engaged in the relationship at hand.

Implications of Shame and Guilt for Psychological Adjustment

Although guilt appears to be the more 'moral' or adaptive emotion when considering social adjustment, is there a trade-off vis-à-vis individual psychological adjustment? Does the tendency to experience guilt or shame leave one vulnerable to psychological problems?

Researchers consistently report a relationship between proneness to shame and a whole host of psychological symptoms, including depression, anxiety, eating disorder symptoms, subclinical sociopathy, and low self-esteem (Harder et al., 1992; Tangney et al., 2007). This relationship is robust across measurement methods and diverse populations.

There is more controversy regarding the relationship of guilt to psychopathology. The traditional view is that guilt plays a significant role in psychological symptoms. Clinical theory and case studies, including those of Freud, make frequent reference to a maladaptive guilt characterized by chronic self-blame and obsessive rumination over one's transgressions. On the other hand, recent theory and research has emphasized the adaptive functions of guilt, particularly for interpersonal behavior. Tangney et al. (1992) argued that once one makes the critical distinction between shame and guilt, there is no compelling reason to expect guilt over specific behaviors to be associated with poor psychological adjustment.

The empirical research is similarly mixed. Studies employing adjective checklist-type (and other globally worded) measures find both shame-proneness and guilt-proneness correlated with psychological symptoms. On the other hand, measures sensitive to the self versus behavior distinction show no relationship between proneness to 'shame-free' guilt and psychopathology (Tangney et al., 2007).

Implications of Shame and Guilt for Deterring Transgression and Socially Undesirably Behavior

Shame may be very painful, and it may carry with it considerable psychological baggage, but does it at least motivate people to avoid 'doing wrong,' decreasing the likelihood of transgression and impropriety? There is little direct evidence for this presumed adaptive function of shame when considering actual criminal behavior. To the contrary, research suggests that shame has little effect or may even make things worse. For example, in a study of college undergraduates, Tibbetts (2003) found that criminal offending was negatively related to guilt-proneness, but results for shame-proneness were mixed.

In a community sample of adolescents (Stuewig and McCloskey, 2005), proneness to 'shame-free' guilt negatively predicted delinquency; proneness to 'guilt-free' shame did not. In a large German sample of incarcerated adolescents and young adults (Hosser et al., 2008), single item shame ratings at the outset of incarceration predicted higher recidivism rates, whereas guilt ratings predicted lower recidivism. In large sample of adult jail inmates (Tangney et al., 2011), proneness to guilt was negatively correlated with criminal history, risk assessment measures, and psychological factors known to predict violent and nonviolent criminal recidivism. In contrast, inmates' shame-proneness was unrelated to clinician ratings of psychopathy and violent risk, and positively correlated with self-reported substance use, antisocial personality and criminogenic cognitions. In short, in the domain of crime, the experience of guilt clearly serves important inhibitory functions in ways that shame does not.

When Does Guilt Become Maladaptive?

Guilt is frequently cited as a symptom in such psychological disorders as depression, anxiety, and post-traumatic stress disorder (PTSD). What is the chronic problematic guilt described by so many clinicians? One possibility is that maladaptive guilt experiences are actually feelings of guilt *fused with shame*. When a person initially experiences guilt ('Oh, look at what a horrible *thing* I have *done*') but then magnifies and generalizes the event to himself or herself ('…and aren't I a horrible *person*'), many of the advantages of guilt are lost. In addition to feeling tension and remorse over a specific behavior, he or she is impaired by feelings of contempt and disgust for a bad self. For these reasons, shame-fused guilt may be just as problematic as shame itself.

Problems may also arise when people hold an exaggerated or distorted sense of responsibility for events beyond their control, as in survivor guilt associated with PTSD and other psychological problems (O'Connor et al., 2002; Kubany et al., 1995). Importantly, widely used scenario-based measures such as the Test of Self-Conscious Affect do not capture the propensity to experience guilt over events beyond one's control. In a meta-analysis of studies of shame, guilt, and depression (Kim et al., 2011), measures assessing 'legitimate' guilt (for events for which individuals have responsibility) showed little relationship to depression, whereas measures assessing guilt involving unrealistic responsibility for negative events were positively linked to depression.

Finally, there is some evidence that guilt can motivate self-denial or self-punishment when opportunities for reparation are blocked (Nelissen and Zeelenberg, 2009).

When Is Shame Adaptive?

The picture for shame is not entirely bleak. Some recent research in social psychology suggests that in some domains (e.g., academic) and under some conditions yet to be systemically examined, the experience of shame may motivate efforts to repair or improve oneself (de Hooge et al., 2008, 2010; Thompson and Berenbaum, 2006). In contrast, the conditions under which guilt motivates reparative or prosocial behavior are much broader in circumstance and personal attributes (de Hooge et al., 2007).

Embarrassment

Miller (1996) defines embarrassment as "an aversive state of mortification, abashment, and chagrin that follows public social predicaments" (p. 322). Indeed, embarrassment appears to be the most 'social' of the self-conscious emotions, occurring almost without exception in the company of others.

Causes of Embarrassment

In Miller's (1996) catalog of embarrassing events described by several hundred adolescents and adults, 'normative public deficiencies' (tripping in front of a large class, forgetting someone's name, unintended bodily induced noises) were at the top of the list. But there were many other types of embarrassment situations – awkward social interactions, conspicuousness in the absence of any deficiency, 'team transgressions' (embarrassed by a member of one's group), and 'empathic' embarrassment.

The diversity of situations that lead to embarrassment has posed a challenge to efforts at constructing a comprehensive 'account' of embarrassment. Some theorists believe that the crux of embarrassment is negative evaluation by others (Edelmann, 1981; Miller, 1996). This social evaluation account runs into difficulty with embarrassment events that involve no apparent deficiency (e.g., being the center of attention during a 'Happy Birthday' chorus). Other theorists subscribe to the 'dramaturgic' account, positing that embarrassment occurs when implicit social roles and scripts are disrupted. A flubbed performance, an unanticipated belch, and being the focus of 'Happy Birthday' each represent a deviation from accustomed social scripts.

Lewis (1992) distinguished between two types of embarrassment – embarrassment due to exposure and embarrassment due to negative self-evaluation. According to Lewis (1992), embarrassment due to exposure emerges early in life, once children develop a rudimentary sense of self. When children develop standards, rules, and goals (SRGs), a second type of embarrassment emerges – 'embarrassment as mild shame' associated with failure in relation to SRGs.

Functions of Embarrassment

Although there is debate about the fundamental causes of embarrassment, there is general agreement about its adaptive significance. Gilbert (1997) suggests that embarrassment serves an important social function by signaling appeasement to others. When untoward behavior threatens a person's standing in an important social group, visible signs of embarrassment function as a nonverbal acknowledgment of shared social standards, thus diffusing negative evaluations and the likelihood of retaliation. Evidence from studies of both humans and nonhuman primates supports this remedial function of embarrassment (Keltner and Buswell, 1997).

Embarrassment and Shame

Is there a difference between shame and embarrassment? Some theorists essentially equate the two emotions. A more dominant view is that shame and embarrassment differ in intensity

of affect and/or severity of transgression. Still others propose that shame is tied to perceived deficiencies of one's core self, whereas embarrassment results from deficiencies in one's presented self.

Research suggests that shame and embarrassment are indeed quite different emotions – more distinct, even, than shame and guilt. For example, comparing adults' personal shame, guilt, and embarrassment experiences, Tangney et al. (1996b) found that shame was a more intense, painful emotion that involved a greater sense of moral transgression. But controlling for intensity and morality, shame and embarrassment still differed markedly along many affective, cognitive, and motivational dimensions. When shamed, people felt greater responsibility, regret, and self-directed anger. Embarrassment was marked by more humor, blushing, and a greater sense of exposure.

Individual Differences in Embarrassability

As with shame and guilt, people vary in their propensity to experience embarrassment. These individual differences are evident within the first years of life, and are relatively stable across time. Research has shown that embarrassability is associated with neuroticism, high levels of negative affect, self-consciousness, and a fear of negative evaluation from others. Miller (1996) has shown that this fear of negative evaluation is not due to poor social skills, but rather a heightened concern for social rules and standards.

Pride

Of the self-conscious emotions, pride received the least attention, until recently. Most early research on pride comes from developmental psychology, particularly in the achievement domain. For example, Stipek and Recchia (1992) observed developmental changes in the criteria children use for evaluating success and failure, in the types of situation that lead to pride, and in the importance of praise from others.

Mascolo and Fischer (1995) defined pride as an emotion "generated by appraisals that one is responsible for a socially valued outcome or for being a socially valued person" (p. 66). They theorized that pride serves to enhance people's self-worth and to encourage future behavior that is socially valued.

Two Types of Pride?

Both Tangney (1990) and Lewis (1992) suggested that there are two types of pride. Paralleling the self versus behavior distinction of guilt and shame, Tangney (1990) distinguished between pride in self ('alpha' pride) and pride in behavior ('beta' pride). Similarly, Lewis (1992) distinguished between pride (arising from attributing one's success to a specific action) and hubris (pridefulness arising from attributions of success to the global self). Lewis (1992) viewed hubris as largely maladaptive, noting that hubristic individuals are inclined to distort and invent situations to enhance the self, which can lead to interpersonal problems.

Research over the past decade at multiple levels adopting multiple methods (Tracy and Robins, 2007) strongly supports this distinction between 'authentic' (or behavioral) pride and hubristic pride. And like shame and guilt, hubristic and authentic pride do not seem to be equally moral or adaptive emotions. At both the trait and state level, authentic pride is associated with a likable and socially desirable characteristics and prosocial behaviors, whereas hubris is associated with a more negative characteristics and antisocial behaviors (Tracy and Robins, 2007). For example, at the trait level, the two types of pride show divergent relations with achievement, mental health, social behavior, and relationship functioning (Carver et al., 2010; Tracy et al., 2009). People prone to authentic pride tend to be low in depression, trait anxiety, social phobia, aggression, hostility, and rejection sensitivity; and high in self-control, goal-engagement, relationship satisfaction, attachment security, dyadic adjustment, and social support. In contrast, people prone to hubristic pride are more impulsive, anxious, hostile, and inclined to engage in a range of other antisocial misbehaviors (e.g., drug use, petty crimes), and they report poorer relationship adjustment and social support.

Cultural Differences in Shame, Guilt, and Pride

Shame, guilt, and pride are emotions experienced in reference to self. To the extent that the nature of the self differs across cultures and nations (Markus and Kitayama, 1991), one might expect parallel differences in the experience and implications of self-conscious emotions. The research described thus far has been largely conducted in a Western context.

Most studies cross-cultural studies of the self-conscious emotions have compared Asian-Americans and Caucasian-Americans. Research consistently shows that Asian-Americans report more shame than their non-Asian-American counterparts, whereas cultural differences in guilt have been mixed (Lutwak et al., 1998; Tracy and Matsumoto, 2008). Regarding pride, there are clear cultural differences in the value attached to the emotion. Pride is a highly valued and sought out emotion in individualistic cultures (i.e., the United States, Australia, the Netherlands), whereas it is viewed as negative and undesirable in several collectivistic cultures (i.e., China, Spain, Taiwan; Eid and Diener, 2001; Mosquera et al., 2000).

Few studies have moved beyond a consideration of mean differences in proneness to shame, guilt, and pride, to consider the possibility of cultural differences in the *implications* of self-conscious emotions. Shame-proneness appears to be similarly associated with anger, depression, and anxiety among Asian-Americans and non-Asian-Americans, whereas the correlates of guilt are mixed. Two studies have examined cross-cultural continuities and discontinuities in the correlates of shame and guilt, comparing American-born children with Asian-born children living in their country of origin. Bear et al. (2009) reported a positive link between shame and anger among American children, but no such relationship among Japanese children. Comparing children residing in Japan, Korea, and the United States, Furukawa et al. (2012) found substantial group differences the propensity to experience self-conscious emotions. Japanese children scored highest on shame-proneness, Korean children scored highest on guilt-proneness, and the US children were highest on pride. Although it was hypothesized that shame would be less problematic among Japanese children relative to

those raised in Korea and the United States, because shame is more normative and would therefore be less painful in the self-critical Japanese culture, there were surprisingly few differences in the relationship of shame to aggression-related cognitions, emotions, and behavior.

New Directions in Research on the Self-Conscious Emotions

Much research has emerged on self-conscious emotions over the past two decades, but much work remains. Several promising new directions stand out, ensuring that the next two decades of research will be even more exciting than the developments described thus far.

First, researchers have become interested in the phenomena of 'vicarious' and group-based self-conscious emotions. Shame, guilt, embarrassment, and pride are typically experienced in reference to one's own attributes or behaviors. But people can also vicariously experience of self-conscious emotions owing to the actions of *other* individuals or groups. Some of the earliest research in this area focused 'empathic' or vicarious embarrassment (see Miller, 1996). More recently, investigators have examined 'vicarious' or 'group-based' shame and guilt (Lickel et al., 2007; Welten et al., 2012) which in many ways parallels findings regarding personal shame and guilt experiences. For example, as with personal guilt experiences, group-based guilt has been associated with empathy (Zebel et al., 2004) and a motivation to repair or make amends (Lickel et al., 2005). And as with personal shame experiences, vicarious group-based shame (but not guilt) has been linked to a desire to distance oneself from the shame-eliciting event (Lickel et al., 2005; Johns et al., 2005). Nonetheless, there are some indications that vicarious or group-based shame may have a 'kinder, gentler' side than personal shame. For example, under some circumstances, group-based shame appears to motivate a desire to change the image of the group in a proactive fashion (Lickel et al., 2006).

Second, much more research is needed on the phenomenon of 'guilt-tripping' and other efforts to use self-conscious emotions as a form of social control. Sharkey (1991) made important inroads in our understanding of 'intentional embarrassment.' To date, only a handful of studies have explicitly examined guilt induction (Baumeister et al., 1995; Vangelisti et al., 1991). These studies indicate that deliberate attempts to induce guilt occur relatively frequently, especially in the context of close relationships, and especially in response to real or perceived neglect. But many other questions remain. For example, what are the relative costs and benefits inducing shame, guilt, and embarrassment? How do those costs and benefits vary as a function of transgression, type of relationship, and personality characteristics of the inducee? Are some people more vulnerable than others to guilt (or shame or embarrassment) inductions?

Third, researchers have begun to explore context- or domain-specific shame and guilt. For example, researchers concerned with the psychology of eating disorders have assessed feelings of shame specifically in reference to one's body. 'Body shame' has been consistently associated with self-objectification and eating disorder symptoms (e.g., Tiggemann and Boundy, 2008). Other researchers have examined the nature and implications of trauma-related guilt cognitions, such as false beliefs about responsibility or preoutcome knowledge, which are reliably associated with symptoms of depression among trauma survivors (Kubany et al., 1995).

Fourth, exciting discoveries are being made about the psychobiological correlates of the self-conscious emotions. For example, increases in shame (but not guilt or general negative affect) coincide with increased proinflammatory cytokine activity (Dickerson et al., 2004). Experiences of shame have also been linked to elevated cortisol (Gruenewald et al., 2004) as well as distinct patterns of cardiovascular reactivity (Herrald and Tomaka, 2002). Regarding brain regions associated with self-conscious emotions, activation of medial prefrontal cortex and the posterior superior temporal sulcus have both been implicated in the experience of embarrassment and guilt, based on functional magnetic resonance imaging research (for a review, see Robins and Schriber, 2009).

Conclusions

Understanding the self-conscious emotions is critical to understanding the self. Psychological researchers have made much progress since the first comprehensive volume addressing the science of shame, guilt, embarrassment, and pride (Tangney and Fischer, 1995) but there remain many unanswered questions ripe for inquiry. Perhaps more than other emotions, the measurement of self-conscious emotions poses some real challenges. A number of measurement methods have been developed in recent years (see Robins et al., 2007; for a review), but the coming decades will no doubt see improvements in our ability to assess these emotions, enhancing this burgeoning area of research.

See also: Cross-Cultural Psychology; Culture and Emotion; Emotion and Expression; Emotion, Evolution of; Emotion, Neural Basis of; Emotions, Psychological Structure of; Indigenous Psychology; Shame and the Social Bond.

Bibliography

Baumeister, R.F., Stillwell, A.M., Heatherton, T.F., 1994. Guilt: an interpersonal approach. Psychological Bulletin 115, 243–267.

Baumeister, R.F., Stillwell, A.M., Heatherton, T.F., 1995. Interpersonal aspects of guilt: evidence from narrative studies. In: Tangney, J.P., Fischer, K.W. (Eds.), Self-Conscious Emotions: Shame, Guilt, Embarrassment, and Pride. Guilford Press, New York, pp. 255–273.

Bear, G.G., Uribe-Zarain, X., Manning, M.A., Shiomi, K., 2009. Shame, guilt, blaming, and anger: differences between children in Japan and the US. Motivation and Emotion 33, 229.

Carver, C.S., Sinclair, S., Johnson, S.L., 2010. Authentic and hubristic pride: differential relations to aspects of goal regulation, affect, and self-control. Journal of Research in Personality 44, 698–703.

Dickerson, S.S., Kemeny, M.E., Aziz, N., Kim, K.H., Fahey, J.L., 2004. Immunological effects of induced shame and guilt. Psychosomatic Medicine 66, 124–131.

Edelmann, R.J., 1981. Embarrassment: the state of research. Current Psychological Reviews 1, 125–138.

Eid, M., Diener, E., 2001. Norms for experiencing emotions in different cultures: inter- and intranational differences. Journal of Personality and Social Psychology 81, 869–885.

Furukawa, E., Tangney, J.P., Higashihara, F., 2012. Cross-cultural continuities and discontinuities in shame, guilt, and pride: a study of children residing in Japan, Korea and the U.S. Self and Identity 11, 90–113.

Gilbert, P., 1997. The evolution of social attractiveness and its role in shame, humiliation, guilt, and therapy. British Journal of Medical Psychology 70, 113–147.
Gruenewald, T.L., Kemeny, M.E., Aziz, N., Fahey, J.L., 2004. Acute threat to the social self: shame, social self-esteem, and cortisol activity. Psychosomatic Medicine 66, 915–924.
Harder, D.W., Cutler, L., Rockart, L., 1992. Assessment of shame and guilt and their relationship to psychopathology. Journal of Personality Assessment 59, 584–604.
Herrald, M.M., Tomaka, J., 2002. Patterns of emotion-specific appraisal, coping, and cardiovascular reactivity during an ongoing emotional episode. Journal of Personality and Social Psychology 83, 434–450.
Hosser, D., Windzio, M., Greve, W., 2008. Guilt and shame as predictors of recidivism: a longitudinal study with young prisoners. Criminal Justice and Behavior 35, 138–152.
de Hooge, I.E., Breugelmans, S.M., Zeelenberg, M., 2008. Not so ugly after all: when shame acts as a commitment device. Journal of Personality and Social Psychology 95, 933–943.
de Hooge, I.E., Zeelenberg, M., Breugelmans, S.M., 2007. Moral sentiments and cooperation: differential influences of shame and guilt. Cognition and Emotion 21, 1025–1042.
de Hooge, I.E., Zeelenberg, M., Breugelmans, S.M., 2010. Restore and protect motivations following shame. Cognition and Emotion 24, 111–127.
Johns, M., Schmader, T., Lickel, B., 2005. Ashamed to be an American? The role of identification in predicting vicarious shame for anti-Arab prejudice after 9-11. Self and Identity 4, 331–348.
Joireman, J., 2004. Empathy and the self-absorption paradox II: self-rumination and self-reflection as mediators between shame, guilt, and empathy. Self and Identity 3, 225–238.
Keltner, D., Buswell, B.N., 1996. Evidence for the distinctness of embarrassment, shame, and guilt: a study of recalled antecedents and facial expressions of emotion. Cognition and Emotion 10, 155–171.
Keltner, D., Buswell, B.N., 1997. Embarrassment: its distinct form and appeasement functions. Psychological Bulletin 122, 250–270.
Kim, S., Thibodeau, R., Jorgensen, R.S., 2011. Shame, guilt, and depressive symptoms: a meta-analytic review. Psychological Bulletin 137, 68–96.
Kitayama, S., Markus, H.R., Matsumoto, H., 1995. Culture, self, and emotion: a cultural perspective on 'self-conscious' emotion. In: Tangney, J.P., Fischer, K.W. (Eds.), Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride. Guilford Press, New York, pp. 439–464.
Kubany, E.S., Abueg, F.R., Owens, J.A., Brennan, J.M., Kaplan, A.S., Watson, S.B., 1995. Initial examination of a multidimensional model of trauma-related guilt: applications to combat veterans and battered women. Journal of Psychopathology and Behavioral Assessment 17, 353–376.
Leith, K.P., Baumeister, R.F., 1998. Empathy, shame, guilt, and narratives of interpersonal conflicts: guilt-prone people are better at perspective taking. Journal of Personality 66, 1–37.
Lewis, H.B., 1971. Shame and Guilt in Neurosis. International Universities Press, New York.
Lewis, M., 1992. Shame: The Exposed Self. Free Press, New York.
Lewis, M., Sullivan, M.W., Stanger, C., Weiss, M., 1989. Self-development and self-conscious emotions. Child Development 60, 146–156.
Lickel, B., Rutchick, A.M., Hamilton, D.L., Sherman, S.J., 2006. Intuitive theories of group types and relational principles. Journal of Experimental Social Psychology 42, 28–39.
Lickel, B., Schmader, T., Curtis, M., Scarnier, M., Ames, D.R., 2005. Vicarious shame and guilt. Group Processes and Intergroup Relations 8, 145–147.
Lickel, B., Schmader, T., Spanovic, M., 2007. Group-conscious emotions: implications of others' wrongdoing for identity and relationships. In: Tracy, J.L., Robins, R.W., Tangney, J.P. (Eds.), The Self-Conscious Emotions: Theory and Research. Guilford, New York, pp. 351–370.
Lindsay-Hartz, J., 1984. Contrasting experiences of shame and guilt. American Behavioral Scientist 27, 689–704.
Lutwak, N., Razzino, B.E., Ferrari, J.R., 1998. Self-perceptions and moral affect: an exploratory analysis of subcultural diversity in guilt and shame emotions. Journal of Social Behavior and Personality 13, 333–348.
Markus, H.R., Kitayama, S., 1991. Culture and the self: implications for cognition, emotion, and motivation. Psychological Review 98, 224–253.
Mascolo, M.F., Fischer, K.W., 1995. Developmental transformation in appraisals for pride, shame, and guilt. In: Tangney, J.P., Fischer, K.W. (Eds.), Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride. Guilford, New York, pp. 64–113.
Mosquera, P.M.R., Manstead, A.S.R., Fischer, A.H., 2000. The role of honor-related values in the elicitation, experience, and communication of pride, shame, and anger: Spain and the Netherlands compared. Personality and Social Psychology Bulletin 26, 833–844.
Miller, R.S., 1996. Embarrassment: Poise and Peril in Everyday Life. Guilford Press, New York.
Nelissen, R.M.A., Zeelenberg, M., 2009. Moral emotions as determinants of third-party punishment: anger, guilt, and the functions of altruistic sanctions. Judgment and Decision Making 4, 543–553.
O'Connor, L.E., Berry, J.W., Weiss, J., 2002. Guilt, fear, submission, and empathy in depression. Journal of Affective Disorders 71, 19–27.
Robins, R.W., Noftle, E.E., Tracy, J.L., 2007. Assessing self-conscious emotions: a review of self-report and nonverbal measures. In: Tracy, J.L., Robins, R.W., Tangney, J.P. (Eds.), The Self-Conscious Emotions: Theory and Research. Guilford Press, New York, pp. 443–467.
Robins, R.W., Schriber, R.A., 2009. The self-conscious emotions: how are they experienced, expressed, and assessed? Social and Personality Psychology Compass 3, 887–898.
Sharkey, W.F., 1991. Intentional embarrassment: goals, tactics, and consequences. In: Cupach, W.R., Metts, S. (Eds.), Advances in Interpersonal Communication Research—1991. Personal Relationships Interest Group, Normal, IL, pp. 105–128.
Sheikh, S., Janoff-Bulman, R., 2010. The "shoulds" and "should nots" of moral emotions: a self-regulatory perspective on shame and guilt. Personality and Social Psychology Bulletin 36, 213–224.
Stipek, D.J., Recchia, S., 1992. Self-evaluation in young children. Monographs of the Society for Research in Child Development 57 (1), R5–R83. Serial No. 226.
Stuewig, J., McCloskey, L.A., 2005. The relation of child maltreatment to shame and guilt among adolescents: Psychological routes to depression and delinquency. Child Maltreatment 10, 324–336.
Stuewig, J., Tangney, J.P., Heigel, C., Harty, L., McCloskey, L.A., 2010. Shaming, blaming, and maiming: functional links among the moral emotions, externalization of blame, and aggression. Journal of Research in Personality 44, 91–102.
Tangney, J.P., 1990. Assessing individual differences in proneness to shame and guilt: development of the self-conscious affect and attribution inventory. Journal of Personality and Social Psychology 59, 102–111.
Tangney, J.P., 1995. Shame and guilt in interpersonal relationships. In: Tangney, J.P., Fischer, K.W. (Eds.), Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride. Guilford Press, New York, pp. 114–139, 343–367.
Tangney, J.P., Fischer, K.W. (Eds.), 1995. Self-Conscious Emotions: Shame, Guilt, Embarrassment, and Pride. Guilford Press, New York.
Tangney, J.P., Miller, R.S., Flicker, L., Barlow, D.H., 1996a. Are shame, guilt and embarrassment distinct emotions? Journal of Personality and Social Psychology 70, 1256–1269.
Tangney, J.P., Stuewig, J., Mashek, D.J., 2007. Moral emotions and moral behavior. Annual Review of Psychology 58, 345–372.
Tangney, J.P., Stuewig, J., Mashek, D., Hastings, M., 2011. Assessing jail inmates' proneness to shame and guilt: feeling bad about the behavior or the self? Criminal Justice and Behavior 38, 710–734.
Tangney, J.P., Wagner, P., Gramzow, R., 1992. Proneness to shame, proneness to guilt, and psychopathology. Journal of Abnormal Psychology 101, 469–478.
Tangney, J.P., Wagner, P.E., Hill-Barlow, D., Marschall, D.E., Gramzow, R., 1996b. Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. Journal of Personality and Social Psychology 70, 797–809.
Thompson, R.J., Berenbaum, H., 2006. Shame reactions to everyday dilemmas are associated with depressive disorder. Cognitive Therapy and Research 30, 415–425.
Tibbetts, S.G., 2003. Self-conscious emotions and criminal offending. Psychological Reports 93, 101–126.
Tiggemann, M., Boundy, M., 2008. Effect of environment and appearance compliment on college women's self-objectification, mood, body shame, and cognitive performance. Psychology of Women Quarterly 32, 399–405.
Tracy, J.L., Cheng, J.T., Robins, R.W., Trzesniewski, K.H., 2009. Authentic and hubristic pride: the affective core of self-esteem and narcissism. Self and Identity 8, 196–213.
Tracy, J.L., Matsumoto, D., 2008. The spontaneous expression of pride and shame: evidence for biologically innate nonverbal displays. Proceedings of the National Academy of Sciences 105, 11655–11660.
Tracy, J.L., Robins, R.W., 2007. The psychological structure of pride: a tale of two facets. Journal of Personality and Social Psychology 92, 506–525.
Tracy, J.L., Robins, R.W., 2008. The nonverbal expression of pride: evidence for cross-cultural recognition. Journal of Personality and Social Psychology 94, 516–530.
Vangelisti, A.L., Daly, J.A., Rudnick, J.R., 1991. Making people feel guilty in conversations: techniques and correlates. Human Communication Research 18, 3–39.
Welten, S.C.M., Zeelenberg, M., Breugelmans, S.M., 2012. Vicarious shame. Cognition & Emotion 26, 836–846.
Zebel, S., Doojse, B., Spears, R., 2004. It depends on your point of view: implications of perspective-taking and national identification for Dutch collective guilt. In: Branscombe, N.R., Doosje, B. (Eds.), Collective Guilt: International Perspectives. Cambridge University Press, New York, pp. 148–168.

Self-Control Theory and Crime

George E Higgins and Dustin Osborne, University of Louisville, Louisville, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of Gottfredson and Hirschi's (1990) self-control theory. The article details the development. The development pertains to the role of parents and schools. Gottfredson and Hirschi argue that peer association is not an issue with their theory. This article discusses the issues surrounding this viewpoint. In addition, it discusses gender differences. Each of these topics is applied to criminal and deviant behavior.

Attempts to discern the causes of crime have occupied criminologists for many years. These attempts generally manifest themselves in new theories of crime, or more commonly in support of those that are already established. These theories have a variety of foci, ranging from characteristics unique to the individual to those that relate to the environment in which they reside. Although some theories are targeted at specific environments or forms of offending, recent developments can be considered to be more general in nature. That is, they purport to explain a range of crime, occurring in a variety of settings. One such theory, Gottfredson and Hirschi's General Theory of Crime (1990), suggests that this behavior occurs as a result of low self-control, and that levels of self-control are influenced early in life through the socialization process. This article will relay the core tenets of the theory, including the relationship between self-control and crime and the influence that parents and other institutions have on the development of self-control. It will then proceed to provide a brief overview of the supporting literature, and conclude with a discussion of potential policy implications arising from its propositions.

Gottfredson and Hirschi's General Theory of Crime

Whereas most criminological theories seek to detail the factors that contribute to crime, self-control theory seeks to understand what keeps the majority of society from participating in it. Gottfredson and Hirschi (1990) view such behavior as resulting from an inability to refrain from acting when opportunities to offend arise, simply due to the fact that these opportunities present the possibility of pleasure for the offender. In this sense, they view all individuals (even those who commit criminal acts) as being able to rationally calculate the potential pleasure and pain associated with an act, and in the event that the pleasure outweighs the costs, able to act upon their desires. Costs can take many forms, and include potential criminal sanctions, loss of job or status, and/or harm to personal relationships. Most individuals are able to comprehend the magnitude of these costs and refrain from committing the deviant act. However, some individuals seem to lack this ability. Gottfredson and Hirschi contend that this occurs due to a lack of self-control. In other words, some individuals are impulsive and shortsighted. These characteristics lead their rational processes to be bounded in a sense, meaning that the only factors that they weigh are those immediate to the situation. Since immediate benefits are much easier to comprehend than long-term costs, deviant acts become all the more likely for these individuals.

The Development of Self-Control

Gottfredson and Hirschi suggest that self-control is developed early in life, primarily through the socialization process. This process depends on a number of parties, ranging from parents (posited to be most important) to social institutions. Since parents are most responsible for the development of the child, their role is especially important in the socialization process. According to their conceptualization, four specific parental factors contribute to the development of self-control: attachment, behavior monitoring, behavioral evaluation, and corrective actions. Attachment is especially important due to the fact that without it, none of the other actions are likely to be taken. In essence, parents who bond with their children are more likely to take an interest in their development and behavior. This interest allows them to fulfill the second requirement, that behavior is monitored on a regular basis. Due to their young age and inexperience with societal norms, children lack an innate understanding of how to act in a proper manner. Parents are charged with observing their actions, and gaining an understanding of their overall behavior. This information-gathering process allows them to fulfill the third component, behavioral evaluation, and to take note of that which they do not deem proper. Once this behavior is noticed, corrective action can be taken in an attempt to show children their wrongdoings and prevent reoccurrences in the future. This process of observation and correction, if consistent, ultimately allows children to gain their own understanding of what is expected by society.

It is important to note that this four-part process must occur early in the child's development. According to Gottfredson and Hirschi, self-control is established in most instances by the age of 8 and few changes can occur after this stage. Thus, children with low self-control due to poor parenting practices are unlikely to alter their level at a later date even if parenting is improved. Children, due to impulsivity and shortsightedness, will be inclined to make bounded decisions, and thus be at greater risk for antisocial or criminal behavior than their peers who have parents who took an interest in their socialization.

While parents have the unique ability to shape a child's behavior due to the sheer amount of time that they spend with

them, they are not alone in the socialization process. Schools and other social entities (e.g., churches) are in a unique position to frame behavior as well. Once children enter into the educational system – whether it be in preschool or in kindergarten – they spend a significant portion of their time interacting with teachers and other school staff. This time involves those individuals taking a unique interest in the child. This interest manifests itself not only in providing quality education but also in ensuring proper behavior. Since effective classrooms depend on cooperation from students, teachers are charged with observing and regulating behavior. Improper behavior – or that antisocial in nature – is quickly addressed through means of punishment available within the school. Furthermore, information can be transmitted to the misbehaving child's parents in order for them to take external action. Ideally, this 'team effort' of behavioral observation and correction should allow the socialization process to take hold much more easily than if just the parents were responsible for it.

Unfortunately, not all school settings are conducive to behavior regulation. Schools featuring poor management and high levels of disorder are unlikely to assist in the process. On the other hand, even those schools and teachers who take appropriate interest in the child are not necessarily guaranteed to have parental cooperation. Parents who have not bonded appropriately with the child – and who have little interest in their away-from-home behavior – are unlikely to take corrective actions. Thus, while the socialization process is undertaken by teachers and other staff, it is unlikely to result in the development of high self-control since the parents are not reinforcing the norms at home.

A General Theory of Crime

As mentioned within the introduction, Gottfredson and Hirschi's theory is considered a general theory of crime. In other words, they posit that it is applicable to all forms of offending, as well as to all individuals and groups. Furthermore, they believe low self-control to be the primary cause of offending, with other factors culminating in it (e.g., poor parenting practices). This is unique to the extent that a review of the most popular theories of crime reveals a plethora of potential factors influencing the decision to commit deviant acts. Thus, the potential for one theory to have applicability to all criminal decision making offers a great deal of promise to the field.

In addition to this potential benefit, the theory is designed in a manner to allow testing in any number of settings and with nearly any form of data. Many theories require longitudinal data in order to show some manner of causality, whereas self-control theory can be tested using cross-sectional data. Since self-control is deemed to be stable throughout the life course (following the age of 8), queries regarding the individual's impulsivity, short-sightedness, and preference for easy, immediate gratification allow researchers to gather the information necessary to tie their level of self-control to any criminal acts that they may have committed. The development of a scale of self-control by Grasmick et al. (1993) has greatly assisted in this regard, and continues to be used within the field today.

Peer Groups and Low Self-Control

Having now briefly discussed the development of low self-control and Gottfredson and Hirschi's conceptualization of a general theory of crime, attention is turned to discussing its ramifications. As mentioned, low self-control manifests itself in deviant acts due to a shortsightedness on the part of the afflicted individual. These acts begin relatively early in life, though they may be relatively minor at first (e.g., acting out within the school setting). It is the introduction to deviant peer groups that is likely to set in motion the path to more serious acts. While some theories view these groups as the primary cause of offending – as they introduce individuals to the proper techniques required for it – Gottfredson and Hirschi suggest that it is low self-control that leads individuals to join such groups in the first place. In other words, those with low self-control are drawn to similar individuals due to the fact that they view them as fun and willing to take chances much as they are.

Combining these individuals into one peer group may lead to a reinforcement of sorts, and make the commission of crime easier. The suggestion of other theoretical premises that ideas are learned within these groups is not dismissed by self-control theory. Nor is the fact that strength in numbers may exacerbate offending. In fact, both these factors are viewed as being in play. However, it is stressed that this alone is not enough for individuals to embark upon a life of criminality. Their low level of self-control makes them more likely to associate with this type of group, just as it makes them more likely to commit deviant acts. The combination of both these likelihoods only serves to increase the potential for crime.

The Role of Gender in Self-Control Theory

As criminological theory has developed, attention has been directed at the role that gender plays in offending. That is, it is suggested by many that differences do exist between males and females. Gottfredson and Hirschi have received some criticism for their perceived slight to this notion due to the fact that they argue that the same socialization processes are at play in the criminality of both girls and boys. However, the theory does address potential gender issues by suggesting that girls are more likely to feature higher levels of self-control, and are thus less likely to participate in deviant behavior. Research indicates that girls are significantly less likely to offend throughout the lifecourse, reinforcing this proposition. The question must thus be asked, why do girls feature these higher levels? Gottfredson and Hirschi suggest that parents take a greater interest in the behavior of these girls than they do in their male counterparts. This is due to the fact that they are more concerned with potential misbehavior by them and the consequences that such misbehavior might bring. Attachment is thus much higher, and the socialization process is facilitated by it. Having now discussed the core components of self-control theory, attention is now turned to support for theory, with specific focus on recent research within the field.

Support for Self-Control Theory

Over three decades have passed since Gottfredson and Hirschi put froth their General Theory of Crime. During this time a rather large number of researchers have tested their core tenets, with a great deal of support emerging. Perhaps the most exhaustive overview of this research can be found in

a meta-analysis published by Pratt and Cullen (2000). The researchers found that while Social Learning Theory explained a portion of deviant behavior, low self-control constituted a significant predictor across a wide range of samples and behaviors. Furthermore, this predictive power existed regardless of the way in which self-control was measured, offering additional support for the theory. (An additional discussion of the influence of self-control regardless of measurement can be found in the work of Tittle et al. (2003).) In the time since this analysis was published, a plethora of new studies have emerged, offering both additional support and unique insights into the role that self-control plays in deviant acts. These studies can be grouped based on the core tenets of the theory: the role of parenting, development of low self-control, gender differences, and the influence of low self-control on peer group affiliation and deviant behavior. Empirical support for each of these tenets will now be briefly reviewed.

Support for the Role of Parenting

As discussed, parenting plays the key role in the socialization process, and thus in the development of higher levels of self-control. This notion has been tested time and again, and has shown itself to be rather well-supported (Hay, 2001; Gibbs et al., 1998; Gibbs et al., 2003). It appears as if those parents who bond with their children – and thus monitor and control their behavior – facilitate in the development of higher levels of self-control (Finkenauer et al., 2005). Trajectory analysis has further revealed that levels of self-control may increase linearly through the years of early development, and shares a significant relationship with parenting at each point in time (Vazsonyi and Huang, 2010). Thus, consistent application of the monitoring and corrective processes detailed earlier in this piece appears to be incredibly important in ensuring higher levels of self-control once the child reaches the age of 8.

Some research has been directed at determining if any additional factors beyond parenting and other social institutions play a part in the development of low self-control. This research has specifically focused upon community context, and the role that certain neighborhood factors (e.g., poor economic conditions) play upon the development of the individual. For example, Wikstrom and Sampson (2002) argue that the collective efficacy of a community plays a role in behavioral identification and thus correction. This suggestion has received some support (Pratt et al., 2004). However, when controlling for individual characteristics it appears that the role of community context may be a nonsignificant predictor – further supporting the role of parents (Gibson et al., 2010).

Of unique interest is recent evidence suggesting that levels of self-control may be transmitted from parents to child (Boutwell and Beaver, 2010). This transmission is posited by some to be genetic in nature, thus making low levels of self-control less reliant upon socialization than upon biological predisposition (Beaver et al., 2010; Wright et al., 2008). With that said, this postulation is to date understudied, and the potential exists that both socialization and biological factors may work together in the development of self-control (Boutwell et al., 2013).

Support for the Stability of Self-Control

Recall that self-control was posited by Gottfredson and Hirschi (1990) to be fully developed by the age of 8 and remain relatively stable throughout the lifecourse. This proposition has been met with mixed results. Supporting their conceptualization, it has been found that a significant proportion of people remain relatively stable in their level of self-control after the age of 8 (Higgins et al., 2009). However, this does not apply to all, as others have found that a smaller group has the potential for variation well into adolescence (Hay and Forrest, 2006; Hay et al., 2010). This variation can feature improvement to higher levels of self-control, or regression to lower levels. Variation can occur for a variety of reasons, including victimization. For example, it has been found that individuals who experience victimization have the potential for their levels of self-control to be reduced in the near term due to the strain that it imposes (Agnew et al., 2011). Peer association may play a role as well, as information learned within peer groups may lead to changes in control over time, especially for those experiencing relatively low levels of self-control to begin with (Jennings et al., 2013). Ultimately, it may be useful to view self-control as relatively stable, with the potential for certain factors to produce variation over time.

Support for Gender Differences

Gottfredson and Hirschi (1990) suggest that gender differences exist within their theory to the extent that girls are more likely to be properly socialized through parental monitoring than are boys. Research on the link between self-control and deviant behavior in general suggests that boys are more likely to feature low self-control than girls, partially supporting this hypothesis (Shekarkhar and Gibson, 2011). However, recent investigations into impulsivity reveal that while girls are more likely to feature these higher levels, better parental monitoring may not be the cause. For example, Chapple and Johnson (2007) suggest that boys simply experience differential effects in relation to less constrained familial environments. In other words, when monitoring and control are not present, boys are more likely to develop lower levels of control than their female counterparts. With that said, an understanding of gender differences on the development of self-control remains somewhat fleeting. The suggestion has been raised that the way in which girl's self-control is measured and modeled needs improvement (Chapple et al., 2010; Higgins, 2004). Thus, until a better understanding of all the variables that influence self-control is achieved, gender differences are likely to remain a mystery of sorts.

Support for Peer Groups and Deviant Behavior

Gottfredson and Hirschi's (1990) conceptualization of self-control and its influence on behavior has been applied to a range of deviant acts, and is generally met with great support. Recall the proposition that peer groups are not the origin of deviant behavior, but that those with low self-control are drawn to groups composed of similar individuals. Research suggests that this phenomenon does in fact exist (Wolfe and Higgins, 2009; McGloin and Shermer, 2009). Furthermore, it may occur

as a result of their rejection – due to their predisposition – by more traditional peer groups (Chapple, 2005). In other words, the inability of these individuals to associate with nondeviant peers leads them to take refuge in groups of others with low self-control. Once in these groups, the opportunity for offending is only magnified as new avenues are presented to them.

While a complete examination of all forms of antisocial actions explained by low self-control is beyond the scope of this article, it is important to note that it can be applied to nearly any type of behavior. Pratt and Cullen's (2000) meta-analysis of research published throughout the 1990s revealed strong initial support for this notion, as the studies included varied in the behavior that they focused upon. In the time period since, it has been applied to everything from violent offending to imprudent behavior (i.e., academic fraud, binge drinking) (Piquero et al., 2005; Reisig and Pratt, 2011). Furthermore, it has application in relation to relatively modern forms of offending such as software piracy (Higgins et al., 2008). Perhaps most supportive of the relationship between low self-control and deviant behavior is the work of Vazsonyi et al. (2001). Empirical analysis of data from four countries revealed that low self-control played a similar role in the decision to offend across all individuals. Furthermore, it predicted deviant behavior in relation to a range of offenses, from vandalism to drug use. Taken as a whole, it appears that the relationship between self-control and antisocial propensities is rather strong in nature.

Conclusion

In summation, support for the role of self-control in offending seems rather robust, as the good majority of Gottfredson and Hirschi's propositions are backed by the literature. Parenting and schools do seem to play key roles in the development of self-control, with poor monitoring resulting in lower levels of it. Though some variation exists in levels of self-control throughout adolescence, it appears that for the majority of individuals it is relatively stable past the age of 8. Furthermore, low self-control can be tied to a range of deviant acts, and is in most instances shown to be one of the primary causes of such behavior. With that said, much work is still to be done. Gender differences in the development of self-control remain poorly understood, and recent suggestions that biological factors may be at play in the transmission of it from parent to child are deserving of further consideration.

In addition, Hirschi (2004) has since attempted to reconceptualize his stance on control, suggesting that bonds and human agency (rational calculation) play a greater role than previously suggested. Furthermore, he contends that self-control is not so much a personality trait of the individual that remains constant over time, but rather something to be considered in the decision-making process. This development is likely to lead to a reanalysis of the other facets of the theory, and include new questions presented in relation to gender, stability, and the role of parenting and social institutions. At the least it should present researchers with new opportunities for investigation, offering the potential for our understanding of the relationship between control and offending to be further developed in the coming years.

Bibliography

Agnew, R., Scheuerman, H., Grosholz, J., Isom, D., Watson, L., Thaxton, S., 2011. Does victimization reduce self-control? A longitudinal analysis. Journal of Criminal Justice 39 (2), 169–174.

Beaver, K.M., Ferguson, C.J., Lynn-Whaley, J., 2010. The association between parenting and levels of self-control: a genetically informative analysis. Criminal Justice and Behavior 37 (10), 1045–1065.

Boutwell, B.B., Beaver, K.M., 2010. The intergenerational transmission of low self-control. Journal of Research in Crime and Delinquency 47 (2), 174–209.

Boutwell, B.B., Franklin, C.A., Barnes, J.C., Tamplin, A.K., Beaver, K.M., Petkovsek, M., 2013. Unraveling the covariation of low self-control and victimization: a behavior genetic approach. Journal of Adolescence 36 (4), 657–666.

Chapple, C.L., Vaske, J., Hope, T.L., 2010. Sex differences in the causes of self-control: an examination of mediation, moderation, and gendered etiologies. Journal of Criminal Justice 38 (6), 1122–1131.

Chapple, C.L., Johnson, K.A., 2007. Gender differences in impulsivity. Youth Violence and Juvenile Justice 5 (3), 221–234.

Chapple, C.L., 2005. Self-control, peer relations, and delinquency. Justice Quarterly 22 (1), 89–106.

Finkenauer, C., Engels, R.C., Baumeister, R.F., 2005. Parenting behaviour and adolescent behavioural and emotional problems: the role of self-control. International Journal of Behavioral Development 29 (1), 58–69.

Gibson, C.L., Sullivan, C.J., Jones, S., Piquero, A.R., 2010. "Does it take a village?" Assessing neighborhood influences on children's self-control. Journal of Research in Crime and Delinquency 47 (1), 31–62.

Gibbs, J.J., Giever, D., Higgins, G.E., 2003. A test of Gottfredson and Hirschi's general theory using structural equation modeling. Criminal Justice and Behavior 30 (4), 441–458.

Gibbs, J.J., Giever, D., Martin, J.S., 1998. Parental management and self-control: an empirical test of Gottfredson and Hirschi's general theory. Journal of Research in Crime and Delinquency 35 (1), 40–70.

Gottfredson, M., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Grasmick, H.G., Tittle, C.R., Bursik, R.J., Arneklev, B.J., 1993. Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. Journal of Research in Crime and Delinquency 30 (1), 5–29.

Hay, C., Forrest, W., 2006. The development of self-control: examining self-control theory's stability thesis. Criminology 44 (4), 739–774.

Hay, C., 2001. Parenting, self-control, and delinquency: a test of self-control theory. Criminology 39, 707–736.

Hay, C., Meldrum, R., Forrest, W., Ciaravolo, E., 2010. Stability and change in risk seeking: investigating the effects of an intervention program. Youth Violence and Juvenile Justice 8 (2), 91–106.

Higgins, G.E., 2004. Gender and self-control theory: are there differences in the measures and the theory's causal model? Criminal Justice Studies 17, 33–55.

Higgins, G.E., Jennings, W.G., Tewksbury, R., Gibson, C.L., 2009. Exploring the link between low self-control and violent victimization trajectories in adolescents. Criminal Justice and Behavior 36 (10), 1070–1084.

Higgins, G.E., Wolfe, S.E., Marcum, C.D., 2008. Digital piracy: an examination of three measurements of self-control. Deviant Behavior 29 (5), 440–460.

Hirschi, T., 2004. Self-control and crime. In: Baumeister, R.F., Vohs, K.D. (Eds.), Handbook of Self-regulation: Research, Theory, and Applications. Guilford Press, New York, pp. 537–552.

Jennings, W.G., Higgins, G.E., Akers, R.L., Khey, D.N., Dobrow, J., 2013. Examining the influence of delinquent peer association on the stability of self-control in late childhood and early adolescence: toward an integrated theoretical model. Deviant Behavior 34 (5), 407–422.

McGloin, J.M., Shermer, L.O., 2009. Self-control and deviant peer network structure. Journal of Research in Crime and Delinquency 46 (1), 35–72.

Piquero, A.R., MacDonald, J., Dobrin, A., Daigle, L.E., Cullen, F.T., 2005. Self-control, violent offending, and homicide victimization: assessing the general theory of crime. Journal of Quantitative Criminology 21 (1), 55–71.

Pratt, T.C., Cullen, F.T., 2000. The empirical status of Gottfredson and Hirschi's general theory of crime: a meta-analysis. Criminology 38 (3), 931–964.

Pratt, T.C., Turner, M., Piquero, A.R., 2004. Parental socialization and community context: a longitudinal analysis of the structural sources of self-control. Journal of Research in Crime and Delinquency 41, 219–243.

Reisig, M.D., Pratt, T.C., 2011. Low self-control and imprudent behavior revisited. Deviant Behavior 32 (7), 589–625.

Shekarkhar, Z., Gibson, C.L., 2011. Gender, self-control, and offending behaviors among Latino youth. Journal of Contemporary Criminal Justice 27 (1), 63–80.

Tittle, C.R., Ward, D.A., Grasmick, H.G., 2003. Self-control and crime/deviance: cognitive vs behavioral measures. Journal of Quantitative Criminology 19 (4), 333–365.

Vazsonyi, A.T., Huang, L., 2010. Where self-control comes from: on the development of self-control and its relationship to deviance over time. Developmental Psychology 46 (1), 245.

Vazsonyi, A.T., Pickering, L.E., Junger, M., Hessing, D., 2001. An empirical test of a general theory of crime: a four-nation comparative study of self-control and the prediction of deviance. Journal of Research in Crime and Delinquency 38 (2), 91–131.

Wikstrom, P.O., Sampson, R.J., 2002. Social mechanisms of community influences on crime and pathways in criminality. In: Lahey, B.B., Moffitt, T.E., Capsi, A. (Eds.), The Causes of Conduct Disorder and Serious Juvenile Delinquency. Guilford, New York, pp. 118–152.

Wright, J., Beaver, K., Delisi, M., Vaughn, M., 2008. Evidence of negligible parenting influences on self-control, delinquent peers, and delinquency in a sample of twins. Justice Quarterly 25 (3), 544–569.

Wolfe, S.E., Higgins, G.E., 2009. Explaining deviant peer associations: an examination of low self-control, ethical predispositions, definitions, and digital piracy. Western Criminology Review 10, 43–55.

Self-Determination Theory

Edward L Deci and Richard M Ryan, University of Rochester, Rochester, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.L. Deci, volume 11, pp. 7886–7888, © 2001, Elsevier Ltd.

Abstract

Self-determination Theory (SDT) is a motivational theory of personality, development, and social processes that examines how social contexts and individual differences facilitate different types of motivation, especially autonomous motivation and controlled motivation, and in turn predict learning, performance, experience, and psychological health. SDT proposes that all human beings have three basic psychological needs – the needs for competence, autonomy, and relatedness – the satisfaction of which are essential nutrients for effective functioning and wellness. Satisfaction of these basic needs promotes the optimal motivational traits and states of autonomous motivation and intrinsic aspirations, which facilitate psychological health and effective engagement with the world.

Self-determination Theory (SDT) is a theory of human motivation that examines a wide range of phenomena across gender, culture, age, and socioeconomic status. As a motivational theory, it addresses what energizes people's behavior and moves them into action, as well as how their behavior is regulated in the various domains of their lives. SDT's explanations are focused at the psychological level (rather than the sociological or physiological levels), thus using human perceptions, cognitions, emotions, and needs as predictors of regulatory, behavioral, developmental, and experiential outcomes (e.g., Ryan and Deci, 2000b).

Central to the theory is the important distinction between two types of motivation – *autonomous motivation* and *controlled motivation*. Traditionally, motivation theories have treated motivation as a unitary concept focusing just on the total amount of motivation people have for behaviors in order to predict how vigorously they will engage in those behaviors, and many contemporary theories of motivation still do. SDT, in contrast, has always put its primary emphasis on the types of motivation people have for various behaviors. The theory maintains that, although knowing the amount of motivation people have for behaviors may allow one to predict the amount or quantity of the behaviors they will exhibit, assessing types of motivation is necessary for predicting the quality and maintenance of those behaviors.

When people are autonomously motivated, they act with a full sense of willingness and volition, wholly endorsing that which they are doing because they find it either interesting and enjoyable, or consistent with their deeply held, integrated values. Autonomous motivation is typically accompanied by the experience of positive affect, flexibility, and choice. In contrast, when people's motivation is controlled, they act out of coercion, seduction, or obligation. They tend to experience pressure and compulsion, rather than concurrence and choice. Much of the self-determination research has examined either (1) antecedents of these types of motivation, at the developmental or the situation levels, or (2) the concomitants and consequences of the different types of motivation. We address these two types of motivation and their various subtypes in turn, although first we speak briefly about the philosophical assumptions upon which self-determination was built.

The Meta-Theory of SDT

Within psychology, the meta-theoretical assumptions underlying different theories tend to fall along a continuum from mechanistic to organismic. At one end, mechanistic meta-theories assume that humans are passive and thus require some force operating on them to move them into action. Classic theories of this sort include operant theory (Skinner, 1971), which assumes that external reinforcers, such as contingently provided food and water, as well as their derivatives such as money, are what move people into action. When reinforcers are provided after a person emits a behavior in the presence of a contingency, the reinforcer will strengthen the association between the contingency and behavior, making the behavior more likely to occur the next time the contingency is present.

At the other end of the continuum, organismic meta-theories assume that humans are by nature active organisms who can motivate themselves to act on their environment and elaborate and strengthen their understanding and behavioral repertoire. Piaget's (1971) theory of cognitive development is an example of a classic theory built upon organismic assumptions. This approach assumes that people are naturally inclined to expand their cognitive, knowledge structures by seeking stimuli that are optimally discrepant from their existing structures. The assimilation schema is the means through which these new schema are developed, and the organization principal is the process that integrates the new schema into people's existing structures.

Many current theories, such as the social cognitive and social learning theories (e.g., Bandura, 1996) fall somewhere between the ends of the continuum with some aspects of the theories reflecting somewhat mechanistic, and others reflecting somewhat organismic, meta-theoretical assumptions. However, the primary lacuna in these theories is the failure to recognize that people have an innate developmental tendency – that is, that they inherently act to elaborate themselves and to integrate these new experiences. Simply stated, from the organismic perspective, development is not something the environment does to people, although of course the environment does affect people's development; rather development is something that people do for themselves by acting on stimuli either that they seek out or that are thrust upon them.

Self-determination theory (SDT) is decidedly organismic at its core. It assumes that humans are active, working to integrate new material into their own sense of self, but also, and importantly, that the environment can either provide nutrients for this integrative process, or can disrupt and impair the process. Thus, it is the dialectic between the active, growth-oriented organism and the social context that is the central explanatory nexus for motivation, behavior, and development. The degree to which people's motivation is autonomous versus controlled, at either the developmental or situation level, depends on the degree to which the active integrative process functions more versus less successfully, in part depending on whether the social environment supports versus thwarts the integration.

Autonomous Motivation

The early motivation research that led directly to SDT differentiated intrinsic motivation from extrinsic motivation (e.g., Deci, 1971). Intrinsic motivation means people are engaging in an activity because they find it interesting, enjoyable, or fun. The play of young children is a perfect example of intrinsically motivated behavior. The children are active and engaged in a very natural way. They will often be working out some inner agenda, such as expressing feelings, albeit without any deliberate intention to do so. They are simply doing what they find interesting to do, and in the process they are learning and growing. Adults are also intrinsically motivated for some activities, typically their leisure-time pursuits. Because intrinsic motivation is a natural internal motivation involving interest and enjoyment, it is not necessary to motivate people to do what they find intrinsically interesting. They simply do those behaviors and this intrinsic motivation is the prototype of autonomous motivation.

Intrinsic motivation is often discussed in contrast to extrinsic motivation. The latter type of motivation involves a contingency between the target behavior and some separable consequence desired by the individual. What are referred to as reinforcers in operant theory can be thought of as extrinsic motivators. People are often extrinsically motivated by the pursuit of rewards such as money or prizes, the avoidance of noxious stimuli, or the desire for social approval. Colloquially, the classic extrinsic motivators are the 'carrot and stick.'

Intrinsic and Extrinsic Motivation

Early intrinsic motivation research examined the effects of extrinsic rewards on intrinsic motivation. At issue is whether giving extrinsic rewards to people who are intrinsically motivated for an activity would enhance their intrinsic motivation for the activity. That of course would be desirable. Perhaps, however, the two are additive in which case the motivations would not affect each other but would add together to form total motivation. That too would be a good outcome. But the third option is that there is a negative interactive effect between extrinsic rewards and intrinsic motivation. That would mean that giving extrinsic rewards to someone doing an interesting activity would actually diminish the person's intrinsic motivation for the activity.

The initial research indicated that in fact extrinsic rewards do undermine intrinsic motivation (Deci, 1971). People who did an interesting activity to get a monetary reward were subsequently less intrinsically motivated than others who did the same activity without getting the reward. This was a very controversial finding because it highlighted negative effects of rewards, which is no doubt the archetype of motivation in most people's minds, and of course in the behavioral psychologies. So, in the subsequent decades there were many additional tests of this phenomenon. A meta-analysis of more than 100 experiments confirmed that tangible rewards do in fact undermine intrinsic motivation for a behavior, especially if the rewards are contingent on the behavior, expected while doing it, and relatively salient (Deci et al., 1999). On the other hand, the research and meta-analysis also showed that positive feedback enhanced intrinsic motivation. Additional experiments showed, as summarized by Ryan and Deci (2000a), that threats of punishment, deadlines, evaluations, and surveillance all undermined intrinsic motivation, whereas providing people with choice, as well as acknowledging their feelings and perspectives, tended to enhance their intrinsic motivation.

In drawing these findings together and providing a theoretical account, Deci and Ryan (1985) posited that integral to intrinsic motivation are two fundamental psychological needs – that is, the needs for autonomy and competence. Stated differently, people have inherent psychological needs, just as they have basic physiological needs (e.g., oxygen, food, and water), and satisfaction versus thwarting of the basic psychological needs can have a range of positive versus negative consequences. Specifically, people may become dependent on and controlled by external events such as tangible rewards, threats, deadlines, and surveillance, thus having their need for autonomy thwarted and their intrinsic motivation undermined by these events. In contrast, choice and the acknowledgment of their internal perspective have been found to increase people's sense of autonomy, thus enhancing their intrinsic motivation. Concerning feedback, positive feedback tends to affirm people's sense of effectance, thus satisfying their need for competence and enhancing their intrinsic motivation; whereas, negative feedback, which has been found to diminish intrinsic motivation, is thought to have its effects by thwarting the need for competence.

Internalization of Motivation

The fact that the most typical extrinsic motivators have been reliably found to decrease intrinsic motivation – humans' natural, inherent type of autonomous motivation – raises the question of whether it is possible for people to be autonomous while being extrinsically motivated. Ryan et al. (1985) addressed that issue with the concept of internalization, which had been an important concept in developmental psychology for many years, and they suggested that people tend to internalize material endorsed by significant others in order to satisfy a basic psychological need for relatedness. The Ryan et al. idea was that extrinsic contingencies, which are external to people, could be taken in by the individuals and integrated into their sense of self. If that were to happen, people could behave from their own sense of self and thus be autonomous with respect to motivations that had originally been external.

However, the researchers pointed out that internalization, which is a natural part of the integrative process, may not always function wholly effectively so motivations sometimes get only partially internalized and thus not fully integrated. Therefore, they suggested, internalization can be understood in terms of a continuum of autonomy, in which the more fully some value or regulation is internalized the more the accompanying behavior will be enacted autonomously. They discussed four types of extrinsic motivation that resulted from different degrees of internalization.

External regulation is the classic type of extrinsic motivation. People behave because of external contingencies that have not been internalized. Thus, external regulation is the least autonomous and most controlled form of extrinsic motivation. *Introjected regulation* results from people having partially internalized an extrinsic motivation – that is, having taken it in but not really accepted it as their own. Introjection includes being motivated by contingent self-esteem, guilt, or ego-involvement. Introjected regulation has been found empirically to be accompanied by experiences and consequences similar to those associated with external regulation. Thus, although this type of extrinsic motivation is internal to the person, it is still quite controlling. A more fully internalized form of extrinsic motivation is referred to as *identified regulation* because it involves people identifying with the personal value and importance of the behavior for themselves and thus accepting it as their own. Subsequently, they will regulate themselves for related behaviors relatively autonomously. Finally, the fullest type of internalized extrinsic motivation is labeled *integrated regulation*. It involves people having integrated new identifications with other aspects of their own integrated sense of self – that is, with other identifications, values, and needs. With integrated regulation, people act with a full sense of volition and choice. Empirically, identified regulation is more closely related to integrated regulation than it is to introjected regulation. Thus, identified and integrated regulations, as well as intrinsic motivation, are all considered relatively autonomous forms of motivation.

With the elaboration of extrinsic motivation in terms of the degree of internalization, and thus of autonomy, it became clear that the distinction between autonomous and controlled motivation was the most useful and appropriate as the primary distinction in SDT. Accordingly, autonomous motivation comprises external and introjected forms of extrinsic motivation, whereas controlled motivation comprises identified and integrated forms of extrinsic motivation, along with intrinsic motivation. In this, one sees that some types of extrinsic motivation (identified and integrated) are relatively autonomous along with intrinsic motivation and one type of internal motivation (introjected) is relatively controlled. As such, neither the intrinsic–extrinsic distinction, nor the internal–external distinction works as cleanly and effectively as the autonomous-controlled distinction.

Autonomy across Cultures

Various psychologists have argued that the concept of autonomy is a Western concept that is not relevant to Eastern cultures (e.g., Iyengar and DeVoe, 2003); however, numerous studies have shown that the experience of autonomy is also essential in various Eastern cultures such as South Korea, China, and Japan. For example, Chirkov et al. (2003) did a study in Korea, Russia, Turkey, and the United States, finding that when people enacted behaviors consistent with either individualism or collectivism and did so autonomously, they evidenced high well-being, but when their enactment was prompted by controlled motivation, they evidenced poor well-being, regardless of the culture or the values. In other words, autonomy is necessary for a high level of psychological wellness in Eastern as well as Western cultures.

Amotivation

Clearly, both autonomous and controlled motivations are types of motivation, so to whatever degree people have one or the other or both of these, the people will be motivated. In contrast, the concept of *amotivation* refers to people having no intentionality or motivation. People tend to be amotivated for a behavior when they do not feel competent to do it or when they do not value the outcomes that are likely to follow from the behaviors.

Many motivation theories use as their primary distinction: being motivated versus unmotivated. SDT, however, has a tripartite differentiation of autonomous motivation, controlled motivation, and amotivation. Considerable research has used questionnaires that assess these concepts at the levels of specific behaviors, such as stopping tobacco use, or of domains, such as doing schoolwork. However, the tripartite conceptualization has also been examined at the more general personality level.

Causality Orientations

The concept of general causality orientations refers to three individual difference variables related to people's understanding of the causality for their behaviors and the degrees to which, motivationally, they are generally oriented in these ways. The three causality orientation dimensions are the autonomous orientation, the controlled orientation, and the impersonal orientation (Deci and Ryan, 1985). When people are high in the autonomous orientation, they tend to focus on information in the environment and within themselves that they can use in making choices, and they tend to have a high level of autonomous motivation. When people are high in the controlled orientation, they tend to focus on controls and pressures in the environment and within themselves that tell them what they should do, and to a substantial degree their behavior is controlled. When people are high in the impersonal orientation, they tend to focus on cues in the environment and within themselves that signify their incompetence and inability to obtain desired outcomes, and they tend to be amotivated a good deal of the time.

Each person has each of these orientations to some degree, so people are not categorized as being one type of person or another. Rather, the three orientations are all operative to differing degrees, and each orientation influences some of their behaviors and experiences. The autonomous orientation is related to self-esteem and self-actualization; the controlled

orientation is related to public self-consciousness and the type-A coronary-prone behavior pattern; and the impersonal orientation is related to self-derogation and depressive symptoms.

Basic Psychological Needs

Earlier we introduced the concept of fundamental psychological needs, which refers to essential, universal nutrients for psychological health and well-being. We mentioned that satisfying the needs for competence and autonomy appeared to be particularly important for maintaining intrinsic motivation, and also that people tend to internalize extrinsic motivation in order to experience satisfaction not only of these needs but also the need for relatedness.

Indeed, the concept of basic psychological needs is central to SDT because it specifies the nutrients that are essential in the environment to support and facilitate people becoming more autonomously motivated, experiencing greater psychological and physiological wellness, and performing more effectively (Gagné and Deci, 2005; Ryan et al., 2008b). Many dozens of studies, summarized in a recent meta-analysis (Ng et al., 2012) have confirmed that, across cultures, genders, and circumstances, when people experience satisfaction of the three basic psychological needs, they do indeed evidence greater health and well-being.

The importance of supports for psychological need satisfaction has been shown concurrently and developmentally. First, some studies have shown that when the current social environment supports one or more of the basic needs, people tend to be more autonomously motivated in that situation. Examples of this are the experiments showing that both choice and acknowledging people's feelings in a particular situation enhanced their intrinsic motivation for the target activity at that time (Deci et al., 1999). When environments supported people's autonomy, they became more intrinsically motivated. Similarly, an experiment (Deci et al., 1994) showed that supports for autonomy in a particular situation facilitated internalization and integration in that situation.

Second, the promotion of autonomous motivation and wellness has been shown developmentally. For example, studies have shown that when the social contexts of either homes or classrooms were autonomy supportive of young students, the students tended, over time, to develop stronger identifications with the importance of doing schoolwork – that is, they internalized this value and regulation more fully (Grolnick and Ryan, 1989). As noted already, the concept of general causality orientations concerns people's general individual differences with regard to autonomous and controlled motivation and amotivation. This concept of causality orientations is viewed as a developmental outcome – that is, it is said to result from the mix of supporting versus thwarting of the basic psychological needs during one's developmental years. When all of the needs are satisfied over time, in homes, schools, and elsewhere, while children are growing up, they tend to develop a relatively strong autonomy orientation. When the competence and relatedness needs are supported, but the need for autonomy is thwarted, people tend to develop a fairly strong controlled orientation, and when all of the needs are relatively thwarted, people tend to develop a high level of the impersonal orientation. In short, the satisfaction versus thwarting of the basic psychological needs for autonomy, competence, and relatedness explains the enhancement versus undermining of intrinsic motivation, the internalization of extrinsic motivation, and the development of general causality orientations.

Hedonic and Eudaimonic Wellness

Basic psychological needs theory was developed initially to address the issue of psychological wellness, although it has a central role in all of SDT, for satisfaction of the psychological needs has been shown to maintain and enhance intrinsic motivation, to promote internalization of extrinsic motivation, and to facilitate development of the causality orientations. As well, it has consistently been shown to promote well-being. Within that field, a distinction is frequently made between hedonic well-being and eudaimonic well-being (Ryan et al., 2008a). Hedonic well-being refers to being happy, to having high positive affect and low negative affect. Eudaimonic well-being, in contrast, refers to living one's life in a full and deeply satisfying way, actualizing one's human potentials. Eudaimonic living will often be accompanied by positive affect, but people will also experience negative affect when that is appropriate to the moment, as when something sad has just happened. Satisfaction of the basic psychological needs has been shown to be crucial for eudaimonic well-being.

Need Support, Motivation, and Outcomes

Much of the research examining the consequences of autonomous motivation, controlled motivation, and amotivation has been done in various applied settings, including homes, schools, workplaces, medical clinics, athletic settings, psychotherapy offices, and virtual worlds. In short, many hundreds of studies have shown that, across these domains in many countries, outcomes tend to be most positive when the social environment has been supportive of the basic needs and when the target individuals have been autonomously motivated.

For example, when elementary school teachers were more need supportive, their students were more intrinsically motivated and showed higher self-esteem; when instructors in a college course were more need supportive the students became more autonomously motivated for the course over the semester, understood the material better, and received higher grades in the course. And a study of medical students showed that when their instructors were more need supportive, the students internalized the course material more fully. In general, when the social context is more need supportive and people are more autonomously motivated for learning, they learn in a deeper more conceptual way, whereas, when the context is controlling or the learners are controlled in their motivation, they tend to do well at memorizing facts, but they show low levels of conceptual understanding (Ryan and Deci, 2009).

A study of employees in a work setting showed that when managers became more need supportive the employees were more satisfied with their jobs and more trusting of the company. A study of investment bankers showed that when

managers were more autonomy supportive, their employees performed better at their jobs and also showed better psychological well-being. These employees having higher autonomous causality orientations also contributed to their performing well and feeling good. In general, work settings that are more need supportive and workers who are more autonomously motivated have been found to yield more positive work outcomes (e.g., Baard et al., 2004).

In the domain of health care, for example, research has found that when physicians were need supportive, the patients were more autonomously motivated to take their medications and in turn showed greater medication adherence. Other research showed that when medical practitioners were more need supportive for diabetic patients, the patents' health improved (Ryan et al., 2008b).

In research on virtual worlds, results indicate that when players experience more satisfaction of the basic psychological needs while playing the games they are more intrinsically motivated for and more immersed in the games (Rigby and Ryan, 2011). They have also shown that much of the basis for players' aggression during or following game play is a function of their needs having been thwarted while playing.

Goals and Aspirations

The concept of goals has been perhaps the most common motivational concept in the psychological literature on motivation since the 1960s, when cognitive processes have been the central explanatory approach to psychology. Goals are outcomes that people value and hope to attain when engaging in particular behaviors. In SDT, although psychological needs is the most important explanatory concept, goals also have an important place. Specifically, SDT has focused on the degree to which people place value on what are called extrinsic life goals or aspirations, such as wealth, fame, and image, relative to intrinsic life goals such as personal growth, relationships, and community.

Research has shown that when people value the extrinsic aspirations more strongly than the intrinsic aspirations, they tend also to display poor psychological health, whereas when they value the intrinsic aspirations more strongly, they are psychologically healthier (Kasser and Ryan, 1996). These results have been consistently replicated in varied groups and cultures, and the explanation of the findings supported by SDT research is that the pursuit and attainments of the intrinsic goals of self-exploration, meaningful relationships, and community contributions tend to provide direct satisfaction of the basic psychological needs, whereas pursuit and attainment of the extrinsic goals of material possessions, social recognition, and attractive image are at best indirectly satisfying of the basic needs and may even be antagonistic to them.

Furthermore, research has shown that people tend to value the extrinsic aspirations when they have had a low level of basic psychological need satisfaction during their years as children, but they tend to value intrinsic aspirations when they have had a high level of basic satisfaction during those years (Kasser et al., 1995).

As well, studies found that if people's goals are manipulated experimentally, having intrinsic goals made salient led to better learning and performance than having extrinsic goals made salient. When, for example, business students who were learning about communications were told that it would help them learn about themselves, which is an intrinsic goal, their learning and performance was better than when the students were told that it would help them make more money, which is an extrinsic goal (Vansteenkiste et al., 2004).

Close Relationships

The concepts of SDT have also been used to examine close personal relationships such as best friends and romantic partners. Much of the SDT research has focused on the importance of autonomy in close relationships. Some theorists have argued that to have a satisfying close relationship people need to give up autonomy in service of the dyad, but SDT has argued that autonomy, as well as relatedness and competence, must be satisfied within a relationship in order for the relationship to be high quality and truly satisfying.

Studies have shown that indeed the degree to which people experience autonomy in a particular relationship predicts the degree of attachment security in that relationship. Across several relationships (e.g., mother, father, best friend, romantic partner) people experience considerable variability in the degree to which their need for autonomy is satisfied with different partners, and similarly they experience different degrees of attachment security. For each of the partners, autonomy need satisfaction directly predicted security of attachment in the relationship (La Guardia et al., 2000).

A different study examined best friend relationships and found that mutuality of providing autonomy support was important for satisfying friendships. In other words, when a person received autonomy support from a friend it contributed to the person's attachment security, emotional reliance, relationship satisfaction, and well-being, a set of findings that was true for each partner. Further, however, when a person gave autonomy support to the friend, not only did the friend benefit, but the person actually benefitted from the giving to the partner. So, both receiving autonomy support and giving autonomy support within a friendship benefits each partner in that relationship (Deci et al., 2006).

Summary

SDT is a motivational theory that differentiates between autonomous and controlled types of motivation and proposes that autonomous motivation leads to higher quality behavior and experience, especially for heuristic activities. It also differentiates between intrinsic and extrinsic aspirations and proposes that intrinsic aspirations are associated with greater well-being and better performance. Further, the theory specifies three basic psychological needs, those for autonomy, competence, and relatedness, which have been found to be universally essential for psychological health and well-being. Further, satisfaction of the basic needs not only promotes psychological health but also enhances intrinsic motivation, facilitates internalization of extrinsic motivation, supports the development of autonomous causality orientations, and strengthens intrinsic

relative to extrinsic aspirations. Across the domains of people's lives, the people are more optimally motivated, perform better, and feel healthier if their basic psychological needs are satisfied.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Control Behavior: Psychological Perspectives; Eudaemonism; Expectancy-Value-Cost Model of Motivation; Gender and Academic Motivation; Grit; Interest, Psychology of; Motivation and Actions, Psychology of; Motivation in Australian Aboriginal Populations; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivation, Learning, and Instruction; Personal Projects; Race and Academic Motivation; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self and Emotional Development in Adulthood and Later Life; Self-Regulation in Adulthood; Teacher Motivation.

Bibliography

Baard, P.P., Deci, E.L., Ryan, R.M., 2004. Intrinsic need satisfaction: a motivational basis of performance and well-being in two work settings. Journal of Applied Social Psychology 34, 2045–2068.

Bandura, A., 1996. Self-Efficacy: The Exercise of Control. Freeman, New York.

Chirkov, V., Ryan, R.M., Kim, Y., Kaplan, U., 2003. Differentiating autonomy from individualism and independence: a self-determination theory perspective on internalization of cultural orientations and well-being. Journal of Personality and Social Psychology. 84, 97–110.

Deci, E.L., 1971. Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology 18, 105–115.

Deci, E.L., Eghrari, H., Patrick, B.C., Leone, D.R., 1994. Facilitating internalization: the self-determination theory perspective. Journal of Personality 62, 119–142.

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668.

Deci, E.L., La Guardia, J.G., Moller, A.C., Scheiner, M.J., Ryan, R.M., 2006. On the benefits of giving as well as receiving autonomy support: mutuality in close friendships. Personality and Social Psychology Bulletin 32, 313–327.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-Determination in Human Behavior. Plenum, New York.

Gagné, M., Deci, E.L., 2005. Self-determination theory and work motivation. Journal of Organizational Behavior 26, 331–362.

Grolnick, W.S., Ryan, R.M., 1989. Parent styles associated with children's self-regulation and competence in school. Journal of Educational Psychology 81, 143–154.

Iyengar, S.S., DeVoe, S.E., 2003. Rethinking the value of choice: considering cultural mediators of intrinsic motivation. In: Murphy-Berman, V., Berman, J.J. (Eds.), Nebraska Symposium on Motivation: Cross-Cultural Differences in Perspectives on Self, vol. 49. University of Nebraska Press, Lincoln, NB, pp. 129–174.

Kasser, T., Ryan, R.M., 1996. Further examining the American dream: differential correlates of intrinsic and extrinsic goals. Personality and Social Psychology Bulletin 22, 280–287.

Kasser, T., Ryan, R.M., Zax, M., Sameroff, A.J., 1995. The relations of maternal and social environments to late adolescents' materialistic and prosocial values. Developmental Psychology 31, 907–914.

La Guardia, J.G., Ryan, R.M., Couchman, C.E., Deci, E.L., 2000. Within-person variation in security of attachment: a self-determination theory perspective on attachment, need fulfillment, and well-being. Journal of Personality and Social Psychology 79, 367–384.

Ng, J.Y.Y., Ntoumanis, N., Thøgersen-Ntoumani, C., Deci, E.L., Ryan, R.M., Duda, J., Williams, G.C., 2012. Self-determination theory applied to health contexts: a meta-analysis. Perspectives on Psychological Science 7, 325–340.

Piaget, J., 1971. Biology and Knowledge. University of Chicago Press, Chicago.

Rigby, C.S., Ryan, R.M., 2011. Glued to Games: The Attractions, Promise and Perils of Video Games and Virtual Worlds. Praeger, New York.

Ryan, R.M., Connell, J.P., Deci, E.L., 1985. A motivational analysis of self-determination and self-regulation in education. In: Ames, C., Ames, R.E. (Eds.), Research on Motivation in Education: The Classroom Milieu. Academic Press, New York, pp. 13–51.

Ryan, R.M., Deci, E.L., 2000a. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemporary Educational Psychology 25, 54–67.

Ryan, R.M., Deci, E.L., 2000b. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Ryan, R.M., Deci, E.L., 2009. Promoting self-determined school engagement: motivation, learning, and well-being. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook on Motivation at School. Routledge, New York, pp. 171–196.

Ryan, R.M., Huta, V., Deci, E.L., 2008a. Living well: a self-determination theory perspective on eudaimonia. Journal of Happiness Studies 9, 139–170.

Ryan, R.M., Patrick, H., Deci, E.L., Williams, G.C., 2008b. Facilitating health behavior change and its maintenance: interventions based on self-determination theory. The European Health Psychologist 10, 2–5.

Skinner, B.F., 1971. Beyond Freedom and Dignity. Knopf, New York.

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K.M., Deci, E.L., 2004. Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. Journal of Personality and Social Psychology 87, 246–260.

Relevant Website

www.selfdeterminationtheory.org.

Self-Development in Childhood and Adolescence

Susan Harter, University of Denver, Denver, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The nature of children's and adolescents' concepts of self change with age, in large part due to cognitive-developmental changes. Young children's self-concepts are limited to specific, concrete representations (e.g., I can run fast, I know my numbers). In middle and later childhood, children's self-descriptions represent trait labels (e.g., I am smart because I do well in Social Studies, Language Arts, and Math). At adolescence, self-descriptions become increasingly abstract (e.g., I am intelligent, athletic). In addition to domain-specific self-concepts, the ability to evaluate one's overall worth as a person emerges in middle childhood. The level of such global self-esteem varies tremendously across children and is determined by how adequate they feel in domains of importance as well as the extent to which significant others (e.g., parents and peers) approve of them as a person. Low self-esteem is associated with many psychological liabilities including depressed affect, lack of energy, and hopelessness about the future. However, efforts to promote positive self-esteem can be problematic if they are based on unrealistically positive judgments, namely, overestimations of one's competencies. I review the debate about Self-Esteem Movement, which has taken place largely in our schools, which has been criticized for fostering unrealistic perceptions of children's competencies and, in collusion with some parents, a sense of entitlement. Moreover, one must consider cross-cultural differences in the nature of self-esteem, which are legion. Certain cultures (e.g., China) do not even have a term for self-esteem. Plus, instruments designed for Americans are often inappropriate, in terms of content and psychometric properties, leading to faulty interpretations.

Introduction

Beginning in the second year of life, toddlers begin to talk about themselves. With development, they come to understand that they possess various characteristics, some of which may be positive ('I am smart') and some of which may be negative ('I am unpopular'). Of particular interest is how the very nature of such self-evaluations changes with development as well as among individual children and adolescents across two basic evaluative categories: (1) domain-specific self-concepts, i.e., how one judges one's attributes in particular arenas, e.g., scholastic competence, social and (2) global self-esteem (for a complete treatment of self-development in childhood and adolescence, see Harter, 1999, 2012).

Developmental shifts in the nature of self-evaluations are driven by changes in the child's cognitive capabilities. Cognitive-developmental theory and findings (see Piaget, 1960,1963; Fischer, 1980) alert us to the fact that the young child is limited to very specific, concrete representations of self and others, for example, 'I know my ABCs' (see Damon and Hart, 1988; Harter, 1999). In middle to later childhood, the ability to form higher order concepts about one's attributes and abilities (e.g., 'I am smart') emerges. There are further cognitive advances at adolescence, allowing the teenager to form abstract concepts about the self that transcend concrete behavioral manifestations and higher order generalizations (e.g., 'I am intelligent').

Developmental Differences in Domain-Specific Self-Concepts

Domain-specific evaluative judgments are observed at every developmental level. However, the precise nature of these judgments varies with age. Five common domains in which children and adolescents make evaluative judgments about the self have been identified: scholastic competence, physical competence, social competence, behavioral conduct, and physical appearance (Harter, 1999). The types of statements vary, however, across age periods, early childhood, middle childhood, later childhood, and early, mid-, and late adolescence, in keeping with the cognitive abilities of each age period.

Early Childhood

Young children provide very concrete accounts of their capabilities, evaluating specific behaviors. Thus, they communicate how they know their ABCs, how they can run very fast, how they are nice to a particular friend, how they do not hit their brother or sister, and how they possess a specific physical feature such as pretty blond hair. Of particular interest in such accounts is the fact that the young child typically provides a litany of virtues, touting his or her positive skills and attributes. One cognitive limitation of this age period is that the young child cannot distinguish the wish to be competent from reality. As a result, they typically overestimate their abilities because they do not yet have the skills to evaluate themselves realistically. Another cognitive characteristic that contributes to potential distortions is the pervasiveness of all-or-none thinking. That is, evaluations are either all positive or all negative, but primarily, all positive.

Middle Childhood

As the child grows older, the ability to make higher order generalizations in evaluating his or her abilities and attributes emerges. The child may observe that he or she is good at sports, in general. This inference can further be justified in that the

child can describe his or her talent at several sports (e.g., good at soccer, basketball, and baseball). Thus, the higher order generalization represents a cognitive construction in which an overarching judgment (good at sports) is defined in terms of specific examples that warrant this conclusion. Similar processes allow the older child to conclude that he or she is smart (e.g., does well in math, science, and history). The structure of a higher order generalization about being well behaved could include such components as obeying parents, not getting into trouble, and trying to do what is right. A generalization concerning the ability to make friends may subsume accounts of having friends at school, making friends easily at camp, and developing friendships readily upon moving to a new neighborhood. The perception that one is good looking may be based on one's positive evaluation of one's face, hair, and body.

Later Childhood

During later childhood, all-or-none thinking diminishes and the aura of positivity fades. Thus, children do not typically think that they are all virtuous in every domain. The more common pattern is for them to feel more adequate in some domains than others. For example, one child may feel that he or she is good at schoolwork and is well behaved, whereas he or she is not that good at sports, does not think that he or she is good looking, and reports that it is hard to make friends. Another child may report the opposite pattern.

Contributing to this advance is the ability to engage in social comparison. Children can use comparisons with others as a barometer of the skills and attributes of the self. In contrast, the young child cannot simultaneously compare his or her attributes to the characteristics of another in order to detect similarities or differences that have implications for the self. Although the ability to utilize social comparison information for the purpose of self-evaluation represents a cognitive-developmental advance, it also ushers in new, potential liabilities. With the emergence of the ability to rank-order the performance of other children, all but the most capable children will necessarily fall short of excellence. Thus, the very ability and penchant to compare the self with others makes one's self-concept vulnerable, particularly if one does not measure up in domains that are highly valued. The more general effects of social comparison can be observed in findings revealing that domain-specific self-concepts become more negative during later childhood, compared to early childhood.

Early Adolescence

For the adolescent, there are further cognitive-developmental advances that alter the nature of domain-specific self-evaluations. Early adolescence brings with it the ability to create more abstract judgments about one's attributes and abilities. Thus, one no longer merely considers oneself to be good at sports but to be athletically talented. One is no longer merely smart but views the self more generally as intelligent, where successful academic performance, general problem-solving ability, and creativity might all be subsumed under the abstraction of intelligence. Abstractions may be constructed in the other domains. In the domain of behavioral conduct, there will be a shift from the perception that one is well behaved to a sense that one is a moral or principled person. In the domains of social competence and appearance, abstractions may take the form of perceptions that one is popular and physically attractive.

These illustrative examples all represent positive self-evaluations. However, during adolescence (as well as in later childhood), judgments about one's attributes will also involve negative self-evaluations. Thus, certain individuals may judge the self to be unattractive, unpopular, unprincipled, etc.

Mid-Adolescence

Of particular interest is the fact that when abstractions emerge, the adolescent typically does not have total control over these new acquisitions, just as when one is acquiring a new athletic skill (e.g., swinging a bat and maneuvering skis) one lacks a certain level of control. In the cognitive realm, such lack of control often leads to dramatic shifts in self-perceptions across situations or time. For example, the adolescent may conclude at one point in time that he or she is exceedingly popular but then, in the face of a minor social rebuff, may conclude that he or she is extremely unpopular. Moreover, adolescents develop multiple selves, how they are different in different role relationships, for example, cheerful with peers, depressed with parents, self-conscious on a date, etc. However, these vacillations and different selves cause distress, perceived psychological conflict, and confusion about just which is their 'true self' (see Harter, 2012).

Late Adolescence

Gradually, adolescents gain control over these self-relevant abstractions such that they become capable of more balanced and accurate self-representations (see Harter, 1999, 2012). They develop more advanced cognitive skills to deal with the confusion and the conflict between their multiple selves. Two strategies have been documented in the author's own work. As adolescents advance to more mature levels of cognitive development, they develop higher levels of abstractions that can integrate the seeming contradictions of the previous period. For example, the realizations that they are depressed with parents but cheerful with friends can be subsumed under the more advanced abstraction that they are 'moody,' thus integrating and justifying what were previously inconsistent self-perceptions. Second, older adolescents normalize what were former contradictions, describing how it would be 'weird' to be the same with your mother as you are on a date, or with friends at a ball game (see Harter, 2012).

Global Self-Esteem

The ability to evaluate one's worth as a person also undergoes developmental change. The young child simply is incapable, cognitively, of developing the verbal concept of his or her value as a person. This ability emerges at the approximate age of

8 years. However, young children exude a sense of value or worth in their behavior. The primary behavioral manifestations involve displays of confidence, independence, mastery attempts, and exploration (see Harter, 1999, 2012). Thus, behaviors that communicate to others that children are sure of themselves are manifestations of high self-esteem in early childhood.

At about the third grade, children begin to develop the concept that they like, or do not like, the kind of person they are (Harter, 1999; Rosenberg, 1979). Thus, they can respond to general items asking them to rate the extent to which they are pleased with themselves, like who they are, and think they are fine, as a person. Here, the shift reflects the emergence of an ability to construct a higher order generalization about the self. This type of concept can be built upon perceptions that one has a number of specific qualities, for example, that one is competent, well behaved, attractive, etc. It can also be built upon the observation that significant others are approving and supportive, for example, parents, peers, and teachers, think highly of the self. This process is greatly influenced by advances in the child's ability to take the perspective of significant others (Selman, 1980). During adolescence, one's evaluation of one's global worth as a person may be further elaborated, drawing upon more domains and sources of approval, and will also become more abstract. Thus, adolescents can directly acknowledge that they have high or low self-esteem, as a general abstraction about the self.

Individual Differences in Domain-Specific Self-Concepts as Well as Global Self-Esteem

Although there are predictable cognitively based developmental changes in the nature of how most children and adolescents describe and evaluate themselves, there are striking individual differences in how positively or negatively the self is evaluated. Moreover, one observes different profiles of children's perceptions of their competence or adequacy across the various self-concept domains, in that children evaluate themselves differently across domains. Consider the profiles of four different children. One child, Child A, may feel very good about her scholastic performance, although this is in sharp contrast to her opinion of her athletic ability, where she evaluates herself quite poorly. Socially she feels reasonably well accepted by her peers. In addition, she considers herself to be well behaved. Her feelings about her appearance, however, are relatively negative. Child A also reports very high self-esteem. Another child, Child B, has a very different configuration of scores. This is a boy who feels very incompetent when it comes to schoolwork. However, he considers himself to be very competent, athletically, and feels well received by peers. He judges his behavioral conduct to be less commendable. In contrast, he thinks he is relatively good-looking. Like Child A, he also reports high self-esteem.

Other profiles are exemplified by Child C and Child D, neither of whom feel good about themselves scholastically or athletically. They evaluate themselves much more positively in the domains of social acceptance, conduct, and physical appearance. In fact, their profiles are quite similar to each other across the five specific domains. However, judgments of their self-esteem are extremely different. Child C has very high self-esteem, whereas Child D has very low self-esteem. This raises a puzzling question: how can two children look so similar with regard to their domain-specific self-concepts but evaluate their global self-esteem so differently? The author turns to this issue next, in examining the causes of global self-esteem.

The Causes of Children's Level of Self-Esteem

The understanding of the antecedents of global self-esteem has been greatly aided by the formulations of two historical scholars of the self, William James (1892) and Charles Horton Cooley (1902). Each suggested rather different pathways to self-esteem, defined as an overall evaluation of one's worth as a person (see reviews by Harter, 1999, 2012; Rosenberg, 1979). James focused on how the individual assessed his or her competence in domains where one had aspirations to succeed. Cooley focused on the salience of the opinions that others held about the self, opinions that one incorporated into one's global sense of self (*see* Self: History of the Concept).

Competence–Adequacy in Domains of Importance

For James, global self-esteem derived from the evaluations of one's sense of competence or adequacy in the various domains of one's life relative to how important it was to be successful in these domains. Thus, if one feels one is successful in domains deemed important, high self-esteem will result. Conversely, if one falls short of one's goal in domains where one has aspirations to be successful, one will experience low self-esteem. One does not, therefore, have to be a superstar in every domain to have high self-esteem. Rather, one only needs to feel adequate or competent in those areas judged to be important. Thus, a child may evaluate himself or herself as unathletic; however, if athletic prowess is not an aspiration, then self-esteem will not be negatively affected. That is, the high self-esteem individual can discount the importance of areas in which one does not feel successful.

The author has directly examined this explanation in research studies by asking children to rate how important it is for them to be successful (Harter, 1999). The findings reveal that high self-esteem individuals feel competent in domains they rate as important. Low self-esteem individuals report that areas in which they are unsuccessful are still very important to them. Thus, the discrepancy between high importance and perceptions of inadequacy contribute to low self-esteem. In contrast, the congruence between importance and feelings of success leads to higher self esteem.

Incorporation of the Opinions of Significant Others

Another important factor influencing self-esteem can be derived from the writings of Cooley (1902) who metaphorically made reference to the 'looking-glass self.'

According to this formulation, significant others (e.g., parents and peers) were social mirrors into which one gazed in order to determine what they thought of the self. Thus, in evaluating the self, one would adopt what one felt were the judgments of these others whose opinions were considered important. Thus, the approval, support, or positive regard from significant others became a critical source of one's own sense of worth as a person. For example, children who receive approval from parents and peers report much higher self-esteem than children who experience disapproval from parents and peers.

The author's findings (Harter, 1999, 2012) reveal that both these factors, competence in domains of importance and the perceived support of significant others, combine to influence a child's or adolescent's self-esteem. Thus, those who feel competent in domains of importance and who also report high support, rate themselves as having the highest self-esteem. Those who feel inadequate in domains deemed important and who also report low levels of support rate themselves as having the lowest self-esteem.

Cross-Cultural Considerations

Globalization and the proliferation of psychological research on the international scene is a welcome new trend. However, one must appreciate that there are clear differences between Western and Eastern conceptions that much be recognized, before Westerners encourage Eastern cultures to embrace Western frameworks, paradigms, and instruments (Harter, 2012). Clearly, self-esteem is a value touted by Americans, as witnessed by the many attempts to protect and enhance it (Harter, 2012; Leary, 2004). However, high self-esteem, based on individual accomplishments, is not revered as a psychological commodity in most Eastern cultures (see Harter, 2012). In fact, in some languages, namely, Chinese and Japanese, there is no equivalent term for 'self-esteem.' Self-esteem is simply not a core value and therefore is not represented or codified in their respective languages.

This leads to numerous problems when instruments designed for Americans are administered in many other countries (see Harter, 2012). First, the content may not be appropriate, meaning at the domain or item level. Most noticeable is that self-esteem items do not resonate with such populations. As a result, this interferes with the psychometric properties in that American instruments administered in such countries lack reliability and validity, and the factor structures do not hold up. Question formats such as the one employed in our own instruments pull for social comparison, which is frowned upon in many countries, leading to lower scores than would be anticipated given objective data about levels of achievement and competencies in certain cultures. See Harter (2012) where these different issues are addressed in more detail. There, the author urges that those in different cultures take stock of these problems and perhaps initially designed interview questions to address their specific research questions and hypotheses that are more culturally sensitive and will lead to more meaningful findings.

To Raise or Not to Raise Self-Esteem within both Society and Our Schools

For two decades, considerable debate and controversy has surrounded this question. Initiatives have been proposed to raise self-esteem at the national and state levels, as a cure-all for many societal ills. Parallel interventions at the classroom level have been urged by some educators who presume that self-esteem enhancement should motivate students to attain higher levels of academic achievement. The initiative adopted by the State of California, in 1990, reflects one of the most ambitious efforts to tackle this issue at a legislative level (California Task Force to Promote Self-Esteem and Personal Responsibility, 1990). This particular initiative became a public marker of the 'Self-Esteem Movement.' Its message was also a subtext in the playbooks of many educators.

California state legislators expressed their conviction that by raising the self-esteem of its citizenry, many social ills would be significantly reduced, for example, crime, teen pregnancy, delinquency, drug abuse, and school underachievement. Such a widespread intervention would lead to better societal, as well as personal, outcomes. The California initiative vaulted the topic of self-esteem into the national limelight, as the media scrambled to cover the frenzied efforts to alter self-perceptions in the service of the greater cultural good. Swann (1996) likens the quest for high self-esteem to the search for the Holy Grail, pointing to the dizzying array of programs that proliferated. Unfortunately, these efforts did not have their intended effects, based upon very disappointing findings that did not produce the desired outcomes. Self-esteem, particularly if it was unrealistically enhanced, did not represent the anticipated panacea that would cure our social ills (see Eccles and Roeser, 2009).

How the self-esteem movement played out on the stage of America's classrooms, as an educational microcosm of society, is of particular relevance in terms of its implications for the youth. A major focus of many educational interventions has been the enhancement of self-esteem, given the argument that student achievement would improve, as a result. These contentions have fueled an intense debate that is reflected within the educational and psychological literature (see review in Harter, 2012). Proponents have touted the advantages of enhanced self-esteem, whereas opponents have vociferously argued that such efforts are not only misguided and useless, at best, but destructive, at worst. What are the primary arguments, pro and con?

The claims of the proponents are perhaps the most straightforward, given the belief and insistence that raising the self-esteem of students will produce higher levels of achievement and foster academic success, primarily measured by grades and test scores (see Tice and Gailliot, 2006; for thoughtful reviews of the dispute).

Those who believe in the value of enhancing self-esteem among students first argue that while admittedly modest, the correlation between these two constructs is sufficient to warrant educational intervention. Second, these proponents also contend that the enhancement of self-esteem will

produce educational benefits, that is, global perceptions of one's worth as a person act causally to promote gains in academic achievement.

Opponents of educational interventions first point out that the magnitude of the correlation between self-esteem and academic achievements is relatively trivial and thus without serious implications. Second, they challenge the presumed directionality of effects. That is, although the two constructs may bear a modest correlation to one another, it is just as likely that successful achievement in the classroom (as indexed by good grades and high standardized test scores) represents the cause of higher self-esteem, not the effect. These are the general arguments that define each position. Next, the author turns to the more nuanced considerations.

A major concern addressed by many opponents is that seeking to boost self-esteem can artificially destroy any incentives for students to devote effort toward scholastic goals (Tice and Gailliot, 2006), particularly if self-evaluations are unrealistic. Inflated self-esteem provides no incentive for self-improvement. Moreover, the practice of inviting students to make unconfirmed self-affirmations as to their worth ('I am smart,' 'I am awesome,' 'I am special') do not appear to be effective.

Damon (1995) was also among the first to be critical of what he considered an overemphasis by educators on promoting high self-esteem, particularly when these efforts lead to an inflated sense of one's self-worth. Damon has viewed such efforts as not only misguided but potentially detrimental, arguing that they divert educators from the active teaching of skills and deprive students of the thrill that results from personal accomplishment. It has been Damon's contention that self-esteem has become an overrated commodity and that the effusive praise that parents and teachers heap on children to make them feel good about themselves is often met with suspicion by the children themselves. Moreover, it interferes with their attention to the goal of building specific skills in the service of genuine achievement. It does a grave educational injustice to encourage children to think well of themselves without having earned it. Cote (2009) has joined the ranks of opponents, decrying the 'cult of self-esteem' that has gripped schools in this country.

'Feel-good activities,' a mainstay of many of the self-esteem interventions to emerge in the 1980s, have produced few gains, despite the widespread commercial packaging of programs designed to enhance self-esteem. They demand that teachers suspend corrective feedback and remove challenging material, 'dumbing down' the curriculum (Damon, 1995). Such activities divorce self-esteem from palpable achievements that result from realistic attributions to the student's personal effort. As a result, they confound self-esteem with egocentrism, narcissism, arrogance, conceit, and self-delusion (see Harter, 2012). That said, proponents of these packaged programs claim that the self-affirming affective exercises and techniques will extend, not only to heightened school performance, but to broader societal goals to help curb such teenage social ills as pregnancy, drug use, and delinquency. However, there is little evidence to warrant these claims (see Covington, 2006). Moreover, given my own experience in the schools over the years, many teachers who were forced to administer what they called touchy-feely exercises complained that this was not their job definition, rendering them less than effective if they attempted to mouth the self-esteem mantra.

Furthermore, the goal of self-enhancement strategies, for its own sake, interferes with the educational goals of learning and mastery. As Covington (2006) observes, enhanced achievement is more likely if children are motivated to learn for the sake of learning, to stimulate their own curiosity, to challenge themselves, and to set personal goals that involve skill improvement. If the reasons behind school performance are more self-promotional and defensive in nature, for example, striving to succeed in order to avoid failure with its implication that one is incompetent, academic motivation will wane because of the fear that one cannot succeed.

Perhaps one of the strongest and most vocal opponents of self-esteem enhancement programs within the school system is Baumeister and colleagues (Baumeister, 1996; Baumeister et al., 2003). Baumeister's devastating critique is captured by this assertion: "At worst, the pursuit of high self-esteem is a foolish, wasteful, and self-destructive enterprise that may end up doing more harm than good" (Baumeister, 1996: p. 120). What is the basis for his contentions?

To begin with, Baumeister is critical of the California initiative to promote self-esteem, which professed to represent a social vaccine to cure or hopefully inoculate citizens against the evils of crime, violence, substance abuse, teen pregnancy, child abuse, chronic welfare dependency, and educational failure. He rightfully asserts, consistent with the considerable evidence on narcissism, that for many in this country, self-esteem is already inflated.

We seem to be revisiting an earlier movement, from the 1960s and 1970s, what the author has called the Camelot era of self-esteem interventions, which invoked simplistic attempts to enhance self-esteem that were unsuccessful. These urged general self-affirmations (e.g., "I'm OK, you're OK" messages) and affective feel-good exercises that were designed to impact self-esteem directly. The contemporary analog can be observed among educators, coaches, and parents who have fallen prey to the 'good job!' mantra. Children and adolescents are lavishly and indiscriminately praised for their efforts, independent of whether they warrant such positive feedback.

Recent findings (reviewed in Harter, 2012) reveal that narcissism among the youth has systematically increased over the past three decades, given that society, in general, and schools, in particular, have both perpetuated messages and practices that promote self-enhancement, self-serving biases, and a sense of entitlement. Affirmations of 'specialness' and educational practices such as grade inflation and 'dumbing down' the curriculum prompt unrealistic views of competence and contribute to narcissistic self-promotion. In May of 2012, in a risky and courageous high school graduation speech, the speaker shocked students and outraged parents by his blatant contention that "you students are not special!" It was a wake-up call that caught the attention of the national media and reactions went viral. Thus, the bottom line is

that our educational system and child-rearing practices should cultivate behaviors that produce self-views that are both realistic (that is, based on objective evidence) and adaptive (that is, promote activities that predict long-term adjustment to the demands of society).

Conclusions

Two types of self-representations that can be observed in children and adolescents were distinguished, evaluative judgments of competence or adequacy in specific domains and the global evaluation of one's worth as a person, namely, overall self-esteem. Each of these undergoes developmental change based on age-related cognitive advances. In addition, older children and adolescents vary tremendously with regard to whether self-evaluations are positive or negative. Within a given individual, there will be a profile of self-evaluations, some of which are more positive and some more negative. More positive self-concepts in domains considered important, as well as approval from significant others, will lead to high self-esteem. Conversely, negative self-concepts in domains considered important, coupled with lack of approval from significant others, will result in low self-esteem. Self-esteem is particularly important since it is associated with very important outcomes or consequences. Perhaps the most well-documented consequence of low self-esteem is depression. Children and adolescents (as well as adults) with the constellation of causes leading to low self-esteem will invariably report that they feel depressed, emotionally, and are hopeless about their futures; the most seriously depressed consider suicide. Thus, it is critical that one intervenes for those experiencing low self-esteem (see suggestions in Harter, 1999).

Positive self-esteem can be a psychological commodity, a resource that is important for us to foster in our children and adolescents if we want them to lead productive and happy lives. However, unrealistically inflated positive views of one's competencies and worth as a person can also compromise one's personal growth. It can lead to self-aggrandizing tendencies that can merge into narcissism, which can eventually alienate others and lead to blame and hostile behaviors (see Harter, 2012, 2012). Thus, our responsibility should be to foster realistic self-perceptions and provide support to encourage children and adolescents to develop the necessary skills toward that goal.

See also: Identity in Childhood and Adolescence; Self-Concepts: Educational Aspects; Self: History of the Concept.

Bibliography

Baumeister, R.F., 1996. Should schools try to boost self-esteem? Beware of the dark side. American Educator 43, 14–19.

Baumeister, R.F., Campbell, J.D., Krueger, J.L., Vohs, K.D., 2003. Does high self-esteem cause better performance, interpersonal success happiness, or healthier life styles? Psychological Sciences in the Public Interest 4, 1–44.

California Task Force to Promote Self-Esteem and Personal and Social Responsibility, 1990. Toward a State of Self-esteem. California Statement of Education, Sacramento, CA.

Cooley, C.H., 1902. Human Nature and the Social Order. Charles Scribner's Sons, New York.

Cote, J.E., 2009. Identity formation and self-development in adolescence. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, third ed., vol. 1. Wiley, New York, pp. 266–304.

Covington, M., 2006. The self-worth theory of achievement motivation: findings and implications. Elementary School Journal 85, 5–20.

Damon, W., 1995. Greater Expectations: Overcoming the Culture of Indulgence in America's Homes and Schools. Free Press, New York.

Damon, W., Hart, D., 1988. Self-understanding in Childhood and Adolescence. Cambridge University Press, New York.

Eccles, J.S., Roeser, R.W., 2009. Schools, academic motivation, and stage-environment; fit. In: Lerner, R.M., Steinberg, L. (Eds.), The Handbook of Adolescent Psychology, Individual Bases of Adolescent Development, third ed., vol. I. Wiley, New York, pp. 404–434.

Fischer, K.W., 1980. A theory of cognitive development: the control and construction of hierarchies of skills. Psychological Review 87, 477–531.

Harter, S., 1999. The Construction of the Self: A Developmental Perspective. Guilford Press, New York.

Harter, S., 2012. The Construction of the Self: Developmental and Sociocultural Foundations. Guilford Press, New York.

Harter, S., 2012. I-self and Me-self processes affecting developmental psychopathology and mental health. In: Cicchetti, D. (Ed.), Handbook of Developmental Psychopathology. third ed. Wiley, New York.

James, W., 1892. Psychology: The Briefer Course. Henry Holt, New York.

Leary, M., 2004. The Curse of the Self: Self-awareness, Egotism, and the Quality of Human Life. Oxford University Press, Oxford, UK.

Piaget, J., 1960. The Psychology of Intelligence. Littlefield, Adams, Patterson, NJ.

Piaget, J., 1963. The Origins of Intelligence in Children. Norton, New York.

Rosenberg, M., 1979. Conceiving the Self. Basic Books, New York.

Selman, R.L., 1980. The Growth of Interpersonal Understanding. Academic Press, New York.

Swann Jr., W.B., 1996. Self-traps: The Elusive Quest for High Self-esteem. Freeman, New York.

Tice, D.M., Gailliot, M., 2006. How self-esteem relates to the ills and triumphs of society. In: Kernis, M.H. (Ed.), Self-esteem: Issues and Answers. Psychology Press, New York.

Self-Directed Care and Individual Budgets for Older People: Lessons from the UK

Jill Manthorpe, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Policy ambitions to put older people at the center of their own care are being developed internationally. These enable people to express their choices in, and control over, the support they require to make life good. This article outlines the common approaches to this trend and summarizes evidence about the implications of personal budgets, cash for counseling, consumer-directed care, and self-directed support. It discusses the evidence base related to older people as consumers and the implementation of schemes that have been evaluated. The article debates these developments in the context of social science theories.

Introduction and Background

One of the major changes to the care and support of older people in the developed world is that of 'cash for care' schemes or personal allocations of funding to meet care needs (Ungerson and Yeandle, 2007). These are affecting the lives of many disabled older people and their family carers. The changes introduced have been greatly influenced by disability movements and disability politics. They are reforming the social systems of 'care for the aged' and have influenced the social science discipline of gerontology by reminding it of the linkages to disability theory, practice, and politics.

This article draws on the international literature (see Ottmann et al., 2013; Rummery, 2011; Alakeson, 2010) but particularly the current story of the transformation of adult social services in England and Scotland, to illustrate the potential for the social and behavioral sciences to make sense of this change. The aims of the transformation to personal budgets or self-directed support (SDS) (the terms used in the UK to cover 'cash for care' schemes) are to broaden choice and control for people using social care services – generally assistance with activities of daily living (DH, 2008). Overall, it is referred to in England as the policy of personalization. The objective is to improve outcomes for older people and their carers (family caregivers). Central government in England has used the overarching term personalization to direct local government to transform its assessment, allocation systems, service provision, monitoring, and reviews of people needing care and support (HM Government, 2007). In Scotland, the term self-directed support (SDS) is more commonly employed (Scottish Government, 2013). Terms such as consumer-directed care are used to cover similar changes in other countries such as Australia and the United States (US) (Ottmann et al., 2013).

Under the United Kingdom (UK) welfare state system, disabled adults or others in need of care and support are entitled to assistance from local authorities if they meet eligibility criteria around need and risk although they must pay a means-tested contribution if their income or resources are above a minimum level. The essence of the change being made by the policy of personalization is that individuals are encouraged to be more involved in assessing their needs for support; that they are told how much money they are likely to be allocated to meet these needs, and that they can choose how to meet their needs. In early pilots of this approach this money was referred to as an individual budget. At the time of writing in many areas it is now being termed a personal budget or Direct Payment. In light of the sums available to them, older people and other eligible adults devise a support plan to meet their needs, deploying this in a variety of ways, including taking this money in the form of a Direct Payment (cash paid to the person or their proxy), or as a combination of this with local authority commissioned services, or using it to draw up a contract themselves with an organization or individual that provides care as part of its business. The person's support plan has to be agreed with the local authority (often by social work professionals) and is reviewed. As this article reports, there is concern that positive outcomes for older people may be small in scale or not widespread, and some of these issues are explored in this article.

The Evidence Base

These developments rest on evidence that existing services and systems do not meet people's needs effectively, particularly older people who may have combinations of need and frailty that change frequently and unpredictably (Feltoe and Orellana, 2013). Older people report that some social care services funded or provided by local authorities (such as care at home or day-care center) do too little to meet their needs and often arrive too late, offer limited choice, seem paternalistic in nature, and reduce rather than promote independence. Older people are less likely to actively participate in assessment and more likely to underestimate their needs than other disabled adults. The 'lottery' of provision and support reported (Manthorpe et al., 2008) suggests current systems do not meet individuals' needs equitably and agism is often seen to underpin such inequity.

Not surprisingly, given the changes required to set in place the new system of personal budgets, there have been widespread calls for training of frontline staff. Moreover, given the extent of change a number of training initiatives, information events and written and electronic materials are being developed to inform older people of their options, to tell them about the new system and to describe the responsibilities accompanying the exercise of choice and control. This may be seen as part of the 'responsibilization' of citizens under which people have

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28044-6

both rights and responsibilities, which include the taking on of responsibilities previously assumed by the state or professionals. Coproduction is another term used to describe this aspiration for a change in professional and client (user) dynamics (Needham and Carr, 2009).

In parallel, the risks of personal budgets have been highlighted by those with an interest in avoiding the abuse and neglect of older people. The discovery or social construction of the social problem of 'elder abuse' has been another key development in social gerontology in the past few decades (Dixon et al., 2010). Social science theories have also been used to describe the ways in which risk is constructed and can be managed proportionally (Manthorpe and Moriarty, 2010). Personal budgets have been described as 'inherently risky' by some professionals who believe that some older people, for example, with dementia, are open to exploitation in new systems where monitoring and review are not sufficiently sophisticated to be able to identify and respond to possible risks of harm or abuse (Manthorpe and Samsi, 2012).

Evidence and Evaluations

These developments and these theoretical perspectives have been informed by social science evaluations. A major UK study used a randomized control trial (RCT) methodology, meaning that the quality of the evidence is good. In 2005 the English Department of Health (DH) commissioned a wide-ranging evaluation of the 13 individual budget (IB) pilot projects (the Individual Budgets Evaluation Network – IBSEN) (see Glendinning et al., 2008). This evaluation investigated the outcomes, costs and cost-effectiveness of IBs compared with conventional methods of funding, commissioning and service delivery. Comparisons were made for all user groups pooled together and for individual user groups. They were based on data collected 6 months after individuals started receiving their IBs and on in-depth interviews with a subsample roughly 2 months after they had decided to accept an IB.

IBSEN's findings about IBs and *older people* were different from those for other user groups (people under retirement age with learning disabilities; physical disabilities or mental health problems) (see Moran et al., 2013). Older people:

- Were significantly less likely than people in other user groups to report higher aspirations as a result of their experience of the IB process.
- Reported lower psychological well-being than those in the comparison group. (There was some complication in the interpretation of this finding, because it was influenced by whether the interviews were completed by proxy respondents rather than the older people themselves.)
- Provided evidence that there were no significant cost differences between the IB and comparison groups, both for the pooled sample and for older people separately. Across all user groups combined there was some indication of higher use and higher costs of health care services by people in the IB group relative to the comparison group, and also higher care coordinator (care manager/social worker) costs.
- Provided little evidence of cost-effectiveness differences between IBs and standard or traditional support arrangements for older people when concentrating on social care outcomes as the measure of effectiveness. When looking at psychological well-being, standard support arrangements were marginally more cost-effective than IBs.
- Reported themselves in interview they generally did not want the 'additional burden' of planning and managing their own support, with responsibilities for managing a budget, making new arrangements, and managing the risks associated with directly employing care workers rather than having them provided by an agency or local authority.

One hypothesis developed to explain these differences was that it might be too early for older people to have developed the confidence to assume the greater control that comes with holding an IB. Practitioners such as social workers reported that older people had reservations and concerns about IB payments, namely lack of confidence; and fluctuating, poor health and progressive disabilities that would perhaps require multiple adjustments. Another hypothesis was that local authorities might not have had the opportunity within the short timeframe of the IB pilots to implement arrangements suiting older people. For example, many older people receiving IBs opted for care-managed 'virtual budgets' (whereby older people did not manage any money themselves and the arrangements of their support plan were in the hands of a social worker or care manager) and thereby the scope for them to feel in control might have been limited. Practitioners noted that older people often called for help in crisis; and support had to be put in place immediately; not being developed in conjunction with the older person as much as it ideally could have been. A third hypothesis was that the lower average levels of money received as 'virtual' budget or as a Direct Payment by older people, and the predominant use of IBs to fund personal care, offered fewer opportunities to develop tailored support that could impact positively on outcomes. In other words, the inherent agism of support for older disabled people was thrown into sharp relief when it was possible to compare allocations for younger and older disabled people.

However, other evidence has given cause for concern about the experiences of older people while personal budgets are developing nationwide. These highlighted the need to ensure that older people are not disadvantaged by the very policies intended to foster better care and lead to better outcomes. Woolham and Benton (2012), for example, concluded from their detailed study of the impact and outcomes of personal budgets (termed SDS) in the county of Northamptonshire that there were limited benefits and great costs for older people of these new arrangements. Older people receiving SDS in their study were less likely to feel in control; to feel empowered; to feel they were getting the right type of support than other age groups; to feel they had the final say about how the money should be spent and displayed less evidence of positive impact. Overall costs of SDS were higher for all, especially for older people. In the context of evidence-based practice, these studies need to be set alongside powerful personal accounts of the benefits of personal budgets for some older people (see for example Newbronner et al., 2011; Feltoe and Orellana, 2013). A recent systematic review found low-quality evidence for consumer-directed care and notes that while it appears to increase satisfaction with care and community service use it has little effect on clinical outcomes for older people (Low et al., 2011).

Implementation

Personalization has provided great scope for social scientists to research services for older people and has increased understanding of lived experience, care dynamics, and service outcomes. Studies have included major national trials (see Glendinning et al., 2008), local evaluations (Woolham and Benton, 2012), and qualitative explorations (Newbronner et al., 2011). A range of studies has also focused on implementation of government policy and used personalization of an example of complexity and nonlinear development. In particular, implementation research has pointed to the blurring of exhortation and guidance:

> Implementation of personalisation in its broadest sense has proven difficult to measure and record in practice. There are some activities and targets which have been used to indicate that a local authority or provider is taking a personalised approach to provision, but we know that it is as much about culture change and attitude of staff as about process and procedure. Policy documents related to personalisation in social care are, therefore, a mixture of description of the vision of personalisation combined with gentle persuasion aimed at local authorities and providers to implement change.
>
> Feltoe and Orellana, 2013: p. 12

Microlevel Understandings

It is not surprising that mixed views are emerging. Social care support for older people reflects the circumstances in which they first encounter publicly funded social care, often at times of crisis or change. In this section microperspectives are used to explore how older people with experiences of IBs (now personal budgets in the main in England and SDS in Scotland) are able to exercise choice by 'managing the money' themselves, rather than having a budget managed by the local authority (for details of this see Feltoe and Orellana, 2013). **Box 1**, for example, sets out what an older person is able to spend their personal budget on in one local authority area and displays not only the extent but also the limits of choice. It is important to understand the meanings behind such conditionality; in Australia, for example, the prohibition on employing family or friends may have its advantages (minimizing role blurring) but also disadvantages (reducing choice) (Low et al., 2012). In the US evidence from qualitative interviews suggests that poor older people may see this economic transaction – employing family members – as a reciprocal activity and that they value the enhanced relationship (San Antonio et al., 2006).

Box 1 Local stipulations for a personal budget (anonymized local authority)

A person managing a budget is required to:

- set up a separate bank account so that the personal budget can be paid to them by the local authority every 4 weeks;
- be responsible for paying for their support, depending on what support or services are arranged;
- make all payments using check or direct debit. No cash transactions are to be made with the personal budget money;
- spend their money on a flexible basis and therefore need to keep track of how their money is spent; and
- keep a record of how their money is spent, and are expected to keep receipts and bank statements.

The social science literature on choice as applied to the subject of personal budgets was summarized by Stevens et al. (2011). As they noted, if people with social care needs have sufficient cash to enable them to exercise choices over their care and support, this theoretically alters the power dynamics of relationships between older people, professionals, and the organizations in which they work, in relation to needs and how to meet them. It is further argued that increasing individuals' choice over the services they use or the people they employ at microlevel may affect public decisions about priorities and resource allocation (macrolevel). However, they cautioned that a focus on increasing choice at the expense of other concerns, such as variable levels of support for different groups and decreasing the range of affordable options to meet needs of different kinds ran the risk of increasing inequity.

The Benefits of Personal Budgets

Nonetheless older people generally see personal budgets as giving them more freedom of choice and control, enabling them to get support when they want it; and sometimes enabling this support to be more flexible and creative (Moran et al., 2013). They also appreciate knowing what money is available. In contrast, practitioners have observed that while these might be the aspirations of personal budgets, they are also a way to control costs for the local authority. One key change is the move from traditional day center or day services care to 'day opportunities' which are hard to define and assess, and may be laden with value judgments (Manthorpe and Moriarty, 2013).

Sadly, the social sciences have not generally been greatly interested in researching 'low-level' support for older people so one difficulty in relation to social care services of all types is that reports covering usage and user experiences are often small scale and the context of the service is not always clearly described, making it difficult to know what is offered at the service, what is used and by whom, and the outcomes intended, achieved or unachieved. Day services exemplify this. Some day care is described as 'low level' and therefore comparisons with other day-care centers would need to ensure this was taken into account and whether 'low level' describes the needs of older people using the service or the support provided. Patterns of activities or therapy and take-up may vary. Few studies enable us to see the older people using day-care services in the context of other parts of their lives; so we do not always know whether the older person receives support from other parts of the social care system or from family networks, for example. Small ethnographic studies are valuable but they cannot make up for this wider lack of knowledge about service aims, costs, and systems. The impact of this lack of evidence is that the outcomes of personal budgets are difficult to compare to 'treatment as usual.'

Maximizing the Skills of Older People to Manage Their Personal Budget

At microlevel, much depends on older people's familiarity with financial matters generally and if they are accustomed to managing their incomes and savings, if any, should they take on financial management. If people already hold bank accounts then dealing with a new form of credit and debit is likely to be easier than for those unused to managing their money in such ways, such as people living more 'cash in hand.' This may be class-related since it also extends to people's habits in meeting bills and their other financial obligations. One of the problems with mild cognitive impairment or a dementia and other frailty is that older people may find money management more difficult. The social and behavioral sciences are beginning to tell us more about what aspects of financial management are affected by cognitive impairment such as memory, calculation, and estimates of risk (Manthorpe and Samsi, 2012).

Some older people in touch with social care services want to know the details of the rules of personal budgets, such as how they can be used, what they can be used on, and how they may increase or decrease as circumstances change. In England there are numerous examples of people paying for personal care and housework (particularly cleaning), respite breaks, food, transport, ironing, travel, household repairs, computer packages, and lessons (see Newbronner et al., 2011; Feltoe and Orellana, 2013). **Box 2** sets out the guidance provided to older people by one local authority.

On the other hand, some older people have complained that they are not permitted to use their personal budget as they choose. **Box 3** sets out the limits placed by one local authority.

Much of this knowledge touches upon the need for financial skills; particularly organizational abilities; not only to run a financial system generally but also being able to manage change and the unexpected. Numeracy skills are not only important in terms of Direct Payment management but also as a way of knowing and understanding the processes agreed with the local authority. Being able to manage budgets, records, prioritize and spread expenditure resonates with the theory of 'responsibilization' or of 'coproduction' (Needham and Carr, 2009). The areas requiring financial capacity or competence are listed below, and **Box 4** identifies specific elements needed to engage with banking services:

- How to save money each month to put toward something.
- Costs – e.g., what is a 'reasonable' cost of a care worker?
- What expenses come out each month of the Personal Budget?
- Knowing incoming and outgoing amounts every month.
- Having a simple table (spreadsheet) to outline the year's costings and a simple calculator/formula to help social workers see how a personal budget would be spent.
- Providing tips/advice/tools for effective record keeping.
- Tools to assist with this practically – calculator, computer, and so on.

Making Personal Budget Systems Easier for Older People

Not surprisingly, there have been concerns that older people find such systems difficult and the recent years have seen developments to improve access to personal budgets and better support for older people in particular. Evidence from one long-standing state Cash and Counseling Demonstration and Evaluation (CCDE) (Arkansas) program in the US is that building in flexibility and support structures, such as representatives,

Box 2 How can I use my Personal Budget?

Your Personal Budget must be used to meet your needs, as agreed in your Support Plan. Some examples of how you can use your Personal Budget are:

- helping you with personal care, such as bathing and getting dressed
- helping you with day-to-day tasks at home
- helping you get out and about
- helping you to find and pay for day activities, such as a course or a club
- getting help to find work
- paying for equipment such as a computer, which may help you to be more independent
- anything else that helps you achieve what is expected in your Support Plan.

Source: London local authority, 2009.

Box 3 Stipulations around the use of Personal Budget: users are *not* permitted to

- pay for anything illegal
- gamble
- buy alcohol or food
- pay a family member who lives with you to care for you
- pay for anything that the National Health Service (NHS) would pay for
- pay for care outside the UK for longer than 4 weeks (other than in exceptional circumstances)
- pay utility (e.g., fuel) bills, rent or mortgage

Please note, other sources of funding should be explored prior to the allocation of a Personal Budget.

Source: London local authority, 2009.

Box 4 Understanding banking: necessary competencies

- Choosing the right account for the personal budget and understanding it.
- Being aware of the costs of running the account, bank charges, overdraft charges, operations of standing orders, direct debits, etc.
- Arranging the user's contribution to budget/bank accounts.
- Understanding of the time frames around any delays in paying money in – allow time before withdrawal.
- Understanding of the processes and implications of telephone/internet banking (systems for access to account).
- Understanding bank statements, arranging for them to be in paper/large print.
- Rules for changing account.
- Safeguarding against fraud.

consultants, and fiscal intermediaries, safeguards consumers and program funding alike.

One development is greater publicity for positive and personal examples. Recent evidence on day opportunities among older people using personal budgets (funded by local authorities) offers some illustrations of the ways in which people may use such money to choose activities that are tailored to their identity or interests. Personal budgets and self-directed support may offer increased opportunities for people to have greater choice in terms of how they spend their time and with whom. However, it will be important to assess whether closures of traditional day services have a differential impact on people, especially those with very severe disabilities and their carers who may find it hard to access the day opportunities of mainstream and specialist social and leisure activities even with very intensive support.

To provide greater confidence for older people and those supporting them that there are options for great choice and control that will improve rather than diminish their quality of life; 'theories of organizational change' (see Iles and Sutherland, 2001) provide helpful perspectives, particularly those suggesting that thinking and doing are mutually informative. First, the calls for a safety net of support need thought as well as action so that professionals can respond to the immediate concerns, crises, and breakdowns of arrangements that are likely to occur among older people, their support networks and social care providers. Community development approaches may foster the building of close-knit communities but by their very nature these are locally determined and more informal than organized. While the importance of these has been discussed in relation to older people in rural areas (Manthorpe and Stevens, 2009), where care workers may live far from their employers (Dale and Brown, 2005), such concerns are general in nature.

Greater nuances in the reported experiences of older people receiving personal budgets also highlight the heterogeneity of older people. Some older people do not report personal budgets as transformational, as expressed by the respondents in Woolham and Benton's study (2012) some aspects may be preferable but some compromises may be necessary. A mixed picture may seem more authentic to older people and their families than accumulations of resolutely 'transformed lives.' It may help to make sense of what older people report to be 'hassle' around new arrangements, and their instincts that, of all age groups, they do not consider all change as an improvement, nor does it happen without costs or some negative outcomes.

A third theory that may help to understand personalization in practice relates to the professional 'gaze' or surveillance of older people. In service terminology, this refers to the type and level of monitoring around personal budgets and what to do if 'misuse' occurs. It is evident that practitioners have regretted the missed opportunities to scrutinize the tensions of competing policy 'agendas' around choice and control with safeguarding people from abuse under the banner of personalization (Manthorpe et al., 2009; Hunter et al., 2012). Older people are not used to scrutiny in respect of disability benefits and may associate closer monitoring with the reinvention of the 'snoopers' of the Poor Law system. While debates about surveillance in old age have often concentrated on it in institutional settings (Kenner, 2008) personalization throws up potentially new forms of regulation and self-regulation. This includes the 'responsibilization' thesis of becoming employers with the legalities of checks and controls that this status incurs.

These matters, of safeguarding and of monitoring, cast light on the new relationships being forged between older people and social workers (and between older people and the state). On the one hand, personal budgets are likely to intensify consumer behavior and consumer identities among some older people and their families (see Stevens et al., 2011) and consumerism in pensions (Mann, 2006) can be seen as a forerunner of personalization of care and support. Systems of resource allocation will be clarified and this may lead to an increase in numbers of older people and their families feeling that it is worth the 'hassle' or trouble of engaging with public forms of social care. This form of empowerment will make demands upon practitioners who have often seen themselves as possible advocates for older people or have acted as gatekeepers to resources.

Finally, the need for 'responsibilization' and consequently financial literacy among older people has largely been confined to debates around income and resources in or approaching retirement. This suggests the value of seeing personalization as part of the political economy of old age (Mann, 2006). This influential gerontological theory continues to offer helpful insights into the macrolevel of personalization. The development of options for social care presents challenges for older people who may face them at times of increasing frailty. It will be important to consider ways in which older people may develop understanding, skills and confidence in consumer roles prior to illness or disability or may prefer less individualistic transactions.

Conclusion

The discourse of choice and control tends to veer on the hypothetical and abstract; far less on the microlevel financial transactions, the management of the money, and day-to-day decisions. Aspects such as contract law, employment obligations, and training all feature when describing the administration of personal budgets, but less often articulated are the interactions of managing a disability, coping with changes in income, relationship building and sustaining, or nonfinancial transactions. There is a danger that, in stressing the work and microlevel financial aspects of personal budgets, the wider potential for personalizing care and support may evaporate. However, insights from theories of the political economy of old age suggest the value of understanding shifts in and connections between financial and individual responsibilities. 'Training' to be an older person in need of care and support may seem overly burdensome and characteristic of a highly atomized society where social support and interdependence are no longer marks of citizenship. Personal budgets and their equivalents are seen as a cornerstone to achieving the policy aim of enabling older people and their carers to take control and exercise choice over their care and support. To achieve these benefits, a variety of approaches to supporting older people in the process of planning and management seem necessary (see Newbronner et al., 2011) The question remains: will sufficient resources be available for

adequate personal budgets and to employ professionals who can sustain the flexible and creative approaches needed to support different groups in achieving their desired outcomes? Netten et al. (2012) concluded that a policy with potential for considerable improvements in people's lives is taking place in a context that undermines its capacity for success.

Acknowledgments and Disclaimers

This article was funded by the Department of Health under its funding of the Social Care Workforce Research Unit but the views contained in this article are those of the author.

See also: Aging and Health in Old Age; Community Care in England; Disability Movement; Disability and Disabling Practices; Disability, Geography of; Disability: Social and Psychological Aspects; Disability: Sociological Aspects; Elder Abuse; Elder Care; Implementation Science; Old Age: Definitions, Theory, and History of the Concept.

Bibliography

Alakeson, V., 2010. Issues in International Health Policy: International Developments in Self-directed Care. The Commonwealth Fund, New York.

Dale, S., Brown, R., 2005. The Effect of Cash and Counseling on Medicaid and Medicare Costs: Findings for Adults in Three States, Final Report. Mathematica Policy Research, Inc., Princeton, NJ. http://mathematica-mpr.net/publications/pdfs/cashandcounseling3states.pdf.

DH (Department of Health), 2008. Transforming Social Care. Local Authority Circular LAC (DH) (2008) 1. Department of Health, London.

Dixon, J., Manthorpe, J., Biggs, S., Mowlam, A., Tennant, R., Tinker, A., McCreadie, C., 2010. Defining elder mistreatment: reflections on the United Kingdom study of abuse and neglect of older people. Ageing & Society 30 (3), 403–420.

Feltoe, E., Orellana, K., 2013. Making Managed Personal Budgets Work for Older People. What Older People Want, What Works and How to Achieve the Best Outcomes. Age UK, London.

Glendinning, C., Challis, D., Fernandez, J.-L., Jacobs, S., Jones, K., Knapp, M., Manthorpe, J., Moran, N., Netten, A., Stevens, M., Wilberforce, M., 2008. Evaluation of the Individual Budgets Pilot Programme: Final Report. Social Policy Research Unit, University of York, York. http://php.york.ac.uk/inst/spru/pubs/1119/.

HM Government, 2007. Putting People First. The Stationary Office, London.

Hunter, S., Manthorpe, J., Ridley, J., Cornes, M., Rosengard, A., 2012. When self-directed support meets adult support and protection: findings from the evaluation of the SDS test sites in Scotland. Journal of Adult Protection 14 (4), 206–215.

Iles, V., Sutherland, K., 2001. Organisational Change: A Review for Health Service Managers, Professionals and Researchers. NIHR NCC Service Delivery and Organisation, London. http://www.netscc.ac.uk/hsdr/files/project/SDO_FR_08-1001-001_V01.pdf.

Kenner, A.M., 2008. Securing the elderly body: dementia, surveillance, and the politics of "aging in place". Surveillance & Society 5 (3), 252–269.

Low, L.F., Yap, M., Brodaty, H., 9 May 2011. A systematic review of different models of home and community care services for older persons. BMC Health Services Research 11, 93. http://dx.doi.org/10.1186/1472-6963-11-93.

Low, L.-F., Chilko, N., Gresham, M., Barter, S., Brodaty, H., 2012. An update on the pilot trial of consumer-directed care for older persons in Australia. Australasian Journal on Ageing 31, 47–51.

Mann, K., 2006. Three steps to heaven? Tensions in the management of welfare: retirement pensions and active consumers. Journal of Social Policy 35 (1), 77–96.

Manthorpe, J., Moriarty, J., 2010. Nothing Ventured, Nothing Gained: Risk Guidance for People with Dementia. Department of Health, London.

Manthorpe, J., Moriarty, J., 2013. Examining day centre provision for older people in the UK using the Equality Act 2010: findings from a scoping review. Health & Social Care in the Community. http://dx.doi.org/10.1111/hsc.12065. Article first published online: 17 August 2013.

Manthorpe, J., Moriarty, J., Rapaport, J., Cornes, M., Iliffe, S., Clough, R., Bright, L., OPRSI, 2008. There are wonderful social workers but it's a lottery. British Journal of Social Work 38 (6), 1132–1150.

Manthorpe, J., Stevens, M., 2009. Increasing care options in the countryside: developing an understanding of the potential impact of personalization for social work with rural older people. British Journal of Social Work 40 (5), 1452–1469.

Manthorpe, J., Stevens, M., Rapaport, J., Harris, J., Jacobs, S., Challis, D., Netten, A., Knapp, M., Wilberforce, M., Glendinning, C., 2009. Safeguarding and system change: early perceptions of the implications for adult protection services of the English individual budgets pilots – a qualitative study. British Journal of Social Work 39 (8), 1465–1480.

Manthorpe, J., Samsi, K., 2012. 'Inherently risky?': personal budgets for people with dementia and the risks of financial abuse: findings from an interview-based study with adult safeguarding coordinators. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs023. First published online 27 March 2012.

Moran, N., Glendinning, C., Wilberforce, M., Stevens, M., Netten, A., Jones, K., Manthorpe, J., Knapp, M., Fernandez, J.-L., Challis, D., Jacobs, S., 2013. Older people's experiences of cash-for-care schemes: evidence from the English individual budget pilot projects. Ageing & Society 33 (5), 826–851.

Needham, C., Carr, S., 2009. Coproduction: An Emerging Evince Base for Adult Social Care Transformation. Social Care Institute for Excellence (SCIE), London.

Netten, A., Jones, K., Knapp, M., Ferandez, J.L., Challis, D., Glendinning, C., Jacobs, S., Manthorpe, J., Moran, N., Stevens, M., Wilberforce, M., 2012. Personalisation through individual budgets: does it work and for whom? British Journal of Social Work 42 (8), 1556–1573.

Newbronner, L., Chamberlain, R., Bosanquet, K., Bartlett, C., Sass, B., Glendinning, C., 2011. Keeping Personal Budgets Personal: Learning from the Experiences of Older People, People with Mental Health Problems and Their Carers. Social Care Institute for Excellence (SCIE), London.

Ottmann, G., Allen, J., Feldman, P., 2013. A systematic narrative review of consumer-directed care for older people: implications for model development. Health & Social Care in the Community. http://dx.doi.org/10.1111/hsc.12025.

Rummery, K., 2011. A comparative analysis of personalisation: balancing and ethic of care with user empowerment. Ethics and Social Welfare 5 (2), 138–152.

San Antonio, P., Eckert, J.K., Simon-Rusinowitz, L., 2006. The importance of relationship: elders and their paid family caregivers in the Arkansas cash and counseling qualitative study. Journal of Applied Gerontology 25, 31–48.

San Antonio, P., Simon-Rusinowitz, L., Loughlin, D., Eckert, J.K., Mahoney, K.J., Ruben, K.A.D., 2009. Lessons from the Arkansas cash and counseling program: how the experiences of diverse older consumers and their caregivers address family policy concerns. Journal of Aging & Social Policy 22 (1), 1–17.

Scottish Government, 2013. Self Directed Support Act. Scottish Government, Edinburgh. http://www.scotland.gov.uk/Topics/Health/Support-Social-Care/Support/Self-Directed-Support/Bill (accessed 08.07.13.).

Stevens, M., Glendinning, C., Jacobs, S., Moran, N., Challis, D., Manthorpe, J., Fernandez, J.-L., Jones, K., Knapp, M., Netten, A., Wilberforce, M., 2011. Assessing the role of increasing choice in English social care services. Journal of Social Policy 40 (2), 257–274.

Ungerson, C., Yeandle, S. (Eds.), 2007. Cash for Care in Developed Welfare States. Palgrave Macmillan, Basingstoke.

Woolham, J., Benton, C., 2012. The costs and benefits of personal budgets for older people: evidence from a single local authority. British Journal of Social Work. http://dx.doi.org/10.1093/bjsw/bcs086. First published online 20 July 2012.

Self-Efficacy

August Flammer, University of Berne, Bolligen, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Purposive actions as well as positive self-esteem presuppose corresponding self-efficacy beliefs, that is, beliefs that one has the capacity to efficiently use means that lead to desired states. On the opposite, people who perceive themselves as helpless are unhappy and are not motivated for actions. The concept of self-efficacy, as it is mostly in use today, emerged between the forties and the eighties (Heider, Rotter, Seligman, Weiner, Bandura, E. Skinner).

This article treats the main concepts related to self-efficacy, their theoretical and historical roots, their functions and practical uses, as well as developmental and educational/therapeutic aspects. The individual development of self-efficacy beliefs has its roots in early contingency experiences (starting in the first weeks after birth), in the use of degrees of freedom and in the experience of success and failure depending on appropriate causal attributions. These basic experiences also hint at the most effective educational and psychotherapeutic approaches. In front of an actual task, adults trust in their necessary abilities mainly on the basis of their memory of former experiences of comparable success (or failure, respectively). In comparison, adolescents and children are more optimistic than adults and give themselves more often a chance, even if the task is totally new and unknown to them.

Concepts

Everything that happens is caused to happen (Aristotle). Making changes means being a cause or providing a cause that produces a change. True causal changes are difficult to identify, because we are mostly confronted with causal chains and because most changes are due to a multitude of causes or conditions. A car accident may be considered as due to exceed speed, but it may also depend on street conditions, lighting, car conditions, car handling, weather, and so on. What is *the* cause? In everyday life people often choose the most salient condition or an unusual condition to be identified as cause. Here, we prefer the term condition(s). An effect is contingent upon a condition or a set of conditions if it always occurs if the condition or the set of conditions is met. Such conditions may be sufficient but not always necessary for producing the effect.

In this article we are interested in human actions as necessary conditions of change. The success of actions, too, depends on conditions. Considering person-related conditions of effective actions, we differentiate aspects like knowledge, initiative, perseverance, intelligence, experience, physical force, help from others, and more. Thus, instead of saying that an actor is able to produce a certain effect, we can say more elaborately that an actor is endowed with certain means or conditions that enable him or her to attain certain goals. Figure 1 represents a simple two-set causal chain: An actor who activates certain conditions that produce a certain effect.

We say that individuals (as well as groups) are in control of specific goals if they are able to produce the corresponding changes (horizontal line of Figure 1). More elaborately, people are in control of a specific goal if they are aware of the necessary contingencies and if they are competent enough to make these contingencies work (both diagonal lines in Figure 1). Instead of control, Bandura (1977) introduced the word efficacy, more specifically: self-efficacy. We use control and self-efficacy interchangeably.

Controlling means putting control into action. This is not equivalent to having control or being in control: People are in control of certain states or changes of affairs, if they can put control into action, even if they do not. For example, somebody has control over buying a new bicycle (having enough money, knowing where and how to execute a purchase), even if he or she does not buy it – even if he or she never in life buys a bicycle.

People may indeed be able to control certain events without knowing it; they will then eventually miss possible chances to activate such control. On the other hand, people may believe that they have control over a certain range of affairs, but in fact do not. That may make them feel good as long as there is no need to put this control into work, but they will be frustrated if they try to put their believed control into action and possibly regret not having prepared themselves for real self-efficacy.

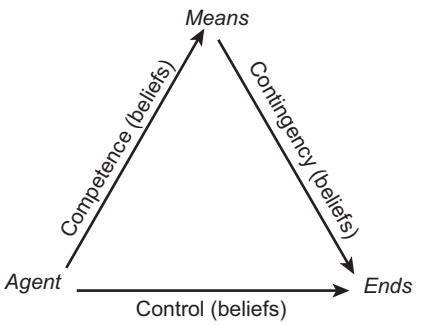

Figure 1 Means-ends relations and agency as components of control. Adapted from Skinner, E.A., Chapman, M., Baltes, P.B., 1988. Control, means-ends, and agency beliefs: a new conceptualization and its measurement during childhood. Journal of Personality and Social Psychology 54, 117–133; Flammer, A., 1990. Erfahrung der eigenen Wirksamkeit [Experiencing One's Own Efficacy]. Huber, Bern, Switzerland.

Obviously, it is important that people not only have control, but that they also know that they have control; they need corresponding control beliefs. Not being in control of an important situation is equivalent to being helpless in this respect (Seligman, 1975). The same is true, if a person only believes that he or she be not in control of an important situation. It has been proved that the psychological effects of helplessness (HL) are different depending on whether the helpless person believes being helpless for ever (chronic HL), whether being helpless is unique (personal vs universal HL), and whether HL is related to a specific domain or to most domains of life (specific vs global HL). Erroneously believing in personal and chronic HL in important domains is a tragic buffer against changes to the better. On the other side, chronic and universal HL over weather conditions does not hamper the person's self-value, because nobody can control the weather. But personal HL over important events that other people are in control of is difficult to accept (cf severe bodily handicaps). In sum, the psychological consequences of believing not to be in control over important domains are fourfold. Severely helpless people are (1) deeply sad about not having control, (2) demotivated to take initiatives or to invest effort and perseverance, (3) cognitively blind for any alternative or better view of the state of the world, and (4) devaluate themselves.

Obviously, at least in subjectively important domains, we prefer self-efficacy to helplessness. As long as we are not completely demotivated we either strive for self-efficacy in important domains (by fighting, learning, or training) or search for compensation. A common type of compensation consists of seeking help or delegating personal control (indirect control or proxy control), for example, to pay a gardener for caring one's garden or to put a doctor in charge with one's health, or to pray to God for a favor in a seemingly hopeless situation.

Another way of compensating noncontrol is to use secondary control (Rothbaum et al., 1982). While control as discussed above (i.e., primary control) consists of making the world fit with one's goals and aspirations, secondary control accommodates personal aspirations or personal interpretations of the actual state to make them fit with the world. As an example, a person who was not admitted to physics studies of his or her University may later 'find out' that working day and night in physics laboratories makes one miss the beauties of the rest of the world.

Very often people act in groups or teams. Take a soccer team as an example. Victory depends on the actions of each individual (even on the actions of the opposed team's individuals, etc.). However, each individual's action is really effective only if it is coordinated with the others' actions. Self-efficacy belief therefore also depends on the trust in others' capabilities and their coordination. The same is true for each team mate: Each of them has to believe in himself or herself and even in the efficacy as well in the self-efficacy beliefs of each other. Thus successful teams have to collectively care for their collective self-efficacy (Bandura, 2000; Stajkovic et al., 2009).

History of the Concept of Self-Efficacy

In the 1950s, Rotter (1954) proposed the concept 'locus of control,' meaning the place where control of desired reinforcement for behavior is exerted. Internal control means that the person gains reinforcement by his or her own sovereign actions. External control means that agents outside the person determine whether reinforcement occurs, possibly powerful others, situational conditions, chance, or luck. Rotter and his associates have developed valid measuring instruments that have been used in thousands (!) of studies demonstrating that internal locus of control is positively correlated with almost all desirable attributes of humans.

Fritz Heider (1944), an Austrian researcher of social perception, studied the subjective interpretations of (i.e., 'causal attributions' for) observed actions and did so by using the concepts of internality and externality. The true origin of an observed action is attributed either to the person (internal) or to person-independent conditions (external). In the first case, a person is responsible for the action outcome, in the second case he or she is not. Heider's work has triggered a large research tradition on causal attributions, whereby cause is understood in a very large sense (see above). Results of this research were of use for successful differentiations of subjective interpretations concerning experiences of control and of helplessness (*see* Attributional Processes: Psychological). Consequently, attribution theory remains an important element of self-efficacy theory.

Modern self-efficacy theory goes beyond Rotter's theory insofar as it is more differentiated (e.g., contingency vs competence, primary vs secondary, personal vs universal, chronic vs transient, specific vs global), distinctively referred to specific domains of actions (e.g., health, school), and elaborated to also include more personality aspects like motivation and development.

Self-Efficacy as an Important Element of a Happy and Successful Person

Individuals with high self-efficacy beliefs also report strong feelings of well-being and high self-esteem in general. Self-efficacy beliefs influence motivation, perseverance, social attitude, health, recovering from illness, learning efficiency, and so on (Bandura, 1997; Flammer, 1990; Maddux, 2009). It feels good to be good. Such persons are willing to take initiatives in related domains, to apply effort if needed, and persevere in efforts as long as they believe in their efficacy. Clearly, such initiatives may sometimes be unsuccessful, but more often they provide learning chances and lead to new experiences opening up new perspectives of the world and of the personal development (snowball effect). Optimistic persons with a high sense of self-efficacy do not give up easily and are able to overcome difficulties that others do not.

It is well documented that very low self-efficacy beliefs can be a basis of depression, especially if it is coupled with the conviction that not only the concerned person can change the situation but that nobody else is able to help (chronic and universal helplessness, i.e., hopelessness).

As to gender differences, in average – but only in average – females are less confident of their self-efficacy than males; the other side of the coin is that males are more often overconfident than females. It is a common finding that girls often outperform boys in school but still hold lower self-efficacy beliefs than

boys. As a consequence, girls less often show self-enhancing attribute patterns than boys, that is, have a stronger tendency to attribute failure to insufficient abilities and success to external factors. This is especially true in achievement domains that were long hold as masculine. Possibly – hopefully – new socialization patterns (new self-experience and more veridical and accurate feedback) will make these gender differences disappear (e.g., Lloyd et al., 2005).

In highly self-efficient individuals, potentially stressful situations produce less subjective stress, because such individuals are confident in having the competencies for finding ways to overcome hurdles. However, while self-efficacy beliefs act as buffers against stress, they can also – indirectly – produce stress insofar as they can induce overly ambitious individuals to assume more responsibility than they are able to cope with. Either the goals are in fact beyond their competence or the sheer quantity of chosen goals to be attained exceed the persons capacity.

Self-efficacy beliefs have also been reported to exert a positive influence on recovery from surgery or illness and on healthy lifestyles. If you believe that your health fully depends on fate or on medical art and effort, you do not personally think and care about what in your life-style could be changed to become or to stay healthy. However, if you think that personal health is at least partly a personal performance, you try to do your best.

High self-efficacy beliefs not only often root in success experience, they also produce success (Schunk and Pajares, 2009; Zimmerman and Schunk, 2009). In the school domain the circular effects between belief, effect, and belief again are easily apparent. Those who believe in themselves invest more effort and as a result of it experience more success, which in turn boosts the belief in further success. Likewise, students who are convinced of their incompetency refrain from effort (or simply feign effort) and consequently have little success, which in turn proves that they are right in not believing in their self-efficacy (circulus vitiosus). Not surprisingly, these good and bad circles have also been studied systematically in the area of sports, especially in high-performance sports (Feltz et al., 2008). Athletes as well as coaches are well advised taking lessons from these insights.

Interestingly, it has been demonstrated repeatedly and in several cultures that in most domains healthy and happy individuals tend to slightly overestimate themselves. If people are asked to compare themselves with the general population of their kind and indicate whether they are about the same (0) or healthier or more good looking or brighter (1 or even 2) or whether they feel they are below average (−1 or even −2), the average of the answers is clearly above 0. This has also been found with respect to self-efficacy: In average, people feel more self-efficient than average. Furthermore it has been found that realistic estimation of self-efficacy is rather typical for persons vulnerable to depressed mood, and clear underestimation increases the chance for a clinical (reactive) depression. Slightly overestimating one's own self-efficiency is healthy and boosting development toward the better, which allows for even higher expectations (snowball effect). However, there is a limit somewhere: Major overestimation can result in painful and harmful clashes with the reality and finally severely undermine realistic judgments.

The Development of Self-Efficacy Beliefs

Even the newborn baby has efficacy, but not really efficacy beliefs. The basic structure of the self-efficacy beliefs develops within the first 3 or 4 years. According to Flammer's (1995, 2009) analysis, the infant's development toward the basic understanding of self-efficacy proceeds through a developmental sequence consisting of the acquisition of (1) the basic event schema, (2) the elementary causal schema, (3) the understanding of personally producing effects, (4) the understanding of success versus failure in aiming at nontrivial goals, and (5) the discovery of being not only the origin of one certain change but also capable of producing such changes more generally.

The basic event schema consists in being aware that certain classes of events happen. These events are mostly very complex like: adequately sucking, smelling, tasting, and drinking milk, or moving the legs, perceiving rhythmical movements in the body, and seeing a suspended toy swinging above the head. The fact that such movements are becoming more 'rounded' and adapted, that is, that learning takes places, indicates that the baby experiences what happens. By the way, this is what Karl Bühler almost 100 years ago has defined as 'Funktionslust' and what Piaget not much later has theoretically interpreted as primary circular reaction.

The elementary causal schema is reached when babies start to recognize that two or more events hang together in the sense that one follows the other. The pleasure provided by tickling games or by seeing and then hearing objects banging on a hard surface shows that the baby is expecting effects, that is, contingency.

Soon after the second stage, the baby starts to experience herself as producing effects. She realizes that certain effects are produced by others that and some effects are their own. If they like it, then they may repeat the same action again and again; and later on the sheer visibility of a pertinent object or situation can trigger the resumption of the causal action. This is what Piaget called secondary circular reaction. Very soon, these are followed by tertiary circular reactions: Infants not only try to reach goals but also test possible means or causes through varying their actions and finding out what different changes they produce. Clearly, not all actions are successful, but this does not matter much to the baby so far, she simply tries again. Success and failure, however, have much more impact on the next stage.

Infants in their third year or so are not only happy in producing effects, but show pride in successfully reaching nontrivial goals like building a high tower with wooden game blocks. A high tower done by her mother may be admired, but a high tower done by one-self magnifies the ego. Reversely, failure despite hardly trying to reach an important and publicly visible goal results in shame: Such infants may corporally sink down in their chair and lower the eyes. Apparently, self-efficacy is more than reaching effects, it means something to the self-value.

Finally, the belief of one's self-efficacy generalizes. It is not just this or that effect one has successfully reached or not, but it is one's own personality that is effective, successful – or not with all the further consequences alluded to above.

Obviously, this stagewise development proceeds in domains that are accessible by the infant so far. A major opening of new domains comes with school entrance. As to

the domain of school success, within the second half of the first decade of life, the child learns more and more differentiations of means toward the same ends. Thus, he or she gradually abandons a global concept of simply being or not being able and singles out – probably in this sequence – (1) the factor effort (more effort is needed to solve tasks, which is a typical lesson to be learned early in school), (2) the factors individual ability and task difficulty (higher difficulty requiring more ability; higher ability allows mastering more difficult tasks), and finally (3) the understanding of the compensatory relation between effort and ability (it is possible to reach the same goals by being less capable but more hard working; not yet mastering certain tasks calls for learning and acquiring higher ability).

In adolescence and in early adulthood more lessons have to be learned; more and more domains become accessible to personal control due to increased cognitive, physical, or economic strength and social power. This is exciting, indeed. However, individuals have to permanently select from the choices that are offered to them (Flammer, 2009). Trying to control everything results in overburdening. One thing is to deselect control domains because they compete with higher priority control domains; another thing is to be forced to renounce control because no accessible contingencies seem to exist. As long as there are enough attractive alternatives available, renouncing is not very painful, but it can severely hurt handicapped individuals and old people when they lose control of important domains.

Old people are well advised not to resign too early but eventually to search for compensations. Such compensations consist of artifacts of all kinds (from memory aids to hearing aids), but they also include the above-mentioned compensations like indirect control (social resources) and secondary control. Indeed, it seems that the extent and the importance of secondary control increase with the lifetime (Heckhausen and Schulz, 1995). In old age, more and more control is taken over by care persons. A few homes for old people even take over so much control over their patients (this Latin expression means endure, bear, tolerate) that M. Baltes and Silverberg (1994) have even suggested that people in such homes adjust better if they give up personal control at all. Alluding to the concept of learned helplessness, they called such behavior learned dependency. Learned dependency helps to avoid certain social conflicts; the only remaining personal control then is the control over giving in.

The Microgenesis of Self-Efficacy Beliefs

If a person starts an action, he or she is convinced or at least hopes to be able to successfully perform. Remembering earlier experiences is probably the most reliable basis for the actual belief. However, if the person is confronted with a concrete difficult and possibly difficult task, how does he or she come to believe in being able or not being able to solve it? – This is the microgenetic question (as opposed to ontogenesis (lifelong development of the individual), the German Gestaltists called the actual or short-term building up or emergence of an idea or of a perception as 'Aktualgenese,' which we translate as microgenesis).

Even if people have had appropriate past experiences, relevant knowledge has to be activated, retrieved, or reconstructed from memory. And even then the result may be an underestimation or an overestimation of one's actual control. As an example, clinically depressive persons tend to remember more negative experiences than positive ones – even if compared with the proportion of really experienced success and failures – thus negatively biasing their expectations for further actions even more.

In a series of studies, the author has interviewed adolescent and adult subjects about several possible tasks and challenges. After the response (yes or no) he eventually asked them in a seemingly casual way 'Why?'. The results indicated that individuals most often refer to dispositional attributes, that is, ready-made generalized beliefs (i.e., 'because I am strong'). Second in frequency were references to personally experienced episodes (surprisingly, there were almost no references to models, i.e., to other people who did the same, successfully or not). However, when later asked about their confidence in believing to be able or not to be able, the subjects indicated more confidence in those control beliefs for which they had referred to concrete experiences than in those based on dispositional attributes. This in spite of the fact that they referred to dispositions more often than to experienced episodes. Also, the latencies before the yes/no answers were shorter, when the subsequent reference answer was episodical rather than dispositional. Taken together with the finding that in the oral format the episodic answers were more frequent than in the written format, these results indicate that people prefer the episodic foundation of their self-efficacy beliefs, but often do not have ready access to relevant remembered episodes. That episodical reference is taken as trustworthier is also shown by the result that people are also more confident in other people's assertion of competence if they refer to repeated former experiences than if they refer either to their general abilities or to one unique experience.

There is also an interesting interaction between ontogenesis and microgenesis. The author and his coworkers have also confronted elementary school children with the same frame of questioning. Unlike adults, children often simply asserted how they would do (i.e., being totally confident that they are able). If questioned again, they often answered – very optimistically – that they would try very, very hard. In one respect, this corresponds with earlier data from adolescents: Compared with adults, adolescents referred more to dispositions, when they positively believed in their self-efficacy, but less when they believed in their nonefficacy. It seems that adolescents – like children – are still very optimistic as to their self-efficacy unless they have experienced specific failures. They are probably not only right but also serve their developmental chances.

Another source of one's own capacity estimates (as well as overestimate and underestimate) is the actual mood state. Positive mood raises control self-efficacy belief, negative mood lowers it, which obviously has the potential of vicious and eventually virtuous circles. Failure produces disappointment and bad mood; bad mood makes failure more salient, so that the individual either avoids challenges or chooses unsolvable tasks; this produces failures and the repeated 'perception' of one's own deficiencies. The reverse is true as well: Positive

mood induces initiative, effort, and perseverance, which may lead to success, and success raises the mood – and adds another positive event to be encoded in memory.

Educational and Therapeutic Aspects

Parents and educators can influence not only the actual self-efficacy beliefs but also life-long generalized trust in one's competence. Contingent behavior by the caregivers is crucial already within the first weeks of life. Caregiver's behavior should be predictable, that is, contingent upon the baby's actual behavior – even more upon the baby's perceptions, feelings, and desires. Contingent behavior fosters not only children's happiness, but also their curiosity and their willingness to learn. This requires high sensitivity toward the child.

Fortunately, parental empathy – a basic condition for contingent behavior – is a natural gift in the majority of attentive parents. But contingent and reinforcing behavior is not always easy. External stress, health, and too frequent changes in caregiving teams make the educational task difficult. In addition, if the baby shows unorganized behavior (so-called difficult babies), the caregiver's task is not only difficult but often less rewarding.

Infants and children need freedom for experimentation. Caregivers are to give the necessary freedom, but to prevent from dangerous experimentation. With the infant getting older, the caregiver has to help her to understand success and failure adequately (causal attribution): An action might be too difficult at a certain age or the young child may be encouraged to learn or simply to try again and invest more effort (Schneewind, 1995).

Clearly, self-efficacy beliefs need to be carefully maintained or even boosted in adolescents and in adults as well. Giving a good job interview, delivering a public speech, or successfully participating in a sports competition can be much helped by a sound belief in one-self. The same is true for elite sports and the systematic and enduring training that needs perseverance and optimism even in case of setbacks. Feltz et al. (2008) are a perfect example of how to train systematically self-efficacy beliefs of athletes, teams, and even coaches. Evidently, not only individuals need to believe in their own efficacy, but teams as well.

Psychotherapy with persons whose self-efficacy beliefs are severely undermined is difficult. Helpless people not only tend to interpret failures to their disadvantage, they also often play down their contribution to eventual success. They therefore lack or even avoid initiatives. Persons who have lost self-confidence immunize themselves against being hurt by not trying anymore, by fatalistical attributions, and even by recalling their biography in a way that is consistent with and justifies their dysfunctional beliefs. Teaching and trying to convince them that they are really capable despite their belief not to be does not help much. Helping them to recall from memory prior success experiences instead of being impressed by failures only is (a bit) more efficient. More efficient are new and successful experiences.

Conclusion

Within the past decades, theory and research have established self-efficacy beliefs as important elements in the understanding of human action and human well-being in a very large sense. The theory is rather well established, empirically founded, and linked to a multitude of other scientific domains. Further steps may include more differentiation between life domains, between cultures and – if possible – between historical times. Given the importance of life-long high self-efficacy beliefs, the developmental conditions surely merit more attention by researchers.

See also: Collective Efficacy and Crime; Conscious Control During Childhood, Development of; Self-Efficacy and Health; Self-Efficacy: Education Aspects.

Bibliography

Baltes, M.M., Silverberg, S.B., 1994. The dynamics between dependency and autonomy. In: Featherman, D.L., Lerner, R.M., Perlmutter, M. (Eds.), Life-span Development and Behavior. Erlbaum, New York, pp. 41–90.

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavioral change. Psychological Review 84, 191–215.

Bandura, A., 1997. The Exercise of Control. Freeman, New York.

Bandura, A., 2000. Exercise of human agency through collective efficacy. Current Directions in Psychological Science 9, 75–78.

Feltz, D.L., Short, S.E., Sullivan, P.J., 2008. Self-efficacy in Sport. Human Kinetics, Champaign, IL.

Flammer, A., 1990. Erfahrung der eigenen Wirksamkeit [Experiencing One's Own Efficacy]. Huber, Bern, Switzerland.

Flammer, A., 1995. Developmental analysis of control beliefs. In: Bandura, A. (Ed.), Self-efficacy in Changing Societies. Cambridge University Press, New York, pp. 69–113.

Flammer, A., 2009. Entwicklungstheorien [Theories of Development], fourth ed. Huber, Bern, Switzerland.

Heckhausen, J., Schulz, R., 1995. The life-span theory of control. Psychological Review 102, 284–304.

Heider, F., 1944. Social perception and phenomenal causality. Psychological Review 51, 358–374.

Lloyd, J.E.V., Walsh, J., Yailag, M.S., 2005. Sex differences in performance attributions, and achievement in Mathematics. Canadian Journal of Education 28, 384–408.

Maddux, J.E., 2009. The power of believing you can. In: Lopez, S.J., Snyder, C.R. (Eds.), Oxford Handbook of Positive Psychology, second ed. Oxford University Press, New York.

Rothbaum, F., Weisz, J.R., Snyder, S.S., 1982. Changing the world and changing the self: a two-process model of perceived control. Journal of Personality and Social Psychology 42, 5–37.

Rotter, J.B., 1954. Social Learning and Clinical Psychology. Prentice-Hall, Englewood Cliffs, NJ.

Schneewind, K.A., 1995. Impact of family processes on control beliefs. In: Bandura, A. (Ed.), Self-efficacy in Changing Societies. Cambridge University Press, New York, pp. 1–45.

Schunk, D.H., Pajares, F., 2009. Self-efficacy theory. In: Wenzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 35–54.

Seligman, M.E.P., 1975. Helplessness. Freeman, San Francisco.

Skinner, E.A., Chapman, M., Baltes, P.B., 1988. Control, means- ends, and agency beliefs: a new conceptualization and its measurement during childhood. Journal of Personality and Social Psychology 54, 117–133.

Stajkovic, A.D., Lee, D., Nyberg, A.J., 2009. Collective efficacy, group potency, and group performance: meta-analyses of their relationships, and test of a mediation model. Journal of Applied Psychology 94, 814–828.

Zimmerman, B.J., Schunk, D.H., 2009. Self-regulated Learning and Academic Achievement, second ed. Routledge, New York.

Self-Efficacy and Health

David P French, School of Psychological Sciences, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by A. Bandura, volume 20, pp. 13815–13820, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the literature on self-efficacy in producing health-related behavior. It shows that the evidence that self-efficacy predicts many such behaviors is strong, but the evidence of causation is not entirely compelling. It evaluates the evidence on how best to increase self-efficacy, and on the mechanisms by which self-efficacy influences behavior. Throughout, evidence is used to evaluate propositions derived from relevant theory, i.e., social cognitive theory.

Introduction

Definition

Self-efficacy has been defined as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997: p. 3). Thus, self-efficacy is an individual's perceptions of control over his or her ability to successfully complete a behavior or behaviors, in order to bring achieve his or her goals.

Historical Development

Self-efficacy, as a concept, was emerged from social learning theory, which has several variants (e.g., Rotter, 1954; Bandura, 1977a). Social learning theories were proposed to contrast with previous learning theory approaches to understand behavior, which emphasized the key role of conditioning through reinforcement of a person's behavior. By contrast, the key insight these various social learning theories shared is that, as the name suggests, people often learn by observing others, i.e., by learning within a social context. An influential series of studies for social learning theory were Bandura's 'Bobo doll' studies, where nursery school children saw adults hitting inflatable dolls with bats (Bandura et al., 1963). The main influence on whether the children copied this behavior was the extent of reward for the others, indicating that the children were learning by observation of their social context, not by being personally reinforced, the form of learning emphasized by learning theory.

This work had demonstrated that outcome expectancies were important in determining behavior, i.e., beliefs that a behavior would lead to a particular outcome. Social learning theory subsequently incorporated discussion of beliefs about capability to perform the behavior, i.e., self-efficacy, which were also shown to be important (1977b). The precise nature of the factors involved, how they were brought about, and how they influenced behavior were elaborated in the subsequent social cognitive theory (Bandura, 1986).

Role of Self-Efficacy in Social Cognitive Theory

According to social cognitive theory (Bandura, 1986), the behavior that individuals choose to carry out is chosen as it helps them achieve their goals (see **Figure 1**). Self-efficacy is hypothesized to be an important determinant of these behaviors, in combination with outcome expectancies, i.e., beliefs about the likely outcomes of the behavior. For example, if a person has a goal of weight loss, he or she may consider that not eating candy would help them achieve that, which is an outcome expectancy. In this example, whether a person attempts to reduce eating candy would depend upon whether they believe they can successfully do so, i.e., their self-efficacy with regard to this behavior. It is important to note that, high self-efficacy for performing a behavior on its own is unlikely to produce a behavior where a person has a low outcome expectancy, i.e., they do not believe that performing that behavior will bring about a particular goal. Equally, if a behavior leads to a goal that is not valued by that person, e.g., they do not wish to lose weight, levels of self-efficacy will be irrelevant.

The importance of self-efficacy, according to social cognitive theory, lies in the insight that a person only believing that performing a behavior, e.g., not eating candy, will lead to a desired outcome, e.g., losing weight, is not sufficient to produce behavior. A person must also have a sufficient level of self-efficacy, i.e., must believe that he or she has the capability to not eat candy for the behavior to be brought about. There is often much variation in levels of self-efficacy for many common health-related behaviors, given that they are difficult to carry out, e.g., stopping smoking, avoiding eating tempting sugary or fatty foods, or increasing the levels of physical activity.

Role of Self-Efficacy in Multiple Theories of Behavior

Self-efficacy has subsequently been recognized as one of the most important constructs within psychology, with a wealth of applications in health, educational, and work settings, among others (Bandura, 1997). Within the health sphere, it is now a core construct, being adopted by many leading theories of behavior, including the transtheoretical model (Prochaska and Velicer, 1997), protection motivation theory (Rogers, 1983), and the theory of planned behavior (Ajzen, 1988) in the form of the closely related construct of perceived behavioral control.

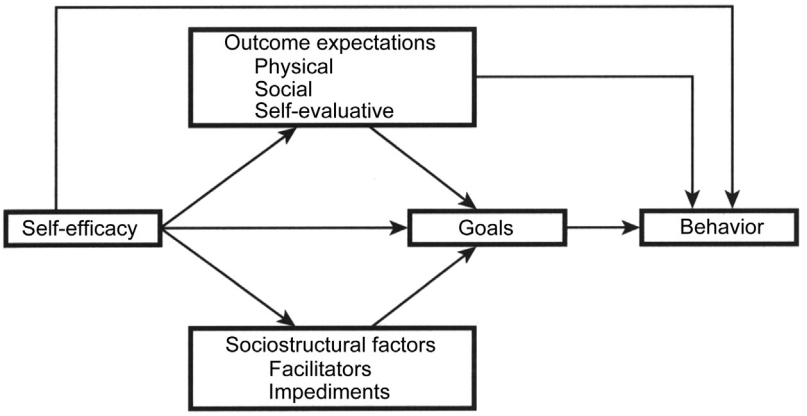

Figure 1 Paths of influence through which perceived self-efficacy and other key social cognitive factors regulate motivation and performance accomplishments.

Role in Determining Health-Related Behavior

Evidence Required to Establish Causation by Self-Efficacy

It is important to be precise when describing the role of self-efficacy in bringing about behavior. A construct such as self-efficacy can be a correlator or predictor of subsequent behavior without necessarily being causally involved in producing that behavior: they could both be causally produced by a third construct, such as social support. There is stronger evidence for causation when self-efficacy mediates change of behavior following an experimental intervention; although, this is not sufficient to demonstrate causation (see French, 2013). It is widely agreed that one of the most effective methods of increasing self-efficacy is for a person to successfully perform that behavior (Bandura, 1997; see also below). Given that, an intervention may change behavior via another mechanism, e.g., by forming effective plans, which is likely to increase self-efficacy in turn. In this scenario, mediation analysis may suggest that self-efficacy is the mechanism responsible for changing behavior when it is not.

Evidence for Self-Efficacy as a Determinant of Health-Related Behavior

There is overwhelming evidence that self-efficacy is a reliable predictor of many health-related behaviors, including healthy eating, screening, safer sexual behavior, and smoking, and also of changes in many health-related behaviors (Luszczynska and Schwarzer, 2005). There is also good reason to believe that it can play a causal role in behavior change, although the evidence for this is not entirely compelling (French, 2013).

As an example of this, there is a wealth of evidence that self-efficacy is a consistent correlate of physical activity in adults (Bauman et al., 2012). However, self-efficacy had mediated the effects of few interventions to increase physical activity (Rhodes and Pfaeffli, 2010). Two good reasons for this pattern of results are many interventions fail to change self-efficacy in the first place, which makes them essential uninformative about the role of self-efficacy in changing behavior and many interventions do not increase physical activity; there are no effects to mediate (French, 2013). However, there are some studies, which provide convincing evidence of mediation, and grounds for belief that self-efficacy played a causal role. For example, one intervention study specifically targeting increasing self-efficacy as a method to increase walking behavior, and produced large increases in objectively assessed walking (Darker et al., 2010). Notably for the present purposes, the intervention was delivered in a single session, and a large change in self-efficacy was observed over the course of this session. This change in self-efficacy mediated the effects of the intervention on walking behavior subsequent to this intervention session. As well as satisfying the statistical tests of mediation, the timings of intervention and measurement in this study make explanations for this mediation other than self-efficacy having a causal role in changing behavior, unlikely.

It is worth noting that there is stronger evidence that self-efficacy is a determinant of many physical activity interventions for older adults (van Stralen et al., 2009), and solid evidence that self-efficacy is a mediator of many interventions with children and adolescents (Lubans et al., 2008). In general however, there is a need for further well-designed intervention studies to further test the hypothesis that changes in health-related behavior can be brought about by increasing self-efficacy.

Reciprocal Determinism

Much research on social cognitive theory has been concerned with the role of individual factors such as self-efficacy in producing behavior. Despite this, the theory proposes that there are three major classes of determinants (internal personal factors, behavior, and the external environment), which all influence each other (Bandura, 1986). Thus, if a person changes his or her behavior, due to having high levels of self-efficacy, it can in turn influence his or her environment, e.g., a person quitting smoking makes in more likely that others close to them will also quit (Christiakis and Fowler, 2008), thereby shaping their environment to make it more supportive of not smoking.

Equally, the theory emphasizes the crucial role of the environment in shaping individuals and their behavior. Much behavior, including health-related behavior, is socially

patterned (Steptoe et al., 2010), e.g., there is a clear decline in smoking status associated with higher incomes. There are likely to be many reasons for these patterns, but by highlighting the influence of the environment on behavior, the theory highlights the limitations of interventions that try to change an individual's unhealthy behavior without also trying to alter the environment, which makes that unhealthy behavior more likely. It also indicates that low self-efficacy may often be a consequence of adverse environments, e.g., a person who lives in a neighborhood where shops do not sell fresh fruit and vegetables faces more barriers for eating a healthy diet, and would have to expend more effort in doing so than a person living in a more supportive environment.

What Is the Best Method for Increasing Self-Efficacy?

Bandura's Four Sources of Information about Self-Efficacy

Bandura (1977b) proposed four main sources that people draw upon when evaluating their levels of self-efficacy: mastery information, vicarious experience, persuasion, and affective/physiological feedback.

Mastery experience is generally held to be the strongest and most convincing source of information (Bandura, 1997): to the extent that a person has successfully performed a behavior in the past, such as exercising regularly, they will tend to be confident that they can do so in the future. Similarly, if a person has repeatedly failed at performing a behavior in the past, e.g., quitting smoking, then his or her self-efficacy for future quitting will tend to be low. Vicarious experience refers to seeing another person performing the behavior, as with the example cited earlier of hitting a bobo doll. Its influence on self-efficacy for that behavior will be strongest when the other person is seen as being similar in terms of salient features. For example, seeing a famous sportsman modeling engagement with exercise is unlikely to lead to much increase in self-efficacy for exercise in a fairly sedentary person in middle age. In contrast, seeing a person successfully engaging with exercise is more likely to increase self-efficacy, where that person is perceived as being similar in terms of levels of activity, age, and gender especially where other personally salient barriers are similar, e.g., a chronic condition is shared. Persuasion is generally held to be the least convincing source of information (Bandura, 1997), consisting of messages designed to convince that the behavior is easy or under a person's control or that the messages have the capability to perform the behavior. Finally, affective/physiological feedback refers to the influence of internal sensory or emotional cues on evaluations of self-efficacy. For example, if a person trying to quit smoking is feeling depressed or lethargic, it may make him or her feel that the demands of quitting are unmanageable, and hence beyond his or her control. Similarly, feeling tired and achy following an intense burst of exercise in a sedentary person may convince that person that he or she does not have the capability to exercise on a regular basis.

It is important to note that for all these sources of information about self-efficacy, the key factor is a person's interpretation of that information, not the information per se that is important in forming that person's evaluation of self-efficacy (Bandura, 1977b).

Which Behavior Change Techniques Lead to Increases in Self-Efficacy?

Subsequent to Bandura's work, other sources of self-efficacy have been proposed, notably imagery (Maddux, 2002). In particular, there is some evidence that encouraging people to imagine successful sporting performance can enhance self-efficacy, and possibly performance (e.g., Callow and Waters, 2005).

To test Bandura's theorizing, reviews have assessed whether there are differences in the extent to which interventions change self-efficacy, according to whether behavior-change techniques (BCTs) are included or not. One review identified interventions aiming to increase lifestyle or recreational physical activity (Ashford et al., 2010), and coded the contents based on a coding scheme derived from the four sources proposed by Bandura (1977b). This review found that techniques involving mastery experience or persuasion were the most commonly used, being used by respectively, 33 and 34 interventions out of a total of 37 interventions included in the review. Interventions using persuasion produced less of an increase in self-efficacy than interventions that did not include persuasion. In contrast, those interventions which involved vicarious experience of success by others were less common, being used in only nine interventions, but were associated with larger increases in self-efficacy when compared with those that did not involve vicarious experience.

More recent reviews have used standardized taxonomies of BCTs, which are general in scope, and not linked to self-efficacy specifically. These reviews have identified those techniques, such as action planning (i.e., getting a person to specify where, when, and how to perform the behavior), that were associated with larger increases in self-efficacy for physical activity when included in interventions with nonclinical populations (Williams and French, 2011). Thus, these reviews allow the identification of apparently effective BCTs that are derived from theories other than social cognitive theory. Further, these reviews have also examined, which BCTs are associated with increases in physical activity. A notable finding is that there is a high degree of association ($r = 0.7$) between change in self-efficacy for physical activity and physical activity behavior in nonclinical populations, as well as a high degree of association ($r = 0.7$) across change in physical activity and self-efficacy when examined by BCT (Williams and French, 2011). Thus, within this population, those BCTs that were associated with larger increases in self-efficacy were also associated with larger increases in physical activity; conversely, those BCTs that were not associated with changes in self-efficacy also tended not to be associated with changes in physical activity. It should, however, be noted that a review of studies with obese participants did not find such an association, suggesting that self-efficacy for physical activity may not be as important as a determinant of physical activity in such populations (Olander et al., 2013).

How Does Self-Efficacy Determine Behavior?

According to Bandura (1977b), self-efficacy is a determinant of behavior, as it affects attempts at both initiating attempts at changing a behavior, and persistence of attempts at behavior

change. In terms of initiation, people who are low in self-efficacy with regard to a particular behavior will be disinclined to attempt that behavior. For instance, those who believe they cannot quit smoking will be reluctant to make a quit attempt. Further, as low feelings of self-efficacy for important goals result in lowered affect (cf Carver and Scheier, 1982), people tend to avoid even thinking about attempting that behavior. Thus, they will not tend to seek out information about changing this behavior, even when such information could potentially be of benefit, such as the helpful role of pharmacological agents that reduce cravings for nicotine in dramatically increasing the success rates for quit attempts. This motivated lack of information seeking is a contributor to 'defensive processing' in the face of information about the harms of a behavior: where a person thinks that he or she cannot change a behavior, tend to ignore or otherwise derogate messages about the harms of that behavior (Croyle et al., 1997).

In addition to being more likely to initiate an attempt at changing behavior, those high in self-efficacy are more likely to persist even in the presence of obstacles and aversive experiences. Hence, a smoker who believes that he or she can successfully quit is likely to make more attempts, as he or she will attribute the failure of a quit attempt to an unstable and potentially controllable factor, e.g., drinking too much at a party, rather than making a stable dispositional attribution: they are the kind of person who cannot stop smoking. Similarly, they may be more likely to attribute bodily signs such as nicotine craving as indicating that they are successfully removing the effects of smoking from their bodies, rather than as a sign that their body needs nicotine.

More recently, theorists have proposed that people may have self-efficacy relating to specific actions that constitute an overall behavior. Specifically, in the area of addictions, Marlatt et al. (1995) proposed five distinct types of self-efficacy: resistance, harm reduction, action, coping, and recovery self-efficacies, relating to addictive substances, such as tobacco and alcohol. Each of them relate to feelings of control, overrespectively, avoidance of addictive substances, minimizing risk once having tried the substance, quitting/abstaining from taking the substance, coping with barriers that arise from abstaining, and dealing with lapses in abstinence. More recently, refinements to this scheme have been proposed by others, with relevance to health behaviors in general (e.g., Luszczynska and Schwarzer, 2003). They proposed instead pre-action self-efficacy, relating to self-efficacy relating to formation of intentions to change, and two postintentional forms of self-efficacy: maintenance self-efficacy, which is similar to Marlatt et al.'s (1995) coping self-efficacy, and recovery self-efficacy, in line with Marlatt et al.'s (1995) definition, although it has been called relapse self-efficacy in later work. These latter distinctions are based upon the idea of there being distinct psychological processes involved in forming intentions and changing and maintaining behavior, in line with the Health Action Process Approach (HAPA; Schwarzer, 1992). The HAPA builds upon earlier theorizing within social cognitive theory, but emphasizes the role of stages of behavior change, and provides greater elaboration of the processes involved following formation of intentions. There is empirical evidence to support the utility of the distinctions between these different classes of self-efficacy in the maintenance of behavior (e.g., Kassavou et al., 2014).

Conceptual Issues

Specificity versus Generality of Self-Efficacy

An ongoing issue regarding the conceptualization of self-efficacy concerns the most appropriate level of specificity for the construct, and at which it should be measured and its effects assessed (Bandura, 1997). Bandura's work generally conceptualizes self-efficacy as being specific, in terms of either a behavior, or components of that behavior (1997). The logic of this is that a person may well be very confident that he or she can exercise regularly but not prevent him or her from eating sweet or fatty foods; in contrast, another person may feel in the control of their ability to avoid eating unhealthily but not feel in the control of increasing their levels of physical activity. These two individuals may therefore attempt to achieve the same goal (maintaining a healthy weight) by performing different behaviors. Following similar reasoning, as a function of self-efficacy, is to help people overcome barriers to them achieving their goals, it makes sense to examine self-efficacy for different behaviors separately, as the barriers are likely to be different.

Despite this, some theorists have proposed examining self-efficacy at a broader level, e.g., self-efficacy for achieving weight loss, which is clearly a goal rather than a behavior. At its most broad, the construct of general self-efficacy has been proposed (Schwarzer and Jerusalem, 1995), which is an overall stable sense of competence to overcome some barriers.

Distinguishing Self-Efficacy from Other Control Constructs

The issue of generality of self-efficacy highlights that there is a multitude of other constructs relating to control: Skinner (1996) identified over 100 constructs relating to control that were in use. Given this, it is important to be clear on how self-efficacy differs from other commonly used constructs.

One key distinction is between control over means of achieving ends, i.e., control over behavior, and control over the ends themselves, i.e., control over ends. Skinner (1996) argues that Bandura unambiguously defined self-efficacy in terms of control over behavior in his earlier writing (e.g., Bandura, 1977), but shifted over time to control over ends (e.g., Bandura, 1997). This has two implications. First, that self-efficacy, as latterly defined, becomes akin to Rotter's (1954) concept of locus of control, which relates to how in-control people feel of events that affect their lives. Secondly, and crucially, it has been argued that most measures of self-efficacy conform to latter definitions regarding control over ends (Williams, 2010). That is, measures of self-efficacy assess not only self-efficacy as intended, but also outcome expectancies. Williams (2010) argues that this invalidates research showing that measures of self-efficacy predict or mediate behavior, as it is possible that the effects are due to outcome expectancies.

Another distinction that has generated much discussion relates to the relationship of self-efficacy to the construct of perceived behavioral control in the theory of planned behavior. The author of the theory of planned behavior proposed that the differences were of negligible importance (Ajzen, 2002).

In light of these discussions, a distinction has been made between self-efficacy, considered to be confidence in their ability to engage in the behavior, and perceived controllability as perceptions of control over the behavior. It has been proposed that perceived behavioral control contains a mixture of both perceived control and self-efficacy. It is certainly the case that self-efficacy is much more predictive of behavior, such as binge–drinking, than is perceived control (see Cooke et al., 2014). However, there is evidence that commonly used measures of both self-efficacy and perceived behavioral control contain items that can be considered a mixture of both (Johnston et al., 2014).

Clarifying Conceptual Confusion

Taken as a whole, the criticisms that have been made of the conceptualization and operationalization of self-efficacy flag up that there is a greater need for clarity. Future measures of self-efficacy should be clearer whether they are concerned with confidence over ability to bring about behaviors or confidence in ability to bring about goals or other outcomes. Future measures of self-efficacy should avoid containing items that could be considered as perceived control. Most importantly, there is a need to develop measures that are not contaminated by also assessing outcome expectancies, so the unique effects of outcome expectancies and self-efficacy can be assessed.

Conclusion

Self-efficacy is a key construct in many theories of health-related behavior. It is consistently predictive of the performance of many of these behaviors. There is also some evidence that it plays a causal role for some of these behaviors, but further research is needed to provide stronger tests of this hypothesis. In particular, future research should focus on disentangling the role of self-efficacy and outcome expectancies in causing behavior, by using measures of self-efficacy that are contaminated by outcome expectancies.

See also: Alcohol Use among Young People; Explanatory Style and Health; Health Behavior, Psychosocial Theories of; Health Behaviors; Physical Activity and Health Promotion; Physical Activity and Health; Smoking and Health; Weight Regulation.

Bibliography

Ajzen, I., 1988. Attitudes, Personality and Behavior. Open University, Milton Keynes.
Ajzen, I., 2002. Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. Journal of Applied Social Psychology 32, 665–683.
Ashford, S., Edmunds, J., French, D.P., 2010. What is the best way to change self-efficacy to promote physical activity? A systematic review with meta-analysis. British Journal of Health Psychology 15, 265–288.
Bandura, A., 1977a. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.
Bandura, A., 1977b. Self-efficacy: toward a unifying theory of behavioral change. Psychological Review 84, 191–215.
Bandura, A., 1986. Social Foundations of Thought and Action. Prentice Hall, Englewood Cliffs, NJ.
Bandura, A., 1997. Self-Efficacy: The Exercise of Control. WH Freeman & Company, New York.
Bandura, A., Ross, D., Ross, S.A., 1963. Imitation of film-mediated aggressive models. Journal of Abnormal and Social Psychology 66, 3–11.
Bauman, A.E., Reis, R.S., Sallis, J.F., Wells, J.C., Loos, R.J.F., Martin, B.W., 2012. Correlates of physical activity: why are some people physically active and others not? Lancet 380, 258–271.
Callow, N., Waters, A., 2005. The effects of kinesthetic imagery on the sports confidence of flat-race horse jockeys. Psychology of Sport and Exercise 6, 443–459.
Carver, C.S., Scheier, M.F., 1982. Control theory: a useful conceptual framework for personality-social, clinical and health psychology. Psychological Bulletin 92, 111–135.
Christiakis, N.A., Fowler, J.H., 2008. The collective dynamics of smoking in a large social network. New England Journal of Medicine 358, 2249–2258.
Cooke, R., Dahdah, M., Norman, P., French, D.P., 2014. How well does the theory of planned behaviour predict alcohol consumption? A systematic review and meta-analysis. Health Psychology Review, in press.
Croyle, R.T., Sun, Y., Hart, M., 1997. Processing risk factor information: defensive biases in health-related judgments and memory. In: Petrie, K.J., Weinman, J.A. (Eds.), Perceptions of Health and Illness. Harwood, Amsterdam, pp. 267–290.
Darker, C.D., French, D.P., Eves, F.F., Sniehotta, F.F., 2010. An intervention to promote walking amongst the general population based on an "extended" theory of planned behaviour: a waiting list randomised controlled trial. Psychology and Health 25, 71–88.
French, D.P., 2013. The role of self-efficacy in changing health-related behaviour: cause, effect or spurious association? British Journal of Health Psychology 18, 237–243.
Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., Pollard, B., 2014. Discriminant content validity (DCV): a quantitative methodology for assessing content of theory-based measures. British Journal of Health Psychology 19, 240–257.
Kassavou, A., Turner, A., Hamborg, T., French, D.P., 2014. Predicting maintenance of attendance at walking groups: testing constructs from three leading maintenance theories. Health Psychology 33, 752–756.
Lubans, D.R., Foster, C., Biddle, S.J.H., 2008. A review of mediators of behaviour in interventions to promote physical activity among children and adolescents. Preventive Medicine 47, 463–470.
Luszczynska, A., Schwarzer, R., 2003. Planning and self-efficacy in the adoption of and maintenance of breast self-examination: a longitudinal study on self-regulatory cognitions. Psychology and Health 18, 93–108.
Luszczynska, A., Schwarzer, R., 2005. Social cognitive theory. In: Conner, M., Norman, P. (Eds.), Predicting Health Behaviour. Open University Press, Maidenhead, pp. 127–169.
Maddux, J.E., 2002. Self-efficacy: the power of believing you can. In: Snyder, C.R., Lopez, S.J. (Eds.), Handbook of Positive Psychology. Oxford University Press, New York, pp. 277–287.
Marlatt, G.A., Baer, J.S., Quigley, L.A., 1995. Self-efficacy and addictive behaviour. In: Bandura, A. (Ed.), Self-efficacy in Changing Societies. Cambridge University Press, New York, pp. 289–315.
Olander, E.K., Fletcher, H., Williams, S., Atkinson, L., Turner, A., French, D.P., 2013. What are the most effective techniques in changing obese individuals' physical activity self-efficacy and behaviour: a systematic review and meta-analysis. International Journal of Behavioral Nutrition and Physical Activity 10, 29.
Prochaska, J.O., Velicer, W.F., 1997. The transtheoretical model of health behaviour change. American Journal of Health Promotion 12, 38–48.
Rogers, R.W., 1983. Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo, J.T., Petty, R.E. (Eds.), Social Psychophysiology: A Source Book. Guilford Press, New York, pp. 153–176.
Rhodes, R.E., Pfaeffli, L.A., 2010. Mediators of physical activity behaviour change among adult non-clinical populations: a review update. International Journal of Behavioral Nutrition and Physical Activity 7, 37.
Rotter, J.B., 1954. Social Learning and Clinical Psychology. Prentice Hall, Englewood Cliffs, NJ.
Schwarzer, R., 1992. Self-efficacy in the adoption and maintenance of health behaviors: theoretical approaches and a new model. In: Schwarzer, R. (Ed.), Self-efficacy: Thought Control of Action. Hemisphere, Washington, DC, pp. 217–242.

Schwarzer, R., Jerusalem, M., 1995. Generalized self-efficacy scale. In: Weinman, J., Wright, S., Johnston, M. (Eds.), Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs. NFER-NELSON, Windsor, UK, pp. 35–37.

Skinner, E.A., 1996. A guide to constructs of control. Journal of Personality and Social Psychology 71, 549–570.

Steptoe, A., Gardner, B., Wardle, J., 2010. The role of behaviour in health. In: French, D., Vedhara, K., Kaptein, A.A., Weinman, J. (Eds.), Health Psychology, second ed. BPS Blackwell, Oxford.

van Stralen, M.M., Vries, H.D., Mudde, A.N., Bolman, C., Lechner, L., 2009. Determinants of initiation and maintenance of physical activity among older adults: a literature review. Health Psychology Review 3, 147–207.

Williams, D.M., 2010. Outcome expectancy and self-efficacy: theoretical implications of an unresolved contradiction. Personality and Social Psychology Review 14, 417–425.

Williams, S.L., French, D.P., 2011. What are the most effective intervention techniques for changing physical activity self-efficacy and physical activity behaviour – and are they the same? Health Education Research 26, 308–322.

Self-Efficacy: Education Aspects

Dale H Schunk, The University of North Carolina at Greensboro, Greensboro, NC, USA
Maria K DiBenedetto, The City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-efficacy, or perceived capabilities for learning or performing actions at designated levels, is grounded in Bandura's social cognitive theory that postulates reciprocal interactions among personal, behavioral, and social and environmental factors. Self-efficacy has been shown to influence various achievement outcomes, including motivation, learning, and self-regulation. Calibration, or how well self-efficacy corresponds to performance, can affect students' motivation and achievement. Self-efficacy is important for teachers as well as students. Areas for future self-efficacy research include assessment, developmental changes, cultural comparisons, and applications to learning with technology.

Self-efficacy refers to perceived capabilities for learning or performing actions at designated levels (Bandura, 1997). The operation of self-efficacy has been explored in various domains, including education, business, athletics, careers, health, and wellness. Researchers have found that self-efficacy influences individuals' learning, motivation, and self-regulation (Bandura, 1997; Pajares, 1996; Schunk and Pajares, 2009).

Self-efficacy is grounded in Bandura's (1986) social cognitive theory, which postulates reciprocal interactions among personal (e.g., cognitions, beliefs, skills, effects), behavioral, and social and environmental factors (Schunk, 2012). Social cognitive theory reflects a view of human agency in which individuals exert a large degree of control over the outcomes of their actions. They hold beliefs that allow them to influence their thoughts, feelings, actions, social interactions, and aspects of their environments. Thus, teachers who want to improve their skills enroll in courses and seek mentoring from more-experienced teachers. But at the same time, people are influenced by their actions and facets of their social environments. Teachers are apt to change their behaviors in response to advice from their mentors.

According to social cognitive theory, strategies for increasing well-being can be aimed at improving emotional, cognitive, or motivational processes, increasing behavioral competencies, or improving aspects of one's social and physical environments. Teachers are responsible for promoting learning among their students. Using social cognitive theory as a framework, teachers can improve their students' emotional states and correct their faulty beliefs and habits of thinking (personal factors), raise their academic skills and self-regulation (behaviors), and alter classroom features and social interactions (social and environmental factors) to ensure student success.

This article discusses the role of self-efficacy in education contexts. Following a review of self-efficacy theory, factors affecting self-efficacy and the measurement and prediction of self-efficacy are discussed. Applications of self-efficacy to self-regulation and teaching are explained. The article concludes with suggestions for future research.

Self-Efficacy Theory

Self-efficacy – a key variable in social cognitive theory – is hypothesized to influence behaviors and environments and in turn be affected by them (Bandura, 1986, 1997; Schunk, 2012). Students with high self-efficacy for learning are apt to engage in self-regulation (e.g., set goals, use effective learning strategies, monitor their comprehension, evaluate their goal progress) and create effective environments for learning (e.g., eliminate or minimize distractions, find effective study partners). In turn, self-efficacy can be influenced by the outcomes of behaviors (e.g., goal progress, achievement) and by inputs from the environment (e.g., feedback from teachers, social comparisons with peers).

Bandura (1997) postulated that people acquire information to gauge their self-efficacy from their performance accomplishments, vicarious (e.g., modeled) experiences, forms of social persuasion, and physiological indexes (Figure 1). Learners' performances provide the most reliable information for assessing self-efficacy because they are tangible indicators of one's capabilities. Successful performances raise self-efficacy, whereas failures can lower it, although an occasional failure (success) after many successes (failures) may not have much of an impact.

Individuals acquire much information about their capabilities through knowledge of how others perform (Bandura, 1997). Similarity to others is a cue for gauging self-efficacy (Schunk, 1995). Observing similar others succeed can raise observers' self-efficacy and motivate them to try the task because they may believe that if others can do it they can as well, although a vicarious increase in self-efficacy can be negated by subsequent performance failure. Persons who

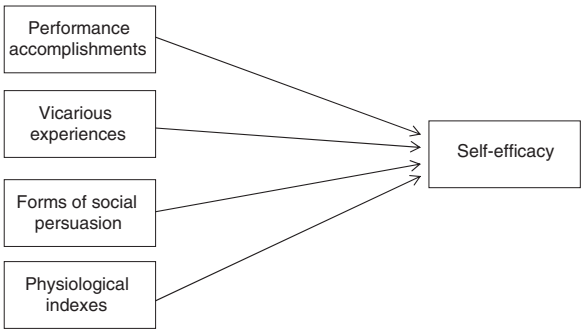

Figure 1 Sources of self-efficacy information. Reproduced from Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

observe similar peers fail may believe they lack the competence to succeed, which may not motivate them to attempt the task.

Individuals also develop self-efficacy from the social persuasions (e.g., "I know you can do it") they receive from others (Bandura, 1997). Social persuasions must be credible and cultivate people's beliefs in their capabilities for successfully attaining outcomes. Although positive feedback can raise self-efficacy, the increase will not endure if students subsequently perform poorly (Schunk, 1995). Conversely, negative persuasions can lower self-efficacy.

Individuals acquire self-efficacy information from physiological and emotional states, such as anxiety and stress (Bandura, 1997). Strong emotional reactions to a task provide cues about anticipated success or failure. When students experience negative thoughts and fears about their capabilities (e.g., feeling nervous thinking about taking a test), those reactions can lower self-efficacy and trigger additional stress and agitation that help ensure the inadequate performance. Learners may feel more efficacious when they perceive that they are less anxious about academic outcomes.

Although self-efficacy is important, it is not the only influence on behavior. Social cognitive theory contends that no amount of self-efficacy will produce a competent performance when the necessary skills to succeed are lacking (Schunk, 1995). Students' *values* (perceptions of importance and utility of learning) also can affect behavior (Wigfield et al., 2004). Even students who feel highly efficacious in science may not take science courses that they believe are not germane to their goal of becoming a medical doctor. Also important are *outcome expectations*, or beliefs about the anticipated outcomes of actions (Bandura, 1997). Students typically engage in activities that they believe will result in positive outcomes and avoid actions that they believe may lead to negative outcomes, even when they feel efficacious about attaining the latter. Assuming requisite skills and positive values and outcome expectations, social cognitive theory predicts that self-efficacy is a key determinant of individuals' motivation, learning, self-regulation, and achievement (Schunk, 2012).

Self-efficacy can influence the choices people make and the actions they pursue (Bandura, 1997). Individuals tend to select tasks and activities in which they feel competent and confident and avoid those in which they do not. Unless people believe that their actions will produce the desired consequences, they have little incentive to engage in those actions.

Self-efficacy also helps determine how much effort students expend, how long they persist when confronting obstacles, and how resilient they are in the face of adversity. Students with a strong sense of self-efficacy approach difficult tasks as challenges to be mastered rather than as threats to be avoided. They set challenging goals and maintain strong commitment to them, heighten and sustain their efforts in the face of failure, and quickly recover their sense of self-efficacy after setbacks.

Distinctions with Other Variables

Other variables bear some conceptual similarity to self-efficacy (Schunk and Zimmerman, 2006). *Self-concept* refers to one's collective self-perceptions formed through experiences with and interpretations of the environment and influenced by reinforcements and evaluations by others (Shavelson and Bolus, 1982). Self-concept typically is viewed as hierarchically organized with a general self-concept on top and subarea self-concepts below, and multidimensional elements, such as self-confidence and self-esteem (Marsh and Shavelson, 1985; Schunk and Pajares, 2009). Self-perceptions of specific competencies influence subarea self-concepts (e.g., subject areas such as history and biology), which in turn combine to form the academic self-concept. Because self-efficacy involves perceived capabilities in specific areas, it should contribute to self-concept (Schunk and Pajares, 2009).

Self-esteem is a general affective evaluation of one's self that often includes judgments of self-worth (Schunk and Pajares, 2009). Like self-concept, it differs markedly from self-efficacy. Self-efficacy involves questions of can (e.g., 'Can I write this essay?'), whereas self-esteem reflects questions of feel (e.g., 'How do I feel about myself as a writer?'). One's beliefs about what one can do may bear little relation to how one feels about oneself, as seen in students with high self-efficacy for learning but low self-esteem because they are socially unpopular with peers.

Perceived control (or *agency*) also differs from self-efficacy. In Bandura's (1997) social cognitive theory, agency is the capability to control one's life events. One's system of personal agency includes self-efficacy and outcome expectations (discussed earlier). Although self-efficacy is a key component of personal agency (Bandura, 1997), it is not the only one. A responsive environment is necessary for self-efficacy to exert its effects. Students may feel self-efficacious for learning yet make no effort to do so because they believe that in their present environment learning will not be rewarded.

The notion of perceived control figures prominently in the theory of planned behavior (Ajzen, 1985, 1991). In this theory, the intention to engage in a behavior is its primary determinant. In turn, intention is affected by the perceive value of the outcome of the behavior, social norms, and pressures from others to perform the behavior, and perceived behavioral control (i.e., perceived difficulty in performing the behavior). The effects on behavior of variables such as self-efficacy are mediated through their influence on intention. Some research shows that perceived behavioral control corresponds closely to self-efficacy (Bandura, 1997). Furthermore, placing a heavy burden on intention as the primary determinant of behavior is risky, because there are many reasons why people do not always act based on their intentions. Education researchers have shown that self-efficacy can affect achievement both directly and indirectly through other variables (e.g., persistence, effort; Pajares, 1996; Schunk and Pajares, 2009).

Self-efficacy differs from *self-confidence*, or a general belief about one's capabilities that often does not specify the object of the belief (e.g., 'She is highly self-confident'). Self-efficacy is situated within Bandura's (1986) social cognitive theory and has a clear and specific meaning. Although self-confident individuals often are self-efficacious, there is no automatic relation between these variables. People who know they will fail at a task are highly self-confident but have low self-efficacy (Bandura, 1997).

Factors Affecting Self-Efficacy

Research has explored the effects of instructional and other classroom processes on self-efficacy (Schunk and Pajares, 2009). This research shows that instructional, social, and environmental factors inform students about their learning, and they use this information to assess their self-efficacy for continued learning.

At the outset of an activity, students' self-efficacy for learning is a function of their prior experiences, personal qualities (e.g., abilities, attitudes), and social supports. The latter include the extent that teachers, parents, and others encourage students to learn, facilitate their access to resources necessary for learning (e.g., materials, facilities), and teach them self-regulatory strategies that enhance skill development.

As students engage in learning activities, they are influenced by personal factors (e.g., goal setting, cognitive information processing) and situational variables (e.g., feedback, social comparisons). These influences provide students with cues about how well they are learning. Self-efficacy is enhanced when students believe they are performing well. Lack of success or slow progress will not necessarily lower self-efficacy if students believe they can perform better, such as by expending greater effort or using more-effective learning strategies (Schunk, 1995). In turn, self-efficacy enhances motivation, learning, self-regulation, and achievement.

Experimental research supports these hypothesized relations with students in different grade levels (e.g., elementary, middle, high, postsecondary), with diverse abilities (e.g., regular, remedial, gifted), and in different content areas (e.g., reading, writing, mathematics, computer applications; Schunk, 1995; Schunk and Ertmer, 2000). Some instructional and social processes that raise self-efficacy are having students pursue proximal and specific goals, exposing children to social models, providing students with performance and attri butional feedback, teaching students learning strategies, having learners verbalize strategies while they apply them, linking students' rewards to their learning progress, and having students self-monitor and evaluate their learning progress. These processes differ in many ways, but they all help to inform students of their learning progress, which raises self-efficacy.

Prediction of Education Outcomes

Research shows that self-efficacy for learning or performing tasks correlates positively and significantly with subsequent achievement on those tasks (Pajares, 1996; Schunk and Pajares, 2009). Correlations between academic self-efficacy and performance in investigations in which self-efficacy corresponds closely to the criterion task have ranged from .49 to .70; direct effects in path analytic studies have ranged from $\beta = .349$ to .545 (Pajares and Urdan, 2006). Self-efficacy explains approximately 25% of the variance in the prediction of academic outcomes beyond that of instructional influences. Using meta-analytic procedures, Multon et al. (1991) found that self-efficacy related to academic performance ($r_\mu = .38$) and accounted for 14% of the variance. Stronger effects were obtained with highly congruent self-efficacy and skill measures. Self-efficacy also correlates positively with indexes of self-regulation (e.g., goal setting, strategy use, performance monitoring; Schunk and Pajares, 2009).

Based on findings from meta-analysis, Brown and Lent (2006) reported that self-efficacy is a significant predictor of education interests and mediates the relation of prior performance and ability measures to interests. Using meta-analysis with self-beliefs that include self-efficacy and more-global measures, such as self-concept, Valentine et al. (2004) found a small positive influence of self-beliefs on achievement and that the influence was stronger with self-beliefs specific to academic domains (e.g., self-efficacy).

Researchers have used causal models to test the predictive and mediational power of self-efficacy. Using path analysis, Schunk (1981) found that self-efficacy exerted a direct effect on children's achievement and persistence in mathematics. Pajares and Kranzler (1995) demonstrated that mathematics self-efficacy has as powerful and direct effect on mathematics performance as does cognitive or mental ability. Zimmerman and Bandura (1994) found that self-efficacy affected achievement directly as well as indirectly through its influence on goals. Schunk and Gunn (1986) found that children's mathematical achievement was directly influenced by use of effective strategies and self-efficacy. Relich et al. (1986) showed that self-efficacy exerted a direct effect on achievement and that instruction had both a direct and an indirect effect on achievement through self-efficacy (Figure 2).

Self-Efficacy Assessment

The earliest self-efficacy studies were conducted in clinical settings. Bandura et al. (1977) administered a self-efficacy and a behavioral test to adults with snake phobias. The test items consisted of progressively more threatening encounters with a snake (e.g., touch it, allow it to sit in one's lap). For the self-efficacy assessment, participants designated which tasks they felt they could perform and rated how sure they were that they could perform the tasks they felt they could perform. To measure generality, participants rated the same tasks with a type of snake different from the type used on the test. This is a microanalytic research methodology; self-efficacy and skill were assessed at the level of specific tasks.

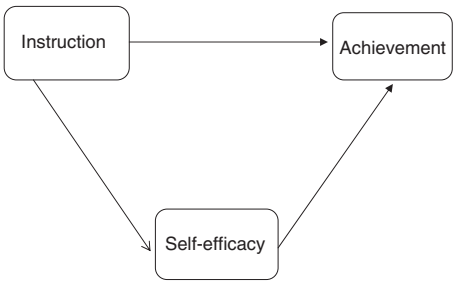

Figure 2 Self-efficacy effects on achievement. Reproduced from Relich, J.D., Debus, R.L., Walker, R., 1986. The mediating role of attribution and self-efficacy variables for treatment effects on achievement outcomes. Contemporary Educational Psychology 11, 195–216.

Participants were not asked for a general rating of how well they felt they could deal with snakes.

Early education studies used a similar methodology. Schunk (1981) gave children with low mathematical skills self-efficacy and skill tests. For self-efficacy, children were shown pairs of long-division problems; the two problems within each pair were comparable in form and difficulty. For each pair, they judged how certain they were that they could solve problems of that type. For the skill test, children attempted to solve problems that corresponded in form and difficulty to those on the self-efficacy test.

Since Bandura's (1977) initial formulation of self-efficacy, researchers have moved beyond his definition of self-efficacy as a domain-specific variable and assessed self-efficacy at more-general levels with such items as, "How well can you get teachers to help you when you get stuck on school work?" and, "How well can you study when there are other interesting things to do?" (Schunk and Pajares, 2009). To arrive at these types of judgments, students must integrate their perceptions across different situations (e.g., involving mathematics, science, social studies, and so forth).

There is some research evidence for a generalized sense of self-efficacy (Smith, 1989). Certain education conditions might foster it. School curricula are structured for positive transfer; for example, learning how to divide requires knowing how to estimate, subtract, and multiply. Students who perform well in mathematics might approach division with higher self-efficacy than those who have encountered difficulties.

Even in the absence of specific curricula structure, generalization also might occur when students believe that the new learning will require skills that they previously have mastered. Thus, writing a research paper requires identifying a topic, conducting a literature review, synthesizing and organizing information, and drawing conclusions. Students likely have used these skills to write papers in English. Students who feel competent in the requisite skills may have higher self-efficacy for writing a research paper than those who question their capabilities.

Finding evidence of generality would not refute the subject specificity of self-efficacy, but it is important to determine how students weigh and combine efficacy information to arrive at generalized judgments. Developmental factors may be an issue, because we should expect that with development students should be better able to assess their capabilities in component areas and determine the types of skills needed to succeed.

Calibration

The issue of *calibration*, or how well self-efficacy relates to actual performance on the corresponding tasks (Schunk and Usher, 2011), has been studied by researchers. When people judge that they are capable of performing a task and then perform it, or when they judge that they are incapable of performing it and cannot perform it, they are well calibrated because self-efficacy accurately predicts performance. Conversely, when people judge that they are capable of performing a task but do not perform it, or when they judge that they are incapable of performing a task but then perform it, they are poorly calibrated because of the lack of correspondence between self-efficacy and performance.

Calibration involves intentions, as stressed by the theory of planned behavior (Ajzen, 1991). In discussing implementation intentions, Gollwitzer and Sheeran (2006) noted that even with positive attitudes and social norms people may not perform behaviors if they doubt their capabilities of doing so successfully.

Calibration is educationally important. Motivation suffers when students who overestimate their capabilities perform poorly. Although self-efficacy that slightly exceeds what one can do is desirable because such overestimation can raise effort and persistence (Bandura, 1997), recurring overestimation can lead to continued failure with resulting decrements in students' motivation and learning.

Overestimation can result when students do not fully understand the demands of the task. Greater task experience informs students of the skills needed to succeed. Calibration also can be affected by instructional and social factors (Schunk and Pajares, 2009). Although instructional practices that provide information about skills required for the task can increase calibration (Schunk, 1981), such practices also can lower it. Students in low-ability groups from which they cannot move may feel demoralized and perform poorly, even though they feel efficacious about learning. Teachers who indiscriminately encourage students (e.g., "Come on, I know you can do this.") without ensuring that they learn skills may produce highly efficacious students who lack the skills to succeed.

The social cultures of schools also may affect calibration (Schunk and Pajares, 2009). Students may perform less than their best – and lower than their self-efficacy would predict – so that they do not become socially isolated as a consequence of being perceived by their peers as overly intelligent.

Self-Efficacy for Self-Regulation

Self-efficacy has been applied extensively to *self-regulation* (or *self-regulated learning*), which refers to learning that results from students' self-generated thoughts, feelings, and behaviors that are systematically oriented toward the attainment of their learning goals (Zimmerman, 2000). Self-regulation involves students' goal-directed activities that they instigate and sustain, such as focusing on task demands, applying effective strategies to learn, establishing productive social and work environments, assessing learning progress, and making strategic adjustments as needed. Developing and maintaining a sense of self-efficacy for self-regulated learning motivates students and promotes their learning (Schunk and Usher, 2011).

Zimmerman (2000) developed a self-regulated learning model that highlights the role of self-efficacy (Figure 3). This cyclical and recursive model includes three phases: forethought, performance, and self-reflection. The *forethought* phase involves processes students engage in prior to learning and includes motivational beliefs and task analysis. Key motivational beliefs are self-efficacy, outcome expectancies, intrinsic interest, and goal orientations (reasons why students want to learn). Task analysis includes goal setting (short- and

long-term goals) and strategic planning, or deciding what methods to use. Although each of these processes has been shown to be related to achievement, self-efficacy is a strong predictor and affects goal setting and strategic planning (Schunk, 2012; Zimmerman and Bandura, 1994).

The forethought phase initiates the learning activities that occur during the *performance* phase, during which students systematically and actively engage in learning. Key self-regulatory processes are self-control and self-observation. Students exert self-control by using such strategies as imagery, self-instruction, attention focusing, and others targeted at reaching goals (Zimmerman, 2002). Self-observation includes self-monitoring and self-recording learning progress. Students gain progress information from their perceptions and feedback from others (e.g., teachers, peers, parents). Research suggests that students who are highly self-efficacious about their capability to learn sustain their efforts and adapt their performances better than those with lower self-efficacy (Pajares, 2008).

On the basis of their self-monitoring and feedback from others, students form self-judgments and experience self-reactions in the *self-reflection* phase. Self-judgments include self-evaluations and attributions of performance to perceived causes (Schunk, 2012). Self-evaluations are based on performance standards that may derive from previous performances, performances by others (e.g., teachers), or absolute standards (Zimmerman, 2002). Causal attributions may reflect controllable causes, such as strategy use and effort, or uncontrollable ones, such as luck or ability (Weiner, 1985). Students also react to their performances with self-satisfaction and adaptive or defensive responses. Self-satisfaction refers to the level of contentment students feel about their performance relative to a standard or goal. Adaptive or defensive responses include emotional reactions to performances. Students who react defensively exhibit apathy, helplessness, procrastination, and cognitive disengagement for future learning to preserve images of self-worth (Garcia and Pintrich, 1994). Students who respond adaptively make adjustments to subsequent self-regulatory behaviors that may include modifying their motivational beliefs and task analyses.

As a consequence of self-reflections, learners may continue working on the task (i.e., the performance control phase) or return to the forethought phase for new goal setting and planning, thus forming the recursive loop in the self-regulation cycle. Students who are self-efficacious about their self-regulatory capabilities are likely to sustain efforts and persist even when they form negative evaluations (Schunk, 2012). These students may attribute performance to strategy use and effort and make adjustments in the forethought phase as they plan for subsequent learning.

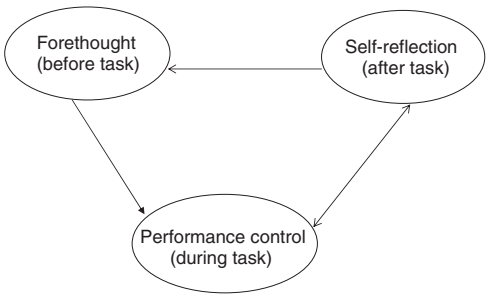

Figure 3 Phases of self-regulated learning. Reproduced from Zimmerman, B.J., 2000. Attaining self-regulation: a social cognitive perspective. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, San Diego, pp. 13–39.

Although self-efficacy is an integral motivational belief in the self-regulatory cycle, it is the development of self-regulatory competence that explains how self-efficacy becomes an internal motivational force behind self-regulated learning. Schunk and Zimmerman (1997) formulated a model of development of self-regulatory competence as consisting of four levels: observation, emulation, self-control, and self-regulation. This is not a stage model in which learners move to the next level after satisfying the criteria of the preceding level. Rather, learners may freely move among the levels while engaged in the same task as they develop increasing self-regulatory competence.

Initially, social models – such as teachers and peers – serve as important sources of self-regulatory behavior and self-efficacy. At the *observation* level, students observe models as they perform actions and learn many self-regulatory processes, such as strategic planning, self-monitoring, and performance adaptations (Zimmerman, 2000). In addition, students may observe models as they receive rewards, such as praise and good grades. These observations vicariously build students' self-efficacy for learning.

At the *emulation* level, students practice the behaviors previously demonstrated by the model and imitate the general pattern or style with assistance from the model (Zimmerman, 2000). Students receive reinforcement from the model and from their own senses and motoric feedback. Students' perceptions of their ability to discriminate and differentiate the self-regulatory processes needed to complete a task, along with the model's feedback, provide them with a sense of competence.

In these first two levels, learners derive information about their self-efficacy primarily from external (social) sources. It is at the third level of *self-control* when students begin to experience self-efficacy from within. Schunk and Zimmerman (1997) describe the self-control level as when learners are able to use strategies independently of the model. Students have internalized what they previously have observed and emulated; however, they are still using the representational patterns of the model to perform behaviors. During this level, students begin to experience self-efficacy internally as they are able reflect upon their capability of matching their work against the standards displayed by the model.

Students have reached the *self-regulation* level when they are able to systematically adapt their performance to different environmental and personal conditions (Zimmerman, 2000). At this level, students are motivated by their self-efficacy beliefs. They now are capable of initiating the use of strategies, making adjustments to behavior based on the situation, and evaluating their performance with the understanding that they are capable of competently making changes as needed. As self-regulated students engage in the three cyclical phases of self-regulated learning, they substantiate their sense of self-efficacy and regulate their thoughts, feelings, and behaviors toward goal attainment.

Teacher Self-Efficacy

Self-efficacy is applicable to teachers as well as students. *Teacher* (or *instructional*) *self-efficacy* refers to personal beliefs about one's capabilities to help students learn (Tschannen-Moran et al., 1998; Woolfolk Hoy et al., 2009). Self-efficacy should influence teachers' activities, effort, and persistence. Teachers with low self-efficacy avoid planning activities they believe exceed their capabilities, do not persist with students having difficulties, expend little effort to find materials, and do not reteach content in ways students might better understand (Ashton and Webb, 1986). Teachers with high self-efficacy are more likely to develop challenging activities, help students succeed, and persist with students who have trouble learning. These teacher behaviors enhance student learning and substantiate teachers' self-efficacy for helping students learn.

Ashton and Webb (1986) found that teachers with higher self-efficacy were likely to have a positive classroom environment (e.g., less student anxiety and teacher criticism), support students' ideas, and meet the learning needs of all students. High teacher self-efficacy is positively associated with teachers' use of praise, individual attention to students, and monitoring students' learning progress, as well as with higher student achievement.

Collective teacher self-efficacy represents teachers' beliefs that their capabilities as a group can enhance students' learning (Henson, 2002). Teachers develop collective self-efficacy when they work collaboratively to achieve common goals (performance accomplishments), learn from one another and have mentors who serve as role models (vicarious experiences), receive encouragement and support from administrators (forms of social persuasion), and work together to cope with difficulties and alleviate stress (physiological indexes). As collective self-efficacy is strengthened, teachers remain motivated to collaborate to improve students' learning. Collective self-efficacy bears a positive relation to teachers' job satisfaction (Caprara et al., 2003).

Future Directions

Researchers will continue to test Bandura's (1997) contention that self-efficacy beliefs can affect virtually every aspect of people's lives – their thoughts, feelings, and actions. One important research issue concerns the generality of self-efficacy. As discussed earlier, researchers have moved away from Bandura's contextual perspective and assessed self-efficacy with measures of greater generality. Decontextualized self-efficacy assessments that lack consistency with the criterion task can distort the influence of self-efficacy. To be explanatory and predictive, self-efficacy measures should reflect the task demands within the domain being analyzed. Evaluating the appropriateness and adequacy of a self-efficacy measure requires making theoretically informed and empirically sound judgments that reflect an understanding of the domain under investigation and its different features, the types of capabilities the domain requires, and the range of situations in which these capabilities might be applied. Self-efficacy measures then can be evaluated for their appropriateness using the level of specificity of their items, the range of task demands that they include, and their correspondence with the outcomes of interest (Bandura, 2006).

Research is needed on the role of self-efficacy at different developmental levels. In assessing self-efficacy, learners must integrate information from the environment, their prior experiences, and their beliefs about the present situation. Such complex cognitive processing may exceed the capabilities of younger children, which can decrease the predictive utility of self-efficacy. Most self-efficacy research has been conducted on older children, adolescents, and adults. Determining how younger children weigh and combine self-efficacy information from different sources will substantiate its relevance for learners of all ages.

There is a need for more research studies using students from different cultures. Most self-efficacy research studies have sampled from students in the United States. Cross-cultural studies will expand our understanding of the operation and generality of self-efficacy. Klassen (2004) reviewed several studies and found that self-efficacy tended to be lower for students from non-Western cultures (e.g., Asian and Asian-immigrant students) than for students from Western cultures (e.g., Western Europe, Canada, the United States). In some cases, the more modest self-efficacy expressed by the non-Western students predicted academic outcomes better than the higher self-efficacy of the Western students. The high level of achievement of Asian students may be a product of socialization practices that promote close parent–child relationships. In these cultures, relational efficacy (i.e., the confidence that students have in their familial and social relations), as well as social support from parents, has a powerful influence on students' academic performance. For these students, the lower levels of academic self-efficacy do not translate into lower academic performance.

Research is needed on self-efficacy as students learn from technology. The explosion of technology in instruction through electronic and distance education shows its potential to foster learning in ways that formerly were unknown (Brown, 2006). Researchers are investigating which strategies are helpful for learning from technology and how to teach students those strategies. Similar research is needed on how learning from technology affects students' self-efficacy for learning. Thus, students' performance accomplishments using technology should help build their self-efficacy. But students also learn through social networking, which suggests that other sources of self-efficacy information (vicarious, persuasive) also can build and sustain self-efficacy and motivation. A better understanding of how use of technology affects students' self-efficacy will have implications for the design of learning activities.

Conclusion

Theory and research support the important role of self-efficacy in learning, motivation, and self-regulation. From instructional and social and environmental sources, students gain information that affects their self-efficacy, and in turn, self-efficacy influences various achievement behaviors. The operation of self-efficacy in education contexts will continue to be clarified as researchers investigate such issues as the

generality of self-efficacy measures, how students' capabilities to process self-efficacy information changes with development, how self-efficacy functions in different cultures, and the role of self-efficacy in learning from technology.

See also: Academic Self-Concept and Achievement; Cognitive Development: Mathematics Learning and Instruction; Educational Learning Theory; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Learning to Learn; Mathematics Education; Self-Concepts: Educational Aspects; Self-Regulated Learning: Theories, Measures, and Outcomes; Teacher Behaviours and Student Outcomes; Teacher Education.

Bibliography

Ajzen, I., 1985. From intentions to actions: a theory of planned behavior. In: Kuhl, J., Beckman, J. (Eds.), Action-Control: From Cognition to Behavior. Springer, Heidelberg, Germany, pp. 11–39.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Ashton, P.T., Webb, R.B., 1986. Making a Difference: Teachers' Sense of Efficacy and Student Achievement. Longman, New York.

Bandura, A., 1977. Self-efficacy: toward a unifying theory of behavioral change. Psychological Review 84, 191–215.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice Hall, Englewood Cliffs, NJ.

Bandura, A., 1997. Self-efficacy: The Exercise of Control. Freeman, New York.

Bandura, A., 2006. Guide for creating self-efficacy scales. In: Pajares, F., Urdan, T. (Eds.), Self-Efficacy Beliefs of Adolescents. Information Age Publishing, Greenwich, CT, pp. 307–338.

Bandura, A., Adams, N.E., Beyer, J., 1977. Cognitive processes mediating behavioral change. Journal of Personality and Social Psychology 35, 125–139.

Brown, J.S., September/October 2006. New learning environments for the 21st century: exploring the edge. Change 38, 18–24.

Brown, S.D., Lent, R.W., 2006. Preparing adolescents to make career decisions: a social cognitive perspective. In: Pajares, F., Urdan, T. (Eds.), Self-efficacy Beliefs of Adolescents. Information Age Publishing, Greenwich, CT, pp. 201–223.

Caprara, G.V., Barbaranelli, C., Borgogni, L., Steca, P., 2003. Efficacy beliefs as determinants of teachers' job satisfaction. Journal of Educational Psychology 95, 821–832.

Garcia, T., Pintrich, P.R., 1994. Regulating motivation and cognition in the classroom: the role of self-schemas and self-regulatory strategies. In: Schunk, D.H., Zimmerman, B.J. (Eds.), Self-regulation of Learning and Performance: Issues and Educational Applications. Erlbaum, Hillsdale, NJ, pp. 127–153.

Gollwitzer, P.M., Sheeran, P., 2006. Implementation intentions and goal achievement: a meta-analysis of effects and processes. Advances in Experimental Social Psychology 38, 69–119.

Henson, R.K., 2002. From adolescent angst to adulthood: substantive implications and measurement dilemmas in the development of teacher efficacy research. Educational Psychologist 37, 127–150.

Klassen, R.M., 2004. Optimism and realism: a review of self-efficacy from a cross-cultural perspective. International Journal of Psychology 39, 205–230.

Marsh, H.W., Shavelson, R., 1985. Self-concept: its multifaceted, hierarchical structure. Educational Psychologist 20, 107–123.

Multon, K.D., Brown, S.D., Lent, R.W., 1991. Relation of self-efficacy beliefs to academic outcomes: a meta-analytic investigation. Journal of Counseling Psychology 38, 30–38.

Pajares, F., 1996. Self-efficacy beliefs in achievement settings. Review of Educational Research 66, 543–578.

Pajares, F., 2008. Motivational role of self-efficacy beliefs in self-regulated learning. In: Schunk, D.H., Zimmerman, B.J. (Eds.), Motivation and Self-regulated Learning: Theory, Research and Applications. Taylor & Francis, New York, pp. 111–139.

Pajares, F., Kranzler, J., 1995. Self-efficacy beliefs and general mental ability in mathematical problem-solving. Contemporary Educational Psychology 20, 426–443.

Pajares, F., Urdan, T. (Eds.), 2006. Self-efficacy Beliefs of Adolescents. Information Age Publishing, Greenwich, CT.

Relich, J.D., Debus, R.L., Walker, R., 1986. The mediating role of attribution and self-efficacy variables for treatment effects on achievement outcomes. Contemporary Educational Psychology 11, 195–216.

Schunk, D.H., 1981. Modeling and attributional effects on children's achievement: a self-efficacy analysis. Journal of Educational Psychology 73, 93–105.

Schunk, D.H., 1995. Self-efficacy and education and instruction. In: Maddux, J.E. (Ed.), Self-efficacy, Adaptation, and Adjustment: Theory, Research, and Application. Plenum Press, New York, pp. 281–303.

Schunk, D.H., 2012. Social cognitive theory. In: Harris, K.R., Graham, S., Urdan, T. (Eds.), APA Educational Psychology Handbook, Theories, Constructs, and Critical Issues, vol. 1. American Psychological Association, Washington, DC, pp. 101–123.

Schunk, D.H., Ertmer, P.A., 2000. Self-regulation and academic learning: self-efficacy enhancing interventions. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, San Diego, pp. 631–649.

Schunk, D.H., Gunn, T.P., 1986. Self-efficacy and skill development: influence of task strategies and attributions. Journal of Educational Research 79, 238–244.

Schunk, D.H., Pajares, F., 2009. Self-efficacy theory. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 35–53.

Schunk, D.H., Usher, E.L., 2011. Assessing self-efficacy for self-regulated learning. In: Zimmerman, B.J., Schunk, D.H. (Eds.), Handbook of Self-regulation of Learning and Performance. Routledge, New York, pp. 282–297.

Schunk, D.H., Zimmerman, B.J., 1997. Social origins of self-regulatory competence. Educational Psychologist 32, 195–208.

Schunk, D.H., Zimmerman, B.J., 2006. Competence and control beliefs: distinguishing the means and ends. In: Alexander, P.A., Winne, P.H. (Eds.), Handbook of Educational Psychology, second ed. Erlbaum, Mahwah, NJ, pp. 349–367.

Shavelson, R., Bolus, R., 1982. Self-concept: the interplay of theory and methods. Journal of Educational Psychology 74, 3–17.

Smith, R.E., 1989. Effects of coping skills training on generalized self-efficacy and locus of control. Journal of Personality and Social Psychology 56, 228–233.

Tschannen-Moran, M., Woolfolk Hoy, A., Hoy, W.K., 1998. Teacher efficacy: its meaning and measure. Review of Educational Research 68, 202–248.

Valentine, J.C., DuBois, D.L., Cooper, H., 2004. The relation between self-beliefs and academic achievement: a meta-analytic review. Educational Psychologist 39, 111–133.

Weiner, B., 1985. An attributional theory of achievement motivation and emotion. Psychological Review 12, 1–14.

Wigfield, A., Tonks, S., Eccles, J.S., 2004. Expectancy value theory in cross-cultural perspective. In: McInerney, D.M., Van Etten, S. (Eds.), Big Theories Revisited. Information Age Publishing, Greenwich, CT, pp. 165–198.

Woolfolk Hoy, A., Hoy, W.K., Davis, H.A., 2009. Teachers' self-efficacy beliefs. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Routledge, New York, pp. 627–653.

Zimmerman, B.J., 2000. Attaining self-regulation: a social cognitive perspective. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, San Diego, pp. 13–39.

Zimmerman, B.J., 2002. Becoming a self-regulated learner: an overview. Theory into Practice 41, 64–70.

Zimmerman, B.J., Bandura, A., 1994. Impact of self-regulatory influences on writing course achievement. American Educational Research Journal 31, 845–862.

Self-Esteem

Christian H Jordan, Wilfrid Laurier University, Waterloo, ON, Canada
Virgil Zeigler-Hill, Oakland University, Rochester, MI, USA
Jessica J Cameron, University of Manitoba, Winnipeg, MB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-esteem has been studied extensively by psychologists, with thousands of empirical articles exploring its potential causes, consequences, and correlates. Despite this considerable attention, disagreements persist over how best to define and conceptualize self-esteem. Most psychologists view it as a global evaluation of the value of the self or self-worth. Self-esteem is also dimensional with 'low self-esteem' and 'high self-esteem' representing the ends of a continuum. Self-esteem's enduring fascination reflects, in part, its associations with important life outcomes, such as psychological adjustment, academic success, physical health, and relationship satisfaction. The extent to which self-esteem actually causes these outcomes, however, remains controversial.

What Is Self-Esteem?

Self-esteem can be conceptualized at global or domain-specific levels. Global self-esteem is an overall evaluation of self-worth, whereas domain-specific self-evaluations reflect self-appraisals within more circumscribed domains (e.g., academic, social, or appearance self-esteem; Shavelson et al., 1976). Global self-esteem and domain-specific self-evaluations are typically positively correlated, raising questions about how they relate to each other. Early conceptualizations – such as that of philosopher and psychologist William James (1890) – suggest that global self-esteem reflects an averaging of specific self-evaluations, weighted by their subjective importance to the individual (Marsh, 1986). It is now clear, however, that domain-specific self-evaluations are also colored by global self-esteem (Brown et al., 2001) such that individuals with higher self-esteem view their specific attributes more positively in part because of their high self-esteem.

Self-esteem is distinguished from self-concept. Self-esteem is most typically viewed as reflecting self-feelings (e.g., 'I like myself'), rather than self-knowledge. The self-concept rather is viewed as reflecting self-knowledge or self-beliefs (e.g., 'I am outgoing,' 'I am smart'). This distinction is not universally accepted, however, and can become blurry (e.g., people have strong feelings about their self-beliefs, such as being heavily invested in seeing oneself as competent; Marsh and Craven, 2006). The self-concept has clear implications for self-esteem. Individuals who possess more positive self-beliefs tend to report feeling better about themselves overall than those with less positive self-beliefs (e.g., Segal, 1988). There are also, however, important indirect links between the self-concept and self-esteem (Showers and Zeigler-Hill, 2006). Early models of the self often viewed it as a single monolithic entity, but more modern conceptualizations typically consider the self-concept to be multifaceted, consisting of a variety of self-aspects defined by such factors as situations, roles, and psychological states of the individual. This multifaceted view of the self-concept highlights the importance of self-concept structure (i.e., how self-concept content is mentally organized) for moderating the connection between self-concept content and self-esteem. Self-concept structure can vary in a number of ways, such as its complexity or the degree to which positive and negative beliefs about the self are integrated with each other versus being compartmentalized. Self-concept structure determines the accessibility of specific self-beliefs, which affects their ability to influence self-esteem (see Showers and Zeigler-Hill, 2012).

Measuring Self-Esteem

Self-esteem is most commonly measured by self-report scales. Because self-esteem is considered to be a subjective judgment, asking people how they feel about themselves is a sensible and direct way to assess self-esteem. For children, the most commonly used self-report measure is Coopersmith's (1967). Self-Esteem Inventory and for adults it is Rosenberg's (1965) Self-Esteem Scale. More recently, Robins et al. (2001) created a single-item self-esteem scale to assess global self-esteem, which simply assesses agreement with the statement, "I have high self-esteem." This single-item measure demonstrates impressively high correlations with longer, more established measures of self-esteem. Other researchers have developed multidimensional measures of self-esteem. Tafarodi and Swann (2001), for example, devised the Self-Liking and Self-Competence Scale to assess distinct dimensions of global self-esteem in order to provide, in their view, a more comprehensive assessment of self-esteem.

People also possess distinct trait and state self-esteem. Whereas trait self-esteem refers to relatively stable and enduring self-evaluations, state self-esteem refers to momentary self-evaluations that can fluctuate across time and situations. The two most common methods for assessing state self-esteem are Heatherton and Polivy's (1991) State Self-esteem Scale and MacFarland and Ross' (1982) state self-esteem assessment. In both scales, respondents are asked to consider their feelings in the present moment. However, in Heatherton and Polivy's (1991) State Self-esteem Scale, test-takers report their self-evaluations in three domains: appearance, performance, and social acceptance. The items reflect

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25090-3

cognitive evaluations in each of these domains, such as 'I feel unattractive.' In contrast, the state self-esteem measure developed by MacFarland and Ross (1982) asks respondents to indicate how much certain emotional terms describe their present feelings. Some of these emotional terms represent low state self-esteem (e.g., 'shame') whereas others represent high state self-esteem (e.g., 'pride').

Implicit Self-Esteem

Self-esteem has traditionally been studied in terms of explicit self-esteem – fully conscious, deliberative self-evaluations that are measured by self-report scales. This approach to studying self-esteem may be limited, however, because people may be unwilling or unable to report all aspects of their self-evaluations (Greenwald and Banaji, 1995). Accordingly, researchers have developed several nonreactive or implicit measures of self-esteem, including the Name-Letter Task (Koole et al., 2001) and a self-esteem variant of the Implicit Association Test (IAT; Greenwald and Farnham, 2000). The Name-Letter Task is premised on the observation that people value things associated with themselves, and so people tend to like the letters of their name, particularly their initials, more than people whose names do not share those letters. The IAT examines people's reaction times in associating positive and negative stimuli with the self. Some theorists suggest that implicit self-esteem, as assessed by these measures, reflects a construct that is distinct from explicit self-esteem with unique implications for psychological functioning (see Zeigler-Hill and Jordan, 2010). Implicit self-esteem is considered to be relatively automatic, associative, uncontrollable, and intuitive in nature. Implicit self-esteem, in combination with explicit self-esteem, predicts important psychological outcomes, such as defensiveness and aggression. Some researchers have questioned the validity of implicit measures of self-esteem, however (Buhrmester et al., 2010).

The Importance of Self-Esteem

Relation of Self-Esteem to Important Life Outcomes

Interest in self-esteem began to build steadily during the 1970s as research began to emerge suggesting that low self-esteem was linked to a variety of social problems such as drug abuse, unemployment, academic underachievement, and violence. The 'self-esteem movement' was in full swing by the 1980s as exemplified by the California Task Force to Promote Self-esteem and Personal and Social Responsibility (1990), which focused on raising the self-esteem of the citizens of California in the hope that this would reduce the social problems plaguing California at that time. Various self-esteem enhancement programs have been implemented over the years but they have not had the sort of benefits originally envisioned by the self-esteem movement. There continues to be intense debate about the role of self-esteem in important life outcomes, with some researchers arguing that self-esteem is important for productivity and well-being (e.g., Swann et al., 2007) and others arguing that it has severely limited value (e.g., Baumeister et al., 2003).

Associations with Psychopathology and Well-Being

Self-esteem is closely linked with psychopathology. This connection is evident in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; American Psychiatric Association, 2013), which contains various references to self-esteem and terms that are related to self-esteem (e.g., 'grandiose sense of self-importance'). Low self-esteem is included as a diagnostic criterion or associated feature for a variety of disorders including major depressive disorder and bulimia (see Zeigler-Hill, 2011). There are two primary models concerning the connection between low self-esteem and psychopathology. The first model is the 'vulnerability model of low self-esteem,' which suggests that low self-esteem serves as a risk factor for psychopathology. The second model is the 'scar model of low self-esteem,' which argues that low self-esteem is a consequence of psychopathology rather than one of its causes.

Self-esteem is also consistently related to subjective well-being: people with higher self-esteem report greater life satisfaction and happiness than those with lower self-esteem (Baumeister et al., 2003). Moreover, people with lower self-esteem experience more negative moods and fewer positive moods than those with higher self-esteem (MacDonald and Leary, 2012). Some have argued that the greater well-being reported by higher self-esteem individuals might be derived from their tendency to rely on more effective coping mechanisms and their tendency to see the world more optimistically (Baumeister et al., 2003). However, the link between self-esteem and well-being might not be so straightforward. For one thing, the association between self-esteem and well-being seems influenced by cultural variables. For example, the link between these variables is stronger in individualistic cultures (e.g., United States) than in collectivist cultures (e.g., China; Diener and Diener, 1995), suggesting that high self-esteem is not universally tied to greater life satisfaction.

Associations with Physical Health

People with higher self-esteem report greater physical health and an enhanced ability to recover after illnesses compared with those with lower self-esteem (see Stinson et al., 2008). Though it may be tempting to conclude that being ill reduces self-esteem, the causal relationship appears to go in the other direction. There are three main reasons why lower self-esteem might lead to health detriments. First, people with lower self-esteem are less likely to engage in behaviors that promote physical health (e.g., Smits and Kee, 1992), and thus their habits tend to undermine good health. Second, lower self-esteem individuals' physiological experience of stress is apt to erode their physical health. Specifically, lower self-esteem individuals are more likely to experience elevated and prolonged cortisol levels after stress (Pruessner et al., 1999), which might explain why they suffer greater health problems after periods of stress. Third, lower self-esteem might lead to detriments in physical heath because low self-esteem individuals experience poorer quality interpersonal relationships (Stinson et al., 2008). Positive, supportive relationships foster more optimal physical functioning (e.g., lowered cortisol levels). Lacking such relationships may thus predispose lower self-esteem individuals to suffer poorer physical health.

Associations with Academic and Career Success

Is self-esteem associated with success? Most research investigating this question has focused on performance at school and work. In the case of academic performance, people with higher self-esteem report somewhat greater academic success than those with lower self-esteem (Hansford and Hattie, 1982). This partly reflects the fact that high self-esteem predicts the subjective experience of academic achievement (i.e., individuals with higher self-esteem perceive themselves to be better students) but is not strongly related to objective indicators of academic achievement such as grade point average (Baumeister et al., 2003). In addition, mounting evidence suggests that the association between self-esteem and academic performance may be driven by the association of each of these variables with intelligence, family background, and academic achievement in earlier grades (e.g., Maruyama et al., 1981). For example, being more intelligent seems to lead to both greater academic success and higher self-esteem, and thus, the association between self-esteem and academic performance is actually spurious.

Higher self-esteem is also associated with greater success at work. Research suggests, however, that self-esteem is a consequence, not a cause, of performing well at work (Pierce and Gardner, 2004). In their review of this literature, Baumeister et al. (2003) concluded that although self-esteem does not directly lead to better job performance, those with higher self-esteem might still perform somewhat better in the work force. Backing this view are several experimental studies revealing that higher self-esteem individuals exhibit better coping skills that would directly relate to occupational performance. For instance, higher self-esteem individuals are more likely to persist on difficult tasks than lower self-esteem individuals, especially when such persistence is likely to be rewarded. Pierce and Gardner (2004) suggest that organization-based self-esteem (i.e., self-appraisals as a 'good' organizational member) and job performance have a reciprocal relationship: Self-esteem is enhanced by success on the job and this enhanced self-esteem contributes to positive organizational behaviors.

Associations with Crime, Delinquency, and Aggression

Is self-esteem associated with detrimental behavior, such as delinquency, crime, and aggressive behavior? Although much of the research to date has produced conflicting findings, recent longitudinal studies provide a more rigorous test of the causal association between self-esteem and destructive behaviors. This research demonstrates that lower self-esteem is associated with self-reported delinquent behaviors and that these reports are corroborated by students' teachers (Donnellan et al., 2005). Importantly, when these researchers assessed delinquent behavior 2 years later, those with lower self-esteem were again more likely to engage in negative behaviors than their higher self-esteem counterparts. Criminal behavior seems to follow the same pattern: Having lower self-esteem in adolescence predicts a greater propensity to engage in criminal behavior in adulthood (Trzesniewski et al., 2006). For both delinquent and criminal behavior, controlling for other variables known to be associated with these behaviors does not alter the strength of self-esteem's predictive power nor does self-esteem seem to be a consequence of engaging in delinquent or criminal acts.

If lower self-esteem is linked with increased delinquent and criminal behavior, it seems logical to conclude that lower self-esteem should be associated with aggressive behavior. Indeed, an impressive longitudinal study of adolescents suggests that this is the case (Donnellan et al., 2005). However, laboratory experiments have demonstrated a more complex relationship between self-esteem and aggression. Baumeister et al. (1996) conclude that it is actually people with higher self-esteem who will act more aggressively than individuals with lower self-esteem but only when their positive self-views are threatened. Numerous studies both in and outside the lab have confirmed this view: people with higher self-esteem and narcissists who report excessively high self-esteem are more likely to aggress against others than their lower self-esteem counterparts when self-views are threatened.

The Protective Function of Self-Esteem

People with high self-esteem handle negative events better than those with low self-esteem. In particular, individuals with lower self-esteem experience more prolonged distress after negative events such as failure or social rejection (Brown, 2010). In fact, Brown suggested that "high self-esteem functions primarily to enable people to fail without feeling bad about themselves" (Brown, 2010: p. 1389). Self-affirmation theory suggests one reason why this might be the case. It argues that people can maintain their self-esteem in the face of negative outcomes by affirming their positive qualities, even in unrelated domains (e.g., a person rejected romantically may console himself with his occupational success). Because people with high self-esteem perceive themselves positively in more domains than low self-esteem people, they have more psychological resources on which to draw when their self-esteem is threatened (Spencer et al., 1993). Indeed, when confronted by failure, high self-esteem people are more likely to spontaneously focus on their personal strengths and suppress thoughts of their personal weaknesses, compared to low self-esteem individuals (Dodgson and Wood, 1998). Self-esteem may even protect people from the anxiety that can stem from the knowledge of one's own mortality. Humans have the arguably unique capacity to understand that they will die one day. This knowledge creates the potential for intolerable anxiety. Terror Management Theory, however, posits that self-esteem is part of a psychological system that allows people to cope with existential anxiety (Pyszczynski et al., 2004). According to Terror Management Theory, people construct worldviews – culturally specific ways of understanding the world and our place in it – to create a sense of stability and meaning and offer the possibility of symbolic immortality. In this view, high self-esteem results when people believe they are living up to the standards prescribed by their worldviews (i.e., when they view themselves as being 'good' people). Self-esteem thus serves an anxiety-buffering function. Consistent with Terror Management Theory, people with higher self-esteem experience less anxiety, are less prone to deny their vulnerability to early death, and respond less defensively to reminders of death.

The Role of Self-Esteem in Interpersonal Relationships

Self-esteem is rooted in interpersonal and social functioning. Indeed, mounting evidence suggests that self-esteem is best

defined as a 'sociometer' or gauge of social acceptance and relational value (see MacDonald and Leary, 2012). From this perspective, self-esteem is not merely a consequence of interpersonal functioning but drops in self-esteem triggered by rejection serve to alert people to act to repair social relationships. In support of this theory, rejection consistently leads to drops in state self-esteem and poor social relationships lead to reduced trait self-esteem over time. As such, lower self-esteem individuals experience more loneliness, report less social acceptance and support from others and see their ongoing relationships as being of poorer quality relative to those with higher self-esteem.

Do lower self-esteem individuals create their poor relationships? The answer to this question is rather complicated. On the one hand, lower self-esteem individuals may report poorer social ties simply because they tend to see the world more negatively in general (Baumeister et al., 2003). From this perspective, underreporting the quantity and quality of relationships may simply stem from their tendency to see everything in a more negative light. In support of this argument, several studies have reported that self-esteem is unrelated to actual acceptance from others. For instance, Campbell and Fehr (1990) asked participants to engage in a series of first-meetings and report their perceptions of how much their interaction partners liked them and how much they liked their interaction partners. Although lower self-esteem individuals reported lower perceptions of acceptance than higher self-esteem individuals, actual liking was unrelated to self-esteem. Lower self-esteem individuals are also apt to miss indicators of acceptance from dating partners. For example, people with lower self-esteem believe their romantic partners see them more negatively than they actually do and they are less likely to see the support that is provided by a romantic partner after a difficult self-disclosure (see MacDonald and Leary, 2012). Some theorists have argued that it is not simply a matter of lower self-esteem individuals neglecting the positivity coming from their romantic partners. Rather, low self-esteem individuals might actively discount their romantic partners' efforts, assuming that they are insincere attempts to appease them rather than being genuine gestures of caring (Lemay and Clark, 2008).

On the other hand, however, self-esteem may indeed be linked with actual behaviors that affect the quantity and quality of social bonds. For example, the partners of those with lower self-esteem report experiencing less relationship satisfaction and commitment than those partnered with higher self-esteem individuals (Robinson and Cameron, 2012), suggesting an actual behavioral difference between such relationships. Murray et al. (2006) outline how self-esteem differences in social behavior may unfold. According to their risk regulation theory, situations of risk or threat elicit a motivational conflict between the desire to protect the self from possible rejection and the desire to achieve greater closeness with another person. In such situations, lower self-esteem individuals will behave in a self-protective manner whereas higher self-esteem individuals, who have greater resources to cope with disappointment, will behave in a more relationship-promoting manner. Research has supported these claims in both ongoing romantic relationships and initial interactions with a potential romantic interest. For example, self-esteem differences in likable behavior and perceptions of acceptance from an attractive person are only observed when the risk of potential rejection is present and when such risk is removed, self-esteem differences disappear or reverse (Cameron et al., 2010). Thus, when risk is relatively low, self-esteem appears to be unrelated to interpersonal outcomes but when interpersonal risk is high – as may often be the case in romantic relationships – lower self-esteem individuals behave in a manner that impedes or harms their interpersonal bonds. Thus, risk regulation theory not only predicts the actual interpersonal behaviors of those with lower and higher self-esteem but may also explain the conflicting findings in this literature.

The Pursuit of Self-Esteem

People generally appear to be strongly motivated to achieve or maintain positive self-views. They display a variety of cognitive biases and behavioral tendencies – such as attributing successes to ability but failures to bad luck – that seem calibrated to promote self-esteem (Campbell and Sedikides, 1999). People may be fundamentally motivated to see themselves as essentially moral and competent individuals, and a number of deleterious human tendencies may serve this goal, including closed-mindedness, rationalizing decisions, and prejudice (Sherman and Cohen, 2006). Indeed, if people have an overall positive sense of self-esteem affirmed (e.g., by reflecting on an important personal value), they display fewer of these self-serving tendencies; they are, for example, more open-minded and less prone to rationalization. Perhaps as a consequence of pervasive self-serving biases, most people have 'positive illusions' or somewhat unrealistic positive self-views (Taylor and Brown, 1988).

Is the pursuit of high self-esteem adaptive? Positive illusions may foster greater psychological well-being (Taylor and Brown, 1988), but this point remains controversial. Indeed, some theorists suggest that pursuing self-esteem is costly (Crocker and Park, 2004). Self-determination theory posits that people have fundamental needs for competence, relatedness, and autonomy that must be fulfilled in order for people to thrive (Deci and Ryan, 1995). From this perspective, the pursuit of self-esteem (which is seen as a common motive but not a fundamental need) can have negative psychological consequences because it interferes with the fulfillment of basic needs. Consistent with this possibility, students whose self-esteem depends on their academic achievement (i.e., whose feelings of self-worth plummet when they perform poorly in school) are more interested in doing well in exams than in mastering the course material (Crocker et al., 2003). As a consequence, they may have less intrinsic interest in learning.

Fragile Self-Esteem

Actively pursuing self-esteem might thus have negative consequences. In addition, not all high self-esteem is equally beneficial. There is increasing recognition that high self-esteem is heterogeneous, that individuals who report having high self-esteem are not all equivalent psychologically (Kernis, 2003).

An influential distinction has been that between secure and fragile forms of high self-esteem (Jordan and Zeigler-Hill, 2013). Secure high self-esteem is viewed as being well-anchored and stable. Individuals with secure high self-esteem acknowledge their shortcomings yet accept themselves nonetheless. Fragile high self-esteem, in contrast, is viewed as inflated, vulnerable to threat, and in constant need of validation. Individuals with fragile high self-esteem are preoccupied with affirming their tenuous positive self-views and are prone to be defensive and aggressive. To distinguish between secure and fragile self-esteem, researchers have relied on three primary indicators: whether high self-esteem is unstable, contingent, or paired with low implicit self-esteem (Kernis and Paradise, 2002). High self-esteem that is unstable fluctuates markedly across time and contexts. There is a great deal of evidence that people with unstable high self-esteem are more defensive and aggressive than those with stable high self-esteem. Contingent self-esteem depends on meeting specific standards or expectations in self-relevant domains and available evidence (though limited) links it to defensiveness. Lastly, among individuals who report having high self-esteem (i.e., have high explicit self-esteem), those with low implicit self-esteem are more defensive and self-enhancing (see Jordan and Zeigler-Hill, 2013).

Where Does Self-Esteem Come From?

Etiology

Given the importance of early childhood experiences for shaping how individuals view themselves, it seems quite likely that these early experiences will be associated with self-esteem. Consistent with this possibility, DeHart et al. (2006) found that the self-esteem levels of individuals in early adulthood were associated with their memories of how their parents interacted with them as children, an association corroborated by the parents' recollections. Beyond parenting styles, it is also possible that the family environment may influence self-esteem through the genetic contributions of both parents (Neiss et al., 2002). Self-esteem appears to be connected to temperament such that those who are predisposed to experience negative affect or to be more reactive to stressful life events may be more likely to develop lower self-esteem if they are exposed to aversive social environments during early childhood. The foundation of self-esteem appears to be established during childhood but feelings of self-worth can certainly change throughout the life span as individuals experience positive and negative life events (e.g., acceptance and rejection).

Developmental Changes

Individuals with relatively high levels of self-esteem early in their lives tend to report higher levels of self-esteem than their peers throughout their lives (Trzesniewski et al., 2013). Despite this consistency across individuals, average levels of self-esteem have been observed to change systematically across the life span such that feelings of self-worth are relatively high during childhood before dropping precipitously at the beginning of adolescence (Robins et al., 2002). Self-esteem follows a quadratic (inverted-U) trajectory from this point such that it increases throughout adolescence, young adulthood, and middle adulthood before reaching its peak around 50–60 years of age and then declining in old age (e.g., Orth et al., 2012).

Origins of Self-Esteem

One might ask why people have self-esteem at all. What are its potential evolutionary origins? The sociometer model posits that self-esteem evolved as a gauge or index of social inclusion such that low self-esteem serves as a signal that one is at risk of exclusion and must take action to maintain social acceptance (Leary and Downs, 1995). Consistent with the sociometer model, feelings of self-esteem are closely related to feelings of social inclusion such that state self-esteem decreases when people are socially excluded (Leary et al., 1995). Kirkpatrick and Ellis (2001) extended the sociometer model by arguing that self-esteem actually represents the operation of a suite of psychological mechanisms (or sociometers) that evolved to meet the distinct adaptive challenges associated with mate selection, establishing and maintaining status and rank, and the formation of alliances. Within this model, humans evolved different, functionally specific self-evaluations to monitor their mate value, social dominance, prestige, inclusion, and collective worth.

Status-tracking models of self-esteem – such as the sociometer model – have focused on the influence that perceived standing has on the self-esteem of the individual without addressing the possibility that self-esteem also influences how others perceive the individual. That is, self-esteem may serve as an interpersonal signal (see Cameron et al., 2013, for an extended review of this issue). The status-signaling model of self-esteem (Zeigler-Hill, 2012) addresses the signaling property of self-esteem by suggesting that the feelings of self-worth exhibited by individuals are likely to influence how they are perceived by others in their social environments. This basic effect has been observed in a variety of studies showing that manipulating the ostensible self-esteem level of a target influences how he or she is evaluated on dimensions such as romantic desirability (Zeigler-Hill and Myers, 2011).

Improving Self-Esteem

Self-esteem is considered to be an integral component of various forms of psychotherapy including humanistic therapy, motivational enhancement therapy, and cognitive-behavioral therapy (Mruk and O'Brien, 2013). For example, cognitive-behavioral therapy influences feelings of self-worth by focusing on the recognition and correction of cognitive errors such as all-or-none thinking, overgeneralization, and selective attention to negative life experiences. A large number of laboratory manipulations have been developed to temporarily alter self-esteem levels, with some of these techniques using classical conditioning (Baccus et al., 2004) or asking individuals to attend to positive stimuli (i.e., smiling faces) while ignoring negative stimuli (i.e., frowning faces; Dandeneau and Baldwin, 2004). Although an array of self-esteem enhancement programs have been implemented in schools, the utility of these programs remains uncertain.

See also: Aggression, Social Psychology of; Cultural Intelligence and Competencies; Culture and the Self: Implications for Psychological Theory; Dark Triad: The "Dark Side" of Human Personality; Depression; Hope and Hopelessness; Identity and Identification, Social Psychology of; Implicit Association Test; Implicit Social Cognition; Mastery Learning; Neuroticism; Optimism and Pessimism; Personality Disorders; Personality and Adaptive Behaviors; Personality and Self; Personality and Values at Work; Personality: Historical and Conceptual Perspectives; Self and Identity Development During Adolescence across Cultures; Self-Concept: From Unidimensional to Multidimensional and Beyond; Self-Determination Theory; Self-Efficacy; Selfconscious Emotions, Psychology of; Social Comparison, Psychology of.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Baccus, J.R., Baldwin, M.W., Packer, D.J., 2004. Increasing implicit self-esteem through classical conditioning. Psychological Science 15, 498–502.

Baumeister, R.F., Campbell, J.D., Krueger, J.I., Vohs, K.D., 2003. Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? Psychological Science in the Public Interest 4, 1–44.

Baumeister, R.F., Smart, L., Boden, J.M., 1996. Relation of threatened egotism to violence and aggression: the dark side of high self-esteem. Psychological Review 103, 5–33.

Buhrmester, M.D., Blanton, H., Swann, W.B., 2010. Implicit self-esteem: nature, measurement, and a new way forward. Journal of Personality and Social Psychology 100, 365–385.

Brown, J.D., 2010. High self-esteem buffers negative feedback: once more with feeling. Cognitive and Emotion 24, 1389–1404.

Brown, J.D., Dutton, K.A., Cook, K.E., 2001. From the top down: self-esteem and self-evaluation. Cognition and Emotion 15, 615–631.

California Task Force to Promote Self-esteem and Personal and Social Responsibility, 1990. Toward a State of Self-esteem. California State Department of Education, Sacramento, CA.

Cameron, J.J., MacGregor, J., Kwang, T., 2013. Badge of honor or mark of shame: self-esteem as an interpersonal signal. In: Zeigler-Hill, V. (Ed.), Self-esteem. Psychology Press, London, pp. 145–162.

Cameron, J.J., Stinson, D.A., Gaetz, R., Balchen, S., 2010. Acceptance is in the eye of the beholder: self-esteem and motivated perceptions of acceptance from the opposite sex. Journal of Personality and Social Psychology 99, 513–529.

Campbell, J.D., Fehr, B.A., 1990. Self-esteem and perceptions of conveyed impressions: is negative affectivity associated with greater realism? Journal of Personality and Social Psychology 58, 122–133.

Campbell, W.K., Sedikides, C., 1999. Self-threat magnifies the self-serving bias: a meta-analytic integration. Review of General Psychology 3, 23–43.

Coopersmith, S., 1967. The Antecedents of Self-esteem. Freeman, San Francisco.

Crocker, J., Luhtanen, R.K., Cooper, M.L., Bouvrette, S.A., 2003. Contingencies of self-worth in college students: theory and measurement. Journal of Personality and Social Psychology 85, 894–908.

Crocker, J., Park, L.E., 2004. The costly pursuit of self-esteem. Psychological Bulletin 130, 392–414.

Dandeneau, S.D., Baldwin, M.W., 2004. The inhibition of socially rejecting information among people with high versus low self-esteem: the role of attentional bias and the effects of bias reduction training. Journal of Social and Clinical Psychology 23, 584–602.

Deci, E.L., Ryan, R.M., 1995. Human autonomy: the basis for true self-esteem. In: Kernis, M.H. (Ed.), Efficacy, Agency, and Self-esteem. Plenum Press, New York, pp. 31–49.

DeHart, T., Pelham, B.W., Tennen, H., 2006. What lies beneath: parenting style and implicit self-esteem. Journal of Experimental Social Psychology 42, 1–17.

Diener, E., Diener, M., 1995. Cross-cultural correlates of life satisfaction and self-esteem. Journal of Personality and Social Psychology 68, 653–663.

Dodgson, P.G., Wood, J.V., 1998. Self-esteem and the cognitive accessibility of strengths and weaknesses after failure. Journal of Personality and Social Psychology 75, 178–197.

Donnellan, M.B., Trzesniewski, K.H., Robins, R.W., Moffitt, T.E., Caspi, A., 2005. Low self-esteem is related to aggression, antisocial behavior, and delinquency. Psychological Science 16, 328–335.

Greenwald, A.G., Banaji, M.R., 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychological Review 102, 4–27.

Greenwald, A.G., Farnham, S.D., 2000. Using the implicit association test to measure self-esteem and self-concept. Journal of Personality and Social Psychology 79, 1022–1038.

Hansford, B.C., Hattie, J.A., 1982. The relationship between self and achievement/performance measures. Review of Educational Research 52, 123–142.

Heatherton, T., Polivy, J., 1991. Development and validation of a scale for measuring state self-esteem. Journal of Personality and Social Psychology 60, 895–910.

James, W., 1890. The Principles of Psychology. Henry Holt, New York.

Jordan, C.H., Zeigler-Hill, V., 2013. Fragile self-esteem: the perils and pitfalls of (some) high self-esteem. In: Zeigler-Hill, V. (Ed.), Self-esteem: Current Issues in Social Psychology. Psychology Press, New York, pp. 80–98.

Kernis, M.H., 2003. Toward a conceptualization of optimal self-esteem. Psychological Inquiry 14, 1–26.

Kernis, M.H., Paradise, A.W., 2002. Distinguishing between secure and fragile forms of high self-esteem. In: Deci, E.L., Ryan, R.M. (Eds.), Handbook of Self-Determination Research. University of Rochester Press, Rochester, NY, pp. 339–360.

Kirkpatrick, L.A., Ellis, B.J., 2001. An evolutionary-psychological approach to self-esteem: multiple domains and multiple functions. In: Clark, M., Fletcher, G. (Eds.), The Blackwell Handbook of Social Psychology, Interpersonal Processes, vol. 2. Blackwell Publishers, Oxford, pp. 411–436.

Koole, S.L., Dijksterhuis, A., van Knippenberg, A., 2001. What's in a name: implicit self-esteem and the automatic self. Journal of Personality and Social Psychology 80, 669–685.

Leary, M.R., Downs, D.L., 1995. Interpersonal functions of the self-esteem motive: the self-esteem system as a sociometer. In: Kernis, M.H. (Ed.), Efficacy, Agency, and Self-esteem. Plenum, New York, pp. 123–144.

Lemay, E.P., Clark, M.S., 2008. "Walking on eggshells": how expressing relationship insecurities perpetuates them. Journal of Personality and Social Psychology 95, 420–441.

Leary, M.R., Tambor, E.S., Terdal, S.K., Downs, D.L., 1995. Self-esteem as an interpersonal monitor: the sociometer hypothesis. Journal of Personality and Social Psychology 68, 518–530.

MacDonald, G., Leary, M.R., 2012. Individual differences in self-esteem. In: Leary, M.R., Tangney, J.P. (Eds.), Self and Identity, second ed. Guilford Press, New York, pp. 354–377.

MacFarland, C., Ross, M., 1982. Impact of causal attributions on affective reactions to success and failures. Journal of Personality and Social Psychology 43, 937–946.

Marsh, H.W., 1986. Global self-esteem: its relation to specific facets of self-concept and their importance. Journal of Personality and Social Psychology 51, 1224–1236.

Marsh, H.W., Craven, R.G., 2006. Reciprocal effects of self-concept and performance from a multidimensional perspective: beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science 35, 63–78.

Maruyama, G., Rubin, R.A., Kingsbury, G.G., 1981. Self-esteem and educational achievement: independent constructs with a common cause? Journal of Personality and Social Psychology 40, 962–975.

Mecca, A.M., Smelser, N.J., Vasconcellos, J., 1989. The Social Importance of Self-esteem. University of California Press, Berkeley, CA.

Mruk, C.J., O'Brien, E.J., 2013. Changing self-esteem through competence and worthiness training: a positive therapy. In: Zeigler-Hill, V. (Ed.), Self-esteem. Psychology Press, London, pp. 163–179.

Murray, S.L., Holmes, J.G., Collins, N.L., 2006. Optimizing assurance: the risk regulation system in relationships. Psychological Bulletin 132, 640–666.

Neiss, M.B., Sedikides, C., Stevenson, J., 2002. Self-esteem: a behavioral genetic perspective. European Journal of Personality 16, 351–367.

Orth, U., Robins, R.W., Widaman, K.F., 2012. Life-span development of self-esteem and its effects on important life outcomes. Journal of Personality and Social Psychology 102, 1271–1288.

Pierce, I.L., Gardner, D.G., 2004. Self-esteem within the work and organizational context: a review of the organization-based self-esteem literature. Journal of Management 30, 591–622.

Pruessner, J.C., Hellhammer, D.H., Kirschbaum, C., 1999. Low self-esteem, induced failure and the adrenocortical stress response. Personality and Individual Differences 27, 477–489.

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., Schimel, J., 2004. Why do people need self-esteem? A theoretical and empirical review. Psychological Bulletin 130, 435–468.

Robins, R.W., Hendin, H.M., Trzesniewski, K.H., 2001. Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg self-esteem scale. Personality and Social Psychology Bulletin 27, 151–161.

Robins, R.W., Trzesniewski, K.H., Tracy, J.L., Gosling, S.D., Potter, J., 2002. Global self-esteem across the life span. Psychology and Aging 17, 423–434.

Robinson, K.J., Cameron, J.J., 2012. Self-esteem is a shared relationship resource: additive effects of dating partners' self-esteem levels predict relationship quality. Journal of Research in Personality 46, 227–230.

Rosenberg, M., 1965. Society and the Adolescent Self-image. Princeton University Press, Princeton, NJ.

Segal, Z.V., 1988. Appraisal of the self-schema construct in cognitive models of depression. Psychological Bulletin 103, 147–162.

Shavelson, R.J., Hubner, J.J., Stanton, G.C., 1976. Validation of construct interpretations. Review of Educational Research 46, 407–441.

Sherman, D.K., Cohen, G.L., 2006. The psychology of self-defense: self-affirmation theory. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 38. Academic Press, San Diego, CA, pp. 183–242.

Showers, C.J., Zeigler-Hill, V., 2006. Pathways among self-knowledge and self-esteem: how are self-esteem and self-knowledge linked? Are these links direct or indirect? In: Kernis, M.H. (Ed.), Self-esteem Issues and Answers: A Source Book of Current Perspectives. Psychology Press, New York, pp. 216–223.

Showers, C.J., Zeigler-Hill, V., 2012. Organization of self-knowledge: features, functions, and flexibility. In: Leary, M.L., Tangney, J. (Eds.), Handbook of Self and Identity, second ed. Guilford, New York, pp. 105–123.

Smits, M.W., Kee, C.C., 1992. Correlates of self-care among the independent elderly: self-concept affects well-being. Journal of Gerontological Nursing 18, 13–18.

Spencer, S.J., Josephs, R.A., Steele, C.M., 1993. Low self-esteem: the uphill struggle for self-integrity. In: Baumeister, R.F. (Ed.), Self-esteem: The Puzzle of Low Self-regard. Plennum, New York, pp. 21–36.

Stinson, D.B., Logel, C., Zanna, M.P., Holmes, J.G., Cameron, J.J., Wood, J.V., Spencer, S.J., 2008. The cost of lower self-esteem: testing a self-and-social-bonds model of health. Journal of Personality and Social Psychology 94, 412–428.

Swann Jr., W.B., Chang-Schneider, C., McClarty, K.L., 2007. Do people's self-views matter? Self-concept and self-esteem in everyday life. American Psychologist 62, 84–94.

Tafarodi, R.W., Swann Jr., W.B., 2001. Two-dimensional self-esteem: theory and measurement. Personality and Individual Differences 31, 653–673.

Taylor, S.E., Brown, J.D., 1988. Illusion and well-being: a social psychological perspective on mental health. Psychological Bulletin 103, 193–210.

Trzesniewski, K.H., Donnellan, M.B., Moffitt, T.E., Robins, R.W., Poulton, R., Caspi, A., 2006. Low self-esteem during adolescence predicts poor health, criminal behavior, and limited economic prospects during adulthood. Developmental Psychology 42, 381–390.

Trzesniewski, K.H., Donnellan, M.B., Robins, R.W., 2013. Development of self-esteem. In: Zeigler-Hill, V. (Ed.), Self-esteem. Psychology Press, London, pp. 60–79.

Zeigler-Hill, V., 2011. The connections between self-esteem and psychopathology. Journal of Contemporary Psychotherapy 41, 157–164.

Zeigler-Hill, V., 2012. The extended informational model of self-esteem. In: De Wals, S., Meszaros, K. (Eds.), Handbook on Psychology of Self-esteem. Nova, Hauppauge, NY, pp. 211–226.

Zeigler-Hill, V., Jordan, C.H., 2010. Two faces of self-esteem: implicit and explicit forms of self-esteem. In: Gawronski, B., Payne, B.K. (Eds.), Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. Guilford Press, New York, pp. 392–407.

Zeigler-Hill, V., Myers, E.M., 2011. An implicit theory of self-esteem: the consequences of perceived self-esteem for romantic desirability. Evolutionary Psychology 9, 147–180.

Self-Organizing Dynamical Systems

Nigel Gilbert, David Anzola, Peter Johnson, Corinna Elsenbroich, Tina Balke, and Ozge Dilaver, Centre for Research in Social Simulation, University of Surrey, Guildford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.A.S. Kelso, volume 20, pp. 13844–13850, © 2001, Elsevier Ltd.

Abstract

The concept of self-organization in social science is reviewed. In the first two sections, some basic features of self-organizing dynamical systems in general science are presented and the origin of the concept is reconstructed, paying special attention to social science accounts of self-organization. Then, theoretical and methodological considerations regarding the current application of the concept and prospective challenges are examined.

Self-organizing dynamical systems (SODS) is a concept widely used in the physical and life sciences, with promising applications in social science. Because there is an extensive literature on the topic elsewhere, the focus here is on the prospective use of the concept in social science, understood in a broad sense to be those disciplines that study any manifestation of human behavior.

Basic Features

Several definitions of self-organization can be found in the literature. Although a consensual transdisciplinary definition is unlikely, and maybe also undesirable, some features are shared by most accounts of the concept.

Pattern Formation

The basic idea of self-organization relates to pattern formation. Identification of patterns, however, is not a straightforward matter. Observational criteria such as the geometric distribution of cell states in a fractal serve well as a starting point for simple systems. For more complicated systems like societies, robust criteria regarding organization are required. Capturing that meaning with a definition that does not then require further qualification is very difficult. Self-organization has been framed as an increase in coordination, a reduction in entropy, a movement from simple to complex structures, or a decrease in degrees of freedom. It is not yet clear to what extent these criteria can be translated from their original domains to social science and if they can provide any insightful information about traditionally studied social processes, such as normalization, stratification, specialization, etc.

Whether or not a change in self-organization can be quantified is another key question. The more clearly the level of organization in a system is quantified, the more objective the organizational criteria devised become. In social science, however, full quantification is considered difficult or unlikely. Measurement in many cases relies on qualitative or stylized perceptions of the system. The observer's subjective point of view and interpretation is important. Comparative studies have shown, for example, that traditional societies have subsystems with high levels of organization and complexity, and, yet, because of the past observers' biases and prejudices, for a long time these societies were understood to be simple and wrongly labeled 'primitive.'

Autonomy

Autonomy covers the 'self' in self-organization. Any change in a self-organized system must come from inside the system, and must be caused by the many individual entities, or interactions between these entities, rather than from a single controlling or external force. This does not require that the system be closed, with no input from outside, but rather that any input must not be a controlling force. Social systems pose two particular challenges for this feature of self-organizing systems. First, there are asymmetric power relations between and within levels; second, the boundaries of social systems and subsystems are difficult to draw. Regarding the former, humans are intentional actors and can have designs on controlling or influencing a system through individual or institutional action, for example, by the implementation of public policy. Regarding the latter, setting the system and subsystem boundaries is crucial. The outside and inside of the system need to be clearly identifiable in order to make sure any controlling force is not ignored via undue manipulation of the boundary of the system. However, social science often ignores or is vague about boundaries, both for theoretical and methodological reasons. For example, it is common to model countries' economies as closed systems, ignoring all the external influence, because it is easier methodologically. Furthermore, it can be difficult to draw a boundary around even simple social subsystems. The rules that define the membership of social groups or who exactly is to be included in a 'family' can be vague and contested.

Robustness and Resilience

Robustness and resilience are features usually ascribed to SODS. Although these two terms are often used interchangeably, robustness refers to a system's ability to resist change, whereas resilience refers to a system's ability to adapt or recover from change. Both concepts could be used for the same process, depending on the observer's perspective, meaning it is not always clear whether a social system is robust or resilient. For

example, an economy may experience an external shock in the form of a natural disaster, but no major changes in its structure may be observed. It will be unclear whether this is because it resisted the shock or adapted to it. It is difficult to devise objective criteria to quantify the difference between the two.

Dynamics

Dynamical systems are those systems whose main features or properties are not time invariant. SODS change over time and these changes are given by the nonstatic character of the constitutive elements and relations. Change in SODS is usually the result of path-dependent processes, meaning that the system's future states are highly dependent on past states. This is why understanding the dynamics is important. However, the exact role that the concept of dynamics plays in the study of SODS is dependent on the conceptualization of self-organization. Some portray self-organization as leading to an equilibrium or attractor state, whereas others highlight the processes and dynamics of a system. Take the case of cooperation. The focus might be on the process of interaction that allows for the emergence of a structure of cooperative behavior in the first place or on the changes produced by this structure in the long run, for example, the propagation of cooperative behavior. In the first approach, the system dynamics lead to and maintain a state of self-organization, whereas in the second case, the organization generates the system's dynamics. Although not philosophically incompatible, the two views focus on different aspects of the dynamic of self-organization.

SODS can include both positive and negative feedbacks loops. The former creates a stabilizing effect that maintains the current system properties, whereas the latter destabilizes systems by creating self-reinforcing processes. Feedback may have thresholds around which the behavior of the system will be significantly different. The effects on organization may not be gradual, but occur in a modular, step-wise, or hierarchical fashion.

Historical Development

Ideas similar to the concept of self-organization can be traced back to many important figures in Western thought. Yet, after the first explicit reference in Kant's *Critique of Judgment*, the concept did not play a major role in science until it was picked up by Ashby (1947) in the context of cybernetics. He used the concept in his discussion on whether it was possible to produce self-induced change in the organization of a machine.

The contemporary use, however, is the result of developments in analytical tools, technology, and theory. Analytically, it required the development of a nonlinear mathematical framework, for it avoids the stringent formalization assumptions of traditional scientific methods (Strogatz, 1994). The application of nonlinear mathematics to the study of dynamical systems has benefited greatly from an increase in computational capabilities. Technological developments in hardware and software have allowed researchers to process large amounts of data and also to find numerical solutions for nonlinear equations. Important developments in information theory and processing have also played a significant role, as they provided insights about the context, conditions, prospects, and limitations of information processing in artificial and natural systems. The idea of cellular automata, for example, has played a major role in the understanding of self-organizing dynamics (e.g., Wolfram, 1994; Mandelbrot, 1982).

The key theoretical development was the articulation of a comprehensive theory of systems, building on important efforts such as von Bertalanffy's (1969) work on general system theory. The concept of self-organization is currently studied as part of the wider framework of complexity science (Bar-Yam, 1997). Social science, as a general area of inquiry, has not contributed much to complexity science nor, with some exceptions, has it benefited much from these developments.

There have been cases of straightforward translation of the complexity framework to social science, especially in disciplines with a preference for formal methods, such as economics. In addition, some areas have used ideas and methods from complexity science, adding some social theory into the mix, to form a subdiscipline that has become known as computational social science (Gilbert, 2010). These efforts, however, have not led to a general growth in interest in these topics or to a proper cross-fertilization with the mainstream.

Some of the concepts associated with complexity and self-organization have also been used independently in social science. For example, Luhmann's (1995) system theory uses concepts such as self-reference and autopoiesis, but frames them in a particular understanding of symbolic communication that is not compatible with all aspects of the current complexity framework. In other cases, the connection has been based on discursive rather than scientific criteria. For example, connections established between deconstruction and chaos theories were not grounded on methodological similarity, but on the way these approaches were said to challenge the scientific status quo (Cilliers, 1998).

Despite this lack of connection with the contemporary usage in complexity science, it could be argued that social science has approached the problem of self-organization in its own way. The notion of self-organization has been applied both to the case of separated entities that become joined after certain dynamics and to the case of the rearrangement of joint entities. The question about the formation of the State that has puzzled social researchers since the first contractualists in the seventeenth century is an inquiry about self-organization because the 'contract' has mostly played a metaphorical role. Likewise, questions about 'good' organization have been a matter of interest for social scientists since the emergence of the disciplines. In economics, Adam Smith's 'invisible hand' refers to the manner in which the economy organizes itself to maximize social welfare, although each economic agent is only trying to increase his or her own welfare. This notion of self-organizing optimal distribution has dominated the development of the discipline and has had huge importance in political discourse. Likewise, the first sociologists focused their theorizing on how societies could or were coping with the effects of the demographical changes that followed the transition to modernity. Weber's 'iron cage,' for example, was understood as providing 'good' organizational principles regarding power and authority in modern societies.

Not all social theories are articulated in the same terms or grounded in the same philosophical foundations. Some of

them are clearly in opposition to the idea of self-organization. Yet, examples of self-organizing dynamics can be found in almost every social setting. Sometimes, the reference is hard to identify because it is surrounded by narrative and metaphors, as with the case of the 'social contract.' At other times, the idea has unnecessary or conflicting philosophical connections. Nonetheless, for most of these accounts, the links with complexity theory can be traced and a reformulation in terms of a complexity theory framework is possible.

Theoretical Considerations

When dealing with self-organizing social systems, one needs to avoid teleological explanations of social processes, that is, explaining social processes in terms of societal goals or final causes. The most well-know teleological approach in social science is functionalism. This approach started as a biological analogy, with social changes studied in terms of function and dysfunction. The teleological character was given the assumption of a functional end state, such as social integration. A key criticism, stressed by conflict theorists, is that there is no reason why integration should be a natural result of the organizational process. It is questionable whether any final aim or goal of a social system can be identified. Although explicit teleological formulations are uncommon now, some forms of teleology still persist implicitly in more formal reasoning, such as causal explanation.

Removing teleology is just the first step. The 'self' in self-organization when studying social phenomena requires more consideration. Societies have several levels of organization. They have cultures, fashions, markets, legal, moral, and social norms and spatial manifestations (towns, cities, farmland, etc.). If self-organization means only that a system is not externally organized, societies can easily be considered self-organizing. However, if self-organization is defined as a system where order or coordination increases without an agent either inside or outside the system producing this order, some caveats arise. Some parts of social systems come about or are shaped by the purposeful or intentional behavior of individuals or institutions within the system. For example, markets are usually understood as self-organizing. However, most markets are regulated (and thus organized) by externally imposed legal rules. Likewise, an institutionally explicitly sanctioned Highway Code affects the organization of highway traffic. In both examples, the autonomy of the subsystem can be questioned.

These issues can be resolved in different ways. One option is to dismiss purposeful or intentional behavior by deeming it unimportant for the overall behavior of the system. By following this approach, social science could benefit from an almost straightforward application of the framework for the analysis of SODS developed in disciplines outside the social realm. This is the path taken by disciplinary areas such as sociophysics (Galam, 2012) and econophysics (Mantegna and Stanley, 2000). Another option is to consider agents' intentions as having local effects only. This is the approach taken by social theories highlighting the effect of unintended consequences of individual purposive action. A third option is to consider individual intentions as socially determined. This view is commonly associated with macro or structural approaches in social science.

The underlying differences between these options, and many others that could be found in between, stem from the different ways that social scientists have approached core philosophical dualisms, such as micro–macro, agency–structure, and subjectivism–objectivism. Although philosophical consensus around these issues is unlikely, operationalization can help advance the discussion in significant ways. Where all these theories converge, despite the differences in philosophical foundations, is in the identification of 'action' as social science's primary epistemological building block. The operationalization of action is important for the study of SODS. A bottom-up approach is distinctive in the contemporary study of complex systems. It is the interaction between a system's components that gives rise to self-organizing dynamics. Human actors are the key component of social systems and individual action is the key element of social interaction.

Methodological Considerations

Questions about operationalization are solved in science through modeling. Models play a major role in science because they make phenomena accessible for scientific treatment. A scientific model is a simplified representation of particular features that makes a phenomenon more easily understandable and measurable. Most social phenomena have a scale and level of complexity that makes them difficult to understand by direct observation and experimentation, and models can be useful for the same reasons as in science. Moreover, in social science, there are technical or moral reasons that often prevent social scientists from experimenting directly on the social world, while experimenting with the corresponding models is possible.

In what regards to human action, scientific models of self-organization deal mostly with the subject of intentionality. Human beings perform actions based on particular beliefs, attitudes, desires, and purposes. The cognitive foundations, decision heuristics, and contexts for all these can be quite different. Yet, in order to be formalized and operationalized, intentionality has to be simplified and detached from its symbolic and material needs and implications. The intention or meaning attached to human action is articulated in terms of reflexive capabilities. Individual action is formalized as a reactivity–reflexivity pair. Humans share reactivity with other basic entities of complex systems. This is understood as the basic capacity to respond to external stimuli. Reflexivity is where social methods usually differ from other disciplines. In models of social phenomena, reflexivity is associated with the articulation of decision heuristics. The complexity of these decision heuristics and the general role of reflexivity in the model depend on the level of detail and realism of the cognitive structure attached to it.

Simple atom-like models, such as those typical of econophysics or sociophysics, have no cognitive structure. Behavior in this case is just reactive. In game theory, when strategies such as tit for tat are introduced, actors have a more complex cognitive structure due to the addition of a capacity for memory (e.g., to memorize previous actions). Alternatively, agents can

have very complex symbolic communication including a proper language. At this latter level of development of the cognitive structure, agents can reflect on some of the most complex principles of interaction, such as double contingency. The complexity of the cognitive structure can also be modeled through the nonsymbolic paradigm. For example, in neural networks, complex information processing can transform syntactical structures according to different semantic purposes.

Simulation Models of SODS

Conventional quantitative statistical methods have difficulty in accounting for the dynamical aspects of social phenomena. They have relied mostly on cross-sectional (i.e., single moment in time) analyses, with powerful longitudinal methods developed only during the late twentieth century. Accounting for the processual character of a phenomenon is sometimes not possible because there are not sufficient longitudinal data, or data sets are not compatible, too complex, or unreliable. Some of these difficulties can be avoided by simulation because it allows modeling one process by another process. Simulation has played an important role in the study of SODS because of its focus on the dynamics of the system. The two styles of simulation that have most often been applied to SODS are system dynamics and agent-based models.

System dynamics models consist of sets of equations that can be solved analytically or numerically, depending on the complexity of the phenomenon analyzed. In most cases, the self-organizing character of the phenomenon is determined by a dynamic equilibrium, influenced by the initial conditions and the effect of feedback loops. For example, Martinez Coll's (1986) system dynamics model is a recreation of Hobbes' seminal problem of order. Hobbes' original formulation refers to the way individuals compete for resources with each other in a human group with no rules. The model consists of a set of equations that determine individual strategies regarding the possession and sharing of goods. In the simulation, successful strategies can be inherited, imitated, or learned from others. Self-organization occurs in the model as a result of these adaptive interactions. The results of the model show a convergence of agent strategies, varying according to the strategies' distribution at the beginning of the simulation.

Agent-based modeling uses a different methodological principle. In these models each agent in a model is a self-contained program that can take autonomous decisions according to its perception of the environment. It is considered a promising tool for the exploration of self-organizing dynamics in the social world because it allows for the direct coding of heterogeneous agents that can be set to interact with each other and with an artificial environment as the simulation progresses. System self-organizing properties are generated in agent-based models from the interactions of these autonomous agents without central control. As agents are influenced by their perception of the environment, system properties can in turn influence decisions, potentially producing feedback loops.

Both simulation techniques have their drawbacks. Social science has struggled historically to delineate a theory of properties and measurement. It is not clear whether social variables, such as 'being poor,' have the same character and status as physical variables, such as 'having weight.' Even if they can be considered as variables in the same way as physical variables, it is unclear how to measure them (Michell, 2007). The need to define formally what is being measured and the relations between the measurements imposes limitations on what can be modeled. Equation-based models face challenges in this regard due to the restrictions of mathematical formal language. By relying on computer code, more flexible syntactically and semantically than mathematical language, agent-based modeling bypasses some of the formalization limitations, but faces additional problems regarding the adequacy and robustness of its results, for comprehensive verification and validation techniques are still being developed.

Examples of SODS

There is a plethora of examples of models of self-organizing systems in the social sciences at differing levels of complexity. We describe two examples that conceptualize their social processes as a self-organizing dynamics.

The first example is a simulation model of residential segregation developed in Schelling (1971). Agents of two types are distributed randomly on a checkerboard. An agent can perceive the eight neighboring cells around itself and has a preference about the maximum number of dissimilar agents in its neighborhood, its 'tolerance threshold.' If the number of dissimilar agents is greater than the threshold, an agent is deemed unhappy and will move to a different cell. The selection of the new destination can result from different criteria. In Schelling (1971), the agents move to the nearest spot that satisfies their tolerance threshold, but many implementations of this model relax this movement criterion, letting agents simply move to a random other spot on the checkerboard (see e.g., Wilensky, 1997, 1999). In each cycle of the simulation, each agent moves to a new spot if it is unhappy in its existing one. An agent moving into a new spot changes the existing neighborhood, creating a path-dependent dynamic. This model revealed the existence of a critical and surprisingly low tolerance threshold, at about 30%, above which an initially random distribution of agents converges to a segregated pattern as shown in Figure 1. If the clustering of same colored agents is interpreted as an increase in the organization in the system, this model shows how segregation dynamics can be an SODS. The simple preference of individuals 'not to be in a minority' together with the perception of their direct neighborhood and their possibility to move leads to high levels of clustering and in turn high levels of organization, without any external influence.

The second example concerns opinion dynamics, i.e., patterns of opinions (such as political views) in a population. There are numerous models of opinion dynamics, one of which is the bounded confidence (BC) model of the social influence on an agent's opinion (Hegselmann and Krause, 2002). In the BC model, a person can only be influenced by others who are not 'too strange' or are deemed 'reasonable' by the agent, that is, not differing too much from one's own opinion. Thus, the range of opinions in which an agent has confidence is bounded. The BC model consists of a set of

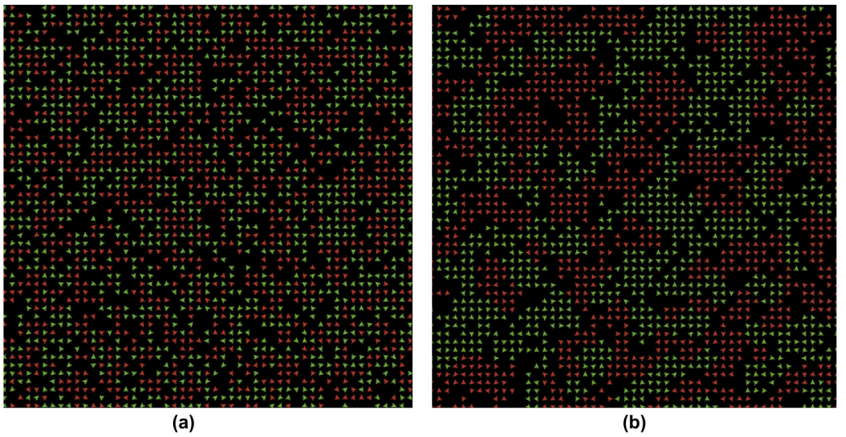

Figure 1 Schelling's segregation model in (a) the initial randomized spatial distribution and (b) the stable segregated pattern obtained after several iterations at tolerance threshold of 30% (Wilensky, 1997, 1999).

individuals, $i, j \in I$, each of which has a certain initial opinion, given by a real number, $x_i(t_0) \in (0,1)$. The profile of all opinions at time t is

$$x(t) = x_1(t), \ldots, x_i(t), x_j(t), \ldots, x_n(t).$$

Each individual i takes into account only those individuals whose opinions are not too distant, expressed by the interval $|x_i(t) - x_j(t)| \leq \varepsilon$ (confidence interval). The set of individuals that i takes into account at time t is

$$I(i, x(t)) = \{j \mid |x_i(t) - x_j(t)| \leq \varepsilon\}.$$

The opinion of individual i at time $t+1$ is the average of the opinions of those individuals in $I(I, x(t))$. Using this simple specification, the macroscopic dynamics of opinion assimilation can be studied.

If the alignment of opinions is interpreted as an increase in organization, the BC model shows how opinion dynamics can be an SODS. The level of opinion alignment reached is dependent on the size of the confidence interval ε. There are three general resulting opinion profiles: (1) plurality, a range of surviving opinions; (2) polarization, two opinions survive at a relatively large distance from each other; and (3) consensus, only one opinion survives (see Figure 2). There are phase transitions at values of $\varepsilon = 0.15$, above which there is polarization, and at $\varepsilon = 0.25$, above which there is consensus.

Both models show how social process can be modeled as SODS, with the increase in the organization of the system represented by a clustering dynamics. The difference is that the segregation model produces clustering in physical space, while the BC model of opinion dynamics produces clustering in a Cartesian space. The self-organizing dynamic in Schelling's model is the result of agents changing their physical location according to predefined rules of interaction, whereas in Hegselmann and Krause's model, the self-organization occurs because of behavioral change produced by processes of social influence.

Conclusions and Future Outlook

Despite the lack of a consensus about the definition of SODS, the concept provides interesting insights for social science. Concepts such as SODS can help to overcome the limitations of more conventional approaches in addressing process, time, and dynamics.

To develop a more comprehensive framework and definition of SODS, cross-fertilization of theory and methods is needed between mainstream social science and the emerging field of complexity science. As increasingly accessible computational power and improvements to software become available, it will become easier for social researchers to deploy modeling tools that can allow them to simulate and understand the many self-organizing phenomena that one sees in the social world.

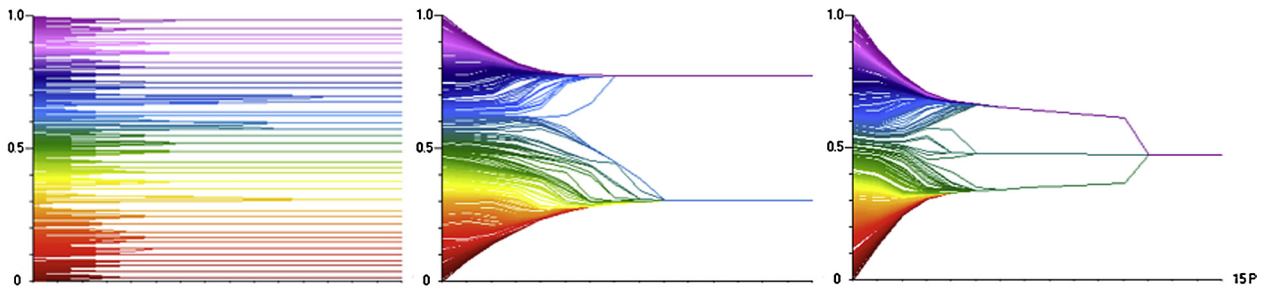

Figure 2 Plurality (left), polarization (center), and consensus (right) in the BC model of opinion dynamics (Hegselmann and Krause, 2002). Individuals are arranged along the *y* axis and are color-coded according to their opinion (from y = 0, red, to y = 1, magenta). The *x* axis represents time.

Bibliography

Ashby, William Ross, 1947. Principles of the self-organizing dynamic system. Journal of General Psychology 37, 125–128.

Bar-Yam, Yaneer, 1997. Dynamics of Complex Systems. Addison-Wesley, Reading, Mass.

Cilliers, Paul, 1998. Complexity and Postmodernism. Routledge, London.

Coll, Martinez, 1986. A bioeconomic model of Hobbes' 'state of nature'. Social Science Information 25, 493–505.

Galam, Serge, 2012. Sociophysics. Springer, New York.

Gilbert, Nigel (Ed.), 2010. Computational Social Science. Sage, London.

Hegselmann, Rainer, Krause, Ulrich, 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of Artificial Societies and Social Simulation 5. Available at http://jasss.soc.surrey.ac.uk/5/3/2.html.

Luhmann, Niklas, 1995. Social Systems. Stanford university press, Stanford.

Mandelbrot, Benoit, 1982. The Fractal Geometry of Nature. Freeman, New York.

Mantegna, Rosario, Stanley, Eugene, 2000. An Introduction to Econophysics. Cambridge university press, Cambridge.

Michell, Joel, 2007. Measurement. In: Turner, Stephen, Risjord, Mark (Eds.), Philosophy of Anthropology and Sociology. Elsevier, Amsterdam, pp. 71–119.

NetLogo Segregation model Wilensky, Uri, 1997. Center for Connected Learning and Computer-based Modeling. Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/models/Segregation.

NetLogo Wilensky, Uri, 1999. Center for Connected Learning and Computer-based Modeling. Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/.

Schelling, Thomas, 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1, 143–186.

Strogatz, Steven, 1994. Nonlinear Dynamics and Chaos. Perseus Books, Cambridge, Mass.

Von Bertalanffy, Ludwig, 1969. General System Theory. George Braziller, New York.

Wolfram, Stephen, 1994. Cellular Automata and Complexity: Collected Papers. Addison-Wesley, Reading, Mass.

Self-Regulated Learning

Philip H Winne, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Learners self-regulate learning when they examine the learning environment and monitor learning processes to inform decisions about managing learning processes. Under the assumption that learners are agents with inherent capability to make and act on decisions, self-regulated learning (SRL) is ubiquitous. Decisions learners make about regulating learning, however, can promote or impede achievement and other valued outcomes. Research on SRL and its roles in education faces multiple challenges. Evidence suggests productive SRL can be fostered through evidenced-based designs for learning activities and environments, by promoting particular beliefs among teachers, and helping learners discover value in education.

Learners in school, on the job, and pursuing their hobbies study a wide range of information – the meaning of terms such as inertia or tempo, differences among political systems and between gasoline and diesel engines, how to prevent backflow in drainage systems, and more. Beyond huge variation in topics, learners also differ enormously as learners. They begin learning with different backgrounds of knowledge. They seek various goals and differ in persistence if challenges arise. Amid a host of individual differences, at least one feature is common. As learners prepare for, engage in, and reflect on learning, they make decisions. Within the broad field of learning science, self-regulated learning, often abbreviated as SRL, encompasses theories about and research on decisions learners make about learning, factors affecting this decision-making, how frequently and easily learners make decisions, and how attributes of these elements relate to other attributes of learning such as motivation and achievement.

Characteristics of SRL

Learning is a process, not a state. Learning involves actions, mostly cognitive but also overtly behavioral, that operate on information in ways learners intend to build knowledge or set a stage for building knowledge. Viewing learning as cognitive operations that learners regulate has several key corollaries (Winne, 2010).

Information is not inherently knowledge but a learner can operate on information to transform it into knowledge. Rehearsing is one common cognitive operation; it involves mentally repeating information, such as a term's definition, or repeatedly performing an action, such as the motions in tying a bowline knot. Another common cognitive operation is translating information into a different format. Drawing a parabola to represent an algebraic expression of a quadratic equation is an example. Choosing operations is a decision that regulates learning. Therefore, SRL is a metacognitive event (Nelson and Narens, 1990) in which cognitive operations and environmental affordances that might be used in learning are the objects of decision-making about how to transform information into knowledge.

A learner's decisions about how to learn are not random choices. Decisions are directed by goals, particular outcomes a leaner forecasts will or very probably will result if a particular pattern of operations is applied. Bandura (1977) labeled this an outcome expectation. Judging the probability that a particular outcome is produced by applying particular operations is different from judging the probability those operation(s) can be carried out. This latter judgment is about capability; it is Bandura's concept of self-efficacy. Thus, SRL is inherently motivational in three senses (Winne and Marx, 1989). First, learners select which goals to pursue. Second, learners choose how to operationalize their pursuit of goals. Third, learners decide how long to pursue a goal that has not yet been achieved. As learners consider these motivationally related choices, they form attributions for outcomes and experience affective reactions. A learner may explain success as due to effort and find other than not in developing knowledge. Having accurately forecast outcomes or succeeded in translating an algebraic expression into a graphical form may satisfy the learner's quest for understanding.

Particular decisions may be motivated by external conditions, such as a parent's promise of reward for a good score on a test or a teacher's warning of extra homework for a poor score. Nonetheless, learners are the ones who choose goals and, given constraints and resources in their environment, how they will seek goals. The additional impact (an economist would say, marginal utility) of various outcomes may have different weights in the learner's system of values. A learner may judge rehearsing more efficient but not as effective as translating as a means to develop knowledge. But a graph may be preferred over algebraic form. That the learner has such choices to make and makes them gives rise to a property called agency. In the process of SRL, the learner expresses agency by deciding how learning unfolds.

In summary, SRL concerns how learners strive to build knowledge by metacognitively monitoring and regulating mental content and factors in the external environment. The objects of metacognitive monitoring include cognition, decision-making, planning, motivation, affect, and emotion. Targets for metacognitive control include: (1) fundamental cognitive operations such as rehearsing or translating, as well as larger-grained learning strategies; (2) motivationally oriented thoughts about reasons for engaging in learning (e.g., mastery or performance goal orientation) and the explanations (attributions) and value(s) of outcomes generated by engaging in

learning; (3) perceptions of feelings (affect, emotion); and (4) factors in the external environment, such as opportunities to obtain help and whether to store information externally in notes. After a stage is set for a learning episode, as the episode unfolds, decisions the learner makes in SRL alter these conditions. SRL is a dynamic process of adaptation in support of learning.

Models of SRL

Models identify and organize key factors at play in a process. Given SRL's wide scope, it is predictable that multiple models have been proposed, first appearing in the late 1980s and early 1990s alongside growth in the wider field of learning science. In differing ways, models of SRL share the characteristics described earlier. But differences among them are important.

Boekaerts' Model

Boekaerts' model addresses "students' ability to select, combine, and coordinate cognitive strategies in an effective way" (Boekaerts, 1999: 447) provided that students perceive real choices or 'alternative action paths' (p. 449) for learning. Paths are shaped by cognitive strategies students believe will develop knowledge, metacognitive knowledge and skills for directing those cognitive strategies, and goals and resources that create conditions within which learners engage with tasks.

Boekaerts postulates two inherent priorities as bases for deciding among action paths. One priority is extending knowledge and skills that expand personal resources for learning and other valued activities. Serving this priority activates a mastery mode in which applying learning strategies spends effort and potentially risks other forms of personal capital, such as self-concept or social standing. The second priority is preserving resources and averting damage to well-being. Serving this priority situates a learner in coping mode where strategies are redirected and opportunities to extend personal resources are set aside. Most learning involves tension between these priorities, motivating prudent SRL.

The hub of Boekaerts' model is an appraising system that weighs real-time information from three sources as a task unfolds. One source is the learner's integrated view of a task's inherent features (e.g., intrinsic cognitive load), instructions and other conditions established by a teacher or peer group, tangible resources (e.g., time, access to information resources), and the social context. A second source of information appraised is the learner's stockpile of knowledge, task specific skills, and metacognitive skills relevant to the task. The third collection of information appraised concerns the student's self, e.g., views about actual and ideal self, motivational elements, and values. The appraisal system weighs this complex, multidimensional bundle of information and assigns priorities to the mastery mode and the coping mode.

Boekaerts' model also describes phases in a learning episode: identifying conditions that bear on the situation, interpreting them, appraising them automatically (because this situation is characterized in a habitual way) or deliberately, setting goals and, finally, engaging with the task. While a learning episode usually unfolds sequentially across phases, a learner may retreat if significant changes arise regarding elements of the situation. Thus, an action path may be revised in the midst of engagement.

Efklides' Model

In Efklides' model of SRL (Efklides, 2011), macrolevel and microlevel factors affect interactions among metacognitive, motivational, and affective processes. Macrolevel factors, also labeled the person level of SRL, are stable characteristics that generalize across tasks, situations, and time. Conventionally, these are traits. Examples are a learner's self-concept and epistemological beliefs. At the microlevel, specific features of a task situated in a particular context interact with person-level factors. The influence of person-level factors recedes as work unfolds, and subjective experience is frequently updated. At the microlevel, topics of metacognition that shape SRL include ongoing thinking, feelings, emotions or physiological states denoting effort exertion during task processing (Efklides, 2011: 7). Because SRL operates in both levels, it is simultaneously a top-down (person) process guided by goals and a dynamic bottom-up (task × person) process. That task × person process successively revises bases for regulating learning as the state of work is updated and, particularly, as the learner subjectively experiences the flow of activity.

Cognition and affect/motivation are foci for regulation in Efklides' model; hence, its name, the metacognitive and affective model of SRL or MASRL. Three kinds of thought contents are distinguished. First, metacognitive knowledge spans what a learner knows about cognition, conditions under which forms of cognition are engaged and the profile of motivational and affective factors associated with task conditions the learner recognizes, and the products constructed in the task. Second, metacognitive experiences are moment-to-moment changes the learner perceives as work unfolds. This includes awareness that something relevant is known even though it cannot be explicitly recalled (i.e., a feeling of knowing), affect arising about the rate and amount of progress, and expectations about whether a particular cognitive tactic might be effective at this point in the task. Third, metacognitive skills are rule-based forms of knowledge that join metacognitive experiences to metacognitive knowledge about controlling how work on a task is organized. An example is a cognitive tactic such as, if the task now seems vague (a metacognitive experience), then review directions for clues about how to proceed (a metacognitive skill).

Efklides posits that SRL has three phases. In task representation, the learner builds a mental representation of a task. This representation is colored by the learner's person-level traits and may differ from an observer's account. Based on this task analysis, cognitive processing is applied to work on the task in phase two. In the third performance phase, a product is generated. Throughout all three phases, as learners monitor metacognitive experiences, adaptions may be made to cognition, metacognition, affect, and effort.

Pintrich's Model

In Pintrich's model (2000, 2004; see also Schunk, 2005), decisions about regulating learning involve factors in four

categories: cognition, motivation, behavior, and context. These factors are addressed over four phases.

In phase 1, cognition generates forecasts about how a learning task may unfold, activates prior knowledge and metacognitive knowledge, and sets goals for the task. Motivational processing focuses on identifying reasons for engaging in this task, i.e., the learner's goal orientation. Other motivational processing judges efficacy for carrying out the cognitive and metacognitive operations perceived essential to do the task, judging the value of and interest in the task's products, and developing an overall judgment of how difficult it will be to complete the task. In the behavioral arena, the learner considers how to allocate time and effort to elements in the task, and how to gather data about progress. Learners also examine whether features of context, such as instructions a teacher provides or social factors, can be regulated during work.

Phase 2 of Pintrich's model is monitoring. In the cognitive arena, judgments of learning and feelings of knowing are rendered. Motivational topics monitored include changes to efficacy expectations, attributions that explain why the task is or is not progressing, and emotional reactions. Having set standards for time and effort to assign the task, learners track these factors to detect deviations. Also, changes to contextual factors are scanned.

Phase 3 is exercising control over factors in each of the four areas. For example, if learning is judged subpar, learners may decide to review content, adjust efficacy expectations, revise the rating of task difficulty, and search for supplementary materials or seek out help.

In the fourth phase of Pintrich's model of SRL, learners react to and reflect on their experiences with the task overall. Again, each category of regulation is considered. Cognitively, learners may realize they might have fared better by activating a broader sample of knowledge related to the task versus 'diving in.' Motivationally, they may attribute low activation of knowledge as explaining why the pace of work was slow, suggesting more effort would have more satisfying results. In the behavioral arena, time may be deemed more important than first considered and, in the contextual category, it may be noted that asking the teacher for more clarity about how parts of a task are weighted would allow better allocations of time.

Pintrich's model highlights motivation by articulating it throughout all four phases. Particular emphasis is given to the learner's achievement goal orientation, i.e., whether a learner works to build expertise for its intrinsic value or to develop competence as a commodity that can be traded for social position or other external rewards.

Winne and Hadwin's Model

Winne and Hadwin's (2013) model of SRL has four phases. In phase 1, the learner surveys conditions that might affect work on a task. The model distinguishes internal cognitive conditions from external conditions. Internal conditions include stable motivational factors such as self-efficacy or achievement goal orientation, knowledge and skills that relate to the task at hand, metacognitive knowledge about when and why to apply tactics for learning, and other knowledge of self the learner judges relevant. External task conditions are factors the learner recognizes in the environment that are expected to influence work on the task. Examples include tools available, elements of social context such as whether peers are knowledgeable and willing to help and, importantly, objectives set for the learner. Overall, conditions inform the learner about the context for SRL in subsequent phases.

Based on information assembled in phase 1, in phase 2 the learner sets goals. Some goals concern qualities of cognitive operations that might apply to the task, such as how easy they are to carry out or their history of generating successful results. Product goals concern results at midpoints during the task and on completing it. Goals provide standards for monitoring how well results match intended products, setting a stage for regulation.

Having established goals, the learner completes phase 2 by planning now to accomplish them. Learning tactics are selected and assembled into more complex learning strategies. A learning tactic is an If–Then rule: If conditions of the task at a particular moment have a particular profile, Then apply a particular cognitive operation. For example, if a list of items needs to be recalled in sequence, then invent a serial mnemonic such as ROY G BIV referring to colors of the spectrum ordered by wavelength. Tactics and strategies have qualities such as how easily they can be applied, how predictably they generate expected products and how rapidly they yield products. These standards can be used to monitor cognition 'on the fly.'

Armed with goals and plans, the learner works on the task itself in phase 3. During engagement, learners who are actively SRL frequently metacognitively monitor operations and midpoint products relative to goals set in phase 2. When a discrepancy is detected, metacognitive control is applied to make adjustments.

After a task is completed or, possibly, abandoned, in phase 4 the learner considers whether to make major adaptations in the event tasks like this might arise in the future. Every element of the preceding phases can be a target for regulation. For example, regarding phase 1, a learner may judge more effort needs to be spent surveying conditions or that self-efficacy should be recalibrated.

Winne and Hadwin consider SRL is ubiquitous, not an approach to learning that learners turn off and on. Recursion is a key feature of their model; that is, products generated in any phase can be reinput to that phase to refine is products. Also important is their view that, while phases usually unfold in serial order, learners may retreat or skip a phase.

Zimmerman's Model

Zimmerman's model of SRL derives from a view that "Self-regulated learning strategies are actions and processes directed at acquiring information or skill that involve agency, purpose, and instrumentality perceptions by learners … for students' strategic actions to be described as self-regulated, one must know their academic goals and perceptions of efficacy" (Zimmerman, 1989: 329). Zimmerman's view has roots in Bandura's theory of reciprocal interactions among a person, features of the person's environment and the person's intentional behavior which, in the case of SRL, is to learn. To engage in SRL, Zimmerman stipulates that a learner (1) monitors how

learning is carried out, (2) judges qualities of that engagement or results it produces, and (3) reacts to that judgment when there are differences between standards for engagement and results. Results include inherent motivational and emotional elements. A fourth quality is important. Learners proactively search for opportunities to learn not just subjects they study but also how to improve the process of learning.

SRL events in Zimmerman's model unfold over three serial phases. In a foresight phase, the learner analyzes an assigned or a self-set task. This leads to setting goals and planning how to reach goals in a strategic way. In the foresight phase, the learner generates four key perceptions. One, self-efficacy, concerns the learner's perception of ability to perform actions in the plan. Second, the learner predicts results anticipated by performing those actions. This is an outcome expectation. A third perception concerns the learner's intrinsic interest in or valuing of this outcome. The fourth perception concerns why the learner intends to carry out this task, i.e., a learning goal orientation.

After forethought, in phase 2, the learner begins work on the task itself. Various cognitive processes are applied such as imagery, self-talk, focusing attention, and recalling and applying selected strategies. As processes are enacted, the learner monitors well results correspond to goals. Importantly, Zimmerman notes learners may deviate from plans, experimenting 'on the spot' if a new tack is predicted to improve on the original plan.

In phase 3 of Zimmerman's model, learners reflect on qualities of results actually achieved and make attributions to explain why results turned out as they did. Emotional reactions accompany attributions, as described in Weiner's (2010) theory of motivation.

Zimmerman emphasized that SRL depends on knowledge of how to learn and skill in applying that knowledge. His model also meshes cognitive and metacognitive activities with the learner's motivation, especially as it is shaped by perceptions of self-efficacy and the satisfaction experienced when reflecting on learning activities.

Methodologies and Issues in Researching SRL

An ideal method for researching SRL would generate reliable information about all aspects described by models of it. Also, because SRL is a process unfolding over time, temporal qualities of events and patterns in the succession of events should receive focus. Research has not yet fully surmounted challenges these issues present.

Multiple methods have been used to observe features of SRL. The most common are self-reports, teacher judgments, and traces. Full reviews of these and other methods, and methodological issues associated with them are available in Winne (2010), Winne and Perry (2000) and Winne et al. (2011).

Self-Report Measures of SRL

Self-report methods ask learners to describe their SRL. These are overwhelmingly prevalent in research. Within self-report methods, three main techniques are used: think-aloud protocols, interviews, and inventories.

Think-Aloud Protocols

Think-aloud protocols are grounded in Ericsson and Simon's (1993) research on using verbal reports as data. In a think-aloud protocol, the learner is instructed to talk about what goes on mentally from the moment the task is presented until it is complete or the researcher stops the learner's work. The gist of instructions is, 'Describe everything you can about how you are thinking as you work on the task.' Learners' utterances are recorded and later coded. Counts of and patterns among features of SRL are noted, including goals set, feelings that arise, plans, standards used to monitor, evaluations produced by monitoring, and so on (e.g., see Azevedo et al., 2010).

Interview Protocols

Interviews about SRL can focus on hypothetical but everyday learning tasks or features of SRL that the learner recalls about a task just completed. Information is solicited by guiding questions and follow-ups may be posed if the learner's responses are vague or mention specific features. For example, the Self-Regulated Learning Interview Schedule (SRLIS) developed by Zimmerman and Martinez-Pons (1986) poses a task: "Most students find it necessary to complete some assignments or prepare themselves for class at home. Do you have any particular methods for improving your study at home?" If the learner cannot respond or appears to need prompting, a follow up is offered: "What if you are having difficulty? Is there any particular method you use?" (Zimmerman and Martinez-Pons, 1988: 285). Learners' responses usually are categorized using a preestablished system. Or, responses may be analyzed for emergent themes and features.

Inventories

Because of the relative ease, time saved and simpler scoring that can be realized by administering inventories, they are the most common self-report methods in researching SRL. Inventories typically set a stage for learners to recall SRL events at a relatively coarse grain size, such as 'When you learn…' or 'In this course….' One common inventory is the Learning and Study Strategies Inventory (Weinstein et al., 1987) a standardized, normed 77-item self-report inventory "designed to measure use of learning and study strategies" (Weinstein et al., 1987: 2) by undergraduate students. Example items are statements such as 'I try to interrelate themes in what I am studying' and conditional relations such as 'When work is difficult I either give up or study only the easy parts.' Learners respond using a 5-point scale ranging from 'not at all typical of me' to 'very much typical of me.'

The most common inventory used in research is the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991). Its 81 items are organized into distinct subscales: intrinsic goal orientation, extrinsic goal orientation, task value, control of learning beliefs, self-efficacy for learning and performance, test anxiety, rehearsal, elaboration, organization, critical thinking, metacognitive self-regulation, time and study environment, effort regulation, peer learning, and help seeking. The MSLQ was designed "to assess college students' motivational orientations and their use of different learning strategies for a college course" (p. 3) regarding actions such as, 'I make good use of my study time for this

course' and conditional relations such as, 'When reading for this course, I make up questions to help focus my reading.' Learners are to 'answer the questions about how you study in this class as accurately as possible' using a 7-point scale anchored by 'not at all true of me: and very true of me.' Instructions and wording of items are often changed slightly to adapt to other situations than the college setting or an entire course.

Methodological Issues about Self-reports

Data generated by self-report methods require the researcher to infer how well a learner's report matches a mental event. The validity of these inferences is challenged by several factors (see Winne and Perry, 2000; Winne et al., 2011). In think-aloud protocols, researchers often find that a learner stops speaking for periods of time while working on a task. Commonly, the researcher prompts the leaner, e.g., 'Please keep speaking.' It is unknown whether SRL events were happening but not reported while the learner was silent.

In think-aloud and interview protocols, it is unknown whether learners are biased to report some kinds of events but not others. The representativeness of a learner's reports is inestimable.

Interview protocols and inventories may introduce information a learner otherwise did not know or would not have considered without the prompts these methods use. If this happens, learners may construct responses consistent with how they believe they might have acted, either because it accords with their view of themselves as a learner or because they judge it socially desirable to have behaved that way. Also, because human memory is constructive, interview and inventory reports may sharpen or dampen features of SRL, or even lead to reporting invented SRL events. How much reports vary from actual SRL is difficult to estimate.

In the case of a think-aloud report, listening to one's oral report about the state of work on a task may trigger metacognitive monitoring that otherwise would not have occurred had the report not be generated. Except for think-aloud reports, the context of SRL events is at least temporally removed from the event per se; or, it may be abstracted in ways that lead to 'average' representations of SRL rather than particular ones. This potential reduction in variance is at odds with a fundamental representation of regulation in SRL as an If–Then–Else rule governed by local and specific conditions that comprise the If element in this representation (Winne, 2010, 2011).

Teacher Ratings of Learner's SRL

Researchers have asked teachers to judge a learner's SRL. Zimmerman and Martinez-Pons' (1988) protocol, the Rating Student Self-Regulated Learning Outcomes: A Teacher Scale, is representative. It asks teachers to rate their students' SRL using 12 of 14 categories on the SRLIS; categories of behavior that usually occur outside the teacher's opportunity to see it, e.g., structuring the environment before starting homework, were omitted. Teachers rate learners using a 5-point scale: 1 = never, 5 = always. Items span learners' usage of a strategy, such as seeking teacher assistance or seeking information. Other items ask teachers to judge observable results of using a strategy, such as whether the assignments were completed on or before due dates and whether the learner set goals and made plans. Teachers' ratings and learners' self-reports on the SRLIS correlated 0.70. However, issues about the validity of inferences regarding learners' self-reports affect interpretations of this correlation, as well.

Traces

Traces are artifacts learners create as they work a task, i.e., traces are accretion data (Webb et al., 1966). Well-designed traces support strong inferences about occurrences of SRL, the context of SRL, and patterns among events that comprise SRL. Suppose a learner is searching the Internet to gather information for a project and is using online software to bookmark Web sites, highlight information within Web sites, develop terms to build a glossary and make notes about information (see Winne and Hadwin, 2013). Suppose in the midst of making a particular note, the learner opens a note made yesterday. A logged record the software makes of this event – a trace – can be interpreted as the learner having metacognitively monitored knowledge while constructing today's note, judging content in a particular note made yesterday would be helpful, and demonstrating motivation to review that information by finding and opening that note, a trace of metacognitive control. Traces of such events interspersed with other kinds of events wherein the learner does not review previous notes while making new ones supports an inference that the learner is self-regulating knowledge construction by metacognitively monitoring information in new notes and exercising metacognitive control to assemble information in old notes with new information. It is logical to infer the learner has a clear outcome expectation and judges there is sufficient incentive to apply that effort.

Traces, particularly when software records them (Winne, 2010), are less susceptible than other methodologies (except full video recording of a learner's work) to challenges in observing SRL. First, context for engaging in SRL can be more fully characterized. Second, a fuller sample of SRL events and their temporal qualities (onset, duration) can be captured without significantly intruding on the task itself. Third, patterns across traces can be investigated to develop more complex accounts of SRL. Fourth, learners' fallible and possibly biased recall of SRL is avoided. A potential important weakness of traces is that, like all data, the researcher must infer what elements of SRL a trace represents. Careful design of traces can lessen but not erase this concern.

Research Findings

Methodological issues affecting SRL research should be remembered in considering findings reported in research involving SRL.

SRL is not always beneficial. Learners sometimes report selecting methods for studying because they are easy to use or to accelerate progress even if they believe other effortful and less rapid methods for studying – the so-called desirable difficulties – would be better for learning (see Bjork et al., 2013). Similarly, procrastination and other self-handicapping techniques, such as when learners report intentionally choosing not to study so they have an excuse for poor performance (see Midgley et al., 1996), are forms of SRL

because learners intentionally decide how to regulate their engagement in (or disengagement from) learning. These decisions, however, depress achievement.

In general, learners' reports of engaging in SRL-related events have moderate positive correlations with academic achievement (e.g., Nota et al., 2004). Supporting or training learners to apply SRL-related actions can improve achievement (e.g., Michalsky, 2013; Sadi and Uyar, 2013) especially when SRL training is part of a multifaceted intervention that provides learners with tactics and strategies for accomplishing academic tasks, such as mathematics (Leidinger and Perels, 2012) and writing (Brunstein and Glaser, 2011).

Correlational and some experimental research suggest several facets that merit attention in fostering learners' use and development of productive SRL. These include designing learning tasks that are complex enough to afford genuine decision making but not so challenging as to lower self-efficacy or overload cognition (Perry, 1998), helping teachers develop beliefs that productive SRL can be supported by directly teaching learners about cognitive tactics and strategies (Dignath-van Ewijk and van der Werf, 2012), designing learning environments that increase learners' intrinsic valuing of school (van der Veen and Peetsma, 2009), and facilitating learners' attention to opportunities to metacognitively monitor and control their learning (Ifenthaler, 2012; Lawanto and Santoso, 2013).

See also: Affect-Regulation Motivation; Decision Making, Psychology of; Emotional Regulation; Health Self-Regulation, Motivational and Volitional Aspects of; Motivation, Learning, and Instruction; Self-Efficacy and Health; Self-Efficacy: Education Aspects; Self-Efficacy; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations.

Bibliography

Azevedo, R., Moos, D.C., Johnson, A.M., Chauncey, A.D., 2010. Measuring cognitive and metacognitive regulatory processes during hypermedia learning: issues and challenges. Educational Psychologist 45, 210–223.

Bandura, A., 1977. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.

Bjork, R.A., Dunlosky, J., Kornell, N., 2013. Self-regulated learning: beliefs, techniques, and illusions. Annual Review of Psychology 64, 417–444.

Boekaerts, M., 1999. Self-regulated learning: where we are today. International Journal of Educational Research 31, 445–457.

Boekaerts, M., Niemivirta, M., 2000. Self-regulated learning: finding a balance a learning goals and ego-protective goals. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, San Diego, CA, pp. 417–450.

Brunstein, J.C., Glaser, C., 2011. Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders composition skills: a randomized controlled trial. Journal of Educational Psychology 103, 922–938.

Dignath-van Ewijk, C., van der Werf, G., 2012. What teachers think about self-regulated learning: investigating teacher beliefs and teacher behavior of enhancing students' self-regulation. Education Research International 2012, 1–10.

Efklides, A., 2011. Interactions of metacognition with motivation and affect in self-regulated learning: the MASRL model. Educational Psychologist 46, 6–25.

Ericsson, K.A., Simon, H.A., 1993. Protocol Analysis: Verbal Reports as Data, second ed. MIT Press, Cambridge, MA.

Glaser, C., Brunstein, J.C., 2007. Improving fourth-grade students' composition skills: effects of strategy instruction and self-regulation procedures. Journal of Educational Psychology 99, 297–310.

Ifenthaler, D., 2012. Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. Journal of Educational Technology and Society 15, 38–52.

Lawanto, O., Santoso, H., 2013. Self-regulated learning strategies of engineering college students while learning electric circuit concepts with enhanced guided notes. International Education Studies 6, 88–104.

Leidinger, M., Perels, F., 2012. Training self-regulated learning in the classroom: development and evaluation of learning materials to train self-regulated learning during regular mathematics lessons at primary school. Education Research International 2012, 1–14.

Michalsky, T., 2013. Integrating skills and wills instruction in self-regulated science text reading for secondary students. International Journal of Science Education 35 (11), 1846–1873.

Midgley, C., Arunkumar, R., Urdan, T.C., 1996. "If I don't do well tomorrow, there's a reason": predictors of adolescents' use of academic self-handicapping strategies. Journal of Educational Psychology 88, 423–434.

Nelson, T.O., Narens, L., 1990. Metamemory: a theoretical framework and new findings. Psychology of Learning and Motivation 26, 125–173.

Nota, L., Soresi, S., Zimmerman, B.J., 2004. Self-regulation and academic achievement and resilience: a longitudinal study. International Journal of Educational Research 41 (3), 198–215.

Perry, N.E., 1998. Young children's self-regulated learning and contexts that support it. Journal of Educational Psychology 90, 715–729.

Pintrich, P.R., 2000. The role of goal orientation in self-regulated learning. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic, San Diego, CA, pp. 451–502.

Pintrich, P.R., 2004. A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review 16 (4), 385–407. Springer-Verlag.

Pintrich, P.R., Smith, D.A.F., Garcia, T., McKeachie, W.J., 1991. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ) (Technical Report No. 91-B-004). University of Michigan, School of Education, Ann Arbor, MI.

Sadi, O., Uyar, M., 2013. The relationship between self-efficacy, self-regulated learning streategies and achievement: a path model. Journal of Baltic Science Education 12, 21–33.

Schunk, D.H., 2005. Self-regulated learning: the educational legacy of Paul R. Pintrich. Educational Psychologist 40 (2), 85–94.

van der Veen, I., Peetsma, T., 2009. The development in self-regulated learning behaviour of first-year students in the lowest level of secondary school in the Netherlands. Learning and Individual Differences 19 (1), 34–46.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures. Rand-McNally, Skokie, IL.

Weiner, B., 2010. The development of an attribution-based theory of motivation: a history of ideas. Educational Psychologist 45, 28–36.

Weinstein, C.E., Schulte, A., Palmer, D., 1987. LASSI: Learning and Study Strategies Inventory. H & H Publishing, Clearwater, FL.

Winne, P.H., 2010. Improving measurements of self-regulated learning. Educational Psychologist 45, 267–276.

Winne, P.H., 2011. A cognitive and metacognitive analysis of self-regulated learning. In: Zimmerman, B.J., Schunk, D.H. (Eds.), Handbook of Self-regulation of Learning and Performance. Routledge, New York, pp. 15–32.

Winne, P.H., Hadwin, A.F., 2013. nStudy: tracing and supporting self-regulated learning in the Internet. In: Azevedo, R., Aleven, V. (Eds.), International Handbook of Metacognition and Learning Technologies. Springer, New York, pp. 293–308.

Winne, P.H., Marx, R.W., 1989. A cognitive processing analysis of motivation within classroom tasks. In: Ames, C., Ames, R. (Eds.), Research on Motivation in Education, vol. 3. Academic Press, Orlando, FL, pp. 223–257.

Winne, P.H., Perry, N.E., 2000. Measuring self-regulated learning. In: Boekaerts, M., Pintrich, P.R., Zeidner, M. (Eds.), Handbook of Self-regulation. Academic Press, Orlando, FL, pp. 531–566.

Winne, P.H., Zhou, M., Egan, R., 2011. Designing assessments of self-regulated learning. In: Schraw, G., Robinson, D.H. (Eds.), Assessment of Higher-Order Thinking Skills. Information Age Publishing, Charlotte, NC, pp. 89–118.

Zimmerman, B.J., Martinez-Pons, M.M., 1986. Development of a structured interview for assessing student use of self-regulated learning strategies. American Educational Research Journal 23 (4), 614–628.

Zimmerman, B.J., Martinez-Pons, M., 1988. Construct validation of a strategy model of student self-regulated learning. Journal of Educational Psychology 80, 284–290.

Zimmerman, B.J., 1989. A social cognitive view of self-regulated academic learning. Journal of Educational Psychology 81, 329–339.

Zimmerman, B.J., 1990. Self-regulated learning and academic achievement: an overview. Educational Psychologist 25, 3–17.

Self-Regulated Learning: Theories, Measures, and Outcomes

Barry J Zimmerman, City University of New York Graduate Center, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-regulated learning refers to how students become masters of their own learning processes. Neither a mental ability nor a performance skill, self-regulation is instead the self-directive process through which learners transform their mental abilities into task-related skills in diverse areas of functioning, such as academia, sports, music, and health. This article will define self-regulated learning and describe the intellectual context in which the construct emerged, changes in researchers' emphasis over time as well as current emphases, methodological issues related to the construct, and directions for future research.

Defining Self-Regulated Learning

Self-regulated learning involves metacognitive, motivational, and behavioral processes that are personally initiated to acquire knowledge and skill, such as goal setting, planning, learning strategies, self-reinforcement, self-recording, and self-instruction. A self-regulated learning perspective shifts the focus of educational analyses from students' learning abilities and instructional environments as fixed entities to students' self-initiated processes for improving their methods and environments for learning. This approach views learning as an activity that students do for themselves in a proactive way, rather than as a covert event that happens to them reactively as a result of teaching experiences. Self-regulated learning theory and research is not limited to asocial forms of education, such as discovery learning, self-education through reading, studying, programmed instruction, or computer-assisted instruction, but can include social forms of learning such as modeling, guidance, and feedback from peers, coaches, and teachers. The key issue defining learning as self-regulated is not whether it is socially isolated but rather whether the learner displays personal initiative, perseverance, and adaptive skill in pursuing it.

	Definitions
Self-regulation processes	Self-initiated cognitions or emotions that can affect learning and performance positively or negatively, such as setting effective or ineffective goals.
Self-regulation strategies	Self-initiated methods designed to enhance learning, such as multistep plan to solve story problems in mathematics.
Self-regulation phases	A description of self-regulation processes that are used before, during, and after efforts to learn.
Self-monitoring	Metacognitive tracking of one's self-regulatory processes or outcomes during learning and performance.
Self-efficacy	Involves motivational beliefs regarding personal competence to accomplish a task, such as correcting a sentence for grammatical errors.
Self-regulation of performance	Involves personal efforts to control adverse behaviors and/or emotions during learning, such as impulsivity or anxiety.
Self-regulation events	Measures of self-regulation that are applied during efforts to learn, such as think-aloud protocols or personal logs.

Continued

Most contemporary self-regulation theorists have avoided absolute distinctions between internal and external control of learning and have envisioned self-regulation in broader, more interactive terms. Students can self-regulate their learning not only through covert cognitive means but also through overt behavioral means, such as selecting, modifying, or constructing advantageous personal environments or seeking social support. A learner's sense of self is not limited to individualized forms of learning but includes self-coordinated collective forms of learning in which personal outcomes are achieved through the actions of others, such as family members, teammates, or friends, or through use of physical environment resources, such as tools. Thus, covert self-regulatory processes are viewed as reciprocally interdependent with behavioral, social, and environmental self-regulatory processes.

Self-regulation is defined as a variable process rather than as a personal attribute that is either present or absent. Even the most helpless learners attempt to control their functioning, but the quality and consistency (i.e., quantity) of their processes are low. (Self-regulatory processes are self-initiated cognitions or emotions that can affect learning and performance positively or negatively, such as setting effective or ineffective goals.)

Novice learners rely typically on naive forms of self-regulatory processes, such as setting nonspecific distal goals, using nonstrategic methods, inaccurate forms of self-monitoring (Self-monitoring is metacognitive tracking of one's self-regulatory processes or outcomes during learning and performance.), attributions to uncontrollable sources of causation, and defensive self-reactions. By contrast, expert learners display powerful forms of self-regulatory processes, especially during the initial phase of learning. Students' efforts to self-regulate their learning have been analyzed in terms of three cyclical learning phases. (Self-regulation phases are a description of self-regulation processes that are used before,

during, and after efforts to learn.) Forethought phase processes anticipate efforts to learn and include self-motivational beliefs, such as self-efficacy (Self-efficacy involves motivational beliefs regarding personal competence to accomplish a task, such as correcting a sentence for grammatical errors.), outcome expectations, intrinsic interest, as well as task analysis skills, such as planning, goal setting, and strategy choice. Performance phase processes seek to optimize learning efforts and include use of time management, imagery, self-verbalization, and self-observation processes. Self-reflection phase processes follow efforts to learn and provide understanding of the personal implication of outcomes. They include self-judgment processes, such as self-evaluation and attributions, and self-reactive processes, such as self-satisfaction and adaptive/defensive inferences. Because novice learners fail to use effective forethought processes proactively, such as proximal goal setting and powerful learning strategies, they must rely on reactive processes occurring after learning attempts that often have been unsuccessful. Such unfortunate experiences will trigger negative self-evaluations, self-dissatisfaction, and defensive self-reflections – all of which undermine self-motivation necessary to continue cyclical efforts to learn. By understanding self-regulation in this cyclical interactive way, qualitative as well as quantitative differences in process can be identified for intervention.

Intellectual Context for Self-Regulated Learning Research

Interest in students' self-regulated learning as a formal topic emerged during the 1970s and early 1980s out of general efforts to study human self-control. Promising investigations of children's use of self-regulatory processes like goal setting, self-reinforcement, self-recording, and self-instruction, in such areas of personal control as eating and task completion prompted educational researchers and reformers to consider their use by students during academic learning. Interest in self-regulation of learning was also stimulated by awareness of the limitations of prior efforts to improve achievement that stressed the role of mental ability, social environmental background of students, or qualitative standards of schools. Each of these reform movements viewed students as playing primarily a reactive rather than a proactive role in their own development. In contrast to prior reformers who focused on how educators should adapt instruction to students based on their mental ability, sociocultural background, or achievement of educational standards, self-regulation theorists focused on how students could proactively initiate or substantially supplement experiences designed to educate themselves.

Interest in self-regulation of learning emerged from many theoretical sources during the 1970s and 1980s. For example, operant researchers adapted the principles and technology of Skinner for personal use, especially the use of environmental control, self-recording, and self-reinforcement. Their preference for single-subject research paradigms and time series data was especially useful for individuals seeking greater self-regulation of learning. During this same time period, phenomenological researchers shifted from monolithic, global views of how self-perceptions influenced learning to hierarchical, domain-specific views and began developing new self-concept tests to assess functioning in specific academic domains. The research of Hattie, Marsh, Shavelson, and others was especially influential in gaining new currency for the role of academic domain self-concepts in learning. During this era, social learning researchers, such as Bandura, Schunk, and Zimmerman, shifted their emphasis from modeling to self-regulation and renamed the approach as social cognitive. They identified a new task-specific motive for learning, self-efficacy belief, and they linked it empirically to other social cognitive processes, such as goal setting, self-observation, self-judgment, and self-reaction. During this era, researchers, such as Corno, Gollwitzer, H. Heckhausen, Kuhl, and others, resurrected volitional notions of self-regulation to explain human efforts to pursue courses of learning in the face of competing events. In their view, self-regulatory control of action can be undermined by ruminating, extrinsic focusing, and vacillating which interfere with the formation and implementation of an intention.

Also during the 1970s and 1980s, suppressed writings of Vygotsky were published in English that explained more fully how children's inner speech emerges from social interactions and serves as a source of self-control. Cognitive behaviorists, such as Meichenbaum, developed models of internalization training on the basis of Vygotsky's description of how overt speech becomes self-directive. During this same era, cognitive constructivists shifted their interest from cognitive stages to metacognition and the use of learning strategies to explain self-regulated efforts to learn. The research and theory of Flavell played a major role in effecting this transition by describing self-regulation in terms of metacognitive knowledge, self-monitoring, and control of learning.

Research on self-regulation was also influenced by the emergence of goal theories during the 1970s and 1980s. Locke and Latham showed that setting specific, proximal, challenging but attainable goals greatly influenced the effectiveness of learners' efforts to learn. Theorists such as Ames, Dweck, Maehr, Midgley, and Nicholls identified individual differences in goal orientations that affected students' efforts to learn on their own. These researchers found that learning or mastery goal orientations facilitated persistence and effort during self-directed efforts to learn, whereas performance or ego goal orientations curtailed motivation and achievement. During this same period, another perspective emerged that focused on the role of intrinsic interest in learning. Deci, Harackiewicz, Lepper, Ryan, Vallerand, and others demonstrated that perceptions of personal control, competence, or interest in a task were predictive of intrinsic efforts to learn on one's own. This focus on intrinsic motivation was accompanied by a resurgence of research on the role of various forms of interest in self-directed learning by a host of scholars in North America, Europe, as well as elsewhere, such as Eccles, Hidi, Krapp, Renninger, Schiefele, Wigfield, and others. During these same decades, research on self-regulation of learning was influenced by cybernetic conceptions of how information is processed. Researchers such as Carver, Scheier, and Winne demonstrated the important role of executive processes, such as goal setting and self-monitoring, and feedback control loops in self-directed efforts to learn.

Changes in Emphasis over Time

Before the 1980s, researchers focused on the impact of separate self-regulatory processes, such as goal setting, self-efficacy, self-instruction, volition, strategy learning, and self-management with little consideration for their broader implications regarding student learning of academic subject matter. Interest in the latter topic began to coalesce in the mid-1980s with the publication of journal articles describing various types of self-regulated learning processes, good learning strategy users, self-efficacious learners, and metacognitive engagement, among others. By 1991, a variety of theories and some initial research on self-regulated learning and academic achievement was published in special journal issues and edited textbooks (Zimmerman and Schunk, 2001).

These accounts of academic learning, which described motivational and self-reactive as well as the metacognitive aspects of self-regulation, spurred considerable research. By the mid-1990s, a number of edited texts had been published chronicling the results of this first wave of descriptive research and experimental studies of self-regulated learning. The success of these empirical studies stimulated interest in systematic interventions to students' self-regulated learning and the results of these implementations emerged in journal articles and textbooks by the end of the 1990s.

Emphases in Current Theory and Research

Role of Self-Motivation

There is much current interest in understanding the self-motivational aspects of self-regulation as well as the metacognitive aspects. Self-regulated learners are distinguished by their personal initiative and associated motivational characteristics, such as higher self-efficacy beliefs, mastery learning goal orientations, favorable self-attributions, and intrinsic motivation, as well as by their strategic and self-monitoring competence. The sources of students' self-motivation are of both practical as well as theoretical importance. On the practical side, researchers often confront apathy or helplessness when they seek to improve students' use of self-regulatory processes, and they need viable methods for overcoming this lack of motivation. On the theoretical side, researchers need to understand how motivational beliefs and accompanying affect interact with learning processes in a way that enhances students' initiative and perseverance. A number of self-regulation models have included motivational and learning features as interactive components, such as Pintrich's self-schema model, Boekaert's three-layered model, Kuhl's action–state control model, and Bandura, Schunk, and Zimmerman's cyclical phase model. These models are designed to transcend conceptual barriers between learning and motivational processes and to understand their interaction. For example, causal attributions are not only expected to affect students' persistence and emotional reactions but also adaptations in their methods of learning. Each of these models seeks to explain how learning can become self-motivating and can sustain effort over obstacles and time. The role of motivation in these models of self-regulated learning was compared and contrasted in 2008 in an edited text (Schunk and Zimmerman, 2008). More than 13 sources of motivation were linked to self-regulated learning theoretically and empirically in this research.

Effects of Self-Regulation Training

A second issue that has attracted considerable attention during the last decade is the acquisition of self-regulatory skills in deficient populations of learners. Training programs for these skills have been conducted with diverse populations and age groups of students. These interventions have involved a variety of formats, such as separate strategy or study skill training classes, strategy training linked explicitly to subject matter classes, such as in mathematics or writing, or personal trainer services in classrooms and counseling centers.

One of the most systematic efforts to teach self-regulated learning processes for writing has been undertaken by Harris, Graham, and their colleagues. Their intervention, named the Self-regulated Strategy Development (SRSD) model, was designed to teach various genres of writing skill to learning disabled students who ranged in educational level from elementary to high school. This instructional approach sought to instill self-regulatory processes, such as goal setting, self-monitoring, self-evaluation, self-instruction, self-reinforcement, by using instructional processes such as modeling and fading of cognitive scaffoldings. More than 40 experimental and quasi-experimental studies have been conducted using this model, and a *meta*-analysis revealed that the effect size of SRSD training was large (Graham and Perin, 2007).

A second *meta*-analysis was conducted with 48 training studies of nondisabled primary school children involving regular teachers or experimental trainers (Dignath et al., 2008). Gains were reported in not only academic performance but also in strategic functioning and motivation. The size of the effect of overall self-regulatory training on academic outcomes was medium. However, those studies that provided metacognitive training designed to control, monitor, regulate, and reflect on learning produced a large effect size. Interventions that gave motivational feedback regarding their strategic learning also produced a large effect size, and studies that provided the combination of metacognitive and motivational training strategies yielded a very large effect size. Interestingly, the *meta*-analysis revealed that experimental trainers were more effective than trained teachers in enhancing students' achievement. A subsequent study by Dignath and Büttner (2008) expanded the data base to include secondary school students as well as the primary school students. These researchers found that the overall effect of self-regulatory training on secondary students' achievement was medium in size, albeit slightly smaller than the effect on primary school youngsters. The effect of metacognitive training with secondary students was large, but the effect of motivational training could not be tested due to limited number of studies of secondary students that included this component. These *meta*-analyses established clearly that children as young as primary schoolers profited significantly from self-regulatory training that focused on metacognitive and motivational components of self-regulated learning.

A fourth *meta*-analysis was conducted by Sitzmann and Ely (2011) with 369 self-regulation work-related training studies

involving nondisabled college students or employees who were taught job- or education-relevant knowledge and skills. Sixteen self-regulatory constructs were identified across these studies, and nine of these constructs were significant predictors of students' achievement: goal level, self-efficacy, metacognitive strategies, attention, time management, environmental structuring, motivation, effort, and attributions. The best predictors (i.e., goal level, persistence, effort, and self-efficacy) jointly accounted for 17% of the variance in students' achievement. The effects of self-regulatory training generalized across trainee populations and differences in length of training in these studies. The researchers recommended future research that does not depend on self-reported data but instead involves measures that can capture the dynamic nature of self-regulated learning.

Methodological Issues

Initial efforts to measure self-regulated learning processes relied on inventories in which students are asked to rate their use of specific learning strategies, various types of academic beliefs and attitudes, typical methods of study, as well as their efforts to plan and manage their study time. Another method was the use of structured interviews that involved open-ended questions about problematic learning contexts, such as writing a difficult essay, completing math assignments, and motivating oneself to study under challenging circumstances. The latter form of assessment requires students to create their own answers and requires trained coders to recognize and classify various qualitative forms of self-regulatory strategies. (Self-regulatory strategies are self-initiated methods designed to enhance learning, such as multistep plan to solve story problems in mathematics.) Although both of these approaches have reported substantial correlations with measures of academic success, they are limited by their retrospective or prospective nature. As such, they depend on recall or anticipatory knowledge rather than on actual functioning under taxing circumstances.

To avoid these limitations, recent research efforts have focused on studying self-regulation during actual learning 'events' using measures such as trace logs, think-aloud protocols, direct observations, structured diaries, and microanalysis (Zimmerman and Schunk, 2011). Trace logs refer to environmental indicators of cognition, such as computer records of requests for information when learning online. Think-aloud protocols involve recordings of students' verbalizations of their thoughts and cognitive processes while learning, such as self-directed comments about changes in strategies. Direct observations pertain to records of students' actions and comments gathered by trained coders, such as the frequency of requests for help. Structured diaries refer to prescribed records that are completed by the learner themselves, such as the quality and quantity of their studying. Microanalysis focuses on answers to specific questions that are posed to assess well-established self-regulatory processes and motivational beliefs or feelings at key points before, during, and after feedback cycles. These event measures seek to provide a window on students' self-regulatory actions and reactions as they occur in real time and dynamic environmental settings.

There is recent evidence that event measures of self-regulatory processes derived from structured diaries can surpass the validity of summary measures obtained from interviews (Nandagopal and Ericsson, 2011). Diary measures were more predictive than interview measures of differences in self-regulation processes between high-, average-, and low-achieving upper division college students majoring in science. More specifically, these researchers discovered that shifts in the timing and the adaptation of self-regulatory processes by these advanced students to varying class requirements (e.g., reading vs test preparation) were more predictive of their achievement than their overall use of these processes. However, there is other research (Winne and Jamieson-Noel, 2002) indicating that even when diary measures focused on detailed learning events, the calibration of these self-reported measures with traces of the students' actual behavior could be a problem. Clearly further research is needed that employs diverse methodologies for assessing self-regulated learning events. (Self-regulation events are measures of self-regulation that are applied during efforts to learn, such as think-aloud protocols or personal logs.)

Future Directions for Research and Theory

Role of Computer Technology

The computer has been widely embraced as an ideal instrument to study and enhance students' self-regulation of learning. Software program menus can be faded when they are no longer needed and students' performance processes and outcomes can be logged in either hidden or overt files. Computers can provide invaluable feedback to the experimenter or the learner because performance outcomes can be stored, analyzed, and graphed in countless ways to uncover underlying deficiencies. Three notable efforts to employ computer technology to study self-regulation of students' learning have emerged to date.

Winne has developed a specially designed computer environment to assess and facilitate self-regulation by students during studying. This software, which is labeled gStudy, provides learners with self-regulatory tools, such as note-taking, that can be used to study different topics. For example, an assigned reading for a history course can be downloaded into gStudy, and as the student reads the text, key passages can be highlighted, cut and pasted to a separate text file, table, or diagram. The software also provides a menu of templates for diagramming or organizing the notes. The purpose of gStudy is to teach students how to 'bootstrap' higher levels of self-regulation during studying.

Azevedo and his associates have cautioned that learning in contemporary hypermedia settings is especially challenging to educators because it is nonlinear, multirepresentational, and open-ended. To learn effectively in these environments, students need to use self-regulatory processes, such as planning, knowledge activation, metacognitive monitoring, and reflection. To implement these processes, the researchers developed MetaTutor, a hypermedia environment designed to model, trace, and foster students' self-regulation of learning. Toassess the effectiveness of MetaTutor training, the investigators have developed think-aloud protocols that can be related to computer traces (e.g., accessing specific hyperlinks) as well as to electronic measures of science achievement.

Anastasia Kitsantas and Nada Dabbagh (2010) published a teacher's manual for teaching self-regulation via supportive computer learning technologies, such as Skype, Weblogs, Wikis, Facebook, and blackboard. Six self-regulatory processes were taught: goal setting, task strategies, self-monitoring and self-evaluation, time management, and help seeking. A four-level sequential training model was employed involving observation, emulation, self-control, and self-regulation. For example, goal setting was taught initially through modeling and emulation before it was practiced and adapted to the students' learning context. Although research has been supportive of the effectiveness of these computerized efforts, *meta*-analytic studies of their relative effects are needed.

Role of Teachers

There is growing research that teachers play an important role in children's self-regulation during class as well as with homework. *Meta*-analytic research (Dignath et al., 2008) has revealed regular teachers were less effective than experimental trainers in conveying self-regulated learning processes. However, there is evidence (Stoeger and Zigler, 2008) that regular school teachers could be trained to convey self-regulated learning skills to fourth graders as part of their regular mathematics assignments during an intensive 5-week intervention. In research conducted with older at-risk students enrolled in basic mathematics courses at a technical college, instructors who were trained to provide self-regulatory instruction produced significant results (Zimmerman et al., 2011). Compared to students of untrained instructors, students who received self-regulatory training displayed more accurate self-judgments and higher achievement on periodic exams during the semester and on a final departmental exam. Clearly, teachers can be trained to convey self-regulated learning processes to students ranging from elementary to collegiate level, but research on long-term effects of such instruction is limited to date. This issue is in need of further research.

Out-of-School Influences

There is well-established research showing that out-of-school factors such as peer groups, families, and part-time employment strongly affect school, but there is only limited evidence to date that these social groups have a significant influence on students' self-regulatory development. However, Brody and his colleagues found that parental monitoring and setting standards regarding their children's performance was very predictive of the children's academic as well as behavior self-regulation, which in turn was predictive of their cognitive and social development. Another study of parental processes by Martinez-Pons found that parental modeling and support for their children's self-regulation was predictive of the youngsters' success in school. There is additional research regarding the psychosocial origins of self-regulatory competence in academic learning.

Grolnick and her colleagues have studied the effects of parental autonomy support and involvement on their children's sources of motivation, such as perceived competence, control understanding, and self-regulation beliefs. Autonomy support refers to parents' valuing and supporting their children's initiation and problem solving activities. These researchers reported the results of several correlational studies indicating a significant predictive path from parental processes to their children's achievement via their children's motivational beliefs. These researchers advanced the hypothesis that children's self-regulatory motives are developed by their parents before they enter school and continue to be supported afterward. A 3-year longitudinal study of parental involvement was conducted with third, fourth, and fifth grade elementary school students. It was found that parents that increased their involvement over time had children who increased in personal competence motivation, increased reading and math grades, and decreased their learning problems over time. Clearly there is a growing body of research regarding the important role of parents on children's self-regulation of learning, but future research is needed on the impact of peers, especially classmates, on students' self-regulation of their learning.

Self-Regulated Learning versus Self-Regulation of Performance Processes

A final issue requiring further research involves the distinction between self-regulation of learning processes and self-regulation of performance (Self-regulation of performance involves personal efforts to control adverse behaviors and/or emotions during learning, such as impulsivity or anxiety.) processes. The former construct refers to personal efforts to acquire advantageous knowledge and skill, such as proficiency in writing or sport. The latter construct refers to personal efforts to control adverse behaviors, such as impulsivity, failures in volition, and the inability to control emotions or addictions. The latter deficiencies in performance have been attributed to a lack of self-control or self-discipline and have been the focus of study by developmental and clinical researchers.

These behaviors have been defined operationally as 'self-controlled' when they are diminished through conscious effort, such as delay of gratification and relapse prevention. Although the context differs between learning settings (acquiring positive behaviors) and performance settings (avoiding negative ones), there is no inherent reason why measures of self-control should not be predictive of students' academic achievement. There is research (Duckworth and Seligman, 2005) indicating that students' self-discipline measures are significant predictors of students' academic achievement, but there is also evidence that self-regulated learning measures may predict better (Zimmerman and Kitsantas, 2014). Future research needs to address the relation between learning and performance measures of self-regulation.

Conclusion

After more than three decades of research, self-regulated learning has emerged as a widely studied construct assessing self-initiated and self-sustained efforts to acquire metacognitive knowledge and task-related skill. At the outset, self-regulation researchers focused on separate processes such as goal setting, but successful findings led them to consider the interaction of these processes with others in more complete models. In addition to metacognitive processes, integrative models

included self-motivational processes. Various integrative self-regulatory models were compared in a series of edited books and journal articles, which attracted the attention of educators seeking to enhance their students' proactivity in academic and nonacademic areas of functioning. The effectiveness of these interventions has been assessed primarily with questionnaires and interviews, and several *meta*-analyses of theses interventions have reported encouraging results. However, important questions remain to be answered regarding real time measures of self-regulatory processes, and new methodologies have emerged that can capture and explain this dynamic data in greater detail.

See also: Academic Engagement; Avoidance and Approach Motivation: A Brief History; Health Self-Regulation, Motivational and Volitional Aspects of; Motivation and Actions, Psychology of; Motivation in Youth Sport and Physical Activity: Developmental Perspectives; Motivation, Learning, and Instruction; Personal Projects; School Achievement: Motivational Determinants and Processes; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Self-Regulation in Adulthood; Teacher Motivation.

Bibliography

Dignath, C., Büttner, G., 2008. Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. Metacognition and Learning 3, 231–264.

Dignath, C., Büttner, G., Langfeldt, H., 2008. How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. Educational Research Review 3, 101–129.

Duckworth, A.L., Seligman, M.E.P., 2005. Self-discipline outdoes IQ in predicting academic performance of adolescents. Psychological Science 16, 939–944.

Graham, S., Perin, D., 2007. A meta-analysis of writing instruction for adolescent students. Journal of Educational Psychology 99, 445–476.

Kitsantas, A., Dabbagh, N., 2010. Learning to Learn with Integrative Learning Technologies: A Practical Guide for Academic Success. Information Age, Greenwich, CT.

Nandagopal, K., Ericsson, K.A., 2011. An expert performance approach to the study of individual differences in self-regulated learning activities in upper-level college students. Learning and Individual Differences. http://dx.doi.org/10.1016/j.lindif.2011.11.018.

Schunk, D.H., Zimmerman, B.J. (Eds.), 2008. Motivation and Self-Regulated Learning: Theory, Research, and Applications. Lawrence Erlbaum, New York.

Sitzmann, T., Ely, K., 2011. A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. Psychological Bulletin 137, 421–442.

Stoeger, H., Zigler, A., 2008. Evaluation of classroom-based training to improve self-regulated learning in time management tasks during homework activities with fourth graders. Metacognition and Learning 3, 207–230.

Winne, P.H., Jamieson-Noel, D., 2002. Exploring students' calibration of self reports about study tactics and achievement. Contemporary Educational Psychology 27, 551–572.

Zimmerman, B.J., Kitsantas, A., 2014. Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. Contemporary Educational Psychology 39, 145–155.

Zimmerman, B.J., Moylan, A., Hudesman, J., White, N., Flugman, B., 2011. Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. Psychological Test and Assessment Modeling 53, 108–127.

Zimmerman, B.J., Schunk, D.H. (Eds.), 2001. Self-Regulated Learning and Academic Achievement: Theoretical Perspectives, second ed. Springer, New York.

Zimmerman, B.J., Schunk, D.H. (Eds.), 2011. Handbook of Self-Regulation of Learning and Performance. Routledge, New York.

Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations*

Edmond P Bowers, G John Geldhof, Paul A Chase, and Richard M Lerner, Tufts University, Medford, MA, USA
Steinunn Gestsdóttir, University of Iceland, Stakkahlíð, Reykjavík, Iceland
Jennifer Brown Urban, Montclair State University, Montclair, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Self-regulation is a topic of enormous theoretical and empirical interest to developmental scientists, as evidenced by the number of articles, review chapters, and books or monographs pertinent to self-regulatory processes that have appeared in the literature in recent years. Self-regulation refers to the rules or laws governing individual conduct (the functioning of the self) and is understood to pertain to the individual's engagement with his or her ecology or context. Self-regulation processes include the cognitive, emotional, and behavioral facets of the individual; involve both organismic (physiological) and intentional functioning; reflect both fluid and crystallized individual attributes; and are the personological component of the mutually influential relations between the individual and the context (represented as individual ↔ context relations, or what Brandtstädter terms 'developmental regulations'). Accordingly, self-regulation constitutes the fundamental process of human development and encompasses much more than other self-governance constructs, such as self-control, grit, soft skills, or noncognitive skills. As such, self-regulation constitutes a process with both basic adaptive significance and important implications for human health.

The regulation by individuals of their relationship with their complex and changing physical, social, cultural, and historical context is the key problem for successful development across life (Baltes et al., 2006). Arguably, the understanding of the system involved in linking individuals and contexts becomes the essential intellectual challenge for developmental science.

Indeed, Posner and Rothbart (2000) have stated that "understanding self-regulation is the single most crucial goal for advancing an understanding of development" (p. 427). Accordingly, to (1) understand development, especially as conceived of within a relational developmental systems perspective (Overton and Mueller, 2012) that focuses on the mutually influential relations between individuals and contexts, or individual ↔ context relations (Brandtstädter, 1998) and, centrally, to (2) appreciate the role of a person's own contributions to this development, focus should be placed on the role of an individual's actions in regulating the course of engagement with the context and in fostering constancy and change (in actualizing plasticity) across life. Self-regulation is, then, a fundamental focus of developmental science across the life span. The features of developmental regulation may be especially salient for adaptation at times of individual or ecological transition. Given the transitions of the adolescent period, this time of life is especially important for maintaining or furthering self-regulations that enhance adaptive functioning. In addition, it represents a naturally occurring ontogenetic laboratory in which to study the structure and function of self-regulation processes.

Across their ontogeny, humans actualize a rich potential for cognitive and behavioral plasticity (Lerner, 1984). The evolutionary gains in complexity (anagenesis) that underlie human plasticity have come at a price, however, that is, neotenous development (Gould, 1977). In other words, the development of humans' eventually high-level cognitive and behavioral capacities is ontogenetically protracted. Other organisms, whose nervous systems have low ratios of association-to-sensory fibers (A/S ratios; Hebb, 1949), are more stereotyped in their eventual, final level of ontogenetic functioning (Schneirla, 1957). That is, stimulus input is highly correlated with behavioral output. These organisms are adapted to ecological niches in which they can survive and reproduce despite the fact that their behavior is regulated strictly by their context. Their relatively low level of ontogenetic plasticity (and their low A/S ratio) solves the problem of the regulation of organism – context relations and thus of adaptation (Hebb, 1949; Schneirla, 1957).

For humans, however, the situation is quite different. The complexity of their nervous systems and the multiple levels of their contexts mean that there is not necessarily one adaptive relation between context and behavior; which behaviors are requisite for adaptation are uncertain. As a consequence, although plasticity affords vast variation in behavior, the evolutionary status of humans means that the selection of adaptive options from within the array of behaviors available to them constitutes the key challenge in human development. As such, self-regulation pertains to all aspects of adaptation, as individuals alter their behaviors – as well as thoughts, attention, and emotions – to react to different contexts and modulate their reactions to and actions aimed at influencing their contexts (Shonkoff and Phillips, 2000). Accordingly, self-regulation involves multiple levels of functioning, ranging from physiological functions to complicated, intentional thought processes to organismically based or purposeful actions. The latter actions may be designed to either enact strategies to attain the aims of or goals for selected transactions with their context or to compensate effectively when goals are blocked or initial actions fail (Baltes et al., 2006; Geldhof et al., 2010; Gestsdottir and Lerner, 2008; McClelland et al., 2010). Self-regulation, then, may be defined as "the ability to flexibly activate, monitor, inhibit,

* This work was supported in part by grants from the John Templeton Foundation.

persevere and/or adapt one's behavior, attention, emotions and cognitive strategies in response to direction from internal cues, environmental stimuli and feedback from others, in an attempt to attain personally-relevant goals" (Moilanen, 2007, p. 835).

Self-Regulation in Adolescence

Across the life span, individuals live in complex physical, social, cultural, and historical contexts. To thrive, individuals make decisions about how to act in ways that meet personal needs and environmental demands (Brandtstädter, 1998). In other words, to be adaptive, individuals must act in ways that support their own healthy functioning and are of benefit to their context. Arguably, these requirements for adaptation undergo significant developmental change during the second decade of life, and changes during these years can set the course for future development (Gestsdottir and Lerner, 2008).

During adolescence, individuals need to accomplish several adaptive tasks involving self and context. First, they must establish and then build knowledge about the requirements for personal adjustment, given their particular characteristics of physiological, physical, psychological, and behavioral individuality. They also must learn the demands for adaptation present in their specific ecological niche. Second, during this portion of life, a foundation needs to be established, and then enhanced, to attain the cognitive and behavioral skills needed to set the goals necessary for survival and, even more, for thriving. Thriving involves taking actions that optimize one's chances for a life marked by health and positive exchanges with one's world (Lerner et al., 2001).

To accomplish these ends, the developing abilities for strategic thinking and executive functioning need to be coupled with the actions required for turning life goals into reality, that is, into successful personal adjustment and ecological adaptation (Baltes et al., 2006). These adaptive tasks are especially complicated during adolescence. Developing links between thinking and action in the service of adaptation occurs during adolescence in the midst of emerging and, in many cases, rapidly changing developmental trajectories. For instance, such trajectories involve changing neurological, cognitive, emotional, somatic, and behavioral characteristics that evolve in relation to normative and often nonnormative changes in key life contexts (e.g., families, peer groups, schools, and communities).

The goal of engaging in these adolescent adaptive tasks is optimizing one's developmental trajectory. When one is moving toward an end state or goal that is uncertain, how does one select the best means to get there? This question reflects the fundamental adaptive challenge of the first two decades of life (Geldhof et al., 2010; Gestsdottir and Lerner, 2008).

How might adolescents meet this challenge? As is the case for people at all points across the life span, adolescents are influenced by their environments and, in addition, they simultaneously select, create, and change these settings (Lerner, 1982). This bidirectional, or reciprocal, relation between individual and context creates 'circular functions' (Schneirla, 1957) in ontogeny and, through such relations, individuals become shapers of the contexts influencing them. Thus, individuals become producers of their own development (Lerner, 1982). Accordingly, during adolescence, individuals are active agents in their *developmental regulations* – that is, they are active agents in mutually influential relationships among the organizational levels within a dynamic, developmental system. These relationships involve variables from integrated biological through historical (temporal) levels. These integrated interlevel relations may be represented as Level 1 ↔ Level 2 and, when focused on the developing individual and his or her active, changing context, these developmental regulations may be represented as individual ↔ context relations.

When these mutually influential relations are beneficial for both the person and his or her context, they may be termed *adaptive* developmental regulations (Brandtstädter, 1998). Given the foundational and developmental knowledge acquisition and skill attainment required for thriving during adolescence, however, and the embeddedness of these requirements in multiple and complex trajectories across the developmental system, the challenges to manifest adaptive developmental regulations are enormous (and arguably ontogenetically distinct) during the adolescent portion of the life span.

Organismic and Intentional Self-Regulation

In the last 10–15 years, the theoretical focus on self-regulation has been coupled with burgeoning research about the ways in which the developing physiological, psychological, and behavioral attributes of adolescents coalesce to provide the means for these attributes to actively contribute to mutually beneficial relations with their contexts. The attributes through which the individual contributes to these developmental regulations may be termed *self-regulation*.

Clearly, our conception of self-regulation encompasses two integrated but nevertheless distinguishable processes. As noted, these processes involve either primarily physiological (or organismic) processes as well as primarily intentional processes. Gestsdottir and Lerner (2008) have explained that *intentional self-regulations* (ISRs) are contextualized actions that are aimed actively toward harmonizing demands and resources in the context of personal goals to attain better functioning and to enhance self-development. ISR is characterized by goal-directed behaviors. ISR processes are more readily available to consciousness than processes and structures of *organismic regulation*, which are broad, ontogenetically consistent (i.e., relatively continuous) attributes of a person that involve biologically based, physiological structures, and functions that contribute to the relationship an individual has with the environment. Such organismic characteristics (e.g., hypothalamic control of body temperature, circadian rhythms, pubertal timing, and temperamental attributes, such as threshold of responsivity or quality of mood) are under no or limited control of the person, and they do not involve intentional efforts of the person to regulate his or her individual ↔ context exchanges.

Both organismic and ISR processes must be integrated across adolescence for adaptive developmental regulations to exist and for the developing person to thrive, not only during the second decade of life but also across the transition into and throughout the adult years. To date, however, such an integrated and life-span approach to self-regulation during adolescence has not been fully formulated. Such an integrative

approach would require scholarship about each of the two domains of self-regulation processes as well as about their systematic interrelation within and beyond adolescence.

ISR in Adolescence

Brandtstädter (1998) has noted that a major difference between infancy and early childhood and later developmental periods – for instance, adolescence – is that the intentional actions of young children do not involve an intention to promote their development. This notion of a personal future and the intentional actions that result from such a conception begin to formulate in adolescence. Brandtstädter (1998) has explained that

"[d]uring the transition to adolescence…the individual's conceptions of self and of personal future become articulated enough to guide intentional activity. External…demands originating in the familial and larger social context become increasingly internalized and integrated into process of self regulation and self-evaluation; with the progression from a…external mode of developmental control to an increasingly intentional and autonomous mode of intentional self-development, a new and higher level in the regulation of ontogeny is reached. This reflexive-intentional mode has been given rather short shrift in developmental research" (pp. 826–827).

Moilanen (2007) made a similar argument when she differentiated between short-term self-regulation and long-term self-regulation, which she believes to be characteristic of adolescence and to represent a developmental shift in functioning. To exhibit long-term self-regulation, the young person must control impulses and direct effort over longer periods of time. This effort may involve substantial planning and, for the most part, is acquired in adolescence (Moilanen, 2007).

ISR requires the adaptive application of fluid regulatory abilities as well as the knowledge and application of crystallized self-regulation skills. Fluid self-regulatory abilities represent highly cognitive processes that emphasize growth during the first part of the life span and subsequent declines throughout adulthood and old age. Components of fluid self-regulation have been summarized as executive function and include, but are not limited to, working memory, cognitive disengagement, sustained attention, and cognitive inhibition. Crystallized self-regulatory skills represent experiential shaping of the cognitive affective processing system as discussed by Mischel (Mischel et al., 1972) and others (e.g., Geldhof et al., 2010). Crystallized components of self-regulation include, for instance, expectancy, value beliefs, and action-control beliefs.

Accordingly, ISR is a construct that includes other self-governance constructs, such as soft skills, self-control, delay of gratification, or grit. These latter constructs tend to be conceptualized and measured either in laboratory situations (e.g., delay of gratification) or in regard to situation-specific reactions to distracting or off-task stimulation (e.g., delay of gratification, self-control), to pertain to a narrow band of self-governance behaviors (e.g., persistence in the face of blocked behaviors, grit), or to constitute mislabeled or incorrectly conceptualized self-governance processes (e.g., labeling the crystallized and fluid cognitive skills involved in goal management and executive functioning as noncognitive skills is a misnomer). In turn, ISR involves (1) selecting and managing positive and important life goals; (2) using strategies to obtain the resources, or engaging in the executive functions, required to optimize one's chances of reaching selected goals (and here, self-control or delay of gratification may be some of the strategies used for optimization); and (3) having the ability to compensate effectively in the face of failure or of a blocked or lost goal (e.g., making a new goal selection in the face of loss may be a useful compensation behavior).

Considerable theoretical burden then is placed on ISR in constituting the individual component of the adaptive relationship between the individual and the context (i.e., adaptive individual ↔ context relations) requisite for both individual thriving and a positive social world. It may be that ISR does more than just enable the individual to draw from and contribute positively to the context that is supporting the positive development of the individual.

Features of Intentional Self-Regulation in Adolescence

Gestsdottir and Lerner (2008) and Gestsdottir et al. (2011) have reviewed the burgeoning literature describing the structural and functional changes in ISR across adolescence. Much of this research has used the selection (S), optimization (O), and compensation (C) (or SOC) model. Baltes, Baltes, and colleagues (e.g., Baltes et al., 2006; Freund and Baltes, 2002) developed a measure to conceptualize and index, respectively, ISR. To illustrate the features of ISR in adolescence, and to point to the impact of ISR on adaptive developmental regulations during this period, it is useful to focus on the data pertinent to the SOC model.

As presented in Gestsdottir et al. (2011), SOC scores covary positively with positive youth development (PYD) and negatively with problem behaviors. This work also identified the structure of SOC over the adolescent period. Using data from the longitudinal 4-H Study of PYD (J. Lerner et al., 2012), an assessment in early adolescence (in grades 5, 6, and 7, approximately ages 10–12 years) found that SOC scores were best represented through a global structure rather than through three differentiated processes. By grade 8, however, evidence increased for a more tripartite-like structure of elective selection (ES), optimization, and compensation. By grade 10, the presence of a four-component structure of ISR can be confirmed: ES, optimization (O), compensation (C), and loss-based selection (LBS). Even though the SOC processes are not fully differentiated or developed, even by grade 10, in each study, SOC scores correlated positively with indicators of PYD and negatively with substance use, delinquency, and depressive symptoms.

Furthermore, within and across time (i.e., at all grades) Gestsdottir et al. (2011) noted that links between SOC scores and indicators of positive development and problem behaviors were all in expected directions. Moreover, the relationship among scores for the SOC processes and for developmental outcomes (positive development, drug abuse, delinquency, and depression) was strongest for the optimization and

compensation components. In addition, SOC scores were more strongly related to indicators of positive development than to problematic behaviors, which is consistent with the idea that SOC skills contribute to successful life management – that is, to the selection and attainment of goals that can further healthy development.

Urban et al. (2010) indicated that the 4-H Study data set also was used to examine possible links between self-regulatory processes and characteristics of the ecology in which the person is embedded. For instance, analyses employing data from grades 5 to 7 illustrated the complex interplay between individual and context that may underlie the adaptive developmental regulations promoting positive development. Exploring whether the ISR ability of youth interacted with participation in out-of-school time (OST) activities in predicting PYD among adolescents living in neighborhoods with relatively low levels of ecological assets, 4-H Study-based research found that youth in these settings who had the greatest capacity to self-regulate (i.e., youth with the highest SOC scores) benefited the most from involvement in OST activities. The strength of these relationships was particularly the case for girls, however. For boys, there were fewer significant relationships found among self-regulation, OST activity involvement, and either PYD scores or scores for depression and risk behaviors (Urban et al., 2010).

In turn, Bowers et al. (2011) examined the development of ISR across seven years of adolescence (grades 5–11) to ascertain whether distinctive patterns of ISR development existed, whether these trajectories differed in relation to several grade 5 parenting characteristics, and whether ISR trajectories were linked to positive and negative developmental outcomes at grade 11. The results indicated that across the developmental period spanning grade 5 to grade 11, the majority of youth in the sample experienced a steady decline in global SOC. By the fourth wave of data collection (grade 8 or approximately 14 years of age), a subset of youth experienced a marked decrease in reported global SOC, whereas another group reported an increase in global SOC. A fourth trajectory group initially reported very low global SOC across the early grades but reported near-average SOC in later grades. Overall, group comparisons indicated that the elevated SOC and steady-decline groups reported higher levels of parental warmth, monitoring, and school involvement at grade 5, as well as higher levels of PYD and contribution at grade 11. In comparison, the late-onset group had significantly lower levels of all three parenting measures at grade 5 than the steady-decline and elevated groups, but they did not differ from these two groups in outcomes at grade 11. The reverse was true for the pronounced-decline group. Although not significantly different in the measures of parenting at grade 5, the pronounced decline group was marked by lower levels of PYD and contribution at grade 11.

Future Directions

On the basis of our review of the literature on ISR in adolescence, future research should focus on the relationships among ISR, ecological assets, and thriving in adolescence. In particular, research should take a systematic approach to the study of the longitudinal, bidirectional relationships among individual strengths and contextual assets. These nuanced analyses should seek to ascertain the presence and bases of interindividual differences in the system of relationships among ISR, ecological assets, thriving, and risk and problem behaviors. Several of the studies we have discussed point to sex differences, which should be examined further.

In addition, the 4-H Study sample (which most of the studies reviewed used) does not include representative samples of racially and socioeconomically diverse youth. As such, an important direction for future research will be to study the role of ISR in thriving among adolescents who vary in age, race, socioeconomic status, religion, and a host of other demographic categories (e.g., immigrant status, single- vs. two-parent households) as well as sex. These demographic categories are potential markers of variation in the availability of ecological developmental assets in the lives of youth. As such, future research should begin to build a knowledge base to answer questions about what assets, for youth with which characteristics or backgrounds, are linked to PYD in the face of particular facets of ISR (e.g., the SOC processes of ES, O, C, or LBS).

When asking such questions, the phrase 'characteristics or backgrounds' should be understood to include features of organismic self-regulation – for instance, temperament attributes, pubertal characteristics, or sleep–wake cycles (circadian rhythms) – of particular salience in understanding individual differences among adolescents (Gestsdottir and Lerner, 2008). In other words, a key direction for future research about the impact of ISR processes during adolescence is to consider the interactions among attributes of both the organismic and the intentional facets of the self-regulation process.

Future research would profit by considering the relationship between SOC and thriving across a range of environments – for example, resource-restricted vs. resource-abundant settings or normative environments vs. settings that have been beset by either natural disasters (floods, earth quakes, or hurricanes) or human-made calamities (e.g., wars). In addition to exploring a range of environments, it would be fruitful to explore how an individual's SOC scores relate to those of their peers. For example, do youth with relatively lower SOC scores exhibit higher levels of thriving if their closest peers score high on measures of self-regulation? Does the self-regulation of one's peers serve as a protective factor for youth? This focus will allow researchers to extend these questions to a range of contextual settings not yet considered in ISR research, either in adolescence or across other portions of the life span.

In addition to issues of sampling of people and settings, the study of ISR would gain from other methodological innovations. The 4-H Study of PYD has assessed ISR using the SOC questionnaire. Although the data derived from the 4-H study attest to the psychometric quality of the measure, particularly in regard to evidence for its concurrent and predictive validity, problems of method variance accounting for some of the findings cannot be eliminated unless questionnaire findings are triangulated with results derived from the use of other types of measures (e.g., data associated with ecological or experimental observations, interviews, or expert ratings).

Use of multiple methods to assess ISR processes may enable more differentiated knowledge about the structure and function of these processes. For instance, in regard to the functions

of the four SOC structures (ES, LBS, O, and C), theory specifies that successful development encompasses the coordinated integration on selection, optimization, and compensation (Freund and Baltes, 2002; Marsiske Lang et al., 1995). For instance, optimization efforts most likely lead to higher levels of functioning only when they are focused on a delineated number of domains of functioning instead of being diffused among many domains. Similarly, selection per se does not ensure high achievement if no goal-relevant means are applied (e.g., wanting to have high peer status but using no means – athletic, social, or academic – to attain that status will not make someone popular). Finally, the adaptiveness of compensation needs to be seen in the context of the entire goal-system (e.g., "How many other goals are there that need resources for optimization?" or "How important, relative to other goals, is a threatened goal?") and the availability of resources. It does not appear to be adaptive to put a lot of one's resources into a relatively unimportant domain of functioning at the cost of having to neglect more important goals (Freund et al., 1999).

The conceptualization of ISR as the coordinated integration of selection, optimization, and compensation implies that these interactions share attributes. Few studies, however, have examined whether selection, optimization, and compensation interact in the prediction of adolescent outcomes. Instead, a summative SOC score often is used in such prediction analyses (e.g., Zimmerman et al., 2007) or the subscales are entered without considering interactions (Gestsdottir et al., 2009). Therefore, it is important that future research systematically test for the presence of such interactions.

In the conduct of such tests, a more accurate and comprehensive picture will be obtained of the ways in which ISR behaviors are linked to adolescent development outcomes. In addition, disaggregating and assessing the interactions among SOC scores would allow researchers to create SOC profiles (e.g., one could use configural frequency analysis to test whether certain profiles of selection, optimization, and compensation behaviors characterize youth from different contexts and whether these profiles are related to outcomes (von Eye et al., 2010)). Furthermore, such disaggregation would allow for the examination of interactions among individual SOC behaviors and contexts.

In the 4-H Study reports reviewed earlier, selection was not related to any outcomes when controlling for optimization and compensation (which, in turn, were both strongly related to youth outcomes). This lack of relationship could be the result of having omitted the assessment of interactions among selection and contextual factors. Some 4-H Study data suggest that such assessments would be valuable. Napolitano (2010) studied 4-H Study youth during grade 8. He found that selection was important for some of these youth – that is, selection was important for individuals living in settings with low levels of ecological assets. Napolitano suggested that youth living in these settings have relatively few viable options for gaining thriving-related resources and, as such, selection of these resources would be of great benefit. In turn, however, in settings with relatively high levels of ecological assets, youth have many different options for attaining positive development-related goals and, as such, selection may not be as salient. These ideas are provocative and are consistent with the findings of Urban et al. (2010). Both the Urban et al. (2010) and the Napolitano (2010) findings point, then, to the usefulness of more studies that assess how different components of ISR processes may, within differentiated sectors of the context, interact in the promotion of thriving.

If processes of ISR are to fully meet the theoretical burden placed on them in regard to linking with ecological developmental assets in the promotion of thriving in adolescence, then research needs to move beyond observing the role of ISR processes within descriptive, longitudinal studies. Theory-predicated, intervention research (e.g., involving randomized control trials) is needed to assess whether enhancement of the abilities of youth to intentionally regulate their own behaviors can result in improvements in their goal selection, pursuit, and management skills and, ultimately, their characteristics of positive development. Given the consistency of support for the role of SOC processes in thriving that has been derived from the 4-H Study data set and, more generally, the evidence associated with developmental systems theories-related research about the plasticity, the potential for systematic change, that characterizes adolescent development (J. Lerner et al., 2009), optimism is warranted that successful community-based intervention programs seeking to promote gains in SOC skills can be implemented. Indeed, Lerner (2004) has noted that effective youth development programs are marked in part by the inculcation through effective mentoring of the sorts of selection, optimization, and compensation skills involved in the SOC model of ISR. The success of such programs in fostering the growth of SOC-related skills suggests that it is reasonable to launch intervention efforts to enhance thriving through mentoring that is aimed at elevating the ISR skills of diverse youth.

To test the potential effectiveness of such interventions, additional simulation-based studies may be warranted. Systems science-based methodologies, such as system dynamics, can be utilized to test the relative contribution of individual self-regulatory abilities and ecological assets to the thriving of youth. These systems-based methodologies can be used to test potential policy levers and may aid in decision making regarding the investment of resources in elevating individual's self-regulation skills and enhancing the availability of resources in community settings (Urban et al., 2011).

Conclusions

If future research is marked by these methodological and substantive innovations, then developmental scientists will be able to gain an important, new understanding of how individuals may play a positive, productive role in their own development. Such knowledge will enable scientists to contribute to a key applied implication of all approaches to PYD (J. Lerner et al., 2009). That is, by enhancing the positive connections between diverse youth and their contexts, all young people can be put on a more positive pathway across their adolescent years.

See also: Cognitive Control (Executive Function): Role of Prefrontal Cortex; Executive Functions During Childhood,

Development of; Health Self-Regulation, Motivational and Volitional Aspects of; Life Management, Developmental Psychology of; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood.

Bibliography

Baltes, P.B., Lindenberger, U., Staudinger, U.M., 2006. Life span theory in developmental psychology. In: Lerner, R.M., Damon, W. (Eds.), Handbook of Child Psychology Theoretical Models of Human Development, sixth ed., vol. 1. John Wiley & Sons Inc, Hoboken, NJ, US, pp. 569–664.

Bowers, E.P., Gestsdóttir, S., Geldhof, J., Nikitin, J., von Eye, A., Lerner, R.M., 2011. Developmental trajectories of intentional self regulation in adolescence: The role of parenting and implications for positive and problematic outcomes among diverse youth. Journal of Adolescence 34 (6), 1193–1206.

Brandtstädter, J., 1998. Action perspectives on human development. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology Theoretical Models of Human Development, fifth ed., vol. 1. Wiley, New York, pp. 807–863.

Freund, A.M., Baltes, P.B., 2002. Life-management strategies of selection, optimization and compensation: measurement by self-report and construct validity. Journal of Personality and Social Psychology 82, 642–662.

Freund, A.M., Li, K.Z.H., Baltes, P.B., 1999. The role of selection, optimization, and compensation in successful aging. In: Brandtstädter, J., Lerner, R.M. (Eds.), Action and Self-development: Theory and Research through the Life-span. Sage, Thousand Oaks, CA, pp. 401–434.

Geldhof, G.J., Little, T.D., Colombo, J., 2010. Self-regulation across the life span. In: Lamb, M.E., Freund, A.M. (Eds.), Social and Emotional Development. Handbook of Life-span Development, vol. 2. Wiley, Hoboken, NJ, pp. 116–157. Editor-in-chief: R. M. Lerner.

Gestsdottir, S., Lerner, R.M., 2008. Positive development in adolescence: the development and role of intentional self-regulation. Human Development 51 (3), 202–224.

Gestsdottir, S., Lewin-Bizan, S., von Eye, A., Lerner, J.V., Lerner, R.M., 2009. The structure and function of selection, optimization, and compensation in adolescence: Theoretical and applied implications. Journal of Applied Developmental Psychology 30, 585–600.

Gestsdottir, S., Urban, J.B., Bowers, E.P., Lewin-Bizan, S., Lerner, J.V., Lerner, R.M., 2011. Intentional self regulation, ecological assets, and thriving in adolescence: A developmental systems model. New Directions for Child and Adolescent Development 133, 61–76.

Gould, S.J., 1977. Ontogeny and Phylogeny. Belknap Press of Harvard University Press, Cambridge, MA.

Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.

Lerner, J.V., Bowers, E.P., Minor, K., Lewin-Bizan, S., Boyd, M.J., Mueller, M.K., Schmid, K.L., Napolitano, C.M., Lerner, R.M., 2012. Positive youth development: Processes, philosophies, and programs. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Developmental Psychology (2nd ed.). Handbook of Psychology, Vol. 6. Editor-in-chief: I. B. Weiner. Wiley, Hoboken, NJ, pp. 365–392.

Lerner, J.V., Phelps, E., Forman, Y., Bowers, E.P., 2009. Positive youth development. In: Lerner, R.M., Steinberg, L. (Eds.), Individual Bases of Adolescent Development (3rd ed.). Handbook of Adolescent Psychology, Vol. 1. Wiley, Hoboken, NJ, pp. 524–558.

Lerner, R.M., 1982. Children and adolescents as producers of their own development. Developmental Review 2, 342–370.

Lerner, R.M., 1984. On the Nature of Human Plasticity. Cambridge University Press, New York.

Lerner, R.M., 2004. Liberty: Thriving and Civic Engagement among American Youth. Sage, Thousand Oaks, CA.

Lerner, R.M., Freund, A.M., De Stefanis, I., Habermas, T., 2001. Understanding developmental regulation in adolescence: the use of the selection, optimization, and compensation model. Human Development 44, 29–50.

Marsiske, M., Lang, F.R., Baltes, M.M., Baltes, P.B., 1995. Selective optimization with compensation: Life span perspectives on successful human development. In: Dixon, R.A., Bäckman, L. (Eds.), Compensation for Psychological Defects and Declines: Managing Losses and Promoting Gains. Erlbaum, Hillsdale, NJ, pp. 35–79.

McClelland, M.M., Ponitz, C.C., Messersmith, E.E., Tominey, S., 2010. Self-regulation: the integration of cognition and emotion. In: Overton, W.R. (Ed.), Cognition, Biology, and Methods across the Life Span. Handbook of Life-span Development, vol. 1. Wiley, Hoboken, NJ, pp. 509–553. Editor-in-chief: R. M. Lerner.

Mischel, W., Ebbesen, E.B., Zeiss, A.R., 1972. Cognitive and attentional mechanisms in delay of gratification. Journal of Personality and Social Psychology 21, 204–218.

Moilanen, K.L., 2007. The adolescent self-regulatory inventory: the development and validation of a questionnaire of short-term and long-term self-regulation. Journal of Youth and Adolescence 36, 835–848.

Napolitano, C., 2010. Context and Adolescent Intentional Self Regulation: Testing the Positive Youth Development Model. Unpublished Master's Thesis. Tufts University, Medford, MA.

Overton, W.F., Müller, U., 2012. Meta-theories, theories, and concepts in the study of development. I. B., Weiner (Editor-in-chief). In: Lerner, R.M., Easterbrooks, M.A., Mistry, J. (Eds.), Handbook of Psychology. Developmental Psychology, vol. 6. Wiley, New York, pp. 19–58.

Posner, M.I., Rothbart, M.K., 2000. Developing Mechanisms of Self-Regulation. Development and Psychopathology 12 (3), 427–441.

Schneirla, T.C., 1957. The concept of development in comparative psychology. In: Harris, D.B. (Ed.), The Concept of Development: An Issue in the Study of Human Behavior. University of Minnesota Press, Minneapolis, MN, pp. 78–108.

Shonkoff, J.P., Phillips, D.A. (Eds.), 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academy Press, Washington, DC.

Urban, J.B., Lewin-Bizan, S., Lerner, R.M., 2010. The role of intentional self regulation, lower neighborhood ecological assets, and activity involvement in youth developmental outcomes. Journal of Youth and Adolescence 39 (7), 783–800.

Urban, J.B., Osgood, N., Mabry, P., 2011. Developmental systems science: Exploring the application of non-linear methods to developmental science questions. Research in Human Development 8 (1), 1–25.

von Eye, A., Mair, P., Mun, E.-Y., 2010. Advances in Configural Frequency Analysis. The Guilford Press, New York.

Zimmerman, S., Phelps, E., Lerner, R.M., 2007. Intentional self-regulation in early adolescence: Assessing the structure of selection, optimization, and compensations processes. European Journal of Developmental Science 1, 272–299.

Self-Regulation During Early Childhood Across Cultures, Development of

Daina Crafa, Douglas Mental Health Institute, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.B. Kopp, pp. 13862–13866, © 2001, Elsevier Ltd.

Abstract

Self-regulation is a construct that represents the development of children's abilities to follow the everyday norms and practices that are embraced by their caregivers, teachers, peers, and cultures. It is a complex and highly individualized ability that relies on multiple physiological, psychoneurobiological, and cognitive processes and is shaped by socialization and caretaker instruction. This article provides a general overview of the social history and mechanisms underlying self-regulation.

Historical and Clinical Contexts

Self-regulation is a construct that represents the development of children's abilities to follow the everyday norms and practices that are embraced by their caregivers, teachers, peers, and cultures. Self-regulation is an intrinsic part of the socialization process, especially during the formative early years of life. Most children develop self-regulation abilities by the time they are 5 years old. While young children initially have trouble with self-inhibition, they strengthen these skills while simultaneously developing control over facial and emotional expressions (Carlson, 2005). Children who do not develop these skills at a young age are at higher risk for developing behavioral disorders later in life. Although a majority of children are able to self-regulate as young as 2–3 years old, some children begin to exhibit dysregulated behaviors instead. These behaviors vary across cultures but may include resistance to requests made by adults, difficulty with routines, high levels of activity, and unfocused attention often associated with an inability 'to wait' or a poor temperament. The early emergence of dysregulated behaviors appear to act as a predecessor to the development of mental health issues later in life.

Perhaps for this reason, the modern movement of research aiming to understand self-regulation began in the late twentieth century as clinical investigations. Previous findings, primarily from North America and Western Europe, on topics such as impulse control and resistance to discipline implicated factors such as 'ineffective parenting,' as well as distortions in the child's own cognitive or language skills. With rare exceptions, few attempts were made at the time to understand antecedents, culture- or age-related correlates, and consequences of early self-regulated or dysregulated behaviors. However, as a more authoritarian and less authoritative parenting style began to emerge at the turn of the century, a more comprehensive and integrative perspective was assumed to broaden the developmental model and provide insights about multiple contributors to problematic behaviors. Family and social standards were increasingly emphasized because of their relevance for self-regulation. These studies focused on the characteristics and consequences of authoritative parenting, children's autonomy needs, and understanding how the self-determined needs of the individual balance with the value based requirements of social–cultural groups. Such concerns formed the foundation of early self-regulation research.

Most recently, advances in physiological, cognitive, and brain-based methods have shifted this research toward an examination of multileveled explanations describing mechanisms of compliance and self-regulation. Meanwhile, the increasing cultural diversity in the populations of many of the countries prominently represented in research seem to have raised awareness of the role-differing cultural norms as well as neighborhood environments may play in developing social expectations and shaping behaviors.

Three Intertwined Definitions of Self-Regulation

Self-regulation refers to the development of children's ability to follow everyday customs and valued norms embraced and prescribed by their parents or caretakers and others. Self-regulation is a vital constituent of the *socialization* process. A broad construct, self-regulation encompasses a diverse set of behaviors such as compliance, the ability to delay actions when appropriate, and modulation of emotions, motor, and vocal activities as suitable to norm-based contexts. The *self* in self-regulation is an essential feature, and involves a sentient self – one that recognizes and understands the reasons for standards and evaluates one's own actions in relation to others' feelings and needs. Thus the hallmark of self-regulation is the ability to act in accordance with various family and social values in the absence of external monitors, across a variety of situations, but neither slavishly nor mindlessly.

A second meaning of self-regulation refers to various *physiological or psychobiological processes* that function adaptively to situational demands, and often do not involve normative standards. Examples include the regulation of intensity of arousal subsequent to emotion-producing events, or the control of centrally regulated physiological and perceptual systems to novel events. Centrally regulated systems include electrodermal responses, vagal tone and heart rate, and attention control. Variation in these responses is often studied for links to temperament styles, control of arousal states, and emotion control and coping (e.g., Ochsner and Gross, 2005; Kashdan et al., 2006; Goldin et al., 2008).

A third view of self-regulation comes from recent perspectives in cognitive science, which posits that motivation and

executive functions share roles in the *cognitive processes* underlying self-regulation. Here, with an emphasis on an individual's goals and the processes that shape the person's actions, a major role is given to the self's role and to evaluation of gains and losses. Autonomy, control, self-integrity, and efficacy are essential in order for the individual to be effective in the pursuit of desired goals. Within the motivational framework, substantial attention is directed to understanding how adults facilitate or inhibit the child's autonomy.

Socialization

Despite changing emphases in socialization, three principles remain a constant; these have implications for the development of self-regulation. First, most sociocultural groups expect primary authority figures, such as parents and other family members, to begin the active process of socialization. These individuals actively expose older infants and toddlers to family norms and practices. Later, these socializing agents will be joined by others who typically include teachers, peers, friends, and neighbors. New socializing agents may reinforce previous norms, expand the interpretation of norms, or introduce new ones. Children are expected to adapt their behavior accordingly.

Second, caretakers tend to be mindful – perhaps implicitly – of a child's developmental capabilities, the family's functional needs, and sociocultural standards when they teach norms. Protection of the child from harm seems to be the most salient *initial* child-rearing value. Then the content of caretakers' socialization efforts progresses to other family concerns, and from there to the broader context of neighborhood, community, and societal norms. Across sociocultural groups, there are assumptions that toddlers and preschoolers will learn a few specific family and neighborhood norms, whereas school-aged children will adopt family and cultural customs and moral standards, formal laws, and economic values. Adolescents are expected to prepare themselves for social, emotional, and economic independence using as infrastructure the norms of the family and culture.

The third principle is that socialization and self-regulation are adaptive, bidirectional processes. Children are not passive recipients in the socialization process. With the growth of their own cognitive skills and recognition of their own self needs, children show an increasing desire to have a say in defining the everyday customs that they encounter, and the ways customs are followed. Thus each succeeding generation in modern sociocultural groups modifies the content or enactment of customs and norms in some way. An implication for children is that effective self-regulation represents thoughtful, rather than indiscriminate, adherence to family practices and social norms.

Early experiences with caretakers is not the only social factor that influences self-regulation. Cultural scripts and social expectations can also contribute to self-regulation. Parental cautions about normative standards extend beyond the immediate family to peers, neighbors, and teachers, and focus on conventions and moral values, among other topics. Teachers, neighbors, and peers also play large roles in shaping behaviors during early childhood. In many cultures, e.g., Cameroonian Nso, child-rearing is shared between many women in a village.

Physiological and Psychobiological Processes

In addition to social learning, emotion regulation also relies on physiological resources, such as oxygen, and be influenced by physiological changes, such as fluctuations in hormones and heart rate. Negative emotions appear to require more resources than positive emotions according to many current studies. For example, regulation of negative emotions increases the need for increased oxygen consumption and ventilation, as well as causing elevated heart rate (Muraven and Baumeister, 2000; Gailliot and Baumeister, 2007; Lane et al., 2011). Although regulation of positive emotions use these resources too, they do not appear to be as physiologically taxing overall. Similarly, neurobiological studies demonstrate different brain activity in children when regulating negative emotions (Lewis et al., 2006). In particular, negative emotion regulation appears to consistently affect brain areas and processes associated with response inhibition. These brain processes appear to strengthen after early childhood, when more brain regions – such as the ventral prefrontal cortex associated with self control – become involved (Lewis et al., 2006).

Cognitive Processes of Motivation and Executive Function

In recent years, studies increasingly examine executive function and conscious control over one's own thoughts and behaviors. These are not single processes but rather composites of multiple processes, including working memory, inhibition, planning, and error correction. These processes are generally associated with the problem solving necessary for pursuits of goal-oriented behaviors (Carlson and Wang, 2007). The ability to flexibly self-regulate in order to pursue goals has been observed in young children and is considered necessary for successes later in life (e.g., Carlson, 2005; Meltzoff, 2007).

The Interrelationship between Social, Physiological, and Cognitive Processes

Successful self-regulating employs a combination of adequate social, physiological, psychobiological, and cognitive processes to interact appropriately with the surrounding sociocultural environment. For example, the adaptation of social and cultural elements may not be possible without enough physiological and psychobiological resources or the proper cognitive abilities to execute the task (Blair and Diamond, 2008). Such deficits may result in poor self-regulation exhibited through behavioral problems. Despite these underlying and converging factors, self-regulation is ultimately measured by behavioral output. Whether a child is able to control his or her behaviors and comply with the pressures and demands of peers, family, and life.

Many countries employ early intervention therapies and educational programs designed to address emerging problems in young children and to teach children compensatory skills for managing impulses toward dysregulation that may be caused by any number of underlying factors (Graziano et al., 2007). These programs have led to increased academic or professional success for many children. However, these programs also have normalizing effects that are not ubiquitously valued cross-culturally and potentially lead to better outcomes for certain learning styles.

Self-Regulation: The Early Years

The construct of self-regulation has heightened awareness of dysregulation and its long-term implications. Findings from related research point to long-standing problems when young children have difficulty with compliance, delay, and impulse control. To date, however, the developmental antecedents of these problems appear to be complex and context-dependent. However, two important developments occur in the second and third years that are key for the growth of self-regulation. These are emergent selfhood and the control of attention and consciousness.

It is well known that effective norm-based behaviors require reconciliation of conflicts between self-goals and social norm demands. Among mature individuals, these conflicts are often met with reflection about courses of action, informal appraisals of costs and benefits, and plans for making reparations should they be necessary. This complex cognitive strategy is not available to young children. Understanding how very young children begin to sort through competing self and social goals may provide insights into effective paths to self-regulation.

With respect to conscious learning about social norms, this almost certainly occurs, but when and how are not well understood. However, consciousness demands psychic energy, so it is in children's best interests to sort out those family and social norms that require relatively habitual responses from those that demand additional attention or memory efforts from themselves. How children learn to differentiate and classify standards may yield understanding about strategies useful for modulating behaviors in novel norm-based contexts, and why some children falter in these situations.

Concomitant Developmental Factors

The construct of self-regulation has heightened awareness of dysregulation and its long-term implications. Findings from related research point to long-standing problems when young children have difficulty with compliance, delay, and impulse control. To date, however, the developmental antecedents of these problems appear to be complex and context-dependent. However, two important developments occur in the second and third years that are key for the growth of self-regulation. These are emergent selfhood and the control of attention and consciousness.

With respect to conscious learning about social norms, this almost certainly occurs, but when and how are not well understood. However, consciousness demands psychic energy, so it is in children's best interests to sort out those family and social norms that require relatively habitual responses from those that demand additional attention or memory efforts from themselves. How children learn to differentiate and classify standards may yield understanding about strategies useful for modulating behaviors in novel norm-based contexts, and why some children falter in these situations.

A Self-Regulation Developmental Trajectory

Learning to deal with the norm-related requests of elders is a challenge, especially for young children who still have many cognitive and language limitations. It is not surprising that self-regulation takes years to mature into an effective process. An extensive database on topics related to self-regulation in the preschool years suggests that many additional factors contribute to self-regulation – from correlates including gender (many girls learn behavioral controls sooner than boys), temperament characteristics such as controlled attention and inhibition of motor acts, verbal skills, effective self-distraction, and competent use of strategies (Eisenberg and Fabes, 1992; Metcalfe and Mischel, 1999). The biosocial complexity of self-regulation has much room for error and compensation.

Active efforts by adults to socialize children gradually dovetail with children's growth of their own behavioral repertoire. Self-regulation gradually emerges when children understand the reasons for values and standards, possess a cognizant self, are able to suppress an immediate goal that runs counter to family norms, and voluntarily assume responsibility for their own actions across a variety of situations. This autonomous, sometimes conscious, effort to frame activities with normative values stands in contrast to early forms of compliance that do not rely on knowledge or self-awareness.

Processes underlying self-regulation require some ability to adapt to new or spontaneous social contexts, expectations, and demands. Effective self-regulation is by definition an adaptive, developmental process in that children have to discover how to meet their own self needs while in general following societal standards across many settings. The task is a changing one because each age period is associated with new socialization demands imposed by family, neighbors, teachers, other individuals, and the larger social context.

See also: Affect-Regulation Motivation; Defiant Behavior During Adolescence across Cultures; Emotional Regulation; Health Self-Regulation, Motivational and Volitional Aspects of; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulated Learning; Self-Regulation During Adolescence: Variations Associated with Individual–Context Relations; Self-Regulation in Adulthood.

Bibliography

Blair, C., Diamond, A., 2008. Biological processes in prevention and intervention: the promotion of self-regulation as a means of preventing school failure. Development and Psychopathology 20, 899–911.

Carlson, S.M., 2005. Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology 28, 595–616.

Carlson, S.M., Wang, T.S., 2007. Inhibitory control and emotion regulation in preschool children. Cognitive Development 22, 489–510.

Eisenberg, N., Fabes, R.A., 1992. Emotion, regulation, and the development of social competence. In: Clark, M.S. (Ed.), Review of Personality and Social Psychology, Emotion and Social Behavior, vol. 14. Sage, Newbury Park, CA, pp. 119–150.

Gailliot, M.T., Baumeister, R.F., 2007. The physiology of willpower: linking blood glucose to self-control. Personality and Social Psychology Review 11, 303–327.

Goldin, P.R., McRae, K., Ramel, W., Gross, J.J., 2008. The neural bases of emotion regulation: reappraisal and suppression of negative emotion. Biological Psychiatry 63, 577–586.

Graziano, P.A., Reavis, R.D., Keane, S.P., Calkins, S.D., 2007. The role of emotion regulation in children's early academic success. Journal of School Psychology 45, 3–19.

Kashdan, T.B., Barrios, V., Forsyth, J.P., Steger, M.F., 2006. Experiential avoidance as a generalized psychological vulnerability: comparisons with coping and emotion regulation strategies. Behaviour Research and Therapy 44, 1301–1320.

Lane, A.M., Wilson, M.G., Whyte, G.P., Shave, R., 2011. Physiological correlates of emotion-regulation during prolonged cycling performance. Applied Psychophysiological Biofeedback 36, 181–184.

Lewis, M.D., Lamm, C., Segalowitz, S.J., Stieben, J., Zelazo, P.D., 2006. Neurophysiological correlates of emotion regulation in children and adolescents. Journal of Cognitive Neuroscience 18, 430–443.

Meltzoff, A.N., 2007. "Like me": a foundation for social cognition. Developmental Science 10, 126–134.

Metcalfe, J., Mischel, W., 1999. A hot/cool-system analysis of delay of gratification: dynamics of willpower. Psychological Review 106, 3–19.

Muraven, M., Baumeister, R.F., 2000. Self-regulation and depletion of limited resources: does self-control resemble a muscle? Psychological Bulletin 126, 247–259.

Ochsner, K.N., Gross, J.J., 2005. The cognitive control of emotion. Trends in Cognitive Sciences 9, 242–249.

Self-Regulation in Adulthood

Alexandra M Freund and Marie Hennecke, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.F. Baumeister, volume 20, pp. 13859–13862, © 2001, Elsevier Ltd.

Abstract

Self-regulation refers to the extent to which people influence, modify, or control their own behavior (including thoughts and feelings) according to goals or standards. Goals are a key concept in models of self-regulation as they direct and guide cognition and behavior over time and across situations and thereby help to organize and structure perception, thought, and behavior into meaningful action units. However, the setting and pursuit of goals also requires self-regulation. Although the self-regulatory function of goals does not seem to change with age, the way goals are set and pursued does. The current article elaborates these processes in the framework of the model of selection, optimization, and compensation. Adapting to life-span changes in the availability of resources that are necessary for goal achievement, adults shift their primary goal orientation from the attainment of gains to the maintenance of functioning and the avoidance of losses. Moreover, goal focus appears to shift from a stronger focus on the outcomes of goal attainment to a stronger focus on the process of goal pursuit. These shifts in goal orientation and goal focus help self-regulation when faced with developmental goals and contexts requiring the optimization of functioning and performance in young adulthood, and the compensation of impending or actual decline and losses in later phases of adulthood.

Introduction

Self-regulation is a popular concept with more than 850 000 hits on Google Scholar (06/2013). The term refers to a variety of different processes located within the individual such as goal setting, planning, goal pursuit, progress monitoring, action control, delay of gratification, emotion control, or the resolution of goal conflict. Despite the heterogeneity of these processes that are subsumed under the term self-regulation, they have one function in common: They are all concerned with internal processes regulating thoughts, feelings, and actions. Very broadly, self-regulation can be defined as the extent to which people influence, modify, or control their own behavior (including thoughts and feelings) according to goals or standards (Baumeister and Heatherton, 1996). Although large parts of the current literature use the terms self-regulation interchangeably with self-control, some authors distinguish between these concepts. For instance, in Kuhl's conceptualization, self-control refers to processes that support the maintenance of an active goal, whereas self-regulation refers to "maintaining one's actions in line with one's integrated self" (Kuhl and Fuhrmann, 1998: 15). Similarly, Inzlicht and Schmeichel (2012) define self-control as a person's willpower, i.e., their self-regulatory *capacity* or *resources*.

Most of the literature converges on a definition of self-regulation as the processes by which a person aligns his or her behavior, thoughts, and feelings with personal goals. The concept of goals is therefore crucial for understanding self-regulatory processes. Goals can be defined as cognitive representations of desired (or dreaded) outcomes that a person intends to attain (or avoid) through action. They serve as knowledge structures that entail information about desired outcome states and the means that are functional for attaining them (Kruglanski et al., 2002). Goals direct and guide cognition and behavior over time and across situations and thereby help to organize and structure perception, thought, and behavior into meaningful action units (Freund, 2007). In that sense, goals foster self-regulation. However, goals also require self-regulation at all phases of the motivational process:

1. *Goal setting* (i.e., committing to investing means into attaining a desired outcome) requires the deliberative act of mentally simulating the outcomes of alternative goals and comparing their desirability. Such mental simulations are effortful and require self-regulation in that it would be less effortful to just take the first goal that presents itself or give in to impulses that promise immediate rewards.
2. *Goal pursuit* (i.e., the actual implementation of means in order to achieve a certain outcome) and maintenance (i.e., persisting in goal pursuit over time) require the monitoring of goal progress in order to determine if the means are adequate or need to be corrected and the investment of resources such as time, energy, or money. The initiation and maintenance of goal pursuit requires self-regulation in that competing behavioral options (that are often more pleasant than the goal-relevant behavior) have to be warded off.
3. If a goal is not achieved, one needs to decide at some point in time whether to disengage from the goal or to continue trying. Thus, the process of *goal disengagement* necessitates the reassessment of the likelihood of achieving the selected goal and the desirability of the outcome vis-à-vis opportunity costs (i.e., which other goals cannot be pursued because resources are invested in the target goal). Self-regulation is needed in order to not just give up goal pursuit at the first signs of setbacks or stagnation but, at the same time, to let go of a goal that has a very small likelihood of success or requires more resources than justified by the value of the outcome. Moreover, goal disengagement likely also requires emotion regulation (i.e., downregulating the emotional importance of the target goal).
4. Disengagement from a goal is often accompanied by goal *reengagement*. This can be either a new goal or a previously abandoned goal. To take up a previously abandoned goal might require even higher self-regulation than selecting

a goal for the first time because one has to overcome the obstacles or reasons that led to abandoning the goal in the first place.

We will later describe some of the intriguing findings that speak to the effectiveness of self-regulation. Before doing so, we want to emphasize the significance of self-regulatory processes from a developmental perspective.

Historical Overview: Emergence of the Recognition of the Active Role of the Individual in Shaping One's Life

The term self-regulation implies that people actively contribute to their developmental pathways and are not entirely under the control of external stimuli or biological processes. Although hardly controversial nowadays, it took a revolution – the 'cognitive revolution' – to introduce this notion into psychology in the 1960s when the field was dominated by behaviorism of the Watsonian or Skinnerian type. The importance of self-regulation for describing and predicting human behavior was stimulated by research by Mischel (1968) and Bandura (1969). Their research has convincingly shown that situational, external stimuli are but one factor influencing behavior and that, when attempting to understand human behavior, it is necessary to take sources of behavior regulation into account that lie within the individual (i.e., self-regulatory processes).

Developmental psychology was dominated by the view that development is best described either in terms of biologically determined maturation (i.e., unfolding of a biological program or blueprint) or as the result of a specific learning history in the sense of classic and operant conditioning. The latter view is exemplified in Watson's famous statement that if one were to give him a dozen babies he could – by arranging the environmental conditions – produce whatever adults one would wish to see (Watson, 1925). Watson also noted that he was overstating the role of the environment but he wanted to juxtapose the introspective approach to psychology focusing on inner states such as feelings and thinking with the behaviorist position. Although the interaction of biological and environmental factors for development was also stressed early on, for instance, by Piaget, the *proactive* role of the individual in creating his or her development has become more recognized since the 1980s (e.g., Brandtstädter and Lerner, 1999; Lerner and Busch-Rossnagel, 1981). In this perspective, development is viewed as the coconstructive interplay of biology, culture/society, and the individual (Baltes, 1997). On the one hand, individuals proactively shape their environment and place themselves into certain contexts. On the other hand, individuals react to their environment by continuously adapting and reacting to their physical, cultural, and social environment.

One important way in which individuals play a proactive role in their development is by choosing, committing to, and pursuing a set of personal goals such as wanting to become a teacher or deciding to have a child. The kind of goals a person develops and commits to is embedded in sociocultural, biological, and phylogenetic factors (Freund, 2007). The less behavior is under control of these factors, the more degrees of freedom a person has to choose what personal goals he or she wants to pursue and how. In other words, the less societal or biological influences structure and determine which, when, and how to pursue goals, the more influential are self-regulatory processes (Freund et al., 2009). Although less relevant for self-regulation, it might take a lot of self-control (following Kuhl's definition) to comply with external demands and norms, particularly when the respective goals do not match personal goals, preferences, and values.

In the following, we describe how self-regulatory processes shape development and how they themselves develop across adulthood.

Self-regulation across Adulthood: Personal Goals as Building Blocks for Life Management throughout the Life Span and into Old Age

Action-theoretical approaches view goals as central building blocks of self-regulation because goals structure and direct attention, emotions, and behaviors. Similarly, a number of developmental models have stressed the importance of goal-related processes and self-regulation for development, particularly in adulthood (i.e., the model of selection, optimization, and compensation (SOC) by Baltes and Baltes, 1990; the model of optimization in primary and secondary control by Heckhausen et al., 2010; the model of assimilative and accommodative coping by Brandtstädter and Renner, 1990). As one example, we present below the model of SOC. The *SOC model* (e.g., Baltes and Baltes, 1990; Freund, 2008) posits that, throughout the life span, individuals adapt to and bring about internal and external changes over time through the processes of SOC.

The importance of *selection* of goals is based on the assumption that biological, social, and individual opportunities and constraints specify a range of alternative domains of functioning at any given age and across the life span. For instance, the likelihood of women to conceive a child peaks in early adulthood and sharply declines in the late 1930s and 1940s. Most women are aware of this biological window of opportunity to become pregnant and take it into account when thinking about the goal to have a child. Similarly, educational goals are closely related to the opportunities offered by societies to obtain schooling, secondary, and tertiary education at a given age. Finally, examples for individual opportunities and constraints for personal goals are personality traits such as extraversion (making it easier to pursue social goals) or motivational dispositions such as being highly power motivated (facilitating the pursuit of goals involving making decisions and leadership). Biological, social, and individual factors interact in outlining the range of potential goals. The number of potential goals usually exceeds the amount of internal and external resources available to an individual, necessitating a reduction of the range of goals by *selecting* a subset on which to focus one's resources. Selection refers to the development, elaboration, and commitment to a limited number of goals. This does not only allow focusing one's resources but also gives direction to development (specialization). Moreover, goals organize behavior into action sequences and reduce the complexity of a given situation as they guide attention and behavior.

According to the SOC model, by delineating the range of goals on which to focus one's resources, selection is the first important step for successfully managing one's life through self-regulation. By specifying desired outcomes, selection sets standards against which to compare one's actual state of functioning, and the discrepancy can serve as a guide for behavior initiation, maintenance, or change (Carver and Scheier, 1998). In other words, to actually achieve the selected goals, *optimization* needs to take place. Optimization denotes the process of acquiring, refining, coordinating, and applying the means or resources that are necessary to reach the goals envisioned. The importance of optimization is illustrated in the research on performance that generally shows that performance is, to a large degree, a function of the acquisition and investment of goal-relevant skills and resources. In particular, the research on expertise has demonstrated that high performance depends to a significant degree on the deliberate practice of skills, including the acquisition of new skills, their refinement, their integration, and automatization. Thus, optimization is related to higher levels of performance and functioning that, in turn, are related to the accumulation of resources. Regarding self-regulation, optimization can be regarded as the second key concept of self-regulation as it refers to the actual investment of means into the attainment of selected goals.

The third component of the SOC model is *compensation*. Similarly to optimization, compensation is aimed at meeting the goal standards set by the selected goals. However, unlike optimization, compensation aims at counteracting losses in previously available goal-relevant means. This involves the substitution of lost means by using previously unused means or by acquiring new means. The objective is to maintain functioning or minimize losses when faced with resource losses.

A second form of managing losses is to restructure one's goals by focusing resources on the most important goals, adapting goal standards, or substituting no longer achievable goals (*loss-based selection*). The SOC model posits that loss-based selection is an important process of self-regulation as it allows focusing (or redirecting) resources adaptively when other means for the maintenance of functioning (compensation) are not available or would be invested at the expense of more promising goals. Loss-based selection involves two steps: (1) disengagement from the goal that is affected by losses, and (2) reengagement into alternative goals. Research has shown that disengagement without reengagement is not particularly adaptive when confronted with losses or blocked goals (Wrosch et al., 2003). This result underscores the important role of goals for self-regulation. The processes of disengagement are not yet well understood. Central questions concern if and how people can deliberately downgrade the emotional value and importance of goals in order to be able to 'let them go' without feeling one has failed which might undermine self-efficacy and feelings of control that are needed for further goal pursuit. Another question concerned with loss-based selection is how people decide whether to persist in goal pursuit despite setbacks and losses (compensation) or to disengage from further goal pursuit (Brandstätter and Schüler, 2013). The main issue in this regard is whether this decision resembles the choice between alternative goals during goal setting or is psychologically different because of anticipated or encountered losses threatening the maintenance of goals. This question taps into the psychological asymmetry of wanting to achieve gains or trying to avoid losses (e.g., Freund and Ebner, 2005). We will address the changes of gains and losses across adulthood next.

The SOC model stresses the importance of goals and adaption to developmental ecologies for self-regulation across the life span. Age-related changes in gains and losses in resources necessitate the flexible use of goal-related strategies. During childhood and through early adulthood, the ratio of gains to losses is highly positive. Across adulthood, however, losses become more prevalent and outweigh gains in old age. Adapting to this change in the availability of resources, goals shift from a predominant orientation on gains in youth and early adulthood to an increasing importance of maintenance and the avoidance of losses in middle adulthood and old age. This age-related shift is related to subjective well-being. One explanation of the age-differential effects of goal orientation toward gains and losses is that, with increasing age, the maintenance of functioning and avoidance of losses takes on a more positive meaning. Taken together, these results concerning goal orientation underscore that it is not only important for self-regulation *that* people set and pursue goals, but *how* they represent their goals depending on the availability of resources (in this context, gain vs loss orientation).

Is the shift from a primary goal orientation toward gains to an increasing importance of maintenance and the avoidance of losses also reflected in age-related differences in goal pursuit in the service of achieving higher levels of functioning (optimization) or maintaining functioning in the face of loss or decline (compensation)? On the level of self-report, optimization and compensation decrease from middle adulthood to old age. This is probably due to the fact that both processes refer to the investment of goal-relevant means that decline with age. Importantly, however, optimization and compensation are both related to subjective indicators of successful self-regulation (e.g., subjective well-being, absence of loneliness, subjective health) across adulthood and even into very old age. However, speaking for age-differential investment of resources in optimization and compensation, behavioral studies show that older adults pursue goals more persistently when they have to compensate for a loss, whereas younger adults are more persistent when aiming at optimizing their performance. A summary of the empirical studies on SOC is provided by Freund (2008).

Goal Setting across Adulthood

Probably one of the most influential models of self-regulation is Carver and Scheier's (e.g., 1998) cybernetic feedback model. The model defines goal setting as the process that creates a negative discrepancy between the current and a desired future state, which the person strives to overcome by employing goal-relevant means. It furthermore assumes that feedback loops on different levels of the goal hierarchy inform us about how well we do in reducing the discrepancy between the current and the desired future state. According to this model, self-regulation is the process of striving for an equilibrium between an actual and a desired state. One of the features that make this model

very fruitful is that it elaborates general processes of self-regulation that are not content-specific and that can be applied equally well to goals of varying levels of abstractness and including short- and long-term goals. Basically, the model predicts that people regulate the effort they put into pursuing a goal based on goal progress. Progress is defined as the rate of discrepancy reduction between the actual state and the desired end-state. If the person proceeds to the end-state faster than desired, positive affect arises. If progress is below the desired rate, negative affect arises. The affect accompanying goal pursuit, then, serves to regulate the selection and investment of goal-relevant means (i.e., positive affect: continuing to invest the same means; negative affect: changing the intensity of means investment or switching to different means).

From a developmental perspective, however, one of the shortcomings of this model is that it cannot explain goal setting and change in goals over time. If equilibrium is the desired end-state, why would anyone set goals for themselves and strive for change? According to a life-span perspective, social age-related expectations provide a reference for the kinds of goals that should be pursued at a given point in life. In addition, social expectations reflect age-related opportunity structures that guide a person toward goals for which resources are available and that, therefore, have a higher likelihood of being achieved than nonnormative goals. In this sense, age-related expectations represent important external constraints and guide goal setting across the life span and, thereby, impact self-regulation (Freund, 2007).

The Impact of Time Perspective on Goals

Age itself is not a meaningful explanation for psychological change across the life span. However, aging has psychological implications such as the accumulation of experience, the loss of cognitive and physical resources, and the restriction of future time perspective. These changes impact motivation and self-regulation in predictable ways. Research in the realm of socioemotional selectivity theory (e.g., Carstensen et al., 1999) has focused on the impact of time perspective on motivation. As people get older, their future time perspective becomes more restricted. Carstensen and her colleagues have proposed that viewing one's future as limited also impacts the kind of goals a person wants to pursue. Accordingly, younger adults (or people with an extended future time perspective) tend to strive for what Carstensen has labeled 'expansive goals,' i.e., goals with benefits that extend into the far future, specifically acquiring knowledge and making novel experiences. In contrast, older adults (or other persons with a more limited future time perspective, e.g., terminally ill patients) tend to favor goals with more immediate benefits, specifically experiencing meaningful positive emotions in the 'here and now.' These motivational differences are important for goal setting and pursuit in the realm of social relations. Socioemotional selectivity theory suggests that older adults, due to their shorter future time perspective, select their social contacts to be emotionally rewarding and meaningful, resulting in a smaller social network with close relationships than that of younger adults, who, due to their longer future time perspective, are less selective and socialize with others for emotional as well as informational reasons.

Generalizing the preference of older adults for emotionally satisfying social relations to information processing in general, socioemotional selectivity theory postulates – and provides empirical evidence – that older adults show a bias in attention and memory for positive compared to negative stimuli (Reed and Carstensen, 2012). This seems to suggest that older adults might habitually select goals that are directed at positive outcomes as both attention and memory favor positive information. However, recent evidence points to the flexibility of the preferential processing of positive or negative information depending on the goal of information processing: When attempting to minimize risks or potential losses, older adults seem to prefer negative information more than younger adults do (Depping and Freund, 2013). Taken together, research on the positivity effect points to age-related changes in the regulatory function of positive and negative information. Moreover, it underscores the important self-regulatory function of goals not only for behavior but also for information processing.

Setting Multiple Goals across Adulthood

People tend not to pursue single goals but strive for multiple different goals at a time. However, the availability of resources such as time, energy, attention, money, or social support restricts the number of goals that individuals can attend to at any given time. In line with the SOC model, several studies support that as a reaction to age-related resource losses, older adults are more selective and therefore report less personal goals than younger adults (e.g., Riediger and Freund, 2006). Other studies, however, did not report the age-related reduction in the number or self-reported personal goals (e.g., Ogilvie et al., 2001). These diverging results might be due to methodological differences in the way people report their goals: For example, people can report goals on different levels of abstraction. Whereas one person might list the single goal of 'increasing my physical health,' another person might list several more subordinate goals that nevertheless could be subsumed under this single goal, e.g., 'eat more healthily,' 'sleep more,' 'exercise two times a week.' It might be the case that the specific instructions in these studies interact with the level of abstraction of younger and older adults' goals.

Selectivity is not only reflected in the number of goals a person pursues. Goals also differ with regard to how much they facilitate each other. For instance, the pursuit of the goal 'running a marathon' might also help to accomplish the goal of 'losing weight.' Goals are more likely to facilitate each other when means of goal pursuit are multifinal as expressed in the proverbial "killing two birds with one stone." In contrast, intergoal conflict occurs when the pursuit of one goal (e.g., running a marathon) impairs the accomplishment of another goal (e.g., being more productive at work in order to receive a promotion) because both goals necessitate different means that draw on the same resources (e.g., time). Again underscoring the importance of selection, people who focus their goals on central life-domains also report higher goal facilitation. This contributes to self-regulation as facilitation leads to

higher engagement in goal pursuit and goal achievement. In contrast, goal conflict is related to lower everyday and habitual emotional well-being (Riediger and Freund, 2006). Interestingly, older adults show more intergoal facilitation and less conflict than younger adults. This might help them to better use their (restricted) resources efficiently for the pursuit of multiple goals.

Another way to pursue multiple goals in the face of limited resources is derived from Carver and Scheier's cybernetic feedback model, namely balancing. Balancing is a strategy that describes how people invest into the pursuit of one goal until a subjectively sufficient degree of goal progress has been achieved, and then turn to another goal. Balancing occurs only if people focus on goal progress rather than goal commitment (Fishbach and Dhar, 2005). Goal-directed action fosters the subsequent allocation of resources to other goals only if it is subjectively interpreted as an indicator of progress. If, in contrast, goal-directed action is interpreted as an indicator of goal commitment, resources will be continually allocated to the goal at hand. A focus on commitment thereby promotes the persistent pursuit of one goal rather than the balancing of multiple goals. As of yet, there exist no studies on age-related changes in the tendency to interpret investment into goals as goal progress or commitment, and the associated tendencies to balance multiple goals or sequence them.

Goal Focus: The Path or the Goal

Recently, the concept of goal focus has received considerable attention. It refers to the way people think about their goals and the goal components they tend to focus on (Freund et al., 2012). Goals encompass the cognitive representations of desired outcome states and the means that are functional for attaining them (Kruglanski et al., 2002). The extent to which a person focuses on these outcome states or the means may differ at any given time. A process focus denotes the higher salience of the means of goal pursuit (i.e., the 'how' of goal pursuit), whereas an outcome focus refers to a higher salience of the desired outcomes of goal attainment (i.e., the 'why' of goal pursuit). According to Vallacher and Wegner's optimality hypothesis (e.g., Vallacher et al., 1989), people perform better if they represent difficult, complex, and unfamiliar tasks in a concrete, process-related way that breaks these tasks down into their single steps. In contrast, performance in simple and familiar tasks benefits if they are represented in an abstract, outcome-related way that focuses on their meaning and effects. Extending this research to goals that are not technically difficult but pose self-regulatory challenges, studies have shown that focusing on the means rather than the outcomes of goal pursuit increases the pursuit of difficult goals such as dieting, exercising, or learning for an exam (for an overview of this research see Freund et al., 2012).

How does goal focus change across adulthood? A number of theoretical arguments speak for a shift from outcome to process focus as people get older. First, the developmental tasks of younger adults have an inherent outcome-oriented aspect. Young adults are still in the process of 'becoming,' i.e., their success is often measured in terms of the outcomes they achieve and their age-related appropriateness. For example, young adults strive to receive a college degree, find a job, find a partner, build a home, and found a family. During middle and into old adulthood, goals shift from attaining new and better outcomes to maintaining one's health, cognitive and physical resources, and close social relationships, and avoid losses. This shift in goal orientation from achieving gains to maintenance of functioning or the avoidance of losses also fosters a shift from focusing on (new) outcomes to a stronger process focus.

Moreover, older adults' shorter future time perspective might foster the adoption of a process-focused mind-set. On the one hand, the means of goal pursuit tend to be more concrete and context-embedded than the desired outcomes. They have to be implemented 'now' whereas the outcomes of goal pursuit are by definition a consequence, sequentially later in time. As for most goals, except for those that are pursued for purely intrinsic reasons (because their pursuit is fun), desired outcomes are psychologically more distant – their attainment is a consequence of goal pursuit and happens later in time. As older adults have a more restricted future time perspective, the process of pursuing a goal in the now and here might be of higher importance and more salient to them than the promise of achieving desired outcomes in the future. Focusing on the present might also come with a stronger focus on the process and the attempt to make this process as pleasant as possible. Empirical results support the hypothesis of an increase in process focus across adulthood (see Freund et al., 2012).

Outlook

The concept of self-regulation lies at the heart of models of successful development across adulthood. These models stress the importance of the setting, pursuit, and maintenance of goals as well as the disengagement from unavailable goals. Whereas the self-regulatory function of goals and goal-related processes does not seem to change with age, the content, orientation, and focus of goals adapt to life-span changes in resources and contexts.

Although models of successful development have acknowledged that the contexts in which people live are of major importance for their development, the exact processes and mechanism of how individuals use self-regulation to deal with varying contextual factors, constraints, demands, and opportunities, in order to promote successful development, have not yet been addressed. During childhood, adolescence, and young adulthood, biological constraints and social expectations regarding the kinds of goals to be pursued are relatively strong. In late adulthood and old age, there are fewer normative age-related expectations. This relatively greater social freedom might strengthen the importance of self-regulation through self-set personal goals. Going back to the distinction between self-control and self-regulation put forth by Kuhl, self-control might be more important in young adulthood, when external demands are stronger, and self-regulation might become more important in old age, when external demands are weaker. At the same time, however, resources needed for shaping one's environment according to personal goals decline in old age. Although it has been shown that older adults are still successful in pursuing their goals (e.g., Hennecke and Freund,

2010), the vicissitudes of old and very old age probably set boundaries to the extent that individuals can self-regulate their development.

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Control Behavior: Psychological Perspectives; Coping across the Lifespan; Health Self-Regulation, Motivational and Volitional Aspects of; Mastery Learning; Motivation and Actions, Psychology of; Self and Emotional Development in Adulthood and Later Life; Self-Regulated Learning: Theories, Measures, and Outcomes.

Bibliography

Baltes, P.B., 1997. On the incomplete architecture of human ontogeny: selection, optimization, and compensation as foundation of developmental theory. American Psychologist 52, 366–380.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives in successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, pp. 1–34.

Bandura, A., 1969. Principles of Behavior Modification. Holt, Rinehart, & Winston, New York.

Baumeister, R.F., Heatherton, T.F., 1996. Self-regulation failure: an overview. Psychological Inquiry 7, 1–15.

Brandtstädter, J., Lerner, R.M. (Eds.), 1999. Action and Development: Origins and Functions of Intentional Self-development. Sage, Thousand Oaks, CA, pp. 401–434.

Brandtstädter, J., Renner, G., 1990. Tenacious goal pursuit and flexible goal adjustment: explication and age-related analysis of assimilative and accommodative strategies of coping. Psychology and Aging 5, 58–67.

Brandstätter, V., Schüler, J., 2013. Action crisis and cost-benefit thinking: a cognitive analysis of a goal-disengagement phase. Journal of Experimental Social Psychology 49, 543–553.

Carstensen, L.L., Isaacowitz, D.M., Charles, S.T., 1999. Taking time seriously: a theory of socio-emotional selectivity. American Psychologist 54, 165–181.

Carver, C.S., Scheier, M.F., 1998. Self-regulation of Behavior. Cambridge University Press, Cambridge.

Depping, M.K., Freund, A.M., 2013. When choice matters: task-dependent memory effects in older adulthood. Psychology and Aging 28, 923–936.

Fishbach, A., Dhar, R., 2005. Goals as excuses or guides: the liberating effect of perceived goal progress on choice. Journal of Consumer Research 32, 370–377.

Freund, A.M., 2007. Differentiating and integrating levels of goal representation: a life-span perspective. In: Little, B.R., Salmela-Aro, K., Nurmi, J.E., Phillips, S.D. (Eds.), Personal Project Pursuit: Goals, Action and Human Flourishing. Erlbaum, Mahwah, NJ, pp. 247–270.

Freund, A.M., 2008. Successful aging as management of resources: the role of selection, optimization, and compensation. Research in Human Development 5, 94–106. http://dx.doi.org/10.1080/15427600802034827.

Freund, A.M., Ebner, N.C., 2005. The aging self: shifting from promoting gains to balancing losses. In: Greve, W., Rothermund, K., Wentura, D. (Eds.), The Adaptive Self: Personal Continuity and Intentional Self-development. Hogrefe & Huber, Ashland, OH, pp. 185–202.

Freund, A.M., Hennecke, M., Mustafic, M., 2012. On means and ends: process and outcome focus. In: Ryan, R.M. (Ed.), The Oxford Handbook of Human Motivation. Oxford University Press, New York, NY, pp. 280–300.

Freund, A.M., Nikitin, J., Ritter, J.O., 2009. Psychological consequences of longevity: the increasing importance of self-regulation in old age. Human Development 52, 1–37.

Heckhausen, J., Wrosch, C., Schulz, R., 2010. A motivational theory of life-span development. Psychological Review 117, 32–60.

Hennecke, M., Freund, A.M., 2010. Staying on and getting back on the wagon: age-related improvement in self-regulation during a low-calorie diet. Psychology and Aging 25, 876.

Inzlicht, M., Schmeichel, B.J., 2012. What is ego depletion? Toward a mechanistic revision of the resource model of self-control. Perspectives on Psychological Science 7, 450–463.

Kuhl, J., Fuhrmann, A., 1998. Decomposing self-regulation and self-control: the volitional components inventory. In: Heckhausen, J., Dweck, C.S. (Eds.), Motivation and Self-regulation across the Life Span. Cambridge University Press, Cambridge, pp. 15–49.

Kruglanski, A.W., Shah, J.Y., Fishbach, A., Friedman, R., Chun, W.Y., Sleeth-Keppler, D., 2002. A theory of goal systems. In: Advances in Experimental Social Psychology, vol. 34. Academic Press, San Diego, CA, pp. 331–378.

Lerner, R.M., Busch-Rossnagel, A. (Eds.), 1981. Individuals as Producers of Their Development: A Life-Span Perspective. Academic Press, New York.

Mischel, W., 1968. Personality and Assessment. Wiley, New York.

Ogilvie, D.M., Rose, K.M., Heppen, J.B., 2001. A comparison of personal project motives in three age groups. Basic and Applied Social Psychology 23, 207–215.

Reed, A.E., Carstensen, L.L., 2012. The theory behind the positivity-effect. Frontiers in Emotion Science 3, 339.

Riediger, M., Freund, A.M., 2006. Focusing and restricting: two aspects of motivational selectivity in adulthood. Psychology and Aging 21, 173–185.

Vallacher, R.R., Wegner, D.M., Somoza, M., 1989. That's easy for you to say: action identification and speech fluency. Journal of Personality and Social Psychology 56, 199–208.

Watson, J.B., 1925. Behaviorism. Norton, New York.

Wrosch, C., Scheier, M.F., Miller, G.E., Schulz, R., Carver, C.S., 2003. Adaptive self-regulation of unattainable goals: goal disengagement, goal re-engagement, and subjective well-being. Personality and Social Psychology Bulletin 29, 1494–1508.

Semantic Differential

Andrea Ploder and Anja Eder, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After its introduction in the 1950s, the semantic differential (SD) received enormous attention among psychologists and other social and behavioral scientists. Especially during the 1960s, its methodological details and possible fields of application were broadly and critically discussed. In 1969, David R.J. Heise reports more than 1000 papers on SD – a huge amount given the publication rates of that time. A good deal of this attention might be due to its promise to 'repair' a vital defect of behaviorist thinking, its smart use of factor analysis (a fashionable method at that time), and its connectivity to prominent debates on meaning in linguistics and philosophy of language. Especially Charles E. Osgood's claim for a universal structure of meaning in the EPA (evaluation, potency, and activity) structure points to the heart of the debate on semantic relativism.

In the early 1970s, behaviorism lost its position as the leading paradigm of psychology whereas cognitivism gained increased attention. The interest in the SD decreased. Still, SD joined the group of standard scaling techniques during the 1980s and 1990s. Today, it is applied by many disciplines worldwide, such as psychology, sociology, linguistics, marketing research, etc.

History of the Semantic Differential

The semantic differential (SD) was introduced and mainly developed by the US psychologist Charles E. Osgood (1916–91). It is a type of semantic rating scale measuring the connotative meaning of concepts like terms, objects, events, activities, ideas, etc. It captures the affective and cognitive components of respondents' attributions to selected concepts on a multidimensional level. Different from other rating scales, the SD is universally applicable – it measures associations, motivations, emotions, attitudes, etc., for almost every concept.

The original SD questionnaires contained a list of seven-point scales (−3; 0; +3) between bipolar, contrasting adjectives (e.g., good–bad, light–dark, social–unsocial) and a neutral zero point. Merely the poles of SD scales are labeled and the scales' intervals are assumed to be of equal distance (Figure 1).

The major purpose of the SD is to understand the connotative meaning of concepts, by measuring the connotative similarities, and differences within and between respondents' ratings. Factor analysis is the central procedure to analyze the proximity and distance between SD ratings and to determine the different dimensions by which they are represented in semantic space (Figure 2). In polarity profiles, the distribution of response patterns, the hetero- and homogeneity of the respondents' evaluations, is illustrated graphically (see Figure 3) and provides the basis for further correlation-based analysis.

Today, the SD is a standard scaling technique in the social and behavioral sciences. It is most prominent in psychology, but also used in linguistics, sociology, marketing research, and other empirically oriented disciplines.

Tornado

Fair ---- : ---- : ---- : ---- : ---- : ---- : ---- Unfair

+3 +2 +1 0 –1 –2 –3

Figure 1 Semantic differential on the tornado and fairness, with labeled poles, equidistant scale intervals, and a neutral zero point. Reproduced from Osgood, Ch. E., Tzeng, O. (Eds.), 1990. Language, Meaning, and Culture: The Selected Papers of C.E. Osgood. Praeger Publishers, p. 9.

Methodological Predecessors within the Social Sciences: Other Types of Scaling

Rating scales came up during the 1920s and 1930s, a period of formative methodic and methodological inventions. Due to their broad applicability, scaling procedures were quickly adopted in a lot of disciplines within the social and behavioral sciences. The SD was not the first scaling procedure for the measurement of attitudes. Still prominent are the Thurstone scale (a 'method of equal appearing intervals'), the Likert scale (a 'method of summative ratings'), and the Guttman scale (a 'method of scale analysis' coming up in the 1950s) (Edwards and Kilpatrick, 1948).

A Thurstone scale consists of statements, which are judged along graduations, i.e., 'very agree' to 'very disagree.' The concept was introduced in 1928 by the US psychologist Louis L. Thurstone. Thurstone scales are also assumed to be equidistant, but they do not operate with bipolar adjectives. Thurstone also was one of the founders of multiple factor analysis, a central method for the analysis of SD scales (Thurstone, 1948). The US social psychologist Rensis Likert (1932) introduced a five-point rating scale. Like the SD, it contains a neutral midpoint. Though, Likert scales do not use adjective lists but phrase positive or negative statements which together make up for a set of items (a 'summated rating scale'). Today, the Likert scale is by far the most famous and widely used scaling procedure within the social and behavioral sciences. A Guttman scale is dichotomous, each item being responded with either 'yes' or 'no.' It was developed by the US-Israelian sociologist Louis Guttman in 1944 and contains a set of phrased items, arranged in a consecutive order. Whoever agrees to an item on a higher level needs to agree to all items on the levels below.

The idea to measure not only attitudes but *meaning* with scales was brought up by the neobehaviorists John A. Glaze (1928) and Clark L. Hull (1933), who were among the first to scale associations. In a late autobiographical communication to a colleague, Osgood stresses the enormous influence of Hulls work on his thinking: "I had the heady feeling that

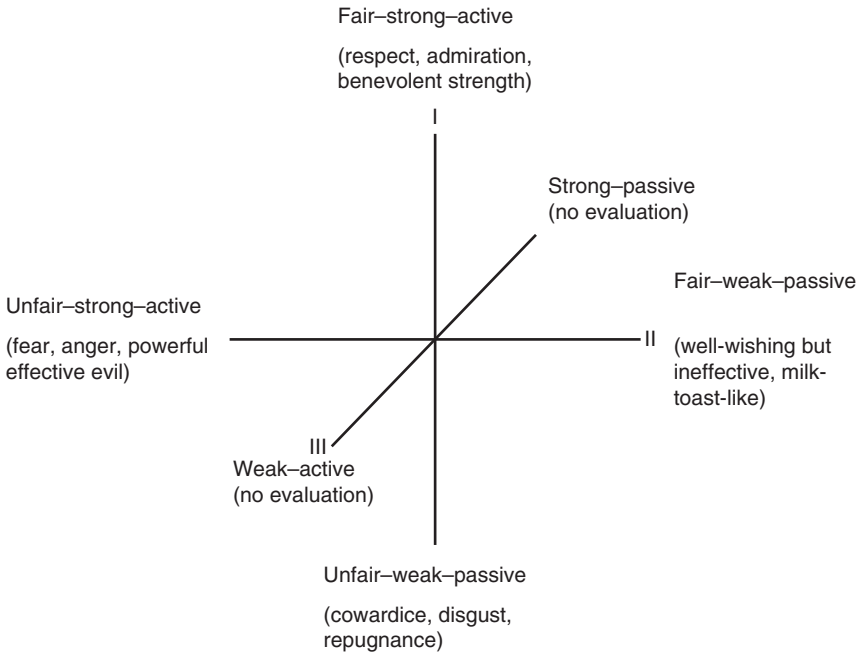

Figure 2 Semantic space resulting from SD data on political judgments. Reproduced from Osgood, Ch. E., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana, p. 121.

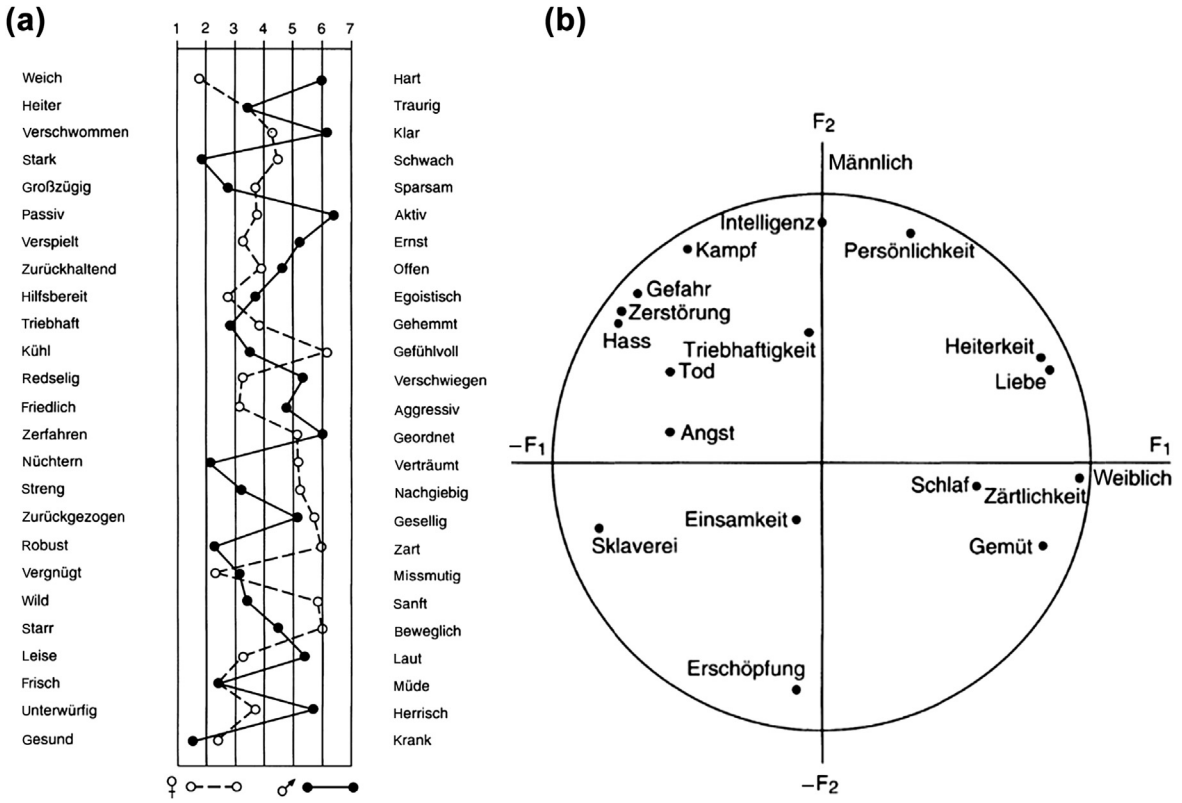

Figure 3 J. Bortz/N. Döring (2006: 186): (a) polarity profiles of concepts ‘masculine’ and ‘feminine,’ according to Hofstätter, P., 1973. Einführung in die Sozialpsychologie. Alfred Kröner Verlag, Stuttgart, p. 259; (b) two-dimensional conceptual system.

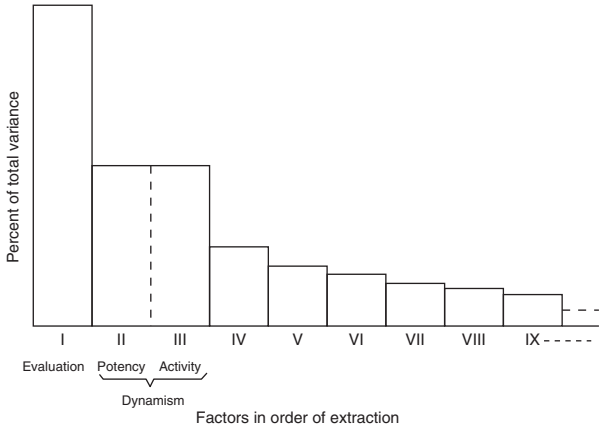

Figure 4 Factor analysis. Extraction of three leading factors: Evaluation, Potency, and Activity (EPA). Reproduced from Osgood, Ch. E., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana, p. 73.

here – with appropriate elaborations designed to bring *meaning* to the force – lay the key to even the most complex human behaviors, including language and social behavior generally" (quoted from Osgood/Tzeng, 1990: xi). Osgood's neobehaviorism can be read as an extension of Hull's approach to psychology. The psychologist Charles I. Mosier (1941) started to use bipolar adjectives in 11-point scales, covering an evaluative dimension (Osgood et al., 1957: 17) (see Chapter 3, Figure 4).

The development of scaling techniques between the 1920s and 1950s was accompanied by considerable efforts in statistical procedures, i.e., the introduction of correlation coefficients (Spearman, Tau, Pearson), factor analysis (Spearman; Thurstone), latent structure analysis, Cronbach's alpha, etc.

Theoretical and Institutional Context

The SD came up during a time of theoretical and methodological struggle within US-American psychology. When Osgood wrote his initial paper on the topic, the leading approaches in US psychology were learning theory and behavior therapy, both in the tradition of behaviorism. In line with its major theoretical conviction that human behavior can be exhaustively studied and explained with reference to its outside manifestations, behaviorist research was focused on observation and the measurement of physical reactions. Starting in the early 1950s, it got increasingly criticized – from neobehaviorists as well as representatives of *Gestalt psychology*, from linguists and philosophers of language. Being himself a neobehaviorist, Osgood was convinced of the general merits of behaviorism but saw at least one major problem: its inability to study mental events, which do not result in clearly observable patterns of behavior. The SD can be read as a methodological attempt to solve this problem.

When Osgood first started to work with the SD, semantics (i.e., the study of meaning in linguistic expressions) was a widespread topic in many areas of scientific engagement. What has later been referred to as the 'linguistic turn' concerned not only linguistics and philosophy of language, but also many other disciplines. From the late-1940s onwards, also psychology developed a strong interest in language (e.g., Skinner, 1957; Thurstone, 1948) in which Osgood was engaged from the very beginning. In 1952, the US Social Science Research Council (SSRC) established a Committee on Linguistics and Psychology (CLP) chaired by John B. Carroll and Osgood until its dissolution in 1961 (see Murray, 1998: 22f.). The committee was part of a program to support behaviorist social science within the SSRC.

At the University of Illinois in Urbana, Osgood was able to work with a research group focused exclusively on his project – the development of a theory and method for the measurement of meaning. He had the opportunity to do calculations with the ILLIAC 1, one of the very first electronic digital computers for nonmilitary use in the United States (see Osgood et al., 1957: i).

Osgood's most important reference point outside of psychology was psycholinguistics, a subfield of linguistics coming up in the late 1930s. With his work in the CLP, he was at the forefront of this discipline, focused on the psychological background of the acquisition and use of language capacities. In the early years of psycholinguistics, its general approach was behaviorist learning theory. From the early 1950s onwards, this focus changed. In 1959, the linguist and philosopher Noam Chomsky published a review of Burrhus F. Skinners' (1957) book on *Verbal Behavior*, where he convincingly showed that the acquisition and use of language cannot be studied from a merely behaviorist perspective. This finding considerably weakened behaviorism and was an important step toward the so-called 'cognitive turn' in psychology. Osgood, however, tried to establish a theoretical and methodological connection between behaviorist psychology and the psycholinguistic discussion of language and meaning. In 1966, when the attack against behaviorist psycholinguistics reached its provisional climax, he remained on the (neo)behaviorist side (Osgood and Tzeng, 1990: 18).

One of Osgood's central arguments is that structures of meaning are culturally universal. When he started to develop the SD, this idea was common sense among many scholars in different disciplines. A lot of psychologists searched for cross-cultural patterns during that time (e.g., Ekman and Friesen, 1971). In the late-1950s, the universality thesis got increasingly contested in linguistics and philosophy of language. Especially the writings of Edward Sapir and Benjamin Lee Whorf supported some kind of linguistic relativism. Their basic argument was that language affects thought and not the other way round. This implies that language cannot be the result of a universal structure of the human mind. It must be the result of contingent natural and social circumstances under which a language develops. The debate on linguistic universalism and relativism was a central issue in linguistics and philosophy of language at that time, a number of prominent scholars being involved. As we will see, Osgood never believed in the relativity thesis and hardly ever considered any arguments in its support. Instead, he dedicated most of his attention during the 1960s and 1970s to an empirical proof for a specific universal structure of affective meaning – the EPA (evaluation, potency, and activity) structure.

Despite of many possible connections, Osgood was not strongly engaged in contemporary philosophical debates on

meaning. In his 1952 and 1957 writings, he discusses several concepts but fairly ignores every approach that does not refer to behavior. His major point of reference is the behaviorist interpretation of Charles S. Peirce's pragmatist semiotics by Charles W. Morris.

Development and Early Career (1940s and 1950s)

The origins of the SD trace back to research on synesthesia – a neural linkage of two or more sensory or cognitive pathways. This linkage can result in the perceptive entanglement of very different entities – such as 'the number two,' 'green,' and 'warm.' With help of a primitive SD (two poles without scale), Theodore F. Karwoski (an early teacher and intellectual guide of Osgood, Osgood/Tzeng, 1990: 3), Henry S. Odbert, and others found out that some 'quasisynesthetic' phenomena are very widely spread. Their findings indicate that many semantic relations are shared by a very high percentage of people, independent from their cultural background (e.g., large–loud, near–fast, bright–happy, bad–dark, etc.). In the early 1940s, Osgood worked on synesthetic thinking together with Karwoski (Karwoski et al., 1942) and started to develop a scaling technique together with his friend and colleague Ross Stagner (Osgood and Stagner, 1941; Stagner and Osgood, 1941). In 1946, Stagner and Osgood used this technique for a study on social stereotypes. Previous studies had suggested a continuum between the poles, so they introduced a seven-step scale with a neutral zero point. In order to find out more about the forces behind the association game, they conducted a large test with 100 college students and 1000 items. They applied factor analysis and found three central factors with maximal differentiating power: evaluation, strength (later: potency), and activity (**Figure 4**). In the dimension of evaluation, adjective pairs like 'good–bad' and 'beautiful–ugly' showed high factor loadings. Pairs like 'large–small' and 'strong–weak' clustered in a separate dimension, which Stagner and Osgood called strength. Pairs like 'fast–slow' and 'active–passive' were assigned to the factor of activity. Together, the three factors made up the so-called EPA structure. In the following years, Osgood and his colleagues found many similar factor structures in other projects and started to assume that EPA represents the 'general dimensions of meaning.'

In 1952, Osgood introduced the SD as a tool for objective measurement of meaning. He was convinced that meaning is an important issue for behaviorist psychology but could not be captured with the methodological tools of the time. He discussed several theories of meaning and ended up with the following definition: "Words represent things because they produce some replica of the actual behavior toward these things, as a mediation process. This is the crucial identification, the mechanism that ties particular signs to particular stimulus-objects and not to others" (204). This conception is clearly neobehaviorist. It admits that cognitive processes affect human behavior but – apart from that – stays within the framework of behaviorist thinking. According to this definition, the central semantic relation ties a sign to an object – and the object stimulates a certain behavior. If the object is not at hand, the sign produces a cognitive replication – not of the object, but of the behavior toward the object – which we call meaning. In other words, the meaning of a sign is a *mediator* between the sign and a particular stimulus–response. For Osgood's purpose, it is crucial that the cognitive replication of a sign does not exhaust in its denotative meaning but always carries additional associations – its *connotative meaning*. In his widely received textbook on *Methods and Theory in Experimental Psychology* (1953), he explained why this kind of cognitive response is most likely to be the carrier of meaning: it is easy to perform, requires little energy, and therefore needs only little reinforcement to be learned. In the SD, Osgood exploited this idea methodologically, assuming that every mediational meaning response operates within a universal semantic space, which his scales claim to represent.

After a discussion of several methodological possibilities for the measurement of meaning, Osgood advocated a scaling approach. Scaling, he argued, provides for "controlled association in which the nature of the association is specified by definition of the scales (…) but the direction and intensity of association is unspecified" (222). Finally, he discussed the origins, logic, and merits of the SD and called for further research on its objectivity, reliability, validity, sensitivity, comparability, and utility (230 ff.). As will be shown, a lot of papers answered to that call. Especially between 1964 and 1974, the number of publications on the merits and limits of the SD were enormously high (see **Figure 5**).

In a 1954 case report for the *Journal of Abnormal and Social Psychology* (*A Blind Analysis of a Case of Multiple Personality Using the Semantic Differential*), Osgood and Zella Luria applied the SD to data of a multiple personality case. The material they worked with was data on patients with multiple personalities. The case of 'Eve' was very prominent in 1950s psychology (several teams of researchers worked on the case and in 1957, the story of 'Eve' is even be made into a film), and also this paper received wide attention. Osgood and Luria compared the semantic structure of Eves' three personalities, based on six SD testings with 15 concepts and 10 scales. The research design was very experimental and established a new area of application for the SD: clinical personality and psychotherapy research. Osgood did not systematically follow this line of research, but he used the data and his major findings in his 1957 book and came back to the case in the 1970s (see Osgood et al., 1976).

Until today, the most important work of reference for the SD is *The Measurement of Meaning* – a book published in 1957 by Osgood, George J. Suci, and Percy H. Tannenbaum. The volume summed up the central insights from the first years of research on the SD and can be read as a systematic introduction to its theory, methodology, application, and evaluation. Towards the end, the authors addressed some of the most prominent points of criticism and gave an outlook to important tasks for future SD research. The book received good reviews and reached the status of an academic bestseller.

Prominent Variations of the SD in the German-Speaking Countries

Although Osgood counts as the founding father and uncontested expert on the SD, other scholars developed prominent variations. Especially in the German-speaking countries, two other concepts received a lot of attention: The

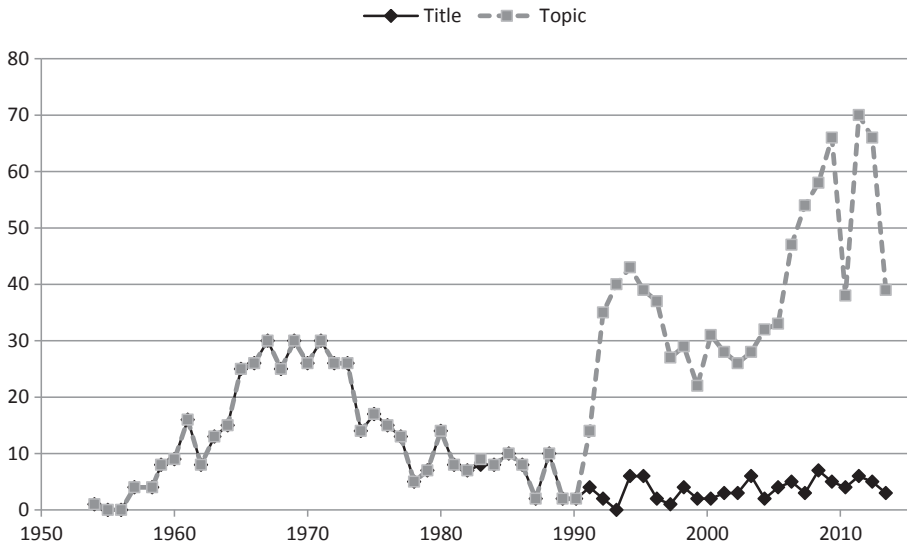

Figure 5 Number of SSCI articles on the SD, 1950–2013. Data source: Social Science Citation Index ('Semantic Differential' in title (black); 'Semantic Differential' as a topic (gray).

Polaritätenprofil (or 'polarity profile') (Hofstätter, 1955/56) and the Eindrucksdifferenzial (Ertel, 1965).

The polarity profile was developed by the Austrian Social Psychologist Peter Robert Hofstätter in the course of the late-1950s. During his stay at the *Massachusetts Institute of Technology* and *The Catholic University of America*, Hofstätter got in touch with Osgood's writings and started to set up new bipolar adjective lists. Himself a researcher of groups, roles, and stereotypes, Hofstätter focused on group-specific differences toward meaning. The most famous example of a polarity profile, according to Hofstätter, included 25 contrasting adjective pairs measuring attributions to males and females (see **Figure 3**). It rated men's and women's qualities from 'soft' to 'tough' or 'submissive' to 'authoritative,' each on a seven-point scale. With this polarity profile, Hofstätter contributed to the research on prejudice and stereotypes as well as the sense of self and awareness of others (1973: 366). Analytically, Hofstätter started with a correlation of respondents' answers within and between polarity profiles. This correlation suggested a dimensionality, which he proved with help of factor analysis in a second step. Hofstätter's analysis resulted in two to three dominant factors as well, but he assessed the maximum four-dimensional system of connotative meanings according to Osgood as being 'primitive' (1973: 264). Hofstätter developed a three-dimensional model of personality including several factors beyond the EPA structure. According to him, at least 10–15 factors would be needed to capture the individual dimensionality of personality (Hofstätter, 1973). Furthermore, he assumed the semantic space to be historically and culturally grown (1963), not universal to all human beings, as Osgood had concluded.

A second influential variation of the SD is the *Eindrucks differential* (ED), developed by the psychologist S. Ertel in 1964/65. It contains 18 bipolar adjective couples and, in accordance with Osgood's EPA structure, assigns six adjective pairs to each dimension of the semantic space. In contrast to Osgood and Hofstätter, Ertel emphasized the *emotional* component of meaning. While Osgood had reduced affective meaning to the dimension of Evaluation, Ertel assumed emotionality to be central for the whole of the semantic space – including the dimension of potency and activity. Ertel agreed with the EPA structure of the semantic space, however develops an alternative interpretation.

One of the major differences between the US-American and European scientific communities of this time can thus be framed as a matter of style with significant methodological consequences. Most US scholars relied on Osgood's tested SD scales, whereas European scholars selected feature lists far more intuitively, in the manner of Hofstätter and Ertel (Schäfer and Fuchs, 1975: 119). In the European applications and validations of the SD technique, one can find references to Osgood, Hofstätter, and Ertel and, beyond, to national scholars who apply the SD in their 'own' cultural context.

Prosperity and Critique (1960s)

An SSCI count (Social Science Citation Index.) of articles concerned with the SD indicates an impressive prosperity during the 1960s and 1970s. **Figure 5** illustrates the results from a broad search on 'semantic differential' in all registered works of the index as well as a limited search in the contributions' titles, covering the time period from the 1954 to 2013. The vast majority of results are articles in journals, i.e., *The Journal of Psychology*, *Psychological Reports*, *Public Opinion Quaterly*, *Educational and Psycholgical Measurment;* occasional contributions were meeting abstracts, notes, and book reviews. Before the 1980s, almost all SD related articles addressed the SD in title, indicating a discussion of the technique as such. A close look into the papers shows that, most of them did not only apply the method but also dealt with its methodological background and/or question its empirical validation. In the 1990s, the number of articles dealing with SD tremendously increased, yet most of them merely applied the SD technique

rather than questioning its methodological implications. This increase is (in part) due to the growing number of countries and journals included in the Social Science Citation Index, in particular from the 1990s on.

Already at the beginning of the 1960s, the SD was widely recognized, in particular among US psychologists. Within this period, Osgood himself continued to verify and advance the technique as documented in his *Studies on the generality of affective meaning systems* (1962), *Comparative Study of Cultures* (1964), and the *Whys and Wherefores of E, P, and A* (1969). Therefore, the center of SD research remained at the University of Illinois, where Osgood, Tannenbaum, Suci, Luria, M.S. Miron, J. Lyle, and others further developed the theoretical and methodological framework of SD and practically applied it in their fields. Various articles on the SD were written at other US-American Universities, i.e., Washington, Michigan, Connecticut, Pennsylvania, Arizona, Berkeley, etc., covering different subfields of Psychology. In 1969, Osgood and Snider edited a sourcebook including nearly 1500 concepts for SD studies, which was also widely used.

The SD was discussed in various other fields during the 1960s, especially in education, marketing, communication, and theater studies. In Marketing Research, William A. Mindak was one of the first who dealt with *Fitting the Semantic Differential to the Marketing Problem* (1961). He used the SD to measure consumers' reactions to brands, products, companies, and their images. Robert F. Kelly and Ronald Stephenson (1967) in their approach focus on the analysis of consumers' decisions as support for retailers and consumer service institutions. In the field of theater studies, the major proponent of the SD technique was Raymond G. Smith who developed his own SD scales throughout the 1960s. Referring to Osgood's scales, he introduced 'speech correction concepts' (1962), tried to validate SD and tested its factorial structure in this specific field of research.

According to the SSCI, the most widely recognized contributions in the 1960s were Osgood's *Semantic Differential Technique in the Comparative-Study of Cultures* (1964) published in the American Anthropologist and David R.J. Heise's methodological discussion *Some Methodological Issues in Semantic Differential Research* (1969) published in the Psychological Bulletin.

Osgood (1964) is a central text for the proof of the universal (cross-cultural) validity of the EPA structure of affective meaning. By the end of the 1950s, Osgood claimed to have proven the generality of EPA for American English, but many critics doubted its transferability to other cultures and languages. In his 1964 paper, Osgood compared the United States with 16 other countries and language groups. From that time on, he established strong international networks, and the SD research teams around him became more and more international. The results of these studies show congruent results among the language groups in terms of the three dimensions – Evaluation, Potency, and Activity – which led Osgood to confirm his assumption of a generalized and universally valid EPA structure. In 1965, Osgood and the Japanese psychologist Yasumasa Tanaka compared three groups of people with different linguistic and cultural background who rated 24 *perceptual* signs by SD scales. Factor analysis brought about four different dimensions: dynamism, evaluation, warmth, and weight. This result was stable across all cultural groups and encouraged Osgood to engage in further EPA-related projects throughout the 1970s.

Critical methodological discussion and empirical testing of the SD in practical applications trace back to the 1950s and continued throughout the prosperous period. Still, they were particularly high during the 1960s. In his influential methodological discussion, David R.J. Heise (1969) referred to more than 1000 works on the SD, most of them dealing with methodological issues. Heise conceded that the SD is a useful standard method for social psychological research (1969: 421) but systematically highlighted the leading methodological questions and results of empirical tests carried out in the 1950s and 1960s: Among them, there were (1) the metric assumptions of SD scales (equal intervals between bipolar adjectives and the zero point), which Heise overall considered to be inadequate for some applications dependent on the concrete research work; (2) the different explanations for the sources of variance in ratings found by various scholars. In this context (a) bias errors due to scale-checking styles and social desirability were tested and (b) random errors were proven by means of reliability analyses, i.e., test–retest reliabilities. According to Daniel Kahneman – who wrote a dissertation on the SD in 1961 – variance is often partitioned as follows: "one-tenth due to subject-scale interaction, that is due to differences between subjects in the use of scales, one quarter due to bias and/or deviations of subjects' true scores from the population true scores, one quarter due to momentary deviations of subjects from their own true scores and two-fifths due to random error" (Kahneman quoted from Heise, 1969: 412); (3) Heise dealt with the structure of SD scales and stressed the empirically proven prominence of the three dimensions in semantic space: Evaluation, Potency, and Activity (EPA). Critique toward the EPA structure, in particular, included the number of factors and the number of concepts, which both, due to several empirical tests, were said to be underrated by Osgood. Some scholars found four or five factors, others discovered three subdimensions within Evaluation (functional, hedonic, and ethical value) (Heise, 1969: 413). Other scholars, in turn, found only two dimensions in their research. Moreover, it was individual differences, which in many cases counted for differences in the factorial results; a fact, which according to Heise deserved more attention. Heise concluded that Osgood's three dimensions were highly relevant, though further research would be needed to investigate the size and character of the semantic space; and (4) Heise referred to the assumption of concept–scale interactions, based on the different relevance of varying concepts or affected by the 'stimulus environment' (418) due to which he called for further validation.

Consolidation and Internationalization (Early 1970s)

Throughout the 1970s, psychology's leading paradigm gradually shifted from behaviorism to cognitivism. Although cognitive aspects had already been included in neobehavioral frameworks (i.e., in Edward C. Tolman's learning theory), it

took cognitivism many years to take over. As mentioned earlier, SD technique's methodological background is rooted in the neobehaviorist tradition: it includes behaviorist ideas but clearly refuses classical behaviorism. The SD survived the 'cognitive revolution' and became a standard technique of scaling. In the early 1970s, its prominence further increased, methodological discussion continued as well as concrete applications, reliability, and validity testing. Furthermore, it got increasing attention outside psychology and outside of the United States: The SSCI shows articles from Venezuela, Brazil, Italy, Spain, Germany, Russia, Japan, India, Czechoslovakia, etc., thus both, an internationalization and increasing differentiation in various fields of research took place.

One of the most recognized articles on the SD in the 1970s was published by Kalman J. Kaplan (1972) and dealt with the *Ambivalence-Indifference Problem in Attitude Theory and Measurement*. Kaplan modified the SD in order to draw a distinction between ambivalent and indifferent responses. While Osgood understood the midpoint of SD scales as being *neutral*, based on the assumption of reciprocal antagonist bipolarity, Kaplan raised the question of *ambivalent* responses. In his modification, he suggested to separate SD scales into a positive- and a negative component-scale. "Three independent measures are thus generated by this technique: first, the usual bipolar attitude (A), second, its positive or "liking" component (Ap), and third, its negative or "disliking" component (An)" (366). The different components thus could be correlated. Based on the component model, Kaplan furthermore developed a model, which allows to seperate the 'total effect,' 'ambivalence,' and 'polarization' of SD scales. Kaplan's considerations on scalings' midpoints finally turned out to be of great importance for scaling techniques in general.

Osgood himself had a clear focus during the 1970s: he wanted to proof the universality of the EPA structure. He continued and expanded his works on *Cross-Cultural Universals of Affective Meaning* (1975), published together with psycholinguists and his (former) research affiliates M.S. Miron and B. May. Their enthusiasm for cross-cultural comparisons can be understood from different perspectives: In the 1970s, worldwide networking in the field of research but also in politics and other social areas started anew. Many disciplines were interested in differences and similarities between cultures in order to help preventing further conflict. In this line, scholars were also fascinated by the cross-cultural similarity of human beings after World War II, as they provided a plausible justification of the 'new' egalitarian views. Last but not least, the framework requirements of technical progress and IT were a highly important prerequisite to manage the huge amounts of country comparative data in the 1970s.

In their 1975 book, Osgood, Miron, and May stressed the fact that "cross-national research, by its very nature, must be entangled in the cultural and political matrix of the world at the time" (7). Osgood, Miron, and May broadly discussed the conceptual and methodological limitations of their own cross-national research in the field of cross-linguistics. They reflected on the 'westernization' of research meaning – i.e., the application of Western concepts to the analysis of non-Western countries. To deal with this bias, they asked native speakers from each language/culture community to translate the scales, initiated deep-going pretesting procedures for each country, and made efforts in sampling technique and data collection to solve the problems of within-country representativeness and between-country comparability. More than 20 countries (e.g., Afghanistan, Belgium, Costa Rica, Delhi, Germany, Hungary, Italy, Iran, Japan, Malaysia, Mexico, Sweden, Turkey, the United States, etc. (1975: 194)) participated in this new study, Indo-European language groups being overrepresented. The superordinate goal of the project was to prove the cross-cultural validity of the EPA structure. Osgood, Miron, and May started with systematically controlled data from the English-speaking countries and included the non-English speaking countries in a second step. As their findings clearly supported the idea of a general EPA structure, they assumed that "the EPA-system found for English-speaking Americans is linguistically and culturally universal for the human species" (1975:112).

Apart from this, Osgood advanced another sideline of his early work: the application of the SD as a tool for clinical psychology and therapy of schizophrenia. In 1976, he published a 'continuation' of the blind analysis from 1954 together with Luria and S.W. Smith. On behalf of the therapist R.F. Jeans, a new case of multiple personality, 'Evelyn' was systematically compared with the 'three faces of Eve,' in order to improve the SD as a tool for clinical purposes. The 15 concepts used for the case of 'Eve' are added on by 10 further concepts, specific to the case of 'Evelyn' who, according to Jeans, inherited three personalities, 'Mary,' 'Gina,' and 'Evelyn.' Osgood, Luria, and Smith possessed basic information on the patient's family background and anamnesis. As Eve, Evelyn rated herself and her family members along Osgood's SD scales, with different answers for each personality within the three dimensions of the semantic space: evaluation, potency, and activity. Osgood, Luria, and Smith systematically analyzed the specific response patterns of each personality of 'Evelyn' and compared it to the case of 'Eve.' They discussed the results and interpretations of the blind analysis with the therapist Jeans, who agreed with their findings. He used them for his therapeutic work and even included them in one of his books on diagnosis and therapy of multiple personalities (1976: 274).

In 1975, the German sociopsychologist Reinhold Bergler published an anthology in which he, Brigitte Diehl, Albert Fuchs, Reinhard Oppermann, and Bernd Schäfer commented on the status quo of international research on the SD and discussed its methodological foundation and fields of practical applications. Bergler questioned the uncritical application, lack of validation, and theoretical deepening and demanded to make the method itself a central object of investigation (1975: 11ff.). This demand sums up the superior concern of the anthology. In particular, he called for systematic intra- and interindividual analysis as well as investigations considering a longitudinal perspective in order to capture group differences and meanings' changes over time (1975: 15). He stressed the

importance of individually selecting the investigated population and their specific associations, and consequently called for 'obligatory' pretesting methods (1975: 23). In this line, Bergler and his colleagues criticized the universal application of SD by both Osgood and Hofstätter. Among others, Bergler, Schäfer, and Fuchs, sharply questioned the reduction of the semantic space to three dimensions and stressed the existence of intercultural differences in language use. Given the demand to improve SD scaling, Schäfer and Fuchs, on the one hand, called for more extended testing of SD scales' selectivity, to introduce coefficients in order to prove intraconcept variability. On the other hand, they suggested tests for the selectivity of respondents' ratings (1975: 131ff.).

Differentiation and Pluralization of Applications (Late 1970s, 1980s and 1990s)

In the late-1970s, methodological interest in the SD decreased. It became part of the standard repertoire in a lot of disciplines, still predominantly in psychology, but did – as a method – not receive a lot of attention anymore. One reason might be that all relevant arguments had been made and there was nothing left to say. The concept proved to be useful for some research contexts, documented in a great variety of different applications of the SD, however inappropriate for others. Osgood's argument for the universality of the EPA structure did not convince his colleagues and therefore seemed to be forgotten soon. Scholars continued to apply the SD technique, independent from its methodological, historical, and theoretical background. Another reason might be the growing interest in cognitive psychology, which absorbed the attention of the community and outdated some general convictions of Osgood's work. Beyond, new scaling techniques (e.g., Rasch scales and Mokken scales) were established, which attracted various scholars and confronted them with new questions.

Summary and Today's Significance

After its introduction in the 1950s, the SD received enormous attention among psychologists and other social and behavioral scientists. Especially during the 1960s, its methodological details and possible fields of application were broadly and critically discussed. In 1969, Heise reported more than 1000 papers on SD – a huge amount given the publication rates of that time. A good deal of this attention might be due to its promise to 'repair' a vital defect of behaviorist thinking, its smart use of factor analysis (a fashionable method at that time), and its connectivity to prominent debates on meaning in linguistics and philosophy of language. Especially Osgood's claim for a universal structure of meaning in the EPA structure pointed to the heart of the debate on semantic relativism.

In the early 1970s, behaviorism lost its position as the leading paradigm of psychology whereas cognitivism gained increased attention, and the interest in the SD decreased. Still, SD joined the group of standard scaling techniques during the 1980s and 1990s. Today, it is applied by many disciplines worldwide, such as psychology, sociology, linguistics, marketing research, etc.

See also: Attitude Measurement; Factor Analysis and Latent Structure Analysis: Overview; Measurement Theory: History and Philosophy; Psycholinguistics: Overview; Reliability and Validity Assessment: New Approaches; Reliability: Measurement; Scaling and Classification in Social Measurement.

Bibliography

Bergler, R. (Ed.), 1975. Das Eindrucksdifferential. Theorie und Technik. Huber, Bern.
Bortz, J., Döring, N., 2006. Evaluation und Forschungsmethoden für Human- und Sozialwissenschaftler. Springer, Heidelberg.
Bergler, R., Diehl, B., Fuchs, A., Oppermann, R., Schäfer, B., 1975. Das Eindrucksdifferential. Theorie und Technik. Huber, Bern, pp. 119–137.
Chomsky, N., 1959. A review of B.F. Skinner's verbal behavior. In Language 35 (1), 26–58.
Edwards, A.L., Kilpatrick, F.P., 1948. Scale analysis and the measurement of social attitudes. Psychometrika 13 (2), 99–114.
Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17 (2), 124–129.
Ertel, S., 1964. Die emotionale Natur des "semantischen" Raumes. Psychologische Forschung 28, 1–32.
Ertel, S., 1965. Standardisierung eines Eindrucksdifferentials. Zeitschrift für experimentelle und angewandte Psychologie 12, 22–58.
Glaze, J.A., 1928. The association value of non-sense syllables. Pedagogical Seminary and Journal of Genetic Psychology 35, 255–269.
Heise, D.R., 1969. Some methodological issues in semantic differential research. Psychological Bulletin 72, 406–422.
Hofstätter, P., 1955/56. Dimensionen des mimischen Ausdrucks. Zeitschrift für experimentelle und angewandte Psychologie 3, 505–529.
Hofstätter, P., 1963. Differentielle Psychologie. Alfred Kröner Verlag, Stuttgart.
Hofstätter, P., 1973. Einführung in die Sozialpsychologie. Alfred Kröner Verlag, Stuttgart.
Hull, C.L., 1933. The meaningfulness of 320 selected nonsense syllables. American Journal of Psychology 45, 730–734.
Kahnemann, D., 1961. An Analytical Model of the Semantic Differential (Unpublished dissertation). University of California, Berkeley.
Kaplan, K.J., 1972. On the ambivalence-indifference problem in attitude theory and measurement: a suggested modification of the semantic differential technique. Psychological Bulletin 77 (5), 361–372.
Karwoski, T.F., Odbert, H.S., Osgood, Ch. E., 1942. Studies in synesthetic thinking: II. The role of form in visual responses to music. Journal of General Psychology 26, 199–222.
Kelly, R.F., Stephenson, R., 1967. The semantic differential: an information source for designing retail patronage appeals. Journal of Marketing 31 (4), 43–47.
Likert, R., 1932. Technique for the measurement of attitudes. Archives of Psychology 140, 5–55.
Mindak, W.A., 1961. Fitting the semantic differential to the marketing problem. Journal of Marketing 25 (4), 28–33.
Mosier, C.I., 1941. A psychometric study of meaning. Journal of Social Psychology 13 (1), 123–140.
Murray, S.O., 1998. American Sociolinguistics. Theorists and Theory Groups. John Benjamins B.V.
Osgood, Ch. E., Stagner, R., 1941. Analysis of a prestige frame of reference by a gradient technique. Journal of Applied Psychology 1941 (25), 275–290.
Osgood, Ch. E., 1952. The nature and measurement of meaning. Psychological Bulletin 49, 197–237.
Osgood, Ch. E., 1953. Method and Theory in Experimental Psychology. Oxford University Press, New York.
Osgood, Ch. E., Luria, Z., 1954. A blind analysis of a case of triple personality using the semantic differential. Journal of Abnormal and Social Psychology 49, 579–591.
Osgood, Ch. E., Suci, G., Tannenbaum, P., 1957. The Measurement of Meaning. University of Illinois Press, Urbana.

Osgood, Ch. E., 1962. Studies on the generality of affective meaning systems. American Psychologist 17, 10–28.
Osgood, Ch. E., 1964. Semantic differential technique in the comparative study of cultures. American Anthropologist 66, 171–200.
Osgood, Ch. E., Snider, J.G., 1969. Semantic Differential Technique: A Sourcebook. Aldine, Chicago.
Osgood, Ch. E., 1969. On the whys and wherefores of E, P, and A. Journal of Personality and Social Psychology 12, 194–199.
Osgood, Ch. E., May, W.H., Miron, M.S., 1975. Cross-cultural Universals of Affective Meaning. University of Illinois Press, Urbana.
Osgood, Ch. E., Jeans, R.F., Luria, Z., Smith, S.W., 1976. The three faces of Evelyn: a case report. Journal of Abnormal Psychology 1976 (85), 247–286.
Osgood, Ch. E., Tzeng, Oliver (Eds.), 1990. Language, Meaning, and Culture: The Selected Papers of C.E. Osgood. Praeger Publishers.
Schäfer, B., Fuchs, A., 1975. Kriterien und Techniken der Merkmalsselektion bei der Konstruktion eines Eindrucksdifferentials. Bergler, R. (Hg.), Das Eindrucksdifferential. Huber, Bern.
Skinner, B.F., 1957. Verbal Behavior. Acton, Copley Publishing Group.
Smith, R.G., 1962. A semantic differential for speech correction concepts. Speech Monographs 29, 1.
Stagner, R., Osgood, Ch. E., 1941. An experimental analysis of a nationalistic frame of reference. Journal of Social Psychology 1941 (14), 389–401.
Tanaka, Y., Oyama, T., Osgood, Ch. E., 1965. A cross-culture and cross-concept study of the generality of semantic spaces. Journal of Verbal Learning and Verbal Behavior 1963 (2), 392–405.
Thurstone, L.L., 1948. Psychological implications of factor analysis. American Psychologist 3 (9), 402.

Semantic Memory

Julie S Snowden, Manchester Academic Health Sciences Centre, Institute of Brain, Behaviour and Mental Health, University of Manchester, UK; and Greater Manchester Neuroscience Centre, Salford Royal NHS Foundation Trust, Salford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Semantic memory refers to our conceptual knowledge of the world. Understanding of semantic memory has come from several sources: cognitive studies of healthy individuals, computational modeling, patients with disordered semantic memory due to brain disease, and brain imaging and stimulation. The converging evidence indicates that semantic memory involves distributed brain networks, which, at least in part, are linked to the sensory processes involved in perception, action, and language. Whether there is also representation in amodal format remains an area of contention. Knowledge of the world, beyond word and object meanings, is a challenge for future studies of semantic memory.

What is Semantic Memory?

Semantic memory represents a person's conceptual knowledge and understanding of the world. It includes knowledge of the meaning of words, objects, and other percepts experienced directly through the senses as well as the more abstract encyclopedia of facts that one acquires over a lifetime. Knowing that the word 'water' refers to a liquid that is equivalent to 'acqua' in Italian and 'agua' in Spanish is part of semantic memory. So too is knowing that water quenches the thirst, is used for making tea, is slippery when frozen, and can scald when boiled, and that it is possible to drown in water if you cannot swim. Semantic memory encompasses all that one knows. It is crucial in enabling a person to function and interact with the outside world.

Historical Background: Development of Knowledge about Semantic Memory

Semantic-Episodic Memory Distinction

The delineation of semantic memory as a distinct form of memory is credited principally to Endel Tulving who distinguished it from episodic memory (Tulving, 1972). Semantic memory was defined as information that is shared between people within a culture and not tied to a specific temporal or spatial context. By contrast, episodic memory was construed as memory for events defined in terms of time and place and unique to an individual person. Thus, knowing that London is the name of the city where the 2012 Olympic Games were held is part of semantic memory, whereas remembering a visit to London to attend an Olympic event constitutes episodic memory.

Early Language Models of Semantic Processing

Tulving originally construed semantic memory in terms of a mental thesaurus, and hence essentially dealing with linguistic concepts. The language-related focus reflected much of the early work on semantic memory, and indeed, 'semantics' remains a central issue in understanding language.

Hierarchical network model

The term 'semantic memory' had been used prior to 1972 by computer scientist Ross Quillian, who was attempting to program machines to understand language. A central question was how meaning is represented. In his hierarchical model (Collins and Quillian, 1969), concepts were represented as nodes in a network, each associated with a number of properties. In the case of the concepts canary, bird, and animal, for example, the subordinate level concept 'canary' is more closely linked to the superordinate concept 'bird' than the overarching superordinate 'animal.' The finding that it takes longer to verify the sentence 'A canary is an animal' than 'A canary is a bird' was taken as support for the hierarchical account. However, by analogy, the model ought also to predict that it would take longer to verify 'A dog is an animal' than 'A dog is a mammal.' Yet, this is not so. Moreover, the model does not explain why it is easier to verify prototypical examples of a category, such as 'A canary is a bird,' compared to atypical examples, such as 'A penguin is a bird.'

Spreading Activation Model

The 'spreading activation' theory of semantic processing represented a less rigidly structured revision of Quillian's network model. It incorporated the idea of semantic distance, highly related concepts being located closer together than unrelated concepts. The model also allowed different types of links, including class membership 'X is a Y,' attributes 'X has a Y' as well as negative information 'X is not,' 'X cannot Y.' The model proposed that when a concept is processed, activation spreads out along the paths of the network, the activation from the source node becoming progressively more attenuated with outward spread.

Feature Comparison Models

An alternative to the spreading activation model was the feature comparison model, of which major proponents were Smith et al. (1974). The model assumed that the meaning of a concept is determined by a set of defining and characteristic features, and that verification of a sentence such as 'A canary is a bird' involves comparing the features of 'canary' with those of 'bird.' The model correctly predicts that prototypical

examples of a category will be verified faster than atypical examples because of the greater overlap in features. A limitation is that the model does not readily account for differences between the sentences 'A canary is a bird' and 'A bird is a canary.'

These early models share the common principle that concepts contain component properties or features, a notion that underpins contemporary thinking about semantic memory.

Semantic Memory Is More than Words

Schemas and Scripts

In the 1970s, a number of scientists including Marvin Minsky, David Rumelhart, and Roger Schank argued that semantic memory must encompass more than simple word concepts and adopted approaches to comprehension that involved larger chunks of information, referred to as frames, schema, or scripts. Schank and Abelson (1977) proposed that people build up schema/scripts that represent commonly experienced activities and events. Thus, a schema for buying goods in a shop would incorporate the notion of exchanging goods for money. A schema for going to a restaurant would include the notion of choosing from the menu, ordering from a waiter, and so on. Schema varies according to culture. In some cultures, the purchase of goods would incorporate the notion of bartering, whereas in other cultures it would not. The notion of schema based on culturally relevant experience enables people to understand sentences and draw appropriate inferences when information is not explicitly stated. For example, the sentence 'John dropped into the Queen's Arms on his way from work' is likely to be interpreted by a British listener as meaning that John went to a pub called the Queen's Arms where he had a drink and then he went home afterward (not that he went to Buckingham Palace in London and fell into the arms of the Queen). Schema models of semantic memory have been explored through computer modeling. They continue to have relevance, discussed later.

When Semantic Memory Breaks Down

Semantic Dementia

Early studies of semantic memory came from the fields of computer science and experimental psychology, and human studies involved healthy volunteers. A crucial development came from an entirely different quarter. In 1975, Elizabeth Warrington, neuropsychologist working at the National Hospital, London, described a series of patients who had difficulty recognizing objects as the result of degenerative brain disease (Warrington, 1975). A distinction had long been drawn between two forms of disordered visual object recognition: apperceptive agnosia, characterized by a difficulty in integrating visual sensory information into a coherent percept, and associative agnosia, characterized by loss of understanding of percepts that are themselves normal. Warrington's patients showed the associative type. Warrington appreciated, however, that the patients' loss of understanding extended beyond object recognition; they also had difficulty understanding object names. That is, they could no longer remember the meaning of words and objects. Yet, those patients were not amnesic in the traditional sense. Warrington characterized the patients' disorder as a 'selective impairment of semantic memory.' Since then, the multimodal yet circumscribed nature of patients' disorder has been widely reported (e.g., Hodges and Patterson, 2007; Snowden et al., 1996). Snowden introduced the term 'semantic dementia' to capture the widespread nature of the semantic breakdown.

Impact of Semantic Loss

In semantic dementia, words are gradually lost from a person's vocabulary and the conceptual distinctions between related words are progressively eroded. People with semantic dementia make semantic errors in naming, such as 'dog' for lion and 'banana' for lemon, because they can no longer discriminate conceptually between a dog and a lion and between a banana and lemon. They show loss of understanding of words spoken by others and may ask what words mean (e.g., 'lion, what's lion?'). Communication becomes severely compromised.

People with semantic dementia may misuse objects because, although they perceive objects normally, they no longer understand what each item is for. They may be perplexed by environmental sounds (e.g., a clap of thunder). They may attempt to eat inedible objects because of inability to discriminate edible from inedible substances. They may ignore danger signals such as the smell of burning, the taste of stale food, and the rapid approach of a fast car because those signals carry no significance. The consequences underscore the fundamental practical role of semantic memory.

There are, nevertheless, areas of preservation. Despite failing to recognize goods in the supermarket, people with semantic dementia may remember the aisle and shelf from where they obtained goods on a prior supermarket visit. They may not recognize the face or name of their hairdresser, yet have no difficulty remembering the time of their appointment at the hairdresser's and how to get there. Such apparently surprising disparities reinforce Tulving's theoretical distinction between semantic (conceptual) and episodic (event) memory.

Breakdown in semantic memory is not restricted to semantic dementia. It may occur in association with a variety of brain disorders, including Alzheimer's disease, herpes simplex encephalitis, and temporal lobe epilepsy. However, semantic dementia provides the most striking and purest example of semantic breakdown, and therefore constitutes a natural model for understanding semantic memory. Much of what is currently known about semantic memory has come from patients with semantic memory disorders.

Technology Provides a Window on to the Neural Basis of Semantic Memory

In recent years, advances in brain imaging have provided a complementary source of information to patient data, opening up the possibility of exploring directly the anatomical basis of semantic memory. Converging evidence emphasizes the importance of the temporal lobes. Disordered semantic memory due to a variety of etiologies is associated with damage to the temporal lobes. Functional imaging studies, using positron emission tomography and functional magnetic

resonance imaging, show activation of temporal regions when healthy volunteers carry out semantic tasks, and transcranial magnetic stimulation of temporal regions disrupts the efficiency of semantic processing.

The precise regions within the temporal lobe that are most critical have been subject to debate. Some have argued for the importance of the anterior temporal lobes (Visser et al., 2009), not least because of the severe anterior temporal lobe atrophy associated with semantic dementia (Mummery et al., 2000). Others have implicated posterior temporal regions (Martin, 2007). The relative contribution of these brain regions is directly relevant to contemporary models of semantic memory, considered below. Of importance too is the question of the role of the two hemispheres (see Section Modalities of Knowledge) and of the brain regions outside the temporal lobe (see Section Categories of Knowledge).

Theoretical Questions That Have Informed Current Thinking about Semantic Memory

The following are some key neuropsychological questions that have shaped our current state of knowledge about the organization of semantic memory.

Is Superordinate Category Knowledge More Resistant to Disruption than Subordinate Knowledge?

An early question, directly pertinent to Quillian's hierarchically organized model of semantic memory, was whether superordinate category knowledge is more resistant to disruption from brain disease than lower order levels of knowledge, representing information about properties or attributes. It is certainly true that semantic dementia patients are able to classify words or pictures according to their broad class (e.g., animals versus objects; edible versus nonedible) better than they can categorize within a class (e.g., cats versus dogs; apples versus oranges). Nevertheless, it does not necessarily follow that breakdown of semantic memory follows a systematic path from most specific to most general. The strict hierarchical model would assume that access from a lower to higher level of the hierarchy requires traversing intervening levels. A robin should not be correctly classified as an animal if it cannot first be classified as a bird. Yet, patients with semantic disorder do indeed classify robins as animals, without knowing that they are birds. Moreover, they may recognize their own object without recognizing alternative versions of the same object, which would not be predicted if generic category knowledge were preserved.

An alternative argument, to account for the superior ability to classify across compared to within a class, is that if knowledge of some semantic features is lost then the remaining impoverished information will be better able to support broad (e.g., categorical) rather than fine-grained (i.e., exemplar) distinctions. This is particularly so because the features most likely to be retained are those that occur with greatest frequency, being shared by most members of the category (e.g., 'has legs' for animals), whereas those most likely to be lost are those least common and shared by fewest members of the category (e.g., 'has stripes'). Hence, impoverished feature knowledge may be sufficient to allow a tiger to be identified as an animal as opposed to a fruit but not to be distinguished from a dog. Such an account, which is in tune with current thinking, reflects a shift in emphasis away from organizational hierarchy per se toward the types of feature information required for specific semantic tasks.

Modalities of Knowledge: Can They Break Down Selectively?

Semantic dementia affects all sensory modalities. It impairs the ability to understand the meaning of words, faces, objects, tactile stimuli, nonverbal environmental sounds, tastes, and smells. An important question is the relationship between modalities. Does decline inevitably occur in parallel across modalities or can one modality be affected more than another? The question has commonly been framed in terms of whether there is a unitary or multiple semantic systems. However, there are more far-reaching implications. A unitary system assumes that concepts are represented in amodal form, divorced from modality of input. By contrast, a multiple-systems view proposes systems linked to a particular modality; for example, verbal semantics versus visual semantics.

This question posed by neuropsychologists parallels a central debate in the cognitive psychology literature of the 1970s and 1980s. Authors such as Fodor (1975) took the view that concepts are represented in the brain in a formal, amodal, and abstract manner, unrelated to the brain processing of sensory motor functions. The assumption was that modality-specific perceptual systems process information up to the level of structural description, which permits access to a unitary, amodal semantic system, where no trace of sensory motor mechanisms persist. By contrast, others such as Allport (1985) argued that concepts are represented in the brain in the same format in which they were constructed by the sensory motor system.

It is indisputable that people with semantic dementia may fail to recognize an object name (e.g., glove) while continuing to recognize a picture of a glove or the glove itself. One obvious interpretation is that this is evidence of modality-specific meaning systems in the brain (McCarthy and Warrington, 1988). Damage to an amodal semantic system would not predict such dissociated performance. An alternative explanation, however, is that the disparity is simply an artifact of task difficulty. There is a more direct mapping between surface form and meaning in the case of pictures and objects: they give direct clues to meaning, whereas words are arbitrary labels. Patient's degraded knowledge may be sufficient to support performance on the object but not the word recognition task. Less easy to dismiss as artifact are findings of dissociated knowledge of famous faces and names. Faces and names are equally arbitrary. Neither gives direct cues to identity. Moreover, superior performance has been demonstrated in both directions: better knowledge of names in some patients and faces in others. Furthermore, there is a systematic relationship with site of atrophy: poorer knowledge of names associated with greater left- than right-sided temporal lobe atrophy and poorer knowledge of faces with the converse. Such findings argue for the contribution of both hemispheres in semantic memory, and importantly, a differential role of the left and right hemispheres, with the left temporal lobe being more important for verbal and the right

for visual semantics (see Gainotti, 2012 for review of evidence). Proponents of an 'amodal semantics' view offer the counterargument that dissociations reflect impaired connectivity to modality-specific systems. Thus, left temporal lobe damage would have greater impact on connectivity with language systems located in the left hemisphere (Mion et al., 2010). The debate remains central to contemporary models of semantic memory representation.

Categories of Knowledge: Can They Break Down Selectively?

A crucial early observation was that some patients with disordered semantic memory had disproportionate impairment for certain categories of knowledge (e.g., knowledge about animals or fruits) with relative sparing of other categories (e.g., household utensils).

Early debate rested on whether such dissociations were genuine or artifacts of differences in frequency, familiarity, or typicality of the items for which knowledge was tested. It is now established that real dissociations do indeed occur, since they persist even when confounding factors are rigorously controlled. Moreover, dissociations are observed in both directions: superior knowledge of living things in some patients and nonliving thing in others, better abstract word knowledge in some and concrete word knowledge in others, better verb than noun knowledge in some and the reverse in others, and better number than word knowledge in some and the reverse in others. Crucially, category differences are mirrored by anatomical differences. Disproportionate disruption to knowledge about living things has been associated with damage to temporal lobes/ventral pathways and nonliving things to frontoparietal/dorsal pathways. Disproportionate noun impairment is linked to temporal and disproportionate verb impairment to frontal lobe dysfunction. Number knowledge depends crucially on the parietal lobes, which are spared in semantic dementia.

The interpretation of such dissociations in knowledge is nevertheless controversial. An influential view (the functional-sensory hypothesis proposed by Warrington and colleagues) was that category dissociations result from differences in the weightings of properties that define a concept and the sensory processing channels through which information was originally acquired (Warrington and Shallice, 1984). Household objects such as 'cup' and 'bowl,' it would be argued, are distinguished largely by their different functions, whereas fruits such as 'orange' and 'banana' are distinguished predominantly on the basis of sensory properties such as color, shape, texture, taste, and smell. A disruption to knowledge about sensory attributes would result in disproportionately severe impairment for animals compared to household objects, whereas disruption to knowledge about functional attributes would produce the reverse. The underlying assumption is that information is organized in the brain by property rather than category.

A contrasting account (the domain specificity hypothesis), advanced by Caramazza and Shelton (1998), argued that different categories of knowledge are supported by distinct neural networks, so they are regionally segregated in the brain. The finding in semantic dementia that number knowledge (including understanding of quantity and arithmetical skills) is remarkably preserved in comparison with knowledge of words would be a good example of apparent domain specificity. So too would the observed dissociation in some patients between understanding of nouns and verbs. More problematic for the domain specificity account is that, for the most part, category differences are relative. Patients with disordered semantic memory typically show partial information loss about concepts across multiple domains, a finding that would be predicted by the functional-sensory account because of its emphasis on relative weightings of attributes rather than category per se. Regardless of their relative merits, what is crucial is that both *functional-sensory* and *domain specificity* carry the notion of distributed representation of semantic memory across different brain regions, a notion that is supported by brain imaging and underpins current thinking about semantic memory.

Beyond the Functional-Sensory Distinction: Attribute Knowledge and Its Neural Basis

Warrington and colleagues recognized early on that the broad distinction between living and nonliving things was an oversimplification, and within the framework of the functional-sensory account a wide range of dissociations ought to be possible, reflecting relative degrees of disruption to particular sensory or nonsensory attributes. The conceptual distinction between a lemon and a lime, it would be argued, relies heavily on color, taste, and smell, but little on visual form, sound, or movement. By contrast, form, sound, and movement are relevant dimensions in the conceptual distinction between a dog and a cat. Some nonliving things, such as a cup and a jug, are manipulable, so afford specific motor actions with respect to them (e.g., grasping, pouring), whereas other nonliving things, such as a house, are nonmanipulable so are not conceptually associated with a motor action. The evidence indicates that patterns of breakdown and dissociations in performance in semantically disordered patients are indeed more subtle and diverse than the gross categorical distinctions or division between living/nonliving things would imply. Semantic dementia patients have, for example, been reported to show disproportionately impaired knowledge about color-related compared to form-related words, and about action words referring to face movements and speech acts than actions performed with hand and arm (Pulvermüller et al., 2010).

There is a wealth of functional neuroimaging data that complement the patient data (Martin, 2007; Martin and Chao, 2001; Thompson-Schill, 2003). The brain regions active during object identification have been found to depend upon the intrinsic properties of the object presented. Thus, animal naming has been found to activate occipital regions involved in visual processing in addition to temporal regions, whereas naming tools activate the left premotor area involved in object manipulation. Moreover, retrieval of knowledge about specific attributes of an object involves activation of the same brain regions (albeit with a slight anterior shift) as those involved in perceiving and using those objects. Thus, retrieving visual attribute knowledge (e.g., whether a camel has a hump) involves activation of visual association cortex, color knowledge retrieval activation of ventral temporal cortex involved in color processing, retrieval of size

information activation of parietal regions involved in spatial processing. Pulvermüller et al. (2010) showed that words relating to actions involving a specific part of the body activate motor areas that control movements of that body part.

Current Status of Knowledge and Areas of Controversy

The converging neuropsychological and neuroimaging evidence has led to common agreement that semantic memory involves distributed brain networks. Conceptual knowledge of an object such as an 'apple' includes knowledge of its sensory properties (form, color, taste, smell, and tactile sensation) as well as the actions associated with it. The brain regions involved correspond (or at least overlap) with those involved in sensory perception and action (Martin, 2007; Pulvermüller, 2012). The implication is that object concepts are grounded in and emerge from weighted activity within property-based brain regions.

A principal area of contention is whether this distributed multimodal system (or systems) is sufficient to account for semantic knowledge. Proponents of the 'amodal semantics' view argue that it is not (Patterson et al., 2007). They reason that a primary function of the semantic system is to generalize across concepts that have similar significance, even though their sensory properties may be dissimilar. To do this requires a higher order amodal semantic representation that abstracts away from modality-specific attributes. The anterior temporal lobes have been proposed as a semantic hub, where information is distilled, permitting learning of conceptual relationships. The distinction between distributed multiple-modal and distributed + amodal hub views, delineated by Patterson et al. (2007), is illustrated in Figure 1.

There are good reasons for suspecting that the anterior temporal lobes may have a special role. Semantic dementia,

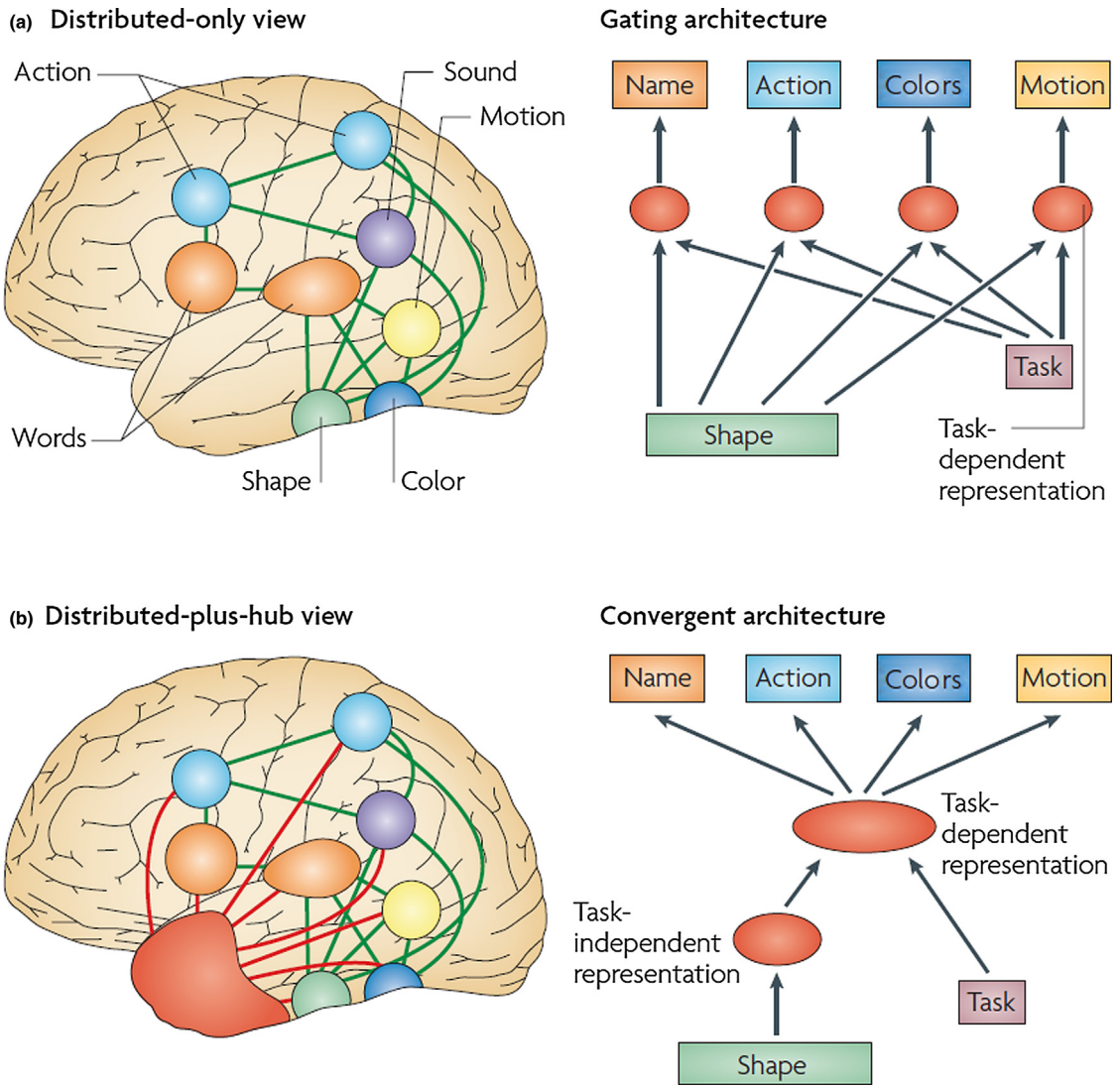

Figure 1 Two theoretical positions regarding the neuroanatomical distribution of the cortical semantic network and schematic models based on these views. Reprinted by permission from Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8, 976–987, Macmillan Publishers Ltd. © 2007.

which involves bilateral damage to anterior temporal lobes, is unique. No other disorder gives rise to such profound and pervasive semantic loss. Nevertheless, the temporal lobe degeneration is extensive leading to widespread loss of neural connectivity extending beyond anterior temporal regions. Moreover, even if crucial, it does not necessarily follow that the role of the anterior temporal lobes is in amodal representation. Antonio Damasio construed the anterior temporal lobes as areas of convergence, distinct convergence zones being dedicated to the association of different aspects of knowledge (Damasio et al., 2004). The finding of visual–verbal, right–left hemisphere differences in semantic dementia patients' recognition of famous faces and names would favor the notion of modality-specific convergence zones rather than a unitary amodal hub. The anterior temporal lobes might serve an integrative role, without entailing amodal representation. The precise site of a putative hub is unclear. Initially ascribed to the temporal pole, there is some neuroimaging evidence for a crucial role of anterior fusiform gyri (Mion et al., 2010). Visser et al. (2012) proposed two areas of convergence, one in middle temporal gyrus that integrates visual and auditory information and another in anterior temporal region identified as transmodal and interpreted as the core hub for coherent concepts.

The way that distributed information is integrated into coherent concepts is likely to remain a crucial area of research.

The Relationship between Semantic and Episodic Memory

Semantic dementia patients have impaired semantic knowledge and relatively preserved autobiographical memory, whereas classical amnesic patients show the reverse. Such clinical observations provide a persuasive argument for the utility of the distinction between semantic and episodic memory systems. Yet, does this necessarily mean that semantic and episodic memory systems are functionally independent? It would be reasonable to suspect that one might impact on the other. New semantic learning is indeed slowed in patients with episodic memory impairment, and performance on formal episodic memory tests is impaired in semantic dementia patients. This is unsurprising. It is harder to learn a list of words if they carry no meaning.

There are, however, additional findings that are less predictable and may seem counterintuitive (Snowden, 2002). In semantic dementia, the insidious loss of knowledge means that at any one stage of illness patients will know some things and not others. It is well established that factors such as frequency of occurrence, familiarity, and typicality are predictors of what is lost and retained. Thus, patients will understand 'dog' and 'horse' at a stage when they no longer recognize 'gnu' or 'bison.' Notwithstanding this general principle, semantic dementia patients who show an inability to name and comprehend simple terms such as pen and scissors may still generate surprisingly low-frequency vocabulary in conversation when it pertains to their own experience (e.g., "I attended the president's inauguration when I lived in Connecticut"; "I am church warden. I help prepare the chalice for the Eucharist"). Investigation of this phenomenon has shown a robust effect of personal familiarity (not equivalent to general frequency or familiarity of a word or object). Semantic dementia patients show superior recognition of names of people and places that have personal experiential relevance; they recognize their own objects better than alternative exemplars of the same object; they recognize objects better if they are in their usual location compared to a novel location. Importantly, their understanding of objects is influenced by their personal use, even if this is idiosyncratic. Thus, they may recognize a clothes peg as 'something used to close the breakfast cereal packet' and not as 'something used to hang clothes out to dry' if that is how they use clothes pegs in their daily lives. Patients' knowledge of the world appears to be increasingly narrow and personalized, linked to their own experience.

In Tulving's formulation, semantic memory was construed as information divorced from a temporospatial context. Yet, the findings in semantic dementia patients suggest their conceptual understanding of the world is very closely tied to their experiential autobiographical memories, including temporospatial context. Experiential memories support semantic knowledge. There is evidence too that semantic dementia patients are impaired in their ability to think prospectively about future events, reinforcing the notion of a functional interrelationship between semantic and episodic memory. Such findings suggest a more dynamic and interactive relationship between memory systems than often thought.

Future Directions in Semantic Memory Research

Much semantic memory research has centered on concrete words and objects. This is understandable not least because variables of interest can be readily manipulated and rigorously controlled. Yet, theoretical accounts must ultimately explain abstract as well as concrete concepts. They need to encompass affective as well as sensory properties of concepts. A narrow focus on semantic memory as a stable network of properties of words and objects may overlook a fundamental aspect of our knowledge: what we know comes from our experience. The precise nature of the interplay between semantic knowledge and autobiographical experience requires further exploration.

Do Schema Models Have a Role?

Schema approaches to semantic memory in the 1970s were motivated by the belief that study of word meanings alone cannot capture the richness and complexity of semantic knowledge. Patients with semantic dementia give cause to reflect again on this view. Such patients may, on occasion, generate a word appropriately in the context of spontaneous conversation (e.g., 'I'll put my shoes on'; 'I'll go and get my glasses'), yet in formal naming and word comprehension tests can neither name nor comprehend those nouns 'shoes' and 'glasses.' This apparent anomaly is not a reflection of performance variability. Semantic dementia patients are remarkably consistent in their performance, reflecting their degraded knowledge. If they cannot name or understand a word on one occasion, then they will be unable to do so on subsequent occasions. The findings point to the importance of another

factor: context. Patients may generate a word in the context of a propositional utterance, when they cannot do so in a standard naming task.

Putting on one's shoes or retrieving one's glasses constitutes a well-rehearsed goal-directed activity. So too is the verbal expression of intention to put on shoes or find one's glasses. It is possible that under conditions of profound degradation of semantic knowledge what are least vulnerable to disruption are the specific, well-rehearsed scripts of one's daily life. Experiential scripts might form the foundation upon which abstracted levels of meaning derive.

Concluding Remarks

The last decade has seen dramatic increase in the understanding of semantic memory. Patients with semantic memory disorders and sophisticated brain imaging paradigms have provided a window into the organization of semantic memory and its neural basis. However, there is still much to understand. Knowledge that London was the site of the 2012 Olympic Games is not readily divisible into sensory properties or motor actions. It is a reminder that semantic memory is more than an isolated set of word and object meanings: it is the wealth of experiential knowledge that enables us to function in the world around us.

See also: Agnosia; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Semantics; Temporal Lobe.

Bibliography

Allport, D.A., 1985. Distributed memory, modular subsystems and dysphasia. In: Newman, S.K., Epstein, R. (Eds.), Current Perspectives in Dysphasia. Churchill Livingstone, Edinburgh, pp. 207–244.

Caramazza, A., Shelton, J., 1998. Domain-specific knowledge in the brain: the animate-inanimate distinction. Journal of Cognitive Neuroscience 10, 1–34.

Collins, A.M., Quillian, M.R., 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247.

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., Damasio, A.R., 2004. Neural systems behind word and concept retrieval. Cognition 92, 179–229.

Fodor, J., 1975. The Language of Thought. Harvard University Press, Cambridge, MA.

Gainotti, G., 2012. The format of conceptual representations disrupted in semantic dementia: a position paper. Cortex 48, 521–529.

Hodges, J.R., Patterson, K., 2007. Semantic dementia: a unique clinicopathological syndrome. Lancet Neurology 7, 1004–1014.

Martin, A., 2007. The representation of object concepts in the brain. Annual Review of Psychology 58, 25–45.

Martin, A., Chao, L.L., 2001. Semantic memory and the brain: structure and processes. Current Opinion in Neurobiology 11, 194–201.

McCarthy, R.A., Warrington, E.K., 1988. Evidence for modality-specific meaning systems in the brain. Nature 334, 428–430.

Mion, M., Patterson, K., Acosta-Cabronero, J., Pengas, G., Izquierdo-Garcia, D., Hong, Y.T., et al., 2010. What the left and right anterior fusiform gyri tell us about semantic memory. Brain 133, 3256–3268.

Mummery, C.J., Patterson, K., Price, C.J., Ashburner, J., Frackowiak, R.S.J., Hodges, J.R., 2000. A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. Annals of Neurology 47, 36–45.

Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? the representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8, 976–987.

Pulvermüller, F., 2012. Meaning and the brain: the neurosemantics of referential, interactive and combinatorial knowledge. Journal of Neurolinguistics 25, 423–459.

Pulvermüller, F., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P.J., Patterson, K., 2010. The word processing deficit in semantic dementia: all categories are equal, but some categories are more equal than others. Journal of Cognitive Neuroscience 22, 2027–2041.

Schank, R.C., Abelson, R., 1977. Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Associates, Hillsdale, NJ.

Smith, E.E., Shoben, E.L., Ripps, L.J., 1974. Structure and process in semantic memory: a featural model for semantic decisions. Psychological Review 81, 214–241.

Snowden, J.S., 2002. Disorders of semantic memory. In: Baddeley, A.D., Kopelman, M.D., Wilson, B.A. (Eds.), The Handbook of Memory Disorders, second ed. John Wiley & Sons, Chichester, UK, pp. 293–314.

Snowden, J.S., Neary, D., Mann, D.M.A., 1996. Frontotemporal Lobar Degeneration. Frontotemporal Dementia, Progressive Aphasia, Semantic Dementia. Churchill-Livingstone, London.

Thompson-Schill, S.L., 2003. Neuroimaging studies of semantic memory: inferring "how" from "where". Neuropsychologia 41, 280–292.

Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.

Visser, M., Jefferies, E., Embleton, K.V., Lambon Ralph, M.A., 2012. Both the middle temporal gyrus and the ventral anterior temporal area are crucial for multimodal semantic processing: distortion corrected fMRI evidence for a double gradient of information convergence in the temporal lobes. Journal of Cognitive Neuroscience 24, 1766–1778.

Visser, M., Jefferies, E., Lambon Ralph, M.A., 2009. Semantic processing in the anterior temporal lobes: a meta-analysis of the functional imaging literature. Journal of Cognitive Neuroscience 22, 1083–1094.

Warrington, E.K., 1975. The selective impairment of semantic memory. Quarterly Journal of Experimental Psychology 27, 635–657.

Warrington, E.K., Shallice, T., 1984. Category-specific semantic impairments. Brain 107, 829–854.

Semantic Similarity, Cognitive Psychology of

Ulrike Hahn, Birkbeck, University of London, London, UK
Evan Heit, University of California, Merced, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the range of different approaches to capturing semantic similarity. Specifically, multidimensional scaling models and featural models (e.g., Tversky's contrast model) are introduced alongside newer approaches such as structural alignment accounts, context vectors, connectionist models, and generative probabilistic models. In addition, references are made to several cognitive abilities in which semantic similarity plays a role, including categorization, inductive reasoning, and memory.

Semantic similarity refers to similarity based on meaning as opposed to form. The term is used widely throughout cognitive psychology and related areas such as psycholinguistics, memory, reasoning, and neuropsychology. For example, semantically similar words are confusable with each other, and can prime each other, with the consequence that verbal memory performance is heavily dependent on similarity of meaning (*see* False Memories, Psychology of). In the context of reasoning, people draw inductive inferences on the basis of semantic similarity, for example, inferring that properties of cows are more likely to be true of horses than to be true of hedgehogs (Heit, 2000).

However, semantic similarity is not only central to psychology, it also plays an ever-increasing role in many practical computing contexts, particularly in the areas of natural language processing and information retrieval, making its study a more and more interdisciplinary endeavor.

But what is semantic similarity? In experimental work, semantic similarity is often estimated directly from participants' explicit judgments. However, there are several advantages to representing semantic similarity within a standardized framework or model. These advantages include the greater consistency and other known mathematical properties that can be imposed by a formal model, and the possible benefits of data reduction and the extraction of meaningful dimensions or features of the stimuli being analyzed. The history of research concerning semantic similarity can be captured in terms of the succession of models that have been applied to this topic. The models themselves can be divided into generic models of similarity and a number of more specialized models aimed at particular aspects of semantic similarity.

General Models of Similarity

Although the following accounts are meant to address similarity in general terms, they can be readily applied to semantic similarity. The two classical, generic approaches to modeling similarity are spatial models of similarity and Tversky's (1977) contrast model. More recently, structured approaches have addressed limitations of both of these classical accounts.

Spatial Representations

Spatial models seek to represent similarity in terms of distance in a psychological space (Shepard, 1980). An item's position is determined through its coordinate values along the relevant dimensions; nearby points thus represent similar items, whereas distant items are psychologically very different. For example, a spatial representation of mammal categories might represent different animals on dimensions of size and ferocity, with lions and tigers being close on both of these dimensions (**Figure 1**). The relevant space is derived using the technique of multidimensional scaling (MDS), a statistical procedure for dimensionality reduction. MDS works from experimental participants' judgments, typically in matrices of proximity data such as pairwise confusions or similarity ratings between items. Spatial models have been used widely for visualization purposes and as the heart of detailed cognitive models, e.g., of categorization and recognition memory (Nosofsky, 1991). Less widely used are related statistical procedures such as hierarchical

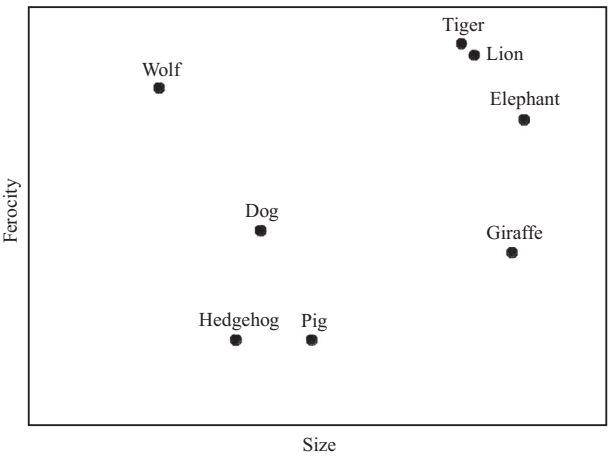

Figure 1 Multidimensional scaling representation of mammals.

clustering (Shepard, 1980), although they have also been used in the context of semantic similarity.

The Contrast Model

The traditional alternative to spatial models is Tversky's (1977) contrast model. This account was developed to address perceived limitations of spatial models, which bring these into conflict with behavioral data (but also see Nosofsky, 1991). Chief among these are violations of the so-called metric axioms that underlie any spatial scheme, such as asymmetries in human similarity judgments that are not well captured by spatial distance, which should be symmetrical. For example, people judged the similarity of North Korea to China to be greater than the similarity of China to North Korea. In the contrast model (eqn [1]), the similarity between items *a* and *b* is a positive function of the features common to both items and a negative function of the distinctive features of *a* and also the distinctive features of *b*. *Each of these three feature sets is* governed by a weighting parameter, which allows the model to capture asymmetries according to the nature of a particular task. According to the focusing hypothesis, greater attention is given to distinctive features of the first item in a comparison than of the second item, hence $\alpha > \beta$. When China is more familiar than North Korea, having more known distinctive features, then the similarity from China to North Korea should be lower than the similarity of North Korea to China:

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \qquad [1]$$

Although applications of the contrast model to the modeling of specific cognitive tasks are fewer than those of spatial models, an early application to semantic similarity can be found in Ortony (1979), who applied the model to distinctions between literal and metaphorical similarity.

Structured Representations

Despite the efforts aimed at establishing the superiority of either the contrast model or spatial models, it has been argued that both approaches share a fundamental limitation in that they define similarity based on oversimplified kinds of representations: points in space or feature sets. Arguably, most theories of the representation of natural objects, visual textures, sentences, etc. assume that these cannot be represented in line with these restrictions. Instead, they seem to require structured representations: complex representations of objects, their parts and properties, and the interrelationships between them. Descriptions such as SIBLING-OF (Linus, Lucy) and SIBLING-OF (Lucy, Linus) cannot be translated in an obvious way to either lists of features or points in space that represent the similarity between these two items as well as their differences from BROTHER-OF (Linus, Lucy). (See also Fodor and Pylyshyn (1988) for a critique of attempts in connectionist networks to represent relational structure in a featural framework.)

The perceived need for accounts of similarity to work with structured representations has given rise to the structural alignment account (see Markman (2001) for an overview), which has its roots in research on analogical reasoning. Structural alignment operates over structured representations such as frames (*see* Schemas, Frames, and Scripts in Cognitive Psychology) consisting of slots and fillers. The comparison process requires that at least some of the predicates, that is relations such as ABOVE (x, y) are identical across the comparison. These identical predicates are placed in correspondence. The alignment process then seeks to build maximal structurally consistent matches between the two representations.

Structural alignment has been implemented in a variety of computational models (e.g., Falkenhainer et al., 1990; Goldstone and Medin, 1994) that have been used to capture behavioral data, especially similarity judgments and analogical reasoning. Experimental results have supported an important prediction of the structural alignment account, that there will be a greater impact of alignable differences compared with nonalignable differences. Nonalignable differences between two representations are elements of one object that have no correspondence in the other. In contrast, alignable differences refer to representational elements that have corresponding roles in the two representations but fail to match exactly. For example, imagine two tables, one with a flower on top and the other with a bowl on top in addition to a chair beneath it. Comparing the two scenes, flower versus bowl would be an alignable difference, whereas the chair would be a nonalignable difference.

Transformational Accounts of Similarity

Most recent among the general approaches to similarity is the transformational account. This views the similarity between two objects as a function of transformation distance, that is, the complexity of transforming ones representation of one object into that of the other (e.g., Hahn et al., 2003). For example, the sequences 1 2 3 4 5 and 2 4 6 8 10 are comparatively similar because a simple operation ($\times 2$) suffices to convert the former into the latter, whereas the relationship between 1 2 3 4 5 and 3 5 7 9 11 is more complex ($\times 2 + 1$), giving rise to a greater distance, and hence, reduced similarity. This account generalizes featural or spatial accounts in as much as the addition, substitution, or deletion of features or shifts along continuous valued dimensions can be seen as particular types of transformations. However, a much broader class of transformations (such as, for example, those important in perceptual contexts such as object constancy like rotation, scaling, shear, etc.) are argued to be psychologically relevant, and this can lead to diverging predictions with featural or spatial models, because such transformations may naturally affect many stimulus aspects at once. In such cases, where a simple transformation gives rise to many individual changes at once, transformational accounts predict that items will be seen as similar, whereas featural or spatial models that track these individual changes will predict dissimilarity. Unlike featural and spatial models, the transformational account is also not restricted to particular types of representation, and measures of transformation distance can be defined both over simple feature lists (or continuous spaces), and over structured representations. Like the featural account, and in contrast to simple spatial models, the transformational account also allows asymmetric similarity. Transformations may be simpler in one direction than in the other, and there is experimental

evidence to suggest that asymmetric similarities may be created by priming directional transformations (Hahn et al., 2009).

Psychological research on transformations and similarity has – to date – focused on perceptual stimuli, but a number of specialized measures of semantic similarity can be conceptualized in transformational terms.

Specialized Models of Semantic Similarity

Semantic Differentials

Leaving behind generic models of similarity, semantic similarity has also been captured through a variety of specially built models and approaches. The first of these is Osgood et al.'s (1957) semantic differential. This approach used psychometric techniques to compute the psychological distance between concepts. Participants were required to rate concepts on 10–20 bipolar scales for a set of semantically relevant dimensions, such as positive–negative, feminine–masculine. The ratings of words on these dimensions would essentially form their coordinates in semantic space. The approach is thus related to spatial models of similarity with the difference primarily in the way that the semantic space is derived. Whereas models based on MDS require pairwise proximity data for all concepts of interest, the semantic differential approach requires that all concepts of interest are rated on all relevant dimensions. Therefore, the semantic differential method suffers from the difficulty that the relevant dimensions must be stipulated in advance. Still, semantic differentials have been widely used, largely because the data are straightforward to collect, analyze, and interpret.

Semantic Networks and Featural Models

The work of Osgood et al. motivated the semantic feature model of Smith et al. (1974) which addressed how people verify statements such as 'a robin is a bird' and 'a chicken is a bird,' considering the defining and characteristic features of these concepts. Verifying the latter statement would be slower owing to different characteristic features for chicken versus bird. The semantic feature model was developed in contrast to Collins and Quillian's (1969) semantic network model. In this model semantic meaning is captured by nodes that correspond to individual concepts. These nodes are connected by a variety of links representing the nature of the relationship between the nodes (such as an 'IS–A' link between robin and bird, indicating that a robin is a bird). Semantic similarity (or distance) is captured in terms of the number of links that must be traversed to reach one concept from another. Although the semantic network model was a groundbreaking attempt at capturing structured relations, it still suffered from some problems. For example, the model correctly predicts that verifying 'a robin is an animal' is slower than verifying 'a robin is a bird' owing to more links being traversed for the first statement. However, this model did not capture typicality effects such as the difference between robin and chicken.

While Collins and Quillian's (1969) model plays a largely historic role in psychology today, the concept of semantic networks and, more generally, graph-based representations of meaning, is very much alive in a variety of practical contexts. Computer scientists concerned with information retrieval, in particular in the context of the World Wide Web, routinely make use of ontologies represented in graph structures. Likewise computational linguists work frequently with graph-based lexical representations; chief among these is the widely used WordNet (see e.g., Miller, 1995; Fellbaum, 2005), a computer-based, large-scale, lexical database of English. WordNet contains nouns, verbs, adjectives, and adverbs in a thesaurus-like fashion, whereby words are grouped into sets of cognitive synonyms. These so-called 'Synsets' are then interlinked by conceptual semantic and lexical relations, such as IS–A links, and part–whole relations.

A variety of formal measures of similarity defined over graph structures have been proposed. Broadly, these measures can be classed as node- or edge (link)-based. The simple measure described in the context of Collins and Quillian's model (i.e., the number of links that need to traversed between two nodes) is a typical example of an edge-based measure. By contrast, node-based measures are based on the identification of common superordinate nodes for the two concepts (nodes) being compared. A prime example of this are so-called information content measures (e.g., Resnik, 1999). In the kinds of graph structures we are considering here, nodes represent unique concepts, which can be thought of as consisting of a certain amount of information. Given then that an edge represents some form of direct association between two concepts, one might view the similarity between two concepts as the extent to which they share information in common. One way of operationalizing this notion of shared information is by identifying it with common nodes further up in the hierarchy that subsume the two nodes in question. Typically, only the first superordinate node that subsumes both concepts is considered. The information content of a node (i.e., concept) is equated with its probability of occurrence, as estimated by its frequency within a large corpus of text. So the similarity between two concepts is then defined as the information content value of the first shared superordinate concept. Finally, hybrid measures combining node- and edge-based components have also been examined (e.g., Jiang and Conrath, 1997).

Though most of this research on graph-based measures of similarity has been conducted outside of psychology, putative measures of semantic similarity are routinely evaluated against benchmark human data sets such as those of Miller and Charles (1991). At the same time, the boundaries between disciplines are becoming increasingly blurred where large-scale models of semantic similarity are concerned, and popular graph-based measures can now also be found in recent research aimed specifically at psychological questions about semantic representation (see e.g., Johns and Jones, 2012).

High-Dimensional Context Spaces

Arguably the presently most popular approach to measuring semantic similarity within psychology is the use of context vectors. These corpus-derived measures of semantics and semantic similarity are successors to the idea of spatial representations of similarity. As an example, Burgess and Lund's (1997) hyperspace analog to language model (HAL) learns a high-dimensional context from large-scale linguistic corpora that encompass many millions of words of speech or text. The

model tracks lexical co-occurrences throughout the corpus and from these derives a high-dimensional representational space. The meaning of a word is conceived of as a vector. Each element of this vector corresponds to another word in the model, with the value of an element representing the number of times that the two words co-occurred within the discourse samples that constitute the corpus. For example, the vector for 'dog' will contain an element reflecting the number of times that the word 'bone' was found within a given range of words in the corpus. These vectors can be viewed as the coordinates of points (individual words) in a high-dimensional semantic space. Semantic similarity is then a matter of distance between points in this space. Several such usage-based models have been proposed to date: similar in spirit to HAL, for example, is Landauer and Dumais's (1997) Latent Semantic Analysis. Latent semantic analysis (LSA) is presently the most widely used of the usage-based models, last but not least, because the project website allows researchers to retrieve similarities between words from a simple user interface without the need to run the model themselves. LSA is also the most widely considered usage-based model in recent comparisons between competing approaches to semantic similarity (see the section Model Evaluation). The most recent addition to the family of context-based models, Jones and Mewhort's (2007) BEAGLE (for Bound Encoding of the Aggregate Language Environment) combines co-occurrence information with statistical information about transitions between words within a holographic memory system. Once again, words are represented by vectors based on patterns of use, but a lexical vector contains the superposition of both a word's context and order information across all the sentences in which it has been encountered. This gives a more fine-grained semantic representation. At the same time, the model seeks to provide a computationally more efficient, and hence psychologically more plausible approach to the acquisition of these representations.

The basic approach of co-occurrence-based models such as HAL, LSA, or BEAGLE takes to its logical consequence Wittgenstein's famous adage of 'meaning as use.' Its prime advantage over related approaches such as MDS-based spatial models of similarity and the semantic differential lies in the ability to derive semantics and thus measures of semantic similarity for arbitrarily large numbers of words without the need for any especially collected behavioral data. The ability of these models to capture a wide variety of phenomena, such as results in semantic priming and effects of semantic context on syntactic processing, has been impressive.

Generative Probabilistic Models

This approach to semantic similarity is, once again, fundamentally statistical in orientation. Like usage-based models, generative probabilistic models are derived from large corpora of text. They can be seen as probabilistic generalizations of usage-based models whereby texts within a corpus are construed as a probabilistic weighting over a set of discourse topics, and topics themselves correspond to probability distributions over words. Conversely, the distribution of words across topics can then be viewed as a representation of word meaning, and the semantic similarity between words is derived from comparison of these distributions across topics. The topics themselves are learned within a hierarchical Bayesian model (Blei et al., 2003; Griffiths et al., 2007). For example, the model may infer latent topic areas such as 'sport,' characterized by a high frequency of occurrence for words such as 'team,' 'play,' 'game,' and so on, or a topic such as 'legal system,' characterized by the frequent occurrence of words such as 'trial,' 'legal,' 'attorney,' and so on. These models are generative models in the sense that the selection of words found in the actual texts are seen as (probabilistically) generated from the underlying topics which the model has inferred. One advantage of topic models is that the generality of the underlying modeling framework allows them to combine readily semantic information with other sources of linguistic or perceptual knowledge.

Connectionist Approaches

The final approach to semantic similarity to be discussed shares with context-based models a statistical orientation, but connectionist modeling has been popular particularly in neuropsychological work on language and language processing. In connectionist models, the semantics of words are represented as patterns of activations, or banks of units representing individual semantic features. Semantic similarity is then simply the amount of overlap between different patterns; hence these models are related to the spatial accounts of similarity. However, the typically nonlinear activation functions used in these models allow virtually arbitrary re-representations of such basic similarities. The representation schemes utilized in these models tend to be handcrafted rather than derived empirically as in other schemes, such as MDS and high-dimensional context spaces. Frequently, they simply involve randomly generated representations; as demonstrated by Johns and Jones (2010), this may inadvertently influence model performance if the distributions of semantic similarities so generated do not match those associated with real words. However, it is often only very general properties of these semantic representations and the similarities between them that are crucial to a model's behavior, such as whether these representations are 'dense' (i.e., involve the activation of many semantic features) or 'sparse,' so that the actual semantic features chosen are not crucial. For example, this distinction between dense and sparse representation has been used to capture patterns of semantic errors associated with acquired reading disorders (Plaut and Shallice, 1993) and also patterns of category specific deficits following localized brain damage (Farah and McClelland, 1991). Rogers and McClelland (2004, 2008), in particular, have used simple connectionist models of semantic representation to explain a wide range of phenomena in conceptual development, as well as adult cognition in both typical and atypical adult populations, such as semantic dementia. The fact that insights from modeling have been linked to both neuropsychological and neuroimaging data (see Rogers et al., 2004; Patterson et al., 2007) means further that connectionist notions of semantic similarity have provided a bridge to the ever-increasing interest in semantic representation within cognitive neuroscience.

Model Evaluation

Models of semantic representation are routinely evaluated against human behavioral data. Chief among these are word association norms, priming data, and similarity ratings. Given the burgeoning interest both in semantic representation and semantic similarity, model evaluation increasingly also involves model comparisons. One difficulty here is that many of the models introduced above are very sensitive to parameterization. Bullinaria and Levy (2007), for example, show how factors such as the size of the training corpus or size of the context window affect model performance for usage-based models, and very simple measures may outperform more complex models when given sufficient amounts of training data (Recchia and Jones, 2009). This makes it difficult to obtain robust answers to the question of which model is 'better.' However, given the scale of these models, systematic exploration of relevant parameters is, at the very least, daunting.

This provides an impetus toward the comparative evaluation of more general model properties. Themes from the literature on general models of similarity thus recur within the context of specialized models of semantic similarity. Both the issue of structured representations, and potential violations of the metric axioms underlying spatial models have been raised in critiques of vector-based co-occurrence models such as LSA; for example, Griffiths et al. (2007) provide evidence that semantic similarity may be asymmetric. However, in the same way that proponents of MDS-based spatial models of similarity have sought to account for evidence of asymmetric similarities by appeal to characteristics of the processes (such as categorization) that draw on these similarities (e.g., Nosofsky, 1991), proponents of vector-based co-occurrence models have sought to show how asymmetries may arise naturally when the underlying vector space is coupled with a simple choice rule governing the actual task (Jones et al., 2011).

Computational complexity is another issue that has been highlighted in model comparisons (e.g., Recchia and Jones, 2009), and, now that the utility of distributional statistics in the context of semantics is well established, there is likely to be increased consideration of ways in which such information may efficiently be exploited in ways that are realistic given human information processing limitations.

Finally, models are being evaluated with respect to very general considerations about the nature of human semantic memory. The question of how best to measure or model semantic similarity is necessarily interlinked with the question of how semantic information is represented in the first place. While early accounts of semantic memory (Tulving, 1972) assumed amodal representations, there has been a wealth of research suggesting that semantic memory is experientially grounded and that semantic representations involve, or at least are linked with, representations of relevant sensorimotor experiences (see e.g., Barsalou, 2008). An early critique of usage-based models by Glenberg and Robertson (2000) focused on the fact that these models do not connect with perceptual experience. In contrast to featural models, which may include perceptual features in their representations of objects, context vectors do not connect directly with the world: knowing that 'dog' is similar to 'cat' in terms of patterns of linguistic co-occurrence is insufficient to allow one to identify either cats or dogs. Recently, modeling work such as Andrews et al. (2009) and Johns and Jones (2012) has tried to meet this challenge by showing how distributional information about context may be combined with perceptual information (features) leading to better predictions of semantic similarity.

Conclusion

There is a wide variety of models of semantic similarity available. Underlying this array of approaches is a classic tension that is as yet unresolved. Many of the models reviewed can be classed as loosely spatial in nature. These models have all been applied extensively to behavioral data; yet there are fundamental limitations to spatial approaches with respect to their representational capacity, made clear by the successes of the contrast model and the structural alignment approach. This leaves this area of research with two contrasting and seemingly incompatible strands of models, each of which successfully relates to data. One possibility for reconciliation is that spatial models and structured representations capture different aspects or types of semantic similarity: one automatic, precompiled, effortless, and in some ways shallow, and one the result of more in-depth, or line, or metacognitive processing. Such a distinction between types of processing of similarity would be compatible with the success of models based on, for example, context spaces in capturing phenomena such as semantic priming, and the success of models such as those based on structural alignment in capturing phenomena such as analogy, or the understanding of novel concept combinations. Whether such a distinction will take shape or whether there will one day be a single, all-encompassing account remains for future research. Further development of models of semantic similarity, however, will be able to benefit not only from computational ideas developed both within and outside psychology, but also from the explosion of interest in semantic representation and semantic similarity within cognitive neuroscience (see e.g., Binder et al., 2009 for a review).

See also: Lexical Semantics; Semantics.

Bibliography

Andrews, M., Vigliocco, G., Vinson, D.P., 2009. Integrating experiential and distributional data to learn semantic representations. Psychological Review 116, 463–498.

Barsalou, L., 2008. Cognitive and neural contributions to understanding the conceptual system. Current Directions in Psychological Science 17, 91–95.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A review of 120 functional neuroimaging studies. Cerebral Cortex 19, 2767–2796.

Bullinaria, J.A., Levy, J.P., 2007. Extracting semantic representations from word cooccurrence statistics: a computational study. Behavior Research Methods 39, 510–526.

Burgess, C., Lund, K., 1997. Modeling parsing constraints with high-dimensional context space. Language and Cognitive Processes 12, 177–210.
Collins, A.M., Quillian, M.R., 1969. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior 8, 240–247.
Falkenhainer, B., Forbus, K.D., Gentner, D., 1990. The structure-mapping engine: algorithm and examples. Artial Intelligence 41, 1–63.
Farah, M.J., McClelland, J.L., 1991. A computational model of semantic memory impairment: modality-specificity and emergent category-specificity. Journal of Experimental Psychology: General 120, 339–357.
Fellbaum, Christiane, 2005. WordNet and wordnets. In: Brown, Keith, et al. (Eds.), Encyclopedia of Language and Linguistics, second ed. Elsevier, Oxford, pp. 665–670.
Fodor, J., Pylyshyn, Z., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.
Glenberg, A., Robertson, D., 2000. Symbol grounding and meaning: a comparison of high dimensional and embodied theories of meaning. Journal of Memory and Language 43, 379–401.
Griffiths, T., Steyvers, M., Tenenbaum, J., 2007. Topics in semantic representation. Psychological Review 114, 211–244.
Goldstone, R.L., Medin, D.L., 1994. The time course of comparison. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 29–50.
Hahn, U., Chater, N., Richardson, L.B.C., 2003. Similarity as transformation. Cognition 87, 1–32.
Hahn, U., Close, J., Graf, M., 2009. Transformation direction influences shape similarity judgments. Psychological Science 20, 447–454.
Heit, E., 2000. Properties of inductive reasoning. Psychonomic Bulletin and Review 7, 569–592.
Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of ROCLING X. ROCLING, Taiwan, pp. 18–33.
Johns, B.T., Jones, M.N., 2010. Evaluating the random representation assumption of lexical semantics in cognitive models. Psychonomic Bulletin and Review 17, 662–672.
Johns, B.T., Jones, M.N., 2012. Perceptual inference through global lexical similarity. Topics in Cognitive Science 4, 103–120.
Jones, M.N., Mewhort, D.J.K., 2007. Representing word meaning and order information in a composite holographic lexicon. Psychological Review 114, 1–37.
Jones, M.N., Gruenenfelder, T.M., Recchia, G., 2011. In defense of spatial models of lexical semantics. In: Proceedings of the Annual Meeting of the Cognitive Science Society.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.
Markman, A.B., 2001. Structural alignment, similarity, and the internal structure of category representations. In: Hahn, U., Ramscar, M. (Eds.), Similarity and Categorization. Oxford University Press, Oxford, UK.
Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM 38, 39–41.
Miller, G.A., Charles, W.G., 1991. Contextual correlates of semantic similarity. Language and Cognitive Processes 6 (1), 1–28.
Nosofsky, R.M., 1991. Stimulus bias, asymmetric similarity and classification. Cognitive Psychology 23, 94–140.
Ortony, A., 1979. Beyond literal similarity. Psychological Review 86, 161–180.
Osgood, C.E., Suci, G.J., Tannenbaum, P.H., 1957. The Measurement of Meaning. University of Illinois Press, Urbana, IL.
Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8, 976–987.
Plaut, D.C., Shallice, T., 1993. Deep dyslexia: a case study of connectionist neuropsychology. Cognitive Neuropsychology 10, 377–500.
Recchia, G.L., Jones, M.N., 2009. More data trumps smarter algorithms: comparing pointwise mutual information to latent semantic analysis. Behavior Research Methods 41, 657–663.
Resnik, P., 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130.
Rogers, T.T., McClelland, J.L., 2004. Semantic Cognition: A Parallel Distributed Processing Approach. MIT Press.
Rogers, T.T., McClelland, J.L., 2008. A précis of semantic cognition: a parallel distributed processing approach. Behavioral and Brain Sciences 31, 689–714.
Rogers, T.T., Lambon Ralph, M.A., Garrard, P., Bozeat, S., McClelland, J.L., Hodges, J.R., Patterson, K., 2004. The structure and deterioration of semantic memory: a neuropsychological and computational investigation. Psychological Review 111, 205–235.
Shepard, R.N., 1980. Multidimensional scaling, tree-fitting, and clustering. Science 210, 390–397.
Smith, E.E., Shoben, E.J., Rips, L.J., 1974. Structure and process in semantic memory: a featural model for semantic decisions. Psychological Review 81, 214–241.
Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.
Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.

Relevant Websites

http://lsa.colorado.edu – LSA.
http://www.cs.princeton.edu/∼blei/topicmodeling.html – Topic Models.
http://wordnet.princeton.edu/ – WordNet.

Semantics

John I Saeed, Centre for Language and Communication Studies, Trinity College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Semantics is the study of meaning communicated through language and is usually taken to be one of the three main branches of linguistics, along with *phonology*, the study of sound systems, and *grammar*, which includes the study of word structure (*morphology*) and of sentence structure (*syntax*). This article surveys some of the main topics of current semantics research.

Introduction

Traditionally, the main focus of linguistic semantics has been on word meaning, or lexical semantics. Since classical times writers have commented on the fact, noticed surely by most reflecting individuals, that the meaning of words changes over time. Such observations are the seeds of etymology, the study of the history of words. Over longer stretches of time, such changes become very obvious, especially in literate societies. Words seem to shift around: some narrow in meaning such as English *queen* which earlier mean 'woman, wife' but now means 'wife of a king.' Others become more general, while still others shift to take on new meaning or disappear altogether. Words are borrowed from language to language. The study of such processes is now part of historical semantics (Allan and Robinson, 2011). Another motivation for the study of word meaning comes from dictionary writers as they try to establish meaning correspondences between words in different languages, or in monolingual dictionaries, seek to provide definitions for all the words of a language in terms of a simple core vocabulary. In lexicology, similarities and differences in word meaning are a central concern.

The principled study of the meaning of phrases and sentences has only become established in linguistics relatively recently. Thus it is still common for descriptive grammars of individual languages to contain no separate section on semantics other than providing a lexicon. Nonetheless it has always been clear that one can identify semantic relations between sentences. Speakers of English know from the semantics of negation that nominal negation has different effects than sentence negation, so that *No-one complained* may aptly be used to answer *Who complained?* while *Someone did not complain* may not. Two other sentences may seem to say essentially the same thing, even be paraphrases of each other, yet one may be more suited to one context than another, like the pair *Bandits looted the train* and *The train was looted by bandits*. A single sentence may be internally inconsistent such as *Today is now tomorrow* or seem to be repetitive or redundant in meaning such as *A capital city is a capital city*. Another feature of sentence meaning is the regularity with which listeners draw inferences from sentences and often take these to be part of the meaning of what was said. Some inferential links are very strong, such as entailment. Thus we say that *Bob drank all of the beer* entails *Bob drank some of the beer* (assuming the same individual Bob, beer, etc.), because it is hard to think of a situation where acceptance of the second sentence would not follow automatically from acceptance of the first. Other inferential links are weaker and more contextually dependent: from the utterance *Bob drank some of the beer* it might be reasonable to infer *Bob didn't drink all of the beer* but it is possible to think of situations where this inference would not hold. We might say that a speaker of the first sentence is implying the second, in a certain context. Speakers of all languages regularly predict and use such inferential behavior to convey their meaning, such that often more meaning seems to be communicated than is explicitly stated. All these aspects of sentence meaning and are currently under study in various semantic frameworks.

Semanticists share with philosophers an interest in key issues in the use of language, notably in reference. We use this term to describe the way in which speakers can pick out, or name, entities in the world by using words as symbols. Many scholars, especially formal semanticists, accept Frege's distinction between reference (in German, *Bedeutung*) and sense (*Sinn*); see Frege (1980). Reference is the act of identifying an entity (the referent) while sense is the means of doing so. Two different linguistic expressions like *the number after nine* and *the number before eleven* differ in sense but they both share the same referent, *ten*. For semanticists it is particularly interesting to study the various mechanisms that a language offers to speakers for this act of referring. These include names like *Dublin*, nouns like *cat* which can be used to refer to a single individual *your cat* or a whole class *Cats are carnivorous*, quantified nominals like *many cats*, *some cats*, *a few cats*, etc. and so on. Linguists as well as philosophers have to account for language's ability to allow us to refer to nonexistent and hypothetical referents like *World War Three*, *the still undiscovered cure for cancer*, *the end of the world*.

Semanticists also share interests with psychologists, for if sense is the meaning of an expression, it seems natural to many semanticists to equate it with a conceptual representation. Cognitive linguists in particular (for example, Talmy, 2000; Langacker, 2008), but also some generative linguists (Jackendoff, 2002), seek to explore the relationship between semantic structure and conceptual structure. One axis of the debate is whether words, for example, are simply labels for concepts, or whether there is a need for an independent semantic interface that isolates just grammatically relevant elements of conceptual structure. Many languages make grammatical distinctions corresponding to the conceptual distinctions of gender and number but few involve distinctions of color or between different animal species. If certain aspects of concepts are more relevant to grammatical rules, as is also claimed by Pinker (1989), this may be justification for a semantic interface.

Approaches to Meaning

Even in these brief remarks we have had to touch on the crucial relationship between meaning and context. Language of course typically occurs in acts of communication and linguists have to cope with the fact that utterances of the same words may communicate different meanings to different individuals in different contexts. One response to this problem is to hypothesize that linguistic units like words, phrases, and sentences have an element of inherent meaning that does not vary across contexts. This is sometimes called inherent, or simply, sentence meaning. Language users, for example, speakers and listeners, then enrich this sentence meaning with contextual information to create the particular meaning the speaker means to convey at the specific time, which can then be called speaker meaning. One common way of reflecting this view is to divide the study of meaning into semantics, which becomes the study of sentence meaning, and pragmatics, which is then the study of speaker meaning, or how speakers use language in concrete situations. This is an attempt to deal with the tension between the relative predictability of language between fellow speakers and the great variability of individual interpretations in interactive contexts. One consequence of this approach is the view that the words that a speaker utters underdetermine her intended meaning.

Semantics as a branch of linguistics is marked by the theoretical fragmentation of the field as a whole. The distinction between formal and functional approaches, for example, is as marked in semantics as elsewhere. This is a large subject to broach here but see Givón (1995) and Newmeyer (1998) for characteristic and somewhat antagonistic views, though Newmeyer (2010) strikes a conciliatory note. One important difference is the attitude to the autonomy of levels of analysis. Are semantics and syntax best treated as autonomous areas of study, each with its own characteristic entities and processes? A related question at a more general level is whether linguistic processes can be described independently of general psychological processes or the study of social interaction. Scholars in different theoretical frameworks will give contradictory answers to these questions of micro- and macro-autonomy. Autonomy at both levels is characteristic of semantics within generative grammar (see, for example, Chomsky, 1995). Functionalists like Halliday (1994) and Harder (1996) would on the other hand argue against micro-autonomy, suggesting that grammatical relations and structure cannot be understood without reference to semantic function. They also seek motivation for linguistic structure in the dynamics of communicative interaction. A slightly different external mapping is characteristic of cognitive semantics (for example, Langacker, 2008), where semantic structures are correlated to conceptual structures.

Another dividing issue in semantics is the value of formal representations. Scholars are divided on whether our knowledge of semantics is sufficiently mature to support attempts at mathematical or other symbolic modeling, indeed on whether such modeling serves any use in this area. Partee (2004), for example, defends the view of formal semanticists that the application of symbolic logic to natural languages, following in particular the work of Montague (1974), represents a great advance in semantic description. Jackendoff (2002) on the other hand acknowledges the value of formalism in semantic theory and description but argues that formal logic is too narrow to adequately describe meaning in language. Other scholars, such as Wierzbicka (1996) and Goddard (2008), view the search for formalism as premature and distracting.

There has been an explosive increase in the research in formal semantics since Montague's (1974) proposal that the analysis of formal languages could serve as the basis for the description of natural languages. Montague's original theory comprised a syntax for the natural language, say English, a syntax for the logical language into which English should be translated (intensional logic), rules for the translation, and rules for the semantic interpretation of the intensional logic. This and subsequent formal approaches are typically referential (or denotational) in that their emphasis is on the connection of language with a set of possible worlds, including the real, external world and the hypothetical worlds set up by speakers. Crucial to this correspondence is the notion of truth, defined at the sentence level. A sentence is true if it correctly describes a situation in some world. In this view, the meaning of a sentence is characterized by describing the conditions which must hold for it to be true. The central task for such approaches is to extend the formal language to cope with the semantic features of natural language while maintaining the rigor and precision of the methodology (see van Benthem and Ter Meulen, 2010 for a comprehensive overview of research in this paradigm).

Research in cognitive semantics presents an alternative strategy. Cognitive semanticists reject what they see as the mathematical, anti-mentalist approach of formal semantics. In their view, meaning is described by relating linguistic expressions to mental entities, conventionalized conceptual structures. These semanticists have proposed a number of conceptual structures and processes, many deriving from perception and bodily experience and in particular, conceptual models of space. Proposals for underlying conceptual structures include image schemas (Johnson, 1987), mental spaces (Fauconnier, 1997), and conceptual spaces (Gärdenfors, 1999). Cognitive semanticists are sympathetic to usage-based accounts such as Taylor (2012), which relates linguistic representations to stored memories of actual language events. Another focus of interest is the processes for extending concepts and here special attention is given to metaphor. Lakoff (1987) and Johnson (1987) have argued against the classical view of metaphor and metonymy as something outside normal language, added as a kind of stylistic ornament. For these writers metaphor is an essential element in our categorization of the world and our thinking processes. Cognitive semanticists have also investigated the conceptual processes which reveal the importance of the speaker's perspective and construal of a scene, including viewpoint shifting, figure-ground shifting, and profiling (Langacker, 2008).

Topics in Sentence Semantics

Many of the semantic systems of language, for example, tense, aspect, mood, and negation, are marked grammatically on individual words such as verbs. However, they operate over the whole sentence. This 'localization' is the reason that

descriptive grammars usually distribute semantic description over their analyses of grammatical forms. Such semantic systems offer the speaker a range of meaning distinctions through which to communicate a message. Theoretical semanticists attempt to characterize each system *qua* system, as in, for example, Verkuyl's (1993) work on aspect and Hornstein's (1990) on tense. Typological linguists try to characterize the variation in such systems across the world's languages, as in the studies of tense, aspect, and mood by Comrie (1976, 1985), Binnick (1991), Bybee et al. (1994), Dähl (1985, 2000), and Frawley (2006). We can sketch some basic features of some of these systems.

Tense

Tense is the linguistic marking of time and allows a speaker to locate a described situation relative to some reference point in time. The basic reference point is the time of speaking or writing. Tense is thus described as a deictic system since it is part of a language's system for establishing spatial, temporal, and human zones of reference around the speaker. For example, in English the deictic words *here*, *now*, *I*, *you* provide points of reference, the contents of which are establishing within a particular context of communication. Tense is often associated with verbs but may be marked by other elements, even in English where temporal adverbs may carry a similar function, as in the contrast between *We let the apartment last year*, *We let the apartment every year*, *We let the apartment next week*. Most grammatical tense systems allow the speaker to describe situations as prior to, concurrent with, or following the act of speaking. So in English, we have the three tenses: past, future, and present. These are basic tenses, usually marked on a verb and can be represented by a diagram like **Figure 1** (using *understand* as an example), in which time is metaphorically represented as line moving left to right, and the symbol represents the time of the act of speaking or writing.

When events are located by tenses in this way against the act of speaking, they are termed absolute tenses. More complicated time references are built on this simple system. By using complex tenses in English, for example, a speaker can identify a time in the past or future and use that time as the reference point for an event in its own past or future, as in examples like *By last Friday he had interviewed seven candidates*, or *Before the end of this century we will have discovered a cure for cancer*. These are often called relative tenses since they are parasitic on the basic system. Such a subdivision of the past or future can be represented in a simple diagram like **Figure 2** where represents the extra reference point in time.

Reichenbach (1947) proposed a system for describing these absolute and relative tense possibilities. In it the time of the act of speaking (or utterance time) is termed S and the time of the

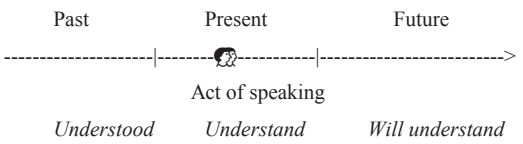

Figure 1 Simple tenses.

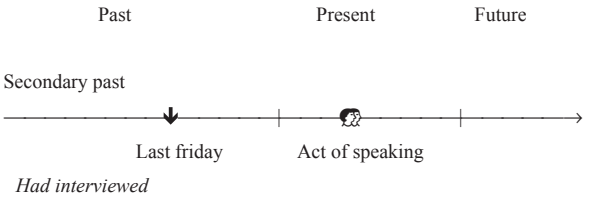

Figure 2 Complex past tense in *By last Friday he had interviewed seven candidates*.

described event, E. Thus present tense has the two overlapping, written S, R, while the other simple tenses involve a precedence, shown as E_S for the past and S_E for the future. As we have seen, the complex tenses require an extra reference point other than S, called in this system the reference time, R. In simple tenses R is simultaneous with E but in complex tenses R may be before or after E. So in **Figure 2** the event (the interview) occurs before the reference time (last Friday) and the reference time occurs before the act of speaking. These relations can be represented as (E_R) and (R_S), which then describes the past perfect in English. Reichenbach's system is developed further in Klein (1994) (see Comrie, 1985; Hornstein, 1990 for discussion of this system to describe the range of tenses found in languages).

The marking of temporal reference by tense is usually supported in real-life communication by pragmatic inferences. Wilson and Sperber (2012b) discuss how temporal ordering or sequencing is often inferred in context, for example, in deciding whether the clauses in *John left and Joan got angry* are to be taken as sequential. Similarly, intervals of time are often unspecified and left to be inferred in context as in the possible time intervals inferred for their examples *I have had breakfast* and *I have been to Tibet*.

Situation Type and Aspect

Situation type and aspect are terms for a language's resources that allow a speaker to describe the temporal 'shape' of events. The term situation type is used to describe the system encoded in the words of a language, while aspect is used for the grammatical systems which perform a similar role. Distinctions of situation type, for which the term Aktionsart is also used, typically allow speakers to describe a situation either as static as in *The bananas are ripe* or as dynamic as in *The bananas are ripening*. Here the state is the result of the process but the same situation can be viewed as more static or dynamic as in *The baby is asleep* and *The baby is sleeping*. As these examples show, this distinction is lexically marked: in English, for example, adjectives are typically used for states and verbs for dynamic situations. There are, however, a group of stative verbs such as *know*, *understand*, *love*, *hate*, which describe static situation types. There are a number of semantic distinctions typically found amongst dynamic verbs, for example, the telic/atelic (bounded/unbounded) distinction and the punctual/durative distinction. Telic verbs describe processes which are seen as having a natural completion, which atelic verbs do not. A telic example is *Matthew was growing up* and an atelic, *Matthew was drinking*. If these processes are interrupted at any point we can automatically say *Matthew drank* but not *Matthew grew up*.

However atelic verbs can form telic phrases and sentences by combining with other grammatical elements so that *Matthew was drinking a pint of beer* is telic. Durative verbs, as the term suggests, describe processes that last for a period of time while punctual describes those that seem so instantaneous that it has no detectable internal structure, as in the comparison between *The man slept* and *The light flashed*. As has often been observed, if an English punctual verb is used with a durative adverbial, the result is an iterative meaning, as in *The light flashed all night*, where we understand the event to be repeated over the time mentioned.

Situation type typically interacts with aspect. Aspect is the grammatical system that allows the speaker choices in how to portray the internal temporal nature of a situation. An event, for example, may be viewed as closed and completed, as in *Joan wrote a book* or as an ongoing process, perhaps unfinished, as in *Joan was writing a book*.

The latter verb form is described as being in the progressive aspect in English but similar distinctions are very common in the languages of the world. In many languages we find described a distinction between perfective and imperfective aspects, used to describe complete versus incomplete events; see Bybee et al. (1994) for a survey. As mentioned above, aspect is intimately associated both with situation type and with tense. In Classical Arabic the perfective is strongly associated with past tense (Comrie, 1976; Binnick, 1991), while in Russian the imperfective is used to describe future events. In English, for example, stative verbs are typically not used with progressive aspect, so that one may say *I know some French* but not *I am knowing some French*. Staying with the progressive, when it is used in the present tense in English (and in many other languages) it carries a meaning of proximate future or confident prediction as in *We're driving to Los Angeles* or *I'm leaving you*. The combination of the three semantic categories of tense, situation type, and aspect produces a complex system that allows speakers to make subtle distinctions in relating an event or describing a situation.

Modality

Modality is a semantic system that allows speakers to express varying attitudes to a proposition. Semanticists have traditionally identified two types of modality. One is termed epistemic modality, which encodes a speaker's commitment to, or belief in, a proposition, from the certainty of *The ozone layer is shrinking* to the weaker commitments of *The ozone layer may/might/could be shrinking*. The second is deontic modality, where the speaker signals a judgment toward social factors of obligation, responsibility, and permission, as in the various interpretations of *You must/can/may/ought to borrow this book*. These examples show that similar markers, here auxiliary verbs, can be used for both types. When modality distinctions are marked by particular verbal forms, these are traditionally called moods. Thus many languages, including Classical Greek and Somali, have a verb form labeled the optative mood for expressing wishes and desires. Other markers of modality in English include verbs of propositional attitude as in *I know/believe/think/doubt/that the ozone layer is shrinking* and modal adjectives as in *It is certain/probable/likely/possible that the ozone layer is shrinking*.

A related semantic system is evidentiality, where a speaker communicates the basis or source for presenting a proposition. In English and many other languages this may be done by adding expressions like *allegedly, so I've heard, they say*, etc. but certain languages mark such differences morphologically, as in Makah, a Nootkan language spoken in Washington State (Jacobsen, 1986, 10):

Wiki•caxaw	'It's bad weather' (seen or experienced directly)
Wiki•caxakpi•d	'It looks like bad weather' (inference from physical evidence)
Wiki•caxakqadi	'It sounds like bad weather' (on the evidence of hearing)
Wiki•caxakwa•d	'I'm told there's bad weather' (quoting someone else)

Aikhenvald (2004) surveys such evidential systems across the languages of the world.

Semantic Roles

This term describes the speaker's semantic repertoire when deciding how to relate participants in a described event. One influential proposal in the semantics literature is that each language contains a set of semantic roles, the choice of which is partly determined by the lexical semantics of the verb selected by the speaker. A characteristic list of such roles is:

AGENT:	Initiator of some action, typically capable of acting with volition
PATIENT:	Entity undergoing the effect of some action, typically undergoing a change of state
THEME:	The entity which is moved by an action, or whose location is described
EXPERIENCER:	An entity aware of the action or state described by the predicate but not in control of it
RECIPIENT:	The entity which receives something
BENEFICIARY:	The entity benefiting from some action
INSTRUMENT:	The means by which an AGENT causes something to come about
LOCATION:	The place in which something is situated or takes place
GOAL:	The entity toward which something moves
SOURCE:	The entity from which something moves
STIMULUS:	The entity causing an effect in the EXPERIENCER

In an example like *Harry immobilized the tank with a broomstick*, the entity Harry is described as the AGENT, the tank as the PATIENT, and the broomstick as the INSTRUMENT.

These roles have also variously been called deep semantic cases, thematic relations, participant roles, and thematic roles.

One concern is to explain the matching between semantic roles and grammatical relations. In many languages, as in the last example, there is a tendency for the subject of the sentence to correspond to the AGENT, the direct object to a PATIENT or THEME, and an INSTRUMENT often occurs as a prepositional phrase. Certain verbs allow variations from this basic mapping, for example, the pattern we find with English verbs like *break*: *The boy broke the window with a stone* (Subject = AGENT), *The stone*

broke the window (Subject = INSTRUMENT), *The window broke* (Subject = PATIENT). Clearly verbs can be arranged into classes depending on the variations of mappings they allow and not all English verbs pattern like *break*. We can say *The admiral watched the battle with a telescope* but *The telescope watched the battle* and *The battle watched* sound decidedly odd.

From this literature emerges the claim that certain mappings are more natural or universal. One proposal is that, for example, there is an implicational hierarchy governing the mapping to subject, typically such as: AGENT > EXPERIENCER > THEME/PATIENT > INSTRUMENT > LOCATION. In such a hierarchy, each left element is more preferred than its right neighbor, so that moving rightward along the string gives us less expected subjects. The hierarchy also makes certain typological claims: if a language allows a certain semantic role to be the subject, it will allow all those to its left. Thus, if we find that a language allows the role INSTRUMENT to be the subject, we predict that it allows the roles to the left, but we do not know if it allows LOCATION subjects.

One further application of semantic roles is in lexical semantics where the notion allows verbs to be classified by their semantic argument structure. Verbs are assigned semantic role templates or grids by which they may be sorted into natural classes. Thus, English has a class of transfer, or giving verbs which in one type includes the verbs *give, lend, supply, pay, donate, contribute*. These verbs encode a view of the transfer from the perspective of the AGENT and may be assigned the grid <AGENT, THEME, RECIPIENT>, where the underlined argument is realized as the subject, as in *The committee donated aid to the famine victims*. A second subclass of these transfer verbs encodes the process from the perspective of the RECIPIENT. These verbs include *receive, accept, borrow, buy, purchase, rent, hire* and have the grid <RECIPIENT, THEME, SOURCE>, as in *The victims received aid from the committee*.

Entailment, Presupposition, and Implication

These terms relate to types of information a hearer gains from an utterance but which are not stated directly by the speaker. These phenomena have received a lot of attention because they seem to straddle the putative divide between semantics and pragmatics described above and because they reveal the dynamic and interactive nature of understanding the meaning of utterances. Entailment describes a relationship between sentences such that on the basis of one sentence, a hearer will accept a second, unstated sentence purely on the basis of the meaning of the first. Thus sentence A entails sentence B if it is not possible to accept A but reject B. In this view, a sentence like *I bought a dog today* entails *I bought an animal today*; or *President Kennedy was assassinated yesterday* entails *President Kennedy is now dead*. Clearly these sentential relations depend on lexical relations: a speaker who understand the meaning of the English word *dog* knows that a dog is an animal; similarly the verb *assassinate* necessarily involves the death of the unfortunate object argument. Entailment then is seen as a purely automatic process, involving no reasoning or deduction, but following from the hearer's linguistic knowledge. Entailment is amenable to characterization by truth conditions. A sentence is said to entail another if the truth of the first guarantees the truth of the second and the falsity of the second guarantees the falsity of the first.

Presupposition, on the other hand, is a more difficult notion. In basic terms the idea is simple enough: that a speaker communicates certain assumptions aside from the main message. A range of linguistic elements communicates these assumptions. Some, such as names, and definiteness markers, like the articles *the* and *my*, presuppose the existence of entities. Thus *James Brown is in town* presupposes the existence of a person so called. Other elements have more specific presuppositions. A verb like *stop* presupposes a preexisting situation. So a sentence *Christopher has stopped smoking* presupposes *Christopher smoked*. If treated as a truth conditional relation, presupposition is distinguished from entailment by the fact that the former but not the latter survives under negation: *Christopher has not stopped smoking* still presupposes *Christopher smoked* but the sentence *I didn't buy a dog today* does not entail *I bought an animal today*.

There are a number of other differences between entailment and presupposition that cast doubts on the ability of a purely semantic, truth conditional account of the latter. Presuppositions are notoriously context sensitive, for example. They may be canceled without causing an anomaly: a hearer can reply *Christopher hasn't stopped smoking, because he never smoked* to cancel the presupposition by what is sometimes called metalinguistic negation. This dependency on context has led some writers to propose that presupposition is a pragmatic notion, definable in terms of the set of background assumption that the speaker assumes is shared in the conversation (see Beaver, 1997 for discussion).

A third type of inference is Grice's conversational implicature (1975, 1978). This is an extremely context-sensitive type of inference which allows participants in a conversation to maintain coherence. So given the invented exchange below:

A: Did you give Mary the book?
B: I haven't seen her yet.

it is reasonable for A to infer the answer 'no' to her question. Grice proposed that such inferences are routinely relied on by both speakers and hearers and that this reliance is based on certain assumptions that hearers make about a speaker's conduct. Grice classified these into several different types, giving rise to different types of inference, or from the speaker's point of view, what he termed implicatures. The four main maxims are called Quality, Quantity, Relevance, and Manner (Grice, 1975, 1978). They amount to a claim that a listener will assume, unless there is evidence to the contrary, that a speaker will have calculated her utterance along a number of parameters: she will tell the truth, try to estimate what her audience knows and package her material accordingly, have some idea of the current topic, and give some thought to her audience being able to understand her. In our example above, it is A's assumption that B's reply is intended to be relevant that allows the inference 'no.'

Implicature has three characteristics: firstly, that it is implied rather than said; secondly, that its existence is a result of the context, i.e., the specific interaction. There is no guarantee that in other contexts *I haven't seen her* will be used to communicate *no*. Thirdly, implicature is cancellable without

causing a contradiction. Thus the implicature 'no' in our example can be canceled if B adds the clause *but I posted it to her last week*.

These three notions entailment, presupposition, and implicature can all be seen as types of inference. They are all produced in conversation and are taken by participants to be part of the meaning of what a speaker has said. They differ in a number of features and crucially in context sensitivity. The attempt to provide a unified analysis of them all is currently a challenge to semantic and pragmatic theories (see Wilson and Sperber, 2012a for discussion of such a unified approach).

Future Developments

Although semantics remains theoretically a very diverse field it is possible to detect some shared trends which seem likely to develop further. One is a move away from a static view of sentences in isolation, detached from the speaker/writer's act of communication toward dynamic, discourse-based approaches. This has always been characteristic of functional approaches to meaning but has also been noticeable in formal approaches as they move away from their more philosophical origins. Among prom-inent examples of this we might mention Discourse Representation Theory (Kamp et al., 2011) and Dynamic Semantics (Groenendijk et al., 1996). This move has put further pressure on decisions about the border to be drawn between semantics and pragmatics and in particular about semantic underspecification (see Recanati, 2004; Cappelen and Lepore, 2005; Preyer and Peter, 2005 for characteristic discussions).

Another development which seems likely to continue is a closer integration with other disciplines in cognitive science. In particular, computational techniques seem certain to make further impact on a range of semantic enquiry, from lexicography to the modeling of questions and other forms of dialogue. A subfield of computational semantics has emerged and will continue to develop (see van Eijck and Unger, 2010 for example).

See also: Deixis; Evidentials in Linguistics; Language and Poetic Structure; Lexicology and Lexicography; Linguistic Presupposition; Valency and Argument Structure in Syntax.

Bibliography

Aikhenvald, Alexandra Y., 2004. Evidentiality. Oxford University Press, Oxford.
Allan, Kathryn, Robinson, Justyna A. (Eds.), 2011. Current Methods in Historical Semantics. Mouton de Gruyter, Berlin.
Beaver, David I., 1997. Presuppositions. In: van Bentham, J., ter Meulen, A. (Eds.), Handbook of Logic and Language. Elsevier, Amsterdam, pp. 939–1008.
van Benthem, Johan, Ter Meulen, Alice (Eds.), 2010. Handbook of Logic and Language, second ed. Elsevier, Amsterdam.
Binnick, Robert I., 1991. Time and the Verb: A Guide to Tense and Aspect. Oxford University Press, Oxford.
Bybee, Joan, Perkins, Revere, Pagliuca, William, 1994. The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World. University of Chicago Press, Chicago.
Cappelen, Herman W., Lepore, Ernest, 2005. Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism. Blackwell, Oxford.
Chomsky, Noam, 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Comrie, Bernard, 1976. Aspect. Cambridge University Press, Cambridge.
Comrie, Bernard, 1985. Tense. Cambridge University Press, Cambridge.
Dähl, Östen, 1985. Tense and Aspect Systems. Blackwell, Oxford.
Dähl, Östen (Ed.), 2000. Tense and Aspect in the Languages of Europe. Mouton de gruyter, Berlin.
van Eijck, Jan, Unger, Christine, 2010. Computational Semantics with Functional Programming. Cambridge University Press, Cambridge.
Fauconnier, Gilles, 1997. Mappings in Thought and Language. Cambridge University Press, Cambridge.
Frawley, William (Ed.), 2006. The Expression of Modality. Mouton de Gruyter, Berlin.
Frege, Gottlob, 1980. In: Geach, Peter, Black, Max (Eds.), Translations from the Philosophical Writings of Gottlob Frege. Blackwell, Oxford.
Gärdenfors, Peter, 1999. Some tenets of cognitive semantics. In: Allwood, Jens, Gärdenfors, Peter (Eds.), Cognitive Semantics: Meaning and Cognition. John Benjamins, Amsterdam and Philadelphia, pp. 12–36.
Givón, Talmy, 1995. Functionalism and Grammar. John Benjamins, Amsterdam and Philadelphia.
Goddard, Cliff, 2008. Cross-linguistic Semantics. John Benjamins, Amsterdam.
Grice, H. Paul, 1975. Logic and conversation. In: Cole, Peter, Morgan, Jerry (Eds.), Syntax and Semantics 3: Speech Acts. Academic Press, New York, pp. 43–58.
Grice, H. Paul, 1978. Further notes on logic and conversation. In: Cole, Peter (Ed.), Syntax and Semantics 9: Pragmatics. Academic Press, New York, pp. 113–128.
Groenendijk, J., Stokhof, M., Veltman, F., 1996. Coreference and modality. In: Lappin, S. (Ed.), The Handbook of Contemporary Semantic Theory. Blackwell, Oxford, pp. 179–214.
Halliday, M.A.K., 1994. An Introduction to Functional Grammar, second ed. Edward Arnold, London.
Harder, Peter, 1996. Functional Semantics: A Theory of Meaning, Structure and Tense in English. Mouton de Gruyter, Berlin and New York.
Hornstein, N., 1990. As Time Goes by. MIT Press, Cambridge, MA.
Jackendoff, Ray, 1990. Semantic Structures. MIT Press, Cambridge, MA.
Jackendoff, Ray, 1996. Semantics and cognition. In: Lappin, S. (Ed.), The Handbook of Contemporary Semantic Theory. Blackwell, Oxford, pp. 539–560.
Jackendoff, Ray, 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press, Oxford.
Jacobsen Jr., William H., 1986. The heterogeneity of evidentials in Makah. In: Chafe, Wallace, Nichols, Johanna (Eds.), Evidentiality: The Linguistic Coding of Epistemology. Ablex, Norwood, New Jersey, pp. 3–28.
Johnson, Mark, 1987. The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press, Chicago.
Kamp, Hans, van Genabith, Josef, Reyle, Uwe, 2011. Discourse representation theory. In: Gabbay, Dov M., Guenthner, Franz (Eds.), Handbook of Philosophical Logic, second ed., vol. 15. Kluwer, Dordrecht, pp. 125–394.
Klein, Wolfgang, 1994. Time in Language. Routledge, Chapman and Hall, London.
Lakoff, George, 1987. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago.
Langacker, Ronald W., 2008. Cognitive Grammar: A Basic Introduction. Oxford University Press, Oxford.
Lappin, Shalom (Ed.), 1996. The Handbook of Contemporary Semantic Theory. Blackwell, Oxford.
Lehmann, W., 1992. Historical Linguistics, third ed. Routledge, London.
Montague, Richard, 1974. In: Thomason, Richmond H. (Ed.), Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven and London.
Newmeyer, Frederick J., 1998. Language Form and Language Function. MIT Press, Cambridge, MA.
Newmeyer, Frederick J., 2010. Formalism and functionalism in linguistics. Wiley Interdisciplinary Reviews: Cognitive Science 1 (3), 301–307.
Partee, Barbara H., 1996. The development of formal semantics in linguistic theory. In: Lappin, S. (Ed.), The Handbook of Contemporary Semantic Theory. Blackwell, Oxford, pp. 11–38.
Partee, Barbara H., 2004. Reflections of a formal semanticist. Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee. Blackwell, Oxford, pp. 1–25.
Pinker, Steven, 1989. Learnability and Cognition: The Acquisition of Argument Structure. MIT Press, Cambridge, MA.
Preyer, Gerhard, Peter, Georg, 2005. Contextualism in Philosophy: Knowledge, Meaning and Truth. Clarendon Press, Oxford.
Recanati, François, 2004. Literal Meaning. Cambridge University Press, Cambridge.
Reichenbach, H., 1947. Elements of symbolic logic the Free Press. Collier-Macmillan, New York, London.
Rosner, M., Johnson, R., 1992. Computational Linguistics and Formal Semantics. Cambridge University Press, Cambridge.

Talmy, Leonard, 2000. Toward a Cognitive Semantics, 2 vols. MIT Press, Cambridge, MA.
Taylor, John R., 2012. The Mental Corpus: How Language Is Represented in the Mind. Oxford University Press, Oxford.
Verkuyl, H.J., 1993. A Theory of Aspectuality: The Interaction between Temporal and Atemporal Structure. Cambridge University Press, Cambridge.
Wierzbicka, Anna, 1996. Semantics, Primes and Universals. Oxford University Press, Oxford.
Wilson, Deirdre, Sperber, Dan, 2012a. Meaning and Relevance. Cambridge University Press, Cambridge.
Wilson, Deirdre, Sperber, Dan, 2012b. In: Wilson, Sperber (Eds.), Pragmatics and Time, pp. 169–186.

Semiotics

Klaus B Jensen, University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Semiotics is the interdisciplinary study of signs, particularly their social origins, uses, and consequences. Rooted in philosophy, logic, and linguistics, the field has developed since the 1960s as an internationally influential approach to research on contemporary culture and communication. Revisiting classic epistemological issues, semiotics further has contributed to the theory of science by clarifying the relationship of a third form of inference – abduction – to deduction and induction. In the context of digital networks, semiotics helps to specify the several different categories of signs that enter into computer-mediated social interaction.

Semiotics is an interdisciplinary field that studies "the life of signs within society" (Saussure, 1959/1916: 16). While the term 'signs' is most commonly used to refer to the constituents of verbal language and other concrete vehicles of communication, in semiotics it further denotes any means of knowing about or representing some aspect of reality – through perception, cognition, or social interaction. Accordingly, semiotics exhibits a variety of family resemblances with disciplines such as philosophy, psychology, and sociology. From the 1960s, semiotics became an internationally influential approach to research on culture and communication within the humanities and social sciences. This article describes the classical origins of semiotics, exemplifies its applications to contemporary society, outlines its implications for the theory of science, and, in conclusion, considers an extension and potential redefinition of semiotics in the context of digital computers.

Origins: Logic and Linguistics

A Classical Heritage

One of the first accounts in Western philosophy and science of signs and their functions in representation and communication was advanced by Aristotle in the fourth century BCE:

> Spoken words are the symbols of mental experience and written words are the symbols of spoken words. Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images.
>
> Aristotle, 2007: n.p.

Aristotle identified the minimal elements of most theories of knowledge, classical as well as modern: things in the world, mental impressions, and oral or written expressions about the things as well as the impressions. These elements – their status and interrelations – have preoccupied philosophy, and have continued to challenge natural, human, and social sciences. At issue has been the existence of a shared reality, the historical and cultural variability of human experience, and the relationship between subject and object, and between subject and subject, as mediated by signs, in some sense of that word.

Aristotle had suggested two main senses. First, the mental impressions that individuals hold could be considered signs that represent certain phenomena in the world to them. Second, spoken and written expressions are signs through which people are able to represent and communicate a particular understanding of these phenomena to others. The first sense – mental impressions – recalls a classical understanding of signs, not as words or images, but as naturally occurring evidence that something is the case. For example, fever is a sign of illness, and clouds are a sign that it may rain, insofar as these signs are interpreted by humans. The second sense – means of communication – points to the distinction that came to inform modern sciences between conventional signs, whether verbal language or mathematical notations, and natural signs or sense data.

An undercurrent in the history of ideas has addressed the ambiguities that follow from these several conceptions of what is a sign. A specific statement concerning natural signs and language as two varieties of one general category of signs was presented by St. Augustine (*c.*400 CE). Grounded in Christian cosmology of the medieval period, this view was elaborated with reference to an understanding of nature as God's Book and as an analogy to the Bible, or the Good Book. Both verbal and natural signs could be seen to bear witness to a divinely sanctioned order or 'great chain of being.' During the seventeenth century, the term 'semiotic,' along with a modern and more differentiated conception of signs, first appeared in John Locke's *An Essay Concerning Human Understanding* (1690). While proposing a science of signs in general, Locke, in fact, restricted his focus to "the most usual" signs (Clarke, 1990: 40), namely, verbal language and logic. Linguistics and logic became the two main disciplines feeding into contemporary semiotics, as represented by two founding figures: Ferdinand de Saussure and Charles Sanders Peirce.

Peircean Logic and Semiotics

Charles Sanders Peirce (1839–1914) was the first thinker to recover this undercurrent from the history of ideas in an attempt to develop a general semiotics, which he referred to as 'semeiotic' (for key texts, see Peirce, 1992–98). Peirce understood his theory of signs as a form of logic, which would inform a comprehensive philosophical system examining the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95033-5

nature of both being (ontology) and knowledge (epistemology). The key to this system is Peirce's definition of the sign in terms of its three aspects:

> A sign, or *representamen*, is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the *interpretant* of the first sign. That sign stands for something, its *object*.
>
> Peirce, 1931–58: 2.228

One important implication of this definition is that signs always serve to mediate between objects in the world (material and nonmaterial) and concepts in the mind. In contrast to skepticist positions from Antiquity to postmodernism, Peirce's point was that signs are not *what* we know, or all that we know, but *how* we come to know what we can justify saying that we know, in science as well as in everyday life. Peircean semiotics, thus, married a classical, Aristotelian notion of realism with the modern, Kantian position that humans necessarily construct their understanding of reality in particular cognitive categories.

Peirce further emphasized that human understanding is not a singular act of internalizing reality, but a continuous process of interpreting reality with a view to action, what he called 'semiosis.' This multistep process is witnessed, for example, in scientific discovery, but also in the continuous coordination of everyday life. Figure 1 illustrates the process of semiosis, indicating how any given interpretation (interpretant) itself serves as a sign in the next stage of a process that gradually differentiates the interpreter's understanding of objects in reality, and which, in principle, is unending. Although Peirce's outlook was that of a logician and a natural scientist, semiosis can be taken to also refer to those processes of communication by which cultures are maintained and societies reproduced and, to a degree, reformed.

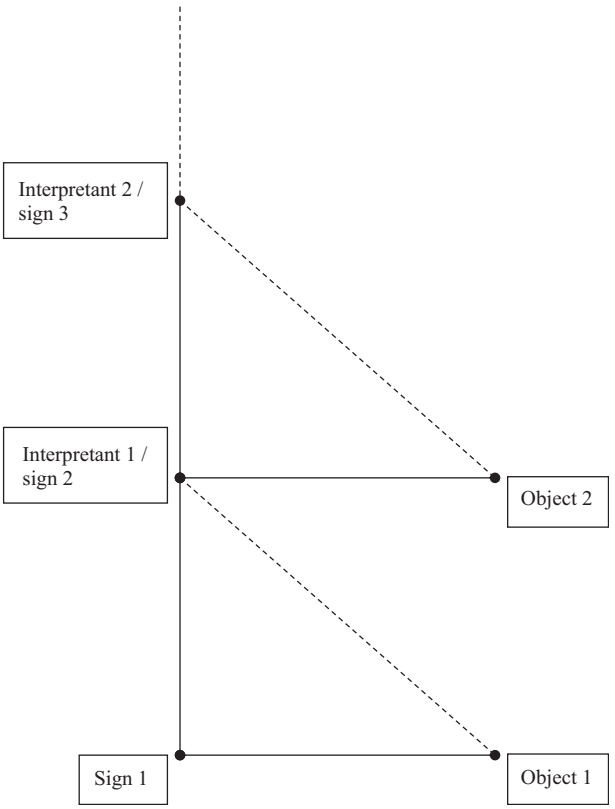

Figure 1 The process of semiosis.

One of the most influential elements of Peirce's semiotics outside logic and philosophy has been his categorization of different types of signs, particularly icon, index, and symbol. Icons relate to their objects through resemblance (e.g., a realistic painting); indices have a causal relation (e.g., the sound of a drum beating); and symbols have an arbitrary relation to their object (e.g., the words of this article). Importantly, signs as actually used by humans in social contexts typically incorporate all of these three aspects to varying degrees. A photograph resembles what it depicts; it results from the impact of physical reality on a recording device; and it commonly symbolizes something beyond the immediate depiction, for instance, in advertising. Different disciplines and fields have employed these sign types as analytical instruments in order to describe the ways in which humans perceive, represent, and act in and on social reality.

In an interdisciplinary perspective, the relationship between semiosis and social action is key. Peirce was a founder not just of semiotics, but of pragmatism, as well, together with William James, John Dewey, and other thinkers. Semiosis informs beliefs, which anticipate actions. Along with semiotics, pragmatism has proven a central legacy for social sciences and cultural studies in recent decades (Dickstein, 1998).

Saussurean Linguistics and Semiology

The other main figure in the development of semiotics, Ferdinand de Saussure (1857–1913), focused on verbal language, which Peirce had described as belonging to one type of signs, namely symbols. The main achievement of Saussure was to outline a framework for modern linguistics. In contrast to the emphasis that philology had placed on the diachronic perspective of how languages change over time, Saussure proposed to study language as a system in a synchronic perspective. Language as an abstract system (*langue*) could be distinguished, at least for analytical purposes, from the actual uses of language (*parole*). The language system, according to Saussure, has two dimensions. Along the syntagmatic dimension, letters, words, phrases, etc., combine to make up meaningful wholes, and each of these units has been chosen as one of several possibilities along a paradigmatic dimension, for example, the verb 'choose' instead of 'select.' This combinatorial system helps to account for the unique flexibility of language as a medium of social interaction.

The intersection of the paradigmatic and syntagmatic axes gives rise to two wider forms of expression and representation – metaphor and metonymy – not just in literature, but in communication and culture generally. Metaphor is the outcome of an unexpected paradigmatic choice that affects the content of the communication as a whole (the syntagm) and, thus, activates additional frames of interpretation. For example, during the 1980s, the American politician and civil rights activist, Jesse Jackson, served as the spokesperson for a 'rainbow coalition'

that aimed to join various disempowered groups, particularly non-white communities, into a new political force. This term represents a significant contrast to the traditional reference to the US population as entering into a 'melting pot.' While the melting pot eliminates variety, the rainbow derives its essential quality and beauty from the difference and complementarity of multiple colors. Speeches and other communications would activate these meanings for a particular political purpose. Metonymy, in turn, is the process by which a single sign evokes the full syntagm to which that sign belongs. When 'the White House' was said to comment on the 'rainbow coalition,' the reference to the building would reference both the American presidency and the authority and vested interests associated with it (Drotner et al., 1996: 195).

To be precise, Saussure referred to the broader science of signs, as cited in the introduction to this article, not as semiotics, but as 'semiology.' Like Locke, and later Peirce, Sausure used the classical Greek term for sign, *semeion*, to coin a nomenclature for a new analytical discipline. The two different terminologies are explained, further, by the fact that Peirce and Saussure had no knowledge of each other's work. Moreover, in Saussure's own work, the program for a comprehensive semiology remained undeveloped, appearing almost as an aside compared to his main enterprise of linguistics. It was during the consolidation of semiotics as an interdisciplinary field from the 1960s that this became the agreed term, overriding Peirce's semeiotic and Saussure's semiology, as symbolized by the formation in 1969 of the International Association for Semiotic Studies. It was also during this period that Saussure's systemic approach to language was redeveloped to apply to the study of culture and society.

One important legacy of Saussure has been his account of the arbitrariness of the linguistic sign. This type of sign is said to have two sides, a 'signified' (concept) and a 'signifier' (the acoustic image associated with it), whose relationship is arbitrary or conventional, as suggested by the wealth of terms in different languages for the same phenomenon. This argument has sometimes been construed in relativist terms, as if sign users were paradoxically free to choose their own meanings and, hence, almost destined to remain divorced from each other and any consensual reality. Or, they might be in a position, individually or as subcultures, to reject prevailing conceptions of social reality. The point is, rather, that the linguistic system as a whole is arbitrary, but fixed by social convention. When applied to studies of cultural forms other than language, specifically images, or when extended to the description of social structures as semiotic systems, the principle of arbitrariness has posed certain analytical difficulties, as illustrated in contemporary studies of media, communication, and culture.

Applications: Media, Communication, and Culture

It was the application of semiotics to questions of meaning beyond logic and linguistics that served to consolidate semiotics as a recognizable, if still heterogeneous field from the 1960s onwards. Drawing on additional traditions of aesthetic and social inquiry – from rhetoric and hermeneutics to critical theory – this emerging field began to operationalize the study of 'the life of signs within society.' The objects of analysis ranged widely, from classical artworks to popular culture, and from premodern societies to modern mass media. At the same time, studies were united by a common emphasis on culture, in the broad sense of worldviews that orient social interaction. The work of Claude Lévi-Strauss (1963/1958) on structural anthropology was characteristic of, and highly influential on, this perspective. Structuralism in studies of culture and communication suggested that certain shared, but largely preconscious systems of interpretation help to explain both the ways in which individuals orient themselves and act in specific contexts, and the cultural horizons of an entire society or epoch. It was Saussure's approach to signs as an abstract system (*langue*) which manifests itself in concrete uses (*parole*) that provided the model for structuralism in several fields. The general model was extended, for example, to theories about social power (Althusser, 1977/1965) and about the unconscious understood as a language (Lacan, 1977).

Of the predominantly French scholars who were instrumental in the consolidation of semiotic forms of analysis, Roland Barthes, along with A.J. Greimas, stands out as a particularly innovative and systematic theorist. His specific model of two levels of signification, outlined in **Figure 2**, has been one of the most widely applied approaches to the relationship between concrete sign vehicles, such as texts and images, and the ideologies or myths that they articulate and may reinforce. Building on Louis Hjelmslev's (1963/1943) formal linguistics, Barthes (1973/1957) suggested that the combination of signifier and signified (expressive form and conceptual content) in one sign (e.g., a picture of a black man in a French uniform saluting the flag) may, in a next step, become the expressive form of a further, ideological content (e.g., that French imperialism is not a discriminatory or oppressive system). Barthes' critical agenda, shared by much

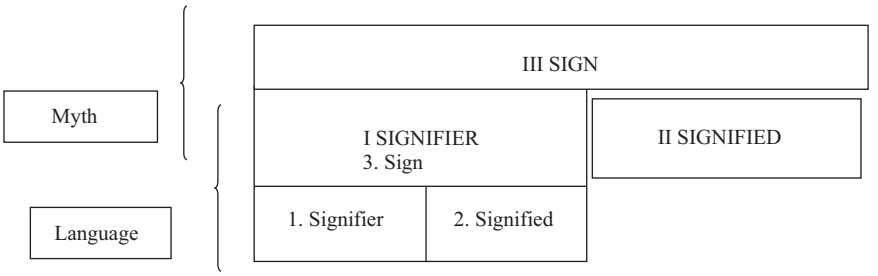

Figure 2 Two levels of signification. Adapted from Barthes, R., 1973. Mythologies. Paladin, London. (Original work published 1957.)

semiotic scholarship since then, was that this discursive mechanism serves to naturalize particular worldviews, while oppressing others, and should be exposed through analytical deconstruction.

In a current perspective, social sciences can be seen to have appropriated and redeveloped semiotics in two different ways. On the one hand, semiotics offers rigorous, yet nuanced methodologies for examining how signs enter into social interactions at the individual, group, and institutional levels. Overlapping, in part, with discourse analysis, such methodologies represent the predominant application of the tradition outside philosophy and the subspecialty of semiotics. On the other hand, some scholars have adopted semiotics as a comprehensive theoretical framework, suggesting that not just texts and images, but also societies and psyches, even biology and cosmology, may be conceptualized as semiotic processes (Posner et al., 1997–98).

This last position points to an abiding dilemma for semiotics, namely, its scope and domain of relevance. A conception of semiotics as a theory-of-everything holds several pitfalls, and has given rise to reservations and criticisms. For one thing, such an extension and overlay of logical and linguistic concepts onto other fields and disciplines may conflate different levels of analysis. For another thing, semiotic studies have sometimes focused on the formal aspects of signs and semiotic processes to such an extent that the substantive aspects of the social or cultural domains in question have been neglected. A case in point is research on visual communication, which, in some instances, has referred to Saussure's principle of arbitrariness, deriving from verbal language, to account for both still and moving images. Early work by Christian Metz (1974) did specify that the medium of film does not constitute a language system, analogous to verbal language, but a language or form of communication drawing on several semiotic registers. Later work has reemphasized that audiovisual media, to a significant extent, operate on different perceptual and cognitive modalities than speech and writing, communicating through natural rather than conventional signs (Messaris, 1994).

A related reservation arises from the ambition of much semiotic research to arrive at conclusions regarding the difference that signs make in social practice. Given the predominant focus of semiotic studies on texts and other cultural artifacts, inferences from texts to readings to effects on individuals, groups, or institutions may not be warranted. The textual focus may also exaggerate the extent to which signs in themselves, rather than the social conditions circumscribing them, determine the life of both signs and societies. In recent decades, these concerns have led to efforts at formulating a social semiotics, or sociosemiotics (for overview, see Cobley and Randviir, 2009), integrating semiotic methodology with social theory and communication studies (Hodge and Kress, 1988; Jensen, 1995).

Theory of Signs and Theory of Science

The ambiguous scope of semiotics helps to explain both its continuing influence since Antiquity and the fact that, mostly, it has remained an undercurrent in the theory of science. The understanding, articulated in Aristotle's writings, that signs constitute evidence for an interpreter of something else that is at least temporarily absent – being interfaces with underlying phenomena and processes – can be taken as a civilizational leap that has enabled scientific reflexivity as well as social organization beyond the here and now. The concept of signs, further, has informed the crucial distinction, again deriving from Aristotle, between necessary and probable relations in logical inferences. Returning to the examples above, fever is a sure sign that the person is ill, since illness is a necessary condition of fever. But clouds are only a probable sign of rain. Empirical social sciences have been built on the premise that probable signs can lead to warranted inferences about social events and processes.

The relationship between a theory of signs and a theory of science has been explored primarily from a Peircean perspective. Peirce himself made important contributions to logic and to the theory of science generally, notably by developing 'abduction' as a third form of inference alongside deduction and induction. Abduction introduces a new rule to account for what may appear as surprising facts, and is central to both scientific innovation and everyday reasoning. Moreover, the wider philosophical tradition of pragmatism, which Peirce inaugurated, has emphasized the interrelation between knowledge and action, conceiving belief as an orientation toward action (Joas, 1993). By contrast, the Saussurean tradition has typically bracketed issues concerning "relations between the word and the world" (Jakobson, 1960: 19).

In future research, semiotics is likely to remain a source of inspiration in diverse scientific domains, as it has been for 2500 years. In view of the record from the past century, the field seems less likely to develop into the sort of full-fledged discipline that at least some of its practitioners have sought to promote. One opportunity for semiotics is to consolidate an interdisciplinary framework for exploring how different disciplines and fields conceive of their data and their concepts as signs. The semiotic heritage offers systematic analytical procedures and means of reflexivity regarding the role of signs in society as well as in the social sciences. Another opportunity is for semiotics to refocus attention on a distinctive type of signs that are currently changing social infrastructures around the world: digital signs.

Meta-Signs and Meta-Communication

The digital computer is a semiotic machine *par excellence*. In addition to being one more medium of representation and interaction, computers are able to integrate and simulate all previous media forms. As such, they are meta-media (Kay and Goldberg, 1999/1977). Because computers are programmable, they comprise new types of signs, while simultaneously reproducing familiar signs under different technological circumstances. Some implications of computing for semiotics can be laid out with reference to two concepts: meta-signs and meta-communication (see further Jensen, 2010).

The interface between computers and semiotics has been a topic of inquiry for some time (Andersen, 1990). Digital computers produce a special kind of what computer scientists refer to as 'data' – metadata, or data about data. In semiotic terms, these can be understood as meta-signs, or signs about

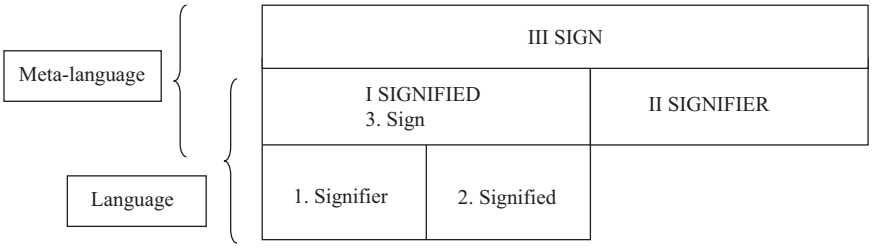

Figure 3 Language and meta-language. Adapted from Barthes, R., 1973. Mythologies. Paladin, London. (Original work published 1957.)

signs. Examples include the title, editor, authors, publisher, ISBN number, contents, index, and so on, of the present encyclopedia. In the case of computers, meta-signs indicate not only which items of information are available in a given work, but also which ones are accessed, by whom, at what time, in which formats, and, perhaps, in which combinations, sequences, and contexts of use. Importantly, meta-signs are documented in and of the operation of digital systems, such as search engines. While this feature facilitates the process of searching, it also enables surveillance of the searcher.

It should be emphasized that although meta-signs are key to digital communication systems, they also enter into other forms of communication, including face-to-face interaction. Gregory Bateson (1972) developed the argument that all communication depends on two additional forms of meta-communication. First, we introduce meta-linguistic information into our interactions to specify, for instance, that we mean a reference to clouds to serve as an illustration of a type of sign. Second, we meta-communicate about our relationships as communicators, clarifying whether we are having an academic discussion about semiotics or, alternatively, are referring to clouds as concrete signs that may make us remind each other to bring umbrellas.

A comparable meta-perspective was outlined by Hjelmslev (1943/1963) in parallel with his account of connotation languages, as redeveloped by Barthes (1957/1973) as a tool of semiotic analysis. Connotation languages and meta-languages have different, but complementary relations to their common point of reference, which is first-order language – 'language' as commonly understood and used by humans day in and day out. Connotation languages build on language, and are themselves languages or vehicles of communication, as exemplified by Barthes' myths. They make sense and make a point. Meta-languages, in comparison, describe language: they are not languages in the ordinary sense, but languages about languages, for instance, syntactical or semantic descriptions of the English language.

Figure 3 summarizes the principle of meta-languages. Compared to Barthes' first model (**Figure 2**), this last figure inverses the relation between signifier and signified at the second level of description. Connotation languages add to the meaning or codification of first-level content; meta-languages contribute to configuring the social relationships that people enter into with reference to this content. A linguist, for instance, takes the analyst's rather than the user's stance vis-à-vis language.

Hjelmslev had qualified his account of meta-languages by making an antecedent distinction between scientific and nonscientific languages. Meta-languages were described as scientific languages, defined by their formal operations (but presumably accessible, above all, by expert users of language, such as linguists). In some media and modalities, however, meta-languages can be used by anyone. Ordinary users of digital media, freely and easily, employ a wide variety of such meta-languages, which serve to frame their communicative interactions with others and the contents of those interactions: they tag the blog entries of others; they forward a news story to friends from a website via an embedded email service; they customize their own profiles at social network sites; and they pull a later push of information to themselves through an RSS (Really Simple Syndication) feed. In each case, users add a layer of information that is not part of the content proper, but which facilitates a whole range of communicative interactions.

This relevance of meta-languages, evidently, was not anticipated by Hjelmslev, who devised his models before the coming of digital computing. Also Barthes' appropriation of Hjelmslev's connotation languages for the study of contemporary culture represented an unanticipated application of a general insight. Hjelmslev conceived languages as systems – systems of expression and second-order systems that either build on or describe such systems. Barthes' accomplishment was to transfer this approach to the study of communication as a practice in social contexts. Connotation languages can be examined not only in analytical retrospect, but as they are expressed and take effect. Equally, meta-languages can be understood not merely as analytical systems, but as widespread practices of meta-communication. Meta-languages hold analytical potentials far beyond verbal language and linguistics, reminding interdisciplinary research of the rich heritage of semiotics and its continued relevance for the study of the life of signs in changing societies.

See also: Audiences, Media; Mass Media, Representations in; Narrative, Sociology of; Photography as a Medium; Visual Images in the Media.

Bibliography

Althusser, L., 1977. For Marx. Verso, London. (Original work published 1965.)

Andersen, P.B., 1990. A Theory of Computer Semiotics: Semiotic Approaches to Construction and Assessment of Computer Systems. Cambridge University Press, Cambridge.

Aristotle, 2007. On Interpretation. http://etext.library.adelaide.edu.au/a/aristotle/interpretation/complete.html (retrieved 15.08.07.).

Barthes, R., 1973. Mythologies. Paladin, London. (Original work published 1957.)

Bateson, G., 1972. Steps to an Ecology of Mind. Granada, London.
Clarke, D.S., 1990. Sources of Semiotic. Southern Illinois University Press, Carbondale, IL.
Cobley, P., Randviir, A., 2009. Introduction: what is sociosemiotics? Semiotica 173 (1–4), 1–39.
Dickstein, M. (Ed.), 1998. The Revival of Pragmatism. Duke University Press, Durham, NC.
Drotner, K., Jensen, K.B., Poulsen, I., Schrøder, K.C., 1996. Medier og Kultur [Media and Culture]. Borgen, Copenhagen, Denmark.
Hjelmslev, L., 1963. Prolegomena to a Theory of Language. University of Wisconsin Press, Madison, WI (Original work published 1943).
Hodge, B., Kress, G., 1988. Social Semiotics. Polity Press, Cambridge.
Jakobson, R., 1960. Closing statement: linguistics and poetics. In: Sebeok, T.A. (Ed.), Style in Language. MIT Press, Cambridge, MA, pp. 350–377.
Jensen, K.B., 1995. The Social Semiotics of Mass Communication. Sage, London.
Jensen, K.B., 2010. Media Convergence: The Three Degrees of Network, Mass, and Interpersonal Communication. Routledge, London, New York.
Joas, H., 1993. Pragmatism and Social Theory. University of Chicago Press, Chicago.
Kay, A., Goldberg, A., 1999. Personal dynamic media. In: Mayer, P.A. (Ed.), Computer Media and Communication: A Reader. Oxford University Press, Oxford, pp. 111–119 (Original work published 1977).
Lacan, J., 1977. The Four Fundamental Concepts of Psychoanalysis. Penguin, Harmondsworth, UK.
Lévi-Strauss, C., 1963. Structural Anthropology. Penguin, New York (Original work published 1958).
Messaris, P., 1994. Visual "Literacy": Image, Mind, and Reality. Westview Press, Boulder, CO.
Metz, C., 1974. Language and Cinema. Mouton, The Hague.
Peirce, C.S., 1931–58. Collected Papers. Harvard University Press, Cambridge, MA.
Peirce, C.S., 1992–98. The Essential Peirce. Indiana University Press, Bloomington, IN.
Posner, R., Robering, K., Sebeok, T.A. (Eds.), 1997–98. Semiotics: A Handbook of the Sign-theoretic Foundations of Nature and Culture. Walter de Gruyter, Berlin.
Saussure, F. d, 1959. Course in General Linguistics. Peter Owen, London (Original work published 1916).

Semiparametric Models

Joel L Horowitz, Northwestern University, Evanston, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Much empirical research in the social sciences is concerned with estimating conditional mean functions. The most frequently used estimation methods assume that the conditional mean function is known up to a set of constant parameters that can be estimated from data. Such methods are called 'parametric.' Their use greatly simplifies estimation and inference but is rarely justified by theoretical or other a priori considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading. Semiparametric methods reduce the strength of the assumptions required for estimation and inference, thereby reducing the opportunities for obtaining misleading results. In addition, semiparametric methods mitigate certain disadvantages of fully nonparametric methods that make no assumptions about the shape of the conditional mean function. This article describes three important semiparametric models for conditional mean functions. They are single index, partially linear, and additive models. These models are compared with parametric and fully nonparametric models. An example based on real data illustrates the pitfalls of parametric models and the advantages of semiparametric models.

Much empirical research in the social sciences is concerned with estimating conditional mean functions. For example, labor economists are interested in estimating the mean wages of employed individuals, conditional on characteristics such as years of work experience and education. The most frequently used estimation methods assume that the conditional mean function is known up to a set of constant parameters that can be estimated from data, possibly by ordinary least squares (OLS). Models in which the only unknown quantities are a finite set of constant parameters are called 'parametric.' The use of a parametric model greatly simplifies estimation, statistical inference, and interpretation of the estimation results, but is rarely justified by theoretical or other a priori considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading. Semiparametric statistical methods reduce the strength of the assumptions required for estimation and inference, thereby reducing the opportunities for obtaining misleading results. These methods are applicable to a wide variety of estimation problems in economics and other fields.

Introduction

A conditional mean function gives the mean of a dependent variable Y conditional on a vector of explanatory variables X. Denote the mean of Y conditional on $X = x$ by $E(Y|x)$. For example, suppose that Y is a worker's weekly wage (or, more often in applied econometrics, the logarithm of the wage) and X includes such variables as years of work experience and education, race, and sex. Then $E(Y|x)$ is the mean wage (or logarithm of the wage) when experience and the other explanatory variables have the values specified by x. As an illustration, the solid line in **Figure 1** shows an estimate of the mean of the logarithm of weekly wages, log W, conditional on years of work experience, EXP, for white males with 12 years of education who work full time and live in urban areas of the North Central USA. The estimate was obtained by applying a nonparametric method (explained in Section Nonparametric Models) to data from the 1993 Current Population Survey (CPS). The estimated conditional mean of log W increases steadily up to approximately 30 years of experience and is flat thereafter.

In most applications, $E(Y|x)$ is unknown and must be estimated from data on the variables of interest. In the case of estimating a wage function, the data consist of observations of individuals' wages, years of experience, and other characteristics. The most widely used method for estimating $E(Y|x)$ is not the nonparametric method mentioned previously, but rather a method that assumes that $E(Y|x)$ is known up to finitely many constant parameters. This gives a 'parametric model' for $E(Y|x)$. Often, $E(Y|x)$ is assumed to be a linear function of x, in which case the parameters can be estimated by OLS, among other ways. A linear function has the form $E(Y|x) = \beta' x$, where β is a vector of coefficients. For example, if x consists of an intercept and the two variables x_1 and x_2, then β has three components, and $\beta' x = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. OLS estimators are described in many textbooks. See, for example, Goldberger (1998).

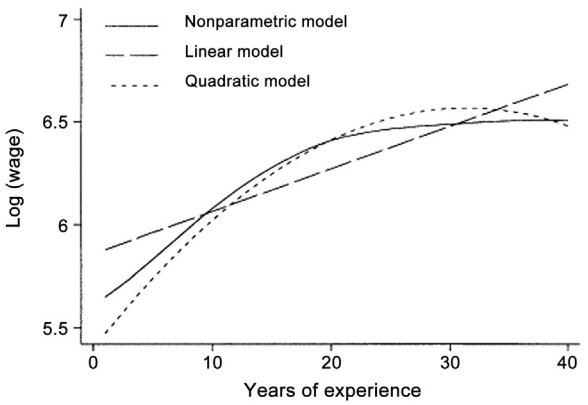

Figure 1 Estimates of $E(\log W|\text{EXP})$.

The OLS estimator of $E(Y|x)$ can be highly misleading if $E(Y|x)$ is not linear in the components of x, that is, if there is no β such that $E(Y|x) = \beta' x$. This problem is illustrated by the dashed and dotted lines in **Figure 1**, which show two parametric estimates of the mean of the logarithm of weekly wages conditional on years of work experience. The dashed line is the OLS estimate, that is, obtained by assuming that $E(\log W|\text{EXP})$ is the linear function $E(\log W|\text{EXP}) = \beta_0 + \beta_1\text{EXP}$. The dotted line is the OLS estimate that is obtained by assuming that $E(\log W|\text{EXP})$ is quadratic: $E(\log W|\text{EXP}) = \beta_0 + \beta_1\text{EXP} + \beta_2\text{EXP}^2$. The nonparametric estimate (solid line) places no restrictions on the shape of $E(\log W|\text{EXP})$. The linear and quadratic models give misleading estimates of $E(\log W|\text{EXP})$. The linear model indicates that $E(\log W|\text{EXP})$ steadily increases as experience increases. The quadratic model indicates that $E(\log W|\text{EXP})$ decreases after 32 years of experience. In contrast, the nonparametric estimate of $E(\log W|\text{EXP})$ becomes nearly flat at approximately 30 years of experience. Because the nonparametric estimate does not restrict the conditional mean function to be linear or quadratic, it is more likely to represent the true conditional mean function.

The opportunities for specification error increase if Y is binary. For example, consider a model of the choice of travel mode for the trip to work. Suppose that the available modes are automobile and transit. Let $Y = 1$ if an individual chooses automobile and $Y = 0$ if the individual chooses transit. Let X be a vector of explanatory variables such as the travel times and costs by automobile and transit. Then $E(Y|x)$ is the probability that $Y = 1$ (the probability that the individual chooses automobile) conditional on $X = x$. This probability will be denoted $P(Y = 1|x)$. In applications of binary response models, it is often assumed that $P(Y|x) = G(\beta' x)$, where β is a vector of constant coefficients and G is a known probability distribution function. Often, G is assumed to be the cumulative standard normal distribution function, which yields a 'binary probit' model, or the cumulative logistic distribution function, which yields a 'binary logit' model. The coefficients β can then be estimated by the method of maximum likelihood (Amemiya, 1985: pp. 115–127). However, there are now two potential sources of specification error. First, the dependence of Y on x may not be through the linear *index* $\beta' x$. Second, even if the index $\beta' x$ is correct, the 'response function' G may not be the normal or logistic distribution function. See Horowitz (1993, 1998) for examples of specification errors in binary response models and their consequences.

Many investigators attempt to minimize the risk of specification error by carrying out a 'specification search' in which several different models are estimated and conclusions are based on the one that appears to fit the data best. Specification searches may be unavoidable in some applications, but they have many undesirable properties and their use should be minimized. There is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investigator's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based. This is because inference following a specification search must be conditional on the selected model, whereas the standard statistical theory of inference is unconditional.

The rest of this article describes methods that deal with the problem of specification error by relaxing the assumptions about functional form that are made by parametric models. The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. These are described in Section Nonparametric Models. They assume that $E(Y|x)$ is a smooth function but make no other assumptions about its shape or functional form. However, nonparametric methods have important disadvantages that seriously limit their usefulness in applications. Semiparametric methods, which are described in Section Semiparametric Models offer a compromise. They make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. In addition semiparametric methods avoid the most serious practical disadvantages of nonparametric methods.

Nonparametric Models

In nonparametric estimation, $E(Y|x)$ is assumed to satisfy smoothness conditions such as differentiability, but no assumptions are made about its shape or the form of its dependence on x. Härdle (1990) and Fan and Gijbels (1996) provide detailed discussions of nonparametric estimation methods. One easily understood and frequently used method is called 'kernel estimation.' This method was used to produce the solid line in **Figure 1**. To describe the kernel method simply, assume that X is a continuously distributed, scalar random variable. Let $\{Y_i, X_i: i = 1, \ldots, n\}$ be a random sample of n observations of (Y, X). Let K be a probability density function that is bounded, continuous, and symmetrical about zero. For example, K may be the standard normal density function. Let $\{h_n\}$ be a sequence of positive numbers that converges to 0 as $n \to \infty$. For each $n = 1, 2, \ldots$ and $i = 1, \ldots, n$ define the function $w_{ni}(\cdot)$ by

$$w_{ni}(x) = \frac{K[(x - X_i)/h_n]}{\sum_{i=1}^{n} K[(x - X_i)/h_n]} \qquad [1]$$

Then the kernel nonparametric estimator of $E(Y|x)$ is

$$H_n(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i \qquad [2]$$

$H_n(x)$ is a weighted average of the observed values of Y. Observations Y_i for which X_i is close to x get higher weight than do observations for which X_i is far from x. It can be shown that if $h_n \to 0$ and $nh_n/\log n \to \infty$ as $n \to \infty$, then $H_n(x) \to E(Y|x)$ with probability 1. Thus, if n is large, $H_n(x)$ is likely to be very close to $E(Y|x)$. Härdle (1990) provides a detailed discussion of the statistical properties of kernel nonparametric estimators.

Nonparametric estimation minimizes the risk of specification error, but the price of this flexibility can be high. One important reason for this is that the precision of a nonparametric estimator decreases rapidly as the number of continuously distributed components of X increases. This phenomenon is called the 'curse of dimensionality.' As a result

of it, impracticably large samples are usually needed to obtain acceptable estimation precision if X is multidimensional, as it often is in social science applications. For example, a labor economist may want to estimate mean log wages conditional on years of work experience, years of education, and one or more indicators of skill levels, thus making the dimension of X at least 3. (*See* Nonparametric Regression for further discussion of the curse of dimensionality.)

Another problem is that nonparametric estimates can be difficult to display, communicate, and interpret when X is multidimensional. Nonparametric estimates do not have simple analytic forms, so displaying and interpreting them can be difficult. If X is one- or two-dimensional, then the estimate of $E(Y|x)$ can be displayed graphically as in **Figure 1**, but only reduced-dimension projections can be displayed when X has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of the estimate of $E(Y|x)$.

Another problem with nonparametric estimation is that it does not permit extrapolation. That is, it does not provide predictions of $E(Y|x)$ at points x that are outside of the support (or range) of the random variable X. This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, in nonparametric estimation, it can be difficult to impose restrictions suggested by economic or other theory. Matzkin (1994) discusses this issue.

Semiparametric methods permit greater estimation precision than do nonparametric methods when X is multidimensional. In addition, semiparametric estimates are easier to display and interpret than nonparametric ones and provide limited capabilities for extrapolation and imposing restrictions derived from economic or other theory models.

Semiparametric Models

The term 'semiparametric' refers to models in which there is an unknown function in addition to an unknown finite dimensional parameter. For example, the binary response model $P(Y=1|x)=G(\beta'x)$ is semiparametric if the function G and the vector of coefficients β are both treated as unknown quantities. This section describes two semiparametric models of conditional mean functions that are important in applications. The section also describes a related class of models that have no unknown finite-dimensional parameters but, like semiparametric models, mitigate the disadvantages of fully nonparametric models.

In addition to the estimation of conditional mean functions, semiparametric methods can be used to estimate conditional quantile and hazard functions, binary response models in which there is heteroskedasticity of unknown form, transformation models, and censored and truncated mean- and median- regression models, among others. Space does not permit discussion of these models here. Horowitz (1998, 2009), Li and Racine (2007), and Powell (1994) provide more comprehensive treatments in which these models are discussed.

Single-Index Models

In a semiparametric single-index model, the conditional mean function has the form

$$E(Y|x) = G(\beta'x) \qquad [3]$$

where β is an unknown constant vector and G is an unknown function. The quantity $\beta'x$ is called an 'index.' The inferential problem is to estimate G and β from observations of (Y, X). G in eqn [3] is analogous to a link function in a generalized linear model, except in eqn [3] G is unknown and must be estimated.

Equation [3] contains many widely used parametric models as special cases. For example, if G is the identity function, then eqn [3] is a linear model. If G is the cumulative normal or logistic distribution function, then eqn [3] is a binary probit or logit model. When G is unknown, eqn [3] provides a specification that is more flexible than a parametric model but retains many of the desirable features of parametric models, as will now be explained.

One important property of single-index models is that they avoid the curse of dimensionality. This is because the index $\beta'x$ aggregates the dimensions of x, thereby achieving 'dimension reduction.' Consequently, the difference between the estimator of G and the true function can be made to converge to zero at the same rate that would be achieved if $\beta'x$ were observable. Moreover, β can be estimated with the same $n^{-1/2}$ rate of convergence that is achieved in a parametric model. Thus, in terms of the rates of convergence of estimators, a single-index model is as accurate as a parametric model for estimating β and as accurate as a one-dimensional nonparametric model for estimating G. This dimension reduction feature of single-index models gives them a considerable advantage over nonparametric methods in applications where X is multidimensional and the single-index structure is plausible.

A single-index model permits limited extrapolation. Specifically, it yields predictions of $E(Y|x)$ at values of x that are not in the support of X but are in the support of $\beta'X$. Of course, there is a price that must be paid for the ability to extrapolate. A single-index model makes assumptions that are stronger than those of a nonparametric model. These assumptions are testable on the support of X but not outside of it. Thus, extrapolation (unavoidably) relies on untestable assumptions about the behavior of $E(Y|x)$ beyond the support of X.

Before β and G can be estimated, restrictions must be imposed that ensure their identification. That is, β and G must be uniquely determined by the population distribution of (Y, X). Identification of single-index models has been investigated by Ichimura (1993) and, for the special case of binary response models, Manski (1988). It is clear that β is not identified if G is a constant function or there is an exact linear relation among the components of X (perfect multicollinearity). In addition, eqn [3] is observationally equivalent to the model $E(Y|x)=G^*(\gamma+\delta\beta'x)$, where γ and $\delta \neq 0$ are arbitrary and G^* is defined by the relation $G^*(\gamma+\gamma v)=G(v)$ for all v in the support of $\beta'X$. Therefore, β and G are not identified unless restrictions are imposed that uniquely specify γ and δ. The restriction on γ is called 'location normalization' and can be imposed by requiring X to contain

no constant (intercept) component. The restriction on δ is called 'scale normalization.' Scale normalization can be achieved by setting the β coefficient of one component of X equal to one. A further identification requirement is that X must include at least one continuously distributed component whose β coefficient is nonzero. Horowitz (1998, 2009) gives an example that illustrates the need for this requirement. Other more technical identification requirements are discussed by Ichimura (1993) and Manski (1988).

The main estimation challenge in single-index models is estimating β. Given an estimator b_n of β, G can be estimated by carrying out the nonparametric regression of Y on $b_n'X$ (e.g., by using the kernel method described in Section Nonparametric Models). Several estimators of β are available. Ichimura (1993) describes a nonlinear least squares estimator. Klein and Spady (1993) describe a semiparametric maximum likelihood estimator for the case in which Y is binary. These estimators are difficult to compute because they require solving complicated nonlinear optimization problems. Xia et al. (2002) and Xia (2006) discuss computation of the estimator of Ichimura (1993).

Powell et al. (1989) describe a 'density-weighted average derivative estimator' (DWADE) that is non-iterative and easily computed. The DWADE applies when all components of X are continuous random variables. It is based on the relation

$$\beta E[p(X)\partial G(\beta' X)/\partial X] = -2E[Y\partial p(X)/\partial X] \qquad [4]$$

where p is the probability density function of X and the second equality follows from integrating the first by parts. Thus, β can be estimated up to scale by estimating the expression on the right-hand side of the second equality. Powell et al. (1989) show that this can be done by replacing p with a nonparametric estimator and replacing the population expectation E with a sample average.

To achieve an $n^{-1/2}$ rate of convergence of b_n to β, the DWADE of Powell et al. (1989) requires p to have more derivatives when the dimension of X is large than when it is small. A practical consequence of this property is that the accuracy of the DWADE tends to decrease as the dimension of X increases. This problem can be overcome by using an average derivative estimator proposed by Hristache et al. (2001a). See, also, Cattaneo et al. (2014).

Ichimura and Lee (1991) and Hristache et al. (2001b) investigate a multiple index generalization of eqn [3]. Horowitz and Härdle (1996) extend the DWADE to models in which some components of X are discrete. They also give an empirical example that illustrates the usefulness of single-index models. Ichimura and Lee (1991) investigate a multiple index generalization of eqn [3].

Partially Linear Models

In a partially linear model, X is partitioned into two nonoverlapping subvectors, X_1 and X_2. The model has the form

$$E(Y|x_1, x_2) = \beta' x_1 + G(x_2) \qquad [5]$$

where β is an unknown constant vector and G is an unknown function. This model is distinct from the class of single-index models. A single-index model is not partially linear unless G is a linear function. Conversely, a partially linear model is a single-index model only in this case. Stock (1989, 1991) and Engle et al. (1986) illustrate the use of eqn [5] in applications. Identification of β requires the 'exclusion restriction' that none of the components of X_1 are perfectly predictable by components of X_2. When β is identified, it can be estimated with an $n^{-1/2}$ rate of convergence regardless of the dimensions of X_1 and X_2. Thus, the curse of dimensionality is avoided in estimating β.

An estimator of β can be obtained by observing that eqn [5] implies

$$Y - E(Y|x_2) = \beta'[X_1 - E(X_1|x_2)] + U \qquad [6]$$

where U is an unobserved random variable satisfying $E(U|x_1,x_2) = 0$. Robinson (1988) shows that under regularity conditions β can be estimated by applying OLS to eqn [6] after replacing $E(Y|x_2)$ and $E(X_1|x_2)$ with nonparametric estimators. The estimator of β, b_n, converges at rate $n^{-1/2}$ and is asymptotically normally distributed. G can be estimated by carrying out the nonparametric regression of $Y - b_n'X_1$ on X_2. Unlike b_n, the estimator of G suffers from the curse of dimensionality; its rate of convergence decreases as the dimension of X_2 increases.

Härdle et al. (2000) provide a more detailed discussion of partially linear models. Xia and Härdle (2006) consider a version of eqn [5] in which $G(X_2)$ is replaced with the single-index form $G(\gamma' X_2)$, where γ is a constant vector.

Nonparametric Additive Models

Let X have d continuously distributed components that are denoted $X_1, \ldots, X_d$. In a nonparametric additive model of the conditional mean function,

$$E(Y|x) = \mu + f_1(x_1) + \cdots + f_d(x_d) \qquad [7]$$

where μ is a constant and $f_1, \ldots, f_d$ are unknown functions that satisfy a location normalization condition such as

$$\int f_k(v) w_k(v) \mathrm{d}v = 0, \quad k = 1, \ldots, d \qquad [8]$$

where w_k is a nonnegative weight function. An additive model is distinct from a single-index model unless $E(Y|x)$ is a linear function of x. Additive and partially linear models are distinct unless $E(Y|x)$ is partially linear and G in eqn [5] is additive.

An estimator of $f_k (k = 1, \ldots, d)$ can be obtained by observing that eqns [7] and [8] imply

$$f_k(x_k) = \int E(Y|x) w_{-k}(x_{-k}) \, dx_{-k} \qquad [9]$$

where x_{-k} is the vector consisting of all components of x except the kth and w_{-k} is a weight function that satisfies $\int w_{-k}(x_{-k}) \mathrm{d}x_{-k} = 1$. The estimator of f_k is obtained by replacing $E(Y|x)$ on the right-hand side of eqn [9] with nonparametric estimators. Linton and Nielsen (1995) and Linton (1997) present the details of the procedure, which is called 'marginal integration,' and extensions of it. Under suitable conditions, the estimator of f_k converges to the true f_k at rate $n^{-2/5}$ regardless of the dimension of X. Thus, the additive model provides dimension reduction. It also permits extrapolation of $E(Y|x)$ within the

rectangle formed by the supports of the individual components of X. Hastie and Tibshirani (1990) discuss an alternative estimation procedure called 'backfitting.' This procedure is widely used, but its asymptotic properties are not yet well understood.

Linton and Härdle (1996) describe a generalized additive model whose form is

$$E(Y|x) = G[\mu + f_1(x_1) + \cdots + f_k(x_d)] \qquad [10]$$

where $f_1, \ldots, f_d$ are unknown functions and G is a known, strictly increasing (or decreasing) function. Horowitz (2001) and Horowitz and Mammen (2007, 2011) present estimators of a version of eqn [10] in which G is unknown. Both forms of eqn [10] achieve dimension reduction. When G is unknown, eqn [10] nests additive and single-index models and, under certain conditions, partially linear models.

To achieve an $n^{-2/5}$ rate of convergence of the estimators of the f_k's, the marginal integration method requires the f_k's to have more derivatives when the dimension of X is large than when it is small. As with the DWADE, a practical consequence of this property is that the accuracy of the marginal integration estimator tends to decrease as the dimension of X increases. Mammen et al. (1999), Horowitz and Mammen (2004), and Yu et al. (2008) present estimators that overcome this problem. Wang and Yang (2007) and Song and Yang (2010) present estimation methods for time series data.

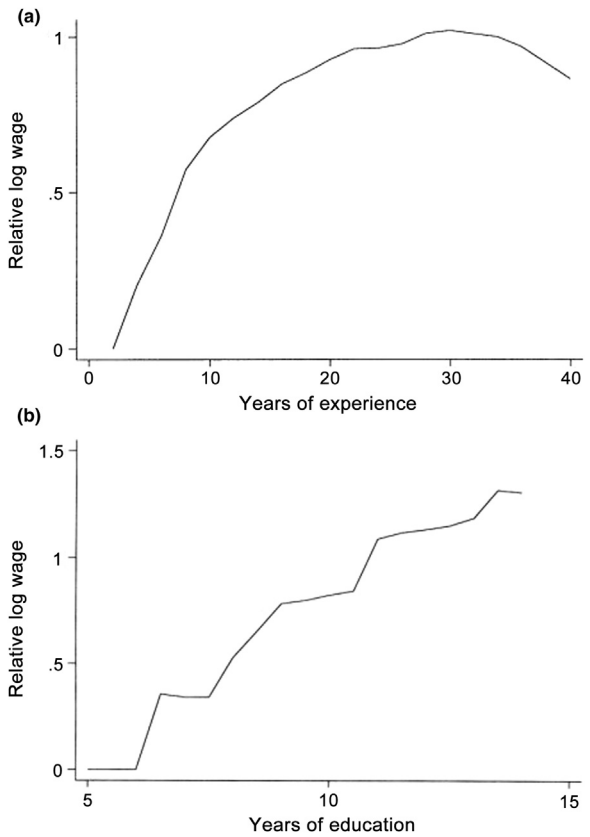

Figure 2 (a) Estimate of f_{EXP} in additive nonparametric model of $E(\log W|\text{EXP,EDUC})$. (b) Estimate of f_{EDUC} in additive nonparametric model of $E(\log W|\text{EXP,EDUC})$.

The use of the nonparametric additive specification (eqn [7]) can be illustrated by estimating the model

$$E(\log W|\text{EXP}, \text{EDUC}) = \mu + f_{\text{EXP}}(\text{EXP}) + f_{\text{EDUC}}(\text{EDUC})$$

where W and EXP are defined as in Section Introduction, and EDUC denotes years of education. The data are taken from the 1993 Current Population Survey and are for white males with 14 or fewer years of education who work full time and live in urban areas of the North Central US. The results are shown in **Figure 2**. The unknown functions f_{EXP} and f_{EDUC} are estimated by the method of Linton and Nielsen (1995) and are normalized so that $f_{\text{EXP}}(2) = f_{\text{EDUC}}(5) = 0$. The estimates of f_{EXP} (**Figure 2(a)**) and f_{EDUC} (**Figure 2(b)**) are nonlinear and differently shaped. Functions f_{EXP} and f_{EDUC} with different shapes cannot be produced by a single-index model, and a lengthy specification search might be needed to find a parametric model that produces the shapes shown in **Figure 2**. Some of the fluctuations of the estimates of f_{EXP} and f_{EDUC} may be artifacts of random sampling error rather than features of $E(\log W|\text{EXP,EDUC})$. However, a more elaborate analysis that takes account of the effects of random sampling error rejects the hypothesis that either function is linear.

Conclusions

This article has described several semiparametric methods for estimating conditional mean functions. These methods relax the restrictive assumptions made by linear and other parametric models, thereby reducing (though not eliminating) the likelihood of seriously misleading inference. The value of semiparametric methods in empirical research has been demonstrated. Their use is likely to increase as their availability in commercial statistical software packages increases.

Bibliography

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.
Cattaneo, M.D., Crump, R.K., Jansson, M., 2014. Small bandwidth asymptotics for density-weighted average derivatives. Econometric Theory 30, 176–200.
Engle, R.F., Granger, C.W.J., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relationship between weather and electricity sales. Journal of the American Statistical Association 81, 310–320.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Chapman & Hall, London.
Goldberger, A.S., 1998. Introductory Econometrics. Harvard University Press, Cambridge, MA.
Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge, UK.
Härdle, W., Liang, H., Gao, J., 2000. Partially Linear Models. Springer, New York.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.
Horowitz, J.L., 1993. Semiparametric and nonparametric estimation of quantal response models. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, vol. II. Elsevier, Amsterdam, pp. 45–72.
Horowitz, J.L., 1998. Semiparametric Methods in Econometrics. Springer, New York.
Horowitz, J.L., 2001. Nonparametric estimation of a generalized additive model with an unknown link function. Econometrica 69, 599–631.
Horowitz, J.L., 2009. Semiparametric and Nonparametric Methods in Econometrics. Springer, New York.
Horowitz, J.L., Härdle, W., 1996. Direct semiparametric estimation of single-index models with discrete covariates. Journal of the American Statistical Association 91, 1632–1640.

Horowitz, J.L., Mammen, E., 2004. Nonparametric estimation of an additive model with a link function. Annals of Statistics 32, 2412–2443.

Horowitz, J.L., Mammen, E., 2007. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. Annals of Statistics 35, 2589–2619.

Horowitz, J.L., Mammen, E., 2011. Oracle-efficient nonparametric estimation of an additive model with an unknown link function. Econometric Theory 27, 582–608.

Hristache, M., Juditsky, A., Spokoiny, V., 2001a. Direct estimation of the index coefficient in a single-index model. Annals of Statistics 29, 595–623.

Hristache, M., Juditsky, A., Polzehl, J., Spokoiny, V., 2001b. Structure adaptive approach for dimension reduction. Annals of Statistics 29, 1537–1566.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58, 71–120.

Ichimura, H., Lee, L.-F., 1991. Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, UK, pp. 3–49.

Klein, R.W., Spady, R.H., 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 387–421.

Li, Q., Racine, J.S., 2007. Nonparametric Econometrics. Princeton University Press, Princeton.

Linton, O.B., 1997. Efficient estimation of additive nonparametric regression models. Biometrika 84, 469–473.

Linton, O.B., Härdle, W., 1996. Estimating additive regression models with known links. Biometrika 83, 529–540.

Linton, O.B., Nielsen, J.P., 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika 82, 93–100.

Mammen, E., Linton, O., Nielsen, J., 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Annals of Statistics 27, 1443–1490.

Manski, C.F., 1988. Identification of binary response models. Journal of the American Statistical Association 83, 729–738.

Matzkin, R.L., 1994. Restrictions of economic theory in nonparametric methods. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam, pp. 2523–2558.

Powell, J.L., 1994. Estimation of semiparametric models. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam, pp. 2444–2521.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 51, 1403–1430.

Robinson, P.M., 1988. Root-N-consistent semiparametric regression. Econometrica 56, 931–954.

Song, Q., Yang, L., 2010. Oracally efficient spline smoothing of nonlinear additive autoregression models with simultaneous confidence band. Journal of Multivariate Analysis 101, 2008–2025.

Stock, J.H., 1989. Nonparametric policy analysis. Journal of the American Statistical Association 84, 567–575.

Stock, J.H., 1991. Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. In: Barnett, W.A., Powell, J., Tauchen, G. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, UK, pp. 77–98.

Wang, L., Yang, L., 2007. Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Annals of Statistics 35, 2474–2503.

Xia, Y., 2006. Asymptotic distributions for two estimators of the single-index model. Econometric Theory 22, 1112–1137.

Xia, Y., Härdle, W., 2006. Semi-parametric estimation of partially linear single-index models. Journal of Multivariate Analysis 97, 1162–1184.

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X., 2002. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society, Series B 64 (Part 3), 364–410.

Yu, K., Park, B.U., Mammen, E., 2008. Smooth backfitting in generalized additive models. Annals of Statistics 36, 228–260.

Sensation and Perception: Direct Scaling

Martha Teghtsoonian, Smith College, Northampton, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Two techniques for the direct scaling of perceived intensity – category and magnitude judgments – are described. The major findings associated with direct scaling are the establishment of Stevens's power law of perceived intensity, the establishment of characteristic exponents for the several sensory continua, and the discovery of the transitivity of exponents. Major problems of direct scaling, and of the body of work based on it, include the failure of category and magnitude scales to agree, the instability of exponents, and the difficulty of integrating results of scaling suprathreshold magnitudes with data on sensory thresholds. Possible resolutions of these difficulties are presented.

One aim of psychophysics is to measure the perceived intensity of any given stimulus level. We easily hear that a tone at a given sound pressure level has a particular loudness and that a tone at a higher level sounds louder still. The measurement problem lies in the question 'How much louder?' Direct scaling is a particular way of answering that question: the observer is asked to assign numbers corresponding to the perceived intensities, the subjective magnitudes, of given stimulus intensities.

History

Fechner's Solution

G. Fechner proposed a solution to the question of 'How loud?', albeit an indirect one, in 1860 (Boring, 1950). He accepted Weber's law – that the difference threshold (the physical size of a difference in intensity needed to tell one signal from another) grows in proportion to signal intensity, determining a constant of proportionality, the Weber fraction, unique to each sensory continuum. Fechner assumed that all difference thresholds on a given sensory continuum are subjectively equal and therefore can serve as the unit of measurement for scales of sensation. He concluded that subjective magnitude grows by constant differences as stimulus magnitude grows by constant ratios, a logarithmic relation now known as Fechner's law.

Category Rating and Magnitude Scaling

In Fechner's laboratory, there also evolved a direct approach to measuring subjective magnitude. This was the method of absolute judgment (later called category rating), in which the subject assigns stimuli to categories according to some subjective aspect. Originally designed to study esthetic judgment, it was adapted by 1920 to study perceived intensity; early findings using this method supported Fechner's law.

In the 1950s, S. S. Stevens proposed another version of direct scaling in which the subject estimates the apparent magnitude of a signal by assigning to it a numerical value – hence, magnitude estimation.

Methods

The direct scaling methods require an observer to judge a series of stimuli varying along (at least) one dimension. For instance, the experimenter might present a pure tone at several sound pressure levels covering a 1000:1 range and request judgments of loudness, or present a disk of light at several luminances covering a 10:1 range and request judgments of brightness.

Category Rating

The observer is instructed to assign each stimulus to a category, which may be designated descriptively (very soft, soft,…, very loud) or numerically (1, 2,…, 7); the assignments should create subjectively equal intervals. The allowable response values, including the extreme values, are specified by the experimenter. Usually, the number of categories ranges from 5 to 11, although smaller and larger values are sometimes used. Typically, the stimulus range does not exceed 10:1, although it may do so.

Magnitude Estimation and Production

In magnitude estimation, the observer is instructed to assign to each stimulus a (positive) number that reflects its apparent magnitude. There are no limits on allowable response values. Originally, one stimulus value was designated as standard and a numerical value assigned to it as modulus. Later, however, the use of designated standard and assigned modulus was abandoned, and the observer was instructed to choose any numerals appropriate to represent apparent magnitudes. The stimulus range may be 100:1 and is often greater.

In magnitude production, stimulus and judgmental continua are interchanged: the experimenter presents a set of numbers, and the observer produces values on some intensive continuum to effect a subjective match for each presented number.

Cross-Modal Matching

The use of numbers, which provoked many objections, can be avoided by instructing the observer to match intensities on one physical continuum directly to intensities on another (Stevens and Mack, 1959). For example, the observer may adjust luminances to match sound pressure levels so that the brightness of a disk of light seems equivalent to the loudness

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43087-4

of a tone, or the observer may produce handgrip forces to match odor intensities so that the perceived effort is equivalent to the strength of smell.

Achievements of Direct Scaling

A Quantitative Phenomenology

These methods made possible a quantitative phenomenology of the various sensory systems. For example, in hearing, the relevance of variables such as frequency and intensity to the ability of the observer to detect a sound or to tell the difference between two sounds had long been known. Furthermore, by having the observer match tones of varying frequency and intensity to a fixed reference tone, equal loudness contours were established. What remained unknown were the loudness relations among these contours. If it can be assumed that the numerals used by observers correspond directly to the number scale, then magnitude estimation provides a reference scale: we can say that one tone sounds twice, or 10 times, or half as loud as another. Magnitude estimation has been used to map the effects of variables known to influence perception for loudness, brightness, tastes and smells, and tactile sensations, as well as internal states such as perceived effort, fatigue, and satiety (Stevens, 1975). It has also been used to study continua without a physical measure (e.g., seriousness of crime, severity of punishment, and their relationship) (Sellin and Wolfgang, 1964).

Category rating has been employed to study the cognitive integration of perceptual dimensions (Anderson, 1981). A version of the category scale, the Borg scale (Borg, 1998) has seen wide use in practical settings.

Intermodal Comparisons

Direct scaling also provides the capability to compare and contrast phenomena that occur in several modalities, such as temporal and spatial summation, sensory adaptation and recovery, and spatial inhibition (brightness contrast and auditory masking), at levels above absolute threshold (Marks, 1974). For example, the smell associated with a persisting olfactory stimulus of constant intensity diminishes in strength over time (adaptation). That diminution can be assessed by asking an observer to assign numbers to the apparent magnitude of the smell at different elapsed times, e.g., after 5, 10, 30, 60 s,…; in this way, the course of sensory adaptation can be traced for this stimulus and for others of differing initial intensities. When the procedure is repeated in a different modality, the parameters of the adaptation curves can be compared.

Stevens's Power Law

The third major achievement of direct scaling is the discovery that, to at least a first approximation, equal stimulus ratios produce equal judgmental ratios. This nearly universal relation is called Stevens's power law, after S. S. Stevens who established it: subjective magnitude is proportional to stimulus intensity raised to some power. Further, that exponent takes distinctive values for each stimulus continuum, ranging from 0.3 for luminance to 2.0 or more for electric shock. When the exponent is less than 1.0, the sensory system in question is compressive – a wide physical range becomes a much smaller perceptual range. When the exponent is greater than 1.0, the sensory system is expansive – a small physical range becomes a much larger perceptual range.

Furthermore, these magnitude exponents predict cross-modal matching exponents: if continua A and B have magnitude exponents a and b, and if a cross-modal match of B to A is obtained, the resulting exponent is a/b. Indeed, the cross-modal exponents for a variety of continua are connected in a transitive network at the level of the individual observer (Daning, 1983).

Problems of Direct Scaling

The Instability of the Exponent

Critics have characterized the search for exponents characteristic of each sensory continuum as futile, since obtained exponent values are (1) so dependent on contextual factors (e.g., range of stimuli, sequence in which stimuli occur, and numerals chosen by the observer) (Lockhead, 1992) and (2) subject to the so-called regression effect (the value of the exponent relating two continua depends on which serves as the target and which as the matching continuum). A reply to these critics (Teghtsoonian and Teghtsoonian, 2003) argued that there are several answers to these criticisms. First, some of these objections, on closer examination, are not supported by data. Second, experiments can identify the sources of some biases and show ways in which they can be controlled. Third, the understanding of context effects and biases can be extended by models of judgmental data. For example, DeCarlo (2005), in an examination of the regression effect, provided theoretical and empirical support for the conjecture that the geometric mean of the two exponents in a cross-modal matching experiment gives an unbiased estimate of the true exponent.

Category vs. Ratio Scales

Although both category rating and magnitude estimation purport to give scales of sensation magnitude, they do not, in general, agree with each other: both scales can be characterized as power functions, but the exponents are not the same. Much discussion has centered on whether subjects can make two kinds of judgments or whether the experimenter transforms a single kind of magnitude judgment by treating it as a ratio or an interval. However, an experimental determination of the sources of variance in the type of scale produced showed that instructions to judge either ratios or intervals accounted for almost none of the variance; much more important were such methodological variables as whether judgmental end points were assigned or free and whether the range of stimuli was small or large. With free end points and a large range, category judgment and magnitude estimation produce the same results, and the obtained power functions have stable exponents characteristic of magnitude estimation (Montgomery, 1975).

Local vs. Global Psychophysics

A long-standing conundrum in psychophysics has been the relation between the local (thresholds, both absolute and

differential) and the global (measurements of subjective magnitude at levels above the absolute threshold) (Luce and Krumhansl, 1988). Fechner believed he had combined the two by using difference thresholds as subjective units to yield a logarithmic scale of sensation. Stevens (1961) proposed, while honoring Fechner, to repeal his law and to substitute a power scale of sensation; he believed that one could not derive measures of sensation magnitude from threshold determinations. Teghtsoonian (1971, 2012) reconciled these two positions by arguing that both difference thresholds and power law exponents are indices of dynamic range (the range from the smallest stimulus discernible to the greatest stimulus to which the sensory system can respond differentially). He proposed that there is a common scale of sensory magnitude for all perceptual continua and that the observer's dynamic range for each continuum maps completely onto that common scale. For example, the least sound intensity experienced may have the same subjective magnitude as the least luminance, and the greatest sound intensity to which the auditory system responds may have the same subjective magnitude as the greatest luminance. Thus the mapping of widely divergent dynamic ranges for the several perceptual continua onto a single subjective magnitude range will determine their power law exponents, and the mapping of widely divergent difference thresholds onto a single subjective difference threshold will determine their Weber fractions.

Use of Numerals as Responses

Critics have questioned whether the numerals assigned by the observer can be taken to correspond to the number scale. Narens (1996) and Luce (2002) have proposed axiomatizations of direct scaling (magnitude estimation and ratio production, respectively) and have shown that two conditions – commutativity and multiplicativity – must be satisfied by the observers' use of numerals if those numerals can be taken to establish a number scale. Empirical studies have shown in at least two cases that scaling judgments are commutative but not multiplicative (Zimmer, 2005).

These findings do not vitiate the achievements outlined above – the establishment of a quantitative phenomenology, the possibility of cross-continuum comparison of perceptual phenomena, and the ubiquity of Stevens's power law. They leave open the question with which we began: Can the observer's responses tell us how much louder (or brighter or heavier) one stimulus seems than another? We do not yet know how to measure sensation, and it may be that it is not measurable.

See also: Axiomatic Theories; Measurement Theory: Conjoint; Psychophysical Theory and Laws, History of; Signal Detection Theory; Stevens, Stanley Smith (1906–73).

Bibliography

Anderson, N.H., 1981. Foundations of Information Integration Theory. Academic Press, New York.
Borg, G., 1998. Borg's Perceived Exertion and Pain Scales. Human Kinetics, Champaign, IL.
Boring, E.G., 1950. A History of Experimental Psychology, second ed. Appleton-Century-Crofts, New York.
Daning, R., 1983. Intraindividual consistencies in cross-modal matching across several continua. Perception & Psychophysics 33, 516–522.
DeCarlo, L.T., 2005. On bias in magnitude scaling and some conjectures of Stevens. Perception & Psychophysics 67, 886–896.
Lockhead, G.R., 1992. Psychophysical scaling: judgments of attributes or objects? Behavioral and Brain Sciences 15, 543–558.
Luce, R.D., Krumhansl, C.L., 1988. Measurement, scaling, and psychophysics. In: Atkinson, R.C., Herrnstein, R.J., Lindzey, G., Luce, R.D. (Eds.), Stevens' Handbook of Experimental Psychology Perception and Motivation, second ed. vol. 1. Wiley, New York.
Luce, R.D., 2002. A psychophysical theory of intensity proportions, joint presentations, and matches. Psychological Review 109, 520–532.
Marks, L.E., 1974. Sensory Processes: The New Psychophysics. Academic Press, New York and London.
Montgomery, H., 1975. Direct estimation: effect of methodological factors on scale type. Scandinavian Journal of Psychology 16, 19–29.
Narens, L., 1996. A theory of magnitude estimation. Journal of Mathematical Psychology 40, 109–129.
Sellin, J.T., Wolfgang, M.E., 1964. The Measurement of Delinquency. Wiley, New York.
Stevens, J.C., Mack, J.D., 1959. Scales of apparent force. Journal of Experimental Psychology 58, 405–413.
Stevens, S.S., 1961. To honor Fechner and repeal his law. Science 133, 80–86.
Stevens, S.S., 1975. Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects. Wiley, New York.
Teghtsoonian, R., 1971. On the exponents in Stevens's law and the constant in Ekman's law. Psychological Review 78, 71–80.
Teghtsoonian, M., Teghtsoonian, R., 2003. Putting context effects into context. In: Berglund, B., Borg, Elisabet (Eds.), Fechner Day 2003. Proceedings of the Nineteenth Annual Meeting of the International Society for Psychophysics. Cyprus, Lanarca Bay.
Teghtsoonian, R. 2012. The standard model for perceived magnitude: a framework for (almost) everything known about it. American Journal of Psychology, 125, 165–174.
Zimmer, K., 2005. Examining the validity of numerical ratios in loudness fractionation. Perception & Psychophysics 67, 369–379.

Sensation Seeking: Behavioral Expressions and Biosocial Bases

Marvin Zuckerman, University of Delaware, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of questionnaire scales to measure sensation seeking, first as a general trait and then as one with two or four facets, is described. Behavioral correlates of the trait include: volunteering for unusual or risky experiences, exceptional dangerous sports, fast and reckless driving, variety of sexual partners and activities, smoking, drinking and drugs, and risky or stressful vocations. Sensation seeking also influences preferences in art, media, music, movies, and television with a preference for novel, intense, and arousing themes like sex and violence.

There is a strong genetic influence in sensation seeking, compared to other personality traits. Biochemical correlates include testosterone, dopamine, and serotonin, and the genes and enzymes that regulate them. Sensation seeking also correlates with cortical reactivity to intensity of stimuli. Parallel studies of other species suggest that sensation seeking has evolved from a strong approach trait, including exploration, and a weak behavioral inhibition mechanism. Similar biological mechanisms underlie the behavioral dispositions in human and nonhuman species.

Definition

Most definitions of personality traits are based on little more than a theoretical concept and the content of the items making up the scale. However the ultimate definition of a construct should go beyond an operational definition tied to a single method of assessment. It should be based upon the body of research relating the construct to external measures derived from the theory of the construct. The construct definition may change and evolve as we test the *construct validity* of our initial and subsequent measures (Cronbach and Meehl, 1955).

Sensation seeking (SS) was initially based on Hebb's (1955) model of an optimal level of cortical arousal (OLA) for learning and performance with the added postulation of individual differences in reticulo-cortical reactivity (Zuckerman, 1969). Subsequent research over the years led to a broader construct with a different biological basis (Zuckerman, 1979, 1994, 2007). The OLA theory was based on the relationship between stimulus intensity and cortical arousal. Newer research on the content of the predictor scales and the predicted phenomena led to the inclusion of stimulus novelty as well as intensity in the qualities of stimuli motivating SS and avoidance. The current behavioral definition of the trait is:

> Sensation seeking is a trait defined by the seeking of varied, novel, complex, and intense sensations and experiences, and the willingness to take physical, social, legal, and financial risks for the sake of such experience.
>
> Zuckerman, 1994: p. 27

Three books (Zuckerman, 1979, 1994, 2007) describe the methods of assessment of the trait and the behavioral and psychobiological research on the trait. The last of these (Zuckerman, 2007) emphasizes the research on risky behavior. Not all behavioral characteristic of high sensation seekers is risky: for instance, their seeking of vicarious experience in movies, art, music, and media with the qualities of stimuli they find pleasurably exciting even if the content is negative (Zuckerman, 2006). The biosocial basis of SS has evolved and changed with new methods and findings. Aside from the three SS books it is described in two other volumes (Zuckerman, 1991, 2005) on the psychobiology of personality in general.

Different names have been given to traits assessed by other measures but quite similar to SS in theory and content. Examples are: change seeking, stimulus variation seeking, excitement seeking, fun seeking, venturesomeness, and novelty seeking (NS). Lack of attention to *convergent and discriminant validation* (Campbell and Fiske, 1959) has led to a needless proliferation of constructs primarily distinguished by their labels rather than their substance. Of course there is always some variation due to specific measures but this method variance is generally not of much interest in personality science. However facets within a major trait may be of theoretical interest and have differential prediction. Such facets have been developed for SS and will be described in the next section.

Development of Sensation Seeking Scales

Form II

Items were written for the first experimental form of the sensation seeking scale (SSS) based on the theory of individual differences in OLA or intensity of stimulation, although some reflecting a need for novelty were also included, as were items endorsing a desire to engage in risky sports or activities (thrill seeking). Factor analyses of items yielded a general sensation seeking (GSS) factor defined by the first unrotated factor. The GSS was the basis of the SSS II (Zuckerman et al., 1964). The forced-choice item response form was used in an attempt to control social desirability which was considered an important determinant of test results at that time.

Forms IV and V

In the development of form II we noticed some indications of additional subfactors but there were not enough items in form I to ascertain their content through rotation of factors. Experimental form III included a wider diversity of items although we

were not certain what content form would describe these factors. Rotation did show four reliable subfactors or facets constituting SSS form IV (Zuckerman, 1971).

1. Thrill and adventure seeking (TAS) items express a desire to seek excitement through unusual or exceptional sports such as skydiving or scuba diving.
2. Experience seeking (ES) items describe a desire to have novel experiences through travel, drugs, and a spontaneous, nonconforming lifestyle with unusual, exciting friends.
3. Disinhibition (Dis) items refer to seeking variety and arousal through social and sexual stimulation, lively parties, and social drinking.
4. Boredom susceptibility (BS) items express an aversion to lack of change and variety of experiences and boring people.

Form IV also included the GSS scale from form II as a measure of the overall factor.

Form V was developed from the items in form IV, selecting the highest loading 10 items from each rotated factor in new samples of American and British subjects (Zuckerman et al., 1978). The factors were reliable and there was little or no loss of internal reliability in the shortened scales. Instead of the GSS, a general factor was assessed as the total of the scores on the four subscales. The four-factor structure of SSS using the original pool of items has been consistently replicated in other countries and languages (Zuckerman, 1994). SSS developed from analyses of the broad personality questionnaires.

Impulsive Sensation Seeking

Factor analyses of scales measuring a broad sample of traits associated with biological factors in personality consistently yielded five major factors: sociability, neuroticism, impulsive sensation seeking (ImpSS), aggression, and activity, comprising the Zuckerman–Kuhlman Personality Questionnaire (ZKPQ; Zuckerman et al., 1993).

The ImpSS scale differed in three important ways from the SSS II to V scales. First, the response format was changed from the forced-choice form to a simpler true–false form. Second, references to specific activities like drinking, drugs, and sex were eliminated and converted to more general statements. The specific items had been confounded in the many studies of actual behavior like drinking, taking drugs, and sexual experience. Third, because of the limited sampling of items specific to the four subscales of the SSS these four factors could not appear. Instead, SSS items were highly related to a type of impulsivity: acting on impulse without planning or anticipation of risk. The Imp and SS components of ImpSS may be scored separately or together. The Total scale has 19 items.

ImpSS from the Cross-Cultural Shortened Form of the ZKPQ (ZKPQ-50-cc)

Aluja et al. (2006) developed a short form of the ZKPQ in four countries, each one using a different language (English, French, German, and Spanish). Reliable and similar factors were found across languages. The 10-item ImpSS scale alpha reliabilities ranged from 0.72 to 0.74 in the four samples. The shortened ImpSS scale contains eight SS items but only two Imp items; therefore these two components should not be scored separately.

The SSS from the Zuckerman–Kuhlman–Aluja Personality Questionnaire

Aluja et al. (2010) developed a new form of the ZKPQ containing four facets for each of the five major factors. Many new items were written for the experimental form but as with the ZKPQ not using items with specific content for drinking, drugs, or sex. The item response format was a 4-point Likert scale. Possible facets for the SS factor included the four factors from the original SSS plus an impulsivity facet, like the one in the ZKPQ and an antisocial SS facet. Factor analyses of items within factors yielded three of the original SSS facets (TAS, ES, and Dis) but the fourth facet was a combination of BS and impulsivity items (BS/Imp). Antisocial SS did not emerge as a fifth facet. Factor analyses of the 20 facets yielded the postulated structure with all four facets loading highest on the appropriate factor and much lower on the other four factors.

Up to now the majority of research on sensation seeking has been done with the SSS-V and its subscales. However since the SSS in the ZKPQ contains substantially the same four subscales without the confounding content of the SSS-V and the sometimes difficult forced-choice response format, it is recommended that future research consider using the 40-item SSS from the Zuckerman–Kuhlman–Aluja Personality Questionnaire (ZKA-PQ).

Behavioral Expressions

Volunteering

The first studies of sensation seeking involved the prediction of reactions to the experimental situation of sensory deprivation (SD) (Zuckerman, 1969). The theory of OLA predicted that persons with high needs for stimulation would be more stressed by a situation where they were deprived of external stimulation. High sensation seekers did become more bored and restless during SD, but an unpredicted finding was that the volunteers for SD were disproportionally drawn from the high SS subjects. Asked why they volunteered, they replied that they expected to have hallucinations and experience unusual mental states. The need for novelty was more important than the need for intensity. In subsequent studies of volunteering, high sensation seekers tended to volunteer for other experiments involving hypnosis, taking hallucinogenic drugs, and encounter groups, but not for more mundane experiments in learning or social psychology.

Within the U.S. Army (Jobe et al., 1983) and Israeli Army and population, volunteers for risky missions or jobs were high sensation seekers. Among soldiers exposed to combat in Israel during the Yom Kippur War, those decorated for bravery in combat were higher than other battle veterans in sensation seeking (Neria et al., 2000). The sensation seeking trait plays a role beyond patriotism, loyalty to comrades, and other motives of heroism. What of physical risky behavior outside of war?

Sports

Replicated studies have shown that those who engage in exceptional and risky sports like mountain climbing, skydiving, scuba diving, whitewater canoeing, and hang gliding are high sensation seekers, but the larger number who engage in ordinary sports like football, baseball, and tennis, and physical education students are only average on this trait. Those who engage in sports requiring dedicated training or endurance like golf and marathon running tend to be low in sensation seeking. The sports that attract high sensation seekers involve novel or unusual and sometimes intense sensations and experiences outside the realm of ordinary life activities. However there is one more common life activity where high sensation seekers find expression despite the physical and legal risks involved.

Driving

Both self-report and observational studies show that high sensation seekers tend to drive recklessly at high speeds, following too closely to cars in front of them, driving while intoxicated, driving with frequent and abrupt lane changes, and generally aggressive driving. From a review of 40 studies, Jonah (1997) concluded that sensation seeking was significantly related to such indices of reckless driving in the great majority of these studies. Observations of drivers (Burns and Wilde, 1995; Heino, 1996) over preselected routes show that high sensation seekers do drive faster and more recklessly than low sensation seekers. Burns and Wilde also showed a correlation with records of past violations.

Sex

Risk-taking in sexual behavior is the source of both extreme pleasure and pain in human relationships. Novelty and variety enhance arousal, and familiarity tends to reduce it. High sensation seekers tend to seek sensation through a variety of sexual partners and activities (Zuckerman et al., 1976). Hoyle et al. (2000) reviewed all studies (up to 1999) relating sexual risk-taking to major personality traits. Sexual risk-taking was defined by number of partners, unprotected sex, and high-risk sexual encounters (like sex with a stranger). Sensation seeking was related to all three categories of risk, and correlations with the SSS were higher than for nearly all other personality traits.

Unprotected sex with many partners has always posed a health risk in terms of sexually transmitted diseases, but beginning in the mid-1990s the AIDS epidemic posed a new risk, particularly for men who have sex with other men. Kalichman et al. (1994) devised two types of SSS to study this group of men. The sexual sensation seeking scale (SSSS) items describe seeking intense sensation through sex although they are not specific to homosexual or heterosexual sex. The nonsexual experience seeking scale (NSES) is a GSS or ES scale with no reference to sex. Three high-risk behaviors were predicted using these scales: many sexual partners, anal sex without condoms, and the use of alcohol or drugs just prior to or during sex (Kalichman et al., 1994). The studies are summarized in Zuckerman (2007, Table 5.2: p. 161). In most studies both the SSSS and the NSES correlated with the three risk factors. In several other studies the General, Dis, and BS SSS also correlated with the three risk behaviors. The nonuse of condoms during anal sex is motivated by the need for more intense stimulation, and the need for many partners by the arousal produced by novelty. The use of alcohol before sex reduces inhibition by the awareness of risk and some drugs used during sex also enhance the intensity of arousal. Sensation seeking can be a deadly trait for those engaging in such risk-taking during sex. Drugs can themselves be a major reward in sensation seeking as next described.

Smoking, Drinking, and Drugs

Correlations among different types of risk-taking show a central intercorrelated core of smoking, drinking, drugs, and sex (Zuckerman and Kuhlman, 2000). The relation between sensation seeking and smoking was first studied by college students in the 1970s (Zuckerman, Tushup, and Finner). Since then the association has been replicated in American college students, the broader American population, Norwegian army recruits and high-school students, the Dutch general population, and Israeli adolescents (see Zuckerman, 1994, 2007, for descriptions of particular studies). Wider publicity on the serious health risks of smoking has reduced the overall incidence of smoking in the younger generations, but the relationship with sensation seeking persists from the 1970s to the present. Recognition of high sensations seekers, as those most involved in smoking, has guided those who design ads and interventions to prevent and discourage smoking in young adolescents (Donohew et al., 2004).

Several studies of preadolescents (ages 10–11) show that sensation seeking, as part of an undercontrolled, impulsive personality, predicts drinking in adolescent and young adult years. Sensation seeking is associated with early onset of drinking, frequency and quantity of drinking, binge drinking, alcohol abuse, and dependence. It is also associated with antisocial tendencies in alcoholics. Hittner and Swickert (2000) did a meta-analysis of 61 studies of the relationship between alcohol use and sensation seeking. The effect size for the SSS Total score was modest but highly significant, but the effect was highest for the Dis subscale ($r = 0.40$). There is a problem in confounding between some items in the Dis scale and reports of drinking but the effect size was the same in studies which removed these items and studies which did not do so (e.g., Finn and Hall, 2004). Studies using the ZKPQ ImpSS scale, which does not contain content confounding items, yielded an effect size as large as that for Dis in the SSS.

The relationship between personality traits and drinking may be mediated by motives for drinking. Cooper et al. (1995) developed a four-factor model for drinking motives: (1) social drinking (parties, celebrations); (2) coping (reduction of anxiety and depression in dealing with problems); (3) enhancement (increasing pleasure and excitement by getting 'high'); and (4) conformity (fitting in and being accepted and liked in a drinking group).

Sociability E is associated primarily with social motivation, N with coping, and sensation or excitement seeking to enhancement motives. Coping and enhancement are both related to heavy drinking but coping is more associated with drinking problems.

Alcohol and marijuana are the earliest and most widely used substances but because marijuana provides some novel sensations and experiences it is more widely used among high sensation seekers. Segal et al. (1980) studied substance use in large samples of college students and naval personnel. Among college students, marijuana users were higher on Dis and ES (confounding items removed from both) than both abstainers and those who used alcohol only but they were not different from those who used other drugs as well. Among the naval personnel, the major difference was between marijuana users and polydrug users rather than alcohol-only users. Later studies of college, community, and drug abusers show that sensation seeking is more related to the number of drugs used than to any specific drug.

Although heavy drinking and marijuana use are the most common forms of substance use and abuse, some high sensation seekers go on to harder drugs including stimulants (cocaine, amphetamine, methamphetamine), suppressants (morphine, opium, heroin, methadone, narcotics), and psychedelics (LSD, MDMA, or 'ecstasy'). Although the different classes of drug have different effects, all are related to sensation seeking as motive and trait (Zuckerman, 2012). Adams et al. (2003) found that NS (a scale highly correlated with ImpSS) was higher in stimulant users than in substance abusers who preferred sedatives such as opiates or narcotics. However it has been repeatedly shown that sensation seeking is more highly related to number of drugs used than specific drugs used. Stimulant use is more highly related to polydrug use than heroin use. However even heroin users value the 'rush,' a highly stimulant effect of the drug immediately after injection.

Sensation seeking predicts outcome of treatment in cocaine addiction (Ball, 1995; Patkar et al., 2004). High SSS is related to poor attendance, high dropout rates, and failed urine tests. Apparently treatment cannot supply a substitute for the positive arousal supplied by drugs. The use of psychedelics is related to the need for novel experience. The effects are different every time they are used and include hallucinatory-type experiences. Segal and Singer (1976) asked drug users about their motives for using particular drugs. Nearly three-quarters of LSD users said that they used the drug to provide a new experience. A fifth of the users said that they were motivated by curiosity. A large study in Europe found that ecstasy users were higher than nonusers on the ImpSS scale (Benschop et al., 2003).

Vocations

Firefighters tend to score high on the SSS. There are periodic periods of high excitement and risk in firefighting. Police and military personnel are average in sensation seeking but, as discussed previously, those among them who volunteer for highly risky jobs or engage in risky behavior like car pursuits are high sensation seekers.

Stressful work, even that not involving personal risk, seems to attract high sensation seekers but is aversive to low sensation seekers. Thus air-traffic controllers have periods of high stress when many flights are competing for space alternating with low activity periods (Musolino and Hershenson, 1977). There is a danger that their high BS during the nonstress periods could lead to inattention posing a risk for air-flight passengers. Medical personnel who choose to work in emergency rooms and rape crisis counselors also tend to score high on sensation seeking in jobs which are sometimes highly stressful.

Entertainment Preferences

Many choices of high sensation seekers have no risk but reflect the basic attraction to sensations and experiences (vicarious) that reflect their taste for novelty, intensity of sensation and arousal, and complexity (Zuckerman, 2006). In movies and television, and art and photography, they are attracted to erotic and violent content. In art they like expressionist, abstract, surreal styles, and even the surprise and humor of Pop art. They also need variety and are quickly bored as indicated by channel switching when watching television (Perse, 1996; Schierman and Rowland, 1985). In music, the high sensation seeker tends to favor the intensity of rock music. These preferences cross the auditory and visual arts. For instance, preferences for violent-abstract painting and hard rock music at one pole and realistic painting and 'easy-listening,' quiet melodic music at the other constitute a dimension related to sensation seeking, particularly disinhibition (Rawlings et al., 2000).

Biological Influences

In the first half of the twentieth century most personality theorists tended to attribute variance in personality traits to parental treatment of children and later life experiences. This bias in what came to be called the 'nature versus nurture' debate was not based upon research but reflected the psychodynamic influences among most theorists and their use of anecdotal, retrospective case history approaches. Behavior genetics, first applied in the area of intelligence, offered a statistical method for examining the relative influences of nature (genetic) and environment while holding the other constant. The method was largely based on comparisons of degrees of relationship between siblings within families. Twins are of particular value because identical twins share all of their genes whereas fraternal twins only share half of the genes that vary within populations, but both types of twins are presumed to share a similar family environment. This allows for a comparison of genetic influence controlling the influence of shared environment.

Genetics

Johnson et al. (2008) did a gigantic meta-analysis for over 50 years of biometric studies organized around the Big-Five trait classification (McCrae and Costa, 2003). There are three surprises in the results. One is the strong influence of genetics as given in the heritability estimates in the five major traits as calculated from studies of twins, whether raised together or adopted into different families and raised without contact during their formative years. Genetic influences account for nearly half of the total variance for most of the traits. The second is the weak influence of shared environment, including the shared family. Environment is important in the remaining part of the variance but it is the nonshared environment or experiences that specifically affect one of the twins and not the other. The third is the narrow range of

variation in results of studies. Some had suggested that some traits are primarily determined by genetics whereas others are primarily due to the family environment. The results showed a narrow range of variation in heritability of specific traits, between 40% and 50%.

In this context the heritability for sensation seeking lies at the high end of the range for personality traits for the total score on SSS-V heritability was 58% (Fulker et al., 1980). For twins adopted soon after birth and raised in different families without contact, the heritability was 59% (Hur and Bouchard, 1997; Lykken, personal communication, 1992). Since the twins were raised in different environments, the lack of difference in heritability indicates no evidence of an effect for shared environment, Eysenck (1983). Hur and Bouchard (1997) and Koopmans et al. (1995) analyzed the heritabilities of the subscales of the SSS-V. Across all three studies, the mean heritabilities for Dis, TAS, and ES ranged from 50% to 57%. BS was lower, 43%, probably because of the lower reliability of this subscale. Except for BS, the heritabilities of the subscales are close to those for the Total score.

Molecular Genetics is the study of specific gene variants. The evidence of moderate to high heritabilities for personality traits using the methods of behavior genetics suggested that it might be possible to find a few genes that were strongly related to each trait. No one expected only one gene to be related to each trait because the studies showed additive genetic effects for most traits.

The first successful study was an association between the trait of NS (highly related to sensation seeking, Zuckerman and Cloninger,1996) and the dopamine receptor 4 gene (DRD4, Ebstein et al., 1996). The DRD4 has two common variations in Western populations: a short form with four repeats of the base sequence and a long form with seven repeats. The long form was associated with high NS scores and the short form with low to moderate scores. Not all replications were successful. The most recent meta-analysis, including 48 studies published between 1996 and 2006, showed an association between both NS and extraversion and the DRD4 (Munafo et al., 2008). However another variant of the DRD4, the C-521T, was associated with NS and impulsivity but not with E. Apart from the problematic nature of the gene–trait association itself these gene variants only account for 3–4% of the trait variance. Obviously there must be many more genes involved and many may be of too small an effect to be detected.

Genes may determine specific behavioral traits rather than a broad measure like the trait ones we use as convenient summaries. The DRD4, for instance, has been associated with attention problems and sensation seeking behaviors in young children (Schmidt and Fox, 2002), attention deficit hyperactivity disorder (Faraone et al., 2001), and drug and alcohol use in young adults (Ebstein and Kotler, 2002). The A1 allele of the DRD2 dopamine receptor is found with greater frequency in alcohol, cocaine, and methamphetamine users (Noble, 1998).

The DRD4 has also been associated with sexual desire, function, and arousal (Ben-Zion et al., 2006). In primarily heterosexual and homosexual males, the DRD4 was associated with same-sex experience in the heterosexuals and opposite-sex experience in the homosexuals (Hamer, 2002). This could represent the attraction to novelty in sexual experience beyond that expressed in the variety of partners within the established sexual preference.

Biochemistry

The hormone testosterone is the driving force in sexual arousal in men. Given its absence or very low levels, sexual arousal is difficult. Daitzman and Zuckerman found that plasma testosterone was related to sensation seeking, particularly the disinhibition subscale. These findings were later replicated by Aluja and Garcia (2008), Aluja and Torrubia (2004), and Gerra et al. (2000), but not by Dabbs (2000), who used a salivary measure of the hormone. However investigators have found that testosterone is associated with a broader range of approach traits including extraversion, impulsiveness, assertiveness, rebelliousness, and aggression; it is also associated with the number of sexual partners. The evolutionary role of testosterone is demonstrated in a study of a polygynous culture that found that extraversion was related to high testosterone, polygynous marriage, and number of offspring in men (Alvergne et al., 2010).

Monoamine oxidase (MAO) is an enzyme that regulates the levels of monoamines in human and primate brain. It comes in two forms: type A preferentially oxidizes the neurotransmitters, serotonin and norepinephrine, whereas type B regulates brain dopamine (Shih et al., 1999). MAO-B can be obtained from blood platelets in humans although the relation between brain and platelet MAO is unknown. Many studies (9 out of 14) found a significant negative correlation between MAO-B and sensation seeking (Zuckerman, 1994). Low levels of MAO-B are also found in personality and clinical disorders characterized by impulsivity and sensation seeking such as antisocial personality disorder and bipolar mood disorder. Low levels are also found in those who abuse alcohol, tobacco, and drugs, all usually associated with high sensation seeking. All drugs, but particularly stimulant drugs, release the catecholamines, dopamine and norepinephrine in the brain.

Dellu et al. (1996) compared two strains of rats, one highly reactive to novelty (HR) and another low reactive (LR), or novelty avoidant, on self-administration of amphetamine and general exploratory behavior across a variety of situations. The HR rats self-administered amphetamine much more than the LR rats, were more exploratory, and autopsy showed higher levels of dopamine activity in the nucleus accumbens, a reward center in the brain where stimulant drugs have their major euphoric effects. In humans, dopamine is assessed indirectly through reactions of dopamine indices to dopamine stimulants like amphetamine. Studies relating sensation seeking to these indices have had mixed results. One problem with the studies of humans is that the measures of dopamine are not specific to brain areas. Dellu et al. (1996) found that dopamine in the nucleus accumbens correlated positively with exploratory behavior in rats but dopamine in the prefrontal cortex showed results in the opposite direction. Studies of responses to dopamine stimulants in relation to sensation seeking in humans have had mixed findings (e.g., Depue, 1995; Gerra et al., 2000; Netter et al., 1996; Wiesbeck et al., 1996). Brain imaging studies have showed a relation between sensation seeking

and amphetamine-stimulated arousal in the ventral striatum (containing the reward-sensitive nucleus accumbens) (Leyton et al., 2007; Stoops et al., 2007).

The biochemical theory of ImpSS suggests that the behavioral trait is based on an interaction between a HR dopamine approach system and an inhibition system based on serotonin (Zuckerman, 1995). Serotonin in limbic system pathways is generally associated with behavioral inhibition (Soubrie, 1986). Blunted serotonin reactivity to serotonin stimulants is characteristic in high sensation seekers (Depue, 1995; Hennig et al., 1998; Netter et al., 1996). Thus there is tentative emerging support for the biochemical/neurological theory of ImpSS.

Psychophysiology

The first biological theory was based on arousal and arousability of the cortical brain as regulated by the reticular activating system (Zuckerman, 1969). This is why the first studies focused on the electrophysiology of the brain at rest and in reaction to stimuli varying in novelty and/or intensity (Zuckerman, 1990). Studies of basal arousal levels of EEG, skin conductance, and heart rate showed no consistent differences between high and low sensation seekers in basal arousal, thus providing no evidence for a theory that high sensation seekers were chronically underaroused and needed intense or novel stimulation to reach an OLA.

However skin conductance and heart rate measures showed that high sensation seekers had a stronger response to a novel stimulus (orienting reflex) but not to subsequent repetitions of the stimulus to which they habituated at the same rate as low sensation seekers. These differences were enhanced by the emotional intensity and sensation seeking relevance of the content of textual stimuli. Novel and emotionally intense stimuli elicit a stronger autonomic response in high sensation seekers that is associated with positive affective tone.

The evoked cortical response is some function of stimulus intensity. However, the relationship between stimulus intensity and the amplitude of an early component of the EP has been shown to vary with sensation seeking, particularly disinhibition (Zuckerman, 1990; Zuckerman et al., 1974). The slope tends to be positive and strong in the high sensation seekers and weak or negative in the low sensation seekers. The results suggest some kind of cortical inhibition in response to high-intensity stimulation in the lows. The EP relationship to stimulus intensity is called 'augmenting' in the high sensation seekers and 'reducing' in the lows. Augmenting and reducing are also markers for similar behavioral characteristics in cats, such as approach to or avoidance of novel stimuli and impulsivity, or capacity to restrain response for reward (Saxton et al., 1987).

Evolution

Evolutionary psychology has been put forward as a new paradigm within which many or all fields of psychology, including personality, may be interpreted (Buss, 2009). But in fact there is a difference between personality and evolutionary theory. The former deals largely with individual variation within species and the latter with variation between species or subspecies. The former interprets variation at the more basic levels as expressive of genetic and biological variations and their influences in behavioral traits defining temperament. The latter assumes a genetic basis for evolved behavioral mechanisms but is less concerned with the specific genetic/biological bases for these differences.

I have postulated a possible evolutionary source for the trait of sensation seeking based not only on the adaptive value of behavioral component traits like explorativeness, sexual variety seeking, and risk-taking. But the evidence of evolution extends downward to other species, particularly those close on the evolutionary tree, like primates. However, mere similarity of behavioral traits across species leans too much on anthropomorphism unless we can show shared genetic/biological traits related to the similar behavioral traits within and between species.

Correlations between sensation seeking and behavioral traits have been found for hormones (testosterone), neurotransmitters, and their enzymes and controlling genes (dopamine, serotonin, MAO-B), and psychophysiological traits (EP augmenting/reducing), in humans and in monkeys, cats, rats, and mice. The findings suggest a broad trait of approach vs withdrawal, or sensation seeking vs. avoidance, or boldness vs. shyness, related to individual differences in most species of mammals, although it may even extend down to insects (Zuk, 2011). Of course these within-species differences are relative to between-species differences, such as that between predators and prey. Our hominid ancestors were both so that sensation seeking selection favored a balance between sensation seeking and avoidance. However of all the hominid species originally in Africa ours may have been the first to leave their familiar habitat to explore and settle in the furthest reaches of the world. Homo sapiens is a sensation seeking species, but one with a large range of variation in this trait within our species.

See also: Five Factor Model of Personality, Assessment of; Gender Differences in Personality and Social Behavior; Personality Assessment; Personality, Biological Models of; Personality, Evolutionary Models of; Personality, Trait Models of; Risk: Social Psychological Perspectives.

Bibliography

Adams, J.G., Heath, A.J., Young, S.E., Hewitt, A.J., Coreley, R.R., Stallings, M.C., 2003. Relationships between personality and preferred substance and motivations for use among adolescent substance abusers. American Journal of Drug and Alcohol Abuse 29 (9), 691–712.

Aluja, A., Garcia, L.F., 2008. Sensation seeking, sexual curiosity, and testosterone in inmates. Neuropsychobiology 51, 28–33.

Aluja, A., Kuhlman, M., Zuckerman, M., 2010. Development of the Zuckerman-Kuhlman-Aluja Personality Questionnaire (ZKA-PQ): a factor/facet version of the Zuckerman Kuhlman Personality Questionnaire (ZKPQ). Journal of Personality Assessment 92, 416–431.

Aluja, A., Rossier, J., Garcia, L.F., Angleitner, A., Kuhlman, M., Zuckerman, M., 2006. A cross-cultural shortened form of the ZKPQ (ZKPQ-50-CC) adapted to English, French, German, and Spanish languages. Personality and Individual Differences 41, 619–628.

Aluja, A., Torrubia, R., 2004. Hostility-aggresiveness, sensation seeking and sex hormones in men: reexploring their relationship. Neuropsychobiology 50, 102–107.

Alvergne, A., Jokela, M., Lummaa, 2010. Personality and reproductive success in a high-fertility population. PNAS Proceedings of the Natiional Academy of Sciences of the United States of America 107, 11745–11750.
Ball, S.A., 1995. The validity of an alternative five-factor measure of personality in cocaine abusers. Psychological Assessment 7, 148–154.
Ben-Zion, I.Z., Tessler, R., Cohen, L., Lerer, E., Raz, Y., Bachman-Melman, R., Gritsenko, I., et al., April 18, 2006. Polymorphisms in the Dopamine D4 Receptor Gene (DRD4) Contribute to Individual Differences in Human Sexual Behavior: Desire, Arousal and Sexual Function. Molecular Psychiatry Advance Online Publications, p. 5.
Benschop, A., Rabes, M., Korf, D.J., 2003. Pill Testing-Ecstasy and Prevention: A Scientific Evaluation in Three European Cities. Rozenberg Publishers, Amsterdam, the Netherlands.
Burns, P.C., Wilde, G.S., 1995. Risk-taking in male taxi drivers: relationships among personality, observational data and driver records. Personality and Individual Differences 18, 267–278.
Buss, D.M., 2009. How can evolutionary psychology successfully explain personality and individual differences. Perspectives on Psychological Science 4, 359–366.
Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multi-trait-multi-method matrix. Psychological Bulletin 56, 81–105.
Cooper, M.L., Frone, M.R., Russell, M., Mudar, P., 1995. Drinking to regulate positive and negtive emotions: a motivational model of alcohol use. Journal of Personality and Social Psychology 69, 990–1005.
Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Builletin 52, 281–302.
Dabbs, J.M., 2000. Heroes, Rogues and Lovers. McGraw-Hill, New York.
Dellu, F., Piazza, P.V., Mayo, W., LeMoal, M., Simon, H., 1996. Novelty-seeking in rats-biobehaviioral characteristics and possible relationships with the sensation seeking trait in man. Neuropsychobiology 34, 136–145.
Depue, R.A., 1995. Neurobiological factors in personality and depression. European Journal of Personality 9, 413–439.
Donohew, L., Bardo, M.T., Zimmerman, R.S., 2004. Personality and risky behavior: communication and prevention. In: Stelmack, R.M. (Ed.), On the Psychobiology of Personality: Essays in Honor of Marvin Zuckerman. Elsevier, New York, pp. 223–245.
Ebstein, R.P., Kotler, M., 2002. Personality, substance abuse, and genes. In: Benjamin, J., Ebstein, R.P., Belmaker, R.H. (Eds.), Molecular Genetics and the Human Personality. American Psychiatric Association Publishers, Washington, DC, pp. 151–163.
Ebstein, R.P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blain, D., et al., 1996. Dopamine D4 Receptor (DRD4) exon III polymorphism associated with the human personality trait of novelty seeking. Nature Genetics 12, 78–80.
Eysenck, H.J., 1983. A biometrical-genetical analysis of impulsivity and sensation seeking behavior. In: Zuckerman, M. (Ed.), Biological Bases of Sensation Seeking, Impulsivity, and Anxiety. Erlbaum, Hillsdale, NJ, pp. 1–27.
Faraone, S.V., Doyle, A.E., Mick, E., Biederman, J., 2001. Meta-analysis of the association between the 7-repeat allele of the dopamine D(4) receptor gene and attention deficit hyperactivity disorder. American Journal of Psychiatry 158, 1052–1057.
Finn, P.R., Hall, J., 2004. Cognitive ability and risk for alcoholism: short term-memory capacity and intelligence moderate personality risk for alcohol problems. Journal of Abnormal Psychology 113, 569–581.
Fulker, D.W., Eysenck, S.B.G., Zuckerman, M., 1980. A genetic and environmental analysis of sensation seeking. Journal of Research in Peronality 14, 261–281.
Gerra, G., Zaimovic, A., Giucastro, G., Maestr, D., Monica, C., Sartori, R., et al., 1998. Serotonergic function after ecstasy in humans. International Clinical Psychopharmacology 13, 1–9.
Gerra, G., Zaimovic, A., Ferri, M., Marzocchi, G.F., Timpano, M., Zambelli, U., et al., 2000. Neuroendocrine corrrelates of temperament traits in abstinent opiate addicts. Journal of Substance Abuse 11, 337–354.
Gosling, S.D., John, O.J., 1999. Personality diemensions in nonhuman animals: a cross-sectional species review. Science 8, 69–75.
Hamer, D.H., 2002. Genetics of sexual behavior. In: Benjamin, J., Ebstein, R.P., Belmaker, R.H. (Eds.), Molecular Genetics and the Human Personality. American Psychiatric Publishing, Washington, DC, pp. 257–272.
Hebb, D.O., 1955. Drives and the C.N.S. (conceptual nervous system). Psychological Review 62, 243–254.
Heino, A., 1996. Risk Taking in Car Driving: Perceptions, Individual Differences and Effects of Safety Incentives. University of Groningen Doctoral Dissertation. Published by A. Heino.
Hennig, J., Kroeger, A., Meyer, B., Prochaska, H., Krien, P., Huwe, S., Netter, P., 1998. Personality correlates of +/− pinodol induced decreases in prolactin. Pharmacopsychiatry 31, 19–24.
Hittner, J.B., Swickert, R., 2000. Sensation seeking and alcohol use: a meta-analytic review. Addicttive Behaviors 31, 1383–1401.
Hoyle, R., Fejfar, M.C., Miller, J.D., 2000. Personality and sexual risk-taking: a quantitative review. Journal of Personality 68, 1203–1231.
Hur, Y.-M., Bouchard Jr., T.J., 1997. The genetic correlation between impulsivity and sensation seeking traits. Behavior Genetics 27, 455–463.
Jobe, J.B., Holgate, S.H., Sopransky, T.A., 1983. Risk taking as motivation for volunteering in a hazardous experiment. Journal of Personality 51, 95–107.
Johnson, A.M., Vernon, P.A., Feiler, A.R., 2008. Behavior genetic studies of personality: an introduction and review of the results of 50+ years of research. In: Boyle, G.J., G.Matthews, Saklofsky, D.H. (Eds.), Personality Theory and Assessment, vol. 1. Sage, Los Angeles, CA, pp. 145–173.
Jonah, B.A., 1997. Sensation seeking and risky driving: a review and synthesis of the literature. Accident Analysis and Prevention 29, 651–665.
Kalichman, S.C., Johnson, J.R., Adair, V., Rompa, D., Multhauf, K., Johnson, J., Kelly, J., 1994. Sexual sensation seeking: scale development and predicting AIDS-risk behavior among homosexually active men. Journal of Personality Assessment 62, 385–397.
Koopmans, J.R., Boomsma, D.I., Heath, A.C., Lorenz, J.P.D., 1995. A multivariate genetic analysis of sensation seeking. Behavior Genetics 25, 349–356.
Leyton, M., aan het Rot, M., Booij, L., Baker, G.B., Young, S.N., Benkelfat, C., 2007. Mood-elevating effects of d-amphetamine and incentive salience: the effect of acute dopamine precursor depletion. Journal of Psychiatry and Neuroscience 32, 129–136.
Lykken, D.T., 1992. Personal Communication.
McCrae, R.R., Costa Jr., P.T., 2003. Personality in Adulthood: A Five-Factor Perspective, second ed. Guilford Press, New York.
Munafo, M.R., Yalcin, B., Willis-Owen, S.A., Flint, J., 2008. Association of the dopamine D4 receptor (DRD4) and approach related personality traits: meta-analysis and new data. Biological Psychiatry 63, 197–206.
Musolino, R.F., Hershenson, D.B., 1977. Avocational sensation seeking in high and low risk taking occupations. Journal of Vocational Behavior 10, 358–365.
Neria, Y., Solomon, Z., Ginzburg, K., Dekel, R., 2000. Sensation seeking, wartime performance, and long-term adjustment among Israeli war veterans. Personality and Individual Differences 29, 921–932.
Netter, P., Hennig, J., Roed, I.S., 1996. Serotonin and dopamine as mediators of sensation seeking behavior. Neuropsychobiology 34, 155–165.
Noble, E.P., 1998. The D2 dopamine receptor gene: a review of association studies in alcoholism and phenotypes. Alcohol 16, 33–45.
Patkar, A.A., Murray, H.W., Mannelli, P., Gottheil, E., Weinstein, S.P., Vergare, M.J., 2004. Pre-treatment measures of impulsivity, aggression, and sensation seeking are associated with treatment outcome for African-American cocaine dependent patients. Journal of Addictive Diseases 23, 109–122.
Perse, E.M., 1996. Sensatiion seeking and the use of television for arousal. Communication Reports 9, 37–48.
Rawlings, D., Barrantes i Vidal, N., Furnham, A., 2000. Personality and aesthetic preference in Spain and England: two studies relating sensation seeking and openness to experiences to liking for paintings and music. European Journal of Personality 14, 553–576.
Saxton, P.M., Siegel, J., Lukas, J.H., 1987. Visual evoked potential augmenting-reducing slopes in cats-2: correlations with behavior. Personality and Individual Differences 8, 511–519.
Schierman, M.J., Rowland, G.L., 1985. Sensation seeking and selection of entertainment. Peronality and Individual Differences 5, 599–603.
Schmidt, L.A., Fox, N.A., 2002. Molecular genetics of temperamental differences in children. In: Benjamin, J., Ebstein, R.P., Belmaker, R.H. (Eds.), Molecular Genetics and the Human Personality. American Psychiatric Association Publishing, Washington, DC, pp. 257–272.
Segal, B.S., Huba, G.J., Singer, J.F., 1980. Drugs, Daydreaming and Personality. Studies of College Youth. Erlbaum, Hillsdale, NJ.
Segal, B.S., Singer, J.L., 1976. Daydreaming drugs and alcohol use in college students. A factor analytic study. Addictive Behaviors 1, 227–235.
Shih, J.C., Chen, K., Ridd, M.J., 1999. Monoamine oxidase. From genes to behavior. Annual Review of Neuroscience 22, 197–217.
Soubrie, P., 1986. Reconciling the role of central serotonin neurons in human and animal behavior. Behavioral and Brain Sciences 9, 319–364.
Stoops, W.W., Lile, J.A., Robbins, C.G., Martin, C.A., Rush, C.R., Kelly, T.H., 2007. The reinforcing, subject rated performance, and cardiovascular effects of d-amphetamine: influence of sensation seeking status. Addictive Behaviors 32, 1177–1188.
Wiesbeck, G.A., Widarz, N., Mauer, C., Jacob, F., Boening, J., 1996. Sensation seeking, alcoholism and dopamine activity. European Psychiatry 11, 87–92.
Zuckerman, M., 1969. Theoretical formulations. In: Zubek, J.P. (Ed.), Sensory Deprivation: Fifteen Years of Research. Appleton-Century-Crofts, New York, pp. 407–432.

Zuckerman, M., 1971. Dimensions of sensation seeking. Journal of Consulting and Clinical Psychology 36, 45–52.
Zuckerman, M., 1979. Sensation Seeking: Beyond the Optimal Level of Arousal. Erlbaum, Hillsdale, NJ.
Zuckerman, M., 1990. The psychophysiology of sensation seeking. Journal of Personality 54, 313–345.
Zuckerman, M., 1991. Psychobiology of Personality. Cambridge University Press, Cambridge.
Zuckerman, M., 1994. Behavioral Expressions and Biosocial Bases of Sensation Seeking. Cambridge University Press, New York.
Zuckerman, M., 1995. Good and bad humors: biochemical bases of personality and its disorders. Psychological Science 6, 325–332.
Zuckerman, M., 2005. Psychobiology of Personality, Second Edition, Revised and Updated. Cambridge University Press, New York.
Zuckerman, M., 2006. Sensation seeking in entertainment. In: Bryant, J., Vorderer, P. (Eds.), Psychology of Entertainment. Erlbaum, Mahwah, NJ, pp. 367–387.
Zuckerman, M., 2007. Sensation Seeking and Risky Behavior. American Psychological Association, Washington, DC.
Zuckerman, M., 2012. Psychological factors and addiction: personality. In: Shaffer, H.J. (Ed.), APA Addiction Syndrome Handbook, Foundations, Influences and Expressions of Addiction, vol. 1. American Psychological Association, Washington, DC, pp. 175–194.
Zuckerman, M., Cloninger, C.R., 1996. Relationships between Cloninger's, Zuckerman's and Eysenck's dimensions of personality. Personality and Individual Differences 21, 283–285.
Zuckerman, M., Eysenck, S., Eysenck, H.J., 1978. Sensation seeking in England and America: cross-cultural, age, and sex comparisons. Journal of Consulting and Clinical Psychology 40, 139–149.
Zuckerman, M., Kolin, I., Price, L., Zoob, I., 1964. Development of a sensation seeking scale. Journal of Consulting Psychology 28, 477–482.
Zuckerman, M., Kuhlman, D.M., 2000. Personality and risk-taking: common biosocial factors. Journal of Personality 68, 999–1029.
Zuckerman, M., Kuhlman, D.M., Joireman, J., Teta, P., Kraft, M., 1993. A comparison of three structural models for personality: the big three, the big five, and the alternative five. Journal of Personality and Social Psychology 65, 757–768.
Zuckerman, M., Murtaugh, T.T., Siegel, J., 1974. Sensation seeking and cortical augmenting-reducing. Psychophysiology 11, 535–542.
Zuckerman, M., Tushup, R., Finner, S., 1976. Sexual attitudes and experience: attitude and personality correlations and changes produced by a course in sexuality. Journal of Consulting and Clinical Psychology 44, 7–19.
Zuk, M., 2011. Sex on Six Legs: Lessons on Life, Love, and Language from the Insect World. Houghton Mifflin Harcourt, New York.

Senses, Anthropology of the

David Howes, Concordia University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The anthropology of the senses takes a cultural approach to the study of the senses, arguing that perception is a social as well as physical and psychological act. It explores the particular meanings associated with different senses and sensations across cultures and how these give rise to different sensory paradigms and practices for understanding and relating to the world. By so doing it challenges many of the assumptions and conclusions of Western sensory psychology (psychophysics) with regard to the number, nature, and uses of the senses. The anthropology of the senses has also been responsible for promoting a sensory approach to the study of culture. As a result, sensory reflexivity has become increasingly integral to good ethnography. Many domains can be explored and understood more fully when approached from a particular sensory or multisensory angle, including religion, medicine, material culture, kinship, apprenticeship, and even language.

The anthropology of the senses starts from the premise that the ways we use our senses, and the ways we create and understand the sensory world, are mediated by culture. Perception is, therefore, not only a matter of physical experience and personal associations, but also of social significance. The senses are constructed and lived differently in different cultures and historical periods. Hence, when we examine the meanings vested in various sensory faculties and sensations across cultures:

> we find a cornucopia of potent sensory symbolism. Sight may be linked to reason or to witchcraft, taste may be used as a metaphor for aesthetic discrimination or for sexual experience, an odor may signify sanctity or sin, political power or social exclusion. Together, these sensory meanings and values form the *sensory model* espoused by a society, according to which the members of that society 'make sense' of the world, or translate sensory perceptions and concepts into a particular 'worldview'. There will likely be challenges to this model from within the society – persons and groups who differ on certain sensory values – yet this model will provide the basic perceptual paradigm to be followed or resisted. (Classen, 1997: p.402)

The first task for the sensory anthropologist is accordingly to examine how the senses are fashioned in a particular culture, and to guard against projecting the sensory model of his or her own society (and class) onto the sensory life of the culture under study (Howes, 1991).

The Hierarchy of the Senses

The traditional Western sensory model holds that the senses are five in number, and that each sense has its proper sphere. Sight is concerned with colors, hearing with sound, smell with odors, and so on. The senses are also informally ranked, with sight being considered the 'highest' most informative, objective, rational, aesthetic, and 'civilized' of the senses. Sight is followed by hearing, 'the second sense.' Both sight and hearing are conventionally set off from the 'lower,' subjective, emotional, unaesthetic, and animalistic or 'brute' senses of smell and taste and touch.

The ranking of the senses is typically allied with the ranking of social groups. For example, the traditional Western association of the male sex with the 'higher' senses of sight and hearing supported the notion that men are naturally fitted for such activities as exploring, ruling, and studying or writing, while the association of the female sex with the proximity senses relegated women to the home, and made them mistresses of the kitchen, the nursery, and the bedroom. Such was the power of this categorization that those women who challenged the sensory division of labor (e.g., by writing or painting instead of cooking and sewing) faced considerable social opprobrium until well into the twentieth century (Classen, 1998).

As regards social class, the following description of French fishermen from a travel book of the 1920s illustrates how class may be sensed and the senses classed:

> [The fishermen] are strong, brown-skinned creatures, moving with a supple animal grace. Only a few have any ambition or education. Their life is primitive, hardworking, and simple, concerned with very little beyond the needs of food and shelter. (Peck, 1929: p.39)

The implication here is that the fishermen are like animals who, supposedly, live mostly through the lower senses, seeking food for their bellies and a warm place to sleep.

Another typical marker of class distinction is the capacity to distinguish between sensations within the same sensory register, and display a preference for those which are deemed to be more 'refined,' such as dark versus 'loud' colors in clothing, or haute cuisine versus 'greasy' fare in food (Bourdieu, 1984).

As regards race, the model of the five senses was used to categorize 'the five races of Man' in the grand taxonomy proposed by the German embryologist Lorenz Oken in the early nineteenth century:

1. The skin-man is the black, African.
2. The tongue-man is the brown, Australian-Malayan.
3. The nose-man is the red, American.
4. The ear-man is the yellow, Asiatic-Mongolian.
5. The eye-man is the white, European (Gould cited in Howes, 2009: p.10).

Oken's ascending scale of 'sensory perfection' (and skin pigmentation) was not based on any intrinsic propensities of the peoples concerned, but rather on their social ranking within the European imagination. His ostensibly biological categorization of senses and peoples was actually shot through with cultural values.

The senses continued to figure prominently in the ongoing effort to establish racial differences in the French physical anthropology and British psychological anthropology of the second half of the nineteenth century. One of the major issues debated by the members of the Société d'anthropologie de Paris (SAP), for example, was how the objectivity of racial categorization could be guaranteed if, as was assumed, perception is a physiological process and therefore potentially tainted with subjectivity? The 'solution' lay in exteriorizing the process of observation by (1) subjecting it to diverse protocols designed to neutralize the 'personal equation,' such as determining 'la bonne distance' from which to gauge the color of the iris (considered to be an essential marker of racial difference) and using Broca's chromatic scale to record judgments; and, (2) deploying a range of newly invented technologies (from the ophthalmoscope to the esthesiometer) which could substitute for the observer's own senses while objectifying those of the 'primitive' (Dias, 2004).

In addition to measuring the senses, the members of the SAP dedicated themselves to determining the localization of the various sensory and other faculties in the topography of the brain. The cultural hierarchy of the senses and its associated values clearly guided their research in this connection, as evidenced by the many fantastical observations they made (Dias, 2004).

The members of the 1898 Cambridge Anthropological Expedition to the Torres Strait (the first expedition of its kind) also employed a huge battery of tests – Zwaardemaker's olfactometer, Politzer's Hörmesser, Lovibund's tintometer, and so on – to gauge the sensory acuity of the natives. They did so based on the hypothesis that the senses (particularly the 'lower senses') of 'primitive man' would prove more acute than those of Europeans due to their mental inferiority and animalistic tendencies. The presumption that sensibility varied inversely with intelligence was a Eurocentric conceit of the first (and worst) order. The evidence of the tests was inconclusive but, astonishingly, this did little to unseat the hypothesis (Berry et al., 2011: ch. 9; Howes, 2012: 634–637).

It bears underlining that while some members of the expedition, most notably W.H.R. Rivers, entertained the notion that environmental (and not just physiological) factors might have a role to play in shaping the test results, no attempt was made to elicit the indigenous sensory model or study native perceptual practices.

From Measurement to Meaning

The critique and collapse of the concept of race in early twentieth century anthropology led to a corresponding decline in the explicit attention paid to the senses. This was due partly to the association of the study of the senses with the construction of the discredited racial taxonomies, and partly to the life of the senses having become the stuff of travel writing. Anthropology, which was striving to establish itself as a science, did not wish to be perceived as focused on 'superficial' sensations. Interestingly, one of the major critics of the racial hypothesis, Franz Boas, pointed to an alternate way forward for the study of the senses in a little note called 'Alternating Sounds' (Boas, 1889). This piece emphasized the importance of attending to the cultural patterning of sense impressions.

Boas' directive was finally operationalized in the 1950s when his former student, Margaret Mead, and various members of her circle, started to explore the notion that people "not only hear and speak and communicate through words, but also use all their senses in ways that are equally systematic … to taste and smell and to pattern their capacities to taste and smell, so that the traditional cuisine of a people can be as distinctive and as organized as a language" (Mead and Métraux in Howes, 2003: p.11). This promising opening was, however, overshadowed by the rise of language-based models of culture in the 1960s and 1970s (e.g., culture as language game, as discourse, as text), and the reduction of ethnography to writing in the 1980s. The 'writing culture' movement virtually obliterated the idea of sensing cultures. It did so by promoting 'the model of the text' (which compounded the logocentrism of conventional Western thought) and focusing attention exclusively on questions of interpretation, representation, authority, and style (Howes, 2003: ch. 1).

However, ongoing empirical research in cross-cultural psychology on such topics as color classification and susceptibility to pictorial illusions kept the senses from being entirely lost to view (Berry et al., 2011: ch. 9). The media theories of Marshall McLuhan and Walter Ong also had an influence, sensitizing anthropologists to the impact of different communications technologies on the 'balance' or 'ratio of the senses' (Carpenter, 1973; Howes 1991: ch. 11). The corporeal turn of the 1980s and 1990s, which highlighted the unity of mind and body (thereby deobjectifying perception), and promoted a more phenomenological approach to the study of cultural experience, was a further contributing factor with respect to bringing the senses back into focus.

As a result of the new emphasis on 'embodiment' (in place of 'representation'), anthropologists came increasingly to take stock of their own senses, and in the process discover things about the sensory models of the people they studied. For example, Paul Stoller underwent an acoustic epiphany in the course of his apprenticeship as a sorcerer among the Songhay of Niger. "You must learn how to hear, or you will learn little about our ways" (Stoller, [1984] 1989: ch. 7) his Songhay mentor admonished him following an incident where he sat there looking when he should have been listening. Stoller ([1984] 1989: ch. 1) also related how he was sensitized to Songhay gustatory codes after being served a particularly bad-tasting sauce by a disgruntled member of his host's household. He went on to write a discerning piece on the sauces of Songhayland, how they are distinguished on the basis of flavor and consistency, and how a cook can play with the ingredients to convey messages (Stoller, [1990] 2005).

While dramatic incidents of the sort Stoller relates may constitute a turning point for some, it is important to recognize that the conventional ethnographic method of 'participant observation' pioneered by Bronislaw Malinowski has always involved exercising the senses to some degree

(Howes, 2003). As François Laplantine observes in *Le social et le sensible*:

> The experience of fieldwork is an exercise in sharing the sensory world [*partage du sensible*]. We observe, we listen, we speak with others, we partake of their cuisine, we try to feel what they experience along with them. (Laplantine, 2005: p.11, my translation)

Good ethnography will therefore tend to be sensuous ethnography (Herzfeld, 2001). The anthropology of the senses itself, however, focuses on eliciting the sensory model of the culture (or subculture) under study and registering the full range of perceptual practices that underpin that model. This *cultural* approach is the complete opposite of the psychophysical approach of the 1898 Cambridge Expedition.

A Cultural Approach to the Study of the Senses

The enumeration, elaboration, socialization, and interrelation of the senses are four topics of interest to sensory anthropology. Taking a cultural approach to the study of these topics can reveal many issues which a psychological (including phenomenological) approach would miss or fail to comprehend adequately, as will be shown in the following.

Enumeration

Consider how the Cashinahua of Peru divide up the sensorium. They distinguish six sensoria or percipient centers – eyes, ears, skin, hands, genitals, and liver – and six kinds of sense-data or knowledge. 'Liver knowledge' pertains to the emotions, since it is in the liver that joy or sorrow, fear or hope, are felt. 'Ear knowledge' concerns social codes, perhaps reflecting the centrality of orality to the transmission of social values. 'Eye knowledge' has to do with perceiving the spiritual insides or substance of persons, animals, and things, as opposed to their surface. (This mode of perception, which is a characteristic of dreams is also helped by the ingestion of hallucinogens.) It is the skin that registers surfaces – the feel of the environment (including sun, wind, rain, and forest) and the behavior patterns of people and animals. 'Skin knowledge' is what enables the Cashinahua to navigate the jungle, locate prey, and interact with others.

'Hand knowledge' is divided by gender: in the case of women, the hands are the means by which knowledge of weaving, pottery-making, and cooking skills enter the body, while in men's case, the hands pick up how to hunt with bow and arrow or chop down a tree. 'Genital knowledge' has to do with sex and procreation, but more fundamentally it is bound up with awareness of mortality and immortality as embodied in one's children. Interestingly, when asked if the brain has any role to play in processing and retaining all this knowledge, the Cashinahua, while appreciating the question, thought not, rather 'the whole body knows' (Howes, 2009: p.2).

Cashinahua sensory psychology is fascinating for its denial of brain-based cognition and insistence on the modularity of the body, rather than of mind (Fodor, 1983). It also assorts functions rather differently than Western sensory psychology. For example, the Cashinahua treat emotions as perceptions in the case of 'liver knowledge' and also mix in sensori-motor functions with perception in the case of 'hand knowledge.' 'Ear knowledge' has a decidedly moral tinge to it since it has to do with conformity to social codes, and so cannot be considered strictly psychophysical either. Meanwhile, the Cashinahua appear to leave out taste and smell while attributing an exaggerated role to the skin and a very restricted role to vision. In effect, the Cashinahua feel with their skin what most Westerners would think of themselves as seeing with their eyes, and the Cashinahua see into places (e.g., the 'souls' of persons), which most Westerners would not think accessible to the senses.

These apparent exaggerations and diminutions, confusions and distinctions are all reciprocal, of course. There is no Archimedean point from which to determine how the senses function. The best an anthropologist can do is to cultivate the capacity to be 'of two sensoria' about things – their own and that of the culture under study (Howes, 2003: pp. 10–12).

If Western sensory psychology is to achieve more than parochial status, it will need to do the same as social anthropology, and acknowledge that there is no one true way to categorize the senses. W.H.R. Rivers, who played a key role in the professionalization of anthropology *and* experimental psychology in Britain in the early twentieth century, could have made a difference in this regard, if he had been more attentive to different ways of understanding the senses across cultures. Instead he assumed that the Western model provided the standard. Following his involvement with the 1898 Cambridge Expedition, he went on to conduct a series of psychophysical experiments among the Toda of India, and published the results in the first issue of the *British Journal of Psychology* (Rivers, 1905). Rivers never inquired into or tried to adapt his experiment to the Toda sensory model. Had he inquired he might have discovered a model that resembled the eightfold plan of the sensorium in the Brihadaranyaka Upanishad:

> (1) prana (breathing organ, i.e., nose; also 'breath of life'); (2) the speech organ; (3) tongue (taste); (4) eye (color); (5) ear (sounds); (6) mana (thought, mind, inner organ); (7) hands (work); and (7) skin (sense of touch). (Elberfeld, 2003: p. 483)

This model is noteworthy for its inclusion of mind and speech among the senses, as well as its privileging of olfaction (which has to do with breathing being the sense of reflection in India – e.g., in yoga – as sight is the sense of reflection in the West). But Rivers, set on his own sensory model, did not perceive any of this, and so the anthropological study of the senses remained dominated by Western biases for many decades to follow.

Elaboration

The anthropology of the senses is historically indebted to the phenomenology of perception. However, it departs from the latter by virtue of its focus on the 'performative elaboration' of the senses (Geurts, 2002: p. 5) as opposed to their 'phenomenological reduction' (Ingold and Howes, 2011).

Compare the experience of taste in, say, North America and India. Taste (*rasa*) is highly elaborated on multiple levels in the sensory model of Hindu India. The Sanskrit gustatory vocabulary consists of six terms, whereas the English taste lexicon has but four (five if one counts the Japanese loanword *umami)*. Ayurvedic medical theory recognizes a range of 'postdigestive tastes,' which suggest a prolongation of gustation and something much deeper than the experience of acid reflux or what in English are called aftertastes. Ayurvedic medicine also holds that the virtue of a medicine resides in its flavor whereas the taste of a drug is not counted among its therapeutic properties in contemporary Western medicine. Emotional states, of which there are eight according to the *Natyashastra*, are classified as savors in India whereas English-speakers assimilate them to 'feelings' and are more inclined to identify them by color (e.g., red with rage, green with envy, etc.). Audiences in India want to 'taste the rasa' of a performance, not just see a play or hear a concert, and there is (again) something more going on in this connection than what an English-speaker means by 'musical taste.' Taste differences are so important in India that they determine an individual's rank in the caste system (with its progressive vegetarianism) in contrast to the Anglo-American tendency to insist that meat be at the center of every meal. And not only is the social order grounded in dietary distinctions, but the very order of the universe is held to depend on regular gustatory exchanges between humans and gods. In the ritual practice of *prasad*, for example, one makes an offering of food to the deities and then, once the latter have consumed its essence or spirit, eats the leftovers (Pinard, 1991; Goswamy, 2005; Freeman, 2011). How would an English-speaking phenomenologist even begin to access any of these multiple senses of taste in India?

The phenomenologist would be on more familiar ground in the domain of sight. He or she could point out that sight is a spatial sense in contrast to hearing, which is temporal, that sight is a distance sense in contrast to touch, which is reciprocal, that sight fosters objectivity (the separation of subject and object) in contrast to taste, which involves fusion, and so on (Jonas, 1970). But then the whole edifice crumbles once cultural practice – the *performance* of the senses – is factored in. In the West, the gaze may appear to be unilateral and objectifying, but in India it is the *mutuality* of the gaze that is emphasized. The practice of *darsan*, for example, involves going to see and be seen by a deity in the form of an idol. This is understood to be auspicious rather than objectifying (Pinard, 1991).

The phenomenologist might carry on by suggesting that perhaps sight is a 'metaphorical touch' in India or, better, invoke the vision, which French phenomenologist Maurice Merleau-Ponty had toward the end of his life of the 'prereflective unity' of the senses and 'the flesh' as all-encompassing. However, while the senses might all come together in the consciousness of the individual subject, and so be thought to collaborate, they also conflict, as when a stick in water appears crooked to the eye but feels straight to the hand, or when a forest looks and sounds still and peaceful but the smell of a wolf signals danger. Cultural conflicts in sensation occur when, for example, a holy person considered to be ugly in appearance emits a sweet odor of sanctity (Classen et al., 1994). In fact, such conflicts happen all the time, albeit subtly (Stewart, 2007). Making sense of the world depends on attending to sensory differentials and breaking with the abstractions and essentialisms of phenomenology. (Many anthropologists identify as phenomenologists and think of themselves as doing phenomenology even when they have clearly had to slough off its strictures to make the observations they do. On the other hand, some do go on practicing phenomenology regardless.)

Socialization and Interrelation

In *Culture and the Senses*, Kathryn Geurts presents an in-depth study of how the socialization of the senses contributes to the development of the self, the acquisition of specific behavioral orientations, and the reproduction of cultural identity among the Ewe-speaking Anlo people of Ghana. This socialization begins even before birth with expectant parents picturing their unborn child as seated on a 'stool' in the womb (the stool being the placenta) practising the arts of posture and balance.

From the moment they can sit and hold their heads up without support, Anlo infants are implored to 'Balance! Balance!' A child's joints are continuously massaged, beginning with its first bath, to ensure agility of movement and inculcate a flexible attitude. As a minority ethnic group living in a resource poor region of Ghana, the Anlo-Ewe must be flexible. As one proverb puts it: "If you visit the village of the toads and you find them squatting, you must squat too" (Geurts, 2002: p. 96). In a related vein, the Ewe language contains over 50 terms for different kinaesthetic styles, or ways of walking, and these are used as a gauge of character. One wants to stride like a lion, *kadzakadza*, not meander, *lugulugu* (Geurts, 2002: p. 74)

The Ewe term for 'sense' is *seselelame* which, literally translated, means 'feel-hear feel-hear at flesh inside.' The audio-haptic connotations of this term, along with its accent on the interior milieu (interoception), reveal much about the focal sensory orientation of this culture. Interestingly, the Anlo hold that loss of hearing is the gravest form of sensory impairment, because with this loss would come a disruption to one's sense of balance. The maintenance of balance is key to what it means to be a person the Anlo way. And, when life gets out of balance, as happens more often than not, Anlo have a tendency to fold inwards, or curl up into a ball. This posture is captured in the very pronunciation of the word by which the Anlo designate themselves as people, and also refers to a particular moment in their migration history when the ancestor who led them out of slavery collapsed from exhaustion and despair. As one of Geurts' female informants noted, "being Anlo meant that you felt that way [i.e., rolled up] a lot, but you always had to unroll, or come out of it, and that gave you a feeling of strength" (Geurts, 2002: p. 118).

Parenthetically, balance is not normally counted as a sense in Western culture, and, on the whole, looks are more important than walks. Significantly, in many Western countries, expectant parents' first encounter with their unborn child usually comes in the form of an ultrasound image. If they are of a certain class, they may play Mozart recordings to the fetus.

The newborn child will continue to be exposed to Mozart and be entertained with colorful, geometric toys and Baby Einstein® videos while lying in its crib or other enclosure. The child's first book will likely be one of the countless childrens' books on *The Five Senses* which, by training the child to focus on the external (exteroceptive) senses and process each sensation separately, will have the effect of suppressing any interoceptive and/or synaesthetic tendencies it might have (van Campen, 2007). In the classroom, the child will be trained to sit still for long periods of time solely looking and listening. Resisting can only result in failure. The youth will go on to construct an identity for itself using social media.

A Sensory Approach to the Study of Culture

Geurts' work shows that exploring a given culture's sensory model – the ways in which the senses are discriminated, elaborated, and related to or combined with one another in practice – can illuminate many other aspects of the culture, such as its moral code and identity. This discovery underlines a point made by Constance Classen: "The broad range of applications for a sensory analysis of culture indicates that the anthropology of the senses need not be a 'sub-field' within anthropology, but may provide a fruitful perspective from which to examine many different anthropological concerns" (1997: p. 409). This point was picked up on and developed by Michael Herzfeld (2001), whose discussion in turn inspired the editors of *Ethnographic Fieldwork: An Anthropological Reader* to include a section on 'Sensorial Fieldwork' in their anthology. As they put it:

> sensorial fieldworkers [should] make the study of the entire sensorium indispensable to other domains of ethnographic inquiry, such as economics or politics … Just as attention to gender and reflexivity is now part and parcel of most ethnographic work, so the entire range of senses should be of similar concern. (Robben and Sluka, 2007: p. 388)

The fruits of this realization can be seen in the proliferation of sense-based methods of inquiry in anthropology, which both challenge and complement the language- or text-based models of earlier decades (Howes 1991: Epilogue; Feld, 2004; Bendix and Brenneis, 2005; Pink, 2009; Howes, 2012; Howes and Classen, 2013). For example, kinship studies, which used to be dominated by the analysis of kinship terminologies, has morphed into 'skinship' studies, which examines the sensory dimensions of the gestures by which kin relate to each other and distinguish themselves from nonkin (Gregory, 2011). Museum studies are increasingly centered on 'experiencing the [multiple sensory] properties of things' in place of 'contextualizing' them solely by means of labels or simply placing them on display (Lamp, 2004; Dudley, 2012). Food studies have transitioned from a concern with 'deciphering' a meal to an interest in how meals are 'experienced,' and how distinctions of flavor may structure social relations (Walmsley, 2005; Sutton, 2010). Religious studies are being rethought in the light of the notion of religions as 'sensational forms' (Meyer, 2009). Even anthropological linguistics has been sensualized as the old debates over how language 'determines' perception, have been supplemented by a new, more productive focus on the senses 'in' language (Classen, 1993: ch. 3; Majid and Levinson, 2011).

Some have called for the reconciliation of language- and sense-based approaches to culture (Porcello et al., 2010). This would be a welcome development, and there is precedent for it in the fact that many cultures treat speech as a sense (Howes, 2009: pp. 3–4). Perhaps therefore this opposition can be overcome. What of the other opposition touched on at various points in the preceding account – that is, the rift between sensory anthropology (with its openness to indigenous constructions of the sensorium) and Western sensory psychology (with its psychophysical reductionism)? The emergent field of neuroanthropology, with its notion of the 'encultured brain' (Lende and Downey, 2012) is one area where there could be a fruitful conversation between sensory anthropologists and their counterparts in the brain sciences. Here, though, it is important to bear in mind (and be prepared to answer) the Cashinahua's challenge to the cerebrocentrism of the latter approach.

By way of illustration, consider the Brazilian martial art of capoeira as described by Greg Downey, who is both a seasoned anthropologist and one of the leading scholars of neuroanthropology. In *Learning Capoeira* (2005) and other works, Downey relates how he was sensitized to diverse highly specific visual and auditory techniques and kinaesthetic practices in the course of his apprenticeship to a master capoerista. Capoeira is famous for the cunning subterfuges, lightning-quick strikes, and extraordinary feats of balancing while upended (i.e., imbalanced) displayed by its adepts. Success at fending off attacks depends on cultivating what is known as the 'sideways glance' – that is, on the practitioner dispersing his or her attention over the entire visual field and suppressing the natural reflex tendency to visually track moving objects or actions. This is a fascinating example of how 'culture tunes our neurons' (Sacks quoted in Howes, 2005: p. 21) or, in other words, how cultural practice overwrites the way the brain is wired, since our natural tendency is to track objects and movements with our eyes. Downey elaborates on what this technique of perception involves:

> A practitioner must pay careful attention to the whole body of his or her adversary all at once because some capoeiristas practice movements that are disjointed. Rather than being grounded in the cooperative action of the entire body, movements involve only a single part in isolation, making it very hard to anticipate what will happen from observing other body parts; that way attacks can also be thrown from unforeseen angles. (Downey, 2007: p. 229)

The phenomenologist would be flummoxed by this description, but the Cashinahua might just find it of interest, given the parallels to their own theory of the modularity of the sentient body.

See also: Apprenticeship, Anthropological Aspects; Body: Anthropological Aspects; Cognitive Neuroscience;

Cross-Cultural Psychology; Ethnography in Applied Social Research; Experimentation in Psychology, History of; Knowledge, Anthropology of; Sociality In Anthropology.

Bibliography

Bendix, R., Brenneis, D. (Eds.), 2005. The senses. Etnofoor 18 (1).

Berry, J.W., Poortinga, Y.H., Breugelmans, S.M., Chasiotis, A., Sam, D.L., 2011. Cross-Cultural Psychology: Research and Applications, third ed. Cambridge University Press, Cambridge, MA.

Boas, F., 1889. On alternating sounds. American Anthropologist 2 (1), 47–54.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste (R. Nice, Trans.). Cambridge University Press, Cambridge, MA.

Carpenter, E., 1973. Oh, What a Blow That Phantom Gave Me! Bantam Books, Toronto.

Classen, C., 1993. Worlds of Sense: Exploring the Senses in History and across Cultures. Routledge, London.

Classen, C., 1997. Foundations for an anthropology of the senses. International Social Science Journal 153, 401–412.

Classen, C., 1998. The Color of Angels: Cosmology, Gender and the Aesthetic Imagination. Routledge, London.

Classen, C., Howes, D., Synnott, A., 1994. Aroma: The Cultural History of Smell. Routledge, London.

Dias, N., 2004. La mesure des sens: les anthropologues et le corps humain au XIXe siècle. Flammarion, Paris.

Downey, G., 2005. Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art. Oxford University Press, New York.

Downey, G., 2007. Seeing with a 'sideways glance': visuomotor 'knowing' and the plasticity of perception. In: Harris, M. (Ed.), Ways of Knowing: Anthropological Approaches to Crafting Experience and Knowledge. Berghahn Books, Oxford.

Dudley, S. (Ed.), 2012. Museum Objects: Experiencing the Properties of Things. Routledge, London.

Elberfeld, R., 2003. Sensory dimensions in intercultural perspective and the problem of modern media and technology. In: Hershock, P., Stepaniants, M., Ames, R. (Eds.), Technology and Cultural Values. University of Hawai'i Press, Honolulu.

Feld, S., 2004. Places senses, senses placed: toward a sensuous epistemology of environments. In: Howes, D. (Ed.), Empire of the Senses: The Sensual Culture Reader. Berghahn Books, Oxford.

Fodor, J., 1983. The Modularity of Mind: An Essay on Faculty Psychology. The MIT Press, Cambridge, MA.

Freeman, R., 2011. Taste. Material Religion 7 (1), 132–139.

Geurts, K., 2002. Culture and the Senses: Bodily Ways of Knowing in an African Community. University of California Press, Berkeley, CA.

Goswamy, B.N., 2005. *Rasa*: delight of the reason. In: Korsmeyer, C. (Ed.), The Taste Culture Reader: Experiencing Food and Drink. Berghahn Books, Oxford.

Gregory, C., 2011. Skinship: touchability as a virtue in East-Central India. Hau: Journal of Ethnographic Theory 1 (1). www.haujournal.org.

Herzfeld, M., 2001. Anthropology: Theoretical Practice in Culture and Society. Blackwell, Oxford.

Howes, D. (Ed.), 1991. The Varieties of Sensory Experience: A Source Book in the Anthropology of the Senses. University of Toronto Press, Toronto, ON.

Howes, D., 2003. Sensual Relations: Engaging the Senses in Culture and Social Theory. University of Michigan Press, Ann Arbor, MI.

Howes, D. (Ed.), 2005. Empire of the Senses: The Sensual Culture Reader. Berg, Oxford.

Howes, D. (Ed.), 2009. The Sixth Sense Reader. Berg, Oxford.

Howes, D., 2012. Re-visualizing anthropology through the lens of *the Ethnographer's Eye*. In: Heywood, I., Sandywelll, B. (Eds.), The Handbook of Visual Culture. Berg, Oxford.

Howes, D., Classen, C., 2013. Ways of Sensing: Understanding the Senses in Society. Routledge, London.

Ingold, T., Howes, D., 2011. Worlds of sense and sensing the world. Social Anthropology 19, 313–331.

Jonas, H., 1970. The nobility of sight: a study in the phenomenology of the senses. In: Spicker, S. (Ed.), Philosophy of the Body. Quadrangle Books, New York.

Lamp, F. (Ed.), 2004. See the Music, Hear the Dance: Rethinking African Art at the Baltimore Museum of Art. Baltimore Museum, Maryland, MD.

Laplantine, F., 2005. Le social et le sensible: introduction à une anthropologie modale. Téraèdre, Paris.

Lende, Daniel H., Downey, G. (Eds.), 2012. The Encultured Brain: An Introduction to Neuroanthropology. The MIT Press, Cambridge, MA.

Majid, A., Levinson, S.C. (Eds.), 2011. The Senses in Language and Culture, 6. The Senses and Society (1).

Meyer, Birgit (Ed.), 2009. Aesthetic Formations. Palgrave, New York.

Peck, A.M., 1929. A Vagabond's Provence. Dodd, Mead & Company, New York.

Pinard, S., 1991. A taste of India. In: Howes, D. (Ed.), The Varieties of Sensory Experience: A Source Book in the Anthropology of the Senses. University of Toronto Press, Toronto, ON.

Pink, S., 2009. Doing Sensory Ethnography. Sage, London.

Porcello, T., Meintjes, L., Ochoa, A.M., Samuels, D., 2010. The reorganization of the sensory world. Annual Review of Anthropology 39, 51–66.

Rivers, W.H.R., 1905. Observations on the senses of the Todas. British Journal of Psychology 1, 321–395.

Robben, A., Sluka, J. (Eds.), 2007. Ethnographic Fieldwork: An Anthropological Reader. Blackwell, Oxford, UK.

Stewart, K., 2007. Ordinary Affects. Duke University Press, Durham, NC.

Stoller, P., 1989 [1984]. The Taste of Ethnographic Things: The Senses in Anthropology. University of Pennsylvania Press, Philadelphia, PA.

Stoller, P., Olkes, C., 2005 [1990]. Thick sauce: remarks on social relations of the Songhay. In: Korsmeyer, C. (Ed.). The Taste Culture Reader, Berg, Oxford.

Sutton, D., 2010. Food and the senses. Annual Review of Anthropology 39, 209–223.

Van Campen, C., 2007. The Hidden Sense: Synesthesia in Art and Science. The MIT Press, Cambridge, MA.

Walmsley, E., 2005. Race, place and taste: making identity through sensory experience in Ecuador. Etnofoor 18 (1), 43–60.

Sentence Comprehension, Psychology of

Charles Clifton, Jr., University of Massachusetts, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of sentence comprehension focuses on how listeners and readers, guided by their knowledge of the grammar of their language, map recognized words onto the meanings of sentences. Psycholinguists have developed offline and online tasks for studying how listeners and readers comprehend sentences. This article reviews some of these tasks and some of the phenomena they have identified, including immediacy, structure-sensitivity, ambiguity resolution, lexical and frequency effects, and the effects of context and prosody. It outlines some of the theories that have been developed to account for these phenomena, and concludes by discussing some phenomena of how sentences are comprehended in realistic contexts.

Comprehending sentences is a process of mapping form (speech or printed text) onto sentence meaning. This article examines the cognitive challenges posed by this process, describes some experimental techniques that psycholinguists have used to understand it, summarizes some basic empirical phenomena and theories of how people identify the structural relations among words in a sentence (how sentences are 'parsed'), and concludes by discussing how sentences are interpreted in context.

The Tasks of Sentence Comprehension

Words are stored as units in lexical memory, but there are just too many sentences for there to be a 'lexicon' of sentences. A reader or listener has to construct or 'compose' the meaning of a sentence out of the meanings of the words in it, constrained by their grammatical relationships. Sentences, that is, must be built compositionally. This compositional process maps seemingly small differences in wording onto very large differences in meaning: Consider the sharp contrast between *The man helped the child to first base* and *The man helped the child on first base*. Listeners and readers also go beyond identifying the grammatical structure of a sentence and the meaning that it explicitly supports, and identify some of the propositions it implies. Social knowledge and knowledge about the world can guide such inferences. If you read *The man paid for the car*, you probably infer that he owns it; if you read *The man paid for the room*, you do not make the same inference. And if you hear someone say *Could you close that window* you recognize it as a request, not as a question about your abilities.

The task of mapping from print or sound to meaning begins with recognizing words, which is by itself not a simple task (*see* Visual Word Recognition, Neurocognitive Psychology of). The information that is retrieved when a word is identified makes an important contribution to sentence meaning (consider the difference between *to* and *on* in the example from the previous paragraph). Another example is the contrast between *Tom gave his dog biscuits* and *Tom gave him dog biscuits*. The difference between *his* and *him* signals a difference in the 'case' of the pronoun, an aspect of its morphology that carries information about the relation it has to a verb or some other 'case-assigning' word in a sentence. The form *him* signals accusative case, which means that the pronoun has some immediate relation to the verb, in this case the indirect object or 'recipient' of the action denoted by the verb. The form *his*, on the other hand, signals possessive or 'genitive' case, which means that the pronoun is immediately related to the following noun *dog*. This whole noun phrase *his dog* in turn takes the recipient role. (Actually, some ambiguity still remains: the larger noun phrase *his dog biscuits* could be the direct object, playing the role of 'theme,' with the recipient unspecified. Listeners and readers are often unaware of ambiguities like this.) Morphologically signaled case actually plays a minor role in English, although it plays a very important role in many other languages. Languages like English signal the structural relations between their words primarily by word order (compare *The dog bit the man* and *The man bit the dog*),

The task faced by listeners and readers is made particularly challenging by the ubiquitous ambiguity of language. Ambiguity begins in the speech stream, which often can be segmented at different points into different words (*see* Speech Perception; see also Cutler et al., 1997). A given word form can correspond to different lexical concepts (e.g., the word *bank* may refer to either a financial institution dealing with money or to a river bank). And a string of words can exhibit different structural relations, which may or may not be resolved by the end of a sentence. (When the ambiguity is resolved by the end of the sentence, the ambiguity is 'temporary'; sometimes the ambiguity is not resolved, although the reader or listener may not realize that it remains.) For instance, *I understood the article on word recognition was written by a real expert* has a temporary ambiguity. Listeners' and readers' knowledge of syntax (roughly, the hierarchical relations among words and phrases in a sentence) allows them to take *article on word recognition* to be either the subject of a complement sentence (…*was written by a real expert*) or (at least temporarily) the direct object of the verb *understood*. (If you have any trouble understanding this sample sentence, try adding the complementizer *that* right after *understood*.) Listeners and readers are generally not aware of such ambiguities (although puns force awareness; consider *A will is a dead giveaway*), but that only means that people's cognitive processes adequately resolve the ambiguities in the course of sentence comprehension. Studying how ambiguities are resolved has shed a good deal of light on these cognitive processes (see Section Phenomena of Sentence Comprehension, below).

Listeners' and readers' understanding of a sentence is guided not only by their knowledge of the words it contains but also by the morphology and syntax of their language. (Morphology is the structure of word forms, including the morphemes that determine word meaning plus morphemes that signal part of speech and various linguistic features, e.g., tense and number; syntax refers to the hierarchical relationships among the words and phrases of a sentence, or in other words, how linguistic elements such as words are put together to form constituents such as phrases or clauses.) Discourse context and world knowledge also affect comprehension. For instance, *I want the flowers on the table* means one thing if it is the answer to the question *Where do you want the flowers?* and something quite different if it answers *Which flowers do you want?* The context guides whether *on the table* is syntactically related to *want* or to *the flowers*.

Ways of Studying Sentence Comprehension

Offline Techniques

When psycholinguistics (a field that includes the study of the processes of sentence comprehension) was in its infancy, experimental techniques were largely limited to 'offline' techniques such as what is remembered when a sentence is comprehended (Miller, 1962). (Offline tasks gather information about the products of sentence comprehension, e.g., what remains in memory after a sentence is heard or read; online tasks measure real-time aspects of the process of sentence comprehension, e.g., how fast a particular word is comprehended.) These techniques did uncover some interesting phenomena, e.g., grammatical structure predicts which sentences will be confused in memory, and continue to do so.

One simple offline technique involves having speakers of a language make judgments about the acceptability of candidate sentences. They are typically asked something like, *Is this something that a competent speaker of your language would say*, or, *Is this something you can understand easily?* This is a version of what linguists typically do in their armchairs – make judgments about whether a sentence is good or not, in deciding whether the grammars they write should permit ('generate') the sentence. However, in the hands of experimental linguists and psycholinguists, acceptability judgments can be collected using proper experimental techniques such as controlling and counterbalancing sentence content, using subjects whose judgments are not confounded by their theoretical positions, and obtaining data whose stability can be evaluated using statistical methods (Cowart, 1997). These days, such judgments can even be collected quickly and easily from large numbers of participants using the Internet. Acceptability judgment experiments have supported interesting claims. For instance, Arregui et al. (2006) showed that elliptical sentences that violate a grammatical constraint that the elided material must match the form of its antecedent became more acceptable if the speaker was more likely to have misremembered the antecedent's form. For instance, people more often misremember passive sentences as active than vice versa (Mehler, 1963). Accordingly, they find a sentence like *Mary was scolded by her father, and today her mother did too* to be more acceptable than *Mary's father scolded her, and today her sister was too*. It appears that a listener or reader is more willing to accept a sentence that violates his or her knowledge of grammar if it seems likely that the speaker or writer could have made a mistake that led to the violation.

Various other offline techniques have been used. Some are as simple as asking for interpretations of ambiguous sentences. For instance, one can study how the prosody of a sentence (its rhythm and melody) affects its interpretation. Placing a prosodic boundary (a change in pitch together with lengthening and even a pause) at different points in a sentence can change its syntactic analysis, and if there are multiple prosodic boundaries, their relative sizes matter. For instance, Clifton et al. (2002) played listeners' versions of discourses like *Who arrived? Johnny and Sharon's in-laws* with prosodic boundaries after *Johnny* and after *Sharon's*. When the boundary after *Johnny* was larger than that after *Sharon's*, people said it was Sharon's in-laws and Johnny who came; when it was the same size or smaller, people said Johnny's and Sharon's in-laws came. The long tradition of studying what is remembered from sentences continues to provide useful information (e.g., about how memory for material that is focused – generally informative material, which can be prosodically marked – is enhanced).

Online Reading Techniques

Sentence comprehension seems almost effortless and automatic, but is really a very complex process. Psycholinguists have devised 'online tasks' to probe into the fine-grain details of this smoothly flowing process. Some early psycholinguists used a technique of sentence verification, measuring the time taken to decide whether a sentence was true or false (e.g., when spoken of a picture). Others (e.g., Mitchell, 1984) measured the time readers took to read each word or phrase of a sentence when sentence presentation was under their control (pressing a button brings on each new word), thereby getting a more precise look at the difficulty readers experience word-by-word. Pressing buttons to read word-by-word is not a highly developed and natural skill and may disrupt normal reading processes. However, the development of techniques for measuring eye movements during reading (Rayner et al., 1989; *see also* Eye Movements in Reading) overcame this objection and, together with the self-paced reading technique, has provided very illuminating data about language comprehension (Clifton et al., 2007).

Online Listening Techniques

Psycholinguists have also developed ways of measuring the process of comprehending sentences during listening. One very useful technique measures event-related brain potentials, electrical traces of underlying brain activities measured at the scalp (*see* Cognitive Neuroscience) at critical points in sentences, e.g., when a sentence becomes implausible or ungrammatical or simply difficult to process. There appear to be distinct signatures of different types of comprehension disruption. For instance, implausibility or violation of expectations commonly leads to an 'N400,' a negative-going voltage shift that peaks about 400 ms after the introduction of the implausibility. This allows a researcher to test theories about

just when and how disruption will appear in hearing or reading a sentence. Another illuminating technique (the 'visual world' technique) relies on the fact that people tend to look at what is being talked about. Researchers have used this fact to study such phenomena as the time course with which sentences are interpreted by presenting subjects with displays of objects and seeing where they look as they hear a sentence (Tanenhaus and Trueswell, 2006).

Phenomena of Sentence Comprehension

Immediacy

Early theories of processing, guided by linguistic analyses that posited 'transformational' rules whose domain could be as large as a full clause, proposed that full sentence analysis and interpretation took place only at a clause boundary. These theories were supported by evidence that memory for verbatim information declines across a clause boundary, that clauses resist interruption by extraneous sounds, and that readers pause at ends of clauses and sentences to engage in 'wrap-up' processes (Fodor et al., 1974). However, a great deal of evidence now exists indicating that readers do substantial interpretation on a word-by-word basis, without waiting for a clause boundary. The initial evidence for this claim came from Marslen-Wilson's (1973) work on 'fast shadowing' (repeating what one hears, with no more than a quarter-second delay). He showed that if the shadowed message contained a mispronunciation of a word's end (e.g., *cigareppe*, where *cigarette* is expected), the shadower would fairly often correct it spontaneously. As this happened more often in normal than in scrambled sentences, the listener must be using grammatical structure or meaning within a quarter-second or so to constrain recognition of words.

Evidence from measuring eye movements in reading led to a similar conclusion (*see* Eye Movements in Reading). For instance, if a person reads *While Mary was mending the sock fell off her lap*, the reader's eyes are likely to fixate a long time on the word *fell* or to regress from that word to an earlier point in the sentence. This happens because the verb *fell* is grammatically inconsistent with the normally preferred structural analysis of … *mending the sock*; *the sock* must be the subject of *fell*, not the object of *mending*. Sentence meaning is also processed on a nearly word-by-word basis. Eye movements are disrupted immediately upon reading the anomalous *carrots* in a sentence like *John used a pump to inflate the large carrots last night*. This pattern of results indicates that readers (and presumably listeners) create and evaluate grammatical and meaning relations word-by-word, without waiting for a clause boundary.

It is tempting to hypothesize that listeners and readers always perform a full semantic analysis of sentences word-by-word, without ever delaying this analysis. This 'immediacy hypothesis' may be too strong. Some evidence indicates that, while true lexical ambiguities (e.g., the difference between bank of a river and bank for money) may be resolved word-by-word, sense ambiguities (e.g., the difference between newspaper as something to read and newspaper as an institution) may not be. Similarly, determining the antecedent of a pronoun may be a task that readers do not always complete at the first opportunity. Nonetheless, one secure conclusion is that listeners and readers generally understand a great deal of what a sentence conveys with little or no lag.

Structure Sensitivity

Words are identified almost immediately as they are heard or read, but that is not the end of comprehension. The relations among words are identified almost immediately, constrained by the reader or listener's knowledge of linguistic structure as well as by the discourse context. These structural relations then constrain the possible meanings of a sentence or sentence fragment. The fragment *I understood the article on word recognition* … means one thing if *the article* is taken as the grammatical direct object of *understood*, and something quite different if it is taken as the subject of a new clause (the 'sentential complement' of *understood*).

Many psycholinguists have taken the generative grammars that have been developed following Chomsky (1957) as useful descriptions of the grammatical knowledge that listeners and readers use in identifying sentence structures. One important structure is phrase structure, the division of a sentence into its hierarchically arranged phrases and the labeling of these phrases. For example, *I understood the article on word recognition was written by an expert* is divided into the subject *I* and the verb phrase *understood … expert*; this verb phrase is divided into the verb *understood* and its embedded sentence complement *the article … was written by an expert*; this sentence is similarly divided into subject and verb phrase; etc. Another aspect of structure is the long-distance dependency between 'moved items' or 'fillers' and their 'traces' or 'gaps' (as in the relation between a *which*-phrase and its source, in *Which article did you just read t*, where the *t* stands for the 'trace' of *Which article*). Still other structures involve case relations (e.g., the distinction between accusative and genitive case discussed earlier) and thematic role relations (e.g., the distinction between theme or affected object and recipient in the sentence *Tom gave his dog biscuits*).

Early psycholinguists devoted much attention to the 'psychological reality' of grammatical structures, arguing whether they are really involved in sentence comprehension (Fodor et al., 1974). As psycholinguistic theory developed, it became apparent that the real debate involves not whether these structures are real but how they are identified in the course of sentence comprehension and how, in detail, the mind creates and represents them. It now seems impossible to explain how we understand sentences without theorizing about how people assign structure to sentences. However, as we will see later, there is very lively debate about just how they do identify sentence structure.

Ambiguity Resolution

Many sentences contain ambiguities, temporary or permanent. Listeners and readers are often not aware of them. They seem to use context, knowledge of the speaker or writer, etc., to arrive at a sensible interpretation. But sometimes readers make the wrong interpretation of an ambiguity, and find themselves confused when they realize their error. They are said to have been 'led down the garden path' by the sentence (Frazier, 1987). (In psycholinguistics, 'led down the garden path'

refers to a phenomenon in which incorrect structural relations are assigned, sometimes temporarily, to the words and phrases of a sentence; the term comes from Victorian novels, where the villain leads the innocent young thing 'down the garden path.') A classic example of a garden-path sentence is *The horse raced past the barn fell*. This sentence is nearly incomprehensible; the reader initially takes *raced* to be the main verb, with *The horse* as its subject, but really *the horse* has to be the object of the passive participle *raced*. The sentence is not actually impossible; it is basically just like *The car driven over the cliff exploded*, which lacks the garden path and is easy to comprehend.

The mistakes listeners and readers make can be extremely diagnostic of the decision rules they follow. A simple empirical generalization (which may or may not describe the cognitive process a reader or listener is engaging in; see Section Psychological Models of Sentence Comprehension) is that when a choice between two analyses has to be made, the reader/listener initially favors (1) the choice that is simpler in terms of grammatical structure, or (2) in the absence of complexity differences, the choice that permits new material to be related to the most recently processed old material. In a sentence like *I knew the answer was wrong*, the direct object analysis (… *knew the answer*) is grammatically simpler than the sentence complement analysis (… *knew the answer was wrong*). In the sentence *While Mary was mending the sock fell off her lap*, a reader/listener prefers to relate *the sock* to the recently processed verb *mending* rather than to material that has not yet been received. In both cases, these preferences are disconfirmed by material that follows, resulting in a garden path.

A great deal of experimental work using techniques described earlier has indicated that reading is disrupted and distinctive patterns of scalp electrical potentials are observed at points where garden paths occur. Not all these garden paths are apparent to conscious introspection, but they can be observed experimentally. Normally, the rules listeners and readers use to decide on the grammatical structures of sentences function so smoothly that their operation cannot be observed. But just as the analysis of visual illusions allows insight into the normal processes of visual perception by identifying when they go astray, identifying when a rule for analyzing sentence structure gives the wrong answer can go far in determining what rule is actually being followed.

Lexical and Frequency Effects

The relations among words and phrases that a reader or listener must identify to comprehend a sentence are specified by the grammatical principles of a language, but individual words contain a great deal of grammatically relevant information. The difficulty, and possibly even existence, of garden paths are modulated by properties of the words in a sentence. For instance, the normal preference to take a noun phrase following a verb as its direct object (which leads to difficulty in sentences like *I understood the article on word recognition was written by an expert*) is reduced for verbs that are more often used with sentence complements than with direct objects. A similar phenomenon can be observed for the much-studied ambiguity between a main verb and a reduced relative clause analysis. Consider the sentence fragment *The patient examined* … A reader who analyzes this fragment in the most straightforward way, as subject + main verb, will have been led down the garden path if it continues … *by the doctor*. Such a continuation indicates that *examined* really began a reduced relative clause that modifies *the patient*, so *examined* was not the main verb of the sentence. (If you have trouble with that, put *who was* before *examined*.) The difficulty of such sentences is reduced when the verb is one that strongly requires a direct object. Compare *The dictator captured in the coup was hated* with *The dictator fought in the coup was hated*; the former is easier, presumably because *capture* requires a direct object.

Plausibility and frequency of use also matter. Sentences that contain a temporary main verb versus reduced relative clause ambiguity are easier to understand when the subject is implausible as agent of the first verb but plausible as its theme are easier than sentences in which the subject is a plausible agent (e.g., *The evidence examined by the lawyer was unreliable* is easier than *The defendant examined by the lawyer was unreliable*). (However, this difference in difficulty of comprehension remains when the sentences are made unambiguous by adding *who was*, suggesting that the effect of plausibility should not be attributed to how it blocks the syntactic misanalysis.) Similarly, more frequent constructions are easier to comprehend. For example, if a verb is used relatively frequently as a participle as compared with its use as a simple past tense, the difficulty of the main verb/reduced relative clause garden path is reduced. MacDonald et al. (1994) reviewed these findings and spelled out one theoretical interpretation, which will be considered shortly.

Effects of Context

Sentences are normally not interpreted in isolation, which raises the question of how context can affect the processes by which they are understood. Some researchers have argued that normally observed garden paths could be eliminated when they are not supported by the situational context in which they are heard. Evidence supporting this has come from experiments using the 'visual world' technique (briefly described in Section Online Listening Techniques). Listeners hear an instruction like *Put the apple on the towel into the box* while they are looking at a display of objects. When there is one apple on a towel, an empty towel, and a box, they look first to the apple and then to the towel (and then follow the instruction). But if there are two apples (one of which is on a towel), an empty towel, and a box, they look just at the apple on the towel and not at the empty towel. The existence of the two apples apparently induced the listener to take the phrase *on the towel* as modifying *the apple* and not as an argument of the verb *put* (which is the normally preferred analysis).

Effects of Prosody

Spoken sentences carry one kind of information that is not well represented in print, the prosody (rhythm and melody) of a sentence. Prosody demonstrably influences how a sentence is comprehended. The presence of a prosodic boundary can affect the interpretation of a sentence. It can even eliminate the normal preferences that result in garden paths. For example the proper pattern of prosodic boundaries (including a boundary

after *studied*) can eliminate the normal difficulty of sentences like *Because John studied the material is clearer now* (in which *the material* is taken as the direct object of *studied*). However, there is more to prosody than just putting in pauses. The entire prosodic pattern of an utterance (including the relative size of prosodic boundaries, and the presence and location of points of intonational prominence called 'pitch accents') can affect how it is interpreted (Frazier et al., 2006).

Psychological Models of Sentence Comprehension

Psycholinguists have learned a great deal about the phenomena of sentence comprehension. This does not mean that they agree about the cognitive processes of listeners and readers that produce these phenomena. Early theories of how people understand sentences claimed that the rules that make up a linguistic theory of a language are known (implicitly) to language users, who use the rules directly in comprehending language (see Fodor et al., 1974). This did not work with the rules provided by early transformational grammars. Their rules operated on domains as large as a clause, and it soon became clear that people did not wait for the end of a clause to understand a sentence.

These early direct incorporation theories were replaced by what can be called 'detective style' theories (Clark and Clark, 1977). Listeners and readers were presumed to search for clues of any kind to the relations among words in a sentence. Psycholinguistic theorizing became little more than listing 'cues' that people used.

In the 1970s, a new breed of grammar-based theories developed. These theories added descriptions of decision strategies to the claim that people use the rules provided by the grammar, focusing on how they choose among various alternative rules in analyzing a sentence (Frazier, 1987; cf Frazier and Clifton, 1996; Mitchell, 1994; for further discussion of these theories). They claimed, for instance, that listeners and readers incorporate each new word into a sentence structure in the simplest and quickest possible way. This led to compelling accounts of garden-path phenomena, and the theories under discussion are often referred to as 'garden-path' theories.

In the 1970s and 1980s, linguistic theory moved away from positing broad, generally applicable rules to making claims about information contained in individual lexical items (constrained by some extremely broad general principles). This encouraged the development of 'lexicalist' theories of sentence comprehension (see MacDonald et al., (1994); for the most completely developed statement of such a theory). These theories emphasize the specific contribution of individual words rather than the effects of conformity to generally applicable word configurations. An individual word can provide a variety of kinds of information, including the possible and preferred thematic roles for its arguments, the frequency with which it is used in various constructions, and the entities it can refer to, as well as information about the phrase structure configurations it can appear in. Lexicalist theories propose that all these kinds of information are available and are used in deciding among alternative analyses.

Contemporary theories of sentence comprehension differ in the range of information that, they claim, guides sentence analysis. Garden-path theories are 'modular' in that they claim that only certain necessarily relevant types of information affect initial decisions about sentence structure. Lexicalist theories are generally nonmodular in allowing many different kinds of information to affect these decisions, and are often cast as constraint-based connectionist models in which all sorts of information (grammatical, lexical, contextual, world knowledge, etc.) contribute simultaneously to sentence comprehension.

Garden-path theories explain garden paths elegantly, and provide a more adequate account of how sentence structures are created than lexicalist theories do. On the other hand, they are forced to ascribe the effects of lexical structure, frequency, and context described above to a 'reanalysis' stage, which follows discovery that the initially preferred analysis is incorrect (see Fodor and Ferreira, 1998). Lexicalist theories, on the other hand, can provide a natural account of effects of lexical structure, frequency, and context, and are widely accepted in the psycholinguistic community. But they have not yet been developed in a way that adequately explains where sentence structures come from (it is unappealing to say that large parts of the grammar are stored with each lexical entry) or that explains why structurally complex sentences are overwhelmingly the ones that lead to garden paths (see Clifton and Staub, 2008; for some further criticisms).

Recent years have seen the emergence of 'memory-based' and 'statistical' models of language comprehension. The former (e.g., Lewis and Vasishth, 2005) apply basic principles of short-term memory to how sentences are comprehended. The latter (e.g., Levy, 2008) claim that listeners and readers form word-by-word expectations of upcoming sentence structure, based on the frequency with which they have experienced different structures in different contexts. Sentence comprehension goes more smoothly when a sentence structure is relatively strongly expected than when an unexpected structure is encountered. In some instances, these models have been implemented computationally, and are able to simulate many of the phenomena described above.

Interpretation

This article has focused on how listeners and readers use their knowledge of language to compose sentence meanings from words. While much of the power of language comes from listeners' and readers' abilities to 'follow the rules' and arrive at grammatically licensed interpretations of sentences, it is also clear that people are influenced by the discourse in which the sentence appears. Sentences are understood easily when they fit well into a discourse (e.g., when they accurately signal what is new to the discourse vs. what is old, or when they address a question that was raised implicitly or explicitly by prior discourse). The context in which an utterance is heard or a sentence is read can profoundly influence its interpretation. The meaning of words can be altered depending on the context in which they occur (consider *bank* in the context of a trip to the river to fish vs. a purchase of a house, or *red* in the context of a fire truck vs. a red-haired woman). A listener can determine whether a speaker who says *Can you open this door?* is asking about the adequacy of a job of carpentry versus requesting some

assistance. Speakers can rely on listeners to get their drift when they use figurative language such as metaphors and possibly irony (*My copy editor is a butcher*). And very generally, what a listener or reader takes a speaker or writer to mean may depend on their shared goals and mutual knowledge (cf Clark, 1996). Still, all these varied phenomena of interpretation depend on the listener's/reader's ability to put words together into sentence meanings, which makes the study of the topics considered in this article an important branch of cognitive psychology.

See also: Eye Movements in Reading; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Speech Perception; Visual Word Recognition, Neurocognitive Psychology of.

Bibliography

Arregui, A., Clifton Jr., C., Frazier, L., Moulton, K., 2006. Processing elided verb phrases with flawed antecedents: the recycling hypothesis. Journal of Memory and Language 55 (2), 232–246.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.

Clark, H.H., Clark, E.V., 1977. Psychology and Language. Harcourt Brace Jovanovich, New York.

Clark, H.H., 1996. Using Language. Cambridge University Press, Cambridge.

Clifton Jr., C., Carlson, K., Frazier, L., 2002. Informative prosodic boundaries. Language and Speech 45 (Pt 2), 87–114.

Clifton Jr., C., Staub, A., 2008. Parallelism and competition in syntactic ambiguity resolution. Language and Linguistics Compass 2, 234–250.

Clifton Jr., C., Staub, A., Rayner, K., 2007. Eye movements in reading words and sentences. In: Gompel, R.V., Fisher, M., Murray, W., Hill, R.L. (Eds.), Eye Movement Research: Insights into Mind and Brain. Elsevier, New York, pp. 341–371.

Cowart, W., 1997. Experimental Syntax: Applying Objective Methods to Sentence Judgments. Sage Publications Inc, Thousand Oaks, CA.

Cutler, A., Dahan, D., Van Donselaar, W., 1997. Prosody in the comprehension of spoken language: a literature review. Language and Speech 40, 141–201.

Fodor, J.A., Bever, T.G., Garrett, M.F., 1974. The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar. McGraw-Hill, New York.

Fodor, J.D., Ferreira, F. (Eds.), 1998. Sentence Reanalysis. Kluwer Academic Publishers, Dordrecht.

Frazier, L., 1987. Sentence processing: a tutorial review. In: Coltheart, M. (Ed.), Attention and Performance. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 559–586.

Frazier, L., Carlson, K., Clifton Jr., C., 2006. Prosodic phrasing is central to language comprehension. Trends in Cognitive Sciences 10 (6), 244–249. http://dx.doi.org/10.1016/j.tics.2006.04.002.

Frazier, L., Clifton Jr., C., 1996. Construal. MIT Press, Cambridge, MA.

Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106, 1126–1177.

Lewis, R.L., Vasishth, S., 2005. An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science 29, 375–420.

MacDonald, M.C., Pearlmutter, N.J., Seidenberg, M.S., 1994. Lexical nature of syntactic ambiguity resolution. Psychological Review 101, 676–703.

Marslen-Wilson, W.D., 1973. Linguistic structure and speech shadowing at very short latencies. Nature 244, 522–523.

Mehler, J., 1963. Some effects of grammatical transformations on the recall of english sentences. Journal of Verbal Learning and Verbal Behavior 2, 346–351.

Miller, G.A., 1962. Some psychological studies of grammar. American Psychologist 17, 748–762.

Mitchell, D.C., 1984. An evaluation of subject-paced reading tasks and other methods of investigating immediate processes in reading. In: Kieras, D.E., Just, M.A. (Eds.), New Methods in Reading Comprehension Research. Erlbaum, Hillsdale, NJ, pp. 69–89.

Mitchell, D.C., 1994. Sentence parsing. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, New York, pp. 375–410.

Rayner, K., Sereno, S., Morris, R., Schmauder, R., Clifton Jr., C., 1989. Eye movements and on-line language comprehension processes. Language and Cognitive Processes 4, SI21–SI49.

Tanenhaus, M., Trueswell, J.C., 2006. Eye movements and spoken language comprehension. In: Traxler, M.J., Gernsbacher, M.A. (Eds.), Handbook of Psycholinguistics, second ed. Academic Press, London, pp. 863–900.

Sentencing, Trends, and Disparities: Crime

Mindy S Bradley, University of Arkansas, Fayetteville, AR, USA
Jeffery T Ulmer, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Many scholars have examined sentencing disparities in efforts to explain why and how these differences occur and untangle the legal and extralegal influences on punishment. This article provides an overview of the common criminological perspectives found in the sentencing literature. We then summarize trends in sentencing research over the last three decades, reviewing much of the empirical research on disparities in the areas of race and ethnicity, gender, and mode of conviction. We conclude with a discussion of some of the gaps and unexplored/underexplored areas in the literature, making suggestions for future research in this field.

Introduction

A growing volume of research examines the determinants of criminal punishment and the degree to which differences in sentencing outcomes across types of offenders are legitimate. While there is no objective standard for deciding what constitutes warranted or unwarranted disparity, some differences in sentencing can be considered legally defensible. Many would agree that certain offenses are more serious than others, and that presence of aggravating or mitigating circumstances or other unique aspects of crimes may signal that relatively more or less severe penalties are appropriate. However, the extent to which sentencing disparities are unjustified is an area of considerable scholarly inquiry, and some dissimilarities, such as those related to offender demographic characteristics or mode of conviction, are particularly controversial. Many scholars have examined sentencing disparities in efforts to explain why and how these differences occur and untangle the legal and extralegal influences on punishment. This article provides an overview of the common criminological perspectives found in the sentencing literature. We then summarize trends in sentencing research over the last three decades, reviewing much of the empirical research on disparities in the areas of race and ethnicity, gender, and mode of conviction. We conclude with a discussion of some of the gaps and unexplored/underexplored areas in the literature, making suggestions for future research in this field.

Theoretical Perspectives on Sentencing

Explanations of sentencing can be classified across three general themes: court communities, focal concerns, and organizational efficiency. Although we discuss each separately, we do not mean to imply that these ideas are contradictory. Rather, they may be more appropriately considered as complementary. Although different aspects of the sentencing process are their particular focus, they may well lead to similar or consistent predictions regarding sentencing outcomes. First, the *court community perspective* argues that sentencing and sentencing disparity cannot be understood without taking into account local courts' organizational structure, organizational culture, case-processing strategies, and relationship networks. Research utilizing this approach focuses on the interactional dynamics of courtroom workgroups, conceptualizing courts as networks of action and communication based on participants' common working environment, interdependencies between key stakeholders (typically the prosecutor's office, bench, and defense bar) and local legal and organizational culture. Several scholars follow this theme, emphasizing the role of social contexts of courts and their decision-making processes (for example, see Eisenstein et al. (1988), Savelsberg (1992), Dixon (1995), Ulmer (1997), and Ulmer and Kramer (1998)). They view sentencing disparities as a product of the informal practices and negotiations among court communities, as well as formal policies and legal structures.

Other researchers have expanded on ideas that are compatible with the court community perspective. For example, Savelsberg (1992) discusses how the 'substantive rationality' of local contexts and individual interests, philosophies, and biases may be act to undermine formally rational decision-making criteria like sentencing guidelines. This may be particularly likely to occur when criteria have been externally imposed on local arenas and individuals. These local substantive interests potentially produce variation in the degree and type of sentencing disparity. Similarly, Dixon's (1995) 'organizational context' theory incorporates the court community perspective and extends on its themes. She emphasizes factors such as the bureaucratization of prosecutors' offices and the judges' bench as variables that exert particular influence on sentencing and sentencing disparity, even under sentencing guidelines.

Second, the *focal concerns perspective* is viewed as extension of the court community approach. Focal concerns explanations (Steffensmeier et al., 1998; Steffensmeier and Demuth, 2000) conceptualize sentencing as a process that occurs within the framework of uncertainty, as court decision-makers formulate attributions about offenders and crimes and act to balance practical and safety concerns. The focal concerns perspective is consistent with Albonetti's (1991) 'uncertainty avoidance' and causal attribution explanations of sentencing and court decision-making in that both highlight the uncertainty and insufficient information regarding the recidivism risk and rehabilitative potential of offenders. Albonetti (1991) describes

sentencing as a situation of bounded rationality in which court actors make highly consequential decisions with insufficient information. Drawing from attribution theory in social psychology, she argues that, as a means of reducing uncertainty, decision-makers fall back on attributions about reoffending risk and/or rehabilitation potential that can be linked to race, gender, and other social status stereotypes. This can then result in extralegal sentencing disparity connected to these statuses. Similarly, focal concerns theory describes how court actors manage uncertainty by making attributions about case and defendant characteristics in terms of four focal concerns: blameworthiness, dangerousness, rehabilitative potential, and practical constraints and consequences. In the abstract, the use of and reliance on these focal concerns are universal, but the meaning, relative emphasis and priority, and situational interpretation of them are embedded in local court community culture, organizational contexts, and politics. The important point for sentencing disparity is that while court community actors may make attributions and interpretations regarding the focal concerns primarily on the basis of legally relevant factors, they may also draw on racial, gender, class, or age stereotypes about defendants. For example, some judges may perceive young black males as particularly dangerous or lacking rehabilitative potential compared to other types of defendants, and sentence accordingly.

A third approach focuses on the organizational prioritization of expedient case processing, suggesting that heightened emphasis on the pragmatic disposition of cases places particular importance on plea bargaining. Extending Michels (1915) principle of the 'imperatives of organizational maintenance' to criminal justice organizations, Dixon notes the importance of efficient processing of cases, positing that *organizational efficiency* may be an overriding priority in sentencing decisions. In other words, Dixon (1995) defines this perspective as situating "efficient case disposition as the operational goal that maintains a stable and orderly sentencing system" (p. 1162). Drawing on this work, Engen and Steen (2000) delineate an organizational efficiency model, stressing that the importance of organizational efficiency may supersede formal rules like laws or sentencing guidelines, or substantively rational concerns like perceived blameworthiness and dangerousness. Prosecutors and judges achieve efficient case processing by encouraging guilty pleas and avoiding trials (Engen and Steen, 2000: p. 1363):

> The goal of criminal justice processing, according to this perspective, is to induce guilty pleas as an organizationally efficient means to an end. Thus offenders who plead guilty are rewarded with less severe sentences than they could otherwise have received.

Although different aspects of the sentencing process are their particular focus, each of the three approaches may lead to similar or consistent predictions regarding sentencing outcomes. For example, the organizational efficiency perspective predicts that a fundamental source of sentencing disparity will be differences across mode of conviction. Specifically, defendants who plead guilty will be sentenced more leniently than comparable individuals convicted via trial (even when controlling for differences in crime severity and prior criminal histories). Neither focal concerns nor court community perspectives refute this idea. Rather, proponents of focal concerns theory would agree with the potential for disparity based on mode of conviction. By conceptualizing pressures for organizational efficiency as one facet of the focal concern of practical constraints and consequences, focal concern proponents agree that concerns regarding the timely and efficient disposition of cases may result in sentencing disparity between those who plead guilty and those who are convicted by trial. Similarly, throughout much of the research applying the court community and organizational context frameworks, the need to process cases in a timely manner, the desire to avoid court backlogs, and the collective practice of rewarding guilty pleas and/or penalizing trials are significant themes. Notably, the court community and organizational context perspectives emphasize *variation* between local courts in the perceived importance of efficiency as a practical constraint and consequence.

Empirical Research on Sentencing: Racial/Ethnic Disparities

In the 1983 National Academy of Science assessment of sentencing research by Blumstein et al. (1983), Hagan and Bumiller review 51 studies on race and sentencing. They detail the many definitional and methodological challenges that prevent scholars from reaching collective conclusions, including weaknesses in defining and distinguishing between legal and extralegal variables, differences in sentence-severity rankings, problems with sample selection, and variation in measurement techniques. Overall, the authors found that contemporary studies were increasingly including legally relevant influences of offense severity and prior record in their designs. These variables and other processual measures, such as bail status, defense counsel, and plea, appeared to explain much of the influence of race and socioeconomic status on sentencing outcomes. Hagan and Bumiller were careful to note that these studies did not produce fewer findings of racial discrimination; rather, they identified many structural contexts where discrimination exists. However, the strength of the reported relationships between race and sentencing often seemed weak since these variables were at opposite ends of the causal chain. The processual approach to the sentencing issue is important because it highlights ways in which mediating variables like employment and bail status can be incorporated in a disadvantaging way into the sentencing process.

Reports estimate that, in 2009, African-American males were over six times more likely to be incarcerated than white males. African-American females were approximately three times more likely to be imprisoned than white females (Guerino et al., 2012). Thus, despite trends that indicate the *magnitude* of racial differences in incarceration risk is decreasing, minorities still continue to be disproportionately overrepresented in correctional populations. While many factors may affect sentencing disparities, explanations of racial and ethnic differences are key areas of public and scholarly debate. Some have argued that more severe sentences imposed on minorities are artifacts of differences in offending patterns and histories across groups. That is, they argue that the tougher

sentences imposed on minority defendants are a consequence of the higher severity of their crimes and prior criminal histories, as well as other legally relevant factors courts draw upon to determine appropriate sentences. However, critics contend that racial/ethnic discrimination in the justice process perpetuates a system in which minority defendants are punished more harshly than white defendants who have committed similar offenses and have comparable criminal histories.

Two later comprehensive reviews (Spohn, 2000; Zatz, 2000) summarized research on sentencing up to 2000, primarily focusing on studies of disparity based on defendant social characteristics. Much of this research did provide evidence of racial/ethnic differences in punishment. For example, in her review of 32 studies using state-level data, Spohn (2000) found that 41 of 95 black vs white estimates (43.2%) were significant indicators of harsher sentences for black defendants. Only four estimates suggested blacks were treated more leniently. Of the 29 effects, 8 reported significantly more severe sentences for Hispanics than whites, while only two effects indicated Hispanics were treated more leniently. In addition, 15 of 22 effects in eight federal court studies showed significant black disadvantages, while 10 of 21 effects indicated Hispanic disadvantages. Similarly, Zatz (2000) concluded that the empirical literature demonstrates distinct racial disadvantages among less serious offenses and low-level felonies, where prosecutors have the greatest amount of discretion in charging and plea-bargaining decisions.

Reviews of the direct effect of race on sentence severity reveal decidedly mixed results, it appears that disparity is more apparent in the decision to sentence offenders to prison than in the length of sentence imposed (Chiricos and Crawford, 1995). Specifically, Spohn (2000) found that, out of the 117 estimates of race effects on sentence severity, 53% indicated that blacks were more likely to receive an active prison sentence, while only 31% indicated that blacks were more likely to receive a longer sentence length.

Both reports also highlighted research that not only determined whether 'race/ethnicity mattered' or 'gender mattered,' but went on to investigate when and how these offender characteristics affect sentencing. The authors reviewed studies investigating how the influences of race, ethnicity, and gender mutually conditioned one another, and were also conditioned by other, more legitimized, factors. For example, Spohn's (2000) review provided a quantitative summary of interaction effects in racial/ethnic disparity studies, including 11 studies examining whether race and/or ethnic effects were conditioned by other social statuses or characteristics, especially gender and/or age (p. 462). All 11 provided some evidence of such effects. In particular, four studies observed sentencing disadvantages for young black and/or Hispanic males, four revealed sentencing disadvantages for unemployed young black or Hispanic males, one provided evidence of disadvantages for poor minorities, and one found disadvantages for low education minorities. Steffensmeier et al. (1998) examined the combined effects of race, age and gender on the in/out decision and sentence length; they determined that young black males were punished more harshly than any other age-race-sex combination.

In addition, much of the research in the 1990s finds that race/ethnicity differences in sentencing were conditional on criminal history, offense type or severity, case-processing factors, jurisdiction characteristics, and victim characteristics. In her review, Spohn notes that eight studies found that blacks or Hispanics with more serious criminal histories were sentenced more severely. Ten studies concluded that race/ethnic differences were conditioned by type or severity of crimes. Three found that racial disadvantages were greatest in less serious crimes, and six found that racial disadvantages were pronounced in drug offenses. In addition, six studies found that race/ethnicity effects were conditioned by case-processing factors. Two studies found that blacks were sentenced more severely if they were detained prior to trial, and one study concluded that blacks represented by a public defender received harsher sentences than those retaining a private attorney. Three studies revealed that black defendants who pled guilty were sentenced more severely than comparable black defendants convicted by trial.

Also, both at the state and federal levels, studies consistently find that sentence severity varies by jurisdiction (e.g., across county or circuit courts). Furthermore, Ulmer and Kramer (1996) explicitly compared sentencing in three Pennsylvania counties. They found significant racial disparities in only two counties. In addition, Steffensmeier and Demuth (2000) revealed that disparities may vary by type of offense, and were most pronounced among drug offenses in US courts.

Finally, Spohn notes that one study found that blacks who victimized whites received more severe sentences than other victim-offender combinations. Other studies have also found that race interacts with other variables and affects sentence severity only in some types of cases (Spohn and Cederblom, 1991). Similarly, Zatz (2000) notes that patterns associated with the victim-offender racial dyad is strongest and most consistent among homicide and rape cases. However, she suggests that the significantly harsher penalties for black male offenders among other offenses (such as burglaries, which may have white victims) is emblematic of the generalized threat that black males are believed to pose. Other studies have also found that race only interacts with other variables and affects sentence severity in some types of cases (Spohn and Cederblom, 1991).

More recent reviews of the sentence literature have been provided by Mitchell (2005) and Ulmer (2012). Since 2000, the majority of research on decisions to incarcerate (in/out decisions) finds that blacks are more likely than whites to receive prison sentences (Steffensmeier and Demuth, 2000; Ulmer and Johnson, 2004). However, not all studies find this; Weidner et al. (2005) find that black offenders are neither more nor less likely than comparable white offenders to receive active prison sentences vs jail sentences or noncustodial sentences. Notably, the majority of studies do not find that statistically significant race/ethnic differences in sentence length (e.g., LaFrentz and Spohn, 2006). In fact, some studies indicate that white defendants receive longer sentences relative to black defendants (Wooldredge et al., 2005). There are some studies that do find a disadvantage for blacks in sentence length (Ulmer and Johnson, 2004; Shermer and Johnson, 2010), depending on the type of offense (Steffensmeier and Demuth, 2000).

In a meta-analysis synthesizing 71 published and unpublished studies, Mitchell (2005) found significant racial

disparities in the likelihood of incarceration; African-American defendants were more likely than white defendants to be sent to prison. With regard to the length of sentences ordered by the court, African-Americans were generally sentenced more harshly than white defendants. However, this effect was small and highly variable. Importantly, Mitchell noted, "the observed differences between whites and African-Americans were small, suggesting that discrimination in the sentencing stage is not the primary cause of the overrepresentation of African-Americans in U.S. correctional facilities" (p. 462).

A few works examining the severity of conviction offenses (or, conversely, charge reductions) generally do not find that black offenders are convicted of more serious charges than white offenders (Albonetti, 1991; Wooldredge et al., 2005; Shermer and Johnson, 2010). However, some suggest that prosecutors are more likely to charge black defendants under mandatory minimums, although it is not clear whether they significantly increase sentence severity (Engen, 2009). Others find that white defendants are more likely to receive substantial assistance departures (Hartley et al., 2007) and downward departures (Mustard, 2001) in federal courts. However, research examining the severity of conviction offenses (or, conversely, charge reductions) generally does not find that black defendants are convicted of more serious charges than whites (Albonetti, 1991; Wooldredge et al., 2005; Shermer and Johnson, 2010). Unfortunately, evidence regarding these and other important presentencing decisions, especially charge reductions, is still relatively limited.

Overall, though, research examining the effect of race/ethnicity on sentencing has yet to resolve this debate and achieve a collective social scientific consensus. Some studies have shown that minorities are sentenced more harshly than whites (Kramer and Ulmer, 1996; Spohn et al., 1981–1982; Zatz, 1984), even after crime seriousness, prior criminal record, and other legal variables are taken into account. Other studies have found either no significant racial differences (Klein et al., 1990) or that blacks are treated more leniently than whites (Gibson, 1977). Still other research has concluded that race influences sentence severity indirectly through its effect on variables such as bail status (LaFree, 1985a), type of attorney (Spohn et al., 1981–1982), or type of disposition (LaFree, 1985b; Spohn, 1992), or that race interacts with other variables and affects sentence severity only in some types of cases (Spohn and Cederblom, 1991).

Empirical Research on Sentencing: Sex/Gender Disparities

While research on gender differences tends to find that women are sentenced more leniently than men, some studies suggest that this is not always the case. Daly and Bordt's (1995) analysis of the statistical findings of research published until 1990 found that women received less severe sentences than men in approximately half of the studies examined, mixed or no effects in 25% of the studies, and harsher sentences in 25% of the studies. In their review, Daly and Bordt (1995) found that women were more likely to receive harsher sentences if charged with felony offenses, if the cases were tried in felony courts, and if the sentencing occurred in urban areas.

Many recent studies find that men are more likely than women to receive a prison sentence vs nonprison sentence (Mustard, 2001; Spohn and Holleran, 2000; Steffensmeier et al., 1998; Weidner et al., 2005; Wooldredge et al., 2005). Studies also find that men are more likely to receive longer terms of incarceration (LaFrentz and Spohn, 2006; Shermer and Johnson, 2010). However, some works fail to find significant gender effects on sentence length (Wooldredge et al., 2005) or the likelihood of incarceration (Curry et al., 2004). In addition, while Albonetti (2002) finds that women drug offenders tend to receive more lenient sentences in comparison to their male counterparts overall, they are disproportionately disadvantaged by increases in offense level.

Much of the scholarly discussion surrounding gender and sentencing centers on how gender conditions leniency. That is, authors explore the conditional effects of gender and other demographic variables. Zatz (2000) is careful to note that the relationship between leniency and gender is not universal-highlighting studies that attempt to distinguish among which women and under what conditions this association exists. She discusses the work of Kruttschnitt (1984), arguing that gender-based family roles may make women more economically dependent on male partners and other family members. Thus formal sanctions may be deemed less necessary to control women's behavior, as they are more subject to informal social controls. In addition, she details the familial-based paternalism explanation provided by Daly (1989), which advocates greater leniency to mothers than nonmothers in the interest of avoiding the pragmatic challenges of incarcerating women with children. Bickle and Peterson (1991) found that the effect of family variables was conditional on race. Specifically, for black mothers, simply having children was not sufficient. Rather, black women must have demonstrated that they were 'good mothers' in order to receive leniency from the court. Similarly, Spohn's (2000) review provided a quantitative summary of interaction effects, including studies examining whether race and/or ethnic effects were conditioned by other social statuses or characteristics, especially gender and/or age (p. 462). All interaction studies provided some evidence of such effects. In particular, four studies found sentencing disadvantages for young males who were black and/or Hispanic males, and four revealed sentencing disadvantages for unemployed young black or Hispanic males.

Authors also explore the conditional effects of gender and process-related variables or on explaining how decision-making may disproportionately disadvantage men/advantage women throughout stages of the justice system. Studies find that women receive leniency at different sentencing stages. Men are less likely to be receive diversion options vs prosecution in drug cases (Alozie and Johnston, 2000), and are less likely to have their charges dismissed (Kingsnorth et al., 2002; Miethe, 1987) or reduced (Miethe, 1987; Mustard, 2001; Shermer and Johnson, 2010).

Empirical Research on Sentencing: Disparities across Mode of Conviction

Critics (e.g., Smith, 1986) have argued that 'plea bargains' are somewhat of a myth, in that pleading guilty may not result in

actual 'bargains' for defendants. Indeed, some offenders who plead guilty would have been acquitted or convicted of lesser offenses had they exercised their right to trial. In addition, some studies failed to find evidence that sentences significantly differed across mode of conviction (Eisenstein and Jacob, 1977). However, the bulk of evidence since the 1980s indicates that, among persons who are convicted, whether defendants pled guilty or were found guilty by trial has a unique and significant influence on their sentence (Blumstein et al., 1983; Ulmer and Bradley, 2006; Kramer and Ulmer, 2009). Moreover, although the size of the plea/trial disparity may vary, the mode of conviction (usually defined as simply whether the defendant pleaded guilty or was convicted at trial) is frequently among the strongest predictors of sentence severity, in some instances rivaling the effects of 'legally relevant' variables such as offense severity, weapon use, or prior convictions and generally having a greater impact than offender status characteristics such as gender, race/ethnicity, or age (Albonetti, 1991, 1998; Ulmer et al., 2010; Ulmer and Johnson, 2004; Miethe and Moore, 1985).

Many studies on other aspects of sentencing include controls for mode of conviction in their models. Although not the focus, by controlling for mode of conviction, these studies imply that whether a defendant has pled guilty or been convicted by trial may have a unique and significant influence on his/her sentence. In much of these works, there appears to be a 'trial penalty,' in that trial-convicted defendants are sentenced more severely than guilty-pled defendants (for some examples, see Albonetti, 1991; Dixon, 1995; Spohn et al., 1981–1982; Steffensmeier et al., 1998; Steffensmeier and Demuth, 2000; Steffensmeier and Hebert, 1999; Steffensmeier et al., 1993; Ulmer and Kramer, 1996) Most, although not all, of the studies that do focus on the effects of different modes of conviction conclude that those found guilty by trial, particularly jury trials, receive more severe sentences than similarly situated defendants who plead guilty (for example, Brereton and Casper, 1982; Johnson, 2003; Ulmer, 1997; Ulmer and Bradley, 2006). For example, in an examination of sentencing guidelines across several states, King et al. (2005) found significant 'process discounts' for some offenders. Reductions in sentence were linked to defendant conduct beyond the offense, including pleading guilty and providing assistance to law enforcement.

Moreover, offender and case characteristics may moderate the size of the trial penalty. Early research by Rhodes (1979) finds that the size of plea-trial sentencing differences varies by type of offense. He found sentencing differences across mode of conviction in robbery convictions, but not among convictions for other thefts and assault. Later studies provide further support for the notion that losing at trial can be particularly deleterious for some kinds of offenders, including African-American defendants and defendants with more extensive criminal histories (for example, Smith, 1986; Ulmer, 1997). Relatedly, Johnson (2003), in a comparison of bench vs jury convictions, found that mode of conviction conditioned the impact of race on disparities. Specifically, African-Americans were least likely to obtain downward guideline departures following trials, especially jury trials.

In addition to other factors, assessments of defendants' culpability and moral character affect determinations of appropriate punishment as judges evaluate their potential for reoffending (see also Albonetti, 1991). Thus, findings with regard to mode of conviction are often framed in terms of exposure to negative information. Recently, Ulmer and Bradley (2006) apply these concepts to examine variation in the size of the trial penalty for serious violent offenders. They suggest that pleading guilty helps defendants avoid disclosing negative information, or 'bad facts' (see Ulmer and Bradley, 2006; Brereton and Casper, 1982; Flemming et al., 1992; Ulmer, 1997). Specifically, trials provide occasions for the disclosure of information, wherein court actors hear information that adversely impacts assessment of a defendant's character. During the course of trials, prosecutors 'dirty up' defendants – they provide judges with vivid descriptions of crimes, emotional victim testimony, and detailed information regarding the defendant's past wrongs and misbehavior (Ulmer and Bradley, 2006). This information, coupled with the increased publicity and attendant public scrutiny associated with trials, may discourage judicial leniency (Brereton and Casper, 1982).

Most researchers agree that the institutionalized practice of 'rewarding' those who plead guilty and 'penalizing' those who lose at trial serves a variety of administrative or organizational needs including, but not limited to, the need for efficiency in case processing (Dixon, 1995; Engen and Steen, 2000). Other have also stressed that the plea process provides a means by which courtroom workgroups are able to pursue substantive justice; fair punishments that take into account the unique circumstances of each case, such as the defendant's culpability or danger to the community (Ulmer and Bradley, 2006), and that can even mitigate the effects of harsh sentencing laws (Mather, 1979).

Consistent with this administrative or bureaucratic justice interpretation, Dixon (1995) found that the size of the plea-trial disparity was conditional on the level of court bureaucratization. Later examinations of federal sentencing found that district-level caseload pressure significantly increases the size of trial effects on both the likelihood of sentence departures (Johnson et al., 2008) and sentence severity (Ulmer et al., 2010). Other studies have provided support for this idea, finding that trial penalties among serious violent offenders tended to be larger in counties with larger caseloads (Ulmer and Bradley, 2006) and lower trial rates (Kramer and Ulmer, 2009).

Limitations

It has long been recognized that sentencing is conditioned by selection processes in earlier justice system decisions. Several chapters of the Blumstein et al. (1983) report, as well as others called for consideration of sentencing in the contexts of earlier criminal justice decisions, particularly pretrial detention, charging, and conviction. The years since the Blumstein et al. (1983) report have seen comparatively few studies that examine sentencing in the context of earlier decisions. Consequently, evidence regarding the roles of race and gender in association with presentencing decisions, especially charge reductions, is still relatively limited. One explanation for this relative dearth is the lack of available data.

A critical development instrumental to the proliferation of sentencing research was the advent of accessibility of collected by sentencing commissions in sentencing guideline states and the federal system during the late 1980 and 1990s. These data were disseminated by sentencing commissions. Sentencing commissions operate under legal mandates, which require that they collect and make such data available. As a result, researchers had access to high-quality postconviction, jurisdiction-wide data for substantial numbers of cases. These data facilitated progress by providing substantially improved measures of offense severity and criminal history, which had often been measured crudely in earlier research, including multiple sentencing-related outcomes, and very large samples drawn from large numbers of contextually diverse counties (for states) or district courts (with federal data).

Unfortunately, guideline sentencing data are limited to convicted offenders, and do not include preconviction information such as the type and number of original charges (and rarely include pretrial detention). Data sets that include preconviction data for large numbers of cases and multiple local jurisdictions have traditionally been difficult to obtain. Relatedly, the Blumstein et al. (1983) report called for sentencing research to examine the effects of defendant socioeconomic status variables on sentencing outcomes, and to disentangle the potential influence of race and ethnicity vs socioeconomic factors in sentencing. Defendant socioeconomic data and/or data on individual court actors has generally been hard to come by in the most widely used guideline data sets. Thus, studies that disentangle socioeconomic factors from other defendant characteristics (especially race or ethnicity) have not been as common.

Furthermore, we know relatively little about sentencing variation between criminal justice actors, especially judges and prosecutors, and the effects of decision maker characteristics in explaining variation in sentencing. Only a handful of studies have examined sentencing variation between judges, and the effect of judge characteristics, and even fewer studies have examined between-prosecutor variation.

Finally, the large majority of criminological literature on sentencing focuses on courts in the United States. We generally know very little about sentencing in other countries, especially non-Western ones. Thus, increased attention to sentencing in international contexts is a major need for criminology.

See also: Crime Trends and Debates; Crime and Gender; Crime, Race and Ethnicity; Imprisonment: Sociological Aspects; Punishment; Wrongful Convictions.

Bibliography

Albonetti, C.A., 1991. An integration of theories to explain judicial discretion. Social Problems 38, 247–266.

Albonetti, C.A., 2002. The effects of the 'safety valve' amendment on length of imprisonment for cocaine trafficking/manufacturing offenders: mitigating the effects of mandatory minimum penalties and offender's ethnicity. Iowa Law Review 87, 401–433.

Albonetti, C.A., 1998. The role of gender and departures in the sentencing of defendants convicted of white-collar offenses under the federal sentencing guidelines. Sociology of Crime, Law and Deviance 1, 3–48.

Alozie, N., Johnston, W., 2000. Probing the limits of the female advantage in criminal processing: pretrial diversion of drug offenders in an urban county. Justice System Journal 21, 239–259.

Brereton, D., Casper, J., 1982. Does it pay to plead guilty? Differential sentencing and the functioning of criminal courts. Law and Society Review 16, 45–70.

Bickle, G., Peterson, R.D., 1991. The impact of gender-based family roles in sentencing. Social Problems 39, 371–394.

Blumstein, A., Cohen, J., Martin, S., Tonry, M.H. (Eds.), 1983. Research on Sentencing: The Search for Reform, vol. II. National Academy Press, Washington, DC.

Chiricos, T.G., Crawford, C., 1995. Race and imprisonment: a contextual assessment of the evidence. In: Hawkins, Darnell F. (Ed.), Ethnicity, Race and Crime. State University of New York Press, Albany, NY, pp. 281–309.

Curry, T.R., Lee, G., Rodriguez, S.F., 2004. Does victim gender increase sentence severity? Further explorations of gender dynamics and sentencing outcomes. Crime & Delinquency 50, 319–343.

Daly, K., 1989. Rethinking judicial paternalism: gender, work-family relations, and sentencing. Gender and Society 3, 9–36.

Daly, K., Bordt, R., 1995. Sex effects and sentencing: an analysis of the statistical literature. Justice Quarterly 12, 141–176.

Dixon, J., 1995. The organizational context of criminal sentencing. American Journal of Sociology 100, 1157–1198.

Eisenstein, J., Jacob, H., 1977. Felony Justice: An Organizational Analysis of Criminal Courts. Little, Brown, Boston, MA.

Eisenstein, J., Flemming, R., Nardulli, P., 1988. The Contours of Justice: Communities and Their Courts. Little, Brown, Boston, MA.

Engen, R., 2009. Assessing determinate and presumptive sentencing: making research relevant. Criminology and Public Policy 8, 323–335.

Engen, R.L., Steen, S., 2000. The power to punish: discretion and sentencing reform in the war on drugs. American Journal of Sociology 105, 1357–1395.

Flemming, R.B., Nardulli, P.F., Eisenstein, J., 1992. The Craft of Justice: Politics and Work in Criminal Court Communities. University of Pennsylvania Press, Philadelphia, PA.

Gibson, J.L., 1977. Race as a determinant of criminal sentences: a methodological critique and a case study. Law & Society Review 12, 455.

Guerino, P., Harrison, P.M., Sabol, W.J., 2012. Prisoners in 2010 (NCJ 236096). Bureau of Justice Statistics, Washington, DC.

Griffin, T., Wooldredge, J., 2006. Sex-based disparities in felony disposition before and after sentencing reform in Ohio. Criminology 44, 893–924.

Hagan, J., 1977. Criminal justice in rural and urban communities: a study of the bureaucratization of justice. Social Forces 55, 597–612.

Hagan, J., Bumiller, K., 1983. Making sense of sentencing: a review and critique of sentencing research. In: Blumstein, Alfred, Cohen, Jacqueline, Martin, Susan, Tonry, Michael H. (Eds.), Research on Sentencing: The Search for Reform, vol. II. National Academy Press, Washington, DC, pp. 1–54.

Hartley, R., Maddan, S., Spohn, C., 2007. Prosecutorial discretion: an examination of substantial assistance departures in Federal crack-cocaine and powder cocaine cases. Justice Quarterly 24, 382–407.

Johnson, B., 2003. Racial and ethnic disparities in sentencing departures across modes of conviction. Criminology 41 (2), 449–488.

Johnson, B.D., Ulmer, J.T., Kramer, J.H., 2008. The social context of guideline circumvention: the case of federal district courts. Criminology 46, 737–783.

King, N., Soule, D., Steen, S., Weidner, R., 2005. When process affects punishment: differences in sentences after guilty plea, bench trial, and jury trial in five guideline states. Columbia Law Review 105 (4), 960–1009.

Kingsnorth, R., Macintosh, R., Sutherland, S., 2002. Criminal charge or probation violation? Prosecutorial discretion and implications for research in criminal court processing. Criminology 40, 553–579.

Klein, S., Petersilia, J., Turner, S., 1990. Race and imprisonment decisions in California. Science 247, 812–816.

Kramer, J.H., Ulmer, J.T., 2009. Sentencing Guidelines: Lessons from Pennsylvania. Lynne Rienner Publishers, Boulder, CO.

Kramer, J.H., Ulmer, J.T., 1996. Sentencing disparity and departure from guidelines. Justice Quarterly 13, 81–106.

Kruttschnitt, C., 1984. Sex and criminal court dispositions: the unresolved controversy. Journal of Research in Crime and Delinquency 21, 23–32.

LaFree, G.D., 1985a. Official reactions to Hispanic defendants in the Southwest. Journal of Research in Crime and Delinquency 22, 213–237.

LaFree, G.D., 1985b. Adversarial and nonadversarial Justice: a comparison of guilty pleas and trials. Criminology 23, 289–312.

LaFrentz, C., Spohn, C., 2006. Who is punished more harshly in federal court? The interaction of race/ethnicity, gender, age, and employment status in the sentencing of drug offenders. Justice Research and Policy 8, 25–56.

Mather, L., 1979. Plea Bargaining on Trial? The Process of Criminal Case Disposition. Lexington Books, Lexington, MA.

Michels, R., 1915. Political Parties. Free Press, Glencoe, IL.

Miethe, T.D., 1987. Charging and plea bargaining under determinate sentencing: an investigation of the hydraulic displacement of discretion. Journal of Criminal Law and Criminology 78 (1), 155–176.

Miethe, T.D., Moore, C.A., 1985. Socioeconomic disparities under determinate sentencing systems: a comparison of pre-guideline and post-guideline practices in Minnesota. Criminology 23 (2), 337–363.

Mitchell, O., 2005. A meta-analysis of race and sentencing research: explaining the inconsistencies. Journal of Quantitative Criminology 21 (4), 439–466.

Mustard, D.B., 2001. Racial, ethnic, and gender disparities in sentencing: evidence from the U.S. federal courts. Journal of Law and Economics 44, 285–314.

Myers, M.A., Talarico, S., 1986. Urban justice, rural injustice? Urbanization and its effect on sentencing. Criminology 24, 367–390.

Rhodes, W., 1979. Plea bargaining: its effect on sentencing and convictions in the District of Columbia. Journal of Criminal Law and Criminology 70, 360–375.

Savelsberg, J., 1992. Law that does not fit society: sentencing guidelines as a neo-classical reaction to the dilemmas of substantivized law. American Journal of Sociology 97, 1346–1381.

Shermer, L.O., Johnson, B., 2010. Criminal prosecutions: examining prosecutorial discretion and charge reductions in U.S. federal district courts. Justice Quarterly 27, 394–430.

Spohn, C., 2000. Thirty years of sentencing reform: the quest for a racially neutral sentencing process. In: Policies, Processes and Decisions of the Criminal Justice System. Criminal Justice 2000, vol. 3. U.S. Department of Justice, Washington, DC.

Smith, D.A., 1986. The plea bargain controversy. Journal of Criminal Law and Criminology 77, 949–968.

Spohn, C., 1992. An analysis of the 'Jury Trial Penalty' and its effect on Black and White defendants. The Justice Professional, 93–112.

Spohn, C., Cederblom, J., 1991. Race and disparities in sentencing: a test of the liberation hypothesis. Justice Quarterly 8, 305–327.

Spohn, C., Holleran, D., 2000. The imprisonment penalty paid by young unemployed Black and Hispanic male offenders. Criminology 38, 501–526.

Spohn, C., Gruhl, J., Welch, S., 1981–1982. The effect of race on sentencing: a re-examination of an unsettled question. Law & Society Review 16, 71–88.

Steffensmeier, D., Demuth, S., 2000. Ethnicity and sentencing outcomes in U.S. federal courts: who is punished more harshly? American Sociological Review 65, 705–729.

Steffensmeier, D., Kramer, J.H., Streifel, C., 1993. Gender and imprisonment decisions. Criminology 31, 411–446.

Steffensmeier, D., Ulmer, J.T., Kramer, J.H., 1998. The interaction of race, gender, and age in criminal sentencing: the punishment cost of being young, black, and male. Criminology 36 (4), 763–798.

Steffensmeier, D., Hebert, C.G., 1999. Women and men policymakers: does the judge's gender affect the sentencing of criminal defendants? Social Forces 77, 1163–1196.

Uhlman, T., Walker, N.D., 1980. "He takes some of my time, i take some of his": an analysis of judicial sentencing patterns in jury cases. Law and Society Review 14 (2), 323–341.

Ulmer, J.T., 2012. Recent developments and new directions in sentencing research. Justice Quarterly 1, 1–38.

Ulmer, J.T., 1997. Social Worlds of Sentencing: Court Communities under Sentencing Guidelines. State University of New York Press, Albany, NY.

Ulmer, J.T., Bradley, M.S., 2006. Variation in trial penalties among serious violent offenders. Criminology 44, 631–670.

Ulmer, J., Johnson, B.D., 2004. Sentencing in context: a multilevel analysis. Criminology 42 (1), 137–177.

Ulmer, J.T., Kramer, J.H., 1996. Court communities under sentencing guidelines: dilemmas of formal rationality and sentencing disparity. Criminology 34 (3), 383–407.

Ulmer, J.T., Kramer, J.H., 1998. The use and transformation of formal decision-making criteria: sentencing guidelines, organizational contexts, and case processing strategies. Social Problems 45 (2), 248–267.

Ulmer, J.T., Eisenstein, J., Johnson, B.D., 2010. Trial penalties in federal sentencing: extra-guidelines factors and district variation. Justice Quarterly 2, 560–592.

Weidner, R., Frase, R.S., Schultz, J.S., 2005. The impact of contextual factors on the decisions to imprison in large urban jurisdictions: a multilevel analysis. Crime and Delinquency 51, 400–424.

Wooldredge, J., Griffin, T., Rauschenberg, F., 2005. (Un)anticipated effects of sentencing reform on the disparate treatment of defendants. Law & Society Review 39, 835–874.

Zatz, M.S., 1984. Race, ethnicity, and determinate sentencing. Criminology 22, 147–171.

Zatz, M.S., 2000. The convergence of race, ethnicity, gender, and class on court decision making: looking toward the 21st century. In: Policies, Processes and Decisions of the Criminal Justice System. Criminal Justice 2000, vol. 3. U.S. Department of Justice, Washington, DC.

Sequential Statistical Methods

Donald A Berry, University of Texas, Houston, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13922–13927, © 2001, Elsevier Ltd.

Abstract

When using a sequential design the data that accrue can affect the future course of the experiment. For example, observations made early in an experiment can determine treatment assignment later in the experiment. The other principal example is when the accumulating results are used to determine when to stop the experiment. This article presents some of the methodology associated with both types of sequential experiment. Of particular importance in the sequential setting is whether one's statistical perspective is frequentist or Bayesian. This article describes the role of statistical philosophy in sequential experimentation, as well as its impact on scientific inference and decision making.

Statistics plays two fundamental roles in empirical research. One is in determining the data collection process: the experimental design. The other is in analyzing the data once it has been collected. For the purposes of this article, two types of experimental designs are distinguished: *sequential* and *nonsequential*. In a sequential design the data that accrue in an experiment can affect the future course of the experiment. For example, an observation made on one experimental unit treated in a particular way, may determine the treatment used for the next experimental unit. The term 'adaptive' is commonly used as an alternative to sequential. In a nonsequential design the investigator can carry out the entire experiment without knowing any of the interim results.

The distinction between sequential and nonsequential is murky. An investigator's ability to carry out an experiment exactly as planned is uncertain, as information that becomes available from within and outside the experiment may lead the investigator to amend the design. In addition, a nonsequential experiment may give results that encourage the investigator to run a second experiment, one that might even simply be a continuation of the first. Considered separately, both experiments are nonsequential, but the larger experiment that consists of the two separate experiments is sequential.

In a typical type of nonsequential design, 20 patients suffering from depression are administered a drug and their improvements are assessed. An example sequential variation is the following. Patients' improvements are recorded 'in sequence' during the experiment. The experiment stops should it happen that at least nine of the first 10 patients, or no more than one of the first 10 patients improve(s). On the other hand, if between two and eight of the first 10 patients improve then sampling continues to a second set of 10 patients, making the total sample size equal to 20 in that case. Another type of sequential variation is when the dose of the drug is increased for the second 10 patients should it happen that fewer than four of the first 10 improve.

Much more complicated sequential designs are possible. For example, the first patient may be assigned a dose in the middle of a range of possible doses. If the patient improves then the next patient is assigned the next lower dose, and if the first patient does not improve then the next patient is assigned the next higher dose. This process continues, always dropping the dosage if the immediately preceding patient improved, and increasing the dosage if the immediately preceding patient did not improve. This is called an 'up-and-down' design.

Procedures in which batches of experimental units (such as groups of 10 patients each) are analyzed before proceeding to the next stage of the experiment are called 'group-sequential.' Designs such as the up-and-down design, in which the course of the experiment can change after each experimental unit responds are called 'fully sequential.' So a fully sequential design is a group-sequential design in which the group size is one. Designs in which the decision of when to stop the experiment depends on the accumulating results are called 'sequential stopping.' Using rules to determine which treatments to assign to the next experimental unit or batch of units is called 'sequential allocation.'

Designs of most scientific experiments are sequential, although perhaps not formally so. Investigators usually want to conserve time and resources. In particular, they do not want to continue an experiment if they have already learned what they set out to learn, and this is so whether their conclusion is positive or negative, or if finding a conclusive answer would be prohibitively expensive. (Where the investigator discovers that the standard deviation of the observations is much larger than originally thought is an example of an experimental design that would be prohibitively expensive to continue because the required sample size would be large.)

Sequential designs are difficult or impossible to use in some investigations. For example, results might take a long time to obtain, and waiting for them would mean delaying other aspects of the experiment. Suppose one is interested in whether grade-schoolers diagnosed with attention deficit hyperactivity disorder (ADHD) should be prescribed Ritalin. The outcome of interest is whether children on Ritalin will be addicted to drugs as adults. Consider assigning groups of 10 children to Ritalin and 10 to a placebo, and waiting to observe their outcomes before deciding whether to assign an additional group of 10 patients to each treatment. The delay in observation means that it would probably take hundreds of years to get an answer to the overall question. The long-term nature of the endpoint means that any reasonable experiment addressing this question would necessarily be non-sequential, with large numbers of children assigned to the two groups before any information at all would become available about the endpoint.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42075-1

Analyzing Data from Sequential Experiments – Frequentist Case

Consider an experiment of a particular type, say one to assess extrasensory perception (ESP) ability. A subject claiming to have ESP is asked to choose between two colors. The *null hypothesis* of no ability is that the subject is only guessing, in which case the correct color has a probability of 1/2. Suppose the subject gets 13 correct out of 17 tries. How should these results be analyzed and reported? The answer depends on one's statistical philosophy. Frequentists and Bayesians take different approaches. Frequentist analyses depend on whether the experiment's design is sequential, and if it is sequential the conclusions will differ depending on the actual design used.

In the nonsequential case the subject is given exactly 17 tries. The frequentist *P*-value is the probability of results as extreme as or more extreme than those observed. The results are said to be 'statistically significant' if the *P*-value is less than 5%. A convention is to include both 13 or more successes and 13 or more failures. (This 'two-sided' case allows for the possibility that the subject has ESP but has inverted the 'extrasensory' signals.) Assuming the null hypothesis and that the tries are independent, the probabilities of the number of successes is binomial. Binomial probabilities can be approximated using the normal distribution. The z-score for 13 out of 17 is about 2, and so the probability of 13 or more successes or 13 or more failures is about 0.05 (the exact binomial probability is 0.049), and so the results are statistically significant at the 5% level and the null hypothesis is rejected.

Now suppose the experiment is sequential. The frequentist significance level is now different, and it depends on the actual design used. Suppose the design is to sample until the subject gets at least four successes and at least four failures – same data, different design. Again, more extreme means more than 13 successes (and exactly four failures) or more than 13 failures (and exactly four successes). The total probability of these extreme values is 0.021 – less than 0.049 – and so the results are now more highly significant than if the experiment's design had been nonsequential.

Consider another sequential design, one of a type of group-sequential designs commonly used in clinical trials. The experimental plan is to stop at 17 tries if 13 or more are successes or 13 or more are failures, and hence the experiment is stopped on target. But if after 17 tries the number of successes is between five and 12, then the experiment continues to a total of 44 tries. If at that time, 29 or more are successes or 29 or more are failures then the null hypothesis is rejected. To set the context, suppose the experiment is nonsequential, with sample size fixed at 44 and no possibility of stopping at 17; then the exact significance level is again 0.049. When using a sequential design, one must consider all possible ways of rejecting the null hypothesis in calculating a significance level. In the group-sequential design, there are more ways to reject than in the nonsequential design with the sample size fixed at 17 (or fixed at 44). The overall probability of rejecting is greater than 0.049 but is somewhat less than $0.049 + 0.049$ because some sample paths that reject the null hypothesis at sample size 17 also reject it at sample size 44. The total probability of rejecting the null hypothesis for this design is actually 0.080. Therefore, even though the results beyond the first 17 observations are never observed, the fact that they might have been observed makes 13 successes of 17 no longer statistically significant (since 0.08 is greater than 0.05). The three designs above are summarized in Table 1. The table includes a fourth design in which the significance level cannot be found.

To preserve a 0.05 significance level in group-sequential or fully sequential designs, investigators must adopt more stringent requirements for stopping and rejecting the null hypothesis; that is, they must include fewer observations in the region where the null hypothesis is rejected. For example, the investigator in the above study might drop 13 successes or failures in 17 tries and 29 successes or failures in 44 tries from the rejection region. The investigator would stop and claim significance only if there are at least 14 successes or at least 14 failures in the first 17 tries, and claim significance after 44 tries only if there are at least 30 successes or at least 30 failures. The *nominal significance levels* (those appropriate had the experiment been nonsequential) at $n = 17$ and $n = 44$ are 0.013 and 0.027, and the overall (or adjusted) significance level of rejecting the null hypothesis is 0.032. (No symmetric rejection regions containing more observations allow the significance level to be greater than this but still smaller than 0.05.) With this design, 13 successes out of 17 is not statistically significant (as indicated above) because this data point is not in the rejection region. The above discussion is in the context of significance testing. But the same issues apply in all types of frequentist inferences, including confidence intervals.

The implications of the need to modify rejection regions depending on the design of an experiment are profound. In view of the penalties that an investigator pays in significance level that are due to repeated analyses of accumulating data, investigators strive to minimize the number of such analyses. They shy away from using sequential designs and so may miss opportunities to stop or otherwise modify the experiment depending on accumulating results.

What happens if investigators fail to reveal that other analyses did occur, or that the experiment might have continued had other results been observed? Any frequentist conclusion that fails to take the other analyses into account is meaningless. Strictly speaking, this is a breach of scientific ethics when carrying out frequentist analyses. But it is difficult to find fault with investigators who do not understand the subtleties of frequentist reasoning and who fail to make necessary adjustments to their inferences.

For more information about the frequentist approach to sequential experimentation, see Whitehead (1992).

Table 1 Summary of experimental designs and related significance levels

Stopping rule	*Significance level*
After 17 observations (nonsequential)	0.049
After at least 4 successes and 4 failures	0.021
After 17 or 44 observations, depending on interim results	0.08
Stop when 'you think you know the answer'	Undefined

Analyzing Data from Sequential Experiments – Bayesian Case

When taking a Bayesian approach (*see* Bayesian Statistics) (or a likelihood approach), conclusions are based only on the observed experimental results and do not depend on the experiment's design. So the murky distinction that exists between sequential and nonsequential designs is irrelevant in a Bayesian approach. In the example considered above, 13 successes out of 17 tries will give rise to the same inference in each of the designs considered. Bayesian conclusions depend only on the data actually observed and not otherwise on the experimental design (Berger and Wolpert, 1984; Berry, 1987).

The Bayesian paradigm is inherently sequential. Bayes' theorem prescribes the way learning takes place under uncertainty. It specifies how an observation modifies one's state of knowledge (Berry, 1996). Moreover, each observation that is planned has a probability distribution. After 13 successes in 17 tries, the probability of success on the next try can be found. This requires a distribution, called a 'prior distribution,' for the probability of success on the first of the 17 tries. Suppose the prior distribution is uniform from 0 to 1. (This is symmetric about the null hypothesis of 1/2, but it is unlikely to be anyone's actual prior distribution in the case of ESP because it gives essentially all the probability to *some* ESP ability.) The predictive probability of a success on the 18th try is then $(13+1)/(17+2) = 0.737$, called 'Laplace's rule of succession' (Berry, 1996: p. 204). Whether to take this 18th observation can be evaluated by weighing the additional knowledge gained (having 14 successes out of 18, with probability 0.737, or 13 successes out of 18, with probability 0.263) with the costs associated with the observation.

Predictive probabilities are fundamental in a Bayesian approach to sequential experimentation. They indicate how likely it is that the various possibilities for future data will happen, given the data currently available. Suppose that after 13 successes of 17 tries one is entertaining taking an additional 27 observations. One may be interested in getting at least 30 successes out of the total of 44 observations – which means at least 17 of the additional 27 observations are successes. The predictive probability of this is about 50%. Or one may be interested in getting successes in at least 1/2 (22) of the 44 tries. The corresponding predictive probability is 99.5%.

The ability to use Bayes' theorem for updating one's state of knowledge and the use of predictive probabilities makes the Bayesian approach appealing to researchers in the sequential design of experiments. As a consequence, many researchers who prefer a frequentist perspective use the Bayesian approach in the context of sequential experimentation. If they are interested in finding the frequentist operating characteristics (such as significance level and power), these can be calculated by simulation.

The next section (Section Sequential Allocation of Experiments: Bandit Problems) considers a special type of sequential experiment. The goals of the section are to describe some of the calculational issues that arise in solving sequential problems and to convey some of the interesting aspects of sequential problems. It takes a Bayesian perspective.

Sequential Allocation of Experiments: Bandit Problems

In many types of experiments, including many clinical trials, experimental units are randomized in a balanced fashion to the candidate treatments. The advantage of a balanced design is that it gives maximal information about the differences between treatments. In some types of experiment, including some clinical trials, it may be important to obtain good results on the units that are part of the experiment. Treatments – or *arms* – are assigned based on accumulating results; that is, assignment is sequential. The goal is to maximize the overall effectiveness – of those units in the experiment, but also, perhaps, including those units not actually in the experiment but whose treatment might benefit from information gained in the experiment.

Specifying a design is difficult. The first matter to be considered is the arm selected for the initial unit. Suppose that the first observation is X_1. The second component of the design is the arm selected next, given X_1 and also given the first arm selected. The third component depends on X_1 and the second observation X_2, and on the corresponding arms selected. And so on. A design is *optimal* if it maximizes the expected number of successes. An arm is optimal if it is the first selection of an optimal design.

Consider temporarily an experiment with n units and two available arms. Outcomes are dichotomous: arm 1 has success probability p_1 and arm 2 has success probability p_2. The goal is to maximize the expected number of successes among the n units. Arm 1 is standard and has known success proportion p_1. Arm 2 has unknown efficacy. Uncertainty about p_2 is given in terms of a prior probability distribution. To be specific, suppose that this is uniform on the interval from 0 to 1.

If $n = 1$ then the design requires only an initial selection, arm 1 or arm 2. Choosing arm 1 has expected number of successes p_1. Choosing arm 2 has conditional expected number of successes p_2, and an unconditional expected number of successes, the prior probability of success, which is 1/2. Therefore arm 1 is optimal if $p_1 \geq 1/2$ and arm 2 is optimal if $p_1 \leq 1/2$. (Both arms – and any randomization between them – are optimal when $p_1 = 1/2$.)

The problem is more complicated for $n \geq 2$. Consider $n = 2$. There are two initial choices and two choices depending on the result of the first observation. There are eight possible designs. One can write a design as $\{a;\ a_s,\ a_F\}$, where a is the initial selection, a_s is the next selection should the first observation be

Table 2 Possible designs and associated expected number of successes

Design	*Expected number of successes*
{1; 1, 1}	$\mathbf{2p_1}$
{1; 1, 2}	$p_1 + p_1^2 + (1-p_1)/2$
{1; 2, 1}	$p_1 + p_1/2 + (1-p_1)p_1$
{1; 2, 2}	$p_1 + 1/2$
{2; 1, 1}	$1/2 + p_1$
{2; 1, 2}	$1/2 + (1/2)p_1 + (1/2)(1/3)$
{2; 2, 1}	$\mathbf{1/2 + (1/2)(2/3) + (1/2)p_1}$
{2; 2, 2}	$\mathbf{2(1/2) = 1}$

Table 3 The optimal expected proportion of successes for selected values of n and for fixed $p_1 = 1/2$

n	1	2	5	10	20	50	100	200	500	1000	10 000
Proportion of successes	0.500	0.542	0.570	0.582	0.596	0.607	0.613	0.617	0.621	0.622	0.624

Table 4 The breakeven values of p_1^* for selected values of n

n	1	2	5	10	20	50	100	200	500	1000	10 000
p_1^*	0.500	0.556	0.636	0.698	0.758	0.826	0.869	0.902	0.935	0.954	0.985

a success, and a_F is the next selection should the first observation be a failure. To find the expected number of successes for a particular design, one needs to know such quantities as the probability of a success on arm 2 after a success on arm 2 (which is 2/3) and the probability of a success on arm 2 after a failure on arm 2 (which is 1/3). The possible designs and their associated expected numbers of successes are given in Table 2.

It is easy to check that only three of these expected numbers of successes (shown in bold) are candidates for the maximum. If $p_1 \geq 5/9$ then $\{1; 1, 1\}$ is optimal; if $1/3 \leq p_1 \leq 5/9$ then $\{2; 2, 1\}$ is optimal; and if $p_1 \leq 1/3$ then $\{2; 2, 2\}$ is optimal. For example, if $p_1 = 1/2$ then it is optimal to use the unknown arm 2 initially. If the outcome is a success, then a decision is made to 'stay with a winner' and use arm 2 again. If a failure occurs, then the decision is made to switch to the known arm 1.

Enumeration of designs is tedious for large n. Most designs can be dropped from consideration based on theoretical results (Berry and Fristedt, 1985). For example, there is a breakeven value of p_1, say p_1^*, such that arm 1 is optimal for $p_1 \geq p_1^*$. Also, one need consider only those designs that continue to use arm 1 once it has been selected. But many designs remain. Backward induction can be used to find an optimal design (Berry and Fristedt, 1985).

Table 3 gives the optimal expected proportion of successes for selected values of n and for fixed $p_1 = 1/2$. Asymptotically, for large n, the maximal expected proportion of successes is 5/8, which is the expected value of the maximum of p_1 and p_2. Both arms offer the same chance of success on the current unit, but only arm 2 gives information that can help in choosing between the arms for treating later units.

Table 4 gives the break-even values of p_1^* for selected values of n. This table shows that information is more important for larger n. For example, if $p_1 = 0.75$ then arm 1 would be optimal for $n = 10$, but it would be advisable to test arm 2 when $n = 100$; this is so even though arm 1 has probability of 0.75 of being better than arm 2.

When there are several arms with unknown characteristics, the problem is still more complicated. Optimal designs may well indicate selection of an arm that was used previously and set aside in favor of another arm because of inadequate performance. For the methods and theory for solving such problems, see Berry (1972), and Berry and Fristedt (1985). The optimal designs are generally difficult to describe. Berry (1978) provides easy to use sequential designs that are not optimal but that perform reasonably well.

Suppose the n units in the experiment are a subset of the N units on which arms 1 and 2 can be applied. Berry and Eick (1995) consider the case of two arms with dichotomous response and show how to incorporate all N units into the design problem. They find the optimal Bayes design when p_1 and p_2 have independent uniform prior distributions. They compare this with various other sequential designs and with a particular nonsequential design: balanced randomization to arms 1 and 2. The Bayes design performs best on average, of course, but it is robust in the sense that it outperforms the other designs for essentially all pairs of p_1 and p_2.

Further Reading

The pioneers in sequential statistical methods were Wald (1947) and Barnard (1944). They put forth the sequential probability ratio test (SPRT), which is of fundamental importance in sequential stopping problems. The study of the SPRT dominated the theory and methodology of sequential experimentation for decades.

For further reading about Bayesian versus frequentist issues in sequential design, see Berger (1986), Berger and Berry (1988), Berger and Wolpert (1984), and Berry (1987, 1993). For further reading about the frequentist perspective, see Chow et al. (1971) and Whitehead (1992). For further reading about Bayesian design issues, see Berry (1993), Berry and Stangl (1996), Chernoff and Ray (1965), Cornfield (1966), and Lindley and Barnett (1965). For further reading about bandit problems, see Berry (1972, 1978), Berry and Eick (1995), Berry and Fristedt (1985), Bradt et al. (1956), Friedman et al. (1964), Murphy (1965), Rapoport (1967), Rothschild (1974), Viscusi (1979), and Whittle (1982/3). There is a journal called *Sequential Analysis* that is dedicated to the subject of this article.

See also: Experimental Design: Overview; Experimental Design: Randomization and Social Experiments; Psychological Treatments: Randomized Controlled Trials; Quasi-Experimental Designs.

Bibliography

Barnard, G.A., 1944. Statistical Methods and Quality Control. Report No. QC/R/7. British Ministry of Supply, London.

Berger, J.O., 1986. Statistical Decision Theory and Bayesian Analysis, second ed. Springer, New York.
Berger, J.O., Berry, D.A., 1988. Statistical analysis and the illusion of objectivity. American Scientist 76, 159–165.
Berger, J.O., Wolpert, R.L., 1984. The Likelihood Principle. Institute of Mathematical Statistics, Hayward, CA.
Berry, D.A., 1972. A Bernoulli two-armed bandit. Annals of Mathematical Statistics 43, 871–897.
Berry, D.A., 1978. Modified two-armed bandit strategies for certain clinical trials. Journal of the American Statistical Association 73, 339–345.
Berry, D.A., 1987. Interim analysis in clinical trials: the role of the likelihood principle. American Statistician 41, 117–122.
Berry, D.A., 1993. A case for Bayesianism in clinical trials (with discussion). Statistics in Medicine 12, 1377–1404.
Berry, D.A., 1996. Statistics: A Bayesian Perspective. Duxbury Press, Belmont, CA.
Berry, D.A., Eick, S.G., 1995. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. Statistics in Medicine 14, 231–246.
Berry, D.A., Fristedt, B., 1985. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, London.
Berry, D.A., Stangl, D.K., 1996. Bayesian methods in health-related research. In: Berry, D.A., Stangl, D.K. (Eds.), Bayesian Biostatistics. Marcel Dekker, New York, pp. 1–66.
Bradt, R.N., Johnson, S.M., Karlin, S., 1956. On sequential designs for maximizing the sum of *n* observations. Annals of Mathematical Statistics 27, 1060–1070.
Chernoff, H., Ray, S.N., 1965. A Bayes sequential sampling inspection plan. Annals of Mathematical Statistics 36, 1387–1407.
Chow, Y.S., Robbins, H., Siegmund, D., 1971. Great Expectations. Houghton Mifflin, Boston.
Cornfield, J., 1966. Sequential trials, sequential analysis and the likelihood principle. American Statistician 20, 18–23.
Friedman, M.P., Padilla, G., Gelfand, H., 1964. The learning of choices between bets. Journal of Mathematical Psychology 1, 375–385.
Lindley, D.V., Barnett, B.N., 1965. Sequential sampling: two decision problems with linear losses for binomial and normal random variables. Biometrika 52, 507–532.
Murphy Jr., R.E., 1965. Adaptive Processes in Economic Systems. Academic Press, New York.
Rapoport, A., 1967. Dynamic programming models for multistage decision making tasks. Journal of Mathematical Psychology 4, 48–71.
Rothschild, M., 1974. A two-armed bandit theory of market pricing. Journal of Economic Theory 9, 185–202.
Viscusi, W.K., 1979. Employment Hazards: An Investigation of Market Performance. Harvard University Press, Cambridge, MA.
Wald, A., 1947. Sequential Analysis. Wiley, New York.
Whitehead, J., 1992. The Design and Analysis of Sequential Clinical Trials. Horwood, Chichester, UK.
Whittle, P., 1982/3. Optimization over Time, vols. 1 and 2. Wiley, New York.

Serial Order, Representation of

Eric Heuer, University of Hawaii at Hilo, Hilo, HI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The ability of an organism to monitor and maintain information about the nature of the world around them is a critical life function. Over the past several decades, a substantial body of evidence has been amassed that highlights the properties, functions, and neural substrates of serial order memory in both humans and nonhuman animal subjects. This article focuses on examining what is known about the neural basis of serial order memory and its dependence upon the prefrontal cortex.

Serial order working memory, the ability to monitor the temporal occurrence of stimuli and/or events, is a function essential to higher-order cognitive processing and the planning/execution of complex actions. Furthermore, this process has been consistently shown to be dependent upon the integrity of the dorsolateral prefrontal cortex (dlPFC), via human and animal studies, such that damage or maldevelopment of this area impairs performance of serial order tasks (Petrides, 2005). The following sections will highlight the evidence from human and animal studies, which collectively, characterize the neural basis of serial order memory, and highlight what is known about the roles of the dlPFC and medial temporal lobe structures in supporting serial order memory.

Prior to the discussion of serial order memory it is important to acknowledge that this term has been used to define a wide range of cognitive functions, which are far beyond the scope of this article. This article will focus on serial order memory as it pertains to working memory, specifically: the intentional mental maintenance, monitoring, or manipulation of the temporal order of stimuli that can then be utilized for the planning and execution of subsequent actions. In many contexts, serial order can also be used to describe relative familiarity, aspects of episodic memory, and the ability to repeat motor sequences that have been learned over many trials. This article aims to describe the basis of serial order working memory by highlighting neuropsychological investigations of both human and nonhuman animal subjects.

Human Serial Order Memory

Effects of Focal Brain Lesions

Historically, one of the most powerful techniques in neuropsychology has been to quantify the behavior of interest in human subjects following focal brain lesions. The utility of patients with focal brain lesions is particularly evident when considering what has been learned from Tan (Paul Broca's patient), Phineas Gage, and Patient HM. While these case studies of single individuals are few and far between, the knowledge that has been gained from the meticulous study of these individuals has proven invaluable in its ability to inform and direct research in this field.

An early study by Petrides and Milner (1982) examined the effects of frontal and temporal brain lesions on the performance of a self-ordered pointing task (SOPT) in human subjects. The SOPT was based upon a serial response task originally designed and implemented for use in investigating the consequences of frontal cortical excisions in nonhuman primates (Pinto-Hamuy and Linck, 1965; Brody and Pribram, 1978). The SOPT directly measures serial order memory by requiring subjects to monitor their behavioral selections over a series of trials. Initially, subjects are shown an array of images, as few as six and as many as 12 individual images, and were asked to point to any single image. Following image selection a delay is imposed, after which the subject is presented again with the entire image set, albeit in a new spatial orientation. On subsequent trials, the subject is asked to select an image they had not selected on the previous trial(s), until all images have been selected across the series of trials. The object of the task was for the subject to select each image only once over the series of trials, thus requiring the subject to monitor the images they have selected as a new spatial array is presented on each trial to preclude the use of a spatial strategy. Data collected in the aforementioned investigation demonstrated that damage to the frontal cortex was sufficient to impair serial order memory, a finding that has been confirmed in two independent investigations of human subjects with focal frontal cortical brain lesions (Wiegersma et al., 1990; Kesner et al., 1994).

It should also be noted that there is evidence to suggest that the temporal lobe may play a critical role in the regulation of serial order memory. Several studies have characterized serial memory impairments in individuals with medial temporal lobe lesions (Petrides and Milner, 1982; Dennis et al., 1988; Mayes et al., 2001). However, given that the subjects in these studies consisted of patients with large temporal lobe resections due to intractable epilepsy in two cases (Petrides and Milner, 1982; Dennis et al., 1988) and a single amnesic subject (Mayes et al., 2001), caution should be used in drawing strong inferences from these data. Cumulatively, these studies support a role for the medial temporal lobe in serial order memory and simultaneously substantiate the need for further investigation into the precise role of the individual medial temporal lobe structures in serial order memory. In one such case, a study quantified memory in three clinical populations: patients with selective lesions of the frontal cortex, Korsakoff's Syndrome amnesiacs, and non-Korsakoff's amnesiacs (Shimamura et al., 1990). Congruent with previous studies,

that investigation found impaired memory for word sequencing but intact word recognition and recall in the frontal cortical lesion group. In contrast, both of the amnesiac patient groups were impaired on word sequencing, recall, and recognition. Cumulatively, these findings further support the argument that the frontal cortex is a critical mediator of serial order memory, while acknowledging that the temporal cortex is a component of the neural network supporting serial order memory. To further elucidate the role of the prefrontal cortex (PFC) in serial order memory, neuroimaging studies that employed serial order memory tasks in human subjects will be detailed.

Neuroimaging

The advent of neuroimaging techniques, including functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), has enabled neuropsychologists to glean a greater understanding of the neural substrates of behavior. In the case of serial order memory, the implementation of neuroimaging approaches enabled an independent investigation of the role the dlPFC plays in supporting serial order memory. While subsequent neuroimaging studies may have been driven by a desire to substantiate the activity of the dlPFC in serial order memory, it also enabled the identification of additional contributors to the neural network supporting serial order. One of the earliest neuroimaging studies of serial order memory utilized $H_2^{15}O$ PET to measure regional cerebral blood flow while subjects performed one of two behavioral tasks: Task One required the subject to remember sequences of numbers that they had generated (self-ordered), while Task Two required the subject to remember sequences of numbers that were given by the experimenter (externally ordered) (Petrides et al., 1993). The goals of this PET study of serial order memory were two-fold: to investigate the role of the frontal cortex in serial order memory and to compare the activation of the frontal cortex, as evidenced by changes in regional cerebral blood flow, between self- and externally ordered numerical sequences. In this case, the data demonstrated that the frontal cortex, specifically the dlPFC, was significantly activated in both serial order memory conditions. Furthermore, the PET analysis also revealed no difference in the magnitude or location of dlPFC activation between the two conditions, which suggests that the dlPFC is activated when sequential information is being monitored, irrespective of the origin of the sequence. Expanding upon this finding, Stern et al. (2000) utilized abstract black and white images in their study of serial order memory to preclude subjects from employing a verbal strategy to monitor the sequence of information presented. The elimination of the verbal component to the serial order task provided for a 'purer' measure of serial order memory and which should more clearly identify brain structures selectively activated by sequence monitoring. Analysis of the fMRI data from this experiment identified, once again, a selective activation of the dlPFC related to the monitoring of the serial position of stimuli. Together, these studies support the assertion that the activation observed in the dlPFC in serial order memory is a function of the monitoring requirements of each of the three tasks, as opposed to the nature of the information to be remembered.

It is important to note that while the dlPFC has been consistently identified as a critical mediator of serial order memory, this structure does not function independently and rather is a member of an interdependent cellular network supporting this memory. Specifically, the aforementioned neuroimaging studies have detected consistent activation associated with serial order memory in the following structures: ventrolateral PFC, frontopolar cortex, premotor cortex (in verbal serial order tasks), posterior parietal cortex, and medial temporal lobe. The role that each of these structures plays in supporting serial order memory has, however, not yet been substantiated (these structures will be revisited in: 2. *Serial Order Memory: Nonhuman Animal Studies*). To further explore the neurobiological basis of serial order, an examination of the developmental literature will be highlighted in the following, specifically examining the functional emergence of serial order proficiency and concomitant maturation of the frontal cortex.

Effects of Development

In parallel to examinations of cognitive function in healthy adults, investigators have frequently examined the time course of the developmental maturation of cognitive functions from birth through early adulthood. Studies that track the sequential progression of neuropsychological development are particularly informative due to the dynamic nature of the maturing nervous system. Theoretically, cognitive functions should emerge relatively consistently across the developmental window of a species and parallel the time course of the maturation of the structure(s) that support them. Fortunately, due to the exhaustive efforts of researchers examining the developing nervous system at multiple levels and developmental time points, the delayed maturation of the PFC and the simultaneous emergence of associated cognitive functions has been extensively characterized (Goldman-Rakic, 1987; Teffer and Semendeferi, 2012). Cumulatively, the literature supports the notion that the brain, in particular the frontal cortex, continues to develop through young adulthood as a function both of experience-dependent plasticity and strong genetic influence (Giedd et al., 2010).

Studies of serial order across development have further substantiated the protracted development of the frontal cortex and its associated cognitive functions. A cross-sectional study by Crone et al. (2006) sought to examine the functional development of serial order memory in three groups of children/adolescents: 8–12, 13–17, and 18–25 years of age. In this study, the investigators compared the performance and brain activation elicited while performing two serial order memory tasks of subjects. The two tasks differed in that the first required subjects to monitor the serial presentation of representational drawings in the order that they were presented (i.e., if a truck was presented first in the series of three objects and they were asked what position the truck occupied they would answer: one). In the second task, subjects were required to mentally reverse the order of the representational drawings (i.e., if a truck was presented first in the series of three objects and they were asked what position the truck occupied they would answer: three). The mental reversal of visual stimuli requires subjects to manipulate the order of the drawings that were presented to them and as such should require greater neural resources to

complete successfully. An examination of the results of this cross-sectional developmental study identified age-related increases in performance on the serial order task and corresponding increases in activation of the dlPFC during task performance. Additionally, the data also showed that while both serial order tasks produced robust activations of the dlPFC, the task that required the subjects to mentally manipulate serial position information produced a stronger activation of the dlPFC. However, the youngest group, aged 8–12 years, failed to differentially activate the dlPFC in the manipulation versus maintenance conditions, which corresponded to a reduction in the accuracy on the manipulation trials. These data further support the notion that the postnatal development of the dlPFC is what enables age-related increases in proficiency in these types of tasks. To further explore the relationship between dlPFC activation and memory accuracy, the task from the aforementioned experiment was modified to examine the effects of increasing memory load across two age groups: 11–13 and 18–25 (Jolles et al., 2011). In this case, the original task employed by Crone et al. (2006) was modified to examine the effect of memory load by varying the number of stimuli to be remembered in the sequence from a fixed number of three stimuli, to series of three, four, and five stimuli. The primary finding of this investigation confirmed the earlier report of age-related increases in serial order memory performance, which was correlated with the magnitude of dlPFC activation. In contrast, the investigators found increased activity in the ventrolateral PFC of the younger children, perhaps as a compensatory mechanism for inadequate dlPFC functionality. However, while the accuracy of subjects decreased with increasing memory load, no significant differences in the activation pattern of the dlPFC were observed. Taken together, these data support the notion that competency in serial order memory matures over a long developmental timescale and parallels the functional capacity of the dlPFC.

As is the case in many fields, cognitive psychologists frequently examine the pathology associated with disease states to try to ascertain what is 'normal.' Thus, an examination of serial order memory in psychiatric and neurological diseases, which have stereotypical neuropathology, including the frontal and temporal cortices, may provide valuable insights into the basis of serial order.

Effects of Mental Illness

Consistently, severe disorders, especially neurodevelopmental and neurodegenerative have been documented to show comorbid cognitive impairments related to the neuropathology of the corresponding brain structures. For example, impairment in working memory is now acknowledged as a consistent biomarker in schizophrenia, likely due to the gross abnormalities and associated dysfunction of the dlPFC of affected individuals (Barch and Ceaser, 2012). Not surprisingly, studies have found impaired serial order memory in individuals with schizophrenia (Fraser et al., 2004) and their impairment on serial order discriminations parallels their impairment on other prefrontal-dependent measures, including the Wisconsin Card Sorting Task (Schwartz et al., 1991). The findings of impaired serial order memory and corresponding frontal cortical neuropathology provide additional support for the dlPFC as a critical mediator of serial order memory, although it is possible that the deficits in working memory performance may also be related to altered dopamine neurotransmission to the dlPFC (Lodge and Grace, 2006).

The assertion that a reduction in dopamine transmission to the PFC could be responsible for the observed impairment in schizophrenia is substantiated by findings showing impaired serial order memory in Parkinson's disease (Shipley et al., 2002). At the same time, it must also be acknowledged that the mechanisms for these two types of dysfunction could be entirely unrelated. Furthermore, the impairment of serial order memory in Parkinson's disease does not directly correlate with other disease-related measures, such as tremor, and likely precedes the major motor pathology (Shipley et al., 2002). Finally, it is quite possible that the etiology of the impairment may be dramatically different in neurodevelopmental and neurodegenerative diseases such as schizophrenia and Parkinson's disease, respectively.

In a third example, while Alzheimer's disease (AD) is neuropsychologically typified by impaired episodic recognition, and semantic memory, investigations into AD have highlighted additional cognitive impairments associated with the disorder. In particular, it seems that impaired serial order memory is a consistent finding across the AD spectrum, from mildly impaired to late stage AD (Johnson and Kesner, 1997). Furthermore, a recent study sought to explore the nature of the serial order deficit in AD and other types of dementia, specifically vascular dementia (Hampstead et al., 2010). The results of the comparative dementia study found that while both AD and vascular dementia patients were impaired on serial order judgments, the temporal order impairment in the vascular dementia group was significantly more pronounced. As the primary initial observed neuropathology in vascular dementia typically is localized within the frontal cortex, as compared to the prototypical temporal lobe pathology of AD, it is not surprising that the vascular dementia patient group has been found to be more impaired on serial order tasks.

Collectively, the data from psychiatric and neurological disorders with respect to serial order memory provide invaluable confirmation of the sensitivity of serial order memory with further evidence of its functional dependence upon the dlPFC. First, the severity of the temporal order memory impairment is far greater in schizophrenia and vascular dementia than in Parkinson's disease or AD and is likely due to the extent of neuropathology within the frontal cortex of both schizophrenia and vascular dementia patients. Secondly, it is likely that the impairment in serial order memory in Parkinson's disease and AD is due to damage to the ventral tegmental dopaminergic system and temporal lobe, respectively, which may account for the more moderate impairment characterized in these two disorders. Thirdly, it seems that while the monitoring of serial information is a functional property of the dlPFC, it is clear that other regions, including the temporal lobe, mesolimbic dopamine system, and the posterior parietal cortex, are critical members of the neural circuit supporting serial order memory. For further insight into the neuropsychological substrates of serial order memory, an examination of the experimental animal literature is necessary.

Serial Order Memory in Nonhuman Animals

As has been the case with most neuropsychological areas of study, investigators frequently turn to a comparative animal model system to supplement the direct study of neuropsychology in human subjects, for a variety of reasons. While it is true that studies of patients with focal brain lesions can prove invaluable in enhancing our understanding of brain structure–function relationships, as evidenced by Patient HM, the relative paucity of such individuals necessitates alternative approaches. As such, the fundamental advantage to employing nonhuman models of neuropsychological function lies in the ability of the investigator to examine the direct consequence of selective brain lesions and its impact on cognitive function(s).

Nonhuman Primates

The use of nonhuman primate models of higher cognitive function has proved an invaluable resource for investigators who are interested in the neural basis of behavior. Nonhuman primates are an advantageous model system due to the degree of homology in neural architecture, protracted cognitive/physical development, and the ability to assay cognitive functions in a manner similar, or in some cases identical, to the way it is assessed in human subjects (Passingham, 2009). Studies of serial order memory in nonhuman primates date back as far as the early twentieth century to those on the effects of focal brain lesions in the chimpanzee (*Pan troglodytes*) (Jacobsen et al., 1935). The initial study by Jacobsen sought to examine the consequence of frontal lobe ablation on the performance of a sequence of responses to retrieve a food reward. In two separate paradigms it was demonstrated that the bilateral, but not unilateral, frontal lesions were sufficient to impair the performance of a sequence task that involved the monitoring of task progression to determine the direction or location of subsequent responses. In other words, bilateral frontal lesions impaired performance when the subject was required to track the serial progress of their own responses to determine the appropriate action that will lead to reward. A subsequent study, conducted in the same laboratory, examined the consequence of both lateral prefrontal cortical and inferotemporal lesions in the rhesus macaque (*Macaca mulatta*) on a serial response task. The results of this experiment demonstrated that frontal but not inferotemporal lesions were sufficient to impair an internally guided sequential task, confirming the earlier finding of Jacobsen in the chimpanzee (Pinto-Hamuy and Linck, 1965).

Years later, a resurgence in the focus of studies of serial order in humans sparked renewed interest in examining the basis of serial order memory in nonhuman primates. A preliminary study by Gower (1992) demonstrated that serial order memory in the macaque monkey is comparable to humans, both in terms of primacy/recency effects and the delay/temporal distance-based decrement in performance. Furthermore, a series of studies sought to definitively identify the locus of serial order memory tested rhesus macaques with focal lesions of the dlPFC on both self- and externally ordered serial order memory tasks (Petrides, 1991a,b, 1995). Collectively, this series of studies demonstrated that the dlPFC is a critical mediator of serial order memory, such that lesions of the dlPFC, but not other frontal cortical areas, selectively impair performance on self- and externally ordered serial tasks. The data collected in these experiments, which examined the performance of macaque monkeys on multiple independent serial order tasks, are particularly compelling in that they clearly demonstrated that the monitoring of information was the critical factor supported by the dlPFC, regardless of the nature of the information to be monitored. A pair of single-unit electrophysiological studies in the macaque monkey provided further support, for the role of the dlPFC, in supporting serial order memory. Cumulatively, the electrophysiological studies of self-ordered serial memory tasks reported selective task- and load-dependent increases in neuronal firing within the dlPFC, further substantiating the findings of the earlier lesion studies (Ninokura et al., 2003; Hasegawa et al., 2004).

The initial and still only study by Kimble and Pribram (1963) demonstrated that large hippocampal lesions were sufficient to impair performance on a serial response task, a finding that was later replicated in human subjects (Petrides and Milner, 1982). However, due to the large size of the lesions it is not possible to say whether the observed impairment in serial order working memory was due to damage of the hippocampus, adjacent cortical areas in the medial temporal lobe, or a combination of the two. To date, no study has measured the effects of selective hippocampal lesions on the performance of serial order memory tasks in adult nonhuman primates. Two recent studies have shown that neonatal lesions of the hippocampus are sufficient to impair the performance of both self- and externally guided serial order memory tasks (Heuer and Bachevalier, 2011, 2013). However, because the hippocampal lesions were performed in neonates the interpretation of these findings is inherently complicated. The observed impairment in serial order memory could have resulted from lack of a functional hippocampus in adulthood, a disruption of the normal development of the dlPFC due to hippocampal damage, or a combination of these two possibilities. The complexity of interpreting the results of neonatal hippocampal lesions is further substantiated by the neonatal hippocampal lesion model of schizophrenia, in rats, which demonstrates that early hippocampus damage results in impaired medial prefrontal cortical function in adulthood (Lipska, 2004). Given the relative paucity of data in primates as to the impact of hippocampal lesions on serial order memory, the focus of this article will now shift to rodent models of cognition.

Rodents

In parallel to nonhuman primate studies of serial order memory, extensive research has also been conducted in rodent species, specifically rats (*Rattus norvegicus*). Before discussing the results of the rodent data, there are some differences in the cognitive assessment, memory capacity, and neuroanatomy of the rat as it pertains to the administration and interpretation of serial order tasks that should be addressed. In particular, it is important to note the differences in studies of serial order memory in rats as compared to those of human and nonhuman primates. First, as rodents have well-developed spatial and olfactory abilities, the serial order tasks in

rodents typically employ a radial-arm maze, olfactory stimuli, or combinations of the two. Secondly, while both humans and nonhuman primate subjects can monitor and discriminate between adjacent items in lists of five stimuli, the literature suggests that the rats' ability to discriminate adjacent stimuli in a sequential list is limited, at best. The inherent limitation of the rat to discriminate between adjacent stimuli may suggest that it is utilizing a strategy other than the monitoring of stimuli (e.g., relative familiarity or recency). Thirdly, while the rat has proven to be a useful animal model for investigations of cognition in a multitude of domains, caution must be taken when interpreting studies of PFC function in this subject as the homology between rodent and primate PFC is still a subject of debate (Uylings et al., 2003; Seamans et al., 2008). While it is important to view the following data with these considerations, it is also important to acknowledge the contribution of these data to our current understanding of serial order memory.

In an effort to substantiate findings of impaired serial order memory in human subjects following frontal lobe damage, Kesner and Holbrook (1987) designed a series of serial order assessments for rats utilizing spatial position in an eight-arm radial maze. The development of this adaptation, whereby spatial position was utilized as an identifier enabled the first studies of serial order memory in the rat. Mirroring findings in human subjects with lesions of the dlPFC, this study demonstrated that lesions of the medial PFC, in the rat, were sufficient to impair serial order memory, which was later confirmed in a study by Chiba et al. (1997). Specifically, the authors examined the performance of rats on both novel and repeated order spatial sequences. Interestingly, while medial prefrontal cortical lesioned subjects were impaired on both sequences, the magnitude of impairment was diminished in the repeated sequences. One possible explanation for the reduced impairment of repeated sequences is that repetition of the sequence allowed the information to be transferred into an alternative memory store, which may not be dependent upon the functional integrity of the medial PFC.

Given the absence of data detailing the impact of selective hippocampal lesions on serial order performance in primates, any information that the rodent data could provide on this topic is even more valuable. While it might have been expected that hippocampal lesions would impair serial order memory performance to some degree, the results of this study indicated that the serial order impairment following hippocampal lesions was more severe than that observed following the medial prefrontal cortical lesions (Chiba et al., 1994). While one could argue that the observed impairment on the serial order memory task was due to the spatial component of the task, a variation of the serial order memory task that used odor instead of spatial position mitigates this possibility (Kesner et al., 2002).

Conclusion and Future Directions

The fundamental goal of this article was to try to connect what is known about the basis of serial order memory, specifically the neural substrates that support and regulate its function across studies of humans, monkeys, and rodents. At its core, the capacity of an organism to monitor sequential events, be it their own actions or changes in their environment, is a critical function of species survival and proliferation. While not overemphasizing the importance of serial order memory, it is hard to imagine functioning successfully in the world without this capacity.

In examining the literature from the early studies of Jacobsen (1935) to the more recent developments of automated assessments of multiple-item serial order memory (Templer and Hampton, 2013), it is clear that there is a lot that remains to be explored in terms of the basis of serial order memory. At present, while it is reasonable to assert that the dlPFC is the functional locus of serial order memory, the contributions of other brain structures in serial order memory remain largely unknown.

See also: Amnesia: General; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Executive Functions During Childhood, Development of; Functional Magnetic Resonance Imagery, Analysis of; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Retrieval; Temporal Distinctiveness Models of Memory; Temporal Lobe; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Barch, D.M., Ceaser, A., 2012. Cognition in schizophrenia: core psychological and neural mechanisms. Trends in Cognitive Sciences 16, 27–34.

Brody, B.A., Pribram, K.H., 1978. The role of frontal and parietal cortex in cognitive processing: tests of spatial and sequence functions. Brain 101, 607–633.

Chiba, A.A., Kesner, R.P., Gibson, C.J., 1997. Memory for temporal order of new and familiar spatial location sequences: role of the medial prefrontal cortex. Learning & Memory 4, 311–317.

Chiba, A.A., Kesner, R.P., Reynolds, A.M., 1994. Memory for spatial location as a function of temporal lag in rats: role of hippocampus and medial prefrontal cortex. Behavioral and Neural Biology 61, 123–131.

Crone, E.A., Wendelken, C., Donohue, S., Van Leijenhorst, L., Bunge, S.A., 2006. Neurocognitive development of the ability to manipulate information in working memory. Proceedings of the National Academy of Sciences of the United States of America 103, 9315–9320.

Dennis, M., Farrell, K., Hoffman, H.J., Hendrick, E.B., Becker, L.E., Murphy, E.G., 1988. Recognition memory of item, associative and serial-order information after temporal lobectomy for seizure disorder. Neuropsychologia 26, 53–65.

Fraser, D., Park, S., Clark, G., Yohanna, D., Houk, J.C., 2004. Spatial serial order processing in schizophrenia. Schizophrenia Research 70, 203–213.

Giedd, J.N., Stockman, M., Weddle, C., Liverpool, M., Alexander-Bloch, A., Wallace, G.L., Lee, N.R., Lalonde, F., Lenroot, R.K., 2010. Anatomic magnetic resonance imaging of the developing child and adolescent brain and effects of genetic variation. Neuropsychology Review 20, 349–361.

Goldman-Rakic, P.S., 1987. Development of cortical circuitry and cognitive function. Child Development 58, 601–622.

Gower, E.C., 1992. Short-term memory for the temporal order of events in monkeys. Behavioural Brain Research 52, 99–103.

Hampstead, B.M., Libon, D.J., Moelter, S.T., Swirsky-Sacchetti, T., Scheffer, L., Platek, S.M., Chute, D., 2010. Temporal order memory differences in Alzheimer's disease and vascular dementia. Journal of Clinical and Experimental Neuropsychology 32, 645–654.

Hasegawa, R.P., Blitz, A.M., Goldberg, M.E., 2004. Neurons in monkey prefrontal cortex whose activity tracks the progress of a three-step self-ordered task. Journal of Neurophysiology 92, 1524–1535.

Heuer, E., Bachevalier, J., 2011. Neonatal hippocampal lesions in rhesus macaques alter the monitoring, but not maintenance, of information in working memory. Behavioral Neuroscience 125, 859–870.

Heuer, E., Bachevalier, J., 2013. Working memory for temporal order is impaired after selective neonatal hippocampal lesions in adult rhesus macaques. Behavioural Brain Research 239, 55–62.

Jacobsen, C.F., 1935. Functions of the frontal association area in primates. Archives of Neurology and Psychiatry 45, 565–575.

Jacobsen, C.F., Wolfe, J.B., Jackson, T.A., 1935. An experimental analysis of the functions of the frontal association areas in primates. Journal of Nervous Mental Disease 82, 1014.

Johnson, D.L., Kesner, R.P., 1997. Comparison of temporal order memory in early and middle stage Alzheimer's disease. Journal of Clinical and Experimental Neuropsychology 19, 83–100.

Jolles, D.D., Kleibeuker, S.W., Rombouts, S.A., Crone, E.A., 2011. Developmental differences in prefrontal activation during working memory maintenance and manipulation for different memory loads. Developmental Science 14, 713–724.

Kesner, R.P., Gilbert, P.E., Barua, L.A., 2002. The role of the hippocampus in memory for the temporal order of a sequence of odors. Behavioral Neuroscience 116, 286–290.

Kesner, R.P., Holbrook, T., 1987. Dissociation of item and order spatial memory in rats following medial prefrontal cortex lesions. Neuropsychologia 25, 653–664.

Kesner, R.P., Hopkins, R.O., Fineman, B., 1994. Item and order dissociation in humans with prefrontal cortex damage. Neuropsychologia 32, 881–891.

Kimble, D.P., Pribram, K.H., 1963. Hippocampectomy and behavior sequences. Science 139, 824–825.

Lipska, B.K., 2004. Using animal models to test a neurodevelopmental hypothesis of schizophrenia. Journal of Psychiatry & Neuroscience 29, 282–286.

Lodge, D.J., Grace, A.A., 2006. The hippocampus modulates dopamine neuron responsivity by regulating the intensity of phasic neuron activation. Neuropsychopharmacology 31, 1356–1361.

Mayes, A.R., Isaac, C.L., Holdstock, J.S., Hunkin, N.M., Montaldi, D., Downes, J.J., Macdonald, C., Cezayirli, E., Roberts, J.N., 2001. Memory for single items, word pairs, and temporal order of different kinds in a patient with selective hippocampal lesions. Cognitive Neuropsychology 18, 97–123.

Ninokura, Y., Mushiake, H., Tanji, J., 2003. Representation of the temporal order of visual objects in the primate lateral prefrontal cortex. Journal of Neurophysiology 89, 2868–2873.

Passingham, R., 2009. How good is the macaque monkey model of the human brain? Current Opinion in Neurobiology 19, 6–11.

Petrides, M., 1991a. Functional specialization within the dorsolateral frontal cortex for serial order memory. Proceedings Biological Sciences/The Royal Society 246, 299–306.

Petrides, M., 1991b. Monitoring of selections of visual stimuli and the primate frontal cortex. Proceedings Biological Sciences/The Royal Society 246, 293–298.

Petrides, M., 1995. Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. Journal of Neuroscience 15, 359–375.

Petrides, M., 2005. Lateral prefrontal cortex: architectonic and functional organization. Philosophical Transactions of the Royal Society of London – Series B: Biological Sciences 360, 781–795.

Petrides, M., Alivisatos, B., Meyer, E., Evans, A.C., 1993. Functional activation of the human frontal cortex during the performance of verbal working memory tasks. Proceedings of the National Academy of Sciences of the United States of America 90, 878–882.

Petrides, M., Milner, B., 1982. Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. Neuropsychologia 20, 249–262.

Pinto-Hamuy, T., Linck, P., 1965. Effect of frontal lesions on performance of sequential tasks by monkeys. Experimental Neurology 12, 96–107.

Schwartz, B.L., Deutsch, L.H., Cohen, C., Warden, D., Deutsch, S.I., 1991. Memory for temporal order in schizophrenia. Biological Psychiatry 29, 329–339.

Seamans, J.K., Lapish, C.C., Durstewitz, D., 2008. Comparing the prefrontal cortex of rats and primates: insights from electrophysiology. Neurotoxicity Research 14, 249–262.

Shimamura, A.P., Janowsky, J.S., Squire, L.R., 1990. Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. Neuropsychologia 28, 803–813.

Shipley, B.A., Deary, I.J., Tan, J., Christie, G., Starr, J.M., 2002. Efficiency of temporal order discrimination as an indicator of bradyphrenia in Parkinson's disease: the inspection time loop task. Neuropsychologia 40, 1488–1493.

Stern, C.E., Owen, A.M., Tracey, I., Look, R.B., Rosen, B.R., Petrides, M., 2000. Activity in ventrolateral and mid-dorsolateral prefrontal cortex during nonspatial visual working memory processing: evidence from functional magnetic resonance imaging. Neuroimage 11, 392–399.

Teffer, K., Semendeferi, K., 2012. Human prefrontal cortex: evolution, development, and pathology. Progress in Brain Research 195, 191–218.

Templer, V.L., Hampton, R.R., 2013. Cognitive mechanisms of memory for order in rhesus monkeys (*Macaca mulatta*). Hippocampus 23, 193–201.

Uylings, H.B., Groenewegen, H.J., Kolb, B., 2003. Do rats have a prefrontal cortex? Behavioural Brain Research 146, 3–17.

Wiegersma, S., Van Der Scheer, E., Human, R., 1990. Subjective ordering, short-term memory, and the frontal lobes. Neuropsychologia 28, 95–98.

Serial Verb Constructions

Mark Sebba, Lancaster University, Lancaster, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 20, pp. 13927–13930, © 2001, Elsevier Ltd.

Abstract

The term 'serial verb construction' is applied to a range of apparently similar syntactic constructions in different languages, which have in common the fact that they contain two or more verbs within a single phrase or clause, without overt signs of coordination. This article discusses problems of identification and description of serial verbs, their distribution in the world's languages, some competing syntactic analyses, the range of semantic functions which serial verbs perform, their importance in the field of Creole studies, and the historical development of 'serial verbs' into members of other syntactic categories such as prepositions and complementizers.

Definition and Importance

The term 'serial verb construction' (SVC) is usually applied to a range of apparently similar syntactic constructions in different languages, in which several verbs occur together within one clause or unit, without evidence of either subordination or coordination of the verbs. For example, sentences like (1), from the West African language Twi, are typical of what most scholars would regard as an SVC:

(1) Kofi de pono no baae
Kofi take-PAST table the come-PAST
'Kofi brought the table' or more literally,
'Kofi took the table and came with it'

SVCs have attracted attention from linguists for three main reasons. First, syntacticians have found them interesting because the possibility of having more than one main (i.e., nondependent, nonauxiliary) verb within a clause or clause-like unit challenges traditional assumptions that a clause contains exactly one predicate (see Foley and Olson, 1985: pp. 17, 57). Second, because of their prevalence both in languages of West Africa and in certain Creole languages, they have become central to debates among Creole language researchers, being seen either as evidence for the importance of substrate (i.e., African mother-tongue) influence on the Creole, or as deriving from universal properties of human language which are manifested in the creolization process (*see* Pidgin and Creole Language Varieties). Third, serial verbs show a tendency to be reanalyzed as other grammatical categories (complementizers and prepositions), with implications for the processes of historical change.

While different researchers use different criteria for defining serial verbs, the following set would probably be accepted by most and is based on one described by McWhorter (1997: p. 22) as a "relatively uncontroversial distillation of the conclusions of various scholars":

(2) (a) an SVC contains only one overt subject;
(b) it contains no overt markers of coordination or subordination;
(c) it expresses concomitant actions (either simultaneous or consecutive) or a single action;
(d) it falls under one intonational contour;
(e) it has tense–modality–aspect marking on none of the verbs, or on one only (usually the first or last in the series) or on all of them; and
(f) it contains a second verb, which is not obligatorily subcategorized for by the first verb.

Even the above six conditions leave it unclear where exactly to draw the line around SVCs, particularly in languages which have little morphology. In some cases of what may appear to be SVCs, we may be dealing simply with unmarked coordination of verb phrases expressing simultaneous actions.

Serial Verb Types

Serial constructions can be classified into different types on the basis of the functions of the verbs in the series relative to one another.

The sentences in (3) provide examples of different functional types of SVC which have been discussed in the literature.

(3) (a) *Directional complements*: V_2 is *go*, *come*, or another common intransitive motion verb which functions to indicate the directionality of the action denoted by V_1.
Olú gbé àga wá
Olu take chair come
'Olu brought a chair' (Yoruba, West African)

(b) *Other motion verb complements*: similar to type (a), but V_2 may be transitive motion verb with its own object expressing the goal of the action denoted by V_1
Kofi tow agyan no wɔɔ Amma
Kofi throw-PAST arrow the pierce-PAST Amma
'Kofi shot Amma with the arrow'

(c) *Instrumental constructions*: V_1 is *take* or a semantically similar common verb, while V_2 denotes an action performed with the aid of the object of V_1.
Kofi teki a nefi koti a brede
Kofi take the knife cut the Bread
'Kofi cut the bread with a knife' (Sranan, Caribbean Creole)

(d) *Dative constructions*: V_2 is usually *give*, with an object which denotes the (semantic) indirect object of V_1.

Ogyaw ne sika mãã me
he-leave-PAST his money give-PAST me
'He left me his money' (Twi, West African)

(e) *Comparative constructions*: a verb meaning *pass* or *surpass* is used with an object to express comparison. In this case 'V_1' is often in fact an adjective.
Amba tranga pasa Kofi
Amba strong pass Kofi
'Amba is stronger than Kofi' (Sranan, Caribbean Creole)

(f) *Resultative constructions*: V_2 denotes the result or consequence of an action denoted by V_1.
Kofi naki Amba kiri
Kofi hit Amba kill
'Kofi struck Amba dead' (or more literally, 'Kofi hit Amba and killed her') (Sranan, Caribbean Creole)

(g) *Idiomatic constructions* (*lexical idioms*): These are cases where the meaning of the verbs together is not derivable from the meanings of the verbs separately. They are not found in all serializing languages, but some of the West African group are particularly rich in them.
Anyi-Baule (West African):
bɔ 'hit' nĩã 'look' bɔ…nĩã 'say, tell'
Yoruba (West African):
là 'cut open' yé…'understand' là…yé 'explain'

Distribution in the World's Languages

Serial verb constructions have been identified in many languages but it is clear that they occur in areal clusters and are not spread evenly among the languages of the world. So far serializing languages have been identified in the following areas: West Africa (Kwa and related language families), the Caribbean (Creole languages which have a historical relationship with Kwa languages), Central America (e.g., Misumalpan), Papua New Guinea (Papuan languages and Tok Pisin Melanesian Pidgin English), South-east Asia (e.g., Chinese, Vietnamese, Thai). There are isolated reports of SVC-like constructions from elsewhere. How credible these are depends on which criteria are adopted to define SVCs.

Grammatical Analyses

Numerous researchers have put forward proposals for grammatical analyses of SVCs. The analyses may be classified into two types, with some degree of overlap. *Semantic* analyses typically seek to account for SVCs in terms of how they break down verbal concepts into more basic semantic components (take–carry–come for *bring*, for example). *Syntactic* analyses differ mainly according to whether they treat SVCs as a phrase structure phenomenon (usually under a version of X-bar theory), as a form of complementation, subordination, or secondary predication (i.e., involving more than one clause-like unit), or as a lexical phenomenon (e.g., as single but disjoint lexical items).

The following phrase structure (4) proposed for SVCs in the literature is typical of those advocated by a number of researchers:

(4) VP → V XP VP where X is N or P

This produces a right-branching tree with a theoretically infinite number of verbs in a series. Such a structure would allow lexical relations and relations like 'subject-of,' and 'object-of,' to hold between verbs in the series, and captures the intuition that verbs (or verb phrases) function as complements to verbs earlier in the sequence.

Some researchers have treated SVCs as involving complementation or subordination, with clauses or clause-like units embedded within the VP and dependent on a higher verb. For example, Larson (1991: pp. 198–205) has argued in favor of analyzing serial constructions as a case of 'secondary predication.' A structure similar to that for *Carol rubbed her finger raw* in English can also, he says, account for serial combinations of the *take…come* and *hit…kill* types – (a) and (f) above.

Foley and Olson (1985) propose a distinction between 'nuclear' and 'core' serialization. In nuclear serialization the serial verbs all occur in the nucleus of the clause, in other words the verbs form a single unit, which must share arguments and modifiers. The core layer of the clause consists of the nuclear layer plus the core arguments of the verb (Foley and Olson, 1985: p. 34). In 'core' serialization, two cores each with own nucleus and arguments, are joined together. Languages may exhibit both kinds of serialization or just one. Examples (a) and (b) below (Foley and Olson, 1985: p. 38) illustrate this difference; (a) is an example of core, and (b) of nuclear, serialization:

(5) (a) fu fi fase isoe
he sit letter write
'He sat down and wrote a letter' (Barai, Papuan)
(b) Fu fase fi isoe
He letter sit write
'He sat writing a letter'

This evidence suggests that 'serialization' may not be a unitary phenomenon, and although a single account which explains the two or more different construction types (such as Foley and Olson's) would be welcome, it may be that separate analyses may be needed as more different types of SVC come to light.

The Creolists' Debates

SVCs are among the most salient grammatical features of many Creoles of the Caribbean area, which distinguish them grammatically from their European lexifier languages. As such they have attracted attention from creolists, who, finding similar structures in West African languages which historically are connected with the Creoles in question, have claimed SVCs are evidence of substrate influence on the grammar of the Creoles concerned.

An alternative view has been offered by Bickerton (1981) and Byrne (1987), who regard the similarities between SVCs in the Creole and West African languages as coincidental. Instead, they argue, SVCs result in Creoles from universal principles of grammar, and are a consequence of the rudimentary verbal structure (V but not VP) which exist, they say, in a Creole at the earliest stage of development. They point out the existence of SVCs or serial-like structures in other pidgins and Creoles such as Tok Pisin (New Guinea Pidgin) and

Hawaiian Creole English, which have no historical connections with West Africa.

McWhorter (1997) argues against Bickerton and Byrne, claiming that the Caribbean Creoles not only share SVCs with West African languages, but that their SVCs are structurally similar to each other in ways that SVCs from other areal language groupings are not. Using the nuclear/core distinction proposed by Foley and Olson (see above), he argues that Papuan languages serialize at the nuclear level, while Kwa languages, Caribbean Creoles, Chinese, and other Southeast Asian languages serialize at the core level. Considering also the range of pidgins and Creoles, which lack SVCs altogether, he concludes that "SVCs have appeared around the world in precisely the creoles which had serializing substrata" (1997: p. 39).

SVCs have thus become a major site of contention between 'substratists' and 'universalists' in Creole studies, with both sides remaining committed to their own positions (*see* Pidgin and Creole Language Varieties).

SVCs and Language Change

Particular verbs which participate in serial constructions in some cases appear to have been reanalyzed as members of another syntactic category, e.g., from verb to preposition *(give → for)* or complementizer (*say → that*). In such cases, the position of the verb usually makes it amenable to such a reanalysis (e.g., if it typically occurs immediately before its objects, it may be reanalyzed as a preposition if the semantics encourage this interpretation). Reanalyzed serial verbs lose, to varying degrees, their verbal properties and take on, again to varying degrees, the morphological characteristics of their new category. Lord (1973) describes cases of reanalysis of verbs as prepositions, comitative markers, and a subordinating conjunction in Yoruba, Gã, Ewe, and Fon. Verbs with the meaning '*say*' are susceptible to reinterpretation as complementizers (introducing sentential clauses), while those with the meaning '*finish*' have a tendency to be interpreted as completive aspect markers (*see* Grammaticalization).

Related Constructions

Lack of a satisfactory definition of 'serial verb construction' makes it difficult to decide what may count as a related kind of structure. Chinese has a range of verbal structures which bear resemblances to SVCs. One of these, the class of *co-verbs*, is widely believed to be the result of the reanalysis of serial verbs as prepositions. Others, so-called *resultatives*, resemble secondary predications.

The Bantu languages, though related to the serializing languages of West Africa, do not seem to have SVCs. However, some have verbal chains where the second and subsequent verb have different marking from the first. In many Bantu languages, the verb meaning *say* is homophonous with a complementizer. These phenomena remain to be explained by a general theory.

See also: Syntax (General); Valency and Argument Structure in Syntax.

Bibliography

Bickerton, D., 1981. Roots of Language. Karoma, Ann Arbor, MI.

Byrne, F., 1987. Grammatical Relations in a Radical Creole: Verb Complementation in Saramaccan. Benjamins, Amsterdam.

Foley, W.A., Olson, M., 1985. Clausehood and verb serialization. In: Nichols, J.A., Woodbury, A.C. (Eds.), Grammar Inside and Outside the Clause: Some Approaches to Theory from the Field. Cambridge University Press, Cambridge, UK, pp. 17–60.

Larson, R.K., 1991. Some issues in verb serialization. In: Lefebvre, C. (Ed.), Serial Verbs: Grammatical Comparative and Cognitive Approaches. Benjamins, Amsterdam, pp. 184–210.

Lefebvre, C.J. (Ed.), 1991. Serial Verbs: Grammatical, Comparative and Cognitive Approaches. Benjamins, Amsterdam.

Lord, C., 1973. Serial verbs in transition. Studies in African Lingusitics 4 (3), 269–296.

McWhorter, J.H., 1997. Towards a New Model of Creole Genesis. Peter Lang, New York.

Sebba, M., 1987. The Syntax of Serial Verbs: An Investigation into Serialization in Sranan and Other Languages. Benjamins, Amsterdam.

Service Economy, Geography of

Brita Hermelin, Linköping University, Linköping, Sweden
Grete Rusten, University of Bergen, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the development and role of services in the economy, the spatial relations of services as workplace and infrastructure. Topics include how services are organized and positioned in production systems, spatially distributed and operate across space. This article illustrates and reflects on the key question for service research in geography, namely the relationships between (1) integrating the practice of face-to-face service operations with partners and (2) providing services over distance and different locations (urban and rural, domestic and foreign) through the means of information and communication technology (ICT). The article concludes with an outlook on future research directions; clearly, service research reflects major societal issues and discourses, including new financial models, innovation, and environmental projects.

Three main sections follow: Part 1 is a broad discussion of the definitions and economic role of services; Part 2 focuses on producer services emphasizing position, development and roles in production systems; and Part 3 offers a discussion of recent research areas and prospects for future research on services.

Part 1. The Wider Setting

Conceptual Approach

Prompted by the deindustrialization of Western economies, research published in the 1970s attempted to capture key elements of the rapid and substantial growth of the service sector. Researchers in this area proposed several definitions of services; for example,

> Production of "goods" results in a material product that is typically storable and transportable. Production of "services" results in an output that is not storable and usually requires direct interaction with the customer. (Stanback, 1979: p. 5)

Hill's (1977) definition of services from the same period emphasizes the processes and effects of service activities. "A service may be defined as a change in the condition of a person or a good belonging to some economic unit, which is brought about as the result of the activity of some other economic unit, with the prior agreement of the former person or economic unit" (Hill, 1977: p. 318).

These definitions share an understanding of the direct, personal interactions that occur between service providers and clients. This process of coproduction involves "embodied expertise and emotional labor" (Bryson, 2007), which implies that services occur mainly in local markets, where face-to-face interactions are possible. In general, spatial relations in the provision of services are related to services' intangible features, namely whether a service is transportable through information and communications technology (ICT) modes or must be provided through face-to-face interactions. ICTs have drastically changed the market conditions concerning accessibility and exposure. This balance between requirements of *to be close* and *to be relatively insensitive to distance* is continuously changing and varies strongly among different types of service activities. This is also an issue that has attracted much research in geography.

One method researchers have approached the problem of understanding service sector has been to define the sector in terms of subcategories. Daniels (1985), for instance, used destination of output (producer or consumer services), type of occupation (white-collar or blue-collar), and organizational form (public or private). Also mentioned should be the conventional categorizations resembling templates for statistical classifications. Examples include "distributive services (communication, trade, transport); producer services (banking, business services, real estate); social services (education, health, postal services, public and nonprofit services) and personal services (hotels, repairs, restaurants, travel and others)" (Singelmann, 1978; Browning and Singelmann, 1975). Of these four categories, producer services, in particular, have received a great deal of research attention (e.g., Bryson and Daniels, 2007). Accordingly, one main section of this article is devoted to this segment of services, with particular emphasis on advanced producer services (APSs). Service sectors such as finance, retail, and cultural sectors are included in other articles in this encyclopedia and therefore are treated here only in summative terms.

Finding a comprehensive definition of services has been a less prominent concern in the more recent literature. Instead, the service approach argues for more integrated perspectives based on the assumption that services are activities integrated into all types of industries, regardless of the industrial sector. Arguments based on theories of service-dominated logics (Vargo and Lusch, 2008) and 'servification' of economic activities characterize many studies (Toivonen and Tuominen, 2009). Integrated perspectives have also been mentioned in discussions about combining a focus on knowledge-intensive business services (KIBS) bought in the market with knowledge-intensive service activities that may take place within the firm (Aslesen and Isaksen, 2007).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72050-2

Traditional production companies that no longer make their products in-house but instead purchase all production capacity from others are a further example of how the distinction between manufacturing and services is unclear. Yet, a different definitional approach that could bring fruitful insight to service research would be to change the traditional focus seen from the lens of the service provider to the lens of the customer (Edvardsson et al., 2005).

The Role of the Service Industry in the Economy

Increasing research on the service sector signals the increasing importance of such economic activities. One direct (and readily observed) impact is the dominance of the service sector in the job market. If we examine Western industrial economies and the structure of the job market, we can see that the United States was a forerunner in restructuring to a service economy. Since the mid-1960s, service jobs have come to dominate, gradually transforming Western countries into service economies (Daniels, 1985). This shift includes both commercial and public sectors.

Figure 1 illustrates the proportion of jobs in the service sector for a selection of Organization for Economic Cooperation and Development (OECD) countries. It includes the countries with the highest (80% or more) and lowest percentage of workers in the service sector. Those countries with the highest percentage of service jobs include smaller economies such as Luxembourg and the Netherlands, as well as large economies such as the United States and the United Kingdom. OECD countries with the smallest percentage of jobs in the service sector include former Eastern Bloc countries and Turkey. On an average, 72% of the jobs in OECD countries are service jobs.

Most research on services has been based in the empirical context of the older industrialized economies. As services have more recently expanded in Asia, so has research interest shifted in empirical focus (e.g., Daniels et al., 2005; Park, 2006; Ström, 2005, 2006).

There are doubts about the benefits of such expansion of increasingly service-oriented economies. First, there are arguments concerning limited productivity growth in service-dominated economies. If services are coproduced in face-to-face interactions, it is difficult for productivity to grow and eventually the growth getting difficult to measure. However, the organizational forms for the production and distribution of services vary immensely and there are service activities with strong, measurable productivity growth. Such services include standardized operations that have been able to replace staff with computers in recent years (banks, retail, travel and transportation, etc.) (Bryson et al., 2004: pp. 20–21).

Other doubts about the benefits of service-based economies center on the tradability constraints with limited value as an export industry. The concern is that the changing balance between manufacturing and services (with the service sector predominating) threatens the overall export volume of national economies. However, this line of reasoning is empirically untrue. Over time, service companies have developed into major transnational corporations doing international business. Over half of the companies on the Forbes Global 2000 list (the world's largest public companies) are in the service industry segment. Banking and other financial services companies dominate the list, with more than 500 companies. Other service industries with more than 50 major international companies on the list are retail, transportation, construction, telecommunications, media and business services. Other categories of companies cut across the service–manufacturing divide, such as health care equipment and services, with more than 50 major global companies on the list.

Although tradability in the service industry has increased substantially, this increase does not fully correspond to the growing role this sector plays in the economy in terms of its share of GDP or employment. In OECD countries in 2008, exports of services were 23% of all exports (OECD/Eurostat, 2011).

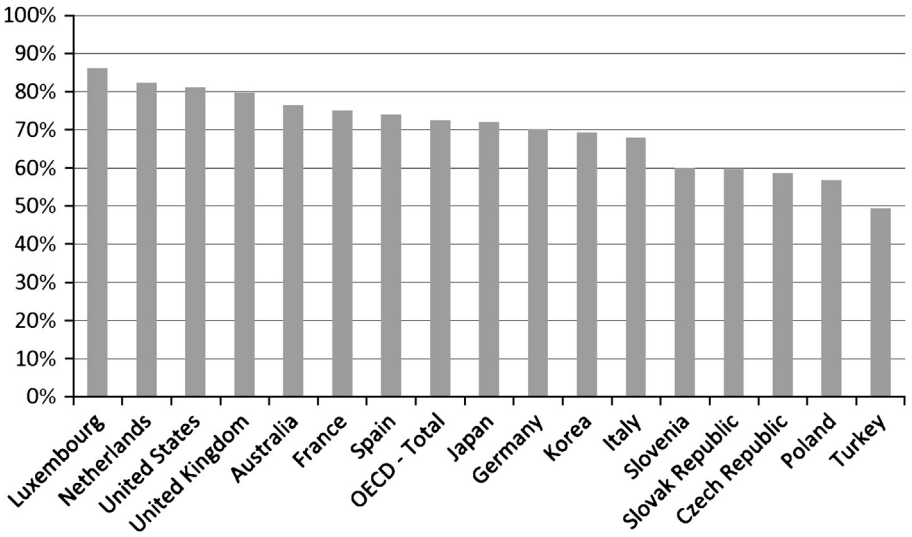

Figure 1 Percentage of workforce employed in services in a selection of OECD countries in 2011. Source: OECD Statistics OECD StatExtracts/Short-Term Labour Market Statistics. http://stats.oecd.org/Index.aspx?DataSetCode=STLABOUR.

Many different and mutually reinforcing factors are behind the internationalization of services (Daniels, 1993). These factors include the political liberalization and disappearance of national barriers to different types of international exchange (financial capital, direct investments, commodities, labor, etc.), standardization, the development of ICTs to facilitate communication and monitor spatially dispersed organizations, networks and relationships between suppliers and clients. The extent to which services are ICT-mediated and ICT-coded in a digital form further enhances their suitability for international markets (Bryson, 2007). This development leads to a discussion about the international division of labor for service tasks and about organizational forms that include offshoring or global sourcing.

As the discussion above indicates, the service industry contributes to the job market, to export incomes, and to the economy as a whole. In addition, more qualitative discussions cite the importance of the service industry (via the effects of its output) for regional development. For instance, service providers can increase the productivity and improve the competence of their clients' firms, as well as facilitate specialization of production networks. Firms in various sectors can also use services as a tool to differentiate their offerings from those of competitors (Kujala et al., 2013).

From a production network perspective, the growth of service jobs or service firms is partly an illusion in that, rather than being produced in-house, services are provided by external suppliers. This raises the issue of outsourcing, through which firms "switch from internal sourcing of inputs to external independently owned providers" (Bryson, 2007: p. 31) However, increased externalization of services is not just a matter of reorganization. Increased demand also relates to the recent demand for different types of services. Increasing complexity, technological developments, competition, expectations, and spending power in a globalized economy are all processes that explain this development. This applies to ICT services, education services and management, among other specialized and more routinized services.

The Service Industry, Uneven Development, and Regional Specialization

Although the expansion of the service industry to the international scale is an important trend, some services, including many consumer services and public sector services, cater to primarily local or regional markets. With spatial distributions that mimic the population settlement pattern, such services make important contributions to local and regional job markets. Other service activities, including APSs and public services, are mainly concentrated in larger urban areas. Such services were the focus of office location research in the 1970s, which investigated, inter alia, office relocation possibilities away from southeast England, a 'hot' issue because of the urban congestion and concomitant diseconomies of scale in London (e.g., Goddard, 1975; Daniels, 1979). More recently, authors have discussed the emergence of knowledge-intensive and creative industries in inner cities (Hutton, 2004). Studies have also illustrated the suburbanization of services along major infrastructure systems following the increasing costs for such services in city centers (e.g., Hermelin, 2007; Illeris, 1997).

Research in the United States on rural-based advanced services includes work on classic self-employed (Lone Eagles) and small (High Flyers) firms (Beyers and Lindahl, 1996). Such firms may operate from home with service clients all over the United States and abroad. Their means of operation include online communication as well as site visits to their clients' firms. Some home-based service firms examined in this work also had an urban office address that the owners visited from time to time. In South Korea, Park (2006) examined how self-employed experts form virtual teams, selected to match the needs of their clients in different locations. The regional specialization of international industries is another example of how client hot spots for services can occur outside of large urban areas (Bryson and Rusten, 2005).

Recent studies have also problematized theories of geographic clustering and innovation capacities, which maintain that innovation requires the colocation of actors and resources. The spatiality of innovation processes has different geographies in different types of milieus, such as metropolitan or more sparsely populated and remote regions (Doloreux and Shearmur, 2012).

Labor Force

Labor is the main resource and capital of service companies. The educational background of employees in service industries varies from relatively little education in the retail and transportation industries to high average levels in sectors such as business services, finance, and public administration. Relative to workers in insecure and temporary jobs, the service work force comprises some of the most privileged segments of the labor market in terms of incomes and other fringe benefits.

Highly competitive service companies and service functions, such as call centers, look for low-cost labor. The cost of labor varies among locations and may relate to aspects of workers such as gender. Low-skilled labor-oriented industries and activities faced with fierce price competition have produced a labor market with poorly paid workers and insecure jobs (Tickell, 1999).

Linda McDowell has elaborated on how service jobs involve different aspects of bodily performance and what is termed 'bodywork': "From sex work, through elder care to university teaching, an embodied performance is a central part of the exchange" (McDowell, 2009: p. 64). This embodiment of work typically means face-to-face interactions between workers and customers that require emotional labor. The example of 'empathetic' women workers in health care illustrate the gendered understanding of care delivered by female workers in the emotional work sector.

Research on the international trade of services has also focused on specific service experts, including work on expatriation in professional legal firms (Beaverstock, 2004) and studies of international law firms and their work and communication arrangements (Jones, 2007). It also includes research on growth strategies and modes of entry in a study of global headhunting firms (Faulconbridge et al., 2008) and studies of advertising firms operating

on different geographical levels in relationship with clients (Hermelin, 2009).

Part 2. Advanced Producer Services in Focus

High-order or APSs are widely acknowledged to contribute to the creation of wealth in an economy. Although the definition of APS varies, it is possible to find common ground as summarized by Beyers (2001): "Business, legal, and engineering and management services are included in most studies, while the financial services are often considered to be part of the producer services. It is less common to consider services to transportation, wholesaling and membership organizations as a component of the producer services" (p. 13930). Different authors also use different labels for this sector, sometimes equating it with the KIBS concept.

The main argument for the importance of APS is the increasing significance of knowledge, learning, and innovation in modern economies. APSs are important components in mediating, customizing, and developing knowledge and contributing to the learning process among client firms. KIBS are key constituents of innovation systems (Cooke and Leydesdorff, 2006) as well as key vectors of knowledge transmission (MacPherson, 2008). Though KIBS act as enablers, they are also innovators in their own right, developing new combinations of knowledge, services, ways of accessing markets, and ways of producing and delivering services (Miles, 1993).

ICT and Service Activities across Organizations and Space

ICT services, one sector of APS, is an important example on how such companies are important for innovations systems. ICTs provide organizations with a new type of flexibility in work arrangements (Coffey and Bailly, 1991). The ICT tool became a very efficient way for business to access information and communicate online across space.

Many innovations in services involve combinations of ICTs (e.g., ICT hardware, software, 3D scanners, webcameras, and smart boards) with organizational changes, including different types of self-service systems (e.g., automatic telling machines (ATMs), check-in systems at airports and tax reporting systems). This has resulted in more time-efficient and less labor-intensive operations, as well as in facilitating customer projects, design, distribution, reporting, training and control of data across systems, and geographical distances. The advent of 3D technologies for example, enables engineers, architects, and designers to better visualize products and thus to integrate their different areas of expertise more closely in tailoring the details of the production process (Rusten and Bryson, 2010).

Technological shifts in how services are organized have actually pushed some work from the service provider to the user. The implication of this technological rationalization is a drastic cut in number of service offices, such as bank units, travel agencies, and post offices. These cutbacks have also led to a more spatially centralized location pattern.

It is also important to address the interdependencies among technologically oriented innovation projects (innovations in services) and research and development, technological services, and Technology Transfer Organizations (services for innovation). The interdependencies do, for example, concern the forms and conditions of entrepreneurship and service innovations. Monnoyer-Longé (2010) uses empirical evidence from Toulouse to discuss the structure of these types of service firms and how firms are supported by the local economy. Related topics concern what parts of an innovation process are internal versus external to an organization, (Bryson and Monnoyer, 2002; Gallouj, 2002). Sundbo (1997) has studied the level of innovations and ways of organizing among service firms with empirical evidence from Denmark.

Outsourcing is about the transferring of in-house activities to external providers that may be geographically near or distant, and offshoring refers to those cases where the external providers are in another country. The occurrence of service offshoring expanded with the introduction of call centers in the 1970s and 1980s, but since then, such offshoring has come to involve more knowledge-intensive services. Programming, accounting services, engineering and medical diagnostic services can all be electronically transmitted over long distances. As a result, ICT enables organizations to become more flexible in how work is arranged spatially (Coffey and Bailly, 1991). In some cases, it was a way of cutting costs or getting around the problem with scarce local access of expertise (e.g., places with recruitment problems and dearth of local service providers).

Cost savings, a wish to focus on the core business and gaining access to additional capacity and competence are among the motivations for outsourcing or offshoring (Wernerheim, 2012). Internal service solutions and shared service centers are other examples of business models that offshoring vendors use in the spatial organization of these projects. Market trends influence the way businesses manage their operations. There are reasons to expect that continued development, adoption and integration of technologies would lead to increased offshoring, as more information is automated and digitized.

Geographers have examined the way client firms and service providers find each other. Sourcing strategies (such as where and how to search) are partly the result of internal structural elements, including motivation, knowledge, previous experience, and, above all, financial position. Rusten et al. (2005) explain how multisite companies that use a professional purchasing department for the whole company tend to weaken ties to local service providers that the local units of the company had established. Awareness of local qualities will be less evident if such a purchasing department is located elsewhere. These new forms of sourcing knowledge have eroded some of the obvious advantages associated with clusters as many services can easily be accessed from elsewhere. Yet another alternative is negotiating procurements through brokers. The broker may procure a single service or a bundle of different services packaged together. This bundle can be directly supplied by a single supplier or coordinated by a broker that chooses among different suppliers. Using a bundling strategy in bid announcements is a way to reduce the number of service providers and transfer some of the management costs in organizing such projects to a third party. Contractual issues

can include questions concerning the size, length, quality, responsibilities, and other details about the delivery of services. A further way of limiting the number of suppliers can be to group several sites under the same contract. The implication of winner that takes all is twofold. First, companies and areas receiving contracts may face capacity problems and need to increase their number of staff. Second, those that do not win contracts might have to reduce their staff or close their operations. Both effects will create a more unstable and unpredictable situation for the workforces and lead to greater mobility between companies and locations.

An important organizational reform that challenged the quality, efficiency, and division of labor between public and private services has been the modernization of the service sector with new public management (NPM). Starting in the United Kingdom and New Zealand in the beginning of the 1980s, NPM later spread to several other countries. The organizational changes this modernization brought expanded the field of business services to accommodate issues such as practice of tendering for bids, contract management, performance measurements, training, and legal matters. NPM research includes discussion about the efficiency of NPM reforms to rural areas (Hanlon and Rosenberg, 1998).

The Sarbanes–Oxley Act (SAX) came after the accounting scandal in the US service industries in 2001 (Enron, Arthur Andersen) and regulates all companies American and foreign that are listed on US stock exchanges. To avoid conflicts of interest and to increase transparency, this Act introduced several regulations to raise auditing standards and encourages organizational and legal separation between consulting and accounting services. In the same period, several mergers and acquisitions among multinational consulting firms created service providers with expertise in management and ICT (Bryson and Rusten, 2005). Examples are PricewaterhouseCoopers, sold to IBM in 2002, and several acquisitions within Cap Gemini and KPMG. A European version of SAX came with the 8th Company Law Directive in the European Union in 2006.

Another regulation that has promoted services was the Directive on Services in the Internal Market, which aimed to stimulate growth in the service industries in Europe by removing some legal and administrative barriers. The service directive includes a range of services to both business and private individuals, including business-related services, trade, food, and accommodation, but excluding finance, temporary agency services, temporary work agency services, and social services provided by the State (http://ec.europa.eu/internal_market/top_layer/services/index_en.htm). The list of services that were included and excluded once again challenges conventional conceptual understandings of services that undergirded research in this field. Figure 2 below describes major technological, organizational, and regulative changes that have challenged the conceptual understanding of services and have thrown up new issues for research.

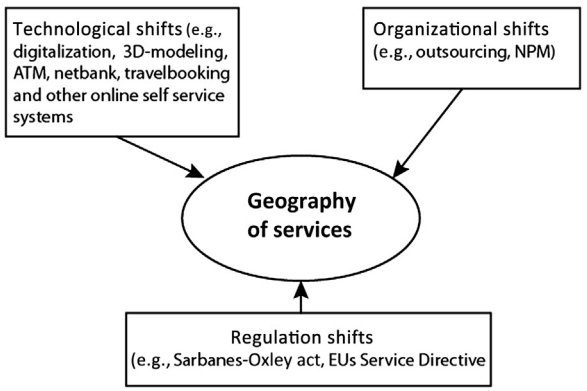

Figure 2 Development of major factors forming the understanding of and research on services.

Part 3. Recent and Future Research Directions

Future directions for research can be traced through scholarly conferences and networks. In this context, the European Research Network on Services and Space (RESER), which was founded in Lyon in 1991 (Bonamy et al., 1993; http://www.reser.net/) is an important forum. Its annual conference is a meeting place for geographers, economists, and other social scientists conducting service research.

Recent publications in major journals of geography confirm the likely continuance of a focus on APSs although also less-established themes are represented. These upcoming fields also include discussions on new organizational forms like for-profit education services and joint projects of private and public sector organizations (Hall and Appleyard, 2011; Siemiatycki, 2012). Recently published articles on ecosystem services have also led discussions to address environmental policy more widely (e.g., Dempsey and Robertson, 2012). These examples will once again challenge our understanding about what services are.

See also: Economic Geography; Economic Restructuring: Geographic Aspects; Finance, Geography of; Globalization: Geographical Aspects; Industrial Geography; International Business: Geographic Aspects; Labor, Geography of; Market Geographies.

Bibliography

Aslesen, Heidi Wiig, Isaksen, Arne, 2007. New perspectives on knowledge-intensive services and innovation. Geografiska Annaler Series B-Human Geography 89B (Suppl. 1), 45–58.

Beaverstock, Jonathan V., 2004. Managing across borders: knowledge management and expatriation in professional legal services. Journal of Economic Geography 4, 157–179.

Beyers, William B., 2001. Service economy, geography of. In: International Encyclopedia of Social and Behavioral Sciences. Elsevier, London, pp. 13930–13935.

Beyers, William B., Lindahl, David P., 1996. Lone eagles and high fliers in rural producer services. Rural Development Perspectives 12, 2–10.

Bonamy, Joel, Daniels, Peter W., Illeris, Sven, Phillippe, Jean, 1993. Editors' preface. The Service Industries Journal 13 (1), 1–2.

Browning, Harley L., Singelmann, Joachim, 1975. The Emergence of a Service Society: Demographic and Sociological Aspects of the Sectoral Transformation of the Labor Force in the U.S.A. National Technical Information Service, Springfield, VA.

Bryson, John R., 2007. The 'second' global shift: the offshoring or global sourcing of corporate services and the rise of distanciated emotional labour. Geografiska Annaler Series B-Human Geography 89B (Suppl. 1), 31–43.

Bryson, John R., Daniels, Peter W. (Eds.), 2007. The Handbook of Service Industries. Edward Elgar, Cheltenham.
Bryson, John R., Daniels, Peter W., Warf, Barney, 2004. Service Worlds: People, Organizations, Technologies. Routledge, London.
Bryson, John. R., Monnoyer, Marie-Christine, 2002. Understanding the relationship between services and innovation: the RESER review of the European service literature on innovation 2002. The Service Industries Journal 24 (1), 205–222.
Bryson, John R., Rusten, Grete, 2005. Spatial divisions of expertise: knowledge intensive business service firms and regional development in Norway. The Service Industries Journal 25, 1–19.
Coffey, William J., Bailly, Antoine S., 1991. Producer services and flexible production: an exploratory analysis. Growth and Change 22, 95–117.
Cooke, Phillip, Leydesdorff, Loet, 2006. Regional development in the knowledge-based economy: the construction of advantages. Journal of Technology Transfer 31, 5–15.
Daniels, Peter W., 1979. Spatial Patterns of Office Growth and Location. Wiley, London.
Daniels, Peter W., 1985. Service Industries. A Geographical Appraisal. Methuen, London.
Daniels, Peter W., 1993. Service Industries in the World Economy. Blackwell, Oxford.
Daniels, Peter W., Ho, Kong Chong, Hutton, Thomas, 2005. Service Industries and Asia-Pacific Cities: New Development Trajectories. Routledge, London.
Dempsey, Jessica, Robertson, Morgan M., 2012. Ecosystem services: Tensions, impurities, and points of engagement within neoliberalism. Progress in Human Geography 36, 758–779.
Doloreaux, David, Shearmur, Richard G., 2012. Collaboration, information and the geography of innovation in knowledge intensive business services. Journal of Economic Geography 12, 79–105.
Edvardsson, Bo, Gustafsson, Anders, Roos, Inger, 2005. Service portraits in service research: a critical review. International Journal of Service Industry Management 16, 107–121.
Faulconbridge, James, Hall, Sarah, Beaverstock, Jonathan, 2008. New insights into the internationalization of producer services: organizational strategies and spatial economies for global headhunting firms. Environment and Planning A 40, 210–234.
Gallouj, Faiz, 2002. Innovation in services and the attendant old and new myths. The Journal of Socio-Economics 31, 137–154.
Goddard, John, 1975. Office Location in Urban and Regional Development. Oxford University press, Oxford.
Hall, Sarah, Appleyard, Lindsay, 2011. Commoditising learning: cultural economy and the growth of for-profit business education service firms in London. Environment and Planning A 43, 10–27.
Hanlon, Neill T., Rosenberg, Mark W., 1998. Not-so-new public management and the denial of geography: Ontario health-care reform in the 1990s. Environment and Planning C: Government and Policy 16 (5), 559–572.
Hermelin, Brita, 2007. The urbanization and suburbanization of the service economy: producer services and specialization in Stockholm. Geografiska Annaler 89B (S1), 59–74.
Hermelin, Brita, 2009. Producer service firms in globalizing cities: the example of advertising firms in Stockholm. The Service Industries Journal 29, 457–471.
Hill, T.P., 1977. On goods and services. Review of Income and Wealth 23, 315–338.
Hutton, T., 2004. The new economy of the inner city. Cities 21 (2), 89–108.
Illeris, Sven, 1997. The changing location of service activities in the Copenhagen region. Geografisk Tidskrift/Danish Journal of Geography 97 (1), 120–131.
Jones, Andrew M., 2007. More than managing across borders? The complex role of face-to-face interaction in globalizing law firms. Journal of Economic Geography 7, 223–246.
Kujala, Jaako, Ahola, Tuomas, Hikuri, Satu, 2013. Use of services to support the business of a project-based firm. International Journal of Project Management 31, 177–189.
MacPherson, Alan, 2008. Producer service linkages and industrial innovation: results of a twelve-year tracking study of New York state manufacturers. Growth and Change 39 (1), 1468–2257.
McDowell, Linda, 2009. Working Bodies: Interactive Service Employment and Workplace Identities. John Wiley & Sons, Chichester.
Miles, Ian, 1993. Services in the new industrial economy. Futures July/August 25 (6), 653–672.
Monnoyer-Longé, Marie-Christine, 2010. Entrepreneurship and service innovation, a challenge for local development. In: Gallouj, Faiz (Ed.), Innovation and Services. Edward Elgar Publishing, Cheltenham, pp. 573–587.
OECD/Eurostat, 2011. Summary tables: main categories of services and zone totals. In: OECD Statistics on International Trade in Services 2010. Detailed Tables by Service Category, vol. I. OECD Publishing.
Park, Sam Ock, 2006. Service worlds and the dynamics of economic spaces. In: Harrington, James W., Daniels, Peter.W. (Eds.), Knowledge-Based Services: Internationalization and Regional Development. Ashgate, Aldershot, pp. 15–40.
Rusten, Grete, Bryson, John R., 2010. Industrial Design, Competition and Globalization. Palgrave Macmillan, Basingstoke.
Rusten, Grete, Bryson, John R., Gammelsæter, Hallgeir, 2005. Dislocated versus local business service expertise and knowledge: the acquisition of external management consultancy expertise by small and medium-sized enterprises in Norway. Geoforum 36, 525–539.
Siemiatycki, Matti, 2012. Public-private partnership networks: exploring business-government relationships in United Kingdom transportation projects. Economic Geography 87, 309–334.
Singelmann, Joachim, 1978. From Agriculture to Services: The Transformation of Industrial Employment. Sage, Beverly Hills, CA.
Stanback Jr, Thomas M., 1979. Understanding the Service Economy. Employment, Productivity, Location. The Johns Hopkins University Press, Baltimore and London, p. 5.
Ström, Patrik, 2005. The Japanese service industry: an international comparison. Social Science Japan Journal 8, 253–266.
Ström, Patrik, 2006. Internationalization of Japanese professional business service firms: dynamics of competitiveness through urban localization in Southeast Asia. In: Harrington, James W., Daniels, Peter W. (Eds.), Knowledge-Based Services: Internationalisation and Regional Development. Ashgate, Aldershot, pp. 153–174.
Sundbo, Jon, 1997. Management of innovation in services. The Service Industries Journal 17, 432–455.
Tickell, Adam, 1999. The geographies of services: new wine in old bottles. Progress in Human Geography 23, 633–641.
Toivonen, Marja, Tuominen, Tiina, 2009. Emergence of innovations in services. The Service Industries Journal 29, 887–902.
Vargo, Stephen L., Lusch, Robert F., 2008. Service-dominant logic: continuing the evolution. Journal of the Academic Marketing Science 36, 1–10.
Wernerheim, Michael C., 2012. Does service offshoring boost productivity? Some Canadian evidence on causation. The Service Industries Journal 32, 551–569.

Relevant Websites

http://ec.europa.eu/internal-market/services – European Commission-Service Directive.
www.reser.net – RESER European Association for Research on Services.

Settlement and Landscape Archaeology

Gary M Feinman, The Field Museum, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Settlement pattern archaeology and the investigation of ancient landscapes, especially when systematically implemented, have been some of the most significant archaeological innovations of the last half century. These studies have shed new light on the emergence of hierarchically organized and urban societies in regions around the world, while also providing new perspectives on the history of human–environmental interactions. This article reviews the roots of these regional archaeological approaches, their theoretical underpinnings, and some of the key empirical contributions.

Settlement pattern and landscape approaches are central to the mission of contemporary archaeology (Kowalewski, 2008; Renfrew, 2003, p. 313; Sabloff and Ashmore, 2001, p. 14). Through archaeological surveys, they provide a regional perspective on behavioral change that has been key to the transition from normative to relational or populational perspectives on the past. Although these studies have had the greatest impact on the understanding of arid and semiarid areas, they have been employed under a range of conditions. There is no correct way to survey; however, methodological procedures and analytical strategies must be guided by environmental conditions, available resources, and research goals. The most successful studies to date have been those in which significant and sustained time and labor investments have been made.

For contemporary archaeology, settlement and landscape approaches represent an increasingly important focus that is vital for a core mission of the discipline to describe, understand, and explain long-term cultural and behavioral changes. Despite this significance, relatively few syntheses of this topic have been undertaken (cf Ammerman, 1981; Billman and Feinman, 1999; Fish and Kowalewski, 2009; Kowalewski, 2008; Parsons, 1972). Yet settlement and landscape approaches provide the only large-scale perspective for the majority of premodern societies. These studies are reliant on archaeological surface surveys, which discover and record the distribution of material traces of past human presence/habitation across a landscape (*see* Survey and Excavation (Field Methods) in Archaeology). The examination and analysis of these physical remains found on the ground surface (e.g., potsherds, stone artifacts, house foundations, or earthworks) provide the empirical foundation for the interpretation of ancient settlement patterns and landscapes.

Historical Background

Although the roots of settlement pattern and landscape approaches extend back to the end of the nineteenth century, archaeological survey has only come into its own in the post-World War II era. Spurred by the analytical emphases of Steward (1938), Willey's Virú Valley Archaeological Survey (1953) provided a key impetus for settlement pattern research in the Americas. In contrast, the landscape approach, which has a more focal emphasis on the relationship between sites and their physical environments, has its roots in the United Kingdom. Nevertheless, contemporary archaeological studies indicate a high degree of intellectual cross-fertilization between these different surface approaches.

Early Foundations for Archaeological Survey in the Americas and England

The American settlement pattern tradition stems back to scholars such as Morgan (1881), who queried how the remnants of Native American residential architecture reflected the social organization of the native peoples who occupied them. Yet the questions posed by Morgan led to relatively few immediate changes in how archaeology was practiced, and for several decades few scholars endeavored to address the specific questions regarding the relationship between settlement and social behavior that Morgan posed. When surface reconnaissance was undertaken by archaeologists, it tended to be a largely unsystematic exercise carried out to find sites worthy of excavation.

In the United Kingdom, the landscape approach, pioneered by Fox (1923), was more narrowly focused on the definition of distributional relationships between different categories of settlements and environmental features (e.g., soils, vegetation, and topography). Often, these early studies relied on and summarized surveys and excavations that were carried out by numerous investigators using a variety of field procedures rather than more uniform or systematic coverage implemented by a single research team. At the same time, the European landscape tradition generally has had a closer link to romantic thought as opposed to the more positivistic roots of the North American settlement pattern tradition (e.g., Sherratt, 1996).

The Development of Settlement Archaeology

By the 1930s and 1940s, US archaeologists working in several global regions recognized that changing patterns of social organization could not be reconstructed and interpreted through empirical records that relied exclusively on the excavation of a single site or community within a specific region. For example, in the lower Mississippi Valley, Phillips et al.

(1951) located and mapped archaeological sites across a large area to analyze shifting patterns of ceramic styles and settlements over broad spatial domains and temporal contexts.

Yet the most influential and problem-focused investigation of that era was that of Willey in the Virú Valley. Willey's project was the first to formally elucidate the scope and potential analytical utility of settlement patterns for understanding long-term change in human economic and social relationships. His vision moved beyond the basic correlation of environmental features and settlements as well as beyond the mere definition of archetypical settlement types for a given region. In addition to its theoretical contributions, the Virú program also was innovative methodologically, employing (for the first time in the Western Hemisphere) vertical air photographs in the location and mapping of ancient settlements. Although Willey did not carry out his survey entirely on foot, he did achieve reasonably systematic areal coverage for a defined geographic domain for which he could examine changes in the frequency of site types, as well as diachronic shifts in settlement patterns.

Conceptually and methodologically, these early settlement pattern projects of the 1930s and 1940s established the intellectual underpinnings for a number of multigenerational regional archaeological survey programs that were initiated in at least four global regions during the 1950s and 1960s. In many ways, these later survey programs were integral to the theoretical and methodological reevaluations that occurred in archaeological thought and practice under the guise of 'the New Archaeology,' or processualism. The latter theoretical framework stemmed in part from an expressed emphasis on understanding long-term processes of behavioral change and cultural transition at the population (and so regional) scale. This perspective, which replaced a more normative emphasis on archetypical sites or cultural patterns, was made possible to a significant degree by the novel diachronic and broad scalar vantages pieced together for specific areas through systematic regional settlement pattern fieldwork and analysis.

Large-Scale Regional Survey Programs

During the 1950s through the 1970s, major regional settlement pattern programs were initiated in the heartlands of three areas where early civilizations emerged (Greater Mesopotamia, highland Mexico, and the Aegean), as well as in one area known for its rich and diverse archaeological heritage (the Southwest United States). The achievements of the Virú project also stimulated continued Andean settlement pattern surveys, although a concerted push for regional research did not take root there until somewhat later (e.g., Billman and Feinman, 1999; Parsons et al., 1997).

Beginning in 1957, Robert M. Adams (e.g., 1965, 1981) and his associates methodically traversed the deserts and plains of the Near East by jeep, mapping earthen tells and other visible sites. Based on the coverage of hundreds of square kilometers, these pioneering studies of regional settlement history served to unravel some of the processes associated with the early emergence of social, political, and economic complexity in Greater Mesopotamia. Shortly thereafter, in highland Mexico, large-scale, systematic surveys were initiated in the area's two largest mountain valleys (the Basin of Mexico and the Valley of Oaxaca). These two projects implemented field-by-field, pedestrian coverage of some of the largest contiguous survey regions in the world, elucidating the diachronic settlement patterns for regions in which some of the earliest and most extensive cities in the ancient Americas were situated (e.g., Blanton et al., 1993; Sanders et al., 1979). After decades, about half of the Basin of Mexico and almost the entire Valley of Oaxaca were traversed by foot (Balkansky, 2006).

In the Aegean, regional surveys (McDonald and Rapp, 1972; Renfrew, 1972) were designed to place important sites with long excavation histories in broader spatial contexts. Once again, these investigations brought new regional vantages to areas that already had witnessed decades of excavation and textual analyses (Galaty, 2005). Over the same period, settlement pattern studies were carried out in diverse ecological settings across the US Southwest, primarily to examine the differential distributions of archaeological sites in relation to their natural environments, and to determine changes in the numbers and sizes of settlements across the landscape over time. More recently, new projects have introduced systematic, broad-coverage, archaeology settlement pattern studies to other regions, including parts of northern China (e.g., Feinman et al., 2010; Peterson et al., 2010; Underhill et al., 2008) and Madagascar (Wright, 2007). In each of the global areas investigated, the wider the study domain covered, the more diverse and complex were the patterns found. Growth in one part of a larger study area was often timed with the decrease in the size and number of sites in another. And settlement trends for given regions generally were reflected in episodes of both growth and decline.

Each of these major survey regions (including much of the Andes) is an arid to semiarid environment. Without question, broadscale surface surveys have been most effectively implemented in regions that lack dense ground cover, and therefore the resultant field findings have been most robust. In turn, these findings have fomented long research traditions carried out by trained crews, thereby contributing to the intellectual rewards of these efforts. As Ammerman (1981, p. 74) has recognized, "major factors in the success of the projects would appear to be the sheer volume of work done and the experience that workers have gradually built up over the years."

Settlement Pattern Research at Smaller Scales of Analysis

Although settlement pattern approaches were most broadly applied at the regional scale, other studies followed similar conceptual principles in the examination of occupational surfaces, structures, and communities. At the scale of individual living surfaces or house floors, such distributional analyses have provided key indications as to which activities (such as cooking, food preparation, and toolmaking) were undertaken in different sectors (activity areas) of specific structures (e.g., Flannery and Winter, 1976) or surfaces (e.g., Flannery, 1986, pp. 321–423). In many respects, the current emphasis on household archaeology (e.g., Santley and Hirth, 1993; Wilk and Rathje, 1982) is an extension of settlement pattern studies. Both household and settlement pattern approaches have fostered a growing interest in the nonelite sector of complex societies, and so have spurred the effort to understand societies as more than just undifferentiated, normative wholes.

At the intermediate scale of single sites or communities, settlement pattern approaches have compared the distribution of architectural and artifactual evidence across individual sites. Such investigations have clearly demonstrated significant intrasettlement variation in the functional use of space (e.g., Hill, 1970), as well as distinctions in socioeconomic status and occupational history (e.g., Blanton, 1978). From a comparative perspective, detailed settlement pattern maps and plans of specific sites have provided key insights into the similarities and differences between contemporaneous cities and communities in specific regions, as well as the elucidation of important patterns of cross-cultural diversity (e.g., Arnauld et al., 2012).

Contemporary Research Strategies and Ongoing Debates

The expansion of settlement pattern and landscape approaches over the last decades has promoted the increasing acceptance of less normative perspectives on cultural change and diversity across the discipline of archaeology. In many global domains, archaeological surveys have provided a new regional-scale (and in a few cases, macroregional-scale) vantage on past social systems. Settlement pattern studies also have yielded a preliminary means for estimating the parameters of diachronic demographic change and distribution at the scale of populations, something almost impossible to obtain from excavations alone. Nevertheless, important discussions continue over the environmental constraints on implementation, the relative strengths and weaknesses of different survey methodologies, issues of chronological control, procedures for population estimation, and the appropriate means for the interpretation of settlement pattern data.

Environmental Constraints

Although systematic settlement pattern and landscape studies have been undertaken in diverse environmental settings including heavily vegetated locales such as the Guatemalan Petén, the eastern woodlands of North America, and temperate Europe, the most sustained and broadly implemented regional survey programs to date have been enacted in arid environments. In large part, this preference pertains to the relative ease of finding the artifactual and architectural residues of ancient sites on the surface of landscapes that lack thick vegetal covers. Nevertheless, archaeologists have devised a variety of means, such as the interpolation of satellite images, the detailed analysis of aerial photographs, and subsurface testing programs, that can be employed to locate and map past settlements in locales where they are difficult to find through pedestrian coverage alone (e.g., Chase et al., 2012). In each study area, regional surveys also have to modify their specific field methodologies (the intensity of the planned coverage) and the sizes of the areas that they endeavor to examine to the nature of the terrain and the density of artifactual debris (generally nonperishable ancient refuse) associated with the sites in the specified region. For example, sedentary pottery-using peoples generally created more garbage than did mobile foragers; the latter usually employed more perishable containers (e.g., baskets and cloth bags). Consequently, other things being equal, the sites of foragers are generally less accessible through settlement and landscape approaches than are the ancient settlements that were inhabited for longer durations (especially when ceramics were used).

Survey Methodologies and Sampling

Practically since the inception of settlement pattern research, archaeologists have employed a range of different field survey methods. A critical distinction has been drawn between full-coverage and sample surveys. The former approaches rely on the complete and systematic coverage of the study region by members of a survey team. To ensure the full coverage of large survey blocks, team members often space themselves 25–50 m apart, depending on the specific ground cover, the terrain, and the density of archaeological materials. As a consequence, isolated artifact finds can occasionally be missed. But the researchers generally can discern a reasonably complete picture of settlement pattern change across a given region. Alternatively, sample surveys by definition are restricted to the investigation of only a part of (a sample of) the study region. Frequently, such studies (because they only cover sections of larger regions) allow for the closer spacing of crew members.

Archaeologists have employed a range of different sampling designs. Samples chosen for investigation may be selected randomly or stratified by a range of diverse factors, including environmental variables. Nevertheless, regardless of the specific sampling designs employed, such sample surveys face the problem of extrapolating the results from their surveyed samples to larger target domains that are the ultimate focus of study. Ultimately, such sample surveys have been shown to be more successful at estimating the total number of sites in a given study region than at defining the spacing between sites or at discovering rare types of settlement. The appropriateness of sample design can be decided only by the kinds of information that the investigator aims to recover. There is no single correct way to conduct archaeological survey, but certain methodological procedures have proved more productive in specific contexts and given particular research aims.

Chronological Constraints and Considerations

One of the principal strengths of settlement pattern research is that it provides a broadscale perspective on the changing distribution of human occupation across landscapes. Yet the precision of such temporal sequences depends on the quality of chronological control (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology). The dating of sites during surveys must depend on the recovery and temporal placement of chronologically diagnostic artifacts from the surface of such occupations. Artifacts found on the surface usually are already removed from their depositional contexts. Finer chronometric dating methods generally are of little direct utility for settlement pattern research, since such methods are premised on the recovery of materials in their depositional contexts. Of course, chronometric techniques can be used in more indirect

fashion to refine the relative chronological sequences that are derived from the temporal ordering of diagnostic artifacts (typically pottery).

In many regions, the chronological sequences can be refined only to periods of several hundred years in length. As a result, sites of shorter occupational durations that may be judged to be contemporaneous in fact could have been inhabited sequentially. In the same vein, the size of certain occupations may be overestimated as episodes of habitation are conflated. Although every effort should be made to minimize such analytical errors, these problems in themselves do not negate the general importance of the long-term regional perspective on occupational histories that in many areas of the world can be derived from archaeological survey alone. Although the broad-brush perspective from surveys may never provide the precision or detailed views that are possible from excavation, they yield an encompassing representation at the population scale that excavations cannot achieve. Adequate holistic perspectives on past societies rely on the multiscalar vantages that are provided through the integration of wide-ranging archaeological surveys with targeted excavations.

Population Estimation

One of the key changes in archaeological thought and conceptualization over the past half century has been the shift from essentialist/normative thinking about ancient societies to a more populational perspective. But the issue of how to define past populations, their constituent parts, and the changing modes of interaction between those parts remains challenging at best. Clearly, multiscalar perspectives on past social systems are necessary to collect the basic data required to estimate areal shifts in population size and distribution. Yet considerable debate has been engendered over the means employed by archaeologists to extrapolate from the density and dispersal of surface artifacts pertaining to a specific phase to the estimated sizes of past communities or populations.

Generally, archaeologists have relied on some combination of the empirically derived size of a past settlement, along with a comparative determination of surface artifact densities at that settlement, to generate demographic estimates for a given community. When the estimates are completed for each settlement across an entire survey region, extrapolations become possible for larger study domains. By necessity, the specific equations to estimate past populations vary from one region to another because community densities are far from uniform over time or space (e.g., Fang et al., 2004). Yet due to chronological limitations, as well as the processes of deposition, disturbance, and destruction, the techniques for measuring ancient populations remain coarse-grained. Although much refinement is still needed to translate survey data into quantitative estimates of population with a degree of precision and accuracy, systematic regional surveys still can provide the basic patterns of long-term demographic change over time and space that cannot be ascertained in any other way.

The Interpretation of Regional Data

Beyond the broad-brush assessment of demographic trends and site distribution in relation to environmental considerations, archaeologists have interpreted and analyzed regional sets of data in a variety of ways. Landscape approaches, which began with a focused perspective on humans and their surrounding environment, have continued in that vein, often at smaller scales. Such studies often examine in detail the placement of sites in a specific setting with an eye toward landscape conservation and the meanings behind site placement (Sherratt, 1996). At the same time, some landscape studies have emphasized the identification of ancient agrarian features and their construction, use, and long-term implications for anthropogenic environmental impacts (e.g., Fisher et al., 2009).

In contrast, other settlement pattern investigations have employed a range of analytical and interpretive strategies. In general, these have applied more quantitative procedures and asked more comparatively informed questions. Over the last 40 years (e.g., Johnson, 1977), a suite of locational models derived from outside the discipline has served as guides against which different sets of archaeological data could be measured and compared. Yet debates have arisen over the underlying assumptions of such models and whether they are appropriate for understanding the preindustrial past. For that reason, even when comparatively close fits were achieved between heuristically derived expectations and empirical findings, questions regarding equifinality (similar outcomes due to different processes) emerged. More recently, theory-building efforts have endeavored to rework and expand these locational models to specifically archaeological contexts with a modicum of success. Continued work in this vein, and the integration of some of the conceptual strengths from both landscape and settlement pattern approaches are requisite to understanding the complex web of relations that govern human-to-human and human-to-environment interactions across diverse regions over long expanses of time.

Looking Forward: The Critical Role of Settlement Studies

The key feature and attribute of archaeology is its long temporal panorama on human socioeconomic formations. Understanding these formations and how they changed, diversified, and varied requires a regional/populational perspective (as well as other vantages at other scales). Over the last century, the methodological and interpretive tool kits necessary to obtain this broadscale view have emerged, diverged, and thrived. The emergence of archaeological survey (and settlement pattern and landscape approaches) has been central to the disciplinary growth of archaeology and its increasing ability to address and contribute to questions of long-term societal change. Why and how did inequality, demographic nucleation, and concentrations of political power become central to the current human career (*see* States and Civilizations, Archaeology of) and in what ways did these historical processes differ and parallel each other across the globe (e.g., Peterson and Drennan, 2012; Smith et al., 2012).

Yet settlement pattern work has only relatively recently entered the popular notion of this discipline, long wrongly equated with and defined by excavation alone. Likewise, many archaeologists find it difficult to come to grips with

a regional perspective that has its strength in (broad) scalar representation at the expense of specific reconstructed detail. Finally, the potential for theoretical contributions and insights from settlement pattern and landscape approaches (and the wealth of data collected by such studies) has only scratched the surface. In many respects, the growth of regional survey and analysis represents one of the most important conceptual developments of twentieth-century archaeology. Yet at the same time, there are still so many mountains (literally and figuratively) to climb.

See also: Chronology, Stratigraphy, and Dating Methods in Archaeology; States and Civilizations, Archaeology of; Survey and Excavation (Field Methods) in Archaeology.

Bibliography

Adams, R.M., 1965. Land behind Baghdad: A History of Settlement on the Diyala Plains. University of Chicago Press, Chicago.

Adams, R.M., 1981. Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates. University of Chicago Press, Chicago.

Ammerman, A.J., 1981. Surveys and archaeological research. Annual Review of Anthropology 10, 63–88.

Arnauld, M.C., Manzanilla, L.R., Smith, M.E. (Eds.), 2012. The Neighborhood as a Social and Spatial Unit in Mesoamerican Cities. University of Arizona Press, Tucson, AZ.

Balkansky, A.K., 2006. Surveys and Mesoamerican archaeology: the emerging macroregional paradigm. Journal of Archaeological Research 14, 53–95.

Billman, B.R., Feinman, G.M. (Eds.), 1999. Settlement Pattern Studies in the Americas: Fifty Years since Virú. Smithsonian Institution Press, Washington, DC.

Blanton, R.E., 1978. Monte Albán: Settlement Patterns at the Ancient Zapotec Capital. Academic Press, New York.

Blanton, R.E., Kowalewski, S.A., Feinman, G.M., Finsten, L.M., 1993. Ancient Mesoamerica: A Comparison of Change in Three Regions, second ed. Cambridge University Press, Cambridge, UK.

Chase, A.F., Chase, D.Z., Fisher, C.T., Leisz, S.J., Weishampel, J.F., 2012. Geospatial revolution and remote sensing LiDAR in Mesoamerican archaeology. Proceedings of the National Academy of Sciences of the United States of America 109, 12916–12921.

Fang, H., Feinman, G.M., Underhill, A.P., Nicholas, L.M., 2004. Settlement pattern survey in the Rizhao area: a preliminary effort to consider Han and pre-Han demography. Indo-Pacific Prehistory Association Bulletin 24, 79–82.

Feinman, G.M., Nicholas, L.M., Fang, H., 2010. The imprint of China's first emperor on the distant realm of eastern Shandong. Proceedings of the National Academy of Sciences of the United States of America 107, 4851–4856.

Fish, S.K., Kowalewski, S.A. (Eds.), 2009. The Archaeology of Regions: A Case for Full-coverage Survey. Percheron Press, Clinton Corners, NY.

Fisher, C.T., Hill, J.B., Feinman, G.M. (Eds.), 2009. The Archaeology of Environmental Change: Socionatural Legacies of Degradation and Resilience. University of Arizona Press, Tucson, AZ.

Flannery, K.V. (Ed.), 1986. Guilá; Naquitz: Archaic Foraging and Early Agriculture in Oaxaca, Mexico. Academic Press, Orlando, FL.

Flannery, K.V., Winter, M.C., 1976. Analyzing household activities. In: Flannery, K.V. (Ed.), The Early Mesoamerican Village. Academic Press, New York, pp. 34–47.

Fox, C., 1923. The Archaeology of the Cambridge Region. Cambridge University Press, Cambridge, UK.

Galaty, M.L., 2005. European regional studies: a coming of age? Journal of Archaeological Research 13, 291–336.

Hill, J.N., 1970. Broken K Pueblo: Prehistoric Social Organization in the American Southwest. University of Arizona Press, Tucson, AZ.

Johnson, G.A., 1977. Aspects of regional analysis in archaeology. Annual Review of Anthropology 6, 479–508.

Kowalewski, S.A., 2008. Regional settlement pattern studies. Journal of Archaeological Research 16, 225–285.

McDonald, W.A., Rapp Jr., G.R. (Eds.), 1972. The Minnesota Messenia Expedition. Reconstructing a Bronze Age Environment. University of Minnesota Press, Minneapolis, MN.

Morgan, L.H., 1881. Houses and House Life of the American Aborigines. US Department of Interior, Washington, DC.

Parsons, J.R., 1972. Archaeological settlement patterns. Annual Review of Anthropology 1, 127–150.

Parsons, J.R., Hastings, C.M., Matos, R., 1997. Rebuilding the state in highland Peru: herder–cultivator interaction during the Late Intermediate period in Tarama–Chinchaycocha region. Latin American Antiquity 8, 317–341.

Peterson, C.E., Lu, X., Drennan, R.D., Zhu, D., 2010. Chiefly communities in Neolithic Northeastern China. Proceedings of the National Academy of Sciences of the United States of America 107, 5756–5761.

Peterson, C.E., Drennan, R.D., 2012. Patterned variation in regional trajectories of community growth. In: Smith, M.E. (Ed.), The Comparative Archaeology of Complex Societies. Cambridge University Press, Cambridge, UK, pp. 88–137.

Phillips, P., Ford, J.A., Griffin, J.B., 1951. Archaeological Survey in the Lower Mississippi Alluvial Valley, 1941–47. Peabody Museum of Archaeology and Ethnology, Cambridge, MA.

Renfrew, C., 1972. The Emergence of Civilisation. The Cyclades and the Aegean in the Third Millennium BC. Methuen, London.

Renfrew, C., 2003. Retrospect and prospect: Mediterranean archaeology in a new millennium. In: Papadopoulos, J.K., Leventhal, R.M. (Eds.), Theory and Practice in Mediterranean Archaeology. Cotsen Institute of Archaeology, UCLA, Los Angeles, pp. 311–318.

Sabloff, J.A., Ashmore, W., 2001. An aspect of archaeology's recent past and its relevance in the new millennium. In: Feinman, G.M., Price, T.D. (Eds.), Archaeology at the Millennium: A Sourcebook. Springer, New York.

Sanders, W.T., Parsons, J.R., Santley, R.S., 1979. The Basin of Mexico: Ecological Processes in the Evolution of a Civilization. Academic Press, New York.

Santley, R.S., Hirth, K.G. (Eds.), 1993. Prehispanic Domestic Units in Western Mesoamerica. CRC Press, Boca Raton, FL.

Sherratt, A., 1996. 'Settlement patterns' or 'landscape studies'? Reconciling reason and romance. Archaeological Dialogues 3, 140–159.

Smith, M.E., Feinman, G.M., Drennan, R.D., Earle, T., Morris, I., 2012. Archaeology as a social science. Proceedings of the National Academy of Sciences of the United States of America 109, 7617–7621.

Steward, J.H., 1938. Basin Plateau Aboriginal Sociopolitical Groups. Bureau of American Ethnology, Washington, DC.

Underhill, A.P., Feinman, G.M., Nicholas, L.M., Fang, H., Luan, F., Yu, H., Cai, F., 2008. Changes in settlement patterns and the development of complex societies in southeastern Shandong, China. Journal of Anthropological Archaeology 27, 1–29.

Wilk, R.R., Rathje, W.L. (Eds.), 1982. Archaeology of the Household: Building a Prehistory of Domestic Life. American Behavioral Scientist 25 (6), 611–724.

Willey, G.R., 1953. Prehistoric Settlement Patterns in the Virú Valley, Peru. Bureau of American Ethnology, Washington, DC.

Wright, H.T., 2007. Early State Formation in Central Madagascar: An Archaeological Survey of Western Avaradrano. Memoirs, 43. Museum of Anthropology, University of Michigan, Ann Arbor, MI.

Sex Crimes

Mary A Finn and Sebastian Czech, Georgia State University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The definitions, nature, and extent of sex crimes committed against adults (rape/sexual assault and sex trafficking/prostitution) and children (sexual abuse, incest, pedophilia, and child sex trafficking) are presented. Three common explanations for sex offending are provided: feminist theory of rape, economic theory of sex trafficking, and pathological/individualistic explanations of child sexual abuse. Emerging research areas on juvenile sex offenders and female sex offenders, as well as sex offending in the contexts of the military, on college campuses, and in correctional settings (prisons, jails, juvenile confinement facilities) are presented.

Introduction

Sexual victimization of adults and children is a complex and universal problem that occurs in multiple cultures and societies. For both child and adult victims of sex crimes, involvement in sexual activity is unwanted or forced in order to satisfy another person's sexual gratification, need for power and control, or financial benefit. Information on the prevalence and incidence of sex crimes, whether collected through surveys of community or from official reports of victimization, underestimate the true extent of the phenomenon. The majority of perpetrators of sex crimes know their victims, and females are more likely to be victims and males more likely to be perpetrators, regardless of the age of the victim. The explanations for sex offending are varied, often dependent upon the sexual maturity (age) of the victims. Most recently, research attention has shifted to examining sex offending by women and by youth, and federal legislation has been enacted to address the extent of sexual assaults that occur in institutional settings, such as the military, college campuses, and correctional institutions.

Sex Crimes: Definitions, Nature, and Extent

Sex Crimes against Children

Protection of children from sexual abuse and other types of maltreatment is the joint responsibility of state child protection and law enforcement agencies. Sexual abuse is defined as the "involvement of the child in sexual activity to provide sexual gratification or financial benefit to the perpetrator, including contacts for sexual purposes, molestation, statutory rape, prostitution, pornography, exposure, incest, or other sexually exploitive activities" (U.S. Department of Health and Human Services, 2012: 125). Most states have laws that mandate certain professionals and institutions report suspected child abuse/neglect to appropriate legal authorities. As such, the action or inaction of parents and caretakers are the focus of such laws, and intervention efforts of child protection agencies include prevention of maltreatment and family preservation. However, not all abuse suffered by children is perpetrated by their parents or caretakers, but is instead perpetrated by peers, strangers, or acquaintances. If child abuse persists or worsens in the home, or if child abuse is perpetrated by someone outside the family, criminal charges may be levied. Thus, information on sexual victimization of children is available from reports made to child protection agencies and from official records of police, prosecutors, and courts.

The widespread prevalence of child sexual abuse internationally was documented initially by Finkelhor (1994) and more recently by Pereda et al. (2009). Child sexual abuse is a significant public health problem. Few epidemiological studies have been conducted to date, and many lack methodological rigor. Finkelhor (1994) analyzed reported prevalence rates in 21 countries and concluded that the prevalence of sexual abuse during childhood averaged around 20% for women and between 5% and 10% for men. More recently, Pereda et al. (2009) conducted a meta-analysis of prevalence studies in the field of sexual abuse, comprised of 65 original articles about child sexual abuse in 22 countries, and found significant difference in the prevalence of child sexual abuse by gender; the mean prevalence rate in men was 7.9% and in women was 19.7%. In addition, they noted that prevalence rates varied by continent, with rates highest in Africa and lowest in Europe. South Africa reported the highest prevalence rates for both men (60.9%) and women (43.7%), whereas France reported the lowest rates for both men (0.9%) and women (0.6%). Only three countries (i.e., Jordan, South Africa, and Tanzania) reported prevalence rates of child sexual abuse among men above 20%, whereas eight countries (i.e., Australia, Costa Rica, Israel, South Africa, Sweden, Switzerland, Tanzania, and the United States) reported prevalence rates of child sexual abuse among women above 20%.

Within the United States, sexual abuse of children is one of four major types of child maltreatment monitored through the National Child Abuse and Neglect Data System (NCANDS), an effort funded by federal government to collect and analyze data on child abuse and neglect from the states annually. In its most recent report, *Child Maltreatment 2011*, sexual abuse was the least common form of maltreatment, with less than 10% of children sexually abused. Of the 61 472 victims who suffered sexual abuse, 26.3% were 12–14 years of age and 21.8% were 15–17 years of age. Among the 885 003 perpetrators associated with maltreating a child, 6.2% (54 906) engaged in the sexual abuse of children. Further, 9 out of 10 of the perpetrators of child abuse and neglect were the child's parents, relatives, or the unmarried partners of the child's parents.

Additional information on sexual abuse of minors (those under 18 years of age) is available from the National Crime

Victimization Survey (NCVS), an annual self-report victimization survey in which persons aged 12 and older were drawn from a nationally representative sample of households in the United States. In 2011, an estimated 244 000 rape/sexual assault victimizations occurred, representing less than 4% of all criminal victimizations reported (Truman and Planty, 2012). Further analysis of NCVS data for a 15-year period reported that over 91% of all victims of rape or sexual assault were female. In comparing the sexual victimization rates across three time periods (1994–98, 1999–2004, 2005–10), the report concluded that rates of sexual victimization declined over time and by age group, with 12–17-year-olds reporting highest rate of sexual victimization (Planty et al., 2013). The rate of rape/sexual assault victimizations of youths aged 12–17 declined from 7.0 victimizations per 1000 youth in 1994 to 2.2 victimizations per 1000 youths in 2010 (White and Lauritsen, 2012).

Conducted between January and May 2008, the National Survey of Children's Exposure to Violence (NatSCEV) recorded the victimization experiences of a nationally representative sample of 4549 children aged 1 month to 17 years living in the contiguous United States and found that 1 in 16 (6.1%) of those surveyed were victimized sexually in the previous year and nearly 1 in 10 (9.8%) over their lifetimes. Sexual victimizations included attempted and completed rape, sexual assault by an adult or peer, sexual exposure (exhibitionism) by an adult or peer, and sexual harassment. Adolescents aged 14–17 were the most likely to be sexually victimized and girls were more likely than boys to be sexually victimized (Finkelhor et al., 2009b).

In addition to reports from surveys, additional information on sex crimes is available from the National Incident-Based Reporting System (NIBRS), an expansion of the Federal Bureau of Investigation's (FBI) Uniform Crime Reports (UCR). The UCR compiles information on crimes reported and arrests completed by law enforcement agencies throughout the United States. NIBRS collects a wider range of information on victims, offenders, and circumstances for a greater variety of criminal incidents than the UCR, however not every law enforcement jurisdiction reports data into NIBRS, so the data may not be representative of the United States. NIBRS provides information on sex crimes against children including sexual assault, statutory rape, child pornography, and prostitution. Finkelhor and colleagues have issued a series of reports on sex offending and victimization experiences of juveniles as reported in NIBRS. First, in an analysis of data collected for 2004 that included 29 participating states or more than 4 million criminal incidents, Finkelhor et al. (2009a) reported that juvenile sex offenders (those aged 6–17) numbered 89 000, comprised more than one-quarter of all sex offenders (juvenile and adult), and comprised more than one-third of sex offenders who victimized juveniles (those under 18). The vast majority (93%) of juvenile sex offenders were male and nearly half (46%) were between the ages 15 and 17. Compared to adults who committed sex offenses against minors, juvenile offenders were more likely to offend in groups, more likely to offend against acquaintances, more likely to victimize boys, more likely to victimize younger children (under age 12), and more likely to commit sodomy or fondling than rape. NIBRS data collected in 2000, that included 19 states or more than 2.8 million criminal incidents, indicated approximately 2900 incidents of child pornography were known to police. Nearly 6 in 10 victims were female, 1 in 4 were members of the offender's family, and 6 out of 10 victims were 12–17 year old (Finkelhor and Ormrod, 2004a). Additional information from NIBRS indicated that nearly 13 814 incidents involving prostitution were known to police in the 4-year period (1997–2000). Of these, about 5% of the incidents involved juvenile offenders, juvenile victims, or both juvenile victims and juvenile offenders. Compared to adult prostitution, juvenile prostitution incidents were more likely to involve male offenders, to involve a group rather than a single offender, and to occur in large cities (Finkelhor and Ormrod, 2004b). Finally, based on the analysis of data collected for NIBRS from 1996 to 2000, nearly 7557 persons aged 7–17 were victims of statutory rape. In more than 6 out of 10 of the statutory rape incidents, the victim and offender were acquaintances. Males accounted for approximately 5% of all statutory rape victims. In the vast majority of statutory rapes, victims and offenders were of the opposite sex (Troup-Leasure and Snyder, 2005).

Incest

Incest is commonly defined as "sexual relations between persons so closely related that their marriage would be illegal or forbidden by custom" (Flowers, 2006: 77). Finkelhor (1978) estimates that anywhere between 75% and 90% of all incest cases go unreported. Among reported cases of incest, the vast majority of aggressors are adult men and the victims are female children, typically between the ages of 6–11. Mother–son or mother–daughter incest is the least common type. Sibling sexual abuse is the least recognized form of incest and its prevalence is not well documented (Ballantine, 2012). However, it is estimated that the rate of sibling incest may be five times the rate of parent–child sexual abuse (Finkelhor, 1980). In most sibling incest, the aggressor is an older male child and the victim is a younger female child.

Pedophilia

Pedophilia is a term often used interchangeably with child sexual abuse or child molestation. However, pedophilia is a clinical diagnosis referring to a sexual preference for children over adults that often includes fantasies, and may not result in actual contact with a child (i.e., noncontact sexual abuse). The Diagnostic and Statistical Manual (DSM-V) used to guide clinicians modified the diagnosis of pedophilic disorders to include three types: the pedophilic type (attracted to prepubescent children, generally younger than 11), the hebephilic type (attracted to pubescent children, usually ages 11 through 14), and the pedohebephilic type (attracted to both) (Harvard Medical School, 2010). Pedophiles are either severely distressed by these sexual urges, experience interpersonal difficulties because of them, or act on them. In general, most individuals who engage in pedophilia are male; the ratio is estimated to be 10 males to every one female (Cohen and Galynker, 2009). However, there is growing interest in female sex offenders (Hall and Hall, 2007). Most of the studies on pedophilia have involved men convicted of crimes against children, and as such may not represent the population of persons with pedophilic tendencies (Harvard Medical School, 2010). Many typologies or classifications of pedophiles have

been developed, including whether they are attracted to a particular age range and/or sex of child, whether they target only family members, and whether child pornography and/or a computer was used to engage the child in sexual activity (Hall and Hall, 2007). Researchers have noted many misconceptions surround child sex offenders including that all are pedophiles, that they target strangers, that all were victims of child sexual abuse themselves, that all have high rates of recidivism (Richards, 2011).

Child Sex Trafficking/Commercial Sexual Exploitation of Children/Minors

Child sex trafficking has become an important topic of both domestic and international interest. While variously defined, child sex trafficking, also referred to as the commercial sexual exploitation of children or minors (CSEC/CSEM), includes at its core sexual abuse of a child by another person who receives money for providing another person sexual access to that child. The child may be engaged in prostitution, nude dancing, stripping, pornographic film/images, or other forms of transactional or fantasy sex (Flowers, 2001). By most estimates, child pornography is a multibillion dollar global enterprise with the United States not only generating the biggest demand (yielding an estimated $6 billion annually), but also producing 85% of the world's supply (Flowers, 2001). According to the United Nations, the United States ranks second in the world (behind Italy) as a destination or market for sexual exploitation for both children and adults (Monzini, 2004). Annual estimates on the scope of sexual exploitation of children in the United States range from as few as 14 500–17 500 (Markon, 2007) to as many as 300 000–500 000 (Estes and Weiner, 2001). However, reliable information on the nature and extent of child sex trafficking in the United States is scarce (Finkelhor and Ormrod, 2004a,b). The UCR, NIBRS, and NCVS do not provide a national statistical picture of the extent of child sex trafficking (Finkelhor and Ormrod, 2000). A recent report by the National Institute of Justice (NIJ) stated that the number of known cases is growing; between 1998 and 2004 almost 300 000 calls made to the National Center for Missing and Exploited Children's CyberTipline reported CSEC (Albanese, 2007). CSEM is a complex problem that primarily occurs in large urban areas in the United States (Finkelhor and Ormrod, 2004a), and affects almost equally male and female teens (Curtis et al., 2008; Finkelhor and Ormrod, 2004a). Runaways and vulnerable youths with a history of physical or sexual abuse also are targets of sexual exploitation (Wolak et al., 2008).

Sex Crimes against Adults

Adults, unlike children, rely primarily on the criminal justice system to respond to reports of sexual victimization, thus information on the extent of adult victimization and offending is available from sources discussed previously (NCVS, UCR, and NIBRS). Data analyzed from the NCVS has noted that between 1994 and 2010 the rates of sexual victimization have declined over time for all age groups (Planty et al., 2013). In 2010, about 270 000 rape or sexual victimizations (a rate of 1.8 per 1000 females) occurred. In 9 out of 10 incidents: the victim was a female, a single offender was involved, and no weapon was used. Only 35% of rape/sexual assaults were reported to the police in 2010.

The UCR collects information on attempted and completed acts of forcible rape, defined as the carnal knowledge of a female forcibly and against her will. This definition excludes male victims and victims who are under the age of consent to engage in sex. The UCR broadened its definition of forcible rape in 2011 to include vaginal or anal penetration with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim. UCR data reported after January 2013 will reflect this definition. Estimated rates of rape are not broken down by age of victim. For 2011, the estimated number of forcible rapes was 83 425 and the rate of forcible rapes was 52.7 per 100 000 persons (UCR, 2012). For 2011, 19 491 arrests for forcible rape were reported. The UCR figures are significantly lower than those reported in the NCVS. Sexual assault continues to be one of the most underreported crimes, so the UCR figures are not a valid estimate of its true extent, as evident by the higher number reflected in the NCVS.

The UCR also provides information on prostitution incidents involving adults, and in 2013 will report information on incidents of human trafficking, including sex trafficking. Both prostitution and sex trafficking involve the sale of sex for money (e.g., commercial sex). However, the key distinction is that in sex trafficking, force, fraud, or coercion is used to compel a person to engage in the sale of sex for money and in prostitution such elements are not present. The U.S. Department of State (2005) estimates 600 000–800 000 persons are trafficked for sex and labor annually across international borders worldwide; recent estimates suggest that 14 500–17 500 individuals are trafficked annually into the United States (U.S. Department of State, 2006). It is believed that the majority are trafficked for sex. The Uniform Crime Report indicated that law enforcement agencies made 62 670 arrests for prostitution in 2010; two-thirds of those arrests were of women (Snyder, 2012).

The Centers for Disease Control (2011) published results of a nationally representative survey of adults aged 18 or older in the United States on both lifetime and annual prevalence rates of their experiences with sexual violence victimizations. The National Intimate Partner and Sexual Violence Survey (NISVS) found that 1%, or approximately 1.3 million women, reported being raped by any perpetrator in the 12 months prior to taking the survey. Nearly 1 in 5 women (18.3%) and 1 in 71 men (1.4%) in the United States had been raped in their lifetime. More than half of female victims of rape reported that the perpetrator was an intimate partner and more than half of male victims reported being raped by an acquaintance. The majority of both female and male victims reported experiencing violence from one perpetrator and that perpetrator was most likely male.

Explanations of Sex Offending

The medical model has dominated the American perspective on what causes sex offending and how best to treat it (Zilney and Zilney, 2009). There is no one single cause of sex offending. In today's society, when thinking of what motivates sexual offending, one usually thinks that it is motivated by the need

for men to have power and control over the victim. However, when one examines the range of sex crimes that are committed, especially when the victim or perpetrator is a minor, and as the age difference grows between the victim and perpetrator, other explanations may be needed. Research on typologies or subtypes of sex offenders began by differentiating between offenders who targeted children versus adults, with further refinements thereafter. In addition, the typologies of child sex abusers and rapists that have developed over the years often sought to identify the motive for sex offending, as doing so was seen as critical to effective intervention. Explanations for sex offending against children are dominated by individual-level explanations that attempt to identify a pathological factor, such as increased sex drive or aggression; the presence of deviant sexual arousal, interests, or preferences; and deficits in intimacy or possession of dismissive attachment styles.

Similarly, research of men convicted of rape argued that sexually aggressive men differed from other men in antisocial tendencies, nonconformity, impulsivity, and hypermasculinity, but such differences faded when comparison was made to admitted rapists who were not convicted (Meyer, 2000). Feminist theorists argue that sexual violence is caused by the social structure. Specifically, patriarchal society posits that men are both structurally and ideologically superior to women. Women in patriarchal cultures are socialized to be submissive and inferior to men, who are entitled to access and regulate of women's sexuality (Zilney and Zilney, 2009). A major tenet of feminist's conceptualization of rape is that it is a violent act motivated by power and anger and not by sexual desire.

Finally, prostitution and sex trafficking are viewed largely as crimes motivated by economics (Kara, 2009; Wheaton et al., 2010), with rational choice theory applied to the crime of trafficking. Rational choice theory posits that individuals will engage in rational choice in the marketplace, and compare costs and benefits in the market in order to obtain the highest profit or level of well-being. In the context of sex trafficking, sellers (traffickers/pimps/managers) offer buyers (employers/johns/customers) differentiated products (vulnerable individuals/prostitutes/sex workers) based on price and preference. Economic profit is the driving motive for sex trafficking. Powerless and marginalized populations (in which women and children are disproportionately represented) are primarily targeted for economic exploitation.

Emerging Issues in Sex Crimes

Female Sex Offenders

Sex offending by women was originally been believed to be a rare event; however, females are involved in upward of 5% of sex offending (Bunting, 2007). Further, incidents of female sex offending are believed to be underreported because victims are embarrassed and/or fear that their disclosures will be dismissed as fabrications due to cultural assumptions about femininity and the glorification of sexual activity between adult women and underage men (Oliver, 2007). Female sex offenders often use strategies to subdue their victims and take advantage of victims with either physical force, methods of exploitation, or verbal abuse (Fisher and Pina, 2013). Female sex offenders, compared to male sex offenders, typically target family members and often engage in offending with a male co-offender (Oliver, 2007). Female sex offenders are less likely to repeat their offending behavior and as a result, are treated more leniently by the courts as they are seen as less of a threat (Sandler and Freeman, 2011).

Juvenile Sex Offenders

A significant minority of child sexual abuse (20–50%) is committed by adolescents not adults (Veneziano and Veneziano, 2002). While research on this population is growing, far less is known about juvenile sex offenders than their adult counterparts. Juvenile sex offenders vary in terms of the age of onset of offending and the age of the victims targeted for sexual abuse. Generally speaking, adolescent sex offenders target those younger in age, females, and persons that they know. Juvenile sex offending is often linked with conduct disorder, antisocial behavior, or substance abuse (Gerardin and Thibaut, 2004). According to Gerardin and Thibaut (2004), juvenile sex offenders were often victims of sexual or physical abuse; studies report prevalence of prior abuse among juvenile sex offenders to range from 40% (Ryan et al., 1996) to 80% (Friedrich and Luecke, 1988; Ryan et al., 1987). Additional research has examined if juveniles who engage in sex offending are likely to continue in that pattern of criminal activity into adulthood. Zimring et al. (2007) in their longitudinal study found no evidence that juvenile sex offenders specialize in sex offending as adults, and that the best predictor of adult sex offending was frequency of police contact and not sex offending as an adolescent.

Sexual Assaults in Institutional Settings

The Military

The official definition of sexual assault in the armed forces includes "intentional sexual contact, characterized by use of force, physical threat or abuse of authority or when the victim does not or cannot consent (Army Regulation 600-20, as cited in Hillman, 2009)." Attention to the issues of sexual violence in the military occurred during the Vietnam War and gained momentum after the publication of Susan Brownmiller's groundbreaking book, *Against Our Will* in 1975. Recent research suggests that military-on-military sexual violence is a significant part of military crime and it persists across lines of rank and branches of the service (Hillman, 2009).

Information on the extent and nature of sexual violence in the military comes in two forms: actual reports filed to command and estimates based on anonymous claims. Military services received 3374 reports of sexual assaults during federal fiscal 2012 (1 October 2011 to 30 September 2012), and 26 000 persons made an anonymous claim of sexual assault but chose not to report the incident U.S. Department of Defense (2013). Estimates based on 2010 reports indicate that males comprised about 55% of sexual assault victims. Survivors may be reluctant to report and perpetrators may be improperly disciplined due to a military climate often described as supportive of victim-blaming, lacking in accountability, and misogynistic.

College Campuses

Given the average age of traditional college students (18–24) coincides with the age of active criminal offending, coupled with a lack of supervision/guardianship and increased

exposure to consumption of alcohol and illicit drugs, rape/sexual assault is an identified issue that many colleges are required to address. Originally passed in 1990, the *Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act* is a federal law requiring colleges and universities across the United States to disclose information about crime on and around their campuses. The U.S. Department of Education hosts a Web site that allows the public to retrieve information on the incidents of crime reported to police for single colleges or group of colleges. The Clery Act has been amended several times (1992, 1998, 2000, and 2008). Most recently in the reauthorization of the Violence Against Women Act, the Campus Sexual Violence Elimination Act (Campus SaVE) was approved, which amends the Jeanne Clery Act and affords additional rights to campus victims of sexual violence, dating violence, domestic violence, and stalking.

Similar to reports of sexual assault in the broader community, the number of sexual offenses reflected in official reports to undercount its true occurrence. There were fewer than 500 forcible sexual offenses reported on college campuses in the United States in 2011 (UCR, 2012); the estimated number of women enrolled in college in fall 2011 was over 5 million. The NIJ has sponsored three separate surveys to measure the nature and extent of sexual assaults on college campuses (Fisher et al., 2000; Krebs et al., 2007, 2011). The rates of sexual assault varied somewhat across the studies due to the different time frames and populations studied (first-year college freshmen, all students enrolled, historically black colleges and universities). Overall, findings indicated that slightly fewer than 20% of women reported experiencing a completed or attempted sexual assault since entering college, and nearly 5% reported experiencing a forcible sexual assault since entering college.

Correctional Institutions: Adult Prisons/Jails and Juvenile Facilities

The *Prison Rape Elimination Act* (PREA) was federal legislation passed in 2003 in an attempt to combat the growing problem of rape and sexual assault in prisons and jails across the United States. PREA applies to all correctional facilities, including prisons, jails, juvenile facilities, military and Indian country facilities, and Immigration and Customs Enforcement (ICE) facilities. The act established guidelines to prevent rape in custodial confinement. The Bureau of Justice Statistics is required to conduct a comprehensive statistical review and analysis of incidents and effects of prison rape (National PREA Resource Center, 2013).

The Bureau of Justice Statistics (Beck et al., 2013a,b) in a series of reports produced annually provides summaries of the incidents of prison rape reported in more than 7600 adult correctional facilities. For fiscal year ending 30 June 2012, data indicated that 4.0% of state and federal inmates reported one or more incidents of sexual victimization by another inmate or facility staff; about 2.0% (estimated 29 300) state and federal prisoners reported an incident involving another inmate. Among jail inmates, 3.2% of inmates reported one or more incidents of sexual victimization by another inmate or facility staff; about 1.6% of jail inmates (estimated 11 900) reported an incident with another inmate. Rates of victimization were higher among inmates who identified as nonheterosexual. Information on the victimization experiences of juveniles held in adult prisons and jails were included. Nearly 1.8% of juveniles aged 16–17 held in adult prisons and jails reported being victimized by another adult, and 3.2% of juveniles aged 16–17 held in adult prisons and jails reported being victim of sexual misconduct by facility staff.

The Bureau of Justice Statistics reported on the 2011–12 incidents of prison rape that occurred in over 300 juvenile facilities in the United States. Findings indicate that 9.5% of adjudicated youth in custody reported one or more incidents of sexual victimization by another youth or staff. The vast majority (8 out of 10) victims reported more than one incident. The majority of victims reported an incident involving staff, notably female staff. Juvenile victims reported that in about one-third of the cases, facility staff initiated sexual contact but closer to half of the time the victim reported initiating sexual contact. The different methods used to encourage youth participation in sexual contacts included the threat of force, offer of protection, and the promise of drugs and alcohol. Similar to adults victimized in custody, youths who identified themselves as being nonheterosexual had higher rates of victimization.

Conclusion

A significant number of people are victims of sex crimes in their lifetime and current estimates of the prevalence and incidents of such criminal victimization do not accurately reflect how widespread this problem is. It appears from available information that the victims of sex crimes are predominantly female, the perpetrators are predominately male, and the victims and perpetrators are known to each other. There is no single cause of sex offending and explanations vary from focusing on the individual-level pathology of offenders to societal-level explanations noting patriarchal social structure. While adult men are considered the primary perpetrators of sex crimes, recent research has noted that women are involved in a minority of sex offending (often as co-offenders with men and typically when the victim is a family member) and that adolescents commit a significant amount of sex crimes against their peers or younger children. Sexual victimization of persons in the armed forces, on college campuses, and in correctional facilities has been the focus of much attention and legislation has been enacted to prevent and enhance responses to such sexual victimization.

See also: Pornography; Rape and Sexual Coercion; Repetitive Sex Offenders, Treatment of.

Bibliography

Albanese, J., 2007. NIJ Special Report: Commercial Sexual Exploitation: What Do We Know and What Do We Do about It? U.S. Department of Justice, Washington, DC.

Ballantine, M., 2012. Sibling sexual abuse – uncovering the secret. Social Work Today 12, 18.

Beck, A., Berzofsky, M., Caspar, R., Krebs, C., 2013a. Sexual Victimization in Prisons and Jails Reported by Inmates, 2011–2012. Bureau of Justice Statistics, Washington, DC.

Beck, A., Cantor, D., Hartge, J., Smith, T., 2013b. Sexual Victimization in Juvenile Facilities Reported by Youth, 2012. Bureau of Justice Statistics, Washington, DC.

Brownmiller, S., 1975. Against Our Will: Men, Women, and Rape. Fawcett Books, New York.

Bunting, L., 2007. Dealing with a problem that doesn't exist? Professional responses to female perpetrated child sexual abuse. Child Abuse Review 16 (4), 252–267.

Centers for Disease Control, 2011. National Intimate Partner and Sexual Violence Survey. Centers for Disease Control, Atlanta, GA.

Cohen, L., Galynker, I., 2009. Psychopathology and personality traits of pedophiles: issues for diagnosis and treatment. Psychiatric Times, p. 1. http://www.psychiatrictimes.com/display/article/10168/1420331.

Curtis, R., Terry, K., Dank, M., 2008. The Commercial Sexual Exploitation of Children in New York City. Center for Court Innovation, New York. http://www.courtinnovation.org/sites/default/files/CSEC_NYC_Executive_Summary.pdf.

Estes, R., Weiner, N., 2001. The Commercial Sexual Exploitation of Children in the U.S., Canada, and Mexico. University of Pennsylvania, Philadelphia, PA.

Friedrich, W., Luecke, W., 1988. Young school-age sexually aggressive children. Professional Psychology: Research and Practice 2, 155–164.

Finkelhor, D., 1978. Psychological, cultural and family factors in incest and family sexual abuse. Journal of Marital and Family Therapy 4, 41.

Finkelhor, D., 1980. Sex among siblings: a survey on prevalence, variety, and effects. Archives of Sexual Behavior 9, 171.

Finkelhor, D., 1994. The international epidemiology of child sexual abuse. Child Abuse and Neglect 18, 409.

Finkelhor, D., Ormrod, R., 2000. Characteristics of Crimes against Juveniles. U.S. Department of Justice, Washington, DC.

Finkelhor, D., Ormrod, R., 2004a. Child Pornography: Patterns from NIBRS. U.S. Department of Justice, Washington, DC.

Finkelhor, D., Ormrod, R., 2004b. Prostitution of Juveniles: Patterns from NIBRS. U.S. Department of Justice, Washington, DC.

Finkelhor, D., Ormrod, R., Chaffin, M., 2009a. Juveniles Who Commit Sex Offenses against Minors. U.S. Department of Justice, Washington, DC.

Finkelhor, D., Turner, H., Ormrod, R., Hamby, S., Kracke, K., 2009b. Children's Exposure to Violence: A Comprehensive Survey. U.S. Department of Justice, Washington, DC.

Fisher, B., Cullen, F., Turner, M., 2000. Sexual Victimization of College Women. Criminal Justice Abstracts, EBSCOhost, n.p. (viewed 22.07.13.).

Fisher, N., Pina, A., 2013. An overview of the literature on female-perpetrated adult male sexual victimization. Aggression and Violent Behavior 18 (1), 54–61.

Flowers, R., 2001. The sex trade industry's worldwide exploitation of children. The Annals of the American Academy of Political and Social Science 575, 147.

Flowers, R., 2006. Sex Crimes: Perpetrators, Predators, Prostitutes, and Victims, second ed. Charles C. Thomas Publishers, Springfield, Illinois.

Gerardin, P., Thibaut, F., 2004. Epidemiology and treatment of juvenile sexual offending. Pediatric Drugs 6 (2), 79–91.

Hall, R., Hall, R., 2007. A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. Mayo Clinic Proceedings 4 (82), 457.

Harvard Medical School, 2010. Pessimism about pedophilia. Harvard Mental Health Letter 1, 27.

Hillman, E., 2009. Front and center: sexual violence in U.S. military law. Politics & Society 37 (1), 101–129.

Kara, S., 2009. Sex Trafficking: Inside the Business of Modern Slavery. Columbia University Press, New York.

Krebs, C., Lindquist, C., Warner, T., Fisher, B., Martin, S., 2007. Campus Sexual Assault (CSA) Study, Final Report. National Institute of Justice, Washington, DC.

Krebs, C., Lindquist, C., Barrick, K., 2011. Historically Black College and University Campus Sexual Assault (HBCU-CSA) Study. National Institute of Justice, Washington, DC.

Markon, J., 23 September 2007. Human Trafficking Evokes Outrage, Little Evidence. Washington Post p. A01.

Meyer, J., 2000. Root Causes of Sexual Assault. Colorado Coalition Against Sexual Assault. Available: http://www.janetmeyer.com/articles/SexualAssault/Root_Causes%20of%20Sex%20Assault.pdf.

Monzini, P., 2004. Trafficking women and girls and the involvement of organized crime in western and central Europe. International Review of Victimology 11, 73.

National PREA Resource Center, 2013. Prison Rape Elimination Act. Available: http://www.prearesourcecenter.org/about/prison-rape-elimination-act-prea.

Oliver, B., 2007. Preventing female-perpetrated sexual abuse. Trauma, Violence and Abuse 8 (1), 19–32.

Pereda, N., Guilera, G., Forns, M., Gómez-Benito, J., 2009. The prevalence of child sexual abuse in community and student samples: a meta-analysis. Clinical Psychology Review 29, 328–338.

Planty, M., Langton, L., Krebs, C., Berzofsky, M., Smiley-McDonald, H., 2013. Female Victims of Sexual Violence, 1994–2010. U.S. Department of Justice, Washington, DC.

Richards, K., 2011. Misperceptions about sex offenders. Trends and Issues in Crime and Criminal Justice 1, 429.

Ryan, G., Lane, S., Davis, J., Isaac, C., 1987. Juvenile sex offenders: development and correction. Child Abuse and Neglect 11, 385–395.

Ryan, G., Miyoshi, T., Metzner, J., Krugman, R., Fryer, G., 1996. Trends in a national sample of sexually abusive youths. Journal of the American Academy of Child and Adolescent Psychiatry 33, 17–25.

Sandler, J., Freeman, N., 2011. Female sex offenders and the criminal justice system: a comparison of arrests and outcomes. Journal of Sexual Aggression 17 (1), 61–76.

Snyder, H., 2012. Patterns and Trends: Arrests in the United States, 1990–2010. Bureau of Justice Statistics, Washington, DC.

Troup-Leasure, K., Snyder, H., 2005. 'Statutory Rape Known to Law Enforcement', Juvenile Justice Bulletin. U.S. Department of Justice, Washington, DC.

Truman, J., Planty, M., 2012. Criminal Victimization, 2011. U.S. Department of Justice, Washington, DC.

Uniform Crime Report, 2012. Crime in the United States: Forcible Rape. http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/violent-crime/rapemain_final.pdf.

U.S. Department of Defense, 2013. Sexual Assault Prevention and Response Office. Annual Report 2012. Available from: http://www.sapr.mil/.

U.S. Department of Health and Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau, 2012. Child Maltreatment 2011. Available from: http://www.acf.hhs.gov/programs/cb/research-data-technology/statistics-research/child-maltreatment.

U.S. Department of State, 2005. The Facts about Human Trafficking for Forced Labor. Available from: http://2001-2009.state.gov/documents/organization/51174.pdf (retrieved 14.11.06.).

U.S. Department of State, 2006. Victims of Trafficking and Violence Protection Act of 2000: Trafficking in Persons Report. U.S. Government Printing Office, Washington, DC.

Veneziano, C., Veneziano, L., 2002. Adolescent sex offenders: a review of the literature. Trauma, Violence & Abuse 3, 247–260.

Wheaton, E., Schauer, E., Galli, T., 2010. Economics of human trafficking. International Migration 48 (4), 114–141.

White, N., Lauritsen, J., 2012. Violent Crime against Youth, 1994–2010. U.S. Department of Justice, Washington, DC.

Wolak, J., Finkelhor, D., Mitchell, K., Ybarra, M., 2008. Online "predators" and their victims: myths, realities, and implications for prevention and treatment. American Psychologist 63, 111.

Zilney, L., Zilney, L., 2009. Perverts and predators. Rowman and Littlefield Publishers, Inc., Plymouth.

Zimring, F., Piquero, A., Jennings, W., 2007. Sexual delinquency in Racine: does early sex offending predict later sex offending in youth and young adulthood? Criminology and Public Policy 6, 507–534.

Sex Life Challenges: The Finnish Case

Osmo Kontula, Population Research Institute, Family Federation of Finland, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In international comparison, Finland is a country where social well-being and gender equality are exemplarily progressed. One could assume that this had caused positive impact also on sexual well-being. This assumption was tested by utilizing nationally representative sex surveys that have been conducted in Finland during several decades. The results are challenging after there was found increasing lack of female sexual desire, decreasing frequency of sexual intercourse, and unwanted trends in female orgasms. Some explanations for these unexpected results were given.

Introduction

Finland is one of the few countries where there is available nationally representative surveys of sexual activities and values among the adult population. According to many international indicators, social progress is well advanced in Finland. In relation to social well-being, the European Quality of Life Survey gave the highest rates in happiness in Europe for Finland and Denmark. This has some implications for sexual values and activities. Sexual images and values are evolving into a more positive and more liberal approach.

Women have a unique position in Finland in international comparison. The rate of women working full time is in Finland, the highest in Western Europe – partly thanks to extensive public childcare that is a subjective right for everybody. Based on the Gender Equity Index, introduced by Social Watch, Finland is number one in gender equity in the world together with Sweden. In education, the results of the Program for International Students Assessment (PISA) evaluation (60 countries), OECD, Finland has been several times number one in the world. Women outnumber men in higher education and comprehensive sex education is at the top level in Europe (Kontula, 2010). This social and educational progress has created positive circumstances for sexual activities also among aging population (Kontula and Haavio-Mannila, 2009; Kontula, 2013).

During the last decades, there was a major progress also in sexuality issues in Finland. Two nationally representative surveys of the sexual behavior and sexual attitudes in 1971 and 1992 showed that people's attitudes had become more liberal, sexual behavior more equal, women sexually more active, and that especially women's sexual satisfaction had increased during the 20 years between these two surveys. The increasingly more copious, open, and versatile treatment of sexuality in the different media was estimated to have been one of the causes of this positive change (Kontula and Haavio-Mannila, 1995; Kontula and Kosonen, 1996).

Public discourse defines acceptable forms of sexuality and delivers disapproval for 'unacceptable' forms, as well as for deeds that have been deemed illegal. In the West, an individual's right to sexual self-determination now represents an important sexual standard, as long as it does not conflict with someone else's right to sexual self-determination. In our culture, sexual interaction occurs as the outcome of a negotiation between two individuals; society is generally unwilling to interfere with these mutual understandings and agreements. The lines are mostly drawn on a 'just-in-case' basis, not because there is any persuasive evidence of the postulated harms (Kontula, 2009).

Over the past 20 years, the key shift in sexual culture in the West has been the opening up of the private sexual sphere into something that is now part of the public sphere. This is manifested in the public proliferation of images of scantily clad people, intimate stories about well-known personalities, and new technological breakthroughs in pornography (private webcams, annual sex expos, and so on). Sex and nudity are a natural and everyday part of our public culture. At the same time, marriage and cohabitation have lost their power to determine whether a particular sexual relationship is acceptable or not. Sex is for everyone, even though not everyone has an equal opportunity to engage in it.

In many respects sexual trends in Finland before the 2000s correspond with research data compiled previously in Europe on the same topic (Sandfort et al., 1998). The broader shift that has occurred in the West has meant a greater number of sexual partners before forming a committed relationship, lower level of commitment in relationships, increases in masturbation, lifetime partners and parallel relationships, more frequent intercourse, increasing commonness of oral and anal sex, and increased condom use. Finland has deviated from the European trends only in that neither intercourse frequency nor condom use increased simultaneously.

The aim of this article is to present the greatest present-day challenges in sexual life in Finland based on the results of the latest national sex surveys. These challenges are increasing lack of female sexual desire, decreasing frequency of sexual intercourse, and unwanted trends in female orgasms.

Method

The FINSEX study (Kontula, 2009) has tracked Finnish sexual trends since the 1970s up to the present. Some survey questions that map out sexual behavior in youth yield data all the way from the 1940s, a wealth of knowledge that is unique even worldwide.

Four national sex surveys were conducted in Finland in 1971, 1992, 1999, and 2007. Each survey's sample was drawn from the Central Population Register, so that all Finns had an

equal opportunity to be selected into the sample. In these four population surveys, respondents aged 18–54 in 1971 and 18–74 in 1992, 1999, and 2007 were selected. Numbers of respondents were 2152 (1971), 2250 (1992), 1496 (1999), and 2590 (2007).

Response rates were 91% (1971), 76% (1992), 46% (1999), and 43% (2007). The higher response rates in 1971 and 1992 were due to face-to-face interviews at respondents' homes. In 1999 and 2007 the data collection was carried out by Statistics Finland as a mailed survey (because of lower costs), which caused lower response rates.

The impact of lower response rates in the 1999 and 2007 studies on comparability with the 1971 and 1992 surveys has been evaluated particularly by analyzing the ways in which people in a particular birth cohort have responded to the same questions concerning their own youth. The representativeness and comparability of the data in relation to 1992 data remained quite good, except in the case of male respondents over the age of 55. The 1999 and 2007 findings (mailed surveys) provide a slight underestimation of men above the age of 55 (sexual initiation somewhat later, and sexually a bit more monogamous in their life time), compared with the entire age group of the respondents.

More detailed information on the sampling, interviewing, and questionnaires is available in Kontula and Haavio-Mannila (1995), Haavio-Mannila et al. (2001), Haavio-Mannila and Kontula (2003), and Kontula (2009).

Results

Gender Differences in Sexual Desire and Increasing Lack of Female Sexual Desire

The first challenge in sex life in Finland is the major gender difference in sexual desire and increasing numbers in lack of female sexual desire. Compared with Baumeister et al.'s (2004) list of indicators of sexual desire, this study utilized respondents' own assessments regarding sexual desire, preferred frequency of sex, and lack of sexual desire.

In 2007, for the first time, Finns were asked to evaluate their own sexual desire in response to the following question: "How often do you experience sexual desire?" The most typical response among both men and women was that they experienced sexual desire several times a week. In terms of frequency, this amounted to 14 times per month. On average, men experienced sexual desire much more frequently than women, on average 21 times per month (medium value 12), whereas women experienced it on average 9 times per month (medium value 4). Young people, compared with older people, were much more likely to feel desire – men, on average, every day, and women, a few times a week. Nearly one in three young men, but only 5% of young women reported feeling sexual desire more than once a day (Figure 1).

Compared with young and middle-aged women, men in the same age groups experienced sexual desire twice as frequently. For men in the older age group, sexual desire was four times as frequent as for women. These gender differences in sexual desire are supported by earlier international research (Baumeister et al., 2004).

An interesting observation in terms of male and female sexual desire is that the level of desire reported by men is compatible with the level of desire in women approximately 20 years younger, and vice versa. This may explain relationships where an older man has found a much younger spouse. If sex and desire are important in this type of couple formation, the choice may not only be appealing on the basis of feelings, but also on the basis of rational reasoning.

In this way, a young woman may get herself a partner whose sexual desire does not significantly exceed her own. Based on these averages, it would be much more difficult for an aging woman to find a young man whose sexual desires are well suited to her own.

This experience gap between the sexes was also visible in responses to a question about whether or not respondents were satisfied with the frequency of intercourse in their current relationship. Approximately 60% of middle-aged women were satisfied with the frequency of sex in their relationship, but only one-third of men of the same age felt similarly. Two-thirds of men and one-third of women would have liked more frequent intercourse. In other words, middle-aged women had been much more successful than men in satisfying their need for sexual intercourse.

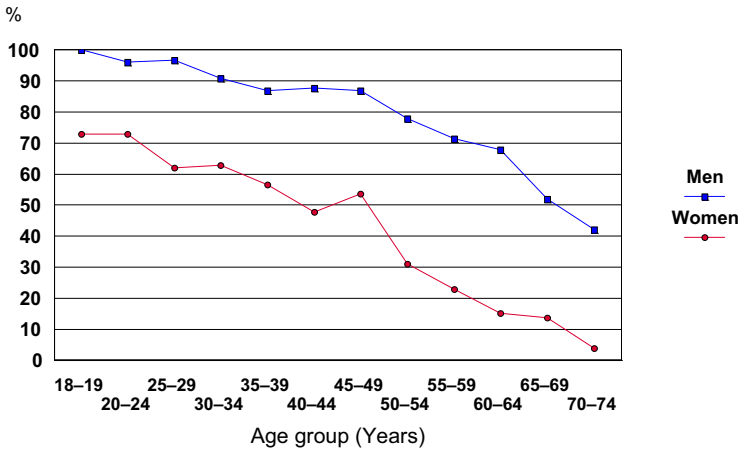

Figure 1 Feels sexual desire at least a few times a week.

The desire for intercourse between married women and men differed to a substantial extent only in relationships that had lasted more than 10 years. In long-term relationships, men were much more likely than women to desire sexual intercourse more frequently. On the other hand, the proportion of men wanting more frequent intercourse was highest in relationships that had lasted slightly more or less than 10 years, and highest among women in relationships that had lasted slightly less than 10 years. To the extent that sexual dissatisfaction is an important factor in explaining marital infidelity, this stage of marriage is a particularly critical period.

In 1992, 1999, and 2007, the FINSEX survey included a question asking respondents whether they had had a problem with lack of sexual desire in themselves or in a partner within the last year. Lack of desire was defined here as a problem, if a respondent reported that the issue arose at least fairly frequently, either in the case of the respondent him or herself or in the partner. The introduction to the question specifically addressed the incidence of various sexual problems (Figures 2 and 3).

Lack of own sexual desire was significantly more common among women than men. In 2007, 41% of women and 12% of men had experienced lack of sexual desire at least fairly frequently. The reported prevalence of lack of desire increased among men and women between 1992 and 2007, but only among women did the trend continue also from 1999 until 2007. As a result, the desire gap between women and men has grown even more.

In a number of international studies, the proportion of people who have no interest in intercourse has fluctuated between 10 and 15%. In the present study as well, 14% of women reported experiencing lack of sexual desire very frequently.

Men reported a partner's lack of sexual desire significantly less frequently than women themselves reported feeling lack of desire. It seems quite likely that men were not always fully aware of the lack of desire on the part of their partners, who had been able to conceal it from them. In this way, women were protecting their own sexual image in the eyes of a male partner and had avoided hurting his feelings with repeated complaints about not feeling desirous.

The Declining Frequency of Intercourse

Second challenge in sex life in Finland is a declining frequency of sexual intercourse. Of all men, 61% and of all women, 53% had had sex within the last week. The gender difference here stems from the greater number of aging female widows who no

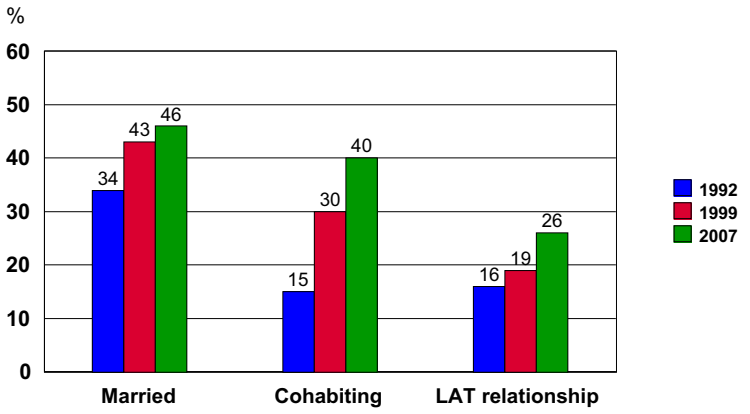

Figure 2 Lack of own sexual desire somewhat frequently in relationship during the last year, women.

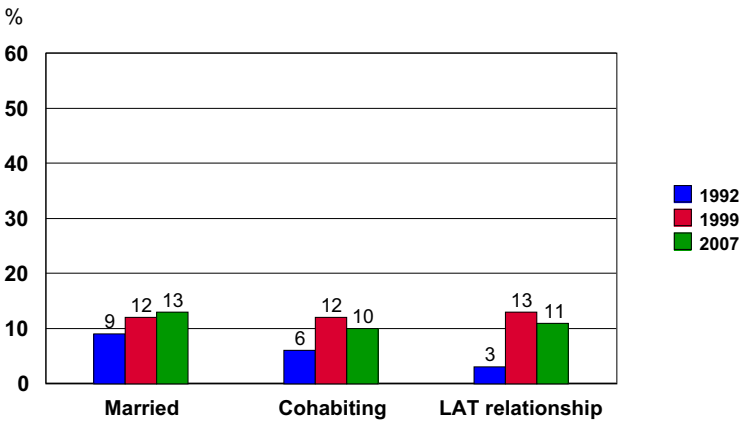

Figure 3 Lack of own sexual desire somewhat frequently in relationship during the last year, men.

longer had the opportunity or desire for a sexual relationship. Compared with figures from 1999, the numbers had declined by several percentage points, and compared with the early 1990s and 1970s, the decline was about 10 percentage points. In other words, fewer Finnish people reported that they had had sex within the last week.

Sixty-six percent of young men and 62% of young women had had sex during the last week. Among middle-aged respondents the proportions were 65 and 60%, respectively, and among the oldest age group 53 and 35%, respectively. Thus, with the exception of the oldest group, approximately two in three Finnish people had had sex at least once the week before. The numbers for middle-aged respondents had fallen by nearly 10% from the 1999 results and approximately 15% from 1992. The decline of sexual intercourse was rather significant among middle-aged respondents.

When the frequency of intercourse reported by Finns is converted into numbers, it turns out that the average man was having sexual intercourse 4.8 times per month, the average woman 4.3 times per month. When also taking into account the median frequency of sexual intercourse, it is further revealed that half of all men and women had sexual intercourse at most once every 2 weeks. Based on this, Finnish people do not have very active sex lives (Figures 4 and 5).

The changes in frequency of sexual intercourse over the 8 years are thought provoking. In 2007, both men and women were having less sexual intercourse by an average of once per month than in 1999. Put another way, the frequency of sexual intercourse among Finns had declined by approximately 20%. Frequencies were down for young and middle-aged people compared with all previous years. Only among the oldest age group did the frequency rise from 1992 (no data is available for 1971), but even those levels declined again after 1999.

The frequency of intercourse declined in all relationship types. In 2007, married men were having sexual intercourse 4.6 and women 4.3 times per month. This amounts to a decrease of one session of intercourse per month. Cohabiting respondents were having sexual intercourse approximately six times per month, a veritable crash from previous figures; a 25% drop, translated to two events of sexual intercourse. Living-together relationships had the highest frequency of intercourse, more than seven times per month, or nearly twice per week. Even this figure had declined by one. The frequency among single people was very irregular and averaged less than

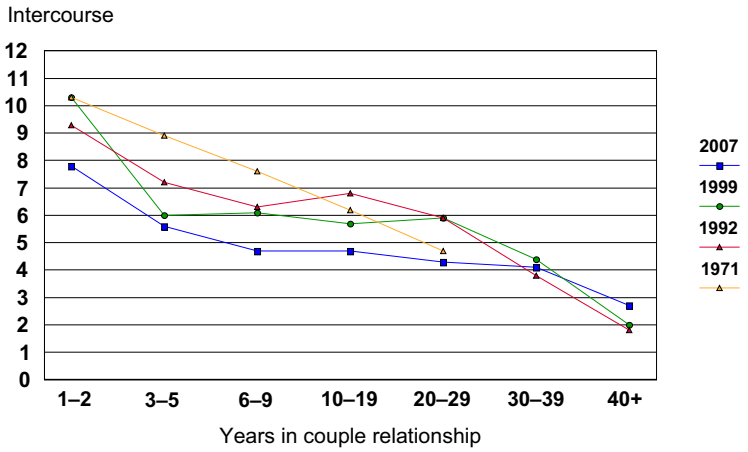

Figure 4 Number of intercourse in month, women.

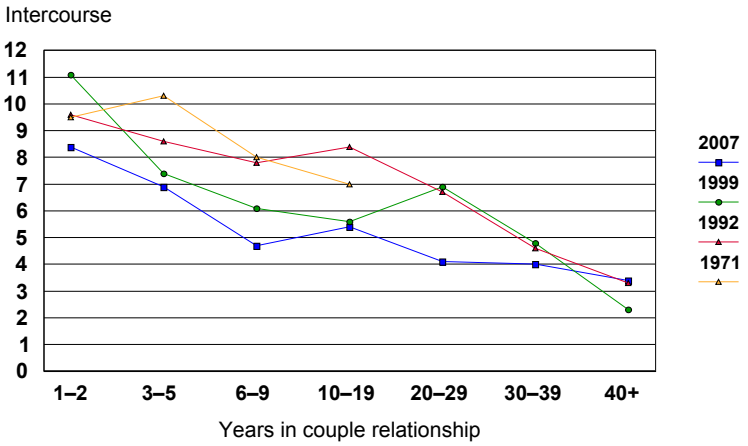

Figure 5 Number of intercourse in month, men.

once a month. Even that represents a slight drop. The decrease in intercourse frequency has affected every category of relationship in Finland.

A whole year or more had passed since the last time that 10% of men and 14% of women had had sex. Among the oldest age group, the figures were 17 and 32%, respectively. These people were usually single-dwellers or single, of whom two-thirds of men and half of women had been without sexual intercourse for more than 1 year. At the same time, there were also marriages where factors like illness or injury had caused the cessation of sex. Of married men and women, 8% had had no sex in the last year.

Sexual Pleasure and Decreasing Frequency of Orgasms

A third challenge in sex life in Finland are declining trends in sexual satisfaction and orgasms. Ninety-four percent of men reported having an orgasm always or almost every time they had intercourse, and one in two reported having an orgasm every single time. Only 5% of men had difficulty having orgasms. A few percent rarely experienced an orgasm when having intercourse. Orgasms are not a given even to all men.

For women, having an orgasm from intercourse was much less guaranteed than for men. Forty-six percent of women said that they always or nearly always had an orgasm when having intercourse, but only 6% always had one. Nearly one-fifth of women had an orgasm approximately half of the time and about one-third had one fairly infrequently at most.

Based on the study's findings orgasticity was unfortunately headed in a worse direction over the 8 years following 1999. This applies to the responses of both men and women, young and middle aged (Figure 6).

In 2007, 9% of women reported never having had an orgasm from intercourse. Difficulties experiencing orgasms affect a large proportion of women. In the earlier studies, the proportion of women who had never experienced an orgasm from sexual intercourse was 5%, lower than in 2007.

There is much to think about and wonder in only 42% of young women (53% still in 1999) reporting that they usually had an orgasm during intercourse, whereas 40% said that they had them fairly infrequently at most.

According to the 2007 findings, 13% of young women had never had an orgasm from intercourse. Fifty-one percent of middle-aged women usually had an orgasm during intercourse (60% in 1999). One-third of middle-aged women and slightly more than one-third of older women had an orgasm fairly infrequently at most. Difficulties having an orgasm among middle-aged women had become as common as it had been in the early 1970s, after a period of temporary improvement. The situation among young women was clearly even worse than it had been in the early 1970s.

The proportion of people who had an orgasm the last time they had intercourse was close to the proportion of those who said that they generally had an orgasm from intercourse. Of men, 92% and of women, 55% had an orgasm the last time they had intercourse. It speaks to the confusion surrounding the concept of 'orgasm' that nearly one in ten women was not able to say whether or not she had had an orgasm the last time she had intercourse.

The findings indicate that women differ from one another greatly in terms of their tendency or capacity to experience orgasms. A significant portion of women experienced persistent problems having orgasms from intercourse, whereas many (12%) found it easy to have multiple orgasms. The inequality in sexual enjoyment among women was drastically greater than among men. It is intriguing why women are experiencing greater, not fewer, problems with orgasms, even though the opportunities for sexual enjoyment now seem better than ever.

Discussion of Sexual Challenges in Finland

It is a great challenge to try to explain why in Finland, where gender equality and social welfare are highly progressed in international comparison, the lack of sexual desire is increasing, frequencies of sexual intercourse are decreasing especially in the middle age, and difficulties to experience orgasm in sexual intercourse are increasing. It was assumed that sexual well-being should improve along the progress of other social well-being.

Sexual desire is a key factor distinguishing men from women in sexuality issues. According to a number of measures,

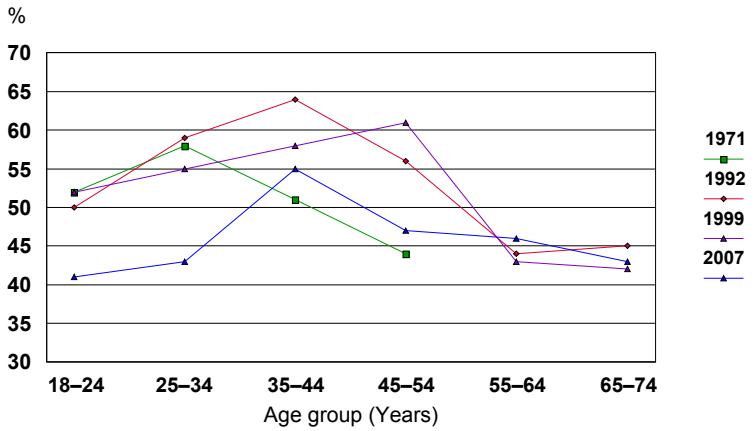

Figure 6 Orgasm during sexual intercourse almost always or usually women.

male sexual desire is manifested at least twice as often as female desire, and men would like to have sex twice as often as women. In every age group, the frequency of sexual intercourse that would have sufficed to satisfy male respondents was higher than for female respondents. An additional dimension comes from the repeated lack of sexual desire reported much more frequently by women than men. On the basis of available data it is not possible to make the claim that the gap between male and female desire would be shrinking in Finland since the 1970s. All of these factors have naturally had practical consequences for sexual interaction between men and women. Need for negotiations and different compromises in love making are increasing.

The findings reveal that women were able to fulfill their desires for sexual intercourse more successfully than men, having the frequency of sex in the relationship that they wanted. This suggests that women were more frequently gatekeepers and thus better able to decide when and how often the couple had sex. This role model continues to persist in Finnish relationships.

A particular challenge lies in solving the problem of the significant increase in women's loss of sexual desire. It is not a case of decreasing desire in long-term relationships, of becoming fed up and bored. The issue of libido loss had also expanded to the relationships of young adults as well as new relationships.

There is an intriguing association behind loss of sexual desire. When the incidence of lack of desire was analyzed in the different surveys by actual frequency of sexual intercourse, it seems that lack of desire did not vary by frequency category. What does this mean? It means that if the actual frequency of sexual intercourse reported by women had not dropped so much in recent years, the incidence of lack of desire among women apparently would not have been any greater than previously. Lack of intercourse decreases the desire to have intercourse. This suggests that female sexuality is much like an unused machine – it must first be oiled before it is functional again.

The motivation to have sex was not affected by any deterioration in the quality of intercourse, except on the part of orgasms, and did not cause people to invest less in intercourse. In 2007, the same proportion of women as in earlier surveys had found their latest sexual intercourse very enjoyable, and the proportion of men who did had actually grown.

Why do men have higher sexual desire than women? One explanation could be based on the social exchange theory (Sprecher, 1998). Within this theoretical framework one could think that the greater rewards that men achieve from sexual interaction could explain their stronger sexual desires. This might make sense in that men experience orgasms so much more often and easily in intercourse compared with women. In other words, men might be more motivated to seek intercourse as something that offers them particular rewards. This viewpoint is supported also by the finding that women who enjoyed sexual intercourse, and got more pleasure out of their sex lives than other women, were also less likely to experience lack of sexual desire.

This theory of 'pursuing rewards' still does not explain the differences between men and women in a comprehensive way, because even women who always had orgasms during intercourse desired to have sex somewhat less frequently than similar men. This perspective nevertheless opens up the possibility of matching male and female desire to suit one another better. If women were to enjoy intercourse more and have orgasms more regularly, the desire gap between the genders should decline.

A surprising and thought-provoking finding and challenge in the study is the decline in intercourse frequency in the 2000s. For example, from 8 years before, there was a 10–15% drop in respondents of all ages who reported that they had had sex within the last week. Only 45% of men and 40% of women had had sexual intercourse in the last week. In terms of frequency figures, Finnish respondents had lost one session of sexual intercourse per month, or 20%, compared with figures 8 years prior. Also compared to the early 1970s, people were having less sex now. Frequency of sexual intercourse declined most in the age group of 25–40-year-olds. The frequency had remained the same for people aged over 50. What could explain the change?

The decline in intercourse frequency that has taken place in Finland is not entirely unique. For example, researchers in the United States and Germany found already in the 1990s that when measured by intercourse frequency, people were not very sexually active. The finding conflicts with the media images that depict passionate, tempestuous sex lives, and the worry about the over-sexualization of our lives. The sex lives of most real people are somewhat modest.

The most typical wish expressed by Finnish respondents was to have sexual intercourse twice a week. People who had achieved this rate were also generally happy with the amount of sex they were having. Most respondents came up short, having only half the sex they would have liked. For this reason, more and more Finns expressed dissatisfaction with their sex life as a whole. Quality was no substitute for quantity in this case. The study's results do not support the claim that the frequency of sexual intercourse is not important from the standpoint of sexual happiness. Certainly, there will be individual exceptions to this.

When measured through the frequency at which Finns had intercourse, their sex lives seemed on their way to growing passivity. The picture looked different when using another measure of sexual activity on a par with sexual intercourse – masturbation. When looking at how many men and women of various ages had been sexually active (sexual intercourse and/or masturbation) within the last week, it turned out that, in 2007, Finnish men and women were approximately as sexually active as they had been according to the two previous surveys conducted in the 1990s. Compared with middle-aged respondents in the 1970s, people were clearly even more sexually active. Further, respondents aged 55–74 years proved to be more sexually active in the last two surveys, compared with the early 1990s.

One of the study's key findings was that sexual activity among Finnish people was shifting to a significant degree from sexual intercourse to masturbation, also among couples. Naturally, this invites questions as to why it has happened. At least in those relationships where both partners are busy with work, possibly on different work schedules, this may be the case. In part, though, masturbation serves to compensate for the decrease in intercourse. The scope is not very broad, based

on the finding that masturbation was already increasing fast at the time when the frequency of intercourse in relationships had not yet significantly changed.

There are those who see the proliferation of masturbation as a sign of sexual emancipation. Masturbation is autonomous sex and does not require a demanding level of interaction with a partner. It's easy and presents no performance anxiety. Masturbation is suited to busy lifestyles and a culture of interaction that demands ever higher levels of sexual performance. As sexual gratification is nevertheless becoming increasingly important to people, the role of masturbation inevitably grows, as indicated by the findings of this study.

Slowly, people have begun to emancipate themselves from their feelings-based relationships, moving from shared sex toward a more individualistic experience of sexual pleasure – masturbation or self-pleasuring. Masturbation has its own justification, and it is increasingly practiced alongside established relationships. With its help, some people are able to escape the high demands imposed by our modern sexual etiquette, with its joint negotiations and situational considerations of the appropriateness of taking sexual initiative. For these people, masturbation is not a substitute for too little sex, but a different and easier form of sexual pleasure. To some extent, this is also a manifestation of a partial distancing between sex and love.

Social and public discourse on the subjects of gender and sexuality have for some time underscored women's right to sexual enjoyment. This enjoyment is often linked to the positive context of a couple relationship and the promotion of sexual well-being. At the same time, public discourse loads additional expectations on men to offer their female partners proper sexual fulfillment. Men's obligation to please their spouse is already paralleled by their earlier obligation to bear responsibility for the family's livelihood.

The great mystery of Finnish sexuality is why difficulty having orgasms has increased in the twenty-first century, at a time when information about how to better achieve orgasms proliferates. Women's and health magazines are full of instructions for the pursuit and cultivation of sexual pleasure and sexuality education has improved substantially. Women's sexual rights have been acknowledged already for a long time.

The proportion of young and middle-aged women who had orgasms through sexual intercourse nearly always or most of the time had fallen by 10 percentage points, and even among men, the proportion of men who always reach orgasm through intercourse had dropped by 15%. There was up to a 20% drop in cohabiting women who had orgasms through intercourse most of the time. In young women, this was associated with the clearly increasing difficulties of having orgasms in the first years of a relationship. Hence, fewer young women considered their sexual relations 'highly enjoyable.'

Women's responses indicated that they experienced more orgasms when they felt that the sex was initiated together with the partner. This may represent one key reason for why women are finding it more difficult to have orgasms. Among young women in particular, more and more felt that the latest sexual intercourse had not been initiated together, but that the impetus had come solely from the partner. Of course, this also speaks to the decline in the willingness of these women to initiate sex to their partner. Add to this young women's growing difficulties becoming aroused and the decline in the frequency of sexual intercourse, and we are approaching the core that reflects toward both sexual desire and the ability to have orgasms. When someone is not fully engaged in love making, it is no surprise that the enjoyment it brings is not what it used to be.

The persistent risk discourse may have created a situation in which women increasingly view sexual interaction through a more rational lens, rather than casting themselves body and soul into enjoying sexual experiences with a partner and realizing their own desires. Excessive rationalism is the biggest enemy of orgasms. Simply put, thinking does alight desire, but orgasms come when thinking ceases. The inability to implement this formula may be a key cause for the trouble that young women especially are increasingly having with orgasms.

The image that emerges from this study may bring solace to many who can now see that numerous others are just as inactive in bed as themselves, and that the sex other people are having is not that exotic. The label of postmodernism that accompanies sex in social discourse is absent in people's real, everyday lives.

See also: Heterosexuality; Marital Quality and Health; Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sexual Dysfunction, Epidemiology of; Sexual Motivation; Sexuality Over the Life Course.

Bibliography

Baumeister, R.F., Catanese, K.R., Vohs, K.D., 2004. Is there a gender difference in strength of sex drive? Theoretical views, conceptual distinctions, and a review of relevant evidence. Personality and Social Psychology Review 5, 242–273.

Haavio-Mannila, E., Kontula, O., Kuusi, E., 2001. Trends in Sexual Life: Measured by National Sex Surveys in Finland in 1971, 1992 and 1999 and a Comparison to a Sex Survey in St. Petersburg in 1996. Working Papers E 10/2001. The Population Research Institute. The Family Federation of Finland, Helsinki.

Haavio-Mannila, E., Kontula, O., 2003. Single and double sexual standards in Finland, Estonia and St. Petersburg. Journal of Sex Research 40, 36–49.

Kontula, O., Kosonen, K., 1996. Sexuality changing from privacy to the open – a study of the Finnish press over the years from 1961 to 1991. Nordisk Sexologi 14, 34–47.

Kontula, O., 2009. Between Sexual Desire and Reality: The Evolution of Sex in Finland. The Population Research Institute D49/2009. The Family Federation of Finland, Helsinki.

Kontula, O., Haavio-Mannila, E., 2009. The impact of aging on human sexual activity and sexual desire. The Journal of Sex Research 46, 46–56.

Kontula, O., 2010. The evolution of sex education and student's sexual knowledge in Finland in the 2000s. Sex Education 10, 373–386.

Kontula, O., 2013. Sexuality among older adults. In: Baumle, A.K. (Ed.), International Handbook on the Demography of Sexuality, International Handbooks of Population, vol. 5. Springer, pp. 195–214.

Kontula, O., Haavio-Mannila, E., 1995. Sexual Pleasures: Enhancement of Sex Life in Finland, 1971–1992. Hampshire, Dartmouth, Aldershot.

Sandfort, T., Hubert, M., Bajos, N., Bos, H., 1998. Sexual behaviour and HIV risk: common patterns and differences between European countries. In: Hubert, M., Bajos, N., Sandfort, T. (Eds.), Sexual Behaviour and HIV/AIDS in Europe: Comparisons of National Surveys. UCL Press, London, pp. 403–426.

Sprecher, S., 1998. Social exchange theories and sexuality. The Journal of Sex Research 35, 32–43.

Sex Offenders

Malcolm Cowburn, Sheffield Hallam University, Sheffield, UK
Steve Myers, University of Salford, Salford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work with sex offenders operates at the intersections of sociology, psychology, and criminology. While key criminal justice and corrections agencies operate with convicted offenders, social work(ers) may be involved with unconvicted people who have behaved in a sexually harmful manner. Popular conceptions of the sex offender embodied in media reportage provide a distorted account. Psychological and sociological theories relating to the etiology of sex offending are reviewed and an integrated public health approach in which social work plays an important role is outlined.

Introduction

Social work practice with sex offenders is informed by both psychological and sociological research and theory. Psychological approaches focus on working with individual offenders whereas sociological perspectives locate sexual offences and sex offenders within a wider social context. This article considers both approaches and how they influence social work with sex offenders.

For the purposes of this article, the International Association of Schools of Social Work (IASSW) (2001) and International Federation of Social Workers (2012) definition of social work is adopted:

> The social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilizing theories of human behavior and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work.
> http://www.iassw-aiets.org/international-definition-of-social-work

Applied to social work with sex offenders, this definition moves beyond the individual into a social and environmental context and it employs both psychological and sociological theories in its interventions with an emphasis on values. These aspects are referred to throughout this article. However, knowledge about people who commit sex crimes and the nature of their crimes is essential and Lancaster (1997) argues that teaching about sex offenders should form a part of every social work curriculum. She notes:

> The reality of child protection responsibilities means that social workers are inevitably brought into contact with people who have committed sexual offences, not just those who are offended against. A much more proactive and truly preventative stance could be developed by encouraging those at pre-qualifying stage of their training to engage with the difficulties of working with perpetrators of sexual abuse.
> (Lancaster, 1997: p. 93)

Lancaster points to the importance of understanding sex offenders for those working with victim-survivors: it provides insight into the victim-survivor's experience and sensitizes the social worker to potentially abusive scenarios within families and beyond. Building on the work of Lancaster, Myers (2008) agrees that knowledge about sex offenders is important for social workers and it is rapidly changing. He considers that knowledge about sex crimes and the people that commit them has a wider relevance in social work education beyond the mere subject matter, being particularly important in aiding reflection on the emotional impact of the work in relation to professional values and also in developing 'epistemic reflexivity.' In two exhaustive reviews of research, policy, and practice, Grady (2009a,b) highlights the central place that social workers in the United States play in working with sex offenders who may have a mental health diagnosis. Moreover, Grady and Abramson (2011: p. 441) note that the "likelihood of social workers in all practice settings finding themselves working with some aspect of sexual offending is a product of increasing sexual violence, which the World Health Organization (WHO) has described as a global crisis" (Krug et al., 2002). In their study of postgraduate social work education in the United States they point to "the increasing prospect that new practitioners, whatever their practice setting, will be called on to work with sexual offenders, offenders' families, or the survivors of sexual offenses" (Grady and Abramson, 2011: p. 441).

It is clearly important that social workers have knowledge and awareness of sex offending, therapeutic programs, and the values underpinning practice. The remainder of this article introduces these issues, but before doing so clarifies issues of terminology.

Definitions and Terminologies

> … terminology … is not simply a neutral statement of the situation, but reflects particular ideological and knowledge claims about the nature of the problem, its causes and consequent practices to deal with it.
> (Myers, 2008: p. 208)

Sex offenders are conventionally defined as people convicted of a sexual offense. In considering social work with sex offenders this definition is inadequate because only a small proportion of people charged with a sexual offense are convicted of that offense (Percy and Mayhew, 1997). Moreover, most people who

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28082-3

are sexually harmed by others do not report the incident(s), thus people who are convicted of sexual offenses are likely to represent the 'tip' of a far larger 'iceberg' (population) of people who sexually harm others. Percy and Mayhew (1997) estimate that the number of sex crimes committed is likely to be 15 times greater than the number reported to authorities. This distinction is important because, according to how social work is constituted and organized in different parts of the world, it is likely that social workers will be aware of unreported sex crimes and working with people who are sexually harming others but are not convicted of sexual offenses.

Moreover, across the globe legal definitions and penal responses to sex crimes differ considerably. Nelken (2009) points to the dangers in comparing systems of criminal justice by adopting either an ethnocentric approach or a relativist stance. He highlights (p. 294) the "Anglo-American tendency to assume what others do in foreign places and foreign languages is less important, and that they too are bound to come in line eventually," whereas a relativist approach considers comparison as a futile activity asserting that legal systems and criminal justice approaches only have relevance to the time and place where they operate. Jehle (2012) outlines the difficulties in studying attrition rates for reported sex crimes across Europe and in presenting credible comparative data.

Apart from the technical issues of definition there is a wider issue related to how professional discourses construct what they define. Social work is at the intersection of a number of very powerful professions; particularly psychology and psychiatry, where disciplinary discourses can obscure the fact, the people who commit sex offences are people. Writing of a commonly used word in the UK, Myers notes:

> 'Paedophile' is a particularly pernicious word in social work, which cannot fail to be used in a derogatory way *to reduce someone to their abusive actions*, thus locating the problem behaviour in deviance and psychopathology.
> (Myers, 2008: p. 208 – emphasis added)

This article is written in ways that recognize the humanity of people who sexually harm others. While the harmful acts that they commit are inhumane, these offenders are human. The point that Myers makes concerning the power of professional discourses to separate the sex offender from humanity is one that will be returned to later.

Understanding Sex Offenders: Moral Panics and Research Findings

The most common source of information about sex offenders is the news media (Jewkes, 2004) and much has been written about how the media represent the sex offender as a 'folk devil' whose particular threat creates a 'moral panic.' A folk devil is

> 'A condition, episode, person or group of persons [that] … become defined as a threat to societal values and interests …' and a moral panic presents the folk devil in 'a stylised and stereotypical fashion by the mass media; the moral barricades are manned by editors, bishops, politicians and other right thinking people.'
> (Cohen, 1972: p. 9)

Since 1991 there has been sustained criminological research into the construction of the sex offender as a 'folk devil' in the news media. There are four significant aspects to how the sex offender and sex crime are represented:

1. A stranger in a public place commits the sex crime. By implication people who are known to victims are unlikely to attack them and thus public space is dangerous and private/domestic space (i.e., the home/family setting) is safe.
2. The sex offender is a prolific offender who is offending and being reconvicted frequently.
3. There is no specific consideration of gendered issues.
4. There is little recognition of sex offending as a premeditated process.

Given the power of these representations, it is important to review research to explore their validity.

Stranger Danger: Dangerous Public Space Safe Domestic Space

In 2007, following a review of the research literature, the British Home Office (2007: p. 5) noted that 'the vast majority (at least 75%) of child sex offenders are known to their victims.' In relation to rape, Myhill and Allen (2002) found in their study of the 1998 and 2000 British Crime Surveys that 92% of women who were raped were attacked by someone they knew.

The Sex Offender as Prolific Offender

The view that most sex offenders invariably reoffend has regularly been contradicted by research findings. For example, the Minnesota Department of Corrections (2007: p. 7) note: "Contrary to popular opinion, sex offenders do not always recidivate. In fact, research has consistently shown that recidivism rates are relatively low for sex offenders." From their own data, they found that after 3 years, 7% of the 3166 offenders had been rearrested for a sex offense, 6% reconvicted, and 3% reincarcerated. By the end of the follow-up period (an average of 8.4 years for all 3166 offenders), 12% had been rearrested for a sex offense, 10% reconvicted, and 7% reincarcerated (Minnesota Department of Corrections, 2007: p. 2). Similarly, Harris and Hanson (2004: p. 7) using data from 10 follow-up studies of adult male sex offenders ($n = 4724$) examined 'sexual recidivism,' defined as new charges or convictions for sexual offences. They found that most sex offenders did not repeat sex offences, and that first-time sex offenders were significantly less likely to repeat sex offences than those with previous convictions for sex offences (19% vs 37% after 15 years). These rates are repeatedly found in studies and are strikingly low when compared to other types of offense. The UK Ministry of Justice (2012) published a longitudinal study of a generic cohort of offenders identified by conviction during January and December 2000. The cohort was made up of 480 000 people; 27.9% reoffended within 1 year, 38.9% after 2 years, 53.2% after 5 years, and 58.9% after 9 years (Ministry of Justice (2012)).

Although data, in the above study, are broken down in relation to age (i.e., 'young offender' and 'adult' offender), revealing higher rates for the younger group, there is no specific

breakdown of reconviction rates in relation to ethnicity. Although, there are indications from the USA (Alaska Judicial Council, 2011) that minority ethnic sex offenders have a higher reconviction rates than Caucasian offenders, this cannot be widely generalized and more detailed study is required in this area.

However, while reconviction data provide some indication of patterns of offending behavior they only refer to behaviors that have been processed through criminal courts, as indicated above much harmful behavior is not reported to any authority. It is very difficult to research and specify offenders' unreported sex crimes, although Soothill (2010), referring to recent US studies, points out that an initial conviction for sex crimes may make it harder for a person to avoid subsequent conviction.

Gender Issues

Kemshall (2004: p. 49), reviewing literature from the 1980s to the present decade, notes that "research consistently indicates that female sex offenders make up about 0.5% of all sexual offences against children." The numbers of female offenders who offend against adults are fewer than those who offend against children (Logan, 2008: pp. 491–492). Therefore, it is reasonable to note that men commit the majority of sex offences. Marshall (1997) estimated that by the age of 40, one in 90 of the men born in England and Wales in 1953 would have a conviction for a serious sexual offense. Moreover, it is likely that many of these men will know their victims.

Sex Offences as Premeditated Acts

Here research around the nature of 'targeting' and grooming is important. Olson et al. (2007: p. 241) describe grooming as "the subtle communication strategies that child sexual abusers use to prepare their potential victims to accept sexual contact." This involves controlling both the victim by concealing the threat posed and manipulating the environment to ensure that the offense will occur without disturbance. Campbell (2009: p. 432) highlights that grooming involves "strategies to present a specific image in an interpersonal relationship through the use of language that indicates the sender's 'persona' to the receiver." She identifies five particular grooming strategies employed by offenders (Campbell, 2009: p. 434): Supplication (presenting a helpless public persona); Intimidation (presenting a powerful or harmful persona); Self-promotion (presenting as competent); Ingratiation (presenting likable or affiliative behaviors); Exemplification (presenting a self that is worthy – demonstrating high moral values). The process of offending is both conscious and premeditated; it does not, as many men in therapy groups have said, 'just happen.'

While empirical data can challenge commonly held assumptions about sex offenders, it has also been systematically used to develop and refine theories related to sex offending. Psychologically based theories have primarily been used to develop more effective therapies. Theories that have a more sociological origin have the potential to influence wider social/public policies, however all theories are socially and culturally located. These issues are reviewed in the next section.

Understanding Sex Offenders: Key Theoretical and Therapeutic Approaches

Theories that seek to understand or explain sexual offending are not merely heuristic devices; developed theories suggest ways of reducing sex offending. Psychological theorizing concentrates on the individual offender and seeks to identify (more) effective therapeutic approaches to reduce the risks posed by individual offenders. The ultimate objective of psychological theory is to improve the clinical response to sex offenders. Evidence for an improved clinical response is complex but much attention is focused on reconviction rates. Sociological (including social anthropological) theories consider sex offending as a social phenomenon. The ultimate objective for sociological theory is to locate sex offending as a social problem that requires a social/public policy response. In many ways, social work is located across the boundaries of these two perspectives in that social workers work with both individuals and with communities, and in some cases they adopt a campaigning role about social issues.

Psychological Perspectives

From a psychological perspective, the works of Ward and colleagues are of key importance (e.g., Ward and Maruna, 2007). This work has been concerned to analyze and develop psychological theory in relation to understanding sex offenders. Ward and Hudson (1998: p. 47) noted:

> The major task of a good etiological theory of rape, child molestation, or any other form of sexual offending is to account for its onset, development, and maintenance.

They also noted that there were three different types of theory that sought to explain sex offending:

- Level 1: multifactorial theories offer a complex account of the etiology and continuance of sex offending.
- Level 2: single factor theories focus on one issue to account for the etiology and continuance of sex offending.
- Level 3: micro-level or offense process theories give particular attention to an aspect of offending behavior.

In the mid-/late twentieth century single factor theories were common and often in conflict with one another. For example, feminist theory and evolutionary psychology theory are both described as single factor theories. The former asserts the key importance of patriarchal social conditioning of men; and the latter emphasizes the importance of physiological urges to perpetuate 'the species' (for a review of these issues, see Travis, 2003). Micro-level theories explore specific aspects of the detail of offending behavior and develop appropriate therapeutic strategies. However, the theories that have had increasing influence are multifactorial theories that incorporate the following dimensions: physiological, psychological, emotional, situational, historical (fixed), and current circumstances in accounting for sex offending. A key aspect of these theories is that they have been developed from analyses

of empirical studies of sex offenders, their personal histories, and their offense patterns.

Within the development of this research, the work of Bonta and Andrews (e.g., Andrews et al., 1990, 2011) is of great significance. Their work has focused on identifying factors associated with risk of reoffending and 'criminogenic need' became a key focus of research during the 1980s. Using statistical analytical techniques, they began to identify that both personal history and current circumstances were relevant to understanding risk of reoffending. Historical unchangeable factors (such as childhood events, or criminal convictions) were considered to be static factors that may relate to (re)offending, while current circumstances (e.g., family/intimate relationships, level of education, employment, accommodation, use of leisure time, use of substances, and social relationships) also had a crucial bearing on the likelihood of (re)offending. These latter circumstances were construed as dynamic risk factors because they were amenable to change and linked to (re)offending. These areas were known as areas of 'criminogenic need,' that is, they are related to the start and continuance of criminal behavior, but if they could be changed they might lead to the cessation of criminal behavior.

While Andrews and Bonta's work recognized the importance of static factors in offending, it did not seek to offer an etiological account. Rather, it focused on key issues in current circumstances that were amenable to change and from this focus developed the 'Risk-Need-Responsivity' (RNR) treatment model. There have been other important attempts to offer multifactorial accounts of sex offending, for example, the confluence model of sexual aggression; integrated theory; quadripartite model of sexual aggression; the four preconditions model, the pathways model (for a full review of these approaches see Beech and Ward, 2004) and the Integrated Theory of Sexual Offending (Ward and Beech, 2006). Beech and Ward (2004) have highlighted the importance of developing credible and coherent theories that account for the onset and continuance of sexual offending, for without such theories, they suggest therapeutic programmes and risk assessments are flawed. McNeill (2012) has recently highlighted that these approaches are theories of the *etiology* of sex crime and suggests that more recent theorizing about *desistance* from crime may be of greater value in helping sex offenders to refrain from committing further offences.

In recent years there has developed a considerable body of criminological research in relation to general desistance from crime. Farrall (1995: p. 56) notes "most of the research suggests that desistance 'occurs' away from the criminal justice system. That is to say that very few people actually desist as a result of intervention on the part of the criminal justice system or its representatives." Ward and Maruna (2007: p. 14) comment that it "makes considerable sense to develop models of rehabilitation on the basis of what is known about how reintegration works in the 'natural' environment outside of the criminal justice system." McNeill (2012) notes that there have been three separate perspectives on desistance that are increasingly becoming integrated: (1) the significance of age and maturation; (2) life transitions and social bonds; (3) personal and social identity. With the theoretical focus on what enables an (ex)offender to live a crime-free life attention is focused on slightly different issues, for example, what is a positive lifestyle and how is it attained?

The two dominant psychological approaches to working with sex offenders in the West are the 'RNR' model (Andrews et al., 2011) and the Good Lives Model (GLM) (Ward et al., 2012). Both models assert that they are underpinned by a multifactorial theory of sexual offending and that their interventions are well evidenced. The two models appear to be competing for therapeutic hegemony in the West and the above two papers neatly encapsulate the differences in the models. Both models recognize the importance of assessing risk and link therapeutic programs to identified areas of risk. Advocates of GLM assert that RNR is overpreoccupied with identifying and policing offender risk and ignores the positive potential all offenders have to live a 'good life' (with help). Advocates of RNR assert that there is little empirical evidence available to support the assertions of the efficacy of the GLM model. They also suggest that the GLM does not give sufficient careful attention to identifying and managing risk. Advocates of both approaches vigorously refute criticisms and cite counterevidence.

Sociological Perspectives

Psychological and criminological theories in relation to sex offending focus primarily on the individual offender and seek to link understandings of the 'cause' of offending and desistance to the development of therapeutic programs. Sociological theories widen the focus and consider sex offending as a social rather than an individual phenomenon. Theories reviewed so far are largely based on an understanding of sex crimes that is dependent on and contained within the processes of criminal justice systems. Both 'criminogenic need' and 'desistance' are identified from research on convicted sex offenders. Sociological (including feminist) theorizing of sexual offending looks beyond the individual offender and beyond the criminal justice system. These perspectives are particularly important for social work(ers) to consider because social work often engages with people who have sexually harmed others but have not been convicted of such acts.

Feminist theory recognizes that sex crimes are rarely reported and those that are reported rarely result in a conviction. Kelly et al. (2005) note that findings from the 2001 British Crime Survey indicated that, by the age of 16, 7% of women had suffered at least one serious sexual assault. Five percent had been raped. The study also looked at attrition in relation to rape crime reports (i.e., rapes reported to the police) and rape convictions. They found that of 3527 crime reports of rape only 8% resulted in a conviction. Similar patterns of attrition are noted in relation to all sex crime reports across Europe and beyond (Jehle, 2012). Feminist theory not only widens consideration of sex offences, it also challenges overdependence on what is legally defined as a sex crime. Kelly (1988) presented qualitative data from 60 women who had experienced some type of sexual violation, and from this material she developed a continuum of sexual violence that ranged from unwanted nonphysical approaches through to rape. Kelly's continuum has continued to be a very

influential framework for recognizing and understanding the place of coercive sex within a wider range of personal relationships. Gavey (2005) makes use of this framework in her powerful presentation of the commonplace nature of sexual coercion within Western cultures.

However, the work of Sanday (2003) is of central importance in theorizing sexual offending in a social context. In her work spanning three decades, she has drawn attention to the characteristics of 'rape-free' and 'rape-prone' societies and communities. Whilst some of her work involves comparative anthropological studies of differing societies, she also studied fraternity houses in universities of the United States. Having carefully reviewed her comparative data she notes:

> 'Rape-free' societies, gender relations were marked by respect for women as citizens, significant female power and authority, and the near absence of interpersonal aggression in social relations … 'Rape-prone' societies, social relations were marked by interpersonal violence in conjunction with an ideology of male dominance enforced through the control and subordination of women. … Faced with such data the only reasonable conclusion by empirically minded social scientists must be that culture plays a role in the expression of male sexual aggression.
>
> (Sanday, 2003: p. 337)

Her conclusion is further substantiated by a group of psychological studies, sometimes known as 'rape proclivity' studies. In this area of research, male undergraduates in North American Universities were asked if they would rape if they were guaranteed that there would be no negative consequences for themselves (e.g., criminal prosecution). A consistent finding in these studies is that between 25 and 30% of respondents indicate that they would rape (Cowburn, 2005). Furthermore, a British study (Burton et al., 1998) involving young people ($n = 2039$) found that young men thought it was acceptable to force a woman to have sex in specified circumstances: if she was his wife; his long-term girlfriend; if he was 'so turned on he could not stop'; if nobody would find out; if he had spent lots of money on her; or if she had slept with many men. Another British study (Burman and Cartmel, 2006) of young people's attitudes found that a third of young men and a sixth of young women thought that using violence in intimate relationships was acceptable under certain circumstances.

Emerging from this work, there has also been a growing body of profeminist scholarship. Mac an Ghaill studied young people at secondary/high school, and identified the influences of 'compulsory heterosexuality, misogyny, and homophobia' (1994: p. 9) in shaping young masculine identities. The juxtaposition of compulsory heterosexuality and misogyny points strongly to values and attitudes that support sex offending. Increasingly, men and their behaviors are being placed at the center of attempts to understand sex offending (see, e.g., Dominelli, 1989; Cowburn, 2006).

Theories that do not look beyond clinical utility are limited because sex offending is not confined to a discrete population of offenders, whose risk can be reduced and who can be supported in their desistance from offending. While some sex offenders may receive effective help, there is always a new population emerging. In researched populations of sex offenders, where conviction data are presented, the majority have no previous offences (Cowburn, 2005). Thus, dealing with people who commit sex offences requires more than clinical excellence, it requires consideration of the population of people who harm others without being convicted and it requires consideration of sex offending at a social/public policy level.

Ward and Hudson (1998) comment that 'a good etiological theory' of sex offending should account for 'its onset, development, and maintenance.' While both Andrews, Bonta, and colleagues and Ward and colleagues note the importance of the social context and cultures wherein offending occurs, this is given little or no sustained attention in their theories of sex offending or therapeutic programs, which focus on individual offenders but do not consider social identities; for example, gender, ethnicity, and class. Given that sexual offences occur in social contexts that may contribute substantially to their onset, development, and maintenance, a comprehensive theory must incorporate social and cultural issues and aim to improve not only therapy but also social/public policy responses to sex crime. It is at the intersection of these areas that social work and social workers are involved with both convicted and nonconvicted sex offenders.

Responding to Sex Offenders: Rehabilitation, Public Health, and Social Work

> Rehabilitation, therefore, is not just about sorting out the individual's readiness for or fitness for reintegration. It is as much about rebuilding the social relationships without which reintegration is impossible. Any would-be supporter of rehabilitation has to do more than sort out 'offenders'. S/he needs to mediate relationships between people trying to change and the communities in which change is impeded or impelled; as well as mediate the role and limits of the state itself in the process.
>
> In sum, my argument is that, viewed from the perspective of desistance theory and research, individual-level psychological theories, analysis and prescriptions are best seen as a necessary but an insufficient part of developing an integrated theory of rehabilitation.
>
> (McNeill, 2012: pp. 13–14)

Social work with sex offenders is about both the individual rehabilitation of the sex offender and development of safer communities. The opening quotation from IASSW strongly highlights the personal and the social orientations of social work. Social workers work with people, while seeking to promote social change. Social workers use both psychological and sociological theories to inform their work, which is underpinned by values of human rights and social justice. However, these aspirations may take varying forms according to the professional context in which social workers operate. In relation to working with sex offenders, social workers may be required to work in a correctional setting, in a child and family setting, or with vulnerable adults. In all of these settings, social workers will be required to understand and assess what is happening and respond in an appropriate manner.

The theories and models of sex offending described above focus on individual risk, individual rehabilitation, and the social context in which offending occurs; potentially social

work is involved in all of these areas. Thus, social work is about more than just the rehabilitation of an individual offender; it is about facilitating change at both an individual level (desistance) and a social level, highlighting and changing endemic social and cultural attitudes that support sex offending. In the early years of twenty-first century, there have been attempts by both academics and governments to integrate individual and social perspectives in developing strategies to reduce sex crimes. Richard Laws, a prominent North American clinician and academic, has developed a three- tier 'public health approach' to sex offending. He suggests that it is necessary to (1) prevent harmful behavior from starting; (2) quickly address early signs of coercive/harmful behavior; and (3) provide treatment programs for people who have been convicted and are in the criminal justice system (Laws, 2008: p. 612). The strength of this model is that it acknowledges and addresses both actual sexual harm and the potential to harm sexually that is outside the jurisdiction of the criminal justice system. The challenge for social work is to operate effectively at all levels within this model.

See also: Antisocial Behavior; Child Protection; Children and Families in Social Work; Probation, offenders and Risk; Risk Assessment in Social Work; Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice.

Bibliography

Alaska Judicial Council, 2011. Criminal Recidivism in Alaska, 2008 and 2009. Alaska Judicial Council, Anchorage, Ak. http://www.ajc.state.ak.us/reports/recid2011.pdf (accessed 18.08.13.).

Andrews, D.A., Bonta, J., Hoge, R.D., 1990. Classification for effective rehabilitation: rediscovering psychology. Criminal Justice and Behavior 17 (1), 19–52.

Andrews, D.A., Bonta, J., Wormwith, J.S., 2011. The risk-need-responsivity model: does adding the good lives model contribute to effective crime prevention? Criminal Justice and Behavior 38 (7), 735–755.

Beech, A.R., Ward, T., 2004. The integration of etiology and risk in sexual offenders: a theoretical framework. Aggression and Violent Behavior 10, 31–63.

Burman, M., Cartmel, F., 2006. Young Peoples' Attitudes to Gendered Violence. NHS Health Scotland, Edinburgh.

Burton, S., Kitzinger, J., Kelly, L., Regan, L., 1998. Young People's Attitudes towards Violence, Sex and Relationships: A Survey and Focus Group Study. Zero Tolerance Charitable Trust, Edinburgh.

Campbell, A.M., 2009. False faces and broken lives: an exploratory study of the interaction behaviours used by male sex offenders in relating to victims. Journal of Language and Social Psychology 28 (4), 428–440.

Cohen, S., 1972. Folk Devils and Moral Panics. MacGibbon and Kee, London.

Cowburn, M., 2005. Hegemony and discourse: reconstruing the male sex offender and sexual coercion by men. Sexualities, Evolution and Gender 7 (3), 215–231.

Cowburn, M., 2006. Constructive work with male sex offenders: male forms of life, language games and change. In: Parton, N., O'Byrne, P., Gorman, K., Gregory, M. (Eds.), Constructive Probation Work. Jessica Kingsley, London.

Dominelli, L., 1989. Betrayal of trust: a feminist analysis of power relationships in incest abuse and its relevance for social work practice. British Journal of Social Work 19, 291–307.

Farrall, S., 1995. Why do people stop offending? The Scottish Journal of Criminal Justice Studies 1 (1), 51–59.

Gavey, N., 2005. Just Sex: The Cultural Scaffolding of Rape. Routledge, London.

Grady, M.D., 2009a. Sex offenders part I: theories and models of etiology, assessment and intervention. Social Work in Mental Health 7 (4), 353–371.

Grady, M.D., 2009b. Sex offenders part II: policies that address sex offenders. Social Work in Mental Health 7 (4), 372–384.

Grady, M.D., Abramson, J.M., 2011. Has social work heeded the call? Sex offender content in social work. Social Work Education: The International Journal 30 (4), 440–453.

Harris, A.J.R., Hanson, R.K., 2004. Sex Offender Recidivism: A Simple Question (User Report No. 2004–03). Public Safety and Emergency Preparedness, Canada.

Home Office, 2007. Review of the Protection of Children from Sex Offenders. Home Office, London.

International Association of Schools of Social Work, 2001. International Definition of Social Work. http://www.iassw-aiets.org/international-definition-of-social-work (accessed 01.07.13.).

International Federation of Social Workers, 2012. Definition of Social Work. http://ifsw.org/policies/definition-of-social-work/ (accessed 01.07.13.).

Jehle, J.-G., 2012. Attrition and conviction rates of sexual offences in Europe: definitions and criminal justice responses. European Journal of Criminal Policy & Research 18, 145–161.

Jewkes, Y., 2004. Media and Crime: Key Approaches to Criminology. Sage, London.

Kelly, L., 1988. Surviving Sexual Violence. University of Minnesota Press, Minneapolis.

Kelly, L., Lovett, J., Regan, L., 2005. A Gap or a Chasm? Attrition in Reported Rape Cases. Home Office Research, Development and Statistics Directorate, London.

Kemshall, H., 2004. Risk, dangerousness and female offenders. In: McIvor, G. (Ed.), Women Who Offend. Jessica Kingsley, London, pp. 209–226.

Krug, E., Dahlberg, L., Mercy, J.A., Zwi, A.B., Lozano, R. (Eds.), 2002. World Report on Violence and Health, Geneva, Switzerland [online]. World Health Organization – http://whqlibdoc.who.int/publications/2002/9241545615_eng.pdf.

Lancaster, E., 1997. Perpetrators of sexual crime: what every social work student should know. Social Work Education: The International Journal 16 (4), 91–110.

Laws, D.R., 2008. The public health approach: a way forward? In: Laws, D.R., O'Donohue, W.T. (Eds.), Sexual Deviance: Theory, Assessment, and Treatment. The Guildford Press, New York, pp. 611–628.

Logan, C., 2008. Sexual deviance in females: psychopathology and theory. In: Laws, D.R., O'Donohue, W.T. (Eds.), Sexual Deviance: Theory, Assessment, and Treatment. Guilford, New York, pp. 486–507.

Mac an Ghaill, M., 1994. The Making of Men: Masculinities, Sexualities, and Schooling. Open University Press, Buckingham.

Marshall, P., 1997. The Prevalence of Convictions for Sexual Offending. Home Office Research and Statistics Directorate. Research Findings No. 55. Home Office, London.

McNeill, F., 2012. Four forms of 'offender' rehabilitation: towards an interdisciplinary perspective. Legal and Criminological Psychology 17 (1), 18–36.

Ministry of Justice, 2012. 2012 Compendium of Re-offending Statistics and Analysis: Ministry of Justice Statistics Bulletin. Ministry of Justice, London.

Minnesota Department of Corrections, 2007. Residential Proximity and Sex Offense Recidivism in Minnesota. Minnesota, St. Paul.

Myers, S., 2008. Revisiting lancaster: more things that every social work student should know. Social Work Education: The International Journal 27 (2), 203–211.

Myhill, A., Allen, J., 2002. Rape and Sexual Assault of Women: The Extent and Nature of the Problem. Home Office, London.

Nelken, D., 2009. Comparative criminal justice: beyond ethnocentricism and relativism. European Journal of Criminology 6 (4), 291–311.

Olson, L.N., Daggs, J.L., Ellevold, B.L., Rogers, T.K.K., 2007. Entrapping the innocent: toward a theory of child sexual predators' luring communication. Communication Theory 17 (3), 231–251.

Percy, A., Mayhew, P., 1997. Estimating sexual victimisation in a national crime survey: a new approach. Studies on Crime and Crime Prevention 6 (2), 125–150.

Sanday, P.R., 2003. Rape-free versus rape-prone: how culture makes a difference. In: Travis, C.B. (Ed.), Evolution, Gender and Rape. The MIT Press, Cambridge, MA, pp. 337–361.

Soothill, K., 2010. Sex offender recidivism. Crime and Justice: A Review of Research 39 (1), 145–211.

Travis, C.B. (Ed.), 2003. Evolution, Gender, and Rape. The MIT Press, Cambridge, MA.

Ward, T., Beech, A., 2006. An integrated theory of sexual offending. Aggression and Violent Behavior 11, 44–63.

Ward, T., Hudson, S.M., 1998. The construction and development of theory in the sexual offending area: a metatheoretical framework. Sexual Abuse: A Journal of Research and Treatment 10, 47–63.

Ward, T., Maruna, S., 2007. Rehabilitation: Beyond the Risk Paradigm. Routledge, London.

Ward, T., Yates, P.M., Willis, G.M., 2012. The good lives model and the risk need responsivity model: a critical response to Andrews, Bonta, and Wormwith. Criminal Justice and Behavior 39 (1), 94–110.

Sex Role Development and Education

Angel Nga-Man Leung, Center for Psychosocial Health and Aging, The Hong Kong Institute of Education, New Territories, Hong Kong
Henry KS Ng, The University of Hong Kong, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Nunner-Winkler, volume 21, pp. 13958–13961, © 2001, Elsevier Ltd.

Abstract

This article reviews major theories and empirical evidence on the development of sex (or gender) role. Gender role is different from sex; it is more about the socialization process of what and how men and women should behave. Biological, evolutionary, cognitive developmental, and social learning theories are first described to explain formation of gender stereotypes and gender schemas, followed by discussions on how family, education, workplace, media, and cultural settings would further reinforce the preexisting gender role. Bronfenbrenner's ecological theory is applied to conclude how different social contexts interact and impact on the development of gender role.

Ever since we had knowledge on genetics and the sex chromosomes, gender and sex identification have become not only clearer but also more complicated. The XY sex determination system differentiates the two sexes on the 23rd pair of chromosomes, where males have the XY combination and females have the XX combination. Still, there are cases when individuals are born with other combinations, such as XO (having one X chromosome only) XXX, XXY, or XYY chromosomes (for review, see Blackless et al., 2000). The ambivalence is also seen in external genitals, a much more common clue we rely on in determining newborn babies' gender. Some babies are born with inconsistent internal and external genitals, external genitals that resemble both sexes, or ambiguous genitals. Intersexed babies or children have usually undergone surgeries for 'correction' of their genitals to fit the binary differentiation of the two sexes. Nevertheless, one's sexual identity or gender role seems not to be easily manipulated. One of the most famous cases in how 'assigned sex' does not work would probably be the case of David/Brenda Reimer (1965–2004). David was accidentally castrated when he was 7 months old. To correct the surgical failure, his parents made him undergo hormonal and surgical treatments, and raised him as a girl, Brenda. Not knowing that she was actually a boy, Brenda still preferred to and acted like a boy. At the age of 15, he was told the truth and eventually received multiple surgeries and hormonal treatments to rebuild his penis and manhood. Living his whole life struggling to fit in the role being a male or female, David ended his own life at the age of 38 (Wood, 2011). The tragic story of David Reimer implied two things: (1) biological or genetic factor probably plays a crucial role in defending ones' sexual identity or sex role and (2) the binary definition of sex does not always apply to everyone and may cause unnecessary pain to certain individuals.

Sex, gender (or sex role), and sexual orientation (or sexuality) are constructs that are usually viewed as consistent with each another. In most cultural settings and most of the time, people expect males to process a male sex role, perform in a masculine way, and be interested in females, while females are expected to process a female sex role, perform in a feminine way, and be interested in males. Sex is expected to be binary: a person can be either male or female. However, this is not always applicable to everyone. Besides intersexed individuals, there are also transgendered individuals, who endorse sexual identities that are opposite of their biological sex.

Sex is binary most of the time, but gender is not. Gender refers to how individuals express themselves; even it does not fit into ones' biological sex. Sexuality is also not necessarily to be consistent with one's sex. Compared to biological sex, gender is more about the socialization process of what and how men and women should behave. In other words, it is how one does rather than whom one is. Still, biological sex is inseparable from gender. Biological factors potentiate individuals to behave, while gender allows a flexible variation and input from the environment (e.g., learning experience). If a baby is born to be biologically male, external forces, such as the parents and school, would exert forces at different level to direct his gender schema and behaviors.

In the following sections, several major theories and related empirical evidence are considered.

Biological Factors

The word 'gender' apparently refers to differences rather than similarities between the two sexes. There has been a long history that scholars tried to trace gender differences to the fundamental biological differences. Recent research evidence has supported the genetic effect on sex role and related behaviors. Very often, the society encourages sex typing of children so that they learn their prescribed roles very early. Sex-typed behavior means babies would show different patterns of behavior according to their sex, even earlier than they develop any concrete ideas about gender (Boyd and Bee, 2012). As early as 18–24 months, infants already show some preference over toys for their gender (Thommessen and Todd, 2010). Some children, however, exhibit gender atypical behaviors (GABs) that have attracted attention from psychologists. Knafo et al. (2005) examined the genetic and environmental effects on GABs by studying 5799 twin pairs (3- to 4-year-old twin pairs). Gender atypical children were defined

as those who had high affiliation with the opposite sex roles and behaviors (i.e., feminine boys, masculine girls). The researchers found a modest genetic influence on boys' femininity (37%) and a strong genetic influence on girls' masculinity (82%). The effect of shared environment was much stronger for gender atypical boys (42%) than for such girls (0%). One plausible reason is that parents (and the society) are more tolerant of the boundary crossing of girls than boys, so that they put more effort (external forces) to socialize their sons to adopt masculinity. The genetic effect on GABs was also found in other age groups (7- and 10-year-old children, van Beijsterveld et al., 2006).

One major implication of GAB is its effect on the development of sexual orientation in adulthood. In a recent study, Alanko et al. (2010) reported that GAB in childhood shared common genetic background with that of homosexuality in adulthood, which was stronger for men than women. Bem's (2000) exotic becomes erotic theory suggests a provocative explanation of the linkage. Accordingly, GAB is theorized as an arousing agent because gender atypical children are often alienated by same-sex peers. As one grows up, the physiological arousal may be misattributed to erotic attraction to the same-sex teasers, thus causing homosexuality. As the alienation (and thus physiological arousal) is often stronger for boys, it is not surprising why the GAB-homosexuality linkage is stronger for male.

Two limitations are observed for genetic studies on sex role behaviors. First, although genetic factors account for the development of (atypical) sex role, the specific effective genes are not identified yet. Would it be the same as the 'gay' gene (i.e., the Xq28 region, Hamer et al., 1993, which is inconclusive either)? Without knowing the effective genes, it is impossible to know how genetics transfers its effects to gender role behaviors. Furthermore, gender role (a)typicality is by itself a relative construct that is under the impact of environmental forces. Nothing can be atypical unless there is something that is commonly endorsed (i.e., typical). In view of this, macro-level factors, such as culture and religious beliefs, may play more crucial role in the development of sex role.

As mentioned in the aforementioned paragraph, there are occasions that individuals are born with variations in their XX or XY chromosomes. Wood (2011) and Pressley and McCormick (2007) summarized studies on how cognitive or physical attributes may be affected by hormones or chromosomal combinations as follows: individuals who have XXY chromosomes (Klinefelter syndrome) tend to be taller but with less androgens, and they may exhibit lower spatial abilities but with normal intelligence quotient; individuals who have XO chromosome (i.e., one X chromosome only; Turner syndrome) may not have usual puberty changes and may show poorer spatial skills. Girls with XXX (triple X girls) may be taller than other XX girls, and they have normal puberty and intelligence. In addition, sex hormones and differences in using different lobes of brains between the two sexes are generally believed to be related to some differences between men and women. Nevertheless, previous research on the gender differences of cognitive abilities was inconclusive (e.g., Hyde et al., 1990). In addition, the modest differences seem to be a result of environmental factors more than one's biological sex.

Evolution and Selection of Sex Roles

According to Charles Darwin's theory of evolution (1859), the traits of a species evolve as a result of adaptation to environmental challenges. Individuals who have developed the more adaptive traits are more likely to survive and reproduce, thus passing the traits to the next generation. In other words, challenges from Mother Nature decide which traits to be augmented, diminished, or modified within a species (i.e., natural selection). Over the past two decades, such an idea has been extended to understand human behavior, especially gender roles and sexual behaviors. Evolutionary psychology, as the name suggests, seeks to explain individual differences based on the evolutionary history of our species.

Evolutionary psychology (see Buss and Schmitt, 2011, for review) holds that the psychological mechanisms we currently have are the results of ancestral adaptation. Reproduction is of particular importance because it determines what traits, be it physical or psychological, can be passed to the offspring. In this regard, the different reproductive systems of men and women pose different challenges and economic tradeoffs to them, leading to different sexual behaviors and strategies. Men, whose reproductive system features mass production of sperms and longer active period throughout the lifespan, have little cost in impregnating women. The fact that they do not necessarily need to bear their children also reduces their parental investment. As a result, men usually prefer short-term sexual relationships and multiple sexual partners. In contrast, the reproductive and parental costs of women are much greater: the ova are far rarer in a woman's lifespan, and much resource is devoted to bearing and raising a child. As such, a supportive and committing long-term partner is more important for women to attain reproductive success. It is believed that the different sexual strategies determine why men and women adopt different gender roles. Buss and Schmitt (2011) summarized abundant studies that were believed to support the sexual strategies theory. To sum up, the studies showed that men were more likely and willing to have multiple, short-term sexual relationships with strangers than women. However, it is questioned whether these findings are direct support or just a circular argument of the theory. While the sexual strategies theory aims to explain the different mating strategies of men and women, finding that the two sexes actually differed is but a reiteration of its origin. If gender behaviors and roles are evolved, the most direct evidence should show how the changes occur. In biology, this is attained by fossil records of different species sampled from different layers and periods. Such evidence is seemingly absent in the evolutionary psychology literature.

If gender role is a result of adaptation to reproductive challenges, the aforementioned gender differences should be observed irrespective of the context one is in. In other words, the observed gender role differences should be universally true across different cultures. While earlier studies supported this cross-cultural universality (e.g., Schmitt, 2005), recent research has provided opposite evidence. For example, Perrin et al. (2011) investigated four love and sexual behaviors that were expected to be different across men and women, but only reported a significant difference of desires for relationship support. Moreover, the difference was largely explained by

individual's gender role identity (masculinity vs femininity), implying that the difference is a result of gender role socialization. In a later study, Zentner and Mitura (2012) compared the gender differences in mate preferences across societies of different levels of gender equality. The expected gender difference from evolutionary theory was only observed in nations of low sexual parity (e.g., Turkey, Korean Republic). The difference declined as the gender inequality gap decreased (e.g., Finland, Philippines).

It should be noted that by evolution, Buss and Schmitt (2011) did not assert that gender roles and behaviors are purely genetically determined. If certain gender roles are inherited across generations, it can be 'selected' through nongenetic mechanisms, such as gender role conformity, social conditioning, modeling, and reinforcements. Still, the social constructionist perspective would question whether such a selection theory overinterprets and complicates the simple social influence. For example, the social pressure to conform to gender norms may be termed as a selection of the trait, which in reality does not contribute to the understanding of the gender issues. Meanwhile, it is hard to totally refute the role of genetics in determining gender roles, as reviewed in the previous section. Regarding the biological and evolutionary theories, more direct evidence is required on how gender roles are acquired through biological mechanisms. In the following, we discuss the other side of the argument – the cognitive and social view on gender roles.

The Development of Gender: Cognitive Developmental Theories

Kolhlberg (1966) suggested a cognitive developmental theory on understanding the gender development in children. By the age of 2 years, most infants can identify themselves as boys or girls, which is the first step of forming gender identity. Children gradually understand that gender is something stable and that they would remain as either boys or girls. They have insights about how physical features define the gender of individuals and gain gender stability. For instance, men are those in beard and short hair, while women are those in dress and long hair. Sometimes, children make mistakes by thinking that external outfits will change the gender for individuals; they may think that if a boy wears a dress and lipstick, he will turn into a girl. Such confusions on gender identity among young children reflect their limitations in cognitive functioning, and this can be explained by applying Piaget's (1977) theory. According to Piaget (1977), children in the preoperational stage usually lack the awareness of conversation, which hinders them from maintaining a constant identification of an object with changing appearances. It is usually by the age of 5–6 years that children would attain gender constancy, so that they understand that gender does not change along with physical appearances and that their gender will remain unchanged in their whole life.

The attainment of gender stability and constancy signifies the beginning of the socialization and social learning of gender roles. By identifying themselves as either boys or girls, children learn the appropriate behaviors and preference of their gender and form gender schema, which is a "packet of gender-relevant information, understanding, knowledge, and beliefs" (Levy and Fivush, 1993: p. 132). Such a mental structure affects how gendered information is filtered, processed, and reacted to. Preschoolers make inferences of what people should do based on their rigid gender schemas. Behaviors that violate the schemas become taboos and are avoided (e.g., boys should not play with dolls; girls should not play with guns). The violation of gender schemas is also avoided by limiting plays with same-sexed playmates (vs opposite sex ones) only (e.g., Martin and Little, 1990). The development of gender schemas continues as children age, when they gradually understand the complexity and flexibility of gender roles. In the following, we discuss how gender schema is organized and the different sources that contribute to its formation.

Knowledge of Gender Roles and Behaviors

Gender schemas and stereotypes are important sources of knowledge that guide our behaviors in a gendered context. They allow us to make quick judgments on people and situations based on gender. According to gender schema theory (Bem, 1981), masculinity and femininity are often unconsciously applied to different occupations and activities. To name a few, ballet dancers, flight attendants, and nurses may often be assigned to the feminine schema, while warriors and soldiers are often assigned to the masculine schema. Consider the naming of the *X-Men*, which is a team of mutant superheroes (both men and women) that fights against different evil parties. It would probably be against most people's gender schema when the team is renamed as *X-People*, or the heroines are called *X-Women*.

How is gender knowledge organized in the human mind? Apart from masculinity and femininity, recent research has supported a general model of social cognition that may bring insights into research on gender schema and stereotypes. Fiske and colleagues (see Fiske et al., 2007, for review) theorized two fundamental stereotypes in social cognition, namely, warmth and competence. According to the researchers, the two universal dimensions of perceiving individuals and out-groups serve survival purposes. The warmth stereotype is assessed first, which determines whether the others are friendly or not. Being the primary stereotype, perceived warmth determines whether one would help (active facilitation) or attack (active harming) the perceived. The competence stereotype is then assessed, which tells whether the others have the ability to execute their intentions. Perceived competence, the secondary stereotype, determines whether one would associate with (passive facilitation) or neglect (passive harm) the target. In social cognition, perceivers would map the target according to the two stereotypes and thus lead to different behavioral and affective responses.

The two fundamental stereotypes, when combined, predict different emotions toward the perceived individual or outgroup (Cuddy et al., 2007). When high warmth is paired with high competence, admiration is induced, which leads to univalent facilitation (both active and passive) with the perceived. Another univalent emotion happens with low warmth being paired up with low competence, which results in univalent downward emotions (e.g., contempt) and harming

to the perceived. Ambivalent emotions and behaviors are predicted when either dimension is high. For instance, a group is pitied when it is perceived with high warmth but low competence. This combination leads to active facilitation (helping), but also passive harm (negligence). On the other hand, a group is envied when it is perceived as highly competent but low in warmth. This combination leads to passive facilitation (association) due to its competence. However, the perceived unfriendly intent also attracts active harm (attack).

The fundamental stereotypes of warmth and competence may be applied to gender schema and explain how and why people react to men and women differently in the society. For example, Fiske et al. (2002) used cluster analysis to illustrate how several gendered identities are perceived differently on the two stereotypes. The research showed that housewives are perceived as warm and incompetent, sexy women and gay men are perceived as more competent but less warm, and business women, men, as well as rich and educated people are perceived as most competent but low in warmth. It should be noted that the stereotype map differs across sample of raters and time (see Cuddy et al., 2007).

Although the work by Fiske and colleagues did not intend to study gender perception, the model shows resemblance to gender schema theory. The warmth dimension is analogous to feminine communal traits, while the competence dimension is analogous to masculine agentic traits. Still, the two stereotypes have important implications on gender roles research. First, the theory makes clear prediction on how warmth and competence influence one's affective and behavioral responses to the perceived target. It also highlights the possible ambivalence of social perception, that one can exhibit both facilitation and harm to the perceived target (i.e., pity and envy). Based on this model, future research may investigate how different gendered groups are mapped differently on the two dimensions by individuals of different cultures, socioeconomic status, and gender.

How Gender Is Acquired

Family Factors

Language is an important media for the acquisition of gender roles. Parents, who are socialized with gender schemas, transfer the knowledge to their daughters and sons through different verbal responses. Early studies (e.g., Cherry and Lewis, 1976) have reported that mother–daughter communication features two-way communication (e.g., asking questions, making interpretation), while mother–son communication is dominated by one-way communication (e.g., giving instructions and explanations). Interestingly, Clearfield and Nelson (2006) reported that the different communication styles toward daughters and sons occurred as early as 6 months of age, when the infants were not able to speak yet. The observed mothers also engaged more in the play of their daughters (vs sons), which may subtly encourage social interdependence to them. Fathers, on the contrary, generally have more physical play with their children (e.g., Lindsey et al., 2010).

Parents are perhaps the first source for children to learn gender schemas or stereotypes. From birth, parents treat babies differently based on their sex. Wood (2011) suggested that by conveying different messages to little girls and boys via daily conversations, children learn their gender identity. Children compare their own behavior and contrast it with their parents' behavior and expectations and take up the gender-appropriate behavior. For instance, parents tend to play masculine toys with boys and feminine or neutral toys with girls (Wood et al., 2002).

Social Learning and Observational Learning Theories

The social learning theory (Bandura, 1977) suggested that children imitate others for the appropriate gender roles and behaviors. While children are prewired to imitate whatever comes to their attention, rewards and punishments may moderate whether the learning is successful. Reward, be it direct or vicarious, is applied to reinforce gender-appropriate behaviors. For instance, parents tend to have more rough and tumble plays with boys (vs girls); parents also tend to provide gender-appropriate toys to their children (Levy, 1999), and show more acceptance to boys expressing anger than to girls doing the same (Zahn-Waxler and Polanichka, 2004). Parents are generally less tolerant to cross-gender behavior, which means behaving in an atypical manner to one's gender, among boys than in girls. Boys are more likely to be punished for not being masculine than girls being not feminine, or having gender role transgression (e.g., Sirin et al., 2004). As boys grow into men, studies found that men experience more gender role anxiety than women (Vandello and Bosson, 2013). This might be explained by the social learning theory, as men tend to receive punishment for not showing manhood since young age. As mothers are usually the primary caregivers of children, mothers tend to relate themselves more to daughters than to sons. When boys cannot fully identify themselves by modeling mothers, they recognize themselves as being different from mothers (Wood, 2011). Vandello and Bosson (2013) argued that in most cultural settings, boys can only adapt to 'snormal' or 'healthy' male gender roles by separating themselves from their female caregivers, and to avoid femininity by their best efforts to show that they perform, behave, and think differently compared with females to gain manhood.

As children grow up and build broader social connection outside the family, the learning models and rewards are extended, mainly in school settings. Eder et al. (1995) found that in middle schools, males and females are reinforced differently and they therefore may pick up different gender roles. For instance, males are reinforced by both peers and teachers as being athletes, while females are not; males are encouraged to act aggressively, again by teachers and peers; compared with boys, being physically attractive is valued among girls, while physically unattractive girls are more likely to be picked on; sexually active girls receive defaming comments, while sexually active boys receive social praise and approval; heterosexuality is widely accepted in middle schools; and those who do not act according to the widely accepted gender role would be socially rejected. Teenagers may be socialized to adopt the accepted gender role by conforming to the norm. Any preexisting differences in gender role are further strengthened by observing the hidden rule and reinforcing (or punishing) by following or violating them.

Media

With increasing autonomy and literacy, children may acquire gender schema through more channels other than their parents. In a recent study, Fitzpatrick and McPherson (2010) studied the gendered content of 56 coloring books published in the United States. The researchers reported that male (vs females) characters were more likely to be portrayed as authoritative figures (e.g., adults, superheroes) and showing more gross motor actions. Female characters, on the other hand, were often depicted as static and humanlike. It was noted that most of the printed characters also appeared in other popular media, such as movies and cartoons. This implies that the acquisition of gender schema may also occur in other media. Meanwhile, adults seem to have a selection bias toward gender-typed media such that their gender schemas can be self-reinforced (Knobloch-Westerwick and Hoplamazian, 2012). The researchers reported that both biological sex and gender schemas (masculinity vs femininity) predicted the gender-typed magazines that one would pick. The selection bias also appeared to reinforce the prescribed gender schema. The self-reinforcing nature of gender schema explains why the vicious cycle between gender schema and gendered behavior is so difficult to break.

Gender Role Expectations, from the Education System to Work

Gender role expectations affect how males or females perform in educational settings. "Test performance depends on the test taker's characteristics, as well as the test taker's interpretation of the testing situation" (Massa et al., 2005: p. 109). It has been found that for an embedded test on spatial ability, when men's performance was not affected by their gender role, females who held feminine role performed better when the same embedded test was framed as measuring empathy; on the other hand, when the test is framed as measuring spatial ability, females who held masculine role performed better (Massa et al., 2005). This suggested that gender role is highly relevant to how students view themselves, which may then affect how confident (or well) they perform on cognitive tests in certain ways to make their performance consistent with their assigned gender roles. Stereotype threat may hinder one's educational performances. When women are presented with stereotype threat that women did worse compared with men on a test on math ability, they tended to perform less well (Spencer et al., 1999).The internalized stereotype threat may not only hinder one's performance in schools but also influence the choices males and females make in education.

Despite the fact that differences in actual test scores in math or science tests between male and female students are small, female students tend not to be as confident as male students in their performance in math or science. Pressley and McCormick (2007) summarized a number of findings and suggested that parents and teachers tend to 'cooperate' in decreasing girls' confidence in math or science subjects. For instance, parents tend to offer less help to their daughters than sons; they tend to undermine their daughters' abilities in math and science subjects but care about and encourage their sons' performance in math or science subjects more. Teachers also tend to give less help to girls in class and are more responsive to boys' questions. Boys on average received more interactions with teachers for science or math classes.

For a long period, people held strong stereotypes toward educations for males and females. Compared with boys, more girls are illiterate, do not have access to education, or have narrower choices of education (Silberstang, 2011). This may be related to gender stereotype, cultural ideology, or religious practice. More women are doing lower paid jobs or earning less than men. In less than 100 years ago, women were expected to take 'woman subjects', such as home economics only. In developed countries where gender equality is emphasized, females were gradually encouraged to take up more professional jobs such as social work and teaching (Silberstang, 2011). In 2010, although females graduated with bachelor degrees in the United States outnumbered males (954 891 women vs 713 336 men), 'sex tendencies' in professionals are still present. In the United States, men graduated with the more traditionally 'male' subjects, such as degrees in computer and engineering, outnumbering women for 4 to 5 times, while women graduated with the more traditionally 'female' subjects, such as psychology and outnumbered men for 3 to 4 times (National Science Foundation, 2010).

Women working in traditionally 'male-dominated' settings, such as engineering, may also face more difficulties as stereotypes on women being less competent in these professionals are still held. If women do not perform exceptionally well in 'male-dominated' industries or professionals, they would be viewed as incompetent; however, if they excel, they would be viewed as selfish, cold, and not being welcomed by colleagues (Dean, 2006). Despite the fact that more females than males are awarded with bachelor degrees, fewer females become faculty in academia or take up senior executives post (Silberstang, 2011). For instance, the most 'productive' time for junior faculty members to accumulate publications for attaining tenure collides with the most 'reproductive' time of female faculty members. Rhoads (2010) argued that when both males and females are entitled to maternity leave, male faculty members can make use of the leave to produce papers, while female faculty members cannot as they have to carry out duties as mothers. The worst label to a woman is 'being a bad mother' (Chrisler, 2013). Both the gender role stereotypes on women's abilities and the biological limitations that females have shorter reproductive period compared with men may result in a glass ceiling effect in workplace. Women are more often to face with challenges in work–life–family balance as they are expected to be nurturing, kind, and less motivated to earn money but be more interested in family life. Such widely held beliefs strengthen the preexisting gender inequality from schools to workplace. The biological limitations, gender stereotypes, and cultural expectations for women to devote more time and energy on the 'mother track' would ultimately affect women's upward mobility.

Bronfenbrenner's Ecological Theory and Sex Roles

Bronfenbrenner's ecological theory (1993) is applied here to conclude the development and reinforcement of existing sex roles in a given society. The framework of ecological system emphasizes the bidirectional influences between individuals' characteristics and the different environmental contexts surrounding individuals. There are several 'systems' or

Table 1 Summaries and comparisons of different theories concerning sex or gender role development

	Biological or genetic factors on determining sex	*Evolutionary theory on sexual strategies*	*Cognitive developmental theories on the development of gender role*	*Social learning and observational learning theories of gender role*	*Bronfenbrenner's ecological theory on sex or gender role*
Summaries of theories or factors/comparisons: is sex or gender role development more likely to be	Sex is binary most of the time and is biologically determined	Reproduction is of particular importance because it determines what traits can be passed to the offspring. Men and women pose different challenges and economic tradeoffs to reproduce, leading to different 'selected' sexual behaviors and strategies	Children gradually understand that gender is stable by attaining gender stability and constancy as they grow up	Gender roles are learned by imitating others for the appropriate gender roles and behaviors. While children are prewired to imitate whatever comes to their attention, rewards and punishments may moderate whether the learning is successful	The bidirectional influences between individuals' characteristics and the different environmental contexts (in terms of different systems) surrounding individuals determines one's sex or gender role
Inborn or environmental influence?	More likely to be inborn. Biological factors potentiate individuals to behave	Most likely to be inborn, as certain gender roles are inherited across generations	More likely to be environmental. Children understand gendered behaviors differently in different environments	Most likely to be environmental. Children receive different rewards or punishments for gender-appropriate behavior in different environments	Both inborn and environmental. Both individuals' characteristics and environmental input affect sex or gender role development
Active (individuals can control it) or passive (individuals cannot control it)?	Passive, one cannot control the inborn sex but gender	More likely to be passive, as the 'selected' sex roles are mostly inherited	Both active and passive. Children's understanding of gender depends on their cognitive development. Children can actively seek for cognitive maturation but it is also passively determined by the biological maturation of their brains	Both active and passive. Children actively seek for behavior that leads to rewards, but they cannot control their environment in terms of what behavior is rewarded or punished	Both active and passive. It is relatively difficult to change the outermost system such as the macrosystem (i.e., culture). Nevertheless, individuals may also actively affect the systems around them. For instance, gender equality movement starts with several pioneers in striving for gender equality, which gradually affects other systems
Similar or different (different individuals show similar patterns or different patterns?)	More likely to be similar, as genetic or biological factors tend to be more universal	Similar but different. Gender difference from evolutionary theory was observed in nations of low sexual parity. The difference declined as the gender inequality gap decreased	Similar but different. Progress of brain maturation tends to be universal, but environmental influence on the understanding of gender roles is different	More likely to be different. Based on different experience (positive or negative) that individuals receive environmentally, individuals develop different gender roles	More likely to be different. Different cultures and socioeconomic contexts determine sex or gender role development differently

'contexts' surrounding individuals and affect their life span development. Starting with the slight biological differences between the two sexes such as sex-typed behavior that can be found in infants, sex roles are further reinforced and strengthened through the microsystem, mesosystem, exosystem, and macrosystem. Microsystem is the innermost circle surrounding individuals. Closed family members, teachers, and peers belong to microsystem. Families, peers, and teachers reinforce and strengthen the appropriate gender behavior and gender roles of children. Mesosystem refers to the interactions between microsystems. For instance, both parents and teachers tend to undermine girls' math and science abilities, which in turn strengthens the preexisting gender stereotype that girls do less well than boys. Exosystem means the socioeconomic context, such as media, neighborhood, and legal or social welfare services that are available in ones' social settings; it also involves the links between social settings. For instance, mass media impacts individuals on the ideologies about males and females. Children's experience and perceptions about gender role may also be influenced by the working status of mothers. Women are more likely to go out to work if there are suitable social contexts that provide educational opportunities and childcare services that encourage them to work; this indirectly affects the experience of children with working mothers. For instance, daughters of working mothers aspire for career more (e.g., Hoffman, 1998). Macrosystem, which is the outermost layer of Bronfenbrenner's model, refers to social and cultural ideologies that one is exposed to. For instance, some Chinese societies may still endorse the traditional Confucian beliefs about gender roles. Sayings such as *zoi gaa cung fu*; *ceot gaa cung fu*; *lou loi cung zi* (girls should submit to their fathers; wives should submit to their husbands; mothers should submit to their sons in later years, respectively) and *neoi zi mou coi bin si dak* (ignorance is a girl's virtue) are still commonly used. These cultural beliefs have been constantly strengthened and reinforced among Chinese for thousands of years, through families, media, schools' work settings, and so on. The pervasive effects of such beliefs have influenced the gender role perception of the Chinese for thousands of years, resulted in most Chinese women losing chances to receive education, and not being able to question the orders from men, which in turn reinforces the legitimacy of the dominating roles of men in families, job markets, politics, and ruling power. Summaries and comparisons of different theories concerning sex or gender role development are shown in **Table 1**.

Conclusion

In most cultural settings, men should be masculine, i.e., be aggressive, forceful, competent, dominant, and probably the bread winner, and women should be feminine, i.e., be soft, tender, caring, altruistic, and probably a mother. Such an ideology is presented in most cultural settings and is reinforced through the interconnections between the microsystem, mesosystem, exosystem, and macrosystem of the ecological theory (Bronfenbrenner, 1993)

Achieving and maintaining the culturally accepted manhood or womanhood, however, is not easy for both men and women. Vandello and Bosson (2013) argued that manhood is difficult to gain and is easily lost. To gain masculinity, men have to engage in masculine-typed activities such as risk taking and aggressive behavior, achieving in careers, and avoiding being feminine. They argued that maintaining manhood is a constant anxiety for men, as a single feminine act can destroy the masculine image being built and maintained for a long time. Vandello and Bosson (2013) concluded several studies on creating gender threat for males, and it has been found that males experience high level of anxiety and stress when they were asked to carry out feminine tasks. On the other hand, although womanhood or femininity has been viewed as more 'naturally' associated with females, Chrisler (2013) argued that womanhood is equally difficult to achieve. Womanhood is achieved through pursuing culturally accepted standard of beauty and being good mothers, which are also some big challenges that most females have to face.

Despite recent movements in reducing discrimination against individuals who hold different gender roles and gender identities (such as transgendered individuals) or sexual orientation (lesbian, gay, or bisexual individuals) by providing them equal rights in schools, work, communities, and laws, the traditional binary sex roles of femininity and masculinity still dominate the mainstream culture. "Gender role constraints oppress everyone, and we will all be better off if those constraints are loosened" (Chrisler, 2013: p. 119). It is hard to live a life completely free of gender roles or constraints, nevertheless. It is suggested that more efforts can be placed in 'neutralizing' the effects of gender roles in schools and mass media by not emphasizing the difference between males and females. After all, biology gives two sexes, gender does not necessarily have to be binary, and we can think, behave, and live in being feminine, masculine, androgynous, or completely 'genderless'.

See also: Education (Primary and Secondary Schools) and Gender; Gender Development: Cultural Differences; Gender Differences in Personality and Social Behavior; Gender and Education; Gender and Feminist Studies in Geography; Gender and Place; Gender: Anthropological Aspects; Masculinity, Scripts of; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Alanko, K., Santtila, P., Harlaar, N., Witting, K., Varjonen, M., Jern, P., Johansson, A., von der Pahlen, B., Sandnabba, N.K., 2010. Common genetic effects of gender atypical behavior and sexual orientation in adulthood: a study of finnish twins. Archives of Sexual Behavior 39 (1), 81–92.

Bandura, A., 1977. Social Learning Theory. General Learning Press, New York.

van Beijsterveldt, C.E.M., Hudziak, J.J., Boomsma, D.J., 2006. Genetic and environmental influences on cross-gender behavior and relation to behavior problems: a study of Dutch twins at ages 7 and 10 years. Archives of Sexual Behavior 35 (6), 647–658.

Bem, S.L., 1981. Gender schema theory: a cognitive account of sex typing. Psychological Review 88 (4), 354–364.

Bem, D.J., 2000. Exotic becomes erotic: interpreting the biological correlates of sexual orientation. Archives of Sexual Behavior 29 (6), 531–548.

Blackless, M., Charuvastra, A., Derryek, A., Fausto-Sterling, A., Lauzanne, K., Lee, E., 2000. How sexually dimorphic are we? Review and synthesis. American Journal of Human Biology 12, 151–166.

Boyd, D., Bee, H., 2012. Lifespan Development. Pearson, Upper Saddle River, NJ.
Bronfenbrenner, U., 1993. The ecology of cognitive development: research models and fugitive findings. In: Wonziak, R., Fischer, K. (Eds.), Development in Context: Acting and Thinking in Specific Environments. Erlbaum, Hillsdale, NJ, pp. 3–44.
Buss, D.M., Schmitt, D.P., 2011. Evolutionary psychology and feminism. Sex Roles. 64 (9–10), 768–787.
Cherry, L., Lewis, M., 1976. Mothers and 2-year-olds: a study of sex-differentiated aspects of verbal interaction. Developmental Psychology 12 (4), 278–282.
Chrisler, J.C., 2013. Womanhood is not as easy as it seems: femininity requires both achievement and restraint. Psychology of Men & Masculinity 14 (2), 117–120.
Clearfield, M.W., Nelson, N.M., 2006. Sex differences in mothers' speech and play behavior with 6-, 9-, and 14-month-old infants. Sex Roles 54 (1/2), 127–137.
Cuddy, A.J.C., Fiske, S.T., Glick, P., 2007. The BIAS map: behaviors from intergroup affect and stereotypes. Journal of Personality and Social Psychology. 92 (4), 631–648.
Darwin, C.R., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, first ed. John Murray, London.
Dean, C., 19 December 2006. Women in Science: The Battle Moves to the Trenches. The New York Times pp. D6–D9.
Eder, D., Evans, C.C., Parker, S., 1995. School Talk: Gender and Adolescent Culture. Rutgers University Press, New Brunswick, NJ.
Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82 (6), 878–902.
Fiske, S.T., Cuddy, A.J.C., Glick, P., 2007. Universal dimensions of social cognition: warmth and competence. Trends in Cognitive Sciences 11 (2), 77–83.
Fitzpatrick, M.J., McPherson, B.J., 2010. Coloring within the lines: gender stereotypes in contemporary coloring books. Sex Roles 62 (1–2), 127–137.
Hamer, D.H., Hu, S., Magnuson, V.L., Hu, N., Pattatucci, A.M., 1993. A linkage between DNA markers on the X chromosome and male sexual orientation. Science 261 (5119), 321–327.
Hoffman, L.W., 1998. The effects of the mother's employment on the family and the child. Parenthood in America. In: Proceedings of the Conference Held in Madison, Wiscons, in 19–21 April 1998. Retrieved 25 October 2013, from: http://parenthood.library.wisc.edu/Parenthood.html.
Hyde, J.S., Fennema, E., Lamon, S., 1990. Gender comparisons of mathematics attitudes and affect. Psychology of Women Quarterly 14 (3), 299–324.
Knafo, A., Iervolino, A.C., Plomin, R., 2005. Masculine girls and feminine boys: genetic and environmental contributions to atypical gender development in early childhood. Journal of Personality and Social Psychology 88 (2), 400–412.
Knobloch-Westerwick, S., Hoplamazian, G., 2012. Gendering the self: selective magazine reading and reinforcement of gender conformity. Communication Research 39 (3), 358–384.
Kohlberg, L., 1966. A Cognitive-Developmental Analysis of Children's Sex-Role Concepts and Attitudes. In: Maccoby, E.E. (Ed.), The Development of Sex Differences. Stanford University press, Stanford, CA, pp. 82–172.
Levy, G.D., 1999. Gender-typed and non-gender-typed category awareness in toddlers. Sex Roles 41, 851–873.
Levy, G.D., Fivush, R., 1993. Scripts and gender: a new approach for examining gender role development. Developmental Review 13, 126–146.
Lindsey, E., Cremeens, P., Caldera, Y., 2010. Gender differences in mother-toddler and father-toddler verbal initiations and responses during a caregiving and play context. Sex Roles 63 (5–6), 399–411.
Martin, C.L., Little, J.K., 1990. The relation of gender understanding to children's sex-typed preferences and gender stereotypes. Child Development 61, 1427–1439.
Massa, L.J., Mayer, R.E., Bohon, L.M., 2005. Individual differences in gender role beliefs influence spatial ability test performance. Learning and Individual Differences 15, 99–111.
National Science Foundation, 2010. Table 5–7. Bachelor's Degrees Awarded, by Race/ethnicity, Citizenship, Sex, and Field: 2010. Retrieved 1 November 2013, from: http://www.nsf.gov/statistics/wmpd/2013/pdf/tab5-7.pdf.
Perrin, P.B., Heesacker, M., Tiegs, T.J., Swan, L.K., Lawrence Jr., A.W., Smith, M.B., Carrillo, R.J., Cawood, R.L., Mejia-Millan, C.M., 2011. Aligning Mar and Venus: the social construction and instability of gender differences in romantic relationships. Sex Roles 64 (9–10), 613–628.
Piaget, J., 1977. The Development of Thoughts: Equilibration of Cognitive Structures. Viking, New York.
Pressley, M., McCormick, C.B., 2007. Child and Adolescent Development for Educators. Guilford Press, New York.
Rhoads, S., 2010. Biology contributes to difference between male and female behaviors. In: Miller, K. (Ed.), Male and Female Roles. Gale: Cengage Learning, New York, pp. 22–40.
Schmitt, D.P., 2005. Sociosexuality from Argentina to Zimbabwe: a 48-nation study of sex, culture, and strategies of human mating. The Behavioral and Brain Sciences. 28 (2), 247–275.
Silberstang, J., 2011. Learning gender: the effects of gender-role stereotypes on women's lifelong learning and career advancement opportunities. In: London, M. (Ed.), The Oxford Handbook of Lifelong Learning. Oxford University Press, New York, pp. 370–388.
Sirin, S.R., McCreary, D.R., Mhalik, J.R., 2004. Differential reactions to men and women's gender role transgressions: perceptions of social status, sexual orientation, and value dissimilarity. The Journal of Men's Studies 12, 119–132.
Spencer, S.J., Steele, C.M., Quinn, D.M., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35, 4–28.
Tommessen, S., Todd, B., April 2000. Revising sex differences in play: very early evidence of stereotypical preferences in infancy. In: Paper Presented at the Annual Meeting of the British Psychological Society. Standford-upon-Avon, UK.
Thommessen, S., Todd, B., April 2010. Revisiting sex differences in play: very early evidence of stereotypical preferences in infancy. Paper Presented at the Annual Meeting of the British Psychological Society. Stratford-upon-Avon, UK.
Vandello, J.A., Bosson, J.K., 2013. Hard won and easily lost: a review and synthesis of theory and research on precarious manhood. Psychology of Men & Masculinity 14 (2), 101–113.
Wood, J.T., 2011. Gendered Lives: Communication, Gender, and Culture. Wadsworth Cengage Learning, Boston, MA.
Wood, E., Desmarais, S., Gugula, S., 2002. The impact of parenting experience on gender stereotyped toy play of children. Sex Roles 47, 413–422.
Zahn-Waxler, C., Polanichka, N., 2004. All things interpersonal. In: Putallaz, M., Bierman, K.L. (Eds.), Aggression, Antisocial Behavior, and Violence among Girls: A Developmental Perspective. Guilford Publications, New York, pp. 48–68.
Zentner, M., Mitura, K., 2012. Stepping out of the caveman's shadow: nations' gender gap predicts degree of sex differentiation in mate preferences. Psychological Science 23 (10), 1176–1185.

Sex Selective Abortion

Daniel Goodkind, Independent Researcher, Arlington, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sex selective abortion refers to the decision to terminate a pregnancy based on the sex of the fetus. In societies that have son preference, sex selective abortion is used more commonly after a fetus is found to be female (rather than male), which often results in a sex ratio at birth that is excessively masculine. Several societies have evinced rising sex ratios at birth in recent decades. This article reviews the factors that contribute to sex selective abortion and describes the major demographic and social characteristics of parents who use it. This article also discusses the dynamics of sex discrimination across the life course as well as possible future implications of sex imbalances among adults.

Introduction

Advances in prenatal testing technologies in recent decades have allowed parents to identify the sex of a fetus at an early stage of pregnancy. Sex selective abortion refers to the decision to terminate a pregnancy based on the sex of the fetus. In societies that have son preference, sex selective abortion is used more commonly after a fetus is found to be female (rather than male), which often results in a sex ratio at birth that is excessively masculine. Those distorted sex ratios are then carried forward as these birth cohorts enter young adulthood.

Several societies have evinced rising sex ratios at birth in recent decades. The earliest examples were noted in East Asia during the late 1980s and early 1990s (Gu and Roy, 1995; Goodkind, 1995), although since then such distortions have been observed in parts of South Asia and West Asia (Attané and Guilmoto, 2007; UNFPA, 2012).

This article reviews some of the broad factors that contribute to sex selective abortion. It also notes other factors that may also contribute to sex ratio variation even in the absence of sex selection. This article then describes some demographic and social features of those exhibiting selective abortion, theories of discrimination within the life course, and possible future patterns and implications that might be expected.

Three Factors Contributing to Sex Selection

Researchers have identified three key factors that have led to sex selective abortion in several societies. These factors are listed below, each of which will be summarized in turn.

1. son preference – the desire for at least one son,
2. low or declining fertility – which reduces the chances of having at least one son, and
3. access to prenatal sex-testing technology (and abortion).

The first factor is son preference. In traditional societies, sons were often more valued by parents than daughters because sons provided greater benefits – either economic or social – to the family. This is particularly true in patrilineal societies, where brides tend to move in with grooms, which often implied that daughter's loyalties then belonged to her husband's family rather than her natal family. Thus, in traditional societies, having many sons may be equated with wealth. That said, since sex selective abortion tends to occur at low levels of fertility, a more appropriate way of describing son preference is that parents want to have at least one son.

The second factor is declining fertility. When parents prefer to have fewer children, the number of slots available to have a son declines. Since the chances of having a son are just over 50% at each birth, at low levels of fertility, the concern for having a son at any given birth intensifies. Consider, for instance, that when parents have four children, only about 6% of families would not have a son by chance. However, when parents have only two children, about one-quarter of families would not have a son by chance. Prior research has suggested that 'replacement' levels of fertility (e.g., a total fertility rate, TFR, of 2.1 per woman), is a critical threshold for intensified concerns when parents desire at least one son. Yet at ever lower levels of fertility, the intensified concerns due to demographic change are typically counterbalanced by declining proportions of parents who want at least one son.

The third factor is access to both effective prenatal sex-testing technology as well as abortion services. Even if parents have son preference and their concerns for sons are intensified due to lower fertility, they cannot effectively choose the sex of their child in the absence of such technologies. Major advances in prenatal testing technologies around the 1980s (Bennett, 1983) allowed for the testing of fetal sex, and further advances permitted testing earlier within pregnancy, including within the first trimester. These advances have thus played perhaps the crucial role in the masculinization of sex ratios in many societies.

Other Sources of Variation in Sex Ratios at Birth

The 'normal' range of sex ratios at birth is generally considered to be anywhere between 103 and 107 male births per 100 females (Hill and Upchurch, 1995). The excess of male births is thought to reflect either the greater likelihood of male conceptions and/or the greater likelihood that male fetuses will survive during the prenatal period. Yet given that the range which is considered normal is fairly wide, even in the absence of prenatal testing technologies, other factors contribute to such variation. To cite just one example of variation within a particular society over time, Japan's sex ratio at birth gradually

 http://dx.doi.org/10.1016/B978-0-08-097086-8.31038-8

rose from about 104 in the earlier twentieth century to a peak above 107 in the mid-1960s, only to fall back closer to 105 more recently. The 1960s peak occurred well before reliable prenatal testing is thought to have existed.

Among the many theories proposed to explain such variation are biological ones, such as genetics and maternal age. Broader environmental or behavioral explanations include the timing of coitus within the menstrual cycle, diet, and even altitude. Of course, the evidence underlying some of these proposed theories may reflect statistical or reporting biases. For instance, ratios tend to have large statistical variation. A sample of 10 000 or more births may be required in order to draw valid conclusions about sex ratio differences. Similarly, among possible reporting biases, some societies may underreport female births and children more than male births and children (Goodkind, 2011). An additional consideration, in cases of interracial marriages the ethnic identification of the child may be different depending on whether it is a son or daughter. An appendix is included which summarizes the wide variety of explanations proposed for variations in the sex ratio at birth – no attempt is made here to assess the validity of each one.

Areas Where Sex Ratios Have Risen and Basic Trends

The demographic outcomes of sex selective abortion first became apparent in Asia during the 1980s, most notably in South Korea, where the sex ratio at birth rose from around 106–107 in the mid-1980s to a peak above 115 in the early 1990s (Gu and Roy, 1995). Since the 1990s, China (Goodkind, 2011) and India, the world's two population billionaires, joined this group. This was not surprising to many observers, given strong patrilineal patterns in these societies, as well as the three contributing factors noted early. Less expected were other Eurasian countries that have evinced sex imbalances, such as group, such as Armenia, Azerbaijan, and Georgia. What was surprising about the latter group is that they were not widely known for having strong son preference.

In son-preferring societies, sex ratios at birth differ sharply depending on birth order, as well as the sex distribution of earlier births. The likelihood of sex selection rises dramatically if a daughter is born first. Typically, sex ratios of first births are normal, while second and higher order births then show more distorted ratio. This is because, as mentioned earlier, parents want at least one son and within a smaller family size and may only go beyond their ideal number if they have yet to have a son. There is also evidence of sex selection against male fetuses – some parents may want not only at least one son but also at least one daughter. For parents wanting at least one daughter, after the birth of one or two sons, the sex ratio of the next child may be excessively feminine.

Recently, however, in their most recent censuses, both Vietnam (in 2009) and China (in 2010) showed a notable elevation in the sex ratio of first births. Given the prior evidence that sex selection is limited to second and higher order births, this has come as a surprise. Why the very recent change? One likely explanation concerns the crisscrossing social forces identified earlier. As fertility declines, intensified concerns over reduced slots for having a son are usually counterbalanced by a decline in son preference overall, yet if some proportion of son-preferring parents now truly want only one child, sex selection will occur for the first birth. China's one-child policy would not appear to explain this very recent trend, since that policy had been in place for three decades without any appreciable increase in the sex ratios of first births.

The Substitution Hypothesis and the Role of Technology

Historical studies have documented evidence of both son preference and daughter discrimination in traditional societies that lacked access to prenatal testing technologies. Discrimination in such societies could be manifested through excess mortality of female infants or children (or even female infanticide). One key question is whether sex selective abortion simply adds to these ongoing traditional forms of discrimination or whether its use 'substitutes' for them and thus leads them to decrease. Comparisons of aggregate sex ratio measures of prenatal and postnatal discrimination over time suggest the possibility of both dynamics, but are inconclusive (Goodkind, 1995). As for survey data among individuals, some support for substitution has been found in Taiwan (Lin et al., 2014).

That said, advocates of abortion rights often argue that strategic substitution by parents does exist (Goodkind, 1999), and nearly all researchers seem to agree with a weaker formulation of the substation hypothesis – that prenatal sex-testing technologies have allowed parents to focus their concerns about the sex composition of children during the prenatal period. This weaker formulation may also shed light on the role of technology on son preference. As mentioned earlier, societies such as Azerbaijan and Armenia have recently exhibited large distortions in sex ratios at birth, yet neither society was well known in the past for son preference. It is possible, then, that prenatal sex-testing technologies have made sex discrimination in family formation acceptable in areas where traditional forms of discrimination (e.g., female infanticide) had not been.

Future Trends

For societies exhibiting sex selection, a variety of future trends in the sex ratio at birth have been considered (Attané and Guilmoto, 2007):

1. They continue to rise.
2. They rise, yet thereafter return to normal, as in the case of South Korea.
3. They remain at their current elevated levels.
4. The fall, yet not back to normal levels.
5. They oscillate back and forth.

Although it is too early to tell for sure, empirical evidence suggests that first scenario is unlikely. Sex ratios tend to reach a peak, after which they fall back, although perhaps not back to normal. Different societies may exhibit different patterns.

The Bachelors

We conclude by considering the future. Whether or not sex ratios at birth return to normal, what are some of the possible

social outcomes in societies where child sex ratios have become increasingly masculine? We consider below some of these possible outcomes, many have of which have been well identified (Gutentag and Secord, 1983; Hudson and den Boer, 2004). Although some of these possible outcomes are hotly contested, one outcome is certain – as child cohorts with large sex imbalances mature, there will eventually be large sex imbalances among young adults.

Perhaps the most discussed outcome of sex imbalances involves marriage squeeze. When excessively masculine birth cohorts reach the age of marriage, larger proportions of males will marry at later ages, or not at all. Moreover, in societies where females tend to marry males of a higher social status, it is poor men who will suffer most, since there will be few who are beneath them. Of course, the famine for males is a potential feast for females. And the effect on female marriage patterns is not at all certain. Although a wide selection of husbands could encourage women to marry at younger ages, if they are spoiled for choice and/or want to focus on career options, they may decide to marry at later ages. Marriage squeeze will become a major research topic in coming decades. South Korea was the first country to show substantial distortions in sex ratios at birth by the late 1980s, and distorted birth cohorts should be reaching peak ages of marriage from now until about 2025.

Another possible outcome is that large cohorts of bachelors will bring about increased crime and/or violence. Several researchers suggest that this has already happened (Edlund et al., 2007), although explanation for such findings may have nothing to do with the frustration of young men unable to find a bride (or romantic female partner). Any rise in crime or violence might simply be due to the shifting composition of the population. If young males tend to behave badly (compared to young females) and are more likely to commit crimes, then larger proportions of young males in any society will result in more crime per capita.

Yet one cannot rule out that sex imbalances among adults could have profound social consequences. Consider, for instance, voting. In many societies, voting patterns among women tend to differ compared to men. Thus, to the extent that societies have voting rights, a larger proportion of males in the population could swing elections and referendums in favor of whatever males tend to prefer. And if males are more likely to favor, say, military interventions, then military interventions could be more likely.

Appendix – Summary of Proposed Explanations for Variations in Sex Ratios at Birth

- Biological
  - genes (including possible ethnic differences),
  - maternal age and/or prior children born, and
  - differential benefits by sex of improvements in maternal and fetal health
- Environmental/behavioral
  - altitude,
  - diet, and
  - timing of coitus within the menstrual cycle
- Statistical/reporting
  - small sample size,
  - sex-based differences in reporting completeness (e.g., still birth by sex), and
  - sex differences in identification in multiethnic societies (includes 'minority status' hypothesis).
- Sex selection
  - preconceptive sperm sorting,
  - prenatal testing followed by sex selective abortion, and
  - policies that affect use or outcomes of sex selection.

See also: Abortion, Demography of: United States; Assortative Mating in the Marriage Market; Children, Value of; Families and Households, Formal Demography of; Family Planning Programs: Feminist Perspectives; Fertility Transition: China; Fertility Transition: East Asia; Sexuality, Theories of.

Bibliography

Attané, Isabel, Guilmoto, Christophe, Z. (Eds.), 2007. Watering the Neighbour's Garden. The Growing Female Deficit in Asia. CICRED, Paris.

Bennett, Neil, 1983. Sex Selection of Children. Academic Press, New York.

Edlund, Lena, Li, Hongbin, Yi, Junjian, Zhang, Junsen, 2007. "Sex Ratios and Crime: Evidence from China's One-child Policy", IZA Discussion Paper 3214. Institute for the Study of Labor, Bonn.

Gutentag, Marcia, Secord, Paul F., 1983. Too Many Women? the Sex Ratio Question. Sage Publications, Beverly Hills, CA.

Goodkind, Daniel, 2011. Child underreporting, fertility, and sex ratio imbalance in China. Demography 48 (1), 291–316.

Goodkind, Daniel, 1999. Should prenatal sex selection be restricted: ethical questions and their implications for research and policy. Population Studies 5, 49–61.

Goodkind, Daniel, 1995. On substituting sex preference strategies in East Asia: does prenatal sex selection reduce postnatal discrimination? Population and Development Review 22, 111–125 (Birth in China with Reference).

Gu, Baochang, Roy, Krishna, 1995. Sex ratio at birth in China with reference to other areas of East Asia: what we know. Asia-Pacific Population Journal 3, 17–42.

Hill, Kenneth, Upchurch, Dawn M., 1995. Gender differences in child health: evidence from the demographic and health surveys. Population and Development Review 21 (1), 127–151.

Hudson, Valerie M., den Boer, Andrea M., 2004. Bare Branches: The Security Implications of Asia's Surplus Male Population. MIT Press, Cambridge, MA.

Lin, Ming-Jen, Liu, Jin-Tan, Qian, Nancy, 2014. More missing women, fewer dying girls: the impact of sex selective abortion on sex at birth and relative female mortality in Taiwan. Journal of the European Economic Association 12 (4), 899–926.

UNFPA (United Nations Fund for Population Activities), 2012. Sex Imbalances at Birth: Current Trends, Consequences, and Policy Implications. UNFPA Asia and the Pacific Regional Offices, Bangkok.

Sex Therapy, Clinical Psychology of

Walter Everaerd, University of Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychological interventions, which have become known as sex therapy, have been designed to reverse the symptoms of sexual dysfunction. The research of Masters and Johnson inspired many approaches to sexual problems. Their main focus in therapy is on adequate sexual stimulation and the prevention of fear of failure and performance anxiety. Most of the evidence for the efficacy of sex therapy was produced 20 years ago. A recent meta-analysis failed to show robust evidence-based practice, except for female desire and female orgasmic disorder.

Introduction

Sex therapy is an approach to the treatment of sexual problems. Results of psychophysiological studies of sexual responses in sexually functional subjects in the 1960s allowed gynecologist William Masters and psychologist Virginia Johnson to develop a therapeutic format for the treatment of sexual inadequacy (1970). In this format, sexual problems were openly discussed, there was no time-consuming delving-into-past conflicts, and there were suggestions for a direct reversal of symptoms of sexual failure. Masters and Johnson preferred working with couples, because the interaction within the couple often contributes in important ways to the sexual difficulties. Other therapists have offered treatment to individuals and to groups of individuals or couples. Several other approaches developed from behavior therapy and rational emotive therapy and focused on performance anxiety and fear of failure. Others used interventions from couple and group therapy (Balon and Segraves, 2005; Leiblum, 2007). It is fair to say that nowadays almost all approaches to sexual difficulties incorporate elements from the Masters and Johnson sex therapy format.

Apart from organic causes, sexual problems may be attributed to interpersonal difficulties and to problems with emotional functioning. Masters and Johnson were convinced that 'adequate' stimulation would result in sexual response in an engaged partner. Adequate stimulation feels good, is rewarding, and facilitates focus of attention on those feelings. Originally, adequate stimulation in therapies was restricted to behaviors comprising the classical heterosexual foreplay, coitus, and orgasm sequence (Masters and Johnson, 1964). This criterion for adequate stimulation has been abandoned because of the rich diversity in sexual practices that exist in the real world.

Masters and Johnson hypothesized that fear of failure, and as a consequence becoming a spectator of one's own feelings, is the most important cause for sexual inadequacy. They devised a very ingenious procedure to assist people in becoming engaged in feelings of sexual excitement. Three important steps were specified. In the first step, people learn to accept that mutual touching feels good. In this step, demands for sexual performance, and anxieties related to such demands, are precluded by an instruction not to touch genitals or secondary sex characteristics (e.g., breasts). When the first step leads to acceptance of positive bodily feelings, the second step demands including genitals and breasts in mutual touching and caressing. Masters and Johnson have suggested some variations to accommodate specific sexual problems (e.g., premature ejaculation, erectile problems). The second step has to result in the physiological signs of sexual excitement – vaginal lubrication in women and erection of the penis in men. In the classical sequence, the first and second steps prepared for the third step, which consisted of coital positions, again with variations, and stimulation through coitus to orgasm. Although coitus may be relevant from a reproductive point of view, it is not always and certainly not the only way to experience the ultimate pleasures of sex (Laumann et al., 1999).

Desire, Excitement, and Orgasm

Masters and Johnson, on the basis of their psychophysiological studies, proposed a model of sexual response consisting of three phases: an excitement phase, an orgasm phase, and a resolution phase (Masters and Johnson, 1964). The excitement phase and orgasm phase may be recognized as the second and third step of the classic treatment format. When other therapists began to apply Masters and Johnson's format, it became clear that many patients do not easily engage in interactions prescribed by the sex therapy format. Apparently, an initial motivational step was missing. People who hesitate or avoid intimate interactions may lack desire, which often means that they do not expect the interaction to be rewarding. In 1979, Helen Kaplan proposed adding a desire phase, preceding the three phases specified by Masters and Johnson. Since Kaplan's proposal it has become clear that the prevalence of lack of desire is considerable (Kaplan, 1995). With hindsight, most people accept that lack of desire must be the most important sexual problem. It is a problem for the individual who does not arouse desire in his or her partner. In most instances, not feeling desire is in itself unproblematic, but lack of sexual desire may become a problem in the relationship.

Sex therapy was a fresh and new treatment. Sexual problems were openly discussed, there was no time-consuming delving-into-past conflicts, and there were suggestions for a direct reversal of symptoms of sexual failure. Masters and Johnson preferred working with couples because the interaction within the couple often contributes in important ways to the sexual difficulties. Other therapists have offered treatment to individuals and to groups of individuals or couples.

An alternative to Masters and Johnson's therapy format is the mimicking of normal sexual development through the use of masturbation (Heiman et al., 1976). This has been an important step for many women, especially those who missed this aspect of discovery of their own sexuality. In this approach, people learn to induce sexual excitement through masturbation to eventually apply this 'skill' in interaction with a partner.

Some approaches developed from cognitive behavioral therapy and rational emotive therapy focused on performance anxiety and fear of failure. Others used interventions from couple and group therapy. It is fair to say that nowadays almost all approaches to sexual difficulties incorporate elements from the Masters and Johnson sex therapy format.

Sexual Dysfunctions in Men

Diagnostic Procedures

The aim of the initial clinical interview is to gather detailed information concerning current sexual functioning, onset of the sexual complaint, and the context in which the difficulty occurred. This information gathering may be aided by the use of a structured interview and paper-and-pencil measures regarding sexual history and functioning. An individual and conjoint partner interview, if possible, can provide additional relationship information and can corroborate data provided by the patient. The initial clinical interview should help the clinician in formulating the problem. It is important to seek the patient's agreement with the therapist's formulation of the problem. When such a formulation is agreed upon, the problem may guide further diagnostic procedures.

Many men with erectile dysfunction may be wary of psychological causes of their problem. Psychological causes seem to imply that the man himself is responsible for his problem. This may add to the threat to his male identity that he is already experiencing by not being able to function sexually. Considering the way a man may experience his problem, it can be expected that it will not be easy to explain to him the contribution of psychological factors. A clinician knowledgeable in biopsychosocial aspects of sexual functioning should be able to discuss the problem openly with the patient. Dysfunctional performance is meaningful performance in the sense that misinformation, emotional states, and obsessive concerns about performance provide information about the patient's 'theory' of sexual functioning. When contrasting this information with what is known about variations in adequate sexual functioning, it is often clear that one cannot but predict that the patient must fail. For the clinician, a problem arises when, even with adequate stimulation and adequate processing of stimulus information according to the clinician's judgment, no response results, either at a physiological or a psychological level (Janssen and Everaerd, 1993).

At this point, a number of assessment methods aimed at identifying different components or mechanisms of sexual functioning may be considered. In principle, two main strategies may be followed: In the first, although a psychological factor interfering with response cannot be inferred from the report of the patient, one can still suspect some psychological factor to be at work. Possibly, the patient is not aware of this factor, thus he cannot report on it. Eliminating this psychological influence may result in adequate response. The second strategy applies when even with adequate (psychological) stimulation and processing, responding is prevented by physiological dysfunction. Physiological assessment may then aid in arriving at a diagnostic conclusion. The biopsychosocial approach predicts that it is inadequate to choose one of these strategies exclusively. The fact that sexual functioning is always psychophysiological functioning means that there may always be an unforeseen psychological or biological factor.

Psychological Treatments of Sexual Dysfunctions in Men

Approaches to Treatment

The most important transformation of the treatment of sexual dysfunctions occurred after the publication of Masters and Johnson's (1970) *Human Sexual Inadequacy*. First of all, they brought sex into the treatment of sexual problems. Before the publication of their seminal book, sexual problems were conceived as consequences of (nonsexual) psychological conflicts, immaturity, and relational conflicts. In most therapies for sexual problems, sex was not a topic in the therapeutic transactions. There were always things 'underlying', 'behind,' and 'besides' the sexual symptoms that deserved discussion. Masters and Johnson proposed to attempt directly to reverse the sexual dysfunction by a kind of graded practice and focus on sexual feelings. If sexual arousal depends directly on sexual stimulation, that very stimulation should be the topic of discussion. Here the second important transformation occurred: A sexual dysfunction was no longer something pertaining to an individual; rather, it was regarded as a dysfunction of the couple. It was assumed they did not communicate in a way that allowed sexual arousal to occur when they intend to 'produce' it. Masters and Johnson thus initially considered the couple as the 'problem' unit. Treatment goals were associated with the couple concept: The treatment goal was orgasm through coital stimulation. This connection between treatment format and goals was lost once Masters and Johnson's concept was used in common therapeutic practice. People came in for treatment as individuals. Male orgasm through coitus adequately fulfills reproductive goals, but it is not very satisfactory for many women because they do not easily achieve orgasm through coitus. What have remained over the years, since 1970, are a direct focus on dysfunctional sex and a focus on sexual sensations and feelings as a vehicle for reversal of the dysfunction.

What Masters and Johnson tried to achieve in their treatment model is a shift in their patients' focus of attention. Let us look at one of Masters and Johnson's interventions to elucidate this point. People with sexual dysfunctions tend to wait and look for the occurrence of feelings, instead of feeling what occurs – hence, the spectator role. Their attention is directed toward something that is not there or does not exist, which is frustrating. In simplest form, Masters and Johnson propose to redirect attention by using the following steps: first of all they manipulate expectations by instructing the patient about what is allowed to occur and what is not. It is explained to the patient that nonsexual feelings are to be accepted as a way to accept sexual feelings later on, and therefore sexual areas are excluded

in the initial homework tasks. From a psychological point of view, this manipulation is ingenious; it directs attention away from sex – when you feel a caress on your arm it may be pleasant but (now) it is not sexual – however, at the same time, it defines sexual feelings as feelings in 'sexual areas.'

To attain a direct approach of sexual function, numerous variants of couple, communication, and group therapy have been used. Rational emotive therapy has been used to change expectation and emotions. To remedy biographical memories connected with sexual dysfunction, psychoanalytic approaches have been used as well as cognitive behavior therapy approaches. There are specific interventions for some dysfunctions; for example, premature ejaculation has been treated with attempts at heightening the threshold for ejaculatory release (stop–start or squeeze techniques) (Balon and Segraves, 2005; Leiblum, 2007; Kolodny et al., 1979).

Recently, as a spin-off of research into cardiac vascular smooth muscle pharmacology, drugs have become available which act by relaxing smooth muscles in the spongiose and cavernous structures in the penis. This relaxation is necessary to allow blood flow into the penis, thus causing an erection. Some of these drugs (e.g., sildenafil – Viagra®) support the natural neurophysiological reaction to sexual stimuli. Others act locally in the penis without sexual stimulation. To slow down speed of ejaculation in men with premature ejaculation, serotonin selective reuptake inhibitors (SSRIs) seem to be helpful (Waldinger et al., 1998).

Smooth muscle relaxants like sildenafil are helpful in older men when less naturally occurring transmitters are available. Men with vascular or neurodegenerative diseases (e.g., diabetes, multiple sclerosis) may also benefit from the use of smooth muscle relaxants (Boolell et al., 1996). Although these drugs are very effective, they do not help every man with erectile problems.

Pharmacological treatment for erectile disorder may be an important step in restoring sexual function (Shamloul and Ghanem, 2013). Most couples will need information and advice to understand what they may expect from this treatment. For many, it will not bring the final resolution of their relationship problems. In addition to drug treatment, they will need some form of sex therapy or psychotherapy.

Validated Treatments for Male Sexual Dysfunctions

It has been difficult to get an overview of treatments for sexual dysfunctions because any proposal about how to approach dysfunctions was valid. This has changed through the introduction of criteria for validated or evidence-based practice by the American Psychological Association (APA Presidential Task Force on Evidence-Based Practice, 2006). From a systematic review and meta-analysis of psychotherapies for male sexual dysfunctions, it appears that the state-of-the-art is far from satisfactory (Frühauf et al., 2013). Randomized controlled trials (RCTs) did not show clear evidence for Erectile Dysfunction and Premature Ejaculation. Male Dyspareunia and Male Orgasmic Disorder were never studied in RCTs. Berner and Günzler (2012), although less strict about inclusion of studies, were more optimistic about the efficacy of interventions. Although the evidence-based practice movement should be firmly supported, unqualified support would be disastrous for the practice of the treatment of sexual problems. The care for patients with sexual problems must be continued even without proof according to the rules of 'good clinical practice.' The sensible clinician will learn to be very careful about any claims concerning either diagnostic procedures or treatments, and should support the science that is the background for her or his actions.

Sexual Dysfunctions in Women

Diagnostic Methods

Similar to the procedures in men, initial interviews should help the clinician in formulating the problem and in deciding whether sex therapy is indicated. Since sexual problems can be a consequence of sexual trauma, it is necessary to ask if the woman ever experienced sexual abuse. An important issue is the agreement between therapist and patient about the formulation of the problem and the nature of the treatment. To reach a decision to accept treatment, the patient needs to be properly informed about what the diagnosis and the treatment involve and what is known about the efficacy of the proposed therapy. Depending on the nature of the complaint, the initial interviews may be followed by medical assessments. In contrast to the assessment of men, the psychobiological assessment of women's sexual problems is not well developed (Laan and Everaerd,1995).

Psychological Treatments of Sexual Dysfunctions in Women

Approaches to Treatment

Similar to men, the treatment of sexual dysfunction in women contains many elements from Masters and Johnson sex therapy. As noted before, an important addition, especially in women, is the use of masturbation to discover their own sexuality.

Low sexual desire is generally treated with sensate focus exercises to minimize performance pressure, and communication training. In the treatment of sexual aversion, the focus is on decreasing anxiety, the common core of sexual aversions. Behavioral techniques, like exposure, are most commonly used.

Treatment of sexual arousal disorder generally consists of sensate focus exercises and masturbation training, with the emphasis on becoming more self-focused and assertive in asking for adequate stimulation (Masters and Johnson, 1970).

For the treatment of primary (lifelong) anorgasmia there exists a well-described treatment protocol (Heiman et al., 1976). Basic elements of this program are education, self-exploration and body awareness, and directed masturbation. Because of the broad range of problems behind the diagnosis of secondary (not lifelong) and situational anorgasmia, there is no major treatment strategy for this sexual disorder. Depending on the problem, education, disinhibition strategies, and assertiveness training are used (see Frühauf et al., 2013). It is important to identify unrealistic goals for treatment like achieving orgasm during intercourse without clitoral stimulation.

For dyspareunia (genital pain often associated with intercourse), there are multiple possible somatic and psychological causes. A common picture is vulvar vestibulitis, pain at small inflamed spots at the lower side of the vaginal opening. However, often there is no clear organic cause for the pain.

Treatment should be tuned to the specific causes diagnosed and can vary from patient to patient. Behavioral interventions typically include prohibition of intercourse, finger exploration of the vagina, first by the woman, then by her partner. Sensate-focus exercises may be used to increase sexual arousal and sexual satisfaction. Pelvic floor muscle exercises and relaxation training can be recommended in case of vaginismus or a high level of muscle tension in the pelvic floor.

Treatment of vaginismus commonly involves exposure to vaginal penetration by using dilators of increasing size or the women's fingers. Pelvic floor muscle exercises may be used to provide training in discrimination of vaginal muscle contraction and relaxation, and to teach voluntary control over muscle spasm.

Pharmacological treatment of sexual disorders of women is just beginning. Smooth muscle relaxants have been used in women to ameliorate sexual arousal and as a consequence hypoactive sexual desire. It appears that drugs like sildenafil produce smooth muscle relaxation and increased genital blood flow, but they have no effect on the subjective experience of sexual response. In women with subphysiological levels of testosterone – mainly in the postmenopause – testosterone patches appear to have an effect on mood, energy, and libido.

Validated Treatments for Women's Sexual Dysfunctions

Reviews of treatments for sexual dysfunctions in women following the criteria for validated or evidence-based practice have been published (Heiman and Meston, 1997; Baucom et al., 1998; Günzler and Berner, 2012; Frühauf et al., 2013). Heiman and Meston conclude that treatments for primary anorgasmia fulfill the criteria of 'well-established,' and secondary anorgasmia studies fall into the 'probably efficacious' group. They conclude with some reservations that vaginismus appears to be successfully treated if repeated practice with vaginal dilators is included in the treatment. Their reservations are due to a lack of controlled or treatment comparison studies of vaginismus. All authors conclude that adequate data on the treatment of sexual desire disorder, sexual arousal disorder, and dyspareunia are lacking. Frühauf et al. (2013) endorse the conclusions of Heiman and Meston and note that "according to our results the situation has not changed much" (p. 929).

Although the evidence-based practice movement deserves support, care for patients with sexual problems must be continued even without proof according to the rules of 'good clinical practice.'

Future Directions

Sex therapy bloomed in the 1970s and 1980s, but reviews of evidence-based treatments suggest that developments stagnated and very few new studies have been undertaken. The recent shift to biological approaches will continue, at least for a while. Viagra and testosterone patches will shortly be followed by more centrally acting drugs (e.g., dopamine agonists). The search for drugs has provoked a wide range of studies into the biological basis of sexual function. This work inspires behavioral and cognitive neuroscience studies, which may provide a framework and new tools to better understand sexual emotions and sexual motivation.

See also: Occupational Health; Sexual Dysfunction, Epidemiology of; Sexual Motivation; Sexuality Over the Life course.

Bibliography

APA Presidential Task Force on Evidence-Based Practice, 2006. Evidence-based practice in psychology. American Psychologist 61, 271–285.

Balon, R., Segraves, R.T. (Eds.), 2005. Handbook of Sexual Dysfunction. Taylor & Francis, Boca Raton, FL.

Baucom, D.H., Shoham, V., Mueser, K.T., Daiuto, A.D., Stickle, T.R., 1998. Empirically supported couple and family interventions for marital distress and adult mental health problems. Journal of Consulting and Clinical Psychology 66, 53–88.

Berner, M., Günzler, C., 2012. Efficacy of psychosocial interventions in men and women with sexual dysfunctions—a systematic review of controlled clinical trials. Part 1—the efficacy of psychosocial interventions for male sexual dysfunction. Journal of Sexual Medicine 9, 3089–3107.

Boolell, M., Allen, M.J., Ballard, S.A., Gepi-Attee, S., Muirhead, G.J., Naylor, A.M., Osterloh, I.H., Gingell, C., 1996. Sildenafil: an oral active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. International Journal of Impotence Research 8 (2), 47–52.

Frühauf, S., Gerger, H., Schmidt, H.M., Munder, T., Barth, J., 2013. Efficacy of psychological interventions for sexual dysfunctions: a systematic review and meta-analysis. Archives of Sexual Behavior. http://dx.doi.org/10.1007/s10508-012-0062-0.

Günzler, C., Berner, M., 2012. Efficacy of psychosocial interventions in men and women with sexual dysfunctions—a systematic review of controlled clinical trials. Part 2—the efficacy of psychosocial interventions for female sexual dysfunctionism. Journal of Sexual Medicine 9, 3108–3125.

Heiman, J., LoPiccolo, L., LoPiccolo, J., 1976. Becoming Orgasmic: A Sexual Growth Program for Women. Prentice-Hall, Englewood Cliffs, NJ.

Heiman, J.R., Meston, C.M., 1997. Evaluating sexual dysfunction in women. Clinical Obstetrics and Gynecology 40, 616–629.

Janssen, E., Everaerd, W., 1993. Determinants of male sexual arousal. Annual Review of Sex Research 4, 211–245.

Kaplan, H.S., 1995. The Sexual Desire Disorders: Dysfunctional Regulation of Sexual Motivation. Brunner/Mazel, New York.

Kolodny, R.C., Masters, W.H., Johnson, V.E., 1979. Textbook of Sexual Medicine, first ed. Little, Brown, Boston, MA.

Laan, E., Everaerd, W., 1995. Determinants of female sexual arousal: psychophysiological theory and data. Annual Review of Sex Research 6, 32–76.

Laumann, E.O., Paik, A., Rosen, R.C., 1999. Sexual dysfunction in the United States: prevalence and predictors. Journal of the American Medical Association 281, 537–544.

Leiblum, S.R. (Ed.), 2007. Principles and Practice of Sex Therapy. The Guilford Press, New York.

Masters, W.H., Johnson, V.E., 1964. Human Sexual Response. Little, Brown, Boston, MA.

Masters, W.H., Johnson, V.E., 1970. Human Sexual Inadequacy, first ed. Little, Brown, Boston, MA.

Shamloul, R., Ghanem, H., 2013. Erectile dysfunction. Lancet 381, 153–165.

Waldinger, M.D., Berendsen, H.H.G., Blok, B.F.M., Olivier, B., Holstege, G., 1998. Premature ejaculation and serotonergic antidepressants-induced delayed ejaculation: the involvement of the serotonergic system. Behavioural Brain Research 92, 111–118.

Sexism

Fabio Lorenzi-Cioldi and Clara Kulich, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article first reviews the historical rise and the developments of psychological research on sexism which led to the differentiation between biological sex and cultural gender, and highlights commonalities between gender relations and other status-based intergroup relations. Old-fashioned, overt and explicit forms of sexism are then distinguished from contemporary, more subtle, and implicit forms. The latter are characterized by ambivalence of hostility and benevolence toward women. The final section discusses major behavioral consequences of the different forms of sexism, emphasizing the harmful impact of benevolence on women's feelings of self-worth and competence, and actual performances.

Introduction

Historical Roots: The 'Discovery' of Sexism

Unequal treatment of individuals based on their sex is probably as old as humankind. However, the social recognition that unequal treatment is discriminatory, and the ensuing notion of 'sexism,' is far more recent. Long before gender inequality was recognized as a societal problem, early feminists such as Christine de Pizan in the fifteenth century in France, Mary Wollstonecraft in the seventeenth century in Britain, and more recently Simone de Beauvoir, only to name a few, denounced the arbitrariness and the illegitimacy of the hierarchical organization of gender differences in society. But it was not until a speech on 'Women and the undergraduate' by Pauline M. Leet at an US university in 1965 that the term sexism entered lay people's vocabulary, and became a matter of debates. Leet explained this notion by establishing a parallel with racism, a construct that people were already familiar with:

> When you argue … that since fewer women write good poetry this justifies their total exclusion, you are taking a position analogous to that of the racist — I might call you in this case a 'sexist' … Both the racist and the sexist are acting as if all that has happened had never happened, and both of them are making decisions and coming to conclusions about someone's value by referring to factors which are in both cases irrelevant.
>
> p. 3: quoted in Shapiro, 1985: p. 6

Sexism is a form of prejudice or discriminatory treatment based on a person's sex. The coinage of this term occurred in concomitance with the realization that the persistent social disadvantage of women urged a collective response. Until World War Two, the term 'minority' typically referred to the numerical inferiority of national and linguistic groups. Only in the 1950s its meaning made a significant shift to label groups that are inferior in status, or are stigmatized, such as women, people with disabilities, or the elder. Since then, the notion of minority has become "a general term for all groups subjected to prejudice and discrimination" (Meyers, 1984: p. 1). This was also a starting point for the study of prejudice (e.g., Allport, 1954) and its underlying motivations originating in intergroup relations (e.g., Tajfel, 1981).

When interest in studying sex differences began to grow, sexist ideology was embedded in the questions asked. Society and researchers until the beginning of the twentieth century assumed that physiological and biological characteristics such as males' bigger brain size and physical strength were evidence of men's superiority. Early research used such characteristics to examine their role in predicting intelligence. But extensive research by Terman and his collaborators could not show any evidence of consistent differences in intelligence between men and women, which resulted in the revision of intelligence tests based on this assumption (Terman and Merrill, 1937). Many subsequent empirical studies indulged in identifying and quantifying sex differences in behavior and personality. These efforts were summarized and interpreted by Maccoby and Jacklin in their authoritative book *The Psychology of Sex Differences* (1974). Their comprehensive summary of the literature led them to conclude that most beliefs about existing sex differences, such as 'girls are more sociable than boys' or 'boys are better at cognitive tasks,' were not supported by empirical research. Overall, this research showed more similarities than differences between the sexes. Maccoby and Jacklin, nonetheless, acknowledged the existence of specific differences such as men's superiority in math and spatial aptitudes, and women's superiority in verbal tasks. However, instead of advocating causes related to innate characteristics of the sexes, they suggested that such differences could be adequately accounted for by socialization processes and stereotypes about women and men. As a consequence, further research in the social sciences began to focus on learning processes, rather than biological sex, as factors influencing the observed sex differences.

Maccoby and Jacklin's work was criticized due to its narrative approach. It was not until recently that Hyde (2005) conducted a more sophisticated analysis of this past and the subsequent literature. Despite the use of more rigorous meta-analytic methodologies, her analysis basically corroborated Maccoby and Jacklin's conclusions. The findings allowed for the conclusion that, overall, "Women and men are more alike than they are different" (p. 581), and that differences between the sexes are smaller than the differences that can be found within each sex. Though advocating a provocative *gender similarities hypothesis*, Hyde likewise mentioned exceptions, for instance motoric behaviors and sexuality, but most importantly differences in self-attributions of traits associated with

two key dimensions of stereotypes: *warmth* and *competence*. Women were shown to develop interdependent and relational self-construals, while men appeared to focus on independence and agency. As we know today, these differences derive from the gendered treatment of men and women. Contemporary research on sexism is highly influenced by the distinction between these two types of self-construals, and focuses on the negative consequences of applying them when distinguishing between men and women (e.g., Wood and Eagly, 2010).

The shift from biological approaches of sex differences to social and cultural ones urged a revision of concepts and terminology. Differences between men and women that could be accounted for by stable biological factors are since denoted with the term *sex*, while differences due to more malleable cultural factors are denoted with the term *gender*. An important discovery in empirical research was that social context moderates differences that were thought to be rooted in the biological sex, thus making the term gender more accurate. As Deaux and Major write "Now you see them, now you don't" (1987: p. 369), meaning that in certain contexts we observe differences in attitudes and behaviors between men and women, while in other contexts these differences vanish. A seminal experiment by Goldberg (1968) showed that evaluations of a piece of work (e.g., an article) varied as a function of the presentation of the author as female or male. Thus, it was not the actual competence of the article's author that led to perceptual and stable differences in the article's quality, but the observers' precognitions in the form of gender stereotypes. Although Goldberg's findings have suffered from a lack of replication, the idea of social context as an important moderator of sex differences crucially influenced the upcoming work.

The Social Status Dimension

In virtually all countries around the world, men are in socially dominant positions relative to women. From this observation, and drawing on the evidence of contextual effects on sex differences, most work on sexism has focused on the status differentials between men and women. It is of no surprise that the mere use of the term 'sexism' conveys the idea of discrimination against women, not men. Although there are incidences of sexism against men, that is *reverse sexism*, the consequences of gender discrimination are usually psychologically more harmful for the powerless. References to sexism against men are seldom. The *Ambivalence Toward Men Inventory* (Rudman and Glick, 2008) is one of the few attempts to measure sexist beliefs against men.

Despite differences between the gender status system and other long-term status systems, notably ethnicity, age, or social class, the study of status hierarchies has provided valuable insights for understanding how gender operates at a more general level (see Lorenzi-Cioldi, 2006; Ridgeway, 2011). The assumption that men have higher status and greater power than women is central to many social psychological analyses of sexism (Wood and Eagly, 2010). Stereotypic beliefs about men and women are firmly rooted in society's gendered division of labor, and therefore in the social structure (Guimond et al., 2013; Eagly, 1987). In general, less prestigious roles and occupational domains (e.g., nurse) ask for communal behavior, whereas more prestigious ones (e.g., bank trader) ask for agentic behavior. For instance, Johannesen and Eagly (2002) demonstrated that participants attribute higher scores of positive and negative agency to people, rather than communion, as a function of people's professional status. The diffuse knowledge that men and women are unevenly distributed in the occupational hierarchy boosts the belief that, in actuality, men *are* agentic and women *are* communal. Accordingly, perceivers come to believe that men and women possess the personality characteristics that are required to perform their respective social roles. Overall, then, the empirical evidence shows that men and women are attributed traits and characteristics deriving from implicit assumptions about their social status, overriding to a large extent the role of biological sex. In a seminal demonstration of this central hypothesis of *social role theory*, Eagly and Wood showed that, knowing only the gender of the protagonist of an influence scenario, participants inferred higher status for men than for women. In contrast, when both genders were portrayed in an ostensibly similar occupational role, the impact of gender stereotypes was reduced (see Eagly, 1987). Likewise, Lorenzi-Cioldi demonstrated this preeminence of status cues in the interpretation of male and female behavior in a study in which participants rated two managers and two employees of each sex on agentic and communal traits. Participants matched the traits according to the targets' professional roles, rather than their sex, judging both male and female managers as more agentic than the employees (see Guimond et al., 2013).

A host of analyses of secondary data provides further support to this social role hypothesis. To illustrate, Twenge (2001) examined assertiveness among women using survey data gathered from 1931 to 1993. She found that levels of this agentic personality trait varied according to historical periods, with higher levels corresponding to women's increased access to the labor market. Current societal trends toward more representative numbers of women in higher professional strata allow for the prediction that differences in judgments about men and women tend to fade away (see Diekman and Eagly, 2000). In particular, the greater variety of social roles enacted by women should be conducive to beliefs of an increased agency of women's personalities.

Types of Sexism

As we mentioned in the former section, sexist attitudes are rooted in gender stereotypes that ascribe higher status to men than to women. Thus, the concept and the measurement of sexism are largely concerned with gender asymmetries disadvantaging women. The contemporary literature lingers over a distinction between overt and more subtle forms of sexism. In what follows, we describe developments in the study of prejudice and stereotypes which will inform the understanding of the current construct and measures of sexism in the field of social psychology.

Evolution of Definitions of Prejudice

Early approaches considered prejudice as "an antipathy based upon a faulty and inflexible generalization" (Allport, 1954: p. 9). Prejudice was therefore intrinsically negative and hostile.

However, research showed that most social groups are also ascribed positive characteristics. For instance, Myrdal suggested, already in 1944, that contradicting norms expressed by the American Creed produced ambivalent attitudes about racial minorities. When people express negative prejudice, they tend to express at the same time sympathy, such as concerns for unjust treatment. Katz and his collaborators were the first to systematically investigate the effects of prejudice, sympathy, and the combination of the two (e.g., Katz and Hass, 1988).

The focus on the ambivalence of prejudice was further triggered by difficulties encountered in those days' techniques to measure prejudice. Such measures were explicit, typically asking people to associate traits to various social groups (e.g., aggressive, lazy, intelligent, etc.; see Katz and Braly, 1933). However, after World War Two people's sensitivity to minority issues increased. Egalitarianism became a social norm condemning the use of overt stereotypical generalizations based on group membership. Observations in the 1950s alluded to the fact that, at least in higher social classes, participants knew about the political correctness of egalitarian attitudes (McKee and Sherriffs, 1957). In concomitance with the feminist and antiracism movements in the 1970s, various types of *–isms* gained visibility and became blameworthy. As a consequence, the tendency to endorse overt prejudice against women declined over the years. Yet, in this process, more subtle forms of prejudice came into being, asking for a revision of the definition of this concept and of its measurement (see Maass et al., 2000). Such a revision was first met by Katz and Hass (1988), in the racial domain. In the gender domain, Guttentag and Secord (1983) advocated the necessity to take into account ambivalence arising from protective attitudes toward women. A few years later, Glick and Fiske (1996) presented the first psychometric scale integrating these principles (*Ambivalent Sexism Inventory*).

As a result of these developments, sexism, though still defined in terms of prejudice and discriminatory behavior, now includes combinations of *hostility* and *benevolence*. Accordingly, Glick and Hilt (2000) define prejudice as "*the implicit or explicit attitude that a group deserves inferior status*. The cognitive, emotional, and behavioral aspects of prejudiced attitudes may (for the prejudiced perceiver) be subjectively favorable, unfavorable, or ambivalent in their orientation toward the target group, yet they all serve to promote and maintain that group's subordination." (p. 245).

Ambivalence in Sexism

Early research on sex differences and stereotypes helps to understand the ambivalence of modern manifestations of sexism. The investigation of the content of gender stereotypes has its origin in Terman and Miles' work. Initially, these authors endeavored to examine sex differences in intelligence. But in their subsequent and most known work, *Sex and Personality* (1936), they took a different stand, conceding that research should rather focus on differences between male and female *temperaments*. In their *Test of Attitudes and Interests*, they assumed that masculinity and femininity are opposing ends of a single continuum, that this continuum is anchored in biology, and that temperaments can be assessed by asking men and women about their emotions, knowledge, and opinions. Deviations from the endorsement of the adequate masculine or feminine temperament were then considered as pathologies (i.e., indications of inversion and homosexuality, and of a variety of psychological maladjustments). Parsons and Bales (1955) likewise drew parallels between masculinity and instrumental roles in professions, and between femininity and expressive roles in the family, offering a broader sociological perspective on gender stereotypes. Broverman and collaborators later took up this idea in their *Questionnaire of Sex-Role Stereotypes* (Broverman et al., 1972). Their findings have long been considered as strong evidence for stereotypes portraying women as "less competent, less objective, and less logical than men," and men as "lacking interpersonal sensitivity, warmth, and expressiveness in comparison to women." (p. 75).

In the beginning of the 1970s, however, this one-dimensional (and bipolar) view of the masculinity–femininity construct began to be questioned. Bem and Spence empirically demonstrated that masculinity and femininity could be more adequately represented as two independent dimensions. As a consequence, the avenue was open for an individual to conceive of him/herself as possessing traits belonging to both personality types. With this notion of 'psychological androgyny' as a self-conception allowing for behavioral flexibility and better self-esteem, the simultaneous presence of masculine and feminine traits in each person, regardless of biological sex, was not considered abnormal anymore (see Lorenzi-Cioldi, 1996). Later conceptions of the androgynous person stressed the non-use of a gender-schema (Bem, 1981), supporting the ideal of a gender-blind society (i.e., *we are all the same*) endowed with the potential to overcome sexist attitudes and behaviors.

Let us turn to our initial question of how sexism is becoming increasingly ambivalent. Following the literature and their own investigations, Fiske, Glick, and their collaborators propose that *competence* and *warmth* are key dimensions of group stereotypes in their *stereotype content model* (Fiske et al., 2002). In what can be considered as an extension of Eagly's social role theory, they show that members of high-status groups are typically described as agentic (i.e., influential, independent, competent), while members of low-status groups are mainly described as communal (i.e., warm, caring, nurturing). Hence, this model implies that men and women are considered as complementing each other. But they are not equal. Men are attributed higher status, influence, and power, and receive respect for their agentic qualities, making them suitable candidates for leadership positions and other prestigious social roles. In contrast, women are praised for their interpersonal skills, which confer them positivity (see the so-called 'women are wonderful' effect; Wood and Eagly, 2010). But at the same time, this ostensible liking of women reveals a lack of agency, and thus competence, resulting in disrespect. In an apparent paradox, adulating women for their interpersonal skills undermines their evaluation in situations where competence is demanded. As a result, women are perceived to be suitable for low-status roles (e.g., domestic and health-care activities), and less suitable for high-status roles (e.g., decision-making positions). In sum, the positive connotation of the warmth dimension fuels ambivalence in a context where competence is relevant, and leads to the actual devaluation of women.

Glick and Fiske (1996) argue that the negative component of the stereotype is reflected in *hostile sexism*, and the positive one in *benevolent sexism*. A combination of the two produces ambivalence: The allegedly positive message communicated by benevolence masks the underlying hostile aim to dominate women and to keep women in their place. For example, research has shown that women may be applauded for their performances, yet resources are not allocated, or are even misattributed to others (Guinote and Vescio, 2010). Benevolent sexism expresses itself through protective paternalism, complementary gender differentiation, and heterosexual intimacy (Glick and Fiske, 1996). Protective and complementary aspects of benevolent sexism impact on the perception of women as needing protection and being unable to do things by themselves. Furthermore, they create expectations about the course of social interaction. On the one hand, men are expected to exert power over women, to protect, and to take care of them, putting them into a dependent position. On the other hand, women are compelled to assist this behavior by showing their affection and subordination to men. Indeed, research demonstrated that women perceive benevolent-sexist men as more attractive than hostile-sexist and nonsexist men (e.g., Montañés et al., 2013). Expectations of traditional gender roles, which figure men as protectors and women as *protégées*, should thus create the least conflict and the higher attraction between sexes. In line with this reasoning, experimental findings have shown that women face more difficulties to resist benevolent than hostile treatment (e.g., Becker and Wright, 2011). In addition, heterosexual interdependency and sexual attraction between men and women deepens the influence of such softer forms of prejudice in gender relations. The gender hierarchy is unlikely to be secured only through open hostility and subjugation. It is more efficiently maintained by coupling it with paternalism (Glick and Fiske, 1996; Rudman and Glick, 2008).

Hostile gender relations are learned and applied at an early age. Glick and Hilt (2000) offer a review of the literature on developmental aspects of sexism. At 9 months, children discriminate between male and female faces, and voices. At 2–3 years, they label themselves as boys or girls, and they know gender stereotypes related to objects (e.g., toys, clothes, and work). At kindergarten age, they become aware of power differences between genders, associating boys with power and negative traits (e.g., strong, fast, aggressive, cruel) and girls with fear, helplessness, and positive traits (e.g., needs help, nice, affectionate). This goes hand in hand with the development of negative attitudes toward, and dislike of, the other sex, and avoidance of interaction and play with the other sex (e.g., Maccoby and Jacklin, 1974). By the age of five, children have internalized that gender is a stable category. Accordingly, they begin to put emphasis on their gender identity, and engage in competitive forms of gender differentiation with the aim to show that the ingroup is better than the outgroup (Rudman and Glick, 2008). It is only with the rise of romantic sexual attraction, at about age 11, that gender relations become more complex and ambivalent, because positive emotions mix with the habitual hostility. According to correlational findings, romantic experiences are related to sexism, creating hostility against girls in adolescent girls and benevolence toward girls in adolescent boys (de Lemus et al., 2010). Glick and Hilt (2000) suggest that during puberty boys develop ambivalent feelings toward girls that include both hostility and benevolence. They direct benevolence toward girls who correspond to the feminine stereotype and whom they want to interact with, and hostility toward girls that do not comply with their expectations. In contrast, girls demonstrate awareness of status differences. They perceive higher power in boys, thus experiencing less satisfaction from interactions with boys. Consequently, they hold onto rather hostile attitudes toward boys. However, these attitudes can be undermined when girls enter a heterosexual relationship, which may produce ingroup hostility (de Lemus et al., 2010). Thus, adolescents enter the gender dynamics of adulthood where women are rewarded by men with benevolence when they behave in line with traditional gender roles.

Measuring Sexism

The acknowledgment that the negative effects of prejudice are not solely due to hostile attitudes and behaviors toward women, but can also derive from positive attitudes, was an important step in the development of sexism measures. Such measures can be roughly partitioned into two historical phases. The first phase focused on feminine and masculine personality traits which were supposed to fit biological sex. In the second phase, measures encompassed more explicit attitudes toward women. These measures were often used to generate hypotheses about discriminatory behavior toward women, or to relate them to other measures of social inequality.

First Phase

Following Terman and Miles' (1936) and Broverman et al.'s (1972) attempts to measure male and female temperaments and personalities, research produced much evidence about social attitudes toward the sexes. Jarrett and Sherriffs (1953) offered one of the initial attempts to assess, by means of a list of positive and negative behaviors, attitudes "toward the relative 'value' of men and women in our culture" (p. 41). The authors assumed that 17 of the items were stereotypically male (e.g., "Proportional to their numbers, most responsible for auto accidents in 1939 were ..."), 17 stereotypically female (e.g., "Having the most understanding for the needs of children are ..."), and 24 neutral. Participants judged for each item if it was true for men or for women. Results showed that, overall, participants gave more favorable scores to men compared to women. This finding points to the intriguing fact that both men and women show similar tendencies to value the higher-status group. The same authors conducted another study with similar items, and concluded that women and men hold the same stereotypes. Moreover, they were surprised to find that only 5 of the 24 presumably neutral items were actually judged as neutral. The conclusion was that most behavior is part of the male or female stereotype. In an attempt to overcome some methodological deficiencies of these studies, McKee and Sherriffs (1957) asked psychology students to rate men and women on a series of adjectives. Overall, men were again evaluated more favorably than women, an effect that was even stronger among women than men, suggesting a consensual ascription of legitimacy to the gender hierarchy.

Second Phase

This historical phase is divided up in three generations of sexism measures. The first aims to reveal the content of hostility toward women; the second aims to measure hostility in more indirect ways in order to meet social desirability concerns that make people hide prejudice; finally, the current generation integrates the dynamics of ambivalent sexism. We will only mention a few examples. An exhaustive summary of sexism measures is provided by Beere (1990).

First Generation – Old-Fashioned Attitude Scales

The 1970s and 1980s witnessed the publication of a number of well-designed gender attitudes scales. These scales were widely used and translated into different languages, which allowed for the first time a comprehensive test of their reliability and validity. They can be considered old fashioned because they measure traditional attitudes about gender roles and women's inferiority, notably in competence. Examples are Spence et al.'s *Attitudes Toward Women Scale* (AWS; e.g., "Women should be concerned with their duties of childbearing and house tending rather than with desires for professional or business careers"), Rombough and Ventimiglia's *Old-Fashioned Sexism Scale* (OFS; e.g., "Women should stay at home and care for the children"; "Men make better engineers than women"), and Beere et al.'s *Sex-Role Egalitarianism Scale*, with items assessing attitudes toward women and men in both traditional and nontraditional roles in various domains (marital, parental, employment, social, heterosexual, and educational).

Second Generation – Modern Sexism Scales

Gender stereotypes have been quite stable over time, and traditional expressions of hostile sexism in public discourse and sexual violence are still present in the twenty-first century. Nonetheless, in western societies, the restrictive nature of expectations about men's and women's social roles and behaviors has loosened up after the feminist movements of the 1970s and 1980s. Thus, more indirect ways of expressing sexism have become common. Two widely used instruments Tougas et al.'s *Neosexism Scale* and Swim et al.'s *Modern Sexism* Scale (MSS) illustrate this renewal. Indeed, these instruments tap opinions about egalitarian practices, rather than asking plain questions about women's competence or gender roles. Tougas et al. define neosexism as a "manifestation of a conflict between egalitarian values and residual negative feelings toward women" (1995: p. 843; e.g., "Women will make more progress by being patient and not pushing too hard for change"; "Due to social pressures, firms frequently have to hire underqualified women"). Likewise, subscales from the MSS deal with the *Denial of continuing discrimination* (e.g., "Discrimination against women is no longer a problem in the United States"), the *Antagonism toward women's demands* (e.g., "It is easy to understand the anger of women's groups in America"), and the *Resentment about special favors for women* (e.g., "Over the past few years, the government and news media have been showing more concern about the treatment of women than is warranted by women's actual experiences."). Nonetheless, MSS still includes measures of the traditional, more overt sexism (e.g., "Women are generally not as smart as men").

Third Generation – Ambivalence

The above scales still represent clearly negative statements, either overt or more subtle. None of these measures managed to integrate the ambivalent nature of modern prejudice that was initially called into attention by Guttentag and Secord (1983). This requirement was met by *ambivalent sexism theory* (Glick and Fiske, 1996). Central to this theory is the softer and friendlier type of sexism conveyed through benevolence. Traditional *hostile sexism* is characterized by its overtly negative nature that emphasizes women's inferiority to men, whereas *benevolent sexism* emphasizes status differences by affectionate and patronizing behaviors (Glick and Fiske, 1996; Rudman and Glick, 2008). Benevolent sexism is potentially equally harmful as the hostile, but through a different process. Its ostensibly flattering and friendly nature conceals the underlying expression of dominance, making it more likely to be accepted by women. Benevolent sexism meets social norms that demand that men protect women. It legitimizes status differences, and undermines potential conflicts and tensions.

Hostile and benevolent forms of sexism occur most often side by side. Indeed, benevolence contributes to mask the negativity of hostility (e.g., Barreto and Ellemers, 2013). Thus, the *Ambivalent Sexism Inventory* (Glick and Fiske, 1996) includes a *Hostile sexism* subscale (HS; e.g., "Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for 'equality'") and a *Benevolent sexism* subscale (BS). The latter has three subcomponents that correspond to the above-mentioned characteristics of benevolence: *Protective paternalism* (e.g., "Women should be cherished and protected by men"), *Complementary gender differentiation* (e.g., "Many women have a quality of purity that few men possess"), *and Heterosexual intimacy (*e.g., "*Men* are incomplete without women"). Across 16 nations, Rudman and Glick (2008) found HS and BS to be highly correlated. HS predicted negative trait ascriptions to women, and BS positive trait ascriptions, thus supporting the view of prejudice as an antipathy plus an affectionate patronizing component.

Collectively, the measures described so far can be considered as *explicit* measures of sexism. In concomitance with these measures, various scholars have argued that expression of opinions is highly contingent on social desirability concerns. Furthermore, attitudes may not be consciously accessible. Self-deception may drive people to negate or hide prejudice, or simply be not aware of their actual prejudice. In the 1980s and 1990s, a variety of alternative, mostly *implicit* measures, were developed in response to these new challenges. The most popular measure is the *Implicit Association Test*. This computerized instrument was originally proposed by Greenwald and his collaborators in 1998. Participants either press a key when two stereotypically linked words appear on a screen (e.g., a female name and a feminine profession), or press another key when there is no stereotypical association between the words. If reaction times are faster for stereotypical associations than for nonstereotypical ones, it is inferred that the individual is prejudiced. Other studies have measured unconscious nonverbal behaviors that express prejudice (e.g., de Lemus et al., 2012). Overall, however, clear demonstrations of the ecological validity of such measures are still awaited.

Consequences of Sexism

Sexism sustains social asymmetries in gender relations that typically advantage men. Sexist attitudes and opinions may produce discriminatory treatment such as unequal hiring and promotion practices, but also more extreme forms such as physical aggression and sexual violence. Past research has demonstrated a variety of detrimental discriminatory effects of sexism. The previous discussion of types of sexisms is important because each of the types maintains gender hierarchies in specific ways. In particular, it is important to distinguish between effects of hostile and benevolent sexism. One crucial difference is that HS is more easily challenged than BS. The sexist nature of benevolence is particularly difficult to identify as a corollary of discriminatory behavior (Barreto and Ellemers, 2013). Women are themselves vulnerable to, and thus accepting, this persuasive influence style. Conversely, they may more easily confront hostility. Explicit exertion of power and open acts of hostility are more likely to fuel a motive for collective action and protest (e.g., Becker and Wright, 2011). HS is therefore only effective if exercised infrequently, and/or when it is balanced with benevolence (Guinote and Vescio, 2010). Furthermore, recent research also suggests that confrontation of sexist behavior may have negative consequences on the evaluation of women, and that men judge as less attractive women who do not endorse traditional sexist beliefs (e.g., Barreto and Ellemers, 2013; Montañés et al., 2013). Thus, confrontation of sexism is not an easy option for women.

Another consequence of hostile sexism is the so-called *stereotype threat* effect. Women perform typically worse in stereotypically masculine tasks (e.g., a math test) as their gender membership is emphasized, even by means of subtle reminders (e.g., Spencer et al., 1999). However, similar damaging effects occur with benevolent sexism. A growing body of experimental evidence demonstrates that benevolence is harmful for women's feelings of competence and even actual performances (Dardenne et al., 2007; de Lemus et al., 2012; Guinote and Vescio, 2010). Women may underperform because of their fear of negative treatment if they reveal good performances. Agentic achievements by women are often judged as a violation of prescriptive gender stereotypes, and can thus lead to punishments, sabotage, and unfavorable evaluations, a phenomenon called *backlash* (Eagly and Karau, 2002; Rudman and Glick, 2008). Research on gender stereotypes has shown that the agentic descriptions of men match closely those of the normative white and healthy adults in general, whereas descriptions of women are embedded in group-specific, relational, and communal characteristics (see Wood and Eagly, 2010). Accordingly, the 'think manager–think male' phenomenon (Schein et al., 1996) shows that leaders in general are most often endowed with masculine, agentic characteristics. Female leaders must therefore face the conflicting demands of what it means to be a good leader and of what it means to be a woman (see Heilman et al., 2004). *Role congruity theory* (Eagly and Karau, 2002), as well as the *status incongruity hypothesis* (Rudman et al., 2012), examine antecedents and consequences of perceived violations of gender or status expectations. To avoid punishment, women tend to adapt their behaviors, for example, by negotiating lower salaries for themselves than for others, not promoting themselves in job-interviews, or hiding their performances (e.g., Rudman and Glick, 2008). These behaviors support the status quo of the gender hierarchy, and contribute to ban women from prestigious and high-status activities (Rudman et al., 2012). Thus, sexism does not only impact women's actual performances, but also leads to the downplaying, or the conscious hiding, of successful performances in traditionally masculine domains. Moreover, it has consequences on strategies for identity improvement. Indeed, if men pursue positive distinctiveness using agentic (i.e., status-relevant) dimensions, women, in a parallel process, pursue positive distinctiveness using communal (status-irrelevant) dimensions (e.g., Maccoby, 1998). From the social identity perspective, this difference can be conceived of as a *social creativity* strategy whereby members of a low-status group achieve positive distinctiveness despite their negative standing in the social structure (e.g., Tajfel, 1981).

Finally, at a macro-level, it has been found that national averages in HS and BS are positively related to gender inequality indexes in those nations (Rudman and Glick, 2008). Especially in traditionally male domains, men consider that the work domain is not suitable for women's qualities and competencies, whereas women accuse sexism targeted against them. More generally, the impact of communal stereotypes on women's professional careers is revealed by the consistent and substantive negative correlations between family life (as assessed, for instance, with family size) and professional life. The corresponding correlations typically reach outstanding positive levels among men (cf Lorenzi-Cioldi, 2002).

Conclusion

We conclude this article by emphasizing an important emerging sex difference in the expression of gender stereotypes and sexism. As we mentioned, until the 1970s men and women tended to acknowledge a social male advantage. Men were judged as more agentic (competent, independent, objective) and women as more communal (warm, expressive, interpersonal sensitive). Although sexism remains a complex societal phenomenon nowadays with both sexes expressing sexism against women, women's attitudes clearly evolved more rapidly than men's. Whereas men's attitudes have remained largely traditional, women's attitudes are moving toward more liberal views.

See also: Emotions and Intergroup Relations; Feminist Research and Social Work; Gender Differences in Personality and Social Behavior; Gender and Academic Motivation; Implicit Association Test; Implicit Social Cognition; Intergroup Relations; Prejudice and Discrimination; Racism: Social Psychological Perspectives; Social Psychology; Stereotypes in Social Psychology; Stigma, Social Psychology of.

Bibliography

Allport, G., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.

Barreto, M., Ellemers, N., 2013. Sexism in contemporary societies: how it is expressed, perceived, confirmed and resisted. In: Ryan, M.K., Branscombe, N.R. (Eds.), The SAGE Handbook of Gender and Psychology. Sage, London, pp. 289–305.

Becker, J.C., Wright, S.C., 2011. Yet another dark side of chivalry: benevolent sexism undermines and hostile sexism motivates collective action for social change. Journal of Personality and Social Psychology 101 (1), 62–77.

Beere, C.A., 1990. Gender Roles. A Handbook of Tests and Measures. Greenwood Press, Westport, Conneticut.
Bem, S.L., 1981. Gender schema theory: a cognitive account of sex typing. Psychological Review 88 (4), 354–364.
Broverman, I.K., et al., 1972. Sex-role stereotypes: a current appraisal. Journal of Social Issues 28 (2), 59–78.
Dardenne, B., Bollier, T., Dumont, M., 2007. Insidious dangers of benevolent sexism: consequences for women's performance. Journal of Personality and Social Psychology 93 (5), 764–779.
Deaux, K., Major, B., 1987. Putting gender into context: an interactive model of gender-related behavior. Psychological Review 94 (3), 369–389.
Diekman, A.B., Eagly, A.H., 2000. Stereotypes as dynamic constructs: women and men of the past, present, and future. Personality and Social Psychology Bulletin 26 (10), 1171–1188.
Eagly, A.H., 1987. Sex Difference in Social Behavior: A Social-Role Interpretation. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.
Eagly, A.H., Karau, S.J., 2002. Role congruity theory of prejudice toward female leaders. Psychological Review 109 (3), 573–598.
Fiske, S.T., et al., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82 (6), 878–902.
Glick, P., Fiske, S.T., 1996. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology 70 (3), 491–512.
Glick, P., Hilt, L., 2000. From combative children to ambivalent adults: the development of gender prejudice. In: Eckes, T., Trautner, M. (Eds.), Developmental Social Psychology of Gender. Erlbaum, Mahwah, NJ, pp. 243–272.
Goldberg, P.A., 1968. Are women prejudiced against women? Transaction 5, 28–32.
Guimond, S., Chatard, A., Lorenzi-Cioldi, F., 2013. The social psychology of gender across cultures. In: Ryan, M.K., Branscombe, N.R. (Eds.), The SAGE Handbook of Gender and Psychology. Sage, London, pp. 216–233.
Guinote, A., Vescio, T.K., 2010. The Social Psychology of Power. The Guilford Press, New York, pp. 363–380.
Guttentag, M., Secord, P., 1983. Too Many Women? Sage, Beverly Hills, CA.
Heilman, M.E., et al., 2004. Penalties for success: reactions to women who succeed at male gender-typed tasks. Journal of Applied Psychology 89 (3), 416–427.
Hyde, J.S., 2005. The gender similarities hypothesis. The American Psychologist 60 (6), 581–592.
Jarrett, R.F., Sherriffs, A.C., 1953. Propaganda, debate, and impartial presentation as determiners of attitude change. Journal of Abnormal and Social Psychology 48 (1), 33–41.
Johannesen-Schmidt, M.C., Eagly, A.H., 2002. Diminishing Returns; the effects of income on the content of stereotypes of wage earners. Personality and Social Psychology Bulletin 28 (11), 1538–1545.
Katz, D., Braly, K., 1933. Racial stereotypes of one hundred college students. Journal of Abnormal and Social Psychology 28 (3), 280–290.
Katz, I., Hass, R.G., 1988. Racial ambivalence and American value conflict: correlational and priming studies of dual cognitive structures. Journal of Personality and Social Psychology 55 (6), 893–905.
Leet, P.M., 1965. Women and the Undergraduate. Mimeo.
Lorenzi-Cioldi, F., 2006. Group status and individual differentiation. In: Postmes, T., Jetten, J. (Eds.), Individuality and the Group. SAGE Publications, London, pp. 94–115.
Lorenzi-Cioldi, F., 2002. Les représentations des groupes dominants et dominés [Images of Dominant and Subordinate Groups]. Presses Universitaires, Grenoble.
Lorenzi-Cioldi, F., 1996. Psychological androgyny: a concept in search of lesser substance. Towards the understanding of the transformation of a social representation. Journal for the Theory of Social Behaviour 26 (2), 137–155.
de Lemus, S., Moya, M., Glick, P., 2010. When contact correlates with prejudice: adolescents' romantic relationship experience predicts greater benevolent sexism in boys and hostile sexism in girls. Sex Roles 63 (3–4), 214–225.
de Lemus, S., Spears, R., Moya, M., 2012. The power of a smile to move you: complementary submissiveness in women's posture as a function of gender salience and facial expression. Personality and Social Psychology Bulletin 38, 1480–1494.
Maass, A., Castelli, L., Arcuri, L., 2000. Measuring prejudice: implicit versus explicit techniques. In: Capozza, D., Brown, R. (Eds.), Social Identity Processes. Sage, London, pp. 96–116.
Maccoby, E.E., 1998. The Two Sexes: Growing up Apart, Coming Together. Harvard University Press, Cambridge.
Maccoby, E.E., Jacklin, C.N., 1974. The Psychology of Sex Differences. Stanford University Press, Stanford, California.
McKee, J.P., Sherriffs, A.C., 1957. The differential evaluation of males and females. Journal of Personality 25 (3), 356–371.
Meyers, B., 1984. Minority group: an ideological formulation. Social Problems 32 (1), 1–15.
Montañés, P., et al., 2013. How attractive are sexist intimates to adolescents? The influence of sexist beliefs and relationship experience. Psychology of Women Quarterly 37 (4), 494–506.
Myrdal, G., 1944. An American Dilemma. Harper & Row, New York.
Parsons, T., Bales, R.F., 1955. Family, Socialization and Interaction Process. Free Press, Glencoe, IL.
Ridgeway, C.L., 2011. Framed by Gender: How Gender Inequality Persists in the Modern World. Oxford University Press, New York.
Rudman, L.A., et al., 2012. Status incongruency and backlash effects: defending the gender hierarchy motivates prejudice against female leaders. Journal of Experimental Social Psychology 48 (1), 165–179.
Rudman, L.A., Glick, P., 2008. The Social Psychology of Gender: How Power and Intimacy Shape Gender Relations. The Guildford Press, New York.
Schein, V.E., et al., 1996. Think manager – think male: a global phenomenon. Journal of Organizational Behavior 17 (1), 33–41.
Shapiro, F.R., 1985. Historical notes on the vocabulary of the womens movement. American Speech 60 (1), 3–16.
Spencer, S.J., Steele, C.M., Quinn, D.M., 1999. Stereotype threat and women's math performance. Journal of Experimental Social Psychology 35 (1), 4–28.
Tajfel, H., 1981. Human Groups and Social Categories. Cambridge University Press, Cambridge.
Terman, L.M., Merrill, M., 1937. Measuring Intelligence: A Guide to the Administration of the New Revised Stanford-Binet Tests of Intelligence. Houghton Mifflin, Boston.
Terman, L.M., Miles, C.C., 1936. Sex and Personality: Studies in Masculinity and Femininity. McGraw-Hill, New York.
Tougas, F., et al., August 1995. Neo-sexism: plus ça change, plus c'est pareil. Personality and Social Psychology Bulletin 21, 842–849.
Twenge, J.M., 2001. Changes in women's assertiveness in response to status and roles: a cross-temporal meta-analysis, 1931–1993. Journal of Personality and Social Psychology 81 (1), 133–145.
Wood, W., Eagly, A.H., 2010. Gender. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of social psychology, vol. 1. Wiley, New York, pp. 629–667.

Relevent Websites

http://www.coe.int/t/DGHL/STANDARDSETTING/EQUALITY/ – Council of Europe.
http://ec.europa.eu/justice/gender-equality/ – European Union.
https://implicit.harvard.edu/implicit/ – Measure of Sexism.
http://datatopics.worldbank.org/gender/ – Statistics.
http://www.understandingprejudice.org/links/sexism.htm – Scientific and Teaching.
http://www.catalyst.org/ – Sexism in the Workplace.
http://www.everydaysexism.com/ – Sexism in Everyday Life – Reports.
http://www.unesco.org/new/en/unesco/themes/gender-equality/ – United Nations – UNESCO.

Sexual Behavior and Social Networks

Benjamin Cornwell and Rachel Behler, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Few things are more private than sexual behavior. Yet scholars increasingly see sexual behavior as being shaped by the broad social networks of kin, friends, and other contacts in which individuals are embedded. This article describes research that considers the relevance of these social (not sexual) contacts to a wide range of sexual behaviors, including sexual initiation, sex partner selection, expression of sexual identity, sexual frequency and problems, as well as risky behaviors such as unprotected sex, transactional sex, using drugs while having sex, and finding sexual partners online. We also address the finding that social networks are themselves shaped by these sexual behaviors. We close by discussing new directions for research on this topic.

Due to its implications for individual well-being, public health, fertility, and other important phenomena, research on sexual behavior has tremendous currency in the fields of sociology, demography, epidemiology, and medicine, among others. A rapidly growing literature shows that social networks play a major role in shaping numerous specific sexual behaviors. This work is distinct from but related to analyses of sexual networks themselves – webs of sexual connections among sex partners – by virtue of its focus on the importance of the nonsexual relationships that individuals maintain. As we will show, a variety of very private sexual behaviors – including such things as having unprotected sex and using drugs to enhance sex – are shaped by diverse social contacts including peers, friends, coworkers, neighbors, and family members. From early adolescence to later life, the influence of these types of social contacts is evident in individuals' desires for and attitudes toward sex, expression of sexual identity, exposure to sex partners and sex markets, and even the onset of sexual dysfunction.

The main goal of this article is to review evidence of strong relationships between aspects of individuals' social networks and their sexual behaviors. We begin by briefly discussing some of the mechanisms that are thought to give rise to this association, but we spend most of the article reviewing the evidence.

Why Do Social Networks Matter?

Social networks shape and channel individual behaviors through a variety of processes including social influence, role modeling, social support, and social control (e.g., see Friedkin and Johnsen, 2011). For one, sexual behavior is learned and diffused through close relationships. It is reinforced through network members' behavioral examples, the normative environments they create, and the pressures they exert, as well as individuals' perceptions of them, regardless of how accurate these perceptions are. Social network members can serve as sources of information about specific sexual behaviors (e.g., based on previous experience) as well as about potential sex partners, and they often introduce individuals to their future sex partners. Social network members often also exercise informal control over sexual behavior by creating restrictive normative contexts, by discouraging certain kinds of behaviors, and by exercising other forms of informal social control (e.g., ostracism). Finally, social network members are also responsive and react to each other's behaviors, thereby providing mechanisms by which sexual behaviors may reciprocally affect social networks.

The following sections address foundational and recent work which demonstrates associations between social networks and individuals' sexual behaviors. The remainder of this article is organized topically, addressing sometimes separate areas of research on different kinds of sexual behaviors.

Sexual Intention and Initiation

We begin by addressing sexual behaviors early in the life course. Most research shows that networks of peers shape individuals' sexual intentions and behaviors at an early age. A variety of mechanisms operate here, including adolescents' (perceptions of) network members' attitudes toward sex, their sexual behaviors, and their control efforts. Most studies agree that adolescents who perceive disapproval among their friends and/or parents with respect to adolescents having sex are more likely to remain abstinent or to delay sexual initiation (Buhi and Goodson, 2007). Adolescents' intentions and desires to have sex are also associated with the belief that most same-age peers have had sex (Miranda-Diaz and Corcoran, 2012).

Network members' narratives about romantic relationship trajectories and sexual milestones are also important. Harding (2007) suggests that adolescents from disadvantaged neighborhoods, for example, may have more lenient attitudes toward getting pregnant due to the larger number of competing sexual frames and scripts to which they are exposed. School-level norms are similarly influential in shaping adolescents' age of sexual initiation and risk of pregnancy (Teitler and Weiss, 2000). One study found that the efficacy of virginity pledges among adolescents is highly contingent on the number of other pledgers (Bearman and Bruckner, 2001).

Sexual Orientation and Identity

Social network members provide a normative environment that can be either permissive or intolerant with respect to a given sexual orientation (e.g., homosexuality), and thus may

 http://dx.doi.org/10.1016/B978-0-08-097086-8.35029-2

facilitate or obstruct certain sexual behaviors. Developmental perspectives on sexual orientation likewise argue that sexuality is partly a function of social pressures (e.g., toward conformity) and socialization, noting surprising levels of fluidity in the sexual orientations of young adults over time (see Saewyc, 2011). Other research shows that social network members' perceived emotional supportiveness is a key factor in determining whether sexual minorities disclose their sexual orientation (e.g., Latkin et al., 2012). Those who hide their orientation from close network members may be forced to engage in secretive (e.g., 'down-low') sexual behavior, which can lead to anonymous sex, transactional sex, and increased risk of HIV infection (e.g., Dodge et al., 2008), especially in the case of men who have sex with men (MSM) who also have sex with women.

Most research on the link between social networks and sexual orientation, however, focuses on the reverse process through which sexual orientation shapes social networks. A person's sexual orientation can affect how existing network members interact with them (see Julien et al., 1999). Sexual minorities perceive less support from family members than heterosexuals do, which leads sexual minorities to be more selective about their social network connections, sometimes leading to less kin-centered networks. In a similar vein, the joint social networks of both gay and lesbian partners tend to be larger and more comprised of friends than the joint networks of heterosexual couples.

Sexual Frequency, Satisfaction, and Problems

There is considerable evidence that having sex has numerous physical and mental health benefits (e.g., Brody, 2010). A large body of research on sexual health emphasizes the value of frequent sexual intercourse and satisfaction with sex, especially in committed relationships (e.g., Higgins, et al., 2011). Conversely, some researchers are interested in factors that contribute to sexual problems or dysfunction. Although much of this work focuses on health and medical factors, relationship quality, and other individual and dyadic factors, there is growing awareness of extradyadic social network influences on sexual functioning. We briefly discuss research on these topics in the following sections.

Sexual Frequency and Satisfaction

The extent of sexual activity in which individuals engage is related to their connectedness to nonsexual social contacts. Research shows that having supportive social ties increases the likelihood of engaging in and enjoying sexual activity (e.g., Prairie et al., 2011). They were also more likely to have reported being satisfied with their sexual activity. These associations may reflect benefits of psychosocial factors such as self-esteem, happiness, and life satisfaction for physical and sexual functions. As we will discuss below, maintaining extensive and wide-ranging social networks also increases the likelihood of being satisfied with one's sexual relationships, in part because it yields access to more diverse sex markets and thus increases the likelihood of finding a well-matched sex partner.

Sexual Dysfunction

Sexual problems and sexual dysfunction – for example, lack of interest in sex, difficulty achieving orgasm, erectile dysfunction (ED) – have also been conceptualized as social network issues. Existing research that emphasizes dyadic relationship qualities that diminish sexual arousal response (e.g., incompatible sexual scripts between partners or partner discord) already indirectly implicates sex markets and partner-matching processes (discussed below) in the genesis of sexual dysfunction (e.g., individuals who have smaller social networks find it more difficult to establish a sexually fulfilling relationship). Some research has linked men's risk of ED specifically to how men and their partners are connected through broader social networks. One aspect of social networks that is thought to increase heterosexual men's risk of psychological risk factors such as depression and anxiety is the extent to which men dominate their own social network resources. One study found that when a heterosexual man's female partner has more frequent contact with his network members than he does, he is 92% more likely to report having trouble getting/maintaining an erection and 78% more likely to have trouble achieving orgasm during sex (Cornwell and Laumann, 2011). The authors argue that this network arrangement introduces strain into men's sexual relationships by decreasing partner satisfaction and sense of privacy and autonomy. Little is known about what role social networks play in the genesis of sexual problems in women.

Extramarital Sex

The influence of social networks extends to the practice of extramarital sex. In general, extramarital sex is more likely among men, and most of the work that has been done on the social network determinants of extramarital sex focuses on men. Research on this population documents strong peer influences on extramarital sex. Men whose friends and other confidants engage in extramarital sex (with sex workers or otherwise) are also more likely to engage in extramarital sex (Vanlandingham et al., 1998). Another important aspect of social networks in this respect is the extent of overlap between sex partners' social networks. Network overlap increases social support, sense of couplehood, and the capacity for mutual monitoring. Among both men and women, spouses who share numerous mutual social network members are less likely to have extramarital sex (Treas and Giesen, 2000).

Sex Partner Selection and Sex Markets

Social networks directly shape processes of sex partner selection and matching. Network contacts provide information and advice about potential sex partners, and can make introductions between them. They can also exert pressure to date certain types of people or to pursue certain types of relationships (Laumann et al., 2004). Research suggests that roughly 60% of all partnerships (including marriage, cohabitation, and short-term relationships) result from an introduction by a mutual acquaintance. Mutual friends are responsible for 30% of those introductions, while family members, coworkers, and

classmates together account for the remaining 30% (see Laumann et al., 1994). The most common way people find their partners is through introductions made by their friends. It is worth noting that how these introductions are made is related to the level of commitment within sexual relationships.

Sex Markets and Affiliation Networks

The concept of sex markets reframes the search for sex partners, identifying the intersection of social networks and physical spaces as markets in which participants shop for sex partners interested in specific sexual activities (e.g., casual sex). Sex markets arise from the combination of informal networks of individuals who facilitate introductions to sex partners and organizations and other places that constitute opportunity structures for meeting sex partners. Examples include places that host singles' nights, café speed-dating events, and dance clubs. The 'affiliation networks' of places that form sex markets (see Frost, 2007) shape and are shaped by the sexual behavior of their constituent members. Sex markets are socially constructed by their members – defined by the expectations of their participants and by their embeddedness in particular local social and cultural milieus. Participants hold expectations about the types of sexual relationships (e.g., short-term versus long-term), the sexual scripts, and the 'rules of exchange' that operate within a given market. Accordingly, different places cater to different sex markets. The speed at which sexual activity occurs and which sexual acts take place in the relationship is linked to the place partners were introduced. For example, partners who meet in a bar are more likely to report having sex within the first month of the relationship than those who meet in school (Laumann et al., 1994). Larger spaces, like neighborhoods, are often composed of multiple overlapping sex markets, including adjacent bars that cater to heterosexual or homosexual clientele or public parks that attract people who are looking for anonymous short-term sex (Laumann et al., 2004). Thus, researchers have begun to examine individuals' positions within the networks of places that form sex markets to understand individuals' risk of contracting or spreading sexually transmitted diseases (STDs) such as HIV/AIDS (e.g., Frost, 2007).

Online Social Networks

Online social networks are increasingly used as sex markets. This is especially true of populations in 'thin markets,' or places in which few potential partners exist who meet baseline criteria for matching, such as LGBT people searching for other LGBT partners. Online networks expand an individual's effective search area beyond both the geographic boundaries formed by their neighborhoods and the social boundaries formed by their personal social networks. Men and women in the 35- to 50-year-old age range are the most likely to meet online (Rosenfeld and Thomas, 2012). As of 2009, the Internet was the third-most-common way heterosexuals found their partners, following introductions by friends and meeting in bars. Online meetings accounted for 22% of heterosexual partner meetings and 60% of same-sex partner introductions. This new role of the Internet as a relationship broker has also been accompanied by declines in the role of personal social networks and local places in forging sex partnerships (Rosenfeld and Thomas, 2012). As we will discuss in greater detail below, this is a disturbing trend in light of the fact that people who use online social networks are also more likely to report engaging in a number of risky sexual behaviors, including having unprotected sex, having sex with multiple partners, using drugs during sex, and having sex with a known HIV-positive partner (e.g., see Benotsch et al., 2002; Horvath et al., 2008).

Risky Sexual Behavior

Some research on the role played by social networks in sexual behavior is motivated by public health concerns, especially regarding the spread of STDs like hepatitis and HIV. Much research has focused on how sexual and drug-sharing networks, or 'risk networks,' affect disease diffusion, giving rise to the field of network epidemiology. But researchers are also increasingly interested in the role of other social network contacts who may not be risk network partners (e.g., family members and close friends) in encouraging, facilitating, discouraging, or obstructing sexual behaviors that can increase STI (sexually transmitted infection) transmission (e.g., Tyler, 2008). This includes a range of potentially risky sexual behaviors, including engaging in unprotected sex, sex with multiple partners, engaging in sex while under the influence of drugs or alcohol, and transactional sex.

Unprotected Sex

One of the most-studied questions in this area of research is how social network members shape individuals' beliefs about and use of condoms during sex. Some research highlights the importance of parents. Wolff and Crockett (2011) find that adolescents who receive more support from their parents are less likely to have unprotected sex. Likewise, those who communicate with their parents about sexual behavior are less likely to have unprotected sex (Hadley et al., 2009). Other network members – especially friends – may influence individuals' tendencies to use condoms by serving as role models, by exchanging information, and/or by shaping the normative environment. Choi and Gregorich (2009) find that women are more likely to report partner condom use when they have at least one confidant who encourages condom use. Condom use is also greater among those whose confidants know each other, suggesting either that perceived agreement among confidants or confidants' abilities to exercise informal control affects women's safe sex practices. Other work shows that adolescents whose friends engage in deviant behavior (e.g., smoking and drinking) are more likely to engage in unprotected sex (Wolff and Crockett, 2011), suggesting that more permissive normative environments in general increase the risk of unprotected sex.

Even among at-risk populations, such as sex workers, network influence remains strong. Latkin et al. (2003) found that, among impoverished men and women who are at high HIV risk, having friends who talk about condoms, encourage condom use, and using condoms increases the likelihood of condom use. Condom use is more likely among male clients

of female sex workers when they perceive that at least some of their male social network members also used condoms consistently (e.g., Barrington et al., 2009). Other work shows that MSM are more likely to engage in unprotected anal intercourse when they perceive that their nonsex partner confidants likewise engage in such behavior (Schneider et al., 2013). Structural features of individuals' social networks may also matter. Those who have dense networks are more likely to perceive high levels of condom use among their network members than those whose networks are more sparse (Barrington et al., 2009).

Use of Drugs during Sex

A growing concern among health researchers is the practice of using drugs – such as inhaled amyl nitrate (poppers) and amphetamines – to enhance sexual experience. The use of drugs during sex is associated with risker sexual behavior (e.g., rough sex, unprotected sex) and thus heightened risk of HIV transmission, in part because drugs effectively reduce inhibition and rational decision-making (see Stall and Purcell, 2000). The use of drugs during sex may be greater among individuals who are already at greater risk of STD infection, especially homeless youth and MSM (Stall and Purcell, 2000). Various classes of drugs have become widely established as 'drugs of choice' for engaging in high-risk sex in a number of contexts, especially stimulants, poppers, and ED drugs (e.g., Mansergh et al., 2006). One recent concern is over the use of multiple drug use during sex, as this further increases the risk of unsafe sex practice and HIV infection (Ostrow et al., 2009). Recent research has discovered a strong link between the use of drugs during sex among nonsex partner social network members and individuals' own use of sex drugs. This has to do with individuals' perception of approval of such behavior among their network members, in that the presence of even one enabling social network member in one's confidant network is associated with significantly increased risk of sex-drug use, even after controlling for those network members' perceived use of sex drugs (Schneider et al., 2013).

Multiple Sex Partners

Having sex with numerous people over a short period of time or with multiple partners (simultaneously or in separate but concurrent relationships) is a major risk factor for STI transmission. The research that exists on how social networks shape the likelihood of having multiple sex partners focuses on both social influence processes and sex markets. On the one hand, research on homeless youth suggests that those who have a close social network member (e.g., friend) are less likely to have multiple sex partners (Ennett et al., 1999). On the other hand, a study of impoverished urban African-American women found that the likelihood of having multiple sex partners increases with personal social network size, possibly indicating greater access to potential partners through network channels (Neblett et al., 2011). Finally, the likelihood of having sex with multiple partners is positively associated with the extent of alcohol and drug use among social network members (e.g., Neblett et al., 2011), suggesting that the broader normative environment is important.

Transactional Sex

Exchange-oriented, or 'transactional,' sex refers to the dynamic wherein sex is offered by a participant in return for something else, frequently drugs or money. Transactional sex is often conceptualized in terms of prostitution, but it can also be used to describe less-commercialized forms of sex, such as courting rituals that dictate a 'tit-for-tat' pattern in which gifts are exchanged in return for sexual favors. Transactional sex is frequently associated with risky behavior because of the power dynamic that can prevent the party offering sex from utilizing protection. This is especially true in the context of 'survival sex,' which involves transactional sex as a means acquiring money, food, or even shelter. Social network members may affect transactional sex behavior both through conventional social influence pathways or through sheer abandonment. Survival sex is especially likely among disadvantaged youths who have been alienated by close social network members (especially family), often because of their sexual orientation (Ennett et al., 1999). Research also shows that one's likelihood of engaging in transactional sex is a function of the social network environment. One study reports that women who believed that their core social network members (especially friends) exchanged sex were more than twice as likely to exchange sex as women who did not believe that their friends exchanged sex. At the same time, perceived approval of transactional sex among these women's peers cut their likelihood of exchanging sex by more than half (Davey-Rothwell and Latkin, 2008). It is important to note that gender is a salient dimension of such exchange-oriented sex, as men are much more likely to purchase sex and women are more likely to sell it (Latkin et al., 2003).

Conclusion

This article has highlighted the relevance of broad networks of (nonsexual) social contacts – including kin, friends, peers, and others – to seemingly private sexual behaviors. Social network members create the normative contexts in which sexual behaviors and intentions emerge, they serve as role models, and they can facilitate or impede sexual partnering. Social networks can therefore both facilitate and suppress sexual behaviors ranging from adolescent sexual initiation to survival sex.

There are several key challenges facing researchers in this area. To understand the implications of nonsexual social networks for consequential sexual behaviors and problems, this work needs to be more closely merged with network-epidemiological research on *sexual* networks. An important but largely unaddressed question is to what extent individuals' social and sexual networks overlap. Some social contacts (e.g., friends) may also be sex partners or more distant members of one's sexual network. Overlap between sexual and social networks can occur when people are alienated from their social networks (e.g., families) and/or when social contacts (e.g., friends) become sex partners (and vice versa). As such, estimates of social network influence may actually partly capture sex partner influences. The assessment of the selective emergence and extent of such network overlap is an important direction for research. Related to this, the vast majority of the above research is cross-sectional and is plagued by

endogeneity. Social network researchers are actively working on methods to distinguish between the effects of social networks on behavior and the reciprocal effects of that behavior on the structure and composition of networks. An important step in this direction is the collection of longitudinal data that captures changes in individuals' social and sexual networks.

Finally, the promise of this area of research lies in the prospect that it can be used to attenuate important social problems, such as the spread of HIV/AIDS as well as some of the risky behaviors that contribute to it, such as the use of drugs during sex and survival sex. Fortunately, social networks are figuring increasingly prominently in intervention studies that focus on such behaviors. Recent interventions have explored the prospect of using the power of social network members' influence to increase condom use and other HIV-prevention efforts among adolescents, sex workers, and other at-risk populations (see Barrington et al., 2009; Wang et al., 2011), to reduce transactional sex among women (Davey-Rothwell and Latkin, 2008), and numerous other applications. Greater collaboration between social network researchers and public health researchers is needed to ensure that these interventions harness the full range of mechanisms through which various social network features may affect relevant sexual behaviors.

Acknowledgments

We wish to thank Erin York Cornwell and Linda J. Waite for providing useful suggestions that improved this article.

See also: Adolescent Sexual Risk; Hooking up and Casual Sex; Sexual Dysfunction, Epidemiology of; Sexuality Over the Life Course; Sexuality, Theories of; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Barrington, C., Latkin, C., Sweat, M.D., Moreno, L., Ellen, J., Kerrigan, D., 2009. Talking the talk, walking the walk: social network norms, communication patterns, and condom use among the male partners of female sex workers in La Romana, Dominican Republic. Social Science & Medicine 68, 2037–2044.

Bearman, P.S., Bruckner, H., 2001. Promising the future: virginity pledges and first intercourse. American Journal of Sociology 106, 859–912.

Benotsch, E.G., Kalichman, S., Cage, M., 2002. Men who have met sex partners via the internet: prevalence, predictors, and implications for HIV prevention. Archives of Sexual Behavior 31, 177–183.

Brody, S., 2010. The relative health benefits of different sexual activities. The Journal of Sexual Medicine 7, 1336–1361.

Buhi, E.R., Goodson, P., 2007. Predictors of adolescent sexual behavior and intention: a theory-guided systematic review. Journal of Adolescent Health 40, 4–21.

Choi, K.-H., Gregorich, S.E., 2009. Social network influences on male and female condom use among women attending family planning clinics in the US. Sexually Transmitted Diseases 36, 757–762.

Cornwell, B., Laumann, E.O., 2011. Network position and sexual dysfunction: implications of partner betweenness for men. American Journal of Sociology 117, 172–208.

Davey-Rothwell, M.A., Latkin, C.A., 2008. An examination of perceived norms and exchanging sex for money or drugs among women injectors in Baltimore, MD, USA. International Journal of STD & AIDS 19, 47–50.

Dodge, B., Jeffries, I.V., William, L., Sandfort, T.G.M., 2008. Beyond the down low: sexual risk, protection, and disclosure among at-risk black men who have sex with both men and women (MSMW). Archives of Sexual Behavior 37, 683–696.

Ennett, S.T., Bailey, S.L., Federman, E.B., 1999. Social network characteristics associated with risky behaviors among runaway and homeless youth. Journal of Health and Social Behavior 40, 63–78.

Friedkin, N.E., Johnsen, E.C., 2011. Social Influence Network Theory: A Sociological Examination of Small Group Dynamics. Cambridge University Press, New York.

Frost, S.D.W., 2007. Using sexual affiliation networks to describe the sexual structure of a population. Sexually Transmitted Infections 83 (Suppl. 1), i37–i42.

Hadley, W., Brown, L.K., Lescano, C.M., Kell, H., Spalding, K., DiClemente, R., Donenberg, G., 2009. Parent-adolescent sexual communication: associations of condom use with condom discussions. AIDS and Behavior 13, 997–1004.

Harding, D.J., 2007. Cultural context, sexual behavior, and romantic relationships in disadvantaged neighborhoods. American Sociological Review 72, 341–364.

Higgins, J.A., Mullinax, M., Trussell, J., Davidson, J.K., Moore, N.B., 2011. Sexual satisfaction and sexual health among university students in the United States. American Journal of Public Health 101, 1643–1654.

Horvath, K.J., Rosser, S., Remafedi, G., 2008. Sexual risk taking among young internet-using men who have sex with men. American Journal of Public Health 98, 1059–1067.

Julien, D., Chartrand, E., Begin, J., 1999. Social networks, structural interdependence, and conjugal adjustment in heterosexual, gay, and lesbian couples. Journal of Marriage and Family 61, 516–530.

Latkin, C.A., Forman, V., Knowlton, A., Sherman, S., 2003. Norms, social networks, and HIV-related risk behaviors among urban disadvantaged drug users. Social Science & Medicine 56, 465–476.

Latkin, C., Yang, C., Tobin, K., Roebuck, G., Spikes, P., Patterson, J., 2012. Social network predictors of disclosure of MSM behavior and HIV-positive serostatus among African American MSM in Baltimore, Maryland. AIDS and Behavior 16, 535–542.

Laumann, E.O., Ellingson, S., Mahay, J., Paik, A., Youm, Y., 2004. The Sexual Organization of the City. University of Chicago Press, Chicago, IL.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of the City: Sexual Practices in the United States. University of Chicago Press, Chicago, IL.

Mansergh, G., Shouse, R.L., Marks, G., et al., 2006. Methamphetamine and Sildenafil (Viagra) use are linked to unprotected receptive and insertive anal sex, respectively, in a sample of men who have sex with men. Sexually Transmitted Infections 82, 131–134.

Miranda-Diaz, M., Corcoran, K., 2012. "All my friends are doing it:" the impact of the perception of peer sexuality on adolescents' intent to have sex. Journal of Evidence-Based Social Work 9, 260–264.

Neblett, R.C., Davey-Rothwell, M., Chander, G., Latkin, C.A., 2011. social network characteristics and HIV sexual risk behavior among urban African American women. Journal of Urban Health 88, 54–65.

Ostrow, D.G., Plankey, M.W., Cox, C., Li, X., Shoptaw, S., Jacobson, L.P., Stall, R.C., 2009. Specific sex-drug combinations contribute to the majority of recent HIV seroconversions among MSM in the MACS. Journal of Acquired Immune Deficiency Syndromes 51, 349–355.

Prairie, B.A., Scheier, M.F., Matthews, K.A., Chang, C.-C.H., Hess, R., 2011. A higher sense of purpose in life is associated with sexual enjoyment in midlife women. Menopause 18, 839–844.

Rosenfeld, M.J., Thomas, R.J., 2012. Search for a mate: the rise of the internet as a social intermediary. American Sociological Review 77, 523–547.

Saewyc, E.M., 2011. Research on adolescent sexual orientation: development, health disparities, stigma, and resilience. Journal of Research on Adolescence 21, 256–272.

Schneider, J.A., Cornwell, B., Ostrow, D., Michaels, S., Schumm, P., Laumann, E.O., Friedman, S., 2013. Network mixing and network influences most linked to HIV infection and risk behavior in the HIV epidemic among black men who have sex with men. American Journal of Public Health 103, e28–e36.

Stall, R., Purcell, D.W., 2000. Intertwining epidemics: a review of research on substance use among men who have sex with men and its connection to the AIDS epidemic. AIDS and Behavior 4, 181–192.

Teitler, J.O., Weiss, C.C., 2000. Effects of neighborhood and school environment on transitions to first sexual intercourse. Sociology of Education 73, 112–132.

Treas, J., Giesen, D., 2000. Sexual infidelity among married and cohabiting Americans. Journal of Marriage and Family 62, 48–60.

Tyler, K.A., 2008. Social network characteristics and risky sexual and drug related behaviors among homeless young adults. Social Science Research 37, 673–685.

Vanlandingham, M., Knodel, J., Saengtienchai, C., Pramualratana, A., 1998. In the company of friends: peer influence on Thai male extramarital sex. Social Science & Medicine 47, 1993–2011.

Wang, K., Brown, K., Shen, S.-Y., Tucker, J., 2011. Social network-based interventions to promote condom use: a systematic review. AIDS and Behavior 15, 1298–1308.

Wolff, J.M., Crockett, L.J., 2011. The role of deliberative decision making, Parenting, and friends in adolescent risk behaviors. Journal of Youth and Adolescence 40, 1607–1622.

Sexual Debut

Anthony Paik, University of Iowa, Iowa City, IA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents an overview of sexual debut, which refers to an individual's first experience with vaginal intercourse. Sexual debut represents not only an important life course transition, but also a demographic behavior with significant implications for sexual and reproductive health. This article summarizes cross-national variation and historical trends in sexual debut in both developing and developed countries. It also describes determinants, concomitant behaviors, and consequences of sexual debut.

Sexual debut, or sexual initiation, refers to the first penile–vaginal intercourse during an individual's lifetime. The onset of sexual intercourse is particularly significant because of its implications for gender, the life course, sexual and reproductive health, the formation of cohabiting and marital unions, and human-capital accumulation. There is substantial historical and societal variation in sexual debut as it pertains to its timing, whether this transition occurs with or before marriage, and the degree to which there are gender differences. This article summarizes the following: recent patterns and trends across developed and developing countries since 2000; systematic reviews of risk and protective factors influencing its timing among youth; the partnership contexts and practices under which this transition occurs; and its implications for sexual and reproductive health outcomes, such as the acquisition of sexually transmitted infections, fertility, union formation, and human-capital accumulation.

The onset of sexual intercourse can be understood using one of several conceptual frameworks. First, sexual debut is linked to the transition to adulthood. A life course perspective highlights the social construction of cultural categories linked to physiological aging. From this perspective, sexual debut may be understood as a normative behavior linked to particular life stages (e.g., adolescence or adulthood), the transition from adolescence to adulthood, and physiological processes of aging, including pubertal development (Udry and Campbell, 1994). Second, social-control perspectives emphasize how the onset of sexual intercourse is a behavior governed by major societal institutions, such as the law, religion, and the family (DeLamater, 1981). Sexual debut occurs within the context of legal and moral prescriptions as well as the degree to which societies are permissive of premarital sexual intercourse. Finally, a third major framework for studying sexual debut is through the lens of gender (Schwartz and Rutter, 1998). The onset of sexual intercourse can be viewed as a gendered process or as a way of doing gender. Taken together, these frameworks highlight the importance of aging, gender, and social institutions for defining the normative and nonnormative contexts for the onset of first sexual intercourse.

Cross-National Variation and Trends

Cross-national comparisons indicate that there is substantial variation in the timing and context of sexual debut. **Figure 1** presents proportions of male and female 15–19 years olds who experienced first sexual intercourse based on representative surveys gathered since 2005 across 21 developing countries, Canada, and the United States. It is worth noting that this figure has relatively few data points from Europe, Asia, and South America. There is a large cluster of countries – Kenya, Uganda, Tanzania, Canada, United States, Zambia, Dominican Republic, Malawi, and the Democratic Republic of Congo – where a third to half of male and female 15–19 years olds experienced the onset of first intercourse. In contrast, China is the one country where the vast majority of 15–19 years olds, more than 90%, have not experienced sexual debut. Large proportions of 15–19 years olds did not report the onset of first sexual intercourse in Albania and Rwanda as well.

Figure 1 also shows substantial gender differences in sexual debut among 15–19 years olds in some countries. In Ethiopia, India, Nepal, Senegal, Zimbabwe, Ghana, and Nigeria, 1.5–3 times as many girls and women experienced sexual debut, compared to boys and men of the same age. For example, in India, 28% of female 15–19 years olds have had sexual intercourse, compared to 11% of male youth. This disparity suggests that many sexually active female youth become involved with older men in these countries. In contrast, the proportion of sexually experienced male youth, compared to female youth, is nearly four times as large in Guatemala and twice as large in Moldova.

To understand trends, Figure **2**(a) and 2(b) presents estimates of sexual debut before age 15 among female and male youth by cohorts from the 1990s and 2000s in selected African countries. Prior research found that the proportion of female youth experiencing sexual debut before age 15 decreased when comparing cohorts from the 1970s against the 1990s, but showed an increased proportion of male youth (Wellings et al., 2006). **Figure 2(a)** indicates that in 11 out of 12 African countries, the proportion of female youth who experienced sexual debut by age 15 had declined in the 2000s, compared to the 1990s. Moreover, with the exception of Mozambique, less than one-sixth of female youth became sexually active by age 15 across all of these countries. In contrast, based on **Figure 2(b)**, trends in sexual debut for male youth are less consistent. Most countries have declines in the proportion of male youth who experienced sexual debut by age 15 when comparing cohorts from the 1990s against the 2000s, but Malawi, Mozambique, Rwanda, and Uganda had increasing percentages over time. Still, in 8 out of 11 countries less than

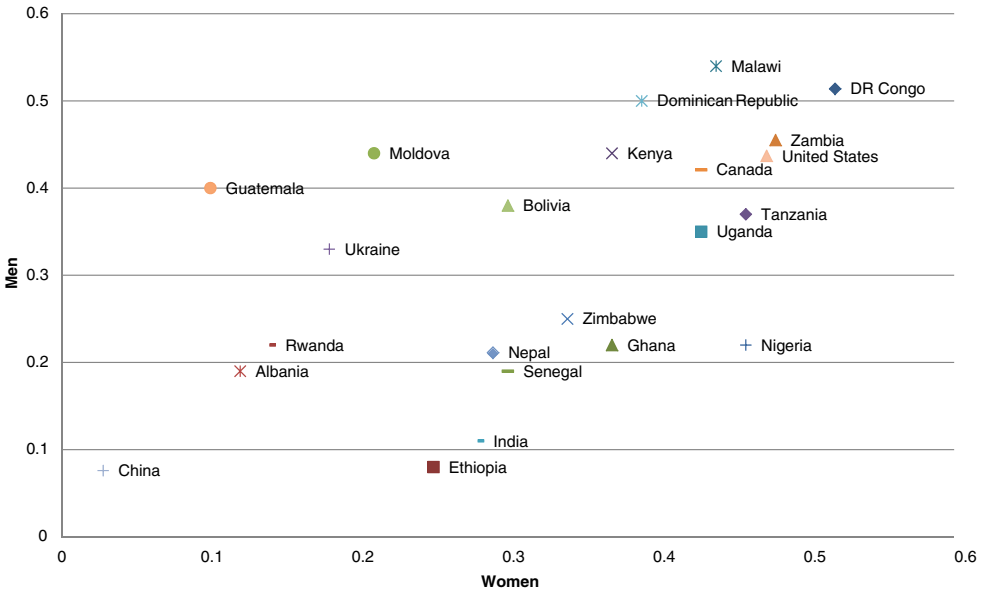

Figure 1 Proportion of 15–19 years olds that ever had intercourse by country and gender (national surveys since 2000). Sources: Anderson, R., Panchaud, C., Singh, S., Watson, K., 2013. Demystifying Data: A Guide to Using Evidence to Improve Young People's Sexual Health and Rights. Alan Guttmacher Institute, New York; Copen, C.E., Chandra, A., Martinez, G., 2012. Prevalence and Timing of Oral Sex with Opposite-Sex Partners among Females and Males Aged 15–24 Years: United States, 2007–2010. National Health Statistics Reports; No 56. National Center for Health Statistics, Hyattsville, MD; Roterman, M., 2008. Trends in teen sexual behaviour and condom use. Health Reports 19, 53–57; Song, Y., Ji, C., 2010. Sexual intercourse and high-risk sexual behaviors among a national sample of urban adolescents in China. Journal of Public Health 32, 312–321.

a sixth experienced sexual debut by age 15 in the 2000s. The long-run trend for female youth in many African countries is an increasing age of first sexual intercourse, whereas for male youth the trend is more mixed.

Longer trend data are available for developed countries. The median age of first intercourse has been declining since the 1950s in a number of developed countries, including Greece, Portugal, Iceland, Denmark, Finland, Norway, France, Switzerland, Great Britain, Belgium, the Netherlands, France, and the United States (Bajos et al., 2010; Finer, 2007; Teitler, 2002). This pattern is larger for female youth due to the fact that male youth tended to become sexually active earlier across all of these countries. However, at least in the United States, data since 2000 indicate that the proportion of sexually active youth is no longer increasing (Santelli and Melnikas, 2010).

A key issue is whether sexual debut is premarital or occurs within the context of a cohabiting or marital union. Many developed countries show a pattern of relatively early sexual debut but delayed union formation. Figure 3 presents median ages of first sexual intercourse and first union formation (cohabitation or marriage) among women aged 20–24, based on demographic health surveys since 2003, which showed considerable variation. For most countries, the median age of first sexual intercourse is lower than the median age of first union, but in three countries – Ethiopia, India, and Nepal – the latter is the same as or lower than the former. This suggests that first intercourse occurs within the context of marriage for many female youth in these countries and that many become married during adolescence. In other countries, such as Mozambique, Uganda, and Tanzania, the median age of first sex is lower than the median age of first marriage, yet both still occur relatively early. In Columbia and Peru, the median age of first sex is more than 3 years lower than the median age of marriage, whereas both occur relatively late in Rwanda, Philippines, and Albania.

Correlates and Contexts of Sexual Debut

This section focuses on nonmarital sexual debut, which is age graded. In an analysis of a recent cohort in the United States, for example, the proportion of female youth who had experienced sexual debut jumped from around 10% at age 14 to nearly 70% by age 18 (Finer, 2007). The substantial age grading of sexual debut reflects its formal regulation via age of consent laws as well as beliefs and attitudes about the appropriateness of becoming sexually active.

Determinants of Sexual Debut

Table 1 lists correlates of the occurrence or timing of adolescent sexual debut based on three recent systematic reviews of the literature in developing countries (Mmari and Blum, 2009), the United States (Buhi and Goodson, 2007), and the United Kingdom (Hawes et al., 2010). In terms of personal factors, older age of menarche and spermarche are associated with delayed sexual debut. For female youth, in particular, age of menarche is one of the major determinants of later age of sexual debut. An explanation for this association is that early pubertal development may cause individuals to treat younger girls as if they were older. Other personal factors include self-esteem, intentions to have sex, and alcohol or drug use. Not surprisingly, intentions to have sex are associated with increased likelihoods to experience early sexual debut, whereas

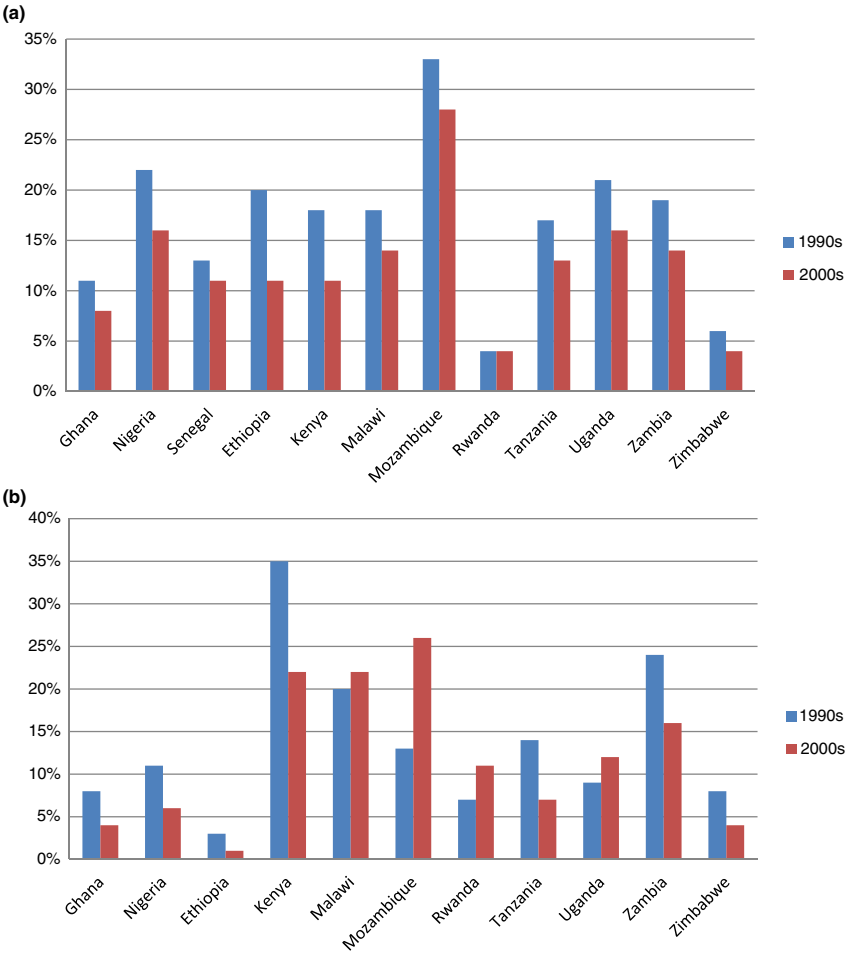

Figure 2 (a) Percent of female youth who had first intercourse by age 15. (b) Percent of male youth who had first intercourse by age 15. Note: National estimates are based on respondents who were 15–24 years old for the 2000s and 20–24 for the 1990s. Sources: Anderson, R., Panchaud, C., Singh, S., Watson, K., 2013. Demystifying Data: A Guide to Using Evidence to Improve Young People's Sexual Health and Rights. Alan Guttmacher Institute, New York; Bankole, A., Singh, S., Woog, V., Wulf, D., 2004. Risk and Protection: Youth and HIV/AIDs in Sub-Saharan Africa. Alan Guttmacher Institute, New York.

self-esteem is associated with delay. Alcohol and drug use are associated with an increased likelihood of early sexual debut. Substance use may not only impair youths' decision making, but also expose them to environments and individuals where sexual activity is more likely to be normative.

A number of family characteristics have been linked to the timing of sexual debut, including parents' background characteristics, family structure, and aspects of parent–child relationships. Higher socioeconomic status (SES) and being raised by two parents is associated with older age of first intercourse, whereas residential mobility and having a mother or sibling who had children early are linked with early sexual debut. Children in higher SES households may tend to have higher opportunity costs of becoming sexually active. Consistent with social control perspectives, parental closeness, parental monitoring, and parental disapproval of sex have all been linked to later age of sexual debut.

School and community-level factors associated with the delay of sexual debut include school enrollment, educational aspirations, academic performance, attending a high-quality school, receiving sex education, and the frequency of religious service attendance. The effect of schooling on the timing of sexual debut is notable; it suggests that educational attainment increases the opportunity costs of becoming sexually active. Schooling is also associated with later formation of a first sexual union. From a policy perspective, this suggests that enhancing educational opportunities in many nations, particularly for female youth, will lead to delayed sexual debut. Another notable association is the impact of sex education on sexual debut in the United Kingdom. Sex education represents an important intervention that could be administered across numerous contexts. It is important to note that the efficacy of abstinence-only sex education is questionable; quasi-experimental studies suggest that it is not effective in comparison to abstinence-plus or comprehensive sex education (Trenholm et al., 2008).

Finally, peer approval of sex and peer sexual behavior are associated with early sexual debut. It is important to note

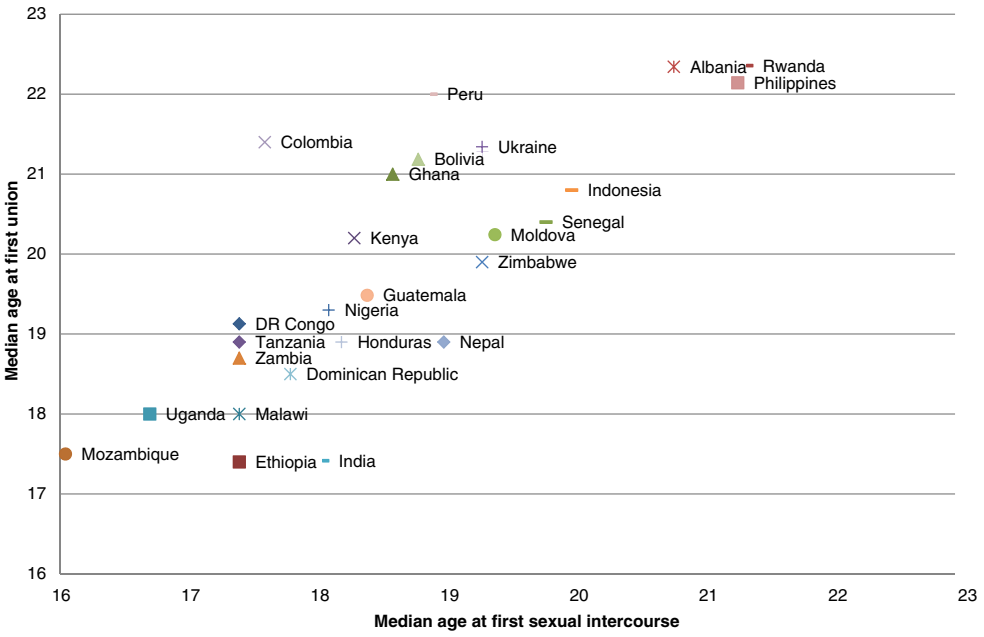

Figure 3 Median age of first intercourse and first sexual union among women (20–24 years old). Source: Anderson, R., Panchaud, C., Singh, S., Watson, K., 2013. Demystifying Data: A Guide to Using Evidence to Improve Young People's Sexual Health and Rights. Alan Guttmacher Institute, New York.

that research in this area has relied on respondents' perceptions of their peers' sexual attitudes and behaviors. As such, correlations between these factors and the timing of sexual debut may reflect projection bias – that is, the tendency of survey respondents to ascribe their attitudes and behaviors on their peers. Research is needed that utilizes peer measures that are not based on survey respondents' perceptions.

Table 1 Correlates of early sexual debut

Variables	*Developing*	*US*	*UK*
Personal			
Age of puberty			–
Intentions to have sex		+	+
Self-esteem		–	
Alcohol or drug use	+		+
Family and social background			
Mother had children at young age			+
Parents' SES			–
Family instability	+		+
Older sibling pregnant as an adolescent	+		
Residential mobility	+		
Parental closeness and relationship quality		–	
Parental control, monitoring, and supervision		–	–
Parental disapproval of sex		–	
Schooling and community level			
School enrollment	–		–
GPA	–		
Educational aspirations	–		
School quality			–
Sex education			–
Religiosity			–
Peer and partner-level			
Perceive friends have sex	+	+	
Peer approval of sex		+	+

+ denotes an increasing risk of earlier sexual debut; – indicates a correlate that delays sexual debut.

Concomitant Aspects of Sexual Debut

In addition to the timing and determinants of sexual debut, there are a number of concomitant aspects. Researchers have focused on the relationship contexts of sexual debut as well as partner characteristics, contraceptive use, alcohol or drug use, degree of wantedness and consensuality, and associated emotions and feelings. In general, younger ages of sexual debut have been linked to unprotected sex, impaired judgment due to alcohol use, and unwanted experiences (Hawes et al., 2010; Paik, 2011). Research also focuses on the relationship contexts of sexual debut, such as whether partners are dating or casual, age differences between partners, and concomitant sex practices. A systematic review of 268 qualitative studies of sexual behavior among youth identified a number of themes related to sexual debut (Marston and King, 2006). Specifically, possession of condoms, which implied to many youth sexual experience and the lack of trust, were particularly stigmatizing for female youth. These beliefs may impede the utilization of condoms at first sexual intercourse. Likewise, displaying chastity was a particularly important theme for female youth.

Consequences of Sexual Debut

The timing and concomitant aspects of sexual debut have been linked to sexual and reproductive health, including sexually transmitted infections, fertility outcomes, and union formation. For example, a systematic review of the literature showed that early sexual debut was associated with an increased risk of HIV among sub-Saharan African women (Stöckl et al., 2012). This increased risk is linked to a number of factors, including subsequent risky sexual behavior (e.g., increased numbers of sex partners), partner characteristics, and possibly biological factors. Also, early sexual debut is associated with a number of fertility outcomes, such as early pregnancies, abortions, and live births (Santelli and Melnikas, 2010), because of the longer exposure period, decreased likelihood of using contraceptives, and increased likelihood of having more partners. Finally, research has linked early sexual debut to early union formation (Miller and Heaton, 1991), lower educational attainment (Sabia and Rees, 2009), and divorce (Paik, 2011).

Current and Emerging Directions in Theory and Research

Methodological Issues or Problems

Sexual debut is correlated with numerous outcomes. Some of these are direct consequences of early sexual debut, whereas the casual linkages can be indirect as well, resulting from increased exposure to sexual activity. In these cases, the causal linkages are fairly clear. Support for causal arguments are less clear for other associations, such as those between sexual debut and subsequent social and demographic behaviors, such as educational attainment, union formation, and divorce. Here, these associations may not be causal, so there is a need for research that can make more credible causal claims linking the timing of sexual debut to subsequent demographic and social behaviors. Also, research on the timing of sexual debut relies on self-reports. However, experimental research suggests that men may overstate their sexual activity, whereas women will underreport it. Few studies have sought to examine the extent of these biases regarding sexual debut.

Future Research

There are a number of current and emerging research issues regarding the timing of sexual debut. First, there is burgeoning interest in the relationship contexts of sexual behavior. For many, sex occurs with the context of sexual unions, such as marriage and cohabiting relationships. Among noncohabiting relationships, sex is often embedded in romantic or serious relationships. Recently, there is increased attention to non-romantic contexts of sex, such as casual sex, but relatively little attention has focused on this issue with respect to sexual debut. A second theme of current interest is the extent to which sexual debut is completely voluntary and wanted. For example, in the United States, a substantial percentage of individuals reported that the onset of first intercourse was not completely wanted (Paik, 2011). There is a strong age grading to this, yet there is a need for more research. Finally, an emerging theme is the need for increased attention to contextual and cultural factors associated with the timing of sexual debut. Research has focused primarily on individual level and social background predictors of the timing of first sex. At the same time, there is substantial cross-national variation, which highlights the importance of varying social contexts and cultural beliefs that give rise to this variation. In sum, sexual debut is not only a significant life course transition, but one that has important policy implications.

See also: Adolescent Sexual Risk; Femininities; Heterosexuality; Hooking up and Casual Sex; Intimate Partner Violence; Masculinity, Scripts of; Methodological Issues in the Study of Sexuality; Sexual Behavior and Social Networks; Sexual Motivation; Sexuality Education; Sexuality Over the Life Course.

Bibliography

Anderson, R., Panchaud, C., Singh, S., Watson, K., 2013. Demystifying Data: A Guide to Using Evidence to Improve Young People's Sexual Health and Rights. Alan Guttmacher Institute, New York.

Bajos, N., Bozon, M., Beltzer, N., Laborde, C., Andro, A., Ferrand, M., Goulet, V., Laport, A., Le Van, C., Leridon, H., Levison, S., Rzafindratsima, N., Toulemon, L., Warszawski, J., Wellings, K., 2010. Changes in sexual behaviours: from secular trends to public health policies. AIDS 24, 1185–1191.

Bankole, A., Singh, S., Woog, V., Wulf, D., 2004. Risk and Protection: Youth and HIV/AIDs in Sub-Saharan Africa. Alan Guttmacher Institute, New York.

Buhi, E.R., Goodson, P., 2007. Predictors of adolescent sexual behavior and intention: a theory-guided systematic review. Journal of Adolescent Health 40, 4–21.

Copen, C.E., Chandra, A., Martinez, G., 2012. Prevalence and Timing of Oral Sex with Opposite-Sex Partners among Females and Males Aged 15–24 Years: United States, 2007–2010. National Health Statistics Reports; No 56. National Center for Health Statistics, Hyattsville, MD.

DeLamater, J., 1981. The social control of sexuality. Annual Review of Sociology 7, 263–290.

Finer, L.B., 2007. Trends in premarital sex in the United States, 1954–2003. Public Health Reports 122, 73.

Hawes, Z.C., Wellings, K., Stephenson, J., 2010. First heterosexual intercourse in the United Kingdom: a review of the literature. Journal of Sex Research 47, 137–152.

Marston, C., King, E., 2006. Factors that shape young people's sexual behaviour: a systematic review. Lancet 368, 1581–1586.

Miller, B.C., Heaton, T.B., 1991. Age at first sexual intercourse and the timing of marriage and childbirth. Journal of Marriage and the Family 53, 719–732.

Mmari, K., Blum, R.W., 2009. Risk and protective factors that affect adolescent reproductive health in developing countries: a structured literature review. Global Public Health 4, 350–366.

Paik, A., 2011. Adolescent sexuality and the risk of marital dissolution. Journal of Marriage and Family 73, 472–485.

Roterman, M., 2008. Trends in teen sexual behaviour and condom use. Health Reports 19, 53–57.

Sabia, J.J., Rees, D.I., 2009. The effect of sexual abstinence on females' educational attainment. Demography 46, 695–715.

Santelli, J.S., Melnikas, A.J., 2010. Teen fertility in transition: recent and historic trends in the United States. Annual Review of Public Health 31, 371–383.

Schwartz, P., Rutter, V.B., 1998. The Gender of Sexuality. Rowman & Littlefield Publishers, New York.

Song, Y., Ji, C., 2010. Sexual intercourse and high-risk sexual behaviors among a national sample of urban adolescents in China. Journal of Public Health 32, 312–321.

Stöckl, H., Kalra, N., Jacobi, J., Watts, C., 2012. Is early sexual debut a risk factor for HIV infection among women in sub-Saharan Africa? A systematic review. American Journal of Reproductive Immunology 69, S27–S40.

Teitler, J.O., 2002. Trends in youth sexual initiation and fertility in developed countries: 1960–1995. Annals of the American Academy of Political and Social Science 580, 134–152.

Trenholm, C., Devaney, B., Fortson, K., Clark, M., Quay, L., Wheeler, J., 2008. Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. Journal of Policy Analysis and Management 27, 255–276.

Udry, J.R., Campbell, B.C., 1994. Getting started on sexual behavior. In: Rossi, A.S. (Ed.), Sexuality across the Life Course. University of Chicago Press, Chicago, IL, pp. 187–206.

Wellings, K., Collumbien, M., Slaymaker, E., Singh, S., Hodges, Z., Patel, D., Bajos, N., 2006. Sexual behaviour in context: a global perspective. Lancet 368, 1706–1728.

Sexual Dysfunction, Epidemiology of

David L Rowland and Beth A Adamski, Valparaiso University, Valparaiso, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past two decades, sexual dysfunction in men and women has been the topic of significant inquiry and research. Included in this article are explanations of each of the dysfunctions and their relationship to the sexual response cycle, discussion of the classification systems, and estimated prevalence rates for both men and women. In addition, the authors discuss common risk factors and mention typical treatment strategies currently available for each of the dysfunctions.

General Orientation

This article focuses on the sexual dysfunctions, that is, problems in responding adequately to achieve a sexually satisfying life, usually within the context of a sexual relationship. Problems typically categorized under the sexual dysfunctions include, for example, a man's inability to get or keep an erection or a woman's inability to reach orgasm – problems specific to adult sexual behaviors that interfere with having a sexually fulfilling relationship. Each type of disorder has its own etiology, causes, and effects, and each is typically treated by a specialist having a deep understanding of the issue as well as effective treatment strategies. Professionals who deal with any of the above disorders typically use diagnostic guides such as the *Diagnostic and Statistical Manual of Mental Disorders* (DSM; American Psychiatric Association, 2000; 2013) or the *International Statistical Classification of Diseases and Related Health Problems* (ICD; World Health Organization, 2013) to identify, classify, and diagnose the problem. New editions have provided revision to the classification systems, as increased understanding and further research of the disorders is constantly developing. This article provides information primarily based on the *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision* (DSM-IV-TR) and ICD-10 classification systems, while noting changes made in the recently released DSM-5 classification system. The ICD-10 classification is closely aligned with the DSM-IV-TR, and the ICD-11, with its anticipated 2015 release, is expected to be organized in a manner similar to the DSM-5.

Understanding Sexual Response

Sexual response is complex, requiring specific preconditions, incorporating a variety of behavioral responses, and including an array of psychosocial and interpersonal factors that relate to emotional, cognitive, and relationship dimensions. However, for purposes of simplification, it is usually divided into three phases: desire, arousal, and orgasm (Masters and Johnson, 1966; Kaplan, 1974). More recent refinement distinguishes 'unprompted' or 'spontaneous' desire from arousability, the latter referring to sexual desire derived from a specific individual, object, or context. Further conceptualization has included separate pain–pleasure dimensions (Schover et al., 1982) as well as attention to other factors such as the feelings, motivations, and attitudes that surround the sexual act (Byrne and Schulte, 1990).

The classification of sexual dysfunctions has evolved, to a large extent, from the conceptual models mentioned above and includes

- a lack of desire, also known as hypoactive sexual desire disorder (HSDD),
- problems with either physiological (e.g., genital) or psychological sexual arousal,
- disorders of ejaculation/orgasm, most typically premature ejaculation (PE) in men and inhibited orgasm in men and women, and
- problems with painful intercourse and/or sexual aversion, both of which may lead to avoidance of sexual intimacy.

Sexual problems are also characterized as either situation-specific (including person) or generalized, and as either lifelong or acquired (through either pathophysiological developments or sexual experiences). As mentioned above, several classification systems are currently in use to define sexual dysfunctions; DSM-IV-TR (American Psychiatric Association, 2000), the relatively new DSM-5 (American Psychiatric Association, 2013), ICD-10 (ICD; World Health Organization, 2013), and International Society of Sexual Medicine (McMahon et al., 2008) are among the more common ones. Each system has categories related to a lack of sexual interest/desire, the inability to become adequately aroused either physically (in terms of genital response) or psychologically (feeling aroused), and problems with orgasm – either reaching orgasm too quickly or reaching it only after a very long time, or perhaps not at all (see **Table 1**). Together, desire, arousal, and orgasm are sometimes referred to as the 'sexual response cycle', and the approach that considers potential biological, psychological, and relationship factors that affect the sexual response cycle is referred to as the 'biopsychosocial' model of human sexual response. An overall summary of risk factors for sexual problems and treatment approaches to those problems, as modeled by the biopsychosocial approach, is provided in **Figure 1**.

While the various components of the sexual response cycle are essentially the same across the sexes, they may be manifested differently as the result of different genital structures. For example, physiological genital arousal is manifested by penile erection in men; the analogous response in women is vaginal lubrication and clitoral erection. In addition, as is apparent

Table 1 Classification of sexual dysfunctions and defining features

	Name	*Defining features*
Disorders of desire/interest and arousal	Male hypoactive sexual desire disorder (HSDD)	Diminished or absent desire for sexual activity
	Male erectile disorder/erectile dysfunction	Difficulty getting and keeping an erection
	Female sexual interest/arousal disorder	Diminished or absent interest in sexual activity, diminished or absent subjective arousal and excitement during sexual activity, difficulty attaining and maintaining adequate genital arousal (e.g., lubrication or vaginal swelling)
	Sexual aversion disorder (SAD) (male and female)	Persistent or frequent strong aversion to or avoidance of sexual activity with a partner *This disorder removed from the DSM-5
Disorders of orgasm/ejaculation	Male orgasmic disorder/inhibited ejaculation (IE)	A delayed or complete inability to reach ejaculation following a normal sexual excitement phase
	Premature ejaculation (PE)	Ejaculating very quickly because the man is unable to delay it
	Female orgasmic disorder (FOD)	Delayed or inhibited orgasm following a normal sexual excitement phase
Disorders of pain	Genitopelvic pain/penetration disorder (male and female)	Recurrent or persistent genital pain that interferes with sexual activity (categorized in DSM-IV-TR as male dyspareunia, female dyspareunia, and female vaginismus)

later in this article, sexual dysfunctions related to each component of the sexual response cycle are distributed differently across men and women, with men having greater problems with erection and ejaculating too quickly, and women having greater problems with sexual desire and reaching orgasm.

The above segmenting of the sexual response cycle into specific phases is helpful to health professionals because problems with each phase have their own etiology and causes and therefore may require their own specific treatment. Nevertheless, these categorizations have little meaning for most lay people who view their sexual experiences as holistic events that are either 'satisfying or not satisfying'. Therefore, when dealing with sexual dysfunctions, the health professional may need to spend significant time helping the client and his or her partner understand not only these elements of sexual response, but also each of their potential causes. Furthermore, although an individual or couple may be dissatisfied with their sexual response and relationship, this does not necessarily signal a dysfunction – specific criteria are used for diagnosing a sexual dysfunction, and unless an individual meets those criteria, he or she may have a sexual problem or be 'sexually dissatisfied', but not 'sexually dysfunctional'. Finally, healthy sexual relationships are not characterized merely by the absence of dysfunctional response. Key elements of healthy sexual relationships include passion, intimacy and caring, and commitment (Sternberg and Barnes, 1988). Passion typically involves such characteristics as sexual feelings, physical attraction, and romantic love. Intimacy deals with dimensions of affection, caring, and expressiveness and the willingness to communicate and share beliefs, attitudes, and feelings. Commitment refers to the decision to be with one partner and to work hard to maintain the relationship. As many sexual problems are rooted in couple's disparate expectations and emotional struggles, including the different ways in which these elements are often played out by each of the sexes, most sexual problems benefit from attention to general relationship factors as well as to specific sexual response issues.

Epidemiology, Terminology, and Related Issues

Epidemiology is a branch of medicine that deals with the frequency, distribution, and spread of disease or health-related events within populations, with attention to those factors that cause, influence, and control them. With respect to sexual dysfunction, epidemiologists are more likely to refer to 'risk' factors than 'causal' factors. That is, certain predisposing biological, psychological, and relationship factors may place a man or woman at higher probability for a particular sexual dysfunction, but these factors do not, in the classic sense of the word, determine or cause the dysfunction. Specifically, men or women having similar sets of risk factors may exhibit very different levels of sexual functionality (or dysfunctionality).

Epidemiologists also distinguish between incidence and prevalence. Incidence is the number of new cases within a population over a given period of time. Prevalence, on the other hand, is the total number of cases within a population at a given time. With respect to sexual dysfunction, prevalence, not incidence, is more typically reported. As indicated previously, the prevalence rate for each defined sexual dysfunction is different, and the prevalence rates differ across the sexes as well.

Determining the prevalence of any sexual dysfunction in men and women is challenging in that the reported occurrence of any sexual problem is subject to numerous factors. Specifically, defining a problem as a sexual dysfunction requires that the functional impairment meet specific criteria – in many instances, a person may view his or her response as problematic, but it may not meet the criteria necessary for a clinical diagnosis. Even more problematic, often the operational criteria for diagnosing a dysfunction differ from one classification system to the next – the use of different criteria will affect

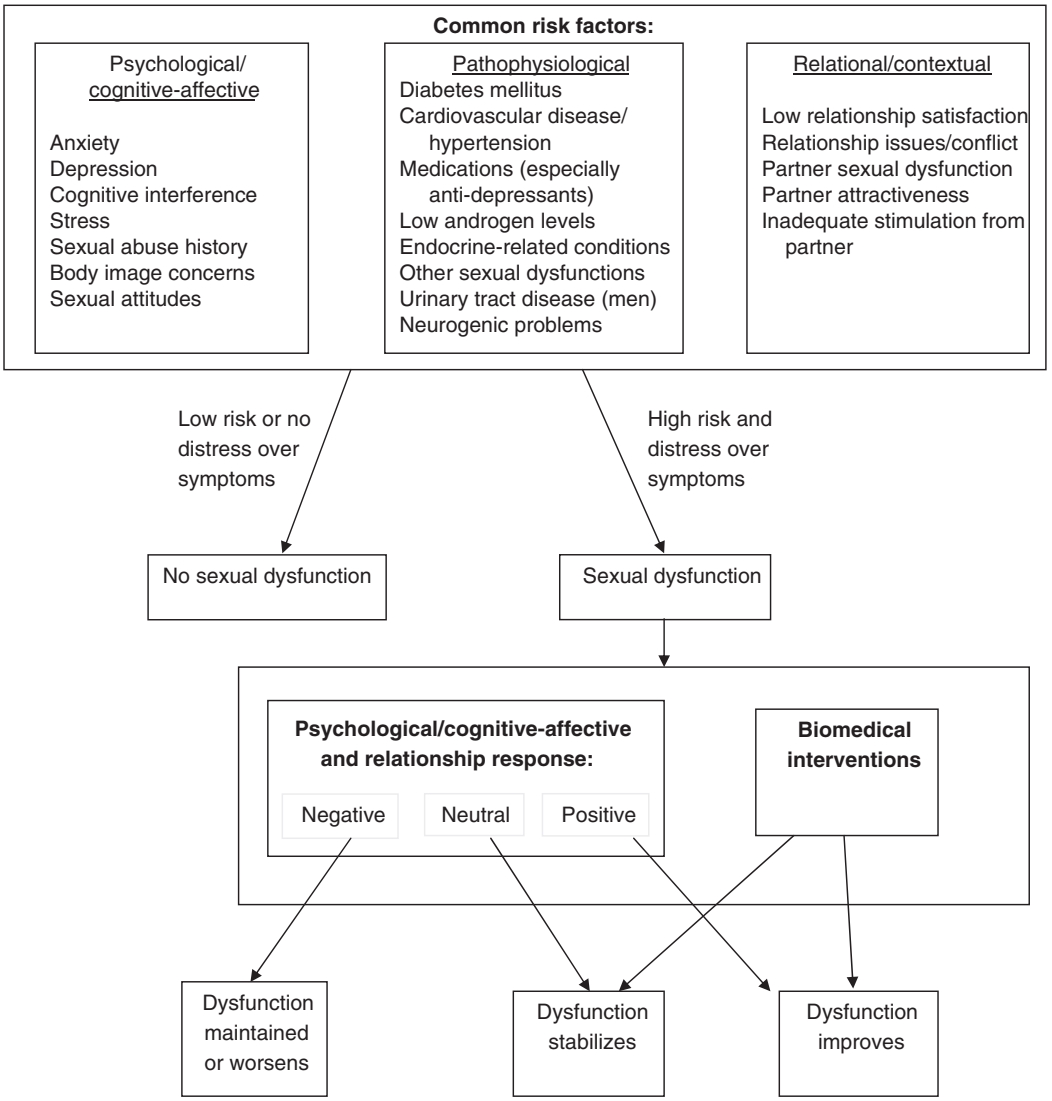

Figure 1 A model illustrating common risk factors for sexual dysfunctions in men and women, including how, based on a biopsychosocial model, psychosocial factors and biomedical interventions may affect the problem and its outcome.

which persons are included in the dysfunctional classification and which are excluded.

Furthermore, for persons to be classified as dysfunctional requires that they recognize the symptoms and potentially define themselves in this category. And, issues of privacy, discretion, and sometimes stigmatization may inhibit people from disclosing what might be viewed as 'inadequacies'. Additionally, a diagnosis of a dysfunction generally requires that the individuals find the symptoms distressing, although not all specialists agree on the need for inclusion of this criterion.

Finally, prevalence is a moving target. For example, for some people, the problem may be transient, as when life stressors or particular sensitivity or vulnerability to specific relationship interactions contribute to or intensify a problem that is initially 'subclinical' – issues that DSM-5 has attempted to address in part by requiring that symptoms be present for a minimum duration of 6 months and by including more precise severity criteria. Changing cultural attitudes, expectations, and acceptance regarding men's and women's sexual problems may influence the willingness to acknowledge and report sexual problems – as such, tendencies for self-reporting problems will change as well. Such factors, along with general health and diet, result in prevalence rates that differ across time and across geographic regions of the world. And, finally, for those sexual dysfunctions that are partly age related (e.g., problems with erection and/or sometimes sexual interest), as the age demographic changes, so also does the prevalence (Rowland, 2012).

Epidemiology of Men's Sexual Dysfunctions

In this section, prevalence rates and risk factors for disorders of desire, arousal (including erectile problems), and ejaculation in men are reviewed, with comment on the prognosis and

treatability of each of the dysfunctions. Pain disorders and sexual aversion are also discussed briefly.

Low Sexual Desire

Recent data from the Global Sexual Attitudes and Behaviors Study (GSSAB; Laumann et al., 2005) suggest a prevalence rate for low sexual desire (also called HSDD) somewhere between 13 and 28%. Several other studies report rates ranging from 0 to about 25%, depending on the nature of the question, the population sampled, and whether the individuals were clinically or self-diagnosed. Taken together, these and other studies (Lewis et al., 2010) suggest a prevalence rate of somewhere between 15 and 25% of men, with higher levels in men over 60 years. With respect to sexual desire, it is important to distinguish between sexual 'desire' and 'interest'. Desire is presumed to be internally driven and physiologically based (at least for men), whereas interest is a more inclusive term that may be affected by such factors as partner attraction, relationship issues, environmental stressors, and so on (Rowland, 2012).

Risk factors related to HSDD in men are often biological and sometimes related to endocrine abnormalities. For example, low levels of androgens, testosterone in particular, are most commonly implicated in low sexual desire. In men, androgens decline with aging, gonadal (testicular) dysfunction, high prolactin levels, and various pathophysiological conditions. A relevant factor is not just the overall level of testosterone circulating in the blood, but the availability of that testosterone (called 'free' testosterone because it is not bound to molecules that prevent it from being biologically active) to exert an action on neural structures.

Sexual desire, as powerful as it is at times, exists within a hierarchy of biological and psychological needs. When other needs are not being met (usually a condition that induces stress), sexual desire often wanes. Thus, general life stressors such as disease, depression, psychobehavioral problems, and crisis typically result in low sexual desire – in such instances the HSDD is secondary to another (primary) condition, indicating that treatment of the primary condition will likely ameliorate the low sexual desire.

Relationship and partner issues sometimes play a role in low sexual desire in men – for example, a man's partner may have a sexual problem (e.g., anorgasmia or pain during sex) or an aversion to sex, or the man may not find his partner sexually attractive. Such conditions may result in low sexual interest, which may appear as low sexual desire. However, such men often find other sexual outlets and therefore the low interest is directed toward a specific partner or situation. That is, sexual desire may be normal, but interest in the partner is diminished or absent. Relationship therapy involving better communication between partners, an expanded repertoire of sexual behaviors, specified sex 'dates', and rekindling sexual interest may help overcome the problem, assuming the couple is motivated to work together to improve their sexual relationship.

Arousal Disorders

In men, the most common arousal problem is getting and keeping an erection, clinically referred to as erectile dysfunction (ED). Seldom do men have a problem feeling sexually aroused when they do have an erection – a disconnection between feeling aroused and being genitally aroused is more likely to occur in women.

ED is fairly common, affecting as many as 30 million American men and approximately 5–50% of all men depending on their age group (see Lewis et al., 2010). Because erectile response assumes well-functioning vascular and neural systems, the prevalence of ED increases with age, as these systems exhibit age-related diminished responsiveness. Thus, the prevalence may be as low as 10% in men under 40 years, with incrementally increasing prevalence to over 50% in men over 70 years.

What is not reflected in these numbers is the percentage of ED men whose etiology is primarily psychogenic rather than biogenic – recent epidemiological studies have not distinguished between the two. Obtaining such data would be challenging, given the overlap between categories and the often transient nature of psychogenic ED. However, the fact that the prevalence increases from a low of about 10% to as high as 50% in older cohorts (an increase that is primarily due to aging and thus biogenic factors) suggests that the prevalence of psychogenic ED is probably no more than 10% in the overall population (Rowland, 2012).

Risk factors may be broadly categorized into one of two domains: (1) pathophysiological, those arising from a general disease state; and (2) psychosocial, those arising from a negative affective state, most often related to relationship/partner and/or performance issues.

Pathophysiological Problems

Given that erection results from engorgement of blood to the penis, diseases or conditions that compromise cardiovascular function have the potential to interfere with erectile response. Furthermore, since vascular response is under the control of the neural system, neuropathy involving the autonomic nervous system may also interfere with erectile response.

- Diseases that affect cardiovascular and/or neural function are typically age related and sometimes fatal. Generally, then, as a biogenically based dysfunction, ED is not a specific disease process, but rather just one manifestation of a more general cardiovascular malfunction (Lewis et al., 2008). As a result, the common approach in the medical community is that a man having ED due to biogenic reasons usually has more significant health issues that require attention. Specific disease conditions likely to produce vascular-based ED include diabetes mellitus, hypercholesteremia, and hypertension, conditions that may result in part from particular long-term life style choices (e.g., smoking, lack of exercise, dietary preferences, and so on). Obesity is also a general risk factor for ED, although it is yet unclear if it represents an independent factor or whether its effect is tied to cardiovascular disease and diabetes.
- Neurogenic problems typically affect erectile response by interfering with autonomic pathways that allow dilation of vessels coursing to the penis, that is, the discrete neural control required to stimulate and control the smooth

muscle that allows increased blood flow to the penis is disrupted. Diseases leading to such erectile problems include Parkinson's, epilepsy, encephalitis, multiple sclerosis, and various other neural degenerative diseases. Strokes, accidents, brain and spinal injury, surgery, and other trauma, depending on their location, extent, and nature, may also affect erectile function either directly or indirectly. Trauma, for example, that directly affects the lower spinal cord is quite likely to disrupt erectile response.

Psychosocial Factors

The way people think, feel, behave, and relate to others can affect erectile response. In many instances, the psychological factors are related to general life events and therefore likely to affect erectile response only indirectly, for example, by attenuating sexual desire or arousal. But in other instances, these factors may be tied directly to concerns about getting an erection. Men having such concerns are likely to feel anxiety about their performance, to expect failure, to make behavioral adjustments such as trying to penetrate the partner in a semiflaccid state, and to overfocus on getting an erection – all strategies that can worsen the problem and interfere with positive dynamics within the relationship.

Currently, effective treatments for men with ED of any origin, biogenic or psychogenic, are available. A class of drugs called the phosphodiesterase 5-inhibitors (or PD-5i: Viagra, Cialis, Levitra, etc.), which increase blood flow to the penis, sometimes used in conjunction with brief sex education and counseling, can help men return to a more satisfactory sexually functional state.

Ejaculatory Disorders

Disorders of ejaculation are many – dry ejaculation, retrograde ejaculation, painful ejaculation, orgasmic anesthesia, and so on – but the most common and bothersome problems center on the timing of the event. Some men ejaculate very quickly because they are unable to delay it – a condition known as premature ejaculation; others have a difficult time reaching ejaculation or are unable to ejaculate at all – a condition known as inhibited ejaculation (IE).

Premature (Early) Ejaculation

Most studies estimate 10–30% of men experience PE at some point in their lives (Rowland, 2011). Variability in estimates arises from the use of different definitions, whether the estimate is based on clinical diagnosis versus self-report by the respondent, and the extent to which the man is sufficiently concerned about the condition to seek treatment – only about half the men experiencing PE actually seek help.

Although data are limited, estimates are that about 75% of men with PE have primary (or lifelong) PE (McMahon, 2002), and that the other quarter has acquired it at some point during their sexual lives. Acquired PE may result from either pathophysiological or psychological factors. About a third of men with PE also report having erection problems, so knowing which problem is primary and which is secondary (the ED or the PE) is important, as treatment needs to begin with the primary problem. Some research indicates that age is not a factor with respect to the prevalence of PE; other research, however, suggests that PE may either increase or decrease with increasing age. Given the inconsistent results, the most parsimonious position is to assume that there is no large change in PE prevalence related to age.

For most men with PE, no obvious pathophysiology exists; that is, no underlying problem in structural or functional anatomy, physiology, or biochemistry can be identified. When PE does have a pathophysiological origin, it is usually fairly easy to identify and is likely to surface during a medical history and examination. For example, among the conditions commonly associated with PE are lower urinary tract (such as prostatitis and urethritis) and endocrine problems, particularly hyperthyroidism. Sometimes, however, the site of the pathophysiology may be far removed from the pelvic or genital area; for example, cerebrovascular accidents have sometimes been linked to PE. Specific drugs may sometimes induce PE, although not necessarily with any reliability. For example, use of L-DOPA, pimozide (an antipsychotic used to control tics), amphetamine, heroin, and over-the-counter drugs that mimic sympathetic activation have on occasion been associated with PE.

Most current thinking views ejaculatory latency as showing 'natural' variation across individuals, suggesting some degree of 'hardwiring' of the ejaculatory reflex and latency. It is plausible that, as with most biobehavioral responses (e.g., eating and emotion), inherent differences in response tendencies exist and that, for the specific biobehavioral response of ejaculation, different men possess predispositions to respond with shorter or longer latencies. However, the actual inherent physiological properties that differentiate men with PE from men with more typical latencies are unknown.

Surprisingly little is known about psychological factors related to PE. The question whether men with PE are somehow fundamentally different from other men has long been a topic of interest. For example, several studies suggest that men with PE may have heightened social-cognitive or affective vulnerability. Specifically, anxiety, a construct considered central to many psychological disorders, has also been postulated to have a role in psychogenic sexual dysfunction, including PE. For example, men with PE on average display more trait- and situation-based anxiety and depression (Corona et al., 2004) than the general population. What remains unresolved is the reason or cause for the anxiety itself – whether it is genetically or developmentally predispositional.

Perhaps not surprising in view of the emotional characteristics and disturbances discussed above, one interpretation for the above associations is that heightened risk for affective problems, inhibited social confidence, and amplified impact of external influencers – including social norms and/or partner feedback – constitutes specific vulnerability in the realm of 'social-affective' functioning, at least for a subpopulation of men with PE (Rowland, 2012). Although the social-affective functioning of men with PE does not fall in the clinical range, this social-affective vulnerability may heighten the probability for developing and intensifying negative (and counterproductive) thought patterns – patterns known to contribute to, maintain, or exacerbate functional impairments in many psychological realms, including PE (Patrick et al., 2007).

More so than the other male sexual dysfunctions, PE is considered a couple's problem: the dysfunction does not typically exist outside the man's relationship with his partner. Thus, therapies often include both partners. In some situations, partner characteristics may also contribute to the problem, for example, when the partner has a sexual aversion, exhibits sexual avoidance, or is concerned about sexual pain. Although such partner issues may lead men to ejaculate rapidly, the PE in this case is actually secondary to the partner's dysfunction.

A number of treatment options are available for PE, including dapoxetine and off-label antidepressants – these drugs affect serotonergic function, thereby delaying ejaculation. Such medications may be used in conjunction with behavioral and cognitive therapies that reduce stimulation to the penis while concomitantly helping the man attend to the genital sensations that signal approaching ejaculation.

Inhibited Ejaculation

The prevalence of delayed ejaculation or IE (the latter being the complete inability to reach ejaculation) is unclear – normative data for defining the duration of 'normal' ejaculatory latency, particularly regarding the right tail of the distribution (i.e., beyond the mean latency to ejaculation), is essentially nonexistent. Furthermore, larger epidemiologic studies have not subdivided men into various types of diminished ejaculatory function. For example, the continuum (and/or overlap) from delayed ejaculation to IE has not been adequately explored.

In general, IE has been reported at fairly low rates in the literature, typically around 3–5%, and thus it has been seen as clinically uncommon. However, based on more recent clinical experiences, some urologists and sex therapists are reporting an increasing incidence of IE (Rowland et al., 2010) leading to newer estimates of anywhere between 3 and 10% (see Lewis et al., 2010). The prevalence of IE appears to be moderately and positively related to age – not surprising in view of the fact that ejaculatory function as a whole tends to diminish with age. However, no large-scale studies have systematically investigated the strength or reliability of this putative relationship.

- In some instances, a pathophysiological or biogenic condition may account for IE since any procedure or disease that disrupts sympathetic or somatic innervation to the genital region has the potential to affect ejaculatory function and orgasm. Thus, spinal cord injury, multiple sclerosis, pelvic-region surgery, severe diabetes, and medications that inhibit α-adrenergic innervation of the ejaculatory system have been associated with IE. Nevertheless, a sizable portion of men with IE exhibit no clear somatic factors that account for the disorder. As the result of their inability to ejaculate, these men also do not experience orgasm.
- Multiple psychosocial explanations have been offered for IE. In some instances, the man may have a strong 'autosexual' orientation that involves an idiosyncratic and vigorous masturbation style – carried out with high frequency – which does not 'match' vaginal stimulation. In other instances, disparity between the reality of sex with the partner and the sexual fantasy (whether or not unconventional) used during masturbation is another potential cause of IE (Perelman and Rowland, 2006). This disparity may involve multiple factors, such as partner attractiveness and body type, sexual orientation, and the specific sex activity performed. In summary, high-frequency idiosyncratic masturbation, combined with fantasy/partner disparity, may well predispose men to experience problems with arousal and ejaculation, resulting in the lack of sufficient psychosexual arousal during coitus to achieve orgasm.

Finally, the evaluative/performance aspect of sex with a partner often creates 'sexual performance anxiety' for the man, a factor that may contribute to IE. With respect to IE, anxiety surrounding the inability to ejaculate may draw the man's attention away from erotic cues that normally serve to enhance arousal. Some such men may be overconscientious about pleasing their partner and lose focus on arousing stimuli (Apfelbaum, 2000).

Few medical interventions exist for men with IE, although a number are currently under investigation. Assuming an etiology that is not biological, effective treatment often involves working with the couple to identify strategies to heighten sexual arousal in the man with IE.

Sexual Pain and Aversion Disorders

Sexual pain disorders in men (male dyspareunia), especially when genital related, are uncommon and most often secondary to a biomedical condition. Estimates for sexual pain in men have been reported at rates of 1–6%, but the reliability of such estimates is not yet established (Lewis et al., 2010). To further complicate matters, DSM-5 has now combined male and female pain disorders into a single category referred to as genitopelvic pain/penetration disorder. Genital lesions and abrasions (sometimes from surgery), prolonged erection after sexual stimulation has subsided (priapism), and pelvic trauma are sometimes sources of discomfort during sex. Insofar as sexual pain is related to an identifiable somatic cause, the prognosis for treatment is very good, assuming the source of the pain can be controlled.

Sexual aversion disorder (SAD) represents more than merely a lack of sexual interest or desire, but is characterized by a strong dislike of sexual intimacy and activity, including the sight, touch, smell, and other sensory aspects of sex as well as intercourse itself. In some instances, different aspects related to sexual intimacy may be differentially affected. Sexual aversion is often manifested through strong negative emotional responses such as disgust and revulsion. As with sexual pain disorders, sexual aversion leads to avoidance of sexual intimacy, but the reasons for this avoidance are quite different for each problem.

Due to the rare occurrence of SAD and the subsequent dearth of research on the problem, reliable prevalence rates are not available. A change in the conceptualization of this disorder is reflected by the fact that it has been removed from DSM-5. Furthermore, because such problems tend to be highly idiosyncratic, treatment is typically more complex and may involve combination therapies that include antianxiety medication and sexual counseling.

Epidemiology of Women's Sexual Dysfunction

In this section, risk factors and prevalence rates for disorders of desire, arousal, and orgasm in women are reviewed and pain disorders are briefly discussed. The prognosis and treatability of each dysfunction are also addressed.

Low Sexual Desire

Prevalence rates for low sexual desire (also called HSDD) in women are significantly higher than for men. The DSM-IV-TR does not define HSDD in a gender-specific way, despite apparent gender differences related to desire. Revisions in the DSM-5 address this issue by combining desire and arousal issues in women (only) into a single category, now referred to as female sexual interest/arousal disorder.

Data from the GSSAB (Laumann et al., 2005) suggest prevalence rates for HSDD somewhere between 26 and 43% in women, in contrast to estimates between 13 and 28% in men. In fact, HSDD probably has the highest prevalence rate among the female sexual dysfunctions, not surprising given the correlation between HSDD and other impairments in sexual functioning. One review of studies found that 64% of women indicated low sexual desire as the cause of any sexual problem (Hayes et al., 2006). However, it is not clear whether and by how much prevalence rates might change, given more recent conceptualizations of sexual desire in women.

Risk factors implicated in HSDD are biological, psychological, and contextual. As with men, low androgen levels may be related to low sexual desire in women. Other sexual dysfunctions may be related to HSDD, as impairment in the other phases of sexual response can impact subsequent desire. Psychological factors include affective disorders and sexual attitudes. Specifically, depressed mood and anxiety are correlated with higher rates of HSDD, as well as the antidepressants used to treat them (especially selective serotonin reuptake inhibitors). A woman's self-reported 'importance of sex' has been associated with desire; women for whom sex is very important are likely to have higher sexual desire. Finally, relational risk factors include relationship problems or poor relationship quality, both often leading to low sexual desire, or a large discrepancy in desire between partners.

Treatment for HSDD may include androgen replacement therapy for women experiencing androgen deficiency. Relationship therapy is often necessary, given that women with desire difficulties are generally in relationships and experiencing distress because their lack of desire is troublesome to them, their partners, or both. The most widely used psychological treatment for HSDD in women is sensate focus therapy, developed by Masters and Johnson (1970), in combination with other approaches. Sensate focus helps the couple learn to enjoy pleasurable sensations while decreasing the focus on goal-directed sex. As part of this approach, women educate their partners on touch and positions which stimulate desire, arousal, and potentially orgasm.

Arousal Disorders

The DSM-IV classification and definition of arousal difficulties in women, referred to as female sexual arousal disorder, has presented several challenges in assessing prevalence. First, cases in which lack of arousal could be attributed to another sexual dysfunction and/or that resulted from the physiological effects substance, medication, or a general medical condition did not qualify as a dysfunction (American Psychiatric Association, 2000). For example, the primary problem of low sexual desire could contribute to a secondary problem of lack of arousal. Although changes in arousal related to menopause are not considered a general medical condition, they may affect genital tissue thickness and elasticity, as well as lubrication. Second, a woman could be experiencing arousal difficulties due to inadequate stimulation, yet few prevalence studies investigated this parameter. Third, the definition dealt only with problems related to physiological response (e.g., adequate lubrication or vaginal swelling) and did not include subjective feelings of arousal.

As previously mentioned, difficulties with interest and arousal in women have now been combined into a single category in DSM-5, addressing many of the challenges related to diagnosis and assessment of prevalence. This new category and definition includes criteria for both subjective and physiological arousal. It also eliminates the need to differentiate whether the primary problem lies with interest or with arousal, as the two were often interdependent.

The range of women reporting problems with lubrication is between 3 and 31%, depending on the time frame used and the age of respondents (Graham, 2010). One review found prevalence rates between 8 and 15% for all women and 21 and 31% for sexually active women (Lewis et al., 2004), with 21% indicating experiencing distress due to these difficulties. It appears that these problems increase with age.

Several pathophysiological or biogenic conditions may account for physiological sexual arousal difficulties in women. Given that estrogen plays a role in maintaining vaginal tissue function and structure, estrogen deficiency has been associated with decreased or delayed vaginal lubrication and vaginal blood flow in postmenopausal or ovariectomized women. Disruptions to the sympathetic nervous system, which plays an important role in genital arousal, have also been implicated – thus conditions such as spinal injury and multiple sclerosis are associated with arousal difficulties. Other medical conditions disrupting arousal function include those that are endocrine related, such as hyperprolactinemia, diabetes mellitus, bilateral ovariectomy, and hysterectomy. Also included are cardiovascular problems such as coronary artery disease and hypertension – more often implicated in male ED but also factors in arousal difficulties in women (Salonia et al., 2006; Meston et al., 2008).

Psychosocial factors related to arousal difficulties in women include marital or relationship dissatisfaction, depression and anxiety (more so 'state' than 'trait' anxiety), interfering thoughts, and negative self-focus including poor body image (Bradford and Meston, 2006). Finally, a history of sexual abuse can contribute to arousal difficulties, as the sensations experienced during sexual activity may actually inhibit arousal.

Few medical interventions are available to women with arousal difficulties; in contrast with their effects on erection in men, vasodilating drugs (such as Viagra) tend to show inconsistent or only small effects (Meston et al., 2006). The EROS

clitoral therapy device (Urometrics, St. Paul, MN, USA), which uses suction to increase vasocongestion, has shown some success in increasing vaginal lubrication, sensation, and orgasm (Meston et al., 2006). Psychological treatments for female arousal difficulties include sensate focus, systematic desensitization, cognitive–behavioral techniques, and communication training.

Orgasmic Disorders

Disorders of orgasm in women, clinically referred to as female orgasmic disorder (FOD), are characterized by delayed or inhibited orgasm. The orgasmic difficulty persists despite a normal sexual excitement phase. Therefore, women who experience problems with orgasm, but who also experience difficulties in desire and/or arousal, would not be diagnosed with FOD. Rather, problems with arousal or desire would be considered the primary diagnosis, with the anorgasmia secondary to these diagnoses. Diagnosis of an orgasmic disorder is also complicated by the variability of the stimulation that can trigger an orgasm and the subjectivity of the experience of orgasm in women, in contrast to men, who generally reach orgasm with ejaculation (Meana, 2012).

Prevalence rates of about 24% have been reported for women with orgasm difficulties, making it the second most commonly reported sexual problem (National Social and Health Life Survey; Laumann et al., 1994). Risk factors for orgasm difficulties are most often psychological. Relationship problems (including conflict and lack of intimacy), anxiety, and obsessive–compulsive tendencies are commonly associated with orgasm difficulties (Meston et al., 2008). Interfering thoughts and worry during sex can also contribute to the problem. Because women often blame themselves for their sexual inadequacies, the anxiety and subsequent difficulties may become self-perpetuating (Meana, 2012). Finally, women generally experience a decrease in frequency of orgasm due to normal, age-related declines in sexual functioning.

Given the presumed psychological nature of FOD, no medical treatments have been found. Relationship and sex therapy may be beneficial in improving relationship quality and addressing inadequate stimulation contributing to the problem (Meana, 2012). One of the most widely used treatments for orgasm difficulties is directed masturbation, which uses cognitive–behavioral techniques to help a woman become familiar with her body and bring herself to orgasm, after which time her partner may become involved in the process (Meston et al., 2004).

Sexual Pain and Aversion Disorders

Although sexual pain is not uncommon among women, it is not well understood and DSM-IV-TR diagnostic categories did not adequately consider their complexity and diversity. In fact, the two classifications of sexual pain disorders in women – dyspareunia and vaginismus – are not reliably differentiated from one another. Dyspareunia is defined as recurrent or consistent genital pain, while vaginismus is defined as recurrent or persistent involuntary spasm of the musculature of the outer third of the vagina (American Psychiatric Association, 2000). The new DSM-5 definition has combined all pain disorders for males and females into one category – genitopelvic pain/penetration disorder – addressing this difficulty in differentiation. However, both the DSM-IV-TR and DSM-5 definitions specify that the pain interferes with sexual activity, although women may well experience genital pain in the absence of sexual activity. Genital pain in women may better be conceptualized as a condition lying on a continuum – from sexual to nonsexual – taking into account various aspects of the pain itself, as well as psychological and contextual factors (Farmer et al., 2008; Meana, 2012).

Given the ambiguity in definitions of sexual pain disorders in women, prevalence rates are difficult to determine. Ranges between 12 and 39% have been suggested for premenopausal women experiencing sexual pain. One study that distinguished the two DSM-IV-TR disorders found a rate of 5% for vaginismus and 28% for dyspareunia, while studies from North America have found a rate of 12% for dyspareunia (Farmer et al., 2008). A study assessing different types of pain symptoms found subgroups and an overall lifetime incidence for sexual pain of 16% (Harlow and Steward, 2003). It is unclear whether prevalence rates vary with age.

Several biomedical risk factors have been correlated with female sexual pain. Frequent or persistent vulvar and/or vaginal infections have been associated with dyspareunia, as has vaginal atrophy (Farmer et al., 2008). Estrogen may be indirectly involved, as women who are taking oral contraceptives, are pregnant, or are in the periovulatory phase of their menstrual cycle have reported increased genital pain. In addition, postmenopausal women with reduced estrogen levels have reported vaginal atrophy, vaginal dryness, and dyspareunia (Farmer et al., 2008). Tissue damage resulting from abdominal or pelvic surgeries or sexual trauma may also lead to female genital pain as do vaginal spasms or contractions (referred to as pelvic floor hypertonicity) and weak pelvic muscles (Reissing et al., 2004). Few psychological risk factors have been identified, but factors such as catastrophizing and hypervigilance may exacerbate an existing condition (Farmer et al., 2008; Meana, 2012). Generally, however, genital pain and affective disorders have been only weakly correlated; it may be that the distress caused by the genital pain contributes to feelings of sadness or anxiety.

The prognosis for pain disorders in women is poor, especially if left untreated, as the problem may worsen due to tissue damage and irritation from continued sexual activity. Fear of the pain itself may intensify the problem (Meana, 2012). Although well-designed treatment studies are sparse, for one subtype of dyspareunia, vulvar vestibulitis syndrome, surgical removal of painful vestibular tissue appears effective (Farmer et al., 2008). Electromyography biofeedback and cognitive interventions may help with the pain catastrophization and hypervigilance, such that women redirect their thoughts away from the sensations and concern about pain. Given the complex and emotional nature of pain, a multidisciplinary approach relying on both medical and psychological interventions may yield the best outcomes (Farmer et al., 2008).

SAD is characterized by a persistent or frequent strong aversion to or avoidance of sexual activity with a partner.

Although categorized under the disorders related to sexual desire in DSM-IV-TR, there have been questions as to whether it truly falls within this category, as women presenting with SAD rarely express concern over their lack of desire (Meana, 2012). Rather, they are more likely to express disgust, fear, or anxiety over engaging in sexual activity. Thus, it is unclear as to whether SAD represents an actual sexual dysfunction or whether it might be better conceptualized as a specific phobia.

Prevalence rates for SAD remain largely unknown and, as with the analogous condition in men, DSM-V has now removed this classification. Nevertheless, clinical observations suggest that the prevalence for SAD is higher for women than for men. In addition, the prognosis for SAD is poor. If conceptualized as a phobia, treatment typically focuses on reduction of fear and anxiety, thereby reducing interference with sexual activity but not necessarily resulting in more enjoyable sex (Meana, 2012).

Summary and Conclusions

The past two decades have witnessed significant strides in understanding the prevalence, etiology, and treatment of men's and women's sexual dysfunctions. Driven partly by pharmaceutical interests and resources, men's dysfunctions such as ED, low sexual desire, and PE have been the object of substantial research, the result being greater awareness of these problems, clearer detail of the etiologies, and improved treatment methods using both biomedical and counseling methods. Other problems such as IE and sexual pain and aversion disorders in men are only now beginning to receive attention.

Women's dysfunctions too have attracted significant attention over the past decade, with the development of new models of sexuality, greater clarity regarding the types of problems, and some moderately effective strategies for overcoming problems. Given that women tend to focus more on relationship issues than men when assessing satisfaction within their sexual relationship, biomedical interventions for sexual desire and arousal tend overall to be less effective, and relationship counseling figures more prominently in successful treatment. On the other hand, anorgasmic women are sometimes served well through self-exploration that includes educational and self-stimulatory components.

Finally, it is necessary to realize that the understanding and classification of sexual problems is far from an exact science. As sexual problems are better understood; as demographics change; as norms, attitudes, and definitions vary over time; and as better sampling methods are developed, prevalence rates are likely to be affected.

See also: Marital Quality and Health; Methodological Issues in the Study of Sexuality; Perversions, Sexual (Paraphilias); Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sex Therapy, Clinical Psychology of; Sexuality Over the Life Course; Sexuality, Theories of; Sexuality: Cultural Aspects; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. Author, Washington, DC. Text Revision.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. Author, Washington, DC.

Apfelbaum, Bernard, 2000. Retarded ejaculation: a much misunderstood syndrome. In: Lieblum, Sandra, Rosen, Raymond (Eds.), Principles and Practice of Sex Therapy, third ed. Guilford Press, New York, NY, pp. 205–241.

Basson, Rosemary, Leiblum, Sandra, Brotto, Lori, Derogatis, Leonard, Fourcroy, Jean, Fugl-Meyer, Kersten, et al., 2003. Definitions of women's sexual dysfunction reconsidered: advocating expansion and revision. Journal of Psychosomatic Obstetric and Gynecology 24, 221–229.

Bradford, Andrea, Meston, Cindy, 2006. The impact of anxiety on sexual arousal in women. Behaviour Research and Therapy 44, 1067–1077.

Byrne, Donn, Schulte, Lise, 1990. Personality dispositions as mediators of sexual responses. Annual Review of Sex Research 1, 93–117.

Corona, Giovanni, Petrone, Luisa, Mannucci, Edoardo, Jannini, Emmanuele, Mansani, Riccardo, Magini, Antonia, Giommi, Roberto, Forti, Gianni, Maggi, Mario, 2004. Psycho-biological correlates of rapid ejaculation in patients attending an andrologic unit for sexual dysfunctions. European Urology 46 (5), 615–622.

Farmer, Melissa, Kukkonen, Tuuli, Binik, Yitzchak, 2008. Female genital pain and its treatment. In: Rowland, David, Incrocci, Luca (Eds.), Handbook of Sexual and Gender Identity Disorders. Wiley & Sons Inc., Hoboken, NJ, pp. 220–250.

Graham, Cynthia, 2010. The DSM diagnostic criteria for female sexual arousal disorder. Archives of Sexual Behavior 39, 240–255.

Harlow, Bernard, Steward, Elizabeth, 2003. A population-based assessment of chronic unexplained vulvar pain: have we underestimated the prevalence of vulvodynia? Journal of the American Medical Women's Association 58, 82–88.

Hayes, Richard, Bennett, Catherine, Fairley, Cristopher, Dennerstein, Lorraine, 2006. What can prevalence studies tell us about female sexual difficulty and dysfunction? Journal of Sexual Medicine 3, 589–595.

Kaplan, Helen Singer, 1974. The New Sex Therapy. Brunner/Mazel, New York, NY.

Laumann, Edward, Nicolosi, Alfredo, Glasser, Dale, Paik, Anthony, Gingell, Clive, Moreira, Edson, Wang, Tianfu, 2005. Sexual problems among men and women aged 40–80 yrs: prevalence and correlates identified in the global study of sexual attitudes and behaviours. International Journal of Impotence Research 17, 39–57.

Leiblum, Sandra, Brown, Candace, Wan, Jim, Rawlinson, Leslie, 2005. Persistent sexual arousal syndrome: a descriptive study. Journal of Sexual Medicine 2, 331–337.

Lewis, Ronald, Fugl-Meyer, Kersten, Bosch, Rund, Fugl-Meyer, Axel, Laumann, Edward, Lizza, Eli, Martin-Morales, Antonio, 2004. Definitions, classification, and epidemiology of sexual dysfunction. In: Lue, Tom, Basson, Rosemary, Rosen, Raymond, Giuliano, Francois, Khoury, Saad, Montorsi, Francesco (Eds.), Sexual Medicine: Sexual Dysfunction in Men and Women. Health Publications, Paris, pp. 39–72.

Lewis, Ronald, Fugl-Meyer, Kersten, Corona, Giovanni, Hayes, Richard, Laumann, Edward, Moreira Jr., Edson, Rellini, Alessandra, Segraves, Taylor, 2010. Definitions/epidemiology/risk factors for sexual dysfunction. Journal of Sexual Medicine 7, 1598–1607.

Lewis, Ronald, Yuan, Jiuhong, Wang, Run, 2008. Male sexual arousal disorder. In: Rowland, David, Incrocci, Luca (Eds.), Handbook of Sexual and Gender Identity Disorders. Wiley & Sons Inc., Hoboken, NJ, pp. 32–67.

Masters, William H., Johnson, Virginia E., 1966. Human Sexual Response. Little, Brown & Co., Oxford, England.

Masters, William H., Johnson, Virginia E., 1970. Human Sexual Inadequacy. Little, Brown & Co., Boston, MA.

McMahon, Chris, 2002. Long term results of treatment of premature ejaculation with selective serotonin re-uptake inhibitors. International Journal of Impotence Research 14 (Suppl. 3), S19.

McMahon, Chris, Althof, Stanley, Waldinger, Marcel, et al., 2008. An evidence-based definition of lifelong premature ejaculation: report of the International Society for Sexual Medicine (ISSM) ad hoc committee for the definition of premature ejaculation. Journal of Sexual Medicine 5, 1590–1606.

Meana, Marta, 2012. Sexual Dysfunction in Women. Hogrefe Publishing, Cambridge, MA.

Meston, Cindy, Hull, Elaine, Levin, Robert, Sipski, Marcalee, 2004. Women's orgasm. In: Lue, Tom, Basson, Rosemary, Rosen, Raymond, Giuliano, Francois, Khoury, Saad, Montorsi, Francesco (Eds.), Sexual Medicine: Sexual Dysfunction in Men and Women. Health Publications, Paris, pp. 783–849.

Meston, Cindy, Seal, Brooke, Hamilton, Lisa, 2008. Problems with arousal and orgasm in women. In: Rowland, David, Incrocci, Luca (Eds.), Handbook of Sexual and Gender Identity Disorders. Wiley & Sons Inc., Hoboken, NJ, pp. 188–219.

Patrick, Donald, Rowland, David, Rothman, Margaret, 2007. Interrelationships among measures of premature ejaculation: the central role of perceived control. Journal of Sexual Medicine 4 (3), 780–788.

Perelman, Michael, Rowland, David, 2006. Retarded ejaculation. World Journal of Urology 24, 645–652.

Reissing, Elke, Binik, Yitzchak, Khalife, Samir, Cohen, Deborah, Amsel, Rhonda, 2004. Vaginal spasm, pain, and behavior: an empirical investigation of the diagnosis of vaginismus. Archives of Sexual Behavior 33, 5–17.

Rowland, David, 2011. Psychological impact of premature ejaculation and barriers to its recognition and treatment. Current Medical Research and Opinion 27 (8), 1509–1518.

Rowland, David, 2012. Sexual Dysfunction in Men. Hogrefe Publishing, Cambridge, MA.

Rowland, David, McMahon, Chris, Abdo, Carmita, Chen, Juza, Jannini, Emmanuele, Waldinger, Marcel, Ahn, Tai, 2010. Disorders of orgasm and ejaculation in men. Journal of Sexual Medicine 7, 1668–1686.

Salonia, Andrea, Briganti, Alberto, Rigatti, Patrizio, Monstori, Francesco, 2006. Medical conditions associated with female sexual dysfunctions. In: Goldstein, Irwin, Meston, Cindy, Davis, Susan, Traish, Abdulmaged (Eds.), Women's Sexual Function and Dysfunction: Study, Diagnosis, and Treatment. Taylor & Francis, New York, pp. 263–275.

Schover, Leslie, Friedman, Jerry, Weiler, Stephen, Heiman, Julia, LoPiccolo, Joseph, 1982. Multiaxial problem-oriented system for sexual dysfunctions: an alternative to DSM-III. Archives of General Psychiatry 39, 614–619.

Sternberg, Robert, Barnes, Michael, 1988. The Psychology of Love. Yale University Press, New Haven, CT.

van Lankveld, Jacques, 2008. Problems with sexual interest and desire in women. In: Rowland, David, Incrocci, Luca (Eds.), Handbook of Sexual and Gender Identity Disorders. Wiley & Sons Inc., Hoboken, NJ, pp. 154–187.

World Health Organization, 2013. http://www.who.int/classifications/icd/en/ (accessed 20.05.13.).

Relevant Websites

www.aasect.org – American Association for Sex Educators, Counselors and Therapists (AASECT).
www.acog.org – American College of Obstetrics and Gynecology (ACOG).
www.auanet.org – American Urological Association (AUA).
www.andrologysociety.com – American Society of Andrology.
www.issm.info – International Society for Sexual Medicine (ISSM).
www.isswsh.org – International Society for the Study of Women's Sexual Health (ISSWSH).
www.sexscience.org – Society for the Scientific Study of Sexuality (SSSS).
www.sstarnet.org – Society for Sex Therapy and Research (SSTAR).

Sexual Harassment: United States and Beyond

Anna-Maria Marshall, University of Illinois, Urbana-Champaign, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by V. Schultz, volume 21, pp. 13981–13987, © 2001, Elsevier Ltd.

Abstract

Beginning in the 1970s, women filed lawsuits in U.S. federal courts arguing that unwanted sexual attention at workplace constituted a form of sex discrimination. After decades of judicial policy-making, the legal definition of sexual harassment remains ambiguous, and the process for offering relief to targets of harassment reflects employers' interests in addition to employee rights. At the same time, sociologists and psychologists have explored the causes of employee's experiences with sexual conduct in the workplace and its impact on the targets. Social scientists have also analyzed organizational responses to sexual harassment, research that raises questions about how legal rights ought best be protected. This article examines sexual harassment law in the United States, and then turns to the social science literature that documents the frequency of harassment, employees' understanding of their rights and their willingness to file claims, and concludes by considering evidence of 'sexual harassment in the workplace' (rather than 'at workplace').

Legal innovation, particularly in the area of civil rights and civil liberties, emerges from the clash of complex social forces. Notably, social movements and advocacy networks make novel claims for the recognition of new rights on behalf of the powerless and the marginalized. These new rights might protect groups from discrimination or enhance their power in more political arenas. However, activists do not exercise complete control over the process or pace of legal innovation, particularly in the judicial arena. Rather, their opponents, who normally have greater resources, resist their claims. The end result is often a new set of rights that embodies the compromises inherent in this adversarial process.

Women and their lawyers followed this path of legal innovation when creating the right to a workplace free from sexual harassment, a development that began in the United States. In the 1970s, an interesting combination of feminists, lawyers, and individual women filed employment discrimination lawsuits arguing that unwanted sexual attention in the workplace constituted a discriminatory working condition in violation of federal employment discrimination statutes (Marshall, 1998). This new interpretation of these statutes generated resistance among employers, especially among their lawyers and human resource professionals. The resulting judicial precedents have generated definitions of sexual harassment that remain ambiguous, even after years of litigation. In addition, the remedies – and the conditions that attach to those remedies – reflect employers' interests as well as employee rights.

The development of the legal claim for sexual harassment in the US coincided with the recognition of many disadvantages women faced in the workplace. Many trace the origins of the concept of sexual harassment to Lynn Farley at Cornell University, who offered a field seminar on "Women and Work" (Elman, 1996; Cahill, 2001; Weeks et al., 1986). Relying on consciousness-raising as a pedagogical technique, Farley and her students realized that everyone in the class had been forced to quit or been fired from a job because of the behavior of male colleagues (Farley, 1978). The group described this experience as 'sexual harassment,' a concept that Farley developed in her book *Sexual Shakedown* (1978). That class gave rise to a small women's rights organization at Cornell, Working Women United, that conducted some of the first surveys of working women and that publicized the problem, getting national attention (Elman, 1996; Cahill, 2001; Weeks et al., 1986). In the late 1970s, larger, multiissue women's organizations in the US, such as NOW and 9 to 5, included sexual harassment on their agendas. They designed public education campaigns to raise awareness among working women and lobbied the U.S. Congress and administrative agencies to regulate workplaces. Later on, these organizations filed amicus briefs in prominent court cases in the US and occasionally supplied lawyers when interesting cases reached appellate courts (Marshall, 2001).

As the legal claim for sexual harassment has developed, sociologists and psychologists have generated a rich research tradition exploring women's experiences with sexual conduct in the workplace. They have identified several different explanations for such conduct and have documented the range of responses to such behavior. In addition, social scientists have analyzed organizational responses to sexual harassment, research that raises questions about how legal rights ought best be protected.

This article describes the origins of the legal claim for sexual harassment in the United States, emphasizing U.S. judicial precedents that define sexual harassment and that specify processes for redressing it. The article also outlines the social scientific accounts of sexual harassment in the workplace, demonstrating the way that law both corresponds to and deviates from employees' experiences. And finally, a brief discussion of the diffusion of sexual harassment laws to other jurisdictions suggests alternative conceptualizations of the harms that may offer new regulatory possibilities.

Defining Sexual Harassment in the US: Legal Rules and Workers' Experiences

Legislative and administrative policy-makers in the US were slow to take an interest in regulating sexual harassment, and not surprisingly, the claim found its way to the courts. Title VII of the Civil Rights Act of 1964 prohibited sex discrimination in

the terms and conditions of employment but does not specifically identify sexual harassment as a form of sex discrimination. The first cases articulated claims for *quid pro quo* harassment, where women's submission to sexual favors was a requirement of the job. At first, federal courts dismissed these cases, arguing that sexual harassment was a personal problem between men and women in the workplace that emerged from the 'natural attraction' of opposite sexes. Thus, this behavior was not a social problem that the judicial system ought to resolve (MacKinnon, 1979).

Federal courts in the US first recognized sexual harassment as a form of sex discrimination in a series of cases in 1976 and 1977 (*Williams v. Saxbe* 1976; *Barnes v. Train* 1977; *Tomkins v. PSE&G* 1977). The courts reasoned that sexual demands placed on working women were a working condition imposed on women but not on men. As such, they constituted sex discrimination. But the courts limited the claim to situations where supervisors explicitly made a promotion, salary, or the job itself contingent on sexual favors.

Later in the 1970s, women began filing lawsuits with more expansive claims, describing working environments rife with sexual jokes and banter, as well as unwanted physical contact that, in some extreme cases, constituted sexual assault. Unlike women making claims for *quid pro quo* harassment, however, these women did not lose their jobs or suffer any tangible job detriment like the loss of a raise or a promotion. Nevertheless, they argued that the working environment imposed a substantial burden on their ability to perform their jobs. Although these claims initially failed in court, the Equal Employment Opportunity Commission promulgated regulations in 1980 that defined sexual harassment as including behaviors such as jokes, threats, and demands for sexual attention when "such conduct [that] has the purpose or effect of unreasonably interfering with an individual's work performance or creating an intimidating, hostile, or offensive working environment" (29 C.F.R. Sec. 1604.11).

The U.S. Supreme Court endorsed these regulations in *Meritor Savings Bank v. Vinson* (1986), in which it defined a hostile working environment as a workplace where employees faced "unwelcome sexual advances, requests for favors, and other verbal or physical conduct of a sexual nature" (*Meritor Savings Bank v. Vinson* 1986, 65). The Court also articulated the standards for establishing a successful claim: "For sexual harassment to be actionable, it must be sufficiently severe or pervasive 'to alter the conditions of [the plaintiff's] employment and create an abusive working environment'" (*Meritor Savings Bank v. Vinson* 1986, 67). Later, the Court clarified that the behaviors had to interfere with the employee's performance of her job duties, although employees did not necessarily have to sustain severe psychological distress (*Harris v. Forklift Systems, Inc.* 1993). Thus, the legal definition of a hostile working environment had both objective and subjective components. The conduct in question had to meet a standard that was clearly objectionable, but the target of the harassment also had to perceive the conduct as being harmful.

The problems associated with linking sexual harassment to sexual conduct became apparent when courts started hearing cases involving same-sex sexual harassment. Given that the premise of the original holdings was that a male supervisor's harassment of a female employee constituted sex discrimination, same-sex harassment fit less comfortably within this legal framework. However, the Supreme Court resolved this question by holding that same-sex harassment violates Title VII as long as the plaintiff can prove the statutory requirements: that the harassment was severe and pervasive and that it was discriminatory (*Oncale v. Sundowner Offshore Services, Inc.* 1998).

The claim for sexual harassment was a novel judicial interpretation of a statute, illustrating the capacity of the courts to address social problems when those problems do not reach the agendas of other policy-makers. However, critics have noted the limitations of these court-crafted policies. Civil libertarians have observed that to avoid liability, employers might adopt restrictive policies that could unnecessarily ban all sexual expression in the workplace (Volokh, 1991; Schultz, 1998). On the other hand, feminists have argued that by tying the claim of sexual harassment to sexual advances, courts have misconceptualized the problem, which is actually more general male domination in the workplace (Abrams, 1997; Franke, 1996). And Vicky Schultz (1998) has argued that tying sexual harassment to sexuality ignored a wide array of other forms of harassment that had nothing to do with sexual advances.

Over the years, sociologists and psychologists have conducted extensive research on the nature and prevalence of sexual conduct in the workplace, offering a perspective on whether the legal definition corresponds to employees' experiences at work. According to most studies, women often face unwanted sexual attention in the workplace. Notably, women remain the primary targets of such behaviors (Uggen and Blackstone, 2004). Experts estimate that between 35% and 50% of all working women have had at least one such experience (Gutek, 1985; Fitzgerald and Ormerod, 1993; Fitzgerald et al., 1988). Still, more recent research has begun to explore sexual harassment of men (Settles et al., 2012; Kearney and Rochlen, 2012; Uggen and Blackstone, 2004) and sexual harassment between employees of the same sex (Bennett, 1996).

Sexual conduct in the workplace comprises a range of different behaviors that vary in their intrusiveness. To capture this range, researchers follow a particular protocol: they first ask respondents whether they have experienced specific sexual behaviors, ranging from crude jokes and remarks to sexual assault. They then ask respondents how often those behaviors occurred and what effect they had. Lastly, respondents are asked whether the behavior constituted sexual harassment (Fitzgerald et al., 1988; Fitzgerald and Ormerod, 1991). This research design allows researchers to examine not just the prevalence of sexual conduct in the workplace, but also the influence of legal rules on employees' perceptions of their own experiences (Marshall, 2003).

By far, the most common form of sexual behavior in the workplace was what Fitzgerald and her colleagues described as 'gender harassment' – conduct that conveys insulting or hostile views of women but which are not necessarily "aimed at sexual cooperation" (Gelfand et al., 1995). This conduct would not meet the objective standards set by the courts for sexual harassment because it is insufficiently 'severe.' The second category, 'unwanted sexual attention,' consists of behavior that is more explicitly sexual in nature, including uninvited conversations about a target's sex life or pressure for dates. If evidence showed that these behaviors were in some way welcomed by targets, they might be considered innocuous

flirtation; when unwelcome, such conduct is more intrusive and threatening to the target. The third category consists of 'sexual coercion' – demanding sexual favors backed by a promise of benefits or a threat of negative consequences. This category would be considered *quid pro quo* harassment in the courts (Gelfand et al., 1995).

Researchers have several competing theories to explain the existence of sexual harassment in the workplace. Some suggest, for example, that sexual harassment is "the product of culturally legitimated power and status differences between men and women" (Welsh, 1999; MacKinnon, 1979). Sex-role spillover theories emphasize the way that potent stereotypes and gender socialization processes extend male dominance over women into the workplace, making women vulnerable to sexual harassment (Tangri et al., 1982; Gutek, 1985). Women are most likely to be harassed in workplaces where the proportion of workers is skewed – either more females or more males – because in such locations, women's 'femaleness' is rendered more conspicuous (Stockdale, 1996; Welsh, 1999). In male-dominated work settings, masculine qualities of toughness and competition are valued more highly, and women's presence is threatening to that workplace culture (Vogt et al., 2007). On the other hand, in female-dominated workplaces, sex roles characterized by nurturing and connection drown out work-related identities (Gutek and Morasch, 1982). Thus, greater contact between men and women at work, particularly when men numerically dominate the workplace, is responsible for more sexual harassment (Gutek et al., 1990; Gruber, 1998).

Others argue that sexual harassment emerges from power differentials between men and women. In this view, sexual harassment is one byproduct of male economic power over women – a technique that men use to maintain that dominance (MacKinnon, 1979; Uggen and Blackstone, 2004). Extensive evidence of female supervisors being harassed by subordinates and peer harassment led some to question whether power was truly a factor in sexual harassment. Having less power in their workplaces, women are more likely to be harassed by both their coworkers (Gutek, 1985) and their subordinates than by their supervisors (McKinney, 1990, 1992; Rospenda et al., 1998). Researchers, however, argued that these forms of contrapower harassment simply reflect the extensive informal power that men have in the workplace (Cleveland and Kerst, 1993; Grauerholz, 1996). "[M]ale peers lack an institutional basis for dominating their female colleagues. For this group, sexual harassment may be one of the few means available for realizing this end" (Lafontaine and Tredeau, 1986). Similarly, characterizing sexual harassment as "an equalizer," McLaughlin et al. (2012) found that female supervisors were most likely to perceive their experiences as sexual harassment and to report such behaviors.

Women have different reactions to their experiences with sexual conduct in the workplace. For many women, such incidents can be devastating to their careers and well-being. Women report that as a result of sexual harassment, they suffer decreased morale on the job, increased absenteeism, and other negative changes in both the quality and the quantity of their work (Fitzgerald and Ormerod, 1993; Coles, 1986). Women also lose their jobs in the aftermath of these experiences, either by resignation, transfer, or even discharge (Coles, 1986; Fitzgerald and Ormerod, 1993). In addition to the work-related problems, targets can also suffer psychological damage because of sexual harassment. In clinical studies, victims report a series of symptoms – fear, anxiety, anger, and depression, for example – similar to those suffered by victims of sexual assault (Koss, 1990; Houle et al., 2011).

On the other hand, critics of this line of research argue that not every working woman reacts in a negative way to sexual expression in the workplace. For example, Christine Williams and her colleagues observed "Workers themselves often conceive of sexual behaviors at work along a continuum, ranging from pleasurable, to tolerable, to harassing" (Williams et al., 1999; Giuffre and Williams, 1994; Dellinger and Williams, 2002; Williams, 1997). For some workers, for example, sexual behavior and joking help to relieve stress (Quinn, 2002). Thus, contextual factors are likely to shape the impact of sexual conduct in the workplace (Williams, 1997; Williams et al., 1999).

Given the wide range of behaviors and the different reactions to those behaviors, it is not surprising that researchers have identified wide variation in targets' willingness to label conduct sexual harassment (Welsh, 1999; Fitzgerald et al., 1997). In most studies, women are more likely than men to label particular forms of conduct sexual harassment (McLaughlin et al., 2012; Uggen and Blackstone, 2004), although recent evidence suggests that men with more egalitarian gender beliefs are more likely than other men to describe behaviors as sexually harassing (Uggen and Blackstone, 2004). These assessments also depend on surrounding situational factors, such as the severity and frequency of the conduct and the status of the harasser (Stockdale et al., 1995; Giuffre and Williams, 1994; Marshall, 2003).

Organizational characteristics, formal policies, and legal rules also shape people's perception of sexual harassment (Tinkler, 2008; Marshall, 2003). For example, Tinkler analyzed people's perceptions of sexual harassment as a function of their familiarity with relevant policies as well as their own personal beliefs about the 'problem of sexual harassment.' She found that men who believed that people were too sensitive about sexual harassment and who had attended a training session were more likely than others to apply the label of sexual harassment to even less intrusive behaviors (Tinkler, 2008).

Sexual behavior at work has not disappeared in the years since sexual harassment was first recognized as a form of employment discrimination. Rather, sexual conduct and conversation remain fixtures of the American workplace. However, research has shown that employees do not automatically consider every such experience as sexual harassment. Rather, they reserve that label for more intrusive kinds of conduct that have a more harmful impact, in effect applying the objective and subjective tests suggested by the legal standards (Marshall, 2003). The next question is how employees choose to respond and how the legal environment shapes their options.

Sexual Harassment in Organizational Context: Grievance Processes in the US

The right to be free from sexual harassment is part of a package of civil rights that purportedly protect and advance equality in the workplace. In the United States, however, the target of the

harassment has the responsibility for enforcing those rights (Farhang, 2010). Courts have delegated authority to employers to design policies and grievance procedures either to prevent sexual harassment or to remedy it once it has occurred. As a result, sexual harassment policies have become relatively formal processes that mimic legal proceedings; they may do more to protect employers than to advance employee rights (Edelman et al., 1999; Marshall, 2005). This section examines sexual harassment policies developed in response to judicial precedent and describes research about how targets of sexual harassment navigate these processes.

Recent judicial activity on sexual harassment has focused on the grievance procedures surrounding employee complaints about harassing behaviors. The employer's grievance process was at issue in *Meritor Savings Bank v. Vinson*, in which the Court suggested that an appropriately designed policy could protect an employer from sexual harassment claims. Although Meritor had such a grievance procedure, it required employees to lodge complaints with their supervisors, which was unacceptable if the supervisor was the harasser. The Court noted that this process was fatally flawed, but that one might be designed that could protect the employer from liability for harassing conduct.

In later companion cases, *Faragher v. Boca Raton* (1998) and *Burlington Industries v. Ellerth* (1998), the Supreme Court provided further guidance about the kinds of grievance procedures that would protect employers from liability. The Court held that employers would be held responsible for their supervisors' harassing behaviors directed at employees resulting in negative consequences. However, in cases where the harasser took no 'tangible employment action' against his victim, the Supreme Court held that employers could defend themselves from liability when they had a proper sexual harassment policy in place and the harassed employee failed to use it.

Although the Court was vague about the details of the policies, human resources professionals and employment lawyers responded to the decision by developing policies that would meet employer's obligations under the new rules (Bisom-Rapp, 1998; Grossman, 2003; Edelman et al., 1999). Writing in professional journals, employment specialists narrowed in on a set of features for these policies that would inoculate employers from liability. For example, the policies had to define prohibited conduct, provide a list of specific examples, and be widely and regularly distributed to all employees and supervisors. The grievance process had to provide multiple avenues for pursuing grievances; it could never require a harassed employee to confront the harasser. Each complaint had to be investigated, even if the complainant was reluctant to pursue it. And finally, employers had to respond promptly to complaints and take corrective action to redress the problem (Grossman, 2003). These features largely addressed the Supreme Court's concern about the appropriate design for a sexual harassment grievance procedure.

In addition, the Court held that employers must show that they had 'exercised reasonable care' to prevent and redress sexual harassment. Human resources professionals urged employers to take the initiative in monitoring harassment. To demonstrate their proactive concern, employers increasingly adopted training programs to educate employees about the limits of acceptable workplace behavior and to teach supervisors about how to handle complaints brought to their attention. The evidence for the effectiveness of these training programs is thin (Tinkler et al., 2007; Tinkler, 2008, 2012). Still, courts have recognized that the adoption of these programs is evidence that employers are 'exercising reasonable care' in preventing sexual harassment (Edelman et al., 1999).

The Supreme Court's decisions in *Faragher* and *Ellerth* did not simply require employers to develop grievance procedures for sexual harassment. The Court also imposed obligations on employees to use the employer's grievance procedure before filing a complaint, or show a good reason why they failed to do so. Thus, if harassed employees "unreasonably failed to take advantage of any preventive or corrective opportunities provided by the employer," their complaints would be dismissed (*Faragher v. Boca Raton* 1998, 807; *Burlington Industries v. Ellerth* 1998, 765). The requirement of 'reasonableness' might have allowed some plaintiffs to argue that lodging complaints was not feasible given the surrounding context of the workplace – fear of job loss, peer pressure, or gender isolation, for example. Yet a study conducted after the decisions were released showed that each time a plaintiff failed to invoke a grievance procedure, the trial court dismissed the complaint (Sherwyn et al., 2001). Thus, the Supreme Court requires that to preserve their rights, targets of sexual harassment must do the one thing that they are least likely to do – use a grievance procedure to resolve their problems.

Through the years and across many different studies, respondents have reported that their first response is to try to ignore the behavior or to avoid their harassers (Merit Systems Protection Board, 1995; Blackstone et al., 2009; Marshall, 2005). When they do choose to respond, targets of sexual harassment have a wide range of options short of filing formal grievances. For example, they might engage in self-help by admonishing conduct they disapprove of, or they may seek support from their colleagues at work or friends and family outside the workplace (Blackstone et al., 2009; McDonald, 2012; Marshall, 2005). Yet these responses would not be enough to pursue a claim for sexual harassment in the courts. According to the standards the Court articulated in *Faragher* and *Ellerth*, targets must invoke their employers' grievance procedures to state a claim for sexual harassment.

Employees only very rarely use formal procedures to complain about their experiences with unwanted sexual attention (Firestone and Harris, 2003; Gutek, 1985; Fitzgerald et al., 1995; Welsh, 1999). Surveyed about their responses to sexual harassment, employees themselves described a variety of reasons for their reluctance to report their experiences. In general, they are ambivalent about the grievance procedures and the personnel who administer them (Marshall, 2005; Welsh, 1999; Fitzgerald et al., 1995). Employees reported three fears: losing their job, being blamed for the incident, or not being believed. Although many agree that sexual harassment policies have reduced the incidence of many forms of unwanted sexual attention (Gruber, 1998; Tinkler, 2008), survey research strongly suggests that employees have serious reservations about the very policy responses that are designed to protect their rights (Welsh, 1999; Marshall, 2005; Tinkler, 2008, 2012; Edelman et al., 1999).

These reservations reflect what employees know about their own workplaces: employment policies and procedures to

redress sexual harassment are firmly embedded in organizational practices and cultures. Thus, they may not be equipped to dismantle the structural problems giving rise to sexual harassment. Designed and implemented by managers, grievance procedures come to reflect the organization's interests, overriding their ostensible goal of protecting employee rights (Edelman et al., 1999; Marshall, 2005). Marshall (2005) demonstrated how, in one workplace, managers actively discouraged women from filing complaints, offering interpretations of underlying incidents that trivialized the harm or erecting barriers to complaints not contemplated by the policies themselves. Based on their own experiences and the experiences of others, the intended beneficiaries of these policies learned that the complaint process would be inhospitable to their claims and sought other, less formal means to address their problems.

Moreover, recent research suggests that sexual harassment policies may be having unintended consequences that could lead to backlash because of their complex impact on gender beliefs. For example, Tinkler and her colleagues showed that a sexual harassment policy triggered male-advantage stereotypes and led participants to think less of both their male and female colleagues (Tinkler et al., 2007). Participants exposed to such a policy found both men and women less competent, less considerate, and lower in status (Tinkler, 2008). Even the sex of the trainers may have an impact on gender beliefs. While female trainers in an experiment prompted men to express more egalitarian views of male and female competence, female trainers also triggered unconscious gender stereotypes among all participants (Tinkler et al., 2014). This line of research raises the very real concern that sexual harassment policies could trigger resistance (Tinkler et al., 2014; Kalev et al., 2006). By activating gender stereotypes that characterize women as fearful of sex and in need of protection, policies may have the unintended consequence of polarizing men and women in the workplace, which may also lead to backlash against all such policies (Tinkler et al., 2014).

The legal rules governing sexual harassment have shaped the U.S. workplace in meaningful ways. Most obviously, employers have complied with their legal obligations and adopted policies that define sexual harassment and provide targets with means for redressing their complaints. Yet by itself, law cannot dictate how these policies actually work. The legal definitions of harassing behavior remain ambiguous, depending on both objective standards of conduct and the target's subjective experience. Policies will not become relevant, therefore, unless a target perceives an experience as sexual harassment. And although the policies play an important role in employees' ability to vindicate their rights, the procedures are administered by managers who are chiefly concerned with protecting the employer's interests.

Sexual Harassment in Comparative Context

Cross-cultural research has demonstrated that sexual harassment is a common problem for working women around the world. Indeed, there are meaningful differences in the targets of sexual conduct at work in other countries, as well as in societal responses to such conduct. In addition, sexual harassment offers interesting insights into the process of policy diffusion. Other jurisdictions have relied on the United States as a model and adopted a discrimination framework as a means to address the problem, while other countries have forged their own solutions, drawing on their own cultural understandings of the causes and harms of sexual harassment.

For example, Uggen and Shinohara (Uggen and Shinohara, 2009) have analyzed the similarities and differences between women experiencing sexual harassment in the US and Japan. Women in both countries were most likely to meet with unwanted sexual attention in manufacturing industries and least likely to experience it in government work. Yet there were important differences. In Japan, the youngest workers reported higher lifetime rates of harassment, while slightly older women in the US reported higher rates of incidents. Uggen and Shinohara argue that this age effect is attributable to the timing of the introduction of the concept 'sexual harassment' in the popular culture. In addition, marriage and family rendered Japanese women less susceptible to harassment. Those effects disappeared, however, when the researchers factored in characteristics of the job. Notably, in Japan, higher income women were more likely to report sexual harassment. Uggen and Shinohara speculate that women with higher incomes work in male-dominated fields, making them more susceptible to this kind of attention.

Efforts to regulate sexual harassment are "refracted through the structure and culture of individual nations" (Uggen and Shinohara, 2009: 225). Kathrin Zippel (2006) has analyzed the processes of policy diffusion from the US to the European Union and its member states. Transnational advocacy networks were responsible for disseminating information about sexual harassment and for crafting proposals for possible regulatory frameworks. They persuaded the European Union to adopt 'soft law' measures – resolutions and recommendations to member states to develop their own policies against sexual harassment. Against this backdrop of 'Europeanization,' feminist officials in German policy-making institutions – 'femocrats' – placed sexual harassment on the political agenda. Feminist opposition to sexual harassment corresponded to trade union narratives about the exploitation of workers. As a result, the 1994 Employee Protection Act prohibited sexual harassment but did so as part of a larger package of employee rights. Similarly, feminist advocacy, policy-makers in Austria adopted the discrimination framework prevalent in the United States. However, the discrimination frame was not successful on the shop floor, where Austrian factory managers, for example, resisted implementing what they viewed as foreign forms of regulation (Cahill, 2001).

As in Germany and Austria, France has a long tradition of union-organized workplaces, and French activists preferred to portray sexual harassment as a violation of the respect and dignity owed to workers (Bernstein, 1994). Yet the French did not regulate sexual harassment as a work-related problem. Rather, sexual harassment is a crime in France (Saguy, 2003). These differences in conceptualizing the harm of sexual harassment offer opportunities to imagine regulatory frameworks that transcend the U.S. reliance on discrimination theories (Bernstein, 1994, 1997; Franke, 1996; Abrams, 1997; Schultz, 1998).

Conclusion

The legal innovation of recognizing sexual harassment as a form of employment discrimination offered women an opportunity to redress an obstacle to equality at work. Yet through the years, that opportunity has been filtered through the structure of the workplace and employers' efforts to protect that structure. The legal definition of sexual harassment has provided both employees and managers with a guideline for sorting behaviors based on the harm that they do, yet may be too limited to address the other forms of gender harassment that do not involve sexual conduct. Moreover, by giving employers authority for designing and implementing grievance processes, courts have limited the protection offered to protect employee rights.

See also: Business Law; Civil Rights; Discrimination and the Law; Disputes, Social Construction and Transformation of; Employment and Labor, Regulation of; Gender and the Law; Law and Everyday Life; Law, Mobilization of; Organizations and the Law; Rights: Legal Aspects.

Bibliography

Abrams, K., 1997. New jurisprudence of sexual harassment. Cornell Law Review 83, 1169–1230.
Bennett, J., 1996. Same-sex sexual harassment. Law & Sexuality: A Review of Lesbian and Gay Legal Issues 6 (5), 1–30.
Bernstein, A., 1994. Law, culture, and harassment. University of Pennsylvania Law Review 142 (4), 1227–1311.
Bernstein, A., 1997. Treating sexual harassment with respect. Harvard Law Review 111 (2), 445–527.
Bisom-Rapp, S., 1998. Bulletproofing the workplace: symbol and substance in employment discrimination law practice. Florida State University Law Review 26, 959–1049.
Blackstone, A., Uggen, C., McLaughlin, H., 2009. Legal consciousness and responses to sexual harassment. Law & Society Review 43 (3), 631–668.
Cahill, M.L., 2001. The Social Construction of Sexual Harassment Law: The Role of the National, Organizational, and Individual Context. Ashgate/Dartmouth, Burlington, VT.
Cleveland, J.N., Kerst, M.E., 1993. Sexual harassment and perceptions of power: an under-articulated relationship. Journal of Vocational Behavior 42 (1), 49–67.
Coles, F.S., 1986. Forced to quit: sexual harassment complaints and agency response. Sex Roles 14 (1–2), 81–95.
Dellinger, K., Williams, C.L., 2002. The locker room and the dorm room: workplace norms and the boundaries of sexual harassment in magazine editing. Social Problems 49 (2), 242–257.
Edelman, L.B., Uggen, C., Erlanger, H.S., 1999. The endogeneity of legal regulation: grievance procedures as rational myth. American Journal of Sociology 105 (2), 406–454.
Elman, R.A., 1996. Sexual Subordination and State Intervention: Comparing Sweden and the United States. Berghahn Books, New York.
Farhang, S., 2010. The Litigation State: Public Regulation and Private Lawsuits in the U.S. Princeton University Press, Princeton, NJ.
Farley, L., 1978. Sexual Shakedown: The Sexual Harassment of Women on the Job. McGraw-Hill, New York.
Firestone, J.M., Harris, R.J., 2003. Perceptions of effectiveness of responses to sexual harassment in the US Military, 1988 and 1995. Gender, Work & Organization 10 (1), 42–64.
Fitzgerald, L.F., et al., 1988. The incidence and dimensions of sexual harassment in academia and the workplace. Journal of Vocational Behavior 32 (2), 152–175.
Fitzgerald, L.F., Ormerod, A.J., 1993. Breaking silence: the sexual harassment of women in academia and the workplace. In: Denmark, F.L., Paludi, M.A. (Eds.), Psychology of Women: A Handbook of Issues and Theories. Greenwood Press/Greenwood Publishing Group, Westport, CT, pp. 553–581.
Fitzgerald, L.F., Ormerod, A.J., 1991. Perceptions of sexual harassment. Psychology of Women Quarterly 15 (2), 281–294.
Fitzgerald, L.F., Swan, S., Fischer, K., 1995. Why didn't she just report him? The psychological and legal implications of women's responses to sexual harassment. Journal of Social Issues 51 (1), 117–138.
Fitzgerald, L.F., Swan, S., Magley, V.J., 1997. But was it really sexual harassment?: legal, behavioral, and psychological definitions of the workplace victimization of women. In: O'Donohue, W. (Ed.), Sexual Harassment: Theory, Research, and Treatment. Allyn & Bacon, Needham Heights, MA, pp. 5–28.
Franke, K.M., 1996. What's wrong with sexual harassment. Stanford Law Review 49 (4), 691–772.
Gelfand, M.J., Fitzgerald, L.F., Drasgow, F., 1995. The structure of sexual harassment: a confirmatory analysis across cultures and settings. Journal of Vocational Behavior 47 (2), 164–177.
Giuffre, P.A., Williams, C.L., 1994. Boundary lines: labeling sexual harassment in restaurants. Gender & Society 8 (3), 378–401.
Grauerholz, E., 1996. Sexual harassment in the academy: the case of women professors. In: Stockdale, M.S. (Ed.), Sexual Harassment in the Workplace: Perspectives, Frontiers, and Response Strategies, pp. 29–50.
Grossman, J.L., 2003. The Culture of compliance: the final triumph of form over substance in sexual harassment law. Harvard Women's Law Journal 26, 3–76.
Gruber, J.E., 1998. The impact of male work environments and organizational policies on Women's experiences of sexual harassment. Gender & Society 12 (3), 301–320.
Gutek, B.A., 1985. Sex and the Workplace. Jossey-Bass, San Francisco.
Gutek, B.A., Cohen, A.G., Konrad, A.M., 1990. Predicting social-sexual behavior at work: a contact hypothesis. Academy of Management Journal 33 (3), 560–577.
Gutek, B.A., Morasch, B., 1982. Sex-ratios, sex-role spillover, and sexual harassment of women at work. Journal of Social Issues 38 (4), 55–74.
Houle, J.N., et al., 2011. The impact of sexual harassment on depressive symptoms during the early occupational career. Society and Mental Health 1 (2), 89–105.
Kalev, A., Dobbin, F., Kelly, E., 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. American Sociological Review 71 (4), 589–617.
Kearney, L.K., Rochlen, A.B., 2012. Mexican-American and Caucasian university men's experience of sexual harassment: a preliminary report. Psychology of Men & Masculinity 13 (3), 264.
Koss, M.P., 1990. The women's mental health research agenda: violence against women. American Psychologist 45 (3), 374–380.
Lafontaine, E., Tredeau, L., 1986. The frequency, sources, and correlates of sexual harassment among women in traditional male occupations. Sex Roles 15 (7–8), 433–442.
MacKinnon, C.A., 1979. Sexual Harassment of Working Women: A Case of Sex Discrimination. Yale University Press, New Haven, CT.
Marshall, A.-M., 2001. A spectrum in oppositional consciousness: sexual harassment plaintiffs and their lawyers. In: Mansbridge, J.J., Morris, A.D. (Eds.), Oppositional Consciousness: The Subjective Roots of Social Protest. University Of Chicago Press, Chicago, pp. 99–145.
Marshall, A.-M., 1998. Closing the gaps: plaintiffs in pivotal sexual harassment cases. Law & Social Inquiry 23 (4), 761–793.
Marshall, A.-M., 2005. Idle rights: employees' rights consciousness and the construction of sexual harassment policies. Law & society review 39 (1), 83–124.
Marshall, A.-M., 2003. Injustice frames, legality, and the everyday construction of sexual harassment. Law & Social Inquiry 28 (3), 659–689.
McDonald, P., 2012. Workplace sexual harassment 30 years on: a review of the literature. International Journal of Management Reviews 14 (1), 1–17.
McKinney, K., 1992. Contrapower sexual harassment: the effects of student sex and type of behavior on faculty perceptions. Sex Roles 27 (11–12), 627–643.
McKinney, K., 1990. Sexual harassment of university faculty by colleagues and students. Sex Roles 23 (7–8), 421–438.
McLaughlin, H., Uggen, C., Blackstone, A., 2012. Sexual harassment, workplace authority, and the paradox of power. American Sociological Review 77 (4), 625–647.
Merit Systems Protection Board, 1995. Sexual Harassment in the Federal Workplace: Trends, Progress, Continuing Challenges.
Quinn, B.A., 2002. Sexual harassment and masculinity the power and meaning of "girl watching". Gender & Society 16 (3), 386–402.
Rospenda, K.M., Richman, J.A., Nawyn, S.J., 1998. Doing power: the confluence of gender, race, and class in contrapower sexual harassment. Gender & Society 12 (1), 40–60.
Saguy, A.C., 2003. What Is Sexual Harassment?: From Capitol Hill to the Sorbonne. University of California Press.
Schultz, V., 1998. Reconceptualizing sexual harassment. The Yale Law Journal 107 (6), 1683–1805.

Settles, I.H., Buchanan, N.T., Colar, B.K., 2012. The impact of race and rank on the sexual harassment of Black and White men in the US military. Psychology of Men & Masculinity 13 (3), 256–263.

Sherwyn, D., Heise, M., Eigen, Z., 2001. Don't train your employees and cancel your 1-800 harassment hotline: an empirical examination and correction of the flaws in the affirmative defense to sexual harassment charges. Fordham Law Review 69 (4), 1265–1304.

Stockdale, M.S., 1996. What we know and what we need to learn about sexual harassment. In: Sexual Harassment in the Workplace: Perspectives, Frontiers, and Response Strategies. Sage Publications, Inc, Thousand Oaks, CA, USA, pp. 3–25.

Stockdale, M.S., Vaux, A., Cashin, J., 1995. Acknowledging sexual harassment: a test of alternative models. Basic and Applied Social Psychology 17 (4), 469–496.

Tangri, S.S., Burt, M.R., Johnson, L.B., 1982. Sexual harassment at work: three explanatory models. Journal of Social Issues 38 (4), 33–54.

Tinkler, J., Gremillion, S., Arthurs, K., 2014. Perceptions of legitimacy: the sex of the legal messenger and reactions to sexual harassment training. Law & Social Inquiry, vol. 39. doi:10.1111/lsi.12065.

Tinkler, J.E., 2008. "People are too quick to take offense": the effects of legal information and beliefs on definitions of sexual harassment. Law & Social Inquiry 33 (2), 417–445.

Tinkler, J.E., 2012. Resisting the enforcement of sexual harassment law. Law & Social Inquiry 37 (1), 1–24.

Tinkler, J.E., Li, Y.E., Mollborn, S., 2007. Can legal interventions change beliefs? The effect of exposure to sexual harassment policy on men's gender beliefs. Social Psychology Quarterly 70 (4), 480–494.

Uggen, C., Blackstone, A., 2004. Sexual harassment as a gendered expression of power. American Sociological Review 69 (1), 64–92.

Uggen, C., Shinohara, C., 2009. Sexual harassment comes of age: a comparative analysis of the United States and Japan. Sociological Quarterly 50 (2), 201–234.

Vogt, D., et al., 2007. Attitudes toward women and tolerance for sexual harassment among reservists. Violence Against Women 13 (9), 879–900.

Volokh, E., 1991. Freedom of speech and workplace harassment. UCLA Law Review 39, 1791–1872.

Weeks, E.L., et al., 1986. The transformation of sexual harassment from a private trouble into a public issue. Sociological Inquiry 56 (4), 432–455.

Welsh, S., 1999. Gender and sexual harassment. Annual Review of Sociology 25 (1), 169–190.

Williams, C., 1997. Sexual harassment in organizations: a critique of current research and policy. Sexuality & Culture 1, 19–43.

Williams, C.L., Giuffre, P.A., Dellinger, K., 1999. Sexuality in the workplace: organizational control, sexual harassment, and the pursuit of pleasure. Annual Review of Sociology 25 (1), 73–93.

Zippel, K.S., 2006. The Politics of Sexual Harassment: A Comparative Study of the United States, the European Union, and Germany. Cambridge University Press, New York.

Sexual Minorities and Sexual Citizenship

Carol Johnson, University of Adelaide, Adelaide, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The concept of sexual citizenship is based on the argument that citizens' rights and entitlements have long been influenced by issues of sexuality, including both heterosexual and same-sex forms. There have been major advances in terms of ensuring that gays and lesbians have access to the same sexual citizenship rights as heterosexuals. However, there are debates over whether sexual citizenship rights focused on same-sex couples still exclude and marginalize some others. There are also questions regarding whether Western concepts of sexual citizenship are applicable to other societies internationally.

The concept of 'sexual citizenship' draws on a broad view of citizenship that sees it as involving much more than just issues of nationality, or which passport a person is entitled to hold. Rather, the broader concept of citizenship is based on the view that being a citizen of a country also means that that country, and its government, should grant its citizens various rights and entitlements. These include the right to be protected from discrimination and the right of equal access to government services and welfare benefits. It is a conception of citizenship that goes back to at least the eighteenth century in Western thought. It has been particularly well developed in European-influenced conceptions of the welfare state, in which government is seen to have a central role to play in looking after the welfare of its citizens.

Nonetheless, the idea of what the rights and entitlements of a citizen should be, and the role of government in relation to them, has always been contentious. For example, many European countries have a broader view of the health and welfare benefits that a citizen should be entitled to, and the role of the state in providing them, than is the case in the United States. Even within Europe, social democratic parties have tended to favor a more extensive welfare state than conservative parties, although the influence of free market economics has led to a winding back of benefits in many countries.

However, it is not just the nature of what the citizen is entitled to that is contentious but also the issue of *which* citizens should be included or excluded from those rights and entitlements. Many of these disagreements have centered around the politics of identity. In other words, there have been debates regarding not just what the rights and entitlements of the majority of citizens should be but also regarding what the rights and entitlements of those individuals who belong to marginalized groups should be. So, for example, there have been struggles over whether women should have the same rights as men and over whether members of marginalized racial groups should have equal rights and entitlements to those of members of powerful racial groups.

Debates over sexual citizenship therefore need to be understood within this broader context of debates over what the rights and entitlements of citizens should be. But what is sexual citizenship? Conceptions of sexual citizenship have been seen by some commentators to cover a very broad range of issues pertaining to sex. In an early landmark article on sexual citizenship, Diane Richardson (2000) identified three aspects of sexual rights, involving (1) sexual practice (from the right to engage in same-sex relations to rights to sexual activity by the disabled), (2) rights of self-definition (e.g., the right to identify as gay or lesbian, or for women to have sexual self-determination and to control their reproductive rights), and (3) rights gained via social and political institutions (e.g., the right for same-sex couples or interracial couples to marry). Related issues include the rights of citizens who identify as transgendered, transsexual, and intersex (Evans, 1993) to have equal rights, access to medical services, and to have their identities legally recognized. Others see sexual citizenship as also including the rights of citizens to be protected from sexual exploitation, including the forms of exploitation involved in pornography and sex trafficking (Hearn, 2006).

However, contemporary political debates over sexual citizenship have been largely dominated by the issue of whether citizens in same-sex relationships should have the same rights and entitlements as citizens in heterosexual relationships. So, for example, Phelan (2001: p. 5) has argued that "lesbians and gay men are not currently citizens in the full political sense" in the United States in that they do not have the same rights to equal participation and equal protection as other citizens. Importantly, the concept of sexual citizenship does not just involve an analysis of the rights of sexual minorities. Underlying many conceptions of sexual citizenship is the recognition that mainstream constructions of citizenship have always involved the sexual (Bell and Binnie, 2000). As feminist political theorists have long pointed out (e.g., Okin, 1979), when modern democratic conceptions of the citizen were first being developed in the eighteenth century, they often revolved around the traditional conception of the citizen as a male head of household with female members of the family being constructed in a subordinate position. (The citizen was also generally constructed as white and property-owning). What many earlier feminist analyses did not adequately draw attention to was that the male head of household was therefore being constructed as predominantly heterosexual, given that this was a household conceived of in heterosexual family terms. Consequently, issues of sexuality lay at the heart of conceptions of citizenship but it was a conception of citizenship that excluded same-sex relationships. So, citizenship was traditionally constructed as 'heteronormative citizenship' (Johnson, 2002) – in other words, it was constructed in ways that assumed a heterosexual norm.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10224-7

The fact that the citizen (as a concept) was constructed as a heterosexual head of family draws attention to the ways in which citizenship has always traversed the public and private spheres. Given the foundations of citizenship in personal life, citizenship issues were not confined to the public sphere of economic and political life (Hearn, 2006) such as government, electoral politics, the market, or the workplace. Similarly, the concept of citizen has always been embodied via forms of gender (and race) as well as sexuality. Indeed, government policies in diverse fields, including welfare services, immigration, and military service have helped to construct modern identities of both the heterosexual and the homosexual (Canaday, 2009).

Initially, struggles over homosexual citizenship rights often reproduced a strong public/private division (Richardson, 2000). Debates over the decriminalization of male homosexuality in the West revolved around arguments that consenting adults should be free to engage in same-sex practices in private without risking criminal charges. Richardson and Monro (2012) have described this stage in the fight for equal citizenship rights as a struggle over mis-recognition; in other words, an argument that gays and lesbians should not be characterized as pathologically deviant (and, in the case of men, criminal). The right to engage in same-sex acts in private was a very limited form of sexual citizenship right, albeit a crucial one given the previous criminal sanctions in many countries. It was also a struggle that continued in many Western countries, such as some states of the United States, until relatively recently, although Britain had decriminalized homosexuality in the 1960s and France had abolished key sodomy laws as long ago as the eighteenth century. Nonetheless, homosexuality is still illegal in many other parts of the world, including parts of Asia and Africa (Tremblay et al., 2011).

However, especially following the decriminalization of male homosexuality in the West, broader issues of sexual citizenship increasingly came to the fore. Debates over the rights of gays and lesbians began to move beyond the right of the individual to be able to engage in sexual activity in private without the threat of criminal charges. They increasingly focused on additional rights, such as the rights of gays and lesbians to live free from discrimination and from homophobic violence. Antidiscrimination measures were successfully fought for in a range of countries, including South Africa and parts of South America (Tremblay et al., 2011). However, these struggles are ongoing even in some Western countries, for example, there is still not a federal antiworkplace discrimination policy in the United States and there are ongoing debates in Australia over whether religious organizations should continue to have exemptions from antidiscrimination legislation.

Increasingly too there were arguments regarding the rights of couples in same-sex relationships to have the same entitlements as those in heterosexual relationships. In other words, the gay and lesbian movement began to challenge heteronormative constructions of the citizen by arguing that same-sex couples should be recognized by the state just as heterosexual couples are and should therefore have the same entitlements as those heterosexual couples. The debates over government recognition also contributed to breaking down the public/private division in other ways as gays and lesbians increasingly argued that they should be able to live open lives, including displaying affection in public, rather than having to 'pass' as heterosexual in public and confining same-sex sexuality to the closet (Johnson, 2002). The late twentieth and early twenty-first century has seen major advances in many European countries, Canada, the United States, Australia, South Africa, and parts of Latin America (Tremblay et al., 2011) in terms of ensuring that same-sex couples can access at least some of the same rights and entitlements as heterosexual couples, although marriage still remains a stumbling block in many countries.

In other words, the gay and lesbian movement increasingly challenged the heterosexual construction of citizen identity and the resulting granting of rights, benefits, and entitlements to heterosexual couples only. (The potential problems that arise from largely accepting the couple as the basis of sexual citizenship will be explored later). The areas that can be impacted on by government recognition of same-sex couple relationships are extensive and range from issues of superannuation benefits and pensions, taxation, immigration, consent to medical treatment, welfare and health benefits, to family rights such as whether parenting relationships are recognized and to adoption rights. In the United States, it was estimated some years ago that there were an estimated 1049 federal rights and entitlements that heterosexual married couples could access, ranging from social security to employment benefits, that same-sex couples were not eligible for (Kotluski, 2004). Depending upon what rights and entitlements are granted to citizens in a particular society, sexual citizenship can therefore have enormous implications. These implications occur whenever government, or other relevant authorities, recognizes a sexual relationship for conferring benefits and entitlements. In some countries that only involves recognizing people in married relationships or in registered civil partnerships but in some other countries (such as Australia) it involves recognizing unmarried people who have lived together for a number of years.

So, sexual citizenship rights are far from an abstract issue. Sexual citizenship rights impact on whether you can marry the person that you love. If you fall in love with someone from another country, sexual citizenship rights can impact on whether you are able to bring your lover to live with you in your country. If couples can combine their incomes for taxation purposes, sexual citizenship rights can impact on the amount of taxation that you pay. If you die, sexual citizenship rights can impact on whether a surviving lover is eligible for a pension or superannuation benefits. Sexual citizenship rights can impact on whether you are able to give consent to your lover's medical treatment if they are too ill to make decisions for themselves or whether you are able to visit them in hospital. In the case of same-sex families, sexual citizenship rights can influence whether same-sex couples have the right to adopt or foster children, whether surrogacy arrangements are recognized and whether parenting relationships by nonbiological parents are recognized.

In other words, sexual citizenship has implications for a range of intimate relationships including both couple relationships and other forms of sexual intimacy. Consequently, sexual citizenship interacts with, and is indeed a form of, another version of citizenship that has been identified, namely, 'intimate citizenship' (Plummer, 2003; Roseneil, 2010). Intimate citizenship recognizes the relationship between citizenship and intimate personal relationships. However,

a difference between intimate citizenship and sexual citizenship is that the conception of intimate citizenship recognizes forms of intimate relationships, including caring relationships, that go beyond the sexual, for example, the relationships of friends, family, and extended kinship relationships. Sexual citizenship also impacts on emotional issues such as love. Consequently, sexual citizenship is a form of 'affective citizenship' (Johnson, 2010), in that it is a form of citizenship shaped by which emotions between individuals are recognized as legitimate by the state, for example, whether same-sex love is recognized. However, as is the case with intimate citizenship, affective citizenship also stretches beyond the sexual to include other forms of personal and public emotions. Consequently, what is distinctive about sexual citizenship, compared with other forms of citizenship, is the focus on the 'sexual' as a component of citizenship. Furthermore, while sexual citizenship can and does intersect with other forms of social marginalization such as gender and race (e.g., when lesbians are not seen to be sufficiently feminine or when black men are constructed as sexually dangerous to white women), the primary focus of 'sexual citizenship' is on the 'sexual.'

Note that the rights associated with sexual citizenship are not just economic and social, they also have cultural ramifications in terms of people's right to be recognized and represented symbolically (Bell and Binnie, 2000), not merely in mainstream political culture, for example, in the discourse of politicians, but also in the media and other forms of popular culture. Sexual citizenship rights can also impact on how people are treated in everyday life, for example, whether people are likely to be subject to homophobic abuse and violence. Here, as elsewhere, issues of sexual citizenship can go beyond issues of state recognition to include forms of social recognition, in terms of social attitudes and how people are treated in everyday life.

Critiques of Sexual Citizenship

While many gay and lesbian organizations internationally have seen the struggle for sexual citizenship as an important part of the struggle for same-sex rights (Tremblay et al., 2011), there are nonetheless a number of critiques that can be made of aspects of sexual citizenship. In particular, critics who draw on queer theory have expressed concern about the normalizing effects of sexual citizenship. After all, having the state recognize particular forms of relationships for the purpose of granting rights and entitlements can construct some forms of personal sexual relationships as the norm and others as deviant. In the process, state recognition can deny legitimacy to some of the diverse forms that same-sex sexual expression can take, including those that may be influenced by class and ethnic factors, and add to the pressures on those who are already stigmatized (Barrett and Pollock, 2011). Many queer critics have pointed out that governments have tended to only recognize marriage-like relationships that could be seen as mimicking traditional heterosexual relationships. As with same-sex marriage itself, the question therefore arises of whether granting sexual citizenship rights potentially marginalizes those gays and lesbians who engage in other forms of sexual relationships, for example, non-monogamous ones that involve multiple sexual partners or sexual relationships that are not long-term sexual partnerships (Butler, 2002). In other words, sexual citizenship rights could be the latest way in which the state constructs 'good' and 'bad' homosexual citizens (Smith, 1994), with only those homosexuals who ape heterosexual marriage relationships being constructed as 'good' homosexuals and others still being constructed as deviant and engaged in antisocial behavior. As Bell and Binnie (2000: p. 30) have put it: "given the heterosexualisation of Citizenship how can rights claims based on citizenship status from sexual minorities be made to work other than by replicating heterosexualized articulations of the 'good citizen.'"

It is not just that some forms of same-sex relationships are potentially privileged over others in ways that can exclude and marginalize non-marriage like relationships. Importantly, Richardson and Monro (2012) have noted that arguments based on the view that citizens who are in same-sex relationships should have the same rights and entitlements as heterosexual citizens can exclude transcitizens. Transgendered, transsexual, or intersex citizens do not fit into either same-sex or traditional heterosexual or gay and lesbian identity categories. Such negative outcomes do not surprise those queer critics who argue that conceptions of citizenship have always been exclusionary (Payne and Davies, 2012) because they were initially developed as ways of delineating national belonging from forms of racial 'otherness.' Furthermore, some proponents of broader forms of intimate citizenship would question the very focus on using sexual relationships as the constitutive feature of the unit of citizenship, pointing out that this can limit the state's recognition of other forms of intimate relationship which might be just as important as, or even more important than, sexual relationships in contemporary society, for example, close friendships (Roseneil, 2010).

In addition, left-wing commentators, whether drawing on queer theory or not, have noted the difficulties involved in trying to develop political strategies for advancing issues of sexual citizenship in a context in which both conceptions of the sexual and of citizenship are being commodified in capitalist societies. Such critiques include Evans' (1993) arguments regarding the way in which sexual identity has been reduced to a consumer choice lifestyle in a way that fragments political dissent and related analyses of the ways in which citizenship has been increasingly privatized under the influence of right-wing social and economic thought, so that benefits which were once delivered by the state are now seen to be obtainable only via the market (Berlant, 1997; Smith, 1994). Similarly, Stychin (2006: p. 913) argues that the recognition of the same-sex couple, via the United Kingdom's civil partnership scheme, involved the privatization of care to same-sex partners and the state reducing its responsibility accordingly. Consequently, those who are same-sex active but who do not engage in state-approved forms of sexual relationships, or who do not identify as gay or lesbian, or who lack necessary economic means, can be excluded from existing forms of sexual citizenship. The concept of sexual citizenship therefore remains a highly contentious one, partly because the very processes of inclusion that many see as a sign of major reform having been achieved can also involve redrawing citizenship boundaries in a way that excludes or marginalizes others. There are also issues regarding whether sexual citizenship as a concept can

adequately include the identities, experiences, and struggles of members of marginalized sexual groups living in diverse countries, especially those outside the West.

Are Western Conceptions of Sexual Citizenship Readily Applicable to Diverse Societies?

Bell and Binnie (2006) have pointed out that analysis of sexual citizenship has always involved having to engage with different geographies of sexual citizenship. As the earlier account suggests, it is difficult enough to give an account of sexual citizenship that covers Western societies given the differences between them. For example, it was noted that some European societies have a far more extensive conception of what the health and welfare entitlements of citizens should be, and the state's role in providing that, than in the United States. The state's forms of recognition of sexual relationships also differ markedly between countries, influenced by different institutions, cultures, and political histories (Smith, 2008; Tremblay et al., 2011). For example, in the United States, couples need to be married in order to access many federal entitlements (Kotluski, 2004). By contrast, in various countries, including some states of the United States, there are forms of civil partnership registration that make couples eligible for relevant government benefits. In Australia, both heterosexual and same-sex couples merely need to have lived together for a number of years in a *de facto* relationship, they do not have to be married for their relationship to be recognized for the purposes of federal government entitlements (Johnson et al., 2011). Similarly, countries have very different ways of recognizing parental relationships in same-sex families, if they do at all. So conceptions of sexual citizenship need to be sufficiently flexible to allow for differences in the ways in which particular countries have developed and constructed citizenship rights.

The situation is even more complicated if one tries to merely apply existing Western typologies and conceptions of sexual citizenship to nonWestern societies. There are potential problems in simply exporting Western concepts of sexual citizenship rights, based on particular assumptions about the respective roles of government and communities, to societies where there is not a clear division between the state and civil society. For example, the division between state and civil society is much less clear in societies from the Middle East to parts of Southeast Asia in which the sharia law operates and in which key legal decisions regarding personal relationships are made in religious rather than state courts. Similarly, in many societies, sexual behavior and expressions of identity may be 'policed' more by community groups, including forms of religious vigilantes in some extreme cases, than by agents of the state. Are these still 'citizenship' issues in same way as in the 'West'? They do still involve broader issues of social rights, or of social citizenship. However, they do not directly involve issues of state-based citizenship rights insofar as government has subordinated its rights to make decisions about such matters to religious powers. Similarly, it would be hard to establish that some issues seen as sexual citizenship issues in the West are sexual citizenship issues in societies where government welfare state entitlements are not so well developed and extensive – and where it therefore is not so easy to establish that state entitlements are being directly impacted by nonrecognition of same-sex relationships. At the least, conceptions and strategies of sexual citizenship used in the West would need to be extended and revised in order to cover such societies, the varying forms discrimination can take, and the diverse sites where political struggles can occur.

Even more fundamentally, Jeffrey Weeks (1999: p. 36) has suggested that the emergence of the sexual citizen is related to a now 'commonplace' development 'at least in the metropolitan heartlands of Western Societies' for members of previously marginalized sexual groups to claim an *identity* on the basis of that sexual identity and corresponding *rights* related to that identity. However, it has been argued that in some societies, such as China, sexuality may not be seen as being such a crucial part of identity as in the West (Chou, 2001). Though it should be noted that this is contentious since others would argue that sexual identity plays a more explicit role than Chou acknowledges in many 'Asian' societies, including Chinese ones (see e.g., Khor and Kamano, 2006). There are also questions about when it is appropriate to designate issues of citizenship as issues of 'sexual' identity as opposed to issues of gender identity. Some matters commonly considered to be 'sexual' in the West are also not necessarily designated in the same way in other societies. For example, Jackson (2001) has argued that some behaviors and attributes that would be constructed in terms of 'sexual' identities in the West can be understood better in terms of (multiple) eroticized 'genders' in Thailand. In other words, one's form of masculinity or feminity is seen to explain one's preference in sexual partners, with eroticism being understood 'as a secondary or derivative component' of one's gender identity (Jackson, 2001: p. 18). Similarly, Atluri (2012) cautions against assuming that Western conceptions of state-sanctioned rights of sexual citizenship can adequately encompass the identities, religious significance, or transgressive public performances of 'sexual' minorities such as Indian Hijras, or adequately address Indian traditions of transgressive sexual politics that involve conceptions of sexual renunciation and sacrifice rather than sexual activism.

Nonetheless, despite such reservations, sexual citizenship is increasingly taking on international forms. This is especially the case as there are increasing calls for same-sex issues to be seen as a crucial part of international human rights regimes, including via the United Nations (Clinton, 2011; Ki-moon, 2012). However, writers such as Atluri (2012) and Sabsay (2012) caution against constructing conceptions of Western superiority in regard to rights, especially ones based on conceptions of a vulnerable, sexually victimized, racial 'other' that needs to be liberated by Western interventions. After all, there is also a long history of Western colonialism, in which the rights of racial 'others' were repressed to ensure colonial rule. In particular, colonial rule often involved the imposition of Western laws outlawing homosexuality, including in societies where homosexuality had sometimes not been specifically criminalized before. Britain's notorious section 377 was imposed on various Asian societies that were part of the British empire, and still exists in India, Malaysia, and Singapore. Indeed, British colonial law also stigmatized Indian Hijras, reducing their traditional religious and social status.

At the least, trying to apply conceptions of sexual citizenship to diverse, non-'Western' societies draws attention to the need

to recognize the different ways in which citizenship is constructed, and the varied relationships between state and society that can exist in different parts of the world. However, it should be noted that there are nonetheless many parts of the world where at least some aspects of existing analyses of sexual citizenship analyses work well. This is particularly the case if one focuses on the heteronormative aspects of sexual citizenship, given the widespread nature of constructions of the citizen as implicitly heterosexual. There are also many parts of the world where the fight for the rights of sexual minorities, even if just for basic issues such as decriminalization of male homosexuality involves battles over rights that other citizens take for granted. Nonetheless, Western gays and lesbians need to respect the decisions of sexual and gender minorities elsewhere about whether they find struggles over 'sexual citizenship' to be the best way in which to formulate their own specific social and political aims. They also need to acknowledge that advances in gay and lesbian rights have not been confined to Europe or North America but have also taken place in countries ranging from South Africa to parts of South America (Tremblay et al., 2011).

The issue of the internationalization of sexual citizenship is further complicated by the fact that there are numerous transnational international sites where issues of sexual citizenship are both located and being contested. The United Nations has already been mentioned. However, as Hearn points out, these transnational sites also include the internet, which is a site where both marginalized groups such as gays and lesbians, are organizing to struggle for sexual citizenship rights, and where diverse issues of abuse of sexual citizenship themselves arise. The latter go far beyond same-sex issues to include, for example, issues involving child pornography or the sexual trafficking of women (Hearn, 2006).

Conclusion

Sexual citizenship has been a very contested site of political struggle, precisely because the sexual has been so crucial in constituting the unit of citizenship in most societies. As Richardson and Monro point out (2012: pp. 12–13), given the history of gays and lesbians being depicted as unnatural and a threat to society "for LGBT people to be recognized and accepted as 'ordinary citizens' who deserve equal rights suggests, therefore, that there has been a profound shift in the meanings attached to sexuality, gender and citizenship." That change has only been won because of long struggles by the gay and lesbian movement. This is not to deny that critiques of the normalizing, and potentially exclusionary, tendencies implicit in even the reformed versions of sexual citizenship have their point. However, given that the alternative is to leave sexual citizenship as heteronormative citizenship, that constructs the heterosexual as the norm, most gay and lesbian organizations (in the West at least) have decided that the benefits of ensuring the political recognition of same-sex relationships outweigh the potential disadvantages. Many theorists of sexual citizenship agree, arguing that the choice of remaining 'noncitizens' is not viable since it would involve harmful forms of exclusion and discrimination (Bell and Binnie, 2000: p. 146). Nonetheless, serious issues remain, including the need to ensure that conceptions of sexual citizenship rights include transgender, transsexual, and intersex issues. Finally, conceptions of sexual citizenship need to be flexible enough to cover the multiple ways in which citizen identities and entitlements, and relationships between the state and society, are constructed in different national, international, and transnational contexts.

See also: Citizenship: Political; Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Heterosexism and Homophobia; Lesbian and Gay Activism in Latin America; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations; Queer Anarchism; Same-Sex Sexualities in Turkey; Sexual Minorities in China; Sexual Minorities in Modern Russia; Sexual Minorities in South Africa; Sexual Minorities in the Maghreb; Sexual Minority Families of Choice and Family Networks: We Are Family; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Atluri, Tara, 2012. The prerogative of the brave: Hijras and sexual citizenship after orientalism. Citizenship Studies 16, 721–736.

Barrett, Donald C., Pollock, Lance M., 2011. Testing a typology of adaptions to same-sex sexual orientation among men. Sociological Perspectives 54, 619–640.

Bell, David, Binnie, Jon, 2000. The Sexual Citizen: Queer Politics and Beyond. Polity, Cambridge.

Bell, David, Binnie, Jon, 2006. Geographies of sexual citizenship. Political Geography 25, 869–873.

Berlant, Lauren, 1997. The Queen of America Goes to Washington City. Duke University Press, Durham.

Butler, Judith, 2002. Is kinship always already heterosexual. Differences 13 (1), 14–34.

Canaday, Margot, 2009. The Straight State: Sexuality and Citizenship in Twentieth Century America. Princeton University Press, Princeton.

Chou, Wah-Shan, 2001. Homosexuality and the cultural politics of Tonghzhi in Chinese societies diversity in Asia. In: Sullivan, G., Jackson, P.A. (Eds.), Gay and Lesbian Asia: Culture, Identity, Community. Harrington Park Press, New York, pp. 27–46.

Clinton, Hillary, 19 December 2011. Secretary Clinton: Gay Rights are Human Rights. IIP Digital, US Department of State video. http://iipdigital.usembassy.gov/st/english/video/2011/12/20111219142526aerdna0.5667492.html#axzz1j4Fzifr0 (accessed 28.12.12.).

Evans, David T., 1993. Sexual Citizenship: The Material Construction of Sexualities. Routledge, London.

Hearn, Jeff, 2006. The implications of information and communication technologies for sexualities and sexualised violences: contradictions of sexual citizenships. Political Geography 25, 944–963.

Jackson, Peter. A., 2001. Pre-gay, post-queer: Thai perspectives on proliferating gender/sex diversity in Asia. In: Sullivan, G., Jackson, P.A. (Eds.), Gay and Lesbian Asia: Culture, Identity, Community. Harrington Park Press, New York, pp. 1–26.

Johnson, Carol, 2002. Heteronormative citizenship and the politics of passing. Sexualities 5, 316–336.

Johnson, Carol, 2010. The politics of affective citizenship: from Blair to Obama. Citizenship Studies 14, 495–509.

Johnson, Carol, Maddison, Sarah, Partridge, Emma, 2011. Australia: parties, federalism and rights agendas. In: Tremblay, M., Paternotte, D., Johnson, C. (Eds.), The Lesbian and Gay Movement and the State: Comparative Insights into a Transformed Relationship. Ashgate, Houndmills, pp. 27–42.

Khor, Diana, Kamano, Saori (Eds.), 2006. 'Lesbians' in East Asia: Diversities, Identities and Resistance. Harrington Park Press, Binghampton, NY.

Ki-moon, Ban, 2012. Message to Human Rights Council Meeting on Violence and Discrimination Based on Sexual Orientation or Gender Identity, Secretary-General Ki-moon, Ban, Geneva (Switzerland), 7 March 2012, 1475, Ban Ki-moon's speeches, United Nations News Centre http://www.un.org/apps/news/infocus/sgspeeches/statments_full.asp?statID=1475 (accessed 28.12.12.).
Kotluski, Davina, 2004. Why You Should Give a Damn About Gay Marriage. Advocate Books, Los Angeles.
Okin, Susan M., 1979. Women in Western Political Thought. Princeton University Press, Princeton.
Payne, Robert, Davies, Cristyn, 2012. Introduction to the special section: citizenship and queer critique. Sexualities 15, 251–256.
Phelan, Shane, 2001. Sexual Strangers: Gays, Lesbians and Dilemmas of Citizenship. Temple University Press, Philadelphia.
Plummer, Ken, 2003. Intimate Citizenship: Private Decisions and Public Dialogues. University of Washington Press, Washington.
Richardson, Diane, 1998. Sexuality and citizenship. Sociology 32, 83–100.
Richardson, Diane, 2000. Constructing sexual citizenship: theorizing sexual rights. Critical Social Policy 20, 105–135.
Richardson, Diane, Monro, Surya, 2012. Sexuality, Equality and Diversity. Palgrave Macmillan, Basingstoke/New York.
Roseneil, Sasha, 2010. Intimate citizenship: a pragmatic, yet radical, proposal for a politics of personal life. European Journal of Women's Studies 17, 77–82.
Sabsay, Leticia, 2012. The emergence of the other sexual citizen: orientalism and the modernisation of sexuality. Citizenship Studies 16, 605–623.
Smith, Anna M., 1994. New Right Discourse on Race and Sexuality: Britain, 1968–1990. Cambridge University Press, Cambridge.
Smith, Miriam, 2008. Political Institutions and Lesbian and Gay Rights in the United States and Canada. Routledge, London.
Stychin, Carl F., 2006. 'Las Vegas is not where we are': queer readings of the Civil Partnership Act. Political Geography 25, 899–920.
Tremblay, Manon, Paternotte, David, Johnson, Carol, 2011. Introduction. In: Tremblay, M., Paternotte, D., Johnson, C. (Eds.), The Lesbian and Gay Movement and the State: Comparative Insights into a Transformed Relationship. Ashgate, Houndmills, pp. 1–12.
Weeks, Jeffrey, 1999. The sexual citizen. In: Featherstone, M. (Ed.), Love & Eroticism. Sage, London, pp. 34–52.

Relevant Websites

http://www.amnesty.org/en/sexual-orientation-and-gender-identity – Amnesty International, Sexual Orientation and Gender Identity.
http://ilga.org/ – International Lesbian, Gay, Bisexual, Trans and Intersex Association.

Sexual Minorities in China

Day Wong, Hong Kong Baptist University, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In postreform China, many sexual minorities have developed an awareness of sexual rights and a sense of belonging in a queer globality. However, as they are subject to the imperative of heterosexual marriage, sexual orientation and identity tend to be fluid and fragmented. The vast majority of sexual minorities enters heterosexual marriage, while some gays and lesbians have begun using the strategy of cooperative marriage. Despite decriminalization and depathologization of homosexuality, homosexuality remains stigmatized as it has been linked to HIV/AIDS, and married homosexual men are held responsible for spreading HIV to the low-risk population. Divisions among sexual minorities are witnessed along urban/rural and class lines.

Definitions and Prevalence

Sexual minorities are commonly defined as groups whose sexual identity, orientation, or practices differ from the majority of the surrounding society. In China where cultural prohibition against sexual minorities is strong and identity politics are not prevalent, it is difficult to draw an unequivocal dividing line between the sexual minority and the larger majority, nor can we know with certainty the size of sexual minorities.

Social science research on homosexuality in China has relied on descriptions of behaviors, desires, and attractions rather than sexual identities. The Sex Civilization Survey, the first nationwide survey of sexual behaviors, collected data from 20 000 subjects in China in 1998–99. Among the married respondents, 0.5% of the urbanites and 2.3% of the rural dwellers reported experiences of same-sex intimacy (Liu and Lu, 2005). The survey was implemented through official organizations and conducted in institutional settings. The response rate was very high, but given the sensitivity of the homosexuality issue in China, underreporting on same-sex relationships can be expected. Another nationwide random survey was carried out in 2001 with university students (Pan and Yang, 2004). To encourage fuller reporting of such sensitive issues, the survey relied on mailed questionnaires, with unidentifiable, return address typed envelopes provided to guarantee anonymity. Among the 961 respondents, 11.4% admitted different levels of homosexual orientation in a psychological sense and 17.6% reported experiences of having physical contact of a sexual nature with persons of the same sex. If being a homosexual is defined by both psychological and physical dimensions, 4.4% had both kinds of experience. Based on these surveys, official estimates of the homosexual population in China range from 2 to 4%. More liberal estimates would say around 7% if erotic same-sex hugging is taken as an indicator, or 11% if the psychological dimension is taken into consideration.

Questions can be raised as to whether those who have had a same-sex experience should be regarded as bisexual or homosexual. Among students who have psychological inclination, but not physical contact, 36.8 and 25.0% of the male and female groups, respectively, had experienced heterosexual sexual intimacy. Even among those who are homosexual in both the physical and psychological sense, 5.6 and 11.5% of the male and female groups, respectively, had experienced heterosexual sexual intimacy. The researchers remarked that rather than being exclusively homosexual, a substantial proportion of the students who had same-sex experience can be regarded as bisexual (Pan and Yang, 2004).

The tendency to have both homosexual and heterosexual experiences is not unique to university students. When looking at the more mature population such as married men, such tendency cannot be regarded as sexual exploration. Based on a 1998 survey, Zhang and Chu (2005) estimated that 70–80% of male homosexuals/bisexuals in the cities have married or will marry a woman. The proportion in the rural population can be as high as 90%. If sexual orientation is defined by sexual desire or attraction rather than sexual behavior, it is far from clear if these individuals are attracted to both sexes or obliged to live out the cultural prescription. In China where people consider it a moral imperative to marry, sexual orientation tends to be fluid and fragmented.

Family Obligations and Imperative of Heterosexual Marriage

China has undergone massive transformation from a command to a market economy since the reform in 1978. The loosening of state control in employment, commodities and the market has enabled Chinese people to enjoy unprecedented freedom. Young people in rural China have left family farming to look for employment opportunities in cities. Urbanites also experience a newfound freedom as a result of the decline of the work unit (*danwei*) system, which used to provide living quarters and food, approve marriage and travel, and exercise control over the lives of individuals. For many sexual minorities, geographical mobility, the emergence of the private housing market, and the market reform as a whole have enabled them to gain privacy and choice in intimate relationships. Living away from parents, being socially anonymous as part of city life, and having economic opportunities to make an independent living, all point toward the possibility of pursuing a sexual life away from the conventional norm of heterosexual marriage. In Western countries, the rise of industrial capitalism and its

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10247-8

resulting urbanization brought about similar opportunities, which enabled homosexual desire to coalesce into a personal identity. Individuals constructed their personal lives and social networks based on same-sex attraction (Weeks, 1977; D'Emilio, 1980/1993). In China, however, family pressure remains a major source of concern for sexual minorities.

With the retreat of the state in the provision of social welfare, the family has been looked upon as an indispensable unit of care. Motivated by the practical issue of how to care for an aging population, the Chinese government has reinvigorated filial piety, a core Confucian virtue. The Law on Protection of the Rights and Interests of the Elderly was passed in 1996 and the Chinese Filial Piety Awards Ceremony is held annually. Chinese sexual minorities do not simply observe filial piety out of compliance to state initiative. In an age of high mobility, fierce competition and unpredictable employment, the family serves as a collective of pragmatic, emotional, and symbolic importance. Parental support is relied on for educational attainment, establishment of social network, and career investment. The implementation of the one-child policy in 1979 has shaped the intergenerational relationship to be one characterized by emotional attachment rather than authoritarian parenting. Sexual minorities feel obliged to fulfill parental expectations in order not to cause them grief or shame.

When one reaches the age of marriage, parents and relatives are often keen to arrange meetings to introduce prospective partners. The intense pressure to marry is felt by most Chinese sexual minorities – the only child generation. Sons, in particular, bear the responsibility of having children and carrying on the family bloodline. The marriage status proves that one is a responsible adult, which saves the face of the parents and facilitates career advancement. Women are in no way better than men in escaping marriage pressure (Kam, 2013). Their freedom to move out of the parental home before marriage or relocate to another city is more restricted in comparison to men. The disadvantaged position of women in the job market and the gender income gap also make it difficult to gain financial independence outside marriage. Although popular media as well as scholars have used 'sexual revolution' to describe the changes in China, marriage remains the prevailing norm. Unlike those in the West who have delayed marriage or elected not to marry, the gap between the ages of first sex and marriage is small (Parish et al., 2007).

Based on ethnographic dialogs with over 520 homosexual males, Pan et al. (1995) found that those who were married gave reasons, which clustered around (1) believing that they must marry, otherwise they would be looked down upon by society, parents, and neighbors; (2) career advancement or leadership position in work units demands that one must be married; (3) their mothers want grandchildren; and (4) they fear old age alone. In the past, homosexual men buried their sexuality in (heterosexual) marriage. Such a strategy is now considered to be unethical and cause a great deal of pain to the marital partners. In China's sexual revolution, sexual happiness is regarded as the key to a harmonious conjugal family, which in turn, would promote social stability. Not only men but also women are entitled to enjoy sexual pleasure. The demand for an active and harmonious sex life leads to increased difficulties for homosexuals who opt to wed. In recent years, the problem of tongqi (wife of homosexual man) has been brought to public attention. It is estimated around 12 million straight women who are unknowingly married to gay men. Support groups have been established for these women, who suffer in a marriage that lacks love and sex.

In order to accommodate their same-sex relationships while minimizing direct conflicts with the family, some gays and lesbians opt for a cooperative marriage. Aided by the Internet, gays and lesbians are now seeking each other out for a public heterosexual marriage, so that they can pursue a same-sex relationship in their private lives. Cooperative marriage is often called 'fake marriage' (*jia jiehun*) or 'ritual/pro forma marriage' (*xingshi hunyin*). Some couples hold a marriage ceremony but do not apply for a marriage license. Other cooperative marriages can vary according to the living and financial arrangements, or the preference for children. Gays and lesbians who look for prospective partners on dating Web sites can discuss and agree on the terms of marriage. A gay newsletter published in 2008 reported that of the 100 respondents, 91% had learned about cooperative marriage, 24% were considering this option, and 20% had already opted for cooperative marriage (Chen, 2009).

The feasibility, strengths, and weaknesses of cooperative marriage are very much under debate. A study by Kam (2013) of the lesbian community in Shanghai showed that for some lesbians, marriage is the only way to move out from their parental home and escape parental control. They understand cooperative marriage as a compromise that fulfills parental expectations. Yet this conformity to the heteronormative model enables them to carve out space for their same-sex relationships. A cooperative marriage may develop as a household that welcomes other queer members for social gatherings. Hence, it has the potential of becoming an extended queer kinship network that includes the two same-sex couples as well as other community members.

Cooperative marriages are not void of the secular concerns found in any marriage. The fact that it is a performative union makes it liable to exemplify mainstream values. Gay men often prefer their marital partners to look feminine and pretty, thus tomboy lesbians are ruled out as a possible candidate. Many also expect their lesbian partners to have children through natural or artificial means. Lesbians require their marital partners to have a good job and own an apartment. Many refuse to consider gay men who are sissy or like to mingle in gay circles. Inequality, exclusion, and conflict are played out in the mate selection process of cooperative marriages. A study by Chen (2009) revealed that some lesbians fall in love with their marital partners in cooperative marriages. Their same-sex partners often feel excluded and jealous especially when lesbians have children with the marital partners and became more involved in the heterosexual families. Thus marriage and family life, which promise public recognition and security, can entice an individual away from the queer world.

Decriminalization, Medicalization, and Stigmatization

Many people believe that there is an increasing acceptance toward sexual minorities, by citing the evidence that homosexuality in China was decriminalized in 1997 and delisted as a mental illness in 2001. However, police intervention in the

activities of sexual minorities is common and sexual minorities continue to be the object of medical surveillance. As Kang (2012) pointed out, decriminalization of homosexuality in China should be understood as an unintended consequence rather than a sign of the victory of sexual rights.

In the Mao era and the early reform period, homosexuality was penalized under the category of hooliganism. The 1979 Criminal Law included forced sodomy and sodomy with minors as 'other hooligan activities' that undermine public order. Although the law did not mention consensual sodomy, the police regarded anal sex between men in both private and public as a type of general hooliganism that deserves arrest. Aside from the criminal penalty, homosexual men who are members of the Communist Party could be punished by having their membership removed. Nonparty members could lose their jobs, be detained without trial or sent to labor camps (Kang, 2012).

As the Chinese economic reform began to take root, legal reform was considered necessary to protect individual and corporate interests, and provide legitimacy for the government. The revised Criminal Law in 1997 emphasizes that all crimes must be expressly prescribed by the law, thus removing hooliganism for being an ill-defined, catchall crime. Sodomy with minors is prohibited under the criminal code of indecent assault. However, it is worth noting that the new law has no mention of forced sodomy between adults. This proves that the legal code in China is dominated by heteronormative assumptions and the issue of homosexuality has never been inside the purview of Chinese lawmakers (Kang, 2012).

The removal of hooliganism does not imply that sexual minorities in China are free from the oppression of the penal system. Aside from criminal law, they are subject to the Public Order Administrative Penalty Regulation. Homosexuals can face administrative penalties for engaging in prostitution, and group and public sex. When arrested by the police, they are often harassed, detained, and charged with fines. While sexual minorities in the United States face random street violence, their counterparts in China face police harassment. Police raids of gay bars, discos, and other establishments seem to be largely motivated by monetary gain, which indicates that the queer community is a victim of bureaucratic corruption.

Homosexuality is subject not only to legal punishment, but also to medical pathologization. Prior to 2001, homosexuality was classified as abnormal psychology and reparative therapy was carried out on some homosexual people. Academics and experts who published on homosexuality might express sympathy toward the suffering of homosexual people; yet they all tended to pathologize homosexuality and recommended scientific sexual education as the preventive measure.

The delisting of homosexuality as a mental illness by the American Psychiatric Association and World Health Organization, in 1986 and 1990, respectively, stirred up debates within the mental health professional community in China. In the *Chinese Classification of Mental Disorders* published in 1989 and 1994, homosexuality continued to be officially defined as a form of sexual disorder. The Chinese community of psychiatry saw itself as a doorkeeper to watch over the health of the national body and safeguard against the invasion of a decadent foreign culture. Although the majority in the field held that homosexuality was 'a strange disease that deserves denunciation,' nevertheless, opinions were diverse, which ranged from 'crime' and 'disorder,' to 'unclear' and 'natural,' and even to 'a matter of minority's oppression.' The continued efforts of activists in China and overseas, which include mobilization of support from the American Psychological Association and American Psychiatric Association, finally led to the removal of homosexuality from the list of mental illnesses by the Chinese Psychiatric Association in 2001 (Wu, 2003).

While the medical discourse in contemporary China no longer condemns homosexuality as a mental illness, homosexuality has been linked to HIV/AIDS, which are considered to be a sign of moral bankruptcy under Western influence. By linking homosexuality to AIDS, public discourses created social fear and reinforced the stigmatization of homosexuality. Chinese sexologist Pan Suiming (cited in Kam, 2013: p. 47) pointed out that AIDS prevention became the reason for the police and other authorities to harass and detain homosexuals. This situation is the result of the entrenched distrust held by legal authorities and public health professionals against homosexual people.

An increasing awareness of the AIDS threat in the 1990s caused Chinese officials, academics, and clinicians to pay attention to the specific population of men who have sex with men (MSM). The first survey on MSM in 1997–98 found that two-thirds of the respondents have never used condoms, half have engaged in intercourse with women, and one-third have been married. Due to the likelihood that HIV would not only spread among MSM, but also to the female population, the researchers warned that the spread of HIV in China may be faster than in the Western countries in the 1990s (Zhang et al., 2000).

Subsequent studies highlight the lack of knowledge and awareness of HIV risk among MSM and the role of public education in AIDS prevention. Other studies, nevertheless, showed that although many MSM have good HIV/AIDS knowledge awareness, high-risk sexual practices are still common and the rate of consistent condom use is low. It is necessary to examine the more deep-rooted cultural forces that shape one's decision to engage in protective practices. Prevention efforts can be hindered because of the stigma associated with homosexuality in the Chinese culture. To avoid exposure of their homosexuality, MSM were found to be reluctant to seek out health services or accept outreach services even though these are free of charge. In addition, the stigma associated with HIV/AIDS can cause further hindrance and lead to a low level of HIV testing.

More in-depth studies have been carried out on the stigma faced by MSM. A qualitative study by Liu and Choi (2006) with MSM in Shanghai found that the stigma faced can be traced to four cultural factors: social status and relationships, family values, perceptions of immorality and abnormality, and gender stereotypes of masculinity. Liu et al. (2011) examined the influence of both collectivism and individualism, which are both promoted in contemporary China, on the development of homosexuality stigma in an inequitable society. MSM who hold vertical collectivism tend to see the importance of social status, accept inequality, and conform to collectivist social norms. Consequently, they are more likely to perceive a higher level of public stigma. Presumably, MSM with an

individualistic orientation place self-interest above collective interest and disregard public attitudes toward their homosexuality. However, this does not apply to vertical individualism whereby individuals want to do their own thing and be the best. With aspiration for success and self-accomplishment, they tend to experience a higher level of self-stigma, that is, the fear of societal discrimination against them.

With regard to the rising trend of HIV infection, national HIV prevalence among MSM had increased from 1.4% in 2001 to 5.3% in 2009. In 2012, the prevalence rate was 6.4%, in which one in six living with HIV was infected through sex between men. While there is an increasing trend of infection among MSM, heterosexual sex takes the lead in transmission. In 2011, 52% of individuals newly infected with HIV reported heterosexual sex as the transmission route, and the prevalence among sex workers is very low (0.3%). However, since a substantial proportion of MSM are married or engage in intercourse with women, researchers reiterate that MSM/bisexual men can act as a bridge for HIV transmission to the low-risk population.

The participation of China in international health organizations requires regular reporting on the knowledge and spread of AIDS and other sexually transmitted diseases (STDs). This has led authorities, however reluctantly at first, to enlist the help of activists and nongovernmental organizations in collecting data about the (male) homosexual population. Government funds are granted to activists and academics for education, outreach, and research. While HIV-related organizations that serve MSM are flourishing, lesbian organizations are relatively invisible and limited by the lack of resources. Although lesbians also need to address the issues that threaten their health and well-being, their concerns are often overlooked by the government. Sometimes they share resources earned by MSM-related organizations to carry out their work. The state has thus unintentionally created queer-friendly institutions that provide support for sexual minorities (Zhou, 2009).

Sexual Identity, Cultural and Social Belonging

With the advent of the Internet, hundreds of lesbian and gay Web sites have emerged in China. Sexual minorities feel at home and that they are the 'natural' majority when they enter queer cyberspace. The cyberspace is particularly important for lesbians since traditional gay venues are highly commercialized and largely catered for gay men who have consumption power. The Chinese cyberspace enables lesbians to develop social networks and deal with real life problems associated with their sexuality (Martin, 2009). Although online commercialization is inevitable, which reinforces hedonist and individualist values (Ho, 2010), the Chinese cyberspace has enabled sexual minorities to express themselves and develop their sexual identities and communities (Wong, 2011).

The Internet makes connection possible between sexual minorities in China and their counterparts in other parts of the world. Although they may not be in actual contact with each other, some Chinese individuals have developed a sense of belonging in a gay globality (Martin, 2009). However, this does not imply a universal sense of sexual identity. On the one hand, self-conscious modern gay and lesbian individuals in China desire alliance with Western gay and lesbian networks and have subtly adopted Western identity politics. On the other hand, they reject Western dominance and seek to defend an authentic Chinese same-sex identity. The tensions of local/global, Chinese/West, and authenticity/modernity are played out in the paradoxical articulation of queer identity in China (Wong, 2007; Ho, 2010).

Sexual minorities in China call themselves *tongzhi* (literally, same will). 'Tong' (literally the 'same') refers to sharing the same political goal of subverting heterosexism while hinting at same-sex orientation. In recent years, the category has expanded to include lesbian, gay, bisexual, transgender, and queer-identified people. The creation of this label represents an attempt to indigenize sexual identity and reclaim a Chinese cultural identity. *Tongzhi*, which means comrade, was used as a term of address in communist China. The term has fallen out of common usage because of its revolutionary connotation and is now mainly used within the official discourse of the Communist Party. The appropriation by Chinese sexual minorities means a parodic political subversion (Bao, 2011).

The identity of tongzhi serves to connect sexual minorities in China, Hong Kong, Taiwan, and other diasporic communities. They have built a larger community of sexual minorities through networking on Chinese-language Web sites and the circulation of books and magazines. In the First and Second Chinese Tongzhi Conferences held in 1996 and 1998 in San Francisco and Hong Kong, respectively, the participants discussed issues that concerned ethnic Chinese sexual minorities. They asserted that confrontational politics in the West should be avoided as family relations and social harmony are of utmost concern to Chinese people. The coming home alternative was created as an indigenous model (Chou, 2000). Coming home emphasizes the process of introducing one's same-sex partner to family as a close friend, leaving the gay identity more unspoken. Parents may refer the partner as a second daughter or son, and the partner is integrated into the family.

In times of intense globalization, coming home and coming out can no longer be treated as opposing strategies. In 2005, Hong Kong activists joined the call to mark the International Day Against Homophobia and organized a rally. Three years later in 2008, the first pride parade was held in Hong Kong, and attracted hundreds of participants from the rest of China, Taiwan, and other regions. Even in these coming out events, the gaining of family acceptance continues to be a major theme, thus reflecting the cultural fusion of the coming out and coming home models (Wong, 2007). Shanghai followed suit in 2009 and held its pride festival every year but without a march, as demonstrations are heavily cracked down in China. In 2013, the first lesbian, gay, bisexual, and transgender (LGBT) march took place in more than 10 cities, although some organizers were arrested and detained.

A slogan in the march in Changsha stated: "No more homophobia, the world belongs to everyone" (*buzai kong tong; shijie datong*) and appeals to a Chinese utopia found in the Confucian Classic of Rites, *datong* (literally great togetherness), which has long influenced Chinese reformers and revolutionaries. The use of the notion of datong enables sexual minorities to claim cultural citizenship, the belonging to the Chinese culture. In China where queer desire has long been seen as either nonexistent or a foreign decadence, sexual minorities

reconstruct indigenous, historical accounts of same-sex relationships. Stories such as 'The Cut Sleeve (*Duan Xiu*),' which is about a Chinese emperor who cut his sleeve because he did not want to wake his male companion who was sleeping on it, and 'The Bitten Peach (*Fen Tao*),' which tells of an ancient ruler of a state who shared a bitten peach with his minister, are often cited as evidence that homosexuality was widely practised in China for thousands of years and had enjoyed wide acceptance. Female tongzhi communities circulated the anecdote about a group of women called rubbing mirror party (*mojing dang*), who were known for same-sex sexual practices and community bonding in late nineteenth-century Shanghai.

Today, *tongzhi* remains a popular identity label in China, but as a result of growing exposure to the transnational gay scenes, some individuals have come to adopt the westernized identity of gay. Some identify themselves in this way: "wo shi ge gay," which means, "I am a gay." Capable of using a mix of Putonghua and English, they are usually university educated, financially stable, middle class, and Internet users. They found that the English word 'gay' denotes happiness and is not as politicized and blunt as the term tongzhi. The use of a foreign term has been associated with an oblique, stylish, and westernized presentation of the self (Wong, 2011). Lesbians in China usually call themselves *lazi* or *lala*. These terms are abbreviations of the transliteration of 'lesbian,' which mark a hybridized discourse of sexual identities. There are still others who dislike the term tongzhi for being too politicized and the term gay for being too foreign. Some men describe themselves as *quannei*, that is, someone in the (tongzhi) circle. This enables them to build connections with each other and expand their social network without politicization of sexual identity.

Thirty years of market reform have increased the visibility of sexual minorities and opened space for diverse ways of negotiating sexual identity and cultural norms. Some people frequent gay venues and accumulate a collective identity as members of a certain homosexual group (*quannei*). In addition to the opportunities for cruising, these gay venues enable friendship development and a sense of social belonging. Others articulate a strong sense of sexual identity and sexual right. They resist heterosexual marriage and pursue an independent homosexual life. They call themselves 'pure gay' or 'pure tongzhi' to markedly distant certain sexual lifestyles, such as bisexuality or cohabitation with a woman (Sun et al., 2006). There are still others who adopt a heterosexual identity even though they enjoy same-sex encounters. They believe that heterosexual marriage is a social obligation, and avoids emotional attachment with same-sex partners. They perform only sexual penetration in same-sex activities, thus concurring to the desired gender role that informs heterosexual identity (Li et al., 2010). Sexual identity in contemporary China is fluid and fragmented, and it is not uncommon for individuals to switch back and forth from a heterosexual to homosexual identity depending on the context (Li et al., 2010).

Quality Citizens, Internal Divisions, and Transgender People

The emphasis on the cultivation of quality citizen–subjects has created division among tongzhi even though queer life in China includes queers from all over the world. The concept of quality (*suzhi*) originated in population control, which aimed to reduce population quantity and raise population quality. Quality refers to the physical, mental, and moral attributes of the citizen–subject that are necessary to ensure national survival. Quality has thus become a widely circulated currency in everyday language, which brings together notions of modernity, knowledge, hygiene, urbanity, wealth, and taste. It is used to exclude rural gay men, some of whom make their living through sex work and are called money boys. Money boys are a triply stigmatized identity, as a rural-to-urban migrant, a prostitute, and a homosexual (Kong, 2010). Quality tongzhi chastises money boys as outsiders who are uncivilized, dirty, and carry greater risk of HIV infection, which spreads to the urban community because of their greed and ignorance. They are considered the main culprits for shaming the entire same-sex community (Jones, 2007; Ho, 2008).

The concept of quality is also used by MSM as a tool to assess the risk of potential sexual partners, such as deception, theft and blackmail, and contracting HIV and STDs (Jones, 2007). Characteristics such as cleanliness, stability, and knowledge are associated with quality and low risk. Those who have attained a higher educational level and who have married are more desirable. Whereas in the West, desirability is granted to those who have come out in public, in China, it is more likely to be granted to those who enact the tongzhi identity privately, but publicly enact the identities of a good family man and a respectable citizen. Those who are unmarried and fully integrated into communities of practice of gay men are perceived to be promiscuous. The concept of quality encourages gay men to lead a double life and serves as a powerful normalizing force within the gay community.

Since China has not yet developed a full-fledged transgender community, transgender people are one of the most misunderstood minorities. Transgenders are still called renyao (literally, human monster) in everyday language. They are rejected by families and the larger society. Men who cross-dress are sometimes featured in shows and performances hosted in gay and lesbian bars. Yet, they feel marginalized and find it difficult to integrate into the tongzhi community. Pi Jun, in his autobiographical essay (2010), shared the experience of exclusion by both lesbians and transsexuals. Jun is a female-to-male transgender who has had a bilateral mastectomy. Since transgender people have a much lower visibility compared with lesbians and gays, and few related Web sites can be found, transgenders mingle with the lesbian groups. However, in recent years, some lesbian groups have refused to accept stone butch or female-to-male transgenders as lesbians. Pun felt that with a male self-identification, he did not belong in the two-step identification of lesbians in which first, I am a woman and second, I love woman. When he came into contact with other female-to-male transsexuals, he again felt excluded, as the transsexuals who underwent male reassignment surgery refused to recognize him as a real man.

The Chinese government claims a progressive attitude toward transgenders since policies have been adopted to grant them civil rights, allow them to change their identification cards, and legally recognize their marriages after sex reassignment surgery. However, Chinese transgenders have a much harder time in gaining social acceptance and personal freedom

than other sexual minorities. Unlike the official depathologization of homosexuality in 2001, gender dysphoria is considered to be first and foremost a mental illness. In 2009, the Chinese government published fairly stringent regulations for anyone who hopes to have sex reassignment surgery (http://shardsofchina.wordpress.com/2012/02/08/sex-in-china-the-third-sex-transgender-people-struggle/). For instance, they are required to live as member of the opposite sex for 5 years and undertake therapy for at least 1 year before such surgery can be considered. To protect marriage and family, the government refuses to grant married persons the right to sex change. Permission from their local police bureau must be obtained in order to confirm that they have no objections to the reassignment surgery and will issue an ID card to the individual for their new gender. These form multiple barriers for enforcing constraints over sex change. Even if successful, life afterward can still be unbearable since the gender on academic certificates and work permits has yet to change. Since the first sex change operation was carried out in 1983, only 2000 Chinese people have been able to become fully transsexual. For ordinary people, a successful transition is a next to impossible struggle, especially when the cost of surgery is estimated to be around 200 000 RMB (32 000 USD).

Conclusion

Unlike their Western counterparts who can live relatively openly as gay, encountering at worst religious opposition and chance of homophobic harassment, Chinese sexual minorities are subject to the imperative of heterosexual marriage. Sexual minorities in China have begun to call for the right to wed, which has met with some sympathy. Li Yinghe, an academic at the prestigious Chinese Academy of Social Sciences, has repeatedly proposed the legalization of same-sex marriages in the national committee of Chinese People's Political Consultative Conference. A famous Beijing activist–film director, Chun Tze Yan, argued that unlike Western individuals who do not hold marriage in high regard and frequently divorce, the Chinese value family and marriage, and that is why there is a strong and practical demand for same-sex marriage. The dual emphasis on gay rights and family values marks the specific character of tongzhi identity and politics.

Factors such as the marriage imperative, normative family obligations, desired gender roles, and the need for social and cultural belonging will continue to shape the formation of sexual identity and politics in China. What complicates the scene is the hegemonic discourse of quality, which causes division among tongzhi along urban/rural and class lines. The discourse of quality creates a system of hierarchy and domination grounded on distinctions between the respectable and the degenerate. Quality tongzhi are those who privately enact the tongzhi identity, but publicly perform the identities of a respectable citizen. Close integration with the gay community is associated with promiscuity, which would exclude them from consideration for cooperative marriages. To address internal divisions and exclusion will require the broadening of the queer political spectrum to more than just a fight for same-sex marriage.

See also: Heterosexism and Homophobia; Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex; Medicine and Health for Sexual Minorities; Sexual Minorities and Sexual Citizenship; Sexual Minority Families of Choice and Family Networks: We Are Family; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Bao, H., 2011. 'Queer comrades': transnational popular culture, queer sociality, and socialist legacy. English Language Notes 49 (1), 131–137.
Chen, Y., 2009. Nu tongxinglian de hunyin he jiating gei chuantong hunyin zhidu dai lai de tiaozhan (Lesbian women's marriages and families: a challenge to the traditional marriage system). Shehui (Society) 4, 107–129.
Chou, W., 2000. Tongzhi: Politics of Same-sex Eroticism in Chinese Societies. Haworth Press, New York.
D'Emilio, J., 1980/1993. Capitalism and gay identity. In: Abelove, et al. (Eds.), The Lesbian and Gay Studies Reader. Routledge, New York, pp. 467–476.
Ho, L.W.W., 2008. Speaking of Same-Sex Subjects in China. Asian Studies Review 32 (4), 491–509.
Ho, L.W.W., 2010. Gay and Lesbian Subculture in Urban China. Routledge, London and New York.
Jones, R.H., 2007. Imagined comrades and imaginary protections: identity, community and sexual risk among men who have sex with men in China. Journal of Homosexuality 53 (3), 83–115.
Jun, P., 2010. Transgender in China. Journal of LGBT Youth 7 (4), 346–358.
Kam, L.Y.L., 2013. Shanghai Lalas: Female Tongzhi Communities and Politics in Urban China. Hong Kong University Press, Hong Kong.
Kang, W., 2012. The decriminalization and depathologization of homosexuality in China. In: Weston, T.B., Jensen, L.M. (Eds.), China in and beyond the Headlines. Rowman & Littlefield, Lanham, MD, pp. 231–248.
Kong, T., 2010. Outcast Bodies: Money, sex and Desire of Money Boys in Mainland China. In: Yau, C. (Ed.), Bending Lives Negotiating Sexualities in Hong Kong and China. Hong Kong University Press, Hong Kong, pp. 17–35.
Li, H., Holroyd, E., Lau, J.T.F., 2010. Negotiating homosexual identities: the experiences of men who have sex with men in Guangzhou. Culture, Health & Sexuality 12 (4), 401–414.
Liu, J.X., Choi, K., 2006. Experiences of social discrimination among men who have sex with men in Shanghai, China. AIDS and Behavior 10 (Suppl. 4), S25–S33.
Liu, D., Lu, L., 2005. Zhongguo Tong Xing Lian Yanjiu. (Study of Homosexuality in China). Zhongghuo shehui chubanshe, Beijing.
Liu, H., Feng, T., Ha, T., Cai, Y., Liu, X., Li, J., 2011. Chinese culture, homosexuality stigma, social support and condom use: a path analytic model. Stigma Research and Action 1 (1), 27–35.
Martin, F., 2009. That global feeling: sexual subjectivities and imagined geographies in Chinese-language lesbian cyberspaces. In: McLelland, M., Goggin, G. (Eds.), Internationalizing Internet Studies. Routledge, New York.
Pan, S., Yang, R., 2004. Xing'ai shinian: quanguo daxuesheng xingxingwei de zhuizong diaocha, (Sexuality of Chinese College Students: A Ten-Year Longitude Nationwide Random Study). Social Sciences Documentation Publishing House, Beijing.
Pan, S., Wu, Z., Gil, V.E., 1995. Homosexual behaviors in contemporary China. Journal of Psychology & Human Sexuality 7 (4), 1–17.
Parish, W.L., Laumann, E.O., Mojola, S.A., 2007. Sexual behavior in China: trends and comparisons. Population and Development Review 33 (4), 729–756.
Sun, Z., Farrer, J., Choi, K., 2006. Sexual identity among men who have sex with men in Shanghai. China Perspectives 64, 2–12.
Weeks, J., 1977. Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present. Quartet Books, London and New York.
Wong, D., 2007. Rethinking the coming home alternative: hybridization and coming out politics in Hong Kong's anti-homophobia parades. Inter-Asia Cultural Studies 8 (4), 600–616.
Wong, D., 2011. Hybridization and the emergence of 'gay' identities in Hong Kong and in China. Visual Anthropology 24 (1/2), 152–170.
Wu, J., 2003. From 'Long Yang' and 'Dui Shi' to Tongzhi: homosexuality in China. Journal of Gay and Lesbian Psychotherapy 7 (1), 117–143.
Zhang, B.C., Chu, Q.S., 2005. MSM and HIV/AIDS in China. Cell Research 15, 858–864.
Zhang, B., et al., 2000. A survey of men who have sex with men: mainland China. American Journal of Public Health 90, 1949–1950.
Zhou, K., 2009. China's Long March to Freedom: Grassroot Modernization. Transaction Publishers, New Brunswick, NJ.

Sexual Minorities in Modern Russia

Brian J Baer, Kent State University, Kent, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the history of sexual minorities in modern Russia, beginning with a discussion of the methodological challenges posed by the interpretation of Western sources, on the one hand, and of Russia's discursive silence, on the other. The history of Russia's sexual minorities is divided into three historical periods: prerevolutionary Russia, Soviet Russia, and post-Soviet Russia. Special attention is paid to the differences between literary and other representations of homosexuality and the descriptions of gay and lesbian life provided in the diaries of Russian gays and lesbians, revealing a deep-seated ambivalence regarding the phenomenon of homosexuality in Russia.

Introduction

Writing the history of Russia's sexual minorities is a particularly challenging task given the fact that so many of the historical accounts were written by Western observers and that these accounts are highly subjective, shaped by the mutually defining nature of the relationship of Russia and the West that emerged in the sixteenth century, a relationship based on relatively limited contact. As Martin Malia comments, "Russia has at different times been demonized or divinized by Western opinion less because of her real role in Europe than because of the fears and frustrations, or the hopes and aspirations, generated within European society by its own domestic problems" (Malia, 1999: p. 8). Native Russian sources, on the other hands, are, in the words of Lloyd Berry and Robert Crummey, 'relatively laconic' (1968: p. xiv), and this is especially true in regard to sexual matters. Indeed, with the exception of the opening and closing decades of the twentieth century, Russians have maintained a discursive silence on the topic, which stands in stark contrast to the voluminous discourse on sex and sexuality generated in the West, described by Foucault in his 'repressive hypothesis.' That silence was reflected not only in the suppression and erasure of gay and lesbian voices but also in the absence of research into the lives of gays and lesbians in Russia that has left the topic of homosexuality in Russia especially vulnerable to Western interpretations that perhaps say more about the Western cultures that have generated them than they do about Russia itself. The following section provides a brief overview of the most influential of these Western interpretations.

The Western Gaze

Western accounts of sexual life in Russia from the Renaissance on display a marked tendency to view Russians as sexual libertines, which betrays an Orientalizing gaze that links an assumption of unfettered sexual behavior to economic and political underdevelopment and political autocracy. Some of the earliest accounts fall within this tradition. First-person accounts in the sixteenth century by Sigismund von Herberstein, ambassador from the German Empire to the court of Ivan IV, and by George Turberville, secretary to the English ambassador to Moscow, as well as the seventeenth-century accounts by Adam Olearius, a member of an embassy from the Duke of Holstein to Muscovy and Persia, insist that sodomy was a widespread practice among the Russians, and Olearius notes, "not only with boys but also with men and horses" (1967: p. 142), and that Russians bore little shame in regard to it. In fact, Olearius notes that these sexual escapades "provide matter for conversation at their carouses" (1967: p. 142). Turberville in a poem dedicated to his 'special friend Master Edward Dancie' describes the Russians as 'a people passing rude, to vices vile inclined,' and says the Russian peasant "will lead a bogard's (bugger's) life. The monster more desires a boy within his bed/Than any wench, such filthy sinne ensues a drunken head" (Cross, 1971: pp. 70–71).

Another important episode in the sexualization (and homosexualization) of Russia occurred in the context of the emerging science of psychoanalysis. Some of Freud's most famous clients were Russians: Lou Andreas-Salomé and Sergei Pankov. Freud diagnosed Pankov, better known as the Wolfman, as a repressed homosexual. Aleksandr Etkind argues that Freud viewed his Russian patients through the lens of contemporary Western stereotypes of Russians: "Sexual freedom was ascribed to Russians" (Etkind, 1997: p. 111). But Freud's patients were not the only source of this view. In 1912, a book was published in France *Confession sexuelle d'un anonyme russe* (The Sexual Confession of an Anonymous Russian), which was a first-person narrative of a Russian aristocrat who resorts to psychoanalysis in his struggle against pedophilic urges; and the rumors surrounding the tsarina's favorite, Grigorii Rasputin, were also of a decidedly sexual nature.

For many European leftists, Soviet communism promised to free individuals from the shackles of Bourgeois prudery and repression. The Austrian psychoanalyst William Reich and the Swiss journalist Ella Maillart saw early Soviet Russia as a kind of sexual paradise, where young people engaged in sexual relations naturally, with a total absence of shame. Reich later changed his opinion and in 1936 published an indictment of the repression of Soviet homosexuals. Similarly, André Gide, who traveled to Soviet Russia in 1934, was quickly disabused of any notion he might have had of Russia as a place of sexual freedom for homosexuals. He documented his disenchantment in his best-selling *Return from the USSR*. Nevertheless, the notion of Russia as sex friendly was picked up again during the Cold War by the historians John Lauritsen and David Thorstad (1974) who argued that the absence of legislation

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10218-1

criminalizing homosexuality in the early Soviet period suggested a tolerance for homosexuality. Simon Karlinsky mounted a vociferous attack against this interpretation of Soviet history in the San Francisco–based *Gay Sunshine Press*, tracing the homophobia of the Soviet regime to Lenin himself. Karlinsky offered an alternative sexual history of Russia, presenting prerevolutionary Russia as a kind of 'golden age' for gays and lesbians. In his zeal, however, Karlinsky made unreflective use of some dubious historical sources, such as Rozanov's *People of the Moonlight*, to argue for the homosexuality of medieval Orthodox saints, such as Boris and Gleb, and Moses of Hungary (1976: p. 1). The association of Russia with sexual freedom was resurrected once again in the late Soviet and early post-Soviet period by Western observers, such as American anthropologist Laurie Essig, American journalist David Tuller, and British novelist Duncan Fallowell, who found in Russia a sexually liberating, queer alternative to the West with its rigid conception of sexual identity (see Baer, 2002).

The opening of archives and the lifting of censorship restrictions, however, have allowed for a more nuanced, evidence-based view of homosexuality in nineteenth and twentieth century Russia.

Homosexuality in Prerevolutionary Russia

It does appear that homosexuality was less rigorously prosecuted in Russia than in the West, at least until the Soviet period. A general attitude of tolerance toward homosexuality appeared to be the norm among the highest echelons of Russian society throughout most of the nineteenth century, as evidenced in Alexander Pushkin's adaptations of ancient Greek and contemporary Arabic poems with homoerotic themes (see Green, 1997) and in Mikhail Lermontov's bawdy Hussar poems. Mikhail Kuzmin's diaries, published only after the fall of the Soviet Union, paint a picture of a thriving gay subculture, and recent research has put an end to the rumors that the composer Petr Chaikovskii committed suicide over a sex scandal with a young boy (Poznansky, 1996). It is now clear that Chaikovskii and his gay contemporaries, including many prominent members of Russia's ruling class, led active and quasi-open gay lives. Nevertheless, Nikolai Gogol's tortured relationship to his homosexuality, which may have led to his early death, as well as the anguished diary entries of Grand Duke Constantine Romanov confessing his extramarital trysts with bathhouse attendants and Chaikovskii's sorrowful letter to his brother about his homosexuality and his decision to marry suggest that there were still powerful religious and social strictures in place against the open practice of homosexuality.

By the end of the nineteenth century, rapid industrialization had moved masses of Russians from the countryside to the cities where they lived beyond the supervision of their families and neighbors; in turn, boarding houses for young, single workers appeared in these cities, and features of a gay subculture began to emerge (see Healey, 2001: pp. 29–49). The Russian *banya* or 'bathhouse' and parks became cruising grounds where the upper and lower classes mingled. Many prominent members of Russian high society, including several Grand Dukes, carried on rather open homosexual liaisons and avoided prosecution. The explorer Nikolai Preobrazhenskii was also a very public figure who led a homosexual lifestyle. The gay and lesbian subculture of the time, as exemplified by the poetry of Aleksei Apukhtin, was characterized by high asceticism and a poetics of euphemism and coded references (Taylor, 2002: p. 12). It was at this time, however, that homosexuality began to enter the public sphere. For example, Vladimir Dmitrievich Nabokov, a prominent public figure in the last years of the empire and father of the writer Vladimir Nabokov, took a public stand against the criminalization of homosexuality on the grounds that the law should not interfere in a person's private life (Engelstein, 1992: pp. 67–70).

In 1905, political and social unrest, expressed in workers' strikes, military mutinies, and peasant protests, swept across the Russian empire. To end the turbulence, a reluctant czar was forced to make concessions that resulted in the Russian Constitution of 1906, which for the first time in Russian history granted Russians freedom of speech and assembly. Restrictive censorship laws were lifted and questions of gender, sex, and sexuality were for the first time in Russian history discussed with relative freedom in the public sphere (see Engelstein, 1992). Homosexuality attained unprecedented visibility in this period. In 1906, the Russian poet and prose writer Mikhail Kuzmin published his novel *Wings*, which recounts the main character Ivan (Vanya) Smurov's symbolic journey toward acceptance of his homosexuality. A success-de-scandale in its time, Kuzmin's novel was only one of many contemporary novels to offer new, frank portrayals of sexual desire. Kuzmin's reputation in Russia, however, is more as a poet than as a prose writer. His poetry, inspired by his homosexual affairs, is marked by a formal sophistication combined with unabashed sensuality, as in the poetic cycle 'The Love of This Summer' (1906), considered by critics to be Kuzmin's artistic credo.

Kuzmin's diaries from 1905 to 1907 reveal a homosexual culture with blurry edges at a time when sexual experimentation was the order of the day. One of Russia's leading Symbolist poets, Viacheslav Ivanov, together with his wife, Lydia Zinovieva-Annibal, hosted a weekly artistic salon in their apartment in a building that came to be known as 'the tower.' Open-minded in matters of sex and sexuality, they were close friends with Kuzmin, who was a habitué of their salon and lived for several years in the same building (Malmsted and Bogomolov, 1999: p. 129). A bisexual, Ivanov produced several poems about his homosexual experiences, which were published in his 1911 book of verse, *Cors Ardens*, under the title 'Eros.' Zinovieva-Annibal made her own contribution to Russian gay literature with her novel *Thirty-Three Freaks* (1907) and her collection of stories *The Tragic Zoo* (1907). But as the titles suggest, these works presented lesbianism in an unflattering light as decadent and perverse. Equally unflattering was the depiction of homosexuality in Fedor Sologub's novel *Petty Demon* (1907), which features a sadistic schoolmaster who sublimates his pederastic urges by whipping the young boys in his charge. The gay writer Riurik Ivniev's collection of poetry *Self-Immolation* (1913) also presents homosexual desire in a perverse, decadent light.

The most sustained treatment of homosexuality in Russia until recently was offered by the writer and philosopher Vasilii Rozanov in a monograph devoted exclusively to homosexuality, entitled *People of the Moonlight* (1911). In this work,

Rozanov exposed and condemned the sexophobia of Russian Orthodox culture. This disgust with sex, Rozanov argues, produced a loss of virility that is embodied in 'spiritual homosexuals.' These are men defined not by their attraction to other men but rather by their lack of attraction to women. While he acknowledges the existence of 'real' homosexuals, most homosexuals he suggests are the products of Russia's sexophobic religious culture. For Rozanov, spiritual homosexuals were a symptom of the decadence of Russian civilization, but the asexuality of Rozanov's homosexuals bore little resemblance to the actual life of gays and lesbians living in Russia at the time, as revealed in the diaries of Konstantin Romanov and Petr Chaikovskii. Two of the most colorful figures of the Silver Age period were Nikolai Kliuev and Zinaida Gippius. Married to the philosopher and writer Dmitrii Merezhkovskii, Gippius was a true gender bender, often dressing as a man and adopting a male persona in some of her lyric poems, perhaps to mask her lesbian desire (Burgin, 1993: p. 85). Gippius' interest in homosexuality was not only personal, but, like that of other members of her milieu, had a philosophical aspect, representing the transcendence of traditional, biologically defined gender identities. As her contemporary Ivanov put it in his diary, "Homosexuality is inseparably connected with humanism" (Malmsted and Bogomolov, 1999: p. 105). Unlike the cosmopolitan Gippius, Kliuev came from the provinces from a family that belonged to the Khlysty sect. Even when he moved to St. Petersburg he continued to dress in peasant garb. Kliuev was one of several 'peasant' poets of the time, but his work was also influenced by the elite Symbolist poets. He was the friend, mentor, and purported lover of Sergei Esenin, who dedicated poems to Kliuev (Chernetsky, 1997: p. 153).

Homosexuality in Soviet Russia

Despite the abovementioned silence of early Soviet law on the subject of homosexuality, the regime never embraced the cause of homosexual liberation, and by the early 1930s, life had become difficult for the gays and lesbians who remained in Soviet Russia. Both Kliuev and Esenin died tragic deaths in the Soviet Union. Esenin committed suicide in 1925, leaving a note written in his own blood. Kliuev was exiled to Siberia where he eventually died. All of Kliuev's unpublished writings, which he had left with an ex-lover Nikolai Arkhipov, were lost when Arkhipov himself was arrested and sent to the gulag (Karlinsky, 1997: p. 151). Kuzmin miraculously avoided prison, dying a natural death in 1936. The members of his circle of intimates, however, did not fare as well. His long-time lover Iurii Yurkun was arrested in 1938 and executed, and the translator Ivan Likhachev, who had been a pall bearer at Kuzmin's funeral, was also arrested and spent many years in the camps, where he translated Baudelaire's *Les fleurs du mal* from memory (Etkind, 1997: pp. 47–48).

Those who were not imprisoned were silenced. The last hurrah of gay-themed literature in Soviet Russia was the miraculous publication in 1929 of Kuzmin's collection *The Trout Breaks the Ice*. This was 1 year after the publication of Konstantin Vaginov's novel *Goat's Song* (1928), which featured the gay character Kostya Rotikov. The lesbian poet Sophia Parnok also found it increasingly difficult to publish her work in the 1920s. Marina Tsvetaeva's poem 'Podruga' (Girlfriend), inspired by her affair with Parnok, was not published until 1979, in the United States. The poet Riurik Ivniev (pseudonym of Mikhail Kovalev), too, ceased publishing homoerotic verse in the Soviet period, and his prerevolutionary poetry was only republished in 1988 (Molnar, 1997: p. 159).

Stalin's consolidation of power in the late 1920s resulted in stricter regulation of what was considered to be 'nonnormative' behavior, and in 1934 homosexuality was criminalized at the same time that abortion was made illegal. But homosexuality had already disappeared from Russian literature by this time. It would also disappear from the Soviet press and as an object of academic study. Kuzmin's final public reading of his poetry in Leningrad in 1928 was perhaps the "last demonstration of Leningrad's homosexuals" (Malmsted and Bogomolov, 1999: p. 349).

While homosexuality disappeared from the public sphere in the Soviet era, the diaries of gays and lesbians who remained in Russia – such as the poets Mikhail Kuzmin and Riurik Ivnev, and the cabaret singer Vadim Kozin – offer an important source of information about the private lives of Russia's sexual minorities. Their diaries document the harassment and everyday humiliations endured by homosexuals at the hands of the regime. (Kuzmin's diary entries of 1905–07 are very different from his entries of 1934.) Many gay, lesbian, and bisexual Russians fled the Soviet Union after the October Revolution and took up residence in various metropolitan centers around the world. Marina Tsvetaeva emigrated to Paris, as did the poet Anatolii Steiger; the poet Valery Pereleshin emigrated first to China, then to Brazil, where he settled. Although the Russian diaspora was only nominally less sexophobic than the Soviet regime, these poets were able to continue working and to find venues in which to publish their work. Pereleshin's magnum opus *Ariel* (1976), which consists of 169 classical sonnets documenting an epistolary romance Pereleshin had with a man in Moscow, was first published in Germany. Gennadii Shmakov, who emigrated to New York City, translated works of gay literature, such as James Baldwin's *Giovanni's Room* and the poems of Constantine Kavafy, while continuing to champion the work of Russia's first openly gay writer, Mikhail Kuzmin, who had been all but forgotten in Soviet Russia.

Homosexuality in Post-Soviet Russia

The breakup of the Soviet Union in the early 1990s was accompanied by the lifting of censorship restrictions, the decriminalization of homosexuality, and the influx of Western NGOs. These changes resulted in new visibility for Russia's sexual minorities and the promise of a lesbian, gay, bisexual, and transgender (LGBT) activist movement. All the trappings of an open gay and lesbian subculture appeared in Russia's largest cities: clubs, magazines, publishing houses, and activist organizations. Moreover, the founding of Russia's first gay Web site, gay.ru, gave provincial gay men and lesbians access to gay-themed books and journals. Although most of these venues were short lived and depended on financial support from the West, they nevertheless made possible the publication of works by new Russian gay and lesbian writers and of translations of Western gay literature, as well as the republication of works by

Soviet and prerevolutionary writers that had been censored by the regime. Some of the most striking events in post-Soviet publishing involved the publication of gay-themed works, such as Kuzmin's frank and witty diaries (1998, 2000) and Kostya Rotikov's gay 'history' of St. Petersburg, *Drugoi Peterburg* (The Other Petersburg, 1997). (Kostya Rotikov is the pseudonym of Iurii Piriutko.) Russia's first anthology of gay literature, *Liubov' bez granits* (Love without Borders, 1997) also appeared at this time; it featured the Russian writers Lermontov, Kuzmin, Zinovieva-Annibal, Ivnev, Esenin, Limonov, and Kharitonov, several of whom were not in fact gay, alongside canonical gay writers of the West, such as Proust and Michelangelo; the cover features a black-and-white reproduction of the Russian artist Aleksandr Ivanov's 1822 painting *Two Models*, portraying two provocatively posed male nudes. These works, which were lauded by some and violently denounced by others, represent early attempts by post-Soviet Russian gays and lesbians to reclaim, or queer, their Russian and Soviet past.

The work of several out gay men – Dmitrii Volchek, Aleksandr Shatalov, Dmitrii Kuzmin, and Iaroslav (Slava) Mogutin – in their work as poets, critics, editors, and publishers has done much to secure the availability of gay- and lesbian-themed literature in today's Russia and to battle long-standing stereotypes that associate homosexuality with asexuality and suffering. However, a deep ambivalence toward homosexuality persists within Russian culture today, reflected in the government's refusal to allow gay pride parades, in new laws to outlaw public discussion of homosexuality, described as 'propaganda,' and in high levels of homophobia. Vitaly Chernetsky describes gays and lesbians as "the most stigmatized and oppressed minority group in contemporary Russia" (2007: p. 146). This ambivalence is also reflected in sensational, often graphic depictions of homosexuality in the mainstream media and in the literary works of non-gay-identified writers designed to shock. In 2001, Elizabeth Wolfe, a reporter for the *Moscow Times*, declared the LGBT activist movement in Russia moribund (10). This is explained not only by opposition from an authoritarian government and an equally authoritarian official church, but by the persistence of Soviet-era sexophobia, manifested in a reluctance on the part of many Russians to discuss issues of sex and sexuality in public, as reflected in new legislation to outlaw 'homosexual propaganda.' All that, combined with the fact that many Russians – straight and gay alike – reject the Western civil rights model predicated on rigid and exclusive identities, makes the future of an open gay and lesbian subculture in Russia uncertain.

See also: Gay and Lesbian Movements; Heterosexism and Homophobia; Queer Geography; Sexual Orientation: Historical and Social Construction.

Bibliography

Baer, Brian James, 2002. Russian gays/Western gaze: mapping (homo)sexual desire in post-Soviet Russia. GLQ 8 (4), 499–520.

Berry, Lloyd, Crummey, Robert, 1968. Rude and Barbarous Kingdom: Russia in the Accounts of Sixteenth-Century Voyagers. University of Wisconsin Press, Madison, WI.

Burgin, Diana Lewis, 1993. Laid out in lavender. Perceptions of lesbian love in Russian literature and criticism of the Silver Age, 1893–1917. In: Costlow, Jane T., Sandler, Stephanie, Vowles, Judith (Eds.), Sexuality and the Body in Russian Culture. Stanford University Press, Palo Alto, pp. 177–203.

Chernetsky, Vitaly, 1997. Sergei Esenin (1895-1925). In: Moss, Kevin (Ed.), Out of the Blue: Russia's Hidden Gay Literature. An Anthology. Gay Sunshine Press, San Francisco, CA, p. 153.

Chernetsky, Vitaly, 2007. Mapping Postcommunist Cultures: Russia and Ukraine in the Context of Globalization. McGill-Queen's University Press, Montreal and Kingston.

Cross, Anthony, 1971. Russia under Western Eyes, 1517-1825. St. Martin's Press, New York.

Engelstein, Laura, 1992. The Keys to Happiness: Sex and the Search for Modernity in Fin-de-Siecle Russia. Cornell University Press, Ithaca, NY.

Etkind, Alexander, 1997. Eros of the Impossible. The History of Psychoanalysis in Russia. Westview Press, Boulder, CO.

Etkind, Efim, 1997. Russkaia perevodnaia poeziia XX veka (Russian translated poetry of the 20th century). In: Etkind, E. (Ed.), Mastera Poeticheskogo Perevoda (Masters of Poetic Translation). Akademicheskii Proekt, St. Petersburg, pp. 5–54.

Green, Michael, 1997. Alexander Pushkin. In: Moss, Kevin (Ed.), Out of the Blue: Russia's Hidden Gay Literature. An Anthology. Gay Sunshine Press, San Francisco, CA, p. 29.

Healey, Dan, 2001. Homosexual Desire in Revolutionary Russia. The Regulation of Sexual and Gender Dissent. University of Chicago Press, Chicago, IL.

Karlinsky, Simon, 1976. Russia's Gay Literature and History (11th-20th Centuries), vols. 29–30. Gay Sunshine (Summer/Fall), 1–7.

Karlinsky, Simon, 1977. Russia's Gay Literature. In: Part II: Controversy, vol. 31. Gay Sunshine (Winter), 26.

Karlinsky, Simon, April 1, 1982. Gay life before the Soviets: revisionism revisited. Advocate 339, 31–34.

Karlinsky, Simon, 1997. Nikolai Klyuev (1887-1937). In: Moss, Kevin (Ed.), Out of the Blue: Russia's Hidden Gay Literature. An Anthology. Gay Sunshine Press, San Francisco, CA, pp. 151–152.

Lauritsen, John, Thorstad, David, 1974. The Early Homosexual Rights Movement (1864-1935). Times Change Press, New York.

Lauritsen, John, Thorstad, David, 1977. Russia's Gay Literature. In: Part II: Controversy, vol. 31. Gay Sunshine (Winter), 25–26.

Malia, Martin, 1999. Russia under Western Eyes. Harvard University Press, Cambridge, MA.

Malmsted, John, Bogomolov, Nikolay, 1999. Mikhail Kuzmin. A Life in Art. Harvard University Press, Cambridge, MA.

Molnar, Michael, 1997. Ryurik Ivnev (1891-1981). In: Moss, Kevin (Ed.), Out of the Blue: Russia's Hidden Gay Literature. An Anthology. Gay Sunshine Press, San Francisco, CA, p. 159.

Moss, Kevin (Ed.), 1997. Out of the Blue: Russia's Hidden Gay Literature. An Anthology. Gay Sunshine Press, San Francisco, CA.

Olearius, Adam, 1967. The Travels of Olearius in 17th-Century Russia (S. Baron, Ed. and Trans.). Stanford University Press, Stanford.

Poznansky, Alexander, 1996. Tchaikovsky's Last Days. A Documentary Study. Clarendon Press, Oxford.

Taylor, Philip, 2002. Foreward. In Three Tales. Aleksei Apukhtin (P. Taylor, Ed. and Trans.). Associated University Presses, Cranbury, NJ, 7–15.

Wolfe, Elizabeth, 2001. Gays gather quietly, out of the political spotlight. Moscow Times (March 27), 10.

Sexual Minorities in South Africa

Zethu Matebeni, University of Cape Town, Rondebosch, Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the demise of apartheid, gay and lesbian people in South Africa have enjoyed equal rights and access to legal reform on many levels. Same-sex partnerships, adoption rights, immigration rights, and other social benefits for same-sex partners have become a norm in South Africa. Though the postapartheid South African constitution grants rights to sexual minorities, the extent of racial and class inequalities within gay politics indicates the need for a revamped movement or politics that will challenge structural injustices and lack of redistribution. Looking at South Africa through the commonly used gay lens is not only limited but also problematic. This lens follows the same forms of exclusion that gay history has revealed. Gender, class, and racial tensions within postapartheid gay politics, as in South Africa as a whole, are easily overlooked when the focus is primarily on freedom and equality.

Historical Roots of Lesbian and Gay Activism in South Africa

Since the 1950s, South Africa has seen an upsurge/interest in gay, lesbian, bisexual, and recently transgender and intersex (LGBTI) issues. Already it must be stated that the priority has been mainly or solely on gay and lesbian issues, with specific attention on white gays and lesbians. Laws and regulations have historically governed sexual and gender minority issues and continue to do so. The Prohibition of Disguises Act 16 of 1969, the Immorality Amendment Act 57 of 1969, the Immorality Amendment Act 2 of 1988, the Alteration of Sex Description and Sex Status, 2003, and others have all been designed to govern genders and sexualities. Alongside there have been social attitudes viewing mainly gays and lesbians as immoral and outcasts. However, while the mode of governing sexual minorities has changed, that is, its legal approach no longer reliant on punishment and exclusion (Cameron, 1995), what remains the same are the stark disparities among sexual minorities themselves.

Cameron (1995: p. 93) argues that the "traditional attitude of intolerance towards gay sexual conduct seems to be deeply ingrained in our legal history." Prior to the end of apartheid, laws governing sexuality prohibited any form of sexual contact between people of the same sex. Such contact was not only outlawed but also regarded as 'unnatural.' Society responded accordingly and stigmatized those involved in such acts branding them as criminals and giving them pariah status. Therefore, we cannot divorce or separate legal struggles from social struggles. These inevitably affect one another. However, what the current South African picture on sexual minorities shows is the disconnect between legal reform and sociopolitical and economic reform with the latter lagging behind the advancement of legal reform. While not difficult to understand why the disparities exist, it is curious that groups who have fought against domination and oppression have remained silent, and most times complicit, on these issues.

The South African Constitution of 1996 is exemplar both for the nation and for the world. It promises and offers rights and freedoms to everyone, regardless of their social category. Accompanying the constitution are significant court judgments and laws that have significantly improved the lives of gay people. A year after the adoption of the constitution, the Cape High Court ruled that the crime of sodomy was unconstitutional. By 1998 sodomy was decriminalized. Many of the subsequent laws, although accessible to all, were a result of legal challenges brought by mostly white gay and lesbian couples. Except for a few, the legal victories focused on economic access or advancement of same-sex couples.

A number of legal victories followed (Judge et al., 2008: pp. 55–57) in 1998 with the court ordering that a lesbian partner could receive benefits from her partner's state medical scheme. In 1999, same-sex life partnerships were allowed to receive the same pension benefits as heterosexual married partners. That same year, the Constitutional Court ruled, "the permanent same-sex life partner of a South African citizen or permanent resident should be granted the same rights as a spouse when it comes to immigration rights." By 2002, same-sex partners could claim equal benefits from the pension funds of their state-employed partners. They could also be joint legal parents of a minor adopted child. A year later, both same-sex partners were allowed to be registered as legal parents of a child born *in vitro* by one partner. The following year a surviving partner of a deceased same-sex person was granted the right to claim damages for loss of support. By 2006, the Constitutional Court confirmed that this would be extended to equal benefits in the case of a partner dying without a will. Within a couple of years, the common-law definition of marriage was declared unconstitutional. This led to the adoption of the Civil Union Act of 2006.

South Africa's Bill of Rights is the world's flagship in the protection and promotion of gay rights. It has made South Africans believe that there are intrinsic commonalities shared by those regarded as 'sexual minorities,' i.e., gays and lesbians. However, in the context of the current postapartheid state, those same groups of people find themselves divided along race, class, gender, and sexuality. Even their commonality as sexual minorities becomes dubious.

A Fractured Movement: The Early Years of Gay Politics in South Africa

From the onset, stark divisions marked the traces of the emergence of a gay and lesbian movement in South Africa.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10245-4

These were on many levels, in relation to politics, race, class, and gender. It appeared quite early on that white gay men would be the major beneficiaries of these divisions. During apartheid, the first signs of a gay movement appeared under the Legal Reform Fund set up in 1966 to oppose antihomosexual legislation in the form of the Immorality Amendment Bill. The fund was established as direct action following a police raid at a famous party in Forest Town, one of Johannesburg's affluent suburbs. Money raised by the fund would go toward legal representation established to investigate the proposed Bill. Although there were a number of lesbians and transsexual people at the party, it has been believed as a gay man's party. Gevisser (1995) calls this the closest equivalent to New York's Stonewall.

Perhaps what likens the subsequent political movements is that both were exclusively white, and both were largely silent about the visibility and role of lesbians, nonwhites, and trans people. This invisibility has remained within what is considered the LGBT movement and activism in South Africa. In reality, transgender and intersex people remain outside the LGBT framework and organize independently. Writing about Stonewall, Christian Williams (2013) states that it has come to be believed that "the middle class white gay men were the real heroes" and thus enjoy the benefits of equality while it was trans people who put their lives at risk. Similarly, the events of the Forest Town raid have been perceived as the resistance of white gay men while trans people and cross dresses who were most visible were arrested for 'masquerading' (Gevisser and Matebeni, 2008).

What was rather curious about the Forest Town party were the events after the arrest of many of Johannesburg's middle and upper class white gays. These events have come to symbolize the fractured state of gay, as well as LGBT politics in South Africa today. The Legal Reform Movement (LRM), set up through the fund, was careful not to align itself with Black antiapartheid movements of the time, as it feared antagonizing the state (Gevisser and Cameron, 1995). These were already indicators of where gay politics of the time would go, toward an apolitical stance unwilling to unsettle government.

In less than 20 years after the raid, the LRM had reinvented itself, only to remain the same. In naming itself, LRM was not explicit and was 'unobtrusive,' although it served the interests of gay people. By 1982 a national organization, Gay Association of South Africa (GASA) was formed, differing from LRM only by name. At this time, it was clear that the agenda for the organization was specifically in the naming of gay (Epprecht, 2012). It soon linked up with the 6010 initiative in Cape Town, as well as with other smaller groups in other parts of the country. GASA, which addressed the psychosocial needs of white gays and lesbians, had an explicit policy of being 'apolitical.'

This was a very crucial time in South Africa as apartheid was at its most volatile and aggressive. Antiapartheid groups were working hard at overthrowing the government. In August 1983, the United Democratic Front (UDF) was launched to resist against apartheid state's increasing repression. While there were gay and lesbian members part of the UDF and other members focusing on a militaristic actions against the apartheid government, GASA, which had formed a year earlier took an apolitical stance. Their mission statement read as follows: "a non-militant, non-political answer to gay needs" (Croucher, 2002: p. 318). So, quite early on, *de facto* white organizations and their constituencies divorced themselves from national politics. Even when GASA's members were in conflict with the state, the organization's stance was to "remain nonaligned in broader South African politics" (Gevisser, 1995: p. 51). So as not to compromise their work by being in conflict with government, GASA refused to speak out against the apartheid state's racist policies. This would later lead to GASA's closure and its exclusion from international gay platforms.

Operating in a tense political climate in South Africa, GASA's mainly white middle-class gay male members were forced to deal with their political conservatism when Simon Nkoli, a black gay man and antiapartheid activist from the township, as well as other black gays and lesbians joined GASA in 1983. Unlike their counterparts in GASA, Nkoli and his followers' sociopolitical conditions were neither uncomplicated nor apolitical. His presence, specifically, was necessarily provocative. He forced GASA to respond to many issues, including the organization's 'apolitical' and segregating stances. Luirink (1998: p. 21) presents a case that GASA was 'apartheid-friendly' in the ways in which it imposed upon its members to change their behavior, 'play it straight, keep a low profile, and to not give offense.' Cock (2003: p. 37) suggests that this ideology was 'patriarchal.'

Emily Craven (2011: p. 30) argues "GASA's supposed policy of remaining apolitical was in fact a very clear political project of accommodation in which there was a decision to negotiate space within the apartheid regime and not oppose it." It was a political stance to benefit white gay interests. Referring to GASA's decision and commitment to voting for the National Party because of its pro-gay attitude, Gevisser (1995: p. 61) argues "far from being apolitical GASA had in fact shown clearly that they would actively support racial repression if it was in the interests of white gay people."

Nkoli, with other members of the UDF, an antiapartheid organization of the 1980s, was arrested after a rent boycott in September 1984. From prison, he sought support from GASA. In 1985, he was charged with treason and was in detention for alleged terrorist activities together with 21 other activists. In response, GASA argued it could not support Nkoli as it claimed that he had played a minimal role in the South African gay movement and because GASA had an 'apolitical' constitution What GASA revealed was its gay political agenda, which was both opportunist (politically) in terms of self-interests and also apolitical (Gevisser, 1995: pp. 48–61). This same framework has taken place also in the fight for equal rights.

In 1988, Nkoli was acquitted and together with mostly black lesbian and gay activists set up the Gay and Lesbian Organisation of the Witwatersrand (GLOW). Unlike previous organizations that had existed, GLOW was overtly political and situated its politics in broader antiapartheid struggles. This was also the time of demonstrations campaigning for democracy and the unbanning of liberation movements. GLOW saw the recognition of human rights as indivisible from other struggles. In actual fact, GLOW's close association with the African National Congress (ANC), South Africa's national liberation movement, sat uncomfortably with many gay people. But as Gevisser (1995: p. 36) notes "even the most anti-ANC of gay

people probably had some sense that their best hope for equality and protection lay with that party." Two years after its formation, GLOW organized the first gay and lesbian pride march in Johannesburg.

Tensions within the South Africa gay rights movement have always existed and have been carried through to the democratic transition. White gays and black gays have been on polar ends. Politically, they have been and remain divided. Due to legal gains, there is an assumption of unity that in reality is a false sense of unity. As Jacklyn Cock (2003: p. 36) poignantly observes "the unity of the gay rights movement should not be overemphasized … the movement was fragmented, splintered politically and divided along race, gender, class and ideological lines."

'Equality for All – Terms and Conditions Apply'

The early 1990s, characterized particularly by black gay and lesbian visibility, saw gay and racial liberation march hand-in-hand. This was exemplified in the first pride march in Johannesburg. At the core of that march, organized by GLOW, was the idea for pride to be a visual symbol for gay and racial liberation. And this was carried out in the speech made by Simon Nkoli at that pride march. Ditsie (2006: p. 19) recounts it as follows:

> I am black and I am gay. I cannot separate the two parts of me into secondary or primary struggles … So when I fight for my freedom I must fight against both oppressions.

It is reasonable to believe that it is at this moment that the already existing tensions and fractures within the lesbian and gay movement solidified. GLOW's successes and ideologies were short lived. By 1994, when the National Coalition for Gay and Lesbian Equality (NCGLE) campaigned for the inclusion of sexual orientation in the new constitution that was being drafted, gay politics had returned to GASA's ideology – not only 'in the interests of white gay people' (Gevisser, 1995: p. 36) but black politics championed by GLOW receded. The NCGLE had orchestrated a community that in actual fact did not exist. Oswin (2007) argues that the coalition, by virtue of being a coalition and thus an umbrella body of between 20 and 75 organizations, managed to create an idea that a community and a unified lesbian and gay movement existed. Gunkel (2010: p. 71) argues that the NCGLE's strategy 'played down differences.' While its success, 'attributed to its single-issue focus' should be applauded, because of this the NCGLE can be likened to the LRM (Cock, 2003: p. 37).

Sexual orientation was included in the equality clause in the Bill of Rights in 1996. Already at that time, there was some dissent. Certain Christian leaders saw the clause as 'undemocratic' (Cock, 2003). For gays and lesbians 'the Constitution created an impression of unity' Craven (2011: p. 32). However, many recognized the existing fractures within the lesbian and gay movement, which would be played out by the infamous 'shopping list,' a litigation strategy that started with decriminalization of sodomy and ended with legal recognition of same-sex marriage (Berger, 2008: p. 18).

Ending with the highly controversial Civil Union Act in 2005, the litigation strategy by the NCGLE (which by then was the Lesbian and Gay Equality Project (LGEP)) to change the law on a number of issues had been considered a white middle-class act. Premiere in the continent and fifth in the world, South Africa allowed gay and lesbian couples to marry or, as many believe, have civil unions or partnerships (Judge et al., 2008). Upon presentation to the public, the Civil Union Bill faced almost outright rejection and opposition. Traditionalists, religious groups and everyone else who were already skeptical of the progressive nature of the constitution and gay rights, argued that same-sex unions would be against the order of nature and African cultural norms (Reid, 2008).

A journalist, John Qwelane, who later became South Africa's ambassador to Uganda (Gerntholtz and Nath, 2010), published a hate speech article lambasting the progressive nature of the constitution and the depreciating morals and values of South African society because of the legalization of same-sex marriage. His article included a cartoon illustration of a man marrying a goat with a headline "when human rights meet animal rights" (Qwelane, 2008: p. 14). The Johannesburg Equality Court later found Qwelane guilty of 'propagated hatred and harm' (News24, 2011).

Although suggested as a poverty-alleviation strategy (Oswin, 2007), the question of same-sex marriage was in actual fact directed at accessing benefits. It was to protect the economic interests of mainly white gay couples. Black gays and lesbians were made to believe that same-sex marriage promoted values of extended familial and community support. It remains unclear how black gays and lesbians have benefited from this law. The LGEP's strategy on marriage, and similarly same-sex marriage, begs to be problematized. It is very odd and worrying that in a country like South Africa, where the state is directly responsible for poverty and gross inequalities, gays and lesbians would seek marriage as a solution to these social ills. The complicit and assimilatory stance that many gays and lesbians have taken toward national politics and the state has made them seem apathetic to wider political struggles.

Oswin (2007: p. 650), following Spivak in Danius and Jonsson (1993), calls the LGEP's stance a "politics of strategic essentialism." She argues that the lobbying efforts deployed to advance constitutional gains have been a "deliberately conservative approach that has been characterized as elitist, unrepresentative, and male dominated" (Danius and Jonsson, 1993). On the one hand, this strategy created a constitution that would protect the rights of lesbians and gay people in South Africa. On the other, it created the category "poor, black gay or lesbian" (Oswin, 2007: p. 666), addressed later.

Perhaps the most interesting period in gay and lesbian politics in South Africa is post 2000. It is at this point that the wild dramatization of 'poor black lesbians and gays' becomes the spectacle that is 'gay' South Africa today. Since the early 2000s two narratives have emerged to symbolize the experiences of gay and lesbian people in South Africa. The one takes shape through the visual display of gay and lesbian pride marches throughout the major towns and cities in the country.

The other, which I deal with later, is the representation of the black lesbian who is the perpetual victim of male sexual violence and murder.

The Politics of Pride

The history of gay and lesbian pride marches goes back to 1991, having started in the streets of downtown Johannesburg with only about 800 people. In just over 20 years almost 20 000 people can be seen taking part in the pride march in the affluent northern suburb of Rosebank. However, the Johannesburg gay and lesbian pride march has simultaneously reinvented itself while remaining a constant reflection of a splintered rainbow nation. This notion of a rainbow nation, while celebrated for displaying the diversity that is South Africa today, is something of an illusion. Gqola (2001) argues that even though symbolic, the rainbow nation hides difference, political struggles, and the sociocultural histories of the South African body politic. The rainbow flag, carried and flown religiously at every gay and lesbian pride march annually, works as a constant reminder of this illusion.

Johannesburg Pride has been the most successful and longest-running pride march/parade in the continent. Over the years, it has claimed various spaces in the city, starting off in downtown Johannesburg and later moving to exclusive suburbs in the northern part of the city. Location of the pride march remains a contestation that marks racial and class differences. The pride march has also afforded different kinds of gay, lesbian, and to some extent transgender people an opportunity to make themselves visible.

At the most recent Johannesburg pride march, a group of black lesbians organized a protest demanding that the pride committee recognize the plight of black lesbians in South Africa. Their request was met by a violent outburst from the white organizers and marchers who demanded that the protesters remove themselves from the 'route' designated for gays and lesbians (SACSIS, 2012). The subsequent events of the clash between white gays and black lesbians have led to the closure of the nonprofit organization that runs Johannesburg pride. As one journalist points out "what transpired (at the Johannesburg Pride in 2012) also served to reveal the deep malady of racism in South Africa" (Schutte, 2012).

For many years, stark racial and class divisions have symbolized Johannesburg Pride that remains a 'fenced off' area in a public park. The fence has similarly marked racial and class boundaries, with black gay and lesbian participants gathering outside the fence in direct protest to the inaccessibility of the commercialized pride venture. In 1 year, a group of pride anarchists stood outside the pride fence in Rosebank wearing T-shirts written "equality for all – terms and conditions apply." A week prior to the pride in Rosebank, many of these same black lesbian and gay people march in the township for Soweto Pride, protesting against hate crimes toward black LGBT people. This is a political march, Soweto Pride, started in 2005 by a Forum for the Empowerment of Women, a black lesbian organization in Johannesburg (Mhlongo, 2008). Only a handful of white gays and lesbians would be seen at this march. These two forms of protest toward what is currently seen as a gay 'parade' have signaled the deep-seated fractures within gay and lesbian politics in South Africa.

As a number of black pride participants have put it, referring to pride parade in Cape Town and Johannesburg:

> … Gay and lesbian people have not been part of eradicating racial domination. Many gay and lesbian people have and continue to be part of the system that maintains the geographical apartheid that is in Cape Town today …. (Ludwig interview, 21 February 2013)

Bringing another form of exclusion that has characterized gay and lesbian politics generally, a trans activist reflects on the ways in which pride marches have ignored trans issues. He argues "transgender people within the LGBT grouping do not find reasons to be proud or even be at the pride marches. The agenda for pride marches has always been about homosexuality. The current challenges related to transgender people accessing identity documents or accessing hormonal therapy have never appeared in the political agendas of pride marches" (Kheswa interview, 21 February 2013).

Current gay politics can no longer continue avoiding structural inequalities and various forms of oppression and racism that exist in South African society. One of the challenges that local gay politics faces in South Africa is the increasing demand for global ideals. It has been argued elsewhere that globalization, capitalism, and commercialization tend to normalize gay and lesbian people for the global market (Gibson-Graham, 2001). When this happens, local politics and resistance shift from a form of 'radicality' to a politics of 'palatability' (Oswin, 2005) that serves the global gay image. This is the one of the issues that has currently faced Johannesburg and Cape Town Pride. An interview with one of the organizers of the antiprotest at Johannesburg illustrates this:

> Johannesburg Pride did not take seriously some of the issues affecting black lesbians in the townships. In particular pride has become a commercial venture benefitting only a few. The politics of resistance and change have taken a back seat, that pride is now just a party, a parade. (Mbandazayo interview, 21 February 2013)

Representation of Black Lesbians in South Africa

The black lesbian occupies a peculiar space in South Africa and the African continent. South Africa is layered with contradictions and complexities. On the one hand, LGBT people form part of the progressive movement and the rainbow nation that is South Africa today. On the other, individual rights to sexual expression, sexual and gender identity are often challenged in the only African state that is praised for protecting diversities. However, there are limits to diversity, at least in the South African context. Those who transgress the boundaries of diversity often get punished in the most gruesome ways. In particular, the existence of minority groups (including women and females) is often under threat. Among South African black lesbian women specifically, one is considered lucky to escape rape, or even murder before their 30th birthday (Matebeni, 2012).

Black lesbian life in South Africa has become synonymous with rape and even death (Schaap and Gim, 2010). The One in

Nine Campaign (2013: p. 1) states "when people in other parts of the world hear about lesbians in South Africa through the media, almost the only thing they are likely to learn is that butch, black, soccer-playing, lesbians in townships are raped, and sometimes killed, by black men who wish to 'correct' them …." Another media report reads, "Being a lesbian in South Africa can be a death sentence" (Turley, 2012). This is the popular representation of black lesbians in South Africa today, both in public discourse and in academic texts.

Black lesbian life has always been portrayed not only as undesirable, but also through the narrative of victimhood, and to a lesser extent, criminality. In this narrative, the black lesbian remains the object that is made to disappear, in various ways. This is seen in the first appearance of a queer female figure, Gertrude 'Gertie' Williams, in a story published in Drum magazine in 1956 and later reprinted in Gevisser and Cameron's Defiant Desire (1995). In this article and ones that followed, Gertie was termed a 'lesbian gangster' and was imprisoned (Chetty, 1995: p. 128). She was a referred to as a criminal – a female figure who could not be seen as a woman. Lisa Downing's (2013) current body of work on female murderers and criminals suggests that in areas like crime women are treated with an unquestioned prejudice. For example, she quotes Jeffner Allens' assertion that can remain unproblematized in the study of female criminality, "a woman is not violent and if violent, a female is not a woman" (Lisa Downing, 2013: p. 165).

In similar vein, a more contemporary look at black South African lesbians shows that they are easily publicly humiliated and punished for being a lesbian. It was in the public eye when Zoliswa Nkonyane, a 19-year-old lesbian, and her friend were approached while walking home by some girls in their township who claimed that the two 'wanted to be raped' because they were 'tomboys' and living as out lesbians. In the altercation, Zoliswa was clubbed, kicked, and beaten to death by a group of men in full view of people walking by (Thamm, 2006). This kind of public violence works in ways that communicate to black lesbians and noncomforming females that their sexuality and gender can 'be done away with' publicly, in the presence of witnesses. In addition, this is a way in which men control not only women's bodies, but also maintain patriarchal spaces. Many other women have been similarly done away with in such violent ways. (Sizakele Sigasa and Salome Masoa, a black lesbian couple raped and murdered in 2007 in Johannesburg; Thokozane Qwabe, a 23-year-old black lesbian raped and murdered in 2007, in KwaZulu-Natal; Sibongile, a black lesbian raped and murdered in 2008 in Cape Town; Eudy Simelane, a black lesbian murdered and suffered attempted rape in 2008 in Johannesburg; Khanyiswa, a black lesbian murdered in 2008 in Port Elizabeth; Ncumisa Mzamela, a 21-year-old black lesbian burnt to death in KwaZulu-Natal in 2010; and the following (with age): Nokuthula Radebe (20), Ntsiki Tyatyeka (21), Khanyiswa Hani (25), Tshuku Ncobo (26), Mpho Setshedi (27), Sanna Supa (28), Hendrietta Thapelo Morifi (29), Phumeza Nkolonzi (18). There are many others whose victimization, rape, and murder remains unreported and undocumented (source www.iranti-org.co.za).)

The heightened and special focus on violence toward black lesbians, in the context of wider forms of violence toward women and gender nonconforming persons in South Africa, has warranted some critique (Hames, 2011; Matebeni, 2013; One in Nine Campaign, 2013). In this narrative, the black lesbian is made to become the victim of a specific and directed form of violence easily deemed 'corrective' or even 'curative.' The concept 'curative/corrective' rape, now used widely by the media and other publics (Muholi, 2004), defines 'curative' rape, as the "rape of women perceived of as lesbian by men as an ostensible 'cure' for their (aberrant) sexualities" (Muthien, 2007: p. 323). Without any form of interrogation, the world as South Africans themselves have stripped the black lesbian of any other existence but the sight of a victim of 'corrective rape,' and eventually murder (Kelly, 2009).

This form of violence toward nonconforming women should not be seen as 'special' or even isolated. Currently, what is regarded as against the 'norm' is seen as worthy of punishment. People, whose sexuality is seen as nonnormative, although not articulated in this language, often fall victims to violent acts or prejudice. Previous cases such as those of women murdered in some townships in South Africa because of their HIV status are exemplary of such violence and prejudice. (Lorna Mlosana, 21 years of age, was raped in a toilet outside a shebeen/tavern in 2003 in Khayelitsha Township, Cape Town. After raping her, her attacker then beat her to death in a fury when she told him she was HIV positive (www.tlac.org.za, 2006). In a similar case, Gugu Dlamini, 36 years of age, from KwaMashu Township outside Durban was murdered in 1998 after she courageously publicly disclosed her HIV status on World AIDS Day. She was accused of shaming her community and suffered repeated threats and assaults that lead to her fatality (www.Beatit.co.za, 1998). The violent attacks and the murders of women disclosing their HIV status are perpetuated by the 'misogynistic notion that women are the primary vector of the HIV virus and thus deserving of death' (Salo and Gqola, 2006: p. 4).

The victim narrative of black lesbians is a problematic limited view of how black lesbians and gender nonconforming female persons experience the fullness of their lives. Rape and forms of sexual torture perpetrated on women's bodies are problematic, not only with regards to what they do to the victims and society. The language in which they are constructed and communicated veils the violent essence of the actual violence.

Gibson-Graham (2001: p. 241) suggest that there are different ways in which the victim narrative or victim role, which is "prescribed by the rape script," can be challenged. The most obvious way would be a refusal to "accept the victim role" (pp. 245, 248). A number of strategies can be put to use to do that, including 'using speech in unexpected ways, diminishing the power of the perpetrator, and rescripting the effects of the rape on the victim.' In particular, they suggest that a change in the representation of rape – to see rape as 'death, an event that is final and lasting in terms of the damage it does to self' (Gibson-Graham, 2001). Black lesbian groups challenge patriarchy and violence in South Africa by deploying various strategies, including legal representations in court as well as advocating for legislations against hate-related crimes.

Sexual Minorities in South Africa Today

The vision for the South African legal framework for sexual minorities has focused on conformity. As De Vos and Barnard (2007: p. 825) point out "considering the structural divisions of our apartheid legacy, it was specifically important from an integrative point of view that same-sex and heterosexual couples be brought together under one common institution … And this form of integration is indispensable as part of the normalization of post-apartheid South African society as a whole in the conformity with the constitutional ideal of openness." This vision, while important and significant in moving sexual minorities forward, has presented a number of challenges. The situation for black and lower class sexual minorities does not reflect the dream and pursuit of freedom for which many have fought. Racial, class, gender, and political tensions have brought many divides within the gay body politic, many of which cannot be easily solidified. It will take gay and lesbian people to rise up and take note of the divisions existing among them. Only through such political consciousness can South Africa become a better place for all.

See also: Gay and Lesbian Movements; Heterosexism and Homophobia; Identity and Identification, Social Psychology of; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Lesbians: Social and Economic Situation; Mass Media, Representations In; Queer Geography; Same-Sex Sexualities in Turkey; Sexual Minorities and Sexual Citizenship; Sexual Minorities in the Maghreb; Sexual Orientation: Historical and Social Construction; Sexual Orientation: Historical and Social Construction; Social Protest; Southern Africa: Sociocultural Aspects.

Bibliography

Berger, Jonathan, 2008. Getting to the constitutional court on time: litigation history of same-sex marriage. In: Judge, Melanie, Manion, Anthony, De Waal, Shaun (Eds.), To Have and to Hold: The Making of Same-Sex Marriage in South Africa. Fanele, Johannesburg, pp. 17–28.

Cameron, Edwin, 1995. Unapprehended felons: gays and lesbians and the law in South Africa. In: Gevisser, Mark, Edwin, Cameron (Eds.), Defiant Desire: Gay and Lesbian Lives in South Africa. Routledge, New York and London, pp. 89–98.

Chetty, Dhianaraj, 1995. Lesbian gangster: excerpted from golden city post and drum magazine. In: Gevisser, Mark, Cameron, Edwin (Eds.), Defiant Desire: Gay and Lesbian Lives in South Africa. Routledge, New York and London, pp. 129–133.

Cock, Jacklyn, 2003. Engendering gay and lesbian rights: the quality clause in the South African constitution. Women's Studies International Forum 26 (1), 35–45.

Craven, Emily, 2011. Racial Identity and Racism in the Gay and Lesbian Community in Post-apartheid South Africa (Unpublished MA Thesis). Wits University, Johannesburg.

Croucher, Sheila, 2002. South Africa's democratisation and the politics of gay liberation. Journal of Southern African Studies 28 (2), 315–330.

Danius, Sara, Jonsson, Stefan, 1993. An interview with Gayatri Chakravorty Spivak. Boundary 2 (20), 24–50.

De Vos, Pierre, Barnard, Jaco, 2007. Same-sex marriage, civil unions and domestic partnerships in South Africa: critical reflections on an ongoing saga. The South African Law Journal 124 (4), 795–826.

Ditsie, Beverley, 2006. Today we are making history. In: De Waal, Shaun, Manion, Anthony (Eds.), Pride: Protest and Celebration. University of the Witwatersrand: GALA, pp. 19–20.

Downing, Lisa, 2013. The Subject of Murder: Gender, Exceptionality, and the Modern Killer. University of Chicago Press, Chicago.

Epprecht, Marc, 2012. Sexual minorities, human rights and public sexual health strategies in Africa. African Affairs 111 (443), 223–243.

Gerntholtz, Liesl, Nath, Dipika, 27 January, 2010. Is Qwelane Good for Uganda? There Must Be a Better Ambassador for South Africa. Times LIVE. http://www.hrw.org/en/news/2010/01/27/qwelane-good-uganda.

Gevisser, Mark, 1995. A different fight for Freedom: A history of South African lesbian and gay organisation – the 1950s to the 1990s. In: Gevisser, Mark, Edwin, Cameron (Eds.), Defiant Desire: Gay and Lesbian Lives in South Africa. Routledge, New York and London, pp. 38–69.

Gevisser, Mark, Cameron, Edwin (Eds.), 1995. Defiant Desire: Gay and Lesbian Lives in South Africa. Routledge, New York and London.

Gevisser, Mark, Matebeni, Zethu, 2008. Jo'burg Tracks: Sexuality in the City. An Exhibition at Museum Africa. GALA, Johannesburg.

Gibson-Graham, Julie and Katherine, 2001. Querying globalization. In: Hawley, John C. (Ed.), Postcolonial, Queer: Theoretical Intersections. State University of New York, New York, pp. 239–275.

Gqola, Pumla Dineo, 2001. Defining people: analysing power, language and representation in metaphors of the new South Africa. Transformation 47, 94–106.

Gunkel, Henriette, 2010. The Cultural Politics of Female Sexuality in South Africa. Routledge, New York.

Hames, Mary, 2011. Violence against black lesbians: minding our language. Agenda: Empowering Women for Gender Equity 25 (4), 87–91.

Judge, Melanie, Manion, Anthony, De Waal, Shaun, 2008. To Have and to Hold: The Making of Same-Sex Marriage in South Africa. Fanele, Johannesburg.

Kraak, Gerald, 2008. Are our lives ok? Reflections on 13 years of gay liberation in South Africa. In: Judge, Melanie, Manion, Anthony, De Waal, Shaun (Eds.), To Have and to Hold: The Making of Same-Sex Marriage in South Africa. Fanele, Johannesburg, pp. 278–283.

Kelly, Annie, 2009. Raped and Killed for Being a Lesbian: South Africa Ignores 'Corrective' Attacks. http://www.guardian.co.uk/world/2009/mar/12/eudy-simelane-corrective-rape-south-africa (accessed 12.03.09.).

Luirink, Bart, 1998. Moffies: Gay Life in South Africa. Ink, Cape Town.

Matebeni, Zethu, 2012. Gone @20 – The Lucky Ones Are Not Yet Born. Black Looks. http://www.blacklooks.org/ (accessed 28.09.12.).

Matebeni, Zethu, 2013. Deconstructing violence towards black lesbians in South Africa. In: Ekine, Sokari, Abbas, Hakima (Eds.), Queer African Reader. Pambazuka Press, Oxford, pp. 343–367.

Mhlongo, Mongezi, 6 October, 2008. Pride Paints Soweto Pink. Women's Net. http://www.womensnet.org.za/news/pride-paints-soweto-pink (accessed 12.03.09.).

Muholi, Zanele, 2004. Thinking through lesbian rape. Agenda 18 (61), 116–125.

Muthien, Bernedette, 2007. Queerying borders: an Afrikan activist perspective. Journal of Lesbian Studies 11 (3), 321–330.

News24, 31 May, 2011. Qwelane Hate Speech Ruling Welcomed. News24. www.news24.com/South Africa/News.

One in Nine Campaign, 2013. What's in a Name? Language, Identity and, the Politics of Resistance. One in Nine Campaign, Johannesburg.

Oswin, Natalie, 2005. Researching 'gay Cape Town', finding value-added queerness. Social & Cultural Geography 6 (4), 567–586.

Oswin, Natalie, 2007. Producing homonormativity in neoliberal South Africa: recognition, redistribution, and the equality project. Signs 32 (3), 649–669.

Qwelane, John, 20 July, 2008. Call me names, but gay is not okay. Sunday Sun, 14.

Reid, Graeme, 2008. 'This thing and that idea': traditionalist responses to homosexuality and same-sex marriage. In: Judge, Melanie, Manion, Anthony, De Waal, Shaun (Eds.), To Have and to Hold: The Making of Same-Sex Marriage in South Africa. Fanele, Johannesburg, pp. 73–86.

Salo, Elaine, Gqola, Dineo Pumla, 2006. Editorial: subaltern sexualities. Feminist Africa 6, 1–6.

Schaap, Jeremy, Gim, Beein, 2010. Video Documentary: Female Athletes Often Targets for Rape. 'E:60'. http://sports.espn.go.com/espn/e60/news/story?id=5177704.

Schutte, Gillian, 8 October, 2012. Johannesburg Gay Pride Parade: Not Much to Be Proud of. The South African Civil Society Information Service (SACSIS). www.sacsis.org.za/site/article/1450.

Thamm, Marianne, 26 February, 2006. Not just another murder. Mail & Guardian Online. www.mg.co.za (accessed 03.04.06.).

Turley, Melissa, 18 October, 2012. South Africa: The Fight for Acceptance in the Rainbow Nation. http://www.globalpost.com/dispatches/globalpost-blogs/rights/South-Africa-lesbians-violence-lgbt (accessed 04.04.13.).

Williams, Christian, 18 February, 2013. Interview with an Actual Stonewall Riot Veteran: The Ciswashing of Stonewall Must End! The TransAdvocate. http://www.transadvocate.com/.

Sexual Minorities in the Maghreb

Jarrod Hayes, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Maghrebian culture going as far back as the classical age of Arabic literature is rich with representations of sexual minorities, and a number of studies in the social sciences examine its importance in Maghrebian societies, even though the latter are often lacking in case studies or field work. Since the 1950s, literature and films in French and Arabic have seen a renewed interest in sexual diversity. And like the several studies on sexuality in the social sciences written since the 1970s, contemporary cultural production has often rewritten the history of sexuality to understand the present. In short, literary and other cultural texts can also be of interest to social and behavioral scientists interested in the Maghreb.

Background

The Maghreb (or predominantly Muslim and Arabo-Berber region of Northwest Africa) has an extensive archive acknowledging sexual diversity. For example, the thirteenth-century *Delight of Hearts*, by 'Tunisian'-born Ahmad al-Tifashi, catalogs the amorous activities of *eunuques*, transgendered folk, go-betweens, prostitutes, libertines of all sexes, adulterers, pederasts, kept boys and women, and *epheboi*. It describes such 'sexual' acts as slapping, genital massage (or rubbing, especially among women), and sodomizing men while they sleep without waking them up. What stands out about its methodology are its generous quotations of literary texts, the most notable of which are the poems of Abu Nuwas, the eighth-/ninth-century Arab-Persian poet known for his praise of wine and the love of boys and women.

Texts such as this have often been deployed by partisans of both sides of the essentialism/social construction debate regarding the history of homosexuality. In the early days of the gay liberation movement, as historians sought evidence for the historic roots of Western lesbian and gay identities, writers such as Marc Daniel saw in classical Arabic literature evidence of gay men throughout history. Gay Sunshine, the same press that included Daniel's study in the collection *Gay Roots*, would only partially translate the French translation of al-Tifashi, offering the 'gay' parts, thereby imposing a hetero-/homosexual binary that is absent in the text as a whole. Likewise, John Boswell, in his foundational study for lesbian and gay studies, would write in 1980:

> Although the Qur'an and early religious writings of Islam display mildly negative attitudes toward homosexuality, Islamic society has generally ignored these deprecations, and most Muslim cultures have treated homosexuality with indifference, if not admiration. Almost without exception, the classical works of Arabic poetry and prose, from Abu Nuwas to the *Thousand and One Nights*, treat gay people and their sexuality with respect or casual acceptance. (pp. 194–195)

From the other side of the essentialism/social construction debate, in the foundational first volume of his *History of Sexuality*, Michel Foucault would reinforce the social constructionist argument when he defined the *scientia sexualis* that characterized Modern Western societies against "the societies – and they are numerous: China, Japan, India, Rome, the Arabo-Moslem societies – which endowed themselves with an *ars erotica*" (p. 57). Along the same lines, Arno Schmitt would write, "But in these societies there are no 'homosexuals' – there is no word for 'homosexuality' – the concept is completely unfamiliar. There are no heterosexuals either" (Schmitt and Sofer: p. 5).

Western (Mis)understandings of Sexuality in the Maghreb

Orientalism

Although deployed in the essentialism/social construction debates, discussions of sexuality in the Arab world have too often been based on a questionable understanding of same-sex eroticism and desire among Muslims. For example, in spite of its constructionist stance, *Sexuality and Eroticism among Males in Moslem Societies*, the 1992 collection that Schmitt edited with Jehoeda Sofer, nonetheless produced a culturally essentialist vision of an absolute Muslim difference when it comes to sexuality, a vision saturated with colonial prejudices and Orientalist stereotypes, including the counting of experiences of sexual tourism as anthropological or sociological evidence.

In his foundational work on Orientalism, Edward W. Said (1978) highlighted the sexual aspects of European understandings of the 'Orient,' itself a colonial invention. Connecting the sexual and epistemological connotations of the word 'penetration,' Orientalist penetration (understanding and knowledge) and its parallel in colonial penetration (military invasion) both come to have sexual implications: "The Middle East is resistant, as any virgin would be, but the male scholar wins the prize by bursting open, penetrating through the Gordian knot despite 'the taxing task'" (p. 309). These implications take on a very literal component in representations of European travelers' sexual experiences in the 'Orient,' because "the Orient was a place where one could look for sexual experience unobtainable in Europe" (p. 190). Furthermore, some of these experiences would have been considered nonnormative from a Western perspective. So when Said writes, "The Orient becomes a living tableau of queerness" (p. 103), although 'queer' did not yet have the implications it does today, this citation may indeed now be read in its contemporary meaning.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.10256-9

In *Desiring Arabs* (2007), Joseph A. Massad characterizes Schmitt and Sofer's representation of Arab and Muslim desires as Orientalist just as he does with Boswell's (164–169). To this list he also adds Stephen O. Murray and Will Roscoe's 1997 *Islamic Homosexualities: Culture, History, and Literature* (Massad: pp. 170–171). Furthermore, in Massad's view, Orientalism continues to have an enduring influence in not only Western but also Middle Eastern and North African writings on sexuality in the Arab world. He criticizes, for example, Moroccan feminist Fatima Mernissi, who according to him, "in classical Orientalist manner, employed seventh- through sixteenth-century Arabic texts to *interpret* and *explain* not medieval but *modern* Arab society" (pp. 154–155). In this critique, he also mentions Fatna Aït Sabbah's *La femme dans l'inconscient musulman: Désir et pouvoir* (1982) (*Woman in the Muslim Unconscious*), whom he suggests is like Mernissi (in fact, the website of the Sorbonne's Institut européen en sciences des religions states that Aït Sabbah is a 'nom de plume' for Mernissi), and he argues that the latter's work is intended for a Western audience. In an analysis of Qur'anic, juridical, erotic, and courtly literature, Aït Sabbah examines representations of women's sexuality and devotes almost half the book to the archetype of the 'omnisexual woman.' This figure reduces woman to her vagina, which is characterized by the ability to suck in all it comes in contact with. The omnisexual woman, who desires giant penises as well as sex with women, can never be satisfied sexually. While this is a set of stereotypes that predominantly apply to women, there are omnisexual men as well.

Massad's list of Orientalist writing on sexuality by Arab writers also includes Tunisian social psychologist Abdelwahab Bouhdiba's 1975 *La sexualité en Islam* (*Sexuality in Islam*). Like Aït Sabbah, Bouhdiba does not focus specifically on the Maghreb but on Muslim societies in general. His work offers a psychological survey of the Qur'an, Islamic law, and classical Arabic literature, especially that of an erotic nature, the first and foremost of which is *Alf layla wa layla* (*A Thousand and One Nights* or *The Arabian Nights*). Bouhdiba's word of choice for same-sex desire is homophilia, and his discussion of it is not ghettoized in a single chapter of the book but is mixed in throughout with discussions of the division between the sexes, marriage, the meaning of body hair, the place of sexuality in heaven, the *hammam* or public bath, the practice of circumcision, veiling, as well as illicit sexual behaviors (which include ones committed between members of the same sex): sodomy, pederasty, cross-dressing, prostitution, and both puritan and obscene folklore. To his citations of classical literature, Bouhdiba adds references to contemporary literature in both Arabic and French. Examples of the latter include the Algerian Rachid Boudjedra's 1969 novel *La repudiation* (*The Repudiation*), which will be discussed again below.

Though not considered by Massad, Algerian anthropologist and psychoanalyst Malek Chebel's *L'esprit de sérail: Perversions et marginalités sexuelles au Maghreb* (1988) (*The Seraglio Spirit: Perversions and Marginal Sexualities in the Maghreb*), in its very title, deploys a stereotypical image from Orientalist discourse (a gesture Massad would no doubt criticize). (The seraglio is an Orientalist reference to the harem – women's portion of the house into which nonrelative males are not admitted – and it often tends to imply inclusion of a large number of female slaves and concubines in addition to wives.) Chebel's work contains seven chapters, one on each of the following topics: (1) homosexuals, perverts, and androgynes; (2) obscene language; (3) the virginity taboo; (4) sexuality and the veil; (5) erotic literature in Arabic; (6) ululations; and (7) circumcision.

Ahistoricism

As in Massad's critique of Mernissi quoted above, a key aspect of Orientalist characterizations of sexuality in the Maghreb involves using the past to explain the present in a way that conflates the two. Since the publication of Johannes Fabian's *Time and the Other: How Anthropology Makes Its Object* (1983), cultural historians and social scientists have been more reluctant to compare non-Western societies and cultures to periods from the Western past. Yet while one might think that such Western ahistorical readings of classical texts could be responsible for such homogenizing accounts of any so-called Arab or Muslim sexuality, a small but significant body of studies of sexuality by Francophone Maghrebians has continued the process of generalizing across vast historical periods by similarly citing historically vast bodies of Arabic (and, increasingly, French) texts to offer accounts of sexuality with similar claims to comprehensiveness. Aït Sabbah, Bouhdiba, and Chebel already discussed above all fall into this category.

Homogenization

Ahistoricism itself might be considered but one part of a more general homogenizing approach to sexualities in the Muslim world or in North Africa and the Middle East as an all-inclusive region. Murray and Roscoe's *Islamic Homosexualities* would likewise survey historical periods from the pre-Islamic era to the present in a much more anti-Eurocentric approach that was also less concerned with a binary opposition between the Islamic world and the West. Nonetheless, their title does reveal a certain bias in that it characterizes the Muslim world as homogeneous to a much greater extent than what obtains in Western scholars' treatment of their own societies. Who, for example, in the West would think of entitling a comparable study on the West 'Christian Homosexualities'?

The Maghreb Writes Back

Social Sciences

While Massad's critique of an Orientalist assertion of a timeless essence regarding Arab sexualities must certainly be acknowledged and valued, there is an alternative way of understanding Bouhdiba's and, especially, Aït Sabbah's and Mernissi's readings of classical texts. This alternative would understand their interpretations as precisely that, rereadings of the past that strategically seek to make interventions in the present. In a Maghrebian context, especially in the periods prior to, during, and after the civil war in Algeria in the 1990s, a number of writers have looked to the past to challenge an Islamist monopoly on Islamic 'tradition' and religious writings as well as to assert a more heterogeneous version of the past as a way of effecting change in the present. Indeed, this is the stated goal of many of Mernissi's writings.

In contrast, one might argue that the same is the goal of more recent Maghrebian writings on sexuality; two in particular also happen to have restricted their assertions to the Maghreb (in the first case, Chebel) and even more specifically Morocco (in the second, Serhane). For example, one might argue that Chebel recuperates the Orientalist cliché of the seraglio in the interest of promoting tolerance for sexual diversity. While the first edition of this work was published before the Algerian civil war was in full swing, its reissue in the midst of that war in 1995 was under a significantly less titillating title: *L'esprit de sérail: Mythes et pratiques sexuelles au Maghreb* (*The Seraglio Spirit: Sexual Practices and Myths in the Maghreb*). One must therefore wonder whether this change in title was meant to make the work more palatable to a wider readership.

While many have focused on Massad's critique of a 'gay international' that attempts to impose Western constructions of sexuality onto the Arab or Muslim world, Chebel might also be associated with the secularist Arabic-language scholars examined by Massad, those who have explored the applicability of the concept of sexuality to the Arab world. In spite of the prominence Chebel gives to 'homosexuality' by discussing it in his first chapter, like Bouhdiba, Chebel contextualizes same-sex desire within a variety of other sexual practices with the effect of normalizing the 'perverse' or, conversely, of demonstrating how central 'perversions' are to Maghrebian sexual and gendered practices. This is what in the West would be called queering as a critical and theoretical practice, one that also involves a certain work of translation, between Arabic and French, between European and Arab or Muslim cultures, between the premodern and the modern. Indeed, many of the contemporary Arabic-language writers whom Massad discusses also consider the long tradition of Arabic 'sexology.'

Massad's extensive work on 'translations' (like those mentioned above) from French or Anglophone works and concepts into Arabic constitutes a much more complex account than what many of his critiques have allowed for. And as Sahar Amer demonstrates (2008), such translation has not always been a one-way street, but has instead 'gone both ways.' Furthermore, Chebel goes further than Bouhdiba in his use of contemporary French-language writers such as Boudjedra, Mohammed Dib, Kateb Yacine, Aicha Lemsine, and Abdelhak Serhane, some of whom will be discussed again below. Chebel has also, by now, produced a considerable body of additional works related to sexuality or eroticism in the Maghreb: *Le corps dans la tradition au Maghreb* (1984) (*The Body in Maghrebian Tradition*), *La féminisation du monde: Essai sur "Les Mille et Une Nuits"* (1996) (The Feminization of the World: Essay on the "Thousand and One Nights"), *Histoire de la circoncision: Des origines à nos jours* (1992) (*The History of Circumcision from Its Origins to the Present*), *Le livre des séductions* (1986) (*The Book of Seductions*), and *L'encyclopédie de l'amour en Islam: Erotisme, beauté et sexualité dans le monde arabe, en Perse et en Turquie* (1995) (*The Encyclopedia of Love in Islam: Eroticism, Beauty, and Sexuality in the Arab World, Persia, and Turkey*). *La féminisation du monde* is especially notable for its treatment of 'homosexuality,' including Chebel's most extensive discussion of lesbianism.

Furthermore, in contrast with the geocultural generalizations of Mernissi and Bouhdiba, over time the social science on sexuality has increasingly circumscribed its purview. The second of the more recent Maghrebian works on sexuality is psychologist Abdelhak Serhane's 1995 *L'amour circoncis* (*Circumcized* Love). (Serhane is most known, however, as a novelist.) As mentioned above, this work, the most recent in a series of Maghrebian studies on sexuality, is also the most circumscribed geoculturally, restricting as it does its purview to the sole country of Morocco, even though, thematically it overlaps a great deal with its more general predecessors. After a first part on the socialization or typical development of the Moroccan child and a second one on the role of women, the place of sexuality in Moroccan tradition, and father- and motherhood, part three treats such topics as circumcision, masturbation, homosexuality, bestiality, the *hammam*, prostitution, and marriage. Again, by placing the institution of marriage at the end of a longer list including nonnormative practices, Serhane echoes Chebel's tactic of 'queering' the normative and normalizing the 'queer.' Serhane also stands out in his emphasis on more contemporary sources, including Mernissi (and with her the field of sociology) as well as Freud (and with him psychoanalysis), the latter of whom Serhane characterizes as antifeminist. Furthermore, Serhane supports his discussion of contemporary Moroccan society with quotations from the Qur'an and other religious texts as well as from classical literature, along similar lines as the writers discussed above.

In spite of its more specific context, then, *L'amour circoncis* covers similar ground as Bouhdiba and Chebel, both of whom Serhane cites. Yet more than Bouhdiba and Chebel, in many of his points Serhane reads like barely veiled allusions to contemporary Maghrebian literary production in French. One example is his description of same-sex sexual abuse in Qur'anic schools, of which an episode is described in Boudjedra's *La répudiation*. Another would be Serhane's considerable discussion of zoophilia, also described in a page-long comparison in Rachid Mimouni's *Le fleuve détourné* (1982) (*The Diverted River*) of the advantages and disadvantages of sex with a goat, cow, and she-ass, the last of which wins this particular 'contest' (pp. 39–40). In fact, in many ways, *L'amour circoncis* reads like an essay version of representations of sexuality in Serhane's novel *Les enfants des rues étroites* (1986) (*Backstreet Boys*), which itself might also remind readers of *La répudiation*. The latter contains the most extensive, detailed, and explicit representation of two boys having sex in view of the entire family of one of them. In his 1983 *Messaouda* (pp. 133–134), Serhane mentions 'nouiba,' which entails boys' taking turns sodomizing one another. The eponymous character of this novel is an androgynous prostitute, initiator of boys, and witch.

Literature

1950s

Well before the 1980s, which saw the publication of Serhane's first novels, the foundational 1950s generation of Maghrebian writers began to distinguish their literary production in French from that of metropolitan France often in the form of colonial childhood narratives. These novels were the first to assert a certain sexual diversity, especially among boys and unmarried men. The first edition of Mouloud Feraoun's *Le fils du pauvre* (1950) (*The Poor Man's Son*) mentions sex between boy

cousins. In Mouloud Mammeri's *La colline oubliée* (1952) (*The Forgotten Hill*), the narrator's cousin has a homoerotic relationship with a musician/dancer. Albert Memmi's *La statue de sel* (1953) (*Pillar of Salt*) describes collective sex play among teenage boys of mixed ethnicities. Formerly the passive partner in anal intercourse, the male protagonist of Driss Chraïbi's *Le passé simple* (1954) (*The Simple Past*) insists on being the active partner with other men as he reaches adulthood.

1960–90

Since the 1950s, representations of sexual diversity have not abated; indeed they have become more frequent. In fact, rare is the Maghrebian writer who has not incorporated representations of male same-sex desire in at least one novel. Others, like Boudjedra and Tahar Ben Jelloun include it in several works. Representations of female same-sex desire are rarer, but far from nonexistent. In the French version of Boudjedra's *La pluie* (1987) (*Rain*), the narrator has a lesbian aunt. The female transgender protagonist of Ben Jelloun's *L'enfant de sable* (1985) (*The Sand Child*) has a sexual encounter with (another) woman after his or her wife dies. As part of her life story, the female narrator of *Les yeux baissés* (1991) (*Downcast Eyes*) describes a homoerotic encounter with another girl that leads to a more sexual dream. Merzak Allouache's novel *Bab el-Oued* (1995) includes an affair between two women (though it is not included in the film version). In less explicit terms, Assia Djebar stages a number of homoerotic encounters within the women's *hammam* in *Ombre sultane* (1987) (*A Sister to Scheherazade*) and in the title story of *Femmes d'Alger dans leur appartement* (1980) (*Women of Algiers in Their Apartment*)

In "Défendre la diversité culturelle au Maghreb" (1991) (Defend Cultural Diversity in the Maghreb), Ben Jelloun also defends representations of sexual diversity and claims that Maghrebian novels in French have been at the forefront of challenging the status quo through sexual transgressions. Yet he himself translated one of the most important contemporary Arabic texts in this regard: Mohamed Choukri's *Le pain nu* (1980), translated into English as *For Bread Alone* (1973), is an autobiographical narrative of growing up in poverty amidst petty crime, drugs, and prostitution, and although most of the prostitution is heterosexual, the male narrator recounts one episode of selling sex to an older man. Since 1982, Boudjedra has written his novels in Arabic first. While there are differences between the Arabic 'originals' and their French 'translations,' same-sex desire is far from absent from the Arabic versions.

1990–Present

After a rich tradition of supposedly heterosexual representations of queer sexualities, a handful of openly homosexual writers began to publish in the 1990s. Beginning in 1995 with *L'enfant ébloui* (*The Dazzled Child*), the Moroccan Rachid O. has written four autobiographical novels in which he relates his cross-generational loves as a boy and his affairs with European men as a young man. In *Un "poisson" sur la balançoire* (1997) (*'Fish' on the Swing*) and its sequel *Une promesse de douleur et de sang* (1998) (*A Promise of Suffering and Blood*), which are explicit enough to be characterized as erotica, Eyet-Chékib Djaziri recounts the boyhood loves and affairs of a binational, French-Tunisian protagonist. The Moroccan Abdellah Taïa has also written a number of autobiographical novels in which the main character's desire for other boys (as a younger boy) and men (as he grows up) plays a prominent role. From novel to novel, he describes same-sex relations in both Morocco and Europe. In *L'armée du salut* (2006) (*Salvation Army*), the narrator explicitly describes his sexual desire for his beloved brother Abdelkébir (the same family names recur from novel to novel)

Although she might deny the label lesbian writer (as she denies those of woman, French, or Algerian writer), Nina Bouraoui sets a lesbian love story in a Swiss girl's boarding school in *La vie heureuse* (2002) (*Happy Life*). In her previous *Garçon manqué* (2000) (*Tomboy*), a female narrator refuses to accept being assigned a gender. Finally, any discussion of Maghrebian gay writing in French must also include Jean Sénac, an Algerian poet of European origins. In spite of openly praising the love of boys, he was a member of the Ben Bella government after independence and, even after being marginalized by Boumédiène's coup in 1965, went on to host a radio show on poetry that rivaled the popularity of television.

Film

There have also been a number of films treating aspects of same-sex desire in the Maghreb. *Rih essed* (*Man of Ashes*) (1986) tells the story of a groom-to-be who remembers being sexually molested as a boy apprentice by his master. By the same director, *Bezness* (1991), whose title is the Tunisian word for 'beach gigolo,' tells the stories of male prostitutes with both men and women clients in Tunisia on vacation. Referred to by some as the 'Algerian Almodovar,' Nadir Moknèche directed one film, *Viva Laldjérie* (*Viva Algeria*) (2004) that contains scenes of public sex between men. *Le fil* (*The String*) (2009) is about a France-based architect who returns to Tunisia and falls in love with his mother's gardener, a monolingual Francophone son of Tunisian parents who has reverse migrated (back) to Tunisia. The architect also enters into an arranged marriage with a lesbian friend so that the son of her and her partner can have a legally recognized father.

Conclusion

In short, Maghrebian culture is rich with representations of sexual minorities, and a number of studies in the social sciences examine its importance in Maghrebian societies, even though the latter are often weak in case studies or field work. In the 1996 gay-pride issue of the Parisian daily *Le monde*, however, a writer going by the name of Anis described an aborted attempt to organize a lesbian and gay movement in Algeria during a brief window of political opening at the end of the 1980s. Placed in a history with Sénac, this historical detail could only suggest what political possibilities lay ahead.

See also: Afrocentric Approaches to Social Work; Gay and Lesbian Movements; Gay and Lesbian Tourism; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Mass Media, Representations In; Middle

East, North Africa: Sociocultural Overview; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Gender; Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations; Queer Geography; Same-Sex Sexualities in Turkey; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Aït Sabbah, Fatna, 1982. Woman in the Muslim Unconscious (Mary Jo Lakeland, Trans.). Pergamon Press, 1984, New York.

Aït Sabbah, Fatna, 14 October 2012. http://www.iesr.ephe.sorbonne.fr/index6643.html.

al-Tifashi, Ahmad, 1981. Les délices des cœurs; ou, Ce que l'on ne trouve en aucun livre (René R. Khawam, Trans.). Phébus, Paris.

al-Tifashi, Ahmad, 1988. The Delight of Hearts; or, What You Will Not Find in Any Book (Edward A. Lacey, Trans.). Gay Sunshine Press, San Francisco.

Amer, Sahar, 2008. Crossing Borders: Love between Women in Medieval French and Arabic Literatures. University of Pennsylvania Press, Philadelphia.

Anis, 22 June 1996. Un homosexuel à Paris. Le monde, p. 15.

Ben Jelloun, Tahar, 1991. Défendre la diversité culturelle au Maghreb. In: L'état du Maghreb, Camille, Yves Lacoste (Eds.). La Découverte, Paris, pp. 271–272.

Boswell, John, 1980. Christianity, Social Tolerance, and Homosexuality: Gay People in Western Europe from the Beginning of the Christian Era to the Fourteenth Century. University of Chicago Press, Chicago.

Bouhdiba, Abdelwahab, 1975. Sexuality in Islam (Alan Sheridan, Trans.). Saqi Books, 1998, London.

Chebel, Malek, 1988. L'esprit de sérail: Perversions et marginalités sexuelles au Maghreb. Lieu Commun, Paris.

Chebel, Malek, 1995. L'esprit de sérail: Mythes et pratiques sexuelles au Maghreb. Payot, Paris.

Daniel, Marc, 1975–1976. Arab civilization and male love. In: Winston Leyland, Gay Roots (Ed.), Twenty Years of Gay Sunshine: An Anthology of Gay History, Sex, Politics and Culture. Gay Sunshine Press, 1991, San Francisco, pp. 33–75.

Fabian, Johannes, 1983. Time and the Other: How Anthropology Makes Its Object. Columbia University Press, New York.

Foucault, Michel, 1976. The History of Sexuality: Volume One: An Introduction (Robert Hurley, Trans.). Random House, 1978, New York.

Massad, Joseph A., 2007. Desiring Arabs. University of Chicago Press, Chicago.

Murray, Stephen O., Roscoe, Will, 1997. Islamic Homosexualities: Culture, History, and Literature. New York University Press, New York.

Said, Edward W., 1978. Orientalism. Random House, 1979, New York.

Schmitt, Arno, Sofer, Jehoeda (Eds.), 1992. Sexuality and Eroticism among Males in Moslem Societies. Harrington Park Press, New York.

Sexual Minority Families of Choice and Family Networks: We Are Family

Emily E Wheeler and Sharon G Horne, University of Massachusetts Boston, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the lesbian, gay, bisexual, transgender, and queer (LGBTQ) community, the concept of 'family' has different meanings. Historically the right to form a traditional family (e.g., two parents and children) with associated legal rights and protections has not been an option for LGBTQ couples. In the absence of formal family arrangements, LGBTQ individuals have created their own family and community structures. Yet, over time, steps have been taken to create and legitimize family rights for LGBTQ people, and as social and political attitudes and norms have changed and laws and policies followed suit, so has the ability of LGBTQ individuals to form legally recognized family structures.

Family of Choice

Family and kinship have long been associated with groups bonded via marriage and birth. The prototypical nuclear family and its companion, the extended family, have been firmly entrenched in the modern history of the American family. It is in the context of this history, as well as the limits placed on lesbian, gay, bisexual, transgender, and queer (LGBTQ) couples by heterosexism and anti-LGBT violence, laws, and policies found throughout the United States, that Kath Weston's (1991) ethnographic study of lesbian and gay individuals living in and around the San Francisco Bay area, *Families We Choose*, challenged this definition within that community.

Weston's (1991) work elicited several important insights about how the 80 men and women she interviewed (in addition to herself, as a reflective participant) conceived of kinship. Her account captured a more fluid conceptualization of family, and one that was not necessarily predicated on marriage and 'blood' relationships. Weston's participants shared a range of responses to their disclosure of their sexual identities to their 'straight families,' 'families of origin,' or 'biological families,' some of which had been maintained through the process of revealing their sexual identity and some of which had not. However, most remarkable about this work was the experience participants shared of their 'gay families,' the titular 'families we choose,' or 'families of choice,' which they defined as the networks of current and former romantic partners and friends.

What was perhaps most provocative and important about Weston's (1991) account was that her intimate descriptions of these family of choice bonds challenged the idea that the strength of family ties is predicated on those of the more traditional nuclear family structure. At the time of its publication, marriage was long from being a legal possibility for same-sex couples in the United States and many states prohibited adoption by second-parents of the same gender. Thus, legally prohibited from traditional 'family' structures and at times having been rejected by families of origin, Weston described the means by which her participants negotiated life through kinship ties to their gay families, including culture and identity building.

These two kinds of family structures (family or origin and family of choice) are not meant to be mutually exclusive, nor are families of choice a replacement for families of origin. Indeed, families of choice may include chosen family of origin members. Family bonds are meant to be characterized by strength and durability, beyond a collection of one's close friends or relations. Families of choice were a means of accessing a particular kind of support around sexual identity and related issues that could not be assumed in other relationships.

Gay and lesbian communities and neighborhoods have also been viewed as part of an extension of family of choice. Levine's (1979) landmark essay 'Gay Ghetto' was the first to describe the phenomenon of urban physical spaces occupied by a community of gay individuals in sociological terms. The term 'ghetto' was meant to elicit comparison to other cultural communities that had been segregated, and gay ghettos, also known as 'gay enclaves,' have most often been associated with the gay male community, albeit not exclusively. Salient examples include the Castro district in San Francisco, Christopher Street in New York, Key West in Florida, and Provincetown and Northampton, Massachusetts. These neighborhoods have served as spaces that allow a freedom of expression not otherwise possible and have also been homes for LGBTQ-based organizations and institutions. Further, and not unlike families of choice, these communities have often served as a means of culture shaping and maintenance by acting as sites of communication and storytelling, as well as collective action (Usher and Morrison, 2010).

LGBTQ Families in the Twenty-First Century

As the legal, political, and social landscape has shifted for the LGBTQ community, so too has the concept of family. Yet families of choice, as Weston (1991) described them, have continued to be important for many LGBTQ individuals. The family of choice concept has found some empirical support; LGB individuals maintain more friends as confidantes as they age than the general population, whereas the general population maintains connection with more family of origin members (Dewaele et al., 2011). Oswald (2002) applied the family resilience model to gay and lesbian families and described two main processes that characterize these networks: intentionality and redefinition. Intentionality refers to the

processes by which gay and lesbian individuals create networks in a nonaffirming environment and includes acts such as choosing family members, creating rituals, and building a sense of community. Redefinition refers to processes by which gay and lesbian people and their family networks are affirmed, and includes such acts as acknowledging the role of heterosexism in their experience and giving family members named identities. Such an approach to family uses the concept of resilience to highlight the ways in which these family networks are a means of coping with the risks and challenges that result from stigma and oppression.

The role of families of choice have also been explored among more specific groups such as aging lesbian women (Ariel, 2008); rural gay men (Cody and Welch, 1997); gay men with HIV (Nord, 1996); African-American men who have sex with men (Kubicek et al., 2012); and gay, lesbian, and bisexual youth (Nesmith et al., 1999). These kinship networks have been described in the literature as a critical source of support for LGBTQ individuals, both for issues and stressors related to sexuality and gender identity.

Many LGBTQ men of color participate in created families that are known as gay families or 'houses.' These social networks can include members who participate as 'parents' or 'grandparents' and mentor and support 'children' who are often younger men who have been rejected by families of origin or who experience sexual stigma in their ethnic communities and racial stigma in majority culture (Phillips et al., 2011). These families are often connected with dance or modeling performances in 'balls' and these networks typically embrace a range of gender expressions and identities. Houses have been found to be important for HIV prevention and intervention efforts within communities of color by helping men to reduce risky sexual behavior (Arnold and Bailey, 2009).

The support of other LGBTQ individuals is of particular importance in an environment in which heterosexism continues to proliferate and experiences of discrimination and harassment on the basis of sexuality and/or gender are all too common (Almeida et al., 2009; Hequembourg and Brallier, 2009; Nadal et al., 2012). Families of choice create alternative environments in which individuals can experience safety and where they can communicate honestly about their sexuality and gender. Much like the way traditional family structures are conceptualized, family of choice relationships offers a stable source of support over time as well.

While family networks have persisted within the LGBTQ community, traditional 'gay enclaves' have experienced more change. In a *New York Times* article, Brown (2007) questioned if such neighborhoods were 'passé' due to the sometime marked changes over the last several decades. In their chapter, Usher and Morrison (2010) described that these areas have experienced a demographic shift as more heterosexual individuals have moved in and LGBTQ residents are more often married and have children or have moved into outer urban neighborhoods and suburbs. In addition, Usher and Morrison point to the economic shifts in these areas, which were once dominated by community organizations and businesses operated by LGBTQ community members but have been increasingly commercialized by mainstream businesses that have driven up rent prices and driven out long-standing neighborhood institutions. While the emerging role of online communities in the LGBTQ community may replace physical LGBTQ space, these changes also elicit questions about the possible detriment of culture-building and collective action when such physical spaces are no longer present.

LGBTQ Families: Marriage and Parenting Legislation

One source of significant change in the LGBTQ community is the evolving legislation regarding rights related to gender and sexuality, in particular, marriage and parenting rights. Marriage is the more visible of these two areas, with same-sex marriage laws having been passed in 17 states and the District of Columbia and civil unions and domestic partnerships legalized in several others since Vermont passed the first civil union law in 2000. Recent attention has been turned to the current federal marriage law in the form of the Defense of Marriage Act, which restricted the definition of marriage to one man and one woman and prohibited the recognition of same-sex unions at the federal level, and the Supreme Court's decision to hear a case arguing the law's unconstitutionality (Condon, 2012). In 11 states and the District of Columbia, adoption of an unmarried partner's child, termed 'second-parent adoption,' is protected under law, and in states where same-sex relationships are recognized legally, parents can access rights afforded to married couples such as stepparent adoption. In some other states, second-parent adoption is not protected by law but is legally recognized. In the remaining states, no legal avenues for second-parent adoption exist. The ramifications of these laws are manifold; issues range from access to a partner's insurance benefits to child custody in the case of injury or death, and they also set cultural standards of inequality by prizing some relationships (legally recognized) over others that are not deemed as valid.

Marriage and parenting rights legislation for LGBTQ families has resulted in steps toward equal rights for members of this community, but the larger movement has also spurred some anti-LGBTQ legislation. Those lobbying against LGBTQ rights or on behalf of anti-LGBTQ legislation may enlist rhetoric that magnifies existing heterosexist attitudes. The families of LGBTQ individuals may be delegitimized in comparison to traditional families (Peterson, 2011) or attacked as less fit or unfit to parent, despite a plethora of research demonstrating otherwise (for a review, see Patterson, 2009).

A growing body of research has investigated the impact of legislative campaigns concerning LGBTQ rights on members of the community. These legal battles are not without consequences for community members and have been associated with increased negative affect, depressive symptoms, and increased stress (Rostosky et al., 2009; Russell, 2004), and experience of negative outcomes has been found among family members of LGBT individuals as well (Arm et al., 2009; Horne et al., 2011). The stressors associated with these legal battles for LGB individuals include facing increased homophobia, increased vigilance, and experiences of division within communities (Russell and Richards, 2003). On the other hand, participation in activism for LGBTQ rights has also been associated with benefits and is considered a means of coping with the experience of discrimination and harassment (Riggle et al., 2010).

In addition to changes in legislation specific to LGBTQ rights, broader social changes have also molded the landscape of family life for heterosexuals and LGBTQ individuals alike. Weeks et al.'s (2001) text, for which they interviewed British LGBTQ men and women, put the contemporary experience of these individuals in the context of the history of families of choice and dominant conceptualizations of family and found space for both similarities with and differences between LGBTQ and heterosexual individuals. Weeks et al. emphasize the freedom that their participants voiced in regards to their experience of family; having been marginalized from dominant conceptualizations of family life, participants were freed from the expectations that come with these ideas and could create their own kind of family instead. In addition, the authors note that both the traditional nuclear family structure and heterosexism, while still predominant, have been increasingly challenged on a societal level, and it is in the context of these changes that contemporary LGBTQ individuals have begun to expand notions of families of choice. Weeks et al. also emphasize a tension between the pressure to embrace traditional family forms and the discomfort or rejection by others experienced as a result of an understanding of marriage and family as cultural institutions that have long been used to demarcate LGBTQ individuals as 'other.'

Parenting by LGBTQ Individuals

Raising children is one aspect of LGBTQ family life that has changed in myriad and creative ways. Whereas at one time many LGBTQ individuals, couples, and other family structures were given no avenue for planning to have children or acknowledge their parental status as openly LGBTQ people, LGBTQ community members are increasingly planning families and raising children openly. A recent Williams Institute analysis reported that an estimated 37% of adults who identify as LGBT have had a child during their lives, and an estimated 19% of current same-sex households include children (Gates, 2013). While these statistics should be read with caution due to the challenges inherent to accurate measurement of the LGBTQ population and the limitation of the sample to couples and households only, they provide some sense of the number of individuals who are navigating the complexities of LGBTQ parenting.

Recent data also suggest that same-sex couples are four times more likely to be raising an adopted child than different-sex parents, and they are six times more likely to be raising foster children (Gates, 2013). The data also suggest that the majority of children raised by same-sex couples are biological or 'step' children, forming what are often termed 'gay stepfamilies.'

Indeed, LGBTQ individuals are becoming parents in a diversity of ways. Patterson and Riskind (2010) provide an excellent review of the literature on the ways in which LGBTQ individuals seek pathways to parenting that include stepfamilies, surrogacy, artificial insemination, fostering and adoption, coparenting, and others, as well as the potentially shifting attitudes toward parenting among community members (see also Green, 2012; Tasker and Patterson, 2007). These trends may suggest the LGBTQ parents who participated in these surveys are increasingly planning children within the context of their relationships, as community members increasingly 'come out' earlier in life (Floyd and Bakeman, 2006) and more avenues for parenting become available.

Yet becoming a parent is by no means easy for LGBTQ persons. Scholars have begun to explore the complexity of the LGBTQ parent experience, including both the aspects of it that are universal for parents and others that are unique to negotiations of systems and relationships that offer differing levels of support specific to gender and sexuality issues (Goldberg, 2010, 2012; Goldberg and Allen, 2012; Hines, 2006; Ryan and Berkowitz, 2009). Parents report encountering heterosexism and discrimination both in institutions and relationships. For many, the need to integrate a parenting identity with their sexual and gender identities is a process because parenting is not necessarily an assumed or 'natural' rite in dominant conceptualizations of minority sexuality. Whereas being 'out' with one's identity is often assumed to be an indication of healthy development, disclosure takes on new meaning for LGBTQ persons who become parents and must navigate interactions with agencies, schools, health care settings, government offices, and other organizations and balance the needs and safety of their children against the price of concealment of identity (Tasker and Patterson, 2007). Further, parenting via biological means is often assumed and given precedence, providing an additional challenge for parents who parent through adoption or stepparenting or other means.

These accounts highlight the ways in which LGBTQ parents are navigating uncharted territory and the benefits of their own resiliency and the support of others. Yet many parents report that some of their friendships within the LGBTQ community weaken or dissolve after becoming a parent and that instead, friendships with other parents who are heterosexual are formed (Goldberg, 2012; Green, 2012). These shifts may reflect differing attitudes toward parenting across LGBTQ and heterosexual communities and/or the prioritization of a parenting identity over sexual or gender identity. Parents have also noted that their transitions to parenthood were accompanied by changes in relationships within their families of origin, including both positive and negative ones.

Conclusion

The history of family for LGBTQ individuals communicates, above all, that the need for the love and support offered by family bonds is a universal one. While members of this community historically have not been provided with opportunities to experience family as it is most popularly conceived, they have nonetheless found ways to establish lasting bonds through families of choice. LGBTQ individuals continue to face heterosexism, discrimination, and harassment on the basis of their gender and sexuality today, but as acceptance of LGBTQ persons has increased and legal recognition of the right to experience traditional or constructed family life has begun to be a reality, the meaning of 'family' in the LGBTQ community has continued to evolve. In order to support the LGBTQ community in its diversity, this transition needs to include the right for all LGBTQ individuals to choose their preferred form of family, including both those arrangements that resemble the

traditional family as well as the novel and rich forms of chosen families commonly found in LGBTQ communities.

See also: Family Law in the United States; Family Policy; Family and Kinship, History of; Family as Institution; Family: The Anthropology of the Concept and Its History; Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals; Gay and Lesbian Movements; Heterosexism and Homophobia; Queer Geography; Sexual Minorities and Sexual Citizenship.

Bibliography

Almeida, J., Johnson, R.M., Corliss, H.L., Molnar, B.E., Azreal, D., 2009. Emotional distress among LGBT youth: the influence of perceived discrimination based on sexual orientation. Journal of Youth and Adolescence 38 (7), 1001–1014. http://dx.doi.org/10.1007/10964-009-9408-x.

Ariel, J., 2008. Women aging together in community. Journal of Lesbian Studies 12 (2–3), 283–292. http://dx.doi.org/10.1080/10894160802161463.

Arm, J.R., Horne, S.G., Levitt, H.M., 2009. Negotiating connection to GLBT experience: family members' experience of anti-GLBT movements and policies. Journal of Counseling Psychology 56 (1), 82–96. http://dx.doi.org/10.1037/a0012813.

Arnold, E.A., Bailey, M.M., 2009. Constructing home and family: how the ballroom community supports African American GLBTQ youth in the face of HIV/AIDS. Journal of Gay & Lesbian Social Services 21, 171–188. http://dx.doi.org/10.1080/10538720902772006.

Brown, P.L., 30 October 2007. Gay Enclaves Face Prospect of Being Passé. The New York Times. http://www.nytimes.com/2007/10/30/us/30gay.html?pagewanted=all&_r=0 (accessed 27.12.12.).

Cody, P.J., Welch, P.L., 1997. Rural gay men in northern New England. Journal of Homosexuality 33 (1), 51–67. http://dx.doi.org/10.1300/J082v33n01_04.

Condon, S., 7 December 2012. Supreme Court to Take up Prop. 8, DOMA. CBS News. http://www.cbsnews.com/8301-250_162-57557877/supreme-court-to-take-up-prop-8-doma/ (accessed 29.12.12.).

Dewaele, A., Cox, N., Van den Berghe, W., Vincke, J., 2011. Families of choice?: Exploring the supportive networks of lesbians, gay men, and bisexual. Journal of Applied Social Psychology 41 (2), 312–331. http://dx.doi.org/10.1111/j.1559-1816.2010.00715.x.

Floyd, F.J., Bakeman, R., 2006. Coming-out across the life course: implications of age and historical context. Archives of Sexual Behavior 35, 287–296. http://dx.doi.org/10.1007/s10508-006-9022-x.

Gates, G.J., 2013. LGBT Parenting in the United States. http://williamsinstitute.law.ucla.edu/wp-content/uploads/LGBT-Parenting.pdf (accessed 15.03.13.).

Goldberg, A.E., 2010. Lesbian and Gay Parents and Their Children: Research on the Family Life Cycle. American Psychological Association, Washington, DC.

Goldberg, A.E., 2012. Gay Dads: Transitions to Adoptive Fatherhood. New York University Press, New York/London.

Goldberg, A.E., Allen, K.R. (Eds.), 2012. LGBT-Parent Families: Innovations in Research and Implications for Practice. Springer, New York.

Green, R.-J., 2012. Gay and Lesbian Family Life: Risk, Resilience, and Rising Expectations. In: Walsh, F. (Ed.), Normal Family Processes. Growing Diversity and Complexity Guilford Press, New York/London, pp. 172–195.

Hequembourg, A.L., Brallier, S.A., 2009. An exploration of sexual minority stress across the lines of gender and sexual identity. Journal of Homosexuality 56 (3), 273–298. http://dx.doi.org/10.1080/00918360902728517.

Hines, S., 2006. Intimate transitions: transgender practices of partnering and parenting. Sociology 40 (2), 353–371. http://dx.doi.org/10.1177/0038038506062037.

Horne, S.G., Rostosky, S.S., Riggle, E.D.B., 2011. Impact of marriage restriction amendments on family members of lesbian, gay, and bisexual individuals: a mixed-method approach. Journal of Social Issues 67 (2), 358–375. http://dx.doi.org/10.1111/j.1540-4560.2011.01702.x.

Kubicek, K., McNeeley, M., Holloway, I.W., Weiss, G., Kipke, M.D., 2012. "It's like our own little world": resilience as a factor in participating in the ballroom community subculture (Online publication). AIDS and Behavior. http://dx.doi.org/10.1007/a10461-012-0205-2.

Levine, M.P., 1979. Gay ghetto. Journal of Homosexuality 4 (4), 363–377. http://dx.doi.org/10.1300/J082v04n04_04.

Nadal, K.L., Skolnik, A., Yinglee, W., 2012. Interpersonal and systemic microaggressions toward transgender people: implications for counseling. Journal of LGBT Issues in Counseling 6 (1), 55–82. http://dx.doi.org/10.1080/15538605.2012.648583.

Nesmith, A.A., Burton, D.L., Cosgrove, T.J., 1999. Gay, lesbian, and bisexual youth and young adults: social support in their own words. Journal of Homosexuality 37 (1), 95–108. http://dx.doi.org/10.1300/J082v37n01_07.

Nord, D., 1996. The impact of multiple AIDS-related loss on families of origin and families of choice. American Journal of Family Therapy 24 (2), 129–144. http://dx.doi.org/10.1080/01926189608251026.

Oswald, R.F., 2002. Resilience within the family networks of lesbians and gay men: intentionality and redefinition. Journal of Marriage and Family 64, 374–383. http://dx.doi.org/10.1111/j.1741-3737.2002.00374.x.

Patterson, C.J., 2009. Children of lesbian and gay parents: psychology, law, and policy. The American Psychologist 64 (8), 727–736. http://dx.doi.org/10.1037/0003-066X.64.8.727.

Patterson, C.J., Riskind, R.G., 2010. To be a parent: issues in family formation among gay and lesbian adults. Journal of GLBT Family Studies 6 (3), 326–340. http://dx.doi.org/10.1080/1550428x.2010.490902.

Peterson, D., 2011. Neoliberal homophobic discourse: heteronormative human capital and the exclusion of queer citizens. Journal of Homosexuality 58, 742–757. http://dx.doi.org/10.1080/00918369.2011.581918.

Phillips, G., Peterson, J., Binson, D., Hidalgo, J., Magnus, M., 2011. House/ball culture and adolescent African-American transgender persons and men who have sex with men: a synthesis of the literature. AIDS Care 23, 515–520. http://dx.doi.org/10.1080/09540121.2010.516334.

Riggle, E.D.B., Rostosky, S.S., Horne, S.G., 2010. Does it matter where you live?: Nondiscrimination laws and the experience of LGB residents. Sexuality Research and Social Policy 7, 168–175. http://dx.doi.org/10.1007/s13178-010-0016-z.

Rostosky, S.S., Riggle, E.D.B., Horne, S.G., Miller, A.D., 2009. Marriage amendments and psychological distress in lesbian, gay, and bisexual (LGB) adults. Journal of Counseling Psychology 56 (1), 56–66. http://dx.doi.org/10.1037/a0013609.

Russell, G.M., 2004. The dangers of same-sex marriage referendum for community and individual well-being: a summary of research findings. Angles 7 (1), 1–4.

Russell, G.M., Richards, J.A., 2003. Stressor and resilience factors for lesbians, gay men, and bisexuals confronting antigay politics. American Journal of Community Psychology 31 (3–4), 313–328. http://dx.doi.org/10.1023/A:1023919022811.

Ryan, M., Berkowitz, D., 2009. Constructing gay and lesbian parent families "beyond the closet". Qualitative Sociology 32, 153–172. http://dx.doi.org/10.1007/s11133-009-9124-6.

Tasker, F., Patterson, C.J., 2007. Research on gay and lesbian parenting. Journal of GLBT Family Studies 3 (2–3), 9–34. http://dx.doi.org/10.1300/J461v03n02_02.

Usher, N., Morrison, E., 2010. The demise of the gay enclave, communication infrastructure theory, and the transformation of gay public space. In: Pullen, C., Cooper, M. (Eds.), LGBT Identity and Online New Media. Routledge, New York, pp. 271–287.

Weeks, J., Heaphy, B., Donovan, C., 2001. Same Sex Intimacies: Families of Choice and Other Life Experiments. Routledge, New York.

Weston, K., 1991. Families We Choose: Lesbians, Gays, Kinship. Columbia University Press, New York.

Sexual Motivation

Martha K McClintock, The Institute for Mind and Biology, The University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As in other mammals, human motivation has multiple components each based on different neuroendocrine mechanisms. Its components are tightly integrated in young and middle adulthood, yet develop and age at different times. Thus they are most easily recognized as separable components during childhood and older ages, and in clinical complaints. The components are sexual sensorimotor reflexes and pleasure, gender identity, sexual attractiveness, sexual valence and salience, sexual receptivity, and proactive sexual desire.

Human sexual motivation is a pleasurable subjective state felt by people as sexual desire, libido, lust, sexual interest, drive, passion, and craving (Giles, 2008). It causes a person to act sexually, guides behavior toward achieving sexual goals, to seek sexual objects and partners and persevere when they are not attained. Yet, sexual motivation does not necessarily culminate in sexual behavior, and, in psychoanalytic theory (Freud, 1930), is a force that can be denied or repressed, and also sublimated into positive activities unrelated to sexuality.

Adult sexual motivation is not monolithic. It has multiple components, which are so tightly coupled in adults that it is hard to imagine one without the others. Yet, these components do not develop at the same time, and from birth to adulthood their developmental unfolding reveals that they are indeed separable (McClintock, 2009). Moreover, neuroendocrine and behavioral research in animals reveals that the multiple components have different brain and hormonal mechanisms. In developmental order, they are as follows: (1) *sexual sensorimotor function* enabling genital arousal, sexual responses, and sensuality; (2) *gender identity*, necessary for acquiring both culturally appropriate sex roles including gender-specific sexual behavior and attitudes; (3) *sexual attractiveness*, culturally specific efforts to make oneself sexually attractive to others; (4) *sexual valence or salience*, perceiving particular people, things, or stimuli as sexually attractive, that is, perceiving a person, an image, a song, or any number of things as erotic; (5) *sexual receptivity*, which is the consent and willingness to engage in intimate sexual behavior with another person; and (6) *proactive sexual desire*, that is, wanting and seeking out sexual interactions with another person or oneself.

Sexual motivation arises from a nexus of biological forces, both physiological and evolutionary, psychological emotions and beliefs and molding by social context and cultural constructs. This article reviews the evidence that these biological, psychological, and social forces play different roles for each of the components, again demonstrating that they are separable parts of a functioning whole. The social, behavioral, and neuroendocrine mechanisms of sexual motivation have been explicated extensively in many vertebrate species and manifest many communalities.

Males typically have the capacity to mate at anytime during a breeding season, whereas female mating and sexual motivation is closely tied to a fertile state, whether it is going into heat or ovulating reflexively in response to mating behavior. Human sexual behavior, and that of a few other primates (Fürtbauer et al., 2011; Wallen, 1990), is notable because women can have sex at any time. This evolutionary advance means that pleasure, recreation, and intimacy are additional important functions of sexual motivation, certainly changing and expanding the relative contributions of biology, psychology, and social context. Where possible, differences and similarities between men and women will be noted, as well as variation within each gender.

Sexual Sensorimotor Function

Sexual sensorimotor function is the most foundational component of sexual motivation, in both developmental and evolutionary time frames. Genital stimulation is pleasurable and motivating. Penile erections are sexual reflexes highly conserved across mammals and many vertebrates. In people, erections develop *in utero* and are obviously frequent in newborn boys, indicating intact spinal reflex arcs (Giuliano and Rampin, 2004). A balance of inhibitory and excitatory brain mechanisms regulates erections. In male rats, the classic model for erectile function in men, the paraventricular nucleus of the hypothalamus is key, integrating information from central and autonomic nervous systems and projecting to the brainstem and spinal cord (Argiolas and Melis, 2005). These oxytocin-containing neurons are modulated by a complex set of interacting compounds including testosterone and neuropeptides such as dopamine, GABA, hexarelin, pro-VGF-derived peptides, and potentially melanocortin (Argiolas and Melis, 2003; Shadiack et al., 2007).

A girl's clitoris develops from the same embryological tissue as the penis (Shafik, 1993). Although unstudied, it is likely that clitoral erections also develop early, and even in adulthood, there is a paucity of research; as of 2014, there had been only 32 research articles on clitoral erections listed in PubMed, in striking contrast to 7586 on penile erections. Likewise, the scrotum and labia develop from the same embryological tissue and are homologous structures responding erotically to touch (Puppo, 2013).

These erections and pleasurable genital sensations are initially elicited by many types of touch and stimuli: clothing, moving air, a nursing mother, a caregiver cleaning genitals, and by infants themselves as they begin to explore their own

unclothed bodies. Toddlers masturbate, continuing through age 5 (Bancroft, 2003; Sandfort and Rademakers, 2000) when the practice wanes. At this young age, masturbation is focused on pleasurable physical genital sensations without a sexual relationship or object of desire, and without the other components of adult sexual motivation such as being sexually receptive to another person. Again, there has been little investigation of the role of masturbation in the development of sexuality in girls (Whitfield, 1989).

In mammals, touch and genital responses in infancy and juvenescence are essential for developing normal adult sexual motivation in both males and females (Cooke et al., 2000; Lenz et al., 2008). Clitoral stimulation in rats is rewarding (Parada et al., 2010), supporting the theory that it plays an important role in the genital sensorimotor component of sexual motivation, although Wallen and Lloyd (2008) argue it has no adaptive advantage. Nonetheless, in humans, these pleasurable sensations and sexual reflexes in infancy and childhood likely undergird the adult sexual sensorimotor reflexes studied in exquisite detail by the pioneering work of Masters and Johnson (1966) and recently elaborated by Argiolas and Melis (2003), Basson (2001), and Hyde et al. (2012).

Social context obviously plays a strong role linking the inherently pleasurable and motivating genital sensations with other components of sexual motivation emerging in adolescence and adulthood. In male rats, juvenile social isolation impairs adult sexual behavior (Cooke et al., 2000). In many cultures and social classes, infants discovering pleasant genital sensations are not heavily discouraged, while in others, genital stimulation is shameful and disgraceful, if not taboo, particularly for girls. Thus, from an early age, the interplay between biological reflexive genital responses and strong cultural and moral views shape adult sexual responses and motivation.

Gender Identity

Gender identity as a boy/man or girl/woman enables acquisition of gender-specific sexual motivation, attitudes, and behavior, a learning process that continues throughout adulthood. Gender identity itself, however, does not appear to be learned. Rather, the current view is that it is determined prenatally, likely by testosterone which masculinizes the brain long after and independently of its masculinizing the genitals (Bao and Swaab, 2011).

This is not to argue pure biological determinism. Social and cultural factors do play a large role, establishing what sexual identities and behaviors are considered normative, natural variants, or taboo. Awareness of different kinds of genitals, and ultimately categorizing other people and oneself based on having a penis or a vagina is a learned cultural norm. Interestingly, learning the distinctions appears to develop independently of children's pleasure in their own genital sensations, albeit at similar ages (Bancroft, 2003).

One of the more dramatic examples of the role of social context comes from cultural comparison of people with 5-alpha-reductase deficiency, which causes genetic males (XY) to be categorized as girls at birth because they have labia and a clitoris. At puberty their sex organs change dramatically. Their testes descend and the clitoris enlarges into a penis, enabling adults to father children. Among the Sambia of New Guinea, where this genetic variation is common, it is accepted that some people are men, some women, and others are 'kwolu-aatmwol,' the gender of people born as girls who change to men at puberty, a gender which confers status as a shaman (Herdt, 1990). In the Dominican Republic, where the genetic variant is also common, it was simply accepted before cultural contact with western researchers that some people change gender at puberty. After contact, the Western view of the United States and Europe prevailed, and such people are now viewed as suffering from a pathological type of sexual development and gender dysphoria. Thus social context and culture shapes the meaning of biological mechanisms, determining how an individual experiences their gender identity and orientation, which contributes to the development of sexual motivations that drive their behavior.

Sexual Attractiveness

The desire to *be sexually attractive* to other people is another aspect of sexual motivation (also termed *attractivity*), that is, to arouse sexual desire in others or have sex appeal. Sexual coloration and displays in animals that reproduce sexually are often dramatic and effective in attracting healthier, more fertile, and more parental mates. North African Houbara bustards display dramatic plumage while running and 'booming' to attract mates; male bowerbirds adjust their built bowers and behavioral displays based on responses to the female (Patricelli et al., 2002). Likewise, there is growing evidence that female showiness affects male mate choice (Amundsen, 2000). Female hummingbirds defend a larger and more attractive territory with colorful and energetic displays; among female redwing blackbirds, the intensity of the red epaulets reflects their health condition. But showiness can come at a cost, even though it is a more effective mating strategy; male Houbara bustards with the more dramatic displays age faster than less showy males.

Humans in particular employ a wide range of behaviors to increase their sexual attractiveness, either to a potential sexual partner or socially. Language, of course, depends entirely on social context and culture, but so do nonverbal behaviors including dress and adornments, cosmetics, gestures, eye contact, facial expressions, posture and gait, even displaying altruistic behavior (Barclay, 2010; Elliot and Pazda, 2012; Knapp et al., 2014; Moore, 2010; Townsend and Wasserman, 1997). Often these nonverbal behaviors emphasize the physical traits that distinguish the two sexes, such as waist to hip ratio and musculature, which are more universal cross culturally in their sexual appeal (Singh, 2004; Dixson et al., 2010).

Attractivity and sexual signaling begins developing early in childhood, between ages 3 and 6, when boys often 'flirt' with their mothers and girls with their father, as noted by Freud (1949). Again, sexual attractiveness is developing well before a child has any concept of actively seeking a sexual partner. With time, however, planning and becoming sexually attractive becomes more than a simple 'drive.' It must involve the cerebral cortex, particularly executive function in the prefrontal cortex (PFC), and is modulated by gonadal hormones (Durante et al., 2008; Grammer et al., 2004; Kalat, 2007).

Sexual Valence and Salience

Noticing and feeling that another person is sexually attractive means that they are perceived with a sexual valence, becoming salient and standing out from all others. *Sexual valence and salience* is essential for sexual arousal, which is increasingly viewed as a responsive process to stimuli, rather than a spontaneous, internal drive (Basson, 2001; Both et al., 2011). In motivational theory, people or things that are sexually attractive are termed 'incentive stimuli' imbued with a positive valence, eliciting 'wanting' and activating approach (Hernández-González et al., 2008).

Sexual valence and salience of other people first develops around the age of 10, before puberty is culturally recognized in the United States. At this age, the first memorable sexual attraction or 'crush' may be for a peer, a teacher, or a movie star (McClintock and Herdt, 1996; McClintock, 2009). On average, boys and girls have this experience at the same age (ranging from 6 to 15 years of age), despite the 2-year sex difference in pubertal development between girls and boys (Herdt and McClintock, 2000). Moreover, people with a homosexual orientation also report that this was the age of their first sexual attraction, despite strong social sanctions to be heterosexual.

Maturation of the adrenal glands, or adrenarche, occurs around age 10, and production of the androgen dehydroepiandrosterone rises into the adult range stimulating pubic hair growth and also allowing the experience of finding someone else sexually attractive to become memorable.

The current view is that sexual orientation, i.e., being gay, straight, or bisexual, is typically determined prenatally (but by a different set of neuroendocrine mechanisms than determine gender identity, Bao and Swaab, 2011). Around 10 years of age, the adrenal hormonal milieu activates these prenatally determined brain circuits, and the objects of one's desire become salient, well before a child has developed the proactive sexual motivation or behavior that will later emerge in adolescence. As adults, not only do gays find pictures of same-sex swimsuit models more sexually attractive than the opposite sex, which are attractive to straight heterosexuals, but self-identified bisexuals find both men and women models sexually attractive (Lippa, 2013).

There are sex differences in brain circuits processing such visual sexual stimuli in adults, with men exhibiting greater metabolic activity in the amygdale than do women (Hamann et al., 2004). Women, however, may have been understimulated in this study, being more sensitive to contextual stimuli, particularly emotional expressions of an affectionate relationship (Leitenberg and Henning, 1995; Rupp and Wallen, 2008) as well as observational stance (participant vs observer, Both et al., 2011; Bossio et al., 2014). Nonetheless, there is consistent neurological evidence for sex differences in responses to sexual stimuli (Karama et al., 2002) as well as effects of gonadal steroids such as testosterone and estradiol (Alexander and Sherwin, 1993; Roney and Simmons, 2008; Rupp et al., 2009; Welling et al., 2007, 2008). In all studies, including neuroendocrine studies of rodent models, the PFC is critical for processing sexual valence and salience. In the rat, it is the orbital PFC that participates in the processing of olfactory sexual signals, while the medial PFC is involved in motive-emotional processing, temporal sequencing, and memory (Hernández-González et al., 2008).

Sexual Receptivity

Sexual receptivity is the willingness to have sex with a partner when given the opportunity. People, particularly women, can be sexually receptive prior to being sexually aroused (the genital sensorimotor component). In this case, arousal builds afterward during sexual activity, rather than before it. This sequence of components contrasts with Masters and Johnson's classic unidirectional linear model of sexual behavior in which desire precedes being receptive and sexual arousal. The need to reconceptualize sexual motivation has been driven by the fact that women in particular may be receptive to engaging in sexual activity for a variety of reasons different from sexual desire or arousal, such as a need for emotional closeness, gratitude, duty, or payment (Basson, 2000; Meston, 2001).

In women, cultural context has different effects on sexual receptivity and its relationship with the other components of sexual motivation. East Asian women living in Canada are less likely than Euro-Canadians to report being sexually receptive to their partners. Among East Asian women, mainstream acculturation is not associated with greater sexual receptivity even though acculturation is associated with greater sexual arousal. Perhaps the East Asian women in these studies were less comfortable reporting sexual receptivity than arousal. Nonetheless, the relative resistance of sexual receptivity to Westernization is striking.

In contrast, the first study of sexuality of Indian women reveals that their sexual receptivity is high. Among 100 married Indian women, fully 75% reported accepting their husband's sexual advances with pleasure, although none initiated sexual activity (Avasthi et al., 2008). Married Han women in mainland China exhibited a different pattern. Greater sexual receptivity was associated with more sexual initiation (Han et al., 2014). Taken together, this cross-cultural comparison supports the independence of sexual receptivity from other components of sexual motivation such as arousal and sexual initiation.

Elegant neuroendocrine research in mammals has elucidated the mechanisms of sexual receptivity, focusing on the female's adopting a mating posture in response to the male, rather than fighting him off. In many species, this is a lordosis reflex, which raises the vagina to facilitate copulation. This reflex is enhanced by estrogen released by ripening ovarian follicles that act peripherally to enlarge the genital sensory field. It also binds in neurons of the ventrolateral portion of the ventromedial hypothalamus, part of a well-characterized system including the posterodorsal medial amygdala, bed nucleus of the stria terminalis, medial preoptic nucleus, and the arcuate nucleus (Micevych and Dewing, 2011). The reflex is regulated within this system by a panoply of neuropeptides including glutamate, GABA, oxytocin, cholecystokinin, and generalized CNS arousal, such as norepinephrine (Flanagan-Cato, 2011; Lee and Pfaff, 2008).

The cultural and neuroendocrine mechanisms of sexual receptivity in men and in male animal models is understudied, perhaps because it plays a small role, but more likely reflecting

a cultural bias. In Soay sheep, rams are more receptive to some ewes that are in heat than others, and thereby choose a mate. Male rats and mice are also more receptive to some females than others, exhibiting large individual differences in how responsive they are to a female in heat (Bharadwaj et al., 2013; McClintock and Adler, 1978).

Proactive Sexual Desire and Solicitation

Proactive sexual desire is not only wanting the reward of erotic pleasure but also making an effort to initiate sexual activity with a partner or by masturbating. Historically, the focus of neuroendocrine research had been on the 'appetitive' sexual motivation of male rodents and not on females. It then became evident that female rodents also actively solicit mating, in addition to being receptive to the male's advances, and actively pace copulation, competing to get the ejaculate of particular males (McClintock, 1984). A new term, 'proceptive,' was proposed for this behavior, as if the female's active sexual motivation was somehow different from the male. A better solution is employing 'proactive sexual desire' as a gender-neutral alternative to reference actively seeking erotic sexual activity by either males or females.

In females, proactive sexual desire has different neuroendocrine mechanisms than sexual receptivity in rodents, nonhuman primates, and humans (Coria-Avila and Pfaus, 2007; López et al., 2009). In female rodents, progesterone from the ovary and brain, in addition to estradiol, plays a key role in acting on the preoptic nucleus (Micevych and Dewing, 2011; Sakuma, 1995). Proceptive solicitation relies on an intact preoptic area, whereas its ablation enhances lordosis, the receptive component. Comparably detailed neuroendocrine mechanisms distinguishing male receptive and proactive sexual motivation have not been done because male receptivity is understudied.

Ovarian hormones also increase proactive sexual desire in women. During the preovulatory sexual phase of the menstrual cycle, women report increased sexual desire, sexual fantasies and dreams, and masturbation (Bullivant et al., 2004; Gangestad and Thornhill, 2008; Tarín and Gómez-Piquer, 2002). However, in sharp contrast to mammals that must be in heat to actively seek mating, this periovulatory increase in women is superimposed on proactive sexual desire sustained levels throughout the ovulatory menstrual cycle. Likewise, in female nonhuman primates that also mate when not fertile, sexual initiation, rather than copulation, is arguably the best behavioral indicator of female proactive sexual desire (Wallen, 1990, 2013).

In hypogonadal men, testosterone increases sexual desire (Finkelstein et al., 2013) as it does in most male mammals. Likewise, testosterone regulates proactive sexual motivation in male rats (Ågmo, 2003; Hamson and Watson, 2004), increasing dendritic spine number of neurons in the medial preoptic area of the hypothalamus in Syrian hamsters (Garelick and Swann, 2014). Androgens may also contribute to women's sexual desire. They are produced tonically in women, as they are in men, from both the ovary and the adrenal. They rise slightly prior to ovulation. As a drug, they benefit some women suffering from hypoactive sexual desire (Davis and Worsley, 2013). However, androgens are not sufficient as they are in men, and their effects on sexual desire require a background of estradiol (Wallen, 2013).

In men and women, testosterone and social interactions with potential partners form a positive feedback loop. Brief social interactions with young women increase testosterone in men, which in turn increases desire (Roney et al., 2003; Roney et al., 2007). Likewise in women, watching the courtship of an attractive man raises their testosterone (López et al., 2009), affecting their sexual desire as well.

Evolutionary theory posits that the sex (i.e., male or female) with less investment in conception and parenting will have higher levels of proactive sexual behaviors than the sex with a large investment, which will benefit by being more selective and choosing their partner, often by being receptive or not. In many mammals, the male fills this role, and is evident among people in many settings (Avasthi et al., 2008; Hendrie et al., 2009).

Women also initiate sociosexual encounters verbally and nonverbally, with a variety of culturally specific behaviors. In the contemporary United States, such behaviors include leaning forward, showing a palm, primping, short darting glance, lip lick, caressing an object, and head toss (Moore, 2002; Perper and Weis, 1987). These behaviors are more common around the time of ovulation even when women are unaware of their own fertile state.

Hormones and other compounds affecting sexuality can have a stronger effect on sexual desire and fantasies, measures of their motivation, rather than on sexual activity, which is obviously constrained by the availability of a sexual partner (Spencer et al., 2004). Likewise, partner availability determines whether proactive sexual motivation is experienced as desire or fantasies. If a woman has a regular sexual partner, then increased sexual motivation can be experienced as a desire for that particular partner. If without a partner, then she is more likely to create one through sexual fantasy.

Sexual desire develops around puberty in both boys and girls, manifest as 'wet dreams.' In girls, they are likely more common around the time of ovulation, as they are in women. But again, understanding at what age such dreams begin is a taboo topic, particularly in girls, precluding research on this natural marker of coordination among components of sexual motivation.

Biologically, puberty lasts a long time, during which the components of sexual motivation become fully integrated. It begins before the first ejaculate and menses that herald the capacity to conceive children and does not end until peak reproductive maturity is attained in the late teens or early 20s, and in women technically not until the first pregnancy and lactation. The capacity for partnerships and social roles develop and become intertwined with all components of sexual motivation, laying the foundation for resilient sexuality that can be sustained through diverse lifespan trajectories and into old age.

Diverse Scales for Measuring Human Sexual Motivation in Men and Women

The ideal study of human sexual motivation would measure each of its distinct components within a person. Simultaneous measurements would reveal the interactions among the components, as well as their differential sensitivity to a person's

cultural, social, psychological, and biological milieu. At the moment, there is no single scale for measuring all components of sexual motivation. Nonetheless, scales have been published for each of the components; Table 1 is arrayed to highlight each component of sexual motivation and to list the published scales focused on that component. Rather than arguing that some types of scales are better than others, a more profitable approach recognizes that they are often focused on different components. Ideally, future research will create a composite scale to enable the simultaneous study of multiple components.

The extant scales are often gender specific and focus on the most common clinical complaints by men and women. For men, the most common focus is on the sensorimotor component, particularly erectile dysfunction and premature ejaculation. For women, scales focus on disorders of hypoactive sexual desire, with only a few distinguishing between the receptive and proactive components of sexual motivation (see Table 1). Modifications of the scales would include sensorimotor questions appropriate for women and clearly distinguish the receptive and proactive components in both men and women.

The power of this approach is demonstrated by the scales that do include items assessing both the sexual sensorimotor component and the proactive sexual desire component, sometimes including receptivity (see Table 1). Among women, the multidimensional structure of the scales distinguished women with and without sexual difficulties. For example, in a normal population, orgasm frequency and ease (the sensorimotor reward driving sexual motivation) was the biggest contributor to the largest factor (eigen value 9.06), but contributed only to a much smaller factor (eigen value = 1.87) among women with sexual dysfunction. Sexual arousal was part of the largest factor in both groups, but among normal women arousal was associated with orgasm while it was associated with desire among women with dysfunction (Wiegel et al., 2005).

There were also national/cultural differences in the factor structure. In Chinese Han women of mainland China, frequency of sexual thoughts were not associated with desire, as they were in a North American population, but rather constituted a smaller factor explaining only 7.66% of the variance. Thus, the reported lower sexual desire among Canadian Asian women may result from utilizing the North American subscale which combines reported desire with thoughts and fantasies. That is, the levels of desire may well be the same, but not expressed as reporting sexual fantasies and thoughts.

Even when measuring the same aspect of sexual motivation, some scales frame items positively and others as a dysfunction. The BISF-W asks "how frequently have you become aroused during the following activities?" (Mazer et al., 2000). The Natsal-SF asks, if you have "felt no excitement or arousal during sex" for a period of 3 months or longer (Mitchell et al., 2012). The BISF-W is thus better suited to assessing a wide range of variation in normal sexual function as well as dysfunction, whereas the Natsal-SF focuses on prevalence of sexual problems.

The scales designed for both men and women typically focus on sexual difficulties. In the sensorimotor component, these included impotence, poor physical arousal, trouble lubricating, and problems with orgasm. The proactive component is assessed by questions about lacking sexual enjoyment or desire. Positive elements included sexual thoughts and drive (see Table 1).

The other components of sexual motivation are not included in general scales for sexual motivation. Thus, separate scales are necessary for each. Gender identity is well measured by Freund's scales, assessing variation in behaviors, self-concept, and wishes typical of the opposite sex (see Table 1). Barelds et al. (2011) developed a scale to measure how physically attractive a person perceives themselves to be as well as to their partner and the general public. Diamond (2011) measures sexual valence and salience, a key component of sexual orientation, not by asking about behavior as Kinsey did, but by focusing on sexual attraction to both men and women, providing a rich measure of a continuum of attraction including degrees of bisexuality.

Sexual receptivity is measured by specific items in scales by Mazer et al. (2000) and Han et al. (2014), although their coding system treats receptivity as part of a continuum with sexual initiation (see Table 1). Nonetheless, single item analyses are supported. Receptivity can also be measured by conceptualizing it in the context of Humphreys and Brousseau's Sexual Consent Scale (2010).

Finally, proactive sexual desire can be measured separately from direct questions about sexual sensorimotor arousal by assessing content of daydreams with the Sexual Daydreaming Scale (Purifoy et al., 1992). Recognizing that sexual assertiveness can be driven not only by sexual desire and motivation, but also by other motivations such as power and reinforcing emotional intimacy, Anderson and Newton (2011) developed the Sexually Assertive Behavior Scale to tease out these diverse motivations among women (see Table 1).

Summary and DSM 5

The argument for components of sexual motivation is strong. The components develop at different ages. They have different neuroendocrine mechanisms. There are strong cultural differences in the associations among components. One component may be dysfunctional while the others are intact. Even though men and women have all of the same components, there are sex differences in neuroendocrine mechanisms that support them and which are most vulnerable to dysfunction.

Nonetheless, the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM 5) has collapsed these components in their criteria for sexual dysfunction among women (IsHak and Gabriel, 2013). In DSM 4, the female disorders of arousal (sensorimotor component) and desire (proactive sexual desire component) were separate. The DSM 5 now has a single diagnosis called 'female sexual interest/arousal disorder.' This amalgamation was intended to recognize the striking sex difference in the temporal relationship between desire and arousal. In men, the pattern is typically linear, with desire preceding arousal, as described by Masters and Johnson (1966). In women, however, they can occur in the opposite sequence, building upon each other.

Not all on the DSM 5 committee agreed that this warranted combining these two components to define a single diagnostic category, especially because the category also includes the other components of sexual motivation delineated in this article: few erotic thoughts or fantasies, low initiation of sexual activity, low responsiveness to a partner's sexual

Table 1 Scales for measuring components of sexual motivation

Scales					*Components of sexual motivation*					
Scale name	*Reference*	*Gender*	*Target population*	*Composite score*	*Sexual sensorimotor function*	*Gender identity*	*Sexual attractiveness*	*Sexual valence and salience*	*Sexual receptivity*	*Proactive sexual desire solicitation*
Brief Index of Sexual Functioning of Women (BISF–W) updated	Mazer et al., 2000; based on Taylor Rosen and Leiblum, 1994	Women	North American, with and without sexual dysfunction	Norms available	Arousal; pleasure/orgasm; difficulties	N.A.	N.A.	N.A.	Thoughts/desire (D1), receptivity/initiation combined (D4): initiation(8), receptivity (9), and desired frequency (12)	
Chinese Brief Index of Sexual Functioning of Women (C-BISF-W)	Han et al., 2014	Women	Mainland China, with and without depression	Validated	Arousal; orgasm; difficulties	N.A.	N.A.	N.A.	Desire only (D1), receptivity/initiation (D6) also includes experiencing pleasure (10) and being able to communicate desires (17)	
Female Sexual Function Index (FSFI)	Wiegel et al., 2005	Women	North American clinic, with and without sexual dysfunction	Reliability, convergent, and discriminant validity	Arousal/orgasm, lubrication, pain;	N.A.	N.A.	N.A.	N.A.	Desire/arousal
British National Survey of Sexual Attitudes and Lifestyles-Sexual Function (Natsal-SF)	Mitchell et al., 2012	Men and women	Survey of community dwellers in Britain	Validates with general population and clinical samples	Negative items; no excitement, trouble with climax	N.A.	N.A.	N.A.	N.A.	Negative items, lacked enjoyment or interest
Derogatis Interview for Sexual Functioning (DISF)	Derogatis and Melisaratos, 1979	Men and women	North American and European community norms	Validated	Arousal, orgasm	N.A.	N.A.	N.A.	N.A.	Thoughts, sex drive
Detailed Assessment of Sexual Arousal (DASA)	Basson and Brotto, 2003	Men and women	North American clinic, with and without sexual dysfunction	Screener	Subjective, pleasure and physical arousal	N.A.	N.A.	N.A.	N.A.	N.A.

Feminine Gender Identity Scale (FGIS)	Freund et al., 1977	Men	North American heterosexuals, homosexuals, and MF transexuals	Validated	N.A.	Variation among men in behaviors, self-concept, and wishes	N.A.	N.A.	N.A.	N.A.
Masculine Gender Identity Scale for Females (MGISF)	Blanchard and Freund, 1983	Women	North American heterosexuals, homosexuals, and FM transexuals	Validated	N.A.	Variation among women in behaviors, self-concept, and wishes	N.A.	N.A.	N.A.	N.A.
Assessment of Attractiveness	Barelds et al., 2011	Men and women	Dutch heterosexual couples	Validated	N.A.	N.A.	Attractive to self, partner, and general population	N.A.	N.A.	N.A.
Assessing Multiple Facets of Attraction to Women and Men	Diamond, 2011	Men and women	North American heterosexuals and homosexuals	Validated	N.A.	N.A.	N.A.	Frequency and intensity of attraction to men and women	N.A.	N.A.
Sexual Consent Scale Revised (SCS-R)	Humphreys and Brousseau, 2010	Men and women	Heterosexual undergraduates North America	Validated	N.A.	N.A.	N.A.	N.A.	Consent to sexual invitation	N.A.
Sexual Daydreaming Scale	Purifoy et al., 1992	Men and women	Homosexual and heterosexual throughout lifespan	Validated	N.A.	N.A.	N.A.	N.A.	N.A.	Multiple facets of sexuality, correlated with daydreaming in general
Sexually Assertive Behavior Scale (SABS)	Anderson and Newton, 2011	Women	North American	Internal consistency, reliability	N.A.	N.A.	N.A.	N.A.	N.A.	Discriminates sexual and other motivations

invitation, low response to sexual cues whether genital or not, and low physical excitement and pleasure and being nonresponsive during sexual activity. Nonetheless, the diagnostic criteria require only three out of six symptoms, not all of them, recognizing that they are indeed separable. Therefore, the 'lumping' of the components of sexual motivation in DSM 5 emphasizes that they are normally tightly coupled in typical sexual functioning, and that dysfunction arises when they are no longer coordinated.

See also: Adolescent Sexual Risk; Biomarkers in Demographic Research; Erotic Capital; Heterosexual Expression at Older Ages; Heterosexuality in Africa; Heterosexuality in China; Heterosexuality; Hooking up and Casual Sex; Methodological Issues in the Study of Sexuality; Sex Life Challenges: The Finnish Case; Sexual Dysfunction, Epidemiology of; Sexuality Over the Life Course; Sexuality in Australia; Sexuality, Theories of; Sexuality: Cultural Aspects.

Bibliography

Ågmo, A., 2003. Unconditioned sexual incentive motivation in the male Norway rat (*Rattus norvegicus*). Journal of Comparative Psychology (Washington, D.C.: 1983) 117 (1), 3–14.

Alexander, G.M., Sherwin, B.B., 1993. Sex steroids, sexual behavior, and selection attention for erotic stimuli in women using oral contraceptives. Psychoneuroendocrinology 18 (2), 91–102.

Amundsen, T., 2000. Why are female birds ornamented? Trends in Ecology & Evolution 15 (4), 149–155.

Anderson, P.B., Newton, M., 2011. Sexually assertive behavior scale. In: Fisher, T.D., Davis, C.M., Yarber, W.L., Davis, S.L. (Eds.), Handbook of Sexuality-Related Measures, third ed. Routledge, New York.

Argiolas, A., Melis, M.R., 2003. The neurophysiology of the sexual cycle. Journal of Endocrinological Investigation 26 (3 Suppl.), 20–22.

Argiolas, A., Melis, M.R., 2005. Central control of penile erection: role of the paraventricular nucleus of the hypothalamus. Progress in Neurobiology 76 (1), 1–21.

Avasthi, A., Kaur, R., Prakash, O., Banerjee, A., Kumar, L., Kulhara, P., 2008. Sexual behavior of married young women: a preliminary study from North India. Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine 33 (3), 163–167.

Bancroft, J. (Ed.), 2003. Sexual Development in Childhood. Indiana University Press, Bloomington, Indiana.

Bao, A.-M., Swaab, D.F., 2011. Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorders. Frontiers in Neuroendocrinology 32 (2), 214–226.

Barclay, P., 2010. Altruism as a courtship display: some effects of third-party generosity on audience perceptions. British Journal of Psychology (London, England: 1953) 101 (Pt 1), 123–135.

Barelds, D.P.H., Dijkstra, P., Koudenburg, N., Swami, V., 2011. An assessment of positive illusions of the physical attractiveness of romantic partners. Journal of Social and Personal Relationships 28 (5), 706–719.

Basson, R., 2000. The female sexual response: a different model. Journal of Sex & Marital Therapy 26 (1), 51–65.

Basson, R., 2001. Human sex-response cycles. Journal of Sex & Marital Therapy 27 (1), 33–43.

Basson, R., Brotto, L.A., 2003. Sexual psychophysiology and effects of sildenafil citrate in oestrogenised women with acquired genital arousal disorder and impaired orgasm: a randomised controlled trial. BJOG 110 (11), 1014–1024. PubMed PMID: 14592587.

Bharadwaj, P., McInnis, C., Madden, A.M.K., Bonthuis, P.J., Zup, S., Rissman, E.F., Park, J.H., 2013. Increased dendritic spine density and tau expression are associated with individual differences in steroidal regulation of male sexual behavior. PloS One 8 (7), e69672.

Blanchard, R., Freund, K., 1983. Measuring masculine gender identity in females. Journal of Consulting and Clinical Psychology 51 (2), 205–214.

Bossio, J.A., Spape, J., Lykins, A.D., Chivers, M.L., 2014. Observational stance as a predictor of subjective and genital sexual arousal in men and women. Journal of Sex Research 51 (3), 303–315.

Both, S., Laan, E., Everaerd, W., 2011. Focusing "hot" or focusing "cool": attentional mechanisms in sexual arousal in men and women. The Journal of Sexual Medicine 8 (1), 167–179.

Bullivant, S.B., Sellergren, S.A., Stern, K., Spencer, N.A., Jacob, S., Mennella, J.A., McClintock, M.K., 2004. Women's sexual experience during the menstrual cycle: identification of the sexual phase by noninvasive measurement of luteinizing hormone. Journal of Sex Research 41 (1), 82–93.

Cooke, B.M., Chowanadisai, W., Breedlove, S.M., 2000. Post-weaning social isolation of male rats reduces the volume of the medial amygdala and leads to deficits in adult sexual behavior. Behavioural Brain Research 117 (1–2), 107–113.

Coria-Avila, G.A., Pfaus, J.G., 2007. Neuronal activation by stimuli that predict sexual reward in female rats. Neuroscience 148 (3), 623–632.

Davis, S.R., Worsley, R., 2013. Androgen treatment of postmenopausal women. The Journal of Steroid Biochemistry and Molecular Biology.

Derogatis, L.R., Melisaratos, N., 1979. The DSFI: a multidimensional measure of sexual functioning. Journal of Sex & Marital Therapy 5 (3), 244–281.

Diamond, L.M., 2011. Assessing multiple facets of attraction to women and men. In: Fisher, T.D., Davis, C.M., Yarber, W.L., Davis, S.L. (Eds.), Sexuality-Related Measures, third ed. Routledge, New York.

Dixson, B.J., Dixson, A.F., Bishop, P.J., Parish, A., 2010. Human physique and sexual attractiveness in men and women: a New Zealand-U.S. comparative study. Archives of Sexual Behavior 39 (3), 798–806.

Durante, K.M., Li, N.P., Haselton, M.G., 2008. Changes in women's choice of dress across the ovulatory cycle: naturalistic and laboratory task-based evidence. Personality & Social Psychology Bulletin 34 (11), 1451–1460.

Elliot, A.J., Pazda, A.D., 2012. Dressed for sex: red as a female sexual signal in humans. PloS One 7 (4), e34607.

Finkelstein, J.S., Lee, H., Burnett-Bowie, S.-A.M., Pallais, J.C., Yu, E.W., Borges, L.F., Jones, B.F., Barry, C.V., Wulczyn, K.E., Thomas, B.J., Leder, B.Z., 2013. Gonadal steroids and body composition, strength, and sexual function in men. The New England Journal of Medicine 369 (11), 1011–1022.

Flanagan-Cato, L.M., 2011. Sex differences in the neural circuit that mediates female sexual receptivity. Frontiers in Neuroendocrinology 32 (2), 124–136.

Freud, S., 1930. Civilisation and its discontents. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud – The Future of an Illusion, Civilization and its Discontents, and Other Works. Hogarth Press, London, pp. 79–80.

Freud, S., 1949. The Ego and the Id. Hogarth Press, London.

Freund, K., Langevin, R., Satterberg, J., Steiner, B., 1977. Extension of the gender identity scale for males. Archives of Sexual Behavior 6 (6), 507–519.

Fürtbauer, I., Mundry, R., Heistermann, M., Schülke, O., Ostner, J., 2011. You mate, i mate: macaque females synchronize sex not cycles. PloS One 6 (10), e26144.

Gangestad, S.W., Thornhill, R., 2008. Human oestrus. Proceedings. Biological Sciences/The Royal Society 275 (1638), 991–1000.

Garelick, T., Swann, J., 2014. Testosterone regulates the density of dendritic spines in the male preoptic area. Hormones and Behavior 65 (3), 249–253.

Giles, J., 2008. The Nature of Sexual Desire. University Press of America, Lanham, Maryland.

Giuliano, F., Rampin, O., 2004. Neural control of erection. Physiology & Behavior 83 (2), 189–201.

Grammer, K., Renninger, L., Fischer, B., 2004. Disco clothing, female sexual motivation, and relationship status: is she dressed to impress? Journal of Sex Research 41 (1), 66–74.

Hamann, S., Herman, R.A., Nolan, C.L., Wallen, K., 2004. Men and women differ in amygdala response to visual sexual stimuli. Nature Neuroscience 7 (4), 411–416.

Hamson, D.K., Watson, N.V., 2004. Regional brainstem expression of Fos associated with sexual behavior in male rats. Brain Research 1006 (2), 233–240.

Han, Z., Gan, Z., Han, H., Chen, J., Li, K., Guan, N., 2014. Validity of the Chinese version of the brief index of sexual functioning for women with a new scoring algorithm and comparison of normative and recurrently depressed Han Chinese population. The Journal of Sexual Medicine 11 (2), 439–446.

Hendrie, C., Mannion, H., Godfrey, G., 2009. Evidence to suggest that nightclubs function as human sexual display grounds. Behaviour 146 (10), 1331–1348.

Herdt, G., 1990. Mistaken gender: 5-alpha reductase hermaphroditism and biological reductionism in sexual identity reconsidered. American Anthropologist 92 (2), 433–446.

Herdt, G., McClintock, M., 2000. The magical age of 10. Archives of Sexual Behavior 29 (6), 587–606.

Hernández-González, M., Guevara, M.A., Ågmo, A., 2008. Motivational influences on the degree and direction of sexual attraction. Annals of the New York Academy of Sciences 1129, 61–87.
Humphreys, T.P., Brousseau, M.M., 2010. The sexual consent scale-revised: development, reliability, and preliminary validity. Journal of Sex Research 47 (5), 420–428.
Hyde, J.S., Byers, E.S., DeLamater, J.D. (Eds.), 2012. Understanding Human Sexuality, fifth ed. McGraw-Hill Ryerson, p. 620.
IsHak, W.W., Gabriel, T., 2013. DSM-5 changes in diagnostic criteria of sexual dysfunctions. Reproductive System & Sexual Disorders 02 (02), 1–3.
Kalat, J.W., 2007. Biological Psychology. Thomson/Wadsworth, Belmont, CA, p. 582.
Karama, S., Lecours, A.R., Leroux, J.-M., Bourgouin, P., Beaudoin, G., Joubert, S., Beauregard, M., 2002. Areas of brain activation in males and females during viewing of erotic film excerpts. Human Brain Mapping 16 (1), 1–13.
Knapp, M.L., Hall, J.A., Horgan, T.G., 2014. Nonverbal Communication in Human Interaction. Cengage Learning, p. 528.
Lee, A.W., Pfaff, D.W., 2008. Hormone effects on specific and global brain functions. The Journal of Physiological Sciences: JPS 58 (4), 213–220.
Leitenberg, H., Henning, K., 1995. Sexual fantasy. Psychological Bulletin 117 (3), 469–496.
Lenz, K.M., Graham, M.D., Parada, M., Fleming, A.S., Sengelaub, D.R., Monks, D.A., 2008. Tactile stimulation during artificial rearing influences adult function and morphology in a sexually dimorphic neuromuscular system. Developmental Neurobiology 68 (4), 542–557.
Lippa, R.A., 2013. Men and women with bisexual identities show bisexual patterns of sexual attraction to male and female "swimsuit models". Archives of Sexual Behavior 42 (2), 187–196.
López, H.H., Hay, A.C., Conklin, P.H., 2009. Attractive men induce testosterone and cortisol release in women. Hormones and Behavior 56 (1), 84–92.
Masters, W.H., Johnson, V.E., 1966. Human Sexual Response. Bantam Books, Toronto; New York, p. 363.
Mazer, N.A., Leiblum, S.R., Rosen, R.C., 2000. The brief index of sexual functioning for women (BISF-W): a new scoring algorithm and comparison of normative and surgically menopausal populations. Menopause (New York, N.Y.) 7 (5), 350–363.
McClintock, M.K., 1984. Group mating in the domestic rat as a context for sexual selection: consequences for the analysis of sexual behavior and neuroendocrine responses. In: Rosenblatt, J.S. (Ed.), Advances in the Study of Behavior. Academic Press, New York, pp. 1–50.
McClintock, M.K., 2009. Sexual development: desire and behavior. In: Shweder, R.A., Bidell, T.R., Dailey, A.C., Dixon, S.D., Miller, P.J., Modell, J. (Eds.), The Child: An Encyclopedic Companion. University of Chicago Press, Chicago.
McClintock, M.K., Adler, N.T., 1978. The role of the female during copulation in wild and domestic Norway rats (*Rattus Norvegicus*). Behaviour 67 (1), 67–95.
McClintock, M.K., Herdt, G., 1996. Rethinking puberty: the development of sexual attraction. Current Directions in Psychological Science 5 (6), 178–183.
Meston, C.M., 2001. Receptivity and personal distress: considerations for redefining female sexual dysfunction. Journal of Sex & Marital Therapy 27 (2), 179–182.
Micevych, P.E., Dewing, P., 2011. Membrane-initiated estradiol signaling regulating sexual receptivity. Frontiers in Endocrinology 2, 26.
Mitchell, K.R., Ploubidis, G.B., Datta, J., Wellings, K., 2012. The Natsal-SF: a validated measure of sexual function for use in community surveys. European Journal of Epidemiology 27 (6), 409–418.
Moore, M.M., 2002. Courtship communication and perception. Perceptual and Motor Skills 94 (1), 97–105.
Moore, M.M., 2010. Human nonverbal courtship behavior–a brief historical review. Journal of Sex Research 47 (2), 171–180.
Parada, M., Chamas, L., Censi, S., Coria-Avila, G., Pfaus, J.G., 2010. Clitoral stimulation induces conditioned place preference and Fos activation in the rat. Hormones and Behavior 57 (2), 112–118.
Patricelli, G.L., Uy, J.A.C., Walsh, G., Borgia, G., 2002. Male displays adjusted to female's response. Nature 415 (6869), 279–280.
Perper, T., Weis, D.L., 1987. Proceptive and rejective strategies of U.S. and Canadian college women. Journal of Sex Research 23 (4), 455–480.
Puppo, V., 2013. Anatomy and physiology of the clitoris, vestibular bulbs, and labia minora with a review of the female orgasm and the prevention of female sexual dysfunction. Clinical Anatomy (New York, N.Y.) 26 (1), 134–152.
Purifoy, F.E., Grodsky, A., Giambra, L.M., 1992. The relationship of sexual daydreaming to sexual activity, sexual drive, and sexual attitudes for women across the life-span. Archives of Sexual Behavior 21 (4), 369–385.
Roney, J.R., Lukaszewski, A.W., Simmons, Z.L., 2007. Rapid endocrine responses of young men to social interactions with young women. Hormones and Behavior 52 (3), 326–333.
Roney, J.R., Mahler, S.V., Maestripieri, D., 2003. Behavioral and hormonal responses of men to brief interactions with women. Evolution and Human Behavior 24 (6), 365–375.
Roney, J.R., Simmons, Z.L., 2008. Women's estradiol predicts preference for facial cues of men's testosterone. Hormones and Behavior 53 (1), 14–19.
Rupp, H.A., James, T.W., Ketterson, E.D., Sengelaub, D.R., Janssen, E., Heiman, J.R., 2009. Neural activation in women in response to masculinized male faces: mediation by hormones and psychosexual factors. Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society 30 (1), 1–10.
Rupp, H.A., Wallen, K., 2008. Sex differences in response to visual sexual stimuli: a review. Archives of Sexual Behavior 37 (2), 206–218.
Sakuma, Y., 1995. Differential control of proceptive and receptive components of female rat sexual behavior by the preoptic area. The Japanese Journal of Physiology 45 (2), 211–228.
Sandfort, T., Rademakers, J. (Eds.), 2000. Childhood Sexuality: Normal Sexual Behavior and Development. Haworth Press, New York, p. 136.
Shadiack, A.M., Sharma, S.D., Earle, D.C., Spana, C., Hallam, T.J., 2007. Melanocortins in the treatment of male and female sexual dysfunction. Current Topics in Medicinal Chemistry 7 (11), 1137–1144.
Shafik, A., 1993. Vaginocavernosus reflex. Clinical significance and role in sexual act. Gynecologic and Obstetric Investigation 35 (2), 114–117.
Singh, D., 2004. Mating strategies of young women: role of physical attractiveness. Journal of Sex Research 41 (1), 43–54.
Spencer, N.A., McClintock, M.K., Sellergren, S.A., Bullivant, S., Jacob, S., Mennella, J.A., 2004. Social chemosignals from breastfeeding women increase sexual motivation. Hormones and Behavior 46 (3), 362–370.
Tarín, J.J., Gómez-Piquer, V., 2002. Do women have a hidden heat period? Human Reproduction (Oxford, England) 17 (9), 2243–2248.
Taylor, J.F., Rosen, R.C., Leiblum, S.R., 1994. Self-report assessment of female sexual function: psychometric evaluation of the Brief Index of Sexual Functioning for Women. Arch Sex Behav. 23 (6), 627–643. PubMed PMID: 7872859.
Townsend, J.M., Wasserman, T., 1997. The perception of sexual attractiveness: sex differences in variability. Archives of Sexual Behavior 26 (3), 243–268.
Wallen, K., 1990. Desire and ability: hormones and the regulation of female sexual behavior. Neuroscience and Biobehavioral Reviews 14 (2), 233–241.
Wallen, K., 2013. Women are not as unique as thought by some: comment on "hormonal predictors of sexual motivation in natural menstrual cycles", by Roney and Simmons. Hormones and Behavior 63 (4), 634–635.
Wallen, K., Lloyd, E.A., 2008. Clitoral variability compared with penile variability supports nonadaptation of female orgasm. Evolution & Development 10 (1), 1–2.
Welling, L.L.M., Jones, B.C., DeBruine, L.M., Conway, C.A., Law Smith, M.J., Little, A.C., Feinberg, D.R., Sharp, M.A., Al-Dujaili, E.A.S., 2007. Raised salivary testosterone in women is associated with increased attraction to masculine faces. Hormones and Behavior 52 (2), 156–161.
Welling, L.L.M., Jones, B.C., DeBruine, L.M., Smith, F.G., Feinberg, D.R., Little, A.C., Al-Dujaili, E.A.S., 2008. Men report stronger attraction to femininity in women's faces when their testosterone levels are high. Hormones and Behavior 54 (5), 703–708.
Whitfield, M., 1989. Development of sexuality in female children and adolescents. Canadian Journal of Psychiatry. Revue canadienne de psychiatrie 34 (9), 879–883.
Wiegel, M., Meston, C., Rosen, R., 2005. The female sexual function index (FSFI): cross-validation and development of clinical cutoff scores. Journal of Sex & Marital Therapy 31 (1), 1–20.

Relevant Websites

http://omicsonline.org/dsm-5-changes-in-diagnostic-criteria-of-sexual-dysfunctions-2161-038X.1000122.php?aid=18508 – DSM-5 Changes in Diagnostic Criteria of Sexual Dysfunctions.
http://cw.routledge.com/textbooks/9780415801751/table-of-contents.asp – The Handbook of Sexuality Related Measures.
http://www.kinseyinstitute.org/research/index.html – The Kinsey Institute.

Sexual Orientation and the Law

Stu Marvel, Emory University School of Law, Atlanta, GA, USA
Martha M Ertman, University of Maryland, Francis King Carey School of Law, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.M. Ertman, volume 21, pp. 13991–13995, © 2001, Elsevier Ltd.

Abstract

The article on sexual orientation and the law tracks historical and recent developments in the area by explaining the historical contingency of sexual orientation, its overlap with issues of gender conformity, and ways that litigation and theoretical research have informed one another. It details the changing social construction of same-sex sexuality from sinful conduct to a maligned status in the late nineteenth century, to a neutral status in the mid-twentieth century, and finally to a focus on postidentity in the late twentieth and early twenty-first century. Discussing developments in various countries, the article first traces the way that litigators and theorists focused on challenging antisodomy laws in the 1980s and early 1990s. It then describes sexual orientation law's shifting focus toward family law issues on a global level, and particularly the push toward same-sex marriage. The article further describes various theoretical approaches to sexual orientation law. It explores contemporary frameworks and tensions in the field, in particular, the attempt to reconcile the postidentity approaches of postmodernism with the identity-based assumptions of conventional legal doctrine, the globalization of sexual and gender identity issues, and an increasing focus on the ways that transgender issues overlap and diverge from mainstream gay and lesbian concerns.

Introduction

The article on sexual orientation and the law tracks historical and recent developments in the area by explaining the historical contingency of sexual orientation, its overlap with issues of gender conformity, and ways that litigation and theoretical research have informed one another. It details the changing social construction of same-sex sexuality from sinful conduct to a maligned status in the late nineteenth century, to a neutral status in the mid-twentieth century, and finally to a focus on postidentity in the late twentieth and early twenty-first century. Discussing developments in various countries, the article first traces the way that litigators and theorists focused on challenging antisodomy laws in the 1980s and early 1990s. It then describes sexual orientation law's shifting focus toward family law issues on a global level, and particularly the push toward same-sex marriage. The article further describes various theoretical approaches to sexual orientation law. It explores contemporary frameworks and tensions in the field, in particular the attempt to reconcile the postidentity approaches of postmodernism with the identity-based assumptions of conventional legal doctrine, the globalization of sexual and gender identity issues, and an increasing focus on the ways that transgender issues overlap and diverge from mainstream gay and lesbian concerns.

Definitions of Sexual Orientation and Gender Identity

Sexual orientation refers to any classification based on sexual desire or conduct, such as heterosexuality, same-sex sexuality, or bisexuality. Sexual orientation describes an individual's physical or emotional attraction to the same or different gender, as well as attractions that may range along the gender spectrum or be variable over the course of one's lifetime (Diamond, 2008). Asexuality is also a sexual orientation, describing a low or absent sexual attraction to others. A person's sexual orientation is distinct from their gender identity and gender expression.

Gender identity refers to a person's internal sense of themselves as male or female (or potentially both or neither). This internal conception may not correspond to a person's body and the gender assigned at birth. Gender identity is self-defined, and emerges from a combination of inherent as well as environmental factors.

The term cisgender marks congruence between the gender one is assigned at birth, one's body, and one's lived experience of that gender; cisgendered people experience a match with societal expectations around their physical and mental expressions of gender identity.

A person who is transgendered experiences a gender identity that does not match up with the gender assigned at birth. Recent usage has favored the term 'trans*' to describe the broadest range of identities that fall outside the cisgender norm. This includes people who identify as transgender, transsexual, transvestite, genderqueer, genderless, two-spirit, and others. Usage without the asterisk generally refers explicitly to transgender men and women. The struggle for self-definition has been an important part of the trans* movement, with focus placed on the use of proper terminology as a symbol of respect and inclusivity. Being trans* does not imply any specific sexual orientation, and may involve identification as straight, gay, lesbian, or bisexual (or something else entirely).

Homosexuality and Heterosexuality as Historically Contingent

The study of the law relating to sexually marginalized people evolved in the 1980s and 1990s out of feminist legal theory, postmodern theory, and critical race theory. A major work informing what may be loosely termed as 'queer legal theory' is

Michel Foucault's *History of Sexuality*, vol. 1 (Foucault, 1978), which posited that sexual orientation is socially constructed rather than naturally or divinely ordained. This point has been further developed by Judith Butler's theory of gender performativity, which posits that identities such as gender are not static, but are produced and sustained through their constant reiteration and performance. (Butler, 1990).

From Conduct to Status to Belief

The social construction of sexual orientation is reflected by the shifting manner in which society and law have viewed same-sex sexuality. The different terms used to describe same-sex sexuality indicate these changes. Prior to the late nineteenth century, same-sex sexuality was viewed as sinful conduct, redeemable through repentance. The terms 'homosexual' and 'heterosexual' did not appear in English until 1892, along with the idea that sexual orientation was a status. These terms are products of German and English medical research on same-sex sexuality, which saw same-sex sexuality as a status, gave it a name, contrasted it with the previously unmarked status of heterosexuality, and assigned it social meaning as a constellation of characteristics based on what a person fundamentally *is* (rather than particular sexual acts that one *does*). As Foucault explains, "[H]omosexuality appeared as one of the forms of sexuality when it was transposed from the practice of sodomy into a kind of interior androgyny, a hermaphrodism of the soul. The sodomite had been a temporary aberration; the homosexual was now a species" (Foucault, 1978: 43). The terms 'invert' or 'homosexual' refer to this medicalized understanding of same-sex sexuality as a sickness rather than sinful conduct. Oscar Wilde's 1895 trials and imprisonment for gross indecency have been credited with introducing into popular culture the status of the male homosexual as effete, artistic, self-centered, and sybaritic (Ertman, 2000).

A third stage, beginning in the 1940s and 1950s, relied on the scientific research of Alfred Kinsey and Evelyn Hooker to contend that same-sex sexuality was a normal variation from opposite-sex sexuality, rather than a disease (Hooker, 1957; Kinsey et al., 1948). The nascent homophile movement adopted the term 'gay' to distance itself from the medical judgment associated with 'homosexual.' A fourth stage, emerging through LGBT (lesbian, gay, bisexual and transgender) and feminist poststructuralist studies in the early 1990s, rejected the very notion of sexual orientation status (evil, neutral or positive), pointing out the indeterminate and potentially subordinating qualities associated with status-based notions of identity. This stage dubbed minority sexual orientations 'queer' (reclaiming the epithet), treating this antistatus as an epistemological category. Queer theory, as this diverse field of study became known, arguably uncouples status from conduct, rejecting essentialist categories to the point that one person can be both queer and engaged in opposite-sex sexuality.

As a deconstructionist project, queer theory remains an intentionally inchoate discipline, definitionally unstable and often rejecting normative aims. This stands in contrast to legal scholarship, which even in its most critical formations, is often oriented toward normative questions and the pursuit of substantive remedies. This tension has made the development of 'queer legal theory' a difficult project. Janet Halley and Andrew Parker have pointed out that queer theory's failure to engage critical legal traditions, and especially the critique of rights, have generated a hostility in centrist legal studies to queer theoretical approaches that do not easily produce a 'policy recommendation' (Halley and Parker, 2007). This queer suspicion of law's essentialist categories has marked the move of LGBT rights into a global setting, with the concern that Western frameworks of gay, lesbian, bisexual, and trans* identity might not be appropriate or welcome in other cultural settings.

Changing Focus of Sexual Orientation Research

Sexual orientation law represents a synergy of gay rights advocacy and theoretical academic writings. Advocates bring cases and support legislation countering antigay discrimination. Scholars both craft theories that inform these efforts and evaluate the cases and statutes that become law. The process of legal reform around antigay discrimination has progressed in the twenty-first century toward an increasing social acceptance of lesbian and gay minorities within most Western, and many non-Western, states. This process requires a society to slowly move from a position of rejecting and persecuting its lesbian and gay minority to one of acceptance and respect by granting equal access to social and political opportunity as full citizens (Wintemute, 2011). It generally has evolved along the following path: (1) decriminalization of same-sex activity; (2) policies or laws prohibiting sexual orientation discrimination; and (3) reform of family law. In step with this process, two important historical areas for reform have involved the repeal of antisodomy laws and the state recognition of same-sex marriage.

Law and culture have also moved toward the recognition of transgender identities and nonbinary gender categories, with some states codifying the acknowledgment of 'third gender' options for self-identification. Nevertheless, substantial discrimination toward trans* individuals still remains. In the US, for example, transgender and gender-nonconforming people experience high rates of economic insecurity, with trans* people of color reporting especially devastating effects from antitransgender bias and persistent structural racism (Grant et al., 2011). Transgender law covers a range of issues related to discrimination in employment, education, immigration, and child custody decisions, as well as high rates of hate crimes and vulnerability to physical and sexual abuse in prison (Goodmark, 2013; Levi and Monnin-Browder, 2013; International Commission of Jurists, 2011). Rights to equality, privacy, and dignity have marked the jurisprudence on gender identity and gender recognition.

On a global level, many nations' courts have ruled against LGBT discrimination under both domestic constitutional rights and the discourse of international law and international human rights, as have transnational legal entities such as the European Union (Stychin, 2004). International documents have also been developed to address issues of sexual orientation and gender identity, most notably the Yogyakarta Principles on the Application of International Human Rights Law in relation to Sexual Orientation and Gender Identity. The

Principles were developed by a group of human rights experts and launched as a global charter for gay rights in 2007. They outline the application of international human rights law to sexual orientation and gender identity, affirming binding international legal standards with which all states must comply.

Antisodomy Laws

Because the criminalization of same-sex sexuality is the most obvious expression of state condemnation, sexual orientation law arose in challenges to antisodomy laws. Initially, much of the legal reform around the decriminalization of sodomy began legislatively. The 1957 Wolfenden Report in the UK proved an important turning point for criminal laws against sodomy. The report argued against equating the sphere of crime with that of sin, insisting that there must "remain a realm of private morality and immorality which is, in brief and crude terms, not the law's business." The Wolfenden recommendations were followed by the British government in 1967 through amendments to the *Sexual Offences Act*, and inspired the American Law Institute's proposal in the *Model Penal Code (MPC)* that states remove homosexuality from the list of offenses. While the *MPC* led a number of American states to repeal sodomy laws, many states nevertheless retained their bans on same-sex sexuality.

Advocates also challenged sodomy laws in court. In *Bowers v. Hardwick*, 478 US 186 (1986), the US Supreme Court upheld Georgia's statute criminalizing sodomy against a constitutional privacy challenge. In *Bowers*, the court declared that the constitution did not confer "a fundamental right to engage in homosexual sodomy." In a vivid concurring opinion, Chief Justice Burger wrote that to recognize sodomy as a fundamental right would be to 'cast aside millennia of moral teaching.' He cited Blackstone (1859) to describe same-sex sodomy as "an offense of 'deeper malignity' than rape, a heinous act 'the very mention of which is a disgrace to human nature.'" Other cases challenged discrimination against gays and lesbians, most notably *Romer v. Evans*, 517 U.S. 620 (1996), which struck down a Colorado constitutional provision that prevented gays and lesbians from obtaining protection against discrimination. The Court held that the provision's 'animus' toward gays and lesbians violated the Equal Protection Clause of the U.S. Constitution by making them "unequal to everyone else" and "a stranger to [Colorado's] laws." Nearly a decade later the Court returned to the issue in *Lawrence v. Texas*, 539 U.S. 558 (2003), which challenged a Texas statute that criminalized homosexual sodomy while leaving heterosexual sodomy legal in the state. The case was decided 5 to 4, overruling *Bowers* and concluding that the state lacked a legitimate interest in regulating the private sexual conduct of consenting adults. By holding that "the state cannot demean [same-sex couples'] existence or control their destiny by making their private sexual conduct a crime," *Lawrence* invalidated sodomy laws in the remaining 13 states that had them on the books.

Other governments have decriminalized same-sex sexuality and banned discrimination based on sexual orientation. The 1996 South African Constitution explicitly prohibited sexual orientation discrimination. The 2000 European Charter of Fundamental Rights includes explicit provisions against discrimination based on sexual orientation. Though the Charter is nonbinding, it frames the European vision on human rights. Several other states have protection against sexual orientation discrimination embedded in their constitutions including Ecuador, Canada, and Brazil, while countries like the Netherlands and Romania have sexual orientation antidiscrimination laws or articles in their Penal Code.

At the opposite extreme, the sodomy ban remains in force mainly in three categories of countries in Asia, Africa, and the English-speaking Caribbean as follows: (1) former colonies of the British Empire that inherited antisodomy legislation as part of English law; (2) countries with Muslim majorities that do not separate Islamic religious law from secular criminal law (the most populous exceptions to this pattern are Indonesia and Turkey); and (3) low-income and middle-income countries (Wintemute, 2011). In 2000 in Malaysia, for example, former deputy Prime Minister Anwar Ibrahim was convicted of sodomy and sentenced to 9 years in prison after a 14-month trial. In late 2013, the Supreme Court of India upheld the constitutionality of Section 377 of the Indian Penal Code, which criminalizes "carnal intercourse against the order of nature," overturning an intermediate appellate court decision. The Ugandan Parliament passed an 'Anti-Homosexuality Bill' in December 2013, which punishes sexual relations between persons of the same sex with a maximum penalty of life in prison.

The Ban on Same-Sex Marriage

Although antisodomy laws are rarely enforced, their existence impedes other antidiscrimination claims, such as gay people's rights to marry or to serve in the military. Advocates and legal scholars, however, have paid far more attention to marriage. Changes in family law around same-sex marriage have occurred through legislative change, judicial rulings based on constitutional guarantees of equality and by direct popular vote through ballot initiatives or referenda.

In 2001, the Netherlands became the first country to allow same-sex couples to marry. At the time of writing (early 2014), 15 other countries have followed suit, including Argentina, Belgium, Brazil, Britain, Canada, Denmark, France, Iceland, New Zealand, Norway, Portugal, South Africa, Spain, Sweden, and Uruguay. In Mexico, same-sex marriage is recognized in all states, but is only performed in three locations. Israel does not allow same-sex marriages to be performed, but does recognize marriages performed in permitting jurisdictions. Other countries offer many rights to same-sex couples but stop short of marriage, including Austria, Croatia, Czech Republic, Ecuador, Finland, Germany, Greenland, Hungary, Liechtenstein, Luxembourg, Slovenia, and Switzerland.

In the United States, starting with Massachusetts in 2003, a total of 19 states and the District of Columbia currently permit same-sex marriage. In 2013, the U.S. Supreme Court struck down as unconstitutional a key part of the Federal Defense of Marriage Act (DOMA), enacted by Congress in 1996, which defined marriage as a legal union between one man and woman. The Court in *United States v. Windsor*, 570 U.S. 12 (2013) held that this provision violated the Fifth Amendment's guarantee of equal protection of the laws because DOMA's purpose was to 'disparage and injure those

whom the state, by its marriage laws, sought to protect in personhood and dignity,' namely married same-sex couples.

Recognition of Gender Identity

While transgender law encompasses a complex range of issues, including health care, access to gender reassignment surgery, employment and housing discrimination, and transgender marriage, a major focus of advocacy and litigation has been on securing the legal recognition of one's preferred gender. Cases of legal recognition arise when trans* individuals seek to change their name or gender designations assigned at birth, and apply for the revision of official identity documents such as birth certificates, passports, and national identity cards (ICJ, 2011). As identification is required for many activities in daily life, the issue is significant, particularly for individuals with a gender presentation at odds with their identifying documents. In 2003, Australia became the first country to allow for a nonbinary designation of sex on a government passport, with a citizen claiming 'X' as their official gender marker rather than male or female. This indeterminate gender marker also became available to citizens of New Zealand in 2013.

Pakistan announced the recognition of a third gender designation on national census and identity cards in 2009, with Bangladesh and Nepal following suit in 2013. In April 2014, the Indian Supreme Court recognized transgender people as a legal third gender. Grounding its decision on rights guaranteed by the nation's Constitution as well as international law, the court determined gender identity to be fundamental to the rights to self-determination, dignity, and freedom. The judgment obliges central and state authorities to address the discrimination and stigma faced by transgender people, and grants individuals the right to legal recognition of their chosen gender identity (*National Legal Services Authority v. Union of India* [Writ Petition (Civil) No. 400 of 2012], delivered 15 April 2014).

Many countries now allow individuals to change their legal gender from male to female, or from female to male, although medical intervention (such as sex reassignment surgery) or psychological evaluation (such as a diagnosis of gender dysphoria) are required steps in most jurisdictions. The majority of European nations allow for name changes and revisions to designated sex on birth certificates, pending the submission of reports from a licensed medical practitioner. South Korea, Japan, and Malaysia also allow for transgender individuals to change their legal gender, although medical intervention is mandatory in most cases. In South Africa, the 2003 Alteration of Sex Description and Sex Status Act similarly permits the issuance of a new birth certificate and identity documents to transgender individuals who have undergone medical or surgical gender reassignment.

In the United States, the authority to determine issues of legal gender rests with the individual states. Idaho, Ohio, Kansas, Tennessee, and Oklahoma do not allow transgender people to change the gender marker on their birth certificate. Even in states such as California, Oregon, Vermont, and Washington that permit the change, however, different agencies follow different policies for the issuance of identity documents, requiring various medical, psychological, and surgical interventions and diagnoses. According to a 2011 study by the National Center for Transgender Equality, among those U.S. citizens who have transitioned gender, only 21% have been able to update all their identification documents and personal records and 33% have updated none of their IDs or records (Grant et al., 2011). Transgender individuals who cannot correct their documentation may face a greater risk of discrimination. In the same national study, 40% of people reported experiencing harassment when presenting identification that did not match their gender identity or expression, while 15% were asked to leave and 3% were attacked or assaulted.

Within the United States, discrimination based on a person's gender identity is increasingly prohibited, most notably in regard to employment facilities and gender-segregated facilities and/or activities. For instance, in the 2012 appeal *Macy v. Department of Justice*, the U.S. Equal Employment Opportunity Commission held that discrimination against an individual because that person is transgender (gender identity discrimination) is tantamount to sex discrimination and therefore prohibited under Title VII.

Theoretical Approaches to Sexual Orientation Law

The emphasis in current theory and research on sexual orientation law is to deconstruct and reconstruct legal regulations that subordinate gay and transgendered people. Under the rubric of theoretical approaches, we include both ideological and methodological approaches.

Ideology

There are diverse ideological approaches in the developing theory of sexual orientation law. A liberal approach would suggest that gay people are sufficiently similar to heterosexuals to justify equal legal treatment, and altering legal rules to allow same-sex marriage or openly gay military personnel will not fundamentally change these institutions. A critical approach, in contrast, contests essentialized notions of identity, seeing identity as socially constructed rather than reflecting essential commonalties among people who engage in same-sex sexuality. It may also critique the inequality inherent in institutions like marriage and the military. This kind of critical approach often seeks a more comprehensive restructuring of legal regulations than simply adding gay people to existing institutions, instead proposing the incorporation of sexual orientation analysis into other antisubordination discourses, such as feminism, critical race studies, and class analysis. Critical theorists might further propose alternatives to marriage, such as a range of relationship recognition devices for all families, including both couples and noncouples like close friends or adult sisters whose lives are intertwined (Polikoff, 2008; Rosenbury, 2007). These views contend that only new institutions can transcend essentialist notions of identity and alleviate the sex/gender hierarchies inherent in marriage.

Emerging Scholarship in Sexual Orientation and Gender Identity

Postmodern and empirical approaches have contributed to the literature on sexual orientation law. An influential queer

theoretical critique challenged the essentialism in immutability claims that are foundational to both legal doctrine and much gay and lesbian advocacy (Halley, 1994). An empirical study, in contrast, compiled data about rates of arrest and prosecution for sexual offenses to show that consensual same-sex sexual activity was unfairly singled out for criminal penalty (Eskridge, 1999). Both of these approaches have helped erode the legal and social legitimacy of criminalizing sodomy and build support for same-sex marriage and other reforms.

The most recent development in the field of law and sexuality has involved the construction of a global discourse of LGBT rights and political community. As LGBT causes found traction in law and policy in many Western nations, the focus of legal struggle has become increasingly internationalized. This shift has functioned to position LGBT rights as a new 'marker of modernity,' creating new political hierarchies and casting states that fail to respect rights around sexual diversity as 'backward and uncivilized' (Rao, 2011). Within these political hierarchies sexuality functions as a site of conflict and transfer point for power, with postcolonial states approaching the management of sex as a tool of governance – often in direct resistance to former colonial powers (Franke, 2004). These conflicts reveal the tension between cosmopolitan and communitarian perspectives, as well as intergenerational divides, the split between urban and rural dwellers, and between those who invoke a 'traditional past' unburdened by Western influences and those who claim a local history of pluralism and tolerance.

The importance of trans* voices and issues of gender identity have also gained increasing prominence in recent years. Critical politics around trans* issues have emerged from legal academia and activism, arguing for a broader agenda than mere law reform. As focused on the experiences of transgender and queer people, Dean Spade has called for attention to wide-ranging issues including "prison abolition, the elimination of poverty, access to full health care, and an end to immigration enforcement" (Spade, 2011). Such critical perspectives reflect the poststructuralist suspicions of law's normativity. They also evidence wariness of the civil rights and 'equality' strategies pursued by mainstream gay and lesbian advocacy organizations, criticizing their reliance upon state institutions of law and administration and their failure to account for the trans* bodies marginalized by such a legal rights framework.

See also: Critical Race Theory; Discrimination and the Law; Equality and Inequality: Legal Aspects; Family Law in the United States; Feminist Legal Theory; Gender and the Law; Queer Theory and Intersectionality; Race and the Law; Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective; Sexual Minorities and Sexual Citizenship; Sexual Orientation: Historical and Social Construction.

Bibliography

Blackstone, W., 1859. Commentaries, vol. IV. Harpe Brothers, New York.
Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
Diamond, L.M., 2008. Sexual Fluidity: Understanding Women's Love and Desire. Harvard University Press, Cambridge, MA.
Ertman, M., 2000. Oscar wilde: paradoxical poster child for both identity and post-identity. Law and Social Inquiry 25, 153.
Eskridge, W., 1999. Gaylaw: Challenging the Apartheid of the Closet. Harvard University Press, Cambridge, MA.
Foucault, M., 1978. The History of Sexuality. In: An Introduction, vol. I. Pantheon, New York.
Franke, K., 2004. Sexual tensions of post-empire, 33 studies in law. Politics and Society 63, 90.
Goodmark, L., 2013. Transgender people, intimate partner abuse and the legal system. Harvard Civil Rights & Civil Liberties Law Review 48, 51.
Grant, J.M., Mottet, L.A., Tanis, J., Harrison, J., Herman, J.L., Keisling, M., 2011. Injustice at Every Turn: A Report of the National Transgender Discrimination Survey. National Center for Transgender Equality and National Gay and Lesbian Task Force, Washington.
Halley, J., 1994. Sexual orientation and the politics of biology: a critique of the argument from immutability. Stanford Law Review 46, 503.
Halley, J., Parker, A., 2007. "Introduction", after sex? On writing since queer theory. South Atlantic Quarterly 106 (3), 413.
Hooker, E., 1957. The adjustment of the male overt homosexual. Journal of Projective Techniques 21, 18–31.
International Commission of Jurists (ICJ), 2011. Sexual Orientation, Gender Identity, and Justice: A Comparative Law Casebook. Geneva.
Kinsey, A., Pomeroy, W., Martin, C., 1948. Sexual Behavior in the Human Male. W.B. Saunders, Philadelphia, PA.
Levi, J.L., Monnin-Browder, E.E., 2013. Transgender Family Law: A Guide to Effective Advocacy. AuthorHouse, Bloomington, IN.
Polikoff, N., 2008. Beyond (Straight and Gay) Marriage: Valuing All Families under the Law. Beacon, Boston, MA.
Rao, R., 2011. Queer in the time of terror. In: Narrain, A., Gupta, A. (Eds.), Law Like Love: Queer Perspectives on Law. Yoda Press, New Delhi, pp. 43–66.
Rosenbury, L., 2007. Friends with benefits? Michigan Law Review 106, 189.
Spade, D., 2011. Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law. South End Press, New York.
Stychin, C., 2004. Same-sex sexualities and the globalization of human rights discourse. McGill Law Journal 49, 951–968.
Wintemute, R., 2011. Same-sex love and Indian penal code §377: an important human rights issue for India. NUJS Law Review 4 (1), 31–65.

Sexual Orientation: Biological Influences

Malvina N Skorska and Anthony F Bogaert, Brock University, St. Catharines, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition by E. Stein, volume 21, pp. 13995–13998, © 2001, Elsevier Ltd.

Abstract

Research into the various influences on sexual orientation development has increased over the past 15–30 years. In this article, the key findings of the research into the biological influences on sexual orientation development are discussed. Specifically, biological influences are discussed in relation to prenatal hormones, genetics, the brain, the body, and the fraternal birth order effect. Future directions for research are also discussed and provide several fruitful lines of research for current and future researchers.

In 1973, 'homosexuality' was removed from the Diagnostic and Statistical Manual of Mental/Psychiatric Disorders. The removal of homosexuality as a mental disorder has had an impact on society, and has stimulated copious amounts of research on the complex trait of sexual orientation over the last 15–30 years. This research not only informs us about sexual dimorphisms (e.g., gender roles), atypical sexual attraction, and nontraditional aspects of sexual orientation (e.g., asexuality), but also provides insight into the definition and origins of sexual orientation. This article will summarize the current state of research into the biological origins of sexual orientation. After an introduction into the definition of sexual orientation and the biological approach to examining sexual orientation development, the literature on several biological influences are discussed in relation to hormones, genetics, the brain, the body, and the fraternal birth order effect. Future directions in this area of research will conclude the article.

What Is Sexual Orientation?

There are many aspects to an individual's sexual preferences. For example, an individual may prefer individuals of a certain age, ethnicity, body type, or personality, to name a few, for their sexual partner(s). A core aspect of an individual's sexual preferences is, however, their sexual orientation. Although there is no consensus on an exact definition of sexual orientation, generally sexual orientation concerns the sex that an individual is interested in for his/her sexual partner(s). Definitions of sexual orientation can include two components: a psychological component concerning attraction to or fantasies about a certain sex, and a behavioral component concerning the sex of a partner with which an individual actually engages in sexual acts. In recent years, the field has shifted toward an emphasis on the psychological component – specifically sexual attraction – in defining sexual orientation.

Most researchers categorize sexual orientation into three main categories: homosexual, heterosexual, and bisexual, although there is debate among researchers whether sexual orientation is categorical or continuous in nature. 'Homosexual' describes an orientation toward the same sex; 'heterosexual' describes an orientation toward the other sex; and 'bisexual' describes an orientation toward both sexes. Labels of straight (or heterosexual), gay (or homosexual man), and lesbian (or homosexual woman) are often used interchangeably with the homosexual and heterosexual categories. Some recent sexual orientation theorists have suggested a fourth main category of sexual orientation, asexual, the absence of sexual attractions for others, although asexual individuals are much less studied than the three traditionally defined groups – gay/lesbian, straight, and bisexual – mentioned above. It is important to note that although these categorization labels for traditional groups (e.g., gay, lesbian, bisexual) are used in several studies, they are confounded with identity labels (e.g., not all individuals who engage in same-sex behavior identify as gay, lesbian, or bisexual). Thus, this article will use the terms same-sex orientation, both-sex orientation, and other-sex orientation to refer to gay and lesbian individuals, bisexual individuals, and straight individuals, respectively. Various epidemiological studies have indicated that approximately 2–5% of men, and approximately 1–2% of women have a predominant or exclusive orientation toward the same sex, approximately less than 1% of women and men have an orientation toward both sexes (LeVay, 2010), and approximately 1% of the population may be asexual (Bogaert, 2004, 2012).

The Biological Approach

Several approaches have been taken to explain the development of variation in sexual orientation. These include psychoanalytic, learning, and biological approaches, with the two former approaches emphasizing a nurture or environment perspective. Recent researchers have placed an increased emphasis on nature or a biological perspective to investigating the cause of sexual orientation. This emphasis is partly due to the failure of the other approaches in explaining certain aspects of sexual orientation development.

An emphasis on biological influences to sexual orientation development was largely borne out of animal studies that suggested early (i.e., prenatal and neonatal) exposure to testosterone and/or estrogen, which are hormones produced by the gonads, have different effects on the brain, adult sexual behavior, and adult sexual partner preference of males and females in species such as rats, monkeys, and sheep. Current models also include the influence of genes in sex determination and sexual differentiation. For example, the *Sry* gene, which is encoded by the Y chromosome, determines whether the gonads develop into testes or ovaries, and the gonads' hormones, especially testosterone released from the testes, influence the physical

development and neural development of males and females. These organizational effects of gonadal hormones also influence adult sexual behaviors and adult sexual partner preferences. Note, that we are using the word influence to indicate that biology exerts some impact on the development of sexual orientation; however, this does not mean that it directly determines sexual orientation in a fixed manner. With this knowledge about the sexual development of male and female animals, coupled with evidence that some animals engage in same-sex- and/or both-sex-oriented sexual behaviors, researchers studying the origins of sexual orientation development in humans began to explore several biological variables and their relation to sexual orientation development in humans. Some of the key findings of this research are outlined in the following sections.

Hormones

To examine the influence of hormones in sexual orientation development, researchers have examined both circulating hormone levels and prenatal hormone levels. Circulating hormone levels (sometimes called activational or reversible effects) refer to hormones that are currently circulating in adults. Prenatal hormone levels (sometimes called organizational or more permanent effects) refer to hormones present in an individual while he/she is in the womb that influence gonadal and brain development in a fetus. Researchers have generally found that there were no differences in circulating testosterone levels between men with a same-sex orientation and an other-sex orientation. However, some studies have found evidence suggesting that more masculine-identifying women with a same-sex orientation, compared to more feminine-identifying women with a same-sex orientation had higher salivary testosterone levels (Please note that a range of gender expression occurs in individuals. There are clear differences on a group level between men and women in gender expression; however, although average differences exist between men and women, there is also a range of variability within men and within women in gender expression. Likewise, although there is evidence that same-sex-attracted individuals are, on average, more gender nonconforming than other-sex attracted individuals, there is also a wide range of variability within same-sex-attracted and other-sex attracted individuals in gender expression.) As there was not much support for the role of circulating levels of testosterone in sexual orientation development, researchers have focused on the possible influence of prenatal hormones on the organization of the fetal brain with respect to the development of sexual orientation. This research assumes that typical levels of hormones (e.g., testosterone) have been altered in prenatal sexual development (e.g., lowered in male fetuses; raised in female fetuses), and that this atypical exposure to prenatal hormones ultimately predisposes an individual to a same-sex orientation later on in life by affecting areas of the brain (see below) implicated in sexual orientation. Researchers cannot ethically manipulate levels of hormones in an experiment on pregnant women; however, other, more indirect or correlational lines of research have generally provided support for the role of prenatal hormones in human sexual orientation development (e.g., Ellis and Ames, 1987; Hines, 2011; Meyer-Bahlburg, 1984).

One avenue of research is to examine other atypical sexual development. For example, girls and women who have the classical form of congenital adrenal hyperplasia experience an increase in prenatal androgen (e.g., testosterone) levels produced by the adrenal glands. As a result, these girls and women have some degree of masculinized genitals. Studies conducted on the relationship between women with congenital adrenal hyperplasia and sexual orientation have found that women with the most severe forms of this condition have, on average, a reduced other-sex orientation (i.e., have an increased same-sex orientation). Boys and men (i.e., these individuals are born with the XY chromosomes) with androgen insensitivity syndrome and 5-alpha reductase deficiency experience a decreased response to androgens, which results in men with feminized genitals who are, as a result, usually reared as girls. The few studies that have been conducted on these girls suggest that as adults these women are generally interested in men for their partners; however, the results are more complicated in girls with 5-alpha reductase deficiency, as they experience masculinization at puberty and some change to be men as adults (e.g., due to cultural or medical factors). In sum, this research provides some support for the role of androgens in sexual orientation development in both women and men (e.g., Hines et al., 2003; Meyer-Bahlburg et al., 2008; Money et al., 1984).

Another avenue of research is via pregnant women who were prescribed hormones during their pregnancies. In the 1940s–60s, pregnant women were prescribed a synthetic estrogen, diethylstilbestrol (DES), to prevent miscarriage. Thus, these women carried fetuses that were exposed to increased levels of estrogen (which can have masculinizing effects prenatally). Some studies on the offspring of these women indicate that women who were exposed to DES were associated with an increase in same-sex orientation compared to women who were not exposed to DES prenatally. Another study did not find this association, although a less sensitive measure of sexual orientation was used. In men, there was no supported relationship of an increase in same-sex orientation with men who were exposed to DES compared to men who were not exposed to DES. Thus, research provides some support for the role of DES (and by implication estrogen) in sexual orientation development in women only (e.g., Meyer-Bahlburg et al., 1995; Titus-Ernstoff et al., 2003).

The last avenue of research related to hormones discussed in this article concerns childhood behavior. Generally, boys play with male-typical toys (e.g., trucks, cars) and girls play with female-typical toys (e.g., dolls, jewelry). Further, boys engage in more rough-and-tumble play than girls, and the majority of boys and girls have same-sex playmates. Retrospective and prospective studies have found that men and women with same-sex orientations recall more atypical patterns of behavior in their childhood with respect to toy preference, activity preference, and playmate choices. Studies conducted on girls with congenital adrenal hyperplasia have found that these girls show increased levels of childhood atypical behavior as well (e.g., Bailey and Zucker, 1995; Berenbaum and Hines, 1992; Meyer-Bahlburg et al., 2004). This suggests that prenatal androgens may influence the development of childhood playmate choices, toy interests, and activity interests. As a result, because individuals with a same-sex orientation are more likely to show increased levels of childhood atypical

behavior, this provides additional support for the notion that atypical exposure to prenatal androgens is a factor related to sexual orientation development. Thus, these lines of research, and other lines of research not presented here, converge on the idea that hormones are one factor that is related to the development of sexual orientation in humans.

Genetics

Twin studies suggest that there is some degree of genetic influence on sexual orientation. That is, these studies find that monozygotic twins, which share approximately 100% of their genes, have higher concordance rates than dizygotic twins, which share approximately 50% of their genes, for a same-sex orientation; although, the exact concordance rates differ depending on the study. Studies also suggest that there is an increased rate of same-sex orientation among biological siblings. Further, there is some support that the chromosomal region Xq28 on the X chromosome is associated with same-sex orientation in men; however, replication of this finding by an independent group of researchers was not successful and the exact genes in this chromosomal region that are relevant to sexual orientation have not yet been identified as of the date of the writing of this review. Another study has also implicated the X chromosome in sexual orientation development in men. Female cells contain two X chromosomes and one of them is randomly inactivated in early development in each cell. Since this inactivation is random, 50% of the X chromosomes should be maternal in origin and 50% should be paternal in origin (i.e., because women receive one X chromosome from their mother and the other X chromosome from their father). Mothers of gay sons displayed skewing in X chromosome inactivation (i.e., a nonrandom amount of X activation was either paternal or maternal) and this skewing increased with an increase in the number of gay sons. Another study using molecular genetic techniques has provided some support for the role of autosomal genes (i.e., genes not on the sex chromosomes), specifically on chromosomes 7, 8, and 10, in men's sexual orientation development. Although research into the genetics of sexual orientation development is not abundant, the existing research (e.g., Bailey and Pillard, 1991; Bocklandt et al., 2006; Hamer et al., 1993) suggests that there is a genetic component to sexual orientation development, with most studies conducted only in men.

From an evolutionary perspective, same-sex orientations are a paradox: genes that underlie same-sex orientations would not be passed down to future generations because same-sex sexual activity does not result in offspring via sexual reproduction. Yet, same-sex orientations exist (and likely have existed throughout human history) and thus, researchers have proposed theories/hypotheses to explain why genes for a same-sex orientation existed and continue to thrive. One hypothesis based on kin selection argues that genes for a same-sex orientation were indirectly selected for because individuals with a same-sex orientation helped increase the reproductive success of their siblings via aiding in resource accretion or caring for their children. This hypothesis has garnered some support, at least in men, through investigations of the *fa'afafine*, a relatively feminine group of men in Independent Samoa who engage in same-sex behavior. Several studies suggest the *fa'afafine* were more likely to aid their kin than men with an other-sex orientation in Independent Samoa. Another hypothesis that has found support suggests that a same-sex orientation was related to an increased maternal fecundity. There are several other hypotheses; however, these hypotheses have yet to be rigorously tested (see Rahman and Wilson, 2003, for a review). Thus, it is clear that research into the maintenance of genes for a same-sex orientation is underway; however, no clear answer to why such genes are propagated has been established (e.g., Vasey and VanderLaan, 2010).

The Brain

Researchers examining brain structures implicated in sexual orientation development have found some promising results. Aside from being the main organ of behavior (including mediating sexual attraction), the brain is an important area to study with respect to biological influences because the brain is affected by prenatal hormones (among other factors). Current biological models of sexual orientation development include the idea that prenatal hormones affect areas of the brain implicated in sexual orientation development, and these affected areas somehow influence the sexual orientation of an adult individual. Generally, it has been found that the interstitial nuclei of the anterior hypothalamus (INAH), specifically INAH-3, is implicated in sexual orientation. The hypothalamus has many functions, including roles in feeding and drinking, regulation of body temperature, and reproduction/mating. Men with a same-sex orientation, and women with an other-sex orientation had the same cell volume in this area of the hypothalamus; however, the cell volumes in men with a same-sex orientation and women with an other-sex orientation were both smaller compared to men with an other-sex orientation (note: women with a same-sex orientation were not examined). Thus, this area seems to have been organized in a more female-typical direction in men with a same-sex orientation. This finding was partially replicated by an independent set of researchers. Homologous sites, such as the sexually dimorphic nucleus of the preoptic area, have been implicated in mating behavior in rodents, and other mammals (e.g., LeVay, 1991).

Other brain areas have been implicated in sexual orientation. For example, the suprachiasmatic nucleus was found to be larger in men with a same-sex orientation than men with an other-sex orientation by one set of researchers; however, the finding is difficult to interpret since the suprachiasmatic nucleus is not known to be involved in anything sexual, but is involved in regulation of circadian rhythms. A study also suggested that the anterior commissure, which is a set of smaller fibers that connect the two cerebral hemispheres of the brain, was larger in men with a same-sex orientation than men with an other-sex orientation. Thus, this area also appeared to be slightly feminized in men with a same-sex orientation, as previous studies indicated that women with an other-sex orientation have a larger anterior commissure than men with an other-sex orientation. Another study indicated that the isthmus, a section of the corpus callosum that connects parts of the parietal lobe and temporal lobe in the left and right

hemispheres, respectively, was larger in right-handed men with a same-sex orientation compared to right-handed men with an other-sex orientation. The isthmus has been shown to be larger in nonright-handed individuals than right-handed individuals (*see* Section The Body). Further, one study demonstrated that men with an other-sex orientation have a right cerebral hemisphere that was larger than the left cerebral hemisphere, while women with an other-sex orientation have left and right cerebral hemispheres that were the same size. Men with a same-sex orientation had the same size left and right cerebral hemispheres, similar to women with an other-sex orientation, and women with a same-sex orientation had a right cerebral hemisphere that was slightly larger than the left cerebral hemisphere, which slightly aligned with the findings in men with an other-sex orientation. Other studies have implicated the perirhinal cortex (involved in olfaction, spatial processing, and memory) in women with same-sex orientation, and the amygdala in men and women with same-sex orientation, with shifts in the sex-atypical way. Thus, it is clear that there are a number of important brain areas associated with sexual orientation, although much of this work has only been conducted in men, is very recent, and requires replication (see Bao and Swaab, 2011; LeVay, 2010, for reviews). Note that although the assumption in this research is that the causal direction is such that the brain influences sexual orientation, the direction of causality is still unclear (i.e., it is possible that sexual orientation may influence the structure of the brain or a third variable influences both brain structure and sexual orientation).

The Body

Some sexual orientation researchers have examined sex dimorphic physical characteristics that are influenced by prenatal hormones, and then examined if men and women with a same-sex orientation differ from men and women with an other-sex orientation on these characteristics. Specifically, these researchers aim to determine whether these characteristics are shifted in the female-like way in men with a same-sex orientation and male-like way in women with a same-sex orientation, compared to men and women with an other-sex orientation, respectively. If physical development in men and women with a same-sex orientation is shifted in this way, then it suggests that prenatal hormones are, along with affecting these physical development characteristics, altering structures of the brain relevant to sexual orientation. This research strategy underlies a number of research programs on sexual orientation development, including 2D:4D finger ratios, handedness, otoacoustic emissions, and height.

Women have a higher ratio of the length of the second digit to the length of the fourth digit (i.e., 2D:4D finger ratio) than men. Some studies have found that prenatal androgen exposure (or androgen/estrogen exposure) underlies this sex difference. With respect to the association of 2D:4D finger ratios and sexual orientation, findings have been mixed. Some studies reported that men with a same-sex orientation had a more female-typical ratio, while others reported no difference or a more male-typical ratio; other studies reported that the ratio was more male-typical in only the right hand. A recent meta-analysis concluded that women with a same-sex orientation had a more male-typical 2D:4D finger ratio, but there was no relationship between sexual orientation and 2D:4D finger ratio and sexual orientation in men (Grimbos et al., 2010).

Boys and men are more likely to be nonright-handed than girls and women, and it has been argued that handedness is determined prenatally, including a possible role for prenatal hormones. Several studies have been conducted on the relationship between handedness and sexual orientation with mixed results. A meta-analysis of 20 such studies concluded that men with a same-sex orientation are 34% more likely than men with an other-sex orientation to be nonright-handed, and women with a same-sex orientation are 91% more likely than women with an other-sex orientation to be nonright-handed (Lalumiere et al., 2000).

Otoacoustic emissions are sounds produced in the cochlea (a structure inside the inner ear) that can be recorded from the outer ear. They can be spontaneously produced (i.e., spontaneous otoacoustic emissions (SOAEs)) or produced after a sound is presented to the ear (i.e., click-evoked otoacoustic emissions (CEOAEs)). Generally, the right ear has more SOAEs than the left ear and women have more SOAEs than men. CEOAEs are stronger in the right ear than in the left ear and women have stronger CEOAEs than men. These sex differences exist in newborns and thus, the sex difference is assumed to be determined prenatally by androgens. Studies have found that the number of SOAEs was decreased and the strength of CEOAEs was weakened in women with a same-sex orientation, similar to the patterns found in men with an other-sex orientation; however, there was no relationship between sexual orientation in men and the number of SOAEs and strength of CEOAEs. Thus, the otoacoustic emission research provides evidence of shifts in the male-typical way in women with a same-sex orientation in terms of SOAEs and CEOAEs; however, shifts in the female-typical way were not found in men with a same-sex orientation (McFadden and Champlin, 2000).

Men are generally taller than women and prenatal hormones have been partially found to underlie this sex difference. The majority of studies conducted on the relationship between height and sexual orientation have found that men with a same-sex orientation were shorter, on average, than men with an other-sex orientation. The relationship between height and sexual orientation in women is less consistent, with the majority of studies not finding a difference in height between women with a same-sex orientation and women with an other-sex orientation (e.g., Bogaert and Blanchard, 1996; Bogaert, 2010).

Overall, the research programs investigating several physical markers of prenatal hormone exposure have found some support for a relationship between 2D:4D finger ratios, handedness, otoacoustic emissions, and sexual orientation in women, as well as support for a relationship between handedness, height, and sexual orientation in men. Indirectly (i.e., because these studies are correlational), these studies add support to the idea that prenatal hormones are a factor in sexual orientation development.

The Fraternal Birth Order Effect

It has been consistently shown in various samples and by independent researchers that men with a same-sex orientation

tend to have an increased number of older brothers than men with an other-sex orientation (i.e., a later fraternal birth order). It has been estimated that each additional older brother increases the chance of same-sex orientation by 33%. This effect has been termed the fraternal birth order effect because the number of older sisters is not associated with sexual orientation in men, and the number of older sisters and older brothers is not associated with sexual orientation in women. Studies have ruled out potential confounding variables to explain the fraternal birth order effect, such as year of birth, age, socioeconomic status, sibship size, and parental age. Further, the fraternal birth order effect has been demonstrated even if biological older brothers were reared in a different household, and independent of the presence of adopted older brothers or older stepbrothers (Blanchard and Bogaert, 1996; Bogaert and Skorska, 2011).

An immunological explanation, termed the maternal immune hypothesis, has been proposed as an explanation for the fraternal birth order effect. Although there is no direct empirical support (yet) for the maternal immune hypothesis, there has been no feasible alternative explanation of the fraternal birth order effect to date. Specifically, the maternal immune hypothesis postulates that the mother's immune system is continuously immunized to male-specific proteins (associated with the Y chromosome, which a mother does not have) with each successive male fetus. The increasing concentration of antibodies the mother develops after each male fetus cross the blood–brain barrier to affect specific areas in the brain associated with sexual orientation development, which would then influence the sexual orientation of the man as an adult. Several lines of research indicate that transfer of material occurs from the fetus to its mother, male-specific substances cause immune responses in women, the relevant male-specific substances play a role in sexual differentiation of the brain (independent of prenatal hormones), and a maternal immune response to male-specific substances affects fetal brain development. These lines of research increase the plausibility of the maternal immune hypothesis in explaining the fraternal birth order effect, and more direct empirical investigations testing this hypothesis are currently underway.

Future Directions

The study of the origins of the complex trait of sexual orientation has come a long way since 1973. Through the efforts of numerous researchers, it is clear that there is increasing evidence that biology influences sexual orientation development in humans (note, again that an 'influence' does not mean that biology determines sexual orientation in a fixed manner). However, the field of sexual orientation research is still quite young and there are several questions that require answering in the future. For example, there is less sexual orientation research in women than in men (e.g., the genetics and brain research has been mostly done in men). To understand how sexual orientation develops requires a good understanding of the developmental processes in both men and women. There is also a lack of studies of individuals who are attracted to both sexes and those attracted to neither sex (i.e., asexual individuals). The field would benefit from conducting more research into these groups of individuals, as they will provide important new insights into sexual orientation development, broadly defined. As well, the findings in studies that have been conducted thus far often require replication in different samples to ensure the effects are reliable and generalizable to the wider population. For example, the majority of studies are on middle-class caucasian Western individuals, with several studies using small samples sizes. Further, many of the studies that have been conducted thus far need to be extended to rule out alternative explanations, as they often only provide indirect support for biological mechanisms. Last, the biological research into sexual orientation development could integrate with research and theory on the environmental origins of sexual orientation to provide a complete picture of how a complex trait such as sexual orientation develops. For example, women's sexuality has been argued to be more flexible in sexuality and sexual orientation compared to men's sexuality (e.g., Baumeister, 2000). With several questions left to answer and numerous studies to be conducted, the future of research into the biological influences on sexual orientation development (and sexual orientation development more generally) looks promising.

See also: Gay, Lesbian, and Bisexual Youth; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Sexual Orientation: Historical and Social Construction; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Bailey, J.M., Pillard, R.C., 1991. A genetic-study of male sexual orientation. Archives of General Psychiatry 48, 1089–1096.

Bailey, J.M., Zucker, K.J., 1995. Childhood sex-typed behavior and sexual orientation: a conceptual analysis and quantitative review. Developmental Psychology 31, 43–55.

Bao, A., Swaab, D.F., 2011. Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorder. Frontiers in Neuroendocrinology 32, 214–226.

Baumeister, R.F., 2000. Gender differences in erotic plasticity: the female sex drive as socially flexible and responsive. Psychological Bulletin 126, 347–374.

Berenbaum, S.A., Hines, M., 1992. Early androgens are related to childhood sex-typed toy preferences. Psychological Science 3, 203–206.

Blanchard, R., Bogaert, A.F., 1996. Homosexuality in men and number of older brothers. The American Journal of Psychiatry 153, 27–31.

Bocklandt, S., Horvath, S., Vilain, E., Hamer, D.H., 2006. Extreme skewing of X chromosome inactivation in mothers of homosexual men. Human Genetics 118, 691–694.

Bogaert, A.F., 2004. Asexuality: prevalence and associated factors in a national probability sample. Journal of Sex Research 41, 279–287.

Bogaert, A.F., 2006. Biological versus nonbiological older brothers and men's sexual orientation. Proceedings of the National Academy of Sciences of the United States of America 103, 10771–10774.

Bogaert, A.F., 2010. Physical development and sexual orientation in men and women: an analysis of NATSAL-2000. Archives of Sexual Behavior 39, 110–116.

Bogaert, A.F., 2012. Understanding Asexuality. Rowman & Littlefield Publications, Inc., Plymouth, UK.

Bogaert, A.F., Blanchard, R., 1996. Physical development and sexual orientation in men: height, weight and age of puberty differences. Personality and Individual Differences 21, 77–84.

Bogaert, A.F., Skorska, M., 2011. Sexual orientation, fraternal birth order, and the maternal immune hypothesis: a review. Frontiers in Neuroendocrinology 32, 247–254.

Ellis, L., Ames, M.A., 1987. Neurohormonal functioning and sexual orientation: a theory of homosexuality-heterosexuality. Psychological Bulletin 101, 233–258.

Grimbos, T., Dawood, K., Burriss, R.P., Zucker, K.J., Puts, D.A., 2010. Sexual orientation and the second to fourth finger length ratio: a meta-analysis in men and women. Behavioral Neuroscience 124, 278–287.

Hamer, D.H., Hu, S., Magnuson, V.L., Hu, N., Pattatucci, A.M.L., 1993. A linkage between DNA markers on the X-chromosome and male sexual orientation. Science 261, 321–327.

Hines, M., 2011. Prenatal endocrine influences on sexual orientation and on sexually differentiated childhood behavior. Frontiers in Neuroendocrinology 32, 170–182.

Hines, M., Ahmed, S.F., Hughes, I.A., 2003. Psychological outcomes and gender-related development in complete androgen insensitivity syndrome. Archives of Sexual Behavior 32, 93–101.

Lalumiere, M.L., Blanchard, R., Zucker, K.J., 2000. Sexual orientation and handedness in men and women: a meta-analysis. Psychological Bulletin 126, 575–592.

LeVay, S., 1991. A difference in hypothalamic structure between heterosexual and homosexual men. Science 253, 1034–1037.

LeVay, S., 2010. Gay, Straight, and the Reason Why: The Science of Sexual Orientation. Oxford University Press, Inc., New York.

McFadden, D., Champlin, C.A., 2000. Comparison of auditory evoked potentials in heterosexual, homosexual, and bisexual males and females. Journal of Auditory Research in Ontycology 1, 89–99.

Meyer-Bahlburg, H.F.L., 1984. Psychoendocrine research on sexual orientation – current status and future-options. Progress in Brain Research 61, 375–398.

Meyer-Bahlburg, H.F.L., Ehrhardt, A.A., Rosen, L.R., Gruen, R.S., Veridiano, N.P., Vann, F.H., Neuwalder, H.F., 1995. Prenatal estrogens and the development of homosexual orientation. Developmental Psychology 31, 12–21.

Meyer-Bahlburg, H.F.L., Dolezal, C., Baker, S.W., Carlson, A.D., Obeid, J.S., New, M.I., 2004. Prenatal androgenization affects gender-related behavior but not gender identity in 5-12-year-old girls with congenital adrenal hyperplasia. Archives of Sexual Behavior 33, 97–104.

Meyer-Bahlburg, H.F.L., Dolezal, C., Baker, S.W., New, M.I., 2008. Sexual orientation in women with classical or non-classical congenital adrenal hyperplasia as a function of degree of prenatal androgen excess. Archives of Sexual Behavior 37, 85–99.

Money, J., Schwartz, M., Lewis, V.G., 1984. Adult erotosexual status and fetal hormonal masculinization and demasculinization – 46,XX congenital virilizing adrenal-hyperplasia and 46,XY androgen-insensitivity syndrome compared. Psychoneuroendocrinology 9, 405–414.

Rahman, Q., Wilson, G.D., 2003. Born gay? The psychobiology of human sexual orientation. Personality and Individual Difference 34, 1337–1382.

Titus-Ernstoff, L., Perez, K., Hatch, E.E., Troisi, R., Palmer, J.R., Hartge, P., Hyer, M., Kaufman, R., Adam, E., Strohsnitter, W., Noller, K., Pickett, K.E., Hoover, R., 2003. Psychosexual characteristics of men and women exposed prenatally to diethylstilbestrol. Epidemiology 14, 155–160.

Vasey, P.L., VanderLaan, D.P., 2010. Avuncular tendencies and the evolution of male androphilia in Samoan *fa'afafine*. Archives of Sexual Behavior 39, 821–830.

Sexual Orientation: Genetic Aspects

Khytam Dawood, Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article will provide an overview of the evidence on the genetics of sexual orientation, including family and twin studies and molecular genetics findings to date. Familial aggregation of sexual orientation has been reported in several family studies of both male and female homosexuality, although the genetic and environmental influences on this familial clustering have not been clearly defined by the largest twin studies published thus far. Recent molecular genetic studies will also be reviewed, including the two main strategies that have been used to date – linkage and association analyses. The implications of recent advances in molecular genetic studies will also be discussed.

Introduction

Sexual orientation describes what is erotically attractive to an individual and is usually consistent with sexual identity that refers to an individual's labeling of self as heterosexual, homosexual, or bisexual; both are typically not consolidated until adolescence or later. Older studies tended to define sexual orientation behaviorally, whereas most current researchers define sexual orientation psychologically. Because sexual attraction and fantasy are less likely than behavior to be constrained by societal pressures, psychological sexual orientation is thought to be a more stable and fundamental trait. A psychological definition of sexual orientation tends to produce lower prevalence figures (Dawood et al., 2009; Laumann et al., 1994) compared to a behavioral one (Chandra et al., 2011), and is used in most of the studies demonstrating familiality and heritability.

This article will provide an overview of the evidence on the genetics of sexual orientation, including family and twin studies and molecular genetics findings to date. Male and female sexual orientation appear to be distributed differently in the general population. For men, sexual orientation typically appears to be bimodally distributed, with most men clustered at the heterosexual end of a continuum, many fewer at the homosexual end, and the fewest men in the bisexual middle (Rieger et al., 2005; Bailey et al., 2000). Female sexual orientation tends to taper gradually from exclusively heterosexual to exclusively homosexual (Bailey et al., 2000). The bulk of the available evidence suggests moderate heritability for male and female sexual orientation (Dawood and Bailey, 2000; Dawood et al., 2009). Female sexual orientation has been studied less extensively, but current studies are consistent with a genetic contribution for women as well (Pattatucci and Hamer, 1995; Kirk et al., 2000). Familial aggregation of sexual orientation has been reported in several family studies of both male and female homosexuality (Dawood and Bailey, 2000), although the genetic and environmental influences on this familial clustering have not been clearly defined by the largest twin studies published thus far, which have produced contradictory results. Recent molecular genetic studies will also be reviewed, including the two main strategies that have been used to date – linkage and association analyses. The implications of recent advances in molecular genetic studies will also be discussed.

Behavior Genetic Studies

Traditionally, behavior genetic research has proceeded in three main stages. *Family studies* are initially conducted to determine whether a trait or characteristic (often called a phenotype) runs in families by comparing rates in families of probands (i.e., individuals possessing the trait) versus families of controls (that typically represent the base rate in the general population). If there is a heritable genetic contribution to a trait, one expects to find familial aggregation. However, the mere demonstration of clustering of a trait in families does not prove a genetic influence, because some traits run in families for environmental reasons.

Twin studies are conducted in order to separate genetic from familial environmental effects by comparing the similarity of monozygotic (MZ, also called identical) and same-sex dizygotic (DZ, also called fraternal) twins who have been reared together. MZ twins share all their genes in common, whereas DZ twins share 50% of their segregating genes on average, just like any other pair of full siblings. Since both twins share the same pregnancy, it is assumed that they experience essentially the same prenatal environment. Because both twins have also been reared together, their postnatal environment is assumed to be approximately equal. Thus, if MZ twins are more similar (i.e., concordant) for a trait than DZ twins, this is assumed to reflect their greater genetic similarity and is evidence that genetic factors influence the phenotype. Twin studies allow one to estimate heritability, defined as the proportion of the variance in expression of the trait due to all genetic influence(s) combined. Results from twin studies support a significant environmental contribution to most behavioral traits since MZ twins are not 100% concordant.

Adoption studies are a further method used to separate the effects of genes from the environment. Adoption produces family members who share family environment but are not genetically related, and vice versa. Thus, this method allows one to estimate the contribution of family environment to family resemblance. Studies on both male and female sexual orientation exist from all three research methodologies.

Family Studies

In the past three decades, several contemporary family genetic studies have been conducted on sexual orientation in men and women. All available studies have focused on the rate of homosexuality in siblings rather than in other first-degree relatives (e.g., parents or offspring), due to the decreased reproduction of homosexual individuals.

Both male and female homosexuality appear to run in families (Bailey and Pillard, 1991; Bailey and Benishay, 1993; Bailey et al., 1993; Bailey and Bell, 1993; Pattatucci and Hamer, 1995; Bailey et al., 1999; Schwartz et al., 2010). The rate of homosexuality among brothers of homosexual males has consistently been around 9% (Bailey and Pillard, 1991). These results have exceeded those for heterosexual controls, as well as the prevalence estimates from recent large-scale epidemiological surveys, suggesting that male homosexuality is familial. Homosexual women also appear to have more homosexual sisters than do heterosexual controls, though the familiality estimates have varied more widely for women. There is a trend for male homosexual probands to have more homosexual brothers than lesbian sisters and for the opposite pattern to be seen for lesbian probands, suggesting that at least some of the familial factors influencing male homosexuality differ from those influencing female homosexuality. However, the degree of cofamiliality (reflecting common familial influences) of male and female homosexuality remains inconclusive (Dawood et al., 2000).

Twin Studies

Several large-scale twin studies have been conducted in the past two decades, and they have been generally consistent in detecting moderate to large heritabilities for both male and female sexual orientation. However, there have been methodological limitations, in particular, several of the large twin studies of sexual orientation recruited probands via advertisements in gay or lesbian publications. Such sampling is likely to result in volunteer bias that affects twin concordances and heritability analyses, though in most scenarios this would be more likely to lead to a false-negative study. More recently, investigators have reported results from population-based twin studies, and the discussion here will focus on these large-scale findings.

Bailey et al. (2000) systematically recruited twins from a twin registry and reported lower twin concordances for homosexuality than in prior studies, although their findings were also consistent with moderate to large heritabilities for male and female sexual orientation. A further analysis of these data using multivariate structural equation modeling estimated heritability of the latent variable of male homosexuality around 30% and for female homosexuality around 50% (Kirk et al., 2000).

Kendler et al. (2000) recruited 756 twin and nontwin sibling pairs from a US national probability sample and found that 32% of MZ twins were concordant for homosexual orientation versus 13% of DZ same-sex twins. The concordance rates for MZ twins were lower than those reported in previous studies that recruited via advertisements in gay publications suggesting that twin pairs concordant for sexual orientation may be more likely to respond to such ads than are twin pairs discordant for sexual orientation. It is also worth noting that unlike previous studies in this area, this study assessed sexual orientation by a single item on a self-report questionnaire.

Bailey et al. (2000) recruited male and female twin pairs systematically from the Australian Twin Registry and assessed their sexual orientation as well as two related traits: childhood gender nonconformity and continuous gender identity. Twin concordances for homosexual orientation were lower than in prior studies with 20% of male MZ twins concordant versus 0% of DZ twins, and 24% of female MZ twins versus 10.5% of DZ twins. Univariate analyses showed that familial factors were important for all traits, but were less successful in distinguishing genetic from shared environmental influences. Multivariate analyses suggested that the causal architecture differed between men and women, and, for women, provided significant evidence for the importance of genetic factors to the traits' covariation.

Most recently, two large-scale twin studies examining sexual orientation in men and women were reported with similar results. Alanko et al. (2010) conducted a study of genetic effects of childhood gender atypical behavior and adult sexual orientation in a population-based sample of 3261 Finnish twins. Quantitative genetic analyses demonstrated that variation in both traits was partly due to genetics, with the rest explained by nonshared environmental effects. Bivariate analyses implicated substantial common genetic and modest common nonshared environmental correlations underlying the co-occurrence of childhood gender atypical behavior and adult sexual orientation (Alanko et al., 2010).

Langstrom et al. (2010) conducted the largest twin study of sexual orientation to date. The authors used data from a population-based survey of all adult twins in Sweden to estimate the influence of genetic and environmental effects on same-sex sexual behavior. A sample of 3826 MZ and DZ same-sex twin pairs were used in the study, which found familial clustering of same-sex sexual behavior in both men and women. The results indicated that same-sex behavior is influenced by both heritable and unique or nonshared environmental variables, with heritability estimates of 0.34–0.39 for men and 0.18–0.19 for women. Further, the authors note that hereditary effects appeared weaker in women, though these were not statistically significant (Langstrom et al., 2010). The study is also noteworthy for achieving greater statistical power and higher response rates than previous such research, providing the most unbiased estimates to date of genetic and nongenetic contributions to same-sex behavior.

Molecular Genetic Studies

Once a solid foundation of support for significant genetic influence on a trait has been built by means of behavior genetics (family, twin, and adoption studies), as has been the case for sexual orientation especially with males, molecular genetic studies are a next logical step. The two primary varieties of these studies are association and linkage designs.

Association studies are based on linkage disequilibrium. This means that a gene variant influencing a trait was initially

associated with specific alleles of nearby polymorphic loci. As generations (and the meioses that produce sperms and eggs) pass, the trait-influencing gene and marker allele may remain statistically associated because their proximity reduces the number of recombinations or crossing-over events that occur between them. An advantage of association tests is that the chromosomal region examined is usually much smaller than the region examined by testing for linkage in families. Association is often more powerful than linkage in that a valid association may be detected in a sample when linkage is not detectable, even when the gene is playing only a modest role. Most association studies in the past were the population-based type where the allele frequencies of a group of unrelated cases were compared against those of a group of unrelated controls, and this is the only type of association study that has been carried out on male sexual orientation. A potential pitfall of population-based case–control studies is that, although some populations appear homogeneous to superficial examination, they are in reality composed of different ancestral human groups, each one potentially with a different allele distribution at the studied loci. If one or more such groups are represented in a largely different proportion in one of the samples of an association data set (i.e., either in the controls or in the cases), false-negative or false-positive association findings may easily arise due to methodological artifact.

Linkage analysis exploits the key biological phenomenon during generation of sperm and eggs of meiotic recombination, or crossing over, during which both the maternally and paternally derived chromosomes lie in close proximity and undergo exchange of genetic material between the homologous chromosomes, e.g., between a paternally derived chromosome and its maternally derived counterpart. The chance of crossing over between two loci (locations on a chromosome) is referred to as the recombination fraction. Genes and other genetic markers (DNA sequence variations known as polymorphisms) that are close together are less likely to be separated by this process than are those that are farther apart. Therefore, they are usually inherited together by the progeny cells, and are considered genetically linked. Due largely to the complexity of the genetic contributions to male sexual orientation and uncertainty regarding key parameters (mode of inheritance, number of relevant genes, etc.), the type of linkage analyses preferred are nonparametric allele-sharing methods, and more specifically, the affected sibling pair (ASP) method. ASP designs measure the frequency with which a genetic marker allele (or variant) is inherited from a particular parent (referred to as IBD, meaning identical by descent) in a pair of siblings both manifesting the trait. The presence of a trait-influencing gene is revealed when the IBD allele sharing between affected siblings exceeds the expected 50%.

Association Studies

Two association studies have been performed to date for male homosexuality. Macke et al. (1993) used a population based case–control method to examine DNA sequence variation in the androgen receptor gene, reasoning that some variants may affect sexual differentiation of the brain. This study employed a sample of 197 homosexual males and 213 unselected (with respect to sexual orientation) male controls with variants of the androgen receptor gene located on the X chromosome (but not at Xq28), and found no significant differences in the distributions of mutations in homosexual and heterosexual men (i.e., no evidence for association). This gene was selected for examination partly due to its location on the X chromosome since there is some evidence for excess maternal transmission of male homosexuality, which would be consistent with X-linked transmission. However, using linkage analysis, the authors showed that sibling pairs concordant for homosexuality were no more likely than chance to share the same androgen receptor allele.

Another reason the androgen receptor was chosen for examination was not due to its position, but rather due to its function, which is to transduce messages from androgens ('male' hormones) to the nucleus of the cell, thus affecting other genes responsive to androgens. In addition, androgens masculinize the brain and behavior, and homosexual individuals are sex-atypical on some behavioral and neural characteristics (Rahman, 2005). In general, many different 'candidate' genes may be nominated for examination by means of association testing, but the strength of their candidacy is often in question (relative to any other gene expressed at some point in the brain, which most are). For example, studies in animal models where gene variation may be introduced and the effects systematically examined may suggest genes to examine in humans, but there is an unresolved question regarding the validity of extrapolating complex behaviors from species sometimes as different from humans as fruit flies.

Dupree et al. (2004) conducted an association study of male sexual orientation to investigate whether differences in the gene encoding the aromatase enzyme influence sexual orientation in men. Aromatase cytochrome P450 (CYP19) is necessary for the conversion of androgens to estrogens, and plays an important role in the sexual differentiation of the brain in rodents. This study found no differences between heterosexual and homosexual men in the expression of aromatase mRNA by microarray analysis, suggesting that variation in this gene is not likely to be a major factor in the development of individual differences in male sexual orientation (Dupree et al., 2004).

Linkage Studies

The findings to date from behavior and molecular genetic studies predict that the genetics of male homosexuality will not be simple, and this prediction is consistent with the results of linkage research thus far. Several linkage analyses of male homosexuality to the X chromosome have previously been reported (Hamer et al., 1993; Hu et al., 1995; Sanders et al., 1998; Rice et al., 1999). These studies have been largely based on the assumption that oligogenic transmission (a 'few' genes contribute) was most likely, and therefore relied on the ASP method of linkage analysis. While the Xq28 chromosomal region in Hamer et al.'s (1993) study showed a significant linkage signal, with supporting data in a second data set from the same group in a follow-up study (Hu et al., 1995), it is still indeterminate (as in many complex traits) whether this finding represents a true positive. An independent group (Sanders et al., 1998) found inconclusive evidence of linkage to Xq28 in 1998, and a third group (Rice et al., 1999) found no support

for linkage to Xq28 in 1999. Combining all four linkage studies, with respective affected sibling pair (ASP) sample sizes of 40 (1993), 33 (1995), 54 (1998), and 52 (1999), yields a 'multiple scan probability' (MSP) of 0.000 03, which is a suggestive *p*-value when considering all of the chromosomes, i.e., the entire genome (Sanders and Dawood, 2003). The replication MSP (excluding the original positive report from 1993) of 0.07 is at the level of a 'trend' and thus not quite statistically significant. This pattern of results is one that has been predicted for complex traits with oligogenic inheritance on the basis of simulation studies: stochastic variation in the degree of cosegregation of any one locus with a trait, which produces variation in the magnitude of linkage findings across samples. Of course, the sample size of the individual linkage studies should be considered a major factor.

Hamer et al.'s (1993) original report consisted of two major analyses: a pedigree study and a linkage study. First, they examined family pedigrees in a 'randomly ascertained' sample of homosexual probands and found that the probands had a high rate of homosexual brothers, 13.5%, relative to the base rate of homosexuality in the general population. Furthermore, their pedigrees showed an excess of homosexual uncles and male first cousins on the maternal side compared to the paternal side, though the difference was not significant. This excess was more pronounced in a subsequent analysis of an additional sample of 38 families with two homosexual brothers. If there are genes for male sexual orientation, then these families should be especially rich with them. The probands' maternal uncles and cousins (through maternal aunts; no homosexual cousins were sons of maternal uncles) had rates of male homosexuality of 10.3% and 12.9%, respectively, compared to rates of 1.5% and 3.1% for paternal uncles and cousins. This pattern of results is precisely what one would expect if an X-linked gene influenced male sexual orientation. In X-linked inheritance, males with the trait inherited the gene from their mothers, and hence have more maternal than paternal relatives with the trait.

Because of the suggestion of X linkage, Hamer et al. (1993) then searched the X chromosome using linkage analysis. Specifically, they looked at the pairs of homosexual brothers without evidence of paternal transmission (e.g., they excluded a few cases in which the father may have been homosexual; the pairs analyzed in the linkage study included all eligible pairs analyzed in both pedigree studies). They examined 22 genetic markers distributed across the X chromosome, in order to see if brothers concordant for homosexuality were also concordant for the markers. For chromosomal region Xq28 at the tip of the long arm of the X chromosome, 33 of 40 pairs of homosexual brothers shared all the markers. This was statistically different from the expected rate (20 of 40), suggesting that a gene influencing male sexual orientation lies within that chromosomal region.

To date, at least four major replication attempts have been reported, including Hu et al. (1995) from Hamer's own research group. Rice et al. (1999) obtained pedigree information from 182 families with at least two homosexual brothers. They failed to find a significant excess of homosexual maternal uncles or cousins. In a subset of 41 sibling pairs, they also failed to replicate the finding of linkage to Xq28. Unlike Hamer et al. (1993), however, Rice et al. did not exclude brother pairs with strong evidence of paternal transmission, and it is unclear how many of their subjects would have met Hamer et al.'s (1993) inclusion criteria.

In contrast, Hu et al. (1995) from Hamer's lab reported a successful replication. In this second study, Hu et al. included data from heterosexual brothers as well as homosexual brothers, and found that brothers' similarity for sexual orientation was statistically related to the sharing of Xq28 markers. That is, not only did homosexual brothers tend to share the markers, but also gay–straight pairs tended not to share the markers. The magnitude of the genetic effect was smaller in Hamer's second study, however, and the result was barely statistically significant.

Additional Molecular Genetic Studies

A host of recent studies (Mustanski et al., 2005; Bocklandt et al., 2006; Ramagopalan et al., 2010) has provided additional evidence supporting a heritable component to male and female sexual orientation, and these molecular studies are discussed in further detail below.

Mustanski et al. (2005) conducted the first full genome-wide linkage scan of sexual orientation in men by genotyping 456 individuals from 146 families with two or more homosexual brothers with 403 microsatellite markers at 10-cM intervals. They failed to replicate linkage to Xq28 in the full sample; however, they reported three new regions that approached the criteria for near significance (7q36) and for suggestive linkage (8p12 and 10q26). These chromosomal regions may be used in future replication studies with new samples and denser linkage maps.

Ramagopalan et al. (2010) reported results from a second genome-wide linkage scan of male sexual orientation in an attempt to replicate the above findings. In lieu of using microsatellite assays, the researchers conducted single-nucleotide polymorphism-based scans for linkage on the basis that these provide significantly greater genomic coverage and information content (Ramagopalan et al., 2010). The study used a Canadian cohort of 55 families with two or more homosexual brothers and reported no evidence for linkage at or around 7q32 as found by Mustanski et al. (2005) and no other genome-wide significant logorithm of odds (LOD) scores. The authors concluded that genes involved in male homosexuality probably exert more modest effects than can be detected by linkage analysis.

The most recent linkage study of male sexual orientation is also by far the largest study of its kind to date. Sanders et al. (submitted for publication) conducted a genome-wide linkage scan using DNA samples collected from 409 independent pairs of homosexual brothers (908 analyzed individuals in 384 families). It is noteworthy that all previous genetic linkage studies of male sexual orientation were conducted on smaller sample sets. Two regions of linkage were identified – the pericentromeric region on chromosome 8, and Xq28, which was implicated in prior research (Hamer et al., 1993; Hu et al., 1995). These new results appear to support the existence of genes on chromosome 8 and the X chromosome that influence the development of male sexual orientation (Sanders et al., submitted for publication).

Using a different approach that also implicates the role of the X chromosome, Bocklandt et al. (2006) measured X chromosome inactivation in a sample of 97 mothers of homosexual men and 103 age-matched control women without homosexual sons. They reported extreme skewing of X inactivation was significantly higher in mothers of homosexual men (13%) compared to controls (4%), and increased in mothers with two or more homosexual sons (23%). These findings support a role for the X chromosome in regulating sexual orientation in some homosexual men (Bocklandt et al., 2006), although it is unclear whether these unusual X-inactivation patterns influence the development of sexual orientation in sons via a direct mechanism such as the fraternal birth order effect (Blanchard, 1997), or whether it is the indirect result of a different mechanism. These results have yet to be replicated.

Higher Fecundity of Female Maternal Relatives of Homosexual Men

Camperio-Ciani et al. (2004) were the first to report a significant increase in the fecundity of female maternal relatives of homosexual men in a family study that also replicated prior research showing that homosexual men have more maternal homosexual relatives (Hamer et al., 1993), and that homosexual men tend to be later born with more older brothers than older sisters (Blanchard, 1997, 2004, 2008). This finding supports an X chromosomal genetic factor that is associated with homosexuality in men and promotes fecundity in female carriers. Since then, several studies have replicated this finding (Iemmola and Camperio-Ciani, 2009; Rahman et al., 2008; Schwartz et al., 2010; VanderLaan and Vasey, 2011; Rieger et al., 2012; Camperio-Ciani and Pellizzari, 2012), as well as a similar pattern in female maternal relatives of bisexual men (Camperio-Ciani et al., 2009; but see Blanchard, 2012). Evidence that genes predisposing to homosexuality may simultaneously increase the reproductive success of heterosexual relatives would also resolve the evolutionary conundrum of the genetic stability of male homosexuality in the general population (Zietsch et al., 2008).

The Fraternal Birth Order Effect

Several excellent reviews of biological research on human sexual orientation have been published within the past decade (Mustanski et al., 2002; Rahman, 2005), which can provide more comprehensive reviews of neuroendocrine and neurodevelopmental theories of sexual orientation. However, perhaps the most replicated finding in sexual orientation research is the fraternal birth order effect in homosexual men (Blanchard, 2008) whereby homosexual men have a greater number of older brothers than heterosexual men do, in diverse community and population-level samples. The estimated odds of being homosexual increase by approximately 33% (from the base rate in the general population) with each older brother, and statistical modeling using epidemiological procedures suggest that one in seven homosexual men owe their sexual orientation to the fraternal birth order effect (Cantor et al., 2002). While purely genetic factors could not explain this effect, recent evidence suggests that the effect is related to prenatal events (Blanchard, 2004; Bogaert, 2006). Currently, a maternal immune response to male-specific, Y-linked antigens is the most plausible explanation for this effect (Bogaert and Skorska, 2011), which becomes stronger with each male pregnancy. However, empirical studies supporting this hypothesis have yet to be conducted.

Directions for Future Research

During the past two decades, a growing body of evidence has accumulated suggesting that familial and genetic factors affect both male and female sexual orientation. The genetic evidence is substantially stronger for male than for female sexual orientation, and multiple genes are likely to influence the development of sexual orientation.

Although ongoing studies investigating genetic sources of variation in sexual orientation will contribute a critical aspect for understanding the origins and development of sexual orientation, perhaps the most interesting topic for future research in this area lies in studying nonshared environmental sources of variation, and the epigenetic relationship between environmental and genetic factors.

The study of discordant MZ twins, who are genetically identical, can be particularly important since nonshared environmental factors are responsible for MZ twins discordant for sexual orientation. For example, future molecular genetic studies could examine epigenetic modifications of DNA between twins using a genome-wide screen of differentially methylated regions to identify potential discrete differences between homosexual and heterosexual MZ twins.

At this stage, few conclusions can be drawn with certainty regarding genetic and environmental determinants of sexual orientation. Important methodological research innovations hold the most potential for furthering our knowledge on the origins and development of human sexual orientation. Future research should also attempt to integrate different biological approaches in order to provide valuable information about the specific pathways by which genes exert their influence on sexual orientation and its correlates.

See also: Behavioral Genetics: Psychological Perspectives; Gender and Social Research; Genetic Studies of Behavior: Methodology; Geriatric Psychiatry; Heterosexuality; Human Mate Choice, Evolution of; Lesbian, Gay, Bisexual and Transgender Aging; Marriage; Sexism; Sexuality and Social Work; Sexuality: Cultural Aspects; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Alanko, K., Santtila, P., Harlaar, N., Witting, K., Varjonen, M., Jern, P., Johansson, A., von der Pahlen, B., Sandnabba, N.K., 2010. Common genetic effects of gender atypical behavior in childhood and sexual orientation in adulthood: a study of Finnish twins. Archives of Sexual Behavior 39, 81–92.

Bailey, J.M., Bell, A.P., 1993. Familiality of male and female homosexuality. Behavior Genetics 23, 313–322.

Bailey, J.M., Benishay, D.S., 1993. Familial aggregation of female sexual orientation. American Journal of Psychiatry 150, 272–277.

Bailey, J.M., Dunne, M.P., Martin, N.G., 2000. Genetic and environmental influences on sexual orientation and its correlates in an Australian twin sample. Journal of Personality and Social Psychology 78, 524–536.

Bailey, J.M., Pillard, R.C., 1991. A genetic study of male sexual orientation. Archives of General Psychiatry 48, 1089–1096.

Bailey, J.M., Pillard, R.C., Dawood, K., Miller, M.B., Farrer, L.A., Trivedi, S., Murphy, R.L., 1999. A family history study of male sexual orientation using three independent samples. Behavior Genetics 29, 79–86.

Bailey, J.M., Pillard, R.C., Neale, M.C., Agyei, Y., 1993. Heritable factors influence sexual orientation in women. Archives of General Psychiatry 50, 217–223.

Blanchard, R., 1997. Birth order and sibling sex ratio in homosexual versus heterosexual males and females. Annual Review of Sex Research 8, 27–67.

Blanchard, R., 2004. Quantitative and theoretical analyses of the relation between older brothers and homosexuality in men. Journal of Theoretical Biology 230, 173–187.

Blanchard, R., 2008. Review and theory of handedness, birth order, and homosexuality in men. Laterality 13 (1), 51–70.

Blanchard, R., 2012. Fertility in the mothers of firstborn homosexual and heterosexual men. Archives of Sexual Behavior 41, 551–556.

Bogaert, A.F., 2006. Biological versus nonbiological older brothers and men's sexual orientation. Proceedings of the National Academy of Sciences of the United States of America 103 (28), 10771–10774.

Bogaert, A.F., Skorska, M., 2011. Sexual orientation, fraternal birth order, and the maternal immune hypothesis: a review. Frontiers of Neuroendocrinology 32, 247–254.

Bocklandt, S., Horvath, S., Vilain, E., Hamer, D.H., 2006. Extreme skewing of X chromosome inactivation in mothers of homosexual men. Human Genetics 118, 691–694.

Camperio-Ciani, A., Corna, F., Capiluppi, C., 2004. Evidence for maternally inherited factors favouring male homosexuality and promoting female fecundity. Proceedings of the Royal Society B: Biological Sciences 271 (1554), 2217–2221.

Camperio-Ciani, A., Iemmola, F., Blecher, S.R., 2009. Genetic factors increase fecundity in female maternal relatives of bisexual men as in homosexuals. Journal of Sexual Medicine 6, 449–455.

Camperio-Ciani, A., Pellizzari, E., 2012. Fecundity of paternal and maternal non-parental female relatives of homosexual and heterosexual men. PLoS ONE 7 (12), e51088.

Cantor, J.M., Blanchard, R., Paterson, A.D., Bogaert, A.F., 2002. How many gay men owe their sexual orientation to fraternal birth order? Archives of Sexual Behavior 31, 63–71.

Chandra, A., Mosher, W.D., Copen, C., Sionean, C., 2011. Sexual behavior, sexual attraction, and sexual identity in the United States: data from the 2006–2008 National Survey of Family Growth. National Health Statistics Reports 36, 1–36.

Dawood, K., Bailey, J.M., 2000. The genetics of human sexual orientation. In: Rodgers, J.L., Rowe, D.C., Miller, W.B. (Eds.), Genetic Influences on Human Fertility and Sexuality: Theoretical and Empirical Contributions from the Biological and Behavioral Sciences. Kluwer Academic Publishers, Boston, MA, pp. 237–252.

Dawood, K., Pillard, R.C., Horvath, C., Revelle, W., Bailey, J.M., 2000. Familial aspects of male homosexuality. Archives of Sexual Behavior 29, 155–163.

Dawood, K., Bailey, J.M., Martin, N.G., 2009. Genetic and environmental influences on sexual orientation. In: Kim, Y.-K. (Ed.), Handbook of Behavior Genetics. Springer, New York, pp. 269–279.

Dupree, M.G., Mustanski, B.S., Bocklandt, S., Nievergelt, C., Hamer, D.H., 2004. A candidate gene study of CYP19 (aromatase) and male sexual orientation. Behavior Genetics 34, 243–250.

Hamer, D.H., Hu, S., Magnuson, V.L., Hu, N., Pattatucci, A.M.L., 1993. A linkage between DNA markers on the X chromosome and male sexual orientation. Science 261, 321–327.

Hu, S., Pattatucci, A.M.L., Patterson, C., Li, L., Fulker, D., Cherny, S., Kruglyak, L., Hamer, D.H., 1995. Linkage between sexual orientation and chromosome Xq28 in males but not females. Nature Genetics 11, 248–256.

Iemmola, F., Camperio-Ciani, A., 2009. New evidence of genetic factors influencing sexual orientation in men: female fecundity increase in the maternal line. Archives of Sexual Behavior 38 (3), 393–399.

Kendler, K.S., Thornton, L.M., Gilman, S.E., Kessler, R.C., 2000. Sexual orientation in a U.S. national sample of twin and nontwin sibling pairs. American Journal of Psychiatry 157, 1843–1846.

Kirk, K.M., Bailey, J.M., Dunne, M.P., Martin, N.G., 2000. Measurement models for sexual orientation in a community twin sample. Behavior Genetics 30, 345–356.

Langstrom, N., Rahman, Q., Carlstrom, E., Lichtenstein, P., 2010. Genetic and environmental effects on same-sex sexual behavior: a population study of twins in Sweden. Archives of Sexual Behavior 39, 75–80.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago.

Macke, J.P., Bailey, J.M., King, V., Brown, T., Hamer, D., Nathans, J., 1993. Sequence variation in the androgen receptor gene is not a common determinant of male sexual orientation. American Journal of Human Genetics 53, 844–852.

Mustanski, B.S., Chivers, M.L., Bailey, J.M., 2002. A critical review of recent biological research on human sexual orientation. Annual Review of Sex Research 13, 89–140.

Mustanski, B.S., Dupree, M.G., Nievergelt, C.M., Bocklandt, S., Schork, N.J., Hamer, D.H., 2005. A genomewide scan of male sexual orientation. Human Genetics 116, 272–278.

Pattatucci, A.M.L., Hamer, D.H., 1995. Development and familiarity of sexual orientation in females. Behavior Genetics 25, 407–420.

Ramagopalan, S.V., Dyment, D.A., Handunnetthi, L., Rice, G.P., Ebers, G.C., 2010. Journal of Human Genetics 55, 131–132.

Rahman, Q., 2005. The neurodevelopment of human sexual orientation. Neuroscience and Biobehavioral Reviews 29, 1057–1066.

Rahman, Q., Collins, A., Morrison, M., Orrells, J.C., Cadinouche, K., Greenfield, S., et al., 2008. Maternal inheritance and familial fecundity factors in male homosexuality. Archives of Sexual Behavior 37 (6), 962–969.

Rice, G., Anderson, C., Risch, N., Ebers, G., 1999. Male homosexuality: absence of linkage to micro satellite markers at Xq28. Science 284, 665–667.

Rieger, G., Blanchard, R., Schwartz, G., Bailey, J.M., Sanders, A.R., 2012. Further data concerning Blanchard's (2011) "Fertility in the mothers of firstborn homosexual and heterosexual men". Archives of Sexual Behavior 41, 529–531.

Rieger, G., Chivers, M.L., Bailey, J.M., 2005. Sexual arousal patterns of bisexual men. Psychological Science 16 (8), 579–584.

Sanders, A.R., Cao, Q., Zhang, J., Badner, J.A., Goldin, L.R., Guroff, J.J., Gershon, E.S., Gejman, P.V., 1998. Genetic Linkage Study of Male Homosexual Orientation. Poster Presentation at the 151st Annual Meeting of the American Psychiatric Association. Toronto, Ontario, Canada.

Sanders, A.R., Dawood, K., 2003. Sexual orientation. In: Nature Encyclopedia of Life Sciences. Nature Publishing Group, London.

Sanders, A.R., Martin, E.R., Beecham, G.W., Guo, S., Dawood, K., Rieger, G., Badner, J.A., Gershon, E.S., Krishnappa, R.S., Kolundzija, A.B., Duan, J., Gejman, P.V., Bailey, J.M. Genome-wide Scan Demonstrates Significant Linkage for Male Sexual Orientation, manuscript submitted for publication.

Schwartz, G., Kim, R.M., Kolundzija, A.B., Rieger, G., Sanders, A.R., 2010. Biodemographic and physical correlates of sexual orientation in men. Archives of Sexual Behavior 39, 93–109.

VanderLaan, D.P., Vasey, P.L., 2011. Male sexual orientation in independent samoa: evidence for fraternal birth order and maternal fecundity effects. Archives of Sexual Behavior 40, 495–503.

Zietsch, B.P., Morleya, K.I., Shekara, S.N., Verweija, K.J.H., Keller, M.C., Macgregor, S., et al., 2008. Genetic factors predisposing to homosexuality may increase mating success in heterosexuals. Evolution and Human Behavior 29, 424–433.

Sexual Orientation: Historical and Social Construction

Jeffrey Weeks, London South Bank University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sexual orientation suggests an essential sexual nature. The task of historical or social constructionist approaches is to suggest that it is this belief that needs investigation. Such approaches seek to do two broad things: to understand the emergence of sexual categorizations (such as the 'homosexual' or the 'heterosexual' in Western cultures since the nineteenth century) within their specific historical and cultural contexts and to interpret the sexual meanings, both subjective and social, which allows people to identify with, or reject, these categorizations. It is thus largely preoccupied not with what *causes* individual desires or orientations, but with how specific definitions develop within their historical contexts, and the effects these definitions have on individual self-identifications and collective meanings.

Sexual orientation is a much used but fundamentally ambiguous concept. Although the concept had its origins in debates over the meaning of the sexual instincts in sexological theories of the nineteenth century, it came into widespread use in discussions of sexuality in the 1970s, largely as a synonym for homosexual desire and object choice, less frequently for heterosexual patterns, although in theory, it encompasses all sexual tendencies. Sexual orientation, as a concept, suggests an essential sexual nature, deeply rooted in the human personality, shaping sexual desire, behavior, and identity. The challenge that historical or social constructionist approaches have taken up is to suggest that these essentialist assumptions are what actually need investigation. Why is contemporary culture so preoccupied with discovering a fixed, true sexual nature in each individual? A more historically grounded analysis seeks to do two broad things: to understand the emergence of sexual categorizations (such as 'the homosexual' or 'the heterosexual' in Western cultures since the nineteenth century) within their specific historical and cultural contexts and to interpret the sexual meanings, both subjective and social, which allow people to identify with, or reject, these categorizations. Such analysis is thus largely preoccupied not with what *causes* individual desires or orientations, but with how specific definitions *develop* within their historic contexts, and the effects these definitions have on individual self-identifications and collective meanings.

The Rise of Historical Approaches to Sexuality

Historical and social explanations of sexual desire and behavior are usually counterposed to what are called essentialist explanations. Essentialist assumptions are central to traditional ways of thinking about sexuality, both in common sense knowledge and in theory. Until relatively recently, most writers on sexuality saw it as an irresistible natural energy, rooted in biology, and barely held in check by a thin crust of civilization. This was foundational for the pioneers of sexology from the late nineteenth century. Essentialist views of sexuality have been inextricably linked with assumptions about gender. The genital and reproductive distinctions between biological men and biological women have been read not only as necessary but also as sufficient explanation for different sexual needs and desires. They appear as the most basic distinctions between peoples. It followed that when speculations began about the origins of homosexuality, it was taken for granted that they lay in a perversion or inversion of the biological direction to the opposite sex. So the male homosexual was seen as having a female soul in a male body, or in the case of lesbianism a male soul in a female body. Later theories sought explanations in hormonal, genetic, or psychological differences, but all tended to assume a problem in the natural direction of object choice. Heterosexuality was assumed to be the normal given state of the human individual, from which everything else was an unfortunate distortion or deviation (Weeks, 2009). The implications of such imagery are deeply conservative, assuming the intractability to change of social institutions and the inevitability of certain givens, such as male/female or heterosexual/homosexual differences. Social explanations, on the contrary, have sought to question and problematize these naturalistic assumptions, and to demonstrate that what was made in history can be changed in history.

The classical starting point for social constructionist approaches is widely seen as an essay on 'The Homosexual Role' by the British sociologist, Mary McIntosh (1968). Its influence can be traced in a range of historical studies from the mid-1970s (Weeks, 1977; Trumbach, 1977), and it has been frequently anthologized (for example, Stein, 1992). What is important about the work is that it asked what was at the time a new question: not, as had been traditional in the *sexological* tradition from the late nineteenth century, what are the *causes* of homosexuality, but rather, why are we so concerned with seeing homosexuality as a condition that has causes? And in tackling that new question, McIntosh proposed an approach which opened up a new research agenda: through seeing homosexuals 'as a social category, rather than a medical or psychiatric one.' Only in this way, she suggests, can the right questions be asked, and new answers be proposed. Using Kinsey (Kinsey et al., 1948), McIntosh makes a critical distinction between homosexual behavior and 'the homosexual role.' Homosexual behavior is widespread; but distinctive roles have developed only in some cultures, and do not necessarily encompass all forms of homosexual activity. The creation of a specialized, despised, and punished role or category of homosexual, such as that which developed in Britain from the early eighteenth century, was designed, she argued, to

keep the bulk of society pure in rather the same way that the similar treatment of some kinds of criminal keeps the rest of society law abiding. McIntosh drew on a variety of intellectual sources, from structural functionalism to dramaturgical approaches, but clearly central to her argument was a form of labeling theory. The creation of the homosexual role was a form of social control designed to minoritize the experience, and protect and sustain social order and traditional sexual patterns. In other words, homosexual identifications were less the result of some inbuilt drive than of social constraints and possibilities.

If McIntosh put on the agenda the process of social categorization, another related but distinctive approach that shaped social constructionism came from the work of John Gagnon and William Simon, summarized in their book *Sexual Conduct: The Social Sources of Human Sexuality* (1974). Drawing again on the work of Kinsey, and symbolic interactionist traditions, they argued that, contra to the teachings of sexology, sexuality, far from being the essence of 'the natural,' was subject to sociocultural shaping to an extraordinary degree. The importance that culture attributes to sexuality may, therefore, they speculated, not be a result of sexuality's intrinsic significance: society may have had a need *to invent* its importance and power at some point in history. Sexual activities of all kinds, they suggested, were not the results of inbuilt, natural drives or instincts but of complex psychosocial processes of development, and it is only because sexual activities are embedded in social scripts that the physical acts themselves become important. These insights suggested the possibility of exploring the complex processes by which individuals acquired subjective meanings in interaction with significant others and the effects of 'sexual stigma' on these developmental processes (Plummer, 1975).

By the mid-1970s, it is possible to detect the clear emergence of a distinctive sociological account of sexuality, with two related concerns. One focused on the social categorization of sexuality, asking questions about what historical factors shaped the labeling and stigmatization of sexual differences, differences which then appeared as natural, but were in fact cultural. This gave rise to a major effort of historical investigation that has traced the emergence of the idea of the 'invert,' 'intermediate type,' 'third-sexer,' 'homosexual,' or 'gay' – terms have varied at different times – as a distinct type of individual. The other concern was primarily with understanding the shaping of subjective meanings of sexuality through socio-psychological processes and enacting historically shaped sexual scripts, which could explain how individuals defined themselves as sexually different in specific historical circumstances.

Following the approaches pioneered by sociologists and social psychologists such as McIntosh and Gagnon and Simon, a third theoretical element now came into play, represented by the work of Michel Foucault (1976/1979). Foucault's introductory essay on *The History of Sexuality* is often seen, misleadingly, as the starting point of historical and constructionist approaches, but there can be no doubt of the subsequent impact of what was planned as a brief introduction to a multivolume study. Like Gagnon and Simon, Foucault appeared to be arguing that 'sexuality' was a 'historical invention.' Like McIntosh, and others who had been influenced by her, he saw the emergence of the concept of a distinctive homosexual personage as a historical process, with the late nineteenth century as the key moment. The concept of the homosexual was not the result of a neutral scientific discovery. It was produced through discourses that worked simultaneously to invent same-sex subjects and regulate them, through a complex of power and knowledge. The process of medicalization, in particular, was seen as a vital explanatory factor: homosexuality was increasingly defined as a specific psychological or medical condition that distinguished some people from others. Like McIntosh, Foucault suggested that psychologists and psychiatrists have not been objective scientists of desire, as the sexological tradition proclaimed; but on the contrary, they have been diagnostic agents in the process of social definition and regulation. But at the same time, his suggestion that people do not react passively to social categorization – 'where there is power, there is resistance' – left open the question of how individuals react to social definitions. How, in a phrase that now became central to the debate, identities are formed in the intersection of social and subjective meanings.

Sexual Behavior, Sexual Categories, and Sexual Identities

The most crucial distinction for historical and social explanations is between sexual behavior (how we practice sexuality), categories (how sexual types are defined), and identities (how we see ourselves and people like us, and distinguish ourselves from others). Kinsey et al. (1948) had shown that there was no necessary connection at all between what people did sexually and how they identified themselves. If, in a much disputed figure, 37% of the male population had some sort of sexual contact with other men to the point of orgasm, yet a much smaller percentage claimed to be exclusively homosexual, then identity had to be explained by something other than sexual proclivity or practice. Sexual identity as an idea was much more complex than it seemed. It proved, nevertheless, to be a powerful mobilizing idea for those who sought to affirm their sense of self and being through a public identification. By the 1970s, many self-proclaimed homosexuals were 'coming out,' in the wake of the new lesbian and gay movement, and defining themselves very clearly as gay, lesbian, or bisexual. Many of those who came out saw in the historicization of the homosexual category a way of understanding how and why same-sex desire had carried such discrimination, prejudice, and hostility. Others who came out, however, were strengthened in their belief that homosexuality was not only intrinsic to their sense of self and social identity but also expressed a distinctive orientation that was essential to their nature. This difference between seeing it as a useful category and seeing it as an essential part of their makeup was at the heart of the so-called social constructionist–essentialist controversy in the 1970s and the 1980s (Stein, 1992). For many of those who came out, a critique of essentialism could also be conceived of as an attack on the very idea of a homosexual identity, a fundamental challenge to the hard-won gains of the lesbian and gay movement, and the claim to recognition of homosexuals as a legitimate minority group. This was the source of the appeal of

subsequent theories of a 'gay gene' or 'gay brain,' which suggested that sexual orientation was wired into the human individual (Wilson and Rahman, 2005). It is probably the case that the majority of the lesbian and gay world, especially in the United States, strongly believes that a separate, natural orientation is the basis of their identities. It seems to correspond with common sense understandings of how individuals feel about their sexual desires.

It is important to make several clear points in response to these constructionist-essentialist debates, where social scientific debates became a marker of social movement differences. First, the distinction between behavior, categories, and identities need not necessarily require the ignoring of questions of causation; constructionist theory merely suspends them as irrelevant to the question of the social organization of sexuality. Foucault himself stated that : "On this question I have absolutely nothing to say" (cited in Halperin, 1995: p. 4). The really important issue is not whether there is a biological or psychological propensity that distinguishes those who are sexually attracted to people of the same gender from those who are not. More fundamental are the meanings that these propensities acquire, however, or why ever they occur; the social categorizations that attempt to demarcate the boundaries of meanings; and their effect on collective attitudes and individual sense of self. Social categorizations have effects in the real world, whether or not they are direct reflections of inbuilt qualities and drives.

The second point to be made is that terms such as constructionism and roles should not be seen in any deterministic or mechanistic fashion. They are in the end no more than heuristic devices to identify and understand a problem in studying sexuality in general and homosexuality in particular. It is transparently obvious that the forms of behavior, practices, identity, institutional arrangements, regulation, beliefs, ideologies, and even the various definitions of 'sexual' vary enormously through time and across cultures and subcultures. A major objective of historical and social constructionist studies of the erotic has been to challenge the taken for granted, to denaturalize sexuality in order to understand its human dimensions and the coils of power in which it is entwined and how it is shaped by historical forces and events, and by individual and collective agency. This is a continuous process. In the wake of positive lesbian and gay identities, there has been a proliferation of other identities, from bisexual to transgender, which work to trouble the binary divides of heterosexuality/homosexuality and masculinity/femininity.

The third point that requires underlining is that, regardless of evidence for the contingency of sexual identities, this should not imply that personal sexual identities, once acquired by individuals, are irrelevant and can readily be sloughed off. That categories and social identities are shaped in history does not in any way undermine the fact that they are fully lived as real by people. The complex relationship between societal categorization and the formation of subjectivities and sexual identities has in fact been a key focus of writing about homosexuality since the mid-1970s. On the one hand, there is a need to understand the classifying and categorizing processes that have shaped our cultural concepts of homosexuality – the law, medicine, religion, patterns of stigmatization, as well as formal and informal patterns of social regulation. On the other, it is necessary to understand the level of individual and collective reception of, and battle with, these classifications and categorizations – the level of subjectivity and agency. The best historical work has attempted to hold these two levels together, avoiding both sociological determinism (you are what society dictates) and extreme voluntarism (you can be anything you want): neither is true (see discussion in Vance, 1989).

Some of the most interesting work on the tension between sociological determinism and voluntarism has attempted to explore the subcultures, networks, urban spaces, or even rural idylls that provided the terrain, and the conditions of possibility, for the emergence of distinctive sexual identities. McIntosh's suggestion that the late seventeenth century saw the emergence of a subcultural context for a distinctive homosexual role in England has been enormously influential. Her rediscovery of the London mollies' clubs has been the starting point of numerous historical excavations (for example, Trumbach, 1977; Bray, 1982). There is now plentiful work that attempts to show that homosexual subcultures and identities existed before the late seventeenth century – for example, in the early Christian world (Boswell, 1980), or in other parts of Europe (see essays in Herdt, 1994) – just as there have been scholars who have argued that we cannot really talk about homosexual identities until the late nineteenth, or even mid-twentieth centuries. There is a real historical debate. As a result, it would now seem remarkable to discuss sexual identities (and their complex relationship to social categorizations) without a sense of their historical and social context. Sexual *identities* are made in history, not in nature.

Heterosexuality and Homosexuality

Identities are not created in isolation from other social phenomena. In particular, a history of homosexual identities cannot possibly be a history of a single homogeneous entity, because the very notion of homosexuality is dependent, at the very least, on the existence of a concept of heterosexuality, which in turn presupposes a binary notion of gender. Only if there is a sharply demarcated difference between men and women in the culture does it become meaningful to distinguish same sex from other sex relationships. Social constructionism originally paid little attention to the construction of heterosexuality. But without a wider recognition that 'the heterosexual' was also a social construction, attempts to explain the invention of 'the homosexual' made little sense. The two concepts are inextricably married together. One of the early attractions of the first volume of Foucault's *The History of Sexuality* was precisely that it both offered an account of the birth of the modern category of homosexual and put that into a broader historical framework. Foucault did this by postulating the invention of sexuality as a category in Western thought, and in delineating the shifting relationships between men and women, adults and children, and the normal and the perverse as constituent elements in this process. Foucault himself was criticized for putting insufficient emphasis on the gendered nature of this process, but this was more than compensated for by others in their developing of feminist critiques of 'compulsory heterosexuality' or 'the heterosexual institution,' with its own complex history (Rich, 1980; Jackson and Scott, 2010).

Central to these debates on the institutionalization of heterosexuality was the perception that although sexuality potentially takes many forms, the ways in which it is lived are never simply a spectrum where homosexuality and heterosexuality sit equally side by side. Historically, sexuality has been structured in dominance, with heterosexuality privileged, and that privilege is essentially male oriented. Homosexuality has been constructed as a subordinate formation within the 'heterosexual continuum,' with male and female homosexuality having a different relationship to the dominant forms. In turn, once this is recognized, it becomes both possible and necessary to explore the socially constructed patterns of femininity and masculinity as power relationships (Connell, 2009). Increasingly, in a range of disciplines, power came to be seen as the key element that needed to be explored and understood.

Although the constructionist debates began within the disciplines of sociology and history, later developments, taking forward both theoretical and political (especially feminist) interventions, owed a great deal to poststructuralist and deconstructionist literary studies, and to the emergence of 'queer studies.' Whereas history and sociology characteristically attempted to produce order and pattern out of the chaos of events, the main feature of the poststructuralist and deconstructuralist (hereafter poststructuralist) approaches at first was to show the binary conflicts reflected in literary texts. The texts are read as sites of gender and sexual contestation, and therefore of power and resistance (Sedgwick, 1990). Sedgwick's work and that of the American philosopher Judith Butler (1990) were in part attempts to move away from the essentialist/constructionist binaries by emphasizing the 'performative' nature of sex and gender. For Butler, there is no originating 'sex' on which gender differences or the homosexual/heterosexual binary are built. On the contrary, gender differences are shaped by their incessant enactment and repetition. They come to seem so fixed, embedded, and embodied because they are constantly acted out in historically given institutional frameworks.

Such poststructuralist work in turn opened up what might be called the 'queer moment,' which radically challenged the relevance of all fixed sexual categorizations. For queer theorists, the perverse is the worm at the center of the normal, giving rise to sexual and cultural dissidence and a transgressive ethic, which constantly works to unsettle binarism and to suggest alternatives.

Whereas early constructionism sought to explain how meanings and identities were shaped in historical circumstances, queer theory seeks to demonstrate how all fixed concepts are inextricably involved in the flow of power. This theoretical position was separate from but contingently linked to a new outburst of sexual radicalism at the end of the 1980s and the early 1990s, which was both a reaction to the acquired immunodeficiency syndrome crisis on the gay world, especially in the United States, and a growing frustration with what was seen as the conservative, assimilationist tendencies of many gay communities (Warner, 1999). The problem with both the poststructuralist theoretical tendencies and the related new forms of militancy was that they were rooted in very special cultural and political contexts, in highly developed Western societies.

Comparative Perspectives

One of the main lessons of historical analyses of heterosexuality and homosexuality is the recognition that such concepts are peculiar to a very specific historical period, from the nineteenth century onward, in a distinct region of the world, largely Western Europe and North America. Much of the debate about the homosexual/heterosexual binary divide was based on the perceived Western experience. Yet from the beginning, comparisons with non-Western sexual patterns were central to historical and constructionist perspectives, and it was immediately obvious how differently sexuality was lived elsewhere. Foucault (1976/1979) compared the Western 'science of sex' with the non-Western 'erotic arts,' and although he has been subsequently critiqued for his own culture-bound assumptions, his addressing of the 'erotic arts' clearly opened the way to more nuanced analysis.

It was the very fact of different patterns of 'institutionalized homosexuality' on a global scale that had formed the starting point of McIntosh's essay. She identified two key regulating elements that could be observed in different parts of the world. The first was the erotic entanglements between adults and nonadults (as in the intergenerational sex, which denoted the passage from childhood to adulthood in some tribal and premodern societies). The second was based on gender inversion (classically as in the case of the native American 'berdache,' where a sort of third sex was seen as emerging, a pattern that can also be traced in the association of male homosexuality with effeminacy, or lesbianism with hypermasculine behavior). Neither pattern necessarily implies the existence of a particular homosexual condition. For some writers, these patterns are the keys to understanding homosexuality in premodern times. Historians have traced the evolution over time of patterns of homosexual life, which have shifted from intergenerational ordering, through categorization around gender and class, to recognizably modern forms of egalitarian relationships (for a historical perspective, see Trumbach, 1998).

It would be misleading, however, to believe that these broad patterns fully encompass the diverse ways in which sexuality is lived in everyday lives at any particular time. Individual identities and social positions are shaped not only by and at the intersection of a host of often conflicting forces, typically dynamics of class, gender, race, and ethnicity, but also embracing a host of other determining influences including nationality, faith, geography, age and generation, ability and disability, as well as sexuality (Taylor et al., 2010). The interlocking, mutually reinforcing impact of various forms of power and domination around race, gender, and sexuality was first articulated by the US and European–Black feminists as it became clear that the universalizing tendencies of white Western feminist analyses, especially around issues like the family, did not speak to the experience of all women. What their intersectionalist analyses sought to do was to show how such experiences worked together to reinforce one another in the uniqueness of subject position: they were lived not as separate forms of oppression but as inextricably interlocking and densely lived experiences, producing their own forms of resistance. The force of these analyses was to highlight the reality that identification is always complex, that subjectivity is messy, and that personhood is inextricably bound up with

relations of power. Some identities and experience of oppression are more important at some time rather than another, giving rise to different forms of political struggle, and different identities and subjectivities.

If this is true within Western cultures, it is even more relevant when sexuality is viewed within a global perspective. Although Altman (2001) has discussed the emergence of the 'global gay,' which is based on the increasingly generalized characteristics of sexual lifestyles in an increasingly globalized world, there is plentiful evidence at the same time of the different ways in which homosexuality and heterosexuality are lived in different cultures across the world, and where the idea of a fixed orientation is remote from everyday experience.

The forces of globalization are bringing into contestation and confrontation different beliefs, behaviors, and assumptions, and in the process are reshaping the context and meanings of intimacy and the erotic. While these globalizing processes affect all parts of the world, their impact is uneven on individuals, groups, states, and regions because they are enmeshed in huge disparities of power and gross inequalities, and this is manifest in persistent inequities between cultures and in continuing sexual injustices, especially against women, children, and lesbian, gay, or transgendered identified peoples. Globalization in many ways accentuates and dramatizes the inequalities that already existed. Members of new global elites have more in common with each other than with the poor in their own nation states. People with the latest hi-tech gadgets rub shoulders with people who have nothing. The new digital technologies may create new forms of inequality, new desires, wants, and needs, while people lose the comforts of their lost worlds, and are forced defenseless or ill equipped into a global marketplace – including the sexual marketplace. Yet, alongside these dangers are new possibilities of links across difference, not least through the explosion of the Internet and social networking, giving rise to new subjectivities, and to claims to social justice and new discourses of human sexual rights (Aggleton et al., 2012). The growing profile of lesbian, gay, bisexual, transgender (LGBT) issues in political and cultural debates across the globe has had complex and often contradictory effects. From Iran to Zimbabwe, Iraq to Uganda, governments and fundamentalist religious movements have used prejudice against homosexuality to mobilize against perceived Western influences and in favor of a highly selective version of traditional values. At the same time, human rights claims have been advanced to support beleaguered LGBT people, and to produce common identifications and solidarities.

Beyond Constructionism

Historical and social constructionism has advanced and changed rapidly since the 1970s. The 'category' that early scholars were anxious to deconstruct has become 'categories' that proliferate in contemporary societies. 'Roles,' neat slots into which people could be expected to fit as a response to the bidding of the agents of social control, have become 'performances' (Butler, 1990) or 'necessary fictions' (Weeks, 1995) whose contingencies demand exploration. 'Identities,' which once seemed categorical, are now increasingly seen as fluid, relational, and hybrid: people are not quite today what they were yesterday, or will be tomorrow. Personal 'narratives,' stories people tell each other in the various interpretive communities to which they belong, are increasingly seen as critical in shaping our identifications. Individual identities, far from being automatic responses to internal orientations, are negotiated in the ever-changing relationship between self and other, within rapidly changing social structures and meanings (Plummer, 1995, 2003).

There is a paradox here. While historical and constructionist approaches have problematized and deconstructed identities, and shown their social origins, it has not stopped an ever-growing proliferation of identifications and subjectivities. Even heterosexuality, which for so long was the silent norm, so taken for granted, like the air we breathe, that it was rarely articulated, has found a new voice. People now talk more explicitly about their heterosexual orientation as a way of defining themselves against assertive homosexual subjectivities. Among those who for long were disqualified for their sexual or gender nonconformity the change has been more striking. Terms such as queer or omnibus descriptions such as LGBTQI (lesbian, gay, bisexual, transgender, queer/querying/questioning, intersex…, the listing is potentially long) attest to the vitality of sexual and gender identifications and battles over meanings. Sexuality is no longer a destiny, but a focus for defining and redefining a creative life.

Sexual orientation may, or may not, be a product of genetics, psychosocial structuring, or environmental pressures. That issue, which has tortured sexology for over a century, may or may not be resolved at some stage in the future. For those who seek historical and social understanding, however, other questions are central: not what causes the variety of sexual desires, preferences, or orientations that have existed in various societies at different times, but how societies shape meanings around sexual diversity and the effects these have on individual lives. That puts the emphasis where it should be, on human endeavor in ever-changing historical circumstances.

See also: Gay and Lesbian Movements; Heterosexism and Homophobia; Identity Movements; Same-Sex Sexualities in Turkey; Sexual Minorities and Sexual Citizenship; Sexual Orientation: Biological Influences.

Bibliography

Aggleton, P., Boyce, P., Moore, H., Parker, R. (Eds.), 2012. Understanding Global Sexualities: New Frontiers. Routledge, Abingdon and New York.

Altman, D., Vance, C., Vicinus, M., Weeks, J., et al., 1989. Homosexuality, Which Homosexuality? GMP Publishers, London.

Altman, D., 2001. Global Sex. University of Chicago Press, Chicago.

Boswell, J., 1980. Christianity, Social Tolerance, and Homosexuality: Gay People in Western Europe from the Beginning of the Christian Era to the Fourteenth Century. University of Chicago Press, Chicago and London.

Bray, A., 1982. Homosexuality in Renaissance England. GMP Publishers, London.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York and London.

Connell, R.W., 2009. Gender. Polity Press, Cambridge.

Foucault, M., 1976/1979. Histoire de la sexualite. In: La Volunte de Savoir, vol. 1. Gallimard, Paris. Published in translation as The History of Sexuality, vol. 1, An Introduction. Allen Lane, London.

Gagnon, J.H., Simon, W., 1974. Sexual Conduct: The Social Sources of Human Sexuality. Hutchinson, London.
Halperin, D., 1995. Saint Foucault: Towards a Gay Hagiography. Oxford University Press, Oxford and New York.
Herdt, G. (Ed.), 1994. Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History. Zone Books, New York.
Jackson, S., Scott, S., 2010. Theorising Sexuality. Open University Press, Maidenhead.
Kinsey, A.C., Pomeroy, W.B., Martin, C.E., 1948. Sexual Behavior in the Human Male. W.B. Saunders Company, Philadelphia and London.
McIntosh, M., 1968. The homosexual role. Social Problems 16 (2), 182–192.
Plummer, K., 1975. Sexual Stigma: An Interactionist Account. Routledge and Kegan Paul, London.
Plummer, K., 1995. Telling Sexual Stories: Power, Change and Social Worlds. Routledge, London and New York.
Plummer, K., 2003. Intimate Citizenship: Private Decisions and Public Dialogues. University of Washington Press, Seattle.
Rich, A., 1980. Compulsory homosexuality and lesbian existence. Signs 5 (4), 631–661.
Sedgwick, E.K., 1990. Epistemology of the Closet. University of California Press, Berkeley and Los Angeles.
Stein, E. (Ed.), 1992. Forms of Desire: Sexual Orientation and the Social Constructionist Controversy. Routledge, London and New York.
Taylor, Y., Hines, S., Casey, M. (Eds.), 2010. Theorizing Intersectionality and Sexuality. Palgrave, London.
Trumbach, R., 1977. London's sodomites: homosexual behavior and western culture in the 18th century. Journal of Social History 11 (1), 1–33. Fall.
Trumbach, R., 1998. Sex and the Gender Revolution, Volume 1, Heterosexuality and the Third Gender in Enlightenment. Chicago University Press, Chicago and London.
Vance, C.S., 1989. Social construction theory: problems in the history of sexuality. In: Altman, D., et al. (Eds.), Homosexuality, Which Homosexuality? GMP, London, pp. 13–34.
Warner, M., 1999. The Trouble with Normal: Sex, Politics and the Ethics of Queer Life. The Free Press, New York.
Weeks, J., 1977. Coming Out: Homosexual Politics in Britain from the Nineteenth Century to the Present. Quartet Books, London.
Weeks, J., 1995. Invented Moralities: Sexual Values in an Age of Uncertainty. Polity Press, Cambridge. Columbia University Press, New York.
Weeks, J., 2009. Sexuality, third ed. Routledge, Abingdon and New York.
Wilson, G., Rahman, Q., 2005. Born Gay: The Psychology of Sex Orientation. Peter Owen, London.

Sexual Politics in the Arab World

Shereen El Feki, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the late twentieth century, attitudes toward nonconforming sexualities and gender roles in the Arab region have hardened, a trend fueled by a rise in Islamic conservatism and exploited by authoritarian regimes. Although political cultures have proved slow to change in the wake of the 'Arab Spring,' a growing freedom of expression is opening space for discreet challenges to the sexual status quo, part of wider debates over human rights and personal liberties in the emerging political order.

Introduction

The Arab region, also called the Middle East and North Africa (MENA), is vast and varied – 22 countries, 370 million people, three major religions, dozens of sects and ethnic groups – with as much diversity within states as there is between them. Yet running across these borders are three red lines – subjects not to be challenged in word or deed. The first of these is politics. The events of the 'Arab Spring', with mass uprisings and violent conflict in Tunisia, Libya, Egypt, Yemen, Bahrain, and Syria, have taken a chisel to this line and begun the process of chipping away at received political wisdoms – even if building functioning democracies is proving, in most cases, more difficult than bringing down decades-old dictatorships. The second red line is religion, and in particular, the region's dominant faith, Sunni Islam. That too is coming into question, with increasing public debate over the role of religion in public life, in response the rise and subsequent travails of political Islam in Egypt, Tunisia, and Morocco, and recent gains by Islamic extremists in Syria, Libya, and Iraq.

The third red line is sex. Or more broadly defined, sexuality: that is, the sexual act and all that goes with it, including gender roles and identity, sexual orientation, pleasure, intimacy, eroticism, and reproduction (World Health Organization, 2006). Across the Arab world today, the only socially accepted sexuality is strict heteronormativity, its cornerstone family-endorsed, religiously sanctioned, state-registered marriage. Anything outside of this context is haram (forbidden), *illit adab* (impolite), *'ayb* and *hchouma* (shameful) – a seemingly endless lexicon of reproof (El Feki, 2014).

Public opinion appears to uphold this sexual status quo. In World Values Surveys, conducted between 2010 and 2014, strong majorities across the Arab region rejected homosexuality, abortion, premarital sex as 'unjustifiable.' This pattern is in stark contrast to publics in Western Europe, Australasia, and the United States, where less than a quarter of those surveyed considered such practices unacceptable. Yet on questions of democracy and its importance to their own lives, publics were far more closely aligned, even with popular disenchantment in the aftermath of the Arab Spring (**Table 1**). It remains the case, as scholars Norris and Inglehart observed in 2002, that the gulf separating Middle East and West "involves Eros far more than Demos" (Norris and Inglehart, 2002).

Sex has long been a dividing line between the West and the Arab world. A keynote of the Orientalist critique since the birth of Islam, Arab cultures were occasionally celebrated, but more often denigrated, for their perceived sexual license by more than a millennium of European observers (Said, 1995; Daniel, 1993). Today, positions are largely reversed, with the Arab, and broader Islamic, world widely criticized for sexual intolerance. It is not just social liberals, within and outside the region, who hold this view. It has also become a central plank of the 'Islamophobic' discourse of conservatives in America and Europe, the self-proclaimed last stand

Table 1 Attitudes towards sex and politics, % agreeing with above statements, select countries

	Premarital sex never justifiable	*Abortion never justifiable*	*Homosexuality never justifiable*	*Democracy absolutely important*
Morocco	69.3	71.7	83.6	58.7
Tunisia	n/a	70.4	87.1	57.1
Jordan	86.8	66.0	85.6	43.2
Lebanon	n/a	44.7	47.7	33.0
Qatar	90.9	76.1	91.7	54.1
Iraq	n/a	47.9	69.9	31.2
Australia	6.0	15.7	13.1	58.7
Germany	8.0	22.5	17.8	58.8
New Zealand	9.5	17.4	19.6	52.6
Sweden	3.6	4.7	7.5	73.5
United States	13.5	22.4	24.0	46.5

Source: World Values Survey 2010–2014.

in the battle between 'Western' values and the depredations of 'radical' Islam, particularly as they relate to the rights of women (Saunders, 2013). And the West, once praised by some nineteenth-century Arab observers for its hard line on same–sex relations, is now seen by many within the region as a radiating source of sexual debauchery from which the region must be shielded.

These conflicts play out in international fora, notably at the United Nations, with battle lines clearly drawn over sexual rights: that is, the freedom to access sexual and reproductive health services and to generate, share, and consume ideas and information about sexuality; the right to choose one's own partner and to be sexually active, or not, in consensual relations; the freedom to decide whether to have children, and when; the right to control one's own body and the liberty to pursue a satisfying, safe, and pleasurable sexual life, without coercion, discrimination, or violence (IPPF, 2008).

Since the 1994 International Conference on Population and Development in Cairo, civil society and governments around the world have struggled to enshrine sexual rights in international agreements, in the face of strong resistance from Arab states, collected under the umbrella of the Organization of Islamic Cooperation (OIC). More recently, further alliances have formed with member states of the African Union, as well as the Vatican and its allies, to oppose recognition of rights on the basis of sexual orientation and gender identity, and to enshrine the legitimacy of 'traditional values' in international contexts. Some of the longest-standing struggles are over the Convention on the Elimination of all Forms of Discrimination Against Women (CEDAW); while 20 of the 22 Arab states have ratified or acceded to the Convention, almost three-quarters of these have also lodged reservations to assorted articles, largely on the grounds of potential conflict with shari'a, which informs the majority of their constitutions. Egypt's Muslim Brotherhood spoke for many conservative opponents in the region when it attacked recent attempts to strengthen CEDAW's provisions as a slippery slope to free love, homosexuality, prostitution, in an effort to 'destroy Islamic ethics,' 'demolish the institution of the family,' ending in 'the last step of intellectual and cultural invasion.'

The rise of Islamic fundamentalism in the Arab region since the 1970s has significantly hardened, and politicized, attitudes toward gender roles and sexuality, for both men and women. While gains have been made in a number of countries on aspects of sexual and reproductive rights, particularly as they relate to marriage (for example, Morocco's 2004 revisions to the *Moudawana*, or family law code, and Egypt's 2008 amendments to the Child Law), these were largely introduced by authoritarian regimes in the absence of wider political, economic and social reforms recognizing and respecting fundamental human rights. The calls for justice, freedom, and dignity which mobilized millions during the Arab Spring – as well as equality, autonomy, integrity, and privacy – are as yet, unanswered and whether such reforms will extend to sexual and reproductive rights remains to be seen. This is a pressing issue for two populations outside the staunchly defended citadel of marriage, on the sharp end of political and social regulation: unmarried youth, and lesbian, gay, bisexual, transgender/transsexual, queer (LGBTQ) men and women.

Youth, Sexuality, and Marriage

Marriage is a pivotal rite of passage for the estimated 100 million people between 15 and 29 years of age in the Arab world (Dhillon and Yousef, 2009). Early marriage remains a risk for poor, less educated, and rural girls across the region – particularly those in desperate circumstances, such as Syrian refugees trafficked into sex work (masquerading as short-term marital unions) by their impoverished families. However, early marriage and motherhood, which was the norm in the most of the region two to three generations ago, is a declining phenomenon, even in countries such as Yemen where tradition and religious conservatism have long militated against setting 18 as the minimum legal age of marriage (UNFPA, 2012).

It is the other end of the singulate mean that is now a widespread source of concern. The singulate mean age of marriage is rising in most of the Arab region, in some cases dramatically: in Morocco, and Tunisia, for example, on average women are first marrying in their late 20s and men in their early 30s (**Table 2**). This delay is due in part to economics, matrimony having become an expensive proposition in the consumer-culture economies of Arab region. Tradition and religion dictate that men and their families cover the costs of marriage, from the big white wedding and beyond, but double-digit unemployment among youth – a major driver of the recent uprisings – means that men are obliged to wait to tie the knot.

At the same time, significant advances in female education, and the small, but growing, presence of women in the workplace is also pushing up the age of marriage of women. Indeed, there is rising moral panic in many countries over the phenomenon of *'anusa*, or spinsterhood – educated women who fail to find a husband, and therefore remain at home, dependent on their families. Divorce is causing similar social anxiety, particularly in the wealthier Gulf states that fear a rising tide of family disintegration and a loss of national identity in the face of mass migration of foreign workers to their rapidly developing economies.

Nuptiality is never very far from politics. In the United Arab Emirates (UAE), for example, government-sponsored marriage funds offer financial subsidies to promote intra-Emirati weddings (in the hope of discouraging marriage with foreigners), and provide further support services in an effort to stave off divorce. Elsewhere, notably Lebanon, legal restrictions on individuals marrying outside their religious or ethnic group within the country, or prohibiting women from passing nationality to their foreign husbands (or the children of such

Table 2 Mean age at first marriage

Country	*Men*	*Women*
Bahrain	27	23
Egypt	29	24
Jordan	30	26
Kuwait	28	25
Morocco	31	27
Palestine	25	20
Tunisia	33	28
Saudi Arabia	25	20
Yemen	25	22

Source: ESCWA, 2004–2012.

unions) face growing challenge from civil society. Nonetheless, fears of the decline of marriage are greatly exaggerated: beyond the headlines, official statistics show that the vast majority of people in the Arab region get, and stay, married, though they are waiting longer now to do so.

As a consequence, there are more single youth in the region than ever before. This is a generation caught between biology and sociology, reaching sexual maturity in a climate that is reluctant to countenance any alternative to sex within marriage. Given rising religiosity among much of the region's youth, a significant but highly speculative number are also turning to alternative forms of marriage within Islam, among them *mut'a* (temporary 'pleasure' marriage, permitted under Shi'i Islam) and *'urfi* (so-called customary marriage, increasingly popular in Sunni-majority countries) in an attempt to give their sexual relations a religious cover, however controversial (Hasso, 2010). For most young people, however, these remain more of a last resort than a lifestyle choice.

Official reluctance to fully face up to this cultural and demographic dilemma makes research on youth sexuality often difficult. However, a window is opening, thanks to human immunodeficiency virus (HIV). From a low base, the Middle East and North Africa (MENA) is now one of only two world regions where HIV is still on the rise: new infections and acquired immunodeficiency syndrome (AIDS)-related deaths have doubled over the past decade (UNAIDS, 2014; **Table 3**). Over the past decade, HIV has provided the socially acceptable context of public health to ask unmarried youth hard questions about their sexual knowledge, attitudes, behaviors and practices.

In Morocco, Algeria, Tunisia, Jordan, Lebanon, West Bank – anywhere large-scale quantitative studies have been conducted – a similar pattern emerges: upwards of half of young men say they are sexually active before marriage, generally making their debut in their mid-to late teens and taking multiple partners along the way. Condoms are rarely used, in large measure because of their popular association with *zina*, which is religiously proscribed sex outside of marriage (El Feki, 2014).

Table 3 HIV/AIDS in the Middle East and North Africa, 1995–2013

Year	*People living with HIV (median estimate)*	*AIDS-related deaths (median estimate)*
1995	42 000	1600
1996	50 000	2100
1997	60 000	2600
1998	70 000	3200
1999	82 000	3900
2000	93 000	4600
2001	110 000	5300
2002	120 000	6100
2003	130 000	7100
2004	140 000	7800
2005	160 000	8800
2006	170 000	9700
2007	180 000	10 000
2008	190 000	11 000
2009	200 000	12 000
2010	210 000	12 000
2011	220 000	13 000
2012	230 000	14 000
2013	230 000	15 000

Source: UNAIDS.

In general, far smaller percentages of young women are willing to admit to premarital sex. This is due to the primacy of female virginity. Sex before marriage is legally prohibited for both sexes – at least on paper – in more than fourth-fifths of Arab states, but law does little to influence people's sexual behavior anywhere in the world. Rather it is patriarchy, buttressed by religious conservatism, that maintains a pervasive double standard, in which women are expected to turn up on their wedding nights with their hymens intact, but men who have sex before marriage are largely indulged.

Indeed, female virginity is rarely a private matter in communities across the region, but instead a source of collective concern and family (particular male) reputation. Increasing reports of honor killing – in Jordan and the Palestinian Territories, for example – are part of a spectrum of gender-based violence across the Arab region coming to light; recent national surveys in Morocco and Tunisia, for example, show that more than a third of young women have experienced violence at some point in their lives, usually at the hands of family members (Haut Commissariat au Plan, 2009; Office National de la Famille et de la Population, 2010). Indeed, families expend enormous attention and effort on the hymen, this precious piece of female anatomy. Among the tools at their disposal is female genital mutilation, thought to 'cool' a woman's sexual drive; widely practiced in countries bordering the Red Sea, and northern Iraq, its capital is Egypt, where more than 90% of ever-married women aged 15–49 have been circumcised (UNICEF, 2013).

Restrictions on girls' physical activity and social mobility are another means of control; virginity testing is also practiced, including age-old customs (such as *dukhla*, in which a bloodied wedding bedsheet is taken as proof-positive of defloration) and modern medical exams, imposed by families and even, in the case of post-Arab Spring Egypt, by the state in order to discipline female protestors. Research shows that young people themselves both uphold, and subvert, this convention by resorting to nonvaginal intercourse, with high rates of anal and oral sex, as well as hymen repair (El Feki, 2014).

Such is the taboo around youth sexuality that it's not just the end, but also the perceived means that are rejected. There is, for example, widespread resistance to sexuality education in schools, despite ample evidence of ignorance among children, as well as their parents, and confusion compounded by the common consumption of Internet porn. Similarly controversial is the provision of sexual and reproductive health services, including contraception and abortion, to unmarried youth.

Laws and government policies are proving slow to change. In spite of all these obstacles, however, women and men across the region are challenging constraints, albeit working along the grain of culture and tradition. In Morocco, for example, civil society is pushing for a gradual liberalization of abortion laws, currently restricted in the vast majority of Arab countries, yet practiced illegally nonetheless, with an estimated 1.7 million unsafe procedures annually in the Middle East and North Africa (World Health Organization, 2011; Evrard, 2014). Morocco is also home to groundbreaking nongovernmental organizations (NGOs) helping unwed mothers to find jobs, reunite with their families and, critically, stay with their children.

On sexuality education, Lebanon is one of a handful of countries in the region working toward a national 'life skills' curriculum (including wide-ranging instruction on sexual and reproductive health), no easy task given teacher embarrassment and other impediments; elsewhere, programs such as Muntada Jensaneya (Arab Forum for Sexuality, Health and Education), among Palestinian communities in Israel, are finding innovative ways to communicate the facts of life in and out of the classroom.

In Tunisia and Egypt, government-sponsored 'youth-friendly' clinics are working to provide reproductive health information and services, with mixed results, in part because of healthcare provider prejudice. More successful is youth-to-youth education, such as the Y-Peer network, under the aegis of the United Nations Population Fund of UNFPA, the United Nations Population Fund (formerly known as the United Nations Fund for Population Activities), with chapters across the region, in which young people provide one another with information on sexual health. And there is also increasing action by civil society groups to tackle gender-based violence, be it pressuring governments to pass legislation (as in Egypt, against sexual violence in public spaces, in Lebanon against domestic abuse and in Morocco on rape), empowering young women on the streets, holding police to account for sexual abuse, or working with young men to understand, and prevent, such violence at its roots.

An explosion in social media since the Arab Spring is also fostering unprecedented challenges to the sexual status quo. Most dramatic, and arguably least successful, have been the FEMEN-inspired postings of nude photos by young Tunisian and Egyptian women. While acting in the name of freedom of expression and bodily autonomy, these incidents have invoked the ire, not only of religious conservatives, but of political liberals as well (Sboui, 2014). More discreet challenges to authority are proving less divisive, among them Morocco's Facebook protest against the arrest of two teenagers for kissing in public, or Lebanon's playful 'Strip for Jacky' campaign against the dismissal of one the country's Olympic athletes on charges of indecency. Recent years have also seen the emergence of new sources of accurate information, and reasoned discussion, in Arabic on sexuality – notably Al Hubb Thaqafa, a pioneering pan-Arab website which harnesses the power of Facebook, Twitter, and YouTube to stimulate informed and respectful debate about sex, not just as a problem to be managed, but as a pleasure to be celebrated – an openness not seen since the golden age of Arabic erotica a millennium ago.

LGBTQ Populations

Also under pressure is the estimated 2–3% of the region's population who have sex with their own sex or have a different gender identity – some (though by no means, all) of whom self-identify as LGBTQ, that is, lesbian, gay, bisexual, transgender/transsexual or queer (Abu-Raddad et al., 2010). Stigma, discrimination, and violence – by the state, communities, and families – is pervasive. While patriarchies pump up the pressure on women, the rise of Islamic fundamentalism in recent decades has also narrowed the line on what it means to be a 'real man.' There is little tolerance for appearance or behavior perceived as 'gay' in most quarters, and the age-old prejudice against the passive, penetrated male lives on.

Contemporary mainstream Sunni and Shi'i thinking condemns both male and female homosexual acts. Despite a history of more open interpretation by religious scholars within the region, and a lively debate in today's wider Islamic world on possibilities of accommodating same–sex relations, very few Arab voices – notably Tunisian and Syrian academics – openly argue for alternative approaches within the region's dominant faith (Kugle, 2010). Nor is this a uniquely Muslim phenomenon; the many Christian Churches, and conservative Jewish sects, of the region take a similarly hardline against those who break the heteronormative mold.

Politics adds fuel to the religious flame. Within the region, homosexuality has become a weapon in political battles in post-Arab Spring: as an accusation, used by both Islamists and their more ostensibly secular counterparts in Egypt, Tunisia, and Syria, to tar opponents; and as yet another tool to prop up political power by resisting any challenge to the sexual or social status quo as a Western-inspired, if not directly organized, attempt to undermine 'traditional' Arab and Islamic values. In recent years, the status of alternative sexual orientations and gender identities has become a hallmark of liberal democracy. As a consequence, Western aid conditionality, while not yet openly applied to the Arab region, is serving to further strengthen the association of LGBTQ rights with a 'foreign' agenda, fusing homo- and xenophobia in a volatile mix. The question of homo- and transsexual rights has even entered the long-running Arab–Israeli conflict, with a fierce debate on comparative freedoms for LGBTQ individuals on either side of the fence, and their bearing on the legitimacy of a future Palestinian state.

This popular and politically expedient association poses a particular challenge for civil society in the Arab region, whose work largely depends on foreign funding. There are now more than dozen NGOs working to empower LGBTQ populations, scattered across Morocco, Algeria, Tunisia, Egypt, Palestine, Northern Iraq, and Lebanon, the epicenter of LGBTQ civil society, with almost half-a-dozen groups offering information, support and other services. While still relatively few in number, and with an outreach still largely limited to educated, middle class men and women, this is a threefold increase in activity across the region since 2010. Such growth is due in large measure to an increasing willingness on the part of young activists to test the boundaries of freedom of expression and assembly, even in the face of rising political Islamism (as in Morocco), a clampdown on civil society (as in Egypt), and violent conflict in much of the region.

Social media is key to this development. Online sites and apps, such as GayRomeo and Grindr, are popular with gay men across the region; Facebook and other Internet fora are particularly valuable for the minority of women with access to the Internet, whose physical and social ambit is far more constrained by family and community control than that of their male counterparts. And it has also transformed collective organization, in sexuality as in politics. The granddaddy of all LGBTQ organizations in the region, Helem, based in Lebanon, grew out of an online discussion group in the mid-2000s. Some of the newest groups exist only online – in the form of Internet magazines, as in Morocco, or online radio, as in Algeria – providing information, support, and opportunities for discussion rarely available in Arabic, even on, socially accepted married sexual life. Other groups, such as Al Qaws and Aswat, supporting queer Palestinian

women, Meem has closed doors, mirror their extensive on-the-ground activities with a strong online presence. While technology opens new doors, it also presents new challenges: in particular, online security, and the urgent need to protect activists, be it from police entrapment or family outrage.

Such Internet Internet activity complements several offline regional initiatives launched since the beginning of this decade. Among them is the Arab Foundation for Freedoms and Equality (AFE), which trains human rights advocates, particularly those with an interest in LGBTQ rights; the Regional Network on HIV/AIDS for men-who-have-sex-with men (M-Coalition); and MantiQitna, an annual retreat for LGBTQ activists from across the Middle East and North Africa, focused on networking, capacity building, and advocacy. These collaborations are part of a broader strategy by LGBTQ activists to address the violence, stigma, and discrimination experienced by their peers by weaving their message into wider efforts secure fundamental human rights for all, rather than an exclusive focus on a sexual identity politics. In Tunisia, for example, several new NGOs, working on recognition and respect for personal freedoms, and minority rights, have incorporated alternative sexualities and gender identities in their purview from the outset.

Law is a major challenge. More than four-fifths of countries in the Arab region have penal codes that explicitly criminalize homosexual acts. Some, like that of Sudan, specify male sodomy; others, like that of Algeria, sanction both male and female same–sex relations; others, as in Lebanon and Syria, condemn sexual relations "against the order of nature" (Al Farchichi and Saghiyeh, 2009). Burdens of proof vary widely, and it is usually poor men on the receiving end of the state's attention. Punishments can range from fines and imprisonment to the death penalty, in the case of Sudan, Saudi Arabia, and Yemen where strict interpretations of shari'a are the law of the land.

Nor are men and women necessarily immune in the few Arab states, among them Jordan and the West Bank, without specific articles outlawing homosexuality. In Egypt, for example, legal provisions on 'habitual debauchery' and prostitution are routinely used to prosecute gay men, most famously in the Queen Boat incident of 2001 and more recently in the arrest of more than 70 homosexual men and male-to-female transgendered individuals, swept up in a politically motivated crackdown in the run-up to the 2014 presidential elections. In Iraq there is as yet no state prosecution of homo- or transsexual men and women, yet they face gruesome reprisals and rough justice in Islamist-dominated strongholds beyond government control.

The law can punish in subtle ways as well. Legal codes and policies across the Arab region that penalize sex work and vaguely defined 'public indecency' or HIV infection among migrant workers, or consensual adult sexual activity outside of marriage, or the provision of sexual and reproductive health information, have a disproportionate effect on women and sexual minorities. And then there is sanction through inaction: for example, bureaucratic foot-dragging and red tape which prevents transsexual men and women from changing their sex on official documents, thereby exposing them to a range of difficulties, including prosecution where the law punishes transgendered appearance or behavior, as in Kuwait. Or, more commonly, a failure on the part of authorities to protect those who cross the heterosexual line from verbal or physical abuse, blackmail, job dismissal, and other human rights violations, especially when the perpetrators are within the family or tribe, bedrock of Arab societies.

In the current political climate, changing the law on the books is difficult – all the more so in countries with new constitutions, such as Egypt, Libya, and Iraq, which invoke Islam as a cornerstone. Nonetheless, the experience of NGOs in Lebanon shows that it is possible to mitigate the law in practice. For example, ongoing, informal interaction between Helem and police and lawyers, along with raising awareness of LGBT individuals of their rights, is relieving some of the pressure. And in the courts, strategic litigation is turning the law on its head: over the past 5 years, two landmark cases, involving a gay and transgendered man, have seen judges in Lebanon rule that Article 534, with its condemnation of sexual acts "contrary to nature" does not apply to such individuals who fall within the spectrum of nature's rich variety (Jernow, 2011).

Health is another problem as well as a point of progress. Men-who-have-sex-with-men are among the hardest hit by the region's emerging HIV epidemic, with infection rates exceeding 10% in a number of countries. Treatment rates are also the lowest in the world, with fewer than 20% of those in need of anti-retroviral therapy receiving the medication they require (UNAIDS, 2014) As with youth populations, HIV is paving the way to quiet government approval for research, funding, and outreach to vulnerable groups, including male and female sex workers. Public health also allows religious leaders, who are key to shifting mainstream attitudes, to constructively engage on questions of homo- and transsexuality, although translating these conversations into positive action from the pulpit, or the minbar, is an ongoing challenge.

Ironically, the biggest single health challenge facing LGBTQ men and women are health-care providers themselves. Homophobia is rife: one study in Lebanon, for example, found that more than half of physicians surveyed refused to treat homosexual men, preferring to send them onto psychiatrists for reparative therapy, also a popular practice in Egypt and the Gulf (Al-Kak, 2009). Again, NGOs are slowly changing the landscape. Lebanon is home to Marsa, the Arab region's first LGBTQ-friendly sexual health clinic, and similar operations are taking root in other countries, among them Morocco. Outreach to doctors is also bearing fruit. Last year, for example, the Lebanese Psychiatric Society openly condemned reparative therapy as malpractice, putting in place measures to discipline its practitioners. And Lebanese physicians themselves have also banded together in a new NGO to combat stigma and discrimination against LGBTQ patients.

Media is also key to shaping views. Mainstream press, movies, talk shows, and soap operas in the Arab region generally take a dim view of homo- and transsexuality. Such attitudes are reflected in language, *shadh* or 'deviant' being the most common term used to describe men-who-have-sex-with-men. One of the notable successes of LGBTQ activists in Lebanon has been to shift media to more neutral terms, such as *mithliyya* for homosexuality. Lebanon is now home to one of the few openly gay performers in the Arab region and its authors and filmmakers are part of small but growing trend – noticeable also in North Africa – toward more sympathetic and nuanced portrayals of alternative sexualities and gender identities.

In Egypt, for example, most press were quick to curry political favor and align with government officials in condoning the 2014 arrests of homo- and transgendered men, who were left to languish in prison in the absence of popular pressure for their release. In contrast, thanks to civil society, a 2013 TV program featuring homosexual men caught in flagrante at a gay cruising cinema in Beirut had different outcome. Following the broadcast, police raided the venue, arresting 30 mainly poor, migrant workers, who were then subjected to anal examinations – a scientifically and ethically dubious forensic technique to 'prove' homosexual acts, popular across the Arab region. However, other Lebanese media, with which Helem had worked on such issues, were quick to denounce the tests. As a result, not only were the men released, but the country's national medical association also condemned the practice, threatening to strike off any doctor implicated; the Ministry of Justice subsequently issued a circular instructing police to cease and desist from such procedures, a welcome step towards eliminating the practice.

Conclusion

Among the more remarkable pieces of graffiti in Cairo following the 2011 uprisings was a cartoon of two mustachioed men, in policemen's caps, kissing each other, under the slogan, 'Cops are Gays.' Given recent clashes between protesters and security forces, this was intended as an insult, not a sociological observation. Within days, however, unknown artists had transformed the moustaches into rainbows and replaced the hate-filled text with a more constructive insight: 'Homophobia is not revolutionary.'

If only it were as easy to change attitudes on the ground. The Arab world is by no means alone in its sexual dilemmas, as recent furores over homosexuality in Uganda, sexual violence in India, or abortion in Spain clearly attest. But some features stand out. Across the region, sex is bound up in shame – particularly for women – which makes it a powerful tool of political and social control. Sexual freedoms are hard to exercise when family interests trump individual choice. And when appearance counts for more than reality – when virginity is defined by a piece of anatomy rather than a state of chastity, for example, or when prostitution masquerades as marriage.

While there have been small gains in sexual rights since the Arab Spring, there are limits to growth, even in as relatively open an environment as Beirut, when sexuality remains a taboo; when women's rights are curtailed, in law as well as practice; when youth unemployment limits autonomy; and when fundamental political, economic and social rights are still far from reality for all but an elite across the Arab region. Bringing sexuality into the broader debate on individual rights and personal freedoms is key to ushering in change in both political and personal life. But that is the work of a generation at least.

Acknowledgment

This article draws on the author's book, *Sex and the Citadel: Intimate Life in a Changing Arab World*.

See also: Afrocentric Approaches to Social Work; Ancient Egypt, Archaeology of; Female Genital Mutilation; Gay and Lesbian Movements; Global Violence Against Women; Honor Killing: Multidimensional and Multilevel Perspectives; Honor and Shame; Islam: Middle East; Islamic Fundamentalism; Islamic Law; Middle East, North Africa: Sociocultural Overview; Middle Eastern and North African Legal Traditions; Near Middle East/North Africa Studies: Culture; Near Middle East/North Africa Studies: Gender; Near Middle East/North Africa Studies: Politics; Near Middle East/North Africa Studies: Religion; Near Middle East/North Africa Studies: Society and History; Same-Sex Sexualities in Turkey; Sexual Minorities and Sexual Citizenship; Sexual Minorities in the Maghreb; Sexuality: Cultural Aspects; Transgender, Transexual, and Gender Variant Individuals.

Bibliography

Abu-Raddad, L., Akala, F.A., Semini, I., Riedner, G., Wilson, D., Tawil, O., 2010. Characterizing the HIV/AIDS Epidemic in the Middle East and North Africa: Time for Strategic Action. World Bank, Washington, DC.

Al Farchichi, W., Saghiyeh, N., 2009. Homosexual Relations in the Penal Codes: General Study Regarding the Laws in Arab Countries with a Report on Lebanon and Tunisia. Helem, Beirut.

Al-Kak, Faysal, 2009. Homophobia in Clinical Services in Lebanon: A Physician Survey. Helem, Beirut.

Dhillon, N., Yousef, T., 2009. Generation in Waiting: The Unfulfilled Promise of Young People in the Middle East. Brookings Institution Press, Washington, DC.

Daniel, N., 1993. Islam and the West: The Making of an Image. Oneworld, Oxford.

Evrard, A., 2014. The Moroccan Women's Rights Movement. Syracuse University Press, Syracuse.

El Feki, S., 2014. Sex and the Citadel: Intimate Life in a Changing Arab World. Vintage, London.

Hasso, Frances S., 2010. Consuming Desires: Family Crisis and the State in the Middle East. Stanford University Press, Stanford, CA.

Haut Commissariat au Plan, Royaume du Maroc, 2009. Enquête nationale sur la prévalence de la violence a l'égard des femmes au Maroc. Haut Commissariat au Plan, Rabat.

International Planned Parenthood Federation, 2008. Sexual Rights: An IPPF Declaration. International Planned Parenthood Federation, London.

Jernow, Ali, 2011. Sexual Orientation, Gender Identity and Justice: A Comparative Law Casebook. International Commission of Jurists, Geneva.

Kugle, S., 2010. Homosexuality in Islam. Oneworld, Oxford.

Norris, P., Inglehart, R., 2002. Islam and the West: Testing the 'Clash of Civilizations' Thesis. John F. Kennedy School of Government at Harvard University, Cambridge, MA.

Office National de la Famille et de la Population, 2010. Enquête nationale sur la prévalence de la violence a l'égard des femmes en Tunisie. Office National de la Famille et de la Population, Tunis.

Said, E., 1995. Orientalism: Western Conceptions of the Orient. Penguin, London.

Saunders, D., 2013. The Myth of the Muslim Tide. Vintage, London.

Sboui, Amina, 2014. Mon Corps M'appartient. Dectire, Paris.

UNAIDS, 2014. The Gap Report. UNAIDS, Geneva.

UNFPA, 2012. Marrying Too Young: End Child Marriage. UNFPA, New York.

UNICEF, 2013. Female Genital Mutilation/Cutting: A Statistical Overview and Exploration of the Dynamics of Change. UNICEF, New York.

World Health Organization, 2006. Defining Sexual Health: Report of a Technical Consultation on Sexual Health, January 28–31, 2002, Geneva. World Health Organization, Geneva.

World Health Organization, 2011. Unsafe Abortion: Global and Regional Estimates of the Incidence of Unsafe Abortion and Associated Mortality in 2008. World Health Organization, Geneva.

Relevant Websites

www.ohchr.org/en/hrbodies/cedaw – Committee on the Elimination of Discrimination against Women.

www.lmarabic.com

www.worldvaluessurvey.org – World Values Survey.

Sexual Risk Behaviors

Leickness C Simbayi, Human Sciences Research Council, Cape Town, South Africa; and Psychiatry and Mental Health, The University of Cape Town, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sexual risk behaviors place either of the sexual partners involved at increased risk of some negative health outcomes such as increasing the chance of acquiring or transmitting disease especially sexually transmitted infections including human immunodeficiency virus or having an unwanted pregnancy. They include early sexual debut, unsafe sexual practices such as having unprotected (i.e., without a condom) vaginal, anal, and oral sex, having multiple sexual partners, engaging in group sex, engaging in dry sex or vaginal douching, and having sex while bleeding. There are also some culturally sanctioned sexual risk behaviors. Safer sex practices such as abstinence, sexual fidelity, and condom use offer solutions to avoid the negative health outcomes mentioned.

Introduction

Sexual risk behaviors encompass all those practices associated with sexual intercourse irrespective of sexual preferences that place either of the sexual partners involved at increased risk of some negative health outcome such as increasing the chance of acquiring or transmitting disease especially sexually transmitted infections (STIs) including human immunodeficiency virus (HIV) or an unwanted pregnancy. Common sexual risk behaviors include early sexual debut, unsafe sexual practices such as having unprotected (i.e., without a condom) vaginal and anal sex, unprotected oral sex, having multiple sexual partners, and engaging in group sex, dry sex or vaginal douching, and having sex while bleeding (Leclerc-Madlala et al., 2009). In addition, using unreliable methods of birth control, or using birth control inconsistently also constitutes sexual risk behavior as it increases the risk of occurrence of unwanted pregnancies. These behaviors facilitate STI/HIV transmission through mixing of blood, sperm, and vaginal fluids between or among sexual partners whereas pregnancies result from the fertilization of eggs in females by sperm from males after engaging in sexual intercourse (UNAIDS, 2013). STIs themselves have serious consequences such as (1) increasing the risk of HIV acquisition threefold or more; (2) stillbirth, neonatal death, low birthweight and prematurity, sepsis, pneumonia, neonatal conjunctivitis, and congenital deformities; (3) cervical cancer; and (4) pelvic inflammatory disease, adverse pregnancy outcomes, and infertility (WHO, 2013), while unwanted pregnancies are limited to numerous maternal and child health problems (WHO, 2005). This article focuses primarily on sexual risk behaviors associated with STIs including HIV, and leaves out those linked to unwanted pregnancies as the two sets of behaviors are largely identical in nature. It concludes by discussing safer sex practices such as abstinence, fidelity, and condom use that are known to reduce acquiring or transmitting disease especially STIs including HIV as well as unwanted pregnancies.

Early Sexual Debut

Initiating sex at an early age increases lifetime risk to HIV (UNAIDS, 2013) and has also been linked to other HIV risk factors including alcohol use, having unprotected sex, unplanned pregnancy, and having multiple sexual partners (UNAIDS, 2013). Possible explanations for young people to engage in early sex include peer pressure, low self-esteem, pessimism, sexual coercion, and transactional sex (see Shisana et al., 2009). Others include exposure to television programs with plots around love, sex, and money (Ankomah et al., 2011). Infection occurs primarily due to tearing and bleeding that occur during sex because of the immaturity of adolescent females' reproductive organs. Furthermore, early sexual debut implies that the potential period of exposure to STIs/HIV is lengthened because it is linked to more frequent sexual intercourse, more lifetime STIs, and more sexual partners (Donenberg et al., 2003). Typically young people who have sex at an early age know very little about their own sexual and reproductive health let alone how to protect themselves from acquiring STIs and HIV as they tend to have unprotected sex. Early sexual debut is also linked to less consistent contraceptive use (Geary Waszak et al., 2008) and unplanned pregnancies (Baumgartner et al., 2009).

Unprotected Vaginal Sex

One of the sexual risk behaviors that people of all ages engage in all over the world is having unprotected sex (UNAIDS, 2013). That is, they either do not use male/female condoms correctly and consistently every time they have sexual intercourse as they feel they can trust their sexual partners. If either of the sexual partners is infected with either a classical STI such as trichomoniasis, chlamydia, gonorrhea, human papillomavirus (genital warts), syphilis, genital herpes, or HIV the risk for the uninfected partner to also become infected is high. Among the main reasons for the poor use of condoms that is observed is the prevailing negative attitude toward condom use among both partners. Furthermore, and perhaps more important is the poor access to condoms, especially female ones, which are generally more expensive than male ones (Guerra and Simbayi, 2014). In addition, condom use with primary partners – either spouses or steady partners – tends to be lower than condom use with nonregular partners as well as being lower among older persons (Peltzer et al., 2012; Shisana et al., 2009).

Unprotected Anal Sex

Unprotected anal sex regardless of sexual preferences is also considered sexual risk behavior (UNAIDS, 2013). While it is a common sexual practice among men who have sex with men, it is not as widely practiced among heterosexual couples although it is believed to occur worldwide as part of normal sexual behavior (e.g., see Halperin, 1999). There is a need to highlight the risk for STI/HIV associated with this sexual behavior, if practiced unprotected, i.e., without condoms. More disturbing is that female adolescents avoid vaginal sex to retain their virginity but unfortunately engage in heterosexual anal sex instead under the misguided belief that anal sex is not as risky as vaginal sex (Leclerc-Madlala, 2001, 2002).

Although usually denied, anal sex between men in prison takes place in most countries especially where no conjugal visits are permitted (UNAIDS, 2013). Thus, unprotected sexual intercourse among incarcerated men, whether coercive or consenting, puts those involved at risk for STI/HIV infection. In accounts of prison sex, where sexual favors are used in exchange for cigarettes, drugs, or protection from violence and death, incarcerated adolescents more frequently report previous sexual intercourse, intercourse with multiple sexual partners, rape, gang rape, and inconsistent condom use. As a result of their higher lifetime numbers of sexual partners and their failure to use condoms, youth in detention have higher rates of STIs. Upon release from prison, the victim is likely to transcend his aggressive sexual attitude to raping other members of his community in an attempt to regain his manhood. Given the unwillingness of many authorities to acknowledge that unprotected anal sex is fairly common among men held in prisons, and to make condoms available for use in such institutions, prisons continue to be major breeding grounds for new STIs/HIV infections.

Unprotected Oral Sex

Oral sex involves giving or receiving oral stimulation (i.e., sucking or licking) to the penis (fellatio), the vagina (cunnilingus), or the anus (anilingus) (CDC, undated). People can also acquire or transit STIs/HIV via oral sex although it is a much lower risk activity than sexual intercourse involving vaginal or anal sex, especially for the recipient. This is only true, of course, if the person giving the oral sex has no sores in their mouths or the person receiving it either bleeds or has sores on their sex organs. Similar precautions such as use of condoms or dental dams can be taken during oral sex to prevent STI/HIV infection via this route.

Intergenerational Sex

One very common practice found in many countries in the world involves young girls having sex with men who are their seniors by more than 10 years (UNAIDS, 2013). This is also known as age mixing. Intergenerational sex relationships are mainly driven by poverty where young girls are attracted by material rewards from the older partner (Leclerc-Madlala, 2008; Ott et al., 2011). The promise of gifts and other financial and social benefits, including the potential for upward social mobility, can affect early sexual debut and motivate girls to engage in age-disparate sexual relationships, transactional sex, or marriage at a young age (UNICEF, 2013). This practice of early or arranged marriages is common among various cultures throughout the world as well as in some religions. Where HIV prevalence is high, these factors combine with increased biological vulnerability, limited knowledge about HIV, and low-risk perception to enhance the risks and limit the choices girls and their families make. Given the fact that it currently takes approx. 10–12 years between initial infection and the manifestation of AIDS, most young women in Southern Africa can be assumed to have been infected with HIV during their early teens. Secondly, young girls are often coerced by older men including male schoolteachers into having their first sexual experience with them thus putting them at risk for STI/HIV infection (Shell and Zeitlin, 2000).

Coital Bleeding

There are some reports of people in South Africa engaging in sex when there is blood. The bleeding occurs for different reasons ranging from menstruation by the female, physical injury to the vagina or vaginal walls during intercourse, or STIs such as genital ulcers. Clearly such behavior puts those involved at risk for STIs/HIV infection (Kalichman et al., 2006; O'Farrell and Hoosen, 2007). This probably also occurs worldwide albeit at times inadvertently. There is therefore a need to either promote condom use or to refrain from sex when there is such genital bleeding.

Multiple and Concurrent Sexual Partnerships

Having multiple sexual partners, whether concurrently or through serial monogamy with a high turnover of sexual partners increases the likelihood of exposure to STIs/HIV infection through expanding sexual networks. This is because the likelihood of HIV transmission in sexual networks increases as a result of high viral loads that occur in the early phases of new HIV infection. Having concurrent sexual partners is thought to exacerbate this risk (Halperin and Epstein, 2007; Mah and Halperin, 2010). However, it has recently been argued that concurrency does not fully explain the high HIV epidemic in Sub-Saharan Africa (e.g., see Lurie and Rosenthal, 2009; Sawers and Stillwagon, 2010). In particular, Sawers (2013) has argued that concurrent partners do not raise individuals' risk of acquiring HIV any more than having a nonconcurrent partner. This is because the contributory effects of concurrency appear to be diminished in high HIV prevalence settings where cultural heritage of polygyny also still tends to legitimize sex through informal concurrent partnering among males (see Leclerc-Madlala et al., 2009). Nevertheless, it remains crucial to reduce exposure to multiple sexual partnerships, irrespective of concurrency (Tanser et al., 2011).

Although not substantiated, group sex involving a person who is infected by an STI or HIV should also have a similar effect (Sam Friedman, self-communication 2012). Just as in the other instances described, it is imperative that people engaged in such behavior also use condoms correctly and consistently whenever they engage in sexual intercourse.

Sex Work

Sex work or prostitution involving both sexes occurs in many countries throughout the world. In some countries where sex work is illegal, it occurs underground. Sex workers have few or no rights and are taken advantage of by those who pay for their services. Even in countries with high HIV prevalence, those who pay for their services will often insist on paying more to have unprotected sex than use condoms to protect themselves from STI/HIV infections. Consequently, female sex workers also end up becoming pregnant and are in turn forced to terminate their unwanted pregnancies. One of the ways to reduce the risk on unprotected sex involved in sex work is through legalizing sex work and ensuring that sex workers are provided with health education including the need for correct and consistent use of condoms (UNAIDS, 2013).

Dry Sex or Vaginal Douching

Dry sex or vaginal douching involves cleaning the vagina with either a variety of substances including antiseptics and detergents or inserting traditional herbs into the vagina and drying it with paper or cotton prior to or after having sexual intercourse (Louria et al., 2000; Morar et al., 2003). Some drying and tightening agents are used as 'love potions' to retain the affections of a partner. Dry sex is meant to increase the pleasure that men derive during sexual intercourse. This is because vaginal wetness is seen as indicative of female infidelity, uncleanliness, possible infection, and looseness on the part of the woman rather than natural preparedness to have sex (see Leclerc-Madlala et al., 2009). The main problem with this practice lies with the fact that douching disturbs the vaginal flora or causes extensive local irritation and inflammation of vaginal walls thereby enhancing the sexual transmission of HIV infection (see Leclerc-Madlala et al., 2009). Unfortunately such practices still exist even today thus putting women at risk of STI/HIV infection.

Substance Abuse

There is growing evidence of the strong link between substance abuse especially alcohol and HIV/STI infection (Kalichman et al., 2007; Parry et al., 2010, 2009). Indeed, unwanted pregnancies have also been attributed to substance use. According to this hypothesis, alcohol intake increases sexual risk-taking behavior. Accordingly, the argument has been that when alcohol, and any other illicit drug is consumed in excessive amounts, it inhibit people's ability to engage in safer sex practices such as using condoms correctly and consistently. Thus, persons under the influence of alcohol or other illicit drugs are unable to protect themselves from being infected by HIV/STIs when having sexual intercourse with an infected person. With both alcohol and drug use on the increase globally, the HIV/STIs via this route are also bound to increase. In particular, the use of 'hard' drugs such as cocaine, amphetamines, heroin, mandrax (i.e., Methaqualone), and cannabis is most problematic, particularly among youths (see Peltzer et al., 2009). There is therefore a need to emphasize the links between alcohol or illicit drug consumption and unsafe sex among young adults in order to prevent the spread of HIV/STI infection as well as preventing unwanted pregnancies.

Cultural Rites of Death

There are two main customs that are widely practiced by most indigenous cultures especially in Sub-Saharan Africa. The first is known as *levirate*. This refers to a situation in which the brother of a deceased man is obliged to marry his brother's widow, and the widow is obliged to marry her deceased husband's brother (e.g., for a brief discussion about this practice in Swaziland, see Nxumalo, 1999). The second is known as *sororate*. As noted earlier, a widower or sometimes a husband of a barren woman marries his wife's sister in order to ensure that the relationship developed by the two families produces some offspring and hence does not dissolve (also see Nxumalo, 1999). In the case where one partner in the new relationship is infected with HIV, it increases the risk of the new partner to become infected. This is especially so when the departed spouse died because of an AIDS-related illness.

Indigenous Healing Practices

In most countries in Sub-Saharan Africa and indeed throughout the world, modern (i.e., western) and traditional (i.e., indigenous) healing systems co-exist side by side (see Hopa et al., 1998). Even though the western system is dominant in terms of both official status and scientific acceptability, indigenous healing practices are more widely accessible and used due to pervasive traditional health beliefs, which are followed particularly by people who live in rural areas with poor western medical services. Consequently, an overwhelming majority of the populations of most if not all of the countries in Sub-Saharan Africa and also in many other places throughout the world consult with traditional healers for treatments of most ailments. More interestingly, many people have found a way to utilize what both systems have to offer to them. One of the indigenous healing practices that is thought to be partly responsible for the spread of HIV infection involves the healers either having sex with many of their clients or recommending that their clients should have sex with virgins as part of some of their treatments (Dr M. Guma, personal communication, December 2013). Such practices enhance the chance of STI/HIV infection.

Safer Sex Practices

Three main safer sex practices that are recommended to reduce the risk of STI/HIV infection and falling pregnant are (1) sexual abstinence, (2) avoiding multiple sexual partnerships, and (3) using condoms, both male and female types, correctly and consistently every time one has sex. However, this is easier said than done especially among those who have reached the age of consent within their specific cultures, as they are considered mature enough to have sex and the drive to have sexual intercourse is primordial. Many people find it very difficult to practice safe sex and consequently unsafe sexual practices occur quite frequently.

The only way to completely eliminate all these risks is to abstain entirely from sexual activity. Nevertheless, sexual abstinence is difficult especially for those who have attained the age of consent in any culture. Instead, delaying sexual debut remains an important focus for STI/HIV prevention, especially for adolescents. Some interventions that promote primary or secondary abstinence and delayed sexual debut can reduce an individual's risk of being exposed to STIs/HIV although their effectiveness remains doubtful (see Kantor et al., 2008; UNAIDS, 2013). Research has found that reasons for delaying early sex among young females tend to include religious injunction against premarital sex, prevention of HIV infection, fear of pregnancy, and linked to this, the fear of dropping out of school (Ankomah et al., 2011). These factors should be considered when developing STI/HIV prevention programs.

Reducing partner turnover and overall numbers of sexual partners is a key component of an effective STI/HIV prevention response (UNAIDS, 2013). The basic message is the promotion of fidelity to one sexual partner whose STI/HIV status is known. Programs are needed that discourage multiple sexual partnerships, or – where this is not possible – to emphasize the consistent use of condoms. In particular, it is critical to address the social norms, which sanction the involvement of males in multiple sexual partnerships. This issue needs to be taken up by whole communities and should not be left to individuals to be dealt with.

Condoms, both male and female, are currently the only available and most effective technology to prevent HIV and other STIs, as well as unintended pregnancies, among sexually active people (UNAIDS, 2013). Increasing the availability of female condoms (Guerra and Simbayi, 2014; UNAIDS, 2013) and promoting consistent condom use remain important focal areas for change. Male condoms are inexpensive, cost effective, their use does not require assistance of medical or healthcare personnel, and they can be utilized by anyone who is sexually active. Correct and consistent use of both male and female condoms is an integral component of combination HIV prevention strategies that individuals can choose at different times in their lives to reduce risks of sexual exposure to HIV and other STIs, or as a dual protective method.

In terms of sex work, ways of making sex safe for both sex workers and their clients is critical. This includes both improving the way they are treated by the law and improving their access to health care services, including the provision of, or access to, male and female condoms as well as lubricants (UNAIDS, 2013). Similarly, ways need to be found to encourage people to drink alcohol responsibly, especially if they intend to have sex. Recent work in South Africa has shown that multilevel interventions that target both men who drink in informal drinking places (or 'shebeens') and the surrounding communities may help reduce risks for STI/HIV infections (see Kalichman et al., in press).

Finally, as cultures continue to evolve it is imperative that cultural practices such as polygyny, dry sex, early marriage, cultural rites of death, and traditional healing practices involving sex with a client are reconsidered among those cultures that practice them in order to reduce the risk of STI/HIV infection and also unwanted pregnancies. This will require working closely with traditional leaders who are the custodians of culture in many societies the world over.

Conclusions

In this article we presented several sexual risk behaviors, which facilitate STI/HIV transmission through mixing of blood, sperm, and vaginal fluids between or among sexual partners. Some of these sexual practices also carry the risk of unwanted pregnancies. The article then discussed safer sex practices to prevent both STI/HIV infection and unwanted pregnancies. Recommendations were also made on how to also reduce the impact of sex work, alcohol, and culturally sanctioned sexual practices. It is hoped that the article has shed some light on the important issue of sexual risk behaviors and their health and social implications for both social and behavioral scientists and the public at large.

See also: Afrocentric Approaches to Social Work; Alcohol Use among Young People; Anthropological Approaches to the Study of Education: The United States and Beyond; Culture as a Determinant of Mental Health; Drug Use and Abuse: Psychosocial Aspects; Globalization and Health; Gynecological Health: Psychosocial Aspects; HIV Risk Interventions; Healing Practices: Alternative and Complementary; Health Behaviors; Health Education and Health Promotion; Health Psychology; Health Risk Perception; Health and Illness: Mental Representations in Different Cultures; Infectious Diseases: Psychosocial Aspects; Public Health as Social Science; Sexuality Education; Sexuality: Cultural Aspects; Social Epidemiology.

Bibliography

Ankomah, A., Mamman-Daura, F., Omoregie, G., Anyanti, J., 2011. Reasons for delaying or engaging in early sexual initiation among adolescents in Nigeria HIV prevalence. Adolescent Health, Medicine and Therapeutics 2, 75–84.

Baumgartner, J.L., Geary Waszak, C., Tucker, H., Wedderburn, M., 2009. The influence of early sexual debut and sexual violence on adolescent pregnancy: a matched case-control study in Jamaica. International Perspectives on Sexual and Reproductive Health 35 (1), 21–28.

CDC. Oral Sex and HIV Risk. http://www.cdc.gov/hiv/risk/behavior/oralsex.html (accessed 29.12.13.), undated.

Donenberg, G.R., Brynat, F.B., Emerson, M.A., Wilson, B.A., Pasch, B.A., 2003. Tracing the roots of early sexual debut among adolescents in psychiatric care. Journal of the American Academy of Child and Adolescent Psychiatry 42 (5), 594–608.

Geary Waszak, C., Baumgartner, J.L., Tucker, H., Johnson, L., Wedderburn, M., Wagman, J., 2008. Early Sexual Debut, Sexual Violence, and Sexual Risk-taking Among Pregnant Adolescents and Their Peers in Jamaica and Uganda. FHI Working Paper Series, USA.

Guerra, F.M., Simbayi, L.C., 2004. Prevalence of knowledge and use of the female condom in South Africa. AIDS & Behaviour 18 (1), 146–158.

Halperin, D.T., 1999. Heterosexual anal intercourse: prevalence, cultural factors, and HIV infection and other health risks. Part I. AIDS Patient Care 13 (12), 717–730.

Halperin, D.T., Epstein, H., 2007. Why is HIV prevalence so severe in Southern Africa? The role of multiple concurrent partnerships and lack of male circumcision: Implications for AIDS prevention. The Southern African Journal of HIV Medicine 8 (1), 19–25.

Hopa, M., Simbayi, L.C., Du Toit, C., 1998. Perceptions on integration on traditional and western healing in the new South Africa. South African Journal of Psychology 28 (1), 8–14.

Kalichman, S.C., Simbayi, L.C., Cain, D., Cherry, C., Jooste, S., 2006. Coital bleeding and HIV risks among men and women in Cape Town, South Africa. Sexually Transmitted Infections 33 (9), 551–557.

Kalichman, S.C., Simbayi, L.C., Kaufman, M., Cain, D., Jooste, S., 2007. Alcohol use and sexual risks for HIV/AIDS in Sub-Saharan Africa: systematic review of empirical findings. Prevention Science 8 (2), 141–151.

Kalichman, S.C., Simbayi, L.C., Cain, D., Carey, K.B., Carey, M.P., Eaton, L., Harel, O., Mehlomakhulu, V., Mwaba K., Randomized community-level HIV prevention intervention trial for men who drink in South African alcohol-serving venues. European Journal of Public Health. [Epub ahead of print 18 November 2013], in press.

Kantor, L.M., Santelli, J.S., Teitler, J., Balmer, R., 2008. Abstinence-only policies and programs: an overview. Sexuality Research & Social Policy 5 (3), 6–17.

Leclerc-Madlala, S., 2001. Virginity testing: managing sexuality in a maturing HIV/AIDS epidemic. Medical Anthropology Quarterly 15 (4), 533–552.

Leclerc-Madlala, S., 2002. On the virgin cleansing myth: gendered bodies, AIDS and ethnomedicine. African Journal of AIDS Research 1, 87–95.

Leclerc-Madlala, S., 2008. Age-disparate and intergenerational sex in Southern Africa: the dynamics of hypervulnerability. AIDS 22 (Suppl. 4), 1–9.

Leclerc-Madlala, S., Simbayi, L.C., Cloete, A., 2009. The sociocultural aspects of HIV/AIDS in South Africa. Chapter 2. In: Rohleder, et al. (Eds.), HIV/AIDS in South Africa 25 Years On: Psychosocial Perspectives. Springer, New York, pp. 13–25.

Levin, R.J., 2005. Science update – wet and dry sex – the impact of cultural influence in modifying vaginal function. Sexual and Relationship Therapy 20 (4), 465–474.

Louria, D.B., Skurnick, J.H., Palumbo, P., Bogdon, J.D., Rohowsky–Kochan, C., Denny, T.N., Kennedy, C.A., 2000. HIV heterosexual transmission: a hypothesis about an additional potential determinant. International Journal of Infectious Diseases 4 (2), 110–116.

Lurie, M.N., Rosenthal, S., 2009. Concurrent partnerships as a driver of the HIV epidemic in sub-Saharan Africa? The evidence is limited. AIDS & Behaviour 14 (1), 17–24.

Mah, T.L., Halperin, D.T., 2010. Concurrent sexual partnerships and the HIV epidemic in Africa: evidence to move forward. AIDS & Behavior 14 (1), 11–16.

Morar, N.S., Ramjee, G., Gouws, E., Wilkinson, D., 2003. Vaginal douching and vaginal substance use among sex workers in Kwazulu-Natal, South Africa. South African Journal of Science 99, 371–374.

Nxumalo, M., 1999. Women's health, sociocultural practices, and HIV/AIDS in Swaziland. In: Hope Sr., K.R. (Ed.), AIDS and Development in Africa. Haworth Press, New York, pp. 59–68.

O'Farrell, N., Hoosen, A.A., 2007. Coital bleeding and HIV risks in South Africa [letter]. Sexually Transmitted Diseases 34 (2), 122.

Ott, M.Q., Barnighausen, T., Tanser, F., Lurie, M.N., Newell, M.L., 2011. Age-gaps in sexual partnerships: seeing beyond 'sugar daddies'. AIDS 25 (6), 861–863.

Parry, C., Rehm, J., Poznyak, V., Room, R., 2009. Alcohol and infectious diseases: an overlooked causal linkage? Addiction 104, 331–332.

Parry, C.D.H., Rehm, J., Morojele, N.K., 2010. As there a causal relationship between alcohol and HIV? Implications for policy, practice and future research. African Journal of Drug & Alcohol Studies 9 (2), 81–91.

Peltzer, K., Simbayi, L., Kalichman, S., Jooste, S., Cloete, A., Mbelle, N., 2009. Drug use and HIV risk behaviour in three urban South African communities. Journal of Social Sciences 18 (2), 143–149.

Peltzer K., Swartz S., Naidoo P., Ramlagan S., Mlambo M., Sanger N. et al., 2012. Evaluating the value that young people attach to lovelife and investigating factors impacting the life choices and risk tolerance of youth, unpublished research report.

Sawers, L., 2013. Measuring and modeling concurrency. Journal of the International AIDS Society 16, 17431.

Sawers, L., Stillwagon, E., 2010. Concurrent sexual partnerships do not explain the HIV epidemics in Africa: a systematic review of the evidence. Journal of the International AIDS Society 14 (34). http://dx.doi.org/10.1186/1758-2652-13-34.

Shell, R.C.H., Zeitlin, R., 2000. Positive outcomes: the chances of acquiring HIV/AIDS during school going years in the Eastern Cape, 1990–2000. Social Work Practitioner – Researcher 12 (3), 139–154.

Shisana, O., Rehle, T., Simbayi, L.C., Zuma, K., Jooste, S., Pillay-van Wyk, V., Mbelle, N., Van Zyl, J., Parker, W., Zungu, N.P., Pezi, S., the SABSSM III Implementation Team, 2009. South African National HIV Prevalence. Incidence and Communication Survey 2008: A Turning Tide Among Teenagers? HSRC Press, Cape Town.

Tanser, F., Bärnighausen, T., Hund, L., Garnett, G.P., McGrath, N., Newell, M.L., 2011. Effect of concurrent sexual partnerships on rate of new HIV infections in a high-prevalence, rural South African population: a cohort study. Lancet 378 (9787), 247–255.

UNAIDS, 2013. UNAIDS Report on the Global AIDS Epidemic 2013. UNAIDS, Geneva.

United Nations Children's Fund (UNICEF), 2013. Towards an AIDS-free Generation – Children and AIDS: Sixth Stocktaking Report. UNICEF, New York.

WHO, 2005. The World Health Report 2005–Make Every Mother and Child Count. Chapter 3: Planning Pregnancies before They Even Happen. WHO, Geneva. http://www.who.int/whr/2005/chapter3/en/index3.html (accessed on 29.12.13.).

WHO, 2013. Sexually Transmitted Infections (STIs). Fact sheet N°110. http://www.who.int/mediacentre/factsheets/fs110/en/ (accessed on 29.12.13.).

Sexuality and Social Work

Mark Henrickson, Massey University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sexuality has been described as an umbrella term that relates to a private dimension of people's lives, but a number of social work authors propose that sexuality must also be understood in the cultural and sociopolitical contexts of cisgendered heteronormativity. Social work's core values include social justice, human rights, and the dignity and worth of the individual, and are well positioned to address these issues in practice, education, and research environments. This article introduces a taxonomy of sexuality and gender, briefly explores sociopolitical contexts that pose critical challenges to sexual and gender minorities, and considers social work responses to these challenges.

Introduction

Sexuality has been described as an umbrella term that "relates to the private dimension in which people live out their sexual, intimate and/or emotional desires" (Dunk-West and Hafford-Letchfield, 2011: p. 2). The place of sexuality within social work is contested. A number of social work authors propose that this private dimension must be understood in cultural and sociopolitical contexts as well as at the individual level. The interplay of these dimensions is reflected in the way sexuality has been treated in the social work literature.

In the social work literature, sexuality has usually been understood as having to do with gender and sexual minorities, people who identify (or who are labeled) with terms like gay, lesbian, bisexual, or transgender. Some social work authors have portrayed these sexualities as cultural variants, where the social work student and practitioner learn something about sexual and gender minorities and their needs in order to be more culturally competent in their practice (Mallon, 2008). More recently, this liberal notion has been critiqued by those who claim that such cultural models of social work education and practice are rooted in heterosexual understandings of the world, and do not take into account the fundamental issue of heteronormativity (Hicks, 2008), that is, the assumption that heterosexuality – opposite-sex attraction and desire, and the legal and social rights, institutions, and privileges that go with it, such as marriage and child rearing – is natural, inevitable, or desirable (Montgomery and Stewart, 2012: p. 164). Cultural variant approaches have also been criticized for problematizing sexual and gender minorities. Some social work authors have taken a pragmatic critical approach that comprises both of these perspectives (Cosis Brown and Cocker, 2011).

Although sexuality has been undertheorized in the social work literature (Fish, 2008), social work, as a discipline whose core values include social justice, human rights, and the dignity and worth of the individual, is well positioned to address these issues in practice and policy, academic and research environments. This article introduces a taxonomy of sexuality and gender, encourages looking beyond Western biases of contemporary sexual and gender identity discourse, and briefly explores some sociopolitical contexts that pose critical challenges to sexual and gender minorities. Finally, some social work responses to these challenges will be considered.

Taxonomy

The taxonomy of sexuality and gender is complex. It draws on the psychological, political, and social work literature, as well as the growing interdisciplinary and international literature of queer theory. A full exploration of essentialist and social constructionist approaches to sexual identity (see Hammack, 2005) is beyond the scope of this article, but understanding the debate is helpful in understanding the development of sexual and gender identity theories. In brief, an essentialist position posits that individuals are born with a sexual identity, such as hetero- or homosexual, and that a major developmental task for such individuals is to discover and claim that essential identity. Social constructionist theories posit that sexual identities are largely constructed by their social and cultural contexts, and that people construct their identities in conformity to those contexts.

In the last few decades, the life course perspective that integrates these two approaches has emerged. Concepts such as heterosexual, homosexual, bisexual, and even male, female, and transgender may seem to be obvious at first glance, and in everyday use their meanings may seem clear. But there are a number of problems with these commonly used words. These apparently simple words refer to complex categories that comprise biology (DNA), anatomy (the so-called reproductive organs and secondary sexual characteristics), behaviors (who we have sexual activity with), attractions (who we are sexually attracted to), emotions (who we fall in love with), and fantasies (daydreams and night dreams). These words have often been used in ways that assume that these different attributes are aligned with each other, and are static and unchanging throughout our lives. Drawing on interdisciplinary research we now understand that people's identities are not static and that they can evolve and change over the course of their lives. The life course perspective of human development (Hammack and Cohler, 2009) proposes that biological desire motivates behavior, and individuals "internalize the sexual story possibilities" (Hammack, 2005: p. 281) of a culture in order to create lived identities. Individuals not only recognize their essential desires but also learn cultural stories to construct their evolving identities.

Challenging Binaries

The words male and female carry with them the assumption that people's genders are binary, that is, that people are either

one or the other. This notion is challenged by social work transgender theorists (Teich, 2012; Alleyn and Jones, 2010; Markman, 2011), who propose that sex (that is, biology and anatomy), gender identity (the individual experience of femaleness and maleness), and gender expression (behavior) can best be expressed as continua, rather than as discrete binaries. Some people experience themselves as gender syntonic, or cisgendered, that is, their experience of their gender matches the biological sex with which they were labeled at birth (Teich, 2012: p. 15), and some people experience themselves as gender dysphoric, gender dissonant, transgendered, or transsexual, that is, they do not identify with the sex to which they were assigned at birth. Some people are born with unambiguous sexual characteristics, and others are born with ambiguous sexual characteristics (Cornnwall, 2010), some of whom identify as intersex, or 'gender fluid' (Solomon, 2012: p. 602). Transgendered, intersexed, and gender fluid people may or may not choose consistently to present or express themselves publicly as male or female, and may or may not choose to modify their bodies with hormones, transformation surgery, or other means.

In the same way that gender is now becoming understood as a continuum, so also sexual behavior, attraction, emotion, and fantasy are also more appropriately understood as continua. The venerable Kinsey studies (Kinsey et al., 1948, 1953) introduced this idea in the mid-twentieth century, and in a more evolved form, this continuum concept is increasingly supported by research, particularly in women (Diamond, 2005; Hines, 2012; Rosario et al., 2006). For social workers to assume that a client who presents to the social worker as partnered to someone of the opposite sex is sexually active only with that partner, or only with others of that partner's sex, or to assume that the sexual (or even gender) identity of a long-term client will remain unchanged over time is to risk overlooking a significant part of the ecology of the client.

Sexual or gender minority individuals will probably experience themselves as different in some fundamental ways from their birth parents, and will in all likelihood develop an epistemological framework to understand that difference: they will quickly learn to question what they are told about people like themselves, and to validate knowledge primarily through their own experiences, rather than through what they are told by others. They will learn to reject the vertical identity provided by affiliation with their parents or caregivers and develop more horizontal identities (Solomon, 2012). Another salient feature of a sexual or gender minority epistemology is the decision to disclose. Most sexual or gender minorities cannot be identified by the casual observer by any outward appearance or presentation: they must choose to disclose, either by a verbal statement or in some other obvious way in each new environment, with each new person. That sexual and gender minorities must make a decision to disaffiliate with a cisgendered heteronormative world also shapes the way these individuals interact in the world. If a sexual minority client sits down in an agency waiting area and sees only family-life magazines, or overhears remarks about heteronormative family life from administrative staff, then the agency has established a heteronormative environment, and the responsibility to disclose a nonconforming identity now falls to the client.

A Global Context

Notions of individual sexual identities are largely perceived as Western constructs, where sexual identity becomes the entire or primary identity of the individual, rather than a matrix of intersecting social identities such as culture, tribe, or other social status. Social workers should not assume that these sexuality constructs are globally shared. In the Asia–Pacific context, for example, there is a multifarious array of social and cultural roles, behaviors, identities, and sexual and gender expressions that are quite different from Western-style lesbian, gay, bisexual, and transgender individual identities (Lim, 2008; Peletz, 2009; Murray, 2002). Yet, in multinational urban areas like Hong Kong or Singapore, Western-style sexual identities also exist (Martin et al., 2008). Similar observations can be made about other parts of the world (Whitaker, 2006; Murray and Roscoe, 1998). One challenge to local cultural and indigenous expressions of identity is the Internet, where English and its hegemonic Western identity constructs have been the dominant language. In an increasingly migrant world, social workers who use only Western constructs will miss key nuances to individual identities. The language of 'queer' by sexual and gender minorities is being reclaimed as an insider term, a "destabilizing rebellion against the formation of identities around fixed poles of gender or sexual orientation" (Shernoff, 2008: p. 143).

Heterosexual Privilege

The most significant critique of the prevailing social work taxonomy of sexual and gender minorities is that these notions are othering (Dominelli, 2002: p. 18) and exclusive. Such terms indicate putative cisgendered heterosexual power over sexual and gender minorities to name, objectify, pathologize, and confer privilege on cisgendered heterosexuality as both normal and normative. When social workers use words like lesbian, gay, or transgendered, they may be reproducing those categories, and thereby identifying them as inferior (Fish, 2008). A position of heteronormativity is thereby adopted from which to know about sexual and gender minorities as objects of knowledge. Thus, heteronormativity and heterosexuality remain unexamined and invisible (Hicks, 2008: p. 68), as well as privileged.

It is for these reasons that social workers, particularly those who work in multicultural environments, will wish to avoid objectifying language such as lesbian, gay, and bisexual: this vocabulary is not universally meaningful. In addition, lengthy and reductionist acronyms (e.g., LGBTIQQ) that attempt to be inclusive inevitably omit some possible identities, particularly indigenous cultural identities. How many heterosexuals would respond well to being labeled 'Het' in a social work text or client record? The language 'sexual and gender minorities' may be the most inclusive language, and can readily be used in practice and research until individuals and communities indicate which terms they prefer to use about themselves.

Sociopolitical Contexts: Sexuality and Human Rights

Sexuality must be understood in sociopolitical contexts. For social work, this also means that sexuality must be understood with a view to the fundamental social work principles of social

justice, human rights, and the dignity and worth of persons (International Federation of Social Workers, 2004). Human immunodeficiency virus (HIV), the increasing visibility of sexual and gender minorities, and the current debate about legal recognition of same-sex relationships have brought into sharp relief the statutory situations of sexual and gender minorities around the world. Although there have been significant gains in many countries and jurisdictions, the enduring existence of sexual and gender minorities in cultures and countries around the world does not by any means imply universal acceptance or even tolerance. Uncritical assumptions about what is normal has social, criminal, and too often fatal consequences for sexual and gender minorities – the ultimate expression of heterosexual privilege and dominance. The late lesbian writer Adrienne Rich argued that "heterosexuality has had to be imposed, managed, organized, propagandized and maintained by force" (1983, cited in Fish, 2008: p. 188).

State-Enforced Conformity

There are a number of countries in the world where consensual adult same-sex relationships are punishable by whipping (e.g., Malaysia), imprisonment (e.g., Myanmar/Burma, Tanzania, Uganda), torture (e.g., Uganda), or death (e.g., Iran, Mauritania, Qatar, Saudi Arabia, Sudan, Yemen, and parts of Nigeria) (Amnesty International, 2008). Although homosexuality is a capital crime in Iran, Iran has a liberal approach to transsexuality; Iran has the second highest rate of gender reassignment surgery in the world (after Thailand). Transsexuality is constructed as a medical condition and the state provides up to half the cost of transformation surgery. This means that many sexual minorities are pressured to undertake gender reassignment surgery to live their identities (Najmabadi, 2011). In 2007, Kuwait enacted a law that made it a criminal offense 'to imitate the appearance of the opposite sex' (Amnesty International, 2008: p. 12). Although the High Court of Delhi (India) declared unconstitutional Section 377 of its 1861 Penal Code, a legacy of the British Raj, which criminalized 'voluntary carnal intercourse against the order of nature' (Peletz, 2009), that decision was itself overturned by the Indian Supreme Court in December 2013.

In November 2012, the Uganda Parliament revived its infamous 'Kill the gays' bill (as it was known; Okeowo, 2012); in February 2013, Ugandan President Musevini signed a version of the law that criminalized same-sex sexual activity for both men and women with penalties ranging to life in prison, although it did not include the death penalty. This action followed similar legislation in Nigeria enacted in January 2014.

It is not that only sexual behaviors or appearances have drawn the attention of legislators. Although homosexuality was decriminalized in Russia in 1993, the Russian Duma passed a law in June 2013, which criminalized providing information about homosexuality to people under the age of 18 years. This law was the focus of much protest associated with the 2014 Winter Olympics in Sochi, Russia.

In some countries and regions, social workers and social work organizations have chosen to speak out against such political and cultural oppression and violence. In others, they have supported oppressive laws. In yet others, they have been forced, or have strategically chosen, to remain silent.

Socially Enforced Conformity

Even where oppressive statutes are not enforced, or where laws are enlightened, climates of public intolerance and fear can be created. South Africa is the first country in the world to offer sexual minorities equal protection in its constitution, but this protection does not stop political leaders from referring to sexual minority communities in stigmatizing language (Francis and Msibi, 2011: p. 162). South African lesbians as well as transgender men and heterosexual women who 'defy traditional roles' have lived in fear of so-called corrective rape and murder because the victims 'do not conform to a stereotype heteronormative gender role' (Bavoso, 2012). Reports of corrective rape have also come from Uganda (Okeowo, 2012).

HIV

There can be no doubt that one of the most significant influences on the contemporary history of sexual and gender minorities, particularly among men who have sex with men (MSM), has been HIV. The acquired immunodeficiency syndrome pandemic in developed nations was first identified among MSM in the late 1970s (then called GRID). In most developed nations, sexual and gender minority communities led the way in community organizing prevention through education, political advocacy, and a caring and compassionate response to people whose lives were irrevocably altered or ended by HIV. Social workers and social work researchers – particularly sexual and gender minority social workers – have contributed extensively to the global responses to HIV. Just as HIV spurred a massive global pharmacotherapeutic research effort with flow-on effects to other conditions, and served as a model for advocacy and policies in people affected by other chronic conditions, so it also spurred global social work research into areas such as effective prevention education models, treatment adherence, care and management, a reconceptualization of family and care, and the effects of stigma, which have had enduring benefits.

Religion

Another area where sexuality and gender are frequently contested is religion. Social workers are often anxious about trespassing into the domain of religions. Yet social workers are often members of religious groups and are informed by religious values systems (Cosis Brown and Cocker, 2011). Certainly, gender and sexual minorities are affected by religious values, both directly when they are included or excluded, and indirectly, as other people read about experiences of inclusion or exclusion. Religious organizations may use texts or traditions to establish boundaries about what is acceptable or tolerable behavior around individuals or groups (Henrickson, 2009). Faith-based discrimination, exclusion, and disgust are inappropriate in any setting, and are particularly inappropriate in a social worker. Wilton has written that "in cases where [faith] comes into conflict with a practitioner's ability to deliver respectful and appropriate care, there is a clear obligation to prioritize professional standards over personal belief or morality" (Wilton, 2000, cited in Cosis Brown and Cocker, 2011: p. 91). Attitudes such as 'I do not approve of their lifestyle but I would still work with them' are equally inappropriate, for it is unlikely that a sexual or gender minority individual would want to work with someone with

such judgments. Claims by social workers that they have a right to judgmental views justified by religion and that their views are not being respected have been addressed (Melendez and LaSala, 2006). Social workers can promote social justice approaches by providing assistance in developing advocacy, providing informed testimony and other support for decision-making processes within religious bodies, and assisting affected individuals to find alternative support or safety when it is clear that a particular religious body is actively unsafe. While in many places religious bodies are not governed by laws of the state, they are not immune to public opinion, and all religions have evolved their beliefs, theologies, and practices over time.

A Social Work Response

Antioppressive practice (Dominelli, 2002; Baines, 2011) and the companion discourse of antidiscriminatory practice have been the most frequently utilized theoretical and practice social work responses in the domain of sexuality (Hines, 2012). Brown has proposed, however, that antidiscriminatory discourse "has had very little to offer lesbians and gay men" (2008, as cited in Cocker and Hafford-Letchfield, 2010: p. 2002). Antioppressive practice appears ideally suited to challenge the power inequalities that exist between cisgendered heteronormative majorities and sexual and gender minorities. However, a critique of antioppressive practice is that it often places sexuality within a matrix of structural oppressions, such as racism, sexism, ageism, ableism, and classism (Massaquoi, 2011), and in these matrices, sexuality is often ranked low, hidden, or ignored. We have seen that labeling sexual and gender minorities is an act of othering that can itself be constructed as oppressive. Antioppressive practice as usually interpreted, then, may be a necessary but of itself insufficient response for social work in the domain of sexuality.

Social workers must critically examine their assumptions about heteronormativity, and interrogate those assumptions and their othering vocabulary, so that social work practice, education, and research do not replicate or reinforce heteronorms and privileges. This reflexive and liberating approach appropriately problematizes groups and individuals who exclude, rather than those who are excluded, and places responsibility for restoring equity and justice on sexual and gender majorities, rather than exclusively on sexual and gender minorities. Such an approach may mean that in an academic environment, where sexuality issues are not adequately included, cisgendered heterosexual teachers carry the responsibility for initiating discussions about heterosexual privilege in the same way that culturally dominant groups carry responsibility for initiating discussions about racism. Such proactive approaches are not only entirely consistent with, but are also arguably required by the shared code of ethics of the international social work organizations. These approaches also encourage critical self-examination that will liberate cisgendered heterosexual social work educators, practitioners, and researchers from roles, behaviors, attitudes, and assumptions that confine them.

Practice

Cocker and Hafford-Letchfield (2010: p. 2004) ask whether it is possible for social work to deliver an equitable service for lesbians and gay men. Their response is that there is work to do, and that this work must begin by improving partnerships in service delivery and research, and by transferring power from professionals to service users and their communities. As we have seen, social workers in all areas of practice have been largely informed by the discourses that problematize sexuality, where sexual identity is seen as the problem rather than the heteronormative environment in which the sexual or gender minority individual lives (Hicks, 2008). Child protection practitioners must engage even young children about their self-understandings, for sexual and gender minority children may experience themselves as different even when they may not have a language for that difference. To place a sexual or gender minority child with carers who are unwilling or unable to support that child's growing awareness of difference will put that child at as much risk as any other violence. Assessing same sex couples for suitability for foster care or adoption should focus not merely on the same-sex part of the relationship, but also on the overall suitability of the candidates.

Other practice areas present particular issues:

- In schools, sexual and gender minority children are at risk of bullying; even where the individual is not a member of a sexual or gender minority, the perception of membership may present sufficient risk. Antibullying interventions benefit all young people.
- Sexual and gender minorities face a unique constellation of challenges in health, mental health, and substance misuse settings, including increased risk of self-harm (D'Augelli and Patterson, 2001). The use of so-called reparative therapy, which attempts to change the orientations of sexual minority individuals, is highly disputed (Grace, 2006; Becksted, 2012), and has been called unethical (Bright, 2004).
- Social work with transgendered individuals in prisons and domestic violence shelters present particular challenges, but steps must be taken to ensure the safety and well-being of both transgender and nontrans people (Teich, 2012).
- Social workers in migrant and refugee settings must be mindful of legal and social restrictions about sexual and gender minorities in sending countries; stigmatized identities will likely have been concealed. Alternatively, some individuals may claim refugee status based on the way sexual and gender minorities or people perceived to be at risk for HIV are criminalized or in danger in their home countries.
- Aging individuals who are members of sexual or gender minorities demonstrate particular resilience and satisfaction compared to aging heterosexuals (Henrickson and Neville, 2012); nevertheless social workers in aged care settings should be mindful of the different social needs of different aging communities.
- Individuals who have multiple minoritized statuses should be supported to express their identities in ways that include and integrate all aspects of their identities (Henrickson, 2006).
- Social workers in policy-making settings who are working to establish legal and social equality for sexual and gender minorities will want to be mindful of the recent debate about 'homonormativity', an approach where the "resolution of the 'problem' of homosexuality is found through gradual legal and civil changes resulting in assimilation" (Hicks, 2008: p. 68). In many instances, gay and lesbian

rights groups adopt this gradualist model of change, even though it does not radically contest the foundations on which cisgendered heterosexual privileges are constructed.

Practice example 1: Mr Tomacelli is 81 years old and has been living in an aged care facility for the last 6 years. His only visitor has been his 50-year-old son, Paul, who has recently had to move further away, and can now visit only occasionally. At admission, Bob, the intake social worker, asked Mr Tomacelli if he was married or single. He responded that he had been married. He also said that his partner had died nearly 20 years ago, but they had raised a child together. Bob recorded that Mr Tomacelli was a widower with one child. Lately Mr Tomacelli has appeared withdrawn and depressed; he does not engage with other residents, and does not participate in any of the activities offered at the facility. He avoids Bob. Terry, another social worker at the facility, has informally engaged with Mr Tomacelli. Eventually, Mr Tomacelli discloses to Terry that he married young, and divorced after a few years. His life partner of nearly 25 years was Alessandro, who died almost 20 years ago. Mr Tomacelli and Alessandro raised Alessandro's son Paul together. The twentieth anniversary of Alessandro's death is approaching, and now that Paul has moved away, Mr Tomacelli has no one in his life with whom to share this important anniversary. Because homosexuality was illegal for most of Mr Tomacelli's life. He does not talk with strangers about his relationship with Alessandro, and does not consider himself gay. At this stage of his life, he is extremely lonely, and has only his memories. During the intake interview, Bob should not have worked from a heteronormative assumption that Mr Tomacelli had only one partner, or that the partner was female. Bob could have asked an open question about relationships in Mr Tomacelli's life, or invited him to talk about the important people in his life. Bob could have understood that an intake interview may need to extend over time, particularly as sexual minorities need time to assess whether individuals are trustworthy. Even if Bob had thought to ask, "Are you gay?", Mr Tomacelli would have denied it, since he does not use this language about himself. The responsibility fell to Mr Tomacelli to disclose, and since he did not know Bob, he did not do so. Terry's more informal engagement allowed Mr Tomacelli the time to assess how Terry might react to his disclosure, to share more about his life, to talk about the key people in his life, and how much they meant to him.

Practice example 2: Aliyah arrived in the country as a refugee a few weeks ago. She had been living in a designated refugee camp for 2 years following her escape from her village. The rebels had killed all of her family, raped her, and left her for dead. She saw her survival as both a blessing and a curse, but was able to make her way to an internationally protected safe area. Now, 2 years later, she has been relocated to a new country. The experienced social worker at the refugee support agency had read the official record and was extremely sympathetic about the trauma Aliyah had endured, but it was clear that she was not telling her whole story. Over time the social worker expressed only support, and gradually gained Aliyah's trust. Eventually she was able to tell the story of the massacre in her village: not only were her parents and extended kin killed, but also her children, and her female partner. None of her family knew about Aliyah's partner, but she suspects that her partner's brother may have betrayed them to the rebels. She now has great difficulty in trusting anyone. However, after her disclosure, the social worker helped Aliyah to make contact with a local support network for sexual minority women. The social worker's nonjudgmental attitude, support, and openness to hearing all of Aliyah's story gave her the confidence to share more of her story. Now Aliyah has the choice to access appropriate support for all aspects of her life, and not just the obvious traumas that she has experienced. The word lesbian has no meaning for Aliyah. She has been in relationships with both men and women, although at the time of the massacre, her primary emotional relationship was with her female partner.

Education

We know that education is effective in changing social work student attitudes about sexual minorities, and a number of authors have described their experiences offering courses and curricular content about heterosexism and sexual and gender minorities (Chonody et al., 2009; Francis and Msibi, 2011). Using an approach that problematizes heteronormative privilege rather than sexual or gender identities, sexual and gender minority individuals are not forced to disclose their identities in environments that may not be safe: although there is support for the approach that coming out in the classroom is a necessity for the sexual minority social work educator, there is also recognition that there is risk to openly disclosing educators, including lower student confidence, student perceptions that the lecturer is of poorer character, and lower credibility than heterosexual colleagues (Gates, 2011).

Sexual and gender minority students, even those who perform well, may arrive in the classroom with a history of experiences of harassment, discrimination, and even violence. Social work educators will need to make explicit statements to all students about their inclusion of sexual and gender minorities. Although it risks perpetuating the discourse of sexual and gender minority as a cultural other, Van Den Bergh and Crisp (2004) set out a useful attitudes, knowledge, and skills heuristic for social work education. Such content should be consistent across social work educational programs within a country, and assessed by professional regulatory bodies. An open attitude and the creation of safe spaces can be expressed in lectures, case presentations, case examples, classroom discussions, and examinations. Field education settings should be assessed to ensure that they are safe for sexual and gender minority students, and where necessary field supervisors should be upskilled, or alternative sites found.

Research

Jeyasingham (2008) advocates for more scholarship that identifies the ways day to day practice and social apparatuses operate to reinforce homophobia in social work contexts. Meezan and Martin (2009) have published an accessible and useful guide to research with sexual and gender minorities. As in practice, improving partnerships in research, transferring power from professionals to research participants and their communities, and avoiding othering language will benefit not only communities, but also social work theory building.

Conclusion

Sexuality inheres within all social work practice areas. Because of its core ethical principles, social work is well positioned to interrogate assumptions that identity is binary, stable, and fixed, and to challenge vocabularies of othering; to avoid uncritical assumptions about what is normal; and to oppose oppressive and lethal state policies and practices. Social workers must not problematize any minoritized identities but instead uphold core social work principles such as social justice, human rights, and the dignity and worth of the individual. Social work practitioners, educators, and researchers can and must take up the

possibly painful and difficult risk of critical self-examination and become aware of places where they, their colleagues, and students participate in the oppression of sexual and gender minorities, including where they are agents of oppressive or exclusive state policies. Social workers in all practice areas must become aware of cisgendered heterosexual privilege and the way that privilege supports existing oppression and exclusion, and maintains hegemony over sexual and gender minorities.

See also: Critical Social Work Practice; Human Rights and Social Work; Social Work Theory.

Bibliography

Alleyn, C., Jones, R.L., 2010. Queerying care: dissident trans identities in health and social care settings. In: Jones, R.L., Ward, R. (Eds.), LGBT Issues: Looking beyond Categories. Dunedin Academic Press, Edinburgh.

Amnesty International, 2008. Love, Hate and the Law: Decriminalising Homosexuality. Amnesty International, International Secretariat, London.

Baines, D. (Ed.), 2011. Doing Anti-oppressive Practice: Social Justice Social Work. Fernwood, Halifax.

Bavoso, John, 2012. Course-correcting the conversation surrounding 'corrective rape'. Diplomatic Courier [Online]. Available: http://www.diplomaticourier.com/news/regions/africa/920 (accessed 07.12.12.).

Becksted, A. Lee, 2012. Can we change sexual orientation. Archives of Sexual Behavior 41, 121–134.

Bright, Chuck, 2004. Deconstructing reparative therapy: an examination of the processes involved when attempting to change sexual orientation. Clinical Social Work Journal 32, 471–481.

Chonody, J.K., Rutledge, S.E., Seibert, D.C., 2009. College students' attitudes toward gays and lesbians. Journal of Social Work Education 45, 499–512.

Cocker, Christine, Hafford-Letchfield, Trish, 2010. Out and proud? Social work's relationship with lesbian and gay equality. British Journal of Social Work 40, 1996–2008.

Cornnwall, Susannah, 2010. Sex and Uncertainty in the Body of Christ: Intersex Conditions and Christian Theology. Equinox, London.

Cosis Brown, Helen, Cocker, Christine, 2011. Social Work with Lesbians and Gay Men. Sage, London.

D'Augelli, A.R., Patterson, C.J. (Eds.), 2001. Lesbian, Gay and Bisexual Identities and Youth: Psychological Perspective. Oxford University Press, Oxford, UK.

Diamond, L.M., 2005. A new view of lesbian subtypes: stable versus fluid identity trajectories over an 8-year period. Psychology of Women Quarterly 29, 119–128.

Dominelli, Lisa, 2002. Anti-oppressive Social Work Theory and Practice. Palgrave Macmillan, Basingstoke, UK.

Dunk-West, Priscilla, Hafford-Letchfield, Trish (Eds.), 2011. Sexual Identities and Sexuality in Social Work, Surrey. Ashgate, England.

Fish, Julie, 2008. Far from mundane: theorising heterosexism for social work education. Social Work Education: The International Journal 27, 182–193.

Francis, Dennis, Msibi, Thabo, 2011. Teaching about heterosexism: challenging homophobia in South Africa. Journal of LGBT Youth 8, 157–173.

Gates, Trevor G., 2011. Coming out in the social work classroom: reclaiming wholeness and finding the teacher within. Social Work Education: The International Journal 30, 70–82.

Grace, André P., 2006. The charisma and deception of reparative therapies: when medical science beds religion. Journal of Homosexuality 55, 545–580.

Hammack, Phillip L., 2005. The life course development of human sexual orientation: an integrative paradigm. Human Development 48, 267–290.

Hammack, Phillip L., Cohler, Bertram (Eds.), 2009. The Story of Sexual Identity: Narrative Perspectives on the Gay and Lesbian Life Course. Oxford University Press, New York, NY.

Henrickson, Mark, 2006. Ko wai ratou? Managing multiple identities in lesbian, gay and bisexual New Zealand Maori. New Zealand Sociology 21, 251–273.

Henrickson, Mark, 2009. Sexuality, religion and authority: towards reframing estrangement. Journal of Religion and Spirituality in Social Work: Social Thought 28, 48–62.

Henrickson, Mark, Neville, Stephen, 2012. Identity satisfaction over the life course in sexual minorities. Journal of Gay and Lesbian Social Services 24, 80–95.

Hicks, Stephen, 2008. Thinking through sexuality. Journal of Social Work 8, 65–82.

Hines, Jude M., 2012. Using an anti-oppressive framework in social work practice with lesbians. Journal of Gay and Lesbian Social Services 24, 23–39.

International Federation of Social Workers, 2004. Ethics in Social Work, Statement of Principles [Online]. International Federation of Social Workers, Bern, Switzerland. Available: http://www.ifsw.org/en/p38000324.html (accessed 02.05.06.).

Jeyasingham, Dharman, 2008. Knowledge/ignorance and the construction of sexuality in social work education. Social Work Education: The International Journal 27, 138–151.

Kinsey, A.C., Pomeroy, W.B., Martin, C.E., 1948. Sexual Behavior in the Human Male. W.B. Saunders Co., Philadelphia.

Kinsey, A.C., Pomeroy, W.B., Martin, C.E., Gebhard, P.H., 1953. Sexual Behavior in the Human Female. W.B. Saunders, Philadelphia.

Lim, Song Hwee, 2008. How to be queer in Taiwan. In: Martin, F., Jackson, P.A., McLelland, M., Yue, A. (Eds.), AsiaPacifiQueer. University of Chicago Press, Urbana, IL.

Mallon, Gerald P. (Ed.), 2008. Social Work Practice with Lesbian, Gay, Bisexual and Transgender People. Routledge, New York.

Markman, Erin R., 2011. Gender identity disorder, the gender binary, and transgender oppression: implications for ethical social work. Smith College Studies in Social Work 81, 314–327.

Martin, Fran, Jackson, Peter A., McLelland, Mark, Yue, Audrey (Eds.), 2008. AsiaPacifiQueer: Rethinking Genders and Sexualities. University of Illinois, Urbana.

Massaquoi, Notisha, 2011. Crossing boundaries to radicalize social work practice and education. In: Baines, D. (Ed.), Doing Anti-oppressive Practice: Social Justice Social Work. Fernwood, Halifax, NS.

Meezan, William, Martin, James I., 2009. Handbook of Research with Lesbian, Gay, Bisexual and Transgender Populations. Routledge, New York.

Melendez, M.P., LaSala, M.C., 2006. Who's oppressing whom? Homosexuality, Christianity and social work. Social Work 51, 371–377.

Montgomery, S.A., Stewart, A.J., 2012. Privileged allies in lesbian and gay rights activism: gender, generation and resistance to heteronormativity. Journal of Social Issues 68, 162–177.

Murray, Stephen O., 2002. Pacific Homosexualities. Writers Club Press, San José, CA.

Murray, Stephen O., Roscoe, W. (Eds.), 1998. Boy-Wives and Female Husbands: Studies in African Homosexualities. Palgrave, New York.

Najmabadi, Afsaneh, 2011. Transing and Transpassing across Gender Walls in Iran [Online]. Feminism without Borders. Available: http://stealthishijab.com/2011/06/02/transing-and-transpassing-across-sex-gender-walls-in-iran/ (accessed 24.12.12.).

Okeowo, Alexis, 2012. Out in Africa: a dangerous fight for gay rights. The New Yorker. Condé Nast, New York.

Peletz, Michael G., 2009. Gender Pluralism: Southeast Asia since Early Modern Times. Routledge, New York.

Rosario, M., Schrimshaw, E.W., Hunter, J., Braun, L., 2006. Sexual identity development among lesbian, gay, and bisexual youths: consistency and change over time. Journal of Sex Research 43, 46–58.

Shernoff, Michael, 2008. Social work practice with gay individuals. In: Mallon, G.P. (Ed.), Social Work Practice with Lesbian, Gay, Bisexual and Transgender People, second ed. Routledge, New York.

Solomon, Andrew, 2012. Far from the Tree: Parents, Children and the Search for Identity. Scribner, New York.

Teich, Nicholas M., 2012. Transgender 101: A Simple Guide to a Complex Issue. Columbia University Press, New York, NY.

Van Den Bergh, N., Crisp, C., 2004. Defining culturally competent practice with sexual minorities: implications for social work education and practice. Journal of Social Work Education 40, 221–238.

Whitaker, Brian, 2006. Unspeakable Love: Gay and Lesbian Life in the Middle East. University of California Press, Berkeley.

Sexuality and Space

Lynda Johnston, University of Waikato, Hamilton, Aotearoa, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sexuality and space are mutually constituted – space makes sexuality and sexuality makes space. Sexuality – as with other identity categories like gender and race – is not fixed or static, but shifts and changes depending upon politics, place, and space. Sexuality and space scholarship can be found in social, cultural, feminist, and queer geographical subdisciplines. This scholarship questions the ways in which sexual identities conform to, and/or resist, certain societal norms. Attention to the intersections of sexual identities with other identities – such as gender, race/ethnicity, age, and (dis)ability – highlights current and ongoing debates in studies of sexuality and space.

Sexuality and space are mutually constituted – space makes sexuality and sexuality makes space. Sexuality – as with other identity categories like gender and race – is not fixed or static, yet is constantly shifting across various cultural and social landscapes. Sexuality and space are also highly contested and the ways in which we live out body/place relationships are political. Over the last three decades geographers have recognized that the body is a politicized site of contestation, struggle, and possibilities. The materiality of bodies is both fluid and fixed, constricted and freed by gender, sex, sexuality, race, class nationality, and disability. In other words, our understanding of sexuality, sexed bodies, and sexual identities is socially crafted, *and* laden with knowledge and politics that are particular to time and place.

Studies of sexuality and space map, connect, and integrate sexed bodies not just through the most obviously sexed public spaces such as bars, casinos, and city red light districts, but also through all spaces. Think, for example, about sexuality and the space of homes (Morrison, 2012). Homes are sites of a range of variety of sexual relations. Nuclear families, extended families, gay and lesbian homes, polyamorous homes, and young people living together are just a few examples of sexual relations that are constituted within intimate and often private home spaces. Whether it is a bar, casino, red light district, or a home sexual politics permeate space. Equally, there are no spaces that sit outside of sexual politics. Sex, space, and politics cannot be disentangled, and geographers have been unraveling some of the complexity and diversity that constructs the embodied experiences of sex and sexuality.

It is not just 'real' or material space that matters to sexed bodies; virtual spaces are equally significant to constructions of sexuality. There are millions of web sites on every kind of sexual identity and sexual performance, for example, pornography, romance, 'find a bride', gay, lesbian, bisexual, and transgendered web sites. Online social media, such as Facebook, is another example of virtual space providing new opportunities for sexual identities, connections, interactions, and communities to emerge. Virtual spaces may enable people to try out new sexual identities unrestricted by the limitations of their real bodies. As technology continues to develop, new levels of communication in nonphysical connections between people are enhanced. Erotic connections are formed and people engage in sexual intimacies, get married, have affairs, and so on, all online.

In what follows I first outline some of the earliest scholarship on sexuality and space. These informal beginnings sparked a deep interest in identity, place, and power. Second, attention is turned to the influence that feminist and queer theories have had on sexuality and space research. During this period, geographers in particular began to question the ways in which heterosexual and homosexual identities conformed to certain societal norms. Finally, the intersections of sexual identities with other identities – such as gender, race/ethnicity, age, and (dis)ability – are offered by way of highlighting current and ongoing debates in studies of sexuality and space.

Geographies of Sexualities: Beginnings

Scholarship on sexuality and space started to appear in the late 1970s. Geographers interested in pursuing this research agenda tended to meet informally because sexuality was deemed not to be a topic that geographers should engage in, but by the end of the 1970s some publications could be found. By the 1980s – when positivism started to lose its grip on the discipline – more sexuality and space studies emerged (Lauria and Knopp, 1985). The 1980s is perhaps best known for the period in which the intersections of class and gay identity dominated scholarship as studies considered the role of gay men and gentrification in US cities (Lauria and Knopp, 1985).

The first lesbian urban geography scholarship appeared in the early 1990s and this period is particularly notable for the groundbreaking research by Gill Valentine (1993). Also significant during the 1990s were the number of studies that discussed topics such as 'eroticized topographies' and sexual identities (Bell, 1994). Much of the 1990s literature on space and sexuality was collected together in one volume – David Bell and Gill Valentine's *Mapping Desire* (1995) – that included topics such as sexualized landscapes, homes, suburbs, lesbian neighborhoods, red light districts, rural spaces, Jamaican ragga, and gay resistance, to name just a few examples. This was a period of scholarship in which it was asserted that sexual politics permeate all space – private and public, urban and rural, at the macro and micro level. Notably, these studies show the desire to draw on feminist, poststructural, and queer

theories in order to examine 'difference.' Many of these studies challenged the notion of fixed identities and spaces by focusing on the proliferation of identities, rather than structural inequalities. Bell and Valentine (1995: 4) frame this early work by noting:

> The women's, gay and civil rights movements emerged in North America and Europe in the 1960s and 1970s on a wave of social and political upheaval. But despite a growing awareness amongst geographers in the following decade of the need to study the role of class, gender and ethnicity in shaping social, cultural and economic geographies, sexualities were largely left off the geographical map.

Bell and Valentine's edited collection was vitally important in helping put sexuality on the geographical map. This collection is known for its empirically rich case studies and less for its advancement of sexuality theories. It opened up, however, pathways for scholars to focus on spatial theories of desire. During the mid-1990s queer politics and activism emerged when numerous governments failed to respond adequately to the AIDS crisis (particularly in the United Kingdom and the United States). New forms of queer activism emerged in order to raise awareness and support for people with AIDS. Geographers have made effective use of both queer politics and queer theory to further understand the performance of sexed bodies and sexual identities in space.

It is not surprising that this early work reflects a disparity between theory and empirics as it reflects the ways in which knowledge and power intersect. Notably geographers have been slow to question taken-for-granted sexualities, such as heterosexuality. The pioneering scholarship of Adrienne Rich's (1980) and her focus on 'compulsory heterosexuality' has been picked up by some geographers who have been able to show the ways in which most people unconsciously use heterosexuality to structure their lived everyday spaces. The socially constructed 'naturalness' of heterosexuality is produced at a variety of scales: babies are expected to be born male or female (not all are); girls and boys (in modern industrial contexts) are expected to grow up to be mothers and fathers and create nuclear households; and a set of prescribed gender norms are attached to these heterosexual roles (Butler, 1990). This normalizing of heterosexuality is also incorporated into the spaces of geographical disciplines which means heteronormativity pervades the ways in which geographers investigate place and space (Rose, 1993). Some have charted the ways in which heterosexism pervades the institutional practices of geography (and creates heterosexist scholarship (Binnie, 1997; Rose, 1993)). Overall, these scholars argue that heterosexism underpins geographical notions of self, place, gender, and sexuality. Hence, the 1990s can be recognized as a period of sexuality and space scholarship which is noted for the critiques of studies that reproduced hegemonic versions of heterosexuality – including family life – and made these studies appear natural, legitimate, innocent, and even morally right.

The term 'heteronormative' is now a widely used concept. Gayle Rubin (1993) was one of the first feminists to first use the concept, and since then geographers, such as Hubbard (2008: 643), have researched places and spaces that are inscribed by 'a normative heterosexuality, in which it is assumed individuals' sexual identities conform to a social norm of heterosexual love, sex, and reproduction'. Heterosexual norms are constructed and reproduced in almost every space and place. Some heterosexualities are considered deviant (for example, sex workers), while others are valorized and held up as exemplar identities that are enshrined in biological discourses (for example, husband and wife couples).

The work of Heidi Nast (1998) is influential here, as she shows how heteronormativity can be found within an array of everyday geographies from the body to the globe. She draws on modern nationalist discourses – languages and imagery – to show the ways in which nation states depend on a type of pure representation of the maternal and of nuclear families. Some nation states are represented using phallic imagery and language that glorify the Oedipal nuclear family father while other nation states represent their territory using images and language of the maternal. Furthermore, in the context of industrial interests the nuclear family is promoted as the core setting for reproducing and socializing industrial workers.

A great deal of discussion has also addressed the notion of 'homonormativity.' It has been adopted to show how certain performances of homosexuality incorporate the logic of heteronormativity. Becoming a normative homosexual usually means gaining the privileges associated with neoliberal politics and practices that are aligned with heterosexuality and class privileges (Browne et al., 2007). Such practices produce a new 'homonormative' identity, which gains popularity as more radical performances of homosexuality become marginalized. 'Acceptable' gay public spheres shrink and typically become defined as consumption spaces and gentrified neighborhoods. A new kind of normativity is emerging as countries around the world debate whether to change laws so that all people, regardless of sex, sexual orientation, or gender identity will have the opportunity to marry if they so choose. These public debates are often centered on the ways in which religious institutions regulate normative sexualities. The concern is that marriage becomes the defining relationship upon which access to social justice and forms of belonging are then measured. These state legislated couple normalities need to be questioned using queer theories.

Queer Spaces, Circuits, and Fields: Questioning Normativities

The turn to queer geographies (Binnie, 1997) is, in part, a reaction against heterosexist geographies, but also an opportunity to distance the subfield of sexuality and space from feminist geography. Queer geographies, however, have benefited from feminist geographers' critical engagement with gender and sexuality. Feminist geographers have been quick to question the authority assigned to science and 'objective' claims of 'universal truth.' A very good example of the use of queer theories by feminist geographers is the work of Gibson-Graham (1996). In the article 'Querying Globalization' Gibson-Graham (1996) draw on feminist theorizations of sexual identities to upset the naturalness and legitimacy of global capitalism. They bring together Marxian theoretical constructions of capital and rape discourses that construct women as victims. Marxian theories often script capitalism as

violently irrepressible, and as discourse 'capitalism' gathers strength from heterosexualized language. Gibson–Graham (1996: 146) attempt to make "globalization less genital, less phallic, by highlighting various points of excess in its inscriptions – places where the inscription can be seen as uncontrollable or indeterminate, or as potentially inscribing noncapitalist identity."

Sexuality has a profound effect on the way people live in, and interact with, space and place. In turn, space and place affect people's sexuality. Gill Valentine (2001: 5) offers an example of the way place and sexuality are mutually constituted, explaining:

> Spatial visibility (e.g., in terms of the establishment of so-called gay ghettos or various forms of street protest or Mardi Gras) has been important to the development of lesbian and gay rights. In turn, these performances of sexual dissidents' identities (re)produce these spaces as lesbian and gay spaces in which sexual identities can be, and are forged.

What Valentine's example makes clear is that space and sexuality do not just interact with, or reflect, each other but rather are mutually constituted. Geographers have noted that place matters – where one is located, what language is being used, and the privileges or lack of – to the production of sexuality and space scholarship. For example, sexuality and space knowledge is different 'down under' in Aotearoa, New Zealand, and Australia to that produced elsewhere.

Queer geographies question the relationships between gender, sex, and sexuality and assert that we have multiple genders and sexualities. At the core of this scholarship is the understanding of sexuality as a social process that allows for expressions of desire (in all directions and beyond dichotomizing discourses). Queer geographers have become dissatisfied with essentializing, limiting, neoliberal, and often Westernized sexual categories. They have turned to fluid, unstable, and ambiguous conceptualizations of sexualities. Spaces, like sexualities, are also considered fluid and ambiguous. This is a move away from "reified phenomena such as 'gay gentrification' and 'gay neighborhoods' (which also tended to be celebrated) to understanding more fluid, ambiguous and contingently sexualized spatialities such as circuits and fields" (Knopp, 2007: 49).

Many articles on sexuality and space are sprinkled throughout a variety of social, cultural, and feminist geography literature. For example, studies such as 'the closet' (Brown, 2000), which is used as both a metaphor but also as a 'spatial practice of power-knowledge' that takes up the notion of spatial scale. The body, city, nation, and globe are spaces in which the closet may be a place of creative and transformative sexual and political resistance to heteronormativity (Brown, 2000).

There are also a number of edited (e.g., Browne et al., 2007) and authored books (e.g., Johnston and Longhurst, 2010; Hubbard, 2011) on the topic of sexuality and space. Topics explored in this literature include the pink economy, gay tourism, gay pride parades, bisexuality, sexuality in rural areas, sexuality and the city, sexuality and citizenship, the globalization of sexuality, HIV and AIDS, and sexuality and ethnicity.

In Browne et al.'s (2007) edited collection *Geographies of Sexualities*: *Theories*, *Practices and Politics*, they say 'geographies of sexualities' is a young field but has 'blossomed' over the past 10 years. It is an area that is now taught in a number of undergraduate courses, at least in countries such as the United Kingdom, the United States, Canada, Australia, and New Zealand, and has succeeded in striking up conversations with a range of other subdisciplinary areas such as cultural and social geography, gender studies, migration studies, and tourism. It's important to note that sexuality and space is not a discrete intellectual field but one that has emerged through its relation with other disciplinary areas.

Sexuality and Other Identities: Queerer Possibilities

In a recent review of sexuality and space, Michael Brown (2012) notes a number of 'intersectional anxieties' that are worth raising as a way of charting future research. He turns his attention to the ways in which sexuality and space studies have (and have not) intersected with other axis of identity, such as gender, race, class, age, (dis)ability, and religion. A consensus is building that if studies of sexualities and space are to be truly queer, then they cannot focus solely on sexuality; rather, connections with other social identities must be considered (see Oswin, 2004). This kind of critique has been leveled at queer theories and activism (Browne et al., 2007).

Geographers stress, not surprisingly, how place brings different identities together (see Peake, 2010 for an account of the intersections of sexuality and race). Geographers have begun to take seriously the notion of intersectionality. The concept of intersectionality can be traced back to the early 1990s in feminist legal scholarship. Kimberle Crenshaw's (1991) article 'Mapping the margins: intersectionality, identity politics and violence against women of color' highlighted the ways in which a black woman is oppressed by both racism *and* patriarchy in ways that black men are not. She argued further that the danger of focusing on a one-dimensional identity and experiences of inequality associated with that one identity – meant researchers risk marginalizing other identities and hence erasing the ways in which people become positioned differently from one another. Crenshaw (1991), therefore, urged researchers to pay attention to the specificities of experienced oppressions.

Some geographers (Olund, 2010) reject Crenshaw's early argument of intersectionality because of the tendency to think of identities as separate (rather than enmeshed) and hence, ironically, essentializing. Using the term 'anticategorical intersectionality', it is argued that the metaphor of intersection is premised on an empirical or conceptual separation of two identities.

So how has sexuality intersected with other identities to produce different spaces? Brown (2012) notes that most scholarship looks at the importance of place in bringing different identities together and less research has been done looking at the convergence of identities across spaces. The scholarship that does exist is uneven.

One of the most researched intersections is sex, gender, and place. A number of studies group together lesbians and gays as single populations (Mulligan, 2008; Waitt, 2005). Also

prominent is work on sexuality and space that separates lesbians (Holmes, 2009; Johnston, 2007; Podmore, 2006) and gay men (Waitt, 2005) at a variety of spatial scales, from bodies and homes to communities, institutions, nations, and the globe.

After many cries for critical scholarship on heterosexuality and space, there are now publications that attend to the intersection of genders and heterosexual identities (Hubbard, 2008; Morrison, 2012). Some have focused on the ways in which particular women may be disciplined and controlled in heteronormative and patriarchal ways, and attention to heterosexual men is a recent trend. New and changing forms of masculine heterosexuality can be found in lifestyle magazines, for example, where sexual prowess is a key component of contemporary masculine identities. It is notable that some towns and cities appear to encourage sexualized rituals more easily than others. Newcastle in England, for example, has a reputation as a party city and is a site that is associated with excessive drinking, wild stag, and hen nights. Similarly, seaside resort towns have become tourist centers of stag and hen parties. These places accommodate visitors that wish to consume adult entertainment usually associated with premarriage celebrations, such as strip clubs, lap dancing, sex toy shops (and the public display of sex toys), cross-dressing, and so on (Hubbard, 2008). In much of this work geographers seek to highlight the privileges *and* oppressions at the intersections of genders and sexualities.

Trans identities have added further complexity to the above scholarship of gender and sexualities. Recent studies by geographers have shown how transphobia and sexism are experienced by trans people in homonormative gay and lesbian spaces, and some have suggested alternative places where support and security can be found. This scholarship challenges any notions that gender is stable and, at the same time, illustrates the intersection of gender and sexuality (see, in particularly Petra Doan's (2010) powerful and moving autoethnographic account of the persistent 'tyranny' of gendered spaces). What is clear from this scholarship is the political urgency to continue questioning, challenging, and subverting the binary system of gendered and sexed spaces. Trans geographies is one way to do this (Browne et al., 2010; Doan, 2010).

Sexuality and space scholarship that recognizes race and ethnicity may highlight the ways homonormative spaces can privilege whiteness and as a consequence are deeply racist (Caluya, 2008). Attention too has turned to the prevailing whiteness of queer scholarship. Jasbir Puar (2006) has provided a compelling account on American homonormativities after 9/11. Other researchers have shown how homophobic discourses are put into operation in order to seek ethnic and national purity by, for example, reproducing narratives of Irishness (and attempting to exclude gay, lesbian, bisexual, transgender, and intersex (GLBTI) performances) at the St Patrick's Day parade in New York City (Mulligan, 2008) and in forming a national closet in Latvia (Waitt, 2005).

Class and sexuality is another fruitful intersection that is receiving more attention in sexuality and space scholarship (Richardson et al., 2009). Jon Binnie (2011: 21) takes an opportunity to highlight three themes "(1) the classing of epistemology and methodology within sexuality studies, (2) the relationship between culture and the economy in framing research on class and sexuality, (3) the significance of space, place, and intersectionality." Privilege and oppression can be illuminated when considering the intersections of class positions, gender, and sexed identities. Yvette Taylor (2011) delves deeply into class divisions of lesbian identity and spaces. She shows how classed identities fracture collective lesbian identities both emotionally and materially. Others have highlighted the class privileges that align with homonormative gay men and lesbian spaces and identities (Holmes, 2009). The confluence of poor, working class, and heterosexuality has also been considered in, for example, Nepal (Richardson et al., 2009).

Work on sexuality, space, and age is another intersection of identities that is beginning to gain prominence. There are accounts of the way in which heterosexual patriarchal discourses in, for example, Vietnam work to control young women. Bringing together race, sexuality, age, and education, Mary Thomas (2011) considers the space of a high school campus for 26 Latina, Armenian, Filipina, African-American, and Anglo girls. Her research shows that while the girls use the 'everyone just wants to get along' discourse, they also perform racism and uphold spatial segregation. Young gay, lesbian, and bisexual youth are the focus of Andrew Gorman-Murray's (2008) study into Australian family homes. Here Gorman-Murray moves beyond geographical literature that has predominantly represented heterosexual nuclear family homes as spaces of exclusion, violence, and abuse for young gays, lesbians, and bisexuals. He notes that while these injustices exist, geographers must also pay attention to how queer youth may queer heterosexual homes and "provide a space for the fluorescence of nonheterosexuality within an apparently heteronormative site" (Gorman-Murray, 2008: 31).

There are important connections between disabilities and sexualities. Geographers have studied the way in which nonnormative sexualities have been categorized as illness, or disease and some studies highlight the challenges that disabled bodies pose to hetero and homonormativities. Work also exists that explores health issues for gay men. (There is enormous scope to examine the interconnectedness between place, (dis) abilities, impairment, and sexual identities although geographers have yet to make a substantial or sustained contribution to this scholarship.) The association of physical ableism with hegemonic notions of beauty and sexuality needs critical examination, as well as research that critiques spatially embodied normativities.

Conclusion

In the first edition of this encyclopedia, Heidi Nast (2001: Health Education and Health Promotion) writes "sexuality-geography research needs to be more theoretically and empirically attuned to differences produced through normative and oppositional constructions of race, gender, class, disability age, religious beliefs, and nation." Her words hold true today. While much contemporary sexuality and space scholarship is attuned to intersections and axis of difference, there is still room for more research to examine these connections.

Over the past decade a substantial corpus of research on sexuality and space has emerged. This research is crucial in understanding the production and expression of sexual identities and sexual practices, whether they be gay, lesbian, bisexual, trans, intersex, queer, asexual, polyamorous, or heterosexual. Despite the rapid growth in work on sexuality and space over the past decade, there is still a great deal of homophobia and heterosexism within the academy. The topic of sex and sexuality cannot just be added to an unchanged geographical discipline, rather the heternormativity of the academy needs challenging in order to expose its absences and institutionalized homophobia. Of course this homophobia and heterosexism also exists well beyond the academy in everyday spaces. Our debates as researchers of sexuality and space shape our disciplinary homes, as well as every other space we may be in. In order to realize and acknowledge the diversity of sexualities, geographers must continue to talk about sex, bodies, erotics, love, pleasure, power, pain, place, and space. It is vitally important to discuss topics such as 'asexuality, polyamory, homeless people having sex, a small town that became known as the sex change capital of the world, and bestiality' in order to push the boundaries of what geographers perceive to be important and relevant sites of investigation (Johnston and Longhurst, 2010: 161). A continued interrogation of sexuality and space is required in order to further understand the myriad of ways in which power, knowledge, and subjectivities encompass sexed, sexy, sensual, and sexual bodies.

See also: Aging and Older People, Geography of; Children's Geographies; Consumption, Geography of; Critical Realism in Geography; Cultural Geography; Development Geographies; Disability, Geography of; Gender and Feminist Studies in Geography; Health, Geography of; Landscape; Postcolonial Geography; Qualitative Methods in Geography; Race and Racism, Geography of; Refugees: Geographical Aspects; Rural Geography; Social Geography; Social Movements, Geography of; Space and Social Theory in Geography; State, Geography of; Surveillance and Privacy, Geography of; Urban Geography; Youth, Geography of.

Bibliography

Bell, D., 1994. Erotic topographies: on the sexuality and space network. Antipode 2 (1), 96–100.
Bell, David, Valentine, Gill (Eds.), 1995. Mapping Desire: Geographies of Sexualities. Routledge, London.
Binnie, J., 1997. Coming out of geography: towards a queer epistemology. Environment and Planning D: Society and Space 15, 223–237.
Binnie, J., 2011. Class, sexuality, and space: a comment. Sexualities 14, 21–26.
Brown, M., 2000. Closet Space: Geographies of Metaphor from the Body to the Globe. Routledge, London.
Brown, M., 2012. Gender and sexuality I: intersectional anxieties. Progress in Human Geography 36 (4), 541–550.
Browne, K., Jason, L., Brown, G. (Eds.), 2007. Geographies of Sexualities: Theories, Practices and Politics. Ashgate, Aldershot.
Browne, K., Nash, C., Hines, S., 2010. Introduction: towards trans geographies. Gender, Place and Culture 17, 573–577.
Butler, J., 1990. Gender Trouble. Routledge, New York.
Caluya, G., 2008. 'The rice steamer': race, desire and affect in Sydney's gay scene. Australian Geographer 39, 283–292.
Crenshaw, K., 1991. Mapping the margins. Stanford Law Review 43, 1241–1299.
Doan, P.L., 2010. The tyranny of gendered spaces – reflections from beyond the gender dichotomy. Gender, Place and Culture 17, 635–654.
Gibson-Graham, J.K., 1996. The End of Capitalism (As We Knew it): A Feminist Critique of Political Economy. Blackwell, Cambridge, MA. Republished in 2006 by Minnesota Press, Minneapolis, MN.
Gorman-Murray, A., 2008. Queering the family home: narratives from gay, lesbian and bisexual youth coming out in supportive family homes in Australia. Gender, Place and Culture 15, 31–44.
Holmes, C., 2009. Destabilizing homonormativity and the public/private dichotomy in North American lesbian domestic violence discourses. Gender, Place and Culture 16, 77–95.
Hubbard, P., 2008. Here, there, everywhere: the ubiquitous geographies of heteronormativity. Geography Compass 2 (3), 640–658.
Hubbard, P., 2011. Cities and Sexualities. Routledge, London.
Johnston, L., 2007. Mobilizing pride/shame: lesbians, tourism and parades. Social and Cultural Geography 8, 29–44.
Johnston, L., Longhurst, R., 2010. Space, Place and Sex: Geographies of Sexualities. Rowman and Littlefield, Lanham, MD.
Knopp, L., 2007. On the relationship between queer and feminist geographies. The Professional Geographer 59 (1), 47–55.
Lauria, M., Knopp, L., 1985. Toward an analysis of the role of gay communities in the urban resistance. Urban Geography 6 (2), 152–169.
Morrison, C.-A., 2012. Heterosexuality and home: intimacies of space and spaces of touch. Emotion, Space and Society 5 (1), 10–18.
Mulligan, A.N., 2008. Countering exclusion: the 'St. Pats for all' parade. Gender, Place and Culture 15, 153–167.
Nast, H., 1998. Unsexy geographies. Gender, Place and Culture 5, 191–206.
Nast, H., 2001. Sexuality and geography. In: Smelser, N., Baltes, P. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam, The Netherlands, pp. 14021–14026.
Olund, E., 2010. 'Disreputable life': race, sex, and intimacy. Society and Space 28, 142–157.
Oswin, N., 2004. Towards radical geographies of complicit queer futures. Acme 3, 79–86.
Peake, L., 2010. Gender, race, sexuality. In: Smith, S., Pain, R., Marston, S.A., Jones III, J.P. (Eds.), Handbook of Social Geographies. SAGE, London, pp. 55–76.
Podmore, J.A., 2006. Gone 'underground'? Lesbian visibility and the consolidation of queer space in Montréal. Social and Cultural Geography 7, 595–625.
Puar, J., 2006. Mapping US homonormativities. Gender, Place and Culture: A Journal of Feminist Geography 13, 67–88.
Rich, A., 1980. Compulsory heterosexuality and lesbian existence. Signs 5, 631–660.
Richardson, D., Poudel, M., Laurie, N., 2009. Sexual trafficking in Nepal: constructing citizenship and livelihoods. Gender, Place and Culture 16, 259–278.
Rose, G., 1993. Feminism and Geography: The Limits of Geographical Thought. University of Minnesota Press, Minneapolis, MN.
Rubin, G., 1993. Thinking sex: notes for a radical theory on the politics of sexuality. In: Abelove, H., Barale, M.A., Halperin, David M. (Eds.), The Lesbian and Gay Studies Reader. Routledge, London, pp. 3–44.
Taylor, Y., 2011. Sexualities and class. Sexualities 14, 3–11.
Thomas, M., 2011. Multicultural Girlhood: Racism, Sexuality, and the Conflicted Spaces of American Education. Temple University Press, Philadelphia, PA.
Valentine, G., 1993. (Hetero)sexing space: lesbian perceptions and experiences of everyday spaces. Environment and Planning D: Society and Space 11 (4), 395–413.
Valentine, G., 2001. Social Geographies. Prentice Hall, Harlow, UK.
Waitt, G., 2005. Sexual citizenship in Latvia: geographies of the Latvian closet. Social and Cultural Geography 6, 161–181.

Sexuality and the Internet

Aaron Ben-Ze'ev, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cyberspace offers an alternative romantic (and sexual) environment that provides people with easy access to many tempting options. It is a kind of huge, dynamic, electronic bedroom loaded with imaginative interactions. This novel environment has a significant impact upon offline romantic activities. This impact is expressed in increased opportunities, greater self-disclosure, decreased vulnerability, greater satisfaction, lesser commitment, increase in violations of boundaries, and reduced exclusivity. Cyberspace provides technical tools that facilitate the opportunity to conduct several romantic (and sexual) relationships at the same time. Although cyberlove and cybersex are likely to become more popular, they will not replace offline relationships, but rather complement them.

Cyberspace is a very popular new environment in which people pursue and conduct romantic and sexual activities (Goldberg et al., 2008). While in December 1995, only 0.4% of the world population used the Internet, in 2013, it has risen to about 40% (http://www.internetworldstats.com/emarketing.htm).

In this article, I will first briefly describe the nature of cyberlove and cybersex. Then I will examine how these online activities affect people's romantic and sexual behavior, values and attitudes. I will focus on the following major changes in the romantic (which includes the sexual) realm: Increased opportunities, greater self-disclosure, decreased vulnerability, greater satisfaction, lesser commitment, increase in violations of boundaries, and reduced exclusivity. All these changes make cyberspace a more exciting romantic environment, one which has far-reaching implications on sexual behavior, values and attitudes not merely within this environment, but also in offline circumstances.

Like the ordinary romantic environment, cyberspace is not homogenous and there is no 'one theory fits all' approach; the phenomenon of online relationships requires a more complex analysis (Whitty, 2013).

Cyberlove and Cybersex

Cyberlove and cybersex are common and attractive romantic and sexual activities (Ross, 2005). Cyberlove is a romantic relationship consisting mainly of computer-mediated communication. Despite the fact that the partner is physically remote and may be anonymous, the emotion of love is experienced as fully and as intensely as in an offline relationship.

In a broad sense, cybersex (or in slang, 'cybering'), refers to all types of sexually related activities offered in cyberspace. In this sense, the viewing of sexually explicit materials on the Internet is also a type of cybersex. In a narrower sense, which I will use here, cybersex refers to a social interaction between at least two people who are exchanging real-time digital messages in order to become sexually aroused. People send provocative and erotic messages, with the purpose of bringing each other to orgasm as they masturbate together in real time. These messages are typically sent via a private communication, such as an e-mail, instant message, or a chat room. The messages may be of various types – video, audio, and text-based. Although here I mainly refer to text-based cybersex, many of the claims are valid for other types of online communication.

In text-based cybersex people describe body characteristics to one another, verbalize sexual actions and reactions, and make believe that the virtual happenings are real. Cybersex requires the articulation of sexual desire to an extent that would be most unusual in face-to-face encounters. In cyberspace, that which often remains unspoken must be put into words (Doering, 2000). When people are involved in cybersex, they cannot actually kiss each other, but nevertheless the kiss they may send is emotionally vivid and its emotional impact is often similar to that of an actual kiss.

I argue that the active personal role in cyberspace makes this environment more exciting and seductive than that of daydreams, erotic novels, or X-rated movies; hence the temptation to engage in sexual activities is great. The personal imaginary interaction is very seductive. Since the line separating passive observation from full interaction has already been crossed in cybersex, it becomes easier to blur the line separating imagination from reality (Ben-Ze'ev, 2004; see also De Mul, 2010).

Participants in cyberlove take the reality of cyberspace seriously. Thus, people speak of their cybermates or even their online husbands or online wives. People have even got cyber-married and vowed to remain faithful to each other. Some women have claimed that they do not want to engage in cybersex with the first person who asks them, since they want to save their virtual virginity for the right man. Similarly, some say that they do not want to have a one-night cyberstand, but rather wish to have a more extended and meaningful online sexual affair (Levine, 1998: pp. 135–136). Nevertheless, in light of the dynamic and restless nature of cyberspace, such brief sexual activities are very common. The illusory nature of cyberspace does not diminish the need to resort to the same illusory methods used in offline circumstances, such as faking cyberorgasms.

I argue that the presence of interactive characteristics in the imaginary realm of online relationship is a tremendous revolution in personal relationships, as it enables people

to reap most of the benefits associated with offline relationships without investing significant resources (Ben-Ze'ev, 2004).

The interactive revolution in online romantic and sexual relationships has promoted both greater social interaction and more solitary activities. In comparison with standard fantasies, online relationships involve more social activities with other people. However, in comparison with offline relationships, many romantic activities are performed while someone is all alone sitting in front of a computer. Take, for example, cybersex. Compared with offline masturbation, cybersex (like phone sex) is a much more social interaction, as it is done while communicating with another person. While in offline masturbation, orgasm comes courtesy of the person's own hands and mind, in cybersex orgasm also comes courtesy of another person's mind. Cybersex narrows the gap between masturbation and offline sex, as it involves the active contribution of another person. However, compared with offline sexual relationships, cybersex is less social and it can in fact reduce the need for actual social interactions. Moreover, the virtual nature of cyberspace often fails to satisfy real needs. As a married woman wrote to her online lover, "I want a lover who actually touches me."

I turn now to discuss some of the major changes in the romantic and sexual realms that are due to the introduction of the Internet: increased opportunities, greater self-disclosure, decreased vulnerability, greater satisfaction, lesser commitment, increase in violations of boundaries, and reduced exclusivity.

Increasing Opportunities

Cyberspace is an alternative, available romantic and sexual environment that provides people with easy access to many available and desired options.

It is easy and inexpensive to contact desired and willing partners and to perform desired actions. It is easier to find attractive romantic and sexual partners in cyberspace than in bars, shopping malls, or supermarkets; moreover, it is easier to identify available and willing people.

Cyberspace is also highly available in the sense that it is immediately accessible for most people. The great accessibility and convenience of cyberspace make people feel comfortable about entering this space and staying there. One does not have to invest much in order to step into this imaginative, romantic and sexual paradise. Millions of people are eagerly waiting for everyone on the Internet every moment of the day. They are available and it is easy to find them. It must be remembered, however, that, as in offline life, most of those people will not be suitable. The great availability of online romantic and sexual partners enables frequent novel changes, and this makes cyberspace a more dynamic, volatile, and exciting romantic and sexual place. A kind of huge electronic bedroom, as someone remarked.

The many romantic and sexual opportunities in the current cyberspace society make this the best and worst of times for lovers. It is indeed both a happy and a difficult time for lovers – happy in that available, willing potential lovers are all around; difficult in that maintaining a loving committed relationship is harder than before as alternative romantic options are easier to explore and to realize.

Greater Self-Disclosure and Lesser Vulnerability

Two seemingly contrasting features of online relationships are: greater anonymity and greater self-disclosure. Anonymity is associated with concealment, which is contradictory to self-disclosure. However, since greater anonymity reduces vulnerability, it typically facilitates greater self-disclosure, which in turn increases familiarity and intimacy.

Research indicates that there is faster and more profound self-disclosure in online communication than in face-to-face meetings (Joinson, 2003; Tidwell and Walther, 2002). In online relationships people can be partially or fully anonymous: people can conceal their true identity or important aspects of it. Anonymity in online relationships facilitates self-disclosure as it reduces the risks involved in disclosing intimate information about oneself. People can express themselves more freely since they are more anonymous, less accountable, and hence less vulnerable. In the anonymity (or semianonymity) of cyberspace, it is much easier to disclose one's true feelings. Accordingly, it is more likely that in cyberspace the process of self-disclosure will be linear, moving in a unidirectional and cumulative fashion from nondisclosure to near full disclosure. In offline circumstances, the opposing urges to reveal and to conceal makes people oscillate between guarded self-concealment and candid self-disclosure.

Online self-disclosure resembles the 'strangers on a train' phenomenon, where people sometimes share intimate information with their anonymous seatmate (McKenna et al., 2002). Since anonymity in cyberspace is greater than on a train, revealing intimate personal details is more common in cyberspace. Online relationships enable people to hide behind a form of communication that is somewhat 'removed from life.' It is easier to open up to a faceless stranger that you do not have to look at while revealing your secret or to see the next morning. In online relations are similar in this regard: people can freely express their emotions and become emotionally close without being vulnerable. Accordingly, it is also easier to fall in love on the Net.

Despite the reduced vulnerability in cyberspace, the online agent can be hurt as well. In this regard, two major aspects are significant: (1) most of the many high hopes that cyberspace generates are not fulfilled – thereby causing frequent and profound disappointments; and (2) profound self-disclosure leaves the agent's mind naked, without any masks to protect her – and this is a highly vulnerable position for anyone.

Online relationships typically have fewer practical implications than have offline relationships; hence, participants in these relationships are less vulnerable. Indeed, in offline relationships, people tend not to reveal much intimate information until they feel safe. In cyberspace, people are ready to disclose more intimate information since they assume that anonymity and spatial distance reduce the risk of harmful consequences. Sex in cyberspace is safe not only in the physical sense, since viruses cannot travel in that space, but also in the psychological sense that is provided by being anonymous.

Since feeling safe is a major precondition for sexual arousal, the safety provided by cyberspace may explain the vast interest in cybersex (Bader, 2002: pp. 263–276).

Greater Satisfaction and Lesser Commitment

Imagination, which plays a major role in cyberspace, makes this space very seductive since it can free people from the limits imposed by their bodies and contexts (Dreyfus, 2001). Imagination has hardly any practical constraints, and it enables one to depict oneself and the other in a most positive and seductive light – much better than in reality. Online imagination can also complete, in an exciting manner, details upon which our online companions might remain vague. Imagination can therefore entertain our wildest fantasies. The imaginary nature of cyberspace makes it easier to idealize the other; and idealization is an essential element in romantic love and sex (Allbright, 2007; Ben-Ze'ev, 2004).

Emotional imagination can be compared to receiving a salary without earning it. Emotional imagination enables us to acquire various affective benefits, such as feeling better, without carrying out the relevant tasks required for obtaining such benefits (Ben-Ze'ev, 2000). For example, in sexual fantasies one may simulate the pleasure of intercourse with an attractive person without having to build an actual relationship and carrying the burden associated with this. It is obvious that there is no free lunch and we have to pay a price for this unearned emotional salary. The cost will be high if people repeatedly use tactics such as gaining emotional benefits from imaginary relationships; people may, for example, increase the risk of distorting reality to the extent that coping with it becomes not easier, but harder.

Cyberspace is a highly imaginative and active space. Romantic imagination also plays a significant part in offline circumstances, but in cyberspace its role is considerably greater; 'virtual reality' can actually augment fantasy. The many attractive online alternatives in cyberspace, coupled with the imaginative freedom that cyberspace offers, might make a current offline relationship less valued and offline partners less attractive.

Increasing Boundaries Violations

Drawing boundaries is essential for human communities: living with other people necessitates limiting our desires. However, globalization, in which cyberspace is a central arena of action, is essentially an act of crossing, cracking and breaking boundaries. Openness, which is an important aspect of globalization, implies boundary deregulation. Thus, since the activities in cyberspace cut across territorial borders, the feasibility and legitimacy of laws based on geographic boundaries are undermined (Almog, 2002).

Drawing romantic boundaries is particularly difficult since many people perceive love as uncompromising ('There are no mountains high enough, to keep me from getting to you'), and unconditional ('Love will always prevail'). People in love occasionally feel chained by normative boundaries preventing them from acting in accordance with their wilder passions. When in love, people want to 'fly on the wings of love,' where the earthly boundaries do not exist. Cyberspace allows people to crack and violate many normative boundaries with considerably reduced risk (Ben-Ze'ev and Goussinsky, 2008).

To decrease the boundaries violation, people may want to make their normative boundaries more flexible, yet this in turn may weaken the safeguards against further violation. Take, for example, cybersex, where romantic and sexual boundaries are much more flexible than in offline circumstances. This flexibility has not reduced the number of offline violations of boundaries but rather increased it. With the expanded use of the Internet, romantic and sexual cheating has increased. Moreover, even if the sexual cheating is limited to online, partners can lose trust in their loved ones and identify themselves as victims of trauma (Schneider et al., 2012).

The flexible nature of boundaries in cyberspace is not necessarily immoral. On the contrary, living by adhering strictly to strict boundaries may be immoral, as it does not take into account the unique, specific, personal and circumstantial aspects of the lover. In this regard, Stephen Toulmin (1981) argues that "A morality based entirely on general rules and principles is tyrannical and disproportionate … only those who make equitable allowances for subtle individual differences have a proper feeling for the deeper demands of ethics."

Reducing Exclusivity

Romantic, and especially sexual, exclusivity has been regarded as the hallmark of romantic relationships. The current social trend leans toward reducing exclusivity in relationships, and this trend is being reinforced by behavior in cyberspace, where romantic boundaries are very flexible. Reducing romantic exclusivity conflicts with partiality, which is one of the basic emotional characteristics and which enables the agent to focus his resources on one or very few objects. This reduction also goes against the heart of ideal love, i.e., the perception that the beloved is the one and only person suitable for the lover's profound love. On the other hand, such a reduction in exclusivity enhances a different characteristic, that of the need for change and novelty, which generates emotions.

The technology associated with online relationships makes it easier, more convenient, and safer to reduce exclusivity. Modern technology continues to improve the methods available for distant relationships. It does not merely provide more options of meeting willing people, but it also offers a more comfortable way to pursue several romantic relationships at the same time.

The practice of whetting your appetite online while eating at home is a significant element in the process of violating marriage's monopoly on sex for married people. As cybersex is seen as a lesser sin – since it can be considered merely a process of talking that involves no actual physical encounter – some offline partners will tolerate or even support it. Letting your partner know about, and even watch, your sexual activity with another person is significant in the sense that the committed couple knowingly accepts that sexual exclusivity is not a one-degree category that should never be violated. Sexual

exclusivity is thus seen as a continuum, and, in some circumstances, certain points along that continuum may be violated.

As it is difficult to entirely avoid the huge amount of available tempting alternatives that cyberspace provides, the attitude toward romantic affairs is likely to change. As online infidelity and adultery radically increase, it may cause a qualitative change in the agreed meanings of 'infidelity' and 'adultery.' The criteria for inclusion within these categories will become more limited. Thus, harsh descriptions such as 'betrayal' are beginning to fade and softer words, such as 'a parallel relationship,' are becoming more popular. Nevertheless, it is highly unlikely that the chaotic and dynamic nature of cyberspace will completely replace the more stable nature of actual space, as people cannot thrive in complete chaos.

Cyberspace enables participants to conduct several online affairs while still maintaining the primary offline relationship. People may have an online relationship that is different in nature from their offline relationship, and hence the two may not be in conflict – at least not significantly enough to force the termination of one of them. If the offline relationship is more pragmatic, with only a dash of romance, the online affair may be full of romance and with little or no practical implications. One relationship may be mainly sexual, while in the other the sexual aspect may be marginal. Since the two types of relationship fulfill different needs, their coexistence may be beneficial.

In order to reduce the risk of ruining their primary offline relationship, some married people may accept their partner having an online sexual affair, but not an online romantic affair. Others may further limit the sexual affair to a one-night cyberstand. All such limitations are intended to minimize the harm done to the primary relationship. It seems, however, that a more substantial change in our emotional makeup is required for coping with these sexual opportunities while still maintaining some stability in the primary relationship.

The Internet presents a serious threat to monogamous relationships in general, and marriage in particular, since it facilitates not merely pleasurable sexual activities but deep romantic relationships as well. A one-night cyberstand is more available and easier to keep secret, and people are not oblivious to this advantage. On the other hand, the conditions for nourishing a deep loving relationship have also been improved. The Internet provides a most enjoyable and efficient means by which various people get to know each other intimately without the distractions of external factors, such as appearance, age, geographical distance, race, nationality, religion, or marital status. This will increase the number of international, intercultural, and interreligious marriages, ultimately modifying global social norms – in the main, making them more flexible (Barak and Fisher, 2002).

The Internet is not unique in providing opportunities for meaningful romances and superficial sexual affairs; its main novelty is that it offers an environment in which – given a wise and moderate use of cyberspace – both types of relationship can coexist. Indeed, some findings suggest that Internet users who engage in online chats are usually people who are satisfied with their offline social relations and for them, socializing in online chat rooms does not substitute for offline emotional or social needs; rather, it complements their offline relations (Maheu and Subotnik, 2001; Peris et al., 2002).

The Internet encourages other types of changes in romantic relationships. Thus, the prominence of verbal communication in online affairs is likely to increase the value placed on intellectual abilities in romantic interactions. Accordingly, online romances are likely to depend more on the combination of emotions and intelligence, including verbal and intellectual features such as wit, a sense of humor, and good articulation. The value of personal characteristics – as compared to the value of external appearance – will probably increase. The change may have rather, less impact once most computers are equipped with a Webcam, but it will be evident even then as long as distant online relationships continue to be based upon written text.

The Alternative Romantic Environment

Cyberspace provides an alternative environment to the actual one. People do not live exclusively in one environment; rather, they move from one environment to the other. Cyberspace enables participants to explore exciting romantic alternatives without necessarily violating significant personal commitments. Cyberspace provides an outlet for developing alternative emotional ties, without completely ruining the primary offline relationship. When people confuse cyberspace with the actual world, the issue of commitment becomes problematic and emotional and moral difficulties emerge.

The great seductiveness of cyberspace and the effortlessness of becoming involved in online affairs also entail risks: People are easily carried away and the risk of addiction is high. Moreover, the gratification from cybersex is often higher than from offline sexual contacts and it is generated not merely in cases of unsatisfying offline sexual activities (Laier et al., 2013). Like other types of addiction, cyberspace does not merely satisfy needs, but creates new needs that often cannot be met. This may lead to a lower probability of being satisfied with one's romantic lot.

The Internet encourages various types of other changes in romantic relationships. Thus, the prominence of verbal communication in online affairs increases the value placed on intellectual abilities in romantic interactions. Indeed, the weight of conversation – which is essentially an intellectual activity – is far greater in online relationships than in offline relationships. Accordingly, online romances are likely to depend more on a combination of emotions and intelligence, including verbal and intellectual features such as wit, a sense of humor, and good articulation. The value of personal characteristics, as opposed to the value of external appearances, will probably increase. This change may have rather less impact once more conversations are conducted with a Webcam. Still, it is likely that written communication will remain central for many years.

It is highly unlikely that, in the future, either offline or online romantic relationships will cease to play an important role in people's lives. The increased lure of the Net lowers the likelihood that those with access to the Net will restrict themselves solely to offline relationships. However, since online relationships lack some basic romantic activities, such as touching and actual sex, satisfying offline relationships will continue to be considered to be an upgraded and more

fulfilling relationship. Learning to integrate cyberspace and cybersex with actual-space in the romantic domain will be a major task for our society (Wysocki, 2012). Indeed, initial indications suggest that more than one-third of marriages in America now begin online. In addition, marriages that began online, when compared with those that began through traditional offline venues, were slightly less likely to result in a marital break-up (separation or divorce) and were associated with slightly higher marital satisfaction among those respondents who remained married. This suggests that the Internet may be altering the dynamics and outcomes of marriage itself (Cacioppo et al., 2013).

Today, dreams are no longer the major tool for imagining a better situation. Cyberspace has become more efficient for such a purpose. Cyberspace allows two lovers to feel as if they are directly connected – as if their bodies do not interfere, allowing their hearts to be in direct communication; they feel as if their souls are melting. People often describe their online relationship as 'dreaming while awake'. People delight in these dreams. However, a life of mere dreams and sleep are dangerous because they are not connected to reality. Online romantic relationships are valuable when they complement, rather than replace offline relationships. The value of dreams, as well as that of Cyberspace, is that they are interspersed with reality.

Social networks, such as Facebook, have become a central place for initiating and maintaining new romantic (including sexual) relationships. The new technological means of communication have shaped romantic relationships. People are all potentially in connection with each other, which provide further depth to their online relationships. The possibility of having so many friends further undermines the value and feasibility of romantic exclusivity.

Online relationships usually have the characteristic of 'unfinished business' since as long as they are not transformed into offline relationships, there is something missing from them. In this sense, they are similar to an extended period of courtship, and accordingly emotional intensity remains high even for a long period. A paradoxical aspect in this regard is that although online relationships are romantically and sexually intense because of, among other factors, their incompleteness, such incompleteness involves the wish to transform the relationship into a more complete one – something that usually decreases the intensity and may lead to the termination of the relationship.

Future changes will inevitably modify present social forms such as marriage and cohabitation, and current romantic practices relating to courtship, casual sex, committed romantic relationships, and romantic exclusivity. In light of what we have learned from recent history, we can expect the further relaxation of social and moral norms; this process should not be considered a threat, as it is not online changes that endanger romantic relationships, but our inability to adapt. Catherin Hakim (2012) argues that "If the pill made pre-marital sex among young people a lot easier, the internet facilitates playfairs among older married people."

Cyberspace has a significant role in maintaining intimate relationships between partners who live apart and meet only intermittently. This is also true of those marriages in which couples are compelled to live apart for some time, usually because they have jobs in different places. Commuter marriage is a rapidly growing form of distant relationship. Such relationships also exist among dating couples and in extramarital affairs; in these groups, cybersex, conducted through either regular or video chats, is very popular and enables the partners to maintain a certain level of intimacy.

Online affairs may make it easier to deal with the problem of regret, as it offers lower-risk affairs. Such affairs provide the excitement associated with romantic and sexual affairs, hence avoiding romantic dissatisfaction in the short term and regret in the long term. The lower degree of risk and moral criticism associated with such affairs also enables participants to avoid short-term regret and actual harm. Avoiding romantically exciting affairs may succeed in preventing short-term regret about the harmful consequences of our actions. As time goes by, however, regret over inaction is stronger, and avoiding pleasant actions becomes more and more difficult.

In a way, cyberspace forces people to face reality – a reality that often includes one's actual, disappointing romantic relationship. One cannot avoid unpleasant reality by leading oneself to believe that better alternatives are not available. Cyberspace takes an intermediate position between reality and sheer fantasy. The escape into cyberspace is not essentially an escape into fantasy; it is often an escape into a semireality that presents our everyday life in a more realistic manner. Sometimes, offline relationships can benefit from such an arrangement. Indeed, people testify that their online affair has been a beneficial, learning experience and has helped them to recognize many positive aspects of their offline relationship.

One way to prevent long-term regret is to encourage people to act on their impulses more often; people should focus less on the short-term consequences of their action and more often 'just do it.' However, this advice could be dangerous as the status quo is usually safer than trying something new. Awarding greater significance to unwise actions might promote risky behavior and decrease chances of survival. Cyberspace provides a way to cope with this dilemma: it lets people act on their impulses, hence reducing long-term regret, and it significantly reduces the risks involved in such behavior, thus reducing the chance of short-term regret.

What Is Next?

An endless search in cyberspace for a better alternative brings frustration and misery rather than happiness. Settling for one's lot, which is an important way to achieve profound happiness, is even more imperative in the cyberspace era. However, if you compromise and accept something much less than your dream demands, it can ruin your happiness as well. We need to find the optimal balance between trying to achieve our dreams and being able to be satisfied with our lot. Current society requires people to make more compromises, since we are aware of more promising alternatives (as compared with our own current situation); however, it is more difficult to make those compromises, since the alternatives are easier to achieve.

Future lovers may have to make more compromises and their frustration in not achieving the imaginary perfect partner might increase. But those lovers have also a greater chance of being satisfied with their romantic relationships in a way that

will not cause them to feel that they are making significant compromises.

Cyberspace poses specific difficulties in finding a romantic framework that will suit all lovers. Nevertheless, love itself is flourishing as a result, among other things, of the use of the Internet. It is more possible now than ever before to find the one you have been waiting for all your life; it is more difficult now to stay with this person for ever.

The test that future society will face is whether it can complement ordinary romantic activities, just as the telephone complements ordinary social activities, or whether they will merely replace these activities with less valuable ones, as television has frequently done.

See also: Culture and Emotion; Emotions, Sociology of; Internet and Privacy; Internet and Social Media: Anthropological Aspects; Oversharing: The Eclipse of Privacy in the Internet Age; Sexual Behavior and Social Networks; Sexual Motivation; Symbolic Boundaries; Wisdom, Psychology of.

Bibliography

Allbright, J.M., 2007. How do I love thee and thee and thee: self-presentation, deception and multiple relationships online. In: Whitty, M.T., Baker, A.J., Inman, J.A. (Eds.), Online Matchmaking. Palgrave Macmillan, London, pp. 81–93.

Almog, S., 2002. From Sterne and Borges to lost storytellers: cyberspace, narrative, and law. Fordham Intellectual Property, Media and Entertainment Law Journal 13, 1–34.

Bader, M.J., 2002. Arousal: The Secret Logic of Sexual Fantasies. St. Martin's Press, New York.

Barak, A., Fisher, W.A., 2002. The future of Internet sexuality. In: Cooper, A. (Ed.), Sex and the Internet: A Guidebook for Clinicians. Brunner-Routledge, New York, pp. 263–280.

Ben-Ze'ev, A., 2000. The Subtlety of Emotions. MIT Press, Cambridge, MA.

Ben-Ze'ev, A., 2004. Love Online: Emotions on the Internet. Cambridge University Press, Cambridge.

Ben-Ze'ev, A., Goussinsky, R., 2008. In the Name of Love: Romantic Ideology and Its Victims. Oxford University Press, Oxford.

Cacioppo, J.T., Cacioppo, S., Gonzaga, G.C., Ogburn, E.L., VanderWeele, T.J., 2013. Marital satisfaction and break-ups differ across on-line and off-line meeting venues. Proceedings of the National Academy of Sciences of the United States of America 110, 1–6.

De Mul, J., 2010. Cyberspace Odyssey: Towards a Virtual Ontology and Anthropology. Cambridge Scholars Publishing, Cambridge.

Doering, N., 2000. Feminist views of cybersex: victimization, liberation, and empowerment. CyberPsychology and Behavior 3, 863–884.

Dreyfus, H.I., 2001. On the Internet. Routledge, London.

Goldberg, P.D., Peterson, B.D., Rosen, K.H., Sara, M.L., 2008. Cybersex: the impact of a contemporary problem on the practices of marriage and family therapists. Journal of Marital and Family Therapy 34, 469–480.

Hakim, C., 2012. The New Rules: Internet Dating, Playfairs and Erotic Power. Gibson Square, London.

Joinson, A.N., 2003. Understanding the Psychology of Internet Behavior: Virtual Worlds, Real Lives. Palgrave Macmillan, Hampshire.

Laier, C., Pawlikowski, M., Pekal, J., Schulte, F.P., Brand, M., 2013. Cybersex addiction: experienced sexual arousal when watching pornography and not real-life sexual contacts makes the difference. Journal of Behavioral Addictions 2 (2), 100–107.

Levine, D., 1998. The Joy of Cybersex: A Guide for Creative Lovers. Ballantine Books, New York.

Maheu, M.M., Subotnik, R.B., 2001. Infidelity on the Internet: Virtual Relationships and Real Betrayal. Sourcebooks, Naperville, IL.

McKenna, K.Y.A., Green, A.S., Gleason, M.E.J., 2002. Relationship formation on the Internet: what's the big attraction? Journal of Social Issues 58, 9–32.

Peris, R., Gimeno, M.A., Pinazo, D., Ortet, G., Carrero, V., Sanchiz, M., Ibanez, I., 2002. Online chat rooms: virtual spaces of interaction for socially oriented people. CyberPsychology and Behavior 5, 43–51.

Ross, M.W., 2005. Typing, doing, and being: sexuality and the internet. Journal of Sex Research 42, 342–352.

Schneider, J.P., Weiss, R., Samenow, C., 2012. Is it really cheating? Understanding the emotional reactions and clinical treatment of spouses and partners affected by cybersex infidelity. Sexual Addiction and Compulsivity 19, 123–139.

Tidwell, L.C., Walther, J.B., 2002. Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: getting to know one another a bit at a time. Human Communication Research 28, 317–348.

Toulmin, S., 1981. The tyranny of principles. The Hastings Center Report 11, 31–39.

Whitty, M.T., 2013. Online romantic relationships. In: Amichai-Hamburger, Y. (Ed.), The Social Net. Oxford University Press, Oxford, pp. 62–78.

Wysocki, D.K., 2012. Cybersex, computer sex addiction, and cyberpornography. In: Bryabt, D. (Ed.), Routledge Handbood of Deviant Behavior. Routledge, London, pp. 272–276.

Sexuality: Cultural Aspects

Daniel Jordan Smith, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines cultural aspects of sexuality and reviews the contributions of social science research to this field of knowledge. Two features are emphasized: (1) the social construction of sexuality, which produces immense cross-cultural diversity and (2) the extent to which societal preoccupations with sexuality are often about much more than sex itself, providing fertile data for social analysis. The article is organized into six substantive sections representing significant domains of scholarship in this area. These sections focus on premarital sex, marriage, extramarital sex, gender and sexuality, non-heterosexuality, and the global AIDS epidemic.

Introduction

Sexuality is both an intimate and very public domain of human social life, essential to biological and social reproduction. Two dimensions of the relationship between culture and sexuality are fundamental. First, despite its basis in human biology, sexuality is culturally constructed in powerful ways. Sexual identities and orientations, the meanings of sexual behavior, and the social consequences of these behaviors vary tremendously across cultural contexts. Second, in every society, the cultural meanings of sexuality are about much more than sex. In order to understand the reasons why people adopt various sexual identities and why they engage in sexual behavior in the ways that they do, it is necessary to situate sexuality in a broader social context. Societal norms and practices around sexuality, along with popular interpretations and political responses to them, are often implicit means by which people debate, contest, or simply give expression to many other issues besides sexuality. Studying the cultural aspects of sexuality is not just a way to understand sexuality better, but it can also be a revealing window onto culture and society more generally.

The social sciences have contributed immensely to the realization that sexuality is culturally constructed. This notion builds on the broader conception of the social construction of reality, but knowledge about the cultural underpinnings of sexuality draws, in particular, on insights from feministic scholarship, from comparative and cross-cultural studies of human sexuality – particularly in anthropology (Rosaldo et al., 1974; Davis and Whitten, 1987) – and more recently, from queer studies. In criticizing patriarchal social structures, exposing a male-privileged division of labor in society, and challenging accepted 'knowledge' about men and women in both popular culture and science, feminist scholars pointed to the centrality of control over women's bodies and their sexuality as a vehicle for reproducing gender inequality (Ortner and Whitehead, 1981). Essential to the feminist critique was the understanding that both gender and sexuality were in large measure human-made, socially constructed categories. The implication is that sexuality is not simply biological destiny, but a domain of political and social action, where shared cultural meanings have significant effects.

If feminist critiques of gender inequality provided one foundation for a broad understanding of the cultural construction of sexuality, anthropological studies of sexuality across a vast array of societies provided another enormous body of evidence showing that sexuality is highly variable across cultural contexts. The wide diversity of identities, beliefs, and practices around sexuality catalogued and analyzed by anthropologists and other cross-cultural researchers has demonstrated conclusively that sexuality is a deeply social domain of human life. The challenge in the social sciences is no longer to prove sexuality's culturally constructed aspects, but to investigate what these are, explain what accounts for differences across societies, and evaluate the consequences – across a huge range of contexts.

Whether in history or in the contemporary world, human communities commonly focus on – in many cases, one might even argue, obsess – over things related to sexuality in ways that are not adequately explained by the stakes associated with sexuality itself. As indicated above, feminist scholars have demonstrated incisively how societal preoccupations with women's bodies and sexuality often serve to express, enact, and enforce a much broader politics of gender (Fausto-Sterling, 2000). But societal obsessions about sexuality are not only about women. For example, in many places, public controversies over homosexuality or non-heterosexuality focus more on male sexuality than female sexuality, and these concerns have been interpreted by social scientists as enacting and enforcing societal norms about masculinity (Connell, 1987; Gutmann, 1996). But popular and political concerns about sexuality are often about much more than gender. Societal preoccupations with sexuality – whether they take the form of local rumors, frenetic media coverage, or legislative debates over proposed laws related to sexuality – are often a means by which much wider concerns about social life are expressed. People – indeed communities and whole nations – frequently conduct much bigger social and political debates in the idiom of discourses about sexuality. For example, heated debates about gay marriage in the U.S., about rape in India, and about homosexuality in Uganda, to name just three recent cases, while important issues in their own right, are also vehicles by which these societies express and contest even larger issues.

The cultural landscape of human sexuality is, of course, vast. This article briefly summarizes some of the cultural aspects of sexuality in relation to six topics: premarital sex, marriage, extramarital sex, gender, non-heterosexuality (or the commonly used western acronym of LGBTQ sexuality – lesbian, gay, bisexual, transgender, and queer), and the AIDS

epidemic. In each section the themes of the cultural construction of sexuality and the wider meanings of discourses about sexuality are employed to organize and limit what is said about otherwise huge topics in and of themselves. The section on the AIDS epidemic concludes the article both because it is arguably the newest major arena of sexuality research and because it exemplifies so clearly how societal preoccupations with sexuality involve stakes and have consequences much wider than sexuality itself.

Premarital Sex

Across most societies, if not all, marriage (see below) is the socially prescribed sphere in which sexual relations are expected and approved of, above all, because of its function as the means for biological and social reproduction. And yet at the same time, the institution of marriage and the social rules that accompany it have also created the very possibility of what is widely known as 'premarital sex.' Cultural preoccupations with premarital sex in the contemporary world are driven, at least in part, by a modern demographic reality: across nearly all societies people are marrying at later ages, creating a much longer window of opportunity for premarital sex (Westoff, 2003). Until relatively recently in human history the gap between puberty and marriage was short. The value placed on formal schooling and higher education for both men and women around the world, and other accompanying social changes, have expanded the premarital period and stimulated cultural responses that are complex and often contradictory. In many societies, premarital sexual activity is seen by young people as a marker of modern identity and simultaneously condemned (especially by older generations) as an emblem of problematic social change, connoting failures of collective morality, and gender and generational stability under siege (Smith, 2000).

On the one hand, studies in many different cultural contexts have shown that young people feel being sexually active before marriage is part of becoming an adult; but much more than that, premarital sex is a way that young people represent themselves as educated, modern, urban, and in tune with contemporary fashions. Even in contexts where premarital sex comes with both health risks – such as unwanted pregnancy or HIV infection (see more below) – and social risks – such as sanctions imposed by parents, kin, and even some peers – evidence suggests that the cultural appeal of premarital sex looms large in young people's sexual decision-making and behavior. While social contexts vary dramatically, the combination of later age at marriage and the widespread (arguably global) association of premarital sex with modern youth identities has meant that young people around the world are increasingly engaged in premarital sex and navigating its positive and negative cultural consequences.

While young people in many societies seem drawn to the social rewards that are conferred by being sexually active before marriage, they also sometimes face significant cultural backlash, mainly because premarital sexuality (especially female premarital sexuality) is often seen as a harbinger, symbol, or consequence of a much wider array of social changes about which people are often anxious, including rural-to-urban migration, increases in female wage labor, growing levels of educational attainment, and the proliferation of global media (Utomo and McDonald, 2009). Collective cultural responses to changes that threaten generational hierarchies, challenge established gendered divisions of labor, or create other new inequalities are often voiced, in part, through discontents about young people's behavior, particularly their sexual behavior. While social anxiety about sexual behavior is arguably a timeless human preoccupation, given sexuality's elemental role in social reproduction, widespread contemporary concerns about premarital sex reflect specific concerns about social reproduction in a world of urbanization, growing inequalities associated with the expanding reach of capitalism, a changing gendered division of labor, and a long duration between puberty and marriage.

Research on the cultural aspects of premarital sex is often closely tied to the (real or imagined) implications of young people's sexual behavior for their health, whether because of unwanted pregnancy, sexually transmitted infection, or HIV and AIDS (see more below). The literature on premarital sexual behavior documents a wide variety of practices associated with specific cultural contexts (Widmer et al., 1998). Some of this scholarship shows very particular health consequences related to specific cultural practices (e.g., initiation rites that include circumcision). But across many different settings, it is the common tension between cultural values that encourage young people to engage in premarital sex and social mores that prohibit or punish it that influence many of the scenarios that lead to negative health outcomes.

Marriage

As noted already, in nearly every cultural context marriage is the most positively sanctioned domain in which sexual relations are expected to occur. Although sex in marriage is universally expected across cultures, the rules are culturally specific. Given that human societies are deeply invested in the cultural rules about sexuality because sex is elemental to both biological and social reproduction, sexuality in marriage is especially important because it is the institution through which reproduction is most intensely regulated. Classic anthropological theories about the function of marriage in human social organization, whether they emphasize alliance or descent, point to its central role in social reproduction. Throughout much of human history, in many societies, sexuality, marriage, and reproduction were (and to an extent remain) inextricably connected.

Even in contexts where procreation is seen as the primary purpose of sex in marriage, cultural norms and understandings of proper sexual conduct vary tremendously (Thornton and Fricke, 1987). In polygynous societies, for example, a man's sexual access to multiple wives is commonly tightly prescribed. In many contexts, the importance of pleasure and its expression in marital sex, including women's pleasure, is highly valued, while elsewhere it is considered inappropriate to emphasize or overtly express physical enjoyment. In some settings all sex, including marital sex, is considered potentially polluting and dangerous (Wardlow, 2006).

The perceived relationship between sex and marriage has changed considerably in recent decades. The idea that love and romance should be a basis for marriage has spread globally,

one factor in precipitating a decrease in arranged marriages (Hirsch, 2003; Coontz, 2006). Although cultural differences are still powerful – and push these new forms of marriage in locally inflected directions – in general, the rise of what has been called 'modern marriage,' 'choice marriage,' 'love marriage,' 'companionate marriage,' or 'romantic marriage' has meant that sexual intimacy, communication, and mutual pleasure in marriage have become more explicitly valued. These global trends seem to elevate the importance of sex itself – apart from its reproductive function – in contemporary marriages worldwide. This pattern coincides with, and indeed may be partly explained by, demographic transition, in which parents are having few children. The relationship between spouses has become more central in the entire endeavor of family making. That being said, rising rates of divorce may also accompany the increasing emphasis on romantic intimacy as a basis for marriage, for when feelings of love subside or transform with time, the basis of the marriage may seem to be weakened as well.

Finally, it is worth noting that societal debates about marriage – which seem to animate religious discourse, popular political discussions, and so-called 'culture wars' in many societies – can be constructively interpreted as conversations and contests over the relationship between sexuality and social reproduction. Whether it is the concern over rates of divorce, gay marriage, teenage pregnancy, forced marriage, or 'sugar daddies' (wealthy, often married men having sex with economically dependent single young women), when societies are collectively preoccupied about issues related to sex and marriage, they are typically voicing concerns about much wider issues of social reproduction – about the kind of values, practices, people, and institutions they will have in the future.

Extramarital Sex

If the fact and social importance of marriage create and shape the possibility and cultural meanings of premarital sex, the same is true of extramarital sex. Perhaps the most widespread characteristic of extramarital sex across cultures is that it is simultaneously proscribed and commonly occurring. Many explanations for the prevalence of extramarital sex – especially men's extramarital sex, which is believed in most societies to be more prevalent – either explicitly or implicitly assume a biological basis. Humans – or at least men – are often presumed to be wired in ways that make monogamy and fidelity difficult. Interestingly, these naturalizing narratives prevail in many societies, even when they are not connected directly to science (or explanations meant to pass for science).

But culturally attuned research on extramarital sexuality has shown that such behavior can be explained in social terms: it is socially organized and in various ways culturally rewarded; it has social functions and meanings that cannot accurately be reduced to natural instincts or drives; and patterns of extramarital sex change over time and vary significantly across contexts. For example, recent comparative research on men's extramarital sexuality has found the social factors like migration for employment (often separating husbands and wives for long periods over significant distances), the growing importance of conspicuous consumption as a marker of class, as also men's participation in same-sex peer group activities, all serve to increase the likelihood of men's extramarital sexual relationships (Hirsch et al., 2009). In short, men's infidelity is not simply biological destiny, but a culturally constructed arena of human activity.

Less research has studied women's extramarital sexual behavior. Arguably this is because women's extramarital sexuality is more stigmatized, and therefore more hidden (and perhaps also less common), and not necessarily because women are biologically different, or at least certainly not solely because they are biologically different. Rates of extramarital sex by gender have not been empirically verified across most cultures. Indeed, the difficulty of such research attests precisely to the culturally charged quality of sexuality. To the extent that research has explored women's extramarital sexual behavior, a diversity of cultural values and patterns of behavior across societies is apparent (Wood and Eagly, 2002).

Like premarital sex, extramarital sex can become the object of social scrutiny, a problem that people in many different societies will point to as emblematic of (or sometimes even a causal factor in) a wider range of social problems. The fact that in many cultures women's participation in premarital and extramarital sex draws more collective ire than men's is a result of the persistence of gender inequality and the role of cultural mores about sexuality in creating and maintaining gender double standards more generally.

Gender and Sexuality

Now largely taken for granted, one of the fundamental contributions of feminist scholarship was to demonstrate that sex and gender are not the same thing (Fausto-Sterling, 2000; Rubin, 2004). Rather than biological sex-determining gendered identities and roles, gender was revealed to be a highly variable, culturally constructed domain, deeply influenced by and consequential for social life. But even as feminist scholarship and a rich body of literature in gender studies more broadly have examined gender as a socially constructed category, sexuality remains deeply connected to gender and connects biological sex and gender in powerful ways. Although biological sex does not mechanistically determine gender, sex and sexuality are nonetheless important components of the social construction of gender.

As such, though gender and sexuality are distinct concepts and different dimensions of human social life, they are often inextricably intertwined. If it can be said that most, if not all, human societies are preoccupied – often even obsessed – with sexuality as an arena of social discourse and moral regulation, it can also be usefully hypothesized that a good deal of the fixation on sexuality is about defining, defending, contesting, and reproducing a gendered division of labor that is fundamental to social organization, and to the production and reproduction of various forms of hierarchy and inequality (Connell, 1987). Social and moral concerns about sexuality are a foundation for cultural constructions of gender differences and the social and political consequences that follow.

One of the predominant findings in studies of the cultural aspects of sexuality as it relates to gender is that patriarchal social organization in many ways depends upon shaping and

controlling women's participation in public and political arenas through regulation of their sexuality (Kandiyoti, 1988; Inhorn, 1996). Patterns of gender inequality maintained through the regulation of women's sexuality can be overt and explicit, such as through laws or religious edicts, or – and perhaps more commonly – they can be informal, tacit, and so taken for granted that they often go unnoticed as forms of social control. When popular concerns about women's sexual morality and behavior become objects of popular discourse and vehicles for the expression of social discontent, as happens so frequently across the globe, they are often proxies for wider social and political concerns. Women's sexual bodies and behavior come to stand for larger tensions and contests regarding social reproduction. Many forms of social inequality are connected to, and arguably depend upon, maintaining aspects of gender inequality.

While feminist critiques and a focus on women's oppression comprise major strands of social science research on the cultural aspects of sexuality as they relate to gender, over the last couple of decades a growing body of work has focused on the social construction of masculinity (Connell, 1995; Gutmann, 1997). Ideas and practices around manhood are equally implicated in wider social and political issues. Much of this work suggests that masculinity is more fragile than it is stereotypically conceived of. Further, men's sexual bodies and behavior are objects and arenas wherein not only masculinity itself is made and expressed, but also – as with women's sexuality – domains where much of the work of social reproduction is conducted. To the extent that social life – including fundamental features of social inequality – depend on ideas about gender and a gendered division of labor in society, and to the extent that the cultural construction of gender mobilizes ideas about sexuality for its power, men's sexuality is as implicated as women's sexuality, although as the dominant and often unmarked category it is less studied.

Non-Heterosexuality

Culturally oriented research on what is referred to in many western societies as LGBTQ (lesbian, gay, bisexual, transgender, and queer) sexuality has demonstrated that it is instructive to analyze social ideas and practices related to non-heterosexual orientations as being situated in a culturally constructed domain in which ideas and practices vary tremendously across social contexts. Although the preferred terms vary for speaking about sex that is not between 'straight' partners, this article uses the concept of 'non-heterosexuality' to encompass the range of orientations that are glossed in Europe and North America as LGBTQ. Like other socially marked sexual identities and behaviors, part of the reason non-heterosexuality receives so much popular attention across societies is because public discourse about it frequently stands for larger anxieties around gender politics, the role of families in reproduction, forms of economic privilege associated with marriage, and much more (Weston, 1993). Scholarship on the cultural aspects of non-heterosexuality has demonstrated not only huge cross-cultural variation, but also a common pattern in which highly charged public debates and widespread homophobia signal much wider worries about society.

One of the signature contributions of cross-cultural social science research on non-heterosexuality has been to show that Western ideas and assumptions about homosexuality and LGBTQ identities are not necessarily shared in other contexts, and patterns of behavior vary significantly (Reddy, 2006). For example, in many societies, the notion of a gay identity is not common (although globalization has certainly made gay identities available to some people in the non-Western world). While in many Western contexts non-hetereosexual practices are associated with both a defined sexual orientation, and even an identity, in many settings people participate in same-sex behaviors while still maintaining what might be called hetero-normative identities and behavior, such as marrying and having children with members of the opposite sex.

In Western societies, advocates for homosexual rights frequently point to evidence for the biological basis for homosexuality. It is well beyond the scope of this article to say much about this, but it is worth noting that Western understandings of biology as a fundamental basis for explaining behavior certainly contribute to an environment in which claiming that homosexuality is biologically based would appear to be a strong argument for combating stigma associated with it. Such an observation says nothing about actual genetic or biological realities; it simply points out that the relative importance of biology as a basis for recognizing and conferring legitimacy is itself culturally variable (Fausto-Sterling, 2000). While some cross-cultural variation in non-heterosexual practices (and beliefs about homo-, bi-, and transsexuality) might well be partly a response to the stigma associated with them in many settings, the extent of the variation certainly suggests that – like heterosexuality – non-heterosexuality is a profoundly socially constructed arena of human behavior.

The stigma surrounding non-heterosexuality that is common in so many societies stems, at least in part, from the ways that public discourse about it is linked to concerns about social reproduction. When societies engage in animated – and often ugly – debates about whether to punish homosexuality with imprisonment, or even death (as has been occurring in various African countries in recent years), or when national debates about gay marriage create polarizing divisions, such as in the United States, what is at stake is not simply attitudes toward homosexuality, but people's concerns about social reproduction more generally. In this regard, non-heterosexuality is a fertile topic through which societies debate, contest, reproduce, and change what they collectively want for the future. Concerns about social reproduction – about the kinds of values, norms, people, and institutions that will constitute society – are voiced and contested in culture wars over those seen as aberrant. Homophobia is not just about sex, it is about society. While people may be mistaken in associating or blaming homosexuality and other stigmatized forms of sexual behavior for a range of societal ills, these connections nonetheless make such prejudices harder to undo.

Sex, Culture, and AIDS

Some scholars have argued that nothing in history has generated more research and scholarship about the cultural aspects of sexuality than the global AIDS pandemic (Parker, 2001). The

AIDS research industry devoted to understanding the cultural aspects of transmission has, of course, produced a plethora of knowledge about culture and sexuality. This research has contributed to understandings about the variety of cross-cultural beliefs and practices related to sexuality, confirming the basic social science insight that sexuality is a highly contextual, fundamentally socially constructed arena of human activity. Further, recent scholarship has demonstrated that the stigma associated with AIDS, challenges faced (and sometimes caused) by public health prevention and treatment apparatuses, and the factors associated with HIV risk are all fundamentally bound up in social processes that are about much more than sexually-transmitted infection routes (Setel, 1999; Smith, 2014).

It is impossible to summarize the thousands of articles and books about AIDS, sexuality, and culture here, but what stands out from this huge literature is the absolute necessity of situating people's sexual decision-making and behavior in a specific cultural context. Every form of sexual activity occurs in situations where economic and social circumstances are significant and cultural (and often religious) meanings matter. The most insightful literature on sexuality in the era of AIDS has shown that it is important not to allow assumptions about sexuality – whether they are popular local ideas or prevailing scholarly orthodoxy – to go unexamined. Over and over again, research has shown that reality is different than what is assumed.

A few examples from studies of sexuality in the context of AIDS are important reminders that good social science research on the cultural aspects of sexuality is vital in order to understand properly why people behave the ways they do, in addition to generating insights into how it might be possible to design and implement effective public health interventions that affect such behavior. For example, so-called transactional sex, where sex is exchanged (typically by women) for some sort of material support (typically from men), is often seen as a potential driver of HIV risk and infection, and as stemming from women's economic destitution. But research has shown that transactional sex frequently involves aspirations over and above mere survival; relationships commonly entail deeper social and affective connections; material support can signal emotion as much as economics; and women often exhibit much more agency than is stereotypically assumed (Hunter, 2010). Addressing the risk of HIV infection in such relationships requires confronting these complex dynamics.

Perhaps nothing illustrates more clearly how culturally mediated understandings of sexuality influence behavior than does the response of many people (particularly in Africa, although in a variety of other societies) to the promotion of condoms as a form of protection from HIV infection. Because AIDS is a stigmatized disease, and because in many societies the epidemic has come to be understood as symbolic of a kind of collective social and moral crisis, individual infection is often interpreted as a sign of personal ethical failure, and even as sin. In this context, ironically, condoms have become symbols of the disease and their use can be seen as signaling that users are the kind of people (or are having the kind of sex) that could spread HIV (Preston-Whyte, 1999). As a result, condom use is sometimes eschewed not because people don't want to prevent HIV infection, but because the cultural meanings of condoms make their use problematic (Smith, 2014).

A final point drawn from reviewing much of the literature on culture and sexuality as they relate to the AIDS pandemic provides an important cautionary lesson with regard to research not only on AIDS, but also on the cultural aspects of sexuality more generally. In distinguishing good and bad scholarship about culture and sexuality, it is important to be aware of the difference between research that offers culturally contextualized explanations and work that might be described as 'culturalist.' Culturalist accounts tend to attribute sexuality and sexual behavior to some sort of timeless notion of culture that often invokes tradition and fails to situate the cultural aspects of sexuality in their social, political, and economic contexts. The idea of 'African sexuality' as it has been invoked to explain the continent's AIDS epidemic is one of the most egregious examples of culturalist fallacy: both scholarly research and media representations of the epidemic have often blamed so-called traditional African sexual practices for the high rates of infection there. While it is necessary to avoid such fallacies, situating sexuality and sexual behavior in cultural context is essential not only to combat the AIDS epidemic, but also to facilitate understanding the human condition more generally.

Conclusion

Two fundamental aspects of the relationship between sexuality and culture organized this article: (1) the fact that sexuality is a culturally constructed domain of human behavior and (2) the reality that cultural preoccupations with sexuality are often about much more than sex. These two aspects are connected. Ultimately, each reflects the extent to which the cultural dimensions of sexuality are related to the centrality of sexuality in social reproduction. Sexuality is such a culturally significant domain because it is fundamental to the way that societies create and reproduce not only people, but also the values, institutions, and life-ways that are the very fabric of social life.

Editor's Note

There are two entire sections of this Encyclopedia, one on Heterosexuality and another on LGBT Studies, virtually all of the articles of which could be plausibly cited as cross-references to this article. Particularly relevant articles are listed in *See also* section below.

See also: Adolescent Sexual Risk; Gender, Feminism, and Sexuality in Archaeological Studies; Heterosexuality; Hooking up and Casual Sex; Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview; Sexuality Over the Life Course; Sexuality and Social Work.

Bibliography

Connell, R.W., 1987. Gender and Power: Society, the Person and Sexual Politics. Stanford University Press.

Connell, R.W., 1995. Masculinities. University of California Press.

Coontz, S., 2006. Marriage, a History: How Love Conquered Marriage. Penguin Books.

Davis, D.L., Whitten, R.G., 1987. The cross-cultural study of human sexuality. Annual Review of Anthropology 16, 69–98.
Fausto-Sterling, A., 2000. Sexing the Body: Gender Politics and the Construction of Sexuality. Basic Books.
Gutmann, M.C., 1996. The Meanings of Macho: Being a Man in Mexico City. University of California Press.
Gutmann, M.C., 1997. Trafficking in men: the anthropology of masculinity. Annual Review of Anthropology 26, 385–409.
Hirsch, J.S., 2003. A Courtship after Marriage: Sexuality and Love in Mexican Transnational Families. University of California Press.
Hirsch, J.S., Wardlow, H., Smith, D.J., Phinney, H., Parikh, S., Nathanson, C.A., 2009. The Secret: Love, Marriage, and HIV. Vanderbilt University Press.
Hunter, M., 2010. Love in the Time of AIDS: Inequality, Gender, and Rights in South Africa. Indiana University Press.
Inhorn, M.C., 1996. Infertility and Patriarchy: The Cultural Politics of Gender and Family Life in Egypt. University of Pennsylvania Press.
Kandiyoti, D., 1988. Bargaining with patriarchy. Gender & Society 2 (3), 274–290.
Ortner, S.B., Whitehead, H. (Eds.), 1981. Sexual Meanings: The Cultural Construction of Gender and Sexuality. Cambridge University Press.
Parker, R., 2001. Sexuality, culture, and power in HIV/AIDS research. Annual Review of Anthropology 30, 163–179.
Preston-Whyte, E., 1999. Reproductive health and the condom dilemma: Identifying situational barriers to HIV protection in South Africa. In: Caldwell, et al. (Eds.), Resistances to Behavioural Changes to Reduce HIV/AIDS Infections, pp. 139–155.
Reddy, G., 2006. With Respect to Sex: Negotiating Hijra Identity in South Asia. University of Chicago Press.
Rosaldo, M.Z., Lamphere, L., Bamberger, J. (Eds.), 1974. Woman, Culture, and Society. Stanford University Press.
Rubin, G., 2004. The traffic in women: notes on the "political economy" of sex. In: Rivkin, J., Ryan, M. (Eds.), Literary Theory: An Anthology. Blackwell, pp. 770–794.
Setel, P., 1999. A Plague of Paradoxes: AIDS, Culture, and Demography in Northern Tanzania. University of Chicago Press.
Smith, D.J., 2000. "These girls today na war-o": premarital sexuality and modern identity in southeastern Nigeria. Africa Today 47 (3), 99–120.
Smith, D.J., 2014. AIDS Doesn't Show its Face: Inequality, Morality, and Social Change in Nigeria. University of Chicago Press.
Thornton, A., Fricke, T.E., 1987. Social change and the family: comparative perspectives from the West, China, and South Asia. Sociological Forum 2 (4), 746–779.
Utomo, I.D., McDonald, P., 2009. Adolescent reproductive health in Indonesia: contested values and policy inaction. Studies in Family Planning 40 (2), 133–146.
Wardlow, H., 2006. Wayward Women: Sexuality and Agency in a New Guinea Society. University of California Press.
Westoff, C.F., 2003. Trends in Marriage and Early Childbearing in Developing Countries (No. 5). ORC Macro, MEASURE DHS.
Weston, K., 1993. Lesbian/gay studies in the house of anthropology. Annual Review of Anthropology 22, 339–367.
Widmer, E.D., Treas, J., Newcomb, R., 1998. Attitudes toward nonmarital sex in 24 countries. Journal of Sex Research 35 (4), 349–358.
Wood, W., Eagly, A.H., 2002. A cross-cultural analysis of the behavior of women and men: implications for the origins of sex differences. Psychological bulletin 128 (5), 699–727.

Sexuality Education

Sara S Marques and Norman A Constantine, University of California, Berkeley, CA, USA; and Public Health Institute, Oakland, CA, USA
Eva S Goldfarb, Montclair State University, Montclair, NJ, USA
Jane Mauldon, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of the different forms of formal sexuality education. It includes definitions of key terminologies used in this field, a review of the standards and frameworks developed to guide sexuality education programs, and a brief critical review of the evidence of effectiveness for different approaches to sexuality education. It also discusses the current public policy landscape for sexuality education.

Introduction

Sexuality education is a lifelong intentional process by which people learn about themselves and others as sexual, gendered beings from biological, psychological, and sociocultural perspectives (Goldfarb and Constantine, 2011). Sexuality education exists in some form in every country and culture. Formal sexuality education is often implemented through school curricula, and may also be offered by health care, faith-based, or community-based organizations. Much of sexuality education, however, is informal, occurring within the process of sexual socialization. Through this process individuals acquire attitudes, values, ideas, and patterns of behavior related to sexuality by observing and interacting with family members, peers, and other members of their community, and by consuming electronic and print entertainment, advertising, news, and public information that have explicit or implicit sexual content.

Most formal sexuality education programs around the world focus on adolescents, serving individuals ranging in age from the preteen years into the early twenties. Sexuality education can be an important tool in helping adolescents navigate the biological, cognitive, and social transitions that transform how they look, think, feel, and interact with others. However, most formal sexuality education does not have such a wide-ranging agenda. In many countries and communities, the concept of sexuality education is highly polarizing and political. In these contexts, the policies that undergird and fund sexuality education programs often reflect the interests and concerns of organized pressure groups. This can result in programs that are restricted in content as well as their purposes, which may include fostering a preference for sex within marriage, discouraging premarital sex, teen pregnancy or teen childbearing, lowering youths' risks of becoming infected with HIV or another sexually transmitted infection, or helping youth avoid sexually coercive behavior.

Terminology and Definitions

Across countries, the terminology used to describe sexuality education policies and programs conveys information regarding the lens through which sexuality education is viewed. In Eastern Europe, for example, family life education is a common terminology. Similarly, Nigeria uses a National Family Life and HIV Education Curriculum (UNESCO, 2010). In some countries, apprehension about HIV/AIDS has resulted in sexuality education programs being subsumed within national HIV prevention programs. For example, with concerns rising about HIV/AIDS in the 1990s, Portugal changed the name of its national program from Program for Personal Development to Program for Health Education, thus embedding their sexuality education program into a more explicitly health oriented framework (Parker et al., 2009). In the United States, the most common forms of sexuality education are generally referred to as abstinence-only (or sometimes abstinence-only-until-marriage) and abstinence-plus. In addition, youth development programs with a pregnancy postponement or sexual risk reduction goal are also sometimes viewed as sexuality education programs. For example, the positively evaluated Carrera Pregnancy Prevention Program has been described as "age-appropriate sex education with access to medical, mental health, academic support, financial literacy, and career planning services" (Carrera, 2013).

In general, these different terms and corresponding approaches can be categorized into two broader categories that reflect historic tensions between competing visions in the field – 'sex education' and the more limited 'comprehensive sexuality education' (Goldfarb, 2009). Sex education has generally referred to instructional approaches focused on preventing pregnancy and disease primarily through controlling young people's behaviors. The course content of sex education typically encompasses anatomy, sexual behavior, reproductive functioning, and methods of disease and pregnancy prevention.

Comprehensive sexuality education, on the other hand, is based on the view that individuals' physical and mental health can be enhanced by a holistic and scientifically grounded understanding of sexuality. Preventing pregnancy and disease represent only one component of a broader goal of fostering sexual health. Comprehensive sexuality education approaches also include a substantial focus on topics such as gender issues, relationships, sexual development, sexual orientation, and pleasure. In this article, the more generic term 'sexuality education' is used to encompass both the sex education and comprehensive sexuality education approaches, as well as other forms of education that address sexual health and sexuality.

Abstinence-Only and Abstinence-Plus Sex Education

Approaches classified as sex education can be further categorized into two divergent categories: abstinence-only and abstinence-plus sex education. Abstinence-only sex education has as its exclusive purpose, teaching the social, psychological, and health gains of abstaining from sexual activity, including, among other things, the idea that a mutually faithful, monogamous relationship within the context of marriage is the expected standard of human sexual activity (Trenholm et al., 2008). Programs under this umbrella explicitly exclude content related to contraception or condoms – other than to note failure rates – as well as safer sex, abortion, and many other topics. These programs tend to focus on religiously based concepts such as chastity, virginity, purity, and the sanctity of marriage.

Abstinence-plus sex education (also known as abstinence-based sex education) promotes the avoidance of risky sexual behaviors, stressing abstinence as the best way to reduce rates of unintended pregnancies and sexually transmitted diseases (STDs). Using a disease prevention, or harm reduction approach, such programs are distinguished from abstinence-only programs by their public health underpinnings and specifically their inclusion of information and instruction about contraception and other protection. These programs are often incorrectly referred to as comprehensive because of the addition of these topics, leading to confusion with the term comprehensive sexuality education. True comprehensive sexuality education embraces a much wider range of goals and topics than just those directly related to disease and pregnancy prevention.

Comprehensive Sexuality Education

Viewing sexuality as a multifaceted force in human life, comprehensive sexuality education programs go beyond disease and pregnancy to address human development, gender issues, sexual orientation and identity, love, sexual health problems, sexual attraction, pleasure, parenting, and sexual and reproductive rights. In addition, these programs are often designed to develop skills in critical thinking, media literacy, and communication. One example of a guiding paradigm for comprehensive sexuality education is the Circles of Sexuality model, developed by Dailey (1981) and adapted by Goldfarb and Constantine (2011). This model identifies five overlapping components of human experience – sensuality, intimacy, sexual identity, sexual health and reproduction, and sexualization – that combine to create the totality of human sexuality.

Controversy attaches to whether comprehensive programs foster or teach sexual morality. Religiously inflected concepts such as chastity and virtue are not typically presented. However, abstinence can be encouraged as a developmentally appropriate and self-protective choice for youth. In addition with the inclusion of discussions (and sometimes skills development) about couple communication and gender relations, comprehensive programs tend to focus more than abstinence-only or abstinence-plus programs on the mindsets and skills that protect against sexual coercion and foster sexual autonomy.

Comprehensive sexuality education has been widely adopted in schools in the Netherlands and Sweden, as well as in other countries of Western Europe and some developing countries (Berne and Huberman, 1999; International Sexuality and HIV Curriculum Working Group, 2009; Parker et al., 2009). In the United States, public schools generally do not follow a comprehensive sexuality education model but private schools and community institutions may, as illustrated by the six-level lifespan sexuality education program, Our Whole Lives, published by the Unitarian Universalist Association and the United Church of Christ.

Addressing Diverse Groups of Learners

Because young people are developmentally and otherwise diverse within even a narrow age range, a 'one size fits all' strategy is rarely appropriate in sexuality education. If learners feel that the norms of their group or culture are not being addressed in a sexuality education program, they will be less likely to internalize its messages. Complicating matters, many people identify with more than one cultural group. They may, for example, relate simultaneously to cultural values and customs of their ethnic group, the strictures of their religion, and the imperatives of their sexual orientation, all of which may conflict.

Some groups of youth are chronically underserved by sexuality education programs. Among the groups least well-served are sexual minority youth – including those who identify as lesbian, gay, bisexual, transgender, or questioning (LGBTQ) – and learners with disabilities.

LGBTQ Youth

Most sexuality education programs, unless specifically designed for LGBTQ youth, employ a heteronormative framework for almost every discussion about contraception, safer sex, relationship building, resistance skills, and sexual violence or abuse, among other topics. Scenarios used for role-play, critical thinking, or values-clarification activities inevitably involve a male and female. Sometimes, well-intentioned programs will use gender-neutral names such as Corey and Jesse, but the content and issues raised are consistently based on a heterosexual paradigm. Furthermore, most sexuality education programs address the topics of sexual orientation, gender identity, and homophobia minimally, if at all. This is especially true of the abstinence-only programs. As a result, LGBTQ youth often find themselves alienated, marginalized, and invisible.

People with Disabilities

Two common and mutually contradictory myths about people with disabilities are that they are not sexual at all and that they are unable to control their sexuality. Research has shown that these and other misconceptions foster discomfort among families and educators about offering sexuality education to people with disabilities. Because children with disabilities are at a greater risk for sexual exploitation and abuse, the first inclination among parents and caregivers often is to keep them ignorant in an effort to protect them. Not educating young

people with disabilities, however, increases rather than decreases their risk. People with disabilities have the same range and intensity in their sexual needs and their desires for love and intimacy as their nondisabled peers. Recognizing disabled students in sexuality education communicates this important message to all.

Public Policies and Legal Frameworks

A variety of public policies converge to create the legal framework for sexuality education and to shape its provision in any given location (see **Box 1** for examples of relevant policy areas). These policies establish a foundation that helps determine the relative ease or difficulty with which sexuality education can be implemented, who receives it, and in what form it is available (Constantine, 2008; Parker et al., 2009).

The public policy context for sexuality education extends from the international level to the local level. At the international level, declarations ratified by United Nations member countries establish global support for sexuality education from diverse perspectives including economic development, gender equity, and health (see **Box 2**). Similarly, regional conventions have been developed to create a shared vision of sexuality education. One such declaration was signed in Mexico City in 2008 by health and education ministers from Latin America. The Latin American Ministerial Declaration mandates school-based sexuality and HIV education throughout the region (UNAIDS, 2008). These declarations provide an important impetus for the establishment of national programs in many countries. However, sexuality education is a long way from being institutionalized in every country, and the amount and nature of local programming typically reflects a country's income levels as well as its political culture and health needs (UNESCO, 2012).

There is a great variety in sexuality education policies across, and often within, countries, states, and school districts. National agencies responsible for sexuality education policies include ministries of education and of health and departments focused on HIV/AIDS, family planning, youth, sports, and others. In Europe, the Ministry of Education is most commonly involved in cooperation with one or more departments.

Some countries, such as Nigeria, have created a national curriculum and have taken on responsibility for its implementation at the central level (UNESCO, 2010). In the United States, however, while sexuality education funding is provided by federal agencies, responsibilities to specify and implement program approaches are conveyed to states and then to individual school districts.

School-based delivery of sexuality education is common across the globe, but countries' policies and programs also differ substantially in the specification of curriculum content, location of implementation, designation of the educator, required educator training, age(s) at which sexuality education should be provided, and in whether parental permission is required. A few countries employ an integrated approach to sexuality education, one in which content is spread across several subject areas. In the United States, however, most students receive sexuality education as a part of biology or health classes (Hoff et al., 2000).

Many countries rely on staff or volunteers from nongovernmental organizations, or a combination of school and community organizations to implement sexuality education programs. Some programs designate peer education as their mode of delivery for sexuality education (e.g., India, Indonesia, Bangladesh; UNESCO Bangkok, 2012).

Policies generally specify the time frame during which sexuality education should be delivered to students through designating either the age or the grade (i.e., schooling level) at which students should receive it. Experts generally agree that an early and developmentally appropriate start to sexuality

Box 1 Policy areas relevant to sexuality education

Policy areas specific to school-based sexuality education:

- Governmental agency responsibility
- Mandatory provision of sexuality education
- Content areas
- Age at which provided
- Parental notification (opt-in and opt-out)
- Person responsible for teaching
- Training required for teacher
- Confidentiality

Related policy areas:

- Age of consent for sexual and reproductive health care services
- Abortion
- Contraceptive provision including condom availability
- Age of mandatory schooling
- Nondiscrimination
- HIV/AIDS testing and treatment
- Testing and treatment for STDs
- Needle exchange

Box 2 International conventions related to sexuality education

1994 International Conference on Population and Development (ICPD) Programme of Action
1999 Overall review and appraisal of the implementation of the Programme of Action of the International Conference on Population and Development (ICPD + 5)
1995 United Nations Fourth World Conference on Women Platform for Action
2000 United Nations Committee on Economic, Social, and Cultural Rights
2001 General Assembly Special Session on HIV/AIDS, Declaration of Commitment on HIV/AIDS
2003 United Nations Committee on the Rights of the Child
2006 United Nations Convention on the Rights of Persons with Disabilities

education is important. In Europe, of the 25 countries for which data are available, 14 countries require that sexuality education begin at 10 years of age or younger, 7 countries require that it begin between 11 and 12 years of age, and 4 countries require that it begin after 12 years. In practice, a majority of these countries implement sexuality education starting around 12 years of age (Parker et al., 2009). In a 28-country, Asia–Pacific study, fewer than half of the countries implemented sexuality education at the primary school level (UNESCO Bangkok, 2012).

The choice between opt-in, opt-out, or no parent permission required in sexuality education is often set by national or local policy, rather than by individual schools. Opt-in policies are those that require parents or guardians to actively provide permission for their children to participate in sexuality education, whereas opt-out policies require notification of parents, but put the onus on the parents to actively request that their children be excluded from participation. Although many countries do have policies that allow parents to exclude their children from sexuality education, others (e.g., Germany) have mandated sexuality education as a part of the national curriculum and, therefore, do not provide parents the option to exclude their children from the sexuality education curriculum.

Additional policies that address the wider context in which sexuality education is implemented, including policies such as confidentiality, availability to youth of reproductive health services, testing and treatment availability, among others, are also important aspects of the policy landscape for sexuality education.

A Case Study: United States

The United States offers an example of a decentralized system that has several layers of sexuality education policies ranging from the national level to the local level. Mandates for sexuality education exist in 23 of the 50 states and in the District of Columbia, whereas 34 states require the provision of HIV education. Twenty-six states require that sexuality education stress abstinence, 11 states require that abstinence be covered, 18 states require that sexuality education include information on contraception, and 19 states require that HIV education include information on condoms (Guttmacher Institute, 2013).

This diversity across state policies reflects national controversy surrounding sexuality education. The heated debates triggered by the topic illustrate its sensitive and often politicized nature, and a lack of consensus about what content should be included and who should teach it (Constantine, 2008; Constantine et al., 2007; Irvine, 2002; Luker, 2007; Parker et al., 2009). These debates have contributed to, as well as resulted from frequent policy shifts nationally and within states. For example, currently in 2013, federal funds for sexuality education in the United States flow to states under two different laws that reflect distinct and mutually inconsistent priorities. The Title V abstinence-only-until-marriage education law (the Abstinence Education Provision of the 1996 Welfare Law, P.L. 104–193, 1998) has for nearly 20 years funded states' abstinence-only sexuality education programs. In contrast, since 2012, states may apply for federal funds through the Personal Responsibility Education Program (PREP) – part of the Patient Protection and Affordable Care Act (P.L. 111–148) – for abstinence-plus sexuality education with a focus on pregnancy outcomes. Thus, states may choose to apply for federal abstinence-plus (PREP) funds, abstinence-only (Title V) funds, or both, even though the implementation guidelines for the two program approaches are incompatible.

State policies generally specify the amount and type of sexuality education that schools provide, determine course content, and set the terms of parental authority over children's participation. Like national policies, some states' policies have lurched from one philosophical approach to another. For example, in 2012, Wisconsin shifted from a more comprehensive approach to sexuality education to a more restrictive policy that reestablished an abstinence-only-until-marriage approach for the state's secondary school students (SIECUS, 2010). On the other hand, in 2005, Oregon made a shift from a risk-focused approach toward a youth development model that embraces sexuality as a natural part of adolescence (Nystrom et al., 2013).

Educational approaches typically vary across school districts within parameters set by state law, as illustrated by the California Comprehensive Sexual Health and HIV/AIDS Prevention Education Act of 2003 (Education Code ss. 51930–51939). This law mandates a minimum of 2 hours of HIV prevention education across grades 7 through 12 and further requires that all sexuality education taught in public school districts or otherwise supported by state funds be age-appropriate, bias-free, medically accurate, and include discussion of all contraception methods approved by the US Food and Drug Administration. Furthermore, materials must be appropriate for students of all genders, sexual orientations, ethnic and cultural backgrounds, and accessible for English language learners and students with disabilities. An evaluation of the implementation of the 2003 California law, however, found that many school districts do not fully comply with these mandates. Students continue to receive incomplete, and in some cases inaccurate, information about sexuality through their school-based sex education programs (Combellick and Brindis, 2011).

Frameworks and Standards

In multiple separate attempts to create a common vision, bring cohesion to the field, and to provide guidance for those seeking to develop and implement sexuality education, various organizations have developed frameworks or standards to help sexuality education. These frameworks correspond to either abstinence-only, abstinence-plus, or comprehensive sexuality education and make apparent the philosophical as well as practical differences that separate these three broad approaches.

An Abstinence-Only Sex Education Framework

Unlike the two more expansive approaches that are discussed below, abstinence-only sex education in the United States is shaped by a legislatively mandated framework that is explicitly prescriptive. The Title V abstinence-only-until-marriage

education law (discussed above) defines abstinence education as an educational or motivational program with the following eight characteristics:

- Has as its exclusive purpose, the teaching of social, psychological, and health gains to be realized by abstaining from sexual activity.
- Teaches abstinence from sexual activity outside marriage as the expected standard for all school-age children.
- Teaches that abstinence from sexual activity is the only certain way to avoid out-of-wedlock pregnancy, STDs, and other associated health problems.
- Teaches that a mutually faithful and monogamous relationship in context of marriage is the expected standard of human sexual activity.
- Teaches that sexual activity outside of the context of marriage is likely to have harmful psychological and physical effects.
- Teaches that bearing children out-of-wedlock is likely to have harmful consequences for the child, the child's parents, and society.
- Teaches young people how to reject sexual advances, and how alcohol and drug use increases vulnerability to sexual advances.
- Teaches the importance of attaining self-sufficiency before engaging in sexual activity.

These features have been used as criteria for states to receive Title V funds for their programs. Exactly how closely a funded program must align with these eight points has been vigorously debated over the years, and federal program guidance on compliance has changed with successive presidential administrations.

An Abstinence-Plus Sex Education Framework

Abstinence-plus education has a longer history than abstinence-only, and a larger body of evaluative studies is employed to support effectiveness claims for curricula-based abstinence-plus programs. The National Campaign to Prevent Teen and Unplanned Pregnancy (Kirby, 2007) has published a set of 17 desirable characteristics for sexuality education programs, which include three pertaining to curriculum content:

- Focus on clear health goals – the prevention of STDs/HIV, pregnancy, or both.
- Focus narrowly on specific types of behavior leading to these health goals (e.g., abstaining from sex, using condoms or other contraceptives), give clear messages about these types of behavior, and address situations that might interfere with them and how to avoid those situations.
- Address sexual psychosocial risk and protective factors that affect and change sexual behavior (e.g., knowledge, perceived risks, values, attitudes, perceived norms, and self-efficacy) and change them.

International Comprehensive Sexuality Education Frameworks

The comprehensive sexuality education frameworks discussed next reflect the missions of the three organizations that developed them. The World Health Organization (WHO) framework (WHO, 2010) aims at reconceptualizing sexual health as a core element of overall health rather than as merely a component of reproductive health. Centered around standards and principles of international human rights agreements, this framework calls for complementary action across multiple domains, including education, to support population sexual health. The United Nations Population Fund framework (UNFPA, 2010), developed from a convened 2010 meeting of global leaders in sexuality education, likewise intends to advance human rights, gender equality, and sexual and reproductive health. Finally, the International Planned Parenthood Federation (IPPF, 2010) framework also employs a rights-based approach.

World Health Organization framework. This framework asserts that sexual health recognizes "the desire of individuals and couples of all sexual orientations and any background to have fulfilling and pleasurable sexual relationships, and its concerns go beyond fertility and reproduction to encompass issues such as sexual dysfunction and disability, and violence related to sexuality" (WHO, 2010: p. iv). According to this framework, comprehensive approaches to sexuality education are likely to focus on:

- prevention of sexually transmitted infections and HIV, contraception, and the mechanics of fertility and reproduction;
- the role of pleasure, eroticism, and satisfaction;
- gender differences, inequalities and human rights, and about the negative and positive effects of gender norms;
- the importance of responsibility and joint decision-making, and training in communication and negotiation skills; and
- sexual and gender identity and sexual choice.

United Nations Population Fund (UNFPA) report. This report (UNFPA, 2010) affirms that comprehensive sexuality education is a human right, arguing that all young people should have access to effective sexuality education programs that are comprehensive, rights-based, and gender sensitive. According to this report, measures of program effectiveness should track changes in gender equality and human rights, as research shows these to be correlated with better sexual and reproductive health. The UNFPA report offers the following recommendations for comprehensive sexuality education:

- Address norms around gender and sexuality, and promote equality, empowerment, nondiscrimination, and respect for diversity.
- Develop indicators that go beyond the biomedical to measure effectiveness, including nonhealth indicators, and evidence of other valuable outcomes of sexuality education.
- In school systems, integrate sexuality, gender, and citizenship objectives into educational goals, and incorporate them across the curriculum.

International Planned Parenthood Federation (IPPF) *framework*. This framework identifies seven essential components of comprehensive sexuality education: gender, sexual, and reproductive health and HIV, sexual rights and sexual citizenship, pleasure, violence, diversity, and relationships. The framework uniquely identifies pleasure as one of the essential components of sexuality, asserting "the right of all young people to enjoy sex and express their sexuality in the way that

they choose" (IPPF, 2010: p. 5). Not offering detailed messages or objectives, the IPPF framework has been purposefully designed for broader application than just in schools. IPPF acknowledges that sexuality education happens formally and informally, in a variety of settings including youth clubs, hair salons, waiting rooms, and often by peer educators. Thus, IPPF created this framework to be a force for collaboration among different youth-serving agencies and nongovernmental organizations.

United States Comprehensive Sexuality Education Frameworks

National Guidelines Task Force framework. Led by the Sexuality Information and Education Council of the United States (National Guidelines Task Force, SIECUS, 2004), a task force of educators, health professionals, and representatives from national US organizations proposed guidelines for comprehensive and developmentally appropriate school-based sexuality education for kindergarten through 12th grade. The guidelines are organized into six key concepts: human development, relationships, personal skills, sexual behavior, sexual health and society, and culture. The guiding vision for this framework is a portrait of a 'sexually healthy adult,' which includes a list of behaviors corresponding to sexual health relating to each of the key concepts. The guidelines identify the primary goal of sexuality education as the promotion of sexual health, and specify four areas to be addressed in pursuit of this goal: providing accurate information related to sexual matters; exploring students' attitudes, values, and insights; cultivating skills in interpersonal relationships; and fostering responsibility (National Guidelines Task Force, SIECUS, 2004).

Future of Sexuality Education national standards. A set of minimal and essential standards has been developed by a team of experts from the fields of health education, sexuality education, public health, public policy, and sexual health advocacy through the leadership of Future of Sexuality Education (FoSE) consortium of leading US sexuality education organizations. The intent of these standards is to foster wider and more consistent provision of sexuality education across the United States, and to help educators make optimal use of the limited time that is allocated to sexuality education in US schools (which averages 3.1 h across all elementary grades, 6 h in middle school, and 8.1 h in high school, according to the FoSE report). The standards include seven topic areas: anatomy and physiology, puberty and adolescent development, identity, pregnancy and reproduction, STDs and HIV, healthy relationships, and personal safety (Future of Sex Education Initiative, 2012).

Evidence of Effectiveness

An extensive body of research exists on the effectiveness of sexuality education interventions in reducing adolescent risky sexual behavior. Hundreds of individual studies have been conducted, along with numerous syntheses, including narrative reviews, programs-that-work lists, and meta-analyses (statistical cumulations of results across multiple evaluations). A large proportion of this research is from the United States.

In any study of program effectiveness, the choice of primary outcomes is critical. Behavior-change outcomes, such as reducing sexual activity and increasing condom and contraception use, have been largely prioritized in program evaluations in this field. Notably missing from this research are outcomes of positive sexuality and behavior, for example, the ability to enter into and maintain healthy relationships, strong communication skills, experiencing positive romantic and sexual relationships, and asserting one's sexual rights and respecting the rights of others. This is consistent with the predominance of research focused on abstinence-plus and abstinence-only curricula and programs, as opposed to fully comprehensive sexuality education.

The largest and most rigorous sexuality education evaluation to date was a United States congressionally mandated 5-year randomized trial of abstinence-only education programs (Trenholm et al., 2008). This study of four popular abstinence-only programs showed no significant differences between program participants and control students on the primary outcomes of number of sexual partners, expectations to abstain, and reported rates of pregnancy, births, and STDs. These results were reinforced by a large-scale observational study employing the Centers for Disease Control and Prevention's National Survey of Family Growth, which also indicated no differences in teen pregnancy rates among adolescents who reported having received abstinence-only sex education, versus those who reported receiving no sex education (Kohler et al., 2008). This study also provided evidence of positive outcomes to those who reported having received abstinence-plus curricula – half the risk of teen pregnancy compared to those who reported abstinence-only sex education and 40% of the risk compared to those who reported no sex education, as well as moderately decreased likelihood of engaging in vaginal sex compared with those who received no sex education.

Studies to test the effectiveness of specific individual abstinence-plus sex education programs have yielded less compelling evidence. Although numerous evaluations have reported positive effects of abstinence-plus programs, most of these evaluations have taken place at the local school district, or community level, and they suffer from some of the common methodological flaws and biases typically associated with these types of evaluations (see Constantine, 2013, for a critical review). Some narrative reviews have led to conclusions of effectiveness for the best of these programs, and some reviews have resulted in lists of recommended evidence-based programs and strategies. But rigorous research syntheses have yielded more cautious conclusions. For example, the authors of the two most prominent meta-analyses conducted in this area, sanctioned by the Campbell Collaboration and the Cochrane Collaboration, concluded that the evidence of effectiveness for abstinence-plus programs in lowering rates of sexual initiation and pregnancy risk was inconsistent, and these authors also commented on the dearth of high-quality evidence on program effectiveness in this area (Scher et al., 2006; Oringanje et al., 2009).

These mixed findings, based on syntheses of multiple evaluations of abstinence-plus programs, seem to contradict the readily available lists of evidence-based programs compiled and publicized by national policy and advocacy

groups, and by the US federal government. In general, the criteria of acceptable evidence of effectiveness necessary for a program to be included on these lists are not stringent. For example, regardless of the number of evaluations conducted and the number of outcomes tested in each evaluation, only one statistically significant outcome is typically required to earn a place on the list. This is referred to as the one study, one outcome fallacy (Constantine, 2013). The problem with this approach is illustrated by a list of 28 programs certified as evidence-based by the US Department of Health and Human Services' for its Teen Pregnancy Prevention Program funding eligibility. According to the Coalition for Evidence-Based Policy (2010), an independent research-use watchdog group, only 2 of the 28 listed programs had strong evidence of effectiveness in prevention of pregnancy or STDs.

In spite of weaknesses in the research base and its policy uses, several broad conclusions are supported by the full body of research evidence on sexuality education effectiveness. First, abstinence-only approaches as typically conceived and implemented have shown limited potential. Second, abstinence-plus approaches that directly focus on promoting behavioral change and include instruction on condoms and contraception methods have better potential, and evidence is accumulating of some modest positive effects overall. Third, with few exceptions, the effectiveness of specific individual abstinence-plus programs and curricula is not well supported by the available evidence. This does not necessarily mean that these programs are ineffective, just that the nature and quality of the available research is inadequate to answer questions at the individual program level. Fourth, true comprehensive sexuality education approaches have not been well studied. Overall, it is fair to say that the full potential to enhance adolescents' sexual health and development through sexuality education is not being realized by the currently popular abstinence-only and abstinence-plus models.

See also: Adolescent Sexual Risk; Methodological Issues in the Study of Sexuality; Sexual Debut; Sexuality Over the Life Course.

Bibliography

Berne, Linda, Huberman, Barbara, 1999. European Approaches to Adolescent Sexual Behavior & Responsibility: Executive Summary and Call to Action. Advocates for Youth, Washington, DC.

Carrera, Michael, March 12, 2013. Teenage pregnancy. New York Times (accessed 11.07.13.).

Coalition for Evidence-Based Policy, 2010. HHS's Evidence-Based Teen Pregnancy Prevention Program. Available from: www.coalition4evidence.org (accessed 14.05.13.).

Combellick, Sarah, Brindis, Claire, 2011. Uneven Progress: Sex Education in California Public Schools. University of California San Francisco, Bixby Center for Global Reproductive Health, San Francisco, CA.

Constantine, Norman A., 2008. Converging evidence leaves policy behind: sex education in the United States. Journal of Adolescent Health 42, 324–326.

Constantine, Norman A., 2013. Intervention effectiveness research in adolescent health psychology: methodological issues and strategies. In: O'Donohue, William T., Benuto, Lorraine T., Woodward Tolle, Lauren (Eds.), Handbook of Adolescent Health Psychology. Springer, New York, NY, pp. 295–322.

Constantine, Norman A., Slater, Jana Kay, Carroll, Julie A., 2007. Motivational aspects of community support for school-based comprehensive sexuality education. Sex Education 7, 421–439.

Dailey, Dennis, 1981. Sexual expression and ageing. In: Berghorn, F., Schafer, D. (Eds.), The Dynamics of Ageing: Original Essays on the Processes and Experiences of Growing Old. Westview Press, Boulder, CO, pp. 311–330.

Future of Sex Education Initiative, 2012. National Sexuality Education Standards: Core Content and Skills, K–12 [a Special Publication of the Journal of School Health]. Available from: www.futureofsexeducation.org (accessed 8.03.13.).

Goldfarb, Eva S., 2009. A crisis of identity for sexuality education in America: how did we get here and where are we going?. In: Schröder, Elizabeth, Kuriansky, Judy (Eds.), Sexuality Education: Past, Present, and Future, vol. 1 Praeger, Westport, CT, pp. 8–30.

Goldfarb, Eva S., Constantine, Norman A., 2011. Sexuality education. In: Brown, B. Bradford, Prinstein, Mitchell J. (Eds.), Encyclopedia of Adolescence, vol. 2. Elsevier Academic Press, London, UK, pp. 322–331.

Guttmacher Institute, 2013. State Policies in Brief: Sex and HIV Education. Available from: http://www.guttmacher.org (accessed 8.03.13.).

Hoff, Tina, Greene, Liberty, McIntosh, Mary, Rawlings, Nicole, D'Amico, Jean, 2000. Sex Education in America: A View from inside the Nation's Classrooms. Kaiser Family Foundation, Menlo Park, CA.

International Planned Parenthood Federation (IPPF), 2010. IPPF Framework for Comprehensive Sexuality Education (CSE). Available from: http://www.ippf.org (accessed 6.03.13.).

International Sexuality and HIV Curriculum Working Group, 2009. It's All One Curriculum: Guidelines and Activities for a Unified Approach to Sexuality, Gender, HIV, and Human Rights Education. The Population Council, Inc, New York.

Irvine, Janice M., 2002. Talk about Sex: The Battles over Sex Education in the United States. University of California Press, Berkeley, CA.

Kirby, Douglas, 2007. Emerging Answers 2007: Research Findings on Programs to Reduce Teen Pregnancy and Sexually Transmitted Diseases. National Campaign to Prevent Teen and Unplanned Pregnancy, Washington, DC.

Kohler, Pamela K., Manhart, Lisa E., Lafferty, William E., 2008. Abstinence-only and comprehensive sex education and the initiation of sexual activity and teen pregnancy. Journal of Adolescent Health 42, 344–351.

Luker, Kristen, 2007. When Sex Goes to School: Warring Views on Sex – and Sex Education – Since the Sixties. Norton and Company, New York.

National Guidelines Task Force, Sexuality Information and Education Council of the United States (SIECUS), 2004. Guidelines for Comprehensive Sexuality Education: Kindergarten through 12th Grade, third ed. Available from: http://www.siecus.org (accessed 9.03.13.).

Nystrom, Robert J., Duke, Jessica E.A., Victor, Brad, 2013. Shifting the paradigm in Oregon from teen pregnancy prevention to youth sexual health. Public Health Reports 128, 89–95.

Oringanje, Chioma, Meremikwu, Martin M., Eko, Hokehe, Esu, Ekpereonne, Meremikwu, Anne, Ehiri, John E., 2009. Interventions for preventing unintended pregnancies among adolescents. Cochrane Database Systematic Reviews 4.

Parker, Rachel, Wellings, Kay, Lazarus, Jeffrey V., 2009. Sexuality education in Europe: an overview of current policies. Sex Education 9, 227–242.

Scher, Lauren S., Maynard, Rebecca A., Stagner, Matthew, 2006. Interventions intended to reduce pregnancy-related outcomes among adolescents. Campbell Collaboration 12.

Sexuality Information and Education Council of the United States (SIECUS). 2010. State Profiles: A Portrait of Sexuality Education and Abstinence-Only-Until-Marriage Programs in the States (Fiscal Year 2010 Edition). Available from: http://www.siecus.org (accessed 11.03.13.).

Trenholm, Christopher, Devaney, Barbara, Fortson, Kenneth, et al., 2008. Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. Journal of Policy Analysis and Management 27, 255–276.

UNAIDS, 2008. Ministerial Declaration: Preventing through Education: 1st Meeting of Ministers of Health and Education to Stop HIV and STIs in Latin America and the Caribbean. Available from: http://data.unaids.org (accessed 12.03.13.).

UNESCO, 2010. Levers of Success: Case Studies of National Sexuality Education Programmes. Available at: www.unesco.org (accessed 4.03.13.).

UNESCO, 2012. Comprehensive Sexuality Education: The Challenges and Opportunities of Scaling-up. Available from: www.unesco.org (accessed 13.03.13.).

UNESCO Bangkok, 2012. Review of Policies and Strategies to Implement and Scale Up Sexuality Education in Asia and the Pacific. Available from: www.unesco.org (accessed 14.03.13.).

UNFPA, 2010. Comprehensive Sexuality Education: Advancing Human Rights, Gender Equality and Improved Sexual and Reproductive Health. Available from: http://www.unfpa.org (accessed 7.03.13.).

World Health Organization (WHO), 2010. Developing Sexual Health Programmes: A Framework for Action. Available from: http://www.who.int (accessed 9.03.13.).

Relevant Websites

www.advocatesforyouth.org – Advocates for Youth.
www.guttmacher.org – Alan Guttmacher Institute (AGI).
http://crahd.phi.org – Center for Research on Adolescent Health and Development (CRAHD).
www.fordfoundation.org – Ford Foundation Program on Sexuality and Reproductive Health and Rights.
www.ippf.org – International Planned Parenthood Federation (IPPF).
www.popcouncil.org – Population Council.
www.siecus.org – Sexuality Information and Education Council of the United States (SIECUS).
www.unesco.org – United Nations Educational, Scientific, and Cultural Organization (UNESCO).
www.unfpa.org – United Nations Population Fund (UNFPA).
www.worldsexology.org – World Association of Sexual Health (WAS).
www.who.int – World Health Organization (WHO).

Sexuality in Australia

Juliet Richters, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Australia is largely secular, with about 50% reporting a religion in surveys. Attitudes are tolerant, with a majority regarding homosexual acts and abortion as acceptable, and few disapproving of premarital sex. However, sex outside a committed relationship is widely disapproved of. Use of contraception is high. States differ in legislation and policing of the sex industry. Just over half of adolescents have sexual intercourse before they leave high school. Condom use at first intercourse is high, but many adults do not use condoms even with casual partners. Sex education is very patchy. HIV has remained largely concentrated among homosexually active men.

Introduction

The international impression of sexuality in Australia may be one of a free and easy beach culture. The reality is more complex, reflecting an ongoing tension between a tolerant set of secular attitudes and a repressive desire to control sexuality, keeping it within bounds and behind closed doors. The impression of a freely hedonistic culture is drawn on by the designer of the front cover of Frank Bongiorno's recent history of sex in Australia (2012), which shows a naked woman running carefree into the ocean. In reality, naked swimming or sunbaking is illegal in Australia, except at a few designated nude beaches. Even babies and toddlers are expected to wear swimming costumes at beaches and swimming pools, though it is claimed that this is to protect them from paedophiles rather than to protect the public from the sight of their genitals.

The Cultural Climate

Media

Media treatment of sexuality – or at least of research into sexual health – is fairly balanced and calm. Response to the launch of the first large national survey of sexual behavior and attitudes (the Australian Study of Health and Relationships, ASHR; Smith et al., 2003a) in 2003 included a double page spread in the broadsheet *Sydney Morning Herald* and serious coverage in other newspapers and radio programs. Headlines such as 'We're a broadminded lot, most of us' (*Sydney Morning Herald*, 9 April 2003) were common. Journalists now draw on these representative sample data rather than the Kinsey figures, which had become almost folklore, treated as eternal fact despite being very located in the time and place of mid-twentieth-century America. Journalists gave enormous attention to statistics on age at first intercourse, perhaps reflecting the concern among parents about the sex lives of their children, and to frequency of intercourse. Published statistics about sex are a benchmark against which people evaluate their own performance as adequate or inadequate.

Even the later publications from this project that might have prompted more disapproval, such as a journal article on bondage and discipline, 'sadomasochism' or dominance, and submission (Richters et al., 2008), were greeted with respect or amusement (rather than derision or outrage) in the media. The very fact that Australia's first major sex survey was directly government funded is significant. A decade earlier, the first UK study (Johnson et al., 1994) had its funding blocked by the Thatcher government and was eventually run with funding from the private Wellcome Trust. At the same time the first major United States survey (Laumann et al., 1994) had its funding blocked by the first Bush administration after having successfully gone through the usual process of scientific peer review, and was run with a greatly reduced sample size.

Media sex scandals are rare in contemporary Australia; the public mostly regard prominent people's sex lives as their own affair, rather than evidence of thorough-going moral turpitude. A 2013–14 scandal about a trade union official, Craig Thomson, who used his union credit card to pay a brothel, focused not so much on the sexual nature of the services but on his fraud in using union money for personal purchases. Before 27 years, former Prime Minister Malcolm Fraser was caught without his trousers in a Memphis hotel corridor, but this did not bring about his downfall.

Attempts to undo the careers of prominent people using sexual scandal have met with limited success. High Court Judge Michael Kirby, whose career started before the legalization of gay sex, when being 'out' as gay was not an option, was accused in 2002 of using an Australian Government car to cruise rent boys in Sydney. It was a foolish target to accuse of such a lack of probity, and the claim was disproved (Bongiorno, 2012: pp. 288–289). Former Justice Kirby, who is proud to bring his lifetime partner to official events, is now a greatly respected senior gay man. His name has been suggested as a possible next governor (Queen's representative) of New South Wales (NSW).

Attitudes

Antidiscrimination laws introduced by gay liberation activists in the 1970s and 1980s have had long-term effects on attitudes; even though Australians with less education, from non-English-speaking backgrounds, and people with a religion are less likely to be tolerant of homosexuality (Rissel et al., 2003a). The concentration of openly gay people in public service jobs and academe – where antidiscrimination is perhaps taken more seriously than in the private sector – has led to slow changes such as state health department funding for antiviolence projects

to protect gay people and to the normalization of same-sex relationships within the *de facto* category. The conservative government of Prime Minister John Howard altered the *Marriage Act* and the *Family Law Act* in 2004 to explicitly exclude same-sex partnerships from legal marriage. In response, campaigns for legal same-sex marriage have achieved majority public support even though the main practical reasons for campaigning for gay marriage in the US do not apply in Australia, as it is already illegal to discriminate against a person (e.g., by refusing to let a dwelling, or having different rules for insurance entitlements) on the grounds of (homo)sexuality or marital status. Members of same-sex couples are assessed on their joint incomes when applying for means-tested government benefits, just as heterosexuals are, whether legally married or not.

Although surveys in Australia have shown that Australians are more tolerant of abortion and homosexual sex than Britons and Americans (Michael et al., 1995; Scott, 1998; Johnson et al., 1994), they respond equally disapprovingly to questions about people in committed relationships having sex with someone else. For example, in ASHR, 78% of respondents agreed that "Having an affair when in a committed relationship is always wrong" (Rissel et al., 2003a), and in the 2009 Australian Survey of Social Attitudes, 89% agreed that "a married person having sexual relations with someone other than his or her husband or wife" was always or almost always wrong. This is largely matched by behavior. In ASHR, only 5% of men and 3% of women in regular relationships for more than a year had had sex with anyone other than the regular partner in the past year (Rissel et al., 2003b).

Sexual Health and Sexual Illness

Australia has received worldwide recognition for its response to the emergence of HIV/AIDS in the 1980s. Due to a combination of favorable circumstances, and the early actions of a few people who fought for legal and health policy changes such as antidiscrimination law and freely available injecting equipment (Baxter, 1992; Blewett, 2003), Australia does not have a widespread HIV epidemic. The strong response by the gay community led to a drop in partner numbers and support for condom use before government information and prevention programs were in place. Despite recent increases in diagnoses, the prevalence of infection remains around 1 in 1000 (Kirby Institute, 2013). Even among prisoners, HIV rates are low.

Early preventive activity to encourage 100% condom use among sex workers was successful, and STI testing rates remain high. It has been said that chlamydia rates are higher in Canberra schoolgirls than among sex workers in Australia (Professor Basil Donovan, at the Australasian Sexual Health Conference 2010). Some brothels and workers, with the support of funded sex worker health organizations, have attempted to normalize condom use for fellatio as well as for vaginal (or anal) intercourse, but compliance with this is mixed, with some establishments advertising 'bareback blow jobs'.

Chlamydia notification rates have been climbing inexorably since tests such as the polymerase chain reaction test became available in the late 1980s, allowing testing of urine samples. Chlamydia is now the most notified communicable disease, much higher than the next most commonly notified infections, gastroenteritis in an institution and influenza. Although the ongoing rise in chlamydia does not appear to have led to a corresponding increase in hospital admissions for pelvic inflammatory disease, chlamydia notifications act as an indicator of exposure to infection, and are unsurprising given the high rates of unprotected sex, especially among heterosexuals.

Australia's success in responding to HIV/AIDS and addressing STIs, especially among homosexually active men, has come at a cost. A focus on disease means that program success can be measured, but it has also meant that broader sexual health and well-being is neglected. In most states, policies for HIV/STI control remain separately from family planning, sex education, maternal health, and sexual assault. The exception is indigenous sexual health policies, which acknowledge that a 'holistic' or integrated approach is necessary.

Research and prevention programs focus on high-risk groups such as gay men, sex workers, and indigenous people and to a lesser extent adolescents and young people, largely neglecting the sexual concerns of the huge majority of people in Australia. This focus worked well in the early years of the HIV/AIDS response, as the epidemic was focused among gay men. The desire to control the sexuality of young people is apparent, as elsewhere, though somewhat more realistic, with only minor influence of the US abstinence push in the 2000s.

Contraception

In ASHR of women exposed to the risk of pregnancy, 95% used some form of contraception (Richters et al., 2003). Of those who did not, the most common reason was some version of an abdication of deliberate pregnancy planning: 'leave it to chance', 'do not care', 'forget', or 'have never got pregnant'. The next most common reason was experience of side effects or contraindications. No woman said she wanted contraception but could not access it. Prescriptions for low-income patients are subsidized, making hormonal contraceptives relatively accessible. Higher use of sterilization in regional and remote areas points to a possible difficulty with access for temporary methods.

In a country where about 50% of people report no religion in telephone surveys (though 78% nominated one in the 2011 Census, presumably because they regarded it as an official document that should reflect their baptism or confirmation status), religion is not a powerful influence on social policy. In ASHR, no women explicitly cited religious objections as a reason for not using contraception – though discomfort of religious origin might be expressed through general reluctance to plan pregnancies or a lingering fear that contraception is unnatural or unhealthy (Richters et al., 2003).

Although there are still people whose contraceptive practice (or lack of it) dismays doctors, it could be argued that many women in Australia have fewer children than they desire rather than too many (Cannold, 2005; Macken, 2005).

Childbearing

Rising rates of recourse to assisted conception techniques including *in vitro* fertilization are influenced by Australia's

pattern of delayed childbearing, which is common to other developed countries where effective contraceptives are easily available. The median age of first-time mothers increased by 2 years over the 1990s and 2000s. Since 2005, more babies have been born to women of 35–39 years of age (once regarded as 'elderly primiparas') than to women of 20–24 years of age. The effect of this is increasingly to make early childbearing a phenomenon concentrated among working-class or underclass women, certain ethnic communities and indigenous women. The higher maternal and perinatal risks associated with early childbearing, as in most women who bear children early tend to have poor social support, give ammunition to those arguing that early childbearing (whether 'early' means before 16, 18, 20, or even 25 years) should be prevented, whatever the childbearing choices of the young women concerned, who may well value babies above further education, or realistically estimate their chance of employment as very low. In other words, their opportunity costs for childbearing are much lower than for the educated middle class.

Abortion

Legal abortion is available in all states, though it remains on the Crimes Act in some states, legalized only by court precedent when performed by a medical practitioner. Thus moral debates about abortion affect its availability and cost. Abortion is not treated in the same way as other gynecological procedures and may incur out-of-pocket costs to the patient of several 100 dollars.

Abortion statistics are hard to come by, as there is no formalized recording mechanism and many abortions are paid for privately without recourse to claims on Medicare. A rough estimate is that in Australia about half of pregnancies are intended, and about half of unintended pregnancies are terminated. Termination rates – i.e., the proportion of total pregnancies that are deliberately terminated – are higher in young women under 25 and women over 45, presumably because pregnancies are less likely to be welcomed by women still in full-time education or not yet in a long-term relationship, and by women who had completed childbearing many years earlier.

The issue has been avoided by politicians for decades as it is seen as a 'no-win' topic, likely only to alienate some noisy opponents without gaining extra support from other voters. Partly as a result of controversy about abortion, there is a very heavy emphasis in clinical practice on fail-safe contraceptive use even for women in whom pregnancy would be very unlikely.

Sex Education

Some form of sex education is part of the school syllabus in all Australian states, but the reality of provision varies widely. Science courses cover mammalian and human reproduction, but lack practical information such as conception and do not engage with issues such as sexual harassment or homosexuality.

In NSW, for example, the high school Personal Development, Health and Physical Education syllabus includes material on sexual health, including STI prevention and contraception, but this is not taught in all schools. As well as state schools, attended by 65% of the high school (12–18 years of age) population, there are systemic Catholic schools and independent schools, often also run by religious institutions, which sometimes do not teach sex education at all, or present reproductive information solely as something that students may need later in their married lives.

Teenagers complain that they are taught 'plumbing' but that it is 'scientific' and not useful to real life (McKee et al., in press). Thus although excellent material for teaching sexuality education was developed in the 1980s and earlier by family planning associations and other bodies, its emphasis on values clarification and engagement with sexual diversity has not been carried through into the reality of most adolescents' time at school even today.

Young People

National surveys of the sexual behavior and attitudes of secondary school students have been conducted every 5–6 years since 1992. These have reflected quite high levels of knowledge about HIV/AIDS – presumably a result of the mandated teacher training and school programs of the late 1980s and the 1990s – but less knowledge about other STIs (Smith et al., 2009).

Despite such programs, folk concepts of risk are resilient, so that even educated young adults cling to the notion that one cannot acquire an STI from someone who is well-off and clean, or in one's own social circle (Crosbie et al., 2013). This leads to high rates of unprotected intercourse at all ages, in the context of high rates of sex with casual partners and sequential monogamy (one-on-one relationships) without STI testing in between. Even for sex that people themselves define as 'casual', only about 40% of heterosexuals always use condoms. Gay men are more diligent condom users; 87% always use condoms for anal sex with casual partners (de Visser et al., 2003a). In gay community surveys, of those with casual partners, 38% report having had any unprotected anal intercourse in the previous 6 months (de Wit et al., 2013).

The school surveys have also revealed a pattern of experience of sexual practices – or acquisition of sexual skills, to put it another way – over the teen years. A majority (71%) of Year 10 students (aged 15–16) surveyed in 2008 had experienced sexual kissing, but only 27% had had intercourse. Of Year 12 students (aged 17–18), 56% had had intercourse and 58% had had oral sex. Most students reported positive feelings about sex, but 38% of young women reported having had unwanted sex at some point (Smith et al., 2009).

Students who leave school earlier, both those who drop out and those who leave to attend vocational college, tend to initiate sex earlier. University students tend to have first intercourse at a later age, and a significant proportion, especially students from Asian backgrounds (many of whom have attended school overseas) are virgins when they commence university. Many have inaccurate knowledge about HIV risks in their home countries (Song et al., 2005) and may be ill-equipped to operate as sexually active independent adults in an Australian university context, for example, by seeking prescriptions for

contraception (Burchard et al., 2011). Australia has large numbers of overseas tertiary students, mostly from eastern and southeast Asia.

National surveys of adults have revealed that the age at first intercourse has fallen over the past several decades, so that people who were born in the 1940s had a median age at first intercourse of 18 years for men and 19 years for women, but for those born in the 1980s the median age was 16 years (Rissel et al., 2003c). As 16 years is the legal age of consent to intercourse in most states, this means that a large minority of teenagers start having intercourse at an age when it is technically illegal. Although many doctors and family planning clinics would be unlikely to make an issue of this unless they thought the teen patient was at serious risk of abuse (Bartholomew, 2009), teenagers may not be aware of this. A disjunction between socially normal (i.e., common, even if not publicly accepted) behavior and the law remains.

In recent decades, young people have been more likely to have oral sex, and some initiate it before they have had intercourse, seeing oral sex as sexual play or 'not real sex'. This is replacing an older view of noncoital practices (except perhaps for manual sex or 'petting' as foreplay) as more risqué and more forbidden than intercourse itself, despite the risks of pregnancy and much higher risks of disease transmission associated with coitus. Anal intercourse is a less common practice, though again the cohort born after 1950 is more likely to have tried it. One reason for the increase may be that the response to HIV/AIDS in the 1980s brought both condoms and anal intercourse into public discourse. Smutty talk about 'frenchies' or Wetchex (a brand name) of condoms and hushed mentions of 'unnatural practices' in divorce cases reported in afternoon papers were replaced by open arguments and information about HIV prevention policy in broadsheet newspapers and on television.

Even parents and teachers approaching sexual issues from a feminist perspective seem to fear oral sex as a possible locus for male exploitation in demanding unreciprocated or unrewarding fellatio from girls, rather than seeing it as a form of sexual play in which girls may experiment sexuality while avoiding both the physical risks of intercourse and its social importance in defining relationship status – or conversely defining girls as sluts if they have intercourse outside a girlfriend–boyfriend context. The double standard is receding; for example, young women report more similar sexual practices and partnering to young men (Grunseit et al., 2005; de Visser et al., 2003b). But qualitative studies repeatedly encounter stories of 'slut-shaming' that constrain young women's sexual agency while rewarding young men for 'scoring' (Albury, 2002). Fear of being thought a slut can make it difficult for young women to openly buy or carry condoms, access adult advice, attend doctors or negotiate fairly with males.

Teenagers from 15 years of age may request their own Medicare card to access subsidized health care without their parents' involvement, though many are not aware of this. Prescriptions for low-income patients are further subsidized, making oral contraceptives relatively accessible. Nonetheless, for teenagers, the cost – and scariness – of condoms, doctors' visits, and prescriptions are a disincentive to contraceptive use, particularly in rural and remote areas where confidentiality is a problem.

The drop in the age of first intercourse is less significant to sexual risk patterns, however, than the rise in the age of marriage or entry into a long-term committed relationship. In the 1950s and 1960s, many people married the first or second person they had sex with (Cook, 2004). Nowadays this is rare, and young adults expect a decade or more of freedom and 'playing the field' before settling down. (This may also be affected by increasing housing unaffordability in cities; high home ownership among young couples, long a component of the 'Australian Dream', is now a thing of the past.) This change is reflected in the lower lifetime numbers of sexual partners reported by those currently in their 60s than those in their 30s, even though the older people have had longer in which to acquire them (de Visser et al., 2003b; unpublished ASHR2 data).

Australia's teenage birth rate in 2009, at 16.5 per 1000 girls of 15–19 years of age, was not as high as that in the UK (25) or the US (42), though considerably higher than Korea, China, Japan, Sweden, and Italy (3.5–6.8) (United Nations, 2013).

Indigenous Sexual Health

Demographically and politically, Australia is an immigrant country formed over the last 200 years, originally by British colonists and settlers, and carries the history of British law and institutions. Only about 2% of the population identifies as indigenous – Australian Aboriginal or Torres Strait Islander – and it is unclear how or to what degree indigenous sexual cultures have influenced mainstream Australian attitudes to sex. Precontact traditional Aboriginal societies were patriarchal and violent; women were promised in marriage, often in childhood, according to the rules of complex kinship patterns. William Buckley, an escaped convict who lived for years among the Aboriginals in the early nineteenth century, reports many of the violent skirmishes he witnessed among Aboriginals as being due to the escape or capture of a young woman by one group or another. It is arguable how much of this intergroup violence was partly produced by social disruption as a result of colonization.

Today's indigenous Australians have considerably worse health on average than the average Australian, resulting in markedly lower life expectancy. Sexual health is also worse, both in urban indigenous people and in rural settings. This disadvantage in sexual health is exacerbated by lack of health services in remote communities (Fairley et al., 1997).

High rates of casual sex – though with no higher numbers of partners than the general population – may put Aboriginal people at higher risk (Bryant et al., 2011). High mobility may also play a part in the epidemiology of STIs in Aboriginal communities (Hui et al., 2013). Consistent condom use in casual sex is low, but no lower than in the general population (Bryant et al., 2011).

It has also been argued that traditional health cultures of more remote Aboriginals, developed over millennia among naked, nomadic tribes, are not conducive to health in settled communities (Sutton, 2006). It is possible that traditional cultures involving separation of 'men's business' and 'women's business', notions of 'shame' and violent ways of dealing with emotional conflict, continue to affect Aboriginal sexual health,

especially women's health. Sexual and domestic violence, together with drug and alcohol problems, are serious issues for many Aboriginal communities (Arabena, 2006; Larkins et al., 2007).

Sex Work

Prostitution is treated differently by the law in each Australian state. Models vary from a decriminalized system to one where all sex work is illegal. Victoria has a system where some brothels are regulated and licensed, ensuring better working conditions and access to free condoms, but where many unlicensed brothels operate illegally and are difficult for outreach services and researchers to access. Decriminalization is associated with much better access by sexual health support services (Harcourt et al., 2010; Seib et al., 2009).

In NSW, the Sex Workers Outreach Project, operating under the auspices of the AIDS Council with funding from the state government, is a peer-run organization with staff (some of whom speak Asian languages) who have very good access to brothels of all types and to independent workers (Harcourt et al., 2010). Decriminalization has not been without its problems. In NSW, local authorities must authorize land use for all businesses, and some local councils have sought to refuse brothels planning permission on 'moral' grounds. Some businesses that are denied permission have continued to operate in breach of planning rules, and some have subsequently won cases in the Land and Environment Court appealing against council decisions (Julie Bates, Urban Realists, pers. comm.).

The Australian experience makes clear that legalizing sex work does not increase the size of the industry, which is driven by many factors such as client demand (not high in the general population in Australia, given the stigma of paying for sex and the ready availability of amateur sex), supply of potential workers (influenced by population levels of education, job availability, and immigration), and specific local needs such as high proportions of single men, for example in mining. About one man in six (15.6%) among men of 16–59 years of age said in the ASHR survey in 2001–02 that they had ever paid for sex, and only 1.9% had done so in the past year; rates by state did not suggest that illegality was a deterrent (Rissel et al., 2003d). The small size of the industry and the growing acceptability of sex work have led to better research, including research into clients. Unlike much of the research internationally, it does not assume that being a sex worker is necessarily both damaging and evidence of prior damage, not so much a job but a stigmatizing risk factor or poor health outcome in itself. Even though many sex workers have arrived in the industry with a history of sexual abuse, homelessness, or drug use, sex work can be a way of reclaiming agency in relations with men that is preferable to a self-image as a victim. Many of the dangers of sex work are direct results of its illegality, so that in criminalized settings workers cannot telephone the police if a client becomes violent, and are subject to police corruption such as demands for free services accompanied by threats of arrest.

The uneven but increasing respectability of sex work has led to attempts to make sexual services more readily available to people (usually men) who seek assistance because of illness or disability. At this specialized and professionalized end of the industry, there is an overlap between sex work and sexual surrogacy; some workers do both.

Homosexuality

Australia's pattern of social understandings of homosexuality is shared with countries of northwestern Europe. That is to say, attitudes to homosexuality and to gay men in particular are different from those around the Mediterranean, the middle East, Asia, Africa, and Latin America, in which men are more likely to be expected to marry women whatever their personal sexual tastes, and people think less of men who do not do so, and regard openly homosexual men as unmanly (Parker et al., 2000).

In Australia, at least among the educated middle class of largely Anglo-Celtic backgrounds, homosexuality is seen as primarily an issue of being attracted to people of the same sex. Educated people mostly regard gay men and lesbians as a variant of normal, just like heterosexuals except that they are attracted to people of the same sex. Thus two women in a same-sex relationship are both seen as lesbian and both men in a same-sex relationship are seen as gay, whether or not one of them is noticeably more dominant or masculine in manner and appearance.

Homosexuality is popularly regarded as an intrinsic element of personality, and as something one is born with, rather than as immorality or illness. Thus people's tolerance of same-sex acts and of homosexuality as identity or preference tends to overlap. In ASHR, 37% of men agreed that sex between adult men was always wrong, reflecting a lesser taboo on same-sex acts between women but only 21% that sex between adult women was always wrong. Among women, about a quarter disapproved of same-sex acts, whether between men or women (Rissel et al., 2003a).

Although not many same-sex-attracted people nowadays grow up imagining themselves to be the only one of their type, some (particularly in rural areas) may perceive that their only choices are either to join the flamboyant urban gay life of the Sydney Gay and Lesbian Mardi Gras (if they were accepted by the gay community) or to hide their feelings in order to be accepted by their peer groups. Over recent years, opportunities for making contact with potential sexual partners via the Internet and social media have reduced the cohesion of the gay community, which was once dependent upon gay community newspapers and venues in 'gay' districts. At the same time, the Internet and social media have made it easier for isolated young people to find out about others like themselves, to make contact with community groups, and to obtain health-related information.

Lesbians tend to be less visible to health and social policy than gay men, and face difficulties in getting their community needs met. For example, smoking rates are much higher among lesbians than other women, yet no community-specific smoking reduction or cessation programs have been run for lesbians (Mooney-Somers et al., 2012).

Despite the public and media attention given to homosexuality, there are more adult virgins than exclusive homosexuals (Smith et al., 2003b). Celibacy has been called the 'new deviancy' (Hart and Wellings, 2002), and this is probably also true in Australia.

See also: Australia, Sociocultural Overviews: Australian Settler Society; Australian Aboriginal Society and Culture: An Overview; Children, Rights of: Cultural Concerns; Intimate Partner Violence; Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sexual Risk Behaviors; Sexuality: Cultural Aspects.

Bibliography

Albury, K., 2002. Yes Means Yes: Getting Explicit about Heterosex. Allen & Unwin, Sydney.

Arabena, K., 2006. Preachers, policies and power: the reproductive health of adolescent Aboriginal and Torres Strait Islander peoples in Australia. Health Promotion Journal of Australia 17, 85–90.

Bartholomew, T., 2009. How general practitioners determine young people's rights. Youth Studies Australia 28, 5–13.

Baxter, D., May 14, 1992. Opening address at 'AIDS—have we got it right?', Australian Doctors Fund conference, Sydney. www.adf.com.au/archive.php?doc_id=100 (accessed 14.02.07.).

Blewett, N., 2003. AIDS in Australia: The Primitive Years. Reflections on Australia's Policy Response to the AIDS Epidemic. Australian Health Policy Institute, University of Sydney, Sydney.

Bongiorno, F., 2012. The Sex Lives of Australians: A History. Black Inc., Melbourne.

Bryant, J., Ward, J., Worth, H., Hull, P., Solar, S., Bailey, S., 2011. Safer sex and condom use: a convenience sample of Aboriginal young people in New South Wales. Sexual Health 8, 378–383.

Burchard, A., Laurence, C., Stocks, N., 2011. Female international students and sexual health: a qualitative study into knowledge, beliefs and attitudes. Australian Family Physician 40, 817–820.

Cannold, L., 2005. What, No Baby? Why Women Are Losing the Freedom to Mother, and How They Can Get It Back. Fremantle Arts Centre Press, Fremantle, WA.

Cook, H., 2004. The Long Sexual Revolution: English Women, Sex, and Contraception 1800–1975. Oxford University Press, Oxford.

Crosbie, K., Richters, J., Hooker, C., Leask, J., 2013. Filthy fingernails and friendly germs: lay concepts of contagious disease transmission in developed countries. In: Banwell, C., Ulijaszek, S., Dixon, J. (Eds.), When Culture Impacts Health: Global Lessons for Effective Health Research. Elsevier, San Diego, CA, pp. 67–83.

Fairley, C.K., Bowden, F.J., Gay, N.J., Paterson, B.A., Garland, S.M., 1997. Sexually transmitted diseases in disadvantaged Australian communities [letter]. Journal of the American Medical Association 278, 117–118.

Grunseit, A., Richters, J., Crawford, J., Song, A., Kippax, S., 2005. Stability and change in sexual practices among first-year Australian university students (1990–1999). Archives of Sexual Behavior 34, 555–566.

Harcourt, C., O'Connor, J., Egger, S., Fairley, C.K., Wand, H., Chen, M.Y., Marshall, L., Kaldor, J.M., Donovan, B., 2010. The decriminalisation of prostitution is associated with better coverage of health promotion programs for sex workers. Australian and New Zealand Journal of Public Health 34, 482–486.

Hart, G., Wellings, K., 2002. Sexual behaviour and its medicalisation: in sickness and in health. British Medical Journal 324, 896–900.

Hui, B.B., Gray, R.T., Wilson, D.P., Ward, J.S., Smith, A.M.A., Philip, D.J., Law, M.G., Hocking, J.S., Regan, D.G., 2013. Population movement can sustain STI prevalence in remote Australian indigenous communities. BMC Infectious Diseases 13, 188.

Johnson, A.M., Wadsworth, J., Wellings, K., Field, J., Bradshaw, S., 1994. Sexual Attitudes and Lifestyles. Blackwell Scientific Publications, Oxford.

Kirby Institute, 2013. HIV, Viral Hepatitis and Sexually Transmissible Infections in Australia: Annual Surveillance Report 2013. Kirby Institute, University of New South Wales, Sydney. www.kirby.unsw.edu.au/surveillance/2013-annual-surveillance-report-hiv-viral-hepatitis-stis (accessed 14.02.07.).

Larkins, S.L., Page, R.P., Panaretto, K.S., Scott, R., Mitchell, M.R., Alberts, V., McGinty, S., 2007. Attitudes and behaviours of young indigenous people in Townsville concerning relationships, sex and contraception: the 'U Mob Yarn Up' project. Medical Journal of Australia 186, 513–518.

Laumann, E.O., Gagnon, J.H., Michael, R.T., Michaels, S., 1994. The Social Organization of Sexuality: Sexual Practices in the United States. University of Chicago Press, Chicago and London.

Macken, D., 2005. Oh No, We Forgot to Have Children! How Declining Birth Rates Are Reshaping Our Society. Allen & Unwin, Sydney.

McKee, A., Watson, A.-F., Dore, J., in press. 'It's all scientific to me': focus group insights into why young people do not apply safe sex knowledge. Sex Education.

Michael, R.T., Gagnon, J.H., Laumann, E.O., Kolata, G., 1995. Sex in America: A Definitive Survey. Warner, New York.

Mooney-Somers, J., Deacon, R.M., Richters, J., Price, K., Léon de la Barra, S., Schneider, K., Parkhill, N., 2012. Women in Contact with the Sydney Gay and Lesbian Community: Report of the Sydney Women and Sexual Health (SWASH) Survey 2006, 2008 and 2010. ACON and Centre for Values, Ethics and the Law in Medicine (VELIM), University of Sydney, Sydney.

Parker, R., Barbosa, R.M., Aggleton, P., 2000. Framing the Sexual Subject: The Politics of Gender, Sexuality, and Power. University of California Press, Berkeley, CA.

Richters, J., Grulich, A.E., de Visser, R.O., Smith, A.M.A., Rissel, C.E., 2003. Sex in Australia: contraceptive practices among a representative sample of women. Australian and New Zealand Journal of Public Health 27, 210–216.

Richters, J., de Visser, R.O., Grulich, A.E., Rissel, C.E., Smith, A.M.A., 2008. Demographic and psychosocial features of participants in bondage and discipline, 'sadomasochism' or dominance and submission (BDSM): data from a national survey. Journal of Sexual Medicine 5, 1660–1668.

Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., Smith, A.M.A., 2003a. Sex in Australia: attitudes toward sex in a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 118–123.

Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., Smith, A.M.A., 2003b. Sex in Australia: selected characteristics of regular sexual relationships. Australian and New Zealand Journal of Public Health 27, 124–130.

Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., Smith, A.M.A., 2003c. Sex in Australia: first experiences of vaginal intercourse and oral sex among a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 131–137.

Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., Smith, A.M.A., 2003d. Sex in Australia: experiences of com sex in a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 191–197.

Scott, J., 1998. Changing attitudes to sexual morality: a cross-national comparison. Sociology 32, 815–845.

Seib, C., Debattista, J., Fischer, J., Dunne, M., Najman, J.M., 2009. Sexually transmissible infections among sex workers and their clients: variation in prevalence between sectors of the industry. Sexual Health 6, 45–50.

Smith, A.M.A., Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., 2003a. Sex in Australia: the rationale and methods of the Australian Study of Health and Relationships. Australian and New Zealand Journal of Public Health 27, 106–117.

Smith, A.M.A., Rissel, C.E., Richters, J., Grulich, A.E., de Visser, R.O., 2003b. Sex in Australia: sexual identity, sexual attraction, and sexual experience among a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 138–145.

Smith, A., Agius, P., Mitchell, A., Barrett, C., Pitts, M., 2009. Secondary Students and Sexual Health 2008: Results of the 4th National Survey of Australian Secondary Students, HIV/AIDS and Sexual Health. Australian Research Centre in Sex, Health & Society, La Trobe University, Melbourne.

Song, A., Richters, J., Crawford, J., Kippax, S., 2005. HIV and sexual health knowledge and sexual experience among Australian-born and overseas-born students in Sydney. Journal of Adolescent Health 37, 243e9–243e14.

Sutton, P., 2006. The politicisation of disease and the disease of politicisation: causal theories and the indigenous health differential. 8th National Rural Health Conference. http://nrha.ruralhealth.org.au/8thNRHC/Papers/KN_sutton,%20peter.pdf (accessed 14.02.07.).

United Nations, 2013. Table 10. Live births by age of mother and sex of child, general and age-specific fertility rates: latest available year, 2000–2009. In: World Summary, Demographic Yearbook. http://unstats.un.org/unsd/demographic/products/dyb/dyb2009-2010/Table10.pdf (accessed 14.02.07.).

de Visser, R.O., Smith, A.M.A., Rissel, C.E., Richters, J., Grulich, A.E., 2003a. Sex in Australia: safer sex and condom use among a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 223–229.

de Visser, R.O., Smith, A.M.A., Rissel, C.E., Richters, J., Grulich, A.E., 2003b. Sex in Australia: heterosexual experience and recent heterosexual encounters among a representative sample of adults. Australian and New Zealand Journal of Public Health 27, 146–154.

de Wit, J., Mao, L., Holt, M., Treloar, C. (Eds.), 2013. HIV/AIDS, Hepatitis and Sexually Transmissible Infections in Australia: Annual Report of Trends in Behaviour 2013. Centre for Social Research in Health, University of New South Wales, Sydney. https://csrh.arts.unsw.edu.au/media/CSRHFile/1_ARTB_2013_report.pdf (accessed 14.02.07.).

Relevant Websites

www.abs.gov.au – Australian Bureau of Statistics.
www.afao.org.au – Australian Federation of AIDS Organisations.
www.latrobe.edu.au/arcshs – Australian Research Centre in Sex, Health & Society.
www.ashr.edu.au – Australian Study of Health and Relationships.
http://aussa.anu.edu.au/data.php/ – Australian Survey of Social Attitudes.
https://csrh.arts.unsw.edu.au/ – Centre for Social Research in Health.
https://www.kirby.unsw.edu.au – Kirby Institute for Infection and Immunity in Society.
www.shfpa.org.au – Sexual Health and Family Planning Australia.
www.swop.org.au – Sex Workers Outreach Project.
www.mardigras.org.au – Sydney Gay and Lesbian Mardi Gras.
www.touchingbase.org – Touching Base.

Sexuality Over the Life Course

Linda Waite, University of Chicago, Chicago, IL, USA
Cameron Charme, NORC, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on the development of sexuality during childhood and how it is transformed over the life course. We discuss relevant findings from research on sexuality in childhood, through young adulthood and midlife, to older adulthood. We point to sexual scripts, the socially learned set of sexual desires and conduct, as a lens for understanding sexuality (*see* Scripts of Masculinity). We highlight life transitions and turning points, sexual agency, cumulative advantage and disadvantage, and the role of culture, gender, race, ethnicity, and social class in experiences of sexuality over the life course.

People continuously acquire and interpret emotional, physical, and intellectual experiences, which shape their lives in various ways. Events that occur early in life can directly or indirectly influence behavior and attitudes later. The environment in the home, race/ethnic background, social class, sexual identity, and religious beliefs have the potential to dramatically influence life trajectories (Carpenter and DeLamater, 2012). In turn, these factors and others shape sexuality over life stages, beginning at childhood and adolescence and continuing through menopause and old age.

Life transitions, either socially positive or negative, involve the movement from one social role to another. Major transitions that affect sexuality include sexual debut, cohabitation and marriage, the birth of children, relationship disruption, and the illness and death of one's spouse or partner (*see* Sexual Debut; Marital Quality and Health). These turning points have the potential to alter sexual attitudes, values, and behavior. The Gendered Sexuality over the Life Course (GSLC) Model, developed by Carpenter and DeLamater (2012) moves beyond life cycle stages to focus on the processes through which sexuality unfolds throughout life. This model will be used to highlight the interrelationships between different stages of sexual development. The GSLC model suggests a consideration of the impact of combinations of life transitions and experiences, either positive or negative. For example, sexual abuse as a child may act as a turning point and may interact with other experiences or transitions. Browning and Laumann (1997) showed that women who were sexually abused as children were more likely than nonabused children to have consensual sex before they were 16, bear children before 18, and contract sexually transmitted infections (*see* Sexually Transmitted Infections: Social Network Analysis). The authors conclude that each negative life transition reinforced negative sexual scripts learned from past experiences and redirected sexual trajectories; whereas, the positive transitions reduced disadvantages and sometimes created advantages for future choices.

Sexual Agency in a sociohistorical context is another key element of the GSLC model. Sexual agency describes how people attempt to shape their life trajectories in both the long term and short term to reflect their values and goals. Thus, a heterosexual single male who does not believe in waiting for marriage for sex enacts sexual agency by actively seeking sexual partners, whereas a woman who believes in waiting to have sex until marriage enacts sexual agency by declining his advances. *Sociohistorical context and generation*, both elements of the GSLC model, are important factors in sexual attitudes and behavior (Joyner and Laumann, 2001). The sexual and gender revolutions of the 1960s transformed sexual attitudes and behaviors, ushering in a period of liberal sexual values that led to earlier sexual debut, increases in unmarried childbearing, more frequent relationship turnover, and a wider variety of sexual practices than in previous generations (Das et al., 2012)

Childhood Sexuality

Although the life course of sexuality begins in childhood, most research has focused on adolescent and adult sexuality. Very little attention has been paid to normative or typical sexual behavior in young children (Thigpen, 2012). The literature that has focused on childhood sexuality puts a strong emphasis on adult–child sexual contact and sexual aggression in childhood. Many studies have found evidence of long-term social and emotional problems throughout life as a result of child–adult sexual contact. Child-to-child sexual aggression, abusive behavior that has both sexual and aggressive components committed by a child upon another child, has similar negative effects. It is thought that the majority of child-to-child aggressors are acting as a result of their own sexual victimization (Thigpen, 2012).

Normative Sexuality in Childhood

Recent studies of children aged 2–12 have expanded our knowledge of normative sexual behavior at younger ages. These studies show that, in early childhood, children typically engage in self-stimulatory behaviors, are interested in sexual topics, and show interest in viewing the genitals of the same and opposite sex. Genital manipulation and masturbatory behaviors are higher among boys than girls (Thigpen, 2012). As children become older, they develop physically, cognitively, and socially in ways that foster interest in and knowledge of sexual behavior. They also form and revise attitudes toward sex. Parenting behavior and attitudes affect those of children;

liberal parenting styles are associated with increased masturbation, higher frequency of touching parent's genitals, and increased overall sexual behavior (Thigpen, 2012). This highlights the importance of context for children's sexual behavior.

Cultural and Social Differences in Childhood Sexuality

Sexual behavior in childhood is shaped by biological, psychological, and social factors operating within specific cultural, social, and political contexts. Cross-cultural comparisons can reveal the effects of these contexts. Goldman and Goldman (1982) compared sexual knowledge among Australian, English, North American, and Swedish children aged 5–15. They found that Swedish children were most knowledgeable about sexuality, perhaps due to sexuality education in schools combined with permissive sexual beliefs and attitudes (*see* Sexuality Education). In the American context, sexual beliefs and attitudes have been shown to vary across socioeconomic, racial, and ethnic groups. African-American children tend to make their sexual debut at a younger age than white, Hispanic, or Asian children. They also tend to have higher rates of sex in adolescence and have more partners in adulthood than whites and Hispanic Americans (*see* Sexual Debut; Thigpen, 2012). Two large-scale studies compared the sexual behavior of African-American and white children aged 7–12 using the Child Sexual Behavior Inventory (CSBI), a parental report of child sexual behavior (Friedrich et al., 1991; Thigpen, 2009). The CSBI consists of 38 measures in 9 sexual behavioral domains. These domains include boundary problems, exhibitionism, gender role behavioral, self-stimulation, sexual anxiety, sexual interest, sexual intrusiveness/aggression, sexual knowledge, and voyeuristic behavior (Friedrich et al., 2001). Parents of white children reported that they displayed 19 sexual behaviors compared to 13 for African-American children. Self-stimulatory behaviors, consisting of genital touching and masturbation, were significantly more prevalent among white boys, whereas among African-American boys it was more common to signify an interest in sex or in girls, such as talking flirtatiously or wanting to watch nudity or sex on TV (Thigpen, 2012). White girls showed a broader range of behaviors and a higher prevalence of sexual behaviors. Self-stimulation and voyeuristic behavior, such as wanting to look at nude photos, was higher among white girls. African-American girls had a higher prevalence for wanting to watch movies that included nudity or sex (Thigpen, 2012).

Sexuality Development in Adolescence

Adolescence is an extremely important stage in the sexual life course (*see* Adolescent Sexual Risk). Biological and physiological changes occur rapidly during puberty, as do changes in/development of cognitive and social skills. During adolescence, boys and girls develop and hone their ability to analyze interpersonal relations and social organization. Most important, during adolescence children and teens gain an understanding of the social and cultural meaning of sexuality. They also acquire sexual scripts. Sexual debut may take place in adolescence or may be delayed until the late teens or young adulthood (*see* Sexual Debut). The timing of sexual debut is influenced by the adolescent's self-esteem, intentions, and drug use. Those who have low self-esteem tend to delay the onset of sexual behaviors. Likewise, teens who experiment with drugs and alcohol face an increased likelihood of early sexual debut (*see* Sexual Debut). Higher intelligence operates as a protective factor against early sexual activity during adolescence, and lower intelligence, to a point, is a risk factor. Higher intelligence is associated with postponement of the initiation of the full range of partnered sexual activities, from holding hands to intercourse (Halpern et al., 2000).

Physiological maturity affects the age at sexual debut; those with earlier age of menarche and spermarche tend to transition into partnered sexual behavior earlier than others (*see* Sexual Debut). Adolescent children are reaching sexual maturity at a much earlier age than their counterparts did a century ago, especially in the United States (Russell et al., 2012). Advances in nutrition and hygiene have contributed to this change. In the United States, public policy laws were established to limit adolescent sexual activity, as it is seen as a social problem. With adolescents becoming sexually mature and sexually active at earlier ages, much of the research on adolescent sexuality has focused on the attendant problems, such as violence, problems for emotional health, sexually transmitted infection, and pregnancy. Joyner and Udry (2000) compare change in levels of depression for adolescents with and without romantic involvement. They find that males and females who become romantically involved experience a larger increase in depression than their counterparts who do not; however, females experience an even larger increase in depression than males in response to romantic involvement. These results suggest that females' greater vulnerability to romantic involvement explains a large part of the emerging sex differences in depression during adolescence.

Sexual Agency

Learning to wield sexual agency is an important developmental task in adolescence. Adolescents must learn to negotiate affectionate and intimate relationships with appropriate others as part of a successful transition to adulthood. They must decide on and manage a sexual debut and all that sexuality entails. Sexual agency, the belief in one's ability to act upon sexual needs in a relationship, is an essential tool for adolescents and adults (Russell et al., 2012). In the United States, high rates of sexual assault, coercion, unintended pregnancies, and sexually transmitted infections may reflect an absence of sexual agency (Russell et al., 2012). The development of sexual agency is influenced by the social context in which it occurs, including the race, ethnicity, and social class of the individual.

Cultural and Social Differences in Adolescent Sexuality

Sexual trajectories through adolescence differ substantially by race, ethnicity, and social class. African-American females initiate partnered sex at an earlier average age than whites and Latina females. Asian-American teens are the most likely to delay sexual debut. Latina adolescents are, on average, 4 years younger than their partners. This is a larger age gap between partners than seen among adolescents of other races. Latino

and African-American adolescents are less likely than white teens to use contraceptives. The African-American population has the highest risk of sexually transmitted infections, HIV/AIDS, and pregnancies, compared to Latinos and white Americans (Russell et al., 2012). However, the birth rate for African-American girls aged 15–19 fell dramatically from 118 per 1000 in 1991 to 47 per 1000 in 2011, and is now about the same as the rate for Hispanic teens. Still, both African-American and Hispanic teens have birth rates about twice as high as those for white teens and four times as high as Asian teens (Martin et al., 2011). Sexual trajectories also vary widely across social and political contexts. Compared to German teens, American teens use contraceptives less, have higher rates of pregnancy, and have more abortions (Russell et al., 2012).

Social class also plays a major role in the sexual development and health of adolescents. Research on social class often focuses on the poor because it is believed that these adolescents are more sexually active and at a greater risk for negative consequences than are those from more privileged backgrounds. Tolman (2002) studied girls from urban environments who had experienced racism and poverty. He found that these girls felt less comfortable talking about their sexuality among peers and parents than girls from suburban areas. Also, the majority of teens who have contracted AIDS are from poor families, victimized prostitutes, or runaways (Russell et al., 2012). Adolescents from poor families use contraceptives less frequently, which can result in higher pregnancy rates. Girls from higher socioeconomic status families may also experience hardships related to sexuality in their adolescent years. For privileged adolescents, the strains of emerging sexuality tend to take different forms; the emphasis on being thin in upper-middle class communities may lead girls to develop a negative self-image, develop eating disorders, and become malnourished (Russell et al., 2012).

Adolescents are generally taught about the physical changes to expect during puberty and recommendations for healthy sexual behavior. Formal sexuality education is usually implemented through school curricula and may also be offered by religious institutions or community groups. However, most sexuality education is informal, occurring through interactions with parents, peers, or others, and through television, movies, print, and electronic media. Through these interactions, and with varying effectiveness, adolescents develop attitudes, values, ideas, sexual scripts, and sexual agency (*see* Sexuality Education). Sexuality education in schools in the United States often does not explicitly include information on contraception or about how to regulate sexuality in the presence of relationships.

Intersecting Trajectories: The Effect of Divorce on Adolescents

The sexual life course trajectories of parents intersect with those of their adolescent children to affect their attitudes and behavior. Adolescents whose parents divorced are more likely to approve of premarital intercourse than are adolescents whose parents remained together, a finding shown consistently in a meta-analysis study done by Jeynes (2006). The findings showed larger effects on positive attitudes toward premarital intercourse than on actual engagement in premarital intercourse (Jeynes, 2012). Research shows that parental divorce acts as a turning point in the sexual life course of children. Children of divorced parents are more likely to have children themselves prior to marriage, more likely to be unmarried and unpartnered as young adults, and more likely to divorce if married (McLanahan and Sandefur, 1994; Waite and Gallagher, 2000).

Sexualities from Young Adulthood through Midlife

College and Casual Sex

The transition from high school to college, especially for those students who move from the parental home to college housing, is a critical period for young adults to reconsider and perhaps change their sexual values and behavior (Wade and Heldman, 2012). For many college students, sex, drugs, alcohol, and partying are the dominant nonacademic activities (Wade and Heldman, 2012).

'Hooking up' is a term used to describe, among college students, a casual sexual encounter that does not always involve intercourse (*see* Hooking up and Casual Sex). These hookups may involve alcohol and may involve strangers, acquaintances, or friends. The casual aspect of hooking up allows college students to be sexually active without the burden of establishing a relationship or friendship (*see* Hooking up and Casual Sex; Taylor, 2013). Previous research has assumed that the men were pushing the hookup culture, but recent social critics have argued that female college students may perceive benefits from casual sex and freedom from the time demands of more serious relationships (Rosin, 2012). A study of 60 college women at the University of Pennsylvania found that most women are focused primarily, at least for the short term, on academic and professional goals. The hookup culture provides a way for these women to have a (casual) sex life without taking the time away from their studies (Taylor, 2013).

Consequences of Casual Sex

The hookup culture may come with substantial consequences, especially for young women. Many women report that their male partners' sexual pleasure is prioritized over their own. By hooking up, women are also taking the risk of being labeled negatively if they engage in casual sex with multiple partners (*see* Hooking up and Casual Sex). College students are more at risk of becoming victims of sexual assault than their peers not enrolled in college (Wade and Heldman, 2012). A study of 48 first-year students enrolled in a sexuality-related writing course provided insight into their sexual experiences. Many students reported that the transition from high school to freshman year was difficult. Some of the participants who were virgins upon starting college stated that they engaged in intercourse to 'fit in.' Other women reported that it was nearly impossible to have a cross-gender friendship because the men would only think about sex (Wade and Heldman, 2012).

Alcohol and parties form the cornerstone of the hookup scene. Alcohol removes one's inhibitions and often leads students to partake in uncharacteristic behavior that is directly opposed to their values and beliefs (Wade and Heldman, 2012). Parties create opportunities for sexualized dancing

and making out, which often lead to hooking up (*see* Hooking up and Casual Sex). Many students reported engaging in intercourse or oral sex because they were drunk. Even when sexual encounters were consensual, many of the students described the experience as unsatisfying. The male's pleasure was often prioritized over the female's, and many young men seemed to expect intercourse from every sexual encounter (Wade and Heldman, 2012).

Midlife Years and Older Adulthood

The period of the life course between adolescence and old age tends to be treated as homogenous 'midlife' when, in fact, many changes take place between ages 20 and 60. These years can include the establishment of stable partnerships, marriage, birth of children, career employment, as well as the loss of these roles through divorce, job loss, children leaving home, and changes in sexual function and behavior (Das et al., 2012). Trajectories of sexuality tend to diverge during midlife for men and women, and by social and cultural factors such as race and ethnicity. Trajectories are also affected by changes in partnership, the onset of chronic diseases, and, declines in physical, social, and cognitive functioning (Lindau et al., 2007; Laumann and Waite, 2008). Those who maintain an active sex life throughout these years tend to have better physical and emotional health and greater well-being (*see* Heterosexual Expression at Older Ages).

The availability of a partner through marriage or cohabitation is one of the chief predictors of sexual activity during the midlife stage. From age 18 to 35, women are more likely to have a stable partner than men, but after age 35, men are more likely to be partnered than women. Women aged 57–85 are less likely than men of their same age to be in a marital or other intimate relationship and this disparity increases dramatically with age. About 85% of men aged 57 are partnered compared to about 69% of women of the same age. By age 85, about 65% of men and only 29% of women are married or cohabitating. Among men and women of the same age, men with a spousal or other intimate relationship were more likely to be sexually active than women with such a relationship. This is due, in large measure, to the age difference between partners; men in their early 70s tend to be married to women in their mid-to-late 60s. Among the older adults who were not in a relationship, only 22% of men and 4% of women reported being sexually active in the previous year (Lindau et al., 2007).

The share of older adults who were sexually active in the past year declines with age, both among all adults and among those with a partner (*see* Heterosexual Expression at Older Ages). At ages 57–65, almost all partnered men and about 80% of partnered women were sexually active, but by ages 81–85 these numbers had fallen to about half of partnered men and 20% of partnered women (Lindau et al., 2007).

For sexually active adults aged 57–85, the frequency of sex was lower among those who were aged 75–85 among younger persons. However, even in this oldest age group, 54% of sexually active persons reported having sex at least two to three times per month, and 23% reported having sex once a week or more. Fifty-eight percent of sexually active people in the 57–64 age group reported engaging in oral sex, as compared with 31% in the 75–85 age group (Lindau et al., 2007).

The prevalence of masturbation, like that of sexual activity with a partner, was lower at older ages and was higher among men than among women. Poorer health was also associated with a lower likelihood of masturbation among women. Fifty-two percent of older men and 25% of older women with a spousal or other intimate relationship reported masturbating in the previous 12 months, as compared with 55% of older men and 23% of older women without a current spousal or other intimate relationship (Lindau et al., 2007).

Menopause, Infertility, and Miscarriage

Menopause, a major reproductive event and important turning point in the sexual life course of women, occurs on average between the ages of 48 and 52. Menopause is the cessation of menstruation caused by the fluctuation of hormones in the ovaries and bloodstream. Infertility is a direct result of menopause and can have major effects on the emotional well-being of women. The hormone fluctuations also can occur for several years prior to and after the onset of menopause (Dillaway, 2012). Hormone fluctuations and the end of fecundity can bring physical, psychological, and emotional changes and challenges.

Dillaway interviewed 98 women between the ages of 38 and 65 about their feelings pertaining to menopause and sexuality. Some women stated that they were relieved to be free of the risks of pregnancy and childbearing. Several participants stated that they felt relief from the cessation of menstruation. A sexual revitalization was also a common theme in the reports of postmenopausal women. One-third of the women in the study reported having had reproductive difficulties, such as miscarriages or infertility. These women reported having both a negative and positive outlook on menopause and their reproductive difficulties and these differed by race. The majority of the African-Americans stated that while miscarriages and abortions were emotionally devastating, they felt that they became emotionally stronger as a result. The white women who reported problems with pregnancy and childbearing more often stated that after infertility and miscarriages they were glad to reach menopause and move on. Women in this study who either married late or remarried often wanted to have children for their first time or with their new husbands. For them, menopause led to both reproductive problems and the accompanying emotional distress.

Physical Health and Sexuality

Physical health plays a major role in the ability to engage in sexual activity. Physical injuries, diabetes, cardiovascular disease, cancer, and medications for many chronic conditions may alter sexual trajectories at various points in the life course. The decrease in blood flow caused by cardiovascular disease or diabetes may lead to erectile dysfunction, as may treatments for prostate cancer. Some types of antidepressants also reduce sexual desire. Physical ailments such as arthritis and obesity can interfere with mobility and functional health, reducing capacity for sex and leading to sexual dysfunction (*see* Heterosexual Expression at Older Ages; Karraker and DeLamater, 2013).

In a study of older adults aged 57–85, women report their most prevalent sexual problems: 43% reported low sexual desire, 39% reported vaginal dryness, and 34% reported an inability to climax. For the men, erectile dysfunction affected 39 and 14% stated that they used medications or supplements to help with their sexual dysfunctions.

The success and popularity of medications for sexual dysfunction may redefine sexuality among older adults. In addition to alleviating sexual problems, the availability of these medications and the discussions about them in the media may reduce the negative stereotypes that surround the sexuality of older adults. In addition, demand for such medications may lead to future research on and development of treatments for other health issues connected to sexual dysfunction.

Mental health conditions also affect sexual activity and overall well-being. Anxiety, depression, bi-polar disorder, and stress tend to affect sexual interest and sexual activity adversely. Recent research has also shown that mental health conditions are more commonly associated with women's sexual problems than with men's (*see* Heterosexual Expression at Older Ages). It is important to note that the onset of a mental health problem might be associated with other underlying health issues, perhaps undiagnosed.

Living with Illnesses and Disabilities

Midlife may bring the onset of chronic disease and the associated loss of function and disability. The extent to which illness and disability are transitions in the life course depends on characteristics of the individual, their relationships and network of support, and the environment in which they live. In a study of men with spinal cord injuries, Bender (2012) examined the impact of this life course transition on the ability to establish and maintain sexual relationships. He concluded that financial status, social support, marital status, and age at occurrence were crucial determinants of men's ability to adapt to the situation.

Bender found that the participants' responses to disability could be grouped into four types: adaptation, performance, conformity, and relinquishment. The men who were in the adaptation group showed the highest levels of control over their sexual trajectories and satisfaction. These men tended to be young at the time of their injury, had high levels of education, and adequate financial security. They tended to be open about their sexual dysfunctions and limitations in relationships (Bender, 2012). The men in the performance group typically tried to present themselves as normal and unaffected by their disability. They felt able to pursue and end relationships on their own terms. Maintaining control over their sexual trajectories, they felt, allowed them to be viewed as desirable and emotionally stable to women they met (Bender, 2012). Unlike the adapters and performers, the conformers had low sexual agency and were older at the time of injury. These men found it difficult to understand their injury and how it would affect their current and future relationships. They tended to conform to the stereotype that older men with a disability cannot maintain a sexual relationship (Bender, 2012). The men in the relinquishment group had a fear of intimacy and tended to avoid dating and sex. These men lacked social support and financial security. They tended to have anxiety and fear about being able to perform sexually (Bender, 2012).

Sexuality in Later Life

Considerable attention has been paid to the onset of sexual activity in the early stages of the life course, and less to sexuality at older ages, including the cessation of partnered sex at older ages. Recent studies suggest that about a third of adults aged 57–85 have been sexually inactive for at least the last year. Karraker and DeLamater (2013) find that two thirds of these sexually inactive older adults believe their sexual relationships with their spouses are finished. For many of these individuals, this transition marks the end of their partnered sexual lives. Others may find a new partner after being widowed or divorced, or will resume a sexual relationship with a spouse who has recovered from a disability or illness.

Continued sexual involvement at older ages depends on the physical capacity to engage in sex, but recent research has shown that it is the health of the *husband* that matters; couples in which the wife is in poor health but the husband's health is good are much more likely to remain sexually active than couples in which the husband is in poor health but the wife's health is good.

Older adults who are happier with their marriage, and those who are remarried, tend to have sex more often than those in a first marriage or those who are less satisfied. This is also true at younger ages. Karraker and DeLamater (2013) find that although happier marriages tend to be sexually active marriages in midlife, at older ages this is less often the case; those in sexually inactive marriages are about as happy with their relationship as those who remain sexually active. Older adults in long-standing marriages may fall back on a lifetime of shared activities, shared children, and memories of an active sex life in the face of declining physical or cognitive function of one or both partners.

See also: Adolescent Sexual Risk; Heterosexual Expression at Older Ages; Hooking up and Casual Sex; Marital Quality and Health; Masculinity, Scripts of; Sexual Debut; Sexuality Education; Sexually Transmitted Infections: Social Network Analysis.

Bibliography

Bender, Alexis A., 2012. Secrets and magic pills: constructing masculinity and sexual "normalcy" following spinal cord injury. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.

Browning, Christopher R., Laumann, Edward O., 1997. Sexual contact between children and adults: a life course perspective. American Sociological Review, 540–560.

Carpenter, Laura M., DeLamater, John D. (Eds.), 2012. Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.

Das, Aniruddha, Waite, Linda J., Laumann, Edward O., 2012. Sexual expression over the life course: results from three landmark surveys. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.

Dillaway, Heather, 2012. Reproductive history as social context: exploring how women talk about menopause and sexuality at midlife. In: Carpenter, Laura M.,

DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.
Friedrich, William N., Grambsch, Patricia, Broughton, Daniel, Kuiper, James, Beilke, Robert L., 1991. Normative sexual behavior in children. Pediatrics 88 (3), 456–464.
Friedrich, William N., Fisher, Jennifer L., Dittner, Carrie Anne, Acton, Robert, Berliner, Lucy, Butler, Judy, Damon, Linda, Davies, W Hobart, Gray, Alison, Wright, John, 2001. Child sexual behavior inventory: normative, psychiatric, and sexual abuse comparisons. Child Maltreatment 6 (1), 37–49.
Goldman, Ronald, Goldman, Juliette, 1982. Children's Sexual Thinking: A Comparative Study of Children Aged 5 to 15 Years in Australia, North America, Britain and Sweden. ERIC.
Halpern, Carolyn Tucker, Joyner, Kara, Udry, J Richard, Suchindran, Chirayath, 2000. Smart teens don't have sex (or kiss much either). Journal of Adolescent Health 26 (3), 213–225.
Jeynes, William H., 2006. The impact of parental remarriage on children: a meta-analysis. Marriage & Family Review 40 (4), 75–102.
Jeynes, William, 2012. A meta-analysis on the relationship between parental divorce family structures and adolescents' attitudes and behaviors regarding premarital intercourse. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.
Joyner, Kara, Laumann, Edward O., 2001. Teenage sex and the sexual revolution. In: Sex, Love, and Health in America: Private Choices and Public Consequences, pp. 41–47.
Joyner, Kara, Udry, J. Richard, 2000. You don't bring me anything but down: adolescent romance and depression. Journal of Health and Social Behavior, 369–391.
Karraker, Amelia, DeLamater, John, 2013. Past-year sexual inactivity among older married persons and their partners. Journal of Marriage and Family 75 (1), 142–163.
Laumann, Edward O., Waite, Linda J., 2008. Sexual dysfunction among older adults: prevalence and risk factors from a nationally representative US probability sample of men and women 57–85 years of age. The Journal of Sexual Medicine 5 (10), 2300–2311.
Lindau, Stacy Tessler, Schumm, L Philip, Laumann, Edward O., Levinson, Wendy, O'Muircheartaigh, Colm A., Waite, Linda J., 2007. A study of sexuality and health among older adults in the United States. New England Journal of Medicine 357 (8), 762–774.
Martin, Joyce A., Hamilton, Brady E., Ventura, Stephanie J., Osterman, M.J., Kirmeyer, Sharon, Mathews, T.J., Wilson, Elizabeth C., 2011. Births: final data for 2009. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System 60 (1), 1.
McLanahan, Sara, Sandefur, Gary, 1994. Growing up with a Single Parent: What Helps, What Hurts. Harvard University Press, Cambridge, MA.
Rosin, H., 2012. The End of Men: And the Rise of Women. Riverhead Publishing.
Russell, Stephen T., Van Campen, Kali S., Muraco, Joel A., 2012. Sexuality development in adolescence. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.
Taylor, Kate, 2013. Sex on campus: she can play that game, too. In: New York Times. The New York Times Company, New York.
Thigpen, Jeffry W., 2009. Early sexual behavior in a sample of low-income, African American children. Journal of Sex Research 46 (1), 67–79.
Thigpen, Jeffry W., 2012. Childhood sexuality: exploring culture as context. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.
Tolman, Deborah L., 2002. Dilemmas of Desire: Teenage Girls Talk about Sexuality. Harvard University Press, Cambridge.
Wade, Lisa, Heldman, Caroline, 2012. Hooking up and opting out: negotiating sex in the first year of college. In: Carpenter, Laura M., DeLamater, John D. (Eds.), Sex for Life: From Virginity to Viagra, How Sexuality Changes Throughout Our Lives. New York University Press, New York and London.
Waite, Linda J., Maggie Gallagher, 2000. The case for marriage: why married people are happier, healthier, and better off financially. Doubleday.

Sexuality, Theories of

Elke D Reissing and Heather VanZuylen, School of Psychology, University of Ottawa, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to provide a brief review of the historical origin of theories of human sexuality followed by a review of modern theories including sociobiological and evolutionary psychological theories, behaviorism, social learning theories, and cognitive theories. Next, social theories including social exchange theory, social constructionism, and script theory are examined. Last, theories which integrate multiple perspectives are discussed, such as biopsychosocial approaches and the dual control model of sexuality. The authors conclude that drawing on multiple, empirically validated theories provides the basis for researching and understanding complex and nuanced yet parsimonious explanations of sexuality.

The Importance of Theories

According to Larsen and Buss (2008), theories provide three useful tools for researchers. First, they provide a guide that helps determine the types of information that researchers should collect, the questions that are useful, and a direction for research. Secondly, theories organize previous findings and provide explanations for patterns in descriptive information. Finally, theories allow researchers to make predictions, which allow the testing of hypotheses. Larsen and Buss also provide standards to evaluate scientific theories, including comprehensiveness, testability, parsimony, and the steering of researchers toward making new discoveries. Scientific theories, therefore, need to accommodate and explain all relevant observations and be able to be falsified. Furthermore, they should be efficient at explaining observations with as few tenets as possible, lead researchers toward new lines of inquiry, and facilitate research in the field by integrating information from different sources into coherent concepts. As the different theories of sexuality are discussed, it is important to be mindful of the elements of genuine theories and the criteria by which they are evaluated. With the importance of theories in mind, early sexuality research will be discussed to provide a historical context to the diversity of contemporary theories that have developed over the past several decades.

Early Sexuality Research

Sexuality theories are relatively new, as sexuality has long been the domain of religion, morality, and philosophy rather than the sciences. The earliest scientific approaches to sexology were medical in nature, and thus early sexuality research was largely conducted by physicians and psychiatrists, particularly in Germany (Carroll, 2013). The medical model allowed for researchers to treat sexuality as a subdiscipline of biology and avoided some of the perceived stigma associated with sex. However, the medical model predominantly focused on sexual pathologies. During the 1800s, much of this research was summed up in *Psychopathia Sexualis* by Heinrich Kaan (1844), which was subsequently expanded upon by Richard von Krafft-Ebing (1886).

Not all early researchers were satisfied with the medical approach. Iwan Bloch founded the *Journal of Sexology* in 1914, in an attempt to draw on historical and anthropological accounts toward a broader understanding of sexual phenomena (Bullough, 1994). Some contemporaries of Bloch, such as Magnus Hirschfeld, Katharine Bement Davis, and Margaret Higgins Sanger used sexology research to foster social change, for example, to increase acceptance of homosexuality, female sexuality, masturbation, and birth control (Carroll, 2013).

Albert Moll published *Untersuchungen über die Libido sexualis*, the first psychological theory of sexuality, in 1897. He argued that the libido had two fundamental drives, *Detumeszenztrieb* (individualistic gratification) and *Contrectationstrieb* (intimacy with others). Although an open critic of Sigmund Freud, he shared Freud's belief that sexuality developed before puberty and that children had libidos (Bullough, 1994).

Although Freud had many contemporaries, his fame brought the psychology of sexuality to the forefront. Freud (1924; see also Petersen and Hyde, 2010) proposed a 'stage model,' whereby the libido was focused on different erogenous zones throughout childhood and these stages, (the oral, anal, phallic, latent, and genital stages) had to be successfully resolved; otherwise individuals could develop fixations and maladaptive personality traits in adulthood. Although criticized for being untestable, sexist, heterosexist, and based on information from only nonnormative case studies (Petersen and Hyde, 2010; Nelson, 2010), Freud's theories have given rise to more contemporary theories, such as feminist psychoanalysis (e.g., Chodorow, 1989) and the object-relations theory (Person, 1980).

Although not all aspects of these historical theories are currently accepted, it is important to acknowledge the work of these pioneers as they laid the foundations of modern sexology and sexual theory. In the remainder of the article, modern theories including biological, learning, cognitive, sociological, and integrated theories will be discussed with regard to their pioneers, basic tenets, examples of seminal research, and predominant critiques.

Sociobiology and Evolutionary Psychology

This theoretical perspective examines the biological causes of behavior, including the use of animal models, genetic analyses, and the study of hormones and neurological substrates (Wilson, 1978). Evolutionary psychology applies the forces of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.35030-9

evolution and natural selection to social behaviors and proposes that certain behaviors are maintained because they are adaptive to the reproductive success of an organism or genetically similar organisms (Confer et al., 2010). In humans, particular ancestral behaviors allowed humans to survive and raise more viable children; therefore the biological bases for that behavior were inherited, thus increasing the frequency of those genes and behaviors (Roberts, 2012). Evolutionary psychology has been applied to a number of human sexual behaviors, including the preference for monogamy and coparenting (a rarity in mammals), which are linked to an increase in receptor sites and activity of the neurotransmitters, oxytocin, and vasopressin (Fisher, 2012). It has also been applied to sex offenders and the adaptive potential of rape (Camilleri, 2012; see also Buntin, this work, for more information on sexual violence), and gender differentiated popularity of romance novels and pornography (Salmon, 2012).

At the core of many evolutionary insights into sexuality is the Sexual Strategies Theory (SST; Buss, 1998), which asserts that there are two main mating strategies, short-term and long-term. Ancestrally, reproduction involved a higher cost for females who are burdened with pregnancy and nursing, compared to the minimal cost involved for men in the donation of sperm. Therefore the primary mating strategy of women is long-term (i.e., forming a long-term relationship), as it is in a female's best interest to carefully choose a partner with resources, who is interested in contributing to raising their child. For a man, a risk of this long-term strategy is that he may be investing resources in a child who is not genetically related. Therefore, men have a greater interest in short-term mating strategies (i.e., casual sex) than women, as this increases the number of potential children (see England & Ronen, this work, for more information on casual sex). However, having greater interest in casual sex does not negate the ability for men to form long-term relationships and raise children, as this strategy maximizes the chances of survival for the child.

Hypotheses generated by this theory have received some support. For example, based on SST, men and women respond differently to various types of infidelity in heterosexual relationships. Women show greater concern if their partner is emotionally attracted to another woman (potentially investing in her); while men show significantly more concern if the partner is physically attracted to another man (potentially unfaithful); (Sabini, 2004; *see* Heterosexual Sexuality: Partner Bond Formation, Sexual Attitudes, Sexual Practices, and Sexual Function over the Life Course).

Evolutionary psychology and sociobiology have been criticized as emphasizing the need to understand ancestral conditions in order to predict presumably more complex, current behaviors. Yet many of the predictions and findings of evolutionary psychology are in fact in line with more proximal theories, such as social and learning theories (Gannon, 2002). Some have argued that the distal causes of behavior are not testable and are less relevant than observable social- and cognitive-based hypotheses (Deaux, 1999). Additionally, when results contradict the evolutionary paradigm, they are sometimes dismissed with the explanation that current conditions differ from ancestral ones (Gannon, 2002). Other theorists have criticized the idea that much of humans' past involved constancy in sexuality and mating choices (Foucault, 1978). Despite these criticisms, evolutionary psychology continues to be a burgeoning field with rich explanatory power, especially when studied as basic predispositions in conjunction with theories that take the social environment into account.

Behaviorism

Behaviorism is the study of how behaviors come about in response to stimuli in the environment, particularly the study of processes associating outcomes with different stimuli through operant or classical conditioning. Behaviorists consider the mind of the child as a blank slate with a few inborn abilities aimed at basic survival and the ability to learn through association (Allgeier and Allgeier, 1998). Thus, people are highly malleable by the environment. *Operant conditioning*, pioneered by Skinner (1974), proposes that behaviors become more frequent when reinforced and diminish when punished. Rewards that satisfy biological or intrinsic needs, such as sex, are primary reinforcers. Thus, behaviors that lead to sex or sexual pleasure will become more frequent and behaviors that interfere with obtaining sex will diminish. Notably, sex is both a behavior and a reinforcer, indicating that sexuality has both innate and learned qualities. For example, when young children might play with their genitals in front of a parent and this behavior may be punished. Thus the young person learns what is sexually pleasurable but also learns to restrict this activity to socially acceptable situations (Allgeier and Allgeier, 1998). Operant conditioning has been used to explain some cases of compulsive sexual behavior; for example, someone who persistently uses sexual behavior to cope with stress is engaging in an intrinsically rewarding activity, while also relieving a negative emotional state, which is also rewarding. This can lead to a positive feedback cycle that increases sexual activity until it may become uncontrollable (Carnes, 1983).

Classical conditioning was developed by Pavlov, who observed that after repeatedly pairing a stimulus that did not elicit a response (neutral stimulus) with a stimulus that did cause a response (unconditioned stimulus; unconditioned response), a link between the two stimuli was formed, such that the neutral stimulus (now a 'conditioned' stimulus) presented on its own, could elicit the same response as the unconditioned stimulus (conditioned response). Thus the association from one stimulus can be transferred to another through repeated pairing; for example, the smell of a lover's perfume or cologne may become arousing through repeated pairings of sexual interest and the odor (Allgeier and Allgeier, 1998). In a similar manner, an aversive stimuli experienced during sexual arousal can lead to a decrease in sexual responses, even when the aversive stimulus is removed (Both et al., 2008). Classical conditioning has been used to explain some of the complexities of sexuality, such as having fetishes or sexual phobias. For example, in therapy, systematic desensitization was used to help patients to disentangle negative associations with sex. The opposite process was used to try to reduce patients' atypical or problematic sexual attractions (Hogben and Byrne, 1998). Despite being empirically supported, learning theories were somewhat limited and did not necessarily capture the full range of human sexuality. Social learning

theory emerged to account for observational learning, values, cognitions, and social interactions.

Social Learning Theory

Social Learning Theory was popularized by Bandura (1986), and encompasses how observations in the social world impact behaviors and cognitions. Although classical and operant conditioning are still relevant, Bandura also proposed modeling, where an observer views a person's actions and their consequences. If that person is rewarded, the observer is more likely to mimic those actions. However, if the action is punished, the observer learns to expect negative consequences from that behavior. Modeling may take place immediately (direct imitation) or the consequences of the behavior are remembered later (observational learning) (Nelson, 2010). For example, premarital sex may be rewarded in some communities and punished in others; and by observing these consequences, members of these communities learn the expectations of acceptable sexual behavior. In fact, the belief that peers are also sexually active is a strong predictor of previous engagement in sexual activity (Christopher et al., 1993; *see* Sexual Debut). Thus, socialization, culture, and media contribute to the development of expectations of the consequences of behaviors and one does not need to experience reinforcement or punishment personally in order to be affected. Bandura (1977) proposed that after successfully performing a task, one experiences a sense of self-efficacy. In the sexual context, the affective experience and timing of first intercourse may be of particular importance, as it has a strong impact on sexual self-efficacy, which is positively correlated with healthy sexual adjustment in later life (Reissing et al., 2011, 2005).

Although heavily supported through empirical research, the primary criticisms of the social learning theory include that it assumes passivity in the nature of the observer who recieves rewards or punishments, that the theory does not explain why certain behaviors are rewarded or punished, and that it fails to explain why some individuals will not conform to social norms (Nelson, 2010). The following theories will address some of these concerns.

Cognitive Theories

Cognitive theories share the basis that one's thoughts and interpretations of the environment are prone to errors, distortions, and incorrect assumptions. Cognitive theories are readily translated into therapeutic interventions, especially through cognitive-behavioral therapy, which seeks to address cognitive distortions and maladaptive assumptions to develop healthier interpretations of, and reactions to events. For example, sex offenders, such as child molesters, tend to have distorted thinking, including what Ward and Keenan (1999) refer to as *implicit theories* about the world, which include viewing children as sexual beings entitled to sexual fulfillment; seeing other adults as dangerous; believing that sexuality is fundamentally uncontrollable; and assuming that there is little harm and potentially even benefits associated with sexually abusing a child. The majority of sex offender treatment programs in the United States and the United Kingdom take a cognitive-behavioral approach in teaching offenders to change their thoughts and behavior patterns in order not to reoffend. These programs are considered empirically supported in reducing recidivism (Beech et al., 2013). Thus, how people think can have a strong impact on their behaviors and how they express their sexuality.

Althof et al. (2005) reviewed how cognitive factors, such as a lack of sexual confidence, anxiety, depression, inadequate sexual information, guilt, and low self-esteem can also trigger or maintain more commonplace sexual problems, such as low sex drive and erectile dysfunction. For example, Barlow (1986) theorized that anxiety in men with erectile dysfunction caused cognitive interference, distractibility, and an inadequate processing of sexual stimuli. Mental distractions from the sexual stimuli are then detrimental to the physical sexual responses of the man. Common cognitive distractions include performance anxiety and feelings of inadequacy. Cognitive-behavioral therapy has also been used to address distorted thinking, anxiety, and maladaptive thoughts in women with hypoactive sexual desire disorder with nearly half the women benefiting from this type of treatment (Althof et al., 2005). Overall, how people think has been implicated in a number of sexual problems and targeting cognitions has become an integral part of sex therapy.

Social Theories

As discussed thus far, theories of sexual behavior can explain sexuality as developing according to biological sex, hormones, genetics, and what has been adaptive in humanity's ancestral past (sociobiology and evolutionary psychology), what is learned through rewards, punishment, and associations between stimuli (operant and classical conditioning), what is observed (social learning theory), and how thought processes can affect sexual functioning (cognitive theories). Thus, these theories are based on the individual and his or her development, biology, learning experience, and cognitions. Although these theories have acknowledged how interacting with the environment is important, the following theories, including the social exchange theory, cultural research, social script theory, and symbolic interactionism emphasize the interpersonal and cultural aspect of sexuality as central to its formation.

The Social Exchange Theory

The social exchange theory has been applied to many different situations; in the context of sexuality, the theory aims at explaining why two particular people may engage in sexual behavior, while another two may not. The theory therefore places the focus predominantly on the relational context. It is placed here under social theories to highlight the strong influence of interpersonal power dynamics, gendered power in society, and because it takes an interpersonal orientation in analysis; however, it is also compatible with evolutionary mate-selection strategies. The social exchange theory posits that human sexuality can be analyzed with the concepts of

economics – that people attempt to maximize their rewards while minimizing costs, that interaction with others is a series of exchanges, and that people who receive rewards feel obligated to reciprocate (Sprecher, 1998).

Within heterosexual culture, the expectation is that men desire sex more than women and thus women have the ability to grant access to something desired. This creates what is referred to as the Principle of Least Interest (Waller and Hill, 1951), which stipulates that the party that has the greater interest in an outcome has less power in negotiations. Thus, women have more power in sexual negotiations and may request additional rewards in exchange, including love, status, money, and influence (Baumeister and Mendoza, 2011). Restricting the availability of sex and increasing its value is therefore of particular importance in cultures where women are disadvantaged in access to income, education, and political representation. Indeed, Baumeister and Mendoza (2011) found that of 37 countries, those which had greater gender equality reported increased casual sexual activity, less abstinence promotion, more premarital sex, and younger age of first intercourse. Thus, as predicted, when women have access to resources through nonsexual means, the 'price' of sexual activity is lower and sex is less restricted.

On an interpersonal level, Byers and Wang (2004) reported that people were more likely to stay in a relationship where they found that the rewards were high and the costs were low. The researchers further noted that partner satisfaction and relationship stability benefited from both partners experiencing a balance of rewards and costs. If one partner felt the other was benefiting more, then for this partner, the temptation to seek rewards outside the relationship increased. This temptation was decreased however, if few or no more appealing alternatives existed (*see* Satisfaction and Commitment in Long-Term Heterosexual Relationships).

The social exchange theory has been criticized on several fronts. One major obstacle in the empirical evaluation of the concept is the subjective nature of costs and rewards, which may differ in value between different people, over time, or through comparisons with other people or rewards. The social exchange model also typically tests how exchanges affect relationships and sexual variables; meanwhile, very little research has been conducted on how the quality of a relationship affects exchange variables. The social exchange theory nevertheless demonstrates that sexual decision-making is grounded in a broader social context that regulates the risks and benefits of sex. It also demonstrates how broad social movements, such as gender equality and feminism, can alter the sexual landscape and affect decision-making.

Social Constructionism

Social constructionism posits that reality is socially constructed, that is, defined and shaped through social interactions, cultural expectations, and language. The degree to which theorists accept objective or biological truths about reality varies widely. Some social constructionists acknowledge that sexuality has an underlying biological drive (i.e., Berger and Luckmann, 1966), while others argue that there is no objective biological essence to sexuality and that it is completely socially defined (i.e., Foucault, 1978). Given that the vast majority of sociological research is social constructionist to some extent, cultural research will first be discussed as an introduction to sociological research, and then, more specifically, script theory and symbolic interactionism will be discussed as particular theories in the field of sociology.

Sociocultural Research

In general, sociological theories examine how social institutions, socialization processes, culture, and language affect sexuality. For example, cross-cultural research demonstrates that women are capable of experiencing sexual pleasure; however, not all cultures recognize the role of pleasure in sexual activity for women. In Uganda, the So people teach women that sex is something to be endured in order to conceive children and no effort is made to encourage sexual pleasure. In contrast, the Mangaian people located in Polynesia are expected from childhood to learn how to give and receive sexual pleasure, and women in this culture are expected to have more orgasms than men and sexual relations before marriage (Allgeier and Allgeier, 1998).

In a cross-cultural study of 190 societies, Ford and Beach (1951) found that attitudes toward homosexuality varied greatly, and thus prevalence also differed. Of 78 less developed societies, the majority (49) approved or tolerated same-sex sexual activity in one form or another. There were other differences, such as whether sexual activity was public (but never in front of children), whether inflicting pain on your partner invokes sexual excitement, or whether masturbation was a regular part of sexuality. Standards of beauty also vary widely from one culture to another, although typically what makes a woman beautiful is more explicitly denoted (*see* Heterosexuality in Africa; Kontula, this work, for discussion of sexuality in Finland; Ng, this work, for discussion of sexuality in China; Richters, this work, for discussion of sexuality in Australia). Alternatively, sexuality has the potential to impact society; as Reiss (1986) argues, even sexually liberal societies place importance on the potential for sexual activity to lead to self-disclosure and pair-bonding. The stability of these interpersonal bonds is important for the formation of families, which are central to societies.

Sociological views of sexuality have been criticized by some theorists who question that social structure or socialization could have any causal role in behavior when compared with the backdrop of sociobiology and evolutionary psychology. Instead, biological drives and tendencies affect culture (Eagly and Wood, 2003). Moreover, sociological theory sometimes focuses on outliers, with cultures that are the exception to the rule taking central place, while the commonalities of the vast majority of cultures may be ignored.

Script Theory

Script theory was pioneered by Simon and Gagnon (1986, 2003) and explored how society defines expectations of sexual behavior. As an attempt to order social behavior, it is important to be able to anticipate the actions of another person. By having predetermined narratives of how sexual

interactions are supposed to occur, people can better predict another person's behavior and act accordingly, making social situations more comfortable.

Sakaluk et al. (2013) conducted a qualitative study on current sexual scripts. They found that dominant heterosexual scripts are typically determined by gender and include assumptions that men are always ready for sex and are the initiators, while women are required to be more reserved, experience less sexual desire, and be the gatekeepers for sex (*see* Femininities). This creates an additional requirement that men are expected to be sexually skilled, knowledgeable, and persuasive, and a lack of skill may be damaging to masculinity (*see* Scripts of Masculinity). However, a new and emerging sexual script for women is emerging that accepts female sexual skills, especially in oral sex. Yet, women are not expected to reveal how they learned these skills to maintain an illusion of inexperience and virginity (Sakaluk et al., 2013). Sexual scripts are also changing the way young adults view relationships and increasingly, noncommitted sexual relationships are becoming the norm rather than exception (*see* Adolescent Sexual Risk). Wentland and Reissing (2011) reported on young adults reliably distinguishing four different casual sexual relationships. These relationships differ in terms of scripts of initiation, maintenance, and termination (Wentland and Reissing, 2012).

Sexual script research has been criticized as lacking predictive power for sexual behavior. Although expectations of behavior are acknowledged, what actually occurs may differ (Bancroft, 2009). Research has yet to determine how sexual scripts affect behaviors such as sexual risk-taking, sexual satisfaction, condom and contraceptive use, and experiencing sexual problems (Sakaluk et al., 2013). On a societal level, sexual scripts are still an important part of determining attitudes toward sexuality, 'appropriate' feelings, and expectations of sexual behavior. They are part of how one learns through socialization, the times and ways in which sexuality is proper and rewarded, as well as when it will be judged as inappropriate (Allgeier and Allgeier, 1998).

Integrated Theories

The previous theories differ widely, ranging in level of analysis from genetic to cultural, but are all derived from data collection and observation. Therefore, which theory is correct? Is sexuality predominantly inborn or shaped through socialization, learning, and culture? Many current theorists acknowledge the empirical basis for multiple theories and choose to integrate different levels of analysis and perspectives in order to create nuanced and comprehensive theories. The following approaches and theories are examples of how these different theoretical lenses can be synthesized into coherent and complex understandings of human sexuality.

The Biopsychosocial Approach

Rather than a predictive theory, the biopsychosocial approach is a paradigm in which biological, social, and psychological aspects of a person are taken into account in order to adequately describe behavior. From this perspective, any clinical intervention for a person with sexual problems should address the biological underpinnings of behavior, the social environment, and the personal psychology of the person (Rossi, 1994; *see* Epidemiology of Sexual Dysfunction). This approach is valuable in that it argues that all levels of analysis are necessary in order to address an issue, from the genetic, hormonal, and evolutionary, to an individual's development, experiences, and cognitions, as well as the interpersonal, social, and cultural aspects of a person's behavior.

A clinician working with a female client with low sexual desire, for example, must investigate whether there are biological reasons such as side effects of prescription medication. Developmental factors may include childhood sexual abuse or a sexually conservative upbringing affecting the experience of the desire for sexual activity. Interpersonal factors may include a lack of communication or comfort in her relationship. If an obstacle exists at any level, then it must be addressed in order to maximize the effectiveness of treatment (Denman, 2004). The biopsychosocial approach is not just clinical in nature; for example, it has been used to analyze sexual behavior such as sexual risk-taking (Bancroft, 2000), how disability affects sexuality (Lindau et al., 2003), HIV/AIDS transmission (Cohen, 1990), and the etiology, maintenance, and treatment of sexual pain disorders (Van Lankveld et al., 2010).

The biopsychosocial approach is difficult to criticize as it considers multiple layers of analyses. It relies on the researcher or clinician to develop ways of integrating theories. In one situation a biological emphasis may be of more use, while in another the psychological perspective may be more appropriate. The approach also does not attempt to distinguish which of multiple psychological theories a researcher should choose from, such as psychoanalytic, cognitive, or behavioral. As a result, there is a risk of atheoretical reasoning and focus on the individual in a predominantly disorder-driven analysis.

The Dual Control Model

The Dual Control Model of Sexuality (DCM; early ref) is an example of a stand-alone theory of sexual motivation that integrates biological mechanisms with psychosocial factors in a very specific, empirically grounded manner (*see* Sexual Motivation). The DCM (Bancroft and Janssen, 2000; see also Bancroft et al., 2009) originates from Gray's (1987) research into Behavioral Activation Systems (BAS) and Behavioral Inhibition Systems (BIS). These two systems help to regulate behavior and motivation through exciting (BAS) and inhibiting (BIS) various responses. The Dual Control Model holds that there is individual variability in thresholds for excitation and inhibition, such that some people are predisposed toward excitation, while others are more predisposed toward inhibition. Bancroft et al. (2009) posit that this variability is usually not problematic unless sexual excitation is particularly high, combined with low inhibition or vice versa. In this case, individuals are at risk for

developing sexual problems such as sexual impulsivity or hypoactive sexual desire.

This model is an integrated theory because it acknowledges both the inherited neurobiology of excitation and inhibition systems in the brain, as well as the effect of cognitions on behavior and the influence of culture. Bancroft and Janssen (2000) developed a questionnaire for excitation and inhibition for men outlining key factors affecting male sexuality: general excitation, inhibitory tone (or general levels of inhibition), and inhibition due to external threats (such as fear of performance failure). Therefore, they argued that men with centrally mediated erectile dysfunction would respond better to psychological therapy if they were prone to inhibition resulting from perceiving threats, whereas those with high inhibitory tone would respond better with pharmaceutical interventions.

As a relatively new theory, more research must be carried out to confirm the initial, promising findings. Although the model was initially constructed using data mostly from men (Rowland, 2006), the development of a DCM questionnaire for women will foster increased research (Graham et al., 2006).

Other Theories to Consider

Despite exploring a range of theories outlining biological, learning, cognitive, social, and integrated theories of sexuality, this list is not exhaustive. For example, political movements resulting in theories such as *Feminist Theory* (e.g., Beasley, 2005) have paved the way to explore how power dynamics and gender roles shape sexuality. Additionally, *Queer Theory* (e.g., Beasley, 2005; Green, 2007) has helped researchers explore how sexual identities are constructed and regulated to produce normative expressions (e.g., heterosexual and monogamous) while discplining and restricting non-normative expressions. Other social constructionist theories include *Symbolic Interactionism* (e.g., Longmore, 1998), which posits that the social world is constructed using symbols; when people interact with the physical world they are constantly interpreting and applying symbolic meanings, which then affect how people relate to themselves and others.

Moreover, this article has not considered the wide variety of theories relating to relationships, the components of love (Sternberg, 1986), and styles of attachment (Feeney and Noller, 1990). The multitude of theories in respect to love, attachment, and relationships are worthy of an entire article of their own.

Conclusion

Despite all the theories, paradigms, and approaches that are available to researchers, modern sexuality research has been criticized for not having a strong theoretical basis (Ainsworth and Baumeister, 2012; Bancroft, 2009; Weis, 2009). As theory guides the formulation and testing of hypotheses, this is problematic. Without theory, many studies are rich in descriptive information, but lack explanatory power. Moreover, information from various sources, however elegantly derived, will need integration to generalize to other individuals, situations, and circumstances to be actually helpful in furthering our understanding of sexuality.

It behooves researchers and theorists to face the challenge of acknowledging empirical support for many different theories and to find a parsimonious explanation to advance sexological science.

See also: Adolescent Sexual Risk; Femininities; Heterosexuality in Africa; Heterosexuality in China; Heterosexuality; Hooking up and Casual Sex; Intimate Partner Violence; Masculinity, Scripts of; Satisfaction and Commitment in Long-Term Heterosexual Relationships; Sex Life Challenges: The Finnish Case; Sexual Debut; Sexual Dysfunction, Epidemiology of; Sexuality in Australia.

Bibliography

Ainsworth, S.E., Baumeister, R.F., 2012. Theories of Sexology: What Have We Learned from College Students? In: McAnulty, R.D. (Ed.), Sex in College. ABC-CLIO, LLC, Santa Barbara, CA, pp. 45–62.

Allgeier, E.R., Allgeier, A.R., 1998. Sexual Interactions: Basic Understandings. Houghton Mifflin Company, Boston, MA.

Althof, S.E., Leiblum, S.R., Chevret-Measson, M., Hartmann, U., Levine, S.B., McCabe, M., Rodrigues, O., 2005. Psychological and interpersonal dimensions of sexual function and dysfunction. Journal of Sexual Medicine 2, 793–800. http://dx.doi.org/10.1111/j.1743-6109.2005.00145.x.

Bancroft, J., 2000. Individual differences in sexual risk taking: a biopsychosocial theoretical approach. In: Bancroft, J. (Ed.), The Role of Theory in Sex Research. Kinsey Institute for Research in Sex, Gender, and Reproduction, Bloomington, IN, pp. 177–212.

Bancroft, J., 2009. Human Sexuality and Its Problems, third ed. Elsevier Limited, Toronto, ON.

Bancroft, J., Janssen, E., 2000. The dual control model of male sexual response: a theoretical approach to centrally mediated erectile dysfunction. Neuroscience and Biobehavioral Reviews 24, 571–579.

Bancroft, J., Graham, C.A., Janssen, E., Sanders, S.A., 2009. The dual control model: current status and future directions. Journal of Sex Research 46, 121–142. http://dx.doi.org/10.1080/00224490902747222.

Bandura, A., 1977. Social Learning Theory. Prentice Hall, Englewoods Cliffs, NJ.

Bandura, A., 1986. The Social Foundations of Thoughts and Actions. Prentice-Hall, Englewood Cliffs, NJ.

Barlow, D.H., 1986. Causes of sexual dysfunction: the role of anxiety and cognitive interference. Journal of Consulting and Clinical Psychology 54, 140–148.

Baumeister, R.F., Mendoza, J.P., 2011. Cultural variations in the sexual marketplace: gender equality correlates with more sexual activity. Journal of Social Psychology 151, 350–360. http://dx.doi.org/10.1080/00224545.2010.481686.

Beasley, C., 2005. Gender & Sexuality: Critical Theories, Critical Thinkers. Sage Publications, London.

Beech, A.R., Bartels, R.M., Dixon, L., 2013. Assessment and treatment of distorted schemas in sexual offenders. Trauma, Violence, & Abuse 14, 54–66. http://dx.doi.org/10.1177/1524838012463970.

Berger, P., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, Garden City, NY.

Both, S., Laan, E., Spiering, M., Nilsson, T., Oomens, S., Everaerd, W., 2008. Appetitive and aversive classical conditioning of female sexual response. Journal of Sexual Medicine 5, 1386–1401. http://dx.doi.org/10.1111/j.1743-6109.2008.00815.x.

Bullough, V., 1994. Science in the Bedroom: A History of Sex Research. HarperCollins Publishers, New York.

Buss, D.M., 1998. Sexual strategies theory: historical origins and current status. Journal of Sex Research 35, 19–31.

Buss, D.M., Larsen, R.J., Westen, D., Semmelroth, J., 1992. Sex differences in jealousy: evolution, physiology, and psychology. Psychological Science 3, 251–255. http://dx.doi.org/10.1111/j.1467–9280.1992.tb00038.x.

Byers, E.S., Wang, A., 2004. Sexuality in close relationships from the exchange perspective. In: Harvey, J.H., Wenzel, A., Sprecher, S. (Eds.), Handbook of Sexuality in Close Relationships. Lawrence Erlbaum, Mahwah, NJ, pp. 203–234.

Camilleri, J.A., 2012. Evolutionary psychological perspectives on sexual offending: from etiology to intervention. In: Shackelford, T.K., Weekes-Shackelford, V.A. (Eds.), The Oxford Handbook of Evolutionary Perspectives on Violence, Homicide, and War. Oxford University Press, New York, pp. 173–196.

Carnes, P., 1983. Out of the Shadows: Understanding Sexual Addiction. Hazelden, Centre City, MN.

Carroll, J.L., 2013. Sexuality Now: Embracing Diversity, fourth ed. Wadsworth, Belmont, CA.

Chodorow, N.J., 1989. Feminism and Psychoanalytic Theory. Yale University Press, New Haven, CT.

Christopher, F.S., Johnson, D.C., Roosa, M.W., 1993. Family, individual, and social correlates of early Hispanic adolescent sexual expression. Journal of Sex Research 30, 68–72.

Cohen, M.A., 1990. Biopsychosocial approach to the human immunodeficiency virus epidemic: a clinician's primer. General Hospital Psychiatry 12, 98–123. http://dx.doi.org/10.1016/0163-8343(90)90020-D.

Confer, J.C., Easton, J.A., Fleischman, D.S., Goetz, C.D., Lewis, D.M., Perilloux, C., Buss, D.M., 2010. Evolutionary psychology: controversies, questions, prospects, and limitations. American Psychologist 65, 110–126. http://dx.doi.org/10.1037/a0018413.

Deaux, K., 1999. An overview of research on gender: four themes from three decades. In: Swann, J.W., Langlis, J.H., Gilbert, L.A. (Eds.), Sexism and Stereotypes in Modern Society. American Psychological Association, Washington, DC, pp. 11–31.

Denman, C., 2004. Sexuality: A Biopsychosocial Approach. Palgrave MacMillion, New York.

Eagly, A.H., Wood, W., 2003. The origins of sex differences in human behavior: evolved dispositions versus social roles. In: Travis, C.B. (Ed.), Evolution, Gender and Rape. Massachusetts Institute of Technology, Cambridge, MA, pp. 265–304.

Feeney, J.A., Noller, P., 1990. Attachment style as a predictor of adult romantic relationships. Journal of Personality and Social Psychology 58, 281–291. http://dx.doi.org/10.1037/0022-3514.58.2.281.

Fisher, H.E., 2012. Serial monogamy and clandestine adultry: evolution and consequences of the dual human reproductive strategy. In: Roberts, S.C. (Ed.), Applied Evolutionary Psychology. Oxford University Press, New York, pp. 93–114.

Ford, C.S., Beach, F.A., 1951. The History of Sexuality. Harper & Brothers, Toronto, ON.

Foucault, M., 1978. The History of Sexuality. An Introduction, vol. 1. Pantheon, New York.

Freud, S., 1924. An Introduction to Psychoanalysis. Permabooks, New York.

Gannon, L., 2002. A critique of evolutionary psychology. Psychology, Evolution & Gender 4, 173–218. http://dx.doi.org/10.1080/1461666031000063665.

Graham, C.A., Sanders, S.A., Milhausen, R.R., 2006. The sexual excitation/sexual inhibition inventory for women: psychometric properties. Archives of Sexual Behavior 35, 397–409. http://dx.doi.org/10.1007/s10508-006-9041-7.

Gray, J.A., 1987. The Psychology of Fear and Stress. Cambridge University Press, New York.

Green, A.I., 2007. Queer theory and sociology: locating the subject and the self in sexuality studies. Sociological Theory 25, 26–45. http://dx.doi.org/10.1111/j.1467-9558.2007.00296.x.

Halperin, D.M., 2003. The normalization of queer theory. Journal of Homosexuality 45, 339–343. http://dx.doi.org/10.1300/J082v45n02_17.

Hogben, M., Byrne, D., 1998. Using social learning theory to explain individual differences in human sexuality. Journal of Sex Research 35, 58–71.

Kaan, H., 1844. Psychopathia Sexualis. Apud Leopoldum Voss, Leipzig, Germany.

von Krafft-Ebing, R., 1886. Psychopathia Sexualis. Verlag von Ferdinand Enke, Stuttgart, Germany.

Larsen, R.J., Buss, D.M., 2008. Personality Psychology. McGraw Hill, New York.

Lindau, S.T., Laumann, E.O., Levinson, W., Waite, L.J., 2003. Synthesis of scientific disciplines in pursuit of health: the interactive biopsychosocial model. Perspectives in Biology and Medicine 46 (Suppl. 3), S74–S86. http://dx.doi.org/10.1353/pbm.2003.0069.

Longmore, M.A., 1998. Symbolic interactionism and the study of sexuality. Journal of Sex Research 35, 44–57.

Moll, A., 1897. Untersuchungen über die Libido Sexualis, vol. 1. Fischer's medizinische Buchhandlung, H. Kornfeld, Berlin.

Nelson, A., 2010. Gender in Canada, fourth ed. Pearson Canada, Toronto, ON.

Person, E.S., 1980. Sexuality as the mainstay of identity: psychoanalytic perspectives. Signs 5, 605–630.

Petersen, J., Hyde, J.S., 2010. Gender differences in sexuality. In: Chrisler, J.C., McCreary, D.R. (Eds.), Handbook of Gender Research in Psychology: Gender Research in General and Experimental Psychology. Springer, New York, pp. 471–491. http://dx.doi.org/10.1007/978-1-4419-1465-1_23.

Reiss, I.L., 1986. Journey into Sexuality: An Exploratory Voyage. Prentice-Hall, Englewood Cliffs, NJ.

Reissing, E.D., Andruff, H.L., Wentland, J.J., 2011. Looking back: the experience of first sexual intercourse and current sexual adjustment in young heterosexual adults. Journal of Sex Research 49, 27–35. http://dx.doi.org/10.1080/00224499.2010.538951.

Reissing, E.D., Laliberté, G.M., Davis, H.J., 2005. Young women's sexual adjustment: the role of sexual self-schema, sexual self efficacy, sexual aversion and body attitudes. Canadian Journal of Human Sexuality 14, 63–75.

Roberts, S.C., 2012. Applying evolutionary psychology. In: Roberts, S.C. (Ed.), Applied Evolutionary Psychology. Oxford University Press, New York, pp. 1–6.

Rossi, A.S., 1994. Eros and caritas: a biopsychosocial approach to human sexuality and reproduction. In: Rossi, A.S. (Ed.), Sexuality across the Life Course. University of Chicago Press, London, pp. 3–36.

Rowland, D.L., 2006. The psychobiology of sexual arousal and response: physical and psychological factors that control our sexual response. In: McAnulty, R.D., Burnette, M. (Eds.), Sex and sexuality: Sexual function and dysfunction, vol. 2. Praeger Publishers, Westport, CT, pp. 37–66.

Sabini, J., 2004. Emotional responses to sexual and emotional infidelity: constants and differences across genders, samples, and methods. Personality and Social Psychology Bulletin 30, 1375–1388. http://dx.doi.org/10.1177/0146167204264012.

Sakaluk, J.K., Todd, L.M., Milhausen, R., Lachowsky, N.J., Undergraduate Research Group in Sexuality, 2013. Dominant heterosexual sexual scripts in emerging adulthood: conceptualization and measurement. Journal of Sex Research. http://dx.doi.org/10.1080/00224499.2012.745473. Advance online publication.

Salmon, C., 2012. The pop culture of sex: an evolutionary window on the worlds of pornography and romance. Review of General Psychology 16, 152–160. http://dx.doi.org/10.1037/a0027910.

Simon, W., Gagnon, J.H., 1986. Sexual scripts: permanence and change. Archives of Sexual Behavior 15, 91–120. http://dx.doi.org/10.1007/BF01542219.

Simon, W., Gagnon, J.H., 2003. Sexual scripts: origins, influences, and changes. Qualitative Sociology 26, 491–497. http://dx.doi.org/10.1023/B:QUAS.0000005053.99846.e5.

Skinner, B.F., 1974. About Behaviorism. Random House, New York.

Sprecher, S., 1998. Social exchange theories and sexuality. Journal of Sex Research 35, 32–43.

Sternberg, R.J., 1986. A triangular theory of love. Psychological Review 93, 119–135. http://dx.doi.org/10.1037/0033-295X.93.2.119.

Van Lankveld, J.D.M., Granot, M., Weijmar Schultz, W.C.M., Binik, I.Y., Wesselman, U., et al., 2010. Female sexual pain disorders. Journal of Sexual Medicine 7, 615–631. http://dx.doi.org/10.1000/j1743-6109.2009.01631.x.

Waller, W.W., Hill, R., 1951. The Family: A Dynamic Interpretation. Warner Books, New York.

Ward, T., Keenan, T., 1999. Child molesters' implicit theories. Journal of Interpersonal Violence 14, 821–838. http://dx.doi.org/10.1177/088626099014008003.

Weis, D.L., 2009. The need to integrate sexual theory and research. In: Wiederman, M.W., Whitley, J.B. (Eds.), Handbook for Conducting Research on Human Sexuality. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 7–24.

Wentland, J.J., Reissing, E.D., 2011. Taking casual sex not too casually: arriving at consensus definitions for different types of casual sexual relationships. Canadian Journal of Human Sexuality 20, 75–91.

Wentland, J.J., Reissing, E.D., September 2012. Friends with Benefits: Are They Really All that Casual?. Symposium presentation presented at the 39th annual meeting of the Canadian Sex Research Forum. Ottawa, Ontario.

Wilson, E.O., 1978. What is sociobiology? Society 15, 10–14. http://dx.doi.org/10.1007/BF02697770.

Sexually Transmitted Infections: Social Network Analysis

Yoosik Youm, Yonsei University, Seo-dae-moon-ku, Seoul, South Korea

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since its inception in the mid-1980s, social network analysis has changed the scene of sexually transmitted infection (STI) research in fundamental ways and brought about a paradigm shift. This article surveys the roles of sexual, social, and online networks in STI dynamics. Some challenges for future research are also considered.

In order to identify studies on sexually transmitted infections (STIs) that employed the social network approach, I searched PubMed for papers that contained the term 'social network.' Then, recognizing that there are other terms that are interchangeable with social network, I searched for papers that contained any of the following four terms in the title or abstract: 'social network,' 'social contact,' 'sexual network,' and 'sexual contact.' By reading each abstract, I determined whether the paper was about STIs. This quick survey, which was not comprehensive in any sense, is summarized in Figure 1. From this figure, we can infer three trends. First, there has been a sharp increase in STI studies containing these four terms since the mid-1970s. Second, out of the four related terms, 'sexual contact' appeared first in the early 1970s and continued to be the term most frequently used. Third, the other three terms began to appear in the mid-1980s and continued to grow in frequency except for 'social contact,' which appeared in only a very limited number of papers throughout the period examined. Since 2009, the difference in frequencies between 'sexual contact' and 'sexual network' (or 'social network') has shrunk significantly.

Closer examination of the papers counted in Figure 1, however, puts the picture in a somewhat different perspective. Although papers containing the term 'sexual contact' began to appear very early, the appearance of papers that actually adopted the social network approach did not appear until 1985. Before the social network approach was integrated into STI studies in concrete forms, the term 'sexual contact' was used to mean 'sexual intercourse,' as opposed to other types of interaction or contact leading to STI transmission. For example, In 1985, a case of phthiriasis palpebrarum was examined that was contracted by hand contact as opposed to sexual contact (Dornic, 1985). Another study from that year examined the possibility of acquired immune deficiency syndrome (AIDS) transmission through contacts with household objects without sexual contact (Kaplan et al., 1985).

A Newly Emerged Paradigm: The Social Network Approach

A very early adoption of the social network approach could be found in Rothenberg's study on core group transmission of

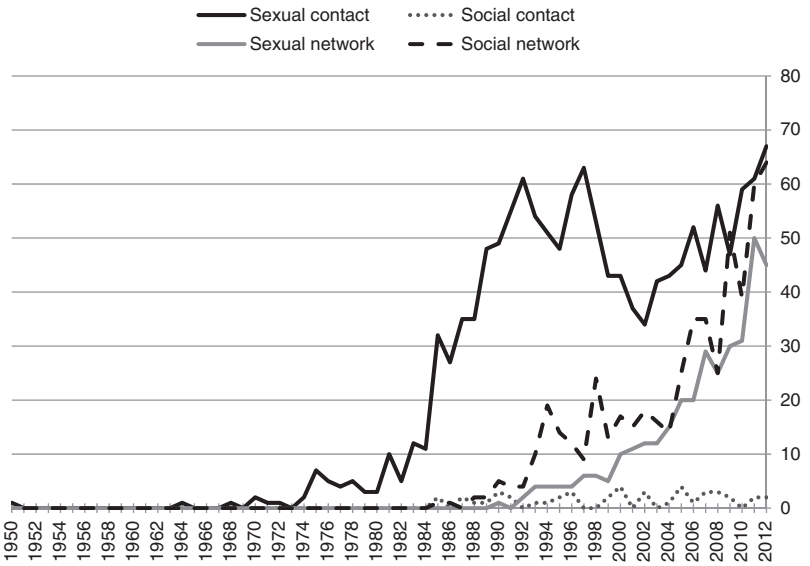

Figure 1 The number of STI studies that contained one of the following four terms in the title or abstract: 'sexual network,' 'sexual contact,' 'social network,' and 'social contact.' If a certain paper included more than one term, the paper was counted multiple times.

gonorrhea in the upstate New York area (Rothenberg, 1983). He identified core geographic areas and concentric circles of diminishing incidence in the city. Although he did not explicitly focus on certain social factors above and beyond individual-level risk factors, his results implied that the disease was not universally prevalent but concentrated in specific areas; thus, transmission or contact was by no means random. Instead, sexual contact was selective with regard to race and ethnicity or geographic area, and it must be essential to trace transmission paths (or networks) in addition to individual-level risk factors to understand diffusion dynamics.

In 1985, Potterat and colleagues developed this line of thought to a fuller extent and defined gonorrhea as a 'social disease' (Potterat et al., 1985). After researchers began to model core groups as small sections of the population that are very sexually active and thus responsible for transmission (Hethcote et al., 1982; May, 1981; Yorke et al., 1978), many empirical studies began to prove the existence of such core groups and to recognize that individuals' promiscuity was not sufficient to account for the breakout of STI epidemics (Bell et al., 1985; Fulford et al., 1983; Perine et al., 1985). In addition to individual-level risk factors, Potterat and colleagues realized that the pattern of sexual associations (or sexual networks) between core groups and other population groups must be taken into account; these authors, therefore, "attempted to make social groups the unit of analysis" (Potterat et al., 1985: p. 32).

Also in 1985, Klovdahl explicitly proposed utilizing the social network method to examine the spread of infectious diseases (Klovdahl, 1985). He argued that a basic proposition of social network research is that "the *structure* of a network has consequences for its individual members and for the network as a whole *over and above* effects of characteristics and behavior of the individuals involved" (Klovdahl, 1985: p. 1204, emphasis original). By applying social network methods to data on 40 patients with AIDS, he confirmed that an infectious agent was involved in its spread. He also argued that transmission dynamics are contingent on the structure of the relevant networks. Therefore, he concluded that increasing knowledge of network structures will help us to examine epidemic potential and develop containment strategies.

In the decades following the advent of the social network approach in the mid-1980s, many studies continued to apply this approach to STI dynamics and to focus on the structure of sexual networks (or ties) that exist *between* individuals rather than attributes and behaviors of individuals (e.g., Anderson et al., 1990; Aral and Holmes, 1990; Laumann and Youm, 1999; Morris et al., 1995; Morris and Kretzschmar, 1997; Rothenberg and Potterat, 1988; Rothenberg et al., 1998; Service and Blower, 1995). As Aral observed, this line of study defined a new paradigm for STI research (Aral, 1999). I will examine the role of sexual and social networks in STI transmission from the perspective of this paradigm. At the end of the discussion, I will also discuss some challenges for future research. To begin, I will describe the structural patterns of sexual networks that are known to be essential in the spread of STIs.

Structural Characteristics of Sexual Networks

We can approach sexual networks from several hierarchical perspectives, ranging from individual (or ego) to global. First, the network characteristics of a focal individual may be of interest. For example, the size of the sexual network (the number of sex partners) of a specific individual belongs to this category. Also of potential interest are the characteristics of dyadic sexual relationships such as the age difference or geographical distance between sexual partners. Furthermore, triadic sexual networks are sometimes relevant; these are sexual network patterns that exist among three people or areas. For example, if a certain community keeps a high volume of sexual ties with two other communities that are rarely sexually connected to each other, the first community is of great interest as a potential bridge. Finally, population- or group-level characteristics of a sexual network as a whole are also essential to understanding STI dynamics. For example, we might want to examine the mean or standard deviation of the number of sex partners making up the global sexual network. Ego, dyadic, and triadic network characteristics are often useful for identifying who is at higher risk of infection or who is most responsible for the continuation of STI epidemics or even pandemics. Global network properties could be utilized to determine the distinctive features of STI transmission in a certain population or group. In this section, I describe structural characteristics widely known to be critical in the spread of STIs.

Ego-Level Network Property: Size

Having a large number of sex partners is recognized as one of the most robust risk factors for infection. Numerous studies confirmed that people with multiple partners are more likely to become infected than are people with only one sex partner (Anderson et al., 1994; Ericksen and Trocki, 1992; Joffe et al., 1992; Melnick et al., 1993). The core group, one of the key concepts in early STI models, was itself defined by members' sexual promiscuity (May, 1981; Yorke et al., 1978). Therefore, the number of sex partners of a focal person is very much in line with the traditional approach of clinical medicine, which focuses on individual-level risk factors. In this sense, the number of sex partners of a focal person does not have any effect above and beyond individual behaviors.

Examining the number of partners of a focal person's partners, however, enables us to see STI dynamics in a completely different light. One of the basic propositions of the network approach to studying STIs is that a focal person's probability of becoming infected is dependent not only on his or her own risk factors but also on those of his or her partners. **Figure 2** illustrates this point based on a hypothetical sexual network. Person A and Person B share an identical level of individual risk and the same number of sex partners.

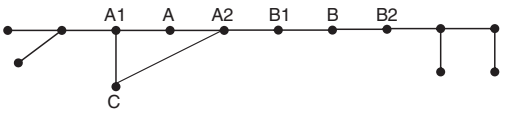

Figure 2 A hypothetical sexual network.

However, both partners of Person A (Person A1 and A2) have three sex partners, while both partners of Person B (Person B1 and B2) have only two sex partners. Other things being equal, Person A is more likely to be infected than Person B not because his or her own level of risk is higher but because his or her partners' level of risk is higher than that of Person B's partners.

Laumann and Youm explained the racial and ethnic differences in the prevalence of STIs in the United States based on this proposition (Laumann and Youm, 1999). They showed that even after controlling for other risk factors, including the number of sex partners, African Americans were about five times more likely to be infected by any bacterial STI than were whites. They proposed that this fact might be attributable to different mixing patterns between African Americans and whites. Using a nationally representative sample, these authors showed that in the African American population, there are many more sexual matches between core partners (having four or more sex partners in the previous year) and peripheral partners (having only one sex partner in the previous year) compared to whites. "For whites, if they are peripheral (instead of being core), they are 180 times more likely to have white peripheral partners (rather than white core partners). However, if they are African American peripherals, they are only 33 times more likely to have African American peripheral partners (rather than African American core partners)" (Laumann and Youm, 1999: p. 255). In other words, there were about six (= 180/33) times more sexual ties between core and peripheral partners within the African American population compared to whites. This means that even if both a peripheral African American and a peripheral white have only one sex partner, the former should have a higher risk of infection because his or her partner is six times more likely to be core.

Dyadic Network Property: Disassortative vs Assortative

One of the most examined questions surrounding STI epidemics is why certain subpopulations or areas have a relatively higher prevalence of infection. As one answer, many studies have singled out the mixing patterns in a sexual network. Other things being equal, disassortative mixing between subgroups eases the spread across the subgroups, while assortative mixing across subgroups contains the diffusion within each subgroup. For example, the relatively high prevalence of STIs among African Americans was explained by disassortative mixing between core and peripheral partners inside the African American population. Furthermore, the prevalence rate within the African American population is kept high due to disassortative mixing between racial and ethnic groups in the United States; African Americans tend to maintain sexual ties only with African Americans rather than with Hispanics or whites (Laumann and Youm, 1999). In a nutshell, STIs spilled over from the core to the entire African American population through disassortative mixing among African Americans, and infections were contained within the African American population because of racial and ethnic disassortative mixing in the United States.

While the previous example employed two dimensions of mixing patterns (racial–ethnic and sexual activity level, i.e., core and peripheral), other dimensions have also proven to be key in understanding STI dynamics. It is widely known that one of the reasons why there was not a more serious AIDS pandemic was the strong disassortative mixing with regard to sexual orientation: the small number of sexual ties between men who have sex with men (MSM) and people of other sexual orientations. In addition to sexual orientation, the dimensions of age and area are of great interest. Age-related disassortative-mixing patterns were confirmed to contribute to larger and faster diffusion of STIs. Younger women, especially adolescents, whose sexual partners were older men reported higher STI incidences in both Africa and the United States (Aral et al., 1999; Darroch et al., 1999; DiClemente et al., 2002; Gregson et al., 2002; Kelly et al., 2003). The effects may reveal a power imbalance in the dyad (or couple), with the older male dominating the sexual negotiation process, tending to lead to more unprotected sex. Conversely, middle-aged women with much younger partners are more frequently diagnosed with STIs in the United States (Kraut-Becher and Aral, 2006). In addition, disassortative mixing with regard to age among MSM seemed to fuel the spread of HIV in the United States. This observation might have resulted from the higher prevalence of HIV among older MSM (Service and Blower, 1996; Morris et al., 1995).

The last mixing pattern to be reviewed is disassortative mixing with regard to area or geographic distance. People who traveled to remote cities and visited sex workers regularly were given special attention because of their potential to spread STIs to geographically distant areas. For example, the behavioral characteristics and STI prevalence of long-distance truck drivers have been examined in Brazil (Lacerda et al., 1997), China (Chen et al., 2006), India (Gawande et al., 2000; Manjunath et al., 2002), Thailand (Morris et al., 1996), and the United States (Valway et al., 2009).

Although disassortative mixing generally makes the spread of STIs easier and faster, there is a tradeoff: In some cases, the number of successive infections decreases, making the diffusion weaker. To take a previous example, frequent sexual relations between core and peripheral partners could hinder transmission if peripheral people who become infected from core partners do not spread STIs further because of their small number of sexual partners. In other words, the basic reproduction rate (the number of people expected to be infected by one person with the STI) must exceed one in order for the STI to continue to exist in a population; disassortative mixing that is sufficiently strong could diminish this rate to below one depending on the transmission parameters, including latent duration, treatment lag, the probability of transmission per sex act, and so on.

Triadic Network Property: Betweenness

Many previous studies have focused on who is more likely to be infected, while only a small number have paid attention to who is more likely to infect other people – the infected vs the transmitter. Without a network approach, the distinction between the two appears meaningless, but the probability of being infected could differ from the power to infect. In **Figure 2**, we concluded that Person A is more likely to be infected than Person B although they have the same number of

sex partners. In order to prevent sexually transmitted disease (STD) transmission, however, protecting Person B might be more important than protecting Person A, as Person B occupies the only path between the two subgroups in the network. In contrast, Person A1 can transmit an STI to Person A2 through Person C without necessarily infecting Person A, thus showing the existence of redundant ties. In this sense, Person B is a much more powerful transmitter than Person A. The important role of bridging people in STI diffusion has been emphasized and empirically confirmed by some studies (Adams et al., 2013; Aral, 2000; Helleringer and Kohler, 2007; Kerani et al., 2003; Nordvik et al., 2007; Youm et al., 2009). For example, Kerani and colleagues confirmed the spatial bridges for gonorrhea and chlamydia infection in the Unites States (Kerani et al., 2003). Using a type of social network analysis called respondent-driven sample (RDS) in a study of drug users in the Chicago area, Youm and colleagues revealed eight *hidden bridging communities*: areas with below-average AIDS prevalence but above-average bridging activities. These communities were largely ignored by public health officials because of their low AIDS prevalence, although their high bridging activities were a key factor in the continuation of the infection in the Chicago area (Youm et al., 2009). The existence, size, and effect of bridging populations are empirical questions. One study showed that in Seattle, the proportion of gonococcal and chlamydial infections attributable to bridge populations was relatively small (Aral et al., 1999).

The challenge in studying bridging activities in STI dynamics is to find an appropriate measure. One of the most widely used network measures for these types of bridging activities is 'betweenness' (Freeman, 1977). The original betweenness score is measured as the frequency with which ego falls between other dyadic nodes (or people) on their shortest paths. An ego that is positioned between other people can facilitate or slow down the transmission; therefore, betweenness is crucial to understanding STI transmission dynamics and to developing effective preventive strategies. The original betweenness measure, however, is not adequate for STI research because it is based on shortest paths that do not allow any repeated points or links. For example, in **Figure 2**, the diffusion process A–A1–C–A2 is not considered when the betweenness score is calculated, given that only the shorter path, A–A2, is taken into account. In addition, reinfection cannot be considered when the betweenness score is calculated because the node (person) cannot appear multiple times in the diffusion process. For example, the diffusion A–A1–A cannot be included despite the fact that multiple infections are possible, except for viral infections to which the host becomes immune. Furthermore, if we change the unit of analysis from the individual to the group or community, multiple infections become the rule rather than the exception: Community A transmits an STI to B, and the STI goes back to community A.

One alternative way to measure bridging activities in STI transmission is to release the assumption of the shortest path. Instead of a *path* that does not allow any repeated nodes or lines, a *walk*, which is an alternating sequence of nodes and lines where an unlimited number of repeated nodes and lines is allowed, was proposed and termed 'walk-betweenness' (Youm, 2010). Walk-betweenness measures the proportion of walks that go through the ego among total walks from the source to the target. Although walk-betweenness is better than path-betweenness for STI research, it is still a universal measure of diffusion rather than one specifically devised for STIs. We need to continue to develop various ways to appropriately measure bridging activities in STI dynamics, incorporating various STI parameters such as transmission probability, infective duration, and latent duration.

Across-Time Network Property: Concurrency

To this point, we have examined the structural characteristics of sexual networks without considering the time dimension. Even when two people have the same number of sex partners, a person who maintains concurrent partnerships plays a different role in the spread of STIs with compared to a person who practices sequential monogamy. The potential role of concurrency was proposed and examined in mathematical models, which showed that concurrent partnerships can dramatically change STI dynamics and increase the number of infected people exponentially (Kretzschmar and Morris, 1996; Morris and Kretzschmar, 1995, 1997). One study also confirmed that sharp racial and ethnic differences in STI prevalence could be partly explained by high rates of concurrent partnerships among African Americans; the authors stated that the observed levels of concurrency and assortative racial mixing would produce a 2.6-fold racial disparity (Morris et al., 2009). Concurrency as a characteristic of a global sexual network in a population appears to fuel the spread of STIs in that population.

Concurrency as a property of individual sexual networks, however, has produced more complicated evidence. Some studies found that people with concurrent partners were more likely to have STI infections. For example, among US adolescents, having a greater number of concurrent partners was associated with STD diagnosis and exposure even after controlling for the number of sex partners (Rosenberg et al., 1999). Another US adolescent study, however, found that teens in sequential or concurrent relationships were more likely to report having an STD than were single-relationship teens (odds ratio (OR) = 2.3 and 3.9, respectively), but they were not statistically different from each other (Kelley et al., 2003). A representative sample of adults from the Chicago area even generated the opposite result; among adults with many lifetime sex partners (13 or more since the age of 18), people who had concurrent partners were less likely to report any STI (OR = 0.35) even after controlling for other risk factors, including the number of lifetime sex partners (Youm and Laumann, 2002). The authors suspected that the answer might lie in the different levels of social embeddedness between concurrency and serial monogamy: "If someone has many sex partners but with no overlap, each partnership is likely to be short and thus casual, approximating one-night stands between strangers with all the risk that entails. In this sense, concurrency may imply longer and thus less risky relationships for people with many sex partners" (Youm and Laumann, 2002: p. 695). This interpretation suggests that future studies may need to measure the duration of each partnership in addition to concurrency.

The Role of Social Networks in STI Transmission

Although Figure 1 shows a consistent increase in the number of publications that included the term 'social network' in the title or abstract since the mid-1980s, many of the counted papers adopted social network analysis and applied it to 'sexual network' data rather than to nonsexual social network data. However, sexual behaviors are among the social acts that contain very deep social meanings and occur under extremely serious social contexts. Thus, in order to understand risky sexual behaviors, it is necessary to examine the social milieux or social networks that produce, reinforce, and change sexual attitudes and behaviors (Bajos, 1997; Delor and Hubert, 2000). Focus group interviews revealed social reasons why certain people in North Carolina engaged in various risky sexual behaviors even when they knew they were dangerous: pervasive economic and racial oppression, lack of community recreation, boredom, and so on (Adimora et al., 2001). In this section, I intend to focus on the role of social (nonsexual) networks in shaping and changing sexual behaviors through factors such as normative pressure, social support, and proper information.

Normative Pressure or Social Support

Depending on the characteristics of members who exert pressure and support, social ties can either encourage or discourage risky sexual behaviors. For example, family members usually will strongly depress risky sexual behaviors. Among black MSM (BMSM), those with a higher family network proportion (having two or more family members in one's close network) were associated with less sex and drug use (OR = 0.38) and less participation in group sex (OR = 0.25) (Schneider et al., 2012). In addition, using a sample of 337 low-income urban African American women with main sexual partnerships, a study concluded that women who had larger nonsexual social networks and more family members in that network were less likely to be in concurrent partnerships (Grieb et al., 2012). Strong control exerted by sexual partners is another factor that strongly restrains unsafe sexual behaviors. One study based on a representative sample in Chicago revealed that among people with a moderate number of lifetime sex partners (1–12), people without strong dyadic (partner) control were more likely to be infected by any STI (Youm and Laumann, 2002).

In contrast to the positive effects of family members or partners on protective sexual behavior, normative pressure from friends could have the opposite effect. One longitudinal study examined how childhood and early adolescent (ages 6–15) peer and nonpeer social networks relate to sexual attitudes and behaviors in adolescence and young adulthood (ages 12–25); it revealed that having a large number of male peers in childhood is related to heightened sexual activity, sexual preoccupation, increased sexual pressure, and risky sexual behavior in adolescence (Noll et al., 2000). Another study used a representative sample of 305 heterosexually active homeless men utilizing meal programs in the Skid Row area of Los Angeles; this study found that the respondents' perception that their social ties (or network alters) engage in risky sex was associated with increased odds of their own concurrency (OR = 3.72) (Wenzel et al., 2012).

Information Provision

In addition to normative pressure and various kinds of social support, social networks are also venues for delivering necessary sexual information in a timely manner (Sprecher and McKinney, 1993). It is not unusual for adolescents to lack proper information about STIs or to underestimate their risk (Mellanby et al., 2000). While adolescents may need information regarding the risk factors of STIs, such as the number of sex partners, condom use, or intravenous (IV) drug use, adults may want information about finding and dealing with potential casual sex partners. A qualitative study on 14 Australian men who acquired HIV while traveling overseas found that social norms and social networks were essential to negotiating sex and social relations in unfamiliar foreign settings (Brown et al., 2012).

People especially rely on informal social networks for concrete information about sexual adventures, such as places where they can find casual but relatively safe sex partners, because "such information is more likely to be available, credible, trustworthy, timely, and specific in such networks" (Youm and Laumann, 2002). This type of information is more needed by people who are very sexually active than by people who have only a moderate number of sex partners.

One study implied that two distinctive advantages of social networks, control and information, could be effective for different subgroups (Youm and Laumann, 2002). Social control may be specifically effective for people with a moderate number of sex partners, given that, in order to avoid STIs, they do not need a constant information feed about potential venues for sexual activities. People with many sex partners, however, may receive protection from information but not from control. People may give up their control over their friends' sexual behaviors once those friends are known to be very sexually active.

Tools for Effective Preventive Programs

Because at-risk people are likely to connect to each other through intimate social networks, the use of such networks could be effective for recruiting, monitoring, and protecting people at potential risk. The US Centers for Disease Control and Prevention (CDC) did a pilot study in which people testing positive or negative (but still at risk) for HIV were required to recruit their social, sexual, and illicit drug partners (network associates (NAs)). They were able to recruit a substantial number of NAs with a high rate of newly diagnosed HIV infection and to do so efficiently (CDC, 2005). In a similar study funded by the CDC, the HIV prevalence of 5.6% among those recruited was significantly higher than the approximately 1% identified in other counseling, testing, and referral sites (Kimbrough et al., 2009).

This recruitment strategy is especially effective for small, hidden populations at high risk, such as MSM or IV drug users who are hard to reach using traditional recruitment methods. BMSM in San Francisco were asked to recruit any of their social contacts who were also BMSM, and 23% of the recruited BMSM

living with HIV were unaware of their infection (Fuqua et al., 2012). This type of snowball sampling was effective for purposes of preventive activities but not ideal for academic research, as the resultant sample should be strongly biased by the characteristics of the seeds who initiated the sampling. We can adjust for this bias to some extent by adopting new sampling procedures developed for social network analysis (Heckathorn, 1997, 2002). Sampling will be discussed in more depth when challenges for future research are presented in this article. For now, let us turn to a newly emerging type of social network: those that are online.

Online Networks as Socially (Un)embedded Places

The sexual-partnering scene has changed dramatically since the advent of the Internet. In particular, various new social networking services (SNSs), where online users are also creators of content, represent an unprecedented phenomenon at the societal level, and relatively little has been discovered about the impact of SNSs on sexual partnering or STI transmission dynamics.

It is intuitive to observe that online partnering leads to more risky behaviors given that many online partnerships are scarcely embedded in one's social surroundings; hence, such partnerships become more similar to sexual encounters between strangers. Many studies have confirmed that a substantial portion of MSM seek sexual partners online, and those who do so are more likely to have an STI history and to engage in high-risk sexual behaviors, including intentional barebacking (sex without condom use) (Benotsch et al., 2002; Elford et al., 2001; Halkitis and Parsons, 2003; Klausner et al., 2000; Liau et al., 2006; McFarlane et al., 2000; Toomey and Rothenberg, 2000).

However, many previous studies ignored the possibility of selection bias; the Internet per se did not increase risky behaviors, but people who relied on the Internet were, from the start, more likely to engage in hazardous behaviors. Some studies have reported the same level of risky behaviors between online and offline partnerships among MSM who looked for sex partners online (Bolding et al., 2005; Chiasson et al., 2007; Hirshfield et al., 2004). Studies on populations outside MSM, meanwhile, produced opposite results. Among men who have sex with women, people who reported Internet partners were less (relative risk = 0.66) likely to have *Chlamydia trachomatis* or *Neisseria gonorrhoeae* infections (Al-Tayyib et al., 2009). Even when the partnering begins online, if much online interaction occurs before the offline meeting and, thus, the partners are tightly embedded into each other's social milieux, the partnership can be safe (Rietmeijer and McFarlane, 2009). Thus, we must avoid making a simple dichotomous distinction between online and offline. Instead, closer examinations of the diverse embeddedness levels of online settings must be emphasized in future research.

Two major characteristics of the online world, relatively high anonymity (or little embeddedness) and the low cost of information diffusion, could represent an opportunity for preventive programs and activities. Online, it is possible to transmit socially stigmatized information, such as advertisements for STI tests, knowledge about STI infections, and partner notification for potential infection, at little cost (for a relatively complete list of online prevention activities, please see Rietmeijer and McFarlane, 2009). More time is needed to judge the effectiveness of online prevention efforts, but some early results do not look promising. Several studies on online STD partner notification programs revealed a degree of effectiveness (e.g., Ehlman et al., 2010) but in a very limited range (Plant et al., 2012; Kerani et al., 2011; Levine et al., 2008; Rietmeijer et al., 2011). One of the most recent studies evaluated InSPOT.org, the first self-contained partner notification website, which was launched in 2005 with funding from the Los Angeles County Department of Public Health Sexually Transmitted Disease Program and AIDS Healthcare Foundation. It was revealed that less than 2% of the respondents who were recruited at a sexual health clinic in Los Angeles had ever sent a notification using the online service (Plant et al., 2012).

The very limited effects of online networking sites on prevention and treatment are not due to there being a small number of visitors to these sites (Rietmeijer, 2012). Rather, it may result from the very nature of online networks: anonymity, low commitment, and low trust. While STI-related information is easily transmitted and advertised quickly and cheaply through online sites, people are not committed to respond unless the information is delivered to the right people at the right time. Unlike offline (social) networks that naturally can provide necessary information in a timely, credible, and specific manner, online networks distribute information in an anonymous, inopportune, and universal manner. A study found that the effectiveness of a syphilis partner notification program, part of the Los Angeles Gay and Lesbian Center's Sexual Program, greatly improved after the placement of community-embedded disease intervention specialists (CEDIS): 15% (21) of 140 partners elicited were identified as having early syphilis and started on treatment, compared to 0% of 13 partners elicited before the placement of CEDIS (Rudy et al., 2012). It is necessary to integrate offline social network support with online networking services for the best preventive outcomes.

Challenges for Future Network Studies

The last several decades have seen a dramatic increase in the use of social network analysis for STIs, beyond a mere increase in the number of studies. This article's discussion of the roles of social network approaches in STI research, although by no means comprehensive, suffices to offer proof of a paradigm shift. Now it has become the rule rather than the exception to take into account the characteristics of sexual partnerships when analyzing STI diffusion dynamics. The primary unit of analysis in STI research has steadily extended outward from individuals to dyadic couples to triadic relationships to global network characteristics. It is certain that future studies of STIs will adopt more elaborate measures, tools, and approaches to conduct social network analysis more fully. In this section, I discuss some challenges for future studies.

There is a need for new analytical tools. First, as discussed in this article, many of the original social network measures were developed for universal purposes, without specific applications

in mind. Some of them, moreover, were specifically designed for economic settings to analyze networks among individuals in firms and industries. The network dynamics involved in STI transmission are quite different from those seen in economic contexts. Thus, direct application of off-the-shelf network measures to STI research should be limited if possible. We are waiting for new measures of centrality, mixing patterns, and bridging activities that are specifically customized for STIs, taking various STI parameters into account. Second, we must adopt new tools to examine the characteristics of global networks. Unlike ego-centric sexual networks, global sexual networks cannot guarantee the independence assumption that is required for statistical analysis, as ties are not independent of each other. Social network analysis has seen a continuous improvement of stochastic models in terms of relaxing the independence assumption in various forms. Various types of exponential random graph models were advanced, the most recent being the development of the Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) for the model (Robins et al., 2007; Snijders et al., 2006; Wang et al., 2013; Wasserman and Pattison, 1996). We must stay attentive to the development of tools to analyze global networks.

Another hurdle is posed by the need to identify a proper sampling scheme. Many studies still adopt nonrepresentative sampling to collect sexual (social or drug) network data such as contact tracing, targeted sampling, or time venue sampling. The resultant network data usually have a strong bias toward people at high risk, reducing their suitability for rigorous scientific study (Laumann and Youm, 1999). Two alternatives are presently possible. First, we can conduct representative sampling of the general population. Examples include the National Health and Social Life Survey (Laumann and Youm, 1999), the National Survey of Family Growth (Kraut-Becher and Aral, 2006), the National Health and Nutrition Examination Survey (Allsworth et al., 2009), and the National Longitudinal Study of Adolescent Health (Pflieger et al., 2013). Representative sampling, however, requires a substantially larger investment of time and money. Furthermore, it is not suited for studying small, hard-to-reach populations such as MSM or IV drug users; from the start of sampling, it is quite hard to identify people who belong to such small and hidden populations. For such populations, snowball sampling has been widely used, but as in other forms of nonrepresentative sampling, the final sample usually has a large bias toward the seeds, or the respondents initiating the snowballing.

Fortunately, RDS was proposed to adjust for the biases of snowball sampling (Heckathorn, 1997, 2002) and has been adopted by many STI studies (Jenness et al., 2011; Johnson et al., 2009; Mimiaga et al., 2007; Neaigus et al., 2013; Youm et al., 2009). Although RDS provides statistical estimates and variances, whether the assumptions for the inference can be met in practice is not certain (Lansky et al., 2012; McCreesh et al., 2012; Rudolph et al., 2011, 2013). However, other types of applications such as risk factor identification, efficient network data collection, and population size estimation are also possible, and these applications require fewer (or no) sampling assumptions (White et al., 2012).

Another data-related issue is the integration of biomarkers. Given the strong social stigma attached to STIs and the high proportion of asymptomatic STI cases, self-reported STI status tends to be underestimated, and biomarkers from physical examination for STIs are helpful for more accurate estimation of respondents' STI status (Kraut-Becher and Aral, 2006). In addition, data on sexual networks from surveys are inevitably incomplete and egocentric: respondents tend to omit information about some of their sexual partners. Furthermore, we usually do not and cannot interview all of the sex partners of respondents. Under the assumption that individuals infected by the same strain of an infectious agent are more likely to have a sexual link, molecular typing can produce a sexual network: a biomarker-produced sexual network (Day et al., 1998; Hesse et al., 1995; Ota et al., 2011; Wylie et al., 2005). Given that many health-related survey studies have begun to accommodate biomarkers in their data sets, we can expect to see more STI data sets where network surveys are combined with biomarkers. There is a need to identify tools to build a more complete and accurate sexual network by combining a survey-based sexual network with a biomarker-produced sexual network.

See also: Clustering and Cohesion in Networks: Concepts and Measures; Core–Periphery Patterns in Networks; Fertility and Social Networks; Method of Moments and Generalized Method of Moments; Models of Influence in Networks; Models of Power in Exchange Networks; Network Analysis; Network Centrality, Measures of; Sexual Behavior and Social Networks; Social Integration, Social Networks, and Health; Social Network Algorithms and Software; Social Network Analysis; Social Networks, Algebra of; Social Networks.

Bibliography

Adams, J., Moody, J., Morris, M., 2013. Sex, drugs, and race: how behaviors differentially contribute to the sexually transmitted infection risk network structure. American Journal of Public Health 103, 322–329.

Adimora, A.A., Schoenbach, V.J., Martinson, F.E., Donaldson, K.H., Fullilove, R.E., Aral, S.O., 2001. Social context of sexual relationships among rural African Americans. Sexually Transmitted Diseases 28, 69–76.

Allsworth, J.E., Ratner, J.A., Peipert, J.F., 2009. Trichomoniasis and other sexually transmitted infections: results from the 2001–2004 National Health and Nutrition Examination Surveys. Sexually Transmitted Diseases 36, 738–744.

Al-Tayyib, A.A., McFarlane, M., Kachur, R., Rietmeijer, C.A., 2009. Finding sex partners on the internet: what is the risk for sexually transmitted infections? Sexually Transmitted Infections 85, 216–220.

Anderson, J.E., McCormick, L., Fichtner, R., 1994. Factors associated with self-reported STDs: data from a national survey. Sexually Transmitted Diseases 21, 303–308.

Anderson, R.M., Gupta, S., Ng, W., 1990. The significance of sexual partner contact networks for the transmission dynamics of HIV. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 3, 417–429.

Aral, S.O., 1999. Sexual network patterns as determinants of STD rates: paradigm shift in the behavioral epidemiology of STDs made visible. Sexually Transmitted Diseases 26, 262–264.

Aral, S.O., 2000. Behavioral aspects of sexually transmitted diseases: core groups and bridge populations. Sexually Transmitted Diseases 27, 327–328.

Aral, S.O., Holmes, K.K., 1990. Epidemiology of sexual behavior and sexually transmitted diseases. In: Holmes, K.K., Mardh, P.A., Sparling, P.F.,

Wiesner, P.J. (Eds.), Sexually Transmitted Diseases, second ed. McGraw Hill Information Services Company, New York, pp. 19–36.
Aral, S.O., Hughes, J.P., Stoner, B., Whittington, W., Handsfield, H.H., Anderson, R.M., Holmes, K.K., 1999. Sexual mixing patterns in the spread of gonococcal and chlamydial infections. American Journal of Public Health 89, 825–833.
Bajos, N., 1997. Social factors and the process of risk construction in HIV sexual transmission. AIDS Care 9, 227–237.
Bell, T.A., Farrow, J.A., Stamm, W.E., Critchlow, C.W., Holmes, K.K., 1985. Sexually transmitted diseases in females in a juvenile detention center. Sexually Transmitted Diseases 12, 140–144.
Benotsch, E.G., Kalichman, S., Cage, M., 2002. Men who have met sex partners via the Internet: prevalence, predictors, and implications for HIV prevention. Archives of Sexual Behavior 31, 177–183.
Bolding, G., Davis, M., Hart, G., Sherr, L., Elford, J., 2005. Gay men who look for sex on the Internet: is there more HIV/STI risk with online partners? AIDS 19, 961–968.
Brown, G., Ellard, J., Mooney-Somers, J., Hildebrand, J., Langdon, T., 2012. HIV risk among Australian men travelling overseas: networks and context matter. Culture, Health & Sexuality 14, 677–690.
Center for Disease Control, 2005. Use of social networks to identify persons with undiagnosed HIV infection – seven U.S. cities, October 2003–September 2004. Morbidity and Mortality Weekly Report 54, 601–605.
Chen, X.S., Yin, Y.P., Gong, X.D., Liang, G.J., Zhang, W.Y., Poumerol, G., Shi, M.Q., Wu, S.Q., Zhang, G.C., 2006. Prevalence of sexually transmitted infections among long-distance truck drivers in Tongling, China. International Journal of STD and AIDS 17, 304–308.
Chiasson, M.A., Hirshfield, S., Remien, R.H., Humberstone, M., Wong, T., Wolitski, R.J., 2007. A comparison of on-line and off-line sexual risk in men who have sex with men: an event-based on-line survey. Journal of Acquired Immune Deficiency Syndromes 44, 235–243.
Darroch, J.E., Landry, D.J., Oslak, S., 1999. Age differences between sexual partners in the United States. Family Planning Perspectives 31, 160–167.
Day, S., Ward, H., Ison, C., Bell, G., Weber, J., 1998. Sexual networks: the integration of social and genetic data. Social Science and Medicine 47, 1981–1992.
Delor, F., Hubert, M., 2000. Revisiting the concept of "vulnerability". Social Science and Medicine 50, 1557–1570.
DiClemente, R.J., Wingood, G.M., Crosby, R.A., Sionean, C., Cobb, B.K., Harrington, K., Davies, S.L., Hook 3rd, E.W., Oh, M.K., 2002. Sexual risk behaviors associated with having older sex partners: a study of black adolescent females. Sexually Transmitted Diseases 29, 20–24.
Dornic, D.I., 1985. Ectoparasitic infestation of the lashes. Journal of the American Optometric Association 56, 716–719.
Ehlman, D.C., Jackson, M., Saenz, G., Novak, D.S., Kachur, R., Heath, J.T., Furness, B.W., 2010. Evaluation of an innovative internet-based partner notification program for early syphilis case management, Washington, DC, January 2007–June 2008. Sexually Transmitted Diseases 37, 478–485.
Elford, J., Bolding, G., Sherr, L., 2001. Seeking sex on the internet and sexual risk behaviour among gay men using London gyms. AIDS 15, 1409–1415.
Ericksen, K.P., Trocki, K.F., 1992. Behavioral risk factors for sexually transmitted diseases in American households. Social Science and Medicine 34, 843–853.
Freeman, L.C., 1977. A set of measures of centrality based on betweeness. Sociometry 40 (1), 35–41.
Fulford, K.W., Catterall, R.D., Hoinville, E., Lim, K.S., Wilson, G.D., 1983. Social and psychological factors in the distribution of STD in male clinic attenders. III Sexual activity. British Journal of Venereal Diseases 59, 386–393.
Fuqua, V., Chen, Y.H., Packer, T., Dowling, T., Ick, T.O., Nguyen, B., Colfax, G.N., Raymond, H.F., 2012. Using social networks to reach black MSM for HIV testing and linkage to care. AIDS and Behavior 16, 256–265.
Gawande, A.V., Vasudeo, N.D., Zodpey, S.P., Khandait, D.W., 2000. Sexually transmitted infections in long distance truck drivers. The Journal of Communicable Diseases 32, 212–215.
Gregson, S., Nyamukapa, C.A., Garnett, G.P., Mason, P.R., Zhuwau, T., Carael, M., Chandiwana, S.K., Anderson, R.M., 2002. Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. Lancet 359, 1896–1903.
Grieb, S.M.D., Davey-Rothwell, M., Latkin, C.A., 2012. Social and sexual network characteristics and concurrent sexual partnerships among urban African American high-risk women with main sex partners. Aids and Behavior 16, 882–889.
Halkitis, P.N., Parsons, J.T., 2003. Intentional unsafe sex (barebacking) among HIV-positive gay men who seek sexual partners on the internet. AIDS Care 15, 367–378.
Heckathorn, D.D., 1997. Respondent-driven sampling II: a new approach to the study of hidden populations. Social Problems 44, 174–199.
Heckathorn, D.D., 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Social Problems 49, 11–34.
Helleringer, S., Kohler, H.P., 2007. Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. AIDS 21, 2323–2332.
Hesse, S., Low, N., Molokwu, B., Uttley, A.H., Jephcott, A.E., Turner, A., Pozniak, A.L., 1995. Molecular typing as an epidemiological tool in the study of sexual networks. Lancet 346, 976–977.
Hethcote, H.W., Yorke, J.A., Nold, A., 1982. Gonorrhea modeling: a comparison of control methods. Mathematical Biosciences 58 (1), 93–109.
Hirshfield, S., Remien, R.H., Humberstone, M., Walavalkar, I., Chiasson, M.A., 2004. Substance use and high-risk sex among men who have sex with men: a national online study in the USA. AIDS Care-Psychological and Socio-Medical Aspects of AIDS/HIV 16, 1036–1047.
Jenness, S.M., Begier, E.M., Neaigus, A., Murrill, C.S., Wendel, T., Hagan, H., 2011. Unprotected anal intercourse and sexually transmitted diseases in high-risk heterosexual women. American Journal of Public Health 101, 745–750.
Joffe, G.P., Foxman, B., Schmidt, A.J., Farris, K.B., Carter, R.J., Neumann, S., Tolo, K.A., Walters, A.M., 1992. Multiple partners and partner choice as risk factors for sexually transmitted disease among female college students. Sexually Transmitted Diseases 19, 272–278.
Johnson, C.V., Mimiaga, M.J., Reisner, S.L., Tetu, A.M., Cranston, K., Bertrand, T., Novak, D.S., Mayer, K.H., 2009. Health care access and sexually transmitted infection screening frequency among at-risk Massachusetts men who have sex with men. American Journal of Public Health 99 (Suppl. 1), S187–S192.
Kaplan, J.E., Oleske, J.M., Getchell, J.P., Kalyanaraman, V.S., Minnefor, A.B., Zabala-Ablan, M., Joshi, V., Denny, T., Cabradilla, C.D., Rogers, M.F., et al., 1985. Evidence against transmission of human T-lymphotropic virus/lymphadenopathy-associated virus (HTLV-III/LAV) in families of children with the acquired immunodeficiency syndrome. The Pediatric Infectious Disease Journal 4, 468–471.
Kelley, S.S., Borawski, E.A., Flocke, S.A., Keen, K.J., 2003. The role of sequential and concurrent sexual relationships in the risk of sexually transmitted diseases among adolescents. The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine 32, 296–305.
Kelly, R.J., Gray, R.H., Sewankambo, N.K., Serwadda, T., Wabwire-Mangen, F., Lutalo, T., Wawer, M.J., 2003. Age differences in sexual partners and risk of HIV-1 infection in rural Uganda. Journal of Acquired Immune Deficiency Syndromes: JAIDS 32, 446–451.
Kerani, R.P., Golden, M.R., Whittington, W.L., Handsfield, H.H., Hogben, M., Holmes, K.K., 2003. Spatial bridges for the importation of gonorrhea and chlamydial infection. Sexually Transmitted Diseases 30, 742–749.
Kerani, R.P., Fleming, M., Deyoung, B., Golden, M.R., 2011. A randomized, controlled trial of inSPOT and patient-delivered partner therapy for gonorrhea and chlamydial infection among men who have sex with men. Sexually Transmitted Diseases 38, 941–946.
Kimbrough, L.W., Fisher, H.E., Jones, K.T., Johnson, W., Thadiparthi, S., Dooley, S., 2009. Accessing social networks with high rates of undiagnosed HIV infection: the social networks demonstration project. American Journal of Public Health 99, 1093–1099.
Klausner, J.D., Wolf, W., Fischer-Ponce, L., Zolt, I., Katz, M.H., 2000. Tracing a syphilis outbreak through cyberspace. JAMA: The Journal of the American Medical Association 284, 447–449.
Klovdahl, A.S., 1985. Social networks and the spread of infectious diseases: the AIDS example. Social Science & Medicine 21, 1203–1216.
Kraut-Becher, J.R., Aral, S.O., 2006. Patterns of age mixing and sexually transmitted infections. International Journal of STD & AIDS 17, 378–383.
Kretzschmar, M., Morris, M., 1996. Measures of concurrency in networks and the spread of infectious disease. Mathematical Biosciences 133, 165–195.
Lacerda, R., Gravato, N., Mcfarland, W., Rutherford, G., Iskrant, K., Stall, R., Hearst, N., 1997. Truck drivers in Brazil: prevalence of HIV and other sexually transmitted diseases, risk behavior and potential for spread of infection. AIDS 11 (Suppl. 1), S15–S19.
Lansky, A., Drake, A., Wejnert, C., Pham, H., Cribbin, M., Heckathorn, D.D., 2012. Assessing the assumptions of respondent-driven sampling in the national HIV Behavioral Surveillance System among injecting drug users. The Open AIDS Journal 6, 77–82.
Laumann, E.O., Youm, Y., 1999. Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation. Sexually Transmitted Diseases 26, 250–261.
Levine, D., Woodruff, A.J., Mocello, A.R., Lebrija, J., Klausner, J.D., 2008. inSPOT: the first online STD partner notification system using electronic postcards. Plos Medicine 5, 1428–1431.

Liau, A., Millett, G., Marks, G., 2006. Meta-analytic examination of online sex-seeking and sexual risk behavior among men who have sex with men. Sexually Transmitted Diseases 33, 576–584.

Manjunath, J.V., Thappa, D.M., Jaisankar, T.J., 2002. Sexually transmitted diseases and sexual lifestyles of long-distance truck drivers: a clinico-epidemiologic study in south India. International Journal of STD and AIDS 13, 612–617.

May, R.M., 1981. The transmission and control of gonorrhea. Nature 291, 376.

McCreesh, N., Frost, S.D., Seeley, J., Katongole, J., Tarsh, M.N., Ndunguse, R., Jichi, F., Lunel, N.L., Maher, D., Johnston, L.G., Sonnenberg, P., Copas, A.J., Hayes, R.J., White, R.G., 2012. Evaluation of respondent-driven sampling. Epidemiology 23, 138–147.

McFarlane, M., Bull, S.S., Rietmeijer, C.A., 2000. The Internet as a newly emerging risk environment for sexually transmitted diseases. JAMA: The Journal of the American Medical Association 284, 443–446.

Mellanby, A., Phelps, F., Lawrence, C., Tripp, J.H., 2000. Teenagers and the risks of sexually transmitted diseases: a need for the provision of balanced information. 1992. Sexually Transmitted Infections 76 (Suppl. 1), S43.

Melnick, S.L., Burke, G.L., Perkins, L.L., Mccreath, H., Gilbertson, D.T., Sidney, S., Hulley, S.B., 1993. Sexually transmitted diseases among young heterosexual urban adults. Public Health Reports 108, 673–679.

Mimiaga, M.J., Goldhammer, H., Belanoff, C., Tetu, A.M., Mayer, K.H., 2007. Men who have sex with men: perceptions about sexual risk, HIV and sexually transmitted disease testing, and provider communication. Sexually Transmitted Diseases 34, 113–119.

Morris, M., Kretzschmar, M., 1995. Concurrent partnerships and transmission dynamics in networks. Social Networks 17, 299–318.

Morris, M., Kretzschmar, M., 1997. Concurrent partnerships and the spread of HIV. AIDS 11, 641–648.

Morris, M., Zavisca, J., Dean, L., 1995. Social and sexual networks: their role in the spread of HIV/AIDS among young gay men. AIDS Education and Prevention 7, 24–35.

Morris, M., Podhisita, C., Wawer, M.J., Handcock, M.S., 1996. Bridge populations in the spread of HIV/AIDS in Thailand. AIDS 10, 1265–1271.

Morris, M., Kurth, A.E., Hamilton, D.T., Moody, J., Wakefield, S., The Network Modeling Group, 2009. Concurrent partnerships and HIV prevalence disparities by race: linking science and public health practice. American Journal of Public Health 99, 1023–1031.

Neaigus, A., Reilly, K.H., Jenness, S.M., Hagan, H., Wendel, T., Gelpi-Acosta, C., 2013. Dual HIV risk: receptive syringe sharing and unprotected sex among HIV-negative injection drug users in New York City. AIDS and Behavior 17 (7), 2501–2509.

Noll, J.G., Trickett, P.K., Putnam, F.W., 2000. Social network constellation and sexuality of sexually abused and comparison girls in childhood and adolescence. Child Maltreatment 5, 323–337.

Nordvik, M.K., Liljeros, F., Osterlund, A., Herrmann, B., 2007. Spatial bridges and the spread of chlamydia: the case of a county in Sweden. Sexually Transmitted Diseases 34, 47–53.

Ota, K.V., Ng, L.K., Melano, R.G., Martin, I.E., Brown, E.M., Richardson, S.E., Fisman, D.N., Low, D.E., Jamieson, F.B., 2011. Identification of sexual networks through molecular typing of quinolone-resistant *Neisseria gonorrhoeae* in Ontario, Canada. Sexually Transmitted Diseases 38, 811–814.

Perine, P.L., Handsfield, H.H., Holmes, K.K., Blount, J.H., 1985. Epidemiology of the sexually transmitted diseases. Annual Review of Public Health 6, 85–106.

Pflieger, J.C., Cook, E.C., Niccolai, L.M., Connell, C.M., 2013. Racial/ethnic differences in patterns of sexual risk behavior and rates of sexually transmitted infections among female young adults. American Journal of Public Health 103, 903–909.

Plant, A., Rotblatt, H., Montoya, J.A., Rudy, E.T., Kerndt, P.R., 2012. Evaluation of inSPOTLA.org: an internet partner notification service. Sexually Transmitted Diseases 39, 341–345.

Potterat, J.J., Rothenberg, R.B., Woodhouse, D.E., Muth, J.B., Pratts, C.I., Fogle 2nd, J.S., 1985. Gonorrhea as a social disease. Sexually Transmitted Diseases 12, 25–32.

Rietmeijer, C.A., 2012. Online interventions for sexually transmitted infection/HIV prevention—what next? Sexually Transmitted Diseases 39, 346–347.

Rietmeijer, C.A., McFarlane, M., 2009. Web 2.0 and beyond: risks for sexually transmitted infections and opportunities for prevention. Current Opinion in Infectious Diseases 22, 67–71.

Rietmeijer, C.A., Westergaard, B., Mickiewicz, T.A., Richardson, D., Ling, S., Sapp, T., Jordan, R., Wilmoth, R., Kachur, R., Mcfarlane, M., 2011. Evaluation of an online partner notification program. Sexually Transmitted Diseases 38, 359–364.

Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P., 2007. Recent developments in exponential random graph (p*) models for social networks. Social Networks 29, 192–215.

Rosenberg, M.D., Gurvey, J.E., Adler, N., Dunlop, M.B., Ellen, J.M., 1999. Concurrent sex partners and risk for sexually transmitted diseases among adolescents. Sexually Transmitted Diseases 26, 208–212.

Rothenberg, R.B., 1983. The geography of gonorrhea: empirical demonstration of core group transmission. American Journal of Epidemiology 117, 688–694.

Rothenberg, R.B., Potterat, J.J., 1988. Temporal and social aspects of gonorrhea transmission: the force of infectivity. Sexually Transmitted Diseases 15, 88–92.

Rothenberg, R.B., Potterat, J.J., Woodhouse, D.E., Muth, S.Q., Darrow, W.W., Klovdahl, A.S., 1998. Social network dynamics and HIV transmission. AIDS 12, 1529–1536.

Rudolph, A.E., Crawford, N.D., Latkin, C., Heimer, R., Benjamin, E.O., Jones, K.C., Fuller, C.M., 2011. Subpopulations of illicit drug users reached by targeted street outreach and respondent-driven sampling strategies: implications for research and public health practice. Annals of Epidemiology 21, 280–289.

Rudolph, A.E., Fuller, C.M., Latkin, C., 2013. The importance of measuring and accounting for potential biases in respondent-driven samples. AIDS and Behavior 17, 2244–2252.

Rudy, E.T., Aynalem, G., Cross, J., Ramirez, F., Bolan, R.K., Kerndt, P.R., 2012. Community-embedded disease intervention specialist program for syphilis partner notification in a clinic serving men who have sex with men. Sexually Transmitted Diseases 39, 701–705.

Schneider, J., Michaels, S., Bouris, A., 2012. Family network proportion and HIV risk among black men who have sex with men. Journal of Acquired Immune Deficiency Syndromes: JAIDS 61, 627–635.

Service, S.K., Blower, S.M., 1995. HIV transmission in sexual networks – an empirical analysis. Proceedings of the Royal Society B: Biological Sciences 260, 237–244.

Service, S.K., Blower, S.M., 1996. Linked HIV epidemics in San Francisco. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 11, 311–313.

Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S., 2006. New specifications for exponential random graph models. Sociological Methodology 36, 99–153.

Sprecher, S., McKinney, K., 1993. Sexuality. Sage Publications, Newbury Park.

Toomey, K.E., Rothenberg, R.B., 2000. Sex and cyberspace-virtual networks leading to high-risk sex. JAMA: The Journal of the American Medical Association 284, 485–487.

Valway, S., Jenison, S., Keller, N., Vega-Hernandez, J., Hubbard Mccree, D., 2009. Risk assessment and screening for sexually transmitted infections, HIV, and hepatitis virus among long-distance truck drivers in New Mexico, 2004–2006. American Journal of Public Health 99, 2063–2068.

Wang, P., Robins, G., Pattison, P., Lazega, E., 2013. Exponential random graph models for multilevel networks. Social Networks 35, 96–115.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks. 1. An introduction to Markov graphs and p. Psychometrika 61, 401–425.

Wenzel, S.L., Rhoades, H., Hsu, H.T., Golinelli, D., Tucker, J.S., Kennedy, D.P., Green, H.D., Ewing, B., 2012. Behavioral health and social normative influence: correlates of concurrent sexual partnering among heterosexually-active homeless men. AIDS and Behavior 16, 2042–2050.

White, R.G., Lansky, A., Goel, S., Wilson, D., Hladik, W., Hakim, A., Frost, S.D., 2012. Respondent driven sampling—where we are and where should we be going? Sexually Transmitted Infections 88, 397–399.

Wylie, J.L., Cabral, T., Jolly, A.M., 2005. Identification of networks of sexually transmitted infection: a molecular, geographic, and social network analysis. Journal of Infectious Diseases 191, 899–906.

Yorke, J.A., Hethcote, H.W., Nold, A., 1978. Dynamics and control of the transmission of gonorrhea. Sexually Transmitted Diseases 5, 51–56.

Youm, Y., 2010. A sociological interpretation of emerging properties in STI transmission dynamics: walk-betweenness of sexual networks. Sexually Transmitted Infections 86 (Suppl. 3), iii24–iii28.

Youm, Y., Laumann, E.O., 2002. Social network effects on the transmission of sexually transmitted diseases. Sexually Transmitted Diseases 29, 689–697.

Youm, Y., Mackesy-Amiti, M.E., Williams, C.T., Ouellet, L.J., 2009. Identifying hidden sexual bridging communities in Chicago. Journal of Urban Health 86 (Suppl. 1), 107–120.

Shadow Work: Measurement

Friedrich Schneider, Johannes Kepler University of Linz, Linz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article the main focus lies on the shadow economy labor force in OECD, developing, and transition countries. Besides informal employment in the rural and nonrural sector also other measures of informal employment like the share of women and men are shown. The most influential factors on the shadow economy labor force are tax policies and state regulation, which, if they rise, increase both. Furthermore, the discussion of the recent microstudies underlines that economic opportunities, the overall burden of the state (taxes and regulations), the general situation on the labor market, and unemployment are especially crucial for an understanding of the dynamics the shadow labor force.

Introduction

Fighting informal (illegal) or shadow employment has been an important policy goal in most countries during recent decades. In order to do this one should have knowledge about the size and development of shadow economy labor force as well as the reasons why people are engaged in shadow economy activities. In this mostly empirically orientated article I am mainly concerned with the size and development of black activities or undeclared work, and the shadow economy labor force.

This article is organized as follows: Section Some Theoretical Considerations about the Shadow Economy Labor Force and Labor Market presents theoretical considerations about the shadow economy labor force and labor market. In Section Shadow Economy Labor Force – Empirical Results a discussion of the size and development of the shadow economy labor force and its various aspects is undertaken. In Section Shadow Economy and Unemployment the interaction between the shadow economy and unemployment is analyzed. Finally Section Conclusions concludes.

Some Theoretical Considerations about the Shadow Economy Labor Force and Labor Market

Having examined the size, rise and fall of the shadow economy in terms of value added over time, the analysis now focuses on the 'shadow labor market,' as within the official labor market there is a particularly tight relationship and 'social network' between people who are active in the shadow economy. (Pioneering work in this area has been done by L. Frey (1972, 1975, 1978, 1980), Jöreskog and Goldberger (1975), Cappiello (1986), Lubell (1991), Pozo (1996) and Tanzi (1999). One of the latest surveys or background paper is Hazans (2011) dealing with informal workers across Europe.) Moreover, by definition every activity in the shadow economy involves a 'shadow labor market' to some extent: (Compare also one of the latest OECD reports with the title *Is Informal Normal: Toward More and Better Jobs* by the OECD (2009).) Hence, the 'shadow labor market' includes all cases, where the employees or the employers, or both, occupy a 'shadow economy position.'

Why do people work in the shadow economy? In the official labor market, the costs firms (and individuals) have to pay when 'officially' hiring someone are tremendously increased by the burden of tax and social contributions on wages, as well as by the legal administrative regulation to control economic activity. In various OECD countries, these costs are greater than the wage effectively earned by the worker – providing a strong incentive to work in the shadow economy.

More detailed theoretical information on the labor supply decision in the underground economy is given by Lemieux et al. (1994) who use microdata from a survey conducted in Quebec City (Canada). (Compare also the microstudy of the shadow labor market for Germany by Haigner et al. (2013).) In particular, their study provides some economic insights regarding the size of the distortion caused by income taxation and the welfare system. The results of this study suggest that hours worked in the shadow economy are quite responsive to changes in the net wage in the regular (official) sector. Their empirical results attribute this to a (mis-) allocation of work from the official to the informal sector, where it is not taxed. In this case, the substitution between labor market activities in the two sectors is quite high. These empirical findings indicate, that "participation rates and hours worked in the underground sector also tend to be inversely related to the number of hours worked in the regular sector" (Lemieux et al., 1994: p. 235). These findings demonstrate a large negative elasticity of hours worked in the shadow economy with respect both to the wage rate in the regular sector as well as to a high mobility between the sectors.

Illicit work can take many forms. The underground use of labor may consist of a second job after (or even during) regular working hours. A second form is shadow economy work by individuals who do not participate in the official labor market. A third component is the employment of people (e.g., clandestine or illegal immigrants), who are not allowed to work in the official economy. Empirical research on the shadow economy labor market is even more difficult than that of the shadow economy on the value added, since one has very little knowledge about how many hours an average 'shadow economy worker' is actually working (from full time to a few hours, only); hence, it is not easy to provide empirical facts (For developing countries some literature about the shadow labor market exists; compare Dallago (1990), Pozo (1996), Loayza (1996), Chickering and Salahdine (1991), and OECD (2009).).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.94036-4

Kucera and Roncolato (2008: p. 321) also deal with informal employment. They address two issues of crucial importance to labor market policy:

1. The intensive labor market regulations as one (major) cause of informal employment, and
2. the so-called 'voluntary' informal employment. Kucera and Roncolato give a theoretical overview on both issues and also a survey of a number of empirical studies, in which mainly the effect of official labor market regulations on informal employment is analyzed, where they find a significant and quantitatively important influence.

Shadow Economy Labor Force – Empirical Results

Worldwide Aspects – Latest Results

The following results of the shadow economy labor force are based on the OECD and World Bank database on informal employment in major cities and in rural areas, as well as on other sources mentioned in the footnotes of this article and the tables. The values of the shadow economy labor force are calculated in absolute terms, and as a percentage of the official labor force, under the assumption that the shadow economy in rural areas is at least as high as in the cities. This is a conservative assumption, since in reality it is likely to be even larger. (The assumption that the shadow economy labor force is at least as high in rural areas as in major cities, is a very modest one and is supported by Lubell (1991). Some authors (e.g., Lubell (1991), Pozo (1996), and Chickering and Salahdine (1991)) argue that the illicit labor force is nearly twice as high in the countryside as in urban areas. But since no (precise) data exists on this ratio, the assumption of an equal size may be justified arguing that such a calculation provides at least minimal figures.) Survey techniques and, for some countries, the MIMIC method, and the method of the discrepancy between the official and actual labor force are used for estimation.

One of the latest studies is the OECD (2009) (see also OECD (2008)) one with the title *Is Informal Normal?*, which provides worldwide figures. This OECD study (the following results and figures are taken from the OECD (2009), executive summary) concludes that in many parts of the world and over the period 1990–2007 informal employment is the norm, not the exception. More than half of all jobs in the nonagricultural sectors of developing countries – over 900 million workers – can be considered informal. If agricultural workers in developing countries are included, the estimates size to roughly 2000 million people. In some regions, including sub-Saharan Africa and South Asia, over 80% of nonagricultural jobs are informal. Most informal workers in the developing world are self-employed and work independently, or owe and manage very small enterprises. According to the OECD study (2009), informal employment is a result of both, people being excluded from official jobs and people voluntarily opting out of formal structures, e.g., in many middle-income countries incentives drive individuals and businesses out of the formal sector.

To summarize, this OECD study clearly comes to the conclusion that informal is really the norm or the normal case. There are 1800 million people working in informal jobs, compared to 1200 million who benefit from formal contracts and social security protection. Informal economic activity, excluding the agricultural sector, accounts for three-fourth of the jobs in sub-Saharan Africa, more than two-thirds in South and Southeast Asia, half in Latin America, the Middle East, and North Africa, and nearly one-fourth in transition countries. If agriculture is included, the informal share of the economy in the above mentioned regions is even higher (e.g., more than 90% in South Asia). Also, this OECD study comes to the result that more than 700 million informal workers 'survive' on less than $ 1.25 a day and some 1200 million on less than $ 2 a day. The study also concludes that the share of informal employment tends to increase during economic turmoil. For example, during the Argentine economic crisis (1999–2002), the countries' 'official' economy shrank as by almost one-fifth while the share of informal employment expanded from 48 to 52%. One can clearly conclude that even under strong economic growth, the share of nonagricultural employment and informal employment is strongly rising.

OECD Countries

In Table 1 the estimates for the shadow economy labor force in highly developed OECD countries (Austria, Denmark, France, Germany, Italy, Spain, and Sweden) are shown. (Shadow economy labor force consists of estimated full-time 'black' jobs, including unregistered workers, illegal immigrants, and second 'black' jobs.) In Austria the shadow economy labor force has arrived at 500 000–750 000 or 16% of the official labor force (mean value) in the years 1997–98. In Denmark the development of the 1980s and 1990s shows that the part of the Danish population engaged in the shadow economy ranged from 8.3% of the total labor force (in 1980) to 15.4% in 1994 – quite a remarkable increase of the shadow economy labor force; it almost doubled over 15 years. In France (in the years 1997/98) the shadow economy labor force reached a size of between 6 and 12% of the official labor force or between 1.6 and 3.2 million in absolute figures. In Germany this figure rose from 8 to 12% in 1974 to 19 and 22% (8 millions) in 1997/98, respectively. For France and Germany this is again a very strong increase in the shadow economy labor force. In other countries the amount of the shadow economy labor force is quite large, too: in Italy 30–48% (1997–98), Spain 11.5–32% (1997–98), and Sweden 19.8% (1997–98). In the European Union about 30 million people are engaged in shadow economy activities in 1997–98 and in all European OECD countries 48 million people work illicitly.

These figures demonstrate that the shadow economy labor market is lively and may provide an explanation, why, for example, in Germany, one could observe such a high and persistent unemployment up to the year 2007.

Additionally, Table 1 contains a preliminary calculation of the total GDP per capita (including the official and the shadow economy GDP per capita) in US-$. In all countries investigated, total GDP per capita is much higher – on average in all countries around 40%. This clearly shows that the productivity in the shadow economy is roughly as high as in

Table 1 Estimates of the size of the 'shadow economy labor force' in some OECD countries 1974–98

Countries	*Year*	*Official GDP per capita in US-$*[a]	*Total economy (shadow economy plus official GDP per capita in US-$)*	*Size of the shadow economy (in % of official GDP) currency demand approach*[b]	*Shadow economy labor force in 1000 people*[c]	*Shadow economy participants in percentage of official labor force*[d]	*Sources of shadow economy and shadow economy labor force*
Austria	1990–91	20 636	25 382	5.47	300–380	9.6	Schneider (1998a,b)
	1997–98	25 874	29 630	8.93	500–750	16.0	and own calculations
Denmark	1980	13 233	18 658	8.6	250	8.3	Mogensen et al. (1995)
	1986	18 496	26 356	9.8	390	13.0	and own calculations
	1991	25 946	36 558	11.2	410	14.3	
	1994	34 441	48 562	17.6	420	15.4	
France	1975–82	12 539	17 542	6.9	800–1500	3.0–6.0	De Grazia (1983) and
	1997–98	24 363	34 379	14.9	1400–3200	6.0–12.0	own calculations
Germany	1974–82	11 940	17 911	10.6	3000–4000	8.0–12.0	De Grazia (1983),
	1997–98	26 080	39 634	14.7	7000–9000	19.0–23.0	Schneider (1998a,b) and own calculations
Italy	1979	8 040	11 736	16.7	4000–7000	20.0–35.0	Gaetani-d'Aragona
	1997–98	20 361	29 425	27.3	6600–11 400	30.0–48.0	(1979) and own calculations
Spain	1979–80	5 640	7 868	19.0	1250–3500	9.6–26.5	Ruesga (1984) and
	1997–98	13 791	19 927	23.1	1500–4200	11.5–32.3	own calculations
Sweden	1978	15 107	21 981	13.0	750	13.0–14.0	De Grazia (1983) and own
	1997–98	25 685	37 331	19.8	1 150	19.8	calculations
European	*1978*	*9 930*	*14 458*	*14.5*	*15 000*	–	*De Grazia (1983) and*
Union	*1997–98*	*22 179*	*32 226*	*19.6*	*30 000*		*own calculations*
OECD	*1978*	*9 576*	*14 162*	*15.0*	*26 000*	–	*De Grazia (1983) and own*
(Europe)	*1997–98*	*22 880*	*33 176*	*20.2*	*48 000*		*calculations*

[a]Source: OECD, Paris, various years.
[b]Source: Own calculations from Schneider and Williams (2013).
[c]Estimated full-time jobs, including unregistered workers, illegal immigrants, and second jobs.
[d]In percent of the population of 20–69 years of age, survey method.

the official economy – a clear indication, that the work effort (i.e., the incentive to work effectively) is as strong in the shadow economy as in the official one. In general these results demonstrate that the shadow economy labor force has reached a remarkable size in the developing countries as well as in highly developed OECD countries, even though the calculation still might have many errors.

Developing and Transition Countries

There have been some newer studies with respect to estimate the size and development of the shadow economy labor force (see also Schneider (2005, 2011), Schneider and Enste (2000), Feld and Schneider (2010), Schneider et al. (2010), Williams and Windebank (2001a,b), Williams (2010a,b, 2011a,b), and Hazans (2011).). Kucera and Roncolato (2008: p. 321) deal with informal employment. They address issues of crucial importance to labor market policy; first, the intensive labor market regulation is one major cause of informal employment, and second, the so-called voluntary informal employment. Kucera and Roncolato give a theoretical overview on both issues and also a survey of a number of empirical studies, in which the effect of the official labor market regulations on informal employment is analyzed, where they find a significant and quantitatively important influence.

In **Table 2** the share of informal employment in total nonagricultural employment by 5-year period and by region is presented. From the table one clearly sees that in all regions the share of informal employment has remarkably increased over time. The share of informal employment in South- and Middle-American countries in the period of 1985–89 was 32.4% and increased in the period of 2000–07 to 50.1%. In 34 Asian countries informal employment rose in the period of 1985–89 from 55.9 to 70.2% from 2000 to 2007. In the 42 African countries the share of informal employment (in percent of total nonagricultural employment) was 40.3% from 1985 to 1989, and increased to 60.5% in 2000–07. **Table 2** clearly demonstrates that there is a very strong positive trend in the share of informal employment (in percent of total nonagricultural employment).

Table 3 provides the share of informal employment in total nonagricultural employment by country, region, and gender. If one splits up the share of informal employment (in percent of total nonagricultural employment) by gender, I generally observe, that the share of women is significantly higher than the share of men. In North Africa (countries Algeria, Morocco, Tunisia, Egypt) the share of informal employment of women is 43.3% and the one of men 49.3% over the period 1990–99. In sub-Saharan Africa the share of women is 84.1%, the one of men 63.0%. In Latin America the share of women is 56.2% and the share of men 47.1%. Only in the region of West Asia and in the transition countries the share of men of informal employment is higher than the one of women. In West Asia (countries Lebanon, West Bank and Gaza Strip, Syria, Turkey, Yemen) the share of women is 31.1%, the share of men 43.4%. In the Transition countries (Kyrgyzstan, Moldova,

Table 2 Share of informal employment in total nonagricultural employment by 5-year period in percentage

	Average share of informal employment in % of local nonagricultural employment over			
Region	*1985–89*	*1990–94*	*1995–99*	*2000–07*
22 South- and Middle-American countries	32.4	35.4	40.3	50.1
34 Asian countries	55.9	60.4	65.4	70.2
42 African countries	40.3	47.1	52.4	60.5
21 Transition countries	30.9	32.3	35.4	40.2

Source: OECD, 2009. Is Informal Normal? Towards More and Better Jobs, OECD, Paris, pp. 34–35; Charmes, J., 2002. Informal Sector, Poverty and Gender: A Review of the Empirical Evidence. Paper prepared on behalf of Wiego (Women in Informal Employment: Globalizing and Organizing). OECD and World Bank, Washington and Paris; Charmes, J., 2007. Informal Sector Poverty and Gender, a Review of Empirical Evidence. Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing). OECD and World Bank, Washington and Paris; Charmes, J., 2008. Informal Sector Poverty and Gender, a Review of Empirical Evidence. Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing). OECD and World Bank, Washington and Paris for the ILO Women and Men in the Informal Economy, 2002. For the most recent period: Heintz, E., Chang, G.B., 2007. Report of Informal Employment for the ILO, ILO Geneva for the ILO, and for West Asia; Charmes, J., 2007. Informal Sector Poverty and Gender, a Review of Empirical Evidence, Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing), OECD and World Bank, Washington and Paris; Charmes, J., 2008. Informal Sector Poverty and Gender, a Review of Empirical Evidence. Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing). OECD and World Bank, Washington and Paris.

Russia) the share of women is 22.3% and the share of men 27.2%. I also realize some remarkable differences. In general the share of informal employment is rather large worldwide and certainly has severe policy implications.

Shadow Economy and Unemployment

(This part is taken from Feld and Schneider (2010).) Although there has been some discussion on the size of the shadow economy labor force and on its reasons, comparatively little attention has been given to the relationship between unemployment and working in the shadow economy. As Tanzi (1999) points out, "the current literature does not cast much light on these relationships even though the existence of large underground activities would imply that one should look more deeply at what is happening in the labor market" (p. 347). The objective of the article by Bajada and Schneider (2009) is to examine the extent of participation in the shadow economy by the unemployed. Their paper has investigated the relationship between the unemployment rate and the shadow economy. Previous literature on this topic has suggested that the relationship between these two variables is ambiguous, predominantly because a heterogeneous group of people working in the shadow economy exists and there are also various cyclical forces at work, such that they produce a net effect that is weakly correlated with unemployment. In their paper they have provided a suggestion for disentangling these cyclical effects, so as to study the component of the shadow economy that is influenced directly by those who are unemployed. They refer to this effect as the 'substitution effect,' which typically increases during declining periods of legitimate economic activity (and increasing unemployment). Equipped with this approach for measuring the 'substitution effect,' they discover that a relationship exists between changes in the unemployment rate and shadow economy activity.

By examining the growth cycle characteristics of the 'substitution effect' component of the shadow economy Bajada and Schneider (2009) determine that the growth cycles are

Table 3 Share of informal employment in total nonagricultural employment, by country, region, and gender (in percent), 1990s and 2000s

	1990–99		*2000–07*	
Region	*Women*	*Men*	*Women*	*Men*
North Africa	43.3	49.3		
Algeria	40.6	43.1		
Morocco	46.8	44.0		
Tunisia	39.2	53.2		
Egypt	46.5	56.9	38.6	47.2
Sub-Saharan Africa	84.1	63.0	77.1	62.6
Benin	97.3	87.0		
Chad	95.2	59.5		
Guinea	86.7	65.6		
Kenya	83.1	59.1		
Mali			89.2	74.2
South Africa	58.4	43.6	64.9	51.0
Latin America	56.2	47.1	59.5	55.4
Bolivia	74.4	55.0		
Brazil	67.3	54.7	52.3	50.2
Chile	43.9	30.9		
Colombia	44.0	34.1		
Costa Rica	48.0	42.1		
Dominican Republic	49.7	46.5		
Ecuador			76.9	73.2
El Salvador	68.6	45.7		
Guatemala	69.4	46.5		
Honduras	65.5	73.6		
Mexico	55.0	54.3	53.5	47.8
Panama	40.8	35.5	50.4	48.7
Peru			72.0	65.1
Venezuela	47.3	46.7	52.1	47.5
South and Southeast Asia	72.7	70.2	73.4	71.2
India	85.7	82.9		
Indonesia	77.2	78.0		
Philippines	73.4	70.8		
Thailand	54.3	49.1		
West Asia	31.1	43.4	35.4	44.4
Lebanon			60.0	44.4
West Bank and Gaza Strip			20.2	46.8
Syria	34.6	42.8		
Turkey	19.1	29.1	32.2	33.4
Yemen	39.7	58.2	29.3	52.8
Transition countries			22.3	27.2
Kyrgyzstan			40.9	47.1
Moldova			18.4	28.0
Russia			7.6	9.6

Source: OECD, 2009. Is Informal Normal? Towards More and Better Jobs. OECD, Paris, p. 47; Charmes, J., 2002. Informal Sector, Poverty and Gender: A Review of the Empirical Evidence. Paper prepared on behalf of Wiego (Women in Informal Employment: Globalizing and Organizing). OECD and World Bank, Washington and Paris, for the ILO Women and Men in the Informal Economy, 2002. For the most recent period: Heintz, E., Chang, G.B., 2007. Report of Informal Employment for the ILO. ILO Geneva for the ILO, and for West Asia.

symmetric (in terms of steepness and deepness) and that changes in the unemployment rate, whether positive or negative, had similar impacts on changes in the substitution effect component. They suggest that the shadow economy is a source of financial support during periods of unemployment for those genuinely wanting to participate in the legitimate economy. Although this does not exclude the possibility that long-term unemployed may also be participating in the shadow economy, it would appear that short-term fluctuations in unemployment directly contribute to short-term fluctuations in the shadow economy.

When Bajada and Schneider consider the various unemployment support programs across 12 OECD countries, there appears to be no real systematic relationship between the generosity of the social security systems and the nature of short-term shadow economic activity by the unemployed. Even the various replacement rates across the OECD countries appear to have little consequence on the rate at which the unemployed take on and cut back shadow economy activity. There is however some evidence to suggest that extended duration spell in unemployment lasts anywhere between less than 3 months to approximately 9 months.

On the whole Bajada and Schneider argue that dealing with unemployment participation in the shadow economy as a way of correcting the inequity it generates, is best handled by more stringent monitoring of those receiving unemployment benefits rather than reducing replacement rates as a way of encouraging reintegration into the work force. A strategy of reducing replacement rates would not only fail to maintain adequate support for those experiencing financial hardship during periods of unemployment, it is likely to have little impact on reducing participation by the unemployed who are willing and able to engage in shadow economy activity.

Conclusions

In this article some of the most recent developments in research on the shadow economy labor force and undeclared work in highly developed OECD, developing, and transition countries are shown.

The discussion of the recent literature shows that economic opportunities for employees, the overall situation on the labor market, not least unemployment are crucial for an understanding of the dynamics of the shadow economy. Individuals look for ways to improve their economic situation and thus contribute productively to aggregate income of a country. This holds regardless of their being active in the official or the unofficial economy.

If I come back to the headline of this article "Work in the shadow: Some Facts" I clearly realize that we have some knowledge about the size and development of the shadow economy labor force. For developing and transition countries, the shadow economy labor force has reached a remarkable size according to OECD (2009) estimates, which is, that in most developing countries the shadow economy labor force is higher than the official labor force. What we do not know are the exact motives, why people work in the shadow economy and what is their relation and feeling if a government undertakes reforms in order to bring them back into the official economy. Hence, much more microstudies are needed to obtain a more detailed knowledge about people's motivation to work either the shadow economy and/or in the official one.

See also: Informal Sector; Informal Work in Developing Countries; Informal and Underground Economics.

Bibliography

Bajada, C., Schneider, F., 2009. Unemployment and the shadow economy in the OECD. Review Economique 60 (4), 1011–1033.

Cappiello, M.A., 1986. Proposita di bibliografia ragionata sull'economia sommersa nell'industria (Italia 1970–82). In: Bagnasco, A. (Ed.), L'altra metà dell'economia, La ricerca internazionale sull'economia Informale. Liguori, Naples, pp. 307–345.

Charmes, J., 2002. Informal Sector, Poverty and Gender: A Review of the Empirical Evidence. Paper prepared on behalf of Wiego (Women in Informal Employment: Globalizing and Organizing). OECD and World Bank, Washington and Paris.

Charmes, J., 2007. Informal Sector Poverty and Gender: A Review of Empirical Evidence. Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing). OECD and World Bank, Washington and Paris.

Charmes, J., 2008. Informal Sector Poverty and Gender: A Review of Empirical Evidence. Paper prepared on behalf of Wiego (Women in Formal Employment Globalizing and Organizing). OECD and World Bank, Washington and Paris.

Chickering, L.A., Salahdine, M. (Eds.), 1991. The Silent Revolution – the Informal Sector in Five Asian and Near Eastern Countries. An International Center for Economic Growth Publication (ICS Press), San Francisco.

Dallago, B., 1990. The Irregular Economy: The "Underground Economy" and the "Black Labor Market". Publishing Company, Dartmouth, UK.

De Grazia, R., 1983. Le travail clandistin: Situation dans les pays industrialisés à économie de marché. Bit, Genève.

Feld, L.P., Schneider, F., 2010. Survey on the shadow economy and undeclared earnings in OECD countries. German Economic Review 11 (2), 109–149.

Frey, L., 1972. Il Lavoro a Domicilio in Lombardia, Giunta Regionale Lombarda. Assessorato al Lavoro, Milan.

Frey, L., 1975. Il potenziale di lavoro in Italia. In: Documenti ISVET. No. 50.

Frey, L., 1978. Il lavoro nero nel 1977 in Italia. In: Tendenze Della Occupazione. No. 6.

Frey, L., 1980. Introduzione all'analisi economica del lavoro minorile. Economia del Lavoro 1–2, 5–16.

Gaetani-d'Aragona, G., 1979. I sommersi. Nord e Sued 7 (1), 26–46.

Haigner, S., Jenewein, S., Schneider, F., Wakolbinger, F., 2013. Driving forces of informal labor supply and demand in Germany. International Labor Review 152 (3 -4), 507–524.

Hazans, M., 2011. Informal Workers across Europe, Evidence from 30 Countries. Discussion Paper and Background Paper, Worldbank, P112988. University of Latvia.

Heintz, E., Chang, G.B., 2007. Report of Informal Employment for the ILO. ILO, Geneva.

Jöreskog, K.G., Goldberger, A.S., 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association 70 (351), 631–639.

Kucera, D., Roncolato, L., 2008. Informal employment: two contested policy issues. International Labor Review 147 (3), 321–348.

Lemieux, T., Fortin, B., Fréchette, P., 1994. The effect of taxes on labor supply in the underground economy. American Economic Review 84 (3), 231–254.

Loayza, N.V., 1996. The economics of the informal sector: a simple model and some empirical evidence from Latin America. Carnegie-Rochester Conference Series on Public Policy 45, 129–162.

Lubell, H., 1991. The Informal Sector in the 1980's and 1990's. OECD, Paris.

Mogensen, G.V., Kvist, H.K., Körmendi, E., Pedersen, S., 1995. The Shadow Economy in Demark 1994: Measurement and Results. Study No. 3. The Rockwool Foundation Research Unit, Copenhagen.

OECD, 2008. Declaring Work or Staying in Underground: Informal Employment in 7 OECD Countries. OECD Employment Outlook. OECD, Paris, (Chapter 2).

OECD, 2009. Is Informal Normal? Towards More and Better Jobs. OECD, Paris.

Pozo, S. (Ed.), 1996. Exploring the Underground Economy: Studies of Illegal and Unreported Activity. W.E. Upjohn, Institute for Employment Research, Michigan.

Ruesga, B.S.M., 1984. Economia oculta y mercado de trabajo: Aproximación al caso espaniol. Información Commercial Espaniola, 55–61.

Schneider, F., 1998a. Further Empirical Results of the Size of the Shadow Economy of 17 OECD-Countries over Time. Paper to be presented at the 54. Congress of the IIPF Cordoba, Argentina and Discussion Paper. Department of Economics, University of Linz, Linz, Austria.

Schneider, F., 1998b. Stellt das Anwachsen der Schwarzarbeit eine wirtschaftspolitische Herausforderung dar? Einige Gedanken aus volkswirtschaftlicher Sicht. Mitteilungen des Instituts für angewandte Wirtschaftsforschung (IAW) I (98), 4–13.

Schneider, F., 2005. Shadow economies around the world: what do we really know? European Journal of Political Economy 21 (4), 598–642.

Schneider, F. (Ed.), 2011. Handbook on the Shadow Economy. Edward Elgar Publishing Company, Cheltenham, UK.

Schneider, F., Enste, D., 2000. Shadow economies: size, causes and consequences. Journal of Economic Literature 38 (1), 73–110.

Schneider, F., Williams, C.C., 2013. The Shadow Economy. IEA, London.

Schneider, F., Buehn, A., Montenegro, C.E., 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461.

Tanzi, V., 1999. Uses and abuses of estimates of the underground economy. Economic Journal 109 (3), 338–347.

Williams, C.C., 2010a. Evaluating the nature of undeclared work in South-Eastern Europe. Employee Relations 32 (3), 212–226.

Williams, C.C., 2010b. Special variations in the hidden enterprise culture: some lessons from England. Entrepreneurship and Regional Development 22 (5), 403–423.

Williams, C.C., 2011a. A critical evaluation of competing representations of informal employment: some lessons from England. Review of Social Economy 69 (2), 211–237.

Williams, C.C., 2011b. Reconceptualising Men's and Women's undeclared work: evidence from Europe. Gender Work and Organization 18 (4), 415–437.

Williams, C.C., Windebank, J., 2001a. Beyond profit motivated exchange: some lessons from the study of paid informal work. European Urban and Regional Studies 8, 49–61.

Williams, C.C., Windebank, J., 2001b. Reconceptualizing paid informal exchange: some lessons from English cities. Environment and Planning A 33, 121–140.

Shamanism

Donald Joralemon, Smith College, Northampton, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14032–14035, © 2001, Elsevier Ltd.

Abstract

Shamanism is a tradition of part-time religious specialists who establish and maintain personalistic relations with specific spirit beings through the use of controlled and culturally scripted altered states of consciousness (ASC). Shamans employ powers derived from spirits to heal sickness, to guide the dead to their final destinations, to influence animals, and forces of nature in a way that benefits their communities, to initiate assaults on enemies, and to protect their own communities from external aggression. The following paragraphs cover common features of shamanic traditions, techniques for inducing ASC, and popular versus academic uses of the term (e.g., in New Age religions, in symbolic anthropology). The article ends with a quick review of 'dead ends and detours' in shamanism studies, including speculations about shamans' mental status and the place of shamanism in cultural evolution. It calls for a refocusing of the term to a more restricted definition.

Shamanism is a tradition of part-time religious specialists who establish and maintain personalistic relations with specific spirit beings through the use of controlled and culturally scripted altered states of consciousness (ASC). Shamans employ powers derived from spirits to heal sickness, to guide the dead to their final destinations, to influence animals and forces of nature in a way that benefits their communities, to initiate assaults on enemies, and to protect their own communities from external aggression.

Shamans exercise mastery over ASC and use them as a means to the culturally approved end of mediating between human, animal, and supernatural worlds. Shamans draw pon background knowledge, conveyed through myth and ritual, which renders intelligible the potentially chaotic experience of ASC. The criterion of control helps to distinguish shamanism from the use of ASC in other traditions.

Shamanism has long been a subject of inquiry and controversy in diverse academic disciplines, with many hundreds of accounts of shamanic practices published by the early 1900s. It has also been a topic of spiritual interest to the wider public in Europe and North America for the past several decades. Depending on the perspective taken, shamanism is either the most archaic and universal form of human spirituality or a culturally distinct religious complex with historical roots in Siberia and a path of diffusion into North and South America. The latter view is adopted in the following.

Siberian Origin and Common Features

The term 'shaman' is drawn from seventeenth century Russian sources reporting on an eastern Siberian people, the Evenk (Tungus). In the classic work on the subject by the historian of religion Mircea Eliade, 'shamanism' is used to refer to a complex of beliefs and practices diffused from northern Asia to societies in central and eastern Asia, as well as through all of North and South America. Eliade's comparative work specifies a core constellation of ideas common to shamanic traditions. The classic shamanic initiation involves the novice's selection – frequently unwanted and resisted – by a spirit(s), a traumatic and dangerous series of ordeals, followed by a death and rebirth that sometimes involves violent dismembering and subsequent reconstitution of the fledgling shaman's body. The archetypal shamanic cosmology is vertically tiered, with earth occupying the middle level and a cosmic tree or world mountain serving as a connecting path for shamans to travel to other cosmic planes (up or down) in pursuit of their 'helping spirits.'

Shamans use a variety of techniques to enter the ASC in which they communicate with their spirit helpers: sensory deprivation (e.g., fasting, meditation), repetitive drumming and/or dancing, and ingestion of substances with psychoactive properties (e.g., plant hallucinogens). The last of these means is particularly important in the shamanism of Central and South America, where an impressive array of plants and animal-derived chemicals have been used for their hallucinogenic properties (e.g., peyote, datura, virola, and poison from the *Bufo marinus* toad).

Among the most commonly reported characteristics of the shaman's ASC are transformations into animals, and flights to distant places. Shamanic animals include the jaguar (lowland South America), the wolf (North American Chukchee), and bears, reindeer and fish (Lapp). Shamanic flights are undertaken to recapture the wandering souls of the sick, to intercede on behalf of hunters with spirits that control animal species, and to guide the dead to their final destination.

The fact that shamans are often called upon to heal the sick should not lead to the conclusion that they are always concerned with the well being of their fellows. While the Hippocratic Oath of Western medicine prohibits physicians from doing harm, shamans frequently engage in actions meant to sicken, if not kill, their adversaries. This is an important corrective to romanticized images of the shamanic vocation.

Popular and Professional 'Shamanism'

Beginning in the 1960s there developed a convergence in the professional and popular interests in shamanism. Inspired by the experimentation with hallucinogenic drugs on the part of American and European youth, attention was drawn to the physiology and psychology of altered states of consciousness.

In academic circles, this led to speculations about the neurochemical foundations of hallucinatory experiences and their potential therapeutic benefits. Investigations were undertaken into the imagery of shamanic states of consciousness to determine how much could be attributed to universal, physiologically related sensory experiences. One often cited collection of essays, published in the journal *Ethos* (Prince, 1982), sought to link the shaman's ASC to the production of the naturally occurring opiates known as endorphins. Scholars argued that this might account for the healing effect of shamanic therapies.

In the popular cultures of North America and Europe, stress came to be placed on shamanism as an avenue to self-exploration, a means by which persons from any culture could advance their quest for spiritual understandings. The legacy of this development is to be found in the forms of 'neoshamanism' among New Age religions, which draw eclectically from more traditional shamanic practices. Native North and South American societies have been special sources of inspiration for practitioners of neoshamanism, but not always with the willing support of the indigenous people.

At the same time, and in response to the evolution of research in symbolic anthropology, anthropologists have contributed increasingly sophisticated ethnographic analyzes of the metaphysical underpinnings of shamanism in specific societies. These anthropologists concentrate their attention on the complexities of shamanic cosmologies and the relationships between shamanic symbolism and crucial features of the natural world (e.g., important animal and plant species, meteorological phenomena, and geographic landmarks).

An especially detailed analysis of a shamanic tradition comes from the Warao of the Orinoco Delta in Venezuela. The German-born and American-trained anthropologist Johannes Wilbert documents the mythic cosmology that supports the work of three distinct kinds of shamans, each of which are responsible for the care and feeding of specific gods. The highest ranking shaman, the Wishiratu ('Master of Pains' or Priest Shaman), responds when the Kanobos (Supreme Beings) send sickness to a Warao village. The Bahanarotu is also a curer, but has a special responsibility for a fertility cult centered on an astonishingly complex supernatural location, the House of Tobacco Smoke, where four insect couples and a blind snake play a game of chance under the watchful eye of a swallow-tailed kite. The Hoarotu (Dark Shaman) has the unpleasant but essential task of feeding human flesh to the Scarlet Macaw, the God of the West. The preferred food of the other gods is tobacco smoke, which is also the vehicle by which Warao shamans enter ASC – by hyperventilating on long cigars made of a strong black leaf tobacco.

Research Directions

Dead Ends and Detours

There have been a number of dead ends and detours in shamanism studies. An unfortunate amount of energy was spent in fruitless discussions of the mental health of shamans – whether or not their abnormal behavior justified the application of psychiatric diagnoses (e.g., schizophrenia). Most scholars now recognize that the culturally patterned nature of shamans' behavior and the positive value placed on their social role make the application of mental illness labels inappropriate.

Another distraction came when cultural evolutionists hypothesized that shamanism was an intermediary stage in the developmental sequence of religious forms, in-between magic and institutionalized religion. Attempts to treat shamanism exclusively as an archaic expression of human religiosity flounder on the perseverance and even revitalization of shamanic traditions in cultures around the world. For example, the remarkable resiliency of shamanic practices is evident in the continuity between contemporary *curanderismo* (curing) on Peru's north coast and shamanic iconography dating to the pre-Hispanic Chavin culture (*c.*900–200 BC).

Positive Directions

A still limited amount of research has been directed toward the important question of how effective are shamanic treatments. This work has been troubled by serious theoretical issues (e.g., what constitutes 'efficacy') and thorny methodological questions (e.g., is the double blind, randomized trial possible under the unusual circumstances of shamanic rituals?). The previously cited endorphin theory has not been confirmed ethnographically. Another approach has adapted psychiatric questionnaires to before and after treatment interviews with shamanic patients.

A fruitful line of investigation has focused on the relationship between shamanism and culturally constructed notions ofgender. Shamans sometimes 'bend' gender roles (e.g., transvestitism) and the sicknesses they treat can be entangled in gender-based conflicts. Some scholars suggest that where women dominate in the shaman role there is a less adversarial model of supernatural mediation than with male shamans. The best examples of gender-focused research come from East Asia (especially Korea), where the majority of shamans are women, and from the Chilean Mapuche, among whom female and cross-dressing male shamans have predominated since at least the sixteenth century.

A final body of research worth noting focuses on the capacity of shamanic traditions to survive and thrive even under the most disruptive social and political conditions. Agents of culture change, whether they are Soviet-era ideologues in Siberia or Christian missionaries in the Amazon, have been repeatedly frustrated in their attempts to obliterate shamanic practices. In recent decades, shamans have even become central figures in the politics of ethnicity and in anti-development protests. Scholars have shown that shamans accomplish this survival feat not by replicating ancient traditions, but by continuously reinventing them in the light of new realities and competing symbolic structures. While this strategy may occasion debates about what constitutes an 'authentic' shaman, it is nevertheless the key to shamanism's continuing success.

A Plea for Terminological Precision

The most serious threat to the academic study of shamanism lies in the broad application of the term to any religiously inspired trance form. To so dilute the concept as to make it

applicable to Kung trance dancers, New Age spiritualists, and Mexican Huichol peyote pilgrims is to render it meaningless. The value of a scientific term is that it groups together phenomena that are alike in significant regards, while distinguishing those that are different. The promiscuous use of the term shaman will ultimately leave it as generic as 'spiritual healer,' and just as devoid of analytic value.

Bibliography

Atkinson, J.M., 1992. Shamanisms today. Annual Review of Anthropology 21, 307–330.

Bacigalupo, A. M. Shamans of the Cinnamon Tree, Priestesses of the Mood: Gender and healing among the Chilean Mapuche. Manuscript in preparation.

Brown, M.F., 1989. Dark side of the shaman. Natural History 11, 8–10.

Eliade, M., 1972. Shamanism: Anchaic Techniques of Ecstasy. Princeton University Press, Princeton, NJ.

Furst, P.T. (Ed.), 1972. Flesh of the Gods: The Ritual Use of Hallucinogens. Praeger, New York.

Hultkrantz, A., 1992. Shamanic Healing and Ritual Drama. Crossroad, New York.

Joralemon, D., 1990. The selling of the shaman and the problem of informant legitimacy. Journal of Anthropological Research 46, 105–118.

Joralemon, D., Sharon, D., 1993. Sorcery and Shamanism: Curanderos and Clients in Northern Peru. University of Utah, Salt Lake City, UT.

Kalweit, H., 1988. Dreamtime & Inner Space: The World of the Shaman, first ed. Shambhala, Boston, MA.

Kendall, L., 1985. Shamans, Housewives, and Other Restless Spirits: Women in Korean Ritual Life. University of Hawaii Press, Honolulu, HI.

Kleinman, A., Sung, L.H., 1979. Why do indigenous practitioners successfully heal? Social Science & Medicine 13 (1B), 7–26.

Peters, L.G., Price-Williams, D., 1980. Towards an experiential analysis of shamanism. American Ethnologist 7, 398–418.

Prince, R., 1982. Shamans and endorphins: Hypotheses for a synthesis. Ethos 10 (4), 409–423.

Reichel-Dolmatoff, G., 1975. The Shaman and the Jaguar. Temple University Press, Philadelphia, PA.

Shirokogoroff, S.M., 1935. The Psychomental Complex of the Tungus. Kegan, Paul, Tranch and Trubner, London.

Silverman, J., 1967. Shamanism and acute schizophrenia. American Anthropologist 69, 21–31.

Sullivan, L.E., 1988. Icanchu's Drum: An Orientation to Meaning in South American Religions. Macmillan, New York.

Wilbert, J., 1993. Mystic Endowment: Religious Ethnography of the Warao Indians. Harvard University Press, Cambridge, MA.

Shame and the Social Bond

Thomas Scheff, UCSB, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article proposes that shame may be the hidden key to understanding our civilization, in the sense that shame or its anticipation is virtually ubiquitous, yet, at the same time, usually invisible. C.H. Cooley's idea of the looking glass self implies that shame and pride can be seen as signals of the state of the social bond. Theoretical work by Cooley and Erving Goffman imply ubiquity, and empirical studies by Norbert Elias and Helen Lewis provide support. The findings of Elias and Lewis also suggest that shame is usually invisible; Elias stated this proposition explicitly. Like other emotions, such as fear, shame can be recursive, acting back on itself (shame about shame). In unusual circumstances, limitless recursion of shame may explain extreme cases of silence or violence.

The psychologist Gershen Kaufman is one of several writers who have argued that shame is taboo in our society:

> American society is a shame-based culture, but shame remains hidden. Since there is shame about shame, it remains under taboo. …The taboo on shame is so strict that we behave as if shame does not exist (Kaufman, 1989).

Kaufman's phrase, shame about shame, turns out to have meaning beyond what he intended: just as fear can lead to more fear, causing panic, shame about shame can loop back on itself to various degrees, even to the point of having no natural limit. Recursion of shame will be discussed further below (for a discussion of the recursion of thought, see Corballis, 2007). Suppose that shame is usually hidden, as suggested by the idea of taboo. What difference would it make? Before taking up this issue, it will first be necessary to understand what is meant by the term shame.

Definitions of Shame

A linguistic way of hiding shame would be to misname all but the most intense or obvious occurrences. The word shame is defined very narrowly in English: an intense crisis response to inadequacy or misbehavior. In English also, unlike most other languages, shame is kept distinct from less intense siblings, notably, from embarrassment. Other languages treat shame as a family of feelings that extends into everyday life. In Spanish, for example, the same word, *verguenza*, is used to mean both shame and embarrassment, and in French, the term *pudeur*, which is translated into English as modesty, is considered a part of the shame family.

There is a social definition of shame in maverick psychoanalysis and in sociology that defines shame broadly, in a way that includes both embarrassment and guilt, and many other shame variants. Erikson (1950) rejected Freud's assumption that guilt was the primary moral emotion for adults. He argued instead that shame was more elemental in that it concerned the whole self, not just one's actions.

This idea was expanded by the sociologist Helen Lynd, who outlined the crucial importance of shame in the constitution of the self and in social life. She was the first to recognize the need for a *concept* of shame that would be clearly defined in order to avoid the misconceptions of vernacular usage.

The next step was taken by the psychologist Silvan Tomkins, who proposed that shame plays a central role in behavior. In his volume on the 'negative affects' (1963, vol. II), he devoted almost *500 pages* to a detailed discussion of shame and humiliation. This treatment dwarfs his discussion of the other emotions. His examples of shame imply a broad conception. Indeed, he argued explicitly that embarrassment, shame, and guilt should be recognized as members of a single affect family, as I do here. How does this family enter into the daily life of modern societies?

The Looking Glass Self

In his first and most general book, Erving Goffman made a surprising claim:

> There is no interaction in which participants do not take an appreciable chance of being slightly embarrassed or a slight chance of being deeply humiliated (1959, p. 243).

This statement occurs only in passing toward the end of the book. Like most of his generalizations, there is little further development, not directly at least. This one asserts that *all* interaction carries with it the risk of exposure to a painful emotion. One of Goffman's main ideas, *impression management*, has a similar implication. The reason we spend such time and care managing our impressions, Goffman argued, is to avoid embarrassment as best we can. Cooley had laid the groundwork for the idea that human life is haunted, if not controlled, by shame, although Goffman does not cite him in this regard.

In two brief statements, Cooley implied that both inner and outer human life produces emotions and that both social and self process *always* leads to either pride or shame.

A. "We live in the minds of others without knowing it." (p. 208).
B. "[The self] seems to have three principal elements:
   1. The imagination of our appearance to the other person
   2. The imagination of his [or her] judgment of that appearance
   3. Some sort of self-feeling, such as pride or [shame]." (p. 184).

Sentence A describes the basis for all social relationships, a theory of mind. Intersubjectivity is built into human nature, yet modern societies make it almost invisible. Small children learn to go back and forth between own point of view and imagining that of the other(s). By the age of five or six, they have become so adept and lightening quick that they forget they are doing it. In Cooley's words, it's like "the ground that bears us up when we walk," taken for granted. The intersubjective bond to others becomes invisible, at least in Western societies.

Human communication is built upon intersubjective guessing because actual speech is complex, fragmented, contextualized, and therefore, if understood literally, usually misunderstood or uninterruptable. To try to understand, the listener must move back and forth between own point of view and the imagined point of view of the speaker. G.H. Mead (1934) called it 'role-taking,' although it is not about roles, but points of view.

Occasionally, role-taking will be referred to in ordinary conversation, but only casually and in passing. For example, one might say to a friend, "We both know that (such and such…)." Even here, in the very act of imaginative mind reading, we usually are not conscious that we are doing it.

Paragraph B proposes a *looking glass self*, the way in which social relationships give rise to a self, which, in turn, leads to pride or shame. Cooley's approach implies that all social interaction produces either pride or shame and that these emotions are social as well as psychological. They signal the state of the relationship: either *connectedness* (pride) or *disconnect* (shame). It further implies that the degree of solidarity or alienation (connect/disconnect) determines the prevailing emotional states in a society.

The paragraph also suggests how, in three steps, either pride or shame might be present all of the time, whether one is alone or with others. The two passages together imply a basis for what will be called the Elias–Lewis conjecture: *shame or its anticipation is ubiquitous in modern societies, yet it also has become largely invisible*. But what about pride?

Cooley's and Goffman's treatment of the looking glass self are quite parallel. Cooley's few concrete examples all concern shame rather than pride. Goffman provided hundreds of examples of impression management, but none ending in genuine pride. There are 16 mentions of either shame or embarrassment in the text of *Presentation of Self in Everyday Life* (1959), but only three of pride. Moreover, all three of the pride mentions were in long quotes by other authors, in which pride was largely incidental. Goffman himself uses only shame terms, not pride. Why did both authors omit it?

Though neither Cooley nor Goffman name the kind of civilization they analyzed, it is clearly the current one, a modern, rather than a traditional society. Perhaps modernity gives rise to their single focus on shame. Shame is a signal of disconnect, alienation. Relationships in modern societies strongly tend toward alienation, and therefore to the ubiquity of shame.

Modernity is built on a base of individualism, the encouragement to go it alone, no matter the cost to relationships. Persons learn to act as if they were complete in themselves and independent of others. This feature has constructive and creative sides, but it has at least two other ones: alienation and the hiding of shame.

Emphasis on individual rationality is a key institution in modern societies. Another is the suppression of the social–emotional world in favor of thought and behavior. One of the many outcomes of this suppression is that emotion vocabularies in modern languages are ambiguous and misleading, so that they tend to hide alienation. For example, in the English language, love is defined so broadly that is often used to hide disconnection (*Women Who Love Too Much*). There are also many other ambiguities, confusions, and deceptions. Since shame is elaborately hidden and disguised, a close examination of the verbal, gestural, and contextual details may be needed to uncover it.

In traditional and Asian societies, the central importance of shame is taken for granted. Indeed, in some Asian societies, such as Japan, it is seen as the central emotion. In a traditional society like the Maori, shame (they call it *whakamaa*) is also treated as the key emotion. Indeed, the whole approach to shame and relationships in this article would be seen as platitudes by the Maori, news from nowhere. But in Western societies, treating shame as highly significant in everyday life is counterintuitive and even offensive.

Western societies focus on individuals, rather than on relationships. Emerson, because of his emphasis on self-reliance as an antidote to blind conformity, was one of the prophets of individualism: "When my genius calls, I have no father and mother, no brothers or sisters." In extreme contrast, in a traditional society, there is *nothing* more important than one's relationships. Freeing up the individual from their relational/emotional world has been at the core of modernization. Since one's relationships and emotions do not show up on a resume, they have been de-emphasized to the point of disappearance. But shame and relationships do not disappear; they just assume hidden and disguised forms.

Individualism is the dominant theme of all relationships in Western societies. This focus disguises the web of personal and social relationships that sustain all human beings. The myth of the self-sustaining individual, in turn, reflects and generates the suppression and hiding of shame and pride. Since pride and shame, or at least their anticipation, are the predominant emotions in social interaction, suppression supports the status quo, the myth of the self-contained individual. But the obverse is that as we become aware of the massive amounts of emotions and disguising of emotion that occur in social interaction, *we can make visible what is otherwise invisible, the state of any given relationship or set of relationships*.

Pride, Shame, and Alienation

The confusion of English vernacular is obvious in the case of pride, since dictionaries and usage both imply two contradictory meanings. The first meaning is negative: pride is interpreted as egotism ('Pride goeth before the fall'). When we say that someone is proud, it is likely to be condemnatory. *False pride* might be a better name for this kind of self-feeling to distinguish it from true pride.

The second meaning is positive: a favorable view of self, but one that has been earned. This kind of pride is genuine, authentic, justified. However, even adding these adjectives does not completely eliminate the negative flavor. In English, the

word pride is often tainted by its first meaning, no matter how impressive the justification.

Individualism also causes endless obfuscation about shame. The primary confusion is the practice of leaving out the social component that arises from the looking glass self: viewing ourselves negatively because we imagine that we are viewed that way by another person or persons.

Both in vernacular and in scholarly usage, shame typically is assumed to be only an internal matter, condemning oneself. But the looking glass self contains both the internal result and the external source. The typical definition of shame in psychological studies involves gross dissatisfaction with self. Cooley's usage includes this part, but also the social component, imagining, correctly or incorrectly, a negative view of self by others.

Cooley's idea of the social source of shame and pride suggests that these emotions are signals of the *state of a relationship*. As indicated earlier, whatever the substantive basis for shame, the actual violation or occasion, a more general component is the state of the bond: *true pride signals a secure bond (connectedness), shame a threatened one (disconnect)*. This definition virtually always includes the substantive cause of shame, whatever it might be, since the causes of shame themselves are usually shared with one's whole society.

Since modern societies produce alienation at many different levels, emotions and relationships are deeply hidden. Shame, in particular, becomes invisible, even for most social and behavioral researchers. A taboo is implied in the many studies of shame that do not use the forbidden word at all. Instead, the focus is on one of the many shame cognates (Retzinger, 1995, lists hundreds). One such cognate is the word awkward, as in 'it was an awkward moment for me.' A further way of hiding shame is to behaviorize it: there are many studies of feelings of *rejection*, *loss* of social status, and the *search* for recognition.

I have just begun to study the occurrence of shame and other emotion terms in millions of digitalized books from 1800 to 2000. These data are called Ngrams by its makers, Aiden (2011). My initial finding is that the use of the word shame has decreased threefold during the 200 years covered by the Ngrams. This issue is somewhat complex, since phrases that use the word do not refer to actual shame: 'What a shame!' means exactly the same thing as 'What a pity.' I will describe all of the findings in a later report.

Lewis's study of shame in psychotherapy sessions, to be discussed later, has received many citations, yet they usually ignore or misinterpret her main findings. She complained to me once that people praise her book but do not read it. Similarly, groups headed by Paul Gilbert and George Brown have published several studies of shame, but with little response.

On the other hand, Evelin Lindner has been able to organize a worldwide following for the study of themes identical to those discussed here. Her success may be due, at least in part, to avoiding the s-word, especially in titles, not only for her organization (*Human Dignity and Humiliation Studies*), but also for her books. Another instance is the work of Robert W. Fuller. He has been able to speak to large audiences all over the world using title words like Lindner's and avoiding the s-word.

The taboo on shame seems to have weakened in the last 10 years among researchers. The downward slope for the word shame has slowed in the Ngrams. But it continues to exert a powerful influence in the vernacular and even in research: shame is still close to being unspeakable and unprintable. The next section will outline a theory that can be used to explain this taboo and the possibility that it can have destructive effects.

Cooley's Brief Examples and Goffman's Lack of Explicit Theses

Cooley offered only brief explication of his two propositions, as in this passage that introduces his idea that we usually do not know that we are living in the minds of others and producing emotions. We only realize it, he states, in extreme or unusual situations:

> Many people of balanced mind…scarcely know that they care what others think of them, and will deny, perhaps with indignation, that such care is an important factor in what they are and do. But this is an illusion. If failure or disgrace arrives, if one suddenly finds that the faces of men show coldness or contempt instead of the kindliness and deference that he is used to, he will perceive from the shock, the fear, and the sense of being outcast and helpless, that he was living in the minds of others without knowing it, just as we daily walk the solid ground without thinking how it bears us up (1922, p. 208).

In the following passage, Cooley explains how the looking glass self generates shame:

> The comparison with a looking glass hardly suggests the second element, the imagined judgment, which is quite essential. The thing that moves us to *pride or shame is* not the mere mechanical reflection of ourselves, but an imputed sentiment, the imagined effect of this reflection upon another's mind. This is evident from the fact that the character and weight of that other, in whose mind we see ourselves, makes all the difference with our feeling. We are *ashamed* to seem evasive in the presence of a straightforward man, cowardly in the presence of a brave one, gross in the eyes of a refined one, and so on. We always imagine, and in imagining share, the judgments of the other mind. A man will boast to one person of an action – say some sharp transaction in trade – which he would be *ashamed* to own to another (1922, pp. 184–185, emphasis added).

This discussion suggests less abstract situations. In the following passage, Cooley refers to particular, though fictional, events in novels, but without quoting any of them in detail:

> In most of [George Eliot's] novels there is some character like Mr. Bulstrode in *Middlemarch*….whose respectable and long established social image of himself is shattered by the coming to light of hidden truth (1922, p. 208).

Cooley's statement, since it is abstract, gives only a slight sense of how catastrophic the shattering of the social image is, and how far it reaches. In the novel, Bulstrode's wife, Dorothea, although blameless, stands by her disgraced husband. The novel provides detailed particulars so that the reader is alerted to the full force of public humiliation. Using Bulstrode's instance to make his point is somewhat of a departure from Cooley's tendency to abstain from description. However, he

does not go so far as to quote the passage and comment on how the details in it relate to his thesis, as Goffman does.

Here, for example, is a quotation showing one way Bulstrode's disgrace reaches to his wife. Cooley could have used this to illustrate the particulars of his thesis:

> When she had resolved to [stand by her husband], she prepared herself by some little acts which might seem mere folly to a hard onlooker; they were her way of expressing to all spectators visible or invisible that she had begun a new life in which she embraced humiliation. She took off all her ornaments and put on a plain black gown, and instead of wearing her much-adorned cap and large bows of hair, she brushed her hair down and put on a plain bonnet… (Eliot, 1900, p. 338).

Dorothea prepares for a public stripping of her dignity by discarding her socially acceptable appearance, replacing it with what might have been prison or funeral clothing. By only referring to events like this one, rather than quoting them, Cooley was unable to describe the full force of his ideas.

Goffman, on the other hand, freely used a great multitude of concrete examples. His wealth of detailed events may be the key to his popularity and his importance. They remind readers of their own instances: 'That's like me!' They can also be used to illustrate many of Cooley's theses.

Here is an example of a Goffman instance that illustrates Cooleyean themes, with numbers added to help the reader keep track:

> Knowing that his audiences are capable of forming bad impressions of him [1], the individual may come to feel ashamed [2] of a well-intentioned honest act merely because the context of its performance provides false impressions that are bad. Feeling this unwarranted shame, he may feel that his feelings can be seen [3]; feeling that he is thus seen, he may feel that his appearance confirms [4] these false conclusions concerning him. He may then add to the precariousness of his position by engaging in just those defensive maneuvers [impression management] that he would employ were he really guilty. In this way it is possible for all of us to become fleetingly for ourselves the worst person we can imagine that others might imagine us to be (1959, p. 236, numbering added).

This instance is somewhat difficult to understand because it is so complex. It would have helped if Goffman had been even more detailed. Suppose a jokester colleague at the office creates a forbidden sound by pressing a whoopee cushion just as you sit down at your desk. You are embarrassed (2), because you imagine that your colleagues think it was you who made the sound (1). Even though you are not the culprit, you blush (3) because you imagine the others in the office think it was your inappropriate action (4).

In this paragraph, Goffman suggested four very brief internal steps, three of which involve living in the mind of the other. Perhaps it was examples like these that led Bourdieu (1983) to call Goffman 'the discoverer of the infinitely small.' The minuteness about Goffman's particulars like this one is the time scale: perhaps portions of a second for each event that is described. So far the discussion has been about theories of shame. Next comes two studies that seem to support the theories.

Elias on European History

In an extraordinary study over hundreds of years of European history, sociologist Norbert Elias analyzed etiquette and education manuals in five different languages. (*The Civilizing Process* 1939. First translated from German into English in 1978.) There are two main themes: as physical punishment decreased, shame became increasingly dominant as the main agent of social control; and as shame became more prevalent, it also became almost invisible because of taboo.

The following excerpt gives the flavor of Elias's study. It is from a nineteenth-century work (von Raumer, 1857) that advises mothers how to answer the sexual questions their daughters ask:

> Children should be left for as long as possible in the belief that an angel brings babies. If girls should later ask how children come into the world, they should be told that the good Lord gives the mother her child…"You do not need to know nor could you understand how God gives children." It is the mother's task to occupy her daughters' thoughts so incessantly with the good and beautiful that they are left no time to brood on such matters. A mother … ought only once to say seriously: "It would not be good for you to know such a thing, and you should take care not to listen to anything said about it." A truly well brought-up girl will from then on feel shame at hearing things of this kind spoken of (1978, p. 180).

Elias First Interprets the Repression of Sexuality in Terms of Hidden Shame:

> An aura of embarrassment…surrounds this sphere of life. Even among adults it is referred to officially only with caution and circumlocutions. And with children, particularly girls, such things are, as far as possible, not referred to at all. von Raumer gives no reason why one ought not to speak of it with children. He could have said it is desirable to preserve the spiritual purity of girls for as long as possible. But even this reason is only another expression of how far the gradual submergence of these impulses in shame and embarrassment has advanced by this time (1978, p. 180).

Elias raises a host of significant questions about this excerpt, concerning its motivation and its effects. His analysis goes to what may be a key causal chain in modern civilization: denial of shame and of the threatened social bonds that both cause and reflect that denial.

> Considered rationally, the problem confronting him [von Raumer] seems unsolved, and what he says appears contradictory. He does not explain how and when the young girl should be made to understand what is happening and will happen to her. The primary concern is the necessity of instilling 'modesty' (i.e., feelings of shame, fear, embarrassment, and guilt) or, more precisely, behavior conforming to the social standard. And one feels how infinitely difficult it is for the educator himself to overcome the resistance of the shame and embarrassment which surround this sphere for him (1978, p. 181).

Elias's study suggests a way of understanding the social transmission of the taboo on shame and the social bond. The adult teacher, von Raumer, in this case, is not only ashamed of

sex, he is ashamed of being ashamed. The nineteenth-century reader, in turn, probably reacted in a similar way: being ashamed, and being ashamed of being ashamed, and being ashamed of causing further shame in the daughter. von Raumer's advice was part of a social system in which attempts at civilized delicacy resulted and continue to result in an endless chain reaction of hidden shame.

Elias understood the significance of the denial of shame to mean that shame goes underground, leading to behavior that is outside of awareness:

> Neither rational motives nor practical reasons primarily determine this attitude, but rather the shame (*scham*) of adults themselves, which has become compulsive. It is the social prohibitions and resistances within themselves, their own superego, that makes them keep silent (1978, p. 181).

Like many other passages, this one implies not only to a taboo on shame, but to the actual mechanisms by which it is transmitted and maintained.

Helen Lewis's Study of Psychotherapy Sessions

Helen B. Lewis, a research psychologist, used a systematic method to locate verbal emotion indicators in many transcriptions of psychotherapy sessions. She seems to have been unaware of Elias's study. She found, to her surprise, that shame/embarrassment was by far the most frequent emotion, occurring more often than all the other emotions combined. She also found that these instances of shame/embarrassment, unlike joy, grief, fear, or anger, were virtually never mentioned by either the client or the therapist. She called the unmentioned instances 'unacknowledged shame.' Her findings provide support, at the word by word level, for Elias's thesis of the prevalence and invisibility of shame at the historical level.

She also found that the shame in these episodes seemed to be hidden in two different ways. *Overt, undifferentiated shame* (OUS) involved painful feelings that were hidden behind terms that avoided the s-word (Elias used the word 'circumlocutions'). *Bypassed* shame involved rapid thought, speech, or behavior, but little feeling. OUS is marked by pain, confusion, and bodily reactions such as blushing, sweating, and/or rapid heartbeat. One may be at a loss for words, with fluster or disorganization of thought or behavior, as in states of embarrassment.

Many of the common terms for painful feelings appear to refer to OUS: feeling peculiar, shy, bashful, awkward, funny, bothered, or miserable; in adolescent vernacular, being freaked, bummed, or weirded out. The phrases 'I felt like a fool,' or 'a perfect idiot' are prototypic. Some of the substitute terms involve phrases. The example used earlier, 'an awkward moment': It's not me that embarrassed (denial), but the moment that is awkward (projection). As indicated above, Retzinger's article (1995) lists over a 100 substitute words and phrases.

Bypassed shame is manifested as a brief painful feeling, usually fleeting, followed by obsessive and rapid thought or speech. A common example: one feels insulted or criticized. At that moment (or later in recalling it), one might experience a very brief jab of painful feeling, followed immediately by imaginary replays of the offending scene. The replays are variations on a theme: how one might have behaved differently, avoiding the incident, or responding with better effect. One is obsessed.

It seems to me that Lewis's use of a systematic method to detect emotion terms and cognates might have led to underreporting of shame episodes. Her finding of episodes of bypassed shame would be one reason, since the method she used would be better at locating OUS terms, sometimes missing the obsessive talk and/or thinking that characterize bypassed shame.

Elias's method was unsystematic, and for that reason, probably much broader than Lewis's. He examined all topics that frequently occur in the books he examined: sexuality, body functions, modesty, delicacy, manners, embarrassment, and what he named 'sociogenetic fear.' By the last phrase he was not referring to fear in the sense of a response to physical danger. Rather, he used it as another way of referring to shame. This kind of sidestepping of the s-word occurs frequently in everyday conversations: 'I fear rejection' has nothing to do with physical danger: it usually means 'I anticipate shame.'

Both the study by Elias and the one by Lewis can be seen as hinting that shame might be ubiquitous yet invisible in modern societies, but neither makes that point explicitly.

How Secret Shame Can Lead to Silence or Violence

Goffman also added a further thesis to the looking glass self, a fourth step to the three proposed by Cooley: managing (such as hiding) shame that could not be avoided. Furthermore, there is a fifth step barely hinted at by Goffman: hiding shame during the fourth step can generate a fifth step in the form of behavior. Helen Lewis (1971) noted that shame may result in withdrawal or even depression, on the one hand, or anger and aggression, on the other. The work of Retzinger (1991), the psychiatrist Gilligan (1997), and the sociologist Websdale (2010) follow-up on the latter direction. These four studies show how the escalation of anger and violence is caused by hidden shame.

> The emotion of shame is the primary or ultimate cause of all violence… The different forms of violence, whether toward individuals or entire populations, are motivated (caused) by *secret* shame (1997, pp. 110–111).

Gilligan's theory is of great interest, since it proposes an emotional cause for both interpersonal and mass violence. Websdale's (2010) study of 211 cases of familicide found strong support for Gilligan's thesis. A finding of particular interest in his study was the sizable minority of perpetrators who had what he calls a *civic-respectable* (C-R) style, in contrast to the majority whose style was angry and aggressive.

The C-R killers had no history of violence and little evidence even of anger. They were almost all middle class men (and a few women) who had lost their jobs. They hid the fact by continuing to leave the house every weekday as if going to work. What they did during their absence was to plan in detail

the killing of their family, and often, themselves. Proud of their abilities as a breadwinner, they apparently could not bear the humiliation of being jobless.

This C-R style of violence, it seems to me, has deep parallels to the preparation of nations for wars of revenge, as was the case of France preparing to make war on Germany in the period 1871–1914 (Scheff, 1997). Especially for the leaders, both shame and anger are carefully hidden behind a veil of rationality. The Bush administration may have been deeply embarrassed by the 9/11 attack during their watch and their helplessness to punish the attackers. The invasion of Iraq on the basis of false premises might have served to hide their shame and anger. The idea of the looking glass self, especially when it is expanded to at least five steps, can serve to generate a large group of general propositions about both interpersonal and collective behaviors.

Neither Cooley nor Goffman dealt with the idea of justified pride, nor have many others. Cooley discussed pride and vanity (1922, pp. 230–237), but his version of pride confounds it with egotism, the usual case in vernacular English. Tracy et al. (2009) have recently noted this confound, distinguishing between what they call authentic (justified) and hubristic pride (egotism).

Conclusion

This article has reviewed four master studies of shame and its place in human life. Two of the studies were articles that imply a theory of shame, giving rise to what might be called the Cooley/Goffman conjecture. Two were empirical studies that seem to support the theory. In its most general terms, it proposes that shame or its anticipation is virtually ubiquitous, yet usually invisible in modern societies. This idea seems to be strongly supported by the substantial empirical studies made by Elias of macrohistory and Lewis of the microworld of social interaction. The theory needs to be further and more directly tested. If further studies support it, perhaps social–behavioral science and psychiatry need to proceed in a new direction.

See also: Aggression, Social Psychology of; Alienation, Sociology of; Honor and Shame

Bibliography

Aiden, Ezra, 2011. http://books.google.com/ngrams.
Bourdieu, P., 1983. Forms of capital. In: Richards, J.C. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York.
Cooley, Charles H., 1922. Human Nature and Conduct. Scribners, New York.
Corballis, Michael, 2007 May–June. The uniqueness of human recursive thinking. American Scientist 95, 240–248.
Elias, Norbert, 1978. The Civilizing Process. Blackwell, London.
Eliot, George, 1900. Middlemarch. Sully and Kleinteich, New York.
Erikson, Erik, 1950. Childhood and Society. Norton, New York.
Gilligan, James, 1997. Violence – Reflections on a National Epidemic. Vintage Books, New York.
Goffman, Erving, 1956. Embarrassment and social organization. American Journal of Sociology 62, 264–274.
Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Doubleday/Anchor Books, Garden City, NY.
Goffman, Erving, 1961. Asylums. Anchor Books, New York.
Goffman, Erving, 1961a. Fun in games. In: Encounters: Two Studies in the Sociology of Interaction. Bobbs-Merrill, Indianapolis, pp. 15–81.
Goffman, Erving, 1967. Interaction Ritual: Essays on Face-to-face Behavior. Pantheon Books, New York.
Kaufman, Gershen, 1989. The Psychology of Shame. Springer, New York.
Lewis, Helen B, 1971. Shame and Guilt in Neurosis. International Universities Press, New York.
Mead, George. H, 1934. Mind, Self, and Society. University of Chicago Press, Chicago.
Retzinger, Suzanne, 1991. Violent Emotions. Sage, Newbury Park.
Retzinger, Suzanne, 1995. Identifying shame and anger in discourse. American Behavioral.
Scheff, Thomas, 1984 June. The taboo on coarse emotions. Review of Personality and Social Psychology.
Scheff, Thomas, 1990. Microsociology. University of Chicago Press, Chicago.
Scheff, Thomas, 1994. Bloody Revenge: Nationalism, War, and Emotion. Westview, Boulder. Re-issued by iUniverse (2000).
Scheff, Thomas, 1997. Emotions, the Social Bond, and Human Reality: Part/Whole Analysis. Cambridge University Press, Cambridge.
Scheff, Thomas, 2000. Shame and the social bond: sociological theory. Sociological Theory 18, 84–98.
Scheff, Thomas, 2011 May. Scientism is social/behavioral science. Contemporary Sociology.
Scheff, Thomas, Retzinger, Suzanne, 1991. Violence and Emotions. Lexington Books, Lexington, Mass. Re-issued by iUniverse (2001).
Tomkins, Silvan, 1963. Affect, Imagery, Consciousness, vol. II. Springer, New York.
Tracy, Jessica, Joey, Cheng, Richard, Robins, Kali, Trzesniewski, 2009. Authentic and hubristic pride: the affective core of self-esteem and narcissism. Self and Identity 8 (2 & 3), 196–213.
von Raumer, Wilhelm, 1857. The Education of Girls (cited in Elias, 1978).
Websdale, Neil, 2010. Familicidal Hearts: the Emotional Style of 211 Killers. Oxford University Press, Oxford.

Shared Belief: Philosophical Aspects

Raimo Tuomela, University of Helsinki, Helsinki, Finland; and University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The importance of the notion of shared belief has been emphasized by philosophers, economists, sociologists, and psychologists at least since the 1960s. This article concentrates on the strongest kind of shared belief, to be called 'mutual belief.' While a belief can be shared in the trivial sense of two or more people having the same belief, mutual belief requires the kind of strong sharing that requires at least the participants' awareness (belief) of their similar belief. Mutual belief is one central kind of collective attitude; examples of others are collective intentions, wants, hopes, and fears. Understandably, collective attitudes are central explanatory notions for the social sciences, as one of their tasks is to study collective phenomena, including various forms of collective thinking and acting. As to mutual beliefs, they serve to characterize social or intersubjective existence in a sense not relying on the making of agreements or contracts. Thus, many social relations, properties, and events – including intentional communication – arguably involve mutual beliefs. Furthermore, characterizations of many other collective attitudes depend on the notion of mutual belief. This article discusses various kinds of shared belief and the problem of whether and when several iterated beliefs are needed. While the emphasis is on individual-mode beliefs, it is pointed out that in full-blown group contexts, special group-mode (or 'we-mode') beliefs will be required of the group members.

Introducing the Concept of Shared Belief

The importance of the notion of shared belief has been emphasized by philosophers, economists, sociologists, and psychologists at least since the 1960s (earlier contributors include Schelling, 1960; Scheff, 1967; Lewis, 1969; Schiffer, 1972). This article deals with the strongest kind of shared belief, to be called 'mutual belief.' While a belief can be shared in the trivial sense of two or more people having the same belief, mutual belief requires the kind of strong sharing that requires at least the participants' awareness (belief) of their similar belief. Mutual belief is one central kind of collective attitude; examples of others are collective intentions, wants, hopes, and fears. Understandably, collective attitudes are central explanatory notions in the social sciences, as one of the tasks of these sciences is to study collective phenomena, including various forms of collective thinking and acting (illustrations are given in this article).

There is some special, additional interest in the notion of shared belief as mutual belief. First, mutual beliefs serve to characterize social or intersubjective existence in a sense that does not rely on the participants making agreements or contracts. Thus, many social relations, properties, and events arguably involve mutual beliefs (cf. Lewis, 1969; Ruben, 1985; Lagerspetz, 1995; Tuomela, 1995). As a simple example, think of the practice of two persons, *A* and *B*, shaking hands. It presupposes that *A* believes that *B* and *A* are shaking hands, and that *A* also believes that *B* believes similarly; and *B* must believe analogously. Similarly, communication has been argued by many philosophers, especially Grice, to involve mutual belief (cf. Schiffer, 1972; Grice, 1989). Second, characterizations of many other collective attitudes in fact depend on the notion of mutual belief (cf. Balzer and Tuomela, 1997).

In social psychology and sociology, theoreticians often speak about *consensus* instead of mutual belief. The notion of consensus – with the core meaning 'mutual belief' – has been regarded as relevant to such topics as public opinion, values, mass action, norms, roles, communication, socialization, and group cohesion. It can also be mentioned that fads, fashions, crazes, religious movements, and many other related phenomena have been analyzed partly in terms of shared beliefs, consensus, shared consensus, mutual belief, or some similar notions. As pointed out by the sociologist Scheff (1967), such analyses have often gone wrong because they have treated consensus merely as a shared first-order belief. Thus, as Scheff argues, consensus as mere first-order agreement does not properly account for 'pluralistic ignorance' (where people agree but do not realize it) or 'false consensus' (where people mistakenly think that they agree). Scheff proposes an analysis in terms of levels of agreement corresponding to a hierarchy of so-called loop beliefs (e.g., in the case of two persons *A* and *B*, *A* believes that *B* believes that *A* believes that something *p*). As is easily seen, pluralistic ignorance and false consensus are second-level phenomena. The third level will have to be brought in when speaking about people's awareness of these phenomena. Other well-known social psychological notions requiring more than shared belief are Mead's concept of 'taking the role of the generalized other,' Dewey's 'interpenetration of perspectives,' and Laing's metaperspectives (see, e.g., Scheff, 1967).

There are two different conceptual–logical approaches to understanding the notion of mutual (or, to use an equivalent term, common) belief: (1) the *iterative* account, and (2) the *reflexive* or *fixed-point* account. According to the iterative account, mutual belief is assumed to mean iterable beliefs or dispositions to believe (cf. Lewis, 1969, Chapter 2 and for the weaker account in terms of dispositions to come to believe, Tuomela, 1995, Chapter 1). In the two-person case, mutual belief amounts to this according to the iterative account: *A* and *B* believe that *p*, *A* believes that *B* believes that *p* (and similarly for *B*), *A* believes that *B* believes that *A* believes that *p* (and similarly for *B*), and the iteration can continue as far as the situation demands. In the case of loop beliefs, there is accordingly mutual awareness only in a somewhat rudimentary

sense. As will be seen, in many cases one needs only two iterations for functionally adequate mutual belief: *A* and *B* believe that *p*, and they also believe that they believe that *p*. However, there are other cases in which it may be needed to go higher up in the hierarchy.

The fixed-point notion of mutual belief can be stated as follows: *A* and *B* mutually believe that *p* if and only if they believe that *p* and also believe that it is mutually believed by them that *p*. No iteration of beliefs is at least explicitly involved here. Correspondingly, a clear distinction can be made between the iterative or level account and the fixed-point account (to be briefly commented on in Section Mutual Belief as Shared We-Belief).

One can speak of the individual (or personal) mode and the group mode of having an attitude such as a belief or intention. This article deals with the general notion of mutual belief, be it in the individual mode or in the group mode. Some remarks on the present distinction are anyhow appropriate here. The group-mode sense, expressible by 'We, as a group, believe that *p*,' requires that the group in question is collectively committed to upholding its mutual belief or at least to keeping the members informed about whether it is or can be upheld. This contrasts with mutual belief in an aggregative individual mode involving only personal commitments to the belief in question. When believing in the group-mode sense, group members are accordingly committed to a certain shared view of a topic and to group-mode thoughts such as 'We, as a group, believe that *p*.' Group-mode beliefs are central in the analysis of the kinds of beliefs that structured social groups such as organizations and states have. According to the 'positional' account defended by Tuomela (1995), the group members authorized for belief or view formation collectively accept the views that will qualify as the beliefs the group has. These views are group-mode views accepted for the group and are, strictly speaking, acceptances of something as the group's views rather than beliefs in the strict sense.

Mutual Beliefs More Precisely Characterized

This section will be concerned with what mutual beliefs involve over and above plain shared beliefs, and Section Mutual Beliefs and the Level Problem takes up the problem of how many layers of hierarchical beliefs are conceptually or psychologically needed. (The discussion below draws on the treatment in Tuomela (1984, 1995).)

Consider now the iterative account starting with the case of two persons, *A* and *B*. Recall from Section Introducing the Concept of Shared Belief that, according to the standard iterative account, *A* and *B* mutually believe that *p* if and only if *A* believes that *p*, *B* believes that *p*, *A* believes that *B* believes that *p* (and similarly for *B*), *A* believes that *B* believes that *A* believes that *p* (and similarly for *B*), and so on, in principle *ad infinitum*. In the general case, the account defines that it is mutually believed in a group, say *G*, that *p* if and only if (1) everyone in *G* believes that *p*, (2) everyone believes that everyone believes that *p*, (3) everyone believes that everyone believes that everyone believes that *p*, (4) and so on *ad infinitum*. The word 'everyone' can, of course, be qualified if needed and made dependent on some special characteristics; for example, it can be restricted to concern only every fully fledged, adequately informed, and suitably rational member of the group.

A major problem with the iterative account is that it seems in some cases to be psychologically realistic. It loads the people's minds with iterated beliefs that people, after all, do not experientially have and possibly, because of lack of rationality and memory, cannot have. Thus, the account must be improved to make it better correspond to psychological reality. One way to go is to operate partly in terms of lack of disbelief. A person's lack of disbelief that *p* is defined as it is not the case the person believes the negation of *p*. One can try to define mutual belief schematically by saying that it is mutually believed in *G* that *p* if and only if everyone believes that *p*, iterated *n* times, and from level $n+1$ on everyone lacks the disbelief that *p*. With $n = 2$ (viz., two levels actually present), this definition says that mutual belief amounts to everyone believing that *p* and everyone believing that everyone believes that *p* and that everyone lacks the disbelief that *p* from the second level on.

The basic reason for using this amended iterative approach is that no more levels of positive belief than people actually have should be required. Provided that the value of *n* can be established, the present account satisfies this. It must be assumed here that the agents in question are to some degree intelligent and rational as well as free from emotional and other disturbances so that they, for instance, do not lack a higher-order belief when they really ought to have one in order to function optimally.

While the iterative account amended with the use of the notion of lack of disbelief seems viable for some purposes, an alternative analysis is worth mentioning. This analysis amends the original iterative account in that it requires iterative beliefs up to some level *n* and from that level on requires only the disposition to acquire higher-order beliefs in appropriate conditions. These appropriate conditions, serving to determine the value of *n*, include both background conditions and more specific conditions needed for acquiring the belief in question. First, the agents in *G* must be assumed to be adequately informed and share both general cultural and group-specific information; especially, they must have the same standards of reasoning so that they are able to 'work out the same conclusions' (see Lewis, 1969, Chapter 2; Heal, 1978). Second, they must be sufficiently intelligent and rational, and they must be free from cognitive and emotional disturbances to a sufficient degree (so that there are no psychological obstacles for adding the $n+1$st order belief when a genuine mutual belief is involved and for not adding it, or indeed for adding its negation, when a genuine mutual belief is not involved). A typical releasing condition for the disposition to acquire a higher-order belief would be simply that the agents are asked about higher-order beliefs (or are presented with other analogous problems concerning them).

There is not much to choose between the two amended accounts, and the matter will not be discussed further here. With the conceptual machinery at hand, one can deal with, for instance, the mentioned phenomena of pluralistic ignorance and false consensus. An account of pluralistic ignorance, concerning something *p*, must obviously include the idea that everyone believes that *p*. Second, it must say that the agents do not have the belief that they agree. Compatibly with this, it can still be required that they do not disbelieve that they believe

that p. As to false consensus, it must be required that not everyone believes that p and that, nevertheless, it is believed by everyone that everyone believes that p.

Mutual Beliefs and the Level Problem

Let us next discuss the level problem. Given the iterative account, the problem is how many iterative levels of belief are conceptually, epistemically, or psychologically needed for success in various cases. The criteria here are functionality and success. For instance, how many levels of belief does successful joint action require? The level question is difficult to answer in general terms. However, it can be conjectured that in certain cases at least two levels (viz., $n = 2$), relative to the base level, are needed for mutual belief, but no more. In some other cases, this is not enough. Generally speaking, there are very seldom epistemic or psychological reasons for going beyond the fourth level. For many agents, level $n = 4$ probably is psychologically impossible or at least very difficult to handle in one's reasoning.

The base level is level 0, at which the sentence or proposition p is located, and p will concern or be about an agent's or agents' actions, dispositions to act, intentions, beliefs, or something related; or, alternatively, p can be (or concern) a belief content. In the analysis below, it is accepted that the beliefs – at least those going beyond the level $n = 2$ – need not be proper occurrent or standing beliefs or even subconscious ones but only dispositions to form the belief in question (see Audi, 1994 for the relevant notion of a disposition to form a belief).

The claim that $n = 2$ is not only necessary but also sufficient in many typical cases is allowed to involve social loop beliefs where the content of mutual belief could be, for instance, that A believes that B believes that A will act in a certain way, or believes that such and such is the case. What the base level will be is relative to the topic at hand. That at least two levels are needed can be regarded as a conceptual truth (related to the notion 'social existence'); and that in many typical cases no more is needed or, in some other cases, four layers are needed can be regarded as general psychological theses concerning the actual contents of people's minds in those situations.

The present view allows that a need may arise to add layers indefinitely. There follows a recursive argument to show that such a situation is possible in principle. Given an analysis of mutual belief in terms of iteration of beliefs such as ours, then at each level under consideration one may start asking questions about people's realizing (or believing) such and such, where 'such and such' refers to the level at hand. This game obviously can go on indefinitely as far as the involved agents' reasoning capacities (or whatever relevant capacities) permit. This is an inductive argument (in the mathematical sense – the first step is obvious) for the open and infinite character of the hierarchy of nested beliefs.

To discuss the level problem in more concrete terms, let us consider mutual belief in the case of two agents, A and B. First, we let our base-level sentences be related to the agents' performances of their parts as follows:

(a) A will do his part of X (in symbols, $p(A)$);
(b) B will do his part of X ($p(B)$).

Next, let us assume:

(i) A believes that (a) and (b);
(ii) B believes that (a) and (b).

This is a social or mutual recognition of the agents' part-performances in the context of joint action.

The sentence p in (a) and (b) may alternatively concern other things than action. Thus it may represent a belief content, for example $p =$ The earth if flat, and we will consider this case below.

Considering joint action, it can be argued that $n = 2$ is both necessary and sufficient for successful performance. In this example, A will obviously have to believe that B will do her part. This gives her some social motivation to perform her own part. Concentrating on our reference point individual A and genuine joint action X (in which the agents' performances of their parts are interdependent), A must also believe (or be disposed to believe, at any rate, viz., come to form the belief if asked about the matter) that B believes that A will do her part. For otherwise, A could not believe with good reason that B will do her part, for B would – if she does not so believe – lack the social motivation springing from her belief that A will do her part. However, if she did lack that motivation, A could not on this kind of ground defensibly form the belief that B will perform her part. This argument shows that it must be required that the agents have a loop belief or at least a disposition to have a loop belief:

(iii) A believes that (i) and (ii);
(iv) B believes that (i) and (ii).

For instance, (iii) gives the loop 'A believes that B believes that $p(A)$,' and this means that in the present kind of case, $n = 2$ is necessary and normally also sufficient (relative to the chosen base level).

An analogous claim can be defended in the case of belief. Here the defense can be given in terms of the recognition by the agents that they share a belief. Suppose not only that each member of the Flat Earth Society believes that the Earth is flat, but also that – because of being organized into a belief group – the members all share this information. This second-order belief indicates the first level at which social and mutual belief can be spoken of in an interesting sense, for at this level people recognize that the others in the group believe similarly, and this establishes a doxastic social connection; at level one that recognition is, however, missing. This is a typical case of a social property, and again $n = 2$ is both necessary and normally sufficient for acting in matters related to the shape of the Earth *qua* a member of the society.

Suppose now the members of the Flat Earth Society begin to reflect upon this second-order belief of theirs and come to form the belief that indeed it does exist. (Perhaps a sociologist has just discovered that they share the second-order belief and told them about it.) They might wonder how strong this second-order awareness is, and so on. All this entails that on some occasions, third-order beliefs may be needed, although the standard case clearly would be second-order beliefs accompanied by a disposition to go higher up if the need arose. Analogously, in the case of joint action in some very special cases, for example, double loop beliefs may be needed for rational action.

The reader should be reminded that not all social notions depend on mutual belief. For instance, latent social influence and power do not require it, nor does a unilateral love relation between two persons.

Mutual Belief as Shared We-Belief

What have been called 'we-attitudes' are central for an account of social life (see Tuomela, 1995, Chapter 1; Tuomela and Balzer, 1999). A we-attitude is a person's attitude (e.g., belief, want, or fear) concerning something p (a proposition or sentence) such that (1) this person has the attitude in question and believes that (2) everyone in the group has the attitude and also believes that there is a mutual belief to the effect that (2) in the group. Of these, clause (1) is required because there cannot of course be a truly shared attitude without all the members participating in that attitude. Clause (2) gives a social reason for adopting and having the attitude A, and (3) strengthens the reason by making it intersubjective. (A shared we-attitude can be either in the group mode or in the individual mode; cf. Section Introducing the Concept of Shared Belief.)

A shared we-belief is a we-belief (be it in the individual or in the group mode) that (ideally) all the group members have. To take an example, a shared we-belief that the Earth is flat entails that the group members believe that the Earth is flat, and also that it is a mutual belief in the group that the group members believe that the Earth is flat.

Shared we-beliefs can be related to the reflexive or fixed-point account of mutual belief. According to the simplest fixed-point account, mutual belief is defined as follows: it is a mutual belief in a group that p if and only if everyone in the group believes that p and also that it is mutually believed in the group that p. It can thus be seen that the account of mutual belief given by the fixed-point theory is equivalent to the definiens in the definition of a shared we-belief. In the fixed-point approach, the syntactical infinity involved in the iterative approach is cut short by a finite fixed-point formula that is an impredicative construct in which the joint notion to be 'defined' already occurs in the definiens. Under certain rationality assumptions about the notion of belief, it can be proved that the iterative approach, which continues iterations *ad infinitum*, gives the fixed-point property as a theorem (see Halpern and Moses, 1992 and for a more general account, Balzer and Tuomela, 1997).

The fixed-point account is in some contexts psychologically more realistic, as people are not required to keep iterative hierarchies in their minds. Note, however, that it depends on context whether the iterative approach or the fixed-point approach is more appropriate. Thus, in the case of successful joint action, at least loop beliefs must be required.

See also: Collective Action; Collective Behavior, Sociology of; Collective Beliefs, Sociological Explanation of; Collective Identity; Collective Memory, Anthropology of; Collective Memory, Psychology of; Gene–Culture Coevolution; Human Cognition, Evolution of; Religion, Sociology of; Religion: Evolution and Development.

Bibliography

Audi, R., 1994. Dispositional beliefs and dispositions to believe. Nous 28, 419–434.
Balzer, W., Tuomela, R., 1997. A fixed point approach to collective attitudes. In: Holmstrom-Hintikka, G., Tuomela, R. (Eds.), Contemporary Action Theory, vol. II. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 115–142.
Dunin-Keplicz, B., Verbrugge, R., 2010. Teamwork in Multi-Agent Systems: A Formal Approach. Wiley, Chichester.
Grice, P., 1989. Studies in the Ways of Words. Harvard University Press, Cambridge, MA.
Halpern, J., Moses, Y., 1992. A guide to completeness and complexity for modal logics of knowledge and belief. Artificial Intelligence 54, 319–379.
Heal, J., 1978. Common knowledge. Philosophical Quarterly 28, 116–131.
Lagerspetz, E., 1995. The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions. Kluwer Academic Publishers, Dordrecht, The Netherlands.
Lewis, D., 1969. Convention, A Philosophical Study. Harvard University Press, Cambridge, MA.
Ruben, D.-H., 1985. The Metaphysics of the Social World. Routledge, London.
Scheff, R., 1967. Toward a sociological model of consensus. American Sociological Review 32, 32–46.
Schelling, R., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, MA.
Schiffer, S., 1972. Meaning. Oxford University Press, Oxford, UK.
Tuomela, R., 1984. A Theory of Social Action. Reidel, Dordrecht, The Netherlands.
Tuomela, R., 1995. The Importance of Us: A Philosophical Study of Basic Social Notions. Stanford Series in Philosophy. Stanford University Press, Stanford, CA.
Tuomela, R., 2007. The Philosophy of Sociality: The Shared Point of View. Oxford University Press, New York (Paperback edition 2010.).
Tuomela, R., Balzer, W., 1999. Collective acceptance and collective social notions. Synthese 117, 175–205.

Sherrington, Sir Charles Scott (1857–1952)

Eike D Schomburg, Georg August University, Gottengen, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14043–14046, © 2001, Elsevier Ltd.

Abstract

The scientific life of Sherrington, born in London in 1857 died in Eastbourne in 1952, is characterized by two main floruits. The first period (1884–1933) with positions in London, Liverpool, and Oxford was occupied by physiological and partly histological investigations of the nervous system. Beside the observation and detailed description of innumerable nervous phenomena, particularly in motor control, Sherrington's, most important contribution to neuroscience resulted from his unique capability to condense his immense knowledge on sensory and motor functions and on the structure of the nervous system in an almost visionary synopsis of the integrative action of the nervous system. His experimentally well-based theories and concepts formed a greatly recognized secure foundation for future developments in neuroscience. During his second period (up to his retirement in Oxford, 1935, and later until his death) Sherrington was mainly dedicated to the philosophy of the central nervous system. Based on his comprehensive knowledge of the nervous system and the field of biology in general, he approached the age-old problem of brain/body–mind relation. Rejecting the assumption of a 'heavenly' origination he defined mind, recognizable by the resulting behavior, as a function of the complex integrative action of the central nervous system.

> Like this old Earth that lolls through sun and shade,
> Our part is less to make than to be made.
>
> Sherrington, 1940, p. 259

To characterize the broad span of Sherrington's engagement he should be described as 'Pathologist, Physiologist, Philosopher and Poet.' It is somewhat arbitrary to pick out only two of them because it is the combination of all these elements that made him both great and unique. However, hardly any other physiologist since 1800 made such a fundamentally important contribution to the knowledge and to the development of concepts in modern neurophysiology. Moreover, when advanced in years, he was the first to attempt a contribution to the solution of the age-old problem of brain/body and mind on a well-founded neuroscientific basis.

The Life

Born in London on 27 November 1857, Sherrington was brought up in Ipswich and Norwich. From his earliest years he felt a strong attraction to poetry, a love which never left him throughout his life. Some of his poems were published in various magazines and later were partly collected in a small volume (Sherrington, 1925).

In 1876 Sherrington first enrolled at St. Thomas' Hospital Medical School in London and in 1881 matriculated at Cambridge University, where he received his BA in 1884 and his MD in 1885. During his undergraduate years he started his research in neuroanatomy and neuropathology, publishing his first papers on these studies in 1884 together with John Newport Langley (Langley and Sherrington, 1884). This was the beginning of a very productive publishing activity with almost 320 scientific publications (for Sherrington's complete bibliography see Fulton, 1952). During his initial postgraduate period at Cambridge Sherrington's scientific interests concentrated on neurohistology and pathology. This period was interrupted by expeditions to epidemics of Asian Cholera in Spain and Italy. The autopsied material taken during these expeditions he analyzed together with J. Graham Brown and Charles S. Roy in England (1885) and with Rudolf Virchow and Robert Koch in Germany (1886). In 1890 Sherrington was appointed lecturer in physiology at St. Thomas' Hospital in London where he mainly performed clinical work. A year later he received the position of the Professor-Superintendent at the Brown Institute at the University of London, where he became remarkably productive with a variety of publications which reflected his pathological, but also his increasing neurophysiological interest. In 1895 Sherrington received his first appointment to a full professorship of physiology at the University College in Liverpool, where he investigated intensely the organization of the spinal cord.

At age 56 (in 1913) Sherrington entered probably the busiest period of his scientific career when he was appointed to the chair of physiology at Oxford University, from which he did not retire until 1935, when he was 78. It was during his Oxford time that he, together with Edgar Adrian, received the Nobel Prize for Medicine in 1932 (the greatest tribute among numerous honours and rewards, like, for example, the Presidency of the Royal Society 1920–25, some 21 international honorary doctorates, and many medals and honorary memberships of British and international societies and academies, among these the corresponding membership of *l'Institut de France*). Three of his students at Oxford were later Nobel Prize winners: Howard Walter Florey, John Carew Eccles, and Ragnar Granit.

In the early 1930s Sherrington performed his last physiological experiments and he felt free to turn to his other great love, the philosophy of the nervous system. Having retired to his old hometown of Ipswich in 1935, his scientific interest was increasingly enriched by philosophical considerations that became evident in the Gifford Lectures, held between 1937 and 1938. Obviously, it was also Sherrington's interest in Natural

Philosophy that caused him to work on Goethe (Sherrington, 1942) and Jean Fernel, a French physician and philosopher of the sixteenth century (Sherrington, 1946).

Though his body was crippled by arthritis, his intellectual capacities remained undimmed until his death in Eastbourne on 4 March 1952.

The Physiologist

Regarding the nervous system as a functional unity resulting from its well-organized parts, Sherrington investigated a variety of motor and sensory systems: the topographical organization of the motor cortex, the binocular integration in vision, as well as the different receptor systems. However, his main efforts concentrated on the spinal cord and its reflex function in motor control. He explicitly admitted that spinal reflex behavior is not 'the most important and far-reaching of all types of nerve behavior' but represents the advantage that 'it can be studied free from complication with the psyche' (Sherrington, 1906/1947, Foreword, p. xiii).

In the spinal cord, isolated from the brain, he carefully analyzed almost all spinal reflex-types from the 'simple' two-neurone reflex arc of the stretch reflex (first described by him), via the complex partly nociceptive flexion and crossed extension reflex and co-ordinated rhythmic scratching and stepping reflexes, up to goal-directed reflex movements of the limbs towards irritating stimuli. Right from the beginning he demonstrated that spinal reflexes are not following a stereotyped all-or-nothing behavior, but that the adaptability of reflexes, based on graduated amplitude, irradiation, and mutual interference between different reflex types, is of great importance for motor control. Sherrington was the first to suggest that inhibition had at least the same functional importance for motor coordination – and other nervous functions – as excitation, not only at the spinal level but also at the brain level and for descending motor commands. This view also opened up new insights into the background of motor disorders observed in various neurological syndromes.

Though some of the reflexes and other motor and sensory phenomena described by Sherrington had been noted earlier, his experimental observations and their analyses were altogether more exact and careful than those of former investigations. Thus, he disproved, e.g., several of Pflügers's reflex laws and the assumption of Descartes (1677) and John and Charles Bell (1826) that reciprocal inhibition is localized peripherally in the muscle. However, the higher experimental sophistication alone would not for sure have received that particularly high enduring recognition. Indeed, Sherrington's most important contribution to neuroscience resulted from his unique capability to condense his immense knowledge on sensory and motor functions and on the structure of the nervous system (partly derived from an intense contact with the famous Spanish neurohistologist Santiago Ramón y Cajal) in an almost visionary synopsis of the integrative action of the nervous system, which laid the foundation for the great progress of neuroscience in the twentieth century and which has lost none of its basic validity. This synopsis made his book *The Integrative Action of the Nervous System* (1906) a milestone in neurophysiology, which has been compared 'in importance with Harvey's' 300 years older '*De motu cordis* in marking a turning point in the history of physiological thought' (Fulton, 1952, p. 174). For Sherrington, spinal reflexes did not have a 'purpose' but a 'meaning.' He replaced the formerly presumed 'soul' of the spinal cord (e.g., Legallois, 1812; Pflüger, 1853) as the driving force for reflexes and even voluntary movements by an astonishing exactly considered function of spinal interneuronal activity. He replaced the Cartesian reflex behaviorist view by the still-relevant view that the performance of coordinated movements as a basis for animal and 'conscious' human behavior requires an integration of 'intrinsically' generated brain functions with 'extrinsically' induced reflex functions. In this context he suggested that the pure 'apsychical' reflex behavior without a contribution of 'mind' loses in importance with phylogenetic ascendance, playing the smallest role in man: "The spinal man is more crippled than is the spinal frog" (Sherrington, 1906/1947, Foreword, p. xiv).

As a true scientist Sherrington never felt that he had arrived at any incontrovertible truth and in a typically unpretentious understatement he concluded: "I have dealt very imperfectly with a small bit of a large problem.... The smallness of the fragment is disappointing" (1931 p. 27); this a statement from a man about whom Frederick George Donnan wrote in 1921 to the Nobel Committee: "In the field of physiology and medicine, Professor Sherrington's works remind one of that of Kepler, Copernicus and Newton in the field of mechanics and gravitation" (in Schuck et al., 1962, p. 310).

The Philosopher

Like for example, Descartes before and Eccles after him, Sherrington realized that mere experimental investigation of the nervous system was not sufficient to solve the fundamental problem of the body/brain–mind relation, particularly the difficulty of explaining how it actually works. In the Rede Lecture on *The Brain and its Mechanism* (Sherrington, 1933) in which he considered the complexity of the brain functions in motor control and behavior and in which he tentatively approached this problem, Sherrington took a fully dualistic view. He even negated any scientific right to "conjoin mental experience with the physiological.... The two ... seem to remain disparate and disconnected" (1933, pp 22–23).

But later in the Gifford Lectures, which he held in Edinburgh in 1937–38 (published as *Man on his Nature* in Sherrington, 1940 and as an extensively revised new edition in 1951) he took up the great enigma of the brain–mind relation in the light of his new neuroscientific knowledge. Rejecting the assumption of a mysterious mind of a 'heavenly' origin he stated, "Ours is an earthly mind which fits our earthly body" (Sherrington, 1940, p. 164). 'Mind' for Sherrington was coupled to motor acts, 'mind' only being recognizable via motor acts. He suggested that 'mind' is graduable, that a recognizable 'mind' is developing in the developing brain in humans and to varying degrees in animals. Regarding the relation between the two concepts of 'energy' (assuming 'matter' also as 'energy' in the physical sense) and 'mind', he concluded that 'mind' as a function of the living brain is mortal, while the 'energy' the matter – is immortal, just changing its state when the brain is dying. In his discussions he entirely avoided the complicating

problem of an 'immortal soul', which he, as being himself oriented towards Natural Religion, left to the Revealed Religions: "When on the other hand the mind-concept is so applied as to insert into the human individual an immortal soul, again a trespass is committed" (Sherrington, 1940, p. 355).

When Sherrington with something like resignation stated, "But the problem of how that [body-mind] liaison is effected remains unsolved; it remains where Aristotle left it more than 2000 years ago" (Sherrington, 1906/1947, Foreword, p. xxiii) he probably underestimated the contribution of his step towards the direction of a solution by defining 'mind' as a function of the brain requiring a complex integrative action of the nervous system; a view which was taken up and further developed by one of his students, Sir John C. Eccles (1970, Eccles and Gibson, 1979).

And Now?

As stated in the section 'The Philosopher' the experimentally well-based theories and concepts, which Sherrington had creatively developed with great visionary imagination, formed a well-recognized secure foundation for future developments in the neurosciences.

The half-century after Sherrington's death has brought an immense increase in neuroscientific knowledge. First, particularly, the technique of microrecordings from nerve cells and the analysis of ionic membrane mechanisms confirmed and extended the knowledge and hypotheses based on Sherrington's activities on brain and spinal functions in sensory perception, motor control, and behavior. Then new techniques allowed for analytic investigation of more and more detailed neuronal structures, functions, and interactions down to the molecular level. New gentechnical approaches enabled the 'production' of innumerable variants of mice with defined genetic defects at the molecular level.

However, the bewildering mass of results originating from these experiments opened up a host of new questions. What is required now is a new 'Sherringtonian' approach which comprises a complex critical analysis of today's entire neuroscientific knowledge (from molecular mechanisms up to the level to complex behavior) in an integrative synopsis and which develops new concepts to explain the sophisticated brain functions and their relation to mind as expressed by behavior. Despite the exponential increase in detailed neuroscientific knowledge, we have not really come closer to a solution of that problem since Sherrington.

See also: Cognitive Neuroscience; Consciousness, Neural Basis of; Functional Brain Imaging of Language Processes; Motor Control Models: Learning and Performance; Motor Cortex; Neural Plasticity; Neural Representations of Direction (Head Direction Cells); Theory of Mind During Infancy and Early Childhood Across Cultures, Development of; Topographic Maps; Vestibulo-Ocular Reflex, Adaptation of the.

Bibliography

Bell, J., Bell, C., 1926. Anatomy and Physiology of the Human Body. Edinburgh.
Descartes, R., 1677. De homine ed. by de la Forge. Amsterdam.
Eccles, J.C., 1970. Facing Reality. Springer-Verlag, New York.
Eccles, J.C., Gibson, W.C., 1979. Sherrington – His Life and Thought. Springer International, Berlin.
Fulton, J.F., 1952. Sir Charles Scott Sherrington, O.M. Journal of Neurophysiology 15, 1667–1690.
Granit, R., 1966. Charles Scott Sherrington – An Appraisal. Nelson, London.
Langley, J.N., Sherrington, C.S., 1884. Secondary degeneration of nerve tracts following removal of the cortex of the cerebrum in the dog. Journal of Physiology 5, 49–65.
Legallois, M., 1812. Expériences sur la principe de la vie. D'Hautes, Paris.
Pflüger, E., 1853. Die sensorischen Funktionen des Rückenmarks der Wirbeltiere. Springer, Berlin.
Schück, H., Sohlman, R., Osterling, A., Liljestrand, G., Westgren, A., Siegbahn, M., Schou, A., Ståle, N.K., 1962. Nobel, the Man and His Prizes. The Nobel Foundation, Elsevier, Amsterdam.
Sherrington, C.S., 1906/1947. The Integrative Action of the Nervous System, second ed. Yale University Press, New Haven, CT, with a new Foreword, 1947.
Sherrington, C.S., 1925. The Assaying of Brabantius and Other Verse. Oxford University Press, London.
Sherrington, C.S., 1931. Quantitative management of contraction in lowest level coordination. Hughlings Jackson Lecture. Brain 54, 1–28.
Sherrington, C.S., 1933. The Brain and its Mechanism. Cambridge University Press, Cambridge, UK.
Sherrington, C.S., 1940/1951. Man on his Nature, second rev. ed. Cambridge University Press, Cambridge, UK, 1951.
Sherrington, C.S., 1942. Goethe on Nature and on Science. Cambridge University Press, Cambridge, UK.
Sherrington, C.S., 1946. The Endeavour of Jean Fernel. Cambridge University Press, Cambridge, UK.

Shils, Edward (1910–95)

Stephen Turner, University of South Florida, Tampa, FL, USA
Steven Grosby, Clemson University, Clemson, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Edward Shils was a prominent American sociologist and social theorist who spent much of his career in Britain. He was the translator of Karl Mannheim and collaborator with Talcott Parsons. His own social theory concentrated on the relation of primary groups and intellectuals to the center of society, which he conceived of in terms of its charismatic character. Unlike Parsons, he was especially concerned with the conflicts between the social attachments of people, and especially with those involving the transcendental. These conflicts were the basis of his account of intellectuals and their response to their own society.

Edward Albert Shils was born in Massachusetts on 1 July 1910 and died in Chicago on 23 January 1995. His father was a Russian Jewish immigrant, a cigar maker, and an employer. His mother was born in Pennsylvania. Neither was educated beyond the eighth grade. The family name was originally Shilo, a variant of a Hebrew word meaning 'his gift.' Shils, having grown up in Philadelphia, attended the University of Pennsylvania, graduating in 1932, having majored in French literature. He trained as a social worker at the New York School of Social Work, and was employed as a social worker during his training, in Chicago. In Chicago, he became involved with the sociologist Louis Wirth, with whom he studied German sociological theory. He was also involved with the economist Frank Knight, who had done the first translation of Weber into English, *General Economic History* (1927), taking a seminar on Weber's *Wirtschaft und Gesellschaft* organized by Knight. Shils also translated other works by Weber (and Simmel) that circulated among students. These eventually became the basis for the translation of three of Weber's most important methodological essays, *Max Weber on the Methodology of the Social Sciences* (1949). During the 1930s, Shils was employed on grants and spent a year at Columbia, before being appointed to the teaching staff of the (Undergraduate) College of the University of Chicago.

He played a central role at The University of Chicago in the development of the famous sequence of social science courses that were part of the general education of undergraduates, and during the 1940s recruited an especially distinguished set of teachers for them, including David Reisman, Barrington Moore, and Sylvia Thrupp. He was especially involved in an advanced course called Freedom and Order, which dealt with the more philosophical aspects of social science.

During the war, Shils, while in England, along with Morris Janowitz, interrogated captured U-boat crew. Out of these interrogations emerged his study of the social relation of small army units as primary groups, without very little, if any, ideological underpinnings. Shils was appointed for part of the year to the London School of Economics in 1946, and was appointed in 1947 to the Committee on Social Thought, and continued in this capacity until his death. From this period to his death, he held joint appointments at the University of Chicago and several universities in England, in particular at the Cambridge University.

Shils wrote in a variety of fields, most of which he came to out of personal acquaintance. At the London School of Economics, he was exposed to the students of the Colonial Empire, notably India. This led to a long engagement with Indian intellectuals and his monograph *The Intellectual between Tradition and Modernity: The Indian Situation* (1961a), and in the 1950s to participation in the New Nations project at the University of Chicago. In the late 1940s, he was appointed to a committee of scientists concerned with the atomic bomb. This led to a long-term engagement with science and science policy, and to his book *The Torment of Secrecy* (1956). His relation with Michael Polanyi, beginning in 1946, led him to a distinctive interest in the social nature of science, to involvement in the Congress for Cultural Freedom, and ultimately in 1962 to the founding of the journal *Minerva*, which drew many of its contributors, especially in the early years, from the participants in the struggle against the Marxian picture of science.

Although Shils was an intellectuals' intellectual, with broad and penetrating engagement in many different scholarly areas, he was not a 'public intellectual' in the conventional sense. Nonetheless, he received many public honors, including the Jefferson Lecturer in 1979 and the Balzan Prize in 1983. He was intensely and personally engaged with leading intellectuals, scientists, and literary figures throughout the world. He was an important behind-the-scenes participant in the *Bulletin of Atomic Scientist* and the journal *Encounter*, and played a large role in *The American Scholar*, as a board member. At dinner in his home, one might find at the table scholars like Arnaldo Momigliano, Owen Chadwick, Hugh Trevor-Roper, the sinologist Michael Loewe, or Wolfgang Mommsen. He was a close friend of Raymond Aron and of the novelist Dan Jacobson and Nobel Laureates Saul Bellow and V.S. Naipaul, and a champion of the Indian author Narit Chaudhari. He brought Bellow to the Committee on Social Thought at the University of Chicago, and Bellow drew on his persona to create some of his most memorable characters. Despite the fact that he lived in this rarified atmosphere both in Cambridge and Chicago, he had a strong sense of empathy with ordinary people, tradesmen, and shopkeepers, and was highly sensitive to their attempts to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61284-9

be respectable and honorable in their dealings with others. This sensitivity can be seen in his published scholarly work, for example, in his humanistic understanding of the *Calling of Sociology* (1961b) and the collection of his articles published under the title *The Constitution of Society* (1972).

Early Writings

Shils differed from his contemporaries Talcott Parsons and Robert Merton in being equally immersed in both the intellectual traditions of American sociology, through his experiences at Chicago, and European sociology. Shils was a great admirer of Robert E. Park and took courses, including one from Ernest Burgess based on the famous Green Bible, Park and Burgess' *Introduction to the Science of Sociology* (1921). He was a student in Park's last seminar in the department, on Collective Behavior. The main line of his intellectual development, and the beginnings of his own distinctive picture of society, came from Charles Horton Cooley and the idea of the primary group. Nevertheless, his thorough study of European thinkers and their intellectual traditions extended his initial concerns radically and in novel directions. His mentor in the 1930s was Louis Wirth, who employed him on a project dealing with the German theorists of the methodological foundations of sociology, and he also worked on a study of the occupational structure of Chicago. At the same time, however, he came into close contact with the economist Frank Knight, who was an admirer of Max Weber. He also became acquainted with Alexander von Schelting, a German refugee who had written a study of Weber's concept of ideal types. Through Knight's seminar on Weber, he became an avid reader of Weber and an active participant in the project of translating Weber, not only the latter's methodological essays but also, along with Max Rheinstein, Weber's writings on law from *Wirtschaft und Gesellschaft* (1954). Through Wirth he became the cotranslator of Mannheim's *Ideology and Utopia* (1936) and the translator of Mannheim's *Man and Society in an Age of Reconstruction* (1940). His career was also helped by the notice of Robert Hutchins, the controversial and charismatic president of the University of Chicago.

In the early 1940s, he served in the US intelligence services (the Office of Strategic Services, predecessor to the Central Intelligence Agency), in England, where he became personally acquainted with Karl Mannheim. He was also exposed to a new strand of thought, which was developing in response to the ideologies that played a role before and during the war. In part, this strand reflected the problem of defining the aims of the war; in part, it represented the attempt by liberal refugees from central Europe to articulate a response to fascism and communism in the nonideological English context. In the course of the decade, Shils became associated with the London School of Economics, establishing relations with Popper, Hayek, Polanyi, and T.S. Eliot, and became aware of the work of Michael Oakeshott. He had already expressed skepticism over Mannheim's views on planning. In London, he also became aware of the issues over the 'planning of science' and freedom of inquiry in science that later occupied much of his time and effort through his editorship of the journal *Minerva*.

His formative experience as an independent researcher was his wartime studies of captured Wehrmacht soldiers and deserters, who were being interviewed to discover the sources of unit cohesion. In the paper *Cohesion and Disintegration of the Wehrmacht in World War II* (1948), published with Morris Janowitz, Shils argued that personal bonds between unit members, not ideology or loyalty to Hitler, accounted for cohesion. Shils later extended the insights of this research in the paper *Primary Groups in the American Army* (1950).

This research, involving the relation between ideology and personal relations, pointed toward a preliminary understanding of society that Shils was later to elaborate in several directions. The conclusion to the cohesion paper was that most men are members of the larger society by means of identifications, which are mediated through the human beings with whom they have personal relations. What was still obscure to him was what these identifications with the larger society were and how they worked. Nevertheless, the model had some significant negative implications. Primary group bonds were critical to the emotional life of people and to their sense of Weness. Ideology was not what held groups together on its own. The "limited capacity of many human beings to sustain a direct, continuous, and intense relationship to the center of their society" ruled out such models of society (1975: p. xxvi). It was also true that some people were sensitive to ideology, to the pull of larger units or ideas, and others had primary group connections with these people, and this could be a formidable source of cohesiveness. Or an ideological movement could fail for the same reason – the inability to develop primary group bonds.

In the study of the Wehrmacht, Shils also discusses themes that become prominent later, including the content of the primary group relations themselves, notably, in the case of the German army, the overwhelming emphasis on honor. He also shows a strong interest in the relation of personality structure and the capacity for different kinds of primary group relations. In this case, discipline and the values and psychological inclinations of Germans of the era fit well with and facilitated unit cohesion. This reasoning reflected the interest in psychoanalysis of Harold Lasswell. The issue reappears in his thorough criticism (1954) of *The Authoritarian Personality* (Adorno et al., 1950).

Shils followed this publication with a consequential collaboration with Talcott Parsons on the nature of values and norms, a definitional enterprise still cited in the interdisciplinary literature on norms. This association marked Shils professionally in the minds of sociologists as a 'Parsonsian.' This association was a double-edged sword: it placed Shils in a position of centrality in academic sociology, but obscured his own originality, and exposed him to the strong negative reaction to Parsons that came to define much discussion of 'sociological theory' over the next two decades. But Shils, who respected Parsons personally, and initially regarded *The General Theory of Action* (Parsons and Shils, 1951) on which they collaborated as an advance, soon became dissatisfied with Parsons' penchant for abstract classificatory schemes, and returned to the elaboration of his own understanding of society.

The collaboration with Parsons had pointed to the lacuna in his own understanding: a sense of what it was that people who

identified with the larger society were identifying with, and what modes of identification existed. His fascination with the question of what binds people was expressed again shortly in his study of the meaning of the coronation of Queen Elizabeth II in 1953, which took his thinking in another direction. Shils had been impressed with Rudolph Otto's *The Idea of the Holy* (1923), Anders Nygren's *Agape and Eros* (1953), and with Weber's ideas about charisma. Shils had a long history of skepticism about the idea that societies were held together by ideology. The Coronation pointed to a different kind of tie, different as well from the primary group.

Shils attempted to resolve the muddle of what kinds of tie held societies together in various ways, in unpublished papers. But despite spending much of his thought on the problem, and operating with a general picture of society, Shils was never entirely happy with his formulations, and never published a comprehensive 'theory of society' that articulated this picture. He was fascinated with the deviations from the core model: the fact that intellectuals, for example, were typically alienated from and often hostile to the central institutions of their own societies. He was also involved in the study of the patched-together 'New Nations' that were emerging out of decolonization, whose states lacked the central institutional status in relation to other social ties characteristic of European nation states, and were often led by intellectuals oriented to Europe.

The key ideas eventually emerged in an incomplete form in *Primordial, Personal, Sacred, and Civil Ties* (1957), where Shils reiterated the negative implications of his understanding that "modern society is no lonely crowd, no horde of refugees from freedom. It is no *Gesellschaft*, soulless, egotistical, loveless, faithless, utterly impersonal and lacking in any integrative forces other than interest or coercion" (1957: p. 112). But his argument also implied that neither is it a Parsonsian system governed by internalized norms nor an all-pervading culture: "the cultural value or belief system of a society can be lived up to only partially, intermittently, and only in an approximate way" (1957: p. 111). The pull of other demands and needs prevents it; immediate needs and authorities are more important.

Shils often wrote about his thought in an autobiographical way. He described "Primordial, Personal, Sacred, and Civil Ties" as a "rambling tale of my intellectual wanderings." This was never intended as a summary of his theoretical image of society, but it comes closest to being one: he mentions a manuscript 'Love, Belief, Civility,' which was intended to be his contribution to an understanding of "the bonds which hold a society together" (1957: p. 112). The book, which would have contained his own 'theory of society,' never appeared. The core of this project was a puzzle inherited from Cooley. Cooley assumed that the ethos of the primary group could be adopted as the ethos of the public life of the larger society; Tönnies contrasted the two and generally regarded them as irreconcilable. Shils turned to Weber to resolve the conflict by extending the notion of charisma. Weber intimated. Shils thought that charisma was at the base of rational legal authority. And this intimation could be generalized to make sense of the different attachments of people not only to their primary group but also to the central institutions of modern society and to ultimate values. Mayo saw that the functioning of corporate bodies depended on the functioning of primary groups. And this fit with Shils own observations of the military, and of ideological groups. In each case, there was a special significance of linking persons with a special degree of attunement to central institutions and ultimate values, who connected the enthusiasm of the primary group with the corporate body.

This basic insight remained with him, but he recognized that it did not differentiate sufficiently between types of ties between people and between forms of attachment to such things as sacred objects, ultimate values, and collective identities. He also recognized the need for a better understanding of collective life, which he could now formulate in terms of 'centers' and 'peripheries.' In the 1950s, he wrote about this in a way that was congruent with the Parsonsian doctrine of the central value system of a society. But Shils was even then construing this in terms of the sacred and charismatic character of central institutions and the values they represented.

Unlike Parsons, Shils was sensitive to the conflicts that arose between various attachments, and especially to the way in which intellectuals appeared to be hostile to the governing institutions and values of their own societies. This phenomenon became a major theoretical focus for him, as he attempted to incorporate both contrasts and attachments within his understanding of center and periphery.

This basic structure of the relation of primary groups to higher things was an unchanging and unambiguous feature of Shils' image of society. A more elusive aspect was his account of the higher things themselves, exemplified by his idea that every society has a center and a periphery that is oriented in some sense to that center, but often in sporadic and negative ways. At some times, Shils comes close to the Parsonsian dogma that society is held together by the existence of a central value system. But Shils discussed the issues of consensus in ways that both extended this idea and conflicted with it. What preoccupied Shils, for most of his work, were cases in which orientations conflicted with one another, and traditions that operated largely independently of the central traditions of a society, such as the tradition of science. In the case of intellectuals, the conflict between their orientation to transcendent ideals and the prosaic, compromise-filled practices of getting and spending of everyday life was the source of their rejection of their own societies.

Shils was particularly fascinated with intellectuals in the developing world, whom he encountered in great numbers at the London School of Economics and during his many trips to India. And the comparison between Western polities and the 'New Nations' overlaid on old societies that had been produced by decolonization sharpened his views of intellectuals, who were torn between their sense of ownership of their own societies, their ambivalence toward traditional culture and their yearning for a deeper connection to it, their attachment to 'aboard' the traditions of oppositional politics, and in the many cases of underemployed intellectuals, by their exclusion from power and desire to exercise it. Each of these attachments can be understood as an orientation to a center, but the orientations conflicted, leading to disappointment and an attraction to ideological critique.

At the theoretical level, Shils gradually came to understand these attachments in terms of the sacred and charisma. Controversially, he radicalized Weber's notion of charisma to apply not only to central institutions but also to anything with

an element of transcendence or superiority, including power itself. This allowed him to account for the myriad tensions between the center of a society and that society's peripheries. It also allowed him to assimilate a great deal of conventional sociology, such as studies of occupational prestige, into the basic picture. The vicissitudes of the center and of the charisma of various transcendent objects were means of accounting for such things as the decline of deference and the long historical rise of egalitarianism, which could be seen as a transfer of charisma to the individual.

This evolving image of society went far beyond notions of consensus, and Shils in his retrospective writing dissociated himself from the emphasis on consensus, noting that in the 1930s he had been skeptical of the idea, accepted by many of his contemporaries impressed by the apparent successes of communism and fascism, that modern societies needed an integrating ideology. But he was also responding to events: he considered the period of the 1950s in the United States to have involved a high level of societal consensus. His experience of the conflict over the war in Vietnam and the vociferous response to Nixon led him to conclude that this consensus was broken.

Shils continually struggled with the question of what and how the center, collectivity, and society should be understood. He inherited from Knight and Weber a strong skepticism about collective concepts. But he was equally skeptical about the reduction to individual rationality, which he regarded as Hobbism, as is clear from his book *Tradition* (1981). We-ness was, he thought, a genuine phenomenon, but also a mysterious one. He also believed in the existence of a realm of "symbolic constructions which depict the ultimate nature of things" (1997a: p. 202) toward which intellectual life was oriented. Reconciling these commitments was not only difficult: perhaps it was impossible. But as a guide to the conflicting commitments characteristic of social life itself, it enabled Shils to identify the grounds for such apparently anomalous phenomenon as the hatred of intellectuals for their own societies and the attractions of ideology.

Shils saw the rise of ideological politics and the rise of a category of intellectuals free from dependence on patronage as closely linked. But he saw that the ideological impulse itself arose from among the same sources as social consensus, namely, an orientation to sacred central things – and in particular, in the West, to the Judaic–Christian tradition itself, with its strong appeal to the notion of justice and its correlated sense of the injustice of the earthly order. This common heritage facilitated Millenarianism and often a Manichaenism that provided a focus for a sense of grievance and injustice. Intellectuals, by virtue of their own traditions and the fact that they were not responsible for the practical realities of decision, were free to indulge in criticism of the established order in light of these grievances, and to construct intellectual justifications for their grievances.

In *Ideology and Civility* (1958), Shils, paralleling such related thinkers as Michael Oakeshott, contrasted ideological politics to the mundane politics of civility. The latter was a politics limited by the sense of the possible, by a respect for the different opinions of others, by the recognition that the achievement of one good is normally linked to negative consequences for other goods and that virtue is not realized solely in one form, and by a respect for the limits of authority and power. But the balances of this kind of politics were difficult to maintain, especially under the pressure of ideological movements that did not respect these balances. In particular, the intrinsic conflicts between the collective and individualistic forms of liberalism are difficult to balance.

Shils also wrote extensively on issues that grew out of the atomic movement, on science, and on cultural issues, notably the issues of mass culture that excited commenters on the Left during the postwar period. These writings were responses, sometimes polemical, to particular intrusions of ideology into scholarship or social science, or literary criticism. The period from the end of World War II to the early 1960s during which most of these writings were published was particularly fraught, with former communists and liberals responding to the hard Left on the issues of the day, very often in the form of literary and cultural rather than overtly political arguments.

Shils had, from the 1930s, an interest in Left-wing movements, had discussions with communists at the University of Chicago, and was well-informed about ideological matters. He was not tempted himself, as many of his contemporaries were, and was critical from the beginning of the naiveté of Mannheim about the implementation of social 'plans,' a skepticism he explicitly linked to Frank Knight. After the war, he was acutely sensitive to the influence of the Stalinist Left over intellectuals. At the same time, his involvement in the Atomic Scientists movement, and earlier in peace movements, placed him far from the 'right.' He appreciated the saying that the only serious politics was conservative liberalism and liberal conservatism. His political views were similar to those of his friend, the sociologist Raymond Aron.

The experience of the student movement in the late 1960s and the early 1970s and the debates over the war in Vietnam as well as the divisive politics associated with the Nixon Presidency altered Shils views and his intellectual direction. He sensed that the consensual character of American society of the 1950s had permanently changed. While insisting that there be no toleration for the disruption of the university, he, nevertheless, advocated, in private advice to the President of the University, Edward Levi, lenience toward students involved in disciplinary proceedings related to the protests over the war. But he also took the view that the academic ethic itself was under grave threat. He published originally in Minerva a translation of Weber's writings on the subject, "Max Weber on Universities: The Power of the State and the Dignity of the Academic Calling in Imperial Germany" (1974), and wrote, as a member of the International Council on the Future of the University, *The Academic Ethic* (1974) as an attempt to define this ethic and identify the challenges to it.

The bulk of his later writing, after 1976, concerned the university, intellectuals, and memoirs of the great scholars whom he had known. He also edited volumes on Cambridge women academics and on the scholars of The University of Chicago, in an attempt to record the personal side of these communities and describe the character of great scholars who had fulfilled the academic ethic. His own recollections of individual scholars form an important part of the historical record of the intellectual life of the twentieth century, and are an important contribution to the genre. Three posthumous collections of his later writings capture well the intellectual preoccupations of this remarkable person and scholar for the

period after the earlier publication of those three volumes of his collected papers: *The Intellectuals and the Powers* (1972), *Center and Periphery* (1975), and *The Calling of Sociology* (1980). To those volumes of his collected papers should be added *Tradition* (1983). The three posthumous collections are *The Calling of Education* (1997a), which brings together Shils' reflections on the academic ethic, the character of the modern university, and his important Jefferson Lectures of 1979; *The Virtue of Civility* (1997b), which contains his later reflections on the constitution of modern society; and *A Fragment of a Sociological Autobiography* (2006), which revolves around Shils' reflections on the path of his intellectual preoccupations throughout his life.

See also: Aron, Raymond (1905–83); Authoritarian Personality; Mannheim, Karl (1893–1947); Military Sociology; Parsons, Talcott (1902–79); Social Scientists as Experts and Public Intellectuals; Weber, Max (1864–1920).

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper and Row, New York.

Mannheim, K., 1936. Ideology and Utopia: An Introduction to the Sociology of Knowledge (E.A. Shils, L. Wirth, Trans.). Kegan Paul, Trench, Trubner & Co., London. Harcourt, Brace & Co., New York.

Mannheim, K., 1940. Man and Society in an Age of Reconstruction: Studies in Modern Social Structure (E. Shils, Trans.). Kegan Paul, Trench, Trubner & Co., London. Harcourt, Brace & Co., New York.

Nygren, A., 1953. Agape and Eros (P.S. Watson, Trans.). SPCK, London (first published 2 vols. 1930, 1936).

Otto, R., Harvey, J.W., 1923. The Idea of the Holy. Oxford University Press, Oxford (first published 1917).

Park, R.E., Burgess, E.W., 1921. Introduction to the Science of Sociology. The University of Chicago Press, Chicago.

Parsons, T., Shils, E., 1951. Toward a General Theory of Action. Harvard University Press, Cambridge.

Parson, T., Shils, E., Naegele, K.D., Pitts, J.R. (Eds.), 1961. The Theories of Society: Foundations of Modern Sociological Theory, vols. 1–2. The Free Press of Glencoe, Inc., New York.

Shils, E.A., Janowitz, Morris, 1948. Cohesion and disintegration in the Wehrmacht in World War II. Public Opinion Quarterly 12 (2), 280–315.

Shils, E.A., 1950. Primary groups in the American army. In: Merton, R., Lazarsfeld, P.F. (Eds.), Continuities of Social Research: Studies in the Scope and Method of "The American Soldier". The Free Press, Glencoe, IL, pp. 19–39.

Shils, E.A., 1954. Authoritarianism 'Right' and 'Left'. In: Christie, R., Jahoda, M. (Eds.), Studies in the Scope and Method of 'The Authoritarian Personality'. Free Press, Glencoe, IL, pp. 24–49.

Shils, E.A., 1956. The Torment of Secrecy: The Background and Consequences of American Security Policies. The Free Press, Glencoe, IL.

Shils, E.A., 1957. Primordial, personal, sacred, and civil ties: some particular observations on the relationship of sociological research and theory. British Journal of Sociology 8 (2), 130–145.

Shils, E.A., 1958. Ideology and civility: on the politics of the intellectual. Sewanee Review 66 (3), 450–480.

Shils, E.A., 1961a. The Intellectual Between Tradition and Modernity: The Indian Situation. Mouton, The Hague.

Shils, E.A., 1961b. The calling of sociology. In: Shils, E.A., Parsons, T., Naegele, K.D., Pitts, J.R. (Eds.), Theories of Society: Foundations of Modern Sociological Theory. The Free Press of Glencoe, New York, pp. 1405–1448.

Shils, E.A., 1972. The Constitution of Society. The University of Chicago Press, Chicago.

Shils, E.A., 1975. Center and Periphery: Essays in Macrosociology. University of Chicago Press, Chicago.

Shils, E.A., 1980. Center and Periphery: Essays in Macrosociology. The University of Chicago Press, Chicago.

Shils, E.A., 1981. Tradition. The University of Chicago Press, Chicago.

Shils, E.A., 1984. The Academic Ethic: The Report of a Study Group of the International Council on the Future of the University. The University of Chicago Press, Chicago.

Shils, E.A., 1997a. The academic ethic. In: Grosby, S. (Ed.), The Calling of Education. The University of Chicago Press, Chicago, IL.

Shils, E.A., 1997b. The virtue of civility. In: Grosby, S. (Ed.), The Virtue of Civility: Selected Essays on Liberalism, Tradition, and Civil Society. Liberty Fund, Indianapolis, IN.

Shils, E.A., 2006. In: Grosby, S. (Ed.), A Fragment of a Sociological Autobiography: The History of My Pursuit of a Few Ideas. Transaction Press, New Brunswick, NJ.

Weber, M., 1927. General Economic History (F. Knight, Trans.). George Allen & Unwin, London.

Weber, M., 1949. The Methodology of the Social Sciences. In: Shils, E., Finch, H.A. (Trans. and Eds.). The Free Press, Glencoe, IL.

Weber, M., 1954. Max Weber on Law in Economy and Society. In: Rheinstein, M., Shils, E.A. (Eds. and Trans.). Harvard University Press, Cambridge, MA.

Weber, M., 1974. Max Weber on Universities: The Power of the State and the Dignity of the Academic Calling in Imperial Germany. In: Shils, E.A. (Ed. and Trans.). The University of Chicago Press, Chicago.

Relevant Websites

http://www.anb.org/articles/14/14-01145.html – American National Bibliography Online: Shils, Edward Albert.

digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1058…slisfrp/59 – Bibliography of the Published Works of Professor Edward Shils.

http://www.independent.co.uk/news/people/obituaries-professor-edward-shils-1569959.html – The Independent Obituary: Edward Shils.

http://www.newcriterion.com/articles.cfm/In-memoriam–Edward-Shils–1910-1995-4233 – The New Criterion in Memoriam Edward Shils.

http://www.commentarymagazine.com/article/the-intellectuals-and-the-powers-and-other-essays-by-edward-shils/ – Review *The Intellectuals and the Powers and Other Essays*.

http://chronicle.uchicago.edu/950202/shils.shtml – The University of Chicago Chronicle Obituary: Edward Shils.

http://en.wikipedia.org/wiki/Edward_Shils – Wikipedia Category: Edward Shils.

Shklar, Judith N (1928–92)

Andreas Hess, School of Sociology, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Judith Shklar's political theory defies traditional labels. It emerged, as one observer has pointed out, between the 'end of ideology' discussions of the 1950s and the early 1990s discussion of the 'end of history.' Shklar's writings contributed significantly to modernizing political thought by arguing for a new, more skeptical, and stripped-down version of liberalism that brings political theory and real-life experiences closer together.

The Formative Years (1928–50)

Judith Shklar was born Yudita Nisse in 1928 in the Latvian capital of Riga. Her parents belonged to the wealthy Jewish elite of the city. They were not religious, which meant that for the children a modern French-style education was favored, including the teaching of classical and modern languages, literature, and the sciences. Additionally, private education continued at home.

By 1939 Riga began to look more and more like a trap. The Nisse family escaped to Sweden in a lucky moment just before the Hitler–Stalin pact, which led first to the Russian occupation of Latvia, followed by the German invasion of Latvia a few months later. However, Sweden did not look too safe either after the Nazi's occupation of neighboring Norway. The Nisse family realized that in case of a Nazi invasion they would not be safe from persecution, arrest, or deportation. Equipped with false papers they escaped via the Soviet Union and Japan to North America. On arrival in Seattle, the family was first detained as illegal aliens but eventually discovered by a Rabbi who had inspected the camp. After their release, the family finally settled in Montreal.

Judith perfected her English reading and writing skills while attending a Protestant School. Later, at McGill University she encountered a gifted teacher, Frederick Watkins, an American Rousseau specialist who was to become her supervisor. Watkins was so impressed with the intellectual abilities of his student that he recommended her to Harvard. In 1950, together with her newly wed partner Gerald Shklar, Judith moved to Cambridge, Massachusetts near Boston.

In the Aftermath of Totalitarianism (1950–64)

At Harvard, the German-American scholar Carl Joachim Friedrich (1901–84) became Shklar's supervisor and mentor. Friedrich, who taught comparative politics, showed Shklar the ropes; he gave her space and the opportunity to develop her own teaching and research profile.

Shklar found herself falling between several stools. She was too young to belong to the group of German émigré scholars that had begun to take hold in institutions of higher learning; at the same time she did not fully belong to the established Harvard crowd either. She discovered that she was a particular kind of refugee and therefore somewhat of an outsider – an insight that stayed with her for the rest of her life and that would have a decisive impact in the development of her political theory. As an 'exile from exile' Shklar would remain a lifelong skeptic. She consistently questioned the repository of wisdom of the Western humanist canon and its underlying sense of human progress and improvement. Having faced the totalitarian threat herself she also had trouble understanding the common intellectual treatment of Fascism and the Western liberals' tolerance for Stalin.

Shklar's interests leaned to the study of modern ideologies rather than to a formally oriented comparative politics in the way that her supervisor Friedrich promoted it. For her thesis she opted for a topic that sounded very Hannah Arendt-like in theme and scope. The original intention had been to discuss the Holocaust and the limits of understanding and explaining the incomprehensible and the constraints this imposes for the development of modern political theory. For her thesis, entitled 'Fate and Futility: Two Themes in Contemporary Political Theory' (unpublished PhD, Radcliffe, 1955), Shklar was awarded Harvard's Radcliffe Prize and the Birkhead Prize of the American Political Science Association.

Shklar's first book, *After Utopia*, was a revised version of the PhD dissertation (Shklar, 1957). The book's stated aim was to assess the state of contemporary political theory and political philosophy after the extreme experiences of totalitarianism and the Second World War. Like Arendt, Shklar observed a theoretical *stasis*; yet in contrast to Arendt's melancholia for the classics, Shklar argued for a fresh start and an altogether more rational attempt at political theoretical conceptualization. She saw a need for more realism in political theory at a time when an old-fashioned liberalism had "become unsure of its moral basis, as well as increasingly defensive and conservative" (Shklar, 1957: p. viii). Her critical remarks were not only directed at Cold War liberals but also directed at Friedrich and Arendt who continued to repeat the mantra of totalitarianism almost *ad nauseam* but who seemed unable to explain what liberal democracy and political theory could actively do in order to become more attractive and convincing, particularly in light of the ideological competition.

In 1963 the question of tenure came up. After years of teaching and after some promising publications, documented through a prize-winning dissertation, a book with Princeton University Press, a second book soon to be published with Harvard University Press, and being the recipient of a Guggenheim fellowship Shklar could simply not understand why the Department was leaving her dangling. In the end Shklar, who was financially independent, decided to ask the university

whether the process could be short-cut by offering her a permanent half-time lectureship instead. The University agreed and in 1963 Shklar was finally offered employment as a lecturer, a peculiar arrangement that would last as late as 1970.

Shklar's second book, *Legalism*, appeared in 1964. Between the so-called Hart–Fuller debate of the late 1950s, which discussed the role of legal positivism in the context of the Nuremberg Trials, and the discussion surrounding Arendt's *Eichmann in Jerusalem*, it seemed well placed. An analysis of the trials provided an opportunity to discuss the relationship between law, morality, and politics. More specifically, Shklar argued that formalized justice should not be regarded as an end in itself but also that just procedures cannot create a democratic order on its own. This had been the illusion of the judges in Nuremberg: if only the laws were seen to be operating, the impression of functioning democratic institutions and politics would follow from that. Against such assumptions Shklar argued that it was a legalistic illusion to think that in the end the example of the Nuremberg Trials could actually prescribe some future international criminal law and court system: "The trial was a tribute both to the intellectual limitations of legalism as an ideology and to the real political value which legalism has in practice" (p. 147).

Shklar points out that it was true that the trials took place in a legal vacuum. There was no properly functioning system of international criminal law. In this context it was a sense of justice that turned out to be the main driving force behind the trials. This was very much against the notion of positive law that maintains the distinction between legal procedures and political aims and values such as justice. What has happened then was that the new crimes committed and the trials conducted in response to barbaric deeds had led to an interpretation that tried to put together again what had in reality become separated. Put differently, after the experience of totalitarianism, it became harder to suggest that legal procedures, justice, and politics were not linked at all.

The Making of a Political Theorist (1964–86)

Shklar's next book, *Men and Citizens*, was on Rousseau (Shklar, 1969). The study of the Genevan philosopher was not that far removed from her earlier concerns and epistemological interests as it might appear on first sight. Engaging with Rousseau could not only lead to a critique of classic liberalism, but *ex negativo* his arguments touched on a number of themes that, according to Shklar, went to the very heart of a critical theory of liberalism – something that students who saw Rousseau only as an intellectual precursor to Marx seemed to have forgotten.

Shklar observed that for Rousseau utopia served foremost as a critique of both religious notions and contemporary society; utopias were practical orientation points: they educate the senses and guide judgment. At the same time Rousseau knew that a simple return to nature or past utopias was impossible. Once mankind had embraced the division of labor, including developed forms of property structures and inequalities, there was no turning back. Utopian idylls like Sparta and the tranquil rural Swiss home only confirmed that mankind was on a restless search, driven to find a home to rest, while actually being unable to encounter one in reality.

According to Shklar, Rousseau had learned his psychology, his 'history of the human heart,' mainly from Locke. The genetic method freed him from having to deal with actualities or real historical facts. The method consisted of the description of a civilizational drama, which has to be understood as a kind of repeated *ur*-condition, acted out time and time again. However, while man was responsible for maintaining and prolonging the situation, he had arrived at this situation innocently. The societal chains were there first: man, although born equal, had to deal with inequality and constraints from the very beginnings of the civilizing process.

In order to fully reveal his human potential, for Rousseau certain 'psychological' capacities in man have to exist: the ability to make choices, the use of imagination, memory, the capacity of self-perfection, and the ability to feel pity at the sight of human suffering. The human mind appeared thus to be in a permanent struggle: on one side we find these positive psychological capacities and on the other side the social and identifiable constraints – the tension that results in human suffering. Only deep introspection and insightful self-education, including searching for one's freedom by physical travel, provided an escape.

The final part of Shklar's study was devoted to Rousseau's idiosyncratic use of language and style. It was the French philosopher's 'organizing images' that explained in large part his popularity and success. Rousseau had the advantage of having been a thinker who, unlike many modern academic theorists, was never confined to narrow specialisms. His language was thus far removed from abstraction without ever considering academic jargon.

Shklar's *Men and Citizens* contained a scarcely veiled self-critique. In *After Utopia* Shklar had pronounced modern political theory to be in deep crisis. With Rousseau, Shklar showed that political theory, rightly conceived, might provide a way out. It connected the past to the present and, if studied and practised carefully, could reveal a 'surplus meaning.' Shklar had demonstrated that Rousseau had discussed all the ingredients that modern political theory was concerned with, in particular the theme of justice and the concomitant need to avoid extreme inequality. His philosophy could be used to show the continuous relevance of modern republican ideas and themes such as the formative experiences of individuals as expressed in virtues and vices, ideas that remained also important to liberal thought. Furthermore, an engagement with Rousseau's philosophy could help us to think about the appropriate body politic and those governmental structures that would guarantee freedom from fear and cruelty.

As with Rousseau, the preoccupation with Hegel derived from preparing and teaching a core course on the Enlightenment at Harvard. *Freedom and Independence* (Shklar, 1976) was a critical analysis of Hegel's phenomenology. For Shklar, Hegel merged not only aspects of passion and reason, he was also the first philosopher to consider the idea that the progress of human thought could be explained by systematically studying social interaction. The analysis of conflict between the individual and his social and natural environment, and how thought emerged from that confrontation, remains his greatest contribution. What emerges is, according to Shklar, a common concern for how individual and social environment are linked and how they interacted. As to his speculations about how to

solve the problem, however, Shklar was less sure. Hegel had little time for Rousseau's educational suggestions, while Rousseau would certainly have objected to the achievements and advantages of *Bildung* as described by Hegel. Hegel's reservations *vis-à-vis* Montesquieu were of a different kind: Hegel accepted the latter's distinctions between the ancients and the moderns; however, the real distinction between the two, according to Shklar, was that Montesquieu clearly favored reform and the proper functioning of political institutions based on honor and decent behavior while Hegel favored a change of mind and morals, a different 'spirit.'

While Shklar remained unimpressed by the Prussian's attempt to explain the intellectual foundations on which modern civil society could be built, there were still some important insights to be gained from interpreting the text in a different way by reading it as a kind of political psychology. Most critics seem to have agreed with Shklar. George Kateb praised and singled out Shklar's interpretation of Hegel's moral concerns and the weight she gave to the philosopher's discussion of cultural, political, and social ingredients that are necessary to establish a moral community (Kateb, 1987). Equally the Hegel expert Z.A. Pelczynski stressed Shklar's dissection of the political-cultural dimensions, such as the moral costume and the critical self-reflection, necessary to maintain a modern political system (Pelczynski, 1977).

After having been employed as a lecturer for many years, Shklar's status, like that of a dozen other women at Harvard came under scrutiny. A critical internal report had been filed, which was soon followed by swift action by the university; it led finally to the promotion of those female Harvard members of staff who had been denied tenure until then – Judith Shklar being one of them. More recognition followed. Shklar was elected to President of the American Society for Political and Legal Philosophy in 1982, and in 1983 she became Vice-President of the American Political Science Association. The same year Shklar spend half a year as Visiting Fellow at All Souls in Oxford. This was followed by a year as Pitt Professor of American History and Institutions at Cambridge University.

Shklar's own experience of the threat of totalitarianism and fear had made it difficult for her to continue pursuing political theory in a business-as-usual fashion; for her the time for great systems, ideas, and theorems was over. What was needed was to reflect about problems in a way that would pay its dues to the extreme experiences of the century but at the same time acknowledge that the world had not just suddenly come to a halt. Shklar was most interested in those formative processes that used to be the theme of classical republican thought. However, she also realized that what was needed was to re-think the old themes of vices and virtues in the context of modern liberal democracy. Shklar's declared aim was to bring political theory closer to the social and political realities of modern society by discussing modern vices not just in the political context but in the social realm as well. This was not easy, particularly since the boundaries between the spheres always shifted.

In *Ordinary Vices*, Shklar argued that in contrast to other liberal conceptions who aim at a *summum bonum*, what she termed the modern and specifically American imprinted 'liberalism of fear,' refers only to a *summum malum* – cruelty (and, in response to cruelty, how to avoid suffering) (Shklar, 1984: p. 7ff). The other vices (hypocrisy, snobbery, arrogance, betrayal, and misanthropy) need to be ranked in relation to the first, cruelty. Shklar further maintained that there were many aspects to vices and to conceive them simply in either simple affirmative or negative terms would not do justice to the complexity of the problem. Hypocrisy, for instance, may, generally speaking, not be well regarded. Yet it allows people to pretend to be better than they are and it would therefore be highly problematic for a liberal democracy to take public measures to abolish it. A similar argument could be made in the case of snobbery, which is hardly a crime, let alone always necessarily an antidemocratic attitude. Treason and betrayal seem to be more problematic; yet again, no one-size-fits-all solution can serve as a response to the complexity of these two vices. What one person sees as treason to the fatherland, another person might see as a positive contribution or even regard as the truly sane response to cruelty (resistance to totalitarian regimes comes to mind here immediately). Adultery, another example of betrayal, can occur between married or unmarried partners and often does destroy confidence. However, a private breach of trust does not necessarily mean that trusting relationships with other people are impossible. Even misanthropy does not always have to have negative consequences. It may merely be a personal protest against the brutal way the world sometimes works. It might even have some positive effects as Shklar illustrates by referring to examples from Shakespeare to Swift.

Looking at the various vices nothing justified putting cruelty first more than the experiences of the twentieth century. As a consequence "liberal democracy becomes more a recipe for survival than a project for the perfectibility of mankind" (p. 4). In other words, liberalism should not be seen as an "amoral free-for-all" but rather as a defense of a delicate achievement in which we continue to "endure contradiction, complexity, diversity, and the risks of freedom" (p. 5).

As Patrick Riley, a former student of Shklar, noted, *Ordinary Vices* was "the *summa* of Shklarism," "a fusion of psychological confidence and moral scepticism" (Riley in Memorial Brochure, 1992: p. 102). In an earlier review, Riley had already commended Shklar for having written a "brilliant, psychologically insightful, and deeply humane book … In it Montaigne and Hegel meet to wonderful effect: Montaigne supplies the 'ordinary vices'… and Hegel brings the method of illumination" (Riley, 1985: p. 610).

Ordinary Vices had been inspired by Montaigne, but Montesquieu was never far off Shklar's intellectual radar screen. In a monograph for Oxford University Press's series 'Past Masters' she took the opportunity to put forward her interpretation of Montesquieu, particularly her view that the French philosopher did not stress rights so much than a mental predisposition toward an early form of a "liberalism of fear" (p. 89). For Shklar, Montesquieu's early liberalism of fear is the attempt to avoid above all cruelty, unnecessary suffering and brutal punishment. The French thinker favored a rather minimalist conceptualization free of utopian dimensions; its aim and major purpose was to achieve a decent society marked by the absence of fear.

Shklar also discussed the way Montesquieu was received in the newly emerging American republic. She argued that if he had lived long enough the French Enlightenment philosopher

would have been pleasantly surprised at the echo that *The Spirit of the Laws* received among the American Founding Fathers, particularly in the context of the debates surrounding the United States Constitution. Shklar describes Montesquieu as a mediator not only between Europe and America but also as a "bridge between a traditional and a modern idea of constitutional government" (p. 111). For Americans it did not matter from where Montesquieu came and in which context he wrote. In a very different environment and social and political context, his new theory about the separation of powers could still seem to make perfect sense, even though it was a misreading of the actual English constitution.

Shklar's study of Montesquieu was welcomed, especially by scholars who had long complained that the American-French dimension had often been neglected when studying the republican tradition. Shklar had rightly pointed out "that Pocock's reconstruction of the 'Atlantic republican tradition' had given us *Hamlet* without the Prince by somehow forgetting France" (Wright, 2007: p. 152). This was obviously something that Shklar's Montesquieu interpretation tried to correct.

Shklar also devoted a number of essays to political theory's methodological problems. 'Learning without Knowing,' for example, was the attempt to tackle two problems, that of 'surplus meaning' and that of the type of knowledge that the humanities and social sciences produce (Shklar, 1998a). When old ideas are interpreted in changing contexts, particularly in times of reform and democratization, this poses a challenge to which history and the history of ideas have to respond. According to Shklar, the sociologization of the history of ideas ran into the risk of paying too much attention to quantitative detail. She warned that specialization could narrow the mind (Shklar, 1998a). At the same time one should not go to the other extreme by conflating values and intentions of the more radical kind with good history writing. Detachment from politics was part of the ethics of responsibility as Weber, whom Shklar referred to, knew only too well (p. 126). Shklar conceded that Weber's ideal was hard to achieve. Particularly on-science practitioners will never gain the same level of security of knowing and knowledge (p. 126). The demand for producing useful knowledge in the humanities and social sciences was therefore loaded with risks and ambiguities. In such a situation Shklar recommended organized skepticism and self-control against prophets and gurus (p. 128).

In another essay Shklar voiced her concern about the interpretative turn in the social sciences (Shklar, 1998b). If intellectual influences, intentions of the writer, and specific historical constellations had no bearing on the actual texts in question, as, for example, Ricoeur has argued, the texts itself would be subject to a rather impoverished reading. Not taking intentions, desires, and wishes of actors involved into account would render not only social action meaningless for the participants, but for the more professional observers of social action (such as social anthropologists, historians, and sociologists) the object of their study would simply disappear. To see the reading of the text as something that is identical with human action blurs all boundaries of the different forms or appearances of existence, i.e., any social action could simply be regarded as 'reading.' Shklar suspected that the real motive behind such argumentation is not only to establish hermeneutics as a leading epistemological instrument in the modern social science and arts, but also to defend the intellectual hegemony and almost absolute position of the interpreter. The thus 'secularized,' revamped idea of the hermeneutic circle serves no-one but the new absolute position of the interpreters themselves.

Transatlantic Encounters of a Different Kind (1987–90)

While Shklar's trilogy on Rousseau, Hegel, and Montesquieu can be read as an attempt to make sense of modern European conceptualizations of liberty and its 'psychological' preconditions, a preoccupation with American political theory and the American intellectual tradition meant obviously not just touching on questions of liberty under different social and political circumstances but also taking transatlantic communication and exchanges into account. Shklar's 'optimal marginality' was one of the reasons why she remained skeptical of answers that seemed to her to be either too one-dimensional or too complacent, particularly when they appeared as grand narratives, be it in the form of American exceptionalism or in the form of a rather arrogant 'whig' history of American liberalism.

Many of Shklar's American writings were directly or indirectly an engagement with Hannah Arendt. Shklar admired Arendt but Shklar also felt that something was missing in her work. She often agreed with what Arendt had to say, the unprecedented social and political conditions under which the American Republic was first conceived, the success the Founding Fathers had in absorbing the revolutionary spirit and energy of the people and in preserving the country's political institutions, and the relative success and achievements of the American Revolution when compared to the problems that resulted from the French Revolution. But Shklar also identifies what was completely absent in Arendt's account: that the American Republic had to fight a Civil War in order to secure the earlier promises of the Constitution, particularly in relation to the question of citizenship. Arendt also appeared weak when it came to contemporary issues. According to Shklar, she had little knowledge of post-Civil War history, Southern politics, race relations, and the more contemporary constitutional issues, which dealt with citizenship.

Shklar also differed from Arendt in her particular interpretation of the novel environment that the Americans found themselves in. Using an expression first coined by Emerson, Shklar distinguished between the so-called 'party of hope' and 'the party of memory.' The former position was represented by Tom Paine and Thomas Jefferson and referred to the rejection of and radical rupture with almost everything that related to European history (p. 96ff); Paine and Jefferson not only insisted firmly on living in the present, but their attitude also allowed them to nurture bright hopes for the future. The latter position was represented by John Adams and to a certain extent by James Madison. Adams and Madison suggested taking a closer look at the past in order to find out exactly what had gone wrong with classic European republicanism. This view was much more pessimistic in its outlook than Jefferson's and Paine's; in essence, it was what led this faction of the Founding Fathers to construct certain built-in security mechanisms within

the Constitution. As Shklar pointed out, both views, along with the public discussion of them, shaped what Lipset has called the First New Nation. There was thus not one monolithic way to come to terms with a new beginning. Shklar's interpretation threw a more sophisticated light on the beginnings of the republic than Arendt's did.

In 1990, at the annual meeting of the American Political Science Association in San Francisco, Shklar delivered a groundbreaking presidential speech entitled 'Redeeming American Political Theory' in which she tried to situate the emergence and history of American political ideas in the context of real power struggles (Shklar [1991]/1998). She insisted that political ideas and reflections were not limited to academic or scientific discourse but could also be encountered outside of academia, most prominently perhaps in the form of literature and in popular perceptions and interpretations. In the context of her discussion of the specific *Americaness* of political ideas she referred especially to Emerson, Hawthorne, or Adams to make a case against a rather narrow and limited discourse of both political science and political theory. The main thrust of her speech, however, was directed against the taken-for-granted assumptions of American exceptionalism and liberalism, and that of the relevance of political theory for political science. According to Shklar, liberal exceptionalism was unable to deal adequately with the darker side of American history. The alternative that Shklar presented was not only much more plausible it also showed that political theory had practical and moral implications and at times, as in the case of Emerson, broad educational impact.

Shklar had always been unhappy with a political theory that did not pay attention to specific historical contexts and constellations out of which political ideas emerged. At the same time she was aware that to abstract and universalize from specific historical situations would be equally problematic. A masterful example of a much more balanced discussion is her essay 'Positive Liberty, Negative Liberty in the United States' (Shklar, 1998c). Against Berlin's notion of negative liberty, Shklar stresses that at least in America rights emerged as means that would allow people "to realize goals *against* others" (p. 111f, emphasis in the original). In America, insists Shklar, the fight between master and slave was not just a Hegelian fantasy but real. For her, the fight against slavery was a fight for "freedom itself," it was "a way of political life" (p. 112). The claiming of rights was an inseparable part of that struggle. As Shklar showed, the idea of rights was much more important to the history of American liberalism than the Europeans' liberal distinction between positive and negative liberty (Shklar, 1998c).

In *Faces of Injustice* (1990), which followed on from her deliberation of the limits of negative liberty, Shklar suggested a change of perspective: injustice might be better understood as something that is not just the negative counterpart to justice. Injustice must be studied as a phenomenon in its own right, something that needed to be conceptualized in its own peculiar dimension. This needed to be done independently of 'grand theory' constructions in search of a *summum bonum*. With her critique Shklar aimed at a defense of a barebones liberalism which remained compassionate while at the same time remaining antisystemic in its conceptualization.

Shklar argued that in contrast to advocating distributive models there existed always a skeptical political theory tradition, which provided guidance and contained critical reflections in relation to injustice. For her, the skeptical tradition had the advantage that it regards injustice not just as an absence of justice or just procedures but as occurring "continuously within the framework of an established polity with an operative system of law, in normal times" (p. 19). The skeptics knew this – hence their interest in discussing concrete cases and situations. In a way skeptical philosophers like Plato, Augustine, and Montaigne were first and foremost exemplary story tellers, not system builders.

Shklar concluded that to give injustice its due demands a different perspective and different type of narrative. A good start would be to identify, recognize, and listen to victims of injustice. In effect, such a new critical approach could tell us more about the many faces of injustice rather than following the false hope of striving for an ever more perfect state of justice, including the idea of a perpetual amelioration of the laws.

In *American Citizenship: The Quest for Inclusion* (1991), Shklar elaborated on some of the arguments from *The Faces of Injustice* by exploring the concepts of political and social rights. She was aware that political membership remained of primary importance as it constituted the referential framework for all other aspects. She showed that to insist only on formal membership and citizenship was to neglect the rich connotations of citizenship. Shklar's argument suggested that political belonging (as expressed in voting) depends on earning. After all, one had to be a part of a political community first before the standing and status that derives from being able to provide for oneself can be acknowledged. However, this did not necessarily imply that the social world and earning one's living have merely derivative functions. In contrast, Shklar suggested that there always existed a strong relationship between the two. With reference to American history, Shklar showed that citizenship would be an empty shell if deprived of these concrete social connotations.

Reflections on Obligation and Exile (1991–92)

Toward the end of her life, Shklar had come to realize that the theme of exile and emigration had the potential of throwing an entire new light on the history of obligation and loyalty. To be sure, Shklar had always felt that the story of exile sat uncomfortably with the ideal of belonging. In a discussion of Michael Walzer's *Interpretation and Social Criticism* she had criticized Walzer's model interpreter whose criticism supposedly worked differently because in contrast to the 'discoverer' or 'inventor' type who either founded a religion or a paradigm (the prophet or discoverer), or functioned like a neutral and somewhat distant 'constitutional legislator' (the inventor), the interpreter belonged to a community, could work from within, and was therefore 'less manipulative' (Shklar, 1998: p. 378). For Shklar, Walzer's argument was too much group- or community-centred; it gave too much credit and had too much confidence in the shared understanding and values that underlie community life. While Shklar did not exclude the possibility that such communities could exist, for her such

a presupposition did not sit easily with the experiences of conflict, or the pluralism of interests in modern representative democracies, nor did it give enough weight to individuals and their right to hold opinions *as individuals*.

For Shklar the perspective of exile threw a different light on obligations. Exile, so she noted, was a fundamental human experience that had been dealt with extensively in history and literature. However, surprisingly little had been said about it in terms of political theory – something that Shklar attempted to remedy (Shklar, 1998: p. 57). The exile perspective allowed her to explore and discuss various aspects and themes of political thought – conditions for submission to rules, political obligation, etc. – from a different angle. It also implied a form of extreme decision making, often engaging "the entire personality" (p. 59).

According to Shklar, the fate of modern exiles showed some distinct features. As Shklar points out, Athens and Rome, at least in their early democratic and republican periods, emphasized the public character and just procedures of their polities and demanded both obligation and loyalty to them. In contrast, many modern states have failed by turning state apparatuses into organizations where "governmental illegality" violated and disrupted civil society and the political process to the effect that "political loyalty may survive but not obligation to obey the law" (p. 48). Shklar stressed that exiles, both those who left but also those who went into 'internal exile' and tried to maintain their clear private conscience, had absolutely no obligations to such a state, country, or government. The difference between classical and modern exiles was predominantly marked not by the effort to avoid civil war but was based on the individual's critical judgment and the conscientious rejection to follow what is demanded of him or her. To put it differently, massive exit virtually tested the legitimacy of the modern state (p. 59).

Shklar was painfully aware that today's emigrants and exiled face even more challenging conditions. In many instances there is no host country to offer asylum. There often are no foreign exile and no country to escape to. Most of today's refugees find themselves 'in pure limbo' a situation which evokes moral concern, even moral outrage. Those who are outraged and show solidarity find themselves in a solitary Thoreau-like situation, that is, neither can they join a liberating force nor is it always possible to identify fully with the many refugees due to the lack of detailed knowledge, physical distance, culture, and language etc. As Shklar rightly observes, there is no 'we' here, and, in any case, what exactly could loyalty and obligation refer to, or what kind of notions could they entail in such a situation?

For Shklar exclusionary practice is often the mother of loyalties of often dubious quality and must sometimes be worrying. One conclusion that can be drawn from the modern experience is that cultural and national cohesion remain overrated ideas, eventually causing and even prolonging the conditions in which old and new injustices thrive. Shklar is aware that citizenship might not be the solution to all problems but it still remains a first important step toward preventing injustice, for "injustice not only cancels obligations and undermines loyalties, however resilient the alternatives many seem, it also engenders the conflicts between obligation and the affective ties that bind us" (p. 55).

Shklar's Legacy

Judith Shklar died of a heart attack in 1992. In a commemorative service Stanley Hoffmann, Shklar's best friend at Harvard, pointed out that they both were "survivors from the European catastrophe" (Hoffmann, 1992, in Memorial Brochure: p. 13). Both had become "liberals without illusions" (Hoffmann, 1992) who abhorred the deadly consequences of ideologies but who also valued political theory traditions that served as a bulwark against cruelty and potential atrocity. Above all, Shklar was distinguished by her curiosity and honesty, something that also shone through in her teaching and writing. Hoffmann stressed how important it was to see the connection between Shklar's "experience as a refugee as well as her later life as an immigrant in America and her outlook on power" (Hoffmann, 1993: p. 174). There exist two political science approaches: "those who study power because they like to exert it and those who study it because they fear it – those who would like to ride the horse of power and those who are scared of being trampled by it" (Hoffmann, 1993). Shklar clearly belonged to the latter category. For Hoffmann, Shklar's most important contribution lay in studying the qualities, the moral costume, and the political-psychological dimension of citizens. Montesquieu's observation that "character and government constantly mold each other" was at the heart of her preoccupation (p. 178).

In the same commemoration John Rawls pointed out that there were three central aspects to Shklar's work: a barebones liberalism, whose intellectual foundations go back to Montaigne and Montesquieu; an interest in moral psychology, which was less interested in individual perception than in its political dimension; and a certain form of skepticism *vis-à-vis* grand theories (Rawls in Memorial Brochure, 1992: p. 7). Dennis Thompson, a former student turned colleague, concurred with Rawls but added that while Shklar's skepticism was directed against grand theory, this was born out of common sense about grand theory's own limited historical horizon and function.

Another commentator, James Miller, saw Shklar mainly as a moral psychologist. As such she shared many concerns with fellow political theorists, particularly Isaiah Berlin, Hannah Arendt, and Michael Walzer; however, she also disagreed with them profoundly. While Shklar respected Berlin but disliked his rather snobbish attitude and his attempt to blend into the UK elite, Hannah Arendt, in turn, got lost in republican abstractions in her book *On Revolution* (a book that, as Miller rightly points out, she hated) and Michael Walzer was too communitarian for her taste. While Shklar shared some characteristics and habitus forms with fellow refugees and immigrants, she never became a joiner. She would pursue her own Montaigne-informed, pyrrhonic-style course in which her own attempts did not build up to any grand theory but instead convincingly managed to show that "we are strangers to one another and [that] we are too ignorant to judge each other" (Miller, 2000: p. 818).

Such a 'pyrrhonic' form of thinking might be explained through Shklar's own experience of exile, which developed into a habit of thought. Jonathan Allen observed that "exile left another, deeper imprint on Shklar's work." It led her "to adopt an outlook on politics characteristic of those minorities peculiarly vulnerable to exile, victimization, and the exercise of

arbitrary power." It finally turned Shklar's liberalism of fear into "a liberalism for grown-ups" (all quotes in Allen, 1998: pp. 446–550).

See also: Arendt, Hannah (1906–75); Citizenship: Political; Hegel, Georg Wilhelm Friedrich (1770–1831); Immigration, Emigration, and Citizenship; Immigration: Political Aspects; Liberalism: Political Doctrine and Impact on Social Science; Migration: Cultural Aspects; Political Science, History of; Rousseau, Jean-Jacques (1712–78); Totalitarianism.

Bibliography

Allen, J., 1998. Liberalism for grown-ups. Government and Opposition 33 (4), 544–550.

Hess, A., 2014. The Political Theory of Judith N. Shklar: Exile from Exile. Palgrave Macmillan, New York.

Hoffmann, S., 1992. Tribute. In: Memorial Tributes to Judith Nisse Shklar, 1928–1992 (brochure), p. 13.

Hoffmann, S., 1993. Judith Shklar as Political Thinker. Political Theory 211 (2), 172–180.

Kateb, G., 1987. Review of freedom and independence. The American Political Science Review 72 (1), 232–233.

Miller, J., 2000. Pyrrhonic liberalism. Political Theory 28 (2), 810–821.

Pelczynski, Z.A., 1977. Review of freedom and independence. Political Theory 5 (1), 127–130.

Rawls, J., 1992. Tribute. In: Memorial Tributes to Judith Nisse Shklar, 1928–1992 (brochure), p. 7.

Riley, P., 1985. Review of ordinary vices. The American Political Science Review 79 (2), 610–611.

Riley, P., 1992. Tribute. In: Memorial Tributes to Judith Nisse Shklar, 1928–1992 (brochure), p. 102.

Shklar, J.N., 1957. After Utopia: The Decline in Political Faith. Princeton University Press, Princeton, NJ.

Shklar, J.N., 1964. Legalism: An Essay on Law, Morals, and Politics. Harvard University Press, Cambridge, Mass.

Shklar, J.N. (Ed.), 1966. Political Theory and Ideology. Macmillan, New York.

Shklar, J.N., 1969. Men and Citizens: A Study of Rousseau's Social Theory. Cambridge University Press, Cambridge.

Shklar, J.N., 1976. Freedom and Independence: A Study of the Political Ideas of Hegel's Phenomenology. Cambridge University Press, Cambridge.

Shklar, J.N., 1984. Ordinary Vices. Harvard University Press, Cambridge, MA.

Shklar, J.N., 1987. Montesquieu. Oxford University Press, Oxford.

Shklar, J.N., 1990. Faces of Injustice. Yale University Press, New Haven.

Shklar, J.N., 1991. American Citizenship: The Quest for Inclusion. Harvard University Press, Cambridge, MA.

Shklar, J.N., 1998a. Learning without Knowing. In: Hoffmann, S. (Ed.), Political Thought and Political Thinkers. The University of Chicago Press, Chicago, pp. 105–131.

Shklar, J.N., 1998b. Squaring the hermeneutic circle. In: Hoffmann, S. (Ed.), Political Thought and Political Thinkers. The University of Chicago Press, Chicago, pp. 76–103.

Shklar, J.N., 1998c. Positive liberty, negative liberty in the United States. In: Hoffmann, S., Thompson, D.F. (Eds.), Redeeming American Political Thought. The University of Chicago Press, Chicago, pp. 111–126.

Wright, J., 2007. Montesquieuean moments: *The Spirit of the Laws* and republicanism. Proceedings of the Western Society for French History 35, 149–169.

Yack, B. (Ed.), 1994. Liberalism without Illusions – Essays on Liberal Theory and the Political Vision of Judith N. Shklar. The University of Chicago Press, Chicago.

Short- and Long-term Memory during Infancy and Early Childhood across Cultures

Patricia J Bauer, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Memory is a multidimensional construct that is evident from early in life yet undergoes a protracted course of development. This article features elaboration of the construct of memory with emphasis on distinctions between short- and long-term memory. It also features discussion of data on developmental changes in memory. Neurodevelopmental and social and cultural influences of change in memory are highlighted.

The singular noun – *memory* – implies a single entity. In fact, memory is not a unitary construct. There are many different kinds and types of memory, each with its own distinctive characteristics. The focus of this article is short- and long-term memory. As the names imply, short-term memory lasts only seconds or minutes before fading away, whereas long-term memories may last a lifetime. A major goal of this article is to describe the course of development of these different types of memory over the first years of life. A second goal is to describe some of the mechanisms of development, including changes in neural systems and social and cultural factors.

Short-Term and Working Memory

Short-term memory operates over brief spaces of time: on the order of 10–30 s. Working memory also is temporally limited but differs from short-term memory in another way: it involves manipulation of information during temporary storage.

Verbal Tests and Developments in Childhood

Tasks of short-term memory require encoding and temporary maintenance of information. Tests of working memory impose the additional constraint that during the brief delay, the subject must operate on the information in some way. An oft-used test of short-term memory involves listening to a string of digits (e.g., 8-2-4-5-7-1) and then repeating them after a brief delay (8-2-4-5-7-1). To make this simple task a measure of working memory, the requirement to reverse the order of the sequence is imposed. In this case, the correct response is 1-7-5-4-2-8.

Short-term and working also have capacity limitations that change with development. Whereas the memory span for adults is seven items (±2), the short-term memory spans for children 9, 7, 5, and 2 years of age are estimated to be six, five, four, and two items, respectively (Dempster, 1981). Critically, in both adults and children, memory span varies as a function of domain-specific knowledge. In domains in which children have greater expertise than adults, the typical age-related advantage for adults relative to children is reversed (e.g., Chi, 1978).

A variety of paradigms have been used to examine the simultaneous maintenance and manipulation of information in working memory, including the *reading span task* in which subjects read a series of sentences and attempt to remember the sentence-final word of each (e.g., Daneman and Carpenter, 1980). Performance on these tasks shows the same pattern of age-related improvements with age as are apparent on tests of short-term memory. Typically, the absolute number of items that can be retained in working-memory tasks is reduced relative to short-term memory tasks, however.

Nonverbal Tests and Developments in Infancy

A common means of testing the duration of short-term or working memory in nonverbal or preverbal organisms is with the *delayed-response task*. The task involves hiding an object in one of multiple locations, imposing a delay, and then allowing the subject to retrieve the object. Based on this task, the temporal extent of working memory in infancy increases from 0 s at 8 months to 8 s at 12 months (e.g., Diamond and Doar, 1989).

Estimates of short-term or working-memory capacity can be obtained from the *change-detection paradigm*. Infants observe two monitors side-by-side, each with colored squares that blink on and off. On one monitor the stimuli remain constant across trials whereas the color of the stimulus on the other monitor changes on each trial. If infants notice the color change, they look longer at the novel stimulus. The capacity of memory is evaluated based on the number of squares in the display. Capacity increases from 1 to 2–3 elements over the first year of life (Oakes et al., 2007).

Long-Term Memory

Long-term memories may last an indefinite period of time and the capacity of long-term memory seemingly is infinite. A major distinction in long-term memory is between that which is deliberate and strategic and that which is incidental. Incidental long-term memories are of unique episodes and experiences (episodic memory) and may be of high personal relevance and significance (autobiographical memory).

Deliberate and Strategic Memory

Especially in academic settings, individuals may consciously work to maintain memories and use strategies to do so. Although there is evidence of deliberate memory even in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23064-X

infancy (see Roebers, 2014), the most striking age-related changes occur in the later preschool and elementary school years. During this period, children become more strategic in their memory behavior, employing a greater number and variety of strategies and deriving greater benefit from them (see Bjorklund et al., 2009; for a review).

In a seminal study, Flavell et al. (1966) found that verbal rehearsal – repetition of the names of to-be-remembered picture stimuli – was more common in older than younger children. Whereas only 10% of the 5-year-olds engaged in rehearsal, 85% of the 10-year-olds did. The older children also remembered more. Subsequent studies observed that even when younger children rehearse, they do so less effectively. Younger children are more likely to repeat only a subset of the stimuli they are required to remember (*passive* rehearsal) whereas older children repeat more items (*active* or *cumulative* rehearsal; e.g., Guttentag et al., 1987). When young children are trained to use the more effective cumulative rehearsal, recall improves, though not to the levels observed in older children (e.g., Cox et al., 1989).

Use of sorting or clustering of related items at the time of study and/or test shows similar developmental trends. Relative to older children, younger children are less likely to spontaneously group similar items together into categories as they are studying them and at the time of recall, they are less likely to report items in groups or 'clusters' from the same category (e.g., Bjorklund et al., 1992). When younger children are explicitly instructed to sort, they use the strategy and experience a concomitant increase in the number of items recalled, yet still do not recall as many items as older children (e.g., Carr and Schneider, 1991). By later in the school years, the repertoire of strategies that children use has expanded substantially, children use the strategies spontaneously, and they derive mnemonic benefit from them. The more strategies used, the greater the level of recall.

Incidental Episodic Memory

Long-term episodic memory supports retention of information about unique events that can be located in a particular place and time (Tulving, 1972). It has a protracted course of development, beginning in infancy and extending into adolescence.

Episodic Memory in Infancy

Over the first years of life there are developmental changes in long-term memory for specific episodes along a number of dimensions. Because infants are preverbal, a common means of testing long-term memory is with nonverbal imitation-based tasks in which objects are used to demonstrate an action or sequence of actions that infants are invited to imitate. The task is an accepted analog to verbal report (e.g., Bauer, 2007; Carver and Bauer, 2001).

A salient change in episodic memory over the first 2 years of life is the length of time over which it is apparent. At 6 months, infants remember for 24 h whereas by 9 months, they remember over delays as long as 5 weeks (Lukowski and Bauer, 2014; for a review). Thirteen- to fourteen-month-olds remember actions over delays of 4–6 months (Bauer et al., 2000; Meltzoff, 1995), and by 20 months, children remember for as many as 12 months (Bauer et al., 2000). Infants also recall the temporal order of multistep sequences, though retaining order information presents a cognitive challenge to young infants, in particular. Age-related differences in ordered recall are especially apparent under conditions of greater cognitive demand, such as when less support for recall is provided, and after longer delays (Bauer et al., 2000).

Over the first 2 years there also are changes in the robustness of memory for specific episodes. For instance, there are changes in the number of experiences that seem to be required in order for infants to remember. Whereas 6-month-olds require six exposures to events in order to remember them 24 h later, 14-month-olds require only a single exposure to support recall over 4 months (Bauer, 2013; for a review). Another index of memory robustness is the extent to which it is disrupted by interference. One form of interference that has been studied in infancy is changes in context between encoding and retrieval. With development, infants are increasingly able to show evidence of memory even over changes in (1) the size, shape, color, and/or material composition of the objects used in demonstration vs test; (2) the appearance of the room at the time of demonstration vs test; (3) the setting for demonstration of the modeled actions vs test; and (4) the individual who demonstrated the actions and the individual who tested for memory for the actions (see Lukowski and Bauer, 2014; for a review).

Episodic Memory in Early Childhood

Beginning in the third year, verbal means of assessment become a viable means of testing long-term memory. This permits testing of controlled laboratory events as well as events from outside the laboratory. Early studies of young children's memories for the events of their own lives focused on everyday, routine events. The research revealed 'minimalist,' yet nevertheless accurate, reports by children as young as 3 years of age. The children's reports included actions common to the activities and almost invariably, the actions were mentioned in the temporal order in which they typically occurred (Nelson and Gruendel, 1986). The actions also were those most central or important to the event, suggesting that children appreciated the goals of many of the events of their lives.

Young children also form memories of unique events. For example, in Hamond and Fivush (1991), 3- and 4-year-old children recalled a trip to Disney World they had taken either 6 or 18 months previously. The amount they remembered did not differ as a function of the delay. Moreover, the older and younger children did not differ in the overall amount of information they reported about the event. The older children tended to provide more elaborate responses to memory questions, relative to the younger children, however.

There are also suggestions of developmental changes in the length of time over which memories of unique events are retained. For example, Quas and her colleagues assessed children, 3–13 years of age, for their memories for a painful medical procedure that they had experienced between 2 and 6 years (Quas et al., 1999). Children who were older at the time of the procedure provided clearer evidence of recall of it. In contrast to the older children's detailed reports, the reports of the younger children were vague. Moreover, none of the children who had been 2 years of age at the time of the procedure demonstrated clear memory for it later, whereas

most of the children who had been at least 4 years old at the time of the procedure remembered it.

With age, children report different types of information: older children and adults focus on what is unique or distinctive about an experience whereas younger children seem to comment on what is common across experiences. For example, in Fivush and Hamond (1990), in response to the interviewer's invitation to talk about going camping, after providing the distinctive information that the family had slept in a tent, a 2.5 year old went on to report on the more typical features of the camping experience. In total, only 48% of the information that the children reported was distinctive. This pattern changes over time: By 4 years of age, children report about three times more distinctive than typical information (Fivush and Hamond, 1990). Greater focus on the distinctive features of events is associated with increases in the ability to remember episodes that are truly unique from one another.

Incidental Autobiographical Memory

Changes in how long unique episodes are remembered and in the distinctiveness of events remembered are critical ingredients in the development of a particular type of memory for specific events that becomes apparent during the preschool years, namely, *autobiographical* memory. Autobiographical memories are of events and experiences that make up one's life story or personal past. They differ from 'run-of-the-mill' episodic memories in that they are infused with a sense of personal involvement or ownership in the event. They are memories of events that happened to one's self; in which one participated; and about which one had emotions, thoughts, reactions, and reflections. It is this feature that puts the 'auto' in '*auto*biographical.'

Emergence in the Preschool Years

Over the preschool years, children's memories take on more and more autobiographical features (see Bauer, 2007, 2014, for discussions). For example, from a very young age, children include references to themselves in their narratives: "*I* fell down." With age, they increasingly pepper their narratives with the subjective perspective that indicates the significance of the event for the child (Fivush, 2012). For example, they go beyond comment on the objective reality of 'falling down' to convey how they felt about the fall: "I fell down and *was so embarrassed* because everybody was watching!" It is this subjective perspective that indicates the significance of the event to one's self.

There are also changes in the marking of events as having taken place at a specific place and time. For instance, children increasingly include specific references to time, such as 'on my birthday' or 'last summer' (Pathman et al., 2013). Markings such as these not only establish that an event happened at a time different from the present, but they begin to set up a timeline along which an organized historical record of when events occurred can be constructed. Children also include in their narratives more orienting information, including where events took place and who participated in them (e.g., Fivush and Haden, 1997). These changes serve to distinguish events from one another, thereby making them more distinctive.

Finally, over the preschool years, children include in their narratives an increasing number of elements that suggest rich detail that would contribute to the sense of vivid recollection. They include more intensifiers ("Cause she was *very* naughty"), qualifiers ("I *didn't like* her video tape"), elements of suspense ("And *you know what*?"; examples from Fivush and Haden, 1997; see also Ackil et al., 2003). In Fivush and Haden (1997), for example, children went from using approximately four descriptive terms per event at age 3.5 years to using 12 such terms per event at 6 years. The result is a much more elaborate narrative that brings both the storyteller and the listener to the brink of reliving the experience.

Changes in Childhood and Adolescence

Age-related changes in autobiographical reports continue throughout the elementary school years and beyond. One obvious change is in the breadth or completeness of children's narratives. Like a good newspaper story, a 'good' autobiographical narrative includes the *who*, *what*, *where*, *when*, *why*, and *how* of events. Younger children include fewer of these essential elements, relative to children 10–11 years of age (e.g., Bauer and Larkina, 2013; Bauer et al., 2007).

Even once children's narratives become more complete, there is continued development in autobiographical narratives. During the transition from late childhood to adolescence, personal narratives become more elaborated and they come to feature more biographical information that links events to one another (Habermas and Paha, 2001). During this same period individual identity becomes a critical issue, and personal stories begin to take on greater significance (e.g., Habermas and Bluck, 2000). With the cognitive and social advances of adolescence, individuals are able to organize and organize memories of their past experiences into a life narrative, integrating a sense of past, present, and future into a personal timeline (Habermas and Bluck, 2000).

Mechanisms of Change in Short- and Long-Term Memory

A complete understanding of the mechanisms of developmental change in short- and long-term memory ultimately will entail multiple levels of analysis, from the neural systems that support memory to the impact of culture on what is regarded as important to remember and how it is to be expressed.

Neural Systems of Memory

Early in life, a significant determinant of memory is the developmental status of the neural system responsible for it. The formation, storage, and later retrieval of memory traces depend on a multicomponent network of brain structures, including some in the temporal lobe and others in the cerebral cortex. When we experience an event, the sights, sounds, smells, and tactile sensations associated with it register in primary and association regions distributed across the cortex. For this distributed representation to endure as a memory trace, it must be stabilized or *consolidated* for long-term storage (McGaugh, 2000). Later retrieval of a memory trace depends on

prefrontal cortex, in particular (e.g., Eichenbaum and Cohen, 2001; Zola and Squire, 2000).

Some aspects of the temporal-cortical network that supports memory develop early. Others undergo a protracted course of development that extends into adolescence (see Bauer, 2006, 2007, 2014; Richman and Nelson, 2008, for discussions). The long, slow course of development of the neural structures and network responsible for memory means that throughout infancy and for much of childhood, the processes by which new memories are formed, stabilized, and integrated into long-term storage are relatively ineffective and inefficient. As a result, memories are vulnerable to disruption and interference and thus to forgetting (see Bauer, 2009; for discussion). The younger the child, the greater the vulnerability. The vulnerability of memory traces in childhood contributes to more rapid forgetting of events and experiences, ultimately resulting in the phenomenon of childhood amnesia (the relative paucity of memories of events from the early years of life: Bauer and Larkina, 2013; Bauer et al., 2007).

Social and Cultural Influences on the Development of Memory

From some of its earliest manifestations, memory is influenced by the social world (Nelson and Fivush, 2004). Social influences from the family to the larger cultural group shape what children ultimately come to view as important to remember and even how they express their memories.

Familial Factors

Young children learn to craft personal narratives by participating in conversations about past events with more skilled partners, typically their parents. Adults exhibit two different styles when they engage in memory talk with their children (Fivush, 2014; for a review). Parents who frequently engage in conversations about the past, provide rich descriptive information about previous experiences, and invite their children to 'join in' on memory conversations, are said to use an *elaborative* style. In contrast, parents who provide fewer details and instead pose specific questions to their children (e.g., "What was the name of the restaurant where we had breakfast?"), are said to use a *repetitive* or *low elaborative* style.

Stylistic differences have implications for children's memory narratives. Children of parents using the elaborative style report more about past events than children of parents using the repetitive or low elaborative style (see Fivush, 2014; for a review). The elaborative style is thought to provide the young child with, in effect, a 'prototype' of the narrative frame into which information about the actors, actions, intentions, and affective experiences of those involved in an event is inserted. The advent of the narrative organization allows the child to begin to construct the personal history that is the hallmark of autobiographical memory. Consistent with this suggestion is evidence of cross-lag correlation from maternal style variables at 40 and 46 months to child verbal report variables at 58 and 70 months of age: Maternal use of a more elaborative style facilitates children's independent narrative accounts (Reese et al., 1993). Effects of elaborative style on children's memory extend beyond the reminiscing context to imitation-based tasks (Bauer and Burch, 2004). Analogous effects are observed in deliberate memory tasks (Larkina et al., 2008).

Cultural Factors

The larger cultural milieu also influences the development of memory (see Wang, 2014). Most of the research that makes this point has involved comparisons of children from Eastern and Western culture groups. It has been found that the early autobiographical memory reports of children from Asian cultures include fewer references to themselves and fewer personal evaluations, relative to reports from children in the United States. In addition, the autobiographical narratives of children from Asian cultures tend to feature generic as opposed to specific events, and they are shorter and less detailed, relative to those provided by children in the United States (e.g., Wang et al., 2000). As reviewed by Wang (2014), these differences may be associated with variability in cultural perspectives on the value and goal of reminiscing. Indeed, Asian dyads engage in talk about the past less frequently than American dyads, and Asian mothers are less elaborative than American mothers, especially when talking about emotional experience (e.g., Wang et al., 2000).

Conclusions

Short- and long-term memory are apparent even in infancy yet undergo protracted courses of development. In the case of short-term and working memory, developmental changes are apparent in the duration over which information can be held in mind and manipulated and in the capacity of the memory systems.

Long-term memories may be either deliberate and strategic or incidental. Incidental memory may be of everyday or routine events or of specific events located in place and time, so-called episodic memories. Episodic memories of personally relevant or significant events are considered autobiographical. Both the number and efficacy of memory strategies increase over the preschool years and especially in childhood. Children experience concomitant improvements in the amount of information they remember. Episodic and autobiographical memory also become more robust over the preschool years and beyond, resulting in a larger corpus of memories of unique and distinctive events, many of which are personally relevant and even self-defining.

Developments in memory are multiply determined, by factors ranging from the neural structures and network that subserves memory, to the social and cultural influences on what is considered important to remember and express. Neural system developments result in changes in the vulnerability of memories to forgetting, contributing to the development of childhood amnesia. Social and cultural factors contribute to individual and culture group differences in the robustness of autobiographical memories across the lifespan.

See also: Amnesia: General; Amnesia: Psychogenic; Autobiographical Memory During Early Childhood and Cultural Variations; Autobiographical Memory; Childhood Memories; Cognitive Development in Infancy, Neural Mechanisms of; Early

Memory Development and Cultural Variations; Episodic Memory; Genetic Approaches to Memory; Infancy and Human Development; Learning and Memory, Neural Basis of; Memory Development During Infancy and Early Childhood across Cultures; Short-Term Memory: Psychological and Neural Aspects; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Ackil, J.K., Van Abbema, D.L., Bauer, P.J., 2003. After the storm: enduring differences in mother-child recollections of traumatic and nontraumatic events. Journal of Experimental Child Psychology 84, 286–309.

Bauer, P.J., 2006. Constructing a past in infancy: a neuro-developmental account. Trends in Cognitive Sciences 10, 175–181.

Bauer, P.J., 2007. Remembering the Times of Our Lives: Memory in Infancy and Beyond. Erlbaum, Mahwah, NJ.

Bauer, P.J., 2009. The cognitive neuroscience of the development of memory. In: Courage, M.L., Cowan, N. (Eds.), The Development of Memory in Infancy and Childhood, second ed. Psychology Press, New York, pp. 115–144.

Bauer, P.J., 2013. Memory. In: Zelazo, P.D. (Ed.), Oxford Handbook of Developmental Psychology, Body and Mind. vol. 1. Oxford University Press, New York, pp. 505–541.

Bauer, P.J., 2014. The development of forgetting: childhood amnesia. In: Bauer, P.J., Fivush, R. (Eds.), The Wiley-Blackwell Handbook on the Development of Children's Memory. Wiley-Blackwell, West Sussex, UK, pp. 519–544.

Bauer, P.J., Burch, M.M., 2004. Developments in early memory: multiple mediators of foundational processes. In: Lucariello, J., Hudson, J.A., Fivush, R., Bauer, P.J. (Eds.), Development of the Mediated Mind: Culture and Cognitive Development. Essays in Honor of Katherine Nelson. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 101–125.

Bauer, P.J., Burch, M.M., Scholin, S.E., Güler, O.E., 2007. Using cue words to investigate the distribution of autobiographical memories in childhood. Psychological Science 18, 910–916.

Bauer, P.J., Larkina, M., 2013. Childhood amnesia in the making: different distributions of autobiographical memories in children and adults. Journal of Experimental Psychology: General. http://dx.doi.org/10.1037/a0033307. Advanced on-line publication.

Bauer, P.J., Wenner, J.A., Dropik, P.L., Wewerka, S.S., 2000. Parameters of remembering and forgetting in the transition from infancy to early childhood. Monographs of the Society for Research in Child Development 65 (4, Serial No. 263).

Bjorklund, D.F., Coyle, T.R., Gaultney, J.F., 1992. Developmental differences in the acquisition and maintenance of an organizational strategy: evidence for the utilization deficiency hypothesis. Journal of Experimental Child Psychology 54, 434–448.

Bjorklund, D.F., Dukes, C., Brown, R.D., 2009. The development of memory strategies. In: Courage, M.L., Cowan, N. (Eds.), The Development of Memory in Infancy and Childhood. Taylor & Francis, New York, pp. 145–175.

Carr, M., Schneider, W., 1991. Long-term maintenance of organizational strategies in kindergarten children. Contemporary Educational Psychology 16, 61–72.

Carver, L.J., Bauer, P.J., 2001. The dawning of a past: the emergence of long-term explicit memory in infancy. Journal of Experimental Psychology: General 130, 726–745.

Chi, M.T.H., 1978. Knowledge structures and memory development. In: Siegler, R.S. (Ed.), Children's Thinking: What Develops? Erlbaum, Hillsdale, NJ, pp. 73–95.

Cox, B.D., Ornstein, P.A., Naus, M.J., Maxfield, D., Zimler, J., 1989. Children's concurrent use of rehearsal and organizational strategies. Developmental Psychology 25, 619–627.

Daneman, M., Carpenter, P.A., 1980. Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior 19, 450–466.

Dempster, F.N., 1981. Memory span: sources of individual and developmental differences. Psychological Bulletin 89, 63–100.

Diamond, A., Doar, B., 1989. The performance of human infants on a measure of frontal cortex function, the delayed-response task. Developmental Psychobiology 22, 271–294.

Eichenbaum, H., Cohen, N.J., 2001. From Conditioning to Conscious Recollection: Memory Systems of the Brain. Oxford University Press, New York.

Fivush, R., 2012. Subjective perspective and personal timeline in the development of autobiographical memory. In: Berntsen, D., Rubin, D.C. (Eds.), Understanding Autobiographical Memory: Theories and Approaches. Cambridge University Press, New York, pp. 226–245.

Fivush, R., 2014. Maternal reminiscing style: the sociocultural construction of autobiographical memory across childhood and adolescence. In: Bauer, P.J., Fivush, R. (Eds.), The Wiley-Blackwell Handbook on the Development of Children's Memory. Wiley-Blackwell, West Sussex, UK, pp. 568–585.

Fivush, R., Haden, C.A., 1997. Narrating and representing experience: preschoolers' developing autobiographical accounts. In: van den Broek, P., Bauer, P.J., Bourg, T. (Eds.), Developmental Spans in Event Representation and Comprehension: Bridging Fictional and Actual Events. Erlbaum, Mahwah, NJ, pp. 169–198.

Fivush, R., Hamond, N.R., 1990. Autobiographical memory across the preschool years: toward reconceptualizing childhood amnesia. In: Fivush, R., Hudson, J.A. (Eds.), Knowing and Remembering in Young Children. Cambridge University Press, New York, pp. 223–248.

Flavell, J.H., Beach, D.R., Chinsky, J.H., 1966. Spontaneous verbal rehearsal in a memory task as a function of age. Child Development 37, 283–299.

Guttentag, R.E., Ornstein, P.A., Siemens, L., 1987. Children's spontaneous rehearsal: transitions in strategy acquisition. Cognitive Development 2, 307–326.

Habermas, T., Bluck, S., 2000. Getting a life: the emergence of the life story in adolescence. Psychological Bulletin 126, 748–769.

Habermas, T., Paha, C., 2001. The development of coherence in adolescent's life narratives. Narrative Inquiry 11, 35–54.

Hamond, N.R., Fivush, R., 1991. Memories of Mickey Mouse: young children recount their trip to Disneyworld. Cognitive Development 6, 433–448.

Larkina, M., Güler, O.E., Kleinknecht, E., Bauer, P.J., 2008. Maternal provision of structure in a deliberate memory task in relation to their preschool children's recall. Journal of Experimental Child Psychology 100, 235–251.

Lukowski, A.F., Bauer, P.J., 2014. Long-term memory in infancy and early childhood. In: Bauer, P.J., Fivush, R. (Eds.), The Wiley-Blackwell Handbook on the Development of Children's Memory. Wiley-Blackwell, West Sussex, UK, pp. 230–254.

McGaugh, J.L., 2000. Memory—A century of consolidation. Science 287, 248–251.

Meltzoff, A.N., 1995. What infant memory tells us about infantile amnesia: long-term recall and deferred imitation. Journal of Experimental Child Psychology 59, 497–515.

Nelson, K., Fivush, R., 2004. The emergence of autobiographical memory: a social cultural developmental theory. Psychological Review 111, 486–511.

Nelson, K., Gruendel, J., 1986. Children's scripts. In: Nelson, K. (Ed.), Event Knowledge: Structure and Function in Development. Erlbaum, Hillsdale, NJ, pp. 21–46.

Oakes, L.M., Ross-Sheehy, S., Luck, S.J., 2007. The development of visual short-term memory in infancy. In: Oakes, L., Bauer, P.J. (Eds.), Short- and Long-Term Memory in Infancy and Early Childhood: Taking the First Steps toward Remembering. Oxford University Press, New York, pp. 75–102.

Pathman, T., Larkina, M., Burch, M.M., Bauer, P.J., 2013. Young children's memory for the times of personal past events. Journal of Cognition and Development 14, 120–140.

Quas, J.A., Goodman, G.S., Bibrose, S., Pipe, M.-E., Craw, S., Ablin, D.S., 1999. Emotion and memory: children's long-term remembering, forgetting, and suggestibility. Journal of Experimental Child Psychology 72, 235–270.

Reese, E., Haden, C.A., Fivush, R., 1993. Mother-child conversations about the past: relationships of style and memory over time. Cognitive Development 8, 403–430.

Richman, J., Nelson, C.A., 2008. Mechanisms of change: a cognitive neuroscience approach to declarative memory development. In: Nelson, C.A., Luciana, M. (Eds.), Handbook of Developmental Cognitive Neuroscience, second ed. The MIT Press, Cambridge, MA, pp. 541–552.

Roebers, C.M., 2014. Children's deliberate memory development: the contribution of strategies and metacognitive processes. In: Bauer, P.J., Fivush, R. (Eds.), The Wiley-Blackwell Handbook on the Development of Children's Memory. Wiley-Blackwell, West Sussex, UK, pp. 865–894.

Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.

Wang, Q., 2014. The cultured self and remembering. In: Bauer, P.J., Fivush, R. (Eds.), The Wiley-Blackwell Handbook on the Development of Children's Memory. Wiley-Blackwell, West Sussex, UK, pp. 605–625.

Wang, Q., Leichtman, M.D., Davies, K., 2000. Sharing memories and telling stories: American and Chinese mothers and their 3-year-olds. Memory 8, 159–177.

Zola, S.M., Squire, L.R., 2000. The medial temporal lobe and the hippocampus. In: Tulving, E., Craik, F.I.M. (Eds.), The Oxford Handbook of Memory. Oxford University Press, New York, pp. 485–500.

Short Term Memories, Theories of

Evie Vergauwe and Nelson Cowan, University of Missouri-Columbia, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses how information that is no longer available for perception is maintained mentally. To do so, the psychological concepts of short-term memory and working memory are introduced. Next, two different research traditions in the literature are discussed, together with the key theories and findings of each tradition. From that, we will conclude that the general way to maintain information over a short period of time is by focusing attention on it. For verbal material, such as a phone number, people have an extra mechanism available that consists in articulatory rehearsal.

What is short-term memory (STM)? One very popular example of STM is the capacity to remember a telephone number from the phone book over the short term, that is, at least till one is able to dial it. Within psychology, the limited-capacity system underpinning human's ability to maintain mentally information that is no longer available for perception is referred to as STM. Maintaining an address, a vehicle registration plate, a grocery list, or driving directions are just some of the most frequent examples of the need to maintain some information over a brief period of time in our daily life. How is one able to keep information in memory when it is no longer physically available, that is, when one can no longer hear, see, feel, taste, or smell the information to be maintained?

As will be seen, the concept of STM is very closely related to another concept in psychology, that of working memory. Investigators differ in their definitions but the authors adhere to the following popular distinction. While the concept of STM specifically refers to the ability of storing information over a short period of time (a matter of seconds), working memory refers to the ability to maintain mentally information in an active and accessible state, while concurrently and selectively processing some additional information (Baddeley and Hitch, 1974). As such, while both concepts refer to maintaining information over the short term, working memory is specifically concerned with doing so while processing new incoming information. Thus, if one were to maintain a phone number while just sitting there doing nothing else, psychologists would say that the person is using STM. If, however, one were trying to remember the same phone number while watching a music clip on YouTube, psychologists would say the person is using working memory. Of course, this does not necessarily mean that the person is using a whole other set of cognitive mechanisms and processes to fight against forgetting of the phone number. It just means that, depending on the exact conditions in which one is trying to maintain information, the psychological concept one would refer to could be either STM or working memory.

The ever-changing and very dynamic environment we are currently living in means that one is most often trying to maintain information while doing something else. Other examples of using working memory in daily life are paying cash, which requires one to maintain how much one already gave while calculating how much more one should give to pay the total amount; taking notes in class, which requires one to maintain and write down the beginning of a sentence while listening to the ongoing stream of information; or simply being on Facebook, which frequently puts one in a situation in which he or she has to maintain the content of an ongoing chat conversation with one friend while processing the incoming news feeds of other friends.

From Short-Term Memory to Working Memory

While the concepts of STM and working memory both refer to maintaining information over a rather short period of time, other faculties have been proposed for maintaining information over longer periods of time. The distinction between a short-term and a long-term component of memory had been proposed by William James (1890), who distinguished between a temporary primary memory and a more durable secondary memory, and gained popularity among psychologists when Donald Hebb suggested in 1949 the existence of two types of memory, STM and long-term memory (LTM), with the former depending on temporary electrical activity in the brain and the latter on more durable neurochemical changes in the brain, a view that was supported by Brown (1958) and Peterson and Peterson (1959). Miller (1956) suggested that the capacity of STM is limited so that people can retain only about seven items. Some alternative proposals notwithstanding (e.g., Melton, 1962), by the late 1960s, the distinction between STM and LTM was widely accepted in the memory field. Moreover, the distinction had been incorporated in one of the most influential models of human memory, the model of Atkinson and Shiffrin (1968, 1971).

The Atkinson–Shiffrin (1968) model (also known as the multimemory model, the multistore model, or humorously as the modal model because it was typical of a number of models of the era) proposed three separate memory stores, differing in their function, capacity, and duration. Information was assumed to flow from the environment through a series of brief sensory registers, which are part of the perceptual system, into a single short-term store. This limited-capacity short-term store was thought of as a unitary system responsible for encoding information into and retrieving information from a long-term store, this third type of store having a much larger capacity and capable of more or less permanent storage. Importantly, the unitary short-term store was conceived as an active memory, an operating

memory capable of both storing and manipulating information. That is, Atkinson and Shiffrin assigned the role of working memory to the short-term store by assuming that it could handle both processing and storage of information instead of only storing information.

There was, however, little empirical evidence backing up this idea. Also, Shallice and Warrington (1970; Warrington and Shallice, 1969) reported the case of a patient showing severe impairment on STM tasks, an impairment that could be accounted for by assuming a defective short-term store. However, if the unitary short-term store is indeed a working memory system that handles both processing and storage of information, then the patient should not only show impaired STM performance, but also a general cognitive processing dysfunction. No such general impairment was found; the patient showed preserved learning and comprehension abilities. This suggested that processing and storage are handled by different systems.

In their seminal 1974 paper, Baddeley and Hitch decided to examine this issue in detail. Instead of studying patients, they studied performance in normal subjects using a dual-task paradigm. In such a paradigm, subjects are required to perform two tasks at the same time. Baddeley and Hitch asked participants to maintain series of digits while performing processing tasks such as reasoning or language comprehension. They reasoned that maintaining series of digits would use a substantial amount of the storage capacity of a limited capacity working memory system. If this working memory system is also used for general processing, then reasoning and language comprehension performance should suffer from the concurrent memory load. This should be so because, when they are performed under concurrent memory load, less capacity would be available for the processing tasks. Contrary to this prediction, it was observed that reasoning and language comprehension was not disrupted by having to maintain three digits concurrently and were only moderately disrupted by having to maintain six digits concurrently. This pattern of results suggested that even if there is some overlap between the system supporting general cognitive processing and the system supporting storage, there seems to be some capacity available for processing that was not taken up by a concurrent memory load of six digits (even though that is near the amount a typical adult can remember and repeat, as shown by Miller, 1956). From this, Baddeley and Hitch concluded that, while there might be a central workspace devoted to processing and storage, storage of information is mainly handled by subsystems that are independent from this central workspace. It was proposed to replace the unitary short-term store by the concept of working memory, which was supposed to be a multiple-component system.

The core of the multiple-component model of working memory of Baddeley and Hitch was assumed to consist of a limited work space, which can be divided between processing and storage activities. In contrast to the unitary view of Atkinson and Shiffrin, however, Baddeley and Hitch did not consider working memory as one entirely flexible system of which all resources can be devoted to either storage or processing activities. Instead, they suggested that the central flexible system interacted with a more peripheral system that can be used for storage, but that is not available for general processing. Once the capacity of the peripheral system is exceeded, some of the capacity or time of the flexible central general-purpose system needs to be devoted to storage. The peripheral component was thought of as a capacity-limited phonemic buffer (also referred to as phonemic rehearsal system and phonemic loop) able to store speechlike information in serial order without drawing on the central component. As such, the observations of Shallice and Warrington (1970; Warrington and Shallice, 1969) could be accounted for by assuming that only the phonemic buffer of the patient was impaired, while the central component of his working memory system was intact. Besides the phonemic buffer for verbal information (e.g., letters, digits, words, and sentences), which is known now as the phonological loop, the existence of a second, comparable but separate, peripheral system was postulated for the storage of visuospatial information (e.g., shapes, colors, locations, and orientations), a system which is known now as the visuospatial sketchpad. In sum, the first widely used model of working memory comprised a central flexible domain-general work space, which can be divided between processing and storage activities (i.e., central executive component) and two domain-specific storage-only components, one for verbal information (i.e., phonological loop) and one for visuospatial information (i.e., visuospatial sketchpad). The first component can be referred to as the central level, the latter components as the peripheral level of working memory.

A Central Level and a Peripheral Level in Working Memory

The Baddeley and Hitch (1974) chapter was an extraordinary inspiration to an ever growing number of researchers in cognitive psychology and cognitive neuroscience. Accordingly, as Saito and Towse (2007) noted, a survey of academics identified the chapter as one of the 100 most influential papers in cognitive science in the last millennium (Millennium Project). At least part of the breadth and extent of its impact and of its longevity has been attributed to the fact that the concept of working memory can be used in two distinguishable ways (e.g., Saito and Towse, 2007; Towse et al., 2007). Indeed, throughout the working memory literature, two different conceptions of working memory emerge. Working memory has been used as a general unitary concept (which Baddeley, 1986; refers to as WMG for working memory general) playing an important role in human information processing by serving both processing and storage operations in ongoing cognition. This WMG conception of working memory can be differentiated from using the working memory concept as a specific model (which Baddeley, 1986; refers to as WMS for working memory specific) with its specific architecture, structure, and processes involved in tasks requiring WMG. Broadly speaking, researchers using the WMG conception have been inspired by the *idea* of working memory introduced by Baddeley and Hitch, whereas researchers using the WMS conception have been inspired by the model of working memory introduced by Baddeley and Hitch in 1974. These

different approaches have resulted in two different research traditions of working memory, each with their respective theories, research methodologies, assumptions, and implications. Indeed, while the WMS tradition has mainly focused on the peripheral level of working memory, the focus of the WMG tradition has primarily been on the central level of working memory, which is closely related to attention. Interestingly, as seen in the following, the traditions have different answers to our question of how information is maintained over a short period of time. Both traditions have, of course, their merits. After presenting key theories and findings of each tradition, it is shown that their answers to our question are actually quite compatible.

Focusing on the Peripheral Level in WMS: Domain-Specific Maintenance Mechanisms

Within the WMS tradition, researchers have mainly aimed at characterizing the peripheral systems of working memory while the central level remained less studied. Concerning the phonological loop, four basic findings and phenomena associated with verbal serial recall led to propose a fractionation of the verbal storage system into two subcomponents. First, the presence of irrelevant and unattended speech disrupts verbal serial recall, an effect referred to as the irrelevant speech effect (e.g., Colle and Welsh, 1976; Salamé and Baddeley, 1982). Second, recall of a series of words or letters is worse when these words or letters sound alike, that is, are phonologically similar, an effect referred to as the phonological or phonemic similarity effect (e.g., Baddeley, 1966; Conrad and Hull, 1964). Third, recall of a series of words is worse for series of long words than for series of short words and this word length effect appears to be caused by the length of time required to pronounce the words rather than by the number of syllables or letters they contain (e.g., Baddeley et al., 1975b). Finally, the requirement to utter continuously an irrelevant sound such as 'the, the, the' (i.e., articulatory suppression) disrupts verbal serial recall (e.g., Baddeley et al., 1984).

Accounting for the ensemble of these phenomena was possible by assuming that the phonological loop consists of two separate subcomponents: (1) a passive phonological store that can hold verbal information in its phonological code for about 2 s and (2) an articulatory rehearsal process that counteracts time-related forgetting of information held in the phonological store by rehearsing its content subvocally. Furthermore, it was assumed that whereas nonauditory verbal information needs to be recoded in a phonological format by the articulatory rehearsal process before it can enter the phonological store, auditory verbal information has direct access to the phonological store. Hence, unattended irrelevant speech goes directly into the phonological store where it interferes with the verbal material stored in there for further recall. As a result, recall performance is impaired. Furthermore, when verbal items stored in the phonological store are phonologically similar, they will have a large number of overlapping phonological features and, as a result, get confused. Hence, their recall will be less accurate. The effect of concurrent articulatory suppression and word length, on the other hand, are accounted for by the articulatory rehearsal loop. Indeed, the utterance of an irrelevant sound is assumed to block the process of articulatory rehearsal, meaning that verbal material stored in the phonological store cannot be maintained through that facility. Finally, the word length effect is thought of as resulting from the fact that longer words take longer to be rehearsed by the articulatory loop and, as such, delay the maintenance of other words through subvocal rehearsal, which results then in forgetting of these words.

Inspired by the substantial work and, scientifically speaking, fast progress in understanding the structure and functioning of the phonological loop, researchers started to explore more in detail its visuospatial counterpart using comparable techniques. A first step in understanding the visuospatial sketchpad was to establish experimentally its existence. Unlike the phonological loop, the existence of a separate system for visuospatial material was postulated by Baddeley and Hitch (1974) more out of theoretical necessity than based on extensive empirical evidence, although there was some; notably a study by Brooks (1968) showing that verbal recall is most readily disrupted by concurrent vocal activity while visuospatial recall is most readily disrupted by concurrent spatially monitored activity (see also Baddeley et al., 1975a). The rationale behind the experimental work that aimed at demonstrating that a visuospatial sketchpad was indeed needed to account for working memory functioning (e.g., Farmer et al., 1986; Logie, 1986; Logie et al., 1990) was the following. Researchers reasoned that, if two separate peripheral domain-specific working memory systems exist, then it should be possible to interfere with each of these systems selectively. To test this, studies (1) used verbal interference techniques such as articulatory suppression and irrelevant speech to demonstrate that these techniques have a disruptive effect on concurrent verbal activities, but not on concurrent visuospatial activities and (2) developed visuospatial interference techniques to demonstrate that these visuospatial analogs have a disruptive effect on concurrent visuospatial activities, but not (or less) on concurrent verbal activities. For example, Logie (1986) used an STM task as primary task to selectively interfere with. Participants were presented with a list of concrete words that had to be recalled after a 10-s retention interval. The domain of the memory task was manipulated by asking participants to maintain the series of words either verbally (i.e., rehearsing the words subvocally) or visually (i.e., creating a visual image of each word and integrating these images into one interactive image). During the presentation and recall of the series of words, participants were either presented with no additional information, with line drawings of common objects and animals on screen for which participants were instructed to ignore them without closing their eyes or looking away, or with spoken names of common objects and animals presented through headphones for which participants were instructed to ignore them. Hearing irrelevant speech was thought to provoke the known irrelevant speech effect on the verbally maintained words, whereas viewing irrelevant pictures was assumed to provoke a similar effect on the visually maintained words. The predicted pattern of selective interference was observed. Irrelevant speech impaired recall performance of verbally maintained words, but not that of

visually maintained words. Irrelevant viewing, on the other hand, impaired recall performance of visually maintained words while it had very little effect on recall performance of verbally maintained words. It was concluded that a visuospatial sketchpad in working memory was needed and the next step was to explore its structure and functioning.

Some studies had demonstrated that concurrent movements disrupt maintenance of spatial information (Baddeley and Lieberman, 1980; Smyth and Pendleton, 1989) and that viewing irrelevant pictures disrupts maintenance of visual information (Logie, 1986; Quinn and McConnell, 1996). These findings seemed to corroborate Logie's proposal of a visuospatial sketchpad that, like the phonological loop, can be fractionated into two subcomponents. Indeed, Logie (1995) suggested that the visuospatial memory system consists of a passive visual store (the visual cache) and an active spatial rehearsal process (the inner scribe), with the latter being related to the control of movement. Logie and Marchetti (1991) explored this idea in a recognition task in which subjects had to maintain either color shades (visual memory) or location sequences (spatial memory). The retention interval between presentation and recognition was filled with either the presentation of irrelevant pictures (visual interference) or a concurrent movement task (spatial interference). The viewing of irrelevant pictures was thought to selectively interfere with the visual passive component, whereas the concurrent movement was thought to selectively interfere with the spatial active component. This selective interference pattern was indeed observed in that spatial memory performance was disrupted by the concurrent movement task, but not by the concurrent irrelevant pictures, while visual memory performance was impaired by concurrent irrelevant pictures but unaffected by concurrent movements.

The findings of Logie and Marchetti (1991) appeared to confirm the fractionation of the visuospatial sketchpad into a passive visual store and an active spatial movement-related rehearsal process. Irrelevant visual input was thought to have direct access to the visual cache where it interferes with the visual material that is stored in there during retention (e.g., color shades). As a result, visual memory performance is impaired when irrelevant visual pictures are presented during retention. The disruptive effect of concurrent movements on spatial memory performance is accounted for by the operation of the active spatial rehearsal process, which is assumed to be related to control of movement. As such, the requirement to perform movements during retention of spatial material blocks the spatial rehearsal process, meaning that spatial material can no longer be maintained by this mechanism with poorer memory performance as a consequence. In the same way, the disruption of spatial recall performance by a perceptuomotor task (e.g., spatial tracking; e.g., Baddeley and Lieberman, 1980) can be explained by assuming that the active spatial rehearsal process was being blocked by concurrent spatial tracking.

Thus, within the WMS tradition, research has resulted in a domain-based fractionation of the peripheral system into a verbal and a visuospatial subsystem, each comprising a rehearsal mechanism independent from the central level of working memory. As such, within the WMS tradition, the answer to our question would be that information is maintained over short periods of time in domain-specific buffers by rehearsing them using a domain-specific mechanism. Importantly, these domain-specific maintenance mechanisms at the peripheral level of working memory are typically assumed to act independently from the central level (e.g., Baddeley and Logie, 1999; Cocchini et al., 2002; Duff and Logie, 2001). Thus, when one tries to maintain a phone number, one would use articulatory rehearsal, which consists in subvocally repeating the digits without the need of focused attention. It should be noted that, although most researchers would agree on the existence of some domain-specific mechanisms, independent from attention and devoted to maintenance, the precise conception of these mechanisms is still a matter of debate (see Jalbert et al., 2011a,b; Jones et al., 2007; Lewandowsky and Oberauer, 2008 for alternative views on the phonological loop; see Klauer and Zhao, 2004; Repovs and Baddeley, 2006; Vergauwe et al., 2009; Wood, 2007 for alternative views on the visuospatial sketchpad).

Focusing on the Central Level in WMG: Domain-General Maintenance Mechanisms

While the WMS tradition has focused on the peripheral level of working memory, the WMG tradition has mainly focused on the central component of working memory as proposed by Baddeley and Hitch (1974). Remember that, in the original multiple-component model, the central work space was proposed to be devoted to processing and storage. In other words, even though domain-specific maintenance mechanisms would exist at the peripheral level, Baddeley and Hitch left open the possibility of there also being a domain-general mechanism of storage at the central level. Interestingly, while this central domain-general storage faculty was left largely unexplored within the WMS tradition (but see more recent research, e.g., Baddeley et al., 2010, 2011), the possibility of domain-general storage was central in the embedded-processes model proposed by Cowan (1988, 1995, 1999).

Cowan described the information processing system as embedded processes: (1) the LTM system, (2) within which, a subset is currently activated, that is, in a temporarily heightened state of activation, and (3) within which, a subset is currently in the focus of attention. While activated memory was supposed to include diverse features of information and to be limited by decay over time and by interference with other information that is alike. The focus of attention, however, was proposed to include coherent, integrated pieces of information (i.e., chunks, see Miller, 1956; for constant recall limited to several chunks see Chen and Cowan, 2009; Cowan et al., 2012) and was assumed to be limited in capacity to about 3–4 chunks (Cowan, 2001). In such model, the limited storage capacity observed in humans is assumed to reflect the limited capacity of the focus of attention that serves as a domain-general storage mechanism. In line with that, several studies have shown that there is a constant storage limit of about 3–4 items in working memory. For example, Luck and Vogel (1997) used a visual-array comparison task in which an array of objects such as colored squares is briefly displayed on computer screen. Next, after a brief delay, a second array is shown and participants had to decide whether the arrays were

the same or different. Typically, performance dropped severely for arrays with more than three to four colored squares and this limit was found to be constant across the number of features in the stimuli. As noted by Cowan (2011), this limit was also found to be constant within an individual, across variations in the complexity of the stimuli (Awh et al., 2007) and the number of items in the array (Zhang and Luck, 2008). Importantly, this limit at about three to four items has not only been observed for briefly presented visual arrays, but also for lists of spoken material. Cowan (2001) suggested that the limit of three to four items occurs under certain circumstances in which items are integrated objects and the circumstances discourage rehearsal or chunking of the items to reduce the load. In line with that, there are several studies that show that memory for spoken lists is limited at about four items in adults when they maintain the information under concurrent articulatory suppression, a technique known to prevent the use of articulatory rehearsal (for an overview, see Table 2 in Cowan, 2001).

If this limit does indeed arise from the capacity limit of a domain-general storage device, then one should be able to demonstrate a fixed capacity limit across different domains of working memory. For example, if one is trying to maintain visual colors and spoken digits at the same time, one should find a limit at four items in total, such as three colors plus one spoken digit. It is important to note that such a trade-off in memory capacity across domains would not be expected if only domain-specific maintenance mechanisms would exist because, in that case, the digits would be stored in the verbal subsystem and the colors in the visuospatial subsystems, with very little influence from one subsystem on another because they are supposed to operate independently from attention and independently from each other. A constant capacity across domains was observed by Saults and Cowan (2007). Adults were able to maintain about four visual items or, when it was combined memory for spoken digits, four items in total (e.g., three colors plus one digit). Other studies also showed that verbal and visuospatial memory tasks interfere with each other (e.g., Morey and Cowan, 2004, 2005), suggesting that there is indeed a domain-general storage faculty in working memory. In the model of Cowan, the focus of attention would serve this role.

If the domain-general storage device is indeed closely related to attention, then one should observe that maintenance of information in this device is disrupted by other attention-demanding distractions that do not necessarily require maintaining some information, regardless of the specific domain to which that information pertains. Accordingly, researchers have shown that memory for visual arrays is disrupted by the requirement to identify tones that are presented to participants while they are trying to maintain the visual array (Morey and Bieler, 2012; Stevanovski and Jolicoeur, 2007). Similarly, Chen and Cowan (2009) have demonstrated that memory for spoken lists is disrupted by performing a simple three-choice visual reaction time task.

The question of whether concurrent attention-demanding activities disrupt ongoing maintenance activities, across the different domains of working memory, has been addressed explicitly within the framework of the time-based resource sharing model of working memory (TBRS, Barrouillet et al., 2004, 2007, 2011). According to this model, both processing and storage rely on attention. Attention being a limited-capacity resource has to be shared between the two main functions of working memory that are processing and storage. Next, the model assumed that, while attention is focused on information, it activates that information, but when attention is switched away from information, the memory traces of that information decay over time. In order to fight against this time-based forgetting, one needs to bring the information back into the focus of attention so that the memory traces get reactivated (see also Cowan, 1995). Importantly, as the TBRS model assumes that the focus of attention can only be on one element at a time, it cannot process and store information at the same time. This means that, while attention is captured by a processing task, it is not available for maintaining information and thus, information can get lost. Thus, the TBRS model predicts that as the proportion of time during which attention is captured by a processing task gets bigger, more information will get lost with poorer recall performance as a result. This proportion of time is called cognitive load.

To test this, new computer-paced complex span tasks were created. Complex span tasks require the retention of a series of items while concurrently performing a processing task and were developed by researchers assuming that any good measure of working memory should involve both processing and storage activities in order to capture the essential dual function of working memory. For example, Daneman and Carpenter (1980) created the reading span task that combines sentence verification with memory for series words, Case et al. (1982) created the counting span that combines counting dots with memory for series of numbers, and Turner and Engle (1989) created the operation span that combines solving math problems with memory for words. It has been shown that scores in complex span tasks have larger and more consistent correlations with higher order cognition than scores in simple span tasks that only require maintenance of information (e.g., Ackerman et al., 2005; Daneman and Carpenter, 1980; Engle et al., 1999).

Using similar kinds of tasks in which the pace of processing was imposed by the computer, Barrouillet and colleagues showed that recall performance is a linear function of the cognitive load of concurrent processing when processing and storage pertain to the same domain, either both verbal (e.g., Barrouillet et al., 2004, 2007, 2011) or both visuospatial (Vergauwe et al., 2009). Importantly, however, increasing the cognitive load of concurrent processing also disrupted maintenance for information that pertained to another domain of working memory: less letters were recalled when the cognitive load of a concurrent task consisting in judging spatial configurations increased, and less spatial locations were recalled when the cognitive load of a concurrent task consisting in judging words increased (Vergauwe et al., 2010; Vergauwe et al., 2012). This shows maintenance of information is disrupted by other attention-demanding distractions that do not necessarily require maintaining some information, regardless of the specific domain to which that information pertains.

As noted before, the initial notion within the WMG tradition was that in order to test the capability of the working memory system, both storage and processing had to be tied up

(Daneman and Carpenter, 1980; Case et al., 1982). Accordingly, complex span tasks were constructed in the aim of understanding individual and developmental differences in cognition and it was shown that these measures correlate robustly with measures of cognitive ability, much more than simple span measures (e.g., Daneman and Merikle, 1996), even when the storage and processing tasks come from different cognitive domains, such as spatial versus verbal domains (e.g., Kane et al., 2004: Lépine et al., 2005). However, an alternative interpretation of this set of findings is that one must tap attention-demanding processes to observe strong correlations with ability, and the processing task prevents rehearsal processes that otherwise would circumvent attention. Consequently, a single task can correlate well with cognitive ability when rehearsal is impossible. A good example is running span, in which a list of items is presented that is difficult to rehearse because the list length is not known in advance to the participant. With such a procedure, useful rehearsal is difficult or impossible (e.g., Hockey, 1973) and correlations with cognitive aptitudes are still very high despite the absence of separate storage and processing tasks (e.g., Cowan et al., 2005; Broadway and Engle, 2010).

Taken together, within the WMG tradition, research has resulted in proposing a domain-general maintenance mechanism that relies on attention. As such, within the WMG tradition, the answer to our question would be that information is maintained over short periods of time by focusing attention on the information to be maintained so that it gets reactivated over and over again. Thus, when one tries to maintain a phone number, one would use attentional refreshing, which consists in very briefly rethinking of the digits without the need to articulate them. Thus, it seems that the different traditions have a different answer to our question of how information is maintained over a brief period of time. However, as mentioned before, these views can be quite compatible.

Two Levels for Maintaining Information in Working Memory

A rather straightforward view to this debate would be to enforce the initial idea of Baddeley and Hitch (1974) according to which there is a central and a peripheral level in working memory, each with their own way of functioning independently from each other. In this idea, there would be a domain-general mechanism of storage that relies heavily on attention, that can be applied to maintain all kinds of information and that is disrupted by any concurrent activity insofar this activity requires attention for its execution. Although Baddeley (1986) omitted this level of storage in the interest of parsimony, he has more recently seen the need to revive such a notion in the form of the episodic buffer (Baddeley, 2000), which can include binding between different sorts of codes and served a function similar to the focus of attention of Cowan (1988). Next to this attentional refreshing, there would be peripheral maintenance mechanisms that act independently from attention, that can only be applied to a specific kind of information, and that are only disrupted by concurrent activities that involve information that is alike. Cowan's (1988) activated portion of LTM included interference on the basis of similarity to account for the same kinds of dissociations that Baddeley and Hitch accounted for with the phonological and visuospatial stores.

In line with the idea of central and peripheral levels in working memory, recent studies have shown that verbal material can indeed benefit from two independent mechanisms that operate jointly for maintenance (Camos et al., 2009; Hudjetz and Oberauer, 2007). Interestingly, Camos et al. (2011) showed that adults can adaptively choose between the two mechanisms: when they try to maintain verbal information that is phonologically very much alike, they prefer the attentional refreshing mechanism; when less attention is available, they favor articulatory rehearsal. While several studies have shown the existence of two separate mechanisms for maintaining verbal information, it appears very difficult to demonstrate the existence of a maintenance mechanism for visuospatial material that operates outside the focus of attention (e.g., Morey and Bieler, 2012; Rudkin et al., 2007; Vergauwe et al., 2009).

To conclude, the general way to maintain information over a short period of time seems to be by bringing it back into the focus of attention over and over again so that it gets reactivated instead of forgotten. For verbal material, such as a phone number, people have an extra mechanism available that consists in articulatory rehearsal, probably because this kind of information can be mapped directly onto a familiar spoken response (Baddeley and Larsen, 2007). Although working memory is relatively limited in capability and serves as a kind of bottleneck to what the human mind can accomplish, it does allow a number of strategic work-arounds (chunking, rehearsal, and refreshing) that underlie the awesome ability of humans to solve problems and create new things.

See also: Retrieval; Short-Term Memory: Psychological and Neural Aspects; Visual Memory, Psychology of; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Ackerman, P.F., Beier, M.E., Boyle, M.O., 2005. Working memory and intelligence: the same or different constructs? Psychological Bulletin 131 (1), 30–60.

Atkinson, R.C., Shiffrin, R.M., 1968. Human memory: a proposed system and its control processes. In: Spence, K.W., Spence, J.T. (Eds.), The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 2. Academic Press, New York, pp. 89–195.

Atkinson, R.C., Shiffrin, R.M., 1971. The control of short-term memory. Scientific American 225, 82–90.

Awh, E., Barton, B., Vogel, E.K., 2007. Visual working memory represents a fixed number of items regardless of complexity. Psychological Science 18, 622–628.

Baddeley, A., 2000. The episodic buffer: a new component of working memory? Trends in Cognitive Sciences 4, 417–423.

Baddeley, A., Allen, R.J., Hitch, G.J., 2010. Investigating the episodic buffer. Psychologica Belgica 50, 223–243.

Baddeley, A.D., 1966. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. Quarterly Journal of Experimental Psychology 18, 362–366.

Baddeley, A.D., 1986. Working Memory. Clarendon Press, Oxford, UK.

Baddeley, A.D., Allen, R.J., Hitch, G.J., 2011. Binding in visual working memory: the role of the episodic buffer. Neuropsychologia 49, 1393–1400.

Baddeley, A.D., Grant, W., Wight, E., Thomson, N., 1975a. Imagery and visual working memory. In: Rabbit, P.M.A., Dornic, S. (Eds.), Attention and Performance, vol. V. Academic Press, London, pp. 205–217.
Baddeley, A.D., Hitch, G., 1974. Working memory. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation, vol. 8. Academic Press, New York, pp. 47–89.
Baddeley, A.D., Larsen, J.D., 2007. The phonological loop: some answers and some questions. The Quarterly Journal of Experimental Psychology 60, 512–518.
Baddeley, A.D., Lewis, V.J., Vallar, G., 1984. Exploring the articulatory loop. Quarterly Journal of Experimental Psychology 36, 233–252.
Baddeley, A.D., Lieberman, K., 1980. Spatial working memory. In: Nickerson, R. (Ed.), Attention and Performance, vol. VIII. Erlbaum, Hillsdale, NJ, pp. 521–539.
Baddeley, A.D., Logie, R.H., 1999. Working memory: the multiple component model. In: Miyake, A., Shah, P. (Eds.), Models of Working Memory. Cambridge University Press, New York, pp. 28–61.
Baddeley, A.D., Thompson, N., Buchanan, M., 1975b. Word length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behaviour 14, 575–589.
Barrouillet, P., Bernardin, S., Camos, V., 2004. Time constraints and resource sharing in adults' working memory spans. Journal of Experimental Psychology: General 133, 83–100.
Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., Camos, V., 2007. Time and cognitive load in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 33, 570–585.
Barrouillet, P., Portrat, S., Camos, V., 2011. On the law relating processing to storage in working memory. Pyschological Review 118, 175–192.
Broadway, J.M., Engle, R.W., 2010. Validating running memory span: measurement of working memory capacity and links with fluid intelligence. Behavior Research Methods 42, 563–570.
Brooks, L.R., 1968. Spatial and verbal components of the act of recall. Canadian Journal of Psychology 22, 349–368.
Brown, J., 1958. Some tests of the decay theory of immediate memory. Quarterly Journal of Experimental Psychology 10, 12–21.
Camos, V., Lagner, P., Barrouillet, P., 2009. Two maintenance mechanisms of verbal information in working memory. Journal of Memory and Language 61, 457–469.
Camos, V., Mora, G., Oberauer, O., 2011. Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. Memory and Cognition 39, 231–244.
Case, R., Kurland, D.M., Goldberg, J., 1982. Operational efficiency and the growth of short-term memory span. Journal of Experimental Child Psychology 33, 386–404.
Chen, Z., Cowan, N., 2009. Core verbal working-memory capacity: the limit in words retained without covert articulation. The Quarterly Journal of Experimental Psychology 62, 1420–1429.
Cocchini, G., Logie, R.H., Della Sala, S., MacPherson, S.E., Baddeley, A.D., 2002. Concurrent performance of two memory tasks: evidence for domain-specific working memory systems. Memory and Cognition 30, 1086–1095.
Colle, H.A., Welsh, A., 1976. Acoustic masking in primary memory. Journal of Verbal Learning and Verbal Behaviour 15, 17–32.
Conrad, R., Hull, A.J., 1964. Information, acoustic confusion and memory span. British Journal of Psychology 55, 429–432.
Cowan, N., 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. Psychological Bulletin 104, 163–191.
Cowan, N., 1995. Attention and Memory: An Integrated Framework. Oxford Psychology Series (No. 26). Oxford University Press, New York.
Cowan, N., 1999. An embedded processes model of working memory. In: Miyake, A., Shah, P. (Eds.), Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. Cambridge University Press, Cambridge, MA, pp. 62–101.
Cowan, N., 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences 24, 87–185.
Cowan, N., 2011. The focus of attention as observed in visual working memory tasks: making sense of competing claims. Neuropsychologia 49, 1401–1406.
Cowan, N., Elliott, E.M., Saults, J.S., Morey, C.C., Mattox, S., Hismjatullina, A., Conway, A.R.A., 2005. On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. Cognitive Psychology 51, 42–100.
Cowan, N., Rouder, J.N., Blume, C.L., Saults, J.S., 2012. Models of verbal working memory capacity: what does it take to make them work? Psychological Review 119, 480–499.
Daneman, M., Carpenter, P.A., 1980. Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior 19, 450–466.
Daneman, M., Merikle, P.M., 1996. Working memory and language comprehension: a meta-analysis. Psychonomic Bulletin and Review 3, 422–433.
Duff, S.C., Logie, R.H., 2001. Processing and storage in working memory span. Quarterly Journal of Experimental Psychology 54 (A), 31–48.
Engle, R.W., Tuholski, S.W., Laughlin, J.E., Conway, A.R.A., 1999. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of Experimental Psychology: General 128, 309–331.
Farmer, E.W., Berman, J.V.F., Fletcher, Y.L., 1986. Evidence for a visuo-spatial scratch-pad in working memory. Quarterly Journal of Experimental Psychology 38 (A), 675–688.
Hebb, D.O., 1949. Organization of Behavior. Wiley, New York.
Hockey, R., 1973. Rate of presentation in running memory and direct manipulation of input-processing strategies. Quarterly Journal of Experimental Psychology 25, 104–111.
Hudjetz, A., Oberauer, K., 2007. The effects of processing time and processing rate on forgetting in working memory: testing four models of the complex span paradigm. Memory and Cognition 35, 1675–1684.
Jalbert, A., Neath, I., Bireta, T.J., Suprenant, A.M., 2011a. When does length cause the word length effect? Memory and Cognition 37, 338–353.
Jalbert, A., Neath, I., Suprenant, A.M., 2011b. Does length or neighborhood size cause the word length effect? Memory and Cognition 39, 1198–1210.
James, W., 1890. The Principles of Psychology. Henry Holt, NY.
Jones, D.M., Hughes, R.W., Macken, W.J., 2007. The phonological store abandoned. Quarterly Journal of Experimental Psychology 60, 497–504.
Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., Engle, R.E., 2004. The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology: General 133, 189–217.
Klauer, K.C., Zhao, Z., 2004. Double dissociations in visual and spatial short-term memory. Journal of Experimental Psychology: General 133, 355–381.
Lépine, R., Barrouillet, P., Camos, V., 2005. What makes working memory spans so predictive of high level cognition? Psychonomic Bulletin and Review 12, 165–170.
Lewandowsky, S., Oberauer, K., 2008. The word length effect provides no evidence for decay in short-term memory. Psychonomic Bulletin and Review 15, 875–888.
Logie, R.H., 1986. Visuo-spatial processing in working memory. Quarterly Journal of Experimental Psychology 38 (A), 229–247.
Logie, R.H., 1995. Visuo-Spatial Working Memory. Lawrence Erlbaum Associates Ltd, Hove, UK.
Logie, R.H., Marchetti, C., 1991. Visuo-spatial working memory: visual, spatial or central executive? In: Cornoldi, C., McDaniels, M.A. (Eds.), Mental Images in Human Cognition. Springer, New York, pp. 72–102.
Logie, R.H., Zucco, G.M., Baddeley, A.D., 1990. Interference with visual short-term memory. Acta Psychologica 75, 55–74.
Luck, S.J., Vogel, E.K., 1997. The capacity of visual working memory for features and conjunctions. Nature 390, 279–281.
Melton, A.W., 1962. Editorial. Journal of Experimental Psychology 64, 553–557.
Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–97.
Morey, C.C., Bieler, M., 2012. Visual short-term memory always requires general attention. Psychological Bulletin and Review 20, 163–170.
Morey, C.C., Cowan, N., 2004. When visual and verbal memories compete: evidence of cross-domain limits in working memory. Psychonomic Bulletin and Review 11, 296–301.
Morey, C.C., Cowan, N., 2005. When do visual and verbal memories conflict? the importance of working-memory load and retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition 31 (4), 703–713.
Peterson, L.R., Peterson, M.J., 1959. Short-term retention of individual verbal items. Journal of Experimental Psychology 58, 193–198.
Quinn, J.G., McConnell, J., 1996. Irrelevant pictures in visual working memory. Quarterly Journal of Experimental Psychology 49, 200–215.
Repovs, G., Baddeley, A.D., 2006. The multi-component model of working memory: explorations in experimental cognitive psychology. Neuroscience 139, 5–21.
Rudkin, S.J., Pearson, D.G., Logie, R.H., 2007. Executive processes in visual and spatial working memory tasks. Quarterly Journal of Experimental Psychology 60, 79–100.
Saito, S., Towse, J., 2007. Working memory as a construct in cognitive science: an illustrious past and a highly promising future. Psychologia 50, 69–75.
Salamé, P., Baddeley, A.D., 1982. Disruption of short-term memory by unattended speech: implications for the structure of working memory. Journal of Verbal Learning and Verbal Behaviour 21, 150–164.
Saults, J.S., Cowan, N., 2007. A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. Journal of Experimental Psychology: General 136 (4), 663–684.
Smyth, M.M., Pendleton, L.R., 1989. Working memory for movements. Quarterly Journal of Experimental Psychology 41, 235–250.

Shallice, T., Warrington, E.K., 1970. Independent functioning of verbal memory stores: a neuropsychological study. Quarterly Journal of Experimental Psychology 22, 261–273.

Stevanovski, B., Jolicoeur, P., 2007. Visual short-term memory: central capacity limitations in short-term consolidation. Visual Cognition 15, 532–563.

Turner, M.L., Engle, R.W., 1989. Is working memory task-dependent? Journal of Memory and Language 28, 127–154.

Towse, J.N., Hitch, G., Horton, N., 2007. Working memory as the interface between processing and retention: a developmental perspective. Advances in Child Development and Behavior 35, 219–251.

Vergauwe, E., Barrouillet, P., Camos, V., 2009. Visual and spatial working memory are not that dissociated after all: a time-based resource sharing account. Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 1012–1028.

Vergauwe, E., Barrouillet, P., Camos, V., 2010. Do mental processes share a domain-general resource? Psychological Science 21, 384–390.

Vergauwe, E., Dewaele, N., Langerock, N., Barrouillet, P., 2012. Evidence for a central pool of general resources in working memory. Journal of Cognitive Psychology 24, 359–366.

Warrington, E.K., Shallice, T., 1969. The selective impairment of auditory verbal short-term memory. Brain 92, 885–896.

Wood, J.N., 2007. Visual working memory for observed actions. Journal of Experimental Psychology: General 136, 639–652.

Zhang, W., Luck, S.J., 2008. Discrete fixed-resolution representations in visual working memory. Nature 453, 23–35.

Short-Term Memory: Psychological and Neural Aspects

Giuseppe Vallar, University of Milano-Bicocca, Milan, Italy; and IRCCS Istituto Auxologico Italiano, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Short-term memory' refers to a number of systems, with limited capacities, which secure the temporary (in the range of seconds) retention of a variety of materials. Short-term memory systems are distinct from the more peripheral and short-lasting components of echoic memory for auditory stimuli, and iconic memory for visual stimuli. The three more extensively investigated short-term memory systems include: a phonological component, concerned with the retention of verbal material; a visual component, involved in recognition memory for objects; and a spatial component, retaining information about location. The general functional architecture of short-term memory includes a store and a rehearsal process, which preserves the temporary trace. These three retention systems are functionally independent, have discrete neural correlates in the association cortical areas, and may be selectively disrupted by brain damage. Phonological short-term memory is a mainly left-hemisphere-based system, including the posterior–inferior parietal (store component), and premotor (rehearsal component) regions. Memory for spatial location is a mainly right-hemisphere-based system, comprising the association occipital, posterior parietal, and dorsolateral frontal cortices ('dorsal visual stream'). Visual recognition memory for objects is associated with a network including the association occipital, temporal, and frontal cortices ('ventral visual stream'), possibly with a greater contribution from the left hemisphere. Short-term memory systems, in addition to supporting temporary retention for a variety of purposes, participate in the long-term learning of new information. One such case is the acquisition of new vocabulary in native or second language by children and adults.

Short-Term Memory

'Short-term memory' refers to a number of systems with limited capacity (in the verbal domain, roughly the 'magical' number 7 ± 2 items, see Miller, 1956), concerned with the temporary (in the range of seconds) retention of a variety of materials. 'Short-term memory' is separate from more peripheral retention systems, which store material for a shorter amount of time, in a format that is closer to the physical features of the stimulus: acoustical echoic, and visual iconic memory. Current knowledge of the functional and anatomical organization of short-term memory in humans, and its role in cognition, is presented, drawing data from three main sources of evidence: (1) behavioral studies in neurologically unimpaired individuals and in brain-injured patients with selective deficits of immediate retention; (2) correlations between the anatomical localization of the cerebral lesion and the short-term memory disorder of brain-damaged patients; and (3) correlations between the activation of specific cerebral areas and the execution of tasks assessing short-term retention in healthy participants. The two more extensively investigated aspects of short-term memory are considered: verbal and visuospatial. 'Short-term memory' is closely related to the concept of 'working memory.' The present article focuses on the 'storage' and 'rehearsal' components of the system, rather than on the cognitive operations and executive functions currently associated with 'working memory.' However, the section devoted to the uses of short-term memory illustrates some working aspects of short-term retention.

Historical Origin of the Construct 'Short-Term Memory'

Suggestions of a distinction between two types of memory, one concerned with temporary retention, and the other having the function of a storehouse for materials, which have been laid out of sight, date back at least to John Locke's *An Essay Concerning Human Understanding* (1690). In 1895, in his *Principles of Psychology*, William James revived the distinction, suggesting the existence of a limited-capacity 'Primary Memory,' embracing the present and the immediate past, and subserving consciousness. Psychological research in the nineteenth century and in the first half of the twentieth century was, however, mainly concerned with the diverse factors affecting learning and retention, in the context of a basically unitary view of human memory. It was only in the 1950s that short-term memory became the object of systematic behavioral studies in neurologically unimpaired participants. In the late 1960s, the fractionation of human memory into a short- and a long-term system became a current view (Atkinson and Shiffrin, 1968; Figure 1).

Functional Architecture of Short-Term Memory

Evidence from Unimpaired Participants

Phonological Short-Term Memory

Three main behavioral phenomena suggest the existence of a discrete limited-capacity system, concerned with short-term retention. (1) Accuracy in recalling a short list of stimuli (e.g., trigrams of consonants or of words) decreases dramatically in a few seconds if the participants' repetition of the memory material (rehearsal) is prevented by a verbal distracting activity, such as counting backward by threes. (2) In immediate serial recall of verbal material (memory span), the participants' repetition performance is limited to about 5–7 items (i.e., the above-mentioned 'magical' number seven). Recall is affected by factors such as phonological similarity and word length: lists of phonologically dissimilar items are recalled better than lists of similar items (phonological similarity effect); lists of short words are recalled better than lists of long words (word length effect). (3) In the immediate free recall of a long sequence of

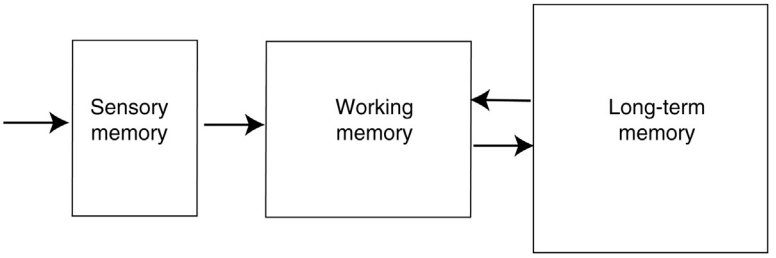

Figure 1 The two-component serial memory model of Atkinson and Shiffrin (1968).

events, such as 15–30 words, exceeding the participants' immediate span, the final five to six stimuli of the list are recalled better that the preceding ones. This 'recency effect' vanishes after a few seconds of distracting activity, and is minimally affected by factors such as age, rate of presentation of the stimuli, and word frequency.

These empirical observations illustrate the main characteristics of 'short-term memory': a retention system with limited capacity, where rapid forgetting occurs, in the time range of seconds, due to interference and temporal decay of the memory trace. Forgetting may be prevented through rehearsal. Material stored in short-term memory has a specific representational format, which, in the case of the extensively investigated verbal domain, involves phonological codes, separately from lexical–semantic representations stored in long-term memory. Each of the three above-mentioned phenomena (short-term forgetting, immediate memory span, and free recall) subsequently proved to be considerably more complex than initially thought. They were also interpreted as compatible with a unitary, single-system, view of human memory. This account ran into difficulties, however, mainly on the basis of neuropsychological evidence.

The functional architecture of phonological short-term memory has been investigated in detail using effects that break down storage and rehearsal subcomponents. The effect of phonological similarity, whereby immediate serial recall of auditory and visual verbal material is poorer for sequences of phonologically similar stimuli than for dissimilar ones, reflects the coding that takes place in the phonological short-term store. The effect of word length, whereby the immediate serial recall of auditory and visual verbal material is poorer for sequences of long words than for short ones, is held to reflect the activity of the process of rehearsal, abolished by 'articulatory suppression,' i.e., a continuous uttering of an irrelevant speech sound. Suppression, while disrupting rehearsal, also reduces immediate memory performance in span tasks. The interaction between phonological similarity, input modality, and articulatory suppression, with suppression abolishing phonological similarity only when the stimuli are presented visually, suggests that rehearsal participates in the process of conveying visual verbal material to the phonological short-term store. Finally, some phonological judgments (such as rhyme and stress assignment) are held to involve the articulatory components of rehearsal, because the performance of neurologically unimpaired participants is selectively impaired by suppression.

Also lexical–semantic representations contribute to immediate retention, particularly in immediate serial recall (i.e., span), which is the task most frequently used in order to assess short-term memory performance. It has long been known that immediate memory span is better for words than for nonwords, namely letter strings which do not possess a preexisting representation in the lexicon (effect of 'lexicality'). Also, high frequency, and imaginable/concrete words are better recalled than low frequency, abstract words. The effects of semantic relatedness or similarity are more complex, with increase or decrease of recall performance having been reported.

Visual and Spatial Short-Term Memory

In the nonverbal domain a similar distinction is drawn between visual (for objects) and spatial (for locations) short- and long-term memory systems, with the relevant representational formats being in terms of shape or spatial position of the stimulus. These systems have a limited capacity of about 3–4 items, which can be retained in the range of seconds (Zimmer, 2008). Also in the visuospatial domain, similarity, recency, and interference effects have been observed. Visual and spatial short-term memories are likely to comprise storage and rehearsal (pictorial and spatial) components. In the case of spatial locations, rehearsal may include spatial attentional mechanisms, maintaining the encoded position of the stimulus in a spatial reference frame (e.g., egocentric), as well as planned movements (e.g., ocular, manual reaching, locomotion) toward the position of the target.

Other Domains: Tactile and Olfactory Short-Term Memory

Other sensory domains have been much less investigated, as compared with visuospatial and auditory-verbal (phonological) short-term memory. Accordingly, less definite conclusions can be drawn. In the tactile/haptic domain it appears that different systems support short- and long-lasting retention, with a distinction broadly similar to that found for the phonological and visuospatial short-term memory systems (Gallace and Spence, 2009). Also for olfactory memory, a distinction between short-term and long-term memory appears to be a valuable framework (White, 2009).

Evidence from Brain-Injured Patients

Studies in patients with localized brain damage provide evidence that supports the independence of short- and

long-term memory systems, conjuring up a double dissociation of deficits. In patients with 'global amnesia,' which may be characterized as a selective impairment of the declarative or explicit component of long-term memory, verbal and visuo-spatial short-term memory are unimpaired: immediate serial span, the recency effect in immediate free recall, and short-term forgetting are within normal limits. Immediate retention may, however, depend on long-term memory when the material to be memorized exceeds the limited capacity of short-term memory, is difficult to rehearse, and rehearsal is prevented, or attention is diverted (Jeneson and Squire, 2011). This functional dissociation suggests a serial organization of the two systems, with temporary retention being a necessary condition for long-term storage. Since the late 1960s, selective impairments of short-term memory have been reported (Baddeley et al., 2002; Vallar, 2006). The more extensively investigated area concerns auditory-verbal (phonological) short-term memory.

Selective Deficits of Phonological Short-Term Memory

The pattern of impairment. These patients show a disproportionate reduction of the auditory-verbal span to an average of less than three items (digits, letters, or words); the recency effect in immediate free recall of auditorily presented lists of words is abolished; short-term forgetting is abnormally rapid. The disorder is modality-specific, with the level of performance being better when the material is presented visually (Figure 2). This input-related dissociation has two main implications: (1) discrete phonological and visual short-term memory components exist; (2) in the input–output processing chain, the phonological short-term store should be envisaged as an input system, rather than as an output buffer store. This fractionation argues against a monolithic view of the system as a single store, which is amodal, namely: not specific for the different sensory modalities, as the model shown in Figure 1. Additional support to this input locus of the system comes from the observation that, in some of these patients, speech production is entirely preserved. Most patients exhibit the detrimental effect of phonological similarity with auditory presentation in span tasks. This is in line with the view that auditory information still gains access to a phonological code, which provides a main contribution to residual immediate retention, in spite of the damage to the phonological short-term store/rehearsal system. More generally, these findings are in line with the view that auditory material has an automatic access to phonological short-term storage, which is a robust process, infrequently entirely disrupted by brain damage. With visual presentation, these patients do not recode phonologically the to-be-recalled information, and there is some evidence for visual coding and retention to occur, in a visual short-term store (see Figure 4). The effect of word length is absent, and this suggests the lack of use or damage to the process of rehearsal.

The role of lexical–semantic long-term memory representations in residual span performance. The residual span performance of patients with defective phonological memory is likely to be supported by lexical–semantic codes, as that of neurologically

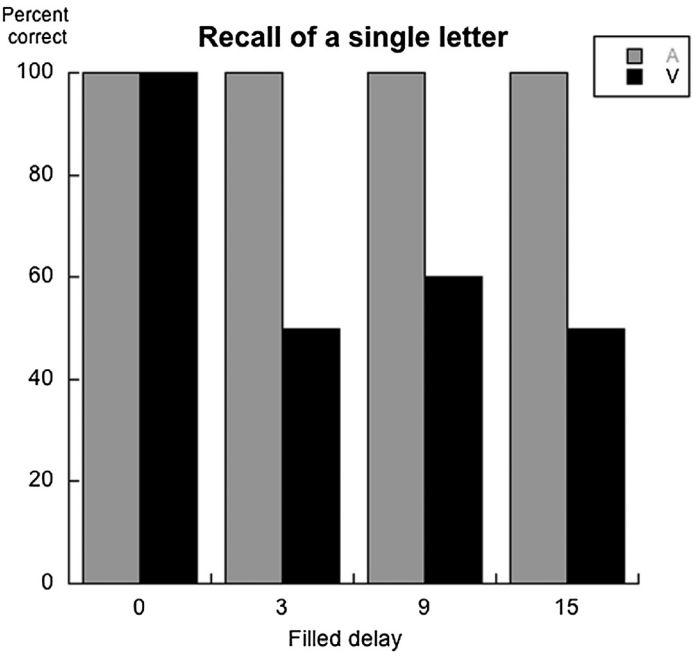

Figure 2 Short-term forgetting of a single letter by left-brain-damaged patient PV, after 3–15 s of delay filled by arithmetic distracting activity. With an auditory input, dramatic forgetting occurred within a few seconds. In the visual modality, the patient's performance was at ceiling. Immediate recall was fully accurate in both input modalities, ruling out perceptual and response-related deficits, differently from the short-term memory disorder. Data from Basso, A., Spinnler, H., Vallar, G., Zanobio, M.E., 1982. Left hemisphere damage and selective impairment of auditory-verbal short-term memory. Neuropsychologia 20, 263–274.

unimpaired participants. In a meta analysis of the auditory immediate serial recall performance of 25 patients with a defective phonological short-term memory, digit span is 2.38 (range 1–3.6), letter span 1.79 (range 1–2.5), and word span 2.00 (range 1.05), with the differences between the three types of stimuli being significant (Baddeley et al., 2002). The better serial recall of words indicates an effect of the lexicality of the stimuli. Conversely, patients with semantic dementia may show a reduction of such an effect in immediate serial recall, while showing comparable effects of phonological similarity (Jefferies et al., 2008).

Phonological short-term memory and long-term learning. Patients with defective phonological short-term memory may show unimpaired long-term verbal (Figure 3), and visuospatial learning. This observation further corroborates the complete independence of short- and long-term memory systems, but is incompatible with a serial organization – in which defective short-term memory systematically entails a long-term memory impairment, as predicted by serial models, such as the one shown in Figure 1 – suggesting a parallel architecture instead. After early perceptual analysis, information may enter short- or long-term memory, either of which may be selectively disrupted by brain damage. However, the learning abilities of these patients are dramatically impaired when the phonological material to be learned does not possess preexisting lexical–semantic representations in long-term memory. This is the case of pronounceable meaningless sequences of letters, such as nonwords, or words of a language unknown to the participant (Figure 3). This deficit of phonological learning indicates that, within a specific representational domain, temporary retention in short-term memory is necessary for a stable long-term learning and acquisition, as predicted by serial models. These findings also suggest interactions between short- and long-term memory systems, with the role of the latter supporting also immediate retention and learning, when preexisting representations of the stimulus are stored in the brain, such as in the case of words known by the participant.

Verbal rehearsal. The process of rehearsal has been traditionally described as an activity that, through rote repetition, refreshes the short-term memory trace, preventing its decay. The precise characteristics of rehearsal have been elucidated in more detail in recent years, particularly in the domain of phonological memory. Rehearsal may be conceived in terms of the recoding of the memory trace from input (auditory-verbal) to output-related (articulatory) representations and vice versa (Figure 4). More specifically, rehearsal of verbal material has long been regarded as 'articulatory' in nature, involving output-related verbal processes – such as the motor programing of speech production in a phonological assembly system or output buffer – the actual articulation of the material to be rehearsed (subvocal rehearsal), or both. Anarthric subjects, who are unable to utter any articulated speech sound, due to congenital disorders or brain damage acquired in adult age, may, nonetheless, show a preserved immediate memory, including verbal rehearsal. These patients have cerebral lesions involving the motor cortex or the descending efferent pathways, directed to neurons in the brainstem, which control the articulatory musculature. This is therefore not necessary for the operation of rehearsal, which should be conceived as a 'central'

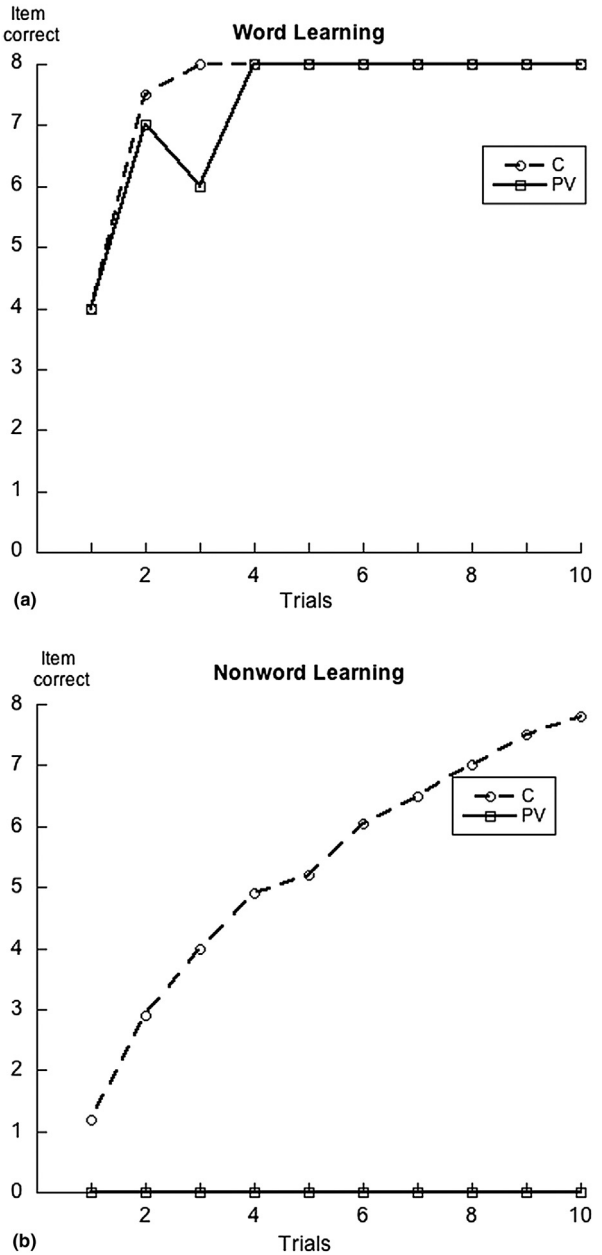

Figure 3 Word (a) and nonword (b) paired-associate learning by patient PV and matched healthy control participants (C), with auditory presentation. The patient's performance was within the normal range with native language (Italian) words, dramatically defective with unfamiliar nonwords (Russian words transliterated into Italian). Data from Baddeley, A.D., Papagno, C., Vallar, G., 1988. When long-term learning depends on short-term storage. Journal of Memory and Language 27, 586–595.

process, not requiring any peripheral muscular activity. Conversely, brain-damaged patients with a selective impairment of rehearsal, due to lesions involving the left premotor cortex, show a defective immediate verbal span, as do patients with damage to the phonological short-term store. Both components of phonological memory contribute to immediate recall, with the phonological store providing the main retention capacity. The ability to perform phonological judgments is disrupted, however, by damage to the rehearsal process, rather

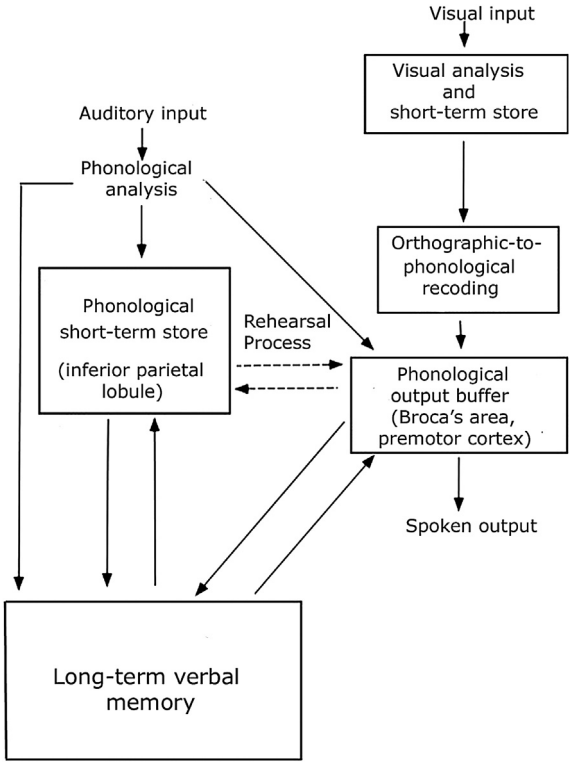

Figure 4 An anatomofunctional multicomponent model of phonological short-term memory (Vallar, 2006).

than to the phonological store. By contrast, a defective rehearsal leaves the recency effect in immediate free recall of auditory lists of words largely unimpaired. This recency effect is severely reduced or absent in patients with damage to the phonological short-term store, which is the main functional component contributing to it.

Orthographic-to-phonological recoding. Patients with defective phonological short-term memory have preserved grapheme-to-phoneme conversion skills, being able, for instance, to read aloud nonwords, namely sequences of letters that do not possess a lexical entry. Patients unable to articulate speech, due to bilateral damage selectively involving the motor cortex, may show a selective impairment of phonological recoding, with the phonological short-term store and rehearsal components being preserved.

Selective Deficits of Spatial and Visual Short-Term Memory

Brain-injured patients may show deficits of short-term retention in the visual or in the spatial domain, even though these patterns of impairment have been explored less extensively than the phonological disorder.

Deficits of spatial short-term memory. Immediate retention of short sequences of spatial locations, as assessed by a block-tapping task (a spatial analogue of digit span), may be selectively defective, with verbal and visual memory being unaffected, in both their short- and long-term components. Spatial long-term learning, assessed using a block-tapping test with a limited number of familiar positions (i.e., 9), may also be preserved. The imaginal deficit of right-brain-damaged patients with left spatial neglect (i.e., the patients' impairment to recollect the left-hand side of visuospatial images, retrieved from long-term memory) may be interpreted in terms of a unilateral disorder of visuospatial short-term memory. More generally, deficits of visuospatial short-term memory may contribute to a number of manifestations of the neglect syndrome: patients may fail to keep track of the explored positions in visual search tasks, revisiting previously explored targets. However, patients with left spatial neglect may show no deficits of visuospatial short-term memory: accordingly, the impairment of this immediate retention system, while possibly contributing to the manifestations of visuospatial neglect, and shaping them, does not appear to be a core component of the syndrome.

Deficits of visual short-term memory. In other brain-damaged patients, the deficit may involve the visual (shape) component of short-term memory, with spatial short-term memory being preserved. Impairments in the short-term visual recognition of unfamiliar faces, objects, voices, and colors have also been described. The disorder of patients with defective visual imagery (as assessed, for instance, by tasks requiring color or size comparisons) may be interpreted in terms of an impaired visual short-term memory store. A deficit of visual short-term memory may also disrupt long-term learning of unfamiliar nonverbal material, as assessed by recognition memory for unfamiliar faces and objects. This extends to the visual domain the conclusion that long-term acquisition requires short-term storage.

The Uses of Short-Term Memory

Is there a use for a system providing the temporary retention of a limited amount of stimuli, besides infrequent situations such as the following? A friend tells us an unfamiliar 8-digit number, which we have to dial on a telephone placed on the other side of a large street and we have no paper and pencil to write it down. The answer is positive. Short-term retention contributes to the stable acquisition of new information in long-term memory. More specifically, phonological short-term memory plays an important role in learning new vocabulary. It also participates in the processes of speech comprehension and production.

Long-Term Learning

The observation that patients with defective auditory-verbal span are impaired in learning unfamiliar pronounceable letter sequences gives rise to the possibility that phonological memory may contribute to a relevant aspect of language development, the acquisition of vocabulary (Figure 3). Similarly, individuals with a developmental deficit of phonological memory are impaired in vocabulary acquisition and in nonword learning. Individuals with a congenital cognitive impairment, such as Down's syndrome, provide an opposite pattern, when their deficit spares phonological short-term memory: acquisition of vocabulary, foreign languages, and

nonword learning are also preserved. Converging evidence from different populations supports this view. Correlational studies in children have shown that the capacity of phonological memory is a main predictor of the subsequent acquisition of vocabulary, both in the native, and in a second language. In normal adult participants, the variables that disrupt immediate memory span (phonological similarity, item length, and articulatory suppression) also impair the acquisition of nonwords. Polyglots have a greater capacity of phonological memory, compared to nonpolyglots, and a better ability to learn novel words. Phonological short-term memory may be considered as a learning device for the acquisition of novel phonological representations, and the building up of the phonological lexicon. A few observations in brain-damaged patients suggest a similar role for visuospatial short-term memory in the acquisition of new visual information, such as unfamiliar faces and objects.

Language Processing

The idea that short-term retention contributes to speech comprehension dates back to the 1960s. Phonological memory may withhold incoming auditory-verbal strings, while syntactic and lexical–semantic analyses are being performed. Patients with defective phonological memory show a preserved comprehension of individual words, as well as of many sentential materials, and a normal ability to decide whether or not sentences are grammatically correct. This may reflect, on the one hand, the operation of on-line lexical–semantic processes, heuristics and pragmatics, and, on the other, the complete independence of syntactic and lexical–semantic processes from phonological memory. Patients are, however, impaired by 'complex' sentences, where 'complexity' may refer to a number of nonmutually exclusive factors, such as a high speed of material presentation, which prevents the immediate build-up of an unambiguous cognitive representation; word order conveying meaning-crucial information (e.g., in sentences in which a semantic anomaly is introduced by a change in the linear arrangement of words: "The world divides the equator into two hemispheres, the southern and the northern"; the more complex sections of the Token Test, in which commands such as "touch the green big square and the small yellow circle" are given to participants); syntactic complexity, and extralinguistic presuppositions biasing the interpretation of the spoken message. Under these conditions, adequate interpretation may require backtracking to the verbatim (phonological) representation of the sentence, temporarily held in phonological memory. This provides a 'backup' or 'mnemonic window' resource for performing supplementary cognitive operations, necessary for comprehension.

Neurofunctional Architecture of Short-Term Memory

Phonological Short-Term Memory

Anatomoclinical correlation studies in brain-damaged patients with a selective impairment of auditory-verbal span indicate that the inferior parietal lobule (supramarginal gyrus) of the left hemisphere, at the temporoparietal junction, represents the main neural correlate of the 'store' component of phonological short-term memory (Vallar, 2006). The frontal premotor regions in the left hemisphere and other structures such as the insula are the major neural correlates of the 'rehearsal' component, even though the available anatomoclinical data are more limited (Figure 3).

Functional neuroimaging studies in normal participants concur with this pathological evidence, to suggest a left-hemisphere-based network. Activation in the left supramarginal gyrus (Brodmann's area (BA) 40) is associated with the 'store' component of short-term phonological memory, activation in the left frontal premotor BA 44 (Broca's area) and BA 6, and in the left insula, with the 'rehearsal' component (Paulesu et al., 1996). In these studies, in line with the behavioral evidence from unimpaired participants and brain-damaged patients, an immediate verbal span task activates both the inferior parietal region (phonological short-term store) and the premotor cortex (rehearsal process) in the left hemisphere. Conversely, rhyme judgments selectively activate the left premotor regions, whose damage, in turn, disrupts the patients' ability to perform this task. These activation and lesion-based data support, from an anatomofunctional perspective, the behavioral distinction between a 'storage' component and a 'rehearsal process' in phonological short-term memory. Furthermore, they qualify 'rehearsal' as a process making use of components (the 'phonological output buffer' of Figure 4) also concerned with the planning (i.e., programing in the left premotor cortex) of articulated speech. Seen in this perspective, phonological memory may be regarded as a component of the language system. Meta analyses of imaging studies in unimpaired participants involved in tasks requiring immediate retention of verbal material support a major contribution from the left hemisphere, including also subcortical structures (thalamus and basal ganglia); right hemisphere regions are also involved. However, when the retention task becomes more complex, with greater executive demands, the hemispheric asymmetry is no longer present (Wager and Smith, 2003).

Visual and Spatial Short-Term Memory

Studies in brain-damaged patients suggest an association between damage to the right hemisphere (and, to a lesser extent, to the left hemisphere) and defective short-term memory for spatial locations, with single items, or sequences of stimuli. There is also some, although less definite, evidence suggesting that damage to the posterior regions of the left hemisphere brings about disorders of visual short-term memory, such as defective immediate retention of sequences of visual stimuli (e.g., lines), and impaired recognition of more than one visual stimulus at a time (defective simultaneous form perception) (Baddeley et al., 2002).

Neuroimaging activation studies in humans support the distinction between primarily spatial (a location, a 'where' component) and visual (object) short-term memory systems (Wager and Smith, 2003; Zimmer, 2008). The main neural correlates of spatial memory for locations include the occipital extrastriate, posterior parietal, dorsolateral premotor and prefrontal cortices. Short-term visual object recognition

memory is associated with activations in a network including the occipital extrastriate, temporal, and ventrolateral premotor and prefrontal cortices. The prefrontal components may also differ in the type of processes they perform: the ventrolateral component in maintenance and retrieval, the dorsolateral component in monitoring and more active processing. Right hemisphere areas (posterior parietal, premotor, and prefrontal) may play a more relevant role in spatial memory for locations. The left inferior-temporal cortex may be particularly involved in visual object short-term memory. These patterns of activation indicate an association of the 'dorsal visual stream' with spatial short-term memory for locations and of the 'ventral visual stream' with short-term visual recognition memory for objects, although their hemispheric lateralization is controversial.

Neurofunctional Models of Short-Term Memory

Before the late 1960s, the prevailing view was that memory is a unitary system. Its fractionation into subcomponents started with the short-term/long-term distinction, such as that of the functional model of Atkinson and Shiffrin (1968), shown in Figure 1. In the following years, multicomponent models of memory were further developed, including subsystems both within long- and within short-term memory. As for short-term memory, subcomponents such as those shown in Figure 4 were specified, in both their functional architecture and neural underpinnings, with a main distinction involving store vs rehearsal subsystems, for verbal, spatial, and object (visual) domains.

This multicomponent view went not unchallenged, due to a resurgence of basically unitary models of memory. One such model is shown in Figure 5 (Ricker et al., 2010). Short-term memory is conceived as a subset of long-term memory, activated by attention and executive processes. The relationship between attention and immediate retention has been a long-standing issue. A unitary position may readily account for the contribution of long-term memory systems to immediate retention, with memory performance in any given specific task resulting from the contribution of the activated neurofunctional subset. Also multicomponent models, however, are open to interactions with long-term memory systems, as the one shown in Figure 4. Furthermore, lesion data from brain-damaged patients suggest that specific cortical regions of the brain are involved in immediate retention. These may be well conceived as an activated subset of long-memory, but then the differences between the unitary and the multicomponential approaches become unclear. This issue is illustrated in Figure 5, in which the store and rehearsal components of short-term memory are shown as specific (both at the functional and the neural levels) activated subsets. In addition, the phonological short-term store is conceived as a passive and automatic system, which needs little attentional resources in order to operate, as also shown by the robustness of the detrimental effects of phonological similarity on immediate serial recall performance in a variety of pathological conditions, ranging from selective deficits of phonological memory (Vallar, 2006), to semantic dementia, transcortical sensory aphasia (Jefferies et al., 2008), and the disexecutive syndrome.

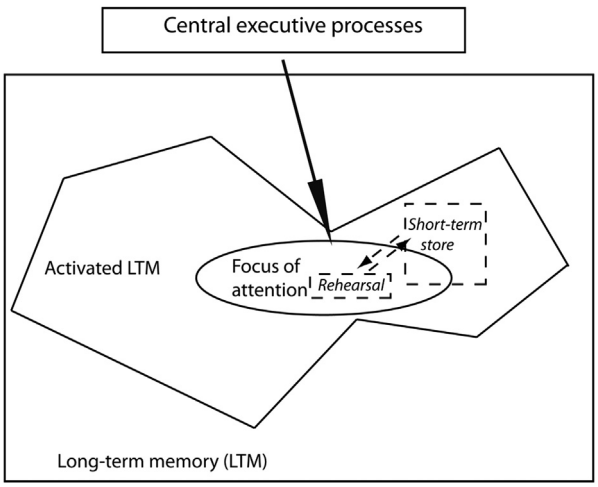

Figure 5 A functional model of short-term memory as a subset of long-term memory, temporarily activated by attentional and executive processes (Ricker et al., 2010). The figure shows the phonological short-term store and articulatory rehearsal components of short-term memory as activated subsets.

In sum, while immediate memory performance depends on the contribution of different resources (in the verbal domain, phonological, lexical–semantic; in the visual domain, perceptual primitives, stored representations of objects; in the spatial domain, egocentric and allocentric reference frames), this does not necessarily imply that the system is unitary. The specificity of the syndromes of defective short-term memory (phonological, spatial, and visual) both behaviorally, and in terms of neural correlates, supports the view that this retention system should be conceived as a set of independent, though related, components, interacting with long-term memory, and with executive, and attentional resources.

'Short-term' memory, as discussed and illustrated so far, overlaps in part with 'working' memory (Aben et al., 2012), with this close relationship having been set since the development of the first models in the later 1960s and early 1970s, distinguishing 'short-' vs 'long-term' storage systems (Atkinson and Shiffrin, 1968). As outlined in this article, 'short-term' memory comprises a number of storage components, and specifically dedicated rote rehearsal and recoding systems, which, in the case of phonological short-term memory, may also be conceived as components of the language system. 'Working memory,' conversely, includes a number of more controlled processes (elaborative rehearsal, coding decisions, and retrieval strategies) which operate on the content of the different 'short-term' stores. Seen in this perspective, 'short-term' memory stores, and dedicated rehearsal processes, could be considered as subcomponents of a 'working' memory system (Baddeley, 2012). 'Working' memory may be also seen as a set of specific executive and control processes (among other systems, e.g., the programing of complex action), which operate on material temporarily held in different limited-capacity 'short-term' stores, with the contribution of rote rehearsal processes. More generally, the 'working' memory processes may be conceived as components of an executive system, mainly based in the

prefrontal cortex, involved in setting up complex, purposeful, and organized behavior (Stuss and Knight, 2012).

Conclusion

Behavioral observations and neuroanatomical evidence from normal participants and brain-injured patients concur in suggesting that 'short-term memory' should be conceived as a multiple-component system with specific functional properties and discrete neural correlates. These systems secure the retention of a limited amount of material in the time range of seconds and contribute to relevant aspects of cognition, such as the long-term learning of novel material, of which no preexisting representations are stored and available. Short-term memory systems closely interact with long-term memory, executive, and attentional processes in supporting immediate retention performance.

Main short-term memory systems include verbal, visual (object), and spatial components, with store and rehearsal processes. The brain bases of short-term memory are mainly cortical regions: (1) a posterior parietal–dorsolateral frontal network for the verbal system, with a major role of the left hemisphere; (2) an occipital extrastriate-posterior parietal–dorsolateral frontal network for the spatial (location) system (dorsal stream), with a major role of the right hemisphere; and (3) an occipital extrastriate-temporofrontal network for the visual (object) system (ventral stream). These networks are largely distinct from those involved in long-term memory performance, which include the medial temporal lobe, diencephalic nuclei, and their connections.

See also: Amnesia: General; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Consciousness, Neural Basis of; Declarative Memory, Neural Basis of; Episodic Memory; Hemispatial Neglect, Neural Basis of; Learning and Memory, Neural Basis of; Olfactory System; Semantic Memory; Speech Perception; Speech Production, Psychology of; Tactile Processing; Visual Imagery; Visual Memory, Psychology of; Visual Streams: Dorsal and Ventral; Working Memory, Neural Basis of; Working Memory, Psychology of.

Bibliography

Aben, B., Stapert, S., Blokland, A., 2012. About the distinction between working memory and short-term memory. Frontiers in Psychology 3, 301.

Atkinson, R.C., Shiffrin, R.M., 1968. Human memory: a proposed system and its control processes. In: The Psychology of Learning and Motivation. Advances in Research and Theory. Academic Press, New York, pp. 89–195.

Baddeley, A.D., 2012. Working memory: theories, models, and controversies. Annual Review of Psychology 63, 1–29.

Baddeley, A.D., Kopelman, M.D., Wilson, B.A., 2002. The Handbook of Memory Disorders, second ed. Wiley, Chichester, England.

Gallace, A., Spence, C., 2009. The cognitive and neural correlates of tactile memory. Psychological Bulletin 135, 380–406.

Jefferies, E., Hoffman, P., Jones, R., Ralph, M.A., 2008. The impact of semantic impairment on verbal short-term memory in stroke aphasia and semantic dementia: a comparative study. Journal of Memory and Language 58, 66–87.

Jeneson, A., Squire, L.R., 2011. Working memory, long-term memory, and medial temporal lobe function. Learning and Memory 19, 15–25.

Miller, G.A., 1956. The magic number seven, plus or minus two: some limits to our capacity for processing information. Psychological Review 63, 81–93.

Paulesu, E., Frith, U., Snowling, M., Gallagher, A., Morton, J., Frackowiak, R.S.J., Frith, C.D., 1996. Is developmental dyslexia a disconnection syndrome? Evidence from PET scanning. Brain 119, 143–157.

Ricker, T.J., AuBuchon, A., Cowan, N., 2010. Working memory. In: Nadel, L., (Ed.), Wiley Interdisciplinary Reviews: Cognitive Science 1, 573–585.

Stuss, D.T., Knight, R.T. (Eds.), 2012. Principles of Frontal Lobe Function, second ed. Oxford University Press, New York.

Vallar, G., 2006. Memory systems: the case of phonological short-term memory. A festschrift for cognitive neuropsychology. Cognitive Neuropsychology 23, 135–155.

Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive Affective and Behavioral Neuroscience 3, 255–274.

White, T.L., 2009. A second look at the structure of human olfactory memory. Annals of the New York Academy of Sciences 1170, 338–342.

Zimmer, H.D., 2008. Visual and spatial working memory: from boxes to networks. Neuroscience and Biobehavioral Reviews 32, 1373–1375.

Relevant Website

http://www.memory-disorders.org/ – The Memory Disorders Research Society.

Shrinking Cities

Robert Beauregard, Columbia University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Persistent population loss from cities around the world has given rise to a new research and policy initiatives meant to combat the consequences of depopulation. This new perspective replaces an earlier concern with urban decline while deemphasizing large-scale, commercial redevelopment projects as an antidote to the lack of growth. This article reviews the origins of the shrinking city idea and surveys the recent and continued depopulation of cities in the United States, Europe, and Asia. It then considers various policy responses that focus on addressing the twinned problems of abandoned housing and the reuse of once-occupied sites, often via efforts to make the city more environmentally sustainable. After discussing arguments for shifting city policy from the pursuit of growth to 'smaller is better,' the cases of Detroit (USA) and Leipzig (Germany) are discussed.

Through the centuries, the populations of cities and towns have expanded and contracted. Until the midtwentieth century, the overall trend was one of growth. Just after World War II, however, large cities (particularly those with manufacturing economies) began to shed residents, businesses, and jobs in alarming numbers. Certain of these cities, places like Cleveland and St. Louis in the United States, did so uninterrupted for decades. The preferred term for this condition was 'urban decline' (Beauregard, 2003). By the end of the century, however, urban living (at least in the United States) had become attractive once again and a new, less-freighted term was needed. Still losing residents, these cities were no longer 'declining,' they were 'shrinking.'

Origins

The term 'shrinking cities' became prominent around 2000 when a group of scholars applied the label to east German cities. The phrase echoed a 1980s discussion of declining industrial cities in the Ruhr area and received international publicity from a project – *Schrumpfende Stadte* – initiated by Philipp Oswalt (2005, 2006). That project gathered descriptions of such shrinking cities as Manchester (UK), Detroit (USA), and Halle/Leipzig (Germany) and produced reports, videos, books, and exhibitions. At approximately the same time, Karina Pallagst at the University of California: Berkeley organized a conference on the topic that launched the Shrinking Cities International Research Network (Martinez-Fernandez et al., 2012). These activities were soon followed by initiatives in Germany (e.g., the 5-year Shrink Smart research consortium funded by the European Union) and in the United States, notably the 'Defining the Future for America's Cities' conference in 2011 in Detroit under the auspices of The American Assembly at Columbia University (New York City) (Mallach, 2012a).

Additional attention was focused on shrinking cities in the United States by the National Vacant Properties Campaign (NVPC) (Leonard and Schilling, 2007). NVPC was later subsumed under the Center for Community Progress (CCP) in Washington, DC, and also gave rise to the Vacant Property Research Network. The CCP works with both governmental and nongovernmental stakeholders to 'reintegrate vacant, abandoned, and blighted properties into the economic and civic life of their communities.' Various pieces of legislation were introduced in Germany (e.g., Stadtumbau Ost) and the United States (e.g., the Neighborhood Stabilization Program of 2008) and this reinforced the belief that shrinking cities, if not qualitatively distinct from declining cities, had to be approached differently. By 2013, a robust network of international researchers and policy activists had emerged to share ideas across national and city boundaries.

Depopulation

Central to the shrinking city discussion is population loss. Undeniable was that while cities had cast off residents in the past, the losses after the 1950s were prevalent across certain types of cities, unprecedented in scale, and (most importantly) persistent. In the United States, substantial population loss from the older central cities began in the 1950s and reached its peak in the 1970s (Beauregard, 2009). During that decade, New York City had an unprecedented drop-off of 824 000 residents (10.4%) with no other city in the United States even close to matching that number – the city second in population loss, Chicago, had a decline of 362 000 residents (10.7%). Over two-thirds of the 50 largest cities lost population in the 1970s. Between 1950 and 1980, the large cities that shrunk experienced an average loss of 72 000 residents while between 1980 and 2000 that number decreased but was still a significant 50 000 people. Between 1950 and 2010, Flint (Michigan) became smaller by 61 000 residents (37%), while after 1990 New Orleans (Louisiana) declined in population by 29%, Cleveland (Ohio) by 18%, and Rochester (New York) by 11%. Urban decay, suburbanization, deteriorated public services, and racial conflict were key contributors to the out-migration of households. Plant closures in heavy manufacturing and the decentralization of light industry, initially to the suburbs but later from the northeast and Midwest regions to the south and west regions, further contributed to the diminishing attractiveness of these places.

In the United Kingdom, population loss was a serious problem beginning in the 1950s. Manchester's peak population occurred in 1931 while Birmingham peaked in 1951 and Sheffield in 1981 (Turok and Mykhnenko, 2007). Thereafter, they all lost residents with Manchester having a 47% decline, down 393 000 residents from 1931. In Europe, population loss from shrinking cities was in part a consequence of the devastation of World War II with industrial restructuring (particularly involving heavy manufacturing and coal mining) also influential. Central to the recent shrinking city phenomenon in Europe was the dissolution of the Soviet Union in 1989 and the breakdown of state socialism that led to a massive migration of people both from the low-wage Communist to high-wage Western countries and, within the postsocialist countries, from rural areas and small and medium-size cities to the major cities such as Prague (Mykhenenko and Turok, 2008; Wiechmann, 2008). The result was a landscape of urban shrinkage that mostly overlapped the areas where manufacturing had anchored the regional economy. An additional contributing factor was the drop in fertility rates below population replacement levels.

Outside the postsocialist world, and between 1970 and 2005, Torino lost 35% of its manufacturing jobs while Sheffield experienced a 74% decline, with corresponding population losses. Halle (Germany), Budapest (Hungary), Bytom (Poland), Torino (Italy), Donetsk (Ukraine), Glasgow (Scotland), and Murmansk (Russia) also shed large numbers of residents (Power et al., 2010). Bologna in Italy lost 24% of its population after 1971, Porto in Portugal fell 20% in population after 1981, and Murmansk in Russia lost a quarter of its residents in the 15 years prior to 2005. In cities where shipping and port activity declined, such as Marseilles (France) and Genoa (Italy), population losses followed. Compared with the United States, and speaking generally, depopulation was not severe, albeit with exceptions like Manchester in England.

Beyond Europe and Russia, former manufacturing cities and cities dependent on a particular industry (whether manufacturing such as steel or resource extraction such as coal mining) experienced a similar loss of residents. Sudbury in Canada, Mount Isa in Australia, Zhouzhuang in China, and Yubari in Japan are just a few of these shrinking cities. In Japan, demographic change has been a major factor in long-term shrinkage while economic instabilities have been more important as regards short-term shrinkage. Between 1990 and 2005, 30 of the 47 prefectures in Japan experienced depopulation. Cities with fewer than 200 000 residents were most prone to shrinkage with company towns such as Hitachi and Muroran particularly vulnerable. The exceptions were the medium-size cities located in the suburban ring of metropolitan areas. Overall, shrinking cities are most often found in declining industrial regions, although sprawl development and suburbanization can also be detrimental for city population growth and, as in the United States, often coincide with industrial restructuring.

While cities like Cleveland (Ohio) and Camden (New Jersey) in the United States and Manchester and Liverpool in the United Kingdom led the way in the period of urban decline, by the 1990s, cities in the United Kingdom had begun to rebound while those in the United States continued to cast off residents and jobs (Beauregard, 2009). In the United States, places like Boston (Massachusetts) and Washington, DC, suffered population losses prior to 1990 and prospered thereafter. Other cities – manufacturing centers with significant coastal or inland ports – continued to shrink. Pittsburgh, Detroit, Cleveland, St. Louis, Baltimore, Philadelphia, and Buffalo topped the list of big cities. Not only had these cities shed residents steadily from 1950 to 1990, they continued to do so after that time. Baltimore, for example, had a 22.5% net lost of residents from 1950 to 1990 and then another 11.5% decline (approximately 85 000 people) between 1990 and 2010. Of the medium-size and small cities, Flint (Michigan), Camden (New Jersey), Chester (Pennsylvania), Youngstown (Ohio), and Rochester (New York) have been the most prominent of those that have become smaller.

To the extent that the collapse of heavy manufacturing and resource extraction were main contributors to population loss (whether in the era of decline or shrinkage), the shrinking cities are concentrated in specific regions within these countries. The industrialized Ruhr area was the site of many of Germany's shrinking cities. In the United States, it was the Midwest region and parts of the northeast around the Great Lakes. And while the Midwest did not suffer an absolute population loss between 1950 and 2010, it grew slower than the west and south regions and went from having 29% of the country's total population to 22%. In Poland, cities in the coal mining and steel-making region of Silesia were particularly vulnerable. And, in England, the cities in the manufacturing and coal mining regions of the north were the major sites of urban shrinkage. In East Germany, the population went from 18.2 million residents in 1990 to 17.0 million in 2002 due to manufacturing decline and the reunification that enabled east Germans to chose where to live. Moreover, certain sparsely populated and agricultural or natural resource regions also lost population. These were areas like northern Finland, eastern Scotland, and northern Sweden.

Although many central cities in the United States have been shrinking, their metropolitan or urban regions have not. Since 1970, about 1 of 10 metropolitan areas have lost residents in any given census enumeration period, with most of them having only episodic rather than persistent losses. The metropolises of Buffalo (New York), Dayton (Ohio), Pittsburgh (Pennsylvania), and Youngstown (Ohio), however, have all been shrinking, as have their central cities. In about one-half of the instances of metropolitan depopulation, the overall population loss was solely a matter of central city population loss. As a generalization, suburban areas tend to grow in metropolitan areas of the United States even if their central cities do not. Nonetheless, a small number of shrinking metropolitan areas have experienced suburban population losses as well.

In Europe, conditions have been somewhat different. Over one-half of all functional urban regions (roughly equivalent to metropolitan areas of the United States) in the European Union had lost population between 1996 and 2001. Suburbanization, however, was less of a factor than deindustrialization. With suburbanization, the city's dominant city often loses residents, whereas industrial collapse coupled with German reunification and the demise of the Soviet Union has led to whole functional urban regions becoming smaller in population.

Overlapping the shrinking cities' phenomenon was the bursting of the housing bubble in 2008, a condition that occurred in numerous countries. During the real estate boom of the early to mid-2000s, with investment capital readily available, developers (and governments) built large residential communities and new towns. Inexpensive and easily obtainable home mortgages, heavy borrowing by homeowners against their home equity, exotic and untested mortgage instruments, speculative anticipation of property value appreciation, and unscrupulous lending practices fueled building activity. With many households having taken on mortgages that were financially untenable if the economy failed to grow rapidly, a bursting of the housing bubble was inevitable. Banks withdrew investment capital and large projects were halted in midconstruction. Mortgage foreclosures increased dramatically, particularly in older cities with weak housing markets but also in rapidly growing metropolitan areas like Phoenix in the United States (Immergluck, 2009). Cleveland and Detroit were particularly hard-hit, leading to further blight, abandonment, and population out-migration.

Shrinkage, then, was not just a product of manufacturing decline, suburbanization, or demographic collapse, but was also brought about by real estate speculation and unscrupulous lending practices. In Spain and Ireland, in the southwest region of the United States, and even in Inner Mongolia (where the government had built new towns), many real estate ventures simply collapsed, failing to attract homebuyers and essentially going into default. Ordos in Inner Mongolia and the Cheng-gong district in Kunming in the Yunnan province in China were two of the most prominent of these developments. These places earned the label 'ghost cities.' Even New York City felt the consequences of the 2008 economic recession with nearly 700 construction projects 'stalled' and thus empty in 2010.

Policy Response

Research and policy on the decline and adaptation of shrinking cities tends to focus on the significant problems posed by abandoned and derelict housing, thereby turning the shrinking city issue into a debate about distressed neighborhoods (Billitteri, 2010; Dewar and Thomas, 2013; Schilling and Mallach, 2012). All of the shrinking cities in the United States and Western Europe have experienced significant numbers of vacant and abandoned homes. (In Eastern Europe, depopulation has mainly reduced overcrowding.) Often clustered in specific (usually low-income and/or formerly industrial) neighborhoods, vacant and abandoned buildings pose health and fire hazards and detract significantly from the public image of the city and the potential reuse of the land on which these buildings sit. Numerous cities have responded with extensive housing demolition efforts, setting goals in the tens of thousands over a short time span. Baltimore (Maryland) in 2002, undertook to demolish 5000 houses per year over 5 years, hoping to reduce its stock of 40 000 vacant buildings. Philadelphia's Neighborhood Transformation Initiative, launched in 2001, demolished 14 000 abandoned buildings (Gurwitt, 2002). Despite such efforts, these shrinking cities still had large numbers of derelict and abandoned homes as residents continued to leave. Moreover, one less-than-desirable consequence of their efforts was block after block of vacant lots, often weed-strewn and unsightly.

In many of these cities, the local housing stock not only consisted of a high percentage of vacant and abandoned buildings but was also a target of speculators who bought properties cheaply in the hopes of taking in rents (without making any repairs or renovating them) or quickly selling them at a profit (Mallach, 2012b). Such speculation encouraged existing residents to sell, and they often then left the city. It also prevented the city government from taking control over these properties and sealing them to prevent vandalism and fires or managing demolition and subsequent clean-up of the site. Local governments were further constrained by the lack of efficient and effective legal tools to acquire and dispose of vacant properties.

To this end, a few, pioneering local governments began to experiment with new models of land banking that enable a central authority to acquire vacant properties before they are purchased by speculators (Mallach, 2004, 2012b). While cities such as Atlanta and St. Louis had chartered land banks in the 1970s and 1980s, more comprehensive land bank models emerged in the 2000s. In Michigan, a reform of municipal tax foreclosure laws gave county-level governments the ability to create the next generation of land banks. With these new powers, land banks could undertake a broader range of activities such as purchasing tax-delinquent abandoned properties, demolishing the buildings, and managing the subsequent reuse, sale to a private owner or transformation of the lot into a community garden. The most famous of these land banks in the United States is the Genesee County Land Bank Authority that took on the responsibility for managing Flint, Michigan's abandoned housing stock. Nevertheless, some local officials remain averse to demolishing usable properties and are hopeful that buyers can be found to inhabit them. This is seldom the case.

Demolition has also played a significant role in Europe for adapting to smaller populations (Bernt, 2009; Wiechmann and Pallagst, 2012). Many former East German cities made the removal of vacant buildings a major redevelopment goal. Of particular challenge in this regard has been the large, Communist-era, housing estates (such as Gruneau outside Leipzig) where ownership changes and out-migration resulted in buildings with little to no financial value. Approximately 8000 flats were demolished in Leipzig and another 7600 in Gruneau, with the latter aimed at reducing densities so that the estates could remain viable.

Housing demolition initiatives in the United States were done in the hope that either investors would then purchase the now-empty lots for new residential and commercial developments or community development corporations (CDCs) would acquire vacant lots for affordable housing or community gardens. By the 1990s, though, it was clear that most private investors had little interest in doing so while CDCs lacked the resources or organizational capacity. City officials and community groups began to grapple with two, harsh facts: first, their cities were unlikely to attract significant numbers of tax-paying and consumer-spending households or of businesses that could provide jobs and pay taxes to the local government and, second, that heavily subsidized, large-scale economic or real estate development projects were

unlikely to reenergize the local economy. Growth was not a presumable solution; stabilization was the more pragmatic goal.

The United States' situation differed from that in the United Kingdom and Germany. In the cities of the United Kingdom, demolition was meant to remove substandard housing and to build lower density housing on the now-vacant sites. As for Germany, Leipzig is an example of a city where demolition was used to balance supply and demand; by 2013, the city was beginning to repopulate its housing stock. And, in Eastern Europe much of the demolition was directed at derelict structures remaining from the socialist era.

Consequently, and primarily for the United States, a key difference between a declining city perspective and a shrinking city perspective was that city officials, community development leaders, and local elites began to talk about the possibility that their city would simply have to adapt to being smaller (Dewar et al., 2013). (By contrast, city policy in Leipzig was always about attracting new investment and expanding the population.) Leading the way in this discussion was Youngstown (Ohio) whose 2005 vision statement and citywide plan recognized that the city was unlikely to experience significant population growth and thus any planning for the future had to think in terms of less development than it previously had (Mock, 2011; Swope, 2006). Even in Youngstown, the mayor under whom the plan had been written maintained his public commitment to growth once the city had been stabilized. Most scholars, researchers, and policy analysts took the position that these cities had to 'think smaller.' Large-scale, ostensibly catalytic projects requiring major capital investment (and deep public subsidies) were viewed with a new skepticism. For elected officials and owners of downtown properties and businesses, such thinking went against their interests. It was difficult for them to imagine a smaller city that would also be prosperous.

For the 'smaller is better' advocates, having fewer residents with more vacant land and less dense development was an opportunity to become more sustainable (Krohe, 2011; Rugare and Schwarz, 2008). Detroit is reported to have 20 mi^2 of vacant land, almost all of which was once occupied by homes, factories, and stores. The vacant land could be used for community gardens, absorb rainwater, support small forests that could counteract air pollution, or be planted to provide noise buffers between residential neighborhoods and industrial areas. Large areas could be allowed to 'go wild' – back to nature – thus supporting birds and animals. Contaminated sites could be replanted with vegetation that 'cleaned' the soil. Buried streams could be 'day-lighted' to channel water run-off. Retention ponds could be built to divert storm water from the sewer system. Greenways – linear parks – could be fitted into the remaining urban fabric. By managing the vacant land, the city could become less dense in a way that fell between an 'urban' and a 'suburban' landscape while being ecologically sustainable.

A good example of this approach is "Reimagining a More Sustainable Cleveland," a project that began in 2007. Its goal was to devise strategies for making sustainable reuse of vacant properties, by positioning these sites within the city's larger, ecological processes. Of the many, community-initiated projects it funded were a vineyard, pocket parks, rain gardens, and neighborhood pathways. High on the list of goals were carbon sequestration, soil restoration, biodiversity, and heat island mitigation.

At the same time, elected officials and government administrators grapple with the problem of how to adapt the infrastructure and public services of the city to its smaller population. Local governments in the United States shrinking cities suffer from falling property values and declining tax revenues and yet have to maintain water mains, streets, schools, and fire stations that are overcapacity for the current population size. Cities do not shrink evenly with one area of the city emptying-out before another begins to lose residents and businesses. Rather, the shrinkage is geographically uneven, with areas only partially abandoned. This forces city governments to provide services in sparsely populated neighborhoods; that is, neighborhoods with fewer people than could be served by the existing infrastructure whether fire stations or sewer lines. One response is to withdraw services (e.g., close the elementary school), while another is to move from universal provision to a fee-for-service delivery model (e.g., having homes and businesses pay individually for the trash that they want removed).

Detroit

Scholars, activists, and practitioners working on shrinking city issues generally agree that stemming population loss, housing demolition, ecological reuse of vacant land, and infrastructure and public service resizing need to be addressed in a coordinated fashion; that is, from a planning and urban design perspective. The Youngstown master plan was a first effort, and in 2013 the Detroit government unveiled the most comprehensive of these endeavors, *Detroit Future City*. A product of the Detroit Works Project, and totally funded by philanthropies, it attempted to align the city's assets with its opportunities, essentially acknowledging the current situation of severe population and job loss and setting stability (not growth) as a goal. The plan further argued that the smaller population might be less important than the diversity of residents and the robustness of the job base. (Detroit still has jobs, but most of them are taken by workers who lived in the suburbs.) It proclaimed that "Detroit can be a vibrant city of 700 000 people or less if deliberate actions are taken to increase family wealth and the earning power of people who are now in poverty."

Detroit occupies a central position in the public debate about shrinking cities (Berkooz, 2010; Galster, 2012; Ryan, 2012). Not only has it experienced severe population loss, falling from a population of 1.85 million in 1950 to 713 000 in 2010, it is often touted as the city with the worst housing stock, the greatest expanse of vacant land, a history of dysfunctional local government, and the most devastated industrial landscape of any shrinking city in the country. A survey in 2009 recorded 33 000 abandoned homes and 90 000 vacant lots. With the decentralization of the automobile industry, the economic base of the city had collapsed. In the 1940s and continuing for decades, racial tensions and overt conflict led to the out-migration of many white households. Blocked from relocating to the suburbs, African–Americans subsequently became geographically concentrated in the city. Detroit became a predominately African–American city with a large proportion of poor people and a government struggling to manage public services and

avoid bankruptcy. In 2013, with fiscal insolvency imminent, the Michigan state government put Detroit in receivership and took control of the city's finances. Flint (Michigan) was already under state financial management.

One source of hope for Detroit, at least as reported in the popular media, was the influx of young artists in search of cheap housing and studio space (Herscher, 2012). A proposal to build a light-rail line from the central business district to the northern suburbs was intended to increase property values (and eventually tax revenues) along a major transportation corridor, but was later changed to a less expensive rapid-bus service. The city had earlier built new stadia for professional baseball and football teams downtown and was successful in attracting gambling casinos, with the latter becoming major contributors to the city's tax revenues. The impact of these large-scale efforts on the city's economy and its residents has been minimal. In 2013, poverty was still entrenched and the city's fiscal stability still elusive. Detroit also attracted photographers fascinated by the industrial devastation, collapsing homes, abandoned schools, and once-occupied but now-vacant lots. The resultant photographs were labeled 'ruins porn' and can be found all over the web.

Detroit was not alone among shrinking cities in the United States in undertaking ostensibly catalytic projects to bolster the local economy (thereby increasing job opportunities and encouraging capital investment) and retain existing residents while attracting new ones. Pittsburgh built new baseball and football stadia, expanded its convention center, and established a downtown arts district anchored by a museum and a symphonic hall. Philadelphia subsidized new residential development in areas where it had undertaken demolitions, launched a major initiative to revitalize the riverfront, transformed a former shipyard into an area of offices and residences, and provided support for expanding its museum district along one of its historic avenues. Washington, DC, built a new baseball stadium on its riverfront and slated the surrounding area for residential and commercial development. It also established a new office district behind its central rail station. With a thriving governmental sector attracting young people to the city, the government made every effort to support and publicize its gentrifying neighborhoods. In all instances, the leadership of these shrinking cities did not fully abandon the belief that large-scale investment would bring back prosperity. Even in cities with vast tracts of vacant and abandoned housing, city governments still adopted policies and distributed subsidies to enable private sector and nonprofit developers to build new housing.

Leipzig

The other side of the shrinking city narrative is occupied by Leipzig in the former East Germany, a city whose population dropped 7.0% (37 000 people) between 1989 and 2001 (Bontje, 2004; Power et al., 2010: pp. 107–129). Here, the focus of scholars and policymakers is not the devastation, but the ability of the German government, supported by European Union funding, to respond to shrinkage. Leipzig suffered both from industrial collapse and suburbanization and, as its population fell, high vacancy rates in the housing sector. Various governmental programs renovated existing housing, developed new uses for industrial buildings, turned vacant land into walking trails and linear parks, and built in-fill residential projects to stabilize neighborhoods. The government also supported development in the central business district with, among other projects, a new shopping mall in a modernized train station, a new building for its internationally recognized trade fair, and an expanded and enhanced airport. On the city's fringe, former coal mining sites were turned into recreational areas, initiatives that extended the German government's efforts in the 1990s to remediate and redevelop the Emscher River valley in the former industrial region of the Ruhr, an initiative that produced the well-known Emscher Landscape Park and Duisburg-Nord Landscape Park. European shrinking cities were less damaged by population loss and seemed better able to manage the consequences of depopulation.

Leipzig was not the only shrinking city in Europe engaged in various redevelopment and economic development initiatives to stem population loss, reenergize the local and regional economies, and make itself more competitive for capital investment. Bilbao in Spain became world-famous after the construction of its Guggenheim Museum designed by the famous architect Frank Gehrey. The museum boosted tourism and changed the image of this once-industrial city. Belfast in Ireland in the 1990s embarked on a regeneration of its mostly derelict and environmentally contaminated port area and established a science park to attract high-technology businesses. Waterfront regeneration, airport expansion, public transportation investments (e.g., a new metro in Bilbao), and technology parks as well as neighborhood renewal and new housing became the major tools for controlling shrinkage.

Conclusion

While New York, Shanghai, London, Sao Paulo, and Amman are struggling to attend to the problems of growth, other cities across the globe are devising policies and programs for managing depopulation. Faced with the prospect of minimal growth in the near-future, these cities can only hope to become smaller in a way that is less physically destructive, less harmful to existing residents, and environmentally sustainable. Even though these cities are only shrinking and not declining, a distinction more rhetoric than reality, they still have formidable challenges to overcome. One of the most daunting in the United States is the increasing poverty of the remaining residents. Hovering over this state of affairs, and unavoidable in 2013 as an economic recession lingered in most of these countries, was the prospect that conditions could worsen.

See also: Cities and the Financial Crisis; Housing and Cities; Industrialization, Typologies and History of; Industrialization; Neighborhood Revitalization and Community Development; Real Estate Development; Suburbanization and Suburbanism.

Bibliography

Beauregard, Robert A., 2003. Voices of Decline: The Postwar Fate of U.S. Cities, second ed. Routledge, New York.

Beauregard, Robert A., 2009. Urban population loss in historical perspective: United States, 1820–2000. Environment and Planning A 41, 514–528.

Berkooz, Corry Buckwalter, 2010. Repurposing Detroit. Planning 76 (9), 26–31.

Bernt, Matthias, 2009. Partnerships for demolition: the governance of urban renewal in East Germany's shrinking cities. International Journal of Urban and Regional Research 33 (3), 754–769.

Billitteri, Thomas J., 2010. Blighted cities. CQ Researcher 20 (40), 941–964.

Bontje, Marco, 2004. Facing the challenge of shrinking cities in East Germany: the case of Leipzig. GeoJournal 61, 13–21.

Dewar, Margaret, Thomas, June Manning (Eds.), 2013. The City after Abandonment. University of Pennsylvania Press, Philadelphia.

Dewar, Margaret, Kelly, Christina, Morrison, Hunter, 2013. Planning for better, smaller places after population loss. In: Dewar, M., Manning Thomas, J. (Eds.), The City after Abandonment. University of Pennsylvania Press, Philadelphia, PA, pp. 289–316.

Galster, George, 2012. Driving Detroit: The Quest for Respect in the Motor City. University of Pennsylvania Press, Philadelphia, PA.

Gurwitt, Rob, 2002. Betting on the bulldozer. Governing 15 (10), 28–34.

Herscher, Andrew, 2012. The Unreal Estate Guide to Detroit. University of Michigan Press, Ann Arbor, MI.

Immergluck, Daniel, 2009. The foreclosure crisis: foreclosed properties and federal policy. Journal of the American Planning Association 75 (4), 406–423.

Krohe Jr., James, 2011. The incredible shrinking city. Planning 77 (9), 10–15.

Leonard, Jennifer R., Schilling, Joseph M., 2007. Lessons from the field – strategies and partnerships for preventing and reclaiming vacant and abandoned properties. Real Estate Review 36 (3), 31–39.

Mallach, Alan, 2004. Abandoned property: effective strategies to Reclaim community assets. Housing Facts and Findings 6 (2:1), 3–7.

Mallach, Alan (Ed.), 2012a. Rebuilding America's Legacy Cities. The American Assembly, New York.

Mallach, Alan, 2012b. Depopulation, market collapse and property abandonment: surplus land and buildings in legacy cities. In: Mallach, A. (Ed.), Rebuilding America's Legacy Cities. The American Assembly, New York, pp. 85–109.

Martinez-Fernandez, Cristina, Audirac, Ivonne, Fol, Sylvie, Cunningham-Sabot, Emmanuel, 2012. Shrinking cities: urban challenges of globalization. International Journal of Urban and Regional Research 36 (2), 213–225.

Mock, Brentin, 2011. Can they save youngstown? New American City 20, 40–45.

Mykhenenko, Vlad, Turok, Ivan, 2008. East European cities – patterns of growth and decline, 1960–2005. International Planning Studies 13 (4), 311–342.

Oswalt, Philipp, 2005. Shrinking Cities. In: International Research, vol. 1. Hatje Cantz, Ostfildern.

Oswalt, Philipp, 2006. Shrinking Cities. In: Interventions, vol. 2. Hatje Cantz, Ostfildern.

Power, Anne, Ploger, Jorg, Winkler, Astrid, 2010. Phoenix Cities: The Fall and Rise of Great Industrial Cities. The Policy Press, Bristol, UK.

Rugare, Steve, Schwarz, Terry (Eds.), 2008. Cities Growing Smaller. Cleveland Urban Design Collaborative, Kent State University, Cleveland, OH.

Ryan, Brent, 2012. Design after Decline: How America Rebuilds Shrinking Cities. University of Pennsylvania Press, Philadelphia, PA.

Schilling, Joseph, Mallach, Alan, 2012. Cities in Transition: A Guide for Practicing Planners. American Planning Association, Chicago, IL. Planning Advisory Service Report 568.

Swope, Christopher, 2006. Smart decline. Governing 20 (2), 46–52.

Turok, Ivan, Mykhnenko, Vlad, 2007. The trajectories of European cities, 1960–2005. Cities 24 (3), 165–182.

Wiechmann, Thorsten, 2008. Errors expected – aligning urban strategy with demographic uncertainty in shrinking cities. International Planning Studies 13 (4), 431–446.

Wiechmann, Thorsten, Pallagst, Karina M., 2012. Urban shrinkage in Germany and the USA. International Journal of Urban and Regional Research 36 (2), 261–280.

Sibling-Order Effects

Frank J Sulloway, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Throughout much of the world, eldest children are privileged, socially and economically, over their younger siblings. For this reason, firstborns tend to exhibit greater filial loyalty compared with laterborns. Because sibling order is a proxy for individual differences in age, size, and status, it affects sibling strategies in competition for parental investment. Differences in sibling strategies in turn exert a modest influence on personality and social behavior, reflecting a Darwinian process by which brothers and sisters seek to reduce direct competition through the adoption of unique family niches. Eldest siblings, who typically occupy the niche of a surrogate parent, tend to be more conscientious, achievement oriented, and conforming to authority than are their younger siblings; whereas laterborns, who cultivate low-power strategies, are often more agreeable, open to experience, and extraverted in the sense of being fun-loving and excitement-seeking. These differences in personality are displayed in various historical trends, as laterborns have often championed radical revolutions and reform movements – including the Protestant Reformation, numerous political revolutions, and radical revolutions in science.

Historically, sibling order has influenced important aspects of social, economic, and political life, and it continues to do so today in many traditional societies. Discriminatory inheritance laws and customs about royal succession that favor firstborns and eldest sons are just two examples. Sibling-order effects have also been documented for a wide variety of behavioral tendencies, although the magnitude and causal interpretation of these effects have been subject to debate. In the Darwinian process of competing for parental favor, siblings often employ strategies that are shaped by order of birth within the family, and these strategies exert a lasting impact on personality. Radical revolutions are particularly likely to elicit differences in support by birth order. These behavioral patterns appear to be mediated by differences in personality as well as by differing degrees of identification with the family.

Social, Economic, and Political Consequences

Many societies – especially in past centuries and in non-Western parts of the world – have engaged in practices that favor one sibling position over another. For example, most traditional societies permit infanticide, especially when a child has a birth defect or when an immediately older infant is still breastfeeding. However, no society condones the killing of the elder of two siblings (Daly and Wilson, 1988). An extensive survey of birth order and its relationship to social behavior in 39 non-Western societies found that the birth of a first child typically increases the status of parents and stabilizes the marriage (Rosenblatt and Skoogberg, 1974). In addition, firstborns generally received more elaborate birth ceremonies, were given special privileges, and tended to exert authority over their younger siblings. In most of these 39 societies, firstborns maintained supremacy over their younger siblings even in adulthood. They also gained control of a greater share of parental property than laterborns, were more likely to become head of their local kin group, and tended to exert greater power in relationships with nonfamily members.

Previously, many Western societies employed sibling order as a means of deciding who inherits parental property or assumes political power. Primogeniture (the policy by which the eldest child or son automatically inherits property or political authority) was the most commonly employed mechanism, but other discriminatory inheritance practices have also been employed. For example, secondogeniture involves leaving the bulk of parental property to the second child or second son, and ultimogeniture involves leaving such property to the last-born or youngest son.

Most variations in inheritance practices by sibling order can be understood by considering the advantages that accrue from such policies, given local economic circumstances (Hrdy and Judge, 1993). For example, primogeniture has generally been practiced in societies where wealth is stable and based on limited land, and where talent does not matter much. Leaving the bulk of parental property to a single offspring avoids subdividing estates and hence reducing the social status of the family patronymic – something that was particularly important in past centuries among the landed aristocracy. However, in Renaissance Venice economic fortunes were generally based on speculative commerce rather than ownership of property, and parents typically subdivided their estates equally so as to maximize the chances of having one or more commercially successful offspring (Herlihy, 1977). Ultimogeniture is a policy often found in societies that impose high death taxes on property. This inheritance practice has the consequence of maximizing the interval between episodes of taxation, thus reducing the overall tax burden.

Even in societies employing inheritance systems that favor one sibling over others, parents have commonly provided more or less equally for their offspring by requiring the child who inherits the family estate to pay a compensatory amount to each sibling. Primogeniture and related practices did not always mean disinheritance – a common misassumption. In medieval and early modern times, however, primogeniture among the landed aristocracy did mean that some younger sons and daughters faced difficult economic and social prospects. Among the medieval Portuguese nobility, landless younger sons and daughters were significantly less likely to marry and to leave surviving offspring (Boone, 1986). Younger sons, for example, left 1.6 fewer children than did eldest sons.

Younger sons were also nine times more likely than eldest sons to father a child out of wedlock. Because they posed a serious threat to political stability in their own country, younger sons were channeled by the state into expansionist military campaigns in faraway places such as India, where they often died in battle or from disease. Similarly, the Crusades can be seen, in part, as a church-state response to this domestic danger (Duby, 1977). The surplus of landless younger daughters in the titled nobility was dealt with by sending them to nunneries.

Personality

Birth-order differences have long been claimed in the domain of personality, although these claims have remained controversial despite considerable research on this topic. Psychologists have investigated the consequences of birth order ever since Charles Darwin's cousin Francis Galton (1874) reported that eldest sons were overrepresented as members of the Royal Society. After breaking away from Sigmund Freud in 1910 to found a variant school of psychoanalysis, Alfred Adler (1927) focused on birth order in his own attempt to emphasize the importance of social factors in personality development. A secondborn, Adler considered firstborns to be 'power-hungry conservatives.' He characterized middleborns as competitive and lastborns as spoiled and lazy.

Since Adler speculated about birth order and its consequences for personality in 1927, psychologists have conducted more than 2000 studies on the subject. Critics of this extensive literature have argued that most of these studies are inadequately controlled for key covariates, such as social class and sibship size; that studies often conflict; and that birth-order differences in personality and IQ seem to have been overrated (Ernst and Angst, 1983). Meta-analysis – a technique for aggregating findings in order to increase statistical power and reliability – suggests a different conclusion. If we consider only those well-designed studies controlling for important background variables that covary with birth order and can introduce spurious cross-correlations, meta-analytic reviews reveal modest but consistent birth-order differences for a wide range of personality traits (Sulloway, 1995, 2010). For instance, firstborns are generally found to be more conscientious than laterborns, a difference that is exemplified by their being more responsible, ambitious, and self-disciplined. In addition, firstborns tend to be more conforming to authority and respectful of parents (Manaster et al., 1998). Firstborns also tend to have higher IQs than their younger siblings – a reduction of about one IQ point is observed, on average, with each increment in birth rank (Zajonc and Mullally, 1997; Kristensen and Bjerkedal, 2007).

These and other birth-order differences in personality and behavior can be usefully understood from a Darwinian perspective on family life (Sulloway, 1996, 2010). Because they share, on average, only half of their genes, siblings will tend to compete unless the benefits of cooperating are greater than twice the costs (Hamilton, 1964). In this Darwinian story about sibling competition, birth order does not exert a direct biological influence on personality. For example, there are no genes for being a firstborn or a laterborn. Instead, birth order is best seen as a proxy for various environmental influences, particularly disparities in age, physical size, power, and status within the family. These physical and mental differences lead siblings to pursue alternative strategies in their efforts to maximize parental investment (which includes emotional as well as physical resources). This perspective on how sibling order shapes personality accords with research in behavioral genetics, which finds that most environmental influences on personality are not shared by siblings and hence belong to the 'nonshared environment' (Plomin and Daniels, 1987). Conceptualized in these terms, the family is not primarily a shared environment but rather a series of microenvironments that differ according to age, sex, birth order, and other sources of individual differences.

Prior to the twentieth century, half of all children did not survive childhood. Even minor differences in parental favor would have increased an offspring's chances of getting out of childhood alive. Because eldest children have already survived the perilous early years of childhood, they are more likely than younger siblings to reach the age of reproduction and to pass on their parents' genes. Quite simply, a surviving eldest child was generally a better Darwinian than a younger child, which is why parents in most traditional societies tend to bias investment, consciously or unconsciously, toward older offspring. An exception to this generalization involves youngest children born toward the end of a woman's reproductive career. These children also tend to be favored, since they cannot be replaced (Salmon and Daly, 1998).

Because laterborns generally receive less parental investment than firstborns, they also exhibit higher mortality rates in many cultures (Hertwig et al., 2002). Given that the evolutionary costs of risk-taking are lower whenever the likelihood of survival is lower, laterborns are predicted to take greater risks than their older siblings. Risk-taking can be considered a generalized personality characteristic and hence should extend to many different types of behavior. One test of this hypothesis has involved a meta-analysis of 24 previous studies of risk-taking in athletics, which included 8340 participants (Sulloway and Zweigenhaft, 2010). After controlling important background factors such as sibship size and social class, laterborns were found to be 1.5 times more likely than firstborns to participate in dangerous sports such as rugby, football, and downhill skiing. Similarly, among 700 brothers who played major league baseball, younger brothers were 1.7 times more likely to attempt to steal bases than their older brothers, 1.4 times more likely to be caught stealing bases, and 1.4 times more likely to successfully complete stolen base attempts.

Even when parents do not favor one offspring over another, siblings compete to prevent favoritism. Siblings do so, in part, by cultivating family niches that correspond closely to differences in birth order. Firstborns, for example, tend to act as surrogate parents toward their younger siblings, which is a good way of currying parental favor. As a consequence, firstborns tend to be more conservative and parent identified than their younger siblings, which is one of the most robustly documented findings in the literature. Laterborns cannot compete as effectively as firstborns for this surrogate parent niche, since they cannot babysit themselves. Laterborns therefore seek alternative family niches that will help them to

cultivate parental favor in different ways. To do so, they must often look within themselves for latent talents that can only be discovered through systematic experimentation. Toward this end, laterborns tend to be more open to experience – that is, more imaginative, prone to fantasies, and unconventional – propensities that offer an increased prospect of finding a valued and unoccupied family niche.

In addition to their efforts to cultivate alternative family niches, siblings diverge in personality and interests because they employ differing strategies in dealing with one another. These strategies are similar to those observed in mammalian dominance hierarchies. Because firstborns are physically bigger than their younger siblings, they are more likely to employ physical aggression and intimidation in dealing with rivals. Firstborns are the 'alpha males' of their sibling group, and they generally boss and dominate their younger brothers and sisters. For their own part, laterborns tend to employ low-power strategies to obtain what they want, including pleading, whining, cajoling, humor, social intelligence, and, whenever expedient, appealing to parents for assistance.

Laterborns also tend to form coalitions with one another in an effort to circumvent the physical advantages enjoyed by the firstborn. Middle children are the most inclined to employ diplomatic and cooperative strategies. Some middle children are particularly adept at nonviolent methods of protest. Martin Luther King, Jr., the middle of three children, began his career as a champion of nonviolent reform by interceding in the episodes of merciless teasing that his younger brother inflicted upon their elder sister.

Only children represent a controlled experiment in birth-order research. They have no siblings and therefore experience no sibling rivalry. Consequently, they are not driven to occupy a particular family niche. Although only children, like other firstborns, are generally ambitious and conform to parental authority, they are intermediate between firstborns and laterborns on most other personality traits. Because age spacing can affect functional sibling order, a firstborn whose next younger sibling is six or more years younger is effectively like an only child. Similarly, large age gaps can make some laterborns into functional firstborns or only children.

That brothers and sisters differ from one another for strategic reasons, rather than randomly, has been shown by studies involving more than one sibling from the same family (Schachter, 1982). In one study, firstborns were found to be significantly different in personality and interests from secondborns, who were significantly different from thirdborns. By contrast, the first and third siblings were not as different as adjacent pairs, presumably because of less competition. This process of sibling differentiation has been termed 'deidentification' and extends to relationships with parents. For example, a firstborn child sometimes identifies more strongly with one parent than another. In such cases, the secondborn tends to identify more strongly with the parent that is not already preferred by the firstborn. This process of sibling differentiation is generally adaptive by reducing direct competition for scarce family resources, exemplifying what Darwin (1859) called the 'principle of divergence.'

The most compelling evidence for birth-order effects in personality comes from studies in which siblings assess each other's personalities (Paulhus et al., 1999; Healey and Ellis, 2007; Sulloway, 1999, 2010). Such within-family designs control for the kinds of spurious correlations that can result from comparing individuals from different family backgrounds. Studies based on such direct sibling comparisons exhibit consistent birth-order effects in the expected direction. When the results are organized according to the Five-Factor Model of personality (Costa and McCrae, 1992), firstborns tend to be more conscientious than laterborns, whereas laterborns tend to be more agreeable, extraverted, and open to experience than firstborns.

Although reasonably consistent patterns for birth order and personality are observed according to the Five-Factor Model of personality, findings are significantly heterogeneous for three of the five personality dimensions. For example, firstborns tend to be more assertive than laterborns, which is an indicator of extraversion, but laterborns tend to score higher on most other facets of extraversion, which include being more fun-loving, sociable, and excitement-seeking. Similarly, laterborns tend to be more open to experience in the sense of being unconventional, whereas firstborns tend to be more open to experience in the sense of being intellectually oriented. Finally, firstborns are more neurotic in the sense of being anxious about status, but laterborns are more neurotic in the sense of being self-conscious.

As measured by direct sibling comparisons within the family, birth-order differences explain about 4% of the variance in personality, less than does gender and substantially more than do age, family size, or social class. In studies controlled for differences in age, sex, social class, and sibship size, siblings are about twice as likely to exhibit traits that are consistent with their sibling positions as to exhibit inconsistent traits. In short, not every laterborn has a laterborn personality (just as some firstborns deviate from the expected trend), but a reasonably consonant pattern is nevertheless present.

One still-unresolved question about birth order and personality is the extent to which within-family patterns of behavior transfer to behavior outside the family of origin. Studies suggest that birth-order effects observed in extrafamilial contexts are about one-third to one-half the magnitude of those manifested within the family (Sulloway, 1999, 2001). Relative to firstborn spouses and roommates, for example, laterborn spouses and roommates are generally perceived to be more agreeable and extraverted, but less conscientious and neurotic. Outside the family of origin, birth-order effects seem to manifest themselves most strongly in intimate living relationships and in dominance hierarchies. These findings are not surprising, since they involve the kind of behavioral contexts in which sibling strategies were originally learned.

Behavior during Radical Revolutions

Some of the most consistent evidence for sibling differences in behavior outside the family of origin comes from social and intellectual history. During socially radical revolutions, laterborns have generally been more likely than firstborns to question the status quo and to adopt a revolutionary

alternative. This difference is observed even after controlling for the fact that laterborns, in past centuries at least, appear to have been more socially liberal than firstborns. As a rule, these effects are context sensitive. During the Protestant Reformation, for example, laterborns were more likely than firstborns to suffer martyrdom in support of the Reformed faith. By contrast, in countries that turned Protestant, such as England, firstborns were more likely than laterborns to suffer martyrdom for their refusal to abandon their allegiance to the Catholic faith. Thus laterborns gave their lives in the service of radical rebellion, whereas firstborns did so in an effort to preserve the waning orthodoxy. There was no generalized tendency, however, for laterborns to become 'martyrs' (Sulloway, 1996).

Similarly, during the French Revolution firstborn deputies to the National Convention generally voted according to the expectations of their social class, whereas laterborn deputies were more likely to vote in the opposite manner. For this reason, there was only a modest tendency for firstborns and laterborns to vote in a particular manner, since voting was generally influenced by other important considerations that determined what it meant to conform or rebel in a rapidly changing political environment.

In response to radical conceptual changes in science one finds similar differences by birth order (Sulloway, 1996, 2009). Firstborns tend to be the guardians of what Thomas Kuhn (1970) has called 'normal' science, during which research is guided by the prevailing paradigms of the day. By contrast, laterborns tend to be the outlaws of science, sometimes flaunting its accepted methods and assumptions in the process of attempting radical conceptual change. During the early years of the Copernican Revolution, which challenged church doctrine by claiming that the earth rotates around the sun, laterborns were five times more likely than firstborns to endorse this heretical theory. Copernicus himself was the youngest of four children. When Charles Darwin, the fifth of six children, became an evolutionist in the 1830s, he was 10 times more likely to do so than a firstborn. During other notable revolutions in science – including those led by laterborns such as Bacon, Descartes, Hutton, Semmelweis, and Heisenberg, and by occasional firstborns such as Newton and Einstein – younger siblings have typically been twice as likely as firstborns to endorse the new and radical viewpoint. Conversely, when new scientific doctrines, such as vitalism and eugenics, have appealed strongly to social conservatives, firstborns have been more likely than laterborns to pioneer and endorse such novel but ideologically conservative ideas. These birth-order effects typically fade over the course of scientific progress, as conceptual innovations accumulate supporting evidence and eventually convince the scientific majority.

In spite of their predilection for supporting radical innovations, laterborns have no monopoly on scientific truth. They were nine times more likely than firstborns to support Franz Joseph Gall's bogus theory of phrenology, which held that character and personality could be read by means of bumps on the head. In their willingness to question the status quo, laterborns run the risk of error through overeager rebellion, just as firstborns sometimes err by resisting valid conceptual changes until the mounting evidence can no longer be denied.

During 'normal' science, firstborns possess a slight advantage over laterborns. Being more academically successful, firstborns are more likely than laterborns to become scientists. They also tend to win more Nobel prizes. This finding might seem surprising, but, according to the terms of Nobel's will, Nobel prizes have generally been given for 'discoveries' or creative puzzle solving in science, not for radical conceptual revolutions. Firstborn scientists who have innovated within the system include James Watson and Francis Crick, who together unraveled the structure of DNA, and Jonas Salk, who developed the polio vaccine. Such historical facts highlight another important point, namely, that firstborns and laterborns do not differ in overall levels of 'creativity.' Rather, firstborns and laterborns are preadapted to solving dissimilar kinds of problems by employing disparate kinds of creative strategies.

Family Sentiments

The kinds of birth-order effects that are observed during radical historical revolutions may depend as much on differences in 'family sentiments' as they do on personality. As Salmon and Daly (1998) have shown, firstborns (and to a lesser extent lastborns) are more strongly attached to the family system than are middle children. Historically, radical revolutions have tapped differences in family sentiments in two important ways. First, being considerably older than their offspring, parents were more likely to endorse the status quo, given that age is one of the best predictors of the acceptance of new and radical ideas (Hull et al., 1978). For this reason, endorsing a radical revolution has usually meant opposing parental values and authority, something that firstborns (and to a lesser extent lastborns) are less likely to do than are middleborns. Second, most radical revolutions, even in fields such as science, have tended to raise issues that are directly relevant to issues of parental investment and discrimination among offspring, including filial loyalty and overall identification with the family system. During the Reformation, for example, Protestant leaders such as Martin Luther strongly advocated the right to marriage by previously celibate clergymen and nuns, who tended to be younger sons and daughters. These Protestant leaders also advocated the egalitarian principles of partible inheritance, by which property – and even political rule – were subdivided equally among offspring (Fichtner, 1989). During the height of the controversies raging over Darwin's theory of evolution, Darwin noted that discriminatory inheritance practices posed a serious impediment to human evolution, remarking to Alfred Russel Wallace: "But oh, what a scheme is primogeniture for destroying natural selection!" (Sulloway, 1996: p. 54).

Even when radical ideology does not play a significant role in conceptual change, and when discriminatory inheritance systems are also not a factor, differences in identification with the family can affect social and political attitudes. As Salmon (1998) has shown experimentally, political speeches appeal differentially to individuals by birth order depending on whether kinship terms are included in the speech. Firstborns and lastborns are more likely to react positively to speeches that employ kinship terms such as 'brother' and 'sister,' whereas

middle children prefer speeches that employ references to 'friends.' To the extent that radical social movements make use of kinship terms, which they often do, birth-order differences in family sentiments will tend to influence how siblings react to radical change.

Conclusion

In past centuries birth order and functional sibling order have influenced such diverse social and political phenomena as royal succession, expansionist military campaigns, religious wars and Crusades, geographic exploration, and inheritance laws. Even today, birth order and family niches more generally are among the environmental sources of personality because they cause siblings to experience the family environment in dissimilar ways. In particular, birth order introduces the need for differing strategies in dealing with sibling rivals as part of the universal quest for parental favor. This is a Darwinian story, albeit one with a marked environmental twist. Although siblings appear to be hard-wired to compete for parental favor, the specific niche in which they have grown up determines the particular strategies they adopt within their own family. Finally, because birth order and family niches underlie differences in family sentiments, including filial loyalty and self-conceptions about family identity, the family has historically supplied a powerful engine for revolutionary change.

See also: Family Size Preferences; Family as Institution; Family: The Anthropology of the Concept and Its History; Galton, Sir Francis (1822–1911); Property: Legal Aspects of Intergenerational Transmission.

Bibliography

Adler, A., 1927. Understanding Human Nature. Greenberg, New York.
Boone, J.L., 1986. Parental investment and elite family structure in preindustrial states: case study of late medieval-early modern Portuguese genealogies. American Anthropologist 88, 859–878.
Costa, P.T., McCrae, R.R., 1992. NEO PI-R Professional Manual. Psychological Assessment Resources, Odessa, FL.
Daly, M., Wilson, M., 1988. Homicide. Aldine de Gruyter, Hawthorne, New York.
Darwin, C.R., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.
Duby, G., 1977. The Chivalrous Society. Edward Arnold, London.
Ernst, C., Angst, J., 1983. Birth Order: Its Influence on Personality. Springer-Verlag, Berlin.
Fichtner, P.S., 1989. Protestantism and Primogeniture in Early Modern Germany. Yale University Press, New Haven, CT.
Galton, F., 1874. English Men of Science. Macmillan, London.
Hamilton, W., 1964. The genetical evolution of social behavior. Parts I and II. Journal of Theoretical Biology 7, 1–52.
Healey, M.D., Ellis, B.J., 2007. Birth order, conscientiousness, and openness to experience: tests of the family-niche model of personality using a within-family methodology. Evolution and Human Behavior 28, 55–59.
Herlihy, D., 1977. Family and property in Renaissance Florence. In: Herlihy, D., Udovitch, A.L. (Eds.), The Medieval City. Yale University Press, New Haven, CT, pp. 3–24.
Hertwig, R., Davis, J.N., Sulloway, F.J., 2002. Parental investment: how an equality motive can produce inequality. Psychological Bulletin 128, 728–745.
Hrdy, S., Judge, D.S., 1993. Darwin and the puzzle of primogeniture. Human Nature 4, 1–45.
Hull, D.L., Tessner, P.D., Diamond, A.M., 1978. Planck's principle. Science 202, 717–723.
Kristensen, P., Bjerkedal, T., 2007. Explaining the relation between birth order and intelligence. Science 316, 1717.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.
Manaster, G.J., Rhodes, C., Marcus, M.B., Chan, J.C., 1998. The role of birth order in the acculturation of Japanese Americans. Psychologia 41, 155–170.
Paulhus, D.L., Chen, D., Trapnell, P.D., 1999. Birth order and personality within families. Psychological Science 10, 482–488.
Plomin, R., Daniels, D., 1987. Why are children in the same family so different from one another. Behavioral and Brain Sciences 10, 1–60.
Rosenblatt, P.C., Skoogberg, E.L., 1974. Birth order in crosscultural perspective. Developmental Psychology 10, 48–54.
Salmon, C., 1998. The evocative nature of kin terminology in political rhetoric. Politics and the Life Sciences 17, 51–57.
Salmon, C.A., Daly, M., 1998. Birth order and familial sentiment: middleborns are different. Human Behavior and Evolution 19, 299–312.
Schachter, F.F., 1982. Sibling deidentification and split-parent identifications: a family tetrad. In: Lamb, M.E., Sutton-Smith, B. (Eds.), Sibling Relationships: Their Nature and Significance across the Lifespan. Lawrence Erlbaum, Hillsdale, NJ, pp. 123–152.
Sulloway, F.J., 1995. Birth order and evolutionary psychology: a meta-analytic overview. Psychological Inquiry 6, 75–80.
Sulloway, F.J., 1996. Born to Rebel: Birth Order, Family Dynamics, and Creative Lives. Pantheon, New York.
Sulloway, F.J., 1999. Birth order. In: Runco, M.A., Pritzker, S. (Eds.), Encyclopedia of Creativity, vol. 1, pp. 189–202.
Sulloway, F.J., 2001. Birth order, sibling competition, and human behavior. In: Holcomb, H.R. (Ed.), Conceptual Challenges in Evolutionary Psychology: Innovative Research Strategies. Kluwer Academic Publishers, Dordrecht and Boston, pp. 39–83.
Sulloway, F.J., 2009. Sources of scientific innovation: a meta-analytic approach. Perspectives on Psychological Science 4, 455–459.
Sulloway, F.J., 2010. Why siblings are like Darwin's finches: birth order, sibling competition, and adaptive divergence within the family. In: Buss, D.M., Hawley, P.H. (Eds.), The Evolution of Personality and Individual Differences. Oxford University Press, New York, pp. 86–119.
Sulloway, F.J., Zweigenhaft, R.L., 2010. Birth order and risk taking in athletics: a meta-analysis and study of major league baseball players. Personality and Social Psychology Review 14, 402–416.
Zajonc, R.B., Mullally, P.R., 1997. Birth order: reconciling conflicting effects. American Psychologist 52, 685–699.

Sign Language

Ursula Bellugi and Edward S Klima, Salk Institute, La Jolla, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14066–14071, © 2001, Elsevier Ltd.

Abstract

What would language be like if its transmission were not based on the vocal tract and the ear? How is language organized when it is based instead on the hands moving in space, and the eyes? Do these transmission channel differences result in any deeper differences? The existence of primary signed languages allow us to inquire about the determinants of language organization from a new perspective. Recent research into primary signed languages passed down from one generation of deaf people to the next has made clear that there are many different signed languages arising independently of one another, and of the spoken languages of their surrounding communities. At their core, spoken and signed languages are found to be essentially the same in terms of organizational principles and rule systems. Nevertheless, on the surface, signed and spoken languages differ markedly. American Sign Language and other signed languages display complex linguistic structure, but unlike spoken languages, convey much of their structure by manipulating spatial relations, making use of spatial contrasts at all linguistic levels. The surface and core properties of spoken and signed languages allow us to examine in depth the neurological basis for language and the human capacity for creating linguistic systems in a new way.

Until recently, most of the scientific understanding of the human capacity for language has come from the study of spoken languages. It has been assumed that the organizational properties of language are connected inseparably with the sounds of speech, and the fact that language is normally spoken and heard determines the basic principles of grammar, as well as the organization of the brain for language. There is good evidence that structures involved in breathing and chewing have evolved into a versatile and efficient system for the production of sounds in humans. Studies of brain organization indicate that the left cerebral hemisphere is specialized for processing linguistic information in the auditory–vocal mode and that the major language-mediating areas of the brain are connected intimately with the auditory–vocal channel. It has even been argued that hearing and the development of speech are necessary precursors to this cerebral specialization for language. Thus, the link between biology and linguistic behavior has been identified with the particular sensory modality in which language has developed.

Although the path of human evolution has been in conjunction with the thousands of spoken languages the world over, recent research into signed languages has revealed the existence of primary linguistic systems that have developed naturally independent of spoken languages in a visual–manual mode. American Sign Language (ASL), for example, a sign language passed down from one generation to the next of deaf people, has all the complexity of spoken languages, and is as capable of expressing science, poetry, wit, historical change, and infinite nuances of meaning as are spoken languages. Importantly, ASL and other signed languages are not derived from the spoken language of the surrounding community: rather, they are autonomous languages with their own grammatical form and meaning. Although it was thought originally that signed languages were universal pantomime, or broken forms of spoken language on the hands, or loose collections of vague gestures, now scientists around the world have found that there are signed languages that spring up wherever there are communities and generations of deaf people (Klima and Bellugi, 1988).

One can now specify the ways in which the formal properties of languages are shaped by their modalities of expression, sifting properties peculiar to a particular language mode from more general properties common to all languages. ASL, for example, exhibits formal structuring at the same levels as spoken languages (the internal structure of lexical units and the grammatical scaffolding underlying sentences) as well as the same kinds of organizational principles as spoken languages. Yet the form this grammatical structuring assumes in a visual–manual language is apparently deeply influenced by the modality in which the language is cast (Bellugi, 1980).

The existence of signed languages allows us to enquire about the determinants of language organization from a different perspective. What would language be like if its transmission were not based on the vocal tract and the ear? How is language organized when it is based instead on the hands moving in space and the eyes? Do these transmission channel differences result in any deeper differences? It is now clear that there are many different signed languages arising independently of one another and of spoken languages. At the core, spoken and signed languages are essentially the same in terms of organizational principles and rule systems. Nevertheless, on the surface, signed and spoken languages differ markedly. ASL and other signed languages display complex linguistic structure, but unlike spoken languages, convey much of their structure by manipulating spatial relations, making use of spatial contrasts at all linguistic levels (Bellugi et al., 1989).

The Structure of Sign Language

As already noted, the most striking surface difference between signed and spoken languages is the reliance on spatial contrasts, most evident in the grammar of the language. At the lexical level, signs can be separated from one another

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52018-2

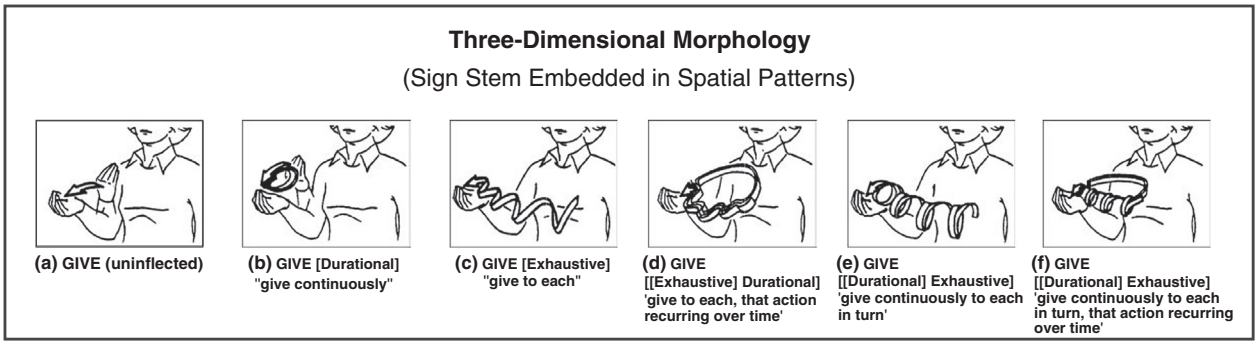

Figure 1 Spatially organized (three-dimensional) morphology in ASL.

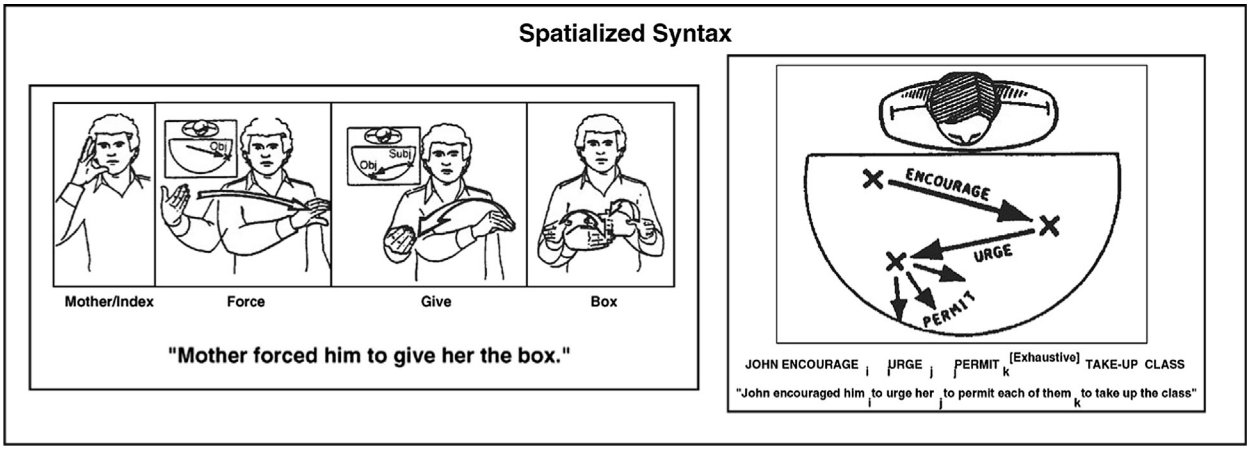

Figure 2 Spatialized syntax.

minimally by manual parameters (handshape, movement, location). The signs for *summer*, *ugly*, and *dry* are just the same in terms of handshape and movement, and differ only in the spatial location of the signs (forehead, nose, or chin). Instead of relying on linear order for grammatical morphology, as in English (*act, acting, acted, acts*), ASL grammatical processes nest sign stems in spatial patterns of considerable complexity (see Figure 1), marking grammatical functions such as number, aspect, and person spatially. Grammatically complex forms can be nested spatially, one inside the other, with different orderings producing different hierarchically organized meanings.

Similarly, the syntactic structure specifying relations of signs to one another in sentences of ASL is also essentially organized spatially. Nominal signs may be associated with abstract points in a plane of signing space, and it is the direction of movement of verb signs between such endpoints that marks grammatical relations. Whereas in English, the sentences 'the cat bites the dog' and 'the dog bites the cat' are differentiated by the order of the words, in ASL these differences are signaled by the movement of the verb between points associated with the signs for *cat* and *dog* in a plane of signing space. Pronominal signs directed toward such previously established points or loci clearly function to refer back to nominals, even with many signs intervening (see Figure 2). This spatial organization underlying syntax is a unique property of visual-gestural systems (Bellugi et al., 1989).

The Acquisition of Sign Language by Deaf Children of Deaf Parents

Findings revealing the special grammatical structuring of a language in a visual mode lead to questions about the acquisition of sign language, its effect on nonlanguage visuospatial cognition, and its representation in the brain. Despite the dramatic surface differences between spoken and signed languages – simultaneity and nesting sign stems in complex co-occurring spatial patterns – the acquisition of sign language by deaf children of deaf parents shows a remarkable similarity to that of hearing children learning a spoken language. Grammatical processes in ASL are acquired at the same rate and by the same processing by deaf children as are grammatical processes by hearing children learning English, as if there were a biological timetable underlying language acquisition.

First words and first signs appear at around 12 months; combining two words or two signs in children, whether deaf or hearing, occurs by about 18–20 months, and the rapid expansion signaling the development of grammar (morphology and syntax) develops in both spoken and signed languages by about 3–3.5 years. Just as hearing children show their mastery and discovery of grammatical regularities by producing overregularizations ('I goed there,' 'we helded the rabbit'), deaf children learning sign language show evidence of learning the spatial morphology signaling plural forms and aspectual forms

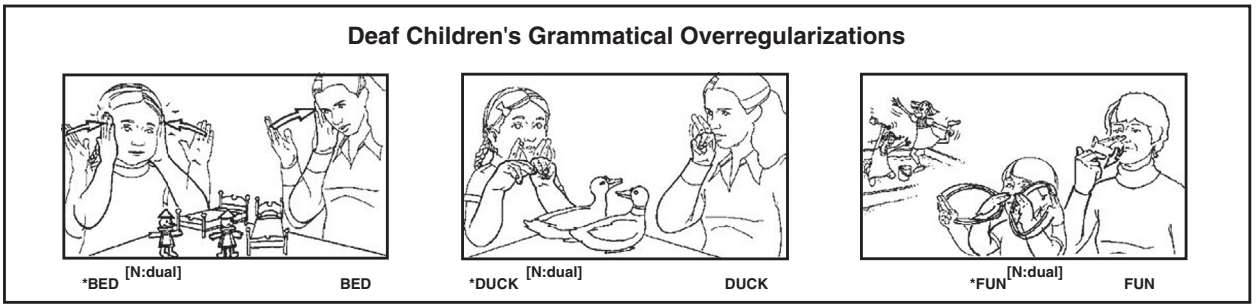

Figure 3 Deaf children's grammatical overregularizations.

by producing overregularizations in spatial patterns (see Figure 3).

Neural Systems Subserving Visuospatial Languages

The differences between signed and spoken languages provide an especially powerful tool for understanding the neural systems subserving language. Consider the following: In hearing/speaking individuals, language processing is mediated generally by the left cerebral hemisphere, whereas visuospatial processing is mediated by the right cerebral hemisphere. But what about a language that is communicated using spatial contrasts rather than temporal contrasts? On the one hand, the fact that sign language has the same kind of complex linguistic structure as spoken languages and the same expressivity might lead one to expect left hemisphere mediation. On the other hand, the spatial medium so central to the linguistic structure of sign language clearly suggests right hemisphere or bilateral mediation. In fact, the answer to this question is dependent on the answer to another, deeper, question concerning the *basis* of the left hemisphere specialization for language.

Specifically, is the left hemisphere specialized for language processing per se (i.e., is there a brain basis for language as an independent entity)? Or is the left hemisphere's dominance generalized to process any type of information that is presented in terms of temporal contrasts? If the left hemisphere is indeed specialized for processing language itself, sign language processing should be mediated by the left hemisphere, as is spoken language. If, however, the left hemisphere is specialized for processing fast temporal contrasts in general, one would expect sign language processing to be mediated by the right hemisphere. The study of sign languages in deaf signers permits us to pit the nature of the signal (auditory-temporal vs visual-spatial) against the type of information (linguistic vs nonlinguistic) that is encoded in that signal as a means of examining the neurobiological basis of language (Poizner et al., 1990).

One program of studies examines deaf lifelong signers with focal lesions to the left or the right cerebral hemisphere. Major areas, each focusing on a special property of the visual-gestural modality as it bears on the investigation of brain organization for language, are investigated. There are intensive studies of large groups of deaf signers with left or right hemisphere focal lesions in one program (Salk); all are highly skilled ASL signers, and all used sign as a primary form of communication throughout their lives. Individuals were examined with an extensive battery of experimental probes, including formal testing of ASL at all structural levels; spatial cognitive probes sensitive to right-hemisphere damage in hearing people; and new methods of brain imaging, including structural and functional magnetic resonance imaging (MRI, fMRI), event-related potentials (ERP), and positron emission tomography (PET). This large pool of well-studied and thoroughly characterized subjects, together with new methods of brain imaging and sensitive tests of signed as well as spoken language allows for a new perspective on the determinants of brain organization for language (Hickok and Bellugi, 2000; Hickok et al., 1996).

Left Hemisphere Lesions and Sign Language Grammar

The first major finding is that so far only deaf signers with damage to the left hemisphere show sign language aphasias. Marked impairment in sign language after left hemisphere lesions was found in the majority of the left hemisphere damaged (LHD) signers, but not in any of the right hemisphere damaged (RHD) signers, whose language profiles were much like matched controls. Figure 4 presents a comparison of LHD, RHD, and normal control profiles of sign characteristics from The Salk Sign Diagnostic Aphasia Examination – a measure of sign aphasia. The RHD signers showed no impairment at all in any aspect of ASL grammar; their signing was rich, complex, and without deficit, even in the spatial organization underlying sentences of ASL. By contrast, signers with LHD showed markedly contrasting profiles: one was agrammatic after her stroke, producing only nouns and a few verbs with none of the grammatical apparatus of ASL, another made frequent paraphasias at the sign internal level, and several showed many grammatical paraphasias, including neologisms, particularly in morphology.

Another deaf signer showed deficits in the spatially encoded grammatical operations which link signs in sentences, a remarkable failure in the spatially organized syntax of the language. Still another deaf signer with focal lesions to the left hemisphere reveal dissociations not found in spoken language: a dissociation between sign and gesture, with a cleavage between capacities for sign language (severely impaired) and manual pantomime (spared). In contrast, none of the RHD signers showed any within-sentence deficits; they were completely unimpaired in sign sentences and not one showed aphasia for sign language (in contrast to their marked nonlanguage spatial deficits, described below) (Hickok and Bellugi, 2000; Hickok et al., 1998).

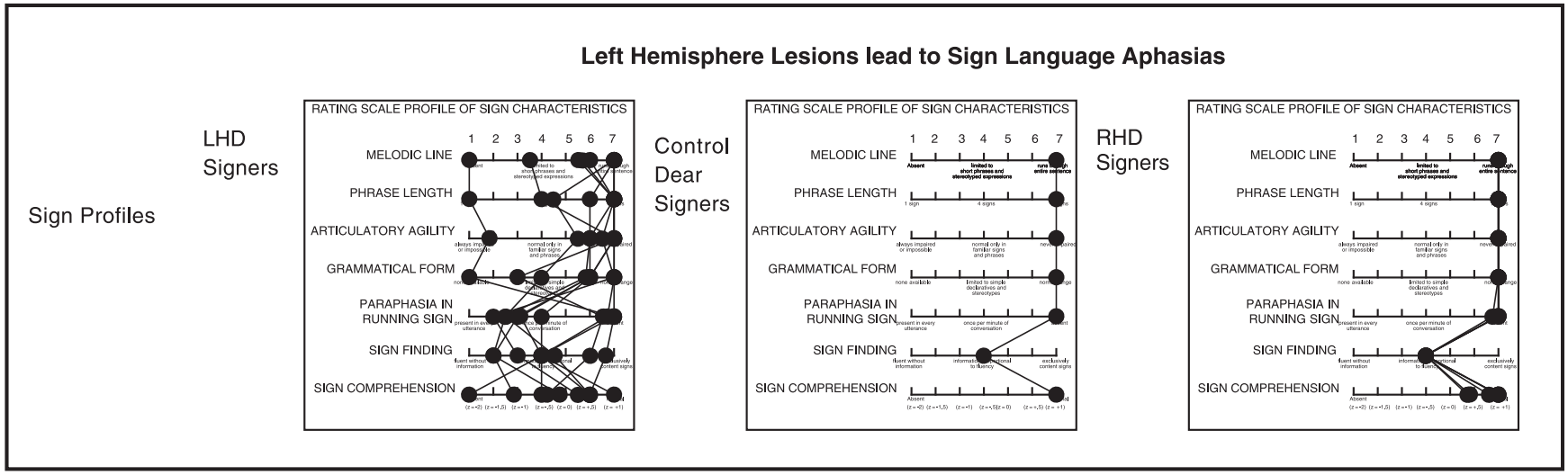

Figure 4 Left hemisphere lesions lead to sign language aphasias.

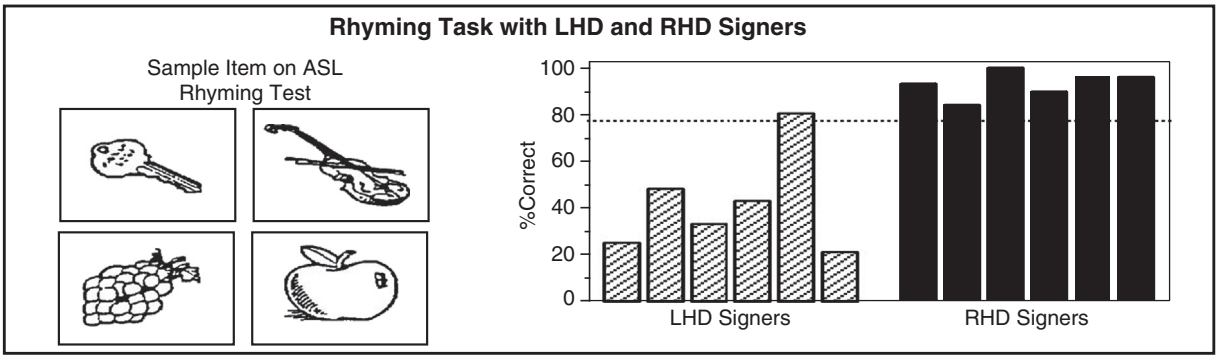

Figure 5 Rhyming task with LHD and RHD signers.

Moreover, there are dramatic differences in performance between LHD and RHD signers in formal experimental probes of sign competence. For example, a test of the equivalent of rhyming in ASL provides a probe of phonological processing. Two signs 'rhyme' if they are similar in all but one phonological parametric value such as handshape, location, or movement. To tap this aspect of phonological processing, subjects are presented with an array of pictured objects and asked to pick out the two objects whose signs rhyme (Figure 5). The ASL signs for *key* and *apple* share the same handshape and movement, and differ only in location, and thus are like rhymed pairs. LHD signers are significantly impaired relative to RHD signers and controls on this test, another sign of the marked difference in effects of right- and left-hemisphere lesions on signing. On other tests of ASL processing at different structural levels, there are similar distinctions between left- and right-lesioned signers, with the right-lesioned signers much like the controls, but the signers with left hemisphere lesions significantly impaired in language processing. Moreover, studies have found that there can be differential breakdown of linguistic components of sign language (lexicon and grammar) with different left hemisphere lesions.

Right Hemisphere Lesions and Nonlanguage Spatial Processing

These results from language testing contrast sharply with results on tests of nonlanguage spatial cognition. RHD signers are significantly more impaired on a wide range of spatial cognitive tasks than LHD signers, who show little impairment. Drawings of many of the RHD signers (but not those with LHD) show severe spatial distortions, neglect of the left side of space, and lack of perspective. RHD deaf signers show lack of perspective, left neglect, and spatial disorganization on an array of spatial cognitive nonlanguage tests (block design, drawing, hierarchical processing), compared with LHD deaf signers. Yet, astonishingly, these severe spatial deficits among RHD signers do not affect their competence in a spatially nested language, ASL. The case of a signer with a right parietal lesion leading to severe left neglect is of special interest: Whereas his drawings show characteristic omissions on the left side of space, his signing (including the spatially organized syntax) is impeccable, with signs and spatially organized syntax perfectly maintained.

The finding that sign aphasia follows left hemisphere lesions but not right hemisphere lesions provides a strong case for a modality-independent *linguistic* basis for the left hemisphere specialization for language. These data suggest that the left hemisphere is predisposed biologically for language, independent of language modality. Thus, hearing and speech are not necessary for the development of hemisphere specialization – sound is not crucial. Furthermore, the finding of a dissociation between competence in a spatial language and competence in nonlinguistic spatial cognition demonstrates that it is the type of information that is encoded in a signal (i.e., linguistic vs spatial information) rather than the nature of the signal itself (i.e., spatial vs temporal) that determines the organization of the brain for higher cognitive functions.

Language, Modality, and the Brain

Analysis of patterns of breakdown in deaf signers provides new perspectives on the determinants of hemispheric specialization for language. First, the data show that hearing and speech are not necessary for the development of hemispheric specialization: sound is *not* crucial. Second, it is the left hemisphere that is dominant for sign language. Deaf signers with damage to the left hemisphere show marked sign language deficits but relatively intact capacity for processing nonlanguage visuospatial relations. Signers with damage to the right hemisphere show the reverse pattern. Thus, not only is there left hemisphere specialization for language functioning, there is also complementary specialization for nonlanguage spatial functioning. The fact that grammatical information in sign language is conveyed via spatial manipulation does not alter this complementary specialization.

Furthermore, components of sign language (lexicon and grammar) can be selectively impaired, reflecting differential breakdown of sign language along linguistically relevant lines. These data suggest that the left hemisphere in humans may have an innate predisposition for language, regardless of the modality. Since sign language involves an interplay between visuospatial and linguistic relations, studies of sign language breakdown in deaf signers may, in the long run, bring us closer to the fundamental principles underlying hemispheric specialization.

See also: Evolution and Language: Overview; Human Language: Evolutionary Precursors; Sign Language: Psychological and Neural Aspects.

Bibliography

Bellugi, U., 1980. The structuring of language: clues from the similarities between signed and spoken language. In: Bellugi, U., Studdert-Kennedy, M. (Eds.), Signed and Spoken Language: Biological Constraints on Linguistic Form. Dahlem Konferenzen, FL, Weinheim/Deerfield Beach, pp. 115–140.

Bellugi, U., Poizner, H., Klima, E.S., 1989. Language, modality and the brain. Trends in Neurosciences 10, 380–388.

Emmorey, K., Kosslyn, S.M., Bellugi, U., 1993. Visual imagery and visual-spatial language: enhanced imagery abilities in deaf and hearing ASL signers. Cognition 46, 139–181.

Hickok, G., Bellugi, U., 2000. The signs of aphasia. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, second ed. Elsevier Science Publishers, Amsterdam, The Netherlands.

Hickok, G., Bellugi, U., Klima, E.S., 1996. The neurobiology of signed language and its implications for the neural organization of language. Nature 381, 699–702.

Hickok, G., Bellugi, U., Klima, E.S., 1998. The basis of the neural organization for language: evidence from sign language aphasia. Reviews in the Neurosciences 8, 205–222.

Klima, E.S., Bellugi, U., 1988. The Signs of Language. Harvard University Press, Cambridge, MA.

Poizner, H., Klima, E.S., Bellugi, U., 1990. What the Hands Reveal About the Brain. MIT Press/Bradford Books, Cambridge, MA.

Sign Language, Disorders of

David P Corina, Cognitive Neurolinguistics Laboratory, Center for Mind and Brain, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Deaf and hearing persons who routinely use a signed language to communicate are subject to many of the same communication impairments that befall hearing users of spoken language. However, the expression of sign language impairment is poorly understood. Here we examine the effects of aphasia, apraxia, and Parkinson's disease on sign language use. We observe that the effects of these disorders are largely homologous to those found in spoken languages. We then consider childhood developmental disorders, and focus on autism spectrum disorder and stuttering. We discuss how the well-known communicative impairments associated with severe autism is expressed in signers with autism. We discuss how errors in copying limb movements seen in children with autism, is realized in sign language. Finally we turn to a discussion of stuttering, a common trait in speech development, which we claim is largely absent in sign language. Taken together the examination of language disorders in users of signed and spoken languages provides a powerful means to understand the common neural substrates of human communication while leaving in stark relief those properties of human communication that are specific to its modality of expression.

Signed languages of the deaf are naturally evolving linguistic systems exhibiting the full range of linguistic complexity found in spoken languages. Contrary to popular belief sign language is not universal. Just as there are a multitude spoken language communities around the world (e.g., speakers of Quechua, Farsi, Portuguese, and English), there are many different sign language communities (e.g., signers of Langue des Signes du Québec, Deutsche Gebärdensprache (DGS), Taiwan Ziran Shouyu, and British Sign Language (BSL), to name but a few). Signed languages are not invented hand symbols that simply represent the words of a spoken language, nor are they elaborate and codified systems of pantomime. They are complete, natural languages in their own right, with forms and grammatical rules that are abstract, linguistically complex, and largely independent of their surrounding spoken language communities. Although individual signed languages often correspond to the same geographical and political regions that separate users of spoken languages (e.g., both German spoken language and DGS are found in Germany, while English and BSL are found in England), sign languages are autonomous of the spoken languages used within these regions (for example, American Sign Language (ASL) is also used in Nigeria and in the Philippines). Deaf communities utilize sign languages just as hearing communities use spoken languages. They are fully complementary forms of human communication. Many signed languages do have auxiliary systems for borrowing elements of spoken languages into a manual form. The most familiar of these are fingerspelling systems in which the hand or hands represent the alphabet (or written character) on the hands.

Deaf individuals are subject to the same disorders that affect language abilities in hearing persons, including developmental language disorders such as autism as well as acquired disorders, such as aphasia. However, given the fact that the transmission and reception modalities of speech and sign languages clearly differ, there is the theoretical possibility that one may observe modality-specific language impairments that only manifest in speech or sign. In this article we provide an introduction to some disorders that are common to both signed and spoken languages. Here we focus on adult neurological disorders, primarily aphasia, apraxia, and Parkinson's disease. We discuss the realization of these sign language disorders in relation to their spoken language counterparts. We then turn to the emerging field of developmental language disorders and briefly explore the impact of autism spectrum disorder (ASD) on signing abilities and finally consider whether there is an analogue to stuttering in a signed language. Much of the research on disordered signing has focused on deaf communities in the United States, hence the majority of the examples are drawn from ASL, but these are suggested to be representative of users of other signed languages. However, as worldwide research on signed languages continues to grow, we look forward to the confirmation of this premise and will be interested to learn about how different properties of signed languages may result in variations of impairment.

The consideration of linguistic impairments in signers provides a unique avenue for understanding whether symptoms associated with a language disorder reflect the unique demands of a language's mode of expression or reflects core linguistic and cognitive properties associated with human languages processing regardless of the modality of expression.

Sign Language Aphasia and Limb Apraxia

Damage to left-hemisphere cortical regions can result in acquired language impairments known as aphasia. Aphasia may differentially affect a patient's ability to produce, repeat, and/or comprehend language. Clinically, the constellation of symptoms associated with aphasia gives rise to distinct subtypes of aphasia. Brain damage may also result in impairment to the oral articulators that impede the ability to speak or sequence speech movements; this motor-based language impairment is termed an apraxia of speech. Theoretically, language deficits which impair the core representational properties of language (i.e., the aphasias) may be separated from purely implementation-based apraxias, in practice, it is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54043-4

often difficult to differentiate these impairments, as lesions are rarely discrete and patients often present with admixtures of symptoms.

Lesions to the left inferior frontal region are often associated with Broca's aphasia whereby a patient may have a severe problem in language production, with less pronounced impairments in language comprehension. Lesions in the posterior superior and middle temporal lobes regions often result in Wernicke's aphasia, which is marked by severe language comprehension problems with fluent, but often mistake-prone (i.e., paraphasic) speech output.

Deaf signers, like hearing speakers, also exhibit language disturbances when left-hemisphere cortical regions are damaged (for a review see Corina, 1998). Case histories coupled with more recent neuroimaging studies of healthy deaf subjects have provided confirming evidence for the importance of the left hemisphere in the mediation of signed languages (MacSweeney et al., 2008). Moreover there is good evidence that within the left hemisphere, cerebral organization in deaf signers follows an anterior/posterior dichotomy for language production and comprehension, respectively, which is familiar from speech. Emerging studies do suggest that sign languages may recruit the inferior and superior parietal lobe regions to a greater extent than is seen in spoken languages. These regions are involved in the guidance of manual and visuomotor behaviors as well as in the encoding of spatial environmental properties which may be used in the service of grammatical distinctions in signed languages such as ASL.

Sign Language Production

In spoken language aphasia, chronic language production impairments are typically associated with left-hemisphere frontal anterior lesions that involve the cortical zone encompassing the lower posterior portion of the left frontal lobe (e.g., Broca's area) and associated white-matter tracts (Goodglass, 1993) and the anterior insula (Dronkers et al., 2000). A classic case study reported in Poizner et al. (1987) provides clues to the role of left-hemisphere prefrontal regions in the mediation of signing. Patient G.D., a deaf signer with a large lesion in a left anterior frontal region encompassing Brodmann's area (BA) 44/45, presented with nonfluent signing with intact sign comprehension. Specifically, G.D.'s signing was effortful and dysfluent, with output often reduced to single-sign utterances. The signs she was able to produce were agrammatic, devoid of the movement modulations that signal morphosyntactic contrasts in fluent signing. As with hearing Broca's aphasics, this signer's comprehension of other's language was undisturbed by her lesion. Both at the single sign and sentence level, her comprehension were on par with that of control subjects.

Sign Language Comprehension

The fluent aphasias are typically manifested as severely impaired comprehension coupled with fluent, but often paraphasic (e.g., semantic and phonemic paraphasic) output. These aphasias are typically associated with lesions to left-hemisphere posterior temporal regions. Wernicke's aphasia, for example, is often associated with damage to the posterior regions of the left superior temporal gyrus. More recent work has suggested the contribution of the posterior middle temporal gyrus in cases of chronic Wernicke's aphasia (Dronkers et al., 2000).

Signers with left-hemisphere posterior lesions also evidence fluent sign aphasia. Two example cases are reported in Poizner et al. (1987) and Corina et al. (1992). These patients presented with severe sign comprehension difficulties in the face of relatively fluent but paraphasic sign output. While damage to left temporal cortex has been demonstrated to impair sign comprehension in some patients (Hickok et al., 2002), the lesions in these two patients did not occur in cortical Wernicke's area proper, but rather involved frontal lobe and inferior parietal areas. In both cases, lesions extended posteriorly to the supramarginal gyrus (SMG). This is interesting, as lesions associated with the SMG alone in users of spoken language do not typically result in severe speech comprehension deficits. These two cases suggest that sign language comprehension may be more dependent than speech on left-hemisphere inferior parietal areas, a difference that may reflect within-hemisphere reorganization for cortical areas involved in sign comprehension.

Sign Paraphasias

Language breakdown following left-hemisphere damage is not haphazard, but systematic and affects independently motivated linguistic categories. An example of the systematic patterning in sign and spoken language breakdown is illustrated by paraphasic errors (Corina, 2000). Paraphasia describes the erroneous substitution of an unexpected word for an intended target. Most semantic paraphasias represent the same part of speech as the target and have a clear semantic relationship to it, for example saying *hand* instead of *foot*. In contrast, phonemic or 'literal' paraphasias refer to the production of unintended sounds or syllables in the utterance of a partially recognizable word (Goodglass, 1993). Phonemic sound substitution in speech may result in another real word, related in sound but not in meaning (e.g., *telephone* becomes *television*). Also attested are cases in which the erroneous word shares both sound characteristics and meaning with the target (*broom* becomes *brush*; Goodglass, 1993).

Examples of ASL semantic paraphasias are well documented (Poizner et al., 1987). For example, subject P.D. produced clear lexical substitutions: BED for CHAIR, DAUGHTER for SON, QUIT for DEPART, etc. In general, the semantic errors of P.D. overlap in meaning and lexical class with the intended targets. Concerning phonemic errors, a priori we may expect to find an equal distribution of phonological errors affecting the four major formational parameters of ASL phonology: handshape, movement, location, and orientation. However, handshape configuration errors are the most widely reported, while paraphasias affecting movement, location, and orientation are infrequent (Corina, 2000). This is noteworthy as handshape is considered a separate autosegmental tier in formal treatment of ASL phonology (Sandler and Lillo-Martin, 2006).

An unusual case of sign paraphasia is reported by Hickok et al. (1996). A life-long signer, R.S., suffered an infarct to the left frontal operculum. Neurological examination revealed an initial expressive aphasia that largely resolved with lingering problems of word-finding and frequent phonemic paraphasia.

In the absence of limb apraxia, R.S. exhibited paraphasia restricted to two-handed signs. For example, for signs that require two hands to assume different handshapes and/or to move independently, R.S. would incorrectly fail to move one of her hands. In other cases, for signs where the appropriate movement of one hand was moved relative to the position of the other, R.S. would produce incorrect relational movements. That is, while the path movement itself was correct, it was not carried out correctly with respect to the spatial relationship of the other hand. In addition, during one-handed signing, she exhibited mirroring of movements and handshapes of the dominant hand on the nondominant hand that was somewhat reduced in degree of movement. This mirroring was not seen during nonlinguistic movement and was qualitatively different from mirror movements effecting distal movement sometimes seen in cases of hemiparesis (Hickok et al., 1996). These paraphasias do not have a clear analogy to errors seen in spoken language. These selective two-handed errors may be taken as an indication that the modality and/or form of a linguistic system may place unique demands on the neural mediation and implementation of a language. These errors can be taken as evidence for language-form specific linguistic impairment.

Induced Paraphasias

Additional insights into the neural control of paraphasic errors have been reported by Corina et al. (1999), who investigated sign language production in a deaf individual undergoing an awake left-hemisphere cortical stimulation mapping procedure for the surgical treatment of epilepsy. During the language-mapping portion of the procedure, the participant was required to sign (with his left hand) the names of line drawn pictures. Disruption of the ability to perform the task during stimulation was taken as evidence that the cortical region in question was integral to the language task (Stemmer and Whitaker, 1998).

Stimulation to two anatomical sites led to consistent naming disruption. One of these, an isolated frontal opercular site, corresponded to the posterior aspect of Broca's area, BA 44. A second site, located in the parietal opercular region, corresponded to the SMG, BA 40. Importantly, the nature of these errors was qualitatively different. Stimulation of Broca's area resulted in errors involving the motor execution of signs. These errors were characterized by a laxed articulation of the intended sign, with nonspecific movements (repeated tapping or rubbing) and a reduction in handshape configurations to a laxed-closed fist handshape. With stimulation of the SMG, S.T. produced both formational and semantic errors. Formational errors were characterized by repeated attempts to articulate the intended targets. Successive formational (i.e., phonological) approximations of the correct sign were common. The semantic errors were characterized as semantic substitutions that are formationally quite similar to the intended targets. One conceptualization of these data is that stimulation to Broca's area has a global effect on the motor output of signing, whereas stimulation to a parietal opercular site, the SMG, disrupts the correct selection of the linguistic components (including both phonological and semantic elements) required for sign naming.

Apraxia and Signed Language

Signed languages provide an important contrast to naturally occurring gestures and limb actions. Unlike naturalistic movements, linguistic motor behaviors are highly symbolic, hierarchically compositional, and conventionalized. The production of signs nevertheless requires that linguistic forms stored in memory be planned and executed, respectively, by the same somatotopic brain regions and manual articulators that mediate physically comparable nonlinguistic gestures. Thus, in principle one might expect that deficits in sign language use may be distinguished clinically from apraxic disturbances of learned movement execution (Geschwind and Damasio, 1985) as well as cases where gesture use and sign language use are uniformly impaired.

Apraxia is defined as an impairment of the execution of a learned movement in response to a stimulus that would normally elicit the movement. Such impairments are subject to the conditions that the afferent and efferent systems involved are intact, and that the difficulty is not due to inattentiveness or lack of cooperation (Geschwind and Damasio, 1985). The left inferior parietal lobe has been considered an area responsible for the representation of learned movement. Lesions associated with the left inferior parietal lobe result in an inability to perform and comprehend gestures, whereas damage more anterior to this region may result in production impairment with sparing of pantomime comprehension (Heilman et al., 1982; Rothi et al., 1991). Apraxic disorders may affect symbolic intransitive movements (such as saluting) as well as everyday movements (like dressing oneself), and may even impede the ability to copy meaningless limb movements. Most researchers acknowledge two main categories of apraxia: ideational apraxia and ideomotor apraxia. In ideational apraxia, the patient is specifically impaired in the ability to recall previously well-established actions, such as those involved in object use. Ideomotor apraxia refers to an inability to imitate an unfamiliar action or series of actions (McCarthy and Warrington, 1990).

Dissociations between sign impairments and ideational apraxias (disturbances in conventionalized gestures and object-oriented pantomimes) are evident. Patient W.L., for example, produced and comprehended pantomime normally but demonstrated marked deficits in sign language production and comprehension (Corina et al., 1992). His gestures were clearly intended to convey symbolic information that he ordinarily would have imparted with sign language. The case of W.L. emphasizes that sign language impairments following left-hemisphere damage are not simply attributable to undifferentiated impairments in the motoric instantiation of symbolic representations, but in fact reflect disruptions to a manually expressed linguistic system. This conclusion has recently been supported by Marshall et al. (2004), who report on the dissociation of sign and gesture in Charles, an aphasic user of BSL. Once again, the ability to use nonlinguistic gesture was intact, while sign language expression showed significant impairment.

In contrast, aphasic subject G.D., who had a large frontoanterior lesion and was dysfluent in sign production, showed impaired ability to produce representational intransitive and transitive movements to command, but nevertheless showed

normal performance on a pantomime recognition test (Poizner et al., 1987). Subject Maureen reported by Marshall et al. (2005), a bilingual user of BSL with a left-hemisphere infarct in the vicinity of the middle cerebral artery exhibited a dense limb apraxia, that like G.D. affected, her signing and the reproduction of both meaningful and meaningless limb movements. She was unable to gesture the use of common objects that she nevertheless used in spontaneous everyday activities.

The sparing of pantomime comprehension in the face of aphasic comprehension impairments emphasizes the fact that language impairments following left-hemisphere damage are not attributable to undifferentiated symbolic impairments. However the fact that limb apraxia can co-occur with sign language production deficits as in the cases of G.D. and Maureen, also remind us of the close coupling between skilled movement abilities and execution of sign language. This point in reinforced by the fact that we have yet to witness a signer who shows spared execution of skilled nonlinguistic movement with a clear impairment in the production of sign language.

Parkinson's Disease

Parkinson's disease is a degenerative disorder of the brain that affects the extrapyramidal motor system and is mainly characterized by progressive tremor, bradykinesia (slowness of movement), and rigidity. Degeneration of the substantia nigra with subsequent decrease of striatal dopamine is responsible for the motor and cognitive disturbances seen in Parkinson's patients (Litvan, 1996). The motor deficits observed in these subjects are not specific to language but are evidenced across the general domain of motor behaviors. In hearing patients with Parkinson's disease, motor system dysfunction affects the speech articulators. The initial problem in the untreated patient is a failure to control respiration for the purpose of speech, and there follows a forward progression of articulatory symptoms involving the larynx, pharynx, tongue, and, finally, the lips (Critchley, 1981). Perceptual features of speech changes include aprosody (i.e., monopitch, monoloudness, diminished stress contrasts, and rate abnormalities), impairment of articulation (such as imprecise consonants), and vocal quality deviations (such as harsh or breathy voice). Physiological studies show that patients with Parkinson's disease have reduced articulatory displacements relative to normal speakers, which may appear as rigidity, and improper coordination of agonist–antagonist muscle pairs or groups (Gentil et al., 1995).

Studies of deaf persons with Parkinson's disease highlight the importance of the extrapyramidal motor system and basal ganglia in the mediation of sign language. The sign language deficits observed in Parkinsonisan signing are an expression of the generalized motor impairments common in Parkinson's disease. Poizner and colleagues have presented several case studies of Parkinsonian signers (Brentari et al., 1995). These signers show disturbed joint use, flattening of phonological distinctions, and timing disruption. Signers with Parkinson's disease have been described as signing in a monotonous fashion, with a severely restricted range of temporal rates and tension in signing, two features which normally serve to signal emphasis. Accompanying restrictions in limb movements are deficits in the motility of facial musculature, which further reduce expressivity in signing.

In Parkinsonian signing the scope of sign movement and precision of articulation are compromised. These changes include the distalization of movements, reductions in contact between sign articulators, and laxing of handshapes. Distalization of movement refers to the displacement of sign movements to joints closer to the extremities. For example, a sign like GIVE contains a path movement away from the signer and is normally articulated by straightening the elbow joint, thus moving the hand away from the body. A distalization of this movement would maintain the joint angle of the elbow and substitute an outward nod of the wrist, also moving the hand away from the signer. The end result is maintained directionality of movement with reduced articulatory demands. However, when this displacement occurs the perceptual saliency of signing suffers.

In Parkinsonian signing, there is often a systematic reduction of sign articulation that results in signs being articulated at the torso (as opposed to distinct, lexically specified locations). For example, Brentari and Poizner (1994) report their subject J.H. as signing BOY, normally articulated on the forehead, at the torso level. In fact, most of J.H.'s signs were articulated near the waist during free conversation. In Parkinsonian signers, physical contacts between the major articulators and body locations may be entirely lost. The lack of contact is consistent with a description of overall smoothing of articulation in signers with Parkinson's disease; maintenance of physical contacts with body locations probably places significant demands on the updating and programming of movement plans used in signing. The handshapes of signers with Parkinson's disease are often quite lax and do not maintain the crisp closure or opening of finger postures seen in normal signing. Importantly, the handshapes observed (and indeed all the articulatory distinctions) are approximations of the intended form, and not incorrect substitutions. Differences in hand orientation have been shown to migrate toward less-marked orientations about the sagittal plane, and reductions in the magnitude of orientation changes required by some signs have been documented (Brentari and Poizner, 1994). These findings are consistent with a general disruption of motor control that leads to phonetic disruptions in signing. This can be contrasted to the phonemic disruption apparent in aphasic signing.

All Parkinsonian patients typically show hesitation and slowing in the initiation and execution of movements. Sequential movements appear to be particularly difficult. Indeed, slowness in movement execution and slowness in switching are common manifestations of basal ganglia disease (Agostino et al., 1992). Signers with Parkinson's disease tend to execute all signs in a slow and uniform temporal rate. Parkinsonian subjects show similar temporal pauses between words and phrases (unlike normal signing); this lack of temporal differentiation adds to the perceptual phenomena of monotonous signing. The lack of differentiation may very well reflect disruption of processes involved in the rapid motor shifts that are required for fluent multiword signing.

In summary, the general movement disorders associated with Parkinson's disease affect sign language abilities and result in phonetic disruption in the articulation of signing. Motor

disruption leads to errors involving timing, scope, and precision of articulation features which are common to both Parkinsonian speech and signing. The study of timing relationships in signing provides further evidence that basal ganglia dysfunction affects not only movement execution (e.g., slowing) but also the switching between movements. Taken together, these studies illustrate the benefits of addressing questions of motor control from a perspective of sign language production.

Developmental Disorders of Sign Language

Studies of developmental disorders that affect sign language use are regrettably rare. The lack of systematic studies is hampered by several factors including the lack of reliable and normative developmental measures for sign language production and comprehension. In addition, the great heterogeneity in language exposure, levels of hearing loss, and educational placement (mainstreamed vs residential school, oral-based vs sign-based curriculums, etc.), coupled with the fact that deafness is often a comorbid symptom of a wide range of developmental disorders, makes systematic studies of subsets of already low-incidence populations difficult. However, some headway is being made (see for example, Mason et al., 2010; Quinto-Pozos et al., 2011).

In considering developmental disorders, one can ask, how do developmental disorders which impact spoken language abilities manifest in a deaf user of a signed language? One can also ask, whether there are, as of yet, unidentified language disorders that are specific to signed languages that do not have a clear analogy to impairments seen in users of spoken languages. Equally interesting is whether there are impairments in spoken languages that do not have clear analogue in signed language.

Autism

ASD consists of a set of neurobiological developmental disorders characterized by communicative and social deficits as well as repetitive, stereotyped behaviors (Centers for Disease Control, 2012a). ASD appears to be more frequent in the deaf population than in the general population (Szymanski et al., 2012; Jure et al., 1991) and there is evidence that severe hearing loss occurs in the ASD population at a rate of about 10 times than that of the general population (Rosenhall et al., 1999).

Clinicians have had interest in the use of visual manual communication as adjunctive therapy for nonverbal hearing children with autism. Early studies suggested that even fairly brief simultaneous communication or sign language training can be an effective means of improving communication skills in low-functioning hearing autistic children. Studies suggest that these children can often learn a small set of sign language vocabulary, with some showing progression to two-sign utterances. However, after the initial success of these early sign combinations, most hearing autistic children make only limited progress in terms of the average length and complexity of their sign utterances (Bonvillian et al., 1981).

Sign researchers have begun to explore the communicative patterns observed in the language of deaf children with autism. Poizner et al. (1987) briefly presents the case of Judith M, a severely autistic deaf signer. Judith M lived in a household with two deaf signing parents and two older deaf brothers, all highly fluent signers. As such the communicative impairments observed Judith M. cannot be attributed to an impoverished language exposure as may occur in cases of deaf (and autistic) signers raised in nonsigning homes. Judith M. began to exhibit autistic symptoms at age 11 months; this included a failure to make eye contact with persons around her and an abnormal interest in inanimate objects over social contacts. She exhibited language problems including a late onset of language use. She began to sign at age 5; however her language output was seldom spontaneous, and her language exhibited echolalia, where she repeated single signs from ASL sentences directed toward her. Her signing was reported as 'stiff' and her sign output was devoid of grammatical complexity seen in normal childhood signed utterances. Her echolalic signing was interspersed with meaningless repetitive movements that appear empty of communicative intent. She rarely made eye contact and tended to avoid social interaction other than to serve her own basic wants and needs. She showed no variation in facial expression or gesturing, nor did she attempt to mime or to communicate in other ways. Family members noted that she appeared generally incapable of abstract thought, and showed no indication that she responded to the feelings of others. Thus despite being immersed in a home environment in which sign communication rather than speech communication is the norm, Judith M. nevertheless presents with a profile of a severely autistic child with characteristic social and communicative impairments.

In a recent report, Shield and Meier (2012) reported a sample of five deaf children with ASD who exhibited a preponderance of sign language errors affecting the formational components of a sign. As noted previously, formational errors of sign production may independently affect the handshape, orientation, movement, and location of a sign. Shield and Meier (2012) reported that ASD children showed a robust tendency to fingerspell with a reversed palm orientation. Since fingerspelling is typically directed outward, such reversals suggest while learning to fingerspell, children with ASD reproduce handshapes as they appear from their own perspective, failing to reorient their own palm so that it faces their addressee. This marked pattern of sign articulation has been reported in the imitation of body movements by hearing children with ASD. A compelling finding from a number of studies is that hearing children with ASD tend to reverse the direction of observed palm orientations when imitating the gestures of others (Williams et al., 2004). These reversed palm orientations suggest that children with ASD imitate gestures as they appear from their own perspective. This has been taken as evidence for autism-specific deficit in self–other mapping, that is, a deficit in the process(es) by which children and adults observe the movements of others and map them onto their own bodies (Rogers and Pennington, 1991).

The studies of sign language in hearing children with ASD indicate that while signs may provide an alternative and at times a beneficial avenue for communication, the mere availability of this manual mode of communication does not fundamentally unlock the pervasive social and language impairments associated with ASD. The rare studies of deaf

signers with ASD provide a fascinating example of how complex manual movements, such as those used during fingerspelling, reflect a pervasive deficit in self-expression.

Stuttering

Stuttering is a disorder in which sounds, syllables, or words are repeated or prolonged, disrupting the normal flow of speech. Speech disruptions may be accompanied by struggling behaviors, such as rapid eye blinks or tremors of the lips. Stuttering affects people of all ages. Developmental stuttering is observed in approximately 5% of children between the ages of 2 and 5 as they are developing language skills. Boys are twice as likely to stutter as girls; as they get older, however, the number of boys who continue to stutter is three to four times larger than the number of girls. Most children outgrow stuttering. About 1% or less of adults stutter (NIDCD, 2012b). Stuttering may also be observed after acute brain damage, and recent studies have identified genes associated with familial developmental stuttering (Kang et al., 2010). However, the causal mechanisms remain poorly understood.

Studies of developmental stuttering have suggested that the left perisylvian region is anatomically more heterogeneous in people who stutter than in controls (Foundas et al., 2001), with recent studies showing reductions in white-matter tracts innervating left sensorimotor cortex (Buchel et al., 2004; Sommer et al., 2002). In contrast, with developmental stuttering, acquired stuttering is more often associated with subcortical lesions, in particular in the basal ganglia, than with lesions in cortical speech and motor regions (Giraud et al., 2008; Carluer et al., 2000).

The prevalence of stuttering in spoken languages has led sign language researchers to inquire whether such impairment is evidenced in naturally occurring signed languages. There are no detailed case study reports of stuttering solely in the manual modality. A drawback in these studies is the lack of an operational definition 'sign stuttering' (see Whitebread (2004) for a proposed hypothetical description of features of sign stuttering based largely on analogy with spoken language stuttering).

Two survey studies polled teachers for the deaf and other professionals knowledgeable about the deaf community and sign language whether they had observed 'sign stuttering' (Silverman and Silverman, 1971). Both of these studies had respondents who stated they had observed various dysfluencies in manual communication, and based on these polls the authors have concluded such disorders exist.

However, there are several reasons why one should be cautious about such claims as both of these studies have serious limitations. For example, in these second-hand reports, it is not clear if the teachers are reporting on individuals who are communicating in a signed language or a mixture of speech and sign. In addition, in the study by Cosyns et al. (2009) in the report of signers with 'stuttering-like dysfluencies ' the youngest reported afflicted individual is 10 years old, which is at odds with the developmental expression of this disorder commonly seen in users of spoken language.

While both Silverman and Silverman (1971) and Cosyns et al. (2009) conclude that stuttering exists in sign language, there are numerous reasons to argue otherwise. First, the lack of available evidence itself is noteworthy. While developmental speech stuttering affects 5% of the populations (NIDCD), despite years of informed studies of sign language acquisition in native deaf children, there seem to be few, if any, credible reports of such a common and pronounced stage of developmental stuttering behaviors in deaf signing children.

While there are some reports that children who stutter may have impaired manual skills, these are most often observed in concert with speech tasks, with increased speech stuttering leading to increased dysfluencies in manual motor behaviors (Saltuklaroglu et al., 2009). This suggests that the manual impairments seen in stutterers are a secondary consequence of the speech dysfluencies. It is of further interest to note that cases of musicians who have experienced stuttering-like behaviors tend to be associated with wind instruments where there is necessity of coordination between phonatory, oral, and manual behaviors.

This is not to say dysfluent signing does not occur, indeed as noted, manifestations of dysfluent sign production have been reported in aphasia, Parkinson's disease, and autism. However to characterize such dysfluency in terms of evidence for stuttering seems premature. The fact that there is such a low prevalence for developmental stuttering in sign language is perhaps overall more informative than trying to assume a homology between all speech and sign disorders. Taken at face value, developmental stuttering is communication disorder that might be specific to speech and is not manifested in a signed language.

Sign Language-Specific Deficits

Sign language researchers have begun to explore the presence of specific sign language disorders in deaf children (Mason et al., 2010; Morgan et al., 2007; Quinto-Pozos et al., 2011). These initial efforts make use of survey and case study approaches to examine a small number of deaf children, who despite having deaf signing parents, have been identified by teachers as having atypical problems in signing. In this way researchers are confident that the deficits are not confounded by variations in language acquisition that may be evident in deaf children who are raised in home where the parents do not know sign language. The goal of these studies is to document the presence of such disorders and to determine whether the deficits observed reflect general cognitive and/or global language impairments (and thus affect both spoken and signed language) or whether the impairments are in fact unique to the sign linguistic system.

In one example, Morgan et al. (2007) reported the case of Paul, a 5-year-old native deaf signer of BSL with a sign language impairment which they entertain as a possible analogue of specific language impairment (SLI) evidenced in a sign language. SLI is diagnosed where a deficit in normal spoken language acquisition is found with no apparent cognitive, social, or neurological cause (Leonard, 1998). Children with SLI typically have difficulty with the acquisition of phonology and morphosyntax.

Pauls' expressive BSL was restricted to short sentences made up of one or two signs with very limited grammar. He mostly produced a series of single signs with few grammatical inflections and few polymorphemic signs. He often used pointing to pictures, himself and others to avoid proper sentence structure.

He tended to use affective facial expression rather than using the face in coordination with the hands to mark questions. His receptive abilities mirrored his expressive profile, whereby he showed normal receptive vocabulary abilities for simple signs, but showed significant impairment on more complex BSL sentences, especially those in which grammatical constructions are used to encode negation, noun–verb distinctions, spatial verbs, and classifiers. These constructions encode meanings through morphosyntactic rules. Interestingly, and in contrast to hearing children with SLI, Paul's phonology was considered age-appropriate. Thus Paul's language profile is characterized as intact phonology and lexicon but with impaired grammar reception and production. Paul's greatest areas of difficulty were with polymorphemic constructions such as agreement morphology and the classifier system. The existence of a version of SLI appearing in a signed language provides evidence that SLI may be a modality-independent impairment in complex language use, but one in which the particular characteristics of the language in question may show specific vulnerabilities. For example, in spoken language SLI may affect both phonology and morphosyntax, while in a signed language the timing characteristic of sign articulation (i.e., phonology) is less perilous, whereas morphosyntax, which is heavily represented in a variety of constructions, remains vulnerable in sign language SLI.

Language disorders, whether acute or developmentally based, affect not only hearing users of spoken languages but users of signed languages as well. By comparing the manifestation of language disorders across users of disparate languages and disparate modalities of expression provides powerful means to determine the causal mechanisms involved in language pathology.

Acknowledgments

This work was supported in part by NSF SBE-1041725 and NIH NIDCD R01 DC011538.

See also: Aphasia; Apraxia; Sign Language.

Bibliography

Agostino, R., et al., 1992. Sequential arm movements in patients with Parkinson's disease, Huntington's disease and dystonia. Brain 115, 1481–1495.

Anonymous, 2012a. Centers for Disease Control and Prevention (Online) Available at: http://cdc.gov/ncbdd/autism/index.html.

Anonymous, 2012b. National Institute on Deafness and Other Communication Disorders: Stuttering (Online) Available at: http://www.nidcd.nih.gov/health/voice/pages/stutter.aspx.

Berardelli, A., et al., 1986. Fast complex arm movements in Parkinson's disease. Journal of Neurology, Neurosurgery and Psychiatry 49, 1146–1149.

Bonvillian, J.D., Nelson, K.E., Rhyne, J.M., 1981. Sign language and autism. Journal of Autism and Developmental Disorders 11, 125–137.

Brentari, D., Poizner, H., 1994. A phonological analysis of a deaf Parkinsonian signer. Language and Cognitive Processes 9, 60–99.

Brentari, D., Poizner, H., Kegl, J., 1995. Aphasic and parkinsonian signing: difference in phonological disruption. Brain and Language 48 (1), 69–105.

Buchel, C., et al., 2004. White matter asymmetry in the human brain: a diffusion tensor MRI study. Cerebral Cortex 14 (9), 945–951.

Carluer, L., et al., 2000. Acquired and persistent stuttering as the main symptom of striatal infarction. Movement Disorders 15 (2), 343–346.

Chiarello, K., Mandel, M., 1982. Aphasia in a prelingually deaf woman. Brain 105, 29–51.

Corina, D.P., 1998. The Processing of Sign Language: Evidence from Aphasia. In: Whitaker, H., Stemmer, B. (Eds.), Handbook of Neurology. Academic Press, San Diego, pp. 313–329.

Corina, D.P., 2000. Some observations regarding paraphasia in American Sign Language. In: Emmorey, K., Lane, H.L. (Eds.), The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima. Lawrence Erlbaum Associates, Mahwah, pp. 493–507.

Corina, D.P., et al., 1999. Functional roles of Broca's area and SMG: evidence from cortical stimulation mapping in a deaf signer. NeuroImage 10 (5), 570–581.

Corina, D.P., et al., 1992. Dissociation between linguistic and nonlinguistic gestural systems: a case for compositionality. Brain and Language 43 (3), 414–447.

Cosyns, M., Van Herreweghe, A., Christiaens, G., Van Borsel, J., 2009. Stutter-like dysfluencies in Flemish Sign Language users. Clinical Linguistics and Phonetics 23 (10), 742–750.

Critchley, E.M., 1981. Speech disorders of Parkinsonism: a review. Journal of Neurology, Neurosurgery and Psychiatry 44, 751–758.

Dronkers, N.F., Redfern, B.B., Knight, R.T., 2000. The neural architecture of language disorders. In: Gazzaniga, M.S. (Ed.), The New Cognitive Neurosciences. MIT Press, Cambridge, pp. 949–958.

Dronkers, N.F., Redfern, B.B., Ludy, C.A., 1995. Lesion localization in chronic Wernicke's aphasia. Brain and Language 51 (1), 62–65.

Fawcett, R.G., 2005. Stroke-associated acquired stuttering. CNS Spectrums 10 (2), 94–95.

Foundas, A.L., et al., 2001. Anomalous anatomy of speech-language areas in adults with persistent developmental stuttering. Neurology 57, 207–215.

Gentil, M., Pollak, P., Perret, J., 1995. Parkinsonian dysarthia. Revue Neurologique 151, 105–112.

Giraud, A.-L., et al., 2008. Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. Brain and Language 104, 190–199.

Goodglass, H., 1993. Understanding Aphasia. Academic Press, San Diego.

Heilman, K.M., Rothi, L.J., Valenstein, E., 1982. Two forms of ideomotor apraxia. Neurology 32 (4), 342–346.

Hickok, G., Bellugi, U., Klima, E.S., 1996. The neurobiology of sign language and its implications for the neural basis of language. Nature 381 (6584), 699–702.

Hickok, G., Love-Geffen, T., Klima, E.S., 2002. Role of the left hemisphere in sign language comprehension. Brain Lang 82 (2), 167–178.

Jure, R., Rapin, I., Tuchman, R.F., 1991. Hearing-impaired autistic children. Developmental Medicine and Child Neurology 33 (12), 1062–1072.

Kang, C., et al., 2010. Mutations in the lysosomal enzyme–targeting pathway and persistent stuttering. The New England Journal of Medicine 362, 677–685.

Kiyaga, N.B., Moores, D.F., 2003. Deafness in Sub-Saharan Africa. American Annals of the Deaf 148 (1), 18–24.

Leischner, A., 1943. Die "Apasie" der Taubstummen. Archiv fur Psychiatrie und Nervenkrankheiten 115, 469–548.

Litvan, I., 1996. Parkinson's disease. In: Beaumont, J.G., Kenealy, P.M., Rogers, M.J.C. (Eds.), The Blackwell Dictionary of Neuropsychology. Blackwell Publishing, Oxford.

Leonard, L., 1998. Children with specific language impairment. MIT Press, Cambridge.

Marshall, J., et al., 2004. Aphasia in a user of British sign language: dissociation between sign and gesture. Cognitive Neuropsychology 21 (5), 537–554.

Mason, K., et al., 2010. Identifying specific language impairment in deaf children acquiring British sign language: implications for theory and practice. British Journal of Developmental Psychology 28 (1), 33–49.

Marshall, J., Atkinson, J., Woll, B., 2005. Aphasia in a bilingual user of British sign language and English: effects of cross-linguistic cues. Cognitive Neuropsychology 22 (6), 719–736.

McCarthy, R.A., Warrington, E.K., 1990. Cognitive Neuropsychology. Academic Press, San Diego.

Morgan, G., Herman, R., Woll, B., 2007. Language impairment in sign language: breakthroughs and puzzles. International Journal of Communication Disorders 35, 95–116.

Poizner, H., Klima, E.S., Bellugi, U., 1987. What the Hands Reveal about the Brain. MIT Press, Cambridge(Massachusetts).

Quinto-Pozos, D., Forber-Pratt, A.J., Singleton, J.L., 2011. Do developmental communicative disorders exist in the signed modality? Perspectives from professionals. Language, Speech and Hearing Services in Schools 42, 423–443.

Rogers, S.J., Pennington, B.F., 1991. A theoretical approach to the deficits in infantile autism. Development and Psychopathology 3 (2), 137–162.

Rosenhall, U., et al., 1999. Autism and hearing loss. Journal of Autism and Developmental Disorders 29 (5), 349–357.

Rothi, L.J.G., Ochipa, C., Heilman, K.M., 1991. A cognitive neuropsychological model of limb praxis. Cognitive Neuropsychology 8, 443–458.

Saltuklaroglu, T., Teulings, H., Robbins, M., 2009. Differential levels of speech and manual dysfluency in adults who stutter during simultaneous drawing and speaking tasks. Human Movement Science 28, 643–654.

Sandler, W., Lillo-Martin, D.C., 2006. Sign Language and Linguistic Universals. Cambridge University Press, New York.

Shield, A., Meier, R.P., 2012. Palm reversal errors in native-signing children with autism. Journal of Communication Disorders 45, 439–454.

Silverman, F.H., Silverman, E., 1971. Stutter-like behavior in manual communication of the deaf. Perceptual and Motor Skills 33, 45–46.

Sommer, M., et al., 2002. Disconnection of speech-relevant brain areas in persistent developmental stuttering. Lancet 360, 380–383.

Szymanski, C.A., Brice, P.J., 2008. When autism and deafness coexist in children: what we know now. Odyssey: New Directions in Deaf Education 9 (1), 10–15.

Stemmer, B., Whitaker, H.A. (Eds.), 1998. Handbook of Neurolinguistics. Academic Press, San Diego.

Szymanski, C.A., Brice, P.J., Lam, K.H., Hotto, S.A., 2012. Deaf Children with Autism Spectrum Disorders (Online) Available at: http://dx.doi.org/10.1007/s10803-012-1452-9.

Whitebread, G., 2004. Gallaudet University. (Online) Available at: http://www.gallaudet.edu/documents/academic/honors/stuttering_thesis_final.pdf.

Williams, J., Whiten, A., Singh, T., 2004. A systematic review of action imitation in autistic spectrum disorder. Journal of Autism and Developmental Disorders 34, 285–299.

Sign Language: Psychological and Neural Aspects

David P Corina, University of Washington, Seattle, WA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14071–14075, © 2001, Elsevier Ltd.

Abstract

This article presents an overview of psychological and neural aspects of sign languages. A brief discussion of cultural and linguistic issues of deafness and the structure of American Sign Language (ASL) provide a framework to understand findings from developmental psychology, psycholinguistics, cognitive psychology, and neuropsychological studies. Signed languages are naturally occurring human languages of deaf communities. Sign language acquisition follows a developmental trajectory similar to spoken languages. The recognition of a sign is influenced by the formational properties of sign language structure and occurs faster than spoken language recognition. While signs are articulated slower than spoken words, the proposition rate for sign and speech is the same. Memory for signs exhibits patterns of interference and forgetting that are similar to those found for speech. Studies of attention and perception indicate that the early use of sign language may enhance certain aspects of nonlanguage visual perception, such as the perception of motion. Neuropsychological and functional imaging studies show that left hemisphere regions are important in both sign and spoken language processing. Left hemisphere damage produces aphasia in signers. The existence of languages expressed in different modalities (i.e., oral–aural, manual–visual) provides a unique opportunity to explore and distinguish those properties shared by all human languages from those that arise in response to the modality in which the language is expressed.

Signed languages of the deaf are naturally occurring human languages. The existence of languages expressed in different modalities (i.e., oral–aural, manual–visual) provides a unique opportunity to explore and distinguish those properties shared by all human languages from those that arise in response to the modality in which the language is expressed. Despite the differences in language form, signed languages possess formal linguistic properties found in spoken languages. Sign language acquisition follows a developmental trajectory similar to spoken languages. Memory for signs exhibits patterns of interference and forgetting that are similar to those found for speech. Early use of sign language may enhance certain aspects of nonlanguage visual perception. Neuropsychological studies show that left hemisphere regions are important in both spoken and sign language processing.

Linguistic Principles of American Sign Language

Language and Deaf Culture

Sign languages are naturally occurring manual languages that arise in communities of deaf individuals. These manual communication systems are fully expressive, systematic human languages and are not merely conventionalized systems of pantomime or manual codifications of a spoken language. Many types of deafness are inheritable and it is not unusual to find isolated communities of deaf individuals, who have developed complex manual languages (Groce, 1985). The term Deaf Community has been used to describe a sociolinguistic entity that plays a crucial role in a deaf person's exposure to and acceptance of sign language (Padden and Humphries, 1988). American Sign Language (ASL), used by members of the Deaf community in the USA and Canada, is only one of many sign languages of the world, but it is the one that has been studied most extensively.

Structure of ASL

A sign consists of a hand configuration that travels about a movement path and is directed to or about a specific body location. Sign languages differ from one another in the inventories and compositions of handshapes, movements, and locations used to signal linguistic contrasts just as spoken language differ from one another in the selection of sounds used, and how those sounds are organized into words. Many sign languages incorporate a subsystem in which orthographic symbols used in the surrounding spoken language communities are represented manually on the hands. One example is the American English manual alphabet, which is produced on one hand and allows users of ASL to represent English lexical items.

All human languages, whether spoken or signed, exhibit levels of structure that govern the composition and formation of word forms and specify how words combine into sentences. In formal linguistic analyses, these structural levels are referred to as phonology, morphology, and the syntax of the language. In this context, phonological organization refers to the patterning of the abstract formational units of a natural language (Coulter and Anderson, 1993). Compared to spoken languages in which contrastive units (for example, phonemes) are largely arranged in a linear fashion, signed languages exhibit simultaneous layering of information. For example, in a sign, a handshape will co-occur with, rather than follow sequentially, a distinct movement pattern.

ASL exhibits complex morphology. Morphological markings in ASL are expressed as dynamic movement patterns overlaid on a more basic sign form. These nested morphological forms stand in contrast to the linear suffixation common in spoken languages (Klima and Bellugi, 1979). The prevalence of

simultaneous layering of phonological content and the nested morphological devices observed across many different sign languages likely reflect an influence of modality on the realization of linguistic structure. Thus the shape of human languages reflects the constraints and affordances imposed by the articulator systems involved in transmission of the signal (i.e., oral vs manual) and the receptive mechanisms for decoding the signal (i.e., auditory vs visual).

A unique property of ASL linguistic structure is the reliance upon visuospatial mechanisms to signal linguistic contrasts and relations. One example concerns the use of facial expressions in ASL. In ASL, certain syntactic and adverbial constructions are marked by specific and obligatory facial expressions (Liddell, 1980). These linguistic facial expressions differ significantly in appearance and execution of affective facial expressions. A second example concerns the use of inflectional morphology to express subject and object relationships. At the syntactic level, nominals introduced into the discourse are assigned arbitrary reference points along a horizontal plane in the signing space. Signs with pronominal function are directed toward these points, and the class of verbs that require subject/object agreement obligatorily move between these points (Lillo-Martin and Klima, 1990). Thus, whereas many spoken languages represent grammatical functions through case marking or linear ordering, in ASL grammatical function is expressed through spatial mechanisms. This same system of spatial reference, when used across sentences, serves as a means of discourse cohesion (Winston, 1995).

Sign Language Acquisition

Children exposed to signed languages from birth acquire these languages on a similar maturational timetable as children exposed to spoken languages (Meier, 1991). Prelinguistic infants, whether normally hearing or deaf, engage in vocal play commonly known as babbling. Recent research has shown that prelinguistic gestural play referred to as manual babbling will accompany vocal babbling. There appear to be significant continuities between prelinguistic gesture and early signs in deaf children exposed to ASL, though questions remain as to whether manual babbling is constrained by predominantly motoric or linguistic factors (Cheek et al., 2001; Petitto and Marantette, 1991).

Between 10 and 12 months of age, children reared in signing homes begin to produce their first signs, with two-sign combinations appearing at approximately 18 months. Some research has suggested a precocious early vocabulary development in signing children. However, these reports are likely to be a reflection of parents' and experimenters' earlier recognition of signs compared to words rather than underlying differences in development of symbolic capacities of signing and speaking children.

Signing infants produce the same range of grammatical errors in signing that have been documented for spoken language, including phonological substitutions, morphological overregularizations, and anaphoric referencing confusions (Petitto, 2000). For example, in the phonological domain a deaf child will often use only a subset of the handshapes found in the adult inventory, opting for a simpler set of 'unmarked' handshapes. This is similar to the restricted range of consonant usage common in children acquiring a spoken language.

Psycholinguistic Aspects of Sign Language

Psycholinguistic studies of ASL have examined how the different signaling characteristics of the language impact transmission and recognition. Signs take longer to produce than comparable word forms. The average number of words per second in running speech is about 4–5, compared with 2–3 signs per second in fluent signing. However, despite differences in word transmission rate, the proposition rate for speech and sign is the same, roughly one proposition every 1–2 s. Compare, for example, the English phrase 'I have already been to California' with the ASL equivalent, which can be succinctly signed using three monosyllabic signs, glossed as *finish touch california*.

Sign Recognition

Studies of sign recognition have examined how signs are recognized in time. Recognition appears to follow a systematic pattern in which information about the location of the sign is reliably identified first, followed by handshape information and finally the movement. Identification of the movement reliably leads to the identification of the sign. Interestingly, despite the slower articulation of a sign compared to a word, sign recognition appears to be faster than word recognition. Specifically, it has been observed that proportionally less of the signal needs to be processed in order to uniquely identify a sign compared to a spoken word. For example, one study reports that only 240 ms or 35% of a sign has to be seen before a sign is identified (Emmorey and Corina, 1990). In comparable studies of spoken English, Grosjean (1980) reports that 330 ms or 83% of a word has to be heard before a word can be identified. This finding is due, in part, to the simultaneous patterning of phonological information in signs compared to the more linear patterning of phonological information characteristic of spoken languages. These structural differences in turn have implications for the organization of lexical neighborhoods. Spoken languages, such as English, may have many words that share in their initial phonological structures (tram, tramp, trampoline, etc.) leading to greater coactivation (and thus longer processing time) during word recognition. In contrast, ASL sign forms are often formationally quite distinct from one another, permitting quicker selection of a lexical unique entry.

Memory for Signs

Efforts to explore how the demands of sign language processing influence memory and attention have led to several significant findings. Early studies of memory for lists of signs report classic patterns of forgetting and interference, including serial position effects of primacy and recency (i.e., signs at the beginning and the end of a list are better remembered than items in the middle). Likewise, when sign lists are composed of phonologically similar sign forms, signers exhibit poorer recall (Klima and Bellugi, 1979). This result is similar to what has been

reported for users of spoken language, where subjects exhibit poorer memory for lists of similarly sounding words (Conrad and Hull, 1964). These findings indicate that signs, like words, are encoded into short-term memory in a phonological or articulatory code rather than in a semantic code.

More recent work has assessed whether Baddeley's (1986) working memory model pertains to sign language processing. This model includes components that encode linguistic and visuospatial information as well as a central executive which mediates between immediate and long-term memory. Given the spatial nature of the sign signal, the question of which working memory component(s) is engaged during sign language processing is particularly interesting. Wilson and Emmorey (2000) have shown word-length effects in signs; it is easier to maintain a cohort of short signs (monosyllabic signs) than long signs (polysyllabic signs) in working memory. They also report effects of 'articulatory suppression.' In these studies, requiring a signer to produce repetitive, signlike hand movements while encoding and maintaining signs in memory has a detrimental effect on aspects of recall. Once again, analogous effects are known to exist for spoken languages and these phenomena provide support for a multicomponent model of linguistic working memory that includes both an articulatory loop and a phonological store. Finally, under some circumstances deaf signers are able to utilize linguistically relevant spatial information to encode signs, suggesting the engagement of a visuospatial component of working memory.

Perception and Attention

Recent experiments with native users of signed languages have shown that experience with a visual sign language may improve or alter visual perception of nonlanguage stimuli. For example, compared to hearing nonsigners, deaf signers have been shown to possess enhanced or altered perception along several visual dimensions, such as motion processing (Bosworth and Dobkins, 1999), mental rotation, and processing of facial features (Emmorey, 2001). For example, the ability to detect and attend selectively to peripheral, but not central, targets in vision is enhanced in signers (Neville, 1991). Consistent with this finding is evidence for increased functional connectivity in neural areas mediating peripheral visual motion in the deaf (Bavelier et al., 2000).

As motion processing, mental rotation, and facial processing underlie aspects of sign language comprehension, these perceptual changes have been attributed to experience with sign language. Moreover, several of these studies have reported such enhancements in hearing signers raised in deaf signing households. These findings provide further evidence that these visual perception enhancements are related to the acquisition of a visual manual language, and are not due to compensatory mechanisms developed as a result of auditory deprivation.

Neural Organization of Signed Language

Hemispheric Specialization

One of the most significant findings in neuropsychology is that the two cerebral hemispheres show complementary functional specialization, whereby the left hemisphere mediates language behaviors while the right hemisphere mediates visuospatial abilities. As noted, signed languages make significant use of visuospatial mechanisms to convey linguistic information. Thus sign languages exhibit properties for which each of the cerebral hemispheres shows specialization. Neuropsychological studies of brain injured deaf signers and functional imaging studies of brain intact signers have provided insights into the neural systems underlying sign language processing.

Aphasia in Signed Language

Aphasia refers to acquired impairments in the use of language following damage to the perisylvian region of the left hemisphere. Neuropsychological case studies of deaf signers convincingly demonstrate that aphasia in signed language is also found after left hemisphere perisylvian damage (Poizner et al., 1987). Moreover, within the left hemisphere, production and comprehension impairments follow the well-established anterior versus posterior dichotomy. Studies have documented that frontal anterior damage leads to Broca-like sign aphasia. In these cases, normal fluent signing is reduced to effortful, single-sign, 'telegraphic' output with little morphological complexity (such as verb agreement). Comprehension, however, is left largely intact. Wernicke-like sign aphasia following damage to the posterior third of the perisylvian region presents with fluent but often semantically opaque output and comprehension also suffers. In cases of left hemisphere damage, the occurrence of sign language paraphasias may be observed. Paraphasia may be formationally disordered (for example, substituting an incorrect handshape) or semantically disordered (producing a word that is semantically related to the intended form). These investigations serve to underscore the importance of left hemisphere structures for the mediation of signed languages and illustrate that sign language breakdown is not haphazard, but rather honors linguist boundaries (Corina, 2000).

The effects of cortical damage to primary language areas that result in aphasic disturbance can be differentiated from more general impairments of movement. For example, Parkinson's disease leads to errors involving timing, scope, and precision of general movements including, but not limited to, those involved in speech and signing. These errors produce phonetic disruptions in signing, rather than higher-level phonemic disruptions that are apparent in aphasic signing (Corina, 1999).

Apraxia is defined as an impairment of the execution of a learned movement (e.g., saluting, knowing how to work a knife and fork). Lesions associated with the left inferior parietal lobe result in an inability to perform and comprehend gestures (Gonzalez Rothi and Heilman, 1997). Convincing dissociations of sign language impairment with well-preserved praxic abilities have been reported. In one case, a subject with marked sign language aphasia affecting both production and comprehension produced unencumbered pantomime. Moreover, both comprehension and production of pantomime were found to be better preserved than was sign language. These data indicate that language impairments following left hemisphere damage are not attributable to undifferentiated symbolic impairments and demonstrate that ASL is not simply an elaborate pantomimic system.

Role of the Right Hemisphere in ASL

Studies of signers with right hemisphere damage (RHD) have reported significant visuospatial disruption in nonlinguistic domains (e.g., face processing, drawing, block construction, route finding, etc.) but have reported only minimal language disruption. Problems in the organization of discourse have been observed in RHD signers, as have also been reported in users of spoken languages. More controversial are the influences of visuospatial impairment in the use of highly spatialized components of the language, such as complex verb agreement and the classifier system. Further work is needed to understand whether these infrequently reported impairments (both in comprehension and production) reflect core linguistic deficits or rather reflect secondary effects of impaired visuospatial processing.

Functional Imaging

Recent studies using functional brain imaging have explored the neural organization of language in users of signed languages. These studies have consistently found participation of classic left hemisphere perisylvian language areas in the mediation of sign language in profoundly deaf, lifelong signers. For example, positron emission tomography studies of production in British Sign Language (McGuire et al., 1997) and Langue des Signes Quebecoise (Petitto et al., 2000) reported deaf subjects activated left inferior frontal regions, regions similar to those that mediate speech in hearing subjects. Researchers using functional magnetic resonance imaging techniques in investigating comprehension of ASL in deaf and hearing native users of signed language have shown significant activations in frontal opercular areas (including Broca's area and dorsolateral prefrontal cortex) as well as in posterior temporal areas (such as Wernicke's area and the angular gyrus) (Neville et al., 1998). Also reported was extensive activation in right hemisphere regions. Subsequent studies have conferred that aspects of the right posterior hemisphere activation appear to be unique to sign language processing and present only in signers who acquired sign language from birth (Newman et al., 1998).

Investigations of psychological and neural aspects of signing reveal strong commonalities in the development and cognitive processing of signed and spoken languages despite major differences in the surface form of these languages. Early exposure to a sign language may lead to the enhancements in the specific neural systems underlying visual processing. There appears to be a strong biological predisposition for left hemisphere structures in the mediation of language, regardless of the modality of expression.

See also: Language Acquisition; Neural Basis of Language Development; Sign Language.

Bibliography

Baddeley, A., 1986. Working Memory. Oxford University Press, New York.

Bavelier, D., Tomann, A., Hutton, C., Mitchell, T., Corina, D., Liu, G., Neville, H., 2000. Visual attention at the periphery is enhanced in congenitally deaf individuals. Journal of Neuroscience 20, RC93.

Bosworth, R., Dobkins, K., 1999. Left-hemisphere dominance for motion processing in deaf signers. Psychological Science 10, 256–262.

Cheek, A., Cormier, K., Repp, A., Meier, R., 2001. Prelinguistic gesture predicts mastery and error in the production of early signs. Language 77, 292–323.

Corina, D., 1999. Neural disorders of language and movement: evidence from American sign language. In: Messing, L., Campbell, R. (Eds.), Gesture, Speech and Sign. Oxford University Press, New York, pp. 27–44.

Corina, D., 2000. Some observations regarding paraphasia in American sign language. In: Emmorey, K., Lane, H. (Eds.), The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 493–507.

Coulter, G., Anderson, S., 1993. Introduction. In: Coulter, G. (Ed.), Phonetics and Phonology: Current Issues in ASL Phonology. Academic Press, San Diego, CA.

Conrad, R., Hull, A., 1964. Information, acoustic confusion and memory span. British Journal of Psychology 55, 429–432.

Emmorey, K., 2001. Language, Cognition, and the Brain: Insights from Sign Language Research. Lawrence Erlbaum Associates, Mahwah, NJ.

Emmorey, K., Corina, D., 1990. Lexical recognition in sign language: effects of phonetic structure and morphology. Perceptual and Motor Skills 71, 1227–1252.

Gonzalez Rothi, L., Heilman, K., 1997. Introduction to limb apraxia. In: Gonzalez Rothi, L., Heilman, K. (Eds.), Apraxia: The Neuropsychology of Action. Psychology, Hove, UK.

Grosjean, F., 1980. Spoken word recognition processes and the gating paradigm. Perception and Psychophysics 28, 267–283.

Groce, J., 1985. Everyone Here Spoke Sign Language. Harvard University Press, Cambridge, MA.

Klima, E., Bellugi, U., 1979. The Signs of Language. Harvard University Press, Cambridge, MA.

Liddell, S., 1980. American Sign Language Syntax. Mouton, The Hague; The Netherlands.

Lillo-Martin, D., Klima, E., 1990. Pointing out differences: ASL pronouns in syntactic theory. In: Fisher, S., Siple, P. (Eds.), Theoretical Issues in Sign Language Research I: Linguistics. University of Chicago Press, Chicago.

McGuire, P., Robertson, D., Thacker, A., David, A., Frackowiak, R., Frith, C., 1997. Neural correlates of thinking in sign language. Neuroreport 8, 695–698.

Meier, R., 1991. Language acquisition by deaf children. American Scientist 79, 60–70.

Neville, H., 1991. Neurobiology of cognitive and language processing: effects of early experience. In: Gibson, K., Peterson, A. (Eds.), Brain Maturation and Cognitive Development: Comparative and Cross-Cultural Perspectives. Aldine de Gruyter Press, Hawthorne, NY, pp. 355–380.

Neville, H., Bavelier, D., Corina, D., Rauschecker, J., Karni, A., Lalwani, A., Braun, A., Clark, V., Jezzard, P., Turner, R., 1998. Cerebral organization for language in deaf and hearing subjects: biological constraints and effects of experience. Proceedings of the National Academy of Science 90, 922–929.

Newman, A., Corina, D., Tomann, A., Bavelier, D., Jezzard, P., Braun, A., Clark, V., Mitchell, T., Neville, H., 1998. Effects of age of acquisition on cortical organization for American sign language: an fMRI study. NeuroImage 7, S194.

Padden, C., Humphries, T., 1988. Deaf in America: Voices from a Culture. Harvard University Press, Cambridge, MA.

Petitto, L., 2000. The acquisition of natural signed languages: lessons in the nature of human language and its biological foundations. In: Chamberlain, C., Morford, J. (Eds.), Language Acquisition by Eye. Lawrence Erlbaum Associates, Mahwah, NJ.

Petitto, L., Marantette, P., 1991. Babbling in the manual mode: evidence for the ontogeny of language. Science 251, 1493–1496.

Petitto, L., Zatorre, R., Gauna, K., Nikelski, E., Dostie, D., Evans, A., 2000. Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. Proceedings of the National Academy of Science 97, 13961–13966.

Poizner, H., Klima, E., Bellugi, U., 1987. What the Hands Reveal about the Brain. MIT Press, Cambridge, MA.

Wilson, M., Emmorey, K., 2000. When does modality matter? Evidence from ASL on the nature of working memory. In: Emmorey, K., Lane, H. (Eds.), The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 135–142.

Winston, E., 1995. Spatial mapping in comparative discourse frames. In: Emmorey, K., Reilly, J. (Eds.), Language, Gesture, and Space. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 87–114.

Signal Detection Theory

Michael Hautus, The University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N.A. Macmillan, volume 21, 14075–14078, © 2001 Elsevier Ltd.

Abstract

Signal detection theory (SDT) provides a framework for the quantification of sensitivity in sensory-perceptual tasks, and can be applied to other real-world applications. The measure d' quantifies the resolving power of the perceptual system, while controlling for the response bias of the judge. Given the same stimuli and judge, resolving power should be invariant irrespective of the psychophysical task employed. SDT models can unify results over different tasks, and different decision strategies – the mapping of perceptual information to a response – allowed for by some tasks.

Signal detection theory (SDT) provides a framework for the analysis of performance in tasks for which stimuli (or events) are confusable. SDT can be applied to applications in engineering, the assessment of sensory or cognitive performance in the fields of psychology and sensory evaluation, the assessment of diagnostic accuracy in the medical sciences, and forecasting in areas such as meteorology.

A distinction is made between two fundamental paradigms: discrimination and detection. The discrimination paradigm quantifies how well individuals (or systems) can resolve the difference between two stimuli, whereas the detection paradigm quantifies an individual's ability to resolve the difference between the presence and absence of a stimulus. Much of the early work in SDT focused on detection. Given that the general application of SDT to both paradigms is essentially the same, the remainder of this article will focus primarily on discrimination.

The Single-Interval Task

The single-interval task is the standard experimental design in SDT. It is also known as the Yes–No task, and the A–Not-A task, in the psychophysics and sensory evaluation literature, respectively. On a trial, one of two stimuli (S_1 or S_2) is presented to the judge, who must indicate whether S_1 or S_2 was presented. Table 1 gives the four possible outcomes for a trial when the judge issues a binary response (e.g., 'S_2' or 'S_1;' 'yes' or 'no;' 'A' or 'Not-A'). Hits and correct rejections are correct responses, and misses and false alarms are incorrect responses. The medical and epidemiology literature uses a different terminology, also given in Table 1.

Consider for Example 1 an auditory experiment designed to measure the ability of a judge to discriminate the difference in amplitude of two tones, S_1 and S_2, with a relatively lower and higher intensity, respectively. On a trial, either S_1 or S_2 is presented to the judge, who must indicate which was presented. If the intensity difference between the tones is small, the sensory evidence upon which the judgment is made will be confusable; that is, across repeated presentations, the sensory magnitude arising from S_1 will sometimes be lower than that for S_2, and at other times it will be higher. This variation in sensory magnitude is a consequence of many factors, one of which is moment-to-moment variability in activity within the sensory system. The degree of sensory variability relative to the difference between the stimuli limits the performance of the judge.

This sensory variability is attributed to the addition of random noise, often called sensory noise, to the presented stimulus. The original terminology of SDT arose from this viewpoint. In a detection task, where a low-level stimulus either is or is not presented on a trial, the stimulus is called the signal. When no stimulus is presented, sensory noise still influences the detection decision; consequently, trials on which stimuli are not presented are called noise trials. When the signal is presented, the sensory noise is also present; hence, these are called signal-plus-noise trials. The combination of sensory noise and stimulus-related neural activity used by the judge in decision making is the sensory evidence.

Returning to Example 1, consider the data in Table 2, which details the frequency of each outcome from the presentation of S_1 and S_2, each 50 times. The frequencies are typically converted to rates; for example, the hit rate, H, is obtained by dividing the frequency of hits by the number of S_2

Table 1 Four outcomes of a trial in the single-interval task as named in psychology (first entry) and medicine (second entry)

	Binary response	
Stimulus	'S_2'	'S_1'
S_2	Hit True positive (sensitivity)	Miss False negative
S_1	False alarm False positive	Correct rejection True negative (specificity)

Table 2 Frequencies (top entry) and rates (bottom entry) of each outcome for Example 1, with 50 presentations each of S_1 and S_2

	Binary response	
Stimulus	'S_2'	'S_1'
S_2	Hits = 37 $H = .74$	Misses = 13 $M = .26$
S_1	False alarms = 18 $F = .36$	Correct rejections = 32 CR = .64

presentations ($H = 37/50 = .74$). The divisor for the false-alarm (F) and correct-rejection (CR) rates would be the number of S_1 presentations. The rates for the two outcomes for S_1 and two outcomes for S_2 each sum to one; thus, the experiment can be summarized by just two rates. Tradition is to use H and F.

The Standard SDT Model

A common assumption about the sensory noise is that it is normally distributed. Another less common assumption is that the variability of the sensory noise is independent of the difference between S_1 and S_2; this is the so-called equal-variance assumption. These two assumptions underpin the standard version of the SDT model, sometimes called the normal–normal equal-variance model (NEV). Figure 1 depicts the decision space for this model. The two distributions represent the sensory evidence arising from S_1 and S_2. In perceptual tasks, this evidence typically represents the moment-to-moment variation in the neural representation of the stimuli. However, it can also represent physical variation in S_1 and S_2 across presentations; this is typical of class-discrimination experiments, in which S_1 and S_2 are drawn from two sets of stimuli (e.g., old and new words in a word recognition experiment), or experiments in which random variation is deliberately added to the stimuli. In medical applications, this evidence can represent the magnitude of a physical measure, sometimes called a biomarker, which may have use in diagnosing the presence (S_2) or absence (S_1) of an abnormality. Such measures commonly show a distribution of magnitudes across patients, or time.

Faced with this evidence, a judge must make a decision. The solid vertical line in Figure 1 is a decision criterion. Evidence that falls above the criterion leads to the response 'S_2' and results in hits and false alarms. 'S_1' is the response given to evidence that falls below the criterion, leading to misses and correct rejections. The areas under the distributions above and below the criterion provide the four outcome rates.

Sensitivity and Response Bias

The most common formulation of the model given in Figure 1 has two parameters. The first, d', is the difference between the means of the two standardized distributions, and represents the ability of the perceptual system to distinguish between S_1 and S_2. The second, x_c, is the value of the evidence axis on which the criterion is located, and represents the judge's propensity for responding one way or the other. x_c is rarely used directly. Instead, a parameter, c, is defined as the distance of the criterion from a point midway between the two distributions, hence $c = x_c - d'/2$. In Figure 1, $d' = 1$ and $x_c = -.69$; hence $c = -1.19$. The shaded areas yield $F = .75$ and $H = .95$. When estimated from data, d' and c are measures of sensitivity and response bias, respectively.

The standard SDT model can now be applied to the data from Example 1. The two parameters can be estimated using the equations $d' = z(H) - z(F)$ and $c = -{}^1/_2\,[z(H) + z(F)]$, where $z(\bullet)$ is the inverse-normal transform. For Example 1, $d' = z(.74) - z(.36) = .64 - (-.36) = 1.00$, and $c = -{}^1/_2\,[z(.74) + z(.36)] = -.14$. The distributions in Figure 1 are appropriate for these data, but with the criterion indicated by the dashed vertical line. For binary data, the model is saturated, and goodness-of-fit (GOF) cannot be assessed.

A common method used to increase the available degrees of freedom of the model is to increase the number of response options, often referred to as confidence ratings. This adaptation of the single-interval task provides more than one F, H pair from a single experiment. For Example 2, consider the same experiment as before, except that now the judge responds on a six-point confidence-rating scale. A response of '1' could indicate certainty that the stimulus presented was S_2, while '6' could indicate certainty that it was S_1. Other values of response indicate levels of certainty between these two end points; for example, '4' may indicate 'uncertain, but think it was S_1. Table 3 shows the frequencies of each response to the presentation of S_1 and S_2. F and H for each response are obtained by cumulating frequencies for all response ratings up to the target response, and then dividing by the total number of presentations of S_2 for H and of S_1 for F. This results in $n_r - 1$ estimates of F, H pairs, where n_r is the number of ratings used on the response scale. The determination of measures of sensitivity and bias from these data benefits from the introduction of a different representation of the model illustrated in Figure 1.

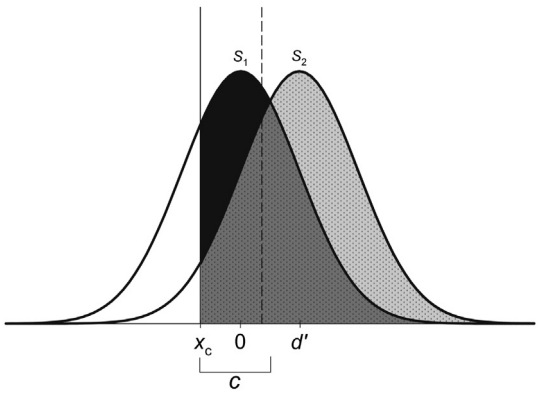

Figure 1 Decision space for the single-interval task.

Receiver Operating Characteristic Curves

The receiver operating characteristic (ROC) curve is obtained by setting d' to a fixed value and varying the criterion, c, over all

Table 3 Frequency of each rating response given to presentations of S_1 and S_2

		'S_2'		Rating response			'S_1'
Stimulus		1	2	3	4	5	6
S_2	Frequency	22	15	6	3	3	1
	H	.44	.74	.86	.92	.98	1.00
S_1	Frequency	5	11	13	10	8	3
	F	.10	.32	.58	.78	.94	1.00

Accumulation of frequencies from left to right, and division of each sum by the number of presentations (50) gives the estimated values of F and H.

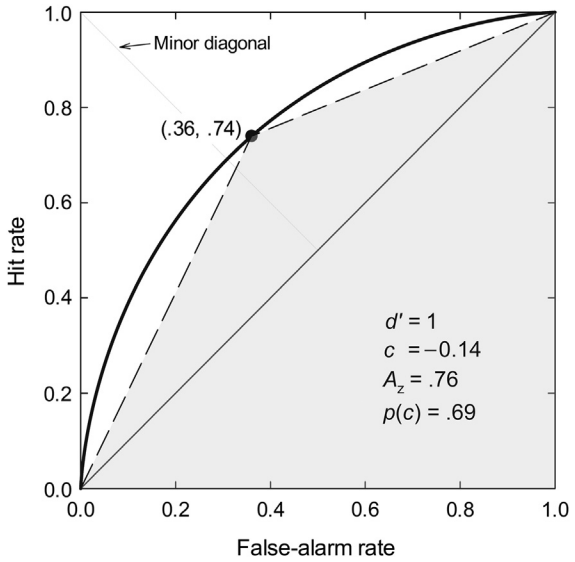

Figure 2 ROC curve for the binary-response data of Example 1.

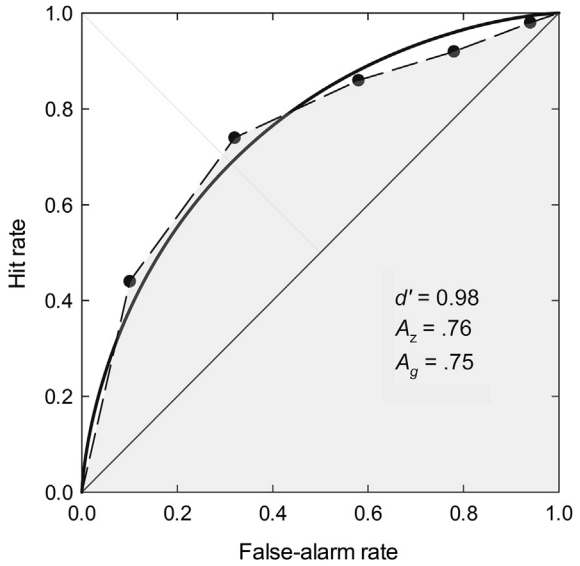

Figure 3 ROC curve for the rating-response data of Example 2.

possible values. Each value of c yields an F, H pair. The ROC curve is a plot of H against F in the unit square. The ROC curve illustrated in Figure 2 has $d' = 1$. Any point on the curve represents a particular value of the criterion. The circle, at $F = .36$, $H = .74$, corresponds to $c = -.14$ and represents the data from Example 1.

The shape of the ROC curve for the NEV model is concave downward, symmetrical about the minor diagonal of the ROC square, and passes through the points (0,0) and (1,1). The concavity of the ROC curve is proportional to d', and while d' itself is not easily quantifiable directly from the ROC curve, this relationship suggests another measure of sensitivity, the area under the ROC curve, A_z (sometimes referred to as area under the curve; AUC). A_z can range between half and one, and is equal to $\Phi(d'/\sqrt{2})$, where $\Phi(\bullet)$ is the standard normal distribution function. Chance performance corresponds to $d' = 0$ and $A_z = .5$, whereas perfect performance has $d' = \infty$ and $A_z = 1$.

Another index of sensitivity, proportion correct ($p(c)$), can be determined from the ROC space. It is the area under an ROC curve formed by straight lines that join the points (0,0), (F,H), and (1,1), and is indicated by the shaded area in Figure 2. Determination of $p(c)$ in terms of F and H using simple geometrical principles yields $p(c) = {}^1/_2\,[H + (1-F)]$. Given that $1-F$ is the CR rate, it is clear that $p(c)$ is half the sum of the proportions for each type of correct response. For the case of Example 1, $p(c)$ is .69. Points at different locations on the same ROC curve will yield different values of $p(c)$, demonstrating that $p(c)$ is not bias-free, unlike d', or the full area under the ROC curve, A_z, which for Example 1 is .76.

The points in Figure 3 define the empirical ROC based on the data from Example 2 (see Table 3). No single ROC curve based on the NEV model could pass through all points. Thus, the best-fitting curve must be determined; note that both F and H are subject to sampling variability.

There are two basic approaches to fitting ROC curves to empirical ROC data. The ordinal regression approach, which models the probability frequencies of ratings, is more widespread, and can be traced back at least to the algorithms published by Dorfman and Alf (1969). The direct modeling approach is a more recent formulation. It directly models the relationship between the frequency of ratings for S_1 and S_2 (Pepe, 2003). Both approaches allow for covariates in the models. While covariate modeling in ROC analysis appears infrequently in the psychological literature, it is prevalent in medical literature.

The ROC curve in Figure 3 was fitted using the ordinal regression approach. The estimate of d' is .98 and A_z is .76. There are also five criterial estimates ($c = -1.91, -1.16, -0.68, -0.10, 0.69$). The GOF of the curve can be assessed using Pearson's chi-square (χ^2). In this case, $\chi^2 = 2.71$, there are 4 degrees of freedom, and yields a probability of .608. This probability – if the standard assumptions are met for χ^2 – can be interpreted as the probability that the data arose from the model, given that the model is correct. In practice, small probabilities (e.g., $p < .001$) are employed to reject a model (Press et al., 1992).

For rating data, $p(c)$ can be generalized by constructing an ROC curve with linear segments, as illustrated in Figure 3, and then calculating the area under it. The two most commonly used names for this measure are A_g and the R-index. This index is equivalent to the Mann–Whitney U statistic (Bamber, 1975). For Example 2, A_g is .75, which in this case is close to A_z at .76.

Alternative Distribution Assumptions

The NEV model is but one valid alternative. A natural generalization of this model is the normal–normal unequal-variance (NUEV) model. This model allows the variance of the S_2 distribution to differ from one (assuming a standardized S_1 distribution), or more generally, the ratio of the variances of S_1 to S_2 to vary from one. NUEV ROC curves can be asymmetrical about the minor diagonal of the ROC square, whereas NEV ROCs are symmetrical. Consequently, the NUEV model

can fit a broader variety of data. However, this model can be problematic.

ROC curves have two major properties. An ROC curve that touches the unit square only at the points (0,0) and (1,1) is said to be regular, and one with a monotonic-decreasing slope as F increases is said to be proper. The majority of empirical ROCs support the premise that ROC curves should be both regular and proper. NUEV ROC curves are regular, but are not proper. Despite this impediment, NUEV ROCs are the most prevalent type reported in the literature. A proper ROC curve can be derived from the NUEV model by transforming the decision variable so that it is monotonic with likelihood ratio (Egan, 1975).

It is not necessary to assume that the underlying distributions for ROC curves (or SDT in general) are Gaussian, as any distribution can replace the Gaussian. A common alternative is the logistic distribution, which affords much simpler mathematics than the Gaussian. In addition, the logistic ROC curve is very similar to the Gaussian ROC curve, so similar that it is difficult to choose which is more appropriate from experimentation. In addition, logistic distributions are strongly connected to choice theory (Luce, 1959) and have strong justification for use in the analysis of categorical data, such as collected in SDT research (Agresti, 2007).

Distributions may be selected based on the known variability in the physical stimulus. If this variability is sufficiently large it will swamp the internal sensory variability, meaning that the evidence upon which decisions are made primarily derives its variation from the physical stimulus. Examples of this include Rayleigh distributions as the basis for the discrimination of auditory Rayleigh noise; χ^2 distributions for the discrimination of auditory Gaussian noise; or Poisson distributions for the detection of light (Egan, 1975). The gamma distribution has received attention in ROC analysis in the medical sciences (Dorfman et al., 1997). Gamma ROCs are both proper and regular. In addition, the gamma distribution holds several special relationships to other distributions, including beta, exponential, and χ^2, all of which can be shown to be special cases of gamma.

Multiple-Interval Tasks

The single-interval task is but one of many paradigms. If S_1 and S_2 are presented consecutively on a trial, the judge could be asked to identify which was S_2, the first or the second stimulus. This is the two-alternative forced-choice (2AFC) task. This task poses the opportunity to consider how a judge maps the sensory evidence onto a response. For the single-interval task, the judge divides the sensory evidence into two 'sections' using the criterion and attaches a response to each section. This mapping of evidence to response is termed a decision strategy. In 2AFC, a judge can employ two distinct decision strategies. The judge could determine the difference between the sensory evidence from the two presentations, and then this difference could be compared to a criterion. If the difference is greater than the criterion, the judge responds 'second,' otherwise 'first.' If the criterion is zero (representing an equal weighting of both sources of evidence), this amounts to choosing the presentation with the greatest magnitude of sensory evidence. This is the decision strategy (DS). An alternative strategy is for the judge to estimate separately the likelihood that each presentation was either S_1 or S_2, and then to compare the ratio of these two likelihoods (second over first presentation) to a criterion. Assuming the criterion is set to one (representing equal evidence either way), if the ratio of likelihoods is greater than one, the judge responds 'second,' otherwise 'first.' This is called the independent-observations decision strategy (IOS). For 2AFC, responses based on any choice of criterion under the DS can be matched exactly by a distinct criterion under the IOS to yield the same performance. The ROC curves based on each decision strategy are identical. For this reason, attention to the decision strategy adopted by a judge is usually not required when determining performance in 2AFC.

Now consider another two-stimulus task. On any trial one of the following sequences of stimuli are presented to the judge: $<S_1, S_1>$, $<S_2, S_2>$, $<S_1, S_2>$, or $<S_2, S_1>$. The judge then indicates whether the two stimuli presented were the 'same' or 'different.' This is the standard same–different task. The judge can employ the DS or the IOS in this task as well. For the DS the judge determines the difference between the sensory evidence arising from each presented stimulus. If this difference is less than a criterion value, the judge responds 'same,' otherwise 'different.' For the IOS, the judge should respond 'same' if both stimuli presented on a trial were more likely to stem from S_1 than from S_2, or if both were more likely to stem from S_2 than from S_1; otherwise respond 'different.'

ROC curves can be generated for the same–different experiment by defining H and F as the probabilities of responding 'different' and 'same,' respectively, given that the stimuli were in fact different. The ROC curves for each decision strategy take on a different form. Figure 4 illustrates ROC curves for the case where $d' = 2$. For the DS, the curves are asymmetrical about the minor diagonal, whereas for the IOS the curves are symmetrical. This indicates that the decisions made under each strategy, given the same values of sensory evidence, are not always the same. Figure 4 also

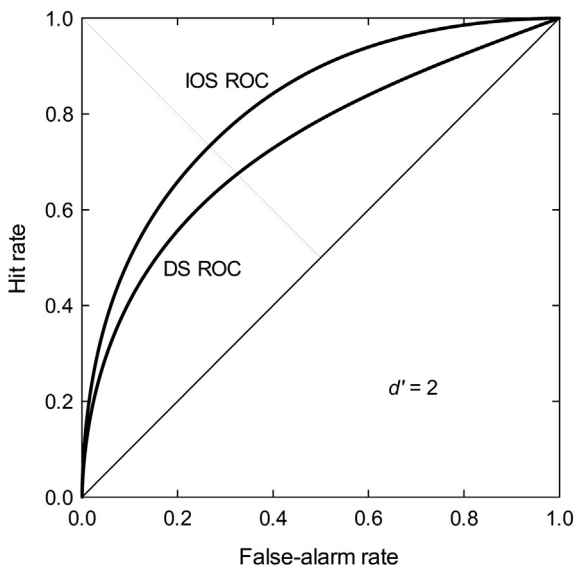

Figure 4 ROC curves ($d' = 2$) for two decision strategies in the same–different task.

shows that the value of H, at all values of F, is higher for the IOS ROC curve than that for the DS. Not only do judges sometimes give different responses under the two strategies, but when the responses do differ, the response from the IOS judgment is more often correct.

The IOS, based on the likelihood ratio, is the optimal decision strategy in a statistical sense, and of all possible strategies yield the highest proportion of correct responses (Green and Swets, 1966). In contrast, the DS is a suboptimal decision strategy. The difference in performance can be substantial; a 20% discrepancy is common. This has practical ramifications. It is not possible to estimate an accurate value of d' for the same–different experiment unless the decision strategy used by the judge is known. Visual research on the shape of the same–different ROC suggests that the DS is used almost exclusively for simple visual stimuli, whereas the IOS is used with more complex visual stimuli (Irwin and Francis, 1995). However, recent research suggests that this pattern may not apply to all modalities; specifically, the IOS may be used for simple gustatory stimuli.

Same–different is not the only task to yield different levels of performance when different decision strategies are used. Tasks belonging to the reminder paradigm also show different levels of performance. These include the single-interval and 2AFC reminder tasks – identical to their namesakes already discussed, except that S_1 is presented at the start of every trial. The ABX experiment, sometimes referred to as matching-to-sample, is another example.

The same–different task shows a difference in shape between the ROC curves derived from each decision strategy. Other tasks show only a difference in the area under the ROC curves for each strategy. This makes it more difficult to determine which strategy a judge uses for these tasks. The most common approach hinges on one of the central tenets of SDT; a valid measure of sensitivity should obtain the same estimate irrespective of the method used to obtain it. In the NEV SDT framework, d' is that measure. Estimates of d' can be obtained under the assumption of the various decision strategies available in a task and can then be compared to d' estimates obtained using a different task for which there is less uncertainty about the decision strategy adopted. Usual candidates for this comparison are the single-interval task, which has only one candidate decision strategy, or the 2AFC task, which has two candidate strategies, but with each theoretically producing the same outcomes (Macmillan and Creelman, 2005; Hautus et al., 2009). This approach may be confounded by factors not yet incorporated into the SDT framework that can differentially affect performance across psychophysical tasks. For example, memory effects (decay or interference) are likely to be larger in tasks with more stimulus presentations per trial. Memory effects on d' have been demonstrated in several studies that manipulated the interstimulus interval in the 2AFC task (Macmillan and Creelman, 2005).

Multiple Stimuli Varying along One Dimension

Consideration, thus far, has been given to discrimination between two stimuli (or two classes of stimuli). The extension of SDT to more than two stimuli is easy when those stimuli are located along a single physical dimension (e.g., sound intensity), or more generally, change along more than one physical dimension but only one perceptual dimension (e.g., loudness). This form of task, with n stimuli and m response options ($m \leq n$), is typically called a classification task. The representation of the decision space is analogous to that of Figure 1. The number of distributions equals n (in Figure 1, $n = 2$), the number of stimuli employed in the experiment, except there are now more than two stimuli. d' can be determined, for any pair of stimuli, as the distance between their means (assuming the NEV model).

Performance on classification tasks turns out to be worse than on two-stimulus discrimination tasks. An explanation is that performance is limited in discrimination tasks primarily by the resolving power of the sensory system, whereas classification tasks are also influenced by context coding, processes involved in assigning labels, which add judgmental variance to the sensory variance (Durlach and Braida, 1969). In an experiment that contrasted these two approaches for determining sensitivity to experimental pain, the classification task ($n = m = 6$) produced estimates of sensitivity over four times less than the discrimination task (Irwin et al., 1994). Similar studies in other modalities confirm the degraded performance due to additional noise associated with decision processes in classification tasks.

Multidimensional Detection Theory

The tasks previously described that involve two stimuli (or classes of stimuli) are appropriate for stimuli that are unidimensional, as well as for stimuli that can vary along more than one dimension (e.g., intensity and frequency). The tasks in the previous section that involve more than two stimuli are only applicable when those stimuli are located along the same single dimension (e.g., intensity). Multidimensional detection theory is the extension of SDT to the case of more than two stimuli, each of which can vary over more than one physical or perceptual dimension.

Multidimensional detection theory can address issues related to the dimensionality of sensory evidence and the degree of independence of criterion placement on those dimensions. Three forms of independence are of note: perceptual independence, perceptual separability, and decisional separability. Perceptual independence refers to the relationship between the physical dimensionality of a stimulus and the sensory evidence arising thereof. For example, if the intensity of a white light is increased gradually the perception of brightness, as expected, increases, but so too does the perceived hue; the light looks more yellow. Here, a single physical dimension gives rise to two dependent perceptual dimensions. When there is more than one physical dimension, perceptual separability holds if the level of one stimulus dimension has no effect on the sensory evidence arising from other dimensions. Stimuli that vary along the physical dimensions of electromagnetic wavelength (hue) and shape give rise to perceptually separable dimensions. Decisional separability holds if the judge's criterion for the sensory evidence relating to one stimulus dimension is not influenced by the sensory evidence arising from other dimensions (Ashby, 1992).

See also: Discrete-State Models of Information Processing; Psychophysical Theory and Laws, History of; Signal Detection Theory, History of.

Bibliography

Agresti, A., 2007. An Introduction to Categorical Analysis. Wiley, New Jersey.

Ashby, F.G. (Ed.), 1992. Multidimensional Models of Perception and Cognition. Erlbaum, New Jersey.

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12, 387–415.

Dorfman, D.D., Alf Jr., E., 1969. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating method data. Journal of Mathematical Psychology 6, 487–496.

Dorfman, D.D., Berbaum, K.S., Metz, C.E., Lenth, R.V., Hanley, J.A., Dagga, H.A., 1997. Proper receiver operating characteristic analysis: the bigamma model. Academic Radiology 4 (2), 138–149.

Durlach, N.I., Braida, L.D., 1969. Intensity perception: I. Preliminary theory of intensity resolution. Journal of the Acoustical Society of America 46, 372–383.

Egan, J.P., 1975. Signal Detection Theory and ROC Analysis. Academic Press, New York.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.

Hautus, M.J., van Hout, D., Lee, H.-S., 2009. Variants of A Not-A and 2AFC tests: signal detection theory models. Food Quality and Preference 20, 222–229.

Irwin, R.J., Francis, M.A., 1995. Perception of simple and complex visual stimuli: decision strategies and hemispheric differences in same-different judgments. Perception 24, 787–809.

Irwin, R.J., Hautus, M.J., Dawson, N.J., Welch, D., Bayly, M.F., 1994. Discriminability of electrocutaneous stimuli after topical anesthesia: detection-theory measurement of sensitivity to painful stimuli. Perception and Psychophysics 55 (2), 125–132.

Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, second ed. Erlbaum, New Jersey.

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C [or C++], second ed. Cambridge University Press, Cambridge.

Relevant Websites

http://bit.ly/xsRwQg – Online ROC calculator at Johns Hopkins.
http://bit.ly/xL7WI2 – ROC Analysis software by Metz.
http://bit.ly/ApyHPy – ROC Analysis software by Pepe Lab.
http://bit.ly/wkohmE – SDT explained with animation at anaesthetist.com.

Signal Detection Theory, History of

Carleton Douglas Creelman, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.A. Swets, volume 21, pp. 14078–14082, © 2001 Elsevier Ltd.

Abstract

Detection Theory is traced from its origins in engineering and statistical decision theory through to its adaptation in experimental psychology, medical diagnosis, and elsewhere. The original displacement of threshold notions is summarized, and early experiments establishing the theory, as well as some adaptations and extensions are described.

Introduction and Background

Signal Detection Theory, as the phrase is used in psychology, medical science, and in many other fields, refers to ways data are analyzed, where detection or identification of an input signal is required. These techniques have roots in the technology of identifying threat from enemy aircraft using RADAR during the World War II (WWII). The path from there to current applications is the task of this article.

The story, as it unfolds in the field of sensory psychology, involves in the first instance, problems of measuring the sensitivity of human visual and auditory systems. The story also involves quantifying the basis of the decisions that are made in conditions with uncertain or noisy input information. The application of Detection Theory involved a genuine revolution in the ways data in many fields are collected and analyzed. Signal Detection Theory has provided new ways to understand how sensory systems process their inputs.

An operator of an early RADAR console attempting to identify an aircraft, watching a display on a cathode ray tube, was faced with a prototypical detection problem. The display was covered with 'snow,' reflections from atmospheric features, and the results of random activity in the vacuum tube circuits processing signals passed from an antenna. This was the visual equivalent of the noise from an FM radio when not tuned to a strong station. Buried in this visual noise might have been a reflection from an aircraft. Noise might mock the signature of an aircraft, and noise might mask an important signal.

To overcome these problems, a first approach might be to increase the strength of the overall input to the RADAR console. However, amplification boosted noise to the same degree as relevant signals. A more sophisticated approach arose, primarily, in the United States, at the Massachusetts Institute of Technology (MIT).

Subsequent to the end of WWII, and the end of tight security around theoretical work in the field, academically connected laboratories, such as those at MIT and the University of Michigan, published work describing ways to analyze faint, noise-contaminated signals. From the Research Laboratory of Electronics at MIT, van Meter and Middleton (1954) published their analysis of the problem of extracting a signal that is imbedded in noise. Davenport and Root (1958) provided a comprehensive text on the engineering issues surrounding signal extraction. At the University of Michigan, Peterson et al. (1954) addressed the same issues, arriving at some of the same theoretical findings, and built on them by adding consideration of the decision processes required. The technique involved two core ideas: detection should involve correlation of a known signal with the noise-masked input and the input signal could be fully characterized by sampling at a rate determined by the bandwidth of the noise and the duration of the observation.

A second part of the detection problem, when the input might or might not contain a signal, was to define the optimal basis for decision. Here, the authors showed that the *likelihood ratio* was the optimal basis for decision, the ratio of the likelihood that the input came from a signal in the noise to that of the noise alone, with no signal present. Such a signal analysis and subsequent decision was characterized by Peterson and Birdsall an *ideal observer*. The notion of an ideal observer looms large later in approaches to human perception and neurological processing.

These notions found their way into psychological laboratories and from there into a much wider community, starting with the Vision Research Laboratories at the University of Michigan.

Vision Research at Michigan

In the early 1950s, two graduate students were working in the Vision Research Laboratory, directed by H. Richard Blackwell. W. P. Tanner, Jr. (known as 'Spike') had spent WWII with the US Navy, working in an electronics facility. After the war, he obtained a Masters degree at the University of Florida and then went on to the PhD program at Michigan. John A. Swets had done his undergraduate degree at Michigan. Tanner, with a young family, was working on the side at the Vacuum Tube Laboratory in the Engineering School. He came into contact with graduate students there, Wes Peterson, an engineering student, and Ted Birdsall, a mathematician. Together they were developing their Theory of Signal Detectability; Tanner saw ways to apply some of their ideas in the experiments he was conducting in the Psychology Department, and Swets quickly saw the importance of this work.

The Vision Research Laboratory at Michigan was equipped with excellent stimulus-generating equipment, with full optical benches devoted to presenting simple visual stimuli focused on backlit screens. Control of experiments was mechanical/electrical; exact timing of stimulus presentation was electronic, and recording observers' responses was on mechanical

counters. Experiments were conducted in small light-proof theaters seating only two paid viewers who reported their responses using pushbuttons. This arrangement, fine stimulus control, with precise timing and response registration using state-of-the-art technology and with observer isolation formed the basic arrangement for psychophysical experimentation elsewhere.

This setting was put together by Blackwell with help from technicians, and with statisticians to formalize ways to analyze data (Blackwell et al., 1954). The aim was to understand how the visual system worked; how it could be studied and thresholds measured, in a fundamental sense.

In the 1950s, psychological theory and experiment were under the thrall of Behaviorism (Watson, 1925). Psychological theorists and experimenters took from Watson the notion that, since internal, cognitive experiential events could not be observed directly, they had no place in the field. Observed behavior was then all that was left for systematic, rigorous scientific study. Subjects in sensory experiments, for instance, could say 'yes' or 'no,' or the equivalent with a button press, and the processes that lead to that response were deemed outside the purview of the theorist.

However since the time of Fechner, the recognized founder of psychophysics (Fechner, 1860), it had been known that seeing a faint stimulus or hearing a faint sound was probabilistic; the 'threshold' was a statistical fiction, an arbitrary point on the continuous increasing function relating stimulus intensity to the proportion of 'yes' responses. That function increased from zero with no stimulus to 100% with a sufficiently large one. Usually, Osgood (1958) in his definitive textbook on experimental psychology captured the issue in Operational terms: "Thresholds only appear in the final statistical treatment of the data." And he concluded, "It is clear that the various psychophysical methods are not merely different ways of measuring the same thing."

Sometimes, however, observers responded 'yes' to an empty stimulus. Laboratory manuals suggested presenting the occasional blank stimulus to be sure that subjects were paying attention. A 'yes' response in this case was treated seriously; it should be treated as 'a lack of rigorous attention' that the observer should be 'waked up' (Titchener, 1901, pp. 103). This all-or-none view of sensory experience, that on any occasion sensory input triggered a sensation or not, the concept, at least on a trial-by-trial basis that an observer experienced a presented stimulus, or not, was at the heart of most approaches to psychophysical measurement.

Blackwell, in his laboratory, approached these problems with two important innovations. First, since observers were known sometimes to respond to blank presentations, called then, 'catch-trials,' they should be accepted as part of the data. Rather than presenting a catch-trial only rarely, as recommended by Titchener, catch-trials were made a regular occurrence, presented on as many as 50% of trials on a random basis. The assumption underlying the analysis of the data was that the observers simply or mistakenly sometimes said 'yes' on a catch-trial because of a mix-up in finger-to-button alignment, or desire to please the experimenter, or some other reason. Thus, the obtained percentage of 'yes' responses, made up of those supposedly actually experienced mixed with the other 'spurious' ones, should be 'corrected' to remove the improper ones. This was because there were a sufficient number of catch-trials to constitute a reliable estimate of how often the observer responded 'yes' when the stimulus had not been seen. The true psychometric function, relating percentage of seen signals to intensity would then go from zero to 100%, rather than, as the uncorrected data showed, starting at the 'false-alarm' rate of response to empty trials at zero stimulus intensity.

The second innovation in Blackwell's laboratory was to eliminate the problem of false alarms entirely, presenting to observers on each trial four successive time intervals, indicated by auditory markers. Only one of the intervals, randomly chosen, contained a visual signal. The observer was then required to indicate a best guess as to which of the four contained the stimulus. This four-alternative forced choice procedure, if the observer saw nothing, would yield a theoretical minimum 25% correct responses, the expected random performance for inability to see anything. So this was used as the 'correction' to make a psychometric function go from zero to 100%, instead of the obtained lower limit of 25% at zero stimulus intensity.

Another important predecessor of Detection Theory, Thurstone (1927) had suggested that the effects of stimuli to be compared were statistically distributed. The idea added by Detection Theory, which Thurstone and his followers missed, was that the dividing line between one response and the alternative could vary, chosen depending on aims of the subject in an experiment. Thurstone's model had the line between choice of one alternative and the other situated at the crossover point of the alternative statistical distributions, and not adjustable according to the situation and objectives of the observer.

The Original Experiments

Swets proceeded to apply notions from the theoretical structure of Peterson and Birdsall, in Blackwell's laboratory. Ongoing discussions with Tanner enriched the work and ultimately led to his PhD thesis being published (Tanner et al., 1956). Tanner's work in the laboratory, on detection of visual signals that were presented randomly over a spatial area, was never published; Tanner left the Vision Research Laboratory and moved to another Michigan laboratory, the Electronic Defense Group. There he established a hearing research laboratory with an active program of research on auditory function, publishing and supervising graduate students and postdoctoral fellows even though he himself did not yet have a doctoral degree.

Tanner and Swets showed that decisions in visual detection tasks appeared to be on the basis of statistically distributed inputs, where a signal added to a background simply moved the mean of an underlying distribution. Furthermore, the data showed that observers were able to evaluate sensory inputs on a continuum, a dimension which could be considered to correspond to Detectability Theory's likelihood ratio. These data questioned the established threshold concept.

Swets' data showed that observers could be made to establish different levels of a 'criterion' for saying 'yes' to a faint visual signal. Criterion could be manipulated either by establishing different pay amounts for a hit (correct identification of the presence of a signal) and fines for a false alarm (incorrect report of a signal when none was present), or by varying the

chance that a signal would be present on a trial. So, for the same stimulus intensity, observers varied their hit rate along with the false-alarm rate, depending on payoffs or signal likelihood. This behavior was consistent with the Peterson and Birdsall predictions, and with the measure of performance d′ (dee-prime), which was found to be consistent across different criterion settings, representing the distance between the means of assumed underlying distributions.

As an aside, Tanner and Swets adopted Peterson and Birdsall's measure *D*, but as psychologists with training in statistics, they preferred the distance measure, which was the square root of *D*. However, 'square root of *D*' was found to be awkward in conversation and in writing, so they shortened the phrase to d′ and the alternative designation stuck.

A further indication of the notion of a continuum of underlying likelihood in psychophysical tasks was Swets' demonstration that observers could maintain several criteria simultaneously, rating each presentation of a stimulus (or nonstimulus catch-trial) on a numerical scale. If experience were all-or-none, such a task would be impossible. The data from the rating experiment corresponded to the receiver operating characteristic (ROC) presented by Peterson and Birdsall. Rating data could be analyzed to yield pairs of hit and false-alarm rates corresponding to each separate criterion. The theoretical relationship, graphically, was a smooth curve from (0,0) to (1.0,1.0) when hit and false-alarm percentages were plotted against each other. An observer in this task who saw stimuli on an all-or-none basis and sometimes guessed, would yield a straight-line ROC, intercepting the vertical false-alarm axis at the 'guessing' rate; the data were satisfyingly on the side of a Detection Theory curvilinear ROC.

A further demonstration of the continuous nature of a sensory input was from the 'second-choice' experiment. These data were summarized in an influential science article by Swets (1961), from the earlier technical report (Swets et al., 1955). In this experiment, a faint spot of light was flashed on an illuminated background during one of four time intervals. Observers were to indicate their choice and then to indicate which of the remaining three intervals might have contained the signal if they might have been wrong in their initial choice. If a threshold all-or-none mechanism were at play, then on those trials when the first choice was incorrect no information would be available, and the second choice could only have been correct by chance, 33.3%; while if a continuous process underlay the decision, then the second choice should have been right more often than chance, and the theory predicted the percentage of correct second guesses, increasing with flash intensity. The data strongly supported the idea of a continuum of information, as opposed to an all-or-none threshold process.

Propagation and Adaptation

Swets, with a new PhD degree, went to Cambridge, to MIT, where J.C.R. Licklider had become an enthusiastic proponent of the new Detection Theory. Swets has credited Licklider for pushing the editor of the Psychological Review to publish the combined theoretical and empirical Tanner and Swets article. While teaching at MIT, Swets joined the acoustical consulting and research firm Bolt, Beranek and Newman on a part-time basis, a connection which eventually led to full-time involvement and ultimately to the title of Chief Scientist.

Tanner's laboratory was joined by graduate student David M. Green, who as an undergraduate had participated as an observer in the visual experiments. This collaboration led to support for the Detection Theory approach in the auditory detection realm, with an influential technical report by Tanner et al. (1956). It is said that at MIT people around Licklider, hearing him talk so often of this work, thought the author to be a single person, named Tanner Swetsngreen. Green, having completed his PhD, joined Swets at MIT. He subsequently taught at the University of Pennsylvania and at the University of Florida.

Other laboratories followed, with focus on application to other areas of human perception. For instance, James P. Egan, at Indiana University began studies of speech processing (e.g., Egan et al., 1956), which focused on Detection Theory analysis, and Egan published a monograph on uses of ROC analysis (Egan, 1975).

Meanwhile researchers in other psychological research areas saw that the concept of a continuum of evidence was an extraordinarily useful framework. One important area was in the study of verbal memory. For instance the idea of verbal recognition as a problem of extracting signal from noise, basically as a correlation process, was proposed by Murdock (1982), and many researchers came to understand the meaning of false alarms in recognition tasks, using ROC analysis. An excellent summary of much of this work was provided by Benjamin et al. (2009). Creelman (1962) proposed a model for human time discrimination, attempting to identify the nature of the internal timekeeper for comparison of short durations.

At the same time, the medical world was introduced to Detection Theory by Metz (1978). In collaboration with Swets, Metz introduced concepts from Detection Theory, pointing out the necessity to consider false alarms as well as hits in diagnostics. Metz introduced the terms currently used in medicine, 'sensitivity,' the proportion of correct identification of disease processes (hit rate) and 'specificity,' the proportion of correct decisions that disease was not present (1.0 – false-alarm rate). These statistically independent measures have now become standard in the medical literature. Metz introduced the ROC curve to his medical colleagues and explained the equivalence of detectability for all pairs of hits and false alarms along the same ROC. The notion that for the same capability to detect, for instance a breast cancer, the diagnostician could select one criterion level or another and thus achieve more hits, but at the cost of more false alarms, was and remains an important consideration for an X-ray reader. Detection Theory requires sensitivity on the part of the diagnostician to the likelihood of disease and to the costs and benefits of treatment (or of failure to treat) in the case of a positive finding. These are the same issues that Swets explored in his original laboratory studies of masked visual stimuli, and which Green, Birdsall, and Tanner demonstrated in the auditory realm.

The Ideal Observer

In building electronic detection systems, the specification of an ideal level of performance was important, limiting efforts

to achieve better-than-ideal results. One aspect of the work of Peterson and Birdsall was to specify, for any combination of signal and masking noise, a maximum possible detection performance, an ideal observer. The combination of optimal signal-processing procedure and decision strategy established a ceiling on possible performance. Tanner and Birdsall (1958) defined *efficiency* of a receiver (or a human listener) as the ratio of the obtained performance to that of an ideal. Tanner then showed (Tanner, 1961) that various ways of specifying the signal to observers, for instance marking the observation precisely, could improve efficiency. Much later Taylor and Forbes (1969) showed that with sufficient information, presented binaurally, near-ideal detection was possible for their human observers. In study of human detection and animal physiological responses, the ideal approach proved useful (Cohn et al., 1975). This work followed upon the much earlier studies by Kuffler et al. (1956), showing that neural rates of firing in response to light stimuli were not fixed, but random.

Multidimensional Extensions

Tanner introduced the concept that detection could be measured for differences in the magnitude of sensory experience, showing that detection of a difference could be considered simply as a detection problem, where the 'signal' to be detected was the difference between two possible sensory inputs. This work was summarized in the PhD thesis that Tanner finally presented (Tanner, 1960). Later Creelman (1963, 1965) showed that Detection Theory measures along unidimensional sensory continua such as loudness and lengths of lines could be shown to be additive. Tanner (1956) extended the analysis, showing that discrimination between auditory signals of different frequencies could be modeled successfully in two dimensions, where detectability of each signal was plotted along a separate axis, and the listener's ability to distinguish between them was measured as the angle between the axes. These ideas have been greatly extended to analysis of multidimensional stimulus sets in the work of Ashby and Townsend (Ashby, 1992; Ashby and Gott, 1988).

Spreading the Word: Books and Conferences

The definitive presentation of Signal Detection Theory, as a theoretical framework and a way to conduct psychophysical experiments, was published by Green and Swets (1966). A more recent overview of theory and experiment is by Macmillan and Creelman (2005). Swets and Picket (1982) summarized the broader applications of Detection Theory in areas as diverse as medical diagnostics and library item retrieval. Swets (1964, Revised Ed. 1988) reprinted a collection of influential articles showing the extent to the time of applications. Elementary introductions have been published by McNicol (1972) and Wickens (2002). Lively discussion among psychologists working in audition over the years has been at meetings of the Acoustical Society of America.

Some Detractors

Attacks on the conceptualizations that underlie Detection Theory, as it is applied to psychophysics, came from several directions. One line of attack was simply to dismiss it as nothing new, as using the 'time worn concepts of probability theory' to complicate the relatively simple field of psychophysics (Stevens, 1964). A more sophisticated approach has been that of Balakrishnan (1999), and he has recently offered an alternative approach to measure sensory discrimination (Balakrishnan and MacDonald, 2011). Krantz (1969) showed that published ROC curve data might be accounted for with a sophisticated multiple-threshold theory. Another line of discussion has involved attempts to free detection measures from the assumptions underlying calculation of d′, particularly the reliance on the assumption regarding the shape of the underlying distributions, the Gaussian normal distribution. These 'nonparametric' approaches (Pollack and Norman, 1964; Donaldson, 1992) were shown by Macmillan and Creelman (1996) in fact to imply distributional underpinnings consistent with the version of the theory developed by Luce (1959), which used the logistic distribution instead of the Gaussian.

Importance for Psychology

Signal Detection Theory, along with Information Theory (Miller, 1956), was an early harbinger of the cognitive revolution in psychology (Neisser, 1967). With its focus on the processes involved in decision making and signal analysis, processes presumed to be carried out by human and animal perceivers, the approach showed that a model specifying internal, assumed complex cognitive functioning could be explored experimentally.

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Mathematical Psychology; Memory Models: Quantitative; Signal Detection Theory.

Bibliography

Ashby, F.G. (Ed.), 1992. Multidimensional Models of Perception and Cognition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Ashby, F.G., Gott, R.E., 1988. Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition 14, 33–53.

Balakrishnan, J.D., 1999. Decision processes in discrimination: fundamental misrepresentations of signal detection theory. Journal of Experimental Psychology: Human Perception and Performance 25, 1189.

Balakrishnan, J.D., MacDonald, J.A., 2011. Performance measures for dynamic signal detection. Journal of Mathematical Psychology 55 (4), 290–301.

Benjamin, A.S., Diaz, M., Wee, S., 2009. Signal detection with criterion noise: applications to recognition memory. Psychological Review 116 (1), 84–115.

Blackwell, H.R., Pritchard, B., Ohmart, J., 1954. Automatic apparatus for stimulus presentation and recording in visual threshold experiments. Journal of the Optical Society of America 44 (4), 322–326.

Cohn, T.E., Green, D.G., Tanner Jr., W.P., 1975. Receiver operating characteristic analysis: application to the study of quantum fluctuation effects in optic nerve of *Rana pipiens*. Journal of General Physiology 66 (5), 583–616.

Creelman, C.D., 1962. Human discrimination of auditory duration. Journal of the Acoustical Society of America 34 (5), 582–593.

Creelman, C.D., 1963. Detection, discrimination, and the loudness of short tones. Journal of the Acoustical Society of America 35, 1201–1205.
Creelman, C.D., 1965. Discriminability and scaling of linear extent. Journal of Experimental Psychology 70, 192–200.
Davenport, W., Root, W., 1958. An Introduction to the Theory of Random Signals and Noise. McGraw-Hill, New York.
Donaldson, W., 1992. Measuring recognition memory. Journal of Experimental Psychology: General 121, 275–277.
Egan, J.P., 1975. Signal Detection Theory and ROC Analysis. Academic Press, New York.
Egan, J.P., Clarke, F.R., Carterette, E.C., 1956. On the transmission and confirmation of messages in noise. Journal of the Acoustical Society of America 28, 536–550.
Fechner, G.T., 1860. Elemente der Psychophysik (Reissued 1964 by Bonset, Amsterdam ed.). Leipzig: Breitkopf & Hartel.
Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.
Krantz, D.H., 1969. Threshold theories of signal detection. Psychological Review 76 (3), 308–324.
Kuffler, E.W., FitzHugh, R., Barlow, H.B., 1956. Maintained activity in the light and dark-adapted cat's retina. Journal of General Physiology 40, 683–702.
Luce, R.D., 1959. Individual Choice Behavior. Wiley, New York.
Macmillan, N.A., Creelman, C.D., 1996. Triangles in ROC space: history and theory of "nonparametric" measures of sensitivity and response bias. Psychonomic Bulletin & Review 3, 164–170.
Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, second ed. Erlbaum, Mahwah, NJ.
McNicol, D., 1972. A Primer of Signal Detection Theory. Allen & Unwin, London.
Metz, C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8 (4), 283–298.
Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63, 81–96.
Murdock Jr., B.B., 1982. A theory for the storage and retrieval of item and associative information. Psychological Review 89, 609–626.
Neisser, U., 1967. Cognitive Psychology. Appleton-Century-Crofts, New York.
Osgood, C.E., 1958. Method and Theory in Experimental Psychology. Oxford University Press, Oxford.
Peterson, W.W., Birdsall, T.G., Fox, G.G., 1954. The theory of signal detectability. Transactions of the IRE Professional Group on Information Theory PGIT-4, 171–212.
Pollack, I., Norman, D.A., 1964. A nonparametric analysis of recognition experiments. Psychonomic Science 1, 125–126.
Stevens, S.S., 1964. Sensory process. In: Rosenblith, W. (Ed.), Sensory Communication. John Wiley & Sons, New York.
Swets, J.A., 1961. Is there a sensory threshold? Science 134, 168–177s.
Swets, J.A. (Ed.), 1964, Revised Ed. 1988. Signal Detection and Recognition by Human Observers. Wiley, New York.
Swets, J.A., Pickett, R.M., 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.
Swets, J.A., Tanner, W.P., Birdsall, T.G., 1955. The Evidence for a Decision-Making Theory of Visual Detection (Technical Report No. 40). University of Michigan, Ann Arbor, MI.
Tanner Jr., W.P., 1956. Theory of recognition. Journal of the Acoustic Society of America 28, 883–888.
Tanner Jr., W.P., 1960. Application of the Theory of Signal Detectability to Amplitude Discrimination. (PhD, University of Michigan). Dissertation Abstracts International, 21-02, p. 364.
Tanner Jr., W.P., 1961. Physiological implications of psychophysical data. Annals of the New York Academy of Science 89, 752–765.
Tanner Jr., W.P., Birdsall, T.G., 1958. Definitions of d' and η as psychophysical measures. Journal of the Acoustical Society of America 30 (10), 922–928.
Tanner Jr., W.P., Swets, J.A., Green, D.M., 1956. Some General Properties of the Hearing Mechanism (Technical Report 36 ed.). University of Michigan: (Electronic Defense Group), Ann Arbor, MI.
Taylor, M.M., Forbes, S.M., 1969. Monaural detection with contralateral cue (MDCC). I. Better than energy detector performance by human observers. The Journal of the Acoustical Society of America 46 (6), 1519–1526.
Thurstone, L.L., 1927. A law of comparative judgment. Psychological Review 34, 273–286.
Titchener, E.B., 1901. Experimental Psychology: A Manual of Laboratory Practice. Macmillan, New York. 1901–1912.
van Meter, D., Middleton, D., 1954. Modern statistical approaches to reception in communication theory. Transactions of the IRE Professional Group on Information Theory PGIT-4, 119–141.
Watson, J.B., 1925. Behaviorism. W. W. Norton, New York.
Wickens, T.D., 2002. Elementary Signal Detection Theory. Oxford University Press, New York.

Significance, Tests of

Nancy Reid, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Tests of significance are used in statistics to assess the agreement between data and models. This article describes how such tests are formulated and used, illustrates this with several examples, and discusses some difficulties in interpretation that have been raised.

A test of significance assesses the agreement between the data and a hypothesized statistical model. The magnitude of the agreement is expressed as an observed level of significance, or *p*-value, which is the probability of obtaining data as or more extreme than the observed data, if the hypothesized model were true. A very small *p*-value suggests that either the observed data are not compatible with the hypothesized model or an event of very small probability has been observed. A large *p*-value indicates that the observed data are compatible with the hypothesized model. As the *p*-value is a probability, it must be between 0 and 1, and a very common convention is to declare values smaller than 0.05 as 'small' or 'statistically significant' and values larger than 0.05 as 'not statistically significant.' The historical rationale for this very arbitrary cut-off point is that the calculation of a *p*-value was difficult, and tables useful for common statistical models were prepared for general use.

To make this more concrete, it is necessary to consider statistical models and the notion of a (simplifying) hypothesis within that model. The theory of this is outlined in Section Model-Based Inference, with some more specific points considered in Section Further Topics. Section Difficulties with Significance Tests provides some brief comments on criticisms of significance tests. This introduction concludes with a highly idealized example to convey the idea of data being inconsistent with a hypothesized model.

Example 1. Students in a statistics class partake in an activity to assess their ability to distinguish between two competing brands of cola, and to identify from taste alone their preferred brand. Each of the 20 students expresses a preference for one brand or the other, but just one student claims to be able to discriminate perfectly between the two. Twenty cups of each brand are prepared by the instructor and labeled '1' and '2.' Each student is presented with a pair of cups and asked to record which label corresponds to Brand A. The result is that 12 students correctly identify the competing brands, although the student who claimed a perfect ability to discriminate identified the brands incorrectly.

The labeling of the cups as 1 or 2 by the instructor was completely random, i.e., cup 1 was equally likely to contain Brand *A* or *B*. The students did not discuss their opinions with their classmates, and the taste testing was completed fairly quickly. Under these conditions, it is plausible that each student has a probability of $\frac{1}{2}$ of identifying the brands correctly simply by guessing, so that about 10 students would correctly identify the brands with no discriminatory ability at all. That 12 students did does not seem inconsistent with guess work, and the *p*-value helps to quantify this. The probability of observing 12 or more correct results if one correct result has probability $\frac{1}{2}$ and the guesses are independent can be computed by the binomial formula as $\left\{\binom{20}{12}+\binom{20}{13}+\cdots+\binom{20}{20}\right\}\left(\frac{1}{2}\right)^{20} = 0.34$: there is no evidence from these data that the number of correct answers could not have been obtained by guessing: in more statistical language assuming a binomial model, the observed data is consistent with probability of success $\frac{1}{2}$.

The student who claimed to have perfect discrimination, but actually guessed incorrectly, argued that her abilities should not be dismissed on the basis of one mistake, so the class carried out some computations to see what the *p*-value for the same observed data would be if the number of pairs of cups was increased. The probability of one or zero mistakes in a set of *n* trials for various values of *n*, is given in Table 1. From this we see that, for example, one or no mistakes in five trials is consistent with guess work but the same result in 10 trials is much less so.

In both parts of this example we assumed a model of independent trials, each of which could result in a success or failure, with constant probability of success. Our calculations also assumed this constant probability of success was 0.5. This latter restriction on the model is often called a 'null hypothesis' and the test of significance is a test of this null hypothesis; the *p*-value measures the consistency of the data with this null hypothesis. In many applications the null hypothesis plays the role of a conservative position that the experimenter hopes to disprove, and one reason for requiring rather small *p*-values

Table 1 Probability of zero or one mistakes in *n* independent Bernoulli trials with probability of a mistake = 0.5

n	*Probability*
5	0.1875
6	0.1094
7	0.0625
8	0.0352
9	0.0195
10	0.0107
11	0.0059
12	0.0032
13	0.0017
14	0.0009
15	0.0002

before declaring statistical significance is to raise the standard of proof required to replace a relatively simple working hypothesis by one that is possibly more complex and less well understood.

As formulated here the hypothesis being tested is that the probability of a correct choice is 0.5, and not the other aspects of the model, such as independence of the trials, and unchanging probability of success. The number of observed successes does not measure such model features, it provides information only on the probability of success. Functions of the data that do measure such model features can be constructed, and from these significance tests that assess the fitness of an assumed model; these play an important role in statistical inference as well.

Model-Based Inference

Models and Null Hypothesis

We assume that we have a statistical model for a random variable Y taking values in a sample space $\mathcal{Y}$, described by a parametric family of densities $\{f(y;\theta);\theta\in\Theta\}$. Tests of significance can in fact be constructed in more general settings but this framework is useful for defining the main ideas. If Y is the total number of successes in n independent Bernoulli trials with constant probability of success, then

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \qquad [1]$$

$\Theta = [0,1]$, and $\mathcal{Y} = \{0,1,\ldots,n\}$. If Y is a continuous random variable following a normal or bell curve distribution with mean θ_1 and variance θ_2^2, then

$$f(y;\theta) = \frac{1}{\sqrt{2\pi}\theta_2}\exp\left\{-\frac{1}{2\theta_2^2}(y-\theta_1)^2\right\} \qquad [2]$$

$\Theta = \mathbb{R}\times\mathbb{R}^+$ (R - real line; R - positive real line), and $\mathcal{Y} = \mathbb{R}$. The model for n independent observations from this distribution is

$$f(y_1,\ldots,y_n;\theta) = \frac{1}{\left(\sqrt{2\pi}\right)^n\theta_2^n}\exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^n(y_i-\theta_1)^2\right\} \qquad [3]$$

$\Theta = \mathbb{R}\times\mathbb{R}^+$, and $\mathcal{Y} = \mathbb{R}^n$. For further discussion of statistical models, *see* Statistical Sufficiency; Distributions, Statistical: Special and Discrete; Distributions, Statistical: Approximations; Statistical: Special and Continuous.

As noted above, we assume the model is given, and our interest is in inference about the parameter θ. While this could take various forms, a test of significance starts with a so-called *null hypothesis* about θ, of the form

$$H_0 : \theta = \theta_0 \qquad [4]$$

or

$$H_0 : \theta\in\Theta_0 \qquad [5]$$

In [4] the parameter θ is fully specified, and H_0 is called a *point* null hypothesis or a *simple* null hypothesis. If θ is not fully specified, as in [5], H_0 is called a *composite* null hypothesis. In the taste-testing examples the simple null hypothesis was $\theta = 0.5$. In the normal model, [2], a hypothesis about the mean, such as $H_0 : \theta_1 = 0$, is composite, as the variance is left unspecified. Another composite null hypothesis is $H_0 : \theta_2 = \theta_1$, which restricts the full parameter space to a one-dimensional curve in $\mathbb{R}\times\mathbb{R}^+$.

A test is constructed by choosing a *test statistic* which is a function of the data that in some natural way measures departure from what is expected under the null hypothesis, and which has been standardized so that its distribution is known either exactly or to a good approximation under the null hypothesis. Test statistics are usually constructed so that large values indicate a discrepancy from the hypothesis.

Example 2. In the binomial model [1], the distribution of Y is completely specified by the null hypothesis $\theta = 0.5$ as

$$f(y) = \binom{n}{y}2^{-n}$$

and consistency of a given observed value y_0 of y, is measured by the p-value $\sum_{y=y_0}^n\binom{n}{y}2^{-n}$, the probability of observing a value as or more extreme than y_0. If y_0 is quite a bit smaller than expected, then it would be more usual to compute the p-value as $\sum_{y=0}^{y_0}\binom{n}{y}2^{-n}$. Each of these calculations was carried out in the discussion of taste testing in the Introduction.

Example 3. In independent sampling from the normal distribution, given at [3], we usually test the composite null hypothesis $H_0 : \theta_1 = \theta_{10}$, by constructing the *t-statistic*

$$T = \sqrt{n}\left(\overline{Y}-\theta_{10}\right)/S$$

where $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $S^2 = (n-1)^{-1}\sum(Y_i-\overline{Y})^2$. Under H_0, T follows a t-distribution on $n-1$ degrees of freedom, and if large values of T are considered evidence against H_0, the p-value is

$$\Pr\left\{T\geq\sqrt{n}\left(\overline{y}-\theta_{10}\right)/s\right\}$$

where $\overline{y}$ and s are the values observed in the sample. (If small values of T are considered evidence against H_0, we would use the other tail of the distribution; see Section Two-Sided Testing for comments on two-sided tests of significance.) This probability needs to be computed numerically from an expression for the cumulative distribution function of the t-distribution. Historically tables of this distribution were provided for ready reference, typically by identifying a few critical values, such as $t_{0.10}$, $t_{0.05}$, and $t_{0.01}$ satisfying $\Pr\{T_\nu\geq t_\alpha\} = \alpha$, where T_ν is a random variable following a t-distribution on ν degrees of freedom. It was arguably the publication of these tables that led to a focus on the use of particular fixed levels for testing in applied work.

Example 4. Assume the model specifies that $Y_1,\ldots,Y_n$ are independent, identically distributed from a distribution with density $f(\cdot)$ on $\mathbb{R}$ and that we are interested in testing whether or not $f(\cdot)$ is a normal density:

$$H_0 : f(y) = \left(\sqrt{2\pi}\right)^{-1}e^{-\frac{1}{2}y^2} \qquad [6]$$

or

$$H_0 : f(y) = \left(\sqrt{2\pi}\right)^{-1}\theta_2^{-1}\exp\left\{-(y-\theta_1)^2/\left(2\theta_2^2\right)\right\} \qquad [7]$$

the former is a simple and the latter is a composite null hypothesis. For this problem it is less obvious how to construct

a test statistic or how to choose among alternative test statistics. Under [6] we know the distribution of each observation (standard normal) and thus of any function of the observations. The ordered values of Y could be compared to their expected values under [6], for example by plotting one against the other, and deviation of this plot from a line with intercept 0 and slope 1 could be measured in various ways. In the case of the composite null hypothesis [7], we could make use of the result that under H_0, $(Y_i - \overline{Y})/S$ has a distribution free of θ_1 and θ_2, and the vector of these residuals is independent of the pair $(\overline{Y}, S^2)$, and then for example compare the sample skewness $n^{-1}\sum_{i=1}^{n}\{(Y_i - \overline{Y})/S\}^3$ with that expected under normality.

Example 5. Suppose we have a sample of independent observations $Y_1,\ldots,Y_n$ on a circle of radius 1 and our null hypothesis is that the observations are uniformly distributed on the circle. One choice of a test statistic is $T_1 = \sum_{i=1}^{n}\cos(Y_i)$, very large positive (or negative) values indicating a concentration of observations at angle 0 (or π). If we instead wish to detect clumps of observations at two angles differing by π then $T_2 = \sum_{i=1}^{n}\{\cos(2Y_i) - 1\}$ would be more appropriate. The exact distribution of T_1 under H_0 is not available in closed form, but the mean and variance are readily computed as 0 and n, so a normal approximation might be used to compute the p-value.

In the examples described above the test statistics are ad hoc choices likely to be large if the null hypothesis is not true; these are called pure tests of significance, and are treated in detail in Cox and Hinkley (1974: Chapter 3); Examples 4 and 5 above are drawn from that chapter. A more sensitive test can be constructed if we have more specific knowledge of the likely form of departures from the null hypothesis. The theory of hypothesis testing formalizes this by setting up a null hypothesis and alternative hypothesis, and seeking to construct an optimal test for discriminating between them (*see* Hypothesis Testing, in Statistics). In the remainder of this section we consider an approach based on the likelihood function.

Significance Tests Based on Likelihood

In parametric models tests of significance are often constructed by using the likelihood function, and the p-value is computed by using an established approximation to the distribution of the test statistic. The likelihood function is proportional to the joint density of the data:

$$L(\theta; y) = c(y)f(y; \theta) \qquad [8]$$

see Likelihood, Methods of Statistical Inference.

We first suppose that we are testing the simple null hypothesis $H_0 : \theta = \theta_0$ in the parametric model $f(y; \theta)$. Three test statistics often constructed from the likelihood function are the Wald or maximum likelihood statistic:

$$w_e = \left(\widehat{\theta} - \theta_0\right)^T j\left(\widehat{\theta}\right)\left(\widehat{\theta} - \theta_0\right) \qquad [9]$$

the Rao or score statistic:

$$w_u = U(\theta_0)^T\left\{j\left(\widehat{\theta}\right)\right\}^{-1}U(\theta_0) \qquad [10]$$

and the likelihood ratio statistic:

$$w = 2\left\{\ell\left(\widehat{\theta}\right) - \ell(\theta_0)\right\} \qquad [11]$$

where in [8], [9], and [10] the following notation is used:

$$\sup_{\theta} L(\theta; y) = L\left(\widehat{\theta}; y\right) \qquad [12]$$

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ U(\theta) &= \ell'(\theta) \end{aligned} \qquad [13]$$

$$j(\theta) = -\ell''(\theta) \qquad [14]$$

The distributions of each of the statistics [8], [9], and [10] can be approximated by a χ^2_k distribution, under the model $f(y; \theta_0)$, with k the dimension of θ in the model. This relies on being able to apply a central limit theorem to $U(\theta)$, and to identifying the maximum likelihood estimator $\widehat{\theta}$ with the root of the equation $U(\theta) = 0$. The precise regularity conditions needed are somewhat elaborate; see for example Lehmann and Casella (1998: Chapter 6) and references therein. The important point is that under the simple null hypothesis the approximate distributions of each of these test statistics is known, and p-values readily computed.

In the case that the hypothesis is composite, a similar triple of test statistics computed from the likelihood function is available, but the notation needed to define them is more elaborate. The details can be found for example in Cox and Hinkley (1974: Chapter 9.3) and the notation above follows theirs.

If θ is a one-dimensional parameter, then a one-sided version of the test statistics given at [8], [9], and [10] can be used instead, as the signed square root of w_e, w_u, or w follows approximately a standard normal distribution.

It is rare that the exact distribution of test statistics can be computed, but the normal or chi-squared approximation can often be improved. These improvements are discussed in Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997), Severini (2000) and Brazzale et al. (2007). One conclusion of this work is that among the three test statistics above, the signed square root of the likelihood ratio statistic w is generally preferred on a number of grounds, including the accuracy of the normal approximation to its exact distribution. This is true for both simple and composite tests of a scalar parameter.

Significance Functions and Posterior Probabilities

We can also use a test of significance to consider the whole set or interval of values of θ that are consistent with the data. If θ is scalar one of the simplest ways to do this is to compute $r(\theta) = \pm\sqrt{w(\theta)}$ as a function of θ, and tabulate or plot $\Phi\{r(\theta)\}$ against θ, choosing the negative root for $\widehat{\theta} < \theta$, and the positive square root otherwise. This significance function will in regular models decrease from one to zero as θ ranges over an interval of values. The θ values for which $\Phi(r)$ is 0.975 and 0.025 provide the endpoints of an approximate 95% confidence interval for θ. This approach is emphasized in Fraser (1991).

In a Bayesian approach to inference it is possible to make probability statements about the parameter or parameters in the model by constructing a posterior probability distribution for them. In a model with a scalar parameter θ based on a prior $\pi(\theta)$ and model $f(y; \theta)$ we compute a posterior density for θ as

$$\pi(\theta|y) \propto f(y; \theta)\pi(\theta) \qquad [15]$$

and can assess any particular value θ_0 by computing

$$\int_{\theta_0}^{\infty} \pi(\theta|y)d\theta$$

the posterior probability of θ being larger than θ_0. This posterior probability is different from a p-value: a p-value assesses the data in light of a fixed value of θ, and the posterior probability assesses a fixed value of θ in light of the probability distribution ascribed to the parameter. Many people find a posterior probability easier to understand, and indeed often interpret the p-value in this way. There is a literature on choosing priors called 'matching priors' to reconcile these two approaches to inference; see Kass and Wasserman (1996), and Datta and Mukerjee (2004); see also Bayesian Statistics. Unfortunately it is not possible to find so-called matching priors that are simultaneously matching for all components of a vector parameter.

Hypothesis Testing

Little has been said here about the choice of a test statistic for carrying out a test of significance. The difficulty is that the theory of significance testing provides no guidance on this choice. The likelihood-based test statistics described above have proved to be reasonably effective in parametric models, but in a more complicated problem, such as testing the goodness of fit of a hypothesized model, this approach is often not available. To make further progress in the choice of test statistics, the classical approach is to formulate a notion of a 'powerful' test statistic, i.e., one that will reliably give small p-values when the null hypothesis is not correct. To do this in a systematic way requires specifying what model might hold if in fact the null hypothesis is incorrect. In parametric models where the null hypothesis is $H_0 : \theta = \theta_0$ the alternative may well be $H_a : \theta \neq \theta_0$. In more general settings the null hypothesis might be H_0 : 'the response variable follows a normal distribution' and the alternative H_a : 'the response variable does not follow a normal distribution.' Even in the parametric setting if θ is a vector parameter it may be necessary to consider what direction in the parameter space away from θ_0 is of interest. The formalization of these ideas is the theory of hypothesis testing, which considers both null and alternative hypotheses and optimal choices of test statistics (*see* Hypothesis Testing, in Statistics; Goodness of Fit: Overview; Methods and Models).

Further Topics

Fixed Level Testing

The problem of focusing on one or two so-called critical p-values is sometimes referred to as fixed-level testing. This was useful when computation of p-values was a very lengthy exercise, and it was usual to provide tables of critical values. It is now usually a very routine matter to compute the exact p-value, which is usually (and should be) reported along with other details such as sample size, estimated effect size, and details of the study design. There is still in some quarters a reliance on fixed level test, with the result that studies for which the p-value is judged 'not statistically significant' may not be published. This is sometimes called the 'file drawer problem,' and a quantitative analysis was considered in Dawid and Dickey (1977).

In several fields it is now standard to make the results of all studies on a given topic, including inconclusive studies, available online. In healthcare the Cochrane collaboration is perhaps the best-known of these. This issue is particularly important for *meta-analysis*, see Section Combining Tests of Significance.

Achievable *p*-values

In some problems where the distribution of the test statistic under the null hypothesis is concentrated on a discrete set, the number of available p-values will be relatively small. This happens with categorical data, especially if the sample size is small. Some authors have argued that for such highly discrete situations a better assessment of the null hypothesis can be achieved by the use of Barnard's mid p-value, which replaces $\Pr\{t(Y) \geq t^0\}$ with $(1/2)\Pr\{t(Y) = t^0\} + \Pr\{t(y) > t^0\}$, where $t(Y)$ is the statistic on which the significance test is based, and t^0 is the observed value in the data; see Agresti (1992) and references therein.

Combining Tests of Significance

The p-value is a function of the data, taking small values when the data are incompatible with the null hypothesis, and vice versa. As a function of y the p-value itself has a distribution under the model $f(y;\ \theta)$ and in particular under the null hypothesis H_0 has the uniform distribution on the interval $(0,1)$. In principle then if we have computed p-values from a number of different datasets the p-values can be compared to observations from a $U(0,1)$ distribution with the objective of obtaining evidence of failure of the null hypothesis across the collection of datasets. This is one of the ideas behind *meta-analysis*; *see* Meta-analysis: Overview; Meta Analysis: Tools. One difficulty is that the studies will nearly always differ in a number of respects that may mean they are not all measuring the same parameter, or measuring it in the same way. Another difficulty is that studies for which the p-value is not 'statistically significant' will not have been published, and thus are unavailable to be included in a *meta-analysis*. This selection effect may seriously bias the results of the *metametaanalysis*. In some fields this selection effect has been lessened by the practice of registering all trials in a given subject area on the internet.

Two-Sided Testing

A point of confusion in the evaluation of p-values for testing scalar parameters is the distinction sometimes made between one-sided and two-sided tests of significance. A reliable procedure is to compute the p-value as the twice the smaller of the probabilities that the test statistic is larger than or smaller than the observed value, under the null hypothesis. This so-called two-sided p-value measures disagreement with the null hypothesis in two directions away from the null hypothesis, toward the alternative that the, say, new treatment is worse than the old treatment as well as better. In the absence of very

concrete a priori evidence that the alternative hypothesis is genuinely one-sided this *p*-value is preferable.

In testing hypotheses about parameters of dimension larger than one, it can be difficult to decide on the relevant direction away from the null hypothesis. One solution is to identify the parameters of interest individually, and carry out separate tests on each of these parameters in turn. This will usually be effective if there are a relatively small number of parameters of interest. In applications involving computation of a very large number of *p*-values, new techniques are needed; these are briefly discussed in the next section.

Difficulties with Significance Tests

Sample Size

In Figure 1 we show the *p*-value for a one-sided *t*-test with an observed value of the *t*-statistic equal to 2.0, as a function of the sample size. In this example, and quite generally, the *p*-value is a decreasing function of the size of the sample, so that a very large study is more likely to show 'statistical significance' than a smaller study. This has led to considerable criticism of the *p*-value as a summary measure. The *p*-value is also sometimes misinterpreted, especially when it is small, as the probability that the null hypothesis is false. Some statisticians have argued that for this reason posterior probabilities are a better measure of disagreement with the null hypothesis; see for example Berger and Sellke (1987) and Schervish (1996).

To some extent the criticism can be countered by noting that the *p*-value is just one summary measure of a set of data, and excessive reliance on one measure is inappropriate. In a parametric setting it is nearly always advisable to provide, along with the *p*-value for testing a particular value of the parameter of interest, an estimate of the effect size, or some relevant parameter of the model, along with an indication of the precision of this estimate. This can be accomplished by reporting a significance function, if the parameter of interest is one-dimensional. At a more practical level, it should always be noted that a small *p*-value should be interpreted in the context of other aspects of the study. For example a *p*-value of less than 0.05 could be based on a very small difference in a study of 10 000 cases or a relatively large difference in a study of 1000 cases. A 1% reduction in an average response may be of substantial importance for one scientific context, and meaningless for others; this needs to be evaluated in that context, and not by relying on the fact that it is 'statistically significant.' Unfortunately, the notion that a study report is complete if and only if the *p*-value is found to be less than 0.05 is fairly widely ingrained in some disciplines, and indeed forms a part of the requirements of some government agencies for approving new treatments.

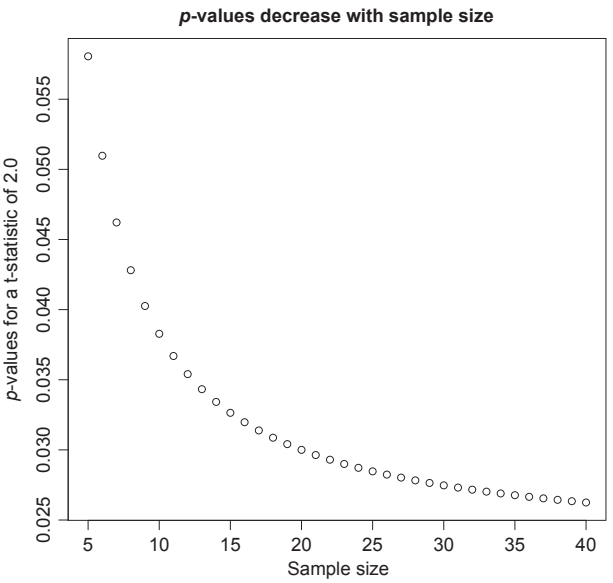

Figure 1 The *p*-value associated with an observed *t*-statistic equal to 2.0, as a function of the sample size.

Multiple Testing

If a number of significance tests are carried out on the same set of data, but the significance level, or *p*-value, that is reported is the smallest of these, then a different analysis is needed. This smallest *p*-value will not have the interpretation of a *p*-value from a single test; for example if we regard *p*-values less that 0.05 as 'significant,' then we would expect to find 5 spuriously significant results in 100 tests, on average.

In some applications there might be two or three thousand tests carried out, all of a similar type. One example arises in image analysis, for example comparing the blood flow in each of several thousand pixels or voxels of a brain scan, under two (or more) conditions. Large numbers of *t*-tests are often conducted in genomic analysis of expression arrays, with again the goal of comparing two conditions. The *p*-value for a single such comparison is no longer a reliable measure of the consistency of the data with the null hypothesis; several *p*-values will be small even when the null hypothesis is true, simply by chance.

There is a large literature on assessments of hypotheses under multiple testing, that has become particularly prominent in biological applications, but also has applications to many other sciences. *See* False Discovery Rate; Multiple Comparisons A good statistical theory reference is Efron (2010), and there are a number of more specialized books, for particular scientific fields, such as Dudoit and van der Laan (2008). In high-energy physics the search for new particles involves looking through a great many energy bands, sometimes called in that context the 'look elsewhere' effect, and it has become conventional to require a so-called '5-sigma' result to claim discovery of a new particle. This relates the observed level of significance to the probability that a normal random variable exceeds its mean by five standard deviations; this probability is 3×10^{-7}.

In many areas of research the delineation of the number of significance tests that have been carried out, but are not reported, is less clear. For example, in a new epidemiological study of the health effects of some environmental agent, several different models may be fitted to the data, including perhaps various transformations of the exposure measurements, different levels of control for confounding, and so on. Again the evidence provided by any single reported significance test needs to be considered in the light of the many other

tests that may have been formally or informally applied to the same data.

The reliance on multiple significance tests forms a part of the increasing concern, in the academic and in the popular press, about the replicability of published scientific research; see for example, Ioannidis (2005), or Economist (2013). Research in this area is ongoing; for a recent discussion in the medical context see Jager and Leek (2013) and the accompanying discussion.

Conclusion

A test of statistical significance is a mathematical calculation based on a test statistic, a null hypothesis, and the distribution of the test statistic under the null hypothesis. The result of the test is to indicate whether the data are consistent with the null hypothesis: if they are not, then either we have observed an event of low probability, or the null hypothesis is not correct.

The choice of test statistic is in principle arbitrary, but in practice might be determined by convention in the field of application, by intuition in a relatively new setting, or by one or more considerations developed in statistical theory. It is convenient to use test statistics whose distributions can be easily calculated exactly or to a good approximation. It is useful to use a test statistic that is sensitive to the particular departures from the null hypothesis that are of particular interest in the application.

A test of statistical significance is just one component of the analysis of a set of data, and should be supplemented by estimates of effects of interest, considerations related to sample size, and a discussion of the validity of any assumptions of independence or underlying models that have been made in the analysis. A statistically significant result is not necessarily an important result in any particular analysis, but needs to be considered in the context of research in that field.

An eloquent introduction to tests of significance is given in Fisher (1935: Chapter II). Kalbfleisch (1979: Chapter 12) is a good text book reference at an undergraduate level. The discussion here draw considerably from Cox and Hinkley (1974: Chapter 3), which is a good reference at a more advanced level. An excellent overview is given in Cox (1977).

See also: Bayesian Statistics; Distributions, Statistical: Approximations; Distributions, Statistical: Special and Discrete; Hypothesis Testing in Statistics; Multiple Comparisons, Statistics of; Statistical Sufficiency.

Bibliography

Agresti, A., 1992. A survey of exact inference for contingency tables. Statistical Science 7, 131–153.
Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall, London.
Berger, J.O., Sellke, T., 1987. Testing a point null hypothesis: the irreconcilability of p-values and evidence. Journal of the American Statistical Association 82, 112–122.
Brazzale, A.R., Davison, A.C., Reid, N., 2007. Applied Asymptotics: Case Studies in Small Sample Statistics. Cambridge University Press, Cambridge.
Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.
Cox, D.R., 1977. The role of significance tests. Scandinavian Journal of Statistics 4, 49–70.
Datta, G., Mukerjee, R., 2004. Probability Matching Priors: Higher Order Asymptotics. Springer-Verlag, New York.
Dawid, A.P., Dickey, J.M., 1977. Likelihood and Bayesian inference from selectively reported data. Journal of the American Statistical Association 72, 845–850.
Dudoit, S., van der Laan, M.J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer-Verlag, New York.
Economist, October 11, 2013. Unreliable Research: Trouble at the Lab. Print Edition.
Efron, B., 2010. Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. (Institute of Mathematical Statistics Monographs). Cambridge University Press, Cambridge.
Fisher, R.A., 1935. The Design of Experiments. Oliver & Boyd, Edinburgh.
Fraser, D.A.S., 1991. Statistical inference: likelihood to significance. Journal of the American Statistical Association 86, 258–265.
Ioannidis, J., 2005. Why most published research findings are false. PLOS Medicine 2, 696–701.
Jager, L.R., Leek, J., 2013. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15, 1–12.
Kalbfleisch, J.G., 1979. Probability and Statistical Inference, vol. 2. Springer-Verlag, New York.
Kass, R.E., Wasserman, L., 1996. Formal rules for selecting prior distributions: a review and annotated bibliography. Journal of the American Statistical Association 91, 1343–1370.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer-Verlag, New York.
Pace, L., Salvan, A., 1997. Principles of Statistical Inference from a Neo-Fisherian Perspective. World Scientific, Singapore.
Schervish, M.J., 1996. P values: what they are and what they are not. American Statistician 96, 203–206.
Severini, T.A., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

Simmel, Georg (1858–1918)

O Rammstedt, University of Bielefeld, Germany
Natàlia Cantó-Milà, Universitat Oberta de Catalunya (UOC), Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents a general overview of the life and thoughts of Georg Simmel from the perspective of the social sciences. After a brief biographical introduction, a discussion of those aspects of his *oeuvre*, which enjoy close proximity with the social sciences, particularly sociology, will constitute the main focus of the article. Together with Max Weber and Emile Durkheim, Simmel is one of the great classical theorists of sociology, and, as will be shown herein, contributed substantially to the development of the discipline, freeing it from needing to be an all-encompassing study of society as a whole. Thus, an overview of Simmel's fundamental theoretical concepts will be delivered, with the aim of illustrating his central relevance to the field of the social sciences.

Georg Simmel: A Biographical Sketch

Georg Simmel was born in the heart of Berlin on 1 March 1858. He was the youngest son of Flora (born Bodstein) and Eduard Simmel, who, although coming from Jewish families, had been baptized into Christianity (he as a Catholic, she as a Protestant). Following the early death of Simmel's father in 1874, the family suffered serious financial difficulties, which, as far as the young Georg was concerned, were overcome thanks to Julius Friedländer, a friend of the family (cofounder of the music publishing company 'Peters'). Friedländer felt a strong sympathy for the young Simmel, and indeed took him under his wings as his protégé. Thus, Georg could attend high school and then university in Berlin, where he studied philosophy, history, art history, and social psychology (*Völkerpsychologie*). Simmel received his degree as Doctor of Philosophy in 1881, but not without difficulty: his first attempt at a doctoral thesis, *Psychological and Ethnological Studies on the Origins of Music*, was not accepted, and he had instead to submit his previous work on Kant, *On the Essence of Matter – Das Wesen der Materie nach Kant's Physischer Monadologie*, which had earned him a prize (Köhnke, 1996). The *Habilitation* (postdoctoral qualification to lecture) came next in 1885, for which he also encountered some controversy, and after this his academic career began immediately as a *Privatdozent* (external lecturer). An extraordinary (*außerordentliche*) professorship without salary followed in 1901, and indeed he had to wait until 1914 before he was offered a regular professorship at the University of Strasbourg, where he remained until his death on 26 September 1918, shortly before the end of the First World War.

During his lifetime, Simmel was a well-known figure in Berlin's cultural world. He did not restrict himself merely to scientific or academic matters, but consistently showed great interest in the politics of his time, including contemporary social problems, and in the world of arts. He sought to be in the presence of, and in contact with, the intellectuals and artists of his day. He married the painter Gertrud Kinel, with whom he had a son, Hans, and maintained friendships with Rainer Maria Rilke, Stefan George, and Auguste Rodin, amongst many others. At his home in Berlin, he organized private meetings and seminars (Simmel's *privatissimo*), whose participants he would choose personally.

Georg Simmel and the Social Sciences

Simmel's contributions to the social sciences are immeasurable. Indeed, together with Max Weber (1864–1920) and Émile Durkheim (1858–1917), he set the foundations of that which today is understood as sociology and as the social sciences in general.

Simmel's interest in the social sciences, especially in sociology, can be traced to the very beginning of his academic career. After attending seminars offered by Moritz Lazarus and Heymann Steinthal (founders of the *Völkerpsychologie* (social psychology)) during his student years, he became a member of Gustav Schmoller's circle, where he became acquainted with the debates concerning the national economy of the time, and for whom he held a conference on *The Psychology of Money* (GSG2 1989, pp. 49–65), which would later constitute the first pillar of one of his major works, *The Philosophy of Money* (Simmel, 1989a). In both of these circles Simmel became more aware of, and sensitized toward, social questions. Schmoller's engagement with social questions, together with Lazarus' and Steinthal's emphasis on the level of *'Überindividualität'* (supraindividuality), and their relativistic worldview and insistence that ethical principles are not of universal validity, as well as Simmel's own interest in Spencer's social theory, all helped to shape the contours of his sociological approach. These various influences were melded together with Simmel's philosophical orientation, particularly his interest in Kant, which yielded a new and rather far-reaching sociological perspective that would evolve extensively throughout his life, being merged, at the end of it, with his increasing interest for the philosophy of life.

Simmel's first sociological work, *On Social Differentiation* (*Über Sociale Differenzierung*: Simmel, 1989c), written at the very beginning of his academic career, was deeply influenced by Herbert Spencer and the ideas of Gustav Schmoller and his circle. Slowly, however, as the 1890s drew on, the admiration Simmel had felt for Spencer's theories turned into rejection,

and he distanced himself from an evolutionary-organicist approach to sociology. In this rejection of such theories, he brought his knowledge of Kant to his sociological thinking, and this would become the basis of his later contact and dialog with the members of the southern-German neo-Kantian school, such as Rickert, Windelband, and with the sociologist who was closest to their ideas: Max Weber. This contact with neo-Kantianism influenced Simmel's approach to the social sciences at the end of the nineteenth century and during the initial years of the 1900s. However he did not become a neo-Kantian, and his interpretation and use of Kant's works in his academic work differed from the tendencies that the neo-Kantians shared.

From contemporary testimonies we know that Simmel was in fact one of the greatest lecturers at the University of Berlin, that his seminars were attended by a great number of students, and that his books were widely read and discussed (Gassen and Landmann, 1993). It is thus difficult to understand why Simmel did not have a more successful academic career. We know from his correspondences and from Gassen and Landmann's attempts to reconstruct Simmel's life and *oeuvre* that Georg Jellinek engaged himself, though with no positive results, in seeking to obtain an ordinary professorship for Simmel at the University of Heidelberg in 1908 (with, following Jellinek's death, a further attempt by Alfred Weber in 1912, and another in 1915). His converted Jewish background (and thus, despite of being German, being pointed at with the label of 'Jewishness') surely played a significant role in the lack of recognition Simmel received from the academic system. However, this was not the only cause for his somehow peculiar academic career. His peculiar and original understanding of science, which diverged greatly from established patterns, as well as his characteristic mode of writing, using essays instead of more 'academic' and standardized forms, contributed to his not being accepted into the rather formal and classical German academic milieu. Both attempts at obtaining a professorship for Simmel were blocked by the bureaucracy of the Grand Dukedom of Baden. In fact the letter of evaluation written by the Berlin historian Schäfer regarding a possible professorship for Simmel in Heidelberg in 1908 has as the primary argument for denying Simmel's potential ability to be a good professor his 'Jewishness,' which, according to Schäfer, too obviously tinged Simmel's character and intellectual efforts with a strong relativism and negativity, which could not be good for any student (Gassen and Landmann, 1993; Köhnke, 1996).

Throughout his life Simmel had to fight against these kinds of accusation, which also came from the more conservative neo-Kantian schools of thought; and he endeavored to build a 'positive relativism' in an attempt to show that he did not question 'absolute pillars' and thus leave us with nothing, but sought instead to show that this sense of 'absoluteness' was also a product of human reciprocal actions and effects (*Wechselwirkungen*), and was not fundamentally absolute. As the planets are not held by any strong hand, but by an equilibrium of forces, and notwithstanding the fact that they cannot fall just like that but are firmly held at their place by that very same equilibrium of forces, so do the values which we strongly believe in, the moral constructions or aesthetic evaluations: they are held by the invisible, and yet, so strong, relational threads that keep them at their place. Such an argument was too much for the society and scientific milieu of the time to accept and forgive, even when Simmel later radically rejected his *Introduction to the Moral Sciences* (*Einleitung in die Moralwissenschaft*: Simmel, 1989b; Simmel, 1991), referring to it as a sin of his youth. This was the work in which his relativism, in the critical, 'negative' sense, had been most fully in bloom (Köhnke, 1996).

Simmel's interest in the social sciences and in the positive relativism (which we nowadays tend to call as 'relationism') accompanied him throughout his life. Actually, despite the fact that Simmel shifted his attention during his career from the central concept of society, toward culture first, and life at the end, his interest for 'society,' 'culture,' and 'life' was present with different intensities all the way through his academic writing. Thus, we cannot be surprised that Simmel, in his letter to Célestin Bouglé of 2 March 1908, wrote: "At the moment I am occupied with printing my *Sociology*, which has finally come to its end," and, sentences later, added that work on this book 'had dragged on for fifteen years' (Simmel Archive at the University of Bielefeld); thus indicating that his engagement with sociology was a long-term project. Considering that *Sociology* (*Sociologie*) was first published in 1908 (and therefore his work on it must have begun around 1893), we can find its first seed in his article *The Problem of Sociology* (*Das Problem der Sociologie*) originally published in 1894.

Simmel must have thought this article to be a significant contribution to sociology, in as much as he endeavored to spread it abroad as much as possible. Hence, the French translation of *The Problem of Sociology* appeared, simultaneously with the original German version, in September 1894. The American translation appeared in the *Annals of the American Academy of Political and Social Science* a year later, and, by the end of the century, the Italian and Russian translations were also in print. The American translation is of particular significance, in as much as Simmel emphasized therein, in a footnote, that sociology involved empirical research, and should not be thought of as an independent offshoot from philosophy, but as a science concerning the social problems of his time. In the Italian translation he delivered an updated version of the text, wherein he introduced clear references to his theoretical polemic with Emile Durkheim.

From his letter to Célestin Bouglé of 15 February 1894, we know that Simmel was, after the completion of *The Problem of Sociology*, quite excited by this new discipline, and that he did not foresee a shift away to any other fields of inquiry in his immediate future. During those years he worked on most of the key sociological areas of research, thus articulating the key social problems of his time within a sociological framework: for example, workers' and women's movements, religion, the family, prostitution, medicine, and ethics, amongst many others. He seemed deeply interested in putting his new theoretical proposals and framework for the constitution of sociology into practice. He realized that the institutionalization of the new discipline would be reinforced by the establishment of journals for the discipline; hence his participation in the *Institut Internationale de Sociologie*, for whom he became vice-president; the *American Journal of Sociology*, and, although only briefly (due to differences with Emile Durkheim), *l'Année sociologique* (Rammstedt, 1992: 4). He also played with the idea of creating his own sociological magazine. Another means of solidifying

the role of sociology within the scientific sphere was through the academia, and he engaged himself in organizing sociological seminars, offering them uninterruptedly from 1893 until his death in 1918. On 15 June 1898, he wrote to Jellinek: "I am absolutely convinced that the problem, which I have presented in the *Sociology*, opens a new and important field of knowledge, and the teaching of the forms of sociation as such, in abstraction from their contents, truly represents a promising synthesis, a fruitful and immense task and understanding" (Simmel Archive at the University of Bielefeld). (The collection of all the letters from and to Simmel, which have been found up to date and are available in the Georg Simmel Archive, have been published in the mean time as Simmel (2008a,b).)

It is clear, however, that Simmel did not work continuously on the *Sociology* for 15 years as, from 1897 to 1900, he worked almost exclusively on *The Philosophy of Money* (*Philosophie des Geldes*), and also found time for the writing and publication of his *Kant* (1904), as well as a reprint of his *Introduction to the Moral Sciences* (1904), which he had intended to rewrite, for he no longer accepted most of the ideas he had presented therein when it was first published in 1892 (although he did not achieve this); the revised edition of *The Problems of the Philosophy of History* (*Probleme der Geschichtsphilosophie* 1905), *The Philosophy of Fashion* (*Philosophie der Mode* 1905), and *Religion* (*Die Religion* 1906/1912) were all worked on by Simmel during this period too (Rammstedt, 1992).

In *The Problem of Sociology*, he questioned for the first time the lack of a theoretically well-defined object of study for the emerging discipline, and sought to develop a specific sociological approach, which would entail a distinct object of study, in order to bestow legitimacy and scientific concreteness to a discipline, under attack from different, and more settled, lines of fire. Fearing his call had not been heard, Simmel endeavored to prove his point by writing a broader work, within which he attempted to put into practice the main guidelines he had suggested as being central to the newly emerging discipline. In this way the monograph *Sociology*, almost 1000 pages long, was cobbled together from various bits and pieces, taken from several essays he had written between the publication of *The Problem of Sociology* in 1894 and its final completion in 1908. Simmel, as can be understood from his letters, was aware of the incompleteness of this work, but rescued it by saying that it was an attempt to realize that which he had suggested almost 15 years earlier, for it had not been noticed enough by the scientific community.

By that time (1908), Simmel did not wish to be thought of as only a sociologist. This was due to sociology not being an established discipline within the academic world, which therefore did not allow him to obtain a professorship in the field (i.e., offering very little recognition). Nevertheless he never quite abandoned the field of sociology and continued to write about religion, women's issues, and the family. He merely broadened his scope to include aesthetics, culture, and *Lebensphilosophie* (the philosophy of life), whilst participating at the same time in the founding of the German Sociological Society (*Deutsche Gesellschaft für Soziologie*), for whom he served as one of the chairmen until 1913.

The writing of *Sociology* marks the end of Simmel's strongly Kantian period, and represents a turning point in his interests. From this moment onward, he refused to be merely labeled as a sociologist, seeking instead to return to philosophy. Indeed, following the publication of *Sociology*, he did not publish any sociological articles for 9 years (although he continued, as has already been mentioned, to offer sociological seminars until his death), instead devoting his efforts to philosophy, history, and the philosophy of art. Hence, as part of Simmel's output from these years we find, amongst other works, *Kant und Goethe* (*Kant und Goethe* 1906/1916), *Goethe* (1913), *Kant* (1904/1913/1918), *The Principal Problems of Philosophy* (*Hauptprobleme der Philosophie* 1911), *The Philosophical Culture* (*Philosophische Kultur* 1911/1918), and *Rembrandt* (1916). This new direction was accompanied and partly motivated by Simmel's acknowledgment of Henri Bergson's *oeuvre*, and his inclination toward *Lebensphilosophie* (the philosophy of life; see Fitzi, 1999; Simmel, 1999a). Thus, 'culture' and, later on, 'life' became the primary focus of Simmel's theoretical work, and consequently 'society' was pushed into a secondary role (without ever totally disappearing).

During these years Simmel occupied himself with the study of artists, of their production, and how the relation between life and its formal expression is crystallized in their work. It was as if Simmel had lost interest in sociology, as he did not write a single line concerning it for years. Yet, unexpectedly in 1917, a year before his death, he wrote *Grundfragen der Soziologie* (*The Fundamental Questions of Sociology*). The impetus for writing this book came from a publisher (*Sammlung Göschen*), who intended to print an introductory work to sociology, and which he invited Simmel to write, because of the success other works of his had enjoyed for the same publisher. If Simmel had indeed distanced himself from all sociological questions, it is likely he would have merely reached for the shelf to his previous works and rewritten them in shorter form. But this was not the case, because, although he utilized older material, Simmel rewrote and redefined his perspective, and in *The Fundamental Questions of Sociology*, scarcely 100 pages long, presented the final stage of his sociological reflections, which melded together his previous study of forms of sociation with a perspective from the philosophy of life. This approach to sociology fell into neglect after his death, awaiting revitalization, harboring a broad scope, and original perspective for new generations of sociologists to rediscover. Simmel's last work (1918) was a philosophical contribution to the philosophy of life (*Lebensphilosophie*), the *Lebensanschauung – View of Life* (Simmel, 1999a).

The Object of Sociology

At a time when sociology was still far from being an established discipline, instead seeking to stand up and open its eyes for the first time, Simmel sought to release it from the burden of being 'the science of society.' According to him this burden was an impossible one for the newborn discipline to carry, in as much as being the science of society meant having to compete with already settled and established disciplines for the legitimacy of its object of study: law and history, psychology, and ethnology, all could argue the case that society was their object, thus leaving sociology with the mere pretense of including elements from them all.

Viewed from this perspective, as an object, society was an all-encompassing matter but, at the same time, it eluded any scientific investigation, just like sand falling through our fingers. Simmel contributed to the demystification of 'society' as some kind of essential entity (as it appeared in nineteenth-century sociology, and remained as such in the sociology of Durkheim and Tönnies), instead explaining it as a dynamic process, a continuous happening, a continuous becoming, which is nothing more than the mere sum of the existing forms of sociation.

Individuals as well as society/ies are not units in and of themselves, though they may appear as self-sufficient units depending on the distance the observer interposes between him-/herself and them (as observed objects). However, Simmel argued in *The Fundamental Questions of Sociology* (Simmel, 1999a: 62–68) that if we merely take individuals and pretend, through adopting a significant distance, to approach 'society,' or a social phenomenon, we will not reach our goal. Therefore, when conducting a sociological inquiry, what we must seek to do is base ourselves on the forms of sociation, which are built in *Wechselwirkung*. Simmel used this concept when either addressing 'interactions' or following the Kantian definition as 'reciprocal actions and effects' (Simmel, 1978). The difference between the two possible meanings of this concept, or, more precisely, the differentiation of the two different concepts hidden behind the same word, has been addressed in the English translation of *The Philosophy of Money*; so the usual mistake of translating *Wechselwirkung* as a direct synonym for interaction has been corrected.

According to Simmel, what actually takes place between individuals will be seen to be that which constitutes the object of sociology, not merely the individuals by themselves, or society as a whole, for, as previously mentioned, society is nothing but the sum of forms of sociation, a continuous process in and for which these forms intertwine and combine themselves to form the whole (Simmel, 1992a: 19). Thus, Simmel defined society as the sum of the forms of sociation which emerge and crystallize in *Wechselwirkung*, i.e., in reciprocal actions and effects.

The Concept of 'Form'

If we take reciprocal actions and effects as our starting point, we achieve only a perspective common to all social and human sciences: sociology as method. In order to construct sociology as an independent discipline, with a specific object of study, it is necessary to analytically differentiate between form and content, that is, 'forms' as means and patterns of interaction between individuals, social groups or institutions, and 'contents' as the concrete practices that are embedded within, and channeled by, forms. Thus, social forms were conceived as being the object of analysis for a scientific sociology, which would only be possible empirically. Contents, on the other hand, should be left to other disciplines, such as psychology or history, to analyze.

Simmel's sociological theory is orientated toward these 'forms of sociation' (concept which he included in the writing of *Sociology*). For individuals to become social, they need to rely on such forms in order to channel their contents, as forms represent their only means of participation in social interaction. Forms are independent from contents, and when analyzing them, it is necessary that they are abstracted from individual, particular participation in concrete interactions: the question of social forms does not include that regarding the specific relationships between the participants involved in concrete interactions; we are dealing only with that which is *between* them, as for an analysis of social forms the particular individuals involved in them are irrelevant. Human beings, with their particular 'contents,' are socialized human beings who have learned to realize and express these contents within a social framework: via 'exteriorizing,' via acting through forms. Hence forms are social objectivations, which impose themselves, with their norms, upon particular individuals, an imposition which could only be annulled by isolation. These impositions and constraints are part of what sociation (and socialization) is a concept which plays a key role in Simmel's theory, for he actually maintained that human beings are not social beings by nature, or, at least, asked us not to take this sociability for granted.

Simmel placed the study of the forms of sociation, as a product of reciprocal actions and effects, at the center of his formal, or pure (*reine*), sociology. The form of competition can be used as an example of what Simmel actually meant by 'forms.' Competition does not imply any specific contents, but a set of 'rules of the game' so to speak that remain the same, independent of those who are competing, or what they are competing for. Competition forms 'contents' (such as job seeking, or looking for the attention of a beloved person, amongst many others), limits the boundaries of actions, and actually brings them into being via giving them a framework and shape in which they can appear, be understood and given an appropriate reply to, in the arena of actions and effects.

This situation of having the choice between multiple forms for channeling one content is stressed appreciatively in Simmel's sociology, particularly in his *Lebensphilosophie*, when he no longer contrasted the concept of form with the concept of content, but with the concept of life. In this later period, forms are the crystallization of the unretainable flow of life, yet also the channels of expressing life, for life cannot be expressed as only itself. Simmel explained this apparent paradox by asserting that life is 'more-life' and also 'more-than-life.' The concept of 'more-life' merely implies that this continuous flow connects every complete moment with the next; 'more-than-life' implies that life is not life if it does not transcend its boundaries, and becomes crystallized in a form. So life becomes art, or science; for example, life becomes externalized and crystallized and hence expressed and fulfilled. For instance, we should understand art as one form of expressing (*äussern*) life through (artistic, aesthetic) forms. Actually uncovering how this perspective can be applied to sociology may appear to be a difficult task; the first step toward clarifying this was made by Simmel himself in his *The Fundamental Questions of Sociology*.

'Me' and 'You'

Deliberately distancing himself from the concept of 'alter ego,' Simmel emphasized the significance the concept of 'you' should

have to all sociological theory. He articulated and embedded this concept within his – sociological – theory of knowledge. In parallel with Kant's question on nature, Simmel formulated this question as 'How is society possible?' – the first and well-known digression of his *Sociology* (1908). In order to answer this question he proposed three a priori. According to the first, 'Me' and 'You' would see each other as 'to some extent generalized' (Simmel, 1992a: 47), and he therefore assumed that each individual's perception of the other (i.e., 'you') would also be generalized to a certain degree. The second a priori affirms that each individual as an 'element of a group is not only a part of the society, but, on top of this, also something else' (Simmel, 1992a: 51) that derives its own uniqueness, its individuality from the duality of being/not being sociated. The third a priori is related to the assignment of each individual to a position within their society (in the sense of the sum of highly differentiated reciprocal actions and effects): 'that each individual is assigned, according to his qualities, to a particular position within his social milieu: that this position, which is ideally suited to him, actually exists within the social whole – this is the supposition according to which every individual lives his social life' (Simmel, 1992a: 59); this is what Simmel meant when he wrote about the 'general value of individuality.'

These sociological a priori, orientated toward role, type, and individual/society, also allow us to understand the central aspects of Simmel's methodology, which can be framed together with the concepts of 'differentiation,' in the sense of division and difference, and of 'dualism,' in the sense of an irreconcilable, tragic opposition of the elements of a social whole.

A Proposal for Three Sociologies

According to Simmel sociology needed to stand as an autonomous science in close relationship with 'the principal problem areas' (Simmel, 1999a: 76–84) of social life, even when these areas were distant from each other. Already in 1895 he asserted that just which name to give to any particular group was quite unimportant, in as much as the real question was to state problems and to solve them, and not at all to discuss the names which we should give to any particular group (Simmel, 1895: 420).

He stressed this orientation into problem areas again in 1917, when he returned to theorizing on *The Fundamental Questions of Sociology*. Thus, as an attempt to approach these main questions (i.e., which relationships exist between society and its elements, individuals) he concentrated on three different sets of problems with concrete examples, particularly in the third part of the first chapter. The first focuses on objectivity as a component of the social sphere of experience; the second includes the actual facts of life, in which, and through which, social groups are realized; and the third emphasizes the significance of society (Fitzi, 1999, 2002; Simmel, 1999a: 84) for grasping and understanding attitudes toward the world and life. These three problem areas correspond to his three proposed sociologies: the 'general sociology,' which, wherever society exists, deals with the central relationships between the individuals and the social constructions resulting from them. The aim of general sociology is to show which reciprocally orientated values these social constructions and individuals possess. He gave an example for general sociology in the second chapter of *The Fundamental Questions of Sociology*, by illustrating the relationships which exist between the 'social and the individual level,' a perspective which, at the same time, sociologizes the masses theory.

The 'pure or formal' sociology focuses upon the multiple forms individuals use to channel contents in their reciprocal actions and effects with others, be it with other people, social groups, or social organizations, and thus they constitute society. According to him, each form wins, entwined in these social processes, 'an own life, a performance, which is free from all roots in contents' (Simmel, 1999a: 106). The example Simmel chose for his formal sociology was the form of 'sociability' (*Geselligkeit*), an elaboration upon this form became the third chapter of *The Fundamental Questions of Sociology*. Finally, the 'philosophical sociology' circles the boundary of the 'exact, orientated toward the immediate understanding of the factual' (Simmel, 1999a: 84) empirical sociology: on the one hand the theory of knowledge (*Erkenntnistheorie*), on the other, the attempts to complement, through hypothesis and speculation, the unavoidably fragmentary character of factual, empirical phenomena (*Empirie*), in order to build a complete whole (Simmel, 1999a: 85). In the fourth chapter, entitled 'The Individual and Society in Eighteenth and Nineteenth Century Views of Life,' Simmel illustrated the abstract necessity of individual freedom, which should be understood as a reaction to the contemporary, increasing, social constraints and obligations. Thus, reference is made to general conflicts between individual and society, which derive from the irreconcilable conflict between the idea of society as a whole, which 'requires from its elements the one-sidedness of a partial function,' and the individual, who knows him/herself to be only partially socialized, and 'him/herself who wants to be whole' (Simmel, 1999a: 123).

Simmel and Contemporary Sociology

Simmel's theoretical significance to contemporary sociology (and to the social sciences in general) resides in the various theories, which explicitly or implicitly are built upon the basis of his works. As examples, it will be sufficient to name symbolic interactionism, conflict theory, functionalism, the sociology of small groups, and theories of modernity. Simmel also introduced into sociology the essay as an academic form of analysis, whilst his digressions within *Sociology* should be mentioned as well, for they have become generally recognized as classical texts; see, for example, his digressions on 'The Letter' (*Brief*), 'Faithfulness' (*Treue*), 'Gratefulness' (*Dankbarkeit*), and 'The Stranger' (*Fremde*). But above all sociology owes Simmel the freedom he gave it from the fixation on the 'individual and society' as an ontic object – in hindsight, a point of no return.

When the first version of this article was written more than 10 years ago (2000), Simmel scholars were still fighting an academic battle for Simmel's recognition as a classical author of sociology, such as Durkheim or Weber. In April 2012, we can

state that this academic battle has been won. Georg Simmel belongs now to the founding fathers of sociology and the social sciences, and is fully recognized as one of the main authors of the classics' generation. This did not happen by chance. It has implied the work of a few sociologists, who have devoted their academic life and work to keeping the memory of Simmel's works, and applying his theoretical lenses to current problems. Otthein Rammstedt discovered Simmel in his years as a student in Frankfurt, especially under the guidance of Gottfried Salomon-Delatour, and he has worked during decades to help young scholars interpret the richness of Simmel's works and legacy, among others breaking stereotypes that were given to students long before they ever read a line of Simmel's original works. Otthein Rammstedt has coordinated and edited Simmel's collected works (24 volumes edited mainly in German which collect the whole of Simmel's works, and all the letters that could be recovered in an arduous research in archives throughout Europe and the USA). The last volume of Simmel's collected works is about to see the light of day, and thus the whole of Simmel's works will be available to the German reading public. At the same time, great efforts have been made to make Simmel's works available in other languages. By the time the first version of this article was published, we were still missing the possibility of a complete translation of *Sociology* into English; 3 years ago this translation was published. David Frisby in the UK and Kurt Wolff and Donald Levine in the USA made great efforts to spread Simmel's works within the English-speaking academic milieu. David Frisby corrected the wrong translation of *Wechselwirkung* for interaction, which had been widely accepted for many years as the only possible way of translating *Wechselwirkung* into English. At the same time, he applied a Simmelian framework to the analysis of modern (and postmodern culture), and city landscapes. Frisby's understanding of Simmel's works as somehow 'impressionistic' and fragmentary was not shared by other Simmel scholars, like, for instance, Otthein Rammstedt, who argued and argues against the stereotype of Simmel's theoretical framework and analyses being fragmentary. At the same time, authors like Klaus Lichtblau or Klaus Christian Köhnke sought to investigate Simmel's relation to his contemporaries, unmasking some myths about Simmel (like for instance that which wrongly stated that he had been a neo-Kantian). Despite minor differences, this generation of scholars worked together to make Simmel's oeuvre available to contemporary social scientists and philosophers. In the meantime, another generation has been able to work upon, and with, Simmel's theories, upon the shoulders of giants – once again. This new generation has focused more prominently upon the relational paradigm in Simmel's works, seeking to apply it, as we understand he would have recommended us to do, to our contemporary sociological problems and questions. (See, for instance, Fitzi, 2002; Papilloud, 2002, 2003; Cantó-Milà, 2005; Phyyhtinen, 2010; Tokarzewska, 2010.)

Furthermore, the anniversaries of the publication of *The Philosophy of Money* first (1900), and *Sociology* later (1908) have also given occasion for the organization of conferences and seminars that have sought to actualize Simmel's works, and discuss the striking validity of most of the theoretical insights he left for us to explore further. (See, for instance, Rammstedt et al., 2003; Rol and Papilloud, 2009.) And, last but not least, Simmel's works have been translated into Spanish, Portuguese, Italian, French, Japanese, Chinese, etc. thus starting a global debate upon his sociological and philosophical legacy.

See also: Capitalism; Cities: Internal Structure; Differentiation: Social; Family as Institution; Groups, Sociology of; Interactionism, Symbolic; Knowledge, Sociology of; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Modernization and Modernity in History; Modernization, Sociological Theories of; Money, Sociology of; Religion, Sociology of; Social Movements, History of: General; Sociological Theory; Sociology: Overview; Symbolic Interaction: Methodology; Urban Sociology.

Bibliography

Cantó-Milà, N., 2005. A Sociological Theory of Value: Georg Simmel's Sociological Relationism. Transcript & Transaction Publishers, Bielefeld & New Brunswick.

Dahme, H.J., Rammstedt, O. (Eds.), 1983. Georg Simmel. Schriften zur Soziologie: Eine Auswahl. Suhrkamp, Frankfurt am Main.

Fitzi, G., 1999. Henri Bergson und Georg Simmel: Ein Dialog zwischen Leben und krieg. Die persönliche Beziehung und der wissenschaftliche Austausch zweier Intellektuellen im deutsch-französischen Kontext vor dem Ersten Weltkrieg (Doctoral thesis). University of Bielefeld.

Fitzi, G., 2002. Soziale Erfahrung und Lebensphilosophie. Georg Simmels Beziehung zu Henri Bergson. UVK, Konstanz.

Gassen, K., Landmann, M., 1993. Buch des Dankes an Georg Simmel. Briefe, Erinnerungen, Bibliographie. Zu seinem 100. Geburtstag am 1. März 1958. Duncker & Humblot, Berlin.

Köhnke, K.C., 1996. Der junge Simmel in Theoriebeziehungen und sozialen Bewegungen. Suhrkamp, Frankfurt am Main.

Papilloud, C., 2002. Le don de relation. Georg Simmel et Marcel Mauss. L'Harmattan, Paris.

Papilloud, C., 2003. La réciprocité. Diagnostic et destins d'un possible dans l'oeuvre de Georg Simmel. L'Harmattan, Paris.

Phyyhtinen, O., 2010. Simmel and the social. Palgrave MacMillan, Basingstoke.

Rol, C., Papilloud, C., 2009. Soziologie als Möglichkeit. 100 Jahre Georg Simmels Untersuchungen über die Formen der Vergesellschaftung. VS Verlag, Wiesbaden.

Rammstedt, O., 1992. Programm und Voraussetzungen der Soziologie Simmels. Simmel-Newsletter 2, 3–21.

Rammstedt, O., Papilloud, C., Cantó-Milà, N., Rol, C., 2003. Georg Simmels "Philosophie des Geldes": Aufsätze und Materialien. Suhrkamp, Frankfurt am Main.

Simmel, G., 1895. The problem of sociology. Annals of the American Academy of Political and Social Science 6, 412–423.

Simmel, G., 1978. In: Frisby, D., Bottomore, T. (Eds.), The Philosophy of Money. Routledge, London.

Simmel, Georg, 1989a. In: Frisby, D., Köhnke, K.C. (Eds.), GSG6: Philosophie des Geldes (1900/1907). Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1989b. In: Köhnke, K.C. (Ed.), GSG 3: Einleitung in die Moralwissenschaft. Eine Kritik der ethischen Grundbegriffe. Erster Band (1892/1904). Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1989c. In: Dahme, H.J. (Ed.), GSG 2: Aufsätze 1887–1890. Über sociale Differenzierung (1890). Die Probleme der Geschichtsphilosophie (1892). Surhkamp, Frankfurt am Main.

Simmel, Georg, 1991. In: Köhnke, K.C. (Ed.), GSG 4: Einleitung in die Moralwissenschaft. Eine Kritik der ethischen Grundbegriffe. Zweiter Band. Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1992a. In: Rammstedt, O. (Ed.), GSG 11: Soziologie. Untersuchungen über die Formen der Vergesellschaftung (1908). Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1992b. In: Dahme, H.J., Frisby, D. (Eds.), GSG 5: Aufsätze und Abhandlungen 1894–1900. Surhkamp, Frankfurt am Main.

Simmel, Georg, 1993. In: Cavalli, A., Krech, V. (Eds.), GSG 8: Aufsätze und Abhandlungen 1901–1908, Band II. Surhkamp, Frankfurt am Main.

Simmel, Georg, 1995a. In: Kramme, R., Rammstedt, A., Rammstedt, O. (Eds.), GSG 7: Aufsätze und Abhandlungen 1901–1908, Band I. Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1995b. In: Behr, M., Krech, V., Schmidt, G. (Eds.), GSG 10: Philosophie der Mode (1905). Die Religion (1906/1912). Kant und Goethe (1906/1916). Schopenhauer und Nietzsche (1907). Surhkamp, Frankfurt am Main.

Simmel, Georg, 1996. In: Kramme, R., Rammstedt, A., Rammstedt, O. (Eds.), GSG 14: Hauptpobleme der Philosophie (1910/1927). Philosophische Kultur (1911/1918). Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1997. In: Oakes, G., Röttgers, K. (Eds.), GSG 9: Kant (1904/1913/1918). Die Probleme der Geschichtsphilosophie, 2 Fassung (1905/1907). Suhrkamp, Frankfurt am Main.

Simmel, Georg, 1999a. In: Fitzi, G., Rammstedt, O. (Eds.), GSG 16: Der Krieg und die geistigen Entscheidungen (1917). Grundfragen der Soziologie (1917). Vom Wesen des historischen Verstehens (1918). Der Konflikt der modernen Kultur (1918). Lebensanschauung (1918). Surhkamp, Frankfurt am Main.

Simmel, Georg, 1999b. In: Köhnke, K.C. (Ed.), GSG 1: Das Wesen der Materie (1881). Abhandlungen 1882–1884. Rezensionen 1883–1901. Suhrkamp, Frankfurt am Main.

Simmel, Georg, 2008a. In: Köhnke, K.C. (Ed.), GSG 22: Briefe 1880–1911. Suhrkamp, Frankfurt am Main.

Simmel, Georg, 2008b. GSG 23: Briefe 1912-1918. Jugendbriefe. Suhrkamp, Frankfurt am Main.

Tokarzewska, M., 2010. Der Feste Grund Des Unberechenbaren: Georg Simmel Zwischen Soziologie und Literatur. VS Verlag, Wiesbaden.

Simon, Herbert (1916–2001)

John-Christopher Spender, ESADE, Universitat Ramon Lull, Barcelona, Spain; and School of Economics and Management, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Simon's reputation and legacy are curious and extensive, but opaque. Many talk of bounded rationality without understanding what he meant by it. His 1978 Nobel Prize in economics shocked the economics community, which did not see him as an economist at all. His PhD thesis, eventually published as *Administrative Behavior* has gone to four editions, yet is not much cited by organization theorists today. His work in artificial intelligence is more identifiable, but now regarded as part of the field's history given he left no theory or model in wide use today – with the little known exception of 'list processing,' more due to his collaborator Allan Newell. This article argues Simon was a philosopher, the first of our neo-Weberian organizational world. Bounded rationality is Simon's core philosophical contribution and it provides a new entry point into understanding how firms are able to create new economic value. It is not a model of the human mind. Our present obsession with positivist methodologies prevents us appreciating Simon's work – but its time will come.

Herbert A. Simon's reputation is broad, particularly in the public discourse for 'bounded rationality' is part of the general lexicon, 'satisficing' less so. Simon's permanent place in intellectual history was marked by the 1978 Nobel Prize in Economic Science, but his scholarly impact and legacy is more difficult to establish (Augier and March, 2004; Spender, 2013). Indeed, few economists ever understood how or why Simon won, especially as Sune Carlson, a noneconomist, wrote the commendation and made little mention of economics. Simon's Turing Award, ORSA/TIMS John von Neumann Theory Prize, and National Medal of Science were easier to understand, though the General Problem Solver Simon developed with Allan Newell is forgotten except to historians of artificial intelligence (AI). Newell and Simon's famous claim to have developed a 'thinking machine' over the 1955 Christmas break is forgotten too, though some recall the pair built an early chess-playing program. Others know Simon as a psychologically informed administrative theorist moving beyond the naive organization and management theories laid out by Max Weber or Frederick Taylor, or the behaviorist psychology of Skinner; hence his American Psychological Association Lifetime Achievement Award. His PhD-based book *Administrative Behavior (AB)*, published in 1947 and translated into a dozen languages, is now in its fourth edition. Carlson's Nobel commendation certainly cited *AB* but implied some of the prize was for *Organizations* (March and Simon, 1958), contributions to the 'behavioral theory of the firm,' and for two papers widely regarded as Simon's most substantial theoretical contributions: *A Behavioral Model of Rational Choice* (Simon, 1955) and *Rational Choice and the Structure of the Environment* (Simon, 1956). But none of these works created much change in economics, indeed the term 'bounded rationality' did not appear in *AB* until the 1958 Second Edition's new Foreword. Nonetheless *AB* has been studied by legions of doctoral students, and some masters' students too. Its Chapter 2, a withering attack on the 'proverbs of administration,' was based on a 1946 paper in *Public Administration Review* (Simon, 1946). Yet the proverbs, bureaucratic practices, and Scientific Management survive unscathed as the management practitioner's – and most researchers' – norm. Note also that the citations to *AB* pale in comparison with those to Michael Porter's 1980 *Competitive Strategy* or Jay Barney's 1991 RBV (*resource-based view*) article, indicating the researchers' current norms. Some specialists are aware of Simon's early work in public administration, also now invisible, and that he was awarded the American Political Science Association's Dwight Waldo Award in 1995.

What happened to these achievements and honors? Our inability to identify Simon's legacy today can be contrasted with the eulogies when he died in 2001. There was widespread recognition that a major scholar had passed on, though his precise contribution was difficult to pin down. Most obituaries focused on his AI work. Some obituarists noted his work helped establish that human beings lacked the reasoning capacity neoclassical economists presumed, but the obituaries did not consider the impact of that conclusion. Remarkably few noted he had inflicted irreparable damage to the neoclassical economists' claims, notwithstanding the string of Nobel Prizes awarded to US-based economists such as Paul Samuelson, Kenneth Arrow, Milton Friedman, Gerard Debreu, Franco Modigliani, Gary Becker, and Robert Lucas – rationalists one and all. Clearly little changed in economics; indeed it may be that Simon's lack of impact arises precisely because of the continued presence and power of Chicago-style economics and its associated neoliberal politics, even as those economists' bafflement in 2008 seemed to endorse Simon's charge.

If so, what can be said today? First, Simon successfully kicked open a door others had merely pointed to, most notably by Frank Knight with his *Risk, Uncertainty, and Profit* (Knight, 1921) and the notion of 'Knightian uncertainty.' If economics is to be a source of useful insight into the human condition, a self-sustaining group of scholars will have to pass through the door marked uncertainty into the 'new economics' that Simon anticipated (Coase and Wang, 2012). In the event, and again there are precedents for this, Simon was not able to go there himself. Rather it was as if he had shouted 'FIRE!' and, showing the audience how to escape, had to sacrifice his own project. If that sounds overly dramatic there were profoundly dramatic dimensions to Simon's scholarly life, career, and work, even as his private life was sublimely stable. Thus the most fruitful

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61286-2

reason to study Simon's enormous *oeuvre* is to clarify the potential research areas he pointed toward rather than the common claim that he overturned much in the several different fields he visited. In today's world of shrunken academic debates that contribute little to public understanding, Simon worked as a medieval scholar might on the broadest canvas imaginable – pondering the nature and limits of human understanding and their impact on our personal, economic, and national lives, a topic scoped out for most of the last three centuries by scholars such as Hume, Locke, Kant, and Marx. In spite of writing a vast amount, Simon seemingly wrote to smaller thinkers – academic economists, business school professors, psychologists, and computer scientists – who generally ignored the deeper implications of his work. In this respect Simon, or rather his potential contribution, was probably one of the earlier casualties in the post-World War II struggle to change the US universities. Simon's academic career began in 1942 and flourished in a period during which the US universities made radical changes to their mission and process, matters now emerging as the crisis of US educational performance is attracting both public and governmental concern and the long-term impact of the US' post-World War II policies are undergoing new assessment. Paradoxically, Simon was a fellow traveler with those whose thinking led to these changes and his Nobel may have been a bittersweet token of their regard.

But to begin at the beginning; the Simon parents were from the community of German immigrants settled in Milwaukee. Simon was a hard-working and exceptionally able son who won a full scholarship to the local large university (University of Chicago) after extensive personal study of political and public administration matters, robustly discussed around the dinner table. He arrived at the university in 1933 and took courses in the humanities and the social, physical, and biological sciences. Robert Maynard Hutchins had been appointed President of the university in 1929 and was transforming it by loosening up the prevailing orthodoxies in degree structure and content. Confronting the traditionalists, Hutchins eliminated the football program as a diverting irrelevance and set the stage for the remarkable period of sociological, legal, and political scholarship that lifted the university into the topmost rank. He left to head up the Ford Foundation in 1951, a matter of considerable later import. Meanwhile Simon set about exploiting the university's new curricular flexibilities by auditing upper level courses in several disciplines. He decided to major in economics but discovered it obliged him to take accounting, which he did not care for. He switched to political theory, dipping his toes into psychology, biology, mathematics, research methods, and philosophy – the renowned polymath was evolving. One of his teachers was Henry Simons, a forgotten giant of Chicago economics, and a contemporary of Frank Knight's. Simon wrote the question that was to define his academic *oeuvre* – "How do human beings reason when the conditions for rationality postulated by neoclassical economics are not met?" – came to him 'forcibly' during one of Simons' price theory courses. As a political science student with good insight, Simon intuited politics and administration, perhaps democratic society itself, would be superfluous were people actually like those the neoclassical economists imagined. The insight was not merely the fruitful germ of a critique of most economics, it implied a novel approach to the question "How might real-world decision making and administration be better understood and improved?" Simon said his overall project was to 'stiffen' the prevailing political and sociological discourse by exploring more scientific approaches; to improve the democratic process by ridding it of mere opinion, dogma, and bluster. In 1935, then aged 19 years, Simon wrote an undergraduate course paper on allocating Milwaukee's municipal funds between playground maintenance – such a planting trees and cutting grass – and 'leadership activities' – planning and running neighborhood programs, what we now call a messy or 'wicked' decision that went well beyond data analysis.

The paper led to collaboration with Clarence E. Ridley, head of ICMA (International City/County Management Association) and a manual for city administrators, published in 1936 to widespread acclaim. So Simon's list of publications began; eventually to number around 650. The manual leveraged into a 3-year Rockefeller Foundation funded project on public administration in California where Simon managed a team of mature researchers, his PhD still not completed, though continuing by special arrangement with the Economics Department at Chicago. Simon married in 1937, a relationship that lasted until his death. In 1942, finishing up his degree and California project, the Simons returned to Chicago where he gained an entry-level position at the Illinois Institute of Technology (IIT). After settling into his teaching, he began to recast his PhD thesis into *Administrative Behavior*, circulating about 100 mimeographed copies. One reader who replied at length was Chester Barnard, and so the principal influence over Simon's thinking came into place as he revised the initial drafts substantially. Both Barnard and Simon, of course, rejected naive rationalism and presumed the manager's task was defined by the nonrational aspects of those being managed – bounded rationality being one descriptor – and the consequent need to display 'leadership' as the practice of synthesizing the fragments known into an actionable picture. *Administrative Behavior*, with a Foreword by Barnard, was published in 1947, again to considerable acclaim. Simon was elevated to Full Professor at IIT and invited into the Cowles Commission (then based in Chicago) to do some of the research they funded. This brought him into the tight group of scholars circulating in and out of RAND in Santa Monica, California that included many future Nobel winners – such as Tjallings and Arrow, who later wrote *The Limits of Organization* (Arrow, 1974). In 1949, Simon was tempted away from IIT to help set up the Graduate School of Industrial Administration (GSIA) at Carnegie Institute of Technology (which became Carnegie Mellon University in 1967). He was to work with Lee Bach, the GSIA Dean, and the psychologist Harold Guetzkow, fellow Chicago graduates. Simon was a hot property and held out for a healthy salary and appointment as Head of the Department of Industrial Administration, one of GSIA's two departments.

Simon's vision was to create a top-notch social science department that could compete with Chicago, Harvard, Wharton, and Stanford, and also with the rising number of think tanks like the Cowles Commission and RAND. Equipped with the most scientific techniques and computing resources, and good research funding from the foundations and government agencies, GSIA would make substantial progress toward a synthesis of organization theory, finance, economics,

psychology, and decision science that the previous generation of grand theorists, such as Weber and Parsons, had failed to achieve. The vision was grounded in Hutchins's and later Merriam's notions of multidisciplinary education, in Simon's experience as a Chicago student, and awareness of government-level interest in applying the social sciences to national issues in the post-World War II era. The vision also reflected onto Simon's notion of the professional academic, for he expected everyone on the GSIA faculty to be able to teach an introductory class in each of the school's subject areas. As a result of the combined intellectual horse-power and politically attuned promotional and fund raising efforts of Bach and Simon's colleagues – including Cooper, March, Holt, Modigliani, and Cyert, GSIA sprang from almost nowhere onto the national intellectual scene. Things were good, scientific successes, and glittering prizes were visible on the horizon.

Then two very different events found Simon stuck between them. First, GSIA's success, expansion, and rising visibility attracted further top caliber economists. But the multidisciplinary Chicago of Bach and Simon's time was being rapidly reconstructed into the famed Economics Department of Hayek, Friedman, Stigler (all Nobel winners) and the libertarian Committee on Social Thought. The rationalist economics star rose rapidly and GSIA's incoming young economists had little time for Simon's multidimensional behavioral synthesizing – seeming to be more waffle than hard science. Even though he was a skilled academic administrator and rainmaker, Simon eventually accepted the inevitable and moved from GSIA to the Carnegie Psychology Department – where he invented 'AI' with Allan Newell. This was not Simon's first 'defeat.' In his first field of public administration he found himself facing a carefully orchestrated academic attack from those – especially Dwight Waldo – who regarded the field as demanding much more than the 'stiffened' rigorous methodologies Simon was promoting. What irony – to be rejected by rationalist economists for his insufficiently rigorous theories of decision making and by public policy leaders for being overly rigorous. But while Simon was a tough guy to work around, given his intellectual aggression and polymath capacity to know something about everything, he was always optimistic. The setbacks, which hurt him for the rest of his life, simply made him change course, not pace.

Simon said his meeting with Allen Newell at RAND in 1952 was the second most important event of his life – after meeting his wife Dorothea in 1936. Newell had been trained by Polya of *How to Solve It* and, understanding heuristics, helped Simon to an amazing breakthrough that few others could have made given their rationalist dispositions; to wit, that computers were not the number crunching devices they were generally thought to be by those (such as Johnny von Neumann) who had used them to do ballistics, statistics, or atom bomb computations, but were actually 'symbol processing' machines – and thus could handle heuristics. Simon realized digital computers could be harnessed to his lifelong project. Given bounded rationality, human thought, as opposed to scientific processes, could never be fully captured by Chicago-style rationalist models. On the other hand, it could be addressed by capturing humans' decision-making heuristics in the psychology lab – for these could then be programmed into a computer that cared not one whit about the scientific status of the relationships it was computing. As a chess player and skilled musician Simon realized what had to happen next. Simon had also sketched a chess program in a 1953 RAND report. He arranged for Newell to join him at Carnegie, notionally as his PhD student, and the two worked with Cliff Shaw, who had worked with von Neumann building the JOHNNIAC computer at RAND. Their first project was the Logic Theorist, programming mathematicians' proof-generation heuristic processes. This required a completely new kind of programming – list processing. The group's creation of the first forms of LISP was indeed a considerable innovation, seldom noted though fundamental to the expert systems, visual recognition, and data mining technologies of today's surveillance society. Simon also began to ask colleagues about human financial decision-making protocols and programmed these, underpinning GSIA's most notable academic achievement, the development of the 'behavioral theory of the firm.' The rapidly progressing AI work with Newell and Shaw came to fruition quickly in the 1955 Christmas claim. But once again such rapid academic advances attracted opposing conservative forces, sometimes irrational forces of ignorance and jealousy, sometimes more thoughtful and effectively critical.

Speculation about the nature of human thought goes back as far as history and thinkers like Aristotle, Descartes, Hume, Locke, and Kant had plenty to say. As news of the 'thinking machine' work of the Carnegie AI group spread, many were shocked and disheartened by what seemed to be philosophically uninformed claims to have broken through into a new understanding of human knowledge and learning. Of course those moving in the machine learning direction were greatly heartened, an especially ironic note struck when Simon recommended Kahneman for a Nobel for the work that he and Tversky had been doing on 'prospect theory,' which owed nothing to Simon's own work. But others were appalled, most famously Herbert Dreyfus, America's leading interpreter of Heidegger's work. Dreyfus, too, was drawn into the formative group of scholars circling RAND that went on to transform the American university. The group was not, as some charge, blindly marching in ideological lockstep, but was more of a constellation of some of the nation's smartest scholars, able to generate and handle major critiques of each other's work. Dreyfus's project was to explore and explain the limitations to digital computing. His 1965 report *Alchemy and AI* caused uproar since it implicitly charged RAND and many in its circle with hubris, perhaps the most serious sin a philosopher can commit. Dreyfus's book – likewise ironic since his brother Stuart was also at RAND working on JOHNNIAC – appeared as *What Computers Cannot Do* (1972). The book sparked immediate public discussion and Simon and Dreyfus plunged into a serious academic firefight, eventually calling a truce that revealed they had been talking past each other all along. But Dreyfus's work gave fresh ammunition to those in the struggle between 'hard' and 'soft' AI that is as alive today as it ever was. 'Hard' are those who believe we shall shortly make machines that are 'more intelligent than Man,' and are pitted against the 'soft' who say this is a seriously unintelligent conception of human intelligence. Simon was never able to escape being framed by this contest, though he had little time for either camp. He proposed, instead, that we understand computers as useful (and increasingly powerful) tools to help us make the

best of our bounded rationality, just as we make pulleys to lift heavy weights and books to help our memory. He had high expectations, of course, and was a fascinated geek. Knowing well that chess was a game of programmable heuristics, he famously anticipated a LISP-programmed computer beating the World Chess Champion, as IBM's Deep Blue did, though he did not anticipate the public's hunger for game shows and thus IBM's Watson winning performance.

The other thing that happened was that the very same foundations that had been so ready to fund the research of those circling around RAND were persuaded to insert themselves into the ideological and methodological debate between the rationalists and their philosophically and methodologically broader opponents. Bach appeared for both sides. Having initially supported Simon's vision for a multidisciplinary business school when Simon arrived at GSIA, Bach's support wavered as the political forces within the academic community and their funders began to line up behind neoliberal rationalism, so further inclining Simon to shift over to the Psychology Department. The RAND-centered group made a huge advance with Arrow's 'impossibility theorem,' published in *A Difficulty in the Concept of Social Welfare*, which many saw as conclusive mathematical proof that socialism could not work. Likewise Hayek's *Road to Serfdom* was flying off the shelves. The US' Cold War answer to Communism and other political dogma was to be hard science and rationalism, 'nothing but the facts, Ma'am.' The power of positivist methodologies was being rapidly expanded by the computer's ability to run ever-more complex statistics, and as a consequence expanding the notion that social, economic, and psychological facts could be framed 'objectively.' These methods were 'on the right side of history,' especially in the social sciences, and in the fact that politicians, city managers, and business people were able to make fact- or evidence-based decisions very differently from their forebears. Management of all types was reconceptualized as a 'science.'

From this point of view the business schools were seriously behind the curve, while GSIA was showing the way to the future. Even where accounting was taken seriously, most business schools treated managerial intuition reverently and the new 'science of management' ignored. Bach met with Donald Kirk David, then Dean at the Harvard Business School, later head of the Ford Foundation, and they eventually persuaded both the Carnegie Foundation and the Ford Foundation to fund inquiries and then issue reports on the parlous state of US business education. The reports were damning. Through their recommendations, and the subsequent funding these same foundation provided to those schools that followed their recommendations, the two 1959 reports had a massive and immediate impact. *Inter alia* they legitimated the priority of quantitative methods and modeling aesthetic that many now see as one of business education's main problems. Faculty members were increasingly selected on the basis of their displays of positivist methodological prowess. Whether they knew anything about the practice of managing or industrial operations became secondary. The business schools' business-experienced people were shown the door. Simon sensed he was being encircled professionally and institutionally, seeing his vision for GSIA rejected, but he also saw that he was being encircled disciplinarily, the answers he offered organization theorists, management theorists, and decision theorists in *AB* being rejected too. His only complaint was buried in his 1967 paper *The Business School: A Problem in Organizational Design* (Simon, 1967). This paper was written from the institutional security of his position in the Psychology Department and his professionally secure global reputation in AI. It is somewhat wistful as well as complex. Simon followed Barnard to remind readers that management is about leadership and effective synthesizing under conditions of uncertainty, not the ever-more rigorous application of rationalist theory. Any school of management that fails to provide its students with some understanding of how to deal with the real world's uncertainty is not meeting its pedagogical obligations. But, and this is the crux of the paper, conjoining education in analysis and synthesis within a single school creates great institutional strain; only the very best Deans will be able to contain it and manage it. It is a 'wicked' management problem. Simon had Lee Bach for a model of Deanship, of course, for no matter whether they agreed or not Bach's achievement in lifting GSIA into the top tier of business schools remains unmatched in US business school history (Khurana, 2007). When the strains are not handled effectively faculty members are driven solely by self-interest, which is to say they work toward academic ends within their own disciplines and ignore the students' need for training to operate and lead in real-world multidisciplinary contexts. Simon labeled this an academic tendency to the 'equilibrium' condition – a sly allusion to the poverty of equilibrium economics – and used the analogy of the rising 'entropy'; what effective Deans must struggle to contain. Today's business schools have not been able to deal with the entropic challenge Simon distilled from his own experience, insights, and disappointments.

Thus Simon's life showed a startling contrast between, on the one hand, the steadiness of his family background, his 50-year devotion to Carnegie Mellon (he lived in the same house throughout and became a CMU Trustee, with the Newell–Simon building named for him), and his steady pursuit of a problematic first intuited at the age of 19, and on the other, the prominent part he played in major disputes in three disparate disciplines: public administration, economics, and AI. Few scholars get to play such a role in one. But Simon's achievement in this regard – bridging three disciplines – says something fundamental about his work. The reason he was able to get to grips with the most fundamental questions in three separate field was not that he was bright, perhaps a genius. It was because he approached all these disciplines as Hume and Locke did, in terms of the fundamental puzzle about the nature of human understanding. In this sense, Simon was a modern philosopher who successfully moved beyond the boundaries most others take for granted and leave unquestioned. Although he wrote that he was an empirical positivist it is clear that he was in no way beholden to any positivist thought or method that dismisses synthesis in favor of rigorous deduction. He also grasped synthesizing, the creative nonlogical process of bringing ideas together, perhaps inductively, when dealing with real world's uncertainties. Understanding the complementarity of the methodological issues and disputes in political theory, sociology, and psychology, Simon managed to construct a sustainable and productive intellectual path between them. The glue that held all this thinking together

was his unswerving faith in the human mind's ability to reason and to create anew when reason stumbled. In this he shared much with John Locke, but there is no evidence that he studied Locke's work. More generally, Simon's grasp of multiple disciplines arose from his ability to escape their particular problematic and see them as instances of the larger struggle to understand human understanding itself.

Simon's legacy can be dimly identified in the several different disciplines he touched – public administration, organization and management theory, AI. But there are no memorials, as Scientific Management is a memorial to Frederick Taylor and his colleagues, or the Turing Machine is a memorial to Alan Turing's insight into the nature of computing machines. Perhaps Simon's most visible thumbprint is 'bounded rationality.' But, as many have noted, it is more often cited than understood. Indeed Simon noted at the time of his Nobel, "bounded rationality seemed to be dying a quiet death, in the US at least. The neoclassicists had clearly won the day." Simon's vision for a practical management science – driven by heuristics and 'best practices' rather than by rationalism – has left no mark. Business schools pay less attention to any of Simon's work. All of which suggests Simon's work – and Barnard's too – should be consigned to the trash can of history as nonrational mysticism that is irrelevant in today's data and analytics driven world.

But if we understand Simon primarily as a philosopher, arguably the first philosopher of the strange world of organizations we have created and now inhabit, he has much to say about our current anxieties. Max Weber lamented the 'disenchantment' of our world, increasingly dominated by rationalized processes and objectified values. Without doubt Simon's philosophizing remains out of place in the neoliberal rationalist world created since World War II. The possibilities of Simon's future influence – his promotion of human value- and faith-penetrated heuristics, not just at the level of democratic society but also at the level of organizational decision making, over the cold and lifeless machine logic of numbers ruling – turns on what happens to us when today's neoliberal dream or nightmare passes into our history. By then there may be little left of the Newmanesque University with its quiet probing and prodigious scholarship. Universities may have become little more than fee-earning think tanks and artificers to the government's chosen programs, and decay as their ideology dies. But if there are survivors, they will find great succor in Simon's work. Unlike the great scholars of the past, such as Hume, Locke, and Kant, Simon brought philosophy freshly into the Information Age. He straddled across the scientific territory of rigorous thought and design and the subjectivist territory of creative social thought. He brought the human heuristic maker together with the computer's extraordinary future. He brought human shortcomings and values together with bureaucratic notions of organization. He brought the power of modern data collection and analysis together with the political imperatives of the democratic process. In short, Simon, a masterly thinker of the modern era, showed us a possible future that dismisses those now calling for retraction into an imagined golden past of trust and higher social values. That we doubt the value of his work today is proof positive of our incapacity to see our presently parlous social, economic, and political situation. But as Santayana suggested, history is unkind to the unthinking – and we shall surely learn.

See also: Administration in Organizations; Heuristics, History of; Uncertainty: History of the Concept.

Bibliography

Arrow, Kenneth J., 1974. The Limits of Organization. W.W. Norton, New York.
Augier, Mie, March, James G. (Eds.), 2004. Models of a Man: Essays in Memory of Herbert A. Simon. MIT Press, Cambridge, MA.
Crowther-Heyck, Hunter, 2005. Herbert A. Simon – the Bounds of Reason in America. Johns Hopkins University Press, Baltimore, PA.
Crowther-Heyck, Hunter, 2006a. Herbert Simon and the GSIA: building an interdisciplinary community. Journal of the History of the Behavioral Sciences 42 (4), 311–334.
Crowther-Heyck, Hunter, 2006b. Patrons of the revolution: ideals and institutions in postwar behavioral science. Isis 97 (3), 420–446.
Crowther-Heyck, Hunter, April–June 2008a. Defining the computer: Herbert Simon and the bureaucratic mind – Part 1. IEEE Annals of the History of Computing 30 (2), 42–51.
Crowther-Heyck, Hunter, 2008b. Defining the computer: Herbert Simon and the bureaucratic mind – Part 2. IEEE Annals of the History of Computing 30 (2), 52–66.
Coase, Ronald H., Wang, Ning, 2012. Saving economics from the economists. Harvard Business Review 90 (12), 36.
Gigerenzer, Gerd, Todd, Peter M., The ABC Research Group, 1999. Simple Heuristics That Make Us Smart. Oxford University Press, New York.
Khurana, Rakesh, Spender, J.-C., 2012. Herbert A. Simon on what Ails business schools: more than 'a problem in organizational design'. Journal of Management Studies 49 (3), 619–639.
Khurana, Rakesh, 2007. From Higher Aims to Hired Hands: The Social Transformation of American Business Schools and the Unfulfilled Promise of Management as a Profession. Princeton University Press, Princeton, NJ.
Knight, Frank Hyneman, 1921. Risk, Uncertainty and Profit. Houghton Mifflin Company, Boston, MA.
March, James G., Simon, Herbert A., 1958. Organizations. John Wiley, New York.
Simon, Herbert A., 1946. The proverbs of administration. Public Administration Review 6 (1), 53–67.
Simon, Herbert A., 1955. A behavioral model of rational choice. The Quarterly Journal of Economics 69 (1), 99–118.
Simon, Herbert A., 1956. Rational choice and the structure of the environment. Psychological Review 63 (2), 129–138.
Simon, Herbert A., 1967. The business school: a problem in organizational design. Journal of Management Studies 4 (1), 1–16.
Simon, Herbert A., 1991. Models of My Life. Basic Books, New York.
Spender, J.-C., 2013. Herbert Alexander Simon: philosopher of the organizational life-world. In: Witzel, Morgen, Warner, Malcolm (Eds.), Oxford Handbook of Management Thinkers. Oxford University Press, Oxford, pp. 297–357.

Simultaneous Equation Estimates (Exact and Approximate), Distribution of

Kimio Morimune, Kyoto University, Kyoto, Japan

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 21, pp. 14101–14106, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

An example of simultaneous equations may consist production and consumption functions of a nation: $Y = a + bK + cL +$ error, and $C = d + eY +$ error. In equations, Y is GDP, K is the capital equipment, L is the labor input, and C is the consumption of a nation. Each equation is modeled to explain the variation of the dependent variable by the right-hand side of the equation. Error terms are added since equations cannot fully explain the variation of the dependent variables. This system does not consist of regression equations since Y is in the right-hand side of the C equation. Estimation of unknown coefficients and the properties of estimation methods are complicated compared with the ordinary least squares estimator. In this article, estimation methods of the whole system, and the asymptotic distribution of the estimators are explained first. Then estimation methods of a single equation are explained, and the asymptotic distribution is given. Using the framework of estimators, the small sample distribution by simulations and exact distribution of the single equation estimators are reviewed. Asymptotic expansions of the distributions of the single equation estimators are reviewed next, and finally the higher order efficiency of the system and the single equation estimators are explained.

A simple example of a system of linear simultaneous equations may consist of production and consumption functions of a nation: $Y = a + bK + cL +$ error, and $C = d + eY +$ error. The variables Y, K, L, and C represent the gross domestic product (GDP), the capital equipment, the labor input, and the consumption, respectively.

These variables are measures of the level of economic activity of a nation. In the production function, Y increases if the inputs K and/or L increase. C increases if Y increases in the consumption equation. Each equation is modeled to explain the variation in the left-hand side 'explained' variable by the variation in the right-hand side 'explanatory' variables. Error terms are added to analyze numerically the effect of the neglected factors from the right-hand side of the equation. These equations are different from the regression equations since the 'explained' variable Y is the 'explanatory' variable in the C equation, and Y and C are simultaneously determined by the two equations. Estimation of unknown coefficients and the properties of estimation methods are not straightforward compared with the ordinary least squares estimator.

In practice, this kind of simultaneous equation system is extended to include more than 100 equations, and regularly updated to measure the economic activities of a nation. It is indispensable to analyze numerically the effect of policy changes and public investments.

In this article, the statistical model and the estimation methods of all the equations are first explained, followed by the estimation methods of a single equation and their asymptotic distributions. Explained next are the exact distributions, the asymptotic expansions, and the higher order efficiency of the estimators.

The System of Simultaneous Equations and Identification of the System

We write the structural form of a system consisting of G simultaneous equations as

$$y_i = Y_i\beta_i + Z_i\gamma_i + u_i \equiv (Y_i Z_i)\delta_i + u_i, \quad i = 1, \ldots, G \qquad [1]$$

where y_i and Y_i are 1 and G_i subcolumns in the $T \times G$ matrix of whole endogenous variables $Y = (y_i, Y_i, Y_{ei})$, Z_i consists of K_i subcolumns in the $T \times K$ matrix of whole exogenous variables Z, β_i and γ_i are $G_i \times 1$ and $K_i \times 1$ column vectors of unknown coefficients, $\delta_i = (\beta'_i, \gamma'_i)'$, and u_i is the $T \times 1$ error term. This system of G equations with T observations is frequently summarized in a simple form

$$YB + Z\Gamma = U \qquad [2]$$

the ith column of which is $y_i - Y_i\beta_i - Z_i\gamma_i = u_i$, i.e., Eqn. [1]. The ith columns of B and Γ may be denoted as b_i and c_i where $(G - G_i - 1)$ and $(K - K_i)$ elements are zero so that $y_i - Y_i\beta_i - Z_i\gamma_i = Yb_i + Zc_i$. Zero elements are called zero restrictions. It is assumed that each row of U is independently distributed as $N(0, \sum)$. The reduced form of Eqn. [2] is

$$Y = Z\Pi + V \qquad [3]$$

where the $K \times G$ reduced form coefficient matrix is $\Pi = -\Gamma B^{-1}$, and the $T \times G$ reduced form error term is $V = UB^{-1}$. Each row of V is assumed to be independently distributed as $N(0, \Omega)$, and then $\sum = B^{-1}, \Omega B^{-1}$.

The definition $\Pi = -\Gamma B^{-1}$, or $-\Pi B = \Gamma$ is the key to identify structural coefficients. Coefficients in β_i are identified if they can be uniquely determined by the equation $-\Pi b_i = c_i$ given Π. This equation is reduced to $-\Pi_0(1, \beta'_i)' = 0$ denoting the $(K - K_i) \times (1 + G_i)$ submatrix in Π as Π_0. (Rows and columns are selected according to the zero elements in c_i and non-zero elements in b_i, respectively.) Given Π_0, this includes $(K - K_i)$ linear equations and G_i unknowns, and β_i is solvable if rank $(\Pi_0) = G_i$. This means $(K - K_i)$ must be at least G_i, or $L = K - K_i - G_i$ must be at least 0. If $L = 0$, β_i is uniquely determined. For the positive L, there are L linearly dependent rows in Π_0 since only G_i rows are necessary to determine β_i uniquely. Once β_i is determined, γ_i is determined by other K_i equations through $-\Pi b_i = c_i$. L is called the number of the degrees of overidentifiability of the ith equation. Structural coefficients are not uniquely estimable if they are not identified (see Statistical Identification and Estimability).

Estimation Methods of the Whole System and the Asymptotic Distribution

The full information maximum likelihood (FIML) estimator of all nonzero structural coefficients δ_i, $i = 1,\ldots, G$, follows from Eqn. [3]. Since it is in a linear regression form, the likelihood function can first be minimized with respect to Ω. Once Ω is replaced by the first-order condition, the likelihood function is concentrated where only B and Γ are unknown. The concentrated likelihood function is proportional to

$$\ln|\Omega_R|, \Omega_R = (Y - Z\Pi_R)'(Y - Z\Pi_R)/T, \quad \Pi_R = -\Gamma B^{-1}, \qquad [4]$$

and all zero restrictions are included in B and Γ matrices. In the FIML estimation, it is necessary to minimize $|\Omega_R|$ with respect to all non-zero structural coefficients.

The FIML estimator is consistent, and the asymptotic distribution is derived by the central limit theorem. Stacking δ_i, $i = 1,\ldots, G$ in a column vector δ, the FIML estimator δ asymptotically approaches $N(0, -I^{-1})$ as follows:

$$\sqrt{T}\left(\widehat{\delta} - \delta\right) \xrightarrow{D} N\left(0, -I^{-1}\right), \quad I = \lim_{T\to\infty} \frac{1}{T} E\left(\frac{\partial^2 \ln|\Omega_R|}{\partial\delta\, \partial\delta'}\right). \qquad [5]$$

I is the limit of the average of the information matrix, i.e., $-I^{-1}$ is the asymptotic Cramer–Rao lower bound. Then the FIML estimator is the best among consistent and asymptotically normal (BCAN) estimators.

The right-hand side endogenous variable Y_i in (1) is defined by a set of G_i columns in (3) such as $Y_i = Z\Pi_i + V_i$. By the definition of V, Y_i or, equivalently, V_i is correlated with u_i since columns in U are correlated with each other. The least squares estimator applied to (1) is inconsistent because of the correlation between Y_i and u_i. Since Z is assumed to be not correlated with U in the limit, Z is used as K instruments in the instrumental variable method estimator. Premultiplying Z' to (1), it follows that

$$\begin{aligned} Z'y_i &= \left(Z'Y_i, Z'Z_i\right)\delta_i + Z'u_i \\ &= \left(0, \ldots, 0, Z'Y_i, Z'Z_i, 0, \ldots, 0\right)\delta + u_i^*, \quad i = 1, \ldots, G \end{aligned} \qquad [6]$$

where the $K \times 1$ transformed right-hand side variables $Z'Y_i$ is not correlated with $u^*{}_i$ in the limit. Stacking all G transformed equations in a column form, the G equations are summarized as $w = X\delta + u^*$ where w and u^* stack $Z'y_i$ and $u^*{}_i$, $i = 1,\ldots, G$, respectively, and are $GK \times 1$. The covariance between $u^*{}_i$ and $u^*{}_j$ is $\sigma_{ij}(Z'Z)$ which is the ith row and jth column sub-block in the covariance matrix of u^*. (The whole covariance matrix can be written as $\sum\otimes,(Z'Z)$ where $\otimes$, signifies the Kroneker product.) Once $\sum$ is estimated consistently (by the 2SLS method explained in the next section), δ is efficiently estimated by the generalized least squares method

$$\widehat{\delta}_{3SLS} = \left\{X'\left[\widehat{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right]X\right\}^{-1}\left\{X'\left[\widehat{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right]w\right\}. \qquad [7]$$

This is the three-stage least squares (3SLS) estimator by Zellner and Theil (1962). The assumption of the normal distribution error is not required in this estimation. The 3SLS estimator is consistent and is BCAN since it has the same asymptotic distribution as the FIML estimator.

Estimation Methods of a Single Equation and the Asymptotic Distribution

An alternative way of estimating structural coefficients is to pick up one structural equation, the ith, in a G-equation system, and estimate δ_i neglecting zero restrictions in other equations. Because of this, the other $(G - 1)$ structural equations can be rewritten equivalently as $(G - 1)$ reduced form equations. The limited information maximum likelihood (LIML) estimator by Anderson and Rubin (1949) applies the FIML method to a $(1 + G_i)$-equation system consisting of (1) and $Y_i = Z\Pi_i + V_i$. This means the first column and the second G_i columns of B are $(1, -\beta'_i)'$ and $(0, I)'$, respectively, and the first column and the second G_i columns of Γ are $(\gamma'_i, 0')'$ and Π_i, respectively. (If we denote Y as (y_i, Y_i, Y_{ei}), Y_{ei} is weakly exogenous in estimating (1) or, equivalently, the structural and reduced form parameters of (y_i, Y_i) given Y_{ei} are variation-free from the reduced form coefficient of Y_{ei}. Then Y_{ei} is omitted from (2), (3), (4), and (8).) Using these limited information B and Γ matrices, we minimize $|\widehat{\Omega}_R|$ with respect to δ_i and Π_i. Defining $P_F = F(F'F)^{-1}F'$ for any full column matrix F, $G = (y_i, Y_i)'(P_Z - P_{Z_i})(y_i, Y_i)$, and $C = (y_i, Y_i)'(I - P_z)(y_i, Y_i)$, it turns out that β_i is estimated by minimizing the least variance ratio

$$\lambda(\beta_i) = \frac{\left(1, -\beta'_i\right)G\left(1, -\beta'_i\right)'}{\left(1, -\beta'_i\right)C\left(1, -\beta'_i\right)'} \qquad [8]$$

γ_i is estimated by the least squares method applied to (1) replacing β_i with the estimator.

The two stage least squares (2SLS) estimator is the generalized least squares estimator applied to Eqn. [6] using $Z'Z$ as the weight matrix. (See Eqn. [10], where k is set to 1.) The assumption of the normal distribution of the error term is not required in this estimation. Both LIML and 2SLS estimators are consistent, and the large sample distribution is

$$\sqrt{T}\left(\widehat{\delta}_i - \delta_i\right) \xrightarrow{D} N\left(0, -I^{-1}\left(\delta_i, \delta_i\right)\right) \qquad [9]$$

where I is calculated similarly to Eqn. [5] but Ω_R is defined using the limited information B and Γ matrices, and only the diagonal submatrix of I^{-1} which corresponds to δ_i is used. (Partial derivatives are calculated with respect to δ_i and columns in Π_i.) This asymptotic distribution is a particular case of Eqn. [5]. Both estimators are consistent and BCAN under the zero restrictions imposed on Eqn. [1].

The k-class estimator $\widehat{\delta}_i(k)$ unifies both LIML and 2SLS estimators (Theil 1961). It is

$$\begin{aligned} \widehat{\delta}_i(k) = &\begin{pmatrix} Y_i'P_zY_i - (k-1)Y_i(I - P_z)Y_i & Y_i'Z_i \\ Z_i'Y_i & Z_i'Z_i \end{pmatrix}^{-1} \\ &\times \begin{pmatrix} Y_i'P_z - (k-1)Y_i'(1 - P_z) \\ Z_i' \end{pmatrix} y_i. \end{aligned} \qquad [10]$$

This is the least squares estimator, the 2SLS estimator, and the LIML estimator when k is 0, 1, and $1 + \lambda$, respectively. There are two important properties in the k-class estimator. It is consistent if $p\lim_{T\to\infty} k = 1$, and is BCAN if $p\lim_{T\to\infty}\sqrt{T}(k - 1) = 0$. If k satisfies these conditions, the k-class estimator is consistent

and BCAN even when $Y_i'(I - P_z)Y_i$ and $Y'_i(I - P_Z)y_i$ are replaced with any matrix and a vector of order $O_P(T)$.

Exact Distributions of the Single-Equation Estimators

Several early studies compared the bias and mean squared errors of OLS, LIML, 2SLS, FIML, and 3SLS estimators by the Monte Carlo simulations since all but OLS estimators are consistent and are indistinguishable. The OLS estimator was often found as reliable as other consistent estimators. Later, the studies went on to t ratios, and the real defect of OLS estimators was found: the deviation from the standard normal distribution is worse than any other simultaneous equation methods. See Cragg (1967) for related papers.

Drawing general qualitative comparisons from simulations is difficult since simulations require setting values of all population parameters. Simulation studies on the small sample properties led to the derivation of the exact distributions, which were expected to permit the drawing of general comparisons without depending on the particular parameter values.

If $n \times 1$ column vectors x_t, $t = 1,\ldots, T$ are independently distributed $N(m_t, \Omega)$, the density function of $\Sigma_{i=1.T} x_i x_i'$ is the noncentral Wishart matrix denoted as $W_n(T, \Omega, M)$ where the noncentrality parameter is $M = \Sigma_{t=1,T}\ m_t m'_t$ (Anderson, 1958, Chap. 13). The study of the exact distribution of the single-equation estimators started from the fact that $G = (G_{kl})$ and $C = (C_{kl})$, k, $l = 1,\ldots, 1 + G_i$ in Eqn. [8] are the noncentral Wishart matrix $W_{1+G}(K - K_i, \Omega, M)$ with the noncentrality parameter matrix $M = \Pi_R' Z'(P_z - P_z)Z\Pi_R$, and the central Wishart matrix $W_{1+G}(T - K, \Omega, 0)$, respectively. The 2SLS, OLS, and LIML estimators of β_i are $G_{22}^{-1}G_{21}$, $(G_{22} + C_{22})^{-1}(G_{21} + C_{21})$, and $(G_{22} - \lambda C_{22})^{-1}(G_{21} - \lambda C_{21})$, respectively, and λ is the minimum root in the polynomial equation $|G - \lambda C| = 0$. Since all estimators are functions of elements in the G and C matrices, their distributions can be characterized by degrees of freedom of the two Wishart matrices, M and Ω matrices.

In deriving the exact density functions, the 2SLS and OLS estimators can be treated in a similar way. In the 2SLS estimator, the joint density of G is transformed into the joint density of $G_{22}^{-1}G_{21}$, $G_{11} - G_{12}\, G_{22}^{-1}\, G_{21}$, and G_{22}. Integrating out $G_{11} - G_{12}\ G_{22}^{-1}G_{21}$ and G_{22} results in the joint density of $G_{22}^{-1}G_{21}$. The resulting density function includes infinite terms, and zonal polynomials when G_i is greater than one. A pedagogical derivation is found in Press (1982, Chap. 5). In the LIML estimator, the joint density of G and C is transformed into that of characteristic roots and vectors. Since $\hat{\beta}$ is rewritten as a ratio of elements in a characteristic vector, the density function is derived by integrating out unnecessary random variables from the joint density function. However, the analytical operations are not easy when there are many endogenous variables. See Phillips (1983, 1985) to find comprehensive reviews on the 2SLS and LIML estimators, respectively.

It was somewhat fruitless to derive exact distributions because these include nuisance parameters and infinite terms. It was difficult to draw general conclusions on the qualitative properties of the estimators from the numerical evaluations of these distributions. See Anderson et al. (1982).

Qualitative properties of the estimators followed from the exact moments of estimators. Kinal (1980) proved that the (fixed) k-class estimator in a multiple endogenous variables case has moments up to $(T - K_i - G_i)$ if $0 \le k < 1$, and up to L if $k = 1$. Mariano and Sawa (1972) proved that, in the $G_i = 1$ case, the mean and variance of the LIML estimator do not exit. (In Monte Carlo simulations, the bias of LIML estimators was often found to be smaller than that of others even though the exact mean is infinite. This showed the clear limitation of the simulation methods.)

Asymptotic Expansions of the Distributions of the Single-Equation Estimators

Asymptotic expansion of the distribution was introduced as an analytical tool which is more accurate than the asymptotic distribution but is less complicated than the exact distributions. For instance, the t ratio statistic, say X, which is commonly used in econometrics, has the density function $f(x) = c\cdot[1 + (x^2/m)]^{-(1+m)/2}$ where c is a constant and m is the degree of freedom under conditions including the normally distributed error terms. Since the mean and variance of X is 0 and $m/(m–2)$, the standardized statistic $Z = \sqrt{(m-2)/m}\cdot X$ has the density $f(z) = c'\cdot[1 + (z^2/(m-2))]^{-(1+m)/2}$ where c' is a new constant. This density function is expanded to the third-order term as

$$f(z) = \phi(z)\left\{1 + \frac{1}{4m}\left(z^4 - 6z^2 + 3\right)\right\} + o\left(m^{-1}\right) \qquad [11]$$

where $\phi(z)$ is the standard normal density function, and the constant is adjusted so that the area under the curve is one. (Since the t distribution is symmetric around the origin, the $O(1/\sqrt{m})$ term does not appear in the right-hand side of the equation.) Rewriting in terms of X, the asymptotic expansion of the t statistic with m degrees of freedom is

$$f(x) = \phi(x) + \frac{1}{4m}\left(x^4 - 2x^2 - 1\right)\phi(x) + o\left(m^{-1}\right) \qquad [12]$$

The first term in the right-hand side is the $n(0, 1)$ density function. The second term in the right-hand side converges to zero if the limit is taken with respect to m. This second term gives deviation of $f(x)$ from $\phi(x)$, and is called the third-order term. For finite m, the third-order term is expected to improve the standard normal approximation. The numerical evaluation of this expansion is easy.

The asymptotic expansions in the simultaneous equation estimators are long and include nuisance parameter matrices such as q below. See Phillips (1983) for a review and Phillips (1977) for the validity of expansion. The asymptotic expansion does not require the assumption of the normal distribution of the error term. Fujikoshi et al. (1982) gave the expansion of the joint density of the estimators of δ_i. In their study, the bias of the estimators is calculated from the asymptotic expansions as

$$AM\left(\sqrt{T}\left(\hat{\delta}_i - \delta_i\right)\right) = \frac{1}{\sqrt{T}}(dL - 1)I^{-1}q + o\left(T^{-1}\right), \quad q = \frac{1}{\sigma^2}\begin{pmatrix}\varpi_{21} - \Omega_{22}\beta_i \\ 0\end{pmatrix}, \qquad [13]$$

where $\hat{\delta}$ is the estimator, d is 1 and 0 fo1r 2SLS and LIML estimators, respectively, and the Ω matrix is partitioned into submatrices conformable with the partitions of G and C matrices. $AM(\cdot)$ stands for the mean operator, but uses the asymptotic expansion for the density function. (Recall that the exact mean does not exist in the LIML estimator.) It is possible to compare the two estimators in terms of the calculated bias. For example, the bias of the 2SLS estimator is 0 when the degree of overidentifiability L is 1.

Further, the mean of the squared errors was calculated from the asymptotic expansions and used to compare estimators. It was proved that the mean squared error of the 2SLS estimator is smaller than that of the LIML estimator when L is less than or equal to 6. Historically, this kind of comparison of 'approximate' mean squared errors goes back to the 'Nagar expansion' by Nagar (1959) and the 'small disturbance expansion' by Kadane (1971). These qualitative comparisons gave researchers some guidance about the choice of estimators.

It was interesting to examine the accuracy of the approximations calculated by the asymptotic expansions of distributions. If the asymptotic expansions were accurate, calculation of the exact distributions could be avoided, and properties of estimators could be found easily from the approximations. However, the approximations were found not to be accurate enough to replace the exact distributions. There are many cases where the asymptotic expansion is accurate when the asymptotic distribution, the first term in the expansion, is already accurate; the asymptotic expansion is inaccurate when the asymptotic distribution is inaccurate. In particular, the asymptotic expansions are inaccurate when

a. The value of asymptotic variance is small;
b. The value of L is large; or
c. The structural error term is highly correlated with the right-hand side endogenous variable.

It is noted, first, $Z'Z/T$ is assumed to converge a nonsingular fixed matrix in the asymptotic theory. Second, for an attempt to improve the accuracy of asymptotic distributions by incorporating large L values, see Morimune (1983). Third, the asymptotic expansions cannot trace closely the skewed exact distributions that happen particularly when the correlation is high.

Higher-Order Efficiency of the System and the Single-Equation Estimators

The asymptotic expansions of distributions can be used in comparing the probabilities of concentration of estimators about the true parameter values. One estimator is more desirable than another if its probability is greater than that of the other. This measure was used in comparing the single-equation estimators, and some qualitative results were derived.

Furthermore, the third-order efficiency criterion was brought in the comparisons. This criterion requires that estimators be adjusted to have the same asymptotic bias as in Eqn. [13]. Then the adjusted estimators are compared, and the maximum likelihood estimator is proved most efficient. It has the highest concentration about the true parameter values in terms of the asymptotic expansion of distribution to the third-order $O(1/T)$ terms. (Akahira and Takeuchi, 1981, for example.) The adjusted maximum likelihood estimator has the smallest mean-squared error at the same time since the difference among estimators is found only in the mean-squared errors.

In whole system estimation, the FIML estimator is third-order efficient. The 3SLS estimator is less efficient than the FIML estimator in terms of the asymptotic probability of concentration once the bias of the two estimators is adjusted to be the same. Morimune and Sakata (1993) derived a simple adjustment of the 3SLS estimator so that the adjusted estimator has the same asymptotic expansion as the FIML estimator to the third-order $O(1/T)$ terms. This estimator is explained by modifying Eqn. [7]. In Eqn. [6], Y_i is replaced by $\hat{Y}_i \equiv Z\hat{\Pi}_i = Z(Z'Z)^{-1}Z'Y_i$ so that the X matrix consists of $Z'\hat{Y}_i$ and $Z'Z_i$. In the modified estimator, we estimate $\tilde{\Sigma}$ and Π by the first round 3SLS estimator and replace $\hat{Y}_i$ in X by $\tilde{Y} \equiv Z\tilde{\Pi}_i$ where $\tilde{\Pi}_i$ consists of proper subcolumns in $\tilde{\Pi} = -\hat{\Gamma}\hat{B}^{-1}$. The new X matrix is denoted as $\tilde{X}$. Finally, the new estimator is

$$\tilde{\delta}_{\mathrm{M3SLS}} = \left\{\tilde{X}'\left[\tilde{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right]X\right\}^{-1}\left\{\tilde{X}'\left[\tilde{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right]w\right\}. \quad [14]$$

This estimator has the same asymptotic expansion as the FIML estimator to the third-order terms and is third-order efficient.

The LIML and 2SLS estimators are simple cases of the FIML and 3SLS estimators, respectively. Then the modified 2SLS estimator which follows from Eqn. [14] has the same asymptotic expansion as the LIML estimator to the third-order term. The LIML estimator and the modified 2SLS estimator are third-order efficient in the single-equation estimation.

Conclusion

Laurence Klein received the 1980 Nobel prize for the creation of a macroeconometric model which is an empirical form of a simultaneous equation system, and for the application to the analysis of economic fluctuations and economic policies. The macroeconometric model became a standard tool to analyze the economies and policies of nations. Trygve Haavelmo received the 1989 Nobel prize for his contribution in the analysis of simultaneous structures. Haavelmo, together with other researchers at the Cowles Commission for Research in Economics, then at the University of Chicago, became the founders of simultaneous equation analysis in econometrics. Part of their research is collected in Hood and Koopmans (1953). Studies on the exact and approximate distributions of estimators came after the research conducted at the Cowles group, and helped to make econometrics rigorous.

Access to computers was the main concern when econometric model-building started spreading all over the world in the 1970s. Since then, computer facilities surrounding econometric model-building have changed greatly. Bulky mainframe computer have been replaced by personal computers. Computer programs were written individually, mostly in Fortran, and were used for regression analyses in model estimation as well as for simulation studies in econometrics theory. The packaged least squares programs replaced the individually written programs later in model estimation. They

run on personal computers and have facilitated greatly the conducting of empirical studies.

See also: Simultaneous Equation Estimation: Overview

Bibliography

Akahira, M., Takeuchi, K., 1981. Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency. Springer, New York.

Anderson, T.W., Rubin, H., 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics 20, 46–63.

Anderson, T.W., Kunitomo, N., Sawa, T., 1982. Evaluation of the distribution function of the limited information maximum likelihood estimator. Econometrica 50, 1009–1027.

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Basmann, R.L., 1957. A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25, 77–83.

Ciavolino, Enrico, Dahlgaard, Jens Jörn, 2009. Simultaneous Equation Model Based on the Generalized Maximum Entropy for Studying the Effect of Management Factors on Enterprise Performance. Journal of applied statistics 36 (7), 801–815.

Cragg, J.G., 1967. On the relative small sample properties of several structural equation estimators. Econometrica 35, 89–110.

Epple, Dennis, McCallum, Bennett T., 2006. Simultaneous Equation Econometrics: The Missing Example. Economic Inquiry 44 (2), 374–384.

Fujikoshi, Y., Morimune, K., Kunitomo, N., Taniguchi, M., 1982. Asymptotic expansions of the distributions of the estimates of coefficients in a simultaneous equation system. Journal of Econometrics 18, 191–205.

Hahn, Jinyong, Hausman, Jerry, 2002. Notes on Bias in Estimators for Simultaneous Equation Models. Economics Letters 75 (2), 237–241.

Hood, W.C., Koopmans, T.C., 1953. Studies in Econometric Methods. Wiley, New York.

Kadane, J.B., 1971. Comparison of k-class estimates when the disturbance is small. Econometrica 39, 723–737.

Kinal, T.W., 1980. The existence of moments of k-class estimators. Econometrica 48, 241–249.

Mariano, R.S., Sawa, T., 1972. The exact finite sample distribution of the limited information maximum likelihood estimator in the case of two included endogenous variables. Journal of the American Statistical Association 67, 159–163.

Mariano, Roberto S., 2001. Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications. In: Baltagi, Badi H. (Ed.), A Companion to Theoretical Econometrics. Blackwell, pp. 122–143.

Morimune, K., Sakata, S., 1993. Modified three stage-least squares estimator which is third-order efficient. Journal of Econometrics 57, 257–276.

Morimune, K., 1983. Approximate distributions of k-class estimators when the degree of over-identifiability is large compared with the sample size. Econometrica 51, 821–841.

Nagar, A.L., 1959. The bias and moment matrix of the general kappa-class estimators of the parameters in simultaneous equations. Econometrica 27, 575–595.

Phillips, P.C.B., 1977. The general theorem in the theory of asymptotic expansions as approximation to the finite sample distributions of econometric estimators. Econometrica 45, 1517–1534.

Phillips, P.C.B., 1983. Exact Phillips. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. 1. North-Holland, Amsterdam. Chap. 8.

Phillips, P.C.B., 1985. The exact distribution of LIML: 2. International Economic Reviews 25 (1), 249–261.

Press, S.J., 1982. Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. Robert E Krieger, Malabar, FL [original edn. 1972, 2nd edn.].

Theil, H., 1961. In: Economic Forecasts and Policy, 2nd edn. North-Holland, New York, pp. 231–232. 334–6.

Zellner, A., Theil, H., 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30, 54–78.

Simultaneous Equation Estimation: Overview

CF Christ, The Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 21, pp. 14106–14110, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Simultaneous equations abound in economics and elsewhere. The simplest economic model uses a supply equation and a demand equation to determine the price of a good and its quantity sold. Neither equation alone can determine either variable when each equation contains both: this is a model with two simultaneous equations. Estimates of the parameters of simultaneous equations are often needed. Parameters can be estimated only if the model and data together are capable of revealing their values, that is, if parameters are identifiable. Examples of identifiable and unidentifiable parameters are presented. For simultaneous equations, least squares regression estimators are typically biased and inconsistent (that is, incorrect on the average in repeated trials, and unlikely to be nearly correct even in large samples). Hence other estimators have been developed that, under suitable conditions, are unbiased in large samples, and consistent. Examples of such estimators are given: indirect least squares, instrumental variables, and two stage least squares. Other such estimators are briefly described: full information maximum likelihood, three stage least squares, and a Bayesian estimator. The particulars of a simultaneous equations model are crucial for its estimation and use, and are unlikely to be quite correct. Relevant tests are briefly discussed.

Simultaneous equations are important tools for understanding behavior when two or more variables are determined by the interaction of two or more relationships in such a way that causation is joint rather than unidirectional. Such situations abound in economics, but also occur elsewhere. Haavelmo (1943, 1944) began their modern econometric treatment.

The simplest economic example is the interaction of buyers and sellers in a competitive market which jointly determines the quantity sold and price. Another important example is the interaction of workers, consumers, investors, firms, and government in determining the economy's output, employment, price level, and interest rates, as in macroeconometric forecasting models. Even when one is interested only in a single equation, it often is best interpreted as one of a system of simultaneous equations.

Strotz and Wold (1960) argued that in principle every economic action is a response to a previous action of someone else, but even they agreed that simultaneous equations are useful when the data are yearly, quarterly, or monthly, because these periods are much longer than the typical market response time.

This article discusses the essentials of simultaneous equation estimation using a simple linear example.

A Simple Supply–Demand Example

Let q and p stand for the quantity sold and price of a good, y and w for the income and wealth of buyers (assumed independently determined), u_1 and u_2 for unobservable random shocks, and Greek letters for unknown constant parameters. Suppose that market equilibrium, in which the price is such that suppliers want to sell the same quantity that demanders what to buy, is described by these linear equations:

$$\text{supply:} \quad q = \gamma_1 + \beta_1 p + u_1, \quad \beta_1 > 0 \qquad [1]$$

$$\text{demand:} \quad q = \gamma_2 + \delta_2 y + \varepsilon_2 w + \beta_2 p + u_2, \quad \beta_2 < 0 \qquad [2]$$

Neither equation alone can determine either p or q, because each equation contains both: these are *simultaneous equations*. Solve them for p and q thus:

$$p = \pi_{11} + \pi_{12} y + \pi_{13} w + (u_2 - u_1)/\Delta \qquad [3]$$

$$q = \pi_{21} + \pi_{22} y + \pi_{23} w + (\beta_1 u_2 - \beta_2 u_1)/\Delta \qquad [4]$$

where

$$\pi_{11} = (\gamma_2 - \gamma_1)/\Delta, \quad \pi_{12} = \delta_2/\Delta, \quad \pi_{13} = \varepsilon_2/\Delta \qquad [5]$$

$$\pi_{21} = (\beta_1\gamma_2 - \beta_2\gamma_1)/\Delta, \quad \pi_{22} = \beta_1\delta_2/\Delta, \quad \pi_{23} = \beta_1\varepsilon_2/\Delta \qquad [6]$$

and

$$\Delta = \beta_1 - \beta_2 \qquad [7]$$

Types of Variables: Structural and Reduced Form Equations

The variables p and q which are to be explained by the model are *endogenous*. Equations [1] and [2] are structural equations. Each of them describes the behavior of one of the building blocks of the model, and (as is typical) contains more than one endogenous variable. Equations [3] and [4] are reduced form equations. Each of them contains just one endogenous variable, and determines its value as a function of parameters, shocks, and explanatory variables (here y and w).

The explanatory variables are predetermined if the shocks for any given period are independent of the explanatory variables for that period and all previous periods; they are *exogenous* if the shocks for each period are independent of the explanatory variables for every period. Thus all exogenous variables are predetermined, but not conversely. For example, the value of an endogenous variable from a previous period

cannot be exogenous, but it is predetermined if the shocks are serially independent.

The Need for Estimation of Parameters

One job of a reduced form equation is to tell how an endogenous variable responds to a change in any predetermined or exogenous variable. Typically, no single structural equation can do this. Another job of a reduced form equation is to forecast the value of an endogenous variable in a future period, based on expected future values of the predetermined and exogenous variables and of the shocks (the latter can safely be set at zero in most cases). Typically, no single structural equation can do this either. If the reduced form equations are to do these jobs, numerical values of their parameters are needed (in this example, values of the π's in eqns [3] and [4]).

Numerical values of the structural parameters are needed as well (in this example, the β's, γ's, δ_2, and ε_2 in eqns [1] and [2]). There are several reasons. First, one wants to understand each of the system's separate building blocks in its own right. Second, in overidentified cases (*see* section 'Estimating One Structural Equation at a Time') better estimates of the reduced form can be obtained by solving the estimated structure (as in eqns [5]–[7]) than by estimating the reduced form directly. Third, if forecasts made by the reduced form are poor, one wants to know which of the structural equations fit the forecast-period's data poorly, and which (if any) fit well. This is discovered by inserting observed values of the variables into each estimated structural equation to obtain an estimate of its forecast-period shock. Then one can revise the poorly fitting structural equation(s) and so try to improve the model's accuracy (*see* section 'The Problem of Choosing and Testing a Model'). Fourth, if one wants to find the new values of the reduced form parameters after a change in a structural parameter, one needs to know the old values of all the structural parameters, as well as the new value of the one that has changed. In the example, one can then use eqns [5]–[7] to compute the reduced form parameters. The critique of Robert Lucas (1976) provides a warning about pitfalls in doing this.

Least Squares Estimators

Least squares (LS) estimators (*see* Linear Hypothesis) of an equation's coefficients are biased if the shocks in each period are not independent of the explanatory variables in all periods. Clearly this is true of eqns [1] and [2], since eqns [3] and [4] show that both u_1 and u_2 influence both p and q. This is typically true of LS estimators of simultaneous structural equations. However, LS estimators often have small variances, so they may sometimes have acceptably small expected squared errors even when they are biased.

LS estimators of the π's in the reduced form, eqns [3] and [4], are unbiased if the shocks in each period have zero mean and constant variance, and are independent of y and w for all periods (so that y and w are exogenous). They have minimum variance among unbiased estimators and are consistent if in addition the shocks are uncorrelated across time. They remain consistent if the shocks are uncorrelated across time but y and w are predetermined rather than exogenous.

The generalized least squares (GLS) method is minimum variance unbiased if the explanatory variables are exogenous but the shocks are correlated across time. This method requires information about the variances and covariances of the shocks.

The Identifiability of Structural Parameters

Having estimates of the reduced form parameters, one can try to make them yield estimators of the structural parameters. In the example this means trying to solve the estimated version of eqns [5] and [7] for estimators of the γ's, β's, δ_2, and ε_2. Denote the LS estimators of the π's by $\widehat{\pi}$. Equations [5] and [6] show that $\widehat{\pi}_{22}/\widehat{\pi}_{12}$ is one estimator of β_1, and $\widehat{\pi}_{23}/\widehat{\pi}_{13}$ is another. Either of them leads to an estimator of γ_1 in eqn [1]. These are indirect least squares (ILS) estimators. They are not unbiased, but they are consistent if the LS reduced form estimators are consistent. The supply parameters β_1 and γ_1 are *identified*, meaning that the data and the model reveal their values (subject to sampling variation). Indeed, they are *overidentified*, because in small samples the two ways to estimate β_1 from $\widehat{\pi}'$s yield different results. (If either δ_2 and ε_2 were zero, there would be only one way, and the supply equation would be *just identified*. If both δ_2 and ε_2 were zero, there would be no way, and the supply equation would be *unidentified*.) Eqns [5]–[7] have an infinite number of solutions for the demand parameters β_2, γ_2, δ_2, and ε_2. Hence the data and the model do not reveal their values; they are unidentified.

Another way to see this is to imagine price and quantity data being generated by intersections of the supply and demand curves in the pq plane. Ignoring shocks, the supply curve is fixed, but the demand curve shifts when y or w changes. Hence the intersections reveal the slope and position of the supply curve as the demand curve shifts. But they reveal nothing about the slope of the demand curve.

The identifiability of parameters is crucial: they cannot be estimated if they are not identified. Reduced form parameters are usually identified, except in special cases. But the identifiability of structural parameters should be checked before one tries to estimate them. The LS formula can be applied to an unidentified structural equation such as eqn [2], but the result is not an estimator of that equation. For a more detailed and general discussion, *see* Statistical Identification and Estimability.

Simultaneous Equations Estimation Methods

Many methods have been developed for estimating identifiable parameters of simultaneous structural equations without the bias and inconsistency of LS. Most of them exploit the assumed independence between shocks and predetermined variables. Some estimate one equation at a time, and others estimate the whole system at once.

Estimating One Structural Equation at a Time

The ILS method mentioned above is one way of estimating one equation that is part of a system. A second way is the instrumental variables (IV) method (Durbin, 1954). LS and

IV will be shown for eqn [1]. Denote the deviation of each variable from its sample mean by an asterisk, for example, $p^* = p - \bar{p}$. Now rewrite eqn [1] in terms of deviations from sample means, thus eliminating the constant term γ_1; multiply it by p^*; sum it over all the sample observations; and divide the sum by $\sum p^{*2}$. The result is the LS estimator *of* β_1:

$$\widehat{\beta}_1 = \sum q^* p^* \Big/ \sum p^{*2} = \beta_1 + \sum u_1^* p^* \Big/ \sum p^{*2} \qquad [8]$$

It is biased and inconsistent because the shock u_1 influences p via eqn [3] so that the error term, the last term in eqn [8], does not equal zero in expectation or in probability limit.

Now rewrite eqn [1] as before in terms of deviations from sample means, but this time multiply it by y^*, sum it over all observations, and divide by $\sum p^* y^*$. The result is an IV estimator of β_1:

$$\widehat{\beta}_{1y} = \sum q^* y^* \Big/ \sum p^* y^* = \beta_1 + \sum u_1^* y^* \Big/ \sum p^* y^* \qquad [9]$$

If y is predetermined it is an *instrument*. In this case the IV estimator is consistent if u_1 has zero mean and constant variance and is serially independent, because then $\sum u_1^* y^*$ is zero in probability limit and $\sum p^* y^*$ is not. Similarly, if w is predetermined it is another instrument, and another IV estimator of β_1 is $\widehat{\beta}_{1w} = \sum q^* w^* / \sum p^* w^*$. Clearly, these IV estimators are equivalent to the ILS estimators of β_1 obtained earlier. For example, the one based on the reduced form coefficients of y is.

$$\begin{aligned} \widehat{\pi}_{22}/\widehat{\pi}_{12} &= \left(\sum q^* y^* \Big/ \sum y^{*2}\right) \Big/ \left(\sum p^* y^* \Big/ \sum y^{*2}\right) \\ &= \sum q^* y^* \Big/ \sum p^* y^* = \widehat{\beta}_{1y} \end{aligned} \qquad [10]$$

Similarly, the other one is $\widehat{\pi}_{23}/\widehat{\pi}_{13} = \widehat{\beta}_{1w}$. For a reduced form equation, IV is the same as LS, because the instruments are precisely the equation's predetermined variables.

The two-stage least squares (2SLS) method (Theil, 1972; Basmann, 1957) is another way of estimating one structural equation at a time. For an overidentified equation it is superior to ILS or IV because it results in one estimator rather than two or more; for a just identified equation it reduces to ILS and IV. It first computes the LS estimator of the reduced form equation for each endogenous variable that appears on the right side of the structural equation to be estimated (this is stage 1); it then replaces the observed data for those endogenous variables by the values calculated from the reduced form, and computes the LS estimator for the resulting equation (this is stage 2). For eqn [1], stage 1 is to compute the LS estimators of the π's in the price eqn [3] of the reduced form; the second stage is to compute $\widehat{p} = \widehat{\pi}_{11} + \widehat{\pi}_{12} y + \widehat{\pi}_{13} w$, substitute this $\widehat{p}$ for p in (1), and compute the LS estimator $\sum q^* \widehat{p}^* / \sum \widehat{p}^{*2}$, which is the 2SLS estimator of β_1. 2SLS is an IV estimator, its instruments are the observed values of the equation's predetermined variables and the reduced-form-calculated values of the endogenous variables from the equation's right side. 2SLS is a *generalized* method of moments (GMOM) estimator (Hansen, 1982). LS and IV are plain MOM estimators.

The k-class estimator (Nagar, 1959) includes as special cases LS, 2SLS, and the limited information maximum likelihood estimator (LIML) (Anderson and Rubin, 1949, 1950). LIML is similar to 2SLS (the two are the same for infinitely large samples), but it has largely been displaced by 2SLS because it requires iterative computations which 2SLS does not, and because it often has a larger variance and sometimes yields outlandish estimates (Kadane, 1971; Theil, 1972).

For an identified structural equation, ILS, IV, LIML, and 2SLS are consistent if the model's explanatory variables are predetermined and the shocks have zero means and constant variances and covariances and are independent across time. An advantage of these methods is that, unlike LS, if applied to an unidentified structural equation they fail to yield estimates.

The Bayesian method of moments (BMOM) method (Zellner, 1998) obtains estimators of linear reduced form and structural equations without making assumptions about the likelihood function of the data. It assumes that the posterior expectations of the shocks given the data are uncorrelated with the predetermined variables, and that the differences between actual and estimated shocks have a covariance matrix of a particular form. Zellner shows how to find optimum BMOM estimators for several different loss functions, including a *precision* loss function which is a weighted sum of squares and cross-products of errors. For this loss function, the optimum BMOM estimator of an identified structural equation belongs to the k-class; it turns out to be LS if the ratio of the sample size to the number of predetermined variables in the model is 2, and it approaches 2SLS in the limit as this ratio grows without limit.

When a structural equation is overidentified, restrictions are placed on the reduced form parameters (such as $\pi_{22}/\pi_{21} = \pi_{32}/\pi_{31}$ in the example, because both ratios must equal β_1), but LS estimators of the reduced form ignore this information. Better estimators of the reduced form, because they use this information, can then be obtained by solving the estimated structure (*see* section 'The Need for Estimation of Parameters').

Estimating a Complete System

If in an identified complete simultaneous equations system the shocks are normally distributed and other suitable conditions are satisfied, an asymptotically efficient but computationally complex method of estimating all the parameters of a complete system is the full information maximum likelihood (FIML) method (Koopmans, 1950; Durbin, 1988). It involves maximizing the joint likelihood function of all the data and parameters, and requires iterative computations. It has largely been displaced by the *three stage least squares* (3SLS) method (Zellner and Theil, 1962), which is also asymptotically efficient but is much easier to compute. 3SLS is an application of the GLS method. 3SLS gets the required information on the shocks' covariances from 2SLS estimates of the shocks; hence the name 3SLS.

Cross-Section vs. Time Series Studies

So far the treatment has concerned time series data, where the observations describe the same country, city, family, firm, etc., at successive periods of time. The same type of analysis is

appropriate for cross-section data, where the observations describe different countries, firms, etc., at a single point of time, with obvious modifications. The analysis has also been extended to panel data, that is, a time series of cross-sections.

The Decline of Simultaneous Equations Models

In recent years the econometric literature has paid diminishing attention to simultaneous equations models and their identifiability. Many modern textbooks give these topics little space, and that only near the end of the book (Davidson and MacKinnon, 1993). Perhaps one reason is that modern computers can handle the estimation of nonlinear models, for which identification criteria are much more complicated and for which lack of identifiability is a much rarer problem.

The Problem of Choosing and Testing a Model

Thus far it has been presumed that the model being estimated is an accurate representation of the real-world process that actually generates the data. This is far from certain; at best it is likely to be only approximately true. Any model has been chosen by someone, perhaps with the aid of economic theory about the maximization of profit or utility, or perhaps because it was suggested by previously observed data. But there is no guarantee that it contains the right variables, or that their endogeneity or exogeneity is correctly stated, or that the correct number of lagged variables has been chosen, or that its equations have the right mathematical form, or that the assumed distribution of its shocks is correct.

One can perform diagnostic tests to see whether the estimated model fits past data well, and whether its calculated past shocks have constant variance and are free of any obvious systematic behavior. But this does not assure that it will do well with future data. In my view, the most stringent and most important test of a model is to expose it to data that were not available when the model was formulated. If it does not describe these data well, it leaves something to be desired. The more new data it can describe well, the more confidence one can have in it. But at any point the best that can be said about a model is that it has done a good job of describing the data that are available thus far.

See also: Fiducial and Structural Statistical Inference; Instrumental Variables in Statistics and Econometrics; Linear Hypothesis; Statistical Identification and Estimability.

Bibliography

Anderson, T.W., Rubin, H., 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics 20, 46–63.

Anderson, T.W., Rubin, H., 1950. The asymptotic properties of estimates of parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics 21, 570–582.

Basmann, R.L., 1957. A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25, 77–83.

Christ, C.F. (Ed.), 1994. Simultaneous Equations Estimation. Edward Elgar, Aldershot, UK.

Cervero, R., Hansen, M., 2002. Induced travel demand and induced road investment: a simultaneous equation analysis. Journal of Transport Economics and Policy 36 (3), 469–490.

Davidson, R., MacKinnon, J.G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.

Durbin, J., 1954. Errors in variables. Revue of the International Statistical Institute 22, 23–32.

Durbin, J., 1988. Maximum likelihood estimation of the parameters of a system of simultaneous regression equations. Econometric Theory 4, 159–170.

Faggian, A., McCann, P., 2006. Human capital flows and regional knowledge assets: a simultaneous equation approach. Oxford Economic Papers 58 (3), 475–500.

Haavelmo, T., 1943. The statistical implications of a system of simultaneous equations. Econometrica 11, 1–12.

Haavelmo, T., 1944. The probability approach in econometrics. Econometrica 12 (Suppl.), 1–115.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hausman, J.A., 1983. Specification and estimation of simultaneous equation models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 1. North-Holland, Amsterdam, pp. 391–448.

Jeanty, P.W., Partridge, M., Irwin, E., 2010. Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. Regional Science and Urban Economics 40 (5), 343–352.

Kadane, J.B., 1971. Comparison of *k*-class estimators when the disturbances are small. Econometrica 39, 723–727.

Koopmans, T.C. (Ed.), 1950. Statistical inference in dynamic economic models. Cowles Commission Monograph, vol. 10. Wiley, New York.

Lucas Jr., R.E., 1976. Econometric policy evaluation: a critique. Carnegie-Rochester Conference Series on Public Policy 1, 19–46.

Moon, M.J., Bretschneiber, S., 2002. Does the perception of red tape constrain it innovativeness in organizations? Unexpected results from a simultaneous equation model and implications. Journal of Public Administration Research and Theory 12 (2), 273–292.

Nagar, A.L., 1959. The bias and moment matrix of the general *k*-class estimators of the parameters in simultaneous equations. Econometrica 27, 575–595.

Strotz, R.H., Wold, H.O.A., 1960. A triptych on causal chain systems. Econometrica 28, 417–463.

Theil, H., 1972. Principles of Econometrics. Wiley, New York.

Zellner, A., 1998. The finite sample properties of simultaneous equations' estimates and estimators Bayesian and non-Bayesian approaches. Journal of Econometrics 83, 185–212.

Zellner, A., Theil, H., 1962. Three stage least squares: simultaneous estimation of simultaneous equations. Econometrica 30, 54–78.

Single-Case Designs: Methodology

Christopher R Berghoff and John P Forsyth, University at Albany, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.P. Forsyth, C.G. Finlay, volume 21, pp. 14116–14121, © 2001, Elsevier Ltd.

Abstract

The central aim of the present article is to review (1) the conceptual basis and rationale guiding the use of single-case designs in experimental and applied contexts and (2) the advantages, disadvantages, and flexible application of the main varieties of single-case design strategies. In so doing, emphasis is placed on the value of a single-case design approach for scientists and practitioners alike, particularly as a means to meet the joint pragmatic goals of prediction and control, and hence scientific understanding and increased accountability.

Basic and more applied branches of behavioral science seek to predict and influence a range of behavioral phenomenon with precision, scope, and depth. In so doing, efforts are made to isolate and control for extraneous sources of variability (i.e., noise), so that prediction and/or influence may be achieved. Methodology is simply a way to address such analytic aims, and depending on the level of analysis one adopts, methods may be applied to address questions based on a collection of observations gleaned from several individuals in a population (i.e., group-based methods) or from each individual case. The latter of these approaches is the focus of single-case methodology (Hayes et al., 1999).

The purpose of this article is to provide a succinct description of a methodological approach well suited for work with single organisms, whether in basic research or applied contexts, and particularly where large N designs or inferential statistics are nonsensical and even inappropriate. This approach, referred to here as *single-case design* (formerly referred to as *single-subject design*) methodology includes several design options involving rigorous and intensive experimental analysis of a single unit, and often replications across several individual cases. No attempt will be made here to describe all available design options or statistical tests, as excellent resources are available on both subjects (see Barlow et al., 2008; Busk and Marascuilo, 1992; Hayes et al., 1999; Huitema, 1986; Iversen, 1991; Kazdin, 2011; Parsonson and Baer, 1992). Rather, the intent is to provide (1) an overview of single-case design methodology and the rationale guiding its use in experimental and applied contexts, (2) a succinct description of the main varieties of single-case designs, including basic guidelines for their application, and (3) to highlight recent trends in the use of such designs.

Single-Case Design: Overview and Rationale

At the core, all scientists ultimately deal with particulars, whether cells, atoms, microbes, or the behavior of organisms. It is from such particulars that the facts of science and generality of findings are derived (Sidman, 1960). This is true whether one aggregates a limited set of observations from a large sample of cases into groups that differ on one or more variables, or situations involving a few cases and a large number of observations of each over time (cf. Hilliard, 1993). The former is characteristic of hypothetico-deductive group design strategies that rely on inferential statistics to support scientific claims, whereas the latter idiographic and inductive approach is characteristic of single-case methodology. As with other research designs, single-case methodology is as much an approach to science as it is a set of methodological rules and procedures for answering scientific and practical questions. As an approach, single-case designs have several distinguishing features.

Subject Matter of Single-Case Designs

Perhaps the most notable feature of single-case methodology is its subject matter and sample characteristics, namely, the intensive study of the behavior of a single (i.e., $N = 1$) unit, such as a person, nonhuman animal, or a classroom (Kratochwill and Levin, 2010). This feature is also a source of great misunderstanding (Furedy, 1999), particularly with regard to the generality and scientific utility of research based on an N of 1. It should be stressed that single-case methodology gets its name not from sample size per se, but rather because the analysis focuses on the behavior of individual units (Perone, 1991), or what Gottman (1973) referred to as *N-of-one-at-a-time* designs. Skinner (1966) described it this way: "… instead of studying a thousand rats for one hour each, or a hundred rats for ten hours each, the investigator is likely to study one rat for a thousand hours" (p. 21); this approach served as the foundation for operant learning (Skinner, 1953, 1957) and, with the increasing popularity of behavior therapy, set the stage for a new approach to treatment development, treatment evaluation, and increased accountability (Barlow et al., 2008).

Though single-case research often involves more than one unit (Cooper et al., 2007; Shadish and Sullivan, 2011), the data are evaluated primarily on a case-by-case basis. The idiographic nature of this methodology is well suited for applied work where practitioners often attempt to influence and change problematic behavior(s) of individual clients (Hayes et al., 1999). Though not a requirement, the inclusion of more than one case helps establish the robustness of observed effects across differentially imposed conditions, individuals,

behaviors, settings, and/or time (Horner et al., 2005). Hence, direct or systematic replication serves to increase confidence in the observed effects and helps establish the reliability and generality of observed relations (Sidman, 1960).

Analytic Aims of Single-Case Designs

As with other experimental design strategies, the fundamental premise driving the use of single-case methodology is to demonstrate control (i.e., influence) via the manipulation of independent variables and analysis of the effects of such variables on behavior (dependent variables). This emphasis on systematic manipulation of independent variables distinguishes single-case designs from other small *N* research such as case reports and case studies, which are often lucidly descriptive, but less controlled and systematic. With few exceptions (see Perone, 1991), the general convention of single-case design is to change one variable at a time and to evaluate such effects on behavior. This *one-variable-at-a-time rule* positions the basic or applied researcher to make clear causal statements about the effects of independent variables compared to a previous state (e.g., naturally occurring behavior during baseline), to compare different independent variables across or within conditions, and only then to evaluate interaction effects produced by two or more independent variables together.

Indeed, single-case methodology is unique in that it typically involves an intensive and rigorous experimental analysis of the behavior (i.e., thinking, emotional responses, overt motor acts) across time, and an equally rigorous attempt to isolate and control extraneous sources of variability (e.g., maturation, history) that may confound and/or mask any effects caused by systematic changes in the independent variable(s) across conditions and time (Sidman, 1960). This point regarding the handling of extraneous sources of variability cannot be overstated. Iversen (1991) and others (Barlow et al., 2008; Sidman, 1960) have noted issues concerning how variability is handled in group design research and single-case methodology. With regard to such issues, Iverson put it this way:

> [in group design research] the fit between theory and observation often is evaluated by means of statistical tests, and variability in data is considered a nuisance rather than an inspiration. By averaging over *N* subjects the variability among the subjects is ignored *N* times. In other words, a data analysis restricted to only the mean generates *N* degrees of ignorance. Single-subject designs are used because the data based on the individual subject are more accurate predictors of that individual's behavior than are data based on an averaging for a group of subjects. (p. 194)

Though the issue of when and how to aggregate data (if at all) is controversial in single-case research (see section Analyzing Outcome Data for a discussion), there is a consensus that variability is to be pinpointed and controlled to the extent possible. This is consistent with the view that variability is imposed and determined from without, and not an intrinsic property of the subject matter under study (Sidman, 1960). Efforts are made deliberately to control such noise so that prediction and control can be achieved with precision and scope.

Measurement Issues of Single-Case Design

Single-case design methodology can also be distinguished by the frequency with which behavior is sampled within conditions and across differentially imposed conditions and time. It is common, for instance, for basic and applied researchers to include hundreds of samples of behavior across time (i.e., hours, days, weeks, sessions) with individual subjects. Yet, a minimum of at least three data points is required to establish level, trend, and variability within a given phase or design element (Hayes et al., 1999). Changes in level (i.e., the mean or median within a condition), trend (i.e., the slope of change), and at times variability (i.e., fluctuations around either the level or trend) from one condition to the next, and particularly changes that are robust, clear, immediate, and stable from behavior observed in a prior condition, help support causal inferences.

Here, stability is a relative term and should not be confused with something that is static or flat. Rather, stability provides a background by which to evaluate the reliability of changes produced by an independent variable across conditions and time. By convention, each condition is often continued until some semblance of stability is observed, at which point another condition is imposed, and behavior is evaluated relative to the previous condition or steady state (cf. Perone, 1991). Such changes, in turn, are continuously evaluated relative to a prior less controlled state (i.e., baseline) or adjacent conditions involving another independent variable. In single-case methodology, such stability can be evaluated within a condition (i.e., from one sample or observation to the next) and across conditions (i.e., changes in level, trend, variability). The pragmatic appeal of such an approach, particularly in applied contexts, rests in allowing the practitioner to evaluate the effectiveness of imposed interventions in real time, and consequently to modify their intervention tactics as the data dictate. Indeed, unlike more formal group-designs that are followed strictly once set in place, single-case methodology is more flexible. It is common for design elements to be added, dropped, and/or modified as the data dictate so as to meet scientific and applied goals. This feature has obvious parallels with how practitioners work with their clients in designing treatment interventions (cf. Hayes et al., 1999).

In sum, all single-case designs focus on individual organisms and aim to predict and influence behavior via (1) repeated sampling and observation, (2) manipulation of one or more independent variables while isolating and controlling sources of extraneous variability to the extent possible, and (3) demonstration of stability within and across levels of imposed independent variables (cf. Perone, 1991). Discussion will now turn to an enumeration of such features in the context of more popular single-case designs.

Varieties of Single-Case Methodology

Most single-case designs can be classified generally as representing two main types: within and between-series designs. Though each will be discussed briefly in turn, it is important to recognize that neither class of design precludes combining elements from the other (e.g., combined-series elements). In

other words, a within-series design may, for some purposes, lead the investigator to add between-series elements and vise versa, including other more complex elements (e.g., interaction effects, changing criterion elements).

Within-Series Designs

The basic logic and structure of within-series designs are simple; namely, to evaluate changes within a series of data points across time on a single measure or set of related measures (cf. Hayes et al., 1999). Stability is judged for each data point relative to other data points that immediately precede and follow it. By convention, such designs include comparisons between a naturally occurring state of behavior (typically referred to as the baseline and denoted by the letter A) and the effects of an imposed manipulation of an independent variable or intervention (denoted by different letters such as B and C). A *simple A–B design* involves sampling naturally occurring behavior (A phase), followed by repeated assessment of responding when the independent variable is introduced (B phase). For example, suppose a researcher wanted to determine the effects of rational-emotive therapy on initiating social interactions for a particular client. If an A–B design were chosen, the rate of initiations prior to treatment would be compared to that following treatment in a manner analogous to popular pre-to-post comparisons in group-outcome research. It should be noted, however, that simple A–B designs are inherently weak in controlling for threats to internal validity (e.g., testing, maturation, history, regression to the mean). *Withdrawal/reversal designs* control for such threats, and hence provide a more convincing case of experimental control.

Withdrawal designs represent a replication of the basic A–B sequence a second time, with the term withdrawal representing the imposed removal of the active treatment or independent variable (i.e., a return to second baseline or A phase). For example, a simple A–B–A sequence allows one to evaluate treatment effects relative to baseline responding. Effects due to treatment should diminish once treatment is withdrawn and the subject or client is returned to an A phase. Other variations on withdrawal designs include B–A–B designs and A–B–A–B–A–B (reversal) designs (see Barlow et al., 2008; Hayes et al., 1999). The A–B–A–B–A–B design provides at least one, or a series, of treatment reinstatement phases (i.e., a return to a second B phase) following withdrawal. Thus, treatment effects should be once again observed after such a switch in phase, providing within-unit replication and contributing to the reliability of the observed effect.

A–B designs and withdrawal/reversal designs are typically used to compare the effects of a finite set of treatment variables to baseline response levels. Data regarding stable response trends are collected across several discrete periods (e.g., time, sessions), wherein the independent variable is either absent (baseline) or present (treatment phase). Phase shifts are data-driven, and response stability determines the next element added to the design; elements that may include other manipulations or treatments either alone (e.g., A–B–A–C–A–B–A) or in combination (e.g., A–B–A–B + C–A–B–A). As the manipulated behavior change is repeatedly demonstrated and replicated across increasing numbers of phase shifts and time, confidence in the role of the independent variable as a cause of such changes increases.

This basic logic of within-series single-case methodology has been expanded in sophisticated and at times complex ways to meet basic and applied purposes. For instance, such designs can be used to test the differential effects of more than one treatment. Other reversal designs (i.e., B–C–B–C) involve the comparison of differential, yet consecutive, treatment interventions across multiple phase shifts. These designs are similar to the above reversals in how control is evaluated, but differ primarily in that baseline phases are not required. Designs are also available that combine features of withdrawal and reversal. For example, A–B–A–B–C–B designs allow one to compare A and B phases to each other (reversal; A–B–A–B), and a second treatment to B (reversal; e.g., B–C–B), or even to evaluate the extent to which behavior tracks a specified behavioral criterion (i.e., changing criterion designs; see Hayes et al., 1999).

Changing criterion designs provides an alternative method for experimentally analyzing behavioral effects without subsequent treatment withdrawal. Criterion is set such that optimal levels given exposure can be met and then systematically increased (or decreased) to instill greater demand on acquiring new repertoires. For example, a child learning to add may be required to calculate four of 10 problems correctly to earn a prize. Once the child successfully and consistently demonstrates this level of responding, the demand increases to 6 of 10 correct problems. Changing criterion designs serves as a medium to demonstrate learning through achieving successive approximations of the end state, and may be one strategy employed when a return to the baseline phase is unlikely (e.g., once acquired, mathematical skills or other behavioral skills are unlikely to revert to a previous baseline state) or when ethical concerns suggest that a return to baseline may be harmful (e.g., reinstating the self-injurious behavior of a child with an autism spectrum disorder).

Between-Series Designs

Between-series designs differ primarily from within-series designs in that data are first grouped and compared by condition, and then by time, whereas the reverse is true for within-series designs. Between-series designs need not contain phases, as evaluation of level, trend, and stability is organized by condition first, but not by time alone (Hayes et al., 1999, p. 176). Designs within this category include *alternating treatment design* and *simultaneous treatment design*. Both allow for the evaluation of multiple (with a minimum of two) treatments concurrently in an individual case.

Alternating treatment designs contain no phases, unlike the previously described within-series designs (Hayes et al., 1999). Rather, concurrent evaluation is made between rapid and largely random alternations of two or more conditions (Barlow and Hayes, 1979). These designs are particularly useful in cases where it would be impractical or unethical to withdraw treatment and in cases where establishing a baseline is unnecessary. Additionally, alternating treatment designs fit well with applied work, as therapists must often routinely target multiple problems concurrently and thus

need to switch rapidly between intervention tactics in the context of therapy.

As with the alternating treatment design, simultaneous treatment designs do not contain phases. In addition, rapid or random alteration of treatment is not required with simultaneous treatment design. What is necessary is that two or more conditions are available simultaneously, with the subject choosing between them. Thus, the simultaneous treatment design is appropriate for situations where one wishes to evaluate the concurrent or simultaneous application of two or more treatments in a single case. While not conducive for evaluating treatment outcome, this particular design is appropriate for evaluating preference or choice (cf. Hayes et al., 1999). For example, multiple reward options could be presented simultaneously (Kratochwill and Levin, 2010). The simultaneous treatment design would not be appropriate for examining the effectiveness of each reward on behavior change, but could be used to evaluate preference for each reward.

Combining Within- and Between-Series Elements

Multiple-baseline designs build upon and integrate the basic logic and structure of within- and between-series elements. The fundamental premise of multiple-baseline designs is to replicate phase change effects systematically in more than one series, with each subsequent uninterrupted series serving as a control condition for the preceding interrupted series. Series can be compared and arranged across behaviors, settings, individuals, or some combination of these (cf. Hayes et al., 1999). Such designs require that the series under consideration be independent (e.g., two functionally distinct behaviors, settings), and that the intervention be administered sequentially beginning with the first series, while others are left uninterrupted as controls. For example, a client might present with three distinct behavior problems all requiring exposure therapy. All behaviors would be monitored during baseline (A phase), and after some semblance of stability is reached the treatment (B phase) would be applied to the first behavior series, while the remaining two behaviors are continuously monitored in an extended baseline. Once changes in the first behavior resulting from treatment achieve stability, the second series would be interrupted and treatment applied, while the first behavior continues in the B phase and the third behavior is monitored in an extended baseline. The procedure followed for the first two series elements is then repeated for the third behavior. This logic can be similarly applied to multiple-baseline designs across individuals or settings (see Barlow et al., 2008; Hall et al., 1970). More complex within-series elements (e.g., A–B–A–B, or B–C–B, or counterbalanced phases such as B–A–B and A–B–A) can be evaluated across behaviors, settings, and individuals.

Note that with multiple-baseline designs the treatment is never withdrawn; rather, it is introduced systematically across a set of dependent variables. Independent variable effects are denoted from how reliably behavior change correlates with the onset of treatment for a particular dependent variable. For such reasons, multiple-baseline designs are quite popular, owing much to their ease of use, strength in ruling out threats to internal validity, built-in replication, and fit with the demands of applied practitioners working in therapy, schools, and institutions. Indeed, such designs can be useful when therapists are working with several clients presenting with similar problems, when clients begin therapy at different times, or in cases where targets for change in therapy occur sequentially.

Recent Trends in the Application of Single-Case Methods

Over 30 journals within psychology and education have recently published single-case design studies (Shadish and Sullivan, 2011) – an encouraging trend compared to the use of such designs over a decade ago (see Forsyth et al., 1999). Yet, single-case methods remain controversial owing, in part, to disagreements regarding proper methodology for examining single-case data, criticisms that single-case designs lack strong methodologies (i.e., lack of randomization), and the need for accountability in applied practice. It is to a discussion of such issues that we now turn.

Analyzing Outcome Data

Historically, visual inspection of plotted data has been the primary method of analyzing single-case data (Parsonson and Baer, 1992), and it was assumed that the researcher-practitioner's focus on control of extraneous causes of error would lead to more Type II errors (i.e., to deny a difference and claim that a variable is not exerting influence over behavior, when it is) compared to more classic Type I errors (i.e., to claim that a variable is exerting influence over behavior, when it is not; cf. Baer, 1977). However, early research suggested that visual inference is unreliable. This is particularly true of datasets based on short series, with a good deal of variability, and autocorrelation (i.e., serial dependence, wherein one observation depends in part on the previous observation; Borckardt et al., 2008), such as that often found in published literature and applied settings (DeProspero and Cohen, 1979; Jones et al., 1978). These and other factors tend to result in inflated Type I errors (Matyas and Greenwood, 1990).

Given possible problems with visual inspection techniques, several calls were made to apply various inferential/randomization statistical approaches to single-case generated data. Though in-depth discussion of these newer methods are beyond the scope of this paper (for reviews see Haardörfer and Gagné, 2010; Manolov et al., 2010), we nonetheless wish to highlight some core concerns emerging from this line of work. First, statistical methods rely on randomization techniques, whereas single-case designs do not (e.g., Haardörfer and Gagné, 2010). Second, such techniques are also impacted by autocorrelation of the data, often resulting in low power and inflated Type I errors (e.g., Borckardt et al., 2008; Haardörfer and Gagné, 2010), particularly when examining short data series (e.g., Manolov et al., 2010). Third, the complexity of some of the proposed analyses may be cumbersome and expensive for practitioners to implement in applied settings. Finally, work on visual inspection techniques suggests that both inter-rater reliability (Hagopian et al., 1997) and accuracy (Fisher et al.,

2003) can be improved through the implementation of training and adherence to structured visual/graphical analysis techniques and correspondence with expert visual analysts. Such improvements in visual inference, coupled with their acceptability, ease of use, and fit with applied practice, make visual/graphical techniques the preferred method when examining single-case data (Haardörfer and Gagné, 2010; Wolery et al., 2010).

Improving Validity

As previously noted, replication of effects across individuals, settings, behaviors, etc. is the traditional strategy used to establish the reliability and generality of inferences gleaned from single-case research designs. Group-based designs, by contrast, tend to employ randomization techniques to improve internal validity. However, randomization need not be exclusive to the domain of hypothetico-deductive group designs. Kratochwill and Levin (2010) have recently provided an excellent introduction to combining both simple and block randomization procedures using single-case methodology. For example, they illustrate implementation of *randomized phase-order designs*, where the sequencing of the A and B phases are randomized using either simple or block randomization for each of the designs presented above. *Randomized phase start-point designs* alter the commencement of each phase, again using either simple or block randomization techniques. In addition, the authors extend both randomized order and start point designs to include replication across units (e.g., multiple individuals, classrooms) and to provide predetermined baseline phases when required.

Instituting randomization techniques will improve the internal validity of single-case designs. However, these advantages come with certain costs (Kratochwill and Levin, 2010). Additional phase changes may be needed to meet desired withdrawal and reversal observations, resulting in increases to time and cost in research settings. Also, perhaps one of the strongest features of single-case designs, particularly in applied settings, is the flexibility offered in allowing the data to guide phase changes and identify new independent variables for manipulation. However, predetermined a priori randomized schedules eliminate such flexibility and may impact the ability to achieve stable behavior levels within a phase. In spite of these limitations, we expect this powerful methodological addition to the scientist-practitioner's arsenal to support and increase the validity of future single-case work.

Accountability and Managed Care

Increasingly, third-party payers are requiring practitioners to demonstrate accountability, that is, to show that what they are doing with their clients is achieving the desired effect (i.e., good outcomes) in a cost-effective and lasting manner. Use of single-case methodology has a place in assisting practitioners in making clinical decisions based on empirical data, monitoring progress and effectiveness of treatment, and in demonstrating accountability to third-party payers and consumers. Further, such methodology, though requiring time and sufficient training to implement properly, is atheoretical and fits nicely with how practitioners from a variety of conceptual schools work routinely with their individual clients (Hayes et al., 1999). Also, single-case designs can be implemented as a method of quickly informing consultants, supervisors, or a treatment team as to the progress and effectiveness of a given intervention (Segool et al., 2007). Thus, there is great potential in single-case methodology for bridging the strained scientist-practitioner gap, and ultimately advancing the science of behavior change and an empirically driven approach to practice and treatment innovation.

Summary and Conclusions

Methodology is a way of knowing. Single-case methodology represents a unique way of knowing that has as its subject matter an experimental analysis of the behavior of individual units. Described here were the assumptions driving this approach and the main varieties of design options available to address basic experimental and applied questions. Perhaps the greatest asset of single-case methodology rests with the flexibility with which such designs can be constructed to meet the joint analytic goals of prediction and control over behavior of individuals. This asset also represents one of the greatest liabilities of such designs in that flexibility requires knowledge of when and how to modify and/or add design elements to achieve analytic goals, and particularly skill in recognizing significant effects, sources, or variability, and when one has demonstrated sufficient levels of prediction and control over the behavior of interest.

See also: Evidence-Based Psychotherapy; Managed Care and Psychotherapy in the United States; Psychological Treatment, Effectiveness of; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Process Research; Training in Clinical Psychology in the United States: Scientist–Practitioner Model.

Bibliography

Baer, D.M., 1977. Perhaps it would be better not to know everything. Journal of Applied Behavior Analysis 10, 167–172. http://dx.doi.org/10.1901/jaba.1977.10-167.

Barlow, D.H., Hayes, S.C., 1979. Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject. Journal of Applied Behavior Analysis 12, 199–210. http://dx.doi.org/10.1901/jaba.1979.12-199.

Barlow, D.H., Nock, M.K., Hersen, M., 2008. Single-case Experimental Designs: Strategies for Studying Behavior Change, third ed. Allyn & Bacon, Boston, MA.

Borckardt, J.J., Nash, M.R., Murphy, M.D., Moore, M., Shaw, D., O'Neil, P., 2008. Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. American Psychologist 63, 77–95. http://dx.doi.org/10.1037/0003-066X.63.2.77.

Busk, P.L., Marascuilo, L.A., 1992. Statistical analysis in single-case research: issues, procedures, and recommendations, with applications to multiple behaviors. In: Kratochwill, T.R., Levin, J.R. (Eds.), Single-Case Research Design and Analysis. Plenum, New York, pp. 159–185.

Cooper, J.O., Heron, T.E., Heward, W.L., 2007. Applied Behavior Analysis, second ed. Pearson/Merrill Prentice Hall, Upper Saddle River, NJ.

DeProspero, A., Cohen, S., 1979. Inconsistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis 15, 415–421. http://dx.doi.org/10.1901/jaba.1979.12-573.

Fisher, W.W., Kelly, M.E., Lomas, J.E., 2003. Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. Journal of

Applied Behavior Analysis 36, 387–406. http://dx.doi.org/10.1901/jaba.2003.36-387.

Forsyth, J.P., Kollins, S., Palav, A., Duff, K., Maher, S., 1999. Has behavior therapy drifted from its experimental roots? A survey of publication trends in mainstream behavioral journals. Journal of Behavior Therapy and Experimental Psychiatry 30, 205–220. http://dx.doi.org/10.1016/S0005-7916(99)00020-8.

Furedy, J.J., 1999. Commentary: on the limited role of the single-subject design in psychology: hypothesis generating but not testing. Journal of Behavior Therapy and Experimental Psychiatry 30, 21–22. http://dx.doi.org/10.1016/S0005-7916(99)00006-3.

Gottman, J.M., 1973. N-of-one and N-of-two research in psychotherapy. Psychological Bulletin 80, 93–105. http://dx.doi.org/10.1037/h0034803.

Haardörfer, R., Gagné, P., 2010. The use of randomization tests in single-subject research. Focus on Autism and Other Developmental Disabilities 25, 47–54. http://dx.doi.org/10.1177/1088357609350745.

Hagopian, L.P., Fisher, W.W., Thompson, R.H., Owen-DeSchryver, J., Iwata, B.A., Wacker, D.P., 1997. Toward the development of structured criteria for interpretation of functional analysis data. Journal of Applied Behavior Analysis 30, 313–326. http://dx.doi.org/10.1901/jaba.1997.30-313.

Hall, R.V., Cristler, C., Cranston, S.S., Tucker, B., 1970. Teachers and parents as researchers using multiple baseline designs. Journal of Applied Behavior Analysis 3, 247–255. http://dx.doi.org/10.1901/jaba.1970.3-247.

Hayes, S.C., Barlow, D.H., Nelson-Gray, R.O., 1999. The Scientist-practitioner: Research and Accountability in the Age of Managed Care. Allyn & Bacon, Boston, MA.

Hilliard, R.B., 1993. Single-case methodology in psychotherapy process and outcome research. Journal of Consulting and Clinical Psychology 61, 373–380. http://dx.doi.org/10.1037/0022-006X.61.3.373.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S., Wolery, M., 2005. The use of single-subject research to identify evidence-based practice in special education. Exceptional Children 71, 165–179.

Huitema, B.E., 1986. Statistical analysis and single-subject designs: some misunderstandings. In: Poling, A., Fuqua, R.W. (Eds.), Research Methods in Applied Behavior Analysis. Plenum, New York, pp. 209–232.

Iversen, I.H., 1991. Methods of analyzing behavior patterns. In: Iversen, I.H., Lattal, K.A. (Eds.), Experimental Analysis of Behavior, Part 2. Elsevier, New York, pp. 193–241.

Jones, R.R., Weinrott, M.R., Vaught, R.S., 1978. Effects of serial dependency on the agreement between visual and statistical inference. Journal of Applied Behavior Analysis 11, 277–283. http://dx.doi.org/10.1901/jaba.1978.11-277.

Kazdin, A.E., 2011. Single-Case Research Designs: Methods for Clinical and Applied Settings, second ed. Oxford University Press, New York.

Kratochwill, T.R., Levin, J.R., 2010. Enhancing the scientific credibility of single-case intervention research: randomization to the rescue. Psychological Methods 2, 124–144. http://dx.doi.org/10.1037/a0017736.

Manolov, R., Solanas, A., Bulté, I., Onghena, P., 2010. Data-division-specific robustness and power of randomization tests for ABAB designs. The Journal of Experimental Education 78, 191–214. http://dx.doi.org/10.1080/00220970903292827.

Matyas, T.A., Greenwood, K.M., 1990. Visual analysis of single-case time series: effects of variability, serial dependence, and magnitude of intervention effects. Journal of Applied Behavior Analysis 23, 241–351. http://dx.doi.org/10.1901/jaba.1990.23-341.

Parsonson, B.S., Baer, D.M., 1992. The visual analysis of data, and current research into the stimuli controlling it. In: Kratochwill, T.R., Levin, J.R. (Eds.), Single-Case Research Design and Analysis. Plenum, New York, pp. 15–41.

Perone, M., 1991. Experimental design in the analysis of free-operant behavior. In: Iversen, I.H., Lattal, K.A. (Eds.), Experimental Analysis of Behavior, Part 1. Elsevier, New York, pp. 135–171.

Segool, N.K., Brinkman, T.M., Carlson, J.S., 2007. Enhancing accountability in behavioral consultation through the use of single-case designs. International Journal of Behavioral Consultation and Therapy 3, 310–321.

Shadish, W.R., Sullivan, K.J., 2011. Characteristics of single-case designs used to assess intervention effects in 2008. Behavior Research Methods 43, 971–980. http://dx.doi.org/10.3758/s13428-011-0111-y.

Sidman, M., 1960. Tactics of Scientific Research: Evaluating Experimental Data in Psychology. Basic Books, New York.

Skinner, B.F., 1953. Science and Human Behavior. Macmillan, New York.

Skinner, B.F., 1957. The experimental analysis of behavior. American Scientist 45, 343–371.

Skinner, B.F., 1966. Operant behavior. In: Honig, W.K. (Ed.), Operant Behavior: Areas of Research and Application. Appleton-Century-Crofts, New York, pp. 12–32.

Wolery, M., Busick, M., Reichow, B., Barton, E.E., 2010. Comparison of overlap methods for quantitatively synthesizing single-subject data. Journal of Special Education 44, 18–28. http://dx.doi.org/10.1177/0022466908328009.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 22

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 22

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 22 *xliii*

VOLUME 22

Situated Knowledge, Feminist and Science and Technology Studies Perspectives 1
Charis M Thompson

Situated Learning (Learning In Situ) 5
Sue Waite and Nick Pratt

Situational Judgment Test 13
Filip Lievens and Britt De Soete

Skills Training: Public Policy Aspects 20
Moira Nelson

Skinner, Burrhus Frederick (1904–90) 25
Marc N Richelle

Skocpol, Theda (1947–) 31
Ian Sheinheit

Slavery and Violence 37
Enrico Dal Lago

Slavery as Social Institution 43
Kevin Bales and Monti N Datta

Slavery: Comparative Aspects 49
Orlando Patterson

Slaves and Slavery, History of 54
Trevor Burnard

Sleep and Health 59
Angelika Schlarb

Sleep and Memory 63
Médhi Gilson and Philippe Peigneux

Sleep Disorders across the Life Span 70
Weidong Wang, Xue Yan, and YanJiao Liu

Sleep Disorders: Psychiatric Aspects 79
Johannes Beck, Erich Seifritz, and Edith Holsboer-Trachsler

Sleep: Neural Systems 87
Michael C Antle

Slow Food as a Social Movement 94
Koen van Bommel and André Spicer

Small Group Research, History of 100
Christian Dayé

Small States, History of 108
Richard Whatmore

Smart Cities 113
Andrea Caragliu, Chiara D Bo, Karima Kourtit, and Peter Nijkamp

Smith, Adam (1723–90) 118
Dominique Bouchet

Smoking and Health 125
Kenneth E Warner

Social Action and Social Justice 132
Sarah Todd

Social and Behavioral Sciences under Dictatorship 139
Olessia Kirtchik and Mariana Heredia

Social and Economic Change: Psychological Challenges for Individuals 147
Rainer K Silbereisen and Martin J Tomasik

Social and Emotional Development in the Context of the Family 156
Rena L Repetti, Meredith S Sears, and Sunhye Bai

Social Beliefs and Attitudes: Personality and the Isms 162
Jeremy Burrus and Lauren Carney

Social Business 168
Enrique Ogliastri, Andrea Prado, Urs Jäger, Antonio Vives, and Ezequiel Reficco

Social Capital 174
Paul F Whiteley

Social Capital in Organizations 181
Lorenzo Bizzi

Social Categorization 186
Craig McGarty, Kenneth I Mavor, and Daniel P Skorich

Social Changes: Models 192
Gudmund Hernes and Kenneth C Land

Social Classification 199
Scott L Washington

Social Cognition 204
François Ric

Social Comparison, Psychology of 210
Ladd Wheeler and Jerry Suls

Social Competence During Adolescence Across Cultures 216
Grace Iarocci and Emily Gardiner

Social Constructionism 222
Viv Burr

Social Constructivism 228
Wolfgang Detel

Social Control and Deviance, History of 235
Elizabeth E Mustaine

Social Democracy 242
Tom P Harsanyi

Social Dilemmas, Psychology of 245
Eric van Dijk

Social Dominance Orientation 250
Andrew L Stewart and Felicia Pratto

Social Ecology 254
Marina Fischer-Kowalski

Social Ecology: The Chicago School 263
Anna-Lisa Müller

Social Enquiry and Action Research for Social Work 269
Purnima George and Danika Syrja-McNally

Social Epidemiology 275
Louis Chauvel and Anja K Leist

Social Epigenetics 282
Douglas L Anderton

Social Exclusion, Social Deprivation and Health 285
Tarani Chandola and Richard Conibere

Social Experiments 291
Larry L Orr and Rebecca A Maynard

Social Geography 301
Sophie Bowlby

Social History 307
Christoph Conrad

Social Identity in Social Psychology 313
Katherine J Reynolds

Social Identity in Sociology 319
Kay Deaux

Social Inequality and Schooling 325
Marie Duru-Bellat

Social Inequality in Cultural Consumption Patterns 331
Koen van Eijck

Social Inequality in History (Stratification and Classes) 338
P Nolte

Social Influence, Social Psychology of 345
Y Ito

Social Influence, Sociology of 348
Robert H Gass

Social Innovation 355
Alex Nicholls and J Gregory Dees

Social Insurance: Legal Aspects 362
Gerhard Igl

Social Integration, Social Networks, and Health 365
Julianne Holt-Lunstad

Social Intelligence and Competencies 371
Susanne Weis and Kristin Conzelmann

Social Intervention in Psychiatric Disorders 380
Wulf Rössler

Social Justice and Criminal Justice 386
Kimberly L Barrett and Michael J Lynch

Social Justice and Social Inequalities 392
Gregg Barak

Social Justice: Historical and Theoretical Considerations 397
Kathleen M Weigert

Social Learning and Culture in Nonhuman Organisms 401
Rachel L Kendal

Social Learning Theory and Crime 409
Jeffrey T Ward and Chelsea N Brown

Social Mechanism 415
Petri Ylikoski

Social Media 421
Bernie Hogan and Joshua Melville

Social Mobility, History of 426
Hartmut Kaelble

Social Movements and Political Violence 430
Lorenzo Bosi and Chares Demetriou

Social Movements, Geography of 437
Byron Miller

Social Movements, History of: General 442
L Tilly

Social Movements: A Social Psychological Perspective 447
John Drury

Social Network Algorithms and Software 454
Ulrik Brandes

Social Network Analysis 461
Michael Steketee, Atsushi Miyaoka, and Maura Spiegelman

Social Network Models: Statistical 468
Philippa Pattison and Stanley Wasserman

Social Networks 473
John Scott and Frans N Stokman

Social Neuroscience of the Twenty-First Century 485
Elizabeth A Necka, Stephanie Cacioppo, and John T Cacioppo

Social Networks, Algebra of 478
Philippa Pattison, Lucia Falzon, and Nectarios Kontoleon

Social Ontology 489
John Latsis

Social Phobia across the Lifespan 493
Stefan G Hofmann, Hannah Boettcher, and Jade Q Wu

Social Properties (Facts and Entities): Philosophical Aspects 500
David-Hillel Ruben

Social Protest 506
Abby Peterson

Social Protest and New Media 512
Sandra González-Bailón

Social Protest Movements and Social Work Practice 518
Carolyn Noble

Social Psychological Theory, History of 525
Peter Hegarty, Sebastian E Bartos, and Katherine Hubbard

Social Psychology 532
Xenia Chryssochoou

Social Psychology, Sociological 538
David Rohall

Social Psychology: Research Methods 544
Lia Figgou and Vassilis Pavlopoulos

Social Question: Impact on Social Thought 553
Giovanna Procacci

Social Referencing during Infancy and Early Childhood across Cultures 556
Christine Fawcett and Ulf Liszkowski

Social Relationships in Adulthood 563
Margaret S Clark, Oriana R Aragon, and Jennifer L Hirsch

Social Resilience in Post-Tsunami Japan 570
Jin Sato

Social Science and Civil Society 576
Eddy Hogg

Social Science and the Media 583
Andrea Römmele and Barbara Pfetsch

Social Science and Universities 588
Sheldon Rothblatt

Social Science Infrastructure: North America (Research and Teaching) 596
Gisèle Yasmeen

Social Science Infrastructure: Africa 604
Freek Cronjé, Johann van Wyk, Jean D Losango, and Samuel Lule

Social Science Infrastructure: Asia 615
Victor R Savage

Social Science Infrastructure: Central and South America 621
Cecilia Mabragaña and Maria Beatriz Gavilán

Social Science Infrastructure: East Asia and the Pacific (Research and Teaching) 631
Takashi Inoguchi

Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations) 637
Gordon CK Cheung

Social Science Infrastructure: European Union (Resourcing and Funding) 643
Thomas König and Helga Nowotny

Social Science Infrastructure: European Union (Social Science Academies and Related Organizations) 648
Thibaud Boncourt

Social Science Infrastructure: North America (Resourcing and Funding) 655
Helmut K Anheier

Social Science Infrastructure: North America (Social Science Academies and Related Organizations) 661
Steven Rathgeb Smith

Social Science Infrastructure: European Union (Research and Teaching) 667
Friedhelm Neidhardt and Georg Thurn

Social Science Learned Societies and Professional Associations 674
Jennifer Platt

Social Science Professions and Professionalization 679
Felice J Levine and Nathan E Bell

Social Science, The Idea of 688
Fred D'Agostino

Social Scientists as Experts and Public Intellectuals 695
Stephen P Turner

Social Simulation: Computational Models 701
Michael W Macy

Social Stratification 706
David B Grusky

Social Structure 713
John L Martin and Monica Lee

Social Structure: History of the Concept 719
Charles Crothers

Social Studies Education 727
Antonio J Castro and Ryan T Knowles

Social Support and Recovery from Disease and Medical Procedures 735
Melika H Stephens and Keith J Petrie

Social System 741
Milan Zafirovski

Social Work and Disasters 753
Michael Cronin and David N Jones

Social Work and Psychodynamic Thought 761
Jerrold R Brandell and Laura J George

Social Work and Refugees 769
Mary Nash

Social Work Education 775
Malcolm Payne

Social Work Ethics 782
Sarah Banks

Social Work in China 789
Angelina WK Yuen-Tsang and Wang Si-bin

Social Work in Eastern Europe 795
Darja Zaviršek

Social Work in Organizations 801
Mark Lymbery

Social Work in Russia 807
Elena Iarskaia-Smirnova and Valentina Yarskaya

Social Work Theory 813
Barbra Teater

Social Work with Racialized Groups: Frameworks for Practice 821
Wanda T Bernard and Ellice Daniel

Social, The: History of the Concept 827
Jean Terrier

Socialism 833
Paul Thomas

Socialism: Historical Aspects 836
Logie Barrow

Socialist Law 844
András Sajó and Mónika Ganczer

Socialist Societies: Anthropological Aspects 849
Katherine Verdery

Sociality In Anthropology 854
Nicholas J Long

Socialization and Education: Theoretical Perspectives 861
Klaus A Schneewind

Socialization in Infancy and Childhood 866
Melinda S Leidy and Ross D Parke

Societal Impact Assessment 873
Reinhard Kreissl, Florian Fritz, and Lars Ostermeier

Societies, Types of 878
Guillermina Jasso

Society and People: History of the Concept 887
Lars Trägårdh

Sociobiology and Sociology: A New Synthesis 892
Richard Machalek and Michael W Martin

Sociobiology, History of 899
Ullica Segerstrale

Sociobiology: Overview 908
Martin Daly

Sociocybernetics 912
Bernd R Hornung

Sociodigital Revolution: Digital Natives vs Digital Immigrants 918
Kai Hakkarainen, Lauri Hietajärvi, Kimmo Alho, Kirsti Lonka, and Katariina Salmela-Aro

Socioeconomic Status and Academic Achievement 924
Kevin KH Chung

Socioeconomic Status and Health 931
Bjørgulf Claussen

Socioeconomic Status and Health Care 937
Marsha Regenstein and Paula M Lantz

Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review 942
Beverly Gaudet and Nicole Letourneau

Sociolinguistics 949
Gerard Van Herk

Sociological Theory 954
Jonathan H Turner

Sociological Traditions 965
Charles Crothers

Sociology and Psychiatry 974
Dirk Richter and Heinz Katschnig

Sociology and the New Institutionalism 979
Victor Nee and Sonja Opper

Sociology, Epistemology of 984
Mario Bunge

Sociology, History of 989
Johan Goudsblom and Johan Heilbron

Sociology: Overview 996
Mitch Duneier

CONTRIBUTORS TO VOLUME 22

Kimmo Alho
Helsinki Collegium for Advanced Studies and Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Douglas L. Anderton
University of South Carolina, Columbia, SC, USA

Helmut K. Anheier
Hertie School of Governance, Berlin, Germany

Michael C. Antle
Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

Oriana R. Aragon
Center for Scientific Teaching, Yale University, New Haven, CT, USA

Sunhye Bai
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Kevin Bales
Wilberforce Institute for the Study of Slavery and Emancipation, University of Hull, Hull, UK

Sarah Banks
Durham University, Durham, UK

Gregg Barak
Eastern Michigan University, Ypsilanti, MI, USA

Kimberly L. Barrett
Eastern Michigan University, Ypsilanti, MI, USA

Logie Barrow
Universitaet Bremen, Bremen, Germany

Sebastian E. Bartos
School of Psychology, University of Surrey, Guildford, UK

Johannes Beck
Center for Affective, Stress and Sleep Disorders (ZASS), Psychiatric Clinics of the University of Basel, Basel, Switzerland

Nathan E. Bell
American Educational Research Association, Washington, DC, USA

Wanda T. Bernard
Dalhousie University, Halifax, NS, Canada

Lorenzo Bizzi
California State University Fullerton, Fullerton, CA, USA

Chiara D. Bo
Università degli Studi di Milano, Milan, Italy

Hannah Boettcher
Boston University, Boston, MA, USA

Thibaud Boncourt
European University Institute, Fiesole, Italy

Lorenzo Bosi
European University Institute, Fiesole, FI, Italy

Dominique Bouchet
University of Southern Denmark, Odense, Denmark; and School of Creative Arts, James Cook University, Townsville, QLD, Australia

Sophie Bowlby
University of Loughborough, Loughborough, UK; and University of Reading, Reading, UK

Jerrold R. Brandell
School of Social Work, Wayne State University, Detroit, MI, USA

Ulrik Brandes
University of Konstanz, Konstanz, Germany

Chelsea N. Brown
University of Texas at San Antonio, San Antonio, TX, USA

Mario Bunge
McGill University, Montreal, QC, Canada

Trevor Burnard
University of Melbourne, Parkville, VIC, Australia

Viv Burr
School of Human and Health Sciences, University of Huddersfield, Huddersfield, UK

Jeremy Burrus
ACT, Iowa City, IA, USA

John T. Cacioppo
The University of Chicago, Chicago, IL, USA

Stephanie Cacioppo
The University of Chicago, Chicago, IL, USA

Andrea Caragliu
Politecnico di Milano, Milan, Italy

Lauren Carney
Educational Testing Service, Princeton, NJ, USA

Antonio J. Castro
University of Missouri, Columbia, MO, USA

Tarani Chandola
University of Manchester, Manchester, UK

Louis Chauvel
Research Unit INSIDE, IRSEI Institute, University of Luxembourg, Walferdange, Luxembourg

Gordon C.K. Cheung
School of Government & International Affairs, Durham University, Durham, UK

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece

Kevin K.H. Chung
The Hong Kong Institute of Education, Tai Po, New Territories, Hong Kong SAR China

Margaret S. Clark
Trumbull College, Yale University, New Haven, CT, USA

Bjørgulf Claussen
Institute of Health and Society, University of Oslo, Blindern, Oslo, Norway

Richard Conibere
University of Manchester, Manchester, UK

Christoph Conrad
University of Geneva, Geneva, Switzerland

Kristin Conzelmann
German Aerospace Center, Hamburg, Germany

Michael Cronin
Monmouth University, West Long Branch, NJ, USA

Freek Cronjé
Bench Marks Centre for CSR, North West University, Potchefstroom, South Africa

Charles Crothers
School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

Fred D'Agostino
The University of Queensland, St Lucia, QLD, Australia

Martin Daly
University of Missouri, Columbia, MO, USA

Ellice Daniel
Ashura Counselling and Consulting, Surrey, BC, Canada

Monti N. Datta
University of Richmond, Richmond, VA, USA

Christian Dayé
University of Graz, Graz, Austria

Kay Deaux
City University of New York, New York, NY, USA

J. Gregory Dees[†]
Duke University, Durham, NC, USA

Britt De Soete
Ghent University, Ghent, Belgium

Chares Demetriou
National Research University, Moscow, Russian Federation

Wolfgang Detel
University of Frankfurt/Main, Frankfurt/Main, Germany

John Drury
School of Psychology, University of Sussex, Falmer, Brighton, UK

Mitch Duneier
Princeton University, Princeton, NJ, USA

Marie Duru-Bellat
Sciences-Po, Paris, France

Lucia Falzon
Defence Science and Technology Organisation, Edinburgh, SA, Australia

Christine Fawcett
Uppsala University, Uppsala, Sweden

Lia Figgou
Aristotle University of Thessaloniki, Thessaloniki, Greece

Marina Fischer-Kowalski
Institute of Social Ecology, Alpen-Adria University, Vienna, Austria

[†] Deceased.

Florian Fritz
Institut für Rechts- u. Kriminalsoziologie, Vienna, Austria

Mónika Ganczer
Hungarian Academy of Sciences, Centre for Social Sciences Institute for Legal Studies, Budapest, Hungary

Emily Gardiner
Simon Fraser University, Burnaby, BC, Canada

Robert H. Gass
California State University, Fullerton, Fullerton, CA, USA

Beverly Gaudet
Faculty of Nursing, University of New Brunswick, Fredericton, NB, Canada

Maria Beatriz Gavilán
Gerencia de Desarrollo Científico y Tecnológico, Buenos Aires, Argentina

Laura J. George
Washington Center for Psychoanalysis, Washington, DC, USA

Purnima George
Ryerson University, Toronto, ON, Canada

Médhi Gilson
Université Libre de Bruxelles, Brussels, Belgium; and FNRS, Belgium

Sandra González-Bailón
Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA

Johan Goudsblom
University of Amsterdam, Amsterdam, The Netherlands

David B. Grusky
Stanford University, Stanford, CA, USA

Kai Hakkarainen
University of Turku, Turku, Finland

Tom P. Harsanyi
Harvard University, Boston, MA, USA

Peter Hegarty
School of Psychology, University of Surrey, Guildford, UK

Johan Heilbron
Centre de sociologie et de science politique de la Sorbonne (CNRS-EHESS), Paris, France

Mariana Heredia
The National Scientific and Technical Research Council/ UBA, Buenos Aires, Argentina

Gerard Van Herk
Memorial University of Newfoundland, St. John's, NL, Canada

Gudmund Hernes
Fafo Institute, Oslo, Norway; and BI Norwegian School of Management, Oslo, Norway

Lauri Hietajärvi
University of Helsinki, Helsinki, Finland

Jennifer L. Hirsch
Yale University, New Haven, CT, USA

Stefan G. Hofmann
Boston University, Boston, MA, USA

Bernie Hogan
Oxford Internet Institute, University of Oxford, Oxford, UK

Eddy Hogg
School of Social Policy, Sociology and Social Research, The University of Kent, Canterbury, UK

Edith Holsboer-Trachsler
Center for Affective, Stress and Sleep Disorders (ZASS), Psychiatric Clinics of the University of Basel, Basel, Switzerland

Julianne Holt-Lunstad
Brigham Young University, Provo, UT, USA

Bernd R. Hornung
Philipps University Marburg, Marburg, Germany

Katherine Hubbard
School of Psychology, University of Surrey, Guildford, UK

Grace Iarocci
Simon Fraser University, Burnaby, BC, Canada

Elena Iarskaia-Smirnova
National Research University 'Higher School of Economics,' Moscow, Russia

Gerhard Igl
Christian Albrechts University, Kiel, Germany

Takashi Inoguchi
University of Niigata Prefecture, Niigata, Japan; and University of Tokyo, Tokyo, Japan

Y. Ito
Keio University, Shonan Fujisawa, Fujisawa, Japan

Urs Jäger
INCAE Business School, Alajvela, Costa Rica

Guillermina Jasso
New York University, New York, NY, USA

David N. Jones
People Need People Consulting, Northampton, UK

Hartmut Kaelble
Institute for History, Humboldt University Berlin, Berlin, Germany

Heinz Katschnig
Ludwig Boltzmann Institute for Social Psychiatry, Vienna, Austria

Rachel L. Kendal
Durham University, Durham, UK

Olessia Kirtchik
National Research University—Higher School of Economics, Moscow, Russia

Ryan T. Knowles
University of Missouri, Columbia, MO, USA

Thomas König
Vienna Science and Technology Fund, Vienna, Austria

Nectarios Kontoleon
Defence Science and Technology Organisation, Edinburgh, SA, Australia

Karima Kourtit
VU University Amsterdam, Amsterdam, The Netherlands

Reinhard Kreissl
Institut für Rechts- u. Kriminalsoziologie, Vienna, Austria

Enrico Dal Lago
National University of Ireland, Galway, Ireland

Kenneth C. Land
Duke University, Durham, NC, USA

Paula M. Lantz
The George Washington University, Washington, DC, USA

John Latsis
University of Reading, Reading, UK; and University of Oxford, Oxford, UK

Monica Lee
University of Chicago, Chicago, IL, USA

Melinda S. Leidy
University of California, Riverside, CA, USA

Anja K. Leist
Research Unit INSIDE, IRSEI Institute, University of Luxembourg, Walferdange, Luxembourg

Nicole Letourneau
Faculties of Nursing and Medicine (Pediatrics and Psychiatry), University of Calgary, AB, Canada

Felice J. Levine
American Educational Research Association, Washington, DC, USA

Filip Lievens
Ghent University, Ghent, Belgium

Ulf Liszkowski
University of Hamburg, Hamburg, Germany

YanJiao Liu
Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

Nicholas J. Long
London School of Economics and Political Science, London, UK

Kirsti Lonka
University of Helsinki, Helsinki, Finland

Jean D. Losango
Bench Marks Centre for CSR, North West University, Potchefstroom, South Africa

Samuel Lule
Social Workers and Anthropologists (ESSSWA), Addis Ababa, Ethiopia

Mark Lymbery
School of Sociology and Social Policy, University of Nottingham, Nottingham, UK

Michael J. Lynch
University of South Florida, Tampa, FL, USA

Cecilia Mabragaña
Gerencia de Desarrollo Científico y Tecnológico, Buenos Aires, Argentina

Richard Machalek
University of Wyoming, Laramie, WY, USA

Michael W. Macy
Cornell University, Ithaca, NY, USA

John L. Martin
University of Chicago, Chicago, IL, USA

Michael W. Martin
Adams State University, Alamosa, CO, USA

Kenneth I. Mavor
School of Psychology and Neuroscience, University of St Andrews, St Andrews, Fife, UK

Rebecca A. Maynard
Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

Craig McGarty
University of Western Sydney and Murdoch University, Australia

Joshua Melville
Oxford Internet Institute, University of Oxford, Oxford, UK

Byron Miller
University of Calgary, Calgary, AB, Canada

Atsushi Miyaoka
Westat, Rockville, MD, USA

Anna-Lisa Müller
University of Bremen, Bibliothekstr, Bremen, Germany

Elizabeth E. Mustaine
University of Central Florida, Orlando, FL, USA

Mary Nash
Massey University, Palmerston North, New Zealand; and Aotearoa New Zealand Association of Social Workers, New Zealand

Elizabeth A. Necka
The University of Chicago, Chicago, IL, USA

Victor Nee
Cornell University, Ithaca, NY, USA

Friedhelm Neidhardt
Wissenschaftszentrum Berlin für Sozialforschung, Germany

Moira Nelson
Lund University, Lund, Sweden

Alex Nicholls
University of Oxford, Oxford, UK

Peter Nijkamp
VU University Amsterdam, Amsterdam, The Netherlands

Carolyn Noble
Victoria University, Melbourne, VIC, Australia

P. Nolte
International University Bremen, Bremen, Germany

Helga Nowotny
Vienna Science and Technology Fund, Vienna, Austria

Enrique Ogliastri
IE Business School, Madrid, Spain

Sonja Opper
School of Economics and Management, Lund University, Lund, Sweden

Larry L. Orr
Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Lars Ostermeier
Zentrum für Technik und Gesellschaft, Technical University Berlin, Berlin, Germany

Ross D. Parke
University of California, Riverside, CA, USA

Orlando Patterson
Harvard University, Cambridge, MA, USA

Philippa Pattison
School of Psychological Sciences, University of Melbourne, Parkville, VIC, Australia

Vassilis Pavlopoulos
National and Kapodistrian University of Athens, Athens, Greece

Malcolm Payne
Manchester Metropolitan University, Sutton, UK

Philippe Peigneux
Université Libre de Bruxelles, Brussels, Belgium

Abby Peterson
University of Gothenburg, Göteborg, Sweden

Keith J. Petrie
University of Auckland, Auckland, New Zealand

Barbara Pfetsch
Communication Theory and Media Effects, Freie Universität Berlin, Garystrasse, Berlin, Germany

Jennifer Platt
University of Sussex, Falmer, Brighton, UK

Andrea Prado
INCAE Business School, Alajvela, Costa Rica

Nick Pratt
Plymouth University, Plymouth, UK

Felicia Pratto
University of Connecticut, Storrs, CT, USA

Giovanna Procacci
University of Milan, Milan, Italy

Ezequiel Reficco
Universidad de los Andes, Bogotá, Colombia

Marsha Regenstein
The George Washington University, Washington, DC, USA

Rena L. Repetti
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Katherine J. Reynolds
Research School of Psychology, The Australian National University, Canberra, ACT, Australia

François Ric
Université de Bordeaux, Bordeaux, France

Marc N. Richelle
University of Liège and Académie Royale de Belgique, Liège, Belgium

Dirk Richter
Bern University of Applied Sciences, Bern, Switzerland

David Rohall
Western Illinois University, Macomb, IL, USA

Andrea Römmele
Hertie School of Governance, Friedrichstrasse, Berlin, Germany

Wulf Rössler
Leuphana University Lüneburg, Lüneburg, Germany; University of Sao Paulo, Sao Paulo, Brazil; and University of Zurich, Zurich, Switzerland

Sheldon Rothblatt
University of California Berkeley, Berkeley, CA, USA

David-Hillel Ruben
University of London, London, UK

András Sajó
Central European University, Budapest, Hungary

Katariina Salmela-Aro
Cicero Learning, University of Helsinki, Helsinki, Finland

Jin Sato
The University of Tokyo, Tokyo, Japan

Victor R. Savage
National University of Singapore, Singapore

Angelika Schlarb
Faculty of Psychology and Sports Science, University of Bielefeld, Bielefeld, Germany

Klaus A. Schneewind
Ludwig-Maximilians-University Munich, Munich, Germany

John Scott
University of Plymouth, Plymouth, UK; and University of Copenhagen, Copenhagen, Denmark

Meredith S. Sears
University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Ullica Segerstrale
Illinois Institute of Technology, Chicago, IL, USA

Erich Seifritz
Zurich University Hospital of Psychiatry, Zurich, Switzerland

Ian Sheinheit
University at Albany, SUNY, Albany, NY, USA

Wang Si-bin
Peking University, Beijing, People's Republic of China

Rainer K. Silbereisen
Center for Applied Developmental Science, University of Jena, Jena, Germany

Daniel P. Skorich
Research School of Psychology, The Australian National University, Canberra, ACT, Australia

Steven Rathgeb Smith
American Political Science Association, Washington, DC, USA

André Spicer
Cass Business School, City University, London, UK

Maura Spiegelman
Westat, Rockville, MD, USA

Michael Steketee
Westat, Jenison, MI, USA

Melika H. Stephens
University of Auckland, Auckland, New Zealand

Andrew L. Stewart
Clark University, Worcester, MA, USA

Frans N. Stokman
University of Groningen, Groningen, The Netherlands

Jerry Suls
Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA

Danika Syrja-McNally
Ryerson University, Toronto, ON, Canada

Barbra Teater
College of Staten Island, City University of New York, NY, USA

Jean Terrier
Institut für Politikwissenschaft, Westfälische Wilhelms-Universität Münster, Münster, Germany

Paul Thomas
University of California, Berkeley, CA, USA

Charis M. Thompson
London School of Economics and Political Science, London, UK

Georg Thurn
Wissenschaftszentrum Berlin für Sozialforschung, Germany

L. Tilly
New School for Social Research, NY, USA

Sarah Todd
School of Social Work, Carleton University, Ottawa, ON, Canada

Martin J. Tomasik
University of Zurich, Zurich, Switzerland

Lars Trägårdh
Ersta Sköndal University College, Stockholm, Sweden

Jonathan H. Turner
University of California Riverside, Riverside, CA, USA

Stephen P. Turner
University of South Florida, Tampa, FL, USA

Koen van Bommel
VU University, Amsterdam, The Netherlands

Eric van Dijk
Leiden University, Leiden, The Netherlands

Koen van Eijck
Erasmus University Rotterdam, Rotterdam, The Netherlands

Johann van Wyk
Bench Marks Centre for CSR, North West University, Potchefstroom, South Africa

Katherine Verdery
Graduate Center, City University of New York, NY, USA

Antonio Vives
Stanford University, Stanford, CA, USA

Sue Waite
Plymouth University, Plymouth, UK

Weidong Wang
Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

Jeffrey T. Ward
University of Texas at San Antonio, San Antonio, TX, USA

Kenneth E. Warner
School of Public Health, University of Michigan, Ann Arbor, MI, USA

Scott L. Washington
University of Chicago, Chicago, IL, USA

Stanley Wasserman
University of Illinois, Urbana, IL, USA

Kathleen M. Weigert
Loyola University Chicago, Chicago, IL, USA

Susanne Weis
Center for Psychological Methods, Diagnostics, and Evaluation, University of Koblenz-Landau, Landau, Germany

Richard Whatmore
School of History, University of St Andrews, St Andrews, UK

Ladd Wheeler
Macquarie University, Sydney, NSW, Australia

Paul F. Whiteley
University of Essex, Colchester, UK

Jade Q. Wu
Boston University, Boston, MA, USA

Xue Yan
Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

Valentina Yarskaya
Saratov State Technical University, Saratov, Russia

Gisèle Yasmeen
Research — Social Sciences and Humanities Research Council of Canada, Ottawa, ON, Canada

Petri Ylikoski
University of Helsinki, Helsinki, Finland

Angelina W.K. Yuen-Tsang
Hong Kong Polytechnic University, Hunghom, Hong Kong

Milan Zafirovski
University of North Texas, Denton, TX, USA

Darja Zaviršek
University of Ljubljana, Ljubljana, Slovenia

This page intentionally left blank

Situated Knowledge, Feminist and Science and Technology Studies Perspectives

Charis M Thompson, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the feminist and science and technology studies roots of situated epistemology are discussed. The idea of combining a commitment to a real world, albeit a constructivist one, with attention to practice and embodiment is assessed. The controversial conferral of agency on nonhuman objects is similarly explicated. Finally, uptake and critique of the notion of situated knowledges is addressed.

The expression 'situated knowledge,' especially in its plural form 'situated knowledges,' is associated with feminist epistemology, feminist philosophy of science, and science and technology studies. The term was introduced by historian of the life sciences and feminist science and technology studies scholar Donna Haraway in her landmark essay *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective* (Haraway, 1991: 183–201). The essay was a response to feminist philosopher of science Sandra Harding's discussion of the "science question in feminism" (Harding, 1986). In her analysis of the potential of modern science to contribute to the goals of feminism, Harding noted three different accounts of objective knowledge in feminist epistemology: feminist empiricism, feminist standpoint, and feminist postmodernism (Harding, 1986: 24–26). Feminist empiricism attempted to replace more-biased with less-biased science. The feminist standpoint, echoing the Marxist tradition from which it derived, stressed the relevance of the social positioning of the knower to the content of what is known. Feminist postmodernism accentuated the power dynamics underlying the use of the language of objectivity in science. Haraway, taking off from Harding, diagnosed a 'Scylla and Charybdis' of temptations between which feminists attempt to navigate on the question of objectivity: radical constructionism and feminist critical empiricism. As she put it, what feminists wanted from a theory of objectivity was "enforceable, reliable accounts of things not reducible to power moves and agonistic, high status games of rhetoric or to scientistic, positivist arrogance" (Haraway, 1991: 188). Her notion of situated knowledges was her attempt to provide just such a theory of objectivity.

Despite the single author provenance, the term resonated with attempts by other scholars in the history, sociology, and philosophy of science to address similar epistemological tensions. The concept was quickly and fruitfully taken up in science and technology studies and feminist theory, provoking a certain amount of reworking in its turn.

'Situated Knowledge' and Objectivity: Constructed Yet Real

The term 'situated knowledge' derives its theoretical importance from its seemingly oxymoronic character, particularly when applied to knowledge about the natural world. It is common to think of modern scientific knowledge as universal, so that it has the same content no matter who possesses it. It is also almost definitional to hold that objective knowledge is warranted by the fact that it captures reality as it really is, rather than being warranted by the situational circumstances out of which the knowledge was generated or discovered (see Shapin, 1994: 1–8 for discussion of, and citations relevant to, various manifestations of these points). Thus, if the law of gravity enables us to make reliable experimental predictions, it is because there is such a thing as gravity that is adequately captured by our scientific understanding; in short, the truth of the knowledge is its own warrant. It is only in the case of false or superseded knowledge that we typically explain what went wrong by reference to faulty assumptions, sloppy work, ill-calibrated equipment, the *Zeitgeist*, or other aspects of the context of discovery. The idea of 'situated knowledge' contests these supposed concomitants of objective knowledge. It suggests that objective knowledge, even our best scientific knowledge of the natural world, depends on the *partiality* of its material, technical, social, semiotic, and embodied means of being promulgated.

Haraway's notion thus had affinities with other feminist epistemologies, which had noted that facts can differ in their content from one time, place, and knower to another (e.g., Collins, 1989). It also had sympathies in common with sociologists of science and scholars of science and technology studies who have suggested that capturing 'reality as it really is' may be dependent on institutional, technical, and cultural norms (Kuhn (1962/1970)), on practice (Clarke and Fujimura, 1992; Pickering, 1992), and attempts to witness, measure, comprehend, or command assent to it (Latour, 1987; Shapin, 1994; Shapin and Schaffer, 1985). All these scholars shared a search for the theoretical resources to do justice to the embeddedness of science and truth. These challenges to conventional views of objectivity brought situated knowledges into conversation with key debates in the philosophy of science around the theory-ladenness of facts (Hesse, 1980: 63–110). Additionally, the suspicion of transcendent universalism entrained an epistemological and political distrust of clear-cut distinctions between subject and object, and a blurring of the distinction between context and content of knowledge or discovery.

Situated knowledges are as hostile to relativism as they are to realism. Haraway describes relativism as "being nowhere while claiming to be everywhere equally" (Haraway, 1991: 191) and realism as "seeing everything from nowhere" (Haraway, 1991: 189), and conceives of them both as 'god-tricks' promising total, rather than partial, located, and embodied, vision. In contrast to realist or relativist epistemologies, Haraway sees the possibility of sustained and rational objective inquiry in the epistemology of partial perspectives. This requires, she maintains, reclaiming vision as a series of technological and organic embodiments, as and when and where and how vision is actually enabled. This crafting of a feminist epistemology of situated knowledges on the basis of vision and partial perspective is noteworthy. The links in the history of science to militarism, capitalism, colonialism, and male supremacy have been theorized around the masculinist gaze of the powerful but disembodied knower disciplining and subjugating the weak by means of a multitude of technologies of surveillance. Feminists have lamented the privilege granted to the visual as a sure basis of knowledge and bemoaned the sidelining in modernity of what some cast as more feminine and less intrinsically violent ways of knowing involving emotion, voice, touch, and listening (Gilligan, 1982). Haraway was concerned that feminists not cede power to those whose practices they wish critically to engage. It is in this spirit that she grounded her feminist solution in an embrace of science and vision, "the real game in town, the one we must play" (Haraway, 1991: 184).

The Feminist Roots of the Dilemma

A tension between emancipatory empiricism and its associated egalitarian or socialist politics, and feminist postmodern constructionism and its associated identity politics, resonates throughout contemporary Western feminist theory. It is a hallmark of those engaged in feminist philosophical and social studies of science that they seek to resolve one or another version of this tension. One horn of the feminist dilemma, according to Haraway, represents the good feminist reasons to be attracted to radical constructionism. Feminist postmodernists, and analysts of science and technology influenced by semiotics (including Haraway herself), helped develop and often appealed to "a very strong social constructionist argument for *all* forms of knowledge claims" including scientific ones (Akrich and Latour, 1992; Haraway, 1991: 184). This position had the benefit of showing the links between power – such things as status, equipment, rhetorical privilege, funding, and so on – and the production of knowledge and credibility. The downside, from the point of view of feminists interested in arguing for a better world, is that the radical constructionist argument risks rendering all knowledges as fundamentally ideological, with no basis for choosing between more and less just ideas, more and less true versions of reality. As Haraway provocatively expressed it, embracing this temptation seemed to leave no room for "those of us who would still like to talk about *reality* with more confidence than we allow the Christian right's discussion of the Second Coming and their being raptured out of the final destruction of the world" (Haraway, 1991: 185).

The second horn of the dilemma, according to Haraway, involves "holding out for a feminist version of objectivity" through materialism or empiricism (Haraway, 1991: 186). Haraway briefly discusses both Marxist derived feminisms and feminist empiricism. Feminisms with Marxist inspirations are several, and their genealogy can be traced in a number of ways. Feminists have long criticized Marxist humanism for its premise that the self-realization of man is dependent on the domination of nature, and for its account of the historical progression of modes of production that grants no historical agency to domestic and unpaid labor (Hartmann, 1981). Some have responded by developing feminist versions of historical materialism (Hartsock, 1983). Feminist standpoint theorists appropriated the general insight, inherited from both Marxist thought and the sociology of knowledge, that one's social-structural position in society – such things as one's class or relation to the means of production, or one's gender or ethnonational characteristics – determines or affects how and what one knows (Smith, 1990).

Likewise, the idea that some social-structural positions confer epistemological privilege has been widely adopted by standpoint theorists and feminist epistemologists arguing for specifically feminine ways of knowing (Rose, 1983). 'Seeing from below,' that is, from a position of subordination, has commonly been theorized by feminists as the position of epistemological privilege, on the grounds that those with little to gain from internalizing powerful ideologies would be able to see more clearly than those with an interest in reproducing the status quo. In Patricia Hill Collins' version of standpoint theory, for example, these insights are used both to validate the knowledges of the historically disenfranchised and to reverse the hegemonic ranking of knowledge and authority, and claim epistemological privilege for African-American women (Collins, 1989).

Psychoanalytic theory, particularly Anglophone object relations theory, inspired some of the early writings on gender and science (Chodorow, 1978; Keller, 1985). Object relations theory attempted to explain the different relation of women and men to objectivity, abstract thought, and science in modern societies. To account for this difference, the theory posited gender-based differences in the socialization of sons and daughters in Western middle-class heterosexual nuclear families. Boys, according to this theory, are socialized to separate from the primary caregiver who is the mother in this normative family scenario. They thus learn early and well by analogy with their emotional development that relational thinking is inappropriate for them; separating themselves from the object of knowledge, as from the object of love, is good. Girls, on the other hand, are supposedly socialized to be like their primary caregiver, so that they can reproduce mothering when their turn comes. Relationality and connectivity, not abstraction and separation, are the analogous ordering devices of girls' affective and epistemological worlds. As applied to objectivity and scientific knowledge, object relations theory seemed to explain to feminists, without resort to distasteful biological determinisms denying women scientific aptitude, why women were excluded from much of science and technology. It also suggested that there were (at least) two distinct ways of knowing, and that much might have been lost in the violence and separation of masculinist science that could be

restored by a proper valuation of the feminine values of connection and empathy (Harding, 1986; Keller, 1983). Like Marxism, psychoanalytic approaches to objectivity gave feminists a means to show the relevance of one's social position to knowledge. Like feminist empiricism, they encouraged the belief in the possibility of an improved, feminist, objectivity (Harding, 1992).

The feminist canon contains a number of empirical studies that have revealed the negative effects of such things as colonialism and stereotypes about race and gender on the production of reliable science (Fausto-Sterling, 1995; Martin, 1991/1996; Schiebinger, 1989; Traweek, 1988). Evelyn Fox Keller's call for "dynamic objectivity" (Keller, 1985: 115–26) and Sandra Harding's demand for "strong objectivity" (Harding, 1992: 244) are exemplary of the aspirations of theoretical feminist empiricism. These projects seek to prescribe scientific methods capable of generating accounts of the world that would improve upon disembodied, masculinist portrayals of science because they would be alert to the practices of domination and oppression inherent in the creation, dissemination, and possession of knowledge. Feminist empiricism nonetheless remains problematic because of its reliance on the dichotomies of bias vs. objectivity, use vs. misuse, and science vs. pseudoscience. The feminist insight of the "contestability of every layer of the onion of scientific and technological constructions" (Haraway, 1991: 186) flies in the face of leaving these epistemological dichotomies intact.

Subjects, Objects, and Agency

Haraway's notion of situated knowledges problematizes both subject and object. Unlike standpoint theories which attribute epistemological privilege to subjugated knowers, and the sociology of knowledge which attributes espitemological privilege to those in the right structural position vis-à-vis a given mode of production, Haraway attributes privilege to partiality. This shift underscores that "situated knowledge" is more dynamic and hybrid than other epistemologies that take the position of the knower seriously, and involves "mobile positioning" (Haraway, 1991: 192)

In situated knowledges based on embodied vision, neither subjects who experience, nor nature which is known, can be treated as straightforward, pre-theoretical entities, "innocent and waiting outside the violations of language and culture" (Haraway, 1991: 109). Haraway maintains that romanticizing, and thus homogenizing and objectifying, the perfect subjugated subject position is not the solution to the violence inherent in dominant epistemologies. As feminists from developing countries have also insisted, there is no innocent, perfectly subjugated feminist subject position conferring epistemological privilege; all positionings are open to critical reexamination (Mohanty, 1984/1991). Subjectivity is instead performed in and through the materiality of knowledge and practice of many kinds (Butler, 1990: 1–34).

Conversely, the extraordinary range of objects in the physical, natural, social, political, biological, and human sciences about which institutionalized knowledge is produced should not be considered to be passive and inert. Haraway says that situated knowledges require thinking of the world in terms of the "apparatus of bodily production." The world cannot be reduced to a mere resource if subject and object are deeply interconnected. Bodies as objects of knowledge in the world should be thought of as "material-semiotic generative nodes," whose "boundaries materialize in social interaction" (Haraway, 1991: 201). The move to grant agency to material objects places the epistemology of situated knowledges at the center of scholarship in science and technology studies (Callon, 1986; Latour, 1987).

Uptake and Critique

Donna Haraway's essay ranks among the most highly cited essays in science and technology studies and has been anthologized. As stated above, situated knowledges is a provocative and rich methodological metaphor with resonances in many quarters. The dialogue between Harding and Haraway continued after the publication of *Situated Knowledges* (Harding, 1992: 119–163). Her epistemology directly influenced, and in turn was influenced by, the work of sociologists and anthropologists of science (Clarke and Montini, 1993; Rapp, 1999), feminist philosophers of science (Wylie, 1999), and practicing feminist scientists (Barad, 1996). In addition, 'situated knowledges' was used as a standard technical term of the field by more junior scholars, including many of Haraway's former students. Over time, the idea of situated knowledges has appeared in more domains, including scholarship on education and the environment (e.g., Lang, 2011; Calvert, 2012).

Critics of situated knowledges have been few. Timothy Lenoir has pointed out that many of the epistemological ideas behind Haraway's situated knowledges are found not only in other major strands of science and technology studies, but also in the work of continental philosophers such as Nietzsche. He likewise critiqued the idea of situated knowledges for its dependence on the apparatus of semiotics (Lenoir, 1999: 290–301). Historian Londa Schiebinger, in her book summarizing the effects of a generation of feminist scholarship on the practice of science, placed Haraway's situated knowledges together with Harding's strong objectivity as attempts to integrate social context into scientific analysis (Schiebinger, 1999). Implicit critiques have been leveled against the limitations of the idea of being situated, for example, in the development of De Laet's and Mol's mobile epistemology (De Laet and Mol, 2000). Sheila Jasanoff and her colleagues argued for bringing differently spatialized entities such as the nation, the local, and the global into the epistemology of science and technology studies, while retaining the insights gained by paying attention to practice, vision, and measurement. These critiques stand more as continuing conversations with, than rebuttals of situated knowledges, however. Scholars and former Haraway students have continued to work with situated knowledges and its limits (e.g., Charis Thompson's (2014) loss of her field-site, and Kim TallBear's (2013) idea of 'rootedness' and 'routedness'). Overall, the idea of situated knowledges remains central to feminist epistemology and science studies and to attempts to understand the role of modern science in society.

See also: Anthropology of Knowledge; Contextual Studies: Methodology; Cultural Studies of Science; Knowledge, Sociology of; Science and Technology, Anthropology of; Scientific Knowledge, Sociology of; Standpoint Theory, in Science.

Bibliography

Akrich, M., Latour, B., 1992. A summary of a convenient vocabulary for the semiotics of human and nonhuman assumblies. In: Bijker, W., Law, J. (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 259–264.

Barad, K., 1996. Meeting the universe halfway: realism and social constructivism without contradiction. In: Hankinson Nelson, L., Nelson, J. (Eds.), Feminism, Science, and the Philosophy of Science. Kluwer, Dordrecht, The Netherlands, pp. 161–194.

Butler, J., 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, London.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and fishermen of St. Brieuc Bay. In: Law, J.J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge. Routledge, London, pp. 196–233.

Calvert, S., 2012. West of Eden: resource wars and nature-cultures in the american West. Review of the american West at risk: science, myths, and politics of land abuse and recovery. Science as Culture 21 (1), 129–134.

Chodorow, N., 1978. The Reproduction of Mothering. University of California Press, Berkeley.

Clarke, A., Fujimura, J. (Eds.), 1992. The Right Tools for the Job: At Work in Twentieth-Century Life Sciences. Princeton University Press, Princeton, NJ.

Clarke, A., Montini, T., 1993. The many faces of RU486: tales of situated knowledges and technological contestations. Science, Technology and Human Values 18, 42–78.

Collins, P.H., 1989. The social construction of black feminist thought. Signs 14 (4), 745–773.

De Laet, M., Mol, A., 2000. The Zimbabwe bush pump. Mechanics of a fluid technology. Social Studies of Science.

Fausto-Sterling, A., 1995. Gender, race, and nation: the comparative anatomy of 'Hottentot' women in Europe, 1815–1817. In: Terry, J., Urla, J. (Eds.), Deviant Bodies: Critical Perspectives on Difference in Science and Popular Culture. Indiana University Press, Bloomington, IN, pp. 19–48.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Haraway, D., 1991. Simians, Cyborgs and Women: The Reinvention of Nature. Routledge, New York.

Harding, S., 1986. The Science Question in Feminism. Cornell University Press, Ithaca, NY.

Harding, S., 1992. Whose Science? Whose Knowledge? Thinking from Women's Lives. Cornell University Press, Ithaca, NY.

Hartmann, H., 1981. The unhappy marriage of marxism and feminism. In: Sargent, L. (Ed.), Women and Revolution. South End Press, Boston.

Hartsock, N., 1983. The feminist standpoint: developing the ground for a specifically feminist historical materialism. In: Harding, S., Hintikka, M. (Eds.), Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science. Reidel, Dordrecht, The Netherlands, pp. 283–310.

Hesse, M., 1980. Revolutions and Reconstructions in the Philosophy of Science. Indiana University Press, Bloomington, IN.

Keller, E.F., 1983. A Feeling for the Organism: The Life and Work of Barbara McClintock. W. H. Freeman and Company, New York.

Keller, E.F., 1985. Reflections on Gender and Science. Yale University Press, New Haven, CT.

Kuhn, T.S., 1962/1970. The Structure of Scientific Revolutions, second ed. Chicago University Press, Chicago.

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Open University Press, Philadelphia, PA.

Lang, J.C., 2011. Troubling knowledge in philosophy of education. Education Theory 61 (1), 75–96.

Lenoir, T., 1999. Was the last turn the right turn? The semiotic turn and A.J. Greimas. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, London, pp. 290–301.

Martin, E., 1991/1996. The egg and the sperm: how science has constructed a romance based on stereotypical male–female roles. In: Laslett, B., Kohlstedt, G.S., Longino, H., Hammonds, E. (Eds.), Gender and Scientific Authority. Chicago University Press, Chicago, pp. 323–339.

Mohanty, C.T., 1984/1991. Under Western eyes: feminist scholarship and colonial discourses. In: Mohanty, C., Russo, A., Torres, L. (Eds.), Third World Women and the Politics of Feminism. Indiana University Press, Bloomington, IN, pp. 51–80.

Pickering, A. (Ed.), 1992. Science as Practice and Culture. University of Chicago Press, Chicago.

Rapp, R., 1999. Testing Women, Testing the Fetus: The Social Impact of Amniocentesis in America. Routledge, London.

Rose, H., 1983. Hand, brain, and heart: a feminist epistemology for the natural sciences. Signs 9 (1), 73–90.

Schiebinger, L., 1989. The Mind Has No Sex? Women in the Origins of Modern Science. Harvard University Press, Cambridge, MA.

Schiebinger, L., 1999. Has Feminism Changed Science? Harvard University Press, Cambridge, MA.

Shapin, S., 1994. A Social History of Truth: Civility and Science in Seventeenth-century England. University of Chicago Press, Chicago.

Shapin, S., Schaffer, S., 1985. Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life. Princeton University Press, Princeton, NJ.

Smith, D., 1990. The Conceptual Practices of Power; a Feminist Sociology of Knowledge. University of Toronto Press, Toronto, ON.

TallBear, K., 2013. Native American DNA: Tribal Belonging and the Falso Promise of Genetic Science. University of Minnesota Press, Minneapolis.

Thompson, C., 2014. Good Science: The Ethical Choreography of Stem Cell Research. MIT Press, Cambridge, MA.

Traweek, S., 1988. Beamtimes and Lifetimes: The World of High Energy Physicists. Harvard University Press, Cambridge, MA.

Wylie, A., 1999. The engendering of archaeology: refiguring feminist science studies. In: Biagioli, M. (Ed.), The Science Studies Reader. Routledge, London, pp. 553–568.

Situated Learning (Learning In Situ)

Sue Waite and Nick Pratt, Plymouth University, Plymouth, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Renkl, volume 21, pp. 14133–14137, © 2001, Elsevier Ltd.

Abstract

This article outlines the origins and meaning initially ascribed to 'situated learning,' drawing out the distinction between learning activity that takes place *in situ* and the epistemologically different idea that knowledge itself can be understood as situated. It details some instructional forms that have dominated the situated learning literature over recent years and which show how those original frames have become expanded and potentially distorted in application. It considers some critiques of situated learning. The focus then shifts to ways in which theorizing learning outside the classroom is developing more generally.

Situated learning refers to the notion that learning is affected by situation; and hence must be of relevance to anyone wishing to think about teaching and learning in and outside of classrooms. While this may seem straightforward and common sense, such a literal definition hides a number of complications, each of which has potential pedagogical implications.

This article first of all outlines the origins and meaning initially ascribed to situated learning. It then details some particular contexts that have dominated the situated learning literature over recent years and which show how those original frames have become expanded and potentially distorted in their application. It considers some critiques of situated learning. The focus then shifts to ways in which theorizing learning outside the classroom is developing more generally.

Learning *In situ*

Dickens' classic novel, *Hard Times* (Dickens, 1854: 1), begins with Mr Gradgrind's declaration that

> … what I want is, Facts. Teach these boys and girls nothing but Facts. Facts alone are wanted in life. Plant nothing else, and root out everything else. You can only form the minds of reasoning animals upon Facts: nothing else will ever be of any service to them.

Such an assertion reflects an assumption about learning as a process of acquisition. While Gradgrind wanted facts, a more contemporary version might also add skills and understanding as well as attitudes and beliefs. However, as Sfard (1998) points out, as long as theories "focus on the 'development of concepts' and 'acquisition of knowledge,' they implicitly agree that this process can be conceptualized in terms of the *acquisition metaphor*" (6, emphasis in original). Involved in this metaphor are many related assumptions about learning, and hence about teaching, including that: learning takes place through and in the mind; context stimulates learning, but is separate from it – means and ends are separate; knowledge, carried in the mind, is transferable; and expertise is a measure of quantity, rather than quality. Furthermore, there are a number of theories about how such acquisition takes place; Gradgrind's passive reception (the image of the empty vessel waiting to be filled comes from this same passage) has been superseded by notions of concept construction, with associated claims about readiness (Piaget, 1952) and then co-construction in a social arena (e.g., Vygotsky and Cole, 1978). Moreover, knowledge can also be of different forms: explicit, propositional, public, or hard; tacit, implicit, private, or soft (see Hildreth and Kimble (2002) and Eraut (2000) for useful reviews of these terms).

In all these ways of conceiving of learning, the individual is the unit of analysis. That is, whatever the mode by which learning takes place, the learning itself resides within and is the possession of the individual. In this individualized conception of learning, the role of situation is part of the *means* by which learning takes place. From this perspective its role is to alter the stimulus through, or mechanism by, which individuals come to acquire new knowledge – the practice of learning. Coupling this with the notion that the context in which something is learned, shapes the *form* of resultant learning and is therefore more, or less, useful in different situations, leads to the common idea that much of what is learned in formal settings (schools, universities, etc.) is 'inert' (Renkl et al., 1996) and/or learned inefficiently. Instead, it is argued, learning should take place in contexts that reflect 'real' situations, either replicated in the formal setting or by taking learners outside so that learning is both stimulated through experiences with which to construct meaning and developed in forms that 'fit' future use.

Learning *as Situated*

In the 1980s, a number of studies began to question this acquisition metaphor of learning and point to problems with the whole idea of context. Largely, these were prompted by observations of learning taking place in nonformal situations. Jean Lave's work with apprentice tailors in Liberia and with supermarket shoppers' mathematics perhaps most famously (though by no means solely – see also, for example, Rogoff, 1990; Nunes et al., 1993) led to the development of the idea of cognition in practice (Lave, 1988; Chaiklin and Lave, 1996) and subsequently to learning situated in practice within communities (Lave and Wenger, 1991). As Lave (1996: 5) notes,

> It is difficult when looking at everyday activity … to avoid the conclusion that learning is ubiquitous in ongoing activity. Situated activity always involves changes in knowledge and action … and "changes in knowledge and action" are what we mean by "learning."

Whereas previous research into cognition had always focused on individuals, albeit acting in social settings, this research refocused on the relationships between people, and between people and artifacts, in social activity. The unit of analysis here is no longer the individual, but the group and the activity undertaken within it; attention is fixed on learning that takes place through social 'community' practice; and learning itself – *what* is learned – becomes a question of how the person changes, not what is acquired, so that questions of identity and change become central to understanding learning. Contrary to learning as acquisition, Sfard (1998) refers here to the (metaphorical) idea of learning as *participation*.

This alternative, situated, perspective on learning involves an important distinction from the notion above. Whereas 'situated learning,' in the sense of learning (as acquisition) taking place *in situ*, has implications for the *means* by which learning takes place, Lave's use of the phrase implies an *epistemological* shift. In this version of 'learning being situated,' the meaning is that knowledge is no longer seen as a possession of the individual to be carried around, unpacked, and made use of across situations, but that meaning itself is 'held' in the relationships between people, artifacts, and activity. In this sense situated learning implies a change not just in the means of learning, but in its *ends* too; so that

> 'cognition' observed in everyday practice is distributed – stretched over, not divided among – mind, body, activity and culturally organized settings (which include other actors).
> (Lave, 1988: 1)

Learning from this perspective is 'embodied'; that is built into the activity of people's practice in social contexts so that it makes more sense to talk about 'knowing' as a process of activity and to situated *meaning*, than knowledge as an object.

It is perhaps worth noting at this stage that defining learning in terms of acquisition or participation reflects an attempt to make sense of something too complex to understand in any single way. In this sense, these theorizations are perhaps best seen as heuristic devices for making sense of how learning works, not claims that they model 'how learning is' (Lea, 2005). The participatory metaphor works at the level of social and cultural interaction. The acquisitive metaphor zooms in on the individual. Continue to zoom in and you come to neuroscience which, "broadly defined investigates the process by which the brain learns and remembers" (Goswami, 2004: 1); zoom out and you reach the sociological level, which aims to make sense of human activity in broad social contexts. All offer different ways to make sense of learning.

The situated view of learning, which has emerged from the focus on interaction of individuals in social practice, has been developed in a number of ways, the best known of which are probably Cultural Historical Activity Theory (CHAT) and Communities of Practice (CoP). The former takes its shape from Engeström who built on the Vygotskian idea of subjects' understanding of objects being mediated by artifacts by setting it in the wider social and historical context of community with its associated rules and divisions of labor. The resulting CHAT model (see, for example, Engeström et al., 1999) aims to focus attention on the whole system of activity within which people act, and the tensions and contradictions that drive such activity. The latter, CoP, comes from Lave and Wenger (1991; also Wenger, 1998) and focuses on the notion of practice in social communities and on learning as the movement from 'legitimate peripheral participation' to more central, expertise. Learning is again seen as change, this time in terms of identity (becoming), community (belonging), meaning (experiencing), and practice (doing) (Wenger, 1998: 5).

It should be evident from the above that the term 'situated learning' represents a contested concept complicated by the notion that 'situated' can refer to both the physical sense of place and authenticity and to the epistemological notion that knowledge itself can be understood to reside in and across people and artifacts in context. In the following sections, different instructional forms associated with theories of situatedness are discussed to illustrate how the concept has expanded and potentially been distorted in practice and we try, as best we can, to distinguish between those that relate to acquisitional learning *in situ* and/or to theories in which meaning itself is situated.

Learning as Apprenticeship

Brown et al. (1989) developed the instructional cognitive apprenticeship model in response to their belief that learning was social rather than individual and that abstract learning was of limited use for real-world contexts as it could not easily be applied. It was intended to make more visible the strategies and processes of an apprenticeship form of learning through authentic problems in school settings. Both expert and learners think out loud so that different ways of approaching the problem are available to all. Learners tackle tasks of increasing challenge supported by an 'expert' and gradually take on more responsibility, moving from legitimate peripheral participation to full participation (Lave and Wenger, 1991). Despite this articulated aspect, the learning retains tacit elements that characterize it as a form of enculturation into ways of doing and behaving rather than as ways of knowing. In Wenger's (1998: 5) terms, it is about understanding learning in terms of belonging, becoming, meaning, and experience. Sadler et al. (2010) reviewed 53 studies of scientific research apprenticeship schemes for secondary and postschool students and suggested that, while broadly successful, attention was needed to the desired outcomes, length of apprenticeship and motivational engagement of those taking part. A new wave of research on this form of situated learning may be stimulated by the UK government's recent revival of apprenticeship as a means of learning through the National Apprenticeship Scheme (2008), after years of declining apprenticeships, which fell from 240 000 in the mid-1960s to 53 000 by 1990 alongside rising post-16 participation in full-time education.

Place-Based Learning

A development in the relationship between situation and learning over the last decade has been in the field of place-based learning. As a challenge to content-driven, instructive educational approaches there has been a push within the environmental education sector for 'hands-on' experiential learning experiences that involve the co-creation of positive environmental and social change. This community interpretation of place-based education is more dominant in the US with proponents such as Gruenewald (2003) and Seaman (2007). The notion of situated learning is employed to describe changes that occur in projects *in situ*, for example, around watershed restoration, community forestry, and other civic ecology activities (Tidball and Krasny, 2007). Within these projects adults act as facilitators of learning and as experienced practitioners in a civic ecology community rather than as instructors. In this way, students develop the capacity to play a meaningful role in shaping their own future and that of their larger community.

In a transnational example of the role of situated learning in civic education, Hoskins et al. (2012) reanalyzed data from the 1999 IEA CIVED transnational study to examine how young people learn positive attitudes toward participation in and knowledge and skills of democracy. They found that learning through social participation, both inside and outside school, and in particular through meaning-making activities shows a strong positive relationship with citizenship knowledge, skills, and dispositions across a wide range of countries. This study is unusual in that it worked with situated learning theory and large-scale, quantitative data.

Turning to science learning through community participation, Fusco (2001) describes a 9-month science project geared toward youth between 12 and 16 years of age, which provided opportunities to participate in real-world activities as they turned a disused community garden into a useable space that included a playground, garden clubhouse, penny store, jungle gym, sandbox, and stage. The usefulness of the 'science' learned was questioned by Fusco (2001), who argues that these opportunities alone do not guarantee that young people will find school (or science) more relevant or be less alienated from school and society – for reasons we outline later in this article. However, a particular turning point for the project was when the young people involved in the project started to realize that the project was not 'fake'; "that they were not simply researching ideas (talking about science) but enacting them (doing science)"(868).

Different aspects of what it might mean to create CoP are considered by Boyer and Roth (2006) who describe their eelgrass research project from a situated learning perspective. They point out that different processes unfold concurrently; people of different ages engage in different actions and participate in different ways with multiple orientations rather than a single teacher, and the setting is both materially and socially heterogeneous and complex in comparison with a classroom setting. Such informal settings, they argue, "therefore provide opportunities to expand forms of participation that are less available in more structured school settings" (1030). A similar claim of expansion of learning opportunities is reported by Nadelson and Jordan (2012) who suggest that field trips as a form of situated learning increase students' interest and knowledge. Placing learning in context, they argue, leads students to become motivated and engaged and are therefore likely to generate greater subject matter interest.

Aguilar and Krasny (2011) applied the CoP theoretical framework to environmental education in their study of learning in after-school environmental education programs for Hispanic youth in the US. They suggest that this framing helps conceptualize the joint enterprise, mutual engagement, and shared repertoire of behavior that characterize some environmental education. It was also found to be useful in identifying and describing learning as bringing about changes in identity formation as a result of participation in the programs.

Border crossing between home and school communities is increasingly seen as a way to engage young people not succeeding at school. For example, Faircloth (2012) demonstrates in two studies how students negotiated an engaged identity-in-practice in the classroom. She employs the concept of cultural modeling derived from Lee (2007) to bridge the differences between school-based expectations and students' funds of knowledge by blending the 'authentic' and 'academic' lives of students. Gutiérrez and colleagues (Gutiérrez, 2008; Gutiérrez and Larson, 2007) defined this as 'third space'; the traditional expectations of schooling are interwoven with the informal identities that students value, which are derived from community, culture, family, and interests and which frequently receive less recognition in school. She suggests that this approach allows students to develop what she refers to as 'hybrid identities,' where personal values enmesh with school expectations. This study illustrates a potential critique of situated learning as a vehicle for enacting social change. The warrant for legitimacy of the practices of any community is a key factor and CoP may exclude as well as allow members to adopt more central positions. Wenger (1998) deals with this through reference to 'negotiability' within 'economies of meaning' to illustrate that communities are not neutral places but invested with power derived from wider social structures.

Problem-Based Learning

Problem-based learning is also associated with situated learning in response to a concern that school learning is divorced from practical utility for everyday living. This is a strong theme in the literature on situated learning linked to the authenticity of the situation of learning. Greeno (1998) and the Middle School Mathematics Through Applications Project Group, Treffers and Beishuizen (1999) in the Realistic Math Education Project in the Netherlands have used real-world contexts for mathematical problems. As in Collins et al. (1989) model, a key aspect is that the mathematical reasoning within design activities, which is often tacit, is made explicit and supported by conceptual mathematics lessons to scaffold the mathematical understanding needed for the practical tasks. Pratt (2011) suggests that this interlacing of the conceptual and practical is necessary for disciplinary competence in essentially highly abstract subjects such as math, but also recognizes that the world of abstract mathematics is, itself, a legitimate 'situation'; math may legitimately be done for its own sake to develop new mathematical ideas.

The importance of foundational subject knowledge was also a factor in a study of cognitive and motivational effects of

learning *in situ* in secondary schools for typically low-performing pupils (Hauptschule) in Germany. Thiel and Ulber (2004) suggested that deficits in problem solving might be due to pupils lacking well-organized subject knowledge, learning strategies and metacognitive skills so that processing of new information, self-evaluation of the learning process, and the vertical transfer of knowledge necessary for the problem solving were hampered. They concluded that more time and scaffolding would be needed for learning *in situ* to be effective. Yet an earlier study by Ulber (2003) had found that pupils with the poorest preconditions were most favorably affected by a situated approach, but speculated that this might be attributable to them attracting more attention from teachers rather than the community of practice per se. These apparent anomalies in studies indicate that the nature of the relationships within a learning community is fundamental to how successful the learning generated in communities may be. Bennett et al. (2005: 9) conducted a systematic review of context-based approaches in secondary science, which yielded only five studies. They cautiously suggested "context-based approaches provide an effective way to interest and motivate pupils in their science lessons."

Virtual CoP

Over the last decade, problem-based learning as a form of situated learning has particularly been adopted by virtual learning developers. The increasing use of technologies by spatially dispersed communities and by those who physically meet has been studied by Wenger et al. (2005), who concluded that virtual communities performed much the same role through communication tools that include instant messaging, whiteboard and video links, e-mailing, discussion boards, wikis, tweets, and blogs that serve to share and organize community knowing.

In contrast to place-sensitive contexts, there has been a shift in understanding from the notion of learning *in situ* in physical space and real time to the use of situated learning in a virtual context. It is debatable whether significant aspects of learning *in situ* may be lost when the concept is applied to a virtual environment; the physicality of the situation is lost and the individual is no longer exposed to the external dynamic and changing nature of the situation. For example, in the eelgrass mapping project, participants in the project had the opportunity for *impromptu* conversations with passersby, were very much aware of the rising tide and the eelgrass (more-than-human aspects of relationality, which we discuss in a later section), and had the opportunity to go off-task (at least the task as dictated by the teacher). The last point is important; there is space and freedom for learning to emerge in unanticipated ways. This represents a tension in terms of the notion of learning as movement from the periphery to central expertise and refers us back to the idea that such communities are inevitably shaped by wider discourses of power and legitimacy. Additional learning events occurred beyond the stated aims and objectives of the projects. An effective side of learning could be argued to be a significant aspect of learning *in situ*, including a sense of well-being that one gains from working with others or the challenges of maintaining positive relationships within their community. Yet Gee (2008: 35–36) argues, however, that this is not lost in the virtual world and that games still engage the emotions, while Garrison et al. (2000) discuss the need to move beyond cognitive presence in online education to also consider social and teaching presence.

Distance Education

Herrington et al. (2006) claim that the design of online courses can be based on the concepts of situated learning and authentic learning. They suggest that successful distance education should be based upon complex and sustained scenarios (authentic task design), which involve students solving problems in realistic situations "resembling the contexts where the knowledge they are learning will eventually be applied" (3). Yet, despite widespread signaling of situated learning and CoP as theoretical underpinning for virtual learning environments, knowledge may still be viewed as something that may be acquired rather than co-constructed ways of knowing. For example, Campanella et al. (2008) state that

> New technologies can effectively contribute to a movement away from the duplication of traditional instructional methods, both in the classroom and at a distance, toward a more resource-based approach to instruction that no longer emphasizes the teacher as the main source of knowledge. (605)

Furthermore, as we have suggested before, what constitutes 'authenticity' is contested. Generally, it refers to the ordinary practices within the cultural context, so is related to what the norms and expectations are in that setting, and how far the tasks that students are asked to perform align with that. However, work done in school (or any other formal setting) is authentic to that context – i.e., it is 'authentically' schoolwork – and this leads to a related problem of defining the community that learners are central *to*. Formal education is a social practice in its own right and therefore learners' main focus is often on being central (i.e., successful) students, not necessarily central to the 'authentic' context of the problem (see, for example, Pratt and Back (2009) and comments below).

For Hung and Chen (2001), 'situatedness' is inherent in e-learning through its 'anywhere and anytime' access. Working from a situated learning perspective, they set out some key features that characterize e-learning environments as 'situated learning' in their view:

- Web-based so that learners can access the learning environment in their embedded and situated contexts – anywhere and anytime
- portable so that they can be used in context
- focused on tasks and projects, thus enabling learning through doing and reflecting-in-action
- focused on depth over breadth, thus enabling learners to analyze communicative 'speech acts'
- create interdependencies between individuals where novices need more capable peers, thus capitalizing on the zone of proximal development and the diverse expertise within the community
- personalized with tasks that are meaningful to the learner in their context and personalized strategies and content by tracking the learner's history, profile, and progress. (8)

However, the relational aspects of learning in a CoP and movement from the periphery to the center are less clearly delineated by them. Recent advances in Web 2.0 technologies for social networking may be increasing relational capacities in e-learning contexts, which Amin and Roberts (2008) suggest offer virtual spaces that are as productive as actual physical proximity.

Game-Based Learning

Technology has also enabled the creation of near real-life virtual worlds, which give players the feeling of living and assuming a role in a real situation, allowing them to communicate, interact, and cooperate with others *in situ* (Shang et al., 2006). Game-based learning and simulations are increasingly used in educational contexts, building on theories of situated learning, experiential learning, and activity theory and, it is argued, "bring at least some of the benefits of situated learning into the conventional classroom setting" (Lunce, 2006: 40). Educational simulations are seen as providing a method for students to check their understanding of the real world by modeling the structure and dynamics of a conceptual system or a real environment.

Herrington and Kervin (2007) suggest that situated learning concepts underpin the design of many educational computer games, including:

- *Authentic context*: A sustained and complex learning environment that reflects the way the knowledge will ultimately be used and allows sustained exploration because problems are ill-defined, so that students have to *find* as well as *solve* them (Young, 1993).
- *Expert* performance: As in learning through apprenticeship, access to such expert thinking and performances allows students to observe the task before it is attempted.
- *Multiple roles and perspectives*: Affords students' opportunities to investigate multiple ideas, roles, and perspectives.
- *Reflection*: Affords students the opportunity to reflect upon a broad base of knowledge to solve problems, and to predict, hypothesize, and experiment to produce solutions.
- *Collaboration and articulation*: Students work in small groups or pairs and talk about their progress as they go about the task.
- *Coaching and scaffolding*: The teacher coaches and scaffolds rather than instructs.
- *Integrated authentic testing*: Assessment is integrated with the activity rather than a separate test being set.

While these draw on situated learning theory, in common with many applications of the term, the implicit conceptualization of knowledge is acquisitional.

Some Critiques of Situated Learning

As might be apparent from the comments and studies above, many of the examples of situated learning as the term is now used in education retain an individual notion of learning. Lang and Canning (2010) analyzed citations of the term in educational research and argue that despite widespread citation, most are tokenistic and few studies build cumulatively upon the theory. They attribute this as due potentially to two factors: that it offers a tool for thinking rather than the basis for empirical inquiry and that it has an effective 'value-based appeal' (299) for professionals to counterbalance behaviorist and self-directed learning theoretical approaches. If, as we suggest, it is used as one lens among others, its usefulness as a tool for thinking is clear.

Edwards (2005) argues for a relational turn that acknowledges the 'collective competence of a community' and the nature of relationships across which that stretches. She recognizes that learning is constrained and/or afforded (Gibson, 1979) by the system of activity within which individuals work. Analysis of systems that frame the community in which learning is coconstructed have increasingly drawn on Engeström et al.'s CHAT to make such relationships more visible (e.g., Daniels et al., 2007; Williams et al., 2007). Indeed, Engeström and Sannino (2010) argue that expansive learning drawing on activity theory is able to look both outward to the interconnectivity of multiple activity systems and inwards to look at the impacts of subjectivity, embodiment, and affect. Hodkinson et al. (2007: 425) suggest that "the learning of individuals can be understood as a process of becoming, through participation in several different learning cultures over time" and, we would add, places.

Gibson's (1979) theory of affordances refers to the impact of the environment on behavior through perceptions and interaction with what is possible within that context. Apart from physical possibilities, ways of behaving are also constrained by the cultural context as knowledge is co-produced by the learner and the situation in the context of a culture or society, so the position of the learner within the culture is an important factor. Therefore teachers need to respect the knowledge communities from which learners come in order to provide bridging capital that will enable them to access alternative cultural capital (Putnam, 2000). Anderson et al. (1996) argue that while Brazilian street sellers, who correctly calculate the cost of items that they sell in the streets, are unable to answer similar questions at school, it does not mean that arithmetic procedures taught in the classroom cannot be used by shopkeepers, pointing to other apparent examples of 'transfer' of learning such as reading in different contexts. Such tensions are not always fully accounted for in papers drawing on situated learning theory.

Furthermore Miller and Zuengler (2011) argue that not all situated learning may work to the good but may lead to reinforcement of hegemonies in classrooms. They cite their study of English language learners in a high school sheltered civics class and how participation of nonproficient English speakers augmented the linguistic capital of students with stronger English language skills. They argue that resistance to participation may actually work to expose how mainstream practices acquire implicit legitimation that needs to be challenged in diverse societies. As Lave and Wenger themselves (1991: 108–109) comment, "the purpose [of communities of practice] is not to learn *from* talk as a substitute for legitimate peripheral participation; it is to learn *to* talk as a key to legitimate peripheral participation." Thus learning that is regarded as situated must include room for the voice of those at the periphery. Power relationships may limit the extent that participation is possible. Tennant (1997) suggests, "In their

eagerness to debunk testing, formal education and formal accreditation, they do not analyze how their omission [of questions about wider social structures shaping affordances for different people] affects power relations, access, public knowledge and public accountability" (79). Social capital gained in such CoP may differentially exclude members. As an example of an inclusive community, Rogoff, Turkanis, and Bartlett studied how teachers, students, and parents co-developed a school curriculum and pedagogy in a school in Salt Lake City, noting that they prioritized 'instruction that builds on children's interests in a collaborative way' including them in planning learning activities and recognizing that "parents and teachers not only foster children's learning but also learn from their own involvement with children" (Rogoff et al., 2001: 3).

What this demonstrates in part is the potential danger of seeing 'situated learning' as a pedagogical recipe – that is, as a model of how learning can (even should) be developed. However, the use of a situated lens with which to consider learning in different contexts is perhaps more likely to support the exposure of these kinds of inequalities. As we have noted above, such awareness of relationships within systems may be made more apparent through CHAT analyses (Williams et al., 2007).

The question also arises about whether school learning should be 'judged' in terms outside its own context. Sherman (2004) maintains that school science is often unfairly criticized as being a pale imitation of professional science, but she contends it should instead be considered in its own right as a disciplinary exposition of scientific principles and skills rather than a means of creating mini scientists. Jurdak (2006) contends that situated problem solving may lead to an appreciation of both the power and limitations of using mathematics in the real world if the authenticity of the problem tasks is balanced with the mathematical requirements of the curriculum. Both these studies emphasize the wider structural framing of CoPs and the consequent limitations on flexibility within them. Wenger's (1998) original criteria were clear that CoPs arise through collective interest of a community around something that matters to people; this condition of common interest and shared goals may sometimes get overlooked in pedagogical structures that are intended to create (or impose) a learning community. Fueling motivation to participate through interest may present problems perhaps when the agenda for learning is set elsewhere.

A tension also exists between the particular knowing in a community related to specific tasks and a general understanding of ways of working together. Anderson et al. (1996) suggest that transfer can and does occur between learning tasks and is related to the amount of practice and the degree of congruence with other learning. They argue moreover that complex problems do not necessarily lead to better learning as sometimes tasks broken down and addressed in their constituent parts may be more easily tackled, especially by those who find school learning more difficult (see, for example, Forest School pedagogy, which breaks tasks into small steps (Knight, 2010; Maynard, 2007)). On the other hand, Ulber's study (2003) suggests that a gestalt approach may be more appropriate for low-performing students. It is important to bear in mind, as we stated earlier, that situated learning theory is best thought of as a way of looking at learning not a recipe for how learning should happen. Many proponents of situated learning would not claim that group learning is always productive or that individual practice cannot contribute to becoming a more successful participant in social practices, but it does suggest that context and purpose should be considered when planning learning and diverse aims for learning are best supported by a diversity of learning situations (Waite and Pratt, 2011).

Furthermore, one of the side effects of stable CoP is reification (Wenger, 1998: 58) so that ways of doing things become rules and documented procedures, thus the community of a school becomes a less fluid context for certain sorts of learning that are not regarded as legitimate within it. This effect may be compounded by assessment practices that focus on individual rather than group achievements. Most assessment practices in Western societies are geared toward individual attainment and fail to take account of cooperative learning within a standards-driven educational climate, let alone the notion that learning itself can be thought of as distributed across individuals and artifacts. Thus learning, when thought of as socially situated, may be difficult to recognize and credit within the prevalent assessment systems that treat knowledge as a thing that can be owned in order to gain advantage over competitors.

Theorizing Learning Outside the Classroom beyond Situated Learning

This article has looked at situated learning as one theory (among others) that may help to make sense of learning in communities in and out of school. It is essentially a relational view of learning and we have noted how relations between members of communities co-construct understanding through apprenticeship, problem solving, and shared interests. Since its inception, there has been a divergence in the evolution of the concept of situated learning stemming from its interpretation as learning *in situ*, which supports the business of using place, problems, and groups to site learning in legitimate and authentic situations, and learning as situated, which encourages educationalists to think about knowing as nonindividual and context-bound. This divergence leads to some interesting debate about the validity of the virtual for CoP and what authenticity might mean. In this and much existing research and practice that draws on situated learning theory, the relational is read as inherently social, between people.

In the place-based strand or learning *in situ*, the situation (or place) itself is understood as holding a set of expectations or norms of practice and contributing to what can be learned there (see, for example, Massey, 2005). Waite (2013) suggests that scope for ways of being and becoming may be constrained by the cultural density of places, so that schools, for example, that have strongly established norms of practice tend to lead to the reproduction of those practices. Outside the classroom, more ambiguous spaces, where the cultural density of norms and practices is lighter, may afford a greater range of creative and less predictable learning opportunities. In addition, places that have stronger cultural densities associated with other CoP such as students' own home community or a community of

scientists or artists may elicit different opportunities for change and learning (see also Gutiérrez, 2008).

A further line of theory and research that is developing from place-based understandings shifts away from an anthropocentric social focus to include the more-than-human in relational aspects of situated learning theory (Wattchow and Brown, 2011). Critical posthumanist thinking extends the idea of significant members of communities and participation to the more-than-human world (Haraway, 2008). Recognition that human beings are not alone in the world promotes an ecological perspective to learning. The acknowledgment that the more-than-human world has much to teach us is an important new direction in the idea of situated learning for the future of our world.

Implications for Learning Inside and Outside the Classroom

From the acquisitive perspective, spatial connotation of learning *in situ* leads to a focus on the ways in which learning in different 'places' can shape knowledge (as acquisition). This might be greater in efficiency (through motivation, enthusiasm, etc.) or the production of forms of learning that 'fit' the context in which learners might then use it, making learning more 'accessible' and 'fit for purpose.' Taking young learners out to a pond to study frogs will stimulate their learning, and provide access to experiences that they would not otherwise get (many of which might be tacit and embodied) and make it more likely that knowledge is developed in forms that are not 'inert' (Renkl et al., 1996) but are more likely to be usable in the future. In this sense, the hope is that they learn 'better' or 'more' than if they had studied the 'same' knowledge about frogs from a book, including acquiring experiences that would not have been accessible to them (such as feelings, smells, etc.), which form part of an effective development and upon which they can draw in the future (Somerville, 2007).

From the participatory perspective, situation demands a focus on the social activity involved with the connotation that knowledge itself is distributed across the context in which the activity takes place. What learners come to know about frogs is not the same knowledge but learned 'better'; it is a completely different form of 'knowing frogs' (Waite and Pratt, 2011).

The implications of these views of situated learning are numerous, but one or two stand out. Firstly, they raise questions about the nature of learning in different contexts – and hence about problem-based, work-based, place-based, and virtual learning. If it is just a matter of place, and/or task, then these should lead to learning that is different from traditional classrooms; but if the focus is on activity – the practices in which learners and teachers are engaged – then what might matter more is not the place in which learning happens, but the forms of engagement that teachers and learners adopt and the relationships between them. Taking children to a park (Waite and Pratt, 2011) or youngsters to a museum (Peacock and Pratt, 2011) may not change learning significantly if the teachers bring their school relationships with them.

Secondly, if, as suggested earlier, schooling can be seen as a social practice in its own right, learning in alternative situations may not be helpful for schooling if the latter does not offer the same social relationships as the alternative. The much maligned notion of 'teaching to the test' becomes a problem only because of the nature of the test, which represents a very particular socially constructed form of (school) activity. Though space does not allow us to explore this, the implications here are far reaching given the techno-rational version of schooling prevalent at the time of our writing.

See also: Academic Self-Concept and Achievement; Co-constructivism in Educational Theory and Practice; Cognitive Styles and Learning Styles; Communities of Practice; Computer-Assisted Instruction; Continuing Professional Development; Cooperative Learning in Schools; Distance Education; Educational Learning Theory; Environments for Education; Intelligence in Cultural, Social and Educational Context; Intrinsic Motivation: A Cultural Perspective; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Lifelong Learning and Its Support with New Technologies; Open Education: Learning and Teaching; School Learning for Transfer; Sustainability: A Dialog of Values; Vygotsky's Theory of Human Development and New Approaches to Education.

Bibliography

Aguilar, O.M., Krasny, M.E., 2011. Using the communities of practice framework to examine an after-school environmental education program for Hispanic youth. Environmental Education Research 17 (2), 217–233.

Amin, A., Roberts, J., 2008. Knowing in action: beyond communities of practice. Research Policy 37, 353–369.

Anderson, J.R., Reder, L.M., Simon, H.A., 1996. Situated learning and education. Educational Researcher 25 (4), 5–11.

Bennett, J., Hogarth, S., Lubben, F., 2005. A Systematic Review of the Effects of Context-Based and Science-Technology-Society (STS) Approaches in the Teaching of Secondary Science. Department of Educational Studies, University of York, York.

Boyer, L., Roth, W.M., 2006. Learning and teaching as emergent features of informal settings: an ethnographic study in an environmental action group. Science Education 90 (6), 1028–1049.

Brown, J.S., Collins, A., Duguid, P., 1989. Situated cognition and the culture of learning. Educational Researcher 18 (1), 32–41.

Campanella, S., Dimauro, G., Ferrante, A., Impedovo, D., Impedovo, S., Modugno, R., Pirlo, G., Sarcinella, L., Stasolla, E., Trullo, C.A., 2008. Situated learning teaching courses in the university environment: a practical experience. WSEAS Transactions on Advances in Engineering Education 8 (5), 603–612.

Chaiklin, S., Lave, J., 1996. Understanding Practice: Perspectives on Activity and Context (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge University Press, Cambridge.

Collins, A., Brown, J.S., Newman, S.E., 1989. Cognitive apprenticeship: teaching the crafts of reading, writing and mathematics. In: Resnik, L.B. (Ed.), Knowing, Learning and Instruction: Essays in Honor of Robert Glaser. Lawrence Erlbaum, Hillsdale, NJ, pp. 453–494.

Daniels, H., Leadbetter, J., Soares, A., MacNab, N., 2007. Learning in and for cross-school working. Oxford Review of Education 33 (2), 125–142.

Dickens, C., 1854. Hard Times. Wordsworth: Printing Press.

Edwards, A., 2005. Cultural historical activity theory and learning: a relational turn. In: TLRP Annual Conference, University of Warwick, November 2005. http://www.education2.bham.ac.uk/documents/research/liw/edwards_tlrp_05.doc (accessed 21.07.12.).

Engeström, Y., Sannino, A., 2010. Studies of expansive learning: foundations, findings and future challenges. Educational Research Review. 5, 1–24.

Engeström, Y., Miettinen, R., Punamäki, R., 1999. Perspectives on Activity Theory. Cambridge University Press, Cambridge.

Eraut, M., 2000. Non-formal learning, implicit learning and tacit knowledge in professional work. In: Coffield, F. (Ed.), The Necessity of Informal Learning. Policy Press, Bristol, pp. 12–31.

Faircloth, B.S., 2012. 'Wearing a mask' vs. connecting identity with learning. Contemporary Educational Psychology 37 (3), 186–194.

Fusco, D., 2001. Creating relevant science through urban planning and gardening. Journal of Research in Science Teaching 38 (8), 860–877.

Garrison, D.R., Anderson, T., Archer, W., 2000. Critical inquiry in a text-based environment: computer conferencing in higher education. The Internet and Higher Education 2 (2–3), 87–105.

Gee, J.P., 2008. Learning and Games. In: Salen, K. (Ed.), The Ecology of Games: Connecting Youth, Games, and Learning. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. The MIT Press, Cambridge, MA, pp. 21–40.

Gibson, J.J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, MA.

Goswami, U., 2004. Neuroscience and education. British Journal of Educational Psychology 74 (1), 1–14.

Greeno J.G. and Middle School, Mathematics through Applications Project Group, 1998. The situativity of knowing, learning, and research. American Psychologist 53 (1), 5–26.

Gruenewald, D., 2003. The best of both worlds: a critical pedagogy of place. Educational Researcher 32 (4), 3–12.

Gutiérrez, K.D., 2008. Developing a sociocritical literacy in the third space. Reading Research Quarterly 43 (2), 148–164.

Gutiérrez, K.D., Larson, J., 2007. Discussing expanded spaces for learning. Language Arts 85 (1), 69–77.

Haraway, D., 2008. When Species Meet. University of Minnesota Press, Minneapolis and London.

Herrington, J., Kervin, L., 2007. Authentic learning supported by technology: 10 suggestions and cases of integration in classrooms. Educational Media International 44 (3), 219–236.

Herrington, J., Reeves, T.C., Oliver, R., 2006. Authentic tasks online: a synergy among learner, task, and technology. Distance Education 27 (2), 233–247.

Hildreth, P.M., Kimble, C., October 2002. The duality of knowledge. Information Research 8 (1). http://informationr.net/ir/8-1/paper142 (accessed 27.07.12.).

Hodkinson, P., Biesta, G., James, D., 2007. Understanding learning cultures. Educational Review 59 (4), 415–427.

Hoskins, B., Janmaat, J.G., Villalba, E., 2012. Learning citizenship through social participation outside and inside school: an international, multilevel study of young people's learning of citizenship. British Educational Research Journal 38 (3), 419–446.

Hung, D., Chen, D.T., 2001. Situated cognition, Vygotskian thought and learning from the communities of practice perspective: implications for the design of web-based e-learning. Education Media International 38 (1), 4–11.

Jurdak, M., 2006. Contrasting perspectives and performance of high school students on problem solving in real world situated, and school contexts. Educational Studies in Mathematics 63 (3), 283–301.

Knight, S., 2010. Forest Schools and Outdoor Learning in the Early Years. Sage Publications, London.

Lang, I., Canning, R., 2010. The use of citations in educational research: the instance of the concept of situated learning. Journal of Further and Higher Education. 34 (2), 291–301.

Lave, J., 1988. Cognition in Practice. Cambridge University Press, Cambridge.

Lave, J., 1996. The practice of learning. In: Chaiklin, S., Lave, J. (Eds.), Understanding Practice: Perspectives on Activity and Context. Cambridge University Press, Cambridge.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge.

Lea, M.R., 2005. 'Communities of practice' in higher education: useful heuristic or educational model? In: Barton, D., Tusting, K. (Eds.), Beyond Communities of Practice. Language, Power and Social Context. Cambridge University Press, Cambridge, pp. 180–197.

Lee, C., 2007. Culture, Literacy, and Learning. Teachers' College Press, New York.

Lunce, L.M., 2006. Simulations: bringing the benefits of situated learning to the traditional classroom. Journal of Applied Educational Technology 3 (1). Spring/Summer 2006, 37–45.

Massey, D., 2005. For Space. Sage Publications, London.

Maynard, T., 2007. Forest schools in Great Britain: an initial exploration. Contemporary Issues in Early Childhood 8 (4), 320–331.

Miller, E.R., Zuengler, J., 2011. Negotiating access to learning through resistance to classroom practice. Modern Language Journal 95, 130–147.

Nadelson, L.S., Jordan, J.R., 2012. Student attitudes toward and recall of outside day: an environmental science field trip. Journal of Educational Research 105 (3), 220–231.

National Apprenticeship Service, 2008. http://www.apprenticeships.org.uk/About-Us/National-Apprenticeship-Service.aspx (accessed 21.07.12.).

Nunes, T., Schliemann, A.D., Carraher, D.W., 1993. Street Mathematics and School Mathematics. Cambridge University Press, Cambridge.

Peacock, A., Pratt, N., 2011. How young people respond to learning spaces outside school: a sociocultural perspective. Learning Environments Research 14 (1), 11–24.

Piaget, J., 1952. The Origins of Intelligence in Children. International Universities Press, New York.

Pratt, N., 2011. Mathematics outside the classroom. In: Waite, S. (Ed.), Children Learning Outside the Classroom. From Birth to Eleven. Sage Publications, London, pp. 80–93.

Pratt, N., Back, J., 2009. Spaces to discuss mathematics: communities of practice on an online discussion board. Research in Mathematics Education 11 (2), 115–130.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.

Renkl, A., Mandl, H., Gruber, H., 1996. Inert knowledge: analyses and remedies. Educational Psychologist 31, 115–121.

Rogoff, B., 1990. Apprenticeship in Thinking: Cognitive Development in Social Context. Oxford University Press, Oxford.

Rogoff, B., Turkanis, C.G., Bartlett, L. (Eds.), 2001. Learning Together: Children and Adults in a School Community. Oxford University Press, New York.

Sadler, T.D., Burgin, S., McKinney, L., Ponjuan, L., 2010. Learning science through research apprenticeship: a critical review of the literature. Journal of Research in Science Teaching. 47 (3), 235–256.

Seaman, J., 2007. Taking *things* into account: learning as kinesthetically-mediated collaboration. Journal of Adventure Education and Outdoor Learning 7 (1), 3–20.

Sfard, A., 1998. On two metaphors for learning and the dangers of choosing just one. Educational Researcher 27 (2), 4–13.

Shang, J., Jong, M.S.Y., Lee, F.L., Lee, J.H.M., 2006. VISOLE: a new game-based situated learning paradigm. In: Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06). http://www.computer.org/comp/proceedings/icalt/2006/2632/00/263201082.pdf (accessed 12.07.12.).

Sherman, W., 2004. Science studies, situatedness, and instructional design in science education: a summary and critique of the promise. Canadian Journal of Science, Mathematics and Technology Education 4 (4), 443–465.

Somerville, M., 2007. Becoming frog: a primary school place pedagogy. In: Paper Presented at Australian Association for Research in Education conference 2007. www.aare.edu.au/07pap/som07443.pdf (accessed 27.07.12.).

Tennant, M., 1997. Psychology and Adult Learning. Routledge, London.

Thiel, F., Ulber, D., 2004. Cognitive and motivational effects of situated learning in schools. In: Paper Presented at the European Conference on Educational Research, University of Crete, 22–25 September 2004. http://www.leeds.ac.uk/educol/documents/00003840.htm (accessed 19.07.12.).

Tidball, M.E., Krasny, K.G., 2007. From risk to resilience: what role for community greening and civic ecology in cities? In: Wals, A. (Ed.), Social Learning towards a Sustainable World: Principles, Perspectives, Praxis. Wageningen Academic Publishers, The Netherlands, pp. 149–164.

Treffers, A., Beishuizen, M., 1999. Realistic mathematics education in the Netherlands. In: Thompson, I. (Ed.), Issues in Teaching Numeracy in Primary Schools. Open University Press, Buckingham, pp. 27–38.

Ulber, D., 2003. Differential effects of situated learning on secondary school students. In: Paper Presented at the European Conference on Educational Research, University of Hamburg, 17–20 September 2003. http://www.leeds.ac.uk/educol/documents/00003498.htm (accessed 19.07.12.).

Vygotsky, L.S., Cole, M., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, London.

Waite, S., 2013. 'Knowing your place in the world': how place and culture support and obstruct educational aims. Cambridge Journal of Education 43 (4), 413–434.

Waite, S., Pratt, N., 2011. Theoretical perspectives on learning outside the classroom: relationships between learning and place. In: Waite, S. (Ed.), Children Learning Outside the Classroom. From Birth to Eleven. Sage Publications, London, pp. 1–18.

Wattchow, B., Brown, M., 2011. A Pedagogy of Place: Outdoor Education for a Changing World. Monash University Press, Clayton, Victoria.

Wenger, E., 1998. Communities of Practice: Learning, Meaning and Identity. Cambridge University Press, Cambridge.

Wenger, E., White, N., Smith, J.D., Rowe, K., May 2005. Technology for communities. Communities 2, 1–15.

Williams, J., Davis, P., Black, L., 2007. Sociocultural and cultural-historical activity theory perspectives on subjectivities and learning in schools and other educational contexts. International Journal of Educational Research 46 (1–2), 1–7.

Young, F.Y., 1993. Instructional design for situated learning. Educational Technology Research and Development 41 (1), 43–57.

Situational Judgment Test

Filip Lievens and Britt De Soete, Ghent University, Ghent, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we give an overview of situational judgment tests (SJTs) as selection instruments. Their history, basic characteristics, and development are presented. The available research evidence regarding their reliability, construct-related validity, criterion-related validity, incremental validity, subgroup differences, and test-taker perceptions is also reviewed. As a general conclusion, the increasing popularity of SJTs in personnel selection seems to be accredited to their potential to capture a variety of constructs and for different purposes. Additionally, SJTs are able to predict several job-related and/or academic criteria while at the same time offering prospects permitting to select for diversity.

History and Definition

During World War II, military selection psychologists were in need of a tool to select competent soldiers to join the armed forces. They developed a job test that consisted of detailed and realistic descriptions of challenging military situations. All descriptions were situations that armed forces were likely to encounter while on the job. After reading each situation, recruits were presented with several potential reactions to the given threat or challenge and they were asked which reaction they considered the most effective response (Northrop, 1989). The instrument turned out to be a success. On the one hand, it gave a realistic job preview of what was to come, thereby discouraging recruits with an unfavorable person–organization fit and lowering attrition rates for the army. On the other hand, the tool enabled to measure recruits' judgment skills in job-related settings, thereby significantly facilitating competent new soldier selection. The aforementioned instrument can be considered one of the first situational judgment tests (SJTs). After World War II, several similar tests were designed to capture supervisory potential (e.g., Bruce, 1974; Cardall, 1942; File, 1945; Greenberg, 1963). In 1990, Motowidlo and colleagues framed the SJT as a new alternative measurement procedure for personnel selection (Motowidlo et al., 1990) and thereby reinvigorated interest in SJTs among scientists and practitioners. SJTs present test-takers with realistic job situations, followed by potential response options out of which candidates have to select the most appropriate response (Motowidlo et al., 1990). SJTs are considered measurement tools that aim to capture job-related competencies and skills (Lievens et al., 2008). **Figure 1** shows an example SJT item.

Similar to assessment centers and work samples, SJTs are simulation-based instruments. Simulations are based on the

You and someone from another department are jointly responsible for coordinating a project involving both departments. This other person is not carrying out his share of the responsibilities. What would you most likely/least likely do?

1. Discuss the situation with your manager and ask him to take it up with the other person's manager.
2. Remind him that you need his help and that the project won't be completed effectively without a full team effort from both of you.
3. Tell him that he is not doing his share of the work, that you will not do it all yourself, and that if he does not start doing more, you'll be forced to take the matter to his manager.
4. Try to find out why he is not doing his share and explain to him that this creates more work for you and makes it harder to finish the project.
5. Get someone else from his department to help with the project.

Figure 1 Example situational judgment test item. Motowidlo, S.J., Hooper, A.C., Jackson, H.L., 2006. A theoretical basis for situational judgment tests. In: Weekley, J., Ployhart, R. (Eds.), Situational Judgment Tests. Jossey-Bass, San Francisco, pp. 57–81.

behavioral consistency logic (Lievens and De Soete, 2012). That is, the assumption that candidates' performance during the selection procedure will be consistent with their future performance on the job. The difference between SJTs and other simulation-based measurement instruments lies in their level of fidelity. Fidelity can be defined as the extent to which the selection procedure mirrors the actual job situation (Callinan and Robertson, 2000). As assessment centers and work samples require actual behavior during the selection phase, such instruments can be considered high-fidelity simulations. High-fidelity simulations are more expensive to administer and are therefore generally used in small samples, during later selection stages. In contrast, SJTs are described as low-fidelity simulations (Motowidlo et al., 1990) as they do not require test-takers to display actual behavior but instead confront them with written descriptions of realistic job situations. As a result, SJTs can be administered to large applicant groups in preliminary selection stages. Since their reintroduction by Motowidlo and colleagues in 1990, SJTs have become attractive selection instruments for practitioners who are looking for cost-effective instruments for measuring a wide variety of predominantly interpersonally oriented constructs.

SJT Development

The development of SJTs typically consists of three stages (Motowidlo et al., 1990). In the first stage, the stimulus material is developed. It is advisable to start this phase with a job analysis, so that the knowledge, skills, abilities, and other characteristics (KSAOs) are identified, which are considered to be crucial for job performance. For each of these competencies, critical incidents of work situations are gathered from subject matter experts (i.e., incumbents, their supervisors, clients) or from archival sources. As a next step, test developers select the best nonredundant critical incidents from the total pool and rewrite them into test items of similar length and format.

Second, the response options are developed. To this end, all items are presented to a different group of subject matter experts or to inexperienced employees, which are asked to formulate possible responses to the given job situations. In this stage, test developers aim to collect a satisfactory amount of responses for every single item, and these responses should capture a wide range of effectiveness.

As a third and last step, the SJT's scoring key is developed. Test developers mostly opt for a rational or an empirical scoring key. When developing a rational scoring key, subject matter experts are asked to either identify the best and the worst response options or to rate all responses per item on their effectiveness. Unlike the rational scoring key, an empirical scoring key does not use expert judgments. Instead, the SJT is administered among a large pool of incumbents, whose responses are then linked to a criterion (e.g., job performance). SJT responses that are mostly chosen by high performing employees are labeled as 'correct' or 'highly effective,' whereas the opposite takes place for SJT responses that are not endorsed by high-performing employees or selected by low-performing individuals.

SJT Design Considerations

Although most SJTs are designed according to the predefined steps described above, test developers and their clients also face various design decisions, which may impact on the specific outlook of the SJT developed. These decisions are typically driven by the available resources and the objectives (hiring, promotion, training, recruitment, etc.) one wants to accomplish with the SJT.

A first set of decisions concerns the characteristics of the item. A first aspect of the item stem that varies between SJTs is the *item length*. Some SJTs provide test-takers with a simple and short situational description (e.g., 'A colleague in your team dodges her duties and you have to take over a lot of her tasks.'), whereas others present the participant with detailed descriptions of the job situation and its background. Similarly, items may vary in their level of *contextualization*. Some SJTs are built to measure KSAOs for specific jobs, whereas others are meant to measure generic competencies or skills across several jobs on the same level. As a result, SJTs in the former category will often be characterized by a higher degree of contextualization than the latter types. Highly contextualized SJT items provide specified information about the organizational setting and job context, which is illustrated by mentioning job-specific equipment or by using job terminology. A final item-related difference between SJTs is their level of *interactivity*. Whereas traditional SJTs present each test-taker with the same (sequence of) items, interactive or so-called 'branched' or 'nested' SJTs take into account the test-taker's response to former items to decide, which items will be presented subsequently (Olson-Buchanan et al., 1998). Accordingly, test-takers are confronted with the consequences of their response actions. For example, if a participant decides to fire an unproductive employee in one item, the next item may deal with an uproar in the team as a consequence of the dismissal. The main advantage of these branched SJTs is that they permit to mirror the dynamics of an actual interaction, while maintaining a certain degree of standardization (Lievens et al., 2008).

Design decisions should also be made regarding the SJT's *stimulus and response modality*. Modality refers to the way the information is presented to the candidate (for stimuli) or the manner in which the candidates are required to answer (for responses). Regarding stimulus modality, SJTs traditionally have a text-based format that presents test-takers with written descriptions of job situations. As calls have been made to develop more realistic measurement instruments (e.g., Lane and Stone, 2006), i.e., instruments that show a greater resemblance with the actual job for which they are selecting applicants, written SJT stimuli can be replaced by video-based or multimedia stimuli (Chan and Schmitt, 1997; Weekley and Jones, 1997). In such SJTs, test-takers are presented with short videos of job-related situations. At a critical point, the video freezes and the test-taker is subsequently required to select the most appropriate response to the presented situation.

Regarding response modality, SJTs traditionally have a text-based multiple choice format, which present test-takers with a predetermined set of written responses. However, recently also alternative response modalities have made their entrance in the SJT domain. For instance, Kanning et al. (2006) experimented with video-based instead of written response options.

More recently, Crook and colleagues introduced the single-response SJT (Crook et al., 2011; Motowidlo et al., 2009). Single-response SJTs confront test-takers with only one response option, which they have to rate on its effectiveness by means of a Likert scale. The main advantage of single-response SJTs is their less labor-intensive test design. In contrast to traditional multiple choice SJTs, single-response SJTs require only one group of subject matter experts for test design, which are asked to come up with critical incidents that struck them as extremely effective or ineffective. Based on this information, both item stems and responses can be derived, and the second development stage of traditional SJTs can be skipped.

A third decision concerns the SJT's *response instructions*. Both knowledge-based and behavioral-based response instructions are frequently used (McDaniel et al., 2007). Knowledge-based response instructions ask the candidate to display their knowledge of the responses' effectiveness by selecting the best/worst response option ('What is the best/worst answer?') or by rating each response option on its perceived effectiveness. As a result, SJTs with knowledge-based instructions are measures of maximal performance. In contrast, SJTs with behavioral-based response instructions ask candidates to report how they would respond in the presented situation ('What are you most likely to do?') and are therefore considered as typical performance measures.

Finally, a fourth set of decisions deals with the SJT's *scoring key*. As has been mentioned earlier, empirical and rational scoring keys are the most common choices for scoring keys. Therefore, hybrid forms have also been developed. For instance, an extant empirical scoring key is given to experts to make sure that 'it makes sense.' Alternatively, one might start with an expert-scoring key and later on validate it empirically.

An Evidence-Based Evaluation of SJTs

Since Motowidlo launched the SJT as an alternative measurement procedure for personnel selection in the early 1990s (Motowidlo et al., 1990), research on SJTs has made great strides forward. Today, over 60 studies have been devoted to the development and psychometric properties of SJTs. On the basis of the available research, we review below SJTs in terms of their reliability, construct-related validity, criterion-related validity, incremental validity, subgroup differences, and test-taker perceptions.

Are SJT Scores Reliable?

To assess whether or not SJTs scores are reliable, most prior research has focused on internal consistency reliability. For example, a meta-analysis of McDaniel et al. (2001), based on 39 different SJTs and over 10 000 participants, found internal consistency reliability coefficients ranging from 0.43 to 0.94. A more recent study of Catano et al. (2012) revealed a corrected weighted mean internal consistency of 0.46 based on 56 coefficients.

The variety in internal consistency reliability coefficients is in the first place a function of the instrument's length. SJT scores based on more items generally demonstrate higher internal consistency reliability. In addition, Ployhart and Ehrhart (2003) found that response instructions may serve as a second driver of internal consistency variability in SJTs. More specifically, SJT response instructions requiring the participant to rate each response on its effectiveness by means of a Likert-like scale resulted in the highest internal consistency reliability coefficients. Instructions that asked participants to choose two response options (e.g., 'What is your most/least likely response?' or 'What is the most/least effective response?') displayed somewhat lower internal consistency reliability coefficients, and instructions requiring participants to select one single response option (e.g., 'What is your most likely response?' or 'What is the most effective response?') resulted in the lowest internal consistency reliability coefficients.

Besides test length and response instructions, the often observed heterogeneous nature of many SJTs is probably one of the main reasons for their low internal consistency reliability (Lievens et al., 2008). As most SJTs are heterogeneous at the item level and aim to capture a plethora of constructs, internal consistency may therefore not be the most preferred means to evaluate SJT scores' reliability. Instead, researchers have proposed that alternate-form reliability or test–retest reliability may be more appropriate in the case of SJTs (Whetzel and McDaniel, 2009). Unfortunately, both of these reliabilities are rarely reported. One recent exception is the study of Catano et al. (2012), which reported test-retest reliabilities ranging from 0.66 to 0.82.

What Do SJTs Measure?

As SJTs have been used to capture a wide variety of constructs, answering this question is a challenge. Over the past years, SJTs have been developed to measure several competencies as diverse as entry-level managerial skills (Motowidlo et al., 1990), leadership skills (Oostrom et al., 2012), team work skills (Prewett et al., 2013), emotional intelligence (Libbrecht and Lievens, 2012; Libbrecht et al., in press; Sharma et al., 2013), interpersonal skills of medical students (Lievens and Sackett, 2012), aviation pilot judgment (Hunter, 2003), personal initiative (Bledow and Frese, 2009), and integrity (Becker, 2005; De Meijer et al., 2010).

In an attempt to develop a content-based typology, Christian et al. (2010) identified and clustered all construct domains assessed by SJTs. They found that about 70% of the extant SJTs aim to capture either leadership or interpersonal skills. To assess which factors determine SJT performance, several meta-analyses have demonstrated a relatively high correlation between SJT performance on the one hand and cognitive ability measures or personality traits on the other hand. In terms of SJTs scores' relation with cognitive ability scores, meta-analyses reveal correlation coefficients ranging from $r = 0.32$ (McDaniel et al., 2001) to $r = 0.46$ (McDaniel et al., 2007). Four moderating factors for the relationship between SJT performance and cognitive ability can be derived from the literature. First, SJTs that are developed on the basis of a thorough job analysis seem to display higher correlations with cognitive ability than SJTs that do not start from a job analysis (McDaniel et al., 2001). Second, the more detailed the SJT question, the lower the SJT's correlation with cognitive ability appears to be (McDaniel et al., 2001). Third, several studies have emphasized the importance of the stimulus

format for the constructs assessed. For instance, SJTs with a video-based stimulus format demonstrate a lower correlation with performance on cognitive instruments than paper-and-pencil SJTs (Chan and Schmitt, 1997; Lievens and Sackett, 2006). Finally, in their meta-analysis in 2007, McDaniel et al. (2007) identified the response instructions as the fourth important influencer of cognitive saturation. That is, SJTs with knowledge-based response instructions display higher correlations with cognitive ability than SJTs with behavioral tendency instructions.

SJT scores have also found to be correlated with personality. McDaniel et al. (2007) demonstrated that an SJT's relation with personality traits is a function of its response instructions. That is, SJTs with behavioral response instructions show significantly higher correlations with Agreeableness ($r = 0.37$), Conscientiousness ($r = 0.34$), and Emotional Stability ($r = 0.35$) as compared to SJTs with knowledge-based response instructions ($r = 0.19$, $r = 0.24$, and $r = 0.12$, respectively).

What Do SJTs Predict?

One of the most important evaluation criteria when deciding whether or not to use selection instruments refers to the extent to which they are related to job-related performance domains. Meta-analytic research has confirmed the expectation that SJTs are valuable predictors of job performance. A first meta-analysis on the criterion-related validity of SJT scores revealed a validity coefficient of $r = 0.34$ for predicting job performance (McDaniel et al., 2001). As one of the most important moderators of criterion-related validity, McDaniel and colleagues demonstrated the presence or absence of a job analysis to be an influential factor. SJTs that were based on a thorough job analysis displayed higher validity than SJTs that did not use a job analysis as their starting point. Several years later, a second meta-analysis was undertaken, which included more data than the former and which revealed an estimated population criterion-related validity coefficient of $r = 0.26$ (McDaniel et al., 2007). The most recent meta-analysis on criterion-related validity of SJTs thus far was published in 2010. On the basis of 84 studies, Christian et al. (2010) found validity coefficients of $r = 0.38$ for SJTs measuring teamwork, $r = 0.28$ for SJTs on leadership, and $r = 0.25$ for SJTs capturing interpersonal skills. Additionally, two extra moderators of SJT criterion-related validity were identified. First, the importance of careful predictor-criterion matching was emphasized as SJTs measuring specific competencies (e.g., interpersonal skills) demonstrated the highest validity for predicting matching criteria domains (e.g., interpersonal job performance, see also Lievens et al., 2005). Second, it was found that an SJT's stimulus modality may influence its criterion-related validity. More specifically, video-based SJTs displayed higher criterion-related validity coefficients than text-based SJTs for predicting interpersonal skills (see also Lievens and Sackett, 2006).

Apart from showing sufficient criterion-related validity in employment settings, SJT scores have also found to be useful for predicting academic success. For example, Lievens and colleagues repeatedly demonstrated that SJTs are good predictors of academic performance among student physicians (Lievens, 2013; Lievens et al., 2005; Lievens and Sackett, 2006, 2012). In a similar vein, Oswald et al. (2004) developed an SJT for predicting student performance. Their SJT aimed to capture 12 dimensions of college student performance and was successful in predicting academic success.

Finally, SJT scores have demonstrated consistent incremental validity over and above scores on more established selection instruments (Clevenger et al., 2001). This implies that SJTs succeed in predicting unique variance in criterion performance, which cannot be accounted for by other predictors. Various studies have demonstrated incremental validity of SJTs for predicting job or academic performance over cognitive ability, job knowledge, job experience, and personality (Chan and Schmitt, 2002; Clevenger et al., 2001; Lievens and Patterson, 2011; McDaniel et al., 2001, 2007; O'Connell et al., 2007; Oswald et al., 2004; Weekley and Jones, 1997, 1999). In their meta-analysis, McDaniel et al. (2007) estimated the incremental validity of SJTs over cognitive ability between 3 and 5%, with somewhat higher incremental validity coefficients when behavioral instead of knowledge-based response instructions were used. Additionally, it was found that SJTs provide incremental validity over personality, varying from 6 to 7%, with the highest incremental validity for SJTs with knowledge-based response instructions. Over a combination of cognitive ability and personality, SJTs provided incremental validity of 1–2%. Besides response instructions, also the criterion type has been proven to be an influential factor of incremental validity variability (O'Connell et al., 2007). For instance, O'Connell and colleagues found incremental validity for SJTs over cognitive ability but not over personality for predicting contextual performance, whereas this was not the case for predicting task performance. More research is needed to confirm this finding on a more general level, though.

Recently, conceptual progress has been made to explain why SJTs are valid predictors. According to Motowidlo and Beier (2010), SJTs measure general as well as specific knowledge of the costs and the consequences of job-related actions. General knowledge refers to implicit knowledge about the relationships between expressions of personality traits and effective job performance (i.e., implicit trait policies). That is, such general knowledge refers to people's judgments about the costs and benefits of engaging in courses of action as response to SJT situations. Specific knowledge is different from general knowledge because it is limited to specific job situations, which may include exceptions to successful trait policies (e.g., specific job situations where assertive reactions are more valued than agreeable reactions). The combination of both knowledge types captured by SJTs is assumed to be predictive of effective performance.

What about SJTs and Diversity?

An additional reason for the popularity of SJTs in the field of personnel selection is that they are introduced as 'alternative' predictors, which might display smaller subgroup differences than cognitive ability tests. So far, the research evidence is a bit more mixed. Ethnic subgroup differences on SJTs have varied from approximately one standard deviation (Chan and Schmitt, 1997) to almost zero (Olson-Buchanan et al., 1998), with Caucasians obtaining higher scores than black ($d = 0.38$), Hispanic ($d = 0.24$), Asian ($d = 0.29$), and

European minority ($d = 0.38$) test-takers (De Meijer, 2008; Whetzel et al., 2008).

Some factors may explain the great variety in subgroup differences assonated with SJT scores. First, meta-analytic research identified cognitive loading as one of the most important drivers of ethnic subgroup differences in SJT performance. Cognitive loading refers to the extent that SJT performance correlates with performance on a cognitive ability test. Similar to assessment centers and work samples (Bobko et al., 2005; Dean et al., 2008; Roth et al., 2008), SJTs with a higher cognitive loading display substantially larger ethnic subgroup differences (Roth et al., 2013; Whetzel et al., 2008). Second, the personality loading, i.e., the correlation between the SJT and each of the Big Five personality factors, has been identified as a smaller driver of ethnic subgroup differences in SJTs, so that black–white and Asian–white differences in SJT performance are smaller when the SJT displays a higher correlation with emotional stability and Hispanic–white differences are smaller when the SJT displays a higher correlation with conscientiousness and agreeableness (Whetzel et al., 2008). Third, SJT response instructions have been shown to influence ethnic performance differences. Knowledge-based response instructions ('What is the best response?') generally lead to larger ethnic subgroup differences than behavioral tendency response instructions ('How would you respond?'), which is most likely due to the greater cognitive loading of SJTs with knowledge-based response instructions (Whetzel et al., 2008). A fourth moderator concerns the SJT response process. SJTs that require candidates to rate each response option's effectiveness by means of a Likert scale are more susceptible to ethnic subgroup differences because SJT performance is influenced by the white–black difference in extreme scoring preferences. By controlling for elevation and scatter in responses, black–white performance differences on this SJT type are substantially reduced. In addition, this correction is particularly promising as it simultaneously improves validity (McDaniel et al., 2011).

In terms of gender differences in SJT performance, a meta-analysis demonstrated that female test-takers on average perform slightly higher than male test-takers ($d = -0.11$; Whetzel et al., 2008). Larger female advantages occur when the SJT is more correlated with agreeableness and conscientiousness (Weekley et al., 2004; Whetzel et al., 2008).

Do Candidates Like SJTs?

A last aspect for evaluating SJTs relates the test perceptions. In general, simulation instruments, such as assessment centers, work samples, and SJTs, are more favorably received by applicants than cognitive ability tests, personality inventories, biodata, and integrity tests (Hausknecht et al., 2004; Oostrom et al., 2010). Some researchers have identified specific SJT factors, which may influence test perceptions or attitudes. One of the most important drivers of test perceptions in SJTs thus far, seems to be the stimulus modality. Richman-Hirsch et al. (2000) examined the effect of stimulus modality on test perceptions and attitudes for three content-wise identical conflict management SJTs. The first SJT had a paper-and-pencil format. The second SJT was identical to the first but used a computer screen and an automatic page turner to display the information. The third SJT presented test-takers with exactly the same scenarios but in video format. Applicants watched videos of job-related conflict situations and were asked to select the best response out of a set of four written options. Results demonstrated that test-takers reported significantly more favorable face validity perceptions and test attitudes for the video SJT as compared to the other two formats. Furthermore, there was no difference in reported perceptions and attitudes between the paper-and-pencil SJT and the computerized version, which suggests that computerizing test content is not sufficient to influence test perceptions. Along the same lines, Chan and Schmitt (1997) compared a written SJT with a content-wise identical video variant. They found significantly higher face validity perceptions for the latter SJT. Finally, Kanning et al. (2006) examined which SJT factors improve test perceptions. They discovered that changing the stimulus or response modality from written to video format alone was not enough to increase test perceptions, but that interactive SJTs with video stimuli as well as video response options received the most favorable test perceptions.

Epilogue

In the last two decades, SJTs have become a popular selection instrument that is widely used by practitioners and intensively studied by researchers. Their popularity in the field of personnel selection can largely be accredited to the potential of SJTs to capture a variety of constructs, in diverse settings, and for different purposes. Additionally, research has repeatedly proven that SJTs are able to predict several job-related and/or academic criteria while at the same time offering prospects permitting to select for diversity.

Moreover, the future of SJTs for research and practice is looking bright. At the risk of being self-promoting, below we give a brief overview of some of our own research priorities in the next years. One avenue that we see as increasingly important is the use of SJTs in cross-cultural settings. To this end, research is needed to examine the intercultural transportability and robustness of SJTs (see Lievens, 2006; for an overview). Second, calls have been made to increase our conceptual understanding of the different components of SJTs and their effects on key selection outcomes. Here we advocate the use of a building block approach (e.g., Arthur and Villado, 2008; Lievens et al., in press). That is, instead of treating SJTs as holistic entities, the SJT method can be conceptualized as a combination of various predictor method factors (i.e., response instructions, stimulus format, response format, etc.). By keeping other factors constant, researchers are able to identify the impact of specific SJT factors on the criterion of interest (e.g., construct-related validity, criterion-related validity, subgroup differences, Chan and Schmitt, 1997; Lievens et al., in press; Lievens and Sackett, 2006). Finally, we believe that the SJT domain would benefit from incorporating novel presentation and response formats. Successful examples are 3D animated and avatar-based SJTs (Fetzer, 2012). Another example is the development of video-based SJTs with an open-ended response modality, namely

either a written (responding in a text box) or behavioral (responding through a Webcam) response format (Lievens et al., in press). Clearly, there exist a plethora of opportunities to create new SJT formats and hybrid SJTs in the future. Both practitioners and researchers should join forces to implement and examine them in the future.

See also: Personality Assessment, Faking and; Personnel Selection, Psychology of; Social Intelligence and Competencies.

Bibliography

Arthur, W., Villado, A.J., 2008. The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. Journal of Applied Psychology 93, 435–442. http://dx.doi.org/10.1037/0021-9010.93.2.435.

Becker, T.E., 2005. Development and validation of a situational judgment test of employee integrity. International Journal of Selection and Assessment 13, 225–232. http://dx.doi.org/10.1111/j.1468-2389.2005.00319.x.

Bledow, R., Frese, M., 2009. A situational judgment test of personal initiative and its relationship to performance. Personnel Psychology 62, 229–258. http://dx.doi.org/10.1111/j.1744-6570.2009.01137.x.

Bobko, P., Roth, P.L., Buster, M.A., 2005. Work sample selection tests and expected reduction in adverse impact: a cautionary note. International Journal of Selection and Assessment 13, 1–10. http://dx.doi.org/10.1111/j.0965-075X.2005.00295.x.

Bruce, M.M., 1974. Examiner's Manual: Supervisory Practices Test, rev. ed. Author, Larchmont, NY.

Callinan, M., Robertson, I.T., 2000. Work sample testing. International Journal of Selection and Assessment 8, 248–260. http://dx.doi.org/10.1111/1468-2389.00154.

Catano, V.M., Brochu, A., Lamerson, C.D., 2012. Assessing the reliability of situational judgment tests used in high-stakes situations. International Journal of Selection and Assessment 20, 333–346. http://dx.doi.org/10.1111/j.1468-2389.2012.00604.x.

Cardall, A.J., 1942. Preliminary Manual for the Test of Practical Judgment. Science Research, Chicago.

Chan, D., Schmitt, N., 1997. Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. Journal of Applied Psychology 82, 143–159. http://dx.doi.org/10.1037/0021-9010.82.1.143.

Chan, D., Schmitt, N., 2002. Situational judgment and job performance. Human Performance 15, 233–254. http://dx.doi.org/10.1207/S15327043HUP1503_01.

Christian, M.S., Edwards, B.D., Bradley, J.C., 2010. Situational judgement tests: constructs assessed and a meta-analysis of their criterion-related validities. Personnel Psychology 63, 83–117. http://dx.doi.org/10.1111/j.1744-6570.2009.01163.x.

Clevenger, J., Pereira, G.M., Wiechmann, D., Schmitt, N., Harvey, V.S., 2001. Incremental validity of situational judgment tests. Journal of Applied Psychology 86, 410–417. http://dx.doi.org/10.1037/0021-9010.86.3.410.

Crook, A.E., Beier, M.E., Cox, C.B., Kell, H.J., Hanks, A.R., Motowidlo, S.J., 2011. Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. International Journal of Selection and Assessment 19, 363–373. http://dx.doi.org/10.1111/j.1468-2389.2011.00565.x.

Dean, M.A., Bobko, P., Roth, P.L., 2008. Ethnic and gender subgroup differences in assessment center ratings: a meta-analysis. Journal of Applied Psychology 93, 685–691. http://dx.doi.org/10.1037/0021-9010-93.3.685.

De Meijer, L.A.L., 2008. Ethnicity Effects in Police Officer Selection: Applicant, Assessor, and Selection-method Factors (Unpublished doctoral dissertation). Erasmus University, Rotterdam.

De Meijer, L.A.L., Born, M.P., van Zielst, J., van der Molen, H.T., 2010. Construct-driven development of a video-based situational judgment test for integrity a study in a multi-ethnic police setting. European Psychologist 15, 229–236. http://dx.doi.org/10.1027/1016-9040/a000027.

Fetzer, M.S., April 2012. Current research in advanced assessment technologies. In: Symposium Presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

File, Q.W., 1945. The measurement of supervisory quality in industry. Journal of Applied Psychology 29, 323–337. http://dx.doi.org/10.1037/h0057397.

Greenberg, S.H., 1963. Supervisory Judgment Test Manual. U.S. Civil Service Commission, Washington, DC.

Hausknecht, J.P., Day, D.V., Thomas, S.C., 2004. Applicant reactions to selection procedures: an updated model and meta-analysis. Personnel Psychology 57, 639–683. http://dx.doi.org/10.1111/j.1744-6570.2004.00003.x.

Hunter, D.R., 2003. Measuring general aviation pilot judgment using a situational judgment technique. International Journal of Aviation Psychology 13, 373–386. http://dx.doi.org/10.1207/S15327108IJAP1304_03.

Kanning, U.P., Grewe, K., Hollenberg, S., Hadouch, M., 2006. From the subjects' point of view: reactions to different types of situational judgment items. European Journal of Psychological Assessment 22, 168–176. http://dx.doi.org/10.1027/1015-5759.22.3.168.

Lane, S., Stone, C.A., 2006. Performance assessments. In: Brennan, B. (Ed.), Educational Measurement. American Council on Education and Praeger, Westport, CT, pp. 387–431.

Libbrecht, N., Lievens, F., 2012. Validity evidence for the situational judgment test paradigm to emotional intelligence measurement. International Journal of Psychology 47, 438–447. http://dx.doi.org/10.1080/00207594.2012.682063.

Libbrecht, N., Lievens, F., Carette, B., Côté, S.C., 2014. Emotional intelligence predicts success in medical school. Emotion 14, 64–73.

Lievens, F., 2006. International situational judgment tests. In: Weekley, J., Ployhart, R. (Eds.), Situational Judgment Tests. Jossey-Bass, San Francisco, pp. 279–300.

Lievens, F., 2013. Adjusting medical admission: assessing interpersonal skills via situational judgment tests. Medical Education 47, 182–189. http://dx.doi.org/10.1111/medu.12089.

Lievens, F., Buyse, T., Sackett, P.R., 2005. The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. Journal of Applied Psychology 90, 442–452. http://dx.doi.org/10.1037/0021-9010.90.3.442.

Lievens, F., De Corte, W., Westerveld, L. Understanding the building blocks of selection procedures: effects of response fidelity on performance and validity. Journal of Management. http://dx.doi.org/10.1177/0149206312463941, in press.

Lievens, F., De Soete, B., 2012. Simulations. In: Schmitt, N. (Ed.), Handbook of Assessment and Selection. Oxford University Press, pp. 383–410.

Lievens, F., Patterson, F., 2011. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. Journal of Applied Psychology 96, 927–940. http://dx.doi.org/10.1037/a0023496.

Lievens, F., Peeters, H., Schollaert, E., 2008. Situational judgment tests: a review of recent research. Personnel Review 37, 426–441. http://dx.doi.org/10.1108/00483480810877598.

Lievens, F., Sackett, P.R., 2006. Video-based versus written situational judgment tests: a comparison in terms of predictive validity. Journal of Applied Psychology 91, 1181–1188. http://dx.doi.org/10.1037/0021-9010.91.5.1181.

Lievens, F., Sackett, P.R., 2012. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. Journal of Applied Psychology 97, 460–468. http://dx.doi.org/10.1037/a0025741.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., Grubb, W.L., 2007. Situational judgment tests, response instructions, and validity: a meta-analysis. Personnel Psychology 60, 63–91. http://dx.doi.org/10.1111/j.1744-6570.2007.00065.x.

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A., Braverman, E.P., 2001. Use of situational judgment tests to predict job performance: a clarification of the literature. Journal of Applied Psychology 86, 730–740. http://dx.doi.org/10.1037//0021-9010.86.4.730.

McDaniel, M.A., Psotka, J., Legree, P.J., Yost, A.P., Week, J.A., 2011. Toward an understanding of situational judgment item validity and group differences. Journal of Applied Psychology 96, 327–336. http://dx.doi.org/10.1037/a0021983.

Motowidlo, S.J., Beier, M.E., 2010. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. Journal of Applied Psychology 95, 321–333. http://dx.doi.org/10.1037/a0017975.

Motowidlo, S.J., Crook, A.E., Kell, H.J., Naemi, B., 2009. Measuring procedural knowledge more simply with a single-response situational judgment test. Journal of Business and Psychology 24, 281–288. http://dx.doi.org/10.1007/s10869-009-9106-4.

Motowidlo, S.J., Dunnette, M.D., Carter, G.W., 1990. An alternative selection procedure: the low-fidelity simulation. Journal of Applied Psychology 75, 640–647. http://dx.doi.org/10.1037/0021-9010.75.6.640.

Motowidlo, S.J., Hooper, A.C., Jackson, H.L., 2006. A theoretical basis for situational judgment tests. In: Weekley, J., Ployhart, R. (Eds.), Situational Judgment Tests. Jossey-Bass, San Francisco, pp. 57–81.

Northrop, L.C., 1989. The Psychometric History of Selected Ability Constructs. US.S. Office of Personnel Management, Washington, DC.

O'Connell, M.S., Hartman, N.S., McDaniel, M.A., Grubb III, W.L., Lawrence, A., 2007. Incremental validity of situational judgment tests for task and contextual job performance. International Journal of Selection and Assessment 15, 19–29. http://dx.doi.org/10.1111/j.1468-2389.2007.00364.x.

Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A., Donovan, M.A., 1998. Interactive video assessment of conflict resolution skills. Personnel Psychology 51, 1–24. http://dx.doi.org/10.1111/j.1744-6570.1998.tb00714.x.

Oostrom, J.K., Born, M.P., Serlie, A.W., van der Molen, H.T., 2010. Effects of individual differences on the perceived job relatedness of a cognitive ability test and a multimedia situational judgment test. International Journal of Selection and Assessment 18, 394–406. http://dx.doi.org/10.1111/j.1468-2389.2010.00521.x.

Oostrom, J.K., Born, M. Ph, Serlie, A.W., Van der Molen, H.T., 2012. Implicit trait policies in multimedia situational judgment tests of leadership skills: can they predict leadership behavior? Human Performance 25, 335–353. http://dx.doi.org/10.1080/08959285.2012.703732.

Oswald, F.L., Schmit, N., Kim, B.H., Ramsay, L.J., Gillespie, M.A., 2004. Developing a biodata measure and situational judgment inventory as predictors of college student performance. Journal of Applied Psychology 89, 187–207. http://dx.doi.org/10.1037/0021-9010.89.2.187.

Ployhart, R.E., Ehrhart, M.G., 2003. Be careful what you ask for: effects of response instructions on the construct validity and reliability of situational judgment tests. International Journal of Selection and Assessment 11, 1–16. http://dx.doi.org/10.1111/1468-2389.00222.

Prewett, M.S., Brannick, M.T., Peckler, B., 2013. Training teamwork in medicine: an active approach with role play and feedback. Journal of Applied Social Psychology 43, 316–328. http://dx.doi.org/10.1111/j.1559-1816.2012.01001.x.

Richman-Hirsch, W.L., Olson-Buchanan, J.B., Drasgow, F., 2000. Examining the impact of administration medium on examinee perceptions and attitudes. Journal of Applied Psychology 85, 880–887. http://dx.doi.org/10.1037//0021-9010.85.6.880.

Roth, P.L., Bobko, P., Buster, M., 2013. Situational judgment tests: the influence and importance of applicant status and targeted constructs on estimates of Black–White subgroup differences. Journal of Occupational and Organizational Psychology 86, 394–409. http://dx.doi.org/10.1111/joop.12013.

Roth, P.L., Bobko, P., McFarland, L., Buster, M., 2008. Work sample tests in personnel selection: a meta-analysis of Black-White differences in overall and exercise scores. Personnel Psychology 61, 637–661. http://dx.doi.org/10.1111/j.1744-6570.2008.00125.x.

Sackett, P.R., Lievens, F., 2008. Personnel selection. Annual Review of Psychology 59, 419–445.

Sharma, S., Gangopadhya, M., Austin, E., Mandal, M.K., 2013. Development and validation of a situational judgment test of emotional intelligence. International Journal of Selection and Assessment 21, 57–73. http://dx.doi.org/10.1111/ijsa.12017.

Weekley, J.A., Jones, C., 1997. Video-based situational testing. Personnel Psychology 50, 25–49. http://dx.doi.org/10.1111/j.1744-6570.1997.tb00899.x.

Weekley, J.A., Jones, C., 1999. Further studies of situational tests. Personnel Psychology 52, 679–700. http://dx.doi.org/10.1111/j.1744-6570.1999.tb00176.x.

Weekley, J.A., Ployhart, R.E., Harold, C.M., 2004. Personality and situational judgment tests across applicant and incumbent settings: an examination of validity, measurement, and subgroup differences. Human Performance 17, 433–461. http://dx.doi.org/10.1207/s15327043hup1704_5.

Weekley, J., Ployhart, R.E., 2006. Situational Judgment Tests. Jossey-Bass, San Francisco.

Whetzel, D.L., McDaniel, M.A., 2009. Situational judgment tests: an overview of current research. Human Resource Management Review 19, 188–202.

Whetzel, D.L., McDaniel, M.A., Nguyen, N.T., 2008. Subgroup differences in situational judgment test performance: a meta-analysis. Human Performance 21, 291–309. http://dx.doi.org/10.1080/08959280802137820.

Skills Training: Public Policy Aspects

Moira Nelson, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes factors influencing investments in skills training as well as related outcomes in terms of participation rates and benefits on an individual and societal level. Market mechanisms largely drive human capital investments. According to this logic, workers and employers conduct a cost–benefit analysis, weighing expected payoffs against expected costs and investing accordingly. At the same time, market failures and collective action problems may occur in the provision of training. Various policies and institutions have been shown to resolve such problems and boost training. Whereas participation rates vary widely by firm characteristics, high rates of participation help countries specialize in high-quality production and achieve high job satisfaction.

Introduction

Various factors shape skills investments and their related benefits on an individual and societal level. Following largely from market mechanisms of supply and demand, skill investments are made which increase worker productivity and thereby wages for employees and profits for employers. This logic goes some way in explaining individual and overtime trends. At the same time, there are many reasons to suspect training to be undersupplied, due to informational deficiencies, capital market imperfections, and the risk that workers do not stay with the training firm long enough to repay the firm in the form of productivity increases. With respect to the issue of worker mobility, Becker theorized that employers should only provide training in specific skills. Available data, however, shows employers do invest in general training and various theories offer explanations for why this is the case, oftentimes with reference to the concepts of market failure and collective action problems. Moreover, overeducation also occurs, resulting in the existence of a significant mismatch between the required and actual skills of the workforce.

Participation in skills training varies widely across types of individuals and countries. Younger workers enter into training more often because the horizon of expected payoffs is longer and foregone earnings lower. More productive workers are trained more often because the wage premia is higher. Training rates vary depending on the structure of initial vocational training, continuing training policies, and economic context. If countries are able to realize high rates of participation in training, various benefits are identified including a comparative advantage in high-quality production and high job satisfaction.

Markets for Education and Training

Economic theory reaching back at least 200 years to Adam Smith tells us that competitive markets provide incentives for employers and workers to invest in training. Training occurs because the future (discounted) payoffs outweigh the costs (Becker, 1980). Analyses often consider the direct costs of training as well as the opportunity costs. The following model expresses well the cost–benefit analysis of training,

$$B_{\mathrm{i}}\left(1-(1/r)^{N}\right)\Big/r > K + Y_0 \qquad [1]$$

whereby B_{i} indicates the earnings gained due to schooling, which are presumed to be constant throughout one's working life, N is to the years left in one's career, r is the discount rate on future gains, K is the direct costs of schooling, and Y_0 is foregone earnings during the schooling period. As long as discounted earnings gained from schooling is higher than the sum of the costs of training and foregone earnings, training takes place. The model allows us to derive a few stylized facts about when training will occur which are often upheld in empirical studies. Younger people are expected to engage in more training because they have a longer horizon of higher returns and lower foregone earnings due to seniority pay scales (as N increases, benefits to train increase). Participation among young people indeed tends to be higher, although more disaggregated analysis suggests that effects vary by individual characteristics and type of training.

Skill training activity is also expected to increase when the costs of training decrease (as K decreases, benefits to training increase). Public subsidization of training is generally supported, although theoretically justifying such action requires proof that market mechanisms fail to meet demand which is empirically quite challenging. A randomized experiment using adult training vouchers showed that the provision of such vouchers increased training activity (Messer and Wolter, 2009). At the same time, deadweight costs were also quite substantial at around 50% which underscores the risk that subsidizing training does not actually increase participation on the aggregate.

Finally, eqn [1] predicts that workers who expect a high wage premium from training are more likely to participate (when B_{i} is high relative to Y_0) (Heckman et al., 1999). The evidence strongly supports the expectation that more productive and well-educated workers are more like to be training. There are many different reasons given for this finding. In addition to expecting higher returns, such workers tend to receive more intrinsically motivated. Employers thus hold a distinct bias toward training these workers since the payoff is relatively higher.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.75010-0

Wage premiums are also influenced by economic changes that alter demand for different types of skills. Technological changes and globalization, for instance, generally lead to increases in the level of skills demanded in the labor market. As demand for more highly trained workers increases without a parallel increase in supply, the wage premium associated with higher training increases. Conversely, the entrance of large cohorts of well-educated workers into the labor market has the effect of pushing down the wage premium. This effect is visible in the United States in the 1970s and can also be seen in Korea and Japan (Kim and Topel, 1995).

Causes of Underinvestment in Skills Training

Various theoretical perspectives on skills training challenge the expectation of traditional economic theory that markets always lead to the optimal level of training activity. A central insight here is that encouraging skills training relies on policies and regulations that make workers and employer aware and secure of the future benefits of investing in training.

Concepts of market failures and collective action problems are useful in explaining deficiencies in the market provision of training. Market failures arise due to lack of information about the potential gains from or lack of financing for training and collective action problems occur because the gains to training cannot be fully internalized by the investor. Particularly as skill needs change quickly in response to technological advancements, difficulties may arise in estimating the appropriate scale and content of new training programs. Furthermore, the payoffs from training may be underestimated by either workers or employers. There may be capital market problems whereby workers are unable to get financing for training which would deliver returns in the future (Dougherty and Tan, 1999). The International Adult Literacy survey reveals that time constraints are by far the main reason for nonparticipation.

Beyond market failures, collective action problems arise in the provision of skills training essentially because workers are mobile and can leave the firm which provided them with training. Employers therefore face a risk in training workers, because training stands to improve workers' productivity in competing firms. Often used to analyze incomplete contracting with relation to physical goods, the concept of asset specificity has been used to understand collective action problems in relation to skill investments. Specific assets are those that cannot be readily applied to a different purpose or whose return depends heavily on the cooperation of others. The degree of skill specificity has been widely used to explain employer incentives to train.

In a seminal contribution, Gary Becker (1993) sees the specificity of skills as largely inherent to the skill itself and expresses the willingness of employers to invest in workers' skills as a function of their specificity. The inclination of employers to fund training increases proportionally to the specificity of the skills being delivered. In Becker's account, market failures do not occur and the appropriate level of training always takes place, so long as workers are able to finance training (Acemoglu and Pischke, 1999).

Many have questioned the main claim of Becker's theory that firms only pay for firm-specific skills training, largely due to the observation that large-scale investments in general skills training indeed take place. Some go as far to claim that even investments in technically specific skills are in some way general if they signal that the worker is competent and motivated, in which case one may contend that all types of training are *de facto* general to some degree. Other modifications to Becker's argument have been made. The costs associated with employee turnover, labeled 'transaction costs,' lead both the employer and the employee to seek alternatives to termination via compromised solutions within the firm. 'Monopsonistic' labor markets allow employers to restrict workers' mobility regardless of their skill sets. General training can also be seen as a side payment, provided to encourage the worker to partake in training in specific skills (Kessler and Lülfesmann, 2006). A more compressed wage structure also leads to higher training by increasing the returns to training for the firm. Perhaps through their influence on the wage structure, strong unions are also related to a higher incidence of training. One can generalize that firms provide general training when carrots or sticks are used to limit worker mobility. Economists discuss these factors which limit mobility as producing labor market imperfections (Acemoglu and Pischke, 1999, 1998).

The Varieties of Capitalism approach sees various social policies and regulatory institutions as supporting investments in specific skills (Hall and Soskice, 2001). With respect to industry-specific skills, unemployment insurance provides workers with incentives to invest while standard-setting procedures ensure the uniformity of training and monitoring mechanisms protect against poaching. This approach also recognizes that *workers* need incentives to invest in firm-specific skills, a point not raised in Becker's more employer-centric account and sees employment protection as providing such an incentive.

To summarize, market failures and collective action problems arguably lead to suboptimal levels of training activity. Becker presumed that this would not occur so long as workers invested adequately in general skills training. Yet workers face considerable informational, time, and financial constraints in undertaking such training. Moreover, most training is not solely firm specific, not least because high levels of firm-specific skills signal the general capacity to learn and be productive in the workplace. Various contributions show that market imperfections, created informally in the workplace or institutionalized through such policies as unemployment insurance, provide incentives for workers and employers to invest in training despite associated risks.

Data on Participation Rates and Trends within and between Countries

The term 'coordinated market economies' refers to countries predominantly in continental and northern Europe which exhibit strong welfare states along with high investment in specific types of skills. In contrast, 'liberal market economies,' including the Anglo-Saxon countries, exhibit lean welfare states and excel in the provision of general training. These differences in training activity have important implications for working conditions and comparative advantage in product markets.

The coordinated market economies exhibit higher levels of vocational training, and investment in this type of training is

sustained by institutions that bring employers, workers, and policymakers together and provide them with incentives to compromise over the future orientation of training policy. Finegold and Soskice (1988) importantly recognized the high-skill equilibrium in Germany and the low one in the United Kingdom and traced these differences in skill investments back to the complex net of socioeconomic institutions in the former and the lack thereof in the latter. Crouch et al. (2001) identify three types of socioeconomic institutions related to training: state led, corporatist networks, and markets. In state-led systems, most vocational training takes place in schools (e.g., Sweden, France). In corporatist settings training is done in firms (e.g., Austria, Germany), and in market-based systems there is little vocational training at all. Technological change and tertiarization challenge vocational training systems in diverse ways (Busemeyer and Trampusch, 2011).

Estimating participation rates in training for all categories of workers is complicated because of the variety of training types provided and the lack of systematic data collection. Training may be long term and involve a formal training contract. However, most training is informal which frustrates getting a sense of trends (Lynch, 1994). Skills are often gained through a process of trial and error, by imitating what other workers do and receiving informal supervision from more senior employees.

The US Bureau of the Census' Current Population Surveys (CPS) provides data on workplace training in the United States and reveals that a large proportion of training is indeed carried out informally. In Europe, the Continuing Vocational Training Survey (CVTS) conducted by the European Union provides detailed data about participation in training. The type of training is broken down into five categories: in-work vocational training, self-learning, quality circles, job rotation, and training at conferences/lectures/workshops/seminars. The correlation between these different types ranges between 0.71 and 0.91, with an average of 0.82, which suggest that trends in informal forms of training are a good proxy for more formal training activities.

The CVTS also reveals that training occurs most often in two industries, 'electricity, gas, and water supply' as well as 'financial intermediation,' and least often in 'hotels and restaurants.' Training activity also varies significantly by firm size with larger firms almost always providing training to at least some employees and medium-sized (50–249) and small (10–49) firms offering proportionally less training, respectively. In some countries the gap between different sized firms is small (e.g., Norway, UK) while in others it is quite large (e.g., Greece, Bulgaria).

Participation tends to be slightly procyclical, although there is evidence that training might actually increase during economic downturns. Reconciling these findings, Brunello (2009) explains that employers train existing workers more during downturns and slow the recruitment of new workers.

Public and Semi-Public Policies on Training

Public policies promote training by delivering financial support as well as regulating both the content and completion of training through curricula development and certification, respectively. In most cases, training is financed privately either by the employer or by the worker. However, there are also ways in which the costs of training are socialized. The state often provides funds for training through general taxation, although the amount provided varies tremendously across countries. The Nordic countries are widely known as high spenders. Besides providing funds, the state may also make training financially viable by offering tax rebates to training firms.

Tax levies on firms are another way in which to raise funds for training. Levies may be legally enforced by the state, as in France, or coordinated by employer organizations, as in the Netherlands. Levies typically call on firms to pay a particular percentage of wages into a common pool. Other times risk pooling occurs in more informal ways. For instance, in Germany larger firms historically 'overtrain' to the benefit of smaller firms and therefore subsidize the training costs of smaller firms to some extent.

The state and employers are also heavily involved in regulating curricula and certification. Curricula are often reformed in response to shifts in supply and demand, as periods of tight labor markets provide incentives to broaden occupational profiles. Technological change and more fluid labor markets can also be seen to stimulate consolidation of occupational profiles, since identifying the overlap between different skill sets becomes more important. Certification is key in communicating the skills that training has provided. Flexible certification procedures have become more renowned in the past few decades, such as accreditation of prior learning, modularization, and national qualification frameworks.

Estimates of the Amount Invested in Training

A common way to estimate training costs is the ingredients method according to which all the relevant components, or ingredients, of training are identified, the costs of each estimated, and all of them summed (Levin, 1983; Mincer, 1994). Ingredients usually include workers' and trainers' time, instruction materials and facilities. In reality, many difficulties exist in estimating the costs of various ingredients (Tsang, 1999). Calculating costs in this ways relies on detailed information about how workers spend their time, which is typically collected in the form of a diary. Clearly a costly method, it is not used often, particularly for large samples.

Another way to estimate training costs is to deduce the amount of training conducted from the rate of earnings growth over the career (Mincer, 1962). Evidence indeed shows that training firms experience larger wage growth than nontraining firms. At the same time, earnings growth may result from other factors besides training, which may lead this approach to overestimate investments in training in the case of general skills. For instance, workers often change jobs to receive better pay and earnings growth in this case would be the effect of bargaining power rather than training. In the case of specific skills, however, this approach would underestimate the amount of training provided since earnings would not rise in line with productivity increases. Studies that show that the effect of training is higher for productivity than on wages support this view (Ballot et al., 2006; Dearden et al., 2006).

The net costs of training are widely shown to be negative during the apprenticeship period. German apprenticeship training was estimated at 12 300 DM in 1985 (Franz and Soskice, 1995) and, with different data based on a 1991 survey, found to exceed for certain 7500 DM (Bardeleben et al., 1995). A recent study of Swiss firms reveals that net training costs did not have an impact on the number of trainees, despite the relatively high risk of poaching in the Swiss context (Muehlemann et al., 2007). This finding cautioned against the subsidization of apprenticeship training.

The CVTS also provides rich data on spending on continuing training. In 2005, spending as a percentage of labor costs was highest in Denmark (2.7%) and France (2.3%) and lowest in Latvia (0.8%) and Greece (0.6%). Additional data break down costs by hour, individual, industry, size class, and type of cost.

Given that credit constraints are recognized as a major reason why training does not occur, many countries have negotiated cofinancing arrangements to boost participation. The costs of training are often made tax deductible. Learning accounts and training vouchers are also widely used, whereby the state or the employer usually matches contributions made by participants. Additional incentives are usually made for low-skilled workers.

The Benefits of Training for the Individual and Society

The relationship between participation in training and higher wages is strongly supported by the literature. Research suggests that the gross returns to wages from a year's worth of vocational training is approximately 10% per year or more (Mincer, 1994).

Often unclear, however, is whether training drives wage increases or if those who receive training were simply more productive in the first place. Disentangling these effects is tricky because of selection bias. Workers with a higher relative level of education are far more likely to enter a training program than others (Brunello and Comi, 2004). As a result, training often has the effect of increasing labor market segmentation.

In terms of apprenticeship training, there is a tension between firms' use of apprentices as an immediate source of productivity, by substituting normal workers with apprentices, and as an investment whereby employers devote energy to improving the skill base of apprentices with an eye on creating the future workforce. In a recent study of Germany, firms are shown to follow both strategies (Mohrenweiser and Backes-Gellner, 2010). Firms that use a substitution method tend to use apprentices as cheap labor, as evidenced with data from Norway (Askilden and Nilsen, 2005) and the Netherlands (Smits, 2006). Smits (2006) finds that firms which follow an investment strategy provide higher quality training.

Rather than comparing quality of apprenticeship training, other studies distinguish those countries which invest highly in vocational skills, the coordinated market economies discussed in Section Causes of Underinvestment in Skills Training, to countries which tend to provide general skills training, the liberal market economies. Within this group, Germany is often singled out and the high level of investment in apprentices. Such training also leads to a comparative advantage in high-quality diversified production (Streeck, 1992) which depends on highly training workers (Soskice, 1999). Others point out that the vocational skills are oftentimes provided by schools rather than firms, an empirical tendency often over looked in the literature (Busemeyer, 2009). Given the challenges in encouraging firms to provide general skills, the school component of vocational programs is a propitious way to expand this type of training, particularly as economic change calls for a stronger base in precisely these types of skills. In Denmark, while most training occurs in firms, the state has often used vocational schools to expand the general skills component of training and thereby facilitate adjustment to economic change (Nelson, 2011).

Beyond the need for stronger general skills, various economic changes necessitate lifelong learning. Technological change and other competitive pressures create the need to invest frequently in updating one's skill profile to remain competitive on the labor market. Expanding opportunities to acquire skills and clear certification mechanisms improve the ability of countries to adjust to this changed context (Nelson, 2010).

A particular way to promote learning involves restructuring workplace relations to facilitate discretionary learning. Discretionary learning refers to a form of work organization whereby workers receive considerable autonomy yet also responsibility to independently resolve issues that arise, often by teaching themselves new tasks. The Nordic countries exhibit the highest levels of discretionary learning which partially explains their smooth adaptation to the knowledge economy and low numbers of Taylorist jobs in these countries (Lundvall and Lorenz, 2011a,b).

Skill Availability

The onset of the knowledge economy significantly alters the skills demanded in the labor market. Many of the new jobs being created in high-wage services involve advanced cognitive skills. Interpersonal and communication skills are also in high demand. Technological change and intense global competitive pressures also make updating one's skill set as job profiles and modes of production continue to change. Both strong initial education and access to continuing training are necessary for remaining competitive in the labor market. If education reforms do not keep pace with rising skill needs, the risk of a risk shortage increases.

Yet the phenomenon of overeducation also exists, whereby the increase in educational attainment outstrips the rising skill needs of employers. Overeducation derives largely from competition over jobs whereby potential applicants attempt to gain a competitive advantage by gaining an additional qualification.

Skills mismatch refers to the gap between the skills required by a worker to complete their job and the skills that they have. It is often measured as the combined proportion of the workforce that is undereducated and that which is overeducated. The ILO (2013) estimates that there is a mismatch of between 30 and 50% of the workforce in European countries. Although the sample differs, a report by the OECD suggests the proportion to be closer to 50 (Quintini, 2011).

Concluding Remarks

The issue of skills training has received much attention in recent years. Classical economic theory contends that the optimal level of training is provided through market mechanisms. That is, training always occurs if and only if there is reason to expect a net gain in terms of productivity. For supporters of this perspective, subsidizing training leads to high deadweight losses and is an inappropriate use of government funds.

There are nevertheless reasons to expect market failures and collective action problems to lead to a suboptimal level of training activity. Workers and employers likely face enormous hurdles in recognizing valuable training opportunities and allocating the time and resources necessary to ensure successful completion of these programs. Moreover, most training involves the provision of more general skills which generate insecurities among employers about workers being poached by competing firms. To relieve this fear, the literature illuminates the positive role that various informal and formal policies and institutions can play. Workers and employers may find mutually beneficial solutions among themselves or employers may manipulate their monopsony power to retain workers. Various social policies, labor market regulations, and other institutions are able to reduce the risk facing workers and employers investing in either general or specific skills.

Whereas caution must be exercised in deciding how and when to subsidize and regulate training activity, there appear to be considerable payoffs to both individuals and society more generally from realizing high rates of participation in training. Training is strongly related to higher wages and, if numbers are sufficiently high, countries are able to realize a comparative advantage in high-quality production. At the same time, the high rates of skill mismatch in the workforce today suggest that there is room to improve the alignment between training programs and market needs. Targeted training policies are particularly salient today since the economic crisis has triggered a new rise in skills mismatch (ILO, 2013).

See also: Adult Education and Lifelong Learning: The US Experience and Beyond; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Education and Economic Growth; Education and Employment; Education, Economics of; Vocational Education.

Bibliography

Acemoglu, Daron, Pischke, Jorn-Steffen, 1999. Beyond Becker: training in imperfect labour markets. The Economic Journal 109 (453), 112–142.

Acemoglu, Daron, Pischke, Jörn-Steffen, 1998. Why do firms train? theory and evidence. Quarterly Journal of Economics 113 (1), 79–118.

Askilden, Jan Erik, Nilsen, Øivind Anti, 2005. Apprentices and young workers: a study of the Norwegian youth labour market. Scottish Journal of Political Economy 52 (1), 1–17.

Ballot, Gérard, Fakhfakh, Fathi, Taymaz, Erol, 2006. Who benefits from training and R&D, the firm or the workers? British Journal of Industrial Relations 44 (3), 473–495.

Bardeleben, Richard, Beicht, Ursula, Fehér, Kálmán, 1995. Betriebliche Kosten Und Nutzen Der Ausbildung: Repräsentative Ergebnisse Aus Industrie, Handel Und Handwerk. Bertelsmann.

Becker, Gary Stanley, 1980. Investment in human capital: effects on earnings. In: Human Capital, second ed. Chicago.

Becker, Gary Stanley, 1993. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press, Chicago, London.

Brunello, Giorgio, 2009. The effect of economic downturns on apprenticeships and initial workplace training: a review of the evidence. Empirical Research in Vocational Education and Training 1 (2), 145–171.

Brunello, Giorgio, Comi, Simona, February 2004. Education and earnings growth: evidence from 11 European countries. Economics of Education Review 23 (1), 75–83.

Busemeyer, Marius R., Trampusch, Christine, 2011. The comparative political economy of collective skill formation. In: The Political Economy of Collective Skill Formation. Oxford University Press, pp. 3–38.

Busemeyer, Marius R., 1 July 2009. Asset specificity, institutional complementarities and the variety of skill regimes in coordinated market economies. Socio-economic Review 7 (3), 375–406.

Crouch, Colin, Finegold, David, Sako, Mari, 2001. Are Skills the Answer? The Political Economy of Skill Creation in Advanced Industrial Countries. Oxford University Press, New York.

Dearden, Lorraine, Reed, Howard, Reenen, John Van, 2006. The impact of training on productivity and wages: evidence from British panel data. Oxford Bulletin of Economics and Statistics 68 (4), 397–421.

Dougherty, Christopher, Tan, Jee-Peng, 1999. Financing training: issues and options. Journal of European Industrial Training 23, 416–436.

Finegold, David, Soskice, David, 1988. The failure of training in Britain: analysis and prescription. Oxford Review of Economic Policy 4 (3), 21–53.

Franz, Wolfgang, Soskice, David W., 1995. The German apprenticeship system. In: Institutional Frameworks and Labor Market Performance. Comparative Views on the German and U.S. Economies. Routledge, London, pp. 208–224.

Hall, Peter, Soskice, David, 2001. Varieties of Capitalism: The Institutional Foundations of Comparative Advantage. Oxford University Press.

Heckman, James J., Lalonde, Robert J., Smith, Jeffrey A., 1999. Chapter 31 The Economics and Econometrics of Active Labor Market Programs, vol. 3, Part 1. Elsevier, pp. 1865–2097.

ILO, 2013. Global Employment Trends for Youth in 2013: A Generation at Risk. International Labour Office, Geneva.

Kessler, Anke S., Lülfesmann, Christoph, 2006. The theory of human capital revisited: on the interaction of general and specific investments. The Economic Journal 116 (514), 903–923.

Kim, D., Topel, Robert H., 1995. Labor markets and economic growth: lessons from Korea's industrialization, 1970–1990. In: Differences and Changes in Wage Structures. Chicago University Press, Chicago.

Levin, Henry M., 1983. Cost-Effectiveness: A Primer. Sage Publications, Beverly Hills, CA and London.

Lundvall, Bengt-Ake, Lorenz, Edward (Eds.), 2011a. Social investment in the globalising learning economy: a European perspective. What Future for Social Investment? Policy Press, London.

Lundvall, Bengt-Ake, Lorenz, Edward (Eds.), 2011b. Social investment: a paradigm in search of a new economic model and political mobilisation. What Future for Social Investment? Policy Press, London.

Lynch, Lisa M., 1994. Training and the Private Sector: International Comparisons. University of Chicago Press.

Messer, Dolores, Wolter, Stefan C., February 2009. Money Matters – Evidence from a Large-scale Randomized Field Experiment with Vouchers for Adult Training. SSRN eLibrary.

Mincer, Jacob, 1962. On-the-job training: costs, returns, and some implications. Journal of Political Economy 70, 50–79.

Mincer, Jacob, 1994. Investment in U.S. Education and Training. Working Paper 4844. National Bureau of Economic Research.

Mohrenweiser, Jens, Backes-Gellner, Uschi, 17 August 2010. Apprenticeship training: for investment or substitution? International Journal of Manpower 31 (5), 545–562.

Muehlemann, Samuel, Schweri, Juerg, Winkelmann, Rainer, Wolter, Stefan C., 2007. An empirical analysis of the decision to train apprentices. LABOUR 21 (3), 419–441.

Nelson, Moira, 2010. The adjustment of national education systems to a knowledge based economy: a new approach. Comparative Education 46 (4), 463–486.

Nelson, Moira, 2011. Continued collectivism: the role of trade self-management and the social democratic party in Danish vocational education and training. In: The Political Economy of Collective Skill Formation. Oxford University Press, pp. 179–202.

Quintini, Glenda, 2011. Right for the Job Over-qualified or Under-skilled? OECD, Paris.

Smits, W., 2006. The quality of apprenticeship training. Education Economics 14 (3), 329–344.

Soskice, David W., 1999. Divergent production regimes: coordinated and uncoordinated market economies in the 1980s and 1990s. In: Continuity and Change in Contemporary Capitalism. Cambridge University Press.

Streeck, Wolfgang, 1992. Social Institutions and Economic Performance: Studies of Industrial Relations in Advanced Capitalist Economies. SAGE Publications.

Tsang, Mun C., 3 January 1999. The cost of vocational training. Education + Training 41 (2), 79–97.

Skinner, Burrhus Frederick (1904–90)

Marc N Richelle, University of Liège and Académie Royale de Belgique, Liège, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

American psychologist, a leading figure of behaviorism, Skinner contributed original laboratory techniques for the study of operant behavior; a formulation of the nature of a science of behavior (radical behaviorism), a selectionist view of the causation of animal and human conduct inspired by the biological evolution model; and proposals for applications derived from his experimental and theoretical work to education, psychological treatment, and social organization. He broadened the scope of the behavioristic approach by including verbal behavior and consciousness in his analysis.

After J.B. Watson, the founder of the behaviorist movement, Skinner has been the most influential, and also the most controversial, figure of behaviorism. His contributions to behavioral sciences are manifold: he designed original laboratory techniques for the study of animal and human behavior, which he put to work to produce new empirical data in the field of learning, and to develop a theory of operant behavior that led him eventually to a general psychological theory; he further elaborated the behaviorist approach, both refining and extending it, in a version of behavioral science which he labeled 'radical behaviorism'; he formulated seminal proposals for applied fields such as education and psychological therapy and pioneered in machine-assisted learning; finally, building upon his conception of the causation of behavior, he ventured into social philosophy, questioning the traditional view of human nature and of the relation of humans to their physical and social environment. This part of his work has been the main source of sometimes violent controversy.

Biographical Landmarks

Skinner was born on 20 March 1904 in Susquehanna (Pennsylvania, USA) in a middle-class family and experienced the usual childhood and adolescence of provincial American life. He attended Hamilton College, which was not a particularly stimulant institution to him. He was first attracted by a literary career, which he soon gave up, after traveling to Europe. He turned to psychology and was admitted at Harvard in 1928. He obtained his PhD in 1931, with a theoretical thesis on the concept of reflex – a first landmark in his reflections on the causation of behavior, an issue he was to pursue throughout his scientific career. He stayed at Harvard five more years, beneficiary of an enviable fellowship, affiliated to the physiology laboratory headed by Crozier. In 1936, he was appointed professor at the University of Minnesota where he developed his conditioning chamber for the study of operant behavior in animals – which was to be known as the Skinner box – and wrote his first book *The Behavior of Organisms* (1938). In 1945, he moved to Indiana University, as chairman of the Department of Psychology. In 1948, he was offered the prestigious Edgar Pierce Professorship in Psychology at Harvard University, where he was to stay until his death on 18 August 1990.

Skinner received during his lifetime the highest national awards an American psychologist could receive, and he was praised as one of the most prominent psychologists of the century, in spite of harsh attacks against some of his ideas, from the most opposite sides of scientific and lay-people circles.

Although most of Skinner's laboratory research was carried out with animal subjects, mainly rats and pigeons, using the so-called operant procedure that will be described hereafter, essentially he was interested, not in animal behavior proper, but in behavior at large, and more specifically in human behavior. After the tradition of other experimental psychologists before him, such as Pavlov and Thorndike, and of most of his colleagues in the behaviorist school of thought, such as Hull, Tolman, or Guthrie, he resorted to animals as more accessible subjects than humans for basic studies on behavior, just as physiologists used to do quite successfully. The relevance of extrapolating from animals to humans is of course an important issue in psychology, still debated today. However, Skinner's main concern was obviously with humans, as evidenced by his literary writings in the field of social philosophy – namely the utopian novel *Walden Two* (1948) and the essay *Beyond Freedom and Dignity* (1971) – as well as by his theoretical endeavors to account for human behavior (*Science and Human Behavior* 1953) and for verbal behavior (*Verbal Behavior* 1957). A detailed autobiography in three volumes has been published by Skinner during his lifetime (1976, 1979, 1983).

Operant Conditioning and the Skinner Box

The operant conditioning chamber, often called the Skinner box, is a laboratory device derived from Thorndike's puzzle box and from the mazes familiar to students of learning in rats by the time Skinner started his career. In its most common form, it consists of a closed space in which the animal moves freely; it is equipped with some object that the subject can manipulate easily, be it a lever for rats, or a small illuminated disk upon which pigeons can peck, and with a food dispenser for delivering calibrated quantities of food. By exploring spontaneously this particular environment, the subject will eventually discover the basic relation between a defined response, pressing the lever or pecking the key, and the presentation of a reinforcing stimulus, a small food reward. The process involved is basically trial-and-error learning, as captured since Thorndike in the Law of Effect.

Learning this association can be facilitated by shaping progressively the desired behavior; in that case, the

experimenter observes systematically the subject's behavior, and reinforces selectively those actions that come closer and closer to the response required. The importance of that procedure will be commented further in relation with the concept of variation and with the issue of teaching. The basic relation here is between an operant response (i.e., a response instrumental to produce some subsequent event) and its consequence (i.e., the reinforcement), rather than between a stimulus and a response elicited by it, as in Pavlovian or respondent conditioning. This simple situation may be made more complex either by introducing so-called discriminative stimuli, the function of which is not to trigger the response at the manner of a reflex, but to set additional conditions under which the response will be reinforced, or by changing the basic one response-one reinforcement link to some more complicated contingencies, for instance requiring a given number of responses to one reinforcement, or the passing of some defined delay. A wide variety of schedules of reinforcement have been so studied, be it for their own sake as sources of information on the lawfulness of behavior (for instance, modern research has applied optimization models borrowed from economics to the study of operant behavior), or as efficient tools for other purposes (such as the analysis of sensory functions in animal psychophysics, of the effects of drugs acting upon the central nervous system in experimental psychopharmacology, or of cognitive capacities).

The operant technique presented two important features by the time Skinner developed it from the 1930s to the 1950s. It emphasized the study of individual subjects through a long period of time rather than groups of subjects for a few sessions, as used to be the case in maze studies and the like. This interest in individual behavior would favor later applications to human subjects in educational and clinical settings. Second, the operations involved soon were automatized by resorting to electromechanical circuits, to be replaced later by online computer control. This led to a level of efficiency and precision unprecedented in psychological research.

The Evolutionary Analogy

Skinner captured the essence of operant behavior in the formula 'control of behavior by its consequences,' and very early he pointed to the analogy between the selection of the response by the subsequent event and the mechanism at work in biological evolution. An increasingly large part of his theoretical contributions were eventually devoted to elaborating the evolutionary analogy (Skinner, 1987). The generalization of the selectionist model to behavior acquisition at the individual level, initially little more than a metaphoric figure, has later gained credentials with the theses of neurobiologists, such as Changeux's *Generalised Darwinism* (1983) or Edelman's *Neural Darwinism* (1987), who both have substantiated in ontogeny selective processes previously reserved to phylogeny. One of the main tenets of Skinner's theory converges with contemporary views in neurosciences.

Skinner extended the selectionist explanation to cultural practices and achievements, converging with some schools of thought in cultural anthropology, such as Marvin Harris, and in the history of science, such as Karl Popper's selectionist account of scientific hypotheses.

In Darwinian evolution, a crucial condition for selection to take place is the presence of variations. At the behavioral level, in the simplified situation of the experimental chamber, these can be observed in the exploring activities of the animal subject, eventually resulting in the discovery of the appropriate response. Once acquired, it might appear as the product of rigid training, eliminating variability, and consequently the possibility of new behavior. In fact, variability itself can be taken as the property to be selected, as shown by experiments in humans and animals. As Skinner argued, creative behavior would fit in the general selectionist scheme.

Biological evolution is responsible for the genetic endowment of organisms. Selection accounts for their specific characters, including important behavioral aspects. Ethology, sometimes called biology of behavior, has been mainly devoted to the study of these aspects, insisting on the necessity of observing animals' behavior in the natural environment, rather than put them in the artificial situation of the laboratory. Ethologists have also been very critical toward the focus on environmental factors dominant in the behaviorist approach. Behaviorists, working on learning processes, were blamed for ignoring the genetic contribution in specific behavior, including in the human species, whose members at birth were viewed as a *tabula rasa*. Karl Lorenz, the major founder of ethology, was the leading figure in this opposition to behaviorism.

American behaviorists did not deny the role of heredity, but it is true that they were slow in coming across with ethology. Skinner seems to have discovered it in the 1960s, and he discussed some of Lorenz's ideas in a paper entitled 'The phylogeny and ontogeny of behavior' (Skinner, 1966). He developed further his view in several writings. Briefly, he did not deny the evolutionary origin of many aspects of behavior, nor did he deny the species-specific differences derived from evolutionary history. Such denial would have been paradoxical from a scientist who proposed to extend the model of biological evolution to individual learning and to cultural history. He viewed his analysis of behavior and the ethological approach as complementary, not competing. It is fair to point that Lorenz later made steps in the same direction (Lorenz, 1981).

Radical Behaviorism

As a behaviorist, Skinner viewed psychology as a branch of natural sciences, more explicitly as a branch of biology, which can deal with its subject matter using the same principles as in other fields of life sciences, be it with specific implementations as required by its particular level of analysis.

Skinner defended a brand of behaviorism quite distinct from the dominant view that prevailed in the second quarter of the century. His radical behaviorism was opposed to methodological behaviorism. For most psychologists, defining their science as the science of behavior, after Watson's recommendation, did not really mean that they had abandoned mental life as the main objective of their inquiry; rather, they had simply resigned themselves to study behavior, because they had to admit that they had no direct access to mental life. Such methodological behaviorism, in fact, remained basically dualistic. In contrast, radical behaviorism is definitely monist,

and it rejects any distinction between what is called mental and behavioral. Skinner is, in this respect, closer to Watson's view than to the position of other influent neobehaviorists of his generation, although he developed a far more sophisticated view of behavior than Watson's. For instance, he rejected the simplistic claim that thought is nothing more than subvocal language, and admitted that not all behavior is directly observable. Part of human behavior is obviously private, or covert, and raises difficult methodological problems of accessibility; but this is no reason to give it different status in a scientific analysis.

Skinner vigorously denounced mentalism not so much because it refers to events that would occur in another space and be of a different substance than behavior, but because it offers pseudo-explanations which give the illusion of understanding what in fact remains to be accounted for. Mentalistic explanations, very common in daily life psychology, were still quite frequent in scientific psychology. For example, in the field of motivation, all sorts of behavior were assigned internal needs as causal agents: not only do we eat *because* we are hungry – a simple statement which, at any rate from a scientist's point of view, requires qualification – but we exhibit aggressive behavior because of some aggression drive, we interact with social partners because of a need for affiliation, we work to successful and creative outcomes because of a need for achievement, and so on. For Skinner, such explanations dispense us from looking for the variables really responsible for the behavior; these variables are more often than not to be found in the environment, and can be traced to the history of the individuals interacting with their environment, social and physical.

Combining this epistemological conception and the results of his empirical research, Skinner developed a theory of human behavior emphasizing the determining role of environmental contingencies on human actions. However, along the lines of the evolutionary analogy, he suggested a model that would account equally well for novelty and creative behavior, as exhibited in art and science productions, and for stabilized, persistent habits adapted to unchanging conditions.

Language and Consciousness

Skinner devoted special attention to verbal behavior, the importance of which in the human species justified special treatment. He presented his book *Verbal Behavior* as 'an essay in interpretation,' proposing a functional analysis of the verbal exchanges composing the global episode of speaker-listener communication. In spite of the violent criticisms expressed by the linguist Chomsky (1959), Skinner's analysis foreshadowed some aspects of the pragmatic approach adopted some years later by many psycholinguists, aware of the insufficiency of formal grammars to account for central features characterizing the use of language. He viewed verbal behavior as shaped by the linguistic community, and exerting a genuine type of control over an individual's behavior, distinct from the action of the physical environment. A large part of human behavior is, in his terms, rule governed, that is, controlled by words, rather than shaped by contingencies, that is, through direct exposure to the physical world. Many behaviors can have one or the other origin: avoidance of fire flame can derive from direct experience with fire, or from warnings received during education. Once endowed with verbal behavior, individuals can use it to describe and anticipate their own behavior, or to develop it in its own right, as in literary composition. Were it not for the unfortunate use of the term 'rule,' which makes for confusion, on one hand, with the word as used in formal linguistics and, on the other hand, with the notion of coercive control, what Skinner was pointing to was akin to a distinction now familiar to contemporary psychology and neurosciences between top-down versus bottom-up causation.

Given the widespread renewal of interest for the scientific study of consciousness in the last 25 years, it is appropriate to ask the question: What was Skinner's view on consciousness? More often than not, behaviorists have been blamed for rejecting the study of consciousness out of the field of scientific enquiry, and for having in fact imposed a taboo on the whole scientific community. Such opinion refers to the supposed negation of consciousness by Watson – while he simply denounced the methodological dead end of relying on conscious introspection in accounting for psychological reality. The alleged taboo has not prevented psychologists of various orientations from taking interest in conscious processes. Nor has it prevented members of the behaviorist school to do so, as was the case of Lashley, then in his behaviorist period, who published a long and important paper on the question (Lashley, 1923).

Be that as it may, Skinner was explicitly concerned with the problem of consciousness. He viewed it as a by-product of verbal activity shaped by the verbal community, providing individual speakers with tools for self-description and for reporting to others and to themselves their actions, feelings, and ideas. This approach is essentially similar with Vygotsky's thesis on the socio-discursive origin of consciousness, adopted and further elaborated by his followers, notably Luria. There is no evidence that Skinner was aware of the convergence with these Russian authors.

Also of interest in relation with the issue of consciousness is the discussion published in 1969 on the now very popular issue in Artificial Intelligence circles: could a machine be conscious? Skinner argues that "… the one currently irreducible difference between men and machines" is that: "They are built differently. The ultimate difference is in componentry … To be conscious or aware of itself as a man is conscious of himself a machine would have to be built like a man and would, of course, be a man." Far from ignoring consciousness, Skinner treated it recurrently with the attention it deserves in contemporary debates.

Education

Skinner's interests in applications covered three main areas: education, psychological treatment, and social practices at large. He dealt with the first two in a technical manner, developing principles and techniques for improving educational and therapeutic practices. His treatment of the third one is more akin to social philosophy than to scientific application, although he viewed it as very consistently rooted in his scientific thinking. At a time school education in the United States

was criticized for its deficiencies and for its inefficiency in competing with technological achievements of the Soviet Union, Skinner, as many other American scientists, inquired into the reasons of that state of affairs. Observing what was going on in any normal classroom, including in reputed schools, he concluded that it was violating blatantly most basic principles of learning derived from laboratory analysis of the learning process. Pupils and students were passively exposed to teachers' monologues rather than actively producing behaviors followed by feedback; there was no attempt to adjust the teacher's actions to individual level and rhythm of learning; negative evaluation based on mistakes and errors prevailed over positive evaluation pointing to progresses achieved; punitive controls, known to be poorly effective in shaping and maintaining complex behavior, were still widely used; general conditions and teaching practices were far from favorable to developing individual talents and creativity. Such criticisms had been made by others, but Skinner differed from them in the analysis of the causes and in the remedies proposed. He did not question the importance of endowing the students with basic knowledge and skills that they need if they are to engage in more complex and original activities. But he thought such skills could be mastered using more efficient methods than those currently in use in the classroom. This was the origin of teaching machines, a term that would raise strong objections on the ground that machines could not lead but to dehumanizing teaching. In fact, Skinner's idea has been implemented since then in computer-assisted learning, which is now widely accepted without any protest, but usually with little or no reference to the pioneering projects of the behaviorist psychologist. The device he designed in the 1950s appears quite primitive compared with modern computers: It was an electromechanical machine, adapted from a record player, built in such a way that it would present in a window to the student successive small frames of the material to be learned, each frame requiring an active answer from the learner. The latter could learn at his or her own individual rhythm, ideally with no or few errors, and eventually reach the end of the program, with the guarantee that the subject matter had been mastered from end to end. Good programs would make exams useless, if exams are just a way to control that the material has been covered and understood. Most important, student and teachers would not waste the few hours they could work together in tasks easily fulfilled using teaching devices; they could devote the time so spared to more constructive activities requiring direct human contact.

In spite of numerous attacks from educational circles, Skinner's project inspired many applications, such as programmed instruction in book form, until modern computers would offer the elegant solution we know today. It also contributed to the development of individualized teaching approaches that favor methods allowing students to learn at their own pace in an autonomous and active way.

Teaching machines are but one aspect of Skinner's contribution to education. In a number of his writings, including his utopian novel, Skinner (1968, 1978) did express his reflections on educational issues. He was concerned especially with the disproportion between the resources devoted to education and the poor outcomes; with the tendency to level down individual differences; with the increasing distance between the school environment and real life; with violence in schools, and other matters which remain crucial issues today, with little improvement.

Behavior Therapy

Treatment of psychological disturbances was another field of application to which Skinner (1955, 1989) made influent contributions. By the middle of the twentieth century, psychopathology had elaborated refined descriptive systems of psychological disturbances and equally sophisticated explanatory models such as the psychoanalytic theory. Contrasting with these, methods for treatment were scarce and their results poor. Psychoanalytic treatment had limited indications and practical limitations. Rogers's nondirective therapy, though quite popular, did not provide convincing results. Psychopharmacology was still in the air.

Skinner did not question the classical categorization of mental illnesses, nor did he propose miracle remedies. He simply suggested to look at them as disturbances in behavior, rather than alterations of hypothetical mental structures, such as the psychic apparatus appealed to by psychoanalysis, of which abnormal behavior would be but observable indicators, or symptoms. Consequently, he proposed to attempt to change undesirable behavior by acting directly upon it, rather than upon underlying structures supposedly responsible for it. This approach was not totally new: behavior therapy had its origins in John Broadus Watson's attempts to treat fear in children by resorting to Pavlovian conditioning and in the theoretical work of some neobehaviorists aimed at transposing some psychoanalytical concepts into learning theory models. What Skinner added was his genuine theoretical elaboration, especially based on his antimentalist stand, and techniques of behavior modification drawn from the operant laboratory, supplementing Pavlovian techniques in use up to then. He also brought into the clinical field a sense of rigor transferred from the laboratory, perfectly compatible with the study of single cases.

Skinner did not practice behavior therapy himself. His influence in the field was indirect, by stimulating pioneering research on psychotic patients and mentally deficient people. He gave a decisive impulse to the development of the behavioral approach to treatment, which soon acquired a major position in clinical psychology and counseling, and which eventually merged with cognitively oriented practices into what are labeled cognitive-behavioral therapies. This appears as a somewhat paradoxical evolution given Skinner's opposition to cognitivism. It suggests that practitioners, being professionally more pragmatic, have perceived complementarities between the two schools of thought rather than sticking to doctrinal divergences. Cognitive-behavioral therapies, claiming their evidence-based status, have gained an important share in psychotherapeutic practices around the world, supplanting psychoanalytically oriented treatments. However, their success varies depending upon the country. For example, on the European continent, while they have been widely accepted in most West countries, cognitive-behavioral therapies have been strongly opposed by a typically French

psychoanalytic tradition, in a conflict with institutional and even political implications.

Social Philosophy

Skinner's social philosophy was rooted in a deep confidence that only science would help us in solving the problems we are facing in our modern societies. What is needed is a science of behavior, which in fact is now available, he thought, and could be applied if only we would be ready to abandon traditional views of human nature. He first expressed his ideas in the novel *Walden Two* (1948). Written 2 years after the end of the World War II, the book describes a utopian community run after the principles derived from the psychological laboratory. It is by no means a totalitarian society, as some critics have claimed. Looked at retrospectively, it is surprisingly premonitory of social issues that are still largely unsolved more than half a century later. For instance, working schedules at *Walden Two* have been arranged so that all tasks needed for production of goods and good functioning of the community are distributed among members according to a credit system which results in an average amount of 24 h per week, avoiding unemployment, abolishing any social discrimination between manual and intellectual work, and leaving many free hours for leisure activities such as sports, arts, and scientific research. Emphasis is put on active practice rather than passive watching, on cooperation rather than competition. The community is not isolated culturally: cultural products from outside such as books or records are of course welcome, but radio programs are filtered to eliminate publicity. Education is active; the school building symbolically has no door separating it from the life and work community; there are no age classes, no humiliating ranking; all learn at their own rhythm, in whatever orientation they feel appropriate, throughout their lifetime. Women enjoy complete equality with men. Waste of natural resources is avoided. Special charges in the management of the community are strictly limited in time, eliminating any risk of political career.

The utopian novel has been for the author the occasion to discuss the practice of democracy, also dealt with in a number of his other writings. He denounced the counterproductive effects of the race or fight for power, generating excessive personalization of politics; the tendency of politicians to reject the responsibility of their failures on the citizens who had elected them; the lack of concern for minorities. By the time of the immediate after World War II, when he was writing *Walden II*, he lucidly qualified the process of election of the presidents of the United States as a "masquerade of democracy." He demystified the subtle ways politicians do mask the control exerted over people and he insisted on the importance of developing appropriate strategies of countercontrol.

Similar themes, plus the frightening concerns with pollution, violence, uncontrolled population growth, nuclear weapons, and the like, are further elaborated in the book *Beyond Freedom and Dignity* (1971) and a number of later articles. In an alarming tone, Skinner points to what he feels is the core of our inability to deal with these issues, that is our obstinacy in keeping a conception of human nature which scientific inquiry shows us to be wrong, and which bars any solution to the problems we are confronted with. We still stick to a view of humans as being the center of the universe, free and autonomous, dominating nature, while we are but one among many elements of nature. As a species, we are the product of biological evolution; as cultural groups, the result of our history; and as individuals the outcome of our interactions with the environment. Because we fail to admit this dependency, and draw the consequences of it, we might put in danger our own future. Freedom, autonomy, and merit are no absolute values: They were forged throughout history, and more often than not they are used to disguise insidious controls, the mechanisms of which should be elucidated if we want to develop countercontrols eventually allowing for the survival of our species.

Skinner and Other Great Contributors to Psychology of His Time

Skinner viewed these various facets of his work as closely related, making for a highly consistent theory of human behavior, in which the critical analysis of social processes in modern society was deeply rooted in the experimental analysis of the behavior of animal subjects in the laboratory. So global an ambition has been criticized, and clearly various aspects of his contribution did not have the same fate. If the operant technique is now a widely used procedure to many purposes in experimental research in psychology and related fields, if his early attempts to build teaching machines appear now as ancestors of computer-assisted learning and teaching, if a number of principles of contingencies analysis are now currently put in practice in behavior therapies, radical behaviorism has been seriously questioned and even shaken by the rise of the cognitivist approach in psychology, while Skinner's social philosophy has been attacked from different fronts both on ideological and scientific grounds.

As most great theory builders of the twentieth century in psychology, from Sigmund Freud and Watson to Jean Piaget, Skinner may be blamed for having reduced the explanation of human nature to a very limited set of concepts and findings, namely those he had forged and observed in his own restricted field of research and reflection, ignoring even other concepts and facts in neighboring fields of psychology, leaving alone of other sciences. It is clear that Skinner has made no attempt at integrating, for example, contributions of developmental or of social psychology, nor those of sociology, cultural anthropology, or linguistics. Such neglects might have been deliberate, legitimated by the will to concentrate on what were, in Skinner's mind, essential points left out by other branches of psychology or other sciences dealing with human societies. However, they might appear as sectarianism to those who favor an integrative and pluridisciplinary approach to the complex objects of human sciences. It cannot be decided whether his influence would have been larger or smaller had he adopted a less exclusive stand.

See also: Behavior Therapy: Psychiatric Aspects; Darwinism: Social; Educational Learning Theory; Educational Philosophy; Evolutionary Epistemology; Freud, Sigmund (1856–1939);

Lashley, Karl Spencer (1890–1958); Lorenz, Konrad (1903–89); Normality and Mental Health; Pavlov, Ivan Petrovich (1849–1936); Piaget, Jean (1896–1980); Psychological Treatment, Effectiveness of; Psychological Treatments, Empirically Supported; Psychology, History of (Twentieth Century); Thorndike, Edward Lee (1874–1949); Vygotsky, Lev Semenovic (1896–1934); Watson, John Broadus (1878–1958).

Bibliography

Bjork, D.W., 1997. B.F. Skinner, a Life. American Psychological Association, Washington, DC.

Changeux, J.-P., 1983. L'Homme neuronal. Fayard, Paris (The Neuronal Man).

Chomsky, N., 1959. Review of Skinner B.F., verbal behavior. Language 4, 16–49.

Edelman, G.M., 1987. Neural Darwinism: The Theory of Neuronal Group Selection. Basic Books, New York.

Lashley, K.S., 1923. The behavioristic interpretation of consciousness. Psychological Review 30, 237–272, 329–353.

Lorenz, K., 1981. Fundamentals of Ethology, New York, Springer.

Modgil, S., Modgil, C. (Eds.), 1987. B.F. Skinner: Consensus and Controversy. Falmer, New York.

O'Donohue, W., Ferguson, K.E., 2001. The Psychology of B.F. Skinner. Sage, London.

Richelle, M., 1993. B.F. Skinner, a Reappraisal. Erlbaum, Hove, London.

Roales-Nieto, J.G., Luciano Soriano, M.C., Prez Alvarez, M. (Eds.), 1992. Vigencia de la Obra de Skinner. Universidad Granada Press, Granada (Robustness of Skinner's Work).

Skinner, B.F., 1938. The Behavior of Organisms. Appleton Century Crofts, New York.

Skinner, B.F., 1948. Walden Two. Macmillan, New York.

Skinner, B.F., 1953. Science and Human Behavior. Macmillan, New York.

Skinner, B.F., 1955. What is psychotic behavior. In: Gildea, F. (Ed.), Theory and Treatment of the Psychoses: Some Newer Aspects. Washington University Studies, St Louis, MO, pp. 77–99.

Skinner, B.F., 1957. Verbal Behavior. Appleton Century Crofts, New York.

Skinner, B.F., 1961. Cumulative Record. Appleton Century Crofts, New York.

Skinner, B.F., 1966. The phylogeny and ontogeny of behavior. Science 153, 1205–1213.

Skinner, B.F., 1968. The Technology of Teaching. Appleton Century Crofts, New York.

Skinner, B.F., 1971. Beyond Freedom and Dignity, first ed. Knopf, New York.

Skinner, B.F., 1978. Reflections on Behaviorism and Society. Prentice-Hall, Englewood Cliffs, NJ.

Skinner, B.F., 1987. Upon Further Reflection. Prentice Hall, Englewood Cliffs, NJ.

Skinner, B.F., 1989. Recent Issues in the Analysis of Behavior. Merrill, Columbus, OH.

Skocpol, Theda (1947–)

Ian Sheinheit, University at Albany, SUNY, Albany, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Theda Skocpol, currently the Victor S. Thomas Professor of Government and Sociology at Harvard University, analyzes and researches the principal historical transformations of society. Skocpol's work is an exemplar of historical comparative research. She has asked, and of course answered, 'big' questions: What are the causes of social revolutions? Why has the United States welfare state developed as it has? What is happening to American civic life and democracy? These three questions have guided the three distinct yet overlapping tracks Skocpol's research has followed. The answers proffered by Skocpol and the concepts she developed and deployed to do so, causal historical analysis, state and polity autonomy, policy feedback, and diminished democracy, are described and discussed.

In a 2013 interview for Salon, an online political news magazine, Skocpol is asked a number of questions regarding the current state of politics in the United States. This interview is notable, not because of its social scientific contribution, but because of Skocpol's continued ubiquitous presence in the United States' political public sphere. Skocpol, currently the Victor S. Thomas Professor of Government and Sociology at Harvard University, is an exemplary and unique scholar that has contributed in a multitude of ways to a variety of disciplines, including Sociology, Political Science, and History, a variety of subfields including historical comparative analysis, comparative politics, and political sociology, and has had a palpable and important influence in public discussions. In this particular interview, Skocpol touches on the most important currents in her intellectual trajectory. That is, state autonomy, the history of U.S. social policy, and the structures and institutions of U.S. democratic practice.

Skocpol has addressed these disciplines, subfields, and concepts in 23 books and edited volumes as well as over 70 academic articles. Her pioneering book on social revolutions, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, is the most influential sociological analysis of revolutions to date. It is also where she first and most poignantly advances her theory of state autonomy. She followed this watershed book with the epic historical comparative account of the United States welfare state in *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Here Skocpol continues to advance her historical comparative method in conjunction with her focus on state autonomy. Since the publication of *Protecting Soldiers and Mothers*, Skocpol has furthered her academic reach by entering into the discussions surrounding civic engagement and declining democracy in the United States. Her book *Diminished Democracy: From Membership to Management in American Civic Life* challenges academic, and popular, arguments alike about the decline of citizen engagement. More recently Skocpol has continued to fortify her intellectual program on U.S. politics through research on health care policy (*Boomerang: Clinton's Health Security Effort and the Turn Against Government in U.S. Politics* and *Health Care Reform and American Politics: What Everyone Needs to Know*), Obama's presidency (*Obama and America's Political Future*), and the rise of the 'Tea Party' (*The Tea Party and the Remaking of Republican Conservatism*).

Theda Skocpol has had a fascinating life and early career. She was born in Detroit, Michigan. She received her undergraduate degree in sociology from Michigan State University. During her time at Michigan State, she worked with a civil rights education project at Rust College in Holly Springs, Mississippi. While working on this project she met her husband Bill Skocpol. Skocpol was deeply influenced by the political environment engulfing American campuses during the 1960s and 1970s (Weir, 2002). After graduating from Michigan State in 1969, she attended Harvard University where she studied with Barrington Moore Jr. (author of the groundbreaking work *The Social Origins of Dictatorship and Democracy*). She graduated from Harvard with her PhD in Sociology in 1975. Following the completion of her PhD, Skocpol became a professor of Sociology at Harvard University. She occupied this position from 1975 to 1981.

In 1980 Skocpol was up for tenure at Harvard University. After a five-to-five vote by the tenure committee, with one abstention, the department decided to deny Skocpol tenure. She would have been Harvard University's first female tenured sociology professor (Schumer, 1981). Skocpol, who had just won the American Sociological Association Award for a Distinguished Contribution to Scholarship for *States and Social Revolutions*, filed a suit of sex discrimination against the department and left for a tenured position at the University of Chicago. She remained at Chicago for 5 years and during that time an internal review at Harvard deemed that discrimination was in fact the cause for Skocpol's tenure denial. In 1986 Harvard offered Skocpol tenure and she returned to be the first tenured female sociology professor at Harvard.

Skocpol has had vast and prolific influence on the fields of sociology, political science, and American politics. In 1998 Skocpol became the Victor S. Thomas Professor of Government and Sociology at Harvard University. From 2000 to 2006, she was the director of Harvard's Center for American Political Studies. And, from 2005 to 2007, she was the dean of Harvard's graduate school of arts and sciences. She has received a number of awards for her work and her career achievements, and in 2007 she received the prestigious Johan Skytte Prize in Political Science.

For contemporary scholars working on the broad topics of states, politics, and historical and societal change, engaging Theda Skocpol's work is imperative. Following from the founders of sociology, Marx, Weber, and Durkheim, Skocpol has been interested in the principal historical transformations of society. Skocpol's work is an exemplar of historical comparative research. She has asked, and of course answered, 'big' questions: What are the causes of social revolutions? Why has the United States welfare state developed as it has? What is happening to American civic life and democracy? These three questions have guided the three distinct yet overlapping tracks Skocpol's research has followed.

What Are the Causes of Social Revolutions?

Skocpol's dissertation addressed what she deemed as problems existent in the scholarship analyzing revolutions. The book that resulted from her dissertation was *States and Social Revolutions*. In this book, Skocpol powerfully maps out the causes of the social revolutions in France, Russia, and China and convincingly argues that comparative historical analysis "is the appropriate way to develop explanations of revolutions that are at once historically grounded and generalizable beyond unique cases" (1979: p. 6). She accomplishes this through comparing, what she proffers as three similar or 'positive' cases, France, Russia, and China, and by comparing these cases to a variety of 'negative' cases. That is, cases that have similar causal variables and sequences but do not have similar results (in particular Russia 1905, Japan, Germany, and England).

To best clarify the appropriate cases under investigation, Skocpol clearly conceptualizes social revolutions. Skocpol defines social revolutions as the occurrence of societal structural change with class upheaval, and the occurrence of political with social transformation. She distinguishes social revolutions from political revolutions and rebellions. Political revolutions "transform state structures but not social structures, and they are not necessarily accomplished through class conflict." Rebellions, "may involve the revolt of subordinate classes [even when successful] but they do not eventuate in structural change" (p. 4). In this way, Skocpol can analyze the "relatively few historical instances" of social revolution with specificity and generalizability. Using positive and negative cases of comparison, we glean what variables are existent during successful revolutions that are not during failed social revolutions, political transformations, or rebellions.

Skocpol was discontented with the trends she had read in the current theory regarding revolutions. The theories existent on revolutions fell into two camps: the Marxist and the pluralist. Skocpol correctly highlights the disconnect present between theory and history within these two theoretical trajectories. She further quashes each of these approaches as they fall short of truly understanding the amalgamation of variables involved in creating the conditions for a successful social revolution. As Skocpol highlights, the leading models for understanding revolution all follow a similar, but misguided, pattern. That is, changes give rise to grievances and disorientation, which leads to a conscious and 'purposive, mass-based movement.' This movement "fights it out with … the dominant class and, if it wins, undertakes to establish its own authority" (pp. 14–15). Skocpol argues that this analysis is reductionistic especially in regard to its emphasis on a 'voluntarist' understanding of revolutions. Skocpol does not completely shunt aside ideological or cultural claims (see pp. 164–171) and she returns to them in 1994 in *Social Revolutions in the Modern World* (see pp. 199–210). However, in part because her analysis is on the macroscopic level, she dismisses an ideological perspective for a nonvoluntarist or nonintentionalist historically grounded structural approach.

Skocpol considers the constellation and sequence of contingent variables at work creating the conditions for a successful social revolution. These variables consist of the structural conditions of the old regime, intranational conflicts, international structures and world historical conditions, and the structure of the institutional apparatus known as the state. Prevailing theories of revolution, at that time, according to Skocpol, focused exclusively on intranational conflict, and reduced the state to a puppet of dominant class interests (Marxists) or as a representation of liberal society (pluralists). Skocpol, on the other hand, argues that the state is a potentially autonomous structure. The state has a logic and interests of its own that are 'not necessarily equivalent or fused with' the dominant class or 'the full set of member groups in the polity' (p. 27). This theoretical and empirical advance over past understandings of social revolutions has inspired a spate of research interested in the state as an autonomous actor (Skocpol, 1985).

In the end, Skocpol finds that revolutions are a three-stage process that happens in three short time periods. First, old regime states are weakened by international conflict in the form of military defeats or economic crises. The former simultaneously discredits rulers and makes them appear vulnerable. The latter upsets elites, who are dependent upon a strong state for their income, as well as the masses, who face deprivation and increases in taxes as the state attempts to generate new revenues. Second, the operation of the old regime's state apparatus is disrupted by elites and masses with impunity. Last, the old regime is killed or forced into exile and the new regime obtains the position of power (see Lachmann, 2013 Chapter 3 for an excellent survey and comparison of Skocpol's methods).

States and Social Revolutions was and remains an exceptional piece of scholarship that has had wide-reaching influence. It was cited as among the most influential books in sociology and political science in 1996, 17 years after its publication. It has been translated into nine different languages. It has inspired a deluge of research on revolutions and is the exemplar for the utilization of comparative historical analysis. Inspired by this work and an increasing desire to re-cement historical comparative methods in sociology, Skocpol wrote, with Margaret Somers, "The Uses of Comparative History in Macrosocial Inquiry" and edited a book titled *Vision and Method in Historical Sociology* (1984). Skocpol, also, coedited the book *Bringing the State Back In* (1985), where you can find some of the research deeply influenced by the theoretical and empirical thrust of the state-centered analysis exemplified in *States and Social Revolutions*. In *Bringing the State Back In*, you can also find Skocpol's programmatic title piece. Equipped with the methods and concepts developed and deployed in her work on social revolutions, Skocpol shifts her attention to the history of United States social policy.

Why Has the United States Welfare State Developed as It Has?

Skocpol's work following *States and Social Revolutions*, beginning in the 1980s, focused on understanding public social policies in the United States. Her contributions in this field have inspired and influenced debates in a similar manner as her earlier work on social revolutions. The first book she published answering questions about the United States welfare state and social policy was a coedited volume (with Margret Weir and Ann Shola Orloff) titled *The Politics of Social Policy in the United States* (1985). In this volume Skocpol and her colleagues further their discussion of the fruitfulness of the state-centered approach in determining what, why, and when policies get implemented. Following this edited volume, Skocpol published, *Protecting Soldiers and Mothers*, her epic analysis of the United States welfare state.

In *Protecting Soldiers and Mothers* (1992), Skocpol links Civil War pensions to the conditions of nineteenth century American politics and then to the reaction against such polices in the progressive era. She focuses her analytical and empirical attention to the conditions with which particular social policies passed or were defeated. Once again, Skocpol is attuned to the complex constellation of variables at work and is interested in debunking the idea, proffered in popular and scholarly perceptions alike, that generous social provisions were suddenly created in response to the great depression era. "At that point, most scholars presume that the United States joined the evolutionary mainstream of western social progress – as a 'laggard' on the universal road toward 'the modern welfare state'" (p. 4). This explanation, according to Skocpol, for what social policies are passed at particular times and defeated at others is tenuous at best. Dissatisfied with current research trends, Skocpol rejects the problematic premise that they all share "about the evolutionary nature of 'the welfare state' and the socioeconomic roots of political processes" (p. 39). Elucidating the state as an autonomous actor in the process and struggle for control over policy implementation, Skocpol argues, we do not fall prey to this same a priori trap.

The sociological puzzle to be pieced together is represented by this question: why did the United States not develop similar welfare states as their European counterparts? Utilizing 'European welfare states' for comparison, Skocpol finds that the United States Civil War pensions at first were, indeed, modest. However, during the time period of 1879–1906 social benefits were traveling toward a 'paternalist' state representing magnanimous social benefits for veterans and the elderly. During this historical moment, the trend of American politics toward 'paternalist' policies did not 'lag' behind their European counterparts. In fact, these provisions, as Skocpol demonstrates, accounted for 41.5% of the federal budget by 1893. Following this time period, however, a shift in the social benefit provisions occurred. Due in part to the perceived corruption by politicians and reformers of these provisions, Civil War pensions did not germinate. Alternatively, they became a warning sign for future similar policies. Skocpol argues that United States social policy then became, in large part because of women's civic associations, a 'maternalist' state. By the 1920s, however, these provisions also lost steam and the United States never "quite crystallized into a full fledged maternalist welfare state" (p. 534).

Three of the historically grounded conceptual ideas, that have had extensive influence, present in *Protecting Soldiers and Mothers* should be highlighted: state autonomy (polity autonomy), policy feedback, and the impact of women's civic associations. First, not surprisingly, Skocpol finds much of the causal activity in the process of policymaking in the state. She broadens her state-centered analysis to more effectively include all "political activities whether carried out by politicians or social groups" (p. 41). She terms this a "polity-centered approach." Political activities are, in congruence with the state-centered approach, "conditioned by the institutional configurations of governments and political party systems" (p. 41). Second, the concept of policy feedback highlights Skocpol's emphasis on historical and political contingency. That is, policies are not only a dependent variable but once instated they have independent causal thrust. A historical comparative framework will have to understand that the introduction of particular policies impacts future policy and political trajectories. This concept was utilized previously by Orloff and Skocpol (1984) and discussed by Skocpol and Amenta (1986).

Third, Skocpol brings much needed attention to the role women's civic association played in the development of a 'maternalist' welfare state in the United States during the early twentieth century. For Skocpol, the deficient development of social politics for American men, when compared to European welfare states and American women, can be historically traced to "federated white women's clubs" (1992: p. 26). Dissimilar to American men reformers, who only took their case to a limited circle of elite legislators, women built a coalition from the ground up. Grassroots organizations that mobilized cross-class coalitions, encompassing 'upper and middle class women, joined by some wives of skilled workers,' were exceptionally successful during a short period of time at the turn of the twentieth century (p. 51). Unlike their male counterparts, Skocpol argues, women in their fight for suffrage created the structural seeds for mobilization. "Tying together clubs based in local legislative districts into statewide and nationwide networks of social communication … these widespread associations were well situated to promote specific kinds of social policies" (p. 56). These policies were embodied by 'maternalist' ideals. The emphasis Skocpol placed on marginal groups,' in this case women, historical influence on policymaking processes was pioneering. The analysis of the state institution and policy feedbacks continued Skocpol's insights on elites who affect policy formation. The addition, and impact found, of women's civic associations adds veracity to the process of policymaking and foreshadows Skocpol's future research projects on civic associations.

Skocpol continues her analysis of the failure of the United States to introduce national social insurance programs in her more recent work in *Boomerang* (1997a), *Health Care Reform and American Politics* (2010), and *Obama and America's Political Future* (2012). In *Boomerang*, Skocpol traces the political and state-oriented constraints that the Clinton administration navigated unsuccessfully. The popular presumption about the 'stupidity' amongst the Clinton administration or the fickleness of American public opinion was not the cause of Clinton's health care failure. Rather, Skocpol focuses our attention on the structure and constraints of the state and the political environment. In the end, as Skocpol demonstrates, the political

repulsion of the signifier 'taxes,' the report of the Congressional Budget Office (CBO) and the decisive power of private insurance firms blocked the Clinton health care bill and created a reaction against government. Jacobs and Skocpol continue this line of thought in *Health Care Reform and America Politics*. They describe the policy feedback mechanisms of a failed Clinton plan's influence on the construction of Obama's Affordable Care Act. In the end, Obama's legislation, according to Jacobs and Skocpol, was the result of a "chain of contingent events that transformed the relationship among physicians, hospitals and insurance companies, and thereby shaped each of those groups' interests and their capacities to act politically alone or in alliance with others" (Lachmann, 2013).

Further research on national insurance programs in the United States is conducted in the book *Obama and America's Political Future* (written by Skocpol with commentaries by Larry M. Bartels, Mickey Edwards, and Suzanne Mettler). The incipient objective of *Protecting Soldiers and Mothers* was to understand New Deal policies in American politics (pp. vii–ix). That book transformed into an exploration of policymaking prior to the New Deal that necessarily brought women to the forefront of historical policy research. The original goal, however, was returned to in her essay "Obama's New Deal, Tea Party Reaction, and America's Political Future" in the book *Obama and America's Political Future*. Obama ran and won in 2008 on the grounds of passing new legislation with similar ideals to Franklin D. Roosevelt (FDR) in his New Deal era. In her essay Skocpol, consistent with her historical comparative analysis, compares Obama's first years in office to FDR's. Skocpol proffers that the timing of the economic downturn, the structure of party polarization, the volatility of the media institution, policy feedback mechanisms and backlash in the form of the emergence of conservatism represented by the Tea Party, created the conditions for Obama's more modest accomplishments (More specifically, Skocpol proffers five reasons for Obama's more modest accomplishments. First, unlike FDR and Hoover, Obama necessarily, because his economic downturn was at an earlier stage, had to cooperate with George W. Bush and his economic advisors. Second, the polarized structure of American politics for Obama was different from FDR's time. Polarization, for Obama, is structured purely along party lines. Experts, therefore, on the opposite side of the political spectrum, that knew the economic downturn would continue, had no desire to make concessions. Third, the structure of the media was in, and still is in, the midst of a massive transformation represented by increasingly successful openly aligned partisan media outlets and economic turmoil. These changes increase complications for a president trying to navigate such a volatile terrain. Fourth, FDR built his policies anew, while Obama, on the other hand, unavoidably had to redirect existing path-dependent government institutions and path-dependent policies. Finally, in part due to the new structure of partisan media, is the emergence of a political opposition, similar to the civic associations of America's past, called the Tea Party (Skocpol and Jacobs, 2011).).

In order to answer questions regarding the United States welfare state, Skocpol necessarily elucidated the complexities of historical contingencies. *Protecting Soldiers and Mothers* demonstrated that universal welfare states do not develop in an evolutionary process. Instead, particular policies, during particular historical moments, are constrained by the structure, relationship, and interaction between a variety of political actors. Comparable to social revolutions, social policies are created when an opening in the state structure coincides with the political action of civic associations, as exemplified by the maternalist policies enacted at the turn of the twentieth century. Health care policy tells the opposite side of this story. Skocpol, inspired by her work on women's civic associations, continued down this road and began thinking about her most recent question.

What Is Happening to American Civic Life and Democracy?

For analytical purposes I have separated this article into three sections representing three distinct questions that Skocpol has answered throughout her career. These three trajectories, however, have overlapped empirically, as are exemplified by the emphasis on women's civic associations in *Protecting Soldiers and Mothers* and the discussion of the contemporary civic associations such as the Tea Party in *Obama and America's Political Future*. Skocpol's influential work on civic engagement deserves some independent attention. In 1999 Skocpol coedited *Civic Engagement in American Democracy*. In this book Skocpol wrote the chapter "Advocates without Members: The Recent Transformation of American Civic Life." In this article Skocpol carefully crafts the alterations occurring in the organization of American civic life. Following this she wrote *Diminished Democracy* (2003). In this book, Skocpol highlights, using primary historical documents, the changes in the institutional structure of civic associations in the United States over a century. Her analysis on American civic engagement, which represents a major progression from the existent literature, is twofold. First, she wants to understand civic decline and diminished democracy and second, consistent with all of her other research tracks, she wants to understand how civic associations in civil society can enact political power and potentially get social policies passed.

Observing the trends in American civic engagement has a long history. Tocqueville was first enamored with American's propensity toward joining civic associations in civil society in the 1830s. Skocpol (1997c) enters into this debate armed with historical insight and evidence to add much needed sophistication. She finds that nearly two-thirds of the largest civic associations in the 1950s, though claiming nonpartisanship, 'were directly involved in legislative campaigns' or political efforts of some sort (Skocpol, 2003: pp. 26–30). Further, members joined local chapters of associations for a myriad of reasons, often including social, not political, benefits. Skocpol argues that this empirical fact is important to democracy. Members active in local chapters learned important democratic skills including "how to run meetings, handle moneys, keep records, …participate in group discussions" and recruit new members. Moreover, with the democratic skills learned in these local chapters members could "move up to responsibilities at the district, state and national levels" (Skocpol, 2011: p. 115).

These civic associations with their federated structure, Skocpol finds, were "highly effective at mobilizing members across the nation to lobby for government programs that

addressed concerns that members raised at the local level" (Lachmann, 2010: p. 153). The political prowess of these types of civic associations is exemplified by the America Legion's successful push for the inclusive and magnanimous legislation represented by the G.I. Bill. Skocpol analyzes the contingent and historical variables at work in the process with which this bill was passed in her article "The G.I. Bill U.S. Social Policy, Past and Future" (1997b). In this article she also discusses the concept of policy feedback to highlight the impact the G.I. Bill has had and will continue to have on future policy possibilities in the United States.

During the twentieth century, however, Skocpol and other scholars (see Putnam, 2000) began observing a decrease in engagement, demonstrated by the dissipation of American civic associations. Skocpol finds that up until the early 1960s Americans were ardent joiners and members of civic associations, especially when compared to their European counterparts. Skocpol brings our attention, diverging from previous scholarship, to the "rapid civic reorganization between the mid 1960s and the early 1990s during which elite, well-educated Americans abandoned cross-class membership federations while launching professionally managed organizations" (Skocpol proffers four reasons for this civic transformation. First, unlike the Civil War, WWI, or WWII, the Vietnam War did not inspire the same fraternity among veterans. Second, the civil and human rights revolutions of the 1960s and 1970s challenged the historical gender roles and racial divides characteristic of past civic associations. Thus, "younger generations were reluctant to join" (Putnam, 2000). Third, as women entered the workforce they were no longer available to cater and work to keep the traditional male civic associations afloat. Fourth, emphasizing state autonomy the government entered into a variety of realms of social life fostering the opportunity for a deluge of new professionally run civic associations.) (Skocpol, 2003, 2011).

Professionally run civic associations are organized in a much different way than those previously prominent. These organizations are not structured as national federations with local chapters. Rather these associations are elite-run advocacy groups utilizing the availability of new technology to administer massive mailings and phone calls. Instead of vibrant local chapters with dues paying members these new associations hinge on a large membership of check senders (Skocpol, 1999, 2003). New professional associations are often elite-run and not intraclass, one of the most important characteristics of the old associations. These professional organizations are not conducive to democratic participation; rather they take the power out of the hands of everyday citizens and place them in the hands of elites. This historical context, the professionalization of civic associations is, according to Skocpol, the major example of diminished democracy.

Skocpol, in her coedited volume (with Lawrence R. Jacobs) *Inequality and American Democracy: What We Know and What We Need to Learn*, addresses the impact this structural transformation of civic engagement has on inequality. Skocpol argues that because experts are more valued and paid, and professionally run advocacy groups only need citizen money to run effectively, there is no need for regular citizens to participate. Second, "because higher-educated Americans have experienced sharply rising incomes in recent decades, they are also the folks who can afford to pay for professionally managed advocacy efforts" (Skocpol, 2004, 2011: p. 116). Historically, popular civic associations that were cross-class advocated for 'inclusive' social policies. Skocpol suggests that there is no reason to think that the new professionally run groups will continue this trend. Thus, Skocpol argues, the structural alterations in civic life have not only diminished democracy but also have increased inequality.

Recognizing an anomaly in the shift toward professionally run advocacy groups in civil society, Skocpol turned her analytical gaze toward the Tea Party. Skocpol with Vanessa Williams, in their book *The Tea Party and the Remaking of Republican Conservatism*, analyze historical and demographic data in conjunction with interview research on the Tea Party they conducted. They find that there are around 900 local Tea Party chapters around the country. Their analysis demonstrates that these are organizations consisting mostly of concerned older white citizens. The issues these citizens underscore are the undeserving 'freeloaders' and immigrants who have not 'earned public support' (2012: pp. 60–68). This, however, does not capture the entirety of the Tea Party. Professionally run advocacy groups that existed prior to the emergence of the Tea Party in 2009 joined forces with the grassroots organizations. The combination of an old civic association and new professionally run advocacy groups, Skocpol argues, has given, and will continue to give, the Tea Party power to elect congruent political officials and impact the process of policymaking in the United States.

Skocpol's research recognizes the complexities of American civic engagement and democracy. Through a historical paradigm consistently attuned to comparison, Skocpol is able to fine-tune and clarify our understanding of the constraints and effects the structure of states and institutions can have on civic engagement and democracy. Simultaneously, Skocpol's analysis is able to illuminate the impact citizens can have on political processes. Skocpol understands that in order to unravel the intricacies of American politics, or the causes of social revolution, elucidating both of these variable sequences is imperative.

Three Tracks, One Goal

Theda Skocpol was not afraid to confront the principal questions of society. In her career she addressed three of the most important questions concerning sociologists, political scientists, and historians alike. These three questions sent Skocpol down three distinct yet overlapping empirical and analytical tracks. Utilizing historical comparative methods, while following all three of these tracks, Skocpol has consistently demonstrated the fruits of an institutionally grounded macro historical sociology. Her influence on the disciplines of sociology, political science, and history is momentous. It must not be forgotten, however, that Skocpol's influence has stretched well beyond academic walls. Any scholar or politician interested in the vicissitudes of political institutions and policies would do well learning from Skocpol's intellectual cannon.

Skocpol's work is and should continue to be a template of rigorous, robust, historically grounded, theoretically engaged scholarship. Her historical comparative analysis is unique in

its depth and breadth. The concepts developed and deployed in her work, causal historical analysis, state and polity autonomy, policy feedback, and diminished democracy have inspired a flood of future scholarship. Her overarching goal is to understand and connect the conflicts existent in historically grounded institutional structures in order to decipher crucial political outcomes. Understanding Skocpol's work is imperative for a practitioner from any discipline studying revolution, social policies, civic engagement, or democracy.

See also: Civil Society and the Public Sphere: History of the Concept; Comparative Method in Anthropology; Governance; Health Politics and Policy; Policy Cycle; Political History: History of Politics; Political Parties, History of; Political Sociology; Poverty Law: United States; Poverty Policy; Revolution, Theories of; State, Sociology of the; Voluntary Associations, Sociology of; Welfare State, History of; Welfare: Philosophical Aspects.

Bibliography

Eidelson, J., 2013. Tea Partyers' Grave Fear: Why They Disdain Young People – Even Their Own! Salon.

Jacobs, L., Skocpol, T., 2005. Inequality and American Democracy: What We Know and What We Need to Learn. Russell Sage Foundation, New York.

Jacobs, L., Skocpol, T., 2010. Health Care Reform and American Politics: What Everyone Needs to Know. Oxford University Press, New York.

Lachmann, R., 2010. States and Power. Polity Press, Malden, MA.

Lachmann, R., 2013. What Is Historical Sociology? Polity Press, Malden, MA.

Orloff, A.S., Skocpol, T., 1984. Why not equal protection: explaining the politics of public social spending in Britain, 1900–1911, and the United States, 1880s–1920. American Sociological Review 49, 726–750.

Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.

Schumer, F.R., 1981. A Question of Sex Bias at Harvard. New York Times.

Skocpol, T., 1979. States and Social Revolutions: A Comparative Analysis of France, Russia, and China. Cambridge University Press, New York.

Skocpol, T., 1984. Vision and Method in Historical Sociology. Cambridge University Press, New York.

Skocpol, T., 1985. Bringing the state back. In: Evans, P., Rueschemeyer, D., Skocpol, T. (Eds.), Strategies of Analysis in Current Research. Bringing the State Back. Cambridge University Press, New York, pp. 3–43.

Skocpol, T., 1992. Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States. Harvard University Press, Cambridge, MA.

Skocpol, T., 1997a. Boomerang: Health Care Reform and the Turn against Government. W. W. Norton and Company, New York.

Skocpol, T., 1997b. The G.I. Bill and U.S. social policy, past and future. Social Philosophy and Policy 14 (2), 95–115.

Skocpol, T., 1997c. The Tocqueville problem: civic engagement in American democracy. Social Science History 21 (4), 455–479.

Skocpol, T., 1999. Advocates without members: the recent transformation of American civic life. In: Skocpol, T., Fiorina, M.P. (Eds.), Civic Engagement in American Democracy. Brookings Institution Press, Washington, DC, pp. 461–510.

Skocpol, T., 2003. Diminished Democracy: From Membership to Management in American Civic Life. University of Oklahoma Press, Norman, OK.

Skocpol, T., 2004. Voice and inequality: the transformation of American civic democracy. Perspectives on Politics 2 (1), 3–20.

Skocpol, T., 2011. Civil society in the United States. In: Edwards, M. (Ed.), The Oxford Handbook of Civil Society. Oxford University Press, New York.

Skocpol, T., 2012. Obama and America's Political Future. Harvard University Press, Cambridge, MA.

Skocpol, T., Amenta, E., 1986. States and social policies. Annual Review of Sociology 12, 131–157.

Skocpol, T., Jacobs, L., 2011. Reaching for a New Deal: Ambitious Governance, Economic Meltdown, and Polarized Politics in Obama's First Two Years. Russell Sage Foundation, New York.

Skocpol, T., Somers, M., 1980. The uses of comparative history in macrosocial inquiry. Comparative Studies in Society and History 22, 174–197.

Skocpol, T., Williamson, V., 2012. The Tea Party and the Remaking of Republican Conservatism. Oxford University Press, New York.

Weir, M., 2002. Theda Skocpol: "probing the institutional roots of politics". PS: Political Science and Politics 35 (4), 769–772.

Slavery and Violence

Enrico Dal Lago, National University of Ireland, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Throughout history, slavery has always been associated with violence. Following the different stages of the slavery experience, it is possible to focus first on the process of enslavement as a violent practice either resulting from warfare or from slave trades, from antiquity through the medieval and modern periods. Then, it is important to look specifically at historical forms of domination and control of the slaves and at the slaves' attempts to resist the violence of slavery through violent acts of resistance and rebellion. In most cases, though, the end of slavery came as a result of pressure brought upon governments by abolitionist movements and by other factors, with the consequent implementation of decrees sanctioning the peaceful transition to emancipation.

Introduction

As a violent system of exploitation and oppression, slavery has always been connected with violence throughout its long history. Scholarly studies on slavery have attempted to define the exact characteristics of the type of violence at the heart of slavery. Notably, Orlando Patterson, in his seminal monograph *Slavery and Social Death* (1982), has compared a range of societies across times and continents, where slavery was present in different shapes and degrees, and has formulated the now famous definition of slavery as "the permanent, violent domination of natally alienated and generally dishonoured persons" (Patterson, 1982: 13). In other words, for Patterson and for most slavery scholars, who followed in his path, the violence of slavery focused specifically on the characteristics of violent domination over individuals who had been uprooted through the process of enslavement and who were kept isolated, or alienated, as if they were reduced to an extension of their owners, in a state of symbolic 'social death.' Therefore, in order to investigate the constituent elements of the complex relationship between slavery and violence as it developed through history, we need to look first at the violent process of enslavement, and then at the types of masters' actual domination and control, and the consequent resistance, violent and nonviolent, to both by the slaves.

Historical Processes of Enslavement

Throughout history, the experience of slavery began in the most violent possible way for slaves with the trauma associated to the experience of enslavement. The violent process of uprootedness at the heart of the trauma of enslavement – which, for Joseph C. Miller, was a direct consequence of the dynamic process and historicized strategy of 'slaving,' or of mobilizing directly controlled human resources – left slaves helpless, violated, and vulnerable (Miller, 2011). In antiquity, the most common form of violent enslavement was through warfare. All the great powers of the ancient world, beginning from ancient Egypt and the Mesopotamian civilizations and ending with classical Greece and Rome, engaged in mass enslavement as a common practice related to warfare and conquest. The largest numbers of enslaved enemies could be found among the captives of great expansionist empires, such as, for example, Assyria, and, on an even larger scale, ancient Rome. Through relentless warfare and conquest, the Roman Republic and then the Roman Empire created a slave system with a flourishing slave trade and ubiquitous markets for the sale of slaves, which encompassed the entire Mediterranean area – effectively, the largest slave system in premodern times. At the height of Rome's expansionist period, Roman armies were able to enslave half a million people every year, so that, by imperial times, 10–15% of the whole population of the Roman Empire (50–100 million people) were enslaved; in Italy alone, as much as 30–40% of the population was made of slaves (McKeown, 2011).

Warfare continued to be the main cause of enslavement also throughout the medieval period. Often, violent enslavement was a direct consequence of the conquest and capture of enemies of a different religion; while Muslims had no qualms about enslaving Christians, Christians, in turn, routinely enslaved practitioners of Islam. A number of both regional and long-distance slave trades flourished in different areas of Europe and the Mediterranean, where slave markets could be found in all the main cities, showing that the business of capturing, selling, and purchasing slaves was a flourishing one during the entire Middle Ages. However, in northern regions, especially in northwestern Europe, between the thirteenth and the fourteenth centuries, slavery was gradually replaced by serfdom, while in eastern and southern Europe it went on for longer. While regions of eastern Europe, such as the borderland Slavic areas and the Black Sea coasts, supplied large numbers of slaves at least until the 1500s, during the same period both Italy and Spain witnessed the flourishing of a long-distance slave trade of Africans practiced by both Christians and Muslims in the Mediterranean (Phillips, 1985). With the quickening of pace in Ottoman expansion, in the 1500s and 1600s, the Christian–Muslim conflict in the Mediterranean resulted in massive enslavement of thousands of Europeans by the Barbary Coast corsairs, who conducted their raids from their bases in Algiers and Tunis; according to Seymour Drescher, in the period 1580–1680 alone, "an average of 7000 Christian captives per year were enslaved in the Maghreb," in North Africa. Most of the enslaved captives were Spanish, Portuguese,

French, and English, and these nations responded in kind by conducting raids for the purpose of enslaving North African Muslims throughout this century and afterward (Drescher, 2009: 28).

Starting from the mid-1400s, for four centuries, Europeans removed forcibly and violently from the West Coast of Africa more than 12 million people for the purpose of transporting them to labor on slave plantations in the New World. Succeeding one another in the control of sections of the Atlantic Ocean, partly relying on preexisting African traditions of slavery, partly creating the preconditions for the rise of African states trading in slaves, and mostly simply kidnapping and uprooting Africans of different ethnicities – particularly in the Gulf of Guinea, in the Gold Coast and Ivory Coast, in Senegambia, and in Angola – Europeans of different nations took the violent practices of enslavement and slave trade to an all-time high (Blackburn, 1997). First the Spanish and Portuguese, then the Dutch and the French, and ultimately the British, the Americans, and the Brazilians engaged in a lucrative forced transportation of slaves through the Atlantic – the 'middle passage' (a common expression used to describe the Atlantic slave trade from Africa to the Americas) – that caused close to 2 million deaths by the time Spain's slave trade, the last one operating, was definitely abolished in 1867. Whichever European power captured them, and whichever final destination in the Americas they were bound to, for all African slaves the 'middle passage' was a horrific ordeal, coming after the already extremely violent and traumatic experiences of being kidnapped, sold, and enslaved. Crammed in the belly of the ships so as to occupy all the available space, chained to one another in 0.5-m-high decks, slaves were left in the dirt for a period that often reached 2 months. Inevitably, the diseases, the exhaustion, the brutal discipline, and the suicides and rebellions provoked a very high mortality rate. For those who survived, though, the arrival in the New World was just the beginning of another traumatic and violent ordeal. If they were not branded and immediately dispatched to their new owners, slaves were cleaned up to mask their wounds and were placed on the auction block of a slave market; there, they were sold to the highest bidder, who would inevitably take them to a life of toiling in the fields or in the house of a plantation or farm (Rediker, 2008).

Slavery, Exploitation, and Control from Antiquity to the Modern Period

Even though, throughout history, there have been many societies that have utilized slavery as a violent and profitable system of oppression and exploitation, the importance of utilizing enslaved labor for the functioning of those societies has changed according to time and place. Following the definition first devised by Moses Finley, and then also by Keith Hopkins, most systems utilizing slavery in the course of history can be described as 'societies with slaves' in which slavery was present, but was not essential to the economic, social, and political structure. Conversely, only a few systems that utilized slavery can be considered 'slave societies,' in which slavery was at the heart of the economic, social, and political structure, and slaveholders and those who aspired to be slaveholders were the majority. For Finley and for a number of other scholars, only Classical Athens, Republican and Imperial Rome, the Early Modern and Modern Caribbean, the American South, and Brazil could be classified as authentic 'slave societies' (Burnard and Heuman, 2011).

In the classical world, slaves were ubiquitous and could be found performing all types of tasks in every dimension of the economy and society. Most of them tended to be household slaves, as in all premodern societies, but a large number of them were employed in back-breaking labor in agriculture. In either case, both in classical Athens and in Republican and Imperial Rome, slaves were subject to indiscriminate violence, and they were essentially at the whim of their masters, as they had little or no legal protection. Similarly to subsequent slave codes, particularly Roman law considered slaves as property, and violent punishment, torture, and murder of slaves by slaveholders were fairly common occurrences. In terms of treatment, though, Rome's household slaves fared probably better than others, especially if they were educated, since they could be given administrative tasks – which allowed them a higher quality of life than that of most slaves. The opposite was true of the slaves employed in the large farms and *latifundia*, and, even worse, in the mines. Similarly to slaves employed in other economically profitable businesses at later times, Roman agricultural slaves, who were essential to the functioning of the Roman economy, were chained, kept in barracks, and forced to hard labor from dawn to dusk under the constant supervision of overseers, who were in charge of their discipline and punishment (McKeown, 2011).

Both in medieval Europe and in the Islamic world, the substantial distinction between domestic and agricultural slavery persisted long after the fall of Rome, and the two types of forced labor maintained similar characteristics to the ones they had in Roman times, even though agricultural slavery was no longer an essential part of a large and integrated slave-based economic system. However, in northern Europe, by the later Middle Ages, serfdom replaced slavery in most agricultural regions, and consequently the legal conditions of agricultural laborers improved also in terms of protection from arbitrary violence and abuse. Conversely, slavery continued longer in the Mediterranean regions, and slaves coming both from eastern Europe and Africa were commonly employed under harsh, violently exploitative, labor conditions in agricultural labor in Palestine's Crusader states, in the Mediterranean islands, and in Spain and Portugal (Phillips, 1985). By the mid-1400s, when the Iberian kingdoms launched into the conquest of the Atlantic islands, and then of the New World, a tradition of violent exploitation of mass slave labor associated with profitable agricultural production, especially of sugar, was well established among southern Europeans and was exported by them into the Spanish Caribbean and Portuguese Brazil. Eventually, other European powers – notably the Dutch, then the French, and the British – also implemented plantation slavery, relying on the increasingly larger labor force violently uprooted from Africa through the transatlantic slave trade, and eventually succeeding one another as the dominant Atlantic slave economies in the Early Modern Period (Blackburn, 1997).

In the final decades of the eighteenth century, the end of colonial rule in North America's British Atlantic possessions, in Portuguese Brazil, and above all in French St. Domingue led to the making of the 'second slavery' (the type of capitalist oriented and highly exploitative slave system that characterized the nineteenth-century United States, Cuba, and Brazil) in the Americas – a slave system that was not necessarily tied to the logics of imperial expansion, but rather to the logics of capitalist profit in the world economy. Specifically in the US South, Cuba, and Brazil, the nineteenth century witnessed the growing and consolidation of the 'second slavery' through the mass organization of slave labor for the production of cotton, sugar, and coffee – the three New World agricultural products that dominated the world market. With the advent of the 'second slavery,' the type of violence that was intrinsic to the very texture of slavery became an indispensable component of a system of labor regimentation whose strict organization, enforced through constant supervision and a variety of disciplinary punishment, guaranteed the masters' extraction of the maximum amount of work from the slaves (Tomich, 2003).

Thus, in this sense, the 'second slavery' represented both the ultimate transformation and also the peak of an evolution of exploitative New World slave systems geared toward the maximization of production of particular agricultural items for sale and profit on the world market. Under the 'second slavery,' in response to the constant rise in market demand, pressure was placed on slaves on a continuously increasing basis in terms of regimentation and pace of work, both of which were enforced through the constant, efficient, and brutal use of violence, in order to guarantee fast and efficient mass production of particularly valuable cash crops. This was especially the case with the exhausting 'gang system,' according to which slaves worked divided in gangs, alternating each other in the different tasks required by the agricultural production in a never-ending routine of work operations that lasted uninterruptedly from dawn to dusk. Together with the consequences of the high level of violence implemented by masters, overseers, and slave drivers, this mindless routine resulted for the slaves, inevitably, in a phenomenon of workers' alienation comparable, with all the obvious differences, to the one that later on characterized twentieth-century factories (Dal Lago, 2012).

Resistance, Rebellion, and Revolution

In all the slave systems that existed in human history, slaves devised a variety of ways to resist the violent oppression brought upon them by slaveholders. Many of these forms of resistance were based on individual actions and were also nonviolent. Regarding slaves working in the fields, depending on the circumstances of the agricultural production and on the nature of supervision, resistance typically manifested itself through slowing or halting the pace of work, feigning illness, and, most dramatically, running away. The latter could sometimes take the form of mass flight, if the circumstances allowed it, as in the case of the thousands of Athenian slaves who fled during the Peloponnesian War (431–404 BC), or as in the case of the hundreds of thousands of runaway slaves in the Confederate South during the American Civil War (1861–65) (Burnard and Heuman, 2011).

In the Americas, especially during the Early Modern Period, mass flight of slaves led to the establishment of long-living 'maroon' communities, sometimes hosting thousands of runaway slaves. The largest of these communities was Palmares, in the captaincy of Pernambuco, in northeastern Brazil. Lasting from the early years of the seventeenth century to 1694, Palmares had a population of 11 000 runaway slaves, and it functioned as an independent kingdom, whose inhabitants fought fiercely in the wars brought upon them by both Portuguese and Dutch authorities in order to defeat them and re-enslave them. Especially in the eighteenth century, similarly to Palmares, even not on the same scale, large 'maroon' communities flourished also in other areas of the Americas, such as the interior of Suriname, eastern Cuba, and inland Jamaica. In all these cases, the existence of independent communities of runaway slaves prompted a reaction resulting in repeated violent attempts to suppress them, also because 'maroons' were instrumental in leading the other slaves into open rebellion against slaveholders. In Jamaica, in particular, the attempts at violent suppression of 'maroon' communities led to two prolonged conflicts: the First Maroon War (1665–1740) and the Second Maroon War (1795–96) (Blackburn, 2011).

The most straightforwardly violent form of slave resistance was outright rebellion. Throughout its history, slavery witnessed numerous examples of slave rebellions of different sizes and with different scopes. The rationale behind them was always the acquisition of freedom through violent means, but sometimes slaves could have a wider scope of overthrowing the entire slave system in an entire region as a first step toward ending slavery itself everywhere. Also, in general, aside from the only successful case of St. Domingue, all violent slave rebellions had negative results in both the short term, with the subsequent violent suppression of the rebels, and also the long term, with the increasing of violence in measures of oppression geared toward preventing the further occurrence of rebellions. The largest economic power founded on slave agricultural labor in the premodern world, ancient Rome, was shaken at the very core of its foundation by three large slave revolts occurring within the space of 100 years in the last two centuries of the Republican period and involving tens of thousands of slaves: the First Sicilian Slave War (137–132 BC), the Second Sicilian Slave War (104–100 BC), and the Spartacus Revolt (73–71 BC). All were violently suppressed by the Roman army with thousands of dead (Bradley, 1989).

In medieval times, an equally large slave revolt involved thousands of enslaved Zanj Africans, who rebelled against their Arab masters in the Abbasid Caliphate in AD 869–883. In the Americas, violent slave revolts occurred almost from the very beginning of the establishment of slave societies and continued throughout the Early Modern and Modern periods until slavery's final abolition. Many of these revolts actually remained at the level of large conspiracies, never materializing into action; many others, though, involved large numbers, from tens, to hundreds, to several thousand slaves. The following were among the most important large-scale slave rebellions in the Americas: Tackey's revolt (1760) and the Baptist War

(1831–32) in Jamaica; the Bahia slave revolts (1807–35) and the Male rebellion (1835) in Brazil; the Aponte (1812) and La Escalera (1843–44) rebellions in Cuba; and the Stono rebellion (1739), Gabriel's rebellion (1800), the German Coast uprising (1811), the Denmark Vesey conspiracy (1822), and the Nat Turner rebellion (1831) in the American South. Ultimately, regardless of the odds against them and the almost inevitable failures they were likely to face, through their violent rebellions, slaves demonstrated – in Gad Heuman's words – "that the enslaved were often prepared to resist their enslavement violently" (Heuman, 2011: 231).

By far, the most important example of violent slave rebellion in world history is the Haitian Revolution in St. Domingue, which occurred at the end of the eighteenth century and effectively decreed the end of the colonial slave system in the New World. In the second half of the eighteenth century, the French colony of St. Domingue was the leading producer of sugar, which was grown in large plantations with hundreds of slaves owned by a small elite of white planters and mulattoes, both resentful of their dependence on France; at this time, also a respectable minority of free blacks was present in the island. The revolutionary events that started in France in 1789 provided the white elite of St. Domingue with the opportunity to seek autonomy from the French metropole through the drafting of a constitution, in 1790, that limited French control. Yet, in 1791, taking advantage of their much superior numbers, St. Domingue's slaves organized a series of revolts that, by 1792, encompassed the entire Haitian territory; as a consequence, most of Haiti's white population was slaughtered in the process. By 1793, self-liberated 'black Jacobin' Toussaint L'Ouverture, acknowledged leader of the Haitian Revolution, was able to proclaim the end of slavery, and then, in 1794, the independence of Haiti – the same year the Jacobin Convention in Paris decreed the end of slavery in all the French colonies (Blackburn, 2011). Even though, later on, the rise of Napoleon in France led to the restoration of slavery in the French Empire and to repeated attempts to establish control on St. Domingue, the newly born Haitian Republic succeeded in maintaining its independence, and thus also its ban on slavery, and was officially recognized as a sovereign state in 1804. Effectively, Haiti's recognition as a nation marked not just the end of the colonial slave system in the Americas, but also, more importantly, the positive outcome of the only successful violent slave revolt in world history; in this sense, therefore, the Haitian Revolution is an exception that has given much food for thought to scholars of slavery (Genovese, 1979).

Peaceful versus Violent Abolitionism

(Abolitionism is the collective term for movements aiming at the abolition of either slavery or the slave trade.) Concerted efforts at opposing slavery and bringing about its abolition are a wholly modern feature in human history. In the past 50 years, scholars have determined through a number of studies that, in the later decades of the eighteenth century, for the first time in history, a combination of factors led to the beginning of attacks on the institution of slavery geared toward influencing the public opinion and with slavery's abolition as their ultimate objective. Among these factors, particularly important were the influence of Enlightenment doctrines of human rights, the rise of the classical political economic theory of capitalism, the increasing religious emphasis of Evangelical confessions on the sin of slavery and the leverage of powerful antislavery Christian groups such as the Quakers, and, above all, the emergence of influential white and black radical antislavery spokesmen and organizations. As a consequence of these momentous developments, nonviolent antislavery associations began to be increasingly vocal especially in the English-speaking world, eventually coalescing in England around a movement for the abolition of the British slave trade, headed by William Wilberforce and Thomas Clarkson. As a result of this crucial nonviolent antislavery activity, the Atlantic slave trade was abolished in the British Empire in 1807; significantly, the following year, also the United States abolished its own slave trade (Blackburn, 2011).

Beginning in the late 1820s, nonviolent antislavery forces in the English-speaking world began to pursue a radical agenda focusing on the achievement of immediate abolition of slavery. In Britain, the Anti-Slavery Society's Agency Committee, headed by Thomas Fowell Buxton, engaged in a massive effort in order to disseminate abolitionist literature and petitions to be signed and presented in Parliament. The result of this effort, together with the fear of constant slave rebellions, and changing economic logics, led the British Parliament to pass the 1833 Emancipation Act, which freed all the slaves in the British Empire. (Emancipation is the official abolition of slavery decreed by a government's act; Manumission is the master's voluntary act of freeing his/her own slaves.) In the following decade, partly stimulated by developments in Britain, the French Republican Government also abolished slavery peacefully throughout the French Empire by 1848. In the United States, similarly to what happened in Britain, a nonviolent abolitionist movement, in this case supported by both black and white activists, had begun its activities in the early 1830s, with William Lloyd Garrison's publication of *The Liberator* and with the 1833 foundation of the American Anti-Slavery Society. Unlike their British counterparts, though, American abolitionists were not successful in the short term, and their plans for persuading the majority of the public opinion to support immediate emancipation did not lead to any concrete result (Davis, 1984).

Even though the American abolitionist movement had always been challenged by some individuals who advocated more violent approaches to the problem of immediate emancipation of the slaves, it was the increasing confrontational nature of the issue of slavery's expansion in the western territories of the United States, in the 1840s that led to the rise in violent abolitionist actions against the southern slave system organized by militant abolitionists, who did not believe any longer in pressurizing the public opinion. Thus, during the 1850s, the decade that preceded the occurrence of the American Civil War, radical antislavery activists such as James Redpath and John Brown openly advocated a mass slave insurrection as the only viable means to bring about a violent end to slavery in the US South. Most famously, in 1859, John Brown embarked in his raid on Harper's Ferry with the intention of seizing an arsenal and provoking a massive slave revolt, and thus precisely with this objective in mind (Davis, 1984).

Peaceful versus Violent Endings of Modern Slavery

In the second half of the nineteenth century, the process leading to emancipation in the United States, which was a direct result of the American Civil War, stood in stark contrast with the end of slavery in both Cuba and Brazil, which occurred in a relatively peaceful manner, more similar to the processes that had led to the 1833 end of slavery in the British Empire and to the 1848 end of slavery in the French Empire. In the second half of the nineteenth century, Cuba was still a colony of the Spanish Empire. In the course of the 1860s, the rise of the first Spanish abolitionist society in 1864, and then the Spanish Government's abolition of the Atlantic slave trade in 1867, did a great deal to erode the Cuban planters' power. Subsequently, the creation of the first Spanish Republic in 1868, and the consequent rise of a movement for Cuban independence by the planters in the eastern part of the island during the Ten Years' War (1868–78) precipitated events. In 1870, in order to curb the planters' separatist effort, the Spanish Government issued the *Moret Law*, according to which both the children born of slave mothers since September 1868 and the slaves who were 60 years old or older were free. Then, the subsequent 1880 *Patronato Law* effectively decreed the end of slavery in Cuba starting from 1886 (Drescher, 2009).

In comparison with Cuba, in Brazil, British pressures on the Brazilian Government had led already to a major change in policy, with the abolition of Brazil's Atlantic slave trade in 1850. Then, Brazil's imperial government issued the 1871 *Rio Branco Law*, which freed the children born of slave mothers. Particularly crucial for further developments were, in the following decade, Joaquim Nabuco's 1880 foundation of the *Sociedade Brasileira Contra a Escravidão* (Brazilian Anti-Slavery Society), together with international pressure, and also the initiative of provinces such as Ceara, which abolished slavery in 1884. In 1885, the imperial government passed the *Dantas-Saraiva-Cotegipe Law*, which freed slaves who were 65 years of age, while increasingly larger numbers of runaway slaves fled plantations and farms, and finally, 3 years later, the imperial government decided to proclaim the end of slavery throughout the country with the *Golden Law* of 1888, which decreed slaves immediately and permanently free. Effectively, the 1888 *Golden Law* was one of only two official governmental acts – together with the 1863 US *Emancipation Proclamation* – that freed the slaves unconditionally and without compensation for slaveholders, unlike the case in all the other slave societies (Blackburn, 2011).

In the United States, slavery – particularly its expansion in the western territories – was, effectively, the main issue in national politics since the 1840s. In 1854, the antislavery forces formed the Republican Party specifically with the objective of preventing the expansion of slavery, and 6 years later Republican candidate Abraham Lincoln won the 1860 presidential election with the majority of the northern votes. Soon after, 11 slaveholding states seceded from the Union and formed the Confederacy, a nation wholly committed to the preservation of slavery, plunging the country into a 4-year-long Civil War (1861–65), which would cost more than 600 000 dead – effectively, the cost in human lives of the southerners' violent resistance to the US Government's enforcement of emancipation (Davis, 2006).

While, initially, Lincoln did not fight the war with the intention of freeing the slaves, a combination of factors – chief among them the protracted length of the war, the actions of the slaves, who fled in mass to Union camps, and the increasing importance of Radical Republicans and Abolitionists in Congress – led gradually to his conviction of the necessity of enacting emancipation. In September 1862, after a major Union victory at Antietam, Lincoln drafted the *Preliminary Emancipation Proclamation*, whose definitive version became the 1 January 1863 *Emancipation Proclamation*. Justified by Lincoln as a war measure, the 1863 *Proclamation* declared immediately and forever freed all the slaves in areas under Confederate control, gave legal backing to the slaves' efforts to free themselves, and gave the Union Army the legal power to help them with military action in their efforts of self-liberation. Soon after, African-American soldiers enlisted in the Union Army, and their number rose to reach almost 200 000 units at war's end. Then, in 1865, Congress passed the *Thirteenth Amendment* to the Constitution, officially abolishing slavery throughout the United States. Thus, unlike emancipation in Cuba and Brazil, and unlike the majority of the processes of emancipation in slave societies in world history – with the notable exception of Haiti – US Slave Emancipation was possible exclusively as a direct consequence of the victory of the Union in a major war fought by antislavery forces against a particularly powerful class of slaveholders, who violently resisted slavery's abolition, with a consequent large loss in human lives (Davis, 2006).

The Resurgence of Violent Slavery in the Twentieth Century

In the twentieth century, violent practices of exploitation employing slave labor reared their head again, largely as a direct consequence of the rise of totalitarian regimes, which plagued the recent history of a number of different countries from Europe to Asia. In all these regimes, a crucial common feature was the creation of systems of labor camps, where captives, prisoners of war, opponents, and dissenters were forcibly enclosed and forced to submit to exhausting labor schedules under strict supervision and the threat of violent punishment – essentially in conditions that replicated, albeit in modern settings and within the context of production efforts often related to the modern needs of technological warfare, the common life of enslaved laborers in the large slave systems of the Atlantic before the abolition of slavery. It is important, therefore, to end this succinct analysis of the historical relationship between slavery and violence by briefly looking at the best-known example of resurgence of violent slavery and employment of slave labor in twentieth-century Europe: the Nazi concentration camps (Davis, 1984).

Concentration camps with the employment of mass forced labor had been present in both the German Empire and the Austro-Hungarian Empire during World War I, and they were set up in Germany already from the very beginning of Hitler's rise to power in 1933. However, with the large-scale expansion of the Third Reich during World War II, the mass employment of forced labor by the Nazi Government became a necessity, as between 1939 and 1945, 13 million Germans were employed in Hitler's army and needed to be replaced and also supported

in the war effort by at least an equal number of forced laborers. Thus, as Nazi power expanded, especially in eastern Europe, millions of civilians and prisoners of war were enslaved and sent to toil in forced labor, either in concentration camps, under the supervision of SS units, or in agricultural fields, or in armament factories. At the heart of this massive effort were the two ideological concepts of *Lebensraum* (living space), which was a large area at the core of the projected Reich, to be conquered by the army and settled by Germans, and of the *Untermenschen* (inferior races), which were the racially defective people that pure Aryans, i.e., Germans, would enslave and force to labor in the conquered territories until they ceased to be useful and would be disposed of in order for them to not threaten German racial purity (Drescher, 2009).

Yet, while the violence of enslavement and the exploitative conditions of forced labor, with the exhausting pace and the threat of discipline and punishment, often resulting in death, were features that make Nazi concentration camps comparable to slave plantations in the New World, the actual value placed upon slave labor was very different in the two situations. For one thing, Nazi slaves were not bought or sold on slave markets, and therefore they were not in themselves a source of capital as slaves in New World plantations. More importantly, Nazi obsession with racial purity effectively prevailed over the pressing economic needs for enslaved labor. In the end, with the Holocaust, as Seymour Drescher has remarked, Nazi Germany, "a state with a greater appetite for forced labour than all previous colonial systems in the Americas combined, destroyed far more potential workers in four years than were lost in the Atlantic transit of Africans in the course of four centuries" (Drescher, 2009: 438).

Conclusion

While violence in general has characterized the history of slavery from its inception, specific modes of violence have characterized the different stages of the experience of slavery for the enslaved according to different times and places. The process of enslavement itself resulted from the violence of uprootedness associated with the practices of warfare conducted for the purpose of conquest or else kidnapping, and then the equally traumatic processes of sale and purchase in the slave markets. Equally violent and traumatic were the actual experiences of laboring under slavery, with its inevitable accompaniment of strict discipline and punishment, especially in those cases in which slave agricultural production was geared toward the maximization of the slaveholders' profit. Violence also characterized a great deal of the slaves' response to their own violent oppression and exploitation, through the staging of slave rebellions of different sizes and with different scopes throughout history. Ultimately, though, in the large majority of the cases, aside from the notable exceptions of the Haitian Revolution and the American Civil War – and leaving aside the twentieth-century slave regimes – the end of slavery came relatively peacefully and as a result of governmental decrees that eased the nonviolent transition from slave to free labor, even though almost always not bringing an end to the ex-slaves' economic exploitation.

See also: Civil Wars; Colonialism, Anthropology of; Holocaust, The; Revolutions, History of; Slavery as Social Institution; Slavery: Comparative Aspects; Slaves and Slavery, History of; Violence, History of; Warfare in History.

Bibliography

Blackburn, R., 1997. The Making of New World Slavery: From the Baroque to the Modern. Verso, London.
Blackburn, R., 2011. The American Crucible; Slavery, Emancipation, and Human Rights. Verso, London.
Bradley, K., 1989. Slavery and Rebellion in the Roman World, 140 BC–70 BC. Indiana University Press, Bloomington, IN.
Burnard, T., Heuman, G., 2011. Introduction. In: Heuman, G., Burnard, T. (Eds.), The Routledge History of Slavery. Routledge, London, pp. 1–15.
Dal Lago, E., 2012. American Slavery, Atlantic Slavery and beyond: The U.S. "Peculiar Institution" in International Perspective. Paradigm Publishers, Boulder, CO.
Davis, D.B., 1984. Slavery and Human Progress. Oxford University Press, New York.
Davis, D.B., 2006. Inhuman Bondage: The Rise and Fall of Slavery in the New World. Oxford University Press, New York.
Drescher, S., 2009. Abolition: A History of Slavery and Antislavery. Cambridge University Press, New York.
Genovese, E., 1979. From Rebellion to Revolution: Afro-American Slave Revolts in the Making of the Modern World. Louisiana State University Press, Baton Rouge, LA.
Heuman, G., 2011. Slave rebellions. In: Heuman, G., Burnard, T. (Eds.), The Routledge History of Slavery. Routledge, London, pp. 220–233.
McKeown, N., 2011. Greek and Roman slavery. In: Heuman, G., Burnard, T. (Eds.), The Routledge History of Slavery. Routledge, London, pp. 19–34.
Miller, J.C., 2011. The Problem of Slavery as History: A Global Approach. Yale University Press, New Haven, CT.
Patterson, O., 1982. Slavery and Social Death: A Comparative Study. Harvard University Press, Cambridge, MA.
Phillips, W.D., 1985. Slavery from Roman Times to the Early Transatlantic Trade. Manchester University Press, Manchester.
Rediker, M., 2008. The Slave Ship: A Human History. Penguin, London.
Tomich, D., 2003. Through the Prism of Slavery: Labor, Capital, and World Economy. Rowman and Littlefield, Lanham, MD.

Slavery as Social Institution

Kevin Bales, Wilberforce Institute for the Study of Slavery and Emancipation, University of Hull, Hull, UK
Monti N Datta, University of Richmond, Richmond, VA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by O. Patterson, volume 21, pp. 14146–14152, © 2001, Elsevier Ltd.

Abstract

Slavery has been a permanent part of the human condition throughout history, though not present in all cultures or societies all the time. This article examines contemporary forms of slavery, recognizing that an estimated 29.8 million persons are enslaved in the world today. This article examines four aspects of contemporary slavery: its definition; estimation of its prevalence; risk factors that support or predict enslavement; and policy measures taken to mitigate its spread.

Slavery has been a permanent part of the human condition throughout history, though not present in all cultures or societies all the time. Though sometimes referred to as an 'institution,' it is not one in the precise meaning of that term, which is used to describe those key organized activities that are and have been present, in one form or another, in all societies and cultures, such as economic exchange, processes of socialization, religious expression, or some form of family organization. Put simply, slavery is common throughout history and in evidence in most countries today, but it is not a universal attribute of societies. Put another way, it cannot be assumed to be an innate characteristic or attribute of human beings.

This article examines contemporary forms of slavery, recognizing that, according to The Global Slavery Index (2013), an estimated 29.8 million persons are enslaved in the world today. This article examines four aspects of contemporary slavery: its definition; estimation of its prevalence; risk factors that support or predict enslavement; and policy measures taken to mitigate its spread.

Defining Slavery

Slavery has existed since the dawn of civilization. Aristotle (2013[350 BCE]) wrote: "The slave is a part of the master in the sense of being a living but separate part of his body." Aristotle's comment illustrates how, at its most basic level, slavery is a relationship between two individuals – the enslaved and the slaveholder, characterized by an asymmetrical power differential in which the slaveholder has complete control over the enslaved. The nature of this control ultimately depends on social, psychological, economic, geopolitical, and cultural factors. The method of this control most often includes an interaction of physical, emotional, and sexual violence. For nearly all enslaved females (and some males), this control means being raped by the slaveholder. This often generates trauma and posttraumatic stress within the enslaved. Indeed, the least studied but crucial attributes of slavery are its physical and psychological sequelae.

Although slavery has existed for thousands of years, contemporary slavery is different from the past in at least two ways. First, the cost of a slave today is very low. In Moldova, for instance, through an illicit transaction, a slave can be purchased today for about $200 USD, and in Haiti for about $50 USD (Skinner, 2009). Two hundred years ago, by contrast, in the American South, the cost of a slave was about $1000 USD, which, adjusted for inflation, is about $45 000 USD today (Bales, 2012). Compared to the past, when slaves tended to be capital investments, their acquisition cost is now so low that they might be thought of as disposable inputs to economic enterprises. For example, it can be more cost-effective for slaveholders to simply replace ill or injured slaves rather than pay for their treatment.

A second way in which contemporary slavery is different from the past is that slavery today is universally illegal. This contrasts with most of human history in which slavery was a legally sanctioned, socially acceptable practice. For example, slavery is mentioned throughout the Old and New Testaments and is simply accepted everywhere in the Bible as a customary social arrangement. The result is that compared to its historically legal form, slavery today has become a hidden criminal activity, supported by governmental corruption and global competition to profit from the lowest marginal costs.

Arriving at an operational, legal, or conceptual definition of contemporary slavery is a challenge, given the substantial disagreement among scholars, policy makers, international organizations, activists, and other stakeholders. Today there are many definitions of slavery; one of the most inclusive is the 2012 Bellagio-Harvard Guidelines on the Legal Parameters of Slavery. These guidelines build upon the 1926 Slavery Convention of the League of Nations, which defines slavery as "the status or condition of a person over whom any or all of the powers attaining to the right of ownership are exercised." The Bellagio-Harvard Guidelines specify that the exercise of "the powers of the right of ownership" "should be understood as constituting control over a person in such a way as to significantly deprive that person of his or her individual liberty. Normally this exercise is supported by and obtained through means such as violent force, deception and/or coercion, with the intent of exploitation through the use, management, profit, transfer or disposal of that person." These guidelines conceptualize slavery as an umbrella term under which the various types of enslavement can be understood.

The practice of slavery has always been based upon rationalizations and justifications on the part of slaveholders. In the past, racist ideas of superiority/inferiority provided one justification for enslaving certain populations. Race, however, is just one of many rationalizations utilized to support slavery

in the past and today, others include religion, ethnicity, and gender. Indeed, any form of social discrimination, at its most extreme, might be taken to two ultimate outcomes: enslavement and/or murder (including genocide).

Perhaps the most widespread form of slavery today is debt bondage (also referred to as bonded labor), in which the rationale for enslavement is an obligation linked to a debt. While some legal instruments and commentators falsely assume that the slavery exists because a debt can never be repaid due to the low wage being earned and the high interest rate being charged, this is incorrect. The foundation of slavery is the violent physical and psychological control exerted by the slaveholder, and once the slave has been obtained through the justifying mechanism of debt (or race or religion or some other mechanism), it is the violent force, fraud, and coercion that maintains the enslavement. Forms of debt bondage have existed for thousands of years and continue to this day. It is substantial in South Asia, particularly within India's long-standing caste system, in which the lowest caste (Dalits) and other 'Scheduled Tribes' are enslaved in agriculture, brick kilns, mills, and factories, sometimes for generations. The plight of those caught in this hereditary form of collateral debt bondage slavery reminds us that forms of slavery cannot always be easily separated into 'new' and 'old' forms. Families in South Asia, sometimes entire villages, are found to be enslaved for many generations, and living without knowledge of their rights or the possibility of freedom.

Another common type of slavery is contract slavery, in which a slaveholder procures apparently legal documentation for a person to work, but once employed, that person is enslaved. False contracts mislead unsuspecting victims to believe that their work is legitimate (Bales, 2012). Examples of contract slavery include: South Asian women offered jobs as waitresses in Japan and offered the requisite legal travel documentation, but who, upon arrival in Japan, are forced into prostitution (Bales, 2008); Mexican immigrants in the United States offered jobs as day laborers picking tomatoes in Florida, but who, upon arrival at a tomato farm, are not allowed to leave of their own free will; and Nepalese migrants offered jobs as construction workers in Qatar, but who then are forced to work without pay and left begging for food. It should be noted that contract slavery is often linked to debt bondage slavery when a worker borrows money from the slaveholder or other brokers. A common scenario is borrowing to cover transportation costs to go from the home country to a destination country, and finding that debt dramatically increased and used as a rationale for enslavement.

Perhaps the most underreported and least studied form of contemporary slavery is forced marriage, in which either one or both parties have not agreed to the act and violence or threats are used to force a person into the 'marriage.' In most instances, the victim is female. Politically, a challenge arises in responding to forced marriage, given that some argue that this practice is sanctioned by cultural norms and parallel legal systems such as tribal, customary, or religious laws, and therefore outside the purview of existing legal prosecution and adjudication.

Another less publicized form of slavery, as defined by one UN Convention as human trafficking, is organ trafficking. This falls under three broad categories: first, when a trafficker forces a victim to give up an organ; second, when a victim contractually agrees to sell an organ but is not paid; and third, when a victim is being treated for an ailment and, unbeknownst to them, an organ is removed (United Nations, 2014). The illegal trade of kidneys, for instance, has risen to the extent that the World Health Organization estimates one kidney is sold every hour on the illicit market (Campbell and Davison, 2012).

Despite the variety in the types of contemporary slavery, the most publicized form in developed nations is sex trafficking, due to its prominence in popular culture, policy circles, and the media. One prominent definition of sex trafficking in the United States comes from its federal Trafficking Victims Protection Act, which defines it as "the recruitment, harboring, transportation, provision, or obtaining of a person for the purpose of a commercial sex act where such an act is induced by force, fraud, or coercion, or in which the person induced to perform such act has not attained 18 years of age" (U.S. Congress, 2000). It should be noted that, on the demand side of sex trafficking, purchasers of commercial sex come from all social strata, including the educated, uneducated, wealthy, poor, rich, and famous. To this extent, mainstream society economically supports the global sex trade. This has generated a conversation particularly within the United States faith-based community about abolishing prostitution as a viable way to end sex trafficking.

Forced commercial sexual exploitation is only one way that slavery interacts with mainstream society. Scholars have documented many instances of the enslaved used in the production of commodities that flow into the global supply chain (Bales and Soodalter, 2009). Originating from numerous countries, slave labor is found in the production of textiles, gemstones, apparel, and agricultural products like cocoa, coffee, and sugar. In the Congo, for instance, armed gangs force the enslaved to dig for minerals such as coltan, cassiterite, wolframite, and tungsten used in the production of cell phones, laptops, and other electronic devices. Slavery is therefore a significant, albeit small, factor in global economic output. At the same time, because many Fortune 500 companies invest in or purchase from subsidiary companies that employ a modicum of slave labor in their product supply chain, it can be assumed that a number of mutual funds also have a small percentage of slave labor in their portfolios. At the aggregate, one estimate suggests that about $44 billion USD per year is made in slavery-derived profits (Belser, 2005).

Estimating the Prevalence of Slavery

Estimating the number of enslaved in the world today is a challenge given its hidden, illicit nature. Slaveholders naturally restrict freedom of movement of the enslaved, such that it is difficult for the enslaved to signal their location and situation to a third party. Moreover, with the possible exception of sex trafficking, many forms of contemporary slavery do not leave a lasting impression on the public eye. This includes domestic workers who appear willingly employed but are held under threat of violence and agricultural workers trapped on farms away from public view. It also includes low-skilled workers in remote villages in developing countries with little access to help.

Scholars are just beginning to apply social science research methods to account for the hidden nature of slavery. One method, although not widely available, is to obtain representative sample survey data of the general public, with the aim of eliciting the true level of slavery across the population. Some antislavery organizations and scholars are pioneering this technique, but the challenge is to elicit unbiased responses from respondents who are asked to comment on a crime (slavery), about which they may feel uncomfortable speaking candidly. Moreover, the implementation of such representative sample surveys on a large scale can be financially prohibitive, thus delaying the commission of such surveys on a widespread systematic basis. Another, more widely applied method has been for separate survey research teams to go into the field and collect separate counts of the enslaved in regions where there is a high probability of enslavement. From these lists, researchers then calculate an estimate of the population proportion enslaved. Yet another method is to mine the relatively large literature on primary and secondary source materials from governments, nongovernmental organizations, and journalistic accounts that have chronicled contemporary slavery. Based on these primary and secondary sources, scholars can also estimate the proportion of the population enslaved, although with a high degree of uncertainty. Still other approaches employ several of these techniques at the same time, to varying degrees of accuracy and success.

A problem common to all of these methodologies, however, is that, because operational definitions of slavery differ across scholars, governments, activists, and other stakeholders, some methods may deliberately or inadvertently exclude some major types of slavery. Forced marriage, for example, is often not counted in most studies of contemporary slavery's prevalence, although some scholars estimate the number of people in forced marriages may be in the millions. This suggests that most estimates of modern day slavery are likely underestimates. As to the extent to which these underestimates differ from the true proportion of the population enslaved remains unknown. Further complicating this methodological challenge is the norm that, at least until very recently, most survey researchers based in NGOs and IGOs have not shared their methods or results in a transparent fashion from which to build a common database.

Several oft-quoted estimates of slavery's global prevalence do exist however; the most common estimate is 27 million slaves. The author of that study, however, relying upon more precise representative sample survey data, recently updated that sum to 29.8 million (Bales et al., 2013). Other studies suggest the prevalence of slavery is closer to 21 million (ILO, 2014). Most likely, the divergence in these counts (21 million vs 29.8 million slaves) is best explained by differences in the operational definition of slavery and the methodology used to estimate the prevalence of the enslaved.

Like the estimation of slavery at the global level of analysis, differences also exist in estimating slavery at the level of the nation-state. The United States is illustrative. One study estimated that at any given time there are 10 000 or more people enslaved in the United States (Free the Slaves, 2005). Another study suggested there are upward of 60 000 people enslaved – an increase by a factor of six (Bales et al., 2013). Such widespread deviation among these studies leaves little room for policy makers to have a firm grasp of what the real prevalence of contemporary slavery might be. Similar variation in the estimated enslaved exists for other nations, such as the United Kingdom. In the meantime, some studies of prevalence are being conducted in West Africa, but are only at the early stages. Most likely, there will also be significant variation in what these estimates yield as well.

A substantial lacuna therefore exists on ways to measure reliably the prevalence of contemporary slavery, at both the level of the nation-state and the global level of analysis. This concern has been underscored by the US National Institute of Justice (NIJ), which conducted a comprehensive review of the antitrafficking literature in 2008 and found, "quantitative methodologies are noticeably scarce" and that, "the dominant anti-trafficking discourse is not evidence based" (Goździak and Bump, 2008). Moreover, the NIJ review found that only 16% of the 218 academic articles on contemporary slavery were both peer reviewed and based on empirical research. This means that the majority of scholarship on contemporary slavery has not been subjected to routine standards of social scientific inquiry, such as evaluating a given hypothesis in light of one or more competing explanations.

A more methodologically robust inquiry into the prevalence of contemporary slavery is needed to shed more light on this hidden crime. Although representative random sampling would be ideal for additional studies, particularly at the global level, its prohibitive cost makes this approach challenging, if not untenable, especially for scholars in the developing world who do not have financial access to large survey polling organizations. A potentially innovative approach, methodological approach, is multiple systems estimation (MSE). The MSE methodology originated in the biological sciences, and has been applied in recent years to problems in epidemiology and estimating civilian casualties in war zones. As Manrique-Vallier et al. explain, "MSE methods attempt to estimate the number of cases that were not included in lists that partially enumerate a closed population." Through generating different lists of potential victims of human trafficking (gathered from social services, law enforcement, immigration officials, etc.), scholars can use MSE statistical modeling techniques to identify the hidden population of the enslaved.

Another way to generate better prevalence estimates of contemporary slavery involves more substantial cooperation among key stakeholders. As Goździak and Bump (2008) recommend: "There is a need to establish a forum where research results can be exchanged between different scholars as well as shared with policy makers and service providers; such a forum can take a form of a specialized publication or an international task force." This is a matter of political will and organizational capacity.

The Risk Factors That Help Explain Why People Are Enslaved Today

In this section, we explore how the amount of modern slavery in a country is related to other important factors such as corruption and human development. It is important to uncover these relationships since they can help determine effective efforts for ending modern slavery. Note that global trends and

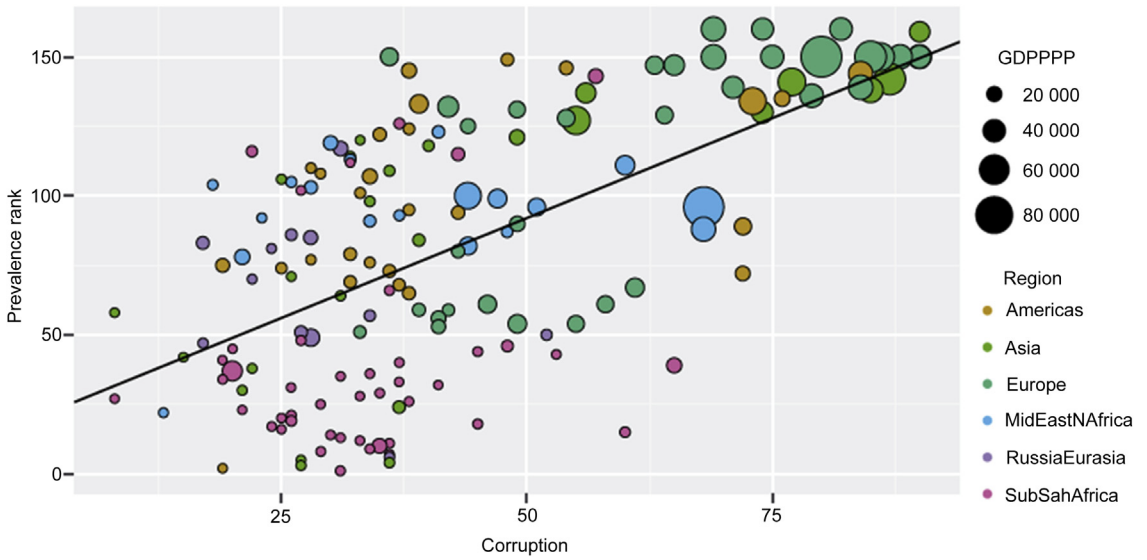

Figure 1 A correlation plot between slavery prevalence and corruption in 2013. Source: Global Slavery Index: http://www.globalslaveryindex.org.

relationships are generalized findings, and do not represent the unique situation of any one country. What is important is that these measurements of co-relationships point to patterns of risk, vulnerability, and interaction in the dynamic crime of modern slavery. Risk itself is measured along five dimensions: the antislavery policies (or lack thereof) of the country where a person lives; the level of human rights afforded to a person in the country in which they live; the level of economic and social development of a country; the stability of the state and law enforcement of the country in which a person lives; and the levels of discrimination a person faces (Bales et al., 2013). Different components of these dimensions can be compared statistically with prevalence levels of slavery across countries.

These statistical correlations tell us two things: first, whether there really is a relationship between the occurrence of modern slavery and another factor; and second, the strength of that relationship. The research literature notes ample examples of corruption being implicated in modern slavery. This is no surprise; if the rule of law is effective, it protects citizens from being enslaved. If the rule of law is corrupted, people are not protected and criminals can pay off officials and act with impunity.

Figure 1 examines the relationship between corruption (as measured by Transparency International) and the prevalence of modern slavery (as measured by the Global Slavery Index) for 162 countries. Note that the data from the Global Slavery Index correspond to the *y*-axis in Figure 1, which ranges from a score of 1 (very low prevalence) to a score of 162 (very high prevalence). Note that the data from Transparency International correspond to the *x*-axis in Figure 1, which ranges from a score of 0 (high corruption) to a score of 100 (very low corruption). Each dot in Figure 1 represents a country. This graph illustrates a strong relationship between the prevalence of modern slavery and corruption ($R^2 = 0.61$), statistically significant at the 0.01 level.

The research literature also frequently draws connections between levels of human development, particularly poverty (including factors such as availability of health care, education, and levels of income) and modern slavery. Figure 2 examines the relationship between human development (as measured by the United Nations Development Programme Human Development Index (HDI) along the *x*-axis) and prevalence of modern slavery (as measured by the Global Slavery Index along the *y*-axis) for 162 countries. Note the upward sloping shape of the graph, which suggests higher levels of modern slavery correlate with lower levels of the HDI (and conversely, lower levels of modern slavery correspond to higher levels of HDI).

The level of a country's human development reflects its economic well-being, which in turn affects the poverty levels and deprivation a citizen might face. On the other hand, higher levels of educational attainment, elevated health care, and the chance of effective citizenship, all tend to allow people to protect themselves from vulnerability to modern slavery. This finding is statistically significant ($p > 0.01$) and the correlation score ($R^2 = 0.74$) points to a strong relationship between modern slavery and development.

A common pathway to slavery is through debt obligations. When access to legal sources of credit is not available, the poor and vulnerable will sometimes turn to other types of lending, even to the mortgaging of their own lives. Access to financial services can be crucial for upward social mobility, which brings the economic stability that helps to protect against enslavement. Figure 3 shows the relationship between access to financial services and the prevalence of slavery for 161 countries. This finding is statistically significant ($p > 0.01$) and the correlation score ($R^2 = 0.63$) suggests a strong relationship.

Current Policy Measures Taken to Mitigate the Spread of Slavery

There is a paradox concerning the policy and legal measures in place to end or mitigate the spread of slavery. Slavery is illegal in every country, it is forbidden by a number of international instruments, and it is one of a handful of human activities that

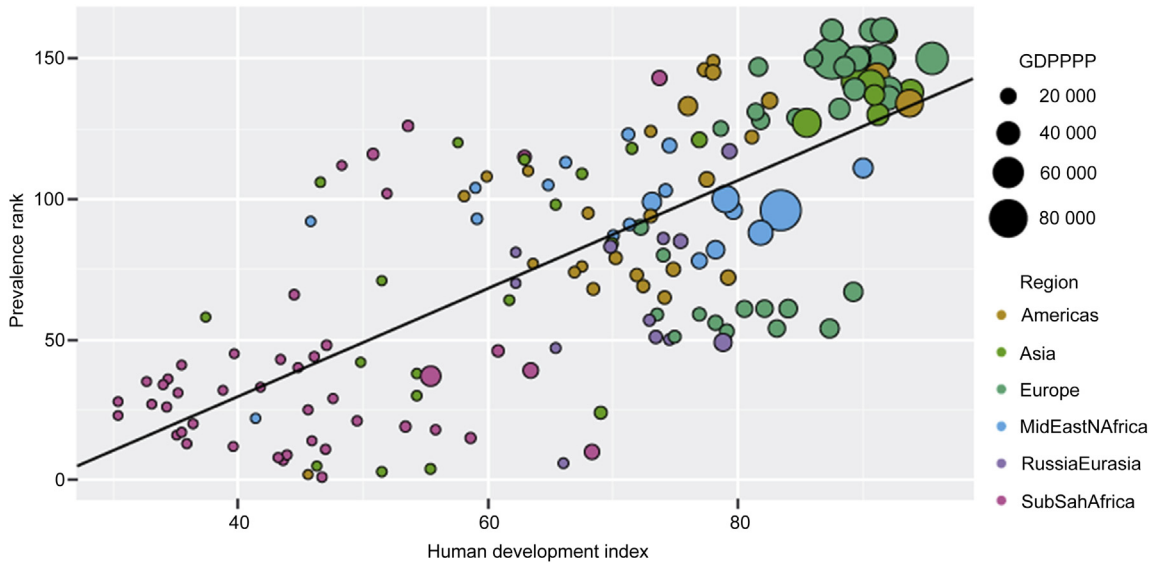

Figure 2 A correlation plot between slavery prevalence and the Human Development Index in 2013. Source: Global Slavery Index: http://www.globalslaveryindex.org.

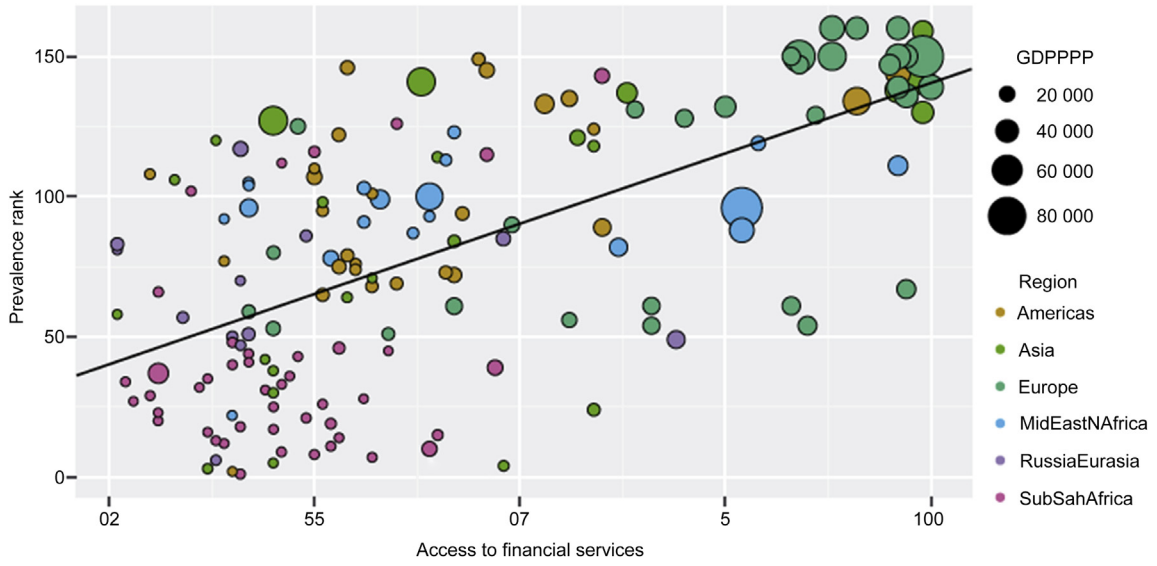

Figure 3 A correlation plot between slavery prevalence and access to financial services in 2013. Source: Global Slavery Index: http://www.globalslaveryindex.org.

are of such severity and the popular consensus of approbation is so great that it is considered a *jus cogens* violation. This type of violation is subject to a fundamental and overriding principle of international law from which no derogation is permitted, also referred to as a 'peremptory norm.' This is a principle accepted by the international community of states that overrides or surmounts other laws. The prohibitions of genocide, maritime piracy, slavery and the slave trade, torture, and wars of aggression and territorial aggrandizement, are generally accepted as *jus cogens* norms. In other words, there is an absolute and global prohibition of slavery and all its related activities.

At the same time, and in the face of this absolute and global prohibition, slavery exists in every country, and policies, laws, and especially the resources needed to address this crime are poor or lacking. Two examples drawn from the United States, one of the most active of national governments in their anti-slavery policy, illustrates this mismatch between the size of the problem and the size of the response. According to one State Department estimate, there are approximately 17 000 new slaves in the United States each year – comparable to the average of around 17 000 homicides in the United States each year. Across all US law enforcement provision (local, state, and federal police/agents), there are about 45 000 trained homicide

specialists, but only about 200 trained antislavery specialists. This disparity in trained personnel results in national homicide clear-up rates of 70–75% and slavery/trafficking clear-up rates of less than 1%. This fundamental pattern is repeated globally across all countries and governments.

Expansion and training of enforcement personnel would be critical to success to mitigation and eradication. Yet, in terms of financial resources, in 2012, US federal expenditure on reducing the criminal activity of slavery and human trafficking is less than that devoted to military bands within the Department of Defense. For comparison, the estimated cost of global eradication of slavery, over a 25-year period, is $11 billion, which is one-half the yearly cost of the US government's War on Drugs. It is important to reiterate that the United States is one of the world's leading countries in addressing slavery and trafficking in policy, law, and resources.

It can be argued that Brazil is a world leader in terms of antislavery policies. There are three distinct provisions that Brazil has enacted that have been shown to be effective. First, in 2002, the government of Brazil initiated a national slavery eradication plan. This brought together all relevant agencies with law enforcement and appropriate nongovernmental organizations as well as unifying and resourcing the legal response. First-year (2003) results were 4789 slaves freed. Second, Brazilian experience also showed the value of specialist federal antislavery squads with attached labor inspection and enforcement personnel. Many, if not most, countries have specially trained antidrug units, but specialist antislavery units, though effective, are rare. Finally, Brazil also established a 'dirty list' of any people or companies that were found to use slave labor. The list is published on the Internet and regularly updated. Being on the list excludes companies from receiving any sort of government funds, grants, or credits. Since much of the process of opening and developing land relies on government tax credits or support, this effectively drives slave-using companies and individuals out of that business. Consumers can determine which companies they would wish to avoid in order to avoid slave-based products.

See also: Discrimination: Racial; International Migration; Minorities; Race Identity; Racial Relations; Slavery and Violence; Slavery: Comparative Aspects; Slaves and Slavery, History of.

Bibliography

Allain, J., et al., 2012. Bellagio-Harvard Guidelines on the Legal Parameters of Slavery. Available from: http://www.qub.ac.uk/schools/SchoolofLaw/Research/HumanRightsCentre/Resources/Bellagio-HarvardGuidelinesontheLegalParametersofSlavery.

Aristotle, 2013[350 BCE]. Politics. University of Chicago Press.

Bales, K., 2008. Ending Slavery: How We Free Today's Slaves. University of California Press.

Bales, K., 2012. Disposable People: New Slavery in the Global Economy, third ed. University of California Press.

Bales, K., Soodalter, R., 2009. The Slave Next Door: Human Trafficking and Slavery in America Today. University of California Press.

Bales, K., et al., 2013. The Global Slavery Index. Available from: http://www.globalslaveryindex.org.

Belser, P., 2005. Forced Labour and Human Trafficking: Estimating the Profits. International Labor Organization, Geneva. Available from: http://www.ilo.org/wcmsp5/groups/public/–ed_norm/–declaration/documents/publication/wcms_081971.pdf.

Campbell, D., Davison, N., 2012. Illegal Kidney Trade Booms as New Organ Is 'Sold Every Hour'. The Guardian. Available from: http://www.theguardian.com/world/2012/may/27/kidney-trade-illegal-operations-who.

Free the Slaves, 2005. Hidden slaves: forced labor in the United States. Berkeley Journal of International Law 47 23. Available from: http://scholarship.law.berkeley.edu/bjil/vol23/iss1/2.

Goździak, E.M., Bump, M.N., 2008. Data and Research on Human Trafficking: Bibliography of Research-based Literature. Institute for the Study of International Migration, Walsh School of Foreign Service, Georgetown University. Available from: https://www.ncjrs.gov/pdffiles1/nij/grants/224392.pdf.

ILO, 2014. Forced Labor. Available from: http://www.ilo.org/global/topics/forced-labour/lang–en/index.htm.

Manrique-Vallier, M., et al., 2003. Multiple systems estimation techniques for estimating casualties in armed conflicts. In: Seybolt, Taylor B., Aronson, Jay D., Fischhoff, Baruch (Eds.), Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict. Oxford University Press pp. 147–164.

Skinner, E.B., 2009. A Crime So Monstrous: Face-to-Face with Modern Day Slavery. Free Press.

United Nations, 1926. Slavery Convention. Available from: http://www.ohchr.org/EN/ProfessionalInterest/Pages/SlaveryConvention.aspx.

United Nations, 2014. Trafficking for Organ Trade. Available from: http://www.ungift.org/knowledgehub/en/about/trafficking-for-organ-trade.html.

United States Congress, 2000. Trafficking in Victims Act. Available from: http://www.gpo.gov/fdsys/pkg/PLAW-106publ386/html/PLAW-106publ386.htm.

Slavery: Comparative Aspects

Orlando Patterson, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14152–14157, © 2001, Elsevier Ltd.

Abstract

In the comparative study of slavery it is important to distinguish between slave holding societies and large-scale, or what Moses Finley called 'genuine,' slave societies. The former refers to any society in which slavery exists as a recognized institution, regardless of its structural significance. Genuine slave societies belong to that subset of slave holding societies in which important groups and social processes become heavily dependent on the institution.

The institution of slavery goes back to the dawn of human history. It remained important down to the late nineteenth century, and persisted as a significant mode of labor exploitation in some Islamic lands as late as the second half of the twentieth century. Remnants of it are still to be found in the twenty-first century in a few areas. Yet it was only in a minority of cases that it metastasized socially into genuine slave societies, though in far more than the five cases erroneously claimed by Keith Hopkins. Thus, the institution existed throughout the ancient Mediterranean, but only in Greece and Rome (and, possibly, Carthage) did genuine slave societies emerge. It was found in every advanced precapitalist society of Asia, but only in Korea did it develop into large-scale slavery. All Islamic societies had slavery, but in only a few did there emerge structural dependence on the institution. In all, there were approximately 50 cases of large-scale slavery in the precapitalist world. With the rise of the modern world, slavery became the basis of a brutally efficient variant of capitalism. There were at least 40 such cases, counting the Caribbean slave colonies as separate units.

This article examines two sets of problems. The first concerns those factors associated with the presence of institutionalized slaveholding. The critical question here is: why is the institution present in some societies yet not in other, apparently similar, ones? The second set of problems begins with the assumption that slavery exists, and attempts to account for the growth in significance of slavery. More specifically, such studies attempt to explain the origins, structure, dynamics, and consequences of genuine slave societies.

In the comparative study of slavery it is important to distinguish between slave holding societies and large scale, or what Moses Finley called 'genuine,' slave societies (Finley, 1968). The former refers to any society in which slavery exists as a recognized institution, regardless of its structural significance. Genuine slave societies belong to that subset of slave holding societies in which important groups and social processes become heavily dependent on the institution.

The institution of slavery goes back to the dawn of human history. It remained important down to the late nineteenth century, and persisted as a significant mode of labor exploitation in some Islamic lands as late as the second half of the twentieth century. Remnants of it are still to be found in the twenty-first century in a few areas. Yet it was only in a minority of cases that it metastasized socially into genuine slave societies, though in far more than the five cases erroneously claimed by Keith Hopkins (1978). Thus, the institution existed throughout the ancient Mediterranean, but only in Greece and Rome (and, possibly, Carthage) did genuine slave societies emerge. It was found in every advanced precapitalist society of Asia, but only in Korea did it develop into large-scale slavery. All Islamic societies had slavery, but in only a few did there emerge structural dependence on the institution. In all, there were approximately 50 cases of large-scale slavery in the precapitalist world. With the rise of the modern world, slavery became the basis of a brutally efficient variant of capitalism. There were at least 40 such cases, counting the Caribbean slave colonies as separate units (for the most complete list, see Patterson, 1982: App. C)

This article examines two sets of problems. The first concerns those factors associated with the presence of institutionalized slaveholding. The critical question here is: why is the institution present in some societies yet not in other, apparently similar, ones? The second set of problems begins with the assumption that slavery exists, and attempts to account for the growth in significance of slavery. More specifically, such studies attempt to explain the origins, structure, dynamics, and consequences of genuine slave societies.

Comparative Approaches to Slaveholding Societies

There is now a vast and growing body of literature on slavery (Patterson, 1977a; Miller, 1993; Miller and Holloran, 2000). Yet, with the notable exception of certain anthropological studies (to be considered shortly) relatively few recent works are truly comparative in that they aim to arrive at general conclusions about the incidence, nature, and dynamics of slavery. Even those few works that compare two slaveholding societies tend to be concerned more with highlighting, through contrast, the distinctive features of the societies under consideration. This highly particularistic trend marks a regrettable departure from earlier studies of slavery.

The evolutionists of the nineteenth century were the first to offer explanations for the presence of slaveholding. Their basic proposition – that there was a close relationship between stages of socioeconomic development and the rise of slavery – has received no support from modern comparative studies (Pryor, 1977). However, some of their less grandiose views have survived later scrutiny. One was their emphasis on warfare, and the demand for labor at certain crucial points in

their scales of development as the important factors (Biot, 1840; Westermarck, 1908). The other was their finding that the socioeconomic role of women was critical. It was claimed, for example, that the subjection of women provided both a social and an economic model for the enslavement of men (Biot, 1840; Tourmagne, 1880).

The most important of the early-twentieth-century theorists was H. J. Nieboer (1910), who broke with the evolutionists with his open-resource theory. His work, which is still influential, was unusual also in its reliance on statistical data. His main hypothesis was that slavery existed to a significant degree only where land or some other crucial resource existed in abundance relative to labor. In such situations, free persons cannot be induced to work for others, so they must be forced to do so. The theory has had lasting appeal, especially for economic historians, since it is both testable and consistent with marginal utility theory. It was revived by Baks (1966) and by the MIT economist, Domar (1970). However, the theory has been shown to have little empirical support from modern cross cultural data (Patterson, 1977b; Pryor, 1977); and Engerman (1973) has questioned seriously its theoretical consistency.

In the course of criticizing Nieboer and Siegel (1945) proffered a functionalist theory which claimed that chattel slavery may be expected to occur in those societies where there is a tendency to reinforce autocratic rule by means of wealth symbols, and where the process results in a rather strongly demarcated class structure. There is little empirical support for this theory, and it verges on circularity. In many pre-modern hierarchical societies with wealth symbols, it was slavery which was the main cause of increased stratification.

The best recent comparative work on slavery has come from historical anthropologists studying Africa. Meillassoux (1975, 1991) and his associates have demonstrated ably from their studies of West Africa and the Sahel just how valuable a nondogmatic Marxian approach can be in understanding the dynamics of trade, ethnicity, status, and mode of production in complex lineage-based societies. The anthropologist Jack Goody (1980), drawing on Baks (1966), began by arguing that 'slavery involves external as well as internal inequality, an unequal balance of power between peoples.' From this unpromising start he enriches his analysis with both the ethnohistorical data on Africa and the cross-cultural statistical data from the Murdock Ethnographic Atlas. Central to his analysis is the role of women. The complex facts of slavery cannot be explained, he argues, 'except by seeing the role of slaves as related to sex and reproduction as well as to farm and production, and in the case of eunuchs, to power and its non-proliferation.' While being valuable, Goody's arguments are confined to Africa, and even with respect to this continent they are insufficient to explain why it was that some African societies came to rely so heavily on the institution while closely related neighboring groups did not.

The economist Frederic Pryor (1977) has come closest to formulating a general theory of premodern slavery using modern statistical techniques applied to cross-cultural data. He distinguishes between social and economic slavery, and argues that different factors explain the presence of one or the other. Central to his theory is the nineteenth-century idea that there is a correspondence or 'homologism' between male domination of women and masters' domination of slaves. He tried to demonstrate that economic slavery was most likely to occur 'in societies where women normally perform most of the work so that the slave and the wife act as substitutes for each other,' whereas social slavery was related to 'the role of the wife in a polygynous situation.' The theory is interesting and robustly argued but is problematic. Where women dominated the labor force there was no need for men to acquire slaves for economic purposes. On the contrary, it was precisely in such situations that slaves, when used at all, were acquired for social purposes. The presumed correspondence between wives and slaves is also questionable. There were profound differences between these two roles. Wives everywhere belonged to their communities, and were intimately kin-bound, whereas slaves everywhere were deracinated and kinless. Wives always had some rights, some power, at least in the protection of their kinsmen, in relation to their husbands and could rarely be killed with impunity, which was not true of slaves. Wives everywhere had honor, while slaves had none.

Far more comparative work is needed for an understanding of the institutionalization of slavery. The main variables involved in any explanation are now well known. They are: The economic and social role of women; polygyny; internal and external warfare; the mode of subsistence (mainly whether pastoral or agricultural); and the mode of political succession. However, the causal role and direction of each of these variables is complex, and their interaction with each other even more so. To take the role of women as an example, in some cases it is their role as producers that is important; in others, their role as reproducers. It is also difficult using static cross-cultural data to ascertain whether low female participation in production is the result of slavery or its cause.

Approaches to the Study of Genuine Slave Societies

Marxist scholars were the first to take seriously the problem of the origins, nature, and dynamics of large-scale slavery (for a review, see Patterson, 1977a). Engels's (1962) periodization view that large-scale slavery constituted an inevitable stage in the development of all human societies was merely one version of nineteenth-century evolutionism, discussed earlier. It dominated East European thought until the deStalinization movement, and is still the orthodox view in mainland China.

More sophisticated, but no less problematic, have been attempts of modern Marxists to formulate a 'slave mode of production.' The most empirically grounded of these attempts is that of Perry Anderson (1974) who argued that 'the slave mode of production was the decisive invention of the Graeco-Roman world, which provided the ultimate basis both of its accomplishments and its eclipse.' There is now no longer any doubt that slavery was foundational for both Athens and Rome, and that these were the first societies in which large-scale or genuine slavery emerged. However, the concept of the 'slave mode of production' overemphasizes the materialistic aspects of slavery, making it of limited value for the comparative study of slavery. Slaves were indeed sometimes used to generate new economic systems – as in Rome and the modern plantation economies – but there are many cases where, even when used on a large scale, there was nothing innovative or distinctive about the economic structure. The

narrow materialist focus not only leads to a misunderstanding of the relationship between slavery and technology in the ancient world, but more seriously, it fails to identify major differences between the Greek and Roman cases, and it is of no value in the study of genuine slave societies in which the structurally important role of slaves was noneconomic, as was true of most of the Islamic slave systems.

Several non-Marxist historical sociologists, drawing also on the experience of ancient Europe, have made important contributions to our understanding of genuine slave societies. According to Weber (1964) slavery on a large scale was possible only under three conditions: '(1) where it has been possible to maintain slaves very cheaply; (2) where there has been an opportunity for regular recruitment through a well-supplied slave market; (3) in agricultural production on a large scale of the plantation type, or in very simple industrial processes.' However suggestive, there is little support for these generalizations from the comparative data. Medieval Korea (Salem, 1978) and the US South (Fogel and Engerman, 1974) disprove (1) and (2). And work on urban slavery in the modern world as well as the relationship between slavery and technology in the ancient world disprove (3) (Finley, 1973).

Finley (1960, 1973, 1981) was the first scholar to grapple seriously with the problem of defining genuine slave societies, which he explicitly did not confine to those in which the slaves were economically important. He also cautioned correctly against too great a reliance on numbers in accounting for such societies. His emphasis on the slave as a deracinated outsider led him to the conclusion that what was most advantageous about slaves was their flexibility, and their potential as tools of change for the slaveholder class. Finlay also offered valuable pointers in his analyses of the relationship between slave and other forms of involuntary labor. And in criticizing Keith Hopkins's (1978) conquest theory of the emergence of genuine slave societies, persuasively he encouraged an emphasis on demand, as opposed to supply factors in the rise of slave society. Romans, he argued, captured many thousands of slaves during the Italian and Punic wars because a demand for them already existed and 'not the other way around.' He postulated three conditions for the existence of this demand: private ownership of land, and some concentration of holdings; 'a sufficient development of commodity production of markets'; and 'the unavailability of an internal labor supply.'

A Framework for the Study of Slave Societies

It is useful to approach the comparative study of slave society with an understanding of what structural dependence means. There are three fundamental questions. First, what was the *nature* of the dependence on slavery; second, what was the *degree* of dependence; and third, what was the *direction* of dependence. Answers to these three questions together determine what may be called the *modes of articulation* of slavery.

The nature of dependence on slavery may have been primarily economic or social, political or militaristic, or a combination of these. Economic dependence was frequently the case, especially in ancient Rome and in the modern capitalistic slave systems. Here the critical questions are: What were the costs and benefits of imposing and maintaining an economy based on slave labor? Stanley Engerman (1973) has written authoritatively on this subject. He distinguishes between the costs of imposition, of enforcement, and of worker productivity. Lower maintenance costs, more constant supply of labor, greater output due to the neglect of the non-pecuniary costs of labor, higher participation rates and greater labor intensity, and economies of scale, are the main factors proposed by Engerman in explaining the shift to slave labor. They apply as much to ancient Rome as they do to the modern capitalistic slave systems of America and the Caribbean.

Military and bureaucratic dependence were the main noneconomic forms, and they could be as decisive for societies as was economic dependence. The rise of Islam was made possible by the reliance on slave soldiers, and the Abbasid Caliphate, along with other Islamic states, was maintained by slave and exslave soldiers and bureaucrats. Although there was little economic dependence on slaves in most Islamic states – with the notable exception of ninth-century Iraq – many of them qualify as genuine slave societies as a result of the politico-military dependence of these regimes on slavery, and the ways in which slaves influenced the character of their cultures, many of the literati also being slaves or descendants of slaves (see Pipes, 1981; Crone, 1981; Marmon, 1999).

The degree of dependence must be taken into account, although it is sometimes difficult to quantify. There has been a tendency to emphasize the size of the slave population over other variables, and while demography is important it can sometimes be misleading. Thus, it has been noted that no more than one in three adults were slaves in Athens at the height of its slave system in the late fifth-century BCE as a way of playing down the importance of slavery. But as M. I. Finley liked to point out, the same was true of most parts of the American Slave South, and no one has ever questioned the fact that it was a large-scale slave system.

Finally, there is the direction of dependence. Even where a society had a high functional dependence on slavery it was not necessarily the case that slavery played an active, causal role in the development of its distinctive character. The institution, though important, was not structurally transformative. In ancient Rome, the Sokoto Caliphate and the slave societies of the Caribbean, the American South and Brazil, slavery was *actively articulated* and transformative. In other cases, however, it was *passively articulated*, the classic instance being Korea during the Koryo and Yi periods where, although a majority of the rural population were at times slaves, there was no significant transformation of the economy, and no impact on the regime's government and culture. The same was true of several of the modern Spanish colonies of Central and South America. During the sixteenth and seventeenth centuries there was marked structural dependence on slavery in Mexico and Central America as well as Peru, and the urban areas of Chile and Argentina, but the institution was not determinative in the course of development of these societies and, as in Korea, when slavery ended it left hardly a trace, either cultural or social (Mellafe, 1975; Palmer, 1976; Klein, 1986; Blackburn, 1997).

These three factors together determined a finite set of modes of articulation, which is a sociologically more useful construct than that of the so-called slave mode of production. Space permits only the most cursory mention of some of the most important such modes of articulation.

The *lineage mode of articulation* refers to those kin-based societies in which large-scale slavery was related critically to the rise to dominance of certain lineages in the process of class and state formation. In some cases, slavery originally served primarily economic ends; in others, mainly social and political ones; but in the majority the institution became multifunctional. This kind of genuine slave society was most commonly found in western and west-central Africa. Warfare, combined with some critical internal factor such as demographic change accounted for the rise of slavery (see Miller, 1977 for the Kongo kingdom). The ideal case of this mode of articulation was the Asante state (Wilks, 1975; Klein, 1981). Slaves were originally incorporated as a means of expanding the number of dependents, a tendency reinforced by the matrilineal principle of descent. The growing number of slaves enhanced the lineage heads who owned them, and facilitated the process of lineage hierarchy and state formation. Later, the role of slaves was greatly expanded to include a wide range of economic activities, including mining.

The *predatory circulation mode* refers to those slave societies in which warfare and raiding mainly for slaves were the chief occupations of a highly predatory elite. The warrior class was usually assisted by a commercial class which traded heavily in slaves. There was usually a high rate of manumission of slaves, who not only contributed to the production of goods, but in their role as freedmen soldiers helped the ruling class to produce more slaves. Thus there was a continuous circulation of persons in and out of slave status as outsiders were incorporated as loyal freedmen retainers, creating a constant need for more enslaved outsiders. Slaves and freedmen often played key roles in, and sometimes even dominated, the palatine service and elite executive jobs.

The contiguous existence of pastoral and sedentary agricultural peoples with different levels of military might was the major factor in the development of this mode of articulation of slavery. The mode is strongly associated with Islam, and there are many examples of it in the Sahel and North Africa. The west–east spread of the Fulani over an area of some 3000 miles over a period of 800 years provides one of the best cases of this mode, on which there is an abundance of historical and anthropological data (Lovejoy, 2000; Pipes, 1981; Meillassoux, 1975, 1991).

The *embedded demesne mode* embraces those patrimonial systems dominated by large landed estates in which slaves were incorporated on a substantial scale to cultivate the demesne land of the lords. Serf or tenant laborers continued, in most cases, to be the primary producers of food for the society, and their rents or appropriated surpluses remained a major source of wealth for the ruling class. However, slaves were found to be a more productive form of labor on the home farms of the lords for the cost–benefit reasons analyzed by Engerman, mentioned above. The landowners got the best of both types of labor exploitation. This was a particularly valuable arrangement where a landed aristocracy needed to change to a new crop but was not prepared to contest the technical conservatism of serfs; where there was a supply of cheap labor across the border; or where there was a high level of internal absenteeism among the landed aristocracy.

This is the least understood or recognized form of advanced slave systems, perhaps because its mode of articulation was usually passive. Many of them were to be found in medieval Europe, especially in France, Spain, and parts of Scandinavia (Dockes, 1982; Verlinden, 1955, 1977; Bonassie, 1985; Anderson, 1974; Patterson, 1991). The slave systems of the western 'states' of eleventh-century England, where slave populations were as high as 20% of the total are likely examples. So, possibly, was Viking Iceland, which may well have had similar proportions of slaves (Williams, 1937; Foote and Wilson, 1970). However, the ideal case of embedded demesne slavery was to be found in Korea, especially during the Koryo and early Yi periods. Here the slave population sometimes exceeded that of other forms of bonded labor (Salem, 1978; Wagner, 1974; Hong, 1979).

The *urban–industrial modes* of articulation were those in which the urban elites came to rely heavily on slaves for support. Slaves played a relatively minor role in agriculture, although they may well have dominated the 'home farms' of certain segments of the ruling urban elites. Slave labor was concentrated in urban craft industries which produced goods for local consumption and exports, as well as the mining sector where it existed.

Slavery emerged on a large scale in such systems as a result of a combination of factors among which were the changing nature and frequency of warfare, conquest of foreign lands, the changing tastes of the ruling class, crises in the internal supply of labor, shifts in food staples, growing commercial links with the outside world, and demographic changes.

This mode of articulation could be either passive or active. To the extent that the character of the urban civilization depended on its urban economy, and to the extent that the economy depended on slave laborers, both manual and technical, to that degree were these systems active in their articulation. The classic case of the active mode of slave articulation were the ancient Greek slave systems, especially Athens of the fifth- and fourth-centuries BC (Finley, 1981; Garlan, 1982; De Ste. Croix, 1981). Typical of the passive mode of urban–industrial articulation were several of the Spanish slave systems of Central and South America during the sixteenth and seventeenth centuries, mentioned above.

The *Roman or urban–latifundic mode:* ancient Roman slavery stands in a class by itself, having no real parallels in either the ancient, medieval, or modern worlds. It came the closest to a system of total slavery in world history. It was distinctive, first, in the sheer magnitude of its imperial power and the degree of dependence on slavery, both at the imperial center and in its major colonial sectors. Second, Rome was unique in the extent of its reliance on slavery in both its rural and urban–industrial sectors. Third, the articulation of slavery was more actively transformative than in any other system, entailing what Hopkins (1978) called an 'extrusion' of free, small farmers and their replacement by slaves organized in gangs on large latifundi. Rome was unusual too, not only for its high levels of manumission, but for the extent to which slaves and slavery came to influence all aspects of its culture.

The *capitalist plantation mode:* contrary to the views of early economic theorists such as Adam Smith, and of Marxist scholars until fairly recently (Genovese, 1965), modern plantation slavery was in no way incompatible with capitalism. Indeed, the rise of capitalism was intimately bound up with this mode of articulation of slavery, and at its height in

nineteenth-century America was as profitable as the most advanced industrial factories of Europe or the northern United States (Fogel and Engerman, 1974). Plantation slavery constituted one version of the worldwide systemic spread of capitalism, in which capital accumulation was advanced through the use of slaves in the peripheral colonial regions, complementing the use of so-called free labor in the metropolitan centers, and of serf and other forms of dependent labor in the semiperipheral areas of Eastern Europe and Latin America (Wallerstein, 1974; Blackburn, 1997).

While plantation slavery bore some organizational resemblance to the ancient slave latifundia, and indeed can be traced historically to late medieval and early modern variants of the ancient model (see Solow, 1991), it was distinctive in its complex, transnational system of financial support, in its production for international export, its heavy reliance on a single crop, its advanced organizational structure, which in the case of sugar involved innovative agri-industrial farms, in the vast distances from which slaves were imported, entailing a complex transoceanic slave-trading system, and in its reliance on slaves from one major geographic region who differed sharply from the slaveholding class in ethnosomatic terms.

However, it is important to recognize that there existed concurrently with the capitalistic plantation mode several other modes of slave articulation that were either precapitalist or at best protocapitalist. As has already been noted, many of the Spanish colonial systems in the Americas relied on forms of slavery that were nonaccumulative and distinctly premodern in their articulation, for example the urban–industrial modes of Mexico and parts of South America which were focused heavily on mining. In the Spanish Caribbean up to the third quarter of the eighteenth century a peculiar premodern form of agri-pastoral slavery prevailed which had little in common with the highly capitalistic slave plantation systems of the neighboring French and British islands, and which had to be dismantled at considerable political and socioeconomic cost in order to make way for the belated capitalistic slave systems that were imposed forcefully during the nineteenth century (Knight, 1970).

While slavery was largely abolished in the Americas through the course of the nineteenth century (Blackburn, 1988; Davis, 1975), pockets of the institution persist into the twenty-first century in northern Africa and parts of Asia (Stearman, 1999). With the possible exception of Mauritania, however, in none of these societies do we find genuine slave systems. It can be claimed, with cautious optimism, that this most inhuman form of societal organization has vanished from the world.

Bibliography

Anderson, P., 1974. Passages from Antiquity to Feudalism. New Left Books, London.
Ayalon, D., 1951. L'Esclavage du Mamelouk. Jerusalem.
Baks, C.J, et al., 1966. Slavery as a system of production in tribal societies. Bijdrgen tot de Taal-, Land en Volkenkunde 122.
Biot, E., 1840. De l'abolition de l'esclavage ancien en Occident. Paris.
Blackburn, R., 1988. The Overthrow of Colonial Slavery, 1776–1848. London.
Blackburn, R., 1997. The Making of New World Slavery. London.
Bonassie, P., 1985. Survie et extinction due regime esclavagiste dans l'Occident du haut moyet age. Cahiers de Civilisation Medievale 28.
Crone, P., 1981. Slaves on Horses. New York.
Davis, D.B., 1966. The Problem of Slavery in Western Culture. Cornell University Press, Ithaca, NY.
Davis, D.B., 1975. The Problem of Slavery in the Age of Revolution, 1770–1823. Cornell University Press, Ithaca, NY.
De Ste Croix GEM, 1981. The Class Struggles in the Ancient Greek World. Cornell University Press, Ithaca, NY.
Dockes, P., 1982. Medieval Slavery and Liberation. University of Chicago Press, Chicago.
Domar, E., 1970. The causes of slavery or serfdom. Journal of Economic History 30, 18–32.
Engels, F., 1962. Anti-Duhring. Foreign Languages Publishing House, Moscow.
Engerman, S., 1973. Some considerations relating to property rights in man. Journal of Economic History 33, 49–51.
Finley, M.I. (Ed.), 1960. Slavery in Classical Antiquity. William Heffer and Sons, Ltd., Cambridge.
Finley, M.I., 1968. Slavery. International Encyclopedia of the Social Sciences, vol. 14. Macmillan, New York.
Finley, M.I., 1973. The Ancient Economy. Chatto and Windus, London.
Finley, M.I., 1981. Economy and Society in Ancient Greece. In: Shaw, B.D., Saller, R.P. (Eds.). New York.
Fogel, R.W., Engerman, S., 1974. Time On the Cross. Boston, vols. 1 and 2. Little, Brown.
Foote, P., Wilson, D., 1970. The Viking Achievement. Hurst and Co., Ltd., London.
Garlan, Y., 1982. Les esclaves en grece ancienne. Maspero, Paris.
Genovese, E.D., 1965. The Political Economy of Slavery. Oxford University Press, New York.
Goody, J., 1980. Slavery in time and space. In: Watson, J.L. (Ed.), Asian and African Systems of Slavery. Oxford.
Hong, S.H., 1979. The legal status of the private slaves in the Koryo Dynasty (in Korean). Han'gukakpo 12.
Hopkins, K., 1978. Conquerors and Slaves. Cambridge University Press, Cambridge.
Klein, A.N., 1981. The two Asantes. In: Lovejoy, P. (Ed.), The Ideology of Slavery in Africa. Sage, Beverly Hills, p. 152.
Klein, H.S., 1986. African Slavery in Latin America and the Caribbean. Oxford University Press, New York.
Knight, F.W., 1970. Slave Society in Cuba during the 19th Century. University of Wisconsin Press, Madison, WI.
Lovejoy, P.E., 2000. Transformations in Slavery: A History of Slavery in Africa. Cambridge University Press, Cambridge.
Marmon, S.E. (Ed.), 1999. Slavery in the Islamic Middle East. Wiener Publishers, Princeton, NJ.
Meillassoux, C., 1975. L'esclavage en Afrique precoloniale. Maspero, Paris.
Meillassoux, C., 1991. The Anthropology of Slavery: The Womb of Iron and Gold. (A. Dasnois, Trans.). Athlone Press, London.
Mellafe, R., 1975. Negro Slavery in Latin America. University of California Press, Berkeley, CA.
Miller, J.C., 1977. Imbangala lineage slavery. In: Miers, S., Kopytoff, I. (Eds.), Slavery in Africa. Madison, WI.
Miller, J.C. (Ed.), 1993. Slavery and Slaving in World History: A Bibliography. Millwood, New York.
Miller, J.C., Holloran, J.R. (Eds.), 2000. Slavery: annual bibliographical supplement. Slavery and Abolition 21 (3).
Nieboer, H.J., 1910. Slavery as an Industrial System. Rotterdam, The Netherlands.
Palmer, C., 1976. Slaves of the White God: Blacks in Mexico, 1570–1650. Harvard University Press, Cambridge, MA.
Patterson, O., 1977a. Slavery. Annual Review of Sociology 3, 407–449.
Patterson, O., 1977b. The structural origins of slavery: a critique of the Nieboer-Domar hypothesis. Annals of the New York Academy of Science 292, 12–34.
Patterson, O., 1982. Slavery and Social Death. Harvard University Press, Cambridge, MA.
Patterson, O., 1991. Freedom in the Making of Western Culture. Basic Books, New York.
Pipes, D., 1981. Slave Soldiers and Islam. Yale University Press, New Haven, CT.
Pryor, F.L., 1977. A comparative study of slave societies. Journal of Comparative Economics 1.
Salem, H., 1978. Slavery in Medieval Korea. Ph. D. dissertation, Columbia University.
Siegel, B.J., 1945. Some methodological considerations for a comparative study of slavery. American Anthropologist 7.
Solow, B. (Ed.), 1991. Slavery and the Rise of the Atlantic System. Harvard University Press, Cambridge, MA.
Stearman, K., 1999. Slavery Today. Wayland Publisher, Hove.
Tourmagne, A., 1880. Histoire de l'esclavage encien et moderne. Paris.
Verlinden C L'Esclavage dans l'Europe Medievale. vols. 1 and 2 (Bruges, 1995; Ghent, 1977)
Wagner, E., 1974. Social stratification in 17th century Korea. Occasional Papers on Korea 1.
Wallerstein, I., 1974. The Modern World System. Academic Press, New York.
Weber, M., 1964. The Theory of Social and Economic Organization. Free Press, New York.
Westermarck, E., 1908. The Origins and Development of the Moral Ideas. Macmillan, London.
Wilks, I., 1975. Asante in the Nineteenth Century. Cambridge University Press, Cambridge.
Williams, C.O., 1937. Thralldom in Ancient Iceland. Chicago, IL.

Slaves and Slavery, History of

Trevor Burnard, University of Melbourne, Parkville, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Wirz, volume 21, pp. 14157–14162, © 2001, Elsevier Ltd.

Abstract

Slavery has existed for millennia. It is both a status – an inferior person with no rights in a society – and a condition, a kind of humanity which was accorded fewer rights than other forms of humanity. A slave was the quintessential outsider. Most societies had slaves but only some societies – ancient Greece and Rome and the plantation colonies of European powers in the Americas – were slave societies in which slavery was essential to the economy and politically sensitive. In modern times, slavery has been strongly connected with race, especially the forced removal of Africans to the Americas to labor in the production of tropical goods.

"O what a rogue and peasant slave am I", William Shakespeare has Hamlet, the doomed Prince of Hamlet exclaim at the end of Act 2 of 'Hamlet'. Hamlet is a prince and so was almost as far removed from an actual slave as it is possible to be. Moreover, neither in Denmark, where the play is set, nor in early seventeenth century England, where the play was performed, was slavery common. But the idea of slavery, as Shakespeare recognized, was well-known even in places where a person would seldom meet a slave let alone become a slave oneself. Shakespeare used the imagery of slavery to get at a deeper truth. To be a slave or to feel like a slave, was to be in the pit of despair because no condition could be worse than being a slave. Slavery was a form of exploitation in which one human was owned by another person and in which the slave hovered uncertainly between the contradictory positions of being both a piece of property and also a person. It was also, as imaginative artists since Aristophanes and thinkers beginning with Aristotle have realized, a state of mind; a status that affected how one thought of oneself, if one was a slave, and how others thought of slaves.

Slavery comes in many culturally specific forms. Common to all these different forms is the fact that slaves are denied the most basic rights. Slaves remain outside the social contract that binds the other members of a given society; they are people without honor and kin, subject to violent domination. They are objects of the law, not its subjects, considered as property or chattel. Hence, we may describe slavery (with Orlando Patterson) as a form of 'social death', and slaves as outsiders kept in a position of institutionalized marginality. Historians accept part of Patterson's concept but are wary of being too dogmatic in seeing slaves as 'socially dead', because such a view tends to deny agency to slaves and tends toward accepting the view often put forward by the master class that enslaved people were akin to beasts or nonhuman property. The trend at present is to see slavery less as a form of social death than as a negotiated relationship. In this relationship masters had most of the power but their power was always contested by enslaved people who had their own imperatives that often conflicted with their masters' desires. Enslaved people formed their own communities with value systems and forms of behavior that differed from the social norms of nonenslaved people; but these communities were set within larger social groupings that exerted their own pull over enslaved people's desires. If slaves were legally 'socially dead' this death was more a legal fiction than a reality. As a host of historians have argued in the last decade, slaves fashioned lives full of meaning for themselves even in the direst circumstances of want and physical and psychological oppression (Patterson, 1982).

Slavery is tied up less with law than with custom. The slave was the quintessential outsider, no matter how dependent a society was on the slaves' labor. That status of outsider was sometimes based on ethnicity, sometimes race, and sometimes residence. In many languages, for example, the term for foreigner also denoted slaves. Nowadays, slavery is considered a most inhumane practice. The United Nations Universal Declaration of Human Rights proclaims: "No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms". Signing the slavery conventions of 1926 and 1956, 122 states have committed themselves to abolish slavery and similar practices such as debt bondage, serfdom, servile marriage, and child labor (Bales, 1999).

In historical times, however, slavery as an institution was accepted in many societies, from early Mesopotamia and Pharaonic Egypt to ancient China, Korea, India, medieval Germany, and Muscovite Russia; from the Hebrews to the Aztecs and Cherokees. Indeed, slavery was a commonly accepted practice throughout Biblical societies and times, and is mentioned favorably as a social practice in both the Old and New Testaments. Nevertheless, according to classical historian Moses Finley there were only five full-fledged slave societies with slaves constituting the majority of people and their work determining the whole economy. All five were located in what we tend to call the West; two in antiquity – Athenian Greece and Rome between the second century BC and AD fourth century; and three in modern times – Brazil, the Caribbean, and the Southern United States. Slavery did exist elsewhere. The Ottoman Empire, for example, was notorious for its slave soldiers, the Janissary, taken from its Slavic neighbors to the North and from Africa. But it was only in the five societies noted above that slavery became the dominant institution in society, shaping all social, economic, and legal relationships. Finley's major contribution to the history of slavery came with his distinction of 'slave societies', such as ancient Greece and pre-twentieth century Brazil, where everything was affected by

slavery and the multitude of societies in which slavery existed but did not form the essential bulwark of social and political life. Most histories of slavery concern 'slave societies' rather than 'societies with slaves'; but many people lived in 'societies with slaves' and often attitudes to slavery were shaped, most of all, in such societies where slavery existed but was of marginal importance. Eighteenth century Western Europe and the Northern colonies of British North America were societies with slaves, rather than slave societies. It was in these societies that the morality of slavery was questioned for the first time. Thus, when discussing the history of slavery we need to also consider antislavery, or attempts to make slavery not normative but peculiar (Davis, 1984; Finley, 1987).

Slavery until the late eighteenth century was so normal as to be almost universal. It was so common that few people thought about whether it was right or wrong. Nevertheless, slavery always attracted historical attention – from Aristotle in ancient times to contemporary commentators. Slavery may be illegal in the twentieth century but millions of people in Asia and Africa still labor in conditions that can only be called slavery. Before the mid-eighteenth century most writers followed Aristotle in thinking of slavery as mostly a positive good, adding to the sum of human happiness even if unfortunate for those caught up in this institution as slaves. Aristotle justified slavery on four grounds. One was racist. He believed that Easterners were intelligent but soft and Northerners were strong and stupid. The latter were ideally suited to hard work but needed to be coerced to do so since they were naturally lazy. They could only be forced to work through coercion and their strength needed to be controlled through the application of extreme force when they disobeyed their owners. It was easy for Europeans to extend this analysis to Africans when after the sixteenth century it became problematic to enslave people of European descent. The second assumption was related to the first: some people had to be controlled for their own good. Just as children had to obey parents, so too it was good if inferior people accepted the commands of their intellectual superiors. Aristotle's third reason was tautological. Once people became slaves they adopted slavish behavior, thus justifying treating them as slaves. Southern slave owners were very fond of this kind of reasoning. Finally, slavery was necessary because without slavery work would not get done. In the plantation colonies of the Caribbean and Brazil such reasoning had great purchase. Working as laborers producing tropical goods, notably sugar, was so disagreeable that no one would do it if they had a choice to be involved in some other activity (Davis, 1966).

Aristotle's defense of slavery prevailed, more or less, as the major historical interpretation of slavery, until the mid-eighteenth century, when for the first time a genuine anti-slavery movement emerged. Ironically, it emerged at the time when the Atlantic slave trade ferrying captive Africans to the Americas was at its height and when plantation slavery in the Americas was most profitable and most important for European prosperity. The writers who began to question slavery, starting with Montesquieu and increasingly including all Enlightenment thinkers, came to question slavery as part of an overall critique of commercial society. Taking their cues from Greece, Rome, and Medieval European experiences, they tended to interpret slavery as an intermediate stage in the general process of historical development. G.W.F. Hegel construed slavery as a sort of development aid for the benefit of Africans. Classical social scientists from Adam Smith to Max Weber identified slavery with backwardness. Marxist thinkers were even more explicit, positing primitive society, slavery, feudalism, capitalism, and communism as five consecutive stages of world history. Thus, slavery and capitalism, slavery and markets, and slavery and technological innovations were considered utterly incompatible (Burnard and Heuman, 2010; Blackburn, 1997).

Studies of US and West Indian slavery have dominated the field of slave studies from its very beginning, although this dominance is beginning to change as scholars start to focus more on Africa and Europe as the places where forms of slavery were first established (Eltis, 2000; Thornton, 1992). Nevertheless, scholars of US slavery, in particular, have set the agenda for research into other domains, such as the transatlantic slave trade and its precedents in the Mediterranean and the Black Sea during the Middle Ages and the Renaissance; and slavery in the Indian Ocean world and in Muslim Africa (Genovese, 1974). Studies of classical slavery have been less dependent on models drawn from studies of New World slave systems. The dominance of studies of US slavery in the historiography of slavery has been unfortunate, because the model for slavery became the Antebellum Southern slave plantation, which was by any terms, highly aberrant within the long history of slavery. Until recently, slavery was treated in essentially ahistorical terms – as an institution that varied little over time and place and about which grand generalizations could be made. The dominant trend in recent historiographies of slavery has been to place slave societies very much in the context of their times. Ira Berlin has been the prime exponent of the necessity of treating slavery as a highly variable institution. Slavery was different when it was first established, when the laws around slavery were uncertain, from what it became once the plantation system had been fully established in the early eighteenth century; and it was different again when the age of revolutions and the abolition of the slave trade in the late eighteenth century focused attention on whether it was moral to deprive individuals and their descendants of their liberty, no matter how valuable their work was to European and American prosperity (Berlin, 1998).

The first historians of slavery tended to be slave apologists. U.B. Phillips may be taken as representative of a larger trend. Writing in the early years of the twentieth century, Phillips argued that despite occasional cruelties, slavery in the United States was mostly benign and indeed was well suited to people of African descent as a result of their innate capacities (Phillips, 1918). Historians have had to fight against the racism that has been implicit in much of the early study of slavery and have had to try, in particular, to write against sources in which evidence was filtered through a lens conditioned by racist prejudice, especially against Africans. The first changes in approach to slavery came after World War II, when scholars such as Kenneth Stampp tried to treat slavery on its own terms. His work tended to stress to an unusual degree the brutality and exploitative nature of slave life in Antebellum America. For some scholars, this brutality reminded them of very recent events, notably Nazi outrages in Eastern Europe against the Jews (Stampp, 1956). Slavery studies took a turn that eventually proved unproductive. Stanley Elkins argued that slavery

should be seen as a quasi-Nazi institution, as a close system of generalized violence and totalitarian control with dehumanizing effects similar to Nazi concentration camps. For Elkins and for other historians in the 1950s and 1960s, slaves were victims, devoid of any power or will of their own. Elkins' efforts at psychological speculation on how slaves coped with brutalization and from the despair of living in a closed system, ultimately were an unprofitable scholarly detour. It has had the unfortunate effect of suggesting that it was not just that slaves were victims but that the slave system was akin to what happened in concentration camps, with slaves having no hope of gaining any psychic maturity (Elkins, 1968). Such an interpretation was not what scholars intended. The result of this approach to slavery (and also the works of econometrists such as Robert W. Fogel and Stanley W. Fogleman, whose much-heralded economic survey of Antebellum Southern slavery in 1974, suggested that slavery was, contra to Adam Smith, an inherently profitable institution) was to suggest both that slavery was not that terrible and also that enslaved people were so under the control of their masters that their very personalities were deformed. Ironically, the historiography of the late 1960s and early 1970s in the United States replicated the kinds of ideas that Antebellum slave holders themselves said characterized enslavement (Fogel and Engerman, 1974).

Ever since that moment in the historiography of slavery, scholars have concentrated on restoring slave agency to the history of slavery. Slavery came to be seen as a negotiated relationship and as an intrinsically unstable process, the terms of which varied with time and place. This process spanned different stages from enslavement and transportation to life in slavery and possible, yet by no means inevitable, social reintegration. There were different ways to overcome the outsider status: resistance and flight, manumission, or the purchase of freedom. Prompted by the American Civil Rights Movement, slave lives, slave culture, and slave resistance developed into privileged subjects of study. Herbert G. Gutman showed that slaves against all odds were living in families, creating enduring networks of kinship and self help (Gutman, 1976). Religion proved to be another major sphere where slaves gained some cultural autonomy. Integrating Christian teachings and African beliefs they developed a richly textured religious life with dance and song as integral parts of worship. The patterns of slave life became a major object of study, with Philip D. Morgan's comparison of eighteenth century South Carolina and Virginia being the best example yet of a comprehensive and sensitive portrait of the numerous ways in which slaves created lives for themselves separate to their masters (Morgan, 1998). Other scholars have focused on resistance, arguing not only that slaves continually resisted the terms and conditions of their enslavement, but also that their resistance was pivotal in shaping important political events such as the Age of Revolutions – a period in which the Haitian Revolution, the only successful slave revolt in history, is increasingly important (Dubois, 2004).

One major theme in the study of slavery has been the effort to determine the numbers of people who were involved in various slave trades. The most determined efforts have been focused on the Atlantic slave trade from West Africa. Philip D. Curtin was the first to attempt a census of the transatlantic slave trade using customs records from the Americas (Curtin, 1969). His calculations have later been complemented and refined, most notably in the Transatlantic Slave Trade Database, a comprehensive accounting of the great majority of voyages that crossed the Atlantic between the fifteenth and nineteenth centuries (Eltis, 2008). Curtin did not have the advantage of later scholars in being able to utilize digital methodology to establish precise numbers, yet his general insights have stood remarkably well. Others tried to extend his census to the older, established Trans-Saharan, Indian Ocean, and Red Sea trades (Vaughan, 2005). Scholars reconstructed slave prices at the points of purchase and sale, carefully tracing changes over time and studying all the other aspects of slave trading and the plantation business. These endeavors engendered seminal new insights. Scholars were able to prove that the length of a voyage was the single most important factor determining mortality during the voyage across the sea, the ghastly Middle Passage, with death rates ranging from 10 to 20%. They showed that pricing was competitive and that supply was responsive to changes in price and that the transatlantic slave trade continued well into the nineteenth century. It ended when Brazil and Cuba emancipated their slaves in the 1880s, the last nations to do so in the Western hemisphere. Scholars also discovered that markets shifted Eastward and Southward along the West African coast, from Senegambia and the Upper Guinea coast to the Bight of Biafra and Northwestern Angola. Imports into Africa varied over time and from place to place, with cotton goods, brandy, guns, and iron tools in high demand everywhere along the coast. With the recent publication of more than 27 000 Atlantic slave trading voyages in one single data set, even nonspecialists have access to parts of the evidence used in the econometric computations. Taken together, studies of the new economic history show that the Marxist thesis that there is an inherent contradiction between slavery and capitalism is untenable, because both slave trading and plantation economies were capitalist to the core, with prices set by the laws of supply and demand (Eltis, 2008; Byrd, 2008; Clarence-Smith, 1989; Manning, 1990; Rediker, 2007).

A major contentious issue among scholars has been the significance of slavery for the Industrial Revolution. In 1944, Eric Williams argued that the profits from the 'triangular trade' had financed the British Industrial Revolution, adding that abolition was a consequence of economic decline (Williams, 1944). This double-pronged 'Williams thesis' became a central tenet of Dependency theory. A close look at hundreds and thousands of slave voyages, however, showed that average profits were smaller than scholars had previously assumed and could never explain why Britain achieved self-sustained industrial growth in the latter half of the eighteenth century. New data on capital formation in that period points in the same direction. Nevertheless, slavery and the plantation economies contributed to British economic growth and the advent of modernity. Slavery reduced the cost of sugar and tobacco for European consumers. It made great fortunes for a few planters and merchants. It also created markets for cheap industrial mass-products. Indeed, sugar plantations with their gang-labor system were 'factories in the fields' with a work discipline prefiguring that of modern industrial plants. The impact of slavery was also felt in the realm of ideas and ideology. It furthered a racially informed negative perception of all things African in the West; and set

the stage for systematic racism against people of African descent (Drescher, 1977).

Recent scholarship has diminished the importance of Antebellum America as the place where slavery was most important. Indeed, the recent turn toward Atlantic history has illustrated that the Antebellum South was indeed a peculiar place, not just in nineteenth century America but in slavery generally. Most Atlantic slave societies did not reproduce themselves naturally, unlike the antebellum South; meaning that they relied on constant inputs of fresh captives from Africa in order to survive. The center of slavery studies has moved backward in time – from the nineteenth to the eighteenth century – and has moved southward, to the Caribbean and Brazil. Most importantly, the key place shaping slavery was Africa. The impact of slavery on Africa has been much debated. An earlier generation of scholars with firsthand experience of colonial domination saw African societies mainly as victims of external forces. They singled out the transatlantic slave trade as a first step toward underdevelopment, as it depopulated whole areas, (or so ran the argument) and induced African rulers to engage in wars for the sake of taking prisoners for sale in exchange for guns, which then triggered new wars. This set of ideas informed Dependency theory as well as world-system analysis and early nationalist African historiography. Walter Rodney even argued that African slavery on the Guinea coast was a result of external demand (Rodney, 1982).

Increased attention to what happened in Africa has changed these assumptions. Scholars such as John Thornton have shown that to understand the Atlantic slave trade means understanding African economics and politics, because it was Africans who controlled the trade, not Europeans (Thornton, 1992). Moreover, just as Europe is a diverse continent, so too was Africa. To understand slavery, it is necessary to understand African diversity, over both space and time. To stay in business, European and American slave traders had to adapt to local circumstances. They were at the mercy of local merchants and local power holders. They hardly ever ventured beyond the ports of call and the forts they built along the West African coast. They also bought women and children, (although the planters considered them a second choice to men), as well as people from areas other than those that the planters preferred. Women constituted up to a third of the slaves sent to the Americas. The female ratio among slaves was higher than among early European immigrants. It ensured an enduring African physical presence (Lovejoy, 1983).

Studies of African slavery have furthered the understanding of these cultural parameters. Slavery was widespread in precolonial Africa and expanded in the nineteenth century despite, or rather because of, the abolition of the transatlantic slave trade by the UK in 1807. Some scholars even argue that West African societies then developed into fully fledged slave societies with 50% and more of their population kept in bondage. First among these were the Sudanic societies from the Futa Jallon to Sokoto and Adamawa, where Islamic revolutions and state-building initiated more intense slave trading than ever before. Yet, African slavery, often referred to as 'household slavery', was markedly differed from chattel slavery. As elsewhere, slaves had to do the hardest work and their owners treated them as outsiders, the first to be sacrificed in public rituals and put to death at the whim of their owner. Nevertheless, it was common to find slaves in positions of albeit borrowed authority, such as traders, officers, and court officials. Elsewhere slave and slave owner worked side by side. Alternatively, slaves lived in villages of their own. They owned property and in some cases, even slaves of their own. More importantly, the slave status was not a fixed status; rather it changed over time along a slave to kinship continuum, with people born into slavery exempt from further sale. Recent scholarship has stressed that women were the preeminent victims of slavery in Africa as they were valued for their labor power and for the offspring they might give birth to. Moreover, one should never forget that power and prestige in African societies depended on the number of followers a person could count on, rather than on the control of land which was not a scarce resource yet, or at least,not everywhere (Lovejoy, 1983; Meillassoux, 1991).

Colonialism has helped to keep the memory of slavery alive in postcolonial societies, from the United States to Brazil, to the Caribbean, and to Africa. In the early nineteenth century, Britain had done its best to end the slave trade. It even stationed a naval squadron for this purpose in West African waters and concluded antislave trade treaties with African rulers who had trouble seeing the benefits of abolition. The European public sometimes saw the imperialist scramble for Africa as an antislavery crusade; but colonial administrations, while forcefully suppressing the slave trade, were reluctant to fight slavery as a social system; rather they closed their eyes and did their best to freeze existing social relations while claiming to bring progress to Africa. When slavery finally ended in Africa, it did so not so much as the result of conscious acts of emancipation but rather because of the emergence of a colonial labor market, with the socially defined older mechanisms for the integration of outsiders easing the transition. The men and women held in bondage eagerly took up alternatives wherever they were viable.

Whereas studies of slavery have become ever more detailed and localized, historians have come to realize that the Middle Passage is more a bridge rather than an abyss or a one-way street, binding different cultures together, a point first made by the anthropologist Melville J. Herskovits (Herskovits, 1941). Slaves came with a history of their own to the Americas; returning Africans and children of former slaves such as Olaudah Equiano, Wilmot Blyden, and James Africanus Horton were among the first to consciously define themselves as Africans. Hailing from different regions of Africa, yet sharing a common plight, they developed an awareness of a common African identity. The books they wrote helped to lay the foundations for Pan-Africanism, a strand of ideas and a political movement central to the freedom struggle of Africans in the twentieth century.

David Brion Davis noted a similar correlation between slavery and freedom in Western religious, legal, and philosophical discourses (Davis, 1966). Hence the double paradox and legacy of slavery: it inflicted death and hardship on millions and millions of people of mostly African descent for the benefit of a few; and while it was instrumental for the rise of racial prejudice, it also shaped the notion of freedom and equality as we know and cherish it today. To consider these global dimensions is the challenge of any further research into the history of slavery, even when dealing with very specific local aspects of the problem. But this is easier said than done because

it requires historians to venture beyond the deeply rooted traditions of privileging the history of a particular nation-state, usually their own, into the open sea of a new comparative history. The fight against slavery is not won so long as debt bondage, servile marriage, forced labor, child labor, trafficking of women, forced prostitution, and other forms of bondage exist in many societies, including the rich countries of the West (Sweet, 2004).

See also: Colonialism, Anthropology of; Colonization and Colonialism, History of; Human Rights, History of; Slavery as Social Institution; Slavery: Comparative Aspects; Trade and Exchange, Archaeology of.

Bibliography

Bales, K., 1999. Disposable People: New Slavery in the Global Economy. University of California Press, Berkeley.
Berlin, I., 1998. Many Thousands Gone: The First Two Centuries of Slavery in North America. The Belknap Press, Cambridge, MA.
Blackburn, R., 1997. The Making of New World Slavery. From the Baroque to the Modern, 1492–1800. Verso, London.
Burnard, T., Heuman, G., 2010. The Routledge History of Slavery. Routledge, London.
Byrd, A., 2008. Captives and Voyagers: Black Migrants across the Eighteenth-Century British Atlantic World. Louisiana State University Press, Baton Rouge.
Clarence-Smith, W.G., 1989. The Economics of the Indian Ocean Slave Trade in the Nineteenth Century. Frank Cass, London.
Curtin, P.D., 1969. The Atlantic Slave Trade: A Census. University of Wisconsin Press, Madison, WI.
Davis, D.B., 1966. The Problem of Slavery in Western Culture. Cornell University Press, Ithaca, NY.
Davis, D.B., 1984. Slavery and Human Progress. Oxford University Press, New York.
Dubois, L., 2004. Avengers of the New World: The Story of the Haitian Revolution. Harvard University Press, Cambridge.
Drescher, S., 1977. Econocide: British Slavery in the Era of Abolition. University of Pittsburgh Press, Pittsburgh, PA.
Elkins, S.M., 1968. Slavery; A Problem in American Institutional and Intellectual Life, second ed. University of Chicago Press, Chicago.
Eltis, D., 2000. The Rise of African Slavery in the Americas. Cambridge University Press, Cambridge, UK.
Eltis, D. (Ed.), 2008. The Trans-Atlantic Slave Trade: A Database on CD-ROM. Cambridge University Press, Cambridge, UK.
Finley, M.I., 1987. Classical Slavery. Frank Cass, London.
Fogel, R.W., Engerman, S.L., 1974. Time on the Cross; The Economics of American Negro Slavery. Little, Brown, Boston.
Genovese, E.D., 1974. Roll, Jordan, Roll: The World the Slaves Made, first ed. Pantheon Books, New York.
Gutman, H.G., 1976. The Black Family in Slavery and Freedom, 1750–1925, first ed. New York Pantheon Books, New York.
Herskovits, M.J., 1941. The Myth of the Negro Past. London Harper & Brothers, New York.
Lovejoy, P.E., 1983. Transformations in Slavery: A History of Slavery in Africa. Cambridge University Press, Cambridge, UK.
Manning, P., 1990. Slavery and African Life. Occidental, Oriental, and African Slave Trades. Cambridge University Press, Cambridge, UK.
Meillassoux, C., 1991. The Anthropology of Slavery: The Womb of Iron and Gold. University of Chicago Press, Chicago.
Morgan, P., 1998. Slave Counterpoint: Black Culture in the Eighteenth-Century Chesapeake and Lowcountry. University of North Carolina Press, Chapel Hill.
Patterson, O., 1982. Slavery and Social Death. A Comparative Study. Harvard University Press, Cambridge, MA.
Phillips, U.B., 1918. American Negro Slavery: A Survey of the Supply, Employment and Control of Negro Labor as Determined by the Plantation Regime. D Appleton, New York.
Rediker, M., 2007. The Slave Ship: A Human History. Penguin, New York.
Rodney, W., 1982. African slavery and other forms of social oppression on the upper Guinea coast in the context of the African slave trade. In: Inikori, J. (Ed.), Forced Migration. The Impact of the Export Slave Trade on African Societies. Africana, New York.
Stampp, K.M., 1956. The Peculiar Institution: Slavery in the Ante-Bellum South, first ed. Knopf, New York.
Sweet, J., 2004. Recreating African Culture, Kinship, and Religion in the African-Portuguese World, 1441–1770. University of North Carolina Press, Chapel Hill.
Thornton, J., 1992. Africa and Africans in the Making of the Atlantic World, 1400–1680. Cambridge University Press, Cambridge, UK.
Vaughan, M., 2005. Creating the Creole Island: Slavery in Eighteenth Century Mauritius. Duke University Press, Durham.
Williams, E.E., 1944. Capitalism and Slavery, first ed. University of North California Press, Chapel Hill, NC.

Sleep and Health

Angelika Schlarb, Faculty of Psychology and Sports Science, University of Bielefeld, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sleep is a vital phenomenon, which fulfills essential functions in human development and affects life as a whole. Sleep-related problems are, therefore, important risk factors for ill-health and affect health at all ages. Factors that affect and are associated with sleep and sleep disturbances are, for example, short sleep, demographic characteristics (age, partnership status, or problems), health conditions (physical activity, obesity), stress, mental disorders, environmental factors, and quality of life. In many cases the reciprocal association between these factors and disturbed sleep makes it difficult to establish cause–effect relationships. Only longitudinal studies will make it possible to shed light on the processes linking sleep and health.

In terms of health behaviors, nutrition, exercise, or stress are widely discussed, whereas sleep in its fundamental role for health is often underestimated. Nevertheless, sleep is vital to physical and mental health, although the exact nature of the relationship between sleep and health remains elusive. Many factors affect the presumed causal link between sleep and health, which are likely to be reciprocal. Several risk factors for sleep disturbances and health impairments, however, have been established. One important first question to ask is what constitutes healthy sleep. Studies (e.g.,National Sleep Foundation) show that in adults, sleep duration should optimally vary between 6.5–9 h of sleep within a 24-h period (e.g., National Sleep Foundation), but it is important to mention that individual needs for sleep may vary substantially. Falling asleep within 15–20 minutes of lying down to sleep and to not have severe sleep interruptions is also considered to characterize healthy sleep, as is the absence of snoring, pauses in breathing, or other nighttime behaviors. Finally, one should feel rested when waking up and be able to be productive throughout the waking hours (National Sleep Foundation).

Variants of Sleep

Short Sleep

There is evidence from experimental studies linking short sleep with impaired appetite control and glucose regulation (Spiegel et al., 1999). In addition, a range of clinical studies have shown that decreased sleep is a risk factor for obesity and diabetes (Knutson and Van Cauter, 2008) and that it affects neuroendocrine stress systems (Meerlo et al., 2008). Decreased sleep duration is associated with poor performance (Van Dongen et al., 2003) and falling asleep at work (Ohayon et al., 1997). Whilst approximately 7 h of sleep is average in adults, there is a wide range in terms of individual sleep duration needed (Tucker et al., 2007). One may ask if it is sleep duration or quality of sleep that matters. When sleeping an average of 7 h a night, sleep quality is better related to health, affect balance, life satisfaction, and feelings of tension, depression, anger, fatigue, and confusion than average sleep quantity. Sleepiness has been shown to be more related to average sleep quality than sleep quantity (Tucker et al., 2007). In line with this, Tamakoshi and Ohno (2004) report that there is no epidemiological evidence that people who sleep 6 h per day and report no daytime sleepiness are at risk for poor health. Although short sleep duration or high variation of sleep has been linked to mortality, Kurina and colleagues (Kurina et al., 2013) found that the available literature does not provide consistent results supporting an association between self-reported sleep duration and mortality. Rather, fragmented sleep, for example caused by sleep apnea, decreases sleep quality and increases sleepiness (Marin et al., 2005).

Influencing Factors

Demographic Characteristics

Demographic characteristics (e.g., age, race, education, occupation, marital status) are also influences and are associated with sleep conditions and sleep problems. The next chapter will focus on age and partnership.

Age

Some research has been initiated covering a range of different age groups. Results show that sleep problems are common at all ages. In the first years of life, sleep duration problems are more common; in school-aged children and adolescents sleep onset problems are often mentioned, whereas with age many sleep-related changes occur. Some studies found that sleep in older people is less deep; they often have difficulties falling asleep, wake up more frequently during the night, awaken earlier in the morning, or report not being able to sleep. Over the last decades it was thought that the elderly need less sleep because of reduced daily activities. Nevertheless, subsequent studies with older people show that nearly 70% of elderly adults reported at least one sleep complaint, and 40% had two or more complaints (Reid et al., 2006). Older adults often change their circadian rhythm and have a high prevalence of sleep apnea and periodic limb movements. Objective tests show that older people are sleepy during the day (Dement et al., 1982). Excessive daytime sleepiness was the best predictor of poor mental and physical health-related quality of life (Reid et al., 2006). Furthermore, older people are often physically impaired, which can in turn affect sleep and sleep

quality (Laugsand et al., 2013). In addition, loneliness, resulting for example, from the loss of a partner, is associated with lower sleep efficiency (Cacioppo et al., 2002).

Partnership

Various studies have shown that marital quality is linked with sleep behavior. Hale (2005) collected data from more than 7000 adults. The study included sleep log data during a 24-h period, showing that unmarried adults were significantly more likely to report shorter sleep duration, compared to married adults. Interestingly, singles also more frequently report long sleep durations. These results show that singles are at high risk to be at the extremes of sleep duration, i.e., to either get too little or too much sleep, both of which are risk factors for poor health (Hale, 2005). In addition to the amount of sleep, some studies also provide evidence of a link between insomnia and lower levels of marital happiness (e.g., Troxel et al., 2009). Grandner et al. (2010) analyzed data of approx. 160 000 adults and found a significant association between marital status and sleep difficulties, such that the highest likelihood of reporting sleep difficulties was found in unmarried groups. Men or women who were divorced or never married had the highest risks (Grandner et al. 2010). Most of these studies are cross-sectional, however. Nevertheless, their results are in line with an earlier longitudinal study, showing that higher marital satisfaction predicted fewer insomnia symptoms, even after controlling for baseline sleep symptoms (Prigerson et al., 1999). Furthermore, Troxel et al. (2010) found in their 8-year longitudinal study of more than 350 adult women, lower sleep quality and longer nighttime awakenings in widows, whereas unmarried women showed longer sleep-onset latency and longer nighttime awakenings than married women.

Health Conditions

In addition to demographic characteristics, health conditions are strongly associated with sleep and sleep disturbances. The next section will focus on physical activity and obesity.

Physical Activity

Physical activity has been associated with better sleep than every other activity. Physical activity or exercise is typically recommended for better sleep by sleep experts (e.g., Lavie, 1996). When people are asked what helps them most to sleep well, more than 30% list exercise as the most important habit or practice (Urponen et al., 1988). But how much exercise is good for better sleep? Exercisers report better sleep than nonexercisers, even though they have the same amount of sleep each night (about 7 h). The study from 2003 (National Sleep Foundation, 2003) showed that people who exercise more frequently per week report fewer insomnia symptoms, less difficulties falling or staying asleep during the night, and lower daytime sleepiness. Furthermore, they wake up more refreshed and rested than people exercising less than once a week. Moreover, in this study, a dose–response pattern was found. People who exercised a minimum of three times per week reported better sleep than those who reported exercising one to two times per week. These results of the National Sleep Foundation study are based on more than 1500 adults showing strong associations of exercising with better sleep. In addition, the study published in 2013 (National Sleep Foundation, 2013) showed that more than 75% of exercisers reported as having good sleep quality compared to 56% of nonexercisers.

Studies with objective measurements show a diverse range of results. While many studies (e.g., Loprinzi and Cardinal, 2011) found positive correlations between various sleep parameters and physical activity (e.g., shorter sleep-onset latency, better sleep quality, fewer insomnia symptoms), there are also reports of null findings (Youngstedt et al., 2003). But how much exercise is healthy? There is no conclusive answer as yet; some studies found negative associations between high levels of activity and sleep (Driver and Taylor, 2000), whereas others found no such associations (Brand et al., 2009). There is evidence that fitness and not physical activity per se is associated with good sleep (Gerber et al., 2010). In summary, there does seem to be an association between sleep and physical activity, but the available results are not clear-cut.

Obesity

The link between sleep and obesity constitutes a wide field of research. During the last decades, not only has the worldwide prevalence of obesity doubled (Finucane et al., 2011) but sleep duration has reduced and poor sleep quality has become a frequent complaint. Sleep loss, especially, has been shown to result in increased hunger and appetite (Morselli et al., 2010). Epidemiological and experimental studies show that short sleep duration and poor sleep quality are risk factors for the development of obesity. In sum, most of the conducted studies found a significant association between short sleep (less than 5 or 6 h per night) and an increased obesity risk (e.g., Buxton and Marcelli, 2010), whereas an additional amount of sleep is correlated with lower BMI scores (Cappuccio and Miller, 2010). Those results are found not only in adults but also in younger age groups. Not only does the amount of sleep play an important role, obstructive apnea (OSA) is another very important issue. Studies show that the prevalence of obesity-related comorbidities such as OSA has increased up to more than 70% in women and 90% in men (Sareli et al., 2011).

In addition to these conditions, pain, heart disease, hypertension, cancer, and diabetes also affect sleep in terms of its duration, architecture, or quality – either as a direct consequence of the disease, or its medication or both.

Stress

Daily hassles and work stress have been shown to be associated with bad sleep and symptoms of insomnia (Hapke et al., 2013). Stress may lead to decreased sleep quality and, therefore, increase the risk for insomnia (e.g., Åkerstedt et al., 2012), with hyperarousal as a cognitive and physiological component playing a fundamental role (Pigeon and Perlis, 2006). Various studies report associations of higher presleep arousal in bad sleepers than in good sleepers (Morin, 2003). Furthermore, experimental studies found an effect of hyperarousal on sleep-onset latency (Wuyts et al., 2012).

Mental Disorders

Mental disorders, particularly depression and anxiety, are highly comorbid with sleep disturbances, as results of cross-sectional

studies show (e.g., Johnson et al., 2006). To date, the mechanisms underlying this comorbidity are still not fully understood. On the one hand, various studies show that individuals with sleep disturbances, especially insomnia symptoms, are more likely to develop future mental disorders (e.g., Wong et al., 2004). On the other hand, existing psychopathology, especially depression and anxiety, can increase the risk for development of adult insomnia (Skapinakis et al., 2013). In addition to anxiety and depression, aggression, delinquent behavior, attention-deficit hyperactivity disorder, and alcohol dependence are currently discussed as being associated with reduced sleep or sleep disturbances, such as insomnia or nightmares (Clinkinbeard et al. 2011; Cortese et al., 2006; Brower, 2003; Johnson and Breslau, 2001). It is still unknown if insomnia is predicted by early psychopathology.

Suicidal Tendency

Sleep is also connected with suicidal ideation, although the exact nature of their association has not been fully explored. Nevertheless, short sleep seems to play a role in risky and suicidal behavior in adolescents (Yen et al., 2010). In addition, insomnia symptoms and nightmares have been found to be associated with suicidal ideation. The duration of insomnia symptoms and nightmares, especially, seems to be significantly associated with suicide risk, independent of current sleep problems (Nadorff et al., 2013). In another study (Nadorff et al., 2012), 81 older adult patients above the age of 65 years were asked about their sleep, symptoms of depression, and suicidal ideation. In contrast to previous studies, insomnia symptoms, but not nightmares, were significantly related to suicidal ideation. This relation was mediated by depressive symptoms.

Environmental Factors

Various environmental factors have been found to impact sleep and decrease sleep quality. Blue light, especially, disturbs sleep, and therefore TV- and computer-time during the late evening should be reduced. In addition, noise negatively affects sleep. Ambient noise level, for example, is an external stimulus that is still processed by the sleepers' sensory functions, despite nonconscious perception of their presence. Exposure to noise varies between traffic, neighborhood noises, aircraft noises, and others. Research concerning environmental noise and sleep cover various fields, such as different situations and environments (at home, intensive care, etc.). Due to noise, sleep-onset latency can be prolonged, night awakenings might occur, and sleep architecture might be disturbed. The consequences of these sleep disruptions and disturbances are as broad as described above (Muzet, 2007).

Quality of Life

Various studies have shown that insufficient sleep is associated with a broad range of health impairments and overall impairment in quality of life. Chronic poor sleep quality, especially, is strongly associated with impaired quality of life. In detail, these impairments seem to be similar to chronic conditions, such as congestive heart failure and major depressive disorder (Philip et al., 2006). This association seems to be stable over time. During pregnancy, sleep problems seem to be the most important factor for reduced quality of life (Da Costa et al., 2009), whereas the successful treatment of sleep disturbances leads to improved quality of life (Sasai et al., 2010).

As seen above, sleep-related problems affect all areas of life, including cognitive development and performance, emotional well-being, work and physical activities, suicidal tendencies, and general physical and mental well-being. However, one has to bear in mind that most of the evidence available is based on cross-sectional and not longitudinal data. In addition, we need to know if increasing sleep duration in the long term has a positive impact on health.

In summary, sleep seems to have a broad range of associations with health. Assessment of sleep should therefore be an important component of general medical and psychological care. Moreover, assessment of sleep in children, adolescents, and adults as well as in the elderly may provide a better understanding of the prevalence of impaired sleep and its diverse implications on these populations in daily functioning.

See also: Chronic Illness, Psychosocial Coping with; Environmental Stress and Health; Mental Health; Physical Activity and Health; Respiratory Disorders: Psychosocial Aspects; Stress, Coping and Health; Weight Regulation.

Bibliography

Anic, G.M., Titus-Ernstoff, L., Newcomb, P.A., et al., 2010. Sleep duration and obesity in a population-based study. Sleep Medicine 11, 447–451.

Åkerstedt, T., Orsinic, N., Petersen, H., Axelson, J., Lekandera, M., Kecklunda, G., 2012. Predicting sleep quality from stress and prior sleep – a study of day-to-day covariation across six weeks. Sleep Medicine 13, 674–679.

Brand, S., Beck, J., Gerber, M., Hatzinger, M., Holsboer-Trachsler, E., 2009. "Football is good for your sleep": favorable sleep patterns and psychological function of adolescent male intense football players compared to controls. Journal of Health Psychology 14, 1144–1155.

Brower, K.J., 2003. Insomnia, alcoholism and relapse. Sleep Medicine Reviews 7, 523–539.

Buxton, O.M., Marcelli, E., 2010. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the United States. Social Science & Medicine 71, 1027–1036.

Buysse, D.J., Angst, J., Gamma, A., Adjdacic, V., Eich, D., Rossler, W., 2008. Prevalence, course, and comorbidity of insomnia and depression in young adults. Sleep 31, 473–480.

Cacioppo, J.T., Hawkley, L.C., Berntson, G.G., Ernst, J.M., Gibbs, A.C., Stickgold, R., Hobson, J.A., 2002. Do lonely days invade the nights? Potential social modulation of sleep efficiency. Psychological Science 13, 384–387.

Cappuccio, F., Miller, M., 2010. The epidemiology of sleep and cardiovascular risk and disease. In: Cappuccio, F., Miller, M., Lockley, S. (Eds.), Sleep, Health and Society: From Aetiology to Public Health. Oxford University Press, Oxford, pp. 83–110.

Clinkinbeard, S.S., Simi, P., Evans, M.K., Anderson, A.L., 2011. Sleep and delinquency: does the amount of sleep matter? Journal of Youth and Adolescence 40, 6–30.

Cortese, S., Konofal, E., Yateman, N., Mouren, M.-C., Lecendreux, M., 2006. Sleep and alertness in children with attention-deficit/hyperactivity disorder: a systematic review of the literature. Sleep 29, 504–511.

Da Costa, D., Drista, M., Verreault, N., Balaa, C., Kudzman, J., Khalifé, S., 2009. Sleep problems and depressed mood negatively impact health-related quality of life during pregnancy. Archives of Women's Health 13, 249–257.

Dement, W.C., Seidel, W., Carskadon, M.A., 1982. Daytime alertness, insomnia and benzodiazepines. Sleep 5, 28–45.

Dew, M.A., Hoch, C.C., Buysse, D.J., Monk, T.H., Begley, A.E., Houck, P.R., et al., 2003. Healthy older adults'sleep precits all-cause mortality at 4 to 19 years of follow-up. Psychosomatic Medicine 65, 63–73.

Driver, H.S., Taylor, S.R., 2000. Exercise and sleep. Sleep & Medicine Review 4, 387–402.

Finucane, M.M., Stevens, G.A., Cowan, M.J., et al., 2011. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. Lancet 377, 557–567.

Gerber, M., Brand, S., Holsbeoer-Trachsler, E., Pühse, U., 2010. Fitness and exercise as correlates of sleep complaints: is it all in our minds? Medicine & Science in Sports & Exercise 42, 893–901.

Gillespie, N.A., Gehrman, P., Byrne, E.M., Kendler, K.S., Heath, A.C., Martin, N.G., 2012. Modeling the direction of causation between cross-sectional measures of disrupted sleep, anxiety and depression in a sample of male and female Australian twins. Journal of Sleep Research 21, 675–683.

Grandner, M.A., Patel, N.P., Gehrman, P.R., Xie, D., Sha, D., Weaver, T., Gooneratne, N., 2010. Who gets the best sleep? Ethnic and socioeconomic factors related to sleep complaints. Sleep Medicine 11, 470–478.

Hale, L., 2005. Who has time to sleep? Journal of Public Health 27, 205–211.

Hapke, U., Maske, U., Scheidt-Nave, C., Bode, L., Schlack, R., Busch, M., 2013. Chronischer Stress bei Erwachsenen in Deutschland. Bundesgesundheitsblatt 56, 749–754.

Jansson, M., Linton, S.J., 2006. The role of anxiety and depression in the development of insomnia: cross-sectional and prospective analyses. Psychological Health 21, 383–397.

Johnson, E.O., Breslau, N., 2001. Sleep problems and substance use in adolescence. Drug and Alcohol Dependence 64, 1–7.

Johnson, E.O., Roth, T., Breslau, N., 2006. The association of insomnia with anxiety disorders and depression: exploration of the direction of risk. Journal of Psychiatric Research 40, 700–708.

Knutson, L., Van Cauter, E., 2008. Associations between sleep loss and increased risk of obesity and diabetes. Molecular and Biophysical Mechanisms of Arousal. Alertness and Attention 1129, 287–304.

Kurina, L.M., McClintock, M.K., Chen, J., Waite, L., Thisted, R.A., Lauderdale, D.S., 2013. Sleep duration and all-cause mortality: a critical review of measurement and associations. Annals of Epidemiology 23, 361–370.

Laugsand, L.E., Strand, L.B., Platou, C., Vatten, L.J., Janszky, I., 2013. Insomnia and the risk of incident heart failure: a population study. European Heart Journal. http://dx.doi.org/10.1093/eurheartj/eht019.

Lavie, P., 1996. The Enchanted World of Sleep. Yale University Press, New Haven, CT.

Leproult, R., Van Cauter, E., 2010. Role of sleep and sleep loss in hormonal release and metabolism. Endocrine Development 17, 11–21.

Loprinzi, P.D., Cardinal, B.J., 2011. Association between objectively-measured physical activity and sleep, NHANS 2005–2006. Mental Health and Physical Activity 4, 65–69.

Magee, C.A., Caputi, P., Iverson, D.C., 2010. Is sleep duration associated with obesity in older Australian adults? Journal of Aging and Health 22, 1235–1255.

Magee, C.A., Caputi, P., Iverson, D.C., 2011. Short sleep mediates the association between long work hours and increased body mass index. Journal of Behavioral Medicine 34, 83–91.

Marin, J.M., Carrizo, S.J., Vicente, E., Agusti, A.G., 2005. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypnopnoea with or without treatment with continuous obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observationa study. Lancet 365, 1046–1053.

Meerlo, P., Sgofio, A., Suchecki, D., 2008. Restricted and disrupted sleep: effects on autonomic function, neuroendocrine stress systems and stress respronsitivity. Sleep Medicine Reviews 12, 197–210.

Morin, C.M., 2003. Role of stress, arousal, and coping skills in primary insomnia. Psychosomatic Medicine 65, 259–267.

Morselli, L., Leproult, R., Balbo, M., Spiegel, K., 2010. Role of sleep duration in the regulation of glucose metabolism and appetite. Best Practice & Research Clinical Endocrinology & Metabolism 24, 687–702.

Muzet, A., 2007. Environmental noise, sleep and health. Sleep Medicine Reviews 11, 135–142.

Nadorff, M.R., Fiske, A., Sperry, J.A., Petts, R., Gregg, J.J., 2012. Insomnia symptoms, nightmares, and suicidal ideation in older adults. The Journals of Gerontology. http://dx.doi.org/10.1093/geronb/gbs061.

Nadorff, M.R., Nazem, S., Fiske, A., 2013. Insomnia symptoms, nightmares, and suicide risk: duration of sleep disturbance matters. Suicide and Life-Threat Behavior 43, 139–149.

National Sleep Foundation, 2003. Sleep in America Poll. National Sleep Foundation, Washington, DC.

National Sleep Foundation, 2013. www.sleepfoundation.org.

Ohayon, M.M., Caulet, M., Philip, P., Guilleminault, C., Priest, R.G., 1997. How sleep and mental disorders are related to complaints of daytime sleepiness. Archives of Internal Medicine 157, 2645–2652.

Patel, S.R., Hu, F.B., 2008. Short sleep duration and weight gain: a systematic review. Obesity (Silver Spring) 16, 643–653.

Philip, P., Leger, D., Taillard, J., Quera-Salva, M.A., Niedhammer, I., Mosqueda, J.G., Bioulac, B., Gerard, D., 2006. Insomniac complaints interfere with quality of life but not with absenteeism: respective role of depressive and organic comorbidity. Sleep Medicine 7, 585–591.

Pigeon, W.R., Perlis, M.L., 2006. Sleep homeostasis in primary insomnia. Sleep Medicine Review 10, 247–254.

Prigerson, H.G., Maciejewski, P.K., Rosenheck, R.A., 1999. The effects of marital dissolution and marital quality on health and health service use among women. Medical Care 37, 858–873.

Reid, K.J., Martinovich, Z., Finkel, S., Statsinger, J., Golden, R., Harter, K., Zee, P.C., 2006. Sleep: a marker of physical and mental health in the elderly. American Journal of Geriatric Psychiatry 14, 860–866.

Sareli, A.E., Cantor, C.R., Williams, N.N., et al., 2011. Obstructive sleep apnea in patients undergoing bariatric surgery: a tertiary center experience. Obesity Surgery 21, 316–327.

Sasai, T., Inoue, Y., Komada, Y., Nomura, T., Matsuura, M., Matsushima, E., 2010. Effects of insomnia and sleep medication on health-related quality of life. Sleep Medicine 11, 452–457.

Skapinakis, P., Rai, D., Anagnostopoulos, F., Harrison, S., Araya, R., Lewis, G., 2013. Sleep disturbances and depressive symptoms: an investigation of their longitudinal association in a representative sample of the UK general population. Psychological Medicine 43, 329–339.

Spiegel, K., Leproult, R., Van Cauter, E., 1999. Impact of sleep dept on metabiloc and endocrine function. Lancet 354, 1435–1439.

Tamakoshi, A., Ohno, Y., 2004. Self-reported sleep duration as a predictor of all-cause mortality: results from the JACC study, Japan. Sleep 27, 51–54.

Troxel, W.M., Buysse, D.J., Hall, M., Matthews, K.A., 2009. Marital happiness and sleep disturbances in a multi-ethnic sample of middle-aged women. Behavioral Sleep Medicine 7, 2–19.

Troxel, W.M., Buysse, D.J., Matthews, K.A., Kravitz, H.M., Bromberger, J.T., Sowers, M.F., Hall, M.H., 2010. Marital/cohabitation status and history in relation to sleep in midlife women. Sleep 33, 973–981.

Tucker, A.M., Dinges, D.F., Van Dongen, H.P., 2007. Trait interindividual differenences in the sleep physiology of healthy young adults. Journal of Sleep Research 16, 170–180.

Urponen, H., Vuori, I., Hasan, J., Partinen, M., 1988. Self-evaluation of factors promoting and disturbing sleep: an epidemiological survey in Finland. Social Science & Medicine 26, 443–450.

Van Dongen, H.P., Maislin, G., Mullington, J.M., Dinges, D.F., 2003. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 26, 117–126.

Wong, M.M., Brower, K.J., Fitzgerald, H.E., Zucker, R.A., 2004. Sleep problems in early childhood and early onset of alcohol and other drug use in adolescence. Alcoholism 28, 578–587.

Wuyts, J., De Valck, E., Vandekerckhove, M., Pattyn, N., Bulckaert, A., Berckmans, D., et al., 2012. The influence of pre-sleep cognitive arousal on sleep onset processes. International Journal of Psychophysiology 83, 8–15.

Yen, C., Kind, B.H., Tang, T., 2010. The association between short and long nocturnal sleep durations and risky behaviors and the moderating factors in Taiwanese adolescents. Psychiatry Research 179, 69–74.

Youngstedt, S.D., Perlis, M.L., O'Brien, P.M., Palmer, C.R., Smith, M.T., Orff, H.J., et al., 2003. No association of sleep with total daily physical activity in normal sleepers. Physiology & Behavior 78, 395–401.

Sleep and Memory

Médhi Gilson, Université Libre de Bruxelles, Brussels, Belgium; and FNRS, Belgium
Philippe Peigneux, Université Libre de Bruxelles, Brussels, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is now widely recognized that sleep plays a pivotal role in the regulation of cognitive functions, especially considering learning and memory processes. In this article, we review key findings and evidences regarding the hypothesis that sleep exerts a promoting effect on plastic processes subtending memory consolidation. Nowadays, experimental findings show that recently learned information not only continues being processed but also can be modulated during sleep, that novel associations can be created during sleep and that emotion plays a modulatory role in sleep-dependent memory consolidation processes.

Introduction

A quarter to a third of human life is spent sleeping, suggesting the utmost importance of this particular state of vigilance. Notwithstanding the empiric recognition of its importance for our welfare, the functions of sleep have long remained poorly understood. Even nowadays it remains a matter of debate. Amongst other nonexclusive functions, growing evidence suggests that sleep participates in brain plasticity and memory consolidation processes (Maquet et al., 2003). Already in 1924, Jenkins and Dallenbach noticed less forgetting for novel verbal information when learning was followed by a period of sleep than after the same amount of time spent awake. However, they ascribed this beneficial effect to a reduced amount of external interference during sleep time, hence a passive shielding contribution of sleep in the process of memory consolidation. Subsequent studies that will be discussed here fostered the stronger hypothesis that sleep actively participates in memory consolidation processes. In this article, we will review key findings and evidences regarding the so-called sleep for memory hypothesis, and highlight open and unsolved issues. Before dwelling into the topic of consolidation, however, sleep and memory concepts will be shortly defined.

Sleep and Memory: Definitions and Key Concepts

Not One, but Different Memories in the Long Term

Popular models of memory organization have conceptualized long-term memory as composed of several specialized subsystems. These systems can be characterized considering the specific features of the information to encode and/or the degree of consciousness needed to integrate or retrieve this information. In this framework, a common distinction is made between declarative and nondeclarative long-term memories.

Declarative memories are verbalizable and consciously available, i.e., the recipient of these memories is usually aware that the stored information exists and is being retrieved. Declarative memory itself encompasses episodic and semantic memory subsystems. Episodic memory stores self-experienced events and facts, and allows replacing memories in their original learning context (e.g., a vivid souvenir of our wedding party). It is mostly subtended by hippocampal activity. At variance, semantic memory refers to the general knowledge that we have about the world, but is usually not anymore associated with the original learning context (e.g., we know that a cow is herbivorous, but there is very little chance that we still remember exactly how and when we have learned this information). A more distributed medial temporal cortical network subtends semantic memory. Importantly, episodic and semantic memories are not fully segregated subsystems, as episodic memories may become semanticized over time or repetition (e.g., we may know having had a great pleasure eating jam at our grandmother's home as a child, but not having in memory a specific and vivid episode associated with this knowledge).

Conversely, nondeclarative memories are usually not easily verbalizable and can be reexpressed (but this is not mandatory) without conscious knowledge. Therefore, nondeclarative memories will be mostly expressed through performance in perceptive, motor, and/or cognitive tasks. Nondeclarative memory encompasses various forms subtended by partially distinct neuroanatomical pathways. Procedural memory is the memory for skills and habits, mostly supported by cortico-subcortical pathways linked to the striatum. Classical conditioning with creation of stimulus–stimulus or stimulus–response associations is associated with amygdala (emotional) and cerebellum (skeletal) responses. And finally, neocortical activity is associated with priming and perceptual learning.

Sleep Cycles and Alternating Sleep Phases

Likewise memory, sleep is not a unitary phenomenon, especially if considered from a neurophysiological perspective. In man, sleep is divided into two main alternating phases, namely nonrapid eye movement (NREM) followed by rapid eye movement (REM) sleep. Four to five 90-min cycles of NREM–REM alternation are usually observed during a normal night of sleep. NREM sleep is further divided into three stages reflecting the increasing depth of sleep: N1, N2, and N3. As compared to wakefulness, electroencephalographic (EEG) activity recorded at the scalp level starts slowing down in the N1 stage with a decrease in alpha rhythms (8–12 Hz) and appearance of theta rhythms (4–7 Hz). Theta rhythms persist throughout stage N2, which is characterized by the occurrence of phasic events

including k-complexes (single sharp waves of large amplitude in the EEG) and spindles (waxing and waning EEG oscillations having a duration of 0.5–3 s in the 12–16 Hz frequency range). Spindle activity presents low intraindividual but high interindividual variability. Stage N3 is also called slow wave sleep (SWS) due to the predominance of slow oscillating waves (0.5–4 Hz) of high amplitude reflecting the high synchrony of discharges over large neuronal populations. At variance during REM sleep, EEG activity is more similar to wakefulness (hence its other name: paradoxical sleep) with a higher-frequency desynchronized activity of low amplitude. Additionally, definite features of REM sleep are muscle atonia and bursts of REMs. The relative proportion of NREM and REM within a sleep cycle varies across a night of sleep due to the progressive dissipation of homeostatic sleep pressure (reflected in slow wave activity (SWA)) and the higher circadian propensity for REM sleep at the end of the night. Consequently, the first half of the night is proportionally richer in SWS–N3 whereas the second part of the night is richer in REM sleep, the N2 stage being present 50% of the time all over the night.

Mechanisms of Sleep-Dependent Memory Consolidation

Memory consolidation refers to the temporal process lasting from a few minutes to several years by which initially fragile memory traces undergo a series of transformations eventually leading to their strengthening into long-term memory stores. Memory consolidation and progressive modification of the acquired information are actually instances of brain plasticity, i.e., the capacity of the brain to adapt its function and structure over time within certain boundaries to accommodate novel experience. In this perspective, reorganization of cortical networks, disinhibition of neuronal assemblies, modification and remodeling of synaptic connections should be the reflections of the processes of memory formation and consolidation. In this respect, experimental evidence suggests that the neurophysiological conditions of sleep are favorable to the consolidation of novel memories and their durable inscription in long-term memory stores, with two main theoretical models.

According to the model of the hippocampal–neocortical dialogue (Buzsáki, 1996; Hasselmo, 1999), novel information is stored during wakefulness within the hippocampus associated with theta–gamma oscillations. Consolidation of hippocampal information takes place during SWS in association with sharp-wave ripple (SWR) activity. Slow oscillations alternating hyperpolarization and depolarization phases trigger the repeated reactivation of hippocampal information through the joint activation of thalamic spindles and hippocampal SWRs, eventually transferring the stored representations toward neocortical stores during the depolarization phase of the slow oscillation (Figure 1). In this framework, sleep spindles play an important role in memory consolidation. If this model was associated with SWS and declarative memory, a less specific variant called the neuronal replay hypothesis proposes that cortical or subcortical activation patterns associated with a novel learning are reactivated during subsequent sleep stages (either NREM or REM stages), eventually leading to the replay of the learned information and its gradual integration and/or transfer into long-term memory stores.

The other main model is the synaptic homeostasis theory proposing that memory consolidation is a local use-dependent synaptic downscaling phenomenon that takes place during sleep as a secondary consequence of the growth of synaptic connections associated with novel learning at wake (Tononi and Cirelli, 2006). During wakefulness, the creation of novel memory traces is associated with locally increased synaptic potentiation and elevated energy and space costs, eventually leading to the saturation of the learning-related neuronal network. In this framework, homeostatically regulated slow oscillations during SWS would offer optimal conditions to downscale synaptic strength to their baseline level, thus restoring synaptic plasticity, while preserving the necessary differentiation between learning-related potentiated and other nonpotentiated synapses. Hence slow oscillations would refine the synaptic weight and indirectly improve the signal-to-noise ratio in learning-related neuronal ensembles (Figure 2).

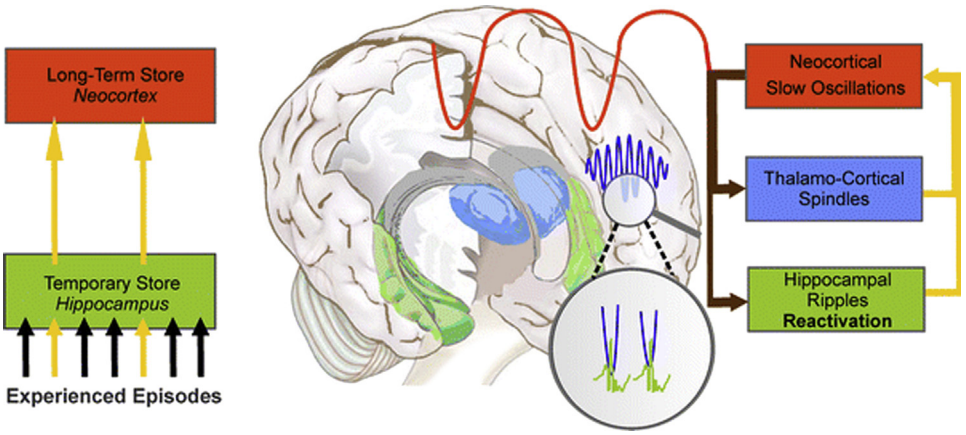

Figure 1 Hippocampo–neocortical dialogue. Hippocampo–neocortical interactions during SWS are under the control of top-down slow cortical oscillations. Depolarization phases in the slow waves coordinate the repeated reactivation of the representation stored in the hippocampus together with SWR and sleep spindles, leading to the formation of 'spindles-ripples' events eventually leading to the transfer of the learned information into neocortical stores. Reprinted with permission from Born, J., Wilhelm, I., 2012. System consolidation of memory during sleep. Psychological Research 76, 192–203.

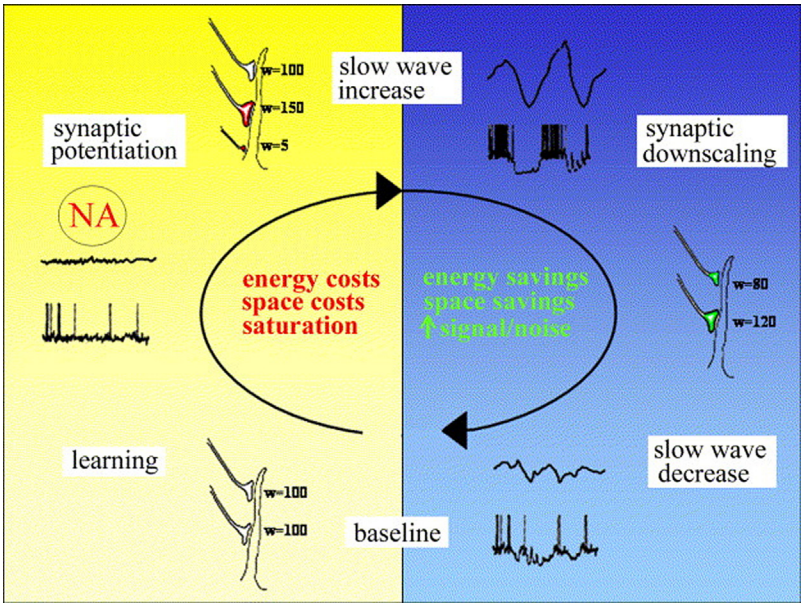

Figure 2 Synaptic Homeostasis Theory. During wakefulness, elevated noradrenaline levels and desynchronized neuronal activity favor information storage. Learning induces synaptic potentiation, reinforcing certain synapses, and creating new ones, a process that costs energy and space and eventually saturates the ability to learn new information. During NREM sleep, homeostatically regulated slow wave oscillations and low noradrenaline levels prevent further potentiation processes. The repeated alternation of depolarization and hyperpolarization phases in slow oscillations progressively downscales synaptic weights until baseline level while keeping the traces of previous experiences. Reprinted with permission from Tononi, G., Cirelli, C., 2006. Sleep function and synaptic homeostasis. Sleep Medicine Reviews 10, 49–62.

Sleep Stages and Memory Systems: A Complex Interaction

Marked neurophysiological differences between sleep stages coupled with the existence of distinct systems in long-term memory entails that interactions are necessarily complex between sleep and memory processes. Accordingly none of the models described above can satisfactorily explain all of reported results in the field, and the search for a single, unitary model of sleep-dependent memory consolidation is not yet and may never be achieved. One possible approach to conceptualize these complex interactions is to consider that NREM and REM sleep entertain privileged relationships with distinct memory systems. In this respect, the dual-process hypothesis postulates that SWS supports the consolidation of declarative and spatial memories whereas REM sleep facilitates the consolidation of procedural memories. Alternatively, the sequential hypothesis proposes that memory consolidation occurs through sequential processes taking place during NREM and then REM sleep. An instance of this theory proposes that the recently learned information is made more resistant to interference during NREM sleep then consolidated and integrated into preexisting knowledge in long-term memory stores during REM sleep.

Arguments in favor of the dual-process hypothesis are provided by studies having compared the effect of postlearning sleep in the first part of the night (SWS predominance) with the effect of postlearning sleep in the second part of the night (REM sleep predominance). These studies evidenced a sleep-dependent retention of declarative spatial and verbal memories after early but not after late sleep. Conversely, sleep-dependent effects on priming and procedural learning were found after late but not after early sleep. On the other hand, sleep fragmentation preserving the overall structure and proportion of sleep stages deteriorates sleep-dependent consolidation processes, in line with the sequential hypothesis. Notwithstanding, other studies have reported outcomes that are difficult to integrate in these models. For instance, nondeclarative visual perceptual and motor tasks have been found to benefit from posttraining SWS and to be associated with spindles activity, whereas consolidation of declarative memories may benefit from posttraining REM sleep especially when the material comprises an emotional dimension. Hence it suggests that besides memory systems, the contribution of particular sleep stages in consolidation processes is also contingent upon the specific features of the material to be consolidated. Notwithstanding these reservations, we will present in separate sections the studies having investigated sleep-dependent consolidation for declarative and non-declarative memory for the sake of legibility and organization of the reading material.

From a methodological standpoint, the relationships between sleep and memory consolidation processes have been investigated using three main approaches (Peigneux et al., 2001; Rauchs et al., 2005). The first and most widely used paradigm is sleep deprivation (vs regular sleep) under the assumption that preventing sleep after learning should have a deleterious impact on memory consolidation processes and/or the underlying neural activity during memory retrieval. Sleep deprivation can be complete (whole night deprivation) or partial if a specific stage of sleep is under scrutiny. Partial sleep deprivation can be conducted by

awakening the participant at each occurrence of the targeted sleep stage (e.g., REM sleep) or using the first vs second half of the night (i.e., early vs late sleep) technique described above. The recall session task then takes place on the next day or later on after recovery of the sleep debt, and memory performance is compared between sleep and sleep-deprived conditions. The second line of research aimed at assessing how learning impacts on neural activity patterns during post-learning sleep, based on the assumption that learning-dependent changes in neurophysiological sleep patterns should reflect ongoing memory consolidation processes. Finally, paradigms using within-sleep stimulations have gained popularity in the recent years. Here, the impact of within-sleep stimulations at nonawakening thresholds was investigated, either during postlearning sleep to modify sleep patterns and boost overnight memory performance or to create novel memories during a sleep episode.

Sleep and the Consolidation of Declarative Memories

Interactions between sleep and declarative memory systems have been classically probed using hippocampus-dependent memory tasks such as learning of sentences, word pairs, object locations, faces, or spatial navigation. Early research using partial sleep deprivation suggests that retention of declarative material (e.g., word pairs) benefit more from sleep in the first half of night (i.e., richer in SWS) than from sleep in the second half of night (i.e., richer in REM sleep), a finding consistently replicated in subsequent studies (Diekelmann et al., 2009). Functional neuroimaging studies also show that sleep-dependent consolidation of declarative memories is accompanied by a functional reorganization in underlying brain activity at retrieval. For instance, delayed retrieval of learned word pairs or pictures is associated with increased hippocampal activity hours to a few days after the postlearning sleep (vs wakefulness) episode, a differential activity that dissipates over 3–6 months to be replaced by increased retrieval-related activity in medial prefrontal areas in participants having slept after the learning episode. Additionally, other functional magnetic resonance imaging (fMRI) studies using spatial navigation tasks showed that even in the absence of observable changes at the behavioral level, sleep after learning eventually promotes a shift from hippocampus- to striatum-based brain activity subtending the retrieval of these spatial memories, suggesting a sleep-dependent evolution toward less costly and more automated striatum-based navigation strategies (Orban et al., 2006).

Buzsáki's (1996) hypothesis of the hippocampo–cortical dialogue during sleep assumed that the progressive transfer of declarative memory-related activity from hippocampal to neocortical areas mostly takes place during NREM sleep. Accordingly, several animal studies (Poe et al., 2010) showed that neurons firing when exploring a new spatial environment are reactivated in the same temporal order (although not always at the same rate) during subsequent NREM sleep. Others also reported learning-related activations during REM sleep. Similarly, reactivation of learning-related hippocampal activity during SWS associated with increased memory consolidation was found in humans both for spatial (Peigneux et al., 2004) and verbal material. Taken together, these results suggest an association between neuronal reactivation during SWS and the offline consolidation of declarative memories.

The acquisition and sleep-dependent consolidation of semantic memories has received less attention than episodic memory consolidation, which may be linked to some extent to the difficulty segregating novel semantic memories from their episodic learning component. Notwithstanding, studies found associations between the acquisition of novel semantic material (e.g., a six-weeks French immersion course or Morse language learning) and increased REM sleep parameters, as well as associations between REM sleep inertia and semantic priming effects in narcoleptic patients. Besides, preserved detection of semantic discordance is possible during various sleep stages (Perrin, 2004). On the other hand, sleep deprivation studies having investigated the sleep-dependent processing of false memories upon the learning of lists of semantically related words have reported contradictory results both suggesting that sleep may favor cognitive abstraction processes typical of semantization, but also that these effects are largely contingent upon the quality of the encoding processes during learning (Lewis and Durrant, 2011). Noticeably, a main caveat in studies focused on the sleep-dependent processing of novel semantic information is that the time elapsed between learning and retrieval phases might be too short to evidence consistent effects, considering that the transfer of information from hippocampus-dependent episodic memory stores to neocortical semantic memory representations is a slow and gradual process that can take months or years to be achieved.

Sleep and the Consolidation of Nondeclarative Memories

The relationships between sleep and consolidation of nondeclarative memories have been mostly assessed using simple perceptual, motor, and visuomotor learning paradigms (Peigneux and Smith, 2010).

At the perceptual level, a prominent role of sleep in the consolidation of visual discrimination abilities has been repeatedly demonstrated. Although initial findings highlighted a predominant role of REM sleep, subsequent studies emphasized a main role of stage 2 sleep (N2) and SWS in posttraining consolidation, those views being reconciled by the demonstration that improvement to the task is associated with both SWS in the first sleep cycle and REM in the last sleep cycle of the night. Complementary studies also showed that repeated practice sessions within a same day actually deteriorate discrimination performance, suggesting neuronal saturation in primary visual cortex projection areas, unless a sleep episode (nap) is allowed. Furthermore, these naps studies found that a 30-min nap stops performance deterioration whereas a 60-min nap brings performance back to its original level and a 90-min nap leads to a performance improvement. As short naps usually only comprise NREM sleep periods and REM sleep can be additionally present in longer naps only, it further supports the hypothesis of a complementary role of sleep stages in the consolidation of visual discrimination abilities, partially in line with the sequential hypothesis. However, sleep-dependent improvements in performance are not consistently present in all perceptual modalities. For instance, whereas

several studies suggested an improvement in auditory discrimination abilities after sleep, others found equivalent performance after a period of wake.

The finger-to-thumb opposition and its keyboard equivalent the finger-tapping task have been widely used to study sleep-dependent motor learning. In these tasks, participants have to repeat as fast and accurately as possible a known, simple sequence of five finger movements. Most studies found a stabilization of performance over offline periods of wakefulness, whereas it increased after postlearning sleep, suggesting a consolidating role of sleep for learned motor representations. Postlearning sleep spindles activity was mostly found to be modified after learning, and to be correlated with performance improvement in these studies. Also, performance $\pm$ 30 min after the end of learning, but not later on during the day, is significantly increased and predictive of the next day's amplitude of improvement, suggesting the functional relevance of early transient phenomena of motor plasticity for sleep-dependent consolidation processes (Hotermans et al., 2006). Posttraining sleep changes have been also associated with more complex forms of motor learning, e.g., trampolining and dancing, or more recently three-ball cascade juggling, again leading to increased posttraining sleep spindle activity. In the motor tasks discussed above, an element to consider is the fact the learner is well aware that he is learning a new motor sequence or skill. Hence learning is relatively explicit, even if verbalizing is less helpful than repeated practice to improve motor performance. Additionally, presence of interactions between hippocampal and striatal activity during learning in a closely similar oculomotor paradigm was associated with subsequent sleep-related improvement in performance, suggesting that tasks in which the hippocampus is involved may to some extent benefit from NREM (stage 2) sleep consolidation processes. This interpretation is consistent with neuroimaging studies having shown the reactivation of learning-related activity and consolidation during posttraining REM sleep after learning higher order perceptual-motor cognitive skills in a probabilistic sequence learning task (i.e., a paradigm of implicit sequence learning) (Maquet et al., 2000). These data support the hypothesis that REM sleep is involved in the reprocessing and optimization of the high-order information contained in the material to be learned, as opposed to its motor component alone. Notwithstanding, other studies using sequence learning tasks have raised several issue. Indeed, some authors found using simpler sequential regularities that sleep is beneficial for sequence learning only when acquisition is explicit, with time alone being sufficient for the consolidation of implicitly acquired sequences. Since studies claiming a sleep-dependency exclusively for explicit sequential material have used deterministic, repeated sequences, and REM sleep-related reactivation studies have used much complex and abstract probabilistic sequences, it is possible that sleep (and especially REM sleep) mostly supports the consolidation of implicitly acquired complex relationship.

Besides sequence learning, other forms of procedural learning have been investigated. In the pursuit rotor task, participants use a cursor or the computer's mouse to track a pseudorandomly moving target (with hidden regularities). Performance to the task deteriorates both after total and late (i.e., filled with REM and N2 sleep) sleep deprivation but not after selective deprivation of REM sleep. Since performance improvement in the task is also associated with posttraining spindle activity, it suggests that offline consolidation is mostly subtended by stage 2 (N2) sleep. However, the initial skill level of participants at learning also determines which stage of sleep is modified after learning, as it has been shown that highly skilled participants present an increase in stage 2 spindles density after learning, which is correlated to performance at retest one week later, whereas performance of less skilled participants is related to changes in REM sleep density. At variance, performance on mirror-tracing tasks is better after late than early sleep and associated with increased REM sleep, but sleep stage 2 fast spindles activity (13–15 Hz) also correlates with skill improvement in this task. Altogether, these results suggest that postlearning modifications of sleep architecture are also contingent upon the skill levels achieved by participants at the end of learning, and that both stage 2 and REM sleep stages contribute to the consolidation of motor skills. Additionally, all motor tasks do not benefit to the same extent from posttraining sleep than motor sequence-based tasks do. Indeed, whereas a sleep-dependent improvement in performance, correlated with increased SWA, was observed using a rotation adaptation task in which subjects learned to adapt a reaching movement to imposed deviations (Huber et al., 2004), other studies merely found time- but not sleep-dependent offline changes in performance using the same task. On the whole, currently available studies indicate that the contributing role of sleep in the offline consolidation of motor abilities may be contingent upon several factors, including the stage of memory formation being considered, the degree of consciousness involved, and the parameters of the task.

Tickling Memories: The Contribution of Within-Sleep Stimulations

Within-sleep stimulation paradigms have been increasingly popular in the last decade. Indeed, this approach has proven particularly well suited to investigate current hypotheses about the neurophysiological mechanisms subtending memory consolidation processes. As stated above, consolidation of declarative, hippocampal-dependent memories has been mostly associated with NREM sleep. In this respect, it is assumed that slow oscillations in the delta frequency range (0.5–4 Hz) trigger the neurophysiological events allowing the transfer and integration of information in long-term memory stores. To test this hypothesis, transcranial Direct Current Stimulation (tDCS) that allows modifying neuronal excitability has been applied at the frequency of slow oscillations (0.75 Hz) during NREM sleep after learning. Results show increased retention of word pairs when tDCS was applied during sleep than after a normal night of sleep, whereas consolidation of a mirror-tracing (procedural memory) task was not affected by tDCS (Marshall et al., 2006). tDCS at other frequencies was not effective, suggesting the specificity of slow oscillations for memory consolidation of declarative memories. Using another approach, it was found that nonawakening presentation during SWS of an odor associated with the material to be

learned (i.e., card pairs in the game 'Memory') actually enhances retrieval performance on the next day, as compared to a normal, undisturbed night of sleep. Again, performance on a procedural memory task (i.e., the finger-tapping task) is not improved in such conditions. Additionally, fMRI recordings showed that odor cuing during SWS after learning triggers a hippocampal response associated with performance improvement in the declarative memory task. A complementary study using a nap paradigm demonstrated that postlearning odor cuing during a 40-min nap actually enhances memories to the level achieved over an undisturbed 90-min nap, whereas an undisturbed 40-min nap was not sufficient to enhance memory performance. Similar results have been obtained using auditory cues rather than odors during postlearning SWS in the context of an object location task, indicating the robustness of the observed effects. Finally in another study, participants learned to produce two melodies in time with moving visual symbols. Only one of the two melodies was then presented during the postlearning nap. Results showed higher performance for the melody represented during sleep than for the other one, suggesting that learning-related auditory stimulation during sleep can effectively and selectively trigger memory consolidation processes. An unsolved question is whether within-sleep stimulations could be used and beneficial in pathological or rehabilitation conditions, especially when considering emotional material.

Another open issue is whether the learning of novel information during sleep is possible. Although the topic was explored already a long time ago under the term of 'hypnopedia,' the few reported results have been far from convincing especially in man. On the other hand, successful conditioning of physiological responses during sleep has been reported in animals (Hennevin et al., 2007). Also, it is established that external stimulations continue to be processed during sleep at various levels of complexity. For instance, auditory event-related potentials are elicited after hearing one's own name both during wake and sleep, and evoked responses upon semantically deviant items are observed in all stages of sleep (Perrin, 2004). Hence many studies demonstrated the ability to process external stimulation while asleep, but failed to demonstrate the learning of new information during sleep. In a recent partial-reinforcement trace conditioning study, however (Arzi et al., 2012), participants were presented pleasant and unpleasant odors associated with two different tones during SWS and REM sleep, while monitoring the automatic sniff volume response. Results show that subsequent presentation of the tones alone during sleep, without the corresponding odors, induced a weak or a strong sniff volume of the response depending on the pleasantness of the odor previously associated with the tones, suggesting that participants successfully learned the odor–tone associations during sleep. Furthermore, conditioning was transferred during wakefulness (but only for stimuli conditioned during NREM sleep), even though the participants were not aware of the association. Although these results are stimulating, it remains to be tested whether the learning of more complex forms such as stimulus–stimulus association is possible during human sleep.

Emotion: A Modulating Factor in Sleep-Dependent Memory Consolidation

In everyday life, learning new information, facts, and events is often embedded within an emotional context. Each of us possesses a series of reminiscing states of mind and emotions associated with a relatively precise recollection of specific events, from the first steps of one's own child to the discovery of the World Trade Center tragedy. Emotion is a peculiar component of episodic memory involving additional neuroanatomical components (e.g., the amygdala complex) that are also activated during REM sleep. In this respect, the sleep-dependent processes subtending the consolidation of emotional memory have received increasing attention. The relationships between sleep and emotion-related memories have been assessed using intrinsically emotional material such as texts, word pairs, pictures, faces, or scenes, alternatively using neutral material associated with an emotional background.

There is a known advantage in the recall of emotional vs neutral material at wake. During postlearning sleep, late (REM-dominant) but not early (SWS-dominant) sleep also benefits more to the retention of emotional than of neutral texts, even when retested up to 4 years later. Conversely, the retrieval of emotional material is disrupted after a night of sleep deprivation. Central and peripheral components of an emotionally negative visual scene are not affected to the same extent by sleep deprivation, suggesting that the relevance of the emotional information to be processed is a selection factor in sleep-dependent consolidation processes. Indeed, whereas memory deteriorates for both components over a wakefulness period, only memory for central, relevant objects is preserved after a same period of time spent sleeping. However, post-training REM sleep may also play a role in protecting emotional memories against a overconsolidation or overboosting effect that would make these too invasive (as it is the case in post-traumatic disorders). Indeed, pharmacological blockade of the strong cortisol release during late (REM) sleep impairs consolidation of neutral but not of emotional texts, and even enhances recall performance for the latter. Hence it suggests that the natural and strong cortisol rise during REM sleep participates in hippocampus-dependent declarative memory consolidation for neutral material, but also protects from an overboosting in the consolidation of amygdala-dependent associations for emotional memories (Wagner et al., 2005).

Recollection of, and familiarity with, a stimulus reflect two different processes in memory. Whereas the former process involves the conscious retrieval of memories in an episodic context, the latter is an implicit expression of a memory trace that is not specifically linked to its context of acquisition. Whether sleep affects differentially or to the same extent recollection and familiarity processes for emotional material is debated. For instance, whereas Hu et al. (2006) observed sleep-related improvements both in response familiarity and recollection measures, others found that such improvement is limited to one of the two processes only. Furthermore, others found that sleep does not affect either dimension of memory at the behavioral level, in line with the hypothesis that emotional memories are of such ecological relevance that their consolidation should take place during posttraining wakefulness already. Besides equal performance at the behavioral level

however, brain imaging data also disclosed stronger connections between the hippocampus and the medial prefrontal cortex during the recollection of emotional material when participants were allowed to sleep after learning, whereas recollection was associated with higher activity in amygdala and occipital areas after sleep deprivation (Sterpenich et al., 2007), suggesting that sleep contributes depotentiating emotional reactivity and integrating memory core information into neocortical networks subtending long-term memory.

According to the sleep to forget and sleep to remember hypothesis (Van der Helm and Walker, 2009), the specific patterns of neurophysiological activity during sleep (and especially during REM sleep) would facilitate the decoupling of declarative memories from their emotional context. In this respect, it is assumed that the high concentration of aminergic neurotransmitters and the increased amygdala activity during wakefulness promote the encoding of emotional memories with their affective context. During subsequent REM sleep, the dramatic rise of cholinergic activity associated with theta oscillations would participate in the reactivation of the novel information, consolidating emotional memories. However, the concurrent decrease in aminergic activity during REM sleep would also lead to a depotentiation of the emotional tone initially associated with the learned information. Supporting this hypothesis, we showed that sleep reduces the facilitating effect of emotional congruency in the background mood context during the learning and retrieval phases, hence suggesting that postlearning sleep effectively releases recently learned memories from their emotional context (Deliens et al., 2012). At variance however, other studies confirmed a role for REM sleep in the processing of emotional memories, but suggested that the affective tone is preserved rather than reduced after posttraining sleep. An important difference between these studies is that an irrelevant emotional context surrounding neutral information was suppressed in the former case, whereas the consolidated information after sleep was intrinsically emotional in the latter case. Hence, sleep may indeed promote the decoupling between recently learned information and its emotional envelope, but only when this emotional information is not or weakly relevant.

Although findings are still inconsistent and the need for further research is warranted, studying the role played by sleep in emotional decontextualization processes is a particularly fascinating area of research considering the prevalence of psychiatric disorders involving mood disturbances that are associated with sleep disorders. For instance, the posttraumatic stress syndrome is specifically associated with REM sleep disturbances. Hence, a better understanding of the REM sleep mechanisms involved in emotion processing could benefit to the possible treatment of such pathologies.

See also: Context Dependent Memory; Declarative Memory, Neural Basis of; Episodic Memory; Genetic Approaches to Memory; Hippocampus and Related Structures; Implicit Memory; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory; Recognition; Retrieval; Retroactive Interference; Semantic Memory; Spatial Navigation; Synaptic Mechanisms and Protein Synthesis in Memory.

Bibliography

Arzi, A., Shedlesky, L., Ben-Shaul, M., Nasser, K., Oksenberg, A., Hairston, I.S., Sobel, N., 2012. Humans can learn new information during sleep. Nature Neuroscience. http://dx.doi.org/10.1038/nn.3193.

Born, J., Wilhelm, I., 2012. System consolidation of memory during sleep. Psychological Research 76, 192–203.

Buzsáki, G., 1996. The hippocampo-neocortical dialogue. Cereral Cortex 6, 81–92.

Deliens, G., Gilson, M., Schmitz, R., Peigneux, P., 2012. Sleep unbinds memories from their emotional context. Cortex 49 (8), 2221–2228.

Diekelmann, S., Wilhelm, I., Born, J., 2009. The whats and whens of sleep-dependent memory consolidation. Sleep Medicine Reviews 13 (5), 309–321.

Hasselmo, M.E., 1999. Neuromodulation: acetylcholine and memory consolidation. Trends in Cognitive Sciences 3, 351–359.

Hennevin, E., Huetz, C., Edeline, J.M., 2007. Neural representations during sleep: from sensory processing to memory traces. Neurobiology of Learning and Memory 87 (3), 416–440.

Hotermans, C., Peigneux, P., Maertens de Noordhout, A., Moonen, G., Maquet, P., 2006. Early boost and slow consolidation in motor skill learning. Learning and Memory 13 (5), 580–583.

Hu, P., Stylos-Allan, M., Walker, M.P., 2006. Sleep facilitates consolidation of emotional declarative memory. Psychological Science 17, 891–898.

Huber, R., Ghilardi, M.F., Massimini, M., Tononi, G., 2004. Local sleep and learning. Nature 430 (6995), 78–81.

Lewis, P.A., Durrant, S.J., 2011. Overlapping memory replay during sleep builds cognitive schemata. Trends in Cognitive Sciences 15 (8), 343–351.

Maquet, P., Laureys, S., Peigneux, P., Fuchs, S., Petiau, C., Phillips, C., Aerts, J., Del Fiore, G., Degueldre, C., Meulemans, T., Luxen, A., Franck, G., Van Der Linden, M., Smith, C., Cleeremans, A., 2000. Experience-dependent changes in cerebral activation during human REM sleep. Nature Neuroscience 3 (8), 831–883.

Maquet, P., Smith, C., Stickgold, R. (Eds.), 2003. Sleep and Brain Plasticity. Oxford University Press, Oxford.

Marshall, L., Helgadottir, H., Molle, M., Born, J., 2006. Boosting slow oscillations during sleep potentiates memory. Nature 444 (7119), 610–613.

Orban, P., Rauchs, G., Balteau, E., Degueldre, C., Luxen, A., Maquet, P., Peigneux, P., 2006. Sleep after spatial learning promotes covert reorganization of brain activity. Proceedings of the National Academy of Sciences of the United States of America 103 (18), 7124–7129.

Peigneux, P., Laureys, S., Delbeuck, X., Maquet, P., 2001. Sleeping brain, learning brain. The role of sleep for memory systems. NeuroReport 12 (18), 111–124.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Philipps, C., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A., Maquet, P., 2004. Are spatial memories strengthened in the human hippocampus during slow wave sleep? Neuron 44 (3), 535–545.

Peigneux, P., Smith, C., 2010. Memory processing in relation to sleep. In: Kryger, M. (Ed.), Principles and Practice of Sleep Medicine, fifth ed. Elsevier, Philadelphia, pp. 335–347.

Perrin, F., 2004. Auditory event-relates studies of information processing during sleep. Psychologica Belgica 44, 43–57.

Poe, G.R., Walsh, C.M., Bjorness, T.E., 2010. Cognitive neuroscience of sleep. Progress in Brain Research 185, 1–19.

Rauchs, G., Desgranges, B., Foret, J., Eustache, F., 2005. The relationships between memory systems and sleep stages. Journal of Sleep Research 14 (2), 123–140.

Sterpenich, V., Albouy, G., Boly, M., Vandewalle, G., Darseau, A., Balteau, E., Dang-Vu, T.T., Desseilles, M., D'Argembeau, A., Gais, S., Rauchs, G., Schabus, M., Degueldre, C., Luxen, A., Collette, F., Maquet, P., 2007. Sleep-related hippocampo–cortical interplay during emotional memory recollection. PLoS Biology 5 (11), e282.

Tononi, G., Cirelli, C., 2006. Sleep function and synaptic homeostasis. Sleep Medicine Reviews 10, 49–62.

Van der Helm, E., Walker, M.P., 2009. Overnight therapy? the role of sleep in emotional brain processing. Psychological Bulletin 135 (5), 731–748.

Wagner, U., Degirmenci, M., Drosopoulos, S., Perras, B., Born, J., 2005. Effects of cortisol suppression on sleep-associated consolidation of neutral and emotional memory. Biological Psychiatry 58 (11), 885–893.

Sleep Disorders across the Life Span

Weidong Wang, Xue Yan, and YanJiao Liu, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is estimated that 10% of the population suffer from sleep disorders, which have profound and widespread effects on human health. Although the diagnosis and treatment of sleep disorders is a relatively young discipline, understanding of the diagnosis, pathophysiology, and treatment of sleep disorders is evolving at a rapid pace. There are around 90 distinct sleep disorders in resent research; most are marked by one of these symptoms: excessive daytime sleepiness, difficulty initiating or maintaining sleep, and abnormal events occurring during sleep. The sleep disorders are associated with increased risk of hypertension, diabetes, obesity, coronary heart disease, heart attack, stroke, and depression. This article focuses on incorporating the essential features of common sleep disorders across the life span.

Sleep is a universal phenomenon that occupies almost one-third of human lifetime (Lee-Chiong, 2004). Sleep influences most physiologic processes in the body, and is, in turn, affected by specific circadian, medical, neuropsychiatric, and behavioral disorders, the sleep environment itself, and medication usage (Lee-Chiong, 2004).

Sleep disorders are common, and at least 10% of the population suffer from a sleep disorder that is clinically significant and of health importance (Ram et al., 2010). Although the diagnosis and treatment of sleep disorders is a relatively young discipline in medicine and psychopathology dating to the late 1970s, understanding of the diagnosis, pathophysiology, and treatment of sleep disorders is evolving at a rapid pace (Roehrs and Roth, 2004). At present, nearly 90 different sleep disorders are listed in the International Classification of Sleep Disorders (ICSD, Ed 2), which can be divided into eight categories (American Academy of Sleep Medicine, 2005). Sleep disorders vary widely in their complexity, their comorbidities, the risks they represent, and the scope of their manifestations (Colten and Altevogt, 2006). The four major sleep complaints include excessive daytime sleepiness; insomnia; abnormal movements, behaviors, or sensation during sleep; and inability to sleep at the desired time (American Academy of Sleep Medicine, 2005).

In addition, sleep problems are associated with health problems, functioning and well-being, work-related indicators, and health care expenditures (Roth et al., 2002). A number of recent studies have demonstrated that individuals with current sleep problems report significantly poorer health, less energy, and worse cognitive functioning than those categorized as having no sleep problem (Roth et al., 2002). The long-term effects of sleep loss and sleep disorders have been associated with a wide range of deleterious health consequences including increased risk of hypertension, diabetes, obesity, coronary heart disease, heart attack, heart failure, stroke, and depression (Colten and Altevogt, 2006).

The principles of sleep disorder treatment include first to find the cause of the sleep disturbance and vigorously treat the primary or comorbid conditions causing the sleep disturbance. If the treatment does not resolve the problem, then treatment should be directed at a specific sleep disturbance that may be caused by various neurological and medical diseases (Chokroverty, 2010).

Sleep Patterns Change across the Life Span

Sleep architecture changes continuously and considerably with age. From infancy to adulthood, there are marked changes in how sleep is initiated and maintained, the percentage of time spent in each stage of sleep, and overall sleep efficiency. Sleep requirements change dramatically from infancy to old age. A general trend is that sleep efficiency declines with age (**Table 1**) (Chokroverty, 2010; Colten and Altevogt, 2006). Such changes will be significantly affected by neurological, environmental, and genetic factors as well as comorbid medical or neurological disorders.

Common Sleep Disorders in Children

The diagnoses of sleep disorders are often overlooked and underdiagnosed in the pediatric population (Faruqui et al., 2011). Sleep problems in preadolescents have been reported for 10–33% of the population (Fricke-Oerkerman et al., 2007; Ipsiroglu et al., 2002; Meijer et al., 2000). An estimated 40% of adolescents report experiencing some forms of sleep problems, including issues of sleep difficulty, snoring, and sleep apnea (Corbo et al., 2001; Gozal and Pope, 2001; Tarasiu et al., 2007).

Sleep disorders in children can be divided into five categories, which are composed of difficulty falling asleep, difficulties with arousal, sleep-related movement disorders, sleep-related breathing disorder, and excessive daytime sleepiness (Carno et al., 2003). Many disorders are included in these categories. There are characteristic symptoms and ages of onset for the different sleep disorders.

Difficulty Falling Asleep

Difficulty falling asleep related to the process of going to sleep is called insomnia. Insomnia among children is often reported

 http://dx.doi.org/10.1016/B978-0-08-097086-8.21027-1

Table 1 Sleep patterns change across the life span

Age group	*Average sleep duration*	*Sleep feature*
Newborns	15–16 h per day	1. Biphasic 2. By 3 months of age, the non-REM (NREM)/REM cycling pattern of adult sleep is established 3. REM sleep constitutes 50% of total sleep time during early infancy 4. Ultradian sleep cycles of REM and NREM sleep states are substantially shorter for infants (50–60 min) 5. Some degree of nighttime waking is normal in young children, particularly during early infancy, when circadian mechanisms are still immature and nighttime feedings a necessity for most infants until about 6 months of age
Child	11 h per day	1. Biphasic 2. REM sleep declining by age 5 years to 20%
Adolescent	9–10 h per day	Monophasic
Adults	7.5–8 h per night	1. Monophasic 2. Ultradian sleep cycles of REM and NREM sleep states are substantially longer for adults (90–100 min)
Elderly people	7 h per night	1. Biphasic 2. Progressive decrease in slow wave sleep (SWS) and REM 3. Earlier bedtimes and wake times 4. Sleep efficiency continues to decline with age

by caretakers and characterized by bedtime resistance, an inability to sleep independently, or both (Colten and Altevogt, 2006). It can manifest as bedtime resistance, in which a child may cry, repeatedly attempt to leave the room/bed, or engage in other behaviors intended to forestall sleep onset. Delayed sleep onset and issues of sleep anxiety are also common, such as needing a parent in the room to fall asleep, the child signals being afraid to sleep alone and/or to sleep in the dark. In school-aged children, this can be a significant problem and may lead to poor school performance (Colten and Altevogt, 2006; Owens et al., 2000).

Children who lack structured bedtime routines or have irregular sleep habits represent a high-risk population. Other influences that may impede settling at bedtime include temperamental characteristics, anxiety, acute illness, medication side effects, and coexisting medical conditions such as attention deficit hyperactivity disorder (ADHD), autism, and other developmental disabilities (Hoban, 2010). Behavioral interventions and establishing good 'sleep hygiene' are the preferred method and goal of treatment (Carno et al., 2003; Hoban, 2010). Use of a structured, age-appropriate bedtime routine helps children sleep.

In addition, night waking is considered problematic when it is excessively prolonged or frequent for age, or when it is excessively disruptive to other members of the household (Hoban, 2010). Overall, about 20% of children below age 2 years exhibit night waking inducing concern by parents (Hoban, 2010). A variety of predisposing influences may be associated with problematic night waking, some of which relate to expected developmental changes during early childhood. Separation anxiety and cosleeping with a parent are associated with increased risk of arousal and awakening during nighttime sleep (Hoban, 2010; Kryger et al., 2005).

Another strong influence of sleep disorder on older school-age children is the well-recognized tendency toward delayed sleep phase type that develops near adolescence, with a reported prevalence of 7–16% (Reid et al., 2004). Delayed sleep phase syndrome (DSPS), one kind of circadian rhythm sleep disorder, may be diagnosed when this tendency results in difficulty maintaining school schedule, as well as chronic or recurrent inability of the child to fall asleep and wake up at conventional and socially acceptable times (American Academy of Sleep Medicine, 2005; Anstead, 2000; Carno et al., 2003; Reid et al., 2004). Besides a result of alterations of circadian timing, there is recent evidence that alterations in the homeostatic regulation of sleep may play an important role in the pathophysiology of DSPS (Watanabe et al., 2003). As common in adolescents, habitually later bedtime and waking time on nonschool nights, end to perpetuate or worsen the tendency toward DSPS (Crowley et al., 2007). Treatment includes a gradual change in the child's sleep time until the desired sleep–wake time is achieved. Both the parents and the child must be committed to the prescribed plan for success (Carno et al., 2003).

Difficulties with Arousals

Parasomnias are episodic disturbances of sleep that happen most often in preschool and school-aged children (D'Cruz and Vaughn, 2001). These aberrant arousals are associated with a spectrum of abnormal behavior or physiologic events that occur in association with sleep, specific sleep stages, or sleep–wake transitions and can include night terrors, sleepwalking, sleep talking, and so on (D'Cruz and Vaughn, 2001; Hoban, 2010). During a parasomnia episode, children are difficult to awaken during the event, and if awakened, demonstrate little or no recall for the episode (D'Cruz and Vaughn, 2001). Multiple parasomnias often occur in the same person at different ages. They are more common in boys than in girls, and often found in conjunction with a positive family history (D'Cruz and Vaughn, 2001; Hublin et al., 2001; Lecendreux et al., 2003; Owens et al., 2000). Limited evidence suggests that obstructive sleep apnea (OSA) or restless legs syndrome (RLS) may represent a treatable precipitant of these parasomnias for some children (Guilleminault et al., 2003). Studies have shown a relationship between the child's stress

level and parasomnias such as night terrors, somniloquy, and body rocking (Laberge et al., 2000).

Night terrors, sleepwalking, and confusional arousal in children are thought to represent forms of incomplete arousal from non-rapid eye movement (REM) sleep. Exacerbations of non-REM arousal parasomnias are often triggered by sleep deprivation or intercurrent illness. Night terrors and confusional arousal affect up to 17% of children (Laberge et al., 2000) but are most commonly observed in toddlers and preschoolers with no dream association or recollection of the incident. These incidents are accompanied by intense autonomic and motor symptoms, which range from moaning, quiet crying, sitting upright in bed, or screams to extreme agitation (Hoban, 2010; Laberge et al., 2000). Patients appear highly confused and fearful. Episodes can last from only seconds to as long as 1 h, but most typically remit within minutes (Carno et al., 2003; Hoban, 2010).

Sleepwalking and sleep talking are more common in school-age children. Sleep talking (somniloquy) is the most frequent parasomnia and can be considered a universal experience (Carno et al., 2003). Although 40% of children exhibit at least one episode of sleepwalking between the ages of 5 and 12 years, frequent sleepwalking affects only 2–3% with a peak prevalence near 10 years of age (Laberge et al., 2000). Episodes of sleepwalking consist of quiet ambulation lasting several minutes, which may include semipurposeful activities such as wandering to some place of the house or urinating in inappropriate locations (Hoban, 2010). Injuries and violent activities have been reported during sleepwalking episodes but generally individuals can negotiate their way around the room (Chokroverty, 2010).

Use of scheduled awakenings is often effective for the treatment of the non-REM arousal parasomnias in children whose sleepwalking or night terrors occur at predictable times of the night (Hoban, 2010). For sleepwalking, security of the environment is important. Drug treatment using clonazepam may be effective for children having particularly frequent or severe forms of sleepwalking and night terrors (Remulla et al., 2004).

Nightmares during childhood are common, and over half of early school-aged children were affected on an occasional basis but only 3% of children show this disturbance more than once a week (Smedje et al., 1999). Different from night terrors, nightmares arise from REM sleep and are thought to result from awakening during a frightening dream. The child awakens fully, and has recollection of the dream content. Persistent nightmares, which are unexpectedly frequent, prolonged, or violent in content, need especially further attention of potential psychological and medical causes (Hoban, 2010). It may be helpful to teach stress modification techniques based on age, which is also beneficial for sleep talking (Howard and Wong, 2001).

Sleep enuresis in children is defined as urinating more than twice a week for children older than 5 years and characterized by recurrent involuntary voiding during sleep at a level that is inappropriate for age (American Academy of Sleep Medicine, 2005). Nocturnal enuresis becomes less common with advanced age, affecting 15% of 3- to 10-year-olds, but only 2% of 13-year-olds (Laberge et al., 2000). It more commonly occurs in boys than in girls. Postulated causes of primary enuresis include immaturity of arousal mechanisms, reduced functional bladder capacity, and genetic predisposition. Children with frequent enuresis often respond to behaviorally based therapies and conditioning programs (Butler et al., 2007; Lottmann and Alova, 2007). Drug treatment like desmopressin (Ferrara et al., 2008) and imipramine (Neveus et al., 2008) is also effective, which is best used within the context of a comprehensive and closely supervised treatment program.

Sleep-Related Movement Disorders

RLS and periodic limb movement disorder (PLMD) are different but related conditions in children and are often comorbid conditions, but either of them may present independently. The RLS criteria in childhood include the adult criteria and additional criteria designed to address the absence or uncertainty of verbal reports (American Academy of Sleep Medicine, 2005). Some children with growing pains may actually have RLS as well (Rajaram et al., 2004). Leg restlessness and growing pains affect 17% and up to 8% of children, respectively (Chervin et al., 2002), but only 2% of children meet diagnostic criteria of RLS (Picchietti et al., 2007). Genetic influences and iron deficiency are suggested as the causes of RLS in the pediatric age group (Muhle et al., 2008; Oner et al., 2007). Childhood PLMD is characterized by excessive periodic limb movements occurring over five times per hour of sleep during polysomnography (PSG), accompanied by clinical complaints of disturbed sleep (American Academy of Sleep Medicine, 2005). The pathophysiology of PLMD is still unknown, but dopaminergic impairment has been implicated (American Academy of Sleep Medicine, 2005). Studies have reported possible association between childhood RLS/PLMD and symptoms of ADHD (Cortese et al., 2005). Treatment options for childhood RLS/PLMD include iron supplementation with low ferritin levels or iron deficiency (Picchietti et al., 2008), or judicious use of dopaminergic agonists (Walters et al., 2000).

Sleep-related rhythmic movement disorder (RMD) is characterized by recurrent and well-stereotyped rhythmic behaviors associated with sleep, which is sleep–wake transition parasomnias that involve rhythmic movement of a large muscle group (American Academy of Sleep Medicine, 2005; Hoban, 2003). Body rocking, head or body rolling, and head banging are some common examples. Episodes of rhythmic movement usually last for several minutes and may recur throughout the night (Hoban, 2010). Rhythmic movements may represent self-comforting or a sleep-inducing behavior (D'Cruz and Vaughn, 2001). Parents may be unaware of the sleep disturbance unless some sort of noise is involved. Sleep-related rhythmic behaviors are observed in a majority of infants, usually between 8 and 18 months, but disappear around 5 years of age in most children (Dyken and Rodnitzky, 1992). Only 3% of children manifest prominent rhythmic behaviors associated with sleep in early adolescence (Laberge et al., 2000). RMD is a self-limited condition that does not need treatment in most children. But for children exhibiting particularly prolonged

or violent forms of RMD, benzodiazepines, behavioral therapies, or use of padding are suggested (Hoban, 2010).

Sleep-Related Breathing Disorder

OSA refers to a breathing disorder during sleep that is characterized by either prolonged partial upper airway obstruction (hypopnea) and/or intermittent complete obstruction (apnea) (American Academy of Sleep Medicine, 2005). OSA in children may not present as overtly as in adults and sometimes associated sleep disturbance can be the most obvious symptom and the reason why the child comes to medical attention in the first place (Wiggs, 2003). The prevalence of OSA is approximately 2% in otherwise normal young children (American Academy of Sleep Medicine, 2005). Although OSA can occur in children of any age, it is most common at preschool ages around 2–8 years, a time coincident with tonsils and adenoids being largest relative to the underlying airway (Colten and Altevogt, 2006). Other genetic, neurological, and craniofacial conditions and obesity are associated with increased risk for the development of childhood OSA. The clinical features of OSA in children include snoring, unusual sleeping positions, headache, and daytime sleepiness (American Academy of Sleep Medicine, 2005). Children with OSA have an increased risk for neurocognitive deficits such as behavioral problems, poor learning, and/or ADHD, which result from the lower oxygen levels during the episodes of sleep apnea (Marcus, 2001). Adenotonsillectomy is the most common treatment of childhood OSA. Otherwise, nasal continuous positive airway pressure (CPAP) is an alternative first-line treatment for OSA in children of all ages (O'Donnell et al., 2006). Other treatments are then advised, such as weight loss, use of supplemental oxygen, positional therapy, or more extensive surgery (Hoban, 2010; Schechter, 2002).

Excessive Daytime Sleepiness

Narcolepsy is a neurologic disorder characterized by excessive daytime somnolence, which may be accompanied by cataplexy, sleep paralysis, and hallucinations at sleep onset or offset. It appears commonly during the second decade of life and after the reported onset of puberty (Kotagal, 1996). Onset of narcolepsy can also have a negative impact on school performance. Misdiagnosis of narcolepsy is common in children owing to similar psychiatric disorders and emotional and behavioral changes. Treatment of narcolepsy in children, in most respects, parallels that of adults (Hoban, 2010). However, large-scale clinical trials have not examined the efficacy and safety of drugs to treat narcolepsy in children and adolescents (Colten and Altevogt, 2006).

Bruxism (Tooth Grinding)

Bruxism often presents between ages 10 and 20 years, but it may persist throughout life, often leading to secondary problems such as temporomandibular joint dysfunction. It is also commonly noted in children with mental retardation or cerebral palsy. The episode is characterized by stereotypical tooth grinding and is precipitated by anxiety, stress, and dental disease (Hoban, 2010).

Common Sleep Disorders in Adults

Sleep disorders have an estimated prevalence of 15–27% in the adult population (Roth et al., 2002). While there are approximately 90 sleep disorders recognized by the medical community (American Academy of Sleep Medicine, 2005), the authors have put together a list of the most common sleep disorders seen in adults as follows. Many people have more than one sleep issue.

Insomnia

Insomnia is defined as a repeated difficulty with sleep initiation, duration, consolidation, or quality that occurs despite adequate time and opportunity for sleep and results in some form of daytime impairment (American Academy of Sleep Medicine, 2005). The prevalence in representative population surveys is 10–14% for chronic insomnia and 20–40% for occasional, or less severe, insomnia (Kryger et al., 2005). The precise causes of insomnia are poorly understood but, in general terms, involve a combination of biological, psychological, and social factors (Colten and Altevogt, 2006). Surveys consistently find that the prevalence of insomnia is higher in women and that it increases with age. In addition, insomnia is more prevalent in those with medical illness, substance users, and those with anxiety or depressive disorders (Roehrs and Roth, 2004; Kryger et al., 2005). Among adults, insomnia complaints typically include reported difficulties initiating or maintaining sleep. Concerns about extended periods of nocturnal wakefulness or insufficient amounts of nocturnal sleep usually accompany these complaints. Consequences of insomnia in adults include an increased likelihood of accidents and absenteeism, diminished performance, impaired judgment and social or vocational functioning, polydrug abuse (including alcoholism), increased use of health care resources, reduction in quality of life, and disturbance in interpersonal relationships. It is important and necessary that short-term sleep disturbance be promptly recognized and appropriately treated before learned habits, attitudes, and coping mechanisms incongruous with sleep become established and perpetuate the sleep disturbance. Insomnia is treatable with a variety of behavioral and pharmacological therapies, which may be used alone or in combination. Therapy has to address the predisposing, precipitating, and perpetuating factors of insomnia (Lee-Chiong, 2004).

Sleep-Disordered Breathing

Sleep-disordered breathing (SDB) refers to a spectrum of disorders that feature breathing pauses during sleep. The most common disorder is characterized by obstructive apneas and

hypopneas (Kryger et al., 2005). Obstructive sleep apnea syndrome (OSAS) is characterized by repetitive episodes of complete (apnea) or partial (hypopnea) upper airway obstruction occurring during sleep (American Academy of Sleep Medicine, 2005; Kryger et al., 2005). These events often result in loud snoring, blood oxygen saturations, and many and recurrent periods of arousal. They typically recur throughout the evening, at times reaching numbers substantial enough to produce sleep fragmentation and subsequent daytime sleepiness (American Academy of Sleep Medicine, 2005). The diagnosis of OSAS requires detection, by PSG, of at least five or more apneas or hypopneas per hour of sleep (Colten and Altevogt, 2006). OSAS tends to affect mainly middle-aged individuals and is strongly associated with obesity. An estimated 4% of men and 2% of women between the ages of 30 and 60 years in the United States have OSAS (Lee-Chiong, 2004). OSAS increases the risk for hypertension, coronary artery disease, heart failure, arrhythmias, cerebrovascular disease, stroke, glucose intolerance, diabetes, and even mortality (Colten and Altevogt, 2006; Lee-Chiong, 2004). In adults, OSA is most effectively treated with CPAP and weight loss (Kryger et al., 2005). Other options, although less effective, include a variety of dental appliances or surgery (e.g., uvulopalatopharyngoplasty) (Kryger et al., 2005).

Sleep-Related Movement Disorders

Restless Legs Syndrome

RLS is a sensorimotor disorder characterized by a strong, nearly irresistible urge to move the legs (American Academy of Sleep Medicine, 2005). It is also associated with paresthesias – uncomfortable feelings – which individuals describe as creeping, crawling, jittery, itchy, aching, or burning feelings. The symptoms are partially or completely relieved by movement. The urge to move and unpleasant sensations worsen during periods of rest or inactivity, especially in the evening and at night, causing most individuals difficulty falling sleep and maintaining sleep (American Academy of Sleep Medicine, 2005). At least 80–90% of RLS patients have periodic limb movement in sleep (PLMS) and may also have periodic limb movement in wakefulness; however, one-third of those with PLMD will have RLS (Chokroverty, 2010; Michaud et al., 2000). The prevalence has been reported at about 5–10% for adult populations but the prevalence of most severe cases is approximately 2.5%, which makes it one of the most common movement disorders and sleep disorders (Chokroverty, 2010; Colten and Altevogt, 2006). In most surveys, the prevalence is greater in women than in men and the disease is chronic and progressive (Chokroverty, 2010). Iron, dopamine, and genetics appear to be primary factors in the pathology of RLS (American Academy of Sleep Medicine, 2005). There are both behavioral and pharmacological treatments for RLS. Four classes of prescription medications are mostly used for the treatment of RLS: dopaminergic agents, benzodiazepines, opioids, or anticonvulsants, in which dopaminergic agents are the primary treatment option for RLS (Colten and Altevogt, 2006). Mild to moderate symptoms can be treated sometimes by lifestyle changes, including maintaining a normal sleeping pattern, taking supplements to manage iron deficiencies, and minimizing consumption of alcohol, caffeine, and tobacco (Colten and Altevogt, 2006).

Periodic Limb Movement Disorder

PLMD is defined as periodic episodes of repetitive, highly stereotyped limb movements characterized by rhythmic extensions of the big toe, dorsiflexions of the ankle, and occasional flexions of the knee and hip, which cannot be explained by any other sleep disorder. Nonetheless, these stereotypical jerks can partially arouse or awaken the individual, leading to sleep fragmentation and complaints of insomnia or hypersomnolence (American Academy of Sleep Medicine, 2005). It was found to have a population prevalence of 4% in studies of five European countries (Kryger et al., 2005). It is also observed in individuals with narcolepsy, rapid eye movement sleep behavior disorder (RBD), OSA, and hypersomnia (Chokroverty, 2010). PLMD is associated with high rates of depression, memory impairment, attention deficits, oppositional behaviors, and fatigue (American Academy of Sleep Medicine, 2005). Similar to RLS, dopaminergic medications are helpful in alleviating the disorder's symptoms.

Narcolepsy and Hypersomnia

Narcolepsy and idiopathic hypersomnia are characterized by a clinically significant complaint of excessive daytime sleepiness that is neither explained by a circadian sleep disorder, SDB, or sleep deprivation, nor caused by a medical condition disturbing sleep (American Academy of Sleep Medicine, 2005). Narcolepsy is characterized by narcoleptic sleep attacks (100%), cataplexy (60–70%), sleep paralysis (25–50%), hypnagogic hallucinations (20–40%), disturbed night sleep (70–80%), and automatic behavior (20–40%) (Chokroverty, 2010). In addition to the major manifestations, patients with narcolepsy may also have four important comorbid conditions: sleep apnea, PLMS, RBD, and nocturnal eating disorder (Chokroverty, 2010). Excessive sleepiness may manifest as pervasive drowsiness, frequent napping, and unexpected and overpowering sleep attacks. Repetitive awakenings, automatic or inappropriate behavior, and memory impairment are associated features. It is a lifelong condition but it generally is less severe and may even disappear in old age (Chokroverty, 2010; Lee-Chiong, 2004). It has a population prevalence of 0.03–0.16% of the general population (Lee-Chiong, 2004).

Excessive sleepiness in persons with idiopathic hypersomnia, which closely resembles narcolepsy syndrome, manifests as extended major sleep episodes (lasting over 8 h) and prolonged naps (up to 1–2 h) (Lee-Chiong, 2004). The sleep pattern, however, is different from that of narcolepsy. The patient generally sleeps for hours but the sleep is not refreshing. The patient does not give a history of cataplexy, snoring, or repeated awakenings throughout the night. Recurrent hypersomnia can manifest in two forms: hypersomnia only or accompanied by binge eating and hypersexuality (Kleine–Levin syndrome) (Colten and Altevogt, 2006; Lee-Chiong, 2004).

The diagnosis of narcolepsy and hypersomnia is based principally on the multiple sleep latency test, which objectively quantifies daytime sleepiness, or by measuring cerebrospinal fluid hypocretin-1 (Kryger et al., 2005). Secondary cases of narcolepsy or hypersomnia are also common, which can occur in the context of psychiatric disorders, central nervous system tumors, Parkinson's disease, multiple sclerosis, head trauma, and genetic disorders (Colten and Altevogt, 2006). Treatment for narcolepsy and hypersomnia is symptomatically based. In most cases, pharmacological treatment (such as modafinil or amphetamine-like stimulants) is needed. Behavioral measures are helpful but rarely sufficient (Colten and Altevogt, 2006).

Circadian Rhythm Sleep Disorders

Circadian rhythm sleep disorders arise from chronic alterations, disruptions, or misalignment of the circadian clock in relation to environmental cues and the terrestrial light–dark cycle (Colten and Altevogt, 2006). The ICSD 2 designated nine different circadian disorders, and the common primary disorders include delayed sleep phase type, advanced sleep phase type, irregular sleep–wake type, free-running type, shift work type, and jet lag type (American Academy of Sleep Medicine, 2005; Colten and Altevogt, 2006). These conditions often result in complaints of insomnia and excessive sleepiness. In addition to physiological and environmental factors, maladaptive behaviors influence the presentation and severity of the circadian rhythm sleep disorders (American Academy of Sleep Medicine, 2005). These disorders may be comorbid with other neurological or psychiatric disorders, making the diagnosis and treatment difficult (Kryger et al., 2005). Treatment for this disorder requires resynchronizing to a more appropriate phase to the 24-h light–dark cycle. In addition to a structured sleep–wake schedule and good sleep hygiene practices, potential therapies include resetting the circadian pacemaker with bright light, melatonin, or a combination of both (Colten and Altevogt, 2006; Reid et al., 2004).

Parasomnias

Parasomnias are undesirable physical, behavioral, or experiential phenomena that occur during entry into sleep, within sleep, or during arousals from sleep (American Academy of Sleep Medicine, 2005). Parasomnias generally manifest as central nervous system activation transmitted into skeletal muscles and the autonomic nervous system channels, often with experiential concomitants, and tend to occur during the transition from one state to another. Parasomnias encompass abnormal sleep-related movements, behaviors, emotions, perceptions, dreaming, and autonomic nervous system functioning, which may result in injuries, sleep disruption, adverse health effects, and psychosocial problems (American Academy of Sleep Medicine, 2005). The ICSD 2 lists 15 items and some of these entities are rare (American Academy of Sleep Medicine, 2005). The common primary sleep parasomnias include disorders of arousal (e.g., confusional arousals, sleepwalking, and sleep terror), RBD, nightmare disorder, sleep enuresis, sleep-related groaning, and eating disorder, most of which are usually common in children (see Section Common Sleep Disorders in Children) (Lee-Chiong, 2004).

Common Sleep Disorders in Older Adults

Sleep disturbances in older adults are often multifactorial. It has been estimated that approximately 50% of older adults complain about difficulty initiating or maintaining sleep (Crowley, 2011; Neikrug and Ancoli-Israel, 2010). However, this higher prevalence of sleep disruption is less a function of age and more a function of other factors that accompany aging, which include the increased presence of medical and psychosocial comorbidities, increased medication use, advances in the endogenous circadian clock, and a higher prevalence of specific sleep disorders (e.g., OSA, PLMS, and RLS) (Bloom et al., 2009). Of major clinical concern is the strong bidirectional relationship between sleep disorders and serious medical problems in older persons (Bloom et al., 2009). Individuals with sleep disorders negatively affect the outcomes of chronic diseases such as hypertension, depression, cardiovascular, and cerebrovascular disease and are associated with increasing mortality and morbidity. Conversely, individuals with any of these diseases are at higher than normal risk of developing sleep problems (Cuellar et al., 2007; Foley et al., 2004; Neikrug and Ancoli-Israel, 2010; Taylor et al., 2007). The most common primary sleep disorders in the elderly include insomnia, SDB, RLS/PLMS, and RBD (Crowley, 2011). Treatment of primary sleep problems can improve the quality of life and daytime functioning of older adults (Neikrug and Ancoli-Israel, 2010).

Insomnia

Numerous studies have shown that the prevalence of insomnia is higher in older adults than in younger adults, with estimates as high as approximately 40% in those aged over 65 years (Ancoli-Israel, 2009; Foley et al., 2004; Paudel et al., 2008). The consequences of insomnia in older adults are substantial and include daytime functional impairments, poor health, decreased physical function, falls, cognitive impairment, and mortality (Ancoli-Israel, 2009). Insomnia is often comorbid with medical and psychiatric illness, such as renal disease, nocturia, cerebrovascular disease, gastrointestinal disease, respiratory disorders, chronic pain and arthritis, neurological diseases, menopause cardiovascular disease, anxiety and depression, as well as multiple medications use in the elderly (Neikrug and Ancoli-Israel, 2010). There are also other factors affecting the ability to sleep in older adults including intake of alcohol, caffeine, and nicotine; circadian rhythm disturbances; and primary sleep disorders, such as SDB, RLS, and RBD (Ancoli-Israel, 2009). While medications are traditionally used to treat insomnia, recent studies have shown that behavioral treatments are more effective and, thus, recommended as the first-line treatment option. In some instances, a combined approach may provide for better

outcomes (Neikrug and Ancoli-Israel, 2010). It is of utmost importance to make the correct medical and psychiatric diagnoses and treat these problems along with appropriate management of the underlying sleep disturbances in older adults (Ancoli-Israel, 2009).

Sleep-Disordered Breathing

SDB is more common in older adults compared to younger adults. Adults 65–90 years of age had a threefold higher prevalence rate than middle-aged adults (Colten and Altevogt, 2006). The prevalence rates of SDB in the elderly were around 55–70% (Bloom et al., 2009; Crowley, 2011; Neikrug and Ancoli-Israel, 2010). Older adults with SDB may also report insomnia, nocturnal confusion, psychomotor function, and daytime cognitive impairment including difficulty with attention and short-term memory loss (Crowley, 2011; Launois et al., 2007). SDB has been associated with nocturia, arterial hypertension, heart failure, atrial fibrillation, diabetes, and stroke, conditions that are more common in the older population (Buchner et al., 2007). Depression has also been found as a common comorbidity in women with SDB (Shepertycky et al., 2005). Older adults are rarely considered candidates for surgery and CPAP is the best approach and first-line treatment for most elderly patients (Bloom et al., 2009; Neikrug and Ancoli-Israel, 2010).

RLS/PLMS

PLMS and RLS are both common in the older adult. The prevalence of RLS/PLMS increases significantly with age, which is experienced by up to 15–34% over the age of 60 years (American Academy of Sleep Medicine, 2005; Cuellar et al., 2007; Neikrug and Ancoli-Israel, 2010). Increasing prevalence of RLS with age may also occur in association with the increasing presence of secondary causes in the aging population, such as age-related declines in dopamine receptors, iron deficiency, and renal failure (Bloom et al., 2009; Crowley, 2011). At the spinal level, lumbosacral narrowing has been reported in some older adults with RLS/PLMS, which is consistent with other data suggesting proprioceptive feedback of limb position is important in the initiation and termination of episodes (Crowley, 2011). Other systemic factors have been proposed, including venous insufficiency in the lower limbs, increased parasympathetic nervous system activity, and iron-deficiency anemia. The recommended treatments for RLS/PLMS in elderly is similar with adults as mentioned before in this article.

REM Sleep-Behavior Disorder

RBD is an important REM sleep parasomnia commonly seen in elderly individuals, which is characterized by abnormal behaviors emerging and intermittent loss of muscle hypotonia or atonia during REM sleep that cause injury or sleep disruption (American Academy of Sleep Medicine, 2005; Chokroverty, 2010). RBD is a male-predominant disorder that emerges after the age of 50 years (American Academy of Sleep Medicine, 2005). The prevalence in the elderly population has been reported around 0.5% (American Academy of Sleep Medicine, 2005). It is seen with increasing prevalence in patients with neurological or neurodegenerative disorder, particularly Parkinson's disease, multiple system atrophy, dementia with Lewy body, narcolepsy, and stroke. PSG demonstrates intermittent loss of REM sleep-associated muscle atonia, with the patient manifesting complex, often violent motor activity associated with dream mentation (Bloom et al., 2009; Chokroverty, 2010). Treatment of RBD involves pharmacologic treatment (e.g., clonazepam and levodopa) and interventions that address environmental safety (Bloom et al., 2009).

Circadian Rhythm Sleep Disorders

In the older adult, factors associated with aging are thought to contribute to the desynchronization of rhythms. The circadian pacemaker itself degenerates with age, which results in less robust rhythms (Reid et al., 2004). With age there is also a gradual decrease in rhythm amplitude, which likely contributes to less consistent periods of sleep–wake across the 24-h day (Reid et al., 2004). The endogenous secretion of melatonin at night is also reduced with age, which results in a weaker circadian rhythm. Therefore, the combination of age-related changes in sleep and circadian rhythm regulation paired with decreased levels of light exposure and activity contribute to the development of circadian rhythm-based sleep disorders in older adults (Bloom et al., 2009). The most common clinical consequence of changes in circadian network of old adults is the advanced sleep phase disorder (ASPD). In middle- to older aged adults, the prevalence of ASPD is estimated at 1–7% (Ando et al., 2002; Bloom et al., 2009). This advanced phase results in sleep and wake times that are several hours earlier than conventional and desired times (American Academy of Sleep Medicine, 2005). Older adults, therefore, feel sleepy in the early evening and awaken in the very early morning hours. Sleep during these times is often normal, but out of sync with the environment. A combination of good sleep hygiene practices and methods to delay the timing of sleep and wake times is usually recommended for the treatment of ASPD (Bloom et al., 2009).

Summary

Although sleep disorders medicine is a relatively young discipline, the understanding of it is evolving at a rapid rate (Roehrs and Roth, 2004). This evolution is seen in the development of diagnostic systems, pathophysiology understanding, signs and symptoms, and finally in the treatment of sleep disorders and comorbid conditions. Treatment of sleep disorders can result in fewer complications of chronic illnesses, increased quality of life, and lower health care costs. It is important for all practicing physicians to understand the implications of disrupted sleep in different age groups and be able to assess accurately and intervene to improve health outcomes and quality through the life span of human beings.

See also: Occupational Health; Sleep Disorders: Psychiatric Aspects; Sleep and Memory; Sleep: Neural Systems.

Bibliography

American Academy of Sleep Medicine, 2005. International classification of sleep disorders. In: Diagnostic and Coding Manual, second ed. Westchester, Illinois.

Ancoli-Israel, S., 2009. Sleep and its disorders in aging populations. Sleep Medicine 10, S7–S11.

Ando, K., Kripke, D.F., Ancoli-Israel, S., 2002. Delayed and advanced sleep phase symptoms. The Israel Journal of Psychiatry and Related Sciences 39, 11–18.

Anstead, M.I., 2000. Pediatric sleep disorders: new developments and evolving understanding. Current Opinion in Pulmonary Medicine 6, 501–506.

Bloom, H.G., Ahmed, I., Alessi, C.A., Ancolilsrael, S., Buysse, D.J., Kryger, M.H., et al., 2009. Evidence-based recommendations for the assessment and management of sleep disorders in older persons. Journal of American Geriatrics Society 57 (5), 761–789.

Buchner, N.J., Sanner, B.M., Borgel, J., et al., 2007. Continuous positive airway pressure treatment of mild to moderate obstructive sleep apnea reduces cardiovascular risk. American Journal of Respiratory and Critical Care Medicine 176, 1274–1280.

Butler, R.J., Holland, P., Gasson, S., et al., 2007. Exploring potential mechanisms in alarm treatment for primary nocturnal enuresis. Scandinavian Journal of Urology 41, 407–413.

Carno, M.A., Hoffman, L.A., Carcillo, J.A., Sanders, M.H., 2003. Developmental stages of sleep from birth to adolescence, common childhood sleep disorders: overview and nursing implications. Journal of Pediatric Nursing 18 (4), 274–283.

Chervin, R.D., Archbold, K.H., Dillon, J.E., et al., 2002. Associations between symptoms of inattention, hyper-activity, restless legs, and periodic leg movements. Sleep 25, 213–218.

Chokroverty, S., 2010. Overview of sleep and sleep disorders. The Indian Journal of Medical Research 131 (2), 126–140.

Colten, H.R., Altevogt, B.M., 2006. Sleep Disorders and Sleep Deprivation, first ed. National Academies Press, Washington.

Corbo, G., Forastiere, F., Agabiti, N., et al., 2001. Snoring in 9- to 15-year-old children: risk factors and clinical relevance. Pediatrics 108 (5), 1149–1154.

Cortese, S., Konofal, E., Lecendreux, M., et al., 2005. Restless legs syndrome and attention-deficit/hyperactivity disorder: a review of the literature. Sleep 28, 1007–1013.

Crowley, K., 2011. Sleep and sleep disorders in older adults. Neuropsychology Review 21, 41–53.

Crowley, S.J., Acebo, C., Carskadon, M.A., et al., 2007. Sleep, circadian rhythms, and delayed phase in adolescence. Sleep Medicine 8, 602–612.

Cuellar, N.G., Rogers, A.E., Hisghman, V., Volpe, S.L., 2007. Assessment and treatment of sleep disorders in the older adult. Geriatric Nursing 28 (4), 254–264.

D'Cruz, O., Vaughn, B., 2001. Parasomnias—an update. Seminars in Pediatric Neurology 8, 251–257.

Dyken, M.E., Rodnitzky, R., 1992. Periodic, aperiodic and rhythmic disorders of sleep. Neurology 42 (Suppl. 6), 68–74.

Faruqui, F., Khubchandani, J., Price, J.H., Bolyard, D., Reddy, R., 2011. Sleep disorders in children: a national assessment of primary care pediatrician practices and perceptions. Pediatrics 128 (3), 539–546.

Ferrara, P., Marrone, G., Emmanuele, V., et al., 2008. Homotoxicological remedies versus desmopressin versus placebo in the treatment of enuresis: a randomised, double-blind, controlled trial (see comment). Pediatric Nephrology 23, 269–274.

Foley, D., Ancoli-Israel, S., Britz, P., Walsh, J., 2004. Sleep disturbances and chronic disease in older adults: results of the 2003 National Sleep Foundation Sleep in America Survey. Journal of Psychosomatic Research 56, 497–502.

Fricke-Oerkerman, L., Pluck, J., Schred, L.M., et al., 2007. Prevalence and course of sleep problems in childhood. Sleep 30 (10), 1371–1377.

Gozal, D., Pope Jr., D.W., 2001. Snoring during early childhood and academic performance at ages thirteen to fourteen years. Pediatrics 107 (6), 1394–1399.

Guilleminault, C., Palombini, L., Pelayo, R., et al., 2003. Sleepwalking and sleep terrors in prepubertal children: what triggers them? Pediatrics 111, e17–e25.

Hoban, T.F., 2003. Rhythmic movement disorder in children. CNS Spectrums 8, 135–138.

Hoban, T.F., 2010. Sleep disorders in children. Annals of the New York Academy of Sciences 1184, 1–14.

Howard, B., Wong, J., 2001. Sleep disorders. Pediatrics in Review 22, 327–341.

Hublin, C., Kaprio, J., Partinen, M., Koskenvu, M., 2001. Parasomnias: co-occurrence and genetics. Psychiatric Genetics 11, 65–70.

Ipsiroglu, D.S., Fatemi, A., Werner, I., et al., 2002. Self-reported organic and non-organic sleep problems in school children aged 11 to 15 years in Vienna. Journal of Adolescent Health 31 (5), 436–442.

Kotagal, S., 1996. Narcolepsy in children. Seminars in Pediatric Neurology 3, 36–43.

Kryger, M.H., Roth, T., Dement, W.C., 2005. Principles and Practice of Sleep Medicine, fourth ed. Elsevier/Saunders, Philadelphia.

Laberge, L., Tremblay, R.E., Vitaro, F., Montplaisir, J., 2000. Development of parasomnias from childhood to early adolescence. Pediatrics 106, 67–74.

Launois, S.H., Pepin, J.L., Levy, P., 2007. OSA in the elderly: a specific entity? Sleep Medicine Reviews 11, 83–85.

Lecendreux, M., Bassetti, C., Dauvilliers, Y., et al., 2003. HLA and genetic susceptibility to sleepwalking. Molecular Psychiatry 8, 14–117.

Lee-Chiong, T.L., 2004. Sleep and sleep disorders: an overview. Medical Clinics of North America 88, 11–14.

Lottmann, H.B., Alova, I., 2007. Primary monosymptomatic nocturnal enuresis in children and adolescents. International Journal of Clinical Practice. Supplement 155, 8–16.

Marcus, C.L., 2001. Sleep-disordered breathing in children. American Journal of Respiratory and Critical Care Medicine 164, 16–30.

Meijer, A.M., Habekothe, H.T., Van Den Wittenboer, G.L., 2000. Time in bed, quality of sleep and school functioning of children. Journal of Sleep Research 9 (2), 145–153.

Michaud, M., Chabli, A., Lavigne, G., Montplaisir, J., 2000. Arm restlessness in patients with restless legs syndrome. Movement Disorders 15 (2), 289–293.

Muhle, H., Neumann, A., Lohmann-Hedrich, K., et al., 2008. Childhood-onset restless legs syndrome: clinical and genetic features of 22 families. Movement Disorders 23, 1113–1121.

Neikrug, A.B., Ancoli-Israel, S., 2010. Sleep disorders in the older adult – a mini review. Gerontology 56, 181–189.

Neveus, T., Tullus, K., Neveus, T., Tullus, K., 2008. Tolterodine and imipramine in refractory enuresis; a placebo-controlled crossover study. Pediatric Nephrology 23, 263–267.

O'Donnell, A.R., Bjornson, C.L., Bohn, S.G., Kirk, V.G., 2006. Compliance rates in children using noninvasive continuous positive airway pressure. Sleep 29, 651–658.

Oner, P., Dirik, E.B., Taner, Y., et al., 2007. Association between low serum ferritin and restless legs syndrome in patients with attention deficit hyperactivity disorder. Tohoku Journal of Experimental Medicine 213, 269–276.

Owens, J.A., Spirito, A., McGuinn, M., Nobile, C., 2000. Sleep habits and sleep disturbance in elementary school-aged children. Journal of Developmental and Behavioral Pediatrics 21, 27–36.

Paudel, M.L., Taylor, B.C., Diem, S.J., Stone, K.L., Ancoli-Israel, S., Redline, S., et al., 2008. Association between depressive symptoms and sleep disturbances in community-dwelling older men. Journal of the American Geriatrics Society 56 (7), 1228–1235.

Picchietti, D., Allen, R.P., Walters, A.S., et al., 2007. Restless legs syndrome: prevalence and impact in children and adolescents–the Peds REST study. Pediatrics 120, 253–266.

Picchietti, M.A., Picchietti, D.L., Picchietti, M.A., 2008. Restless legs syndrome and periodic limb movement disorder in children and adolescents. Seminars in Pediatric Neurology 15, 91–99.

Rajaram, S.S., Walters, A.S., England, S.J., et al., 2004. Some children with growing pains may actually have restless legs syndrome. Sleep 27, 767–773.

Ram, S., Seirawan, H., Kumar, S.K.S., Clark, G.T., 2010. Prevalence and impact of sleep disorders and sleep habits in the United States. Sleep Breath 14, 63–70.

Reid, K.J., Chang, A.M., Zee, P.C., 2004. Circadian rhythm sleep disorders. Medical Clinics of North America 88, 631–651.

Remulla, A., Guilleminault, C., Remulla, A., Guilleminault, C., 2004. Somnambulism (sleepwalking). Expert Opinion on Pharmacotherapy 5, 2069–2074.

Roehrs, T., Roth, T., 2004. Sleep disorders: an overview. Clinical Cornerstone 6 (Suppl. 1C), 6–16.

Roth, T., Zammit, G., Kushida, C., Doghramji, K., Mathias, S.D., Wong, J.M., et al., 2002. A new questionnaire to detect sleep disorders. Sleep Medicine 3, 99–108.

Schechter, M., 2002. Technical report: diagnosis and management of childhood obstructive sleep apnea syndrome. Pediatrics 109, 1–20.

Shepertycky, M.R., Banno, K., Kryger, M.H., 2005. Differences between men and women in the clinical presentation of patients diagnosed with obstructive sleep apnea syndrome. Sleep 28, 309–314.

Smedje, H., Broman, J.E., Hetta, J., et al., 1999. Parents' reports of disturbed sleep in 5–7-year-old Swedish children. Acta Paediatrica 88, 858–865.

Tarasiu, K.A., Greenberg-Dotan, S., Simon-Tuva, T., et al., 2007. Elevated morbidity and health care use in children with obstructive sleep apnea syndrome. American Journal of Respiratory and Critical Care Medicine 175 (1), 55–61.

Taylor, D.J., Mallory, L.J., Lichstein, K.L., et al., 2007. Comorbidity of chronic insomnia with medical problems. Sleep 30, 213–218.

Walters, A.S., Mandelbaum, D.E., Lewin, D.S., et al., 2000. Dopaminergic therapy in children with restless legs/periodic limb movements in sleep and ADHD. Dopaminergic Therapy Study Group. Pediatric Neurology 22, 182–186.

Watanabe, T., Kajimura, N., Kato, M., Sekimoto, M., Nakajima, T., Hori, T., et al., 2003. Sleep and circadian rhythm disturbances in patients with delayed sleep phase syndrome. Sleep 26 (6), 657–661.

Wiggs, L.D., 2003. Paediatric sleep disorders: the need for multidiscipline. The need for multidisciplinary sleep clinics. International Congress Series 1254, 185–190.

Sleep Disorders: Psychiatric Aspects

Johannes Beck, Center for Affective, Stress and Sleep Disorders (ZASS), Psychiatric Clinics of the University of Basel, Basel, Switzerland
Erich Seifritz, Zurich University Hospital of Psychiatry, Zurich, Switzerland
Edith Holsboer-Trachsler, Center for Affective, Stress and Sleep Disorders (ZASS), Psychiatric Clinics of the University of Basel, Basel, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Disturbed sleep is a risk factor for many diseases, among others cardiovascular and psychiatric disorders, with depression and anxiety disorders being most closely related to insomnia. In healthy young subjects, sleep consists of rapid-eye-movement (REM) and non-REM sleep periods that alternate in a cyclic fashion. Sleep abnormalities in patients with major depressive disorders, for example, can be classified as difficulties of sleep continuity, abnormal sleep architecture, and disinhibition of REM sleep. This article gives an overview of neurobiological mechanisms underlying the sleep changes in psychiatric disorders and presents implications for sleep related biomarkers of major depression.

Introduction

Complaints of too little (insomnia) or too much sleep (hypersomnia), or of sleep that is not restorative enough, are termed dyssomnias. About one-third of the general population reports insomnia symptoms, with an increased prevalence in women and the elderly. Disturbed sleep is a risk factor for many diseases, among others type II diabetes mellitus, cardiovascular diseases (Grandner et al., 2012), and psychiatric disorders, with depression and anxiety disorders being most closely related to insomnia. In addition, poor sleep leads to impaired daytime performance in several cognitive domains and a strongly enhanced proneness to accidents. Since chronic dyssomnia most often occurs as a comorbid diagnosis, a thorough evaluation of the patient and his/her sleep complaint is a prerequisite for accurate diagnosis and effective treatment.

The diagnosis of insomnia is based upon the patient's subjective complaint of sleeping too little. Patients report difficulties in initiating or maintaining sleep or of non-restorative sleep, i.e., not feeling well rested after sleep that is apparently adequate in amount. Insomnia may occur as primary or secondary disorder due to another psychiatric condition, a general medical condition, and/or substance misuse. The prevalence of primary insomnia relative to secondary insomnia is actually quite rare.

Hypersomnia includes complaints of excessive sleepiness characterized either by prolonged sleep episodes and/or excessive sleepiness during the day. These symptoms may interfere with social, occupational, and/or other areas of functioning. Like insomnias, hypersomnia may occur as a primary or secondary disorder due to a psychiatric and/or medical condition, and/or substance misuse. Compared to primary sleep disorders, the secondary sleep disturbances are more common among patients in a psychiatric practice or hospital. This is particularly important to bear in mind when evaluating a patient with sleep complaints. Dyssomnia is often associated with psychiatric disorders, such as depression, schizophrenia, anxiety disorders, or personality disorders, and misuse of drugs and/or alcohol. Whenever possible, it is clinically useful to assess which is the primary and which is the secondary disorder. This will guide the treatment and improve preventive measures.

Frequency and Risk Factors

Sleeping problems are among the most frequent complaints a general practitioner must deal with. According to epidemiological studies, 19–46% of the population report sleeping problems, of which 13% suffer from moderate or severe disturbances. In terms of the narrow definition of diagnostic criteria, that is, initial insomnia, interrupted sleep, and disturbance of daily well-being, 1.3% of the population suffers from somnipathy (Holsboer-Trachsler and Kocher, 1996). An aging society, the lifestyle of modern civilization including tension, stress, solitude, shift work, and liability for stress related disorders are major risk factors for disturbed sleep. It has been observed that insomniacs reported more medical problems than good sleepers and had twice as many doctor's visits yearly than good sleepers. Furthermore, in severe insomniacs compared to good sleepers, it was found that insomniacs were hospitalized twice as often as good sleepers over the previous 12 months. It was also observed that insomniacs consumed more medication than good sleepers for various treatments.

A World Health Organization (WHO) collaborative study in 15 different countries found that insomnia is more common in females and increases with age. Indeed, it is thought that half of the population over 65 years of age suffers from chronic sleep disturbances. The WHO study also found that 51% of people with an insomnia complaint had a well-defined International Classification of Diseases-10 (ICD-10) mental disorder (e.g., depression, anxiety, alcohol problems). Two prospective studies in a population of younger and elderly subjects showed that individuals who suffered from insomnia were more than three times as likely to develop a major depression compared to those without sleep problems. A similar result was observed for alcoholism.

In particular, depression and anxiety disorders show a high comorbidity with insomnia. For example, insomnia is present in 70% of mood disorders, preceding the emergence of the latter in nearly half of the cases and not only incurs risk for onset, but also for relapse, recurrence, severity, and treatment of a depressive episode. While disturbed sleep has traditionally been viewed as a mere symptom of these disorders, more recent evidence suggests that insomnia and psychiatric diseases may rather be seen as separate but closely inclined disorders.

Regulation of Normal Sleep

The quantity and quality of sleep is measured with polysomnographic recordings (PSG) that include the electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG). In healthy young subjects, sleep consists of rapid-eye-movement (REM) and non-REM sleep periods that alternate in a cyclic fashion. During the transition from drowsiness to light sleep, a slowing of EEG activity occurs, which marks sleep stage 1. In the following sleep stage 2, sleep spindles and K-complex wave forms appear. Slow-wave sleep (SWS) is characterized by synchronized slow waves that define the sleep stages 3 and 4 (Rechtschaffen and Kales, 1968). REM sleep is characterized by a faster EEG activity, horizontal rapid eye movements, and hypotonus of skeletal muscles (**Figure 1**). The first REM sleep period appears usually after approximately 90 min after the first non-REM period. The major amount of SWS is found before the first REM period. While the first REM period is relatively short, duration of later REM periods increases. One sleep cycle consists of one episode of non-REM sleep and one of REM sleep. In healthy subjects, four to five sleep cycles occur during the night. Growth hormone (GH) is usually secreted during the first non-REM period, whereas cortisol secretion rises during the second half of the night and appears to be associated with increased amount of REM sleep.

At least three major processes are involved in the regulation of normal sleep. According to the two-process model of sleep regulation (Borbély and Achermann, 1999), sleep and wakefulness are influenced by both a homeostatic and a circadian process that interact in a complex way. The homeostatic Process S (related to SWS) increases with the duration of wakefulness prior to sleep onset and augments sleep propensity. The circadian Process C, driven by the internal clock located in the suprachiasmatic nuclei, describes the daily cycle of sleepiness and wakefulness and is related to REM sleep. The third process is the ultradian rhythm of non-REM and REM sleep.

On the neuroanatomical level, the different sleep states are triggered by distinct sets of respective brain regions and local activation patterns of groups of neurons in which sleep-regulating endogenous substances modulate sleep–wake behavior (Saper et al., 2005). While wakefulness is modulated

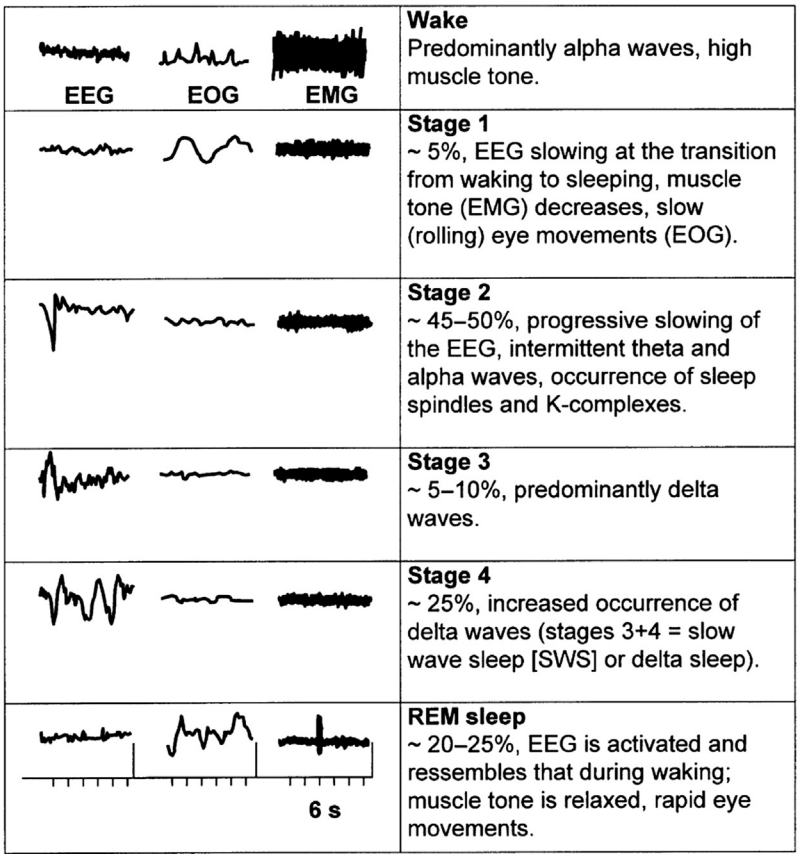

Figure 1 Sleep stages. Rechtschaffen, A., Kales, A., 1968. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Department of Health, Education and Welfare, Neurological Information Network, Bethesda.

by a variety of factors, sleep is promoted by a relatively small amount of sleep-promoting substances including neuropeptides, biogenic amines, prostaglandins, and cytokines. From an evolutionary point of view, the multifactorial sleep regulation makes perfect sense, since dependence on only one specific neuroanatomical location or single factor would render the vital process of sleeping much more vulnerable. The brain stem and the hypothalamus are particularly important brain regions in sleep–wake regulation. Wakefulness is promoted by the ascending arousal system, a network of cell groups that activate the thalamus, and cerebral cortex. Cholinergic nuclei in the upper pons; monoaminergic cell groups including the tuberomammillary nucleus, raphe nuclei, and locus coeruleus; peptidergic neurons in the lateral hypothalamus; and Gamma-aminobutyric acid (GABA)ergic neurons of the basal forebrain contribute to this system (Figure 2). During non-REM sleep, this system is shut off by a key switch in the hypothalamus, the ventrolateral preoptic nucleus (VLPO) (Figure 3). GABAergic and galalinergic neurons in the VLPO are active during non-REM sleep. The cyclic occurrence of REM and non-REM episodes is regulated by a reciprocal interaction of REM-on and REM-off neurons of the brain stem. Pontine cholinergic or glutamatergic REM-on cells induce and maintain REM sleep

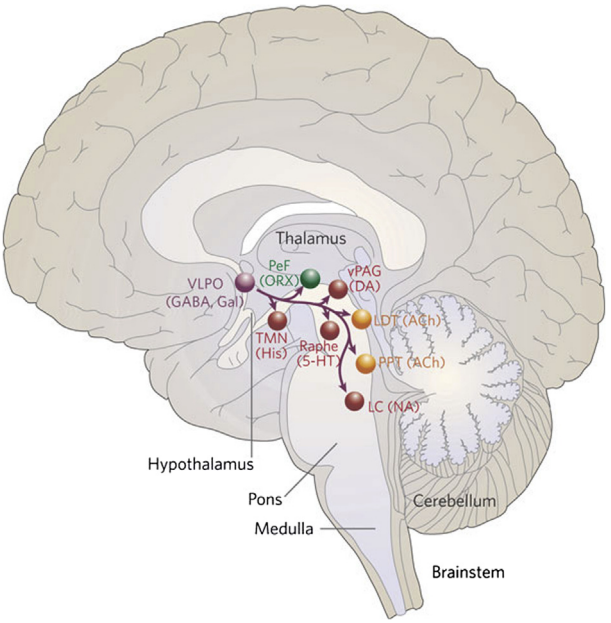

Figure 3 A schematic drawing to show the key projections of the ventrolateral preoptic nucleus (VLPO) to the main components of the ascending arousal system. It includes the monoaminergic cell groups (red) such as the tuberomammillary nucleus (TMN), the A10 cell group, the raphe cell groups and the locus coeruleus (LC). It also innervates neurons in the lateral hypothalamus (LHA; green), including the perifornical (PeF) orexin (ORX) neurons, and interneurons in the cholinergic (ACh) cell groups (yellow), the pedunculopontine (PPT) and laterodorsal tegmental nuclei (LDT). Note that the VLPO neurons lie within the region outlined by von Economo for the anterior hypothalamic lesion that caused insomnia. 5-HT, serotonin; GABA, γ-aminobutyric acid; gal, galanin; NA, noradrenaline; His, histamine. From: Saper, C.B., Scammell, T.E., Lu, J. 2005. Hypothalamic regulation of sleep and circadian rhythms. Nature 437 (7063), 1257–1263.

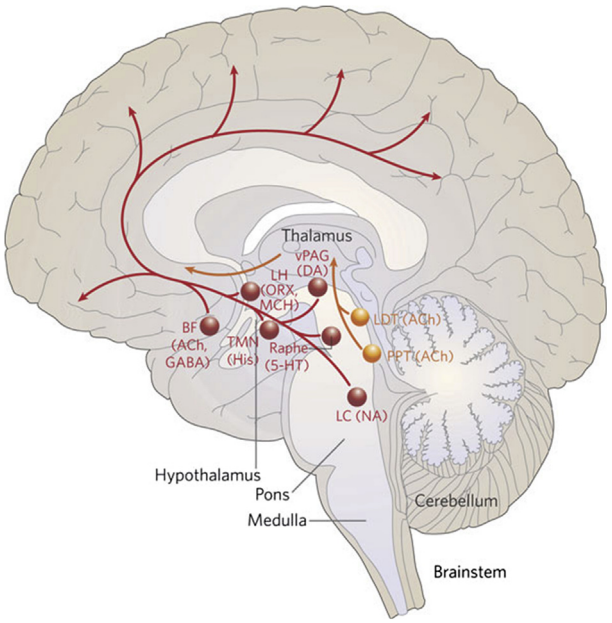

Figure 2 A schematic drawing showing key components of the ascending arousal system. A major input to the relay and reticular nuclei of the thalamus (yellow pathway) originates from cholinergic (ACh) cell groups in the upper pons, the pedunculopontine (PPT), and laterodorsal tegmental nuclei (LDT). These inputs facilitate thalamocortical transmission. A second pathway (red) activates the cerebral cortex to facilitate the processing of inputs from the thalamus. This arises from neurons in the monoaminergic cell groups, including the tuberomammillary nucleus (TMN) containing histamine (His), the A10 cell group containing dopamine (DA), the dorsal and median raphe nuclei containing serotonin (5-HT), and the locus coeruleus (LC) containing noradrenaline (NA). From: Saper, C.B., Scammell, T.E., Lu, J., 2005. Hypothalamic regulation of sleep and circadian rhythms. Nature 437 (7063), 1257–1263.

by phasic activation of occipital cortical regions by the thalamic lateral geniculate body by ponto–geniculate–occipital waves. During wakefulness and non-REM sleep, REM-on cells are inhibited by GABAergic neurons of the deep mesencephalic reticular neurons. A simplified neuroanatomy of sleep–wakefulness is shown in Figure 3. Functional imaging studies revealed that the brain exhibits widespread deactivations during SWS, while REM sleep is associated with increased cerebral blood flow in the thalamus, visual areas, and limbic regions, as well as reduced activity in the dorsal lateral prefrontal cortex (DLPFC), parietal cortex, and the precuneus. These activation patterns seem to correlate with dreaming. While the activation of visual association areas corresponds to dream imagery, activation of the amygdala is in line with dream emotionality, and deactivation of the DLPFC is in accordance with the lack of insight and volition typically experienced in dreams.

Extensive evidence suggests that an important function of sleep is to restore neural capacity within cells and neurocircuitries. Hormones serving restoration and cell growth like GH and prolactin increase their blood level during sleep and important immunological processes are enhanced during sleep. Sleep is not only essential for neural restoration, but also

for memory consolidation and integration (Diekelmann and Born, 2010), emotion regulation, and the preparation for waking life with distinct bidirectional links between sleep and metabolism systems.

Sleep and Psychiatric Disorders

Polygraph sleep research helped paving the way to the modern age of scientific, biological psychiatry. Much of the early sleep research beginning in the mid-1960s was descriptive mapping of events taking place during sleep and laid out the currently rich picture of polysomnographic features in the different psychiatric disorders. Although many of the objective sleep abnormalities associated with psychiatric disorders appear not to be diagnostically specific, these studies have also established sleep measures as neurobiological windows into the underlying pathophysiology associated with psychiatric illnesses. Psychiatric disorders are among the most common primary causes of secondary sleep complaints, particularly of insomnia. Sleep abnormalities may be caused by central nervous system abnormalities associated with psychiatric illnesses as well as by accompanying behavioral disturbances. Patients with depression, anxiety disorders, misuse of alcohol or drugs, schizophrenia, dementia, and personality disorders may complain of difficulty in falling asleep, staying asleep, or of inadequate sleep. Although specific sleep patterns are not necessarily diagnostic of particular psychiatric disorders, there are relationships between certain sleep abnormalities and categories of psychiatric disorders (**Table 1**). Reviews of the literature on sleep in psychiatric disorders documented that findings for patients with affective disorders differed most frequently from those of normal controls (Benca et al., 1992; Steiger and Kimura, 2010).

Sleep Abnormalities in Depression

Among patients with depression, more than 70% complain of insomnia, whereas some suffer from hypersomnia. Sleep abnormalities in patients with major depressive disorders, as assessed by laboratory studies, can be classified as difficulties of sleep continuity, abnormal sleep architecture, and disruptions in the timing of REM sleep. Sleep initiation and maintenance difficulties include prolonged sleep latency (sleep onset insomnia), intermittent wakefulness and sleep fragmentation during the night, early morning awakenings with an inability to return to sleep, reduced sleep efficiency, and decreased total sleep time. With regard to sleep architecture, abnormalities have been reported in the amounts and distribution of non-REM sleep stages across the night, and include increased shallow stage 1 sleep and reductions in the amount of deep, slow-wave (stages 3 and 4) sleep. REM sleep disturbances in depression include a short REM latency (<65 min), a prolonged first REM sleep period, and an increased total REM sleep time, particularly in the first half of the night. While REM latency is clearly affected by age, REM density, in contrast, shows no variation with age, and is enhanced in all investigated age groups in depressed patients. Two studies comparing groups of healthy volunteers and of patients with major depression, anorexia nervosa, and bulimia showed that REM density was enhanced in the depressed patients only, whereas other sleep-EEG variables did not differ. The comparison of patients with depression, never depressed patients with a panic disorder, and healthy volunteers showed that REM sleep duration and REM density were elevated and SWS was reduced only in the patients with depression. REM latency was shortened in both groups of patients when compared to the healthy volunteers and thus appears to be less specific for depression (Lauer et al., 1992).

In the Munich vulnerability study, a prospective high-risk design was applied, that aimed to identify premorbid vulnerability factors for affective disorders (Lauer et al., 1995). Subjects with a positive family history for affective disorders among first degree relatives were investigated. These high-risk individuals who up to the time of the study examination had no psychopathological sign or symptom of depression were compared to healthy volunteers with a negative family history for this or other psychiatric disorders. In the high-risk subjects at the index investigation, REM density was enhanced and the amount of SWS during the first non-REM period was decreased. At a follow-up about 4 years later, these observations were stable (Modell et al., 2002). Every forth subject of the initial sample of high-risk subjects of this study developed an affective disorder during the follow-up period, which is more than twofold the average disease risk. This supports the view that specific traits are closely associated with inherited risk for depression. In these affected high-risk subjects, an increased REM density during the total night and during the first REM period was found in the premorbid sleep EEG. These observations confirm that increased REM density meets all requirements of a vulnerability marker for affective disorder, as it is found in patients with acute depression and in patients after remission. It occurs in healthy high-risk subjects for affective disorders. It is stable over time, and it is predictive for the occurrence of depression and relapse (Hatzinger et al., 2004). REM density has thus been proposed as an endophenotype in family studies and as a biomarker identifying those patients that benefit from specific treatments (Holsboer and Ising, 2010). Simultaneous investigation of sleep EEG and nocturnal hormone secretion showed that in young healthy subjects during the first half of the night SWS and the major peak of GH secretion during 24 h preponderate, whereas the

Table 1 Typical sleep-EEG findings in psychiatric disorders

	Sleep continuity	*SWS*	*REM sleep*	*Hypersomnia*
Depression/dysthymia	↓↓↓	↓↓	↑↑↑	↕
Schizophrenia	↓↓↓	↓↓	↑↑	↕
Anxiety disorders	↓	↔	↔	↔
Alcohol dependency	↓↓	↓↓↓	↑	↔
Personality Borderline disorders	↓	↔	↑	↔
Dementias	↓↓↓	↓↓↓	↔	↕
Eating disorders	↓	↔	↔	↔

Meta-analysis by Benca, R.M., Obermayer, W.H., Thisted, R.A., Gillin, J.C., 1992. Sleep and psychiatric disorders: a meta-analyis. Archives of General Psychiatry 49, 651–668.

second half of the night is dominated by REM sleep and the secretion of the hormones of the hypothalamic–pituitary–adrenocortical (HPA) system, corticotrophin (ACTH) and cortisol. During normal aging and during an episode of depression, changes of sleep-endocrine activity occur. As outlined above, sleep-EEG changes in depression include disinhibition of REM sleep (shortened REM latency, prolonged first REM period, elevated REM density), impaired non-REM sleep (reduced time spent in sleep stage 2 and SWS), and changes of sleep continuity (prolonged sleep latency, frequent nocturnal awakenings, and early morning awakening). Similarly, during aging more shallow sleep develops. ACTH and cortisol secretion are elevated in depressed patients. The amplitude of the nocturnal course of these hormones is dampened in the elderly resulting in an elevated nadir. GH secretion is blunted during aging and in a depressive episode as well. The sleep-endocrine pattern in young healthy subjects and its changes related to depression and aging suggest that there exist endogenous factors, which are common regulators of sleep EEG and hormones, linking, for example, SWS and GH or HPA activity and REM sleep, respectively (Steiger and Holsboer, 1997). The integrated model of normal and impaired peptidergic sleep regulation is summarized in **Figure 4**.

Sleep Abnormalities in Anxiety and Post-traumatic Stress Disorder

Anxiety disorders are often associated with sleep disturbances that exacerbate the symptoms and contribute to relapses (Mellmann, 2006). Anxiety and sleep problems are frequently comorbid conditions particularly in older people (Magee and Carmen, 2010). Post-traumatic stress disorder (PTSD), a chronic disorder caused by a traumatic event and characterized by hyperarousal, intrusions, and avoidance of trauma-related stimuli is closely related to sleep (American Psychiatric Association, 2000). Sleep disorders such as insomnia and nightmares often occur in PTSD and may exacerbate PTSD symptoms during daytime and sleep (Spoormarker and Montgomery, 2008). Insomnia is often a residual symptom after treatment of PTSD (Nappi et al., 2012). Sleep related breathing disorders and periodic limb movements are also prevalent in PTSD. A meta-analysis of polysomnographic studies in PTSD found reduced SWS, increased stage 1 sleep, and increased REM density in PTSD patients compared to controls (Kobayashi et al., 2007). PTSD patients often show REM fragmentation, which includes more but shorter REM periods that have been found to be predictive of PTSD symptom severity. Since involuntary memory intrusions are key symptoms in PTSD, drug treatment of PTSD-associated sleep disturbances may also impact on flashback of traumatized memories (Nappi et al., 2012).

Neurobiology of Sleep: Relevance to Psychiatric Disorders

Inspired by the growing knowledge of the underlying neurobiology of sleep, investigators proposed theories into the pathophysiology of psychiatric disorders. Because depression has been studied more than any other psychiatric syndrome in recent decades, the models have attempted to explain features of sleep in depression, such as short REM latency, decreased delta sleep, and the antidepressant effects of sleep deprivation in depressed patients. Among the most prominent models of sleep changes in depression are the cholinergic–aminergic imbalance hypothesis for depression, the two-process Model of sleep regulation, the phase-advance Hypothesis, the overarousal hypothesis, the REM sleep hypothesis, and the contribution of the HPA system to sleep-EEG changes in depression. Moreover, some of the sleep-EEG parameters have been suggested as biomarkers of treatment response and long-term course of depression.

The 'cholinergic–aminergic imbalance hypothesis' (Janowsky et al., 1972) for depression postulates that depression arises from an increased ratio of cholinergic to aminergic neurotransmission in critical central synapses. Because various features of the sleep of depressed patients (decreased REM latency and delta sleep) have been simulated in normal volunteers by pharmacological probes, the reciprocal interaction hypothesis from basic sleep research and the cholinergic–aminergic imbalance model from clinical psychiatry have been correlated. The reciprocal interaction model assumes that the cycling alternating pattern of non-REM and REM sleep is under the control of noradrenergic/serotonergic and cholinergic neuronal networks. Linking these concepts, it is suggested if depression results from diminished noradrenergic and serotonergic neurotransmission, cholinergic activity would be expected to increase and lead therefore to the sleep disturbances of depression.

Among other neurotransmitters, serotonin plays a role in the pathophysiology of depression and its treatment; most antidepressants ultimately lead to an enhancement of serotonergic neurotransmission that is believed to be associated with clinical improvement. Depression is associated with a disinhibition of REM sleep (shortened REM latency, increased REM density) and serotonin leads to a suppression of REM sleep. Furthermore, there is evidence that the antidepressant effect of sleep deprivation is related to a modification of serotonergic neurotransmission, thus sleep regulation and depression share common physiopathological mechanisms at the serotonergic level. According to the 'two-process model of sleep regulation,' depression is thought to result from, or is associated with, a deficiency of Process S. This model was supported by evidence that EEG power density in the delta frequencies was decreased during sleep in depressed patients compared with controls and by the fact that sleep deprivation increases both Process S and mood. In the extended 'two-process model of sleep regulation,' the interaction of hormones with EEG features has been integrated. Preclinical investigations and studies in young and elderly normal controls and in patients with depression demonstrate that neuropeptides play a key role in sleep regulation. As an example, growth hormone-releasing hormone (GHRH) is a common stimulus of SWS and GH release, whereas corticotropin-releasing hormone (CRH) exerts opposite effects. It is suggested that an imbalance of these peptides in favor of CRH contribute to changes in sleep EEG and endocrine activity during depression. Based on the findings that the HPA axis is disregulated in depression, and that CRH produces EEG-sleep changes reminding to depression on one

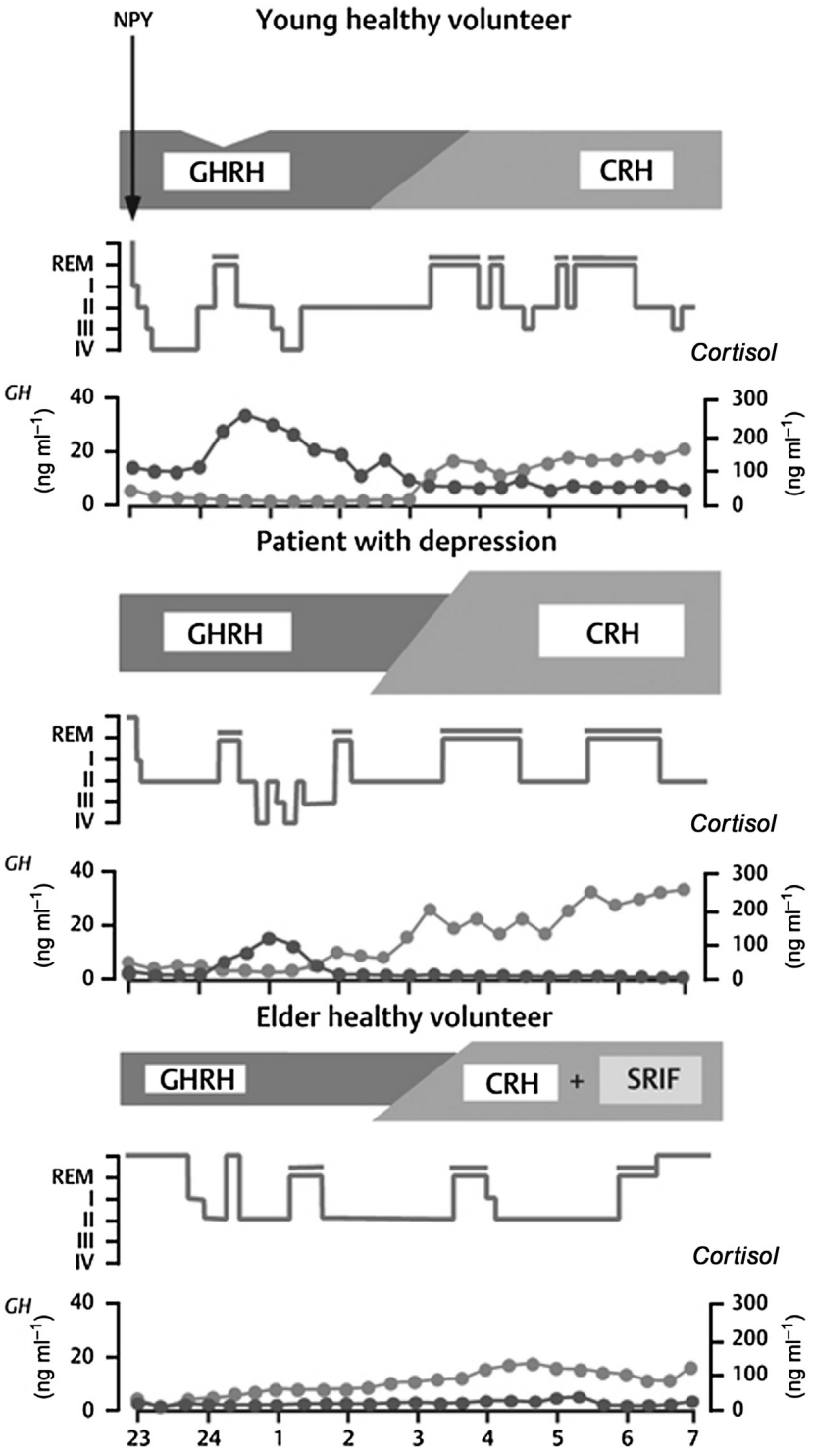

Figure 4 Model of normal and impaired peptidergic sleep regulation. During the first half of the night the influence of GHRH preponderates resulting in the high amounts of SWS and GH. During the second half of the night CRH and its cofactor somatostatin dominate. Changes in the GHRH/CRH balance during depression and aging result in similar alternations of sleep-endocrine activity. GHRH = growth hormone-releasing hormone, CRH = corticotrophin-releasing hormone, SRIF = somatostatin. Adapted from: Steiger, A., 1995. Schlafendokrinologie. Der Nervenarzt 66, 15–27, Figure 2.

hand (SWS reduction, REM sleep disinhibition), and that the somatotrophic system reciprocally interacts with the HPA system (decreased GHRH and SWS), it has been postulated that deficient Process S is associated with deficient GHRH and an overdrive of CRH. Although abnormally high values of cortisol secretory activity normalize after recovery from depression, GH release and several characteristic disturbances of the sleep EEG may remain unchanged. These putative trait-dependent alterations suggest that strategies aiming at restoration of these sleep changes are worthy of exploration for their potential antidepressant effect.

The depressed state appears not only to be associated with a disturbance of sleep–wake homeostasis, but also of circadian function or of their interaction. The 'phase-advance hypothesis' (MacLean et al., 1986) suggests that the phase position of the underlying circadian oscillator is 'phase advanced' relative to the external clock time. This is supported by studies showing that short REM latency (one of the key sleep-EEG changes in depression) could be simulated in normal controls by appropriate phase shifts of hours in bed.

As mentioned above, sleep deprivation has potent antidepressant effects in more than half of all depressed patients. This observation has prompted the hypothesis that depressed patients are 'overaroused.' Later, the 'overarousal hypothesis' has been forwarded and was supported by psychological self-rating studies suggesting that clinical improvement to sleep deprivation in depression may be associated simultaneously with subjective feelings of more energy (arousal), less tension, and more calmness (dearousal). Other data consistent with this hypothesis include the short, shallow, and fragmented sleep patterns; lowered arousal thresholds; and elevated nocturnal core body temperature often seen in depressed patients. This hypothesis has been tested by means of studies of localized cerebral glucose metabolism with 18F-deoxyglucose positron emission tomography in separate studies in depression, sleep deprivation, and the first non-REM period of the night. In these studies, it was found that elevated brain metabolism prior to sleep deprivation predicted clinical benefits in depressed patients and that normalization of these measures is associated with clinical improvement. It was also found that local cerebral glucose metabolism in cingulate and amygdala at baseline was significantly higher in clinical responders than in nonresponders or normal controls. Furthermore, it was shown that glucose metabolic rate was increased during the first non-REM period in depressed patients compared with normal controls. Moreover, these studies demonstrated significant 'hypofrontality,' that is, reduced ratio of frontal to occipital activity, compared with normal controls.

The 'REM sleep hypothesis' of depression has been based on the findings that REM sleep is enhanced in depression and that virtually all antidepressant drugs suppress REM sleep. Early, however not replicated, studies have shown that a sustained remission from depression may be achieved by selective REM sleep deprivation carried out repeatedly every night during about 2 weeks. This treatment modality was not followed because the long-term REM sleep deprivation was too demanding exhausting for the patients. The so-called 'sleep-phase advance' protocol that is scheduling sleeping times in a way that minimizes the occurrence of REM sleep, has been shown to produce a sustained antidepressant effect. However, also this treatment modality is demanding for both the patients and the institutional staff and therefore rarely applied.

Sleep EEG and HPA System: Markers of Antidepressant Treatment Outcome

Some of the above mentioned sleep-EEG variables were suggested as biomarkers and predictors of antidepressant therapy response and clinical outcome (Steiger and Kimura, 2010; Baskaran et al., 2012). Increase of REM latency after 2 days of treatment with amitriptyline was correlated with the clinical outcome after 4 weeks. In a prospective study with the selective serotonin reuptake inhibitor (SSRI) paroxetine and the serotonin reuptake enhancer tianeptine, REM density after 1 week of treatment was a predictor of treatment response (Murck et al., 2003). A follow-up study in depressed patients who had previously participated in an earlier trial with trimipramine treatment revealed associations between sleep-EEG parameters and treatment outcome as well as the long-term course of depression: low sleep continuity, low SWS, and elevated REM density were associated with the number of previous depressive episodes. Increased REM density and decreased SWS at the end of the treatment trial were related to an increased recurrence rate during the follow-up period, indicating a strong association between long-term course of depression and these markers of the sleep EEG (Hatzinger et al., 2004). These sleep-EEG variables were related to impaired HPA system regulation as shown by abnormal dexamethasone/corticotropin-releasing hormone (DEX/CRH) test results. Noteworthy, another study suggested that a failure to normalize the DEX/CRH test within 2.5 weeks of antidepressant treatment is an independent predictor for nonresponse (Ising et al., 2007). Interestingly, associations between unfavorable sleep-EEG patterns, enhanced HPA activity, and more difficult behavioral psychosocial dimension were found already in otherwise healthy kindergarten children (Hatzinger et al., 2008), indicating an interplay between sleep and HPA axis regulation in a dimensional way rather than only in diagnostic categories of health versus disease. Recently, parameters of EEG cordance during REM sleep, a quantitative EEG measure combining complementary information from absolute and relative power of EEG spectra, were found as an early marker of treatment response that differentiated between responders and nonresponders to antidepressant pharmacotherapy.

Future Perspectives

We have outlined just some aspects of sleep research that are relevant to depression. Linking basic and clinical approaches has been one of the 'royal roads' to the neurobiological underpinnings of psychiatric diseases and their treatment in the past, and to narrow the gap between bench and bedside. In addition to whole night PSG recordings, the search for biomarkers in daytime naps might be a promising research strategy, since it has been shown that even very short naps lead to relapses in patients profiting from therapeutic sleep deprivation. Such short sleep periods thus appear to be very closely

related to the neurobiological mechanisms underlying depression.

Based on the findings that sleep-EEG variables like REM density or certain clusters of variables including markers of HPA axis regulation are able to predict treatment response and long-term course of the disease, there is hope that future treatment of depression will be improved with the support of biomarkers that better reflect the underlying pathology than diagnostic attributions that are based only on clinical symptoms (Holsboer, 2008). A combination of biomarkers derived from polysomnography, complemented by neuroendocrine testing, brain imaging, and genomics are promising tools that may ultimately guide treatment strategies that are tailored to the individual patient.

See also: Alcohol-Related Disorders; Anxiety and Anxiety Disorders; Depression; Hypnosis, Psychiatry of; Post-Traumatic Stress Disorder.

Bibliography

American Academy of Sleep Medicine, Iber, C., Anoli-Israel, S., Chesson, A., Quan, S.F., 2007. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine, Westchester, IL.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. DSM-IV-TR®. American Psychiatric Association, Washington, DC.

Baskaran, A., Milev, R., McIntyre, R.S., 2012. The neurobiology of the EEG biomarker as a predictor of treatment response in depression. Neuropharmacol 63, 507–513.

Benca, R.M., Obermayer, W.H., Thisted, R.A., Gillin, J.C., 1992. Sleep and psychiatric disorders; a meta-analyis. Archives of General Psychiatry 49, 651–668.

Borbély, A.A., Achermann, P., 1999. Sleep homeostasis and models of sleep regulation. Journal of Biological Rhythms 14 (6), 557–568.

Diekelmann, S., Born, J., 2010. The memory function of sleep. Nature Reviews Neuroscience 11, 114–126.

Grandner, M.A., Jackson, N.J., Pak, V.M., Gehrman, P.R., 2012. Sleep disturbance is associated with cardiovascular and metabolic disorders. Journal of Sleep Research 21, 427–433.

Hatzinger, M., Hemmeter, U.M., Brand, S., Ising, M., Holsboer-Trachsler, E., 2004. Electroencephalographic sleep profiles in treatment course and long-term outcome of major depression: association with DEX/CRH-test response. Journal of Psychiatric Research 38, 453–465.

Hatzinger, M., Brand, S., Perren, S., Stadelmann, S., von Wyl, A., von Klitzing, K., et al., 2008. Electroencephalographic sleep profiles and hypothalamic–pituitary–adrenocortical (HPA) activity in kindergarten children: early indication of poor sleep quality associated with increased cortisol secretion. Journal of Psychiatric Research 42, 532–543.

Holsboer, F., 2008. How can we realize the promise of personalized antidepressant medicines? Nature Reviews Neuroscience 9, 638–646.

Holsboer, F., Ising, M., 2010. Stress hormone regulation: biological role and translation into therapy. Annual Review of Psychology 61, 81–109.

Holsboer-Trachsler, E., Kocher, R., 1996. Somnipathies: new recommendations for their diagnosis and treatment. Drugs of Today 32 (6), 477–482.

Ising, M., Horstmann, S., Kloiber, S., Lucae, S., Binder, E.B., Kern, N., Künzel, H.E., Pfennig, A., Uhr, M., Holsboer, F., 2007. Combined DEX/CRH test predicts treatment response in major depression – a potential biomarker? Biological Psychiatry 62, 47–54.

Janowsky, D.S., el-Yousef, M.K., Davis, J.M., Sekerke, H.J., 1972. A cholinergic-adrenergic hypothesis of mania and depression. Lancet 2, 632–635.

Kobayashi, I., Boarts, J.M., Delahanty, D.L., 2007. Polysomnographically measured sleep abnormalities in PTSD: a meta-analytic review. Psychophysiol 44, 660–669.

Lauer, C.J., Krieg, J.C., Garcia-Borreguero, D., Özdaglar, A., Holsboer, F., 1992. Panic disorder and major depression: a comparative electroencephalogramic sleep study. Psychiatry Research 44, 41–54.

Lauer, C.J., Schreiber, W., Holsboer, F., Krieg, J.C., 1995. In quest of identifying vulnerability markers for psychiatric disorders by all-night polysomnography. Archives of General Psychiatry 52, 145–153.

MacLean, A.W., Knowles, J.B., Vetere, C., 1986. REM sleep and depression: further use of computer simulation to test the phase-advance hypothesis. Psychiatry Research 19, 25–36.

Magee, J.C., Carmen, C.N., 2010. The relationship between sleep and anxiety in older adults. Current Psychiatry Reports 12, 13–19.

Mellmann, T.A., 2006. Sleep and anxiety disorders. Psychiatric Clinics of North America 29, 1047–1058.

Modell, S., Ising, M., Holsboer, F., Lauer, C.J., 2002. The Munich vulnerability study on affective disorders: stability of polysomnographic findings over time. Biological Psychiatry 52, 430–437.

Murck, H., Nickel, T., Künzel, H., Antonijevic, I.A., Schill, J., Zobel, A., et al., 2003. State markers of depression in sleep EEG: dependency on drug and gender in patients treated with tianeptine or paroxetine. Neuropsychopharmacol 28, 348–358.

Nappi, C.M., Drummond, S.P., Hall, J.M., 2012. Treating nightmares and insomnia in posttraumatic stress disorder: a review of current evidence. Neuropharmacol 62, 576–585.

Rechtschaffen, A., Kales, A., 1968. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Department of Health, Education and Welfare, Neurological Information Network, Bethesda.

Saper, C.B., Scammell, T.E., Lu, J., 2005. Hypothalamic regulation of sleep and circadian rhythms. Nature 437 (7063), 1257–1263.

Spoormarker, V.I., Montgomery, P., 2008. Disturbed sleep in post-traumatic stress disorder: secondary symptom or core feature? Sleep Medicine Reviews 12, 169–184.

Steiger, A., Holsboer, F., 1997. Neuropeptides and human sleep. Sleep 20, 1038–1052.

Steiger, A., Kimura, M., 2010. Wake and sleep EEG provide biomarkers in depression. Journal of Psychiatric Research 44, 242–252.

Relevant Websites

www.sleepassociation.org/ – American Sleep Association.
www.psych.org/ – American Psychiatric Association.
www.mpipsykl.mpg.de/ – Max Planck Institute of Psychiatry.

Sleep: Neural Systems

Michael C Antle, Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.A. Hobson, volume 21, pp. 14178–14185, © 2001, Elsevier Ltd.

Abstract

Our 24-hour day is divided into periods of wake and sleep. Sleep itself is divided into periods of rapid eye movement (REM) sleep and non-REM sleep. A large collection of brain areas exist to produce wake and the various sleep states. These brain areas are networked together to coordinate switching between these states. Homeostatic and circadian factors also influence these sleep areas so that sleep occurs at optimal times to meet physiological needs.

Sleep

We spend approximately one-third of our lives asleep. Sleep is a ubiquitous behavior in mammals and birds, and is found in most other vertebrates. Sleep-like states are even observed in invertebrates such as the fruit fly and pond snail. Although the purpose of sleep is debated, it is clear that it serves a necessary role. Sleep is homeostatically regulated, as sleep loss results in an increase drive to engage in sleep and increased sleep duration once sleep is permitted. Furthermore, total sleep deprivation in rodents eventually leads to death after about 3 weeks.

Definitions

Sleep is a natural state of reduced or absence of consciousness. It is characterized by the following features: (1) a species-typical sleep posture, (2) reduced responsiveness to stimuli, (3) reduced movement, and (4) easily reversible (cf hibernation or coma). While sleep is frequently perceived of as synonymous for 'not awake,' it is a heterogeneous behavior, consisting of two major states: rapid eye movement sleep (REM) and non-REM sleep (NREM), each of which exhibits a variety of unique physiological phenomena. Furthermore, organisms typically oscillate between these two sleep states with a regular period. Modern sleep researchers use electroencephalography (EEG) in conjunction with electromyography and electrooculography to objectively define the various stages of sleep based on characteristic brain oscillations, muscle tone, and eye movements. Respiration, cardiovascular function, body temperature, and hormonal activity are also altered depending on the particular stage of sleep.

Rapid Eye Movement Sleep

REM sleep is characterized by rapid and ballistic medial-lateral movements of the eyes. The EEG pattern closely resembles that of an awake individual with high-frequency and low-amplitude activity. For this reason, REM sleep is also known as paradoxical sleep, as the individual is asleep despite exhibiting a wake-like EEG. REM sleep is also characterized by sequential activation in the pons, lateral geniculate nucleus (LGN), and finally occipital cortex, leading to characteristic ponto-geniculo-occipital (PGO) waves. While in REM sleep we experience muscle atonia in our postural muscles. Breathing rate becomes irregular although it generally increases. Body temperature regulation is diminished during REM sleep. Heart rate and blood pressure increase during REM sleep and exhibit more variability in their rate. Dreams are most frequent and vivid during REM sleep.

Non-REM Sleep

NREM sleep is characterized by gradual slowing of EEG frequency and increases in EEG amplitude. It can be subdivided into various stages, depending on the EEG features. Light sleep (Stage 1) exhibits a slight slowing in EEG frequency compared to waking EEG. Stage two exhibits even more slowing and increased amplitude. Additionally, Stage 2 sleep also exhibits brief high-frequency (12–14 Hz) bursts known as sleep spindles, as well as large voltage low frequency waves known as K-complexes. Stages 3 and 4 are also known as slow wave sleep and are characterized by high-amplitude low frequency oscillations. Eye movements are slow and rolling rather than rapid and ballistic. Body temperature, heart rate, and blood pressure all decrease during NREM sleep, and growth hormone is secreted. Muscle tone is maintained during NREM sleep. There is little dreaming activity in NREM sleep, and typically cognitive content is less than what is experienced during REM sleep.

Historical Investigations

Frederick Bremer conducted the first studies examining the neural regulation of sleep (Bremer, 1935). His passive theory of sleep posited that fatigue diminished sensory input to the central nervous system, and that sleep resulted from this decreased sensory input. He used transections of the cat central nervous system to separate the cortex where EEG was recorded from its various inputs. In the encéphale isolé (isolated brain) preparation a cut was made at the level of the medulla, which preserved all input from the cranial nerves. In this case, normal patterns of sleep and wake EEG were observed. Bremer's second preparation was the cerveau isolé (isolated forebrain), which was performed by making the transaction through the midbrain. In this case, there was preserved sensory input from only the olfactory and optic nerves, and a persistent slow wave sleep EEG pattern was observed. Between these two levels of

transaction lies the reticular system in the core of the brain stem. This area receives collateral input from sensory afferents. When stimulated, persistent wake EEG was observed, and when lesioned, a persistent slow wave sleep EEG pattern resulted (Jones, 2010). This latter observation suggested that sensory input was not sufficient to produce wakefulness, as these lesions preserved sensory input to the central nervous system. More modern studies using lesions and stimulation of various forebrain, hypothalamic, and brain stem regions have revealed that sleep–wake regulation involves the coordinated activity of numerous brain regions, and sleep does not simply result from decreased sensory input due to fatigue.

Homeostatic

The timing of sleep has been proposed to be regulated by two factors: a homeostatic factor and a circadian factor (Borbely, 1982). In this model, a sleep-promoting factor (Process S) accumulates during waking. When Process S accumulates to a specific threshold, a sleep drive is initiated. The levels of Process S then degrade during sleep. Upon falling to another threshold, wake is triggered. These thresholds are regulated by the circadian component, such that the levels at which Process S promotes waking and sleep are time-of-day dependent. While the circadian component has been well established (see below), the search for a homeostatic factor has been more problematic. It had been hypothesized that there may be a humoral factor that promotes sleep in a fashion consistent with Process S. Cytokines, such as interleukin-1, do promote sleep, but this likely has to do with promoting sleep to fight off an infection, and they are not likely involved in sleep regulation. A number of observations rule out a humoral factor that accumulates during wake to induce sleep. First, conjoined twins sharing a blood supply do not show synchronized sleep or sleep states. Furthermore, cetaceans, such as the dolphin, and waterfowl are able to exhibit different EEG states in each cerebral hemisphere, with one hemisphere awake while the other exhibits a slow wave sleep EEG pattern.

The nucleotide adenosine does fit the characteristics of a sleep-promoting homeostatic drive (Basheer et al., 2004; Porkka-Heiskanen et al., 1997). Adenosine can be released as a neurotransmitter, but is also found extracellularly as the byproduct of metabolism of adenosine triphosphate. Adenosine generally inhibits the activity of neurons by activating the A1 receptor. Adenosine levels accumulate during wake, particularly in the basal forebrain regions that contain wake-promoting neurons. Caffeine and other xanthines that promote wakefulness do so by acting as adenosine receptor antagonists.

Neural Mechanisms

Three main regions of the central nervous system play roles in regulating sleep and wake: the brain stem, the basal forebrain, and the hypothalamus. Historical investigations highlighted the importance of brain stem structures for the regulation of wake and REM sleep. The basal forebrain regulates wake. Finally, the hypothalamus plays a role in regulating both sleep and wake, as well as the circadian timing of behavioral state and transitions between sleep and wake. A variety of regions and neurotransmitter play a critical role, summarized below and in Table 1.

Wake

Brain Stem

As highlighted by the early transection studies, the brain stem contains a critical wake-promoting area. A number of brain stem regions combine to give rise to the ascending reticular activating system (ARAS). The first structure to be described was the reticular formation at the core of the brain stem extending from the caudal midbrain to the medulla. This structure is critical for producing wake-like states. Stimulation of the reticular formation in the midbrain elicits a wake-like EEG pattern (Moruzzi and Magoun, 1949). Cells in this region discharge when cortical EEG exhibits a desynchronized low-amplitude pattern during either wake or REM sleep (McCarley and Hobson, 1971). Cholinergic neurons in the pedunculopontine and laterodorsal tegmental nuclei (PPT/LDT) that lie just dorsal to the reticular formation in the pons are also critical for wakefulness. Both of these populations appear to play a role in both REM and wake regulation.

Serotonin

Also in the brain stem are monoaminergic populations that participate in regulation of waking as part of the ARAS. Serotonergic cells in the dorsal and median raphe nuclei are most active during wake, less active during NREM sleep, and are silent during REM sleep. While serotonin neurons are active during wake, they are most active during wake states with lower arousal levels such as during repetitive motor behaviors like chewing or grooming. Serotonin agonists inhibit both NREM and REM sleep. However, serotonin also appears to inhibit cholinergic neurons, and thus it is mildly sedating. Serotonergic cells project to many other regions involved in sleep and wake regulation, including ventrolateral preoptic area (VLPO, see below), histaminergic cells in the tuberomammillary

Table 1 Various brain regions involved in sleep–wake regulation, their corresponding neurotransmitters and relative activity during the various behavioral states

		Activity during stage		
Brain region	*Neurotransmitter*	*Wake*	*NREM sleep*	*REM sleep*
Reticular formation	ACh	↑↑	–	↑↑
Raphe	Serotonin	↑↑	↑	–
Locus coeruleus	Norepinephrine	↑↑	↑	–
Tuberomammillary nucleus	Histamine	↑↑	↑	–
LDT/PPT	ACh	↑	–	↑↑
VLPO	GABA	–	↑↑	↑
eVLPO	GABA	–	↑	↑↑
Basal forebrain	ACh	↑↑	–	↑↑
Lateral hypothalamus	Hypocretin	↑↑	–	–
Suprachiasmatic nucleus	GABA, neuropeptides	High activity during the day Low activity during the night		

nucleus, wake active cells in the basal forebrain, and the circadian clock in the suprachiasmatic nucleus (SCN).

Norepinephrine

Another region of brain stem cells involved in wake regulation is the locus coeruleus, which contains noradrenergic neurons. Like serotonin cells, these neurons are most active during wake, less active during NREM sleep, and silent during REM sleep. They project widely to other areas involved in sleep and wake regulation. They activate cells in the ARAS, serotonergic cells in the raphe, and cholinergic cells in the basal forebrain. Noradrenergic cells also inhibit sleep active neurons in the VLPO and REM active cholinergic neurons in the brain stem.

Basal Forebrain

While brain stem lesions of the ARAS regions initially produce a profound coma-like slow wave sleep state, animals eventually recover an alternating pattern of wake and sleep EEG. This suggests that while the ARAS plays an important role in production of wake, other areas participate as well. The ARAS gives rise to two ascending pathways. The dorsal path projects to the thalamus to regulate cortical arousal, while the ventral path projects to the basal forebrain and hypothalamus where other wake-promoting regions are found.

The basal forebrain contains cholinergic cells that project to the cortex. In particular, the cells in the caudal basal forebrain in regions such as the substantia innominata, nucleus basalis, and preoptic area exhibit higher firing during periods of cortical desynchrony (wake and REM). Activation of cholinergic cells in the basal forebrain increases cortical desynchrony, while inhibition of cholinergic cells here promotes cortical synchrony (Lee and Dan, 2012).

Hypothalamus – Histamine

Histaminergic cells in the tuberomammillary nucleus of the posterior hypothalamus also promote wakefulness. Antihistamine medication used to inhibit allergic reactions or nausea produce drowsiness through blocking H1 receptors. These cells are most active during wake, mildly active during NREM, and silent during REM sleep. These neurons activate cells in the ARAS and inhibit sleep active neurons found in the VLPO. They are activated by input from hypocetin cells in the hypothalamus (see below) and serotonin cells from the raphe nuclei.

Hypothalamus – Hypocretin (Orexin)

The lateral and posterior hypothalamus contains a small population of cells that contain the neuropeptide hypocretin (also known as orexin). These hypocretin cells are critical for normal regulation of wake and sleep, as is evidenced by the sleep disorder narcolepsy. Narcoleptics suffer from sleep attacks and cataplexy (the muscle atonia associated with REM sleep). Most people with narcolepsy have lost the majority of their hypocretin cells, although some heritable forms derive from a mutation in a hypocretin receptor.

Hypocretin cells receive noradrenergic input from the locus coeruleus, as well as cholinergic and serotonergic input. Cholinergic input activates hypocretin cells, whereas serotonin input is generally inhibitory. Noradrenergic input can be either excitatory or inhibitory. Hypocretin cells are active during wake, show diminished activity during NREM sleep, and are silent during REM sleep (Jones, 2005). When active, they excite neurons in a variety of wake active areas, including the tuberomammillary nucleus, cholinergic basal forebrain, locus coeruleus, and reticular formation (Jones, 2010; Sutcliffe and de Lecea, 2002). When activated optogenetically, sleeping animals rapidly wake up (Adamantidis et al., 2007). Optogenetically silencing hypocretin neurons induces slow wave sleep (Tsunematsu et al., 2011). Hypocretin agonists enhance arousal, and even mitigate the negative effects of sleep deprivation on cognitive performance (Deadwyler et al., 2007), while antagonists may be useful to promote sleep in people experiencing insomnia (Neubauer, 2010).

Sleep

Ventrolateral Preoptic Area

The early transections studies highlighted that the forebrain was able to produce slow wave NREM sleep on its own. The location of the major sleep-active neurons is the VLPO. Cell-specific lesions in this region lead to large reduction in sleep, particularly in NREM sleep (Lu et al., 2000). These cells contain the inhibitory neurotransmitters gamma aminobutyric acid (GABA) and galanin. These cells are active during sleep, as assessed by multiunit activity and immediate early gene expression (Brown et al., 2012). Their firing rate is even higher during NREM sleep following sleep deprivation (Szymusiak et al., 1998). Adenosine is able to indirectly activate VLPO neurons by turning off inhibitory input, thus linking the activity of these neurons to a putative homeostatic sleep drive. The VLPO makes reciprocal connections with many components of the ARAS, most notably the histaminergic tuberomammillary nucleus, the noradrenergic locus coeruleus, hypocretinergic cells in the lateral hypothalamus, and the serotonergic raphe nuclei. Acetylcholine and norepinephrine inhibit the activity of VLPO neurons, and serotonin inhibits most VLPO neurons. Histamine and hypocretin appear to indirectly inhibit VLPO neurons by activating local inhibitory interneurons. Conversely, VLPO neurons are able to inhibit most of these areas. It is these mutually inhibitory reciprocal connections that are thought to regulate maintenance of a particular state (e.g., sleep or wake) and transition between states (see flip-flop model below). It is the inhibition of the wake active areas by GABA that is thought to underlie the mechanism of action of benzodiazepine sleeping pills (Jones, 2005).

Median Preoptic Area

The median preoptic area (MnPO) also contains inhibitory GABAergic neurons. While VLPO neurons are most active during sleep, cells in the MnPO appear to be most active in relation to sleep drive. Their firing precedes sleep and is highest during periods of sleep deprivation. These cells also project to many wake active areas, inhibiting their activity. Electrical or chemical activation of the MnPO enhances sleep, whereas silencing of the MnPO enhances wake (Brown et al., 2012).

REM Sleep

REM sleep is a complex state with many unique features. First and foremost, it is a sleep state characterized by unconsciousness

and unresponsiveness. However, cortical activity is wake-like, exhibiting high-frequency, low-power EEG patterns. Some unique electophysiological events occur during REM, such as PGO waves and hippocampal theta activity (4–7 Hz oscillation, Robinson et al., 1977). Additionally, postural muscles are paralyzed during REM sleep. Finally, REM sleep derives its name from unique REMs.

Early transection studies (reviewed in Brown et al., 2012) revealed that cuts between the brain stem and the midbrain eliminated REM sleep electrophysiological patterns in the forebrain (PGO waves, hippocampal theta, low-voltage high-frequency EEG). However, REMs and muscle atonia were preserved, thus highlighting the brain stem as critical for REM sleep regulation.

Cholinergic regions in the brain stem participate in REM sleep. Cells in the laterodorsal tegmental nucleus and pedunculopontine tegmental nucleus are active during both wake and REM sleep, and their firing is correlated with cortical desynchrony. Cells in this region also innervate the thalamus, particularly the lateral geniculate. When these cholinergic brain stem areas are stimulated, PGO wave-like activity can be recorded in the LGN. When these cholinergic brain stem areas are lesioned, PGO waves are lost. Simultaneous recordings in the cholinergic brain stem region and the thalamus during REM sleep reveal that bursting in the LDT/PPT correlates with occurrences of PGO waves.

In rats, glutamatergic neurons just ventral to the LDT/PPT in the sublaterodorsal nucleus (SLD) also appear to be critical for REM regulation. In cats, the homologous cell group is located more caudally, just ventral to the locus coeruleus in the subcoerulear nucleus (SubC). These regions are necessary for muscle atonia, as REM atonia is lost with small discrete lesions of the SLD/SubC. When animals receive larger lesions in these regions, they appear to act out their dreams while in REM sleep. For instance, cats will engage in what looks like predation behavior (stalking and pouncing) while in REM sleep. Cells in the SubC increase their firing prior to atonia, and immediate early gene expression in the SLD is correlated with time spent in REM sleep (Brown et al., 2012; Saper et al., 2010). Electrical stimulation of the SubC can elicit muscle atonia in some cases (Brown et al., 2012). The SLD region also receives input from the VLPO and hypocretin regions of the hypothalamus. Other glutamatergic brain stem regions implicated in REM control are the parabrachial nucleus and precoeruleus regions. These regions project to the forebrain and express immediate early genes during periods of REM sleep (Lu et al., 2006).

Neurons in the reticular formation, particularly in the nucleus pontis oralis (PnO) and nucleus pontis caudalis (PnC), participate in REM regulation. Many cells in these regions are GABAergic and receive cholinergic input from the LDT/PPT. Infusions of cholinergic agonists excite a majority of PnO and PnC neurons, and can induce long periods of REM sleep (Bourgin et al., 1995).

Both serotonergic cells from the raphe and noradrenergic cells from the locus coeruleus appear to inhibit REM sleep. These regions are silent during REM sleep. LDT neurons that are active during REM are silenced when serotonin agonists are applied.

While the circuitry of the brain stem explains much of the neural regulation of REM sleep, areas in the forebrain also appear to participate. Cholinergic neurons in the basal forebrain burst during both wake and REM sleep (Lee et al., 2005), likely contributing to cortical arousal during both of these states. Cells in the extended VLPO region are also involved in REM sleep regulation. While lesions of the central cluster of VLPO cells disrupt primarily NREM sleep, lesions of the medial and dorsal VLPO regions that have more scattered neurons lead to a decrease in REM sleep (Lu et al., 2000). These cells in the extended VLPO show immediate early gene expression that is more highly correlated with REM sleep. These sleep active extended VLPO neurons project to both the locus coeruleus and dorsal raphe, and appear to use the inhibitory neurotransmitter galanin (Lu et al., 2002). These data suggest that the extended VLPO may elicit REM sleep by inhibiting the locus coeruleus and raphe nuclei.

State Regulation

The architecture of sleep and wake is carefully regulated. Species typically exhibit large consolidated bouts of both wake and sleep, and within sleep, animals alternate between consolidated bouts of REM and NREM sleep. As mentioned above, these patterns are modeled in part by the two-process sleep model that contains both a circadian and a homeostatic drive. Actual transitions between wake and sleep, and REM and NREM involve the coordinated change in neural activity of sleep and wake brain regions reviewed thus far.

Circadian

Sleep and wake occur at species-typical times of day. Animals awake at night, such as mice, are nocturnal while those awake during the day, such as humans, are diurnal. Other species exhibit a bimodal activity pattern that tracks dawn and dusk, and are known as crepuscular. The patterning of sleep and wake is controlled by a circadian clock located in the SCN of the hypothalamus (Antle and Silver, 2005). Cells in the SCN are intrinsically rhythmic, due to a molecular clock deriving from autoregulatory transcription–translation feedback loops. The oscillations of these individual clock cells are coordinated, and as a tissue, the SCN provides a temporal signal to the rest of the brain and body to regulate daily oscillations in behavior and physiology. While these circadian rhythms can persist in the absence of time cues, the SCN receives input from the retina so that its oscillations are entrained to the day–night cycle. The SCN outputs to a variety of brain regions involved in sleep and wake regulation, including the VLPO, MnPO, and hypocretin cells. Furthermore, the SCN innervates the subparaventricular zone and dorsomedial hypothalamus, both of which innervate many of the same sleep–wake areas, allowing for modulation of circadian signals (Kriegsfeld et al., 2004; Mistlberger, 2005). The SCN is also able to regulate other areas involved in sleep and wake regulation, such as the histaminergic cells in the tuberomammillary nucleus, through multisynaptic pathways. The SCN also projects to the lateral septum, which regulates the medial septum, which in turn regulates hippocampal theta activity (Bland, 1986).

The SCN is metabolically and electrically most active during the day, regardless of diurnality or nocturnality. The electrical

activity of the circadian clock in rats is also related to sleep state, with elevated activity during REM sleep and lowered activity during NREM sleep (Deboer et al., 2003). Most SCN cells release GABA as their main small molecule neurotransmitter. Most SCN cells also employ one or more peptidergic neurotransmitter as well. These include vasoactive intestinal polypeptide, gastrin-releasing peptide, and arginine vasopressin. The SCN is also able to regulate rhythmicity through a paracrine mechanism, as SCN-lesioned animals regain rhythmicity when they receive fetal SCN grafts that are encapsulated in a material that prevents neuronal reconnection, but allows diffusion of chemical messengers (Silver et al., 1996). Two chemicals have been suggested to participate in this paracrine transmission: transforming growth factor alpha and prokineticine-2 (Mistlberger, 2005).

The circadian clock provides temporal structure to sleep–wake cycles. When the SCN is lesioned, animals still exhibit bouts of sleep and wake, but these bouts are very short and are evenly distributed across the day and night. The circadian clock was originally suggested to promote wake, as it had been observed that arrhythmic animals with SCN lesions exhibited more sleep. One possibility is that more sleep is required once sleep becomes fragmented due to lower sleep efficiency in SCN-lesioned animals. However, the observation of increased sleep in SCN-lesioned animals is not universal, and further studies have since suggested that the circadian clock actively promotes both sleep and wake (Mistlberger, 2005).

On top of regulating behavioral state, the circadian clock can also be influenced by changes in such states. The SCN receives heavy serotonergic innervation from the median and dorsal raphe (Yamakawa and Antle, 2010). Sleep deprivation enhances serotonin output at the SCN (Grossman et al., 2000) and can reset the phase of the circadian clock (Antle and Mistlberger, 2000).

Transitions between States

Wake, REM sleep, and NREM sleep are each associated with activity in specific brain areas, but a mechanism to switch between states is also required. It is the reciprocal innervation patterns between many of these areas that give rise to the brain network that both maintains a particular sleep–wake state and also permits rapid switching between states.

In people, sleep typically lasts for about 8 h. The two-process model explains quite well the timing of sleep onset and offset according to both circadian and homeostatic drives. However, the switching of activity between the various wake and sleep areas needs to be carefully coordinated. A flip-flop model has been proposed that explains how activity of these various areas is coordinated (Saper et al., 2001). The flip-flop switch derives from electrical engineering circuits, and represents a bistable two-path switch. Each state reinforces itself and inhibits the alternate state. In the case of sleep–wake regulation, wake areas are self-reinforcing and inhibit activity in sleep centers, while sleep centers are also self-reinforcing and inhibit activity in wake centers. Circadian and homeostatic factors could influence the activity of wake and/or sleep centers, and would contribute to the trigger to initiate a state transition.

The specific features of the sleep–wake flip-flop switch (**Figure 1**) have noradrenergic signals from the locus coeruleus, serotonergic signals from the raphe and histaminergic signals from the tuberomammillary nucleus on the wake side of the circuit, and GABAergic signals from the VLPO on the sleep side of the circuit. Signals from the hypocretin cells in the lateral

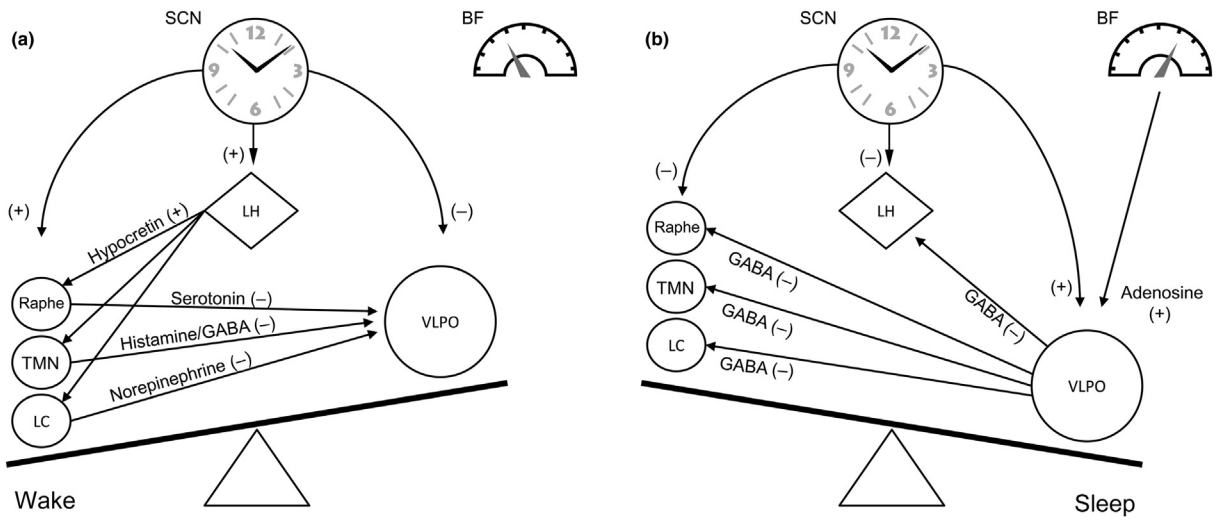

Figure 1 Flip-flop model of regulation of sleep–wake. (a) During wake, the wake centers (raphe nuclei, locus coeruleus (LC), and tuberomammillary nucleus (TMN)) inhibit the activity of the sleep center in the ventrolateral preoptic area (VLPO). The hypocretin neurons in the lateral hypothalamus (LH) reinforce activity in the wake areas. The circadian clock in the suprachiasmatic nucleus may promote wake and inhibit sleep. The homeostatic drive (thermostat, representative of adenosine levels in the basal forebrain) is low, although accumulates during wake. (b) During sleep, the sleep center in the VLPO area inhibits wake areas including the raphe nuclei, TMN, LC, and LH. The circadian clock may inhibit wake and promote sleep. The homeostatic drive is high due to high levels of adenosine in the basal forebrain. Modified from Brown, R.E., Basheer, R., McKenna, J.T., Strecker, R.E., McCarley, R.W., 2012. Control of sleep and wakefulness. Physiological Reviews 92, 1087–1187.

hypothalamus maintain activity in the wake centers, while adenosine serves as a homeostatic signal that influences the activity of the VLPO to promote sleep. The circadian clock can actively promote the state of the flip-flop circuit by directly or indirectly regulating the activity of the hypocretin cells and the VLPO. The key feature of such a flip-flop switch is that animals in a wake state will tend to stay awake, and animals in a sleep state will tend to stay asleep. It is only with significant input to the other side of the flip-flop circuit that the state can be switched, and when such switches occur, they occur rapidly.

Within sleep there are also neural networks that regulate switching between REM and NREM. This switching occurs with an ultradian REM/NREM cycle of about 90–100 min. Typically sleep is initiated through NREM, and bouts of NREM are longer and the cortical EEG shows larger amplitudes and lower frequencies in the first half of the sleep phase. REM bouts get longer as the sleep phase progresses. A number of models have been proposed to account for this switching. Similar to the sleep–wake control, a flip-flop model has also been proposed, where areas active during REM (cholinergic cells) and areas silent during REM (serotonergic and noradrenergic cells) are mutually inhibitory (Lu et al., 2006). Others have suggested the use of a Lotka–Volterra model (McCarley and Hobson, 1975). In classic Lotka–Volterra systems, oscillations in predator–prey population sizes are modeled with oscillations in predator population size lagging prey population sizes. The prey numbers are self-reinforcing, and also promote increases in predator populations. Predator population, however, are inversely related to prey populations (i.e., high predator numbers leads to low prey, and low predator number leads to high prey numbers). The result is a stable oscillation in both predator and prey population sizes. When this model is applied to regulation of REM and NREM, realistic alternations between these states emerges. Specifically, in the case of sleep regulation, the Lotka–Volterra model uses activity of REM-on areas as the prey component, and the activity of REM-off areas as the predator. In this way, areas active during REM slowly activate the REM-off areas, such that their activity eventually reaches a sufficient level to inhibit the REM-on areas. Specifically, serotonergic and noradrenergic cells are active during wake, but slowly reduce their firing during NREM sleep. This gradually disinhibits the cholinergic REM-on cells. When the REM-on cholinergic cells become active, they slowly activate the serotonergic and noradrenergic cells. The REM-on neurons also positively reinforce their own activity to help maintain the REM state until activity in the serotonergic and noradrenergic cells is sufficient to turn off the cholinergic REM-on cells (Brown et al., 2012; McCarley and Hobson, 1975).

Summary

The brain has a large network of areas that orchestrate behavioral state, using a wide variety of neurotransmitters. Brain stem, hypothalamic, and basal forebrain regions are critical for producing wake, sleep, and REM sleep. While early work highlighted the role of brain stem regions for regulating sleep and wake, the hypothalamus and basal forebrain are now recognized as being particularly critical for regulating sleep and wake.

See also: Sleep Disorders across the Life Span; Sleep Disorders: Psychiatric Aspects; Sleep and Health; Sleep and Memory.

Bibliography

Adamantidis, A.R., Zhang, F., Aravanis, A.M., Deisseroth, K., de Lecea, L., 2007. Neural substrates of awakening probed with optogenetic control of hypocretin neurons. Nature 450, 420–424.

Antle, M.C., Mistlberger, R.E., 2000. Circadian clock resetting by sleep deprivation without exercise in the Syrian hamster. Journal of Neurosciences 20, 9326–9332.

Antle, M.C., Silver, R., 2005. Orchestrating time: arrangements of the brain circadian clock. Trends in Neurosciences 28, 145–151.

Basheer, R., Strecker, R.E., Thakkar, M.M., McCarley, R.W., 2004. Adenosine and sleep-wake regulation. Progress in Neurobiology 73, 379–396.

Bland, B.H., 1986. The physiology and pharmacology of hippocampal formation theta rhythms. Progress in Neurobiology 26, 1–54.

Borbely, A.A., 1982. A two process model of sleep regulation. Human Neurobiology 1, 195–204.

Bourgin, P., Escourrou, P., Gaultier, C., Adrien, J., 1995. Induction of rapid eye movement sleep by carbachol infusion into the pontine reticular formation in the rat. Neuroreport 6, 532–536.

Bremer, F., 1935. Cerveau "isolé" et physiologie du sommeil. Comptes Rendus des Séances de la Société de Biologie et des ses Filiales (Paris) 118, 1235–1241.

Brown, R.E., Basheer, R., McKenna, J.T., Strecker, R.E., McCarley, R.W., 2012. Control of sleep and wakefulness. Physiological Reviews 92, 1087–1187.

Deadwyler, S.A., Porrino, L., Siegel, J.M., Hampson, R.E., 2007. Systemic and nasal delivery of orexin-A (Hypocretin-1) reduces the effects of sleep deprivation on cognitive performance in nonhuman primates. Journal of Neuroscience 27, 14239–14247.

Deboer, T., Vansteensel, M.J., Detari, L., Meijer, J.H., 2003. Sleep states alter activity of suprachiasmatic nucleus neurons. Nature Neuroscience 6, 1086–1090.

Grossman, G.H., Mistlberger, R.E., Antle, M.C., Ehlen, J.C., Glass, J.D., 2000. Sleep deprivation stimulates serotonin release in the suprachiasmatic nucleus. Neuroreport 11, 1929–1932.

Jones, B.E., 2005. From waking to sleeping: neuronal and chemical substrates. Trends in Pharmacological Sciences 26, 578–586.

Jones, B.E., 2010. Basic mechanisms of sleep-wake states. In: M.H., K., Roth, T., Dement, W.C. (Eds.), Principles and Practice of Sleep Medicine, fifth ed. Saunders, Philadelphia, PA.

Kriegsfeld, L.J., Leak, R.K., Yackulic, C.B., LeSauter, J., Silver, R., 2004. Organization of suprachiasmatic nucleus projections in Syrian hamsters (*Mesocricetus auratus*): an anterograde and retrograde analysis. Journal of Comparative Neurology 468, 361–379.

Lee, M.G., Hassani, O.K., Alonso, A., Jones, B.E., 2005. Cholinergic basal forebrain neurons burst with theta during waking and paradoxical sleep. Journal of Neuroscience 25, 4365–4369.

Lee, S.H., Dan, Y., 2012. Neuromodulation of brain states. Neuron 76, 209–222.

Lu, J., Bjorkum, A.A., Xu, M., Gaus, S.E., Shiromani, P.J., Saper, C.B., 2002. Selective activation of the extended ventrolateral preoptic nucleus during rapid eye movement sleep. Journal of Neuroscience 22, 4568–4576.

Lu, J., Greco, M.A., Shiromani, P., Saper, C.B., 2000. Effect of lesions of the ventrolateral preoptic nucleus on NREM and REM sleep. Journal of Neuroscience 20, 3830–3842.

Lu, J., Sherman, D., Devor, M., Saper, C.B., 2006. A putative flip-flop switch for control of REM sleep. Nature 441, 589–594.

McCarley, R.W., Hobson, J.A., 1971. Single neuron activity in cat gigantocellular tegmental field: selectivity of discharge in desynchronized sleep. Science 174, 1250–1252.

McCarley, R.W., Hobson, J.A., 1975. Neuronal excitability modulation over the sleep cycle: a structural and mathematical model. Science 189, 58–60.

Mistlberger, R.E., 2005. Circadian regulation of sleep in mammals: role of the suprachiasmatic nucleus. Brain Research Reviews 49, 429–454.

Moruzzi, G., Magoun, H.W., 1949. Brain stem reticular formation and activation of the EEG. Electroencephalography and Clinical Neurophysiology 1, 455–473.

Neubauer, D.N., 2010. Almorexant, a dual orexin receptor antagonist for the treatment of insomnia. Current Opinion in Investigational Drugs 11, 101–110.

Porkka-Heiskanen, T., Strecker, R.E., Thakkar, M., Bjorkum, A.A., Greene, R.W., McCarley, R.W., 1997. Adenosine: a mediator of the sleep-inducing effects of prolonged wakefulness. Science 276, 1265–1268.

Robinson, T.E., Kramis, R.C., Vanderwolf, C.H., 1977. Two types of cerebral activation during active sleep: relations to behavior. Brain Research 124, 544–549.

Saper, C.B., Chou, T.C., Scammell, T.E., 2001. The sleep switch: hypothalamic control of sleep and wakefulness. Trends in Neuroscience 24, 726–731.

Saper, C.B., Fuller, P.M., Pedersen, N.P., Lu, J., Scammell, T.E., 2010. Sleep state switching. Neuron 68, 1023–1042.

Silver, R., LeSauter, J., Tresco, P.A., Lehman, M.N., 1996. A diffusible coupling signal from the transplanted suprachiasmatic nucleus controlling circadian locomotor rhythms. Nature 382, 810–813.

Sutcliffe, J.G., de Lecea, L., 2002. The hypocretins: setting the arousal threshold. Nature Reviews Neuroscience 3, 339–349.

Szymusiak, R., Alam, N., Steininger, T.L., McGinty, D., 1998. Sleep-waking discharge patterns of ventrolateral preoptic/anterior hypothalamic neurons in rats. Brain Research 803, 178–188.

Tsunematsu, T., Kilduff, T.S., Boyden, E.S., Takahashi, S., Tominaga, M., Yamanaka, A., 2011. Acute optogenetic silencing of orexin/hypocretin neurons induces slow-wave sleep in mice. Journal of Neuroscience 31, 10529–10539.

Yamakawa, G.R., Antle, M.C., 2010. Phenotype and function of raphe projections to the suprachiasmatic nucleus. European Journal of Neuroscience 31, 1974–1983.

Slow Food as a Social Movement

Koen van Bommel, VU University, Amsterdam, The Netherlands
André Spicer, Cass Business School, City University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the development of the Slow Food movement. The first section examines the history of the movement and the main changes since its start in 1989. The second focuses on more general themes that account for Slow Food's development. We first examine Slow Food's transformation from a gastronomic to an eco-gastronomic movement and subsequently discuss the movement's potential to create markets around authenticity. Finally, we examine the implications of the increasing globalization of the movement as it broadened its initial Italian focus toward other developed as well as developing countries.

Preoccupation with food, its origins and distribution, has become an emblematic feature of current times among consumers, producers, civil society, and policy makers. Whereas famine and malnutrition remain problems of primary concern in some areas, overproduction and waste are paramount in others. More generally, the dominance of monocultures, chemicals, and large-scale global production of standardized food has been put into question as it potentially ruins taste, health, our environment, the livelihood of farming communities, and overall creates disenchantment with the experience of cooking and eating (Bové and Dufour, 2002; Honore, 2004; Ritzer, 2011; Schlosser, 2002). In response, movements advocating organic food (Skinner, 2007) and fair trade food (Reinecke, 2010) have proliferated. Arguably, one of the most prolific examples of resistance against hegemonic food regimes has been the Slow Food movement and its quest for 'good, clean, and fair' food. What started out as mainly an Italian gastronomic movement attracting foodies has become a global movement that appeals to a broader constituency of environmentalists and people concerned with fair trade and social justice.

As of 2013, Slow Food has around 1500 local chapters (called convivia) and over 100 000 members across more than 160 countries. Founded in 1989, the Slow Food movement has been strongly embedded in its Italian roots. National associations soon emerged in Germany (1992) and Switzerland (1993) to coordinate the activities and convivia in these countries. After the turn of the century, further national associations were set up in the United States (2000), Japan (2004), the United Kingdom (2005), and the Netherlands (2008). Slow Food has its headquarters still in the small town of Bra in the Italian Piedmont region from where global initiatives and projects are coordinated. The association is essentially governed by its International Council and Executive Committee under the continuing presidency of founder Carlo Petrini. An official member-based association in its own right, it can also be seen as a social movement (see e.g., Andrews, 2008; Van Bommel and Spicer, 2011), as it comprises "collective challenges by people with common purposes and solidarity in sustained interaction with elites, opponents and authorities" (Tarrow, 1994: pp. 3–4) and taps into "… a much broader backlash against the high-speed, high-turnover culture of the global food industry" (Honore, 2004: p. 57). We examine the history of Slow Food and focus on three themes that help to explain the development of the Slow Food movement: (1) its broad shift from a gastronomic to an eco-gastronomic movement; (2) the importance of an identity of authenticity in relation to the emergence of a new market for Slow Food; and (3) the implications of the increasing globalization of the movement.

A History of the Slow Food Movement

Early Years

The Slow Food movement's official narrative has it that it was upon realizing that the site of the famous Spanish Steps in Rome would be desecrated by the opening of a McDonalds in 1986, Italian Carlo Petrini instigated a protest. Together with some of his friends, most of them deeply rooted in Italy's Left, he organized a dramatic sit-in that involved eating traditional Italian pasta. This group subsequently decided to form the action group Arcigola (affiliated with the Italian communist party) to protect local Italian cuisine and Slow Food emerged from this action group 3 years later. In fact, Slow Food's history is somewhat more complex, of course, with its roots already in the counterculture movements of the 1960s and 1970s and the growing global appreciation of, and concern for, Italian culinary culture and heritage (for an overview, see e.g., Andrews, 2008; Chrzan, 2004). Still, the new McDonalds was important in instigating the eventual birth of the Slow Food movement, made official with the signing of the Slow Food Manifesto in Paris in 1989. Petrini became president of this new member-based association and remains so until this day.

From the outset Slow Food has set itself firmly against the threat of the erosion of Italian culinary tradition through the globalization and McDonaldization of society (Ritzer, 2011). That is, from its start in 1989 and under the continuing leadership of Petrini the movement has fought against agricultural industrialization and the resulting rise of homogenous and standardized food which subsequently comes at the expense of not only the taste buds but also of local artisan producers and communities, and thus ultimately reduces the pleasure of food.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64060-6

The initial aim of the Slow Food movement was primarily gastronomic, which is stipulated by its founding manifesto:

> A firm defense of quiet material pleasure is the only way to oppose the universal folly of Fast Life … May suitable doses of guaranteed sensual pleasure and slow, long-lasting enjoyment preserve us from the contagion of the multitude who mistake frenzy for efficiency … Let us rediscover the flavors and savors of regional cooking and banish the degrading effects of Fast Food … That is what real culture is all about: developing taste rather than demeaning it.
> (http://www.slowfood.com/about_us/eng/manifesto.lasso. accessed 10.12.13.)

In its early years, Slow Food focused on gastronomy by (discursively) emphasizing the importance of good taste, artisan production, local and traditional cuisine, and the general importance of pleasure and conviviality (Van Bommel and Spicer, 2011). The movement's tactics to achieve its goals were in these years predominantly local and infrapolitical in nature. That is, Slow Food's activities were largely decentralized around the level of the 'convivia' which formed the backbone of the movement. These regionally embedded groups tried to propagate Slow Food by organizing in civil society a variety of grassroots activities such as markets, discussions, school garden projects, taste sessions, and dinners. Engagement with political elites remained rather limited in the movement's formative years. Still, at a more central level, some early projects started to be developed. Most notably, the first edition of Slow Food's flagship biennial event Salone del Gusto (Hall of Taste) was organized in Turin in 1996. This food fair extravaganza brings together a wide and global variety of products and producers adhering to the Slow Food tradition and by 2012 attracted over 220 000 visitors. At the same time, the Ark of Taste was introduced, which is a catalog of gastronomic products "aimed at safeguarding and promoting small-scale fine food purveyors who are threatened by extinction," according to the Ark of Taste Manifesto Petrini 2001.

Growth and Change

Around the turn of the twenty-first century, a shift started to take place around Slow Food as it expanded its reach, focus, and activities. Perhaps most notably, the movement extended its initial focus on gastronomy to what has now become known as 'eco-gastronomy' and structured its message around the notion of 'good, clean, and fair' food. Essentially, the gastronomic focus broadened with additional concerns for the environmental and sustainability aspects of food as well as a commitment to social-economic justice, in particular also for developing countries and other vulnerable food producers. The idea behind it is described by Petrini as "simultaneously observing the principles of taste, respect for the environment, and biodiversity" (cited in Sassatelli and Davolio, 2010: p. 210). Whereas the term eco-gastronomy appears vague, it is possibly exactly this degree of ambiguity that provides it with meaning for a broad range of stakeholders such as gourmets, environmental activists, and advocates of social justice. This broadening of the movement's discourse not only expanded its appeal, it also added a political layer to food that went beyond merely enjoying it. That is, "a gastronome is someone who combines environmental concern with the pursuit of pleasure; an enlightened citizen who will have an expanding knowledge of food in its social and environmental context and an understanding and empathy for the predicaments of the small producer in the era of globalization" (Andrews, 2008: p. 72). In more plain terms, Slow Food's leader argues that "a gourmet who does not think of ecology is stupid, but an ecologist who is not a gourmet is boring" (*The Times*, 4 December 2006).

In terms of Slow Food's actual strategies and tactics, the movement started to combine its initial grassroots activities with more global, political, and centrally coordinated formal activities. For instance, The Ark of Taste was expanded with an operational arm in the form of the Presidia, which started in 1999. Through this initiative, spanning all continents, Slow Food experts search for products or production techniques at the risk of extinction. Relevant experts, producers, technicians, and institutions form a working group once a suitable end-product or processing method has been found. Whereas the exact strategies vary according to project and product, the Presidia revolve around the desire to promote small-scale artisan products and producers, stabilize traditional production techniques, establish stringent production standards, and guarantee a viable future for the food products as well as their regions and ecosystems, all within the confinements of the Slow Food philosophy by combining economic, ecological, social, and cultural considerations.

In 2004, the Terra Madre project was inaugurated. This landmark project aimed to develop an international network of food communities consisting of producers, farmers, consumers, academics, activists, chefs, and local community members and thus created opportunities to form a network of exchange around food that matched Slow Food's idea of eco-gastronomy and the 'good, clean, and fair' ideals. This integrated notion of food communities blurs the boundaries between producers and consumers, with the latter often being labeled co-producer, highlighting the important role of the consumer in the outlook of (global) food chains. Terra Madre is a good example of the politicization of Slow Food, with a considerable number of Terra Madre's projects involving developing countries. At the same time, it shows the movement's increasing globalization. In short, "Slow Food was now an aspiring global political movement which involved producers as well as consumers in its work, and was able to develop a distinctive position on globalization. Its concepts – Good, Clean and Fair, became the organizing principles for this new 'virtuous' globalization" (Andrews, 2008: p. 48). Highlighting its small-scale, sustainable, and local characteristics, Terra Madre was presented by the Slow Food movement as an alternative to more conventional foodways. In particular, Terra Madre resists the latter's prevailing production methods and consumption habits and the resulting degrading environmental and socio-economic effects on the planet and communities.

In addition to the influential Terra Madre project, the first years of the twenty-first century also witnessed the introduction of more central and formalized initiatives such as the opening of Slow Food's University for Gastronomic Science in 2004. Furthermore, the Foundation for Biodiversity started in 2003 and in 2008 instigated a formalized network of farmer's markets following the Slow Food philosophy under the banner

of the Earth Markets project. On a global scale, a range of Terra Madre events and regional Slow Food fairs start to appear from 2005 onward. Slow Food spin-offs dedicated to Cheese (2001) and Fish (2009) emerged under the banner of the Foundation for Biodiversity. A noteworthy example of the more global outreach of the movement, in particular toward the 'Global South' is the start of the '1000 Gardens in Africa' project in 2010. This practical initiative aims to extend Slow Food's long-lasting focus on school gardens by establishing school, community, or family gardens in Africa. The initiative focuses on creating new, or further developing existing, food gardens. The aim is not just to grow traditional and local produce in a sustainable matter, but also to strengthen community relations, agricultural knowledge, and farmer autonomy.

The movement also starts to engage more directly with political elites than before. For instance, its approach is marked by various political involvements with organizations such as the WTO, UN, World Social Forum, EU, and various national governments and NGOs. The internal structure of the movement has grown more complex while political campaigns and collaborations with elites were taking off. At the same time, however, activities at the local convivia level oftentimes remained less political and more embedded in the gastronomy discourse (Sassatelli and Davolio, 2010). Overall, Slow Food went through "…a rapid cultural change, which led to settlements of the organization's structure, new mass media strategies and, most of all, a growing range of activities and institutions to dialogue with. SF broadened its mission from the developed to the developing countries, from consumption to production, from dish to soil and environment" (Sassatelli and Davolio, 2010: p. 210).

In short, the focus, reach, and range of Slow Food's actions grew strongly in between its inception in 1989 and the present. We discuss in more detail the three aforementioned cross-cutting themes that mark the development of Slow Food.

From Gastronomy to Eco-Gastronomy

A first notable development that arises from the history of Slow Food is the transition from a dining club for gourmets primarily relying on local grassroots activities, to a broader movement initiating global political campaigns and attracting gastronomes, environmentalists, and social justice activists. The problem Slow Food faced was that although it attracted a respectable group of enthusiasts, the demand for Slow Food based primarily on gastronomic issues did not sufficiently resonate with a broader audience. This reliance on gastronomy and pleasure also gave Slow Food a stigma of an elitist club for middle-class gourmets rather than a broad movement trying to change the way people more generally feed themselves (see e.g., Andrews, 2008; Donati, 2005; Sassatelli and Davolio, 2010). Thus, it was of key importance to expand Slow Food's appeal while at the same time make sure not to alienate the loyal, and mainly local, early supporters. Many movements face difficulties in balancing these local grassroots level activities with more formal and globally oriented strategies for a broader appeal. In order to maintain coordination and control, it is common for growing organizations that start at the grassroots level to become more formalized, institutionalized, or hierarchical once they grow (Michels, 1958). This potential disconnect between an appeal based on local, informal grassroots activities and a broader more global and formal approach has been remarked upon in the case of Slow Food as well (Chrzan, 2004). Van Bommel and Spicer (2011) have labeled this balancing act of broadening the movement as Slow Food's attempt to balance a 'strategy of autonomy' with a 'strategy of engagement'.

In addition to the range of strategies and activities discussed, and in accordance with previous studies on social movements more generally (e.g., Laclau, 2005; Weber et al., 2008), Slow Food actively also sought to develop a common language with broad appeal. This became known as eco-gastronomy. After 2000, Slow Food tapped into broadly popular discourses situated around biodiversity, fair trade, sustainability, and social justice, while simultaneously preserving its initial focus on gastronomical aspects such as taste, pleasure, local, and artisan food. Interestingly enough, rather than the Slow Food message getting lost and confused in this potpourri of signifiers it arguably got stronger. Where other studies have warned for the threat of the erosion of the movement's identity (Melucci, 1996), it seems that Slow Food has managed to build its case around a central concept (eco-gastronomy) that is abstract and ambiguous enough to attract a broad range of enthusiasts while at the same time maintaining a coherent identity for the movement as a whole. Overall, it can be argued that Slow Food has been able to appeal "to different types of people in very different circumstances. …It seemed to have the ear of restaurateurs, farmers and policy makers, while retaining a radical and principled position sufficient to capture the imagination of anti-global activists and environmentalists" (Andrews, 2008: pp. vi–vii). Overall, Van Bommel and Spicer (2011) note that Slow Food thus managed to elevate itself by pursuing higher values rather than just enjoying food, and appeals to a broad and diverse mix of actors with various novel messages, strategies, and relationships. Slow Food's broad central concept of 'eco-gastronomy', linked with its continuing focus on 'good, clean, and fair' food enabled a scenario in which the movement offered something of value to a broad constituency.

Against Global Standardization: The Rise of Markets of Authenticity

Recent years have seen a proliferation of alternative markets for the likes of organic, local, sustainable, and fairly traded produce (Reinecke, 2010; Weber et al., 2008). On the one hand, this can be explained in the light of fears over genetically modified foods, widespread use of antibiotics, outbreaks of *Escherichia coli* and mad cow disease, the recent horse meat scandals, and the like. On the other hand, these newly emerging markets can also be seen as a countermovement or alternative to the hegemony of today's globalized food industry based primarily on mass produced and increasingly standardized food. Not without merits, this global system of standardized food has also caused an erosion of the connection between food, its locality, and food producers and consumers (Andrews, 2008; DeSoucey, 2010). As noted by Parkins and Craig (2006: p. 100), "such connectedness – of place, people,

history, culture – constitutes a dynamic set of social relations that is under threat by the chances represented by globalization and its capacity to erase cultural specificities."

In this respect, Slow Food's emphasis on 'terroir' is of particular interest. This concept refers to the importance of the cultural traditions and natural conditions of the local place in which the food is situated and forms a vital aspect of Slow Food's discourse. Interestingly, within today's hegemonic food system the consumer has been increasingly distanced from the importance of place for food, and the same applies to the producer in many cases. Slow Food envisions a future in which consumers and producers once again become closely connected with each other, and the locality of food and its production. This is exemplified by Slow Food's insistence on preferring the word 'co-producer' rather than 'consumer'. This term signifies the active role that consumers play in the cultivation, production, and distribution of food. The Slow Food movement can thus be seen as a form of resistance against forces toward the globalization, standardization, and homogenization of food where it "…has attempted to articulate a politics rooted in local identity, traditions and cultures. Such a position, which suggests a distinctive 'politics of place', questions recent accounts of globalization which sees 'the local' as mainly an appendage to 'the global', driven and shaped by global economies, cultures and trajectories" (Andrews, 2008: p. 129).

With this resistance against the global standardization of food, alternative markets have emerged that tap into a desire for authenticity. For instance, through connecting the production and consumption of food with the significance of culture, tradition, and place, the symbolic function of food is a "marker of identity and community" (DeSoucey, 2010: p. 434). This connection between food and nationally or locally based authentic identities has been labeled by DeSoucey as 'gastronationalism' (2010: pp. 433–434). This concept "signals the use of food production, distribution, and consumption to demarcate and sustain the emotive power of national attachment, as well as the use of nationalist sentiments to produce and market food." Whereas DeSoucey's focus was primarily on national identities, more generally her study hints at a turn toward authenticity, a discourse Slow Food alludes to as well in its many initiatives and communications. That is, rather than problematizing food from a point of reference of politico-ethical (fairness) or politico-economic (safety, prices, transparency), Slow Food's appeal builds on a politico-aesthetic dimension (Sassatelli and Davolio, 2010: pp. 226–227). Thus, the identification with fairness, quality, sustainability, and the like are inextricably linked, if not subordinated, with a right and responsibility toward pleasure, enjoyment, and aesthetic experiences around food. Slow Food typically focuses on rediscovering 'real' and authentic experiences, and on bringing back a human face to the production and consumption of food that provides an antidote against the soul destroying standardization of fast food. Slow Food can thus be seen as part of a broader tendency toward valuing local tradition, 'real' experiences, and authenticity to resist global homogenizing forces (see e.g., Boyle, 2003; Lewis and Bridger, 2001).

Slow Food's authenticity appeals to a sense of identification and attachment, further enhanced by its various processes of mobilization through the multiplication of initiatives, strategies, and activities to broaden the movement (Van Bommel and Spicer, 2011). Prior studies on social movements have shown that a consequence of movements' tendency to "cohere around collective identities and build mobilizing structures is that they often spawn new organizational forms and economic practices that sometimes lead to the creation of new markets or industries" (de Bakker et al., 2013: p. 579). In the case of Slow Food, actual markets are created for endangered products that correspond with the Slow Food philosophy of high quality, local, artisan, sustainable, and fairly traded products. For instance, Chrzan (2004) discusses markets for traditional turkey breeds and raw cheeses in the US that are supported by Slow Food initiatives. In addition, Slow Food Presidia projects have generally created, supported, and improved markets for its products and production techniques, such as Robiola di Roccaverano, a raw milk cheese from Piedmont, Huehuetenango Highland Coffee from Guatemala, and Swedish Jämtland Cellar Matured Goat Cheese (see e.g., Fonte, 2006; Peano and Sottile, 2012).

In addition to these micromarkets for particular products, Slow Food contributes to a reappraisal of the hegemonic food system dominated by corporate capitalism. Slow Food's position of linking the importance of pleasure, place, and conviviality with an acknowledgment of the importance of the environment and social justice gives it ample opportunity to alter the discourse around food. Markets often rely on embedded codes and cultural understandings (Rao et al., 2003; Weber et al., 2008) and movements such as Slow Food "often provide the cultural and material resources needed to create new markets for corporate actors" (de Bakker et al., 2013: p. 581). For instance, continuous taste education around its principles aims for consumers to "use their power of choice to modify market relations, in order to make them fairer and more conducive to a good life for all" (Sassatelli and Davolio, 2010: p. 205). Many initiatives seek to develop alternative food networks, liberate and centralize (small-scale) producers, and extend the alternative consumption movement more generally.

Globalization of the Slow Food Movement

The Slow Food movement has sought to globalize its vision of what an authentic relationship with food and those who produce it might be. For instance, the original concerns of the Slow Food movement which were so deeply rooted in Italian regional food cultures rapidly found resonance throughout Europe, and indeed further afield. Furthermore, some basic organizing models such as the convivium proved to be rapidly replicable and could easily spread. Finally, as the movement spread geographically through both developed and developing nations, centralized co-coordinating functions such as the Terra Madre sprung up which would help to hold this rather dispersed movement together. Whereas the Slow Food movement has a global presence, its members are still predominantly found in the Western world. Still, with its projects and initiatives the movement increasingly reaches out to developing nations as well.

As argued by Van Bommel and Spicer (2011), the Slow Food movement proved to be adept at creating a global

structure through shared discourses (such as eco-gastronomy), practices (such as the convivium), and institutions (such as Terra Madre). These discourses, practices, and institutions were replicable into other sectors such as the design of cities and towns (with the Slow Cities movement), travel (Slow Travel), and journalism (Slow Journalism). The result is a loosely connected 'Slow' movement throughout the world which seeks to address generalized problems with the hyper-industrialization of particular sectors through appeal to ideas and experiences of 'Slowness' (Andrews, 2008; Honore, 2004).

As Slow Food and other aligned movements have shifted from being rather local and limited concerns to becoming a global movement which crosses sectors, a number of challenges have appeared. The first is the ongoing tension between the concerns of local convivium and those of more centralized bodies such as national or international offices (e.g., Chrzan, 2004; Sassatelli and Davolio, 2010). For instance, a local convivium may want to pursue idiosyncratic issues that are out of line or in some cases directly contradictory. Furthermore, local convivia may feel like the creation of centralized offices with full time professionals working in areas such as campaign coordination or fund raising goes against the emphasis on local authenticity. The second major tension between local and global chapters is around issues about commercialism and in particular large international food companies (e.g., Sebastiani et al., 2013; Van Bommel and Spicer, 2011). The Slow Food movement was originally founded as a reaction to mass produced food. Many local convivia remain staunch champions of small local producers. However, to garner the resources necessary to support professional central functions and cross-cutting campaigns, the national or global offices of Slow Food have turned to sponsorships and partnerships with well-known international food brands. Some feel these collaborations to be a contradiction with the focus on small local producers and claim that the movement has been effectively co-opted by large corporations, government agencies, and international NGOs. Defenders of collaborations point out that working alongside these large institutions does not give away concerns for locality and authenticity, but affords the movement an opportunity to increase its impact and reach. The final tension exists between attempts to resist the increasing branding and commodification of foodways on the one hand, and the use of the Slow Food brand to promote the movement.

Many participants in Slow Food see the movement as a way of seeking to overcome the encroachment of commercial brands into the process of preparing and eating food. However, as it has grown, Slow Food itself has used many of the techniques of branding to popularize the movement. This branding has led some to express concerns that Slow Food will simply become an upmarket brand which does not genuinely embody many of the more radical aspects of sustainable food as a social movement (Donati, 2005).

Conclusion

Slow Food is an important movement that has challenged the apparent dominance of 'Fast Food' in many countries throughout the world. It has not only done this by establishing a new market niche and a series of institutions that support small farmers as well as adventurous gourmets, it has also articulated a new set of ideas around food, developed new experiences around how food might be consumed and produced, and perhaps most importantly forged new connections among a range of different parties interested in food. Through doing this, the movement has managed to create an alternative form of food production that is not solely driven by large commercial food producers. This has helped local foodways and traditions to be protected in the face of the ongoing march of industrialized food production.

While the Slow Food movement has been associated with a number of significant successes, it also continues to face new set of tensions. These include the tension between focusing on local foodways while also seeking to globalize the movement; between seeking to be authentic to local concerns and also trying to scale up through working with global institutions; and seeking to challenge commercial logics such as branding and attempts to popularize itself through building a widely recognized image and message. Many of these tensions are precisely the kind faced by similar social movements seeking to cater to the desires for locality and authenticity but at the same time build global connections. The ability to manage these tensions and contradictions will probably be decisive in shaping the future of the movement, and others like it. We therefore think that exploring how movements like Slow Food manage these tensions will prove to be a fruitful and exciting area of research in the future.

See also: Agriculture, Economics of; Alternative Food Movements; Ethical Consumers; Food Security and 'Green Revolution'; Food in Anthropology; Food, Sociology of; Social Movements: A Social Psychological Perspective.

Bibliography

Andrews, G., 2008. The Slow Food Story: Politics and Pleasure. Pluto Press, London.

Bové, J., Dufour, F., 2002. The World Is Not for Sale: Farmers against Junk Food. Verso, London.

Boyle, D., 2003. Authenticity: Brands, Fakes, Spin and the Lust for Real Life. Harper Collins, London.

de Bakker, F.G.A., den Hond, F., King, B.G., Weber, K., 2013. Social movements, civil society and corporations: taking stock and looking ahead. Organization Studies 34, 573–593.

Chrzan, J., 2004. Slow food: what, why, and to where? Food, Culture and Society: An International Journal of Multidisciplinary Research 7, 117–132.

DeSoucey, M., 2010. Gastronationalism food traditions and authenticity politics in the European Union. American Sociological Review 75, 432–455.

Donati, K., 2005. The pleasure of diversity in slow food's ethics of taste. Food, Culture and Society: An International Journal of Multidisciplinary Research 8, 227–242.

Fonte, M., 2006. Slow food's presidia: what do small producers do with big retailers? Research in Rural Sociology and Development 12, 203–240.

Honore, C., 2004. In Praise of Slow: How a Worldwide Movement is Challenging the Cult of Speed. Orion, London.

Laclau, E., 2005. On Populist Reason. Verso, London.

Lewis, D., Bridger, D., 2001. The Soul of the New Consumer: Authenticity-What We Buy and Why in the New Economy. Nicholas Brealey Publishing, London.

Melucci, A., 1996. Challenging Codes: Collective Action in the Information Age. Cambridge University Press, Cambridge.

Michels, R., 1958. Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy. Free Press, Glencoe.

Parkins, W., Craig, G., 2006. Slow Living. Berg Publishers, Oxford.
Peano, C., Sottile, F., 2012. Slow Food Presidia in Europe: A Model of Sustainability. Evaluation of the Sociocultural, Agri-Environmental and Economic Results. 2000–2012. Slow Food Foundation for Biodiversity, Bra.
Petrini, C., 2003. Slow Food: The Case for Taste. Columbia University Press, NY, USA.
Petrini, C., 2010. Terra Madre: Forging a New Global Network of Sustainable Food Communities. Chelsea Green Publishing, Vermont, USA.
Rao, H., Monin, P., Durand, R., 2003. Institutional change in Toque Ville: Nouvelle cuisine as an identity movement in French gastronomy. American Journal of Sociology 108, 795–843.
Reinecke, J., 2010. Beyond a subjective theory of value and towards a "fair price": an organizational perspective on Fairtrade minimum price setting. Organization 17, 563–581.
Ritzer, G., 2011. The McDonaldization of Society, sixth ed. Pine Forge Press, Thousand Oaks.
Sassatelli, R., Davolio, F., 2010. Consumption, pleasure and politics slow food and the politico-aesthetic problematization of food. Journal of Consumer Culture 10, 202–232.
Schlosser, E., 2002. Fast Food Nation: What the All-American Meal is Doing to the World. Penguin, London.
Sebastiani, R., Montagnini, F., Dalli, D., 2013. Ethical consumption and new business models in the food Industry. Evidence from the Eataly case. Journal of Business Ethics 114, 473–488.
Skinner, D., 2007. Organising Organic: A Foucauldian Analysis of the Regulation of Organic Food Production (Ph.D. thesis). University of Warwick.
Tarrow, S.G., 1994. Power in Movement: Social Movements, Collective Action and Politics. Cambridge University Press, New York.
The Times, 2006. Slow on the uptake.
Van Bommel, K., Spicer, A., 2011. Hail the snail: hegemonic struggles in the slow food movement. Organization Studies 32, 1717–1744.
Weber, K., Heinze, K.L., DeSoucey, M., 2008. Forage for thought: mobilizing codes in the movement for grass-fed meat and dairy products. Administrative Science Quarterly 53, 529–567.

Small Group Research, History of

Christian Dayé, University of Graz, Graz, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Small group research is a field of empirical research that observes the behavior of groups – and of individuals in groups. These can be natural groups observed in their habitat or 'artificial' groups convened in the laboratory. After blooming in the 1950s, the field receded into the background in both disciplines that originally contributed to it, social psychology and sociology. This contribution discusses the sources relevant to the establishment of the field in the 1940s, introduces three important traditions of small group research, reports some explanations for the field's recession, and presents some newer approaches to the study of small groups.

Introduction

Small group research is a field of empirical research that observes the behavior of groups – and of individuals in groups. These can be natural groups observed in their natural habitat – at the workplace, in the neighborhood, etc. – or 'artificial' groups convened in the laboratory, where for the most part, the group members do not know each other. The field was established in the first decades of the twentieth century, and from the very beginning it attracted researchers from both social psychology and sociology. Its establishment was accompanied with high hopes in the field's impact. It appeared that small group research would be able to develop into a specialty that could span these academic disciplines and guide the further development of both.

The nature of the task at hand requires some deliberate restrictions. Historically, small group research is of Northern American origin. This article is concerned mostly with the history of its early development in Northern America and does not discuss the international diffusion of ideas. For many European countries, however, it holds true that small group research was introduced by two cohorts of scientists. The first cohort consisted of people who emigrated in the course of World War II or earlier due to the Nazi seizure of power and returned to their home countries. The second cohort comprises younger researchers who received post-graduate training in the United States during the 1950s. A second restriction is that conceptual or theoretical differences between approaches must be painted with a big brush. The multitude of different conceptualizations of core notions, like, e.g., the group, and the intricacies that this multitude causes for the theoretical discussion must be neglected (but see Mills, 1967, pp. 10–23, for a thoughtful discussion of various mental models).

A recurrent narrative in the historiography of small group research is that after a first bloom in the 1950s, the field receded into the background. Within the disciplines contributing to it, groups conspicuously – and to most commentators spuriously – moved out of focus. Prior to that, however, small group research was a major player and also a signpost, providing orientation on both intellectual, methodological, and organizational issues to many other lines of research in the Social and Behavioral Sciences (SBS). An important year in the establishment of the field was 1954, when – after a series of influential monographs (especially Bales, 1950; Homans, 1950) and articles emanating from the work at a handful of research centers in the United States – special issues of the *American Sociological Review* (ASR 19 (6), December 1954) and *Sociometry* (17 (2), May 1954) were dedicated to small group research. The latter included an extensive bibliography compiled by Fred L. Strodtbeck and A. Paul Hare (1954), prepared with support by the Ford Foundation's Behavioral Science Division. It documented an increase in magnitude of items on small groups from 0.5 items published per year in the 1890–99 period to 152.5 items in the 1950–53 period. However, these collaborative efforts to create a field that transcends and unifies the various local traditions of small group research, however, were without success; in the 1970s, small group research appeared to be a 'light that failed' (Mullins, 1973, p. 105).

This article is structured as follows: after some remarks on theoretical and methodological sources, three traditions of small group research are introduced and their specificities discussed. The adjacent section reports a variety of arguments put forth in the literature to explain the field's loss of influence after 1970. Further, some contemporary attempts to reestablish small group research as a viable field of research are presented. The concluding remarks attempt an outlook on the future of the field.

Preparing the Ground

Theorizing the Primary Group

Although in many ways related, the history of small group research can be discussed without lengthy references to the conceptual history of the group within SBS or to the theoretical discourse undergirding the research activities. However, some remarks appear appropriate. Social theorizing about the group began earlier, but in the North American scene, Charles H. Cooley's thoughts on primary groups became highly influential and still count as a *locus classicus*. Cooley ([1909]1956, p. 23) defined primary groups as groups "characterized by intimate face-to-face association and coöperation." For the individual, primary groups form the most immediate sources of a socially anchored and therefore more stable worldview. Examples of primary groups include the family, the playgroup of children,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03232-3

the gang, and the neighborhood. They are embedded in society and "to some degree reflect its spirit" (Cooley, 1956, p. 27).

The moral backbone of Cooley's theoretical approach is a concern over the fate of the primary group in modern society. He sees traditional primary groups – mainly the family, but also the neighborhood – challenged by social forces in modern societies. "In our own cities the crowded tenements and the general economic and social confusion have sorely wounded the family and the neighborhood," Cooley ([1909]1956, p. 27) writes, and adds in allusion to social reform efforts that "there is nothing upon which the conscience of the time is more determined than upon restoring them to health." The notion of primary group respectively the perspective on the group as still vital source of social cohesion in modern societies has been influential in several later classic studies of American social science, including *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 1958[1918/1920]), *The American Soldier* (Stouffer, 1949), and *Personal Influence* (Katz and Lazarsfeld, 1955).

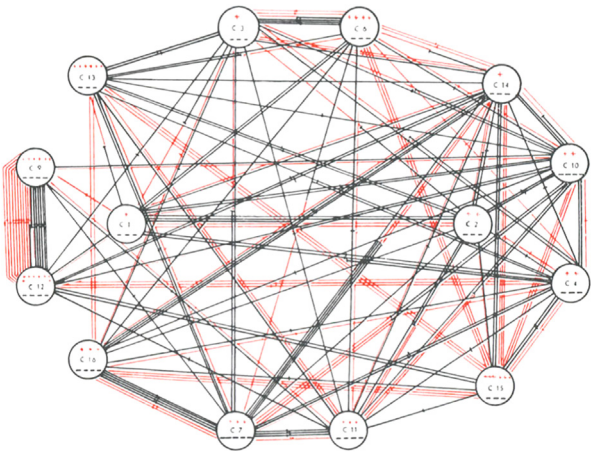

Figure 1 Sociometrical display. Source: Moreno, J.L., 1967. Die Grundlagen der Soziometrie. Wege zur Neuordnung der Gesellschaft, Zweite, erweiterte Auflage. West, Köln, Opladen, p. 256.

Sociometry and Other Early Research on Small Groups

A second source influencing the development of small group research was attempts to systematize the observation of human behavior in groups. Of these, Jacob L. Moreno's sociometry (or sociometrics) probably is the most prominent. While working as a physician in a refugee camp some 20 miles south of Vienna, Austria, during 1916 and 1917, Moreno began to register processes of attraction and rejection within groups. The methodology developed by Moreno in the following decades – mainly after his move to the United States in 1925 – combined a sophisticated psychological theory with a variety of new methods and techniques for observation and analysis, rounded up with a therapeutic thrust.

From the perspective of small group research, the most influential elements of Moreno's system were the methods of observation and analysis. Based on data collected by various means – ranging from direct observation whom a child choses to play with to more elaborate questionnaires – Moreno's *Who Shall Survive?* (1934) introduced visual displays of the relations of attraction, rejection, and indifference within groups (see **Figure 1**). Although less dynamic, with regard to analytical sophistication Moreno's sociometry went beyond the widely acclaimed contemporary studies on norm formation in groups by Muzafer Sherif (1936) or Theodore M. Newcomb (1942). Moreno's attempts to systematically gather data about structures and processes within groups and to display these data visually not only found its continuation in later research on small groups, but also informed social network analysis. However, he came to vigorously oppose the laboratory experiments of both the Lewin and the Bales groups, arguing that it was impossible to capture the dynamics of group life in such artificial, short-term situations (Moreno, 1967, pp. 406–408).

Groups at Work: The Hawthorne Experiments

A third source informing the development of small group research was the Hawthorne experiments, a series of loosely connected tests and experiments at a large factory a few miles west of Chicago, Illinois, that were carried out mainly between 1924 and 1932 (Roethlisberger and Dickson, 2003). Initially designed to investigate the effect of improvements in the working conditions (especially the lighting of the workplace) on the performance of the mainly female workers, the researchers found that what mattered for the productivity of a group was not the objective, presumably improved, conditions. But what then? The interpretation put forth by the team of researchers was that the increase of productivity was caused by the fact that the group was separated from the other workers and treated as something special. "What actually happened was that six individuals became a team and the team gave itself wholeheartedly and spontaneously to cooperation in the experiment. The consequence was that they felt themselves to be participating freely and without afterthought, and were happy in the knowledge that they were working without coercion from above or limitation from below" (Mayo, 1949, p. 64; see also Roethlisberger and Dickson, 2003).

Despite the fact that this interpretation has been fiercely contested ever since (Gillespie, 1991), the Hawthorne experiments received wide acclaim. Probably most important for the trajectory of small group research was that the Hawthorne experiments could be taken as an example of the applicability of social scientific research onto management problems. They established convincingly the relevance of group dynamics research for questions of nonacademic character, be it in business, politics, or the military.

Three Traditions of Small Group Research, 1940s–60s

During its bloom in the late-1940s and 1950s, small group research attracted many talented students. With justifiable restrictions, one can say that three traditions (or schools of thought) shaped the image of the field in these years (Arrow et al., 2000, pp. 11–19; McGrath, 1997). For convenience, these traditions are labeled after universities that played an important part in their consolidation. These are MIT/Michigan, Harvard, and Illinois. Across these traditions, and with

some remarkable exceptions – e.g., the studies of cognitive, dissonance, and informal communication by Festinger et al. (1950, 1956) – most research was carried out in the laboratory and applied experimental designs.

MIT/Michigan

The MIT/Michigan tradition can be traced back to Kurt Lewin's work on group dynamics that he had begun at the University of Iowa's Child Welfare Research Station in the 1930s (Cravens, 1993). Due to the success of his research, Lewin was asked to establish a program on groups at MIT. Lewin followed the call and in 1945 opened the Research Center for Group Dynamics (RCGD) in Cambridge, Massachusetts. After Lewin's premature and unexpected death 2 years later, struggles occurred over the future of the center, and ultimately, a group of RCGD researchers including Dorwin Cartwright, Leon Festinger, John R.P. French, Jr., Ronald Lippit, and Alvin Zander decided to move the center to the University of Michigan, Ann Arbor (http://www.rcgd.isr.umich.edu/history/, accessed 4 February 2014). The process was completed in 1948 and Cartwright became the first director of RCGD at Michigan.

For the small group researchers at Ann Arbor, groups were of interest as 'vehicles for influencing members' (Arrow et al., 2000, pp. 13, 14). They focused on the impact of group dynamics on the individual. Emphasis was put on the forces at work within groups that influenced the opinions, attitudes, or behavior patterns of its members. In the tradition of Ash, Festinger's (1950) work on informal communication started from the assumption that groups exert pressures toward uniformity among its members. These pressures can lead to a change of opinions, beliefs, and attitudes in accordance with the majority, to retention of open discussion of divergent views within the group, or to comparable reactions (see also Festinger et al., 1950, 1956).

Apart from those already mentioned, Kurt Back, Harold H. Kelley, Theodore M. Newcomb, Henry W. Riecken, Stanley Schachter, and John W. Thibaut provided significant contributions to the MIT/Michigan tradition of small group research. Its focus on the impact of groups and group dynamics on the individual members' opinions, beliefs, and attitudes is well documented in one of the tradition's most widely-read publications, *Group Dynamics*, a collection of essays and research articles on small groups edited by Dorwin Cartwright and Alvin Zander (1968, third edition; first edition 1953). How characteristic the emphasis on the impact of group dynamics on the individual was for the MIT/Michigan tradition becomes especially evident if one compares *Group Dynamics* to *Small Groups*, a collection published by researchers from Harvard (Hare et al., 1966; revised edition; first edition 1955).

Harvard

Research in the Harvard tradition did not focus so much on the group's impact on the individual, but rather on the interaction processes occurring within groups. Closer to Moreno's sociometry than the work from the MIT/Michigan tradition, Harvard researchers attempted to measure interaction within groups by quantitative means and to deduce from these measurements some information about the status of the individual group member and of the group as a whole. For instance, Eliot D. Chapple and Carleton S. Coon (1966), in a text included in the aforementioned *Small Groups* (Hare et al., 1966), suggested to test whether or not a group shows equilibrium by observing the rates of interaction between individual members of the group. If these rates remain constant and, after a disturbance, return to a level similar than before, the group has attained equilibrium.

Harvard had been the academic base of the Hawthorne experiments, and some members of its faculty thus were experienced in researching social groups and group effects. This experience obviously spread through many channels. Although its author was only a marginal participant in Harvard's academic life and had received its training at other places, William F. Whyte's study of gangs in Boston's North End, published as *Street Corner Society* in 1943, can in some respect be taken as an early example of the Harvard tradition of small group research. Especially Whyte's serendipitous finding that the results of a bowling evening reflected the status position of the group members shows the interest in group internal status structures that characterizes many works of Harvard affiliates. To the same extent that *Street Corner Society* was a community study (Andersson, 2014), it also was a study on group interaction and performance.

In 1950, two books appeared that consolidated a specific view of small group research at Harvard. The one was George C. Homans's (1950) *The Human Group*, a text that developed a set of core concepts and didactically applied it to the analysis of selected empirical studies, among them some of the Hawthorne experiments, Whyte's *Street Corner Society*, and anthropological studies of 'primitive' and 'modern' communities (e.g., by Bronislaw Malinowski and Raymond Firth respectively by W. Lloyd Warner, Conrad M. Arensberg, and Solon T. Kimball). It has been noted that the adoption of the small group as a core concept for cultural anthropology also happened in the Harvard orbit (Isaac, 2012; Roberts, 1951).

The other, probably even more influential work was Robert F. Bales's (1950) *Interaction Process Analysis*, in which Bales proposed to abstract from the actual content of group interaction and to categorize individual acts according to a predetermined scheme (see **Figure 2**). An act was either a positive reaction (showing solidarity, tension release, or agreement/passive acceptance); an attempted answer; a question, or a negative reaction (showing disagreement/passive rejection, tension, or antagonism). The following years saw Bales' scheme getting integrated into the general social theory developed at Harvard's Department of Social Relations under the leadership of Talcott Parsons (Parsons et al., 1953; especially chapters III and V). Theoretical relations were established with both the pattern variables (Parsons and Shils, 1951) and the then developed understanding of the AGIL scheme.

Apart from those already named, influential contributions came from Edgar F. Borgatta, A. Paul Hare, Fred L. Strodtbeck, and Theodore M. Mills. To all of them, groups were of interest because they "pattern[ed] human interaction" (Arrow et al., 2000, pp. 14–15).

During the 1960s and 1970s, researchers in the Harvard tradition gradually turned away from the aspired heights of a general social theory and became strongly oriented toward social psychology. Considerable effort was dedicated to

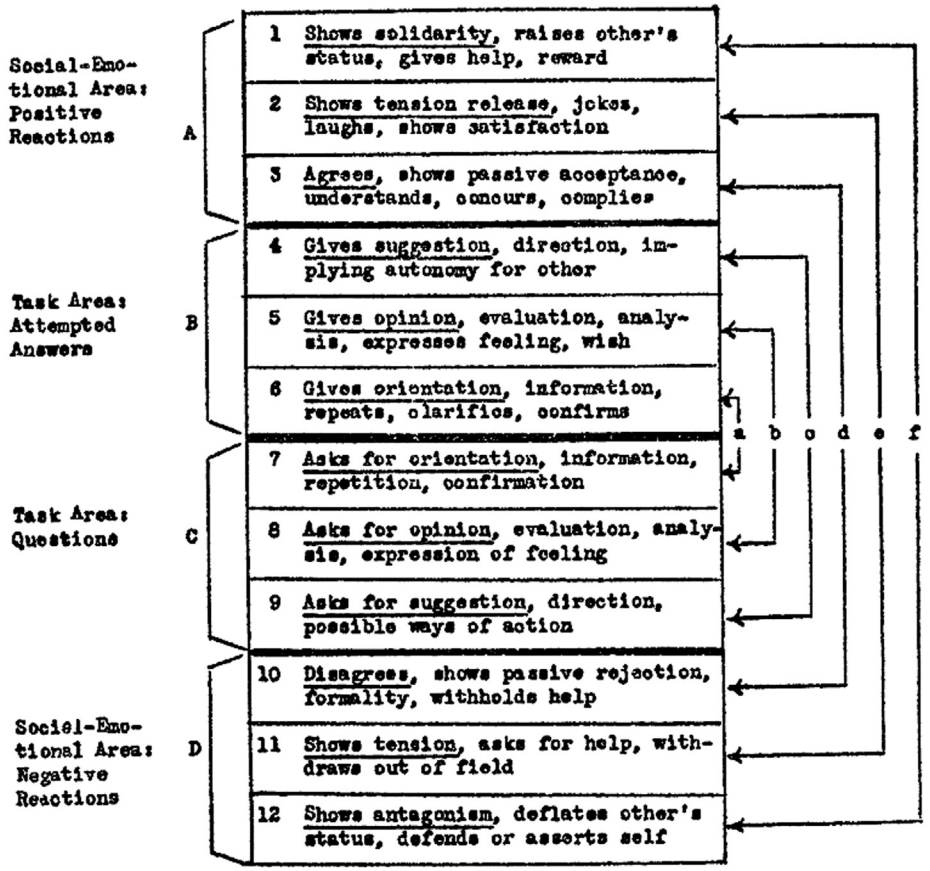

CHART. 1. The System of Categories Used in Observation and Their Relation to Major Frames of Reference.

Key:

a. Problems of orientation
b. Problems of evaluation
c. Problems of control
d. Problems of decision
e. Problems of tension-management
f. Problems of integration

Figure 2 Category system of interaction process analysis. Source: Bales, R.F., 1950. A set of categories for the analysis of small group interaction. American Sociological Review 15, 257–263, p. 258.

promoting the formalization and mathematization of small group research (Berger et al., 1962). After complementing the *Interaction Process Analysis* with a method for assessing value orientations (Bales, 1970), Bales, then director of Harvard's Laboratory of Social Relations, developed the Systematic Multiple Level Observation of Groups (SYMLOG; Bales and Cohen, 1979; Bales, 1999), which "provided the benchmark work on group interaction process for decades." (McGrath, 1997, p. 9) (The Laboratory of Social Relations was a research facility set up by Harvard's Provost conterminously with the Social Relations Department in January 1946, but remained formally independent from the latter.)

Illinois

In contrast to the other two, researchers in the Illinois tradition were more concerned with the quality of group performance. They were interested in groups as "vehicles for performing tasks" (Arrow et al., 2000, pp. 15–16). Authors related back to Floyd Allport's theory of social facilitation, i.e., the phenomenon that individuals perform better in the presence of others performing the same task (Allport, 1920). Allport's approach to social psychology was individualistic and "behaviorist in spirit" (Goodwin, 2012, p. 467). Allport "denied the reality of all traditional social objects" and acknowledged only concepts relating to the individual. Any attempt to explain phenomena by emergent properties of a group was derided by him as "the fallacy of the group" (Danziger, 2000, p. 333).

It has been noted that work in the Illinois tradition "is more diffuse" (McGrath, 1997, p. 10); to a lesser degree than the other traditions have there been a coherent body of research or an intellectual program. Apart from the general perspective on what effects group performance, this tradition was not united by a coherent theoretical system. If there is something comparable to the handbook-like publications in the other traditions – *Group Dynamics* for MIT/Michigan, *Small Groups* for Harvard – then this would be Joseph E. McGrath and Irwin Altman's *Small Group Research* (1966). This book, however, was received as a critical assessment of the field rather than as a signpost for orienting further research

projects. Many projects received support from military funds (e.g., Altman and Haythorn, 1965; Carter et al., 1950; Davis and Hornseth, 1967) and, at least to some extent, followed the interests of the patrons. Research dealt with leadership, communication patterns, problem solving, and decision-making in mock juries (e.g., Nagao and Davis, 1980) and other small groups, for instance in sports and the military.

Why Did Small Group Research Recede?

Many observers have commented on the decline of small group research around 1970, and the literature contains a variety of alternative explanations. These range from theoretical to sociological and cultural explanations.

Theoretical Explanations

One explanatory approach puts an emphasis on the theoretical level. For instance, it has been argued that at the time of their decline, the three traditions had reached inbuilt theoretical and methodological limitations. What had been strengths in the process of establishing the tradition had turned into weaknesses, hindering a fruitful development. The input–output approach used both in the MIT/Michigan and in the Illinois tradition and their demand for experimental rigor led to the neglect of natural groups and of the observation of intragroup processes. The fact that Harvard's IPA, on the other hand, was so profoundly integrated in Parsonian social theory made it less flexible and difficult to adapt to research outside this specific blend of structural functionalism (McGrath, 1997, pp. 10–12).

From the supply side, a new paradigm – social cognition – began to dominate the attention space within social psychology around 1970 (North and Fiske, 2012). With its perspective of humans as information processors and its interest in processes of information encoding, storing, and retrieval, social cognition provided an innovative alternative paradigm and thus indirectly contributed to the decline of small group research (see **Figure 3**). As regards sociology, it has been argued that the relevance of the small group has since the 1970s been paradoxically undermined both by the spread of ethnographic observation and the increasing focus on culture. Although in principle, small groups would have given these new perspectives a natural focus, theorizing went beyond the local spheres. Studies of culture researched groups as mirrors of the general culture in a society, and ethnographies – apparently following Geertz's (1973, p. 22) statement that anthropologists do not study villages, they study *in* villages – were more concerned with exploring more general social phenomena than with describing the processes of social organization within groups. In both cases, "the interactional basis of sociality was downplayed. Groups were studied, but they were not theorized" (Fine, 2012, p. 4).

Sociological Explanations

By the end of the 1950s, small group research appeared to be a fractionated field. There was only little intellectual exchange among the clusters that made up the field (Mullins, 1973, pp. 103–128), and small group research lacked the social fundament required to become a successful and stable specialty within SBS. None of those researchers working on small groups successfully synthesized the various findings and concepts into a coherent and innovative theory, and no one stepped in as the field's intellectual and organizational leader. The various handbooks are a quick proof of the field's lack of theoretical focus: the field's disparity was urgently felt by the small group researchers, but they still could not agree on a single theoretical framework. To Mullins (1973, p. 123), the "causes underlying the failure [of small group research] are both social and intellectual." The causes he lists include: (1) the lack of graduate students; (2) social psychology's interstitial status between sociology and psychology; (3) the early loss of a potential intellectual leader (Kurt Lewin) and the failure to find a competent replacement; and (4) the lack

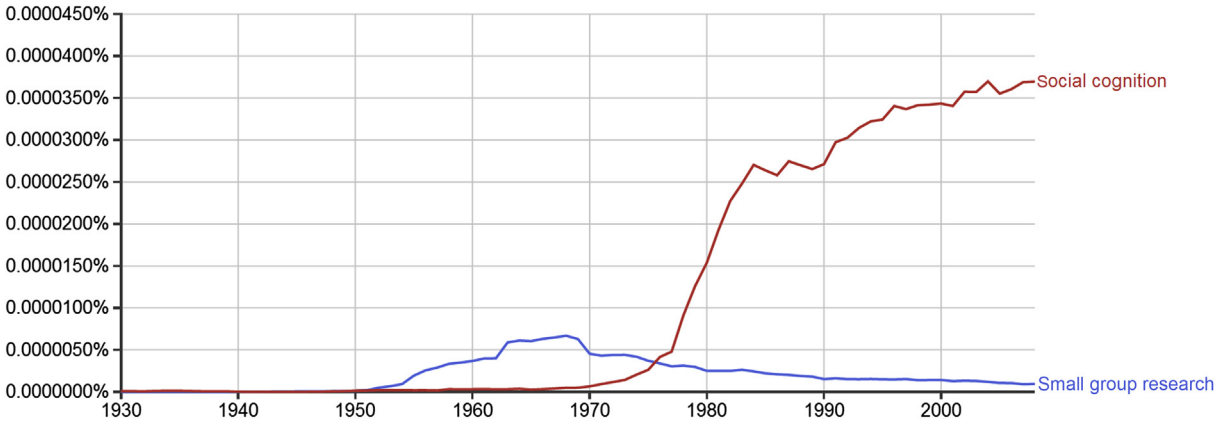

Figure 3 Ngram results for 'small group research, social cognition'. Source: https://books.google.com/ngrams/graph?content=Small+group+research%2Csocial+cognition&year_start=1930&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2CSmall%20group%20research%3B%2Cc0%3B.t1%3B%2Csocial%20cognition%3B%2Cc0, phrase 'small group research, social cognition,' corpus English, years 1930–2008, accessed 12 February 2014. Google Books Ngram Viewer displays a graph showing how often specified phrases have occurred in a corpus of books (e.g., 'British English,' 'English Fiction,' 'French') over the selected years. Clearly not meant to surrogate for a more sophisticated analysis, Ngram Viewer indicates the trajectories of the selected terms or phrases.

of either a program statement or any kind of research direction from a senior, widely respected scholar.

One might also hypothesize whether military and government funding for small group research diminished in the course of the 1960s. With the rise of the computer and the subsequent renaissance of the cybernetic vision of a centralized command and control system, the small group – the team, the squad – might have been seen as becoming less important to military success. However, this hypothesis has not yet been corroborated by sophisticated research.

Cultural Explanations

Several commentators have linked the decline of small group research to cultural implications of the notion of group. I have alluded to some implications above in the passages on Charles H. Cooley. They reverberate many of the connotations of Friedrich Tönnies's distinction between *Gemeinschaft* and *Gesellschaft*, with groups clearly falling on the *Gemeinschaft* side of the dichotomy: they are places for the warm, natural, small, potentially intimate forms of interaction that are so urgently needed to counter the cold, objective, rational, and rationalized forms of interaction in society. On the other hand, society lacks the intimate social control of communities and thus offers the individual more freedom and higher autonomy. Furthermore, groups can lead members onto deviant careers (Park et al., 1984) or could become so important that the 'members' strivings for unanimity override their motivation to realistically appraise alternative courses of action and they engage in groupthink (Janis, 1972). Whether groups are 'a force for good or evil' is a recurrent issue in the history of SBS and philosophy (Levine and Moreland, 2012). And in winding ways, some observers commented, has the fate of small group research been related to the standing of the group within contemporary cultural discourses.

Another form of cultural explanation has been proposed by Ivan D. Steiner, who denied that the decline of group research had something to do with the weakness of the theoretical concept. Rather, he sarcastically proposed, "[i]magine we are viewing a swarm of crawling ants, 99% of whom are moving in the same general direction. We are immediately led to ask what's wrong with those few deviant ants. […] My point is that when society is serene, and only a few wrong-headed deviants disturb our tranquility, we focus our attention on individuals or on large organizations. But when many small segments of society are vying with one another, our attention is drawn to units of an intermediate size. We go where the action is: the prison courtyard, the commune, the assembly line, the committee room, or wherever people assemble and discuss, plot, act, or commiserate. And we find ourselves studying such nebulous things as leadership, cohesion, social facilitation, role systems, and all the rest" (Steiner, 1974, pp. 104–105).

Contemporary Attempts to Revive Small Group Research

While the results presented in **Figure 3** suggest a continuous decrease across the disciplines, studies investigating journal publications conclude that small group research is continued, but moved from social psychology to organizational psychology (Moreland et al., 1994; Sanna and Parks, 1997; see also Stewart et al., 1997). Journals like *Small Group Research* (published since 1970) and *Group Dynamics* (published since 1997) have tried to unify the still scattered field by providing a central publication outlet, but their success in doing so is disputed.

Regardless of its disciplinary location, however, the question has been raised whether small group research requires a new theoretical paradigm in order to regain intellectual attraction and to meet the promises of the field's heyday (Burlingame and Fuhriman, 1997; McGrath, 1997). Out of these discussions emerged an approach that, synthesizing a variety of theoretical elements into a comprehensive paradigm, views groups as complex, adaptive, dynamic systems (Arrow et al., 2000; McGrath et al., 2000): "Rather than simple, groups are complex entities embedded in a hierarchy of levels and characterized by multiple, bidirectional, and nonlinear causal relations. Rather than isolated, groups are intricately embedded within, and have continual mutual adaptation with, a number of embedding contexts. Rather than static, groups are inherently dynamic systems, operating via processes that unfold over time, with those processes dependent both on the group's past history and on its anticipated future. Groups develop as systems over time, and change as a function of changing conditions over time" (McGrath et al., 2000, p. 98).

Another, in many respects complementary, attempt to revive small group research within sociology has been initiated by Gary Alan Fine (especially 2012). Fine conceives of small groups as linking individuals and civil society. They are 'tiny publics' that depend on local cultural understandings – an 'idioculture,' to use Fine's term – that emerge in the life course of the group. Fine studied a broad variety of groups, ranging from Little League baseball teams to fantasy role-playing gamers and mushroom collectors. Fine's sociological approach meets with the emphasis put by psychologists on the dynamic change of groups over time; as a consequence, contemporary small group research appears to be concerned more with natural groups than with adhoc groups in laboratory settings.

Another example of how small group research has recently been used to shape an interesting and innovative argument is Michael P. Farrell's work on collaborative circles. Collaborative circles combine "the dynamics of a friendship group and a work group. At the core is a set of friends who, over a period of time working together, negotiate a shared vision that guides their work" (Farrell, 2001, p. 7). Farrell explores the characteristics of such groups by a series of historical analyses of collaborative circles in art, literature, science, and politics. These analyses show that a set of roles (the leader, the devil's advocate, the peacemaker, the lightning rod, etc.) can be discerned and that usually over time, the members fulfilling these roles change. The development of collaborative circles typically follows a sequence of several stages. After the formative stage, where the core members are introduced to each other, they find themselves united in a rebellion against an authority. In the third stage, the quest stage, they begin to negotiate a new vision that will guide their work in the years to come. This is followed

by a phase of creative work, in which the new vision is explored and developed. In some cases, a fifth phase sees groups take collective action – to organize an exhibition, to found a literary journal, etc. In part because of the efforts and constraints on the individual required for organizing collective action, the sixth phase is typically the separation stage. Finally, group members may come together later in life for a nostalgic reunion stage. Farrell's work is innovative in that it formulates a concept useful for historical analysis that is sophisticatedly anchored in small group research and theorizing. Again, he is concerned with real existing groups, but his data are not derived from participant observation, like e.g., Fine's, but from careful analysis of historical sources – diaries, letters, pamphlets, etc.

Conclusion

A brief outlook on the future of small group research might be appropriate. An increase in scientific concern with small groups will certainly depend on the fruitfulness of the theoretical paradigms available. These must take into account the basic philosophical changes occurring since the 1970s that have left their marks on SBS disciplines, albeit in differing degrees. Although not all aspects of constructivism, postpositivism, poststructuralism, and other lines of critical discourse are relevant, those main points that have entered the mainstream methodological discourse in the SBS should be acknowledged and integrated into the research methodology. The examples introduced in the preceding section do this in at least three ways. They implicitly or explicitly focus on natural groups, emphasize the relevance of time, and construct the relationship between the individual and the group as dynamic phenomena emerging out of interaction. The very concept of membership is problematized. Individuals are not seen as mere group members, but as persons participating in groups. Whether these new theoretical emphases will lead to a rise in the social and infrastructural resources and to the 'critical mass' required to a return of small group research to the forefront of SBS is an open question (Poole et al., 2004).

See also: AGIL, History of; Cognitive Psychology: History; Community and Society: History of the Concepts; Definition of the situation: History of the Concept; Experimentation in Psychology, History of; Network Analysis, History of; Psychology, History of (Twentieth Century); Quantification in the History of the Social Sciences; Quasi- and Field Experiment; Reference Group, History of.

Bibliography

Allport, F.H., 1920. The influence of the group upon association and thought. Journal of Experimental Social Psychology 3, 159–182.

Altman, I., Haythorn, W.W., 1965. Interpersonal exchange in isolation. Sociometry 28, 411–426.

Andersson, O., 2014. William Foote Whyte, Street Corner Society and social organization. Journal of the History of the Behavioral Sciences 50, 79–103.

Arrow, H., McGrath, J.E., Berdahl, J.L., 2000. Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation. SAGE Publications, Thousand Oaks, CA.

Bales, R.F., 1950. Interaction Process Analysis. Addison-Wesley Press, Cambridge, MA.

Bales, R.F., 1970. Personality and Interpersonal Behavior. Holt, Rinehart and Winston, New York; Chicago; San Francisco, etc.

Bales, R.F., 1999. Social Interaction Systems: Theory and Measurement. Transaction Publishers, New Brunswick, NJ, London.

Bales, R.F., Cohen, S.P., 1979. SYMLOG: A System for the Multiple Level Observation of Groups. Free Press, New York.

Berger, J., Cohen, B.P., Snell, J.L., Zelditch, M.J., 1962. Types of Formalization in Small-Group Research. With a foreword by Sanford M. Dornbusch. Houghton Mifflin, Co., Boston.

Burlingame, G.M., Fuhriman, A., 1997. Small group research and dynamical systems theory: conceptual and methodological considerations, a reply to McGrath. Group Dynamics: Theory, Research, and Practice 1, 28–31.

Carter, L., Haythorn, William, Howell, Margaret, Haythorn, W.W., Howell, M., 1950. A further investigation of the criteria of leadership. Journal of Abnormal and Social Psychology 45, 350–358.

Cartwright, D., Zander, A. (Eds.), 1968. Group Dynamics: Research and Theory, third ed. Harper & Row, New York; Evanston; and London.

Chapple, E.D., Coon, C.S., 1966. The equilibrium of groups. In: Hare, A.P., Borgatta, E.F., Bales, R.F. (Eds.), Small Groups: Studies in Social Interaction. Alfred A. Knopf, New York, pp. 54–57.

Cooley, C.H., 1956. The Two Major Works of Charles H. Cooley: Social Organization, & Human Nature and the Social Order. With an Introduction by Robert Cooley Angell. The Free Press, Glencoe, IL.

Cravens, H., 1993. Before Head Start: The Iowa Station & America's Children. The University of North Carolina Press, Chapel Hill, London.

Danziger, K., 2000. Making social psychology experimental: a conceptual history, 1920–1970. Journal of the History of the Behavioral Sciences 36, 329–347.

Davis, J.H., Hornseth, J., 1967. Discussion patterns and word problems. Sociometry 30, 91–103.

Farrell, M.P., 2001. Collaborative Circles: Friendship Dynamics & Creative Work. The University of Chicago Press, Chicago, London.

Festinger, L., 1950. Informal social communication. Psychological Review 57, 271–282.

Festinger, L., Riecken, H.W., Schachter, S., 1956. When Prophecy Fails. University of Minnesota Press, Minneapolis.

Festinger, L., Schachter, S., Back, K.W., 1950. Social Pressures in Informal Groups: A Study of Human Factors in Housing. Harper, New York.

Fine, G.A., 2012. Tiny Publics: A Theory of Group Action and Culture. Russel Sage Foundation, New York.

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Gillespie, R., 1991. Manufacturing Knowledge: A History of the Hawthorne Experiments. Cambridge University Press, Cambridge, England.

Goodwin, C.J., 2012. A History of Modern Psychology, fourth ed. John Wiley & Sons, Hoboken, NJ.

Hare, A.P., Borgatta, E.F., Bales, R.F. (Eds.), 1966. Small Groups: Studies in Social Interaction. Alfred A. Knopf, New York.

Homans, G.C., 1950. The Human Group. Harcourt, Brace and Company, New York.

Isaac, J., 2012. Epistemic design: theory and data in Harvard's department of social relations. In: Solovey, M., Cravens, H. (Eds.), Cold War Social Science: Knowledge Production, Liberal Democracy, and Human Nature. Palgrave Macmillan, New York, pp. 79–95.

Janis, I.L., 1972. Victims of Groupthink. A Psychological Study of Foreign-Policy Decisions and Fiascoes. Houghton Mifflin, Oxford.

Katz, E., Lazarsfeld, P.F., 1955. Personal Influence; the Part Played by People in the Flow of Mass Communications. Free Press, Glencoe, IL.

Levine, J.M., Moreland, R.L., 2012. A history of small group research. In: Kruglanski, A.W., Stroebe, W. (Eds.), Handbook of the History of Social Psychology. Psychology Press, New York, pp. 383–405.

Mayo, E., 1949. The Social Problems of an Industrialized Civilization. With an Appendix on the Political Problem. Routledge & Kegan Paul, London.

McGrath, J.E., 1997. Small group research, that once and future field: an interpretation of the past with an eye to the future. Group Dynamics: Theory, Research, and Practice 1, 7–27.

McGrath, J.E., Altman, I., 1966. Small Group Research: A Synthesis and Critique of the Field. Holt, Rinehart and Winston, New York; Chicago; San Francisco, etc.

McGrath, J.E., Arrow, H., Berdahl, J.L., 2000. The study of groups: past, present, and future. Personality and Social Psychology Review 4, 95–105.

Mills, T.M., 1967. The Sociology of Small Groups. Foundations of Modern Sociology Series. Prentice-Hall, Englewood Cliffs, NJ.

Moreland, R.L., Hogg, M.A., Hains, S.C., 1994. Back to the future: social psychological research on groups. Journal of Experimental Social Psychology 30, 527–555.

Moreno, J.L., 1934. Who Shall Survive? A New Approach to the Problem of Human Interrelations. Nervous and Mental Disease Pub. Co., Washington.

Moreno, J.L., 1967. Die Grundlagen der Soziometrie. Wege zur Neuordnung der Gesellschaft, Zweite, erweiterte Auflage. West, Köln, Opladen.

Mullins, N.C., 1973. Theories and Theory Groups in Contemporary American Sociology. Harper & Row, New York.

Nagao, D.H., Davis, J.H., 1980. The effects of prior experience on mock juror case judgments. Social Psychology Quarterly 43, 190–199.

Newcomb, T.M., 1942. Community roles in attitude formation. American Sociological Review 7, 621–630.

North, M.S., Fiske, S.T., 2012. A history of social cognition. In: Kruglanski, A.W., Stroebe, W. (Eds.), Handbook of the History of Social Psychology. Psychology Press, New York, pp. 81–99.

Park, R.E., Burgess, E.W., McKenzie, R.D., 1984. The City: Suggestions for Investigation of Human Behavior in the Urban Environment. The University of Chicago Press, Chicago; London.

Parsons, T., Bales, R.F., Shils, E.A., 1953. Working Papers in the Theory of Action. Free Press, Glenco, IL.

Parsons, T., Shils, E.A., 1951. Values, motives, and systems of action. In: Parsons, T., Shils, E.A. (Eds.), Toward a General Theory of Action. Harvard University Press, Cambridge, MA, pp. 45–276.

Poole, M.S., Hollingshead, A.B., McGrath, J.E., Moreland, R.L., Rohrbaugh, J., 2004. Interdisciplinary perspectives on small groups. Small Group Research 35, 3–16.

Roberts, J.M., 1951. Three Navaho Households: A Comparative Study in Small Group Culture, Papers of the Peabody Museum of American Archeology and Ethnology. Harvard University, Peabody Museum; Cambridge, MA.

Roethlisberger, F.J., Dickson, W.J., 2003. Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago, the Early Sociology of Management and Organizations. Routledge, London; New York.

Sanna, L.J., Parks, C.D., 1997. Group research trends in social and organizational psychology: whatever happened to intragroup research? Psychological Science 8, 261–267.

Sherif, M., 1936. The Psychology of Social Norms. Harper & Brothers, New York; London.

Steiner, I.D., 1974. Whatever happened to the group in social psychology? Journal of Experimental Social Psychology 10, 94–108.

Stewart, A.E., Stewart, E.A., Gazda, G.M., 1997. Assessing the need for a new group journal. Group Dynamics: Theory, Research, and Practice 1, 75–85.

Stouffer, S.A., 1949. The American Soldier, 2 vols. Princeton University Press, Princeton.

Strodtbeck, F.L., Hare, A.P., 1954. Bibliography of small group research (from 1900 through 1953). Sociometry 17, 107–178.

Thomas, W.I., Znaniecki, F., 1958. The Polish Peasant in Europe and America, 2 vols. Dover, New York.

Small States, History of

Richard Whatmore, School of History, University of St Andrews, St Andrews, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Small states have been the subject of historical enquiry since the writings of Herodotus and Thucydides. Intertwined with the history of empire, the survival or demise of small states has regularly inspired patriotic encomium or tragic lament. Small states threatened by large empires or ambitious neighbors developed numerous survival strategies including alliance, the balance of power, economic specialization, and reliance upon military manliness. All of these were failing by the mid-eighteenth century, leading to a crisis of the small states and the development of new strategies to maintain them in a world increasingly dominated by large commercial empires.

Introduction

Small states have always been threatened by destruction at the hands of more powerful and larger neighbors. That small states were forced in consequence to create survival strategies that might cope with the power imbalance is clear from the earliest histories of Herodotus and Thucydides. The greatness ascribed to the Greek city-states derived in part from a willingness to stand up to the imperial ambitions of Persian Emperors. This pattern has been repeated throughout the recorded history, inspiring glory and causing tragedy in equal measure. The desire for empire and the capacity to establish it can be correlated with the crises of the small states, for whom empire was either inconceivable or undesirable. In Europe more particularly, because of the number of communities across the continent at different times, and because of the lust for empire in the larger states, it might be said that the relationship between small and large states has been still more significant in the shaping of history. The argument can be taken further; the history of modern Europe, with its wars for empire and unions of states in the hope of peaceful coexistence, has concerned the relationship between small and large states. This article identifies some of the strategies frequently adopted by small states in order to survive. It also describes the crisis of the small states in eighteenth-century Europe, after which small states across the globe have found survival more difficult still in the face of large empires with greater political, military, and economic power.

Historical Overview

In the first centuries of the Christian era small states, significantly always likened to faith communities by outsiders, looked to the Roman Empire for their independence. The possibility of maintaining peace by replicating the 'Roman' circumstances of a hegemonic power protecting lesser states was deemed a myth by certain observers in every generation. Yet, however contested, notions of a *pax Romana*, pursued from the inauguration of Augustus as Emperor to the death of Marcus Aurelius, endured into modern times, directly in the form of aspirations to universal monarchy, and indirectly through distinctively Roman ideas about the assimilation of colonies, centering on the introduction of Roman laws and customs to the subject state in return for political privileges. One major fear for small states viewing the history of Rome was that the Romans had tended to protect a state prior to its ultimate subjection. Rome illustrated classic cases of the translation of friendship into dominion, or the transition from a relationship of mutual protection into one of subservience. Overidentification with the culture and politics of a large state was always to be feared. Overreliance upon the economy of a larger neighbor was equally worrying; concentrated economic health was tempting to a large state, which might stimulate its own industries and pay for any aggrandizement by the incorporation of the smaller power.

From the perspective of the small state, trust in its larger neighbor had to be engendered and then maintained. Friendship had to be balanced against the necessary distance associated with continued sovereignty. The tightrope walk of small-state politics always gave way sooner or later, with no small state maintaining its independence in Europe in perpetuity after the fall of Rome. Patterns of policy emerged to prolong the lifespan of small states. Indeed, writers in small states always speculated about the likely future of their state, and this entailed related speculation about the possible futures of larger powers. Comparative study of large states and empires in consequence has tended to be the preserve of authors from small states, as illustrated by such authors as Hugo Grotius, Samuel Pufendorf, Emer de Vattel, and Jean-Charles-Léonard Simonde de Sismondi. These authors recognized that their views were likely to have a large impact upon the policy developed by small-state governments. Several of them were patronized by politicians with a direct interest in shaping policy in a political world of short-term commitments and unforeseen developments.

Prospects for national survival were compounded by uncertainty as to whether neighboring large states were capable of providing genuine security. After the collapse of Rome in the west it was often the case that no major state was capable of aspiring to universal monarchy. Alternatively, universal monarchy was deemed too dangerous a strategy for protecting weaker states, and numerous small states and communities within national sovereignties turned to the Papacy or the Holy Roman Emperor for protection. The history of the Italian maritime republics during the Middle

Ages is instructive here, as the major republics Amalfi, Genoa, Pisa, and Venice, and the lesser maritime powers of Ancona, Gaeta, Molfetta, and Trani, were perpetually involved in securing themselves with the support of shifting coalitions of neighboring or regional powers. The inconstancy of political rule meant that imperial power frequently could not be relied upon to guarantee peace. This was the case in the turbulent fifteenth and sixteenth centuries. Republics such as Genoa then turned to eminent local dynasties such as the Visconti and the Sforza, local princess, French Kings and Emperors for their survival, and in the case of Venice even the Turks. Republics accordingly entered into complicated relationships with larger empires. For example, Ragusa (Dubrovnik) was at various times within the empire of Venice, Hungary, and the Ottomans, while perceiving itself to be an independent republic. As Jeremy Bentham put it in a letter to Prince Adam Czartoryski (21 June 1815), when arguing that his reform proposals were applicable to all kinds of state, "Under the Great Turk was not Ragusa even a republic?"

Complexities of national survival were further compounded by the Reformation, after which small Protestant or Catholic states used the existence of a shared international religion as the basis of preservative offensive or defensive leagues. The impact of Protestant and Catholic leagues of states in international relations after the sixteenth century is a neglected topic in the historiography of political thought. In early modern times more radical sects envisaged communities of the faithful who turned their backs upon contemporary politics at the national level. Small states were often described as churches and the history of certain religious movements to maintain themselves evinced similar concerns and addressed related problems to those experienced by faith communities. One obvious case was that of William Penn and the Quakers. Another was the Pietist Béat Louis de Muralt, who speculated in the 1720s on the relationship between Britain and France for small states before coming to the conclusion that the gradual abandonment of society was the right course for the faithful. In the eighteenth century a variant developed with freemasonry, whose devotees were instructed to accept that the whole world was one great republic, in which every nation was a family, and every particular person a child. For the small communities that continued to exist in state form, alliance continued to be the first means to maintain national independence.

Survival by Alliance

Diplomatic history has today become the central means employed to understand the vicissitudes of the relationship between small states and the monarchical empires of Europe in early modern times. This is right and proper. Contemporaries who were as fascinated by the up and down relationship of major and minor powers looked to the practices of ambassadors, the policy of courts, and the resulting renewal or alteration of established treatises and historic alliances. At times when no power in Europe was seen to be capable of holding the ring, or when the larger states had a manifest and long-term interest in maintaining the *status quo*, sizable states were similarly called upon to play the role of small-state protector in individual cases. Accordingly, much of the history of small states in Europe was concerned with alliances or treaties of peace, either to increase the combined strength of smaller states against larger neighbors, or to shelter within the shadow of an allied empire that was sufficiently powerful not to be overwhelmed by others.

The great means to self-preservation by keeping a general peace was deemed to be the balance of power. From this perspective large states were often described in early modern times as levers within a greater political balance that would ensure peace by pitting states of equal weight against one another. The policy David Hume advised to maintain the balance of power (in an essay of the same name) was alliance with weaker states against the stronger, following the rule pursued by Athens in the war between Thebes and Sparta who 'threw themselves into the lighter scale, and endeavoured to preserve the balance.' Yet as Hume was writing the relative disparity of power between large states meant that the option of treating small states as bulwarks against invasion, or protectorates marking spheres of influence, had come to be questioned. Partly this was due to the nature of war, namely the capacity of able generals to overrun extensive territories with greater speed than in the past. The difficulty of this strategy for the small states of the eighteenth century was twofold. Firstly, the complicated balance of power established by the Treaty of Westphalia (1648) rested on a series of renewable alliances, the breaking of which would have been a direct challenge to the entire system.

Alliance was in consequence a policy undermined by powerful states, so great had the gulf in resources between small and large states become. The most advanced military technology was not seen to be capable of deterring large states' military ambition. Indeed, the solid monopoly by the commercial monarchies of the latest innovations in making war was striking. The bombardment of Genoa with mortars and hotshot from the field guns of the French general Louvois in 1684, simply to apply diplomatic pressure, illustrated the trend. Improvements in the science of siege warfare, exemplified by the career of Sébastien le Prestre de Vauban, made it far easier for monarchies to challenge small states dominated by single cities. As Jean-François Saint-Lambert wrote in the *Encyclopédie* article 'Législateur' (1765), because of their unity of purpose and speed of action, monarchies had an enormous advantage over republics in the pursuit of war.

In alliances between small and large states, smaller states, and especially smaller republics, weaker parties frequently offered economic benefits, such as the stimulation of large-state trade, access to international fairs or markets and, by the eighteenth century, investment in the public debts of large states. When Emperor Frederick Barbarossa considered reviving the empire of Charlemagne and abolished the liberties of the free cities to this end in 1155, he guaranteed the protection of the maritime republics in return for commercial subsidies. In the view of the Scots Jacobite Chevalier Andrew Michael Ramsay, writing in his *Travels of Cyrus* (1730), small states could best survive by commerce, illustrated by the cases of Phoenicia, whose "strength consists in making herself useful and even necessary to all her neighbours," and Tyre, "where commerce is the only support of the state, all the principal citizens are traders; the merchants are the Princes of the republic." The great danger for trading republics was, of course,

the decline of commerce, which "so powerfully operated to the annihilation of the Italian republics as maritime powers."

For some observers the problem was the fluctuating nature of trade itself. The same fate met the 'petty republics of the Hanse towns,' which had 'enriched themselves so greatly by the divisions between Sweden and Denmark' but declined after the union of Calmar. To the eighteenth-century historian, the histories of Crete and Carthage equally revealed that moral, political, and economic dangers were posed for small states when commerce was accompanied by luxury, as was likely to be the case when supplying larger states with trade.

Survival by Virtue

Small states naturally claimed resources of their own to secure their liberty. The major source of self-defense derived from the bonds generated by the very smallness of such communities. Inhabitants identified independence as a supreme political value. Military defense required manliness. Historically, this was most directly associated with the species of small states that called themselves republics. The claim that "virtue and public spirit are the necessary supports of all republican governments," as Montesquieu put it in his *Spirit of the Laws* (1748), was most often cited as the central succor for such states. Every history of a republic that described war against a more powerful enemy employed the trope of martial courage displayed by its citizens because of their patriotic fervor and love of freedom. Ancient examples spoke directly to contemporary times in such histories, with the great examples of the Spartan defense of Greece against Xerxes at Thermopylae and the Athenian destruction of Darius' vast forces at the battle of Marathon finding recent parallels, such as the Dutch independence struggle, which astonished Europe in wars against Spain and then against Louis XIV. In such republics virtue explained their victories and continued survival.

Republican valor embodied manliness. For writers following Livy and Machiavelli, the best way to maintain such a state's independence was to prepare the citizenry for war and plot to create a republican empire, since, as Machiavelli wrote in the second book of his *Discourses*, "Rome made itself strong and powerful by destroying the states round about it, and incorporating strangers within its Citizens." Such claims greatly influenced eighteenth-century thought as John Pocock's *The Machiavellian Moment* abundantly revealed. The sometimes neglected counterclaim, to be found in Pocock's later work, was that all princes ought to protect virtue and that virtue was nowhere the preserve of small states and republics alone. Small states that were not republics equally drew upon the resources of martial virtue and patriotism in their attempts to defend themselves. But what Thomas Paine termed the 'gigantic manliness' associated with republics remained their central means of self-defense and the presumed basis of any success in war.

In *Common Sense* (1776) Thomas Paine wrote, "when republican virtue fails, slavery quickly ensues." Paine was discussing the weakness of Britain, whose constitution was sickly because "monarchy hath poisoned the republic; the crown hath engrossed the commons." Such a state of affairs was contrasted with the virtue demonstrated by the rebellious North American colonists, the modern exemplars of classical republicanism not only because of their rude health and vigor, but also because of the reliance of their economy upon slavery. But, while virtue was seen as a possible basis of a North American Republic, across Europe the leading theme of writers of every persuasion was the natural decline of virtue in republican states. This had been a commonplace in the ancient histories of Polybius, Plutarch, Livy, and Sallust, and was equally prominent in early modern neo-Machiavellian commentary. Virtue was unstable and inconstant, tending to fall prey to the wiles of fortune. Seminal virtue became weak and impaired by inevitable degeneracy through time or the depravity of what John Milton called in *Paradise Lost*, "these latter sinful Ages of the World." An intensely virtuous citizenry could always expect domestic turbulence and warmongering. In the eighteenth century the threat to virtue was greater still because of commerce, the ultimate source of corruption and effeminacy, and the antithesis of the stoic virtues which defined the republican character: steadfastness, loyalty, and self-sacrifice.

The greatest threat, however, to the survival of small republics derived from their military incapacity. By the eighteenth century smaller states could no longer defend themselves against larger states, and especially the commercial monarchies of Britain and France, the empires dominating international politics. In part this was due to the sheer size of monarchical armies and their technological sophistication, which was correlated with the commercial power of a state and its capacity to draw emergency resources from public credit. Although the developing British Empire was seen to be different from that of France, it was equally seen to be potentially devastating for small states. The accusation was leveled that, like all great states, Britain sought to become a universal monarchy. In a related manner it ought to aspire to commercial dominion. In a world in which the relative strength and influence of such states was becoming ever more extensive, no amount of virtue or national patriotism was going to be sufficient for small states to survive intact or untroubled. New ideas were called for to develop means of national survival in the altogether different international political context of the eighteenth century.

Survival by Union

Unions of small states were the long-established solution to problems of monarchical aggrandizement. Unions of states traditionally perceived themselves to be following ancient models, and particularly the union of the tribes that created Israel, and established civil government among the Ancient Hebrews. For much of European history, however, such unions were deemed impractical or failed to be longstanding, the classic case being the union of Calmar between Norway, Sweden, and Denmark (1397), ruptured by the Vasa creation of a separate Swedish monarchy in 1521. The difficulty by the eighteenth century was one of envisaging a union capable of standing against Britain and France, or of adapting itself to a world in which they were the dominant powers. One of the most significant peculiarities of the eighteenth century was the aversion to national unions – such as that of Poland and Saxony or Britain and Hanover – as opposed to large-state

aggrandizement. The Anglo-Scottish union was the great exception, although it was not seen to be a successful model to be imitated by other states until Jacobite challenges had been thwarted. Furthermore, it was seen by its opponents to be another case of a large state extending its empire at the expense of a small state. The message the Anglo-Scottish experiment sent to advocates of union was that uniting small states with larger states, rather than creating conglomerations of small states, might save the weaker sovereignties of Europe, but only at the risk of losing their identity and independence.

The convention by which small states united to form a larger association for defense explained the prosperity of the ancient Greek states, the successful barbarian resistance to Rome between the Rhine and the Danube, and the constancy of republican government in Germany, Holland, and Switzerland. Advocating confederations of states of the same kind, Montesquieu asserted that "the spirit of monarchy is war and enlargement of dominion: peace and moderation is the spirit of a republic." Despite such utterances, Montesquieu was ultimately skeptical about the prospects of a Europe-wide republic or a lasting confederacy of states with different forms of government on the model of the Holy Roman Empire or its microcosm, the Swiss Confederation. Establishment was one difficulty and the maintenance of any confederation another. Such a utopian vision was seen to be the basis of the project for perpetual peace of the Abbé de Saint-Pierre and others.

In the longer term, patriotic and nationalist movements seeking to create more powerful and prominent unitary states, based on shared geographical, cultural, or ethnic identities, sought to challenge the dominion of their larger neighbors, with singular consequences for modern European history. Making small states into large states or into larger unions of states has of course been the preferred solution to the problem of small state weakness in recent times, or at least from the North American founding. Eighteenth-century thought was different because schemes for European or Atlantic unions were then generally described as utopian rather than visionary. With the larger powers seeking to increase national resources in order to undertake global wars against competitor empires, and with a vast number of small and weak sovereignties across Europe, many of which were well known as commercial centers, the temptation for large states to use union as an excuse for the development of empire was too great. National unions involving larger states could rarely be trusted to benefit smaller partners.

The Small-State Crisis of the Eighteenth Century

All authors in the eighteenth century agreed that the small republics could no longer stand alone in war against commercial empires. They had to adapt themselves to the modern lust for trade and empire. In doing so they had to work out means of avoiding collapse in circumstances where Britain or France might seek to conquer them. Alternatively, the small states had to avoid competitors in the east of Europe doing likewise, in the manner of the dismemberment of Poland, which was partitioned between 1772 and 1795 by Prussia, Russia, and the Austrian monarchy. The examples of declining 'sister' states were ceaselessly commented upon as potential futures for the small states and small republics. Poland obsessed Europeans as the great republican monarchy in decline, having too independent, and above all too small, a standing army, monarchs "who had an interest to pursue very different from that of the Republic," and an uncertain division of powers between church and state. The situation of the small communities within the Russian empire was also instructive, because Russia contained so many different forms of state within an absolute monarchy that appeared to function as a federal empire. At the end of the century William Tooke was fascinated by what he called the 'mixed republican monarchy' of 'the Kalmuks and Kirghises'; but the most singular component of the Russian polity he held to be its Sparta – "the military democracy of the Kozaks [Cossacks], the essence and aim of which is war, and even of which we have been witnesses of a corruption, in its denying the other half of the human race all civil and domestic community." Tooke also considered the relationship between religious belief and politics to be significant in Russia, having seen 'the most monstrous polytheism' and 'total unacquaintance with any idea of a supreme intelligence' (*View of the Russian Empire*, 1799). Existence within an absolute monarchical confederation was another possible future for the small republics, but one that none of them recommended or welcomed. Closer to the center of Europe, France and Britain appeared to be perpetually acquiring new territories, with the republic of Genoa ceding Corsica to France in 1764 an apt example, and the British acquisition of Gibraltar in 1713 another.

For some writers there was a natural progression from savagery and barbarism to small dynasties and petty republics, and on to modern monarchy. This was what John Adams was interested in when he wrote that "the checks and balances of republican governments have been in some degree adopted by the courts of princes." Adams was struck by the fact that there were only three innovations in politics that could be said to distinguish ancient and modern times: representations of the people, the separation of the executive and legislative and judicial powers, and the further separation of the legislature itself into branches (*A Defence of the Constitutions of Government of the United States of America*, London, 1787: i–iii).

Adams could not see a great deal of difference between the operation of these principles under monarchies or republics. Such a point of view had led him to be critical of those in North America who drew a sharp distinction between British monarchical traditions and the political ideology of the new republic. When he became the second president of the United States, Adams held that such a fact ought to be in the foreground when adapting the North American federal constitution to a world dominated by European notions of national self-interest. Although Adams remained a republican with respect to North America, the argument that successful small states would do well to model themselves on Britain, or seek the support of empires like Britain in their battles for survival, became more prominent as the nineteenth century progressed.

Small States in the Modern World

After the Congress of Vienna in 1815 the British began to see themselves as the protectors of small states across Europe. As

national confidence increased, reaching a peak during the reign of Victoria, it was increasingly claimed that the small states of Europe could rely upon British arms if larger states attempted to violate their sovereignty. The irony was not lost on international observers that while the British posed as defenders of small states in Europe, they and their European rivals were involved in the extension of their own empires abroad, by colonization or economic dominion. The 'scramble for Africa' is well known, but it had counterparts in every part of the globe during an era of the creation of ever-larger states. If the small states were in decline across Europe in the eighteenth century, the process became global in the nineteenth century.

With regard to Europe, however, the system of alliances established as a result of the Napoleonic Wars required the large states of Europe to go to war if the small states were threatened: this had significant consequences. Sometimes the system of alliances proved sufficiently flexible to prevent war when small states were incorporated into larger empires, as occurred during the processes of German and Italian unification. But the need to defend the small states from attack was equally seen to be a major cause of both the first and second world wars. That they were defending the small and the weak was a claim that could be found on every side in 1914. In other words, as in previous centuries, small states have discovered manifold survival strategies compatible with the existence of global commerce and one or more superpowers in the modern world. At present, nationalist movements in Asia, Africa, and Europe are strong once again. One major difference with previous centuries is the renaissance of ideas about weak or strong federal and confederal unions as a means of preventing the interference of large states. When successful, the small states created have dealt with the some of the problems entailed by their size by adherence to federal unions for trade and the preservation of their sovereignty. At the same time, the existence of such federal unions is held to be destroying national sovereignty. Small states continue to be a political form with a singularly uncertain future in consequence.

See also: Conflict and War, Archaeology of: Behavior and Social Organization; Federalism; Nations and Nation-States in History.

Bibliography

Altorfer, S., 2007. State investment in eighteenth-century Berne. History of European Ideas 33 (4), 440–462.

Barston, R.P., 1973. The Other Powers: Studies in the Foreign Policies of Small States. Allen and Unwin, London.

Bembo, P., 2007. History of Venice (R.W. Ulery, Ed. and Trans.), vol. I. Harvard UP, Cambridge, MA.

Black, J., 1989. The problems of the small state: Bavaria and Britain in the second quarter of the eighteenth century. European History Quarterly 19, 5–36.

Black, J., 1994. European Warfare, 1660–1815. Yale UP, New Haven, CT.

Brady, T., 1995. Turning Swiss: Cities and Empire, 1450–1550. Cambridge UP, Cambridge.

Cooper, A.F., Shaw, T.M., 2009. The Diplomacies of Small States: Between Vulnerability and Resilience. Palgrave Macmillan, Basingstoke [England].

Fisher, H.A.L., 1914. The Value of Small States. Oxford University Press, London.

Fox, A.B., 1959. The Power of Small States: Diplomacy in World War II. University of Chicago Press, Chicago, IL, USA.

Hammond, H., 1951. City-State and World State in Greek and Roman Political Theory until Augustus. Harvard UP, Cambridge, MA.

Höll, O., 1983. Small States in Europe and Dependence. W. Braumüller, Wien.

Hont, I., 2006. The Jealousy of Trade: International Competition and the Nation State in Historical Perspective. Harvard University Press, Cambridge, MA.

Inbar, E., Sheffer, G., 1997. The National Security of Small States in a Changing World. Frank Cass, London.

Kapossy, B., 1998. Le prix de la liberté: idéologie républicaine et finances publiques à Berne au XVIIIe siècle. In: Flouck, F., Monbaron, P.-R., Stubenvoll, M., Tosato-Rigo, D. (Eds.), De l'Ours à la Cocarde. Régime bernois et révolution en pays de Vaud (1536–1798). Lausanne.

Karsh, E., 1988. Neutrality and Small States. Routledge, London.

Kisanga, E.J., Danchie, S.J., 2007. Commonwealth Secretariat, Commonwealth Parliamentary Association. Commonwealth Small States: Issues and Prospects. Commonwealth Secretariat, London.

Kitromilides, P.M., 2006. A Cyprus question before the Cyprus question. In: Kitromilides, Paschalis, Denores, Giorgio (Eds.), A Discourse on the Island of Cyprus and on the Reasons for the True Succession in That Kingdom Venice.

Panke, D., 2010. Small States in the European Union: Coping with Structural Disadvantages. Ashgate, Farnham, Surrey, England.

Pocock, J., 2003. Barbarism and Religion. In: The First Decline and Fall, vol. III. Cambridge UP, Cambridge.

Pocock, J., 1975. The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition. Princeton UP, Princeton, NJ.

Reiter, E., Gärtner, H., 2001. Small States and Alliances. Physica-Verlag, Heidelberg.

Šabič, Z., Bukowski, C.J., 2002. Small States in the Post-cold War World: Slovenia and NATO Enlargement. Praeger, Westport, CT.

Vattel, Emer de, 2008. The law of nations, or, principles of the law of nature, applied to the conduct and affairs of nations and sovereigns [1797]. In: Kapossy, B., Whatmore, R. (Eds.), Liberty Fund, Indianapolis.

Vital, D., 1971. The Survival of Small States: Studies in Small Power/Great Power Conflict. Oxford University Press, London.

Vlcek, W., 2008. Offshore Finance and Small States: Sovereignty, Size and Money. Palgrave Macmillan, Basingstoke [England].

Whatmore, R., 2012. Against War and Empire: Geneva, Britain and France in the Eighteenth Century. Yale University Press, New Haven, CT.

Smart Cities

Andrea Caragliu, Politecnico di Milano, Milan, Italy
Chiara D Bo, Università degli Studi di Milano, Milan, Italy
Karima Kourtit and Peter Nijkamp, VU University Amsterdam, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the vast literature on the concept of Smart City. The concept is first framed in the existing urban economics literature and the ongoing urbanization trends. The article, then, reviews the history of the Smart City concept from its inception, as mostly focusing on information and communication technology endowments as means to foster urban performance, toward the more recent approach based on the concentration of human capital in cities, to, finally, the necessary urban conditions leading to the fast adoption and wise exploitation of e-technologies. Next, the article summarizes the various methodologies and procedures followed in the literature to rank cities according to their degree of smartness. This vast literature is classified in terms of the characteristics pinpointed as major drivers of urban smartness in each work. Finally, the article concludes with a critical review of the Smart City policies drawn by various policy institutions, from the supranational (mainly, EU) to the local administrative level.

Smart Cities in an Open World

Our planet is increasingly turning into an urbanized planet. Cities – or in a more general sense, urban agglomerations – have become the dominant settlement patterns of mankind. Since the year 2007, more than 50% of the world population is living in cities (UNFPA, 2007), with a rapid urbanization rate in Latin America, Africa, and Asia. Consequently, cities tend to become magnets of political and socioeconomic activities, to the extent that modern cities – in particular, megacities – act increasingly as global powerhouses (Neal, 2012). Cities are no longer 'islands of splendid isolation', but global agents with a significant – local and *trans*border – impact. Therefore, in our 'urban century' there is a need to explore and map out the success conditions, strategic profiles, and global impacts of urban agglomerations.

It is noteworthy that cities are not only economic engines, but also part of a *trans*border spatial connectivity constellation. In fact, cities are often part of a functionally and spatially interconnected system of cities. They have to compete and to cooperate in globally interlinked markets to attract clients (business firms, visitors, Foreign Direct Investments (FDIs), knowledge migrants, creative class, etc.). Especially in a knowledge-based economy, skills and innovativeness are often critical success factors for high performing cities. Knowledge intensity, R&D, and innovative minds are a *sine qua non* for a successful modern city (Klaesson et al., 2013).

The increasing scale of cities and their rising economic and political importance also call for efficient governance structures. The traditional hierarchical top–down perspective is being increasingly replaced by participatory, interactive, and self-organizing modes of governance. This new political democratic constellation serves to exploit the development potential of urban agglomerations in a globally competitive and open world. As a consequence, there is a growing need and demand for 'Smart Cities', which are able to cope with both local and global challenges and which are able to employ the available resources (people, energy, land, talent, environment, geographic location) in a way that maximizes overall welfare and sustainable growth. Various important 'grand challenges' (Nijkamp and Kourtit, 2012) inherent in the 'New Urban World' (Kourtit et al., 2014) rise with a view to (stepwise) identification of various multifaceted building blocks that might form altogether a clear measurement and identification of high performing Smart Cities (Nijkamp and Kourtit, 2011). Smart concepts have increasingly become 'en vogue' in recent years. We talk about Smart Cities, smart transport, smart energy, smart grids, smart technology, smart specialization, and smart governance. In many cases, the concept of a Smart City appears to offer an umbrella for an integrated perspective on urban systems characterized by new forms of developing and managing urban technologies and infrastructures that provide efficient and effective solutions to a great many challenges of modern urban life (Kourtit et al., 2013). In particular, such smart solutions serve to enhance the quality of urban services to citizens and business actors, including, e.g., social equity, spatial accessibility, urban sustainability, and socioeconomic vitality.

A wealth of definitions of a 'Smart City' can be found in the literature – sometimes complementary, sometimes contrasting (Deakin, 2010; Torres et al., 2005). In various cases, it is claimed that cities may be smart in the way they exploit the potential of information and communication technologies (ICTs), e.g., for e-government services (Nijkamp and Cohen Blankshtain, 2013), as such services may become vehicles for innovation in cities. Clearly, the measurement of the extent to which ICT offers cities a competitive advantage is fraught with many empirical and methodological problems. In any case, there is a need for operational indicators, which are able to assess differences with respect to a baseline case. The problem of defining a baseline is that not only a clear and measurable urban framework case has to be identified and measured, but also the institutional setting that may be subject to change ('institutional endogeneity'). For instance, the modus operandi of a Triple Helix urban governance model may be influenced by the socioeconomic performance of a city in the previous year

(Etzkowitz and Zhou, 2006). The same holds for the well-known mode 1 and mode 2 research policy configurations based on Gibbons et al. (1994). Kourtit et al. (2013) note the "*tendency to assess Smart City performance not only from an economic perspective but also from an evolutionary, partly noneconomic valuation perspective, in order to suggest solutions for cities to have a competitive edge in a global network economy*" (p. 14). A related concept of Smart Cities can be found in a so-called flagship project 'Smart City' proposed by TNO (2012). This initiative aims to position cities at a competitive global edge in terms of residential and business attractiveness. In this initiative, cities are spatially integrated agents of individual and collective interest, in which information collection, use, and sharing offers an unprecedented value added in terms of quality of life, ecological sustainability, and economic vitality. Smart Cities arise because of intelligent use of digital information, for instance, in domains like human health, mobility, energy use, education, knowledge transfer, and urban governance. The basis of this idea originated from Castells' (1998) contribution on the distinction between 'Space of Places' and 'Space of Flows', next to a 'Space of Information'. All these spaces are integrated in a Smart City. Thus, interfaces of spaces and data are crucial for a Smart City in the TNO conceptualization.

The present article serves to present a concise state-of-the-art overview of the concept of 'Smart Cities', with the aim to provide informed insights into the strength–weakness conditions for 'Smart City' initiatives. The article is organized as follows. First, in Section Definitions and Conceptualization of Smart Cities an overview of the ways the notion of Smart City has been conceptualized and defined will be offered. Next, Section Ranking Procedures for Smart Cities will focus on the different methods used for ranking Smart Cities. Section Policies on Smart Cities will explore the policy implications stemming from the concept of Smart City.

Definitions and Conceptualization of Smart Cities

The profile of a Smart City has, in recent years, become an interesting research topic in urban research and policy (Winters, 2011; Caragliu et al., 2011). It concerns various disciplines, from economics to political science, from smart ecology to smart urban architecture and planning, from smart technological (mainly ICT) development to globalization forces (including international trade). Cities are multifaceted organisms, acting in complex environments, shaped by other urban areas – both neighboring and far away– and by general national and global drivers (Kourtit et al., 2014). This calls for new concepts, which can be difficult to define (Nam and Pardo, 2011), among which the notion of Smart City has recently gained much attention among policy makers, private firms, and researchers. Because of the difficulties in properly defining new urban concepts, it becomes fundamental to narrow the research scope on the concept of Smart City.

Several definitions have emerged over time, each differing from the other in terms of the main 'smart' characteristic deemed as the most relevant. Initial conceptualizations revolved around ICTs as the main pillar around which a city should build its smart pathway. Earlier definitions related to the concept thus include the 'wired city' (Dutton et al., 1987), whereby the focus is on networking the urban space per se and the 'intelligent city' (Komninos, 2002, 2009), which expands this idea by considering also the cognitive element of a digital city and the relationship between individual cognitive skills and the urban information system.

Linking smartness to the availability, development and use of ICT is based on the notion that in order to succeed, cities and urban dwellers must be interconnected. The focus on information technologies and data availability is also related to a policy shift from mainly top–down to increasingly more bottom–up approaches, paving the way for greater importance of citizens' direct participation in the urban governance and their links with city officials. This latter aspect is related to the importance of e-government in the creation of a Smart City (Deakin et al., 2011). This in turn suggests an important role for the human capital component of cities and might also give rise to concerns related to the availability, management, and privacy issues related to 'big data' (Batty, 2012).

This vision of ICT-focused smartness eventually spurred several academic projects and research, not without critiques to the business orientation of this concept. In fact, ICTs have been, from the very inception of this concept, considered at the roots of urban smartness, mostly because of the vast empirical evidence supporting ICT led development (Caragliu, 2013); this attracted the interest of large corporations, aiming at obtaining profits from the interest of urban administrations in such technologies (Vanolo, 2013). This type of critique argues that the Smart City concept would be void of objective content, and would concentrate the attention of citizens and policy makers toward the adoption of ICT technologies, thus ignoring the need of nondigital citizens, or, more importantly, the equal rights of cities that voluntarily choose not to go digital (Hollands, 2008).

Yet, the Smart City concept is not a static idea, and has been recently evolving. Such evolution, which partially tackles these critiques, took place especially in the EU, in relation to the overall sustainability goals of the EU2020 framework, placing more focus on environmental sustainability and clean energy issues. In terms of the conceptualization of Smart City, the focus on environmental issues is incorporated in the concept of 'resilient city' (Newman et al., 2009). This further characterization has also given rise to several studies exploring the need, development, and impact of smart solutions in energy creation and consumption, such as smart meters and smart grids (Clastres, 2011).

The focus on ICT and sustainability in the policy and academic domain are mirrored in the business and consultancy sector, where the interest for the Smart City concept is related to the development of software solutions and applications and management models related to both ICTs and energy. From a purely academic point of view, Smart Cities are often identified in relation to their human capital component and potential (Berry and Glaeser, 2005; Shapiro, 2006; Winters, 2011). The interest on the subject, in the policy, business, and academic worlds, briefly documented above, has given rise to several attempts of providing

a definition of the concept and develop rankings of Smart Cities around the world.

Despite the relevant academic success and the significant investments in Smart City projects at all administrative levels, however, only few works provide a clear-cut definition of what a Smart City really is, in particular extending their characterization beyond the pure technological (viz., ICT-related) content, thus including contextual elements. Because context elements should be expected to explain the success, or failure, of Smart City policies, this article only summarizes four notable definitions (in chronological order) which go truly beyond ICTs as a means to define urban smartness, highlighting the most important elements.

A first comprehensive operational definition of urban smartness is implicitly to be found in Giffinger et al. (2007). Giffinger et al. (2007) provide in fact a classification of European medium-sized cities according to six axes (Smart people, Smart governance, Smart environment, Smart economy, Smart mobility, and Smart living). Their definition reads as follows: "*A Smart City is a city well performing in a forward-looking way in these six characteristics, built on the 'smart' combination of endowments and activities of self-decisive, independent, and aware citizens*" (Giffinger et al., 2007: p. 13).

Caragliu et al. (2011) build on the classification by Giffinger and coauthors, and provide a comprehensive and operational definition of urban smartness. In this case, cities are identified as smart when "*investments in human and social capital and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance*". This definition presents two main advantages. First, it is inspired by an urban production function approach, whereby urban smartness is defined as a precondition to urban economic performance. A clear distinction between inputs and outputs, previously not encompassed in analogous attempts, suggests that smartness can be seen as an intermediate step toward the goal of smart (viz., sustainable) urban growth. Second, the definition decomposes the concept along six dimensions, which can be individually measured and tested, using data from official statistical sources. Therefore, this definition has been among the first ones to be empirically verified (Caragliu and Del Bo, 2012).

More recently, the interplay between technology and its users has been synthesized in Ratti and Townsend (2011), which simplifies the previous approach by stating that "*Truly smart cities will emerge as inhabitants and their many electronic devices are recruited as real-time sensors of daily life*."

Finally, De Souza (2012) concludes that "*A smart city is livable, resilient, sustainable, and designed through open and collaborative governance*." This last definition builds a bridge between previous works focusing on ICTs and context conditions as the main ingredients of urban smartness, and paves the way for future evolutions toward the resilient city.

Clearly, operational definitions must be clarified before measuring and/or ranking cities in terms of smartness, as well as before committing financial resources to policies aiming at stimulating the success of such cities, as they can be a starting point for empirical assessments of the degree of 'smartness' of cities and for sound and valuable policy evaluation exercises.

Ranking Procedures for Smart Cities

The discussion on Smart Cities seems to concentrate nowadays on two major issues:

1. Is a given city a Smart City? This is an empirical research question that can only be answered if clear threshold conditions for the smartness of a city are specified. In reality, we often face no binary conditions, but grades of smartness (e.g., on a scale from 1 to 10).
2. Which city is the smartest city? This question calls for clear benchmark principles, so that a ranking – or even rating – of the various alternative cities under consideration can be obtained. This research question runs parallel to the related question of which city is the most sustainable or most ecofriendly.

The assessment of the degree of smartness of cities leading to a ranking of Smart Cities calls for an operational methodology. The methodological toolbox for a quantitative underpinning of such rankings can be manifold. Examples include spider diagrams to map out the differences or indicators among the set of Smart Cities considered; multivariate analysis to reduce the multidimensional nature of the underlying database, in order to be able to select the key indicators for Smart City achievements; unweighted summation techniques to arrive at an unambiguous ranking order of Smart Cities; multicriteria models to arrive at an unambiguous ranking or rating of the various cities concerned; and, finally, data envelopment analysis to trace the relative ranking of Smart Cites in terms of their relative performance.

Clearly, there may also be a ranking of appropriate benchmark techniques that can be employed to identify the relative profile and performance of a set of Smart Cities. Such a comparative analysis may also be instrumental for Smart City policies. This will be further discussed in Section Policies on Smart Cities.

The study undertaken by the Center of Regional Science, Vienna (Giffinger et al., 2007) aims to rank medium-sized European cities according to their degree of 'smartness'. A Smart City refers here to cities with ICT-oriented industries, including advanced business parks, higher education facilities, and e-government amenities. These aspects are summarized in six characteristic building blocks, as described in the previous section: Smart economy, Smart people, Smart governance, Smart mobility, Smart environment, and Smart living. Each of these six building blocks is decomposed into subcharacteristics. For example, Smart economy (or competitiveness) is subdivided into seven indicators: innovative spirit, entrepreneurship, economic image and trademarks, productivity, flexibility of labor market, international embeddedness, and ability to transform (Rudolf et al., 2010). Through a standardized scoring approach the various European cities in the sample – 70 in total – can then be ranked according to their degree of smartness, with Luxembourg at the top and Ruse on the other end of the spectrum. More fundamental arguments on this type of city ranking methods can be found in Begg (1999).

A more general review of the notion of a Smart City and its positioning in the urban planning literature can be found in Caragliu et al. (2011) (Section Definitions and

Conceptualization of Smart Cities above). The fundamental components of this definition of urban smartness are the widespread use of a networked infrastructure, the city's emphasis on a business-led urban development, its socially including urban residents in the provision of public services, the central role given to high-tech and creative industries, the important role played by social and relational capital, and, finally, social and environmental sustainability as a strategic core orientation for the city.

Kourtit et al. (2013) provide a followup on this ranking, especially from a dynamic competence point of view. They suggest that the environmental and climate neutral capabilities and resources, along with advanced infrastructure, smart logistic systems, smart energy use, accessibility and connectivity, and other spatial facilities, are decisive for the relatively strong competitive position of Smart Cities in the complex force field of urban dynamics. However, much room is left for further empirical verifications and rankings of Smart Cities, in particular linking theory to practice and policy making, with the lenses of the definitions and conceptualizations discussed in Section Definitions and Conceptualization of Smart Cities.

Policies on Smart Cities

Born in the policy context, the notion of Smart City has attracted vast attention both from policy makers and private companies. While the focus of policy makers, due to the original nature of the Smart City concept as ICT-driven, is still on communication technologies, nowadays attention is increasingly being paid to the human and social capital concepts and to environmental sustainability as means of complementing ICT infrastructure. In fact, the only addition of ICT infrastructure to a city cannot per se suffice to determine its success: the broader context conditions need to be conducive to the full adoption, understanding, and wise exploitation of these resources.

Because the concept of Smart City has been mostly studied in a European context, European Union institutions have been focusing most intensely on Smart Cities strategies. Several EU- and local authorities-funded research projects have provided the rationale for the need for policies stimulating urban smartness. Funding bodies and EU institutions include Urbact, Interreg, CORDIS, the European Investment Bank, and the European Commission (the Web sites of the listed projects are added in the Reference list). Other international institutions have also provided evidence about the need for an ICT-driven form of development; these include, among many examples, the Intelligent Community Forum, the Organisation for Economic Co-operation and Development (OECD) and EUROSTAT, whose Oslo Manual (EUROSTAT, 2005) discusses the importance of innovation in ICT industries as a growth driver.

Practically, the largest commitment of public money toward Smart City policies has been made by the European Commission with the Smart Cities and Communities European Innovation Partnership (EC, 2012), which implied so far a total investment of 450 million euros (Falk, 2012). At lower administrative levels, numerous cities, mostly located in the EU28, embarked on Smart Cities projects (to name a few, Amsterdam, Aarhus, Genoa, Lyon, Málaga, Santander, Southampton, and Turin). Notable nonEU examples include Dubai and Singapore.

Along these projects, a relevant role is also played by large ICT companies, with a clear target to diffuse ICT technologies, sensors, advanced heating systems, traffic control devices, etc., to as many urban areas as possible. Multinational Companies with relevant Smart City projects include IBM, Cisco, Hitachi, and Siemens.

Thus, it can be argued that Smart City projects can be classified on the basis of the promoting industry and of the level of aggregation. The most relevant stakeholder is probably the public sector. At the supranational level, EU institutions have committed relevant funds for applied research for the identification of the potential impacts of Smart City policies. This goes under the umbrella of the Smart Cities and Communities European Innovation Partnership.

At the local level, local (city) authorities focused on funding local initiatives targeting ICT infrastructure and the improvement of the context conditions leading to the skilful adoption of such technologies. Such local policies include, among countless recent examples, *Amsterdam Smart City*, Southampton's *Smartcities Card*, and *Smart Santander*. In these cities, funds have been invested, and cooperation efforts put in place, with the aim to improve modern ICT infrastructure and enhance its skilful application by teaching people how to make the best use of it, and show citizens and entrepreneurs what the future of their cities will look like.

Clearly, such government-led efforts could not be carried out without resorting to large corporations willing to invest in Smart City technologies. Multinational companies such as IBM, Cisco, Siemens, and Hitachi have initiated dedicated lines of business for the development of technologies aiming at reducing environmental pollution, fostering a more efficient energy and waste management, and, in general, making the organization of cities easier and more effective.

The public and private sectors do not necessarily interact in a smooth way. In fact, the interplay between the public sector's need for advanced technologies, looking for a better and wiser management of complex urban systems (Caragliu et al., 2011), and the private sector's push for the diffusion and penetration of smart technologies, has also drawn criticism from the academia (Vanolo, 2013; Hollands, 2008).

So far, theoretical definitions of urban smartness have been only partially translated into practical policy making: academic definitions of the concept of Smart City tend, in fact, to be heavily cited within the academic world, and, by the same token, policies tend to avoid making reference to academic definitions. In other words, definitions of urban smartness seem so far to have been only partially shared by academics and practitioners. Besides, the private sector is characterized by socially inefficient incentives toward the diffusion of smart ICT technologies for urban mobility, and traffic and waste management. Externalities emerge in fact in the use of publicly funded ICT infrastructure, whose cost is born only partially by the individual taxpayer, and whose use makes individual citizens and firms more productive. Thus, the present state-of-the-art of the literature on Smart Cities suggests a clear need for bridging theory and practice in shaping sound Smart City policies. Theories and policies related to Smart Cities are still

in their infancy, and call for further research in order to strengthen the concept's theoretical architecture and to shape sound policy making.

See also: Cities: Internal Structure; Development Geographies; Knowledge, Geography of; Local Economic Development; Resilient City; Science and Technology, Social Study of: Computers and Information Technology; Spatial Thinking in the Social Sciences; Urban Geography; Urban Policy in Europe; Urbanization.

Bibliography

Batty, M., 2012. Smart cities, big data. Environment and Planning B: Planning and Design 39 (2), 191–193.

Begg, I., 1999. Cities and Competitiveness. Urban Studies 36 (5–6), 795–810.

Berry, C.R., Glaeser, E.L., 2005. The divergence of human capital levels across cities. Papers in Regional Science 84 (3), 407–444.

Caragliu, A., 2013. Dynamics of knowledge diffusion: the ICT sector in Lombardy. Regional Science Policy and Practice 5 (4), 453–473.

Caragliu, A., Del Bo, C., 2012. Smartness and European urban performance: assessing the local impacts of smart urban attributes. Innovation: The European Journal of Social Science Research 25 (2), 97–113.

Caragliu, A., Del Bo, C., Nijkamp, P., 2011. Smart cities in Europe. Journal of Urban Technology 18 (2), 65–82.

Castells, M., 1998. End of Millennium: The Information Age: Economy, Society and Culture. Wiley-Blackwell, New York.

Centre of Regional Science, 2007. Smart Cities: Ranking of European Medium-sized Cities. Technical University, Vienna.

Clastres, C., 2011. Smart grids: another step towards competition, energy security and climate change objectives. Energy Policy 39 (9), 5399–5408.

De Souza, K.C., 2012. Designing and planning for smart(er) cities. Practicing Planner (Online Magazine of AICP) 10 (4). Winter 2012.

Deakin, M., Lombardi, P., Cooper, I., 2011. The intelcities community of practice: the capacity-building, co-design, evaluation and monitoring of e-gov services. Journal of Urban Technology 18 (2), 13–38.

Deakin, M., 2010. SCRAN: the smart cities (inter) regional academic network supporting the develop-ment of a trans-national comparator for the standardisation of e-government services. In: Reddick, C. (Ed.), Comparative e-government: An Examination of e-government across Countries. Springer Press, Berlin.

Dutton, W.H., Blumler, J.G., Kraemer, K.L., 1987. Wired Cities: Shaping the Future of Communications. Annenberg School of Communications, Washington, DC.

Etzkowitz, H., Zhou, C., 2006. Triple Helix twins: innovation and sustainability. Science and Public Policy 33 (1), 77–83.

European Commission, 2012. Communication from the Commission: Smart Cities and Communities -European Innovation Partnership. Brussels, 10.7.2012, C(2012) 4701 final.

EUROSTAT, 2005. The Oslo Manual. http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/OSLO/EN/OSLO-EN.PDF (accessed 24.07.13.).

Falk, T., 2012. E.U. Invests $450 Million in Smart Cities. http://www.smartplanet.com/blog/cities/eu-invests-450-million-in-smart-cities/3914 (accessed 26.07.13.).

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. The New Production of Knowledge – the Dynamics of Science and Research in Contemporary Societies. Sage, London.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanović, N., Meijers, E., 2007. Smart Cities. Ranking of European Medium-sized Cities. Centre of Regional Science of Vienna. www.smart-cities.eu.

Hollands, R.G., 2008. Will the real smart city please stand up? City 12 (3), 303–320.

Klaesson, J., Johansson, B., Karlsson, C. (Eds.), 2013. Metropolitan Regions. Springer, Berlin.

Komninos, N., 2002. Intelligent cities: innovation, knowledge systems and digital spaces. Routledge, London.

Komninos, N., 2009. Intelligent cities: towards interactive and global innovation environments. International Journal of Innovation and Regional Development 1 (4), 337–355.

Kourtit, K., Nijkamp, P., Partridge, M., 2014. Challenges of the New Urban World. Applied Spatial Analysis and Policy, Forthcoming.

Kourtit, K., Deakin, M., Caragliu, A., Del Bo, C., Nijkamp, P., Giordano, S., Lombardi, P., 2013. An advanced triple -helix network framework for smart cities performance. In: Deakin, M. (Ed.), Smart Cities: Governing, Modelling and Analysing the Transition. Routledge, London.

Nam, T., Pardo, T.A., 2011. Conceptualizing smart city with dimensions of technology, people and institutions. In: Proceedings of the 12th Annual International Conference on Digital Government Research. College Park, MD, pp. 282–291.

Neal, Z., 2012. The Connected City. Routledge, New York.

Newman, P., Beatley, T., Boyer, H., 2009. Resilient Cities: Responding to Peak Oil and Climate Change. Island Press, Washington, DC.

Nijkamp, P., Kourtit, K., 2011. Urban Europe Scoping Document, Joint Programming Initiative. Urban Europe, VU University, Amsterdam.

Nijkamp, P., Kourtit, K., 2012. The 'new urban europe': global challenges and local responses in the urban century. European Planning Studies 21 (3), 291–315.

Nijkamp, P., Cohen-Blankshtain, G., 2013. The importance of ICT for cities: e-governance and cyber perceptions. In: Klaasson, J., Johansson, B., Karlsson, C. (Eds.), Metropolitan Regions. Springer, Berlin, pp. 295–308.

Ratti, C., Townsend, A., Tuesday, August 16, 2011. Harnessing residents' electronic devices will yield truly smart cities". Scientific American. http://www.scientificamerican.com/article.cfm?id=the-social-nexus&print=true (accessed 26.07.13.).

Rudolf, G., Gudrun, H., Hans, K., 2010. The role of rankings in growing city competition, Urban Research & Practice 3 (3), 299–312.

Shapiro, J.M., 2006. Smart cities: quality of life, productivity, and the growth effects of human capital. The Review of Economics and Statistics 88 (2), 324–335.

TNO, 2012. Flagship Project Smart City. TNO, Delft. http://www.tno.nl/content.cfm?context=uitgelicht&content=uitgelicht_content&laag1=1275&item_id=1275&Taal=2.

Torres, L., Vicente, P., Basilio, A., 2005. e-Government developments on delivering public services among EU cities. Government Information Quarterly 22, 217–238.

United Nations Population Fund, 2007. UNFPA State of World Population. http://www.unfpa.org/swp/2007/english/introduction.html (accessed 07.04.11.).

Vanolo, A., 2013. Smartmentality: the smart city as disciplinary strategy. Urban Studies. http://dx.doi.org/10.1177/0042098013494427.

Winters, J.V., 2011. Why are smart cities growing? Who moves and who stays. Journal of Regional Science 51 (2), 253–270.

Relevant Websites

http://urbact.eu/en/projects/innovation-creativity/smart-cities/homepage/ – URBACT.

http://www.northsearegion.eu/ivb/projects/details/&tid=84 – The North Sea Region Programme.

http://ec.europa.eu/energy/technology/initiatives/smart_cities_en.htm – European Commission.

http://ec.europa.eu/regional_policy/thefunds/instruments/doc/jessica/jessica_horizontal_study_smart_and_sustainable_cities_en.pdf – JESSICA for Smart and Sustainable Cities.

http://www.epic-cities.eu/content/smart-cities – EU platform with links to completed and ongoing research projects.

Smith, Adam (1723–90)

Dominique Bouchet, University of Southern Denmark, Odense, Denmark; and School of Creative Arts, James Cook University, Townsville, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.S. Skinner, volume 21, pp. 14189–14195, © 2001, Elsevier Ltd.

Abstract

Adam Smith is often said to have been the founder of the science of economics and the father of liberalism in the sphere of economics. In fact he was neither. He lived at a turning point in western economic and political history, one that was littered with disruptive developments. He came up with a masterful synthesis of the economic knowledge of his period and emphasized both the relative autonomy of these phenomena and their importance in terms of generating wealth from, and in the interests of, everyone. Nevertheless, Smith never denied the moral foundation of economic behavior.

> Political economy, considered as a branch of the science of a statesman or legislator, proposes two distinct objects: first to provide a plentiful revenue of subsistence for the people, or more properly to enable them to provide such a revenue or subsistence for themselves; and secondly, to supply the state or commonwealth with a revenue sufficient for the public services.
>
> Adam Smith, *The Wealth of Nations* Book IV, Introduction (Smith, 1999(1776): Intro. p. 5)

A Scottish Thinker

Adam Smith was born in January 1723 in the small town of Kirkcaldy on the coast of Scotland a few miles north of Edinburgh. He matriculated at Glasgow University in November 1737. Glasgow at the time was a commercial harbor in which tobacco and wool were traded, and ships built. Edinburgh was an important center of intellectual activities. As a 17-year old he went to Oxford University to study philosophy and literature for 6 years. In 1751 he became professor of literature at Glasgow University where from 1753 he took the chair in moral philosophy. He achieved a great scholarly repute giving lectures on ethics, rhetoric, literary style, justice, security and 'cheapness or plenty.' Already then, he was addressing the issue of political economy in his lectures. In connection with his lectures on ethics he wrote his first seminal work: *The Theory of Moral Sentiments*. A work which, besides its great erudition and originality, was written in a clear language, which Edmund Burke was to describe as 'rather painting than writing.' Its publication in 1759 led to Smith's appointment as tutor to the third Duke of Buccleuch. In this capacity he spent nearly 3 years in France from 1764 to 1767 conversing with many leading French intellectuals, among whom were D'Alembert, D'Holbach, Dupont, Helvetius, Morellet, Necker, Quesnay, Turgot, and Voltaire. Back to Scotland in 1767, Smith dedicated most of his time writing the work for which he became well known worldwide: *The Wealth of Nations*. It was published in 1776, soon translated in many languages and had an impact on the politics of Nations. For the rest of his life, Smith was mainly busy with new editions of his work and high charges that he held. He died in 1790 in Kirkcaldy at the age of 67.

Founder of Economics

Adam Smith was one of the earliest thinkers to suggest that laws on a par with those reigning in the world of physics could be applied to economic systems. Indeed he is often said to have been the founder of the science of economics and the father of liberalism in the sphere of economics. In fact he was neither. He lived at a turning point in western economic and political history, one that was littered with disruptive developments. He came up with a masterful synthesis of the economic knowledge of his period and emphasized both the relative autonomy of these phenomena and their importance in terms of generating wealth from and in the interests of everyone. Nevertheless, Smith never denied the moral foundation of economic behavior. The fact that his work inspired after his death a massive body of economic theory, more abstract than ever existed before, should not conceal the fact that Adam Smith himself was more a philosopher than an economist as his inquiries into economics were much more motivated by an understanding of human nature than by an intention to develop an autonomous science. Far from developing a reductionist view of social relations, Adam Smith provided an elaborated critique of a short-sighted political approach to economic life.

While the economy was strongly influenced by market exchange in Adam Smith's time, that part of economic life, which was under the political sway of aristocrats, was still rooted in appropriations, donations, and privileges. The fact that political power was in the hands of the aristocrats, whereas wealth stemmed largely from a market economy, was a major source of tension at the time. In England, which was then the world's dominant power, technical inventions flourished, the population was growing, agriculture had been restructured, social classes were evolving, and there were lively debates on political and ethical matters. European governments were interested mainly in power, war, money, and agricultural production since these could bolster their position. Because of his interest in exchange and in production and consumption of whatever type, Adam Smith brought in a more rational way of contributing to the wealth of nations and a more modern understanding of what society is and how it works. His declared ambition was to enable people to provide themselves

with adequate resources and in so doing to generate a surplus sufficient to finance public services.

In order to do this, Adam Smith analyzed the mechanisms of national growth on the basis of fundamental factors and concepts. Some of these ideas came from his own reading of earlier scholars. For instance, his distinction between productive and unproductive work comes from the physiocrats; the idea that capital is part of a stock that can be allocated to productive work presupposes the idea of stock as put forward by the mercantilists. At any rate, Adam Smith stated that there are three factors of production: land, labor, and capital. He defined wealth to be purchasing power and believed there to be a link between the wealth of each individual and the wealth of everyone in society. He distinguished exchange value from use value, i.e., the usefulness of an object. He recognized that the growth of capital enables division of labor to be augmented and its productivity to be increased. He brought to the fore that which contributes to an increase in wages and demand but also contributes to a decline in profits. He criticized outdated ideas on wealth and political economy. Thus, merely accumulating precious metals and colonies does not, *ipso facto*, increase a nation's wealth. Labor is the real source of wealth. He believed that it is the market which should be the center of attention rather than money or land. Restricting flows of goods and particularly import quotas gives rise to an unsatisfactory allocation of capital and inadequate division of labor. State aid to exporters can have negative effects. Smith was also interested in return on investments made by the state.

The fame of Adam Smith's work spread rapidly. This was largely due to the winds of change mentioned above, but also because of the American war of independence (1775–83), which persuaded England, as well as many other countries, to change their views with respect to international relations and wealth. Adam Smith put forward solutions to the problems of his time. And, since his ideas influenced the political economy of the world's dominant power they quickly became widespread. Smith founded the so-called classical school of economics on his critique of the mercantilist ideas of the previous century and it was this which, in the next century, from David Ricardo to Jean-Baptiste Say, had a big influence on Karl Marx. Marx was later to place great emphasis on the self-regulating nature of the new, capitalist economy, which, in Smith's time, was still embryonic and utopian. Furthermore, in addition to highlighting the fact that workers were not fully paid for their labor, along with describing (in the Book V of *The Wealth of Nations*) the negative impact of the division of labor on workers' mental health, Smith was also the first to put forward an economic definition of social classes and to define profit in relationship to the input of capital. It is worth pointing out, nevertheless, that, as far as he was concerned, the key conflict was not between workers and capitalists, but rather between those who were in favor of competition and those who wanted government protectionism and the establishment of monopolies. It should also be borne in mind that Smith basically ignored the question of money in his search for fundamental economic laws; it was only after the 1929 Great Depression and John Maynard Keynes' work that political economics took a serious interest in monetary aspects once again. Nevertheless, Smith's dynamic systems approach has continued to influence generations of economists and sociologists. This is still very much the case, 200 years after the publication of *The Wealth of Nations*; an example being the French *Regulation School* (Michel Aglietta, Robert Boyer).

Ideological Twists

Just like those of Karl Marx (1818–83) a century later, Adam Smith's ideas have been hijacked for ideological purposes and reduced to the role of rabble-rousing slogans that are a long way from the powerful insights imbued in the original ideas. The facts that the original analyses were rooted in a given historical context with a well thought-through conceptual system are ignored. By taking things out of context not only do the words and quotations lose their meaning, but also they become useless in terms of understanding evolving social phenomena. And yet Smith has provided us with a remarkable synthesis of the economic and political ideas of his time and proposed a radically different roadmap for the future development of our society. Contrary to the currently prevalent ideological caricature, Adam Smith was not an advocate of laissez-faire capitalism who believed that the wealth stemming from the growth of capital would eventually trickle down to those who it exploited. He was a theoretician of the relationships between markets and society who shed light on how an understanding of economic phenomena could contribute to a fresh understanding of government, society and the relationship between them. He drew attention to the importance of productive work, everyday consumption, and national trade. Smith, a moral philosopher facing the intellectual challenges of his time, helped not only to found economics as the science of wealth, but also sociology as the science of social bonds. In doing this, Adam Smith actually helped to bring into existence both the market as it is today and the state. It is unhelpful, though, to reduce his contribution to being merely a champion of laissez-faire.

An Industrial Revolution in Embryo

At the time when Adam Smith was reflecting on the evolution of society, the industrial revolution was in its infancy. In 1769 James Watt patented his steam engine, two years later Richard Arkwight set up the first cotton-spinning mill, and the *spinning jenny* and the *mule jenny* appeared in 1770 and 1779, all increasing productivity while requiring adaptation to new and restrictive collective work systems. Despite these changes, the industrial revolution had still not really taken off by Smith's death in 1790. There were still far more household servants than factory workers and craftsmen, and most work was still carried out at home and without the help of machines. Nevertheless those changes were significant in some parts of the population. They altered human relations, calorific intake, time disciplines, patterns of sociability, and family possessions; they resulted in gains in productivity and social satisfaction (Bayly, 2005: p. 49). Today's economic historians call our attention to the fact that an 'industrious revolution' had begun in the late seventeenth century of a process of household-based resource

reallocation that increased both the supply of marketed commodities and labor, and the demand for market-supplied goods (Vries, 1994). Although Smith himself did not foresee the industrial revolution at all, he did notice this 'industrious' revolution. Smith drew attention to the importance of the national product and the internal market; he saw that if one took the trouble to enhance these, they could become sources of wealth for the whole of society.

A New Conception of Society

In Adam Smith's epoch systems of production, trade and the priorities of rulers were quite unlike those we see today. Notwithstanding this, people had already begun to reinterpret the world in a quite different way from that of a few centuries earlier. A slow disintegration of the traditional social order and its *weltanschauung* had already begun in the thirteenth century, while, from the sixteenth century onwards, the religious foundations, in which political ideas were rooted, had begun to creak ever more loudly. This disintegration accelerated in the seventeenth century, when the possibility of a lay society really came to the fore. As Max Weber (1864–1920) was to put it later, there was a 'disenchantment'; society came to rely only on itself; this was a new way of understanding politics, and it now included economics.

It was not only a 'disenchantment' of society, it was also a new conception of history and of the competition between the concern for salvation and the need for knowledge. Humankind was no longer only to rely on divine grace; a better life was believed achievable by the exertion of its own powers. The scientific revolution provided mankind with a new mastery over the conditions of its existence and suggested that the production of knowledge is cumulative and irreversible; thus the idea of progress was emerging. Also emerging, in Smith's time, was a new understanding of the relations between the elements and the whole of a system: the self-correcting procedure of scientific discovery allowed the scientific enterprise as a whole to flourish in spite of the constant revision and even rejection of particular findings.

Back in 1615 Antoyne de Montchrétien in his *Traité de l'économie politique* (Montchrétien, 1889(1615)) had argued that what matters the most for a community's fate is neither its religious devotion guided by its clergymen nor the courage and military skills of its noblemen, but the occupation of the common people. What people produce and consume, how they organize their work, how clever the artisans and the merchants are – this is what affects the most the general good. Montchrestien was not the first to make that point. Barthélemy de Laffemas who was the commercial advisor of the French king Henri IV has already pointed that out in 1606. Precisely this idea will be at the core of Adam Smith's argumentation 170 years later.

What Smith set down was in fact the conclusion to a lengthy period of philosophical debate concerning the institutions and the regulation of society, once freed from references to the holy writ. A fresh reality had appeared, one requiring a new language, a new way of speaking and seeing, and different behaviors. The main problem of political philosophy at this time was how to come to terms with an autonomous society with no reference to an overarching external authority. What is it that binds a society together? How long has it been so bound? How can we make sure that it will continue as a coherent whole? By the eighteenth century the main issue was to understand how society was *regulated* and to ensure that it had a simple and solid basis.

A New Relation between Morals and Politics

The problem of morals had been handed down from a religious *weltanschauung*, but, from Machiavelli (*The Prince 1513*) onwards, it was looked at in a nonreligious context. In this new, lay context it was no longer a question of teaching mankind what *must* be done with reference to the next world, but rather to understand what man actually is and what can be done in this world with men as they actually are. From the seventeenth century onwards the idea spread that the social contract and the workings of society should be studied *on the basis of natural human emotions and not despite them*. In order to endow society with a solid and universally acceptable foundation, it was therefore necessary to develop a science of *human emotions* on the lines of the *science of reason*, as exhibited by mathematics. This implied casting doubt on all forms of power that are not based on reason.

Adam Smith is part of the philosophical tradition within which morality and politics are inextricably intermingled. Politics was increasingly perceived to be the art of managing human emotions. When he was 30 years of age, Smith became chairman of the moral philosophy department at Glasgow University. Six years later, in 1759 he published his *Theory of Moral Sentiments*. It was his interest in what can drive moral behavior that led Smith to lay greater stress on the economy than on politics. In *An Inquiry into the Nature and Causes of the Wealth of Nations*, published in 1776, he portrayed society as no longer being basically a political system, but rather an economic system within which markets can play a beneficial role. Though the markets of which he speaks did not actually yet exist he sensed that the world was changing. Bolstered by the advancement of scientific explanation during the century of Galileo, it was henceforth possible to envisage a change from mere submission to the dominant authorities to a scientific approach to the mechanisms of society. Market forces, essential to society, could actually be studied and encouraged in order to enable wealth within a nation to grow. That they could be studied and managed indicates that we are a long way from the laissez-faire ideology to which his theory of markets has been reduced. The idea of free markets carried with it the desire for self-regulation of civil society and thus a certain ideal of autonomy. Actually Smith's work contributed to and participated in the elaboration of a market-driven utopia as a principle of social organization in a period of widespread disruption; although the power of the church was on the wane, it was still fairly unthinkable to question the growing authority of sovereign rulers. In the reflections stemming from the bulk of political philosophy, ordinary people's consumption was not yet of interest. In the hands of Louis XIV, consumption was but a political instrument,

a method of ruling. However, political power had developed in different ways in different countries where the transcendental idea of society was no longer extant.

The Birth of Nations

When Smith was born, France was still an absolute monarchy: the French state was not only very firmly in control throughout the country but also garnered its force and its income from this control. This absolute monarchy had systematically eliminated all intermediary groups, all traditional communities, extended families, religious orders, and guilds in a bid to establish a direct link between each individual and the state. Consequently, France as a whole had become a dynamic tool applied to the construction of a monarchical system. Political economics is concerned more with the confirmation of this grip on an area while political philosophy leans more toward a critique of despotism. In Great Britain the state was not so omnipotent and not so ubiquitous. A constitutional monarchy was set up in 1688 and John Locke's (1632–1704) ideas became widespread. Tolls were practically unknown. The inland communication network was less developed. The French state lived off taxes on internal trade, whereas the English state derived its revenue from taxes on wool, which was a key commodity in foreign trade.

In France the state actively sought to expand its internal market in its own interest. By the close of the fourteenth century, a royal permit was required to set up fairs and markets. In Great Britain prior to Smith's period domestic trade was not taxed very much and tolls were almost unknown. Road and fluvial networks were less well developed than in France. That being said, in Smith's time foreign trade was expanding without changing its nature while domestic trade was undergoing radical change in both countries. In summary, a new form of social relationships appeared, including relationships between people, their relationships with geographical space, with the powers that be, and with what was perceived to be possible; these new relationships took on a different shape, notably in terms of the ways in which, from the fourteenth century onwards, the decomposition of the political form of the empire varied spatially. In fact the nation-state and free markets, which appeared were basically a new way of representing and structuring social space.

The Relation between Nation-State and Market

Adam Smith's work fitted into the dynamics of the development of political economics, the subject of which was precisely this relationship between the nation-state and free markets, between economic activity and political action.

For a while those who were trying to understand this relationship looked at the economy basically from the point of view of those in power. Adam Smith looked at the generation of wealth in general, i.e., abstractly and not merely in terms of the sovereign's interest. Before him, not only did those who looked at the economy see it in terms of the sovereign's benefit, but those who looked at trade saw it only from the traders' standpoint. With Adam Smith, however, it was the nation as a whole that became key: *civil society*, as it became known, took on both economic and political importance. And one of the main reasons why this concept took root in Great Britain was of course that civil society was more advanced there. Indeed in France, the state continued to be a system of power and administration, whereas in England the state saw itself more as providing and guaranteeing the rule of law. It was this that enabled James Steuart (1713–80) in particular, in his *Inquiry*, to envisage political economy on the basis of a relatively autonomous civil society and a state which does not merely govern. His fellow countryman Adam Smith went further. He refused to limit himself to the study of what may be useful to lawmakers or statesmen and sought to understand and be of use to civil society in and of itself. (See the introduction to Book IV in *Wealth of Nations*). By doing this Smith became the first person to dilute the political concept of nation in the economic concept of the market. In his view, free markets and not the state were the number-one factor in generating wealth, whereas the distinction between the domestic and the political economy was no longer fundamental. The science of wealth looks at all those phenomena which, in a modern society wherein the rule of law obtains, are worthy of attention if one wishes to improve the material standard of living of everyone in the society.

As Rosanvallon has shown (Rosanvallon, 1999), Smith's conception of markets enabled certain issues to be analyzed in a radically new way. The first inversion of perspective was to enable relationships between nations to be perceived as being complementary rather than antinomic. Whereas *rapports de force* (power relations) between nations are only to the benefit of the dominant power, trade between nations can be beneficial to all. Political wars, when redefined in terms of competitive trading, become a force for peace and, rather than dilapidating wealth, this force contributes to wealth creation for all. The second major shift concerns that which binds society together. Whereas both Hobbes and Rousseau – despite their opposing ideas – see the social contract in terms of submission (to the sovereign for the former and to the general good for the latter), Smith's free market ideas enable one to conceive of a society without a central node and without the phenomenon of submission. Each individual institutes and perpetuates society in his everyday relationships with others. Rising above the lawmakers, the markets regulate not only the relationships between goods, but also a substantial proportion of the relationships among people. Having sought to run government on the basis of absolute truth with the rationale of the sovereign lord as the inevitable starting point, henceforth, it would be necessary to take into account the rational behavior of those who are governed.

It is no longer a question – as in Hobbes and Rousseau – of knowing how to find in the origins of society the legal setup that would keep power in check from the outset. Henceforth we are dealing with a society that works autonomously and one where subordination phenomena contribute to social development. As Michel Foucault commented, starting from Smith on, "the problem is simply how to regulate and limit power within a society in which subordination is already at work. It is

in this way that the question which has obsessed practically all political thought from the end of the 18th century to the present arises, that is to say, the question of the relations between civil society and the state" (Foucault, 2008(2004): p. 309; Foucault, 2004: p. 312).

A Liberal Stance

In fact Smith, just like Marx later on, made, above all, a synthesis of a substantial amount of the research and thinking that was available to him at the time. He did not invent the labor theory of value nor the theory of the division of labor. This latter concept appears not only in a number of publications at the beginning of the seventeenth century, but also in the works of Plato. Many thinkers prior to Smith had already defended free trade and delved into the price mechanisms for allocating scarce resources and regulating trade. Mandeville and Hume had already split the economy off from moral considerations and were both of the opinion that it is through their own interest that people can be ruled. Steuart had already pointed out that markets are not a substitute for the social contract, they are the driving force heralding democracy (Rosanvallon, 1999: p. 49).

Notwithstanding the foregoing, before Smith no one had presented what has come to be called the *liberal doctrine* in such a systematic and well-argued way. It is not merely a question of the *invisible hand*, which, for some of his lesser followers, constitutes his conception of free markets. Smith only used this expression on three occasions in three separate publications and, as Emma Rothschild has shown, in a fairly ironic manner (Rothschild, 2002: Chapter 5). Michaël Biziou also showed that rather than to a benevolent providence, what Smith refers to in these occasions is the fact that individual actions can have unwanted consequences which can be beneficial as well as harmful to society (Biziou, 2003). In fact, long before Hegel, Adam Smith was the first person to understand civil society in economic terms. According to him, it is the economic link between production and consumption in free markets that binds society together. And 'the market' is not limited to a certain type of exchange; according to Smith it is society as a whole that is organized on an exchange basis. The division of labor for Smith plays a fundamental role in the transformation of society and the consumer becomes a citizen of a secular society. Smith was the first to emphasize the need to create a veritable market society because the market of which he spoke did not yet exist in his time. He believed it was necessary to understand its advantages and to facilitate its emergence. He had a rather negative opinion of the merchants vindicated in the fact that, when exempt from state control, merchants' behavior often tended to be harmful to the common good. Nevertheless, Smith was against the mercantilists, who believed that the state represented the interests of everyone, considered traders to be unproductive, and thought that farm produce was more important than all other activities and trade. But Adam Smith was not an advocate of laissez-faire. Nor was he of the opinion that everything depends on free markets. Not only should the state provide a system of justice and infrastructure (roads, bridges, ports, and so on), it should also promote science as "*the great antidote to the poison of enthusiasm and superstition*" (*Wealth of Nations*, vol. II, Book V, Chapter 1). Smith was also in favor of political regulation to encourage movements of goods trade and to resist market related social problems, such as the exploitation of apprentices. We should bear in mind that the factories and working classes analyzed by Marx did not appear until later.

A Scientific Point of View

In fact the abstract system studied and favored by Smith must be distinguished from what was later to be dubbed capitalism, the main feature of which was not the general good, which Smith aspired to, but rather the accumulation of *abstract capital*. Smith took inspiration from the physical sciences to study the systems of relationships. Inspired by Newtonian physics, he sought to break down and rebuild the mechanisms of society on the basis of a minimum number of assumptions, possibly in the hope of identifying a law similar to the law of universal gravitational attraction formulated by Newton (1643–1727). Smith looked for causal relationships and often ruled out numerous background factors in the hope of perceiving fundamental relationships more clearly. Perhaps he thought he would discover some simple principle of organization, especially since he – as his master and friend Hobbes – had great faith in mother nature and believed, as did many of his contemporaries, that mankind would be wise to pay attention to her. He simplified even more so because the market system turned out to be so complex that understanding all its mechanisms in detail would be beyond the intellectual grasp of legislators. Here again, however, Smith did not reduce society to a mere mechanism. What he did rather was to identify factors that were likely to contribute to the *wealth of nations* relegating other factors to the background noise as it were. The workings of the state in his time did not appear to him to contribute much to the *wealth of nations*. In line with the principles developed by the mercantilists and the physiocrats in the interests of sovereign rulers who were no longer subject to religious ethics, state action usually consisted of encouraging any activity that would fill their coffers with money. Smith had far greater confidence in social relationships that work in such a way that everyone gains through them. This was particularly the case since he was one of those in this period of history who thought that man has a natural propensity to help others.

Self, the Other, and the Social

Smith's well-known theory of self-regulating supply and demand first arises in the moral propensity that he describes in his *Theory of Moral Sentiments*, which was published 20 years earlier in 1756. Many people are indeed astonished that Smith should have arrived at economics from his starting point in moral philosophy. However, according to Dupuy, Smith believed that moral philosophy and economics are one and the same science (Dupuy, 1992: p. 101). Smith's ponderings on moral issues took place in a climate of reflection that was very much present in his Scottish entourage. It had appeared as

a reaction against thinkers such as Hobbes who saw only cynicism and selfishness in man. Smith's own *maitre a penser* in Glasgow, Francis Hutcheson (1654–1746), taught that man had an innate tendency to universal goodwill but was torn between its egoistic and altruistic instincts. Another thinker who influenced Smith, David Hume (1711–76), believed that moral judgments are not rooted in reason but in feelings. Smith's original contribution was to lay stress on *self-love*, and to merge self-interest and self-love. He says "Sympathy, therefore, does not arise so much from the view of the passion, as from that of the situation which excites it" (Smith, 2000(1759): p. 7). Smith emphasizes the importance of our imagination. While we obviously never leave our own minds and bodies, we are constantly imagining ourselves in the place of others. He writes "We sometimes feel for another, a passion of which he himself seems to be altogether incapable; because, when we put ourselves in his case, that passion arises in our breast from the imagination, though it does not in his from the reality" (Smith, 2000(1759): p. 7). Thus sympathy does not arise from *selfishness*. In Smith's view, *self-love* is informed by the recognition of others. To put it another way, the Smithian subject needs his fellow men in order to forge an identity. This foreshadows Hegel (1770–1831) who was later to state that one's identity depends on recognition by others.

A Forerunner

Adam Smith prefigures René Girard's and Jacques Lacan's Theory of Desire, since he emphasizes the wish for the approval of others in the feelings and emotions that motivate one's acts (Bouchet, 2011). This is a most modern conception of human relationships, where each individual is supposed to put her- or himself in someone else's situation, imagining that the other does the same. And it is true that, in a market context, each individual will take actions, be a spectator, and also be both a producer and a consumer. One might almost say that the Smithian *Theory of Moral Sentiments* heralds the rise of sociology and foreshadows some of Durkheim's analyses in his *Elementary Forms of the Religious Life* (1912). Thus we see that Smith may be considered to be a precursor of sociology because he gives priority to social relationships. In fact he takes it for granted that it is human relationships that mediate their relationships to objects. (Notwithstanding this, he does consider exchange to be a natural tendency and not a social category.) Smith may also be viewed as an anti-utilitarian, since he shows that men think they are pursuing utility whereas in fact what they seek is recognition and admiration. Here we are far from the caricature of *homo economicus* that has been disseminated by a form of economic liberalism that no longer has much in common with Smithian sociology.

Adam Smith also foreshadowed the theoreticians of the consumer society. He pointed out that we do not need a great deal to meet our basic needs. Anything beyond that is used mainly to attract attention – gain respect or attain recognition others would later say – through *sympathy*. Smith thus conceives of *sympathy* (which has much to do with empathy) as being the foundation of any evaluation of how we feel. In his *Theory of Moral Sentiments* Part I, Section I, Chapter 3, Smith writes "Upon all occasions, his own sentiments are the standards and measures by which he judges of mine" (Smith, 2000(1759): p. 15). And a few pages later he says "Every faculty in one man is the measure by which he judge of the like faculty in another" (Smith, 2000(1759): p. 18).

Thus, according to Smith, empathy encourages men to reconcile themselves with one another, but one cannot ignore moral values. Free markets cannot work without certain moral virtues, notably prudence and justice. Without these, the economic exchanges and the markets of Smiths' scheme, cannot develop. Sympathy and benevolence also play a role. On another tack, Smith believed that the same mechanisms that can contribute to the wealth of nations and – as he saw it – to social cohesion can also give rise to their corruption. It is odd, therefore, that some still portray Adam Smith as an apostle preaching that the market is a superior substitute for politics and morals. It is also worthwhile highlighting that the economic system, which Adam Smith put forward as a utopian model in his time is not ubiquitous. It is a system founded on a certain type of market, a certain type of mutual confidence and self-perception, and a certain type of property and division of labor. Again, in Smith's day, the economy of the Western world was an aristocratic one based on gifts and privilege. Examples of other societies that have experienced different modes of economic development, other types of bond between material and spiritual phenomena, and people abound: ancient Greece, China, India, and Japan, inter alia. Today we can say that the capitalist system could not be studied until it had come into existence, but that the gradual way in which interest and conceptual analysis in and of it has grown has contributed strongly to its expansion and its continued existence. Smith was as much an analyst as a utopian. He was a social theorist who gave the most powerful theoretical boost to the modern economy and who still continues to have a major influence on the way in which people today perceive their relationships with the world about them and among themselves. Social scientist keep going back to his work to find inspiration for an understanding of social change. Thus, it has recently been used by the sociologist Giovanni Arrighi to analyze the origins and dynamics of China's recent extraordinary rise and the ongoing shift of the epicenter of the global economy from North America to East Asia (Arrighi, 2007).

See also: Enlightenment; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Liberalism: Political Doctrine and Impact on Social Science.

Bibliography

Arrighi, Giovanni, 2007. Adam Smith in Beijing. Lineages of the Twenty-first Century. Verso, London and New York.

Bayly, Christopher Alan, 2005. The Birth of the Modern World 1780–1914. Global Connections and Comparisons. Blackwell, Oxford.

Biziou, Michaël, 2003. Adam Smith et l'origine du libéralisme. Presses Universitaires de France, Paris.

Bouchet, Dominique, 2011. Desire. In: Southerton, Dale (Ed.), Encyclopedia of Consumer Culture, vol. 1. Sage, Los Angeles and London, pp. 440–444.

Dupuy, Jean-Pierre, 1992. Le sacrifice et l'envie. Le libéralisme aux prises avec la justice sociale. Calmann-Lévy, Paris.

Foucault, Michel, 2004. Naissance de la biopolitique. Cours au Collège de France 1978–1979. Gallimard/Seuil (Collection Hautes Études), Paris.

Foucault, Michel, 2008(2004). The Birth of Biopolitics. Lectures at the College de France 1978–1979. Palgrave MacMillan, London.

Montchrétien, Antoyne de, 1889(1615). Traité de l'économie politique. Plon, Paris.

Rosanvallon, Pierre, 1999. Le capitalisme utopique. Histoire de l'idée de marché. Seuil (Collection Points Essais N. 385), Paris.

Rothschild, Emma, 2002. Economic Sentiments. Adam Smith, Condorcet and the Enlightenment. Harvard University Press, Cambridge, MA.

Smith, Adam, 1999(1776). The Wealth of the Nations. Books IV–V. Penguin Books, Harmondsworth.

Smith, Adam, 2000(1759). The Theory of Moral Sentiments. Prometheus Books, Amherst, New York.

Smith, Adam, 2002(1759). The Theory of Moral Sentiments. Edited by Knud Haakonssen. Cambridge University Press, Cambridge.

Smith, Adam, 2004(1776). The Wealth of Nations. Volume 1: Books I–III and Volume 2: Books IV–V. Penguin Books, Harmondsworth.

Vries, Jan de, June 1994. The industrial revolution and the industrious revolution. The Journal of Economic History 54 (2), 249–270.

Vries, Jan de, 2008. The Industrious Revolution. Consumer Behavior and the Household Economy, 1650 to the Present. Cambridge University Press, Cambridge, New York, Melbourne.

Smoking and Health

Kenneth E Warner, School of Public Health, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Cigarette smoking is the leading behavioral cause of preventable, premature death, killing 6 million people annually. This article describes the burden of smoking, explains how epidemiological science unearthed the principal tobacco-related disease connections, and summarizes the causes of tobacco consumption. The article then describes the efforts of social and behavioral scientists to improve the effectiveness of smoking cessation techniques and contribute to the formulation of effective tobacco control programs and policies.

Introduction

Worldwide, 1 billion people smoke cigarettes, with nearly 80% of smokers living in low- and middle-income countries (LMICs). According to the World Health Organization (WHO), 6 million people die annually as a consequence. Smoking is the leading behavioral cause of preventable, premature death throughout the world, responsible for one of every 10 adult deaths. If current trends persist, smoking will claim as many as one billion lives during the twenty-first century (WHO, 2012).

The causes of this pandemic of avoidable disease and death are a complex web of physiological, psychological, social, marketing, and policy factors. Social and behavioral scientists of all disciplinary stripes have contributed to identifying and disentangling these factors, although no single model has yet adequately characterized the interactions of these determinants. Social and behavioral scientists have also elucidated important mechanisms by which both private sector and governmental intervention can discourage the initiation of smoking and encourage quitting.

After a brief description of the burden of smoking and how epidemiological science unearthed the principal tobacco-related disease connections, the causes of tobacco consumption are summarized. The efforts of social and behavioral scientists to improve the ability of smokers to quit smoking are then described, including better theoretical characterization of the determinants of success in quitting and the development of more effective cessation interventions. Next examined is how social and behavioral scientists have analyzed the effects of selected, important tobacco control measures and thereby contributed to the formulation and implementation of tobacco control programs and policies. Finally, the future contribution of the social and behavioral sciences to dealing with a potentially even greater cataclysmic crisis in world health during the remainder of the twenty-first century is considered.

Traditional methods for the prevention of 'youth smoking,' such as school health education and the enforcement of minimum age-of-purchase laws, are only touched upon, due in part to limited scientific understanding of how initiation can be effectively discouraged and also because it is covered elsewhere (Lantz et al., 2000; US Department of Health and Human Services, 2012; *see* Substance Abuse in Adolescents, Prevention of; Health Promotion in Schools).

Cigarette Smoking and Disease: Making the Connection

In the major industrialized nations, smoking causes from a sixth to a fifth of all mortality. The two major smoking-related causes of death are lung cancer and coronary heart disease (CHD). In the United States, epidemiologists estimate that smoking accounts for close to 90% of all lung cancer deaths, with lung cancer the leading cancer cause of death in both men and women. Smoking is credited with better than a fifth of CHD deaths. In addition, smoking causes four-fifths of all chronic obstructive pulmonary disease mortality and just under a fifth of all stroke deaths (US Department of Health and Human Services, 2004). The exposure of nonsmokers to environmental tobacco smoke is also a substantial cause of disease and death (US Department of Health and Human Services, 2006).

Mortality attributable to smoking is rising rapidly in LMICs, reflecting their large but more recent adoption of smoking. WHO estimates that nearly 80% of the world's smokers now reside in LMICs. By 2030, WHO projects that four-fifths of smoking-produced death and disease will occur in LMICs (WHO, 2012).

Although the health hazards of tobacco smoking have been suspected for centuries, serious interest is a twentieth-century phenomenon, coincident with the advent of cigarette smoking. Prior to the twentieth century, tobacco was smoked primarily in pipes and cigars, chewed, or used nasally as snuff. Harsh tobaccos made deep inhalation of tobacco smoke difficult. Tobacco probably exacted a modest toll through the nineteenth century.

In the United States, four factors combined in the early twentieth century to make cigarette smoking the most popular, and lethal, form of tobacco consumption. Perfection of the Bonsack cigarette rolling machine in 1884 introduced relatively inexpensive and neater cigarettes to the market. Discovery of the American blend of tobaccos (combining more flavorful tobaccos from Turkey and Egypt with milder American tobaccos) made deep inhalation of cigarette smoke feasible. The development and marketing of Camel cigarettes in 1913, the first American blend cigarette, introduced this new-generation product to the American public with an advertising campaign often credited with inaugurating modern advertising. Finally, cigarettes were included in soldiers' rations

during World War I, permitting soldiers a quick and convenient battlefield tobacco break. Considered effeminate and unattractive prior to World War I, cigarette smoking came home from it a 'manly' behavior. Since then, cigarette smoking has dominated all other forms of tobacco use by far in the United States and in most countries of the world (Brandt, 2007; Proctor, 2011).

Lung cancer was virtually unknown at the beginning of the twentieth century. By 1930, it had become the sixth leading cause of cancer death in men in the United States. Less than a quarter of a century later, lung cancer surpassed colorectal cancer to become the leading cause of male cancer death. It achieved the same status among American women in the mid-1980s, surpassing breast cancer (US Department of Health and Human Services, 2004).

The rapidly growing rate of lung cancer in the early decades of the century spurred a number of epidemiologists and laboratory scientists to begin investigating the relationship between smoking and cancer. Wynder and Graham (1950) published a now classic retrospective epidemiologic analysis that strongly linked the behavior to the disease. Throughout the decade, a series of articles documented similar findings from major prospective (cohort) mortality studies in the United States and the United Kingdom. Scientific groups in both countries soon thereafter published seminal public health documents indicating smoking as the major cause of lung cancer and a cause of many other diseases (Royal College of Physicians of London, 1962; US Public Health Service, 1964).

By the end of the twentieth century, approximately 70 000 studies in English alone associated cigarette smoking with a wide variety of malignant neoplasms and cardiovascular and pulmonary diseases, as well as numerous other disorders. The modern plague of tobacco-produced disease is now rapidly metastasizing – both figuratively and literally – to the world's poorer nations, the result, in part, of increasing affluence and western image-making.

The Causes of Smoking

Smoking affords users a very mild 'high' and addresses a number of often seemingly competing physical and psychological needs – for relaxation or stimulation, distraction or concentration, for example. The crucial element in the sustained use of cigarettes is the addictiveness of nicotine in tobacco smoke. Once addicted, as most regular smokers are, smoking serves the physiological purpose of avoiding nicotine withdrawal. Nicotine affects brain receptors in much the same manner as other addictive substances, and withdrawal shares many of the same unpleasant characteristics (US Department of Health and Human Services, 2004).

In most countries, smoking is initiated during the teen and even preteen years. At the time of initiation, new smokers typically do not appreciate the nature of addiction, much less the addictiveness of nicotine. Further, they tend to be short-sighted and unconcerned about potential distant adverse health consequences (US Department of Health and Human Services, 2012). The combination translates into a large pool of new smokers who have adopted the behavior and become addicted, without considering either the danger or the addictiveness of smoking; a situation that most will come to regret.

The physical effects of nicotine notwithstanding, history teaches that virtually all drug use is socially conditioned; tobacco smoking is no exception. The use of tobacco in the Americas prior to the arrival of Europeans demonstrates this vividly. Although tobacco played a prominent role in most native societies, in some, tobacco smoking was restricted exclusively to the shaman, who used it for medicinal and religious purposes. In other societies, tobacco smoking anointed the official nonsectarian functions of tribal leaders (such as the famous 'peace pipe'). In still others, nearly all males smoked tobacco frequently, for personal and social reasons (Goodman, 1993).

In contemporary society, the initiation of cigarette smoking is often viewed as a rite of passage to adulthood. In many poorer societies, smoking is seen as a symbol of affluence. In virtually all cases, smoking results from role modeling of parents, prominent celebrities, or youthful peers.

Many tobacco control advocates attribute much of smoking's holding power to sophisticated tobacco industry marketing campaigns. In the United States alone, the industry spends billions of dollars per year on advertising and other forms of marketing. The industry insists that the purpose of its advertising is solely to vie for shares of the existing market of adult smokers, a claim greeted with derision by the 'public health' community. With US smokers exhibiting strong brand loyalty and two companies controlling over 70% of the US market, marketing directed at brand switching (or its defensive analog, maintaining brand loyalty) would appear to be a relatively fruitless investment. Tobacco control advocates thus believe that much of the industry's marketing effort is directed toward attracting new smokers, primarily children but also groups of first- and second-generation American adults not yet fully acculturated into American society (e.g., Hispanic females, among whom there is very low prevalence of smoking).

Similarly, the introduction of aggressive Western cigarette marketing into Asian, Eastern European, and African countries has been attacked by the public health community as spurring the growth of smoking among children and other traditionally nonsmoking segments of society (e.g., Asian females). In many of these countries, cigarette advertising was nonexistent or restricted to modest use of print media prior to the introduction of advertising by the major multinational tobacco companies. Before the dissolution of the Soviet Union, there was virtually no advertising of the state-produced cigarettes in the Eastern European countries. That ended with a flourish when multinational tobacco companies entered the newly liberated economies and, in some instances, took over cigarette production from the inefficient state-run enterprises. In Japan, the state tobacco monopoly had never bothered to advertise on television prior to the entry of the multinationals. In short order, competition from the advertising of Western cigarette brands led to cigarettes becoming the second most advertised product on Japanese television. In Africa, the advertising media portray smoking as the indulgence of the much–admired affluent set, with billboards in desperately

poor villages depicting high-society Africans smoking and smiling in convivial social settings.

Although many tobacco control advocates find advertising a convenient villain to attack, widespread smoking has often preceded formal marketing. Tobacco smoking spread like wildfire through much of Europe in the sixteenth and seventeenth centuries. Similarly, male smoking rates in Russia and the Eastern European countries were very high prior to the fall of Communism and the advent of modern cigarette marketing. The effects of advertising are further examined in Section Analysis of the Effects of Tobacco Control Policies.

The Art and Science of Smoking Cessation

In countries in which the hazards of smoking have been well publicized, surveys find that most smokers would like to quit and most have made at least one serious 'quit attempt'; yet relatively few of those who try to quit succeed in any given attempt. In the United States, for example, approximately 70% of smokers report that they want to stop and as many as half the smokers try to quit each year. Yet only 2.5% succeed in quitting each year. Combined with persistent cessation efforts by many smokers, this modest quit rate has created a population in the US in which there are as many former smokers as current smokers. Thus, in the aggregate, quitting has significantly reduced the toll of smoking but the toll remains stubbornly high due to the difficulty of quitting. Concerns about that difficulty are exacerbated by data indicating that adult smoking prevalence in many nations leveled off during much of the past decade – in the United States it did not change at all from 2004 to 2009 and fell only slightly during the next 2 years – and it has even increased in a few countries that previously had managed to reduce prevalence significantly. For example, in Singapore, smoking prevalence rose from 12.6 to 14.3% from 2004 to 2010.

Although most former smokers quit without the aid of formal programs or products (Chapman and MacKenzie, 2010), the widespread desire to quit, paired with its difficulty, has created a small but thriving market for interventions that increase the odds of quitting smoking. Formal cessation interventions range from the distribution of how-to-quit booklets, to mass media cessation campaigns, to individual- and group-counseling programs, to self-administered nicotine replacement therapy products, to use of cessation pharmaceuticals combined with clinical counseling. The efficacy of interventions ranges from sustained quit rates of about 5–7% for the most general and least resource-intensive interventions (e.g., generic cessation booklets) to 30% or more for the most resource-intensive programs that combine sophisticated counseling with the use of pharmaceuticals (Fiore et al., 2008).

Over the years, behavioral scientists have helped refine smoking cessation techniques by evaluating interventions and developing theory applicable to smoking cessation. Preeminent in the domain of theory have been models that characterize how smokers progress, through a series of stages, to contemplate, attempt, and eventually succeed or fail in quitting, with cessation maintenance also examined (Prochaska and DiClemente, 1983). Relating cessation advice to smokers' stages of readiness to change constitutes one way of 'tailoring' cessation messages. Another involves tying cessation advice to the specific motivations and concerns of individual smokers trying to quit. For example, consider a smoker motivated to quit primarily by the high cost of smoking but also concerned about gaining weight. Armed with this knowledge, a cessation counselor can develop specific information on the financial savings the smoker can expect once he or she quits, while suggesting specific strategies to avoid weight gain (or to deal with it if it occurs).

Tailored cessation messages have the obvious virtue of meeting the idiosyncratic needs of individual smokers. In the absence of computer technology, however, they would entail the substantial cost of collecting information on the individual's needs and concerns and then developing individualized (tailored) cessation advice for each person. Computers, however, permit simple and inexpensive collection and translation of information into tailored cessation materials, such as individualized advice booklets and calendars with specific daily advice and reminders (Strecher, 1999). Online programs can provide instant tailored suggestions to help people quit smoking. The concept of tailoring holds great potential to enhance quit rates both in individual counseling situations and in low-cost online programs.

Social and behavioral scientists have also played a central role in defining appropriate medical treatment of smokers. The US Public Health Service's clinical guideline for smoking cessation was produced by an expert committee, including social and behavioral scientists who had worked on smoking cessation as service providers, developers or evaluators of interventions (Fiore et al., 2008). The guideline urges physicians to regularly counsel their smoking patients about the implications of the behavior and encourage them to quit. It concludes that a highly effective cessation approach involves physician counseling to quit, supplemented with use of cessation pharmaceuticals and maintenance advice through follow-up contacts. A cost-effectiveness analysis of an early version of the guideline added economics to the social sciences contributing to understanding optimal smoking cessation therapy (Cromwell et al., 1997).

Despite substantial improvements in the efficacy of cessation treatments, in any given year only a small fraction of the smokers who say they want to quit succeed in doing so; and only a small fraction of these have employed professional or programmatic assistance. As such, in any short-run period of time, the contribution of smoking cessation interventions to reduce the health toll of smoking is modest at best. This source of frustration has led a subset of smoking cessation professionals to explore methods of achieving 'harm reduction' that does not depend on smokers completely overcoming addiction to nicotine. Harm reduction techniques range from helping smokers to reduce their daily consumption of cigarettes to encouraging consideration of the long-term use of lower-risk nicotine-delivery devices, such as nicotine 'gum' or the patch or even low-nitrosamine smokeless tobacco products. Though fraught with problems, harm reduction may be an idea whose time has come (Zeller et al., 2009).

Analysis of the Effects of Tobacco Control Policies

Another way to grapple with the toll created by smoking is to develop policies that discourage the initiation or continuation of smoking or restrict it to areas in which it will not impose risk on nonsmokers (Warner, 2006). Social and behavioral scientists have devoted substantial effort to studying the impacts as well as the processes by which such policies come to be adopted. This section examines the former. (For the latter, see, e.g., Fritschler and Rudder, 2007; Kagan and Vogel, 1993.)

In 1999 the World Bank described and evaluated a wide array of tobacco control policies in countries around the globe (Jha and Chaloupka, 1999). The research documented in that volume and its more detailed companion book (Jha and Chaloupka, 2000) became the evidence base for the policy requirements of the WHO's 'Framework Convention on Tobacco Control' (FCTC) – the world's first global health treaty, now ratified by 177 nations (Framework Convention Alliance, 2013). A major goal of the FCTC is to get LMICs to adopt the basket of evidence-based tobacco control policies that have reduced smoking in developed nations. Those countries that have devised and invested in serious national plans to curb tobacco use have achieved considerable success, beginning with Norway in the 1970s and, more recently, Canada and Australia, each of which has driven smoking prevalence down, faster and lower than the rest of the developed world. Recently, a few countries in Central and South America have developed significant national tobacco control plans, including Uruguay whose aggressive graphic warning legislation has earned the ire of – and a lawsuit from – Philip Morris International. The US Department of Health and Human Services announced a national plan to combat smoking in 2010, the first such national effort in the United States (US Department of Health and Human Services, 2010). California, the state that has achieved most success in tobacco control in the United States, has had an orchestrated state tobacco control program for years (California Department of Health, 2013), although it has struggled through several attempts by the state legislature to redirect funds targeting tobacco control.

Three specific tobacco control policies have commanded the greatest amount of research attention: taxation, smoke-free workplace and public place legislation, and bans on advertising and promotion. We focus below on the social and behavioral research related to each.

Taxation

Research, primarily by economists, has established that taxing cigarettes is among the most effective measures to decrease smoking (Chaloupka and Warner, 2000; Chaloupka et al., 2012). Because the quantity of cigarettes consumed declines by an amount proportionately smaller than the associated tax rise, taxation increases government revenues at the same time that it decreases smoking. In developed countries, economists find that a 10% increase in cigarette price induces approximately a 4% decrease in quantity demanded. In developing countries, the demand impact may be twice as large. Effective dissemination of research findings has made taxation one of the central tenets of a comprehensive tobacco control program and a major component of the Framework Convention on Tobacco Control.

Although the 'bottom line' about taxation is well established, research by economists and others points to concerns that remain unanswered. For example, does taxation discourage the initiation of smoking? Although evidence preponderantly suggests that it does, a few studies challenge the conventional view (Chaloupka and Warner, 2000; Chaloupka, 1999). 'Side effects,' or subtle unanticipated impacts of taxation, warrant additional attention as well. Notably, one study found that increases in cigarette tax in the United States caused some smokers to switch to higher nicotine cigarettes to get their customary dose of nicotine from fewer cigarettes, particularly among younger smokers (Evans and Farrelly, 1998).

In the relatively new field of behavioral economics, a small cadre of psychologists is studying experimentally how smoking responds to incentives. For example, investigators give subjects a daily 'budget' (e.g., play money) with which to buy cigarettes, food, and other commodities, at prices established by the investigators, and evaluate the effects of 'taxation' by raising the price of cigarettes. Findings generally have been quite consistent with those in the mainstream economics literature (Chaloupka et al., 1999).

Behavioral economics can investigate questions not addressable using existing real world data, but the approach also has decided limitations in predicting how people will respond to price changes in the real world. By learning from each other, the two fields can enrich the evidentiary base for future tobacco tax policy.

Smoke-Free Workplace and Public Place Legislation

Smoke-free laws, originally labeled 'clean indoor air laws,' restrict or prohibit smoking in public places and workplaces. Legislation grew out of concerns that the exposure of nonsmokers to environmental tobacco smoke (ETS) could create risk of disease, leading in the United States to rapid diffusion of state laws beginning in 1973. A decade later, similar laws emerged at the local level of government, followed by much stronger state laws beginning in the early 2000s. Indeed, since the first truly comprehensive smoke-free workplace law was adopted in the state of Delaware in 2003, some 30 states have implemented completely smoke-free workplace legislation, laws that cover all restaurants and bars as well as other workplaces. Internationally, over 30 countries have adopted completely smoke-free workplace laws, led by Ireland in 2004.

The scientific knowledge base actually followed early diffusion of clean indoor air legislation, with a number of studies of the relationship between ETS exposure and the risk of lung cancer published in the 1980s (US Department of Health and Human Services, 2006). By the 1990s, the research base had become sufficiently strong for the US Environmental Protection Agency (EPA) to declare ETS a 'Class A Carcinogen' – a proved environmental cause of cancer in nonsmoking humans. The EPA also detailed nonfatal respiratory disease effects of ETS, particularly in children (US Environmental Protection Agency, 1992). Thereafter, research

implicated ETS in heart disease deaths in adult nonsmokers, possibly in an order of magnitude greater than the lung cancer impact (US Department of Health and Human Services, 2006).

Social and behavioral scientists have informed the debate by studying the process of the emergence and diffusion of clean indoor air laws and, subsequently, the effects of such laws on nonsmokers' exposure to ETS and on smokers' smoking behavior. Studies have found generally high levels of compliance with the laws even in bars, and reductions in employees' exposure to ETS, as measured by self-reports, air sampling, and examination of employees' body burdens of cotinine, a nicotine derivative. Other research has demonstrated that smoke-free laws increase the smoking cessation rate and decrease daily smoking by continuing smokers. Research by social scientists has established that hospitality industry venues – primarily restaurants – do not lose revenue following the adoption of smoke-free laws, a finding that has played an important role in debates about the desirability of laws covering restaurants and bars (Warner, 2006). Recently, the most dramatic finding to emerge from analysis of the effects of smoke-free laws is that jurisdictions that adopt such laws see a decrease in hospital admissions for acute myocardial infarctions (Lightwood and Glantz, 2009). Over 100 studies worldwide have documented the phenomenon. The idea that a law prohibiting smoking in public places and workplaces could reduce heart attack deaths by possibly 10–20%, (common findings in the studies), suggests that such legislation may be one of history's most effective public health interventions.

Advertising and Promotion

A diverse group of sociologists, psychologists, economists, marketing specialists, and public health scholars has studied the effects of marketing on smoking. Research by several scholars concludes that modern Western-style advertising creates an imagery that many people, including large numbers of children, find easily recognizable and attractive. There is a strong correlation between children's interest in cigarette marketing campaigns and their subsequent smoking behavior. Temporally, advertising campaigns directed at specific segments of a population have often been followed by significant growth in smoking in the targeted groups, including women in the US in the 1960s, young women in Asia in the 1990s, and children in many countries (Warner, 2006).

Whether any of these associations constitutes a causal relationship is the essential question. The same unidentified factor that makes cigarette advertising attractive to certain children could account for their subsequent smoking, independent of the advertising per se. Similarly, cigarette marketers might foresee an (independently occurring) expansion of a market segment and dive into the advertising void to compete for shares of that new market.

In the absence of the ability to run randomized controlled trials, the empirical analysis of the relationship between cigarette advertising and consumption has been unable to prove a causal connection between the two or to estimate its likely extent (US Department of Health and Human Services, 1989; Warner, 2006). However, strong evidence supporting causality comes from a study reexamining data on the relationship between countries' policies with regard to cigarette advertising (ranging from no restrictions to complete bans) and levels of smoking within those societies (Saffer and Chaloupka, 1999). Blending marketing theory with empirical analysis, this study concluded that a complete ban on cigarette advertising would decrease smoking by about 6%, while partial bans (e.g., banning cigarette ads on the broadcast media) would be unlikely to have an impact on cigarette consumption.

Combined with the previously existing evidence, this study leads to the most plausible interpretation of the relationship between advertising and cigarette consumption. It is a conclusion that probably will satisfy neither tobacco industry defenders of the 'brand-share only' argument, nor tobacco control advocates who condemn marketing efforts as the principal cause of smoking. Cigarette advertising and other forms of marketing probably do increase the amount of smoking in a statistically significant manner. (The amount is likely to vary among societies, depending on the maturity of the smoking market and on familiarity with large-scale marketing campaigns.) The converse, of course, is that advertising and marketing almost certainly do not account for the majority of cigarettes smoked. For this, one must turn to other influences, all less policy-tractable than advertising, including role modeling, peer behavior, and the addictiveness of nicotine and smoking.

Interaction of Policies

As policy research has become more sophisticated, the relationships among policies and their impacts have come to be better understood. Illustrative is research examining the joint effects of tax increases and the adoption of clean indoor air laws. For example, some of the smoking decline credited to tax increases might be associated with a citizenry more interested in reducing smoking. In turn, the latter could be reflected in the adoption of clean indoor air laws. Examining the joint effects of the two policies confirmed researchers' suspicions, thereby suggesting a reduced (but still quite substantial) impact of taxation on smoking. Similarly, disentangling the impacts of state-level tobacco control programs that mix tax increases with media antismoking campaigns is important but analytically challenging. Several research studies pave the way toward better evaluation of multiple-component policies (Warner, 2006). There is widespread agreement that effective tobacco control requires a multiple-component policy environment (Warner, 2000). Hence the need to better understand the interactions among policies.

The Future of Social and Behavioral Science Contributions to Tobacco Control

Social and behavioral science have contributed greatly toward understanding how and why the epidemic of smoking has evolved throughout the twentieth century. They have also elucidated a set of tools that can help society to extricate itself from the tenacious grip of this public health disaster.

With innovations in information and pharmacological technology combining with better insights into human behavior, we can anticipate further improvements in assisting smokers to quit. Although the long-range goal must focus on preventing future generations of children from starting to smoke, helping current adult smokers to quit remains critical to reducing the tobacco-attributable burden of disease and mortality over the next several decades; reductions in youth initiation will not show up in aggregate mortality numbers sooner than 30–40 years following the change in initiation rates.

The effectiveness of efforts to encourage smokers to quit depends to a significant degree on the environment in which smoking occurs. If smoking is increasingly viewed as antisocial, quitting smoking will become easier and more urgent for current smokers. Policy making is society's best means of intentionally altering the environment; and better understanding of the effects of policy interventions and how they come to be adopted will be crucial to shaping the environment with regard to smoking in the coming years.

The multidimensional problems associated with tobacco use vividly illustrate the need for scientists of all disciplinary persuasions to work together. The next great challenge confronting the diverse fields of social and behavioral science in tobacco control lies in learning how to combine the methods and insights of the various disciplines into an integrated whole.

See also: Addictions: General Considerations; Drug Use and Abuse: Psychosocial Aspects; Health Behavior, Psychosocial Theories of; Health Education and Health Promotion.

Bibliography

Brandt, A.M., 2007. The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product that Defined America. Basic Books, New York.

California Department of Health, 2013. California Tobacco Control Program. http://www.cdph.ca.gov/programs/tobacco/pages/default.aspx (accessed 17.03.13.).

Chaloupka, F.J., 1999. Macro-social influences: the effects of prices and tobacco control policies on the demand for tobacco products. Nicotine & Tobacco Research 1 (51), 105–109.

Chaloupka, F.J., Grossman, M., Bickel, W.K., Saffer, H. (Eds.), 1999. The Economic Analysis of Substance Use and Abuse: An Integration of Econometric and Behavioral Economic Research. University of Chicago Press, Chicago.

Chaloupka, F.J., Warner, K.E., 2000. The economics of smoking. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics. Elsevier, Amsterdam, pp. 1539–1627.

Chaloupka, F.J., Yurekli, A., Fong, G.T., 2012. Tobacco taxes as a tobacco control strategy. Tobacco Control 21 (2), 172–180.

Chapman, S., MacKenzie, R., 2010. The global research neglect of unassisted smoking cessation: causes and consequences. PLoS Medicine 7 (2), e1000216. http://dx.doi.org/10.1371/journal.pmed.1000216.

Cromwell, J., Bartosch, W.J., Fiore, M.C., Hasselblad, V., Baker, T., 1997. Cost-effectiveness of the clinical practice recommendations in the AHCPR guideline for smoking cessation. Journal of the American Medical Association 278, 1759–1766.

Evans, W.N., Farrelly, M.C., 1998. The compensating behavior of smokers: taxes, tar, and nicotine. RAND Journal of Economics 29, 578–595.

Fiore, M.C., Jaen, C.R., Baker, T.B., et al., 2008. Treating Tobacco Use and Dependence: 2008 Update. Clinical Practice Guideline. US Department of Health and Human Services, Public Health Service. http://bphc.hrsa.gov/buckets/treatingtobacco.pdf (accessed 17.03.13.).

Framework Convention Alliance, 2013. What is the Framework Convention on Tobacco Control?. http://www.fctc.org/index.php?Itemid=5&id=8&option=com_content&view=article (accessed 15.10.13.).

Fritschler, A.L., Rudder, C.E., 2007. Smoking and Politics: Bureaucracy Centered Policy Making, sixth ed. Prentice Hall, Upper Saddle River, NJ.

Goodman, J., 1993. Tobacco in History: The Cultures of Dependence. Routledge, New York.

Jha, P., Chaloupka, F.J., 1999. Curbing the Epidemic: Governments and the Economics of Tobacco Control. World Bank, Washington.

Jha, P., Chaloupka, F.J., 2000. Tobacco Control in Developing Countries. Oxford University Press, New York.

Kagan, R.A., Vogel, D., 1993. The politics of smoking regulation: Canada, France, and the United States. In: Rabin, R.L., Sugarman, S.D. (Eds.), Smoking Policy: Law, Politics, & Culture. Oxford University Press, New York, pp. 22–48.

Lantz, P.M., Jacobson, P.D., Warner, K.E., Wasserman, J., Pollack, H.A., Berson, J., Ahlstrom, A., 2000. Investing in youth tobacco control: a review of smoking prevention and control strategies. Tobacco Control 9, 47–63.

Lightwood, J.M., Glantz, S.A., 2009. Declines in acute myocardial infarctions after smoke-free laws and individual risk attributable to secondhand smoke. Circulation 120, 1373–1379.

Prochaska, J.O., DiClemente, C.C., 1983. Stages and processes of self-change of smoking: toward an integrative model of change. Journal of Consulting and Clinical Psychology 51, 390–395.

Proctor, R., 2011. Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition. University of California Press, Berkeley.

Royal College of Physicians of London, 1962. Smoking and Health: A Report on Smoking in Relation to Cancer of the Lung and Other Diseases. Pitman Publishing Co., London.

Saffer, H., Chaloupka, F.J., 1999. Tobacco Advertising: Economic Theory and International Evidence. Working Paper No. 6958. NBER National Bureau of Economic Research, Cambridge, MA.

Strecher, V.J., 1999. Computer-tailored smoking cessation materials: a review and discussion. Patient Education and Counseling 36, 107–117.

US Department of Health and Human Services, 1989. Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General. US Department of Health and Human Services Public Health Service, Centers for Disease control. National Center for Chronic Disease Prevention and Health Promotion, Office of Smoking and Health (DHHS Publication No. (CDC)), Rockville, MD, 89–8411.

US Department of Health and Human Services, 2004. The Health Consequences of Smoking: A Report of the Surgeon General. US Dept. of Health and Human Services, Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health http://www.cdc.gov/tobacco/data_statistics/sgr/2004/complete_report/index.htm (accessed 17.03.13.).

US Department of Health and Human Services, 2006. The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, Coordinating Center for Health Promotion, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/preventing-youth-tobacco-use/full-report.pdf (accessed 17.03.13.).

US Department of Health and Human Services, 2010. Ending the Tobacco Epidemic: A Tobacco Control Strategic Action Plan for the U.S. Department of Health and Human Services. Office of the Assistant Secretary for Health, Washington. http://www.hhs.gov/ash/initiatives/tobacco/tobaccostrategicplan2010.pdf (accessed 17.03.13.).

US Department of Health and Human Services, 2012. Preventing Tobacco Use among Youth and Young Adults: A Report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention. US Government Printing Office, Washington, DC. http://www.surgeongeneral.gov/library/reports/preventing-youth-tobacco-use/full-report.pdf (accessed 17.03.13.).

US Environmental Protection Agency, 1992. Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders. US Environmental Protection Agency, Washington.

US Public Health Service, 1964. Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service. US Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control PHS Publication No. 1103, Washington.

Warner, K.E., 2000. The need for, and value of, a multi-level approach to disease intervention: the case of tobacco control. In: Smedley, B.D., Syme, S.L. (Eds.), Promoting Health: Strategies from Social and Behavioral Research. National Academy Press, Washington, pp. 417–449.

Warner, K.E., 2006. Tobacco policy research: insights and contributions to public health policy. In: Warner, K.E. (Ed.), Tobacco Control Policy. Jossey-Bass, San Francisco, pp. 3–86.

World Health Organization, 2012. Tobacco. Fact sheet no. 339 http://www.who.int/mediacentre/factsheets/fs339/en/index.html (accessed 17.03.13.).

Wynder, E.L., Graham, E.A., 1950. Journal of the American Medical Association 143, 329–396.

Zeller, M., Hatsukami, D., Backinger, C., et al., 2009. The strategic dialogue on tobacco harm reduction: a vision and blueprint for action in the U.S. Tobacco Control 18, 324–332.

Social Action and Social Justice

Sarah Todd, School of Social Work, Carleton University, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work is unique as a profession, largely as a result of its commitment to addressing both individual problems and the social conditions that create and sustain those problems. Social action, social justice, and organizing for change are integral to social work, but their prominence has varied throughout the profession's history, largely in relation to the visibility and activity of broader social movements. In contemporary social work, social justice remains centrally important, but is operationalized more in daily practice and rarely linked to social movements.

Introduction

The International Federation of Social Workers and International Association of Schools of Social Work collaborated in 2000 and 2001 to create a shared definition of social work. Included in that definition are the statements that "the social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well being" and that "principles of human rights and social justice are fundamental to social work" (www.ifsw.org). Social work's intellectual core is unique because of its simultaneous concern with the problems faced by individuals, particularly those who are marginalized, vulnerable, and oppressed, and the larger structures and social relations that create and sustain individual struggles (Reamer, 1994). However, this dual commitment is not without tension. The social work definition of social justice is murky and the profession's actions often reflect this lack of clarity. At times, social work has aligned with dominant ideologies that marginalize and harm disenfranchised communities. For example, social workers continue to be heavily engaged in practices of surveilling poor people, particularly around parenting. This surveillance, while couched in supportive language, is invasive and regulatory. As Margolin (1997) notes, in many social work interventions,

> Members of the dominant group are active, and members of the subordinate, passive: one party sees, the other is seen; one writes, the other is written about; one purveys knowledge and direction, the other gratefully absorbs and obeys. And while not all families involved in these programs are on the same economic plane, there is a continuing story of economic deprivation: living below the poverty level appears again and again as clients' distinguishing characteristic.
> p. 176

At other times, and in other contexts, the fact that social work occupies a space so directly linked with those who are disadvantaged by global economic and social inequality has meant that workers within the profession have played the roles of initiator, facilitator, advocate, and ally, concerned with how inequality harms individuals and groups. An example of this is the work of Jayesh Patel, who cofounded the nongovernmental organization Manav Sadhna in Ahmedabad, India, in 1990. Manav Sadhna collaborates with people living in poverty to support children enrolling and attending school, facilitate women's economic independence, and ensure people in the community have enough to eat. Social work's conceptualization of social action and social justice has been most significant in terms of imagining a way for those who work closely with the state to ally with social movements for social justice.

Many social workers work in organizations that receive state funding and also support public community protests against how states sustain and exacerbate economic inequality. Throughout the world, social workers have allied with people in poverty, whether this is through participation in Occupy Movements, activism against austerity measures and the facilitation of community-based survival strategies in Greece, or involvement in United Nations-Habitat programs in Nairobi, Kenya, that seek to improve the overall well-being of people living in slums through addressing human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS), shelter, inadequate services, unemployment, and land tenure.

There are, arguably, three types of social work interventions: those that are focused on maintaining the social order, those that are therapeutic in nature, and those that are focused on social transformation. The social transformative focus in social work is concerned with questions of how power is distributed among people, historical injustices against minority groups, the recognition of disenfranchised groups, and the distribution of resources (Dominelli, 2009). It raises concerns about the ways in which societies historically devalue and stereotype groups of people so as to justify and maintain unjust laws and distribution of resources. The resulting interventions are concerned with "ensuring social justice at individual, group and community levels and in seeking structural change to make this happen" (Dominelli, 2009: p. 13). In seeking structural change, transformative interventions tend to rely on organizing groups for collective action. For example, social workers with the First Nations Child and Family Caring Society have spent years challenging the structural inequalities facing indigenous communities in Canada, including poverty and inequities in access to education, health care, and adequate housing. The organization has initiated a number of collective action campaigns to mobilize citizens to work toward social justice for indigenous people in Canada.

Mapping Social Work's Commitment to Social Justice

Social work's commitment to organizing for change, as opposed to solely intervening in the lives of individuals, has two interrelated historical paths. Much of the profession's investment in collective action comes out of the successes of community-based activism by people of color and people living in poverty, such as the civil rights movements, indigenous people's movements across the globe, and global protests against the exploitation of people and resources in the Global South (Lee, 2011). These social movements have challenged therapeutic approaches to social problems, encouraged the profession to adopt a stronger focus on social justice, and informed the profession's understanding of how one can engage in social action to create change. Their influence has driven social work to challenge oppression through a wide range of issues, including racism, sexism, ableism, colonization, heterosexism, and poverty (Thompson, 2002). In particular, people living in poor communities have repeatedly raised their voices to critique the often dehumanizing and humiliating mechanisms through which various governments respond, or fail to respond, to their needs. In turn, they have challenged social work to critique and challenge the profession's complicity in these systems.

A second historical path started with the late-nineteenth-century settlement house movements, which were first founded in Britain and then exported to North America. The most famous settlement house worker was Jane Addams, a Chicago-based social worker whose innovations in working with disenfranchised communities provided much of the intellectual and practical foundation for the social work profession. Jane Addams preferred collective approaches to resolving social problems, seeing society's harms in sociological rather than psychological terms. This was, and continues to be, a disruption of much of social work practice that has aligned more closely with psychological theories. Settlements involved mostly middle-class people living in impoverished communities and provided a range of community-based interventions to help community members address their own problems. Settlement workers' strategies for change included those that the communities could create themselves, research into the social conditions in which people were living, and social action to achieve state-level legislative changes. Settlement movements engaged local citizens to create support groups and recreational programs, establish health clinics, build playgrounds, create juvenile courts, and organize protests to change labor laws and other key pieces of legislation that were seen as oppressive (Stebner, 1997).

The influence of settlement house movements can be seen in the work done by social workers in the contemporary context. For instance, in Ontario, Canada, community health centers are situated in impoverished communities and have a mandate to serve them. The social workers in these centers support citizen campaigns to affect policy and legislative change and work with community members to organize community kitchens and gardens, homework clubs, recreational programs, and all-candidate debates. Thus, social workers collaborate with citizens to establish groups and programs within the community as a space of 'prefigurative politics,' where disenfranchised persons use the skills they develop in the local sphere to engage with broader social movements (Breines, 1989). When social workers organize for social change, it is often about engaging people who have had little opportunity to influence their own environments, so that they can have a collective experience of affecting change. While social work interventions are usually much smaller scale than those witnessed in social movements, they are the space in which many citizens gain their first experiences in joint action and build their ability to work toward social justice. For example, in many disenfranchised neighborhoods, organizing an all-candidates meeting during elections exposes people to the political platforms of politicians and creates the opportunity for community members to understand what those platforms mean for their daily life. This is often one of the first times when many members of disadvantaged communities engage with the political process and begin to see that by having a political voice they can sometimes impact their community's health and well-being.

While the community-based social work that began in the nineteenth century has had a recent resurgence in certain parts of the world, overall, the social work practice of organizing for change has waxed and waned in popularity. After the settlement house movement and the depression years, there was a strong focus on professionalizing social work combined with the marginalization of social activism. The profession's commitment to social action and social justice then resurged with the social movements of the 1960s and 1970s. For example, social workers were quite active during the women's rights movement of the 1960s and this, in turn, challenged social work as a profession to attend to the ways social policies and social work practices impacted on women's lives. Since the 1980s, writers have observed an enduring discourse of social justice, but one that has been largely focused on day-to-day justice rather than the social, economic, and political systems that sustain inequality (O'Brien, 2010). Social justice work today seems to involve a range of activities including engaging in advocacy, negotiating unjust systems, and treating marginalized people with dignity and respect (McLaughlin, 2011). At times, social work is a profession that pushes social justice agendas, but often social movements are challenging the profession to recommit to social justice. For example, social workers in a particular geographic area may work with people who come together for joint action because of an oppressive issue affecting those who share a physical space or because they have a common experience of identity marginalization. Social workers have often worked closely with communities affected by environmental degradation due to oil sands developments or nuclear energy plants, or in communities that have been adversely affected by the destruction of fish stocks due to international overfishing practices (Lundy, 2011). Social workers such as John Coates (2003) have challenged the profession to center ecology in our understandings and work for justice. More recently, Lena Domnelli (2012) has argued for green social work that integrates concern for human needs with an understanding of physical environments and the biosphere.

In the current neoliberal context of global politics and economics, social workers have found it challenging to be involved in social action. There have been times when the principled statements of the profession's commitment to social justice have seemed marginal in social work courses and have

been rarely integrated into practice (Specht and Courtney, 1994). However, there remains quite a bit of variation around the world in terms of how much schools of social work and social work practitioners have been able to maintain a social justice focus in their pedagogy and work. While much social work practice in the United States and Britain has shifted away from a commitment to creating social change, the profession in Latin America and the Caribbean has created and sustained progressive elements. In other nations, such as the United Kingdom, Canada, and Australia, there is a strong intellectual tradition in critical and structural social work, but the struggle remains to maintain that focus in practice (Ferguson, 2008).

The history of social work also includes a never-ending tension between the notion of professionalization and social activism. There has always been a debate as to whether social work should be part of a movement to achieve social justice or whether it should primarily be a profession with evidence-based practices deployed to help individuals, groups, and communities adjust to their environments (Olson, 2007). This tension is observable when comparing the work of authors such as Martin Davies, who, in his book, *The Essential Social Worker* (1981/1994), understands social work as a centrist profession that is focused on the compassionate care of vulnerable individuals. In contrast, writers such as Maurice Moreau (1979), Dominelli and McLeod (1982), and Fook (1993) conceptualize social work as a profession that attends to how relations of heterosexism, abelism, racism, sexism, and capitalism shape people's individual problems and is thus highly political and collectivist in nature. Professional credibility is important in market economies and essential for gaining political, economic, and social control over an occupation, and the progressive and social activist elements of social work are seen by many as a threat to this credibility (Figueira-McDonough, 2007). Empirical research from the United States has suggested that as social work professionalizes, members become more involved with conventional politics and less involved with broad social movements; however, given growing social movements concerned with global inequality and injustice, social workers may once again repoliticize and organize around structural issues (Reeser and Epstein, 1990; Hoefer, 2012). The challenge for many in the social work profession continues to be how to craft a notion of professionalism that is consistent with a commitment to equality and social justice.

Conceptualizing Social Action and Social Justice

Much work has been done in recent decades to clarify the nebulous terrain of social justice within social work. The profession's conceptualization of social justice is informed by theorists such as John Locke (1988), John Rawls (1999), Carol Gilligan (1982), Grace Clment (1996), Friedrich Nietzsche (2008), and Amartya Sen (2011). The breadth of these authors and their ideas about social justice is reflected in the profession itself having a diverse range of conceptions of social justice. One of the most useful definitions of the concept was developed by Craig (2002), who states that social justice is

> a framework of political objectives, pursued through social, economic, environmental and political policies, based on an acceptance of difference and diversity and informed by values concerned with: achieving fairness, and equality of outcomes and treatment; recognizing the dignity and equal worth and encouraging the self esteem of all; the meeting of basic needs; maximizing the reduction of inequalities in wealth, income and life chances; and the participation of all including the most disadvantaged.
> pp. 671–672

This vision of social justice reflects the profession's dual mandate to attend to the individual and the environment that shapes people's lives. As a result, the profession's understanding of social justice is one in which broad-scale structural justice goes always hand in hand with the quest for individual well-being. We attend to social justice in micro and macro relations of power. This dual focus has created some ambiguity, but it is also the space of innovation that social work brings to the concept of social justice. In addition, the social work profession has worked to refine definitions of social justice so that they can be operationalized.

Social justice is understood to be a realistic choice for individuals and societies. Social justice, in this construction, is not abstract, but is instead achievable through the small and large decisions individuals, institutions, and societies make. The injustices of society are constituted and maintained by people, and so can be challenged and disrupted. Gil's (2008) conceptualization of social justice includes three dimensions: individual human relations, values and social institutions, and global human relations. This last aspect of social justice has been vitally important in recent decades, as the profession has had to grapple with the fact that injustice within a local community can have long links to distant corporations, policies, and nations (Lundy, 2011). While globalization has resulted in the intensification of wealth in a handful of corporate conglomerates in cities such as Toronto, London, New York, and Tokyo, it has also meant austerity, job losses, and a downward shift in wages in many smaller communities around the world (Workman, 2003). As a result, the profession has had to develop new strategies for achieving social justice, with each dimension requiring its own style of intervention. At times social workers are using restorative justice circles within communities to address micro aspects of social justice (Baskin, 2011). At other times, professionals are working with institutions to develop policies and processes that address injustice or aligning with social movements to address global inequality. In this way, social justice is conceptualized as a multidimensional concept that includes concern for the distribution of resources (distributive justice), the individual or community's right to self-determination, the promotion of welfare or well-being, and substantive equality (McLaughlin, 2011). By combining all these elements into a single understanding of social justice, social work draws upon both class-based conceptualizations of inequality and those that have emerged in more recent decades through identity-based social movements (Lundy, 2004; Mullally, 2006). This conceptualization of social justice is one that is flexible, relevant to work with individuals and communities, and achievable. Overall, it remains important to define social justice in a way that all social workers, whether working in therapeutic settings with individuals, as gatekeepers

of state resources, or engaged in social action can see resonating with their practice.

Operationalizing Social Justice: Organizing for Change

When social workers engage in social action to create social change, they do so mainly by organizing community groups. The change that social workers attempt to achieve by using community organizing methods is multifaceted and includes facilitating and organizing communities to meet their own needs more effectively and to make demands on the state or private sector to redistribute resources. Social work organizing practices are shaped by ideas from organizers such as Paulo Freire and Saul Alinsky, whose writings have helped conceptualize the principles of and approaches to organizing for social change, and by contemporary social movements including the feminist, disability, and queer movements.

Paulo Freire's work on popular education has influenced the role that social workers take when organizing for change. In his seminal publication, *Pedagogy of the Oppressed*, Freire (1972) conceptualized the role of a popular educator as one who works with communities to help them develop a critical consciousness of what is happening in their world and how this shapes their daily existence. The practice of popular education is useful for social work because it outlines a process of education that analyses ideological and structural power relations. A popular educator is one who works to establish a relationship with communities based on values of critical compassion and a strong commitment to social transformation. It is a role many social workers take on when working with members of marginalized communities. Popular educators are seen as colearners in the communities with whom they work. This notion is congruent with the practices of the settlement house movement where members of communities were encouraged to learn from one another. The settlement house movement began with the London-based Toynbee Hall, which was founded on the radical idea that academics, activists, and people living in poverty could live and work together to create practical and political strategies to address poverty. Those working in the settlement house movement not only pushed the state to respond to human needs, but also offered services and programming to fill in where the state could not or did not (Bradley, 2009). Freire's notion of praxis provides an opportunity for social workers to develop their reflective skills when working with communities. All effective action must be preceded and followed by a period of reflection that then informs future actions. This phase of reflection includes a focus not only on the pragmatics of any approach used, but also on individual and group emotions. Freire (1990) saw social workers' use of personal stories to inform political action and the profession's ongoing reflection on relations of power, privilege, and empowerment as integral to social transformation. Freirian writing has enriched the field of community organizing within social work so that the core of social action is grounded in the personal stories of disenfranchised people. Individual stories become the collective narratives that express the hopes and fears, needs and strengths that are the basis of social work for change.

Freire's writing also influences social work's awareness of the ways in which poverty and powerlessness erode hope, resulting in silence and disengagement. Within social work, social action is used as a means of reengaging people to create change in their lived environment. This, in turn, rebuilds hope. Freirian approaches to social action have created space to link spiritual dimensions of human experience with revolutionary change (Hope and Timmel, 1984). While social work's notion of social change is purely a secular one, many social workers are motivated toward the profession because of spiritual commitments and many communities rely on their spiritual practices as a source of hopefulness and strength to sustain them during difficult times and through the change process. Freire articulates a way to imagine processes for social change that do not rely on the influence of institutional religion, but can draw upon radical spirituality as a means to foster hopefulness.

In addition to Paulo Freire's writing, social work's approach to organizing for change has been influenced by the writing of Saul Alinsky, who articulated a clear process for social action involving confrontation with power holders. While Alinsky worked in the 1940s to 1960s, many activists still use his strategies to change neighborhoods in the United States, United Kingdom, Australia, and Korea. His approach to social action offers an alternative for social workers who are predominantly schooled in practices of consensus building. While skills for achieving consensus are helpful when organizing, social workers who are engaging in social action often find themselves in a more confrontational relationship with the state and companies that fail to fulfill their obligations to marginalized citizens. Alinsky perceived organizing as a battle between the haves and the have nots and felt that one fought this battle by any means necessary and with little concern for ethics and morals. His writing offers concrete tactics for social workers to use when having to confront opposition. Alisnky's approach to organizing is quite male-centered, focused on formal justice. Since his death, feminists and others have built upon his model, integrating his approach with a more women-centered focus on relationships and the politics and practices of caring (see Stall and Stoeker, 1998).

Like Freire, Alinsky firmly believed that social action had to be grounded in the experience of the people with whom one is working. His approach involved uniting citizens around their own self-interest and protesting in ways that were outside of the usual means through which people express dissent. In particular, he used large-scale, creative techniques that stayed close to his member's experiences, but grabbed media attention and threw his opposition off balance (Alinsky, 1971). Alinsky understood social action in pragmatic terms. He felt that social action needed to target the causes of people's oppression, not the symptoms, and he outlined a process of agitating for change, antagonizing power holders, educating the public, and organizing people. He believed that in order for communities to be able to sustain themselves through the social change process, they needed to be highly structured, with clear roles and responsibilities. One of Alinsky's most underrated yet significant contributions to social work's notion of organizing for change is that he saw the role of the community organizer as a professional activity. By reworking community organizing in this way, he offers an opportunity to imagine professionalism as including a commitment to highly politicized social action.

Social movements have also shaped social work's approach to organizing for change. From feminist movements, social workers have learned to work in a process-oriented way to achieve their goals and to use consensus-based processes. In addition, the feminist movement's view of power as limitless and the strategies used to link people's private experiences to public action have been instrumental in social work's approach to social change in impoverished communities. Finally, social work has drawn extensively on the writing and practice of feminist organizers who have struggled with embracing diversity within the movement (see, Naples, 1997). This component of the movement has challenged social work to create inclusive ways of organizing diverse groups. Practices such as creating caucuses have been integrated into mainstream approaches to organizing in social work.

The consumer mental health movements, antiracist, and disability movements have also challenged the way the profession organizes. And finally, the lesbian, gay, bisexual, transgendered, and queer movements and queer literature have had their own impacts, challenging the profession to consider how our labeling of communities, even with terms then used to mobilize public support and create services, does little to create revolutionary change. These movements have challenged the profession to move away from identity labels, to work with groups and awareness about how social discourses shape the stories communities and individuals tell one another, and to reflect upon how often the social work vision of marginalized communities is one that furthers the profession's agenda rather than their experience (Willis, 2007; Siddiqui, 2011). For example, queer theorists have argued that social work's conception of lesbian, gay, bisexual, transgendered, two-spirited, and intersex youth as struggling and suicidal is used to justify the profession's interventions in their lives, whereas the movement of school-based gay-straight alliance clubs is a community-based initiative built on the vibrant strengths of the community, which does not require professional interventions (Willis, 2007). These are ongoing reminders to the profession that social workers' first job should be to support communities in achieving their own vision.

The influence of key social movements and social activists has helped to refine the way social workers conceptualize social action and organize for change. In particular, social workers organizing for change see their work as professional practice by reimagining professionalism as a role of colearning where professionals bring particular understandings of power and the change process to engage in a process in which the stories of people who are oppressed inform their working relationship and social justice goals that are imagined (Ledwith, 2011). This professionalism is also deeply rooted in the ethical responsibility to align with the most marginalized in society and seek social justice for these groups. Through a process that is grounded in individual storytelling and a critique of the root causes of oppression, social workers organizing for social change continue to focus on the profession's dual mandate of attending to the individual and the structures of oppression that shape their lives.

Contemporary Challenges and Conclusion

While the work of organizing for change remains vital to progressive social work's vision for social justice, contemporary challenges and cautions need to be addressed. Western neoliberal governments have dramatically shifted the responsibility of services and programs from the state to the community, which reflects a significant shift away from the welfare state and the idea that the state has a direct responsibility for the well-being of its most marginal citizens (Kelly and Caputo, 2011).

This shift to community has had a number of worrying effects. First, communities have become known entities in which members are increasingly mapped, classified, and surveilled for the good of a wider community. Anyone can now find out a community's median level of income, the ratings of local schools, the racial and linguistic diversity within a geographic area, the average rate of rental accommodations, and crime statistics. These statistics help to define and reify a community, which, in combination with a 24-h news cycle and instant media reports, means that disadvantaged communities can quickly become stereotyped and stigmatized. In addition, individuals and groups of citizens have had to become more responsible for their own well-being while the state is less responsible. While such a shift aligns well with contemporary language of independence and self-sufficiency, what is obscured here is that all communities require state support to function well and that there are advantages for all when the state takes collective responsibility for services such as clean water provision. The final concern, and possibly the most pressing for social work, is that downloading this responsibility to communities has resulted in community-based agency staff becoming spokespeople for those communities, obscuring the organizations' own interests and agendas and prioritizing those over the community's needs. Community interventions are highly desirable in neoliberal contexts; they are efficient, cost-effective, and yet deny the reality that many local neighborhoods no longer have the capacity and resources to effectively address the problems that they are facing, which are often the result of extralocal forces. As a result, the state increasingly controls the local without taking responsibility for it (Kelly and Caputo, 2011). Social workers can play a key role in this process by working with communities to assess the positive and negative consequences of various community funding opportunities and programs.

The notion of community that is mobilized to affect the neoliberal agenda is a nostalgic one, where communities are imagined as wholly good spaces defined by informal care, dignity, and respect (Joseph, 2002). This is set in opposition to the supposedly dehumanizing and depersonalized state-delivered services that are managed through bureaucracies and institutions. By reifying community in this way, the state is able to pull out of more and more obligations to its citizens. Its place is being taken by an increasing privatization of services that are often marketed in the long lost language of community (see Dominelli, 2004). For instance, privatized services for seniors are marketed using the terminology of homefulness and as retirement or care communities. These nostalgic renderings fail to account for the reality that the local community has always been a space of exclusion and rigorous social control just as much as it has been a space of support and inclusivity (see Dominelli, 2006). Communities tend to take care of those who fit within their norms, and they also struggle with and exclude those who transgress the norms or

expectations of the group. The secondary problem with this nostalgic rendering is that it comes from communities with flourishing agricultural or manufacturing bases from which people were able to meet their basic needs. This is not the reality of many marginalized communities today (Kelly and Caputo, 2011). Contemporary global exploitation has resulted in massive shifts in production and left many communities unable to access the jobs necessary to sustain themselves. Thus, while organizing for change often relies on intervening at the level of community and this holds a great deal of potential for achieving some measure of social justice, it also has its dangers and cautions as this approach to social change can in fact work to sustain broad relations of inequality by distracting people from the root causes of their economic marginalization.

These tensions create a significant amount of pressure on social workers who are organizing communities. That they must perform a careful dance, balancing the basic needs of the community and engaging in the social action necessary to link their local work to broader social justice movements, and social change because a collective intervention is not, in and of itself, necessarily progressive. For social workers' interventions in the community to be oriented toward social justice requires community workers to develop deep alliances with social movements that have a greater distance and independence from the state. The flexibility and critiques of social movements help to challenge the profession continually to reflect and rework its approach to social change to ensure that social workers are achieving social justice. For example, those involved in disability movements, movements of people living with HIV/AIDS, and mental health service users/survivors movements have challenged the profession critically to reflect on how we perceive, involve, and respond to people living with disabilities, people living with HIV/AIDS, and people who use mental health services (Beresford, 2000). Feminists also did this in the 1960s and following their critique they set up services run for and by women (see Dominelli and McLeod, 1982).

There is significant concern that the place of social justice in contemporary social work has been marginalized through new managerial processes and politics. While this has produced a significant strain on the profession that has limited the profession's capacity for social action, social justice seems to remain a priority. The opportunities for social workers to effect change at the broader political and structural levels have been reduced substantially, but practitioners are still operationalizing these issues in their daily practice (O'Brien, 2010). The contemporary struggle of many in the profession is in how to link these daily practices to social action. For those organizing communities for change, the struggle is how to sustain these interventions in the direction of social justice and not to facilitate the state's ongoing negation of its responsibility to care to disenfranchised members of our societies.

Throughout the vast literature on social justice and social work, there is a clearly articulated vision for the future of social work, one that is deeply aligned with the anticapitalist and global justice movements of our time and, through such alliances sustains a radical edge pushing for social change (see Ferguson, 2008). It seems likely that such a radical frame might help to protect against some of the dangerous co-optations of community-based practice that social work has seen happening under neoliberal regimes and to sustain the professional commitment to social action, social change, and social justice.

See also: Antisocial Behavior; Feminist Research and Social Work; Social Movements: A Social Psychological Perspective; Social Work Theory.

Bibliography

Alinsky, S., 1971. Rules for Radicals: A Pragmatic Primer for Realistic Radicals. Random House, New York.

Aronson, J., Hemingway, D., 2011. "Competence" in neoliberal times: defining the future of social work. Canadian Social Work Review 28 (2), 281–286.

Baskin, C., 2011. Strong helpers' Teachings: The Value of Indigenous Knowledge in the Helping Professions. Canadian Scholars Press Inc., Toronto.

Beresford, P., 2000. Service users' knowledges and social work theory: conflict or collaboration. British Journal of Social Work 30 (4), 489–503.

Bradley, K., 2009. Poverty, Philanthropy and the State: Charities and the Working Classes in London. Manchester University Press, Manchester.

Breines, W., 1989. Community and Organization in the New Left, 1962–1968: The Great Refusal. Rutgers University Press, New Brunswick, NJ.

Clment, G., 1996. Care, Autonomy, and Justice: Feminism and the Ethic of Care. Westview Press, Boulder.

Coates, J., 2003. Ecology and Social Work: Towards a New Paradigm. Fernwood, Halifax.

Craig, G., 2002. Poverty, social work and social justice. British Journal of Social Work 32 (2), 669–682.

Davies, M., 1981/1994. The Essential Social Worker. Ashgate, Surrey.

Dominelli, L., 2004. Social Work: Theory and Practice in a Changing Profession. Polity Press, Cambridge.

Dominelli, L., 2006. Women and Community Action, second ed. Policy Press, Bristol.

Dominelli, L., 2009. Introducing Social Work. Polity Press, Cambridge.

Dominelli, L., 2012. Green Social Work. Polity Press, London.

Dominelli, L., McLeod, E., 1982. Feminist Social Work. MacMillan, London.

Ferguson, I., 2008. Reclaiming Social Work: Challenging Neo-liberalism and Promoting Social Justice. Sage, London.

Figueira-McDonough, J., 2007. The Welfare State and Social Work: Pursuing Social Justice. Sage, Thousand Oaks, CA.

Finn, J., Jacobson, M., 2008. Just Practice: A Social Justice Approach to Social Work, second ed. Eddie Bowers Publishing Co., Inc., Peosta, IA.

Fook, J., 1993. Radical Casework: A Theory of Practice. Allen and Unwin, St. Leonards.

Freire, P., 1972. Pedagogy of the Oppressed. Continuum, New York.

Freire, P., 1990. A critical understanding of social work (M. Moch, Trans.). Journal of Progressive Human Services 1 (1), 3–10.

Gil, D., 2008. Meeting universal human needs as the foundation for individual and social development and of social and global justice: comments upon receipt of the Justice Studies Association's 2008 Noam Chomsky Award. Contemporary Justice Review 11 (4), 323–330.

Gilligan, C., 1982. In a Different Voice. Harvard University Press, Cambridge.

Hoefer, R., 2012. Advocacy Practice for Social Justice, second ed. Lyceum Books, Chicago.

Hope, A., Timmel, S., 1984. Training for Transformation: A Handbook for Community Works (Books 1–3). ITDG Publishing, London.

Joseph, M., 2002. Against the Romance of Community. University of Minnesota Press, Minneapolis and London.

Kelly, K., Caputo, T., 2011. Community: A Contemporary Analysis of Policies, Programs and Practices. University of Toronto Press, Toronto.

Ledwith, M., 2011. Community Development: A Critical Approach, second ed. The Polity Press, Bristol.

Lee, P., 2009. The social worker and social action: address delivered before the New York State Conference on Social Work, 1935. Journal of Community Practice 17 (3), 327–333.

Lee, B., 2011. Pragmatics of Community Organizing, fourth ed. Common Act Press, Don Mills.

Locke, J., 1988. In: Laslett, Peter (Ed.), Two Treatise of Government. Cambridge University Press, Cambridge.

Lundy, C., 2004. Social Work and Social Justice: A Structural Approach to Practice. Broadview Press, Peterborough.
Lundy, C., 2011. Social Work, Social Justice and Human Rights: A Structural Approach to Practice, second ed. University of Toronto Press, Toronto.
Margolin, L., 1997. Under the Cover of Kindness: The Invention of Social Work. University of Virginia Press, Charlottesville and London.
McLaughlin, A.M., 2011. Exploring social justice for clinical social work practice. Smith College Studies in Social Work 81 (2–3), 234–251.
Moreau, M., 1979. A structural approach to social work practice. Canadian Journal of Social Work Education 5 (1), 78–94.
Mullally, B., 2006. The New Structural Social Work, third ed. Oxford University Press, Don Mills.
Naples, N., 1997. Community Activism and Feminist Politics. Routledge, New York.
Nietzsche, F., 2008. In: Kaufmann, Walker (Ed.), The Portable Nietzsche. Penguin Books, London.
Olson, J., 2007. Social work's social justice and professional projects: discourses in conflict. Journal of Progressive Human Services 18 (1), 45–69.
O'Brien, M., 2010. Social justice: alive and well (partly) in social work practice. International Social Work 54 (2), 174–190.
Rawls, J., 1999. A Theory of Justice. Oxford University Press, Oxford.
Reamer, F.G., 1994. The Foundation of Social Work Knowledge. Columbia University Press, New York, NY.
Reeser, L.C., Epstein, I., 1990. Professionalization and Activism in Social Work: The Sixties, the Eighties and the Future. Columbia University Press, New York.
Sen, A., 2011. The Idea of Justice. Belknap Press of Harvard University, Cambridge.
Siddiqui, S., 2011. Critical social work with mixed-race individuals: implications for anti-racist and anti-oppressive practice. Canadian Social Work Review 28 (2), 255–272.
Solas, J., 2008. Social justice and social work: what are we fighting for? Australian Social Work 61 (2), 124–136.
Specht, H., Courtney, M., 1994. Unfaithful Angels: How Social Work Has Abandoned its Mission. Free Press, New York.
Stall, S., Stoeker, R., 1998. Community organizing or organizing community? Gender and the crafts of empowerment. Gender and Society 12 (6), 729–756.
Stebner, E., 1997. The Women of Hull House: A Study in Spirituality, Vocation and Friendship. State University of New York Press, New York.
Thompson, N., 2002. Social movements, social justice and social work. British Journal of Social Work 32 (6), 711–722.
Willis, P., 2007. 'Queer eye' for social work: rethinking pedagogy and practice with same sex attracted young people. Australian Social Work 60 (2), 181–196.
Workman, T., 2003. Social Torment: Atlantic Canada in the New World Order. Fernwood Publishing, Halifax.

Relevant Websites

www.fncaringsociety.com – First Nations Child and Family Caring Society (Canada).
www.iassw-aiets.org – International Association of Schools of Social Work.
www.ifsw.org – International Federation of Social Workers.
http://www.manavsadhna.org/ – Manav Sadhna (India).

Social and Behavioral Sciences under Dictatorship

Olessia Kirtchik, National Research University–Higher School of Economics, Moscow, Russia
Mariana Heredia, The National Scientific and Technical Research Council/UBA, Buenos Aires, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Various forms of dictatorship have been a context in which social and behavioral sciences (SBS) have been developing through most of the twentieth century. Nazi and fascist regimes in Europe, Communist single-party states, military juntas in Latin America and elsewhere in the postcolonial world accompanied the crisis of tradition and development of modernity as an alternative to liberal democracy. Dictatorships have thoroughly affected the history of SBS pursuing a policy of repression and control and, sometimes, encouraging a growth of various social science disciplines. The lack of intellectual and institutional *autonomy* is generally endured, though to different degrees and in different aspects, by SBS under dictatorship.

Dictatorship, Social Sciences, and Modernity

The relatively short history of social sciences has been heavily impregnated by uneasy yet pervasive relations with various forms of dictatorship. Dictatorial (i.e., nondemocratic) forms of government have dominated in different regions of the world through most of the twentieth century. According to the Human Security Report (2011), in 1946 only 20 countries (28% of the world's governments) can be counted as democratic, while by the end of the century their number goes up to 79, and continues to grow ever since (by 2008 that share has more than doubled). Dictatorships accompanied the crisis of tradition and development of modernity as an alternative to liberal democracy. The latter, established as a normative model of the modern government, has failed to stabilize itself, with a few notable exceptions, in the biggest part of the world before the end of the twentieth century (O'Donnell, 1972; Brooker, 1995; Kim et al., 2013). Consequently, it should be stressed that nondemocratic forms of government are not an exception, a deviation, or a hazard, as they are often perceived, but a structural part of modernity providing a most common context for the development of social sciences in both negative and positive senses.

In the years following the World War I, and especially during the Great Depression, the first wave of modern dictatorships – of Fascist or para-Fascist nature sharing antiliberal, anticommunist, and nationalist values – touched, for a more or less long period of time, many of the European countries. Nazi Germany's Third Reich and the loyal governments fell after the defeat in World War II, while Communist regimes were spreading in different parts of the world. For the first time, 'dictatorship of proletariat' emanated from the Russian Revolution and was transformed into the Stalinist regime of personal power in the late 1920s. Following World War II, this type of political regime was consolidated in a large part of the Central and Eastern Europe forming the Soviet bloc for nearly half a century. Communist dictatorships also proliferated, in the after-war period, in Southeast Asia (following the example of China), Africa, and some other parts of the Third World. Finally, numerous military dictatorships, stable or volatile, were established through the *coups d'Etat* in Latin America, and elsewhere in the new postcolonial societies.

Modern dictatorships have taken very different forms and have had various ideological inspirations (if any), so that a unique definition or a set of common features might hardly be formulated (Ezrow and Frantz, 2011). Some of them might have possessed at least some of the formal institutions of representative democracy (parliament, elections, etc.), been more or less repressive, to some extent respected the rule of law, and gained a popular support. However, a dictatorship can negatively be defined as a form of government that excludes political pluralism, competition, and prevents citizens from participation in decision-making. Power is often concentrated in the hands of an individual, a group of individuals (i.e., a *junta*), or a single party (merged with the State apparatus). Along with other recurrent features (cult of personality, abuse of individual freedoms, corruption, etc.), a political monopoly constitutes an essential difference between modern dictatorships and liberal democracies. To the extent to which a dictatorship tends to suppress any opposition, any critic, any public deliberation, it gives ground to a largely accepted opinion according to which social sciences can only exist in a democratic society.

Indeed, specific features of dictatorships are far from being neutral for the development of social sciences. The establishment of any dictatorial regime, accompanied by the elimination of political opposition, severely affects the situation of social sciences. Being particularly sensitive to political issues, social sciences undergo a profound restructuration (or even suppression) during the first years of dictatorship. Ideological campaigns and purges, repressions (expulsions and, in extreme cases, massive imprisonments and executions) are a widespread phenomenon. These campaigns might provoke massive exoduses of social thinkers and even entire schools. Karl Mannheim, Paul Lazarsfeld, Ludwig von Mises or Wilhelm Röpke, and the Ordoliberal school, exemplify the emigration of scholars (often of Jewish origin) from Germany and Austria to the United States and other countries. Similar movements, from Eastern Europe, included sociologists, economists, philosophers, and other social scientists opposed to the formation of proletarian dictatorships. These exoduses often engaged major figures as Mircea Eliade (from Romania), Tzvetan Todorov (from Bulgaria), or Pitirim Sorokin (from the Soviet Union), and considerably weakened national schools of social

sciences. Famous 'philosophers' ships' transported thinkers and scientists expelled from the Soviet Union in 1922. Those who stayed were often subjects to 'inner emigration,' dismissal from academic positions, or different forms of compliance with the officials.

The periods of intense repression and unification are normally followed by periods of a relative liberalization concomitant with stabilization of new institutional structures and routines. This 'normalization' can imply certain pluralism and autonomous thinking in the field of social and behavioral sciences (SBS). Nonetheless, unlike in democratic regimes, they can only achieve a reduced autonomy from political power. Public research and education institutions are systematically put under surveillance, and intellectual production is strictly controlled. Furthermore, social scientists cannot overtly discuss their ideas, and their production is mostly isolated from the public space and mass media. In other words, under dictatorship, social sciences can only be legitimately mobilized as specialized knowledge aimed at informing (and legitimating) decision-making, and not as an instance of independent thought able to criticize and denounce power. In the case of nationalist dictatorships, the symbolic function of SBS consists essentially to praise and to defend a national culture against international influences or critics.

The ambition of dictatorships to unify and control public discourses, which prevents social sciences from emancipation, might vary in degree according to the ideological nature and severity of a regime. The most important, and relevant for this discussion, distinction is the one that opposes totalitarian and authoritarian regimes (Linz, 1975). The former often follows a single-party state model, based on an extended social mobilization and a coherent official ideology. It tends to control every aspect of life of its citizens, and to exclude all differing opinions and lifestyles. Hitlerism and Stalinism represent two major figures of historical totalitarianism (Arendt, 1951). The authoritarianism, in turn, is less radical and comprehensive, and can be applied to any political system, ideologically ambiguous, but averse to competitive elections and republican procedures (Linz and Stepan, 1996). In spite of their great variety, due to their relative tolerance to public debate and political opposition as well as the absence of systematic intrusion in individual privacy, most Latin American dictatorships, and arguably some post-Stalinist regimes in Central and Eastern Europe, might be classified as authoritarian (O'Donnell et al., [1986]1991).

As ideal types, these two categories are landmarks in a continuum from the most top-down, monopolistic, and rigid system into different degrees of liberalization and tolerance. They can be also applied to describe different stages of the same regime. At the beginning dictatorships tend to be more rigid and coherent than in later stages. The distinction between totalitarianism and authoritarianism is crucial in a sense that while the former denies any independent intellectual production outside the State and the official ideology, the latter accepts a larger autonomy for social organizations and certain tolerance to opposition. As Stoppino (1983, p. 132) emphasizes, in authoritarian regimes, civil society corporations (as trade unions, pressure groups, lobbyists, or religious associations) persist and the independence of educational system and media is respected to a certain degree. Even political opposition can be tolerated if it is not in open competition with central powers.

Nonetheless, one should emphasize, once more, the variability of historical figures of dictatorships in their relations to social sciences. Differences in the situation of social sciences are related to historical contexts presenting various linguistic, politicoeconomic, and religious constellations. For instance, the Portugal (Salazar) or Spanish (Franco) dictatorships of conservative nature implied an alliance with Catholic Church and a mistrust of social sciences associated with modernist values (Giner, 1990). In turn, Nazi and Stalinist regimes were radically antiliberal and anticonservative in their ambition to promote a truly scientific government of society in line with the party ideology. Contrary to Communist states, later Latin American military regimes were market-friendly and welcomed international cooperation with (capitalist) countries, including external financial support for SBS development. The resulting patterns of institutionalization of social sciences under dictatorships are quite different. In some cases, authoritarian regimes give an impulse to institutional growth and internationalization of social sciences and, in many others, hinder these developments. The existence of developed national traditions of social sciences *before* the establishment of a dictatorial regime is also an important variable.

In what follows, two relevant, yet in many respects opposing, cases are considered: European Communist single-party states and Latin American military dictatorships. These cases are exemplary in a sense that, first, they possess established traditions of different fields of SBS (which have existed before, during, and after dictatorial periods), and, second, dictatorships were not just short episodes of the history of these regions, but were in place for many years or even decades, so that they have left a deep mark on the trajectories of social sciences. Each of these regional cases is characterized by a geographical and temporal unity, as well as politicoeconomical proximity. Common patterns, as well as variations within these cases, will be considered.

European Communist Single-Party States

Scholarship on the SBS under dictatorship in the USSR and Socialist bloc, but also in Germany and other European countries (Italy, Portugal, Spain, and others), is quite scarce. Dictatorship episodes are mostly omitted as obscure and unproductive for social sciences. External observers often depict Socialist sociology as being simply Marxist–Leninist orthodoxy (Keen and Mucha, 1994, p. 1). Scholars with experience of Socialist academia, on the other hand, sometimes lack a critical distance. Balanced accounts considering the sociological contexts of social science histories are rare. Although monographic studies of national disciplinary histories are accessible, attempts at comparative analysis are still in shortage (see, e.g., Koleva, 2002; Voříšek, 2008).

In the history of Socialist SBS, two phases can clearly be distinguished: strong dictatorship (Stalinism), established in the Soviet Union by the end of the 1920s and later imposed to new Socialist states after World War II, and the post-Stalinist period beginning with the Khrushchev 'Thaw' in 1956.

Stalinist Model

In the Soviet Union, non-Marxist approaches in SBS are to be banned from the early 1920s, but a relative pluralism of opinions within this paradigm is tolerated until the end of the decade. The 'Cultural Revolution' imposed by Stalin radically redefines the intellectual and institutional landscape of social sciences. This totalitarian move implies a complete cognitive and institutional unification under the lead of historical materialism. The conformity to Marxist–Leninist doctrine and class principle becomes the main criteria of scientific truth and objectivity. Sociology, psychoanalysis, but also genetics, and some other disciplines, are labeled 'false bourgeois sciences.' As a consequence, their development is brutally stopped in the Soviet Union in the 1930s for a few decades. Economic sciences, which have flourished during the 1920s, are not completely eclipsed, yet they undergo a rupture: the most visible schools and domains of economic research are suppressed. Numerous scholars engaged in these fields (including the prominent economists Nikolai Kondratiev and Alexander Chayanov) are arrested and later executed, many others are sent to Gulag or suspended from teaching and doing research. A new science, political economy of socialism, is painfully and slowly elaborated under the close scrutiny by Stalin himself. Nonetheless, some other branches of SBS scholarship (i.e., the Marxist history of science, in particular the work of Boris Hessen; the activity theory of Alexei N. Leontiev and cultural–historical theory of Lev Vygotsky in psychology) were developing during this period and even attracted the attention of Western colleagues.

This totalitarian model of social science resembles relations between political regime and academia in European Fascist and semi-Fascist states. During the period of dictatorship, the undergraduate teaching of sociology is forbidden in Portugal (Torres, 2010), and its development is virtually eclipsed in Spain from 1939 to 1950s (Giner, 1990). In Germany, sociology is not forbidden, but it has to conform itself to *Völkisch* ideology of the Third Reich: Nazi sociology is promoted against any other conception of social science (Western liberal or Marxist). Seemingly to the Stalinist class (Party) principle, it is the racial or ideological criteria that serve to distinguish between true and false: e.g., Durkheimian sociology is stigmatized as a 'Jewish,' and thus erroneous, social theory, while Max Weber is criticized for his 'liberal' orientations (Klingemann, 1996, p. 55). Methodological nationalism and intellectual autarchy are cultivated in social sciences aimed to contribute to the ideological legitimation and building of a new society. Western social theory is under suspicion in the Soviet Union and it is studied under the label of 'critique of bourgeois theories.' Most of the Western scholarship is classified, and can only be accessed, in the libraries, through a special authorization.

From 1948, the Stalinist model of social science is imposed to most newly established Socialist societies in the Central and Eastern Europe (except for Yugoslavia where Tito refuses to be subordinated to the Soviet Communist party but nevertheless monopolized academia). Social sciences are rebuilt to conform the theoretical framework of historical materialism (part of the Marxism–Leninism). In some cases, especially in Romania, purges take place against 'fascist' social scientists. 'Bourgeois sciences,' especially, sociology ceases to exist institutionally in almost all Eastern Europe at least until the end of the Stalinist period. 'Soviet institutes' are open to propagate the Marxist–Leninist social science. With this goal, Soviet scholarship (textbooks and collections of texts in the field of historical materialism and scientific communism) is translated into most East European languages (Keen and Mucha, 1994, p. 7). 'Fundamentals of Marxism' (Yugoslavia) or 'scientific communism' (USSR) aimed at strengthening ideological integrity, become compulsory subjects in all undergraduate curricula in the Socialist bloc.

In the Soviet Union, as well as in satellite countries, research in social sciences is mainly carried out within the Academy of Sciences, a centrally managed and publicly funded network of research and auxiliary institutions (Graham, 1967). Besides, a network of Party research institutions and schools destined to form political cadres is established with, at the head of the hierarchy, the institute of Marxism–Leninism depending on the Central Committee of the Communist Party. In this highly centralized and hierarchical system, independent or private institutions do not exist, and the publicly funded research and higher learning institutions are directly subordinated to the central Party organs and administrative bodies. The political influence is exercised through laws and regulations, and research planning. Throughout the Socialist period, very close ties, a sort of symbiosis, between party–state bureaucracy and academia are maintained, with the leading role of the Party (Krementsov, 1997, p. 4).

Post-Stalinist Social Sciences

In the years following the beginning of the Khrushchev 'Thaw,' i.e., from the end of the 1950s, social sciences undergo a process of an extensive institutional (re)building in the Soviet Union and satellite countries. Disciplinary infrastructures comprising research institutions, university departments, professional associations and reviews are created on the initiative or with approval of the central Party organs. In particular, sociology (re)appears as an academic discipline, though with a different pace and scale, in all Socialist bloc: first in Poland, Yugoslavia, and Czechoslovakia; and later, during the 1960s and the 1970s, in the Soviet Union, Bulgaria, the German Democratic Republic (GDR), Romania, and Hungary (Keen and Mucha, 1994; Voříšek, 2008). But it is significant that full sociological curricula and university grades in sociology appear quite late, if not only after the fall of dictatorships. As in the West, new social sciences are institutionalized: economic cybernetics (as a Socialist counterpart of mathematical economics), scientific forecasting of economic and technological development (while the Western-style futurism is banned), science studies, area studies (American, African, etc.), and so on.

The rationale behind these efforts is primarily the competition on the ideological front of the Cold War: to increase the presence of Socialist scientists and to disseminate the achievements of Socialist regimes on the international scene. International congresses abroad and inside the Soviet bloc serve this goal. Ideologically viable participants are carefully selected to represent Socialist countries in these meetings. Development of the Third World (from India to Africa) is a particularly important stake in the competition between the two super powers. Experts' activities funded by the State or private

foundations (in the American case), as well as student scholarships and exchanges, are promoted on both sides of the Iron Curtain. Second, social sciences come to be viewed as an important factor of the socioeconomic development: Socialist governments formulate a need for scientific expertise in social and economic matters. Hence, functions of social sciences, economy, and sociology in the first place, are conceived as to study complex social phenomena in order to deliver a 'scientific basis' to the construction of socialism. Various techniques of optimal planning, systems analysis and mathematical modeling of social and economic processes are promoted in order to satisfy this demand. According to this orientation, social sciences are mostly limited to empirical and applied research, which might have immediate utility. Nonetheless, facts of 'reality' established through the empirical research often contradict the official orthodoxy and present a source for dissidence (Levada, 1969; Konrad and Szelenyi, 1979) which is systematically suppressed. Large-scale sociological surveys (e.g., on living conditions of the working class) and opinion polls represent a shift in knowledge production on Communist societies.

A relative political liberalization of the 'Thaw' and reformist movements of the 1960s enlarge the scope of academic discussions, and increase the participation of social scientists in public affairs and policy-making. Reformist governments seeking to modernize the economy engage various social scientists, especially economists, as scientific advisors and launch public discussions on reforms in the media. Nonetheless the degree of academic freedom and openness varies considerably from country to country. Marxist revisionist currents appear in Poland (the so-called 'Warsaw school of history of ideas'), Hungary (Lukacsian sociologists), Yugoslavia (*Praxis* group), etc. Original theoretical developments are to be found under less constraining regimes or on margins of centralized academic systems. As examples we can cite, among others, works by the Estonian cultural historian Mikhail Lotman and the Tartu–Moscow Semiotic School, Polish sociologists led by Stanisław Ossowski.

A special status of economists is to be stressed: they are not only the largest discipline in terms of university chairs, research institutions, and staff, they are often given a greater intellectual autonomy, due to the crucial role of economic performance in the competition of Communist and capitalist systems. Economists grow to political influence in their respective countries, in particular, during the economic reforms of the mid-1960s (Wagener, 1997). Works by East European economists (especially Hungarian and Polish) on market socialism find an echo among Western neoclassical scholars (Bockman, 2011). Along with the Soviet school of optimal planning (led by the 1975 Economics Nobel Prize winner Leonid W. Kantorovitch), substantial contributions are made, just to mention a few, by Oscar Lange, Michal Kalecki, and Włodzimierz Brus (Poland); János Kornai and Peter Thomas Bauer (Hungary); Branko Horvat and Jaroslav Vanek (Yugoslavia); Ota Šik (Czechoslovakia).

During the 'Thaw,' international contacts and cooperation across the Iron Curtain are intensified, especially in strategic domains (applied mathematics, space, computer science, economic development, and so on). Groups of American and European specialists, as well as individual visitors, are admitted to the Soviet Union and other countries of the Socialist bloc from the late 1950s. Western scholarship becomes partly available through scientific journals and translations of seminal works (e.g., the Russian translation of "Economics" by Paul Samuelson is published in 1964), but it has a restricted circulation. The international mobility of Soviet and other Eastern European scholars is also limited and is mostly open to Party members and officials. Contacts and moves of foreign visitors, in turn, are strictly controlled and surveyed. In 'soft dictatorships,' like Poland or Hungary, international exchanges are much more intensive: social scientists (especially economists) are allowed to participate in international conferences abroad and to visit Western universities as invited fellows for longer stays. Student exchanges and other forms of scientific cooperation are frequent. For instance, an international research project on the impact of industrialization on everyday life carried out in mid-1960s in 10 countries, funded by United Nation Educational, Scientific, and Cultural Organization (UNESCO), is directed by the Hungarian sociologist Alexander Szalai. Economists from Hungary, Yugoslavia, and Poland are engaged by international organizations in programs of development in the Third World. In other cases, as in the USSR, Romania, or the GDR, social sciences mostly remain in isolation.

In spite of the intellectual revival and intensive institutional development during the 'Thaw,' newly established social sciences are in a precarious situation until the end of the Communist rule. With the return of a conservative trend after the Prague Spring and the Soviet invasion in 1968, the contraction of academic autonomy and repressions against ideologically dubious 'revisionist' social thinkers are renewed in many Socialist countries. Especially, sociology falls once again in disgrace (except in Hungary or Bulgaria where it develops without interruption). For instance, in the Soviet Union a newly created Institute of Concrete Sociological Studies is closed in 1972. The antirevisionist and anti-Semitic campaigns are followed by a new wave of social scientists' emigration to Western Europe and the United States. Previous developments continue to exist as a 'shadow science' in a form of *samizdat* publications and informal seminars (Keen and Mucha, 1994; Koleva, 2002).

One should wait for Perestroika and the fall of the Berlin Wall for scientific internationalization and political liberalization. In the atmosphere of 'stagnation' (the years of Brezhnev rule from 1968 until mid-1980s), the purity of ideological-theoretical basis is controlled by Party censorship. Some Western scholarship (Parsons' functionalism in sociology or some neoclassical economics) is tolerated, but only if it is reinterpreted along the 'right' ideological lines and to the extent to which it is 'useful.' In order to cope with this ideological pressure, SBS scholars develop different strategies: the use of a special language full of euphemisms, and the ability to read 'between the lines'; the predilection for mathematical formalization and 'pure theory'; the technocratic language of statistics; or a retreat to politically neutral subjects (e.g., a history of prerevolutionary doctrines). As a result of censorship and self-censorship, social sciences – caught between the theoretical dominance of Marxism–Leninism and technicalities of applied research – are hardly able to develop original theories, especially middle-range ones, of Socialist societies. This legacy still persists in social sciences even after the fall of the dictatorial regimes where the autonomy of scholars remains an open question (Bikbov and Gavrilenko, 2002–2003).

Latin American Military Dictatorships

Contrary to Eastern Europe, monographic studies on national disciplinary cases are rare and limited to bigger countries (as Argentina, Brazil, Chile, and Mexico), while comparative analyses are frequent. Several systematic comparisons have examined the autonomy and institutionalization of SBS (Brunner and Barrios, 1987; Calderón and Provoste, 1992; Sonntag, 1988; Trinidade, 2007). An overwhelming majority of comparative books on the subject are published, thanks to foreign support (Canadian International Development Agency, Ford Foundation, UNESCO, and Maison des Sciences de l'Homme).

If compared to the Socialist bloc, a greater variety of experiences in political regimes and SBS institutionalization patterns characterizes the twentieth-century Latin America. Dictatorships were frequently responsible for the SBS instable development, but not necessarily: some nondemocratic governments were SBS friendly, while some democratically elected regimes suspended their development. In this complex twentieth-century SBS history, the 1970s mark a turning point. SBS achieved a certain level of institutionalization, at the same time, many Latin American countries experienced waves of bloody repressions. Most recent dictatorships have often had deep consequences for social sciences persisting today.

Politics and SBS before the 1970s

The transition of most Latin American societies to democracy and modernity was not exempt of deep tensions, which manifested themselves in acute social conflict and political instability. Chile, Colombia, Uruguay, and Venezuela constitute the few exceptions on the continent: a relatively stable political party system, peaceful succession of governments, and opposition tolerance were achieved here long before the 1970s.

The expansion of State intervention following the global crisis of the 1930s gave a powerful impetus for the development of SBS disciplines, especially economics and demography, in México, Brazil, Argentina, Chile, Peru, Colombia. Afterward, the trajectories of social sciences were affected, though unequally, by multiple dictatorial regimes experienced by these countries. Having dissimilar political histories, Mexico and Brazil both provided a steady growth for SBS in public institutions. Mexican one-party dominant system, established after the 1910 peasant revolution, let the *Partido Revolucionario Institutional* monopolize the access to presidency for 60 years. Relatively tolerant to public opposition, Mexico was particularly SBS friendly. Continuity for SBS was also strong in Brazil, even if its political system was less stable. Open elections and civil rights were fairly respected until a new military intervention of 1964–85. In both cases, humanities and social science were associated with national identity and mainly reserved to social elite members. Both countries kept most SBS activities in public institutions and developed distinguished academic centers supported by the State as El Colegio de México or el Museu Social de Río de Janeiro.

Argentina is probably the most extreme example of tensions between democratization and order. Between 1930 and 1983, Argentina saw a succession of civilian and military governments and experienced periods of restrictions on free speech, free assembly, and participation in government. In spite of the massive enrollment of students and academic freedoms, Argentinian public universities suffered from underfunding and recurrent political purges (Buchbinder, 2005). Academic instability prevented specific and legitimate hierarchies to be established and reproduced. 1930 and 1943 dictatorships as well as Juan Domingo Perón's (1945–55) elected government faced new conflicts. Overwhelmingly supported by the masses, but close to Catholic Church and reluctant to universities' autonomy, Perón caused the dismissal and exile of numerous intellectuals and scientists. Amado Alonso (Spanish philologist, future professor, and PhD honoris causa at Harvard University), Bernardo Houssay (future Medicine Nobel Prize winner), Raúl Prebisch (the most famous Economic Commission for Latin America economist) left the country (Buchbinder, 2005, p. 149). Purges did not cease after Perón's overthrow. Those who had lost their positions in public universities came back and expelled those who had arrived later (Neiburg, 1998). Yet, the 1955–65 decade, governed alternatively by civilians and military officials, is considered to be the golden age for SBS in Argentina: new disciplines as psychology, sociology, and economics were created; many positions were opened; university budget increased; the national research council (CONICET) was founded; State agencies recruited many SBS graduates. The modernizing process was interrupted by a new military coup in 1966 that punished especially natural and social sciences. While many scientists and artists left Argentina, those who stayed in the country became closer to Peronists and leftists' parties and radicalized their positions denouncing the alliance of armed forces, social elites, and imperialism.

Despite political diversity, certain common features can be mentioned in the development of SBS in Latin America. On the one hand, as in Eastern Europe, political elites expected from SBS empirical data and useful tools to improve population and territories' management. On the other hand, this ambition was highly supported by international organizations and especially by American foundations since President Kennedy proposed the Alliance for Progress (1961) as a kind of Marshall Plan (1947) for Central and Southern America to prevent Communist expansion in the region (Berman, 1983; Smith, 1991).

Geopolitical concerns and international funding were then at the very origin of the institutionalization of social sciences, but also of their regionalization. US support to Latin American SBS was not exempt of competitors and conflicts. The Communist Party's network, the Catholic Church and European governments were also involved in Latin America to promote different styles of social thought. In 1964, the 'Operation Camelot' became a scandal in Chile, revealing that area scholars, supported by US Army, were supposed to gather information relevant to the US counterinsurgency program (Navarro, 2011). But, for many decades, US influence was indirect. Three main regional organizations, all of them financed by international funds, contributed to the expansion and regionalization of SBS: the United Nation Economic Commission for Latin America (ECLA, created in 1948, mainly composed by economists), the Latin American Faculty of Social Sciences (in Spanish FLACSO, created in 1957, with a main participation of sociologists and political scientists), and the Latin American Council of Social Sciences (in Spanish CLACSO,

created in 1967, with a similar composition). In this context, there was a competition among countries and institutions to attract international support as well as a rivalry between universities and international organizations to keep SBS under their scope (Beigel, 2013).

ECLA gave Latin American SBS a clear mission: development's promotion and technical support. ECLA's founding father, Raúl Prebish, stressed the opposition between the 'center' and the 'periphery,' focusing on the international division of labor based on degrees of industrialization. Between 1950 and 1970, many issues and analytical categories were developed in strong collaboration among Mexican, Chilean, Argentinian, Brazilian, Uruguayan scholars (Love, 1996; Sonntag, 1988) but also with American and European contributions. The structuralist inflation theory (in economic sciences), labor informal sector analyses (in sociology), or bureaucratic authoritarian regimes' conceptualization (in political sciences) are just three main examples of Latin American themes that gained international prestige.

As most of SBS formation was offered in public universities, placed in the core of main cities, open to ideological diversity and with highly politicized students, powerful students' association appeared and made of universities one of the most important social actors in the region. Contrary to technocratic organizations, universities and in particular students associations were closer to critical approaches and frequently to Marxist theories and Communist parties. Some authors conclude that vigorous commitment to social causes and personal engagements with leftist parties were sometimes a more important source of prestige for SBS professors than originality and intellectual strength. Even if students' politicization came earlier in some countries (Argentina and Peru) and later in others (Chile and Mexico), by the 1970s, youth was a crucial factor of Latin American political dynamics.

1970s' Dictatorships and Their Legacy for Latin American SBS

After the Cuban Revolution, a New Left was consolidated in the region, having strong ties with universities and SBS. The clearer expression of academic radicalization was the emergence of the 'dependency theory' that, contrary to the ideas originally promoted by ECLA, accused Western imperialism of the region's underdevelopment (Love, 1996). While SBS scholars in the 1960s invested most of their energy in creating and extending public agencies, a decade later many of them concluded that these efforts were useless without a revolutionary break of imperial ties. Universities were main participants in political mobilization and radicalization. Their students were frequently involved in the emergence of urban and rural guerrilla in Argentina, Bolivia, Brazil, Chile, Colombia, and Peru. Most 'dependentist' scholars (sociologists, economists, historians) had a political engagement in the 1960s.

Even before the military coups, SBS were already victims of violence. Urged by the imminence of revolution, teaching and research activities were subordinated to political commitment. Face to growing mobilization and radicalization of their students, many scholars left public universities behind to create private research centers. Tired of facing threatens and mistreatment, many specialists associated themselves to form a sort of cooperative in order to share a working space and fight together for foreign or private financial support. At the same time, those intellectuals that remained in public universities were often the first victims of paramilitary assassinations.

Due to SBS previous development and Pinochet's government extreme methods, Chile was the most dramatic example of the destructive effect of dictatorships. The fall of Santiago de Chile as a cultural center had deep consequences for the previous regionalization of social sciences. ECLA and FLACSO did not disappear but lost most regional importance and political influence. Due to the reception of refugees, Mexico gained a certain importance in SBS production during the late 1970s and the early 1980s. Accompanying its economic importance, Brazil strengthened its regional scientific relevance, especially in anthropology. Nevertheless, a direct link was established between Latin American scholars and American and European universities. As refugees first, young scholars started following their degrees abroad with no relation to their universities of original attachment. In economics, US dominance was therefore unquestioned. In sociology, the competition between US and European academies remain stronger even today.

Locally, Chile experienced as Argentina and Uruguay an unprecedented violence. Universities were put under surveillance and dictatorships' designated authorities concentrated most prerogatives and were entitled to purge students and professors. Many of them were kidnapped from university locals. Most were executed after 'disappearing.' A few were formally arrested and spent many years in prison or fled the country. Others were fired without explanation and excluded of any public institution of teaching and research. In Uruguay, 40% of professors lost their jobs. Many public research centers as the Uruguayan Instituto de Ciencias Sociales were closed. In many Chilean and Argentinian universities, SBS degrees in sociology, psychology, and anthropology were suppressed or no more inscriptions were accepted. In Chile and Argentina, students' enrollment dropped drastically (Brunner and Barrios, 1987, p. 45). Even libraries were purged of texts written by Marx, Freud, and Einstein as main references of ideological deviance. Less studied, Bolivia, Peru, and Paraguay suffered similar processes once 1970s dictatorships were established.

At the same time, dictatorship caused or reinforced the migration of SBS from public to private institutions (Brunner and Barrios, 1987). The most significant and generalized legacy of Latin American dictatorships was the creation of private research centers disconnected from teaching activities. Considering the importance and stability of public institutions in Brazil, private centers were fewer but produced the most innovative and prestigious research. During dictatorship, these centers exercised a very discreet activity, 'for their own.' Participants experienced these years as working in a 'catacomb university.' In fact, the separation of teaching and research was encouraged by foreign donors and by governments. On the one hand, private centers could only survive, thanks to the symbolic and material support by international organizations. On the other hand, dictatorships' attention was focused on universities and students' belligerency. Studies produced in the private centers were not particularly visible or disturbing.

This migration from public to private research centers had consequences in SBS production. General theory was abandoned in favor of more concrete and specialized studies. Certain themes related to international concerns grew while

others, focused in national reality, regressed. Policy evaluation and counseling were encouraged while descriptive and interpretative approaches declined. Critical intellectual profiles were put at a disadvantage while technocratic orientations were recompensed.

Dictatorship action was not only repressive though. Military governments had an active role in reformatting high educational system. In Chile, Pinochet policy left a persisting legacy: on the one hand, public universities declined; the cost of education was taken as a private responsibility; on the other hand, competition for funding was encouraged among institutions to improve their efficiency. In Argentina, military policy withdrew economic resources from public universities located in big cities to invest them in geographically dispersed research centers. At the same time many prestigious SBS researchers were fired from the CONICET, many right Catholic intellectuals without any academic recognition were recruited or designated by national authorities.

Mexico and Brazil were also touched by SBS political radicalization. The 1968 Tlatelolco square massacre was the most extreme expression of Mexican repression of students' protests. During Brazilian last dictatorship, many students were persecuted, imprisoned or killed. But compared to others, these repressions were pretty selective, and most SBS organization and programs not only survived but also were enlarged (Brunner and Barrios, 1987, p. 42). University autonomy was respected and professionalization did not stop. Since Santiago de Chile declined as a center of SBS excellence, Mexico and Brazil consolidated their influence by receiving first Latin American refugees, students, and scholars later.

As in Eastern Europe, various social science disciplines did not receive an equal treatment depending on the ideological bases of military governments: liberal in economics and mainly Catholic fundamentalists. Sergio Castro and Adolfo Diz were appointed economy minister by Pinochet in Chile and President of the Central Bank by Videla in Argentina respectively. They were accompanied and followed by many US trained economists in the highest state positions. At universities, economics flourished while other disciplines decayed. While leftist economists left public universities for exile or private research centers, liberal economists founded private schools or think tanks collaborating with the military. At the University of Buenos Aires, the degree in 'political economy' was changed into 'economics' and its curriculum was deeply modified to enlarge mathematics and reduce humanities and social sciences. In the long term, dictatorships encouraged a strict delimitation between economics and other social sciences that persists today. While most economists became international and private consultants independent from universities and public agencies, academic institutions, especially public ones, employ most other SBS scholars.

Contrary to more stable democratic nations, certain aspects of previous SBS configuration still remain: the disconnection between excellence in research and teaching, the contrasting sort of economists and other social scientists, the preference for applied studies, the importance of private and international funding for private research centers, the fragmentation of national fields, the importation of general theories and international prestige as a source of distinction due to weakness of other national and regional sources of academic recognition.

Conclusion

In all dictatorships, the intellectual and institutional *autonomy* of SBS is compromised, though to different degrees and in different aspects. Inasmuch as immediate needs of dictatorial regimes are privileged, SBS scholars loose their capacity of freely choosing their subjects and, frequently, their methods. As authorities try to control public discourses, academics loose their right to diffuse publicly their findings or views. A perpetual threat of repressions and dependence on public authorities prevent from the consolidation of autoregulated academic communities and reinforce their fragility vis-à-vis external influences. Nonetheless, one should not overestimate the *specificity* of SBS in dictatorial (autocratic) regimes as compared to democratic ones. The predominance of empirical (statistical) forms of knowledge and orientation toward public policy needs characterizes, to a lesser or a greater extent, relations between the state administration and academia in democratic countries during the twentieth century.

Most common and important features of social sciences under Socialist regimes are as follows: ideological alignment to historical materialism (Marxism–Leninism) under control of censorship; political guidance with respect to legitimate objects of research and theories; hesitations between practical needs (rationalization of government) and fear of dissidence; technocratic and applied orientations of social sciences; a relative isolation and autarchy; auto-limitation in theoretical activity. However, practical and ideological needs of the Socialist states contributed to professionalization and institutionalization of social sciences on a pace with democratic countries in the post-World War II period. Privileged relations with political authorities could enlarge the marge of initiative and autonomy, yet this autonomy was always threatened. Post-factum (or by external observers) SBS under dictatorship are often considered as 'antiscience,' 'pseudoscience,' or 'ideological bluff' like political economy of Marxism–Leninism or Nazi sociology (Lepenies, 1985). A forced compromise with authorities and official ideology was also often a source of a poor public image of social sciences. Nonetheless, an important role of social scientists in the reformist movements across the Socialist bloc is to be stressed. Humanistic and critical discourses, gestated in the liberal atmosphere of 'Thaw,' reappear with force at the moment of Perestroika and democratic transitions.

Latin American dictatorships also show common patterns with regard to SBS: violence and repression against public institutions, persecution of students and intellectual networks; ideological intolerance to critical discourses in public space; indifference with respect to legitimate objects of research and theories apart from economics and religious controversial matters; technocratic and applied orientations of social sciences; strengthening of international networks at the expense of regional and national integration; privatization of research. Similarly to the Communist Eastern Europe, practical needs of the Latin American states contributed to professionalization and institutionalization of social sciences. But unlike in the former, neoliberal policies and internationalization were encouraged here well before the fall of dictatorial regimes. Budget cuts and suppressions of positions in public institutions were specific factors of SBS' economic and symbolic

dependence. On the other hand, international networks were a source of economic and political support for SBS scholars. International connections, social knowledge, and commitment to free thought gave SBS a crucial role in reformist movements and democratic government in the 1980s.

Yet, academic freedoms and self-regulation are never wholly achieved; furthermore, new challenges are faced by social sciences. A shifting situation of SBS alone might be viewed as a good barometer of social and political change in post-dictatorial societies.

See also: Cold War: Impact on Social and Behavioral Sciences; Development and the State.

Bibliography

Arendt, H., 1951. The Origins of Totalitarianism. Schocken Books, New York.
Beigel, M.F., 2013. The Politics of Academic Autonomy in Latin America. Ashgate, Farnham, Surrey, England; Burlington, VT.
Berman, E., 1983. The Influence of the Carnegie, Ford, and Rockefeller Foundations on American Foreign Policy. The Ideology of Philanthropy. State of New York University Press, New York.
Bikbov, A., Gavrilenko, S., 2002–2003. Russian sociology: autonomy under question. Logos 35 (5–6/2002), 1–25, 37 (2/2003), 51–85 (In Russian).
Bockman, J., 2011. Markets in the Name of Socialism: The Left-Wing Origins of Neoliberalism. Stanford University Press, Stanford.
Brooker, P., 1995. Twentieth-Century Dictatorships: The Ideological One-Party States. Macmillan Press, London.
Buchbinder, P., 2005. Historia de las Universidades Argentinas. Sudamericana, Buenos Aires.
Brunner, J.J., Barrios, A., 1987. Inquisición, mercado y filantropía. FLACSO, Chile.
Calderón, F., Provoste, P., 1992. Autonomía, estabilidad y renovación. Los desafíos de las ciencias sociales en América Latina. CLACSO, Buenos Aires.
Ezrow, N., Frantz, E., 2011. Dictators and Dictatorships: Understanding Authoritarian Regimes and Their Leaders. The Continuum International Publishing Books, London.
Giner, S., 1990. Spanish sociology under Franco. In: Giner, S., Moreno, L. (Eds.), Sociology in Spain. Consejo Superior de Investigaciones Cientificas, Madrid, pp. 57–72.
Graham, L., 1967. The Soviet Academy of Sciences and the Communist Party. Princeton, Princeton, NJ.
Human Security Report Project, 2011. Human Security Report 2009/2010: The Causes of Peace and the Shrinking Costs of War. Oxford University Press, New York.
Keen, M.F., Mucha, J. (Eds.), 1994. Eastern Europe in Transformation. The Impact on Sociology. Greenwood Press, Westport, Connecticut-London.
Kim, M., Schoenhals, M., Yong-Woo, K. (Eds.), 2013. Mass Dictatorship and Modernity. Mass Dictatorship in the Twentieth Century. Palgrave Macmillan, London.
Klingemann, C., 1996. Soziologie im Dritten Reich. Nomos Verlagsgesellschaft, Baden-Baden.
Koleva, S., 2002. The disciplinary identity of sociology: profiles of construction (sociology in Poland, Russia and Bulgaria in the 1950s and 1960s). Sociological Problems 74–91 (special issue).
Konrad, G., Szelenyi, I., 1979. The Intellectuals on the Road to Class Power. Harvester Press, Brighton.
Krementsov, N., 1997. Stalinist Science. Princeton University Press, Princeton.
Lepenies, W., 1985. Die drei Kulturen: Soziologie zwischen Literatur und Wissenschaft. Hanser, Munich and Vienna.
Levada, Y., 1969. Lectures on Sociology (in Russian). Informational Reports of the Research Council on Problems of Concrete Sociological Studies (Moscow), 5 (20)–6 (21).
Linz, Juan, 1975. Totalitarian and authoritarian regimes. In: Greenstein, F.I., Polsby, N.W. (Eds.), Handbook of Political Sciences, vol. 3. Addison-Westley, Reading, MA, pp. 175–411.
Linz, J., Stepan, A., 1996. Toward consolidated democracies. Journal of Democracy 7 (2), 14–33.
Love, J., 1996. Economic ideas and ideologies in Latin America since 1930. In: Bethell, L. (Ed.), Ideas and Ideologies in Twentieth Century Latin America. Cambridge University Press, Cambridge, pp. 207–274.
Navarro, Juan José, January 2011. Cold war in Latin America: the camelot project (1964–1965) and the political and academic reactions of the chilean left. Comparative Sociology 10 (5), 807–825.
Neiburg, F., 1998. Los intelectuales y la invención del peronismo. Alianza Editorial, Buenos Aires.
O'Donnell, G., 1972. Modernización y autoritarismo. Paidós, Buenos Aires.
O'Donnell, G., Schmitter, P.C., Whitehead, L. (Eds.), 1991[1986]. Transitions from Authoritarian Rule: Comparative Perspectives. The Johns Hopkins University Press, Maryland and London.
Quelle sociologie derrière le "rideau de fer" ? 1950–1989, 2007/1. Revue d'Histoire des Sciences Humaines 16.
Smith, T., 1991. The alliance for progress: the 1960s. In: Lowenthal, A.F. (Ed.), Exporting Democracy: The U.S. and Latin America. Johns Hopkins University Press, Baltimore, pp. 71–89.
Sonntag, H., 1988. Duda/Certeza/Crisis. La evolución de las ciencias sociales en América Latina. Unesco-Nueva Sociedad, Caracas.
Stoppino, M., 1983. Autoritarismo. In: Bobbio, N., Matteucci, N., Pasquino, G. (Eds.), Diccionario de Política. Siglo XXI, Madrid, pp. 125–136.
Torres, A.C., 2010. Sociology, science and profession: the Portuguese experience. In: Patel, S. (Ed.), The ISA Handbook of Diverse Sociological Traditions. Sage Publications Ltd, New Delhi, pp. 105–113.
Trinidade, H., 2007. Las ciencias sociales en América Latina en perspectiva comparada. Siglo XXI, México.
Voříšek, M., 2008. Antagonist, type, or deviation? A comparative view on sociology in post-war Soviet Europe. Revue d'histoire des sciences humaines 18, 85–113.
Wagener, H.-J., 1997. Second thoughts? Economics and economists under socialism. Kyklos 50 (2), 165–187.

Social and Economic Change: Psychological Challenges for Individuals

Rainer K Silbereisen, Center for Applied Developmental Science, University of Jena, Jena, Germany
Martin J Tomasik, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the consideration of contextual influences on individual development has a long tradition in psychological science, the rapid pace of social, economic, and political change around the globe has spurred new conceptual and empirical research on the dynamics between macro-level change and individual-level adjustment and development. This research is targeting epigenetic, psychological, and social processes by various methodological approaches, such as analysis of archived data, cohort comparisons, natural experiments, comparisons between stages of social change, or the assessment of interindividual differences in response to challenges arising from social change. Empirical findings from different methodological approaches converge in the notion that social change has both comprehensive and polyvalent effects that often follow a prototypical syntax and cascade through the various levels of developmental contexts, which mediate the effects on the individual and moderate them as a function of cultural traditions. More recent research shows that individuals differ in the extent to which they are affected by such change and that these differences are in part attributable to the way demands resulting from social change are perceived and appraised, to how individuals cope with such demands, and to the moderating effects of proximal and distal developmental contexts.

Social change is an umbrella term for various kinds of radical and comprehensive changes concerning the political system of a society, its economic organization, major social institutions, means of production, demographic composition, ecology, and value and belief systems. While social change is obviously not a new phenomenon, many believe that the pace of social change has been accelerating over the recent decades (see Tomasik and Silbereisen, 2012). This acceleration may be rooted in the growing integration of societies across the globe, thereby facilitating the spread of innovations; or it may be due to actions by political elites in less developed countries who want to catch up with technologically more advanced societies.

Until 20 years ago, social change was not a prominent topic of empirical psychological research, but this changed with the breakdown of the Soviet Union and other communist countries around 1990 in Europe and Central Asia. Although this started as a purely political transition, it soon turned into a broadscale economic transformation to market capitalism, with all kinds of concomitant adversities, such as increased unemployment and income disparity, despite many beneficial effects. The effects of the political breakdown were also soon overlaid with new challenges of globalization, which changed traditional means of production and removed many trade barriers resulting in worldwide competition. Formerly protected markets had to open up for products manufactured by cheaper labor, for migration of people from different regions of the world, and for unprecedented idea diffusion around the globe. The large-scale change processes were further influenced by the worldwide financial and economic crisis that started in 2007 with the collapse of global bank systems. This lead to entire regions suffering from economic depression and recession and to a currency crisis in the European Union that is still unresolved. Many countries also had to make major adjustments to their already strained welfare systems due to challenges resulting from demographic change, such as a decline in the birthrate and an increase in the older segment of the populations. Another challenge is the increase in migration across borders with permanent resettlement, as well as transnational migration characterized by belonging to more than one nation-state.

Concerning the causes of the macro-level changes, many explanations have been offered. If we take the breakdown of the communist rule in the former German Democratic Republic (East Germany) as an example, and this may as well apply to other countries dominated by the former Soviet Union, a prominent theory sees the lack of modernization of the entire society as core element behind its collapse. This was compounded by an inept model of central planning of economic activities that were completely out of step with the needs of an industry dependent on innovation and worldwide cooperation. Likewise, modern modes of production require increased autonomy and responsibility, and in response to that politics have to exercise participatory modes of governing. Other explanations have suggested that it was not the state of the economy that was pivotal in the downfall of communist rule but a basic discrepancy between the declared humanistic beliefs of a society and a permanent deprivation of civil liberties. Such discrepancies were made visible to growing segments of the population by the increased interconnectedness of people especially via new means of communication such as internationally broadcast TV programs. It is interesting to note that countries such as China are now struggling with similar problems related to the Internet in spite of the new freedom in the economy and business.

Psychological Dynamics behind Political Change

From a psychological perspective, a rapid transformation of the political system is not possible without a parallel radical decline in the population's support of the status quo. Under ordinary circumstances, people tend to engage in various motivational psychological processes aimed at defending the social system in which they live, even when confronted with

massive failures in its functioning. This seems to be because we all live interdependently in broader social systems and have a tendency to uphold the legitimacy of whatever norms, rules, and social structures inherent in these systems. However, according to the systems justification approach, there are circumstances demonstrated in experimental research that modify this general tendency. External threats to a political system tend to strengthen the justification of the established order, and people typically defend systems on which they depend, often in spite of criticisms they themselves may have. Another condition that strengthens the justification of an established order is the inescapability of a political system, be it through tight border controls or through other constraining measures. Similar to the system justification approach, one can also delineate conditions that lead to collective action in situations of social change, among them the erosion of intergroup barriers that are seen as unjustified.

The aim of such motivational processes, as evidenced by the case of the former communist countries, is a rationalization of the situation in order to make the experience tolerable; here other new research has demonstrated that rationalization of constrained freedoms (seeing them in a most positive light, as suggested by systems justification) rather than taking action (protesting openly) against such circumstances is related to the severity of personal restrictions. The difference probably lies in the level of absoluteness – if the restrictions to freedom apply to all in a complete, certain, and permanent fashion, rationalization is the likely response. If, however, restrictions do not apply to all equally, or are uncertain, and time-limited, then behavior to the opposite of system expectations is more likely. Note that such psychological reactance was promoted by changes in the border regime, and by otherwise increasing contacts between East and West in Germany and elsewhere. The unexpected opening in 1989 of escape routes to the West through neighboring countries, such as Hungary, reduced the tendency for system justification among the East German population.

Finally, a lack of control over one's personal life is often compensated by a belief in the control established by the social system. In other words, commonly shared stability mechanisms may keep a system alive for a long time, but the tipping point may be easily reached once the system's legitimacy is put into question. This is probably what happened with the unprecedented civil strife shortly before the fall of the Berlin Wall because ways to escape, to become less dependent, and to exercise control had become evident. The decline of the ideological barriers of the Cold War also played a role in reducing the external threat. In other words, a major contribution to the openness for social change by the people and in part by the elites of East Germany was due to the gradual erosion of the motivational forces to maintain the status quo and by the promise of reforms that, however, came too late. The failing of reforms had undercut the reality and perception of the absoluteness of all social controls that had helped to stabilize the system by motivational processes. Such processes may also, of course, be contagious – the awareness that absolute political systems similar to their own have been toppled may lead citizens to question the absoluteness of their own system, thereby empowering them to vent their accumulated anger over the failure and injustice of their own system and increasing the likelihood of collective action toward social change.

Social Change a Topic of Psychological Research

Psychology was badly prepared to studying the consequences of macro-level change on individual behavior and development. When we conducted one of the first studies on the consequences of German unification, the issue of social change and related concepts did not really exist in the psychological literature. For example, entering the key words 'social change and well-being' in the PsychINFO data source returned 29 entries up to 1998. The same search from 1998 to 2012, however, returns 219 entries, which represents an impressive increase in research-based activities. As is understandable in a situation of unpreparedness, researchers in psychology fell back to earlier approaches that had stressed the role of the physical and social environment, such as Bronfenbrenner's model of hierarchically organized layers of contexts, from the macro level of societal beliefs via more intermediate contexts such as neighborhoods to the micro context such as the work place or the family and their intersections. Oishi and Graham (2010) see the emerging research on social change since the 1990s as a push to revitalize buried theoretical traditions in psychology that deal with macro-level phenomena in research on human mind and behavior. They deem this as especially true for cultural psychology and its dealing with symbols, meanings and regions, and for social ecology, studying physical, societal, and interpersonal environments. Scientists such as Bronfenbrenner or Markus and Kitayama are seen as utilizing approaches leading directly to research on social change. One has to bear in mind, however, that the contexts, whether cultural or ecological, were traditionally conceived as rather stable entities, whereas social change is by definition a dynamic phenomenon.

The renaissance of socioecological approaches and their transdisciplinary nature comes at a time of another major development in psychology – that of molecular genetics research. As it turns out, both have unique relevance for our understanding of the role of social change and human behavior and its course across the life span. Recent innovations in molecular genetics research for the first time illuminate the actual ways in which the environment influences future psychological pathways, especially through influencing major physiological systems highly relevant for adaptation. More specifically, particular environmental pressures typically related to strains involved in macro-level adversities, such as wartime famine on mothers and their small children (Heijmans et al., 2008) have the capacity to change the expression of DNA via methylation and other molecular processes. They modulate the 'readability' of genetic coding, and thereby influence systems such as the hypothalamic-pituitary-adrenal (HPA) axis that leave a mark on how individuals deal with strain and stress. Certainly such epigenetic influences are not the only ones, and their influence may be moderated by later experiences, but the knowledge that there is a clear biopsychosocial pathway linking adversities typical of social change with individual behavior and development is new. Very recently, researchers also reported evidence that such processes are not limited to childhood.

The role of the stress system is central in the 'experiential canalization' model presented by Blair and Raver (2012). Here, adversities are seen as a source of characteristic restrictions of parenting that translate down to gene expression, and consequently regulate the stress system that is adaptive to the harsh environment being experienced. This leads to a particular pathway of psychosocial development; first by influencing neural connectivity and then by having an effect on behavior core for adjustment, such as self-regulation. The canalization model not only demonstrates ecological mediators (parenting in this case) and biological processes (epigenetic influence of stress system), but also makes an important distinction between short- and long-term effects. The possible negative short-term effect of adversity-stricken environments on self-regulation represents not a deficit, but actually an adaptation to the circumstances that provides benefits by a faster response to threats typical of the environment. In the longer-term, however, such benefits may turn into disadvantages concerning physical and psychological health, if they are not compensated for by benign future experiences.

For psychological concepts on the effects of macro-level social change on individuals, when taken together, the important message of the two strains of research is that the actual experiences of social change, processed through various ecological contexts, are crucial because they leave their traces on core physiological and neurological systems that relate to the ways in which people deal with challenges in general. Against this backdrop it is no big surprise that ways of coping with stress and related concepts play a core role in psychological research on social change.

Lessons Learned from Research on Social Change and Human Development

Studies on political and economic change over the last two decades established a number of empirically validated principles concerning change itself and its consequences for human behavior and development. Considering these results helps psychologists to see how this is embedded in broader societal structures and to understand its dynamics.

Syntax of Change

A good example of the syntax of change is that of German unification. During the transition period following unification, almost all institutions, from welfare to industry, were remodeled after the West German prototype. Several years into unification, however, it became clear that, for instance, the former provision of day care for workers' children actually on the premises of industrial plants had the beneficial effect of attracting skilled workers and their families to the East. More generally, what at first glance appeared to be lack of differentiation between state and economy under the old system, gradually became seen as a possible model for the family-friendly organization of work contexts. In general, the subsequent transformation period was characterized by a renaissance of early child care as a right for all families with children. In essence, such processes can be seen as adaptations – for better or for worse – to 'overshooting' changes of old institutions in the heat of transition.

Role of Cultural Traditions

Cultural traditions built up over decades or even centuries significantly moderate the effects of gross changes in political and economic systems. Same-aged young people who had attended the same secondary school system in the Soviet Union strove quite differently after the breakdown of communism and the Soviet Union if they were raised and now lived in the new Baltic States or in one of the former Central Asian republics. Entrepreneurial skills and activities that, given its history, were traditionally more prevalent in the Baltic States helped to overcome the meanwhile obsolete career paths typical of the Soviet Union (Titma and Tuma, 2005). Similar observations were made by economists who found that old centers of innovative industries quickly became revitalized after German unification. There are many more instances of the role of buried cultural traditions and their revitalization in periods of radical social change, such as the renaissance of Confucian beliefs in Vietnam or the lasting influence of the collectivist cultures in leisure patterns of the young in India's urban middle class that otherwise changed a lot due to industrialization and the effects of globalization. Religion and spirituality also played a prominent role in the breakdown of the (outspokenly secular) Soviet system in Central and Eastern Europe when apparently forgotten orientations and networks came back to life. In cases such as Poland, the church even belongs to those who replace the old elites.

Cascading Effects

Changes on the macro level do not affect individuals directly. Rather they 'cascade down' through various intermediate contexts, ultimately exerting an effect on families, which can be seen as the micro context par excellence. From there, the effect reaches individuals, who are also influenced by their current stage in the life span. A perfect example is provided by research on the effect of social change in a remote valley in Asia (Axinn et al., 2011). Government activities to develop the region by building streets to attract economic development did not immediately change the role of women and the nature of parenting in the traditional communities. Rather, it was the establishment of small workshops that provided paid work and required women to stay outside their homes for the working day. The new experiences had an effect on the appreciation of values such as autonomy, and this then gradually reshaped parenting priorities and also the partnerships toward more equality.

Comprehensive Effects

If radical and comprehensive enough, challenges will have an effect on social, human, and financial capital. In the case of the social change in Europe that followed the upheavals of 1989, it was obvious that people trained in occupations that had been functional under the old system were having difficulties in utilizing their skills, especially in areas like economics and law where qualifications were based on a different value system.

Consequently, such people risked their status within society, because both old and new systems very much based social status on formal educational merits and occupational prestige that, however, were different. This devaluation of human capital, however, did not leave all people helpless, as research by Titma and Tuma (2005) revealed. In the early years after the collapse of the USSR personal agency was found to be more relevant for life success rather than formal qualification or credentials. An example of such agency was the informal selling of scarce but sought-after consumer goods that was based on earlier social ties and that required a high level of personal self-efficacy for success. It was only later in the process, once the new institutional order had been established, that education and other formal credentials became relevant again for one's position in the status hierarchy. It appears that self-efficacy established before the transition even helped people who were close to the former state and party organizations and who subsequently often experienced discrimination and social alienation. Furthermore, since the collapse of the Soviet Union in the late 1980s, there has been a common trend in many postcommunist transition societies toward increasing income inequality, whereby large segments of the population live in poverty while others have become extremely wealthy. In contrast to the 'Kuznet curve,' which posits a natural cycle of economic inequality driven by market forces during social development and transformation, in these postcommunist countries inequality did not decrease even after initial growth. For instance, metropolitan centers gained relative to rural areas, the financial sector increased at the expense of others, and in some countries even expenditures in education and science declined compared to previous times, with negative long-term effects for competitiveness. Overall, these effects are counter-productive for people's trust in the social system.

Polyvalent Effects

Political and economic change is polyvalent concerning consequences for individuals. Therefore, in order to structure possible outcomes for human development, it makes sense to look at the basic tasks and achievements that characterize any particular phase of the life course, often called developmental tasks; in the following, we take young adulthood as an example.

Social change has the potential to affect the ways and means by which individuals resolve the task of embarking on an occupational career, of starting a family or partnership of their own, or becoming an active member of civil society. A review of relevant research have shown that it is not usually the substance or appreciation of the developmental tasks that are affected by social change, except perhaps for the new role of public self-presentation in modern mass media, but that the accustomed steps toward those goals may lose their relevance in the changed societal structure (Tomasik and Silbereisen, 2012). For instance, the combination of high unemployment rates in many countries in Europe, especially those considered to be in transition and the effects of globalization-related adjustments, such as downsizing or unstable work (unemployment rates among young people have reached 40% in some Southern European countries), makes it difficult for young people to follow settled routines for establishing themselves in an occupation. This has negative carryover effects on partnering and family formation, as well as on civic participation. Beyond the effects of social change on core developmental tasks, there are also overarching consequences for people's well-being. In this regard, mental health is of utmost importance for social policies, and it is fair to say that 'threatened well-being,' including an increase in mental disorders, seems to be characteristic of European transformation countries after 1990 (Orosa, 2012). It should be noted that the increase was especially steep in new countries of the European Union, thereby demonstrating that it is probably not only the shock of the 1990s that fueled the changes.

Conceptual Ways of Linking Macro with Micro

Psychological studies on individual consequences of social change require theoretical conceptualizations that reflect the macro level and at the same time model the interindividual processes in response to the manifestations of the macro change. Thus far, psychological research has followed two slightly different routes concerning the outcomes of interest. First, scientists used the specific psychological mechanisms known to influence a particular target behavior of interest, and reformulated particular models to connect additional exogenous variables or design features reflective of social change. Schoon et al. (2007), for instance, who were interested in career transitions in times of social change, used a model entailing parental social class, and parents' expectations for the young adolescent's school motivation and academic success, their job aspirations, the timing of parenthood, with occupational status in adulthood as outcome. Manifestations of social change were represented by individual material hardship that was increasing due to economic challenges and societal unrest in the United Kingdom at that time. The longitudinal data comprised the period from birth until the offsprings' third decade of life.

Second, rather than relying on behavior-specific models, other researchers used a more general conceptualization of processes likely to have an impact on a wide range of different target variables. Conger et al. (2010) used family material hardship to reflect the macro-level conditions at the time of the Iowa farm crisis in the United States. This crisis was characterized by agricultural recession, which included farm bankruptcies, out-migration, and even increased suicide rates. The authors assessed an intraindividual cascade of processes from perceived economic pressure to coping reactions with these strains, resulting in parental distress, marital conflict, and reduced emotional investments, which then hampered academic achievement and social competence in children. The study comprised longitudinal data spanning three generations.

Using these two studies as illustrations serves another purpose – both refer to the landmark research of Glen Elder that he conducted on the Great Depression of the 1930s. He used data from the period but based his work on current concepts (Elder, 1974). This general approach, which characterizes both studies reported, used a circumscribed but potent manifestation of the societal challenges, embedded them via sample and exogenous variables in the larger historical context, and capitalized on established social and developmental psychology models of the target outcomes of interest. To date,

research following such approaches has concentrated on the structure of the life course (timing and sequence of transitions), subjective well-being and mental health (especially interesting because of its dual nature as antecedent and consequence of thriving and success in life), career development, and civic engagement (because it may represent the seed for future collective action to overcome the negative effects of social change). Several other classes of behavior and development were also tackled, such as participation in the new media or leisure in general, but the common denominator is that 'ultimate' outcomes were chosen that are of high relevance for social policy planning (Orosa, 2012).

Methodological Approaches

Closely connected with the conceptual models are methodological approaches, of which four stand out as being the most utilized. First, archived data were analyzed representing cohorts that stand for a historical divergence in macro-structural conditions. The work of Schoon et al. (2007), who used the longitudinal British Birth Cohort Studies of 1958 and 1970, is a prime example of such research. In comparing results between cohorts later in adolescence and adulthood, Schoon could show that personal agency factors, such as job aspirations and educational performance, played a different role in predicting occupational development, depending on the economic prosperity the cohorts had lived through during their formative years. In times of economic stability, job aspirations were of prime importance. In economically precarious times, however, it was educational performance that played the main role. This particular divergence reflects the historical changes on the labor market toward higher skill levels based on extended formal education and advance credentials (Hofäcker et al., 2010).

Second, studies, such as that by Kohn (e.g., 2010), used a natural experiment approach made possible by the gradual diffusion of social change across countries. The countries studied were chosen because they represent advancing steps in a process of social change toward market capitalism. The focus of the research, however, was the role of greater opportunities for self-directed work among people higher in the social stratification hierarchy, which had a subsequent positive influence on personality flexibility. This was demonstrated for the United States and subsequently for Poland in the 1980s before regime change. But one difference was found whereby in Poland but not in the United States, workers' well-being was higher than that of managers, mirroring the greater job protection enjoyed by workers in the former communist Poland. This difference changed rather quickly to the U.S. model after the change from a socialist command economy to market capitalism in Poland. The same change – although more gradual – also happened in other Eastern European transformation countries, such as the Ukraine. Thus, the very fact that the position in the status hierarchy corresponds to higher self-direction and personality flexibility appears to be universal, but it was no longer the workers who revealed higher well-being but the managers as is common in the United States and the West more general. Rather limited change against the backdrop of remarkably stable structures and people indeed seems to be an important characteristic of social change in general (Tomasik and Silbereisen, 2012).

A third approach to studying social change was to conduct planned comparisons between given times in the process of a critical event characterizing radical change in a political and economic system, such as German unification. In the early 1990s, Silbereisen and colleagues began such a program with research comparing same-aged, representative cohorts in young and middle adulthood between 1991 and 1996. During that period of transformation, social institutions in the East related to education, training, and career development in general were replaced with ones that followed the West German model. The effect was an increase not only in the importance of individual choice, but also in personal risk concerning employment prospects. Against this backdrop, the research found that following unification, the timing of first vocational decisions was delayed in the East whereas timings in the West stayed the same. In general, the findings demonstrated the influence of social institutions as pace-setters concerning life-course transitions. However, as expected, in areas where there was no structural change or no obvious role of social institutions, such as the age of first romantic involvement, timing did not change.

Fourth, all the methodological approaches mentioned thus far share the view that the social change addressed was so overwhelming that differences in the extent to which individuals were affected by the change are not really relevant. This was an approach, however, not taken by all. For example, an interdisciplinary group of sociologists, psychologists, and economists, working as part of the largest research consortium on social change in Germany after 1990 (SFB 580), focused on people's particular individual experiences associated with and stimulated by ongoing social change. The emphasis of the psychologists' work was on the internal psychological processing leading to psychosocial outcomes in response to social change.

Recent Research Using a New Conceptualization

This research of the psychologists in the SFB 580 consortium was guided by the 'Jena Model of Social Change and Human Development,' which puts new 'demands' on individuals resulting from change at the core (see Silbereisen et al., 2010). Demands are conceived as proximal manifestations on the individual level of macro-level structural change. These demands are influenced by a cascade of hierarchically organized contexts in which individuals may find themselves and that may themselves be changed through the influence of social change. An example of this is the 'filtering' of risks for individuals' resolution of major developmental tasks by institutions of the welfare state (Hofäcker et al., 2010). For instance, globalization provides challenges for businesses that are passed down to employees, thereby possibly confronting them with unstable conditions which bring with them uncertainties concerning future planning regarding their work and careers. Such demands, however, are not equally distributed across the population. Whereas in some countries like Italy, seniority protects employees from the more negative aspects of globalization, such as job termination, in other countries like Germany, it is the young who are protected against

unemployment. One consequence of this is the dramatic difference in youth unemployment rates. In the sociological terminology of Hedström and Swedberg (1996), demands represent the 'situational mechanisms' that link structural changes on the macro level to the individual and form the point of departure for psychological processing.

Demands and Well-Being

The Jena group studied the relationship between demands in various domains of life and several indicators of well-being. Large representative samples of people in young and middle adulthood from 77 administrative regions in four federal states of Germany were surveyed. The regions represented distal contexts through which social change cascades to the individual. As reported by Silbereisen et al. (2010), demands were assessed by asking participants to endorse statements concerning growing perceived uncertainties in the domains of work, family, and public life over the last few years. The topics were chosen so as to reflect changes in population statistics for the 5 years leading up to 2005. For example, the item, "When considering the past five years, my career plans were often hindered by unforeseen events and circumstances" was associated with the documented increase in staff turnover rates and other relevant indicators of precarious work conditions between 2000 and 2005. A cumulated load index was formed for six items each that were deemed to cover relevant experiences people could have faced due to the ongoing social and economic change.

Based on the filtering concept, and as expected, participants who were employed, married, had high educational attainment, and were living in the more prosperous West of unified Germany showed lower demand loads. The strongest effect was found for current employment status. Moreover, the 77 regions (identified in official European statistics as NUTS-3 units) also showed differences in aggregate demand loads; a variation that reflected differences in actual economic prosperity. Taken together, the demands indeed seem to reflect manifestations of objective circumstances. What applies to demands in the domain of work also was revealed by family demands (such as "When considering the past five years, it is more difficult, given my present life circumstances, to decide whether I want to have a(nother) child or not").

In terms of resource conservation, demands may bring with them the threat of resource loss in important domains of life and thereby produce stress and require action. Certainly, it is known from relevant research that the anticipation of future possible threats constitutes a potent source of stress which is likely to have effects on well-being if not dealt with adequately. Thus, although demands resulting from social change are not specific events but personal concerns, it is feasible to assume that individuals differ in their experience of uncertainties relevant to their resolution of developmental tasks and to the maintenance of their well-being. Recent research, which focused on the relationship between demands and various indicators of well-being, distinguished between a stable 'trait' component and a variable 'occasion' component of life satisfaction. They demonstrated that demands have an effect on the occasion component of life satisfaction, but that the trait component also has an effect on the demand load; the two effects had about the same magnitude. Thus, demands were confirmed as affecting (reducing) well-being independent of the fact that people showing higher well-being also experienced lower loads of demands.

Negotiating Demands of Social Change

If demands related to ongoing social change mean stress for individuals, then how they deal with it should play a role. To assess these ways, the Jena Model used the developmental regulation approach by Heckhausen et al. (2010), which distinguishes between engagement with the demands (working hard to resolve the issues whatever it takes) and disengagement (to redefine or avoid dealing with the demands all together). For example, an individual when facing impaired career perspectives may undertake many activities and invest a lot of effort into retraining, and yet ultimately decide to turn with all effort to a different field with better chances. According to Heckhausen et al., this represents three forms of engagement in pursuing age-typical goals. In contrast, disengagement means that, when faced with an obstacle to achieving a goal, the individual either downgrades the importance of the goal or makes a total change in direction; both behaviors represent two different forms of disengagement. In the sociological terminology of Hedström and Swedberg (1996), engagement and disengagement represent the 'action formation mechanisms' in responding to social change.

Analyses on the German data, whereby demands were assessed by items referring to the experiences reported, revealed that engagement was always more endorsed than disengagement, and that engagement was greater the more demands were appraised as a challenge, as leading to probable gains, and controllable. Higher demand loads corresponded to higher levels of both engagement and disengagement. All this is what one would have expected from earlier research on very different issues, such as complications of physical health or imminent serious life decisions (see Heckhausen et al., 2010). Whether dealing with demands of work or with demands related to family did not make much of a difference. Participants from the more severely affected East of the country revealed lower levels of disengagement in spite of higher levels of demands than in the West. This finding probably means that these people cannot afford to disengage given the fewer reserves, such as family financial assets or opportunities in general, at their disposal, due to generally lower affluence in the East.

Nevertheless it would be inadequate to view engagement as the best option under all possible circumstances. When the Jena group again looked at the relationship between demands and well-being, it became clear – as expected – that engagement or disengagement has a positive effect on well-being only when the action taken is a good match to the individual's circumstances in terms of controllability. More specifically, well-being was less negative given a certain demand load if people judged the demands to be under their control. It seems that engagement or disengagement not congruent with opportunities is harmful.

Moderating Effects of the Context

Of special importance for the Jena Model is the view that the relationship between demands and psychosocial outcomes is

moderated by contextual variables. A prime example is the economic prosperity of the political-administrative regions within Germany and how this affected the people living there. As already mentioned, the sample was structured by 77 of such regions.

Based on other research conducted soon after unification (e.g., Bergmann, 1994), we expected a weaker effect of demands on well-being in regions with lower economic prosperity. In earlier studies we found that economic hardship in the East of the country was not associated with depressive mood, which was contrary to findings for the West (and for many other samples from Western countries). We saw this protective effect as rooted in differences in attributions of the cause of the problems: system failure in the East, individual responsibility in the West. Such differences were repeatedly revealed in surveys for the 1990s. What the Jena group now found in two related studies for work and family demands (Pinquart et al., 2009) confirmed our expectations and lends extra credence to the ecological-contextual approach in general. Utilizing multilevel analyses, the relationship between demand loads and well-being was less negative than the average among those who lived in regions with high unemployment rates.

In interpreting these results, the 'social norm' effect known from economics comes to mind. This means that the relationship between individual unemployment and mental health was weaker when the aggregate unemployment rate was higher: under such conditions, the unemployed feel better about their situation. It is clear that the demands studied are not equivalent to unemployment, but the norm effect was also observed for precarious working conditions and the work demands assessed came close to that.

We also found that the effects of engagement were less positive in economically challenged regions and that under such circumstances disengagement seemed to be even positive for well-being as response to high demands in economically weak regions. To understand this better, we separately analyzed a subsample of those who were affected by really high demand loads and again compared the 77 regions. As expected, the regression lines between satisfaction with work and disengagement were positive in economically disadvantaged regions and negative (as on average) in the other regions. Furthermore, satisfaction increased over time as a function of disengagement in the economically disadvantaged regions but decreased in others.

Benefits of Social Change

Social change is, of course, not limited to negative effects, although many of the studies we have reported concentrated on that, especially given the economic situation in the mid-2000s. An interesting aspect of globalization-related strains on the job market is that employers often seem to compensate increases in performance pressures by offering potential benefits such as growing autonomy in decision-making. Our view was that autonomy may function as a resource to counterbalance the negative effect of demands, as we found. The negative effect of high demands on subjective well-being was balanced by job autonomy. Other resources such as optimism may also play a role by increasing the likelihood of engagement (see Pavlova and Silbereisen, 2013, who demonstrated that even a rather stable disposition can affect how individuals deal with change).

Generalization across Countries and Type of Demands

Demands are conceptualized as rooted in changes on the macro level, affecting major domains of life not only in Germany but also in other countries that have undergone political and economic change. The level of demands may vary (as the comparison of East and West in Germany already revealed), but the effects should nevertheless show commonalities, also concerning demands other than those related to work and family.

Role of National Contexts

The results reported thus far refer to Germany and show that demands indeed are stressors that affect well-being negatively. Welfare systems work as a buffer against the negative consequences of social change in peoples' life, and in this regard Germany can be considered a conservative welfare regime. It is transfer-oriented and thereby attenuates the economic consequences for employees that, for example, result from unemployment. The current Polish system, in contrast, shows strong liberal elements that put an emphasis on the responsibility of the individual rather than the state (see Hofäcker et al., 2010). Against this backdrop, a parallel research program in Poland, which is a country considerably more stricken by the effect of political and economic change than Germany, gave a first confirmation that the Jena Model can be applied more broadly. Analyses on the ways in which people deal with demands revealed beyond many similarities to the German results, a major difference concerning the role of how demands were appraised: whether they were seen as a challenge or a threat, or as a gain or a loss did not play much of a role in Poland. This is likely a consequence of the much less generous welfare transfers common in that country. As such transfers mount to only one-seventh of those in Germany, people cannot afford to make their engagement with demands dependent on favorable appraisals but have to engage with them 'come what may.' This was the case in spite of the fact that demands were perceived as more threatening, and associated with loss and low controllability. Moreover, people responded to loss by disengagement, probably to preserve strength to tackle the many other pressures in their life.

Current social change has many more aspects than those reported here. For example, research in China focused on the effect of new opportunities for self-realization on parenting and found that parents who perceived the ongoing economic changes as positive reported more encouragement of independence in their adolescent children, thereby helping them to adapt to the new economic situation. This effect, however, was more pronounced in urban regions, whereas in rural areas, similar concerns about demands to those studied in Germany prevailed (Chen et al., 2010). Likewise, in a sample of young university graduates from Ghana, the perception of demands in the work domain corresponded with the intention to continue with education, thereby hoping to be better prepared for the new times.

Demands of Active Aging

A different line of research addresses the demographic shifts resulting from increasing life expectations and the 'active aging' policies pursued in many European countries. People beyond statutory retirement age (65 in Germany) are growingly confronted with demands to participate in life-long learning, to maintain their physical fitness, and to complement their future income after work life from state pensions by private pension insurance. Using the Jena model as before but with demand statements adjusted to the purpose, we analyzed demand levels and appraisals, again looking for effects of the economic disparities between East and West of unified Germany. The results revealed that demands were higher in the East among those still in work, and that demands were appraised as more threatening and less as a challenge than in the West. Among those already receiving a pension from the nation-wide public pension insurance fund, however, differences no longer existed between the regions. In other words, differences still existing in the economic opportunities between the two parts of unified Germany are mirrored in different demand levels. This has consequences for psychosocial outcomes, especially when one takes into consideration that older people's negotiation of demands linked to active aging seems primarily to depend on how they appraise these demands.

Conclusion

The main result of the research reported in this article is the role of various levels of contexts in moderating the complex relationships among demands (and benefits), strategies to deal with reduced opportunities and new expectations, resources one can bring to the fore, and psychosocial outcomes that are crucial for adjustment in life. This relationship varies as a function of the distal or proximal ecology, which provides risks and opportunities, and transforms the challenges by their own characteristics. However, it seems as if this moderating role has a double face: living in an economically challenged region eases the negative effects of demands on well-being, which certainly helps people feel better about their situation in the short term. Long term, however, it may undercut engagement thereby hindering the achievement of developmental tasks that would help to promote further options for better adjusted development.

When this line of research on the degree to which individuals are affected by the changes on the macro level and what this means for adjustment began, the topic of social change was almost nonexistent in psychology. The breakdown of communist rule in Eastern Europe, the gradual changes to market capitalism in countries such as China, and the democratic developments in states emerging from political oppression in countries such as South Africa, all played a role in instigating new research initiatives. More recently, the 2008/2009 worldwide economic crisis has lead to new research initiatives that utilize approaches in the tradition of Bronfenbrenner, Elder, Conger, and also the Jena model.

Stress related to globalization and other issues of social change is a major threat to mental capital and well-being, and these effects have to be seen as embedded in the entire life course, with a particular emphasis on the years marking the onset of adulthood. Spectacular evidence is provided by the consequences of rapid mass privatization on mortality in postcommunist countries. There is now evidence that among males such programs were associated with an increase in early mortality rates of about 10%. This effect was probably mediated by employment rates, and when comparing countries among those with higher social capital (participation in social organizations) the effect was reduced.

The challenges on the macro level that are behind the demands of social change are not easy to avoid or to change radically, so that, for social policy, it may be important to plan and implement interventions that support people's capabilities to deal with challenges of social change constructively. Such means have to be tailored to the type of change and to the culture within which it is found, but whatever one wants to do, interventions need to be very definite concerning clarity of expectations, certainty about implementation, and overall opportunities for peoples' striving for success. If, however, situations are blurred concerning what the new expectations resulting from social change actually are, the likelihood that people will fall back to or even 'accentuate' old behaviors rather than revert to what is actually needed given the change on the macro level (Caspi and Roberts, 2001). Whatever is the outcome on individual behavior and development can also have an effect on the macro level; the respective 'transformational mechanisms' (Hedström and Swedberg, 1996) relevant for this process have not yet been the focus of psychological research on social change.

See also: Coping across the Lifespan; Decision Making: Nonrational Theories; Motivation in Australian Aboriginal Populations; Motivation: Life Course and Sociological Perspectives; Personality and Adaptive Behaviors; Personality and Risk Taking; Schooling: Impact on Cognitive and Motivational Development; Self-Regulation in Adulthood; Sociodigital Revolution: Digital Natives vs Digital Immigrants; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Axinn, W.G., Thornton, A., Barber, J.S., Murphy, S.A., Ghimire, D., Fricke, T., Matthews, S., et al., 01 September 2011. Chitwan Valley (Nepal) Family Study: Changing Social Contexts and Family Formation. ICPSR04538–v7. Inter-University Consortium for Political and Social Research (Distributor), Ann Arbor, MI. http://dx.doi.org/10.3886/ICPSR04538.v7.

Bergmann, B., 1994. Erleben und Bewältigen von Arbeitsunsicherheit in Sachsen [Experience of and coping with occupational uncertainty in Saxony]. In: Montada, L. (Ed.), Arbeitslosigkeit und soziale Gerechtigkeit: ADIA-Stiftung zur Erforschung neuer Wege für Arbeit und soziales Leben, vol. 2. Campus Verlag, Frankfurt, Germany, pp. 214–232.

Blair, C., Raver, C.C., 2012. Child development in the context of poverty: experiential canalization of brain and behavior. American Psychologist 67 (4), 309–318.

Caspi, A., Roberts, B.W., 2001. Personality and development across the life course: the argument for change and continuity. Psychological Inquiry 12, 49–66.

Chen, X., Bian, Y., Xin, T., Wang, L., Silbereisen, R.K., 2010. Perceived social change and childrearing attitudes in China. European Psychologist 15, 260–270.

Conger, R.D., Schofield, T.K., Conger, K.J., Neppl, T.K., 2010. Economic pressure, parent personality and child development: an interactionist perspective. Historical Social Research 35, 169–194.

Elder Jr., G.H., 1974. Children of the Great Recession. University of Chicago Press, Chicago.

Heckhausen, J., Schulz, R., Wrosch, C., 2010. A motivational theory of lifespan development. Psychological Review 117, 32–60.

Hedström, P., Swedberg, R., 1996. Social mechanisms. Acta Sociologica 39, 281–308.

Heijmans, B.T., Tobi, E.W., Stein, A.D., Putter, H., Blauw, G.J., Susser, E.S., Slagboom, P.E., Lumey, L.H., 2008. Persistent epigenetic differences with prenatal exposure to famine in humans. Proceedings of the National Academy of Sciences 105, 17046–17049.

Hofäcker, D., Buchholz, S., Blossfeld, H.-P., 2010. Globalization, institutional filters and changing life course patterns in modern societies: a summary of the results from the GLOBALIFE project. In: Silbereisen, R.K., Chen, X. (Eds.), Social Change and Human Development: Concept and Results. SAGE Publications, London, pp. 101–124.

Kohn, M.L., 2010. Class, stratification, and personality under conditions of apparent social stability and of radical social change: a multi-nation comparison. In: Silbereisen, R.K., Chen, X. (Eds.), Social Change and Human Development: Concept and Results. SAGE Publications, London, pp. 190–207.

Oishi, S., Graham, J., 2010. Social ecology: lost and found in psychological science. Perspectives on Psychological Science 5, 356–377.

Orosa, F.J.E., 2012. Psychosocial wellbeing in the central and Eastern European transition: an overview and systematic bibliographic review. International Journal of Psychology. Advance online publication.

Pavlova, M.K., Silbereisen, R.K., 2013. Dispositional optimism forsters opportunity-congruent coping with occupational uncertainty. Journal of Personality 81, 76–86.

Pinquart, M., Silbereisen, R.K., Körner, A., 2009. Perceived work-related demands associated with social change, control strategies, and psychological well-being. Do associations vary by regional economic conditions? Evidence from Germany. European Psychologist 14, 207–219.

Schoon, I., Martin, P., Ross, A., 2007. Career transitions in times of social change: his and her story. Journal of Vocational Behavior 70, 78–96.

Silbereisen, R.K., Pinquart, M., Tomasik, M.J., 2010. Demands of social change and psychosocial adjustment: results from the Jena study. In: Silbereisen, R.K., Chen, X. (Eds.), Social Change and Human Development: Concept and Results. SAGE Publications, London, pp. 125–147.

Titma, M., Tuma, N.B., 2005. Human agency in the transition from communism: perspectives on the life course and aging. In: Schaie, K.W., Elder Jr., G.H. (Eds.), Historical Influences on Lives and Aging. Springer, New York, pp. 108–143.

Tomasik, M.J., Silbereisen, R.K., 2012. Social change and adolescent developmental tasks: the case of post-communist Europe. Child Development Perspectives 6, 326–334.

Social and Emotional Development in the Context of the Family

Rena L Repetti, Meredith S Sears, and Sunhye Bai, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Family relationships shape children's social and emotional behavior and understanding. When parents are warm, sensitive, and balance control with child autonomy, children learn to trust others and to regulate their negative emotions. Both discordant and supportive family interactions contribute to the development of social skills and social cognition. Repeated daily stressors at home increase physiological and emotional reactivity and promote maladaptive coping. Culture, socioeconomic status, and other aspects of the family's broader social context influence these critical dimensions of family life.

In the course of protecting, nurturing, and socializing children, families create the social and emotional fabric of their lives. Families that offer children the opportunity to balance individuality (the ability to assert oneself and a sense of self as separate from others) with connectedness (responsiveness to the needs of others) seem to foster optimal emotional and social development, at least in the subgroup of European-American middle class families that make up the vast majority of research participants (Darling and Steinberg, 1993).

Through observation of and participation in routines, habits, and rituals, children learn the expectations and norms of their cultural group, and gain a sense of belonging and membership (Grusec and Davidov, 2010). It is in the family that children first learn which behaviors and emotion expressions are, and which are *not*, socially acceptable and how to manage complex social situations. For example, children of parents who display high levels of positive affect tend to display more positive emotions among peers, while children of parents who display a lot of negative affect tend to show poorer social competence (Denham et al., 1997). Children growing up in families in which there is more frequent talk about mental states demonstrate more knowledge about emotions and greater social understanding (Dunn, 2007). Beliefs and attitudes critical to social development are also shaped by the behaviors that children observe in their families. For instance, children of substance using parents report more familiarity with substances (e.g., cigarettes, alcohol, and other drugs) and develop more favorable beliefs about their use at an earlier age (Ellis et al., 1997). Similarly, children who observe aggression during marital conflict surmise that aggression is an acceptable method of resolving disagreements, and tend to exhibit more behavioral problems themselves, such as aggression toward peers, especially if they respond to marital conflict with more negative affect or negative cognitive appraisals (Grych and Fincham, 1990; Kitzmann et al., 2003).

In short, families socialize children's behaviors, emotion expressions, beliefs, and attitudes. Family socialization takes place through interactions with genes. On the one hand, environmental inputs, such as family relationships, influence which genes are activated at different points in development. At the same time, many effects of the family environment are modified by genetically based differences between children (Repetti et al., 2011a). Our coverage of the family environment side of the equation is divided into three topic areas. First and foremost, we review the different qualities of family relationships that contribute to social and emotional development. The next two sections call attention to the importance of understanding the broader social ecology in which a family is embedded, and the diversity of different household structures.

Family Relationships

According to family systems theory, the family functions as a dynamic whole. However, most of what we know about the effects of the family on child development is based on the study of two-person dyads: the child's relationships with parents and with siblings, and the relationship between the child's parents. Evidence suggests that each dyad influences the other relationships as well as the functioning of each individual family member. Because it is through their social interactions that family members shape children's emotional and social development, our discussion is organized around these key family relationships – parent–child, sibling, and interparental. The section ends with a description of how child functioning may be shaped by certain stressful qualities of family relationships.

Parent–Child Relationships

This section describes how social and emotional behavior is socialized through interactions with parents, in particular through their provision of protection, support, warmth, affection, and sensitivity, and through their exertion of control and use of discipline. Two final subsections discuss the influence that parent psychological functioning and child behavior have on those interactions.

Parent Sensitivity and Warmth

When parents are warm, sensitive, and protective, children learn to trust adults and to regulate their feelings of distress and negative emotions. Emotion regulation, in turn, facilitates social development because negative emotions fuel antisocial behavior and interfere with the enactment of positive, prosocial behaviors; by regulating their emotional states children are better able to manage social challenges in family and peer settings. Parent caregiving, especially the protection and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23046-8

support they provide through responsiveness to signs of child distress, leads to secure or insecure attachments (Grusec and Davidov, 2007). A secure attachment is characterized by the child's receptiveness to the parent's provision of comfort, the child's tolerance of temporary separation, and his or her willingness to explore his surroundings. Trust in others and a sense of safety form the foundation for the development of social and emotional competence (Ainsworth, 1989). Parental warmth is defined as responsive, supportive, and affectionate parenting that is sensitive to the child's needs. Frequent expressions of approval and positive emotion are two indicators of parental warmth (Darling and Steinberg, 1993). Warmth generally promotes healthy social and emotional development in children, as indicated by children's understanding and regulation of their emotional states, their empathy, prosocial behavior, and level of acceptance by peers.

There are a number of ways that sensitive, warm, and protective parenting facilitates social and emotional development. The availability of a safe outlet for the expression of negative affect in a supportive context facilitates the development of children's emotion regulation skills. In addition, by observing the parent's compassionate behavior the child learns to regulate his or her own behavior and negative emotions and, in particular, learns about empathic responding to others. Empathy, in turn, motivates prosocial behavior and inhibits antisocial behavior (Grusec and Davidov, 2007). When parents are responsive to their needs and wishes, children learn how to share goals and accommodate another's needs. They become engaged in a system of mutual reciprocity, which fosters a cooperative stance in children and compliant behavior with parents, encourages the development of prosocial behavior and conscience (guilt following perceived wrongdoing), and discourages antisocial behavior (Grusec and Davidov, 2007). Dyadic synchrony, or the extent to which a parent–child dyad exhibits reciprocation of positive emotion, mutual responsiveness, and cooperation are associated with fewer behavioral problems and greater social competence in children (Harrist and Waugh, 2002).

Parents also socialize by providing direct guidance when children display distress (Morris et al., 2007). *Emotion coaching* includes labeling children's emotions, responding to children's emotion expressions with empathy and validation, and providing assistance with problem solving. Parents who are aware of their children's emotional states and view expressions of negative emotions as teaching opportunities are more likely to engage in coaching, in comparison to parents who feel discomfort when their children display distress. The children of parents who coach have better peer relationships and emotion regulation skills than those whose parents express disapproval, or are dismissive, of negative emotion.

Parent Control and Discipline

Children learn socially appropriate behavior through discussion and explanation, as well as through the parents' imposition of control and discipline, that is, through the experience of consequences for their actions. It is important to recognize that parents must enforce rules in the face of child resistance and desire for autonomy (Grusec and Davidov, 2007). Baumrind (1966) identified three iconic parent disciplinary styles. *Authoritarian* parents adhere to a strict set of guidelines about what constitutes appropriate behavior and attitudes; they emphasize obedience, not autonomy, engage in unilateral communication without opportunities for child feedback, and employ punitive measures in response to actions or attitudes that stray outside of the established guidelines. *Authoritative* parenting is characterized by high levels of emotional support, clear standards for children's behavior but bidirectional communication and autonomy-granting that allow children opportunities to influence family rules. *Permissive* parents are nonpunitive, accepting, and affirming of the child's affective states and behavior, and tend to place minimal emphasis on rules or demands. Authoritative parenting is associated with greater self-reliance in adolescent offspring, reduced rates of anxiety and depression, and less delinquent behavior, across several ethnic and socioeconomic boundaries (Steinberg et al., 1991). A mild level of parental control that preserves a child's sense of autonomy appears to be most effective in shaping behavior (Grusec and Davidov, 2010).

The effects of parenting behaviors and disciplinary styles on children's social and emotional development appear to vary depending on the child's temperament (level of fearfulness or irritability), age, and sex, as well as the nature of the child's misdeed and the emotion that the parent behavior conveys (e.g., love and concern vs hostility and rejection). In fact, the impact of parental control behaviors may depend less on the specific strategies that a parent uses and more on the child's interpretation of the parent's intent and the acceptability of that message. Sensitive and warm parenting that fosters trust increases the child's compliance, cooperation, and willingness to be guided by parent directives and to accept parental values (Grusec and Davidov, 2007; Maccoby, 2007).

Parent Psychological Functioning

A parent's psychological well being indirectly contributes to children's social and emotional development by helping to shape a family's social environment, particularly the quality of the parent–child relationship. Depression poses significant challenges for engaging in positive interactions with other family members. In mothers, depression compromises the ability to provide warm and sensitive care and is associated with the use of more negative discipline methods and with negative attribution biases regarding their children (Zahn-Waxler et al., 2002). The impact on parent–child interactions helps to account for the poor social and emotional outcomes that are often observed in children of depressed mothers; they are less skilled at regulating their negative emotions, which increases risk for developing internalizing problems (e.g., depression and anxiety), and externalizing problems (e.g., anger, hyperactivity, and impulsivity). Though paternal depression has been examined far less frequently than maternal depression, research suggests that it has a similar effect on children's risk of developing externalizing problems, but the association with internalizing problems may be weaker (Connell and Goodman, 2002). Similar to depression, parents with anxiety, schizophrenia, or antisocial personality disorders show less warmth and responsiveness toward their children. Parents with anxiety disorders also appear to experience

difficulty promoting autonomy in their children, and impose more control over their children's behaviors (Zahn-Waxler et al., 2002).

Child Contributions

Children are not passive members of their families. By triggering certain responses from others and actively seeking out particular situations and settings, children 'evoke' their own social worlds both inside and outside of the family (Grusec and Davidov, 2007). Because it is understood that parent–child interaction is transactional, with each member of the dyad continually influencing the behavior of the other, research on the parent–child relationship has shifted from a top-down perspective to one that emphasizes on the agency of the child as well as the parent (Maccoby, 2007). For instance, research involving children with disruptive behavior problems (e.g., conduct disorder) has found that the problematic behavior detracts from parents' psychological well being, as well as many aspects of parenting: responsiveness, parent–child communication, monitoring, and discipline. In fact, children's misbehaviors predicted subsequent parenting behavior more strongly than parenting predicted children's subsequent behavior (Pettit and Arsiwalla, 2008).

Sibling Relationships

In addition to parents, relationships with brothers and sisters, which are characterized by familiarity and often intense emotions, can also play an important part in social and emotional development. As Dunn (2007) has explained, it is through those familiar and intense interactions that children learn to anticipate another person's intentions and perspective, manipulate others' emotions, deceive, tease, manage conflict, and share an imaginative world in joint pretend. Siblings engage in conversations about why people behave the way that they do, and learn about mental states and feelings as causes of behavior. Sibling relationships can provide a social context for developing prosocial empathetic skills, social understanding, and emotion regulation, and they can serve as a training ground for hostile and aggressive behavior. Brothers and sisters can reinforce each other's aggression by fighting back and escalating conflict. Research indicates that conflict between siblings contributes to children's anxiety, depression, and delinquent behavior, independent of the influence of other family relationships. Developmental scholars have described a process of *sibling deidentification*, whereby siblings find different domains of interest in order to create separate and independent identities (Dunn, 2007).

The Interparental Relationship

So far we have discussed a child's own relationships with family members. However, in two-parent families, the relationship between the two parents has both direct and indirect effects on a child's social and emotional development. Coparenting refers to the way that the members of a couple support or undermine each other's parenting activities and how child care and child rearing responsibilities are apportioned. It is an important dimension of the relationship between two parents with implications for child development, particularly when there is a high level of daily hassles in the home, which appear to amplify elements of dysfunctional coparenting (Maccoby, 2007). A primary mechanism by which the interparental relationship is thought to act on child outcomes is via changes to parent–child interactions. For example, distress and conflict in the couple relationship have been found to be associated with a range of child adjustment problems; the link between couple hostility and adolescent internalizing and externalizing problems appears to be mediated by parental harshness and hostility toward the child, psychological intrusiveness and low acceptance, and parent–child conflict (Bradford et al., 2008). In the short-term, the likelihood of hostile parenting behavior and parent–child conflict rise on days when the parents engage in conflict with each other (Krishnakumar and Buehler, 2000). In the long-term, the effects of couple distress on conflicts with children have significant ramifications for children's emotional distress (i.e., anxiety and depression symptoms), a pattern that has been observed in families from European-American, Asian, and Latino backgrounds (Chung et al., 2009). Here, again, we see children playing an active role in family interactions. Negative emotional reactivity in children predicts increases in marital conflict, which in turn is predictive of children's behavior (Schermerhorn et al., 2007).

Stressful Family Relationships

A family social environment that is characterized by conflict and aggression and cold, unsupportive, or neglectful relationships poses risks for child development because it presents daily stressors to which children must repeatedly mount physiological/neuroendocrine and emotional/behavioral responses (Repetti et al., 2002). In this type of chronically stressful environment, it is particularly challenging to acquire adequate emotional and behavioral self-regulatory skills. In fact, children growing up in these 'risky families' show heightened physiological and emotional reactions to stressors, and use maladaptive coping strategies to regulate their responses. High levels of conflict at home sensitize children to anger; they react with greater distress, anger, anxiety, and fear. Periods of relief or respite from aroused emotional states may be critical to the proper regulation of emotion dynamics. However, chronic or repeated stressors, such as high levels of violence and family conflict, may not allow for sufficient recovery from heightened emotional arousal. The resulting sustained states of emotional arousal may, over time, increase reactivity, creating a sensitized stress-responsive system. Moreover, children from families characterized by conflict, overt anger, hostility, and abuse are more likely to maintain a heightened state of vigilance for signs of anger and are also more likely to respond with distress, anger, and aggression to incidents that involve anger, discord, and tension. Compared to other children, they appear to be hyperprepared for threats and hyperresponsive to those threats, which may contribute to increased reactivity (Repetti et al., 2011b).

Family members' negative mood and physiology can compound the impact that risky families have on child development. Through processes of emotion transmission

(short-term links between one member's momentary negative mood and another member's concurrent or subsequent emotional experience) and physiological coregulation (e.g., linkages between family members' day-to-day fluctuations in levels of the hormone, cortisol, which is associated with stress), the mood and physiology of one family member can be affected by the mood and physiology of another family member (Repetti et al., 2011b).

The Social Ecology of the Family

We shift here from a focus on relationships to factors that shape family life. According to Bronfenbrenner's (1989) social ecological model, the influence that families have on child development can only be understood by taking into account the broader social milieu in which families are embedded. Three dimensions of that larger world and their influence on the family social context are discussed next: culture, socioeconomic status (SES), and stressors from both inside and outside the home.

Culture

Child-rearing practices and patterns of family interaction vary in different cultures. For example, ethnographic studies show that sibling caregiving is much more widespread and significant in many cultures than it is in the United States. In many non-Western communities, these sibling experiences play an important role in socialization for parenthood (Dunn, 2007). Researchers have also identified (1) differences in the parenting styles typically exhibited in different cultures, as well as (2) interactions between parenting behaviors and culture in child social and emotional outcomes. With respect to the first point, middle class European-American families – that is, parents in an individualistic culture that prizes autonomy – most often exhibit authoritative parenting styles, whereas parents from collectivist cultures (e.g., Indian and Egyptian) more often exhibit authoritarian parenting (Rudy and Grusec, 2006). Authoritarian parenting also predicts more fearful, compliant behavior among European-American girls but more assertive behavior among African-American girls (Darling and Steinberg, 1993).

Studies of differences between authoritarian and authoritative parents in collectivist and individualist cultures suggest that authoritarianism has detrimental effects on child outcomes only in individualist cultures, through the mechanism of higher maternal negative emotion and cognition (Rudy and Grusec, 2006). One possible explanation for these findings is that Baumrind's parenting styles (1966) do not entirely capture the 'how' of parent–child interactions: parents may be more obedience-oriented, but this does not necessarily determine whether discipline is delivered with harshness or nurturance. Typical Chinese parenting, for instance, has been described as authoritarian 'training' delivered with involvement, care, and concern for the child's welfare (Chao, 1994), which leads to more positive child outcomes than the authoritarian parenting exhibited in European-American families. However, there is research to suggest that even when examining child outcomes in individualist vs collectivist cultures, authoritative parenting is still associated with more positive social and emotional outcomes than authoritarian or permissive parenting (Sorkhabi, 2005). This is certainly a topic that requires more empirical investigation.

Socioeconomic Status

SES is an important dimension of a family's social ecology. Parents' expectations, values, styles, and behaviors toward their children vary with the family's SES (Hoff et al., 2002). In general, higher-SES parents tend to place a higher value on independence and self-direction for their children. There are corresponding differences in parenting styles and behaviors: lower-SES parents tend to show more authoritarian and permissive styles of parenting, whereas higher-SES parents exhibit the authoritative style more frequently (Kohn, 1969). Overall, the latter group emphasizes democracy and child-centered styles of parenting, whereas lower-SES groups tend to practice more parent-centered styles. As such, lower-SES parents promote less autonomy and cooperation with their children. In addition to these differences in parenting style, lower-SES parents are less able to monitor their children or be as involved with their children's homework and extracurricular activities, in comparison to higher-SES parents. These differences may be in part attributable to extrafamilial stressors that are more common in lower-SES families, such as higher work stress, lack of neighborhood resources, and financial strains.

Economic resources determine what type of nutrition and health care children receive, the quality of a family's housing, as well as the child's exposure to toxins, such as air pollution, and community violence. Low-income families are less able to provide children with high quality educational opportunities and enriching activities, such as books, music, and art, or with access to museums, clubs, and sports teams. The stress of economic hardship and job loss undermine parents' abilities to respond in nurturant ways to their children. Children growing up in impoverished families are more likely to suffer physical abuse, verbal abuse, and neglect, to experience family instability, and to be exposed to poor socialization practices at home, including parenting that is more coercive and less authoritative, warm, and engaged (Patterson and Hastings, 2007; Repetti et al., 2002).

The Impact of Chronic and Daily Stressors on Family Life

Chronic life stress – whether due to the demands of economic strain, marital problems, difficulties at work, or health problems – influences how parents interact with children and their attitudes about parenting. In general, more stressed parents tend to display less positive and more negative affect, less responsiveness, and less consistent discipline toward their children. Parents under chronic stress tend to hold higher and less developmentally appropriate expectations for their children, and view their children as displaying more anger (Crnic and Low, 2002; Morrison Gutman et al., 2005). This observation (that parents who report high levels of chronic stress also exhibit more psychological distress and depressive symptoms) explains how stress is linked to parenting behaviors.

On a day-to-day basis, parenting behavior can change according to events that occur outside the parent–child relationship. Daily stressors such as feelings of being overloaded at

work, interpersonal difficulties at work, and disruptions in the interparental relationship (e.g., marital conflict) affect how parents interact with their children. For example, when parents experience stressful days at work, they are more likely to engage in either social withdrawal – defined as reductions in expressions of emotion, such as affection, and social engagement – or in conflictual or angry interactions with spouses and with children (Repetti et al., 2009). This process is known as 'spillover' because stress in one context (in this case, work) is 'spilling over' into another context (in this case, family interactions; Repetti, 1987).

Family Structure and Diversity

Finally, we note that many children today are reared by single parents, never-married couples, lesbian and gay couples, grandparents, step-parents, adoptive parents or, thanks to advances in alternative reproductive technologies, such as donor insemination and egg donation, by parents to whom they have no genetic tie. Research examining these different family forms indicates that the family resources, social processes, and relationships covered above are much more important for children's social and emotional development than is the family's structure (Patterson and Hastings, 2007). However, economic resources are unevenly distributed across different family structures. Two parents, compared to one, have access to greater economic resources, and have more extensive social support networks; plus two parents can provide support and encouragement for one another.

Increased parenting stress, less access to resources and social support, and the lower quality of parenting that is sometimes found in single-parent families are risk factors for difficulties in children's socioemotional development. Single mothers report more dysfunctional interactions with their children, including more severe disputes, less enjoyment of motherhood, lower overall quality of parenting and more parenting stress, than do mothers with partners (whether opposite or same-sex; Patterson and Hastings, 2007).

Although the mechanisms are not clear, the offspring of heterosexual parents who are married show better adjustment than the offspring of unmarried heterosexual parents. Children with lesbian or gay parents develop in much the same ways as other children. No differences have been observed in their social, emotional, or sexual development, and the predictors of development in children reared by heterosexual and gay parents are the same: child rearing behaviors, parenting style, family management. In general, the genetic relatedness of parent and child and parental sexual orientation do not seem to be important for children's social and emotional development (Patterson and Hastings, 2007).

Conclusions

A tremendous amount of social and emotional development takes place in the context of family relationships. The family provides the social partners and experiences through which children learn about trust, intimacy, empathy, control, and the complex rules of social interaction; the role models for behavior and emotion expression; the 'culture brokers' who introduce children to ways of acting and knowing; and the daily challenges and stressors to which children must learn to respond and cope. Those experiences, in turn, reflect and are shaped by the larger social world in which the family exists. Although our discussion has focused on childhood, the family's influence begins in infancy, and continues through adulthood, when relationships with a new family created through marriage and parenthood, as well as enduring relationships with the family of origin, have profound effects on emotional and social well being (Repetti et al., 2011a).

See also: Early Emotional Development and Cultural Variability; Emotional Development, Effects of Parenting and Family Structure on; Neonatal Social Isolation: Effects on Neural Function.

Bibliography

Ainsworth, M.S., 1989. Attachments beyond infancy. American Psychologist 44 (4), 709–716. http://dx.doi.org/10.1037/0003-066X.44.4.709.

Baumrind, D., 1966. Effects of authoritative parental control on child behavior. Child Development 37 (4), 887–907. http://dx.doi.org/10.2307/1126611.

Bradford, K., Vaughn, L.B., Barber, B.K., 2008. When there is conflict: interparental conflict, parent–child conflict, and youth problem behaviors. Journal of Family Issues 29 (6), 780–805. http://dx.doi.org/10.1177/0192513X07308043.

Bronfenbrenner, U., 1989. Ecological systems theory. Annals of Child Development 6, 187–249.

Chao, R.K., 1994. Beyond parental control and authoritarian parenting style: understanding Chinese parenting through the cultural notion of training. Child Development 65 (4), 1111–1119.

Chung, G.H., Flook, L., Fuligni, A.J., 2009. Daily family conflict and emotional distress among adolescents from Latin American, Asian, and European backgrounds. Developmental Psychology 45 (5), 1406.

Connell, A.M., Goodman, S.H., 2002. The association between psychopathology in fathers versus mothers and children's internalizing and externalizing behavior problems: a meta-analysis. Psychological Bulletin 128 (5), 746–773. http://dx.doi.org/10.1037/0033-2909.128.5.746.

Crnic, K., Low, C., 2002. Everyday stresses and parenting. In: Handbook of Parenting: Volume 5: Practical Issues in Parenting, second ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 243–267.

Darling, N., Steinberg, L., 1993. Parenting style as context: an integrative model. Psychological Bulletin 113 (3), 487–496. http://dx.doi.org/10.1037/0033-2909.113.3.487.

Denham, S.A., Mitchell-Copeland, J., Strandberg, K., Auerbach, S., Blair, K., 1997. Parental contributions to preschoolers' emotional competence: direct and indirect effects. Motivation and Emotion 21 (1), 65–86. http://dx.doi.org/10.1023/A:1024426431247.

Dunn, J., 2007. Siblings and socialization. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization: Theory and Research. Guilford Press, New York, NY, pp. 309–327.

Ellis, D.A., Zucker, R.A., Fitzgerald, H.E., 1997. The role of family influences in development and risk. Alcohol Health and Research World 21 (3), 218–226.

Erel, O., Burman, B., 1995. Interrelatedness of marital relations and parent-child relations: a meta-analytic review. Psychological Bulletin 118 (1), 108–132. http://dx.doi.org/10.1037/0033-2909.118.1.108.

Grusec, J.E., Davidov, M., 2007. Socialization in the family: the roles of parents. In: Grusec, J., Hastings, P. (Eds.), Handbook of Socialization. Guilford Publications, New York, NY, pp. 284–308.

Grusec, J.E., Davidov, M., 2010. Integrating different perspectives on socialization theory and research: a domain-specific approach. Child Development 81 (3), 687–709. http://dx.doi.org/10.1111/j.1467-8624.2010.01426.x.

Grych, J.H., Fincham, F.D., 1990. Marital conflict and children's adjustment: a cognitive-contextual framework. Psychological Bulletin 108 (2), 267–290.

Harrist, A.W., Waugh, R.M., 2002. Dyadic synchrony: its structure and function in children's development. Developmental Review 22 (4), 555–592. http://dx.doi.org/10.1016/S0273-2297(02)00500-2.

Hoff, E., Laursen, B., Tardif, T., 2002. Socioeconomic status and parenting. In: Handbook of Parenting: Volume 2: Biology and Ecology of Parenting, second ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 231–252.

Kitzmann, K.M., Gaylord, N.K., Holt, A.R., Kenny, E.D., 2003. Child witnesses to domestic violence: a meta-analytic review. Journal of Consulting and Clinical Psychology 71 (2), 339–352. http://dx.doi.org/10.1037/0022-006X.71.2.339.

Kohn, M., 1969. Class and Conformity. Dorsey Press, Homewood, IL.

Krishnakumar, A., Buehler, C., 2000. Interparental conflict and parenting behaviors: a meta-analytic review. Family Relations: An Interdisciplinary Journal of Applied Family Studies 49 (1), 25–44.

Maccoby, E.E., 2007. Historical overview of socialization research and theory. In: Grusec, J.E., Hastings, P. (Eds.), Handbook of Socialization: Theory and Research. The Guilford Press, New York, NY, pp. 13–41.

Morris, A.S., Silk, J.S., Steinberg, L., Myers, S.S., Robinson, L.R., 2007. The role of the family context in the development of emotion regulation. Social Development 16 (2), 361–388. http://dx.doi.org/10.1111/j.1467-9507.2007.00389.x.

Morrison Gutman, L., McLoyd, V.C., Tokoyawa, T., 2005. Financial strain, neighborhood stress, parenting behaviors, and adolescent adjustment in urban African American families. Journal of Research on Adolescence 15 (4), 425–449. http://dx.doi.org/10.1111/j.1532-7795.2005.00106.x.

Patterson, C.J., Hastings, P.D., 2007. Socialization in the context of family diversity. In: Grusec, J.E., Hastings, P. (Eds.), Handbook of Socialization: Theory and Research. The Guilford Press, New York, NY, pp. 328–351.

Pettit, G.S., Arsiwalla, D.D., 2008. Commentary on special section on "bidirectional parent–child relationships": the continuing evolution of dynamic, transactional models of parenting and youth behavior problems. Journal of Abnormal Child Psychology 36 (5), 711–718. http://dx.doi.org/10.1007/s10802-008-9242-8.

Repetti, R.L., 1987. Linkages between work and family roles. Applied Social Psychology Annual 7, 98–127.

Repetti, R., Flook, L., Sperling, J., 2011a. Family influences on development across the life span. In: Fingerman, K.L., Berg, C.A., Smith, J., Antonucci, T.C. (Eds.), Handbook of Life-Span Development. Springer Publishing, New York, pp. 745–775.

Repetti, R.L., Robles, T.F., Reynolds, B., 2011b. Allostatic processes in the family. Development and Psychopathology 23 (3), 921–938.

Repetti, R.L., Taylor, S.E., Seeman, T.E., 2002. Risky families: family social environments and the mental and physical health of offspring. Psychological Bulletin 128 (2), 330–366.

Repetti, R.L., Wang, S., Saxbe, D., 2009. Bringing it all back home: how outside stressors shape families' everyday lives. Current Directions in Psychological Science 18 (2), 106–111.

Rudy, D., Grusec, J.E., 2006. Authoritarian parenting in individualist and collectivist groups: associations with maternal emotion and cognition and children's self-esteem. Journal of Family Psychology 20 (1), 68.

Schermerhorn, A.C., Cummings, E.M., DeCarlo, C.A., Davies, P.T., 2007. Children's influence in the marital relationship. Journal of Family Psychology 21 (2), 259–269. http://dx.doi.org/10.1037/0893-3200.21.2.259.

Shipman, K.L., Zeman, J., 2001. Socialization of children's emotion regulation in mother–child dyads: a developmental psychopathology perspective. Development and Psychopathology 13 (2), 317–336.

Sorkhabi, N., 2005. Applicability of Baumrind's parent typology to collective cultures: analysis of cultural explanations of parent socialization effects. International Journal of Behavioral Development 29 (6), 552–563. http://dx.doi.org/10.1080/01650250500172640.

Steinberg, L., Mounts, N.S., Lamborn, S.D., Dornbusch, S.M., 1991. Authoritative parenting and adolescent adjustment across varied ecological niches. Journal of Research on Adolescence 1 (1), 19–36.

Zahn-Waxler, C., Duggal, S., Gruber, R., 2002. Parental psychopathology. In: Handbook of Parenting: Volume 4: Social Conditions and Applied Parenting, second ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 295–327.

Social Beliefs and Attitudes: Personality and the Isms

Jeremy Burrus, ACT, Iowa City, IA, USA
Lauren Carney, Educational Testing Service, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social attitudes and beliefs can have a powerful influence on how one interprets, and respond to, everyday events. Employing a factor analytic approach, Saucier (2000, 2013) has recently developed a taxonomy of social attitudes and beliefs which contains five dimensions. He labeled these dimensions 'isms' because words describing these attitudes and beliefs tend to end in 'ism' (e.g., fundamentalism, relativism, and materialism). The five dimensions of isms are: tradition-oriented religiousness, unmitigated self-interest, communal rationalism, subjective spirituality, and inequality-aversion. In this article we briefly discuss beliefs and attitudes more generally before discussing isms in detail. We then discuss the relationship of Big Five personality dimensions to isms and additional correlates of isms before providing a few areas for future research directions. The article concludes by making an argument that isms are a potentially very important, but highly under-researched, aspect of human individual differences.

Each of us adheres to a set of beliefs and attitudes that, at least in theory, guide the way we interpret events and respond to them. We are consciously aware of most of these beliefs. For instance, when I get into my car in the morning to go to work, I believe that turning the key will start the engine. And so, faithfully, every morning I turn the key, expecting my car to start. Almost invariably, it does. Only when the engine does not start do I start to question the veracity of this belief. Other beliefs are held more implicitly. For example, while some implicitly believe that each of us is born with a fixed amount of intelligence (an 'entity' theory), others believe that intelligence is malleable and can be increased with practice (an 'incremental' theory; Dweck et al., 1995). As another example, people have a tendency to believe that something is more likely to be true the more often they have heard it (e.g., Allport and Lepkin, 1945). At times, these beliefs can have an evaluative quality. I can believe that my soul will enter the afterlife after I die, and I can also believe that it is *good* for my soul to enter the afterlife after I die. Typically, the evaluative components of beliefs are referred to as 'attitudes.'

A particular set of beliefs and attitudes, "… whose referents have shared general societal relevance to many people in religious, economic, political, educational, ethnic, and other social areas" (Kerlinger, 1972: p. 614, as cited in Saucier, 2000) can be referred to as 'social attitudes and beliefs.' For example, people who believe in egalitarianism believe that all people should be treated equally, regardless of their status in the social hierarchy. As such, they will tend to treat all people in a similar fashion. In everyday conversation, we use a range of words to describe such different belief systems, that could be used to describe social (e.g., egalitarianism and ethnocentrism), political (e.g., liberalism and conservatism), and religious (e.g., spiritualism and religionism) attitudes or beliefs, all of which might affect the way we behave. Note that each of the example social attitude or belief ends in 'ism.' Thus, these terms were labeled 'isms' by Saucier (2000).

In the following section, we delve deeper into the origin and structure of beliefs and attitudes. Next, we provide more detail specifically on social attitudes and beliefs, focusing on Saucier's isms dimensions. We then discuss several issues surrounding isms, including their relation to Big Five personality dimensions, their relationship to other constructs, and future directions for research involving isms. Although some empirical findings are discussed, because isms are a relatively understudied construct, much of this discussion relies heavily on theory.

Beliefs

A belief is how sure one is that knowledge acquired (e.g., the Earth revolves around the sun) or a subjective experience (e.g., last night's party was the most fun night of my life) is true (Wyer and Albarracin, 2005). Beliefs can vary in terms of strength, such that the surer one is that a proposition is true, the stronger the belief. They can refer to objects or events from the past (e.g., Facebook was founded on 4 February 2004), the present (e.g., my wife is 5′ 9″), or the future (e.g., all cars will be driverless in 15 years). Beliefs about the future can also be thought of as 'expectations' (Wyer and Albarracin, 2005).

As stated briefly above, beliefs can also be held explicitly or implicitly. On the one hand, explicitly held beliefs are those which one is aware of holding. For example, I have a very strong explicit belief that I had scrambled eggs and coffee for breakfast this morning. On the other hand, unless brought to conscious awareness, one is generally unaware of holding implicit beliefs. Oftentimes, these beliefs consist of basic assumptions regarding the workings of the mind and the world in general, and are sometimes referred to as implicit, or naive, theories (e.g., Dweck et al., 1995; Schwarz, 2004; Wyer and Albarracin, 2005). The entity versus incremental theories of intelligence example from above is one example of an implicit theory. Others include the naive theory that written material that is easier to process (e.g., is written clearly rather than blurrily) has been seen before (e.g., Schwarz, 2004), and that behavior has occurred more frequently if examples of it can be easily brought to mind (Schwarz et al., 1991). An important point in distinguishing explicit versus implicit beliefs is that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.25082-4

implicit beliefs can influence our judgment, and subsequent behavior, without our awareness.

Beliefs can be based in cognition or motivation. Clearly many of the examples of beliefs discussed above are cognitive in nature. They refer to the probability that knowledge acquired through experience about oneself or others, events, places, or objects, is true or correct. People may also hold beliefs because they are motivated to do so. Although people might hold a variety of motivations, a few basic ones have been identified in the psychological literature. For example, people strive to interact meaningfully with other individuals ('communion striving'), to have influence over others ('status striving'), to have control over their lives ('autonomy striving'), and to feel competent ('achievement striving'; see Barrick et al., 2013 for a review). Additionally, people have the need to feel like the world is fair and just (Lerner and Miller, 1978), and also the need to reconcile thoughts of their own impending deaths (Greenberg et al., 1986). Individuals vary on their level of each of these needs; one might have a very low need for communion striving but a high need for achievement striving, whereas another person might demonstrate the opposite profile. Beliefs can be generated that can help ease one's concerns about each of these motivations. For example, the belief, "Even though I have few friends, the friends I do have are very close to me." might be generated to help one feel he or she is achieving progress toward fulfilling the communion striving need. Additionally, the belief, "People have souls that are separate from their physical bodies" might be generated to help ease one's fear of death.

Attitudes

An attitude is an evaluation of a person, object, or event (Eagly and Chaiken, 1993). They consist of an experiential and an instrumental component (Ajzen, 2006). The experiential component represents the valence of the entity and can be measured with scales ranging from *good–bad*, to *pleasant–unpleasant*, to *enjoyable–unenjoyable*. The instrumental component represents the utility of an entity and can be measured with scales ranging from *useful–useless*, to *beneficial–harmful*, and to *valuable–worthless*. Thus, it can be said that I have a positive attitude toward math if I enjoy the experience of doing math, and/or if I feel that math is useful and valuable (e.g., Lipnevich et al., 2011).

Although the relationship of beliefs to attitudes is not necessarily agreed upon (Wyer and Albarracin, 2005), one prominent theory of attitudes states that attitudes develop from beliefs (Ajzen, 1991). That is, beliefs about a person, object, or event become linked to a positive or negative attribute to form an attitude. Because I am pretty sure that I was born, I have a very strong belief that I indeed have a birthday. My attitude toward my birthday, however, may change with age. When I am young, I associate my birthday with positive features (parties and presents), and so I have a positive attitude toward my birthday. Later in life, however, I may begin to associate my birthday with negative features (getting older and thus closer to death), and my attitude toward my birthday may turn negative. These positive or negative associations can come from personal experience or from witnessing (or learning about) others' experiences (e.g., "Social Learning Theory," Bandura, 1977).

If attitudes are truly derived from beliefs, it follows that they, too, have both cognitive and motivational bases. We can develop an attitude because we know (or think we know) something, or we can develop an attitude because we want (or need) to believe something. Similarly, attitudes can be explicitly and implicitly held (Nosek, 2007). As with beliefs, an explicit attitude is one that is aware of holding and can consciously report on. In contrast, one is not always aware of holding an implicit attitude and thus he or she cannot consciously report on them. In fact, they are usually activated automatically. Typically, they are measured with the Implicit Association Test (IAT), which measures the strength of associations between two sets of concepts (e.g., good vs bad; black vs white). Explicit attitudes often agree with implicit attitudes (Nosek, 2007) but sometimes diverge. For example, people often explicitly claim no racial favoritism, although results from the IAT find that they implicitly favor whites to blacks (Nosek et al., 2002).

Social Beliefs and Attitudes: Isms

As briefly described above, social attitudes and beliefs have been coined "Isms" by Saucier (2000). Prior to the year 2000, several social attitudes and beliefs had been identified in the research literature (e.g., Adorno et al., 1950); however, no attempt was made to create a taxonomy of such attitudes and beliefs until Saucier's seminal work.

In an attempt to find a simplified structure within such a wide range of attitudes and beliefs, Saucier (2000) factor-analyzed 266 isms using a population of college students. He found four dimensions of isms, initially labeling them with Greek letters to avoid using misleading terms in categorizing such broad factors. He later felt more comfortable in providing more conventional labels to these dimensions. Saucier (2000) provides a detailed description of the first four dimensions. His descriptions are summarized below and additional examples and definitions are included in **Table 1**.

'Alpha,' or 'tradition-oriented religiousness,' is described as the degree to which an individual reveres religious authority and believes in, and supports, conventional and traditional forms of religion. Cultural traditions are important for those who score high on alphaisms. This value may result in many behavioral manifestations, including a willingness to behave aggressively in the name of traditional sources of authority. It would follow, for example, that many of those who volunteered to participate in the Crusades in the late Middle Ages in the name of the Catholic Church would have scored high on alpha. Taking an example from more recent events, we would hypothesize that the political unrest in the Middle East, and the associated aggressive actions taken, is highly motivated by alphaisms. An example item used to measure alpha is, "Religion should play the most important role in civil affairs."

'Beta,' or 'unmitigated self-interest,' measures the degree to which someone justifies different forms of self-interest, such as materialism or forms of physical pleasure. According to Saucier (2000), it concerns attitudes and beliefs about one's immediate environment. While many alphaisms focus on the past,

Table 1 Example isms and their definitions

Alpha (tradition-oriented religiousness)	
Calvinism	God is all-powerful, and those whom God chooses will be saved by God's grace alone.
Secularism	Religious considerations should be excluded from civil affairs and public education.
Triumphalism	Moral laws are fixed, absolute, and the same everywhere.
Beta (unmitigated self-interest)	
Solipsism	The self is the only thing that can be known and verified.
Materialism	Physical well-being and worldly possessions are the greatest good and highest value in life.
Ethnocentrism	I believe in the superiority of my own ethnic group.
Gamma (communal rationalism)	
Nationalism	I am devoted to one particular nation.
Hedonism	Behavior is motivated by the desire for pleasure and the avoidance of pain.
Fundamentalism	I adhere firmly to fundamental and basic principles.
Delta (subjective spirituality)	
Animism	There are spiritual beings separate, or separable, from bodies.
Anthropomorphism	Many nonhuman things have human motivation and human characteristics.
Pragmatism	I adhere to and am dedicated to the facts.
Epsilon (inequality-aversion)	
Classism	I prefer people of a certain social or economic class.
Communalism	I believe in communal (group or community) ownership of goods and property.
Anarchism	All forms of government are oppressive and undesirable and should be abolished.

Reproduced from Saucier, G., 2000. Isms and the structure of social attitudes. Journal of Personality and Social Psychology 78, 366–385; Saucier, G., 2013. Isms dimensions: toward a more comprehensive and integrative model of belief-system components. Journal of Personality and Social Psychology 104, 921–939.

betaisms tend to be present focused. We hypothesize that an individual high on betaisms would be more likely to sacrifice potential future gains for present ones than an individual low on betaisms. As such, for example, we would conjecture that most environmental activitists would score low on betaisms. One example item is, "The pleasures of the senses are the highest good."

'Gamma,' or 'communal rationalism,' stresses the support of a nation and a common institution which guarantees individual freedom and which emphasizes the use of reason. This dimension covers isms such as patriotism, humanism, and functionalism. We believe that the prototypical coach of a team sport provides an apt example of a person high on gamma. A quick scan of the gamma examples in **Table 1** makes this clear. 'Nationalism' (I am devoted to one particular nation) can be retranslated to 'teamism' (I am devoted to one particular team). 'Hedonism' (Behavior is motivated by the desire for pleasure and the avoidance of pain) is displayed in coaches who motivate players by inducing pain to punish player mistakes (e.g., yelling at them and making them run extra laps). Finally, 'fundamentalism' (I adhere firmly to fundamental and basic principles) is clearly a part of coaching, as one important job of the coach is to teach a sport's basic fundamentals and principles. An example gamma item is, "I believe in government by law with the consent of those people governed."

'Delta,' or 'subjective spirituality,' is the degree to which a person values subjective spiritual or 'paranormal' experiences. Though some psychologists tend to group religion (or alphaisms) and spirituality (deltaisms) into one factor, evidence shows these are two distinct factors (Saucier, 2000, 2013; Saucier and Skrzypińska, 2006). Religious beliefs place emphasis on the opinions of an organized group of people, while spiritual beliefs are formed based on subjective experiences and individual feelings. Saucier (2000) implies that delta has roots in Indian Hinduism, and, in American tradition, Transcendentalism. An important aspect of delta is the belief that divinity is not exclusive to humans; nature also has divine aspect. An example delta item is, "Enlightenment is gained through meditation, self-contemplation, and intuition."

Later, Saucier (2013) again factor-analyzed a large collection of isms, but this time with a wider community sample. He replicated the four factors of alpha, beta, gamma, and delta. However, there was also evidence of a fifth factor, 'epsilon,' or 'inequality-aversion,' relating to egalitarianism. "One pole of the dimension is anchored by item clusters endorsing communal ownership, social welfare programs, and a socialist society; the other pole involves views supportive of domination of one class, group, or nation over others (e.g., elitism, chauvinism, and jingoism). This dimension is concerned foremost with attitudes toward inequalities within one's own society, but this extends secondarily to inequalities between societies" (Saucier, 2013: p. 931). An example item is, "Free market capitalism leads to gross inequalities in income and wealth, which is a great social evil." Thus, the current Survey of Dictionary-based Isms (SDI), developed by Saucier, contains items that measure all five of the dimensions.

It should be clear from the examples provided that isms are indeed forms of beliefs (e.g., Solipsism: "The self is the only thing that can be known and verified.") and attitudes (e.g., Materialism: "Physical well-being and worldly possessions are the greatest good and highest value in life."). As such, it follows that if beliefs and attitudes can be held both explicitly and implicitly, then so should isms. We posit that, if the explicit/implicit distinction is thought of as a continuum, most isms would fall closer to the implicit side of the pole. It seems clear that many isms are influenced by, or are indeed products of; ideas imbued by one's culture that drive one's view of the world.

One definition states that culture has three components that are learned and shared by a group of people: "(1) The *things* people make and use (called artifacts), (2) the ways people *behave* (including language), and (3) the *ideas* people have" (e.g., Caligiuri et al., 2011: p. 2). These components are thought to be relatively stable in nature, although they can change depending on circumstances (Caligiuri et al., 2011). Culture is often explained using the iceberg metaphor (e.g., Herman, 1970), such that most of culture's influence in everyday life is unseen or 'below the surface of the ocean.' The part that can be seen or is above the surface includes the norms, artifacts, and behaviors of specific cultures. Just below the surface are the values and interests held by specific cultures. Finally, deeper below the surface still are the assumptions held by specific cultures. These 'below the surface' aspects of culture appear to be synonymous with isms. It is thought that cultural assumptions influence values and interests, which then influence norms,

artifacts, and behaviors. Isms may represent some combination of culturally influenced assumptions, values, and interests.

The likely relation of isms to cultural values and assumptions calls to mind other taxonomies of cultural values. For instance, Hofstede (1983) has identified several dimensions upon which cultures vary. These include: 'power distance' (acceptance of unequal power); 'uncertainty avoidance' (tolerance for ambiguity); 'individualism versus collectivism' (the degree that people are oriented toward groups); and 'masculinity' (the degree to which men and women's values differ in a society). Another prominent theory, Schwartz's (1992) cultural values, identifies several values that can be organized into two factors: 'organization versus change' and 'self-enhancement versus self-transcendence.' To the best of our knowledge, the relation of isms to Hofstede's dimensions has not been investigated. Saucier (2013) did investigate the relation of isms to Schwartz's dimensions and found that three of the isms dimensions fell within the Schwartz two-dimensional structure, with subjective spirituality and communal relationism falling outside of the structure.

One possible way that isms differ from the beliefs and attitudes discussed above is that, whereas beliefs and attitudes in general vary as to the extent to which they are cognitively or motivationally determined, isms seem to be primarily motivational in nature. That is, each of the isms provided in the examples in this article (see Saucier, 2000, 2013 for a complete list) can be said to speak to one of the fundamental motives described above. For example, to conjecture from the examples listed in **Table 1**:

- *Communalism* (I believe in communal (group or community) ownership of goods and property) relates to communion striving
- *Classism* (I prefer people of a certain social or economic class) relates to status striving
- *Solipsism* (the self is the only thing that can be known and verified) relates to autonomy striving
- *Materialism* (physical well-being and worldly possessions are the greatest good and highest value in life) relates to achievement striving
- *Triumphalism* (moral laws are fixed, absolute, and the same everywhere) relates to the need to feel the world is fair and just
- *Animism* (there are spiritual beings separate, or separable, from bodies) relates to the need to reconcile thoughts of one's own death

If the isms are indeed related to one's basic motivations and needs, this suggests that manipulating these motivations (e.g., making thoughts of death more accessible) should lead to a change in one's endorsement of corresponding isms. Although we do not know of any research that speaks directly to this issue, some naturalistic research does speak indirectly. For example, in the days immediately following the terrorist attacks of 9/11 (a clear example of increasing the salience of one's death) 74–82% of Americans showed their patriotism by displaying the American flag on their home, car, or person (Morgan et al., 2011). This percentage had decreased to 56% by 2003.

Although we speculate that one's endorsement of isms can be manipulated, and thus are malleable, isms are typically thought of as relatively stable individual differences. This may come as a surprise to some given the large literature in social psychology that documents techniques for persuasion and attitude change (e.g., Cialdini, 1993). However, certain types of attitudes have shown considerable stability. Such is the case with vocational interests; attitudes about activities involving goal-related behaviors (Rounds and Su, 2014). Meta-analysis has found that interests stabilize as early as age 12, and remain stable at least through middle adulthood (Low et al., 2005). In fact, they appear to be more stable than personality. One commonality we see between isms and vocational interests are their high relevance to the self. It is likely that highly self-relevant attitudes are more accessible (easier to bring to mind) than less self-relevant attitudes (e.g., Fazio and Olson, 2007). As such, they influence one's interpretation of daily life events and, as they do so, increase in strength over time. The fact that isms have been identified by Saucier as individual differences, and the likelihood that they are relatively stable, may suggest to some that isms might simply be another way to label the Big Five personality traits. This is discussed briefly in the following section.

Isms and Personality

Many have questioned the ways in which personality and belief systems overlap; some believed that various beliefs could be just another dimension of personality (Adorno et al., 1950; Homer and Kahle, 1988). Saucier (2013) found that each of the five ism dimensions significantly correlated with at least one of the Big Five factors, yet they did not correlate higher than $r = 0.25$. This suggests that the isms dimensions do have important associations with the Big Five, but are actually unique from these personality factors. For example, people who were less conscientious were more likely to endorse subjective spiritualism (however, there was no significant correlation between conscientiousness and endorsing traditional religions). People who were more agreeable were more likely to endorse inequality-aversion and be more egalitarian, as one might expect. Openness was found to be negatively correlated with unmitigated self-interest, while people who were less emotionally stable were more likely to endorse traditional religions. Saucier (2013) also examined change in personality characteristics across the span of 10 years and found that unmitigated self-interest was correlated with a negative change in agreeableness, conscientiousness, emotional stability, and openness.

Isms, thus, represent one alternate conceptualization of personality that focuses on individual differences in social attitudes and beliefs. Importantly, the magnitude of correlations of isms and the Big Five personality factors suggests that they do not completely overlap with the Big Five. This suggests that isms might act as an important complement to the Big Five for both illuminating the nature of individual differences and predicting important outcomes.

Other Correlates of Isms

The relative paucity of research on isms means that very little work has been conducted investigating isms's correlates.

Recently, however, Saucier (2013) has examined the relation of isms to previously used measures of social attitudes and beliefs, political preference, and subjective well-being (SWB). He found that several isms correlated significantly with previous measures of social attitudes and beliefs. To provide few examples of correlations that exceeded 0.20, tradition-oriented religiousness and inequality-aversion correlated with right-wing authoritarianism (social conservatism) at 0.76 and −0.29 respectively, and, unmitigated self-interest and inequality-aversion correlated with social dominance orientation (a preference for hierarchy) at 0.22 and −0.54, respectively. Thus, it could be argued that tradition-oriented religiousness and inequality-aversion overlap to such an extent with previously identified constructs so as to be redundant. This is especially the case for tradition-oriented religiousness.

Interestingly, only one ism was found to be significantly correlated with measures of subjective well-being, communal rationalism, which suggests that people who believe in the importance of a fair and just government are also happier. The relationship was weak, however, ($r = 0.13$). Given the oft-replicated finding that those who participate in religious activities tend to have higher SWB (Diener et al., 1999), it is surprising that people who endorsed traditional religious or spiritual beliefs were not found to be significantly happier. Finally, isms were related to political preference, with republicans more likely to be high on tradition-oriented religiousness, and democrats more likely to be high on communal relationism, subjective spirituality, and inequality-aversion.

Future Directions

There are numerous avenues for future research on isms. One important area will be to investigate the development and temporal stability of isms. Presumably, given the discussion of above, one strong influence of ism development is one's culture. However, other influences likely come into play. One obvious influence should be religious upbringing. Others include parental attitudes and parenting style, sociopolitical events that occur during one's formative years, formal and informal educational experiences, and a host of idiosyncratic personal experiences. Furthermore, some social attitudes have been shown to have a genetic component (e.g., Waller et al., 1990), and so, too, could isms.

As is the case with vocational interests, we suspect that isms are relatively stable throughout the lifespan. However, despite prior belief to the contrary, individual differences such as personality have been shown to change throughout the lifespan (e.g., Roberts et al., 2006). As we posited above, one possible mechanism for ism change over the lifespan would be to threaten one's basic needs. For instance, major sociopolitical, or personal life, events that threaten one's belief that the world is fair and just could lead to shifts in one's level of tradition-oriented religiousness.

Another important area for future research will be to replicate the structure of isms across several cultures. To the best of our knowledge, the only attempt at a cross-cultural replication was Krauss's (2006) replication in Romania. In this study, Krauss roughly replicated the first four isms factors. It will be important, however, to replicate the structure in non-European cultures, such as the cultures of Asia, Africa, and Latin America. It is our belief that nearly any pattern of results would be interesting. Replication across all, or most, cultures might suggest a common human evolution of social attitudes and beliefs that are somehow related to the survival of the species. By contrast, distinct isms patterns across cultures would provide important information about the structure of cultural differences.

A final area for future research will be to predict important school and work outcomes with isms. Meta-analysis has found that attitudes predict a wide range of behaviors (Kraus, 1995), and individual difference measures predict important outcomes both at school (e.g., Poropat, 2009) and at work (e.g., Judge et al., 2013). Isms should be no different. Numerous hypotheses can be put forth. Some examples include:

- Tradition-oriented religiousness is negatively related to creativity
- Unmitigated self-interest is negatively related to teamwork
- Communal relationism is positively related to leadership
- Subjective spirituality is positively related to coping skills (this hypothesis is less clear than the others)
- Inequality-aversion is positively related to organizational citizenship behaviors (positive behaviors such as helping coworkers, being cooperative, and putting forth extra effort to get the job done; Borman and Motowidlo, 1993).

If isms are indeed shown to relate to important outcomes, it is unclear about how to best use them to improve outcomes for individuals. One potentially fruitful use would be in providing feedback to students and employees as to how their unstated assumptions and values influence both their behavior and their interpretation of the events that happen at school and work.

Conclusion

Each of us has a set of social attitudes and beliefs that guide our perception and behavior as we traverse life. To witness their power in everyday events, one merely needs to turn on the television or read the news. Once one is aware of these isms dimensions, it is clear that nearly (or probably) all major social and political events are heavily influenced by them. We suspect that they also have a profound influence in nearly all minor everyday events. As such, a further exploration of these dimensions has the potential to greatly understand our understanding of human behavior in general, and our prediction of important performance outcomes more specifically.

See also: Authoritarianism; Big Five Factor Model, Theory and Structure; Dark Triad: The "Dark Side" of Human Personality; Epistemological Beliefs and Motivation; Essentialism; Eudaemonism; Five Factor Model of Personality, Facets of; Optimism and Pessimism; Personality, Trait Models of; Subjective Wellbeing and Culture; Values, Social Psychology of.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harpers, New York.

Allport, F.H., Lepkin, M., 1945. Wartime rumors of waste and special privilege: why some people believe them. Journal of Abnormal and Social Psychology 40, 3–36.

Ajzen, I., 1991. The theory of planned behavior. Organizational Behavior and Human Decision Processes 50, 179–211.

Ajzen, I., 2006. Constructing a TpB Questionnaire: Conceptual and Methodological Considerations. Retrieved from: http://www.unix.oit.umass.edu/∼aizen/pdf/tpb.measurement.pdf.

Bandura, A., 1977. Social Learning Theory. Prentice-Hall, Englewood Cliffs, NJ.

Barrick, M.R., Mount, M.K., Li, N., 2013. The theory of purposeful work behavior: the role of personality, higher-order goals, and job characteristics. Academy of Management Review 38, 132–153.

Borman, W.C., Motowidlo, S.J., 1993. Expanding the criterion domain to include elements of contextual performance. In: Schmitt, N., Borman, W.C. (Eds.), Personnel Selection in Organizations. Jossey-Bass, San Francisco, pp. 71–98.

Caligiuri, P., Noe, R., Nolan, R., Ryan, A.M., Drasgow, F., 2011. Training, Developing, and Assessing Cross-Cultural Competence in Military Personnel (Technical Report No. 1284). U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA.

Cialdini, R.B., 1993. Influence: The Psychology of Persuasion. Morrow, New York.

Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L., 1999. Subjective well-being: three decades of progress. Psychological Bulletin 125, 276–302.

Dweck, C.S., Chiu, C., Hong, Y., 1995. Implicit theories and their role in judgments and reactions: a world from two perspectives. Psychological Inquiry 6, 267–285.

Eagly, A.H., Chaiken, S., 1993. The Psychology of Attitudes. Harcourt Brace Jovanovich, Orlando, FL.

Fazio, R.H., Olson, M.A., 2007. Attitudes: foundations, functions, and consequences. In: Hogg, M.A., Cooper, J. (Eds.), The Sage Handbook of Social Psychology. Sage, London, pp. 123–145.

Greenberg, J., Pyszczynski, T., Solomon, S., 1986. The causes and consequences of a need for self-esteem: a terror management theory. In: Baumeister, R.F. (Ed.), Public Self and Private Self. Springer-Verlag, New York, pp. 189–212.

Greenwald, A.G., McGhee, D.E., Schwartz, J.K.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.

Herman, S.M., September 1970. The organization as an iceberg. In: Paper Presented at the Organizational Development Network Conference, Chicago, IL.

Hofstede, G., 1983. National cultures in four dimensions: a research-based theory of cultural differences among nations. International Studies of Management & Organization 46–74.

Homer, P.M., Kahle, L.R., 1988. A structural equation test of the value-attitude-behavior hierarchy. Journal of Personality and social Psychology 54, 638–646.

Judge, T.A., Rodell, J.B., Klinger, R.L., Simon, L.S., Crawford, E.R., 2013. Hierarchical representations of the five-factor model of personality in predicting job performance: integrating three organizing frameworks with two theoretical perspectives. Journal of Applied Psychology 98, 875–925.

Kerlinger, F.N., 1972. The structure and content of social attitude referents: a preliminary study. Educational and Psychological Measurement 32, 613–630.

Kraus, S.J., 1995. Attitudes and the prediction of behavior: a meta-analysis of the empirical literature. Personality and Social Psychology Bulletin 21, 58–75.

Krauss, S., 2006. Does ideology transcend culture? A preliminary examination in Romania. Journal of Personality 74, 1219–1256.

Lerner, M.J., Miller, D.T., 1978. Just world research and the attribution process: looking back and ahead. Psychological Bulletin 85, 1030–1051.

Lipnevich, A.A., MacCann, C., Krumm, S., Burrus, J., Roberts, R.D., 2011. Math attitudes and mathematics outcomes in U.S. and Belarusian middle school students. Journal of Educational Psychology 103, 105–118.

Low, K.S.D., Yoon, M., Roberts, B.W., Rounds, J., 2005. The stability of vocational interests from early adolescence to middle adulthood: a quantitative review of longitudinal studies. Psychological Bulletin 131, 713–737.

Morgan, G.S., Wisneski, D.C., Skitka, L.J., 2011. The expulsion from Disneyland: the social psychological impact of 9/11. American Psychologist 66, 447–454.

Nosek, B.A., 2007. Implicit–explicit relations. Current Directions in Psychological Science 16, 65–69.

Nosek, B.A., Banaji, M.R., Greenwald, A.G., 2002. Harvesting implicit group attitudes and beliefs from a demonstration website. Group Dynamics 6, 101–115.

Poropat, A.E., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychological Bulletin 135, 322–338.

Roberts, B.W., Walton, K.E., Viechtbauer, W., 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. Psychological Bulletin 132, 1–25.

Rounds, J., Su, R., 2014. The nature and power of interests. Current Directions in Psychological Science 23, 98–103.

Saucier, G., 2000. Isms and the structure of social attitudes. Journal of Personality and Social Psychology 78, 366–385.

Saucier, G., Skrzypińska, K., 2006. Spiritual but not religious? Evidence for two independent dispositions. Journal of Personality 74, 1257–1292.

Saucier, G., 2013. Isms dimensions: toward a more comprehensive and integrative model of belief-system components. Journal of Personality and Social Psychology 104, 921–939.

Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 25. Academic Press, New York, pp. 1–65.

Schwarz, N., 2004. Metacognitive experiences in consumer judgment and decision making. Journal of Consumer Psychology 14, 332–348.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., Simons, A., 1991. Ease of retrieval as information: another look at the availability heuristic. Journal of Personality and Social Psychology 61, 195–202.

Waller, N.G., Kojetin, B.A., Bouchard, T.J., Lykken, D.T., Tellegen, A., 1990. Genetic and environmental influences on religious interests, attitudes, and values: a study of twins reared apart and together. Psychological Science 1, 138–142.

Wyer, R.S., Albarracin, D., 2005. Belief formation, organization, and change: cognitive and motivational influences. In: Albarracin, D., Johnson, B.T., P Zanna, M. (Eds.), The Handbook of Attitudes. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 273–322.

Social Business

Enrique Ogliastri, IE Business School, Madrid, Spain
Andrea Prado and Urs Jäger, INCAE Business School, Alajvela, Costa Rica
Antonio Vives, Stanford University, Stanford, CA, USA
Ezequiel Reficco, Universidad de los Andes, Bogotá, Colombia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social business deals with economic organizations oriented to social value creation. It is a new concept integrating multiple dimensions and meanings, including managerial experiences of public, private, and nonprofit organizations. Social business includes contributions from corporate social responsibility, nonprofit management, social entrepreneurship, and inclusive business (i.e., bottom of the pyramid). The concept is being consolidated by the evolution of relationships between business and society in the twenty-first century.

Introduction

The core concept of social business is economic activities with a priority aim to social value creation. It is a new integrative concept covering multidimensional meanings. There is an overlap of conceptual developments and managerial practices. The concept of social business is the result of contributions from corporate social responsibility (CSR), nonprofit organizations (NPOs), social entrepreneurship (SE), and base of the pyramid/inclusive business (BoP/IB).

CSR was the original concept and, along with environmental sustainability, became a fashion in management during the first decade of the twenty-first century. CSR originally was conceived as ethics and moral obligations, related to altruistic or philanthropic motives, but it has moved to be considered by many as social investments and plain good business. Such claim is still in debate, however. The evolution of CSR included practices of socially responsible investments, stakeholder management, the new philanthropy, and International Organization of Standardization (ISO) standard 26000.

NPOs and management are studied in and include an array of names and institutions, such as nongovernmental organizations (NGOs), charitable organizations, philanthropic foundations, cooperatives, civil society organizations, and so on. The majority of NPOs now get more than half of their income from self-generated income. The literature about NPO has focused on the professionalization, actual social impact, and financial sustainability of the organization. A key part of research has been about strategic alliances between sectors. NPO started by incorporating the best practices of for-profit management tools, but increasingly it is aimed at adapting and developing its own specific management concepts. Furthermore, the blurring of borders between public, private, and civic sectors is challenging NPOs in three ways. First, the changing challenges of the top management in respect to mission, governance, strategy, accountability, and social impact assessment. Second, the need to explore new models to gain resources, such as foundations as a legal form to secure financial resources, revenue-based nonprofits, or volunteering. Third, the discussions on the blurring of sector borders affect the ongoing question of effective education for nonprofit managers.

SE is a new concept undergoing debates about its very nature. It consists of innovative social value creation by organizations from public, private, civic, and hybrid sectors. SE is a source of social change, particularly when the organization can be replicated in other environments. Research on social entrepreneurs often contrasts them with for-profit entrepreneurs, but no conclusive evidence has been reached. Research has focused on the creation of social ventures, the measurement of social impact, and theoretical constructs development. SE is undergoing consolidation as a unified concept, and it is getting contributions from psychology, management, and economics.

BoP/IB are increasingly interchangeable terms, but the BoP was originated by scholars, whereas IB has been used more by practice-oriented and multilateral international organizations. The main idea is to alleviate poverty by integrating low-income citizens as consumers, employees, producers, suppliers, distributors, retailers, and entrepreneurs in the value chain of large companies. The best case of this BoP/IB concept is the microfinance field, in which results have been proven to benefit everyone involved. This approach is directly related to CSR and focus on the management of stakeholders and social capital.

Yunus (2008) restricted the concept of social business to for-profit business with a social aim, while our definition is wider, covering for-profit organizations as well as NPOs that engage in economic activities for social value creation. The following sections discuss the four issues of social business in more detail.

Corporate Social Responsibility

Where Does CSR Come From?

Traditionally businesses have felt the need to contribute to society, particularly in those times in which segments of society have demonstrated strong needs, like in the case of pervasive poverty, immigration, natural disasters, and the like. Company actions represented an extension of the spirit of individual solidarity, framed in the context of philanthropy (love for mankind). Some concerned businesspersons like Carnegie and Rockefeller in the United States and Cadbury in the United Kingdom in the early twentieth century and Bill Gates and Warren Buffet in the early twenty-first century, saw the

importance of complementing the actions (or failures) of governments in promoting society's development and contributed to financing schools, libraries, health facilities, and research among others. This support for society has been carried out with their own money. On many occasions, corporate resources are used to enhance society's well-being, with or without the expectation that they will enhance company profits. When they are related, the term strategic philanthropy often is used. To enhance impact, these resources can be managed with corporate principles and sometimes are called 'philanthrocapitalism' or 'venture philanthropy' (Porter and Kramer, 2002).

Where Is CSR Now?

In the mid-twentieth century, after World War II, the ideas of CSR developed further, and many companies began to see a responsibility toward society – some still out of solidarity but others out of enlightened self-interest. These companies realized that a developed society could be good for business. Research and practice evolved from the explicitly normative and ethics-oriented arguments to implicitly normative and performance-oriented managerial studies and practices (Lee, 2008). At this point, the concepts of CSR (Carroll and Buchholtz, 2011) developed to reflect the fact that companies used corporate resources for the good of society and the corporation to account for the impact on society and the environment of their activities. The modern concept includes concern for the impact on all the parties affected by company activities, in this context called 'stakeholders.' Obviously, this concern takes many different forms and includes many different parties, from employees and customers all the way to communities and the natural environment. In many cases, these stakeholders make their voices heard spontaneously, particularly in the case of corporations with large impacts, but in all cases, it is important for the company to engage in stakeholder management through consultation and to assess impact and obtain feedback.

To guide the implementation of these activities in all types of institutions the International Standards Organization produced the ISO 26000 Guidance for Social Responsibility Standard, which unlike other ISO Norms, is noncertifiable (ISO, 2011).

Within these basic ideas, there are very large variations in company behavior, depending on the context in which they operate; the strength of existing regulations and enforcements; the level of development of civil society, media, and financial markets; and particularly the relative power and willingness of consumers or client to exercise their rights. There can be companies for which the responsibility stops at complying with the law; for others, it will include actions that go beyond the law, either because the law or its enforcement are weak (as in developing countries) or because they see the value of being responsible in the form of risk mitigation of future problems or to gain the favor of the market for responsibility – that is, all stakeholders.

CSR in the second decade of the twenty-first century is a management strategy that not only includes the management of impacts on society but also tries to capture the potential benefits from the responsibility vision. One way to look at this idea is through the concern for a triple bottom line – that is, concern for the financial, social, and environmental results, and not just the traditional financial profitability, developed in the late-twentieth century. Its actual implementation will depend on the corporation's degree of responsible development, the context of time and place, the type of business, reactions of stakeholders, government and civil development, and society's needs and expectations. There are large variations in responsible practices between corporations operating in well-developed markets and in less developed ones, from large corporations to small companies, from broadly held corporations with anonymous, and many times short-term, shareholders to closely controlled and family-owned corporations. The motivations of the corporations and the impact that their responsible practices have on their own performance through the reactions of stakeholders determine their responsibility strategies.

These variations include a core of responsible practices that all companies should have, starting with the compulsory compliance with the law, responsible products and procurement, employee welfare, care for environmental resources among others, but beyond these practices, there are many variants. Some take a broader view of the contribution to society and to economic development through IB, SE, and business at the bottom of the pyramid, which will be discussed in the following section. In many cases, through the extensive use of natural resources, particularly air, water, and energy, corporations have a significant impact on the environment and need to ensure environmental sustainability through mitigating, offsetting, and undoing their impacts (Vezzoli and Manzini, 2010).

To stimulate responsible practices in corporations through their access to capital and to try to capture the expected larger benefits of responsible corporations, a field of socially responsible investment has been developed that seeks to invest directly, or indirectly through investment funds, in these corporations. The selection criteria vary widely, from simple negative screening (eliminating companies with certain products or practices) to positive screening (looking for specific or general responsible practices) (Fung et al., 2010).

Still, in many corporations, to a larger or lesser extent, 'greenwashing' will be found. This is the traditional exploitation of imperfections in the markets and asymmetries in information by some companies to capture the benefits of corporate responsibility without the corresponding costs.

Where Is CSR Going?

In the end, all of these ideas converge into sound management practices. In the future, there should be no need to talk about CSR, or any other responsibility concepts, as all will boil down to good management. CSR simply will be good management. Nevertheless, research will continue to be in segments, given the complexity of the issues. In the near future, competition will force corporations to be responsible if they want to survive. Advances in information technology and dissemination will provide better knowledge to all actors, and the possibilities for stakeholders to influence will be increased, forcing corporations to be responsible. Corporations will endeavor to be financially, socially, and environmentally sustainable.

Managing Nonprofit Organizations

Where Are NPOs Now?

This section seeks to take stock of the selected research on managing NPOs. It starts from the assumption that research on nonprofit management sometimes explicitly, but in most cases implicitly, intends to professionalize the amateur administration assumed for many NPOs.

Although the focus on professionalization has been given since the beginning of nonprofit management research in the 1970s, the research agenda has changed: Originally, scholars mainly focused on transferring proved management instruments from profit enterprises to NPOs, in areas such as accounting, human resource management, or strategic planning. The increasing financial scarcity since the turn of the millennium, however, has strengthened scholars' interests in effective fund-raising and marketing of NPOs (Helmig et al., 2004). Since then, the borders among the social, markets, and public sector have blurred (Evers, 2005).

Where Are NPOs Going?

In light of the difficulties to hold up the former distinction among the three sectors, scholars adapt their research agenda in three ways.

Research on Top Management

The blurring of sector borders challenges executive directors and board directors twofold (Young, 2008). First, scholars argue that NPOs play an increasingly critical role in the development of national economies and societies worldwide. They see a growing importance of NPOs in influencing international affairs in such fields as international trade. The question in this research field is how NPOs can support governmental policy making in solving international social and ecological issues. This rising interest of scholars is documented by the new journal *Nonprofit Policy Forum*, among others, edited by Dennis Young and Linda Serra. Second, corporate philanthropy moves from a focus on social projects to one that emphasizes the bottom line of the corporation. In respect to this new corporate landscape, scholars explore new challenges for NPOs in regard to such issues as competition with profit enterprises and opportunities for cooperation between NPOs and profit enterprises.

These challenges affect the top management in five different fields. First, there is an increasing interest in the 'mission drift' and how or if NPOs find a balance between social mission and financial resources. Second, scholars address the different roles between the board of directors and the executive director. They focus less on normative approaches that describe the ideal governance structure than on questions of how the boards deal with conflicts and goal contradictions. Third, the current research on nonprofit strategy explores the difficulties to implement formal strategy processes within NPOs. There is an increasing interest in understanding how NPOs strategize. Fourth, more scholars explore how NPOs manage being accountable to main stakeholders. A key issue therein is to measure social impact. Fifth, and finally, the rising interest on impact measurement is why social impact assessment is an upcoming research field.

Research on Gaining Resources

Although NPOs traditionally have been able to operate effectively by focusing exclusively on their mission, many now are facing the imperative of adopting businesslike reasoning. This strengthened scholars' interest in nonprofit marketing and fund-raising. The number of contributions in these fields is immense. The discussion on enterprising NPOs leapfrogs the traditional social mission orientation of NPOs. This stream of literature focuses on NPOs that, in part, sell services or products to markets to gain money that they invest in their social services. Scholars also search for new legal forms that answer increasing financial scarcities in which they expect foundations to play a particularly critical role. Besides financial resources, the literature on volunteering focuses on human resources. Scholars explore how NPOs can gain, keep, and lead volunteers.

Because of the stronger integration of social and economic issues within most management functions, there is an upcoming discussion on hybrid NPOs. This stream of literature asks whether NPOs have to manage a single bottom line or multiple bottom lines (Anheier, 2005).

Research on Management Education

Studies confirm that executive leadership transition will be an important focus of study in the nonprofit sector over the next several years. Because of the challenges discussed thus far, the requirements of a nonprofit executive director's job gradually expand from a social focus to a more complex task at the interface of markets and civil societies (Austin et al., 2006a). Executives need to act differently to resolve or reconcile the tensions between businesslike reasoning and the social mission. These increasing job requirements lead to changes in human resources and governance structures in which executive directors play a more important role, similar to the ones in for-profit enterprises. This trend leads to a growing discussion on 'leadership deficit' and the increasing need for more skilled executives and adequate training.

As a conclusion to practitioners, these trends might differ from country to country: In some regions, NPO managers are a 'secret elite' with high social impact. In general, the tendency is to professionalize NPO management. In particular, for managers of highly professionalized NPOs, the transfer of management instruments and concepts proven to be successful in profit enterprises to NPOs and the integration of businesslike reasoning is no longer sufficient. New management concepts are needed that are suitable for executive director's specific management challenges that differ from the challenges of their colleagues in profit enterprises.

Social Entrepreneurship

Where Is SE Now?

Definitions of SE have been developed within different academic fields to refer to for profit, nonprofit, and public sector as well as hybrid organizations combining some or all the above (Peredo and McLean, 2006; Short et al., 2009). Austin et al. (2006b) defined SE "as innovative, social value creating activity that occurs within or across the nonprofit, business, or government sectors" (p. 2). Short et al. (2009) encouraged

scholars to embrace a broad definition of the concept, such as the one proposed by Mair and Marti (2006): a "process involving the innovative use and combination of resources to pursue opportunities to catalyze social change and/or address social needs" (p. 37). A unified definition has not yet emerged, making it hard for the field to establish its legitimacy. What all definitions share, however, is the social enterprises' underlying drive to create social value, rather than shareholder wealth (Austin et al., 2006a,b). For social entrepreneurs, the social mission is explicit and central; wealth is just a means to an end. Social entrepreneurs are change agents in the social sector (Dees, 2007).

One way scholars have tried to set the boundaries of the SE construct is by contrasting it with the traditional or commercial entrepreneurship. The opportunities generated by market failures, the mission of the enterprise, the mobilization of human and financial resources, and the performance measurements are fundamental distinguishing features among these two types of entrepreneurs (Austin et al., 2006b; Dees, 2007). For instance, the embedded social purpose or the nondistributive restriction on surpluses of the social enterprises limits them from tapping into the same capital markets as their commercial counterparts (Austin et al., 2006a,b). Thus, while some of them charge fees for their services, they often have to compete for donations and other types of support in the philanthropic markets (Dees, 2007). The management of human resources in the social enterprises also requires a different approach. Many rely on volunteer labor, and the economics of the enterprise often make it difficult to compensate staff as competitively as commercial initiatives (Austin et al., 2006a).

The social purpose of these enterprises makes measuring their performance a challenging task. Although commercial entrepreneurship can rely on relatively quantifiable measures, such as financial indicators or market share, to measure their performance, social change poses a greater challenge due to nonquantifiability, multicausality, temporal dimensions, and perceptive differences of the social impact generated. Measurement complications influence these enterprises' accountability and shareholder–donor relations. Other differences identified between commercial entrepreneurship and SE include the longer time it takes to achieve goals, the greater number of stakeholders involved, the absence of financial loss for stakeholders, and the nondistribution constraint of the latter (Austin et al., 2006b).

Innovation has also been acknowledged by scholars as an important feature of the social enterprises (Dees, 2007; Peredo and McLean, 2006). A significant range of activities can fall under SE, such as helping increase capacities of actors to solve their own problems, disseminating innovations to serve a widely distributed social need, and building movements to challenge the structural causes of social problems. Social entrepreneurs do not need to be inventors. They can innovate in the way they structure their core programs, in how they assemble the resources, or in the way they ensure funding for their work (Dees, 2007).

SE can occur through multiple organizational and legal forms; the particular form a social enterprise takes should be based on which format would most effectively mobilize the resources needed to address the social issue (Austin et al., 2006a). It is in part due to the heterogeneity of organizational forms that these enterprises can take that academic research on SE has been informed by for-profit entrepreneurship, social issues in management, and public and nonprofit management literature (Short et al., 2009).

At the individual level, empirical evidence of social entrepreneurs being different from its commercial counterpart is scarce. In this sense, social entrepreneurs as traditional entrepreneurs recognize and exploit opportunities to create value, employ innovation, tolerate risk, and refuse to accept limitations in available resources (Dees, 2007; Peredo and McLean, 2006). One important difference between these two types is the importance of the social entrepreneur's networking skills. These skills become fundamental for the success of the enterprise given the wide variety of stakeholders they have to cope with and the limitation of resources they often face in the philanthropic markets (Austin et al., 2006b).

Where Is SE Going?

Despite the growing attention that SE has received by scholars and practitioners in the past two decades, this academic field is considered to be in its infancy (Short et al., 2009). Research in this field has been dominated by anecdotal evidence, case studies, and conceptual papers (Mair and Marti, 2006; Short et al., 2009). For the SE field to progress, the next two decades should be characterized by unity in construct definition and by examining this construct through a variety of established theoretical lenses with clear boundary conditions (Short et al., 2009). In this line, new insights might be gained by applying a quantitative research approach more frequently. Potential avenues for future research can be explored by incorporating other disciplines in the study of SE, such as operations management, economics, marketing, finance, psychology, and accounting, which can inform the field and provide new frameworks of analysis (Short et al., 2009).

Base of the Pyramid and Inclusive Business

Where Does BoP/IB Come From?

Although it has become increasingly common to find the term 'base of the pyramid approach' (BoP, hereafter) used interchangeably with 'inclusive business' (IB), both can be traced to have different origins and to retain some specificities. The former was coined by a group of US scholars, mainly from the Universities of Michigan and Cornell, who pushed the idea that private business had a big role to play in global poverty alleviation. In essence, their message was not new: The CSR movement had been pushing that idea for decades. These intellectuals, however, innovated substantially on two accounts. First, the tone of the message changed: Although previously, the link between poverty and markets had been framed in normative terms, these authors were the first to articulate a convincing business case behind it. Second, and most important, those delivering the message were not activists, but highly credible advisors to corporate America. The result was that the concept stopped being discussed only at the company foundation or the community relations department: For the first time, it made it to the corporate boardroom.

The concept of IB, on the other hand, started to be used intensively by the mid-2000s by various practice-oriented organizations, including those that grouped large corporations – like the World Business Council for Sustainable Development, or the International Business Leaders Forum – and international development organizations – such as the *Stichting Nederlandse Vrijwilligers* (the Netherlands Development Organization, or simply SNV) or the British Department for International Development. Following their lead, various multilateral agencies – such as the United Nations Development Program, the World Bank, and the Inter-American Development Bank – quickly adopted the term and made it part of their agenda.

Where Is BoP/IB Now?

The concept of IB was defined as "one which seeks to contribute toward poverty alleviation by including lower-income communities within its value chain while not losing sight of the ultimate goal of business, which is to generate profits" (SNV & World Business Council for Sustainable Development, 2008: p. 2). This implies integrating low-income citizens as consumers, employees, producers, suppliers, distributors, retailers, and entrepreneurs. To a large extent, the IB concept stresses the role of 'anchor' companies, which open up portions of their value chain to disenfranchised groups to improve their living standards.

In its original formulation, the BoP concept was narrower, as it was presented as a neglected opportunity: a *fortune* to be captured if only companies cared enough to cater to the needs of the poor (Prahalad, 2005). Low-income citizens were seen as the new frontier, *the next 4 billion* of untapped consumers (Hammond et al., 2007). These early writings emphasized the potential of multinational corporations (MNCs) as the organizational form best placed to capitalize on the opportunity and improve the lives of the poor (Lodge, 2002), as they commanded the financial clout, global distribution channels, and brainpower needed to address the challenge. In actuality, even in those early works, the emphasis on large corporations is not absolute. Many of the cases used by Prahalad in Fortune at the BoP are not MNC: Casas Bahía was a large Brazilian Company; Voxiva a small- to medium-size enterprise (SME), and Aravind an NPO. That fact, however, was not acknowledged in the conceptual frameworks developed inductively from the experiences.

In any case, in the ensuing years, these BoP traits were softened. First, BoP authors started to recognize the critical importance of NGOs as 'unconventional partners' in hybrid value chains. Moreover, recent formulations – the so-called BoP 2.0 – expanded the definition of the concept, so as to blur most differences with IB: The emphasis on MNC is gone and the poor are seen in various roles not only as untapped consumers. "BoP ventures span sector and size, including initiatives by large multinational and domestic companies, local small and medium-scale enterprises, and business developed by nonprofit organizations and social entrepreneurs" (London and Hart, 2011: p. 10). The BoP domain is defined as those business models "specifically designed to target the poorest segments of society as consumers, producers, and entrepreneurs" (London and Hart, 2011: p. 2). Two differential shades of meaning remain, however. Although BoP authors have always stressed the need to engage the needy through ad hoc methodologies, the IB writings and practice have assumed more continuity with the status quo. Whereas the BoP success stories are all about new products, novel technologies, and hybrid business models, IB cases are about including vulnerable groups on existing value chains.

A somehow-related shade of meaning between both concepts is their relation with the wider CSR agenda. In Prahalad's (2005) seminal book, CSR is barely mentioned. In those few instances in which it is indeed brought about, the author stressed that what he proposed was something *different* from that: "This was not about philanthropy and notions of corporate social responsibility" (Prahalad, 2005: p. XIV). That emphasis was meant to highlight that there was real opportunity in this market, not just moral duty. The IB agenda, on the other hand, always remained very close to the CSR movement. This should come as no surprise, as the IB's widely cast net (encompassing consumers, employees, producers, suppliers, and distributors) overlaps nicely with the various stakeholders about whom large companies must care. In the original BoP writings, the commercial opportunity is stressed more prominently than the ethical imperative of "giving back to community," the reverse is true in the IB lexicon – although this difference has faded with time. In the last BoP writings, the social and environmental dimensions of the value created appear more prominently than the commercial opportunity. This shift may be related to the fact that for most MNC, profits at the BoP have proved somehow elusive. The evidence regarding the size and robustness of the opportunity is mixed at best. Some authors have expressed skepticism about the real commercial value of the BoP segment, pointing out that the same success stories keep up appearing over and over; some go as far as calling the BoP a 'mirage' (Karnani, 2007: p. 90). To date, few studies have tried to measure up the actual creation of economic value in BoP ventures; the microfinance industry, which has succeeded in delivering substantial financial value while at the same time transforming the lives of the poor, stands as a noteworthy exception.

Where Is BoP/IB Headed?

From the preceding paragraphs, it should be clear that one obvious trend is the convergence between the agendas and terminologies of the BoP and IB approaches. Both retain distinctive characteristics, but these seem to be losing practical relevance. Another related trend is the convergence with the so-called green agenda: Environmental concerns are increasingly prominent on BoP writings, on a par with poverty alleviation – what Hart has called 'the Great Convergence' (Hart, 2010: p. 162).

A not-so-obvious trend is the convergence between the agendas of the BoP and social enterprise. As the social enterprise literature has disengaged from specific organizational forms, accepting that SE can take place not just in nonprofits, but also in SME and even in corporations, few real distinctions separate what BoP and SE authors are studying. Finally, the social innovation approach may in the long run engulf portions of the research that currently are carried out in many of the fields of SE, CSR, IB, or BoP.

Conclusion: Trends and Implications

Whitehead suggested the last thing to be discovered by any science is what the science is really about, and this seems to be true in this case. The concept of social business is in its infancy, but the field has been growing rapidly in multiple interrelated dimensions. Practical implications and managerial initiatives have been strong in parallel with conceptual development. A managerial core on CSR has matured into distinct research areas. Social business, as economic activities for social value creation, will become a core concept integrating the multiple avenues of research and practice. At least four directions focus future development of the field: the consolidation of good management, international regulations, blurring borders among sectors, and the transformation of capitalism.

Social business management is being consolidated by trends to certifications, such as ISO, emerging codes of ethics and malpractice for the managerial profession, interactive learning between sectors, development of techniques and best practices to reduce uncertainties of business and society, and theoretical development of management techniques. Marketing, human resources, informatics, finance, operations, and strategy, among other subfields of management, will become more sophisticated, integrating concerns of the stakeholders. For example, public relations will no longer be just pushing press releases and propaganda, but rather listening to the public and to all stakeholders of a company. Organizations of the private, public, and third sectors will have higher standards and accountability when dealing with employees, consumers, suppliers, shareholders (economic sustainability), governments, communities, and the environment.

The trend of international regulations will be particularly strong on environmental issues (i.e., organic agriculture), human rights, labor fairness, the conduct of war, and governance regulations. Multilateral international organizations will have more control and will make international regulations and agreements more enforceable. Probably ISO 26000 will be certifiable, as the other norms of ISO now are. That way the concerns of social business will be widespread.

Foundations, NGOs, cooperatives, and NPOs normally will be required to be efficient, generate their own income, and evaluate their real social impact. Public sector organizations will be accountable on the real impact of policies. Private companies as well as public organizations and social enterprises will be required to be economically sustainable, achieve social value, and contribute to environmental sustainability, the triple bottom line. Some may be called 'fourth sector' or 'hybrid organizations'; strategic alliances among sectors may increase and cross-learnings will become highly valuable. The trend to the blurring of sectors will mean a convergence of management and a development of management as a professional activity.

The emergence of the social sciences in the nineteenth century witnessed a preoccupation with the decrease of solidarity in capitalist society, or at least a change in the nature of the institutional ties that hold society together. Probably the evolution of capitalism implies changing its very nature. A most successful leader of social business predicts a blooming of social capitalism, a second type of private economic activity oriented to social ends (Yunus, 2008). It may be quite a ways away, but the consolidation of important initiatives of new philanthropy, volunteer work and nonprofits, social investments, IB, and SE as we have seen in this century may result in changes of 'capitalism' as understood by Adam Smith.

Bibliography

Anheier, H., 2005. Nonprofit Organizations. Theory, Management, Policy. Routledge, London, New York.

Austin, J., Gutiérrez, R., Ogliastri, E., Reficco, E., et al., 2006a. Effective Management of Social Enterprises: Lessons from Business and Civil Society Organizations in Iberoamerica. Harvard University Press, Cambridge, MA.

Austin, J., Stevenson, H., Wei-Skillern, J., 2006b. Social and commercial entrepreneurship: same, different, or both? Entrepreneurship: Theory & Practice 30 (1), 1–22.

Carroll, A.B., Buchholtz, A.K., 2011. Business and Society: Ethics, Sustainability and Stakeholder Management, eighth ed. South-Western College Pub, Cincinnati, CT.

Dees, J.G., 2007. Taking social entrepreneurship seriously. Society 44 (3), 24–31.

Evers, A., 2005. Mixed welfare systems and hybrid organizations: changes in the governance and provision of social services. International Journal of Public Administration 28 (9 & 10), 736–748.

Fung, H.-G., Law, S.A., Yau, J., 2010. Socially Responsible Investment in a Global Environment. Edward Elgar Publishers, London.

Hammond, A.L., Kramer, W.J., Katz, R.S., Tran, J.T., Walker, C., 2007. The Next 4 Billion: Market Size and Business Strategy at the Base of the Pyramid. World Resources Institute and International Finance Corporation, Washington, DC.

Hart, S.L., 2010. Capitalism at the Crossroads: Next Generation Business Strategies for a Post-crisis World. Prentice Hall Pearson, NJ.

Helmig, B., Jegers, M., Lapsley, I., 2004. Challenges in managing nonprofit organizations: a research overview. Voluntas 15 (2), 101–116.

ISO, 2011. Discovering ISO 26000.

Karnani, A., 2007. The mirage of marketing to the bottom of the pyramid. California Management Review 49 (4), 90–111.

Lee, M.-D.P., 2008. A review of the theories of corporate social responsibility: its evolutionary path and the road ahead. International Journal of Management Reviews 10 (1), 53–73.

Lodge, G.C., 2002. The corporate key: using big business to fight global poverty. Foreign Affairs 81 (4), 13–18.

London, T., Hart, S.L., 2011. Creating a fortune with the base of the pyramid. In: London, T., Hart, S.L. (Eds.), Next Generation Business Strategies for the Base of the Pyramid: New Approaches for Building Mutual Value. FT Press, Upper Saddle River NJ, pp. xxxvii, 249.

Mair, J., Marti, I., 2006. Social entrepreneurship research: a source of explanation, prediction and delight. Journal of World Business 41 (1), 36–44.

Peredo, A.M., McLean, M., 2006. Social entrepreneurship: a critical review of the concept. Journal of World Business 41 (1), 56–65.

Porter, M., Kramer, M., 2002. The competitive advantage of corporate philanthropy. Harvard Business Review 80 (12), 56–68, 133.

Prahalad, C.K., 2005. The Fortune at the Bottom of the Pyramid: Eradicating Poverty through Profits. Wharton School Publishing, Upper Saddle River, NJ.

Short, J.C., Moss, T.W., Lumpkin, G.T., 2009. Research on social entrepreneurship: past contributions and future opportunities. Strategic Entrepreneurship Journal 3, 161–194.

SNV & World Business Council for Sustainable Development, 2008. Inclusive Business: Profitable business for successful development.

Vezzoli, C.A., Manzini, E., 2010. Design for Environmental Sustainability, second ed. Springer, Berlin.

Young, D., 2008. Alternative perspectives on social enterprise. In: Cordes, J.J., Steuerle, E. (Eds.), Nonprofits Business. The Urban Institute Press, Washington DC.

Yunus, M., 2008. Creating a World Without Poverty: Social Business and the Future of Capitalism. Public Affairs, New York.

Social Capital

Paul F Whiteley, University of Essex, Colchester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.W. Jackman, volume 21, pp. 14216–14219, © 2001, Elsevier Ltd.

Abstract

The French sociologist Pierre Bordieu defines social capital as "the sum of resources, actual or virtual, that accrue to an individual or group by virtue of possessing a durable network of more or less institutional relationships of mutual acquaintance and recognition." There are different definitions of social capital, but his definition captures the key idea that it is a resource which individuals can utilize to further their goals, but which resides at the level of the community. It is created and fostered by relationships between individuals of a voluntary and nonfinancial nature. This article reviews varying definitions of social capital before going on to examine its origins, which can be traced back to both microlevel or individual, and macrolevel or community-based social processes. Finally, the article examines the effects of social capital on society and policy-making, reviewing the evidence that it has benign effects in relation to economic prosperity, educational attainment, health and welfare, crime reduction, and subjective well-being in communities where it is abundant.

Introduction

The concept of social capital has its antecedents in a long tradition of studying the role of communities in human society, which goes back to Greek philosophers such as Plato and Aristotle. Many classic theorists such as Adam Smith, Alexis de Tocqueville, and Emile Durkheim have examined ideas relevant to understanding what is now known as social capital (Bowles and Gintis, 2002). The concept has gained prominence in recent years from the work of three notable theorists: Pierre Bordieu (1986), James Coleman (1988), and Robert Putnam (1993). At its core, social capital emphasizes the importance of community ties between individuals in facilitating and enabling collective action to take place. For Coleman, social capital permits society to do things that it otherwise could not do, and though it is utilized by individuals it originates in relationships between people in society. Coleman wrote: "Like other forms of capital, social capital is productive, making possible the achievement of certain ends that would not be attainable in its absence" (Coleman, 1990: 302).

He illustrated the effectiveness of social capital with the example of the New York wholesale diamond market. In this market it is common for participants to hand over bags of diamonds to each other to permit closer examination, without any insurance or formal contracts. This market is based on trust and it works very effectively. This is largely because the community of traders is dominated by orthodox Jews, who share the same religious beliefs, live in the same highly integrated community, and trust each other. Someone who takes advantage of this trust and cheats risks being ostracized not only by other traders but by the community they live in.

The French sociologist Pierre Bordieu provided as succinct a definition of social capital as anyone. He writes:

> Social capital is the sum of resources, actual or virtual, that accrue to an individual or group by virtue of possessing a durable network of more or less institutional relationships of mutual acquaintance and recognition.
>
> Bordieu and Wacquant, 1992: 119

There is a lively debate about the meaning of social capital with some writers emphasizing its importance to the individual as a resource which they can use to promote their private interests (Lin, 2001), while others emphasize the importance of trust, shared norms, and values in the wider society (Fukuyama, 1995), and yet others like Bordieu highlight the importance of connections between individuals in social networks. The study of social capital was given a big impetus by the publication of *Making Democracy Work* (1993) by Robert Putnam, who examined the effectiveness of democracy in the Italian regions. He explained the large social, political, and economic differences between these regions in terms of their different endowments of social capital.

Putnam found that in regions like Emilia-Romagna in northern Italy, people were sociable, joined voluntary groups, and took part in a wide variety of activities, both formal and informal, which put them in close contact with their fellow citizens. He argued that this built a reservoir of social capital and a willingness to trust others, particularly strangers. In contrast, in the southern region of Calabria civic life was not conducive to building social capital since there were fewer voluntary associations, less social activity, and lower levels of trust. In Emilia-Romagna, citizens developed strong 'norms of reciprocity' and this had a benign effect on many aspects of life in the region including promoting economic prosperity, improving health, assisting educational improvements, encouraging political participation, and enhancing subjective well-being. In this part of Italy people came to expect fair dealing, both from their fellow citizens and also from the regional government. By contrast, in Calabria civic life was not conducive to building social capital, since individuals were not inclined to trust their fellow citizens or the government and as a consequence associational life was neglected. Norms of reciprocity and trust were relatively weak and government did not work very well.

Putnam's argument has its roots in Alexis de Tocqueville's *Democracy in America* (1990(original 1840)) which argued that interpersonal trust and associational life are linked in a virtuous circle, so that as people mix with others in an

informal, voluntary setting, they come to know and trust each other, and this in turn promotes more voluntary activity and civic engagement. As this happens the feelings of trust begin to extend to people beyond the individual's immediate family and circle of acquaintances. These interactions produce what the economists call 'externalities,' which arise when transactions have external effects on others. In this case cooperative behavior spills over into the wider society. Because they know and trust their fellow citizens they are more likely to trust and to help strangers. As social capital develops, society works more efficiently by reducing the transaction and policing costs associated with widespread levels of mistrust between individuals. Societies poor in social capital display only weak associational ties and exhibit limited mutual trust, and as a consequence, they fare worse and work less efficiently than their more fortunate neighbors.

Putnam (2000) drew attention to the 'dark side' of social capital, making the point that the consequences of social capital are not all benign. He makes the distinction between 'bridging' and 'bonding' social capital, the former referring to networks which link diverse communities together, and the latter which keeps them apart. He cites the civil rights movement in the United States and ecumenical religious organizations as examples of the former, while ethnic fraternal groups and exclusive country clubs are examples of the latter. A contemporary example of this would be Northern Ireland which has a lot of social capital in the sense of strong local communities, but they are divided by religious and political differences with Protestant communities having very limited contact with Catholic communities.

Another distinction which has been made in the literature relates to 'linking' social capital which describes connections between individuals in a hierarchical setting with unequal levels of power. Bridging or bonding social capital refers to relationships between individuals at a horizontal level, or between people who are more or less on an equal footing. Linking social capital is very relevant for understanding how people see their government or their superiors in any hierarchical organization.

The social capital idea has been subject to considerable critical scrutiny, both theoretical and empirical (Tarrow, 1996; Claibourn and Martin, 2000; Milner, 2002; Halpern, 2005). One issue concerns the relationship between social capital and financial capital. In the case of financial capital it is depleted by use, so that if an individual spends their savings they will not then be available for investment. However, the same is not true for social capital which is depleted by a lack of use. In this case if individuals cease to take advantage of social capital by working together in communities, then it will depreciate over time. The term 'use it or lose it' very much applies to social capital in a way which is not true of financial resources.

Measuring Social Capital

Just as there are disagreements about the definition of social capital there are disagreements about its measurement (Portes, 1998; Adler and Kwon, 2002). For many researchers the essence of social capital is captured by measures of interpersonal trust (Fukuyama, 1995; Brehm and Rahn, 1997; Pattie et al., 2004). For others it is best measured by the number of voluntary organizations which exist in society and which people are prepared to join (Paldam and Svendsen, 2000). Yet others think that the primary focus should be on measuring the social networks in society, particularly those that facilitate economic prosperity arising from 'weak ties' (Granovetter, 1973; Lin and Erickson, 2008). A number of researchers, including Putnam himself, combine all three measures and define social capital in terms of the sum of formal and informal networks together with levels of interpersonal trust in society (Putnam, 1993; Cote and Healy, 2001).

However social capital is measured, there are wide variations in its incidence both within societies and across countries. This can be illustrated with the assistance of a question asked in the World Values surveys (see http://www.worldvaluessurvey.org). These surveys are aimed at measuring the values, attitudes, and social norms of individuals in a large number of countries. The aim is to understand variations in culture, morality, voluntary activities, religiosity, and a variety of other measures across the world (see Inglehart, 1990). The question is *Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?* The 2010–2012 survey showed that only about 24% of respondents in 51 countries agreed that people could be trusted, while 73% felt that they had to be careful in dealing with other people. **Figure 1** shows the percentages of people who trust, analyzed countrywise.

Figure 1 shows that willingness to trust others varied from 3% in the Philippines to 67% in the Netherlands. Even taking into account the difficulties of comparing very diverse sets of people using different languages and having different cultures this is an amazingly large variation. In some countries like Colombia, pretty much no one trusts anyone else, whereas in other countries like Sweden levels of trust are very high. Insofar as this question captures social capital, then it is extremely diverse.

A glance at **Figure 1** suggests that there are several important correlates of social capital. It appears to be more abundant in Western-style developed democracies with relatively high standards of living, and much less so in poor, developing countries with nondemocratic governments. However, trust is clearly not just a function of income or wealth, since a relatively poor country like Kazakhstan is well above a number of more affluent countries like Spain and Taiwan on the scale. Similarly, despite the heavy focus of research on social capital on the United States, that country is not in the highest trust group which appears to be dominated by other mature democracies such as Australia, Sweden, and the Netherlands. **Figure 1** raises intriguing questions about the sources and origins of social capital, a topic we consider next.

The Origins of Social Capital

There is great diversity in accounts of the origins of social capital with microlevel and macrolevel explanations competing with each other. At the micro, or the level of individuals, some neuroscientists have argued that humans are hardwired to trust each other and to exhibit altruism in their relationships (Grimes, 2003). At the other end of the scale, it

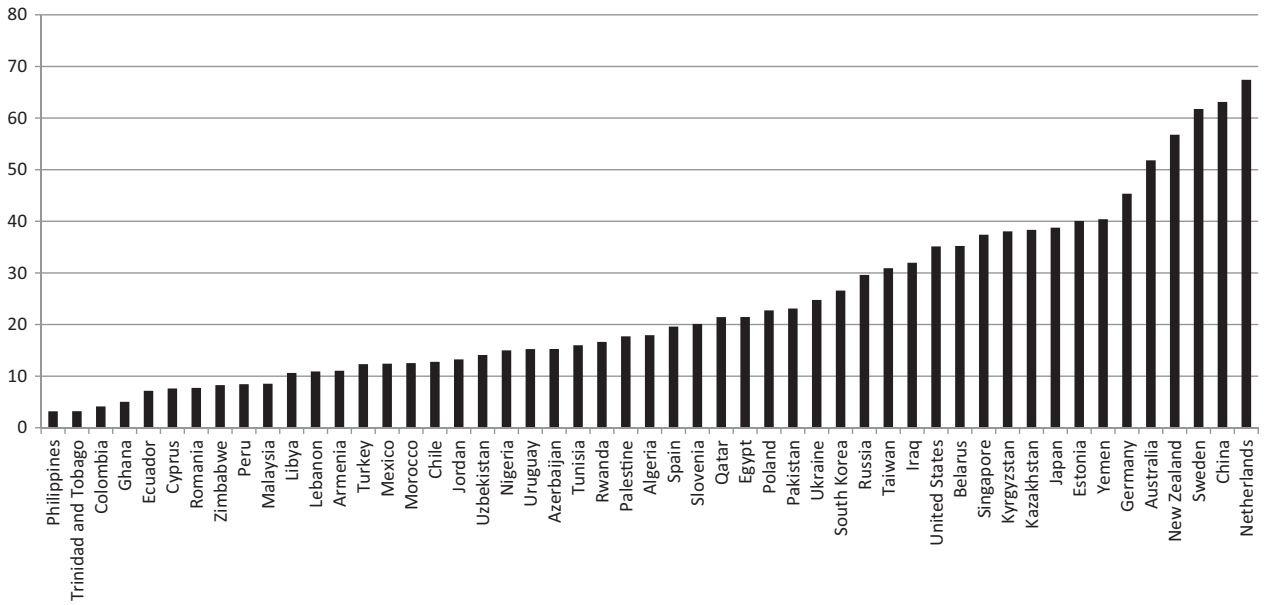

Figure 1 Percentages who trust others in fifty-one countries, 2010–2012. Source: World Values Dataset, 2010–2012.

has been argued the origins of social capital lie in long-standing and sweeping macrolevel historical forces, a position taken by Putnam (1993) in his early work.

If we consider microlevel explanations first, then there is a lot of research in behavioral economics which shows that individuals are not the self-centered utility maximizing entities of traditional economic theory (Camerer, 2003). This can be illustrated by game theory, which is the study of rational decision-making in situations of conflict and cooperation (see Rapoport, 1970). The ultimatum game involves two participants, one of whom makes an offer to the other to divide up pot of money between them (Guth et al., 1982). If the latter accepts the division, then they share out the money and the game ends; if not, then neither of them gets anything. Traditional economic logic would suggest that the proposer should take as much of the money as possible, subject only to the responder being given something. Equally, the latter should accept any division of the money, which gives them a positive amount, on the grounds that something is better than nothing.

However, this is not what happens in a laboratory setting. In numerous experiments proposers offer an average of about 40% of the money with many offering half. Similarly, responders reject offers involving less than 20% of the money about half of the time (Camerer, 2003: 43). The assumption that players maximize their own returns with minimal concessions to other people is clearly falsified by these experiments. This is because issues of fairness and justice, which are hardwired into the great majority of individuals, play an important role in decision-making. Individuals are altruistic and consider the welfare of others.

Trust games get even closer to the idea of social capital than ultimatum games. These are a variant of ultimatum games in which an 'investor' decides how much of a given pot of money should be invested or kept. The invested portion of the money produces a return and this is then divided up by a 'trustee' between the two of them (Berg et al., 1995). The investor has to make a risky decision at the beginning knowing that by investing the money she or he may not get any return if the trustee keeps the proceeds to themselves. Patterns of altruistic behavior are clearly apparent in these games, with completely selfish investors or trustees being quite rare phenomena. Trust occurs and is rewarded, although there are variations in the levels of generosity of participants (Camerer, 2003: 85–86).

These laboratory experiments show that narrow, selfish behavior is not the norm as far as society is concerned. Trust and trustworthiness is hardwired into the human brain and it affects behavior generally, making it an important feature of human society. Having made this point, however, it is readily apparent in **Figure 1** that there are large variations in trust across different types of society, although it is important to remember that trust in response to the simple question in the World Values surveys can be bonding rather than bridging. The experiments suggest that almost all humans trust but trusting strangers, a key feature of social capital, is less common, particularly in some societies. This suggests that the context in which individuals live is very important for understanding the origins of social capital.

At a more macrolevel of analysis, considerable work has been done to investigate how cooperation between individuals evolves over time. This work utilizes the prisoner's dilemma game to model cooperation (Rapoport, 1970). This is one of the best known games in the literature and gets its name from the story used to illustrate the decision problem of the actors involved. The story is that two suspects are arrested on suspicion of committing a crime, but the prosecutor can not convict them without a confession from at least one of them. They are separated and interrogated with the aim of getting one of them to confess and give evidence against the other. If neither confesses they both get a short jail sentence based on a lesser charge which can be proved without a confession. However, if one confesses and agrees to give evidence against the other, he gets a reduced sentence while his companion gets a severe

sentence. If both confess and offer to testify against each other then both get a heavier sentence than they would have if neither had confessed. This is the prisoner's dilemma.

If the game is played only once then the optimal strategy for both prisoners is to confess, on the grounds that they can not trust each other. Whatever their companion does they are better off confessing than taking the risk that they will be the one not to confess. This produces a joint outcome which is worse than the one which would occur if they cooperated with each other and did not confess. This game has been used in a variety of settings, particularly as a model for understanding the 'tragedy of the commons' (Hardin, 1968), or the depletion of the world's finite resources such as the Amazon rainforest and fish stocks in the oceans.

If this game is played repeatedly over time then the prisoner's can learn to trust each other and avoid the heavier sentences by cooperating in a series of plays. Analytical or computable solutions to the problem of how cooperation develops in the sequential game are difficult to find because of the complexities of the problem. With this in mind, the political scientist Robert Axelrod (1984) organized a computer tournament between academic colleagues working in the areas of game theory and decision-making in order to try to find out which strategy is most likely to elicit cooperation in the sequential game.

The winner of this tournament was the relatively simple strategy called 'Tit for Tat,' which involves cooperating on the first move and continuing to do so if your companion reciprocates. However, if he defects by confessing then the individual should immediately retaliate by confessing on the next round of the game. If the companion subsequently reverts to cooperation, then the individual should again reciprocate on the next round. Since cooperation is rewarded and defection is punished then the likelihood of developing cooperative strategies over time is greatly increased. The strategy works because it is simple and easy to understand; it retaliates against bad behavior, but is also forgiving in the sense that it rewards a resumption of cooperation.

This research has generated a lot of experimental work on the problem of cooperation (Axelrod, 1997; Ridley, 1997; McCulloch, 2008) and it is valuable in highlighting the circumstances in which social capital can develop. It is more likely to develop in small communities in which individuals tend to know each other and interact frequently rather than in isolated hunter–gatherer societies where interactions are infrequent outside of the family and tribe. It helps to have sanctions against noncooperative behavior, but they should not be too inflexible and draconian. If sanctions are very large and inflexible they do not encourage a resumption of cooperation if it breaks down at some point, and noncooperative behavior is likely to become entrenched. One implication of this is that social capital is less likely to emerge in large cities where people are anonymous and frequently interact with strangers on a one-off basis. There is convincing evidence to support this hypothesis (Oliver, 2001).

Sociologists have drawn attention to the importance of the family and in the early socialization experiences of individuals in creating social capital. Bowlby (1988) for example, argues that infants benefit greatly from the relationship with their primary caregiver, most commonly the mother and a secure bonding between infant and mother will provide the basis for trust and cooperation in later life. Family breakup involving divorce tends to reduce generalized trust among children (Hall, 1999), the loss of one parent reducing emotional support and also access to wider social networks (Jonsson and Gahler, 1997).

Of course the family is embedded in a wider society, and so weak or fragmented relationships between families in communities will very likely reduce social capital. In a famous study of a village in southern Italy, Edward Banfield (1958) showed that societal relationships were characterized by 'amoral familism' or a persistent failure of individuals to cooperate with others outside of their immediate family. He attributed this in part to high levels of mortality and a fragmented system of land tenure which inhibited collective action. Other factors such as religion, national identities, and individual personality characteristics also influence the creation of social capital (Whiteley, 1999).

While stressing the importance of networks of voluntary associations in the creation of social capital, Putnam (1993) controversially argued that the origins of the large differences in social capital across the Italian regions lay in the medieval history of Italy. In 1231 Frederick II, who had established dominion over southern Italy, established a new constitution which cemented his hold over the region and entrenched an autocratic state (Putnam, 1993: 123). In contrast in the northern city-states such as Florence, Venice, Bologna, and Milan fledgling semidemocratic constitutions emerged, which involved significant citizen participation in government (Putnam, 1993: 124). These historical circumstances laid down different path dependencies in these regions and promoted cooperation between citizens in the north, and alienation between them in the south. This interpretation of what happened in medieval Italy has been challenged (Ferrangina, 2013), but it does draw attention to the importance of the social and historical context in explaining how social capital develops over time.

One of the criticisms of Putnam's analysis was that it played down the importance of inequality in influencing social capital. It has long been recognized that the middle class have larger and more diverse social networks than the working class, and the evidence suggests that the well-off are usually more trusting and more likely to volunteer than the poor (Goldthorpe et al., 1987; Hall, 1999). Edwards and Foley (1998) have drawn attention to the point that networks are only useful to the extent that they provide resources that individuals can utilize. Variations in power and resources in different types of network produce huge differences in the effectiveness of social capital, and **Figure 1** suggests that standards of living clearly influence social capital, and so this is an important point. As Halpern puts it:

> If you are a fairly well-connected middle-class person, it might be that the numerical size of your network may not be all that different from that of Bill Gates, Richard Branson, Prince Al Waleed bin Talal or whoever… But if we instead look at the resources that a billionaires network provides access to compared to our own, we would see a world of difference.
>
> 2005: 23

Overall, it is clear that both microlevel and macrolevel factors are at work in influencing the emergence of social

capital. Individual psychological predispositions play an important role, as does the family but they can be fostered or inhibited by the communities in which people live and by the governments in those societies. On a broader level they are influenced by relationships between the state and individuals. Governance influences social capital, just as social capital influences governance. At the same time many of the effects of social capital as far as policies are concerned are influenced by levels of inequality in society (Wilkinson and Pickett, 2011); so there is a clear link between inequality and social capital.

One of the reasons why the concept of social capital has attracted so much attention is because of the benign benefits which appear to be associated with it. We examine these next.

The Effects of Social Capital

Social capital appears to have major consequences for society. Research suggests that societies where levels of trust and voluntary activity are high tend to perform better on a variety of important social and economic indicators than societies where trust is low and voluntary activity is lacking (Knack and Keefer, 1997; Inglehart, 1999; Whiteley, 2000; Knack, 2002). Putnam (2000) shows that social capital in America delivers many benefits to localities such as greater economic prosperity, better health outcomes, reduced crime, improved education, and even higher levels of life satisfaction or subjective well-being.

To illustrate how these benefits can occur, it is instructive to take the example of social capital and economic growth. It is quite clear that interpersonal trust has a positive impact on economic growth in a variety of countries, controlling for other important determinants of growth (Whiteley, 2000). This works in a number of different ways. Firstly, widespread trust reduces the transaction costs of doing business; there is less need for complex contracts and expensive lawyers to draw up and enforce such contracts if individuals can trust each other. The example of the diamond market in New York is instructive in that regard. Secondly, if people are more trustworthy then crimes such as fraud and theft are likely to be less of a problem and that in turn reduces the costs of doing business. As well as lower costs of crime, social capital also reduces the costs associated with policing and law enforcement.

A third aspect of the effects of social capital is that principal–agent problems are likely to be lower in societies with high levels of social capital. These occur when an employer (the principal) has to trust an employee (the agent) to deliver a service without fully being able to supervise the quality of their work. In this situation, the agent has an incentive to shirk on the delivery of the service, which in turn lowers overall productivity and restricts growth. Finally, in high-trust societies individuals can afford to take more risks when they innovate, by investing in projects that might pay off but which are potentially money losers. Clearly, low-trust societies are likely to be more risk-averse than high-trust societies and this can hamper innovation and economic performance. Overall, we can see that there are many ways in which social capital can influence economic growth.

There is evidence of a decline in social capital in the United States since the 1960s, and after considering various alternatives Putnam (2000) claimed that television was responsible for this development. A similar argument might be developed about the Internet, given that it is now ubiquitous in Western societies. Television usage in the 1950s and 1960s and Internet usage in the 1990s and 2000s became very widespread. Both tend to isolate people in front of their screens, and reduce the time available for interacting and networking which are the mechanisms that promote social capital. Hence the rise of television and the Internet may be weakening social capital. Not surprisingly, though, these claims are controversial. Norris (1996) took issue with the argument that television has reduced social capital, and there is no evidence to suggest that Internet usage has limited activism in voluntary organizations such as political parties (Whiteley, 2011).

One early finding was that while social capital has been declining in the United States, it does not seem to have declined much in Britain (Hall, 1999). However, this conclusion was premature, since it was based on evidence available only up to the 1990s. Subsequent research has shown that social capital has declined in Britain in a similar way to the United States (Grenier and Wright, 2001). An illustration of this can be found in the World Values surveys conducted in 1981 and again in 2005 in Britain. In the 1981 survey some 43% of Britons thought that most people can be trusted, but by a quarter of a century later, this figure had declined to 30% (Whiteley, 2012: 85). A similar point can be made about voluntary associations in Britain. Church membership, party membership, the Women's Institute, and above all trade unions have all been declining in Britain over time (Whiteley, 2012: 83). While large charitable organizations such as the National Trust continue to attract members, these are largely 'cheque book' participants, who pay their dues but do not get actively involved (Jordan and Maloney, 1997).

Another finding relates to the link between local government and social capital in Britain. While spending on local services does not appear to be influenced by social capital, the policy outcomes derived from such services are affected. In the case of education, for example, which is the biggest item in the budget for English and Welsh local authorities, the best pupil performances in GCSE are found in areas with high levels of political participation and associational activity (Pattie et al., 2004: 189–223). Other case studies of real estate development and urban management schemes in the United States have shown that social capital significantly improves the quality of policy delivery (Saegert and Winkel, 1998; Glaeser and Sacerdote, 2000).

These findings all suggest that social capital is an asset to effective governance and policy-making, but its relationship to political participation is more ambiguous. Putnam (2000: 31–64) produced a lot of evidence to support the idea that social capital and political participation have both declined at the same time in the United States. But there are indirect links between social capital and political participation. The civic voluntarism model of political participation has long established that there is a correlation between social status, personal affluence, and education on the one hand and indicators of social capital such as trust and voluntary activity on the other (Parry et al., 1992; Verba et al., 1995). This model shows clearly that social status and affluence are determinants of political participation, which establishes an indirect link between social capital and participation. Translated to the level of the

community, this implies that greater affluence produces higher levels of social capital which in turn produces higher levels of civic involvement.

However there is an ambiguity in this relationship. Eric Oliver (2001) points out that social capital is not the same as affluence and that participation varies systematically across different communities in the United States in a way which is unrelated to their standards of living. In his analysis political participation is greater in economically mixed communities than in either very poor or in very rich communities, even when controlling for individual-level factors such as the survey respondent's income. A similar point can be made about racial mixed communities which tend to have higher rates of participation than homogenous white or black communities.

His argument is that individuals are more likely to have to protect their interests by getting involved in 'mixed' rather than in homogenous communities. Politics in diverse communities involve a battle between individuals with rather different interests, whereas in homogeneous communities the residents tend to have similar interests. If an individual has roughly the same interests as the rest of the people in their community, they can safely leave it to others to make decisions in the knowledge that they are likely to get the mix of taxes and spending they prefer. In this situation the risks of nonparticipation and the potential gains from participation are not that great. However, this is not the case in heterogeneous communities where the political process can produce large differences in policy outcomes. If individuals do not get involved in such communities they may well end up with policies which they regard as highly undesirable. This argument is a challenge to social capital theory, since it is often the most homogeneous communities which have the most social capital.

Conclusion – Where Does Social Capital Go from Here?

The research on social capital is rich and diverse and it contains a number of controversies which are to be expected in relation to such a broad and important concept. As we have seen there are microlevel foundations to social capital and it can be promoted or inhibited by the communities in which people find themselves, but also by broad historical events which can change the trajectories of different societies. It is no surprise for example to see in Figure 1 that ex-communist countries such as Slovenia, Poland, and Russia have less social capital than Western democracies such as the Netherlands and Sweden. Communism with its pervasive surveillance state and refusal to permit independent organizations undoubtedly weakened social capital. Also there is a good deal of evidence in the figure that authoritarian societies have less social capital than democratic societies, a point made by Putnam.

The key challenge for the social capital research agenda is to map out clearly where it is declining and where it is not, as a preliminary exercise to understanding the dynamics of social capital over time. In the past high-quality comparative data on social capital was hard to come by, but international research initiatives such as the European Social Survey, the International Social Survey Programme, and the World Values Surveys have made the task of mapping social capital in diverse societies much easier. On the other hand comparative longitudinal data are still hard to come by, but it is essential to have this if the long-term effects are to be better understood.

Bibliography

Adler, P.S., Kwon, S.W., 2002. Social capital: prospects for a new concept. The Academy of Management Review 27 (1), 17–40.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Axelrod, R., 1997. The Complexity of Cooperation. Princeton University Press, Princeton.

Banfield, E., 1958. The Moral Basis of a Backward Society. The Free Press, Glencoe, IL.

Berg, J.E., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games and Economic Behavior 10, 122–142.

Bordieu, P., 1986. The forms of capital. Readings in Economic Sociology xx, 280–291.

Bordieu, P., Wacquant, L.J.D., 1992. An Invitation to Reflexive Sociology. University of Chicago Press, Chicago.

Bowlby, J., 1988. A Secure Base: Clinical Applications of Attachment Theory. Routledge, London.

Bowles, S., Gintis, S., 2002. Social capital and community governance. The Economic Journal 112, 419–436.

Brehm, J., Rahn, W., 1997. Individual-level evidence for the causes and consequences of social capital. American Journal of Political Science 41, 888–1023.

Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Russell Sage Foundation, New York.

Claibourn, M.P., Martin, P.S., 2000. Trusting and joining? An empirical test of the reciprocal nature of social capital. Political Behavior 22, 267–291.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94 (Suppl.), S95–S119.

Coleman, J.S., 1990. Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge, MA.

Cote, S., Healy, T., 2001. The Well-being of Nations. The Role of Human and Social Capital. Organisation for Economic Co-operation and Development, Paris.

Edwards, B., Foley, M.W., 1998. Civil society and social capital beyond Putnam. American Behavioural Scientist 42, 124–139.

Ferrangina, E., 2013. The socio-economic determinants of social capital and the mediating effects of history: making democracy work revisited. International Journal of Comparative Sociology 54, 48–73.

Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. Hamish Hamilton, London.

Glaeser, E.L., Sacerdote, B., 2000. The social consequences of housing. Journal of Housing Economics 9 (1–2), 1–23.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Grimes, K., May 10, 2003. To trust is human. New Scientist 32–37.

Goldthorpe, J.H., Llewellyn, C., Payne, C., 1987. Social Mobility and Class Structure in Modern Britain. Clarendon Press, Oxford.

Grenier, P., Wright, K., 2001. Social capital in Britain: an update and critique of Hall's analysis. In: Presentation at ARNOVA's 30th Annual Conference. November 29th–December 1st.

Guth, W.R., Schmittberger, Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.

Hall, P., 1999. Social capital in Britain. British Journal of Political Science 29 (Pt 3), 417–462.

Halpern, D., 2005. Social Capital. Polity Press, London.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Inglehart, Ronald, 1990. Culture Shift in Advanced Industrial Society. Princeton University Press, Princeton, NJ.

Inglehart, Ronald, J., 1999. Trust, well-being and democracy. In: Warren, Mark E. (Ed.), Democracy and Trust. Cambridge University Press, Cambridge.

Jonsson, J.O., Gahler, M., 1997. Family dissolution, family reconstitution and children's educational careers: recent evidence for Sweden. Demography 34, 277–293.

Jordan, G., Maloney, W., 1997. The Protest Business: Mobilizing Campaign Groups. Manchester University Press, Manchester.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. The Quarterly Journal of Economics 112, 1251–1288.

Knack, S., 2002. Social capital and the quality of government: evidence from the Sates. American Journal of Political Science 46, 772–785.

Lin, N., 2001. Social Capital: A Theory of Structure and Action. Cambridge University Press, London and New York.
Lin, N., Erickson, B.H. (Eds.), 2008. Social Capital: An International Research Program. Oxford University Press, Oxford.
McCulloch, M.E., 2008. Beyond Revenge: The Evolution of the Forgiveness Instinct. Jossey-Bass, San Francisco.
Milner, H., 2002. Civic Literacy: How Informed Citizens Make Democracy Work. University of New England Press, Hanover.
Norris, P., 1996. Does television erode social capital? A reply to Putnam. Political Science and Politics 29.
Oliver, E., 2001. Democracy in Suburbia. Princetion University Press, Princetion, NJ.
Paldam, M., Svendsen, G.T., 2000. An essay on social capital: looking for the fire behind the smoke. European Journal of Political Economy 16, 339–366.
Parry, G., Moyser, G., Day, N., 1992. Political Participation and Democracy in Britain. Cambridge University Press, Cambridge.
Pattie, C., Seyd, P., Whiteley, P.F., 2004. Citizenship in Britain: Values, Participation and Democracy. Cambridge University Press, Cambridge.
Portes, A., 1998. Social capital: Its origins and applications in modern sociology. Annual Review of Sociology 24, 1–24.
Putnam, R.D., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press, Princeton.
Putnam, Robert D., 2000. Bowling Alone. Simon and Schuster, New York.
Rapoport, A., 1970. N-Person Game Theory: Concepts and Applications. University of Michigan Press, Ann Arbor.
Ridley, M., 1997. The Origins of Virtue: Human Instincts and the Evolution of Cooperation. Viking, New York.
Saegert, S.G., Winkel, G., 1998. Social capital and the revitalization of New York City's low-income housing. Housing Policy Debate 1 (9), 17–60.
Tarrow, Sidney, 1996. Making social science work across space and time: a critical reflection on Robert Putnam's *Making Democracy Work*. American Political Science Review 90, 389–397.
Tocqueville, Alexis de, 1990 (original 1840). In: Democracy in America, vol. 1. Vintage Books, New York.
Verba, S., Lehman Schlozman, K., Brady, H.E., 1995. Voice and Equality: Civic Voluntarism in American Politics. Harvard University Press, Cambridge, MA.
Whiteley, Paul, 1999. The origins of social capital. In: Van Deth, Jan, Marraffi, Marco, Newton, Kenneth, Whiteley, Paul (Eds.), Social Capital and European Democracy. Routledge, London.
Whiteley, Paul, 2000. Economic growth and social capital. Political Studies 48 (3), 443–466.
Whiteley, Paul, 2011. Is the party over? The decline of party activism and membership across the democratic world. Party Politics 17 (1), 21–44.
Whiteley, Paul, 2012. Political Participation in Britain: The Decline and Revival of Civic Culture. Palgrave-Macmillan, London.
Wilkinson, R., Pickett, K., 2011. The Spirit Level: Why Equality Is Better for Everyone. Penguin Books, London.

Social Capital in Organizations

Lorenzo Bizzi, California State University Fullerton, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social capital represents the set of interpersonal relationships which provide resources that create value for individuals in organizations. The most studied forms of social capital describe social relationships in terms of the tie characteristics and the position an individual occupies in the network structure of relationships. Social capital provides resources such as information, control, trust, and support. It affects attitudinal and behavioral outcomes and it is affected by individual dispositions and contextual factors. Social capital can be used to study intra-organizational and inter-organizational relationships.

Social capital refers to the set of interpersonal relationships which create value for individuals in organizations. The concept has been imported to organizational studies from sociology. Sociologists have always been interested in observing how individuals cluster in social groups and they began acknowledging that social relations may be beneficial for the individuals who are embedded in them. The core intuition behind social capital is that social relationships provide resources which can enable individuals to attain desired outcomes. While sociologists focused on studying a wide array of social relationships involving family, friendship, or affiliation to political and social groups, organizational theorists began examining social relationships at work. Scholars in social networks have elaborated on the concept of social capital to specify its forms and identify the way in which they affect individual outcomes. A fundamental contribution provided by social network theorists highlighted that social capital is not merely related to the number of dyadic relationships between pairs of individuals, but it relates to the way in which relationships combine in structural patterns.

There are different forms to specify social capital and different ways in which social capital creates value for individuals. In order to shed light on the concept and implications of social capital, the present article is structured as follows. The first part reviews and discusses the different definitions of social capital. The second part describes the forms of social capital and illustrates which kind of social relations are likely to create value for individuals. The third part explains how social relations create value, describing the resources that social relations give access to. The fourth part reviews the core beneficial outcomes stemming from the exploitation of the resources provided by social relations. The fifth part focuses on the sources of social capital. The last part extends the arguments of social capital to the relationships among organizations.

The Concept of Social Capital

Adler and Kwon (2002) provide the core intuition behind social capital: whereas market relationships are characterized by economic exchanges in which individuals trade goods or services for money, interpersonal relationships are characterized by social exchanges in which individuals exchange favors. Anytime an individual grants a favor to another individual, he or she will receive a 'credit' or goodwill which can be used as a resource to facilitate the attainment of personal outcomes in the future. Social capital theory suggests that interpersonal relations create value for individuals as they provide resources which can be used for the achieving desired outcomes. Yet, although the core argument of social capital is clear and scholars seem to have converging opinions about it, what exactly does the concept of social capital capture? Does it reflect the social relations, their value, or the resources? Scholars have not reached agreement on that, and several conceptualizations have been used to characterize social capital.

In the first conceptualization, social capital in organizations equals social relations, generally described as the structure of social networks. According to Galunic et al. (2012: p. 1215) "social capital is primarily about the relationships that actors maintain, characterized in terms of such features as number, strength, and density." Burt (1992: p. 9) identifies social capital as "friends, colleagues, and more general contacts through whom you receive opportunities to use your financial and human capital." Nahapiet and Ghoshal (1998) acknowledge that one of the most popular conceptualizations of social capital focuses only on the structure of network relationships developed in organizations.

In the second conceptualization, social networks are exogenous to the definition of social capital. Social capital and social networks are two independent but causally related concepts. More specifically, social networks are said to be the cause of social capital. In Inkpen and Tsang (2005) the authors specify that social capital originates from social networks. Gulati and Gargiulo (1999) specify that social networks create social capital and identify social capital as the assets tied to the manager's network. Burt (2001) explicitly mentions that network structures create social capital. Hence, if social networks create social capital, they are exogenous to its definition and social capital is conceptualized as the result of social networks. More specifically, social capital is identified with the resources which are provided by social networks. Adler and Kwon (2002: p. 23) argue that "social capital is the goodwill available to individuals and groups. Its source lies in the structure and content of the actor's social relations. Its effects flow from the information, influence, and solidarity it makes available to the actor." Bourdieu (1985: p. 248) defines social capital as "the aggregate of the actual or potential resources which are linked to possession of a durable network of more or

less institutionalized relationships of mutual acquaintance or recognition."

In the third conceptualization, social capital refers to both social networks and the resources provided with them. In the definition of social capital we find two elements: the social networks and the resources. Here social networks are no longer exogenous to the definition of social capital. From Nahapiet and Ghoshal (1998: p. 243), we read that social capital is defined as "the sum of the actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit. Social capital thus comprises both the network and the assets that may be mobilized through that network." We can read from Boxman et al. (1991: p. 52) that social capital is "the number of people who can be expected to provide support and the resources those people have at their disposal."

Although authors seem to disagree significantly on the theoretical conceptualization of social capital, when it comes to its empirical operationalization we find far more convergence. No matter how scholars define social capital, they most often tend to measure it by capturing the interpersonal relationships that individuals develop in organizations. The resources provided by social capital are mostly assumed to represent the explanatory mechanisms which relate interpersonal relationships to individual outcomes of interest. Yet, not all interpersonal relationships represent social capital. Coleman (1990: p. 302) specifies that "social capital is defined by its function. It is not a single entity, but a variety of different entities having two characteristics' in common: They all consist of some aspect of social structure, and they facilitate certain actions of individuals who are within the structure." As both Burt (1992) and Coleman (1990) acknowledge, social capital is therefore composed of the specific set of interpersonal relationships which create value for individuals in organizations.

The Forms of Social Capital

Which interpersonal relationships are likely to create value for individuals in organizations? Scholars have not reached consensus on which relationships constitute social capital and have developed conflicting arguments highlighting the benefits of different forms of social relations. We can identify three main levels that scholars have considered when studying social capital. The first is the tie level, at which social relationships are described by observing the characteristics of ties linking individuals. A popular way to study the characteristics of ties is through their strength, measured as the intensity or frequency of interactions among individuals. Granovetter (1973) theorized that weak ties tend to be particularly advantageous for individuals and highlighted the mechanisms through which they give access to opportunities. Yet, strong ties could be beneficial as well given that a strong bond between individuals facilitates sharing privileged experiences and information which would not be normally shared through weak ties (Altman and Taylor, 1973).

The second level is the local structure level, which identifies the relationships that an ego (the focal individual) has with his or her alters (the direct connections) and the relationships that these alters have among each other. Scholars are involved in an ongoing debate regarding which type of local structure is more beneficial to individuals. Burt (1992), on one side, argues that brokerage, in which the ego bridges 'structural holes' among unconnected alters, is beneficial to individuals. Coleman (1990), on the other side, argues that closure, in which an ego is tied to alters who are connected to each other, can be more advantageous. Scholars have embraced different positions elaborating alternative theoretical arguments on which type of local structure can create value for individuals and hence qualifies as social capital.

The third level is the global structure level, which captures the position that an individual occupies in the whole network of relationships in the organization. Network centrality has been used as prominent concept and it identifies the extent to which individuals occupy a central position in the network. There are different types of centrality. Betweenness centrality describes the extent to which individuals fall 'in between' others and captures the degree to which shortest paths connecting any pair of individuals in the network pass through the focal actor. Closeness centrality describes the extent to which an individual is 'close' to all others and observes the average path distance between the focal actor and all other actors in the network. Eigenvector centrality describes the extent to which individuals are tied to 'popular' alters, who have many connections. While there is active debate regarding which tie characteristics and which local structure allow more advantages, there is more convergence on the idea that a central position allows better access to resources than a peripheral position.

The Benefits of Social Capital

If we define social capital as the interpersonal relationships which create value for individuals through the provision of resources, it becomes important to identify which resources are provided by which form of social capital. This section identifies the most commonly referenced resources provided by social capital. The first and most prominent resource provided is information. Several forms of social capital provide informational advantages. At the tie level, weak ties are expected to be efficient in transferring information while strong ties are expected to be effective in transferring tacit and complex information (Hansen, 1999). At the local structure level, brokerage positions rich in structural holes are supposed to provide nonredundant information (Burt, 1992). While individuals in a closed structure will be exposed to the same information which redundantly circulates inside the network, individuals in an open structure will be more likely to gain timely access to diverse information. Closed structures with dense relationships are also supposed to give some informational advantages as circulation of information within a closed network tends to be particularly efficient (Coleman, 1990). Last, at the global structure level, network centrality leads to informational advantages as individuals can benefit from being positioned in the midst of multiples flows of information (Mehra et al., 2001).

A second resource provided by social capital is control. Control is typically provided at the local structure by brokerage positions, in which the individual can beneficially exploit the lack of connection between alters (Burt, 1992). Given that the alters cannot exchange information about the broker,

the broker benefits from a positional advantage which makes alters dependent on him or her, allowing negotiation leverage. Ideally, control over others can be exercised also by central positions in the global network as centrality could create conditions of dependence on the central individual while allowing that person to be relatively independent.

A third resource provided by social capital is trust. Trust is mostly provided at the local structure by closure (Coleman, 1990). While brokerage allows the ego to exercise control on the alters without being controlled by them, closure allows alters to exercise control on each other, hence discouraging opportunistic behaviors and facilitating the conditions for the emergence of trust in the network. Trust could also be provided at the tie level as strong relationships entail intimacy and generate the conditions for the development of trust among partners (Altman and Taylor, 1973).

A fourth resource provided by social capital is support. Social relationships can provide instrumental support through referrals. The social network of an individual does not only give him or her information, but it also diffuses information about the individual to all other persons in the network. That information can grant access to opportunities or help to build reputation within the organization. Burt (1992) acknowledges that individuals in brokerage positions are more likely to benefit from referrals. Centrality can also be beneficial in gaining support and in establishing positive attributions and beneficial reputation in the organization (Balkundi et al., 2011). Finally, social capital can provide emotional support as relationships help overcome situations of distress or transfer positive emotions through contagion dynamics (Totterdell et al., 2004). Such support could be provided by strong ties (Altman and Taylor, 1973) as well as by network closure at the local level (Totterdell et al., 2004).

The Outcomes of Social Capital

The resources provided by social capital are likely to be exploited by individuals in order to attain desired outcomes in organizations. Specifying the outcomes the individuals are capable to attain through their interpersonal relationships is fundamental to establishing whether these relationships constitute social capital or not. Social capital has been associated with attitudinal outcomes (Totterdell et al., 2004). Individuals have an innate need for social interaction and social relationships at work can provide the resources to satisfy their social needs and develop positive attitudes toward the job – such as job satisfaction – or toward the organization – such as organizational commitment. The resources of emotional support and trust provided by social capital could be particularly relevant for the prediction of attitudinal outcomes. However, the opportunity to exercise control also generates positive attitudinal responses on individuals. Tie characteristics, local structure and global structure can all exercise predictive effects on the individual's attitudes and more specifically on positive or negative affectivity (Totterdell et al., 2004).

Social capital has also been associated with behavioral outcomes and more specifically with individual performance (Mehra et al., 2001). Social capital could directly or indirectly affect individual performance. It could indirectly affect individual performance through the mediation of positive attitudes: social capital triggers positive attitudes, which fuel motivation and make individuals perform better. Social capital could also directly affect performance through the resources it gives access to. All the resources identified in the previous section are directly likely to explain the relationship between social capital and individual performance. All the forms of social capital identified above have been studied in relation to performance, employing alternative theoretical positions. It is important to acknowledge that social capital can affect different forms of performance. The resources of information, control, trust, or support could allow individuals to better achieve in-role performance, in which they can effectively execute their assigned tasks. The resources of social capital could also allow individuals to better achieve extra-role performance, in which they perform activities that create value for the organization but are not formally expected to be performed. For example, the social capital of individuals could affect their organizational citizenship behaviors (Bowler and Brass, 2006). Another type of behavioral dimension studied in social capital research is creative performance, given that the informational advantages of social relations facilitate the combination of knowledge necessary to generate creative ideas (Obstfeld, 2005).

Another important outcome of social capital is power, or influence. Social capital allows individuals to gain power and influence, which can be used to achieve desired outcomes (Brass and Burkhardt, 1993). In general, the resources of social capital which are more likely to give power and influence are control and information. Controlling others and avoiding the control of others gives one the opportunity for gaining power and exercising influence over others. The information advantage can create conditions of information asymmetries that give opportunity for exercising influence.

Last, social capital allows individuals to secure a wide range of career benefits, such as finding a job (Granovetter, 1974), getting a promotion (Seibert et al., 2001), or getting a higher pay (Seidel et al., 2000). Again, it is fundamental to acknowledging that social capital could directly or indirectly affect career benefits. It can indirectly affect career benefits through the intermediation of the previous outcomes – performance and power. More specifically, individuals may be promoted, get a new job, or get a higher pay because social capital enables them to perform better. Individuals can also achieve career benefits when they exploit the power granted by social capital to better negotiate career conditions. Yet, social capital can also directly affect career outcomes, without the intermediation of performance and power. The informational advantage of social capital can expose individuals to more opportunities for finding a job. The instrumental support mechanism, through the system of referrals, can help an individual find a job, get a promotion or get higher pay. Tie characteristics, local characteristics and global characteristics have been all studied in relation to career outcomes (Seibert et al., 2001).

The Causes of Social Capital

Having established that social capital is beneficial, it becomes relevant to explain the factors that allow individuals to develop it. Research on the causes of social capital is characterized by a tension. In fact, social network theorists initially attracted the

attention of scholars because social capital was claimed to be independent of individual characteristics. While traditional perspectives were explaining individual outcomes through individual attributes, research in social capital began proposing that individual outcomes were not only dependent on the characteristics of individuals but were also a function of the characteristics of the social environment in which the individual is embedded. Therefore, for instance, to understand why someone performs better than others, we do not simply have to examine the characteristics of the individual, but also look at his or her social capital. However, if social capital is developed as a result of individual characteristics, the significance of its role would be diminished. If social capital affects individual outcomes but is dependent on individual attributes, then social capital would just be an intermediate mechanism that helps explain why individual attributes transfer an effect to individual outcomes.

There have been studies suggesting that social capital can depend on the personality characteristics of individuals. Mehra et al. (2001) highlighted that self-monitoring may be the core personality variable which explains how individuals develop social capital. Individuals differ in their capacity to read social situations and to interact with others. Individual dispositions explain why some employees are capable of developing social capital while others are not. The relationship between individual attributes and social capital still needs significant investigation. Although it is necessary to acknowledge that individual dispositions may affect the formation of social capital, it is also important to recognize that individual dispositions may interact with social capital, which can enable or constrain the expression of individual characteristics. As Borgatti et al. (2014) state, it is incorrect to assume that the behavior of individuals depends exclusively on the structure of relations and it is inappropriate to claim that social capital theory denies the role of agency in individuals. Yet, it is probably also incorrect to assume that the behaviors of individuals depend exclusively on their individual characteristics and that the social environment is a mere result of individual dispositions. Individual agency and social structures are two factors which are likely to interplay dynamically in the determination of individual outcomes.

In order to better understand the factors that influence the formation of social capital, it is important to isolate the processes through which social capital is formed. Kilduff and Tsai (2003) posit that social relationships could be formed as a result of two core processes: goal-directed and serendipitous processes. In goal-directed processes, an individual has a specific goal, perceives the need to acquire information and forms relationships in order to help achieve this goal. We can assume that for this type of formation process, the work environment may play a fundamental role in shaping social relations. Brass (1981) has found that social relations are significantly influenced by the work interdependences in the organization, which regulate the need to relate to others and acquire information. In serendipitous processes, an individual spontaneously develops social relations with others without a preexisting instrumental need that motivates the individual to address others. We may argue that personality attributes or similarity among individuals play a role here in the development of social capital.

Social Capital among Organizations

The arguments developed focus on the individual as unit of analysis. Research in social networks in sociology has been mostly concerned with describing the relationships among individuals and the core contributions to the theory of social capital focused on the individual level of analysis. However, social capital cannot only describe how relationships affect individual outcomes but also how relationships affect organizational outcomes. Instead of looking at relationships inside the organization, scholars have elaborated on the concept of social capital and explored how relationships among organizations can affect organizational outcomes.

There are different ways to elaborate on the concept of social capital and explore the relationships among organizations. One way would be to examine the interpersonal relationships among individuals in different organizations. Another way would be to explore the affiliation of individuals with multiple organizations: interlocking directorates could be an example. A third way would be to examine inter-organizational relationships as alliances. The application of the concept of social capital at the organizational level of analysis follows a logic similar to that at the individual level of analysis. There are some significant similarities in observing the effects of social capital at the individual and at the organizational levels of analysis (Brass et al., 2004). A core idea to understand the application of social capital theory to the organizational level of analysis is that of embeddedness (Uzzi, 1997), according to which social relations tend to overlap with economic relations and therefore the study of social capital can effectively explain inter-organizational relations. Even inter-organizational relations are supposed to create value because they provide access to a specific set of resources. The forms of social capital also focus on the characteristics of inter-organizational ties and the organization's position in the local or the global structure. Tie strength, brokerage, closure and centrality all have been used at the individual level as well as at the organizational level. The resources provided by social capital are similar as inter-organizational relationships also give access to information, trust, control or support. There are different outcomes of social capital studied by scholars at the organizational level but the most studied is organizational performance (Baum et al., 2014) and innovation (Ahuja, 2000).

Conclusion

Research in social capital has been and remains one of the most prolific streams of investigation for organizational theorists. The number of publications and the variety of outcomes of social capital provide solid support to the original insight that interpersonal relations create value for individuals as well as for organizations. There are some new directions which researchers are pursuing to expand our understanding of social capital. The mixed evidence in support of different forms of social capital seems to suggest that there is no universal form of social capital which gives benefits to all individuals in any circumstance. For this reason, recent empirical investigations have started to examine the contingency conditions under which specific

forms of social capital are likely to be more or less beneficial (Reinholt et al., 2011). Scholars also have started exploring the downsides of social capital and acknowledging that social capital does not only provide resources, but can also be associated with constraints and limitations that negatively affect individual outcomes (Bizzi, 2013). Scholars have begun investigating new forms of social capital: new directions point at the potential for studying relational pluralism, in which individuals simultaneously share multiple types of relationships with the same individuals (Shipilov et al., 2014). Recent research has also been devoting more attention to the origin of social capital in order to better understand when and how valuable social relations can be generated (Ahuja et al., 2012). To conclude, social capital has already offered significant advancements to our understanding of organizations and promises to continue leading our empirical investigations and enriching our understanding of the social dynamics among individuals working in organizations.

See also: Exchange: Social; Group Processes in Organizations; Group Processes, Social Psychology of; Interorganizational Relationships and Networks; Network Analysis; People in Organizations; Social Network Analysis; Social Networks; Status and Role, Social Psychology of; Status and Role: Structural Aspects.

Bibliography

Adler, P.S., Kwon, S.-W., 2002. Social capital: prospects for a new concept. Academy of Management Review 27, 17–40.

Ahuja, G., 2000. Collaboration networks, structural holes, and innovation: a longitudinal study. Administrative Science Quarterly 45, 425–455.

Ahuja, G., Soda, G., Zaheer, A., 2012. The genesis and dynamics of organizational networks. Organization Science 23, 434–448.

Altman, I., Taylor, D., 1973. Social Penetration: The Development of Interpersonal Relationships. Holt, Rinehart and Winston, New York.

Balkundi, P., Kilduff, M., Hanison, D.A., 2011. Centrality and charisma: comparing how leader networks and attributions affect team performance. Journal of Applied Psychology 96, 1209–1222.

Baum, J.A.C., Cowan, R., Jonard, N., 2014. Does evidence of network effects on firm performance in pooled cross-section support prescriptions for network strategy? Strategic Management Journal 35, 652–667.

Bizzi, L., 2013. The dark side of structural holes: a multilevel investigation. Journal of Management 39, 1554–1578.

Borgatti, S.P., Brass, D.J., Halgin, D.S., 2014. Social network research: confusions, criticisms, and controversies. In: Brass, D.J., Labianca, G., Mehra, A., Halgin, D.S., Borgatti, S.P. (Eds.), Research in the Sociology of Organizations, vol. 40. Emerald Publishing, Bradford, UK, pp. 1–29.

Bourdieu, P., 1985. The forms of capital. In: Richardson, J.G. (Ed.), Handbook of Theory and Research for the Sociology of Education. Greenwood, New York, pp. 241–258.

Bowler, W.M., Brass, D.J., 2006. Relational correlates of interpersonal citizenship behavior: a social network perspective. Journal of Applied Psychology 91, 70–82.

Boxman, E.A.W., De Graaf, P.M., Flap, H.D., 1991. The impact of social and human capital on the income attainment of Dutch managers. Social Networks 13, 51–73.

Brass, D.J., 1981. Structural relationships, job characteristics, and worker satisfaction and performance. Administrative Science Quarterly 26, 331–348.

Brass, D.J., Burkhardt, M.E., 1993. Potential power and power use: an investigation of structure and behavior. Academy of Management Journal 36, 44–470.

Brass, D.J., Galanskiewicz, J., Greve, H.R., Tsai, W., 2004. Taking stocks of networks and organizations: a multilevel perspective. Academy of Management Journal 47, 795–817.

Burt, R.S., 2001. Structural holes versus network closure as social capital. In: Lin, N., Cook, K., Burt, R.S. (Eds.), Social Capital: Theory and Research. Sociology and Economics: Controversy and Integration Series. Aldine de Gruyter, New York, pp. 31–56.

Burt, R.S., 1992. Structural Holes. Harvard University Press, Cambridge, MA.

Coleman, J.S., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.

Galunic, C., Ertug, G., Gargiulo, M., 2012. The positive externalities of social capital: benefiting from senior brokers. Academy of Management Journal 55, 1213–1231.

Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Granovetter, M., 1974. Getting a Job. A Study of Contacts and Careers. Harvard University Press, Cambridge, MA.

Gulati, R., Gargiulo, M., 1999. Where do interorganizational networks come from? American Journal of Sociology 104, 1439–1493.

Hansen, M.T., 1999. The search-transfer problem: the role of weak ties in sharing knowledge across organization subunits. Administrative Science Quarterly 44, 82–111.

Inkpen, A.C., Tsang, E.W.K., 2005. Social capital, networks, and knowledge transfer. Academy of Management Review 30, 146–165.

Kilduff, M., Tsai, W., 2003. Social Networks and Organizations. Sage, London.

Mehra, A., Kilduff, M., Brass, D.J., 2001. The social networks of high and low self-monitors: implications for workplace performance. Administrative Science Quarterly 46, 121–146.

Nahapiet, J., Ghoshal, S., 1998. Social capital, intellectual capital, and the organizational advantage. Academy of Management Review 23, 242–266.

Obstfeld, D., 2005. Social networks, the tertius iungens orientation, and involvement in innovation. Administrative Science Quarterly 50, 100–130.

Reinholt, M., Pedersen, T., Foss, N.J., 2011. Why a central network position isn't enough: the role of motivation and ability for knowledge sharing in employee networks. Academy of Management Journal 54, 1277–1297.

Seibert, S.E., Kraimer, M.L., Liden, R.C., 2001. A social capital theory of career success. Academy of Management Journal 44, 219–237.

Seidel, M.-D.L., Polzer, J.T., Stewart, K.J., 2000. Friends in high places: the effects of social networks on discrimination in salary negotiations. Administrative Science Quarterly 45, 1–24.

Shipilov, A., Gulati, R., Kilduff, M., Li, S., Tsai, W., 2014. Relational pluralism within and between organizations. Academy of Management Journal 57, 449–459.

Totterdell, P., Wall, T., Holman, D., Diamond, H., Epitropaki, O., 2004. Affect networks: a structural analysis of the relationship between work ties and job-related affect. Journal of Applied Psychology 89, 854–867.

Uzzi, B., 1997. Social structure and competition in inter-firm networks: the paradox of embeddedness. Administrative Science Quarterly 42, 35–67.

Social Categorization

Craig McGarty, University of Western Sydney and Murdoch University, Australia
Kenneth I Mavor, School of Psychology and Neuroscience, University of St Andrews, St Andrews, Fife, UK
Daniel P Skorich, Research School of Psychology, The Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Categorization is a process by which people make sense of things by working out similarities and differences. Social categorization is especially important because it explains the way that people make sense of other people and themselves. Studying social categorization helps us to understand the way that people perceive groups and form impressions of individuals. The big ongoing question in the study of social categorization relates to whether this is a process that works primarily by reducing information overload or generating new information.

Categorization is a psychological process that people use to make sense of the world around them. The sensemaking power of categorization comes from the idea that people understand things by working out in what ways those things are similar to some things and different from others. The process is widely studied in both cognitive and social psychology but the real interest for social psychology comes from the fact that people categorize people. That is, categorization is not just the way we make sense of objects in the world around us but it is also the way we make sense of people including ourselves. In other words, our sense of self, or our identity, is derived from a process of understanding who we are similar to and who we are different from.

Categorization is complicated, however, because as well as being a psychological process that psychologists can study it can also be used to refer to an operational process that psychologists can engage in as a scientific or quasiscientific activity. That is, just as people in their everyday lives can perceive categories of people to be, say, ingroup and outgroup members, then so can psychologists be involved in classifying people on the basis of personality type, mental disorders, roles, occupations, and abilities. Psychologists no longer participate in scientifically dubious practices that involve classifying people in terms of race.

One way of expressing this complication is to suggest that there are categories of the mind and there are categories of the world (such as the biological taxonomical category of species). Both sets of categories are important, but it is even more important to try to be clear about which sort of category we are talking about at a particular time. Categories of the world may seem to be more real and verifiable (though, of course, these can be updated as scientific understandings change as they have with the conception of race and as they do from time to time with biological species), but categories of the mind are also real because they have real consequences for the way that people understand the world. Indeed the scientific enterprise of classifying things in the world is made possible by the psychological process of categorization. By acknowledging this and by recognizing that even what seems like the strongest scientific categories are often revised as new evidence becomes available, we are reminded that it is not an easy task to judge the validity and utility of our own or other people's categories. A complex world affords many ways of understanding the same thing, and it is this capacity for mental flexibility rather than rigidity that is the hallmark of categorization as a psychological process.

Part of this flexibility comes from the power of categorization to deal with abstract ideas. A scuffle in the street between two people may be understood as interpersonal violence but if one person is a police officer and the other is a protester then it may be interpreted as routine law enforcement or as a political conflict between the government and its opponents, as state repression of free speech, or as police protecting public order and safety, and so on. All of these ways of making sense of the action rest upon discerning and applying categories and linking the categories to the interpretation of the world.

Social categorization has an additional level of complexity in social psychology because when we categorize people we also categorize entities that can categorize themselves. Consider again the scuffle in the street. A fight between two police officers or two protesters may have a completely different meaning to a fight between police and protesters. In order to understand a conflict between two people who belong to the same category, we need to understand how those people come to see themselves on different sides. In order to understand why those two people are fighting each other (rather than, say, both fighting someone else), we need to assume that those people see each other as different in some relevant way. Our system of categories helps us to understand these issues, and where they do not work well, we need to update them on the basis of new information. Violence between two police officers at a street protest might be explained in terms of a personal dislike or disagreement between the officers, but personal dislike is unlikely to be a good explanation of intergroup conflict. Generally speaking, interpersonal factors tend to provide poor explanations for intergroup processes, and to use intergroup explanations we need to categorize people into groups.

The examples we have used point to the importance of self-categorization of both perceivers and the people they perceive. The explanatory power of self-categorization rests on the power to capture possibilities. By this we mean that self-categories can and do reflect the groups we belong to now but they also help frame expectations about future

relationships. We would expect that two police officers who have a fierce disagreement at a street march could be reconciled if they were to be effectively reminded of their common duties, role, and mission as police officers. A political campaigner can promote a cause by breaking the unity of the other side and recruiting former opponents to join their side. This is only possible if the campaigner is able to imagine alternatives where people who were opponents can be seen to be on the same side. In other words, social categories are informative about selfhood (who we are) but also about possible selves (who we could be).

In this introduction we have sought to point to some of the rich nuances of complexities of social categorization. With this background in place we now turn to some of the core areas of research concern in the area. We frame this discussion in relation to two broad families of research on social categorization: the impression formation family (forming impressions of individuals taking account of their group memberships) and the group representation family (forming impressions of groups).

Before we do this, it is helpful to provide some background on some topics that have informed discussion of categorization in cognitive psychology. These topics are category representation and structure, and category learning and use.

Storing and Using Categories

Three ideas have dominated considerations of category representation and structure. These are the classical view, the probabilistic or prototype view, and the exemplar view. The classical view is that categories have all-or-none defining features that allow one category to be utterly discrete from other categories. While such categories would be logically useful it is difficult to find evidence of rules that define such categories. Wittgenstein's famous example was the concept of game where he set the challenge of defining a set of rules that include things as different as the Olympic Games and noncompetitive children's amusements all as games. The real structure of categories appeared to be that they were fuzzy in that instances were not simply members or nonmembers of a set but that some members were good instances and some were less good, and that instanced might belong to one or more set.

Two successor views to the classical view have dominated since the 1970s. The prototype view championed in the work of Rosch and colleagues holds that categories are based around an abstract summary or ideal called a prototype, and that the variation in typicality in instances is explained in terms of the discrepancies between this abstract summary and those instances: some instances are not seen to be typical members because they are very different to the prototype.

The second major alternative view was the exemplar approach. This approach suggested that a category was not stored in terms of a single abstract summary but as a set of instances (exemplars) that were encoded as traces in memory. Exemplars that were most similar to other exemplars on average would be seen to be more typical. Exemplar-based models were broadly introduced into social psychology by Smith and colleagues (e.g., Smith and Zarate, 1992) but were prefigured in work on outgroup homogeneity.

Both major approaches explained variation in typicality but they also had problems. The cognitive architecture of prototype models seemed implausible in terms of the storage of the prototype in some way separate to a set of instances. The exemplar models converged with work pointing to the distributed nature of cognition but did not offer an account as to why particular sets of instances hang together.

Some solutions were offered in relation to models that pointed to variable or mixed representation. In social psychology, self-categorization theory (SCT) offered a good example of this by positing that people could be perceived to a greater degree as individuals or as group members at different times.

At the same time interest increased in intuitive theory and knowledge as factors that could bind categories together. This was present in cognitive psychology through the work of Medin, Murphy, and Barsalou and others and through a variety of approaches in social psychology (in all cases reflecting, to greater or lesser degree, a rediscovery of the work of Bruner, 1957). A classic finding of this stream of ideas was that a disc that was midway in size between a typical pizza and a typical coin was more likely to be seen to be a pizza because people had the theory that coins were restricted by laws to particular sizes but pizzas could vary with the cook's preference.

The cornerstone of the scientific understanding of category use is Bruner's (1957) perceptual readiness formulation. This formulation suggests that the use of a category – variously referred to as the activation, application or salience, of a category – is constrained by two interacting sets of factors: those that the perceiver brings to the situation, and those present in the stimulus and/or the surrounding environment. In Bruner's model, the first set of factors is captured by the concept of *accessibility*, which refers to the *readiness* of the perceiver to employ a particular category relative to other categories. The second set of factors is captured by the concept of *fit*, which refers to the match between stimulus specifications and a particular category relative to other categories. If a category is particularly accessible, it will require a smaller degree of fit to become activated than a category that is less accessible. If a category is particularly fitting, it will require lower accessibility to become activated than a category that is less fitting. Accessibility and fit therefore jointly determine the activation and use of categories.

Bruner's "accessibility by fit" account of the category use process remains widely accepted with refinements. Higgins (1996), for example, argues that the perceiver factors can be further broken down into *availability* and *accessibility*, while the situational factors can be further broken down into *applicability* and *salience*. Accessibility and applicability refer broadly to the same processes captured by Bruner's accessibility and fit concepts, respectively. Availability refers to the presence in memory of a construct, and salience refers to those properties of a stimulus that draw or hold attention relative to other features. In both cases, these added factors constrain category use through their interaction with the other perceiver and situational factors.

The SCT (Turner et al., 1987) account of category use suggests that a category will become salient as a function of an interaction between *perceiver readiness* and *comparative* and *normative fit*. Perceiver readiness refers to the accessibility of

particular categories in memory relative to others, but also captures the perceiver's long-term theories with regard to categorizing the world, as well as the perceiver's current expectations, motivations, values, goals, and needs. Comparative fit refers to the degree to which stimuli that covary with an organizing principle (such as a category label) are on average more similar to one another than those stimuli are to all the remaining stimuli in the perceiver's current psychological frame of reference. Normative fit refers to the degree to which the category specifications defined by the comparative relations are in the direction expected by the perceiver, based on his or her background theories and previous experiences. The salience model proposes that categories that the perceiver is particularly ready to deploy will require lower comparative and normative fit to become salient than categories that the perceiver is less ready to deploy. Similarly, when comparative fit within a stimulus field strongly favors a particular category, then the perceiver need not be as perceptually ready to deploy that category for it to become salient. The key point within all models that build on the perceptual readiness formulation is that the interaction of perceiver factors and situational factors will act together to determine category use.

Forming Impressions of People

Although category use is widely accepted to function in the way described by the perceptual readiness formulation, some researchers argue that certain types of categories will be automatically accessed and/or automatically applied, irrespective of readiness or fit-type factors. These researchers – who variously describe human perceivers as cognitive misers, mental sluggards, motivated tacticians, and efficiency experts – argue that the categories of race, age, and gender are particularly likely to be used, and that their use will occur involuntarily and outside the perceiver's conscious awareness. From this perspective, categories are thought to be simplifying or streamlining devices that allow a perceiver to quickly and easily relate to a novel person based on that person's group membership. Race, age, and gender are thought to be used automatically because they apply to almost any novel person a perceiver might encounter, because they are visually available immediately upon encountering a novel person, and because the stereotypes associated with them provide perceivers with rich but streamlined chunks of prior knowledge on which to base their judgments.

Research in this tradition has focused on impressions of people as unique exemplars or as group members, depending on the perceivers' goals and motivations. A particularly prominent model is the continuum model of impression formation developed by Fiske and Neuberg (1990). This model endorses the notion that categorization is an effortless process employed by perceivers to simplify perception. The model argues, however, that categorization is only one of two complementary processes in the perceiver's social cognitive toolkit, the other being the process of *individuation*. Individuation is said to be a more effortful, richer process than categorization, one that focuses the perceiver on a target person's individual-level characteristics rather than his or her group memberships. The model posits that impression formation proceeds along a continuum, anchored at one end by categorization and at the other end by individuation, and that movement along the continuum is mediated by a perceiver's motivations, and his or her resultant deployment of available cognitive resources. Perceivers who are unmotivated to form accurate impressions of a target person will use group memberships to make sense of the target, remaining at the proximal end of the continuum. When perceivers are motivated, on the other hand, they will deploy their limited cognitive resources and attempt to form a more individual-level, attributed-oriented impression of the target, thus moving further toward the distal end of the continuum.

A key consideration within the impression formation family of research has been cognitive load. Given that the deployment of cognitive resources is thought to be central to a perceiver's transitioning from categorization to individuation, the availability or otherwise of those resources prior to the deployment of them has been a central focus of research. Research has consistently shown that stretching a perceiver's cognitive resources – known as imposing a *cognitive load* – leads the perceiver to refer more heavily to stereotypes associated with the target's group memberships when describing the target than when the perceiver's resources are not stretched. A typical example of this finding can be found in the experiments of Gilbert and Hixon (1991).

Despite the wealth of research supporting the idea that categorization is driven by a need to reduce cognitive load other views have emerged. The meaning-seeker approach, which arose from the social identity tradition, assumes that perceiving a target in terms of his or her group memberships is an adaptive, useful, and valid way to structure one's understanding of the social world, and not merely a way to deal with complexity, nor simply a way to effectively balance accuracy and efficiency concerns. Categorization may be an effortful process and not simply an energy saving device.

A recent challenge to the primacy of cognitive load has recently been made. Skorich and Mavor (2013) have pointed out that the traditional understanding of social categorization as an efficient group-based process is problematic, because it hides the fact that the source of categorization's purported efficiency can be dissociated from the group-level of representation at which categorization is usually assumed to function. These researchers argue that the categorization and individuation concepts, as traditionally discussed and experimentally operationalized, can be thought of as confounding a group/person dimension and a memory-based/data-driven dimension. Given that the term 'categorization' is always used to mean both 'group' and 'memory-based' and that the term 'individuation' is always used to mean both 'person' and 'data-driven' it is impossible to determine which dimension is the source of the efficiency advantage of categorization relative to individuation. Skorich and Mavor were able to demonstrate that cognitive load privileges memory-based over data-driven processing, not group-level over person-level processing. These findings call into question both the efficiency of categorization as it is traditionally understood and also the traditional 'group as category' understanding of categorization.

Also arising from the impression formation family has been a line of research exploring whether the processing of faces might be related to the efficiency of categorization. Macrae and

his colleagues (e.g., Macrae et al., 2005; Pendry and Macrae, 1994) have shown, for example, that (1) diagnostic cues in a face are important for sex and race categorization, particularly under suboptimal (blurred or inverted) processing conditions, (2) mere exposure to a target face does not necessarily lead that target to be categorized by its sex, (3) multiple social categories are processed within a face without impacting on the categorization outcome, (4) manipulating diagnostic cues can lead participants to falsely respond that a male is female and vice-versa, and (5) perceivers are faster at responding that a face is male or female than they are at responding that a face belongs to a familiar or unfamiliar person, particularly under suboptimal (blurred or inverted) processing conditions. All of these findings have been related back to the idea that categorization is a simple and efficient process, and that its simplicity and efficiency derives from the immediate visual availability of categorization cues, in the face.

Forming Impressions of Groups

An issue of central importance to both prototype and exemplar theories of categorization is the variability within and between categories and what that variability is attributed to. Interest in the perception of social category variability developed extensively under the banner of group homogeneity, with a particular emphasis on the outgroup homogeneity effect (OHE) and associated implication for stereotyping. The source of those perceptions of variability could logically arise from three sources: the target, the perceiver, or the context, which are the focus of three different theoretical perspectives.

The concept of entitativity focuses attention on the target of perception itself; that is a target (group or person, for example) might be perceived to have more or less entitativity (Hamilton and Sherman, 1996). A different literature explores the individual differences in the expectations of the perceiver. The implicit, lay theory of person perception (in the work of Dweck, Levy, and colleagues) suggests that perceivers might be predisposed to think of persons as being either very stable or quite variable, and this perspective has been extended to group perception as well. The third possibility is that perceptions of variability arise primarily from the context in which the judgments are being made. SCT has particularly focused on this aspect as a conceptual complement to the other perspectives though SCT, in fact, broadly embraces all three possibilities. Work on the illusory correlation effect also focuses attention on the context in which the stimuli are perceived, though in quite different ways. We explore in more detail the notion of social category homogeneity in the following section, and then further perspectives based on the target, the perceiver, and the context.

Interest in homogeneity and the link to stereotyping goes back to early twentieth century work on perceptions of the relative homogeneity of stereotypical perception of actual groups, but the observation that other groups tended to be rated higher in homogeneity (and hence stereotyping) than one's own group led to a large body of work on the OHE as an explanatory process in stereotyping.

Several explanations were proposed. One well-explored possibility was that we are more familiar with ingroups than outgroups and that familiarity increases perceptions of heterogeneity (Linville et al., 1986), or that the greater amount of information available for ingroups made it more robust to forgetting effects. Relatedly, outgroups were proposed to be represented primarily in the form of prototypes, while ingroups might be represented as a mixture of prototype and exemplar representations (Park and Rothbart, 1982). These explanations seemed consistent with a tendency to find null effects for gender groups, where presumably we are more equally familiar, compared to racial and other types of groups, but ultimately some gender effects have been found, undermining the importance of familiarity explanations. An alternative, contextual explanation was made by S. Haslam et al. (1995) who argued that in many OHE studies, judgments were made either for ingroups alone (which would be an intragroup context) or of an outgroup (which would be intergroup). The intragroup context may create a focus on variability whereas intergroup may focus on homogeneity. As this line of research developed, several reviews and meta-analyses appeared, but no clear universal explanation emerged. The 'effect' is quite small, which undermines any simple explanation, but most likely several processes, in particular status, personal identity salience, and group size, can contribute to perceptions of group homogeneity.

Other developments of the homogeneity literature explored different ways of understanding the notion of variability. Early work by Linville and colleagues tended to focus on the proportion of group members holding some characteristic, but subsequent work explored the possibility that we might represent variability information in the form of distributions of exemplars or that we have a more complex understanding of the interrelations of attributes associated with different subtypes of a category.

The Entitativity Continuum

A different approach to thinking about category representation developed around the idea of the entitativity of a social target. Campbell (1958) developed the idea of entitativity, based on Gestalt theory concepts, to capture the extent to which the characteristics of a group would make it more likely that an aggregate of members would be seen to have some collective quality – to be seen as an entity. Hamilton and Sherman (1996) revived the entitativity concept as a way of exploring two key issues: the extent to which person and group targets could both be seen as entitative or not, and the implications for underlying cognitive representation and judgment. Hamilton and Sherman argued that when a target is seen as an entity, perceivers are more likely to make online judgments about that target. As each piece of information is processed, it is incorporated into an integrated representation of the target and that any affective judgment of the target will also be constructed during the impression formation. By contrast, if a target is not seen as particularly entitative when impressions are formed, then the pieces of information are more likely to be loosely organized in memory and not connected to a central concept of the target. When perceivers are subsequently in a position requiring a judgment of the given target, those who processed the impressions online would have a better overall memory of

the specifics and a tendency to recall earlier material (which would be closer to the core of the representations in memory). Those who did not process the impressions online would have to actively assemble the impressions at the time of (memory-based, or retrospective) judgment, leading to lower overall recall and more reliance on recent impressions.

The studies by McConnell and colleagues supported several key assertions from Hamilton and Sherman (1996). Person and group targets could both be seen as having some degree of perceived entitativity and to vary along a common continuum representing a common perceptual process. This has implications for the recall of information about both types of targets. It was also possible to manipulate the expectation of entitativity, such that both person and group targets could be explicitly made to be seen as more or less entitative, with associated affects on memory processes and judgments, reflecting an important influence of the context. Finally, McConnell and colleagues found evidence to support the idea that a person target tended to be seen as more entitative by default (in the absence of their explicit manipulations) relative to a group target. Hamilton and Sherman (1996) interpret this latter finding as indicating that persons will be assumed to be entitative but that groups may vary more widely and this has led to a line of research to catalog the perceptions of entitativity of different kinds of groups (e.g., work groups, families, broad sociological categories) and in the process has provided a mechanism for exploring the underlying dimensions on which perceivers may attribute entitativity to a target. Sherman and colleagues have proposed that the online/memory-based processing distinction and the tendency for persons to be seen as higher in entitativity could be a basis for dual-process distinction between person and group impression formation.

There are also some important lines of qualification of these ideas. Crisp and Hewstone (2007) and others have explored the important moderating effects of multiple social categories. The key insight here is that perceiving persons in terms of more than one category membership can have moderating effects on categorization (that may, for example, encourage more nuanced and accommodating views of minorities). In a different vein, N. Haslam and colleagues have emphasized the importance of particular category content. Certain categories are perceived to have forms or qualities that imbue them and make them seem especially different from others. These essentialized categories include biological categories. Good examples of these essentialized categories appear in work on dehumanization where it has been argued that certain categories of people can come to be seen as lacking human qualities and thus to be more like animals or machines (Haslam, 2006).

Controversies and Enduring Issues

The study of categorization continues to offer a number of choice points or unresolved issues. One that we have referred to above is the issue of whether categorization is driven by information reduction or meaning creation and whether there are new intriguing ways to recast that question. A second issue that we have not dealt with here (but see McGarty, 1999) is the question of what form do organized systems of category structures take? The idea that categories are organized into hierarchies (so that lower level categories are completely contained within higher order categories) is appealing (and underpins the work of Dovidio, Gaertner, and colleagues on recategorization), but there are good reasons to doubt its applicability to many category systems of the world and of the mind.

The longevity of category structures and the extent to which they are fixed or stored in the mind is also a good point to debate. It is tempting to think of categories as structures that are stored and retrieved (like books in a library) but this is an outmoded view of cognition. It may be better to think of categories as structures that are spontaneously generated or constructed in response to context.

Finally, we have focused a great deal here on classic social categories that are based on sets of people who share a demographic feature. There are also social categories based on people who share opinion, emotion, or role as we have explored in some of our work. However, is it possible to go deeper than that to address 'social' categories that might not relate to different people sharing the same attribute but to the same person having different attributes that apply to that person at some times and in some contexts? Are the same principles of categorization applicable to these intrapersonal categories, or do instead the principles derived from studying the new forms of social categories lead to a reconsideration of the classic forms? Whatever the answers it is certainly true that the study of social categorization will remain one of the most fascinating fields in social psychology and beyond.

See also: Dehumanization; Essentialism; Identity and Identification, Social Psychology of; Leadership; Self-Categorization Theory; Social Cognition; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Bruner, J.S., 1957. On perceptual readiness. Psychological Review 64, 123–152.

Campbell, D.T., 1958. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. Behavioral Science 3, 14–25.

Crisp, R.J., Hewstone, M., 2007. Multiple social categorization. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 39. Academic Press, Orlando, FL, pp. 163–254.

Fiske, S.T., Neuberg, S.L., 1990. A continuum of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In: Zanna, M.P. (Ed.), Advances in Experimental social Psychology, vol. 23. Academic Press, San Diego, CA, pp. 1–74.

Gilbert, D.T., Hixon, J.G., 1991. The trouble of thinking: activation and application of stereotypic beliefs. Journal of Personality and Social Psychology 60, 509–517.

Hamilton, D.L., Sherman, S.J., 1996. Perceiving persons and groups. Psychological Review 103, 336–355.

Haslam, N.O., 2006. Dehumanization: an integrative review. Personality and Social Psychology Review 10, 252–264.

Haslam, S.A., Oakes, P.J., Turner, J.C., McGarty, C., 1995. Social categorization and group homogeneity: changes in the perceived applicability of stereotype content as a function of comparative context and trait favourableness. British Journal of Social Psychology 34, 139–160.

Higgins, E.T., 1996. Knowledge activation: accessibility, applicability, and salience. In: Higgins, E.T., Kruglanski, A.W. (Eds.), Social Psychology: Handbook of Basic Principles. Guilford Press, New York, pp. 133–168.

Linville, P.W., Salovey, P., Fischer, G.W., 1986. Stereotyping and perceived distributions of social characteristics: an application to ingroup-outgroup perception. In: Dovidio, J.F., Gaertner, S.L. (Eds.), Prejudice, Discrimination, and Racism. Academic Press, Inc., Orlando, FL, pp. 165–208.

Macrae, C.N., Quinn, K.A., Mason, M.F., Quadflieg, S., 2005. Understanding others: the face and person construal. Journal of Personality and Social Psychology 89, 686–695.

McGarty, C., 1999. Categorization in Social Psychology. SAGE, London.

Park, B., Rothbart, M., 1982. Perception of out-group homogeneity and levels of social categorization: memory for the subordinate attributes of in-group and out-group members. Journal of Personality and Social Psychology 42, 1051–1068.

Pendry, L.F., Macrae, C.N., 1994. Stereotypes and mental life: the case of the motivated but thwarted tactician. Journal of Experimental Social Psychology 30, 303–325.

Skorich, D.P., Mavor, K.I., 2013. Cognitive load privileges memory-based over data-driven processing, not group-level over person-level processing. British Journal of Social Psychology 52, 469–488. http://dx.doi.org10.1111/j.2044-8309.2012.02099.x.

Smith, E.R., Zarate, M.A., 1992. Exemplar-based model of social judgment. Psychological Review 99, 3–21.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-Categorization Theory. Basil Blackwell, Oxford.

Social Changes: Models

Gudmund Hernes, Fafo Institute, Oslo, Norway; and BI Norwegian School of Management, Oslo, Norway
Kenneth C Land, Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by G. Hernes, volume 21, pp. 14228–14233, © 2001, Elsevier Ltd.

Abstract

This article defines models of social change in terms of how established social forms – rites, tools, languages, structures, hierarchies, and rules, whether they are simple or complex – are maintained or changed. It commences by citing the distinction of social facts from individual actions, identifies three requirements of theories of social change, shows how the 'Population, Organization, Environment, and Technology' (POET) model can be used to understand the loci of change identified in different models of social change, focuses on actors and structures and actor assumptions and structure assumptions as key elements of models of social change, and notes the distinctions of intentions of actors from the effects of their actions and expectations from desired consequences. The article concludes by identifying examples of formal systems – mathematics, statistics, and computational algorithms – that have been used to develop and assess models of social change.

Introduction

There is an amazing multiplicity of social forms: rites, tools, languages, structures, hierarchies, and rules. How has this variety come about and why is it forever changing? Yet, most of the time, humans behave like those around them. How has this conformity developed? And why is it maintained?

Models of social change address these questions. What generates change, and is it differentiating or uniforming? What maintains established structures, whether they are simple or complex?

Classicists, such as Durkheim, Marx, and Weber, provided models of the transformation as well as the preservation of social forms and analyzed diversification as well as homogenization. This they did, not so much by general theoretical statements, as by laying bare the social logic driving specific historical passages. They can be used to highlight models of social change and provide paragons that present-day studies can be appraised against.

Individual Actions and Social Facts

Durkheim, in *Suicide* ([1897] 1951) sought to explain both constancy and variety. He asked: "Why are suicide rates so stable – indeed, more stable than ordinary death rates – and yet variable systematically between groups?" Why do they take on a new stubborn stability after marked shifts? Durkheim held stable but differing suicide rates to be social facts. They exist relatively independent of the individuals who compose the group on which they are measured. Social facts – e.g., the language we speak, the rites we share, or the streets we walk – endure, Durkheim argued, *sui generis*.

'Social facts' have an existence of their own, like a swarm of moths which can change its shape or shift its position – yet, of course, there are no social phenomena without individuals being part of them; no swarm without the ephemeral movements of individual moths. Put paradoxically, social facts – both changing and constant – are relatively independent of those on whom they depend. Individuals are held in place by their fellows and take their cues from each other. This is important in the study of social change; it is always mediated through individuals who are shaped by their physical and social environment before they can act on it, and who change it by reacting to circumstances that they have brought forth or that are transmitted from the past.

'Social change' is at the core of sociological analysis. Yet, few studies cover the whole cycle of analyzing "how [individuals] react to conditions of their own making and in so doing change these conditions" (Hernes, 1976, p. 513). Rather, they present snapshots of a condition, such as a poll of opinions or their causal determinants. The reasons are many; collecting data as a process unfolds is costly and time-consuming. Data from other sources do not easily lend themselves to analysis over time. Survey techniques often result in a kind of aggregate social psychology rather than studies of structural change; and when a survey restricts itself to "enumerating individual characteristics, [it] treats the individual as if [she/he] were detached from [her/his] environment and hence as an abstraction" (Boudon, 1971, p. 48). Indeed, the methodological successes of survey analysis can lead to conceptual failures. Consequently, most social scientists do not go into deep analysis of social change even if it is at the core of social analysis.

Requirements of Theories of Social Change

From the above observations, three logical requirements of models of social change can be identified (Hernes, 1976). First, it should be possible to use the same basic approach in studying both constancy and change. This implies that a stable structure should be seen as a process in (temporary) equilibrium and that the theory must include an explanation of why the forces or parameters governing a process themselves are regenerated. Structural change must be described in terms of the processes that generate the change. Structural stability must be explained in terms of the processes that not only maintain

the stability, but also maintain the processes that maintain the stability.

Some accounts are static, e.g., 'percentage unemployed.' Yet, from a model point of view, it not only makes better sense, but also makes for more interesting analysis, to conceptualize even 'static' phenomena in terms of processes. For example, if the number unemployed remains stable, this does not imply that the same individuals remain out of work – only that the flows of individuals in and out of employment cancel out. Hence stability can be viewed as a process in equilibrium (Coleman, 1964).

Secondly, models of social change should incorporate intrinsic sources of change. Changes in social systems may be induced from the outside, but all changes cannot be thus induced (Elster, 1971). Even when the impetus for change is imported, the mechanisms transmitting external impacts should be identified. Indeed, when an item is taken from the outside, it should be possible to employ the same kinds of explanations for its adoption as if it had been invented at home. Moreover, the sources of change should also be intrinsic to the theory, in the sense that effects of previous actions provide or change premises for further choices of action.

Lauriston Sharp (1952) gives an example in 'Steel axes for Stone Age Australians.' He describes how the aboriginal Yir Yoront group as late as the 1930s, lived an essentially isolated and independent economic existence based on Stone Age techniques. However, "their polished stone axes were disappearing fast and being replaced by steel axes which came to them in considerable numbers, directly or indirectly, from various European sources to the south" (Sharp, 1952, p. 17). Sharp discusses the changes in the lives of the group that resulted from the replacement of the stone axes around which their culture was built. For example, stone axes were an item of trade and hence important in intertribal relations, as well as in belief systems, celebrations, and totemic ceremonies. Men crafted the stone axes, which were a symbol of their masculinity; yet they were lent to women who used them in their heavy chores. He notes that "The shift from stone to steel provided no major technological difficulties" (Sharp, 1952, p. 20). Sharp then analyzes how this change of tools changed trade relations, gender interactions, and authority structures:

> The most disturbing effects of the steel axe, operating in conjunction with other elements also being introduced from the white man's several sub-cultures, developed in the realms of traditional ideas, sentiments, and values. These were undermined at a rapidly mounting rate, with no new conceptions being defined to replace them. The result was the erection of a mental and moral void which foreshadowed the collapse and destruction of all Yir Yoront cultures, if not, indeed, the extinction of the biological group itself (Sharp, 1952, p. 21).

Here the new technology came from the outside – the impetus for change was exogenous. But its adoption is explained in terms of simple principles at the level of the actors: steel axes were nearly free, good tools, simple to use, and requiring less effort – the same advantages that would accrue if they had been invented locally. But the effects of this *deus ex machina* changed the premises for further choices of action, in this case with devastating aggregate effects; it reduced women's dependence on men, it undermined belief systems, it wrecked havoc with traditional trade relations and their social embellishments.

Finally, as just seen, social change is mediated through individual actors. Hence the third requirement of theories of social change: they must show how macrovariables or structural characteristics affect individual motives, capacities, and choices; and how the choices of actors, in turn, change the macrovariables. For the Yir Yoront, adopting a new tool changed their social structure in ways that were no part of their intention.

Loci of Social Change and the POET Model

Next the loci of change need to be identified. At a most general level of analysis, a classical theoretical model originally developed in human ecology is useful in this task. This model identifies four elements of a human system at the societal level, that is, Population structure, Organization, Environment, and Technology, and hence termed the 'POET' model. Within the POET model, "An ecological account of social change is attempted by referring to such instigating factors as environmental change [E] (whether caused by man or other agencies), changes in size and composition of population [P], introduction of new techniques [T], and shifts in the spatial disposition or organization of competing populations [O]" (Duncan, 1959, p. 683).

At the broad population level of analysis, 'Population structure' [P] refers to the age, race, and sex composition of a society's population and to its spatial location and distribution. Population structure has very significant consequences for demographic processes of reproduction, survival, and growth, i.e., fertility, mortality, and migration and for social processes and structures such as social mobility and inequality.

'Organization' [O] is a multifaceted term that refers to the ways in which a society is organized to maintain itself, i.e., its 'sustenance organization.' A major dimension of sustenance organization is 'sustenance differentiation,' that is, the "arrangement of differentiated parts suited to the performance of a given function or set of functions" (Hawley, 1950, p. 178). Sustenance differentiation is a major component of the division of labor. A second dimension is 'functional interdependence,' which can form the other side of the division of labor by combining with sustenance differentiation (Gibbs and Poston, 1975). A third dimension is the 'volume of sustenance produced by the population,' i.e., "the degree of productivity of the particular configuration of sustenance activities" (Poston and Frisbie, 2005, p. 609), which may include agricultural productivity, retail services productivity, wholesale services productivity, personal services productivity, etc. A fourth dimension is the 'extent of efficiency of the sustenance organization,' which can be measured by the ratio of the amount of sustenance produced to the amount of inputs needed in the production process. A bigger ratio indicates higher efficiency. A final dimension is the "degree to which population members are engaged in sustenance-related pursuits" (Poston and Frisbie, 2005, p. 610). Related measures of this dimension can be the labor participation rate in the

population and the extent to which social status is ascribed or achieved.

The Environment [E] concept in the POET model does not comprise everything external to the phenomenon under investigation, "but only those externalities that, by virtue of the limits they set on the acquisition of sustenance, affect the life chances of an organized population with a given technological repertoire" (Poston and Frisbie, 2005, p. 615). Thus, the environment concept in the POET model is a set of conditions serving as limiting (or enabling) resource for the adaption and growth of societal populations, whose breadth depends on "the technological devices and modes of organization that prevail in a given population" (Schnore, 1958, p. 628; Poston and Frisbie, 2005, p. 615). Based on this definition, social and economic activities (e.g., unemployment rates, industry structure, levels of education and income etc.) are not part of the environment. Rather, the environment concept in the POET model has two broad and distinct dimensions: the physical and the social. The physical environment refers to climate, natural resources, topography, etc. The social environment includes factors, such as the size of a population, proximity to the nearest metropolitan areas, and government influences.

Among the four basic elements of the POET model, 'technology' [T] plays a critical role in the adaptation and growth of human populations to their environments because it sets the boundaries for possible social and economic development (Poston and Frisbie, 2005, p. 611). Technology has three dimensions: material features (tools, capital equipment, machines); information (knowledge, techniques, scientific discovery); and energy (Poston and Frisbie, 2005, p. 612). Gross Domestic Product (GDP) per capita often is taken as an overall index of the hardware, capital-equipment, and information dimension of technology of a society.

Application: Explaining Increasing Economic Inequality

The POET conceptual model originally was developed, and traditionally applied, at a broad level of social organization, the societal–human population level, with its focus on "… the organizational aspects of human populations that arise from their sustenance-producing activities" (Poston and Frisbie, 2005, p. 603). Its comprehensive conceptual framework can be used to organize and synthesize studies of large-scale social change. For instance, a societal trend in the United States and many developed countries since the late 1970s that has received considerable research attention recently, is growing economic inequality, a specific organizational aspect of human populations that arises from sustenance activities. Social scientists have observed this trend in various statistical measures of inequality for decades. Sufficient time has elapsed for the impacts of social changes to become evident, for time series of annual measurements to be constructed, and for conceptual models to be developed, estimated, and evaluated empirically, to identify the important social forces driving the inequality trend. These contributions incorporate elements of the POET model in various ways.

One of these studies is Kristal (2013), which analyzed the decline in labor's share of national income and the rise in the capitalist's share of profits. From the late 1970s through the 2000s, labor's share of national income declined by 6% across the US private (nongovernment) sector of the economy. This overall decline in labor's share resulted from a large decline (5–14%) in construction, manufacturing, and transportation combined with a small increase (2–5%) within the finance and service industries. Kristal's (2013) conceptual model identifies erosion in workers' positional power as the main factor leading to the decline in labor's share of profits and posits that this erosion was partly an outcome of technological change in the form of computerization that favored capitalists and high-skilled workers over most rank-and-file workers. Controlling several parameters of workers' positional power (unionization, capital concentration, import penetration, and unemployment), she studied the direct and indirect effects of computer technology on changes in labor's share in US nonagricultural private and manufacturing industries, finding considerable support for the model.

In brief, the key explanatory variable in Kristal (2013) is a form of exogenous technological change (a specific form of the T component of the POET model) that led to large changes in an income inequality measure of the organizational structure of sustenance activities (a specific measure of an element of the O component of the POET model) in the United States. By comparison, Lin and Tomakovic-Devey (2013) focus on the impacts of changes in industry-level financialization (another element of the O component) – the increasing dominance of the finance sector of the economy as measured by the ratio of financial receipts (interest, dividends, and capital gains) to business receipts from the selling of goods and services – on increasing income inequality in the United States for the years 1970–2008. Three measures of income inequality (labor's share of national income, the income share of top earners, and earnings dispersion among workers, all of which are elements of the O component of the POET model) were studied. Net of controls for deunionization of the labor force, globalization of the world's economy, and capital investment (three elements of the O component of POET) and technological change (a T component), Lin and Tomakovic-Devey (2013) conclude that financialization could account for substantial proportions of the changes across the decades in the three measures of inequality.

A third study, Dwyer (2013), highlights yet another factor to the explanation of economic restructuring and job polarization (an element of the O component of POET) in the US labor market, namely, the growth of care work (another element of the O component of POET). Care work jobs are defined as service jobs that contribute to the health, well-being, or development of other people, services that once were mainly provided within families and kin networks rather than in market exchange. Dwyer (2013) shows that much of the increase in job polarization from 1983 to 2007, as measured by the growth of jobs with pay levels in the lowest and highest job wage quintiles as opposed to middle-wage jobs, and especially the growth of jobs for women, is related to an increase in care work jobs, which also tend to be concentrated at lower and higher ends of scales of job skills and labor market power. The growth in care work jobs, in turn, has been driven by changes in the societal organization of sustenance activities such as increases in female labor force participation and

changes in family structures toward smaller localized kin networks (elements of the O component of POET) and a population age structure that is increasing at the older ages due to increased longevity (an element of the P component POET).

In brief, each of these three studies adds knowledge of the forces that have driven recent increased levels of economic inequality in the United States (with similar patterns having occurred in other developed countries). Yet, the loci of change in all three studies are well encapsulated within the POET conceptual framework: the locus of change in Kristal (2013) is an exogenous change in technology (T); Lin and Tomakovic-Devey (2013) highlight the importance of a change in sustenance organization (O); and Dwyer (2013) focuses on changes in the job structure that have been driven by changes in sustenance organization (O) and population structure (P).

Actors and Structures

While the preceding application focused on societal-level processes, the four basic elements of the POET model also are present at the intermediate and smaller levels of social organization down to the smallest level of social organization, the primary group. Even at this smallest level of social organization, the actors (population) and their social structures (networks of social ties and interactions) and the social structures (organizations) within which they operate (families, neighborhoods, voluntary associations, work places) can change. In addition, the environment of a primary group conceived of as enabling resources for interactions and continued existence/survival of the group, may be of key importance. Similarly, the technology available to the group may affect its forms of interaction, e.g., whether face-to-face or via electronic media.

At smaller (than population) levels of social organization, and for relatively short-time spans of analysis, models of social change may be constructed on the presumption that the environment and technology components of the POET model are fixed and unchanging, although possibly variable across a set of units to be analyzed. The minimal components of models of social change must, therefore, include two sets of abstract elements – actors and structures – to mirror what drives the changes in them (Hernes, 1989, 1999).

Hence, the first element to specify in a model of social change is the 'types of actors' assumed to operate. In the language of the theater, specifying the actors corresponds to the casting of a play. For example, in microeconomic theory two types of actors are distinguished: consumers maximizing utility by spending their resources to obtain goods, and producers maximizing profit by the quantities they market.

The second element to specify is the structure in which the actors are embedded – the staging of a play. For example, in microeconomic theory three main market structures are distinguished: perfect competition where the actors are so many and small that none of them can affect the price; oligopoly, where the producers are so few that they react to each others' decisions; and monopoly, where a single producer can set the price.

One reason for combining actors and the structure within which they operate, is to construct a model so that one can logically work out the systemic effects of actors' choices within the structure – the plotting of the play. The point is to answer questions about what happens to the actors or to the structure as a consequence of the actors reacting to the initial conditions – i.e., the logical sequence of consequences that flow from the assumptions made. In other words, we construct models of actors within a structure in order to understand what the actors are doing and why, to explain what results from their actions and why, and whether the results are intended or not – a logical narrative. A logical narrative may have logical flaws and hence be theoretically unsound. And one such logical narrative may exclude others: it simultaneously empowers our reasoning and limits our comprehension – both our understanding of the actors and the explanation of the results of their interaction.

A second purpose is to compare the logical outcomes of the model to a real-life sequence of events. A compelling argument may be empirically false – the logical narrative may not match the historical record. Hence, if the model is to provide a valid understanding or a tenable explanation, there must be a rough goodness of fit between what the model logically leads us to expect about what will transpire and what actually takes place in the real world. The model cannot just be plausible in theory – it must at the end of the day be admissible by the facts; the observed must agree with the inferred. If the model is ruled out by the facts, it must be discarded and another constructed from different assumptions that provide a more realistic understanding and credible explanation.

For example, the simple model of perfect competition implies that profits will tend toward zero. If we observe that profits in fact are high, we may modify the model to better account for this fact by changing the assumptions, e.g., by allowing producers to change consumers' attitudes (advertising) or by allowing for monopolization – i.e., discarding the model of perfect competition.

Specifying Actor Assumptions and Structure Assumptions

To give a logical portrait of the types of actors involved in a social process, the following kinds of questions must be answered:

1. What do they *want* – what are their preferences or purposes?
2. What do they *know* – what are their beliefs, their potential for changing them or for learning?
3. What can they *do* – what are their capacities or skills?

Sometimes answers are only rudimentarily specified, e.g., in models of contagion where contracting a disease clearly is not something sought but something caught.

Actor assumptions allow us to understand what actors can do, what they actually do and why, or how others react to them and why.

A similar set of questions must be answered to specify assumptions about the structure:

1. What alternatives or constraints do the actors face? These may span from opportunity for migration to household budgets. They include collective properties such as laws and norms and physical barriers such as roads or rivers.

2. What social positions can the actors hold – e.g., organizational roles. What are the correlates of positions – e.g., their status or the hierarchy of wages linked to a given division of labor? Rewards, rights, and burdens are structural correlates of positions – they also define, fix, or freeze the alternatives actors face.
3. What are the distributional characteristics of actors, as distinct from their individual manifestations? For example, the number of other actors may affect options as in perfect competition, or the age distribution may affect outlooks for marriage. The sex ratio and the population pyramid are 'population structural' characteristics.

Specifications of the actors' attributes are needed to understand what they do and why. Specifications of the structure – options and payoffs – within which the actors make choices and act, are needed to explain what happens and why. These two sets of assumptions are needed to construct the logical narrative of systemic effects.

Intentions versus Effects and Expected versus Desired Consequences

The distinction between actors and structure is useful when what happens deviates from what was anticipated. In 1936 Robert K. Merton wrote a famous article on *The unanticipated consequences of purposive social action* (Merton [1936] 1976). A key point is that actors' choices and purposive actions (intentions) may interact to produce unanticipated consequences (effects) that were not part of the intentions. This notion is very useful for describing models or mechanisms of social change. The net outcome of rational choices may be counterintuitive. The way to hell may be paved with good intentions and simultaneous choices based on the best of motives may be made counterproductive by some hidden hex. Sometimes actors are caught by what they have wrought; sometimes they do not master the world they have made. Such unintended effects, counterproductive actions, and counterintuitive results, which flow from the actors' choices rather than their intentions, often produce engaging logical narratives.

The simplest models of change are, of course, when purposive actors produce intended results. What appears is what was planned, what materializes is what was sought. However, sometimes what turns up is not what the actors had in mind. Sometimes their best intentions interact to produce self-defeating results, as when a stampede to get out blocks the exit. Sometimes selfish motives produce positive overall effects over and beyond their expectations, with improved welfare coming as a by-product, as if an invisible hand guides self-interest to produce efficient use of resources. As the argument runs in Adam Smith's classic text:

> He [every individual] generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. I have never known much good done by those who affected to trade for the public good. It is an affectation, indeed, not very common among merchants, and very few words need be employed in dissuading them from it (Smith [1776] 1937, Book 4, Chapter 2).

A key point for the analysis of social change is that what results, whether it appears as the compelling influence from a hidden hex or an invisible hand, flows from the actors' choices, not from their motives. That is why understanding the actors is not sufficient – explanation of the interworkings of their actions beyond their intentions, expectations, or belief is necessary. Simultaneous choices based on the best of motives may be counterproductive. The net outcome of rational choices may be counterintuitive. Silly choices may prove lucky strikes; bad motives may result in good fortunes. The way to hell may be paved with good intentions just as private vices may give public benefits (Mandeville [1714] 1989).

In fact, we can classify types of explanations from the answers to two questions: "Were the consequences expected or not?" and "Were they desired or not?" If the results coincide with what was desired, the actors have produced the intended results.

If the results were not expected, but are desirable when they appear, we have what could be called 'invisible hand' explanations, since they are of the kind illustrated by the quote from Adam Smith above.

The cases where the consequences are expected but not liked by the actors when they turn up, can be termed 'suboptimality' (cf. Elster, 1978): the actors know that when they act in their own interest, they nevertheless collectively end up in an undesired state – the joint outcome of their choices makes them worse off. 'The Tragedy of the Commons' illustrates the logic. A 'commons' is any resource that can be used as though it belongs to all. However, when rational agents use it, such a shared resource is doomed, though this is no part of the users' intent. Everyone gains from using it, none gains anything from not taking out what they can when others do. When utilized near carrying capacity, additional use will degrade its value to all. "Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons" (Hardin, 1968). The inexorable working out of the resource's ruin is Hardin's 'tragedy of the commons.' Social change is the consequence of the sum of self-interested actions. And all could be better off by restraints on their own actions. If such collective restraints are imposed by the actors themselves, this represents a change of political structure. This latter logic and model of structural change was first suggested by Thomas Hobbes ([1651] 1982) who called the commons 'the state of nature' and 'war of all against all,' whereas the joint establishment of an authority to keep all from self-defeating actions he called Leviathan (cf. Hernes, 1999).

The fourth case is where the actors misjudge what others will do. Intentions that could be realized under constant conditions are rendered impossible when everyone seeks to realize them and change the conditions. 'Hog cycles' is an

		Are the consequences desired?	
		Yes	No
Are the consequences expected?	Yes	Intention = consequence	Suboptimality
	No	Invisible hand	Counterfinality

Figure 1 Actions and consequences.

economic example: producers act on the same assumptions and overinvest or underinvest in concert – the results of their decisions appear too late to become premises for their decisions. Erosion is another example – the irreversible destruction of alternatives. With a term Elster (1978) borrows from Sartre, this could be called 'counterfinality.'

The four kinds of models are summarized in **Figure 1**. Sometimes the actors get caught by what they have wrought; sometimes they do not master the world they have made. But independent actions can also turn out to be blessings in disguise.

Formal Models of Social Change

Models of social change may be expressed in a variety of formal systems – mathematics, statistics, and computational algorithms. Space limitations do not permit a systematic exposition, but a few examples can be cited.

Social change models expressed in natural language generally are articulations of qualitative explanations and often are found in studies of broad historical social change, as in the classical studies of Durkheim, Marx, and Weber. Verbal models of social changes also often are starting points for the development of more formal models. Examples are Herbert Simon's (1952) model of verbal theory of the intensity of interaction among members of a social group, the level of friendliness among the members, and the amount of activity carried on by members within the group; and Land's (1970) model of Durkheim's ([1893] 1933) verbal theory of the causes of the division of labor in society. These formal models were expressed in the form of deterministic differential equations that model change in continuous time, in order to provide formal representations of the dynamics of change in the theories of Homans and Durkheim, which potentially could be subjected to quantitative empirical estimation and testing.

Since the publication of James Coleman's (1964) 'Introduction to Mathematical Sociology,' stochastic differential equation models that model dynamics in continuous time in probabilistic form, and, more generally, stochastic models of change in both continuous and discrete time (Bartholomew, 1982) have become widespread. These models often can be conveniently estimated empirically by application of statistical estimation methods such as maximum likelihood (ML) (Wasserman, 2004). Statistical models extending from simple cross-tabulations to regression models to structural equation models (SEM) (Bollen, 1989; Kaplan, 2009) are also widely used formal representations for the study of social change. Indeed, SEM modeling recently has been applied to the ML estimation of stochastic differential equations (Singer, 2012). In addition, a variety of statistical models for time series analysis (Jaffe, 2000) and pooled time series–cross section/panel regression models (Woolridge, 2002), including autoregressive, distributed lag/error correction models (Beck and Katz, 2011) have been extensively applied.

All of these stochastic/statistical models tend to model the temporal dimension as historical or 'period-time.' In addition to period-time, demographic models of population processes have focused on age-time in order to take into account the age dependence of many demographic and social processes, as in classical life table models, or on both age-time and period-time, as in stable population theory (Keyfitz, 1977). A third temporal dimension also often is distinguished in studies of demographic and social change, namely cohort-time, which leads to age–period–cohort (APC) statistical models of change (Ryder, 1965; Mason et al., 1973; Yang and Land, 2013).

The foregoing mathematical and statistical models often are applied to the study of large-scale social change. Models of social change at smaller scales that focus on the motives, expectations, and consequences of actions of individual actors often use network (Wasserman and Faust, 1994; Carrington et al., 2005), game theoretic (Osborne and Rubinstein, 1994; Camerer, 2003), or computational representations (Macy and Skvoretz, 1998).

See also: Action, Theories of Social; Collective Action; Durkheim, Emile (1858–1917); Human Ecology; Marx, Karl (1818–83); Organizations: Unintended Consequences; Sociological Theory; Weber, Max (1864–1920).

Bibliography

Bartholomew, D.J., 1982. Stochastic Models for Social Processes, third ed. John Wiley, New York.
Beck, N., Katz, J.N., 2011. Modeling dynamics in time-series-cross-section political economy data. Annual Review of Political Science 14, 331–352.
Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.
Boudon, R., 1971. The Uses of Structuralism. Heinemann, London.
Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton, NJ.
Carrington, P.J., Scott, J., Wasserman, S., 2005. Models and Networks in Social Network Analysis. Cambridge University Press, New York.
Coleman, J., 1964. Introduction to Mathematical Sociology. The Free Press, Glencoe, IL.
Duncan, O.D., 1959. Human ecology and population studies. In: Hauser, P.M., Duncan, O.D. (Eds.), The Study of Population: An Inventory and Appraisal. University of Chicago Press, Chicago, pp. 678–716.
Durkheim, E., [1893] 1933. The Division of Labor in Society. The Free Press, New York.
Durkheim, E., [1897] 1951. Suicide. The Free Press, Glencoe, IL.
Dwyer, R.E., 2013. The care economy? Gender, economic restructuring, and job polarization in the U.S. labor market. American Sociological Review 78, 390–416.
Elster, J., 1971. Nytt syn på økonomisk historie. Pax, Oslo.
Elster, J., 1978. Logic and Society: Contradictions and Possible Worlds. Wiley, Chichester, UK.
Gibbs, J.P., Poston Jr., D.L., 1975. The division of labor: conceptualization and related measures. Social Forces 53, 468–476.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243--1248.
Hawley, A.H., 1950. Human Ecology: A Theory of Community Structure. Ronald Press, New York.
Hernes, G., 1976. Structural change in social processes. American Journal of Sociology 38, 513–546.
Hernes, G., 1989. The logic of 'The Protestant Ethic'. Rationality and Society 1, 123–162.
Hernes, G., 1999. Real virtuality. In: Hedström, P., Swedberg, R. (Eds.), Social Mechanisms: An Analytic Approach to Social Theory. Cambridge University Press, Cambridge, UK, pp. 74–101.
Hobbes, T., [1651] 1982. Leviathan. Penguin, London.
Jaffe, R.A., 2000. Introduction to Time Series Analysis and Forecasting. Academic Press, New York.
Kaplan, D., 2009. Structural Equation Modeling: Foundations and Extensions, second ed. Sage, Los Angeles.
Keyfitz, N., 1977. Introduction to the Mathematics of Population with Revisions. Addison-Wesley, Reading, MA.
Kristal, T., 2013. The capitalist machine: computerization, workers power, and the decline in labor's share within U.S. industries. American Sociological Review 78, 361–389.
Land, K.C., 1970. Mathematical formalization of Durkheim's theory of division of labor. Sociological Methodology 1970, 257–282.
Lin, K.-H., Tomakovic-Devey, D., 2013. Financialization and U.S. income inequality, 1970–2008. American Journal of Sociology 118, 1284–1329.
Macy, M.M., Skvoretz, J., 1998. The evolution of trust and cooperation between strangers: a computational model. American Sociological Review 63, 638–660.
Mandeville, B., ([1714] 1989). The Fable of the Bees. Penguin, London.
Mason, K.O., Mason, W.M., Winsborough, H.H., Poole, W.K., 1973. Some methodological issues in cohort analysis of archival data. American Sociological Review 38, 242–258.
Merton, R.K., [1936] 1976. The unanticipated consequences of purposive social action. In: Merton, R.K. (Ed.), Sociological Ambivalence. The Free Press, Glenoe, IL.
Osborne, M.J., Rubinstein, A., 1994. A Course in Game Theory. MIT Press, Cambridge, MA.
Poston, D.L., Frisbie, W.P., 2005. Ecological demography. In: Poston, D.L., Micklin, M. (Eds.), Handbook of Population. Springer, New York, pp. 601–624.
Ryder, N.B., 1965. The cohort as a concept in the study of social change. American Sociological Review 30, 843–861.
Schnore, L.F., 1958. Social morphology and human ecology. American Journal of Sociology 63, 620–634.
Sharp, L., 1952. Steel axes for Stone Age Australians. Human Organization 2, 17–22.
Simon, H.A., 1952. A formal theory of interaction in social groups. American Sociological Review 17, 202–211.
Singer, H., 2012. SEM modeling with singular moment matrices. Part II: ML-estimation of sampled stochastic differential equations. Journal of Mathematical Sociology 36, 22–43.
Smith, A., [1776] 1937. The Wealth of Nations. The Modern Library, New York.
Wasserman, L., 2004. All of Statistics: A Concise Course in Statistical Inference. Springer, New York.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York.
Woolridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.
Yang, Y., Land, K.C., 2013. Age–Period–Cohort Analysis: New Models, Methods, and Empirical Applications. Chapman & Hall, New York.

Social Classification

Scott L Washington, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Among social scientists, an emerging field of theory and research concerns the nature and workings of social classification. This topic has been highlighted by Starr (1992: pp. 267–274), among others, and it is now becoming a major point of interest across a variety of disciplines, including psychology, sociology, political science, history, and anthropology. The present article offers a formal definition of the topic. After emphasizing the compatibility of different perspectives, it identifies a number of themes common to the study of ethnicity, race, nationalism, caste, age, class, gender, and sexuality.

Among social scientists, an emerging field of theory and research concerns the nature and workings of social classification. This topic has been highlighted by Starr (1992: pp. 267–274), among others, and it is now becoming a major point of interest across a variety of disciplines, including psychology, sociology, political science, history, and anthropology. The present article offers a formal definition of the topic. After emphasizing the compatibility of different perspectives, it identifies a number of themes common to the study of ethnicity, race, nationalism, caste, age, class, gender, and sexuality.

A Definition

Fundamentally, *classification* refers to "the ordering or arrangement of objects" (Sokal, 1974: p. 1116) under a particular description. By nonexperts, as Sokal (1974: p. 1116) has noted, the term 'classification' is often used as if it were synonymous with the term 'identification'. But for scientific or analytical purposes it is helpful to distinguish between the two. Classification involves the creation of logical sets, or categories. Identification by contrast involves the assignment of objects to logical sets, or categories, "once such classes have been established by prior classification" (Sokal, 1974: p. 1116). The first is technically a precondition for the second. Indeed, it is only on the basis of a system of classification that we can 'identify' certain objects at all: a piece of furniture as being a 'chair', for example, "or a plant as being a 'buttercup'" (Sokal, 1974: p. 1116).

This point is all the more important to emphasize when it comes to systems of *social classification*. What is social classification? And what, if anything, is specifically social about it? In any system of social classification, properly speaking, every object is also a subject. The things classified are capable of classifying things. Using standard notation (where x represents a given set of objects; y, a given set of subjects; $\subset$, a term for inclusion; and $\wedge$, a logical operator meaning 'and'), we can define the topic formally:

$$x \subset y \wedge y \subset x$$

This once again is the defining characteristic of any system of social classification: that all of its objects are at the same time subjects, i.e., objects for whom objects exist.

It is essential to keep the definition being presented clear. Classification – or categorization – is by all accounts a basic process. It allows our minds to synthesize and retrieve information parsimoniously. But when a system of classification involves the ordering or arrangement of people (as opposed to plants, for example), something quite special occurs. An object may now object. The authors are now a part of the story being written, and the categories being used may no longer "aim to achieve [only an] economy of memory" (Sokal, 1974: p. 1116). Ordinarily indispensable to reasoning and communication, categories of thought, within systems of social classification, are simultaneously categories of being: 'human kinds', as Hacking (1995), addressing a similar point, has put it.

Here, it bears repeating, we are dealing strictly with systems of social classification. Other systems of classification (e.g., botanical, astronomical, or zoological systems) are technically *asocial*. The term 'natural' has also been used, but for at least two reasons it can be confusing. First, among philosophers, the term has come to be associated with debates over the existence of 'natural kinds'. These debates are concerned with the question of whether a particular system is artificial, or whether its categories are purely conventional. With respect to such debates, the present article remains agnostic. It is concerned only with the question of whether a system's objects are subjects. Second, many forms of social classification are presumed to be 'natural'. This is a fact, and for our purposes it cannot be ignored. Indeed, it is partly the task of a social scientist to explain how and why such presumptions come to exist. However awkward, the term 'asocial' remains logically consistent ($\sim p \neq p$), and it allows us to avoid many of the traps and controversies that have traditionally plagued a term like 'natural'.

It is also worth noting that we are not dealing with systems in which only some objects are subjects. Such is quite obviously the case with zoological systems. Indeed, although it might include *Homo sapiens*, every object in a zoological system (e.g., a starfish) is not also a subject, i.e., a being capable of using language in a self-conscious way. As a phenomenon (and not simply a phrase), 'social classification' refers once again to something very specific: the classification of people by people.

Psychological, Biological, and Sociological Perspectives

Within the social sciences, for the most part, the study of social classification has been defined by three perspectives: psychological, biological, and sociological. These perspectives belong to a division of scientific labor. They are in theory compatible.

Psychological Perspectives

For many psychologists, social classification has been a long-standing concern. One line of research can be traced back for example to Tajfel, who, in a now classic series of experiments, set out to document the impact of social classification under 'minimal' conditions. The results were quite revealing. The simple act of being classified, Tajfel et al. (1971) discovered, predisposed subjects to accentuate between-group differences, overestimate levels of out-group homogeneity, and develop forms of in-group favoritism, with respect to the allocation of resources as well as the evaluation of a person's conduct. These behaviors were not difficult to elicit. They emerged when subjects were classified temporarily and randomly, and even when the distinctions being introduced by the experimenter were trivial and previously meaningless (Tajfel et al., 1971). Though first published in the 1970s, the work of Tajfel remains influential, and it has since become the basis of an entire theory of intergroup behavior (Tajfel and Turner, 1986).

A more recent line of research – associated most prominently with the work of Fiske, among others (e.g., Macrae and Bodenhausen, 2000; Fiske and Macrae, 2012; Fiske and Taylor, 2013) – has focused on 'social cognition'. This body of scholarship has contributed greatly to our understanding of stereotypes, and through it psychologists have developed a better understanding of the cognitive dynamics behind the perception of groups and people (Hamilton and Sherman, 1996). Advances in neuroimaging have even allowed psychologists to link processes of impression formation to a specific area in the brain: mPFC, a dorsal region of the medial prefrontal cortex (Van Overwalle, 2009). For supplementary tasks (e.g., inferring goals and intentions), other parts of the brain also appear to be relevant, including the amygdala, the temporo–parietal junction, the posterior cingulate cortex, the inferior frontal gyrus, and the superior temporal sulcus, but much of the literature on processing strategies is still being generated (Fiske and Taylor, 2013). A subfield devoted entirely to the neural bases and correlates of social cognition (variously called 'social neuroscience', 'social neuropsychology', 'social-cognitive neuroscience', or 'social-cognitive neuropsychology') is now developing.

Biological Perspectives

As a topic, social classification has also been studied from a biological standpoint. Here evolutionary perspectives have been particularly insightful. As Kurzban et al. (2001) have argued, with respect for example to race, our disposition to classify each other is not unrelated to our development as a species. It can be understood partly as the byproduct of a cognitive machinery that evolved to detect coalitions and alliances. Though evolution might have selected for a basic capacity to encode a trait like gender or age (characteristics pertinent to our ability to reproduce), the same cannot be said of course for all forms of social classification, many of which were unknown to our ancestors. What is appealing about the Kurzban et al. (2001) model is that it can accommodate a phenomenon like race, a form of social classification which appears to have only developed in the modern era. It can also accommodate variation in the degree to which social distinctions become salient. As a hypothesis, the Kurzban et al. (2001) model suggests, a phenomenon like race is all the more likely to become salient when it meaningfully cues patterns of alliance.

Like a number of 'dual-process' theories (Fiske and Taylor, 2013: pp. 31–58), moreover, biological approaches bear a crucial implication: that a satisfactory account of our disposition to classify each other requires a study of the body as well as the mind. Not only after all do systems of social classification operate upon the body. They operate through it. This is why, under most conditions, the activation by a stimulus of a particular category seems 'automatic' (Macrae and Bodenhausen, 2000: pp. 96–102). In some contexts, it is true, a category might be used in a manner that is technical, deliberate, and controlled. But as a general description of the phenomenon, such an account can be misleading. With the exception of cases that are ambiguous or in some way unintelligible, for example, we simply *see* a 'woman'. The process by which we have identified the person as a 'woman' (a very complex process, involving the activation as well as the application of a particular category) is for the most part unconscious. Phenomenologically, it can only be said to exist after the fact.

The word 'see' might also be misleading. Traditionally sight has been treated as the 'noblest' of the senses (as Descartes once referred to it), but when it comes to the perception of objects – and especially the perception of subjects, which is to say, the perception of groups and people (Hamilton and Sherman, 1996) – not all cues are visual. This is not to deny the importance of visual cues. As diurnal creatures, hominids standing ultimately on their hind legs, the earliest humans developed their sensorium in such a way as to privilege cues that were visual. And, despite being the last of our senses to have developed fully, sight is broadly superior in its capacity to digest certain kinds of stimuli, allowing us to assimilate information from without at a far greater rate and distance than any other sense (Smith, 1986). As many of our own intuitions might suggest, however, the eye does not do its work alone, and the perception of groups and people is apt to involve all the senses of a socialized body. A person might not only *look* but *smell*, *feel*, or *sound* like a 'woman' as well.

Sociological Perspectives

Like biological perspectives, sociological perspectives offer a necessary and useful complement to psychological perspectives. For many researchers, at its heart, social classification is a cognitive act, a task performed quintessentially by individuals. But in a very real sense, sociologists have demonstrated, institutions as well as individuals can be said to classify. In the modern world, as Starr (1992), for example, has noted, states are inveterate classifiers. They make and enforce distinctions at a variety of levels, relying characteristically upon a variety of agents. "The design of classificatory systems is typically the prerogative of legislatures, … bureaucratic officials, and governing bodies …, whereas the [task of assigning people to a particular category] … belongs [properly] to … intermediary agents, such as … judges [or] 'street-level' bureaucrats, … who [are expected to] decide cases [or] certify [a person's] eligibility for benefits …" (Starr, 1992: p. 270).

It might seem as if the self-descriptions of people are irrelevant to the classificatory operations of a modern state. But, as Starr (1992: p. 270) is careful to note, statistical and

administrative systems vary with respect to how self-descriptions are handled. "At one extreme, some accept self-descriptions at face value; at the other, some classify and assign cases without attending [to such matters] at all…. Between the[se] two [extremes of course there] are countless variations…." To identify people racially, for example, the United States Census Bureau relied upon enumerators until 1970 (Petersen, 1987). Now self-reporting is the norm. Nor is the design of classificatory systems solely the prerogative of bureaucratic institutions. Financial institutions also classify people (Fourcade and Healy, 2013), along with clinical discourses and institutions of social control (Donnelly, 2000).

Institutions aside, sociological perspectives have also been instrumental in helping us grasp – what one might call, drawing once again upon Starr (1992) – the history and politics of social classification. How do categories come to be defined? Which distinctions are perceived to be legitimate? How and why do systems emerge, persist, change, or disappear? By exploring such questions, sociologists (as well as historians, anthropologists, political scientists, and others) have been able to show concretely that systems of social classification are by no means static, that people fight over who belongs where. This is an important point. To account for practices of social classification, sociological perspectives have taught us, it is not enough to understand how certain actors within a particular system perceive or make sense of each other. It is also necessary to understand how "social actors seek to shape the classificatory system and the pay-offs and risks of being categorized in specific ways" (Starr, 1992: p. 274). These pay-offs and risks are not only material. They can also be symbolic (involving for example the acquisition of a stigma) and emotional as well.

Paradigmatic Forms

While social classification is not always invoked explicitly as a topic, many scholars have already embarked upon a study of its forms. Mauss (1979[1924]: p. 32) famously called for a "catalogue of categories". A similar call might be made for a catalog of social categories, a survey of "all those which it is possible to know man has used". A provisional list would at the very least include ethnicity, race, and nationalism (Brubaker, 2009), of course, but also caste (which deserves to be considered alongside ethnicity, race, and nationalism), age, class, gender, and sexuality.

Age

In modern societies, a person's age is typically understood in chronological terms, and once codified it is likely to form the basis of an explicit policy. Legally, for example, a person might not become an 'adult' until the age of 18. In traditional or nonliterate societies, by contrast, a person's age is typically understood in maturational terms, and the transition from youth to adulthood (though marked at various points by rites of passage) tends not to be codified at all (Kertzer and Keith, 1984). While admittedly stark, such a comparison does have the merit of reminding us that, as Linton (1942) observed long ago, age can be studied as a form of social classification. This does not imply of course that age is just a state of mind. Like all organisms (and primates, in particular), human beings are mortal. They are born at a certain point and will die. But the life course of a human being can be divided into any number of stages, and – whether or not, for example, the notion of being 'middle-aged' is entirely unique to the contemporary period – the systems under which people are classified by an age can be said to vary historically and cross-culturally.

Class

The idea of studying class as a form of social classification social scientists have inherited partly from Bourdieu (1984[1979]: pp. 466–484), who in 1979 insisted that class struggles (i.e., struggles between classes) were ultimately inseparable from 'classification struggles' (i.e., struggles over the definition and meaning of classes). Whether distinctions like 'the upper class' or 'the middle class' are based upon a person's wealth, education, occupation, or lifestyle, the implications of such a view are clear. What a class 'is' (a question debated *ad nauseam* by scholars and laypeople alike) depends in each case on how it comes to be defined. Like any form of social classification, a class may be defined *analytically*, as a scientific concept, or *practically*, as a folk concept. A class may also be defined *officially*, as a concept which has been incorporated into the statistical and administrative practices of a modern state (Starr, 1992). The state's role in the making and unmaking of classes social scientists have explored with respect to a variety of cases, including, for example, *cadres* in France (Boltanski, 1987), *Angestellten* in Germany (Kocka, 1981), and professionals in countries like Britain or the United States (Szreter, 1993).

Ethnicity, Race, Nationalism, and Caste

A growing number of social scientists have now embraced the notion that ethnicity, race, and nationalism can be profitably studied alongside each other. Recently, in fact, an effort has been made to integrate each phenomenon into a single domain, a subfield defined loosely by a concern with cultural processes, social organization, and politics (Brubaker, 2009). This domain might be extended for our purposes to include caste, a form of social classification found principally in India (Berreman, 1972). Studying ethnicity, race, nationalism, and caste alongside each other does not require us to distinguish rigidly between each phenomenon. Indeed, as several academics have come to realize, attempting to establish "a sharp, clean distinction between [the] phenomena we ordinarily associate with race and ethnicity" can be quite unproductive (Brubaker, 2009: pp. 25–28), and it may reflect a concern that is more or less peculiar to the United States. Sometimes conceptions of race overlap with conceptions of ethnicity. Sometimes not. What is important scientifically is that each phenomenon can be studied, alongside nationalism and caste, as a form of social classification.

Gender and Sexuality

As many scholars have noted, gender is virtually 'ubiquitous' as a form of social classification (Maccoby, 1988: p. 755). Typically, conceptions of gender are dimorphic, with all cultures and known languages distinguishing at some level between males and females. But intermediate cases do exist. Socially, for example, a transgender person might be said to fall between

traditional conceptions of masculinity and femininity. Physiologically, a hermaphrodite or a pseudo-hermaphrodite might be said to do the same. Indeed, although rare, a person who is chromosomally male (XY) can have ambiguously female genitalia (androgen-insensitivity syndrome), and a person who is chromosomally female (XX) can have ambiguously male genitalia (congenital adrenal hyperplasia). Whether a female with Turner syndrome (XO) or a male with Klinefelter syndrome (XXY) might also qualify as an intermediate case remains a matter of dispute (Sax, 2002).

Like gender, sexuality can also be understood as a form of social classification, as a means of sorting people with respect to certain desires, behaviors, or orientations. Homosexuality from social scientists has perhaps received the most attention (Plummer, 1981). But homosexuality – as a label as well as a category of experience – has not always existed, and it is by no means the only manner in which a person might be classified sexually. In addition to 'homosexuality', for example, one may speak of 'heterosexuality', 'bisexuality', or 'asexuality'. With respect to sexual interests, more broadly, one may also speak of *zoophilia* (animals), *pedophilia* (children), *hebephilia* (pubescent males or females), *ephebephilia* (adolescent males or females), *androphilia* (adult males), or *gynephilia* (adult females). These are merely possibilities, any one of which (e.g., *pedophilia*) is liable to become entrenched as a category (e.g., 'pedophile').

Common Themes

Despite working in different fields, social scientists have addressed a number of themes common to the study of age, class, ethnicity, race, nationalism, caste, gender, and sexuality. Three, in particular, might be identified.

Membership

How is a person's membership in a particular category determined? Here like others (e.g., Brubaker, 2009: p. 26) we must distinguish between the *indicia* associated with a particular category and the *criteria* defining it. This distinction, it will be recalled, is perfectly consistent with the distinction between 'classification' and 'identification' (Sokal, 1974: p. 1116). Criteria are what allow us to classify a given object. Indicia are what allow us to identify it. For example, when it comes to determining a person's gender, properties like facial hair, clothing, the deepness of a person's voice, and breasts are typically indicia. Genitalia (e.g., a penis) and chromosomal pairings (e.g., XY) are typically criteria. When it comes to determining a person's age, similarly, properties like height and physique – including, among other things, muscle tone, hair, or the tightness of a person's skin – are typically indicia. Chronological markers (e.g., years or months since birth) are typically criteria.

Rarely are the criteria defining a person's membership in a particular category fixed once and for all. They are frequently a stake of struggles, and, as sociological perspectives have reminded us, the problem of establishing which criteria are legitimate is fundamentally political and historical (Starr, 1992). Nor of course is the existence of such criteria incompatible with the use of prototypes (Cantor and Mischel, 1979). A person might be French, as an example, but how 'French' he or she is perceived to be depends on how closely he or she approximates the prototype (and in some cases the stereotype) of a French person. Membership is often a matter of degree. Finally, it is by no means impossible for the same property to be used as a criterion in one context, while being used as an indicium in another. Conceptions of race in the United States and Brazil provide a striking example (Wagley, 1965). In the United States, where a person's ancestry is the reigning standard for determining who is black, skin color has been used as one of many indicia. In Brazil, where a different standard prevails, skin color has been used as one of many criteria.

The distinction between indicia and criteria is strictly analytical, and it can be applied once again to any form of social classification: age, class, ethnicity, race, caste, nationalism, gender, or sexuality. It can also help us understand how a person's membership in a particular category is 'managed' in a natural, everyday setting (a puzzle which has long preoccupied students of face-to-face behavior, particularly sociologists). An older person might attempt to look younger, for example, by manipulating in each case a particular set of indicia: wearing make-up, avoiding certain clothes, resorting to cosmetic treatments, or undergoing forms of plastic surgery. A comparable strategy might be adopted by a transgender person wishing to appear more or less masculine. This is partly what it means for a trait like gender or age to be, as some researchers have put it, 'accomplished' (West and Zimmerman, 1987; Laz, 1998).

Categories and Groups

What is the relationship between categories and groups? These terms are often used interchangeably, but it is important for social scientists to remember that, as Starr (1992: p. 281), Brubaker (2009: p. 28–34), and others have pointed out (e.g., Merton, 1968[1949]: pp. 353–354), a category is not a group. Technically, a *category* is a logical set, a class of objects that is only presumed to belong together. A *group* on the other hand is a self-conscious category, a class of subjects bearing a common point of identification. Under certain conditions, of course, a category of people might become a group, but we cannot assume beforehand that the process has already taken place.

No longer taken for granted, in fact, the 'groupness' associated with a particular category has been treated by many scholars as a variable (Merton, 1968[1949]: pp. 364–380; Starr, 1992: p. 281; Brubaker, 2009: pp. 28–34). Two approaches have typically prevailed. From one angle, researchers have attempted to understand how a group is objectively perceived. This is a problem of social cognition, pertaining chiefly to what psychologists call 'entitativity' (i.e., the extent to which a group is perceived by actors as an entity). From another angle, researchers have attempted to understand how a group is subjectively experienced. This is a problem of social cohesion, pertaining chiefly to feelings of solidarity (i.e., the extent to which a group is experienced by actors as an entity). Among Marxists, such concerns have traditionally assumed the form of debates regarding 'class consciousness'. But the problem of how a 'class-in-itself' might become a 'class-for-itself' is by no means unique to Marx or the making of a 'proletariat' alone. It is a theme common to all forms of social classification, a problem which can be said to exist whenever the objects grouped according to the logic of a particular system are subjects.

Stratification

How might social classification be related to stratification? At an interpersonal level, psychologists have shown (e.g., Locksley et al., 1980), social classification is a necessary but insufficient condition for discrimination, the differential treatment of people belonging individually to a category. At an institutional level, analogously, social classification can be understood as a necessary but insufficient condition for stratification, the ranking or arrangement of people into strata.

This point has been emphasized by Massey (2007: p. 8), who notes that "stratification begins psychologically with the creation of cognitive boundaries that allocate people to social categories", a capacity which appears to be hardwired into the human brain. Massey (2007) focuses in his analysis on race, gender, and class – each, a form of social classification which has functioned in the United States as a principle of stratification. Age too has functioned as a principle of stratification (Riley et al., 1972). But the importance of stratification to other forms of social classification remains an open question. Here it would be prudent to avoid the claim that, on its own, social classification is inherently or uniquely hierarchical. It is not. Though often linked to connotations of superiority or inferiority, social classification is nothing more (nor less) than a potential basis for stratification.

Conclusion

Jespersen (1964[1922]: p. 388) once referred to man as a 'classifying animal'. In any system of social classification, it might now be said, man also exists as a classified animal, an object who is equally a subject. For anyone in the social sciences, the question of what happens when a system's objects are subjects remains fundamental. This question is no less relevant to the study of age than it is to the study of class, ethnicity, race, nationalism, caste, gender, or sexuality. Like Molière's *gentilhomme*, in fact, we may soon realize that we have been speaking prose all our lives, addressing in different forms one and the same phenomenon: social classification.

See also: Age, Sociology of; Class: Social; Discrimination: Racial; Ethnic, Racial, and Nationalist Social Movements; Ethnicity, Sociology of; Heterosexual Sexuality: Partner Bond Formation, Sexual Attitudes, Sexual Practices, and Sexual Function over the Life Course; Identity, Social; Inequality, Social; Nationalism, Sociology of; Prejudice in Society: Psychological Perspectives; Prejudice in Society: Sociological Perspectives; Race Identity; Social Stratification; Theories of Sexuality.

Bibliography

Berreman, G.D., 1972. Social categories and social interaction in urban India. American Anthropologist 74, 567–586.
Boltanski, L., 1987. The Making of a Class: Cadres in French Society. Cambridge University Press, Cambridge.
Bourdieu, P., 1984[1979]. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge.
Brubaker, R., 2009. Ethnicity, race, and nationalism. Annual Review of Sociology 35, 21–42.
Cantor, N., Mischel, W., 1979. Prototypes in person perception. Advances in Experimental Social Psychology 12, 3–52.
Donnelly, M., 2000. What is the object of social control? In: Della Porta, D., Greco, M., Szakolczai, A. (Eds.), Identità, riconoscimento e scambio: Saggi in onore di Alessandro Pizzorno. Laterza, Bari.
Fiske, S.T., Macrae, C.N. (Eds.), 2012. The Sage Handbook of Social Cognition. Sage, Thousand Oaks.
Fiske, S.T., Taylor, S.E., 2013. Social Cognition. Sage, Thousand Oaks.
Fourcade, M., Healy, K., 2013. Classification situations: life-chances in the neoliberal era. Accounting, Organizations and Society 38, 559–572.
Hacking, I., 1995. The looping effects of human kinds. In: Sperber, D., Premack, D., Premack, A.J. (Eds.), Causal Cognition: A Multidisciplinary Debate. Oxford University Press, Oxford.
Hamilton, D.L., Sherman, S.J., 1996. Perceiving persons and groups. Psychological Review 103, 336–355.
Jespersen, O., 1964[1922]. Language: Its Nature, Development, and Origin. Norton, New York.
Kertzer, D.I., Keith, J. (Eds.), 1984. Age and Anthropological Theory. Cornell University Press, Ithaca.
Kocka, J., 1981. Class formation, interest articulation, and public policy: the origins of the German white-collar class in the late nineteenth and early twentieth century. In: Berger, S. (Ed.), Organizing Interests in Western Europe: Pluralism, Corporatism, and the Transformation of Politics. Cambridge University Press, New York.
Kurzban, R., Tooby, J., Cosmides, L., 2001. Can race be erased? Coalitional computation and social categorization. Proceedings of the National Academy of Sciences of the United States of America 98, 15387–15392.
Laz, C., 1998. Act your age. Sociological Forum 13, 85–113.
Linton, R., 1942. Age and sex categories. American Sociological Review 7, 589–603.
Locksley, A., Ortiz, V., Hepburn, C., 1980. Social categorization and discriminatory behavior: extinguishing the minimal intergroup discrimination effect. Journal of Personality and Social Psychology 39, 773–783.
Maccoby, E.E., 1988. Gender as a social category. Developmental Psychology 24, 755–765.
Macrae, C.N., Bodenhausen, G.V., 2000. Social cognition: thinking categorically about others. Annual Review of Psychology 51, 93–120.
Massey, D.S., 2007. Categorically Unequal: The American Stratification System. Russell Sage Foundation, New York.
Mauss, M., 1979[1924]. Real and practical relations between psychology and sociology. In: Brewster, B. (Tr.), Sociology and Psychology: Essays. Routledge and Kegan Paul, London.
Merton, R.K., 1968[1949]. Social Theory and Social Structure. The Free Press, New York.
Petersen, W., 1987. Politics and the measurement of ethnicity. In: Alonso, W., Starr, P. (Eds.), The Politics of Numbers. Russell Sage Foundation, New York, pp. 187–234.
Plummer, K., 1981. Homosexual categories: some research problems in the labeling perspective of homosexuality. In: Plummer, K. (Ed.), The Making of the Modern Homosexual. Barnes and Noble, Totowa.
Riley, M.W., Johnson, M., Foner, A. (Eds.), 1972. Aging and Society, vol. 3: A Sociology of Age Stratification. Russell Sage Foundation, New york.
Sax, L., 2002. How common is intersex? A response to Anne Fausto-Sterling. Journal of Sex Research 39, 174–178.
Smith, A., 1986. The Body. Viking, New York.
Sokal, R.R., 1974. Classification: purposes, principles, progress, prospects. Science 185, 1115–1123.
Starr, P., 1992. Social categories and claims in the liberal state. Social Research 59, 263–295.
Szreter, S.R.S., 1993. The official representation of social classes in Britain, the United States, and France: the professional model and 'Les Cadres'. Comparative Studies in Society and History 35, 285–317.
Tajfel, H., Billig, M.G., Bundy, R.P., Flament, C., 1971. Social categorization and intergroup behavior. European Journal of Social Psychology 1, 149–178.
Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Nelson-Hall, Chicago.
Van Overwalle, F., 2009. Social cognition and the brain: a meta-analysis. Human Brain Mapping 30, 829–858.
Wagley, C., 1965. On the concept of social race in the Americas. In: Heath, D.B., Adams, R.N. (Eds.), Contemporary Cultures and Societies in Latin America. Random House, New York.
West, C., Zimmerman, D.H., 1987. Doing gender. Gender and Society 1, 125–151.

Social Cognition

François Ric, Université de Bordeaux, Bordeaux, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents a brief overview of the social cognition approach that has developed in social psychology during the last few decades. We first provide a definition of social cognition, a description of its specific features, and a discussion of why social cognition should not be considered a mere extension of cognitive psychology. Then the article presents a short history of social cognition through the evolution of the representations of the social thinker. Finally, we describe the distinction between two modes of information processing that are of central importance in social cognition research.

Definition and Features of Social Cognition Approach

Social cognition can be defined as a domain of research studying human thinking and of its relationships with social behavior. More specifically, social cognition refers to the study of the processes by which people make sense of themselves and of others as well as of their social environment, and the implications of this thinking for social behavior (for detailed presentations, see for instance, Bless et al., 2004; Fiske and Taylor, 2007). Thinking should be interpreted here in a broad meaning: the cognitive processes (conscious and unconscious) that underlie the understanding of the social environment. Social cognition is thus interested in what information do people attend in specific situations, how they process, encode, and integrate this information with the stored knowledge, and in turn how these processes influence people's social perception, judgment, and behavior.

The approach of social cognition is characterized by several features. First, social cognition is interested more or less directly in social behavior. Even though it was primarily closely associated with social perception, social cognition would better describe an approach of social psychological phenomena than the study of a specific object. Most aspects of the social behavior can be, and are currently, studied within the perspective of social cognition. Social cognition investigates not only intraindividual phenomena (e.g., development of preferences, attitude; e.g., Gawronski and Bodenhausen, 2006), interactions (dyadic interaction as well as small groups, e.g., Nye and Brower, 1996), intergroup relations (Yzerbyt and Demoulin, 2010), but also the role of culture dimensions in cognitive and social functioning (Kitayama and Uskul, 2011). Moreover, a number of studies in social cognition deal with behaviors that would not have been intuitively categorized as social. For instance, research explores the extent to which performance on various tasks (at school, in institution, or in the lab) may be partially determined by the social context (e.g., Autin and Croizet, 2012; Steele and Aronson, 1995).

Second, social cognition is interested in the cognitive processes that underlie social behavior. Thus, social cognition research borrows models (e.g., associative network models of memory), concepts (e.g., attention, information processing) from cognitive psychology, and methods that have proved to be useful to measure cognitive functioning (e.g., reaction time, memory tasks, neuroimaging, psychophysiological measures; see Klauer et al., 2011). However, as we will see later, this does not mean that social cognition is restricted to the application of cognitive methods and models to the study of social behavior. Social inclusion (via the culture or the position is the social structure) is also considered as constraining cognitive processes (e.g., Kitayama and Uskul, 2011).

Finally, social cognition is interested in people's production of representations of their social environment. These lay theories or beliefs may concern any part of the social environment, the self being included. Their importance lies in the fact that they are supposed to guide judgments and behaviors. For instance, stereotypes, defined as lay beliefs about the psychological and physical attributes shared by the members of a social group, may have a strong impact on the way people judge, and act toward, members of the stereotyped group. Such beliefs can also influence the behavior of the very members of the group. As shown by Steele and Aronson (1995), Black students underperform White students on a verbal task only when their ethnic category (and thus the associated stereotype according to which Blacks perform less on verbal task than Whites) is made salient or relevant for the performance. These beliefs influence behavior beyond the stereotype domain. Recently, Job et al. (2010) demonstrated that theories people possess about the management of cognitive resources have a direct impact on their performance. People generally hold that mental control consumes resources that are limited. Thus, after an experience of mental control (e.g., an attempt to quit smoking or the reading of a complex research article), one should not be able to perform optimally until the resources have replenished. Job and colleagues found that it was indeed the case that participants who had previously exerted self-control in a first task underperformed in a subsequent task those who had not exerted self-control. However, the pattern was reversed for participants who were led to believe that, sometimes, mental efforts have energizing properties.

Social Cognition Is Not a Mere Extension of Cognitive Psychology

There are obvious overlaps between the processes that are studied in cognitive psychology and in social cognition. As already mentioned, both domains of research indeed relied on related concepts and methods. However, the processes studied

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24044-0

by social cognition research are hypothesized to be somewhat different because of the specificities of social objects (including persons).

In contrast to physical reality, which can be described relatively objectively, social reality (including others and oneself) cannot be observed directly and requires inferences and interpretations. One can evaluate the weight of a watermelon, the size of a car, or the material in which a ball is made with the use of an appropriate tool. Different persons in different contexts will certainly arrive at the same conclusion as long as the object and the measurement tools remain unchanged.

However, when we encounter a person, it is impossible to objectively know the person's feelings, his or her intentions and traits, whether one should trust him or her …. We can make up our impression on an individual extremely rapidly on the basis of some physical traits (e.g., Willis and Todorov, 2006), but this impression does probably not reflect reality. We have no possibility to objectively know whether this is the case. Moreover, not only are these features not directly observable, but we know that people are not always the same. People change, or show different features, depending on the situation, leading us to engage in revisions of our impressions. For instance, people do not show the same aspects of themselves when they are at work, when they interact with their mother, or with the person they love. Still more problematic, we know that we, as well as other people, intentionally attempt to influence the impressions we make. If you get an interview for a job, you will make effort to give 'a good impression' to the recruiter and thus show the positive aspects of your personality.

These examples may seem obvious when we consider the encounter of a new acquaintance. However, this also probably works for people we know (but generally meet in the same context). As a critical example, statistics indicate that a large part of the homicides (one-third in Canada in 2010; www.statcan.gc.ca; Hotton Mahony, 2011) are committed by a close relative, probably someone the victim could trust, and much more (70% according to 2010 Canadian statistics) by someone known by the victim.

We have deliberately focused our reasoning on the perception of others because people represent certainly the main 'objects' of our thoughts, judgments, and behaviors. It also applies to some extent to the perception of objects or ideas. They often possess symbolic values (motivational and emotional) that cannot be directly observed and are therefore inferred. For instance, if we consider digits and letters, they do not possess a priori strong social meaning. We can however observe that they do not possess the same value for an individual depending on his or her date of birth, occupation (digits should be more valued for a mathematician), name (e.g., Nuttin, 1985), or culture etc. Some of these properties emerge from the relations between these objects and the individual and have been found to constrain behavioral outputs (e.g., Jones et al., 2004).

The question addressed in this section was whether social cognition could be considered as a mere extension of cognitive psychology? The answer is thus clearly 'no.' The question could even be provocatively reversed: What is not social in cognition? Finally, it is difficult to think of any object that would not possess a social value. And, if it were the case, what would be the generalizability of processes related to such objects? Stated otherwise, in contrast to what is generally thought, it is possible to advance that processes highlighted by social cognition research reflect more the general rule than exceptions.

History of Social Cognition: Metaphors of the Social Thinker

Social cognition emerged from the development of social psychology in the late 1970s and was long considered as a branch of social psychology. Social psychology is assumed to have a long tradition of interest in cognition. Most of the pioneers in social psychology research (e.g., Solomon Asch, Fritz Heider, Kurt Lewin, or Muzafer Sherif) were strongly influenced by the Gestalt theory. As such, they espoused a holistic view of psychological phenomena. According to this view, an external stimulus has an impact on people's judgment and behavior through the subjective representation of the stimulus and not through its physical properties. What seems important in the determination of individuals' thought and behavior is the individual representation of the stimulus that is itself a product of the context in which this stimulus is perceived.

Nonetheless, it is the development of cognitive psychology – its models and methods – that gave the ultimate impetus to the emergence of social cognition. Social psychologists, who were convinced by this cognitive approach, developed a 'social cognitive approach.' The history of this approach can be traced in the early 1980s by three milestones. One of the most prestigious journals of social psychology, the *Journal of Personality and Social Psychology*, included an 'Attitudes and Social Cognition' section devoted to this approach in 1980. Two years later was published the first issue of another psychology journal, Social Cognition, totally dedicated to this domain. Finally, in 1984, Susan Fiske and Shelley Taylor published the first edition of *Social Cognition*, which was the first book of reference for researchers interested in the domain (see also Fiske and Taylor, 1991, 2007, for subsequent editions).

As we will see, the conception of the social thinker and thus social cognition has greatly evolved during the last few decades. This evolution can be illustrated by several metaphors (see Fiske and Taylor, 1991, 2007). The first metaphor presented the social thinker as a 'consistency seeker' (Fiske and Taylor, 1991). According to this view, the social thinker was an individual mainly motivated to maintain coherence in his cognitive system and in his social environment. Cognitive dissonance theory (Festinger, 1957) or balance theory (Heider, 1946) are perfect examples of this conception. Both theories propose that discrepancies between the cognitions in an individual system of thoughts are aversive. As a result, individuals are motivated to resolve these discrepancies and to restore consistency in the cognitive system. During the 1960s and 1970s, emerged the metaphor of the 'naïve scientist' (Fiske and Taylor, 1991) depicting an individual motivated to get an accurate perception of the environment. Research on attribution processes in which researchers attempt to explain how people attribute a cause to an event (Jones and Davis, 1965; Kelley, 1972) is particularly illustrative of this metaphor. Studies have however revealed that on many occasions individuals do not have the motivation or the processing resources necessary to engage in an extensive

processing of all the information available and that they rely on cognitive shortcuts in order to get an acceptable answer at a minimum cognitive cost (Tversky and Kahneman, 1974). From these observations, another metaphor was developed: the social thinker as a 'cognitive miser' (Fiske and Taylor, 1984). In this metaphor, the social agent is essentially motivated to preserve cognitive resources. The shortcomings in the perception of the social reality were thus essentially seen as consequences of a reduced investment in information processing. The work of Hamilton and Gifford (1976) on stereotype formation provides a good example of such a conception as well as an illustration of how in his early years social cognition applied cognitive theories and methods to social phenomena. Chapman (1967) had observed that people tend to overestimate the frequency of association between infrequent events. This effect was explained by the fact that infrequent events are salient and thus attract attention. Because they allocate more attention to these stimuli, people remember them more easily and thus overestimate their frequency when compared with less salient associations. On the basis of this work, Hamilton and Gifford (1976) provide evidence that these processes could offer an explanation for the association that White Americans made between Blacks and negative behaviors. As Blacks are a minority in the United States and that minority members are, by definition, more rare than majority members, they are distinctive and thus attract attention. The same being true for undesirable behaviors, which are rarer than desirable behaviors, people tend to overestimate frequency of undesirable behaviors performed by Afro-Americans.

However, the way that people process information, and ultimately engage in social behavior, is influenced by various motivations. Social interactions are indeed complex and, depending on their current goal, people are able to develop different strategies that will be served by different cognitive processes. Depending on the situation, people mostly want not to make a mistake (be accurate), to be accepted by others, or even to show their superiority over other individuals. Evidence supporting this line of reasoning contributed to the development of a new metaphor in which motivation and affect were given a central role: the 'motivated tactician' (Fiske and Taylor, 1991). This metaphor describes the social thinker is a social agent who possesses multiple goals and who puts in place different strategies in order to reach these goals.

In the last two decades, a growing number of studies has further shown that many of these processes could be automatically triggered by the context, with the individual being unaware of the processes that are implied in their social behavior and of its cause (Bargh, 1997). These studies revealed that concepts, stereotypes (e.g., Bargh et al., 1996; Devine, 1989), norms (Aarts and Dijksterhuis, 2003) and affect (Murphy and Zajonc, 1993) may influence thoughts, social judgment and behavior without participants being aware of this influence (e.g., Dijksterhuis and Aarts, 2010). These findings led to the development of the metaphor of the 'primed actor' (Fiske and Taylor, 2007). In this view, the situation provides the individual with many cues that can be unconsciously and automatically used to guide cognitive, affective, and behavioral outputs. Although relatively automatic, these processes need not to be considered as totally innate. There is now growing evidence that culture, conceptualized in a broad (e.g., Westerners vs Easterners) or in a narrower meaning (e.g., social status), shapes and regulates the way our cognitive system operates, leading to pay attention to different aspects of a situation, to develop different goals, representations, and to make different inferences (Kitayama and Uskul, 2011; Snibbe and Markus, 2005). Thus, social cognition research now tends to describe the social thinker as an individual unconsciously governed by cognitive processes that possess strong motivational and emotional components, and that are partially shaped by culture. A last emerging aspect in recent development of social cognition is the increasing importance of the body in the way we think about, and interact with, our social environment (e.g., Niedenthal et al., 2005).

Social cognition is probably currently one of the most prominent and still promising domains of research in social psychology. However, it is worth noting that interest in social cognition has spread over various domains of study of the human thought and behavior. Social cognition is currently widely studied in related domains, such as developmental psychology, psychopathology, and neurosciences (e.g., Cacioppo et al., 2002).

How Do People Think About Social Issues?

The Different Steps of Information Processing

Four main steps of information processing can be broadly distinguished. First, perceptual inputs are received. Then, information from these inputs is encoded and stored in memory. Finally, this stored information is potentially used to guide behavioral outputs. The processing of the social information does not depart from this general sequence. However, social cognition research is interested in how these different stages are influenced, or constrained, by the social context.

Attention is expected to play a crucial role. Because we cannot attend to all the information provided by the situation, our attention is directed toward information that should thus be treated more deeply, which have the potential for more consequences on subsequent judgment and behavior. Which information will receive more attention? A first answer is information that is particularly salient (e.g., infrequent events, abrupt onsets). However, attention can also be attracted by situational features that serve the current goal of the individual. For instance, participants who are made thirsty by eating salty sweets pay more attention than nonthirsty participants to objects related to drinking behavior (e.g., bottle of water, glass; Aarts et al., 2001).

Another important basic goal of human beings is the preservation of one's own physical safety. As a result, in intergroup context people tend to first allocate more attention to outgroup members, and then to the ingroup (Ito and Urland, 2005). Early attention toward outgroup members would serve a protection motive, that would then turn in attention to ingroup members in the absence of real threat. In other words, people pay more attention to information that is particularly relevant for them and their goals. Therefore, this is no surprise that information that is directly related to the self (e.g., one's own name) attracts attention and is even difficult to ignore (e.g., Alexopoulos et al., 2012).

Attention will generally favor encoding and integration of the information in the stored knowledge about objects, individuals or situations. However, information will be more easily encoded (and integrated for future use) to the extent that it is consistent with the knowledge already stored in memory. As a result, even though blatantly inconsistent information attracts more attention than information consistent with our prior knowledge (because of the revision of the action plan that it may imply for the individual), this information should be recalled to a lesser extent if people do not want or cannot make extra effort to integrate it. Most of the time, people cannot afford to integrate all the incoming information because of a limitation in processing resources and time constraints. Thus, in conditions of limited processing resources, information that is consistent with prior knowledge will be easier to encode (as it requires less attentional capacity), better integrated, and more easily retrieved (e.g., Macrae et al., 1993).

Ultimately, what is stored in memory and activated will guide the behavior of the individual. Of course, only a subset of the information stored in memory will play that role. The behavior will be guided by information that is 'accessible.' Accessibility refers to ease with which one can access to a specific piece of information (e.g., concept) in the cognitive system. This accessibility is a function of the frequency and of recency of the activation of the information. Although accessibility influence the likelihood to engage in a given behavior, this does not mean that the individual will necessary engage in that behavior. Several factors, such as normative pressure or other aspects of the situation, may reduce the likelihood that the activation of such behavioral tendencies translates into overt behavior (Macrae and Johnston, 1998; Strack and Deutsch, 2004). For instance, Devine (1989) showed that low- and high-prejudiced people share the same knowledge about the stereotype of Afro-Americans in the USA, that they use it when it is unobtrusively activated. However, she found that in contrast with high-prejudiced participants, low-prejudiced participants do not use their stereotypical knowledge when asked to explicitly describe their thoughts about Black Americans.

Dual-Process Models of Information Processing

The sequence of information processing that has just been described is general. However, and as the research by Devine illustrates, this sequence could lead to diverging results as a function of the mode of processing on which individuals rely. There is a general agreement in social cognition to consider that people rely on two different modes of thinking. Theoretical descriptions of these two modes of thinking are referred to as 'dual-process models' of information processing (see Evans, 2008). The first mode (called, among others, 'system 1,' 'heuristic,' 'impulsive,' or 'associative') is fast, relatively automatic, requires few processing resources, and allows to get a rapid answer to our environment, whereas the other mode of processing (called, among others, 'system 2,' 'systematic,' 'reflective' or 'propositional,' respectively) is slow, controlled, and consumes cognitive resources.

The first (fast) mode of thinking is considered to be the most commonly used by the social thinker (Bargh, 1997) because it satisfies people's constraints of minimizing their efforts and of providing them fast with an acceptable answer. The processes involved in this mode of thinking are various. However, they globally work the same way: people rely on a subset of the available information that will allow them to develop mental shortcuts to arrive at a judgment or at a decision. For instance, people can rely on their affective feelings in order to infer whether they like something or someone (Schwarz and Clore, 1983) or whether their current thoughts about this object can be relied on without engaging in a costly examination of the specific features of the object or person. In a related vein, people estimate the frequency of an event from the ease with which they can recall exemplars of this kind of event, relying on the 'availability heuristic' (Tversky and Kahneman, 1974). More general knowledge structures can be used to develop rapid judgment. These structures are learned through socialization and can become highly automatized, and sometimes activated automatically upon the perception of the object. As an example, stereotypes can be automatically activated in memory upon the perception of a person, and then guide the perception of the person and interpretation of his or her behavior (Devine, 1989). For instance, the identification of a person as a woman could automatically lead to expect that this person is dependent, not skilled in math, romantic, etc. This last example illustrates a negative aspect of this mode of thinking. However, the same processes can also explain why experts are sometimes able to provide a fast response to a problem.

These cues can be used even when people are not aware of such use. The absence of awareness can take two different forms. People can process the stimulus consciously (e.g., they consciously perceive that the person is a woman), yet do not notice that this categorization will have any impact on their judgment or behavior (e.g., they judge the person as having poor math skills, and not suitable for a job of engineer). This kind of processes is referred to as 'postconscious.' The number and the variety of cues (stimuli) that are (nonconsciously) used as a basis of judgment and behavior are very large. In addition to categories that are particularly salient or socially meaningful (e.g., color of the skin, gender, age) many other cues are used without people knowing that they use them. In the domain of person perception, people can rely on the form of the target's face (Zebrowitz and McDonald, 1991), the degree of pupillary dilatation (Niedenthal and Cantor, 1986), the number of previous encounters with the target (Zajonc, 1968), or the experienced temperature (Williams and Bargh, 2008). Particularly illustrative of this point are the findings showing the identification of the valence (positive vs negative) of an extreme emotion experienced by another person requires the use of the physical position of the body. Interestingly, whereas they efficiently use this information (and not facial expression) to make their predictions, people still believe that they base their judgments on facial expressions (Aviezer et al., 2012).

On some occasions, people can even not be conscious of the information that influences them. These 'preconscious processes' are typically studied with the use of subliminal priming (i.e., a procedure aimed at presenting stimuli unconsciously). Concepts, stereotypes but also goals can be activated preconsciously, with an influence on information interpretation, judgment, and social behavior (Bargh et al., 1996; Devine, 1989; Dijksterhuis and Aarts, 2010). For instance, the

nonconscious activation of the stereotype of Black people (associated with hostility for American participants) leads people to subsequently interpret the behavior of a target as being more aggressive (Devine, 1989) but also to behave more aggressively (Bargh et al., 1996).

In dual-process models, this mode of processing is contrasted with a more controlled one in which people consciously engage in various strategies in order to reach their goal. For instance, if individuals want to get an accurate perception of a situation, they may attempt to take into account the available information and integrate it in order to take the right decision. This is typically the mode of processing that we attempt to use when we have an important decision to take. This mode of thinking is associated with voluntary control, effort, and rationality, and is closer to the lay people's belief concerning the way they think about their social environment. As already mentioned, this belief is however not true. It is indeed difficult to think of a situation in which people would be totally free from pre- or postconscious processes. For instance, recent studies suggest that even the operation of highly controlled processes (executive control) can be partially influenced by preconscious processes (Marien et al., 2012).

When Do People Rely on One or the Other Mode of Thinking?

Social cognitivists consider the fast mode of thinking as the default option because it satisfies resources and time constraints. The relevance of the cues (and the related rules) that are later used in the fast mode is learned through socialization processes and practice. Many activities require voluntary control and consciousness to be learned, but are automatized through experience, to be then executed on the basis of the simple presence of significant cues. A classical example of these processes is when people learn to drive. During the first driving lessons, people should pay attention to many aspects that will be later processed automatically. Thus, one can pass from the slow system to the fast one through practice and experience.

In contrast, people engaged in the fast system can sometimes opt for the slow one. This seems to occur typically in two types of situation. First, there are situations in which the fast system is expected not to provide an acceptable answer regarding the situational requirements. For instance, the resolution of a mathematical problem or the understanding of a complex psychological theory requires that one deliberately control our attention and thoughts. Then, people also attempt to regain control on the way they process the situation when inappropriate decisions, judgments, or behaviors are perceived as having potential important consequences for the individual. For instance, if you envisage buying a new apartment, you should think that it would be wiser to take your time to take into account all the information you can bring, and to examine it carefully before you decide to invest all your savings. In a similar vein, if you want to make a good impression on someone (for a job, for the loved one), you would probably attempt to control your behavior in order to avoid any reaction that would lead to a negative impression. This is not to say that it would lead to more efficient use than relying on other (fast or unconscious) processes (Wegner, 1994; Dijksterhuis, 2004), even though we believe so.

There is great agreement on this distinction between two modes of thinking in social cognition research. There are however several aspects that are still debated. First, depending on the theoretical positions, the two modes of processing are hypothesized to function in parallel whereas for others these two models could not take place at the same time, and should thus be conceived as sequentially organized. A related question is the possible interactions between these two modes of processing, especially when they are conceived as working in parallel (e.g., Gawronski and Bodenhausen, 2006; Strack and Deutsch, 2004). Another question is related to the neuroanatomical structures that support these modes of thinking. Depending on the model, both modes rely on the same structures whereas they refer to different pathways for other theorists.

Conclusion

After three decades of study, social cognition has become a vast domain of research that brought many enlightening findings – all of which could not be brought out in the present article due to space limitations – about the way we construct social reality (see, for more detailed presentations, Bless et al., 2004; Fiske and Taylor, 2007). In the same time, these findings have strongly modified the picture of the social thinker. This picture is far now more complex than is was still few years ago, integrating intra- as well culture-related processes. These evolutions of social cognition underline its great explanatory power as well the potential for promising discoveries that researchers should undertake in this domain.

See also: Attitude Formation and Change; Cognitive Dissonance; Decision Making, Psychology of; Embodied Social Cognition; Group Processes, Social Psychology of; Heuristics in Social Cognition; Implicit Social Cognition; Intergroup Relations; Social Categorization; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Aarts, H., Dijksterhuis, A., 2003. The silence of the library: environment, situational norms, and social behavior. Journal of Personality and Social Psychology 84, 18–28.

Aarts, H., Dijksterhuis, A., de Vries, P., 2001. On the psychology of drinking: being thirsty and perceptually ready. British Journal of Psychology 92, 631–642.

Alexopoulos, T., Muller, D., Ric, F., Marendaz, C., 2012. I, me, mine: automatic attentional capture by self-related stimuli. European Journal of Social Psychology 42, 770–779.

Autin, F., Croizet, J.C., 2012. Improving working memory efficiency by reframing metacognitive interpretation of task difficulty. Journal of Experimental Psychology: General 141, 610–618.

Aviezer, H., Trope, Y., Todorov, A., 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science 338, 1225–1229.

Bargh, J.A., 1997. The automaticity of everyday life. In: Wyer, R.S. (Ed.), The Automaticity of Everyday Life: Advances in Social Cognition, vol. 10. Erlbaum, Mahwah, NJ, pp. 1–61.

Bargh, J.A., Chen, M., Burrows, L., 1996. Automaticity of social behavior: direct effects of trait construct and stereotype priming on action. Journal of Personality and Social Psychology 71, 230–244.

Bless, H., Fiedler, K., Strack, F., 2004. Social Cognition. How Individuals Construct Social Reality. Psychology Press, Hove, UK.

Cacioppo, J.T., Berntson, G.G., Adolphs, R., Carter, C.S., Davidson, R.J., McClintock, M.K., McEwen, B.S., Meaney, M.J., Schacter, D.L., Sternberg, E.M., Suomi, S.S., Taylor, S.E. (Eds.), 2002. Foundations in Social Neuroscience. MIT Press, Cambridge.

Chapman, L., 1967. Illusory correlation in observational report. Journal of Verbal Learning and Verbal Behavior 6, 151–155.

Devine, P.G., 1989. Stereotypes and prejudice: their automatic and controlled components. Journal of Personality and Social Psychology 56, 680–690.

Dijksterhuis, A., 2004. Think different: the merits of unconscious thought in preference development and decision making. Journal of Personality and Social Psychology 87, 586–598.

Dijksterhuis, A., Aarts, H., 2010. Goals, attention, and (un)consciousness. Annual Review of Psychology 61, 467–490.

Evans, J.S., 2008. Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology 59, 255–278.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CT.

Fiske, S.T., Taylor, S.E., 1984. Social Cognition. Addison-Wesley, Reading, MA.

Fiske, S.T., Taylor, S.E., 1991. Social Cognition, second ed. McGraw Hill, New York.

Fiske, S.T., Taylor, S.E., 2007. Social Cognition. From Brains to Culture. McGraw Hill, New York.

Gawronski, B., Bodenhausen, G.V., 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. Psychological Bulletin 132, 692–731.

Hamilton, D.L., Gifford, R.K., 1976. Illusory correlation in interpersonal perception: a cognitive basis of stereotypic judgments. Journal Experimental Social Psychology 12, 392–407.

Heider, F., 1946. Attitude and cognitive organization. Journal of Psychology 21, 107–112.

Hotton Mahony, T., 2011. L'homicide au Canada, 2010. Juristat N 85-002-X. http//www.statcan.gc.ca.

Ito, T.A., Urland, G.R., 2005. The influence of processing objectives on the perception of faces: an ERP study of race and gender perception. Cognitive, Affective, & Behavioral Neuroscience 5, 21–36.

Job, V., Dweck, C.S., Walton, G.M., 2010. Ego depletion – is it all in your head? Implicit theories about willpower affect self-regulation. Psychological Science 21, 1686–1693.

Jones, E.E., Davis, K.E., 1965. From acts to dispositions: the attribution proces in social psychology. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, New York, pp. 219–266.

Jones, J.T., Pelham, B.W., Carvallo, M., Mirenberg, M.C., 2004. How do I love thee? Let me count the Js: Implicit egotism and interpersonal attraction. Journal of Personality and Social Psychology 87, 665–683.

Kelley, H.H., 1972. Causal Schemata and the Attribution Process. General learning Press, New York.

Kitayama, S., Uskul, A.K., 2011. Culture, mind, and the brain: current evidence and future directions. Annual Review of Psychology 62, 419–449.

Klauer, C.K., Voss, A., Stahl, C., 2011. Cognitive Methods in Social Psychology. Guilford Press, New York.

Macrae, C.N., Hewstone, M., Griffiths, R.J., 1993. Processing load and memory for stereotyp-based information. European Journal of Social Psychology 23, 77–87.

Macrae, C.N., Johnston, L., 1998. Help, I need somebody: automatic action and inaction. Social Cognition 16, 400–417.

Marien, H., Custers, R., Hassin, R.R., Aarts, H., 2012. Unconscious goal activation and the hijacking of the executive function. Journal of Personality and Social Psychology 103, 399–415.

Murphy, S.T., Zajonc, R.B., 1993. Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. Journal of Personality and Social Psychology 64, 723–739.

Niedenthal, P.M., Barsalou, L.W., Winkielman, P., Krauth-Gruber, S., Ric, F., 2005. Embodiment in attitudes, social perception, and emotion. Personality and Social Psychology Review 9, 184–211.

Niedenthal, P.M., Cantor, N., 1986. Affective responses as guides to category-based inferences. Motivation and Emotion 10, 217–232.

Nisbett, R.E., Wilson, T., 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 84, 231–259.

Nuttin, J.M., 1985. Narcissism beyond gestalt and awareness: the name letter effect. European Journal of Social Psychology 15, 353–361.

Nye, J.L., Brower, A.M. (Eds.), 1996. What's Social about Social Cognition? Sage, Thousand Oaks, CA.

Schwarz, N., Clore, G.L., 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. Journal of Personality and Social Psychology 45, 513–523.

Snibbe, A.C., Markus, H.R., 2005. You can't always get what you want: educational attainment, agency and choice. Journal of Personality Social Psychology 88, 703–720.

Steele, C.M., Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology 69, 797–811.

Strack, F., Deutsch, R., 2004. Reflective and impulsive determinants of social behavior. Personality and Social Psychology Review 8, 220–247.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.

Wegner, D., 1994. Ironic processes of mental control. Psychological Review 101, 34–52.

Williams, L.E., Bargh, J.A., 2008. Experiencing physical warmth promotes interpersonal warmth. Science 322, 606–607.

Willis, J., Todorov, A., 2006. First impression: making up your mind after a 100-ms exposure to a face. Psychological Science 17, 592–598.

Yzerbyt, V., Demoulin, S., 2010. Intergroup relations. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, Hoboken, NJ, pp. 1024–1083.

Zajonc, R.B., 1968. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology 9, 1–27.

Zebrowitz, L.A., McDonald, S., 1991. The impact of litigants' babyfacedness and attractiveness on adjudications in small claims courts. Law and Human Behavior 15, 603–623.

Social Comparison, Psychology of

Ladd Wheeler, Macquarie University, Sydney, NSW, Australia
Jerry Suls, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social comparison theory deals with self-evaluation of abilities, opinions, and other personal attributes through comparing the self with others. We outline the original theory and the methods used to study it empirically. We discuss how individuals may compare upward or downward and may assimilate their self-evaluations toward or contrast away from the comparison person. We close with a discussion of implications for education, health, and happiness.

Social comparison occurs when an individual compares himself or herself to another person. The dimension on which the comparison is made can be almost anything – attractiveness, tennis skill, wealth, etc.

The comparison might be sought deliberately, might occur in a chance encounter, or might be forced upon the person by the social environment. It is not social comparison when one compares one person to another, as in comparing one neighbor to another; the individual making the comparison must be included in the comparison for it to fit the commonly accepted definition of a social comparison. It is not a social comparison when one compares oneself at Time 1 to oneself at Time 2; this is a temporal comparison (Albert, 1977).

The Original Theory

Leon Festinger's (1954) paper, *A Theory of Social Comparison Processes*, still serves as the basic formulation, although there have been many additions and revisions. Festinger argued that it is necessary for people to know what they are capable of doing and to hold correct opinions about the world. Sometimes people can perform a physical reality test on their abilities and opinions, but that could be dangerous, inconvenient, or impossible. For example, if individuals want to know whether they can swim across a swift river, they could simply try it, but that could be dangerous. If it is a person's opinion that the Russian Revolution would not have occurred without V.I. Lenin, there is simply no physical reality test for the truth of that opinion. Consequently, people often seek *social reality* rather than physical reality for their abilities and opinions, and that social reality is obtained through social comparison. An important difference between opinions and abilities in Festinger's theory is a *unidirectional drive upward* for abilities but not for opinions. Regardless of the ability, people would generally like to have more of it than other people have. With opinions, however, people feel confident when they have the same opinions as others. Thus, all the members of a group can be satisfied when they have the same opinion, but not when they have exactly the same level of ability. The central argument of the theory is that we need to compare with similar people in order to achieve a precise estimate of our abilities and the correctness of our opinions. Consequently, we will try to change others or ourselves in order to achieve similarity, and we will reject others if similarity with them cannot be achieved.

Although the original theory dealt with opinions and abilities, most of the research done after the 1950s has been on abilities or ability like personal attributes. The original theory focused on the accurate evaluation of abilities/opinions, but later research has focused less on the accuracy of evaluation and more on the self-esteem implications of evaluation. We do not just want to know how charming we are; we want to know that we are very charming indeed. Wanting to know certain things about ourselves influences both the comparisons we make and how we interpret or construe those comparisons.

Methods for Studying Social Comparison

The three major methods for studying social comparison are 'selection' approaches, which examine the characteristics of the people with whom individuals choose to compare; 'reaction' approaches, which examine individuals' emotional or cognitive reactions to comparison information; and 'narration' approaches, which examine individuals' self-reports of comparisons made (Wood, 1996). Each of these will be examined below.

The Selection Approach

Individuals are given the choice of two or more comparison targets that differ in theoretically important ways. The choices they make (and related data) provide clues concerning social comparison motives. The first notable example, known as the 'rank-order paradigm,' was introduced to study the unidirectional drive upward. Experimental participants were tested on a (fictitious) desirable or undesirable personality trait, such as 'intellectual flexibility' and were told that they occupied the middle rank in a group of several other participants tested at the same time. Participants were given their own (fictitious) score on the test, as well as the rank of each of the other participants, and were told that they could be shown the score of one other person in the group. The rank of the person participants chose to see could be of someone better off than

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24026-9

themselves (known as an 'upward comparison') or worse off than themselves (a 'downward comparison'). In most of the experiments using this method, participants chose to see the scores of those just above them in the rank order. That is, they wanted to compare their scores to slightly better scores (Wheeler, 1966), supporting the similarity hypothesis and the unidirectional drive upward.

The Reaction Approach

Individuals are presented with social comparison information, and their responses to that information are measured. In what is known as the Mr Clean/Mr Dirty study, job applicants casually encountered another applicant whose appearance was either socially desirable (dark suit, well-groomed, and confident) or socially undesirable (torn and smelly clothes, poorly groomed, and confused). Those participants who encountered Mr Clean showed a decrease in self-esteem measured a few minutes later, whereas those who encountered Mr Dirty showed an increase in self-esteem (Morse and Gergen, 1970). In this research, participants contrasted themselves against the other job applicant, but this is not always what happens. As we will discuss below, sometimes people will assimilate their self-judgments to the comparison standard.

The Narrative Approach

This nonexperimental approach attempts to measure the comparisons that people make or have made in their everyday lives. At the most basic level, participants are sometimes asked whom they have compared themselves with in the past. This places too great a burden on memory, however, and a better method is the *Social Comparison Record*, a diary recording of all comparisons as they occur over a period of 2 weeks, including the dimension of comparison, the direction (e.g., upward, downward), relationship to the target, and mood before and after the comparison (Wheeler and Miyake, 1992). When participants were experiencing negative affect, they made upward comparisons. In contrast, when the same participants were experiencing positive affect, they made downward comparisons. Such a result is consistent with *affect–cognition priming* models in which affect primes (make ready and available) cognitions about the self that are congruent with the affect. Experiencing negative affect primes a person to have negative thoughts about the self and thus to see others as superior to the self, leading to a contrastive upward comparison and further feelings of inferiority and negative affect. This research also showed that there was generally a negative emotional response to upward comparisons, and a positive emotional response to downward comparisons. Finally, downward comparisons were more frequent than upward comparisons.

Free response measures are also narrative in approach. These are spontaneous statements, usually made during the course of an interview or conversation, implying comparisons. Wood et al. (1985) reported that breast cancer patients made many free response comparison statements, most of which seemed to indicate downward comparison, such as "I think I am coping a little better than these other women."

The Changing Face of Similarity

Festinger (1954) argued that we must compare with similar others in order to make accurate self-evaluations, but he was not clear about the kind of similarity that was important. At first, investigators focused on (1) similarity on the dimension under evaluation, as we described for the rank-order paradigm. A person wanting to evaluate his or her speed in messaging would compare to someone who is slightly faster. This really does not tell you how fast you are compared to all other people, or whether you should take up thumb work as a profession; it merely tells you that you are or are not faster than the other person. Moreover, there is circularity in saying that we should compare to similar others – if we already know they are similar, why compare with them? To deal with this problem, Goethals and Darley (1977) proposed the 'related attributes' hypothesis. A person should compare to (2) someone who is similar to him or her on attributes related to and predictive of the dimension to be evaluated. Thus, the person should seek a comparison standard who is equally young and dexterous and who spends the same amount of time each day messaging. The person ought to perform about the same as this comparison standard, and if the person outperforms the standard, they will feel that he or she has an unusual ability. In short, the related attributes hypothesis does allow us to get a better understanding of our capabilities and to choose appropriate competitors.

A later attempt to untangle the question of similarity was based on the realization that Festinger was not so concerned with whether one labeled oneself as having low or high ability, but rather with the action possibilities of having a given level of ability. For example, one of the questions addressed by Festinger's theory was 'Can I do X?' Can I swim across the river? Can I complete a college degree? Can I have both a family and a career? One way to answer such questions is to compare oneself to someone – called a proxy – who seems to have a similar level of ability and who has already succeeded at the task (Wheeler et al., 1997). If people seem to have the same amount or more of the ability required as the proxy does, they could conclude with some confidence that they could also perform the task. There are two ways that people can be confident that they have as much of the underlying ability as the proxy: (1) they previously have competed directly against the proxy and have performed as well as the proxy at his best, or (2) they have observed the proxy performing and are similar to the proxy on attributes related to the ability. In both cases, they have established with some degree of certainty that they are similar in ability to the proxy and thus have the same action possibilities. This allows them to pursue the task with some confidence that they, like the proxy, have the right stuff to succeed. We believe that this is the correct interpretation of why Festinger argued that the comparison target should be similar in order for accurate self-evaluation to occur. Individuals want to evaluate their ability to perform some action successfully, and they determine this by comparing with a proxy of similar ability who has already performed that action.

Lockwood and Kunda (1997), who exposed first year and fourth year university students to a newspaper article about a superstar fourth year student, illustrated this nicely. Participants and superstar were matched on the related attributes of

gender and academic major. In a control condition, participants did not read about any target. The dependent variable was self-ratings on 10 adjectives related to career success. The first year students, with time to work toward the success of the superstar, gave themselves high self-ratings, but the fourth year students, who had missed the chance to be a superstar, did not. The first year students used the superstar as a proxy to determine what they might achieve in the future.

Opinion Comparison

As mentioned earlier, comparison of opinions has not been as actively studied as abilities and traits. This is somewhat surprising because persuasion has been an intense area of social psychological research for decades. The likely explanation is in everyday life, persuasive messages most often involve not just learning the source's (comparative) view but also being exposed to arguments. To identify the pure 'comparison effect,' researchers need to strip the comparison process away from the persuasive content, but this limits ecological validity.

Despite this matter, the related attribute approach devoted some attention to opinion comparison. First, in contrast to Festinger who treated opinions generically, Goethals and Darley distinguished between belief and value opinions. Beliefs refer to verifiable facts, whereas values are personal preferences. Comparisons with similar others on related attributes, such as background and general worldview, are of most utility for value assessment. For beliefs, however, someone who is dissimilar on related attributes was hypothesized to be more useful because they provide a different perspective to 'triangulate on' the truth.

The more recent triadic model (Suls et al., 2000) concurs with Goethals and Darley's prediction about values, but to assess facts, people should prefer to compare with someone who is superior on related attributes (conferring more expertise) not just dissimilar per se. Since few facts are completely value-free, however, someone who has more expertise and also shares the same basic values (i.e., worldview) is probably most preferred. The triadic model refers to such a person as a 'similar expert,' and illustrated by the finding that gay men are more likely to adopt safe-sex practices when they are advocated by someone who is both more knowledgeable and shares their sexual orientation (Kelly et al., 1991).

The new model is called 'triadic' because it adds a third kind of opinion comparison, which concerns predictions about affective responses to future situations (e.g., will I like that new film 'X'?). I can predict my future response with some confidence by learning about the response to 'X' of a proxy who has already viewed the film. The proxy should be considered appropriate if he/she shares related attributes or past pattern of agreement with respect to other films. Suls et al. (2000) presented supportive empirical evidence for the triadic model predictions.

Assimilation and Contrast

Until the 1990s, most students of social comparison theory believed that comparison always led to contrast with the comparison target. That is, an upward comparison (with someone better) lowered a person's self-evaluation, and a downward comparison (with someone worse) raised the self-evaluation, as shown in the Mr Clean/Mr Dirty study. There was even a subsidiary theory called 'Downward comparison principles in social psychology' that argued that when people are psychologically threatened they will deliberately seek out someone worse to compare themselves to in order to relieve their distress (Wills, 1981). This theory was quite influential during the 1980s until accumulating evidence caused it to be abandoned. The basic problem with the theory is that when people are psychologically threatened, they are more inclined to see others as superior to themselves and would have difficulty in finding and using downward comparison targets. We discussed previously the research showing that when people are experiencing negative affect, they are more likely to make upward comparisons rather than downward comparisons (Wheeler and Miyake, 1992). Presumably, this happens because negative affect causes people to perceive themselves negatively and to perceive most other people as superiors to themselves. In short, although downward comparison does generally make people feel better, people who are distressed or threatened find it difficult to find and use downward comparisons.

While believing that comparison usually led to contrast, researchers were aware that the rank-order studies, among others, had shown that people usually compare upward. If comparing upward leads to contrasting feelings of inferiority, why would people do it?

Upward Assimilation Theory

Buunk et al. (1990) asked married individuals and cancer patients to indicate how often they experienced both positive and negative emotions when comparing upward and downward. In the case of cancer patients, the downward comparison target was doing worse with the cancer, and the upward target was doing better. In the case of married individuals, the downward comparison was a marital relationship worse than that of the respondent, and the upward comparison was better. For both cancer and marital quality, the direction of comparison did not dictate how often positive or negative affect was experienced – both directions had positive and negative aspects. The paper was important in showing that upward comparison does not necessarily lead to negative affect, but can also lead to feeling inspired, comforted, happy, and pleased. Collins (1996, 2000), one of the authors of the paper just described, undertook a thorough review of the positive effects of upward comparison and proposed a theory of upward assimilation. She argued that upward comparison could lead to either contrast or assimilation depending upon whether the comparison was construed as showing similarity to, or difference from, the comparison standard. People are more likely to construe themselves as similar to upward targets if they *expect* to be similar to the upward targets. In the rank-order paradigm study discussed above (Wheeler, 1966), participants who expected to be more similar to the person above them in the rank order than to the person below them choose to compare with the person above them. The interpretation was that they had

assumed similarity upward and wanted to confirm that they were indeed similar to that person. According to Collins, some variables that might lead to such expectations of similarity to upward targets are (1) high self-esteem, (2) shared distinctiveness with the upward target (such as belonging to the same club), and (3) the direction of comparison. People with high self-esteem generally think that they are as good or better than others are and so should expect to be similar to upward targets. People who share distinctiveness with the upward target draw the conclusion that if they are similar on one distinctive thing, they are probably similar on performance as well. People who are comparing upward have higher expectation of similarity than those comparing downward because of the unidirectional drive upward. Of course, other variables in addition to self-esteem, shared distinctiveness, and the direction of comparison, such as similarity on related attributes (Goethals and Darley, 1977), might lead to expectations of similarity on performance. Whatever leads us to expect similarity and thus to construe that we are similar to upward comparison targets, the result is an increase of some sort in our own self-evaluations. For example, an academic person with six publications might compare with another academic person with 10 publications. If individuals construe 6 publications to be similar to 10 publications, they will conclude that they are solid (and satisfied) members of that class of academician with 6 to 10 publications. If they construe 6 publications and 10 publications to be dissimilar, they will feel inferior to those with 10 publications. Nosanchuk and Erickson (1985) provided a dramatic example of upward assimilation among bridge players. Players indicated that in a variety of situations, they would prefer to compare themselves to players of similar ability. However, when they named the particular players they would compare with, these players turned out to be objectively better players with more master points. Thus, the bridge players had assimilated their ability to the ability of the superior players.

Upward assimilation is an increase in one's self-evaluation because of making an upward comparison. Although this has often been obscured, it is not simply a happy feeling or feeling positive about the future but is a perceptual change in one's self-evaluation. Collins (1996, 2000) noted that downward assimilation is not likely to occur, because of the unidirectional drive upward. Our need to be better and to improve works against downward assimilation.

The Selective Accessibility Model

Mussweiler (2003) proposed a theory in agreement with that of Collins (1996, 2000) but using concepts from the social cognition literature and giving more detail about the mechanisms involved in social comparison. At the moment of exposure to social comparison information, a person makes a tentative judgment of similarity or dissimilarity to the comparison target. Salient features (such as gender, race, or age) of the comparison target determine this judgment. Then the person searches for information consistent with the preliminary judgment of similarity or dissimilarity. Whether one searches for similarity information or dissimilarity information, one is likely to find it because our view of ourselves is remarkably rich and complicated. It is easy to find information whether one is similar or dissimilar to another person. That information then becomes selectively accessible when we make judgments about ourselves. If we have searched for information that we are similar to the standard, we are likely to assimilate our self-evaluations toward the target. If we have searched for information that we are dissimilar to the target, we are likely to contrast our self-evaluations away from the target.

The selective accessibility model (SAM) is elegantly simple and explains a wide variety of empirical results. Contrast is more likely if the standard is extreme or unattainable, or if the standard belongs to an out-group, both of which would lead to an initial hypothesis of dissimilarity. Assimilation is more likely to occur if there is psychological closeness with the standard, which would lead to a search for similarity. If people prior to comparison are asked to point out similarities between two drawings (i.e., procedural priming), they subsequently will more likely assimilate to the comparison standard than if they have pointed out differences. A shortcoming is that the SAM does not explicitly recognize that self-protective concerns may bias comparisons. For example, there is abundant evidence that people assimilate upward, but there is almost no evidence of downward assimilation. The SAM would argue that upward and downward assimilation are equally likely. In fairness, however, one could modify the SAM slightly and state that one would not search for similarity with a downward standard because of self-esteem concerns.

Mussweiler and colleagues have performed numerous experiments that support the SAM. However, an unfortunate aspect of this research is that adequate control groups are rarely used. Conditions that should lead to assimilation are compared to conditions that should lead to contrast. When there is a difference, the researchers argue that they have confirming evidence for the SAM. However, without a control group that is not exposed to any comparison target, there is no way to know whether assimilation has occurred, contrast has occurred, or both have occurred.

Mussweiler makes an important distinction between subjective and objective self-evaluation. An upward comparison standard has two effects. First, it produces a search for similarity to the standard and thus makes accessible knowledge that the self is similar to the high standard. Second, it serves as a reference point (or scale anchor) for making a subjective self-evaluation. The comparison standard may alter the interpretation of subjective rating scales ('very intelligent' may be interpreted differently in the context of a professor than in the context of an infant). These two influences work in opposite directions when making a subjective judgment. If one compares to an intelligent standard (say, professor) one may search for and find information about the self that indicates high intelligence. If then asked to indicate one's intelligence on a subjective scale from 1 to 10, one may assume that 10 is close to the professor's intelligence and so may contrast oneself away from this endpoint despite having accessible information that one is intelligent. Whether assimilation or contrast occurs depends upon the relative strength of the selective accessibility and the reference point effects. On the other hand, if objective self-evaluations are used, there is

no reference point effect to worry about. Objective self-evaluation questions pertaining to intelligence might inquire about the amount of time taken to solve a crossword puzzle, or the number of unknown words encountered in a typical book. Because of the ambiguity of subjective self-evaluations, Mussweiler always uses objective self-evaluations.

Assimilation and Contrast in Education

Social comparison tendencies should be strong in education settings when students compare their performance with others in their class or school and draw conclusions about their academic ability. The frog pond effect (Davis, 1966) refers to the fact that students attending highly selective schools have lower academic self-concepts than those attending nonselective schools (holding academic ability constant). This has been demonstrated in elementary, middle, and high schools across more than 40 countries (Seaton et al., 2009). The explanation is that people compare with those around them and ignore the larger context. Students in a highly selective school ought to have high academic self-concepts; after all, they have earned a place in the highly selective school. However, when they compare themselves with their classmates, they find that they are not superior in ability, and they ignore the fact that their classmates are near the top of the ability continuum. Zell and Alicke (2009) have referred to this contextual neglect of general comparison information as the 'local dominance effect' and have shown that being the best or worst performer in a group of five people influences self-evaluation more than the information about one's standing among 1500 other individuals. During early development, we compare with family members and small groups of peers, and this continues to be a salient mode of self-evaluation as we mature.

In an international collaboration, Huguet et al. (2009) conducted a highly detailed study of social comparison in French public schools. They showed that the frog pond effect was indeed due to the students' perceived relative standing in the class; those who perceived themselves as having relatively low standing had lower academic self-concepts. Thus, contrastive social comparison caused the frog pond effect. The students also named those classmates with whom they chose to compare grades. They named classmates who on average had slightly better grades, and to the extent that they did so, the students had higher academic self-concepts. Thus, assimilative social comparison occurred independently of contrastive social comparison. The assimilative comparison related to comparisons the students deliberately chose to make, whereas the contrastive comparison related to the comparisons forced on the students by the class average. Of the two effects, the reduction in self-concept due to contrastive comparison was much larger than the increase in self-concept due to assimilative comparisons.

Although students named objectively superior classmates as the ones they usually compared with, they believed that they were similar to these superior others, precisely the results of the Nosanchuk and Erickson (1985) study of bridge players. They had assimilated their own ability upward.

Chronic Illness and Social Comparison

Earlier we mentioned the application of comparison processes to the coping of cancer patients. As illness creates uncertainty and stress, social comparison can play a role in coping and adjustment of medical patients. There have been many descriptive–correlational studies and a smaller set of experiments using the reaction methodology to explore the role of social comparison among medical patients (e.g., Stanton et al., 1999). There are two aspects of illness that can be compared: (1) the severity of the illness and (2) how well one is coping with the illness. A recent literature review (Arigo et al., 2012) concluded that identifying or having an actual contact with someone who is coping better tends to be beneficial for patients, while contrasting or being exposed to information about someone with more severe illness tends to be beneficial. In contrast, downward comparisons about coping and upward comparisons about illness severity seem to be unhelpful. The descriptive–correlational research, based on narration and selection methods, provides the most consistent evidence. Making strong claims from these trends is premature, however. As in the basic research literature, few experimental studies included no-comparison controls to be able to determine whether assimilation or contrast has occurred or both have occurred. The few studies that did include no-comparison groups have not found consistent and statistically significantly greater benefits for upward comparisons vs no-comparison controls, and the evidence for downward comparison is also mixed. The review notes that in most of the experiments, patients learn very little about the comparison patient beyond their relative standing with the patient. Consequently, patients may not perceive them to be appropriate proxies to assess their success in the future. The proxy model suggests that the remedy would be to present higher functioning proxies who share underlying attributes (age, socioeconomic status, life circumstances) with the patients. This kind of research has the potential to inform medical professionals about the kinds of comparison targets and instructions that would be most beneficial for psychoeducational materials (e.g., videos) in medical rehabilitation (see also Kulik and Mahler, 1987).

Social Inequality, Invidious Comparisons, and Health

A provocative finding in research on social determinants of health is what is referred to as the 'socioeconomic gradient' (Marmot, 2004; Wilkinson and Pickett, 2010). Individuals who rank higher in the social hierarchy enjoy better health than those who are below. Of course, poor nutrition, crowding, inadequate medical care, etc., no doubt, play a role. However, the association with health occurs at every level of the hierarchy, not simply below the poverty threshold. Marmot et al. (1991) found that individuals at the highest level of the British civil service had better health than those just below (all levels were well above the poverty line). Wilkinson and Pickett (2010) reported that differences in average income across countries had little effect on health, but where the person falls in the pecking order of their peers does have an effect on health. These researchers have proposed that invidious social

comparisons are the culprit, producing chronic stress among those lower on the social ladder, which may in turn increase the risk of psychological and physical illness.

This is a continuing area of research by health psychologists and epidemiologists as not all of the evidence is consistent, and political economists are looking for other explanations. Whereas some researchers maintain that relative standing influences health when one looks at aggregate-level studies (population or community levels), other researchers argue that only individual-level studies can differentiate between relative and absolute income effects on health (Wagstaff and van Doorslaer, 2000). The individual-level research has not shown definitive evidence implicating relative income. Our previous description of advances in basic theory and research also suggests that caution may be appropriate. The social gradient hypothesis assumes that upward comparisons are universally aversive, but as we have noted certain conditions facilitate upward assimilation. Some members of a population may convert a potentially invidious upward comparison into an inspiration. Income inequality also leads to lower happiness (Oishi et al., 2011). Between 1972 and 2008, Americans were, on average, happier in those years of lesser income inequality than in those years of greater inequality. This was true only of lower income people and was due to the feeling of unfairness and lack of trust rather than to household income itself. This means that the subtleties revealed by recent social comparison theory and research may have important societal consequences for a world in which income inequality is increasing.

See also: Academic Achievement Motivation, Development of; Attitude Formation and Change; Group Processes, Social Psychology of; Health Self-Regulation, Motivational and Volitional Aspects of; Heuristics in Social Cognition; Pay, Compensation, and Performance, Psychology of; Power, Politics, and Influence in Organizations; Schooling: Impact on Cognitive and Motivational Development; Social Categorization; Social Cognition; Social Identity in Social Psychology; Social Psychology: Research Methods; Social Psychology.

Bibliography

Albert, S., 1977. Temporal comparison theory. Psychological Review 84 (6), 485–503.

Arigo, D., Suls, J., Smyth, J., 2012. Social comparisons and chronic illness: research synthesis and clinical implications. Health Psychology Review 6, 1–62.

Buunk, B., Collins, R., Taylor, S., VanYperen, N., 1990. The affective consequences of social comparison: either direction has its ups and downs. Journal of Personality and Social Psychology 59, 1238–1249.

Collins, R.L., 1996. For better or worse: the impact of upward social comparison on self-evaluations. Psychological Bulletin 119 (1), 51–69.

Collins, R.L., 2000. Among the better ones: upward assimilation in social comparison. In: Suls, J., Wheeler, L. (Eds.), Handbook of Social Comparison: Theory and Research. Kluwer Academic/Plenum Press, New York, pp. 159–171.

Davis, J., 1966. The campus as frog pond: an application of the theory of relative deprivation to career decisions of college men. The American Journal of Sociology 72, 17–31.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7, 117–140.

Goethals, R.R., Darley, J.M., 1977. Social comparison theory: an attributional approach. In: Suls, J.M., Miller, R.L. (Eds.), Social Comparison Processes: Theoretical and Empirical Perspectives. Hemisphere, Washington, DC, pp. 259–278.

Huguet, P., Dumas, F., Marsh, H., Regner, I., Wheeler, L., Suls, J., Goethals, G.R., Nezlek, J., 2009. Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): an integrative study. Journal of Personality and Social Psychology 97 (1), 156–170.

Kelly, J., St Lawrence, J., Diaz, Y., et al., 1991. HIV risk prevention following intervention with key opinion leaders in the population: an experimental analysis. American Journal of Public Health 81, 168–171.

Kulik, J., Mahler, H., 1987. Effects of preoperative roommate assignment on preoperative anxiety and postoperative recovery from coronary bypass surgery. Health Psychology 6, 525–543.

Lockwood, P., Kunda, Z., 1997. Superstars and me: predicting the impact of role models on the self. Journal of Personality and Social Psychology 73, 91–103.

Marmot, M.G., 2004. The Status Syndrome. Holt, NY.

Marmot, M.G., Smith, G.D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., Feeney, A., 1991. Health inequalities among British civil servants: the Whitehall study. Lancet 337, 1387–1393.

Morse, S., Gergen, K.J., 1970. Social comparison, self-consistency, and the concept of self. Journal of Personality and Social Psychology 16, 148–156.

Mussweiler, T., 2003. Comparison processes in social judgement. Psychological Review 110, 472–479.

Nosanchuk, T.A., Erickson, B.H., 1985. How high is up: calibrating social comparison in the real world. Journal of Personality and Social Psychology 48, 624–634.

Oishi, S., Kesebir, S., Diener, E., 2011. Income inequality and happiness. Psychological Science 22 (9), 1095–1100.

Seaton, M., Marsh, H., Craven, R.G., 2009. Earning its place as a pan-human theory: universality of the big-fish-little-pond effect (BFLPE) across 41 culturally and economically diverse countries. Journal of Educational Psychology 101, 403–419.

Stanton, A., Danoff-Burg, S., Cameron, C., Snider, P., Kirk, S., 1999. Social comparison and adjustment to breast cancer: an experimental examination of upward assimilation and downward evaluation. Health Psychology 18, 151–158.

Suls, J., Martin, R., Wheeler, L., 2000. Three kinds of opinion comparison: the Triadic model. Personality and Social Psychology Review 4, 219–237.

Wagstaff, A., van Doorslaer, E., 2000. Income inequality and health: what does the literature tell us? Annual Review of Public Health 21, 543–567.

Wheeler, L., 1966. Motivation as a determinant of upward comparison. Journal of Experimental Social Psychology 2 (Suppl. 1), 27–31.

Wheeler, L., Miyake, K., 1992. Social comparison in everyday life. Journal of Personality and Social Psychology 62 (5), 760–773.

Wheeler, L., Martin, R., Suls, J., 1997. The proxy model of social comparison for self-assessment. Personality and Social Psychology Review 1, 54–71.

Wills, T.A., 1981. Downward comparison principles in social psychology. Psychological Bulletin 90, 245–271.

Wilkinson, R., Pickett, K., 2010. The Spirit Level: Why Greater Equality Makes Societies Stronger. Bloomsbury Press, New York.

Wood, J.V., 1996. What is social comparison and how should we study it? Personality and Social Psychology Bulletin 22 (5), 520–537.

Wood, J.V., Taylor, S.E., Lichtman, R.R., 1985. Social comparison in adjustment to breast cancer. Journal of Personality and Social Psychology 49, 1169–1183.

Zell, E., Alicke, M.D., 2009. Contextual neglect, self-evaluation, and the frog-pond effect. Journal of Personality and Social Psychology 97 (3), 467–482.

Social Competence During Adolescence Across Cultures

Grace Iarocci and Emily Gardiner, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we consider what it means to be socially competent within the unique contexts that characterize adolescence. This critical period constitutes an especially salient time for examining social competence, as rapidly maturing developmental capacities have significant implications in terms of adolescents' rapidly emerging new social-cognitive resources, changing social goals, and shifting social focus. We discuss the key developmental milestones of adolescence, and illustrate the transactional nature of the social competence construct. Finally, we consider how socially competent processes and outcomes in adolescence may vary cross-culturally.

Developmental scientists have increasingly acknowledged the important role of social competence in children's everyday lives. Social competence has been studied from different levels of analysis, including at the level of cognition (e.g., perspective-taking), overt behaviors/skills (e.g., use of eye contact), and outcomes (e.g., social status, peer acceptance). However, social competence cannot be reduced to any single index or ability; the active and skillful coordination of multiple lower order as well as contextual processes are necessary in order to adequately meet the social demands of a particular situation (Iarocci et al., 2007).

Social competence is both a developmental phenomenon that can be measured over the course of a child's development (ontogenesis), as well as a characteristic of a particular social encounter where the time scale is in the order of seconds and minutes (microgenesis). Thus, continuities and discontinuities in its development are expected, as children are better able to coordinate abilities and take advantage of resources with increasing age, but may be less competent at certain developmental stages or in specific social contexts.

We must also consider that what constitutes socially competent behavior changes throughout development. Whereas in infancy social competency may be defined in terms of the quality of the parent–child relationship, the social world becomes significantly more complex in adolescence, and social interactions begin to require proficiency in a range of sophisticated social repertoires and skills (e.g., appreciation of social context/subtle social nuances, initiating and maintaining conversations, offering emotional support to others).

Adolescence is an especially salient developmental period for examining social competence. The transition to adolescence is marked by a rapid convergence of significant neurobiological, cognitive, emotional, and social transformation. It is recognized as the second major 'window' of development and change, next only in significance to early childhood. For example, significant advances in cognition, in particular metacognition and representational capacity, as well as competence in abstract thinking and complex problem solving, help youth to develop a differentiated and psychologically rich understanding of their social world. Relevant social interactions, such as those with parents, enhance the continued development of complex social understanding, and foster critical social competencies including empathy and self-understanding. However, entering this new social-psychological phase is accompanied by a change in social focus, as the amount of time spent with parents precipitously drops, while time spent with peers increases. During this transition, the parent–child relationship requires renegotiation as children strive toward autonomy and parents must find new ways of relating to, and being supportive of, their children.

Paradoxically, adolescence is both a time in which greater focus is directed inward toward self-reflection, as well as outward toward consideration of others' perspectives. These advancing reflective processes serve youth well with regard to self-identity formation, a key developmental task during this period, as they begin to consider how they are similar, yet also unique from their peers. During this time, the self is simultaneously differentiated from others and viewed in relation to others. A person's identity is shaped by social experiences (Martin et al., 2008) and therefore depends on the youth's ability to relate to others. Within the peer context, intergroup processes take on added significance as youth work to establish their own identity and status within the group. The steadily increasing centrality of peer and romantic relationships heighten the social challenges for adolescents. Those youth who successfully face these challenges improve their social competence whereas those who are unsuccessful may further diminish both their social status and social competencies (Monahan and Steinberg, 2011).

Social Competence in Adolescence

Social competence is conceptualized as incorporating characteristics of individuals and their environments, whereby these two sources of influence are transactional. Within this framework, social competence involves the active and skillful coordination of multiple processes and resources available to the child. Such processes include basic sensory/perceptual and cognitive abilities (i.e., attention, memory, motivation) fundamental to the development of higher-order social-cognitive processes. For example, sharing attention, understanding that others' thoughts and feelings differ from one's own, and expressing emotions and recognizing them in others, are only a few of the sophisticated social abilities involved in social competence. Furthermore, socially competent behavior also requires that a child coordinate these basic social capacities with available contextual resources

(e.g., opportunities for peer interactions). This higher-order coordination is a critical component of social competence because it permits children to appropriately match their social goals with the demands of the social context. Accordingly, social competence entails the development of strategic processes (i.e., techniques) that allow one to evaluate which resources are best given the demands of a particular task in a particular context. Social learning, through mediation and scaffolding, likely influence the development of these strategies or 'techniques,' as well as surrounding systemic factors. For example, the practices of parents and other relevant social mediators (e.g., teachers) play a significant role in shaping how social strategies are implemented in socially appropriate ways.

Various sociocultural systems influence the development of social competence through both proximal (direct) and distal (indirect) effects (Bronfenbrenner, 2000; Cicchetti and Toth, 1997). Direct influences include the youth's interactions with significant persons or events in their lives (e.g., parenting customs, sibling and peer relations, and teaching practices). Indirect influences are thought to emanate from macrosystems, or global-political contexts; mesosystems, which encompass the patterns, beliefs, and values of the culture in which the child exists; and exosystems, which comprise the various formal and informal social structures in the adolescent's environment, including the neighborhood, schools, and local government policies on education and health (Bronfenbrenner, 2000; Cicchetti and Toth, 1997). Within these systems, risk and protective factors are present throughout the course of development and may operate through distal and/or proximal processes to support or hinder the development of social competence.

With this conceptual framework in mind, we now focus on key developmental tasks (i.e., renegotiating parent–child relationships, developing friendships and intimate relationships, and self and cultural identity achievement) of adolescence that illustrate the transactional nature of social competence. The ways in which these major milestones are achieved both influence the future course of social competence development and are shaped by the adolescent's existing social competencies.

The Parent–Child Relationship

The parent–child interactions provide the primary social learning context from infancy. A variety of social-cognitive and socio-emotional processes, such as emotion regulation and recognition, referencing, gaze following, gesturing, and communication, are first evident in parent–child interactions. For example, between the second and third month of life there are significant developments in parent–infant behavioral synchrony, turn taking, and reciprocity, all of which are precursors to a healthy attachment. In typically developing children between the ages of 3 and 6 years, however, there is a marked decrease in time spent in direct contact with caregivers and a concurrent increase in time spent with peers. As such, peer relationships play an increasingly prominent role in a child's social development during the school years. This trend continues throughout middle childhood as increased availability of social opportunities is coupled with increased social interest, social-cognitive maturation, and independence. Despite these contextual changes, parenting behaviors, specifically those related to supporting children's changing developmental needs, continue to impact the child's burgeoning competence. While there are key parenting behaviors that support growth in social skill acquisition (i.e., warmth, sensitivity, and responsiveness), there are also those that interfere with, or hinder these developmental processes (i.e., parental negativity and overreactivity).

There is generally strong continuity of social competence during earlier developmental periods (from early childhood to middle childhood), yet studies of continuity from childhood to adolescence have yielded inconsistent findings (Obradović et al., 2006). Longitudinal studies have shown that although attachment classification in infancy does not predict attachment beyond childhood, functioning during childhood is significantly related to functioning in adolescence, which predicts future functioning in young adulthood (Simpson et al., 2007). Secure infants who are more socially competent in middle childhood have more secure friendships at age 16, and these individuals demonstrate greater emotional expression in their romantic relationships as young adults. Those who begin as secure are much more likely to experience high-quality relationships across development. Early attachment experiences are foundational and influence one's ability to form intimate and stable bonds over time. The parent–child attachment provides a framework within which the child forms expectations about the predictability of relationships. Those youth who have a secure attachment with their parents are likely to view peer interactions and relationships as similarly predictable and safe, providing a secure context for social exploration. Similarly, these early experiences influence later representations of romantic relationships, suggesting continuities between experiences with primary caregivers and the quality of later attachments (Roisman et al., 2009). Thus, although parents play a less prominent role during the adolescent period due to the changing demands of peer-dominated social interactions, an established secure base of parent–child attachment allows adolescents to form their own secure attachments with friends and romantic partners. Parents continue to be a significant influence in the development of social competence, but the relation between parent–child attachment and social competence is modest. For instance, accounts that include other mediational child variables, such as empathy, efficacy beliefs, emotional regulation, and social information processing, as well as community variables, such as neighborhood context, school factors, extracurricular activities, and religious involvement, may better explain the link between parent–child attachment and social competence.

Friendships and Intimate Relationships

During adolescence, peer relationships typically assume increased significance. The relationships that emerge during this time, such as close friendships and romantic relationships, tend to be less based on shared play activities and more on a higher degree of intimacy and self-disclosure (i.e., sharing of personal thoughts and feelings), greater mutuality, and increased reciprocity. During this period, a key behavioral expression of social competence includes the successful development and maintenance of more intimate peer relationships. This period likely involves both an emergence of new skills (e.g., social problem-solving and abstract reasoning), as well as

further developments in existing social-cognitive processes, such as perspective-taking (theory of mind), emotion processing, and pragmatic language abilities. Whereas gains in social competence enable individuals to better contend with social developmental tasks, the successful negotiation of these tasks, in turn, creates further opportunities for continued social learning and growth. For instance, close friendships with more competent peers provide adolescents with valuable social learning opportunities to refine and practice more sophisticated skills, and to receive important social feedback.

Self and Cultural Identity Achievement

Although the process of identity development may appear to be a solitary endeavor that occurs within the individual adolescent, the social-constructionist perspective on identity is that a self only exists in relationship to other selves (Martin et al., 2008). Related to this notion is that the choices adolescents have are typically constrained by their social situation, and similarly, their decisions are made within the context of societal demands. There are several cognitive shifts in higher-order reasoning about the self and abstract and conceptual ways of perceiving the self and others' perspectives that facilitate the exploration of various options. Alongside these cognitive milestones are important changes in the social landscape including emerging peer groups, cliques, and cross-gender relationships. Within these peer contexts, adolescents learn a great deal about the self, social and group conformity and deviancy, social hierarchies, and ostracism. In addition, moving in and out of different peer contexts provides opportunities for experimentation with 'trying on' various identities. Being included and accepted is important in the formation of one's identity, and the identity experimentation that inclusion allows may have a particularly formative impact on the development of social competence. For example, lonely or socially isolated adolescents are more likely to turn to the Internet to experiment with their identity, and apparently benefit from the relative anonymity that this setting provides in learning how to relate to people.

Social and cultural expectations about adolescence also have a profound impact on self-identity development. Parental and cultural beliefs about the adolescent's autonomy, responsibility, and achievements provide a point of reference for the adolescent. However, experimentation with social rules and structures is common. Through active exploration of alternatives to mainstream social roles, adolescents solidify their identities and eventually firmly adopt or reject certain social and cultural norms. Identity achievement occurs when the adolescent has explored and committed to fundamental aspects of their identity. Supportive parents, schools, and communities that encourage exploration are thought to promote identity achievement. Identity achievement is associated with higher self-esteem, increased critical thinking, and advanced moral reasoning in adolescents.

Some aspects of adolescents' identities, such as commitments to occupational, religious, or ethnic identity, may occur later in development. For example, unlike most adolescents of the majority culture, minority adolescents must consider the degree to which their racial or cultural background will be part of their identity. Identification with a minority culture may result in racism, ostracism, and marginalization, and thus, some minority adolescents are undecided or reluctant to make a firm commitment to a particular ethnic identity. However, developmental and social changes conspire to resolve this crisis in most youth. Whereas in early adolescence minority teenagers may deny any interest in their racial or cultural background, by the latter part of this stage, as moral reasoning becomes more sophisticated and social awareness of the relations between the dominant and subcultures are raised, minority adolescents may be ready to explore their heritage. Positive social interactions with other members of the same culture, and attendance at religious services or cultural events, can increase the adolescents' knowledge and instill a sense of pride in their ethnic background. Achieving a positive ethnic identity has been found to be associated with better grades, higher self-esteem, better relations with family and friends, and reduced physical and relational peer aggression.

Cultural Variations in Social Competence

Social competence is particularly advantageous in modern society as the cultural context becomes increasingly diverse. For adolescents developing in a multicultural environment, being able to effectively navigate the social-cultural demands of multiple and diverse perspectives is likely to be integral to their healthy social adaptation and development. However, to fully understand the impact of social competence on development, we must consider the variations in social relationships, social contexts, and social goals and values across cultures.

Developmental research on social competence specifically, or development more broadly, raises the long-standing concern about the universality of psychological principles and findings. Within a universalist framework, researchers determine whether 'universals' found in commonly studied groups hold true in others. However, we must also consider the varying social resources, demands, goals, and values that are reflective of the diverse environments in which social competence develops. Cross-cultural researchers refer to this period as 'adolescences' (Larson and Wilson, 2004, p. 299) to communicate the greatly divergent trajectories that individuals navigate during this period. This latter perspective is known as a relativist framework. This perspective examines each culture for its uniqueness, and is not concerned with establishing cross-culture commonalities (Iarocci et al., 2009).

In examining how the study of cross-cultural continuity and diversity can be approached, it is first important to consider the question: What is culture? This, too, is a diversely defined and complex construct that exerts significant influence in terms of a society's values, expectations, practices, beliefs, and patterns of interaction. We adopt Larson and Wilson's (2004) comprehensive definition:

> Culture might best be understood as the frame of reference that people begin acquiring in infancy and that molds how they think about themselves and others. It is embedded in the language one learns, one's intimate relationships and daily activities, the religion one is taught, and other existential fundamentals, such as when, where, and how one experiences different emotions. Each culture contains its own indigenous psychology, which includes developmental tasks and precepts, as well as postulates what is 'true, beautiful, good, and normal'. (p. 303)

This explanation highlights the pervasive influence of the cultural environment, and demonstrates how its influence extends across socio-ecological levels over time to become integrated within everyday actions and interactions. Thus, one's behavior cannot be evaluated as competent or not without a deep understanding of the specific social influences within the culture. It is the overarching norms and mores present within a particular society that dictate what is considered appropriate. For example, social behaviors considered critical for successful interaction in individual-oriented Western nations, such as sustained and appropriately modulated eye contact during conversation, may actually be considered socially inappropriate in others, such as in socially oriented Asian countries.

Researchers often draw distinctions between collectivist and individualist cultures. However, we must acknowledge that these are merely labels used to refer to aggregated groups of individuals who are not homogeneous. Modern adaptation to urbanization and globalization has led to the coexistence of elements from both dimensions, at least to some extent. This review is not intended to be exhaustive, and as such, will not delve into this discussion in detail. Instead we aim to provide examples that will highlight how interrelated social components contribute to one's expression of socially competent capacities in diverse contexts.

Diversity in Relationships

Resources that are essential to the development of social competence emerge from the interactions that take place between children and their primary caregiver(s). In Western nations, child-rearing activities primarily fall to the nuclear family unit and positive social outcomes are associated with secure child–caregiver attachment. In cultures where child-rearing may rely on greater involvement from extended family or community members, such as Israeli kibbutzim, hunter-gatherer tribes in African cultures (e.g., Dogon of Mali, Efé of Congo, and Gusii of Kenya), and Aboriginal communities, it is important to consider the diversity in caregiving contexts and the potential implications for social competence. Interestingly, even in such distinct settings, the importance of a secure attachment with the primary caregiver(s) emerges as critical. For instance, infants raised in kibbutzim who are securely attached with their mother or father have been found to attain higher school achievement and demonstrate greater emotional maturity, as well as have higher self-perceived social acceptance and self-esteem in adolescence, all of which enable an individual to participate competently in social interaction (Sagi-Schwartz and Aviezer, 2005).

In Western societies, authoritative parenting has consistently been identified as the ideal style over authoritarian, indulgent, or neglectful parenting. Authoritative parent–adolescent interactions are characterized by high degrees of warmth, control, and autonomy granting, and are associated with positive adolescent outcomes including higher academic achievement, fewer mental health concerns, and stronger self-esteem (Steinberg, 2001), yet this is not universal. Authoritarian parenting, distinguished by high control and low warmth, has been associated with the same kinds of positive outcomes in American minority cultures, such as for African-American children and adolescents, as well as for adolescents in Asian societies, such as China, Japan, Vietnam, and South Korea. In comparison to Western parents, those who utilize nondemocratic approaches (i.e., parents' boundaries are nonnegotiable) may be designated as low in parental warmth. However, close attachments can still be characteristic of these relationships. It is likely the ways in which these high-control approaches are interpreted within the relevant cultural context (i.e., as indicative of interest and concern), as well as how they are representative of broader societal circumstances and values, that mediates these different developmental pathways. For example, authoritarian parenting may be protective in collectivist societies that place great importance on respect for authority and group cooperation. Authoritarian-like parent–child interactions, in which adolescents must submit to parental rules, may be more culturally appropriate in such environments as they prepare children to comply with the implicit group harmony-focused social demands of their society. These mechanisms may also have great practical relevance within North America where neighborhoods characterized by low socioeconomic status, high ethnic and racial diversity, and a high proportion of young people are associated with greater violence and delinquency. In circumstances such as these, parents may adopt high-control approaches in an effort to prevent their children from becoming engaged in risky activities.

Diversity in Social Contexts

Adolescence is a developmental period that involves confronting novel social challenges and demands. These may include navigating the complexities associated with more intimate peer and romantic relationships, as well as adopting greater responsibility in contributing to the family economy. Although the demands facing adolescents in Western and non-Western societies differ, this divergence is actually greatest when examining how imposed expectations vary across genders in diverse cultures. Male and female experiences during this stage can be so "dissimilar as to constitute separate adolescences" (Larson and Wilson, 2004, p. 319). For example, in areas of the Middle East, Latin America, Africa, Asia, and South Asia, entering adolescence is accompanied by a drastic transformation in roles. Males are granted increased autonomy and are encouraged to spend time with peers, whereas females experience restrictions on their independence, as unsanctioned relations with the opposite sex are considered highly inappropriate and may bring shame upon the family (Larson and Wilson, 2004; Lee et al., 2010). Moreover, females are often expected to fulfill caretaking roles by taking on greater domestic and childcare responsibilities. In contrast, males are encouraged to assert themselves, to express their ideas, and to make their own decisions. Although these latter behaviors may appear to be consistent with the goals of those within Western society, research conducted with adolescents in Taiwan identified that gender was a better predictor of autonomous thought than cultural value orientation (i.e., collectivistic or individualistic), demonstrating that these patterns may not be reflective of a West-East dichotomy, but of a stark division in how the two genders are viewed and socialized cross-culturally (Lee et al., 2010).

Social demands and expectations become particularly complicated as non-Western societies become introduced to

Western ideals and practices through urbanization, globalization, and increased access to information (i.e., greater use of the Internet and social media, and exposure to international broadcasting sources). This can result in value system adaptations in younger generations, coupled with the presentation of competing demands imposed by different groups (i.e., peers and family). For example, parents may assume that their adolescent daughters will embrace values from their culture of origin; however, the adolescent may perceive these values as outdated or incongruous with their 'idealized' culture. These conflicts are evident even in groups that have emigrated from non-Western societies to Western countries. For example, adolescent daughters whose parents immigrated to Canada from India, Pakistan, and Bangladesh seem to experience a mismatch between their traditional social cultural expectations and those of the Western setting in which they live. These youth were discouraged from intermingling with their Canadian peers, social interactions were highly supervised, and they reported that they were viewed as inferior in comparison to male peers (Talbani and Hasanali, 2000). Children of immigrants must navigate at least two, at times incompatible, cultures to both successfully meet the social demands of their same-aged native-born peer group as well as maintain ties to their cultural heritage.

Aboriginal groups worldwide may share a similar experience even though their histories involved persecution and colonization rather than emigration. In Canada, the term 'Aboriginal' refers to the country's original inhabitants prior to the European invasion. First Nations (the indigenous peoples of the Americas located within the boundaries of Canada, excluding the Arctic), Inuit (indigenous peoples located in the Arctic), and Métis (persons of mixed European-First Nations ancestry) are collectively referred to as Aboriginal peoples. These groups share a history of oppression and loss of culture that has had a deleterious impact on subsequent generations. Currently, youth strive to participate and succeed in the majority culture, which, in some cases, comes at a significant cost to their own personal as well as the collective well-being. The risk and protective factors associated with identification with ancestral identity have begun to be studied among Aboriginal youth. There is growing evidence that identification with one's cultural heritage is particularly protective for these youth with regard to peer relations and educational and social outcomes.

Diversity in Social Goals and Values

In Western societies, adolescence is typically conceptualized as a developmental phase in which central milestones include asserting greater independence, forming a unique identity that distinguishes adolescents among family members, and shifting social focus to the development and maintenance of stronger peer relationships. However, social goals in adolescence vary across cultures. For example, in non-Western countries, including India and Asia, as well as across Arab societies (e.g., Algeria, Saudi Arabia, Lebanon, Palestine, Egypt, and Jordan), the social goals of adolescence are not peer focused but remain family centered. These adolescents often remain more connected with their family members than their North American counterparts do, viewing their parents as companions, spending more time in their company than with their peers, and confiding with their parents about personal matters. The parent–adolescent relationship renegotiation characteristic of the West is also not a central developmental task in such contexts, as these relationships remain hierarchically organized, with adolescents expecting to defer to parents' authority. In such societies, family interdependence, as opposed to independence, is the goal (Dwairy et al., 2006).

Interdependence is also a central concept around which adolescents in collectivist cultures base their identities. In this setting, the self is defined as a 'we-self,' as opposed to a 'me-self,' and reflects the individual's ability to conform to the values and goals of the broader society. It is less tied to how one is unique or what can be achieved independently. In interdependent contexts, socially desirable, and therefore also socially competent, individual characteristics are those that facilitate harmonious group functioning, such as shyness, cooperation, and vigilance. In stark contrast, these qualities may be pathologized in North America, with numerous studies documenting the socially maladaptive outcomes associated with introversion. In the latter context, adolescents are expected to be assertive and independent, and individuals who effectively communicate such qualities enjoy greater social opportunities. Thus, socially competent behavior must be evaluated with an understanding of the social goals and values of a particular society.

Adolescence is the ideal frame of time to capture the multidimensional and transactional landscape of social competence. The unprecedented changes in social abilities and opportunities during this developmental period shape, and are shaped by, social competence. However, there are important variations in social relationships, social contexts, and social goals and values across cultures that must be considered to fully grasp the impact of social competence on development.

See also: Adolescence, Sociology of; Childhood and Adolescence: Developmental Assets; Friendship During Adolescence and Cultural Variations; Identity and Identification, Social Psychology of; Identity in Childhood and Adolescence; Mass Media and Cultural Identity; Personality Changes During Adolescence Across Cultures; Self and Identity Development During Adolescence across Cultures; Social Identity in Sociology; Stress in Adolescence: Effects on Development.

Bibliography

Bronfenbrenner, U., 2000. Ecological systems theory. In: Kazdin, A.E. (Ed.), Encyclopedia of Psychology, vol. 3. American Psychological Association, Washington, DC, pp. 129–133.

Cicchetti, D., Toth, S.L., 1997. Transactional ecological systems in developmental psychopathology. In: Luthar, S.S., Burack, J.A., Cicchetti, D., Weisz, J.R. (Eds.), Developmental Psychopathology: Perspectives on Adjustment, Risk, and Disorder. Cambridge University Press, New York, pp. 317–349.

Dwairy, M., Achoui, M., Abouserie, R., Farah, A., 2006. Adolescent-family connectedness among Arabs: a second cross-regional research study. Journal of Cross-cultural Psychology 37, 248–261.

Iarocci, G., Root, R., Burack, J., 2009. Social competence and mental health among Aboriginal youth: an integrative developmental perspective. In: Kirmayer, L.J.,

Valaskakis, G.G. (Eds.), Healing Traditions: the Mental Health of Aboriginal Peoples in Canada. UBC Press, Vancouver, Canada, pp. 80–108.

Iarocci, G., Yager, J., Elfers, T., 2007. What gene–environment interactions can tell us about social competence in typical and atypical populations. Brain and Cognition 65, 112–127.

Larson, R., Wilson, S., 2004. Adolescence across place and time: globalization and the changing pathways to adulthood. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. John Wiley & Sons, Hoboken, NJ, pp. 299–330.

Lee, C., Beckert, T.E., Goodrich, T.R., 2010. The relationship between individualistic, collectivistic, and transitional cultural value orientations and adolescents' autonomy and identity status. Journal of Youth and Adolescence 39, 882–893.

Martin, J., Sokol, B.W., Elfers, T., 2008. Taking and coordinating perspectives: from prereflective interactivity, through reflective intersubjectivity, to metareflective sociality. Human Development 51, 294–317.

Monahan, K.C., Steinberg, L., 2011. Accentuation of individual differences in social competence during transition to adolescence. Journal of Research on Adolescence 21, 576–585.

Obradović, J., van Dulmen, M.H.M., Yates, T.M., Carlson, E.A., Egeland, B., 2006. Developmental assessment of competence from early childhood to middle adolescence. Journal of Adolescence 29, 857–889.

Roisman, G.I., Booth-LaForce, C., Cauffman, E., Spieker, S., 2009. The developmental significance of adolescent romantic relationships: parent and peer predictors of engagement and quality at age 15. Journal of Youth and Adolescence 38, 1294–1303.

Sagi-Schwartz, A., Aviezer, O., 2005. Correlates of attachment to multiple caregivers in Kibbutz children from birth to emerging adulthood: the Haifa longitudinal study. In: Grossman, K.E., Grossman, K., Waters, E. (Eds.), Attachment from Infancy to Adulthood: the Major Longitudinal Studies. Guildford Press, New York, pp. 165–197.

Simpson, J.A., Collins, W.A., Tran, S., Haydon, K.C., 2007. Attachment and the experience and expression of emotions in romantic relationships: a developmental perspective. Journal of Personality and Social Psychology 92, 355–367.

Steinberg, L., 2001. We know some things: parent-adolescent relationships in retrospect and prospect. Journal of Research on Adolescence 11, 1–19.

Talbani, A., Hasanali, P., 2000. Adolescent females between tradition and modernity: gender role socialization in South Asian immigrant culture. Journal of Adolescence 23, 615–627.

Social Constructionism

Viv Burr, School of Human and Health Sciences, University of Huddersfield, Huddersfield, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social constructionism is a term used to refer to a number of related theoretical approaches that challenge the underlying assumptions and research practices of mainstream psychology. Its insistence upon the historical and cultural specificity of all knowledge means that it constitutes a radical critique of the discipline. Its focus on the importance of language in the construction of people and events has led to the adoption of several varieties of discourse analysis as its research methods of choice.

Introduction: the Origins of Social Constructionism

Social constructionism is a relatively new term in the social sciences, especially psychology. But the ideas and practices that have cohered around this term in recent decades have a longer history in disciplines such as sociology, philosophy, and linguistics. The key tenet of social constructionism is that our knowledge of the world, including our understanding of human beings, is a product of human thought rather than grounded in an observable, external reality. Although very different in other respects, the philosophers Kant, Nietzsche, and Marx all took this view; important concepts in the sociology of knowledge developed during the twentieth century, such as ideology and false consciousness, reflect this key tenet, focusing on how sociocultural forces construct knowledge and on the kind of knowledge they construct.

But an important contributor to social constructionist thinking has been the microsociological approach of Symbolic Interactionism, cogently laid out in the now classic book *The Social Construction of Reality* (Berger and Luckmann, 1966). In the early twentieth century, George Mead, at the University of Chicago, developed this approach from the earlier work of Herbert Blumer and later published his ideas in *Mind, Self and Society* (1934). Mead had studied at Leipzig University under Wilhelm Wundt, who is widely acknowledged as the founder of experimental psychology. However, Wundt is less well known for his belief in the importance of sociocultural factors, such as myth, folk customs, and religion, in understanding human behavior and experience. Mead took up these ideas in his development of Symbolic Interactionism, which argues that people construct and negotiate identities for themselves and others through their everyday social interactions with each other. Language, as a system of socially shared symbolic meanings, is central to this constructive process. Ironically, John Watson, later to become one of the key expounders of behaviorist psychology, studied under Mead as a postgraduate student. Watson's rejection of the concept of 'mind' and his focus on decontextualized behavior was antithetical to Mead's own position, and their relationship deteriorated. When a new psychology department was set up at Chicago, the split between Watson and Mead was mirrored in the subsequent disciplinary separation of psychology from sociology. This division persists today and consequently few undergraduate psychology students have the opportunity to learn of approaches such as symbolic interactionism and the mainstream discipline continues to operate within a socially decontextualized model of human behavior and experience. In addition to these North American influences, social constructionism has also drawn on the ideas of more recent European thinkers. The work of the historian and philosopher Michel Foucault has been highly influential, as has that of Ferdinand de Saussure in structural linguistics, of the philosopher Jacques Derrida's work on 'deconstruction' and of the psychodynamic theorist Jacques Lacan. The ideas of these writers have been extensively drawn upon in developing social constructionist thinking around the role of language in the construction of human social phenomena.

The influx of some of these ideas into psychology was encouraged in the later half of the twentieth century by the increasing dissatisfaction with and unease about their discipline of a number of social psychologists. What came to be known as the 'crisis in social psychology' (e.g., Harré and Secord, 1972; Armistead, 1974) centered on mounting worries about the way that social psychology's agenda was driven by the needs and motivations of powerful factions (government, the military, and commerce); its theories and research findings often seemed to bring further oppression to relatively powerless and marginalized groups (women, ethnic minorities, working class people, and those of non-normative sexual orientation) and this operated partly through the study of human phenomena in socially decontextualized laboratory environments, since experiments ignored the real-world contexts which, it was argued, give human conduct its meaning. There was a move to attend to this social context, as well as to explore human phenomena from the perspective of psychology's 'subjects' themselves rather than to overwrite this with the perspective and voice of the relatively powerful researcher. These concerns encouraged social psychologists to embrace the ideas already flourishing in neighboring disciplines, including microsociology, and also fed into the increasing call to recognize qualitative research methods as legitimate and fruitful for the discipline of psychology.

Epistemology

Social constructionism poses a radical challenge to the epistemological position assumed by mainstream psychology. Psychology has followed the natural sciences in its vision of scientific endeavor, and this vision of science was born in the Enlightenment historical period.

The Enlightenment was characterized by the search for objective truth through reason and rationality. Science was to challenge the dogma and superstition of the medieval period when knowledge was handed down by religious leaders and was to be unquestioningly accepted. The Enlightenment challenged the church's monopoly on truth and proposed that individual people (increasing numbers of whom, with the invention of the printing press, could read the scriptures and scientific publications for themselves) should be responsible for developing their own knowledge, and discover the truth for themselves; individuals should draw their own conclusions about the nature of the world, and how to live in it, based on objective, scientific evidence. This vision of scientific endeavor, focusing on the systematic observation and recording of phenomena, became known as positivism and is closely allied with realism (sometimes referred to as naive realism), the idea that through our observations of the world, objective reality will be revealed to us. The philosopher Auguste Compte (1798–1857) was a central contributor to the idea that science could discover 'laws' of human nature just as it was doing for the natural world. It is important to appreciate the role of positivist thinking in challenging the power of religious dogma and superstition. Nevertheless, the capacity of science to capture the true nature of the world through empirical investigation is what social constructionism, in its turn, now challenges.

Social constructionism problematizes the claim that what we call knowledge, our current understandings of the nature of the world and its phenomena, is derived from objective, unbiased observation of events; that what exists is what we perceive to exist. Neither does it argue that the problem lies in systematic biases operating within scientific method. The concept of bias in turn rests upon the concepts of truth and accuracy, both of which social constructionism also challenges. The solution is not, therefore, to improve scientific method in order to reduce its biases; truth and accuracy are rejected as meaningful conceptual tools because social constructionism's epistemological position is one of relativism. Sometimes referred to as perspectivism, this position argues that there can never be one objective, final and 'true' account of phenomena; instead what exists are multiple perspectives, potentially as many different versions of events and things as there are people in the world. What we call knowledge, which we regard as facts, can therefore not be compared to an accurate standard to assess its truthfulness; instead, different ways of understanding the world coexist in parallel and none of them can be said to be the truth. Rather than pursuing a search for objective truth and the nature of the real world, social constructionism recommends that we take a critical and skeptical attitude toward our taken-for-granted ways of understanding the world.

Time and Place: Knowledge in Context

One of the earliest and most influential social constructionist publications within psychology came from the North American psychologist Kenneth Gergen in 1973. His paper *Social Psychology as History* argues that all knowledge, including psychological knowledge, is the product of a particular historical period and of a particular cultural location. He reframes social psychology as a historical, rather than scientific, project. The constructionist case is that all human psychological and social phenomena arise out of social life, from the interactions between people. These interactions in turns are given structure and content by the culture in which we live, by our society's economic conditions, and by the power relations in which we are embedded. But these social conditions are subject to constant flux. From time to time, and from location to location, the conditions within which people construct themselves and each other differ and are transformed. It follows, therefore, that there can never be definitive answers to questions about the nature of human and social phenomena. Instead, Gergen recommends that we ask how and why certain constructions come into being. This of course includes the constructions that we call 'theories,' and so the very projects of psychology and social psychology themselves become questions about history and culture. To properly understand the evolution of present-day psychology and social life, we therefore must extend our enquiries beyond the individual into social, political, and economic realms. Social constructionism also challenges the very categories of things and people that characterize our (current, local) thinking and language. We cannot imagine dispensing with categories and dichotomies such as urban/rural, male/female, individual/society, or mental/physical, but social constructionism proposes that these are human constructions rather than objective descriptions of the world and at least some of them would not have made sense to people living in earlier historical times.

Such a view disrupts the notion of scientific progress, the idea that through science we are advancing toward a more and more accurate understanding of the physical and psychological world. Our contemporary conceptions of, say, children, illness, or sexuality are certainly different to what they were 100 or 200 years ago, and may differ radically from ways of thinking in non-western, industrialized cultures. But social constructionism cautions against making the assumption that our current, local ways of thinking are better because they represent a more accurate or truthful account. Such thinking, it is argued, is responsible for the imposition of our own ways of thinking upon other cultures and nations; psychology has been imperialist in its view of other cultures and has colonized them, replacing their indigenous conceptions of human life with western-grown ideas. The model of human beings intrinsic to contemporary psychology is a particularly individualistic one, a model which celebrates and privileges the unique, self-contained person. Social constructionists argue that this individualism became part of the discipline of psychology as it developed and flourished in North America, where the individual is arguably especially valorized (Farr, 1996).

Challenging Psychology's Model of the Person

Antiessentialism

The vision of the person lying at the heart of mainstream psychology is the self-contained individual, an autonomous and agentic person that preexists society and social life. This vision is expressed in Floyd H. Allport's (1924: 12) definition of social psychology: "The science which studies the behavior of the individual in so far as his [sic] behavior stimulates other

individuals, or is itself a reaction to this behavior." This pre-existing individual has a discoverable nature, and the agenda of psychology is its discovery. The content of this individual is described by the various and competing psychological theories that have been developed over the last 100 years or so: personality traits and intelligence, unconscious motivations and drives, learned behaviors and habits, attitudes, and beliefs. All these constructs are similar; however, it that they are thought of as properties of the person, whether inherent or acquired. Social constructionism regards all such concepts as constructions, and like any other constructions, they have arisen in the course of social life in particular sociocultural conditions. This position is referred to as antiessentialism, and stands as a critique of the notion of psychological essences, structures, or qualities that exist at an intrapsychic level as properties of individuals. Social constructionism also therefore challenges the determinism of psychology, its assumption of causality. This is the view that psychological and social phenomena are caused and explained by events or structures either in the individual's environment (such as upbringing or traumatic events) or internal to them (such as personality traits or the actions of hormones).

Rejecting Reductionism

The essentialism and determinism of mainstream psychology are also expressed in its reductionist thinking. Reductionism is the attempt to explain complex events in terms of simpler, lower level ones. Levels of explanation are typically arranged from the biological (genetic material, brain activity, hormonal activity, etc.) through the psychological (behavior and experience) to the social/societal (social structures and social phenomena). In this way, complex social phenomena, such as poverty, crime, and alcoholism, are reduced to and explained by events or structures at the psychological level (such as intelligence and personality) which are in turn reduced to and explained at the biological level, often genetic inheritance or brain structures.

This reductionism is not confined to the discipline of psychology but is increasingly evident in popular understandings of a great range of social and psychological phenomena. The burgeoning discipline of neuroscience has taken centre-stage in popular understandings of human experience, so that we are invited to understand psychological events as caused by events happening at a neurological level. Gergen (2010) is critical of this, arguing that it leads to fatalistic thinking; if behavior is seen as caused by neurological events, over which we have no control, personal agency is an illusion and we may simply blame our wayward brains for the acts we commit. Gergen offers a constructionist challenge to such thinking, arguing that the psychological behaviors, events, or properties that neuroscience purports to explain, for example, aggression are complex, changeable social constructions subject to local interpretations. It is therefore not possible to identify causes of such phenomena, whether neurological or not.

The Centrality of Social Interaction and Discourse

Social constructionism turns on its head the relationship between the individual and society implicit in Allport's definition of social psychology; what we recognize as individual persons cannot exist without a preexistent social network. Human beings are born into a world of social relations, language, norms, and customs and it is this social world that constructs them as recognizable persons who are capable of meaningful conduct. In this sense, it would be legitimate to say that all psychology is social psychology; there is no part of an individual's functioning that does not have its origins in social life in one way or another. As in symbolic interactionism, the very contents of our thoughts, what we call our attitudes and beliefs, our sense of self, all these things arise out of interaction with others. The constructionist John Shotter (1995) writes about 'joint action' as a way of conceptualizing the social origins of thought and behavior. Kenneth Gergen (1999), drawing on the work of Lev Vygotsky and the philosophy of Mikhail Bakhtin, argues for a relational, dialogical approach to the self. He writes of the interconnectedness between people; we are locations in a network of myriad 'voices,' past and present. Gergen has since developed these ideas into a political vision of the self, deconstructing and challenging the self/other dichotomy upon which much conflict arguably rests. The constructive work that produces us as persons takes place in social interaction of all kinds, and language as a key constituent of social interaction is therefore of great importance. Discourse is at the heart of social constructionism, and this term has been used in different ways by two approaches that have developed within social constructionism, which can broadly be called the 'macro' and the 'micro' approach (Burr, 2003).

The macro approach draws heavily on the work of Michel Foucault, who argued that discourses are widely prevalent ways of talking about or otherwise representing (such as in images, written materials, films and even buildings and other physical spaces) people or things and these discourses function to construct and maintain our understanding of those people and things in social life. Through his method of the 'archaeology of knowledge,' he traces the discourses historically producing contemporary constructions of phenomena such as mental illness (Foucault, 1964) and sexuality (Foucault, 1976). Our talk about such things is reframed as manifestations of these discourses, almost as if they were speaking through us. In its most extreme form, this view renders people the hapless carriers of discourses who are also constructed by them in a rather deterministic fashion; our identities become multiple and changeable, distributed across the discourses that construct us. However, taking this approach brings to fore the issue of power, something which psychology has arguably neglected. Prevailing discourses are said to be constructed in the interests of the relatively powerful (in general, white, middle-class men), it is relatively powerful in society that has the greatest opportunity to disseminate such discourses and it is their status (for example, as members of the medical profession, the law, or the state) that authorizes and legitimates them. This approach is therefore able to offer an understanding of how psychology as a discipline has contributed to the oppression and marginalization of some people.

The micro approach is associated principally with work that came to be known as discursive psychology (e.g., Edwards and Potter, 1992). Discourse here is taken to refer to our everyday linguistic practices, particularly spoken interactions. The focus here is not on the constructive power of prevailing discourses, but on the constructive work that people do in interactions in

order to build accounts of themselves and events that are effective for them. The key ideas here were laid out in the now classic publication by Jonathan Potter and Margaret Wetherell (1987). Using their own interview research data, which examined racism in New Zealand, they reject the mainstream social psychological concept of attitudes as an explanatory tool. Focusing on the variability within participants' accounts, they propose instead that interactants' talk is occasioned by the moment-to-moment needs of each party in the conversation; these needs might include, for example, justifying one's actions, attributing blame to another for the outcome of events or creating a good impression.

One danger here is that these two approaches threaten to recreate the mainstream division between the individual and society, with the macro approach being socially deterministic and the micro approach seemingly reinstating the agentic individual. However, the macro and micro approaches are not in themselves incompatible and some have been keen to explore how they might be synthesized. One such attempt is in the concept of 'positioning,' put forward by the philosopher of science Rom Harré (Harré and Van Langenhove, 1999). Here, persons are simultaneously positioned by discourses (the macro approach) and draw on these to position themselves and others within specific interactions (the micro approach).

Research

Social constructionism challenges mainstream psychology's approach to research, critiquing the laboratory experiment, in particular, for its socially decontextualized approach to understanding the person. As social constructionism has developed, various theoretical and methodological preferences have emerged within it, and there are now a number of approaches to the analysis of research data that can broadly be called 'discourse analysis'. The macro and micro approaches outlined above are characterized by their own research questions and associated methodologies, although the latter are predominantly qualitative. Those who wish to focus on the constitutive force of discourse and analyze the power relations implicated in subject positions have favored a Foucauldian approach, and Fairclough's (1995) method of Critical Discourse Analysis has gained popularity for such purposes. This kind of approach has been enthusiastically taken up by some feminists and others who wish to conduct research which exposes the oppressive use of discourse and address the marginalization of relatively powerless people.

Research focusing at the more micro level of social interaction has cohered around the practice of what has come to be known as 'discursive psychology'. The research focus here is everyday language as it is used by ordinary people in the course of social interaction. Drawing on the sociological subdiscipline of ethnomethodology, its focus is upon identifying the processes by which ordinary people construct and build accounts of themselves and events. Early work from this orientation used the conceptual tool of the 'interpretative repertoire,' with research aiming to identify the repertoires used by people when accounting for events, often in research or other interviews. Later developments in discursive psychology have led to a greater focus on identifying the use of a range of rhetorical devices (such as justifications and blamings, three-part lists, and extreme-case formulations) used by interactants to construct their accounts in naturally occurring conversations. To this end, discursive psychologists have typically adopted a version of Conversation Analysis as their preferred research method.

Issues and Debates

Realism vs Relativism

The relativist epistemology that appears to be a key feature of social constructionist thinking is regarded by many as deeply problematic has, in the past, been hotly debated (see the special edition of *Theory and Psychology*, 11 (3), published in 2001). Relativism appears to undercut political efforts to challenge oppressive practices and to take a moral stance since it is impossible to demonstrate the 'truth' of one version of events, the superiority of a particular moral code, over any other; one view is as valid as any other, and 'anything goes'. The difficulty appears to be the impossibility of specifying a reality that exists beyond language. This has led to the development of 'critical realism'. Originating with the philosopher Roy Bhaskar, critical realism argues that social systems are real; they are constructed by people and have real causes and constraints that are external to individuals. Oppressive systems and structures may therefore be critiqued as they have an objective existence beyond discourse. Such a position has been important for those researchers who are keen to analyze power relations, especially in relation to marginalized experiences or groups. Indeed, discursive psychology has been criticized for its failure to contextualize social interaction within broader power structures, while discursive psychologists argue that a Foucauldian approach neglects actual everyday language as used by real people.

However, discursive psychologists claim that social constructionism's relativism has been misunderstood. Edwards et al. (1995) argue that social constructionism does not deny the existence of a real, material reality but that this reality is always brought into social being through language and is transformed in this process. Others disagree that relativism is in itself morally problematic. For example, Gergen (2001) believes that the strength of relativism is that it constantly demands argument and debate. Moral and political positions must be argued for and discussed, and this process is the best way of guarding against any one version of truth becoming legitimated and generally imposed. The heat in this debate now appears to have dissipated, with writers generally adopting and justifying one or other of the above positions or their variants.

Identity and Subjectivity: the Relationship between the Individual and the Social

Early enthusiasm for the Foucauldian approach to discourse within social psychology subsequently raised concerns about the loss of anything we might recognize as the psychological content of the person, and about the danger of replacing the determinism of mainstream psychology with a determinism that is simply social/discursive instead. And despite discursive psychology's focus on the individual use of language in interaction, which focuses on the agentic use of language by skilled social actors, the interior life of the person was effectively

'bracketed off,' much as it had earlier been for behaviorism. Another concern was the absence of any embodied presence of the person within social constructionist thinking. If social constructionists were to understand how people construct meaningful identities for themselves and are able to challenge and resist dominant discourses, the person as a psychological being needed to be reinstated, but in a radically different form to the self-contained, liberal-humanist conception of the individual in the mainstream discipline. An early attempt to do this is represented in the edited collection *Changing the Subject* (Henriques et al., 1984), with contributors drawing on (psychoanalytic) Lacanian and object-relations theory to make good the absence of psychological processes from the socially constructed person. Within social constructionist writing, psychology's key traditional concept of personality became replaced by those of identity and subjectivity, to signal the strong social and discursive origins and situatedness of personhood. However, reinstating interior, psychological life as a valid focus of enquiry was not enthusiastically taken up by social constructionists across the board. Now sensitized to the political implications of psychology's historical individualism, many feared backslide into such ideas and were deeply suspicious of any reference to the self.

Nevertheless, psychologists working within a social constructionist framework have made continued efforts to theorize and remodel psychological subjectivity. The concept of 'subject positions' within discourse has become particularly widely used. Davies and Harré (1990) propose that when people take up positions within discourses (for example, as women or men within discourses of gender, or as gay or straight within discourses of sexuality), these positions inevitably provide us with a particular perspective on the world, with affordances for what can meaningfully be said and done from that subject position. The term 'subjectivity,' Frosh (2003) argues, signifies both being subject to (discourse) and being a (psychological, agentic) subject who acts. This concept has become a key focus of a relatively new approach which has emerged over the last 10–15 years, referred to as psychosocial studies. Following a critical agenda, much work within psychosocial studies draws heavily on psychoanalytic theory and is concerned to address the problematic dichotomy of individual/society in a new understanding of subjectivity. Some see the aim of psychosocial studies as transcending this dichotomy such that its poles "are instead thought of together, as intimately connected or possibly even the same thing" (Frosh, 2003), echoing earlier attempts to provide an alternative to the agency/structure debate within sociology (e.g., Giddens, 1979). However, others (e.g., Jefferson, 2008) argue that the 'psychic' is lost in such a conceptualization. In this context, the terms subjectivity and identity are themselves contentious within psychosocial studies. Originally intended as a new way of conceptualizing the relationship between the psychic and the social (Venn, 2006), some (e.g., Wetherell, 2008) are concerned that the psychic realm, under the label of subjectivity, once more becomes one pole of the old individual-social dichotomy with identity occupying the social pole, representing a backslide into dualistic thinking. Wetherell argues that what we regard as features of both social identity and subjectivity are themselves socially constructed, and recommends 'psycho-discursive practices' as a more fitting conceptual tool.

Social Constructionism in Psychology

In the foreward to their (1998) reissue of *Changing the Subject*, the authors comment that since the time of the first edition, there had been little take-up of social constructionism's challenging ideas by the mainstream discipline. This situation seems to have continued, at least outside of mainland Europe, with academics researching and writing principally within one realm or the other. New academic journals have been developed which welcome social constructionist and discursive work, while many preexisting mainstream journals continue to favor work carried out within a traditional, positivist framework. Social constructionist work within psychology can now be said to be broadly subsumed under the umbrella term of 'critical psychology' (Fox and Prilleltensky, 1997), but there has been relatively little productive dialogue between mainstream psychology and social constructionism/critical psychology, the two paradigms, and their agendas, continuing to lead more or less parallel, self-contained lives. Frosh (2003) notes that the more recently emerging field of psychosocial studies has in the main developed outside of academic psychology, and there has been little dialogue between these.

See also: Critical Psychology; Essentialism; Levels of Analysis in Social Psychology; Social Identity in Social Psychology; Social Psychology: Research Methods; Social Psychology.

Bibliography

Allport, F.H., 1924. Social Psychology. Houghton Mifflin Company, Boston.
Armistead, N., 1974. Reconstructing Social Psychology. Penguin, Harmondsworth.
Berger, P., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Burr, V., 2003. Social Constructionism. Routledge, London.
Davies, B., Harré, R., 1990. Positioning: the discursive production of selves. Journal for the Theory of Social Behaviour 20 (1), 43–63. Reproduced in Wetherell, M., Taylor, S., Yates, S.J. (Eds.), 2001. Discourse Theory and Practice: A reader. Sage, London.
Edwards, D., Potter, J., 1992. Discursive Psychology. Sage, London.
Edwards, D., Ashmore, M., Potter, J., 1995. Death and furniture: the rhetoric, politics and theology of bottom line arguments against relativism. History of the Human Sciences 8, 25–49.
Fairclough, N., 1995. Critical Discourse Analysis: The Critical Study of Language. Longman, London.
Farr, R.M., 1996. The Roots of Modern Social Psychology, 1872–1954. Blackwell Publishing, Malden.
Fox, D., Prilleltensky, I. (Eds.), 1997. Critical Psychology: An Introduction. Sage, London.
Foucault, M., 1964. Madness and Civilization: A History of Insanity in the Age of Reason. Vintage Books, New York.
Foucault, M., 1976. The History of Sexuality: An Introduction. Penguin, Harmondsworth.
Frosh, 2003. Psychosocial studies and psychology: is a critical approach emerging? Human Relations 56 (12), 1545–1567.
Gergen, K.J., 1973. Social psychology as history. Journal of Personality and Social Psychology 26, 309–320.
Gergen, K.J., 1999. An Invitation to Social Construction. Sage, London.
Gergen, K.J., 2001. Construction in contention: toward consequential resolutions. Theory and Psychology 11 (3), 433–441.
Gergen, K.J., 2010. The acculturated brain. Theory and Psychology 20 (6), 795–816.
Giddens, A., 1979. Central problems in Social Theory: Action, Structure, and Contradiction in Social Analysis. University of California Press, Los Angeles, CA.

Harré, R., Van Langenhove, L. (Eds.), 1999. Positioning Theory. Blackwell, Oxford.
Harré, R., Secord, P.F., 1972. The Explanation of Social Behavior. Basil Blackwell, Oxford.
Henriques, J., Hollway, W., Urwin, C., Venn, C., Walkerdine, V. (Eds.), 1984. Changing the Subject: Psychology, Social Regulation and Subjectivity. Methuen, London. Re-issued in 1998 by Routledge.
Jefferson, T., 2008. What is "the psychosocial"? A response to Frosh and Baraitser. Psychoanalysis, Culture and Society 13, 366–373.
Mead, G.H., 1934. Mind, Self and Society. University of Chicago Press, Chicago, IL.
Potter, J., Wetherell, M., 1987. Discourse and Social Psychology: Beyond Attitudes and Behavior. Sage, London.
Shotter, J., 1995. In conversation: joint action, shared intentionality and ethics. Theory and Psychology 5 (1), 49–73.
Venn, C., 2006. The Postcolonial Challenge: Towards Alternative Worlds. Sage, London.
Wetherell, M., 2008. Subjectivity or psycho-discursive practices? Investigating complex intersectional identities. Subjectivity 22, 73–81.

Relevant Website

https://www.brocku.ca/MeadProject/inventory5.html – 'The Mead Project'.

Social Constructivism

Wolfgang Detel, University of Frankfurt/Main, Frankfurt/Main, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The introduction of this article presents a sketch of the three main current versions of social constructivism (social epistemology, strong social constructivism, and social ontology) and of the historical background of social constructivism (Section Social Constructivism – Variants and Background). The following three sections outline these three approaches in some more detail and assess them briefly. Concerning social epistemology, veritistic and nonveritistic approaches to social epistemology are discussed (Sections Social Epistemology (Weak Social Constructivism), The Veritistic Approach to Social Epistemology, and Nonveritistic Approaches to Social Epistemology). Concerning strong social epistemology, views about ontology and the social are addressed (Sections Radical (Strong) Social Constructivism, Strong Social Constructivism and Ontology, and Strong Social Constructivism and the Social). And concerning social ontology, first a currently much debated specific example (social ontology in theories on international relations) and then the general approach to social ontology are characterized (Section Constructivism about the Social (Social Ontology), Constructivism and International Relations, and Nonveritistic Approaches to Social Epistemology).

Social Constructivism – Variants and Background

Facets of Social Constructivism

"Social construction is … an idea that has been cropping up all over the place … and is bitterly fought over in the … culture wars" (Hacking, 2000; VII). The core idea of social constructivism in its broadest sense is that some things are produced (and in this way constructed) by social actions, i.e., by actions that we carry out by interacting with other people. There are certainly many things that are socially constructed in this sense, for instance many artifacts, some emotions, and a number of novels. Social constructionists argue, though, that there are a lot of things that seem to depend, at first sight, exclusively on nature and natural laws such that they could not be changed by human actions but that, on closer examination, turn out to be socially constructed and thus alterable by humans. In these cases, social constructivism is taken by its proponents to be a liberating idea being critical of the status quo. It is in this more special version, not in its broadest sense, that the idea of social construction is bitterly fought over. Even in this special sense social constructivism comes in many different variants. It can be concerned with different kinds of things, notably with (1) beliefs, (2) facts or (3) social relations. And it can imply differently strong commitments as to the social construction of X, in particular that (a) X is not inevitable (against first sight), (b) that X is, in addition to (a), a bad thing, and (c) that X is, in addition to (a) and (b), a thing we would be better of if it were at least considerably be transformed, if not destroyed.

The most fundamental and interesting battles over social constructivism are debates about case (a), since if (a) is false, then (b) and (c) are irrelevant. At the same time, (a) does not imply (b), nor does (b) imply (c). Concerning beliefs, social constructivism is a view about the social nature of knowledge and science (sometimes labeled social epistemology or weak social constructivism). Concerning facts, social constructivism is an idealistic view about the social construction of scientific facts and sometimes even about empirical facts (sometimes labeled radical or strong (social) constructivism). Finally, concerning social relations, social constructivism is a view about the construction of the social realm (sometimes labeled social ontology). Social epistemology and radical constructivism take social relations and social actions as given, using them as explanantia. In contrast, social ontology takes social relations and social actions as explananda and is therefore, from a theoretical point of view, the most fundamental version of social constructivism. In this entry, the discussion of social constructivism will be restricted to case (a) concerning beliefs, facts, and social relations.

Historical Background of Social Constructivism

Historically, social constructivism is rooted in historical materialism developed by Marx. Historical materialism is best understood as a research program, roughly in the sense of the methodology of research programmes proposed by Lakatos, the core of historical materialism consisting of the claim that the economic base of a society determines its superstructure. This core is supposed to generate testable historical theories or hypotheses about what, in various societies, the base and the superstructure are, and about the specific ways in which they are interrelated. The strongest, but not the only possible, interpretation is that the superstructure is causally determined by the base. This approach was taken up by sociologies of science worked out during the twentieth century. Thus, for instance, Karl Mannheim wanted to show how humanities and theories of society are socially determined (Mannheim, 1936); and one of the leading American sociologists of science, Robert Merton, suggested that it is primarily the rise and success of science that is determined by social forces (Merton, 1973) (*see* Social Science, the Idea of).

Another important intellectual source of social constructivism is the historical turn of the twentieth century philosophy of science (Kuhn, 1962; Bachelard, 1934; Canguilhem, 1966;

Foucault, 1969; Hacking, 1990). This movement is vigorously opposed to scientific realism and rejects the claim, so crucial for the traditional philosophy of science, that there are universal rational scientific methods and rules such that the history of science can be rationally reconstructed as an continuous intellectual enterprise constantly improving our scientific knowledge by applying these methods. Instead, proponents of the movement see the history of science and scientific beliefs shattered by contingent breaks and dependent on normative attitudes and hidden nonscientific assumptions that an adequate philosophy of science must aspire to reveal.

A third theoretical background of social constructivism is the program of naturalizing epistemology initiated by Quine (Quine, 1969). The crucial idea of this program is that traditional epistemology centering around the question of the justification of claims to (scientific) knowledge should be replaced by a scientific study of causal processes of belief formation and information. In this way, epistemology is supposed to be transformed into a natural science.

In sum, there are three historical ideas that guide approaches to social constructivism, borrowed from a broadly Marxist sociology of science, from the historical turn of twentieth century philosophy of science, and from the program of naturalizing epistemology, namely that the development of scientific knowledge is (1) determined by social forces, (2) essentially contingent and independent of rational methods, and (3) should be analyzed in terms of causal processes of belief formation.

This is not to say that all versions of social constructivism are guided by all three ideas. On the contrary, there are veritistic approaches relying on the claim (1), but rejecting claims (2) and (3), and nonveritistic approaches entertaining all three claims (1)–(3).

Social Epistemology (Weak Social Constructivism)

Social epistemology rests on the methodological assumption that a sociological analysis of science and scientific knowledge can be empirically fruitful and epistemologically illuminating. This assumption has generated detailed empirical studies of scientific practices (for instance, of what is going on in laboratories on a day-to-day basis). However, there are two sorts of social epistemology: veritistic and nonveritistic social epistemology.

The Veritistic Approach to Social Epistemology

The veritistic approach to social epistemology seeks to identify, and to evaluate, social interactions, practices, and forces that influence – in a positive or negative way – the pursuit of the fundamental aim of traditional science (spelled out in classical philosophy of science), i.e., the production of true justified beliefs (Goldman, 1978, 1987, 1999). On this view, science and education place a high value on having true beliefs rather than false beliefs ('veritistic value') (*see* Truth and Credibility in Science). Social interactions, practices, and forces that influence – in a positive or negative way – the pursuit of truth include communicational acts, institutional structures that frame such acts (like universities, journals publishing peer reviewed articles, or conferences), speech practices of reporting and arguing, ways of access to different sorts of information (including internet resources (Goldman, 2006)), market mechanisms regulating the flow of speech, information technologies (Goldman and Cox, 1996), legal systems, and ways of disseminating information about electoral candidates.

One of the most important social aspects of science is the division of cognitive labor (Kitcher, 1990; Thagard, 1997). For instance, it will optimize the scientific pursuit of truth if different members of a scientific community pursue different strategies of solving a scientific problem, rather than all work on the most promising strategy. From this perspective, science can only make progress if it relies on a social consensus practice consisting, in its core sense, of individual practices common to all members of a scientific community. In a broader sense, the consensus practices include the exploitation of reliable scientific authorities, formal and empirical methodologies and background assumptions (Kitcher, 1993).

Another way of addressing the social aspects of science in a veritistic mood is to develop theories of cultures of knowledge. The starting point of such theories is the idea that cultures of knowledge have to be distinguished from theories and contents of knowledge. Classical theories of knowledge tend to provide definitions of knowledge articulating ideals of knowledge and the pursuit of truth. In contrast, cultures of knowledge are constituted by epistemic practices. Epistemic practices are collections of practices aiming at producing, selecting, organizing, and transmitting knowledge (true justified beliefs). They are a kind of culture if a culture is taken, in general, as a set of practices which are pursued by a group of people, which are taught and learned, handed over, institutionalized within this group, and whose exercise leads to a certain social status, e.g., to membership in a social community. To investigate cultures of knowledge, then, means to analyze the epistemic practices of a group of people and to explore the social realization of these practices. Asking for cultures of knowledge is an attempt not to ask simply for the social conditions in which theories did rise and fall, as traditional social epistemology did, but to look at the complex of epistemic practices, theories and ideals of knowledge, and their social conditions and realizations. The point is not to treat definitions of knowledge as mere ideological epiphenomena, but to relate them to epistemic practices and to relate this complex to its social environment (Detel, 2007; Fried, 2009). Contrary to what one might think at first sight (Longino, 1990, 2002), theories about epistemic practices and cultures of (scientific) knowledge and about the historical changes of such epistemic cultures are compatible with assumptions about a universal scientific rationality (Detel, 2009).

Basically, there is nothing wrong with veritistic approaches to social epistemology. This is not to say, though, that there are no open problems with this view. For example, one of the topics that need to be addressed is in which way the pursuit, not of true beliefs, but of justified beliefs can be socially

determined. In this context, pointing to reliable testimony has been, and must be, discussed (Fricker, 1995, 1998; Goldman, 2001; Lackey, 2006; Pettit, 2006). In addition, the historical change of epistemic practices and cultures of knowledge seems to involve a relativism of rationality and justification (Longino, 1990, 2002; but see Detel, 2009).

Nonveritistic Approaches to Social Epistemology

According to the nonveritistic approach to social epistemology, empirical, historical, and sociological studies of scientific practices show that science and scientific results do not depend exclusively on the objective external world which make scientific beliefs true or false, but rather also (or even mainly or exclusively) on social arrangements resulting from negotiations between scientists taking place in the course of scientific practices. The guiding idea is that science, scientific knowledge, and scientific practices are socially determined. It is in this sense that scientific knowledge is supposed to be socially constructed.

Social constructionists in the nonveritistic mood maintain three fundamental theses (Hacking, 2000):

1. Contingency thesis: Natural sciences could have developed in alternative ways. The way natural sciences have in fact developed is not inevitable, but was to a certain extent contingent (for an important historical case see Shapin and Schaffer, 1985).
2. Nominalist thesis: At most some individual things, but not facts or universals, can be taken to exist independently of humans and their social actions.
3. Stability thesis: Explanations for the stability of scientific beliefs must rely, at least in part (sometimes even exclusively) on social factors that are external to the internal criteria of theory choice maintained in modern natural sciences and articulated in rationalist philosophies of science.

Social constructivism is not a unique specified doctrine, however, but rather a bunch of related studies representing different versions of the general nonveritistic approach.

The earliest prominent version is the Edinburgh school of the sociology of science (sometimes called the strong program of the sociology of science) (*see* Strong Program, Sociology of). The proponents of this school start out by criticizing the Marxist sociology of science for the exclusion of mathematics and natural sciences as a proper subject of the sociology of science. The extended claim is that mathematics and natural sciences are socially determined too. Specifically, it is, according to the Edinburgh sociologists of science, not only the historical development of science, its rise and success, that is influenced by social forces; rather, it is the content of accepted scientific beliefs that is determined by social factors or social interests being involved in scientific practices. There are four methodological principle guiding this sort of social constructivism: (1) causality, i.e., exploring the conditions that causally bring about claims to a certain kind of knowledge. (2) impartiality, i.e., examining successful as well as unsuccessful knowledge claims. (3) symmetry, i.e., using the same types of explanations for successful and unsuccessful knowledge claims alike. (4) reflexivity, i.e., applying the theory to itself (Bloor, 1976/1992; Knorr-Cetina, 1981, 1982; Latour and Woolgar, 1986).

From their sociological studies the Edinburgh sociologists of science conclude that there is no definite unique set of rational methods that guide scientific practices and can be referred to in order to explain how scientific results and beliefs are established. Another result is that a sharp distinction between the context of discovery and the context of justification is not helpful in sociological investigations of scientific practices. In this way, the Edinburgh school rejects fundamental assumptions held by the traditional philosophy of science.

If scientific practices and the development of science cannot be rationally reconstructed, then the sociological analysis of scientific practice must basically be a causal explanation of belief formation and belief validation: it is supposed to show in detail how specific scientific beliefs are established as a result of a causal process proceeding from social conditions, social interests, and negotiations between scientists guiding the specific scientific practice under consideration (Kusch, 2002). Therefore, social constructivism is, in the view of the Edinburgh school, itself a kind of natural science. The domain of this sort of science is the set of all scientific beliefs that get accepted for the time being, independently of whether they are true or false and whether they ultimately prove to be successful or not. But social constructivism, as proposed by the Edinburgh sociologists, can also be formulated as a methodological claim mandating a naturalistic approach to scientific practices; as such, social constructivism is best understood as being part of a naturalized epistemology concentrating on the investigation of social causes of belief formation. In the methodological version, nonveritistic social constructivism does not seem to be committed to ontological claims; however, it remains committed at least to the epistemological claim that scientific knowledge cannot simply be seen as an adequate representation of the external objective world, but must rather be taken to be a result of an extremely complex process involving mainly and most importantly social causes. In this way, social constructivism is not only a form of antirationalism, but also a form of antirealism and in this twofold sense, a form of relativism (*see* Relativism, Cognitive; Relativism: Philosophical Aspects).

Another leading version of a nonveritistic social epistemology is the actor–network theory (Latour and Woolgar, 1986; Latour, 1987) (*see* Actor-Network Theory). The leading thought is that scientific knowledge is an effect of established relations between objects, animals, and humans engaged in scientific practices. An actor is, according to this theory, everything that in some causal way affects the production of scientific statements and theories: not only scientists and their actions, but also, for instance, background assumptions, methodologies, techniques, social rules and institutions, routines, experiments, measurements, and the appropriate instruments as well as scientific texts. For an entity to be an actor in this sense it is obviously not required to have contentful mental states and to be able to perform actions as a kind of behavior describable under some intention. Thus,

there can be many sorts of relations and interactions between actors; in particular, some actors can transform other actors (these transformations are sometimes called translations). A network is a set of actors such that there are relations and translations between the actors that are stable, in this way determining the place and functions of the actors within the network. Once a network has been established it implies a sort of closure that prevents other actors or relations from entering the network, thereby opening the possibility of accumulating scientific knowledge that is taken to be the result of translations within the network. Establishing a scientific belief or theory or fact comes down, from the point of view of the actor–network theory, to placing these actors in a stable network. In this sense, scientific beliefs, knowledge, and theories are taken to be constructed by translations taking place in established networks.

The actor–network theory shares a number of basic assumptions with social constructivism as conceived in the Edinburgh school. Thus, both approaches entertain a naturalistic account of scientific practices, do not presuppose a distinction between true or successful and false or unsuccessful scientific beliefs, and reject the possibility of a rational reconstruction of scientific practices and their outcomes. This is sometimes declared the end of knowledge (Fuller, 1988, 1993).

Nonveritistic social constructivism does not use the term *construction* in the literal sense that some things, and especially facts of the external world, get constructed out of some other things in the course of scientific practices. The crucial idea of a (social) construction of scientific knowledge is, rather, that the process and, indeed, the history of scientific belief formation should not be analyzed in terms of a rational reconstruction that tries to show how the methods of science continuously increase the probability that scientific beliefs are adequate representations of an independent external world. Rather, the process and history of scientific belief formation should be modeled in terms of a number of different factors (including rules, techniques, routines, institutions, and other relations) that affect scientific belief formation – mostly in causal ways that can be studied empirically and naturalistically (*see* Knowledge, Sociology of).

One of the most debated applications of social constructivism is the claim held by many feminists that gender is socially constructed (Anderson, 1995) (see *Gender Studies*). Originally, the most prominent feminist theories introduced a distinction between a naturalistic notion of sex and a social notion of gender to denounce all attempts of deriving properties of gender from properties of sex as mere constructions of gender that do not refer to reality. More recently, in some postmodern feminist accounts of sex and gender, it is also the very distinction between sex and gender, and thus, sex itself, that are supposed to be socially constructed (Lorber and Farrell, 1991; Alcoff and Potter, 1993; Potter, 1993; Charlebois, 2011).

A well-known general objection against nonveritistic approaches to social epistemology is that the leading studies use empirical, sociological, and historical evidence to justify fundamental and far-reaching claims that are supposed to debunk the epistemic authority and universal rationality of the sciences. Obviously, in doing so they rely themselves on exactly those scientific methods they try to debunk. For instance, the claim that "there are no context-free or super-cultural norms of rationality" (Barnes and Bloor, 1982, 27) can only be defended by carefully examining the historical and cultural evidence as well as by applying the hypothetical-deductive method.

Radical (Strong) Social Constructivism

Strong Social Constructivism and Ontology

One of the crucial claims of nonveritistic social epistemology is that mathematics and natural sciences are socially determined too. It is this view that proves fundamental for the further claim of many social constructivists that, in a sense, the world as such is socially constructed. Some constructivists focus not on belief formation, but rather on scientific facts: they want to know in which way such facts are hardened, i.e., get accepted as facts by all researchers in the field (Latour and Woolgar, 1986). It is precisely by shifting the focus from beliefs to facts that some social constructivists are inclined to entertain an ontological version of social constructivism that sees scientists as making up scientific facts.

Proponents of this sort of radical social constructivism rely sometimes on the view (entertained by many different philosophers) that scientific facts are supposed to be what accepted scientific beliefs are about – roughly the content of statements showing up in established scientific textbooks on the field (Quine, 1969). On this view, radical social constructivism has idealistic implications only if scientific facts are understood in a strictly realistic fashion – roughly as the way things are independent of the way we take them to be. Radical social constructionists reject this ontological realism as naive (Bloor, 1996).

From this point of view, proponents of nonveritistic approaches to social epistemology often derive a radical constructivist stance. For instance, according to the actor–network theory establishing a scientific theory comes down to locating all things that have an impact for the production and acceptance of scientific theories (the so-called actors) in a stable scientific actor–network. However, these actors are supposed to include facts and external things. If an actor–network changes, all actors (among them facts and external things) change too. In this sense, not only scientific beliefs, knowledge, and theories, but also facts are taken to be socially constructed by translations taking place in actor–networks.

However, the ontology of strong social constructivism is clearly misguided. Even if one rejects naive ontological realism and concedes that scientific facts are supposed to be what accepted scientific theories are about (as for instance Quine, 1953 did), it does not follow that scientific agreements literally produce scientific facts, but only that scientific agreements justify hypothetical existence claims about some parts of the world. If scientific theories change, then it is not parts of the world but only some of our hypothesized existence claims that change too.

Strong Social Constructivism and the Social

Some radical constructionists maintain that constructivism in the sense of the actor–network theory is social not in the strong sense that social forces that are presupposed to exist largely

independently of scientific practices and have a causal impact on these practices – mainly for the reason that the social is barely better understood than the natural. Rather, constructivism is taken to be social only in the weak sense that as a result of processes taking place in networks as a scientific claim about a distinction between the natural and the social, and so about the function of the social for scientific practices, can eventually be developed (Pickering, 1984, 1992). This branch of radical constructivism rejects the idea of a social determination of scientific knowledge prominent in veritistic approaches to social epistemology as well as, for instance, in the Edinburgh school. This thought involves a confusion, though. It implies the absurd demand that a scientific theory may never rely on claims or distinctions (such as the distinction between the natural and the social realm) that have already been themselves socially constructed.

Finally, among radical social constructionists one might distinguish causal constructionists holding that human activities cause and maintain facts about the world (including scientific facts) from constitutive constructionists holding that facts about the world are really nothing else than human activities (Kukla, 2000). It is not clear, however, whether the leading proponents of strong social constructivism are causal or constitutive constructionists.

Constructivism about the Social (Social Ontology)

Verististic, nonveritistic, and strong social constructivism rely not only on a notion of the social, but also on the assumption that the social realm (including social actions) can be taken to be given. But these kinds of social constructivism neither provide a clear definition of the social nor do they explain in which way social items may exist and so be taken as given for certain theoretical purposes. This is, however, what social ontology aspires to, i.e., social constructivism in then sense of a construction of social entities like social actions, social groups, social status, or social institutions.

Constructivism and International Relations

One of the most recent specific examples of social ontology is the social-constructivist theory of international relations (IR). Traditional IR-theory (in particular neorealism) explores mainly in which ways power-items (for example, military forces or economic capacities) determine international relations, notably international balance of power and interaction between states. Insofar as military forces, economic capacities and similar entities involve a material component, the social-constructivist IR-theory takes traditional and especially neoliberal IR-theory to be a materialist theory. In contrast, social-constructivist IR-theory urges that international relations are basically social, not material. They are a human creation relying on human ideas guiding political actors moving in the international scene. In this sense, international relations are socially constructed: "The claim is not that ideas are more important than power and interest, or that they are autonomous from power and interest. The claim is rather that power and interest have the effects they do in virtue of the ideas that make them up. Power and interest presuppose ideas…Let me [propose] a rule of thumb for idealists: when confronted by ostensibly "materialist" explanations, always inquire into the discursive conditions which make them work" (Wendt, 1999, 135 f.).

As this quote from one of the leading social-constructivist IR-theorists shows, the position is sometimes not sufficiently carefully articulated. In particular, some of the formulations come close to an idealist stance involving the claim that ideas 'makeup' power, interest and, indeed, international relations, and, furthermore, that international relations depend exclusively on discursive conditions and "exist only as an intersubjective awareness among people … It is a set of ideas, a body of thought, a system of norms, which has been arranged by certain people at a particular time and place" (Jackson and Sørensen, 2006; ch.6, 162).

But there are also proponents of social-constructivist IR-theory who emphasize that it is not only, and not even prominently, ideas but rather social norms, rules, and relationships we have with others that determine political behavior, "endow material entities with meaning and purpose and provide us with ends to which power and wealth can be used" (Finnemore, 1996, 128). Still, even in these less idealistic accounts of social IR-constructivism a clearly articulated relation between ideas, actions, institutions, interests, and power is missing. It is general social ontology that presents a picture that is, in this respect, much more transparent.

Constructivism and Social Ontology

In one of the most recent developments of social constructivism, leading proponents of the Edinburgh school try to exploit the thought that scientific knowledge is, after all, something like an institution and that institutions in general are constructed by humans (Bloor, 1996). It seems to follow that scientific knowledge is, as being constructed by humans, socially constructed. The idea that institutions are constructed by human persons is part of what is often called the intentionalist program of social ontology. The aim of this program is to clarify the ontological status of social entities, i.e., of social groups, social status, and institutions. The fundamental claim is that it is on the basis of collective mental states and of actions describable under collective intentions that social entities come into existence and maintain an ontological status. That is to say that human beings produce, that is literally construct social entities by having specific contentful thoughts and by performing certain intentional actions (*see* Cooperation: Sociological Aspects).

It is obvious that this approach is a sort of social constructivism in the sense that it is a theory about the construction of the social. This construction is itself held to be social insofar as it consists largely of normative interactions between humans. But it must be emphasized that this sort of constructivism is restricted to the realm of the social, that it presupposes a sharp distinction between the natural and the social, and that it has nothing to say about the construction of natural facts or physical entities. Most importantly, though, the intentionalist program relies on an intentional and normative vocabulary. In all these respects the program differs fundamentally from traditional versions of social constructivism.

Basically the intentionalist program of social ontology is an individualistic account of the social. It relies on the notion of an action as behavior that is describable under a specific contentful intention. Actually, it is the intention that epistemically identifies and ontologically individuates the given behavior as a certain action. One of the crucial theoretical moves is to introduce notions of collective intentions, beliefs and actions. The first step is to assume that for some persons to be socially related in a specific way is to be treated by most of the other members of their society as being related this way. Social relations are taken to imply rules as well as certain rights. In general, therefore, to treat some persons to be socially related in way S is to treat them as being entitled or committed to perform actions according to the rules and rights derivable from S. One of the next important moves is to introduce the notion of a collective attitude (or, as it is sometimes called, a we-attitude) and, in particular, the notion of a collective intention to perform a specific action. Roughly speaking a person P as member of a set S of persons has a contentful collective attitude of the sort A iff (1) P has the A-attitude herself, (2) P believes that every other member has the A-attitude, and (3) P believes that every other member M of S believes that every member of S other than M has the A-attitude. In particular, if a number of persons P_i intend to perform action A together or collectively, then it must be the case, in addition to (1)–(3), that every P_i intends to perform A by performing a specific action A_i as her specific contribution to the performance of A such that P_i thinks that A_i is necessary for A and results, together with the other specific actions A_j intended to be performed by the other P_j, in performing A, and that P_i believes that the other P_j intend to do the same thing and believe themselves that every other person intends to do the same thing. On the basis of these notions it can be defined that (a) that a number of persons P_i have the collective intention to perform action A iff each P_i has the collective intention to do A by performing a specific A_i, in the sense just explained; and P_i perform a collective action A iff A can be described under some collective intention of the P_i. One might then suggest that a number of persons P_i represent a social group iff they share a collective attitude. Finally, let MR be a set of semantically consistent rules R_i such that a number of persons collectively treat an X as a Y by following R_i, then MR is, in the most basic sense, an institution. In particular, if MR is the set of rules according to which scientific results and scientific knowledge are established, and if the persons in question are trained scientists, then MR is an institution of epistemic or scientific practices, and in this sense, scientific knowledge might be called an institution that is constructed by intentional acts of human beings (Gilbert, 1989; Pettit, 1993; Searle, 1995, 2011; Tuomela, 1995, 2007).

This line of argument is a rough sketch that by no means exhausts the theoretical complexity of the intentionalist program of social ontology. For instance, something has to be said about the mechanisms of the formation of collective attitudes to account for the stability of social groups and institutions; at this point, a theory of consensus and a theory of power has to be implemented into the general account. Furthermore, social ontology has to be related to scientific practices in the social sciences (Mantzavinos, 2009).

In any case, even a rough sketch of this sort of social ontology shows in which way, to put it very generally, collective minds and collective actions can be thought to bring into existence, and to keep in existence, and thus to construct, social entities in a fairly literal manner. And it becomes also clear how and why traditional social constructivists might be inclined to render their position more precise by exploiting the intentionalist program of social ontology. This program is a philosophically sound approach. It makes evident, though, that the collective human mind and collective human actions construct social facts in ways in which they don't construct natural facts, i.e., parts of nature governed by natural laws.

Nonveritistic and strong versions of social constructivism tend to downplay this distinction, and this strategy is certainly a fundamental flaw. Another problem with these versions is that they seem to commit themselves to relying exclusively on a causal, physicalist vocabulary. While it certainly makes sense to examine whether a purely naturalistic account of epistemic practices might be fruitful, it does not follow from this strategy, even if it proves fruitful, that intentionalist strategies must be excluded. On the contrary, it is doubtful whether epistemic practices can be adequately described without using also intentionalistic and normative vocabulary. Finally, it is hard to see how nonveritistic social constructivists can overcome Davidson's point that interpretation theory must be part of theorizing about epistemic practices which implies, in turn, that a notion of rationality cannot be excluded and that, at the most fundamental level, the notions of rationality in different epistemic practices must converge. Obviously, this insight is inconsistent with an unqualified contingency claim and relativism about rationality so central for nonveritistic and strong social constructivism.

See also: Actor-Network Theory; Cooperation: Sociological Aspects; Knowledge, Sociology of; Relativism, Cognitive; Relativism: Philosophical Aspects; Social Properties (Facts and Entities): Philosophical Aspects; Social Science, The Idea of; Strong Program, Sociology of; Truth and Credibility in Science.

Bibliography

Alcoff, L., Potter, E. (Eds.), 1993. Feminist Epistemologies. Routledge, New York.
Anderson, E., 1995. Feminist epistemology: an interpretation and a defense. Hypatia 10, 50–84.
Bachelard, G., 1934. Le nouvel esprit scientifique. Alcan, Paris.
Barnes, B., Bloor, D., 1982. Relativism, rationalism, and the sociology of knowledge. In: Hollis, M., Lukes, S. (Eds.), 1982 Rationality and Relativism. MIT Press, Cambridge, MA.
Bloor, D., 1976/1992. Knowledge and Social Imagery. University of Chicago Press, Chicago, IL.
Bloor, D., 1996. Idealism and the sociology of knowledge. Social Studies of Science 26, 839–856.
Canguilhem, G., 1966. Le normale et le pathologique. PUF, Paris.
Charlebois, J., 2011. Gender and the Construction of Dominant, Hegemonic, and Oppositional Feminities. Lexington Books, Lanham, MD.
Detel, W., 2007. Wissenskultur. In: Schützeichel, R. (Ed.), 2007 Handbuch Wissenssoziologie und Wissensforschung. Universitätsverlag Konstanz, Konstanz, pp. 670–679.
Detel, W., 2009. Wissenskultur und universelle Rationalität. Fried, 181–214.
Fried, J. (Ed.), 2009. Wissenskulturen. Über die Erzeugung und Weitergabe von Wissen. Campus Verlag, Frankfurt/Main.

Finnemore, M., 1996. National Interests in International Society. Cornell University Press, Ithaka NY, London.
Foucault, M., 1969. L'archéologie de savoir. Gallimard, Paris (engl. transl. 1980 Power/Knowledge, Pantheon, NY).
Fricker, E., 1995. Telling and trusting: reductionism and anti-reductionism in the epistemology of testimony. Mind 104, 393–411.
Fricker, E., 1998. Rational authority and social power: towards a truly social epistemology. Proceedings of the Aristotelian Society 19 (2), 159–177.
Fuller, S., 1988. Social Epistemology. Indiana University Press, Bloomington.
Fuller, S., 1993. Philosophy, Rhetoric, and the End of Knowledge. University of Wisconsin Press, Madison.
Gilbert, M., 1989. On Social Facts. Routledge, London.
Goldman, A., 1978. Epistemics: the regulative theory of cognition. The Journal of Philosophy 75, 509–523.
Goldman, A., 1987. Foundations of social epistemics. Synthese 73, 109–144.
Goldman, A., 1999. Knowledge in a Social World. Oxford University Press, Oxford.
Goldman, A., 2001. Experts: which ones should you trust? Philosophy and Phenomenological Research 63, 85–110.
Goldman, A., 2006. The social epistemology of blogging. In: van den Hoven, J., Weckert, J. (Eds.), 2006 Information Technology and Moral Philosophy. Cambridge University Press, Cambridge, pp. 111–122.
Goldman, A., Cox, J., 1996. Speech, truth, and the free market for ideas. Legal Theory 2, 1–32.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge.
Hacking, I., 2000. The Social Construction of what? Harvard University Press, Cambridge, MA.
Jackson, R., Sørensen, G., 2006. Introduction to International Relations: Theories and Approaches. Oxford University Press, Oxford.
Kitcher, P., 1990. The division of cognitive Labor. The Journal of Philosophy 87, 5–22.
Kitcher, P., 1993. The Advancement of Science. Oxford University Press, New York.
Knorr-Cetina, K., 1981. The Manufacture of Knowledge: an Essay on the Constructivist and Contexualist Nature of Science. Pergamon Press, Oxford.
Knorr-Cetina, K. (Ed.), 1982. Science in Context. Sage, London.
Kuhn, T., 1962/1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.
Kukla, A., 2000. Social Construction and the Philosophy of Science. Routledge, London.
Kusch, M., 2002. Knowledge by Agreement. Clarendon Press, Oxford.
Lackey, J., 2006. It takes two to Tango: beyond reductionism and non-reductionism in the epistemology of testimony. In: Lackey, J., Sosa, E. (Eds.), The Epistemology of Testimony. Oxford University Press, New York, pp. 160–191.
Latour, B., 1987. Science in Action. Harvard University Press, Cambridge, MA.
Latour, B., Woolgar, S., 1986. Laboratory Life. Princeton University Press, Princeton.
Longino, H., 1990. Science as Social Knowledge. Princeton University Press, Princeton.
Longino, H., 2002. The Fate of Knowledge. Princeton University Press, Princeton.
Lorber, J., Farrell, S., 1991. The Social Construction of Gender. Sage Publications, Thousand Oaks, CA.
Mannheim, K., 1936. Ideology and Utopia. Trans. L. Wirth and E. Shils. Harcourt, Brace and World, New York.
Mantzavinos, C., 2009. Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice. Cambridge University Press, Cambridge.
Merton, R., 1973. Sociology of Science. University of Chicago Press, Chicago.
Pettit, P., 1993. The Common Mind. Oxford University Press, Oxford.
Pettit, P., 2006. When to defer to majority testimony – and when not. Analysis 66, 179–187.
Pickering, A., 1984. Constructing Quarks. University of Chicago Press, Chicago.
Pickering, A., 1992. Science as Cultural Practice. University of Chicago Press, Chicago.
Potter, E., 1993. Gender and Epistemic Negotiation. In: Alcoff, L., Potter, E. (Eds.), Feminist Epistemologies. Routledge, New York, 121–159.
Quine, W., 1953. On what there is. In: Quine, W. (Ed.), From a Logical Point of View, Cambridge, MA.
Quine, W., 1969. Ontological Relativity and Other Essays. Columbia University Press, New York.
Searle, J., 1995. The Construction of Social Reality. The Free Press, New York.
Searle, J., 2011. Making the Social World: the Structure of Human Civilization. Oxford University Press, Oxford.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air-pump. Princeton University Press, Princeton.
Thagard, P., 1997. Collaborative knowledge. Noûs 31, 242–261.
Tuomela, R., 1995. The Importance of Us: a Philosophical Study of Basic Social Notions. Stanford University Press, Stanford.
Tuomela, R., 2007. The Philosophy of Sociality. The Shared Point of View. Oxford University Press, Oxford.
Wendt, A., 1999. Social Theory of International Politics. Cambridge University Press, Cambridge.

Social Control and Deviance, History of

Elizabeth E Mustaine, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One of the many founding concepts of contemporary social science that arose during the Enlightenment was the idea that society could control the behavior of its citizens by using formal mechanisms of social control. Theoretically, by manipulating the magnitude of punishment or pain guilty persons would experience for their deviant behavior, society could stop offenders before they acted. More specifically, since offenders acted in their own best interests, they would think about the costs and benefits of their behaviors prior to acting. Through punishment, society could make the costs of deviant or criminal behavior outweigh whatever benefits miscreants might otherwise accrue. Moreover, this would translate to more general societal deterrence if the punishments individuals received were certain, swift, and severe. This article reviews the history of these concepts through to the present day.

One could argue that humans live collectively. Is it because we are social creatures? Or, do we need each other to survive the harshness of life on earth? It is likely to be some combination of desire and necessity as well as a variety of additional explanations. What is important here is that humans have benefited from the formation of societies and communities in many ways (e.g., the division of labor, collective efficacy, protection from outsiders, in-group processes to name just a few). Indeed, early philosophers noted that the need to form societies and communities provided so much benefit that people were willing to make sacrifices in order to belong. Typically, when humans come together in groups, these societies/communities necessarily establish norms and values in order to make life more predictable and navigable. Of course, the mere formation of norms almost assures that there will be individuals, who, for whatever reason are unable or unwilling to meet the standards and expectations of their fellow affiliates, and society must deal with these nonconformists. Many academic disciplines have proffered explanations for deviance and social control in order to better understand this seemingly inevitable group phenomenon. Sociology (and its subfield of criminology) is one such discipline. Early on, the study of deviance and social control was seen as fundamental to the understanding of society. Threats to morality and the social order, and institutional processes and efforts to control such encroachments were viewed as vital topics for sociological investigation and discourse. Later, scholars eschewed the study of deviance, ironically viewing the intellectual discourse surrounding deviance and social control as not much better than the nonconformists with which they concerned themselves. As time has passed, the relevance of deviance and social control has returned, and the theoretical explanations for them have become more and more sophisticated. What follows is a brief treatise on the history and development of sociological discourse pertaining to deviance and social control.

Prior to the formation of the discipline of sociology, there were plenty of intellectuals interested in explaining deviance and social control. Early primitive people lived in a world where behavior and events were seen as being the result of otherworldly powers. Plagues, floods, and famines were seen as divine punishments (Newman, 1978). Religious leaders performed rituals in order to appease these vengeful spiritual powers. Deviance was seen as the result of possession, and persons who were suspected of deviant behavior were put through 'trials by ordeal' in order to determine their guilt or innocence. Individuals who survived these ordeals were innocent and those who did not survive were guilty, but they were also dead so the situation had been resolved (e.g., A woman suspected of being a witch was tied up and thrown into the water. If she floated she was not a witch; if she sunk was guilty of witchery). Social control, then, in this environment was more about punishing guilty miscreants and ridding society of them (Newman, 1978).

During the Enlightenment, intellectual discourse shifted and philosophers began to view human behavior and social events as being the result of freethinking and free will. Beliefs such as humans being rational, capable of understanding themselves, and acting in their own best interests took center stage, thereby shifting views about deviant behavior from being otherworldly to the result of clear and calculated thinking. This, obviously, represented a huge change in the underlying philosophy of social control and what societies could do to minimize unacceptable and harmful behavior. From this view, scholars from different paradigms formulated theories that explained deviance and social control using different conceptualizations and perspectives.

One important paradigmatic shift that developed was the idea that society could, through serious contemplation, control the behavior of its citizens by using formal mechanisms of social control (Beccaria, 1764; Bentham, 1789). Theoretically, by manipulating the magnitude of punishment or pain a guilty person would experience for his/her deviant behavior; society could stop offenders before they acted. More specifically, since offenders acted in their own best interests, they would think about the benefits of their behaviors prior to acting. Through the threat of punishment, society could force offenders to consider the costs of their behavior whereby (if assessed correctly) the pain or cost of committing a crime slightly outweighed the benefits of it. Moreover, this would translate to more general societal deterrence if the punishments individuals received were certain, swift, and severe. With this perspective, then, deviance was basically those behaviors that were illegal,

because social control was more formulized and was handed out by the State.

This view that society could impact the behavior of individuals through formal social control mechanisms encouraged further research on how groups or communities could positively influence the behavior of their members. Sociologists saw social control as encompassing all those social practices and arrangements that contributed to the maintenance of social order and conformity among social group members, both formal (e.g., arrest) and informal (e.g., shunning). Early theorists viewed these processes as a reflection of social consensus, thereby characterizing social groups as harmonious, consensual, and concordant. Scholars in the late nineteenth century also viewed social control as the fabric of 'organized society.' Social control, then, encompassed the tools through which society formed and maintained harmony.

One problem, however, with restricting the study of deviance to illegalities is that in communities and neighborhoods, people have varying views about deviance. So, researchers must contend with the varying views that people have about it. Because of this, criminologists often find it difficult to work within a legal framework. There are two primary reasons for this. (1) Often what distinguishes whether or not a particular behavior is legal or illegal is based on a technicality that likely has no bearing on causation (e.g., manslaughter vs tragic accident is often a matter of perspective). (2) The law groups various behaviors together that are not similar in terms of causation (e.g., serial homicide, which includes serial murder and mass murder) (Vold, 1958). For this reason many criminologists focused on the 'nature' of a particular behavior rather than the legal definition of it. For example, Hermann Mannheim (1965) posited the view that criminology should study all antisocial behavior, while Edwin Sutherland (1934) viewed crime as behavior that was harmful to society. Thorsten Sellin (1950) defined crime as the violation of conduct norms, while Schwendinger and Schwendinger (1977) thought of crimes as violations of human rights.

Continuing, deviance theorists often focused on the violation of cultural, group, or societal norms, rules, and expectations. Where they differed was the formulation of those norms and expectations. And, an important distinction among scholars was the difference between 'primary' and 'secondary' deviance (Lemert, 1951). Primary deviance was committed by a wide range of individuals in any population, and the causes of it vary widely. Secondary deviance was deviance that was committed due to the stigma associated with certain actions, and the processes that society used to punish and control it. What most scholars agreed on, however, was that there was nothing constant about social views of crime or deviance, so pinpointing certain actions as criminal or deviance over time was a fruitless endeavor. In spite of these definitional difficulties, early sociological deviance theorists typically saw deviance as the result of 'pathologies,' or things that were wrong with individuals, communities, or societies.

Alongside the development of the perspective that humans had free will was a methodological shift in the science of deviant behavior. Auguste Comte (1865), regularly viewed as the founder of the discipline of sociology, wrote less about deviance and social control, but extensively about positivism as an epistemological perspective. Many theorists were influenced significantly by Comte. And, this shift in methodological stance marked such a tremendous change in the frame of reference of theoretical discourse that it might be likened to a revolution. In the earliest years of sociology, then, the study of deviance was primarily the study of observable or empirically based causes of crime and deviance.

One example of early positivism is Cesare Lombroso (1835–1909) who is viewed as the father of the positivist school of criminology (Lombroso, 1876). Initially, Lombroso examined several hundred Italian criminals and found they had physical characteristics that suggested inadequacy and degeneracy. Such characteristics were asymmetry of the face; abnormally large or small head size; excessive jaw and cheekbone dimensions; ears of unusual size (e.g., small, protruding from the head); abnormal nose (e.g., twisted, upturned, flattened, beaklike, or having swollen nostrils); fleshy, swollen, and protruding lips; and pouches in the cheek like those of some animals. These physical characteristics resembled the appearance of lower or more apelike ancestors; thus the individuals having them were assumed to belong to these earlier evolutionary forms. Even though Lombroso's theory of the atavistic criminal has since been debunked, he is important because he exemplifies early efforts to combine theory and science, the scientific generation of facts that can be explained by theoretical interpretation. Additionally, later in his career, he began including psychological and environmental factors into his deliberations on the criminal man, thereby incorporating multiple factors into his theoretical development.

Another example of positivism was one that considered more sociological explanations. During the late 1700s to early 1800s societies worldwide were undergoing massive transformations, and these changes led theorists to consider what changes in deviance would occur due to emerging social change. To elaborate, during the mid-1800s until the mid-1900s American and European sociologists were studying the large-scale social changes that were taking place (Websdale, 2010). These changes included such immense happenings as urbanization, industrialization, capitalism, immigration or emigration, family, and community. These massive social changes influenced societies, communities, and individuals in equally large parts. For example, as the world became more industrialized, more and more young adults left their families' farms and moved to the city in search of wage labor. This increased the amount of urbanization, and population density, which, in turn, changed the way people thought of 'community.' The influx of immigration also changed the face of communities and neighborhoods and cultural heterogeneity increased. As more and more people were uprooted from their traditional lives, many customs fell to the wayside, and people were thrown into a new normless society. In the more traditional communities, family depended on their neighbors for food or items they could not produce themselves. In the city, people did not necessarily know their neighbors or interact much with them. And, men and women were both seen as making valuable and necessary contributions to the functioning of the family. But, with the onset of city life, the close-knit communities that characterized farming areas fell to the wayside and the economy shifted from one where a man worked a job that (literally) provided the food and other items his family needed to one where having a job that paid

money was a necessity. This, of course, changed traditional gender roles when men left the home to work, and women stayed at home to work. Another change that confused gender norms was that as more young men got jobs in the cities, the ways in which they could express their masculinity had to change also. Additionally, the extended family broke up and the nuclear family became more common. This led to a reduction in the number of children couples had because children became economic liabilities and no longer free labor. All of these changes were happening in societies all over the globe, and even though they brought confusion, anger, turmoil, uncertainty, and even an increase in the amount of deviance and crime, most people adapted and conformed to the new social norms and values.

More sociological causes of crime emerged in Europe during this time as systematic registries of births, deaths, and other official records were collected, maintained, and were seen as relatively accurate. Sociological positivists examined the relationships between crime, poverty, economic depression, malnutrition, low intelligence, resentment, unemployment, and relative social inequality, and found that relationships, while not corresponding 100% of the time, these relationships did, indeed, provide partial explanations for criminal or deviant behavior. One issue that arose among these researchers was how to interpret these relationships. For example, some saw poverty as a cultural or learned helplessness that causes crime, while others viewed poverty as a structural limitation that caused deviance and other socially unacceptable behavior.

This search for the pathological causes of crime, however, grew unsatisfying for many sociologists, who began to delve into other ways to explain crime, criminal behavior, and other forms of socially harmful conduct. By focusing on only those pathological characteristics of individuals, some emerging sociologists felt that many potentially relevant social factors were ignored. It was in this climate, the Chicago School of Sociology was born. Researchers at the University of Chicago in the early to mid-twentieth century aggressively emphasized a break with prior (whom they saw as naïve) theorists who were too narrowly focused on individuals and pathologies. Chicago academics saw the environment as a crucial location for research, and the research to be necessarily done involved more rigorous empirical research such as life histories, ethnographic field studies, and census data.

Early on the Chicago School established its theoretical base as the city. Scholars in Chicago such as Robert Park et al. (1925), found the urban environment, with its tremendous diversity, locational and structural patterns, and varying cultures and lifestyles to be infinitely interesting and researchable. Additionally, they viewed areas within the city as having their own character, functions, and populations. This led to the development of their main theoretical proposition which was that cities were arranged in concentric zones of social life. The central business district made up the center zone of social life in the city. It was surrounded by the 'zone of transition,' which was an area inhabited by the poor, the marginalized, and the recently arrived. Next to the transition zone was a concentric layer of working class residents, followed by the residential zone, made up of people who had moderate amounts of money, regular work, and greater ownership of property. Finally, there was the commuter zone, where people who wanted to flee the largeness, density, and diversity of the city lived.

Clifford Shaw and Henry McKay's longitudinal study of juvenile delinquency in Chicago bore out this zonal perspective (1942). Using a very early sort of mapping analysis, Shaw and McKay constructed a series of maps of Chicago and mapped the residences of juvenile delinquents and various other community characteristics (e.g., the location of demolished buildings, the incidence of tuberculosis, and the residences of truants). They then constructed rate maps which showed the percentage of the population that was delinquent based on the other mapped characteristics. The analysis of these maps indicated several important things: that areas with the most delinquency did not necessarily have the most delinquents; the areas with the greatest rates of social problems and juvenile delinquency tended to be concentrated close to the center of the city; the areas that had the highest rates of juvenile delinquency were located within or immediately adjacent to areas of heavy industry or commerce. These areas also had the highest rates of structural degradation, population decline, low economic status, high percentage of families on welfare, low percentage of families owning their own homes, high infant deaths, high rates of tuberculosis, and high rates of adult criminality. In addition, these areas also had high proportions of black and immigrant residents. Finally, these patterns held even when the ethnic composition of the transitional zone changed (e.g., from Irish to Puerto Rican). These findings provided solid evidence that delinquency was influenced, at least in part, by location and urban culture, and not completely by individual pathologies. Deviance was thus viewed as a product of social disorganization.

Another distinction with the Chicago School was that they not only had a strong organizing theoretical framework, but they had an institutional base. This led to an emphasis on training researchers to continue their theoretical development via continued empirical research. So, describing and understanding the underlying dynamics of social life in the zone of transition was another aspect of the theoretical discourse during the heyday of the Chicago School. The graduate students of these eminent scholars were tasked with going into these deteriorated but highly diverse neighborhoods and gaining access to the worlds of the poor, powerless, and exotic and telling their stories (often in miniscule detail) through ethnographic explorations. Such views characterized many of the books published during this time, which included ethnographies of gangs (Thrasher, 1929), organized crime (Landesco, 1929), the hobo (Anderson, 1923), the prostitute (Reckless, 1933), the taxi-dance hall (Cressey, 1932), the rooming house district (Zorbaugh, 1929), the immigrant neighborhoods (Thomas and Zanaiecki, 1920), and blacks (Drake and Cayton, 1945), just to name a few. This concentration on the transitional zone as being distinct from other zones led to the central underlying theoretical theme, which was that different areas of the city were organized differently.

The white male scholars of the Chicago School interpreted this differential organization as less organized than other zones, and thus began social disorganization theory. This theory, which is still popular today, emphasized that neighborhoods with social characteristics such as lower economic status, lower home ownership, higher percentages of the population that are minority or immigrant, higher population density,

more vacant or boarded-up buildings, trash and litter on the street and in yards, higher amounts of victimless crimes (e.g., prostitution, drug, or gun sales), higher amounts of vandalism and property damage, unkempt lawns, and fewer neighborly interactions were neighborhoods where residents were unsatisfied with their surroundings, were planning to leave as soon as money allowed, and cared less for the upkeep of their community since they had less 'stake' in it. Because of this, these residents were also less collectively oriented, so their abilities to rid the community of negative influences was compromised, and they were unlikely to look out for each other and notice outsiders in their midst (Sampson, 2012). This more apathetic attitude showed in the condition of the neighborhood, and as a result, crime and other forms of deviance flourished.

One of the common criticisms of the Chicago School, however, was that its theorists used their own moral judgments to explain their findings rather than sociological analysis. For example, some critics argued that the social world of gangs, prostitutes, and those who frequent taxi-dance halls was not disorganized, per se, it was just organized differently. In this way, than, it was likely that neighborhoods with greater amounts of crime were not so much disorganized as they were differently organized, in ways that allowed crime or deviance to move in and stay.

Another perspective that grew from the idea that humans had free will, was the Functionalist view of crime and deviance which maintained that crime was more a way of life for some groups. In this way, crime and deviance were 'natural' responses to group and social pressures, and criminals and deviants were normal. As the founder of this perspective, Emile Durkheim (1893, 1897) theorized that among the many social institutions that were functional for society, deviance was an institution that was more complex. Even though deviant behavior could cause harm to individuals, it was still functional for society because it forced people to confront their views about what was acceptable and unacceptable and make decisions about what actions would be condemned and which would be absolved. Durkheim argued that without negative behavior, members of society would not have values or morals and there would be a breakdown in the ability of society to regulate the natural proclivities of individuals. This, in turn, would lead to anomie, or a feeling of being not grounded or unconnected.

Subsequent functionalists used these theoretical principles and developed what is known today as strain theory. Strain theorists, such as Robert K. Merton (1938) and Marshall B. Clinard (1957) proposed that each society has at least several social generated pressures or strains that cause criminal behavior. Not all groups feel these pressures equally, but the groups that experience them the most severely are the ones that are most criminal or deviant. This line of theoretical discourse explains that every culture has some important values or goals that are accepted by nearly everyone. There are also institutionally accepted ways or means of achieving these key goals. However, the structure of most societies does not allow everyone to actually attain these well-accepted goals. So, many will be unable to meet cultural expectations, and this will cause anomie among those individuals as well as in groups where many members fall short. This is what leads to strain. Merton argued that there are various ways that individuals can respond to anomie, depending on that individual's or group's attitude about the cultural goals and means. In highly stable societies, individuals will continue to conform, which is to say they still want to meet cultural expectations using socially approved means. Most crimes emerge from the response called innovation. Persons responding this way continue to want to achieve the cultural goals, but find they are unable to do so using acceptable social means, so they innovate (usually by doing something illegal). Some individuals will use a ritualistic response, which is when they continue to support cultural goals, but essentially give up on ever achieving them, and just go through the motions. Retreatists respond by dropping out of the whole game altogether. Those who choose this method of adaptation are typically psychotics, outcasts, vagrants, drunkards, and drug addicts. Rebellion is the last possible response to the problem of anomie. Here the person responds to his failure and frustration by replacing the values of society with new ones. These new values may be political (over throwing the government), religious (seeking a new state of spiritualism), or any number of other possibilities. But, in the end the rebel ceases to be a part of society and lives with an alternate culture, which is likely to be seen as deviant.

Strain theory has segued into subculture theory and is used to explain social groups that are deviant or criminal. For example, theorists explain gang formation as a collective and innovative solution by a group of underprivileged youth (Cloward and Ohlin, 1960). A group of religious outcasts responds with rebellion and physically removes themselves from society in order to relocate to the mountains of Idaho. A recent group of immigrants that is unable to find adequate work and housing after entering a country forms a community and economic enclave within an urban area where its members can work, live, express their unique cultural beliefs, and do not have to interact much with the larger society.

As a competing perspective, French sociologist Gabriel Tarde (1843–1904) expressed the notion that the great level of harmony in society was as equally worthy of study as penal philosophy (1890). More specifically, Tarde observed that most citizens' behavior was congruent with social norms and values, suggesting that social control strategies were fairly successful. For example, when considering deviance, Tarde wrote that the more interesting sociological puzzle was how people came to live together closely and in harmony. In other words, what might be equally, if not more interesting than explaining nonconformity was to explain such high levels of conformity.

So, even though deviance was an interesting sociological phenomenon, what followers of Tarde found equally compelling was the question of why so many people were cooperative, conforming, and 'normal,' even in times of social upheaval. Scholars during the early twentieth century wondered what could account for the reemergence of stability and harmony in the chaotic modern world (for a discussion of this issue see Buzan and Lawson, 2014)? And, given that the formal instruments of social control (e.g., police, governments) were not the likely answer, because social control could be created and maintained without the use or threat of force, what institutions of society were the primary contributors to social stability?

It is here that sociology provides an assortment of answers. Religion, education, family, the whole apparatus of socialization, politics, the military, and the reestablishment of social norms and values were some of the social institutions that theorists looked toward to understand both deviance and conformity. This perspective also shaped a long-standing disposition of sociology which is to analyze biography and history in order to better understand and provide solutions to social problems (Mills, 1959). More current scholars of this paradigm suggest that within these social institutions, people with a 'stake in conformity' were not deviant (Toby, 1957). The idea here is that most people's behavior is controlled through informal social control (e.g., neighbors and communities issuing 'in-group' or 'out-group' status or glares, exclusion, and disapproval). Because we are social beings, we want to fit in and be a part of the group. If that is threatened, we conform to norms and values in order to be accepted (Nye, 1958). Travis Hirschi (1969) provided an explanation for why people have a 'stake in conformity' by noting that there are four elements that increase the social bonds people have with the members of their families and communities. These elements are having an *attachment* (e.g., affectionate toward, sensitivity to) to others and their communities, being *committed* to conventional society (e.g., attending church, doing well in school, getting a job), *involvement* in conventional activities (e.g., playing on a sports team after school, joining the PTA), and *belief* in obeying the laws.

Another paradigmatic development that was occurring alongside the development of the Chicago School maintained its methodological emphasis on observation, but moved to the forefront the fluidity of deviance and social control. Among these Social Reaction theorists were Edwin Lemert (1951), Howard S. Becker (1963), and Erving Goffman (1959), all of whom became increasingly concerned with the process of becoming deviant; particularly the part played by official institutions of crime control and community informal social control. The theoretical gist here is that individuals who commit a crime or who are deviant once or twice (often unknowingly or with only minor transgressions) are officially labeled deviant, stigmatized by their communities, and become ignored and marginalized. This process often results in a continuation of deviant behavior rather than desistance, because the individual has developed the identity of deviant (or criminal), and becomes anomic and disconnected from society. So, deviance is largely defined by the stigmatizing and punitive reactions of others. This theoretical discourse, while originally ignored, became influential in the 1960s as some groups' cultural beliefs included ideas that governments and the rich had the power to impose their definitions and lifestyles on to the rest of the members of society. These new beliefs likely came to pass as a result of individuals who protested the war were attacked by law enforcement officers, and the emerging subcultures of greater drug use and free sexual expression were frowned upon by more 'respectable' societal members.

Social reaction theory's intellectual roots are the symbolic interactionist view of society. Here, the theoretical essence is that of dramaturgy (Goffman, 1959). Metaphorically, people act out their lives on a stage in front of an audience and they manipulate the audience's view by using or omitting various props, language, and settings to establish an impression. The audience, in turn, sees the props, is aware of the setting, and listens to the actor's use of language and forms an impression. Dramaturgy, then, is a two-way street, and is not always predictable. The actor may attempt to control the impression the audience has of his/her, by using a particular prop, but the audience may have a different association with that prop, thereby forming a different impression than what was intended. These symbolic interpretations are so strong that individuals often behave toward others according to their beliefs about these others. For example, Social Reaction theorists have discussed the impressions society has of the blind, the disfigured, stutterers, strippers, prostitutes, perverts, addicts, alcoholics, the unwashed, and the unwed (particularly mothers). The blind, for example, indicate that people often yell at them, apparently thinking if they are visually impaired, they must also be hard of hearing. Additionally, many people think that strippers, prostitutes, and unwed mothers are morally deficient, preferring to focus on the symbolic meanings of their observable actions rather than considering their hard-luck circumstances. Many researchers have looked at common educational practices that illustrate the bureaucratic process that labels certain children 'handicapped' or 'slow' and then ushers them into special education programs.

This theoretical discourse is easily applied to deviance. People hold many symbolic beliefs about those who behave unconventionally. Therefore, once someone has been deviant (primary deviance), the official processing that occurs, and this leads to stereotypical labeling. Deviant individuals, then, are viewed as moral reprobates, are shunned, excluded, marginalized, and ignored. This leads to a disconnection from society and further deviance (secondary deviance). This process can also take place on a more informal level. For example, if a boy is caught smoking marijuana in his neighborhood, it may happen that his neighbors will begin to think of him as a 'druggie' and stop letting him play with their children, get worried when he is around, and cross the street to avoid him when they are walking. This, in turn, would cause that boy to lose his bond with members of the community and feel isolated from them.

In keeping with this perspective on the world, the methods of research by these sociologists were to emphasize *Verstehen* (Weber, 1946). So, greater understanding could only be had when theorists examined deviant persons and groups from their own points of view. Researchers, then, who attempted to systematize or generalize their findings on unconventionality, were to be distrusted.

A final paradigm developing from early criminologists was heavily influenced by Karl Marx (1848). Theorists of this bent, such as George Vold (1958), Richard Quinney (1970), Jeffrey Reiman (1979), and William Chambliss (1973), drew on the politics of Marx to establish the foundation for their theories of crime. Marx argued that in the organized capitalist state the economy was controlled by the owners of the means of production. The workers' interests were not represented, and the only way the workers could gain any independence from the power of the owners was for the state to take over the means of production. Only when the state was in control would all

citizens' interests and needs be accommodated. Sociologists in the 1970s and later, particularly, were greatly influenced by the distrust of those in power and the assertion that crime and deviance were a result of social authority and labeling. This perspective harkens from the conflict approach to understanding society, and these theorists used radical discourse to suggest that the government itself was actually the cause of crime. The State caused criminal behavior because of its unequal treatment of social minorities, and its self-perpetuating goal of keeping the rich in control of the government and economy and the poor downtrodden and demoralized. Here, then, deviance is variable, as it reflects the opinions of the group that is in power. And, social control is used by those in power to keep down those who are not.

Scholars took a number of directions when applying Marx to the study of crime, deviance, and social control. Sellin (1950) suggested that conflict would arise when cultural subgroups' values and norms collided. As society became bigger and more complex, this collision would be inevitable. Within simple societies, cultural values and norms often become law and represent a consensus in society. But, as society grew and became more complex, subgroups with cultural distinctions would clash and the more powerful subgroups would emerge with their values and beliefs represented in law, while the norms and values of other less powerful cultural subgroups would become illegal or deviant. An extension of this thesis was the proposition that groups in society are always striving for or attempting to maintain their positions of power. One way that more powerful groups kept their positions of dominance was to define others groups as unconventional and in need of society's efforts of social control. This resulted in members from less powerful groups being more heavily represented in prisons, and mental hospitals, which, in turn, led to their greater representation among the homeless, jobless, and those on welfare.

Subsequent conflict theorists raised issues such as how women (Belknap, 2001), GLBTQs (Comstock, 1991), and ethnic minorities (Tonry, 1995) were not represented in traditional theories of crime and deviance because their criminal behavior was different from that of white men and their victimization was ignored, minimized, or ridiculed by official agents of social control. An example of how the State used social control to keep the downtrodden down was noted by Michelle Alexander: many ethnic minorities, particularly Blacks, have been, for all intents and purposes, removed from participating in society because they are imprisoned at high rates, and lose many civil rights as a result (e.g., voting, gun ownership, cannot run for office, have trouble getting jobs) (Alexander, 2010). Theoretical discourse from this perspective, then, clearly illustrates the great mistrust sociologists of the time had for those in powerful positions.

For the last 300–400 years, sociological discourse on social control, crime, and deviance has branched out, providing numerous explanations for such behavior. Each theoretical exploration has undertaken a paradigmatic shift in the conceptualization of deviance and/or social control, or the methods used to study such. As such, perhaps the greatest recent contributions of sociologists to the study of social control and deviance are to use multifactor theories that pull a few of the most valid parts of several theories together, particularly those that examine both individual and society level factors. This also suggests that multimethod approaches will enhance our understanding of deviance. Using observation or ethnography as well as surveys or official data can all be combined to provide more complete accounts and descriptions of this important social phenomenon.

See also: Communities and Crime; Crime, Sociology of; Deterrence Theory: Crime; Prevention of Crime and Delinquency; Punishment: Comparative Politics of; Punishment: Social and Legal Aspects; Punishment; Religion: Morality and Social Control; Self-Control Theory and Crime; Social Ecology: The Chicago School; Social Learning Theory and Crime; Strain Theories and Crime.

Bibliography

Alexander, M., 2010. The New Jim Crow: Mass Incarceration in the Age of Color-Blindness. The New Press, New York.
Anderson, N., 1923. The Hobo: The Sociological Study of the Homeless Man. University of Chicago Press, Chicago.
Beccaria, Cesare, 1764. Of Crimes and Punishments. R. Bell, Philadelphia.
Becker, H.S., 1963. Outsiders, Studies in the Sociology of Deviance. The Free Press.
Belknap, J., 2001. The Invisible Woman: Gender, Crime, and Justice, second ed. Wadsworth, Belmont, CA.
Bentham, Jeremy, 1789. An Introduction to the Principles of Morals and Legislation. Clarendon Press, London.
Buzan, G., Lawson, G., 2014. Capitalism and the emergent world order. International Affairs 90 (1), 71–91.
Chambliss, W.J., 1973. The saints and the roughnecks. Society 11 (1), 24–31.
Clinard, M., 1957. Sociology of Deviant Behavior. Holt, Rinehart, & Winston, New York.
Cloward, R.A., Ohlin, L.E., 1960. Delinquency and Opportunity: A Theory of Delinquent Gangs. Free Press, Glencoe, IL.
Comstock, G.D., 1991. Violence against Lesbians and Gay Men. Columbia University Press, New York.
Comte, A., Bridges, J.H., 1865. A General View of Positivism. Trubner and Co, London.
Cressey, P.G., 1932. The Taxi-Dance Hall: A Sociological Study in Commercialized Recreation and City Life. University of Chicago Press, Chicago.
Drake, St C., Cayton, H.R., 1945. Black Metropolis: A Study of Negro Life in a Northern City. University of Chicago Press, Chicago.
Durkheim, E., 1893 (W.D. Halls, Trans. 1997). In: Coser, Lewis A. (Ed.), The Division of Labour in Society. Free Press, New York.
Durkheim, E., 1897. Suicide: A Study in Sociology. Free Press, New York.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Anchor Books, New York.
Hirschi, T., 1960. Causes of Delinquency. University of California Press.
Landesco, J., 1929. Organized Crime in Chicago. Illinois Association for Criminal Justice, Chicago.
Lemert, E., 1951. Social Pathology: A Systematic Approach to the Study of Sociopathic Behavior. McGraw-Hill, New York.
Liazos, A., 1972. The poverty of the sociology of deviance. Nuts, Sluts, and Preverts Social Problems 20 (1), 103–120.
Lombroso, C., 1876. L'Uomo deliquente. Published in English as Criminal Man According to the Classification of Cesare Lombroso (1911). G.P. Putnam's Sons, New York.
Mannheim, H., 1965. Comparative Criminology, vol. 2. Houghton Mifflin, Boston.
Marx, K., 1848. Manifesto of the Communist Party. The Communist League London, England.
Marx, K., Engles, F., 1848. The Communist Manifesto (S. Moore, D. McClellan, Trans. 1992). Oxford University Press, Oxford.
Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3 (5), 672–682.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, Oxford.
Newman, G., 1978. The Punishment Response. Lippincott, Philadelphia.
Nye, F.I., 1958. Family Relationships and Delinquent Behavior. John Wiley, New York.
Park, R.E., Burgess, E., McKenzie, R.D., 1925. The City: Suggestions for the Investigation of Human Behavior in the Urban Environment. University of Chicago Press, Chicago.
Quinney, R., 1970. The Social Reality of Crime. Transaction Publishers, Rutgers, NJ.
Reckless, W., 1933. Vice in Chicago. University of Chicago Press, Chicago.

Reiman, J., 1979. The Rich Get Richer and the Poor Get Prison. Allyn & Bacon, Boston, MA.
Sampson, R., 2012. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.
Schwendinger, H., Schwendinger, J., 1977. Social class and the definition of crime. Crime and Social Justice 7, 4–13.
Sellin, J.T., 1950. Some Current Issues in Penal Treatment. Hæggström, Stockholm.
Shaw, C., McKay, H.D., 1942. Juvenile Delinquency in Urban Areas. University of Chicago Press, Chicago.
Sutherland, E.H., Cressey, D.R., 1934. Principles of Criminology. AltaMira Press, Lanham, MD.
Sutherland, E.H., 1934. The decreasing prison population of England. Journal of Criminal Law and Criminology 24 (5), 880–900.
Tarde, G., 1890. La Philisophie Pénal (H. Rapelje, Trans. 1968. Penal Philosophy). William Heinemann, London.
Thomas, W.I., Zanaiecki, F., 1920. The Polish Peasant in Europe and America. University of Chicago Press, Chicago.
Thrasher, F.M., 1929. The Gang: A Study of 1313 Gangs in Chicago. University of Chicago Press, Chicago.
Toby, J., 1957. Social disorganization and stake in conformity: complementary factors in the predatory behavior of hoodlums. Journal of Criminal Law, Criminology and Police Science 48, 12–17.
Tonry, M., 1995. Malign Neglect: Race, Crime, and Punishment in America. Oxford University Press, Oxford.
Vold, George B., 1958. Theoretical Criminology. Oxford University Press, Oxford, England.
Weber, M., 1946. Essays in sociology. In: Weber, M., Gerth, H., Mills, C.W. (Eds.), From Max Weber. Oxford University Press, New York.
Websdale, N., 2010. Familicidal Hearts: The Emotional Styles of 211 Killers. Oxford University Press, Oxford.
Zorbaugh, H.W., 1929. The Gold Coast and the Slum: A Sociological Study of Chicago's Near North Side. University of Chicago Press, Chicago.

Social Democracy

Tom P Harsanyi, Harvard University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14267–14270, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The term 'Social Democracy' has been used with a very wide variety of meanings since its introduction in France prior to the Revolutions of 1848. At various times and places, the social-democratic label has been adopted by revolutionary Marxist, democratic-socialist, center-left/liberal, and even conservative-authoritarian political parties and tendencies, usually with ties to blue-collar ('proletarian') labor unions. Most commonly, 'social democracy' is used in connection with the practices and ideas of parties that have belonged to the contemporary Socialist International or its preceding organization, the Second International. As parties calling themselves 'social-democratic' have grown steadily more moderate over time, gradually becoming less and less anticapitalist and antiliberal, usage of the term 'social democracy' has evolved in a similar fashion. In the year 2000 'social democracy' generally refers to a vague set of essentially liberal, center-left ideological and policy orientations, often associated with parties that belong to the Socialist International, that involve some combination of a relatively free-market-based capitalist economy, a fairly well developed welfare-state, a liberal-democratic ('polyarchical') political system, and a relatively internationalist approach to questions of foreign policy, security, and globalization.

The term 'social democracy' can refer either to a form of societal organization, an ideology, a set of public policies, a political party or group of parties, or a broad social and political movement centered on such parties and their allies within different sectors of society. The meaning of 'social democracy' has varied enormously since its original adoption by a small, vaguely-socialist group within the left wing of the French democratic-republican opposition prior to the Revolutions of 1848. At various times and places, the social-democratic label has been adopted by revolutionary Marxist, democratic-socialist, liberal/center-left, and even conservative-authoritarian political parties and tendencies, usually but not always with closer or more distant ties to blue-collar labor unions. Most commonly, the term 'social democracy' is used either in connection with political parties that have belonged to the contemporary Socialist International or its predecessor, the Second International of 1889–1914, or with reference to policies and ideologies close to such parties.

As parties calling themselves social-democratic have grown steadily more moderate over time, gradually becoming less and less anticapitalist and antiliberal, usage of the term 'social democracy' has evolved in a similar fashion. At the turn of the twenty-first century, 'social democracy' generally refers either to the programs and policies of the largely center-left political parties that belong to the Socialist International, or to a vague set of essentially liberal, center-left ideological and policy orientations that are often associated with these parties. While specific policies may vary, such center-left 'social-democratic' approaches share a commitment to a relatively free-market-based capitalist economy, a fairly developed welfare-state, a liberal-democratic ('polyarchic') political system, and a relatively internationalist approach to questions of foreign policy, security, and globalization. All over the world, parties and ideologies of this type are currently among the most important in the vast majority of countries with relatively competitive elections. The term 'social democracy' is rarely any longer used as a self-description by Marxist or other anticapitalist and antiliberal tendencies.

The History and Significance of Social Democracy

Ever since the Revolutions of 1848, which are generally considered to mark the birth of the social-democratic movement, the history of social democracy has been characterized by a profound ideological and tactical struggle between relatively radical, anticapitalist tendencies and relatively moderate, reformist orientations. In this dialectical contest, the radicals have generally adhered to some form of State socialism, usually of a Marxist (but anti-Leninist) variety, and/or some version of decentralized and radical-democratic cooperative socialism ('self-management,' 'economic democracy'); the comparatively moderate reformists, by contrast, have increasingly made their peace with multiclass ('catch-all') parliamentary politics, liberal ideology, and the capitalist system. Over the decades, this internal struggle has interacted with the often complex relationships of social-democratic movements with other political, social, and economic forces within particular societies to give rise to a wide variety of more or less 'socialist' ideological visions and public policies, often very effective and/or highly original, which have directly or indirectly exerted a tremendous influence on virtually all aspects of political, social, economic, and intellectual life in the modern world, particularly in countries with relatively capitalist economic systems and relatively liberal-democratic political systems. Although social-democratic movements and governments have frequently been frustrated in their reformist efforts by powerful domestic or international pressures, the resulting compromises and modifications have often marked significant new developments in the history of their respective countries, and of the world as a whole. Indeed, even when confined to the ranks of the political opposition in a particular country, social democrats have sometimes played a major role in bringing about

political, social, or economic reforms, by means of political pressure, foreign example, or intellectual influence.

Among the most important of these social-democratic contributions to the development of the 'welfare-capitalist' synthesis that has gradually come to dominate an increasing portion of the world since the 1880s, have been the following: (1) the striking (albeit gradual) democratization and humanization of political, social, and economic life; (2) the enormous growth and modernization of both the state and the economy, particularly in the realms of social policy, industrial policy, macroeconomic policy, industrial and labor relations (the much-celebrated 'democratic corporatism' of various Northern European countries, now largely defunct in a more market-oriented age), and recently also environmental policy; (3) the more or less successful promotion of social harmony and ideological consensus within previously much more polarized societies; and (4) strong support for international integration and world peace (although the temptation to join in nationalistic ventures, on the one hand, or in pacifistic crusades, on the other hand, has been a frequent source of tension within the social-democratic movement ever since its inception, most notably during the First World War).

Although self-styled 'social-democratic' parties and ideas have played a leading role in the history of an increasing number of countries around the world since 1880s, the concept of 'social democracy' is generally couched in terms based on the experience of Northern European Social Democratic, Socialist, and Labor parties (particularly those of Sweden and the other Scandinavian countries, as well as Germany, Austria, Great Britain, the Netherlands, and Belgium), because they have usually been the leading forces associated with the 'social-democratic' label, whether in terms of ideological and tactical development, policy-innovation, electoral support, governmental influence, or international prestige.

To the extent that world history since 1848 has been the story of the grand struggle between capitalism and socialism, social-democratic movements and ideas have often played a decisive role in finding a more or less successful middle-ground between the two (frequently under the slogan of some sort of 'Third Way' between unbridled capitalism and state-socialism), particularly in Sweden, Norway, and the other countries of Northern Europe. And to the extent that world history during the same period has been the story of the grand struggle between democracy and dictatorship, social-democratic movements and ideas have contributed greatly to both sides: on the one hand, social democracy spawned radical, antiliberal offshoots that quickly evolved into the rival totalitarian movements of communism and fascism during and after the First World War; on the other hand, social democrats have often played a decisive role in resisting and defeating these totalitarian movements, as well as other antidemocratic political, social, economic, and intellectual forces.

From the 1945 onwards social-democratic parties all over the world have grown steadily more dependent on the success of the welfare-capitalist social-market economy and the support of middle-class voters and relatively privileged elites in both the public and private sectors. This has been the result of two major interrelated processes: first, the continuation of these parties' aforementioned gradual movement toward the liberal center of the ideological spectrum; and second, the search for new bases of support and new reasons for existence under the impact of the gradual economic and cultural disappearance and disaffection of these parties' traditional blue-collar industrial ('proletarian') base. The first of these processes, that of increasing ideological moderation, has largely been due to the combination of mass prosperity, the rise of the welfare-state, profound political and social democratization, the failure of Marxist ideas and State socialist regimes ('real-existing socialism'), the spread of international peace and integration, American political and cultural hegemony, and the all-but-inevitable habituation of social-democratic elites to existing power-structures. The second process, namely the disappearance and disaffection of social democracy's old blue-collar base, has stemmed primarily from the combination of four interrelated developments that are associated with relentless economic and social modernization: first, much greater levels of social and economic mobility and opportunity, partly as a result of successful social-democratic policies; second, the accelerated disintegration and dissolution of inherited social and cultural solidarities and cleavages, including class-consciousness ('individualization,' 'desegregation,' 'depillarization'); third, the drastic weakening of inherited attitudes, beliefs, and values ('postmodernization,' the rise of 'postmaterialist' or 'left-libertarian' values), particularly among highly-educated elites; and fourth, the transition from a predominantly 'Fordist' socioeconomic system based on mass production in 'smokestack' industries, to an increasingly 'postindustrial' economy based on much more sophisticated technology, a much older, more highly-skilled, and more highly-educated labor-force, a burgeoning service sector, much more widespread share-ownership, much higher levels of flexibility in all areas of economic and social life, much greater concern for environmental protection, and much higher levels of international interdependence, integration, and homogenization ('globalization').

The Challenges Facing Social Democracy Today

In the twentieth century, social-democratic parties and elites increasingly staked their claim to rule less on utopian visions of 'socialism' than on their ability to manage and modernize the welfare-capitalist social-market economy more fairly, more humanely, more democratically, and, above all, more competently and more efficiently, than their rivals on right and left. Yet, although egalitarian and statist alternatives to free markets and a fairly rapacious capitalism were in the year 2000 weaker and more uncertain than ever before, challenges from the populist and often vehemently antiliberal and antimodern right, which since the 1980s has increasingly drawn the support of traditionally social-democratic (or communist or Christian Democratic) working-class voters, are often close at hand. Thus social democracy is faced with the difficult task of reconciling the interests and aspirations of an economically and culturally ever-more diverse potential constituency, a task which may prove extremely difficult once the worldwide economic boom at the turn of the twenty-first century dissipates. The danger, then, is that social democracy may become caught between the conflicting demands of relatively successful or privileged and

relatively unsuccessful or underprivileged sectors within increasingly unequal and *anomic* postindustrial societies, with the attendant risk that important groups within either or both sectors may gravitate toward antiliberal and antimodern movements on the left or, more likely, the right.

Editor's Note

Interest in social democracy among political scientists and sociologists has surged since the collapse of the Soviet Union, since on the world-historical stage, social democracy now represents the only realistic alternative to neoliberal capitalism. A work of great importance in this resurgent interest is Anthony Giddens, The Third Way: The Renewal of Social Democracy, first published in 1998 by Polity Press but reissued many times since. Numerous other recent works of interest might be mentioned; among the more commonly cited are: Waller et al., 2013; Meyer and Hinchman, 2007; Sassoon, 2014/1996.

See also: Liberalism: Historical Aspects; Liberalism: Political Doctrine and Impact on Social Science; Political Parties, History of; Socialism: Historical Aspects; Socialism; Welfare State, History of.

Bibliography

Anderson, P., Camiller, P. (Eds.), 1994. Mapping the West-European Left. Verso, London.

Berman, S., 1998. The Social Democratic Moment: Ideas and Politics in the Making of Interwar Europe. Harvard University Press, Cambridge.

Esping-Andersen, G., 1985. Politics Against Markets: The Social Democratic Road to Power. Princeton University Press, Princeton, NJ.

Gamble, A., Wright, T., 1999. The New Social Democracy. Basil Blackwell, Oxford.

Gay, P., 1962. The Dilemma of Democratic Socialism: Eduard Bernstein's Challenge to Marx. Collier, New York (1952).

Giddens, A., 2000. The Third Way and Its Critics. Polity Press, Cambridge, UK.

Gillespie, R., Paterson, W. (Eds.), 1993. Rethinking Social Democracy in Western Europe. Frank Cass, London.

Kitschelt, H., 1994. The Transformation of European Social Democracy. Cambridge University Press, Cambridge, UK.

Lindemann, A., 1983. A History of European Socialism. Yale University Press, New Haven.

Marx, K., 1978. Critique of the Gotha Program. In: Tucker, R. (Ed.), The Marx--Engels Reader. W.W. Norton, New York (1875).

Merkel, W., 1993. Ende der Sozialdemokratie? Machtressourcen und Regierungspolitik im westeuropäischen Vergleich. Campus, Frankfurt am Main, Germany.

Meyer, T., Hinchman, L., 2007. The Theory of Social Democracy. Polity Press, Cambridge.

Notermans, T., 1999. Money, Markets, and the State: Social Democratic Economic Policies Since 1918. Cambridge University Press, Cambridge.

Padgett, S., Paterson, W., 1991. A History of Social Democracy in Postwar Europe. Longman, London.

Piven, F.F. (Ed.), 1991. Labor Parties in Postindustrial Societies. Polity Press, Cambridge.

Przeworski, A., 1985. Capitalism and Social Democracy. Cambridge University Press, Cambridge.

Sassoon, D., 1996. One Hundred Years of Socialism: The West-European Left in the Twentieth Century. The New Press, New York.

Sassoon, Donald, 2014/1996. One Hundred Years of Socialism: The West European Left in the Twentieth Century. I. B. Tauris, London.

Scharpf, F., 1991. Crisis and Choice in European Social Democracy. Cornell University Press, Ithaca (1987).

Waller, M., Coppieters, B., Deschouwer, K. (Eds.), 2013. Social Democracy in a Post-Communist Europe. Routledge, Oxon.

Social Dilemmas, Psychology of

Eric van Dijk, Leiden University, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In social dilemmas, people face the dilemma to further their own interest or the collective interest. This article reviews the literature on social dilemmas, by discussing the main types of social dilemmas, and the main findings of social dilemma research. The focus is on understanding the psychological process, and the possibilities to solve social dilemmas.

Introduction

In social dilemmas, people face the dilemma to further their own interests or to further the collective interest, knowing that both interests collide. Consider, for example, the problem of maintaining collective resources. Resources such as energy, oil, and water, can be regarded as collective resources that should be consumed wisely. In such situations, group members may face the dilemma that, despite the collective interest to restrict consumption, it may be in their personal interest to consume excessively.

The term 'social dilemma' was coined by Dawes (1980) when he presented the first overview of (mostly psychological) research on this mixed-motive setting. When Dawes wrote his overview, the conflict between personal and collective interest had already caught the attention from scientists from diverse disciplines as economics, political science, and sociology. After his review, the attention increased even further. Inspired by game theory, and their particular research interests, scientists developed several distinct dilemma types that all model different aspects of social dilemmas.

Social Dilemma Types

Individuals against Individuals: The Prisoner's Dilemma

The prisoner's dilemma can be considered as the prototype, or blueprint, of social dilemma research. The prisoner's dilemma was designed by Melvin Dresher and Merrill Flood, two scientists at the RAND Corporation, but received its name by Albert W. Tucker (see for an overview on the origins of the prisoner's dilemma game Poundstone, 1992). The dilemma is often used as a metaphor of the cold war, where two blocks (east and west) are in conflict while sharing a common interest (preserve peace). In prisoner's dilemma games, individuals can choose between furthering their own outcome (i.e., the defect option) and furthering their collective interest (i.e., the cooperation option). The collective outcomes are higher if both players cooperate. However, if both players only try to maximize their own outcomes, this potential will not be reached. For these reasons, the dilemma can also be described as a clash of 'individual rationality' (maximize the own outcomes) and 'collective rationality' (maximize the collective outcomes).

Individuals in Groups: The Public Good Dilemma and the Common Resource Dilemma

Whereas the prisoner's dilemma only describes a dyadic setting, the conflict can meaningfully be extended to settings involving more persons (i.e., n-Person dilemmas). Indeed, this is the setting that most scientists have in mind when referring to the social dilemma.

Samuelson (1954) was one of the first to analyze the provision of public goods as a social dilemma in which the interests of individuals are at odds with the collective to which they belong. In its purest from, public goods are characterized by two properties: jointness of supply (i.e., consumption of one does not go at the expense of the consumption of others) and nonexcludability (i.e., nobody can be excluded from consumption). Because the first property will rarely be observed in real life, research has focused on the second property. The fact that one cannot be excluded from consumption – even if one does not contribute to the provision of the public good – raises the question of how one could expect anyone to contribute. This is, for example, problematic in the case of public television, where you can enjoy the fruits (i.e., watch TV) without contributing to its provision. Self-interest alone would lead to underprovision of public goods. In experimental research, this public good dilemma (also referred to as the 'Give-some dilemma'; Dawes, 1980) is generally modeled by endowing group members with money that can be contributed to provide a monetary for the group that is divided equally, regardless of one's own contribution. The bonus is often a linear function of the individual contributions (a linear/continuous public good; e.g., all contributions are multiplied by a factor 2), or a fixed monetary bonus that only becomes available if the combined contributions surpass a certain threshold (a step-level public good).

In 1968, Hardin wrote his seminal paper on 'The Tragedy of the Commons' in which he analyzed the problem of overpopulation from the perspective of social dilemmas. As a metaphor for population growth, he described a group of herdsmen who let their herds graze on a common pasture. The freedom of increasing one's own herd in the end leads to overgrazing and thus to (tragic) destruction of the commons. This analysis inspired researchers to develop the common resource dilemma as a research paradigm where group members can harvest from a common resource that may partially replenish, but will become depleted if overused. With the added feature of partial replenishment, the common

resource dilemma (also referred to as the 'Take-some dilemma'; Dawes, 1980) incorporates the distinction between short-term gains (excessive harvests) and long-term losses (depletion of the resource). The dilemma is generally seen as being at the heart of real-life problems of overconsumption (e.g., depletion of natural resources; overfishing; see also Ostrom, 1990).

Groups against Groups: Intergroup Competition

A special case emerges if two groups are in conflict. Such a dilemma could be described as a prisoner's dilemma between groups if one restricts one's analysis at the group level. If, however, one also incorporates the presence of individuals within groups, the situation becomes even more complex. It is this setting that Bornstein aimed to model when he designed the Intergroup Prisoner's dilemma game (Bornstein, 2003). In this dilemma, two groups are in competition, while the individuals within each group have to decide whether or not to cooperate. If the individual group members contribute to their group, the result can be fierce intergroup competition. In such a setting, the collective interest (i.e., the interest that all group members of both groups) may be served if individual group members decide *not* to cooperate within their group. A prototypical example of this type of dilemma is the setting where two nations are in conflict. Cooperation within each nation to fight the enemy may lead them to engage in war, with the collective threat of many casualties on both sides.

Groups within Groups: Nested Social Dilemmas

With nested dilemmas, the problem of cooperation is extended to an even higher level of complexity. In nested dilemmas, you can find subgroups, which are part of a superordinate group. In large organizations, for example, such a nested setting would refer to employees being members of departments that are situated (nested) within a large organization. The hierarchical and nested structure creates complex interdependencies such that cooperation *within* departments may lead to competition *between* departments that in the end may go at the expense of the organization at large (Wit and Kerr, 2002). Nested dilemmas are widespread. On a larger scale, we see countries nested within continents that together have a common interest to address environmental problems (like global warming or pollution of the oceans).

Groups after Groups: Intergenerational Dilemmas

Intergenerational dilemmas capture both the temporal and group dimension, by addressing the dilemma that people face when deciding whether or not to cooperate to benefit future generations. Global environmental change is one of the main problems that shows how our current state is affected by decisions of the generations that preceded us, and how the decisions we make can affect future generations. In addition to the temporal aspect, that may induce a stronger focus on current than on future outcomes, intergeneration dilemmas may suffer from the low level of identification with those who come after us. One of the main researchers on the psychology of intergeneration dilemmas is Kimberley Wade-Benzoni, who described the intergenerational dilemmas as a case of power asymmetry, where the parties who have control over the decision process (the present generation) are not the parties with most at stake (the future generations) (Wade-Benzoni, 2002).

Understanding Behavior in Social Dilemma Behavior: Main Observations and Explanations

Social dilemma research has identified the crucial factors that can promote or hinder cooperation. Interestingly, some of these promote cooperation in one type of dilemma and decrease cooperation in others.

Group Size

In general, group size has been found to decrease cooperation. This negative relation is mostly attributed to the fact that with increased group size, group members see their individual contributions are less essential. Increased group size leads to diffusion of responsibility, and reduced feelings of efficacy such that in large groups individuals may feel their contribution does not really add much to promote the interest of the group (Kerr, 1996).

Communication

Communication has generally been found to promote cooperation. A recent meta-analysis by Balliet (2010) showed that this is especially true for face-to-face communication. Several mechanisms may account for this positive relation. Communication may promote mutual promise making (and keeping) and enable people to commit themselves to promoting the group interest. Communication may also promote identification with the group. The meta-analysis also indicated that the positive effect of communication is stronger in larger groups, which may be explained by its potential to increase within-group identification. Of course, the group size does pose a boundary condition for effective communication, as in large groups (e.g., groups with 1000 individuals) communication may not be feasible. In this respect, it is telling that the meta-analysis on communication only included groups with group sizes lower than 10. Also note that sometimes communication may change this for the worse, for example, when in intergroup dilemmas within-groups communication increases within-group cooperation, but as a result increases intergroup competition (Goren and Bornstein, 2000).

Tacit Coordination

Even when opportunities for communication are limited, group members may sometimes tacitly coordinate their decisions. This means that even without communication, group members converge in the making of their decision (see for an overview De Kwaadsteniet and Van Dijk, 2012). For tacit coordination to work, group members need to have the feeling that the members in their group have a common understanding about the dilemma they face. Compared to continuous dilemmas, step-level dilemmas offer better prospects for tacit coordination because the threshold may serve as a focal point for coordination (e.g., Abele et al., 2010). An often used

coordination rule is the equality rule, which in step-level public good dilemma prescribes that all members contribute an equal share of the provision threshold. In asymmetric dilemmas (i.e., dilemmas in which group members occupy different positions, for example, when some group members are wealthier than others), alternative coordination rules may emerge. In public step-level dilemmas, for example, people rely on the proportionality rule if positions differ in terms of wealth (prescribing that people should contribute in proportion to their wealth). In addition, research showed that the preference for equality is stronger in the resource dilemma than in the public good dilemma (Van Dijk and Wilke, 1995).

Group Identity

Considering that in social dilemmas, individuals choose between furthering the own interest or the group interest, it makes sense to assume that identification with one's group increases cooperation. Such a positive link between identification and cooperation is indeed often observed (e.g., Kramer and Brewer, 1984). It should be noted that in some types of dilemmas, group identification may hinder the collective interest. In nested social dilemmas, identification with one's subgroup may reduce cooperation with the superordinate level (Wit and Kerr, 2002). In intergroup dilemmas high identification with the own group may result in competition with other groups. In this sense, effects of identification may resemble those of communication (which may not come as a surprise because communication may increase identification). In intergenerational dilemmas the paths of communication and identification need not align. While it is impossible (in the absence of a time machine) to communicate with future generations, we may identify with those who come after us (Wade-Benzoni, 2002) and thereby achieve intergenerational cooperation.

Reciprocity

In single-trial prisoner's dilemmas, a motivation to maximize one's own outcomes is served by the relative straightforward strategy to defect rather than cooperate. This may not hold in multitrial settings, however. Repeated interactions may make cooperation a more advantageous strategy by allowing for reciprocity (Gouldner, 1960), even in the absence of communication. The beneficial consequences of reciprocity were, for example, demonstrated in a tournament, organized by Robert Axelrod (Axelrod, 1980). He invited game theorists from economics, sociology, political science, and mathematics to send him their computerized strategies for a series of repeated prisoner dilemma games. All the submitted strategies, plus an added strategy of randomness, played 200 trials of a prisoner's dilemma against each other, the winner being the strategy that accumulated the highest outcomes. The glorious winner of the tournament was the Tit-For-Tat (TFT) strategy, submitted by Anatol Rapoport. TFT starts off with cooperation, but from then on reciprocates by choosing the same option as the opponent did in the previous trial (i.e., it responds to cooperation with cooperation and noncooperation with noncooperation). Whereas studies on multiple trials indeed show the emergence of mutual cooperation, one may also often observe a decline in cooperation toward the end of the trials (Murnighan and Roth, 1983). Reciprocity is even possible in intergenerational settings, when the current generation, for example, feels an obligation to respond cooperative to the cooperation of a previous generation (Wade-Benzoni, 2002).

Reciprocal strategies may promote mutual cooperation. In more recent years, building on evolutionary insights, the field has also identified reciprocal strategies other than 'just' direct reciprocity between pairs. Indirect reciprocity occurs when we cooperate with people because they cooperated with others. This mechanism works via reputation effects (Wedekind and Milinski, 2000): Building a positive reputation by cooperation may in the end bring you higher outcomes if others select you as the partner they want to cooperate with. Relatedly, cooperation has been interpreted as a costly signal that may attract others by sending the message that one is well-off *and* willing to help others (Van Vugt et al., 2007).

Environmental Uncertainty and Noise

Environmental uncertainty refers to uncertainty about features of the social dilemma, for example, about the exact size of the resource in a common resource dilemma, the threshold in a public good dilemma, or the group size dilemma (see for an overview Van Dijk et al., 2004). Cooperation usually drops under environmental uncertainty. The main explanations for this drop are that people may suffer from biased estimates (e.g., people harvest a lot because they overestimate the size of the resource in a resource dilemma), or opportunistically 'use' the uncertainty to justify their noncooperation (e.g., people justify their excessive harvests by arguing they thought the resource size would be large). In addition, uncertainty may reduce opportunities for tacit coordination (e.g., when the threshold becomes uncertainty) and lead to noncooperation by reducing people's sense of having a common understanding on what to do (De Kwaadsteniet and Van Dijk, 2012).

In more recent years, researchers have begun to study the effects of 'noise,' which basically refers to the difference between intentions and behavior, i.e., to when our intentions do not match with our behavior. Positive noise refers to unintentional cooperation; negative noise refers to unintentional noncooperation. Negative noise is the most severe threat to cooperation because people tend to respond negatively to noncooperation (due to the norm of reciprocity). In the case of negative noise, such a reaction of noncooperation would not be justified. One isolated instance of negative noise may, in 2-party interactions, trigger a chain of noncooperation if people adhere to a strict TFT-like strategy. The solution to this may be to adhere to a more lenient strategy, described as TFT + 1 (i.e., always be a little bit more cooperative than your partner; Van Lange et al., 2002).

Personality

People differ on many things, so it makes sense to assume that they may also differ in the weights they put on their own versus the collective interest. The different weighing of own and others' outcomes is central in research on social value orientations (Messick and McClintock, 1968). This research has identified three main personality types with regard to the

weighing of own and others' outcomes. Most people are competitive, individualistic, or cooperative (also referred to as prosocial); see for the most commonly used method to assess these personality types Van Lange et al. (1997). Individualists are primarily concerned about their own outcomes while competitors and prosocials do consider what happens to others. The competitor's interest is primarily in being better-off than others. Prosocials, are motivated to increase both the own and others' outcomes. A meta-analysis by Balliet et al. (2009) on 82 studies showed that these social value orientations strongly affect cooperation across different types of dilemmas, with prosocials consistently being more cooperative than competitors and individualists.

For dilemmas that include a temporal component, another personality difference seems of special relevance. People differ in their consideration of future consequences (Strathman et al., 1994). The general finding is that those with an eye on the future are more willing to cooperate, for example, by restricting their harvests in a common resource dilemmas to ensure future consumption (Joireman et al., 2009). It stands to reason that similar findings may be observed in intergenerational dilemmas, or other dilemmas with repeated encounters.

Solving Dilemmas

Understanding social dilemma behavior is one thing, solving social dilemmas is another. From the earliest studies on, the focus has been on how cooperation can be increased. Unfortunately, not all factors lend themselves to finding a solution. Personality differences should usually be seen as a given, and thus as a factor to be reckoned with, but not necessarily as key to solving dilemmas. Other factors may more readily be connected to solving social dilemmas. A distinction can be made between individual solutions and structural solutions (Messick and Brewer, 1983). Individual solutions more or less follow directly from the factors discussed above. Individual solutions may, for example, be to lower the group size, promote communication (either with or between groups), increase social identity or identification with the collective, and promote the feeling of having a common understanding (to promote tacit coordination).

Structural solutions are both more drastic and more promising in the sense that they address the heart of the problem: The outcomes structure of the dilemma. If social dilemmas are defined by their interdependence structure (which they are), it makes sense to change the interdependence structure. In their most extreme form, structural solutions may solve the dilemma by eliminating the (negative) interdependence between group members, and/or by making cooperation more attractive than noncooperation.

Leadership

The main conclusion in Hardin's (1968) seminal article on the Tragedy of the Commons was that to prevent the tragedy (of overpopulation) we need institutions that restrict individual freedom (as Hardin described it, we need "mutual coercion, mutually agreed upon"). One of the possible options that the early social dilemma research focused on was the installment of (autocratic) leadership. The installment of leadership (e.g., a central authority) to distribute a common resource over all group members not only removes the individual freedom to decide; it essentially restructures the social dilemma to a non-dilemmatic allocation setting. Such a solution can, of course, be very effective. But the solution to replace individual decision freedom by centralized leadership immediately raises the question of whether group members are prepared to freely give up their decisional freedom. Research shows that group members may (only) be willing to do so after high variance (i.e., when some members are a lot less cooperative than others), and collective failure (e.g., when the common resource becomes depleted) (e.g., Messick et al., 1983). It is interesting to see that preferences for leadership are also dependent on the dilemma structure. In general, the preference for leadership may be stronger in resource dilemmas than in public good dilemma. The call for leadership may be higher in the public good dilemma, however, after collective failure (i.e., if group members realize their individual contributions to the public good have been in vain; Van Dijk et al., 2003).

Sanctions

The most propagated solution to prevent noncooperation and to promote cooperation is to use sanctions. Sanctions can either take the form of penalties for noncooperation (negative sanctions), or rewards for cooperation (positive sanctions). Both rewards and punishments can be effective instruments (Balliet et al., 2011). Under some conditions, however, sanctions may countereffective (and promote noncooperation instead of cooperation). In particular, sanctions may undermine the intrinsic motivation to cooperate, by invoking a business frame. Rather than evaluating settings in ethical terms, sanctions may lead people to evaluate their decisions in terms of what they get out of it themselves (Tenbrunsel and Messick, 1999). If sanctions are (too) low, this may lead to a decrease of cooperation. Sanctions may also lower interpersonal trust, because the introduction of sanctioning system may signal to people that apparently others cannot be trusted to be internally motivated to promote the collective interest (why else would one need a sanctioning system?). Sanctions, may thereby create their own need, such that cooperation may collapse if a once installed sanctioning system is lifted (Mulder et al., 2006).

But even when sanctions are effective, one may wonder to what extent people are willing to install or use such sanctions. Here, the distinction between centralized sanctions and decentralized sanctions is important. Centralized sanctions are installed by one central authority; decentralized sanctions are sanctions that group members impose on each other (Balliet et al., 2011). The early studies on preferences for using sanctions primarily focused on centralized sanctions, and, for example, revealed that people are most likely to contribute to the installment of a sanctioning system if they both have the goal to further the collective outcome, *and* fear that others will not cooperate (Yamagishi, 1986). More recently, research has shifted toward the willingness to use decentralized sanctions. This shift was inspired by a study by Fehr and Gächter (2002). In a public goods setting, they assigned participants to 10 trials with and 10 trials without a sanctioning system. With

a sanctioning system, participants could reduce the outcomes of those who did not contribute. This punishment was not only costly to the punished person, but also to the punisher. The results showed the effectiveness of costly sanctions: Without a possibility to sanction, contributions to the public good were low, and even reduced over trials. With a possibility to sanction, contributions were high, and remained high. Apart from demonstrating the effectiveness of costly sanctioning, this study thus clearly revealed that people are willing to punish their noncooperative fellow group members, even if this means that they lower their own outcomes.

The above primarily highlights the instrumentality motive people may have to use sanctions (i.e., to promote cooperation). Whereas this may suggest that people primarily base their decisions on the (positive) expectations about sanction effects (e.g., I sanction to increase cooperation), emotions – and in particular anger – have been identified as a proximate mechanism (e.g., I punish because I am angry; e.g., Fehr and Gächter, 2002).

Trends in Social Dilemma Research

The previous overview illustrates the wide variety of issues that are covered by, and studied in, social dilemma research. We see the social dilemma logic being applied to diverse topics like intergroup conflict, public good provision, and management of scarce resources, to small-scale and large-scale settings, and to instant and temporal settings. The research on social dilemmas – being inspired by game theory – has traditionally made extensive use of experimental research with clearly defined paradigms. One of the trends has been to develop new paradigms to capture the complexities of social dilemmas we observe in real life (see, e.g., the intergroup dilemma, or the nested dilemma). Inspired by Nobel Prize winner Elinor Ostrom, we also witness more studies in the field. Another trend is that researchers from different disciplines more and more team up in their search for solutions to the social dilemma (see, e.g., the integration of insights from economics, social psychology, and evolutionary theory in the literature on sanctions). Combined, these trends will no doubt contribute to a better understanding of how we can manage the conflict between our own interests and the interests of the collective(s) to which we belong.

See also: Attitude Formation and Change; Cooperation and Competition; Decision Making, Psychology of; Leadership; Social Psychology.

Bibliography

Abele, S., Stasser, G., Chartier, C., 2010. Conflict and coordination in the provision of public goods; a conceptual analysis of continuous and step-level games. Personality and Social Psychology Review 14, 385–401.

Axelrod, R., 1980. Effective choice in the prisoner's dilemma. Journal of Conflict Resolution 24, 3–25.

Balliet, D., 2010. Communication and cooperation in social dilemmas: a meta-analytic review. Journal of Conflict Resolution 54, 39–57.

Balliet, D., Parks, C.D., Joireman, J., 2009. Social value orientation and cooperation: a meta-analysis. Group Processes and Intergroup Relations 12, 533–547.

Balliet, D., Mulder, L.B., Van Lange, P.A.M., 2011. Reward, punishment, and cooperation: a meta-analysis. Psychological Bulletin 137, 594–615.

Bornstein, G., 2003. Intergroup conflict: individual, group and collective interests. Personality and Social Psychology Review 7, 129–145.

Dawes, R.M., 1980. Social dilemmas. Annual Review of Psychology 31, 169–193.

De Kwaadsteniet, E.W., Van Dijk, E., 2012. A social-psychological perspective on tacit coordination: how it works, when it works, (and when it does not). European Review of Social Psychology 23, 187–223.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Goren, H., Bornstein, G., 2000. The effects of intragroup communication on intergroup cooperation in the repeated prisoner's dilemma (IPD) game. Journal of Conflict Resolution 44, 700–719.

Gouldner, A.W., 1960. The norm of reciprocity: a preliminary statement. American Sociological Review 25, 161–178.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Joireman, J., Posey, D.C., Truelove, H.B., Parks, C.D., 2009. The environmentalist who cried drought: reactions to repeated warnings about depleting resources under conditions of uncertainty. Journal of Environmental Psychology 29, 181–192.

Kerr, N.L., 1996. "Does my contribution really matter?". Efficacy in social dilemmas. In: Stroebe, W., Hewstone, W. (Eds.), European Review of Social Psychology, vol. 7. Wiley, Chichester, UK, pp. 209–240.

Kramer, R.M., Brewer, M.B., 1984. Effects of group identity on resource use in a simulated commons dilemma. Journal of Personality and Social Psychology 46, 1044–1056.

Messick, D.M., Brewer, M.B., 1983. Solving social dilemmas: a review. In: Wheeler, L., Shaver, P. (Eds.), Review of Personality and Social Psychology. Sage, Beverly Hills, pp. 11–44.

Messick, D.M., McClintock, C.G., 1968. Motivational basis of choice in experimental games. Journal of Experimental Social Psychology 4, 1–25.

Messick, D.M., Wilke, H., Brewer, M.B., Kramer, R.M., Zemke, P.E., Lui, L., 1983. Individual adaptions and structural change as solutions to social dilemmas. Journal of Personality and Social Psychology 44, 294–309.

Mulder, L.B., Van Dijk, E., De Cremer, D., Wilke, H.A.M., 2006. Undermining trust and cooperation: the paradox of sanctioning systems in social dilemmas. Journal of Experimental Social Psychology 42, 147–162.

Murnighan, J.K., Roth, A.E.R., 1983. Expecting continued play in prisoner's dilemma games: a test of several models. Journal of Conflict Resolution 27, 279–300.

Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, New York.

Poundstone, W., 1992. Prisoner's Dilemma. Oxford University Press, New York.

Samuelson, P.A., 1954. The pure theory of public expenditure. Review of Economics and Statistics 36, 387–389.

Strathman, A., Gleicher, F., Boninger, D.S., Edwards, C.S., 1994. The consideration of future consequences: weighing immediate and distant outcomes of behavior. Journal of Personality and Social Psychology 66, 742–752.

Tenbrunsel, A.E., Messick, D.M., 1999. Sanctioning systems, decision frames, and cooperation. Administrative Science Quarterly 44, 684–707.

Van Dijk, E., Wilke, H., 1995. Coordination rules in asymmetric social dilemmas: a comparison between public good dilemmas and resource dilemmas. Journal of Experimental Social Psychology 31, 1–27.

Van Dijk, E., Wilke, H., Wit, A., 2003. Preferences for leadership in social dilemmas: public good dilemmas versus resource dilemmas. Journal of Experimental Social Psychology 39, 170–176.

Van Dijk, E., Wit, A., Wilke, H., Budescu, D.V., 2004. What we know (and do not know) about the effects of uncertainty on behavior in social dilemmas. In: Suleiman, R., Budescu, D.V., Fischer, I., Messick, D.M. (Eds.), Contemporary Psychological Research on Social Dilemmas. Cambridge University Press, pp. 315–331.

Van Lange, P.A.M., Otten, W., De Bruin, E.C.M., Joireman, J.A., 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. Journal of Personality and Social Psychology 73, 733–746.

Van Lange, P.A.M., Ouwerkerk, J.W., Tazelaar, M.J.A., 2002. How to overcome the detrimental effects of noise in social interaction: the benefits of generosity. Journal of Personality and Social Psychology 82, 768–780.

Van Vugt, M., Roberts, G., Hardy, C., 2007. Competitive altruism: development of reputation-based cooperation in groups. In: Dunbar, R., Barrett, L. (Eds.), Handbook of Evolutionary Psychology. Oxford University Press, Oxford.

Wade-Benzoni, K.A., 2002. A golden rule over time; reciprocity in intergenerational allocation decisions. Academy of Management Journal 45, 1011–1028.

Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.

Wit, A.P., Kerr, N.L., 2002. "Me versus just us versus us all" categorization and cooperation in nested social dilemmas. Journal of Personality and Social Psychology 83, 616–637.

Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51, 110–116.

Social Dominance Orientation

Andrew L Stewart, Clark University, Worcester, MA, USA
Felicia Pratto, University of Connecticut, Storrs, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social dominance orientation (SDO) is an individual's preference for group-based inequality and dominance versus equality and inclusion. SDO has multiple origins, including status, gender, socialization, and temperament and personality. SDO is a strong predictor of intergroup ideologies and behaviors, making it indispensable for studying intergroup relations.

Social dominance orientation (SDO) is an individual difference variable that explains sociopolitical attitudes, prejudice, and intergroup ideologies. As a part of social dominance theory (Sidanius and Pratto, 1999), SDO is an individual's preference for group-based inequality and dominance versus equality and inclusion (Pratto et al., 1994). A person's SDO compels them to adopt intergroup ideologies that either increase intergroup inequality (i.e., hierarchy-enhancing ideologies such as racism) or decrease intergroup inequality (i.e., hierarchy-attenuating ideologies such as feminism). Thus, it attempts to explain – at the psychological/individual level – why people adopt various intergroup attitudes or beliefs that can impact intergroup inequality at the societal level. As a general orientation toward intergroup relations, SDO is relevant across time and cultures (Lee et al., 2011; Pratto et al., 2014).

Origins of Social Dominance Orientation

SDO has multiple origins and causes (Kteily et al., 2011; Sidanius and Pratto, 1999). In the initial theorizing, it was proposed to have at least four causes: (1) group status, (2) socialization, (3) temperament, and (4) gender (Sidanius and Pratto, 1999). People who have high group status (e.g., European Americans, Pakeha in New Zealand, or Brahmins in India) have an interest in dominating surbordinate groups and maintaining their dominant social position, so they tend to be higher on SDO than people who have low group status (e.g., African Americans, Maori in New Zealand, or Dalits in India). Thus, group status is strongly related to SDO (Pratto and Stewart, 2012). Further, a wide variety of socialization processes reinforce which social groups should be at the top of the hierarchy and which social groups should be at the bottom. Media (news, television), educational experiences, social learning, and other factors communicate as to where social groups should be placed in a group-based social hierarchy, which can teach people to desire group-based inequality or equality. Also, some research suggests that different temperaments at birth give rise to differences in personality later in life (Sibley and Duckitt, 2008). These personality differences, such as openness to experience or agreeableness, can lead to different levels of SDO (Sibley and Duckitt, 2008). In a meta-analysis, SDO was related to agreeableness and also openness to experience (Sibley and Duckitt, 2008). Finally, men tend to be higher on SDO than women (i.e., the invariance hypothesis; Sidanius et al., 1994), so gender is an important predictor of SDO (Lee et al., 2011). Correlational research has established the relationships between SDO and these four factors, but greater research is needed to explore its origins.

In the most extensive theoretical treatment of the origins of SDO, Sibley and Duckitt (2008) argue that the development of a toughminded personality increases the development of a 'competitive jungle' worldview, where the world is seen as a dog-eat-dog world where people compete over material and symbolic resources. People's competitive jungle worldview then increases their desire for group-based dominance and inequality. Thus, SDO is proposed to be a function of personality traits and the development of a view of the world as competitive where every person must fight for themselves (Duckitt et al., 2002). Research examining the relationships among personality traits, cultural worldviews, and SDO provides evidence supporting this model (Duckitt et al., 2002).

In all, theory and research on SDO suggest that people's desire for intergroup inequality versus equality has multiple origins and causes, rooted in the development of personality and cultural worldviews.

Construct Validity of SDO

An important question since the theoretical development of SDO is 'What is social dominance orientation?' The construct validity of SDO has been questioned by social identity and system justification theorists (see Ho et al., 2012; Jost and Thompson, 2000; Kteily et al., 2011; Lee et al., 2011).

Since its inception, social identity theorists have criticized the construct of SDO (see Kteily et al., 2011; Lee et al., 2011), characterizing it as a mere reflection of social structure and intergroup attitudes. In contrast to this perspective, Kteily et al. (2011) argue that it is useful to view SDO as a cause and as an effect of intergroup attitudes. In a longitudinal study, participants' negative affect toward outgroups (e.g., Asians, Latinos, Blacks), ingroup friendship preference (e.g., a White person preferring to be friends with another White person versus a Black person), and SDO were measured in 1996 and once again in 2000. The results demonstrate that participants' initial levels of SDO strongly predicted outgroup affect and ingroup friendship preference 5 years later. In contrast, ingroup friendship preference in 1996 did not predict SDO 5 years later, and outgroup affect in 1996 predicted SDO weakly 5 years later (Kteily et al., 2011). This study provides evidence that SDO is a cause of intergroup attitudes and prejudice, but can also be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24093-2

affected by prejudice, though weakly. The results from this study directly contradict the hypothesis that SDO is a mere reflection of social structure and intergroup attitudes and is instead a primary driver of intergroup attitudes.

A second issue raised by social identity theorists is that SDO is a mere proxy for ingroup identification (see Lee et al., 2011). Thus, one hypothesis derived from this perspective is that nations that are motivated to avoid uncertainty should display larger differences in SDO than nations that are not motivated to avoid uncertainty. It has long been demonstrated that uncertainty increases ingroup identification, so if SDO is a mere proxy for ingroup identification, then there should be larger group differences on SDO in nations high on uncertainty avoidance. However, in a meta-analysis investigating 206 samples from 1992 to 2009, nations that are motivated to avoid uncertainty have smaller differences between groups in SDO than nations that are not motivated to avoid uncertainty (Lee et al., 2011). Thus, the meta-analysis demonstrates the exact opposite pattern predicted by social identity theorists who claim that SDO is a proxy for ingroup identification. These results demonstrate that SDO is a unique predictor of intergroup attitudes and is not redundant with ingroup identification or other constructs derived from other theories of intergroup relations.

A third issue with the construct validity of SDO derives from system justification theorists, who characterize SDO as two factors, not one (Jost and Thompson, 2000; Ho et al., 2012). According to system justification theory, people are motivated to justify the self (ego justification), the group (group justification), and the systems in which they live (system justification). For subordinate group members (e.g., Blacks in the United States), there is a conflict between their motivation to see their groups positively and to see their systems positively because the system is oppressive. In order to manage this dissonance, subordinate group members rationalize and justify the system, viewing it positively. In line with this reasoning, Jost and Thompson (2000) argue that the SDO scale (a 16-item questionnaire designed to measure people's desire for group-based dominance versus equality) consists of two factors, viz., group-based dominance and opposition to equality. Group-based dominance is hypothesized to correspond to group justification, while opposition to equality is hypothesized to correspond to system justification. In correlational research, Jost and Thompson (2000) find that these two factors correlate more strongly for European Americans than for African Americans, and they differentially relate to psychological well-being, ethnocentrism, and support for social policies between the two different social groups. However, in the meta-analysis mentioned earlier, Lee et al. (2011) find results that contradict Jost and Thompson (2000). There were no differences between men and women on group-based dominance and opposition to equality, and the correlation between these two subscales was high, $r = 0.47$ (Lee et al., 2011). In ethnic/racial groups, there was a difference on opposition to equality ($r = 0.27$) but not on group-based dominance ($r = -0.08$), and these two subscales were correlated highly, $r = 0.83$ (Lee et al., 2011). System justification theory would predict a difference on group-based dominance and not on opposition to equality, so the findings in ethnic/racial groups also contradict the predictions of system justification theory. The results of this meta-analysis strongly suggest that SDO is a unified factor and does not consist of two subfactors.

Despite the results from this meta-analysis, some social dominance theorists have started to think of SDO in terms of two factors, providing a contradictory message in the literature. For example, Ho et al. (2012) argue that SDO consists of two factors, viz., SDO-D (dominance) and SDO-E (equality). Examining seven samples of participants, the authors find that SDO-D and SDO-E, though highly correlated with one another, differentially predict a series of criterion variables (Ho et al., 2012). For example, SDO-D relates to intergroup competition and aggression (e.g., immigrant persecution or nationalism), while SDO-E relates to subtle legitimizing myths and support for redistributive social policies (e.g., welfare policies and the Protestant work ethic). One explanation for this differential prediction is that there exist different psychological orientations underlying a general construct of SDO, as Ho et al. (2012) argue. However, another methodological explanation is that the actual items on the SDO have a semantic confound. For example, items used to measure SDO-E tend to be con-trait items, viz., items that measure being low on SDO (e.g., 'Group equality should be our ideal'; Pratto et al., 1994). This semantic confound between the SDO-D and SDO-E parts of the scale was addressed in a study using confirmatory factor analysis (Xin and Chi, 2010). In this analysis, the 16 items used to measure SDO were specified to load onto a single latent variable, SDO. A second latent variable was specified and consisted of the con-trait items. In a sample of 370 participants from the United States, Canada, and China, the results demonstrated support not only for a unified construct of SDO but also for a second factor controlling for the semantic confound. This analysis suggests that the SDO scale does measure a unified construct, but that the two supposed halves consist of a semantic confound. Further research is necessary to develop an SDO scale that does not have a semantic confound between the two halves of the scale.

Overall, there is strong evidence that SDO is a unified construct and is neither a proxy for ingroup identification nor a mere reflection of social structure and intergroup attitudes.

The Psychology of Low SDO

Although most of the research on SDO has explored the characteristics of people who are high on SDO (e.g., prejudiced, discriminatory, and punitive), very little research has explored the psychological characteristics of people who are low on SDO (Pratto et al., 2013). Toward this end, Pratto et al. (2014) conducted several studies to understand the psychology of people who desire equality and inclusion. In general, people who are low on SDO tend to be inclusive of social groups. These people tend to include more groups, people, and animals in their scope of moral concern (i.e., they care about the lives and well-being of all humans and nonhuman animals), be more empathic and express more empathy, support policies that include groups (e.g., extending marriage rights to gays and lesbians), and reject policies that exclude groups (e.g., more border patrols and fences at the U.S.–Mexico border). Further,

these people view their religious practices (e.g., fasting, praying) as a way to include other people in their faith, and they dislike group boundaries, segregation, competition, and rigid status differences. In all, people who are low on SDO are broadly inclusive.

Intergroup Ideologies and Intergroup Behavior

Social dominance orientation, as a general orientation toward intergroup relations, has tremendous predictive utility in understanding the commonality among a wide variety of intergroup ideologies and intergroup behaviors (Sidanius and Pratto, 1999). Intergroup ideologies are prescriptive beliefs that describe how people and institutions should behave (Pratto et al., 1999). Because ideologies correspond greatly to intergroup behavior, they are strong predictors of discrimination, collective action, and other intergroup behaviors (Pratto, 1999). However, ideologies are highly variable across cultures and across time (Lee et al., 2011). For example, the racism observed in contemporary U.S. society can be considered to be different from that observed in different periods of U.S. history (e.g., old-fashioned racism; Ho et al., 2012). Because of the cultural and historical variability in intergroup belief systems, SDO is an incredibly useful variable used to explain and to predict these vastly different belief systems (Lee et al., 2011). The SDO scale does not specify any particular groups and its content is temporally stable (i.e., groups exist across all historical periods), so SDO is a useful variable in understanding intergroup relations across time and cultures (Lee et al., 2011). Contrary to some researchers (e.g., Duckitt et al., 2002; Sibley and Duckitt, 2008), SDO is not an ideology because it is not historically or culturally specific. Instead, it explains the commonality of intergroup ideologies, which is that some ideologies increase intergroup inequality whereas other ideologies decrease inequality.

Social dominance orientation is useful in identifying an ideology's functional type (Sidanius and Pratto, 1999). Ideologies have two functional types: hierarchy-enhancing and hierarchy-attenuating. Hierarchy-enhancing ideologies (e.g., racism and sexism) promote intergroup behaviors that increase inequality between social groups (e.g., gender discrimination in the workplace or hate crimes). Hierarchy-attenuating ideologies (e.g., feminism and socialism) prescribe intergroup behaviors that decrease inequality between social groups (e.g., feminist activism or limiting commodity fetishism). People who are high on SDO tend to accept hierarchy-enhancing ideologies, and people who are low on it tend to accept hierarchy-attenuating ideologies (Pratto et al., 1994). For example, people who are high on SDO tend to provide more merit-based reasoning (e.g., Protestant work ethic) when deciding to allocate resources to different social groups, and people who are low on SDO tend to provide more need-based reasoning when allocating resources (Pratto et al., 1999). These ideologies provide people with the reasoning or justification to allocate resources to social groups and to engage in discrimination (Pratto et al., 1999). Also, SDO relates positively to racism, heterosexism, sexism, nationalism, and other hierarchy-enhancing ideologies, but it correlates negatively to support for progressive social policies (Lee et al., 2011).

Conclusion

Social dominance orientation is an important individual difference variable in the study of intergroup relations. SDO predicts a wide variety of ideologies and intergroup behaviors, is culturally and historically general, and offers unique insights into people's attitudes toward intergroup relations and social structure. In general, SDO is people's desire for group-based dominance and inequality versus inclusion and equality. It is determined from temperamental differences in infancy, socialization processes, group status, and gender, and the development of a competitive jungle worldview leads people to adopt a higher level of SDO. SDO is arguably a unified factor and a causal agent in driving intergroup attitudes and behaviors. People who are low on SDO tend to be broadly socially inclusive and desire social policies that include social groups in a society's scope of moral concern and justice. Finally, SDO is an incredibly powerful predictor of a wide variety of ideologies and intergroup behaviors, which makes it indispensable in understanding intergroup relations across time and cultures.

See also: Authoritarian Personality; Collective Action; Collective Behavior, Social Psychology of; Cooperation and Competition; Cross-Cultural Psychology; Identity and Identification, Social Psychology of; Intergroup Relations; Justice: Social Psychological Perspectives; Mass Killings and Genocide, Psychology of; Political Psychology; Racism: Social Psychological Perspectives; Social Identity in Social Psychology; Social Psychology; Stereotypes in Social Psychology.

Bibliography

Duckitt, John, Wagner, Claire, du Plessis, Ilouize, Birum, Ingrid, 2002. The psychological bases of ideology and prejudice: testing a dual process model. Journal of Personality and Social Psychology 83, 75–93.

Ho, Arnold, Sidanius, Jim, Pratto, Felicia, Levin, Shana, Thomsen, Lotte, Kteily, Nour, Sheehy-Skeffington, Jennifer, 2012. Social dominance orientation: revisiting the structure and function of a variable predicting social and political attitudes. Personality and Social Psychology Bulletin 38, 583–606.

Jost, John, Thompson, Erik, 2000. Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and European Americans. Journal of Experimental Social Psychology 36, 209–232.

Kteily, Nour, Sidanius, Jim, Levin, Shana, 2011. Social dominance orientation: cause or 'mere effect'? Evidence for SDO as a causl predictor of prejudice and discrimination against ethnic and racial outgroups. Journal of Experimental Social Psychology 47, 208–214.

Lee, I-Ching, Pratto, Felicia, Johnson, Blair, 2011. Intergroup consensus/disagreement in support of group-based hierarchy: an examination of socio-structural and psycho-cultural factors. Psychological Bulletin 137, 1029–1064.

Pratto, Felicia, 1999. The puzzle of continuing group inequality: piecing together psychological, social, and cultural forces in social dominance theory. In: Zanna, Mark (Ed.), Advances in Experimental Social Psychology, vol. 31. Academic Press, San Diego, pp. 191–263.

Pratto, Felicia, Cidam, Atilla, Stewart, Andrew, et al., 2013. Social dominance in context and in individuals: contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. Social Psychological and Personality Science 4, 587–599.

Pratto, Felicia, Stewart, Andrew, Foels, Rob, et al., 2014. Validation of the short social dominance orientation scale, manuscript submitted for publication.

Pratto, Felicia, Sidanius, Jim, Stallworth, Lisa, Malle, Bertram, 1994. Social dominance orientation: a personality variable predicting social and political attitudes. Journal of Personality and Social Psychology 67, 741–763.

Pratto, Felicia, Stewart, Andrew, 2012. Group dominance and the half blindness of privilege. Journal of Social Issues 68, 28–45.

Pratto, Felicia, Tatar, Deborah, Conway-Lanz, Sahr, 1999. Who gets what and why: determinants of social allocations. Political Psychology 20, 127–150.

Sibley, Chris, Duckitt, John, 2008. Personality and prejudice: a meta-analysis and theoretical review. Personality and Social Psychology Review 12, 248–279.

Sidanius, Jim, Pratto, Felicia, 1999. Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression. Cambridge University Press, Cambridge, UK.

Sidanius, Jim, Pratto, Felicia, Bobo, Lawrence, 1994. Social dominance orientation and the political psychology of gender: a case of invariance? Journal of Personality and Social Psychology 67, 998–1011.

Xin, Ziqiang, Chi, Liping, 2010. Wording effect leads to a controversy over the construct of the social dominance orientation scale. The Journal of Psychology 144, 473–488.

Social Ecology

Marina Fischer-Kowalski, Institute of Social Ecology, Alpen-Adria University, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social ecology is a highly dynamic interdisciplinary research field rooted in both social science and natural science traditions. The common denominator of this research field is not so much a shared label but a shared paradigm. The core axioms in this paradigm are that human social and natural systems interact, coevolve over time, with causality pointing in both directions. Social ecology deals with energy and society, land use and food production, the metabolism of societies, and the environmental impacts of human activities. It offers a conceptual approach to society–nature coevolution that integrates historical and current development processes and future sustainability transitions.

Introduction

Social ecology constitutes a highly dynamic interdisciplinary research area with roots in both the social and natural sciences (see part 2). Whatever the discipline of origin, the common motive for moving into the direction of socioecological research has been to critically respond to decades of disciplinary differentiation and specialization that created a blind spot (Latour, 1991) detrimental to society's capability to properly understand and address its relation to the – seemingly increasingly strained – natural environment.

The common denominator of this research field is not so much a shared label – names extend, beyond social ecology, from human ecology, industrial ecology, and ecological economics to socioecological systems analysis – but a shared paradigm. The core axioms in this paradigm are that human social and natural systems interact, coevolve over time, and substantially influence each other, with causality pointing in both directions. What follows from this paradigm is a need to develop theoretical and epistemological concepts and research methodologies that can capture social and natural structures and processes in an integrated fashion. This challenge has been addressed in different ways and at different levels of depth and consistency across the various strands of social ecology.

In this article we first reconstruct some earlier academic roots of social–ecological thinking. Subsequently, we discuss key research traditions that deal with the biophysical features of human societies, such as energy, land use, and social metabolism. This is followed by a review of approaches to identifying environmental impacts of human activities. The fourth part is devoted to biohistory, and reviews theoretical and empirical efforts to analyze society–nature coevolution. Finally, we turn to issues of regulation and governance, with a particular focus on future sustainability transitions.

Academic Traditions Contributing to the Emergence of Social Ecology

The academic roots of social ecology can be traced back to the nineteenth century, partly because the social and natural sciences had not yet fallen into their respective epistemic boxes that made crossovers so difficult in the twentieth century. There are excellent reviews reconstructing such roots in the political economy of Adam Smith, David Ricardo, Karl Marx, and Thomas Malthus (Martinez-Alier, 1987; Sieferle, 1990; Fischer-Kowalski, 1998) illustrating the debates on the interrelations between population, land, food, technology, and economic development. While Smith, Ricardo, and Malthus remained focused on natural limitations to economic growth (in particular: land), Marx was the first to claim technological development (and thus: human ingenuity) as the key driver of economic growth, thus overcoming natural limitations.

Another influential field was geography. George Perkins Marsh's book *Man and Nature: or, Physical Geography as Modified by Human Action* (1864) inspired at least two major efforts to comprehensively account for human-induced changes in global ecology, namely the Princeton Conference on *Man's Role in Changing the Face of the Earth* (Thomas, 1956) and the 1987 conference at Clark University titled *The Earth as Transformed by Human Action* (Turner et al., 1990). In 1969 the German geographer Neef explicitly talked about the 'metabolism between society and nature' as a core problem of geography (Neef, 1969). Since then, geographers have played a major role in social ecology.

Cultural ecology, as brilliantly reviewed by Orlove (1980), must be regarded as another important predecessor of later social–ecological research. Similar to sociology, the beginnings of cultural anthropology (as in the works of Morgan, 1877) were marked by evolutionism, that is, the idea of universal historical progress from more 'natural,' barbarian social conditions to more advanced, civilized ones. Cultural anthropology subsequently splits into functionalist and culturalist strands. The functionalist strands retained a focus on the society–nature interface. Leslie White, one of the most prominent anthropologists of his generation and an early representative of the functionalist tradition, rekindled interest in energetics. For White, the vast differences in the types of extant societies could be described as social evolution propelled by energy and technology (White, 1943). In contrast, cultural anthropologist Julian Steward's 'method of cultural ecology' also paid attention to the quality, quantity, and distribution of resources within the environment, albeit with a different focus. This is exemplified by his comparative study 'Tappers and Trappers' (Murphy and Steward, 1955) which examines how two tribes that were traditionally reliant on subsistence hunting

and gathering (and some horticulture) completely change their ways of living as a result of changing their metabolism. The authors report an irreversible shift from a subsistence economy to dependence upon trade.

Despite some early calls for 'ecology of human' (Adams, 1935; Darling, 1956; Sears, 1953), biological ecology was reluctant to engage in human ecology prior to the environmental debate of the 1970s (Young, 1974). Moreover, when the first influential texts on human ecology appeared (e.g., Ehrlich and Ehrlich, 1970; Ehrlich et al., 1973), they adopted an approach that remained typical for much of the bioecological work until very recently: namely to view humans as a source of disturbance in ecosystems. Those resulted in a rather narrow concept of human agency which hampered interdisciplinary collaboration. The fact that ecologists tended to favor 'natural' ecosystems over 'human-dominated' ones as study objects may also have contributed to this bias. However, the most important factor was that many biological ecologists simply did not recognize the need to develop a more complex approach that conceptualizes socioeconomic systems as different from natural systems. This was further exacerbated by the dominance of neo-Malthusian ideas in bioecological approaches to human ecology (above all in the work of Paul Ehrlich) unpopular with social scientists. The revived environmental debates of the 1970s changed the situation: major international research programs were launched, including UNESCO's Man-and-the-Biosphere, or the International Human Dimensions Programme for Global Environmental Change, which stimulated and supported research across the 'great divide' (Snow, 1956) between the social and natural sciences.

History as a humanities discipline, in particular the tradition of the Annales School (e.g., Fernand Braudel), paved its own way toward social–ecological reasoning. Braudel viewed the history of the Mediterranean as an outcome of interaction between social and natural processes. Maurice Godelier formulates his core hypothesis in the introduction to *The Mental and the Material* as follows:

> Human beings have a history because they transform nature. It is indeed this capacity which defines them as human. Of all the forces which set them in movement and prompt them to invent new forms of society, the most profound is their ability to transform their relations with nature by transforming nature itself.
>
> Godelier, 1984, p. 1

This way of looking at history relates to the Marxist tradition but transcends it, instead taking ecological and/or coevolutionary directions. The classic reading of Marx leads to a discussion of changing 'modes of appropriation of nature' through the development of new means of production, namely technology. Godelier's reading stresses the fact that human appropriation of nature modifies nature and this modified nature in turn stimulates social change. Godelier thus deviates from mainstream social science approaches by viewing nature as historically variable, attributing societies' historical dynamics to a feedback process from nature. Since the 1970s, the field of environmental history has worked with this fundamental concept of mutuality of nature–society relations (Winiwarter and Knoll, 2007).

For sociology, some commentators see the so-called Chicago School of Human Ecology (Park, Burgess, Duncan) as an entry point to modern social ecology (Park, 1952; Duncan, 1959). This school used concepts and terminology from biological ecology such as hierarchy, competition, or succession to analyze urban development. However, the natural environment was largely reduced to spatial structures in the built environment. For example, Duncan's POET model (population, organization, environment, and technology) described social processes but did not refer to natural processes or conditions, with the exception of space (Beus, 1993). This said, some followers of the German human ecology tradition view this differently (cf Bruckmeier, 2004). Catton and Dunlap (1978) called for sociology to move beyond Durkheim's dictum that "a social fact can be explained only by another social fact" (cited in Beus, 1993, p. 94) and to abandon the 'human exceptionalism paradigm' in favor of an ecological paradigm in which the human species is one among other species on Earth, sharing their susceptibility to nature. This frequently cited appeal, nevertheless, did not immediately give rise to substantially new theoretical or empirical approaches. Almost a decade later, German sociologist Luhmann, who built upon the system-theoretical approach of biologists Maturana and Varela (1975), published an influential book (1986) on society's ability (or lack thereof) to adequately react to disturbances it initiates in its natural environment, even if these disturbances could prove to be detrimental to society in the long run. In the same year, Beck (1986) started to publish on 'risk society,' proposing that late modern society was characterized by the ways in which it creates and handles environmental risks and redistributes their consequences among its members. This starkly contrasts with traditional industrial society's preoccupation with its emancipation from natural forces and efforts to legitimately handle social inequalities. Jänicke (1988) and later Mol and Spaargaren (2000) opened a debate on 'ecological modernization.' They claimed that modern societies are capable of technological and organizational learning to mitigate or solve environmental problems. All these approaches, though, fall short of an epistemic turn toward a more symmetrical view of the social and the natural.

Society's Biophysical Structures

Energy and Society

The idea that control of energy matters for society, even determines the advancement of civilization, has a long tradition in social theory, prominently represented by Herbert Spencer. In his First Principles in 1862 societal progress and developmental gradients between societies are linked to energy consumption: The more energy a society is able to consume, the further advanced it is. Societal progress is based on energy surplus because it enables social growth and differentiation and frees up time for cultural activities beyond work to meet basic vital needs. Cottrell (1955/2009), on the other hand, offered a more nuanced analysis of the relevance of the sources and amounts of available energy for social processes that steered clear of the more narrow energetic evolutionism embraced by Spencer and Morgan. Similarly, physicist Vaclav Smil published periodic compendia on *Energy in the Biosphere*

and Civilization (1991) up to *Energy in Nature and Society* (2008) that compiled encyclopedic knowledge on how energy matters. Another physicist, Robert Ayres, has presented convincing theoretical and empirical evidence that the expenditure of useful work (i.e., exergy) was, and is key to economic growth (Ayres and Warr, 2005). The historian Rolf Peter Sieferle (1982, 2001b) analyzed the rise of the United Kingdom into industrialization and political hegemony as an outcome of its 'subterranean forest,' i.e., its use of coal that gave it access to significantly more energy than if the United Kingdom had had all its territory covered with forest and burned it for energetic use. Currently, the issue of reducing fossil fuel consumption to both address 'peak oil' (and peak fossil fuels more generally, cf Murphy, 2012) and avoid dangerous climate change and its potential consequences for economies and societies, is stimulating much research on the potential consequences of changing society's energy base and possibly reducing the energy intensity of social processes altogether.

Land Use and Food Production

That societies extend over territories and restructure them for social purposes, with severe consequences for social as well as natural processes, is not a new issue among social scientists. However, it has regained considerable momentum of late. Ester Boserup (1965, 1981), both in her continuation and her critique of Neo-Malthusianism, was simultaneously at odds with the development policies of her time by (empirically) demonstrating social learning processes in response to population growth and food scarcity. She was able to show that traditional agriculture had found ways to feed an expanding population by intensifying land use (and not only, as in the Malthusian paradigm, by extending agricultural area). Land-use intensity is an essential aspect of the human use of terrestrial ecosystems. In the course of history, the intensification of land use allowed to overcome Malthusian traps and support population growth as well as improving supply of food and other products dependent on photosynthesis. It helped to achieve increases in agricultural production without requiring proportional increases in the area of agricultural land. On the contrary, intensification ensured that most industrialized countries increased their agricultural output in spite of shrinking agricultural areas in the last decades, if not centuries. In the industrial part of the world, we find reforestation instead of long-term deforestation (although this is partly because of deforestation and 'land grabs' in developing countries). On the other hand, increasing land-use intensity has often been associated with detrimental effects of land use on ecosystem functioning, such as soil degradation, groundwater and air pollution, and biodiversity loss. Such processes have had negative effects on the ability of ecosystems to sustain vital ecosystem services, thereby potentially jeopardizing human well-being in the long run. Under conditions of traditional agriculture, which largely prevailed worldwide until the 1960s, increased food output per unit area was achieved through increased investment of human labor (Boserup, 1981). This, in turn, incentivized high fertility to provide the necessary labor power, thereby contributing to population growth. When fossil fuels use facilitated the industrialization of agriculture (mineral fertilization, pesticides, tractors), this started to change (cf. Fischer-Kowalski et al. 2014a), with agriculture turning from a supplier of energy to a consumer (Pimentel et al., 1973). Emerging toxicological hazards linked to agriculture such as those documented in Rachel Carson's seminal book *Silent Spring* (1962) ran the risk of poisoning the entire food chain. More recently, the debate has revolved more around the risks of genetic engineering (see for example the Nature Special Feature, 2013). Importantly, the issue of land use and land cover change has mobilized a large research community (most recently, the international program on 'Future Earth,' http://www.icsu.org/future-earth) working on food security, diets, carbon emissions, biodiversity losses, climate and habitat change through broad interdisciplinary cooperation.

Social Metabolism/Industrial Ecology

Ecological Economics

While Marx had already recognized the centrality of social metabolism with nature for human labor (Marx, 1867/2010), these considerations resurfaced in Ayres' and Kneese's work in the second half of the twentieth century. They claimed that the common failure of economics results from viewing production and consumption in ways that are at variance with the fundamental law of the conservation of mass (Ayres and Kneese, 1969, p. 283). They argue that uncompensated externalities such as pollution and waste are inevitable unless the following three conditions are met: (1) all inputs into the production process are fully converted into outputs, with no unwanted residuals along the way; (2) all final outputs (commodities) are completely used up through consumption; and (3) property rights guarantee that all relevant environmental attributes are in private ownership and that these rights are exchanged in competitive markets. Given that these three conditions are never fully met in practice, environmental policies targeting wastes and emissions must take into account the full cycle of industrial metabolism to succeed (Ayres and Simonis, 1994).

The Viennese School of Social Ecology has introduced a paradigm for social–ecological research by placing social metabolism within a wider picture of society–nature interaction (**Figure 1**).

This model depicts social metabolism as one of the key links between society's biophysical structures (defined as human population, livestock population, and physical infrastructures,

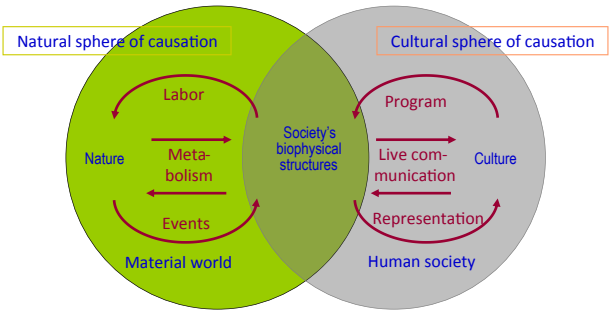

Figure 1 Society–nature interaction and the role of social metabolism. Source: Elaborated after Fischer-Kowalski, M., Haberl, H., 2007. Socioecological Transitions and Global Change: Trajectories of Social Metabolism and Land Use. Edward Elgar, Cheltenham, UK; Northhampton, USA, p. 13 © Institute of Social Ecology (SEC).

including technical equipment) and the natural environment. In order to reproduce its biophysical stocks, society needs continuous flows of energy and materials. At the same time, it expands labor to intervene in nature and modify it according to society's needs (e.g., agriculture, construction activities). Society's biophysical structures are shaped by events that happen in nature. Through culture and communication, society represents these events, interpreting them as rewards for people's efforts (e.g., a large harvest), as catastrophes (e.g., floods), or possibly as irrelevant. In the other direction, there is a cultural program that translates into action. An interesting feature distinct to this model is the overlap between the natural and the cultural–symbolic realms: elements are neither exclusively natural nor purely cultural, but are governed by both natural and cultural factors. In other words, society is a hybrid of the two realms (see also Boyden, 1992). Recent similar conceptualizations used by Liu et al. (2007), who talk about the complexity of coupled human and natural systems, and Becker (2013), who emphasizes the importance of hybrid structures, continue and expand on this paradigmatic tradition.

By conceptually linking metabolic flows with biophysical stocks in this way, it is possible to define boundaries for social systems (vis-à-vis both their natural environment and each other) and to create a consistent metric for material and energy flows for social systems on various scales (local communities, firms, cities, or nation-state). For example, national material flow accounting has become a regular part of public statistics in Japan, the European Union, and elsewhere (Fischer-Kowalski et al., 2011). It provides reasonably reliable annual accounts of natural resource use in physical terms and allows for their comparison across time and with economic accounts (**Figure 2**). On the global scale, adding up all the extraction of natural resources that occurs during a year produces useful figures while, at the same time, avoiding the complexities of international trade networks that distribute these resources around the globe (**Figure 3**).

Figure 2 illustrates an eightfold increase in resource extraction during the twentieth century across the following areas: Biomass, construction minerals, fossil energy carriers, and metals (including industrial minerals). While societies at the beginning of the century reproduced themselves mainly through biomass inputs (i.e., food for humans and animals, firewood), they increasingly turned toward the so-called nonrenewable resources such as fossil fuels and ores (Krausmann et al., 2009). Moreover, biophysical stocks clearly increased between 1900 and 2000, with the human population increasing fivefold. In addition to this substantial population growth, metabolic rate, that is, resource use per person also doubled, from less than 5 tons per person per year to almost 10 tons. This was complemented by an even greater increase in world GDP (at constant prices) and average income per person. This apparent 'decoupling' of resource use and income is mainly due to technological progress that facilitates producing more value with less input, which also makes products cheaper. However, this also feeds into further growth of resource consumption: an hour's work can buy more products.

For nation-states, there is a rich literature comparing resource requirements (e.g., Weisz et al., 2006 for the European Union), resource efficiencies (e.g., Schandl and West, 2010 for Asia and the Pacific), trade patterns (e.g., Dittrich and Bringezu, 2010), and growth in biophysical stocks across countries. On the other end of the metabolic process, there is particular interest in researching greenhouse gas emissions (that can be calculated from fossil fuel use, livestock numbers, and steel and cement production) that either occur directly within countries or result from international trading. Naturally, these emissions are highly relevant for climate policies.

Interest in the metabolism of cities has surged of late, with city planning being identified as an important means to reduce resource consumption while maintaining welfare levels. Substantial amounts of energy for heat and transportation, construction materials, and land can be saved by appropriate spatial structures (Kennedy et al., 2007; Weisz and Steinberger, 2010).

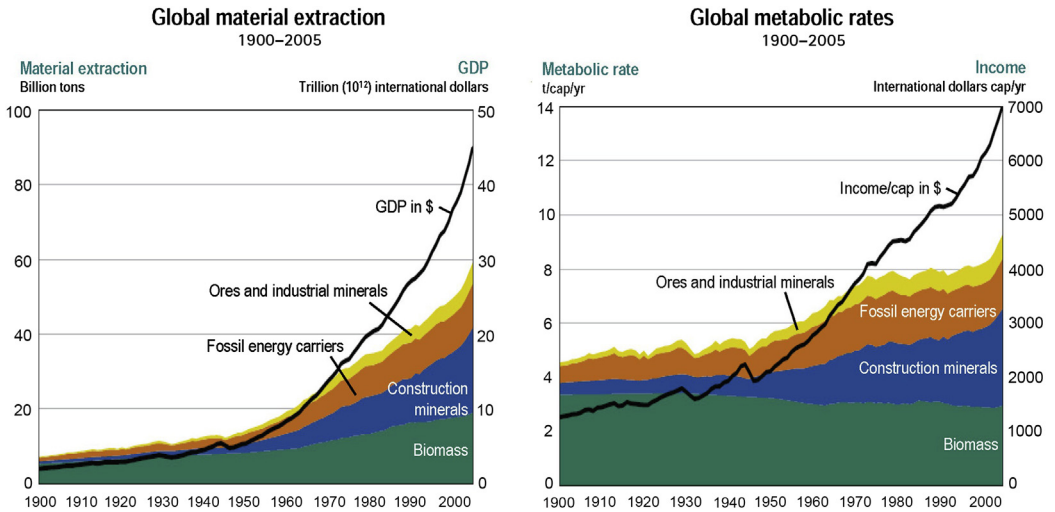

Figure 2 Global social metabolism during the twentieth century. Source: UNEP, 2011. Decoupling Resource Use and Environmental Impacts from Economic Growth. United Nations Environmental Programme, Paris.

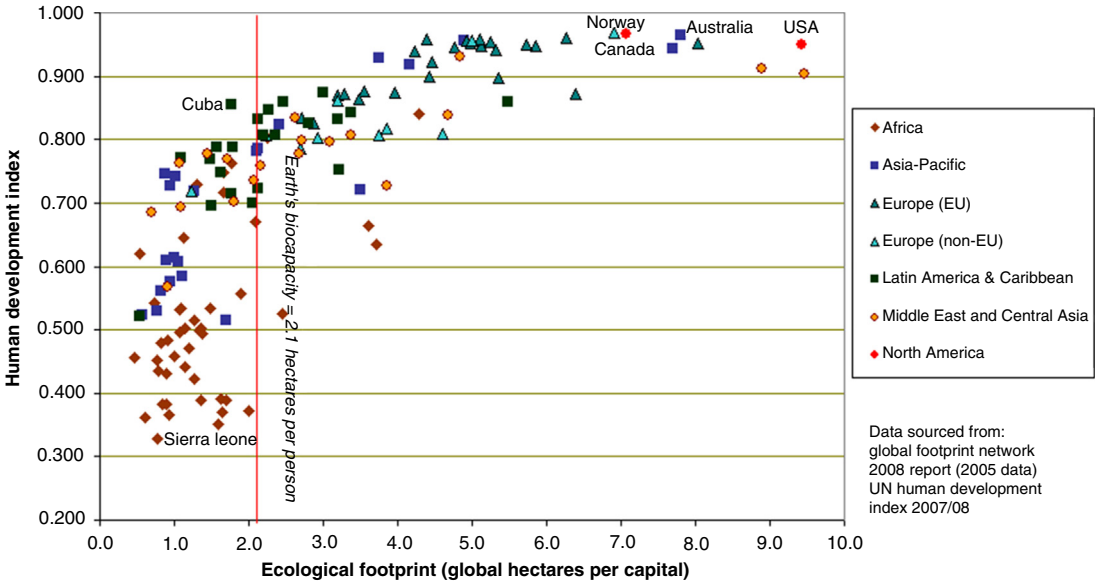

Figure 3 Interrelation between the human development index and the ecological footprint.

A somewhat different metabolic approach is chosen by Giampietro et al. (2012). Their "Multi-Scale Integrated Analysis of Societal and Ecological Metabolism" (MuSIASEM) systematically relates human labor, exosomatic energy use, and economic output to describe the metabolic pattern of various types of social systems (from households to farms to national economies, stratified into sectors). This approach builds on Georgescu-Roegen's (1971) foundational work on 'bioeconomics,' which influenced the emergence of ecological economics.

Identifying Environmental Impacts of Human Activities

In the 1970s, the so-called IPAT equation was developed from a debate about the relative importance of population growth on the one hand, growth in affluence on the other, in determining human impacts upon the environment (Ehrlich and Holdren, 1971; Chertow, 2001). IPAT is the lettering of the following formula:

$$I = P * A * T$$

where I stands for (environmental) Impact, P for Population, A for Affluence, and T for Technology. This formula has been repeatedly applied to estimate various environmental impacts such as land and resource use, pollution, CO_2 emissions, or the ecological footprint (EF) (see below). More statistically elaborate applications use regression analysis to determine the relative weight of the components and to calculate both nonlinearities and interactions (e.g., Dietz et al., 2007). The hope attached to this equation, that technological improvements would neutralize at least some of the ecological damaging effects of population growth and increasing affluence, is not strongly substantiated. In some cases the opposite occurred. For example, Schandl and West (2012) found that the impact of increasing affluence on CO_2 emissions in Asian and Pacific countries was further exacerbated by technology changes. This seems plausible given that economic growth in developing countries typically implies a shift toward using fossil fuels. For a long-term IPAT analysis with similar outcomes see Fischer-Kowalski et al. 2014b.

Another widely used approach to describing human environmental impact is the so-called EF. In 1996, Wackernagel and Rees published the book *Our Ecological Footprint: Reducing Human Impact on the Earth*. Per capita EF, or EF analysis, is a means of comparing consumption and lifestyles, and checking this against nature's ability to provide for this consumption. It does this by assessing the biologically productive land and marine areas required to produce the resources a population consumes and to absorb the corresponding waste using prevailing technology. The EF is expressed in 'global hectares.' Despite several modifications, this indicator has been subject to considerable methodological criticism (cf Gaube et al., 2013), including how different land productivities are taken into account (Haberl et al., 2004), or how trade can be integrated into the picture (Grazi et al., 2007).

This said, it remains one of the most powerful tools for communicating the human impact on the environment. Moreover, it can inform policy by examining to what extent a nation uses more (or less) than is available within its territory, or whether a nation's lifestyle would be replicable worldwide. The footprint can also be a useful tool for educating people about carrying capacity and overconsumption, with the aim of changing behavior. EFs may be used to argue that many current lifestyles are not sustainable. Global EF comparisons also clearly show the deep inequalities in resource use on this planet at the beginning of the twenty-first century.

From a social science perspective, these indicators and analyses of environmental impact cannot adequately capture the complexity on the social system side of the metabolic process and usually lack a coevolutionary perspective. While

they recognize that humans exploit nature, they do not include how nature hits back or what adjustments human societies (are forced to) make in response to environmental problems.

The so-called DPSIR model – D = Drivers, P = (environmental) Pressures, S = States (of the environment), I = (environmental) Impacts, and R = (policy) Responses – as used by the OECD and the European Environmental Agency attempts to address this gap. It assumes a causal chain whereby social causes ('drivers') exert pressures upon the environment that trigger changes, with society subsequently responding to those changes with (presumably ameliorating) policies. Still, while the DPSIR model focuses more on social processes than its predecessors, it does so in a rather narrow way (Stanners et al., 2007).

Biohistory and Society–Nature Coevolution

There is a long tradition in the social and historical sciences to distinguish between qualitatively different forms of societal organization, subsistence or production, or to identify diverse stages of civilization or development. The distinctions drawn, and the criteria upon which these are based, vary considerably but rarely refer to society–environment relations or environmental consequences of human activity.

It has been the special achievement of Rolf Peter Sieferle (1997) to regard different forms of societal organization not only as socially or socioeconomically distinct entities but to systematize them according to their socioecological patterns. This implied considering both social organization (in the widest sense of the word) and concomitant changes in the environment. Key to Sieferle's distinctions is society's key source of energy and its dominant conversion technology for energy production. He distinguishes between the hunting and gathering mode, the agrarian mode (with some subdivisions), and the industrial mode. The energy system of hunters and gatherers is 'passive solar energy utilization.' They live on the products of recent photosynthesis (plants and animals for their food, firewood for heat). That they use fire to cook (rather than grill) their food widens the specter of edibles – but still, only a very small fraction of their environment qualifies as food whose collection requires mobility, both on an everyday basis and seasonally, allowing only for very low population densities. In contrast, the agrarian mode is based upon 'active solar energy utilization,' implying that certain areas are cleared of their natural vegetation and that solar energy on these areas is instead monopolized to grow human food plants. This leads to extensive deforestation of the Earth (and the enrichment of the atmosphere with the CO_2 that previously had been stored in trees and soils), a sedentary way of life as well as a large human labor burden (despite technological improvements to increase yields, Boserup, 1965, 1981). Resulting changes in lifestyle and nutrition allow for higher fertility while the large labor burden motivates having children to share the labor. The resulting growth in population creates high population densities and an expansion of the agrarian mode across the world. Control of territory, tools, livestock, and stored reserves is essential, and frequent territorial conflicts lead to the emergence of specialized classes of people to defend and attack territories, social hierarchies to control them, and urban centers that offer protection and opportunities for trading. In many parts of the world, these agrarian systems developed into major empires and civilizations that emerged, flourished, and in some cases collapsed again (Diamond, 2005; Tainter, 1988).

In the sixteenth century, a new fossil fuel-based energy system started to emerge, supplying society with an unprecedented amount of energy. In the United Kingdom, the use of coal instead of increasingly scarce firewood initiated urban growth and industrial manufacturing. Textile production for export became very profitable and sheep gradually crowded out farmers growing food. The invention of the steam engine finally triggered what is now known as the industrial revolution, which caused large-scale ecological and social transformations. Industrialization continues today, spreading from the core industrialized countries (comprising about 20% of the world population) to the much larger rest of the world at an accelerating speed (Fischer-Kowalski and Haberl, 2007). It remains to be seen whether the exhaustion of fossil fuels, a detrimental transformation of the Earth's climate system, or politically guided change will bring an end to this energy regime. In any case, it is likely that the fossil fuel era will have lasted for a much shorter period of time than the two previous regimes.

Overall, the main appeal of Sieferle's classification is that it sheds light on different functional problems societies face when trying to establish and maintain themselves under specific environmental conditions and the associated evolutionary advantages and drawbacks. Moreover, Sieferle's approach also provides some clues about the possible direction of change. More recently, a new research area has emerged from ecological observations of long-term processes across a global network of local and regional habitats (also called Long-term ecological research sites). In recent times, this research has been extended to also focus on social processes, culminating in the LTSER network (long-term socioecological research, see Singh et al., 2013).

Regulation, Governance and Sustainability Transitions

Good governance of the commons (or lack thereof) is a long-standing socioecological theme. Departing from existing models of collective action, Elinor Ostrom's book, *Governing the Commons* (1990) stimulated a rich strand of research in the area of new institutional economics, which culminated in her winning the Nobel Memorial Prize in Economic Sciences in 2009. The focus of her research was on how people manage common-pool resources such as land and water to maintain long-term sustainable resource yields. Her field studies focused on the management of pastures and irrigation networks by local people and documented how societies have developed diverse institutional arrangements for managing natural resources and avoiding ecosystem collapse. Ostrom (1996) identified a number of 'design principles' of stable local common-pool resource management including:

- Clearly defined boundaries (effective exclusion of external un-entitled parties);
- Collective-choice arrangements that allow most resource appropriators to participate in the decision-making process;
- Effective monitoring by agents who are part of, or accountable to the appropriators;
- A scale of graduated sanctions for resource appropriators who violate community rules and mechanisms of conflict resolution that are cheap and easy to access.

In her later work, these principles have expanded to include a number of additional variables believed to affect the success of systems of self-organization and self-governance, including effective communication, internal trust, and reciprocity.

Ostrom and her many coresearchers developed a comprehensive 'Social–Ecological Systems framework,' within which much of the still-evolving theory of common-pool resources and collective self-governance is now located (Ostrom, 2009). A strong research community that relates to these approaches is the Resilience Alliance. This network of institutions and people publishes the influential open-access journal *Ecology and Society* and defines socioecological systems as

> a multi-scale pattern of resource use around which humans have organized themselves in a particular social structure (distribution of people, resource management, consumption patterns, and associated norms and rules) …Hence, the main aim of resilience management and governance is to keep the system within a particular configuration of states (system 'regime') that will continue to deliver desired ecosystem goods and services and prevent the system from moving into an un-desirable regime from which it is either difficult or impossible to recover, or, to move from a less desirable to a more desirable regime.
> http://www.resalliance.org/index.php/key_concepts

The Frankfurt approach to social ecology deploys the core concept of 'societal nature relations' (*gesellschaftliche Naturverhältnisse*), focusing on the *relations* between society and nature in terms of the various societal *regulations* that define them. Operationally, this approach focuses on *basic* societal nature relations such as freshwater or food supply that are indispensable for meeting basic human needs and for individual and societal reproduction and development. The link to the concept of human needs turns societal nature relations into an irreducibly *normative* concept: the basic societal nature relations ought to be *regulated* in such a way that *all* human are able to meet their basic needs (Becker et al., 2011, p. 79). The Frankfurt approach defines as its 'epistemic object' the 'crisis of society–nature relations' (Becker and Jahn, 2006, p. 19). This definition is highly normative in that presupposes a crisis, that is, a dramatic deviation of the 'is'- from the 'ought'-state of societal nature relations. The purpose of social ecology is thus to generate the knowledge to understand and react to this crisis, with a view to reaching the desired 'ought'-state of societal nature relations. The core research question of social ecology is thus: "How can crisis-ridden societal nature relations be perceived, understood and actively shaped?" (Becker and Jahn, 2006, p. 12). In the 1990s, the German Government established an interdisciplinary social–ecological research program that reinforced this orientation toward basic needs, demanding strong involvement of stakeholders and strengthening the policy relevance of socioecological research in Germany over many years.

The Dutch societal transitions management school adopts a similar approach to managing human-environment systems. In contrast to the resilience alliance tradition, the Dutch school focuses more on interactions between technical and social systems. The core concern are 'wicked' problems in social system functioning that cannot be solved through conventional means but that require a systemic transition. A social–ecological transition is a transition between two dynamic equilibria, that is, a shift from one more-or-less stable state to another. The typical model of a transition is the S-curve, which allows for the distinction of discrete phases: a 'predevelopment phase' in which some processes start to deviate from the dominant pattern; the 'takeoff phase' involving a departure from the original equilibrium; an 'acceleration phase,' where change accelerates in a nonincremental, disruptive, and potentially chaotic manner; and a 'stabilization phase,' where the rate of change is declining and a new dynamic equilibrium is approached (Rotmans et al., 2001, p. 1). The nature of transitional dynamics is described as a generic sequence of mechanisms that result in irreversible changes in the system. 'Niches,' that is, individual technologies, practices, and actors outside or peripheral to the existing regime function as potential drivers of radical innovation (Geels, 2005). Niches emerge and cluster, and by empowering a niche cluster, a niche regime unfolds. This niche regime may then become more powerful, finally dominating and replacing the incumbent regime. Underlying mechanisms are variation and selection, adaptation, emergence, clustering, empowerment, transformation, decay, and building up. Transition management draws together a number of front-runners in a protected environment or arena. To effectively create a new regime, agents need to be at a certain distance from the incumbent regime. However, the continuous link with the regime is also important, requiring buy-in from regime agents that are open to change. The methodological concept of transition management frequently draws on this 'Multi-Level Perspective' on an evolutionary dynamics between a micro (niche), meso (regime), and macro level. Transition management was taken up by the Global Energy Assessment (GEA, 2012) and UNEP for framing its new modeling approach (UNEP, 2013).

To conclude, many of the social ecology approaches reviewed in this article are known and used by natural scientists interested in sustainability questions. However, they appear to be less popular among social scientists who follow more traditional disciplinary pathways. This said, social ecological concepts and methodologies enjoy increasing popularity within the rapidly expanding social–scientific sustainability research community (cf Rau and Fahy, 2013). Greater involvement by social scientists in interdisciplinary debates about humanities' long-term future on Earth would help to further progress a paradigmatic shift toward social–ecological thinking, policy, and practice.

See also: Anthropocene Epoch; Archaeometry; Carrying Capacity of the Environment; Climate Change; Cultural Evolution: Theory and Models; Eco-cities and Sustainable Urbanism; Ecological Economics; Energy Flow Analysis; Energy and Society; Environment and Development; Environmental Economics; Environmental Sociology; Food Security and 'Green Revolution'; Food, Sociology of; Global History: Universal and World; Human Ecology; IPAT (Impact, Population, Affluence, and Technology); Industrial Ecology; Land Change: The Merger of Land Cover and Land use Dynamics; Material Flow Analysis; Mineral Resources, Limits to: The Case of Peak Oil; Modes of Production; Sustainability and Sustainability Science; Sustainable Development: An Economic Perspective; Sustainable Production and Consumption; Time Use and Resource Consumption; Water Resources and Sustainable Water Management.

Bibliography

Adams, C.C., 1935. The relation of general ecology to human ecology. Ecology 16, 316–335.

Ayres, R.U., Kneese, A.V., 1969. Production, consumption and externalities. American Economic Review 59 (3), 282–297.

Ayres, R.U., Simons, U.E., 1994. Industrial Metabolism: Restructuring for Sustainable. Development. United Nations University Press, Tokyo, New York, Paris.

Ayres, R.U., Warr, B., 2005. Accounting for growth: the role of physical work. Structural Change and Economic Dynamics 16 (2), 181–209.

Beck, U., 1986. Risikogesellschaft. Auf den Weg in eine andere Moderne. Suhrkamp, Frankfurt a.M.

Becker, E., 2013. Social-ecological systems as epistemic objects. In: Glaser, M., Krause, G., Ratter, B.M.W., Welp, M. (Eds.), Human-Nature Interactions in the Anthropocene. Potentials of Social-Ecological Systems Analysis. Routledge, London, pp. 37–59.

Becker, E., Hummel, D., Jahn, T., 2011. Gesellschaftliche Naturverhältnisse als Rahmenkonzept. In: Groß, M. (Ed.), Handbuch Umweltsoziologie. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 75–96.

Becker, E., Jahn, T., 2006. Soziale Ökologie. Grundzüge einer Wissenschaft von den gesellschaftlichen Naturverhältnissen. Campus, Frankfurt.

Beus, C.E., 1993. Sociology, human ecology, and ecology. Advances in Human Ecology 2, 93–132.

Boserup, E., 1965. The Conditions of Agricultural Growth. The Economics of Agrarian Change under Population Pressure. Aldine/Earthscan, Chicago.

Boserup, E., 1981. Population and Technological Change – a Study of Long-Term Trends. The University of Chicago Press, Chicago.

Boyden, S.V., 1992. Biohistory: The Interplay between Human Society and the Biosphere – Past and Present. UNESCO and Parthenon Publishing Group, Paris, Casterton Hall; Park Ridge, New Jersey.

Bruckmeier, K., 2004. Die unbekannte Geschichte der Humanökologie. In: Serbser, W. (Ed.), Humanökologie. Ursprünge – Trends – Zukünfte. Ökom Verlag, München, pp. 45–120.

Carson, R., 1962. Silent Spring. Houghton Mifflin Company, Boston.

Catton Jr., W.R., Dunlap, R.E., 1978. Environmental sociology: a new paradigm. The American Sociologist 13 (1), 41–49.

Chertow, M.R., 2001. The IPAT equation and its variants. changing views of technology and environmental impact. Journal of Industrial Ecology 4 (4), 13–30.

Cottrell, F., 1955. Energy and Society. The Relation between Energy, Social Change, and Economic Development. McGraw-Hill Book Company, New York, Toronto, London (Revised edition: 2009).

Darling, F.F., 1956. The ecology of man. American Scholar 25 (1), 38–46.

Diamond, J., 2005. Collapse. How Societies Choose to Fail or Succeed. Viking, New York.

Dietz, T., Rosa, E.A., York, R., 2007. Driving the human ecological footprint. Frontiers in Ecology and the Environment 5 (1), 13–18. See also. www.stirpat.org (STIRPAT: "a research program in structural human ecology") (accessed 24.08.13.).

Dittrich, M., Bringezu, S., 2010. The physical dimension of international trade: part 1: direct global flows between 1962 and 2005. Ecological Economics 69 (9), 1838–1847.

Duncan, O.D., 1959. Human ecology and population studies. In: Hauser, P.M., Duncan, O.D. (Eds.), The Study of Population. The Chicago University Press, Chicago, pp. 678–716.

Ehrlich, P.R., Ehrlich, A.H., 1970. Population, Resources, Environment: Issues in Human Ecology. Freeman, San Francisco.

Ehrlich, P.R., Ehrlich, A.H., Holdren, J.H., 1973. Human Ecology, Problems and Solutions. Freeman, San Francisco.

Ehrlich, P.R., Holdren, J.P., 1971. Impact of population growth. Science 171 (3977), 1212–1217.

Fischer-Kowalski, M., 1998. Society's metabolism. The intellectual history of material flow analysis, part I: 1860–1970. Journal of Industrial Ecology 2 (1), 61–78.

Fischer-Kowalski, M., Haberl, H., 2007. Socioecological Transitions and Global Change: Trajectories of Social Metabolism and Land Use. Edward Elgar, Cheltenham, UK; Northhampton, USA.

Fischer-Kowalski, M., Krausmann, F., Giljum, S., Lutter, S., Mayer, A., Bringezu, S., Moriguchi, Y., Schütz, H., Schandl, H., Weisz, H., 2011. Methodology and indicators of economy wide material flow accounting. State of the art and reliablity across sources. Journal of Industrial Ecology 15 (6), 855–876.

Fischer-Kowalski, M., Krausmann, F., Pallua, I., 2014b. A sociometabolic reading of the Anthropocene: modes of subsistence, population size and human impact upon Earth. Anthropocene Review 1 (2014): 8–33.

Fischer-Kowalski, M., Rotmans, J., 2009. Conceptualizing, observing and influencing social-ecological transitions. Ecology and Society 14 (2), 1–3. http://www.ecologyandsociety.org/vol4/iss2/art3.

Fischer-Kowalski, M., Krausmann, F., Mayer, A., Schaffartzik, A., 2014a. Boserup's theory on technological change as a point of departure for the theory of sociometabolic regime transitions. In: Fischer-Kowalski, M., et al. (Eds.), Ester Boserup's Legacy on Sustainability – Orientations for Contemporary Research. Springer, Dordrecht, Heidelberg, New York, London, pp. 23–42.

Gaube, V., Erb, K., Haberl, H., 2013. Biophysical indicators of society-nature interaction: material and energy flow analysis, human appropriation of NPP and the ecological footprint. In: Fahy, F., Rau, H. (Eds.), Methods of Sustainability Research in the Social Sciences. Sage, London, pp. 114–132.

GEA, 2012. Global Energy Assessment – toward a Sustainable Future. Cambridge University Press and the International Institute for Applied Systems Analysis, Cambridge, UK; Laxenburg, Austria.

Geels, F.W., 2005. The dynamics of transitions in socio-technical systems: a multi-level analysis of the transition pathway from horse-drawn carriages to automobiles (1860–1930). Technology Analysis & Strategic Management 17 (4), 445–476.

Georgescu-Roegen, N., 1971. The Entropy Law and the Economic Process. Harvard University Press, Cambridge, MA.

Giampietro, M., Mayumi, K., Sorman Alevgül, H., 2012. The Metabolic Pattern of Societies. Where Economies Fall Short. Routledge, New York.

Godelier, M., 1984. L'idéel et le matériel. Pensée, économies, societés. Fayard, Paris.

Grazi, F., van den Bergh, J.C.J.M., Rietveld, P., 2007. Spatial welfare economics versus ecological footprint: modeling agglomeration, externalities and trade. Environmental and Resource Economics 38, 135–153.

Haberl, H., Wackernagel, M., Krausmann, F., Erb, K.-H., Monfreda, C., 2004. Ecological footprints and human appropriation of net primary production: a comparison. Land Use Policy 21 (3), 279–288.

Hardin, G., December 13, 1968. The tragedy of the commons. Science 162, 1243–1248.

Jänicke, M., 1988. Ökologische Modernisierung: Optionen und Restriktionen präventiver Umweltpolitik. In: Simonis, U.E. (Ed.), Präventive Umweltpolitik. Campus, Frankfurt/New York, pp. 13–26.

Kennedy, C.A., Cuddihy, J., Engel-Yan, J., 2007. The changing metabolism of cities. Journal of Industrial Ecology 11 (2), 1–17.

Krausmann, F., Gingrich, S., Eisenmenger, N., Erb, K.-H., Haberl, H., Fischer-Kowalski, M., 2009. Growth in global materials use, GDP and population during the 20th century. Ecological Economics 68 (10), 2696–2705.

Latour, B., 1991. Nous n'avons pas jamais été modernes. Essai d'anthropologie symétrique. Editions la Découverte, Paris.

Liu, J.G., Dietz, T., Carpenter, S.R., Alberti, M., Folke, C., Moran, E., Pell, A.N., Deadman, P., Kratz, T., Lubchenco, J., Ostrom, E., Ouyang, Z., Provencher, W., Redman, C.L., Schneider, S.H., Taylor, W.W., 2007. Complexity of coupled human and natural systems. Science 317 (5844), 1513–1516.

Luhmann, N., 1986. Ökologische Kommunikation: Kann die moderne Gesellschaft sich auf ökologische Gefährdungen einstellen? Westdeutscher Verlag, Opladen.

Marsh, G.P., 1864. Man and Nature; or, Physical Geography as Modified by Human Action. Scribners & Sampson Low, London & New York.
Martinez-Alier, J., 1987. Ecological Economics. Energy, Environment and Society, Blackwell, Oxford.
Marx, K., 2010. Capital. A Critique of Political Economy, vol. I. E-Book: www.marxists.org/archive/marx/works/1867-c1/ (Original: 1867).
Maturana, H.R., Varela, F.G., 1975. Autopoietic Systems. A Characterization of the Living Organization. University of Illinois Press, Urbana-Champaign, Illinois.
Mol, A.P.J., Spaargaren, G., 2000. Ecological modernisation theory in debate: a review. Environmental Politics 9 (1), 17–49.
Morgan, L.H., 1963. Ancient Society. World Publications, Cleveland (Original: 1877).
Murphy, D.J., 2012. Fossil fuels: peak oil is affecting the economy already. Nature 483 (7391), 541.
Murphy, R., Steward, J.H., 1955. Tappers and trappers: parallel process in acculturation. Economic Development and Cultural Change 4, 335–355.
Nature Special Feature, 2013. GM crops: promise and reality. Nature 497 (7447), 5–152. http://www.nature.com/news/specials/gmcrops/index.html.
Neef, E., 1969. Der Stoffwechsel zwischen Gesellschaft und Natur als geographisches Problem. Geographische Rundschau 21, 453–459.
Orlove, B.S., 1980. Ecological anthropology. Annual Review of Anthropology 9, 235–273.
Ostrom, E., 1996. Governing the Commons. The Evolution of Institutions for Collective Action. Press Syndicate of the University of Cambridge, Cambridge.
Ostrom, E., 2009. A general framework for analyzing sustainability of social-ecological systems. Science 325 (5939), 419–422.
Park, R., 1952. Human Communities: The City and Human Ecology. Free Press, Glencoe.
Pimentel, D., Hurd, L.E., Bellotti, A.C., Forster, M.J., Oka, I.N., Sholes, O.D., Whitman, R.J., 1973. Food production and the energy crisis. Science 182 (4111), 443–449.
Pomeranz, K., 2000. The Great Divergence: China, Europe, and the Making of the Modern World Economy. Princeton University Press, Princeton.
Rau, H., Fahy, F., 2013. Introduction: sustainability research in the social sciences – concepts, methodologies and the challenge of interdisciplinarity. In: Fahy, F., Rau, H. (Eds.), Methods of Sustainability Research in the Social Sciences. Sage, London, pp. 3–24.
Rotmans, J., Kemp, R., van Asselt, M., 2001. More evolution than revolution: transition management in public policy. Foresight 3 (1), 15–31.
Schandl, H., West, J., 2012. Material flows and material productivity in China, Australia, and Japan. Journal of Industrial Ecology 16 (3), 352–364.
Schandl, H., West, J., 2010. Resource use and resource efficiency in the Asia-Pacific region. Global Environmental Change 20 (4), 636–647.
Sears, P.B., 1953. Human ecology, a problem in synthesis. Science 120, 959–963.
Sieferle, R.P., 1990. Bevölkerungswachstum und Naturhaushalt Studien zur Naturtheorie der klassischen Ökonomie Frankfurt a.M. Suhrkamp.
Sieferle, R.P., 1997. Rückblick auf die Natur: Eine Geschichte des Menschen und seiner Umwelt. Luchterhand, München.
Sieferle, R.P., 2001a. Europe's Special Course: Outline of a Research Program. Breuninger Stiftung, Stuttgart.
Sieferle, R.P., 2001b. The Subterranean Forest. Energy Systems and the Industrial Revolution. The White Horse Press, Cambridge (Original in German: 1982. Der unterirdische Wald. Energiekrise und Industrielle Revolution. C.H. Beck, München.).
Singh, S.J., Haberl, H., Chertow, M., Mirtl, M., Schmid, M. (Eds.), 2013. Long Term Socio-Ecological Research. Studies in Society Nature Interactions across Spatial and Temporal Scales. Springer, Dordrecht, Heidelberg, New York, London.
Smil, V., 1991. General Energetics. Energy in the Biosphere and Civilization. John Wiley & Sons, Manitoba, New York.
Smil, V., 2008. Energy in Nature and Society. General Energetics of Complex Systems. MIT Press, Cambridge, MA.
Snow, C.P., October 6, 1956. The two cultures. New Statesman, 52, 413.
Spencer, H., 1862. First Principles. Williams and Norgate, London.
Stanners, D., Bosch, P., Dom, A., Gabrielsen, P., Gee, D., Martin, J., Rickard, L., Weber, J.-L., 2007. Frameworks for environmental assessment and indicators at the EEA. In: Hak, T., Moldan, B., Dahl, A.L. (Eds.), Sustainability Indicators. A Scientific Assessment. SCOPE 67. Island Press, Washington, Covelo; London, pp. 127–144.
Tainter, J.A., 1988. The Collapse of Complex Societies. Cambridge University Press, Cambridge.
Thomas Jr., W.L., 1956. Man's Role in Changing the Face of the Earth. The Chicago University Press, Chicago.
Turner, B.L.I., Clark, W.C., Kates, R.W., Richards, J.F., Mathews, J.T., Meyer, W.B., 1990. The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years. Cambridge University Press, Cambridge.
UNEP., 2011. Decoupling Resource Use and Environmental Impacts from Economic Growth. United Nations Environmental Programme, Paris.
UNEP., 2013. GEO 5. Global Environmental Outlook. E-Book: http://www.unep.org/geo/GEO5_ebook/index.html.
Wackernagel, M., Rees, W.E., 1996. Our Ecological Footprint, Reducing Human Impact on the Earth. New Society Publishers, Gabriola Island, BC; Philadelphia, PA.
Weisz, H., Krausmann, F., Amann, C., Eisenmenger, N., Erb, K.-H., Hubacek, K., Fischer-Kowalski, M., 2006. The physical economy of the European union: cross-country comparison and determinants of material consumption. Ecological Economics 58 (4), 676–698.
Weisz, H., Steinberger, J.K., 2010. Reducing energy and materials flows in cities. Current Opinion in Environmental Sustainability 2 (3), 185–192.
White, L.A., 1943. Energy and the evolution of culture. American Anthropologist 45 (3), 335–356.
Winiwarter, V., Knoll, M., 2007. Umweltgeschichte. Eine Einführung, Böhlau, Köln.
Young, G.L., 1974. Human ecology as an interdisciplinary concept: a critical inquiry. In: Macfayden, A. (Ed.), Advances in Ecological Research. Academic Press, London, New York, pp. 1–105.

Social Ecology: The Chicago School

Anna-Lisa Müller, University of Bremen, Bibliothekstr, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents social ecology as used in the social and behavioral sciences, with special emphasis on the research conducted by the Chicago School of Sociology. By sketching the central concepts from plant and animal ecology, the author shows how these concepts were transferred to the study of human communities. Additionally, it is shown how social ecology altered the methods used in the social and behavioral sciences. In a final section, the reasons why the approach lost influence are given and personal and thematic successors of the approach presented.

Social ecology describes the conceptual adaptation of plant and animal ecology to the analysis of human communities within the social and behavioral sciences. The approach is also known as human ecology and was particularly used in the early twentieth century to explain the change of societies within their physical environment over time. Its central topic of research was the social and spatial structure of industrial societies, mainly in the context of urbanization. Research with an ecological focus was prominently carried out by sociologists at the University of Chicago from the 1920s to 1930s onward, later known as Chicago School of Sociology. The sociological research using ecological concepts will therefore be in this article's focus.

Origins of Social Ecology

Social ecology as field in the social and behavioral sciences roots in the times when disciplines like sociology, psychology, demography, or economics were still quite young scientific disciplines. In search for theoretical concepts and methodological frameworks, researchers started to use the concepts of ecology for the study of human communities in the late nineteenth and early twentieth centuries. In doing so, the researchers transferred the main concepts and explanatory structures of the (biological) theory of evolution to the changes of human communities. The community as basic unit of analysis in plant and animal ecology was then also used as unit of analysis in the social and behavioral sciences. The scientific analysis of human communities was the scope of the human ecology of that time. The general assumption was that both human behavior and the institutional order of a community were relative to time and space, i.e., that the development of a community had a spatial dimension and processual character. The development of human communities in general and differences between various human communities were explained on the basis of ecological concepts.

Next to the concept of community, concepts like dominance, succession, cooperation, competition, natural areas, and centralization/decentralization were adopted from the theory of evolution. They as well were ecological concepts now used to explain processes of human development and the structure of societies. Roderick McKenzie therefore defines human ecology as the "study of the spatial and temporal relations of human beings as affected by selective, distributive and accommodative forces of the environment" (McKenzie, 1925: p. 63f.).

Ecological Concepts and Their Transfer to Social and Behavioral Sciences

When transferring theoretical concepts from plant and animal ecology to human ecology, the life course of a plant or an animal was used as analogy to describe the development of a human community. The German ecologist Ernst Haeckel described the life course of an individual as consisting of three phases (Haeckel, 1866: pp. 76–81): youth/adolescence, adulthood/maturity, and age/decline. In the first phase, the young organism grows until maturity is reached. In this second phase, the organism is highly differentiated. In the following third phase, decline begins that ultimately leads to the death of the organism.

Translating these phases to human communities was used as a tool to explain the communities' development both in temporal and in spatial terms: during a first phase of development, a community extends in growth, both in numbers and in spatial terms, until reaching a point of equilibrium that corresponds to the human phase of maturity. In this second phase of development, the individuals in the community live in a defined area. After the equilibrium is exceeded, the community tends to decline in growth, again both in terms of numbers and its spatial spread.

The process of growth–stability–decline affects the forces at work in the community: during the first phase, the species of the community stand in competition to each other, each striving for dominance. In the second phase, the equilibrium is accompanied by an established social order where the species cooperate rather than compete in order to survive in the best way possible. In addition, the principle of dominance is at work: a fixed power relation is established that implies a social order in which one group dominates the others. During the third and last phase, the phase of decline, competition increases as the equilibrium is disturbed and social forces are released. Social disorder and economic destabilization are the results.

Main Ecological Concepts Used in the Social and Behavioral Sciences

Communities

The main topic of ecology is the analysis of communities – this is true for plant and animal ecology and for human ecology. In the nineteenth century, ecologists studied plant and animal communities to understand the "complex relations of all animals and plants to each other" (Darwin, 1937: p. 79), also called the "web of life" (Thomson, 1914: p. 263). The basic finding was that the species in a community, understood as the community's units, each have a specific role and are interrelated to the other species in the environment.

This concept of a community whose members share the same physical environment and interact in a specific way was then transferred to human communities (see, e.g., Park, 1915; McKenzie, 1925). Robert Park, one of the representatives of social ecology in sociology and a central figure of the Chicago School of Sociology, differentiates between biotic and cultural communities to describe the different elements of human communities. According to him, human societies vary from plant and animal societies in one respect: on top of an ecological order that builds the fundament of a human community, there exist three further orders: an economic, a political, and a moral order (Park, 1936: p. 14). All three account for the existence of what Park calls "cultural community" (Park, 1936: p. 9), and are the distinctive feature of human societies.

This understanding of human societies as comprising two forms of sociality relates to Ferdinand Tönnies (1991[1887]) differentiation of 'Gemeinschaft' and 'Gesellschaft'. The biotic community, a "Gemeinschaft", forms the basis for "the cultural superstructure" (Park, 1936: p. 13), i.e., the cultural community or 'Gesellschaft', in which complex forms of social control, moral norms, and institutions are responsible for stability, but also social change. By including the ecological concepts of competition and cooperation as characterizing forms of interaction in his model, Park stresses the fact that the human biotic community is, analogous to plant and animal communities, characterized by competition, whereas the cultural community is based on "custom and consensus" (Park, 1936: p. 13). With these differentiations, Park formulates the peculiarities of a decidedly 'human' ecology in contrast to 'plant/animal' ecology.

Natural Areas

By conceptualizing the physical environment of the communities as constitutive feature of their formation, Park, like his colleagues, integrates another important element of ecological theory, the role of the habitat, in the concept of the biotic and cultural communities. The development of communities leads to the formation of rather homogeneous sociospatial areas, in ecological terms 'natural areas' (e.g., Wirth, 1945: p. 485). Natural areas were originally thought of as the physical location of plant and animal communities, and this idea was adopted by human ecologists to describe the living environment of human communities. Natural areas are then understood as areas that are naturally grown environments of a community. In the context of human ecology, 'natural' does not have a biologistic meaning, but the natural areas of communities are natural in that sense that they appear to be natural, but are shaped by human interaction. Within these areas, specific norms, values, codes of acting, and institutions are cultivated. Examples for natural areas within a city are segregated areas of specific communities; prominent descriptions of, for example, gangs (Thrasher, 1927), Jewish communities (Wirth, 1928), and Black communities (Frazier, 1932) were done for Chicago by the Chicago School of Sociology.

Competition and Cooperation

Two central ecological concepts, competition and cooperation, were already mentioned in the above description of the structure of communities. Robert E. Park explicitly refers to Charles Darwin to trace the principle of "competitive cooperation" (Park, 1936: p. 3) as a guiding principle of human life – a principle that, according to Park, is per se sociological (Park and Burgess, 1921: p. 194; Park, 1936: p. 2f.) and was used by Darwin (1937: p. 85) to describe the mutual dependency of species in plant and animal communities.

According to human ecologists, competition and cooperation are features of both plant and animal communities and human communities. Competition is understood as a principle that characterizes the simultaneous existence of species. Species compete for the best position in a given environment, both in terms of food and reproduction. Transferred to the analysis of human communities, competition is mainly understood in economic terms: each individual within a community strives for the best possible economic position.

But as plant and animal ecologists observed, it is not only competition that characterizes communities. In order to survive in the best possible way in a given physical environment, species were thought of as cooperating with other species around them. By cooperating, species were able to improve their position within the community. For human communities, cooperation becomes manifest, e.g., in functional differentiation and economic specialization.

Dominance

The relation of the species within a community is additionally characterized by the principle of dominance. Ecologists assumed that in each natural area there exists one species of dominance. This assumption includes the concept of a hierarchy of species in a given environment: one species dominates the others and is thus the head of a specific order, the other species being subordinate to the dominating species in the respective community. Using the principle of dominance, human ecologists describe the social structure of a community. Transferring the main aspects of the principle of dominance to human communities, Roderick McKenzie argues that each community has its own social order and a dominating form of occupation: in communities of hunters and gatherers, communities tend to be small and of temporary character. In agricultural communities, the communities' character is more permanent. In communities with trade and commerce as dominant form of occupation, they are even more permanent

and concentrate at transportation hubs like streams, railroads, etc. (McKenzie, 1968: p. 5).

Succession

Another principle formulated by plant and animal ecologists is that of succession. If the principle of dominance describes the social structure of a community, the principle of succession describes its temporal-spatial structure. The concept of succession is used by ecologists to explain the growth of plant and animal communities or, in other words, the changes a community goes through in relation to time (change of community) and space (community's habitat). Human ecologists applied this model to human societies to explain the growth of industrial cities like Chicago at the beginning of the twentieth century and to the related changes in the social structure of the society.

Based on this ecological concept, models of city growth were formulated. One influential model was designed by Ernest W. Burgess (1925) who conceptualized the city as comprising of concentric zones. Burgess used the model to explain the growth of industrial cities by emphasizing the processual character of a city's development. To Burgess, a city comprises different zones, each zone being the natural area of a specific community and a certain form of land use. The location of a community is a result of the communities' competition for land use and dependent on the value of land in the respective zone. The growth of population, caused by immigration and a positive birth rate, leads to a spatial expansion of the city over time. This processual development was thought of as being regular and occurring 'in cyclic fashion' (McKenzie, 1968: p. 31).

Centralization/Decentralization

The physical form of the locations of human communities and its development imply processes of centralization and decentralization. The tendency to centralize was understood as characteristic of human communities, resulting from "the tendency of human beings to come together at definite locations for the satisfaction of specific common interests" (McKenzie, 1968: p. 26). In societies of the early twentieth century, the central location was a city's downtown area with shopping, entertainment, and business facilities. In the course of the development of cities, different zones loose and gain importance, thus leading to a process of centralization and decentralization of the whole city (Burgess, 1925; McKenzie, 1968).

Reasons for Using Social Ecology in the Social and Behavioral Sciences

At the beginning of the twentieth century, ecology and its concepts were an influential factor for the formation of social and behavioral sciences. Social scientists were confronted with social phenomena so far unknown to them in the late nineteenth and early twentieth centuries. In the course of industrialization, both the relationships between different individuals in a society and between individuals and their physical environment changed. A rapid growth of cities and the transformation of the social structure were observable.

The adaptation of ecological concepts for the analysis of human societies proved appropriate because of two reasons: the ecological theories describing the development of animal and plant communities gave an analytic framework to explain the interrelatedness (1) of individuals and groups within a community and (2) of the community and its environment. By using ecological concepts of how communities develop, different phases of this process could be distinguished and a temporal dimension could be added to the analysis of human communities. Thus, it was possible to describe the development of a society as a process and not as a mere sequence of static stages. Additionally, by analyzing the interrelation of a community and its physical environment, the spatial dimension of the development of human societies were taken into account.

Within the social and behavioral sciences, social ecology was most frequently used in sociology, namely, by the Chicago School of Sociology (Park and Burgess, 1921). But it also altered the methods and research topics in economics, geography, anthropology, or psychology. In general, understanding human communities as interrelated with their environment was the substantial change in regard to contents in these disciplines.

In sociology, ecological concepts and metaphors helped the young discipline to develop coherent theories and models of the development of human societies. Roderick McKenzie (1925) emphasized the difference between ecology and human ecology and formulated distinctive characteristics of an ecological analysis of living species: to him, the distinguishing skills of humans were the ability to be mobile and to adjust the habitat according to the own needs – skills that plants do not and animals hardly possess (McKenzie, 1925: p. 64). This ability enables human individuals to adjust the environment according to their needs; this alters the relationship between humans and their environment and marks a central difference to plant and animal communities' relation to their respective habitat.

The spatial and temporal dimensions of community development were thus the topics of human ecological research as carried out by the Chicago School of Sociology (McKenzie, 1925: p. 64). Phenomena related to urbanization, central for the studies of the Chicago School, were interpreted using ecological concepts. For example, the city's spatial expansion and its social effects were explained with the help of the concepts of succession and centralization/decentralization.

The forms of using ecological concepts differed: on the one hand, the concepts of ecology were applied to the study of human communities with only slight changes: the community was taken as the basic unit of analysis, and the mutual dependency of the community's members was researched in its relation to the spatial environment and its temporality. On the other hand, the terms of ecology were used in a metaphorical way to describe the human environment: Burgess speaks about the 'metabolism of the city' (Burgess, 1925: p. 48) when describing and explaining the processes of a city's expansion and the mobility of its population. The metaphor of a metabolism refers to the intrinsic logics of change of a city's development.

The merit of social ecology as developed by the Chicago School of Sociology was the fact that it provided a terminology and a methodological framework for the analysis of social processes in industrial societies, especially in the context of urbanization. The sociological concept of segregation, until today widely used by social scientists to explain sociospatial phenomena like the spatial distribution of social groups according to economic capital, was developed on the basis of ecological concepts.

In early demography, an important topic of research concerned the distribution of a community's population within a given environment and during the course of the day. Thus, both the spatial and the temporal distributions of the population were in focus. Understanding the human community as a group that follows the principle of succession within a given environment enabled the researchers to explain mobility in a city: in the course of the development of a community, various social groups compete for land in, e.g., a city. Depending on the social order of the community, each social group is assigned a specific natural area. Parallel to the growth of population, the location of the social groups and their specialization extend. Thus, mobility as phenomenon of community development is bound to demographic developments and economic and social differentiation. Empirical studies researching the spatial patterns of mobility of the city's inhabitants were part of demographic research.

In the 1920s, ecology also influenced the research carried out in geography, with Harlan H. Barrows (1923) even understanding geography as synonymous with human ecology. Here, ecological concepts were used to explain the interrelation of human communities and their geographical environment. In contrast to sociological analyses, it is not only the spatial dimension of a community that is integrated in the analysis (Park, 1967: pp. 69–84), but the geographical environment with its decidedly geographical peculiarities.

The idea that competition and cooperation are structuring elements of community development made social ecology also applicable in the field of economics (Boulding, 1970). In psychology (Barker, 1968; Proshansky et al., 1970) and anthropology (Steward, 1955), ecological concepts influenced the analysis of social behavior and the disciplines' methodology. This includes comparative research of communities, in a diachronic as well as in a synchronic perspective, and the identification of stages in the communities' development patterns.

Methods

Besides influencing the theories of the social and behavioral sciences, ecology also altered the methods applied in the respective disciplines. Park describes understanding and comprehending the characteristics of local community culture as the central challenge of community research. Two different approaches to the study of human communities were established with the help of ecological research questions: a quantitative and a qualitative approach.

The quantitative approach comprised methods to analyze spatial patterns of a community's population. A method used frequently was social area analysis. It was a means to understand phenomena like community patterns and segregation on the basis of census data (e.g., McKenzie, 1923; Shevky and Bell, 1955). In order to improve the results of such studies, Roderick McKenzie even suggested to set up an official statistical office to provide data for research (McKenzie, 1968: Chapter 3).

The qualitative approach was influenced by journalistic practices like on-street-interviewing and study trips to the community in focus. Several empirical studies using fieldwork, observation, and interviews were done at the University of Chicago (Frazier, 1932; McKenzie, 1923; Thrasher, 1927; Wirth, 1928), with this beginning to establish the ethnographic fieldwork as method in the social sciences. Robert Park, working as journalist himself, was an influential figure in establishing methods like participant observation and the usage of fieldnotes as data (e.g., Park, 1967: Chapter 3). In anthropology, a similar change in methods took place: instead of using informants to get data on the phenomenon, researchers started to go to the people themselves.

Generally, there was not a divide between the methods used in human ecological research along the lines of a quantitative vs a qualitative approach. Both sides acknowledged the importance of the respective other set of data and methods used, and often a combination of different methods was proposed.

Reasons for the Loss of Influence

In the course of the twentieth century, social ecology lost influence in the social and behavioral sciences. Four reasons might explain this loss of influence: (1) the approach's inability to deal with phenomena of decline, (2) conceptual problems when dealing with exogenous factors of growth and externally employed forces to control development, (3) a too strong focus on economic factors to explain developments, and (4) the inability to explain internal, subversive processes that counteract processual (in the sense of linear) developments.

Phenomena of Decline

The transfer of ecological concepts to the study of social communities took place during a phase of industrialization and urbanization. The phenomena observed and studied were all phenomena of growth. Burgess developed his model of city growth as an idealized model of Chicago that only works under the condition of growth. Succession – spatial expansion and population growth – is one of its main features. A decline of the community in terms of numbers and spatial dimension is understood as a result of the community's cyclical development and causes economic instability and social disorder and unrest. According to social ecology and using the metaphor of an organism's life cycle, this process is comparable to the death of the organism.

In times of economic and demographic decline and shrinking settlements, observable in many societies in the late twentieth and early twenty-first centuries, the basic ecological assumptions can in fact explain phenomena like shrinking cities and the destabilizing effects of deindustrialization and economic decline. But they cannot be used to analyze the

complex processes in which humans interfere with the processes of decline in order to cope with the effects of population decrease and deindustrialization: as soon as externally controlling forces like global financial crises, natural disasters, or the like enter the stage, ecological analysis reaches its limits.

External Factors to Control Development

In ecology, growth as central aspect of the development of communities is understood as endogenous, i.e., as an internal factor influencing the process of development. Again taking Burgess' model as reference, city growth is accompanied by the expansion and segregation of social groups within the physical environment. The tendency to centralize and decentralize is based on economic factors like land value and not on controlling influences by e.g., city planners. In the course of the twentieth century, city planning was an increasingly important task to cope with the effects of urbanization. In order to improve the living conditions in cities, a number of institutions were set up to control and regulate a city's development.

Economic Focus

Ecology used economic arguments to explain the development and the structure of communities: the more economic capital people possess, the farer away from the city's core they live and the richer the communities in the suburban fringes. Human ecology adopted this argumentation to explain the social and spatial structure of human communities. In the course of disciplinary differentiation, economic explanations were said to be not sufficient any more. The development of human communities and their settlements were increasingly understood as a result of a complex set of economic, social, ecological, political, and cultural factors which made new conceptual approaches necessary.

Forces Counteracting Processual Development

Ecological analysis of communities in general and human ecological analysis in particular understood the structure of communities as homogenous, comprising a fixed hierarchy. Not in focus were the continuous internal struggles for dominance, to be observed in the parallel existence of hegemonic and subversive discourses and in the differing human practices within a community. Thus, neither the various forms of power and counterpower nor the peculiarities of an individual's everyday practices can be taken into account in ecological analyses, among other things because the community is the unit of analysis and not the single individual.

Follow-Up in the Late twentieth Century: the Successors of Social Ecology

In the course of the twentieth century, social ecology lost influence in the social and behavioral sciences. Although the general idea of social ecology – the interrelatedness of human communities and their spatial environment – was still used as explanatory model, research with an explicitly ecological approach was generally secondary. But it also had successors. The successors can be differentiated into a direct modification of social ecology, personal successors, and thematic successors, exemplifying different forms of influence of social ecology. A prominent personal successor is Herbert Blumer. Based on the works of George Herbert Mead, he developed further the concept of symbolic interactionism and was highly influenced by the two Chicago Schools, those of Sociology and of Philosophy. Prominent thematic successors are the sustainability studies and the subdiscipline of environmental sociology. Both disciplines emerged in the second half of the twentieth century.

Modification of Social Ecology

Later researchers modified social ecology in the social sciences as used at the beginning of the twentieth century. In these later researches, the original concepts of ecology and their adaptation for the study of human societies were criticized as being too simplistic (Hawley, 1944). Additionally, a more macro-level approach was taken on the analysis of communities within a given environment, emphasizing the collective dimension of communities (Hawley, 1986). Central topics of research were the organization of society, understanding it as a system, and the interrelation of functionally differentiated parts (Duncan et al., 1959).

Symbolic Interactionism

A different form of follow-up can be seen in the case of symbolic interactionism, embodied by George Herbert Mead and Herbert Blumer. Mead and Blumer, both connected to the Chicago Schools of Sociology and Psychology, developed a theory of human interaction that was later called symbolic interactionism (Blumer, 1969). In general, Mead's and Blumer's interest was in the forms of human interaction within a given environment, a topic not far from the research interests of human ecologists. According to symbolic interactionism, an individual's personality is formed within society, i.e., a human community. A central characteristic of human communities is interaction, among other things with the help of symbols. By constantly acting and interacting, social norms and values are created. Mead understands gestures as the initials of acting and is with this argument critically referring to, among others, Charles Darwin. Thus, both human ecology and symbolic interactionism share Darwin as intellectual predecessor and a general interest in the interrelation of human individuals in a community and a given environment.

Sustainability Studies

Since the 1990s, sustainability studies are a new field in which a renaissance of social ecology can be noted. Interestingly, an early article by Radhakamal Mukerjee (1932) already stresses the central aspects of sustainability studies, identifying them as topics of ecological research. By referring to plant and animal ecology, Mukerjee stresses the fact that nature has its own cycle of development. If humans interfere too much with this cycle, they tend to exhaust natural resources, which ultimately destroy the basis of human life (Mukerjee, 1932: p. 353). By demanding a 'bioeconomic cooperation'

(Mukerjee, 1932: p. 353) of humans and nature, he anticipates the sustainability movements of the late twentieth and early twenty-first centuries. By conceptualizing the society as comprising an ecological, an economic, and a social dimension, Park (1936) formulates what will later be the central topic of sustainability studies.

Environmental sociology, thematically related to sustainability studies, is an additional subdiscipline using ecological concepts (Catton and Dunlap, 1978).

See also: Culture and Economy; Demography, Early History of; Development and Urbanization; Ethnography; Methodology of the History of the Social and Behavioral Sciences.

Bibliography

Barker, R.G., 1968. Ecological Psychology. Concepts and Methods for Studying the Environment of Human Behavior. Stanford University Press, Stanford.

Barrows, H.H., 1923. Geography as human ecology. Annals of the Association of American Geographers 13, 1–14.

Blumer, H., 1969. Symbolic Interactionism. Perspective and Method. Englewood Cliffs, New Jersey.

Boulding, K.E., 1970. Economics as a Science. McGraw-Hill, New York.

Burgess, E.W., 1925. The growth of the city. An introduction to a research project. In: Park, R.E., Burgess, E.W. (Eds.), The City. University of Chicago Press, Chicago, pp. 47–62.

Catton, W.R.J., Dunlap, R.E., 1978. Environmental sociology: emergent forms and issues. American Sociologist 13, 41–49.

Darwin, C., 1937. The Origin of Species, the Harvard Classics. P.F. Collier & Son Corporation, New York.

Duncan, O.D., Schnore, L.F., Rossi, P.H., 1959. Cultural, behavioral, and ecological perspectives in the study of social organization. American Journal of Sociology 65, 132–153.

Frazier, E.F., 1932. The Negro Family in Chicago. University of Chicago Press, Chicago.

Haeckel, E., 1866. Generelle Morphologie der Organismen. Verlag von Georg Reimer, Berlin.

Häußermann, H., Siebel, W., 1994. Gemeinde – und Stadtsoziologie. In: Kerber, H., Schmieder, A. (Eds.), Spezielle Soziologien. Problemfelder, Forschungsbereiche, Anwendungsorientierungen, Rowohlts Enzyklopädie. Rowohlt, Reinbek bei Hamburg, pp. 363–387.

Hawley, A.H., 1944. Ecology and human ecology. Social Forces 22, 398–405.

Hawley, A.H., 1986. Human Ecology: A Theoretical Essay. University of Chicago Press, Chicago.

McKenzie, R.D., 1923. The Neighborhood: A Study of Local Life in the City of Columbus. University of Chicago Press, Chicago.

McKenzie, R.D., 1925. The social ecology to the study of the human community. In: Park, R.E., Burgess, E.W., McKenzie, R.D. (Eds.), The City. University of Chicago Press, Chicago, pp. 63–79.

McKenzie, R.D., 1968. On Human Ecology. Selected Writings, the Heritage of Sociology. University of Chicago Press, Chicago, London.

Mukerjee, R., 1932. The ecological outlook in sociology. American Journal Sociology 38, 349–355.

Park, R.E., 1915. The city: suggestions for the investigation of human behavior in the city environment. American Journal Sociology 20, 577–612.

Park, R.E., 1936. Human ecology. American Journal Sociology 42, 1–15.

Park, R.E., 1967. On Social Control and Collective Behavior. Selected Papers, the Heritage of Sociology. The University of Chicago Press, Chicago and London.

Park, R.E., Burgess, E.W., 1921. Introduction to the Science of Sociology. University of Chicago Press, Chicago.

Proshansky, H.M., Ittelson, W.H., Rivlin, L.G. (Eds.), 1970. Environmental Psychology. People and Their Physical Settings. Holt, Rinehart and Winston, New York.

Shevky, E., Bell, W., 1955. Social Area Analysis. Theory, Illustrative Application and Computational Procedures. Greenwood Press, Westport, CT.

Steward, J., 1955. Theory of Culture Change: The Methodology of Multilinear Evolution. University of Illinois Press, Urbana.

Thomson, A.J., 1914. The Wonder of Life. Henry Holt and Company, New York.

Thrasher, F., 1927. The Gang. University of Chicago Press, Chicago.

Tönnies, F., 1991[1887]. Gemeinschaft und Gesellschaft: Grundbegriffe der reinen Soziologie. reprint of 8th ed. from 1935, 3rd, unrevised ed. Wissenschaftliche Buchgesellschaft, Darmstadt.

Wirth, L., 1928. The Ghetto. University of Chicago Press, Chicago.

Wirth, L., 1945. Human ecology. American Journal Sociology 50, 483–488.

Social Enquiry and Action Research for Social Work

Purnima George and Danika Syrja-McNally, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Critical social enquiry (CSE) connotes research frameworks that originated from the critical social science paradigm and those that have been influenced by feminist, antiracist, anticolonial, poststructural, and postmodern perspectives. This article traces the development of various frameworks of CSE such as action research, participatory action research, and community-based action research associated mainly with critical social science and antioppression research that blends the influences of the critical social science with other theoretical perspectives identified above. These frameworks share a commitment to community participation, social justice, and challenging existing ways of conducting research.

Introduction

The term critical social enquiry (CSE) connotes research frameworks that originated from the critical social science paradigm and those that have been influenced by feminist, antiracist, anticolonial, poststructural, and postmodern perspectives. The initial part of this article presents some of the research frameworks such as action research (AR), participatory action research (PAR), and community-based action research (CBAR) associated mainly with critical social science. The latter part of this article presents the newest form of CSE; anti-oppressive research (AOR) that blends the influences of the critical social science paradigm with the varied theoretical perspectives identified above. All of these frameworks share a commitment to community participation and strive to democratize knowledge, raise awareness of processes of oppression, and engage in social justice (Lavoie et al., 2010). The areas of each approach discussed focus on their basic definitions, roots, methodologies, and limitations.

Critical Social Enquiry

CSE is defined by Cannella and Lincoln (2009) as any research that recognizes the operation of power as reifying unjust and oppressive social conditions as social norms. It is also concerned with issues of identity, which are seen as major shapers of structures of oppression (Cannella and Lincoln, 2009). It is concerned with the emancipation of marginalized people and is firmly political in its intention and process (Alvesson and Sköldberg, 2009; Strega, 2005). Carroll (2004) provides three overlapping images of the critical social inquiry: oppositional, radical, and subversive. He describes the oppositional image of critical inquiry as research, which opposes the status quo through the investigation and critique of social structures and practices of domination. In this image, researchers align themselves politically with the oppressed. The radical image or analysis investigates the challenges researchers faced when exploring the deeper systemic factors influencing oppressions. Thirdly, the subversive image looks for strategies to unsettle the existing arrangements by interrogating 'common sense' binaries that seek to normalize the status quo (Carroll, 2004: p. 3). This process opens doors to an alternative understanding of reality.

Alvesson and Sköldberg (2009) provide an overview of the influence of critical theory in research. They argue that critical theory views people as manipulated, objectified, passive, and conformist in relation to dominant forms of rationality. However, it also views people as having the potential to be autonomous, reflexive, reflective, and critical. Although critical theory is usually thought to be more appropriate for qualitative research, it can also be used in quantitative research.

The history of critical research, which sought to contest entrenched power, can be traced back to the nineteenth century (Carroll, 2004). It originated from sociology, psychoanalysis, philosophy, economics, and aesthetics and was combined into the social science research environment of the Frankfurt school at the end of the 1920s (Alvesson and Sköldberg, 2009). The Frankfurt school critiqued positivism and developed critical theory that was philosophically informed and politically significant (Alvesson and Sköldberg, 2009). Leaders in critical theory included figures such as Marx, Weber, Kant, Hegel, Freud, and Habermas (Alvesson and Sköldberg, 2009). For an example of critical theory written by a social worker refer to Fook (2002).

Since the 1990s, there has been an emerging influence of poststructuralism providing a more practically grounded radicalism to inquiry (Denzin and Giardina, 2009; Ferguson, 2003). Cannella and Lincoln (2009) also note the influence of indigenous and postcolonial perspectives that view the foundations of social enquiry situated in colonial relations of power (refer to Smith, 1999). This new thinking has significantly influenced and transcended the purpose of CSE.

The purpose of CSE has undergone a significant change from its original intent. The purpose has shifted from social justice and empowerment of communities also to include attention on challenging disempowering social relationships that serve to maintain status quo. The purpose of CSE, according to Cannella and Lincoln (2009), is to illuminate the processes of power construction and how these processes disempower groups of people, their knowledges and ways of being. It seeks to uncover myths, reveal hidden truths, and help people change the world for themselves (Strega, 2005). In CSE, knowledge is presumed to be influenced by sociohistorical and political contexts and implicated in and organized around relations of power (Carroll, 2004; Denzin and Giardina, 2009). CSE, therefore, works toward deconstructing discourses, which

maintain these power relations and myths to influence public consciousness and transformation (Cannella and Lincoln, 2009). In this way, CSE opposes not only the traditional notions about the neutral nature of research (Alvesson and Sköldberg, 2009), but it also challenges notions about the neutrality of the researcher (Denzin and Giardina, 2009). Critical social researchers take a stand to change the world, not just interpret it, in ways that resist social injustice and celebrate freedom and democracy (Denzin and Giardina, 2009).

Alvesson and Sköldberg (2009) speak of some of the limitations of CSE. They argue that critical inquiry has the ability to reinforce or question established ideologies without actually adopting a position with either, since it focuses more on preventing research from contributing to dominance rather than on actually overcoming existing dominance.

Another major limitation identified by Ferguson (2003) about CSE and critical theory is its emphasis on negative critique. He argues that the prescribing nature of critical enquiry about what should not be done instead of what can be done and how presents a problem. It may be difficult for change to actually happen in the current sociopolitical climate with the idealized image of practice supported in CSE (Ferguson, 2003).

Action Research

AR is a process that "gives credence to the development of powers of reflective thought, discussion, decision and action by ordinary people participating in collective research on 'private troubles' that they have in common" (Adelman, 1993: p. 8). This definition focuses on developing critical consciousness and developing the capacities of people to take action on problems. It is a collaborative approach that endorses "consensual, democratic, and participatory strategies to encourage people to examine reflectively their problems or particular issues affecting them or their community" (Berg, 2007: p. 224) paying special attention to process. Reason and Bradbury (2008) view AR as less of a methodology, but more of an orientation to inquiry, which seeks to create participative communities focused on action, reflection, and practice in the pursuit of practical solutions to issues of concern. AR is also about creating new forms of understanding for reflection, action, and theory (Reason and Bradbury, 2008).

Adelman (1993) notes, "Kurt Lewin (1890–1947) created the term 'action research' in 1934." It has evolved as a research framework utilizing many different traditions like psychology, anthropology, and sociology (Berg, 2007). Frameworks that use principles of AR include civil rights and feminist movements; liberationist adult and trade union education; postcolonial and critical race theory; and antiwar and ecological protests as well as the student democracy movement (Gayá Wicks et al., 2008).

The purpose of AR is to change or improve what it seeks to study (Berg, 2007: p. 226). In AR, researchers stand with and alongside communities of study (Berg, 2007). Reason and Bradbury (2008: pp. 3–4) discuss critical values that shape AR such as the emphasis on the engagement of people in collaborative and transparent relationships that open up new spaces for dialogue. Two main themes in AR are the importance of practice and life experiences and the web of relationships, events, influences, role models, and experiences in the process of research (Gayá Wicks et al., 2008). Action researchers place life experiences among the primary influences that underpin their research (Gayá Wicks et al., 2008).

Berg (2007) describes three strategies of conducting AR as "technical/scientific/collaborative", "practical/mutual/collaborative/deliberate", and "emancipating/enhancing/critical" (p. 231). In the technical/scientific/collaborative strategy of AR, the researcher identifies a problem using a prespecified theoretical framework and collaborates with a practitioner. The practitioner relays the information of the study to the group. Neither the practitioner nor the community are consulted on the urgency or relevance of the issue, as it solely relies on the expertise of the researcher. In the practical/mutual collaborative/deliberate strategy, the researcher and practitioner define the problem and assess the situation together. The practitioner then informs the community but only in the implementation stage of research. The third strategy, the emancipating/enhancing/critical AR, "embraces a greater concern for empowering and emancipating stakeholders working with the practitioner" (p. 231). As suggested by Berg (2007), this strategy emphasizes involvement of community in identifying and defining issues most critical to their lives as well as working with the community in undertaking research and action on the issue. The emphasis in this strategy is on developing critical consciousness and addressing issues of power between the researcher and the community. While Berg (2007) has described these strategies as theoretically distinct with varying foci, in reality, the research process could look different with researchers using combinations of these strategies. For instance, instead of a researcher being invited to work with a community, a researcher may define the problem with a practitioner and then focus on developing critical consciousness among community members.

Reason and Bradbury (2008) view AR as an emergent process that cannot be predetermined as the methodology of research is developed through the understanding of issues developed by the community and researcher. It uses traditional methods of data collection but have nontraditional orientations and purposes (Berg, 2007).

One of the limitations of AR identified by Reason and Bradbury (2008) is its vulnerability to being co-opted by dominant groups to maintain the status quo. Another limitation of this approach is the existing belief among social scientists that AR is only activism and void of rigorous theories and methods (Levin and Greenwood, 2011). Levin and Greenwood (2011) note that the lack of theoretical and methodological contributions that can be gained through its process can deter researchers from embracing it. Finally, Rutman et al. (2005) critique the approach for its belief that it is possible to achieve a solution that will address the issues experienced by all people involved. This belief fails to consider the operation of power within the community, therefore, potentially affecting engagement of all community members in research and action (Rutman et al., 2005; Tandon, 2008; for an example of AR, see Ku, 2011).

Participatory Action Research

PAR represents a philosophical approach rooted in social justice bringing together inquiry, learning, critical analysis,

community building, and social change (Rutman et al., 2005). It is a critical strategy deeply rooted in the experiences of the colonized 'other' (Carroll, 2004) providing a framework for people who seek to overcome oppressive situations and understand how social forces operate (Tandon, 2008). Terms used by theorists of PAR include empowerment, critical consciousness, transformation, conscientization, dialogue, social action, and participation to characterize different aspects of the research (Tandon, 2008).

PAR is informed by many social struggles and movements (Brydon-Miller et al., 2011). PAR was formed through the efforts of mass social struggles over the independence of colonies in Asia, Africa, and Latin America at the end of World War II (Tandon, 2008). According to Tandon (2008), it was not until the 1970s that PAR became popular as a result of the influence of Paulo Freire (1970) who developed it as a tool for fighting oppression through the practice of adult education and community organizing in Latin America. PAR has also been shaped by the social struggles of feminist, labor activist, civil rights workers, and other activists (Brydon-Miller et al., 2011).

The explicit aim of PAR is to bring about a more just society in which no group or class of people suffer from the deprivation of life's essentials, such as food, clothing, shelter, and where all enjoy basic human freedom and dignity (Alvarez and Gutierrez, 2001). PAR also recognizes the necessity of marginalized people be involved in the process of change (Brydon-Miller et al., 2011). Therefore, PAR's methodology focuses on collaboration, political engagement, and an unfailing commitment to social justice and values the collaboration and knowledge of marginalized people affected by an issue (Brydon-Miller et al., 2011).

According to Gutierrez (2003), PAR projects differ from other types of research in the significance given to process. There are many different ways a researcher can begin a PAR (Gutierrez, 2003). Researchers could either be invited into the community as an outside consultant, be a well-informed member of an organization on the issue, or be invited to work with community members (Gutierrez, 2003). However, Gutierrez (2003) notes that it is not how one begins a research project that is of importance but how the process develops democratically over time.

During the initial part of the project, the researcher engages and organizes the community's attention on an issue similar to a community organizer (Gutierrez, 2003). The researcher then meets with community members to discuss the issue and engage them in deciding a course of action most relevant to them. Researchers and community members then discuss and plan their nature of inquiry, methods of data collection and analysis, and how they will use the findings of their inquiry. Although there has been little written about approaches and methods used by researchers, Photovoice has been a popular method used by many researchers. Photovoice is a method whereby people photograph their own realities and use those photographs to engage in dialogue and explore ways of taking action (Alvarez and Gutierrez, 2001; for information on photovoice, see Wang, 2006).

PAR shares with AOR an understanding that researchers are knowledge producers and are positioned in a power relation for change (Rutman et al., 2005). In PAR, there is less emphasis on locating the researcher than there is in AOR. However, PAR holds the assumption that the researcher and the researched have already developed a critical consciousness of power and that they would not be working together if they were not closely allied (Rutman et al., 2005). This implies that the exercise of locating oneself as a researcher is unneeded during the process (Rutman et al., 2005).

Researchers who believe in PAR have responded to the critique on lack of methodological rigor in the enquiry. Gutierrez (2003) provides insight into how methodological rigor is evaluated for this form of enquiry. She states that validity of data in this approach can be measured by participants' empathy levels, understanding of the intent of a question (face validity), willingness to give needed information (construct validity), triangulation of different methods, and interactive knowledge. It also depends on whether the knowledge shared can act as a catalyst for empowering participants to overcome obstacles to oppression.

Alvarez and Gutierrez (2001) note that many researchers, drawn by the goals and potential of PAR may engage in this sort of research without fully understanding it and subsequently risk reproducing dominance. Alvarez and Gutierrez (2001), Park (1993) and Rutman et al. (2005) argue that without a strong understanding of PAR, researchers risk tokenizing community members. Therefore, it is imperative for researchers and the researched to spend time working through and agreeing on how to conduct research.

Rutman et al. (2005) critique PAR's attempts to create a democratic work environment may not negate power differentials in responsibility, authority, and inability to contribute equally toward research due to life circumstances. Recognizing that there are power differentials in the relationship among the community and between researchers allows for more honest discussions in clarity about roles within a research team. For an example of PAR refer to Gellis (2001).

Community-Based Action Research

The literature on this approach uses different terms such as CBAR, community-based participatory research, and community action research to denote research done with communities. CBAR is an approach to research that focuses on local action and strategies to address specific problems in specific situations (Stringer, 2007). Stringer (2007) notes that cooperation and consensus are the primary orientation of this research and that it provides a climate suitable for groups to work together to achieve a set of goals. Through CBAR, participants learn the skills to engage in research that they can later apply to other issues. CBAR recognizes the importance of this and therefore the importance of members acting as key partners in research (Briar-Lawson et al., 2010).

CBAR evolved out of similar traditions mentioned in the PAR (Reitsma-Street and Brown, 2004). However, CBAR seems to have tailored its focus more on local and community action approaches to enquiry and also emphasizes cooperation and consensus instead of more structural change as is more the goal of PAR (Reitsma-Street and Brown, 2004; Stringer, 2007).

Reitsma-Street and Brown (2004) argue that the purpose of CBAR is community well-being and for Briar-Lawson et al. (2010)

and Stringer (2007), it also means emancipation and community building, such as capacity building, asset enhancing, and social capital development. Its four values are social justice, agency, community connectedness, and critical curiosity for change (Reitsma-Street and Brown, 2004). Through group dialogue people learn about oppression, identify common connections to their own lives, and envisage a social world that accounts for those connections.

Reitsma-Street and Brown (2004) state that fundamental to this research is the careful and continuous negotiation between whose voice is emphasized, what strategies are utilized, and on whose behalf research is conducted. Unique to CBAR, and different from PAR is its intentional effort to include many participants and capture the reactions of multiple audiences in the recognition that among community members, there are "sharp differences in the lifeworlds and futures of its members" (p. 304).

Reitsma-Street and Brown (2004) also describe three approaches to inquiry used in CBAR; the first set uses commonplace qualitative and quantitative methods such as life-history narratives or statistical analysis of communities. The second set uses sustained, democratic dialogue, and other shorter research activities. The third deals with theorizing work in research. Through democratic dialogue on a local level, data is interpreted in a way relevant to different audiences.

Like any other research all of the approaches of CSE whenever undertaken with academic or institutional partner require ethics approval, which may change the way CSE is conducted (Flicker et al., 2007). Minkler (2004) and Flicker et al. (2007) have discussed the limitations of using ethics protocols used for conventional research for community-based research such as CSE. According to them, issues ignored by the current protocols include questions such as who is initiating the research, whether the issue selection is driven by the community, who benefits the most from the research, whose voice is being heard and represented, and what constraints do community members experience regarding involvement. Other ethical issues relate to the sharing and release of findings, use of data and representation of the community through the data, challenges in the action dimensions of research, and insider–outsider tensions (Minkler, 2004; Flicker et al., 2007). However, indigenous researchers have addressed these ethical dilemmas and have created guidelines in response to these issues, for instance in the Canadian Institutes of Health Research guidelines which promote accountability, transparency, and partnership in research (Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, accessed December 2010, http://www.pre.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf).

CBAR values subjective interpretations of relevance and authenticity by the community, qualitative research standards of trustworthiness and congruence, and quantitative statistical standards of reliability, validity, and generalizability as its standard practice (Reitsma-Street and Brown, 2004).

Reitsma-Street and Brown (2004) critique CBAR for its almost exclusive focus on local change efforts and its failure to discuss how one arrives at an understanding of injustice for a certain situation and strategies to change them. CBAR is then at risk of being "co-opted for the benefit of those already holding privilege and power" (p. 316).

A challenge identified by Briar-Lawson et al. (2010) is attaining and maintaining community-based processes. The time commitment required of participants can be a major deterrent to participating, as participants are required to balance many practice and research demands. Stringer (2007) points out the "pressures of bureaucratic life, the limitations of resources, and the competitive push for career advancement" are constantly working against "attempts to humanize and democratize research processes" (p. 190). For an example of CBAR refer to Reitsma-Street and Neysmith (2000).

Antioppressive Research

AOR has emerged as the latest form of CSE. It draws on a number of theoretical frameworks in the pursuit of research and recognizes a relationship between producing knowledge about the world and the actual practice of doing research exists (Potts and Brown, 2005). It also recognizes that knowledge is socially constructed and located within power relations (Potts and Brown, 2005; Strega, 2005). Knowledge is also considered to be situated and subjugated by social location as a result of experiences of privilege and oppression (Moosa-Mitha, 2005).

AOR is as much concerned about the process as it is the outcome. The process itself is viewed as an outcome with the understanding that social justice cannot be achieved without the research process being purposeful and intentional as the focus of the study could potentially objectify 'the oppressed' otherwise (Potts and Brown, 2005: p. 267). Potts and Brown (2005) present three main tenets of AOR. First, AOR is social justice and resistance in the process and outcome. Second, it recognizes that all knowledge is socially constructed and political. Third, the process in AOR is all about challenging oppressive power relationships.

AOR is influenced by critical theories as it makes the connection between knowing and doing (Moosa-Mitha, 2005). However, AOR goes beyond critical theory in recognizing the importance of the subjective and specific including sociohistorical experiences, in emphasizing fluidity and contesting binary constructions of reality (Moosa-Mitha, 2005). AOR has also been influenced by emancipatory approaches to research as it promotes the production of knowledge that supports freedom from oppression and is transformative in the material and social relations of research production (Strier, 2007: p. 860).

The goal in AOR is not to find 'truth' but to find meaning and an understanding of power in order to change situations (Potts and Brown, 2005). AOR acknowledges its value position by taking a stance committed to social justice (Potts and Brown, 2005; Strega, 2005). It views knowledge as power, and since there is a relationship between research and knowledge creation, researchers are cautioned about whether they are supporting or challenging existing power structures (Strega, 2005).

AOR, therefore, aims to redistribute power and ownership over the production of knowledge (Strier, 2007). It transforms the way social researchers frame the focus, scope and, outcome

of their work (Strier, 2007). Research conducted from a neutral or objective location is considered unethical (Herising, 2005) as AOR acknowledges that the worldview of the researcher will shape the researcher's epistemological foundation affecting every step of the research project (Strega, 2005). However, this approach believes this can be avoided if the researchers locate themselves (Herising, 2005).

AOR is 'not methodologically distinctive, but epistemologically distinctive' since it questions the relationships of the knower, the known, and those who want to know (Potts and Brown, 2005: p. 283). Since antioppressive researchers are aware of the dynamics involved in the social construction of knowledge, AOR is not a process to discover knowledge, but a political process to 'cocreate and rediscover knowledge' (Potts and Brown, 2005: pp. 261–262). The question of whose voice is being represented in the research becomes a main focus in critical research (Potts and Brown, 2005; Strega, 2005). These researchers consent with the perspectives of indigenous scholars, critical race theorists, and feminists who have raised important epistemological questions about who is allowed to create meanings about the world, whose voice is validated, and how social identities factor into the decision (Strega, 2005). Strega (2005) argues that social justice researchers consider the answers to the questions: "How can I best capture the complexities and contradictions of the worlds, experiences, or texts I am studying? Whose voice will/does my research represent? Whose interests will it serve? How can I tell if my research is good research? Are choices about resistance and allegiance to the hegemony of Eurocentric thought and research traditions?" (p. 199).

In AOR there is 'no research without relationships', meaning that researchers cannot approach a community with the intention of collecting data and leaving (Potts and Brown, 2005: p. 263). The relationship between researchers and the community is one that could possibly be for life as there is no neat and tidy beginning and end (Potts and Brown, 2005). Through strong relationships, the researcher engages people in all stages of research who then formulate research questions and design the study. The ongoing construction of a study design is done from an epistemological understanding of values and recognizes the importance of subjugated knowledges (Potts and Brown, 2005). Viewed from this perspective, research processes become acts of resistance (Potts and Brown, 2005).

The first site of change usually in AOR is the researchers themselves (Potts and Brown, 2005). Becoming an anti-oppressive researcher means challenging one's own perception of self, history, contexts, and actions (Potts and Brown, 2005). Scholars have written about the importance of transparency and accountability of the researcher to the community (Herising, 2005; Potts and Brown, 2005). It is important that ideological and political biases are made visible and that the researcher continually reflects on when and how their beliefs, values, identities, and knowledges impact the work with various communities (Herising, 2005). This would give indication when those beliefs, values, identities, and knowledges need to be disinherited, disavowed, decentered, disrupted, claimed, reinserted, or centered in the work (Herising, 2005: p. 131). However, such quests for transparency may excuse researchers from attending to the complex interrelationships and colonial context of research (Herising, 2005). Therefore, Herising (2005) argues it is important for researchers to move beyond individualized reflection and reflexivity and take a critical stance by considering the multiple histories that shape dynamics of power.

Strega (2005) provides criteria for assessing AOR. Instead of assessing research from positivist standards of validity and reliability, Streg suggests that criteria such as social justice validity, disseminating back the findings in a manner accessible in format and language to the community, and reflexivity in assessing the researcher's own complicity in systems of domination and subordination should be used for assessing AOR. These ways of evaluating research have been criticized by both positivist and postpositivist researchers. Positivist and post-positivist researchers perceive such criteria for evaluating research as 'attacks on reason and truth' (Seale et al., 2004; Denzin and Lincoln, 2012: pp. 8–9) and hold the opinion that their work is free of bias and subjectivity and hence more scientific. See Potts and Brown (2005) for an example of AOR social work studies.

CSE, AR, and Social Work

The purpose of CSE including AR is congruent with the mission of social work practice in that they both are concerned with the emancipation of marginalized people and creating transformative change (Alvesson and Sköldberg, 2009; Carroll, 2004; Strega, 2005). These are not unlike the concerns of critical social work, which include the emancipation of marginalized people, the elimination of structures of oppression, social justice, and transformative change (refer to Fook, 2002; Potts and Brown, 2005). As such, social work practitioners have been encouraged to embrace research as a core aspect of their practice (Brown and Strega, 2005). Social workers can utilize these forms of enquiry as useful tools for practice by for engaging with communities in understanding and exploring issues of relevance to them, building capacities of communities, and galvanizing communities to take action on issues of oppression and injustice experienced by them. Using CSE social workers can bring forward the subjugated knowledges and otherwise silenced voices of marginalized groups on issues that impact their lives. Participatory, empowerment, and transformative components of CSE can also be effectively used to conduct community appraisals, impact assessments of policies, and program evaluations (for more information on transformative evaluation refer to Fetterman, 2002). Overall, CSE provides opportunities for social workers to participate in transformative change; impacting the lives of community members individually and collectively and ultimately transforming society (see Brown and Strega, 2005: p. 10).

Conclusion

The foregoing discussion traces the evolution of CSE from AR, PAR, and CBAR to AOR. While the philosophical foundation of these frameworks originated from critical social science during the course of their development they have been also influenced by feminist, antiracist, anticolonial, poststructural, and

postmodern perspectives. The common thread that binds these frameworks is the commitment to issues of participation, social justice, antioppression, and transformative change. This commitment of CSE is congruent with the progressive mission of social work and could be effectively used by social workers engaged in effecting change.

See also: Community, Social Context of: The US Case; Critical Reflexivity and Social Work Practice; Critical Social Work Practice; Participatory Action Research in Environmental and Ecological Studies; Participatory Action Research in Social Research; Social Action and Social Justice; Social Justice and Social Inequalities.

Bibliography

Adelman, C., 1993. Kurt Lewin and the origins of action research. Educational Action Research 1 (1), 7–14.

Alvarez, A.R., Gutierrez, L.M., 2001. Choosing to do participatory research: an example and issues of fact to consider. Journal of Community Practice 9 (1), 1–20.

Alvesson, M., Sköldberg, K., 2009. Reflexive Methodology: New Vistas for Qualitative Research, second ed. Sage Publications, London.

Berg, B.L., 2007. Qualitative Research Methods for the Social Sciences, sixth ed. Pearson Education, Inc., San Francisco.

Briar-Lawson, K., Munford, R., Sanders, J., 2010. Challenges and directions in the practice of social work research. In: Shaw, I., Briar-Lawson, K., Orne, J., Ruckkdeschel, R. (Eds.), The SAGE Handbook of Social Work Research. Sage, Thousand Oaks, pp. 463–476.

Brown, L., Strega, S., 2005. Transgressive possibilities. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press/Women's Press, Toronto, pp. 127–152.

Brydon-Miller, M., Kral, M., Maguire, P., Noffke, S., Sabhlok, A., 2011. Jazz and the banyan tree: roots and riffs on participatory action research. In: Denzin, N.K., Lincoln, Y.S. (Eds.), The SAGE Handbook of Qualitative Research, fourth ed. Sage, Thousand Oaks, pp. 27–42.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. http://www.pre.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf (accessed December 2010.).

Cannella, G.S., Lincoln, Y.S., 2009. Deploying qualitative methods for critical social purposes. In: Denzin, N.K., Giardina, M.D. (Eds.), Qualitative Inquiry and Social Justice: Toward a Politics of Hope. Left Coast Press, Walnut Creek, pp. 53–72.

Carroll, W.K., 2004. Unpacking and contextualizing critical research strategies. In: Carroll, W.K. (Ed.), Critical Strategies for Social Research. Canadian Scholars' Press, Inc., Toronto, pp. 303–319.

Denzin, N.K., Giardina, M.D., 2009. Qualitative inquiry and social justice: toward a politics of hope. In: Denzin, N.K., Giardina, M.D. (Eds.), Qualitative Inquiry and Social Justice: Toward a Politics of Hope. Left Coast Press, Walnut Creek, pp. 11–52.

Denzin, N.K., Lincoln, Y.S., 2012. Introduction: the discipline and practice of qualitative research. In: Denzin, N.K., Lincoln, Y.S. (Eds.), The SAGE Handbook of Qualitative Research. Sage, Thousand Oaks, pp. 1–21.

Ferguson, H., 2003. Outline of a critical best practice perspective on social work and social care. British Journal of Social Work 33 (8), 1005–1024.

Fetterman, D., 2002. Empowerment evaluation: building communities of practice and a culture of learning. American Journal of Community Psychology 30 (1), 89–102.

Flicker, S., Travers, R., Guta, A., McDonald, S., Meagher, A., 2007. Ethical dilemmas in community-based participatory research: recommendations for institutional review boards. Journal of Urban Health 84 (4), 478–493.

Fook, J., 2002. Social Work: Critical Theory and Practice. SAGE Publications, London.

Freire, P., 1970. Pedagogy of the Oppressed. Herder and Herder, New York.

Gayá Wicks, P., Reason, P., Bradbury, H., 2008. Living inquiry: personal, political and philosophical groundings for action research practice. In: Reason, P., Bradbury, H. (Eds.), SAGE Handbook of Action Research: Participative Inquiry and Practice, second ed. Sage Publications, London, pp. 15–30.

Gellis, Z.D., 2001. Using a participatory research approach to mobilizing immigrant minority family caregivers. Journal of Social Work Research 2 (2), 268–292.

Gutiérrez, L., 2003. Participatory and stakeholder research. In: Freeman, E. (Ed.), Encyclopedia of Social Work, Supplement. NASW Press, Washington, DC.

Herising, F., 2005. Interrupting positions: critical thresholds and queer pro/positions. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press/Women's Press, Toronto, pp. 127–152.

Ku, H.B., 2011. 'Happiness being like a blooming flower': an action research of rural social work in an ethnic minority community of Yunnan Province, PRC. Action Research 9 (4), 344–369.

Lavoie, C., MacDonald, J., Whitmore, E., 2010. Methods for understanding, learning and social justice. In: Shaw, I., Briar-Lawson, K., Orne, J., Ruckkdeschel, R. (Eds.), The Sage Handbook of Social Work Research. Sage, Thousand Oaks, pp. 298–314.

Levin, M., Greenwood, D., 2011. Revitalizing universities by reinventing the social sciences: building an action research. In: Denzin, N.K., Lincoln, Y.S. (Eds.), The SAGE Handbook of Qualitative Research, fourth ed. Sage, Thousand Oaks.

Minkler, M., 2004. Ethical challenges for "outside" researcher in community-based participatory research. Health Education and Behavior 31 (6), 684–697.

Moosa-Mitha, M., 2005. Situating anti-oppressive theories within critical and difference-centred perspectives. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press/Women's Press, Toronto, pp. 37–72.

Park, P., 1993. What is participatory research? A theoretical and methodological perspective. In: Park, P., Brydon-Miller, M., Hall, B.L., Jackson, T. (Eds.), Voices of Change: Participatory Research in the United States and Canada. Bergin & Garvey, Westport, pp. 65–84.

Potts, K., Brown, L., 2005. Becoming an anti-oppressive researcher. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press, Toronto, pp. 255–286.

Reason, P., Bradbury, H., 2008. Introduction to groundings. In: Reason, P., Bradbury, H. (Eds.), SAGE Handbook of Action Research: Participative Inquiry and Practice, second ed. Sage Publications, London, pp. 1–10.

Reitsma-Street, M., Brown, L., 2004. Community action research. In: Carroll, W.K. (Ed.), Critical Strategies for Social Research. Canadian Scholars' Press, Inc., Toronto, pp. 303–319.

Reitsma-Street, M., Neysmith, S., 2000. Restructuring and community work: the case of community resource centres for families in poor urban neighbourhoods. In: Neysmith, S. (Ed.), Restructuring Caring Labour: Power, Discourse and Everyday Life. Oxford University Press, Toronto, pp. 142–163.

Rutman, D., Hubberstey, C., Barlow, A., Brown, L., 2005. Supporting young people's transitions from care: reflections on doing participatory action research with youth from care. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press/Women's Press, Toronto, pp. 153–180.

Seale, C., Gobo, G., Gubrium, J.F., Silverman, D., 2004. Introduction: inside qualitative research. In: Seale, C., Gobo, G., Gubrium, J.F., Silverman, D. (Eds.), Qualitative Research Practice. Sage, London, pp. 1–11.

Smith, L.T., 1999. Decolonizing Methodologies Research and Indigenous Peoples. Zed Books, London.

Strega, S., 2005. The view from the poststructural margins: epistemology and methodology reconsidered. In: Brown, L., Strega, S. (Eds.), Research as Resistance: Critical, Indigenous & Anti-Oppressive Approaches. Canadian Scholars' Press/Women's Press, Toronto, pp. 199–236.

Strier, R., 2007. Anti-oppressive research in social work: a preliminary definition. British Journal of Social Work 37 (1), 857–871.

Stringer, E.T., 2007. Action Research, third ed. Sage, Thousand Oaks.

Tandon, R., 2008. Participation, citizenship and democracy: reflections on 25 years' of PRIA. Community Development Journal 43 (3), 284–296.

Wang, C., 2006. Youth participation in photovoice as a strategy for community change. Journal of Community Practice 14 (1), 147–161.

Social Epidemiology

Louis Chauvel, Research Unit INSIDE, IRSEI Institute, University of Luxembourg, Walferdange, Luxembourg
Anja K Leist, Research Unit INSIDE, IRSEI Institute, University of Luxembourg, Walferdange, Luxembourg

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S.L. Syme, volume 7, pp. 4701–4706, © 2001, Elsevier Ltd.

Abstract

The search of social determinants of health and disease has advanced substantially over the last decade. We present recent theoretical advancements, methodological approaches, and a selection of empirical evidence for the three main research strands: first, social inequalities can explain health differences. Here, we first focus on hierarchical social stratification with regard to socioeconomic and social class differences influencing health, then we extend the view toward nonhierarchical social stratification with regard to ethnocultural differences, lifestyle, and cognitive and noncognitive abilities. We shortly mention social relations and social network as determinants of health. Last, we use the concept of social times to distinguish age, period, and cohort effects in population health. After presenting evidence on contextual social determinants of health, we close with methodological challenges, social policy implications, and translation to practice.

The Field of Social Epidemiology

Health and disease are not only influenced by biological, genetic factors, and health behaviors, but also to an important extent by social and societal characteristics (Amick, 1995). The discipline systematically investigating the social determinants of health and disease is called social epidemiology. The three main fields assigned to social epidemiology are related to social inequalities, social relations, and social time. In this article, the focus regarding the study of *social inequalities* will be to include both the role of hierarchy in health ('vertical inequality,' Section Social Inequalities in Health and Disease), and to extend beyond the traditional concepts of socioeconomic position and social class and include nonhierarchical differentiations such as ethnocultural differences or lifestyle ('horizontal inequality,' Section Nonhierarchic Social Determinants of Health). After that, *social relations* determining health are presented, which can be related to the French school of social sciences developed by Emile Durkheim (1897), where the study of 'social facts' is rooted in a social theory of integration and of regulation (Berkman et al., 2000; Section Social Relations Pertaining to Health Differences). The study of social inequalities determined by age, another nonhierarchical attribute leading to important social inequalities across cohorts and historical periods and particularly visible in age-period-cohort (APC) analyses, will be expanded in a separate section on the role of *social times* in determining health and disease (Ryder, 1965; Section Social Times in Explaining Health Differences). After reviewing a selection of evidence on contextual social determinants of health (Section Micro- and Macrocontextual Views), we close with perspectives on methodology, social policy implications, and translation to practice (Section Perspectives on Methodology, Social Policy, and Translation to Practice).

A core hypothesis based on extensive sets of observations is that, in virtually all noncommunicable diseases, a socioeconomic gradient can be observed in that lower socioeconomic groups have higher incidence or prevalence rates of disease, develop disability with a greater chance, and die at younger ages than higher socioeconomic groups. As people are born into a certain socioeconomic status or are in lower socioeconomic groups for reasons they cannot (entirely) control, these socioeconomic gradients of health and disease are often referred to as *social* or *socioeconomic inequalities*. In a time of increasing inequalities in many countries (Piketty, 2014), these issues are vital for the future of human health. Today it is almost common knowledge and a central result of this discipline that social inequality leads to bad health (Wilkinson and Pickett, 2010), although it can be argued about the appropriateness of the underlying social interpretations (Goldthorpe, 2010) in both academic and para-academic circles.

Considering the wealth of research in social epidemiology over the last decades, this article can only give a broad overview of some of the main lines of research. In the following, concepts and research on how health is influenced by social inequalities will be presented.

Social Inequalities in Health and Disease

When referring to social inequality and impact on health, two main conceptualizations might be mentioned. The first one pertains to a neutral (smooth) vision of inequality where a *vertical* scale of resources (ultimately: income) generates different health outcomes, conceptualized as *socioeconomic position* (Section Socioeconomic Position). The second one, *social class*, relates to more radical class divides (Section Social Class). In this section, we focus on research on social inequalities as hierarchical differentiation, and particularly consider recent theoretical and methodological advancements. Whereas this 'vertical,' hierarchical aspect of social inequality is the most visible result of social epidemiology, the 'horizontal' or nonhierarchical differentiation must be underlined as a source of investigation, and will be discussed in Section Nonhierarchic Social Determinants of Health.

Socioeconomic Position

For Krieger et al. (1997), *socioeconomic position* is defined as an "aggregate concept that includes both resource-based and prestige-based measures, as linked to both childhood and adult social class position" (p. 345). In this tradition, inequality is mainly a hierarchical series of layers pertaining to unequal size of resources able to affect health. Typical measures of socioeconomic position, based on different degrees of resource and prestige, respectively, are education, income, and occupation. The interchangeable use of these indicators in social inequalities research has often been cautioned, as Krieger (2001a) notes that "(g)iven distinctions between resource-based and prestige-based aspects of socioeconomic position and the diverse pathways by which they affect health, epidemiological studies should state clearly how measures of socioeconomic position are conceptualized" (p. 697). Geyer et al. (2006) show in mortality data of Sweden and Germany that income, occupation, and education reflect different social phenomena and their associations with health outcomes relate to different underlying causal mechanisms.

Social Class

The often acclaimed book of Wilkinson and Pickett (2010) is another typical example of a hierarchical conception of social stratification that had been criticized by Goldthorpe (2010) as a stronger definition of social inequality in terms of structured social classes is needed. This tradition develops a social theory of not only based on individual access to resources in a continuous social scale but to a more holistic (if not Marxist) structuration of this access by a number of classes discontinuously positioned in society, having specific borders and closures able to explain the durability of inequality. In this sense, the term *social class* is defined by Krieger et al. (1997) as a "social category referring to social groups forged by interdependent economic and legal relationships, premised upon people's structural location within the economy" (p. 345). This definition implies that, in fact, social classes extend the view of a health 'gradient' and are (with or without Marxist connotations) not only socioeconomic layers but unequal and mutually hostile status groups of individuals organized on the base of groups sharing social identity (language, lifestyle, elements of recognition), temporal identity (durability of the system of positions, homogamy, social immobility), and even political identity (values and political preferences, unions membership, and political action). In this sense, whereas socioeconomic position is often referred to as social 'ladder' that one can climb up (or down), social class reflects both a hierarchical concept and the creation of 'identity divides' with categorically unequal groups of people. This tradition shows larger explanatory ambitions in terms of system of inequalities of resources and of statuses, resilient to equalizing public policies (Goldthorpe, 2010), but may be more effective when its aims are more modest, as shown by Elo (2009) in her survey of class health inequalities, where the advantage of the second tradition is to better underline the effects of convergence and divergence of economic, educational, cultural, and social resources. Before we continue to the nonhierarchical conceptualization of social stratification, let us first review evidence on (hierarchical) socioeconomic differences in health.

Empirical Evidence

The field of social epidemiology has drastically evolved after several hallmark publications documenting the existence of vast health inequalities, perhaps most prominently presented with data of the Whitehall II study of British civil servants (Marmot et al., 1991). Furthermore, early methodological advances came with different indices of socioeconomic inequality (see Section Perspectives on Methodology, Social Policy, and Translation to Practice for recent methodological advancements). The interested reader is referred to two recent reviews documenting social determinants of health and health inequalities in Europe: Mackenbach et al. (2008) and Marmot et al. (2012). In the beginnings of comparative social epidemiological research, often (general) mortality was used as a health outcome. Among the prominent publications in this field is, for instance, a review by, Subramanian and Kawachi (2004). More recent approaches have benefitted from the availability of national registries on cause-specific mortality, but also on large comparative surveys on self-rated health and medical conditions.

In the last decade, the focus of social inequalities research has shifted to nonmortality health and disease outcomes. Concerning the countless studies applying outcomes such as cardiovascular disease or stroke, we have aimed at providing the most relevant studies and further readings in the respective sections. Large population-representative surveys make it possible to use the indicator of self-rated health, e.g., in research on state-level influences on health (Subramanian et al., 2001) or use self-reports to determine socioeconomic differences in the complex old-age syndrome of frailty.

Some recent research has used biomarkers in order to determine to which extent socioeconomic differences translate into bodily changes (see the concept of embodiment; Krieger, 2001b). A few (very selective) examples of outcomes in this field are socioeconomic differences in obesity, hypertension, or grip strength.

Last, one distinct line of research also makes use of psychosocial outcomes of health, such as life satisfaction or well-being, thereby identifying a socioeconomic gradient in, for instance, depression (Lorant et al., 2003).

Nonhierarchical Social Determinants of Health

Extending the view on social inequalities from a hierarchical perspective, population health is not just about income inequalities and social class divides, but also influenced by individual attributes, both within and across social layers. Thus, a more 'horizontal differentiation' relating to nonhierarchical attributes of the members of a society shows that other factors such as ethnicity, lifestyle, and cognitive and noncognitive abilities produce status and health differences.

Research on nonhierarchical dimensions of social stratification and its capacity in explaining health differences has gained importance during the last decade, and nonhierarchical individual factors contributing to social inequalities have been

ascribed more and more importance. In line with the notion to distinguish between proximal and distal causes for health inequalities, one can differentiate nonhierarchical individual attributes that contribute to inequalities in status and health. One line of argumentation relates to the translation of hierarchical stratification to changes in individual attributes, e.g., the level of *stress* which in turn acts as a cause for health inequalities. A prominent advocate of this line of reasoning is Nancy Krieger (2001a). To illustrate this notion, she has coined the concept of *embodiment*, "a concept referring to how we literally incorporate, biologically, the material and social world in which we live, from conception to death" (Krieger, 2001b: p. 672). Another pathway how socioeconomic status translates to individual attributes is related to the notion of control and related concepts (self-efficacy, helplessness, hardiness), and it can be postulated that feeling in control is related to less stress in the sense of Krieger or that control involves material and immaterial resources such as economic power, information, or prestige. A very recent field delving even deeper into nonhierarchical dimensions of social stratification targets possible interactions of genetic and social factors and is under increasing scrutiny in the field of fertility and reproductive health (Balbo et al., 2013), a cross-borders domain where new perspectives for social epidemiology can be found.

Another line of argumentation is evolving from the search of mediators in the socioeconomic position–health relationship. Past research has shown that socioeconomic position is not able to fully explain the health gradient: It is easy to imagine why those at the bottom of the socioeconomic ladder may fare the worst (the classic line of argumentation being that they may have limited access to health care, not enough resources to afford a healthy diet, or face occupational hazards). Yet health truly follows a gradient in that, compared to those at the top of the socioeconomic ladder, being just beneath the top is in many cases already associated with worse health (Marmot et al., 1991). However, the picture gets more complicated if we extend our view beyond restricted work environments (e.g., the London civil servants' population) and compare hierarchies of members of different cultural contexts (e.g., teachers in schools vs company employees). These complexities aside, the health gradient even within a work environment cannot be fully explained by the mere differential availability of socioeconomic resources. Acknowledging this fact, incorporating nonhierarchical individual attributes into frameworks and studies of social determinants of health has been shown to explain larger parts of the socioeconomic health gradient. While individual nonhierarchical attributes are in many cases related to health outcomes regardless of indicators of social class, it is important to bear in mind that many of these attributes are closely associated, and in many cases interact with indicators of social class. In the following, a selection of relevant research on nonhierarchical social determinants of health is depicted.

Ethnocultural Differences

The most prominent differentiations among nonhierarchical concepts of social stratification are certainly ethnicity and race. These factors may on first glance be individual-level factors, however, an overwhelming body of research concludes that it is often the sociocultural context which defines health advantages and disadvantages of native- and foreign-born people. Native- or foreign-born status has multifaceted health implications such as differential access to health care, differences in common health behaviors of the country of residence, and most prominently strong associations with different socioeconomic position. Ethnic and racial differences in health, thus, cannot be investigated without considering sociocultural context factors contributing to these health differences. As an exhaustive presentation of relevant research is beyond the scope of this article, the interested reader is referred to a comprehensive review on the topic in Anderson et al. (2004). Whereas the main strand of research was driven by the ethnically diverse population of the U.S., recent research also considers ethnic differences in health in European countries with a large migrant population. Applying the life course framework to explain ethnic differences in older-age outcomes, important theoretical and empirical contributions have been made investigating outcomes such as stroke (Glymour et al., 2008).

Lifestyle

These individual attributes relate to (un)healthy *lifestyle*, as important health behaviors (e.g., smoking, physical activity, fruit, and vegetable consumption) in many cases follow a socioeconomic gradient as well (Lantz et al., 2001). As one prominent example for the influence of lifestyle in interaction with socioeconomic position, Krueger and Chang (2008) show that high stress levels and unhealthy behaviors (former smoking, physical inactivity) can independently explain mortality. Further, the combination of these risk factors creates a group of individuals at high risk for premature death, thereby acting as a powerful 'toxin' that can only be identified – and targeted – by including both hierarchical and nonhierarchical dimensions of social stratification in explaining health.

Cognitive and Noncognitive Abilities

The rise of searching for individual-level factors contributing to social inequalities in health has certainly been expedited by theoretical advancements in *life course epidemiology*, tracing back adult health to childhood and even pre- and perinatal influences (Ben-Shlomo and Kuh, 2002). In line with this framework, empirical evidence has been gathered that socioeconomic position may vary across the life course and may have different implications for later health outcomes. In this tradition and with the availability of large data sets spanning large parts or even the whole lifespan of respondents, compelling evidence has been gathered that *intelligence*, assessed in early childhood, contributes to a large extent to adult and old-age health. This led to the debate if intelligence is even the 'fundamental cause' of health inequalities (Gottfredson and Deary, 2004) and to defining the emerging subdiscipline of *cognitive epidemiology* (Deary and Batty, 2007). In this tradition, economists are also increasingly applying the social epidemiological perspective in explaining adult socioeconomic differences and health inequalities with the impact of childhood cognitive and noncognitive endowments (Conti et al., 2010).

Social Relations Pertaining to Health Differences

Considering the wealth of research on the associations between social network and health, we will only give a short overview of the roots of this research and its adoption in social epidemiological research.

A framework on how contextual factors shape health has been developed by Krieger (2001b), the so-called ecosocial theory. Today it is widely recognized that social networks shape health and well-being of their members. In fact, this line of reasoning goes back to Emile Durkheim. In his seminal research on suicidal behavior, Durkheim proposed a double-traits explanation based on regulation (capacity to assign oneself attainable goals) and integration (strength of social ties). Socialist but anti-Marxist, the founding father of the French school of social sciences introduced important explanatory factors. At his time, *economic crisis* – because of its effects of reducing the capacity to achieve goals and of generating frustrations and thus anomic behavior, including suicide – and the emergence of a modern society of isolated individuals lacking social commitments and integrative networks have been detected as important explanations of the expansion of suicide in the late nineteenth-century France.

Indeed, social networks cannot be reduced to the functions they provide, e.g., social support, but they have to be viewed in the larger social and cultural context: Berkman et al. (2000) make a plea for a more Durkheimian view on social networks and health, as "it is critical to maintain a view of social networks as lodged within those larger social and cultural contexts which shape the structure of networks. (…) Only then can we fully consider the multiple pathways by which social networks might so profoundly influence health outcomes" (p. 846).

Social Times in Explaining Health Differences

An important methodological challenge from today's point of view is the integration of *time* in general. The discipline faces complex systems of social causality: Social determinants of health evolve over long time lags that are sometimes spanning several decades between the initial causes, often situated in childhood, and their ultimate consequences in late adulthood. Furthermore, social determinants of health follow complex personal trajectories that are influenced by multidimensional social and other risk factors, possible nonlinearities, and reversibilities. On top of these sources of complexity, 'time' is plural: the concept of 'age' transports information about individual 'somatic' *age* that shapes health outcomes, but is also the reflection of being a member of a *birth cohort*, with all life course experiences from prenatal to childhood throughout adulthood (Ben-Shlomo and Kuh, 2002). In addition, it contains information on *historical period*, thus highly diversifying the definitions of time. Moreover, 'age' and 'period' can compete with regard to the stage of country development. It is therefore desirable – yet difficult – to disentangle the effects of age, period, and cohort on health and disease. As a result, an increasing demand for longitudinal data sets, panel studies, life-event analysis methodology, and sophisticated structural models generates large avenues for methodological developments (Oakes and Kaufman, 2006).

One of the important fields in this domain is the renewal of APC models that intend to ways that disentangle effects related to somatic age, historical period, and birth cohort. From the first 1930s discoveries to Ryder's (1965) conceptualization, and traditional modeling proposals, the field of APC has drastically evolved. After 2005, a large renewal in the domain has been allowed by new available multilevel and general linear models (Yang and Land, 2013).

Researchers supporting this trend recently proposed new investigations of health outcomes such as happiness, obesity, smoking habits, cognitive decay, general as well as cause-specific mortality, to name just a few. APC research aims for instance at detecting birth cohorts at risk that entered adulthood in a period of unsafe health environment, for instance an epidemic outbreak, facing stronger exposure to potentially lethal infections, whereas earlier and later (luckier) cohorts faced a better context of health socialization. This kind of research allows both assessments of public health policies and forecasting on health conditions of aging populations. Extending the view of social inequalities to a comparative perspective across historical time, it is of great interest to know if birth cohorts differ in given advantage or disadvantage compared to earlier or later birth cohorts, that is, if there are social inequalities between generations (Chauvel and Schröder, 2014). The resulting question is the extent to which these inequalities are due to individual-level factors such as educational attainment or, conversely, contextual-level factors such as historical economic situation or welfare regime, and to determine if these disadvantages are, over the life course, permanent or transitory (Chauvel et al., in preparation).

Micro- and Macrocontextual Views

The field of social epidemiology has benefitted from a long tradition of studies investigating contextual influences on health and disease. In the following, a selection of recent evidence on the association of neighborhood factors and health will be presented. After that, comparative evidence on the influence of country, i.e., population-level influences such as types of welfare regimes, will be outlined.

Neighborhood and Area

Extending the social network view on individual health differences briefly outlined before (Section Social Relations Pertaining to Health Differences), a growing body of research investigates health effects of environmental surroundings, area, and neighborhood, applying usually multilevel modeling. Diez Roux and Mair (2010) offer a comprehensive review on how neighborhood can contribute to social and ethnic inequalities in health. They provide a framework that explains how residential segregation by race/ethnicity or socioeconomic position is linked to neighborhood physical environment (e.g., quality of housing) as well as social environment (e.g., safety/violence). This in turn, in interaction with individual behavioral mediators and stress, influences health outcomes (Diez Roux and Mair, 2010). With regard to other health outcomes,

it may be worth to mention a recent study investigating causes for neighborhood inequalities in later-life cognitive function by applying a well-designed multilevel mediation analysis (Clarke et al., 2012).

Neighborhood studies face structural and methodological challenges in the search for true *causes* of health differences related to neighborhood. Oakes (2004) cautiously notes pitfalls of common multilevel models in estimating neighborhood effects and suggests both a counterfactual causal framework and a series of advances to overcome these pitfalls, proposing randomized community trials as gold standard.

Country, Nations, Institutions, and Welfare Regimes

Comparative analyses of socioeconomic gradients in mortality rates show the influence of the welfare regime on the health and health inequalities of its inhabitants (Eikemo et al., 2008). Country of residence does not only reflect different historical steps of socioeconomic development that still influence the present situation (*path dependency*, Pierson, 2000), diverging cultural lifestyles and food habits, but indeed countries are shaped by categorically different ways of organization of institutionalized solidarity and welfare regimes (Esping-Andersen, 1999). Much evidence points to the fact that institutions matter for health and successful societies (Hall and Lamont, 2009). Indeed, the study of comparative decommodification of health systems, insurances, pensions, education, etc. had been able to explain how similar countries in terms of economic development have had divergent outcomes in terms of quality of life and of health performances. Health and life expectancy depend to an important extent on national social policies, such as access to and quality of health care or tobacco control, or simply the level of living standards of older people via public and private pensions systems.

Two recent developments target, first, the use of comparative approaches to directly estimate the differential impact of country and national indicators on health outcomes, and, second, the use of the life course framework to investigate life course influences on later health outcomes.

With regard to the first line of research, the use of large national registries makes it possible to not only consider country of current residence, but also country of origin. Extending research on the so-called U.S. Stroke Belt (a number of U.S. states with high stroke rates), Glymour et al. (2007) show that residence in one of these states in childhood predicts stroke even after moving to a 'healthier' state later in life. The importance of these life course effects matters especially for stroke and other education-dependent outcomes, such as old-age cognitive function or dementia mortality, such as state-level quantity and quality of education.

Another recent comparative study used a summary health policy performance indicator for 43 countries that geographically belong to Europe to determine the extent to which survival is dependent upon national health policies. The authors conclude that health policies importantly contribute to survival by showing that, overall, Sweden, Norway, and Iceland perform best, whereas Ukraine, Russian Federation, and Armenia perform worst (Mackenbach and McKee, 2013).

With regard to the second line of research, life course theory offers plenty of possible hypotheses to test the influence of state-level life course factors on later health outcomes. It is for instance possible that transitory economic conditions such as economic booms or recessions may have long-run effects on later health outcomes. Among the possible mechanisms for this association are diverse factors such as economically influenced changes in nutrition, stress, socioeconomic status, material resources, working conditions, or other factors. Several studies have found effects of economic recessions at different points throughout the life course on later health outcomes, such as economic situation at birth on cardiovascular mortality (Van den Berg et al., 2011) and during working life on older-age cognitive function (Leist et al., 2014).

Closely related to these studies investigating the contextual influences on health differences is the research on successes and failures of social policies in a comparative and longitudinal perspective. In the following, a selection of evidence on the links between social policies and health will be discussed. After that, this article will be closed with pointing to challenges in the translation of evidence into policy changes and practice.

Perspectives on Methodology, Social Policy, and Translation to Practice

Despite all research on social determinants of health, implementing adequate policies to reduce health inequalities is far more difficult than expected. Especially in the U.S., life expectancy drops and morbidity increases despite all knowledge (Berkman, 2009). Not only absolute, but also relative social inequalities persist despite modern welfare regimes, and current theories can only partly explain the pattern of findings (Mackenbach, 2012). Avendano and Kawachi (2014) show that determining the health of a society is a complex question, as multifaceted explanations are needed for the health disadvantages of U.S. Americans compared to people from other high-income countries. Indeed, it is imperative for social epidemiological research today to focus on *social policy assessment* in a context of global social change. Inequality of outcomes in terms of health is certainly one of the strongest challenges of globalizing societies.

Before policy implications can be derived, the search for robust causes of health differences faces a multitude of serious methodological challenges. The introduction and advancement of causal diagrams, i.e., *directed acyclic graphs*, to disentangle causation and correlation in the search for social determinants of health have been valuable to advance our understanding of the pathways through which social factors operate (Glymour, 2006). Social epidemiological research should in the future consider even more than today the role of heterogeneity of effects, i.e., effect modifiers, in determining "what works for whom and when" (Glymour et al., 2013).

These methodological challenges in finding social determinants of health have to be extended in order to determine causation on the macro- and microlevel. In the large domain of causality analysis, further advancements have to be

developed to robustly examine macrolevel causation (Babones, 2008), whereas the panel-based difference-in-differences propensity score matching approach offers a relevant tool to detect microlevel causation (Gebel and Voßemer, 2014). Another relevant example of methodological advancements that social epidemiology should focus on more in the future is to gain evidence if and to which extent social inequalities are caused by effects of age, period, or cohort. The investigation of social times with dynamic analyses and international comparisons are vital tools to detect social policy successes and failures. Then, social epidemiological research has the capacity to robustly detect which societies perform better or worse, how and why, and which social policy changes will work and for which segments of the population in order to improve population health.

See also: Culture as a Determinant of Mental Health; Environmental Stress and Health; Health Outcomes, Assessment of; Income Distribution and Inequality: Measurement Issues; Life Course Epidemiology; Migration and Health; Occupational Health; Personality and Health; Public Health as Social Science; Race: Ethnicity and Health; Smoking and Health; Social Inequality and Schooling; Social Integration, Social Networks, and Health.

Bibliography

Amick, B.C., 1995. Society and Health. Oxford University Press.

Anderson, N.B., Bulatao, R.A., Cohen, B.E., Panel on Race E, and Health in Later Life, National Research Council, 2004. Critical Perspectives on Racial and Ethnic Differences in Health in Late Life. The National Academies Press, Washington, DC.

Avendano, M., Kawachi, I., 2014. Why do Americans have shorter life expectancy and worse health than do people in other high-income countries? Annual Review of Public Health. http://dx.doi.org/10.1146/annurev-publhealth-032013-182411.

Babones, S.J., 2008. Income inequality and population health: correlation and causality. Social Science & Medicine 66 (7), 1614–1626.

Balbo, N., Billari, F.C., Mills, M., 2013. Fertility in advanced societies: a review of research. European Journal of Population/Revue Européenne de Démographie 29 (1), 1–38.

Ben-Shlomo, Y., Kuh, D., 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. International Journal of Epidemiology 31 (2), 285–293.

Berkman, L.F., 2009. Social epidemiology: social determinants of health in the United States: are we losing ground? Annual Review of Public Health 30, 27–41.

Berkman, L.F., Glass, T., Brissette, I., Seeman, T.E., 2000. From social integration to health: Durkheim in the new millennium. Social Science & Medicine 51 (6), 843–857.

Chauvel, L., Leist, A.K., Ponomarenko, V., 1970–2009. Stability or resilience of cohort effects over the life course: assessing the stability of cohort-specific suicide rates in 25 countries (under review).

Chauvel, L., Schröder, M., 2014. Generational inequalities and welfare regimes. Social Forces. sot156. http://dx.doi.org/10.1093/sf/sot156.

Clarke, P.J., Ailshire, J.A., House, J.S., Morenoff, J.D., King, K., Melendez, R., Langa, K.M., 2012. Cognitive function in the community setting: the neighbourhood as a source of 'cognitive reserve'? Journal of Epidemiology and Community Health 66 (8), 730–736.

Conti, G., Heckman, J., Urzua, S., 2010. The education-health gradient. The American Economic Review 100 (2), 234–238.

Deary, I.J., Batty, G.D., 2007. Cognitive epidemiology. Journal of Epidemiology and Community Health 61 (5), 378–384.

Diez Roux, A.V., Mair, C., 2010. Neighborhoods and health. Annals of the New York Academy of Sciences 1186 (1), 125–145.

Durkheim, É., 1897. Le suicide. Étude de sociologie. Félix Alcan, Paris.

Eikemo, T.A., Huisman, M., Bambra, C., Kunst, A.E., 2008. Health inequalities according to educational level in different welfare regimes: a comparison of 23 European countries. Sociology of Health & Illness 30 (4), 565–582.

Elo, I.T., 2009. Social class differentials in health and mortality: patterns and explanations in comparative perspective. Annual Review of Sociology 35, 553–572.

Esping-Andersen, G., 1999. Social Foundations of Postindustrial Economies. Oxford University Press.

Gebel, M., Voßemer, J., 2014. The impact of employment transitions on health in Germany. A difference-in-differences propensity score matching approach. Social Science & Medicine 108, 128–136.

Geyer, S., Hemström, Ö., Peter, R., Vågerö, D., 2006. Education, income, and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against a common practice. Journal of Epidemiology and Community Health 60 (9), 804–810.

Glymour, M.M., 2006. Using causal diagrams to understand common problems in social epidemiology. Methods in Social Epidemiology, 393–428.

Glymour, M.M., Avendano, M., Berkman, L.F., 2007. Is the 'stroke belt' worn from childhood? Risk of first stroke and state of residence in childhood and adulthood. Stroke 38 (9), 2415–2421.

Glymour, M.M., Avendaño, M., Haas, S., Berkman, L.F., 2008. Lifecourse social conditions and racial disparities in incidence of first stroke. Annals of Epidemiology 18 (12), 904–912.

Glymour, M.M., Osypuk, T.L., Rehkopf, D.H., 2013. Invited commentary: off-roading with social epidemiology – exploration, causation, translation. American Journal of Epidemiology 178 (6), 858–863.

Goldthorpe, J.H., 2010. Analysing social inequality: a critique of two recent contributions from economics and epidemiology. European Sociological Review 26 (6), 731–744. http://dx.doi.org/10.1093/esr/jcp046.

Gottfredson, L.S., Deary, I.J., 2004. Intelligence predicts health and longevity, but why? Current Directions in Psychological Science 13 (1), 1–4.

Hall, P.A., Lamont, M., 2009. Successful Societies: How Institutions and Culture Affect Health. Cambridge University Press.

Krieger, N., 2001a. A glossary for social epidemiology. Journal of Epidemiology and Community Health 55 (10), 693–700.

Krieger, N., 2001b. Theories for social epidemiology in the 21st century: an ecosocial perspective. International Journal of Epidemiology 30 (4), 668–677.

Krieger, N., Williams, D.R., Moss, N.E., 1997. Measuring social class in U.S. public health research: concepts, methodologies, and guidelines. Annual Review of Public Health 18 (1), 341.

Krueger, P.M., Chang, V.W., 2008. Being poor and coping with stress: health behaviors and the risk of death. American Journal of Public Health 98 (5), 889–896.

Lantz, P.M., Lynch, J.W., House, J.S., Lepkowski, J.M., Mero, R.P., Musick, M.A., Williams, D.R., 2001. Socioeconomic disparities in health change in a longitudinal study of US adults: the role of health-risk behaviors. Social Science & Medicine 53 (1), 29–40.

Leist, A.K., Hessel, P., Avendano, M., 2014. Do economic recessions during early and mid-adulthood influence cognitive function in older age? Journal of Epidemiology and Community Health 68, 151–158. http://dx.doi.org/10.1136/jech-2013-202843.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., Ansseau, M., 2003. Socioeconomic inequalities in depression: a meta-analysis. American Journal of Epidemiology 157 (2), 98–112.

Mackenbach, J.P., 2012. The persistence of health inequalities in modern welfare states: the explanation of a paradox. Social Science & Medicine 75 (4), 761–769.

Mackenbach, J.P., McKee, M., 2013. A comparative analysis of health policy performance in 43 European countries. The European Journal of Public Health 23 (2), 195–201.

Mackenbach, J.P., Stirbu, I., Roskam, A.-J.R., Schaap, M.M., Menvielle, G., Leinsalu, M., Kunst, A.E., 2008. Socioeconomic inequalities in health in 22 European countries. The New England Journal of Medicine 358 (23), 2468–2481. http://dx.doi.org/10.1056/NEJMsa0707519.

Marmot, M., Allen, J., Bell, R., Bloomer, E., Goldblatt, P., 2012. WHO European review of social determinants of health and the health divide. The Lancet 380, 1101–1129.

Marmot, M.G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E., Feeney, A., Marmot, M., Smith, G.D., 1991. Health inequalities among British civil servants: the Whitehall II study. The Lancet 337 (8754), 1387–1393.

Oakes, J.M., 2004. The (mis) estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social Science & Medicine 58 (10), 1929–1952.

Oakes, J.M., Kaufman, J.S., 2006. Methods in Social Epidemiology. John Wiley & Sons.
Pierson, P., 2000. Increasing returns, path dependence, and the study of politics. American Political Science Review 94 (2), 251–267.
Piketty, T., 2014. Capital in the Twenty-First Century. Belknap Press, Cambridge, MA.
Ryder, N.B., 1965. The cohort as a concept in the study of social change. American Sociological Review, 843–861.
Subramanian, S.V., Kawachi, I., Kennedy, B.P., 2001. Does the state you live in make a difference? Multilevel analysis of self-rated health in the US. Social Science & Medicine 53 (1), 9–19.
Subramanian, S., Kawachi, I., 2004. Income inequality and health: what have we learned so far? Epidemiologic Reviews 26 (1), 78–91.
Van den Berg, G.J., Doblhammer-Reiter, G., Christensen, K., 2011. Being born under adverse economic conditions leads to a higher cardiovascular mortality rate later in life: evidence based on individuals born at different stages of the business cycle. Demography 48 (2), 507–530.
Wilkinson, R., Pickett, K., 2010. The Spirit Level. Why Equality Is Better for Everyone. Penguin Books, London.
Yang, Y., Land, K.C., 2013. Age-Period-Cohort Analysis. New Models, Methods, and Empirical Applications. Interdisciplinary Statistics Series. CRC Press, Taylor & Francis Group, Boka Raton, FL.

Relevant Websites

http://eurohealthnet.eu/
http://www.ehemu.eu/
http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTHEALTHNUTRITIONANDPOPULATION/EXTPAH/0,,contentMDK:20219043~menuPK:460195~pagePK:148956~piPK:216618~theSitePK:400476,00.html
https://www.ihis.us/ihis/ – Harmonized collection of American health status surveys.
http://www.who.int/social_determinants/en/, Key concepts: http://www.who.int/social_determinants/thecommission/finalreport/key_concepts/en/
http://www.who.int/healthinfo/mortality_data/en/

Social Epigenetics

Douglas L Anderton, University of South Carolina, Columbia, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social epigenetics, the study of social influences on, and outcomes of, changes in gene expression that are not attributable to the basic DNA chain has revolutionized both the role of social science in genetics and the framework for integrating biological influences in social science. Between the genotype and the phenotype lies the epigene, which has now been shown to both be responsive to social influences and result in outcomes of social import. Neurobiological studies have demonstrated the importance of social environments and epigenetic change to developmental disorders. Epidemiological epigenetics has demonstrated the influence of social factors on epigenetic change and the importance of epigenetic changes to health and well-being. A model of inherited genetic influence on social behavior has been replaced with a framework in which genetic influence is altered and mediated by social and environmental factors in producing changes in phenotype of social import. Yet, as great as the promise of social epigenetics is, research confronts both the complexity of the mechanisms involved and what is a new, challenging, research environment for many social scientists.

Social Epigenetics

S*ocial epigenetics* refers to the study of the social influences on, or outcomes of, changes in genetic activity (or gene expression) that are not attributable to the basic DNA chain. For most social and behavioral scientists, social epigenetics requires venturing into unfamiliar territory. The conceptual framework, however, is relatively straightforward. One of the ways in which a gene's expression, or effects, can either change or be suppressed, is through *methylation*, in which a methyl group is added to nucleotides in particular regions of the DNA strand. Thus, methylation may result in a genetic expression, or epigene, that differs from those expected from inherited DNA alone, complicating the association of phenotypes with genetic inheritance. Another source of epigenetic change is *histone modification*, where enzyme-catalyzed changes in the molecular structure of proteins that order the DNA into structural units (nucleosomes) can alter the effective transcription of DNA. Both these sources of epigenetic change affect genetic expression. And, the twist that turns this seemingly biological focus into social epigenetics is simply that both methylation and histone modification, and therefore genetic expression, may be impacted by social behaviors and environments. Modified genetic expression, or the epigene, may then, in turn, affect social consequences or behaviors.

What is particularly interesting then is the interaction between social and genetic influences on gene expression, and the potential impact of this interaction on subsequent genetic expression with social impacts. One often repeated illustration of social and behavioral epigenetics is the effects of stress in methylating regions of the receptor for the stress hormone cortisol, generating gene expression that may subsequently alter the ability to cope with stress. Thus, an essentially social event might result in changes in gene expression that then act similarly to a change in genetics to influence a subsequent social behavior. The last major piece of this framework is the durability of epigenetic changes. Epigenetic changes, just as with DNA, can be transmitted with surprisingly high efficiency (up to 100% for some disease-related changes) from one generation to the next, through meiosis and mitosis during reproduction. So, the effects, for example, of a social exposure or event resulting in methylation can extend across generations along with any health, cognitive, or socially significant effects. At the same time, methylated regions may be chemically demethylated. So, epigenetic effects due to methylation, unlike simple genetic mutation, can be reversed or revert over time and generations. Social epigenetics provides a far more complex, yet intriguing and flexible, framework in which to explore the relationship and interaction of social and genetic expressions.

Foundational and Animal Model Epigenetic Research

The word 'epigenetics' was coined in 1942 but predated any detailed understanding of genetic mechanisms, referring only to the general conceptual model of an interaction between inheritance and the environment. The current usage of the word, referring to changes in gene expression outside fundamental changes in inherited DNA, dates only to the mid-1990s. It is not surprising, then, that much of the research that has raised both prospects of, and expectations for, social epigenetics has been basic research on epigenetics mechanisms and largely derived from animal model research.

Curley et al. (2011) reviewed much of the animal model evidence for social influences and epigenetic effects, especially methylation, during developmental life stages. Authors argue the considerable animal model evidence supports an epigenetic intervening mechanism mediating the relationships of social experiences, gene expression, and neurobiological and behavioral effects. Evidence also suggests, however, the complexity of epigenetic effects with different sensitivity to environmental effects across individuals and sexual dimorphism in effects. And, a similar panel review of behavioral epigenetic evidence (Lester et al., 2011), emphasizing histone modification and primarily relying upon animal model research, reached somewhat similar conclusions and cautions regarding the complexity of epigenetic causal chains within human subjects. Much of this existing animal model research that directly

emphasizes social-behavioral impacts has been of a cognitive nature; how epigenetic changes, usually from *in utero* developmental environments, may alter brain function later in life.

Cognitive and Social Epigenetics

Early emphasis on cognitive epigenetics likely dates from findings near the end of the twentieth century emphasizing the role of methylation, or epigenetics, and parental inheritance in the developmental disorder Angelman syndrome. Research on human subjects and neurobiological or cognitive consequences remain, however, limited. Despite a vast amount of literature on epigenetics, one recent review of the research specifically on behavioral epigenetics found only 96 published articles by 2011 (Lester et al., 2011). Still, some opportunities to extend animal model research on cognitive and behavioral effects have been found. Using brain tissue from suicide victims McGowan et al. (2009), among others, created a quasi-experimental comparison of those with reported childhood abuse and those without. This study replicated their animal model work demonstrating methylation of glucocorticoid receptors, and possible later-life stress-coping effects, among those who had suffered childhood abuse. Miller (2010) summarizes this research and several threads of more recent, specifically cognitive, epigenetic research including changes in gene expression through histone modification that are related to learning and memory, as well as age-related changes in memory. This research has spawned, in turn, the exploration of potential treatments for age-related diseases including Alzheimer's.

Social epigenetics and epidemiological epigenetics have a much wider range of interest than cognitive impacts. These interests extend to the impact of social behaviors and circumstance throughout the life course on epigenetic outcomes ranging from direct health consequences to diffuse effects on subsequent social behavior through both cognitive and other pathways (McGowan et al., 2009; Szyf et al., 2008). Research working from social behavior to epigenetic impact, for example, has explored the plausible effect of diet, stress, isolation, socioeconomic circumstances, etc., on both *in utero* and life-course epigenetic outcomes. Research working from the epigenetic to social, has addressed, for example, the plausible association of epigenetic expression on disease prevalence and health, stress coping behaviors, cognitive disorders or even social expression in deviant behavior. The potential for intergenerational nonrecursive effects, for example, the impact of social disadvantage on potential gene expression reinforcing disadvantage in offspring, has also been one of the most intriguing prospects of social epigenetics for future research. And, unlike the limited availability of biological tissues for cognitive or neurobiological behavioral epigenetics, a large number of human epigenetic studies have used a variety of available human tissues, or other bioassays, to focus on epigenetic epidemiology of specific disease outcomes.

Epidemiological Epigenetics

Much of the epigenetic research considering social influences is concentrated on the role of epigenetic methylation in various cancers and the potential for demethylation as a preventative or active treatment. In much of this work social covariates, for example, diet, socioeconomic resources, environmental exposures, neighborhood effects, and life-course behaviors such as childbearing and rearing, are taken as independent variables related to epigenetic change with disease prevalence as the only social outcome considered. The list of diseases and health conditions of social import that are being studied from an epigenetic framework and using social covariates is large, including many cancers, developmental disorders, obesity, Parkinson's, Alzheimer's, cardiovascular diseases, diabetes, and many more (e.g., Tollefsbol, 2012). The effort to find epigenetic damage linked to both social exposures and disease, i.e., social epigenetic epidemiology, is becoming well established and proceeding rapidly. That is not to say that positive findings in this area of research are yet abundant. However, methods have been developed for a growing variety of biological assays, links of social environment to increased methylation have been demonstrated, and methylation associated with genes of epidemiological importance have been identified (e.g., Oberlander et al., 2008). The paradigm shift has passed and this research is now in a phase of normal science. However, as may be expected, an increasing volume of human epigenetic research is also quickly evidencing more complexity than a simple framework might suggest, and evoking caution if not outright skepticism. Miller (2010) notes that researchers looking at diffuse socioeconomic impacts have not had the success they hoped for, epigenetic effects may be complicated by the fact that effects likely vary by different types of tissues, there is more than one way in which gene expression may be modified, and both individual variability and sexual dimorphism appear to exist in epigenetic susceptibility.

Limitations and Future Social Epigenetic Research

The promise of social epigenetics is evident even to critics, "epigenetics could turn out to be at the heart of some of the most vexing problems in society. These ills include the long-term health problems of people raised in lower socioeconomic environments the vicious cycle in which abused children grow up to be abusive parents and the struggles of drug addicts trying to kick the habit" (Miller, 2010). However, as these critics also note, social epigenetics maybe getting well ahead of itself in the simple fact that much of this speculative value is predicated on the results of animal model research and "there's very little evidence in humans that epigenetics connects early life experience to behavioral or health problems later in life" (Miller, 2010). Moreover, the demonstrated relationships between early life socioeconomic status on gene expression in human populations has not been tied to epigenetic mechanisms as a cause of those observable differences in expression. Much of the work on human populations is still limited to exploratory studies seeking effects on specific health outcomes and not driven by specific hypotheses of any plausible mechanism. In short, to prove the promise of social epigenetics, future research must demonstrate the significance of the mechanism in human populations and relevant to social problems and concerns.

The natural complexity of epigenetic science is also not the only limiting factor to social epigenetic research. For the social and behavioral scientist, this research comes with a new set of demands and research roles that ultimately can only be met by a 'big science' collaborative model of research. Although short-term or even cross-sectional studies can begin to provide critical evidence, the most interesting and definitive studies will certainly require longitudinal data of a very demanding nature. In addition to the obvious needs for bioassays and laboratory work, social epigenetics requires unique statistical skills including, most frequently, large-array methylation analyses and, at the same time, skills in collection of longitudinal social and behavioral data of a detailed nature. Any given study may require a nutritionist, endocrinologist, oncologist, neurologist, psychologist, sociologist, etc., simply to formulate a conceptual framework coherently. Advances will of necessity be collaborative, large scale, multidisciplinary, and costly. This simple reality may ultimately be the most limiting factor on the future progress and promise of social epigenetics.

See also: Culture as a Determinant of Mental Health; Epigenetic Inheritance; Epigenetics of Brain Plasticity and Behavior; Epigenetics; Gene–Environment Interplay; Genetic Studies of Behavior: Methodology; Genetics and Society; Occupational Health; Social Epidemiology; Stress: Effects on Neonatal and Child Development; Twin Studies and Epigenetics.

Bibliography

Curley, J.P., et al., 2011. Social influences on neurobiology and behavior: epigenetic effects during development. Psychoneuroendocrinology 36, 352–371.

Lester, Barry M., et al., 2011. Behavioral epigenetics. Annals of the New York Academy of Sciences 1226, 14–33.

McGowan, P.O., et al., 3 March 2009. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. Nature Neuroscience 12, 342–348.

Miller, G., 2 July 2010. The seductive allure of behavioral epigenetics. Science 329, 24–27.

Oberlander, T.F., et al., 2008. Prenatal exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant corticol stress responses. Epigenetics 3, 97–106.

Szyf, M., McGowan, P., Meaney, M.J., 2008. The social environment and the epigenome. Environmental and Molecular Mutagenesis 49 (1), 46–60.

Tollefsbol, T., 2012. Epigenetics in Human Disease. Academic Press, London.

Social Exclusion, Social Deprivation and Health

Tarani Chandola and Richard Conibere, University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The literature on social deprivation, social exclusion, and health forms part of the vast body of evidence on the social inequalities in health. People in the most disadvantaged social groups and communities experience a multitude of different types of social exclusion processes such as homelessness, school exclusions, or long-term unemployment, which in turn results in poor health, disease, and disability. The associations between poor health and different dimensions of deprivation, such as income, area, employment, and educational deprivation are well documented. However, there remains some debate about the causal direction of the association. The links between poor health and social exclusion processes such as discrimination, mental illness, homelessness, and asylum seeking are also well documented. Policy responses have tended to focus on lifting the most deprived and excluded individuals out of poverty. However, others have argued on the need for proportionate universalism, which focuses on reducing the steepness of the social gradient in health, through policy actions that are universal, applicable, and appropriate across different social groups.

Concepts and Definitions Related to Social Exclusion and Social Deprivation

The concepts of social deprivation and social exclusion share a similar focus on the inability of individuals to participate fully in the life of their community or society. The measurement of social deprivation has tended to emphasize a lack of material or financial resources that contributes to a lack of social participation, whereas measures of social exclusion have emphasized the lack of participation to a broader range of social, cultural, and political activities (Levitas et al., 2007).

The literature on social deprivation, social exclusion, and health forms part of the vast body of evidence on the social inequalities in health, which incorporates the deprived and excluded groups, as well as other social groups that are relatively better off (CSDH, 2008). People in the most disadvantaged social groups and communities experience a multitude of different types of social exclusion processes such as homelessness, school exclusions, or long-term unemployment. People who are relatively poor but not excluded may experience similar deprivations, such as unemployment or poor living conditions. However their exposure to the range and length of deprivations may be limited compared with those who are socially excluded (Barnes, 2005).

There are a number of different but related dimensions of social deprivation and social exclusion. Some of the key dimensions are discussed below in relation to their associations with health.

Social Deprivation and Health

Income Deprivation, Poverty, and Health

The links between absolute poverty and health are well-known. Those in absolute poverty lack the income to meet basic human needs such as adequate food, water, and shelter. The consequent malnutrition, poor sanitation, and hygiene then tend to lead to higher risks of poverty-originated illnesses and disease. There is some debate about the causal direction: does poverty lead to ill-health or does ill-health cause poverty. However, most commentators agree that the association is likely to be bidirectional (Wagstaff, 2002). The World Health Organization (WHO) has commissioned reports from both perspectives (CSDH, 2008).

Poverty reduction programs tend to focus on the lack of income dimension of poverty and aim to supplement the income of the poor through direct cash transfers or transfer of goods or services in kind. Some of these transfers are 'means tested,' which require the poor to qualify for such transfers on the basis of an eligibility criterion such as having an income below a certain threshold. However, poverty reduction programs often fail to include improving health as a key priority (Bartlett, 2011).

In recent years, one type of poverty reduction program, Conditional Cash Transfer (CCT) programs, have become increasingly popular. The aim is to reduce poverty by making cash or goods transfers conditional upon the recipient's behaviors that are often health promoting, such as getting regular check-ups at the doctor's office, or receiving vaccinations, or the like. A review of conditional cash transfer programs and health (Lagarde et al., 2007) suggests that conditional cash transfer programs are associated with important health improvements such as increased consumption of nutritious food, increased uptake of preventive health services, and decreased incidence of diarrheal disease. However, the review also highlights a number of potential disadvantages to these programs including their cost-effectiveness. Furthermore, such poverty reduction programs typically pay little attention to the structural determinants of health such as increasing access to quality health care and other services, limited interventions in quality or availability of employment, and limited attention to processes that result in social exclusion (Lagarde et al., 2007).

Area Deprivation and Health

The association between area deprivation and poor health has been noted since, at least, the nineteenth century, when stark life expectancy differences between richer and poorer cities were described (Wohl, 1983). Public health reform in the

nineteenth century was mainly concerned with tackling the harmful effects of the physical and social environment, such as the lack of clean water and air, drainage, sewerage, decent housing, education, and regulated working conditions (Fee and Porter, 1992).

Since the 1990s, there has been a renewed interest in the association between area deprivation and health (Macintyre et al., 2008). There has been an extensive literature on the role of compositional (referring to the characteristics of residents) and contextual (referring to the characteristics of the area) determinants of area inequalities in health (Diez-Roux, 1998).

Most studies observe that even after controlling for individual level deprivation, area deprivation remains associated with poorer health and behaviors (Pickett and Pearl, 2001). Rates of ill-health in deprived areas are greater than rates associated with individual level deprivation, even after accounting for individual health risk behaviors such as smoking and drinking. Although compositional factors have the strongest associations with health and health behaviors, contextual factors in terms of where you live also explains some of the association between area deprivation and health. In addition to compositional and contextual factors, collective factors are another type of explanation for area inequalities in health (Macintyre et al., 2008), which emphasizes the importance of shared norms, traditions, values, and interests.

The 'inverse care law' (Tudor Hart, 1971) or 'deprivation amplification' (Macintyre, 2007) are examples of contextual determinants. These concepts describe how people living in deprived areas are also disadvantaged in terms of access to good quality medical services, such as consultation length, referrals to specialist health care, and health promotion clinics, despite having greater need for such services because of the higher rates of illness and disease in those areas. Environmental injustices are a related idea, which suggest that environmental threats to health (e.g., waste disposal sites, and toxic polluting industries) are more likely to be located in deprived areas occupied by the least powerful in society (Hofrichter, 2000).

One of the main critiques of studies on area deprivation and health is that ecological unit ('neighborhood') is portrayed as a 'mystical black box' (Macintyre et al., 2002), failing to capture clear causal pathways between area deprivation and the health outcomes for the residents. Also the ecological fallacy may operate relationships observed between characteristics of area may not necessarily apply to the individuals living in those areas. Additionally there is the question of scale – do the administrative units of convenience such as census tracts, wards, and output areas really represent a neighborhood or context to which residents identify and belong. Related to this is the Modifiable Areal Unit Problem where conclusions of studies can change according to the arbitrary set of boundaries used (Holt et al., 1996).

Employment Deprivation/Long-Term Unemployment and Health

Unemployment is generally associated with poorer health and lower well-being. A *metaanalysis* of 42 longitudinal studies found that the unemployment is associated with a 63% increased mortality risk than the population as a whole (Roelfs et al., 2011). Furthermore, long-term unemployment is associated with poorer health than short-term unemployment (Nichols et al., 2013). Long-term joblessness could also result in higher mortality (Couch et al., 2013).

There is a long history of research showing that becoming unemployed has large negative effects on mental health, but the evidence on long-term unemployment and worsening mental health is more mixed. There is at least a twofold increased risk of mental health problems among the long-term unemployed, particularly depression and anxiety disorders, compared with the employed (Herbig et al., 2013; Paul and Moser, 2009; Brown et al., 2003). Higher rates of long-term unemployment are associated with increased suicide rates, (Classen and Dunn, 2012), alcohol-related disease, mental illness, and suicide attempts (Browning and Heinesen, 2012). However, there is also some evidence that a longer duration of unemployment is positively correlated with well-being, conditional on being unemployed (Clark and Oswald, 1994).

The mechanism for this increase in risk of mortality and mental health problems is unclear but could be related to income loss, increases in risky health behavior (Browning and Heinesen, 2012), and losses of health insurance coverage (Olson, 1992). In addition, when individuals are out of work for long periods, their skills tend to decline through lack of use, with negative consequences for finding reemployment and potential wages once they find a job. Similarly, being out of work potentially reduces a worker's network of business or social contacts that enables access to jobs. Furthermore, household income decreases during unemployment, and the loss in income becomes larger with longer durations of unemployment. Although most welfare states have some form of unemployment welfare payments, replacement income is often lower than the minimum income needed for healthy living (Morris et al., 2000). In addition, the long-term unemployed may suffer from the stress of being out of work and having to look for work. Lower incomes during long periods of unemployment can affect a person's and their family's well-being. Long-term unemployment within communities could also negatively affect the provision and quality of public services (Nichols et al., 2013).

Despite these plausible mechanisms linking longer periods of unemployment to poorer health, there is considerable debate over the direction of causality. This is because of selection processes in which workers with poor health or disabilities are more likely to become unemployed and less likely to be reemployed. The long-term unemployed may have underlying health problems that result in becoming unemployed. For example, workers with disabilities or mental health problems are more likely to lose their jobs during periods of economic downturns than other, more able bodied workers (Schneider, 2011). While bidirectional causal associations may exist between long-term unemployment and poor health, it is difficult to isolate and estimate the effects from any one direction.

Educational Deprivation/Lack of Qualifications and Health

Childhood deprivation, educational attainment, and later-life outcomes in adulthood are closely linked (DCSF, 2009). Childhood deprivation has a negative impact on educational attainment resulting in increased risk of a child leaving full

time education without any qualifications or skills. Such educational deprivation has indirect and direct consequences on well-being. There may be direct effects on employment and earnings, as people without any qualifications find it harder to find jobs, particularly skilled jobs. The effects of employment deprivation on health have been discussed in the previous section. In addition, lower skilled people are more likely to find employment in physically dangerous occupations, or poorly regulated work with increased risks of accidents (Feinstein et al., 2006). Educational deprivation may also be indirectly associated with lower well-being through effects on health behaviors. Low levels of education may result in a poor understanding the risks associated with risky behaviors such as smoking or drug use (Cutler and Lleras-Muney, 2010). The association between lack of education and low access to quality health care services is also very strong (Cutler and Lleras-Muney, 2010).

Literacy is one of the key skills acquired in basic education (Rajan et al., 2013). Literacy predicts good health in both developed and developing countries (De Walt et al., 2004; Kabir, 2008; CSDH, 2008) and enables a person to engage with the healthcare services and respond to public health campaigns (De Walt et al., 2004: 1232). In poor countries, female illiteracy in particular, is associated with poor health outcomes such as maternal and child mortality (Caldwell, 1986: 184–187; Sen, 1999: 195–198).

Social Exclusion and Health

The concepts of social exclusion and poverty are also highly interrelated. The concept of social exclusion emerged largely in reaction to a narrow definition of poverty largely concerned with income and consumption (Levitas et al., 2007). Poverty can be conceptualized in terms of material as well as social deprivation, whilst social exclusion emphasizes both a person's or a group's inability to participate in social, economic, political, and cultural life and also their relationships with others. And while poverty has a profound effect on some, though not all, of these aspects of social exclusion, there are other important causal factors of social exclusion such as age, disability, ethnicity, gender, and employment status.

Most of the research on social exclusion and health tends to focus on specific aspects of social life that people are excluded from (e.g., income, employment) or particular excluded groups (e.g., the homeless, refugees, specific ethnic, or racial groups). This implicitly conceptualizes social exclusion as a state, rather than a process.

There are many dimensions of social exclusion such as from economic productive processes like employment, good-quality education, living and working conditions, political participation, and cultural resources that can result in social exclusion. Each of these processes can combine and reinforce each other across a person's life course to produce varying degrees of vulnerability and social exclusion.

Discrimination

Discrimination is a key dimension of social exclusion and is also a key cause of health inequity. Discrimination results in unequal access to health services, as well as unequal access to the underlying (social) determinants of health like access to safe water, sanitation, and adequate housing and nutrition.

Direct discrimination occurs when key gatekeepers for services discriminate against the socially excluded. An example is when doctors or health care workers give poorer quality care to the illiterate, the disadvantaged, or the unempowered. This can be manifested through longer waiting times, inadequate diagnosis, the provision of less quality medication, patient segregation, or neglect in patient hygiene or nutrition. Indirect discrimination can manifest, for example, when access to information on health services is typically provided only in the language of the dominant ethnic or racial group or in accordance with dominant cultural practices.

Ethnic minorities and indigenous peoples are often subject to direct and indirect forms of discrimination. Disproportionately high rates of poverty and low rate of employment (or underemployment) among the socially excluded affect their chances of health. They are typically pushed into low-skilled jobs with higher occupational health risks, living in inadequate housing and have less food security. They tend to be poorer than other social groups, making it harder for them to access private health services. Direct and indirect forms of discrimination can interact and result in greater social exclusion. For example, poor, uneducated migrants are at greater risk of exclusion than rich, skilled migrants. Groups such as Roma, migrants, people with disabilities and the very old become particularly vulnerable when they experience multiple exclusionary processes and the social exclusion and discriminatory processes become entrenched (Popay et al., 2008).

Gender discrimination also remains a key health issue, highlighted by the poor state of women's reproductive health rights among socially excluded groups. There is evidence from around the world to show that women from marginalized minorities and indigenous peoples have disproportionately high rates of maternal mortality and often have less access to reproductive health services (MRGI, 2013). In some severe cases, minority women can be coerced or intimidated into making involuntary decisions regarding their sexual and reproductive health, such as sterilization or abortion.

Exclusion from Services and Health/Inverse Care Law

The inverse care law states that "the availability of good medical care tends to vary inversely with the need for it in the population served" (Tudor Hart, 1971). Access to services is also affected by the inverse care law. Those with the greatest need for healthcare use the services the least. The 'Inverse Care Law' has also been used to describe how socially excluded groups are also disadvantaged in terms of access, consultation length, quality of service, likelihood of diagnosis, referrals to specialists or secondary and tertiary services, waiting time, health promotion clinics, and patient enablement processes.

Exclusion due to Disability

Around 15% of the world's population, or estimated 1 billion people, live with disabilities (WHO, 2011). They are the world's largest minority. This figure is increasing through population growth, medical advances, and the aging process,

says the World Health Organization (WHO). In countries with life expectancies over 70 years, individuals spend on average about 8 years, or 11.5% of their life span, living with disabilities.

Persons with disabilities differ in the type and severity of disability, which in turn impacts on their health and well-being. Furthermore, other social determinants of health, such as gender, socioeconomic status, race, and ethnicity, interact with the types of disability and influences how people with disabilities access health care (Sullivan et al., 2006; Alborz et al., 2005). Evidence suggests that persons with disabilities, in general, are more likely than able-bodied counterparts to have multiple and complex health care needs; to perceive their health status as poor; to report having unmet health care needs; and to have lower levels of satisfaction with health care (AHS, 2008).

People with disabilities in developing countries are overrepresented among the poorest people. However, poverty reduction strategies in poor countries have often overlooked the disabled (Miller and Ziegler, 2006). In the developed world, welfare care systems take account of people with disabilities with varying degrees of success. The community model of health care largely operates with most people with disabilities living in the community. This has resulted in people with disabilities becoming more reliant on key gatekeepers to services such as general practitioners to meet their care needs (Caldas de Almeida and Killaspy, 2011). However, these gatekeepers may lack experience and specialized knowledge of complex disabling conditions, or may engender social exclusionary processes through conscious or unconscious discriminatory practices (Parish and Saville, 2006; Ruddick, 2005). The community model of care has also increased care burden stress among the informal carers of people with disabilities, particularly family members and friends (Ouelette-Kuntz et al., 2005).

Exclusion due to Mental Illness

People with mental illness tend to be overrepresented in all socially excluded groups. There is a strong link between mental illness and deprivation, with the proportion of mental health problems being much higher in disadvantaged social groups such as those classified as being in poverty, the homeless, the unemployed, minority ethnic groups, and combinations of these categories (Bonner et al., 2002). People with mental health problems are often socially excluded as a result of social stigma. Such direct discrimination can result in limited social capital, resulting in further social exclusion.

Mental health problems can cause homelessness and can also result in homelessness (Morgan et al., 2007). People with mental health problems may lose secure sources of income which makes it difficult for them to sustain their tenancies. On the other hand, the experience of homelessness is extremely isolating and can engender feelings of powerlessness and hopelessness, which in turn can cause or exacerbate mental health conditions.

Exclusion due to Homelessness

Homelessness is a significant public health concern in the developed and developing world. Access to shelter is a basic human need, so the strong association with poor health is not surprising. Homeless people without any shelter ('rough sleepers') in particular have especially poor health outcomes, and their life expectancy is much lower than the general population (Hwang et al., 2009). However, there are other manifestations of homelessness, such as staying in overcrowded, temporary, or inadequate accommodation which in turn can have a damaging effect on a physical and mental health. These include problems related to respiratory conditions, infectious diseases such as tuberculosis and sexually transmitted infections, drug and alcohol problems. Cohabitation in overcrowded homeless shelters or inadequate housing increases the risk of airborne infections, such as tuberculosis. Social isolation as a result of homelessness results in the loss of social support, which in turn is associated with low compliance with medication (Galea and Vlahov, 2002). Homelessness can also limit a person's access to adequate health or social services, as access may depend on residency in a particular area or city.

Exclusion due to Asylum Seeker or Refugee Status

Refugees and Asylum seekers differ from other groups of immigrants in key characteristics. They are more likely to have been exposed to war, violence, and torture, having experienced forced migration and exile, as well as insecurity in terms of their status in the countries where they seek refuge. Most studies on immigrants find some evidence of the 'healthy immigrant effect' – the health of immigrants tends to be better than that of the general population in both the sending and receiving countries. This effect reflects the selection processes that immigrants face when migrating; immigrants with poor health are less likely to overcome some of the social, economic, and political barriers to migration. In contrast to the general finding of healthy migrant effects, there is evidence that refugee and asylum seekers are at substantially higher risk than the general population for a variety of specific psychiatric disorders (Coid et al., 2008; Jarvis, 2007). For example, those exposed to torture are highly likely to experience symptoms of post-traumatic stress disorder (Steel et al., 2009).

Explanations for this increased risk in psychiatric disorders lie partly in the disruptions to the social support and networks of refugees and asylum seekers when they migrate. Furthermore, during migration, many refugees experience prolonged uncertainty about their citizenship status (Kirmayer et al., 2011). In some countries, harsh conditions in detention centers can lead to a sense of hopelessness among asylum seekers (Silove et al., 2007).

Criticisms of Social Exclusion

The concept of social exclusion has been criticized for being "a broad screen, a curtain, which [hides] problems of desperate destitution" (Gore and Figueiredo, 1997: 44) irrelevant to regions and countries where a majority of people are experiencing severe poverty. This approach suggests that the policy focus should be on alleviating poverty rather than on inclusionary policies. However, as Sen argues (2000: 6) "the real importance of the idea of social exclusion lies in emphasizing the role of relational features in the deprivation of capability"

and thus highlighting the causal relationship between poverty and wider inequalities in societies.

There is some debate over the measurement of social exclusion. Some studies use a series of markers of social disadvantage and differentiation (e.g., social class, employment status, contact with the legal system) to demarcate social exclusion. However, this approach ignores the concept of social exclusion as a process and the importance of agency (Perkins and Repper, 2003). Parr et al. (2004) conceptualizes exclusion and inclusion as subjective states of belonging and involvement in local communities determined by the actions of others in the immediate social milieu. Individuals, in this account, feel more or less included at different times; being excluded is not a static fixed state that can be objectively measured, but a fluid process. This implies a very different approach to studying social exclusion from that used in most studies of social exclusion and health to date.

Policy Responses

The socially deprived tend to have higher levels of ill health compared with more advantaged groups, resulting in a social gradient in health that increases with levels and durations of deprivation (CSDH, 2008). There are marked differences in health between social classes, occupational hierarchies, and status-related social groups. Policy responses to this link between socioeconomic circumstances and health have tended to focus on those in the poorest circumstances and the poorest health, on the most socially excluded, those with most risk factors, and those most difficult to reach (Graham and Kelly, 2004). This is a policy approach that attempts to lift the worst off out of extreme poverty. However, such interventions are likely to help only a relatively small part of the population as those in extreme poverty only constitute a small part of the population (Graham and Kelly, 2004).

Another approach recognizes that the link between social deprivation and health is part of a broader social gradient in health. There are large numbers of people who, while they could not be described as socially excluded, are relatively disadvantaged in health terms. Focusing solely on the most deprived or excluded will not reduce health inequalities sufficiently for this broader group of the population. To reduce the steepness of the social gradient in health, policy actions must be universal and applicable across different social groups. However, the scale and intensity of such policy actions should be proportionate to the level of deprivation and social exclusions. Such policy actions have been termed proportionate universalism (Marmot, 2010).

See also: Disability: Social and Psychological Aspects; Globalization and Health; Health Care, Rationing of; Health Literacy; Justice and Health; Life Course Epidemiology; Migration and Health; Race: Ethnicity and Health; Social Epidemiology; Socioeconomic Status and Health Care; Socioeconomic Status and Health; Unemployment and Mental Health; Urban Life and Health; Violence as a Public Health Problem.

Bibliography

AHS, 2008. Persons with Disabilities: Health Services Literature Review and Community Consultation. Available at: http://www.screening.nhs.uk/getdata.php?id=14488.

Alborz, A., McNally, R., Glendinning, C., 2005. Access to health care for people with learning disabilities in the UK: mapping the issues and reviewing the evidence. Journal of Health Services Research & Policy 10 (3), 173–182.

Barnes, M., 2005. Social Exclusion in Great Britain: An Empirical Investigation and Comparison with the EU. Ashgate, Aldershot.

Bartlett, S., 2011. Poverty reduction strategy papers and their contribution to health: an analysis of three countries. McGill Journal of Medicine 13 (2), 22–28.

Bonner, L., Barr, W., Hoskins, A., 2002. Using primary care-based mental health registers to reduce social exclusion in patients with severe mental illness. Journal of Psychiatric and Mental Health Nursing 9 (5), 585–593.

Brown, D.W., et al., 2003. Associations between short- and long-term unemployment and frequent mental distress among a national sample of men and women. Journal of Occupational and Environmental Medicine 45 (11), 1159–1166.

Browning, M., Heinesen, E., 2012. Effect of job loss due to plant closure on mortality and hospitalization. Journal of Health Economics 31 (4), 599–616.

Caldas de Almeida, J.M., Killaspy, H., 2011. Long-term Mental Health Care for People with Severe Mental Disorders. Available at: http://ec.europa.eu/health/mental_health/docs/healthcare_mental_disorders_en.pdf.

Caldwell, J., 1986. Routes to low mortality in poor countries. Population and Development Review 12 (2), 171–220.

Clark, A.E., Oswald, A.J., 1994. Unhappiness and unemployment. The Economic Journal 104 (424), 648–659.

Classen, T.J., Dunn, R.A., 2012. The effect of job loss and unemployment duration on suicide risk in the United States: a new look using mass-layoffs and unemployment duration. Health Economics 21 (3), 338–350.

Coid, J.W., et al., 2008. Raised incidence rates of all psychoses among migrant groups: findings from the East London first episode psychosis study. Archives of General Psychiatry 65 (11), 1250–1258.

Couch, K.A., et al., 2013. New analyses of worker well-being. In: Polachek, S.W., Tatsiramos, K. (Eds.), Research in Labor Economics. Emerald Group Publishing Limited, pp. 259–305.

CSDH, 2008. Closing the Gap in a Generation. Health Equity through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health, Geneva. Available at: http://www.who.int/social_determinants/thecommission/finalreport/en/index.html.

Cutler, D.M., Lleras-Muney, A., 2010. Understanding differences in health behaviors by education. Journal of Health Economics 29 (1), 1–28.

DCSF, 2009. Deprivation and Education. The Evidence on Pupils in England Foundation Stage to Key Stage 4. London. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/222172/DCSF-RTP-09-01.pdf.

Diez-Roux, A.V., 1998. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. American Journal of Public Health 88 (2), 216–222.

Fee, E., Porter, D., 1992. Public health, preventive medicine and professionalization: England and America in the nineteenth century. Medicine in Society: Historical Essays, 249–275.

Feinstein, L., et al., 2006. What are the effects of education on health? in measuring the effects of education on health and civic engagement: proceedings of the Copenhagen symposium. OECD. Available at: http://www.oecd.org/education/innovation-education/37425753.pdf.

Galea, S., Vlahov, D., 2002. Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. Public Health Reports (Washington, D.C.: 1974) 117 (Suppl.), S135–S145.

Gore, C., Figueiredo, J., 1997. Social Exclusion and Antipoverty Policy: A Debate, vol. 110. International Labour Office.

Graham, H., Kelly, M.P., 2004. Health Inequalities: Concepts, Frameworks and Policy. Available at: http://www.nice.org.uk/niceMedia/pdf/health_inequalities_policy_graham.pdf.

Herbig, B., Dragano, N., Angerer, P., 2013. Health in the long-term unemployed. Deutsches Ärzteblatt International 110 (23–24), 413–419.

Hofrichter, R. (Ed.), 2000. Reclaiming the Environmental Debate: The Politics of Health in a Toxic Culture. MIT Press.

Holt, D., et al., 1996. Aggregation and ecological effects in geographically based data. Geographical Analysis 28 (3), 244–261.

Hwang, S.W., et al., 2009. Mortality among residents of shelters, rooming houses, and hotels in Canada: 11 year follow-up study. BMJ: British Medical Journal 339.

Jarvis, G.E., 2007. The social causes of psychosis in North American psychiatry: a review of a disappearing literature. Canadian Journal of Psychiatry 52 (5), 287–294.

Kabir, M., 2008. Determinants of life expectancy in developing countries. The Journal of Developing Areas 41 (2), 185–204.

Kirmayer, L.J., et al., 2011. Common mental health problems in immigrants and refugees: general approach in primary care. Canadian Medical Association Journal 183 (12), E959–E967.

Lagarde, M., Haines, A., Palmer, N., 2007. Conditional cash transfers for improving uptake of health interventions in low- and middle-income countries: a systematic review. The Journal of the American Medical Association 298 (16), 1900–1910.

Levitas, R., et al., 2007. The Multidimensional Analysis of Social Exclusion. Department of Sociology and School for Social Policy, University of Bristol, Bristol. Available at: http://roar.uel.ac.uk/1781/1/multidimensional.pdf.

Macintyre, S., 2007. Deprivation amplification revisited; or, is it always true that poorer places have poorer access to resources for healthy diets and physical activity? The International Journal of Behavioral Nutrition and Physical Activity 4 (1), 32.

Macintyre, S., Ellaway, A., Cummins, S., 2002. Place effects on health: how can we conceptualise, operationalise and measure them? Social Science & Medicine 55 (1), 125–139.

Macintyre, S., Macdonald, L., Ellaway, A., 2008. Do poorer people have poorer access to local resources and facilities? the distribution of local resources by area deprivation in Glasgow, Scotland. Social Science & Medicine (1982) 67 (6), 900–914.

Marmot, M., 2010. Fair Society, Healthy Live: The Marmot Review. London. Available at: http://www.instituteofhealthequity.org/projects/fair-society-healthy-lives-the-marmot-review.

Miller, U., Ziegler, S., 2006. Making PRSP Inclusive. Available at: http://siteresources.worldbank.org/DISABILITY/Resources/280658-1172608138489/MakingPRSPInclusive.pdf.

Morgan, C., et al., 2007. Social exclusion and mental health: conceptual and methodological review. The British Journal of Psychiatry: The Journal of Mental Science 191 (6), 477–483.

Morris, J.N., et al., 2000. A minimum income for healthy living. Journal of Epidemiology & Community Health 54 (12), 885–889.

MRGI, 2013. State of the World's Minorities and Indigenous Peoples. London. Available at: http://www.minorityrights.org/download.php?id=1293.

Nichols, A., Mitchell, J., Lindner, S., 2013. Consequences of Long-term Unemployment. Washington, DC.

Olson, C.A., 1992. The Impact of Permanent Job Loss on Health Insurance Benefits (No. 684). Department of Economics, Industrial Relations Section, Princeton University.

Ouelette-Kuntz, H., et al., 2005. Addressing health disparities through promoting equity for individuals with intellectual disability. Canadian Journal of Public Health 96, S8–S22.

Parish, S.L., Saville, A.W., 2006. Women with cognitive limitations living in the community: evidence of disability-based disparities in health care. Mental Retardation 44 (4), 249–259.

Parr, H., Philo, C., Burns, N., 2004. Social geographies of rural mental health: experiencing inclusions and exclusions. Transactions of the Institute of British Geographers 29 (4), 401–419.

Paul, K.I., Moser, K., 2009. Unemployment impairs mental health: metaanalyses. Journal of Vocational Behavior 74 (3), 264–282.

Perkins, R., Repper, J., 2003. Social Inclusion and Recovery: A Model for Mental Health Pratice. Baillière Tindall.

Pickett, K., Pearl, M., 2001. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. Journal of Epidemiology & Community Health 55 (2), 111–122.

Popay, J., et al., 2008. Understanding and Tackling Social Exclusion. Final Report to the WHO Commission on Social Determinants of Health from the Social Exclusion Knowledge Network. Available at: http://www.who.int/social_determinants/knowledge_networks/final_reports/sekn_final report_042008.pdf.

Rajan, K., Kennedy, J., King, L., 2013. Is wealthier always healthier in poor countries? the health implications of income, inequality, poverty, and literacy in India. Social Science & Medicine (1982) 88, 98–107.

Roelfs, D.J., et al., 2011. Losing life and livelihood: a systematic review and meta-analysis of unemployment and all-cause mortality. Social Science & Medicine 72 (6), 840–854.

Ruddick, L., 2005. Health of people with intellectual disabilities: a review of factors influencing access to health care. British Journal of Health Psychology 10 (4), 559–570.

Schneider, J.A., 2011. Who are the Long-Term Unemployed in This Recession and what Can Be Done to Help Them?. Washington, DC.

Sen, A., 1999. Development as Freedom. Oxford University Press.

Sen, A., 2000. Social Exclusion: Concept, Application, and Scrutiny. Available at: http://housingforall.org/Social_exclusion.pdf.

Silove, D., Austin, P., Steel, Z., 2007. No refuge from terror: the impact of detention on the mental health of trauma-affected refugees seeking asylum in Australia. Transcultural Psychiatry 44 (3), 359–393.

Steel, Z., et al., 2009. Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: a systematic review and metaanalysis. JAMA: The Journal of the American Medical Association 302 (5), 537–549.

Sullivan, W., et al., 2006. Consensus guidelines for primary health care of adults with developmental disabilities. Canadian Family Physician 52 (11), 1410–1418.

Tudor Hart, J., 1971. The inverse care law. The Lancet 297 (7696), 405–412.

Wagstaff, A., 2002. Poverty and health sector inequalities. Bulletin of the World Health Organization 80 (2), 97–105. Available at: http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0042-96862002000200004&lng=en&nrm=iso&tlng=en (accessed 07.02.14.).

De Walt, D.A., et al., 2004. Literacy and health outcomes: a systematic review of the literature. Journal of General Internal Medicine 19 (12), 1228–1239.

WHO, 2014. Macroeconomics and Health (CMH). World Health Organisation, Geneva. Available at: http://www.who.int/macrohealth/en/.

WHO, 2011. World Report on Disability. World Health Organisation, Geneva. Available at: www.who.int/disabilities/world_report/2011/report/en/.

Wohl, A.S., 1983. Endangered Lives: Public Health in Victorian Britain. JM Dent & Sons Ltd. Available at: http://www.cabdirect.org/abstracts/19842015295.html (accessed 04.02.14.).

Social Experiments

Larry L Orr, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
Rebecca A Maynard, Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides the foundation for understanding social experiments and why it is valuable to conduct them. (For a more detailed discussion of most of the topics included in this article, see Orr (1999).) It begins by defining the term and discussing the rationale for using experimental methods to evaluate social policies, practices, and programs. It then discusses the design and analysis of social experiments, and the issues of internal and external validity of social experiments. The article concludes with a brief discussion of the ethics of such studies and an historical perspective on social experimentation.

What Is a Social Experiment?

A social experiment is a field trial of a policy, practice, program, or other social intervention in which participants are randomly assigned to different 'treatments' or experimental conditions. An experiment may include multiple interventions, or 'treatment groups,' and usually, though not always, a 'control group,' which receives no special treatment. The control group represents the 'counterfactual,' or what would have happened to the treatment group in the absence of the treatment. Therefore, comparing the mean postrandom assignment outcomes of a treatment group with those of the control group yields a measure of the 'impact' of that intervention relative to what the controls would have experienced in the absence of the intervention.

For example, an experiment designed to test two different types of job search assistance (JSA) might randomly assign job-seekers to three groups: a group that receives JSA Type 1 (JSA1), a group that receives JSA Type 2 (JSA2), and a control group, which receives no special assistance. The effect of JSA1 relative to no assistance – its 'net impact' – would be measured by the difference in mean postrandom assignment employment and earnings between the JSA1 group and the control group. Similarly, the effect of JSA2 relative to no assistance would be measured by the difference in mean employment and earnings between the JSA2 group and the control group. The 'differential impact' of JSA1 relative to JSA2 would be measured by the difference in mean outcomes between the two intervention groups. Net impact estimates answer the question, what effect would adopting the intervention have relative to not adopting the program? Differential impacts answer the question, which type of intervention is more effective? In the following section, we elaborate on these ideas and show why social experiments with random assignment have become the 'gold standard' of program evaluation.

Why Experiment?

In popular parlance, 'experimenting' is usually taken to mean implementing a new program, policy, or practice (also referred to throughout this section as *a strategy* or *intervention*) on a small scale to see if it works. If the intervention is truly new, then only by setting up a pilot test is it possible to obtain empirical evidence of its effects. One rationale for social experiments, then, is to create a working model of the intervention so that it can be evaluated. However, determining whether the strategy works requires much more than simply observing it in action.

To see why, it is useful to consider what we mean when we say that a strategy does or does not 'work' and what ways we have of making that determination. We take the question of whether a strategy works to mean whether it is effective in achieving its goals. For most programs, those goals have to do with the effects of the program on its intended beneficiaries. Some interventions have a single goal; others have multiple goals. For example, the goal of a job training program is to increase the earnings of the trainees, whereas a community service program might have multiple objectives, such as providing valuable services to the community, encouraging out-of-school youths to obtain a General Educational Development (GED) certificate, and reducing their antisocial behavior. The behaviors and circumstances of those who are the targets of the intervention under study any time after they are first exposed to the intervention are termed 'outcomes'. (Because some effects can occur almost immediately, while the participant is still in the in the program under study, we define outcomes to include everything that happens after random assignment, not just what happens after the participant leaves the program.) To determine whether the strategy being tested works, we must determine whether it succeeded in changing the outcomes it was intended to affect.

One can, of course, observe the outcomes of interest – for example, postprogram earnings of trainees or GED completion rate of service corps members – simply by following individuals during and after the time when they are exposed to the intervention. But one cannot, on the basis of the observed behavior alone, know what portion of the observed post-program outcomes should be attributed to the intervention. We define the '*impact*' of a program, policy, or practice under study as the difference between the observed outcomes for those subjected to the intervention condition (for example, their postprogram earnings) and *what those outcomes would have been in the absence of the intervention under study*.

Suppose, for example, that 80% of the participants in a job training program obtain jobs within a year of entering the

program. Does that mean that the program is achieving its objective of securing employment for unemployed workers? Not necessarily. Some of the trainees would have found jobs even if they had not participated in the program. If only 20% would have found jobs in the absence of the program, then the program increased the participants' employment rate by 60 percentage points. However, if 80% of those in the program would have found work on their own in the absence of the program, then the program has had no impact on their employment rate. (Of course, in the latter case, the program might have helped them get better jobs, and conceivably it could improve their long-term employment and earnings. Thus, it would be useful also to consider impacts on earnings and possible longer-term outcomes in addition to the 1-year employment rate.)

Similar reasoning applies to other programs, policies, and practices and their outcomes and objectives. Identifying the impacts of particular strategies requires being able to measure the outcomes for those who experienced the particular strategy under study and being able to compare them with what those outcomes would have been had those same individuals *not* experienced the intervention. We can only attribute behavioral changes over and above that base level to the intervention. The fundamental problem of program evaluation, then, is to determine what would have happened in the absence of the intervention. Measuring the actual outcomes of participants is relatively straightforward; not surprisingly, measuring what they would have done in the absence of the intervention under study is much more difficult.

There are, however, a number of different ways to attempt to *estimate* what would have happened to the participants in the absence of the intervention, which we term the *counterfactual* condition. To understand why experimental methods are the preferred approach for estimating what would have happened in the absence of the intervention, it is useful to understand the shortcomings of other approaches. Therefore, we briefly discuss the two principal alternatives to experimental designs: (1) *pre–post designs*, which are also called *reflexive* designs or *one-group pretest–posttest designs* (Campbell and Stanley, 1963, pp. 6–13) and (2) *comparison group designs*, sometimes called *static-group comparison designs* or *double-difference designs*. These designs are two of the most intuitive and common ways to think about whether a particular intervention improves outcomes in the intended ways. (These two broad approaches subsume a large number of specific statistical impact estimation methods. For example, pre–post designs include 'interrupted time series' methods that estimate impacts on the basis of postintervention departures from the preintervention *trend* in the outcome, and comparison groups can be formed through a variety of selection and/or matching procedures. The fundamental characteristics, advantages, and disadvantages of these approaches discussed here apply to all of these more specific estimation methods.)

Pre–Post Designs

A particularly simple way to attempt to determine what would have happened in the absence of the intervention under study is to use the behavior of the participants before they experienced the intervention as a predictor of what their outcomes would have been had they not been exposed to the experimental intervention. In the case of a social development program for young children, for example, we might use the participants' social competence scores in the year before they entered the program as an indicator of what their outcomes would have been in the absence of the program. We would then measure the impact of the program as the change in social competence scores between the year prior to program participation and the year after it. This design has the advantage of requiring data only on the individuals who participated in the program under study.

Importantly, pre–post designs are not always feasible and, even when they are, they have substantial limitations. The outcome of interest might not be defined in the preprogram period. For example, pre–post designs could not be used to estimate the effect of a maternal health program for pregnant women on infant birth weight, which can only be measured after the intervention has been delivered. More importantly, even where preprogram data can be collected, for several reasons behavior in the period prior to exposure to the intervention may not be a good predictor of what would have happened later, in the absence of the intervention.

Environmental factors external to the program that affect the outcomes of interest might change over the time period between the pre- and postmeasures. In the case of a dropout-prevention program, for example, deterioration of job opportunities in the local economy might result in fewer students dropping out of school to go to work, quite aside from any effect of the program. In this case, a pre–post impact measure would erroneously attribute this improvement in school retention to the program rather than to changes in the opportunity cost of attending school.

Of particular importance to interventions focused on children, outcomes may also change over time because of natural *maturation processes*. Consider, for example, the social development program mentioned above, in which preprogram social competence scores are used to estimate the postprogram counterfactual. Even without a special program, children's social skills can be expected to improve over time, e.g., through interaction with parents and siblings at home. Thus, the pre–post impact measure would overstate the effects of the program by including in that measure of program impact those gains in skills that would have occurred anyway.

Even when there are no environmental factors or maturation effects that render pre–post measures poor indicators of the impacts of an intervention, pre–post designs can yield misleading conclusions if, on average, the preprogram period was atypical for subjects exposed to the intervention. Consider, for example, job training programs for workers who have lost their jobs. Even without the program's assistance, many of these workers would have found jobs on their own. Thus, even without the program, some program participants would show improved performance from the period before to the period after program entry. Statisticians refer to this as *regression to the mean*. A pre–post evaluation would attribute these gains in employment to the training program. (This type of effect is most pronounced when program participants are drawn disproportionately from one extreme of the performance distribution, in this case the lower end.)

This phenomenon of individuals participating in programs when their outcomes of interest are at atypical levels is fairly general. Social interventions are often targeted at individuals based on indicators of need that are similar to those that will be the basis for judging the effectiveness of the intervention (for example, health, earnings, or academic performance). To the extent that these needs would be only temporary even without the intervention, selection on need will result in regression to the mean, with need greatest around the time of program entry, then falling over time. In such cases, simple pre–post differences in outcomes will overstate the effect of the program by including the rebound from this temporary period of need, which would have occurred even without the program's help.

Figures 1 and 2 illustrate how reliance on the outcomes of intervention participants alone can create a misleading impression of the effectiveness of an intervention.

Figure 1 shows the time path of welfare benefits received by participants in the Child Assistance Program, a welfare-to-work demonstration (Hamilton et al., 1998). The steep, continuous drop in welfare benefits over a 5-year period would appear to indicate that the program was quite effective in achieving its objective of moving assistance recipients from welfare to employment. This is in fact the type of 'evidence' that is frequently used to demonstrate program effectiveness in the legislative process and in the popular media.

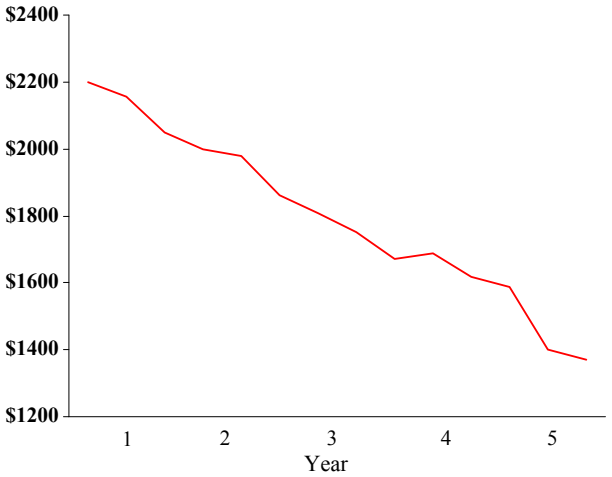

Figure 1 Welfare benefits of participants in a welfare-to-work program.

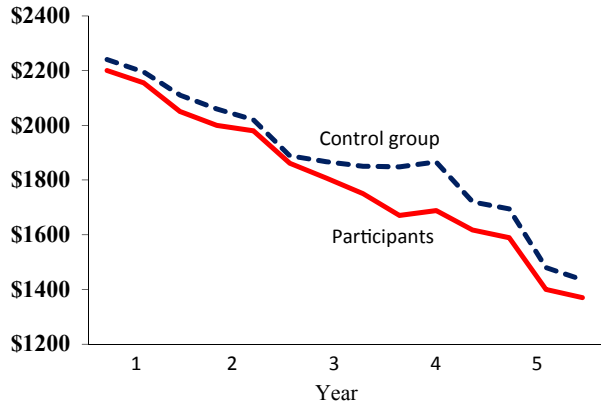

Figure 2 Welfare benefits of participants in a welfare-to-work program and a randomly assigned control group excluded from the program.

As shown in Figure 2, however, a control group of program eligibles who were *excluded* from the program on the basis of a lottery, as part of a social experiment, experienced essentially the same decline in welfare benefits over the same period. This comparison demonstrates that most of the reduction in benefits received by program participants was not attributable to the program, but to the natural process of caseload turnover, as women leave welfare when they find employment or another source of support (e.g., a working partner or spouse or parent), or their children age out of eligibility.

Comparison Group Designs

As the previous example suggests, one way to avoid some of the hazards of pre–post designs is through the use of a comparison group – a group of individuals who are as similar as possible to those selected to experience the intervention under study, except that they are not exposed to the intervention. Under this strategy, the pre–post change in outcomes of the comparison group is used to represent what would have happened to those in the intervention group had they not been exposed to the program, policy, or practice under study.

Ideally, the pre–post change in outcomes of the comparison group will reflect any environmental factors, natural maturation effects, and rebound from atypical behavior or circumstances that affect the outcomes. If that is true, subtracting the pre–post change in outcomes of the comparison group from the pre–post change in outcomes of the intervention group should eliminate these nonintervention factors and yield a more accurate measure of the impact of the program, policy, or practice under study. This is sometimes called the *double-difference* model or, in the language of Campbell and Stanley (1963, pp. 47–50), the nonequivalent comparison group model.

Table 1 illustrates this approach using data from the evaluation of the Job Training Partnership Act program, a national training and job placement program for disadvantaged workers (Orr et al., 1996). The outcome of interest, quarterly earnings, was measured for participants in the quarters before program entry and 18 months later. The first row of the able shows that a simple pre–post measure of impact, using data for participants only, would indicate that the program increased earnings by $281, from $400 per quarter to $681. The second row shows the change in earnings of a sample of low-wage workers who did not apply to the program. Suppose we use that sample as a comparison group. Under the assumption that *in the absence of the program* JTPA participants would have experienced the same increase in earnings as the nonparticipant sample, the comparison group approach estimates the impact of the program as $136 – the change in average earnings from pre- to post-program for the JTPA participant group less the pre–post change for the comparison group (see row three).

The comparison group design thus avoids the erroneous attribution of the full increase in earnings that results pre–post

Table 1 Estimating the impact of JTPA with a comparison group design

Group	*Preprogram earnings*	*Earnings 18 months after program entry*	*Pre–post change*
1. Treatment group (JTPA participants)	$400	$681	$281
2. Comparison group	$1050	$1195	$145
3. Estimated impact (difference in differences)	–	–	$136

to the program in question. Rather, only the change above and beyond that experienced by the comparison group is attributed to the intervention.

Importantly, it is still possible, however, that the net difference in outcomes will not be a fair estimate of the impact of the intervention; the measure is potentially biased due to the way the intervention and comparison groups were selected. Because the pervasiveness of 'selection bias' in nonexperimental comparison group designs is one of the most important reasons for preferring experimental methods, we now examine this phenomenon in more detail. We then return to the estimation of JTPA impacts in light of that discussion.

Selection Bias

The comparison group design is based on the assumption that those in the intervention group would have experienced the same change in outcomes as the comparison group had they not been subject to the intervention being studied. Estimates based on comparison group designs are only as valid as this assumption. Ultimately, one can never be sure how valid that assumption is because it is a statement about something that is inherently unobservable – the experience of the intervention or treatment group if they had not experienced the intervention.

What we do know is that the intervention group is either self-selected or selected by somebody else (for example, students might be selected by teachers) to experience the intervention, whereas the comparison group members were not selected in the same way. Unless those selection decisions are identical for both groups, the two groups may well differ in ways that may or may not be observable. If the difference(s) that led one group to be selected for the intervention and the other not to be selected also lead to differences in the outcomes of interest, the comparison group design will erroneously attribute to the intervention those differences in outcomes that are really due to differences in who was selected for the respective groups. Such errors in attribution are termed *selection bias*.

One can, of course, attempt to match the comparison group to the participants in terms of such personal characteristics as age, race, gender, ethnicity, prior earnings, etc. However, if the two groups differ in *unmeasured* characteristics, such as motivation, native ability, or family and community resources, their outcomes might differ for reasons that have nothing to do with the intervention under study. However, even when the two groups are matched on measured characteristics, selection bias can occur based on those characteristics if the measured status of one group is temporary and that of the other is permanent. Suppose, for example, participants in the JTPA program are matched to a sample of workers in the general population with similar recent employment and earnings. If the trainees applied to the program because they had recently lost their jobs, their low earnings are likely to be temporary, whereas workers in the general population with similar earnings are likely to have permanently low earnings. When the trainees' earnings rebound from their temporarily low levels, it will appear that the program substantially increased their earnings. (This well-known phenomenon, termed the 'preprogram dip' in the job training literature, was first known as the 'preprogram dip,' was first noted by Ashenfelter (1978).) In fact, if the impact of JTPA on quarterly earnings at 18 months after program entry is estimated with a randomly assigned control group rather than a nonexperimental comparison group, it is found to be $53, rather than the $136 shown in **Table 1**. (This estimate is provided for illustrative purposes only; the actual impact estimates in the JTPA evaluation were derived by a more complex regression-based procedure (see Orr et al., 1996).)

The Rationale for Experimental Designs

Random assignment offers a way to create a comparison group that is not systematically different from the group subject to the intervention under study – that is, one that is not subject to selection bias. If assignment to the intervention group or to the comparison group is completely random, selection into one group or the other is *by definition* unrelated to any characteristics of the individual and, therefore, to the outcomes expected in the absence of the intervention. When individuals have been randomly assigned to the experimental conditions, any systematic differences in postrandom assignment outcomes between the groups can confidently be attributed to the interventions, if any, that individuals in the groups received. (By *systematic* differences, we mean any differences that are larger than might be expected on the basis of sampling error. Below, we discuss how one can test whether the between-group difference is likely to be due to sampling error.)

The defining characteristic of a *social experiment* is that it entails random assignment of some pool of individuals (or clusters of individuals such as whole schools or classrooms) to two or more groups that are subject to different program, policy, or practice regimes. Typically, one of these groups is a *control group* that is subject to prevailing, 'business-as-usual' conditions (that is, it is excluded from the experimental intervention). (Experimental control groups are a special type of comparison group. For clarity we will use the term 'comparison group' to mean a nonexperimental comparison group and we use the term 'control group' to denote a comparison group created through random assignment. Occasionally, a social experiment does not include a control group subject to the prevailing conditions, but rather consists of two or more groups created through random assignment that receive different interventions.) In addition, one or more *intervention groups* consist of individuals or clusters of individuals who are assigned to

the intervention being studied. The relevant outcomes of each group are then measured over some follow-up period and the impact of the intervention is estimated as the difference in outcomes between intervention and control groups. Because random assignment assures that exposure to the intervention under study is the only *systematic* difference between the treatment and control groups, the observed difference in subsequent outcomes between those in the intervention and the control groups will be free of selection bias; the difference in means between the two groups will be an unbiased estimate of the impact of the intervention.

By *random assignment* we mean assignment to groups on the basis of a random event, such that (1) the outcome of assignment is unpredictable, and (2) each individual or cluster has a specified probability of being assigned to each group. (Random assignment does *not* simply mean haphazard or arbitrary assignment. In practice, great care must be taken to ensure that each individual or cluster assigned has the prescribed probability of assignment to each group.) The random event can be as simple as the flip of a coin – if the coin comes up heads, the individual is assigned to the treatment group; if it comes up tails, the individual is assigned to the control group. In this example, each individual has a 50% chance of assignment to each group; if large numbers of individuals are randomly assigned, the total sample will be divided approximately evenly between the treatment and control groups, and the groups will have statistically similar characteristics.

In practice, random assignment is usually based on specially designed tables of random numbers or computer algorithms that generate random numbers. In most cases, the unit of randomization (individual or cluster) should be the same as the unit to which the intervention under study pertains. For example, in a study of supplemental education services intended for students with particular needs, individual students would usually be assigned to the intervention or control condition, whereas if the intervention were a new math curriculum, it would be appropriate to randomly assign classrooms. (Special statistical considerations pertain to situations in which clusters of individuals (such as classrooms) are assigned to treatment or control conditions.)

Suppose, for example, that we wish to evaluate an education voucher program such as the one that was implemented and studied in Washington, DC (Wolf et al., 2008). We might begin by setting up the program on a pilot basis and recruiting applicants. Those judged eligible (and still interested after learning more about the program) would then be randomly assigned to an intervention group, which would receive vouchers that could be used to pay for private school tuition, or to a control group, which would not receive vouchers. Over the next several years, we would collect data for both groups on outcomes that might be expected to be affected by receipt of the voucher, such as academic performance of students, parents' and students' perceptions of school safety, and overall satisfaction with the school. The impact of the program on each of these outcomes would be measured by the difference in means between the intervention and control groups over the follow-up period.

Through random assignment, the two groups would be well matched at the point of entry into the program. Therefore, in this example, it is not essential to compare the *changes* in the two groups' outcomes over the follow-up period, as we did in the comparison group design described above. The simple intervention-control difference in postrandom assignment outcomes will, on average, give the same answer as the intervention-control difference in *changes* in outcomes pre-to post-random assignment, because there will be no systematic difference between the two groups at the point of random assignment. Taking the preprogram value of the outcome variable into account will, however, improve the quality or *precision* of the impact estimates.

Although this simple example illustrates the fundamental elements of a social experiment, the range of possible variations on this basic theme is enormous. For example, instead of evaluating a single program, one might wish to compare several different program models or estimate the effects of specific program components. Alternatively, instead of simply estimating the impacts of a particular school voucher program, one might want to study the impacts of programs that offer high and low voucher amounts. Or, instead of studying the effects of the program on applicants who voluntarily apply to the school voucher program, one might wish to study the effects of the program on the entire population eligible for the program in order to see who among those eligible use the vouchers, as well as estimate the average impact across all eligible students.

Designing a Social Experiment

The experimental design consists of specification of the experimental conditions to which subjects are to be randomly assigned and the numbers to be assigned to each condition. The choice of experimental conditions is driven by the policy question to be answered. If a single intervention (new or existing) is to be evaluated, the design will normally consist of a single treatment group, which receives that intervention, and a control group that does not. Testing multiple interventions requires multiple treatment groups, although a single control group will normally suffice as the counterfactual for each treatment group. Impact estimates based on comparison of a treatment group and the control group are termed *net impacts*; they measure the effect of the experimental intervention relative to the status quo. Comparison of the outcomes of one intervention to those of another yields *differential impacts*, the impact of one intervention relative to another. The ability to use a single control group for multiple interventions and the ability to estimate differential impacts are strong reasons to include multiple interventions in the same experiment whenever multiple interventions are relevant for policy.

The proportions of sample assigned to different experimental conditions will affect the precision with which different impacts are measured – i.e., the standard error of the estimate. In general, the smaller the experimental groups being compared, the less precise will be the estimate (i.e., the larger its standard error). Equal allocation to each group will provide equally precise estimates of the net and differential impacts of each intervention tested.

The absolute precision of the estimates will depend not just on the proportional allocation of the sample but on the absolute number of subjects assigned to each experimental condition. For each impact of interest, one can – and

should – compute *ex ante* the *minimum detectable effect, or MDE* – the smallest true effect that can be detected by the design at a specified level of statistical significance and statistical power. ('Statistical significance' measures the risk of concluding that the intervention had an effect when it did not, sometimes termed a Type 1 error. For example, testing impacts at the 0.05 significance level entails a 5% risk of accepting as real a treatment-control difference that occurred by chance, as a result of random assignment. 'Power' measures the likelihood that the design will detect a true effect of a given size. For example, 80% power to detect a $100 increase in earnings means that if the *true* effect of the intervention is to raise earnings by $100, then in repeated sampling and random assignment, we would detect a statistically significant effect 80% of the time.) Small MDEs help assure that impacts that would be relevant for policy are not missed by the experiment just because the sample is too small to detect them. A simplified formula for the MDE is

$$\begin{aligned}\mathrm{MDE} &= k(\alpha, B)\mathrm{SEE}(\widehat{\beta}) \\ &= k(\alpha, B)\sqrt{\frac{\sigma_Y^2}{n_\mathrm{T}} + \frac{\sigma_Y^2}{n_\mathrm{C}}}\end{aligned}$$

where

α = significance level
B = power
$\widehat{\beta}$ = estimated impact
$\mathrm{SEE}(\widehat{\beta})$ = standard error of the impact estimate
σ_Y^2 = variance of outcome Y
n_T, n_C = sample size of treatment and control groups

As can be seen from the formula, the MDE depends critically on the variance of the outcome and the sample sizes of the two groups being compared (in this case, a treatment group and a control group, although it could equally well be two treatment groups). Lower variance outcomes and larger sample sizes yield smaller (more precise) MDEs. The multiplicative factor $k(\alpha, B)$ is a complex function of the significance level (α) and power (B) chosen by the researcher. Lower levels of power (e.g., 70% rather than 80%) reduce the MDE – but at the cost of a higher risk of missing a true effect. Lower levels of significance (e.g., the 0.10 level rather than the 0.05 level) also reduce the MDE – but at increased risk of accepting a chance difference as real.

Analyzing the Experimental Data

Experimental impact estimates are fundamentally comparisons of outcomes between two experimental groups (e.g., treatment and control) at some point after random assignment. Differences in raw means between the treatment and control group are unbiased estimates of the effect of the intervention. More precise (and still unbiased) estimates can be obtained by controlling for chance treatment-control differences in individual characteristics by regressing the outcome on a treatment indicator and a set of covariates X_j measured at baseline:

$$Y_i = a + b_j X_{ji} + cT_i + \varepsilon_i \qquad [1]$$

where Y_i is the outcome for the ith sample member, T_i is a dichotomous treatment indicator for the ith sample member (normally, treatment group = 1), ε_i is a random error term, and a, b_j, and c are coefficients to be estimated. The coefficient of T_i, c, is an estimate of the impact of the experimental intervention on the outcome Y (i.e., the treatment-control difference, controlling for X_{ji}).

The variance of the impact estimate is

$$V(c) = V(\varepsilon_i)\left[\frac{1}{n_\mathrm{t}} + \frac{1}{n_\mathrm{c}}\right] = V(Y_i)(1 - R^2)\left[\frac{1}{n_\mathrm{t}} + \frac{1}{n_\mathrm{c}}\right] \qquad [2]$$

where R^2 is the multiple correlation coefficient for eqn [1], n_t and n_c are the numbers of sample members in the treatment and control groups, respectively, and $V(Y_i)$ and $V(\varepsilon_i)$ are the variances of the outcome and the error term in eqn [1], respectively. As can be seen from this formula, explaining the variance of the outcome is equivalent to an increase in sample size. For example, an R^2 of 0.20 is equivalent to a 25% increase in sample size.

Any baseline characteristics can be used as a covariate. The best covariate is usually the baseline value of the outcome (if it exists), because it captures all of the unmeasured influences on the outcome that do not vary over time. *No* variable that measures behavior or circumstances after random assignment may be used as a covariate because such variables may be affected by random assignment. If so, they may capture some of the treatment–control difference caused by the intervention and result in a biased impact estimate. Thus, for example, it is illegitimate to include measures of the amount of treatment the participant received ('dosage').

A second set of restrictions on impact estimation is that outcomes for both treatment and control groups must be measured in the same way and at the same point in time *relative to the date of random assignment*. Differences between groups in either of these regards may create differences in measured outcomes that are not attributable to the intervention and therefore yield biased impact estimates.

Researchers typically use a t-test based on an estimate of the variance of c to determine whether the estimated impact is larger than could be expected by chance alone. (Note that the expression for the variance of the impact estimate in eqn [2] relies on the population value of the variance of the outcome. Researchers use the sample estimate of that variance to obtain an estimate of the variance of the impact estimate.) Typical significance levels used in social experiments are the 0.10 and 0.05 levels (two-tail test).

Researchers typically estimate impacts on multiple outcomes and conduct 'subgroup analyses' in which they define multiple mutually exclusive, exhaustive subgroups of the sample and estimate impacts on each. This approach creates a 'multiple comparisons' problem. The standard t-test used in such analyses is based on the assumption that the test is performed a single time. When multiple tests are performed, the chance that *at least one* of them will indicate statistical significance by chance alone is higher than the nominal significance level – and if many tests are performed, as they often are, the chance of such a Type I error may be very high. There is at present little agreement within the research community on how this issue should be addressed. An approach that seems to be gaining favor is to separate the estimates into one or

more 'confirmatory' tests of the main impacts that matter for policy and a set of 'exploratory' tests (usually including all subgroup analyses) that are intended to inform the interpretation of the confirmatory tests and/or suggest opportunities for further experimentation. Other suggestions for alleviating the multiple comparisons problem include limiting the confirmatory tests to a small number of central outcomes, using an index of outcomes for the confirmatory test, and/or adjusting the test to take account of the number of impacts estimated. (For the best existing statement of this issue, see Schochet (2008).)

Internal Validity and External Validity of Experimental Estimates

Evaluation methods that provide unbiased estimates of the impacts of a program, policy, or practice *on the population to which they were applied* are termed *internally valid*. Properly implemented social experiments provide internally valid impact estimates. To be most useful for policy purposes, however, impact estimates must have both internal validity and external validity. '*Externally valid*' estimators provide unbiased measures of the impacts of the intervention *on the policy-relevant population* – i.e., the population that would be affected by the policy decision that the evaluation is intended to inform. (The requirement that the estimated effect on the larger population be unbiased implies, in general, that to be externally valid an estimator must also be internally valid.) That is, only if the experiment faithfully implements the program, policy, or practice for a sample that is representative of the policy-relevant population will the findings have external validity and serve as a reliable guide for making decisions about broader adoption of the treatment studied.

The external validity of the study findings can be compromised in a number of ways. For example, for reasons of cost and logistics, experimental samples are usually clustered in a small number of locations – for example, a few schools, welfare offices, housing authorities, or health clinics. This makes it difficult to draw a sample that is truly representative of the policy-relevant population. In fact, experiments are frequently conducted in only one or two localities that cannot claim to be representative of the nation, state, or other jurisdiction for which policy is made. (See Olsen et al. (2013) for a derivation of the bias that may occur when experiments are conducted with nonrepresentative populations.)

In the end, true external validity is an ideal that is almost impossible to attain, if only because the continually evolving policy process represents a moving target. Nevertheless, it is an important ideal to strive for, and in assessing the strengths and weaknesses of alternative evaluation methods or results, it is important to gauge both the external and internal validity of the findings.

Is Experimentation Ethical?

Because they are the only known way to be confident of obtaining unbiased estimates of the impacts of interventions (that is, of avoiding the risk of selection bias), experimental-design studies have become the preferred method of program evaluation among most of the policy research community and among a growing number of policymakers (Baron, 2005; Cook, 2007; Cook and Foray, 2007; Boruch and Mosteller, 2002). But such designs are also subject to a widespread concern, especially among program practitioners: Do they violate ethical standards by denying program services to the control group?

It should be noted at the outset that posing the question this way *presumes* that the program services are beneficial. This is not necessarily the case. Perhaps the most common outcome of careful impact evaluations is a finding of no significant effect of the treatment; in a few instances, Peter Rossi famously expressed this as the Iron Law of Evaluation: "the expected value of any net impact assessment of any large scale social program is zero" (Rossi, 1987). Although this is probably an overly pessimistic assessment, it is true that many social interventions, when carefully evaluated, have not shown evidence of effectiveness. See Orr et al. (1996) and Burtless (1985) for examples of social policy interventions that were found to be outright harmful. In medicine, a number of relatively standard treatments have been found to be harmful when subjected to randomized clinical trials, as in the case of estrogen therapy for postmenopausal women (Women's Health Initiative Steering Committee, 2004). In fact, the very existence of the evaluation is evidence that the agency that funded it is uncertain about the value of the intervention. Thus, the fact that the control group is denied program services does not automatically mean that the control group is disadvantaged by the study.

Nevertheless, it is important to know whether controls are in fact denied services in any meaningful sense. A full discussion of the ethics of social experiments is beyond the scope of this article. (For a more detailed discussion of the ethics of social experiments, see Gramlich and Orr (1975) and Orr (1999).) Here, we simply note that if the supply of services is limited, as in a special demonstration or a budget-constrained program, random assignment can be viewed as simply one way to ration scarce resources. In such cases, in the aggregate, the same number of people will be served whether the fixed supply of services is allocated on a first-come-first-served basis, through staff judgment, or by random assignment of applicants from the larger eligible population. Indeed, random assignment is arguably fairer than other allocation methods because it gives every eligible applicant the same chance of receiving services. Only when all eligible individuals would receive services in the absence of the experiment would random assignment to a control group constitute denial of service in the aggregate. Random assignment to a no-service control group is generally viewed as inadmissible in the context of such 'entitlement' programs.

In considering whether experiments are ethical, it is important not to focus exclusively on the issue of denial of the focal treatment to controls. Other members of society have a stake in whether the experiment is performed. In particular, failure to obtain reliable estimates of the efficacy of an ongoing program, policy, or practice can impose substantial costs on taxpayers. Moreover, failure to identify programs that are ineffective implicitly imposes costs on their intended beneficiaries. Not only do ineffective strategies waste time and create false expectations, but they also consume resources that might otherwise be devoted to more effective ways to address the needs of the

target population. Thus, 'protecting' intended beneficiaries of new initiatives from experiments is not necessarily in their best interest.

A Brief History of Social Experimentation

Dehue (2005) credits the statistician and psychologist Sir Francis Galton with originating the concept of determining effects through statistical comparison of a treated group with an untreated group – essentially, the comparison group design (Galton, 1872). However, the principle of casting of lots to generate comparison groups in assessing the effects of treatments had been suggested over 200 years earlier by the Flemish physician Van Helmont (1662). And in 1835, Pierre Louis clearly proposed the use of comparison groups in medical research, suggested the use of random ('indiscriminate') assignment to treatment, and showed that he understood the relationship between sample size and the precision of the estimated difference in outcomes (Louis, 1835, pp. 75–76; Louis, 1836, pp. 59–60).

It was only after the turn of the twentieth century, however, that researchers began using comparison groups to test the efficacy of alternative social interventions. Led by the educational psychologists, researchers began using treatment and comparison groups (often entire classes of students) to test the effects of classroom circumstances such as fresh versus ventilated air, the gender of the teacher, memorizing methods, and educational measures such as punishing and praising (Dehue, 2005).

As researchers gained experience with the comparison group method, however, they became increasingly concerned – and their critics became increasingly aware – that the measured differences between treatment and comparison groups might reflect more than just the effect of the treatment; they might also include preexisting differences between the two groups. To protect against this form of bias, it became customary to obtain baseline measures of traits, such as intelligence, that might affect the outcomes and then to attempt to match the two groups on these traits. It soon became evident that such matching procedures were not only time-consuming and expensive, but were also far from foolproof; to be sure of an unbiased estimate of the treatment effects, the researcher had to correctly identify and measure *all* the characteristics on which the two groups might differ that could affect the outcomes of interest.

The power of random assignment to eliminate this source of bias by establishing comparable groups was recognized by William A. McCall in his 1923 book *How to Experiment in Education*. Again, the influence of psychology on education research was evident: McCall was a student of, and collaborator with, the renowned psychologist Edward L. Thorndike. The earliest study we have found in which a randomized design was actually applied to form a control group to assess the impact of an intervention was J.E. Walters (1931) study of using college seniors as counselors for 'delinquent' freshmen (Forsetlund et al., 2007, p. 377). (The authors describe the Walters study as the second to use random assignment. However, the earlier study that they identify (Remmers, 1928) actually ranked students by test scores and used alternate names on the list to form the treatment and control groups. Because this is not a random process, it does not meet our definition of random assignment. In fact, alternation had been used over 30 years earlier in a study of serum treatment for diphtheria, in which patients admitted to a hospital on alternate days formed the treatment and control groups (Fibiger, 1898; discussed in Hrobjartsson et al., 1998).)

Over the decade 1925–35, the great statistician R.A. Fisher laid the statistical foundations of experimentation with random assignment in his seminal books *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). Mann (1949) provided the rigorous mathematical basis for many of the methods described in Fisher's seminal books.

In the 1930s and 1940s, random assignment gained increasing favor among educational psychologists as the ideal method for comparing different teaching methods, although the number of randomized experiments was still small. (In an 'extensive search,' Dehue (2005) found one randomized experiment in the 1930s and six in the 1940s. Forsetlund et al. (2007) found three randomized trials in the 1930s and six in the 1940s prior to the well-known randomized controlled trial of streptomycin for pulmonary tuberculosis, which is widely regarded as the first clinical trial in medicine (British Medical Research Council, 1948).) By the 1950s, statistical procedures for the design and analysis of relatively complex experiments were firmly established by researchers and statisticians (see, for example, Mann, 1949; Cochran and Cox, 1950; Lindquist, 1953). Campbell and Stanley's (1963) classic, *Experimental and Quasi-Experimental Designs for Research on Teaching*, clearly laid out all of the issues we have discussed so far (published in 1966 as *Experimental and Quasi-Experimental Designs for Research*).

Application of Experimental Methods to Social Programs, 1960–2000

By the 1960s, the use of the experimental model was sufficiently widespread in education, psychology, and medicine that it was quite natural to apply it to social programs and policies outside those fields. In 1961–64, for example, the Manhattan Bail Bond project used random assignment to test the proposition that many arrestees could successfully be released without bail prior to trial (Botein, 1965). Other randomized tests of law-related programs and procedures undertaken in the 1960s included studies of a variety of approaches to the prevention and treatment of juvenile delinquency; the use of legal counsel in juvenile court; the effects of pretrial conferences; low-stress versus high-stress training for police; alternative penalties for drunk driving; and vocational, surgical, and social rehabilitation for former prisoners (Riecken and Boruch, 1974).

These early applications to social policy received little attention outside the immediate circles of the researchers and funding agencies involved. In the late 1960s and early 1970s, however, a series of four experiments in income maintenance (that is, cash transfers to low-income families) attracted much greater attention, both within the research community and among policy makers. (For an overview and critique of the income maintenance experiments, see Rossi and Lyall (1978).) Launched by the U.S. Office of Economic Opportunity and the Department of Health, Education, and Welfare (HEW),

these studies were novel in several respects. First, their samples were much larger than most of the randomized trials that had gone before, ranging from about 1300 families to nearly 5000 families. More importantly, these were *field* experiments: they involved administering carefully controlled treatments to, and observing the behavior of, these large samples of individuals in the course of their daily lives, not in classrooms, hospitals, psychology laboratories, or other institutional settings. In addition, the question addressed by the experiments – whether receipt of cash transfers (welfare) would cause poor families to stop working – was a highly visible, politically charged issue. These were perhaps the first experimental tests of an issue that national policymakers really cared about.

These highly visible studies of a politically controversial issue prompted other federal agencies to adopt this novel technique in the early 1970s. Large-scale experimental tests of housing vouchers by the Department of Housing and Urban Development, and alternative health insurance plans by the Department of Health, Education, and Welfare (HEW), were explicitly patterned on the income maintenance experiments (Kennedy, 1980; Newhouse, 1993).

In the mid-to-late 1970s and early 1980s, experiments were conducted by the U.S. Department of Labor and other funding agencies to test alternative employment and training services for unemployed workers, welfare recipients, disadvantaged youths, ex-offenders, former substance abusers, and individuals with mild to moderate retardation. In 1981, when federal legislation gave states the option of testing new approaches to the employment of welfare recipients under federally approved waivers of their statewide plans, the Department of Health and Human Services (HHS, formerly HEW) took the opportunity to require that such projects be evaluated using random assignment of eligible program applicants to treatment or control (services-as-usual) condition. This led to 15 years of waiver experiments, leading to the passage of major welfare reform laws in 1986 and 1996. (See HR 3734, 'Personal Responsibility and Work Opportunity Reconciliation Act of 1996,' 9 August 1996. The findings of many of these experiments are summarized in Greenberg et al. (2009).) According to Greenberg and Shroder (1991, 1997, 2004), by 1980, 55 social experiments had been initiated in fields other than medicine and education; by 1995, 195 such experiments had been conducted; and by 2004, at least 273 had been initiated. (Greenberg and Shroder define social experiments as "field studies of social programs in the United States in which there was random assignment of individuals or families to alternative treatments and an emphasis on the measurement of impacts on either market behavior, the receipt of earnings, or transfer payments." Thus, they explicitly do not include in their compendium experiments involving interventions such as drug treatment, medical care, or education programs.)

Application of Experimental Methods to Social Programs, 2001–14

The broad acceptance of social experimentation as the 'gold standard' of program evaluation in the 1980s and 1990s took the form of a number of individual experiments being approved and conducted largely in isolation of one another. In the 2000s, experimentation began to become more institutionalized.

With the institution of the Program Assessment Rating Tool (PART) in 2002, the powerful Office of Management and Budget (OMB) began not only to require that federal agencies systematically evaluate the effectiveness of their programs, but to do so with random assignment wherever feasible. Although PART was discontinued at the end of the Bush administration, OMB under President Obama has continued to press agencies to provide hard evidence of program effectiveness, in the form of randomized trials wherever possible. (See, for example, Zients (2012), a memorandum from the Director of OMB to agency heads, requiring agencies to include in their budget submissions discussion of innovative evaluation approaches and offering technical support and resources to assist them in implementing rigorous evaluations, including random assignment.)

A second important impetus to experimentation came with the establishment of the Institute of Education Sciences (IES), also in 2002. IES's first director, Grover (Russ) Whitehurst, aggressively pursued randomized trials of educational interventions, both existing and proposed. In a field that had seen few social experiments in the previous 40 years, IES launched more than 30 in its first 10 years, through its National Center for Education Evaluation and its network of 10 Regional Education Labs. In fact, virtually all of the major studies initiated by IES during that period were randomized trials. (As noted above, education researchers pioneered the use of random assignment in the 1920s and 1930s. However, in the latter part of the twentieth century, the field largely abandoned the method until it was revived at IES.)

Under the Obama administration, the link between social experimentation and policy has been strengthened by a set of 'evidence-based policy initiatives' – appropriations reserved for funding programs that have been proven effective in well-designed randomized trials. Administered by the Corporation for National and Community Service and the Departments of Health and Human Services, Education, and Labor, these appropriations totaled $1.3 billion in fiscal year 2011 (Haskins and Baron, 2011). The funds are allocated to support evidence-based programs in education, teen pregnancy prevention, home visiting, innovations in community service, and workforce interventions.

See also: Causal Analysis, History of; Ethical Questions in Social and Behavioral Sciences, History of; Experimental Design: Randomization and Social Experiments; Experimentation in Psychology, History of; Field Experiments; Psychology, History of (Early Period); Quasi- and Field Experiment; Quasi-Experimental Designs.

Bibliography

Ashenfelter, Orley, February 1978. Estimating the effect of training programs on earnings. Review of Economics and Statistics 61, 47–57.

Baron, Jon, 2005. Key Items to Get Right when Conducting a Randomized Controlled Trial in Education. Coalition for Evidence Based Policy, Washington, DC.

Boruch, Robert F., Mosteller, Frederick (Eds.), 2002. Evidence Matters: Randomized Trials in Education Research. Brookings Institution Press, Washington, DC.

Botein, B., 1965. The Manhattan Bail Project: its impact in criminology and the criminal law process. Texas Law Review 43, 319–331.

British Medical Research Council, 1948. Streptomycin treatment of pulmonary tuberculosis. British Medical Journal 2, 769–782.

Burtless, Gary, 1985. Are targeted wage subsidies harmful? Evidence from a wage voucher experiment. Industrial and Labor Relations Review 39, 105–114.

Campbell, Donald T., Stanley, Julian C., 1963. Experimental and Quasi-experimental Designs for Research. Rand-McNally, Chicago.

Cochran, William G., Cox, Gertrude M., 1950. Experimental Designs. John Wiley and Sons, New York.

Cook, Thomas D., Foray, Dominique, 2007. Building the capacity to Experiment in schools: a case study of the institute of education sciences in the department of education. Economics of Innovation and New Technology 16 (5), 385–402.

Dehue, Trudy, 2005. History of the control group. In: Everitt, Brian S., Howell, David C. (Eds.), Encyclopedia of Statistics in Behavioral Science, vol. 2. John Wiley & Sons, Ltd, Chichester, pp. 829–836.

Fibiger, J., 1898. Om Serumbehandling af Difteri. Hospitalstidende 6, 309–325, 337–350.

Forsetlund, Louise, Chalmers, Iain, Bjørndal, Arild, 2007. When was random allocation first used to generate comparison groups in experiments to assess the effects of social interventions? Economics of Innovation and New Technology 16 (5), 371–384.

Fisher, Ronald A., 1925. Statistical Methods for Research Workers, first ed. Oliver and Boyd, London.

Fisher, Ronald A., 1935. The Design of Experiments. Oliver and Boyd, London.

Galton, Francis, 1872. Statistical inquiries into the efficacy of prayer. Fortnightly Review XII, 124–135.

Gramlich, Edward M., Orr, Larry L., 1975. "The Ethics of Large Scale Social Experimentation", Ethical and Legal Issues of Social Experimentation. The Brookings Institution, Washington, DC.

Greenberg, David, Shroder, Mark, 1991. The Digest of Social Experiments, first ed. The Urban Institute Press, Washington, DC.

Greenberg, David, Shroder, Mark, 1997. The Digest of Social Experiments, second ed. The Urban Institute Press, Washington, DC.

Greenberg, David, Shroder, Mark, 2004. The Digest of Social Experiments, third ed. The Urban Institute Press, Washington, DC.

Greenberg, David, Deitch, Victoria, Hamilton, Gayle, February 2009. Welfare-to-Work Program Benefits and Costs. MDRC, New York.

Hamilton, William L., Burstein, Nancy R., Long, David, February 1998. Using Incentives in Welfare Reform: The New York State Child Assistance Program. Abt Associates Inc., Cambridge, MA.

Haskins, Ron, Baron, Jon, 2011. Building the Connection between Policy and Evidence: The Obama Evidence-Based Policy Initiatives. NESTA, London.

Hróbjartsson, A., Gøtzsche, P.C., Gluud, C., 1998. The controlled clinical trial turns 100 years: Fibiger's trial of serum treatment of diphtheria. British Medical Journal 317, 1243–1245.

Kennedy, Stephen D., 1980. Final Report of the Housing Allowance Demand Experiment. Abt Associates Inc, Cambridge, MA. Available at: http://www.huduser.org/portal/publications/affhsg/hadefinalreport.html.

Lindquist, E.F., 1953. Design and Analysis of Experiments in Psychology and Education. Houghton Mifflin Co., the Riverside Press, Cambridge, MA.

Louis, Pierre-Charles-Alexandre, 1835. Recherches sur les effets de la saignée dans quelques maladies inflammatoires et sur l'action de l'émétique et des vésicatoires dans la pneumonie. Librairie de l'Académie royale de médecine, Paris.

Louis, Pierre-Charles-Alexandre, 1836. Research on the Effects of Bloodletting in Some Inflammatory Diseases. Hilliard, Gray and Company, Boston.

Mann, H.B., 1949. Analysis and Design of Experiments: Analysis of Variance and Analysis of Variance Designs. Dover Publications, Inc, New York, NY, pp. x–195.

Newhouse, Joseph, 1993. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press, Cambridge, MA.

Olsen, Robert B., Orr, Larry L., Bell, Stephen H., Stuart, Elizabeth A., Winter 2013. External validity in policy evaluations that choose sites purposively. Journal of Public Policy Analysis and Management 32 (1), 107–121.

Orr, Larry L., 1999. Social Experiments: Evaluating Public Programs with Experimental Methods. Sage Publications, Inc, Thousand Oaks, CA.

Orr, Larry L., Bloom, Howard S., Bell, Stephen H., Doolittle, Fred, Lin, Winston, Cave, George, 1996. Does Job Training for the Disadvantaged Work? Evidence from the National JTPA Study. Urban Institute Press, Washington, DC.

Remmers, H.H., 1928. A diagnostic and remedial study of potentially and actually failing students at Purdue University. Bulletin of Purdue University: Studies in Higher Education, IX 28 (12).

Riecken, Henry W., Boruch, Robert F. (Eds.), 1974. Social Experimentation: A Method for Planning and Evaluating Social Intervention. Academic Press, New York.

Rossi, Peter, 1987. The iron law of evaluation and other metallic rules. Research in Social Problems & Public Policy 4, 3–20.

Rossi, Peter H., Lyall, Katharine C., 1978. An overview evaluation of the NIT experiment. Evaluation Studies Review Annual 3, 412–428.

Schochet, Peter Z., 2008. Guidelines for Multiple Testing in Impact Evaluations. Technical Methods Report NCEE 2008–4018. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation, Washington, DC.

Van Helmont, J.A., 1662. Oriatrike, or Physick Refined: The Common Errors Therein Refuted and the Whole Are Reformed and Rectified. Lodowick-Loyd, London, p. 526.

Walters, J.E., 1931. Seniors as counselors. Journal of Higher Education 2, 446–448.

Wolf, Patrick, Gutmann, Babette, Puma, Michael, Kisida, Brian, Rizzo, Lou, Eissa, Nada, Silverberg, Marsha, 2008. Evaluation of the DC Opportunity Scholarship Program: Impacts after Two Years. NCEE, Washington, DC.

Women's Health Initiative Steering Committee, 2004. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the women's health initiative randomized controlled trial. Journal of the American Medical Association 291, 1701–1712.

Zients, Jeffrey D., May 18, 2012. Memorandum to the Heads of Executive Departments and Agencies: Use of Evidence and Evaluation in the 2014 Budget. Office of Management and Budget, Washington, DC.

Social Geography

Sophie Bowlby, University of Loughborough, Loughborough, UK; and University of Reading, Reading, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the growth of and change in social geography since the 1960s. Early preoccupation with positivism and statistical description gave way, first to a behavioral approach and subsequently to a variety of 'radical' approaches. These attempted to challenge the inequalities and injustices of contemporary society. The article examines the subsequent impact of feminist and postmodern ideas and the associated growth of cultural geography on social geography. Over the last 20 years, social geography has had an uneasy relationship with cultural geography and has lost status in the USA. Recent publications concerned with social geography have argued for the importance of some of its traditional concerns and for radical changes in its approach.

Introduction – When and What

Social geography emerged as a significant subdiscipline during the 1960s as part of exciting theoretical and methodological change that geography experienced during that decade. In the 1970s and 1980s, vigorous theoretical debates led to a variety of conflicting approaches to research in social geography. By the early 1980s, the subdiscipline was of major importance within human geography. Indeed, it was claimed that it became "virtually synonymous with the whole field of human geography" (Johnston et al., 1994: p. 562). But the growing prominence of cultural geography since the end of the 1980s has reduced the prominence of social geography.

The rapid growth of the subdiscipline and the significant theoretical debates within it involved shifts in its definition and practice. In 1965, Pahl suggested that it was about "the theoretical location of social groups and social characteristics, often within an urban setting" (p. 82). In 1975, Emrys Jones argued that "social geography involves the understanding of the patterns, which arise from the use social groups make of space as they see it, and of the processes involved in making and changing such patterns" (p. 7). The emphasis on social groups evident in Pahl's earlier definition remains of particular importance, in that what is held to make social geography 'social' is its concern not with individuals but with people as members of groups. However, there are some significant changes between the two definitions. First, Jones' definition introduces the idea that social processes produce spatial patterns and vice versa. This understanding of social geography remained dominant into the 1980s. Second, Jones' phrase "the use social groups make of space *as they see it*" (author's emphasis) signals an increasing interest in people's subjective experiences of space and place – an interest that has become increasingly prominent.

The understanding of the scope and aims of social geography implied in Jones' definition endured until the mid-1980s when it came under increasing attack. These attacks and their implications for the development of social geography make better sense in the context of an account of the type of research that social geographers did from the 1960s to the mid-1980s, presented below.

Engaging with Positivism

Describing the Social Characteristics of Urban Areas

A major focus of research during the late 1950s and 1960s was the statistical description and analysis of the characteristics of the populations of urban areas. Central to this project was the identification of systematic patterns in the location of different social groups. Analysts examined which groups were likely to be segregated from other groups and whether there was evidence for particular patterns of location for different groups within cities.

At this time, there were also moves within human geography to apply positivist philosophy and methodology to the pursuit of geographical knowledge. These moves encouraged, on the one hand, attention to clearly specified theoretical statements and predictions and, on the other, the use of quantitative data and statistical techniques that were considered to ensure rigorous, scientific testing of hypotheses. Also associated with the growing commitment to positivism was a desire for predictive models of the location of human activity and people's 'spatial behavior,' often drawing on neoclassical economics. Geographers interested in social groups and in cities sought to develop predictive theories of the spatial form of cities, including theories of the location of social groups within urban areas.

Geographers looking for theoretical and methodological inspiration turned to the work of urban sociologists. First, they drew upon the work of the Chicago School – including Louis Wirth, Robert Park, and Ernest Burgess – who carried out extensive studies of Chicago in the first half of the twentieth century. They had used maps of a wide range of statistical information in order to identify the 'natural areas' – residential areas with distinctive physical and cultural characteristics – that the Chicago School sociologists argued developed from unplanned processes of competition and cooperation among urban residents. Burgess had developed a general model of urban residential land use as a series of concentric zones focused around the Central Business District.

Second, urban geographers were drawn to the work of Eshref Shevky and Wendell Bell who studied characteristics of the populations of Los Angeles and San Francisco. Shevky and Bell (1955) argued that the 'increasing scale' of society resulting

from the increasing complexity of economic organization and the associated growth of urban areas could be linked to three dimensions of urban residential differentiation: 'social rank' or 'economic status'; 'urbanization' or 'family status'; and 'segregation' or 'ethnic status.' A key feature of their work was the development of 'social area analysis' using data on census tracts. They selected variables to represent the three dimensions and then used these to categorize census tracts in terms of 'high' or 'low' economic, family, or ethnic status. Although the theoretical links between increasing societal scale and the three dimensions: economic, family, and ethnic status were poorly specified and strongly criticized, geographers enthusiastically embraced the use of census data to characterize areas of the city.

The advent of powerful computers in the 1960s allowed geographers to use sophisticated statistical techniques to analyze data on census tracts. Factorial ecology was particularly popular – the use of factor analysis or principal component analysis to identify a few general constructs that underlie a large number of variables. Shevky and Bell's three status dimensions often emerged from the census data. The analyst could also classify census tracts in terms of the underlying constructs and map the results to see whether any systematic spatial patterns in characteristics of census tracts could be identified. These maps were compared with Burgess' concentric model of urban form and comparisons between factorial ecologies of different cities also made. Similar techniques were also used to look for relationships between residential differentiation, location, and patterns of crime, deviance, and mental health.

The availability of census data for many cities (mostly in the 'developed' world) and increasing access to computers allowed social geographers to indulge in an orgy of analysis. However, there were some clear limitations to this work. First, the nature of the constructs identified was limited by the available census data: they could be as much a reflection of the data available as of the social 'reality' of urban areas. Second, because the data related to areas rather than individuals, false associations between variables may have been made (the 'ecological fallacy'). Moreover, differences in the sizes and rules for creating census tracts between countries limited the comparability of the results. Third, although the practitioners of these analyses aspired to work within a positivist framework much of the research was inductive description rather than tests of clearly specified hypotheses linked to well-identified theoretical propositions. Nevertheless, this body of work provided a considerable amount of descriptive information on residential differentiation that has become part of the taken-for-granted knowledge of geography.

The Behavioral Challenge

The descriptive analysis of urban residential characteristics was not the only type of work in social geography done during the 1960s and early 1970s. There were those who, despite their hopes for a positivist approach to the development of theory within geography, were dissatisfied with the widely used approaches to studying human behavior.

In the 1960s, much research in human geography drew its inspiration from models, which posited that people possessed perfect information and were rational and optimizing in their behavior. Dissatisfaction with these assumptions prompted some researchers to turn to psychological theory concerned with cognition and decision making in order to investigate what information people had available to use in spatial decision making and how they used it. The principal method employed in these studies was the questionnaire survey, frequently incorporating some variety of attitudinal and personality test drawn from social psychology. The research carried out within this broad framework is known as 'behavioral geography.'

Work in behavioral geography was concerned with the locational decision making of firms, individuals' decisions over where to live, people's everyday movements, and their information and ideas about space and place. Gould's work on mental maps was particularly influential (Gould and White, 1974). The idea of an incomplete, biased, and personal vision of the spatial world conjured up by the term 'mental map' became part of geographers' ideas about how people made decisions about both everyday movement and migration. Research on how people searched for information in space and developed strategies for choosing locations was also recognized as important in studies of residential mobility and of people's everyday travel behavior. In the former, researchers were influenced by the early work of sociologist Peter Rossi whose 1955 book *Why Families Move* was subtitled *A Study in the Social Psychology of Residential Mobility*. A large number of studies of residential mobility in different cities and countries resulted. In the latter, diary data on travel behavior were used to develop inductive theories about the way people choose and order travel destinations.

Research using the perspectives of behavioral geography – while part of social geography – remains marginal to the field because of its focus on the individual rather than on people as interacting with and partly constituted by their social milieu. However, the importance of people's social milieu is evident in much of the work actually carried out. Researchers often examined contrasts between individuals drawn from different social groups (for example, different income, class, age and, occasionally, gender groups). Others examined the influence of the resources available to individuals on their spatial cognition and so ended up investigating differences between social groups. The research is also significant to social geography because criticisms of its individualistic approach were important to the development of alternatives.

'Radical,' Politicized Approaches

During the early 1970s, many geographers became dissatisfied not only with the philosophical approach of positivism, but also with the lack of critical political engagement that they detected in contemporary research. These concerns grew out of the radical ideas promoted in America and Europe as part of opposition to the American war in Vietnam, antinuclear campaigns (especially in Britain and European countries), civil rights campaigns (particularly in America), and the growing 'rediscovery' of poverty in the prosperous countries of the West. What was needed, it was felt, was research that could suggest how to change society, not merely how to understand it. Positivism was criticized for suggesting that an objective approach to social science was necessary, feasible, and desirable

and that it was possible to find 'universal laws' of human society. Competing 'radical' approaches were proposed; they all shared the desire to challenge the inequalities and injustices of contemporary society although some sought this through changing our valuation of human beings, some through institutional change and some through revolutionary political action. A sketch of some of the main approaches follows. Although here they are separated into demarcated 'boxes,' in reality their boundaries were fuzzy.

Humanistic Geography

The humanistic geography that emerged in the 1970s was animated by a desire to present an alternative to positivism. It shared the behavioral geographers' critique of some positivist approaches as dehumanized – insofar as they did not engage with individuals' subjective experience. However, they added to this the more fundamental criticism that positivist social scientists made false claims to objectivity – by failing to recognize the subjectivity of both the observed and the observer. Here, humanistic geography made a decisive break with positivism and turned to approaches, which require empathy with people's own perception of the values and meanings that inform their action (Jackson and Smith, 1984). Two other important features of humanistic approaches were the insistence that people were active agents shaping their own worlds and a fascination with the links between place and social identity.

Although links have been made between the humanistic geography of the 1970s and a variety of early work in geography, the most immediate influence was the Chicago School. One of the major strands of work focused on understanding the way of life of particular social groups in particular areas of a city (although a few studies of rural areas were also carried out). It used ethnographic methods including participant observation, firsthand observation, and interviews as well as documentary information and census materials. Other researchers used ethnographic methods to explore features of crime and deviance in selected neighborhoods, often using factorial ecology to identify suitable areas for study.

Humanistic geographers also engaged with what were for 1970s geographers the new philosophical approaches of phenomenology and existentialism, with two important consequences. The first was a move toward thinking of the individual as constituted through the social interactions in which he/she takes part. These social interactions and people's everyday behaviors create a taken-for-granted world, which the social scientist can attempt to deconstruct. Second, there was an attempt to avoid prior theorization about the values and meanings of those people who are studied and to be open to understanding the world through their eyes. However, this still allows the findings of such studies to be interpreted within a theoretical or practical context and for disagreement about such interpretations to be fierce.

Another focus was on understanding the cultural meanings of, and values attached to, particular landscapes and built forms. This work employed the methods of hermeneutics (for example, Meinig, 1979). In the 1970s, this work was considered to be on the margins of social geography and soon became seen as part of the revival of cultural geography. Its investigation of the multiplicity and instability of meanings, their communication and reformulation began a tradition of work that has had important implications for contemporary social geography.

Welfare and Weber

David Smith's *Human Geography: A Welfare Approach* (1977) addressed what had begun to emerge as a new approach to human geography of particular significance to social geography. This approach focused on 'who gets what, where, and how' – that is, on issues of social justice and inequality. It therefore involved both arguments about what constitutes social (and often, territorial) justice and attempts to measure empirically various types of social inequality. An early focus was on developing social indicators to measure inequality between social groups living in different areas at a variety of scales ranging from the city to the world. This research was built on the expertise developed in earlier descriptive analyses of census data on cities. It also included measurement of people's spatial access to services (for example health care or education) or proximity to negative factors such as pollution. In order to explain the patterns of inequality or deprivation they identified, researchers developed analyses of the operation of agencies and agents whose actions seemed significant to the distribution of resources.

A substantial body of work was developed around state and private institutions' management of the distribution of resources drawing on the theories of Max Weber. The concept of 'managerialism' was first proposed by Pahl (1975), who suggested that within both the state and private sector there were 'gatekeepers' whose decisions were crucial to people's access to resources. Within social geography, this idea stimulated a wide range of studies of access to housing, emphasizing the constraints on people's choices in the housing market. The theme of choice and constraint was explored in access to health care, education, services for the elderly, parks, and open space and delivery services in rural areas. The ideas of Weber also introduced an increasing interest in conflict between and among institutions and groups. Here social geographers drew upon the ideas of the sociologist John Rex (1968) on housing classes, as well as on Weberian ideas about the operation of bureaucracies and professional groups, to explore conflicts over the distribution of resources.

A different approach to access used the ideas of time geography developed by Hagerstrand (1975). He suggested that time and space are resources people must use to realize the 'projects' they wish to carry out. Since traversing space requires time and many activities require the copresence of individuals or presence at a particular location, not all projects can be realized. People's ability to carry out particular projects is limited by capacity, coupling, and authority constraints. Despite the enthusiasm of some geographers and a few empirical studies, the potential of these ideas remained regrettably underdeveloped. Most empirical use of Hagerstrand's ideas focused on physical limits to access and did not develop the possibilities of a more fully realized social theory of human behavior in space and time.

Although the majority of welfare-focused research was concerned with cities in 'developed' countries, attention was

also paid to inequalities in the rural areas of such countries and to developing countries, Eastern Europe and the former USSR. The study of conflict and constraint often employed ethnographic methods to explore the values and patterns of behavior of gatekeepers and the social groups attempting to access resources. Attempts to answer the 'how' element of welfare geographers' problematic often drew on theories of political economy, which situated explanation of issues such as housing access within an analysis of the operation of capitalist society.

Marxist Approaches

The book that best illustrates the thinking that drew geographers to Marxist theory is Harvey's *Social Justice and the City* (1973), which explains the shift in Harvey's thinking from liberal to Marxist. By the end of the 1970s, Marxist approaches had expanded to create a new orthodoxy. With implications across all geography, Marxist approaches were embraced with particular enthusiasm in social geography.

The analysis of residential differentiation and change had been a central part of social geography since its inception. Marxist theory offered ways of understanding these processes through ideas of class struggle and capital accumulation. Large-scale patterns of investment and disinvestment in the built environment could be interpreted in terms of crises in processes of accumulation. Such analyses were applied to gentrification and suburbanization. Processes of socialization within residential areas could also be theorized as involving the reproduction of particular social subgroups (class fractions).

These early formulations were challenged by a body of socialist–feminist work that developed in the early 1980s. While incorporating many postulates of Marxism, this work criticized the conceptual and practical neglect of gendered power relations and the failure to identify and explore the links between gender and class relations. Critical empirical research on both gentrification and suburbanization sought to explore the ways in which gender relations were involved in the construction of the spatial structure of the city and the implications of these processes for women's oppression in society. In general, this socialist–feminist work was more sensitive than other Marxist approaches to the role of human agency and to humanist claims concerning the partiality of knowledge.

Analysis of access to services was also strongly affected by Marxist theory. Castells' work on collective consumption and social movements was highly influential. Castells (1976) defines collective consumption as the provision of services that are produced and managed collectively and distributed through nonmarket allocation processes. He argued that these were provided through the State because, while necessary to the reproduction of labor power, their provision is not profitable enough to warrant involvement of the capitalist market. Castells made a highly contentious definition of the city as "a unit of collective consumption corresponding more or less to the daily organization of a section of labor power" (1976: p. 148). More important to research in social geography, he spelled out how social movements might arise in response to inadequacies in the provision of collective consumption. Such social movements might fragment class solidarity and reduce the possibility of radical political change or enhance it. Castells' ideas stimulated extensive conceptual and empirical research in both urban sociology and in social geography.

Critiques and Cross Fertilization

By the early 1980s, social geographers had become better versed in social theory and more critical and sophisticated in its application. There were debates and cross fertilization of ideas among researchers working in the three areas identified above and increasing awareness of their shortcomings as well as their strengths. Humanists and feminists raised critiques of Marxist approaches' neglect of human agency, which encouraged geographers' interest in the sociologist Anthony Giddens' 'structuration' theory (Giddens, 1981). Giddens attempted to show how human agency and structures were mutually constitutive within a time-space framework that drew on Hagerstrand's ideas. These critiques encouraged attempts to develop a more 'humanistic' Marxism sensitive to both human agency and contingency. At the same time, those concerned with issues of welfare and social justice were increasingly using some form of political economy approach deriving from Marxism to frame their research. Feminist geographers criticized these approaches for neglecting gender, and a major new body of research by feminist social geographers began to develop. This work involved ethnographic empirical study of the everyday life of women and incorporated and reshaped ideas drawn from welfare geography (in relation to issues of access and inequality), humanistic geography (in relation to women's experience of and social relations within particular places), and Marxist geography (in relation to class and economic power). Some of these feminist works were linked to another major development within geography – the explosive growth of cultural geography during the late 1980s.

This new cultural geography drew on earlier humanistic social geographical research as well as on cultural studies and feminist and postmodern critiques of modernist theory. In particular, modernist theory was criticized for assuming a white, bourgeois, and masculine subject and for its attempts to discover universal truths. Furthermore, postmodern theory insisted that because meaning is constantly created and changed through shared but contested discourses, both individual and group definitions are unstable and contested. These critiques, and empirical studies in the 1980s and 1990s in cultural and economic geography, call into question earlier definitions of social geography and the coherence of the subdiscipline.

Fragmentation and New Theoretical Directions

The later 1980s and the 1990s saw a progressive destruction of the main features of Emrys Jones's (1975) definition of social geography. That definition involved separating social process and spatial pattern and focusing on the use of space and creation of spatial patterns by social groups.

Social Process and Spatial Pattern vs 'Spatiality'

The 1980s saw growing dissatisfaction with the idea of social processes as distinct from spatial patterns, on the grounds that,

because all social interactions necessarily take place in and through space, spatial pattern cannot be isolated from social process. These attacks derived from critiques of positivist analyses of abstract space as an empty container of social processes. Instead, Marxists argued that space was produced through capitalist social processes and through processes of class and other groups' opposition to capitalism. Humanistic and cultural geographers used phenomenological theory and empirical research to argue that understandings of space are produced through people's everyday social activities and do not draw on prior concepts of abstract space. The term 'spatiality' was coined to refer to the inescapable imbrication of space and social practices.

Defining 'The Social'

Feminist work in the 1980s showed that social geography's focus on issues of social reproduction – housing, education, health care, and crime – had been seriously incomplete. Domestic labor and the home had been ignored, as had the significance of gender relations. Moreover, this feminist work also argued that "production and reproduction – the conventionally accepted subjects of economic and social geography respectively – cannot analytically be separated" (Bowlby and McDowell, 1987: p. 305). During the 1990s, the importance of examining the social relations embedded in economic exchange was re-emphasized by a number of researchers. In particular, feminist research on labor markets showed clearly the interpenetration of the social and the economic (Hanson and Pratt, 1995).

The early work of social geographers focused on the implications of people's membership of social groups. The existence of defined social groups and the salience of characteristics such as ethnicity, income, and, much later, gender, were largely taken for granted. The theoretical critiques developed in cultural geography during the later 1980s and early 1990s drew attention to the fluidity, socially constructed, and contested nature of definitions of social groups and the significance of struggles over representations of identity. In the 1970s and early 1980s, despite fierce arguments over explanations for social inequality, the relevance of political economy theories was generally accepted. Postmodern approaches questioned the validity of 'grand theory' such as that of Marx and Weber that had previously been part of the stock in trade of social geographers. Cultural geographers' engagement with the theories of Foucault changed the ways in which 'power' was conceived. Developments in cultural geography have thus required social geographers to examine processes of representation and communication more closely, to revise their theories of the operation of social power, and to question the idea of a single 'social reality' to be uncovered through research.

Furthermore, the rise of interest in nonrepresentational theories in cultural geography from the mid-1990s posed a new challenge to social geographers' understanding of the 'social.' Nonrepresentational theories argued that researchers should go beyond a focus on representation and meaning to consider the complicated ways in which human interactions with each other, with the material world, and with nonhumans create new, emergent capacities to act. Bodily practices are understood to involve precognitive affects created in interactions that are not dependent upon the understanding of the rational subject. Such approaches called into question social geographers' focus on analyzing people's accounts of their experiences and practices as members of social groups and prioritized study of affects and everyday, mundane "doings" (Dewsbury et al. 2002: p. 438).

Making a Difference?

The exciting theoretical concerns posed by cultural geography led to a shift in empirical focus in social geography toward embodied practices and the microprocesses of social interactions and representation through which particular social identities and behaviors are created, valued, and devalued in different places. For many social geographers the growth of cultural geography presented an opportunity to combine many of the themes of the two subdisciplines in fruitful investigation of social inequality and difference and to focus on and destabilize definitions of a wider range of 'social groups' – such as disabled people and gay and lesbian people. In 1988 the '*Social Geography Study Group of the IBG*' (which had been formed in the early 1970s) changed its name to the '*Social and Cultural Geography Study Group*' and 11 years later the journal *Social and Cultural Geography* was founded. These developments exemplified a widespread enthusiasm for the idea that social geographers' tradition of political engagement and ethnographic research could combine positively with cultural geographers' sensitivity to discourse and symbolic expression of difference. Later this combination led to studies of embodied actions and to social and cultural geographers' interest in emotional geographies and geographies of embodiment.

Since 1990, the alliance of social and cultural geography has prompted concerns about the desocializing and depoliticization of social geography and a loss of focus on 'inequality' and 'relevance' – a focus that was central to social geography in the 1970s and 1980s. In the USA and to a somewhat lesser degree in the UK, much research that speaks to those concerns is presented as part of other subdisciplines such as urban geography, population geography, health geography, or geographies of social justice (Del Casino, 2011).

What Future for Social Geography?

Social geography is no longer the 'fashionable and radical' project that it was in the 1970s and early 1980s. Firmly part of the Anglo-American geographic tradition, social geography has largely ignored the intellectual challenges posed by the empirical realities of the developing world and the research of non-Western scholars. In a trenchant critique Peake (2011: p. 768) argues for "interrogation of the institutional whiteness of the discipline and of the white identities of its practitioners" and "openness to the incorporation of non-white knowledges into imaginaries of social geography."

The current period of austerity in the West and growing global and local inequalities pose important political and intellectual questions for geographers concerned with inequality. This challenge may lead to a revival in the importance of the disciplinary identity of social geography or to 'social geography' being carried out under the aegis of other subdisciplines. If social geography can rise to the challenges

Peake poses, it can become part of a movement to reorient human geography as a whole toward a full engagement with both the partiality and the importance of the variety of human geographical concerns and knowledges.

See also: Cognitive Maps; Disability, Geography of; Gender and Feminist Studies in Geography; Geographic Thought, History of; Space and Social Theory in Geography; Spatial Equity; Urban Geography.

Bibliography

Bowlby, S.R., McDowell, L., 1987. The feminist challenge to social geography. In: Pacione, M. (Ed.), Social Geography: Progress and Prospect. Croom Helm, London.

Castells, M., 1976. Theoretical propositions for an experimental study of urban social movements. In: Pickvance, C. (Ed.), Urban Sociology: Critical Essays. Methuen, London.

Del Casino, V., 2011. US social geography, alive and well? Social and Cultural Geography 12 (6), 538–543.

Dewsbury, J., Harrison, P., Rose, M., Wylie, J., 2002. Introduction: enacting geographies. Geoforum 33, 437–440.

Giddens, A., 1981. A Contemporary Critique of Historical Materialism. Macmillan, London.

Gould, P., White, R., 1974. Mental Maps. Penguin, London.

Hagerstrand, T., 1975. Space, time and human conditions. In: Karlqvist, A., Lundqvist, L., Snickars, F. (Eds.), Dynamic Allocation of Urban Space. Saxon House, Farnborough, UK, pp. 3–14.

Hanson, S., Pratt, G., 1995. Gender, Work and Space. Routledge, New York.

Harvey, D., 1973. Social Justice and the City. Edward Arnold, London.

Jackson, P., Smith, S., 1984. Exploring Social Geography. Allen and Unwin, London.

Jones, E., 1975. Readings in Social Geography. Oxford University Press, Oxford, UK.

Johnston, R.L., Gregory, D., Smith, D.M., 1994. The Dictionary of Human Geography, third ed. Blackwell, Oxford, UK.

Meinig, D. (Ed.), 1979. The Interpretation of Ordinary Landscapes. Oxford University Press, New York.

Pahl, R., 1965. Trends in social geography. In: Chorley, R.J., Haggett, P. (Eds.), Frontiers in Geographical Teaching. Methuen, London.

Pahl, R., 1975. Whose City? second ed. Harmondsworth, UK, Penguin. Social geography beyond an Angl-American positionality. Social and Cultural Geography 12 (7), 757–773.

Peake, L., 2011. In, out and unspeakably about: taking social geography beyond an Anglo-American positionality. Social and Cultural Geography 12 (7), 757–773.

Rex, J., 1968. The sociology of a zone in transition. In: Pahl, R.E. (Ed.), Readings in Urban Sociology. Pergamon Press, Oxford, UK.

Smith, D., 1977. Human Geography: A Welfare Approach. Edward Arnold, New York.

Smith, D.P., Browne, K., Bissell, D., 2011. Reinvigorating social geographies? A 'social re/turn' for a changing social world: (re)opening a debate. Social and Cultural Geography 12 (6), 517–528.

Shevky, E., Bell, W., 1955. Social Area Analysis. Stanford University Press, Stanford, CA.

Social and cultural geography, 2011. Forum: Reinvigorating Social Geographies? 12 (6), 517–562.

Social History

Christoph Conrad, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14299–14306, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The term 'social history' refers to a subdiscipline of the historical sciences on the one hand and to a general approach to history that focuses on society at large on the other hand. In both manifestations, social history developed from marginal and tentative origins at the end of the nineteenth and the beginning of the twentieth centuries, and experienced a triumphant expansion from the 1950s to the 1980s. Social historians sought to uncover the relationships between economic, demographic, and social processes and structures, as well as their impact on social stratification, political institutions, social movements, ideologies, and forms of 'public' and 'private' behavior. For the comprehensive variant of 'social history,' the term 'history of society' has gained currency. Class, social status, standard of living, the family, work, social mobility, inequality, modernization, conflict, race, and gender are central keywords in this historiography. Interdisciplinary contact with sociology, especially historical sociology, economics, demography, political science, and, more recently, anthropology and cultural studies has been intensive. The borrowing of concepts and even the adaptation of 'middle range theories' (R.K. Merton) as tools in historical research and writing is a trademark of social historians as compared to traditional narrative historians. Comparative studies have embodied the interest in the application and testing of theories and have brought social history closest to historical sociology.

The term 'social history' refers to a subdiscipline of the historical sciences on the one hand and to a general approach to history that focuses on society at large on the other hand. In both manifestations, social history developed from marginal and tentative origins at the end of the nineteenth and the beginning of the twentieth centuries and experienced a triumphant expansion from the 1950s to the 1980s. Throughout, social history can be best defined in terms of what it wants not to be, or against that to which it proposes an alternative.

On the one hand, social history distinguished itself from political history, which had been dominant during most of the nineteenth and the early twentieth centuries. Consequently, the 'social' in social history meant dealing with the structures of societies and social change; social movements, groups and classes; conditions of work and ways of life; families; households; local communities; urbanization; mobility; ethnic groups etc. On the other hand, social history challenged dominant historical narratives which were constructed around the history of politics and the state or around the history of ideas, by stressing, instead, social change as a core dimension around which historical synthesis and diagnosis of the contemporary world should be organized. With these aims in mind, social historians (including historical sociologists and economic historians) sought to uncover the relationships between economic, demographic, and social processes and structures, as well as their impact on political institutions, the distribution of resources, social movements, shared worldviews, and forms of public and private behavior. For the comprehensive variant of 'social history,' the term 'history of society' (*Gesellschaftsgeschichte*) has gained currency. Not only in Western Europe and North America but also in other parts of the world, the establishment, expansion, and specialization of social history belong among the significant trends of interwar and particularly postwar intellectual and academic life. Nevertheless, in recent years, it shares with other fields of the humanities and social sciences a 'time of doubts' (Chartier, 1993), a period of challenges from interpretative, constructivist, and discursive approaches (Palmer, 1990; Bonnell and Hunt, 1999).

Early Institutionalization (Late Nineteenth–Early Twentieth Century)

Like other generations of historians in the twentieth century that have at some point used the label 'new' for themselves, the claim to innovation by social historians in the 1960s and 1970s obscured the rich diversity of predecessors and competitors, which goes back to the second half of the nineteenth century (Kocka, 1986). In that period, the academic professionalization and disciplinary differentiation of economics, sociology, and political science on the one hand and history on the other, prevailed.

The mutual contact zones or, by contrast, the strength of boundaries, were marked by significant national differences. The resulting interactions and hierarchies between these "fields of knowledge" (Ringer, 1992) differed in respect of the amount of space they provided for historical approaches. The 'Historical School' of political economy in Germany, dominant in the last third of the nineteenth century, is a case in point. It was part of economics and did not gain much recognition in the discipline of history. Some professional historians took part in a process of specialization in 'social and economic history' that kept a critical distance as much from this strand of economics as from the traditional history of great men, events, and ideas. The development of the combined field 'social and economic history' marked a double opposition, on the one hand to the history of political events and on the other hand to an increasingly 'a-historical' economic science.

The creation of a specialized journal in Germany allows one to give a precise date to the development of this subdiscipline of historiography. In 1893 The Journal for Social and Economic

History, (which was continued after 10 years (and still exists today) as the *Vierteljahrschrift für Sozial-und Wirtschaftsgeschichte*) was founded under the direction of the Austrian socialist Ludo Moritz Hartmann and the conservative medievalist Georg von Below. That journal served as a direct reference for Lucien Febvre and Marc Bloch (*see* Bloch, Marc Léopold Benjamin (1886–1944)) when they founded the *Annales* in 1929 in Strasbourg. Not surprisingly, their journal was first called *Annales d'histoire économique et sociale* (1929–39). The international exchanges and mutual observation between some of the protagonists of historiographic innovations in the early decades of the twentieth century were intense. Thus, only 2 years after the French flagship, the Polish journal *Roczniki dziejow spolecznych i gospodarczych* (Yearbooks for social and economic history) appeared. These early journals shared a further characteristic which distinguished them from the mainstream political history around them, namely a keen interest for 'universal,' nonnationalistic history.

In Germany, social and economic history remained a specialized branch without much impact on general history. In England, economic history maintained close links to social themes. After World War II, the French *Annales* historians tended to move their journal away from its earlier identification with this 'subdiscipline.' Similarly, the program represented by Ernest Labrousse for a history of class relations taking as its starting point the economic 'basis,' broadened out more and more into the areas of urban history, quantitative analysis of social strata, and the integral description of regional societies. Where 'social and economic history' continued to be an institutionalized specialty, it did not necessarily figure as an area of major innovations, as was later the case with the 'new' history of the 1960s and 1970s.

These early efforts to establish a specialized field and an alternative view of history with 'society' as the main object, and organizing principle, of analysis were paralleled by similar movements in various countries. In the United States, the *New History* of the turn of the century; in France, Henri Berr with his *Revue de Synthèse (Historique)* starting in 1900 and the reception of Durkheimian sociology by historians; in Germany, the development of a historical sociology by Max Weber and others; and in Great Britain, in the first half of the twentieth century, the impulses of George Unwin and R.H. Tawney for the development of an economic history that included social and cultural factors, were all among the most influential strands of socio-historical thinking (Iggers, 1975).

Two characteristics of these debates and innovations deserve emphasis: the challenge of Marxism in a broad, nonorthodox sense provoked both emulation and animosity. Secondly, it is interesting to note in light of current discussions that, especially in the German context, the concept of 'culture' rather than that of 'society' was conceived as the vantage point for the synthetic organization of knowledge in the social sciences. In the 1950s the representatives of French 'new history' used the term 'history of civilization' to define the broadest possible framework of historical synthesis. Even Trevelyan's (1942) definition of social history as "the history of a people with the politics left out," which is cited as often as it is refuted, provides surprisingly wide room for the 'social,' to implicitly include culture, mentalities, and everyday life. (For a detailed treatment of the above topics *see* Historiography and Historical Thought: Indigenous Cultures in the Americas; History and the Social Sciences).

In recent times a further root of social history has attracted wider attention. It lies in the antimodern nationalist currents that are critical of industrialization which, in particular in the 1920 and 1930s, presented a special attraction for the younger social and human scientists. The interest in rural life, population, and race combined, particularly in Germany, with aggressive nationalism and criticism of the new political borders (Raphael, 1997). This historical writing opposed conventional research subjects and forms; it turned away from the state and classical political history, looked for alternatives to the traditional focus on elites, and took as its example the methods which had already seen much progress in demography, geography, and anthropology. In the 1930s and 1940s, a number of representatives of this so-called *Volksgeschichte* (history of the people) moved closer to National Socialism and became active as wartime political consultants. As some of these historians and social scientists played a significant role in the modernizing of the human sciences in the German Federal Republic, a debate arose as to the extent to which lines of continuity led from the 'history of the people' to social history, and concerning the relation between its representatives' rightist views and their proposed 'structural history' of modern industrial society (Schulze and Oexle, 1999; Conrad, 1999).

Postwar Expansion (1950s–70s)

Not only in Western Europe and North America but also in other parts of the world, the establishment, expansion, and specialization of social history belong among the significant trends of interwar and particularly postwar intellectual and academic life. The spread of the 'new' histories since the 1960s, occurred at a time when systems of higher education in Europe and North America were dramatically expanding. The general increase in academic jobs, student enrolment, and PhD candidates, as well as the massive growth in the publication of books and journals, allowed the various strands of social history to find ample space without replacing or pushing aside more traditional ways of writing and teaching history.

The institutional growth of social history followed different national directions. But it can generally be said that in the 1970s, a high point was reached in its establishment within universities and among the larger public. The most impressive indicator is the founding of new journals concentrated around the year 1975: *Social Science History* (USA, 1975); *Geschichte und Gesellschaft* (West Germany, 1976); *History Workshop Journal* (UK, 1976); *Social History* (UK, 1976). Along with the *Journal of Social History* (USA) founded in 1967 they represented the new 'historical social science,' complementing the older 'school-forming' reviews such as, above all, *Annales ESC* (France, refounded 1946) and *Past and Present* (UK, 1952).

National Variations and International Circulation of Ideas

The triumph of social history in the 1960s and 1970s was a transnational phenomenon. The themes, questions, and methods were astonishingly similar in different countries; the

international specialist public was interconnected through increased contacts. In many places the attractiveness and legitimacy of new approaches were aided by models coming from other countries. Particularly the French *nouvelle histoire*, often described by others as the *Annales* school, developed into a veritable export hit. Already at the international historians congresses in Paris in 1950 and in Rome in 1955, its French representatives and Italian or English allies very much set the tone (Erdmann, 1987). In the following decades, the international circulation of concepts and approaches took on greater strength through the translation of books. A few historiographic models or theoretical influences (such as for instance the work of Fernand Braudel) became widespread in the countries of southern Europe and South America, while others won international recognition via their reception and dissemination on the North American market.

Yet for all that the national differences remained considerable, even in retrospect and after several decades of approaches and exchanges. The strengths of the French school remained – with a few exceptions – limited to the Middle Ages and the early modern period. Not the genesis of industrial society but the explanation of the revolution of 1789 was, and remains, the dominant theme for French historians. By contrast, in Great Britain, an early market society and 'first industrial nation,' historians gave particular importance to these factors. The central approach which English social historians shared with social reformers and social scientists was however focused on the persistence and dynamism of class society.

It is by no means a coincidence that E.P. Thompson in his pioneering work *The Making of the English Working Class* (1963) made decisive strides for social history in this field. And the historical self-examination in other countries can also – though certainly not entirely – be characterized through central motives in their social development: the United States as an immigration and frontier society, in which ethnic cleavages are given more weight than class conflicts; Sweden with its experiences of mass emigration and social crises at the end of the nineteenth and beginning of the twentieth centuries; and finally Japan, which after World War II looked to describe its latest developments in terms of Marxian-inspired modernization models (Conrad, 1999). The political breaches and catastrophes of the twentieth century also formed the node of the new German social history, which was especially influenced by the political and social history initiated by the 'Bielefeld School' historians. The German *Sonderweg*, the German divergence from the West, and the social conditions for the national-socialist rule were the central preoccupations of a whole generation of historians who broke with the 'historicists' (*see* Historicism) and individualized explanatory models (Kocka, 1999; Welskopp, 1999).

Topics, *Neighbors*, Models

The place of social history in the intellectual landscape of the historical and social sciences of the post-World War II era, can be understood by tracing the main trends with respect to preferred research topics and questions and its interdisciplinary alliances, methods, and explanatory models. These trends were most conspicuous in the period from the late 1950s to the 1970s, which witnessed the most pronounced change toward the 'new' social history. The reorientation effected by the spread of social history is most obvious with respect to the subjects chosen for monographs, in the programs of congresses, or in the contents of major journals. The history of social classes and social stratification, in general, entered into the historical mainstream.

In researching social mobility and standards of life in the past, historians in Europe and North America tended to prolong the prehistory of their contemporary societies into the nineteenth century (and later even further back). Labor history (*see* Labor History) left its subaltern and ideological niche and expanded its focus to the realities of work, family, and everyday life. The economic interests and social position of political actors were taken seriously and entered into the 'social interpretation' of great processes like the American and French revolutions, colonialism and imperialism, the development of political parties and social movements, or the causes of war. The fluctuation of harvests, prices and wages, the ebb and flow of population movements and land use, even microbes, and the climate were considered with growing technical sophistication and by an increasing numbers of historians.

Population history with the more micro-oriented historical demography became a leading sector of the new historiography: migration, demographic behavior, and household structure were reconstructed by large research projects. The Cambridge Group for the History of Population and Social Structure founded in 1964 by Peter Laslett, E.A. Wrigley, and Roger Schofield as well as the historical department of the French *Institut National d'Etudes Demographiques* represented the effort of making history 'scientific,' and influenced similar enterprises in Germany and Austria, the Scandinavian countries, and North America among others. Similar aims were held by historically minded sociologists, political scientists, and economists who joined historians in the United States to create a discussion forum under the heading of 'Social Science History' (with conferences since the 1960s and a journal of the same name since 1975).

The flourishing subfields, each with its own journals, congresses, and experts, read like a repertoire of social history themes (Stearns, 1994): labor, demographic, family, urban, rural, ethnic, women's, and gender histories emerged and became aspects of a more encompassing social history . Internal sequences and echoes were to be observed in the establishment of subdisciplines and the proliferation of new topics: the growing emphasis on mentality, agency, subjectivity, and experience since the late 1970s was itself a response to the establishment of abstract, often quantitative, studies on standard of living, family and demography, social and geographic mobility, or political interest groups. Women's historians criticized the unspoken gender assumptions of much of historical demography and family studies, as well as of the history of labor markets and welfare states. The very meaning of the term 'social' was itself reconceived; it lost its direct association with 'socialist' movements or 'social policy' and embraced, instead, the whole abundance of the life-world. The enlarged thematic spectrum of the Amsterdam-based *International Review of Social History* (1956), the Parisian *Le Mouvement Social* (1960), or the German journal *Archiv fuer Sozialgeschichte* (1961) are examples of this transition, which had gained considerable ground by the mid-1970s.

The shift in interdisciplinary alliances mirrored and underpinned topical interests. The nearest neighbors to history were no longer philology, philosophy, or law, but economics, political science, and sociology. For the French *Annales* School, the close relationship with geography and demography was particularly important. The West German focus on the 'political social history' of the late nineteenth and early twentieth centuries, led to an alliance with historical sociology and political science. Various strands of descriptive, often quantitative, economics, social surveys, and sociography are found among the more practical roots of historians who embarked on studies of living standards, family formation, or business cycles. Equally it should not be forgotten that the enormous growth of sociology in the postwar-era held a particular attraction for historians.

The varieties of social and cultural anthropology or ethnology, as it is called in some countries, served as models for a minority of social historians, especially those working on early modern history. The journal *Comparative Studies in Society and History* (founded in 1958) became an important forum for historically minded anthropologists and anthropologically interested historians and social scientists. The relationship between history and anthropology – a discipline which has also changed dramatically during the last three or four decades – is a good indicator for the underlying change of emphasis from 'society' to 'culture' within social history. Natalie Z. Davis's influential collection of essays *Society and Culture in Early Modern France* (Davis, 1975) or the wide impact of Clifford Geertz's writings, especially his 1973 collection *The Interpretation of Cultures*, are cases in point.

Parallel to these shifts, the protagonists and agents of the new social history reflected on the mechanisms of scientific change as such. What appeared to be happening, under their own eyes, and with their help, in the humanities and social sciences seemed to resemble a 'paradigm shift,' as Thomas S. Kuhn had described it for physics in his influential book *The Structure of Scientific Revolutions* (1962). Notwithstanding Kuhn's own doubts as to whether his model could be applied to the social sciences, the 'paradigm shift' became a battle cry in the nascent struggle to establish social history over and against the history of politics, great men, or ideas. The implications of Kuhn's work for the history of science were also welcomed by social historians. Indeed, the emphasis on social and external factors in the development and application of knowledge clearly corresponded to the efforts of social historians to stress context, interests, and structural constraints when dealing with ideas and ideologies (Appleby et al., 1994).

Although, at the time, many were fascinated by the idea of a 'paradigm shift,' in retrospect we see a more subtly differentiated, less abrupt transition. It is certainly true that in individual national academic traditions there were, in part, intense debates as to the 'right' sort of history. And in most cases a political dimension was also present, in which social history was defined as being more to the left and political history more to the right. However, rather than superseding an old paradigm, social historians could work on several levels (institutional, methodological, and thematic) and go about broadening and innovating their field within quite different time frames. If one sees the success story of social history as a pluralistic project fed from many different sources, it becomes easier to understand its expansion and, its later reorientation in the direction of cultural history. In fact, the many social–political and intra-academic forces, which were involved in the rise of social history, have continued to develop independently and gradually cast doubt upon the plausibility of their own basic assumptions.

These definitional clarifications should not, however, obscure the impact of the new social history on the traditions of historiography. The historicist principles of the historians' craft were challenged by analytical, theory-driven approaches. The borrowing of concepts and even the adaptation of 'middle range theories' (R.K. Merton) as tools in historical research and writing, has become a trademark of social historians as compared to traditional, "narrative historians" (Le Goff and Nora, 1985; Fairburn, 1999). The use of comparative methods gained wide currency (Tilly, 1984). Theories of modernization (*see* Modernization, Sociological Theories of) became more attractive, above all, for historians of the nineteenth and twentieth centuries. Even if the explicit use of theories remained, in general, limited, process-oriented terms such as industrialization, urbanization, secularization, and professionalization became ever more widely used in historiography. And they were by no means limited to social history in the more restricted sense. Modernization, industrial society, even democratization, all were part of the meta-narrative governing a good part of social history writ large – albeit with national differences and often heavy criticism from alternative political currents within social history (see for example Judt, 1979). Yet despite evident tension, the height of the social history wave was marked by a good deal of common ground.

The explanatory models proposed by the 'new' histories of the 1950s, the 1960s, and the 1970s shared the aim of explaining collective behavior and social relations by linking them to the structures of the economy and society. Although most opposed the rigidity of official Marxism–Leninism, many historians conceived of society as an architecture which resembled a vaguely materialist interaction between 'basis' and 'superstructure.' In her work on the French revolution, Lynn Hunt explicitly rejected the widespread partition of monographs in three levels of economy/demography, society, and politics or mentality, by simply turning it on its head (Hunt, 1990: pp. 102–103). Interests, world views, political stances were largely understood as collective responses to objective social positions. Since the 1970s, social history has experienced shifts in its basic assumptions about causality that have led to a much less homogenous situation. But there seems to be a widespread refocusing, first on agency and practices, secondly on meaning and identity, and thirdly on microprocesses.

The Will to Synthesis

The social-historical approach to history set other requirements on the writing of an overall historical account as conventional political history. The narrative structure of political events, the periodization by means of wars or revolutions, and the intentionality of a historiography oriented towards the 'big players' could not be simply transposed on to social history. In this new field, syntheses therefore had to contain – whether one liked it or not – a reflection on the character of 'society.' The will to synthesis is, for this reason, part and parcel of social history's

emphasis on a comprehensive approach to history. Here social history sets itself off most clearly from the study of social and economic phenomena within a context, which continues to be defined in political terms.

The authors of large syntheses in social history are, for that reason, also more or less explicitly protagonists of a 'history of society,' or of a *histoire totale* in the French tradition. Mikhail Rostovtzeff's *Social and Economic History of the Roman Empire* (1926) and Marc Bloch's *Société féodale* (1939–40) are both early and distinguished examples. Fernand Braudel, Eric Hobsbawm, and Hans-Ulrich Wehler emphasized, in their own way, this requirement with monumental general histories. But they could do so only by consciously borrowing concepts and models from the neighboring social sciences and by setting out theoretical considerations on social structure and on the motors of social change. Whether these involved models on temporal and geographic structuressuch as Braudel's, or a flexible use of Marxist theories such as Hobsbawm's, or an explicit use of modernization theories such as Wehler's – in each case this type of general history distinguished itself from the conventional historiography in its explicit theoretical orientation also. Yet a clear difference in the degree to which theory governs the historical narrative or the display of arguments remains between these works and those of historical sociologists such as Immanuel Wallerstein, Theda Skocpol, Charles Tilly, or Michael Mann – authors who, nevertheless, share common interests with the historians.

But also at a more humble geographical and temporal level, social historians have continuously endeavored to reconstruct the object of their investigations, society, as a coherent if multifaceted system, sometimes even in a literary way. Above all, however, they continue writing within their respective national historical frameworks (Wehler, 1987–2008). Often this takes the form of collections of articles in which various authors describe the individual aspects of society (a recent example is Thompson, 1990). Even the short, broad-based handbook giving an overview of the state of research belongs in this category. In part, this reflects the tendency to specific research on social classes and strata which leaves out further societal, cultural, and political contexts (Charle, 1994). Although the emergence of the newer directions of gender, experience, and cultural history has increased the fear that social history would become even more fragmented, José Harris (1993), for example, has explicitly endeavored to integrate these very perspectives into a general historical account.

The Revenge of Success

However, social history's expansion in quantitative scale and conceptual scope was large enough to produce serious side-effects. First, specialization increased dramatically. Greater demand for technical expertise, such as quantification, the creation of such subfields as historical demography, or the tendency of methodological approaches like oral history to take on lives of their own, contributed to the differentiation of the field. Second, the proliferation of topics as well as of local or regional units of analysis deemed worthy of a monograph, concurred to create a very colorful map of approaches and subjects. In the late 1980s and afterwards, representatives of social history looking back on the success story of their field could made out the first signs of self-doubt already present in the second half of the 1970s at the highpoint of social history's intellectual and institutional glory (see contributions in Kocka, 1989). More than any other approach to history, social history gradually came to be marked by fragmentation and dispersion. Third, social history, as a general approach, gained access to other historiographical periods and to other disciplines in the humanities. In literary criticism and modern languages the social history of literature became a trademark in the 1970s and early 1980s and, since then, has met with growing skepticism and disillusionment.

Certainly it must be said that these developments represent a tribute to the success of social history. In the course of differentiation and specialization, the new approaches and subfields which had established themselves under the heading of social history, gradually changed their character and underlying priorities (Burke, 1991). What is often described at the end of the twentieth century as a cultural 'turn,' in fact, emerged over the years. Indeed, disillusionment with the social history of the 1970s can best be understood not as a 'paradigm shift' but rather as an incremental, self-reinforcing, sometimes conflictual transition on different, nonunified levels of historical work: research methods, preferred topics and sources, examples taken from other disciplines, new 'buzz words' and concepts, more hidden understandings of causality, agency, and evolution. Without question, the more recent set of shifts from social history to the 'new cultural history,' makes us think harder about the establishment of social history and its concomitant changes of the societal and political context in earlier decades.

The recent 'cultural turn' can, first of all, be understood immanently, that is, intra-academically. In its various manifestations, social history continued in the tradition of an oppositional approach against the historical mainstream. Small wonder that it has attracted attacks from every new wave of critical and revisionist historians as well as from the revival of older (biographic, narrative) forms of historical writing. The concept of gender has in many countries been at the forefront of debates around structure and agency, macrosociological categories and deconstructivist criticism. Particularly in Germany, a fierce debate with aspects of a generational conflict divided the 'history of society' embodied in the Bielefeld School of *Gesellschaftsgeschichte* (Iggers, 1975; Kocka, 1986; Wehler, 1987–2008) and the 'history of everyday life' (*Alltagsgeschichte*) represented by grassroots history workshops and a number of younger academics at the margin of the profession (Lüdtke, 1995).

Yet one must also consider the external, extra-academic factors which in the last decades have changed the relationship of the social sciences to society. The social and political movements of the postwar years made their mark on social history. Especially in the United States, the expansion of themes and problems in social history since the 1960s had been given a strong push by the corresponding conflicts in US society: the movements of Afro-Americans, of students, of women, and of gays and lesbians added to the traditional input from the workers' movement and radical intellectuals. Since the 1980s, the already proverbial trinity of race, class, and gender reflected these social forces in an academic discourse. In Europe, too, a similar, by no means linear, elective affinity could be observed between social movements (the alternative

left, and the women's, ecological, and peace movements) and fresh ideas in history (gender, everyday, and microhistory).

The 1980s and 1990s experienced a strong move toward independence and pluralization of identity politics and the associated attempts at claiming one's own history for oneself. Taken together, these trends have aided the transition away from a history of anonymous social processes and structures. This reorientation became even stronger with growing critique of the ethnocentric dominance of industrial societies in social history leading up to this time. In recent years, the slow inroads which postcolonial discourses have also made into historical science have, on the one hand, furthered the significance of classical social history, this time in the countries of Asia, Africa, and Latin America. On the other hand, the established concepts of modernization, development, progress, and culture were subject to a thoroughgoing critique and redefinition. Finally, the breakdown of actually existing socialism in 1989/1990 and the resulting political and economic transformation led to reflection on the basics of politics and the economy (Kocka, 2000).

The answers to these challenges and interrogations are by no means clear. The voices which speak of a 'crisis' are not restricted to social history but are directed at the historical sciences, if not the humanities in general. Two trends can be made out: beyond all methodological differences, the self-reflection of the academic historian has moved to center stage. The idea that one can produce reliable knowledge about society – contemporary or past – was in 2000 taken far less for granted than a few decades previously (Bonnell and Hunt, 1999). On the other hand, the historicization of all forms of knowledge has become, as it were, a patent remedy in all debates. In opening up toward this sociocultural direction, it is true that social history may have lost out in unity, perhaps in recognizability, but it has gained in intellectual dynamism.

Editor's Note

Although "social history" as a method of "doing history" has faded as a topic of analysis among professional historians in the past 15 years, social historians continue to produce interesting and informative work. Among the more active of the contemporary social historians is Peter Burke. See his A Social History of Knowledge: From Gutenberg to Diderot (Cambridge: Polity Press, 2000); A Social History of the Media: From Gutenberg to the Internet (with Asa Briggs) (Cambridge: Polity Press, 2009). Google turns up "social histories" of a wide variety of social and cultural phenomena: education in England, the Third Reich, care of the insane, the English novel, oral contraceptives, English music, cricket, post-war Japan, drugs, women in broadcasting, the English countryside, etc.

See also: Business History; Cultural History; Economic History; Family and Kinship, History of; Gender History; Historical Demography; History and the Social Sciences; Labor History; Life Course in History; Modernization and Modernity in History; Social Movements, History of: General; Subaltern History; Urban History; Work, History of; Work, Labor: History of the Concept; Working Classes, History of.

Bibliography

Appleby, J., Hunt, L., Jacob, M., 1994. Telling the Truth about History. Norton, New York.
Bonnell, V.E., Hunt, L. (Eds.), 1999. Beyond the Cultural Turn. New Directions in the Study of Society and Culture. University of California Press, Berkeley, CA.
Burke, P. (Ed.), 1991. New Perspectives on Historical Writing. Cambridge University Press, Cambridge, UK.
Charle, C., 1994. A Social History of France in the Nineteenth Century. Berg, Oxford, UK.
Chartier, R., 18 March 1993. Le temps des doutes. Le Monde, Paris vi–vii.
Conrad, S., 1999. Auf der Suche nach der verlorenen Nation. Geschichtsschreibung in Westdeutschland und Japan 1945–1960. Vandenhoeck & Ruprecht, Göttingen, Germany.
Davis, N.Z., 1975. Society and Culture in Early Modern France. Stanford University Press, Stanford, CA.
Erdmann, K.E., 1987. Die Ökumene der Historiker. Geschichte der Internationalen Historikerkongresse und des Comité International des Sciences Historiques. Vandenhoeck & Ruprecht, Göttingen, Germany.
Fairburn, M., 1999. Social History. Problems, Strategies and Methods. Macmillan, London.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Harris, J., 1993. Private Lives, Public Spirit: Britain 1870–1914 (The Penguin Social History of Britain). Oxford University Press, Oxford, UK.
Hobsbawm, E.J., 1972. From social history to the history of society. In: Gilbert, F., Graubard, R. (Eds.), Historical Studies Today. Norton, New York.
Hunt, L., 1990. History beyond social theory. In: Carroll, D. (Ed.), The States of 'Theory'. History, Art, and Critical Discourse. Columbia University Press, New York.
Iggers, G.G., 1975. New Directions in European Historiography. Wesleyan University Press, Middletown, CT.
Iggers, G.G., 1997. Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge. Wesleyan University Press, Middletown, CT.
Judt, T., 1979. A clown in regal purple: social history and the historians. History Workshop Journal 7, 66–94.
Kocka, J., 1986. Sozialgeschichte. Begriff, Entwicklung, Probleme, second ed. Vandenhoeck & Ruprecht, Göttingen, Germany.
Kocka, J. (Ed.), 1989. Sozialgeschichte im internationalen berblick. Ergebnisse und Tendenzen der Forschung. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.
Kocka, J., 1999. Asymmetrical historical comparison: the case of the German Sonderweg. History and Theory 38, 45–50.
Kocka, J., 2000. Historische Sozialwissenschaft heute. In: Nolte, P., Hettling, M., et al. (Eds.), Perspektiven der Gesellschaftsgeschichte. Beck, Munich, Germany.
Le Goff, J., Nora, P. (Eds.), 1985. Constructing the Past: Essays in Historical Methodology. Cambridge University Press, Cambridge, UK.
Lüdtke, A. (Ed.), 1995. The History of Everyday Life. Princeton University Press, Princeton, NJ.
Palmer, B.D., 1990. Descent into Discourse. The Reification of Language and the Writing of Social History. Temple University Press, Philadelphia, PA.
Raphael, L., 1997. Die 'neue Geschichte' – Umbrüche und neue Wege der Geschichtsschreibung in internationaler Perspektive (1880–1940). In: Küttler, W., et al. (Eds.), Geschichtsdiskurs, vol. 4. Fischer, Frankfurt/Main, Germany.
Ringer, F., 1992. Fields of Knowledge. French Academic Culture in Comparative Perspective. Cambridge University Press, Cambridge, UK, 1890–1920.
Schulze, W., Oexle, O.G. (Eds.), 1999. Deutsche Historiker im Nationalsozialismus. Fischer, Frankfurt/Main, Germany.
Stearns, P.N. (Ed.), 1994. Encyclopedia of Social History. Garland, New York.
Thompson, F.M.L. (Ed.), 1990. The Cambridge Social History of Britain, 3 vols. Cambridge University Press, Cambridge, UK.
Tilly, C., 1984. Big Structures, Large Processes, Huge Comparisons. Russell Sage Foundation, New York.
Trevelyan, G.M., 1942. English Social History: a Survey of Six Centuries. Longmans, London, New York.
Wehler, H.-U., 1987–2008. Deutsche Gesellschaftsgeschichte, 5 vols. Beck, München, Germany.
Welskopp, T., 1999. Westbindung auf dem 'Sonderweg'. Die deutsche Sozialgeschichte vom Appendix der Wirtschaftsgeschichte zur historischen Sozialwissenschaft. In: Küttler, W., et al. (Eds.), Geschichtsdiskurs, vol. 5. Fischer, Frankfurt/Main, Germany.

Social Identity in Social Psychology

Katherine J Reynolds, Research School of Psychology, The Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Hardin, volume 11, pp. 7166–7170, © 2001, Elsevier Ltd.

Abstract

In this article different definitions of social identity are outlined that include developmental and sociological approaches. The main focus of the article is on the concept of identity in social psychology and the social identity perspective specifically. The psychological group and social identity processes are outlined along with the conditions that explain which of many identities (self-other similarity, self-stereotyping) will become salient. As part of this endeavor social identification and social identity salience are differentiated and implications for measurement are outlined. It is argued that social identity – the context specific definitions of who 'we' are and what 'we' do – both shape behavior and facilitate behavior change.

Many disciplines in the social and behavioral sciences use the term identity. There is reference to personal identity, social identity, cultural identity, role identity, gender identity, political identity, ethnic identity, and religious identity – to name but a few. In fact based on the social sciences citation index the term identity was listed as a topic 895 times between 1945 and 1970 increasing to 16 125 times between 1991 and 2000 and then to 38 079 times between 2001 and 2010. There is sustained and growing interest in the identity construct. This interest reflects the value and importance of people's group memberships (culture, gender, religion) in explaining phenomena of interest. The growing study of identity also may reflect the availability of instruments to measure the construct in valid and reliable ways. It is also possible that interest reflects a move away from the group (and associated identities) being viewed through a prism of *anticollectivism* – where groups were sidelined because they were considered somehow flawed or inherently inferior to individual understanding (see Reicher et al., 2012 for a recent discussion). For a long time, groups and associated identity have been associated with more primitive 'tribal' psychology where rationality is lost to the emotion of a crowd or mindless conformity to the 'other.' More recently, the group and 'identity' are recognized being important because such processes define standards and shape behavior – what is considered (in)appropriate, just and 'right.' Identity processes are also a vehicle for social change enabling the powerless to be more powerful in addressing inequality and social justice through acting together.

How is identity defined? What is the connection between sociological groups (gender, ethnicity) and identification? Why are these ideas important for the social and behavioral sciences? How is identity measured? Answering these kinds of questions (at least briefly) is the focus of the rest of this article. The article also is framed by the subdiscipline of social psychology and, in particular, explores and highlights social identity as understood within social identity theory and self-categorization theory; often referred to as the social identity perspective (Turner and Reynolds, 2001).

How Is Identity Defined?

The term identity comes from the Latin noun *identitas*, where *tatis* is a derivation of the adjective *idem*, which means 'the same.' There is a focus on sharing a degree of sameness with others on a particular dimension and in order for such sameness to be established comparison is necessary. To complicate matters further the term identity is not used to mean the same thing within psychology or across the social and behavioral sciences. Different literature have used the term in different ways. Even when it comes to use of the term social identity there are divergent meanings, each embedded in particular theoretical frameworks and literatures (e.g., see Brewer, 2001; Deaux, 1993).

Erikson one of the earliest psychologists interested in identity, highlighted two types of identity – ego identity and cultural or social identity. Ego identity refers to a conscious sense of self as unique as well as continuity of experience. A strong sense of self is argued to emerge during adolescence and is central to well-being. One way to think about social identity within this framework is that it captures the extent to which there is a supportive community that validates the identity and gives it strength (e.g., "greater validation of social identity can nurture ego identity," Côté and Levine, 2002: p. 94). More precisely, Erikson argues that "[Ego identity's] most obvious concomitants are a feeling of being at home in one's body, a sense of 'knowing where one is going,' and an inner assuredness of anticipated recognition from those who count" (Erikson, 1968: p. 165).

There is also work on culture and social representations that is related to shaping a sense of self and identity. Being a member of a community means that people come to share the same knowledge, beliefs, theories, conventions, and norms, which give meaning to the environment and guide practices and actions. A common reality of stereotypes, attitudes, opinions, and categorizations is constructed and shared through collective activities (e.g., Augoustinos and Walker, 1995; Chryssochoou, 2004). Such processes underpin the formation of people's 'distal' knowledge, expectations, and theories about the social world and themselves as a function of being a member of a particular 'community' of interacting individuals.

Symbolic interactionism and socialization are central to explaining how the individual becomes shaped by, and attuned to, the social patterns and regularities of their cultural community. There is a trajectory of work through Cooley (1902/1922) and Mead (1934), in particular, where others'

perceptions and reactions are argued to shape one's self-view in a dynamic way. These ideas form the basis of the way identity is understood in sociology where the formation of social identity includes the reflected appraisals of others and others expectations. A generalized conception of one's own self develops through experience and observations of how others act toward the self (the so-called 'looking' glass self). The self is social or public in that it contains the internalized reactions, expectations, and judgments of others (e.g., Turner and Onorato, 1999; Stryker and Statham, 1985). The person enacts these internalized reactions conforming to norms in order to meet others' expectations and achieve the desired outcomes that such conformity affords (e.g., Thoits and Virshup, 1997). In this tradition social identity is an attempt to present the desired views of others as well as define the self to oneself (Deaux and Burke, 2010). The focus in the symbolic interactionist perspective (and related work) is on how the mind is socialized and becomes a reflective object of itself (see Turner and Onorato, 1999 for a more detailed discussion).

In summary, then, the 'social' in identity in the developmental approach, social representations, and sociological literatures refer to others in the environment who impact on the person and the development of self-construal. Identity is self-knowledge gained through experience and reflected appraisals. In sociology there is also an explicit interest in role of groups and socialization of social norms – this work explicitly connects social structure and normative societal or cultural knowledge to explanations of individual behavior. As is outline in more detail in the next section, the psychological group along with a different understanding of social identity may be informative in understanding these processes.

What Is the Connection between Groups and Identification?

In social psychology another layer is added to these understandings of identity. A person may internalize the reflected appraisals of others and the norms and scripts associated with a particular role but not psychologically identify as a member of a particular group in a given situation – 'mothers,' 'social psychologists,' or 'Australians.' As Turner (2000) argues "making a distinction between peoples' relatively stable identifications with particular groups and roles and the salience of any social identity in a specific situation is an important one" (p. 342).

Social identity theory and self-categorization theory focus on the psychological salience of group membership. Self-categorization theory first and foremost is a theory of *psychological* group membership (when is a group a 'group'?). The theory suggests that individuals can view themselves in terms of idiosyncratic personal attributes (e.g., a personal identity) or in terms of social categories (e.g., social identities, such as a social psychologist or an Australian). The concept of personal identity or the personal self ('I') is used to describe situations where individuals perceive themselves to be distinct and different from others (that are available for comparison) and social identity or the social self ('we') refers to an individual's "knowledge that he [or she] belongs to certain groups together with some emotional and value significance to him [or her] of the group membership" (Tajfel, 1972: p. 31).

Importantly, the term social identity refers to psychological groups where an individual perceives him- or herself as being a member because the group is self-relevant and self-defining (Turner, 1982, 1985). Turner (1984) hypothesized that as people defined themselves and others as members of the same category, they would self-stereotype in relation to the category and tend to see themselves as more alike in terms of the defining attributes of the category. This process is referred to as depersonalization. It was argued that it is "the cognitive redefinition of the self – from unique attributes and individual differences to shared social category memberships and associated stereotypes – that mediates group behaviour" (Turner, 1984: p. 528).

It is also recognized that the self is hierarchically organized and can be defined in ways that are more or less inclusive of others (as similar to self) at the individual, subgroup, and superordinate levels. In this exploration of levels of self-categorization, all the different aspects of human experience (as an individual 'I' and group member 'we') are addressed providing a more integrated view of the whole person. Using the same terminology as Kluckhohn and Murray (1953), it is possible to explain how the individual person is like no other person, how the individual person is like some other persons, and how the individual person is like all other persons.

Fundamental SCT research suggests people will only act in the interests of a group to the extent that they see that group as self-defining in a given context (Turner et al., 1994; Oakes et al., 1994). From this perspective, then, a critical question is what determines which identity (and its associated content) will become relevant and meaningful in a particular situation. In self-categorization theory, it is argued that a particular self-categorization and associated norms, beliefs, and behaviors are an outcome of the dynamic interaction between perceiver readiness factors and comparative and normative fit factors. It is the resulting salient self-categorization or identity, that may be as an individual or group member, that is informative in understanding and explaining one's behavior in a given situation.

Perceiver readiness builds on Bruner's (1957) notion of accessibility and relates to the accessibility of particular ways of categorizing based on the perceivers' expectations, motives, values, and goals (Oakes et al., 1994: p. 199). Turner (1987) argues that the "the centrality and evaluative importance of a group membership in self-definition, is a major determinant of accessibility" (p. 55). For example, highly prejudiced people, for whom being 'white' could be an important aspect of their self-definition, may be more ready to categorize stimuli (e.g., a meeting where Anglo-Australians and Indigenous Australians are present) in racial terms. Perceiver readiness can be viewed as 'tuning' the perceiver to particular features in the stimulus environment or social context (Swann and Seyle, 2005).

Readiness is not the only consideration in explaining salience. There must be evidence in the situation that sustains such a categorization. Perceiver readiness and the perceived features of the stimulus environment interact to produce

context-specific categorization. The *meta*-contrast principle is the basis of comparative fit (Turner, 1985). This principle states that in a given situation, stimuli will be categorized as the same when the average differences perceived between them (intraclass differences) are less than the differences between them and other stimuli (interclass differences) within the comparative context.

Categories are formed through a context-specific assessment of similarity and difference. For example, an individual would self-categorize more as a clinical psychologist ('us') rather than individual academic ('I') at a Departmental staff meeting, to the extent that during discussion, the differences between the person and other clinical psychologists were perceived to be less than the differences between clinical psychologists and personality psychologists in that particular social context (e.g., discussions on a new appointment in the Department). Based on the principles of comparative fit, in another situation when social psychologists were included in the frame of reference because say they started to make a pitch for a social academic to fill the new position, clinical and personality psychologists could be seen as more similar ('us') and unified in comparison to the social group ('them'; see Haslam and Turner, 1992; Haslam et al., 1992).

Also impressions of one's own sense of self as a unique and distinct individual ('who am I') are viewed as an act of categorization. An individual may be perceived as outgoing or caring to the extent that the behavioral variability displayed by the person in relation to this characteristic across time and occasions, was less than the variability on this dimension between the individual and other individuals available for comparison (also see Plaks et al., 2003). The formation of a person category follows the same process as the formation of a social category (Mavor, 2004; Oakes, 1987; Oakes et al., 1994; Reynolds and Oakes, 2000).

However, to categorize oneself and others as clinical and personality psychologists or as outgoing or caring, there must not only be similarities within and differences between the stimuli, but these must be in the direction that makes sense for the perceiver on the relevant dimensions of comparison. So there is a content as well as a structural dimension to categorization. The similarities among clinical psychologists (and differences from social psychologists) must make sense based on people's normative expectations about these groups. Normative fit relates to the content dimension, the direction or meaning of the categorization in a particular social environment (see Oakes, 1987).

In summary, the focus for these social psychological theories is on psychological group memberships and social identity – the term in-group is used to describe a group where the norms, values, and beliefs that define the group are internalized and become self-defining in a given context. The theories, knowledge, and normative understandings one has about the world that are shaped by societal position, social roles, and social relations and culture, are only one aspect in explaining social identity salience. It is also necessary to consider the immediate context where the emergent self-categorization or identity that drives behavior will be an outcome of the interaction between 'distal' knowledge (perceiver readiness) and features of the situation (comparative and normative fit).

Why Are These Ideas Important for the Social and Behavioral Sciences?

"A group or role may be chronically self-defining across time, but its influence on behavior is a function of its immediate prepotence in self-perception, which reflects varying motives, goals and circumstances" (Turner, 2000: p. 343). This statement highlights why the social identity concept properly understood is important for the social and behavioral sciences. In the Lewinian tradition, social identity recognizes the group and emergent norms as being important in explaining behavior but also provides new avenues to understanding behavior change (see Lewin, 1951; Reynolds et al., 2010, 2014; Reynolds and Turner, 2006).

Given the importance of this point it is worthwhile explaining in more detail the self-categorization theory argument. In explaining group behavior Turner (e.g., 1987) argues that (1) individuals define themselves as similar to others as a member of the same social category (similar perceivers, in a similar position experiencing the same environment), (2) they learn or develop the appropriate, expected, desirable behaviors that are related to category membership and differentiate it from other categories (a sense of 'us'), (3) they assign the norms and attributes of the category to themselves (internalization) through processes of self-stereotyping, and (4) thus their attitudes, emotions, and behaviors become normative as their category membership becomes salient.

These ideas were extended and further investigated (Turner, 1987, 1991) to show that it is through agreement with other group members 'like us' that one comes to feel one's views are valid, correct, and right. When similar others see the world in the same way this gives us confidence in our own responses. Because other in-group members are viewed as similar to oneself, they become a valid source of information and a testing ground for one's own views on relevant dimensions. It is through social identification and associated social influence that others 'like us' come to play an important role in establishing one's own views, the certainty of such views, and their transmission within the group.

It is argued that it is only within a shared in-group framework where there is a shared social identity that differences in perspective (e.g., criticism, new ideas, deviance) can be resolved through discussion, clarification, and mutual influence. Through these processes in-group members can shape each others' norms, values, and beliefs in significant ways (re)defining 'who we are' and 'what we do.' As different people come to be defined as similar to oneself, they offer new opportunities for social influence and the potential for one's theories, expectations, and beliefs about oneself and the world can change. In a more general sense, though, it is through social identity and the construction of definitions of 'who we are' and 'who we are not' – of making particular definitions of who 'we' are salient and prepotent – that explains how behaviors can become consensualized and transformed into coordinated action.

How Is Identity Measured?

An important distinction has been made between people's relatively stable social identification with particular groups and

social identity salience in a specific situation. Those roles or groups that may be chronically self-defining across time need to be salient in the situation to affect behavior. Postmes et al. (2012) recently argue that social identification is not extensively discussed in either social identity theory or self-categorization theory. The focus of these theories is on recognizing and understanding the emergence of a social identity and its consequences. Social identity is understood as an entity that has meaning psychologically because of its defining characteristics, norms, and relationship with out-groups. Social identity salience refers to how psychologically operative the social identity is in the current context and is important in explaining behavior in a given situation.

In contrast, social identification can be understood as a person's relationship to that entity. Social identification can be thought of as an individual difference measure indicating how important and valued a particular group membership is to self-definition. In self-categorization theory terms it indicates, in part, a person's readiness to see themselves in terms of a particular social category. The salience and internalization of a particular group membership can affect one's readiness to perceive oneself in terms of the particular group in the future (see McGarty, 1999 for further elaboration).

The distinction between social identification and social identity also is important in terms of understanding the 'meaning' or content that defines 'us' in a given situation. Measures of social identification do not (typically) clarify the meaning of the social identity. Two people could indicate to the same degree that being an Australian is important to their identity, they feel pleased to be an Australian, and feel strong ties with other Australians but yet the meaning of the category (the dimensions on which they are self-stereotyping) is fundamentally different. Social identity salience, on the other hand, leads to increased uniformity in self-stereotyping among in-group members as the process of salience itself serves to clarify and consensualize 'meaning' or content. When a social identity becomes salient it will involve comparison with others and categorization of similarity and difference. A shared representation of 'us' and 'them' is likely to emerge as well as clarify of who 'we' are in this context.

Most of the measures that have emerged over recent years concern social identification. During the 1980s and 1990s, there was a focus on unidimensional measures of identification (Haslam, 2004; for a review). Doosje et al. (1995), for example, developed a four-item scale with items covering cognitive (a person's knowledge or awareness of being a member of a particular group; 'I identify with members of (Group X),' 'I see myself as a member of (Group X)'), affective (a person's feeling of affective commitment to the group; 'I feel strong ties to member of (Group X)'), and evaluative (a person's sense of group-based positivity; 'I am pleased to be a member of (Group X)') components of identification. Multidimensional scales started to emerge at the end of the 1990s and have been expanding in number (e.g., Cameron, 2004; Ellemers et al., 1999; Leach et al., 2008; Phinney, 1992; Roccas et al., 2008; Sellers et al., 1998).

Roccas et al. (2008) propose four modes of identification that are conceptually distinct – importance (how much I view the group as part of who I am), commitment (how much I want to benefit the group), superiority (how much I view my group as superior to other groups), and deference (how much I honor, revere, and submit to the group's norms, symbols, and leaders). Superiority is an example where an aspect of social identification moves away from one's relationship to the entity to a belief about the entity itself – it comprises information about social position or group beliefs. Seller et al. (1998) do the same with their Ideology scale that forms part of the Multidimensional Model of Racial Identity measurement.

Postmes et al. (2012) have recently argued that there are advantages in use of single-item measures of social identification. With a single item there is clarity of what is being measured and transparency. It is also efficient to measure constructs of interest with as few items as is possible. Furthermore, short scales can be just as valid and reliable as multi-item and multidimensional measures in particular when the construct under investigation is clearly defined and relatively homogeneous (e.g., Burisch, 1984; Loo and Kelts, 1998). Postmes et al. compared the single item 'I identify with my group' to other multidimensional and single-item measures and found the single item was reliable. This item alone captured much of the concept of social identification.

The measurement of social identity has attracted less attention among researchers (perhaps because of confusion with social identification). Social identity salience (a social identity that is activated or deployed at a given point in time) should be revealed by self-stereotyping in relation to the defining features of the in-group in comparison to the comparison group/s (e.g., women compared to men). Hogg and Turner (1987), for example, showed that when people were organized into mixed-sex groups (men and women) or same-sex groups (men only or women only), individuals were more likely to define themselves in terms of gender and to accentuate their similarity to those of the same gender in the mixed-sex settings as opposed to when only men or women were present (see also Oakes et al., 1991; Onorato and Turner, 2004).

In addition, salience means that there should be greater consensus on the defining content of what it means to be a group member. Along these lines, Haslam et al. (1999) asked participants to reflect on their social category membership as Australians (social identity condition) or to focus on their uniqueness from others (personal identity condition) and completed a checklist identifying words typical of people from Australia before and after a group interaction phase. In the social identity condition (and especially postinteraction), there was greater consensus on the attributes that defined the category of Australian.

When social identity is salient there should be a sense of self-other similarity within the in-group and self-other difference with respect to the comparison group/s. These similarity aspects of social identity salience can then be captured by measures such as 'How similar you think you are to other members of Group X?'; 'How much do you think you have in common with members of Group X?'; 'How similar do you think your values are to members of Group X?'; 'How similar do you think you are to members of Group Y?'; 'How similar do you think members of Group Y are to each other?'; 'How much do you think members of Group X

are similar to one another?'; 'How representative do you think you are of Group X?'; 'How much do you think members of Group X will agree?'

Close inspection of the measurement of social identification and social identity and its salience helps to clarify how these concepts are being defined and used. The development of measurement tools serves to help further clarify these constructs and investigate when and how these concepts are useful in explaining behavior and behavior change.

Conclusion

In this article different definitions of social identity have been outlined that include more developmental and sociological approaches where the focus is on the formation of social identity and its importance for the individual. There is a vast and vibrant exploration of these concepts in these areas that has only really been touched on briefly. The focus of the article has been on social psychology and the social identity perspective specifically. From this perspective the preoccupation is on explaining which of the many group memberships will emerge and become important in predicting behavior in a given situation and its consequences. As part of this endeavor it has been important to differentiate social identification from social identity and its salience. This distinction has implications for measurement. It has been argued that it is social identity salience – the context-specific definitions of who 'we' are and what 'we' do – that both shape behavior and facilitate behavior change.

Acknowledgments

This research was supported by an Australian Research Council grant. We gratefully acknowledge the assistance of Corie Lin in collating the material for this article.

See also: Attitudes and Behavior; Collective Behavior, Social Psychology of; Cross-Cultural Psychology; Culture and the Self: Implications for Psychological Theory; Ethnic Identity, Psychology of; Group Processes, Social Psychology of; Identity and Identification, Social Psychology of; Indigenous Psychology; Intergroup Relations; Leadership; Persuasion Theories; Representations, Social Psychology of; Self-Categorization Theory; Social Categorization; Social Psychology; Values, Psychology of.

Bibliography

Augoustinos, M., Walker, I., 1995. Social Cognition: An Integrated Introduction. Sage, London.

Brewer, M.B., 2001. Ingroup identification and intergroup conflict: when does ingroup love become outgroup hate? In: Ashmore, R., Jussim, L., Wilder, D. (Eds.), Social Identity, Intergroup Conflict, and Conflict Reduction. Oxford University Press, New York, pp. 17–41.

Bruner, J.S., 1957. Going beyond the information given. In: Bruner, J.S., Brunswik, E., Festinger, L., Heider, F., Muenzinger, K.F., Osgood, C.E., Rapaport, D. (Eds.), Contemporary Approaches to Cognition. Harvard University Press, Cambridge, MA, pp. 41–69 [Reprinted in: Bruner, J.S., 1973. Beyond the Information Given. Norton, New York. pp. 218–238.].

Burisch, M., 1984. Approaches to personality inventory construction: a comparison of merits. American Psychologist 39 (3), 214.

Cameron, J.E., 2004. A three-factor model of social identity. Self and Identity 3, 239–262.

Chryssochoou, X., 2004. Cultural Diversity: Its Social Psychology. Blackwell, Malden, MA.

Cooley, C.H., 1902. Human Nature and the Social Order. Scribner's, New York.

Côté, J.E., Levine, C.G., 2002. Identity Formation, Agency, and Culture: A Social Psychological Synthesis. Lawrence Erlbaum, Mahwah, NJ.

Deaux, K., 1993. Reconstructing social identity. Personality and Social Psychology Bulletin 19, 4–12.

Deaux, K., Burke, P., 2010. Bridging identities. Social Psychology Quarterly 73, 315–320.

Doosje, B., Ellemers, N., Spears, R., 1995. Perceived variability as a function of group status and identification. Journal of Experimental Social Psychology 31, 410–436.

Ellemers, N., Kortekaas, P., Ouwerkerk, J.W., 1999. Self-categorisation, commitment to the group and group self-esteem as related but distinct aspects of social identity. European Journal of Social Psychology 29 (2–3), 371–389.

Erikson, E.H., 1968. Identity: Youth and Crisis. Norton, New York.

Haslam, S.A., 2004. Psychology in Organizations: The Social Identity Approach, second ed. Sage Publications, Thousand Oakes, CA.

Haslam, S.A., Oakes, P.J., Reynolds, K.J., Turner, J.C., 1999. Social identity salience and the emergence of stereotype consensus. Personality and Social Psychology Bulletin 25, 809–818.

Haslam, S.A., Turner, J.C., 1992. Context-dependent variation in social stereotyping 2: the relationship between frame of reference, self-categorization and accentuation. European Journal of Social Psychology 22, 251–277.

Haslam, S.A., Turner, J.C., Oakes, P.J., McGarty, C., Hayes, B.K., 1992. Context-dependent variation in social stereotyping 1: the effects of intergroup relations as mediated by social change and frame of reference. European Journal of Social Psychology 22, 3–20.

Hogg, M.A., Turner, J.C., 1987. Intergroup behavior, self-stereotyping and the salience of social categories. British Journal of Social Psychology 26, 325–340.

Kluckhohn, C., Murray, H.A., 1953. Personality in Nature, Society, and Culture. Alfred A. Knopf, New York.

Leach, C.W., Van Zomeren, M., Zebel, S., Vliek, M.L.W., Pennekamp, S.F., Doosje, B., Ouwerkerk, J.W., Spears, R., 2008. Collective self-definition and self-investment: a two-dimensional framework of group identification. Journal of Personality and Social Psychology 95, 144–165.

Lewin, K., 1951. Field Theory in Social Science. Harper Row, London.

Loo, R., Kelts, P., 1998. A caveat on using single-item measures. Employee Assistance Quarterly 14 (2), 75–80.

Mavor, K.I., 2004. Religious Orientation, Social Identity and Attitudes to Homosexuality (Ph.D. Thesis). School of Psychology, The University of Queensland.

McGarty, C., 1999. Categorization in Social Psychology. Sage Publications, London, Thousand Oaks, New Delhi.

Mead, G.H., 1934. In: Morris, C.W. (Ed.), Mind, Self, and Society. University of Chicago Press.

Oakes, P.J., 1987. The salience of social categories. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Riecher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, pp. 117–141.

Oakes, P.J., Haslam, S.A., Turner, J.C., 1994. Stereotyping and Social Reality. Oxford, England: Blackwell.

Oakes, P.J., Turner, J.C., Haslam, S.A., 1991. Perceiving people as group members: the role of fit in the salience of social categorizations. British Journal of Social Psychology 30, 125–144.

Onorato, R.S., Turner, J.C., 2004. Fluidity in the self-concept: the shift from personal to social identity. European Journal of Social Psychology 34, 257–278.

Plaks, J.E., Shafer, J.L., Shoda, Y., 2003. Perceiving individuals and groups as coherent: how do perceivers make sense of variable behavior? Social Cognition 21, 26–60.

Phinney, J.S., 1992. The multigroup ethnic identity measure: a new scale for use with diverse groups. Journal of Adolescent Research 7, 156–176.

Postmes, T.T., Haslam, S.A., Jans, L., 2012. A single-item measure of social identification: reliability, validity, and utility. British Journal of Social Psychology 52 (4), 597–617.

Reicher, S.D., Haslam, S.A., 2006. Rethinking the psychology of tyranny: the BBC prison study. British Journal of Social Psychology 45, 1–40.

Reicher, S., Haslam, S.A., Spears, R., Reynolds, K.J., 2012. A social mind: an appreciation of John Turner's work and its influence. European Review of Social Psychology 23, 344–385.

Reynolds, K.J., Oakes, P., 2000. Variability in impression formation: investigating the role of motivation, capacity, and the categorization process. Personality and Social Psychology Bulletin 26 (3), 355–373.

Reynolds, K.J., Tindall, K., Subašic′, E., 2014. Norms and Behaviour Change: Building Conceptual Clarity to Move Forward, manuscript in submission.

Reynolds, K.J., Turner, J.C., 2006. Individuality and the prejudiced personality. European Review of Social Psychology 17, 233–270.

Reynolds, K.J., Turner, J.C., Branscombe, N.R., Mavor, K.I., Bizumic, B., Subašic′, E., 2010. Interactionism in personality and social psychology: an integrated approach to understanding the mind and behaviour. European Journal of Personality 24, 458–482.

Roccas, S., Sagiv, L., Schwartz, S., Halevy, N., Eidelson, R., 2008. Toward a unifying model of identification with groups: integrating theoretical perspectives. Personality and Social Psychology Review 12 (3), 280.

Sellers, R.M., Smith, M.A., Shelton, J.N., Rowley, S.A.J., Chavous, T.M., 1998. Multidimensional model of racial identity: a reconceptualization of African American racial identity. Personality and Social Psychology Review 2 (1), 18–39.

Stryker, S., Statham, A., 1985. Symbolic interaction and role theory. In: Lindzey, G., Aronson, E. (Eds.), The Handbook of Social Psychology. Random House, New York, pp. 311–378.

Swann Jr., W.B., Seyle, C., 2005. Personality psychology's comeback and its emerging symbiosis with social psychology. Personality and Social Psychology Bulletin 31, 155–165.

Tajfel, H., 1972. Social categorization. In: Moscovici, S. (Ed.), Introduction a la psychologie sociale, vol. 1. Larouse, Paris, pp. 272–302.

Thoits, P.A., Virshup, L.K., 1997. Me's and we's: forms and functions of social identities. In: Ashmore, R., Jussim, L. (Eds.), Self and Identity: Fundamental Issues, vol. I. Oxford University Press, Oxford, pp. 106–133.

Turner, J.C., 1982. Towards a cognitive redefinition of the social group. In: Tajfel, H. (Ed.), Social Identity and Intergroup Relations. Cambridge University Press and Paris: Editions de la Maison des Sciences de l'Homme, Cambridge, pp. 15–40.

Turner, J.C., 1984. Social identification and psychological group formation. In: Tajfel, H. (Ed.), The Social Dimension: European Developments in Social Psychology. Cambridge University Press, Cambridge, pp. 518–538.

Turner, J.C., 1985. Social categorization and the self-concept: a social cognitive theory of group behaviour. In: Lawler, E.J. (Ed.), Advances in Group Processes. JAI Press, Greenwich, Connecticut, pp. 77–122.

Turner, J.C., 1987. A self-categorization theory. In: Turner, J.C., Hogg, M.A., Oakes, P.J., Riecher, S.D., Wetherell, M.S. (Eds.), Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, pp. 42–67.

Turner, J.C., 1991. Social Influence. Open University Press and Pacific Grove, CA.: Brooks/Cole, Milton Keynes, England.

Turner, J.C., 2000. Social identity. In: Encyclopedia of Psychology, vol. 7. London Oxford University Press, pp. 341–343.

Turner, J.C., Oakes, P.J., Haslam, S.A., McGarty, C., 1994. Self and collective: cognition and social context. Personality and Social Psychology Bulletin 20, 454–463.

Turner, J.C., Onorato, R., 1999. Social identity, personality and the self–concept: a self–categorization perspective. In: Tyler, T.R., Kramer, R., John, O. (Eds.), The Psychology of the Social Self. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 11–46.

Turner, J.C., Reynolds, K.J., 2001. The social identity perspective in intergroup relations: theories, themes and controversies. In: Brown, R., Gaertner, S. (Eds.), Handbook of Social Psychology: vol. 4: Intergroup Processes. Blackwell, Oxford, UK and Cambridge, USA, pp. 133–152.

Turner, J.C., Reynolds, K.J., Haslam, S.A., Veenstra, K., 2006. Reconceptualizing personality: producing individuality by defining the personal self. In: Postmes, T., Jetten, J. (Eds.), Individuality and the Group: Advances in Social Identity. Sage, London, pp. 11–36.

Social Identity in Sociology

Kay Deaux, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Abrams, volume 21, pp. 14306–14309, © 2001, Elsevier Ltd.

Abstract

Social identities are definitions of self based on claimed membership in a social category. They incorporate not only cognitive meanings but also emotional significance and motivational implications. Important features of social identification include their multiplicity, the possibilities of change over time, and the range of behavioral consequences that are linked to the endorsement of an identity. Additionally, social identities serve as an important conceptual link between individual psychological processes and the situational and structural features of society.

Social identities are conceptual markers of categorical membership. They define the boundaries (however fuzzy) of a grouping of people and they implicate characteristics that members of that grouping are believed to share. Social identities not only provide cognitive meaning for those who hold them, but they also imply emotional investment in a position and the potential to take actions that are consequential for the identity. Although social identities may emerge from and be reflective of interaction with the group of people who share the identity, they do not require direct contact or interchange with all others who share the category membership. That is, a social identity is fundamentally a self-descriptive psychological position rather than a specific interacting group. Examples of social identities include gender, ethnicity, sexual orientation, nationality, religion, political party, and occupation. In each case, the social identity is a part of self-definition for the person who claims it.

The concept of social identity is ubiquitous in contemporary social science and in popular discourse as well. Within the social and behavioral sciences, theories and investigations of social identity are common to sociology, psychology, and political science, and have emerged in other disciplines, such as economics, as well. 'Identity politics' take the scholarly concept to the streets, grounding it in the day's diversity and civil rights debates while perhaps complicating its scholarly usage. Given this broad sweep, it is perhaps not surprising that some writers have advocated an elimination of social identity from the scholarly lexicon. However, the widespread use of the term suggests that the concept of social identity captures some important elements of human behavior that need more rather than less engagement.

Consider some of the ways in which social identity pervades the analysis of political and social behavior. Immigration debates often highlight issues of national and ethnic identity: for example, should immigrants reject the ethnic identity associated with their country of origin and replace it with the national identity of the new country? Perceived threats to religious identity are often at the heart of prolonged intergroup conflict. Political movements gain strength as more people identify themselves as feminists or peace activists or environmental Greens. In each case, the group label is not simply a way to make distinctions between groups of people, but it is a basis for those who claim the label as self-descriptive to define themselves, their agenda, and their mode of interaction with others.

Numerous theories of social identity have emerged over the years, ever since early scholars such as George Herbert Mead and William James staked out the territory a century ago. Not surprisingly, these theories and investigations reflect disciplinary predilections as well as salient social issues. Sociologists are likely to use symbolic interaction theory as a touchstone, exemplified by the work in a line that extends from Stryker (1980) to Burke and Stets (2009), with many points in between. Psychologists are more apt to turn to Tajfel's (1981) social identity theory and the subsequent developments of self-categorization theory (Turner et al., 1987). At the same time, the phenomena and concepts at issue do not adhere to strict disciplinary boundaries, with psychological process often of interest to sociologists and political scientists and psychologists paying attention to groups and social systems. This disciplinary fluidity reflects the pivotal role that social identity plays in thinking about social behavior, acting as the intermediate link between micro-level psychological processes and macro-level contextual and structural features. For many, the challenge is to integrate the insights from these multiple orientations in a way that can capture both individual and collective processes (Deaux and Martin, 2003; Jenkins, 2004).

My agenda here is not to argue for a particular identity or to sort out theoretical intricacies among theories. Rather, I take as a starting point the validity of social identity as concept and social psychological reality and I focus on four key principles that characterize these identities.

1. Social identities, as a form of self-definition and categorical membership, are inevitably related to but not identical in meaning or function with the categorizations used by others.
2. People typically have multiple social identities that they consider important to their self-definition, which can overlap or be independent of one another, and can be used in different circumstances.
3. Social identities are not static categories but are dynamic entities subject to change over time.
4. Social identities have motivational properties and social consequences.

Each of these principles is developed in more detail in what follows.

Insider and Outsider Perspectives on Social Identities

The categories that serve as the basis for social identities are social constructions that are used in society to create meaningful commonalities and distinctions among people. As such, the meanings associated with any specific category, for example, Australian, Buddhist, or Tea Party member, are to some degree shared by the person who identifies with the category and an observer who is describing someone in that category. Both perspectives draw from common cultural representations to define the identity category; moreover, even the choice of a particular social identity is shaped to a large degree by the set of social categories that is available in a given society at a given time.

Visibility of category markers, for example, skin color, facial features, or religious dress, increases the likelihood that a particular category will be prominent in the repertoire of the holder of the identity as well as the society at large. Further, to the extent that a particular category is salient, it is more likely to shape social interactions; these experiences shared by actor and observer in turn can create more commonality of meaning between the two perspectives. Phenotypic prototypicality, that is, the degree to which a person's physical appearance is considered to be highly representative of the group as a whole, can influence judgments of the importance of racial/ethnic identity, both by observers and by those holding the identity (Wilkins et al., 2010). African-Americans whose faces are considered to be more prototypical of their ethnic group, for example, are believed to be more strongly identified with their ethnic group. Similarly, African-Americans themselves identify more strongly with their ethnic group when their facial features are more prototypically African-American. We can assume that for these visibly categorized people, ethnic identity is more salient in daily interactions than it is for those whose racial markings are less obvious; in turn, greater exercise could result in more centrality accorded to the identity, relative to other identities that might be available but are less often engaged.

Although the visibility of category markers may increase the likelihood that the insider and outsider views of a social identity are similar, possibilities for divergent views abound. A second-generation Asian-American may resent assumptions that their Asian identity is dominant, for example, and questions such as 'Where are you from?' or 'Do you speak English?' can constitute threats to a strongly held identity as an American citizen (Cheryan and Monin, 2005; Wang et al., 2012). Although the inquiry may seem harmless to the asker, targets of the question can be offended and angered by the identity assumptions implied in the question.

Differences between the 'inside' and the 'outside' views of social identity are due in part to the different purposes that they serve. From the outside, a social category is a convenient and indeed unavoidable method of classification, a censuslike means of simplifying the social world by creating a relatively limited set of concepts to deal with a broad array of people and behavior. These beliefs are, in essence, the stereotypes that members of a community to a large degree share. They are a shorthand that sets a priority on simplicity, lack of ambiguity, and low variation within the category. Whereas social categories serve reductionist purposes from the outside, a social identity can be considerably more complex from the insider's perspective, both in the functions it serves and in the range of meaning associated with category membership.

Categorization and self-definition are the first step in social identification, but not the only basis for characterizing and differentiating social identities (Ashmore et al., 2004). As a consequence, people who nominally share the same categorical label, such as man, Moroccan, or Mormon, will not always act the same. Perhaps the best-documented dimension of identity is self-rated importance: how central or significant or strongly held is a social identity for the person? Not all people who claim a social identity are equally committed to that identity; it is the variations in rated importance that help us understand why members of the same category do not necessarily act the same way in what seem to be identity-relevant circumstances. The more strongly a social identity is endorsed, the more likely is the person to interpret events in terms of the socially defined group, to take action on behalf of the group, and to be threatened by situations that challenge the social identity. Beyond the importance dimension, social identities vary in other ways as well. Identities can be more or less embedded in social networks, more or less expressed in behavioral acts, and defined by a wide assortment and not necessarily shared web of experience and meaning. These psychologically based dimensions of difference can be critical in understanding if, when, and how people who have a particular social identity will take action relevant to that identity category.

Differences in meanings and functions of social identities between insider and the outsider views often lead to faulty assumptions. As noted earlier, the observer who infers nationality on the basis of Asian physical features is effectively denying an identity that the second-generation American citizen may hold strongly. Similarly, the observer who imputes meaning to a category that the identity holder does not share (e.g., all feminists burn their bras, all gays are promiscuous) creates problems for social interactions between the two. And yet, although insider and outsider views may be different, they are not completely independent of one another. Indeed, a person's perception of how others view their social group can be a source of self-definition (the classic image of a 'looking glass self').

When insider and outsider views are discrepant, the person whose identity is in question typically has the option of acting in accord with the views of others or diverging from that view to be more consistent with his or her own views (Barreto and Ellemers, 2003). Agreement with the beliefs of others might be chosen for instrumental reasons, for example some personal gain to be experienced from going along with the suppositions or some embarrassment or cost likely to be experienced if one were to diverge. Thus, pretending to be a strong organizational supporter may be preferable to being socially shunted, or staying 'in the closet' may at least temporarily avert criticism or rejection for the gay person interacting with friends or family who are unaware of the sexual preferences.

Choosing to be consistent with one's own identity classification and definition depends on the strength and importance of an identity to its holder, as well as on the circumstances in which identity expression takes place. To the extent that the views of others are unfavorable, as is often the case for minority and stigmatized groups in society, people are less prone to act

in accord with those views. In such instances, the 'looking glass self' may not be an appropriate metaphor, as members of the devalued groups will disregard the views of others in service of maintaining and promoting self-esteem (e.g., Perkins et al., in press).

In sum, the apparent simplicity of social categories and the seeming inevitability of their use can mask the complexity of their construction and the variability of their usage. Further, while insider and outsider views of a social identity category may overlap and interact, the analysis of social identity and its behavioral outcomes requires maintaining a clear distinction between the two.

Multiplicity of Identities

Social identities are defined in terms of identification with and/or membership in specific social groups and categories. Given the complexity of contemporary society, people are likely to have more than one meaningful identity, an assumption that dates back nearly a century to the influential theorizing of the sociologist George Herbert Mead and the psychologist William James. Multiple categories of group membership are instantiated in census reports and survey structures as well as reflected in common discourse. Thus, people can be characterized in terms of their gender, their ethnic or national group, occupation, sexual preference, age, or political position, and numerous other often-used bases of demarcation. Not everyone thinks of him- or herself in terms of each of these possible membership categories; yet most people readily claim to identify with several of them and see these identities as having distinctive features.

Accepting the assumption of multiplicity raises questions as to how various identities might be combined and/or integrated. From the outsider perspective, two or more categories can be combined to create an analytic subcategory, for example, African-American men with college degrees, which in turn can be linked to outcomes of interest, for example, occupational attainment. This approach, however, does not speak directly to the insider process of social identity construction. A more qualitative approach to multiple identities, often represented by the term *intersectionality*, puts its emphasis on the unique characteristics of the combination of identity categories (Crenshaw, 1991; Stewart and McDermott, 2004). By this line of thinking, the juxtaposition of, for example, gender and ethnicity creates emergent properties that are not part of either constituent category on its own. Thus, the experience of being a Latina cannot be fully described by adding together typical descriptors of 'woman' and 'Latino.' Similarly, Muslim women have experiences that are unique to the category combination, rather than determined by religion or gender alone (Deaux and Greenwood, 2013).

Multiplicity of identity and the possible forms of integration are central concerns in the immigration debates, typically focusing on whether identities such as ethnicity or religion can effectively coexist with a national identity, or, alternatively, whether the former must be jettisoned and replaced by a primary identification with the nation state. From a psychological perspective, the concept of biculturalism or multiculturalism is widely accepted, with current discussions centering not on whether a person is capable of holding two or more identities but rather in detailing the various forms that such combinations might take and the social conditions that influence the choices that people are likely to make (Benet-Martínez, 2012; Berry, 1990; Deaux and Verkuyten, 2014; Hong et al., 2000; LaFromboise et al., 1993). In the political realm, in contrast, the possibility of combining identities is often questioned or emphatically denied. In Europe, opponents of multiculturalism often claim that being Muslim is incompatible with a national identification; in the United States, the argument is that being American requires letting go of Latino culture and Spanish language. In part, these different positions reflect the insider–outsider distinction noted earlier, as outsiders simplify the categorization while insiders experience greater complexity and variation. Beyond that distinction, however, the highly charged environment of immigration debates today sets an agenda that is quite detached from the relevant scientific data.

Another question raised by the assumption of multiple identities is how to predict which identities might be more salient and more likely to guide behavior on what occasions. In other words, if one has a choice to make when presenting the self in social interactions, what guides that choice? Several factors contribute to the selection. First, an identity is more likely to guide behavior if it is more important to the individual (Wiley and Deaux, 2011); for example, a strongly identified feminist is likely to view many situations in terms of their possibilities for expressing feminist thought and action, whereas a less strongly identified person might be more selective in the occasions on which feminism is expressed. Immediate circumstances are a second influential factor: similarity of a particular identity with the perceptions of others in the situation (termed the comparative fit) or the normative demands of a situation can increase the salience of a particular identity such that most people who have the relevant identity will act in ways consistent with it. (Think of students and professors in a classroom, or those with a religious identity while in a church, synagogue, or mosque.) The power of situational cues to influence identity salience and in turn affect behavior is dramatically seen in the classic work of Shih et al. (1999), wherein Asian-American women performed better on a math test when their Asian identity was salient and worse when their female identity was salient, consistent with stereotypes about the superior or inferior mathematical skills of Asians and women, respectively.

A third consideration in the choice of identity expression is the strategic concerns of the individual, taking into account the views of the perceived audience and motivated by specific interaction goals, such as being recognized or accepted by the group. From this perspective, people are actors who can consciously (and perhaps unconsciously as well) choose what aspect of themselves to present and will vary their choices depending on their own objectives and their reading of the audience they are facing (Klein et al., 2007; Wiley and Deaux, 2011). As an example, when an Asian-American student is asked by a white investigator whether he or she speaks English, thus questioning the student's American identity, the student will, when given the opportunity, indicate a preference for American over Asian food and recall more American TV shows as an index of their cultural

knowledge (Cheryan and Monin, 2005; Guendelman et al., 2011). In other cases, a person may avoid presenting an identity, assuming the audience will not accept them in that guise, as when a Dominican American immigrant says, "I never say I'm American just because people are not going to believe you are American when you have this whole Spanish look" (Itzigsohn, 2009: p. 133).

The reality of multiple audiences and a charged political atmosphere that intensifies either-or thinking makes the choices of those who value multiple identities much more difficult. Presentational choices made with one audience in mind may cause difficulties when another audience becomes salient and conflict may be avoided only in places where other bicultural people confirm a shared, if complex, reality.

Changes in Social Identities

Some ascribed identity categories, such as gender or race, are generally constant across time, but many other social identities can and do change over the life course. At the most basic level of category inclusion, a person can shift from being a student to being a professor, or from being married to divorced. People also change occupations, join political movements, or assume a new nationality after immigration. These are the most obvious forms of identity change, in which the basic set of memberships is added to or subtracted from. Yet, because social identities are best conceived as having multiple defining elements (Ashmore et al., 2004), all identities can change in ways that are highly significant to the individual but perhaps not evident to the casual observer who uses categorical membership alone to characterize identity.

As one example, consider the shifting meanings – and labels – that have been used by Americans of African descent to describe themselves and to be described by others over the course of the US history. As Philogène (2001) has detailed, the terminology has evolved from Slave to Colored to Negro, terms that all originated in the dominant White discourse, to more recent usage of the terms Black and African-American. Moreover, these changes in social identity were not simply labels, but also involved reformulations of the meanings associated with category membership. Whereas Black was an identity based on racial dichotomization in the United States, that is, Black vs White, the connotations of African-American gave cultural meaning to the identity category and put the group on a par with other ethnically defined groups in the United States.

Why might a person's social identity change, either by shifts in subjective importance or meaning or in more fundamental ways such as the incorporation or rejection of an entire definitional category? Both internal and external pressures can be the impetus for change. Subjective goals and aspirations often lead a person to adopt new social identities, as an occupation or activity group becomes definitional and shared with similar others; conversely, shifting interests can lead to elimination of a previously held identity. Increases in the importance or priority of one identity can decrease the importance of another (although no upper limit on the number of social identities has been established). The acquisition of new information can change the meanings associated with an identity. External forces are also influential. Opportunity structures may change so as to allow a person to take on an identity previously not available, as when women are allowed in the military or immigrants take up residence in a new country. Experiences of discrimination can intensify a person's identification with the stigmatized category (supported by research on what has been termed the rejection-identification hypothesis), as well as change or add to the meaning of the identity itself. Alternatively, perceived loss of status among members of a high-status group can create or intensify identification with a social identity not previously acknowledged. An example of the latter process can be found in the increasing importance attached to white identity among some groups in the United States in the face of changing demographics (Knowles and Peng, 2005; McDermott and Samson, 2005).

More longitudinal work is needed to clarify the parameters of identity change, particularly in the subjective elements that vary among people who share a common category. The evidence is already quite clear, however, that social identities are dynamic rather than static in nature, and that recognition of the possibilities for change is essential to understanding the links between identities and subsequent actions.

Social Identity: Actions, Reactions, and Consequences

Social identities are more than self-descriptive characterizations and a means of positioning oneself in the larger social world. Their considerable importance as a conceptual tool for social scientists stems from their motivational properties, leading to predictable actions and reactions that serve and further a person's goals and a group's impact. People whose social identities are threatened in some way have emotional reactions, engage in cognitive thought processes, and often take actions to avoid or to challenge the threat. People who are stigmatized or devalued because of their group membership often perform less effectively in circumstances where they are being tested. Alternatively, unfair treatment based on group membership can promote collective action on behalf of the group to improve conditions and redress injustice.

Not all who characterize themselves in terms of a particular identity will necessarily take the same actions, nor even take any action at all. The most reliable single indicator of the likelihood of action, involvement, or emotional and cognitive reactions in identity-relevant conditions is the subjective importance of the identity (although more fine-grained predictions are possible when the assessment of self-investment in an identity is further dimensionalized, see Leach et al., 2008). Subjective importance of an identity as a determinant is not independent of contextual influences, of course. Further, because people typically have multiple identities, some of which are equally important to them, particular characteristics of the audience may lead to the selection of one identity over another (Wiley and Deaux, 2011). Turkish/Dutch biculturals, for example, have been shown to emphasize a Turkish identity to a Turkish audience, while presenting a Dutch face to a Dutch audience (Barreto et al., 2003). Such alterations in performance should not be taken as evidence of inconsistency or identity confusion, but rather a demonstration of motivated selective identity enactment.

Areas in which social identities have consequences are extensive, covering a range of individual, intragroup, and intergroup domains. Here I consider just two from this broad array: academic performance under conditions of threat to a social identity; and collective action as both cause and consequence of social identity.

Social Identity and Stereotype Threat

People are not oblivious to the negative views of others. When those negative views are associated with a social category with which one identifies, performance in domain-relevant arenas can be affected, to the detriment of the person holding that identity. This phenomenon, known as stereotype threat (Steele, 2010), has been demonstrated over a variety of identity categories, including race, gender, age, and social class. Groups that perform comparably when stereotypes are not salient, for example, blacks and whites on academic tasks and women and men on math tests, show significant differences when stereotypes about the group are salient (e.g., blacks are academically inferior to whites, women are worse than men in math). Laboratory studies have delineated the conditions and causal factors at work in these situations; fieldwork has shown the impact of stereotype vulnerability on the academic performance of Black and Latino students (Massey et al., 2003) as well as the successful implementation of interventions to counteract stereotype threat effects (Cohen et al., 2012).

Collective Behavior

Classic theories about crowd behavior, dating from the flawed but influential work of Gustave Le Bon, assume that participation in mass movements leads to a loss of individual identity and loss of rationality and control as well. Although this dehumanizing view might appeal to those in power who want to justify their higher status position, close examination of many protest movements suggests far more rationality and focus than Le Bon allowed. As Reicher and Drury (2011) have argued, participation in group action is a process of social identification, involving transformations in self-conceptualization from individual to group, as well as establishing social relations with others who share the social identity. Thus it may be an individual calculation of needs and potential benefits that motivates the initial involvement in group action, but a likely by-product of involvement is taking on a collective identity that can guide and propel further activity. Collective identities, once established, act as important motivators of involvement in social movements, both for groups attempting to gain recognition in society (e.g., gay rights movements, 'gray panthers' groups representing the concerns of older people) as well as established groups working for redistribution of resources (e.g., labor unions, farmers' movement) (Simon and Klandermans, 2001; Simon, 2011).

Collective identification is thus both cause and consequence of sociopolitical activity aimed at improving the conditions and position of a group within the larger society. Social identities are not the only factor in describing either the onset or the outcomes of this activity. They are, however, inescapably involved in understanding both the antecedents of collective action (e.g., perceived injustice in the treatment of one's group, ideological discontent) and the consequences (e.g., feelings of group efficacy, concrete changes in social policy).

Concluding Comments

Social identities emerge from a process of self-categorization, in which people define themselves in terms their membership in a shared social group or category. A social identity is much more than a categorical label, however, as it includes emotional investment in the category, shared meaning systems, and a potential for action consistent with and motivated by the identity. Consequently, to tap the explanatory power of a social identity, it is necessary to unpack the identity category, looking first at the importance the identity has for any particular person and then further at how other dimensions and meanings are implicated.

Social identities also need to be understood in terms of their potential for change over time (both as discrete categories and in terms of more nuanced and less visible changes in importance and meaning) and their flexibility of use, according to situational demands and opportunity structures. Social identities are important not because they are convenient markers of a particular social position, but rather because they have explanatory value in understanding individual variation in a wide range of thoughts, emotional reactions, and behaviors. They are, in essence, an indispensable link between individual psychological process, social interaction, and social systems.

See also: Collective Action; Collective Behavior, Social Psychology of; Collective Behavior, Sociology of; Ethnic Identity, Psychology of; Ethnic, Racial, and Nationalist Movements; Identity Movements; Identity and Identification, Social Psychology of; Identity in Anthropology; Religion: Nationalism and Identity; Self-Categorization Theory; Social Identity in Social Psychology.

Bibliography

Ashmore, R.D., Deaux, K., McLaughlin-Volpe, T., 2004. An organizing framework for collective identity: articulation and significance of multidimensionality. Psychological Bulletin 130 (1), 80–114.

Barreto, M., Ellemers, N., 2003. The effects of being categorized: the interplay between internal and external social identities. European Review of Social Psychology 14, 139–170.

Barreto, M., Spears, R., Ellemers, N., Shahinper, K., 2003. Who wants to know? The effect of audience on identity expression among minority group members. British Journal of Social Psychology 42, 299–318.

Benet-Martínez, V., 2012. Multiculturalism: cultural, social, and personality processes. In: Deaux, K., Snyder, M. (Eds.), The Oxford Handbook of Personality and Social Psychology. Oxford University Press, New York, pp. 623–648.

Berry, J.W., 1990. Psychology of acculturation. In: Berman, J.J. (Ed.), Cross-Cultural Perspectives: Nebraska Symposium on Motivation, vol. 37. University of Nebraska Press, Lincoln, Nebraska.

Burke, P.J., Stets, J.E., 2009. Identity Theory. Oxford University Press, New York.

Cheryan, S., Monin, B., 2005. "Where are you really from?": Asian Americans and identity denial. Journal of Personality and Social Psychology 89 (5), 717–730.

Cohen, G.L., Purdie-Vaughns, V., Garcia, J., 2012. An identity threat perspective on intervention. In: Inzlicht, M., Schmader, T. (Eds.), Stereotype Threat: Theory, Process, and Application. Oxford University Press, New York, pp. 280–296.

Crenshaw, K., 1991. Beyond racism and misogyny: black feminism and 2 live crew. In: Cohen, C.I., Jones, K.B., Tronto, J.C. (Eds.), Women Transforming Politics: An Alternative Reader. New York University Press, New York, pp. 549–568.

Deaux, K., Greenwood, R.M., 2013. Crossing borders: intersectional excursions into gender and immigration. In: Ryan, M.K., Branscombe, N.R. (Eds.), The Sage Handbook of Gender and Psychology. Sage Publications, London, pp. 234–250.

Deaux, K., Martin, D., 2003. Interpersonal networks and social categories: specifying levels of context in identity processes. Social Psychology Quarterly 66, 101–117.

Deaux, K., Verkuyten, M., 2014. The social psychology of multiculturalism: identity and intergroup relations. In: Benet-Martínez, V., Hong, Y. (Eds.), The Oxford Handbook of Multicultural Identity. Oxford University Press, New York, pp. 118–138.

Guendelman, M.D., Cheryan, S., Monin, B., 2011. Fitting in but getting fat: identity threat and dietary choices among U. S. immigrant groups. Psychological Science 22 (7), 959–967.

Hong, Y.Y., Morris, M.W., Chiu, C.Y., Benet-Martínez, V., 2000. Multicultural minds: a dynamic constructivist approach to culture and cognition. American Psychologist 55, 709–720.

Itzigsohn, J., 2009. Encountering American Faultlines: Race, Class, and the Dominican Experience in Providence. Russell Sage Foundation, New York.

Jenkins, R., 2004. Social Identity, second ed. Routledge, London.

Klein, O., Spears, R., Reicher, S., 2007. Social identity performance: extending the strategic side of SIDE. Personality and Social Psychology Review 11, 1–18.

Knowles, E., Peng, K., 2005. White selves: conceptualizing and measuring a dominant-group identity. Journal of Personality and Social Psychology 89 (2), 223–241.

LaFromboise, T., Hardin, L.K.C., Gerton, J., 1993. Psychological impact of biculturalism: evidence and theory. Psychological Bulletin 114 (3), 395–412.

Leach, C., van Zomeren, M., Zebel, S., Vliek, M.L.W., Pennekamp, S.F., Doosje, B., Ouwerkerk, J.W., Spears, R., 2008. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. Journal of Personality and Social Psychology 95 (1), 144–165.

Massey, D.S., Charles, C.Z., Lundy, G.F., Fischer, M.J., 2003. The Source of the River. Princeton University Press, Princeton.

McDermott, M., Samson, F.L., 2005. White racial and ethnic identity in the United States. Annual Review of Sociology 31, 245–261.

Perkins, K., Wiley, S., Deaux, K. Through which looking glass? Distinct sources of public regard and self-esteem among first- and second-generation immigrants of color. Cultural Diversity and Ethnic Minority Psychology, in press.

Philogène, G., 2001. From race to culture: the emergence of African American. In: Deaux, K., Philogène, G. (Eds.), Representations of the Social: Bridging Theoretical Traditions. Blackwell Publishers, Oxford, UK, pp. 113–128.

Reicher, S., Drury, J., 2011. Collective identity, political participation, and the making of the social self. In: Azzi, A.E., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies: A Multidisciplinary Perspective. Wiley-Blackwell, Chichester, Great Britain, pp. 158–175.

Shih, M., Pitinsky, T.L., Ambady, N., 1999. Stereotype susceptibility: identity salience and shifts in quantitative performance. Psychological Science 10, 80–83.

Simon, B., 2011. Collective identity and political engagement. In: Azzi, A.E., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies: A Multidisciplinary Perspective. Wiley-Blackwell, Chichester, Great Britain, pp. 137–157.

Simon, B., Klandermans, B., 2001. Toward a social psychological analysis of politicized collective identity: conceptualization, antecedents, and consequences. American Psychologist 56 (4), 319–331.

Steele, C., 2010. Whistling Vivaldi. Norton, New York.

Stewart, A.J., McDermott, C., 2004. Gender in psychology. Annual Review of Psychology 55, 519–544.

Stryker, S., 1980. Symbolic Interactionism: A Social Structure. Benjamin Cummings, Menlo Park, CA.

Tajfel, H., 1981. Human Groups and Social Categories. Cambridge University Press, Cambridge, UK.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-Categorization Theory. Blackwell, Oxford, UK.

Wang, J., Minervino, C., Cheryan, S., 2012. Generational differences in vulnerability to identity denial: the role of group identification. Group Processes & Intergroup Relations 20 (10), 1–18.

Wiley, S., Deaux, K., 2011. The bicultural identity performance of immigrants. In: Azzi, A., Chryssochoou, X., Klandermans, B., Simon, B. (Eds.), Identity and Participation in Culturally Diverse Societies: A Multidisciplinary Perspective. Wiley-Blackwell, Chichester, UK, pp. 49–68.

Wilkins, C.L., Kaiser, C.R., Rieck, H., 2010. Detecting racial identification: the role of phenotypic prototypicality. Journal of Experimental Social Psychology 46, 1029–1034.

Social Inequality and Schooling

Marie Duru-Bellat, Sciences-Po, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Barnhouse Walters, volume 21, pp. 14309–14313, © 2001, Elsevier Ltd.

Abstract

Social inequality in education is universally observed in countries in which children are brought up in unequal families; they enter school with unequal linguistic, cognitive, and cultural assets and attain unequal levels of achievement. Moreover, they make unequal choices, leading to unequal levels of education and social reproduction. However, the latter varies across countries, demonstrating that the way school itself is organized matters. The timing of the selection process and the degree of segregation within the system are especially important, along with the student social mix they generate. As countries implement policies to fight educational inequality, simply expanding education does not prove efficient since it only delays social selection. More qualitative policies then seem necessary, such as limiting early inequalities and focusing on disadvantaged families, which in turn requires broader social policies.

In democratic societies, education-based meritocracy is supposed to prevail. Since it is now considered illegitimate to attain social status by birth, everyone is meant to achieve it through his or her own efforts and merit, displayed by unequal degrees. Education is the main ideological tool used to justify the remaining inequalities in societies where all individuals are considered equal. So it is crucial that equality of opportunities be guaranteed through school competition. The school system is supposed to level the playing field between students so that no inequalities occur based on social background or other ascribed characteristics such as race or gender. The sociology of education shows that such is not the case. Rather, with students from unequal families, the school system proves unable to counter the inequalities visible at the start and, moreover, often allows them to widen. As a result, in recent decades, sociologists have repeatedly demonstrated that schooling plays a pivotal role in the reproduction – and legitimization – of class privilege and social inequality. However, this is not an inevitability; over the last decade, an increasing number of international comparisons have demonstrated that some countries do indeed succeed in reducing this gap through efficient policies.

Social Background and Persistent Educational Inequality

Both sociological researchers and a variety of international organizations, such as the Organisation for Economic Co-operation and Development (OECD), are quite active in producing figures and analysis regarding the social inequality of achievement in the richest countries. Although one is not discussing the world's poorest countries here, one can nevertheless note that in the said countries, it is not primarily achievement at school that makes a difference but rather access to education, given that schooling is far from universal at both the primary and secondary levels. Moreover, a seminal, albeit slightly dated, study by Heyneman and Loxley (1983) suggested that in the poorest countries, evidence of achievement inequality is seen less in the diversity of student population than in the quality of the schools attended. Thus, the issue of social inequality in schooling would be more relevant in the richest countries where (1) school attendance is nearly universal, and (2) the quality of that which both families and schools offer remains more or less heterogeneous from one student (or site) to another.

Large Social Gaps

A variety of indicators and case studies exist regarding social inequalities of achievement (Tesse et al., 2007). They allow us to observe the percentage of the subgroups attaining some level of education or its correlative degree. For example, in France (one of the least egalitarian countries in Europe along with those in central and eastern Europe), the probability of achieving the 'baccalauréat' (the exit exam at the end of upper secondary school) varies from 88% for the most socially privileged students to 49% for students from working-class backgrounds. Moreover, educational inequalities also take on a qualitative form, with some social specialization into different tracks. Today, among those students who reach the final year of upper secondary school, 45% of socially privileged children are in the scientific track compared to about 20% of students from working-class backgrounds. The latter are much more likely to pursue vocational degrees, thus reducing their chances of pursuing tertiary education.

However, probably due to the dissemination of Rawls' conception of justice, it has been judged increasingly relevant to focus not on the highest level of education achieved but rather on the shared basic level of knowledge and skills needed for one to make a good living. Combined with a growing concern for benchmarking that relies on up-to-date country comparisons, this new trend has resulted in the development of comparative studies focusing on knowledge acquired. The most famous study of this kind is the Programme for International Student Assessment (PISA). It is an internationally standardized assessment implemented by the OECD and administered in schools to 15-year-olds, every 3 years from 2000 (see PISA Technical Report and PISA Data Analysis Manual, 2nd edition, 2009). In 2000, 43 countries were included in the first assessment, and in 2009, 62 countries were included in the fourth assessment. That year, between 4500

and 10 000 students in each country were tested on their native language, mathematics, and science. The aim of the exercises was to assess life skills and not just pure conformity to academic knowledge, because the ability to cope with life is the final objective of schooling rather than the mastery of pure academic content.

These surveys provide a solid empirical basis by which to assess social inequality in achievement, at least in terms of the life skills that schooling imparts to students. A gap between students from different social backgrounds is observed in every country, amounting sometimes to the equivalent of two grade levels. However (according to the 2009 data, see OECD, 2010a), this gap is greater in countries such as France, New Zealand, Belgium, Germany, Chile, among others. Conversely, the gap is much smaller in other countries such as Canada, Finland, Italy, Japan, and South Korea.

Moreover, in many countries, first-generation immigrant students are more likely to be poor performers; on average, they are about twice as likely to be found in the lowest quartile of students. In addition, immigrant students tend to be disadvantaged socioeconomically, thus giving them double handicap. Taking into account their socioeconomic status, the specific impact of an immigrant background remains, confirming the fact that some part of those students' difficulties can be attributed to their immigrant status. This is especially the case in Finland, Germany, and Italy, while it matters less in Canada and the United Kingdom.

With regard to gender, in OECD countries (and in a growing number of countries around the world, with the exception of Africa and South Asia), girls today are achieving higher levels of education. Over all OECD countries, 41.5% of women aged 25–34 years receive a tertiary degree, compared to only 32.7% of their male counterparts. In all OECD countries, they also achieve significantly better scores in PISA literacy tests. The male students in every country achieve better scores in mathematics, but the gap is much smaller than in literacy, and in many countries, the difference is insignificant.

Basic Mechanisms

Although PISA surveys focus on inequalities of achievement, research shows that this is not the only process by which social inequality is generated. Actually, two types of mechanisms that result in two types of inequalities should be distinguished: educational performance on the one hand, and educational choices and ambitions on the other. After taking into account class differences in achievement, the latter mechanism results in persistent social inequalities. These mechanisms are labeled as primary and secondary effects. Primary effects result from the accumulation of social differences in cognitive abilities that are visible when a student first starts school. Secondary effects result from choices and strategies observed given the same level of performance, even if they are never completely independent from the student's level of achievement. Social inequalities then become more pronounced whenever making a choice (choice of subject or track) is required. Actually, choices bring about new specific inequalities: whatever their level of achievement, students do implement varying degrees of self-selection; for instance, average students from lower social backgrounds are unlikely to make choices that will place them in the more demanding tracks, while average students from higher social classes will make those choices.

More broadly, at every branching point, students from the lower classes require a higher probability of success than children of higher social origin before deciding to continue schooling. They are not only more risk adverse but also less motivated since they are able to ensure class maintenance (i.e., achieve the same social position as their father) with a lower level of education, as some sociologists maintain. Boudon (1974) asserts that at every branching point, students and their families make choices based on the future they desire, and by comparing anticipated benefits with risks of failure and costs. They develop strategies based on the unequal resources they possess. Parents from advantaged backgrounds want their children to achieve at least their same career level, which in general requires lengthy schooling. At the same time, they tend to be less risk adverse and the cost issue weighs less in their choice. Even when their child is not brilliant, they will choose the most promising tracks despite the risks in terms of success. Conversely, families from less advantaged backgrounds will be more cautious. Thus, even when the students' academic level is the same, strict self-selection based on social background comes into play. In most countries, these secondary effects play a significant role in social inequality; for example, they account for about 25% of the social selection at the A-levels in Great Britain (Breen and Jonsson, 2005).

The Role of Family

Given that children are raised in unequal environments and in families that possess unequal cultural and economic resources, it is not very surprising to observe such social inequality in schooling. One expects that parents will strive to transfer their knowledge, tastes, and assets to their children so as to secure their own future as well as to reproduce their social status. The French sociologist Bourdieu (with Passeron, 1990) designed a theory that focused on the concept of cultural capital. He maintained that students have varying degrees of familiarity with the dominant culture that school is supposed to impart because they inherit unequal cultural capital from their parents. Students possess specific assets such as cultural codes, modes of conduct, use of language, and also stronger motivations; this is why certain students – the 'inheritors' – perform better at school. However, the British sociologist Sullivan (2000), after achieving a concrete operationalization of the concept of cultural capital, showed that the cultural capital acquired through participation in formal cultural activities (i.e., the theater, art galleries, and concerts) does not significantly alter the development of knowledge and skills that help students to better succeed at school. Students' reading and television habits (when children watch relatively sophisticated programs) prove more important, as they foster intellectual resources that are advantageous at school (assessed early in lower secondary school).

Moreover, the contents of education itself matter. Sociologists such as De Graaf et al. (2000) have suggested that social inequalities may be wider in France and Eastern European countries because they give more importance to the humanities (or highbrow culture), thus giving an advantage

to those students whose parents are well educated and possess this kind of capital. However, this is less the case in countries with curricula geared more toward everyday life skills.

However, according to Bourdieu and Passeron's theory, the education system is not a neutral field divorced from society as a whole. Rather, the culture and standards within the education system reflect society and, in doing so, schools systematically benefit those who have already acquired cultural capital through their family and other social networks. Higher class children easily fit into school culture. Moreover, because the education system is portrayed and widely perceived as being open to all on the basis of talent, many working-class children come to see themselves as intellectually inferior; they accept that they, rather than the system itself, are to blame for their failure. In this way, the education system plays a key role in the cultural reproduction of social inequalities; educational credentials help to reproduce and legitimize social inequalities, as higher class individuals are seen as deserving of their place in the social structure.

Because they were focused on the social function of schooling, these sociologists did neglect the possibility that social inequalities in achievement may also result from true cognitive weaknesses that exist when students enter school. A number of studies (see, for instance, Mercy and Steelman, 1982) show that as early as ages 2 and 3 years, children are already on unequal footing in terms of mastering of language, time, and space, which in turn plays a role in them learning how to read, etc. In the 1970s, the British sociologist Bernstein (1975) while investigating the way in which education reproduces class inequalities produced an analysis of linguistic skills. He argued that children from varying backgrounds develop different language codes (or forms of speech) during their early lives, which in turn affect their subsequent school experiences. His interest was in the systematic differences between the ways in which language is used. He distinguished between a restricted code (a code containing many unstated assumptions that speakers expect others to know, thus making it difficult for nonmembers of the speaker's group to understand) and an elaborated code (a style of speaking in which the meanings of words can be individualized to suit the demands of particular situations). Bernstein proposed that children who have acquired elaborated codes of speech are more able to deal with the demands of formal academic education and easily fit into the school environment. Thus, language is a vehicle of social inequalities between students.

Psychologists show that, across the board, cognitive skills and, more broadly intelligence, develop through social interactions, and that parental practices in a child's daily life do affect his or her cognitive development. Just as disparities in living conditions give rise to unequal health conditions, it has been demonstrated that certain forms of socialization do indeed cause some advantages or disadvantages in children. In particular, adaptable and flexible educational practices (children learn both rules and exceptions that may be justified), and encouraging children to freely explore their surroundings and consider the results of their own actions all have a positive effects on cognitive development. In addition, as parents teach their children that they are personally responsible for what happens to them, they foster self-confidence. The most privileged parents implement what the American sociologist Lareau (2011) calls a 'concerted cultivation,' thus maximizing their children's opportunities to learn and develop, while the least privileged are less active, letting a more 'natural growth' occur.

Those attitudes and practices are linked to school success later on, given that the most favorable practices are more widespread among educated or sufficiently well-off families. That is an important vehicle of social inequality, especially given that school is unsuccessful in compensating for early inequalities and in leveling the playing field. Moreover, as evoked, social differences in risk aversion and educational choices do widen the social gap that results from parents' educational practices. However, with the growing prevalence in sociology of a less deterministic conception of social phenomena and a more assertive conception of the actor, more attention has been given to the context in which these social inequalities are generated.

School Matters

The school, the teachers and their pedagogical practices, the way students are grouped in classes, and, more broadly, the organization of the education system itself, all these contextual dimensions create the framework in which students behave. As Coleman et al. (1966) showed a half century ago, family factors are generally more important predictors of academic success than school factors; however, that is true mainly in countries in which a minimum level of quality of these school factors is warranted. Moreover, school factors are things that are amenable to social policy interventions in ways that family factors are not.

The Story Told by International Comparisons

While the relationships between students' socioeconomic background and achievement prove constant across countries, they also vary in extent. That is, of course, related to the level of socioeconomic inequalities prevailing within the countries themselves. Nordic countries have both low income inequality and equal performances, whereas many English-speaking countries have both large income inequality and large performance inequality. However, the correlation does not hold invariably. PISA data show (in a rather constant way since the 2000 study) that countries where socioeconomic inequality between families is similar may have a varying degree of social disparities in achievement across students. Both New Zealand and Japan have rather wide gaps in income inequality, but educational inequality is much greater in the former that in the latter. The same is true for Belgium and Finland, where income inequality is rather moderate but, once again, educational inequality is much greater in the former than in the latter. In some countries like Australia, both strong income inequality and relatively weak educational inequality exist; conversely, in other countries like Germany, there exists both a great educational inequality and a rather moderate income inequality.

International comparisons are here very heuristic (Janmaat et al., 2013; see also OECD, 2005, 2010a,b). They suggest that some educational systems are indeed capable of compensating for social inequalities, with students proving less unequal

academically than the families they belong to, while others seem to augment the impact of social origin on student performance. Before investigating which characteristics of the education systems may explain this variation, it is to be stressed that, contrary to the widespread common view, there is no trade-off between students' average level and the amount of social inequalities, since the mean performance level is high in the most equal countries (Finland, Japan, and South Korea), even though it is also quite high in some unequal countries (Canada and New Zealand).

Comparisons (such as those led by Green and others on the 2006 data set; see Green et al., 2006) show that all the more equal countries (East Asian and Nordic countries) have comprehensive systems (at the lower secondary level), while among the unequal group (Anglo-Saxon and German countries), 5 of the 10 systems are selective at the secondary level, and four are highly marketized comprehensive systems. More precisely, what the more equal countries have in common, and what is absent in the less equal countries, are structures and processes typically associated with radical versions of comprehensive education: nonselective schools, mixed-ability classes, late subject specialization and measures to equalize resources between schools. All these features contribute to lowering educational inequality.

Another factor that may explain these differences in social inequality is the school intake mix. Measured in terms of average parent job level, wealth, and cultural capital, it has a major impact on the performance of individual students internationally. PISA data show that in English-speaking countries, because the social mix is much more diverse (it varies greatly from one school to another), the school intake mix has a much larger impact on the differences in outcomes than elsewhere. For all OECD countries taken together, school social mix explains 34% of the variance in average outcomes in literacy between schools, whereas in English-speaking countries, its impact is much higher; it explains 64% of the between-school variance in Australia, and 61% in the United States and the United Kingdom. Conversely, in Japan and South Korea, the impact of school status on differences in school outcomes is much lower, explaining 11 and 17% of between-school variance, respectively. Intake effects are also moderately low in Scandinavian countries. Thus, just how much the school attended matters varies across the countries under comparison, and it appears that greater social segregation in school intakes increases educational inequality, whereas increasing the heterogeneity of intakes and narrowing the gaps between schools reduces it. Nordic and East Asian countries appear to have considerably less segregation in school intakes than most of the English-speaking and German-speaking countries and the effect of intake differences on inequalities between schools is much lower.

All in all, while it has been shown that international difference in student performance cannot be attributed to resource differences (Wössman, 2003), the way education systems are organized certainly matters.

The Impact of Organizational Characteristics

Educational research illuminates the ways in which some characteristics of education systems (e.g., the age at first selection, or branching point, the length of the common curriculum, and the social segregation of schools, strongly correlated with the existence of distinct tracks) magnify social inequalities. One reason why is that, in accordance with Boudon's theory, systems with several strategic choices give more latitude to inequalities resulting from the different trade-offs between the costs, risks, and expected benefits that unequal families make for their children.

However, the fact that the systems are differentiated may have an impact in quite a different way, thus putting academic learning at stake. In countries where early selection prevails, children from less advantaged groups tend to attend the weakest tracks and schools. In these settings, the disciplinary climate and teaching conditions are less favorable and the curriculum is less ambitious; fewer resources may also be allocated to these schools. As a result, the gap between levels of academic achievement within such systems widens just as it does between different social groups.

Another factor concerns the impact of the social mix, with the consequence that social inequalities will be larger in countries where the social mix largely differs from one school to another. This is because the composition of the student body itself contributes to creating an environment of inequality. Classmates are resources for one another (the so-called peer effect), meaning that student achievement depends positively on the average achievement in the reference group (the probability that an individual behaves in a certain way increases with the presence of this behavior in the group).

The social mix also impacts the climate of the class and influences which teaching practices are possible. For example, much research shows that students from working-class backgrounds attending a mixed-intake school progress better. Because those students have contact with students who are more adjusted to school norms and have greater cultural resources, they draw benefit from this and are less likely to develop antischool attitudes. They also develop more ambitious educational aspirations. The issue of peer effects is relevant to policy making because it can serve as a basis for relocating students to different schools, with the argument that weak students would benefit if they were in a class with high-performing students. It could even enhance the global efficiency of the system whenever peer effects are higher for low-skilled students (generally, peer effects prove asymmetric, as they favor weaker students).

Teachers also adapt their practices to the student body. They are more demanding in the best tracks and schools, and their expectations tend to foster inequalities present from the outset. In systems that differentiate between students, what they can expect is strictly limited. Actually, some PISA data show (see the analysis of PISA, 2000 and 2003 data sets led by Duru-Bellat et al., 2008) that in countries where school systems are the most differentiated, relatively few students say that their teachers show interest in every student's learning. Similarly, relatively few students say that their teachers help them with learning in most lessons. More broadly, the faith in the capability of every student is less widespread in differentiated systems. Conversely, in an egalitarian country such as Japan, there is a widespread belief in the ability to learn of every pupil.

It should also be noted that while the above-mentioned processes are currently considered as demonstrated, some

debates exist regarding the impact of other characteristics of education systems, like the degree of centralization versus decentralization, the role of competition between schools and of school choice, and the weight of school autonomy versus an important role for the State. However, what is mainly under discussion is the impact of the marketized system (with choice, competition, and autonomy) on students' mean performance. In any case, this impact, be it positive or negative, is never dramatic; however, the fact that it results in increased inequality seems more consensual (Van de Werfhorst and Mijs, 2010).

Last but not least, one may stress that the results of cross-sectional studies such as PISA, and the different policy measures that can be drawn from them can be significantly enhanced through the implementation of a longitudinal component; that is the only way to establish causal links between policies and outcomes and some countries such as Canada or Germany have yet engaged in this kind of study (see OECD, 2010b; Blossfeld et al., 2009).

Democratizing Policies

Around the world, especially in the most advanced economies, education has developed over the most recent decades, with several aims. One is to position the country in a global competition that requires increasingly higher skills, toward the end goal of achieving a knowledge society. Another is to fight against social inequality, which is both wasteful and unbearable in an education-based meritocracy. Thus, the democratization of education has been on the political agendas of most countries (Duru-Bellat, 2004).

The most obvious way to achieve this dual aim seemed to be to simply expand access to education. Obviously, when a whole generation has access to upper secondary school, some social inequality is erased. Thus, lengthening schooling careers may be considered an efficient way to enlarge access to education and, thus, democratize upper secondary and tertiary education. Actually, the expansion of education has indeed resulted in some democratization, especially at the highest level of education reached (although the increased wealth of most countries may have played a part; Breen et al., 2009). But research shows that most often the emergence of social inequalities is only postponed. That is because in order for the most advantaged families to stay ahead, they will often strive to keep their children in school longer. That generates a global trend, often labeled credential inflation, where all students must study longer and longer in order to get the same returns.

This evolution has two broad and unexpected consequences. First, social educational inequality does not really decrease as long as students remain unequally endowed with cultural resources and previous levels of achievement, but rather only shifts it to an upper level so that, ultimately, diplomas remain the gatekeepers of social (im)mobility and inequality is 'maximally maintained,' as Raftery and Hout (1993) suggested two decades ago. As Walters (2000) suggested, growth operates here both as a safety valve and as a counter reform, not allowing things to change. Second, as in economic inflation, this credential fosters a depreciation of the economic worth of the degrees themselves (Büchel, 2003).

Actually, according to Kerckoff's analysis (1995), the function of schooling is twofold; every society channels its members into the social positions through two broad families of institutional arrangements, thereby reproducing the process by which each generation becomes distributed into the unequal positions of society. In the first stage – the institutional arrangement devoted to achieving the socializing function – children are channeled into educational tracks, which results both in social inequalities in achievement and in specific and segregated socialization and training. In the second stage, devoted to achieving the allocation function, students are channeled into jobs based on their credentials; however, this process also relies on other parameters that vary across countries. Despite this illuminating distinction (socialization and allocation), most sociological theories analyzing social reproduction have focused on the crucial socialization function that schools play in this process (notably Bourdieu and Passeron, 1990). The implicit hypothesis is then that getting a university degree is the key to getting a job, as if the allocation process was unable to add any specific additional inequality. Thus, the education level achieved is seen as the main vehicle of social reproduction, through the inequalities of achievement generated within schools. Only recently have some sociologists dared to ask whether it is really so obvious, while fighting social inequalities of achievement is a legitimate aim in and of itself (if education is considered a good and not just a positional asset), it will not dramatically reduce social reproduction.

Here, one may recall Boudon's prediction (1974) that educational inequalities may well decline without causing a corresponding change in the structure of social inequalities. That is because the process of social reproduction itself appears to be intimately embedded in the structural framework of the job market and its evolution. Actually, individuals, whatever their personal attributes, insert themselves into a society whose places are predefined, and even if education is a relatively effective way of accessing the best positions, the definition of these places themselves (and more broadly how society is stratified and unequal) does not fundamentally arise from this action. Similarly, the educational system has no power over the effective market value of the degrees it issues, which itself results from the structural distribution of jobs and the peculiar relationships prevailing between degrees and qualifications. If, as has been observed in most modern countries over recent decades, the expansion of the social structure at its highest end has been slower than the rate at which graduates who might fill this level are produced, an adjustment will be made in the form of credential devaluation.

This crucial role of discrepancies between education (its structures and its evolution) and social reproduction would have one shift the focus to the relationship between education and jobs; what takes place within education would have less influence than the interaction between what the education system 'produces' – the credentials – and social stratification itself. Today, with the rise of youth unemployment, it is all the more valuable to investigate the relationships between degrees and employment as a specific process in the global reproduction process.

Because the concern for youth employment has moved up on the political agenda, education policies and the goal of democratizing education as a whole may see some changes. On the one hand, mandating that all students have the same basic knowledge is the best way to ensure that they find jobs later on. Thus, adopting policies aimed at canceling early tracking and developing common core schooling has been encouraged by the OECD, in conformity with the comprehensive ideal. However, a recent study (Bol and van de Werfhorst, 2013) has shown on a comparative basis that educational systems with a high level of early tracking accompanied by a vocational orientation do improve student allocation in the labor market. As this kind of system is also known to expand educational inequality, there may be a trade-off for systems and policy makers in economic efficiency (through youth employment) and equity of education.

A less ambivalent policy, still more assertively promoted by the OECD, is to give priority to tackling the achievement gap from the very start through early childhood provisions and preschool education. The effects of such programs include better educational achievement as well as, much later on, better employment rates and earnings, crime prevention, satisfying family relationships, and improved health. Therefore, they are often seen as the best methods for producing the highest returns, that is, creating the basis not only for better success at school, but also, more broadly, for better life opportunities. However, some argue that the results of these programs are often uncertain and that it would be more cost-effective to implement actions devoted to parents (e.g., like the parent education programs implemented in Belgium) in order to foster among parents specific forms of educational expertise such as linguistic interactions, learning incentives, reading books, and connecting to schools and teachers.

Of course, one could also argue that it would be still more effective to reduce poverty among the targeted children's families. It is widely recognized that education cannot be the only vehicle used to achieve equity and social cohesion. Education is nested in society as a whole and it seems impossible to achieve true equality of opportunity in an unfair society. Actually, the only countries in which some significant democratization has been observed (Northern European countries) are those that have implemented larger social welfare policies.

See also: Cognitive Development: Child Education; Comparative Research in Education: Iea Studies; Cross-National Comparisons in Education: Findings from Pisa; Educational Sociology; Equity and Education; Families as Educational Settings; Family and Schooling; Inequality, Social; School Choice; Socioeconomic Status and Academic Achievement.

Bibliography

Bernstein, B., 1975. Class, Codes and Control. Routledge and Kegan Paul, London.

Blossfeld, H.-P., Schneider, T., Doll, J., 2009. Methodological advantages of panel studies. Designing the new national educational panel study in Germany. Journal for Educational Research Online 1, 10–32.

Bol, T., van de Werfhorst, H., 2013. Educational systems and the trade-off between labour market allocation and equality of educational opportunity. Comparative Education Review 57 (2), 285–308.

Boudon, R., 1974. Education, Opportunity, and Social Inequality. Wiley, New York.

Bourdieu, P., Passeron, J.-C., 1990. Reproduction in Education, Society and Culture. Sage, London.

Breen, R., Jonsson, J., 2005. Inequality of opportunity in comparative perspective. Annual Review of Sociology 31, 223–243.

Breen, R., Luijkx, R., Müller, W., Pollak, R., 2009. Nonpersistent inequality in educational attainment: evidence from eight European countries. American Journal of Sociology 114, 1475–1521.

Büchel, F. (Ed.), 2003. Overeducation in Europe. E. Elgar, Cheltenham.

Coleman, J.S., et al., 1966. Equality of Educational Opportunity. US Department of Health, Education and Welfare, National Center for Education Statistics, Washington, DC.

Duru-Bellat, M., 2004. Social Inequalities at School and Educational Policies. UNESCO (Fundamentals of Educational Planning).

Duru-Bellat, M., Mons, N., Bydanova, E., 2008. Cohésion scolaire et politiques éducatives. Revue Française de Pédagogie 164, 37–54.

Graaf, N.D., Graaf, P.M., Kraaykamp, G., 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73 (2), 92–111.

Green, A., Preston, J., Janmaat, G., 2006. Education, Equality and Social Cohesion: A Comparative Analysis. Palgrave, Basingstoke.

Heyneman, S., Loxley, W., 1983. The effects of primary school quality on academic achievement across twenty-nine high and low-income countries. American Journal of Sociology 88, 1162–1194.

Janmaat, J.G., Duru-Bellat, M., Green, A., Mehaut, P. (Eds.), 2013. The Dynamics and Social Outcomes of Education Systems. Palgrave, Basingstoke.

Kerckhoff, A.C., 1995. Institutional arrangements and stratification processes in industrial societies. Annual Review of Sociology 15, 323–347.

Lareau, A., 2011. Unequal Childhoods: Class, Race and Family Life. University of California Press, Berkeley.

Mercy, J.A., Steelman, L.C., 1982. Family influence on the intellectual attainment of children. American Sociological Review 47, 532–542.

OECD, 2005. School Factors Related to Quality and Equity. OECD, Paris.

OECD, 2010a. PISA 2009 Results: Overcoming Social Background. OECD, Paris.

OECD, 2010b. Introduction: The Case for Linking PISA with Longitudinal Studies, in Pathways to Success. OECD, Paris. http://dx.doi.org/10.1787/9789264081925-2-en.

Raftery, A.E., Hout, M., 1993. Maximally maintained inequality: expansion, reform and opportunity in Irish education 1921-75. Sociology of Education 66, 41–62.

Sullivan, A., 2000. Cultural capital and educational attainment. Sociology 35 (4), 893–912.

Tesse, R., Lamb, S., Duru-Bellat, M. (Eds.), 2007. International Studies in Educational Inequality, Theory and Policy, 3. Springer, Dordrecht.

van de Werfhorst, H., Mijs, J.B., 2010. Achievement inequality and the institutional structure of education systems: a comparative perspective. Annual Review of Sociology 36, 407–428.

Walters, P.B., 2000. The limits of growth: school expansion and school reform in historical perspective. In: Hallinan, M. (Ed.), Handbook of Sociology of Education. Kluwer Academic/Plenum Publishers, New York, pp. 241–264.

Wössman, L., 2003. Schooling resources, educational institutions and student performance: the international evidence. Oxford Bulletin of Economics and Statistics 65 (2), 117–170.

Social Inequality in Cultural Consumption Patterns

Koen van Eijck, Erasmus University Rotterdam, Rotterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social inequality in cultural lifestyles can be understood using cultural capital theory or status theory. Empirical studies cannot easily produce results that allow for a proper comparison of the merits of these theories. In addition, the central role of habitus in both theories is problematic from a theoretical perspective, as relations between habitus and observed behavior are not straightforward and much of cultural behavior is affected, if not shaped, by variable social contexts.

Introduction

Cultural consumption patterns have solid social roots. They are affected by parental schooling levels and family socialization, level of education, occupational status, and income (Chan, 2010; Mohr and DiMaggio, 1995). All of these determinants are somehow related to class and status, thus causing social inequality in cultural consumption.

The central aim of this contribution is to assess the merits of the main theoretical explanations for this social inequality in cultural consumption by evaluating the existing empirical evidence. The most prominent theoretical accounts, i.e., cultural capital theory and status theory, will be addressed. Difficulties with attempts to empirically settle which one of these theories is more valid will be addressed. In addition, special attention will be paid to the habitus as the linking pin between socioeconomic background and cultural lifestyles. Finally, possible directions for future research are suggested.

Cultural Inequality: Theoretical Accounts

Cultural inequality can be theorized to follow from differences in skills or resources. It takes resources, or capital (cultural, economic, and social) to develop and implement a certain lifestyle. Unequal command of these resources leads to different lifestyles. This ties in with a second common theoretical account, which sees lifestyles as expressing social honor or prestige by demonstrating access to certain, more or less scarce, resources. Such prestige may not only be a mere by-product of a lifestyle, but it may also be a reward that is consciously sought after by engaging in certain esteemed activities.

Knowledge and Resources

According to Bourdieu (1984), cultural lifestyles depend primarily on cultural capital. This cultural capital is acquired during (early) socialization and inculcated in the habitus, which structures subsequent cultural practices. This entails a process that Bourdieu called cultural reproduction, where parents rich in cultural capital transmit this capital to their children, who simultaneously invest and expand it in their educational career and emerge out of this process endowed with a large amount of cultural capital of their own. The cultural skills thus embodied engender an aesthetic disposition that encourages a taste for highbrow culture.

Behind this reasoning lies the assumption that highbrow culture is more complex than popular culture. Thus, in his *information-processing theory*, Ganzeboom (1982) argued that the ability to process complex cultural information is a requirement for highbrow cultural consumption. This argument was based on the work of Gary Becker (1964) on human capital, Bourdieu's notion of cultural capital (Bourdieu and Passeron, 1977), and more psychological ideas from Berlyne (1974). The latter rest upon assumptions regarding the complexity of cultural stimuli, which are hard to assess empirically. However, it does make intuitive sense to assume that the joy derived from cultural consumption depends on people's ability to interpret them. The most pleasurable experiences occur whenever people's skills at deciphering a cultural stimulus match the challenge, or complexity, posed by this stimulus. It thus follows that, in order to enjoy complex culture, one needs the cognitive resources, or skills, to deal with this complexity. These skills depend on intelligence, experience, or creative talent, which are partly reflected in schooling levels, but also enhanced by cultural socialization and acquaintance with the world of culture. This is cultural capital theory's explanation for lifestyle differentiation in a nutshell: people and cultural activities can be matched to one another because differential socialization practices lead to different information processing skills and cultural activities differ in the skills they require.

Status and Social Prestige

People's cultural lifestyles tell us something about their social position. Society at large confers a certain amount of honor to members of specific social classes. Belonging to a certain class can be signaled by displaying the appropriate lifestyle. It is exactly because they indicate the possession of certain skills and resources that cultural lifestyles have a communicative function allowing them to convey a certain status.

At the close of the nineteenth century, Veblen (1953[1899]) argued that lifestyles mainly reflected economic differences. He coined the term *conspicuous consumption* to indicate that the overt possession of expensive consumer goods functioned as an effective mode of status display. Ten years after Veblen's mocking account of the upper classes of his day, Weber (1968

[1909]) argued that economic differences were not the only source of social inequality. He opposed to a one-dimensional view of stratification with his distinction between class and status. Class structure, he argued, was formed by economic relations rooted in labor markets. The status order, on the other hand, was structured by relations of social superiority and inferiority among individuals on the basis of the social honor attributed to them. This honor was conferred on the basis of social (class) position, but descent from founding families or birth into the nobility were also sources of such honor.

Bourdieu (1984) treats status as the symbolic aspect of class structure, but he too argues that class positions are not defined by economic resources alone. Rather than assuming objective class boundaries based on economic structures, Bourdieu focuses on "the structured formation or self-production of class collectivities through struggles that simultaneously involve relationships between and within classes and determine the actual demarcation of their frontiers. Bourdieu replaces the concept of class structure with that of social space, understood as the multidimensional distribution of socially effective forms of power (or capital, be it economic, cultural or social) underlying social positions" (Wacquant, 1991: p. 52). Most notably, Bourdieu introduced cultural capital as a relevant noneconomic source of social inequality and as a tool for social reproduction in a society where the direct intergenerational transmission of economic resources became incompatible with meritocratic ideals (DiMaggio, 1979).

Thus, cultural capital can be seen as both a bundle of knowledge and skills and a status symbol. The latter aspect of lifestyles is represented in Ganzeboom's *status-seeking theory*. Status theory as presented by Ganzeboom is rooted in conflict sociology, taking its inspiration mainly from Weber (1946 [1922]), Veblen (1953[1899]), Collins (1979), and, again, Bourdieu and Passeron (1977). Here, lifestyles are primarily considered as criteria for demarcating and identifying class groupings. Cultural lifestyles are then seen as primarily reflecting the status-rendering characteristics of the behaviors and taste patterns involved. Simultaneously reflecting the honorable investments of time, money, and knowledge, highbrow culture participation has been an excellent way to demonstrate social worth. As long as others believe that this type of culture requires special skills and sensibilities, it also helps legitimate the high-end social positions of the highbrow audience by demonstrating their exquisiteness. Culture can thus be used to exclude people with 'bad' taste and include those who are socially and culturally similar to oneself.

Empirical Tests of the Theories

Above, it has been explained that cultural lifestyles can be seen as either reflections of knowledge, skills, and attitudes, or as indicators of status group membership. In his discussion on the concept of cultural capital, Lizardo also addresses the general distinction between skills vs status seeking. On the one hand, he argues that cultural capital can indeed be defined as "an aptitude or a generalized, transposable (across contexts) skill acquired in the combined realms of the upper-middle class family and the school system" (Lizardo, 2010: p. 310). In addition, he distinguishes what he calls the boundaries point of view, where cultural capital is defined as the institutionalized repertoire of high status signals useful for purposes of marking and drawing symbolic boundaries in a given social context. In the boundary perspective, the link with specific cognitive requirements as to the content of cultural capital is largely absent, since "what counts as cultural capital are those symbolic resources that are actively mobilized by members of groups or class fractions to establish their difference from other groups and thus to devalue the cultural resources and symbolic practices of outsiders" (Lizardo, 2010: p. 310). Bourdieu similarly talked about a 'cultural arbitrary,' indicating that it does not matter much which exact cultural taste the upper classes display; it is in large part the fact that *they* have this taste that makes it a form of cultural, or symbolic, capital. Although, at a more general level, lifestyles reflecting the ability to spend money or to deal with complexity will be more prestigious than lifestyles that require no such investments (Prieur et al., 2008), the symbolic value of cultural lifestyles is defined as a function of the social prestige of its possessors as much as in terms of properties of specific cultural items or activities such as complexity or rarity. This conflation of cultural capital as skills and cultural capital as a status marker makes it hard to empirically determine which theoretical account of social inequality in cultural lifestyles is more valid.

Mapping Social Background Effects

Cultural lifestyles symbolize status positions by pointing at certain skills, attitudes, knowledge, taste, or other resources. A resource that has consistently been found to be a very strong determinant of cultural lifestyles is educational attainment (Kraaykamp and NieuwBeerta, 2000; Van Eijck and Bargeman, 2004). For Ganzeboom (1982), effects of educational attainment substantiate the impact of knowledge. In addition to education, however, occupation and income also show up as recurring sources of lifestyle differentiation. For a proper understanding of the relation between lifestyle and social positions, these prime sources of stratification will be compared and the findings will be linked back to the work of Veblen, Weber, Bourdieu, and Ganzeboom.

According to Lizardo (2008: p. 5), we are witnessing a "systematic transition in the cultural stratification system of the Euro-American West starting in Renaissance Europe from a regime premised primarily on the differential ability to *own* objectified cultural products – with a contemporaneous 'relative democracy' of the ability to appropriate cultural works symbolically – to a cultural stratification regime which devalues the *sole* deployment of the direct appropriation strategy and which emphasizes symbolic appropriation." There can be no doubt that this implies a shift from economic to cultural characteristics as the main determinants of lifestyles. Income should therefore be a less relevant background characteristic than education or other background features related more directly to cultural capital.

Although the literature on cultural lifestyles is very much scattered, a relatively clear picture as to the causes of lifestyle differentiation emerges. Indeed, income seems to be mostly inconsequential when mere preferences or low-profile participation (reading, listening to music) are concerned rather than relatively expensive outdoor practices (Berghman and Van

Eijck, 2009; Chan and Goldthorpe, 2007a; DiMaggio and Useem, 1978; Van Eijck, 1999). Veblen's notions of conspicuous consumption and conspicuous leisure based on income differentiation are clearly not very helpful for understanding lifestyle differentiation today. Income is of relatively little use if we want to understand taste differences. When actual behavior outside one's own home is the dependent variable, income effects are somewhat more common (Bihagen and Katz-Gerro, 2000; Kraaykamp et al., 2007; Van Berkel and De Graaf, 1995). Income seems to only affect activity per sé, distinguishing between, e.g., those who do not attend visual art institutions at all vs those who do so (Chan and Goldthorpe, 2007b). It hardly differentiates between one expressed preference or type of participation and another. Thus, it seems valid to assume that income is a prerequisite for consumption rather than a clear cultural tastemaker, as Ganzeboom (1982) argued.

Occupational class or status is often found to be more important than income (Alderson et al., 2007; Chan and Goldthorpe, 2005; Kraaykamp et al., 2007). Class or status positions clearly represent more than a person's place in a straightforward economic hierarchy. Given the prevalence of status effects over class effects in most of the analyses reported in Chan (2010; see also the contributions to the special issue of *Poetics* (35: 2–3) entitled 'Social status and cultural consumption in seven countries'), social status still seems to play an important role, also after income is controlled for. The importance of status over occupation lends support to a conceptualization of lifestyles as expressions of social group membership, or socio-cultural boundaries, rather than of specific cultural or economic resources.

Yet the most important determinant of cultural lifestyles is education (Chan, 2010). It is not exactly sure what this means, as education seems to enhance all sorts of activities, not just the more complicated or elitist ones, without guaranteeing a very intellectual mode of appropriation. Nevertheless, cultural capital strongly structures lifestyles, as Bourdieu posited. Even if relations are less deterministic than he suggested, the idea of a certain homology of social and cultural positions still has merit. This is further confirmed by the analyses of specific job characteristics. Where income merely affects some forms of participation per sé, education and, to a lesser extent, occupational status additionally differentiate between distinct taste patterns. Finally, much of the analyses referred to here demonstrates that education, income, and occupational status have unequal yet cumulative effects. Social stratification is more than income, occupation, or education: it is a combination of each of these, where education is the element that is most consequential for lifestyle formation, followed by occupation and, lastly, income. This order of importance is in line with the embodied cultural capital regime sketched by Lizardo (2008). The link between taste and education as a cultural structuring principle of social space is and remains very strong.

Cognitive vs Status Indicators

It is difficult to draw strong conclusions regarding the rivaling accounts of cultural lifestyle differentiation from these findings. Status attainment theory and information processing theory often come up with similar predictions, such as positive effects of cultural family socialization or educational attainment on highbrow cultural participation. Clever, detailed hypotheses are needed in order to generate test results that actually differentiate between the two perspectives. What is especially needed is a way to disentangle the different potential effects of education, as schooling has both cognitive (resource) and status (boundary) consequences. This requires detailed measures of cultural capital that allow scholars to distinguish between (effects of) cultural knowledge and (effects of) cultural status. This has been attempted in studies on educational attainment, where distinctions were made between effects of parental reading and parental beaux arts participation (De Graaf, 1986; De Graaf et al., 2000). De Graaf interpreted effects of parental reading as representing an impact of educational resources, or skills, whereas the absence of parental beaux arts participation effects was argued to disprove the relevance of highbrow status culture involvement. Sullivan (2001) provides a more direct measure of cultural capital as information (processing) and finds that the effects of pupil's reading on their scholastic achievement are fully mediated by their vocabulary and cultural knowledge. She concludes that "the reason for the effect of cultural participation on academic attainment is that cultural participation is associated with intellectual resources which help pupils at school" (Sullivan, 2001: p. 910). Cultural knowledge also turned out to be more relevant than beaux arts participation later on in the educational career (i.e., admission to Oxford University: see Zimdars et al., 2009). In a sense, such tests offer what Ganzeboom (1982: p. 188) suggested as a first test, i.e., assessing whether the highest correlations are found between activities that demand equal types of skill and knowledge, or between activities that demand the same kind of formal attendance. If education is more closely correlated with (parental) linguistic skills and knowledge than with (their) beaux arts participation, information theory seems to offer the best explanation for the intergenerational reproduction of educational opportunities.

But why does education, in its turn, enhance cultural participation? Is this a matter of having received relevant instructions or pertinent information too? Nagel and Ganzeboom (2002) assessed the impact of taking arts education classes on participation in legitimate culture. These effects were found to be small and in fact attributable to self-selection (their longitudinal design showed these 'effects' to be present already at age 14, before actual enrollment). Similarly, differences in cultural participation between educational categories were already partly present before the educational career was finished due to differential family socialization (Nagel, 2010). These results should not be too surprising, as there is little emphasis on highbrow culture in school curricula in countries such as Britain, the United States, or the Netherlands (De Graaf et al., 2000; Sullivan, 2001). But here too, elaborate path models would be required to fully disentangle the effects of information processing and status attainment using very specific cultural capital indicators.

Another possible way of disentangling status- and information processing effects would be to focus on instances of status inconsistency or social mobility. Whenever people have schooling levels that do not match their socioeconomic status, there is more room to assess effects of status indicators such as occupational prestige or income and information processing

indicators such as education separately. In such instances, people have different ranks in relevant status dimensions. For example, occupational status only affects cultural taste if one's occupation is high relative to one's schooling level (Ganzeboom, 1982). In that case, we are also most likely to find extra high levels of material consumption (De Graaf, 1991). Similarly, if marriage partners have different status characteristics, De Graaf observed status-maximization effects indicating that the partner with the highest education most strongly determined the cultural participation of the spouses. Such effects are to be interpreted most likely in terms of status seeking rather than information processing.

Finally, looking at changes in cultural lifestyles during the life course might be informative. What happens when people change or lose their job or their spouse? Under such circumstances, their status and social networks may change while their cultural capital does not. Ganzeboom also interprets the strong and lasting effects of early cultural socialization as evidence in favor of information theory, but it may also be argued that, e.g., getting acquainted with highbrow art in the early family context does not only enhance cultural knowledge but a sense of status as well. What he does rightly stress throughout, however, is the need for multivariate analyses because of the intricate relations between family socialization, education, cultural knowledge, and social status. Both theories have their own merits and help explain part of the differences in cultural lifestyles. The main problem is that the social indicators that are most available to sociologists, such as schooling level or occupational status, also combine elements of status and information which makes it difficult to interpret schooling effects as support for information theory or effects of occupational status as support for status-seeking theory. Above all, we seem to need more detailed measures of these background indicators for incorporation into multivariate models.

Criticisms and Controversies

The question whether cultural lifestyles are determined by either cultural resources or status considerations seems hard to resolve empirically. Apart from difficulty in finding appropriate quantitative measures, it can be argued that this dichotomy is conceptually problematic as well. The theories assume that either a cultural disposition or a status position can be largely held responsible for cultural lifestyles. As such, they both assume that a certain, typically class-specific, habitus is translated more or less consistently into observable cultural behavior. This central assumption is problematic for at least three reasons, as will be outlined below.

Deducing the Habitus from Observed Behavior

Since habitus cannot be observed directly, many studies observe cultural behavior and infer from these observations which habitus must be underlying this behavior. For example, if a person attends a ballet performance, it is assumed that this behavior indicates the presence of an aesthetic disposition, or a habitus geared toward highbrow art forms. Several studies have convincingly refuted the validity of this assumption.

In 1993, Halle published his study of the art people put in their homes. He found that members of the upper middle classes were much more likely to display abstract art in their homes while working class families preferred religious iconography. This observation would suggest the existence of distinct class habitus causing these differences in home stylization. However, Halle found that the upper classes wanted their abstract art to be decorative and to symbolize aspects of their everyday lives. Similarly, the reasons why the upper classes display primitive art show a remarkable resemblance to why the working classes put up religious art as both groups attribute symbolic meanings to these works that were related to their own social world. Thus, Halle concluded, distinct behaviors may well imply quite similar habitus.

Although Holt (1997a) does not agree with Halle that his findings imply the invalidity of Bourdieu's theory of taste, elsewhere he does provide evidence of the complexity of the relation between habitus and cultural behavior by criticizing the idea that lifestyles are shaped by universal traits or values and the notion that meanings of cultural behavior can be inferred from intrinsic qualities of the activities engaged in (Holt, 1997b). Thus, he demonstrated that people who share a love of antique furniture may have very distinct reasons for doing so, ranging from enjoying the bargaining involved to true connoisseurship to bricoleur refurbishing. In a similar vein, Bennett et al. (2009) have shown that highbrow cultural behavior can imply very different underlying motivations or dispositions. For example, opera attendance in the United Kingdom is more often part of a work-related social event than an expression of a personal predilection for the genre. And listening to classical music is often described as 'soothing,' 'relaxing,' or something that 'gets the women going' (Bennett et al., 2009: p. 83), implying the application of more or less functional criteria pertaining to mood management (DeNora, 2000) rather than employing aesthetic categories. Again, similar activities or preferences are linked to highly divergent habitus while seemingly distinct preferences do not necessarily imply equally distinct dispositions. In short, habitus cannot be deduced directly from the cultural preferences and practices we observe. This is especially problematic since habitus is often a black box that is assumed or derived from observing something else (either socioeconomic background or cultural behavior), to which it does not relate in a straightforward manner.

Predicting Observed Behavior from the Habitus

Moving the other way around, from the habitus or cultural dispositions to observable cultural taste or behavior, is problematic as well. Lahire (2003) criticizes the theoretical use of dispositions as if they were transferable, always active, and neatly aligned rather than context dependent, regularly on standby mode, and internally incoherent. The activation of people's multiple dispositions depends on opportunity structures that may or may not allow people to actually act in accordance with their dispositions. But a lack of correspondence between what people believe (dispositions) and what they do (action) is not just due to practical barriers keeping people from pursuing their inclinations. According to Lahire (2003), we must acknowledge a more fundamental difference between dispositions to act and dispositions to believe. The latter, beliefs, are inculcated by socialization in the family and a range of institutions, such as school or media of all sorts.

These beliefs are, however, not necessarily confirmed by everyday experiences and therefore differ in their strength. They may be unrealistic in themselves (I must be rich/beautiful/happy/clever), but they may also lack rapport with practices because the latter may be hard to control (eating habits, addictions, and neuroses). That is why Lahire differentiates between dispositions to act and dispositions to believe. Yaish and Katz-Gerro (2012) similarly demonstrate that cultural tastes are more strongly determined by socialization than cultural participation, whereas participation is more determined by economic resources and less by cultural resources than taste. These findings, too, demonstrate that different processes are at work in shaping dispositions to believe (taste) and dispositions to act (participation).

Dispositions to act and dispositions to believe are affected not only by partly different resources, but also by a multitude of social agents and institutions. Therefore, "we cannot act as if a single system of cultural beliefs existed at the same moment in every part of the social world, as if all the social groups tended to fall into line with the legitimate consumption of the most 'cultivated' of the members of the intellectual fractions of the dominant class" (Lahire, 2008: p. 172). As a consequence, cultural lifestyles are hardly ever consistent in terms of cultural legitimacy. Instead, Lahire argues that contemporary cultural lifestyles are dissonant, and that individuals are typical and marginal at the same time as their daily lives propel them to shift from one cultural register to another. Rather than being straightforward embodiments of the taste of their social class, individuals move about in a complex world full of social obligations, routines, secret pleasures, family duties, professional networks, or inaccessible events. When the social world gets more heterogeneous or fragmented, the incorporation of this social world becomes less likely to result in a robust, univocal habitus that functions as a reliable guideline across the different contexts in which people find themselves.

Context and the Flexible Habitus

More than half a century ago, Goffman (1959) illustrated how frequently and often automatically people switch social roles depending on the situation or company they are in. This may occur unconsciously, as part of everyday routines, but people also intentionally manipulate the impression they have on others by adapting their self-presentation to the circumstances at hand. This flexibility not only makes social interaction run more smoothly; it also implies that different schemes of action are called upon in different contexts, thus corroborating the above implication that habitus does not always drive action in a straightforward manner. Thus, Lahire (2008: pp. 181–182) claims that "some individuals sometimes spend more time doing things or consuming cultural products for which they do not have a particularly pronounced 'personal taste,' than they do acting upon those cultural preferences that would express their 'personal cultural identity' […] Cultural practices and preferences are produced when incorporated inherited dispositions, partialities, and competences, meet with fixed institutional or relational contexts; they are not intrinsic characteristics of the individual." Being less heterogeneous in one's actual cultural behavior would require an incredibly strict adherence to an incredibly strong taste. If the concept of habitus is to be used at all, Lahire argues, it should be considered as a compass that is both multiple and sufficiently flexible to allow people to socialize with a heterogeneous set of others and to make it easier to enjoy – or endure – bits of the culture of these others as well.

Collins (2004) takes this reasoning even further by questioning the conception of individuals as having a stable inner set of dispositions, if not identity, that remains constant across a wide range of situations at a more fundamental level. In fact, specific interactions and contexts contribute to this identity rather than revealing it. Thus, Collins (2004: pp. 82–83) argues that "taking part in an ethnic riot […] is not simply a way of acting out a preexisting ethnic identity, but a way of strengthening it, re-creating or even creating it." When evoked repeatedly, the emotions that are linked to specific contextualized roles give "rise to long-term emotional styles, which is a large part of what is meant by the term 'personality'" (Collins, 2004: p. 113). Collins' interaction ritual theory offers a promising way of understanding cultural lifestyles by taking what he calls the point of view of a radical microsociology in order to theorize the large impact of concrete interaction contexts on both the development of dispositions and the way they are translated into cultural behavior. By delving even deeper into the microfoundations of cultural socialization than Lahire, Collins lays bare the cognitive and emotional mechanisms shaping cultural lifestyles, simultaneously accounting for their thoroughly social origins as well as their profoundly contingent expressions.

Future Directions

To conclude, a few directions for future research are suggested on the basis of the ideas presented above. These pertain to searching for lifestyle differentiation among multiple hierarchies, using methods that are sufficiently exploratory to do so, and taking into account the relation between the production and reception/consumption of cultural goods and services.

Accounting for Omnivorous Lifestyles

As we have seen, cultural lifestyles are shaped by many socializing agents and institutions that neither convey consistent signals nor limit their impact to positions on a one-dimensional highbrow-popular dimension. One consequence of this is that lifestyles are unlikely to be consonant. Looking at individual lifestyle patterns, we rather observe cultural dissonance (Lahire, 2008), or omnivorousness (Peterson, 1992), as a mixture of highbrow and popular cultural genres. Although this cultural omnivorousness has been amply observed, it is still unclear how the mixing of highbrow and popular culture should be understood. It seems fruitful to study the mechanisms behind such cultural patterns. On the one hand, omnivores may appropriate a broad range of cultural styles through social mobility, where they accumulate different cultural experiences during their (typically upward) social trajectory (Van Eijck, 1999). This explanation takes into account the growing multitude of social contexts that are relevant (at different points) during people's lives. Alternatively, omnivores may be socialized into highbrow culture and subsequently learn to enjoy popular culture because

their aesthetic disposition allows them to appreciate a broad range of cultural items from a more or less formalist perspective (Lizardo and Skiles, 2012). This approach also emphasizes the different context that offers opportunities to become acquainted with cultural diversity, but puts more emphasis on the aesthetic disposition that both encourages cultural tolerance and allows for a proper decoding of cultural artifacts. Lizardo (2006) is consistent in his emphasis on cultural capital as driver of experience in his study of the impact of cultural tastes on social networks. On the same topic, Mark (1998) reasons more in line with Van Eijck (1999) by stating that sociodemographically heterogeneous networks encourage broader tastes. The latter is also more congruent with the seminal ideas developed by Lahire and Collins. In order to move beyond the assumption of habitus as the automatic or inevitable internalization of more or less homogeneous external circumstances, both suggestions advocate a closer look at the social interaction mechanisms leading to specific lifestyles.

Modeling Cultural Spaces

Cultural omnivorousness has often been linked to the notion of openness (Ollivier, 2008). According to Peterson and Kern (1996: p. 904), omnivorous taste signifies 'an openness to appreciating everything.' This may imply that omnivores differ from Bourdieu's ideal-typical snobs in that they use different criteria for deciding which cultural artifacts are of interest (e.g., Bennett et al., 2009; Johnston and Baumann, 2007), or that they apply the same aesthetic criteria to a wider range of artifacts (Lizardo and Skiles, 2012). In any case, it has become more difficult to rely on analyses that work from assumptions on the meaning of both cultural items (e.g., by constructing scales that are assumed to represent highbrow or popular culture) and socioeconomic indicators (e.g., by computing linear status scales). Studies working along these lines have shed ample light on the prevalence of omnivorousness, but have often operationalized it as the next model of cultural inequality without asking whether other models or dimensions may be at work, perhaps simultaneously. The cultural omnivore opened up one alternative for Bourdieu's theory of taste, but as the mechanisms underlying omnivorousness are only just being explored, scholars should be open to variation on more than just this (former) highbrow-popular dichotomy. There is still much exploratory work to be done in this respect and this may explain the renewed interest in Bourdieu's (1984) original method of multiple correspondence analysis (MCA) (Blasius and Mühlichen, 2010; Rouanet et al., 2000; Roose et al., 2012).

For a number of reasons, MCA can be considered a very suitable technique for a more exploratory analysis of cultural lifestyles. First, its relational approach fits well with the notion that cultural tastes and activities may (increasingly) have multiple meanings or uses (Abbott, 1988; Holt, 1997b; Lahire, 2008; Swidler, 1986). Second, it allows scholars to deal with the notion that different structuring principles seem to be at play at the same time in lifestyle formation (Chan, 2010; Holbrook et al., 2002). Third, it allows for dealing with detailed information on cultural lifestyles, possibly including tastes/dispositions and behavioral indicators simultaneously, which is required for truly exploratory purposes. Finally, it focuses on the underlying dimensions, constructing a cultural space, and thereby provides an interpretation of the findings in terms of relational differentiation that does not lean too heavily on the specific items used. This is necessary because status markers are likely to differ among societies or vary with changes in the field of cultural production (Atkinson, 2011; Van Eijck, 1999), as will be argued below.

Linking Consumption with Production

It is important to note that, for Bourdieu and others (DiMaggio and Powell, 1983; Dowd, 2011), lifestyles are not structured by social class or individual preferences alone. They represent positions in a cultural field that is largely institutionally defined. The distinction between highbrow and popular culture is also shaped by institutions (see also DiMaggio, 1992; Levine, 1988). The educational system, government, and cultural critics classify cultural products and thereby attribute symbolic value to them (Berghman and Van Eijck, 2009; Van Rees, 1987; Verboord, 2003).

DiMaggio (1987) has theorized in more general terms how such symbolic boundaries are determined by characteristics of social structures. Cultural systems vary along several analytically distinct dimensions rather than being classifiable into discrete types. DiMaggio argues that artistic classification systems differ in terms of differentiation (number of genres distinguished), hierarchy (degree of vertical ordering of genres according to prestige), universality (general vs group-specific salience of genres), and ritual strength (defense of genre boundaries by both producers and consumers). Each of these dimensions are affected by macrosocial conditions such as access to higher education, social inequality, correlations between different status attributes, or the degree of interaction between members of different classes. Also, classification systems can be defined by the ritual strength with which boundaries are separated, e.g., boundaries between the production of highbrow and popular culture. For example, Taylor (2009) argues that boundaries between popular and classical music consumption are fading as classical music is increasingly using popular culture marketing strategies such as the creation of superstars or advertising through mass media. The same holds for books, where the boundaries between literary and nonliterary genres (embodied by terms such as literary thriller or fantasy novel) are blurring due to the use of similar marketing strategies and the spread of book-related events such as award ceremonies or special book weeks devoted to literature, children's books, or thrillers. Analyzing such changing characteristics of cultural consumption is necessary to understand cultural lifestyles in a world where the boundary between state-subsidized legitimate culture production and the contemporary culture industries becomes ever more permeable.

See also: Art and Socialisation; Cultural Capital and Education; Cultural Participation, Trends In; Culture, Cognition and Embodiment; Culture, Production of: Prospects for the Twenty-First Century; Embodiment and Culture; Leisure and Cultural Consumption: US Perspective; Symbolic Boundaries.

Bibliography

Abbott, A., 1988. Transcending general linear reality. Sociological Theory 6, 169–186.

Alderson, A., Junisbai, A., Heacock, I., 2007. Social status and cultural consumption in the United States. Poetics 35, 191–212.

Atkinson, W., 2011. The context and genesis of musical tastes: omnivorousness debunked, Bourdieu buttressed. Poetics 39, 169–186.

Becker, G.S., 1964. Human Capital. Columbia University Press, New York.

Bennett, T., Savage, M., Silva, E., Warde, A., Gayo-Cal, M., Wright, D., 2009. Culture, Class, Distinction. Routledge, London.

Berghman, M., Van Eijck, K., 2009. Visual arts appreciation patterns: crossing horizontal and vertical boundaries within the cultural hierarchy. Poetics 37, 348–365.

Berlyne, D.E., 1974. Studies in the New Experimental Aesthetics. Wiley, New York.

Bihagen, E., Katz-Gerro, T., 2000. Culture consumption in Sweden: the stability of gender differences. Poetics 27, 327–349.

Blasius, J., Mühlichen, A., 2010. Identifying audience segments applying the "social space" approach. Poetics 38, 69–89.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgement of Taste. Routledge and Kegan, London.

Bourdieu, P., Passeron, J.-C., 1977. Reproduction in Education, Society, and Culture. Sage, London.

Chan, T.W., Goldthorpe, J., 2005. The social stratification of theatre, dance and cinema attendance. Cultural Trends 14, 193–212.

Chan, T.W., Goldthorpe, J., 2007a. Social stratification and cultural consumption: music in England. European Sociological Review 23, 1–19.

Chan, T.W., Goldthorpe, J., 2007b. Social stratification and cultural consumption: the visual arts in England. Poetics 35, 168–190.

Chan, T.W., 2010. Social Stratification and Cultural Consumption. Cambridge University Press, Cambridge.

Collins, R., 1979. The Credential Society. Academic Press, New York.

Collins, R., 2004. Interaction Ritual Chains. Princeton University Press, Princeton.

De Graaf, N.D., 1991. Distinction by consumption in Czechoslovakia, Hungary, and the Netherlands. European Sociological Review 7, 267–290.

De Graaf, N.D., De Graaf, P.M., Kraaykamp, G., 2000. Parental cultural capital and educational attainment in the Netherlands: a refinement of the cultural capital perspective. Sociology of Education 73, 92–111.

De Graaf, P.M., 1986. The impact of financial and cultural resources on educational attainment in the Netherlands. Sociology of Education 59, 237–246.

DeNora, T., 2000. Music in Everyday Life. Cambridge University Press, Cambridge.

DiMaggio, P., 1979. Review essay: on Pierre Bourdieu. The American Journal of Sociology 84, 1460–1474.

DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.

DiMaggio, P., 1992. Cultural boundaries and structural change: the extension of the high culture model to theater, opera, and the dance, 1900–1940. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago, pp. 21–57.

DiMaggio, P., Useem, M., 1978. Social class and arts consumption: origins and consequences of class-differences in exposure to arts in America. Theory and Society 5, 141–161.

DiMaggio, P., Powell, W., 1983. The iron cage revisited: institutional isomorphism and rationality in organizational fields. American Sociological Review 48, 147–160.

Dowd, T.J., 2011. Production and producers of life styles: fields of classical and popular music in the United States. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51 (special issue), 113–138.

Ganzeboom, H., 1982. Explaining differential participation in high-cultural activities. In: Raub, W. (Ed.), Theoretical Models and Empirical Analyses. ETS Publications, Utrecht, pp. 186–205.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Garden City, New York.

Holbrook, M.B., Weiss, M.J., Habich, J., 2002. Disentangling effacement, omnivore, and distinction effects on the consumption of cultural activities: an illustration. Marketing Letters 13, 345–357.

Holt, D., 1997a. Distinction in America? Recovering Bourdieu's theory of tastes from its critics. Poetics 25, 93–120.

Holt, D., 1997b. Poststructuralist lifestyle analysis: conceptualizing the social patterning of consumption in postmodernity. Journal of Consumer Research 23, 326–350.

Johnston, J., Baumann, S., 2007. Democracy versus distinction: a study of omnivorousness in gourmet food writing. American Journal of Sociology 113, 165–204.

Kraaykamp, G., NieuwBeerta, P., 2000. Parental background and life style differentiation in Eastern Europe: social, political and cultural intergenerational transmission in five former socialist societies. Social Science Research 29, 92–122.

Kraaykamp, G., Van Eijck, K., Ultee, W., Van Rees, K., 2007. Status and media use in the Netherlands: do partners affect media tastes? Poetics 35, 132–151.

Lahire, B., 2003. From the habitus to an individual heritage of dispositions: towards a sociology at the level of the individual. Poetics 31, 329–355.

Lahire, B., 2008. The individual and the mixing of genres: cultural dissonance and self-distinction. Poetics 36, 166–188.

Levine, L.W., 1988. Highbrow/Lowbrow: The Emergence of Cultural Hierarchy in America. Harvard University Press, Harvard.

Lizardo, O., 2006. How cultural tastes shape personal networks. American Sociological Review 71, 778–807.

Lizardo, O., 2008. The question of culture consumption and stratification revisited. Sociologica 2, 1–32.

Lizardo, O., 2010. Culture and stratification. In: Hall, J.R., Grindstaff, L., Lo, M. (Eds.), Sociology of Culture: A Handbook. Routledge, London, pp. 305–315.

Lizardo, O., Skiles, S., 2012. Reconceptualizing and theorizing 'omnivorousness': genetic and relational mechanism. Sociological Theory 30, 263–282.

Mark, N., 1998. Birds of a feather sing together. Social Forces 77, 453–485.

Mohr, J., DiMaggio, P., 1995. The intergenerational transmission of cultural capital. Research in Social Stratification and Mobility 14, 167–199.

Nagel, I., 2010. Cultural participation between the ages of 14 and 24: intergenerational transmission of cultural mobility? European Sociological Review 26, 541–556.

Nagel, I., Ganzeboom, H., 2002. Participation in legitimate culture: family and school effects from adolescence to adulthood. The Netherlands' Journal of Social Sciences 38, 102–120.

O'Hagan, J., 1996. Access to and participation in the arts: the case of those with low incomes/educational attainment. Journal of Cultural Economics 20, 269–282.

Ollivier, M., 2008. Humanist, populist, practical, and indifferent modes of openness to cultural diversity. Poetics 36, 120–147.

Peterson, R.A., 1992. Understanding audience segmentation: from elite and mass to omnivore and univore. Poetics 21, 243–258.

Peterson, R.A., Kern, R.M., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.

Prieur, A., Rosenlund, L., Skjott-Larsen, J., 2008. Cultural capital today: a case study from Denmark. Poetics 36, 45–71.

Roose, H., Van Eijck, K., Lievens, J., 2012. Culture of distinction or culture of openness? Using a social space approach to analyze the social structuring of lifestyles. Poetics 40, 491–513.

Rouanet, H., Ackermann, W., Le Roux, B., 2000. The geometric analysis of questionnaires: the lessons of Bourdieu's La Distinction. Bulletin de Méthodologie Sociologique 65, 5–18.

Sullivan, A., 2001. Cultural capital and educational attainment. Sociology 35, 893–912.

Swidler, A., 1986. Culture in action: symbols and strategies. American Sociological Review 51, 273–286.

Taylor, T.D., 2009. Advertising and the conquest of culture. Social Semiotics 19, 405–425.

Van Berkel, M., De Graaf, N.D., 1995. Husband's and wife's culture participation and their levels of education: a case of male dominance? Acta Sociologica 38, 131–149.

Van Eijck, K., 1999. Socialization, education, and lifestyle: how social mobility increases the cultural heterogeneity of status groups. Poetics 26, 309–328.

Van Eijck, K., Bargeman, B., 2004. The changing impact of social background on lifestyle: culturalization instead of individualization? Poetics 32, 447–469.

Van Rees, K., 1987. How reviewers reach consensus on the value of literary works. Poetics 16, 275–294.

Veblen, T., 1953[1899]. The Theory of the Leisure Class. An Economic Study of Institutions. The Viking Press, New York.

Verboord, M., 2003. Classification of authors by literary prestige. Poetics 31, 259–281.

Wacquant, L.J.D., 1991. Making class: the middle class(es) in social theory and social structure. In: McNall, S.G., Levine, R.F., Fantasia, R. (Eds.), Bringing Class Back In: Contemporary and Historical Perspectives. Westview Press, Boulder, pp. 39–64.

Weber, M., 1946[1922]. Class, status and party. In: Gerth, H.H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York, pp. 180–195.

Weber, M., 1968[1909]. Economy and Society: An Outline of Interpretive Sociology. Bedminster Press, New York.

Yaish, M., Katz-Gerro, T., 2012. Disentangling 'cultural capital': the consequences of cultural and economic resources for taste and participation. European Sociological Review 28, 169–185.

Zimdars, A., Sullivan, A., Heath, A., 2009. Elite higher education admission in the arts and sciences: is cultural capital the key? Sociology 43, 648–666.

Social Inequality in History (Stratification and Classes)

P Nolte, International University Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14313–14320, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article outlines theories of social stratification and class formation in Western thought, from antiquity to the 1960s, with a focus on Marx, Weber, and neo-Weberian approaches. It also looks at the use of concepts of social inequality and class in historiography, particularly in the *New Social History* since the 1960s, with Germany and the United States as examples. This article deals with some of the more recent approaches associated with the work of E.P. Thompson and Pierre Bourdieu among others.

This article sketches the development of theories of social stratification and class formation in Western thought, from antiquity to the 1960s, and especially since nineteenth-century sociology, with a focus on Marx, Weber, and neo-Weberian approaches. It also looks at the use of concepts of social inequality and class in historiography, particularly in the *New Social History* since the 1960s, with Germany and the United States as examples. Since then, the history of stratification and class has been fundamentally challenged and rewritten through concepts and theories of culture. Third, therefore, this article deals with some of those reformulations (e.g., E.P. Thompson, Pierre Bourdieu, or the history of consumption). Some approaches even seem to go beyond class and stratification. Here, the article discusses the impact on historical writing of gender and ethnicity research, of notions of social lifestyles and radicalized individualism, and of ideas about the cultural and linguistic 'construction' of classes.

As a universal fact of all human civilization, social inequality has been shaping advanced societies in a most fundamental way for many thousands of years, especially, since the rise of more complex cultures and political systems in the wake of the 'Neolithic Revolution.' The concept of social inequality refers to an unequal distribution of positions in society, a distribution that is determined by structures of power (e.g., the modern state and its administrative elites), of technology and economy (e.g., industrial technology and capitalism), and of cultural expression and perception (e.g., Christian religion and its attempt to define and explain a 'holy' order of worldly society).

In modern societies, social inequality is often framed as a differentiation of society in distinct groups of similar positions conceivable as 'layers' of the social order on a vertical scale. A general term for this type of social inequality structure is 'stratification.' In the Western tradition, the concept of stratification often – in a more or less explicit way – refers to a ranking of social positions according to the distribution of economic resources; the conflicting groups resulting from this are frequently called 'classes' (in a broad sense). The notion of a society divided into classes, which through their collective action fundamentally shape the dynamic processes of society, has been most widely applied to the specific historical case of Western industrial societies with their market economy, but has also been transferred from here to the description of agrarian societies in ancient or medieval times.

Since around 1970, concepts of stratification and class have been criticized for their neglect of other dimensions of social differentiation and inequality, such as gender, race, or ethnicity. Important as this recent restructuring of the debate on historical (and contemporary) modes of inequality is, stratification and class have been essential concepts for social history in the past (and appear to remain so in the future, if in a modified way). Indeed, it may be said that the development of social history as a field of historical research is hardly conceivable without the analytical concepts of stratification and class. In considering social inequality, stratification, and class in history, two perspectives may be taken. First, stratification is an empirical problem, a problem of historical subject matter. In this regard, one can follow the development of systems of social inequality throughout history, from antiquity to the twenty-first century. Second, stratification refers to an analytical and theoretical problem in (social) history; it appears as a problem of conceptualizing historical processes and structures. In this vein, one can look at concepts, approaches, and methods in research on stratification and class, and consider their development, especially since the nineteenth century, and their usage in different historiographical traditions.

This article will concentrate on the second, the conceptual level. However, both perspectives are intimately linked, since there is no empirical description of inequality without an acute awareness of concepts and language. However, for historians (as opposed to social theorists), all methodological discussion eventually points to empirical research in order to better understand systems of social inequality in the past, and therefore, in the present.

Theories of Social Stratification and Class Formation: Historical Development Through the 1960s

Even before the rise of the modern social sciences with their more elaborate theories of social structure, the explanation (and justification) of inequality was an important part of intellectual thought in advanced, literate societies, often within a general philosophical or theological context. In ancient Athens, Aristotle regarded society as consisting of three hierarchically ordered groups, among which he favored the middle, or *'mesor,'* as being most important for the stability of the *polis*.

Aristotle's scheme of 'stratification' remains influential in its attempts at defining society in terms of 'upper,' 'middle,' and 'lower' classes or strata, among which the middle class is held in highest regard (e.g., in political culture and public discourse in the USA).

In feudalist Western Europe during the Middle Ages, a different kind of tripartite scheme of inequality gained prominence: the idea of a functional differentiation of society into the three orders of Christian clergymen, noblemen, and peasants. With the rise of cities and the emergence of town burghers as an influential group, the functionalist scheme was adapted to new social realities, but it remained a central concept for the interpretation of inequality.

While these theories are themselves a subject of historical inquiry because they tell a lot about the framing and self-understanding of premodern societies (Duby, 1978), a major shift toward modern concepts of stratification and class only occurred in the late eighteenth and nineteenth centuries. In the historical and philosophical works of the Scottish Enlightenment (e.g., Adam Smith, Adam Ferguson, and John Millar), principal stages of civilizational development were associated with particular systems of social inequality and stratified class structure. The language of social inequality that is still in use, with its typical metaphors of 'class' and 'strata,' is heavily indebted to scientific discourses of the eighteenth century, and especially to the enlightenment's quest for 'classification' (Karl v. Linné) and order in the realms of both nature and culture. In the following decades, with Adam Smith and David Ricardo providing important linkages, attention in the theory of stratification increasingly shifted toward a search for the origins of inequality in economic systems of production.

The most powerful, and historically most influential, theory of class as a product of basic economic forces was for a long time Karl Marx' (and Friedrich Engels') concept of Historical Materialism, a theory according to which at each stage of society, an oppressed lower class rises against the economic and political dominance of a ruling class, and eventually overturns the old system through class struggle and revolution. Whatever critical may be said of Marx' theories and its political consequences, it nevertheless became the major reference point for most theories of inequality and stratification until the middle of the twentieth century, and possibly also thereafter. For social history in particular, it provided an extremely suggestive framework for understanding historical change in terms of economic progress, class formation, and societal conflict, instead of referring to diplomacy, politics, and the 'great men' as driving forces of history, as was the case in much traditional nineteenth- and early twentieth-century history. After Marx, theories of inequality based on functional differentiation and the division of labor played an important role for a while (e.g., with Herbert Spencer in the UK, Gustav Schmoller in Germany, and Émile Durkheim in France), but they too were primarily grounded in the idea of economic foundations of particular class systems.

At the beginning of the twentieth century, German sociologist Max Weber in some ways moved class analysis back to Marxian scope and intentions, while on the other hand he developed, if in a fragmentary manner not easy to grasp for later scholars, a broader, historically more flexible, and less dogmatic, framework for the interpretation of social inequality and class structures. Although Weber, like most of his contemporaries, under the impression of industrial capitalism's triumphant advance, acknowledged the enormous impact of economic forces on the molding of society, he was also prepared to systematically consider other dimensions of stratification. Political power, in Weber's view, often played a crucial role in determining positions in a class system; and also inequality for him appeared as fundamentally shaped by 'culture': by general interpretations of the world and the social order as, for example, provided by religion (*Weltbilder*), as well as by the honor, prestige, or status attributed to an individual (or to members of a larger group, such as a specific profession) within the normative setting of a society (see Weber, 1978).

After a brief period of neglect, Weber's concepts of social inequality and social class exerted a strong influence on both American and European social science after 1945. In postwar American sociology and political science in particular, stratification became one of the dominant topics in the field, a topic that all major 'schools' and theoretical currents had to reckon with (see Barber et al., 1968). Despite an enormous variety of approaches and methods, some common features of stratification research in the 1950s and early 1960s stand out, making them attractive for historians. Often, they were not just present-minded, but embedded in a macrosociological and historical framework of thought. Typically, they were interested in linkages between class and political power structures, and in the cultural dimensions of inequality as expressed in neo-Weberian concepts like 'status' or 'prestige' (Bendix and Lipset, 1953; Lenski, 1966).

Since the late 1950s, Weberian and functionalist research in stratification began to overlap with a renaissance of critical theories of society and inequality that were attempting to benefit from the Marxian tradition in an undogmatic manner. Authors like Ralf Dahrendorf or Anthony Giddens were widely read for their innovative intellectual combinations of Marxism and liberalism, or Marxian and Weberian traditions (see Dahrendorf, 1959; Giddens, 1973). In Western social science, the revival of the Marxian tradition proved short lived and mostly faded away in the late 1970s, leaving some strands of cultural Marxism as survivors into the then emerging new age of culturalism in the social sciences. However, Weber's influence, because of the historical depth of his work and a rare combination of flexibility and analytical precision in his categories, never quite vanished, and his work continuously remained inspiring for historians doing research on stratification and class. At the same time, Weber proved remarkably resistant against changing intellectual moods since the late 1970s and 1980s, when classical concepts of stratification and class came under attack. He thus continued to be a starting-point in some of the most important recent approaches in the sociology and social history of inequality and class.

Stratification and Social Classes in Historiography: Trends and Examples

With the rise of social history to intellectual dominance in the 1960s and 1970s in most Western countries, research on the history of stratification and class, mostly in modern societies since the Industrial Revolution of the nineteenth century,

became a central historiographical theme, and even a topic defining what the 'New Social History' was all about. In the general political and academic climate of the 1960s, with its social-democratic air and its resurgence of Neo-Marxian theories, interest for the less privileged groups of industrial society – and especially for the working class – spilled over into history and not only generated a host of special studies on stratification, mobility, and class formation, but also influenced synthetic approaches in social history and the general discourse of historians in that field. While this was a pervasive international tendency, it rendered different results within the national traditions of social history. Therefore, Germany and the US will be considered as two important examples.

Germany

With its strong methodological roots in nineteenth-century historicism and its political conservatism in the historical profession, Germany was a latecomer in the development of postwar social history, and in many respects remained dependent on the reception of theoretical innovation from other countries, such as France, the UK, and the USA. However, due to its strong traditions in social theory, and the theory of class in particular (Marx, Weber), and also because of the great historical significance of class structures in late nineteenth- and early twentieth-century Germany, topics like stratification and class formation quickly took center stage in West German social history during the late 1960s and 1970s. To a larger degree than in other countries, stratification and class emerged as the paradigmatic categories of research in social history (see, e.g., Kocka, 1975), a tendency that came under attack in the 1990s as competing concepts such as gender and ethnicity by many historians were regarded as at least equally important.

For the history of stratification and class in the 1970s and 1980s, two approaches deserve special mention. First, as was the case elsewhere, the history of classes in the beginning concentrated on the less privileged groups and lower strata of society. Here, a particular focus was on the formation of a modern, industrial working class in the nineteenth century, and on the problem of the homogenization of that class from the origins of a multitude of preindustrial, agrarian, and artisanal occupations and interests. In what he termed the 'Weberian application of a Marxian model,' historian Jürgen Kocka studied the formation of a German proletariat, which he conceived as a Marxian class in its basic socioeconomic determination, and at the same time as a Weberian ideal type with which reality can never fully correspond (see Kocka, 1980, 1987, *see* Labor History; Working Classes, History of). Class formation in this approach was considered a historical process progressing through different levels. It proceeded – ideally, though not necessarily in historical reality – from economic homogenization in the workplace toward increasing social and cultural coherence and might end with class consciousness and the political organization of the working class. Since the 1980s, class formation models have also been applied to the upper strata of German society, particularly the bourgeoisie and the professional middle classes, but it appeared that in this case, 'culture' seemed to operate as a much more influential determinant of class than a common economic position.

Second, German traditions in social theory fostered an unusually strong interest in the historical conceptualization of the overall class structure and stratification system, again with an emphasis on the period of German modernization and industrialization from the late eighteenth century onwards. In conceiving social history as the 'history of society,' and the history of society as grounded in structures and processes of social inequality, stratification, and class, German historian Hans-Ulrich Wehler in a very influential project attempted a synthesis of modern German history in terms of a history of inequality and class. His *Gesellschaftsgeschichte* in a systematic way asks for the transformation of the estate system of preindustrial society into an industrial society comprised of social classes, with the bourgeoisie and proletariat in the center, but also giving attention to the nobility and to agrarian classes, such as farmers or farmworkers (Wehler, 1987–95). Typical for much German social history in the 1970s and 1980s, Wehler's approach is methodologically under the heavy influence of Max Weber's categories; and it politically adheres to a vision of an increasingly equal, less class-rifted (if not classless) society.

While most other German authors are not as ambitious and comprehensive as Wehler in their concern with social structures, they have, however, been more inclined to expand the explanatory framework of social inequality into the realm of culture. A new focus of research is on the complicated relationship between social structure on the one hand, and cultural perceptions and practices on the other hand.

The United States

In its self-image, the US has been born as a country of equality, defined by a lack of class and stratification in the European sense, and by the predominance of a broad middle class. The fact that the US has never known an estate system of social order has indeed, in combination with the country's more pragmatist tradition in social thought, exerted a deep influence on social scientific and historical approaches toward class structure and socioeconomic inequality in America. On the other hand, critical strands in social theory and historiography, which were able to identify and discuss inequality of class, appeared earlier than in Germany and remained stronger in the first half of the twentieth century – with the 'Progressive historians' such as Charles Beard being the most important example.

During the 1960s, the blossoming of the 'New Social History,' together with a more self-critical mood in American society in general, prompted a new interest in the history of social class and inequality in America, much in the same way as in other Western countries at about the same time. With regard to approach and method, historians often concentrated on less privileged groups, such as workers and immigrants, and they often did so in case studies of particular cities and through the study of social mobility among the lower strata of urban society (Thernstrom, 1964). For some time around 1970, Neo-Marxist approaches to the study of class and 'class struggle' in American society drew some attention, but on the whole, even leftist historians of class and inequality undertook

much less theoretical effort in conceptualizing their topics than in Germany.

Similar to trends in European social history, however, quantitative methods (as in the study of social mobility) since the late 1970s lost much of their earlier appeal. Instead, the problem of class was more often approached through qualitative, even hermeneutical methods, and through the lenses of class culture and class experience. Also, outside the somewhat narrowing field of working class and labor history, the study of social classes shifted its main focus from the lower classes to the middle classes and their origins in nineteenth-century social change (e.g., Blumin, 1989; Zunz, 1990). This shift may be understood as an attempt to deconstruct the pervasive myth of the American middle class, and at the same time to historically explain the middle-class orientation of twentieth-century America.

In a broader perspective, the study of stratification and class as such after the 1970s retreated from the center of concern for social historians much more quickly, and more pervasively, than in Europe. Obviously, the most important reason for this is the existence of other, competing dimensions of inequality in American society, especially the inequality of race and ethnicity. Class formation in the US rarely occurred in ethnically homogeneous milieus, and often racial and ethnic identity turned out to be more important, and certainly of a much more durable character, than class identity and economic position.

For the same reason, synthetic efforts with generalized views on American society from the perspective of stratification and class were hardly undertaken by American historians. The one large project that probably comes closest to such a synthesis is the two-volume textbook 'Who Built America?' collectively conceived and written by the American Social History Project under the intellectual guidance of labor historian Herbert Gutman (See, American Social History Project 1989–92). The book attempts a synthesis of American history from colonial times to the present by centering its narrative around the development of social classes – particularly the industrial working class – and the struggle of less privileged groups for participation in the economic and political resources of 'mainstream' America. At the same time, the authors' efforts demonstrate that a narrative of American social history from the viewpoint of stratification and class alone cannot make sense; it needs to be expanded by, and connected with, additional categories of inequality, such as race, ethnicity, and gender.

The Challenge of Culture in the History of Stratification and Class

The discussion of the two national examples, especially the case of the US, has already hinted at important new trends in historical research that have begun to change the field in fundamental ways since the 1980s and will probably continue to do so in the future. Although many of the recent tendencies stem from similar motives – e.g., the increasing dissatisfaction with interpretations of class based on the structures of male industrial work – and therefore often overlap in historiographical practice, it is useful to distinguish two horizons of criticism and expansion of the classical stratification and inequality paradigms. Some scholars, for different reasons, fundamentally question the relevance of concepts, such as 'class' and 'stratification' and try to develop alternative concepts, or alternative theoretical and methodological viewpoints on the problem of inequality in history (see Section Beyond Class, Beyond Stratification?). Others consider 'class' and 'stratification' as still important, or even indispensable notions for social history, yet argue that they have to be redefined in order to meet the challenges of social reality and social science at the turn of the twenty-first century. This ongoing redefinition of class is in most cases centered around the idea of 'culture.' Culture, especially in the forms of 'language' and 'experience,' is now widely believed to be the prime factor in determining stratification systems and class structures. However, at a closer look, the challenge of culture differentiates into a number of theoretical and empirical approaches.

Class as Social Action, Experience, and Culture

In 1963, the English Marxist historian E.P. Thompson published his two-volume 'Making of the English Working Class,' which is arguably among the dozen most influential works in twentieth-century social history (Thompson, 1963). At that time, classical approaches to class formation grounded in systematic sociology or political science were still widely in use, and in some countries just beginning to influence the expanding field of social history. Thompson's massive study was rather conventional in the sense of its practical empiricism and its Marxian notion of a working-class winning identity and togetherness in the struggle against its bourgeois class enemy. What was new, however, was Thompson's insistence that 'class' was not a quasi-objective structure determined by certain economic forces, but a very fluid result of practical action and social experience on the part of the lower classes in the decades around 1800. As Thompson largely focused on preindustrial work and living experience, he discovered the roots of class in traditions and reinventions of popular culture. In many respects, Thompson's book was ahead of his time. Reception and debate in Germany, for example, only took off in the 1980s and turned his pioneering work into a major stimulant for the then emerging culturalist turn in social history (*see* Cultural History).

Class Experience: From Production to Consumption

Classical approaches to the history of stratification, from the enlightenment theories of a division of labor to Marx and beyond, have based their identification of certain classes, and their description of the overall inequality structure of society, on the sphere of production, and on the relative position of an individual (mostly seen as a male adult and family head) acquired in that society's system of work and employment. As the nineteenth-century fascination with factory work and industrial production faded and the twentieth century increasingly developed the self-image of a consumer economy and society, it seemed appropriate to search for social spheres of class formation outside the workplace experience: at home, in leisure activities, and especially in

consumption. In American and in British historiography in particular, the history of consumption has been one of the most significant discoveries for the field of social and cultural history in the 1990s (*see* History of Consumption), and several attempts have been made to conceptually link consumption and class, some of which build on much earlier efforts, such as Thorstein Veblen's 'Theory of the Leisure Class' with its ideas about the role of 'conspicuous consumption' for the life of the upper classes.

The notion of consumption can play a key role because of its ability to connect the perspective of economy – the appropriation of goods by groups with different economic resources – with the perspective of culture – insofar goods serve as symbols and as expressions of (individual or collective) identity and lifestyle. It also allows for a more adequate consideration of female and private-life contributions to the formation of social inequality and class. In an important empirical study on Chicago in the 1920s and 1930s, historian Lizabeth Cohen investigated the formation of class identity among ethnic workers and their families through their shared experience of consumption and modern mass culture institutions like radio (Cohen, 1990). As the history of consumption still flourishes at the turn of the twenty-first century, it is likely that its impact on the history of inequality and class will grow even larger in the near future.

Pierre Bourdieu: Class and Cultural Distinction

The first two examples for the cultural expansion of the history of class represent developments within the historical profession, yet trends and new concepts in other social scientific disciplines have also had their impact on the topic. While national traditions of thought remain influential – with Max Weber, for example, still being of prime importance in Germany, and anthropologist Clifford Geertz stimulating American debates – in international perspective one scholar's work in particular has received near unanimous attention and acclaim by social historians during the 1980s and 1990s, namely French sociologist Pierre Bourdieu's theory of class and cultural distinction as developed in his major work *La distinction* (Bourdieu, 1979).

Arguing from his research on contemporary French society, Bourdieu secures a central place in social research for the problem of inequality and even sticks to the concept of class. Yet in a brilliant way he enlarges that concept by understanding class as the result of a complex ensemble of not just economic, but also social and cultural resources or types of 'capital.' Similar to the category of 'prestige' in 1950s American sociology, Bourdieu defines an individual's position in society in terms of his or her 'distinction' and 'habitus,' a concept referring to sociocultural outlook and self-stylization. Though obviously fitting the self-oriented consumer and leisure society of the later twentieth century especially well, the appeal of Bourdieu's approach also extends to the analysis of more traditional societies in the early modern era, and even to non-Western societies, not least because of Bourdieu's own intellectual origins as a social anthropologist. As Bourdieu at the same time draws on 'classical' authors of stratification research, such as Weber and Marx, his work is extremely important in providing a bridge between the classical theoretical tradition and the more recent approaches that center around language, experience, culture, and consumption.

Beyond Class, Beyond Stratification?

While the 'culturalist turn' in the humanities has substantially expanded and reframed perspectives on stratification and class in history since around 1980 other scholars have gone a step further in arguing that the conceptual framework of stratification and class might better be abandoned altogether and replaced by alternative concepts of social order. However, that does not mean that they would deny the existence, and empirical relevance, of stratification systems and class formation under certain historical circumstances. Rather, the argument is that the category of class itself is a historical product of the nineteenth century, and that therefore the continuing conceptual primacy of class endangers the ability of scholars to recognize different lines of social division and inequality.

Gender, Race, and Ethnicity

This point has been most persuasively made by feminist social scientists and historians who have argued that the predominance of class in social history has in fact only been able to describe the male half of the social world – and even that in an inadequate manner, because class formation among male individuals (or, household heads) cannot be conceived of independent from the family and gender systems out of which the sphere of production and patriarchy grows (see Canning, 1992). In a similar way, historians of race and ethnicity have maintained that the stratification of a society in terms of individuals' economic positions never functions as an independent variable in multiracial societies, such as the US, Brazil, or South Africa – nor even in European societies which are increasingly being shaped by their immigrant populations. It has hence become almost a standard method in American, and now also in British social history, to frame the problem of social inequality in the trias of 'race, class, and gender.' In the view of many feminist historians, however, this solution is a dissatisfactory compromise, because it implicitly affirms a conceptual primacy of class by molding the two other concepts after the example of class, or as mere expansions of it (see Scott, 1986).

Therefore, the history of gender is in need of developing specific theoretical foundations out of twentieth-century social theory in order to establish itself as an independent approach in social studies, and not just a part of the 'litany of class, race, and gender' (Scott, 1986: p. 1055). In the near future, we will perhaps see more theoretical efforts, and empirical historical studies, going in that direction, not just in gender history, but also in the history of ethnicity.

From Classes to Lifestyles, 'Milieus,' and Radical Individualism

A completely different line of thought, which also calls traditional concepts of stratification and class into question, has emerged in recent sociological work on contemporary western societies in their 'postindustrial' stage. As industrial work, and

the labor and class relations so intimately connected with it, loses its central status for the definition of modern societies, these societies may no longer be adequately described as stratified class societies. The large, socially coherent and even politically self-conscious groups, such as the 'working class' are considered to be rapidly dissolving, without new, equally coherent group structures building up. Instead, social relations in modern society may be described as based on radical individualization, and a similarity of social status at best finds expression in rather diffuse social 'milieus' (see Beck, 1986; Hradil, 1987). Western societies, in this perspective, are not becoming completely leveled societies with citizens of equal material resources, but the absence of material wants, and a redefinition of the value system toward postmaterial orientations, dramatically diminish the relevance of occupational situation, income, and even class-specific consumption patterns.

As 'social inequality' (in the form of social strata and classes) was the defining element of nineteenth- and early twentieth-century societies, these authors hold, categories like 'risk' or 'information' are defining societies on the verge of the twenty-first century. Although not yet widely received in social history, these new developments in theoretical and empirical sociology will probably have a considerable impact on historical views on inequality and social structure in the future. They offer important insights into the now rapidly expanding field of twentieth-century social history, but it may also be asked whether societies before the industrial age can be interpreted beyond the classical thought system of social estate, class, and stratification.

Constructing Class, Constructing Inequality: From Economic Realism to Cultural Relativism

With the class structure of industrial societies appearing more as an historical exception rather than historical normalcy, and with the interest in the socioeconomic foundations of inequality and class diminishing, historians have become more sensitive to the relativist or constructivist character of social structure in general and of 'class' in particular. Especially in British social history, questions have been asked about the 'language of class,' that is, about the use of concepts, such as class by the historical actors themselves, who were trying to express collective self-identities by communicating new notions of social bonding. With regard to the working classes, the main thrust of this research was on political organization and political action (e.g., in English Chartism), and the objective nature of class was hardly fundamentally doubted (see Stedman Jones, 1983).

Especially in the 1990s, however, and with empirical interest somewhat shifting to the middle classes, the language approach was radicalized under the double influence of cultural and linguistic theory on the one hand, and of 'radical constructivism' in epistemology and the sociology of knowledge on the other hand. Stratification and class were now analyzed as 'imaginary' realms of communication and culture; they figured as notions that expressed 'visions' of the social order as perceived by the contemporaries, rather than realities of economic and power structure as described by social scientists and historians (see Corfield, 1991; Joyce, 1991; Nolte, 2000; Wahrman, 1995). Some of these approaches fundamentally undermine any 'realistic' concept of social inequality and class, and therefore sound like a triumphant signal for the final abandonment of any social scientific or historical study in stratification and class structures. Yet at the same time it can be argued that the language-oriented and constructivist approaches have introduced important expansions to the more traditional methods of social history. They have provided bridges between social and intellectual history and have induced historians to a more careful usage of language and concepts. Also, ironically, in devoting so much energy to the problem of class in historical language and self-consciousness, this work has also underscored the importance of inequality and class in past societies.

Conclusion

After a long period of undisputed reign as a kind of master concept in the study of present and past social structures, theories of stratification and class are now undergoing a period of substantial challenge and criticism, redefinition, and even rejection. The question may thus arise whether theoretical and empirical interest in social inequality will persist in the future, especially in the discipline of history, which is now characterized by a very pronounced shift away from problems of socioeconomic structure in favor of a history of culture, symbolism, and subjectivity.

It is almost certain that inequality, stratification, and class, will lose (and already have lost) some of their original relevance as defining ground for social history as such, when compared with the 1950–70s boom of those topics in the social sciences. In addition, the emerging intercultural debates on the history of non-Western societies may prove that concepts such as 'class' and 'social strata,' which are obviously deeply rooted in the Western tradition of thought and culture since classical antiquity, are not suitable instruments for analyzing Indian, or Central African, societies, be it past or present. As 'postcolonial theories' in the social and cultural sciences describe the 'hybrid' character of such societies, the hybridity of class and competing principles of social relations may also be a distinctive mark of their inequality structures.

On the other hand, three points may be made, which serve as a reminder for the continuing importance, not only of the general notion of social inequality, but also of more specific concepts like stratification and class. First, social inequality – whatever it is grounded on, and in whatever way it is communicated – is an anthropological feature of all human civilization, not only in the West. Second, inequality has often taken the form of a hierarchical order, or 'stratification,' of distinct groups according to their access to power, their rights, and material well-being. Moreover, the hierarchical mode of social differentiation is a widespread mark of the less advanced, less technologically and culturally complex societies in which historians are often especially interested. Third, it is hard not to acknowledge the success, and the astonishing flexibility, of the concept of 'class' for describing systems of inequality. As new and intellectually promising approaches have been developed since around 1980 that link 'culture' with 'class,' and as these approaches are now applied in empirical research, there can be

little doubt about the fact that the history of inequality, stratification, and class will retain a prominent place in the field of social and cultural history.

See also: Class: Social; Equality and Inequality: Legal Aspects; Inequality: Comparative Aspects; Marx, Karl (1818–83); Mobility, Social; Social Stratification; Weber, Max (1864–1920).

Bibliography

Akyeampong, E.K., 2006. Themes in West Africa's History. Ohio University, Athens.

Barber, B., Lipset, S.M., Hodge, R.W., Siegel, P.M., Stinchcombe, A.L., Rodman, H., 1968. Stratification, social. In: Sills, D. (Ed.), International Encyclopedia of the Social Sciences, vol. 15. Macmillan and Free Press, New York.

Beck, U., 1986. Risikogesellschaft. Auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt, Germany.

Bendix, R., Lipset, S.M. (Eds.), 1953. Class Status, and Power: Social Stratification in Comparative Perspective. Free Press, New York.

Blumin, S.M., 1989. The Emergence of the Middle Class. Social Experience in the American City, 1760–1900. Cambridge University Press, New York.

Bourdieu, P., 1979. La distinction. Critique sociale du jugement. Editions de Minuit, Paris.

Canning, K., 1992. Gender and the politics of class formation: rethinking German labor history. American Historical Review 97, 736–768.

Cohen, L., 1990. Making a New Deal: Industrial Workers in Chicago, 1919–1939. Cambridge University Press, Cambridge, UK.

Corfield, P.J. (Ed.), 1991. Language, History, and Class. Blackwell, Oxford.

Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford, CA.

Duby, G., 1978. Les Trois Ordres ou l'Imaginaire du Féodalisme. Gallimard, Paris.

Evers, S., 2002. Constructing History, Culture and Inequality: The Betsileo in the Extreme Southern Highlands of Madagascar. Brill, Leiden.

Giddens, A., 1973. The Class Structure of the Advanced Societies. Hutchinson, London.

Gilbert, D.L., 2008. The American Class Structure in an Age of Growing Inequality. Pine Forge Press, Los Angeles.

Hradil, S., 1987. Sozialstrukturanalyse in einer fortgeschrittenen Gesellschaft. Von Klassen und Schichten zu Lagen und Milieus. Leske und Budrich, Opladen, Germany.

Joyce, P., 1991. Visions of the People: Industrial England and the Question of Class 1848–1914. Cambridge University Press, Cambridge, UK.

Kerbo, H.R., 2000. Social Stratification and Inequality: Class Conflict in Historical, Comparative, and Global Perspective. McGraw-Hill, Boston.

Kocka, J., 1975. Theorien in der Sozial- und Gesellschaftsgeschichte. Vorschläge zur historischen Schichtungsanalyse. Geschichte und Gesellschaft 1, 9–42.

Kocka, J., 1980. The study of social mobility and the formation of the working class in the 19th Century. Le Mouvement Social 111, 97–117.

Kocka, J., 1987. Lohnarbeit und Klassenbildung. Arbeiter und Arbeiterbewegung in Deutschland 1800–1875. J H W Dietz, Berlin, Germany.

Lenski, G.E., 1966. Power and Privilege: A Theory of Social Stratification. McGraw Hill, New York.

Levine, B., et al., 1989–92. Who Built America? Working People and the Nation's Economy, Politics, Culture, and Society, 2 vols. Pantheon, New York.

McIvor, A., 2001. A History of Work in Britain, 1880–1950. Palgrave, Houndmills, Basingstoke, Hampshire.

Midgley, J., Piachaud, D., 2011. Colonialism and Welfare: Social Policy and the British Imperial Legacy. Edward Elgar, Cheltenham, UK.

Mora, S.N., 2006. Exploring Changes in Earnings Inequality during Industrialization: Barcelona, 1856–1905. University of Oxford, Oxford.

Nolte, P., 2000. Die Ordnung der deutschen Gesellschaft: Selbstentwurf und Selbstbeschreibung im 20. Jahrhundert. Beck, Munich, Germany.

Ringer, F.K., 2004. Max Weber: An Intellectual Biography. University of Chicago Press, Chicago.

Scott, J.W., 1986. Gender: a useful category of historical analysis. American Historical Review 91, 1053–1055.

Stedman Jones, G., 1983. Languages of Class: Studies in English Working Class History, 1832–1982. Cambridge University Press, Cambridge, UK.

Thernstrom, S., 1964. Poverty and Progress: Social Mobility in a Nineteenth Century City. Harvard University Press, Cambridge, MA.

Thompson, E.P., 1963. The Making of the English Working Class. Penguin, Harmondsworth, UK.

Turner, S.P., 2000. The Cambridge Companion to Weber. Cambridge University Press, New York.

Wahrman, D., 1995. Imagining the Middle Class. The Political Representation of Class in Britain, 1780–1840. Cambridge University Press, Cambridge, UK.

Weber, M., 1978. Economy and society. In: Roth, G. (Ed.), An Outline of Interpretive Sociology, 2 vols. University of California Press, Berkeley, CA.

Wehler, H.-U., 1987–95. Deutsche Gesellschaftsgeschichte, 3 vols. C H Beck, Munich, Germany.

Zunz, O., 1990. Making America Corporate 1870–1920. University of Chicago Press, Chicago.

Social Influence, Social Psychology of

Y Ito, Keio University, Shonan Fujisawa, Fujisawa, Japan

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7439–7441, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

When an individual's attitude, opinion, cognition, or behavior is coerced it is called 'power,' but when the change is 'voluntary' or 'unconscious' the process is called 'influence.' Social influence is exerted by institutions such as schools and mass media, 'opinion leaders,' and by groups to which people belong or feel attached. The results of empirical research indicate that influence is stronger when the sources of influence are consistent. For example, the influence of mass media is stronger when the content coincides with the opinion of one's 'opinion leader.' When there are discrepancies, the influence of mass media is limited. This explains the mechanism of 'brainwashing' or 'mind control.' However, the mode of influence not only changes attitude through persuasion. The sources of influence can control the salience of people's perception. For example, by determining what is important or what to think at the time of election, mass media and political leaders can influence people's voting behavior. Also, mass media tell people what others think, what is in fashion, and what is passé. Both humans and animals instinctively know that it is safer to belong to a group than to be isolated. Therefore, people tend to be loyal to the norms of the group to which they belong and comply with its mores. In societies where 'honor,' 'shame,' and 'face' are critical virtues, not only the group's members but also 'mere acquaintances' are a source of influence.

When an individual's attitude, opinion, cognition, or behavior is forcefully developed or changed through physical sanctions or punishment, the process at play is called 'power.' However, when the creation or change is 'voluntary' or 'unconscious' the process is called 'influence.' The term 'social influence' in the broadest sense refers to all influences that individuals receive from all types of formal institutions, other individuals, and informal groups.

Formal institutions include schools, mass media, and business corporations. This article focuses on the influence of mass media as an example of the influence of formal institutions, before going on to consider personal influence and informal group influence.

Mass Media Influence

While the influences of home and school are dominant during childhood, the influences of the mass media become increasingly more pervasive after individuals reach adolescence. When radio surged in popularity among the masses during the 1930s, it was feared that mass communications, especially radio, had a powerful and instant effect on a population's attitude and behavior. This fear was reinforced by the success of the Nazis' propaganda machine in Germany and the large-scale panic in the United States in 1938 caused by a single radio program called 'The Invasion from Mars' (Cantril et al., [1940]1966). This radio play, broadcast without any prior warning by Orson Wells's Mercury Theater on CBS, was so realistic and caused such widespread panic that some people actually thought Martians had landed on Earth and sadly committed suicide. As a result, the apparent immediate influence of mass media came to be compared to a fast-acting hypodermic injection or an atomic bomb.

Empirical research on the influence of mass media, however, gradually revealed that it could not be that overwhelmingly influential. It was found that although mass media might reinforce people's existing attitudes and opinions, or even create new ones, they could hardly convince people to change already strongly held beliefs (Klapper, 1960). It was also found that even though mass media information might reach each individual directly, its influence did not. Influence, if any, was thought to reach individuals through so called 'opinion leaders' (Katz and Lazarsfeld, 1955). This view of the mass media's weak influence, however, was later challenged by several new 'powerful media' theories.

One of these theories deals with the 'agenda-setting effect' of mass media (McCombs and Shaw, 1972). Even if the mass media's persuasive power is limited, especially in the short run, they exert a strong influence on people's perception of the salience of each issue. There are 'good news' and 'bad news' stories for everybody and for every organization. If the mass media persistently play up the 'bad news' events surrounding some political candidate or party, it will affect the readers' or viewers' attitudes, and thereby the result of an election. In such a case, the mass media are not actually persuading their readers or viewers to change their minds; they are merely taking advantage of conventional social norms and reinforcing existing attitudes. Nevertheless, mass media can influence individual behavior (such as voting) by setting the priority of people's attention in regard to the issues that are being debated.

Another important role of mass media is that they tell their readers or viewers what other people are thinking. In other words, mass media inform their audience what the prevailing 'public opinion' is regarding each issue. In doing so, mass media tend to amplify the opinions they support. In an extreme case the mass media might treat a minority opinion as if it were that of the majority or some new ineluctable trend. As a result, the holders of the minority opinion, supported by the mass media, become louder and more strident while those of the majority opinion lose their voice and fall increasingly more silent. Thus, through the 'spiral of silence' mechanism orchestrated by mass media, the majority opinion

comes to be replaced by that of the minority (Noelle-Neumann, 1984).

Personal Influence

The 'opinion leader' mentioned above refers not only to professional opinion makers such as journalists, critics, and university professors but to all men and women who influence others around them on a daily basis, in some specific area, such as local politics, fashion, sports, travel, medical care, or even cooking. These opinion leaders are often asked for advice by people who look up to them. To meet such expectations, they expose themselves to the mass media more extensively than their followers. Opinion leaders compare and evaluate opinions and information that are first presented by the mass media. As a result, the influence of mass media first takes place among opinion leaders who in turn regurgitate their latest batch of news to their followers in the form of guidance or advice. Even if a person's opinion is shaped by the mass media, that opinion may be quickly replaced by that of an opinion leader because such people's opinions are trusted more by their followers than are those of the mass media. This process is called the 'two-step flow' of social influence (Lazarsfeld et al., [1944]1968).

Another explanation for this 'personal influence' is provided by the 'cognitive consistency theory.' According to this theory, people are affected by those whom they like because the cognition that one likes someone and the cognition that one's opinion is different from theirs is psychologically inconsistent and therefore uncomfortable. In such a case, it is predicted that people would change their opinion of, or attitudes toward, the person in question (Festinger, 1957; Heider, 1958; Newcomb, 1953).

One of the central studies of social influence was carried out by Milgram (1963). He examined the degree to which individuals could be persuaded to administer 'electric shock' to another person under experimental conditions. The shock was not real, but the compliance to the commands of the experimenter was very real. This study was controversial, and the subject of numerous replications (Quiamzade and Mugny, 2009) and reviews (see Blass, 2004). The effect appears to be robust across cultures (Pascual et al., 2012).

Group Influence

It is known that people tend to conform to the group that they belong to, participate in, or even just happen temporarily to be part of. In other words, the group tends to influence its members. There are several reasons for this conformity. One of them is that other people's behavior in the same group often functions as a stimulus or reminder. For example, if one sees other group members starting to eat, one feels hungry. Another reason is 'fear of isolation.' Not only humans but also many animals know almost instinctively that it is safer to be with the group than to be isolated. The atavistic 'fear of isolation' may be primitive and instinctive but makes sense. It is highly probable that to be isolated costs much more in terms of survival than it would to behave as a group member because, generally speaking, the group has far more information, experience, knowledge, and wisdom than each individual member. There really is safety in numbers.

It is for this reason that each individual usually has several 'reference groups.' 'Reference groups' are similar to 'opinion leaders.' Individuals are loyal to the norms, opinions, and inclinations of their reference groups. These reference groups may be their families, companies, local communities, or governments. Sumner's (1906) notion of an 'in-group' or 'we-group' is a typical example of the 'reference group.' While individuals are loyal to the 'in-group' or 'we-group' they tend to be indifferent or sometimes hostile to the 'out-group,' 'they-group,' or 'othersgroup.'

Social Influence Has Been Examined

Some Japanese sociologists and social psychologists think that what functions as social pressure or constraint on each individual is caused by an abstract social construct called *seken* rather than the influence of specific groups. According to them, individuals are conscious of three different kinds of 'societies.' The first is *miuchi* or *uchi*, which originally meant immediate family but can be applied to close relatives and other immediate-familylike small groups. The second is *tanin* or *soto*, meaning 'others,' 'outsiders,' or 'strangers.' This is not the same as Sumner's 'out-group' because *tanin* is not by any means recognized as a group or a collection of groups. People may be indifferent to *tanin* but are not hostile to it (unlike Sumner's 'out-group'). Finally, *seken* literally means 'world' or 'society.' Specifically, this is the area between the 'immediate family' and the 'others.' This area consists of people whom you know and who know you (other than immediate family members). Friends, neighbors, acquaintances, and relatives are examples.

An important concept used to explain social pressure and constraint is that of shame. Before whom do people usually feel ashamed? People do not feel ashamed before strangers. People do feel ashamed before their immediate family members, but they know that family members will eventually understand and forgive them. It is of *seken* that people are most afraid. *Seken* will remember failure, scandal, or misconduct for a long time and will not forgive as easily as immediate family members might. *Seken* may seem to refer to 'reference groups' but these are different because *seken* is not a group or a collection of groups and people are not always loyal to *seken*.

The same is true of 'keeping face.' Where do people 'lose face'? It is in *seken*. Therefore, individuals are always concerned with *sekentei*. There is no equivalent single word in English, but *sekentei* refers to how your friends, neighbors, acquaintances, and relatives think of you. Thus, according to Japanese sociologists and social psychologists, *sekentei* is a powerful source of social pressure and a ubiquitous constraint on individual attitudes and behavior (Inoue, 1977). This theory may be more applicable to collectivistic societies (where human relations are more important) than to individualistic societies. But even in individualistic societies people do have a sense of 'shame' and 'face,' so this Japanese theory should be helpful in explaining, at least in part, the causes and effects of social influences in individualistic societies too.

See also: Attitudes, Political and Public Opinion; Cognitive Dissonance; Indigenous Psychology; Social Networks.

Bibliography

Blass, T., 2004. The Man Who Shocked the World: The Life and Legacy of Stanley Milgram. Basic Books, New York.

Cantril, H., Gaudet, H., Herzog, H., [1940]1966. The Invasion from Mars: A Study in the Psychology of Panic. Princeton University Press, Princeton, NJ.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Row, Peterson, Evanston, IL.

Heider, F., 1958. The Psychology of Interpersonal Relations. Wiley, New York.

Inoue, T., 1977. Sekentei' no Kozo (The Structure of 'Sekentei'). Nihon Hoso Shuppan Kyokai, Tokyo.

Katz, E., Lazarsfeld, P.F., 1955. Personal Influence. Free Press, Glencoe, IL.

Klapper, J.T., 1960. The Effects of Mass Communication. Free Press, Glencoe, IL.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., [1944]1968. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. Columbia University Press, New York.

McCombs, M.E., Shaw, D.L., 1972. The agenda-setting function of mass media. Public Opinion Quarterly 36 (2), 176–187.

Milgram, Stanley, 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67 (4), 371–378.

Mugny, G., Souchet, L., Codaccioni, C., Quiamzade, A., 2008. Représentations sociales et influence sociale. Psychologie Française 53, 223–237.

Newcomb, T.M., 1953. An approach to the study of communicative acts. Psychological Review 60, 393–404.

Noelle-Neumann, E., 1984. The Spiral of Silence: Public Opinion—Our Social Skin. University of Chicago Press, Chicago.

Pascual, A., Oteme, C., Samson, L., Wang, Q., 2012. Cross-cultural investigation of compliance without pressure the "you are free to…" technique in France, Ivory Coast, Romania, Russia, and China. Cross-Cultural Research 53, 223–237.

Quiamzade, A., Mugny, G., 2009. Social influence and threat in confrontations between competent peers. Journal of Personality and Social Psychology 97, 652–666.

Sumner, W.G., 1906. Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores and Morals. Ginn, Boston.

Social Influence, Sociology of

Robert H Gass, California State University, Fullerton, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social influence involves intentional and unintentional efforts to change another person's beliefs, attitudes, or behavior. Unlike persuasion, which is typically intentional and requires some degree of awareness on the part of the target, social influence may be inadvertent or accidental. Social influence often operates via peripheral processing. Hence, the target may be unaware of the influence attempt. Unlike compliance gaining, which is usually goal directed, social influence is often nongoal directed and the outcomes may be inconsistent with, or unrelated to, a communicator's goals. Social influence encompasses such strategies as indebtedness or reciprocity, commitment, social proof, liking and attractiveness, authority, and scarcity.

Introduction and Definition

Social influence is a fact of everyday life. Humans are, after all, social creatures who spend a good deal of their time trying to influence one another as well as resist each other's influence attempts. Not only is social influence a ubiquitous feature of everyday interaction, it is also prevalent in the boardroom, the courtroom, the classroom, and many other walks of life. As Dahl notes, "social influence is incredibly broad, covering everything from mere presence effects and mimicry to more direct forms of social persuasion" (p. iii). Forsyth (2013) conceptualizes social influence as "interpersonal processes that change people's feelings, thoughts, or behaviors" (p. 305). Social influence may be defined as intentional or unintentional communication that produces change(s) in another's attitudes, beliefs, intentions, motivations, or behaviors. Pratkanis (2007) further adds that social influence is noncoercive in nature. Influence does not involve force.

The term communication is used broadly here and encompasses not only nonverbal cues and physical artifacts (such as clothing and the layout of a room), but cultural norms and socialization processes as well. Change(s) produced via social influence may not be immediately apparent. Because exposure to influence attempts may grow on people, a delayed reaction may occur. As Mason et al. (2007) note "influence models that address only a single time point are limited because influence is an inherently dynamic process—particularly when one takes account of the reciprocal and multidirectional influences that can occur among multiple sources and targets" (p. 281).

Social Influence, Persuasion, and Compliance Gaining

Social Influence versus Persuasion

To fully understand the nature of social influence, it is important to distinguish it from other closely related concepts, such as persuasion and compliance gaining. Persuasion is most commonly viewed as an intentional effort to alter another's attitudes, beliefs, or behaviors (see, for example, Perloff, 2014: pp. 18–19). While inadvertent influence is possible, then, 'accidental' persuasion is not. Some scholars also argue that persuasion requires the conscious awareness of the persuadee (Simons and Jones, 2011). In contrast, social influence does not require the awareness of the target(s) of an influence attempt. Some also argue that for persuasion to occur there must be an observable, measureable effect. Others disagree, noting that effects can be difficult to measure, may not be immediately apparent, and may not be what the persuader originally intended (Gass and Seiter, 2014). Psychological reactance or a 'boomerang' effect, for example, may occur. Research on social influence also emphasizes effects, but there may be latency effects involved. Furthermore, because individuals often use multiple influence tactics simultaneously, it may be difficult to isolate the effects of a particular strategy.

Social Influence versus Compliance Gaining

Compliance gaining is viewed by almost all scholars as an intentional effort to change another's behavior. Compliance gaining is usually other directed. Compliance gaining is also goal directed and results oriented. As its name implies, the emphasis is on compliance or behavioral conformity to the message recommendations. Less emphasis is placed on cognitive changes within the receiver. Like compliance gaining, social influence is other directed. Influence tactics are designed to produce changes in others, although intrapersonal influence may also occur. Unlike compliance gaining, social influence is often nongoal directed and the outcomes may be inconsistent with, or unrelated to, a communicator's goals. For example, a child may model a parent's bad behavior, such as swearing or smoking, contrary to the parent's wishes. With these distinctions in mind, social influence may be seen as an umbrella term that encompasses both persuasion and compliance gaining. Social influence can take place unintentionally or tacitly, may operate at a low level of awareness or without the conscious awareness of a receiver, and the effects of such influence processes may not be readily observable.

Origins of Social Influence Research

Social scientific research on persuasion dates back to Gordon Allport (1935) and Carl Hovland's Yale Attitude Research Program (Hovland et al., 1953). Studies on social influence date back at least 50 years, although they were not integrated

under the rubric of social influence until the 1980s. In this regard, Smith et al. (2011) observed "it is interesting to note that many of the classic studies in social psychology are, at their core, studies of social influence" (p. 599). Research on compliance gaining is more recent and can be traced to Marwell and Schmitt (1967) and Wiseman and Schenck-Hamlin (1981). Research on social influence spans a number of fields including communication, psychology, social psychology, sociology, consumer behavior, and behavioral economics, among others. Studies in this area have examined a wide range of topics including conformity, social proof, normative influence, modeling of behavior, mimicry, psychological reactance, and psychological commitment, to name but a few.

Processes of social influence finally coalesced with the publication of Robert Cialdini's landmark text, *Influence: Science and Practice*, in 1984. The legitimacy of the field of study was established more firmly in 2006 with the launch of the journal *Social Influence*. Although some studies have explored the neurological bases of social influence (Izuma, 2013), the focus here is on communication phenomena – both messages and behavior – that facilitate influence.

Theoretical Frameworks

A number of dual processing models maintain that messages are processed via one of two routes. Central processing (Petty and Cacioppo, 1986a,b), or what some refer to as systematic processing (Chaiken and Trope, 1999) is thoughtful, effortful, and deliberative. Receivers who rely on central processing scrutinize message content carefully. The alternative route, peripheral processing (Petty and Cacioppo, 1986a,b), or what some call heuristic processing (Chaiken and Trope, 1999) relies more on mental shortcuts. Receivers who default to this route tend to respond reflexively to messages. Although receivers may use both routes at once, they typically favor one over the other. Peripheral processing is closely related to what Cialdini calls automaticity, whereby people exhibit knee-jerk reactions that are quick and efficient, but often lead to poor decisions. Simple language triggers, such as the words 'free' or 'save,' for example, may trigger impulse purchases by consumers.

Social influence often takes place via peripheral processing. When people have low involvement or little motivation to pay attention to an issue, they tend to rely on mental shortcuts and superficial thinking. They are less thoughtful and more likely to respond based on appearance cues or social norms. The mere presence of other persons, for example, may make watching a comedian more funny or viewing a sporting event more exciting. Individuals often rely on heuristic cues. A consumer, for example, might assume that the more expensive of two items must be of higher quality. A voter might decide a political candidate is credible based on the candidate's attractiveness. A good deal of social influence, therefore, takes place at a low level of awareness. For example, males rated a female who was wearing a red blouse as more sexually attractive than the same female wearing a blue blouse (Elliot and Niesta, 2008). Color functioned as a peripheral cue in their ratings. As another example, customers in a wine shop purchased more French wine when French music was playing in the background and more German wine when German music was playing (North et al., 1999). Background music thus served as a peripheral cue that activated wine preferences.

In addition to dual processing models, a variety of other theories and models help to account for social influence effects. These will be discussed in relation to particular strategies and tactics, beginning with those identified by Cialdini.

Cialdini's Principles of Influence

In some respects, Robert Cialdini can be credited with founding social influence as a field of study. Although, he did not conduct the original studies, he was the first to systematically integrate research from a variety of disciplines under the heading of social influence. Cialdini (2009) identified six core principles of social influence which are outlined here.

Reciprocity

When one person performs a favor for another, the other person feels obligated to return the favor. One good turn deserves another, as the saying goes. According to Cialdini, this proclivity to reciprocate favors is culturally universal. Performing a favor or giving someone a gift – even if the favor or gift was not requested – creates a sense of obligation or indebtedness in the recipient. Initially, the target may be unaware of the social *quid pro quo*. What is more, such pregiving can result in unfair exchanges. The source enjoys the advantage of deciding on the initial favor as well as the return favor. A neighbor, for example, could retrieve another neighbor's trash cans from the curb on trash day. A few days later, the helpful neighbor might ask to borrow the second neighbor's lawn mower. As the recipient of the uninvited favor, it would be harder for the second neighbor to say no. Similarly, a wealthy voter might make a sizable campaign contribution to a candidate. The donor would not necessarily be buying the candidate's vote, but if the candidate won, the donor could expect some sort of return favor, whether it was a simple photo opportunity, a sit-down meeting, or support for a particular piece of legislation. The norm of reciprocity is a powerful tool for influence. Once established, a social 'I owe you' (IOU) can be difficult to resist without seeming ungrateful or ill-mannered.

Commitment and Consistency

People like to think that they are consistent in their beliefs, attitudes, and behaviors. It bothers them to say one thing and do another. They also do not want to be perceived by others as hypocritical. The need for consistency is partly a psychological drive and partly a social desire to maintain face. When a person perceives that his or her attitudes, statements, or actions are in conflict, the person is motivated to reduce the perceived incompatibility. The person might go about doing this in a variety of ways, including denial, bolstering, rationalizing, differentiating, or other techniques designed to reduce angst.

Public commitments are more binding than private commitments. Hence, wedding rituals such as engagement rings, wedding announcements, public vows, and wedding receptions increase the likelihood that the bride and groom will follow through with the nuptials. Commitment contracts and

devices, such as New Year's resolutions, chastity rings, or swear jars are also ways of increasing an individual's adherence to a particular form of conduct. Cialdini notes that commitments often 'grow legs,' meaning that once a person decides on a particular course of action she or he will persist on that path even if the original motivation for doing so no longer exists.

Social Proof

People often exhibit a herd mentality. When they are unsure of how to behave in social situations, they often look to others as guides for behavior. For example, women in a cafeteria line were much more likely to select a dessert if the person ahead of them did so (Guarino et al., 1994). Social proof also operates online. An initial positive or negative review by one user can bias subsequent reviews (Muchnik et al., 2013).

The more ambiguous the situation, the more likely an individual is to rely on social proof. The more credible the model, the more likely others will follow his or her lead. Following another's lead is also more likely, if the modeler perceives she or he is similar to, or identifies with, the model. Celebrities, peers, groups, or individuals may serve as models. For example, studies show that adolescent exposure to on-screen smoking in movies significantly increases the likelihood that teens will take up smoking (National Cancer Institute, 2008).

Social proof can be informational, based on a desire to be accurate or correct. A person who knew little about fishing, for example, might observe what other anglers were using for bait. Social proof can also be normative, based on a desire to be accepted or fit in. A new employee, for instance, might adapt his or her clothing style to what other employees were wearing.

Word of mouth (WOM), which plays an important role in online marketing, relies on social proof. Consumers often look to others' opinions when deciding what movie to see, where to dine, or what brand of electronics to buy. Online reviews now figure heavily in the decision-making process. A number of Web sites depend on such consumer reviews. Liking a post on Facebook, retweeting another's tweet, or sharing a photo on Pinterest are all forms of online electronic WOM. Social norms marketing, which is used in public awareness campaigns, also relies on social proof. Social network marketing (SNM) seeks to correct people's misperceptions of what is normal or typical behavior, thereby promoting more healthy or safe behavior. College students, for example, often hold misperceptions about alcohol consumption, overestimating the amount of drinking that occurs on campus. Correcting such misperceptions has been shown to reduce binge drinking (Turner et al., 2008).

Liking/Attractiveness/Similarity/Ingratiation/Immediacy

Liking is a pervasive tool for influence. All else being equal, we are more likely to comply with a request by someone we like than someone we dislike. Liking operates in the workplace, in social groups, in retail sales, in doctor–patient interactions, and with complete strangers. As with other influence strategies, liking operates primarily via peripheral processing. Liking may be established in a variety of ways. Perceived similarity is one. We tend to like people who we perceive are similar to us. We identify with people who have similar tastes and interests.

Nonverbal cues are another. Peter Andersen's Direct Effects Model of Immediacy (2004, 2007) posits that nonverbal immediacy behaviors, such as eye contact, smiling, leaning forward, standing closer, and light touch convey warmth and friendliness. Such warm, friendly behaviors facilitate influence attempts. Other research shows that nonverbal mirroring, e.g., mimicking another's nonverbal behavior, can increase rapport, liking, and perceived similarity which, in turn, enhances influence (Guéguen, 2009).

Authority

Authority figures enjoy a decided advantage when it comes to influencing others. Milgram's (1974) classic study demonstrated just how much influence authorities command. When an experimenter, dressed in a white lab coat, holding a clip board, instructed participants to administer increasingly stronger shocks to a stranger in another room, 63% of the participants complied, even though some levels of shock were labeled 'Danger: Severe Shock.' That the participants perceived the experimenter as possessing expert and legitimate authority goes a long way toward explaining the results (Blass and Schmitt, 2001).

Sources who are perceived as being highly credible are more influential than those who are not. Credibility is a perceptual phenomenon; it exists in the eye of the beholder. It is also situational. A person with high credibility in one communication context may lack credibility in another setting. Examples of influential sources abound. Celebrities hold sway over the public, which is why they often land lucrative endorsement deals. Their statements are liked, followed, shared, and retweeted by an admiring public. Politicians also exert leverage on voters. To be sure, not all politicians are revered, but elected officials have considerable clout. Economists, journalists, judges, medical experts, and scientists also wield a good deal of influence.

According to Cialdini (1984), an overreliance on authorities can lead to 'blind obedience' whereby decisions are made without scrutiny and orders are followed without question. The abuse of detainees at Abu Ghraib prison in Iraq exemplifies this tendency. Low-level guards forced detainees to perform degrading acts after being told by higher ups in the chain of command to soften up prisoners for interrogation (Hersh, 2004).

Scarcity

According to the laws of supply and demand, that which is rare is all the more valuable. That which is *perceived* as scarce is also more prized. We want what we cannot have. The scarcity principle is based on this tendency of human nature. It is why advertisers peddle goods with phrases like 'must act now,' 'limited time offer,' or 'limit one per customer' (Aguirre-Rodriguez, 2013). It is why ticket scalpers are able to sell tickets for many times their face value. Scarcity explains why people belong to exclusive groups or clubs, why art prints are sold as 'limited editions,' and why daters often play 'hard to get.' The prospect of not getting what we want, or losing what we already have, creates a sense of urgency. One theoretical explanation for scarcity's effectiveness is based on

psychological reactance (Brehm, 1966) or what is often called 'reverse psychology.' The threat of losing a freedom, having a choice usurped, or foregoing an opportunity triggers defensiveness and a sense of urgency. A quick choice or purchase alleviates the threat.

Scarcity appeals take two forms: limited quantity scarcity (LQS) and limited time scarcity (LTS). An example of the former is a study by Worchel et al. (1975) who found that consumers who were asked to rate a cookie rated it more highly, if it was only one of two left, rather than one of 10. Scarcity may also depend upon limited time. An LTS strategy establishes a limited time 'window' for acting. Either approach can trigger inferences about product quality, price, uniqueness, or desirability. Black Friday sales rely on both strategies; a limited numbers of items are on sale for a limited amount of time. Evidence suggests that LQS appeals prompt a greater sense of urgency and foster stronger purchase intentions than LTS appeals (Aggarwal et al., 2011).

Other Social Influence Strategies

In addition to the six core principles identified by Cialdini, there are many other strategies from which an influencer may choose. Anthony Pratkanis (2007) has compiled an index of 107 social influence tactics, which he groups into four broad categories. One category consists of strategies that seek to establish a favorable climate for influence, the second seeks to foster a favorable relationship with the target, the third focuses on message structure and organization, while the fourth category relies on emotion-laden appeals. Although, there is not sufficient space to discuss each strategy here, some highlights and exemplars from each of the four categories are provided.

Landscaping

Landscaping, the first of Pratkanis (2007) four categories of social influence, involves laying the groundwork for a successful influence attempt. Pratkanis and Aronson (2001) also refer to this approach as 'prepersuasion.' This group of strategies seeks to create a favorable climate for influence. The time, place, or setting is selected to make the target more receptive to an entreaty. A beau who selects a romantic setting in which to 'pop the question' is using landscaping. An employee who shows up early for work each day before asking the boss for a raise, would be engaged in landscaping. Pretrial publicity or the 'spin,' an attorney puts on a client's case is an example of legal landscaping. Some specific examples of this strategy are identified next.

Defining and Labeling an Issue

One means of influencing people is through the power of labeling. An influencer who uses labels effectively can load the dice in his or her favor. For example, referring to a political figure as a 'wonk,' 'pundit,' or 'hack' automatically casts the politician in a negative light. Similarly, assigning idealistic names to legislation, such as the 'Clear Skies Initiative,' or the 'Affordable Health Care Act' imbues the bills with positive connotations. Influencers can also seize the moral high ground via naming and labeling of issues. The term, 'pro-life,' for example, implies that opponents are 'antilife,' while the term 'pro-choice' suggests opponents are 'antichoice.'

In the interpersonal arena, the use of positive social labels ('you are so helpful,' 'I knew I could count on you') can facilitate compliance with requests (Goldman et al., 1985; Kraut, 1973). This is consistent with Becker's labeling theory, which suggests that people tend to live up to the labels assigned them, and with Bem's self-perception theory, which maintains that individuals infer who they are based on what they do. Labeling can increase the likelihood that an individual will draw the inference about him or herself that a persuader intends.

Limited Choices or Options

Another method of 'stacking the deck' in one's favor is to limit the number of options presented. A parent who wants to put a child to bed could issue a directive, "Brush your teeth, then I'll read you a story, then it's your bedtime." Such a dictatorial approach, however, might trigger defensiveness. Offering a choice, such as "Would you rather brush your teeth first, then have a bedtime story, or have a bedtime story, then brush your teeth before bedtime?" gives the child a say in the matter. The options are limited, but the child is less likely to perceive the choice as an infringement on his or her freedom.

Credibility and Social Relationships

The second category of social influence strategies identified by Pratkanis (2007) involves source factors. The reputation of a source and his or her ability to cultivate positive relationships are key to the effectiveness of influence attempts. Source factors are closely related to the twin concepts of authority and liking identified by Cialdini (1984). Two variations on source factors are presented next.

Source Credibility

Source credibility, or *ethos* as it is also known, may well be the most robust form of social influence there is. By way of illustration, a study by Guéguen and Pichot (2001) found that pedestrians were more likely to ignore a 'no crossing' light and follow a well-dressed (high-status) person through an intersection than a poorly dressed (low-status) person. Similarly, shoppers in a music store were less likely to confront or report a well-dressed person who shoplifted (10% intervened) than a shabbily dressed person (39% intervened) (Guéguen, 2003). Credibility is a perceptual phenomenon, e.g., it exists in the eye of the beholder. Credibility is also multidimensional and comprised of three primary dimensions: competence, trustworthiness, and goodwill. Sources who exhibit these dimensions, especially perceived expertise or competence, enjoy significant advantages over their less credible counterparts when it comes to social influence (Pornpitakpan, 2004). The advantages enjoyed by credible sources are accentuated when a target relies on peripheral, rather than central, processing. This is because a person who defers to a source is less likely to pay attention to message content. In contrast, when a source is perceived as lacking in expertise or being untrustworthy, the target is more likely to scrutinize the content of the message.

Plain Folks Appeal

Rather than emphasizing source expertise, an influencer may employ a 'plain folks' appeal. This strategy, also known as a 'common man' appeal, portrays the source as an ordinary person, not a snob, an elitist, or a member of a privileged class. A 'plain folks' appeal may disparage evidence in favor of 'common sense.' What this strategy gives up in the expertise dimension, however, it makes up for in the trustworthiness dimension. A target may identify or empathize more closely with an 'average Joe or Jill' based on perceived similarity. Tom Bodette, the spokesperson for Motel 6, exudes such folksiness when he recites the company motto, "We'll leave the light on for you." Politicians cast themselves into the plain folks role when they don a hardhat at a factory, roll up their sleeves at a country fair, or enjoy a beer at a local watering hole. Voters may see them as more genuine or authentic for doing so.

Message Tactics

A third category of social influence strategies identified by Pratkanis (2007) focuses on message structure and organization. Here the emphasis is on adapting the content of the message to increase its appeal. Many of these strategies, such as implicit versus explicit conclusions, one-sided versus two-sided messages, primacy versus recency, distraction, forewarning, and inoculation comprise the 'bread and butter' of traditional persuasion research. Since explanations of these more traditional strategies are readily available elsewhere, the strategies discussed next are somewhat less well known.

Self-Generated Arguments

When we think about social influence, we tend to think about one party, the source, influencing another party, the target. There is another way, however, which is to have the target convince himself or herself. Self-generated arguments are effective because the target is an active participant in the process. For example, Gregory et al. (1982) gave some residents in a neighborhood reasons for subscribing to a cable service. Other residents were asked to imagine their own uses for the cable service. The results showed that roughly one in five (19.5%) of the residents who were supplied with reasons wound up subscribing, whereas almost half (47.4%) of those who came up with their own reasons subscribed.

Self-generated arguments can be proattitudinal in nature, reinforcing a person's point of view, or counterattitudinal, such as when a person plays the role of devil's advocate. Gordijn et al. (2001) found that proattitudinal advocacy was more effective when an advocate believed his or her position reflected the majority view. Conversely, counterattitudinal advocacy was more effective when an advocate believed his or her position represented the minority view.

Several explanations for the effectiveness of self-generated arguments have been offered. One is that active participation involves central processing, which tends to produce more attitude change than peripheral processing. Another is that when people are asked to develop arguments designed to persuade others, they generate arguments that they find most convincing to themselves. A third explanation is based on cognitive dissonance. Counterattitudinal advocacy, e.g., playing devil's advocate, arouses dissonance leading to attitude/behavior change in the direction of the position advocated.

Message Discrepancy

How much should the position advocated by a source differ from that of the target? If a source's stance on an issue is identical to the target's position, no attitude or behavior change will occur. If the source's stance is radically different from the target's position, the message may be rejected outright. The challenge is to strike a balance between asking for too little change or too much. Social Judgment Theory (Sherif and Sherif, 1967) addresses this issue. Message discrepancy refers to the degree to which a source's position differs from the target's position on an issue. A target's preferred position on an issue is called her or his anchor. The range of positions surrounding a person's anchor that are perceived as tolerable represents that person's latitude of acceptance (LOA). The range of positions that depart so much from the person's anchor that they are deemed unacceptable represents the person's latitude of rejection (LOR). There may be buffer zones as well, known as latitudes of noncommitment (LNC). LNCs are positions about which a target is neutral or, perhaps, indifferent.

For maximum influence, a source should design a message to be as discrepant as possible from the target's anchor, without falling into the target's LOR. The assimilation-contrast phenomenon predicts that messages that fall inside a target's LOA or LNC will be assimilated and perceived as closer to the target's anchor position than they really are. Messages that fall within the target's LOR tend to be contrasted and perceived as farther from the target's position than they really are. By way of illustration, Siero and Jan Doosje (1993) presented Dutch drivers with messages about trade-offs between automobile use and the environment. They found that a moderately discrepant message, e.g., one that fell within the motorists' LNC, produced more attitude change than mildly or moderately discrepant messages.

Emotional Tactics

The fourth category of social influence tactics identified by Pratkanis (2007) involves emotional appeals. Emotional appeals are designed to evoke positive or negative affect. Livingston maintains that "people buy for emotional reasons, then justify their purchases with logic" (p. 16). Typically, emotional appeals function as external inducements. They supplement, or substitute, emotion for reason as a means of changing a target's attitudes or behaviors. Emotional appeals come in a wide variety: embarrassment, empathy, fear, humor, guilt, ingratiation, patriotism, pity, pride, sex, and more.

Ingratiation

Ingratiation has been practiced for millennia. Ingratiatory behaviors are designed to instill positive feelings in a target. "Hey, boss," a flatterer might say, "you look great. Have you lost weight lately?" Watt (1993) found that ingratiators were perceived by their supervisors as being more competent, more

motivated, and more qualified for leadership positions than their noningratiating counterparts. Although, people frown on the use of ingratiation – consider all of its pejorative nicknames: buttering up, brownnosing, bootlicking, sucking up – it is still a common and effective tactic for influence. Chan and Sengupta (2010) found, for instance, that ingratiation is still effective even when a target suspects the ingratiation is insincere. Ingratiation's effectiveness is based partly on liking ("I love your new haircut"), partly on perceived similarity ("I'm a Cubs fan too!"), and partly on social labeling ("You are so thoughtful").

Fear Appeals

Fear appeals are a staple of social influence. They are a fixture of public service announcements and political campaigns. For some parents, they are the bread and butter of child rearing. Some commentators suggest that we are living in a 'culture of fear' and that fear-mongers try to keep the public in a perpetually fearful state. Kim Witte's (1992) Extended Parallel Processing Model is based on a dual processing model. Danger control facilitates constructive responses, while fear control triggers maladaptive behaviors, such as panic. The most effective fear appeals combine high levels of perceived threat with high levels of perceived efficacy. Threat level hinges on the perceived probability, severity, and immediacy of a threat, along with the target's perceived susceptibility or vulnerability to the threat. Efficacy involves both perceived response efficacy, or the availability of a viable means of avoiding the threat, and perceived self-efficacy, or the practical ability of the target to take the recommended action.

Conclusion

To influence is human. Social influence permeates nearly every aspect of daily life. No single chapter can hope to cover all the ways in which people influence one another. This article has provided an overview of some of the most well known, frequently cited forms of influence.

A strength of social influence research is that much of it has been conducted in naturalistic settings. Another strength is that outcome measures have often included naturalistic measures. A third strength of social influence research is that has yielded a wealth of useful, albeit qualified, generalizations, including a number of counterintuitive findings.

An important contribution of research on social influence has been to reveal the variety of ways in which people are swayed implicitly, with little or no awareness. Some may find it disconcerting that humans are not that rational after all. We can blink, sway, and nudge one another into doing things. This realization, however, makes learning about social influence more vital than ever.

Bibliography

Aggarwal, P., Jun, S.Y., Huh, J.H., 2011. Scarcity messages: a consumer competition perspective. Journal of Advertising 40 (3), 19–30.

Aguirre-Rodriguez, A., 2013. The effect of consumer persuasion knowledge on scarcity appeal persuasiveness. Journal of Advertising 42 (4), 371–379. http://dx.doi.org/10.1080/00913367.2013.803186.

Allport, G.W., 1935. Attitudes. In: Murchoson, C. (Ed.), Handbook of Social Psychology. Clark University Press, Worcester, MA, pp. 798–844.

Andersen, P.A., 2007. Nonverbal Communication: Forms and Functions, second ed. Waveland Press, Long Grove, IL.

Andersen, P.A., 2004. Influential actions: nonverbal communication and persuasion. In: Seiter, J.S., Gass, R.H. (Eds.), Readings in Persuasion, Social Influence, and Compliance Gaining. Allyn & Bacon, Boston, pp. 165–180.

Asch, S.E., 1956. Studies of independence and conformity: a minority of one against a unanimous majority. Psychological Monographs: General and Applied 70, 1–70.

Becker, H.S., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press, New York.

Blass, T., Schmitt, C., 2001. The nature of perceived authority in the Milgram paradigm: two replications. Current Psychology 20 (15), 115–121.

Brehm, J.W., 1966. A Theory of Psychological Reactance. General Learning Press, Morristown, NJ.

Chan, E., Sengupta, J., 2010. Insincere flattery actually works: a dual attitudes perspective. Journal of Marketing Research 47, 122–133.

Chaiken, S., Trope, Y., 1999. Dual-Process Theories in Social Psychology. Guilford Press, New York.

Cialdini, R.B., 1984. Influence: Science and Practice. William Morrow and Company, New York.

Cialdini, R.B., 2009. Influence: Science and Practice, fifth ed. Pearson/Allyn & Bacon, Boston, MA.

Dahl, D., 2013. Journal of Consumer Research 40 (2), iii–v. http://dx.doi.org/10.1086/670170.

Eagly, A.H., Chaiken, S., 1993. Process theories of attitude formation and change: the elaboration likelihood and heuristic-systematic models. In: Eagly, A.H., Chaiken, S. (Eds.), The Psychology of Attitudes. Harcourt Brace, Orlando, FL, pp. 303–350.

Elliot, A.J., Niesta, D., 2008. Romantic red: red enhances men's attraction to women. Journal of Personality and Social Psychology 95, 1150–1164.

Forsyth, D.R., 2013. Social influence and group behavior. In: Weiner, I.B. (Editor in-chief), Tennen, H., Suls, J. (Eds.), Handbook of Psychology, vol. 5. Wiley, New York, pp. 305–328.

Gass, R.H., Seiter, J.S., 2014. Persuasion, Social Influence, and Compliance Gaining, fifth ed. Allyn & Bacon/Pearson, Boston.

Goldman, M., Kiyohara, O., Pfannensteil, D.A., 1985. Interpersonal touch, social labeling, and the foot-in-the-door effect. Journal of Social Psychology 125 (2), 143–147.

Gordijn, E.H., Postmes, T., de Vries, N.K., 2001. Devil's advocate or advocate of oneself: the effects of pro- and counter-attitudinal advocacy and numerical support on self persuasion. Personality and Social Psychology Bulletin 27, 395–407.

Gregory, L.W., Cialdini, R.B., Carpenter, K.M., 1982. Self-relevant scenarios as mediators of likelihood estimates and compliance: does imagining make it so? Journal of Personality and Social Psychology 43 (1), 89–99.

Guarino, M., Fridrich, P., Sitton, S., 1994. Male and female conformity in eating behavior. Psychological Reports 75, 603–609.

Guéguen, N., 2003. The effect of shoplifter's status on reporting a crime: an evaluation in a natural setting. Studia Psychologica 45 (3), 221–228.

Guéguen, N., 2009. Mimicry and seduction: an evaluation in a courtship context. Social Influence 4, 249–255.

Guéguen, N., Pichot, N., 2001. The influence of status factors on pedestrians' failure to observe a road safety rule. Journal of Social Psychology 141 (3), 413–415.

Hersh, S.M., 10 May 2004. Torture at Abu Ghraib. The New Yorker. Retrieved from: http://www.newyorker.com/archive/2004/05/10/040510fa_fact?currentPage=all.

Hovland, C.I., Janis, I.L., Kelley, H.H., 1953. Communications and Persuasion: Psychological Studies of Opinion Change. Yale University Press, New Haven, CT.

Izuma, K., 2013. The neural basis of social influence and attitude change. Current Opinion in Neurobiology 23, 456–462. http://dx.doi.org/10.1016/j.conb.2013.03.009.

Kraut, R.E., 1973. Effects of social labeling on giving to charity. Journal of Experimental Social Psychology 9 (6), 551–562.

Livingston, S., November 26, 2007. Emotional intelligence. Adweek 48 (43), 16.

Marwell, G., Schmitt, D.R., 1967. Dimensions of compliance-gaining behavior: an empirical analysis. Sociometry 30, 350–364.

Mason, W.A., Conrey, F.R., Smith, E.R., 2007. Situating social influence processes: dynamic, multidirectional flows of influence within social networks. Personality and Social Psychology Review 11, 279–300. http://dx.doi.org/10.1177/1088868307301032.

Milgram, S., 1974. Obedience to Authority. Tavistock, London.

Miller, R.L., Wozniak, W., 2001. Counter-attitudinal advocacy: effort vs. self-generation of arguments. Current Research in Social Psychology 6 (4), 46–57.

Muchnik, L., Aral, S., Taylor, S.J., 2013. Social influence bias: a randomized experiment. Science 341, 647–650.

National Cancer Institute, 2008. The Role of the Media in Promoting and Reducing Tobacco Use. US Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Bethesda, MD. Tobacco Control Monograph no. 19. NIH Pub. no. 07–6242.

North, A.C., Hargreaves, D.J., McKendrick, J., 1999. The influence of in-store music on wine selections. Journal of Applied Psychology 84 (2), 271–276.

Perloff, R.M., 2014. The Dynamics of Persuasion: Communication and Attitudes in the 21st Century, fifth ed. Routledge, New York.

Petty, R.E., Cacioppo, J.T., 1986a. The elaboration likelihood model of persuasion. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 19. Academic Press, New York, pp. 123–205.

Petty, R.E., Cacioppo, J.T., 1986b. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. Springer-Verlag, New York.

Pornpitakpan, C., 2004. The persuasiveness of source credibility: a critical review of five decades' evidence. Journal of Applied Social Psychology 34 (2), 243–281.

Pratkanis, A.R., 2007. The Science of Social Influence: Advances and Future Progress. Taylor & Francis, New York.

Pratkanis, A.R., Aronson, E., 2001. Age of Propaganda: The Everyday Use and Abuse of Persuasion. Henry Holt and Company, New York.

Sherif, M., Sherif, C.W., 1967. Attitudes as the individual's own categories: the social-judgment approach to attitude and attitude change. In: Sherif, C.W., Sherif, M. (Eds.), Attitude, Ego-involvement, and Change. John Wiley & Sons, New York, pp. 105–139.

Siero, F.W., Doosje, B., 1993. Attitude change following persuasive communication: Integrating social judgment theory and the elaboration likelihood model. European Journal of Social Psychology 23, 541–554.

Simons, H.W., Jones, Jean G., 2011. Persuasion in Society, second ed. Taylor & Francis, New York.

Smith, J.R., Louis, W.R., Schultz, P.W., 2011. Introduction: social influence in action. Group Processes and Intergroup Relations 14, 599–603. http://dx.doi.org/10.1177/1368430211410214.

Turner, J., Perkins, H.W., Bauerle, J., 2008. Declining negative consequences related to alcohol misuse among students exposed to social norms marketing intervention on a college campus. Journal of American College Health 57 (1), 85–93.

Watt, J.D., 1993. The impact of frequency of ingratiation on the performance evaluation of bank personnel. Journal of Psychology 127 (2), 171–177.

Wiseman, R.L., Schenck-Hamlin, W., 1981. A multidimensional scaling validation of an inductively-derived set of compliance-gaining strategies. Communication Monographs 48, 251–270.

Witte, K., 1992. Putting the fear back into fear appeals: the extended parallel process model. Communication Monographs 59, 329–349.

Worchel, S., Lee, J., Adewole, A., 1975. Effects of supply and demand on ratings of object value. Journal of Personality and Social Psychology 32 (5), 906–914.

Social Innovation

Alex Nicholls, University of Oxford, Oxford, UK
J Gregory Dees[†]**,** Duke University, Durham, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social innovation refers to the processes and outcomes that develop a novel approach to addressing a social problem or need. Compared to commercial innovation, it poses some distinctive challenges, particularly with regard to the incentives to invest in it, assessment of performance, and diffusion of effective innovations. Scholars began paying attention to this phenomenon in the late twentieth century, but many open research questions remain.

The term 'social innovation' has had two meanings in the academic literature. In its earliest scholarly uses, primarily in sociology, it was used to refer to the creation of new patterns of human interaction, new social structures, or new social relations. The second focuses on innovations designed to address a social or environmental issue or to meet a specific *social* market failure or need. Often, both types of innovation are combined to establish new patterns of social relations that have positive social outcomes. As Mulgan et al. (2007) argued, social innovations are "social both in their ends and their means" creating "new social relationships and collaborations" as well as responding to specific social needs (p. 2). This article provides an overview of the evolving field of social innovation globally. First, it introduces the term and sets out key definitions. Next, the chapter explores the key contexts and drivers of social innovation and maps the actors who engage with it. Following this, social innovation is located within academic theory. The next section considers some of the distinctive challenges faced by social innovation. The concluding section makes some final observations and sets out a future research agenda.

Introduction

Innovation designed specifically to address social or environmental issues is not new. However, it is only recently that research, deliberative practice, and the formulation of bespoke policy agendas focused specifically on such 'social innovation' have begun to emerge. Today, social innovation is increasingly seen as essential in developing the 'clumsy solutions' needed to tackle today's complex 'wicked problems' across the world (Mulgan et al., 2007; Murray et al., 2010; Nicholls and Murdock, 2011).

Mainstream analyses of the emergence of innovation at the country level have often identified a series of 'waves' based on technological change, usually starting with the Industrial Revolution in the eighteenth century (Kondratiev, 1998). In essence, five waves of modern macroinnovation have been discerned so far, each associated with a particular disruptive technology in the Schumpeterian sense:

- The Industrial Revolution (1771–1829)
- The Age of Steam and Railways (1829–75)
- The Age of Steel, Electricity, and Heavy Engineering (1875–1908)
- The Age of Oil, the Automobile, and Mass Production (1908–71)
- The Age of Information and Telecommunications (1971–)

These waves are usually characterized by step changes in financial markets, productivity growth, urbanization, mass production, knowledge intensity and transfer, and per capita income growth. In other words, *economic* factors are used as the primary evidence in defining such waves of innovation, with social or societal factors being subsidiary or external to the main patterns of disruptive change discerned. More recently, however, *social* innovation has been recognized as a new wave of innovation that gives primacy to systems and processes of change in social relations on the one hand and, on the other, to innovation around the conceptualization, design, and production of goods and services that address social and environmental needs and market failures. Social innovation may represent a step change compared with previous waves of innovation in that it often attempts to disrupt and reconfigure systems themselves rather than simply providing disruption within existing systems. Social innovation can be seen as the simultaneous production of new ideas and new social structures *and* a process of recontextualization of the norms of the public good, justice, and equity.

Today, social innovation cuts across all sectors of society often combining the conventionally disparate logics of private, public, and civil society in new, hybrid organizations and forms of action. For the private sector, there are two dimensions of social innovation: first, the recognition that technological innovations fail if they are not integrated with changes in social relations within the organization and second, as a new agenda for the role of business in society. For the state, social innovation reflects an established tradition of welfare reform based on increased efficiency and effectiveness. It may also challenge the governance status quo in societies by aiming to transform the power structures across social relations that allocate goods and services ineffectively or unequally. For civil society, social innovation involves both internal process of organizational change (such as new legal forms and collaborations) and novelty in external outputs and outcomes (such as new products and services).

[†] Deceased. After a long illness that he had bravely fought, Professor J. Gregory Dees died suddenly on 20 December 2013. With Greg's death, the academic field of social entrepreneurship lost its founder and most respected thinker and many of us lost a dear and kind friend. He is sorely missed but his legacy will endure forever. This article is his last completed work.

Definitions

At its simplest, social innovation represents new ideas that address unmet social needs (Mulgan et al., 2007). However, beyond this, there are two broad schools of thought concerning the definition of social innovation: new social processes and new social outcomes. The first emphasizes changes in social relations and often has a focus on rebalancing power disparities in society. For example, Mumford (2002: p. 253) noted:

> Social innovation…refers to the generation and implementation of new ideas about how people should organize interpersonal activities, or social interactions, to meet one or more common goals.

Westley and Antadze (2010: p. 2) subsequently expanded upon this.

> Social innovation is a complex process of introducing new products, processes or programs that profoundly change the basic routines, resource and authority flows, or beliefs of the social system in which the innovation occurs. Such successful social innovations have durability and broad impact.

Second, social innovation can be seen as the answer to social market failures in the provision of vital public goods. Phills et al. (2008: p. 36) proposed that social innovation is

> A novel solution to a social problem that is more effective, efficient, sustainable, or just than existing solutions and for which the value created accrues primarily to society as a whole rather than private individuals.

The Organisation for Economic Co-operation and Development (OECD) endorsed this definition but also acknowledged the process dimension of social innovation (2011: p. 1),

> Social innovation is distinct from economic innovation because it is not about introducing new types of production or exploiting new markets in itself but is about satisfying new needs not provided by the market (even if markets intervene later) or creating new, more satisfactory ways of insertion in terms of giving people a place and a role in production.

In addition to these two metadefinitions, three *types* of social innovation can be identified that cut across both (see **Table 1**). First, there is *incremental* innovation in existing goods and services to address social need more effectively or efficiently. Second, there is *institutional* innovation that aims to harness or reshape existing social and economic structures to generate new social value and outcomes. Finally, *disruptive* social innovation aims at systems change via political action and is typically driven by social movements aiming to reframe issues to the benefit of otherwise disenfranchised groups.

Social innovation can also be defined in terms of the level of its action or impact from the individual to the systems. Such levels can be mapped against the two main definitions of social innovation focused either on new processes or on new outcomes (see **Table 2**).

Finally, drawing upon theory from design thinking, Murray et al. (2010) set out the key stages of the development of a social innovation as a nonlinear process (see **Figure 1**). This model is characterized by a series of key inflection points where the development of an innovation moves first from prompts and proposals to prototyping (an important part of the design process), then to sustainability and, finally, to scale.

Table 1 Types of social innovation

Level	*Objective*	*Focus*	*Example organization (sector)*
Incremental	To address identified market failures more effectively: e.g., negative externalities and institutional voids	Products and services	Kickstart (low-cost irrigation foot pump) Aurolab (low-cost intraocular lenses) Afghan Institute of learning (female education)
Institutional	To reconfigure existing market structures and patterns to create new social value	Markets	Mpesa (mobile banking) Institute for one world health ('orphan' drugs) Cafedirect (Fair Trade)
Disruptive	To change the cognitive frames of reference around markets and issues to alter social systems and structures	Politics (social movements)	Greenpeace (environmental change) BRAC (micro-finance) Tostan (human rights)

Table 2 Levels of social impact

	Social process (example)	*Social outcome (example)*
Individual	Coproduction (Southwark Circle)	Lost-cost health care (Aravind Eye Hospital)
Organization	Wiki-production (Wikipedia)	Work integration social enterprise (Greyston Bakery)
Network/movement	Open source technology (Linux)	Nontraditional training and education (Barefoot College)
System	Microfinance (Grameen Bank)	Mobile banking (MPESA)

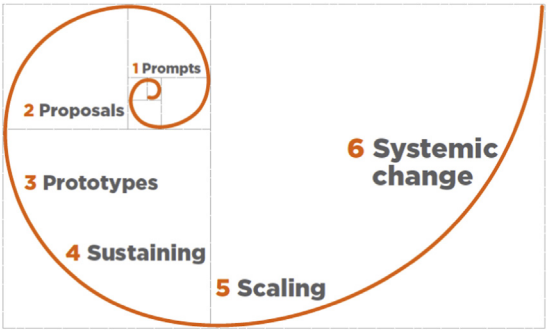

Figure 1 Stages of social innovation development (Reproduced from Murray, R., Caulier-Grice, J., Mulgan, G., 2010. The Open Book of Social Innovation. Young Foundation/NESTA, London).

Contexts

Social innovation can be seen as a response to an acceleration of global crises and the so-called wicked problems characterized by multiple, contradictory analyses and diagnoses (Rayner, 2006; Murray et al., 2010). Examples include the failure of modern state welfare systems (Leadbeater, 1997; Mulgan, 2006a), the impacts of mass urbanization (Moulaert et al., 2007), the failure of conventional market capitalism, increasingly urgent global resource constraints, the impacts of climate change, increased social breakdown, rising life expectancy and associated health and social care costs, growing cultural diversity within and across countries, growing inequality, rising incidences of chronic long-term conditions and pandemics, behavioral problems associated with the 'challenge of affluence,' difficult transitions to adulthood, and endemic reductions in individual happiness and indices of well-being (Mulgan et al., 2007: p. 9). In this context, wicked problems can be seen as highlighting the failure of conventional solutions and established paradigms across all three conventional sectors of society: private sector market failures that ignore pernicious externalities or marginalized populations, public sector thinking in departmental silos, and an absence of focus on performance outcomes in civil society.

An important subset of these 'wicked problems' concerns public sector welfare reform in response to demographic and societal changes combined with the economic realities of rising welfare costs and worsening public finances. In a different reading of the rising interest in social innovation, such action is seen as a necessary (but not always automatic) companion to rapid technological change and economic innovation (Hamalainen and Heiskala, 2007). This conceptualization links to notions of social innovation as a process of reshaping of social relations to maximize productivity and economic development, often framed by the optimistic assumption that the benefits of these changes will be shared equally. Common to both contexts is a series of complex and multifaceted issues that drive innovation not only in processes and outcomes, but also, increasingly, as boundary-blurring activity across the conventional sectors of society. As Murray et al. (2010: p. 3) noted:

> Social innovation doesn't have fixed boundaries: it happens in all sectors, public, non-profit and private. Indeed, much of the most creative action is happening at the boundaries between sectors.

Indeed, social innovation can often be the product of trying to resolve apparent contradictions and tensions across different fields of action. Taken together, the three ideal-type sectors (public, private, and civil society) can be conceptualized as a triad represented in stability as a triangle. Between each of the three ideal-type points lies a range of hybrid organizations that combine elements of the ideal types in different proportions: these represent social innovation as boundary-blurring activities (see Figure 2). Examples include social enterprises that combine business logics and models with social objectives and ownership structures (such as mutual societies and cooperatives) and public–private partnerships that provide new models of welfare provision outside of, but often in tandem with, the state (Bovaird, 2006). Various models of the 'shadow state' exist in which community organizations, as well as larger charities and nongovernmental organizations (NGOs), function as a surrogate state in terms of the functions of an elected government (e.g., the Grameen organizations in Bangladesh that provide education, telecommunications, solar power, health care, employment, and financial services across the country).

The potential of social innovation across sector boundaries has also been demonstrated in practice by the emergence of social innovation 'hubs,' incubators, and networks such as the Social Innovation Exchange that aim to bring together disparate players from across institutional fields to spark new thinking and modeling (http://socialinnovationexchange.org). The value of clustering organizations working across boundaries in a larger innovation ecosystem is also behind the notion of 'Social Silicon Valleys' (Mulgan, 2006b).

Theoretical Traditions

The topic of innovation – like entrepreneurship – has been of interest to scholars for many years. Schumpeter's work in the first half of the twentieth century provided much of the foundational thinking for subsequent innovation research,

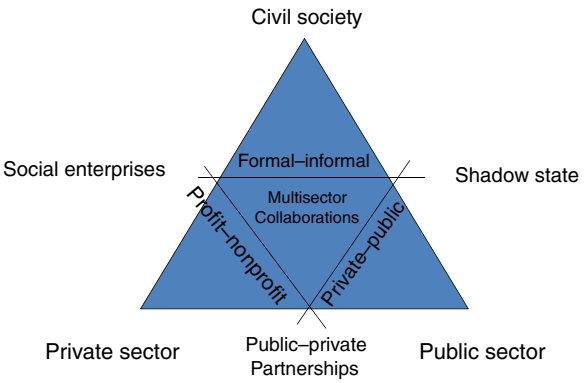

Figure 2 Social innovation: hybridity across boundaries.

particularly within an economic context. Of particular influence have been Schumpeter's (1942) reimagining of Marxian notions of 'creative destruction' caused by systemic innovation and his typology of 'five dimensions' of innovation. The latter provided a conceptual classification used in much of the subsequent analysis of innovation:

1. The introduction of a new product or an improved version of an existing product.
2. The introduction of an improved method of production.
3. The development of a new market (or entry into an existing market for a new player).
4. The development of a new source of supply or supply chain.
5. The more efficient or effective organization of any industry or sector.

Schumpeter's work was also significant in that it introduced the construct of the entrepreneur as endogenous to economic systems in contrast to conventional economic analyses that had largely cast the actor as an exogenous agent of change. Recently, Schumpeter's theory has been extended to include noneconomic forms of entrepreneurship and social entrepreneurship explicitly and broaden the categories of innovation. Even though Schumpeter's work transformed the study of innovation, it has also been problematic in that it has led to the rise of a 'heroic entrepreneur' focus of analysis that can obscure the critical operational role of collaborations, networks, and groups. (This has also been the case in social entrepreneurship research; see Nicholls (2010).)

The next significant contribution to innovation theory came from rural sociology and the analysis of the processes behind the diffusion of innovation in this context. (It is not entirely clear who first coined the term 'social innovation' to refer to novel solutions to social problems.) This research established categories of adopters of innovation – 'innovators, early adopters, early majority, late majority, and laggards' – that are still used today. Faced with a series of industrial crises, stagflation, and increasing overseas competition, a second wave of modern innovation research emerged from the 1970s onward focusing on the effects of new technologies on economic productivity.

In tandem with a range of new theorizing concerning the competitive strategy of firms, the 1980s saw the emergence of innovation studies based firmly within the management discipline and pioneered by von Hippel (1988). This work led to important subsequent contributions by Christensen (1997, 2002) and Chesbrough (2003). In the past 10 years, there has been a plethora of more generalist management books on aspects of innovation, as well as specific studies looking at a single aspect of innovation and its effects.

As part of this proliferation of work on innovation, a specific social innovation literature began to emerge, most notably with attempts by Drucker (1987) and Moss Kanter (1999) to blend commercial and social innovation theory together within a hybrid construct of 'business in society.' (Moss Kanter (1999) developed a model of corporate social innovation that suggested that cross-sector partnerships between not-for-profits and business offered valuable opportunities for innovation as a setting for 'beta testing' new ideas and processes. These pioneering ideas around 'new paradigm partnerships' led directly to subsequent work in 'corporate social entrepreneurship' (see also Collins's (2005) reflections on the similarities between his analysis of 'great' businesses and successful social sector organizations).) However, the distinctiveness of – as well as overlaps between – these two streams of work has been the subject of specific work only in the past 10 or so years. Building upon the analysis of the existing social innovation literature set out by Pol and Ville (2008), two broad conceptualizations of social innovation are identified here: innovation in social relations (that typically reflects process changes) and innovation to address social market failures (that typically reflects outcome changes). A third strand of social innovation theory, focused on macrolevel systems resilience, has also emerged from complexity theory.

Innovation in Social Relations

The largest and most well-developed body of work in this category specifically focuses on innovation that addresses various dimensions of changes in social relations. This literature can be subdivided into five categories of scholarship, as set out below:

- Research Design Challenges: Early work on social innovation developed from within behavioral science with a particular interest in devising 'social change' approaches to tackle key, contemporary social problems, often at a community level (Fairweather, 1967).
- Changes in Social Structures: Hamalainen and Heiskala (2007) argued that it is social innovation processes that ultimately determine the economic and social performance of nations, regions, and industrial sectors and organizations: "Social innovations are changes in the cultural, normative or regulative structures…of society which enhance its collective power resources and improve its economic and social performance" (2007: p. 59).
- Changes in Patterns of Work: Holt (1971) focused on social innovation within organizations, conceived of as new social patterns of employee interaction. This work was echoed in the activities of the Netherlands Centre for Social Innovation 30 years later (see Pot and Vaas, 2008).
- Diffusion of Social Change: from within sociology, there are analyses of the microlevel structures of innovation and diffusion that affect society – for example, how medical innovations spread across groups of clinicians – that have been classed as social innovation. Henderson (1993) was interested in the relationship between social innovation and political change in terms of diffusion processes. He explored how citizen movements catalyze social innovation – conceptualized as distinct from dominant cultural norms – from fluid positions outside of conventional societal structures.
- Urban Studies: There is a significant cluster of work within urban studies exploring innovative responses to social exclusion as social innovation under the heading of Integrated Area Development. Much of this work centers on innovation within social relations in urban contexts, and as a body of work, it explores the potential of public, private, and civil society models; interventions; and interactions. In Moulaert et al. (2007) characterized social innovation as

"a polymorphic constellation of counter-hegemonic movements and initiatives" (p. 196) engaged in active processes of social struggle and change.

Innovation to Address Social Market Failures

This stream of work relates to the outcome-driven model of social innovation already discussed above and focuses on innovation as the means by which new products and services can be provided to underserved market segments. At the macrolevel, this includes the mechanisms by which new markets are created in weak institutional spaces or to address market failures. The latter is conceived of as encompassing failures not only in commercial markets but also in public sector 'markets,' where the state fails to provide public goods, and civil society 'markets,' where charities, not-for-profits, and NGOs fail to provide effective goods and services to their beneficiaries. Such failures typically provide innovation opportunities, but can also provide challenges in terms of reconciling potentially competing institutional logics.

Resilience Theory

A third strand of social innovation theory has emerged recently focusing on its relationship with the overall sustainability and, particularly, the *resilience* of the larger ecosystem within which it evolves. The overall resilience of a system may be examined through the lens of the adaptive cycle, which is graphically represented by an infinity loop encompassing four phases: release, reorganization, exploitation, and conservation. The exploitation and conservation phases in the 'front' loop represent periods of growth and resource accumulation, where change is routine and almost always adaptive, while the release and reorganization phases in the 'back' loop can represent the introduction of novelty, either transformative (radical) or adaptive change, and renewal of the system. The back loop represents, therefore, a precarious moment from the point of view of whether the system remains stable, adapting, and learning but not transforming or whether it is pushed close to a threshold that tips the system into a new stability domain. The new domain may share characteristics with the old stability domain but will have radically different feedback loops, and hence different relationships between the phases (Moore and Westley, 2011). Resilience theory offers a systems-level model of the emergence and dissemination of patterns of social innovation.

Distinctive Challenges of Social Innovation

The structure of capitalism is often credited with providing a powerful incentive for the development and dissemination of commercially viable innovations. As long as markets are relatively open for innovators to sell what they believe people will be willing to pay for, there will be an incentive to innovate. With sufficient profitability, the innovators can attract the capital needed to expand, and other players have an incentive to copy or build on commercially successful innovations, contributing to their diffusion. This system has produced a powerful engine of innovation in capitalist economies.

While some social innovations may have significant commercial potential and benefit from this dynamic, many (if not most) create benefits for society that are not automatically rewarded in the marketplace (Phills et al., 2008). Direct beneficiaries may not be able or willing to pay more than the cost of the innovation, even when the social benefits would make up for such expenditure. And third-party payers may not have the resources (e.g., developing country governments), or may not even know the extent of social benefits created by a given social innovation, to judge how much to pay.

One example is education in developing countries. Innovative curricula and school designs might well improve outcomes in these countries, but without attendance by children in poor families, the improved education will not have its desired widespread social benefits. Attendance by children of the poor can be low for many reasons. For instance, with the rural poor, the parents might be just as happy with their children working on the small farm that constitutes the family's only source of income and food. One solution is paying poor families to send their children to school, tying payment amounts and timing to absentee rates, grade completion, graduation, and even enrollment in tertiary education programs. Yet, some countries can barely afford their education systems now, expecting parents to pay for textbooks, uniforms, and some school fees. The idea that social innovators working in this space can make sufficiently high profits to draw in competitors and create a vibrant market with rapid diffusion seems unlikely, no matter how effective the innovation is at improving educational outcomes. Similar dynamics exist in health care, sanitation, clean water, clean air, and so on. Social, long-term, and intangible benefits are hard to sell to individuals, especially those feeling the day-to-day pressures of poverty. Third parties, such as governments, multilateral agencies, major NGOs, and foundations, can help, but have limits and may not be a viable long-term sustainable source of revenue and profitability for a particular innovation.

Inception

Motivation does not seem to be a major problem. People innovate for many reasons. Research on innovators in business illustrates a wide range of motivations, some of which have nothing to do with making money. There are a variety of motivations for innovation, of which money is only one. It is conceivable that innovators who are driven to solve a social problem (or by other nonpecuniary motivations) will leave the economic dimensions of their innovations as an afterthought. They may be more concerned with developing something that achieves the desired social impact than with something that is economically sustainable and scalable. There are certainly many examples of social innovations that lack sustainable and scalable business models. This is a challenge when business model innovation may be what is needed to create a viable solution. The following are the key questions. Is there more of a disconnect between the innovation and finding a viable business model when the motivation of the innovator is social impact rather than making money? Are there advantages to blending these motivations?

Evaluation

In the case of commercial innovations, the market is the ultimate judge. With social innovation, an extra step is required to decide if the intended social impact is achieved and not offset by other unintended and undesirable side effects. This kind of standard is applied to some commercial products where public safety is an issue (such as pharmaceuticals), but it is crucial for social innovations since market success may not be at all correlated with the social impact of the innovation. Effective evaluation of performance impact is a challenge for social organizations: it can be costly and difficult. Moreover, there are, as yet, no commonly agreed methods or standards of evaluation for social impact, although this is changing with the growing recognition of the importance of this topic by funders. This raises a series of key questions:

- What needs to be measured to establish impact?
- Why should such an evaluation be conducted? Who should pay for it?
- For whom is the data intended? Who decides what counts as a legitimate social impact? In many cases, social innovators may have uncontroversial benefits in mind related to improving education, health status, increasing income opportunities for the poor, and so on. However, some goals are more controversial, such as women's rights, reproductive health decisions, political or ethical content in educational curricula, and more.
- Who should generate/collect the data? How participative should the data collection process be? Does it need to be done by a third party for objectivity?
- When should impact be evaluated?
- How should performance impact best be captured?

Considering each of these questions in turn may allow the most effective and appropriate methodology for a given social innovation to be discerned but will also typically raise difficult issues, not least potential contradictions across different perspectives on the nature and scale of an impact or benefit.

Diffusion and Scaling Impact

Social innovations, even those with demonstrated positive impact, have a reputation for scaling slowly, despite some clear exceptions, such as microcredit. Innovations require change, which commonly meets resistance, and without powerful financial incentives to counter the resistance or clear demand by consumers, it is not surprising that the pace of innovation and its spread will be comparatively slow. It may also be slowed by the fact that in some cases its spread depends on adoption by large bureaucratic institutions such as governments. Consider the example of controlling scurvy in the British Navy that Rogers mentions early in his classic *Diffusion of Innovations*. In this case, even with a compelling controlled trial as early as 1601 showing how lemon juice can prevent scurvy and save many sailors' lives, and another study in 1751, it took until 1795 before the British Navy adopted the innovation as policy and until 1865 before the British Board of Trade adopted the innovation. Commercial innovations can also diffuse slowly. However, at least on the commercial side, if you can demonstrate value to paying customers, and a profitable business model, there are incentives for the innovation to be spread more quickly, especially if it is not fully protected intellectual property. Copycats, clones, and adaptations are likely to arise as others see an opportunity to profit. In the absence of a profitable, self-funding business model, many social innovations struggle to attract the capital and operating funding for expansion. Consider a clinic that provides and monitors tuberculosis treatment that can sustain itself on government funding once that funding is in hand, but it receives funding only months after opening each new site, requiring philanthropic operating capital to start up each new site. The government funding is not enough to repay a loan. Funding is not the only barrier. Some social innovators also report difficulty finding and retaining the necessary talent to scale their innovations, due to the distinctive skills required (often a mix of business skills and technical expertise in the field in which they are operating) and due to limits on compensation because of limited profitability and norms against profiteering from the poor. High salaries in social ventures that get their revenue from governments, aid agencies, or from the poor themselves are frowned upon.

Innovation Ecosystems

Recent literature on innovation has recognized that successful innovations of all kinds require an ecosystem of resource providers, value chain partners, and complementary players. This poses special challenges for social innovators because they are often operating in situations (distressed environments) in which the surrounding ecosystems are weak. Supply chains and distribution networks may be unreliable and underresourced. Infrastructure may be flawed or nonexistent. Institutions that are taken for granted in well-developed markets are often absent or weak. Sometimes social innovators have to fill institutional voids. They cannot count on the profit incentive to induce others to fill the gaps. Institutional voids affect commercial entrepreneurs as well. They are more prominent in social innovation because the spaces in which social entrepreneurs frequently choose to work have weak institutional, economic, and social ecosystems, and because social entrepreneurship is a relatively recent phenomenon with distinct resource and capability needs that have not been well served by existing markets.

Altogether, social innovations face some distinctive challenges in that they are not incentivized or reinforced by the kind of market system that rewards and supports the development of profitable commercial innovations. This raises serious research questions regarding how societies can create the institutional environment to optimize the benefits of social innovation. However, there are promising developments. One example is Social Impact Bonds (SIBs), which represent an interesting example of how incentives can be aligned across disparate groups engaged in social innovation. SIBs use outcome-based payments linked to rigorous performance metrics to align the interests of external investors, government departments (or charitable foundations), and service providers around optimizing interventions that address difficult social issues where success leads to significant social impact.

Failure, and Unintended Consequences

All innovation is risky, and many are bound to fail. If societies want to promote social innovation, they have to tolerate and learn from failure. Yet, there appears to be a norm of hiding failure in the social sector. There seems to be a sense of shame at failing to deliver intended social benefits. The innovator has let down society or a set of needy beneficiaries, and perhaps a generous donor or an investor intent on achieving that social impact. To complicate matters, some failures are unknown because of lack of serious evaluations. This has caused errors to be repeated and progress to be slowed.

Moreover, it should also be acknowledged that social innovation is not, in and of itself, always a positive thing. Indeed, social innovation may have a 'dark side.' This could be evidenced in several ways, for example:

- By socially divisive or destructive objectives and intentions (e.g., secret societies or extreme political parties).
- By unintended consequences that achieve negative social effects (e.g., by excluding some groups from the focus of social goods, services, or change).
- In cases of operational failure, mission drift, or strategic co-option by an external party.

Conclusion

Today, social innovation may offer vital solutions to key global problems such as climate change, the crisis of the welfare state, health pandemics and failures, social dislocation and inequality, and educational failure. Indeed, rather than conceptualizing social innovation as a subset of technological–economic innovation, it may be the case that the reverse is now true.

Social innovation offers a compelling set of examples of hybrid logics and organizational models that offer both microlevel novelty and the potential for macrolevel systems change influence. These models provide useful templates to understand how innovation and change can be operationalized in other settings, as well as generating a range of interesting research questions concerning how to go to scale and the development of ecosystems models of social and environmental change. Indeed, policy makers, private sector leaders, and civil society actors are increasingly recognizing the need for more fluidity across the ideal-type sectors of the Triad model shown above.

However, the study of social innovation remains at an early stage of development. There are many important research topics that require further work including the measurement of innovation outcomes and impacts; the governance and accountability mechanisms for organizations in change processes or systems in flux; the challenges of the 'liability of newness' in terms of societal legitimacy, funding mechanisms, and wider resource strategies for social innovation; and the political processes around successful social innovation. Both in terms of practice and research, social innovation offers rich opportunities for future action.

See also: Business and Society: Social Accounting; Human Rights and Social Work; Innovation; Network Paradigm: Applications in Organizational Science; Organizations: Authority and Power; Social Business; Social Changes: Models; Social Justice and Social Inequalities.

Bibliography

Bovaird, T., 2006. Developing new forms of partnership with the 'market' in procurement of public services. Public Administration 84 (1), 81–102.
Chesbrough, H., 2003. Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business School Press, Boston, MA.
Christensen, C., 1997. The Innovator's Dilemma. Harvard Business School Press, Boston, MA.
Christensen, C., 2002. The rules of innovation. Technology Review 105 (5), 32–38.
Clegg, S., 2010. Flows of Globalization, Keynote Address to 26th EGOS Colloquium.
Collins, J., 2005. Good to Great and the Social Sectors: Why Business Thinking Is Not the Answer. J. Collins, Boulder, CO.
Drucker, P., 1987. Social innovation – management's new dimension. Long Range Planning 20 (6), 29–34.
Fairweather, G., 1967. Methods for Experimental Social Innovation. John Wiley, New York.
Hamalainen, T., Heiskala, R. (Eds.), 2007. Social Innovations, Institutional Change and Economic Performance: Making Sense of Structural Adjustment Processes in Industrial Sectors, Regions and Societies, SITRA. Edward Elgar.
Henderson, H., April, 1993. Social innovation and citizen movements. Futures, 322–338.
Holt, K., 1971. Social innovations in organizations. International Studies of Management and Organization 1 (3), 235–252.
Kondratiev, N., 1998. The works of Nikolai D. Kondratiev. In: Samuels, W., Makasheva, N., Wilson, S. (Eds.), Pickering Masters. Pickering & Chatto, London.
Leadbeater, C., 1997. The Rise of the Social Entrepreneur. Demos, London.
Moore, M., Westley, F., 2011. Surmountable chasms: networks and social innovation for resilient systems. Ecology and Society 16 (1), 5 [online].
Moss Kanter, R., May–June, 1999. From spare change to real change: the social sector as beta site for business innovation. Harvard Business Review, 122–132.
Moulaert, F., Martinelli, F., Gonzalez, S., Swyngedouw, E., 2007. Introduction: social innovation and governance in European cities. European Urban and Regional Studies 14 (3), 195–209.
Mulgan, G., 2006a. The process of social innovation. Innovations: Technology, Governance, Globalization 1 (2), 145–162.
Mulgan, G., 2006b. Social Silicon Valleys: A Manifesto for Social Innovation. Young Foundation, London.
Mulgan, G., Tucker, S., Ali, R., Sanders, B., 2007. Social Innovation: What It is, Why It Matters, and How It Can Be Accelerated. Skoll Centre for Social Entrepreneurship Working Paper, Oxford.
Mumford, M., 2002. Social innovation: ten cases from Benjamin Franklin. Creativity Research Journal 14 (2), 253–256.
Murray, R., Caulier-Grice, J., Mulgan, G., 2010. The Open Book of Social Innovation. Young Foundation/NESTA, London.
Nicholls, A., 2010. The legitimacy of social entrepreneurship: reflexive isomorphism in a pre-paradigmatic field. Entrepreneurship Theory and Practice 34 (4), 611–633.
Nicholls, A., Murdock, A. (Eds.), 2011. Social Innovation: Blurring Boundaries to Reconfigure Markets. Palgrave MacMillan, London.
OECD, 2011. LEED Forum on Social Innovations. Available at: http://www.oecd.org/document.
Phills, J., Deiglmeier, K., Miller, D., 2008. Rediscovering social innovation. Stanford Social Innovation Review (Fall), 34–43.
Pol, E., Ville, S., 2008. Social innovation: buzz word or enduring term. The Journal of Socio-Economics 38, 878–885.
Pot, F., Vaas, F., 2008. Social innovation, the new challenge for Europe. International Journal of Productivity and Performance Management 57 (6), 468–473.
Rayner, S., 2006. Wicked Problems: Clumsy Solutions – Diagnoses and Prescriptions for Environmental Ills. Jack Beale Memorial Lecture on Global Environment. Available at: http://www.sbs.ox.ac.uk/research/Documents.
Schumpeter, J., 1942. Capitalism, Socialism and Democracy. Harper and Row, Cambridge, MA; New York.
von Hippel, E., 1988. The Sources of Innovation. Oxford University Press, New York.
Westley, F., Antadze, N., 2010. Making a difference: strategies for scaling social innovation for greater impact. The Innovation Journal: The Public Sector Innovation Journal 15 (2), 1–19.

Social Insurance: Legal Aspects

Gerhard Igl, Christian Albrechts University, Kiel, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

To reach the goals of social security, the technique of social insurance has been one of the most important tools in the evolution of welfare states. Social insurance, in the sense of Bismarck's social insurance, has undergone major changes. The initial role of social insurance, the social protection of workers, has lost its importance as the personal scope of social insurance has widened. Other forms of social protection, especially state-financed social programs for families, unemployed people, or persons in need, are increasingly prevalent.

Insurance is considered a way of protecting against the consequences of negative future events (risks or contingencies) in a collective manner with a special technique of financing. Social insurance is a kind of insurance organized by public authorities on the basis of legislation, and it is usually compulsory. Its goals include coverage against the consequences of risks like sickness, industrial accident, invalidity, old age, and unemployment for a large part of the population of a nation. To attain these goals, a degree of compulsion is recognized to be necessary. One of the main criteria for distinguishing between social and private insurance is the way of calculating and financing the risk. In the field of social insurance, the risk ceases to be the fundamental criterion since social aims prevail over the economic criteria governing private insurance. Social insurance effects income redistribution, usually described as redistribution between healthy and sick, richer and poorer, younger and older people, or households without and with family members. In other words, social insurance creates a bond of solidarity beyond a mere solidarity of sharing individual risks. The prototype of social insurance was Bismarck's social insurance in Germany (with passage of the Health Insurance Act of 1883). This technique of social protection became widespread and is still considered one of the most important forms of social protection. Because the financial basis of social insurance schemes is earnings related, this technique of social protection will have to change in countries where the role of earnings from work is diminishing.

Social Insurance and Private Insurance

Insurance is a way of providing for the consequences of negative future events (risks or contingencies) in a collective manner with a special technique of financing. Besides the different ways of calculating and financing the risks, social insurance is organized by public authorities through legislation, and is usually compulsory, while private insurance is offered by private enterprises and is usually voluntary. Social insurance seeks to provide coverage against the consequences of risks for a large part of the population of a nation. To attain these goals, a compulsory form of insurance is considered a necessity. Bismarck's social insurance in Germany is in some ways the prototype of social insurance. However, his social insurance covered only workers rather than the general population (Koehler and Zacher, 1982; Peters, 1993).

The differences between social and private insurance can be found in their goals, their techniques of protection, the legal form of their insurance bodies as private or public organizations, and their way of financing the benefits. But there is no clear distinction between them. Sometimes, the goals of social insurance are attained by compulsory private insurance and, in some countries, the bodies responsible for social insurance are private bodies.

In the field of private insurance, the basis of calculation of the premiums is the risk. In the field of social insurance, the risk is less important, and social aims (such as health and welfare) prevail against the economic criteria.

Private insurance, moreover, distributes the burden of an individual risk across a group. Social insurance has the same effects, but it also effects income redistribution between those who are healthy and sick, richer and poorer, younger and older people, or households without and with family members.

Two Characteristic Items of Social Insurance: Financing and Organization

Thus, two characteristic items of social insurance remain, which allow us to distinguish between social insurance and other techniques of social protection. Those interlinked items are the ways of financing social insurance and the organization of social insurance.

Financing

Social insurance benefits were originally financed by contributions levied on earnings and divided between employee and employer, thus expressing the shared responsibility between employer and employee with regard to work-related social risks. State subsidies completed the financing. However, this scheme changed over time. Some countries introduced increasing state subsidies, while others changed the proportion of the contributions burden payable by employer and employee. In Germany, for example, the employees had to relinquish a public holiday to compensate employers for the new financial burden created by the introduction of long-term care insurance. Currently, there are ongoing discussions about changing the financial basis from the earnings of employees to the surplus value of enterprises.

Organization

Originally, social insurance bodies were special bodies that were not directly integrated in the administrative organization of a state. Sometimes, they are coadministered by representatives of employers and employees. This kind of self-government is more prevalent in countries where social insurance bodies are private organizations. In countries with special public authorities for social insurance, but without self-governing or autonomous elements of administration, the ideal image of social insurance is left behind.

Social Insurance and Social Security

The concept of social insurance should not be confused with the concept of social security (cf. UN Declaration of Human Rights, Art. 22), which is an international goal of social policy. Social insurance is considered a special, albeit widespread technique of social protection.

From the perspective of international law, social insurance does not entail special legal problems. Nevertheless, international law scholars have taken a special interest in social insurance as a mechanism of social protection. Thus, many international treaties and other international or regional instruments, especially those involving Europe, address social insurance.

In the field of social security, international law has two main dimensions: (1) coordinating different national social security schemes to guarantee freedom of movement, and (2) defining social standards and aims (e.g., minimum standards, standards to attain, and principles).

The coordination task was originally primarily due to the migration of workers, while the standard defining task began with the Treaty of Versailles, which can be considered the basis of the development of an international social policy (Zacher, 1976).

Fulfillment of International Requirements (International Standards of Social Insurance)

Two kinds of requirements prevail internationally: attainment of minimum standards and achievement of higher levels than a given minimum.

Minimum Standards

The best-known and most basic international instrument to set minimum standards of social security is Convention 102 of the International Labor Organization (ILO) (1952). This convention, ratified by 47 countries, contains minimum standards regarding medical care and benefits in the case of sickness, unemployment, old age, employment injury, maternity, and invalidity, and for survivors. This convention applies to insurance schemes as well as other schemes of protection.

Higher Levels of Social Standards

The European Social Charter (1961) and the Revised European Social Charter (1991) of the Council of Europe, a regional instrument for the European countries, serve as a social standard instrument for the European Union nations. In the frame of the Council of Europe, other social security conventions (e.g., the Code of Social Security) also aim at higher social standards (Becker et al., 2010; Pieters, 1999a).

Social Insurance Benefits

Range of Benefits

The range of risks covered by social insurance schemes, labeled as 'traditional' social risks or contingencies, is described in ILO Convention 102 on minimal standards of social security (see section Minimum Standards). As new social risks develop, it is not clear whether protection will be provided by social insurance schemes or by other means (van Langendonck, 1997; Pieters, 1998).

Social insurance bodies do not provide benefits unique to social insurance. Other bodies of social protection and even private insurance bodies provide comparable or identical benefits. However, there are typical links between social insurance and the benefit provided, especially for those cash benefits designed to replace former earnings (e.g., sickness benefits or old-age pensions). As the insured person pays contributions based on his or her earnings (usually a certain percentage of the income), the benefit replacing the earnings is calculated as a function of the former earnings and is not provided as a flat-rate sum.

Whereas private insurance is almost always paid in cash, social insurance benefits can be either in kind or in cash.

Guarantee of Benefits

In some countries, entitlement and even qualifying periods for social insurance entitlement are protected by the constitution. In the German constitutional doctrine, contributions to the old-age pension scheme are protected by the constitution as property.

Public–Private Mix of Benefits

Initially, social insurance benefits were the only way of providing a subsistence base in the case of risk occurrence. Nowadays, social insurance benefits are increasingly complemented by other means of protection. Mainly in the field of old-age pensions, the interplay of public, occupational, and private protection increasingly serves as a current mode of protection (International Bank for Reconstruction and Development, 1994).

Future Developments in the Field of Social Insurance

To reach the goals of social security, the technique of social insurance was one of the most important tools in the evolution of welfare states. Social insurance, in the sense of Bismarck's social insurance, has undergone major changes. The initial role of social insurance, the social protection of workers, has lost its importance as the personal scope of social insurance has widened. Other forms of social protection, especially state-financed social programs for families, unemployed people, or persons in need, are increasingly prevalent.

Today, there is no unique mode of using social insurance to address social problems. Social insurance schemes rely mainly on financing by contributions and are thus, as regards their financing basis, dependent on the evolution of earnings. This reliance is considered a critical factor in the future financial stability of social insurance schemes and is broadly under discussion (Pieters, 1999b, 2003; van Langendonck, 1997; Berghman, 2002; Brooks and Assaf, 2005).

See also: Health Insurance: Economic and Risk Aspects; Insurance and the Law.

Bibliography

Becker, U., Pieters, D., Ross, F., Schoukens, P. (Eds.), 2010. Security: A General Principle of Social Security Law in Europe. European Law Publishing, Groningen

Berghman, J., 2002. Social Security in Transition. Kluwer Law International, The Hague.

Brooks, R., Assaf, R., 2005. Social Security Reform: Financial and Political Issues. Cambridge University Press, Cambridge.

International Bank for Reconstruction and Development, 1994. Averting the Old Age Crises: Policies to Protect the Old and Promote Growth. Oxford University Press, New York.

Koehler, P.A., Zacher, H.F. (Eds.), 1982. The Evolution of Social Insurance 1881–1981. Studies of Germany, France, Great Britain, Austria and Switzerland, Pinter, London.

Pieters, D., 1993. Introduction into the Basic Principles of Social Security. Kluwer, Deventer, the Netherlands.

Pieters, D. (Ed.), 1998. Social Protection of the Next Generation in Europe. Kluwer Law International, London.

Pieters, D. (Ed.), 1999a. International Impact upon Social Security. Kluwer Law International, London.

Pieters, D. (Ed.), 1999b. Changing Work Patterns and Social Security. Kluwer Law International, London.

Pieters, D. (Ed.), 2003. European Social Security and Global Policies. Kluwer Law International, London, The Hague and New York.

van Langendonck, J. (Ed.), 1997. The New Social Risks. Kluwer Law International, London.

Zacher, H.F., 1976. Internationales und Europäisches Sozialrecht. Verlag R. S. Schulz, Percha/Kempenhausen am Starnberger See, Germany.

Relevant Websites

www.coe.int/T/DGHL/Monitoring/SocialCharter/ – Council of Europe – European Social Charter.

www.ilo.org/ilolex/cgi-lex/convde.pl?C102 – ILO – Social Security (Minimum Standards) Convention, 1952.

www.issa.int – International Social Security Association (ISSA).

www.missoc.org – Mutual Information System on Social Protection/Social Security (MISSOC).

Social Integration, Social Networks, and Health

Julianne Holt-Lunstad, Brigham Young University, Provo, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L.F. Berkman, volume 21, pp. 14327–14332, © 2001, Elsevier Ltd.

Abstract

Social relationships have long been linked to health outcomes. To understand this association various disciplines have approached the empirical understanding of this association from different theoretical and methodological approaches. This article will summarize the historical background of these approaches, the evidence linking social networks to mortality, the theoretical models that explain how this might occur, review the behavioral, psychological, and biological pathways by which social networks influence health, and conclude by providing recommendations on how these may be applied to future research and interventions.

There is now substantial evidence that being socially connected to others not only is relevant to mental health and emotional well-being, but also has a powerful influence on physical health and longevity. Research evidence indicates that having fewer and poorer social relationships is associated with poorer physical health and greater risk for early mortality, while having more and better relationships is associated with better physical health and greater odds of survival. The influence of social relationships on physical health has been examined across a number of different disciplines including psychology, medicine, epidemiology, sociology, anthropology, and public health, from which diverse conceptualizations and measurement approaches have emerged. From this interdisciplinary perspective, multiple pathways have been identified by which social relationships may influence physical health outcomes.

The aims of this article are to (1) review the historical and theoretical perspectives that guide research across diverse disciplines; (2) define the multiple ways in which social relationships have been systematically studied; (3) review the findings that link social relationships to mortality; and (4) provide an integrative guide for future interdisciplinary research and interventions in this field.

Historical and Theoretical Conceptualizations

Across history multiple disciplines have proposed theories that have provided the foundation for empirical investigations of the association between social relationships and health. These include writings from the sociologist Emile Durkheim (1897), psychiatrist John Bowlby, anthropologists Elizabeth Bott and Clyde Mitchell, epidemiologist John Cassel (1976), and psychiatrist Sidney Cobb (1976). These writings give us insight about social integration, social networks, and social support. Each of these is a concept that describes the structure, functions, and processes of social relationships.

One of the earliest historical contributions to the understanding of social relationships and health comes from the renowned French sociologist and philosopher Emile Durkheim (1897/translation 1951). He examined suicide rates across social classes, cultures, religious affiliation, and gender, highlighting the importance of social environment and integration. More specifically, he identified different types of suicide: egoistic, altruistic, anomic, and fatalistic suicide. Egoistic suicide is a result of excessive individualism and detachment from society. Lack of connection to society is thought to be a lack of connection to life, as being a part of society gives meaning in life. Thus, low social integration with society results in low value on human existence and is associated with more egoistic suicide, and high social integration with society results in high value for human existence and is associated with low egoistic suicide.

In 1969, John Bowlby introduced the concept of attachment theory which proposes that it is a universal biological need to connect socially or create social bonds (Bowlby, 1969). Bowlby's theory was influenced by the work of the Nobel Prize winner Konrad Lorenz (1935) who showed that attachment or imprinting (among ducklings) was an innate tendency that increased the odds of survival. Bowlby similarly argued that among humans there is an innate tendency to form a secure attachment to a primary attachment figure (usually mother). This attachment allows for a secure base from which an infant or a toddler can explore; however, threats to proximity to the attachment figure would illicit fear and insecurity. Proximity-seeking behaviors such as crying or smiling are thought to represent evolutionarily adaptive behaviors, as those infants that could maintain proximity to their mother would be more likely to survive and produce offspring of their own. Thus, Bowlby suggests that fear and insecurity around strangers is an innate survival mechanism. Further, Bowlby contends that the reliability and security of the primary attachment figure provides the basis for forming loving, trusting, secure relationships in adulthood. This innate tendency to form attachments to others is thought to be adaptive for survival.

Social Network Theory

The early 1950s gave rise to the consideration of the social network. The term 'social network' has been attributed to sociologist Barnes (1954) as well as the anthropologists Bott (1957) and Mitchell (1969), from whom came some of the first fieldwork done on social networks. Social network theory describes the individual as a node and the relationship to others within the network as ties. A social network then can be illustrated as a network of ties connected by individuals or nodes. This perspective is inherently relational. Social network theory has been applied across many disciplines to examine the

relationships among individuals, family, groups, and even society. According to the social network perspective, there is less importance put on the characteristics of the individual, but rather it is the relationship of ties within the network that is important. Networks are thought to have an influencing pressure on individuals' attitudes and behaviors, which in turn may influence health, and social network analyses are a systematic way of detecting the magnitude of these pressures. Although, epidemiology has used the term 'social network,' it rarely uses social network analysis, the systematic process of mapping the interconnections among people. Social networks have been examined not only from the perspective of epidemiological studies, but also to model health education, disease prevention, spread, and treatment, as well as health-care structure and coordination in an effort to understand the association between social networks and health.

Social Support

In 1976, two influential articles were published that emphasized the health relevance of the qualitative aspects of social networks that gave rise to an emphasis in health psychology on the functional support that social networks may provide. Cobb (1976) defined social support as information from others that one is cared for, loved, esteemed, and part of a mutually supportive network. In his article, he presented evidence suggesting that these social support resources were important in dealing with stressful life events such as pregnancy, hospitalization, and bereavement. Cassel (1976) focused more on the biological processes linking support to health. He argued that social support might best be seen as a protective factor and reviewed studies suggesting that such a relationship factor may modify bodily processes (e.g., blood pressure, endocrine activity) in health-relevant ways. These two reviews highlighted the important role that social support might play on physical health outcomes.

Conceptual Models Linking Social Networks to Health

Since about the 1970s there has been more systematic interest in the health influence of social networks and the social support they provide. In particular, these studies seemed to make competing predictions. Some argued that social relationships were primarily protective under stressful conditions by providing resources (informational, emotional, tangible) that buffer the deleterious effects of acute and chronic stress (stress buffering model). Others contend that social relationships may be associated with protective health effects through more direct means, such as cognitive, emotional, behavioral, and biological influences, that are irrespective of levels of stress (main effects model). Cohen and Wills (1985) reviewed the studies conducted to that point in an effort to test these seemingly competing hypotheses. They concluded that study's characteristics determined when each model was most predictive. Studies that measured the functional aspects of social support supported the buffering model, while studies that measured social integration supported the main effects model. However, subsequent articles suggested that functional measures may also show a direct effect (Uchino, 2004). Thus, functional support may have both stress-related and a direct effect on health outcomes but this may occur via different pathways. As such, the main effects and buffering models are not necessarily competing perspectives but highlight different contexts and processes by which social support operates (Uchino, 2004). Where they converge is in the more 'downstream' processes, such as health behaviors, that ultimately link them to health.

Stress-Related Models

The stress buffering model focused on the possibility that support from social network ties might decrease exposure to negative life events (Gore, 1981). This model suggests that social support is beneficial because network members may provide us with the resources to avoid or reduce exposure to stressors. There are multiple pathways by which social support may reduce stress exposure. First, social support may intervene between stress exposure and stress appraisal to prevent the appraisal of stress (Cohen and Wills, 1985). Second, social support (e.g., informational support) can help individuals make informed decisions that minimize their subsequent stress exposure via proactive coping (Aspinwall and Taylor, 1997). Finally, adequate social support may help decrease exposure to 'secondary stressors' (Pearlin, 1989). For instance, stress at work may, in turn, increase the likelihood of conflict at home. However, if spousal support reduces the stress of work, it may effectively eliminate potential spillover into marital interactions and reduce exposure to such secondary stressors.

The most well-known stress-related model is the buffering model of support (Barrera, 2000; Cohen and Wills, 1985). This model differs from the stress prevention model in that individuals might still experience stress, but social support is predicted to decrease the strength of the association between stress and physical health outcomes (Cohen and Herbert, 1996). This model is based on the possibility that stressors have adverse influence on health behaviors and physical health outcomes but that appropriate support can be an important resource in helping minimize such links. Cohen (1988) has postulated that this stress buffering occurs through a cognitive appraisal process that can, in turn, reduce the normally robust association between stress and health-related outcomes.

A major variant of the stress buffering model is what has been called the matching hypothesis (Cutrona and Russell, 1990). The matching hypothesis predicts that stress buffering is most effective when the type of support matches the needs or challenges of the stressful event. More specifically, it predicts that informational and tangible support should be most effective for controllable events (e.g., preparing for a job interview), whereas emotional and belonging support should be most effective for uncontrollable events (e.g., job layoff; Cutrona and Russell, 1990). This is one of the few theoretical models that highlights how different functional components of support might be related to outcomes based on the stressor context.

Direct Effect Model

The direct effect model postulates that social support is effective more generally regardless of stress levels (Cohen and Wills, 1985). The article by Cohen and Wills found that structural measures of social relationships were more likely to demonstrate direct effects. The direct effect of structural measures was thought to represent its influence on social identity through

direct or indirect social control processes (Thoits, 1983; Umberson, 1987). Direct social control may be seen in pressure to behave more healthily (e.g., get more sleep, eat better, exercise more, see a doctor) from others. Indirect social control can be evidenced by individual engaging in healthier behaviors and avoiding riskier behaviors because relationships provide greater meaning in life. However, there is now evidence that functional support can also have direct effects on outcomes by promoting a sense of connection, self-esteem, and control over life due to knowing that you are cared for and supported by others (Lakey and Orehek, 2011; Thoits, 2011; Uchino, 2004).

Recently, Lakey and Orehek (2011) have proposed relational regulation theory (RRT) to account for direct effects of social support on mental health outcomes. According to RRT, everyday interactions (e.g., talking about events of a typical day, gossip, sports talk) regulate an individual on a daily basis, which may result in positive outcomes like general comfort with that person and a sense of well-being (Lakey and Orehek, 2011). These daily interactions can also serve as the basis for stress buffering or prevention links as individuals generalize their relationship representations to stressful context (i.e., is this a person whom I am comfortable with likely to be supportive during stress?).

Assessment of Social Networks

Social networks, social integration, and social support are often used interchangeably but conceptually represent different approaches to understanding the influence of social relationships on health and longevity. Given the diverse ways in which social relationships may be conceptualized it is not surprising that their influence has been measured in similarly diverse ways. Three major components of social relationships are consistently evaluated: (1) the degree of integration in social networks, (2) the social interactions that are intended to be supportive (i.e., received social support), and (3) the beliefs and perceptions of support availability held by the individual (i.e., perceived social support). The first subconstruct represents the structural aspects of social relationships, and the latter two represent the functional aspects. Notably, these different subconstructs are only moderately intercorrelated, typically ranging between $r = 0.20$ and 0.30. While each of these conceptualizations has been significantly associated with morbidity and mortality, it is thought that each may influence health in different ways. Epidemiological studies that have linked social relationships to mortality can be broadly broken down further into the measurement approaches described in **Table 1**.

Mortality Data on Social Networks, Integration, and Support

There is now substantial evidence for the protective effect of social relationships on risk for mortality from all causes. Some of the first epidemiological evidence was highlighted in the article by House et al. (1988). They summarized data from five prospective studies, including the Alameda County (Berkman and Syme, 1979), Evans County, Tecumseh, Eastern Finland, and Gothenburg studies. Each of these studies was designed to examine large community samples of initially healthy individuals. They assessed their involvement with social relationships, and then followed them over several years to determine whether the level of social integration predicted who would still be alive at the follow-up. Indeed, evidence indicated that those less socially integrated had significantly higher age-adjusted mortality risk relative to those more socially integrated.

Since then, the number of studies examining the effect of social connections (see further description below) on mortality has grown exponentially. In a meta-analysis of 148 independent prospective studies, with data from 308 849 individuals, followed for an average of 7.5 years, results indicate that individuals with greater social connections have a 50% increased

Table 1 Measurement approaches used to assess social relationships

Type of measure		*Description*
Functional		*Functions provided or perceived to be available by social relationships*
	Received support	Self-reported receipt of emotional, informational, tangible, or belonging support
	Perceptions of social support	Perception of availability of emotional, informational, tangible, or belonging support if needed
	Perception of loneliness	Feelings of isolation, disconnectedness, and not belonging
Structural		*The existence and interconnections among differing social ties and roles*
	Marital status	Married vs other
	Social networks	Network density or size, number of social contacts
	Social integration	Participation in a broad range of social relationships; including active engagement in a variety of social activities or relationships, and a sense of communality and identification with one's social roles
	Complex measures of social integration	A single measure that assesses multiple components of social integration such as marital status, network size, and network participation
	Living alone	Living alone vs living with others
	Social isolation	Pervasive lack of social contact or communication, participation in social activities, or confidant
Combined		*Assessment of both structural and functional measures*
	Multifaceted measurement	Multiple measures obtained that assess more than one of the above conceptualizations

odds of survival compared to those who were less socially connected (Holt-Lunstad et al., 2010). The effect was consistent across gender, age, initial health status, and causes of mortality. Importantly, controlling for initial health status helps clarify the direction of the association. Most of the studies were population-based and thus examined initially healthy participants. However, even among those who were ill, greater social connectedness was associated with greater odds of survival. Thus, regardless of initial health status, those who were more socially connected lived longer. Most notably, the overall magnitude of the effect of social connections on risk for mortality was comparable with and in many cases exceeds the effect of many well-established risk factors for mortality (**Figure 1**). For instance, lacking social connectedness carries a risk comparable to smoking up to 15 cigarettes per day, and is greater than alcohol abuse, physical inactivity (sedentary lifestyle), obesity, and air pollution, among others.

The magnitude of risk reduction differed significantly across the type of measurement approach examined. While both structural (Odds Ratio, OR = 1.57) and functional measures (OR = 1.46) were significantly associated with greater odds of survival, living alone (OR = 1.19) and received support (OR = 1.21) were poor predictors of survival. Notably, studies that utilized complex measures of social integration were the best predictors (OR = 1.91). Thus, assessments that took into account the multidimensional aspects of social relationships were associated with a 91% increased odds of survival. Such complex measures of social integration may also be stronger predictors because they are more stable over time than perceptions of support.

There is also reason to believe that the overall effect of social connections on mortality may be a conservative estimate. Most studies made adjustments for initial health and other lifestyle risk factors to establish the independent influence of social connections on risk for mortality. However, social relationships are inextricably linked with multiple pathways associated with mortality. For example, excessive alcohol consumption is associated with poor social support (Peirce et al., 1996; Steptoe et al., 1996); moreover, individuals with positive social relationships engage in less risky health behaviors (Jessor et al., 1998) This includes smoking cessation (Mermelstein et al., 1986), seat belt use, adequate sleep (Brummett et al., 2006), healthy diet (Fuemmeler et al., 2006), regular exercise (Okun et al., 2003), dental hygiene, etc. Therefore, these analyses eliminate one of the pathways by which social relationships may influence risk for mortality. The overall effect may also represent a conservative estimate because most studies failed to measure relationship quality. Thus, estimates likely include both high- and poor-quality relationships. Given that negativity in relationships has been associated with an increased risk for early mortality (Friedman et al., 1995; Tucker et al., 1996), this would suggest that quality does matter. Failure to account for negative or detrimental social relationships within the broad measures utilized across studies may have reduced the overall effect and the effect of positive social relationships on mortality may actually be much larger.

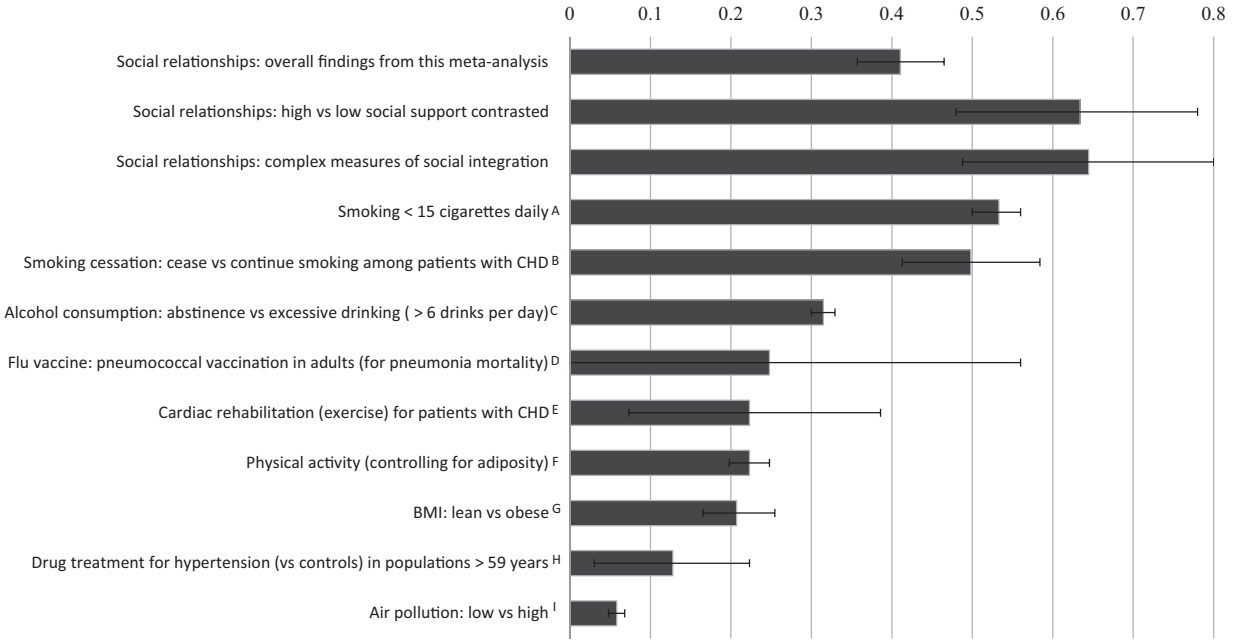

Figure 1 Comparison of odds (lnOR) of decreased mortality across several conditions associated with mortality. Note: Effect size of zero indicates no effect. The effect sizes were estimated from meta analyses; A = Shavelle et al., 2008; B = Critchley and Capewell, 2003; C = Holman et al., 1996; D = Fine et al., 1994; E = Taylor et al., 2004; F,G = Katzmarzyk et al., 2003; H = Insua et al., 1994; I = Schwartz, 1994. CHD, coronary heart disease; BMI, body mass index. Source: Holt-Lunstad J., Smith T.B., Layton J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Med 7 (7), e1000316. doi: 10.1371/journal.pmed.1000316.

Pathways by which Social Relationships May Influence Health

Social relationships can have either direct effects on physiology (e.g., cardiovascular, neuroendocrine, immune functioning) or indirect effects through lifestyle or behavioral effects.

Social Support

Social networks may have a significant influence on health via the provision of various forms of support. Types of support have been categorized into subtypes including informational, emotional, and tangible or instrumental support (Schaefer et al., 1981). Informational support refers to the information or advice that is provided by network members. This may be directly relevant to one's health such as giving recommendations on lifestyle, behavioral, or medical choices; however, it may also be indirectly related to health. For instance, advice or guidance relevant to major life decisions or how to handle a stressful situation may help buffer the subsequence stress response. Emotional support is the love, understanding, and compassion shown by network members. Feeling cared for and validated can influence not only one's appraisal of a potentially stressful situation but also a sense of meaning and purpose in life that, in turn, can influence self-care and less risk taking. Because emotional support is more intimate it is often provided by closer relationships, although other relationships may provide emotional support in circumscribed situations. Instrumental, or often referred to as tangible, support is the help, aid, favor, and direct assistance that others provide. This can include a ride to the doctor, getting groceries, cooking meals, providing money, or caring for children.

Social Influence

Social norms can also provide a powerful influence on health. Although some norms may vary across cultures or social groups, there are often very powerful norms on the acceptability of certain behaviors and violations of these norms are enforced socially via social disapproval and rejection. For instance, many unhealthy or risky behaviors are associated with social disapproval (e.g., smoking, substance abuse, risky sexual practices) while many healthy behaviors are socially approved of (e.g., being physically active, maintaining a healthy weight) and may be associated with reward or praise to reinforce such behaviors. It is important to acknowledge, however, that one's social network can have a detrimental influence as well. For instance, there is some evidence that when a social network is filled with members exhibiting negative health behaviors then there may be social contagion to conform to these unhealthy social norms (e.g., smoking, obesity; Christakis and Fowler, 2013).

Social Participation and Engagement

Another important way in which social networks may influence health is via involvement in that network. Social participation includes activities such as getting together with friends, volunteering, church attendance, belonging to groups, involvement at a community level, and participating in sports, among others. It is through social engagement that social roles, norms, and capital are established and reinforced. Involvement in social groups and activities leads to the establishment of social roles that can provide a sense of meaning and purpose in life. Such engagement may also lead to obligation to others and provision of support, such that an individual becomes the source on whom others can depend on. It is perhaps the interdependence that leads to attachment to one's social group or community, first described by Durkheim, which influences mortality.

Person-to-Person Contact

Importantly, it must be acknowledged that having greater social contact can influence exposure and transmission of infectious diseases. Social network analysis has been utilized to track the transmission of such infectious disease and finds that it is not random but rather can move within social networks. However, it is remarkable that both epidemiological and experimental studies have found that a greater diversity of social relationships is actually associated with better immune functioning (Cohen, 2003).

Biological Pathways

Although social relationships are linked to psychological processes (e.g., stress and depression) that influence health outcomes in their own right (Rozanski et al., 1999); the influence of social relationships on long-term health cannot be completely explained by these processes. Most of the epidemiological evidence statistically controlled for these factors and found social relationships to exert an independent effect on risk for mortality. Reviews of such findings suggest that there are multiple biologic pathways involved (physiologic regulatory mechanisms, themselves intertwined) that, in turn, influence a number of disease endpoints (Uchino, 2006; Uchino et al., 2007; Uchino et al., 1999). The structure and functions of social networks have been linked to health-relevant physiological processes including cardiovascular, neuroendocrine, and immune processes (see reviews: Uchino, 2004; Uchino et al., 1996).

Conclusion

We have now established a robust association between social relationships and mortality. The future generation of research needs to consider both macro- and micro-level processes, and thus a multilevel interdisciplinary perspective is needed. From a macro-level perspective we need to consider cultural norms of relationships, marriage and cohabitation, familial obligation, intergenerational living, societal values, structures, and their respective implications for public health. Similarly, we need a further understanding of the micro-level relationship processes including relationship quality, relationship distress, interpersonal abuse, and interpersonal communication. Similarly, we need to consider the pathways by which such an association occurs at both macro- and micro-level processes. Macro-level processes including social engineering and wellness program participation, and micro-level processes including affect regulation and gene–environment interactions.

A new generation of research will also need to consider the role of technology in understanding social networks and interpersonal relationship functioning and the implications for

public health. Over the last decade or so, there has been a dramatic shift in the way in which people interact socially as a result of advancements in technology. Internet and mobile phone devices, and their associated services, have made it easier for people to communicate and connect with others than ever before. However, there is some evidence to suggest that despite these advances in technology, during this same time period people are reporting greater loneliness and social isolation (McPherson and Smith-Lovin, 2006). Thus, future research will need to consider both the potential benefits and deleterious effects of technology-mediated social networks and social interactions, and the implications for utilizing such technology in interventions.

See also: Caregiver Burden; Coronary Heart Disease: Psychosocial Aspects; Health-Related Support and Self-help Mutual Aid Groups; Occupational Stress Management; Religion and Health: Clinical Considerations and Applications; Social Epidemiology; Social Exclusion, Social Deprivation and Health.

Bibliography

Aspinwall, L.G., Taylor, S.E., 1997. A stitch in time: self-regulation and proactive coping. Psychological Bulletin 121, 417–436.

Barnes, J.A., 1954. Class and committees in a Norwegian Island parish. Human Relations 7, 39–58. http://dx.doi.org/10.1177/001872675400700102.

Barrera, M.R., 2000. Social support research in community psychology. In: Rappaport, J., Seidman, E. (Eds.), Handbook of community psychology. Kluwer Academic Publishers, Dordrecht Netherlands, pp. 215–245.

Berkman, L.F., Syme, S.L., 1979. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. American Journal of Epidemiology 109 (2), 186–204.

Bott, E., 1957. Family and Social Network. Tavistock Press, London.

Bowlby, J., 1969. Attachment and Loss. Hogarth Press, London.

Bolger, N., Zuckerman, A., Kessler, R.C., 2000. Invisible support and adjustment to stress. Journal of Personality and Social Psychology 79 (6), 953–961.

Brummett, B.H., Babyak, M.A., Siegler, I.C., Vitaliano, P.P., Ballard, E.L., Gwyther, L.P., Williams, R.B., 2006. Associations among perceptions of social support, negative affect, and quality of sleep in caregivers and noncaregivers. Health Psychology 25 (2), 220–225. http://dx.doi.org/10.1037/0278-6133.25.2.220.

Cassel, J., 1976. The contribution of the social environment to host resistance: the Fourth Wade Hampton Frost Lecture. American Journal of Epidemiology 104 (2), 107–123.

Christakis, N.A., Fowler, J.H., 2013. Social contagion theory: examining dynamic social networks and human behavior. Statistics in Medicine 32 (4), 556–577. http://dx.doi.org/10.1002/sim.5408.

Cobb, S., 1976. Social support as a moderator of life stress. Psychosomatic Medicine 38, 300–314.

Cohen, S., Herbert, T.B., 1996. Health psychology: psychological factors and physical disease from the perspective of human psychoneuroimmunology. Annual Review of Psychology, 47113–47142. http://dx.doi.org/10.1146/annurev.psych.47.1.113.

Cohen, S., 1988. Psychosocial models of the role of social support in the etiology of physical disease. Health Psychology 7, 269–297.

Cohen, S., 2003. Psychosocial models of the role of social support in the etiology of physical disease. In: Salovey, P., Rothman, A.J. (Eds.), Social psychology of health. Psychology Press, pp. 227–244.

Cohen, S., Wills, T.A., 1985. Stress, social support, and the buffering hypothesis. Psychological Bulletin 98 (2), 310–357.

Cutrona, C.E., Russell, D.W., 1990. Type of social support and specific stress: towards a theory of optimal matching. In: Sarason, B.R., Sarason, I.G., Pierce, G.R. (Eds.), Social Support: An Interactional View. John Wiley and Sons, New York, pp. 319–366.

Durkheim, E., 1897. Suicide. Free Press, New York.

Friedman, H.S., Tucker, J.S., Schwartz, J.E., Tomlinson-Keasey, C., Martin, L.R., Wingard, D.L., Criqui, M.H., 1995. Psychosocial and behavioral predictors of longevity: the aging and death of the 'Termites'. American Psychologist 50 (2), 69–78.

Fuemmeler, B.F., Mâsse, L.C., Yaroch, A.L., Resnicow, K., Campbell, M., Carr, C., Williams, A., 2006. Psychosocial mediation of fruit and vegetable consumption in the body and soul effectiveness trial. Health Psychology 25 (4), 474–483. http://dx.doi.org/10.1037/0278-6133.25.4.474.

Gore, S., 1981. Stress-buffering functions of social supports: an appraisal and clarification of research models. In: Dohrenwend, B., Dohrenwend, B. (Eds.), Stressful Life Events and Their Context. Prodist, New York, pp. 202–222.

House, J.S., Landis, K.R., Umberson, D., 1988. Social relationships and health. Science 241, 540–545.

Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Medicine 7 (7), e1000316. http://dx.doi.org/10.1371/journal.pmed.1000316.

Jessor, R., Turbin, M.S., Costa, F.M., 1998. Protective factors in adolescent health behavior. Journal of Personality and Social Psychology 75 (3), 788–800. http://dx.doi.org/10.1037/0022-3514.75.3.788.

Lakey, B., Orehek, E., 2011. Relational regulation theory: a new approach to explain the link between perceived support and mental health. Psychological Review 118, 482–495.

Lorenz, K.K., 1935. Der Kumpan in der Umwelt des Vogels. Der Artgenosse als auslösendes Moment sozialer Verhaltungsweisen. Journal FürOrnithologie. Beiblatt. (Leipzig) 83, 137–213. http://dx.doi.org/10.1007/BF01905355AU3.

Mermelstein, R., Cohen, S., Lichtenstein, E., Baer, J.S., Kamarck, T., 1986. Social support and smoking cessation and maintenance. Journal of Consulting and Clinical Psychology 54 (4), 447–453. http://dx.doi.org/10.1037/0022-006X.54.4.447.

McPherson, M., Smith-Lovin, L., 2006. Social isolation in America: changes in core discussion networks over two decades. American Sociological Review 71, 353–375.

Mitchell, J.C., 1969. The Concept and Use of Social Networks. Manchester University Press, Manchester, UK.

Okun, M.A., Ruehlman, L., Karoly, P., Lutz, R., Fairlholme, C., Schaub, R., 2003. Social support and social norms: do both contribute to predicting leisure-time exercise? American Journal of Health Behavior 27 (5), 493–507. http://dx.doi.org/10.5993/AJHB.27.5.2.

Pearlin, L.I., 1989. The sociological study of stress. Journal of Health And Social Behavior 30 (3), 241–256. http://dx.doi.org/10.2307/2136956.

Peirce, R.S., Frone, M.R., Russell, M., Cooper, M.L., 1996. Financial stress, social support, and alcohol involvement: a longitudinal test of the buffering hypothesis in a general population survey. Health Psychology 15, 38–47.

Rozanski, A., Blumenthal, J.A., Kaplan, J., 1999. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation 27, 99 (16), 2192–2217.

Steptoe, A., Wardle, J., Pollard, T.M., Canaan, L., 1996. Stress, social support and health-related behavior: a study of smoking, alcohol consumption and physical exercise. Journal of Psychosomatic Research 41 (2), 171–180. http://dx.doi.org/10.1016/0022-3999(96)00095-5.

Schaefer, C., Coyne, J.C., Lazarus, R.S., 1981. The health-related functions of social support. Journal of Behavioral Medicine 4 (4), 381–406. http://dx.doi.org/10.1007/BF00846149.

Thoits, P.A., 2011. Mechanisms linking social ties and support to physical and mental health. Journal of Health and Social Behavior 52, 145–161.

Thoits, P.A., 1983. Multiple identities and psychological well-being: a reformulation and test of the social isolation hypothesis. American Sociological Review 48 (2), 174–187.

Tucker, J.S., Friedman, H.S., Wingard, D.L., Schwartz, J.E., 1996. Marital history at midlife as a predictor of longevity: alternative explanations to the protective effect of marriage. Health Psychology 15, 94–101.

Uchino, B.N., Cacioppo, J.T., Kiecolt-Glaser, J.K., 1996. The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. Psychological Bulletin 119 (3), 488–531.

Uchino, B.N., Uno, D., Holt-Lunstad, J., 1999. Social support, physiological processes, and health. Current Directions in Psychological Science 8 (5), 145–148. http://dx.doi.org/10.1111/1467-8721.00034.

Uchino, B.N., 2004. Social Support and Physical Health: Understanding the Health Consequences of Our Relationships. Yale University Press, New Haven, CT.

Uchino, B.N., 2006. Social support and health: a review of physiological processes potentially underlying links to disease outcomes. Journal of Behavioral Medicine 29 (4), 377–387. http://dx.doi.org/10.1007/s10865-006-9056-5.

Uchino, B.N., Holt-Lunstad, J., Uno, D., Campo, R., Reblin, M., 2007. The social neuroscience of relationships: an examination of health-relevant pathways. In: Harmon-Jones, E., Winkielman, P. (Eds.), Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior. Guilford Press, New York, pp. 474–492.

Umberson, D., 1987. Family status and health behaviors: Social control as a dimension of social integration. Journal of Health and Social Behavior 28, 306–319.

Social Intelligence and Competencies

Susanne Weis, Center for Psychological Methods, Diagnostics, and Evaluation, University of Koblenz-Landau, Landau, Germany
Kristin Conzelmann, German Aerospace Center, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews and integrates research on social intelligence and competence. Firstly, the differences in the definitions of abilities, skills, intelligence, competence, individuals' potentials, and behavior are described. Secondly, theoretical models and measurement methods of social intelligence are contrasted, and requirements for social cognitive ability tasks are established. Thirdly, some representative conceptualizations of social competence are systematically compared and integrated into a framework of socially competent behavior that differentiates between the individual's potential and the resulting behavior. Finally, measurement issues of social skills and competence are discussed (e.g., accounting for situational variability and the question of adequate outcome criteria).

Introduction

During the last decades, new constructs (e.g., emotional and social intelligence, see Mayer et al., 2000; Weis and Süß, 2005 for overviews) in addition to the classical academic understanding of intelligence and personality have been considered capable of describing human attributes more comprehensively. Although still facing methodological problems, emotional intelligence attained the most prominent position (Lievens and Chan, 2010), while social intelligence with the longest history languishes in stagnation (Conzelmann et al., 2013; Landy, 2006; Weis and Süß, 2005). At the same time, social skills and competence received increased attention in applied disciplines such as personnel selection and development, clinical psychology (e.g., autism), and in school settings.

This article describes and integrates the various approaches to specify and assess social intelligence and competence. In the following sections, we deal with the terminology as applied in this article, the construct of social intelligence, and finally we describe and integrate conceptualizations of social competence into a framework of socially competent behavior.

Terminology

Imagine the following situation:

Paul, a team leader in a large corporation, has been asked by his superior, David, to hand in the annual team report earlier than expected. The report is in part the basis for salary increases. At the moment, one of his subordinates, Andrea, is faced with an ongoing divorce. He depends on her contribution to finish the report.

What are the resources Paul needs to accomplish the situation? Apart from abstract academic abilities (e.g., verbal reasoning or retrieval abilities), expert knowledge, and behavioral and cognitive skills (e.g., computer skills or writing skills), his social and emotional abilities and skills can be helpful to accomplish the situation:

1. Social cognitive abilities (e.g., interpret David's verbal and nonverbal behavior in terms of the severity of the situation; anticipate how Andrea will feel when he has to confront her with the problem).
2. Social knowledge (e.g., knowing possible action alternatives or how Paul himself can emotionally cope with the situation in order to function effectively).
3. Social behavioral and cognitive skills (e.g., communication skills such as sending the appropriate nonverbal cues when talking to Andrea, skills in self-regulation such as being capable of applying techniques to adapt to the situation).

All the above-mentioned concepts allow the assessment of individuals as more or less capable, more or less skilled, and so on. Apart from these cognitive and behavioral resources, what are further individual determinants of Paul's behavior?

a. Paul's short-term goals (e.g., Paul might be willing to present himself as an understanding person because he is in love with a colleague, Andrea, and wants to make a good impression).
b. Paul's long-term goals (e.g., Paul might have a job offer that he is planning to accept so that he doesn't care about the deadline).
c. Paul's general interests, motivations, and beliefs (e.g., Paul might believe that everyone is responsible for his or her own good and, thus, does not feel responsible for Andrea).
d. Personality traits (i.e., Paul could be a highly agreeable person increasing the probability for respectful behavior in this specific situation).

Last but not least, the characteristics of the situation as well as the cultural context will influence his behavior (e.g., Paul might adjust his behavior when he finds Andrea drenched in tears after a conversation with her husband. If the corporation has a good working atmosphere, the probability of him being more understanding is greater).

Considering all the aforementioned characteristics and behavioral alternatives, what determines whether Paul can be evaluated as socially competent? Is it enough to possess the cognitive and behavioral resources (1–3) so that he could behave competently if he wanted to? Is he only a socially competent person when he attains his personal goals? Or is it necessary that his behavior, those short- or long-term goals or

his beliefs (a–d) which guided his behavior, are socially acceptable?

The above list of individual resources may not be exhaustive and does not go into the details of the concepts mentioned. It does not address how the different abilities and skills overlap or affect each other. However, it exemplifies the following terminological specifications.

Intelligence versus Competencies

The concepts of intelligence and competence are often applied as synonyms (Süß et al., 2005). Some distinctive features commonly accepted in the literature are the context-specificity of competence (Kanning, 2002; McFall, 1982) and the more general meaning of intelligence constructs across situations or contexts (Süß et al., 2005). Competence appears to be more subject to modification and learning (Rose-Krasnor, 1997), whereas intelligence is comparatively stable over time and seen as hereditary to a large extent (see Deary et al., 2006 for an overview). However, intelligence is often a necessary part of the evaluation of competence (Süß et al., 2005).

Abilities versus Skills

According to Scherer (2007), behavioral or cognitive skills are concrete actions or applications of cognitive operations on concretely defined problems (e.g., driving with a stick shift or applying an algorithm on new data). Skills are acquired in a process of several steps and are finally characterized by an automated series of actions (Ackerman, 1987). Thus, there is reason to infer that skills are either the successful application of situation-specific knowledge or even procedural knowledge itself. In contrast, abilities are only cognitive in nature and represent more general, dispositional capacities (Süß et al., 2005). They typically constitute intelligence constructs (e.g., reasoning ability) and thus, are "either genetically endowed or acquired over a long period of socialization" (Scherer, 2007, p. 103). Besides abilities, skills are typically part of an evaluation of competence.

Potential versus Behavioral Performance

The individual's potential represents all resources of a person, which are necessary to allow competent behavior (i.e., the result) (Süß et al., 2005). Results-oriented approaches understand social competence as effective and/or socially accepted behavior where effectiveness/social acceptance is determined through the degree of goal attainment, specific properties of the situation, and the person who assesses whether and to what degree the behavior points toward social competence (McFall, 1982; Rose-Krasnor, 1997). Whether a person shows effective behavior is not a deterministic function of his or her potential. Additionally, it depends on the person's priorities (e.g., motivation and goals), his or her personality (e.g., extraversion and agreeableness), current states (e.g., fear, exhaustion, and distraction), and on situational and context variables (e.g., broad cultural or group values) (Süß et al., 2005).

While the distinction between potential and result is self-evident in research focusing on abstract academic intelligence (i.e., assessing indicators of job performance as behavioral criterion to prove predictive validity of intelligence constructs), this may not apply in the context of social intelligence, where results-oriented approaches are frequently applied to conclude from behavior to the potential (e.g., Assessment Center exercises; Hoffman et al., 2011).

Requirements for an Intelligence Construct

With the introduction of new constructs, the diversity of labels (e.g., social skills, social and emotional competence and intelligence, and successful intelligence) and assessment approaches (e.g., ability tests, self-report inventories, and Situational Judgment Tests) has increased. However, there are many conceptual and methodological requirements to determine how well an intelligence construct is established (Landy, 2006; Matthews et al., 2005; Schaie, 2001; Weber and Westmeyer, 2001). Some commonly agreed upon requirements established in the literature are (see also Conzelmann et al., 2013):

1. Theory and construct specification
   Matthews et al. (2005) require a priori theoretical considerations on the "localization of the intended construct within the sphere of individual differences" (p. 80). O'Sullivan (1983) demands a clear and nonredundant terminology while Süß (2001) asks for an empirical foundation. Additionally, constructs and subconstructs need to be defined and categorized taking existing hierarchical structures including broad, cognitive determinants (Carroll, 1993) into account which are relatively stable over time and contain a minimum amount of knowledge (Süß, 2001).

2. Operationalizations and validation
   With respect to the psychometric qualities, Matthews et al. (2005) and Weber and Westmeyer (2001) demand psychometrically sound operationalizations congruent with construct specifications (O'Sullivan, 1983). Moreover, performance tests should be applied (Süß, 2001; Weber and Westmeyer, 2001) in the sense of Cattell's conception of T-data (Cattell, 1965). To investigate the convergent and divergent construct validity, adequate operationalizations for the construct of reference have to be selected (Schaie, 2001). Concerning criterion validity, studies have to provide evidence for the incremental predictive validity of heterogeneous external criteria over and above traditional predictor variables (Süß, 2001).

3. Other requirements
   Genetic determination to a reasonable extent is an unquestioned requirement (Austin and Saklofske, 2005), whereas there is no agreement about a positive correlation to age due to increased experience as postulated by Schaie (2001). The latter is in conflict with the requirement of only a minimum amount of knowledge in construct specifications (Süß, 2001). Moreover, it contradicts findings about fluid intelligence yielding a generally curvilinear relationship with age (Schaie, 1994).
   Among others, Austin and Saklofske (2005), Gottfredson (2003) and Neubauer and Freudenthaler (2005) all demand positive correlations between tasks of academic and new intelligence constructs to underpin the idea of the

positive manifold (Jensen, 1998). This view is questioned by some other groups of researchers (e.g., Guttman, 1992; Henry et al., 2005; Horn, 1985). For a comprehensive picture of the entire debate, see Bowman et al. (2002). At the moment, there appears to be no necessity to assume positive correlations with academic intelligence as a 'conditio sine qua non' for new intelligence constructs.

Social Intelligence

In a model of social competence, social intelligence represents one of the core individual resources. Several literature reviews exist (Landy, 2006; Walker and Foley, 1973; Weis and Süß, 2005) which sum up the research history that was castigated by Landy (2006) as a 'long, frustrating, and fruitless search.' The problems are manifold: definitions did not differentiate between cognitive and behavioral requirements, or abilities and skills (Conzelmann et al., 2013). Frequently, self-report inventories were applied (Riggio, 1986). The majority of performance tests were based on verbal material only (e.g., Chapin Social Insight Test; Chapin, 1967) and showed no divergent construct validity with tests of academic intelligence.

The following sections intend to give an overview of the conceptual basis of social intelligence as a cognitive ability construct of existing measurement approaches and of current data.

Theoretical Models

Apart from numerous single definitions, two more comprehensive theoretical approaches are apparent in the literature. Guilford's Structure of Intellect Model (Guilford, 1967) integrates a social content component in addition to abstract academic abilities that results in 30 distinct ability factors of social intelligence (in a revision of the model assumed to be correlated). The entire model, however, was not supported by data (Brody, 2000). Weis and Süß (2005) classify existing definitions as all addressing some social cognitive ability. This resulted in a Model of Social Intelligence based on four presumably correlated social cognitive ability factors (i.e., understanding, memory, perception, and flexibility) plus social knowledge and a higher-order social intelligence factor. Social knowledge, however, is not seen to rely on only cognitive requirements because knowledge contents always depend on situational or cultural influences. For a list of concepts and labels assigned to the four ability domains, see Conzelmann et al. (2013). The four domains were specified as follows:

- Social understanding (also social inference, social interpretation, or social judgment): the ability to understand social stimuli against the background of the given social situation.
- Social memory: the ability to store and recall objectively given social information that can vary in complexity.
- Social perception: the ability to perceive socially relevant information quickly in fairly complex situations.
- Social flexibility: the ability to produce as many and as diverse solutions or explanations as possible for a social situation or a social problem.

The model was partly supported in a multitrait–multimethod study (Weis and Süß, 2007) applying established tests for social understanding and social knowledge and newly constructed tests for social memory. The confirmatory factor model showing the best fit specified three separable but correlated ability factors (i.e., social understanding, social memory, and social knowledge) and one method factor with loadings of all tests based on verbal material (i.e., written language). However, social memory was not sufficiently distinguishable from academic intelligence (Conzelmann et al., 2013) to be considered as a unique social intelligence factor.

The relationship of social to emotional intelligence has not so far been developed in detail. On a broad construct level, several researchers agree on an as yet unspecified overlap (Davies et al., 1998; Mayer et al., 2000). Few studies have investigated the relationship (Davies et al., 1998; Weis and Süß, 2007) and those have found inconsistent results. Regarding the Four Branch Model of Emotional Intelligence (Mayer et al., 2000) and the above-mentioned Model of Social Intelligence (Conzelmann et al., 2013), the domains of perception and understanding are represented in both models. Meanwhile, social contents cover information about emotions but also reach further. However, the question of construct overlap needs more profound empirical clarification based on measures derived from and validated by these or other theoretical models.

Measurement Approaches

In this report, we are only focusing on cognitive ability tests and are omitting self-reported inventories and behavior-based approaches because they do not comply with the aforementioned requirements. Compared to the construction of academic intelligence tests, no comparable body of rules is available for items or tasks of social intelligence. In contrast to tasks such as number series, the item pool is not documented and possibly not even available. The particular tasks that measure social cognitive abilities are not specifically taught in institutional settings. Therefore, numerous characteristics of social intelligence tasks can be identified. **Table 1** provides exemplary psychometric tests representative for the variability of measurement approaches:

- Comprehensive test batteries based on theoretical models (e.g., Magdeburg Test of Social Intelligence, MTSI, Conzelmann et al., 2013) versus single modalities.
- Different task materials (i.e., written and spoken language, pictures, and videos).
- Different item characteristics (i.e., item origin, stimulus properties, and context information).
- Different formal characteristics (i.e., response format and scoring).
- Different queried modalities in social understanding tasks (i.e., what modality has the subject to infer about such as emotions, thoughts, and relationships).

Additionally, **Table 1** contains short global validity information. For reasons of comparison, the Model of Social Intelligence (Conzelmann et al., 2013) serves as a frame of reference for the classifications of task requirements.

Empirical evidence concerning the convergent and discriminant construct validity is equivocal and varies

Table 1 Characteristics of social cognitive ability tasks showing representative example tests

	Taxonomic classifications								
Example test	*Domain*	*Material*	*Item origin*	*Stimulus properties*	*Context information*	*Response format*	*Scoring*	*Queried modality*[a]	*Validity (convergent/ discriminant)*
Social Insight Test (Chapin, 1967)	SU	V	G	EV	AV	MC	GCS	PS	Low/low
Four-Factor Test of Social Intelligence (O'Sullivan and Guilford, 1976)	SU	V P	AR	UV	AV NAV	MC	ECS	E B PS	Low (SI tests), high (tests of same battery)/low
Level of Emotional Awareness Scale (Lane et al., 1990)	SU	V	AR	EV	AV	O	R	E	High/–
Interpersonal Perception Task-15 (Costanzo and Archer, 1993)	SU	F	G	EV	AV	MC	SBS	T R B	Low to medium/high
Diagnostic Analysis of Nonverbal Accuracy Scale (Nowicki and Duke, 2001)	SU	A P	P	EV	NAV	MC	SBS	E	High (with DANVA subtests)/–
Facial Emotion Inspection Time Task (Austin, 2004)	SP	P	AR, P	EV	NAV	RT	SBS	E	High (with response latency tests)/high (verbal academic intelligence)
Situational Test of Emotional Understanding (MacCann and Roberts, 2008)	SU	V	A	EV	AV	MC LR	SBS GCS	E	High/low (verbal academic intelligence)
Situational Test of Emotional Management (MacCann and Roberts, 2008)	SU	V	G	EV	AV	MC LR	GCS	E B	High/low (verbal academic intelligence)
Magdeburg Test of Social Intelligence (Conzelmann et al., 2013)	SU SM SP	V A P F	G	EV (SU) UV (SM, SP)	AV	MC LR RT	SBS (SM, SP) TS (SU)	E T R PT	–/high (SU with academic intelligence, medium (SM and SP with academic intelligence)

Note: SU: social understanding, SM: social memory, SP: social perception, V: written language, A: spoken language, P: pictures, F: video film, G: genuine, AR: artificial, P: posed, EV: equivocal, UV: univocal, AV: available, NAV: not available, MC: multiple choice, LR: Likert ratings, O: open, RT: response latency, R: rating of open answers, SBS: standards based, TS: target scoring, GCS: group consensus scoring, ECS: expert consensus scoring, Low: no sufficient convergent or discriminant validity evidence, High: good convergent or discriminant validity evidence, –: no information available.
[a]Only relevant for tasks of social understanding: E: emotions, T: thoughts, RS: relationships, PT: personality traits, B: future behavior, PS: problem solutions.

substantially in connection with task characteristics just as does the psychometric quality (see Weis and Süß, 2005 for a review). To summarize, applying decontextualized, artificially produced material based on written language resulted in substantial correlations with tasks of abstract academic intelligence, thus questioning the discriminant construct validity (e.g., Chapin Social Insight Test, Chapin, 1967). Additionally, not accounting for the queried modality seemed to impair convergent construct validity (Ambady et al., 2001), just as did different response formats.

Tasks of social understanding that require inferring the mental states of target individuals or offering possible explanations or solutions for complex social problems carry additional problems. Particularly, the application of equivocal stimuli without context information makes inference about any mental state highly difficult or even impossible. Another, so far unsolved problem demonstrated in research on emotional intelligence, is the scoring of tasks when no standard (i.e., objectively correct answer) is available. All scoring methods apart from standards-based scoring (i.e., target scoring and expert or group consensus scoring) carry different problems (see Conzelmann et al., 2013 for a discussion). In short, we consider target scoring rather than expert or group consensus scoring as the best known approximation to objective scoring when questions of social desirability are accounted for and the target's mental state can be defined by means of standardized procedures.

As a conclusion from the above considerations and the empirical literature, the MTSI (Conzelmann et al., 2013) was developed to account for some necessary requirements for social cognitive abilities tasks:

- Theoretical foundation: Task and item construction was founded in the Model of Social Intelligence.
- Multitrait–multimethod/multimodality designs: The systematic variation of task material (i.e., written and spoken language, pictures, and videos) resulted in better interpretable validity evidence and a smaller overlap with the abstract academic intelligence of past studies (Weis and Süß, 2007; Wong et al., 1995).
- Origin of stimulus material: In contrast to artificial or posed material, stimuli with real-life origins (i.e., particularly material other than written language) enhance the ecological validity and the acceptance and promote discriminant construct validity with abstract academic intelligence (e.g., IPT-15, Costanzo and Archer, 1993). Moreover, target information or even objective standards can be sampled for scoring. Transfer to real-life situations is more probable.

- Context information: Tasks with equivocal stimulus material (i.e., tasks of social understanding) should provide enough context information to allow correct inferences. When context information is embedded in stimulus material, it should be varied systematically in order to reduce the context- or situation-specific influence (e.g., through experience or knowledge) to a minimum.

Finally, social intelligence as assessed by existing measurement approaches still lacks evidence for its predictive validity in relevant, real-life outcome criteria. One criterion validity study was conducted by Hampel et al. (2011) applying the MTSI and finding the theory-consistent relations with the level of social anxiety in a general population sample.

Social Competence

We consider social competence as the entity of all individual resources (i.e., the potential) necessary to show social competent behavior in varying types of applied settings (adapted from Kanning, 2002). The following sections give an overview of existing conceptual approaches and integrate them into a framework of social competent behavior. Later, measurement issues are addressed.

Concepts: Comparison and Integration

Definitions of social competence vary substantially according to the spectrum of covered human attributes from just one characteristic (e.g., management of conflict and communication skills) to a complex interaction of assorted variables (see Rose-Krasnor, 1997; Nangle et al., 2010 for reviews). They may either focus on the person's potential, or the behavioral result, or a mixture of both. Additionally, some definitions only incorporate concepts, which allow a evaluation of a person's underlying competence as opposed to definitions that also include individual dispositions or traits (e.g., personality and motivation), not allowing an evaluation of 'better or worse' (Rose-Krasnor, 1997; Schneider et al., 1996). **Table 2** contrasts some selected, comprehensive, and representative approaches.

By definition, the four approaches (**Table 2**) show substantial differences. McFall (1982) developed a process model, which allows for individual characteristics (i.e., skills) besides situational variables, behavior, and the person who evaluates the behavior. Contrary to the other three mentioned, Schneider et al. (1996) conceive social competence as independent from situations, and assessment about the level of competence is only based on the degree of goal attainment. This is in contrast to social acceptance which is relevant in the concepts of Rose-Krasnor (1997) and Kanning (2002).

Table 2 Definitions of social competence

		McFall (1982)	*Schneider et al. (1996)*	*Rose-Krasnor (1997)*	*Kanning (2002)*
Definitions		"Social skills are defined [ ] as the specific abilities that enable a person to perform competently at particular social tasks." (p. 23) "Competence is a generally evaluative term that reflects somebody's judgment, on the basis of certain criteria, that a person's performance on some task is adequate." (p. 13)	"We define social competence as socially effective behavior and its cognitive, affective and conative antecedents. Socially effective behavior is behavior that is instrumental in helping people achieve personal goals that are social in nature." (p. 471)	"… effectiveness in interaction, considered from both self and other perspectives. Social competence is viewed as an organizing construct, with transactional, context-dependent, performance oriented, and goal-specific characteristics." (p. 123)	Entity of a person's knowledge, abilities, and skills which promote socially competent behavior (=context-specific behavior for attaining a personal goal accounting for the social acceptance of the behavior).
Domains	Abilities	(Skills specified as abilities, see definitions above)	Social intelligence (social insight, memory, knowledge)	Emotional understanding	Social perception
	Skills/ knowledge	Decoding skills, decision skills, encoding skills	Social skills, social self-regulation	Problem-solving skills, communication skills, self-control	Behavioral control, assertiveness, communication skills
	Traits	–	Interpersonal personality traits	Motivation, personality traits, prosocial behavior	Social orientation
Orientation		Process	Potential and behavior	Behavior	Potential
Specific to situation		Yes	No	Yes	Yes
Evaluating perspective		Evaluator (x situation x performer)	Goal attainment	Goal attainment and social acceptance	Goal attainment and social acceptance

Table adapted and complemented from Süß, H-M., Weis, S., Seidel, K., 2005. Soziale Kompetenzen [Social Competences]. In: Weber, H., Rammsayer, T. (Eds.), Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie, (Reihe Handbuch der Psychologie) [Handbook of Personality Psychology and Differential Psychology]. Hogrefe, Göttingen, pp. 350–362.

However, leaving the broad level of definitions and looking at the single postulated dimensions of social competence, the approaches overlap. Except for McFall's process model that relies only on social skills, the remaining three approaches incorporate abilities (e.g., emotional understanding and social perception), skills (e.g., decoding skills and self-control), as well as behavioral dispositions (e.g., social orientation, personality traits, and motivation).

Based on the above-mentioned predefinitions, concepts are integrated into a framework of socially competent behavior (see **Figure 1**; modified based on Süß et al., 2005).

Summarizing the above considerations, the framework is specified as follows:

1. Social competence consists of a person's potential and does not include overt behavior in applied setting.
2. The person's potential only includes those constructs and concepts that allow a evaluation of 'better or worse' (i.e., abilities and skills). Any dispositional constructs (i.e., personality and attitudes) are left out although we acknowledge that they ultimately influence the probability of social competent behavior.
3. Within the person's potential, cognitive abilities (i.e., social and emotional intelligence, which are assumed to overlap), skills, and knowledge concepts are clearly differentiated. Abilities are broad, situation independent, relatively stable, and cognitive in nature. Skills are situation specific, subject to modification and learning, and can refer to cognitive operations or behavioral sequences. In contrast to socially competent behavior in applied settings, social behavioral skills are more elementary and instead represent the microscopic level of human behavior. Therefore, the application of skills always stands in relation with situational variables but is typically not influenced by an individual's dispositions and is independent of the person who evaluates the behavior. Declarative knowledge refers to explicit knowledge (e.g., social rules, socially relevant situational scripts, or situation-specific patterns of emotional reactions). Procedural knowledge, typically, is implicit knowledge which cannot be explicitly learned or taught or deliberately recalled (e.g., how to use the tone of voice to convey or trigger a certain feeling, how to apply heuristics in order to choose the best behavioral alternative). The gap between procedural knowledge and the application of it in the form of cognitive or behavioral skills cannot be filled in by drawing upon the literature. This gap appears to be rather small if it exists at all, particularly when measurement opportunities are considered (see next section).
4. Socially competent behavior is predicted by the individual's level of social competence (represented by the solid arrows in **Figure 1**) over and above variables in disposition or temperament. As indicated by the different strength of the

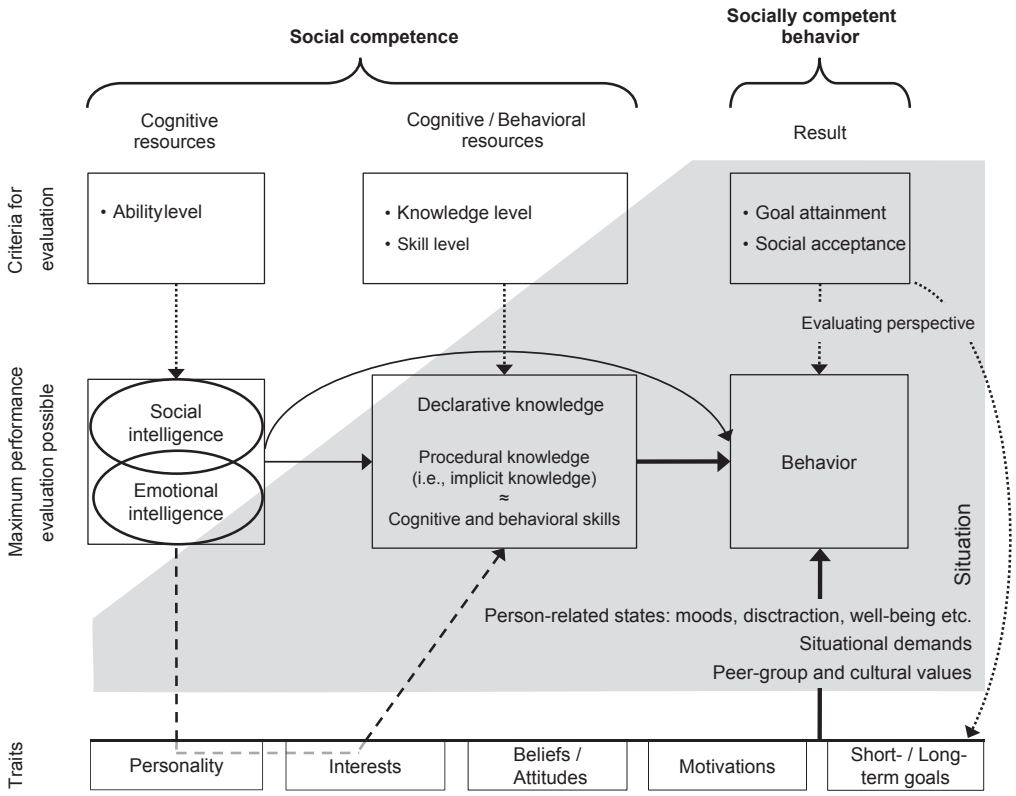

Note. Continuous arrows = prediction (predictor–criterion-relationship)
broken arrow = process of skill / knowledge acquisition
dotted arrow = evaluation (arrowhead points toward behavior/goal to be evaluated)

Figure 1 Framework of social competence and socially competent behavior.

arrows, we assume a larger predictive power of skills and knowledge compared to abilities that are due to construct and measurement similarities. Moreover, social cognitive abilities, social skills, and declarative knowledge are assumed to be positively related constructs.

5. In accordance with the PPIK-Theory (i.e., Process–Personality–Interest–Knowledge, Ackerman, 1996), skills (seen as almost equal to procedural knowledge) and declarative knowledge are acquired by the investment of fluid intelligence (i.e., social and emotional intelligence) into areas of interest and which fit to the individual's personality (indicated by the broken arrows in **Figure 1**).
6. Abilities and declarative knowledge are typically assessed by psychometric tests so that the assessment (indicated by dotted arrows in **Figure 1**) of 'more or less able' or 'more or less knowing' is based on the test score independent of an analyzing perspective (i.e., objectivity). Measuring procedural knowledge or skills is more difficult although still remains independent of subjective interpretation. Whether the demonstration of social behavior is evaluated as socially competent primarily depends on the attainment of the behavioral goals of the individual. Additionally, when social behavior (or its underlying goal) is evaluated based on its social acceptance, the evaluating perspective needs to be considered as well (e.g., individuals from different cultural backgrounds or different peer groups may find differing behavioral actions socially acceptable) (McFall, 1982). These assumptions also mean that an assessment of an individual's social competence based only on his or her behavior in highly complex, applied situations should not be done or only done carefully when a goal is not predetermined (e.g., selection situations) and the situation allows unsystematic influences of conflicting variables that should not be the basis for assessment (e.g., the relevant peer group is not known). In such situations, several reasons for showing or not showing socially competent behavior may exist apart from a lack of social competence.

Measurement and Design Issues

The landscape of measurement approaches for social skills and social behavior is even more heterogeneous than for social intelligence. **Table 3** presents three classes of measurement approaches (i.e., inventories, knowledge tests, and behavioral assessment) cross-classified with constructs including a global assessment of the construction efforts, the possibility of response distortion, and the objectivity. (For a more profound review of measurement approaches to assess social and emotional skills in children and young people, see Humphrey et al., 2011.)

Self-reported and other similar inventories are easily constructed and highly objective but carry the problem of response distortion, particularly in relation with constructs involving evaluation. Implicit knowledge tests (also Situational Judgment Tests, see McDaniel et al., 2007 for a meta-analysis) allow an objective assessment of procedural knowledge and cognitive and behavioral skills that typically cannot be distorted by the subject. The testing principle presents a description of a situation and a number of answering alternatives providing explanations, solutions, or possible behavioral actions. In multiple choice or ratings-based items, individuals have to evaluate the effectiveness of the behavioral actions for solving a given problem or reaching a goal. Although item contents frequently address the behavior, the *knowledge* about behavior is assessed instead ($\approx$procedural knowledge $\approx$ cognitive and behavioral skills). The last category of measurement approaches comprises behavioral observation methods. Observations can be done in experimental or applied settings (e.g., Assessment Center exercises such as role plays). To evaluate social skills or procedural knowledge, specific experimental settings would be required to provoke people to show their ideal behavior and to reduce variance due to states or confounding personality variables in applied settings (e.g., through instructional variations). Socially competent behavior should be assessed in applied settings or, at least, settings with a certain level of complexity so that an average behavioral reaction has little or no chance of occurring. Behavioral observation is a highly complex research field of its own and the problems are multifaceted (i.e., reactivity and participatory observation; social desirability, finding an adequate rating system and inter-rater-reliability; time versus event sampling; and interdependence with exercise; see Hoffman et al., 2011).

Across all three types of measurement approaches, situational variability is a relevant feature. Nangle et al. (2011) provide some examples of (incomplete) classification systems of situations or social tasks and stress their importance not only for purposes of conceptualization and measurement but also for the identification of treatment opportunities. Additionally, specifying the characteristics of the situation also helps in evaluating the degree of generalization of empirical findings.

As pictured in **Figure 1**, socially competent behavior is the outcome of social competence which brings us back to the lack of criterion validity studies for social intelligence. Examining the specifications of the above framework not only provides information about the model itself, but also about the criterion validity of social intelligence given an adequate selection of

Table 3 Categories of measurement approaches to assess procedural knowledge, skills, and behavior

	Procedural knowledge/skills			*Behavior*		
	Construction effort	*Response distortion*	*Objectivity*	*Construction effort*	*Response distortion*	*Objectivity*
Self- and other report	Low	High	High	Low	High	High
(Implicit) knowledge test	High	Low	High		–	
Behavioral observation	High	Low	High	High	Medium	Low

relevant behavioral outcome criteria. In this respect, Lievens and Chan (2010) suggest to determine thoroughly what social intelligence should be valid for and what not. For example, they miss a theoretical basis for investigating the prediction of academic performance by emotional intelligence. In general, the predictive value of social intelligence, knowledge, and skills is different depending on the chosen outcome criterion. Moreover, the selection of outcome criteria needs to consider the hierarchical structures of predictors and criteria (Brunswik, 1956). Only when symmetrical levels in the hierarchy are chosen, the maximum predictability is achieved.

Conclusion

Particularly in the last sections, more questions were posed than answered. And we cannot, at that point, evaluate whether Paul – remember our leading character in the introduction – is a socially competent person without acknowledging several situational characteristics and his behavioral goals which we currently do not know about. To progress substantially in this field of research, several challenges need to be conquered. Conceptual issues have to be addressed empirically (e.g., the suggested framework of socially competent behavior requires empirical substantiation through multivariate research designs). Future research should focus on the development and improvement of measures of not only abilities but also procedural and declarative social knowledge and social skills. Several methodological problems have to be faced of which only one important is the development and use of a scoring procedure that is as objective as possible. In the field of social and emotional intelligence, some objective tests are already being applied (e.g., Conzelmann et al., 2013; Wilhelm et al., 2010), but it will be important to examine to which degree such an objective approach is possible taking into account the aforementioned complexity of the variables. Identifying both, the sources of situational variability (i.e., what are situational attributes that contribute to the variability in social behavior) and adequate and relevant real-life outcome criteria for social intelligence, should be further major aims. The latter is a particularly challenging issue since adequate social criteria are often highly specific and make transfer to practice very difficult. Although we mainly have dealt with the potential of an individual what is considered to be subject of change to only a limited degree, the extent to which trainability is possible (i.e., particularly considering the level of knowledge and skills) has to be investigated.

Last but not least, researchers who want to face these challenges should, at the very beginning of the undertaking, clarify whether social competence is conceived a value-free concept in the sense that not the behavioral goals are evaluated but only the degree of goal attainment. Otherwise, the value of not only goals but also personality characteristics needs to be identified (e.g., against the cultural background) prior to determine the degree of social competence. Facing these challenges and making progress seems to be a hard but worthwhile task. Should there be objective, reliable, and valid measures for social abilities, knowledge and skills, and appropriate methods to train aspects of the discussed constructs in the future, many different applied settings both personal and work could profit.

See also: Emotion, Perception and Expression of; Emotional Intelligence and Competencies; Facial Perception; Intelligence: Central Conceptions and Psychometric Models; Personal Intelligence and Competencies; Situational Judgment Test; Wisdom, Psychology of.

Bibliography

Ackerman, P.L., 1987. Individual differences in skill learning: an integration of psychometric and information processing perspectives. Psychological Bulletin 102, 3–27.

Ackerman, P.L., 1996. A theory of adult intellectual development: process, personality, interests, and knowledge. Intelligence 22, 229–259.

Ambady, N., LaPlante, D., Johnson, E., 2001. Thin-slice judgments as a measure of interpersonal sensitivity. In: Hall, J.A., Bernieri, F.J. (Eds.), Interpersonal Sensitivity. LEA, Mahwah, NJ, pp. 89–101.

Austin, E.J., 2004. An investigation of the relationship between trait emotional intelligence and emotional task performance. Personality and Individual Differences 36, 1855–1864.

Austin, E.J., Saklofske, D.H., 2005. Far too many intelligences? On the communalities and differences between social, practical, and emotional intelligence. In: Schulze, R., Roberts, R.D. (Eds.), An International Handbook of Emotional Intelligence. Hogrefe, Göttingen, pp. 107–128.

Bowman, D.B., Markham, P.M., Roberts, R.D., 2002. Expanding the frontiers of human cognitive abilities: so much more than (plain) g! Learning and Individual Differences 13, 127–158.

Brody, N., 2000. History of theories and measurements of Intelligence. In: Sternberg, R.J. (Ed.), Handbook of Intelligence. Cambridge University Press, New York, pp. 16–33.

Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments, second ed. University of California Press, Berkeley.

Carroll, J.B., 1993. Human Cognitive Abilities. A Survey of Factor-Analytical Studies. Cambridge University Press, New York.

Cattell, R.B., 1965. The Scientific Analysis of Personality. Penguin, Harmondswort, UK.

Chapin, F.S., 1967. The Social Insight Test. Consulting Psychologists Press, Palo Alto, CA.

Conzelmann, K., Weis, S., Süß, H.-M., 2013. New findings about social intelligence – development and application of the magdeburg test of social intelligence. Journal of Individual Differences 34, 119–137.

Costanzo, M., Archer, D., 1993. The Interpersonal Perception Task-15 (IPT-15) [Videotape]. University of California Extension Media Center, Berkeley.

Davies, M., Stankov, L., Roberts, R.D., 1998. Emotional intelligence: in search of an elusive construct. Journal of Personality and Social Psychology 75, 989–1015.

Deary, I.J., Spinath, F.M., Bates, T.C., 2006. Genetics of intelligence. European Journal of Human Genetics 14, 690–700.

Gottfredson, L.S., 2003. Dissecting practical intelligence theory: its claims and evidence. Intelligence 31, 343–397.

Guilford, J.P., 1967. The Nature of Human Intelligence. McGraw-Hill, New York.

Guttman, L., 1992. The irrelevance of factor analysis for the study of group differences. Multivariate Behavioral Research 27, 175–204.

Hampel, S., Weis, S., Hiller, W., Witthöft, M., 2011. The relations between social anxiety and social intelligence: a latent variable analysis. Journal of Anxiety Disorders 25, 545–553.

Henry, P.J., Sternberg, R.J., Grigorenko, E.L., 2005. Capturing successful intelligence through measures of analytic, creative and practical skills. In: Wilhelm, O., Engle, R.W. (Eds.), Handbook of Understanding and Measuring Intelligence. Sage Publications, Thousand Oaks, pp. 295–311.

Hoffman, B.J., Melchers, K.G., Blair, C.A., Kleinmann, M., Ladd, R.T., 2011. Exercises and dimensions are the currency of assessment centers. Personnel Psychology 64, 351–395.

Horn, J.L., 1985. Remodeling old models of intelligence. In: Wolman, B.B. (Ed.), Handbook of Intelligence: Theories, Measurements, and Applications. Wiley, New York, pp. 267–300.

Humphrey, N., Kalambouka, A., Wigelsworth, M., Lendrum, A., Wolpert, M., Aitken, J., 2001. A systematic review of measures of social and emotional skills in children and young people. Educational and Psychological Measurement 71, 617–637.

Jensen, A.R., 1998. The G Factor: The Science of Mental Ability. Praeger, Westport, CT.

Kanning, U.P., 2002. Soziale Kompetenz – definition, Strukturen und Prozesse. Zeitschrift für Psychologie 210, 154–163.

Landy, F.J., 2006. The long, frustrating, and fruitless search for social intelligence: a cautionary tale. In: Murphy, K.R. (Ed.), A Critique of Emotional Intelligence – What Are the Problems and How Can They Be Fixed? Lawrence Erlbaum Associates, NJ, pp. 81–123.

Lane, R.D., Quinlan, D.N., Schwartz, G.E., Walker, P.A., Zeitlin, S.B., 1990. The levels of emotional awareness scale: a cognitive-developmental measure of emotion. Journal of Personality Assessment 55, 124–134.

Lievens, F., Chan, D., 2010. Practical intelligence, emotional intelligence, and social intelligence. In: Farr, J.L., Tippins, N.T. (Eds.), Handbook of Employee Selection. Routledge Academic, New York, pp. 339–359.

MacCann, C., Roberts, R., 2008. New paradigms for assessing emotional intelligence: theory and data. Emotion 8, 540–551.

Matthews, G., Zeidner, M., Roberts, R.D., 2005. Emotional intelligence – an elusive ability. In: Wilhelm, O., Engle, R.W. (Eds.), Handbook of Understanding and Measuring Intelligence. Sage Publications, Thousand Oaks, pp. 79–99.

Mayer, J.D., Salovey, P., Caruso, D.R., 2000. Models of emotional intelligence. In: Sternberg, R.J. (Ed.), Handbook of Intelligence. Cambridge University Press, New York, pp. 396–420.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L., Grubb, W.L., 2007. Situational judgment tests, response instructions, and validity: a meta-analysis. Personnel Psychology 60, 63–91.

McFall, R.M., 1982. A review and reformulation of the concept of social skills. Behavioral Assessment 4, 1–33.

Nangle, D.W., Grover, R.L., Cassano, M., Holleb, L., Fales, J., 2010. Defining competence and identifying target skills. In: Nangle, D.W., Hansen, D.J., Erdley, C.A., Norton, P.J. (Eds.), Practitioner's Guide to Empirically Based Measures of Social Skills. Springer, New York, pp. 3–19.

Nangle, D.W., Grover, R.L., Fales, J., 2011. On the situational variability of social competence and the stability of trait-like conceptions. The Behavior Therapist 34, 40–46.

Neubauer, A.C., Freudenthaler, H.H., 2005. Models of emotional intelligence. In: Schulze, R., Roberts, R.D. (Eds.), An International Handbook of Emotional Intelligence. Hogrefe, Göttingen, pp. 31–50.

Nowicki Jr., S., Duke, M.P., 2001. Nonverbal receptivity: the diagnostic analysis of nonverbal accuracy (DANVA). In: Hall, J.A., Bernieri, F.J. (Eds.), Interpersonal Sensitivity. Theory and Measurement Lawrence Erlbaum Associates, Mahwah, NJ.

O'Sullivan, M., 1983. Measuring individual differences. In: Wiemann, J.M., Harrison, R.P. (Eds.), Nonverbal Interaction. Sage, Beverly Hills, CA, pp. 243–269.

O'Sullivan, M., Guilford, J.P., 1976. Four Factor Tests of Social Intelligence: Manual of Instructions and Interpretations. Sheridan Psychological Services, Orange, CA.

Riggio, R.E., 1986. Assessment of basic social skills. Journal of Personality and Social Psychology 51, 649–660.

Rose-Krasnor, L., 1997. The nature of social competence: a theoretical review. Social Development 6, 111–135.

Schaie, K.W., 1994. The course of adult intellectual development. American Psychologist 49, 304–313.

Schaie, K.W., 2001. Emotional intelligence: psychometric status and developmental characteristics – Comment on Roberts, Zeidner, & Matthews (2001). Emotion 1, 243–248.

Scherer, K.R., 2007. Componential emotion theory can inform models of emotional competence. In: Matthews, G., Zeidner, M., Roberts, R.D. (Eds.), The Science of Emotional Intelligence: Knowns and Unknowns. Oxford University Press, Oxford, pp. 101–126.

Schneider, R.J., Ackerman, P.L., Kanfer, R., 1996. To "act wisely in human relations": exploring the dimensions of social competence. Personality and Individual Differences 4, 469–481.

Süß, H.-M., 2001. Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich [The predictive validity of intelligence in the scholastic and non-scholastic domain]. In: Stern, E., Guthke, J. (Eds.), Perspektiven der Intelligenzforschung [Perspectives of Intelligence Research]. Pabst Science Publisher, Lengerich, pp. 109–136.

Süß, H.-M., Weis, S., Seidel, K., 2005. Soziale Kompetenzen [Social Competences]. In: Weber, H., Rammsayer, T. (Eds.), Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie, (Reihe Handbuch der Psychologie) [Handbook of Personality Psychology and Differential Psychology]. Hogrefe, Göttingen, pp. 350–362.

Walker, R.E., Foley, J.M., 1973. Social intelligence: its history and measurement. Psychological Reports 33, 839–864.

Weber, H., Westmeyer, H., 2001. Die Inflation der Intelligenzen [The inflation of intelligences]. In: Stern, E., Guthke, J. (Eds.), Perspektiven der Intelligenzforschung [Perspectives of Intelligence Research]. Pabst Science Publisher, Lengerich, pp. 251–266.

Weis, S., Süß, H.-M., 2005. Social intelligence – a review and critical discussion of measurement concepts. In: Schulze, R., Roberts, R.D. (Eds.), An International Handbook of Emotional Intelligence. Hogrefe, Göttingen, pp. 203–230.

Weis, S., Süß, H.-M., 2007. Reviving the search for social intelligence. Personality and Individual Differences 42, 3–14.

Weis, S., Seidel, K., Süß, H.-M., 2006. Messkonzepte sozialer Intelligenz – Literaturübersicht und Ausblick [Measurement concepts of social intelligence – Review and future perspectives]. In: Schulze, R., Freund, P.A., Roberts, R.D. (Eds.), Emotionale Intelligenz. Ein Internationales Handbuch [Emotional Intelligence – An International Handbook]. Hogrefe, Göttingen, pp. 213–234.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., Sommer, W., 2010. Individual differences in perceiving and recognizing faces – one element of social cognition. Journal of Personality and Social Psychology 99, 530–548.

Wong, C.T., Day, J.D., Maxwell, S.E., Meara, N.M., 1995. A multitrait-multimethod study of academic and social intelligence in college students. Journal of Educational Psychology 87, 117–133.

Social Intervention in Psychiatric Disorders

Wulf Rössler, Leuphana University Lüneburg, Lüneburg, Germany; University of Sao Paulo, Sao Paulo, Brazil; and University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Serious mental illness can devastate the lives of people who suffer from it. These persons have a right to be treated according to current knowledge. Psychosocial interventions aim at reintegrating the mentally ill into the community and improving their quality of life. The goal of psychosocial intervention is to help disabled individuals to establish the emotional, social, and intellectual skills needed to live, learn, and work in the community with the least amount of professional support. The refinement of all these approaches has achieved a level where they should be made readily available for every person with severe mental disorders.

The objective of psychiatric reforms since the mid-1950s has been to resettle chronically mentally ill patients from large custodial institutions to community settings. Reform practice has changed the perception of mental illness. Enabling disabled persons to live a normal life in the community causes a shift away from a focus on an illness model toward a model of functional disability. Social role functioning including social relationships, work and leisure, as well as quality of life of the persons affected developed as a core target of psychosocial interventions. The overall philosophy of psychosocial interventions comprises two intervention strategies: the first psychosocial strategy is individual-centered and aims at developing the patient's skill in interacting with a stressful environment. The second psychosocial strategy is an ecological approach directed toward developing environmental resources to reduce potential stressors. Most disabled persons need a combination of both approaches. On more specific individual-centered levels, social interventions encompass social skills training, family intervention, and vocational training. Reducing potential stressors in the environment include the provision of sheltered living accommodation and work. Also financial resources have to be secured to meet basic needs. The article explains in detail the theoretical background as well as several psychosocial intervention strategies.

As early as the nineteenth century psychiatrists already suspected that harmful influences from the social environment might promote the appearance of mental illnesses or have a deleterious effect on their course. In that era, the isolation of the mentally ill in the 'healthful' milieu of institutions far removed from their original living environment was thought to be an appropriate means of shielding the patients from the disease-producing influences of their surroundings. The objective of psychiatric reforms since the mid- to late 1950s was to resettle chronically mentally ill persons from large custodial institutions to community settings. The idea behind this was to (re-)integrate patients in their normal living environment as rapidly as possible, because it was thought that their natural social network would make a major supporting contribution to the healing process (Rössler, 2001).

This reform has changed the perception of mental illness. Enabling disabled persons to live a normal life in the community caused a shift away from a focus on an illness model toward a model of functional disability (Grove, 1994). As such, other outcome measures aside from clinical conditions became relevant. Since then social role functioning including social relationships, work, and leisure, as well as quality of life and the family burden, became of major interest for affected individuals living in the community as well as for the health professionals involved in treatment and care of those persons (Rössler, 2000).

In the current scientific discourse, environmental and psychosocial factors are often relegated to the status of a single variable of merely peripheral interest, without further specification. Biological science has indeed made enormous progress toward an understanding of mental disorders, but it should not be forgotten that research in social sciences also yielded a considerable body of knowledge with respect to the psychosocial and environmental factors that influence it (Rössler, 2001).

The relevance of psychosocial and environmental problems is also reflected in the DSM IV and ICD-10. Axis 4 of the DSM (American Psychiatric Association, 1994) and codes Z 55–Z 65 and Z 73 of the ICD-10 (WHO, 1993) are assigned for reporting psychosocial and environmental problems that may affect the diagnosis, treatment, and prognosis of mental disorders. The importance of role functioning when it comes to 'clinical judgment' is also emphasized in the newly launched DSM-5 (Maj, 2013).

Target Population of Psychosocial Interventions

Although the majority of the chronically mentally ill have the diagnosis of schizophrenic disorders, other patient groups with psychotic and nonpsychotic disorders are also targeted by psychosocial interventions. All patients suffering from severe mental illness (SMI) now require psychosocial interventions. The core group is drawn from patients with:

- persistent psychopathology
- marked instability characterized by frequent relapses, and
- social maladaptation (Royal College of Psychiatrists, 1996).

There are other definitions currently used to characterize the chronically mentally ill, but they all share some common

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27067-0

elements, i.e., a diagnosis of mental illness, prolonged duration, and role incapacity.

Furthermore, up to 50% of persons with SMI carry dual diagnoses, especially in combination with substance abuse (Cuffel, 1996). Young adult chronic patients constitute an additional category that is diagnostically more complicated (Schwartz et al., 1983). These patients present complex pictures of symptomatology difficult to categorize within our diagnostic and classification systems. Many of them also have a history of suicide attempts. All in all, they represent a patient population that is mostly difficult to treat.

Models of Vulnerability

There are several vulnerability–stress models forming the theoretical basis into which biological and social intervention approaches of psychiatric treatment, rehabilitation, and care can be integrated. These models were developed mainly for persons with schizophrenic disorders, but they apply essentially to all individuals with SMI.

In the *vulnerability–stress* model it is presumed that socio-environmental stressors superimposed on an underlying and enduring biological vulnerability lead to abnormalities in central nervous system functioning (Zubin and Spring, 1977). In the *vulnerability–stress–competence* model (Nuechterlein and Dawson, 1984), individuals, families, and natural support systems can cope with stressors and have the competence to create a higher threshold for symptomatic illness. In an *interactive–developmental* model (Strauss et al., 1981), the persons affected do not remain in a passive role but actively interact with their environment. They modulate and select the kind of help they get from other people, they choose to comply with the treatment, and they seek or avoid social support and stressful situations.

The interactive–developmental model is the basis for the consumer movement in mental health care. People with mental disorders and their caregivers prefer to see themselves as consumers of mental health services with an active interest in learning about psychiatric disorders and in selecting the respective treatment approaches. Consumerism makes it possible to take up the affected person's perspective and to consider seriously courses of action relevant to them (Kopelowicz and Liberman, 1995).

Conceptual Framework of Psychosocial Interventions

The overall philosophy of social interventions in psychiatric disorders comprises two intervention strategies: the first strategy is individual-centered and aims at developing the patient's skill in interacting with a stressful environment. There is no consensus among researchers as to what this individual-centered approach actually does accomplish. Some understand developing patient's skill as an approach to help disabled persons to compensate for impairments and to function optimally with the deficiencies they have. Others assume that different forms of individual-centered interventions help with recovery from the disorder itself, while the contributing factors to the healing process are not clear (Strauss, 1986).

The second strategy is an ecological approach directed toward developing environmental resources to reduce potential stressors. This approach includes psychosocial interventions concerning housing and occupation as well as measures to build natural networks, as many chronically mentally ill and disabled persons lose close and stable relationships in the course of the disease. Most disabled persons need a combination of both approaches.

In contrast to acute treatment there are mostly no legal powers to enforce psychosocial interventions in long-term care. Thus, the sufferer's autonomy concerning treatment decisions should be respected. Within this frame, therapeutic alliance plays a crucially important role in engaging the affected persons in their own care planning. As such, the therapist should acknowledge that disagreement about the illness between patient and therapist is not always the result of the illness process (Eichenberger and Rössler, 2000).

The International Classification of Functioning, Disability and Health

The long-term consequences of major mental disorders targeted by psychosocial interventions might be described on different levels. A useful framework is provided by the 'International Classification of Impairment, Disability, and Handicaps' (WHO, 1980) and by the revised version of this classification, i.e., the 'International Classification of Functioning, Disability and Health' (WHO, 2001). The revised International Classification of Functioning, Disability and Health (ICF) (4) includes a change from negative descriptions of impairments, disabilities and handicaps to neutral descriptions of body structure and function, activities and participation. A further change has been the inclusion of a section on environmental factors as part of the classification. This is in recognition of the importance of the role of environmental factors in either facilitating functioning or creating barriers for people with disabilities. Environmental factors interact with a given health condition to create a disability or restore functioning, depending on whether the environmental factor is a facilitator or a barrier.

The ICF is a useful tool to comprehend mentally ill in all their dimensions, including impairments at the structural or functional level of the body, at the person level with respect to activity limitations, and at the societal level with respect to restrictions of participation. Each level encompasses a theoretical foundation on which a psychosocial intervention can be formulated.

In addition to acute symptoms, chronically mentally ill patients have impairments in the most basic cognitive processes concerning information input, memory, and abstraction. These cognitive impairments are present even when all symptoms have subsided, and they seem to be enduring markers of a vulnerability to the illness. Persistent and severe impairments lead to persisting limitations, i.e., affected persons will be limited in performing certain roles in their social and occupational environment. Finally, restrictions in participation occur when impairments and persisting limitations place the individual at a disadvantage relative to others in society.

On more specified levels concerning impairments, medication and cognitive training are provided. Effective pharmacotherapy and cognitive training stabilize affected persons so that they are more responsive to learning from their environment. Interventions concerning persisting limitations encompass mainly social skills training and family intervention. Furthermore, social interventions emphasize vocational training suitable to the extent of the handicap.

Reduction of Impairments

Medication

The interaction of medication and psychosocial interventions appear to be more than merely additive, as each can enhance the effects of the other. Some studies suggest that these agents improve the participation of patients on psychosocial interventions (Marder, 2000).

But even when taken regularly, medication is only able to reduce rather than abolish symptoms, and delay rather than prevent relapse. In addition to that, many of the traditional drugs have adverse effects and can weaken patients' ability to perform their social roles or can impair their social and vocational functioning. As such, it is not surprising that noncompliance with medication is one of the most serious problems in the long-term care of persons with SMI.

Several treatment strategies have been developed for improving compliance with medication. For patients with cognitive deficits, the complexity of the medication regimen should be reduced (Gaebel, 1997). Some patients need to be given the medication by a professional service on a regular basis. Depot medication (i.e., intramuscular injections that release the medication slowly into the bloodstream) might be especially useful for patients with fluctuating motivation or insight. It is also useful to involve families and caregivers in increasing compliance with medication.

Yet, many patients living in the community have to take responsibility for their medication themselves. Training in self-management of medication (Eckman et al., 1990) emphasizes the patient's autonomy and increases acceptance of and responsibility for treatment. This also includes variation of medication without consultation within certain limits.

Cognitive Training

With respect to cognitive deficits, consistent results emerged from most studies in so far as remediation of specific deficits through repeated practice or related techniques is successful. Current data suggest that it is effective when provided in a broader context of psychosocial interventions (Medalia and Saperstein, 2013).

One of the most promising comprehensive programs in this regard was developed by Brenner et al. (1992) since the 1980s. They called this program 'integrated psychological therapy' (IPT).

Over the past 30 years, research groups in 12 countries have evaluated IPT in 36 independent studies. Patients undergoing IPT showed significantly greater improvement in all outcome variables than control patients (Roder et al., 2011).

IPT comprises five stages: cognitive differentiation, social perception, communication skills, interpersonal problem solving, and social skills training. It is presented to groups of five to seven patients for 30–60 min two or three times a week. The hierarchically arranged program is aimed at facilitating acquisition and maintenance of more complex skills by remediating cognitive deficits.

Remediation of Persisting Limitations

Skills Training

In recent years, social skills training in psychiatric rehabilitation has become very popular and has been disseminated widely. The most prominent proponent of skills training is Robert Liberman, who has designed systematic and structured skills training since the mid-1970s. Liberman and his colleagues packaged the training in the form of modules which focus on different topics such as medication management, symptom management, substance abuse management, basic conversational skills, interpersonal problem solving, friendship and intimacy, recreation and leisure, workplace fundamentals, community (re-)entry, and family involvement. Each module is composed of skill areas. The symptom management module, for example, encompasses skills areas such as identifying warning signs of relapse, developing a relapse prevention plan, coping with persistent symptoms, and avoiding alcohol and drugs. The skills areas are taught in exercises with demonstration videos, role play and problem-solving exercises, and *in vivo* and homework assignments (Liberman, 1988).

The results of several controlled studies suggest that disabled individuals can be taught a wide range of social skills. Social and community functioning improve when the trained skills are relevant for the affected person's daily life, and the environment perceives and reinforces the changed behavior. Unlike medication effects, benefits from skills training occur more slowly. Furthermore, long-term training has to be provided for positive effects (Kurtz and Mueser, 2008).

Family Intervention

As a consequence of deinstitutionalization, the burden of care has fallen increasingly on the relatives of the mentally ill. It is estimated that up to 65% of mentally disabled individuals live with their families (Goldmann et al., 1981). This is a task many families do not choose voluntarily. Additionally, not all families are equally capable of giving full support to their disabled member, and may not be willing to replace an insufficient health care system (Johnson, 1990). But families also represent support systems, which provide natural settings for context-dependent learning important for recovery of functioning (Schooler, 1995). As such there has been a growing interest in helping families since the beginnings of care reforms (Strachan, 1986).

One area of interest dealt with the expectations of relatives concerning the provision of care. Relatives quite often feel ignored, not taken seriously, and insufficiently informed by health professionals. They also may feel that their contribution to care is not appreciated, or that they will be blamed for the patient's problems. It is certainly not surprising that there is a lot of frustration and resentment among relatives related to the extent of the physical, financial, and emotional family burden. As important research has demonstrated a relationship between a negative affective family atmosphere and an unfavorable course of a mental disorder, the need for helping

families and patients with family intervention programs became clear.

All family intervention programs include an educational aspect covering basic information on the etiology, treatment, and prognosis of schizophrenia. Although family interventions differ in the treatment focus, all family interventions are organized around the principal goal of providing family members with more information about the disorder and strategies for managing common problems. There are more common features: family interventions are concrete, practical, and focus on everyday coping.

Family intervention programs have produced the most promising results. Family intervention is effective for improving outcome (e.g., social functioning) (Koukia and Madianos, 2005). Possibly, family intervention can reduce family burden. Furthermore, the treatment gains are fairly stable (Pilling et al., 2002).

However, it is not clear what the effective components of the different models are. Additionally, family interventions differ in frequency and length of treatment. There are also no criteria for the minimum amount of treatment necessary (Barbato and D'Avanzo, 2000).

Finally, it should be noted that most of the family interventions were developed in the context of Western societies during deinstitutionalization. Family caregiving might be quite different in a different cultural context. This refers to other cultures in total as well as to minority groups in Western societies (Guarnaccia, 1998).

Enhancing Participation by Vocational Rehabilitation

The beneficial effects of work for mental health have been known for centuries (Harding et al., 1987). Vocational rehabilitation therefore has been a core element of psychiatric reforms since its beginnings.

Vocational rehabilitation is based on the assumption that work not only improves activity, social contacts, etc., but may also promote gains in related areas such as self-esteem and quality of life, as work and employment are a step away from dependency toward integration into society. Enhanced self-esteem in turn improves adherence to long-term care of individuals with impaired insight (McElroy, 1987).

Vocational rehabilitation originated in psychiatric institutions, where the lack of activity and stimulation led to apathy and withdrawal among their inpatients. Long before the introduction of medication, occupational and work therapy contributed to sustainable improvements in long-stay inpatients. Today occupational and work therapy are no longer hospital-based, but represent the starting point for a wide variety of rehabilitative techniques teaching vocational skills (Royal College of Psychiatrists/College of Occupational Therapists, 1992).

Vocational rehabilitation programs in the community provide a series of graded steps to promote job (re-)entry. For less disabled persons, brief and focused techniques are used to teach them how to find a job, fill out applications, and conduct employment interviews (Jacobs et al., 1988). In transitional employment a temporary work environment is provided to teach vocational skills that should enable the affected person to move on to competitive employment. But too often the gap between transitional and competitive employment is so large that the mentally disabled individuals remain in a temporary work environment. Sheltered workshops providing prevocational training also quite often prove to be a dead end for disabled persons.

One result of the difficulties of integrating mentally disabled individuals into the common labor market has been the steady growth of cooperatives which operate commercially with disabled and nondisabled staff working together on equal terms and sharing in management. The mental health professionals work in the background providing support and expertise (Grove, 1994). Cooperatives serve as a bridge between sheltered workshops and competitive jobs on the common labor market.

Today, the most promising vocational rehabilitation model is supported employment. In this model, disabled persons are placed in competitive employment as soon as possible and receive all the support needed to maintain their position. Skills are taught in the actual environment, tailored to the job and the individual. If necessary, additional advice is given to the employer. As the course of SMI is unpredictable and job requirements often change, the support provided is neither time-limited nor delivered on a fixed schedule (Wallace, 1998). Participation in supported employment programs is followed by an increase in the ability to find and keep employment (Burns et al., 2007). Today there is sufficient evidence of work having beneficial effects on clinical and social functioning (Catty et al., 2008; Burns et al., 2009; Kilian et al., 2012).

Although findings regarding supported employment are encouraging, some critical issues remain unanswered. Most individuals in supported employment obtain unskilled part-time jobs. Since most studies only evaluated short (12–18 months) follow-up periods, the long-term impact remains unclear.

It must be acknowledged that the integration into the labor market by no means depends exclusively on the ability of the persons affected to fulfill a work role and on the provision of sophisticated vocational training and support techniques, but also on the willingness of society to integrate its most disabled members.

Developing Environmental Resources

Effective community psychiatry requires individualized and specialized treatment, which has to be embedded in a comprehensive and coordinated system of community services. But even when a variety of services is available, they are poorly linked in many cases, and costly duplication may occur.

While developing community support systems it became obvious that there was a need to coordinate and integrate the services provided as each professional involved concentrates on different aspects of the same patient. Therefore, as a key coordinating and integrating mechanism, the concept of case management (CM) originated. CM focuses on all aspects of the physical and social environment. The core elements of CM are the assessment of patient needs, the development of comprehensive service plans for the patients, and arrangement of service delivery (Rössler et al., 1992).

Since the 1980s, a range of different models of CM have been developed which exceed the original idea that CM intends mainly to link the patient to needed services, and to coordinate those services. Today most clinical case managers

also provide direct services in the patient's natural environment. This model is called intensive case management (ICM). ICM is difficult to distinguish from assertive community treatment (ACT).

Stein and Test developed the basic compounds of ACT in the 1970s (Stein and Test, 1980). The original program was designed as a community-based alternative to hospital treatment for persons with severe mental illnesses. A comprehensive range of treatment, rehabilitation, and support services in the community is provided through a multidisciplinary team. ACT is characterized by an assertive outreach approach, i.e., interventions are provided mainly in the natural environment of the disabled individuals (Scott and Dixon, 1995). Research on CM and ACT indicates that these models can reduce time in hospital, but have only small or moderate effects on improving symptomatology and social functioning (Mueser et al., 1998). The debate over the effectiveness and efficiency of CM and ACT is on-going and subject of further trials (e.g., Bramesfeld et al., 2013).

Conclusions

The goal of psychosocial intervention is to help disabled individuals to establish the emotional, social, and intellectual skills needed to live, learn, and work in the community with the least amount of professional support (Anthony, 1979). The refinement of all these approaches as described has achieved a level where they should be made readily available for every person with severe mental disorders.

But we have to be aware that it is a long way to translate research into practice. In 1992, the National Institute of Mental Health, for example, funded the 'schizophrenia patient outcome research team' (PORT) to develop and disseminate recommendations for the treatment of schizophrenia based on existing scientific evidence. The recommendations and further updates address in essence all psychosocial interventions as presented above (Lehman and Steinwachs, 1998; Dixon et al., 2010). Lehman and colleagues assessed the patterns of usual care for schizophrenic patients and examined the conformance rate with the PORT treatment recommendations. The conformance rate was modest. As such, it is obvious that current care practice has to be improved substantially in the light of the outcome research available (Dixon et al., 2010).

Serious mental illness can devastate the lives of people who suffer from it. These persons have a right to be treated according to current knowledge. But mental illnesses are associated with a significant amount of stigma, resulting in social isolation and discrimination in housing, education, and employment opportunities. Mentally ill people are even disadvantaged in receiving appropriate treatment and care, so stigma and discrimination further increase the burden on affected persons and their relatives. Yet, social interventions aim at reintegrating the mentally ill into the community and improving their quality of life. For these reasons a responsive and open-minded environment is indispensable. Therefore, translating research into practice always encompasses combating stigma and discrimination of mental illnesses as well (World Psychiatric Association, 1999).

See also: Behavior Therapy: Psychiatric Aspects; Cognitive Behavioral Therapy; Depression; Education in Psychiatry; Family and Psychiatric Illness: Family Therapy.

Bibliography

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders (DSM IV), fourth ed. American Psychiatric Association, Washington, DC.

Anthony, C.W., 1979. The Principles of Psychiatric Rehabilitation. University Park Press, Baltimore.

Barbato, A., D'Avanzo, B., 2000. Family interventions in schizophrenia and related disorders: a critical review of clinical trials. Acta Psychiatrica Scandinavica 102, 81–97.

Baronet, A.M., Gerber, G., 1998. Psychiatric rehabilitation: efficacy of four models. Clinical Psychology Review 18, 189–228.

Brenner, H.D., Hodel, B., Genner, R., Roder, V., Corrigan, P.W., 1992. Biological and cognitive vulnerability factors in schizophrenia: implications for treatment. British Journal of Psychiatry 161 (Suppl. 18), 154–163.

Bramesfeld, A., Moock, J., Kopke, K., et al., February 15, 2013. Effectiveness and efficiency of assertive outreach for schizophrenia in Germany: study protocol on a pragmatic quasi-experimental controlled trial. BMC Psychiatry 13. http://dx.doi.org/10.1186/1471-244X-13-56.

Burns, T., Catty, J., Becker, T., et al., September 29, 2007. The effectiveness of supported employment for people with severe mental illness: a randomised controlled trial. Lancet 370 (9593), 1146–1152.

Burns, T., Catty, J., White, S., et al., 2009. The impact of supported employment and working on clinical and social functioning: results of an international study of individual placement and support. Schizophrenia Bulletin 35 (5), 949–958.

Catty, J., Lissouba, P., White, S., et al., 2008. Predictors of employment for people with severe mental illness: results of an international six-centre randomised controlled trial. British Journal of Psychiatry 192 (3), 224–231.

Cuffel, B., 1996. Comorbid substances use disorder: prevalence, patterns of use, and course. In: Drake, R.E., Mueser, K.T. (Eds.), Dual Diagnosis of Major Mental Illness and Substance Abuse: Recent Research and Clinical Implications, vol. 2. Jossey-Bass, San Francisco, pp. 93–105.

Dixon, L.B., Dickerson, F., Bellack, A.S., et al., 2010. The 2009 schizophrenia PORT psychosocial treatment recommendations and summary statements. Schizophrenia Bulletin 36 (1), 48–70.

Eckman, T., Liberman, R., Phipps, C., Blair, K., 1990. Teaching medication management skills to schizophrenic patients. Journal of Clinical Psychopharmacology 10, 33–38.

Eichenberger, A., Rössler, W., 2000. Comparison of self-ratings and therapist ratings of outpatients' psychosocial status. Journal of Nervous and Mental Disease 188, 297–300.

Gaebel, W., 1997. Towards the improvement of compliance: the significance of psychoeducation and new antipsychotic drugs. International Clinical Psychopharmacology 12 (Suppl. 1), 37–42.

Goldmann, H.H., Gattozzi, A.A., Taube, C.A., 1981. Defining and counting the chronically mentally ill. Hospital and Community Psychiatry 32, 21–27.

Grove, B., 1994. Reform of mental health care in Europe. Progress and change in the last decade. British Journal of Psychiatry 165, 431–433.

Guarnaccia, P.J., 1998. Multicultural experiences of family caregiving: a study of African American, European American, and Hispanic American families. New Directions for Mental Health Services 77, 45–61.

Harding, C., Strauss, J., Hafez, H., Lieberman, P., 1987. Work and mental illness: I. Toward an integration of the rehabilitation process. Journal of Nervous and Mental Disease 175, 317–326.

Jacobs, H.E., Kardashian, S., Kreinbring, R.K., Ponder, R., Simpson, A., 1988. A skills-oriented model for facilitating employment in psychiatrically disabled persons. Rehabilitation Counseling Bulletin 27, 87–96.

Johnson, D., 1990. The family's experience of living with mental illness. In: Lefley, H., Johnson, D. (Eds.), Families as Allies in Treatment of the Mentally Ill: New Directions for Mental Health Professionals. American Psychiatric Press, Washington, DC, pp. 31–63.

Kilian, R., Lauber, C., Kalkan, R., et al., 2012. The relationships between employment, clinical status, and psychiatric hospitalisation in patients with schizophrenia receiving either IPS or a conventional vocational rehabilitation programme. Social Psychiatry and Psychiatric Epidemiology 47 (9), 1381–1389.

Kopelowicz, A., Liberman, R.P., 1995. Biobehavioral treatment and rehabilitation of schizophrenia. Harvard Review of Psychiatry 3, 55–64.

Koukia, E., Madianos, M.G., 2005. Is psychosocial rehabilitation of schizophrenic patients preventing family burden? A comparative study. Journal of Psychiatric and Mental Health Nursing 12, 415–422.

Kurtz, M.M., Mueser, K.T., 2008. A meta-analysis of controlled research on social skills training for schizophrenia. Journal of Consulting and Clinical Psychology 76, 491–504.

Lehman, A.F., Steinwachs, D.M., 1998. Survey co-investigators of the PORT project. At issue: translating research into practice: the schizophrenia patient outcomes research team (PORT) treatment recommendations; and patterns of usual care for schizophrenia: initial results from the schizophrenia patient outcomes research team (PORT) client survey. Schizophrenia Bulletin 24, 1–10, 11–20.

Liberman, R. (Ed.), 1988. Psychiatric Rehabilitation of Chronic Mental Patients. American Psychiatric Press, Washington, DC.

Marder, S.R., 2000. Integrating pharmacological and psychosocial treatment for schizophrenia. Acta Psychiatrica Scandinavica Supplementum 407, 87–90.

McElroy, E., 1987. Sources of distress among families of the hospitalized mentally ill. New Directions for Mental Health Services 34, 61–72.

Maj, M., 2013. "Clinical judgment" and the DSM-5 diagnosis of major depression. World Psychiatry 12, 89–92.

Medalia, A., Saperstein, A.M., 2013. Does cognitive remediation for schizophrenia improve functional outcomes? Current Opinion in Psychiatry 26 (2), 151–157.

Mueser, K.T., Bond, G.R., Drake, R.E., Resnick, S.G., 1998. Models of community care for severe mental illness: a review of research on case management. Schizophrenia Bulletin 24, 37–74.

Nuechterlein, K.H., Dawson, M.E., 1984. A heuristic vulnerability/stress model of schizophrenic episodes. Schizophrenia Bulletin 10 (2), 300–312.

Pilling, S., Bebbington, P., Kuipers, E., et al., 2002. Psychological treatments in schizophrenia: I. Meta-analysis of family intervention and cognitive behaviour therapy. Psychologie Medicale 32, 763–782.

Roder, V., Mueller, D.R., Schmidt, S.J., 2011. Effectiveness of integrated psychological therapy (IPT) for schizophrenia patients: a research update. Schizophrenia Bulletin 37 (Suppl. 2), S71–S79.

Rössler, W., 2000. Rehabilitation techniques. In: Gelder, M.G., Lopez-Ibor, J.J. (Eds.), New Oxford Textbook of Psychiatry. Oxford University Press, Oxford, pp. 1141–1195.

Rössler, W., 2001. Schizophrenia: psychosocial factors. In: Henn, F., Sartorius, N., Helmchen, H., Lauter, H. (Eds.), Contemporary Psychiatry, vol. 3, Part 1. Springer, Berlin, pp. 121–128.

Rössler, W., Fätkenheuer, B., Löffler, W., et al., 1992. Does case management reduce the rehospitalization rate? Acta Psychiatrica Scandinavica 86, 445–449.

Royal College of Psychiatrists, 1996. Psychiatric Rehabilitation, rev. ed. Gaskell, London.

Royal College of Psychiatrists/College of Occupational Therapists, 1992. Occupational Therapy and Mental Disorders. A Consensus Statement by the Royal College of Psychiatrists and the College of Occupational Therapists. Royal College of Psychiatrists, London.

Schooler, N., 1995. Integration of family and drug treatment strategies in the treatment of schizophrenia: a selective review. International Clinical Psychopharmacology 10 (Suppl. 3), 73–80.

Schwartz, S., Goldfinger, S., Ratener, M., Cutler, D., 1983. The young adult patient and the care system: fragmentation prototypes. New Directions for Mental Health Services 19, 23–36.

Scott, J.E., Dixon, L.B., 1995. Assertive community treatment and case management for schizophrenia. Schizophrenia Bulletin 21, 657–668.

Stein, L.I., Test, M.A., 1980. Alternative to mental hospital treatment: I. Conceptual model, treatment program, and clinical evaluation. Archives of General Psychiatry 37, 392–397.

Strachan, A., 1986. Family intervention for the rehabilitation of schizophrenia: toward protection and coping. Schizophrenia Bulletin 4, 678–698.

Strauss, J.S., 1986. What does rehabilitation accomplish? Schizophrenia Bulletin 12, 720–723.

Strauss, J., Bartko, J., Carpenter Jr., W.T., 1981. New directions in diagnosis: the longitudinal processes of schizophrenia. American Journal of Psychiatry 138, 954–958.

Wallace, C.J., 1998. Social skills training in psychiatric rehabilitation: recent findings. International Review of Psychiatry 19, 9–19.

World Health Organization (WHO), 1980. International Classification of Impairments, Disabilities and Handicaps. WHO, Geneva.

World Health Organization (WHO), 1993. International Statistical Classification of Diseases and Related Health Problems: Tenth Revision. WHO, Geneva.

World Health Organization (WHO), 2001. International Classification of Functioning, Disability and Health. WHO, Geneva.

World Psychiatric Association, 1999. The WPA Global Programme against Stigma and Discrimination because of Schizophrenia. WPA, Geneva.

Zubin, J., Spring, B., 1977. Vulnerability – a new view of schizophrenia. Journal of Abnormal Psychology 86, 103–126.

Social Justice and Criminal Justice

Kimberly L Barrett, Eastern Michigan University, Ypsilanti, MI, USA
Michael J Lynch, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes and explores what is meant by the term social justice. Emphasis is given to social justice issues that pertain directly or indirectly to the study of crime and justice. To accomplish this, general definitions of social justice are described, and the origins of the term social justice are briefly discussed. Next, a series of topics related to the study and practice of social justice are presented and explored in turn. These topics include, though are not limited to, peacemaking criminology, restorative justice, victim advocacy, environmental justice, and prison abolition.

Introduction

Social justice is a broad concept, and has been used in a variety of ways to refer to subjects of interest to criminologists. In this article, we examine the concept 'social justice' and the various ways in which social justice concepts can and have been applied to the study of criminal justice and other related process, such as social justice in law and in the application of law.

To illustrate these different uses of social justice relevant concepts, we examine several areas in which social justice orientations have been employed to illustrate how the primary concern of social justice approaches and measures of equity and fairness can be produced, and the issues that need to be addressed to facilitate social justice outcomes. To accomplish this task, we first review general definitions of social justice. Next, we examine social justice-related literature related to several issues including: peacemaking criminology; restorative justice; victim advocacy; mental health; the prison abolition movement; environmental justice; social work; and public health.

In undertaking this review, it is important for the reader to bear in mind that in contemporary academic literature, the concept of social justice has been extended to a number of fields. Traditionally, criminologists associate the concept of social justice with the work of radical criminologists in the late 1960s and 1970s. The field of social justice research, however, has come a long way since then, and social justice approaches are applied in many different disciplines. To illustrate that point, our discussion is not confined to criminological applications of social justice ideas alone. Indeed, while social justice approaches are certainly important within specific areas of criminological research (e.g., peacemaking criminology; restorative justice; victim advocacy; etc.), important literatures on this topic have emerged on issues, such as environmental justice and equity and public health that should not be ignored because of the ways in which they illustrated the expanded concept of social justice.

Defining Social Justice and the History of Social Justice Research

The idea of social justice can be defined in broad or general terms, and in more restrictive or relative terms. The general definition applies at an abstract level, while relative definitions make the application of social justice concepts clear with respect to given applications of social justice principles.

To begin with the general definition, we can say that the idea of social justice is, in the first place, dependent upon an assumption that equality is a valued social norm. From this it follows that societies are just when they are organized and arranged in ways that facilitate equality, and unjust when they hamper equality. In this general sense, then, we can say that a society takes on social justice values when it is organized in ways that allow its members to achieve, economic, political and social equality. Other broad views of social justice imply that economic, political, and social equality are not simply characteristics of society, but rights of people. That is to say, all people have a right to economic, political, and social equality, and therefore, societies that are organized in ways that impede those forms of equality are unjust societies or societies in which social injustice is evident.

In this broad sense, social justice societies value not only equality, but larger values such as basic human rights all people should be able to obtain. Sometimes, these human rights are defined as those that allow all individuals within a society to live in dignity, and to experience respect.

Historically, this broad concept of social justice can be traced to the work of the Catholic religious scholar, Luigi Taparelli D'Azeglio in the 1840s (Novak, 2000). As a doctrine, the principles of social justice became an official part of Roman Catholic Church teaching beginning in 1931, and thus marked the first extensive evidence of a system of thought committed specifically to establishing conditions for social justice (Novak, 2000). This Catholic social justice tradition has had significant influence in South America, and was employed to promote the development of liberation theology by the Peruvian Roman Catholic priest, Gustavo Gutierrez in 1973 (Berryman, 1987). The religious origins of this social justice emphasis are quite strong, and it was not until the 1960s and 1970s that the concept of social justice began to filter into secular social movements and academic writings.

On the latter point, radical criminologists in the late 1960s and early 1970s, such as Tony Platt, Richard Quinney, Paul Takagi, and Herman and Julia Schwendinger were among the first criminologists to address social justice issues and to illustrate their importance. This was truly a unique approach since

their work was among the first work on the applications of social justice perspectives in any discipline. These individuals were among the group, who formed the journal *Social Justice* (previously called *Crime and Social Justice*), which published its first volume in 1974. These early radicals took a decidedly class-based theoretical approach to social justice, and analyzed the ways in which the economic structure in class societies prevents the achievement of social justice. Over the past four decades, this radical concept of social justice in relation to criminology and criminal justice has expanded in a variety of ways, and has been addressed in relation to peacemaking (Pepinsky and Quinney, 1991; Quinney and Wildeman, 1991), postmodern criminology (Milovanovic, 1997), critical criminology (Arrigo, 2000, 2001), and distributive (Spader, 1988), and restorative justice approaches (Arrigo and Schehr, 1998).

In the narrower sense, social justice concepts and approaches have been applied to specific issues, several of which we review below. These include, for example, discussions of social justice in relation to public health access and the distribution of medical services, with respect to environmental justice or distributive justice issues related to the distribution of toxic hazards and environmental penalties across communities based on their class, race, and ethnic characteristics, and the achievement of social justice for particular groups, such as the mentally ill, prison inmates, school children, women, etc. In this sense, the narrow applications and definitions of social justice are simply illustrations of how general social justice principles apply to specific affected groups that appear to suffer social justice deficits.

With respect to academic analysis of social justice issues and applications of social justice theories, Tyler et al. (1997) argue that four distinct waves of social justice research can be identified. The first wave of research addressed social justice issues from the perspective of relative deprivation theory. This research emphasis drew attention to the feeling of being relatively deprived in comparison to others. In this sense, the relative deprivation approach was not about the absolute qualities of social justice, or the theoretical identification of social justice, but instead focused on the psychological identification of social justice people make when comparing themselves to others. In that view, relative deprivation is a feeling or sense of injustice that stems from feeling or perceptions that one is deprived relative to others with respect to the unequal distributions of social and economic goods, resources, and services. Such feelings can exist at the individual level or at the social level as an indicator of similar psychological orientations felt by a group of like situated individuals.

The second wave of research attended to distributive justice concerns. Distributive justice examines how resources are distributed in social environments, and whether the distribution of those resources is equitable. The main issue here, then, is whether social and economic resources are distributed fairly in reference to some assumption concerning how resources ought to be distributed. Thus, the determination of equity in resource distribution is based on an assumption concerning the defining of fair and equitable. In the broadest sense, one might assume that resources should be fairly or evenly distributed across populations regardless of the personal or group characteristics of individuals. Since the question is one of distribution, the focus is on the end result rather than the process of distribution, though clearly, if the end result is unequal distribution, the distributive process itself is called into question.

During the third wave of research, the focus was narrowed to issues related to procedural justice. Here, the issue is not the end result or the fairness of distributions, but rather the focus is on the process used to make distributional decisions, and whether the process itself is fair and equitable. In this view, inequality in outcomes can be linked to inequality in the distributive process and in access to distributive mechanisms including the law.

The fourth wave drew attention to what Tyler et al. (1997) defined as redistributive justice. Here, we make a distinction between two forms of redistributive justice. In the narrower sense as used by Tyler et al. (1997), redistributive just relates to examining reactions to the breaking of law, a conceptualization of redistributive justice relevant to social psychology and criminology. In other fields, however, redistributive justice is among the broadest forms of social justice, and perhaps one with the longest history. In those areas, redistributive justice is linked to the establishment of societies based wholly in equity principles, and the redistribution, for example, of property is a key concern in socialist and Marxist theory. Since the late 1970s, this concept has also been linked to environmental justice research and movements, and to the observation that environmental laws ought to aid in redistributing environmental benefits and harms equitably.

In the sections that follow, we review social justice perspectives linked to several more specific though often overlapping approaches to illustrate the various ways in which social justice arguments have been applied.

Peacemaking and Social Justice

Peacemaking criminology, which developed from the work of Pepinksy and Quinney, is viewed as a critical element of critical criminology's orientation toward praxis (McEvoy and Gormally, 1997). Praxis is essentially the practice or policy implications, in this case, of critical theories of crime and justice. In the peacemaking view, social justice is achieved through the use of peacemaking policies, including those that can be applied to mediating and restoring the harms caused by crime. Peacemaking practices attempt to address the outcomes of social injustice and oppression. As a criminological approach, it begins with the observation that criminal justice practices are organized in ways that extend the forms of violence to which they supposedly respond, and facilitate the legal application of forms of oppression responsible for crime. In this sense, criminal justice is interpreted as an extension of the forms of injustice that already characterize society. As a form of injustice, crime is viewed as emerging from forms of suffering, such as racism, economic and class inequality, poverty, alienation, and so on, that are part of the organizing structure of a society. In contrast to dominant social values like competition and penal values, such as retribution, peacemaking argues that the pathway to social justice and human fulfillment is establishing conditions for peace and love. For many criminologists, thinking about the problem of crime in this way, particularly with respect to concepts such as peace and love, are off-putting

and beyond the scientific image criminology has long sought to build. In this sense, peacemaking criminology has often been criticized as subjective, utopian, and soft on crime (for criticisms, see Akers and Sellers, 2009). Nevertheless, with respect to issues of praxis, peacemaking criminology is committed to policies including but not limited to: abolition of the death penalty; universal health care; economic redistribution; antidiscriminatory wage policies; drug decriminalization; the replacement of adversarial criminal justice practices with peacemaking mediation (for further discussion, see Sullivan and Tifft, 1998).

Restorative Justice and Social Justice

Restorative justice is an approach of responding to crime that emphasizes repairing the harms caused by criminal acts by way of offender accountability, victim needs, and meaningful change at micro-, mezzo-, and/or macrolevels. This is accomplished by bringing together offenders, victims, and others impacted by criminal act(s) in an effort to cooperatively achieve a just response to a criminal harm. Including victims, offenders, and other community stakeholders in the response to crime, acts to empower the individuals most impacted by the criminal harm. The shared power dynamic illustrates the influence of social justice on restorative justice ideologies. Furthermore, restorative justice approaches promote empathy by design, and aim to be nonadversarial. Many criminologists are acquainted with the term restorative justice via the work of John Braithwaite, and his reintegrative shaming theory.

Restorative justice has inspired a number of policy initiatives, especially directives concerning juveniles and delinquency. For example, the Office of Juvenile Justice and Delinquency Prevention lists and describes five different examples of commonly employed juvenile justice programs associated with restorative justice (Office of Juvenile Justice and Delinquency Prevention, 2013). These initiatives include family group conferences, victim–impact panels, victim–offender mediation, circle sentencing, and community reparative boards. While each of these programs are inspired by the same underlying restorative justice principals, they vary with respect to implementation and practice. For example, family group conferences, victim–offender mediation, and circle sentencing traditionally bring together the victim and offender in person. Victim–impact panels bring together offenders, and persons who have been victimized by similar crimes as those carried out by the offender, or in some cases, secondary victims of the offender's crime are present. Community reparative boards bring together offenders and a trained board of community members who mutually determine sanctions.

In recent years, researchers and practitioners have suggested restorative justice initiatives as an alternative to more punitive or zero-tolerance school disciplinary policies (Morrison and Vaandering, 2012; Nickerson and Spears, 2007). Zero-tolerance school disciplinary policies refer to harsh, draconian school disciplinary practices inspired by the 1990s 'tough on crime' movement, and moral panic surrounding highly publicized school shooting incidents. These policies have been criticized as discriminatory, racist, and a contributing force to the recently termed "School to Prison Pipeline" (Kim et al., 2010). The use of restorative approaches to school discipline, including mental health services, counseling, and other reintegrative initiatives, have been advocated as preferable to suspensions, expulsions, and justice system referrals (Nickerson and Spears, 2007).

Victim Advocacy and Social Justice

Crime victims represent a particularly vulnerable population for a number of reasons. For example, crime victims may endure damaged or lost property, inclusive of wages or earnings. In other instances, crime victims may suffer bodily harm or injury that threatens or changes their health and productivity. Crime victims may suffer emotional trauma, or even the loss of a friend or loved one. Because the damages of crime are associated with a number of these life-altering consequences, crime victims may no longer have access to resources in the same way as they had prior to their victimization. As such, a number of government and non-for-profit agencies have emerged to protect and assist those who have been victimized by crime. In this way, these agencies act to obtain or preserve social justice for crime victims.

Victim advocacy has emerged as one avenue by which crime victims' rights and access to resources are protected. For example, state victim advocacy boards provide compensation to crime victims to help offset medical expenses incurred from their victimization, seek counseling services, or retain an attorney. Government victim advocacy agencies also issue grants to providers that serve crime victims. Non-for-profit entities also advocate for crime victims. For example, the Salvation Army operates a nation-wide domestic violence and rape crisis program, inclusive of a toll free hotline. By serving victims in this way, these programs aim to optimize the health and functioning of crime victims, as it were prior to their victimization. By working to offset the harms imposed by crime, these programs can be understood as protecting social justice for crime victims.

Environmental Justice and Social Justice

For many years, clean air, soil, and drinking water was considered a basic human right, and under the domain of the public commons. However, following the Industrial Revolution, this has changed drastically, with means of production posing very severe and lasting impacts on public health and environmental sustainability. Over time, this has ultimately led to an unequal distribution of environmental hazards – such as industrial sites, toxins like mercury or lead, polluted drinking water – worldwide, leaving some communities more vulnerable to environmental hazards. Likewise, the same environmental 'goods' that were once part of the public commons – clean air, water, soil – now also appear to follow skewed distributions. Unequal access to clean air, water, and soil is interpreted by many as a threat to social justice.

In recent years, researchers and environmental activists have become interested in the nonrandom distributions of environmental 'goods' and 'bads.' In particular, some have

been captivated by the demographic composition of areas surrounding environmental goods and hazard sites. This field has come to be known as environmental justice, and concerns social justice in so far as the sociological processes that determine the distribution of environmental hazards. A number of studies have established that communities that house areas designated by the Environmental Protection Agency as in need of clean up tend to be areas of lower socioeconomic status and also have higher concentrations of minority residents (Saha and Mohai, 2005). Moreover, studies have also demonstrated that communities of higher socioeconomic status and higher concentrations of white residents tend to house desirable environmental properties (Landry and Chakraborty, 2009; Lynch and Stretesky, 2013). For example, a study conducted by Landry and Chakraborty (2009) found that communities of predominantly affluent, white, homeowners were more likely to contain trees in public spaces when compared to communities of predominately economically disadvantaged, black, renters. Lynch and Stretesky (2013) found an inverse relationship between density of citizen-led water-monitoring organizations, and percentage of ethnic and racial minority residents, net of confounders. Because environmental hazards can pose threats to public health and property values, environmental hazards can hinder or impair a person's access to resources. In this way, environmental justice is closely intertwined with social justice.

Public Health and Social Justice

Public health represents another realm that is impacted by social justice, and threatened by social inequalities. Consider that a number of factors that impact human health are decidedly social in nature. These factors include, though are not limited to the environment, access to quality medical services, access to preventive care, health insurance, and health education. To provide one of many examples of intersections between public health and social justice, consider disparities in diet and nutrition. The concept 'food desert' was coined to describe an area where limited access to affordable, healthy food options acts as a contributing factor to spatial and social disparities in health and diet (Beaulac et al., 2009). Studies in public health and urban ecology have documented disparities in the distribution of 'food deserts' across the United States with respect to race and class composition of neighborhoods (Beaulac et al., 2009; Gordon et al., 2011). A systematic review of the literature identifies 49 studies that examine the presence of food deserts in socioeconomically disadvantaged regions, and concludes that there is evidence that food access disparities with respect to race and class exist in the United States (Beaulac et al., 2009). Because of dynamics between races, socioeconomic status, and urban planning appear to interact with access to affordable, healthy foods, public health represents another domain of social justice issues.

Mental Health and Social Justice

Mental health is another domain that evinces social justice concerns. Even though mental health is typically understood as an individual trait, macrolevel patterns in psychological well-being have been observed across demographic groups. These relationships appear nuanced by specific disorder and demographic characteristic (Sheppard, 2002). For example, a positive correlation between economic status and mental health has been observed for decades (Bebbington et al., 2000; Hollingshead and Redlich, 1958). Studies have also demonstrated that women tend to be overrepresented among depression sufferers, while men tend to be overrepresented among schitzophrenics (Sheppard, 2002). In some cases, class has demonstrated to exacerbate race and gender associations with mental health: in one study, working-class women were observed as more likely to be depressed than middle-class women (Brown and Harris, 1978).

The intersections between social justice and mental health also have implications within the criminal justice system. For instance, emotional behavioral disorders tend to emerge at higher rates among inmates than within the general population. Reports from the Department of Justice suggest that nearly half of all inmates in state prisons report some symptoms of mental health disorders (James and Glaze, 2006). By comparison, other accounts suggest that among the general population, approximately 10% of persons report symptoms of mental health disorders (James and Glaze, 2006). One implication of the overrepresentation of mental health conditions among offending populations, is that correctional facilities have become *de facto* mental health service providers. Given diminishing budgets and the high stigmatization of mental health services, correctional facilities may represent the first or only opportunity for some persons to obtain help for mental health conditions. Moreover, as per the Holmes and Rahe (1967) Stress Scale, imprisonment in and of itself can act as a major contributor toward a stress-related mental health breakdown. Given the class biases and disproportionate minority contact that emerges throughout the justice system, combined with gender and ethnicity disparities, both mental health and incarceration share overlapping social justice concerns.

Social Work and Social Justice

The attainment and preservation of social justice has been described as a fundamental value of the social work discipline (Olson et al., 2013). Sheafor and Horejsi (2003) describe social justice as the fairness in the way the most powerful groups in society (governments, corporations) support and protect the basic human rights of all members of a society. To break down this definition further, Sheafor and Horejsi (2003) explain that, in general, the social work discipline recognizes basic, fundamental human rights of all persons. These rights include (though are not limited to) protection from abuse, a right to work and earn income to support oneself, and a right to quality education and health-care services. In addition, these rights operate independently of economic status or other personal achievements; rather, they are basic human rights afforded to members of a society simply by virtue of one's own existence. Because social work recognizes the importance of these basic human rights, social work also recognizes that societies have

a responsibility to protect basic human rights. That is, social work interprets rights and responsibilities as two sides of the same coin (Sheafor and Horejsi, 2003). For example, if people have the right to earn wages to support themselves, society has a responsibility to preserve fair and equal access to gainful employment.

One method that social institutions employ for protection of basic human rights is the criminal law. For this reason, there is considerable overlap between the mission of social work and the function of the criminal law and criminal justice system. For example, society recognizes a right to be free from abuse. Consequentially, laws exist to protect persons from physical harm and abuse. Agents of the criminal justice system have been granted the power to enforce these laws in the interest of protecting citizens from harm. Social workers have served as advocates for the creation of certain laws and social action (e.g., Etta Wheeler is recognized as a social worker who influenced social institutions to take an active role in protecting young people from child abuse), as well as intervening to protect human rights at individual and family levels. However, that society recognizes protections of some harms (e.g., acts of interpersonal violence) over others (e.g., injurious corporate crimes), remains a social justice concern for social workers and criminal justice researchers and practitioners alike (Reiman and Leighton, 2012).

Prison Abolition and Social Justice

According to Mayrl (2013) prison abolition represents a social justice issue that took root in the 1970s, but experienced renewed attention in the late 1990s. Prison abolitionists call for the outright elimination of prisons, and thereby an end to the use of prisons as a means of responding to social problems. Prison abolitionists do not interpret imprisonment as a meaningful or effective response to social ills (e.g., crime). Rather, prison abolitionists view imprisonment as a form of oppression, thereby acting as a mechanism that reproduces and reinforces social inequities. The term 'prison industrial complex' has been used to describe the intersections between the boom in the prison population, the emergence of private, for-profit prisons, and the manipulation of fear of crime by politicians and mainstream media. Parallel to Eisenhower's coining of the phrase "military industrial complex," used to describe the use and production of military arms to advance economic and political agendas, the prison industrial complex has been interpreted as advancing ulterior economic and political agendas (Mayrl, 2013). That is, the prison boom may be better explained as a lucrative business practice, or election campaign strategy, rather than an empirically justified approach to crime control. Because prison populations tend to be individuals of lower socioeconomic status, persons struggling with addiction and other mental health problems, as well as racial and ethnic minorities, prison abolitionists draw attention to the fact that prisons reinforce the marginalization of these minority groups. For these reasons, prison abolitionists interpret imprisonment as a form of oppression, and in turn, a threat to social justice.

Conclusion

In sum, social justice is a broad, interdisciplinary topic that continues to be particularly relevant to criminology. Social justice also continues to captivate practitioners and activists, some of which have galvanized to form influential social justice movements. Many of these movements organize and champion one or more specific social justice issue, and concentrate their efforts on heightening awareness and/or advocating for change. Examples of well-known social justice movements include the US civil rights movement, the women's suffrage movement, and more recently, the occupy wall street movement (Portland State University, 2013). These movements not only acted as powerful forces in shaping US civil rights and liberties, but also educate and inspire policy makers and grass roots social justice movements that continue to develop today.

See also: Criminal Justice and Social Work Practice; Justice and Health; Justice: Social Psychological Perspectives; Social Action and Social Justice.

Bibliography

Akers, R.L., Sellers, C.S., 2009. Criminological Theories: Introduction, Evaluation, and Application, fifth ed. Oxford University Press, New York.
Arrigo, B.A., 2000. Social justice and critical criminology: on integrating knowledge. Contemporary Justice Review 3 (1), 7–37.
Arrigo, B.A., 2001. Critical criminology, existential humanism, and social justice: exploring the contours of conceptual integration. Critical Criminology 10 (2), 83–95.
Arrigo, B.A., Schehr, R.C., 1998. Restoring justice for juveniles: a critical analysis of victim-offender mediation. Justice Quarterly 15 (4), 629–666.
Beaulac, J., Kristjansson, E., Cummins, S., 2009. Peer reviewed: a systematic review of food deserts, 1966–2007. Preventing Chronic Disease 6 (3).
Bebbington, P., Brugha, T., Meltzer, H., Farrell, M., Ceresa, C., Jenkins, R., Lewis, G., 2000. Psychiatric disorder and dysfunction in the UK national survey of psychiatric morbidity. Social Psychiatry and Psychiatric Epidemiology 35, 191–197.
Berryman, P., 1987. Liberation Theology. Temple University Press, Philadelphia, PA.
Brown, G.W., Harris, T.O., 1978. Social Origins of Depression. Tavistock, London.
Gordon, C., Purciel-Hill, M., Ghai, N.R., Kaufman, L., Graham, R., Van Wye, G., 2011. Measuring food deserts in New York City's low-income neighborhoods. Health & Place 17 (2), 696–700.
Hollingshead, A., Redlich, F., 1958. Social Class and Mental Illness. John Wiley, New York.
Holmes, T.H., Rahe, R.H., 1967. The social readjustment rating scale. Journal of Psychosomatic Research 11 (2), 213–218.
James, D.J., Glaze, L.E., 2006. Mental Health Problems of Prison and Jail Inmates. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC.
Kim, C.Y., Losen, D.J., Hewitt, D.T., 2010. The School-to-Prison Pipeline: Structuring Legal Reform. New York University Press, New York.
Landry, S.M., Chakraborty, J., 2009. Street trees and equity: evaluating the spatial distribution of an urban amenity. Environment and Planning A 41 (11), 2651.
Lynch, M.J., Stretesky, P.B., 2013. The distribution of water-monitoring organizations across states: implications for community environmental policing and social justice. Policing: An International Journal of Police Strategies & Management 36 (1), 6–26.
Mayrl, D., 2013. Fields, logics, and social movements: prison abolition and the social justice field. Sociological Inquiry 83 (2), 286–309.
McEvoy, K., Gormally, B., 1997. 'Seeing'is believing: positivist terrology, peacemaking criminology, and the Northern Ireland peace process. Critical Criminology 8, 9–30.
Milovanovic, D., 1997. Chaos, Criminology, and Social Justice: The New Orderly (Dis) Order. Praeger Publishers, Westport.
Morrison, B.E., Vaandering, D., 2012. Restorative justice: pedagogy, praxis, and discipline. Journal of School Violence 11 (2), 138–155.

Nickerson, A.B., Spears, W.H., 2007. Influences on authoritarian and educational/therapeutic approaches to school violence prevention. Journal of School Violence 6 (4), 3–31.

Novak, M., 2000. Defining social justice. First Things, 11–12. http://www.firstprinciplesjournal.com/articles.aspx?article=1760.

Office of Juvenile Justice and Delinquency Prevention, 2013. Restorative Justice. Retrieved 20.06.13 from: http://www.ojjdp.gov/mpg/progTypesRestorativeJustice.aspx.

Olson, C.J., Reid, C., Threadgill-Goldson, N., Riffe, H.A., Ryan, P.A., 2013. Voices from the field: social workers define and apply social justice. Journal of Progressive Human Services 24 (1), 23–42.

Pepinsky, H., Quinney, R. (Eds.), 1991. Criminology as Peacemaking. Indiana University Press, Bloomington.

Portland State University, 2013. Social justice movements. Obtained 22.06.13 from: http://online.ccj.pdx.edu/ccj-careers-resources/criminal-justice-resources/research/social-justice-movements/.

Quinney, R., Wildeman, J., 1991. The Problem of Crime: A Peace and Social Justice Perspective. Mayfield Publishing Company, Mountain View, CA.

Reiman, J., Leighton, P., 2012. The Rich Get Richer and the Poor Get Prison, tenth ed. Allyn & Bacon, Boston, MA.

Saha, R., Mohai, P., 2005. Historical context and hazardous waste facility siting: understanding temporal patterns in Michigan. Social Problems 52 (4), 618–648.

Sheafor, B.W., Horejsi, C.R., 2003. Techniques and Guidelines for Social Work Practice, sixth ed. Allyn & Bacon, Boston, MA.

Sheppard, M., 2002. Mental health and social justice: gender, race and psychological consequences of unfairness. British Journal of Social Work 32 (6), 779–797.

Spader, D.J., 1988. Criminal justice and distributive justice: has the wall of separation been reduced to rubble? Justice Quarterly 5 (4), 589–614.

Sullivan, D., Tifft, L., 1998. Criminology as peacemaking: a peace-oriented perspective on crime, punishment and justice that takes into account the needs of all. Criminal Justice Studies 11 (1–2), 4–34.

Tyler, T.R., Boeckmann, R.J., Smith, H., Huo, Y.J., 1997. Social Justice in a Diverse Society. Westview, Boulder, CO.

Social Justice and Social Inequalities

Gregg Barak, Eastern Michigan University, Ypsilanti, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social justice is a philosophical and transdisciplinary concept that has no particular or distinct disciplinary home. Social justice in theory and practice is part of the general evolution of justice in human civilizations, which is a part of the ongoing struggles against the repression of any people and on behalf of the liberation of all people. The applications of social justice are best understood in relation to the applications of equal justice and restorative justice.

Social justice is a philosophical and transdisciplinary concept that has no particular or distinct disciplinary home. Within and without the domains of criminology, both the struggle for and the study of social justice are underdeveloped and/or unfinished projects. Presently, the constitution of social justice and how to secure human rights for all are not firmly established or agreed upon by most people. Nevertheless, a listing of the basic liberties of social justice includes freedom of ideas, thoughts, and speech; freedom of association; freedom of ethnic and religious culture; freedom from slavery; freedom from racism, sexism, and discrimination; the right to equality and human needs for all; and the redistribution of power, wealth, and status for all, and for entire societies (Sadeghi and Price, 2007).

Over the short term, the struggle for social justice has reflected the historical changes and continuities that helped to form the Universal Declaration of Human Rights (UDHR), adopted by the General Assembly of the United Nations in 1948. Drawing on the battle cry of the French Revolution, 'dignity, liberty, equality, and brotherhood'; on the demands of the Industrial Revolution for political, social, and economic equity; and, finally, on the communal and national solidarity movements associated with the postcolonial era, the articles of the UDHR brought together in one document the universal meanings of human rights.

This struggle for universal human rights within the evolution of justice has not followed anything resembling a linear pathway forward in time and space. Not only has every generation of rights met with resistance, but also major strides on the course to human rights efficacy have also experienced some kind of backlash or setback. To be sure, reactionary forces have not typically nullified each progressive chapter in the acquisition of human rights. In fact, the collective evidence is otherwise: "History preserves the human rights record as each generation builds on the hopes and achievements of its predecessors while struggling to free itself from authoritarianism and improve its social conditions" for the future (Ishay, 2004: p. 4).

Background

Social justice may be thought of simply as "the common good through the equalization of goods or services" (Sadeghi and Price, 2007: p. 4), or it may be more elaborately defined as "the fair distribution of opportunities, rewards and responsibilities in society, as well as principles and institutions for the distribution of meaningful social goods – income, shelter, food, health, education, the freedom to pursue individual goals" (Hudson, 2013: p. 432).

An Italian Catholic priest, Luigi Taparelli, coined the term 'social justice.' In his book, *Theoretical Treatise on Natural Law* (published between 1840 and 1843), and based on the teachings of Thomas Aquinas, he claimed that the components of social justice were products of "natural theology, in which morality, based on religion, ought to be sought after by all people, who, under God, must adhere to their moral beliefs in doing what is right" (Sadeghi and Price, 2007: p. 699). In 1948, the late Father William J. Ferree (1997) in his "The Act of Social Justice" had characterized social justice as the duty in which each person in society is morally obligated to care for the common good of all.

Probably the most influential theory of social justice has been John Rawls' notion of "justice as fairness" in his *A Theory of Justice* (1972), which draws upon both the basic social contract theory and the Kantian philosophical tradition of justice as impartiality. Like most theorists on social justice, Rawls focused his concerns on answering two fundamental questions. "First, how the rules for a fair distribution of meaningful social goods can be determined; second, how much or how little inequality is permissible in a socially just society" (Hudson, 2013: p. 432). In response to the first question, Rawls answers that the rules and institutions of a just society can be obtained only by people acting in an unnatural position of ignorance of their own position, or behind a "veil of ignorance." Accordingly, he is suggesting that a fair distribution can only come about by establishing rules or procedures designed for the purposes of securing impartiality as much as possible. With respect to the question of inequality versus equality or to Rawls's "principle of difference," which maintains that social and economic inequalities are just to the extent that they benefit everyone.

In answering these questions, Rawls sought to strike a balance between equality and freedom. Alternative theories usually prioritize one or the other. For example, Marxist theories line up behind equality; libertarian and neoliberal theories line up behind freedom. A variant of welfare liberalism articulated by Michael Walzer in *Spheres of Justice* (1983) defends pluralism and equality, wherein he argues that diverse goods and services should have different degrees of social distribution on the basis of individual need.

On the 'Evolution of Justice': Equal, Restorative, and Social

Justice is an evolving phenomenon in its ideation, ideology, and implementation. As social constructions and ideal systems, justice discourses and practices cannot be separated from the modern evolution of law, justice, and society – or of what can be described as a three-tier revolution in just-ology or in the entitlement of all human beings to share the exact same formal rights as everybody else regardless of their class, race or ethnicity, gender, religious, national, or sexual affiliation (Ishay, 2004). When each of these models or systems of justice first emerged, they represented a particular stage in the struggle to expand individual rights in relation to both the collective rights of societies and the rights of the nation-states. Ultimately, these rights were extended to all peoples as universal human rights.

The first generation of rights represented the struggle for 'equal justice' or the struggle for 'negative rights.' These called for a restraint from uninhibited state or monarchial power over the individual. This generation of rights was derived from the American and French Revolutions, which both struggled to gain freedom from arbitrary rule. Collectively, these rights, articulated more recently in the Civil and Political Rights of the International Bill of Rights (1966), have shaped governmental control by "rule of law" rather than "rule of man." In the domains of crime and crime control, these rights have at least formally established an impartial and fair enforcement of the procedural and substantive components of the criminal and civil laws.

The second generation of rights represented the struggle for 'restorative justice' or the struggle for 'positive rights.' These called for 'affirmative' actions by and/or on behalf of the state for pursuing the common welfare. Rooted in the struggles to democratize governments and to make them responsible to the needs of people living in 'post' laissez-faire monopoly societies as a whole, these rights have shaped the sociopolitical obligations or minimal duties of the state to facilitate the self-actualization of individual human beings (Barak, 1980). These rights were also collectively articulated in 1966, in the Economic, Social, and Cultural Rights of the International Bill of Rights. In the domains of crime and crime prevention, these have included an emphasis on community social welfare, penitence and redemption, and offender–victim reconciliation.

The third generation of rights is represented by the contemporary worldwide struggle for 'social justice,' or the struggle for universal human rights. Evolving out of the emerging conditions of global interdependence, these rights call for international rules of interaction, such as the establishment of the International Criminal Court in The Hague in 1999. These universal rights also recognize that the delivery of negative and positive rights for all cannot be satisfied within the body of individual states acting alone. In the universal struggle for social justice, world peace, and victim reduction, global emphases have been on ending world hunger, forgiving debt to underdeveloped nations, and treating victims of AIDS/HIV and other debilitating and/or life-threatening diseases, famines, tsunamis, and so forth worldwide.

Equal Justice

In most developed nations like the United States, equal justice assumes the rationality of the prevailing political, economic, and social arrangements in general and of the administration of criminal justice in particular. Within those systems of justice, where criminals are viewed as 'bad' and 'mad,' they are typically disconnected from their social conditions and from their class, racial or ethnic, gendered, and sexual relations. Whether defendants are from at-risk social environments or from privileged milieus, these facts are not formally relevant to the accused or convicted offenders as all are subject to the same 'equal justice' or due process of the law.

Ignoring the effects of social structures and ecologies of crime, these models of equal justice have traditionally not considered the interests of either the injured parties or their perpetrators or the communities from which they originate. Furthermore, the policies of equal justice do not take into account the concept of equal sanctions for unequal offenders or victims. Similarly, the unequal treatment of analogous social injuries or harms perpetrated by the powerful and the powerless, respectively, serve to reproduce the status quo of crime, victimization, and discriminatory justice for many.

The contemporary model of equal justice practices in the United States has its roots in the mid-eighteenth century, when the European Age of Reason or Enlightenment was busy reforming the excessively arbitrary and barbaric justice practices from the medieval period. Although not driven by revenge or vengeance, these utilitarian and practical applications of equal-formal (retributive) justice are repressive as these help to perpetuate a class of marginally 'dangerous' offenders. At the same time, these policies of equal criminal justice downplay the value for most rule violators of flexible sentencing and community alternatives to imprisonment.

Restorative Justice

Fundamentally, restorative justice systems are about decentering punishment in regulatory institutions. At the same time, restorative justice acknowledges the place that punishment has as a means of communicating the 'shamefulness' of the acts in question (Braithwaite et al., 2006). Restorative policies and practices of justice assume that most offenders and nonoffenders, whether perpetrators, victims, or both, fundamentally share an intrinsic humanity (Zehr and Mika, 1998).

Modern practices of restorative justice have their legal roots in the ancient patterns of such diverse cultures as the Sumerian Code of UrNammu (2050 BC), the Hebrew Scriptures and the Code of Hammurabi (1700 BC), the Roman Law of the Twelve Tribes (449 BC), and the earliest collection of the Germanic tribal laws, the Lex Salica (AD 496). Each of these legal systems required that offenders and their families settle accounts with victims and their families, not simply to ensure that injured persons received restitution or compensation but also to ensure that communities were reestablished or restored to their former peace. Similarly, in many precolonial and indigenous societies around the world, sanctions were also primarily compensatory rather than retributive, intended primarily to make victims whole or to restore them to their previous status (Barak et al., 2010).

Restorative justice is being pursued when the focus of justice is on the harms of wrongdoing more than the rules that have been broken, and on the victims and offenders working toward the restoration of the former and the support of the latter in understanding their transgressions and meeting their obligations to make things right again; and is aimed at empowering those affected parties in the processes of justice and dialogue. For example, contemporary systems of Japanese justice, which emphasize 'confession, repentance, and absolution,' are about compensating their victims and restoring community peace (Van Ness and Heetderks-Strong, 1997).

Today, the idea of restorative justice has come to have many different meanings and practices. It has come to be associated with innovations in community mediation, problem-solving justice, victim–offender reconciliation, alternative sentencing, and community service (Daly and Immarigeon, 1998). Regardless of the myriad of practices that seem to be part of a larger movement to incorporate restorative justice programs into communities, they all share in common the desire to move beyond the illegality of the crime to the more fundamental needs of human dignity and respect as well as to the mediated sources of conflict and dispute resolution.

Unlike equal justice systems that revolve around how much pain and suffering have been inflicted by the actions of the wrongdoer, restorative justice systems revolve around how much harm has been repaired or will be prevented ('less harm'). Restorative justice systems are preoccupied with both the interpersonal and larger structural-mutualistic relationships involving offenders, victims, family members, and the neighborhoods from which they are spawned. More specifically, restorative justice emphasizes the recovery of the victim through the processes of redress, vindication, and healing, on the one hand, and of the offender through the processes of fair treatment, reformation, and recompense to the victim, on the other.

Finally, restorative justice relies on both the formal and informal mechanism of social control as it supports the role of government as being responsible for preserving law and order and the role of the community as being responsible for preserving peace and justice. Most importantly, for example, restorative justice involving victim–offender encounters or reconciliation programs tends to humanize both offenders and victims to the other and "permits them substantial creativity in constructing a response that deals not only with the injustice that occurred but with the futures of both parties as well" (Van Ness and Heetderks-Strong, 1997: p. 89).

Social Justice

While equal justice systems revolve primarily around retribution between the offender and the state, and restorative justice systems revolve primarily around reparation and healing between the offender and the victim, social justice systems venture beyond immediate conflicts – perpetrators, victims, and communities – to an account of larger political, economic, and social arrangements. Social justice models incorporate, for example, patterns of social inequality or disadvantage as well as patterns of privilege, power, and advantage, which make people and their communities susceptible or not to the experiences of criminal harm and to the processes of criminalization.

The visions of equity and fairness associated with the policies of social justice are broader and more ambitious than those visions associated with the models of equal and restorative justice. Social justice models expand the notions of conflict and injury beyond what the sovereign law recognizes to include those harms identified as part of an evolving set of human rights, some established in treaties and covenants, and some in international resolutions or tribunals. The violations of any of these fundamental rights constitute what are known in the world community as 'crimes against humanity.'

These offenses generally are committed by policies or practices or by the authorities or agents of the state, such as in the violation of a person and/or a group's inalienable rights to be free, for example, from exploitation, slavery, impoverishment, hatred, discrimination, or genocide. In this context, social justice recognizes harms, injuries, offenses, wrongs, crimes, and violations that are not merely expressions of interpersonal conflict but also expressions of the organizational, institutional, and structural relations of the prevailing political, economic, and social arrangements. Accordingly, to address the underpinning structural inequities or injustices in those arrangements, social justice stresses the importance of public policies of 'harm reduction' that go beyond the confines of the criminal justice system.

Therefore, attention is paid not only to the obvious crimes of the powerful, such as those committed in corporate America or on Wall Street, but also more generally to the economic harm and social victimization caused by all types of deregulation, from environmental to financial; to the lack of social and human capital formation; and to the oppression and exploitation of a growing multiethnic underclass in the United States and many other nations. To the extent that these social forces create crime and crime control systems, in the forms of equal or restorative justice, which are repressive in the sense that they secure a legal order that reinforces inequality and privilege, social justice models advocate a 'war' or intervention against those forces.

In the evolution of justice, what distinguishes social justice from equal justice and restorative justice is that only the former model of justice struggles to escape the confines or shackles of bourgeois legality. The problems with capitalist Constitutions like the United States' are the many claims that equal rights and responsibilities have been substituted for the harsh realities of class domination. In other words, by way of the Constitution, the struggle over the legitimacy of social acts is removed from the plane of morality and society. In turn, these social actions or relationships are relocated in the plane of law and politics. Justice is no longer what is fair but what is legal and politically expedient as decided mostly by lawyers and judges playing by the capitalist rules of the game (Ollman, 1987).

Ultimately, to overcome class domination and/or to achieve social justice, the political economies and societies of the world will have to transition from capitalist formations to socialist formations. Otherwise, social justice for all will remain an unrealized 'pipe dream' trapped in the realities or contradictions of capitalist inequalities. In the meantime, piecemeal or halfhearted reforms in the name of social justice are resumed and advanced.

For example, some proponents of social justice advocate on behalf of 'visionary gradualism' or 'free-market socialism' (Harrington, 1989). Other social justice proponents talk in terms

of democratizing and humanizing capitalism (Shiller, 2011). Grounded in the global principles of feminist, antiracial, and ecological justice, these views still ascribe to the political economies of capitalism. At the same time, these 'social justicians' believe in the eradication of social subjugation, oppression, and exploitation as well as in the establishment of fundamental human rights for all. They oppose unenlightened self-interest, unregulated markets, and unfettered victimization (Barak, 2012).

The Limitations of Equal Justice and the Struggle for Sovereign Justice

As Anatole France wrote about the inefficacy of equal justice in his best-selling *The Red Lily* in 1894, "The law, in its majestic equality, forbids rich and poor to sleep under bridges, to beg in the streets, and to steal bread" (quoted in Sadeghi and Price, 2007: p. 699). Then and now, such laws as those are naturally only relevant to the marginal, to the destitute, to the homeless, to the mentally ill, and so on. Hence, when legal evaluations or assessments of justice – criminal or civil – ignore differences of class, ethnicity, gender, or sexuality in particular and socioeconomic statuses in general, or when they falsely treat unequal persons or corporations under the law as one and the same, they 'blindly' reinforce these differences or inequalities.

Similarly, because of an ideological preference for bourgeois legality with its overemphasis on notions of individual-equal justice, most critiques of the administration of justice fall short. Instead of focusing on the substantive inadequacies of the criminalizing and penalizing processes and the selective and differential outcomes that systemically hurt the powerless and benefit the powerful, most critiques focus on procedural irregularities or on applications of due process and equal protection under the law. The point is that there is a long list of harms and injuries that have been and/or could be prohibited by law that victimize millions of people daily, yet are commonly treated by the legal order as if these actions or omissions were or are 'beyond incrimination' (Barak, 1991a, 1991b, 2012).

Moreover, the consequences of these selective enforcements, differential applications, and inauspicious omissions are counterproductive not only for alleviating pain or reducing harm, but also for obtaining social justice for the most marginalized members of society. For example, the unsuccessful policies of the 'war on drugs,' mandatory sentencing, and the widening of the nets of crime control on less serious offenders have adversely affected African American families and, to a lesser extent, Latino American families. In the case of the former, the large-scale removal of young black males from their communities has contributed to the depletion of the supply of potential marriage partners for young black females and to the accumulation of generational impoverishment. Some criminologists have argued that these social relations of crime, victimization, and punishment have encouraged young female-headed households, creating precisely the types of family formations that have been linked with higher rates of street crime and domestic abuse when not given adequate social support (Currie, 2005; Messner and Rosenfeld, 1994; Foster and Hagan, 2009).

In turn, these countereffective policies coupled with assembly-line, plea-bargained, and bureaucratized equal justice for the marginal classes help to reinforce stereotypic images of criminals. These exclude high-powered corporate executives, emphasizing low-life predators who murder, rob, assault, kidnap, and do drugs rather than white-collar looters who inside trade, contract without bidding, or commit high-stakes securities fraud. To be clear, the point of this critique is not to argue against equal justice in the administration of laws. It is to argue against equal justice as the endpoint of justice rather than as a means to complementing other systems of justice. Indeed, all agents or workers of the various equal justice systems should aspire to act impartially or neutrally, according to both the letter and the spirit of due process and equal protection under the law. However, to achieve social justice, legal-equal justice represents the most elementary of steps.

In addition, equal justice needs to be assisted by the goals and objectives of restorative and social justice if the mass victimization of peoples around the world is to decline. These latter models of justice offer substantial ways to improve the quality of justice beyond the existing legal processes, helping to better curb and reduce all forms of harm and injury – personal, organizational, institutional, and structural – than equal justice systems do.

The Limitations of International Law and the Struggle for Global Justice

The contradictions in the spirit and the enforcement of international law inhibit a reduction in both the worldwide victimization of people and the reproduction of crime and violence. This is because international law does not facilitate the ability of the global community to satisfy the basic human needs and rights of all people. At its core:

> International law is a limitation on autonomy, yet one that has its maximal force when other states are willing to act in its name. The twentieth and twenty-first centuries have shown that when states elect to act 'in the name' of international law… they often do so (or refuse to do so) out of their own political and economic interests rather than within the interests of an international social order. (Mullin and Rothe, 2008: p. 26)

The radical vision of social justice shared below stems from a humanistic, democratic, and participatory world order based on 'mixed' economies of public and private ownership, where people cooperate in both the work and leisure of their local and global communities. It is not that societies will have removed their economic and other forms of competition. Rather, these reciprocal interactions of capital and labor will work on the premises of 'fair' rather than 'free' exchanges of goods and services. These social relations are also underpinned by sustainable and balanced ecosystems. A global order of this kind strives for ecological harmony in the context of climate change and in the ways in which societies and technologies interact with their natural environments to protect all forms of life and habitat (Barak, 2009).

Human Needs

At the beginning of the twenty-first century, we live in a world of shrinking resources but not in a world of scarcity of goods and services. Thus, in a world organized around ecological sustainability, a primary goal of global production of international economic activity should be for the provision of the necessities of life for all of earth's inhabitants. These basics should include but not be limited to food, shelter, healthcare, childcare, education, social services, culture, art, leisure, recreation, work, peace, and security. In brief, this is social production for use and profit by all so that whole groups of people do not have to needlessly suffer.

Human Rights

The objective here is to maximize and extend civil and human rights for all people, which means also curbing abuses and discriminatory actions based on or stemming from differences due to age, ethnicity, gender, sexual orientation, religious affiliation, disability, or migratory status. For example, with regard to sexual orientation, the end of all anti-gay, lesbian, bisexual, and transgender restrictions in the law; or, concerning age, there are special rights that may be attached to children and seniors; regarding women, the reduction and elimination of all forms of sexism, especially that which revolves around sexual harassment, violence, and exploitation; and finally, concerning migratory workers, a global world order that would recognize the rights of people to freely move across borders to live, visit, and work as they choose.

Human Justness

Perhaps the toughest struggles for social justice have to do with obtaining human justness *in* everyday or ordinary public-private contracts, and *for* exceptional or extraordinary atrocities. In the case of the former, do the ways in which contracts or 'right-to-work' laws are written adversely affect workers in general, and current or former inmates in relation to the penal-industrial complex in particular? In the case of the latter, there are the largely inadequate or unresponsive actions to acts of genocidal rape or to other crimes against humanity, such as torture. With domestic and international laws in place to challenge these various forms of harms and social injuries, there is hope that such behavior might be reduced, if not deferred, altogether. Without either changing the fundamental international motivations and opportunities of a global economy and/or establishing the necessary constraints and customary controls to curb these types of antisocial behavior, however, governments and their proxies in specific situations are under little or no pressure to refrain from their continued abuse.

Future Prospects

Dialectically, within an 'evolving' system of social justice and a developing system of capitalism worldwide, there are presently more middle classes emerging than there are middle classes disappearing. At the same time, as the wealth gap expands and inequality accelerates, the competition for uneven wage differences intensifies at home and abroad. Meanwhile, the hegemonic neoliberal policies – austerity, privatization, and deregulation – that are directing both domestic and international development all work against the materialization of redistributing the wealth for all human beings. In addition to expanding or widening the universality of humanitarian law and practice, the struggle for social justice has to consolidate its effort, nationally and supranationally, to resist those harmful policies of hierarchy and privilege. The struggle for social justice should also devise alternative strategies for democratizing and humanizing the vagaries of capitalism.

See also: Civil Liberties and Human Rights; Equality and Inequality: Legal Aspects; Equality: Philosophical Aspects; Freedom, History of; Freedom: Political; Genetics and Social Justice; Genocide and War; Genocide, Historical Aspects of; Genocide: Social and Economic Aspects; Human Genome Project: History and Assessment; Human Rights, Anthropology of; Human Rights, History of; Human Rights: Legal Aspects; Income Distribution and Inequality: Measurement Issues; Income Inequality; Inequality, Social; International Law and Treaties; Justice and Law; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Morality: Evolution of; Rawls, John (1921–2002); Slavery and Violence; Social Inequality in History (Stratification and Classes); Wealth Distribution.

Bibliography

Barak, G., 1980. In Defense of Whom? A Critique of Criminal Justice Reform. Anderson Publishing, Cincinnati, OH.
Barak, G., 1991a. Gimme Shelter: A Social History of Homelessness in Contemporary America. Praeger, Westport, CT.
Barak, G. (Ed.), 1991b. Crimes by the Capitalist State: An Introduction to State Criminality. State University of New York Press, Albany, NY.
Barak, G., 2009. Criminology: An Integrated Approach. Rowman & Littlefield, Lanham, MD.
Barak, G., 2012. Theft of a Nation: Wall Street Looting and Federal Regulatory Colluding. Rowman & Littlefield, Lanham, MD.
Barak, G., Leighton, P., Flavin, J., 2010. Class, Race, Gender, and Crime: The Social Realities of Justice in America, third ed. Rowman & Littlefield, Lanham, MD.
Braithwaite, J., Braithwaite, V., Ahmed, E., 2006. Reintegrative shaming. In: Henry, S., Lanier, M. (Eds.), The Essential Criminology Reader. Westview, Boulder, CO, pp. 288–295.
Currie, E., 2005. Inequality, community, and crime. In: Henry, S., Lanier, M. (Eds.), The Essential Criminology Reader. Westview, Boulder, CO, pp. 299–306.
Daly, K., Immarigeon, R., 1998. The past, present, and future of restorative justice: some critical reflections. Contemporary Justice Review 1 (1), 21–45.
Ferree, W., 1997. Introduction to Social Justice. Paulist Press, Mahwah, NJ.
Foster, H., Hagan, J., 2009. The mass incarceration of parents in America: issues of race/ethnicity, collateral damage to children, and prisoner reentry. Annals of the American Academy of Political and Social Science 623 (May), 179–194.
Harrington, M., 1989. Socialism: Past and Future. Arcade Publishing, Berkeley, CA.
Hudson, B., 2013. Social justice. In: McLaughlin, E., Muncie, J. (Eds.), The Sage Dictionary of Criminology, third ed. Sage, London, pp. 432–433.
Ishay, M., 2004. The History of Human Rights: From Ancient Times to the Globalization Era. University of California Press, Berkeley, CA.
Messner, S., Rosenfeld, R., 1994. Crime and the American Dream. Wadsworth, Belmont, CA.
Mullin, C., Rothe, D., 2008. Blood, Power, and Bedlam: Violations of International Criminal Law in Post-Colonial Africa. Peter Lang, New York.
Ollman, B., Dec 1987. Toward a Marxist Interpretation of the U.S. Constitution. Monthly Review 39 (7).
Rawls, J., 1972. A Theory of Justice. Oxford University Press, Oxford, UK.
Sadeghi, L., Price, B., 2007. Social justice. In: Barak, G. (Ed.), Battleground: Criminal Justice. Greenwood Press, Westport, CT, pp. 697–704.
Shiller, R., 2011. Democratizing and humanizing finance. In: Kroszner, R., Shiller, R. (Eds.), Reforming U.S. Financial Markets: Reflections before and after Dodd-Frank. MIT Press, Cambridge, MA, pp. 1–36.
Van Ness, D., Heetderks-Strong, K., 1997. Restorative Justice. Anderson Publishing, Cincinnati, OH.
Walzer, M., 1983. Spheres of Justice: A Defense of Pluralism and Equality. Basic Books, NY.
Zehr, H., Mika, H., 1998. Fundamental concepts of restorative justice. Contemporary Justice Review 1 (1), 47–55.

Social Justice: Historical and Theoretical Considerations

Kathleen M Weigert, Loyola University Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by V.H. Schmidt, volume 21, pp. 14338–14341, © 2001, Elsevier Ltd.

Abstract

Social justice, generally considered a type or subcategory of justice, has been variously defined. Some authors consider it equivalent to 'distributive justice' while others differentiate it from both general justice and distributive justice. In the latter approach, social justice is seen as a unique type of justice characterized by a focus on the 'common good' and the individual's obligation and right to make a contribution to that (hence, sometimes called 'contributive' justice) while acknowledging the role of the state and civil society to remove barriers that prevent individuals from so doing. Defining it theoretically and determining how it can be advanced empirically, as well as questioning its usefulness in both arenas, are the focus of much contemporary attention.

Theories of Social Justice

While the concept of 'justice' has been around since at least the time of Plato, the term 'social justice' itself did not enter the lexicon until the mid-nineteenth century, through the writing of the conservative Italian Jesuit Luigi Taparelli d'Azeglio (Burke, 2008). Scholars in the Catholic intellectual tradition built on the foundational work about justice in the writings of Thomas Aquinas (who, in turn, built on the work of Aristotle); the link between justice and the 'common good', while complex, is pivotal in this tradition. There are some early twentieth-century attempts to explore the medieval roots of the modern understanding of the idea of social justice as well as to advance the thinking about the concept. The study of the general concept of justice, however, suffered hard times with the turn in philosophy to such approaches as analytic philosophy and logical positivism.

It is in the latter half of the twentieth century that we find renewed interest in theorizing about justice itself. Credit is given especially to the pioneering work of John Rawls for reviving that interest, and with scholars such as David Miller, Iris Marion Young, Amartya Sen, and Martha Nussbaum for extending it, particularly in a more explicitly social vein. Miller (1999) contends that what is involved in social justice is really the distribution of the good and bad things in life, as well as the question of how they should be distributed, thus making social justice an 'aspect' of or equivalent to distributive justice.

Young (2000) is more explicit in establishing social justice as its own concept. She does so by introducing what she calls 'two ideals of social justice', namely, self-development and self-determination, which correspond to two 'general conditions of injustice', namely, oppression (institutional constraint on self-development) and domination (institutional constraint on self-determination). This leads to her definition of social justice as "the institutional conditions for promoting self-development and self-determination of a society's members" (Young, 2000: p. 33). Why this is important is because of what Young claims is the 'goal' of social justice, namely, 'social equality', and here equality should be seen as "the full participation and inclusion of everyone in a society's major institutions, and the socially supported substantive opportunity for all to develop and exercise their capacities and realize their choices" (Young, 1990: p. 173).

Young is building on Rawls, who argues that the main subject of justice is the 'basic structure of society', meaning the arrangement of the major social and political institutions. Young (Young, 2011: p. 70) finds this approach inadequate, however, because Rawls "is looking for a *part* of the society, a small subset of its institutions that is more fundamental than other parts" (Young, 2011: p. 70, italics in original). She contends, however, that "Social structures are not a part of the society; instead they involve, or become visible in, a certain *way of looking* at the whole society, one that sees patterns in relations among people and the positions they occupy relative to one another"(Young, 2011: p. 70; italics in original). This brings Young's approach more in alignment with a fairly standard perspective from the social sciences, where social structure is seen as the relatively stable patterns of social relationships that both emerge from and constrain the actions of individuals, who occupy different and differentiated positions that have particular roles. However defined, issues of agency, access, norms, and power come to the fore as do levels of analysis such as individual-to-individual, intergroup, intrastate, interstate, and global.

The discussion of social structure is rooted, either implicitly or explicitly, in ideas about the human person or more generally about human nature. This in turn leads to ideas about what humans need, desire, and deserve. In his discussion of what persons need, Rawls refers to citizens 'who are fully cooperating members of society' (Rawls, 2001: p. 58). He puts forth the idea of 'primary goods' and distinguishes five kinds: basic rights and liberties; freedom of movement and free choice of occupation; powers and prerogatives of offices and positions of authority and responsibility; income and wealth; and the social bases of self-respect (Rawls, 2001: pp. 58–59). A concrete and comparative perspective is introduced by Sen (2004), who proposed that a 'capability approach' provides a tool for making comparisons about the quality of life of people in different nations. This approach led to the issuance of the 'Human Development Report', which first appeared in 1990, and eventually to the 'Human Development Index', introduced on the twentieth anniversary of that inaugural report. While

Nussbaum states that the capability approach was an attempt to alter the development and policy work, rather than to develop a theory of justice, she (Nussbaum, 2011: p. 18) has developed the idea, preferring to use the plural of capability. The resulting 'capabilities' approach "can be provisionally defined as an approach to comparative quality-of-life assessment and to theorizing about basic social justice" (Nussbaum, 2011: p. 18). In current scholarship, Corning (2011: p. 11) has been perhaps the most explicit about basic needs, arguing that there is a 'science' of human nature which tells us "that each of us is defined, in considerable measure, by an array of 'basic needs' that are essential to our survival and reproductive success, and we come into the world with an orientation toward satisfying these needs" of which there are no fewer than 14 'primary needs'. From that Corning (2011: p. 153) argues that we have the wherewithal to construct a new social contract, namely, a 'biosocial contract'. While clearly a normative theory, it is one built on an empirical foundation, he argues, that demonstrates that humans have an 'innate' sense of fairness that can be advanced through economic and political reforms. Criticism of Corning's approach rests in part on the contention it is a manifestation of the naturalistic fallacy and is basically an essentialist proposition.

The discussion of social structure, human nature, and basic needs brings to the fore two other related issues, the first dealing with opportunity and outcomes, and the second with participation. Regarding opportunity and social justice, the distinction often drawn is between equality of opportunity and equality of outcomes. To understand why some might prefer the former in conceptualizing social justice, we must take into account both the historical context under examination as well as the various cultural values in play. If one is attracted to the individualism tradition in the United States in the early twenty-first century, for example, the leaning might be more toward equality of opportunity. Assuming that the playing field can, in fact, be made level, then, some would argue, equality of opportunity fulfills the extent of the societal obligation to the conditions of social justice; it is up to the individual to do the rest. If, on the other hand, one is more inclined to a communitarian tradition, one might be predisposed to favor the equality of outcomes, believing that the playing field cannot be made level; the historical contingencies are so powerful at this point that some individuals and particular groups, on their own, simply cannot meet their basic needs and contribute to the common good.

This leads fairly readily to the second issue, namely, the concern with 'participation' in the conceptualization of social justice. As noted above, in at least one tradition participation is considered a *sine qua non* of social justice making it distinct from distributive justice. Young's writings, for example, seem to put her in line with this tradition (see also Craig, 2007). Separating out participation from the concept of distributive justice confronts the issue that distributive justice can be seen to imply an exclusive top-down or formal authority perspective: who or what 'decides' (according to whatever laws and policies) to distribute the benefits and burdens to the members of a society? With that approach, the notion of agency tends to redound to the state or societal level, hence either eliminating or at least minimizing the active involvement of individuals and groups to meet their basic needs and contribute to the larger society. While Miller (1999: p. 11) cautions against taking 'distributed' too literally, a case can be made that the argument for making social justice its own category is, at least in part, to emphasize the importance of individual and group agency while not excluding the key role of civil society and the state.

No discussion of social justice can avoid the topic of the role of the state. One basic view is that the state is ultimately responsible for creating and/or maintaining the structures and processes for allocating benefits and burdens in its geographic territory and supporting (or at least not undermining) an environment in which civil society and both individual and group action can flourish. Here the issue of citizenship calls for attention. Rawls (2001) situated his work in the context of a democratic society and spoke about the idea of citizens as free and equal persons in a well-ordered society. While acknowledging the complexities of state boundaries, types, and history, Miller (2003: p. 362) affirms his earlier claim that nations are a special form of community, which, in turn, 'gives the practice of justice among fellow nationals a certain kind of primacy'. And he is quite clear on the importance of paying attention to the beliefs of citizens when theorizing about the nature of social justice. Various critiques have been waged against the idea of social justice being confined to the boundaries of the state (Bell and de-Shalit, 2003). Part of the critique comes from examining the contemporary importance of the economy and the whole discussion of globalization and its implications both for the state as an actor and for the practice of social justice. Several points are worth noting here. One is the formidable challenge that arises if the idea of social justice is broadened to be inclusive of the entire global community. What entity takes on a coordinating, if not a dominating, role in advancing social justice and with what authority? How can social justice be the 'same' in each nation-state (or smaller political entity)? A second, which some would argue is so prevalent in each state that it behooves us to address this deep fissure before attempting to address global social justice, is the issue of the gendered division of society broadly speaking and of labor in particular. At one level, this can be seen as a concrete manifestation of the issue of equality of opportunity and equality of outcomes. An early critic of Rawls' theory in terms of the gendered division of labor is Susan Moller Okin (1987: p. 68), who finds Rawls 'original position' lacking: "if those in the original position did not know whether they were to be men or women, they would surely be concerned to establish a thoroughgoing social and economic equality between the sexes that would preserve either from the need to pander to or servilely provide for the pleasure of the other." Current discussions of 'gender mainstreaming' as a strategy to achieve gender equality (United Nations) and explorations of the roots of gender inequalities in the global economy (Reilly, 2009; Perrons, 2009) add to this concern.

Advancing Social Justice from the Local to Global/ Empirical Research on Social Justice

Understanding and advancing social justice, however it is defined, have occupied scholars and activists alike. Writers in such fields as law (Rand, 2006), clinical practice (Aldarondo, 2007), and public policy (Craig et al., 2008) have

contributed to the discussion as have many scholars in social science disciplines (e.g., Cohen, 1986; Feagin and Vera, 2008; Scherer, 1992). What has oftentimes motivated both scholars and activists in this work is to seek answers to the question: why is there so much human pain and suffering of so many people in the world and what can be done about it? If people believe that the pain and suffering of others is due to luck or chance or to their own actions, on the one hand, no obligation is felt to do something to alleviate their situation. But if people determine that the suffering is 'unjust', on the other, that "implies that we acknowledge the circumstance as grounded at least partly in institutions and the social processes they generate" and "we thereby at least implicitly recognize an obligation to try to improve those social processes" (Young, 2011: p. 33). So even to those seeking greater social justice and finding themselves in the latter category, the question remains: how can this be done? The topic of 'structural injustice' (Young, 2011) or 'structural violence' (Maas Weigert, 2008) is relevant to this discussion; the basic point of these concepts is to indicate that such injustice or structural violence is 'preventable'. In conceiving of social injustice as embedded in multiple policies, private and public social structures, and the actions of a myriad array of individuals, it is still the case that the challenges to discern ways to address that harm remain complicated and both historically and culturally contingent. Designing new policies, creating new structures, and fostering new practices are never easy and always involve contending viewpoints that must be acknowledged as new ways are sought. A perennial worry arises: there will be unintended consequences.

Some theorists and activists argue that it is only through nonviolent action at all levels that positive social change can occur with the hope of at least not making the situation worse and, more proactively, of actually advancing social justice. The work on nonviolent action has developed greatly since the mid-twentieth century and both the theoretical and empirical research on social movements are central to that conversation. An underlying point about nonviolent action is that such action is rooted in the power of the people. Refusing to obey is the beginning of the challenge to the authorities, be they the state or some other entity. That idea of popular support or lack thereof is, of course, central to democratic theorizing. Whether nonviolent withdrawal of support is conceived of as 'pragmatic' or 'principled', it is argued that nonviolent action is the better (if not, only) way to bring about a society that is characterized by greater social justice in processes and structures (see Sharp, 1973). Social movements, of course, can be 'conservative' or 'progressive'. Some believe, for example, that the current structures and processes are, in fact, the most appropriate for bringing about greater social justice; there is nothing 'inherent' in a social movement that says it is or must be 'progressive'.

Future Possibilities for the Concept and Practice of Social Justice

What is the future for the concept and practice of social justice? One challenge is the question of the very usefulness of the concept itself. Does not 'distributive justice' suffice to incorporate what some are trying to claim are key elements that define 'social' justice? Since so many writers seem to equate distributive and social justice, does it help or hinder the theorizing and research to draw a sharp distinction between the two? Those who oppose such combining might argue that the very term 'social' raises more clearly such essential topics as participation, agency, access, and power, and legitimizes those topics as central, not peripheral, to the discussion of social justice. A related concern is the question of content: what can be/must be/should be included in a definition of social justice? If anyone can use the term for whatever particular purposes, does that not weaken the concept itself? Of particular concern: if it is simply identified with the liberal/progressive agenda, does that not defeat or preclude rigorous theorizing and grounded empirical research? Some present an argument that supports the continued use of the concept but contend that the changes in the late twentieth and early twenty-first centuries demand a rethinking of the elements of social justice (Craig, 2007; Cramme and Diamond, 2009).

A second challenge deals with boundary issues. As noted above, Miller (1999: pp. 18–19) has argued that the nation-state has 'a special standing', one where its citizens 'share a common national identity' that results in a political community with three features making 'the application of principles of justice feasible and fruitful': (1) strong bonds of solidarity; (2) a range of shared understandings (regarding principles of desert and of need); and (3) trust backed up by compulsion guiding behavior. Thus he contends social justice is best conceived as applying within national political communities and, therefore, we need to 'understand global justice differently'.

How can the terms social justice and global justice be differentiated? Pitting them against each other does not seem fruitful. Yet it must be acknowledged that they may conflict and in such situations there will be a call for trade-offs (Cramme and Diamond, 2009; Miller, 2009). Determining when and under what conditions trade-offs might be called for is an important task. That, in turn, raises two issues: first, the issue of sovereignty and the nation-state. Is it the case that "the heyday of the nation-state is irrefutably at an end" (Cramme and Diamond, 2009: p. 4)? And if so, what does that mean for social and/or global justice? Second, if the answer to the first question is yes, two related questions emerge: what does the nation-state look like and how does it function in this transition or twilight period (Gautney et al., 2009)? How can a transnational system of governance be created "which complements national authorities in delivering the services and public goods that citizens need to enjoy opportunity and security" (Cramme and Diamond, 2009: p. 14)?

Rather than defining the terms, some social scientists approach the issues through research on particular problems and/or organizing cross-nationally. Banerjee and Duflo (2011), for example, offer research based on randomized control trials to determine what interventions actually 'work' or help in alleviating global poverty. Others are trying to create networks that enable people to work collaboratively in an effort to advance a new vision of 'global social justice'. In the claim of the World Social Forum, for example, 'another world is possible'. In recent years, ideas about 'struggle', 'suffering', and 'solidarity' are drawing explicit attention from such scholars as Feagin and Vera (2008) and Young (2011), as do questions about why people join social movements, locally and globally.

Global movements, around such concerns as climate change and aging societies, are now possible in new ways because the electronic world allows for connections not previously available, even though such means can also be used for greater concentration of power and the diminishment of a practice of greater social justice (Gautney et al., 2009). An example of a platform for connecting social scientists interested in social justice is The Sociological Think Tank, dedicated to "reflection and discussion among sociologists, other social scientists, students, activists, journalists and others who aim to promote human rights, equality, and global and local justice."

New methods and research address nations' capacities to alter the current situations as seen, for example, in the work on 'Sustainable Governance Indicators' by Bertelsmann Stiftung. The World Values Survey, organized as a global network of social scientists and coordinated by the World Values Survey Association, conducts surveys that address worldwide socio-cultural and political change. The aspiration for a more just, sustainable, and humane world finds some support in global data from that survey along with that from the Pew Research Center's Global Attitudes Project, where, for example, one finding is a global concern for growing economic inequality, with most people agreeing that the economic system favors the wealthy, data that address Miller's concern for consulting ordinary citizens in developing theories of social justice. Whether and how such research and networking can ultimately support greater theorizing about, and the advancing of, social justice remains to be seen.

See also: Capitalism: Global; Citizenship, Sociological Aspects of; Class: Social; Equality of Opportunity; Genetics and Social Justice; Income Inequality; Inequality, Social; Labor, Division of; Movements, Social; Nation, Sociological Aspects of; Social Action and Social Justice; Social Changes: Models; Social Influence, Sociology of; Social Justice and Criminal Justice; Social Justice and Social Inequalities; Social Structure; Solidarity, Sociology of; State, Sociology of The; Values, Sociology of.

Bibliography

Aldarondo, E. (Ed.), 2007. Advancing Social Justice through Clinical Practice. Lawrence Erlbaum Associates, New Jersey.

Banerjee, A., Duflo, E., 2011. Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty. Public Affairs, New York.

Bell, D.A., de-Shalit, A. (Eds.), 2003. Forms of Justice: Critical Perspectives on David Miller's Political Philosophy. Rowman & Littlefield, Lanham, MD.

Burke, T.P., 2008. The origins of social justice: Taparelli d'Azeglio. First principles. ISI Web Journal. http://www.firstprinciplesjournal.com/articles.aspx?article=1760&loc=qs (accessed 16.11.13.).

Cohen, R.L. (Ed.), 1986. Justice: Views from the Social Sciences. Plenum Press, New York.

Corning, P., 2011. The Fair Society: The Science of Human Nature and the Pursuit of Social Justice. The University of Chicago Press, Chicago, IL.

Craig, G., 2007. Social justice in a multicultural society: experience from the UK. Studies in Social Justice 1 (1), 93–108.

Craig, G., Burchardt, T., Gordon, D., 2008. Social Justice and Public Policy: Seeking Fairness in Diverse Societies. The Policy Press, Bristol, UK.

Cramme, O., Diamond, P., 2009. Social Justice in the Global Age. Polity, Cambridge, UK.

Feagin, J., Vera, H., 2008. Liberation Sociology, second ed. Paradigm Publishers, Boulder, CO.

Gautney, H., Dahbour, O., Dawson, A., Smith, N., 2009. Democracy, States, and the Struggle for Global Justice. Routledge, New York.

Maas W.K., 2008. Structural violence. In: Encyclopedia of Violence, Peace, & Conflict, vol. 3. Elsevier, Oxford, UK, pp. 2004–2011.

Miller, D., 1999. Principles of Social Justice. Harvard University Press, Cambridge, MA.

Miller, D., 2003. A response. In: Bell, D.A., de-Shalit, A. (Eds.), Forms of Justice: Critical Perspectives on David Miller's Political Philosophy. Rowman & Littlefield, Lanham, MD, pp. 349–372.

Miller, D., 2009. Social justice versus global justice? In: Cramme, O., Diamond, P. (Eds.), Social Justice in the Global Age. Polity, Cambridge, UK, pp. 23–37.

Nussbaum, M., 2011. Creating Capabilities: The Human Development Approach. The Belknap Press of Harvard University Press, Cambridge, MA.

Okin, S.M., 1987. Justice and gender. Philosophy and Public Affairs 16, 42–72.

Perrons, D., 2009. Spatial and gender inequalities in the global economy: a transformative perspective. In: Cramme, O., Diamond, P. (Eds.), Social Justice in the Global Age. Polity, Cambridge, UK, pp. 194–218.

Rand, S., 2006. Teaching law students to practice social justice: an interdisciplinary search for help through social work's empowerment approach. Clinical Law Review 13 (1), 459–504.

Rawls, J., 2001. In: Kelly, Erin (Ed.), Justice as Fairness: A Restatement. The Belknap Press of Harvard University Press, Cambridge, MA.

Reilly, N., 2009. Women's Human Rights: Seeking Gender Justice in a Globalizing Age. Polity, Cambridge, UK.

Scherer, K., 1992. Justice: Interdisciplinary Perspectives. Cambridge University Press, Cambridge, UK.

Sen, A., 2004. Elements of a theory of human rights. Philosophy & Public Affairs 32 (4), 315–356.

Sharp, G., 1973. The Politics of Nonviolent Action. Porter Sargent, Boston, MA.

Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.

Young, I.M., 2000. Inclusion and Democracy. Oxford University Press, Oxford, England.

Young, I.M., 2011. Responsibility for Justice. Oxford University Press, New York.

Relevant Websites

www.sgi-network.org – Bertelsmann Stiftung.
http://www.pewglobal.org/ – Pew Research Center.
http://www.sociologynetwork.org/ – The Sociological Think Tank.
http://www.un.org/womenwatch/osagi/pdf/e65237.pdf – United Nations.
http://www.weforum.org/ – World Economic Forum.
http://www.worldvaluessurvey.org/index_organization – The World Values Survey.

Social Learning and Culture in Nonhuman Organisms

Rachel L Kendal, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article reviews the field of animal social learning and culture, covering how and why it is studied in both captive and wild animals, by researchers with often differing agendas. Our increasing understanding of the adaptive value of social learning and culture is discussed as well as the new avenues for research this opens. The interdisciplinary nature of the field is emphasized, which, among other areas, influences theories regarding modularity in the brain, the evolution of social intelligence, and animal conservation and welfare.

Whether or not nonhuman animals can be said to have 'culture' is a contentious issue that hangs on the definitions of culture and the complexities of animal social behavior. However, it is well established that numerous species have the ability to acquire from others skills, calls, and information concerning predators, mates, or resources. In some instances, this information transmission is sufficient to propagate the diffusion of novel behavior patterns through animal populations and to maintain distinct behavioral traditions between different populations of the same species. These traditions are sometimes regarded as analogous to human culture in many respects and have been described as culture. However, the use of this term indicates little more than that the behavior concerned has been learned socially and does not necessarily imply that the animals concerned are unusually intelligent, or that they have exceptional linguistic or cognitive capabilities.

Social Learning in Animals

Underlying culture in nonhuman animals is a capacity for 'social learning.' The term social learning refers to learning that is influenced by observation of, or interaction with, another animal (typically of the same species) or its products. Although social learning, as an umbrella term, may include active teaching (Hoppitt et al., 2008) or explicit communication, researchers of nonhumans usually mean no more by the term than the acquisition of information from others who inadvertently emit it while going about their business. Social learning is frequently contrasted with individual learning (also known as asocial learning, trial-and-error learning, and innovation), in which animals learn exclusively on the basis of their own experience, without recourse to social cues or guidance. In any instantiation of social learning, the learner or receiver of information is generally described as the 'observer,' whereas the transmitter of information is called the 'demonstrator.' Social learning occurs when the transmission of information between two animals results in the observer learning from the demonstrator. Frequently, the observer learns through social interaction to perform the same behavior as that exhibited by the demonstrator and, if repeated among other animals, this process can result in the spread of a particular behavior pattern and behavioral homogeneity in a population. Less frequently, animals may learn by observation to perform a different behavior from that exhibited by the demonstrator.

Social learning among animals has been of interest to scientists from a number of different disciplines. Psychologists traditionally have been interested in animal learning. Contemporary animal learning theory describes learning in animals according to three types of experience that cause a change in behavior: exposure to a single stimulus (stimulus (S) learning, e.g., habituation or sensitization); to an association between two external stimuli (stimulus–stimulus (S–S) or classical conditioning); or to an association between an external stimulus and an animal's behavioral response (stimulus–response (S–R), response–reinforcer (R–S) learning, or operant/instrumental conditioning). Arguments abound as to whether there is anything different or special about social, compared with asocial learning, and whether social learning can be adequately explained in conditioning terms. However, recent evidence does indicate that social and asocial learning depend upon the same learning mechanisms (reviewed by Heyes, 2011), rather than the alternative, that social learning is dependent upon a set of special-purpose modules evolved many times in different taxonomic groups as adaptive specializations for group living. This evidence includes findings that social learning and asocial learning covary. In birds and primates, social and asocial learning abilities are correlated positively across species even when accounting for differences in body mass, brain volume, phylogeny, and research effort. Likewise, within zebra finches and pigeons, asocial and social learning is correlated. Social learning abilities have also been found in solitary species, the common octopus and red-footed tortoise, which would not be expected if social learning evolved independently to asocial learning, and in response to the selection pressures of social living. Finally, social learning has the same key features across diverse species, including humans. That is, across species, the main types of social learning occur under the same observable conditions as the three main types distinguished for asocial learning (see Section Laboratory Studies of Animal Social Learning). Thus, social and asocial learning depend upon the same learning mechanisms that are taxonomically general rather than specialized. Indeed, Heyes (2011) recently claimed that, at the cognitive level, all that is 'social about social learning' is the biasing or tuning of input mechanisms (perception, attention, and motivation) toward other agents, as the actual long-term encoding of information is mediated

by general learning mechanisms. Transmission biases (or social learning strategies, see Section The Adaptive Value of Animal Social Learning) may be viewed in this light as socially biased input mechanisms that privilege information received via a social versus asocial channel as well as via specific social channels.

Social learning has also been of interest to ethologists and behavioral ecologists because it seems to allow animals to learn about their environments rapidly and efficiently, without having to engage in potentially costly or hazardous learning trials, or expend considerable time and energy exploring the environment. Animals can learn which foods to eat, acquire food processing skills, learn to identify predators, learn which members of the opposite sex to mate with, or develop songs and calls by exploiting the knowledge base of more experienced conspecifics (Heyes and Galef, 1996; Zentall and Galef, 1988). Social learning can be regarded as a shortcut to learning about the environment, but may incur a cost if the acquired knowledge is inappropriate or outdated (Giraldeau et al., 2002). Evolutionary biologists, biological anthropologists, and archaeologists have also studied animal social learning because it is accorded a prominent role in the evolution of human culture. Although definitions of human culture vary, one central feature is the transmission of acquired information between individuals through social learning processes. Many researchers believe that an understanding of the ways in which social learning operates in animal populations will generate insights into the evolutionary roots of human culture (Laland and Galef, 2009). Social learning is regarded as pivotal to a number of social intelligence hypotheses, which maintain that the evolution of high intelligence or large brains was driven by a capacity for social learning and culture (Byrne and Whiten, 1988). Linked to such theories is the suggestion by many psychologists that more complex forms of social learning might be indicative of the animal concerned possessing unusual or sophisticated psychological capabilities (Heyes, 1994). For instance, it has been suggested that for an animal to accurately imitate the motor patterns of another animal it must be conscious, must be capable of taking the other animal's perspective, must be unusually intelligent, or must be capable of intentional action. Although such claims have been hotly disputed, social learning is inextricably linked to other factors believed pivotal in the evolution of humanity's sociocognitive niche (see Section Social Learning and Social Intelligence).

History of Research into Animal Social Learning

The study of social learning in animals dates back to Darwin and flourished as a result of attempts to demonstrate mental continuity between humans and animals in order to prove that humans had evolved. For many Victorian scientists, demonstrations of social (then termed 'imitative') learning in animals were seen as providing evidence of an evolutionary origin of the higher mental faculties of humans. Imitation was regarded as intermediate in psychological complexity between animal instinct and human reasoning (Boakes, 1984). The verb 'to ape' and the phrase 'monkey see monkey do' are manifestations of the then commonly held view that simian primates are frequent imitators of each other's behavior.

Numerous observations of imitative behavior in primates and other animals were reported in the late nineteenth century, but attempts to demonstrate imitation proved problematic. It became apparent that there are many processes that can result in two animals expressing similar behavior. Researchers began to realize that anecdotal stories concerning animals in uncontrolled environments would not provide satisfactory evidence to address this issue. In the first decades of the twentieth century, Edward Thorndike, a leading psychologist, attempted to introduce some scientific rigor into investigations by carrying out controlled experiments exploring imitation in cats, dogs, and monkeys (Galef, 1988). Thorndike's experiments led him to the controversial conclusion that there was no evidence that animals could 'from an act witnessed learn to do an act.' Out of the controversy there emerged two distinct research traditions in the study of social learning: a laboratory experimental tradition, dominated by experimental psychology, and a second tradition that carried out observations of natural populations of animals, dominated by ethologists and primatologists. The laboratory approach has focused on investigating the psychological mechanisms underlying social learning. In contrast, field studies have reported behavior patterns spreading through animal populations or cultural traditions. More recently, researchers developed experimental approaches to the study of the spread of learned information through animal populations, including transmission chain and diffusion studies (Laland et al., 1993; Whiten and Mesoudi, 2008). In addition, mathematical approaches have explored the conditions under which social learning is expected to evolve, and the evolutionary consequences of social learning. Finally, the field has recently seen a marrying of empirical and ethological approaches in the development of statistical methods enabling robust investigation of social learning in wild animals (reviewed in Kendal et al., 2010; see Galef, 1988 for more on the history of research into animal social learning.)

Laboratory Studies of Animal Social Learning

Modern laboratory experiments investigating social learning in animals commonly train animals (called demonstrators) to perform a target behavior (e.g., to push a lever to get a food reward), pair these demonstrators with untrained experimental subjects (called observers) that watch the demonstrator perform the behavior, and then test the observers alone to establish whether they have learned to perform the target behavior (Galef, 1988). Frequently, the rate at which the subjects learn to perform the behavior is compared with that of the control subjects, which have received different forms of observational experience. Such experiments have established that there are a number of ways that animals can learn socially, the variants reflecting the nature of the observer's learning and the type of information transmitted from demonstrator to observer. For instance, in the lever-pressing example, the observers might learn at an accelerated rate relative to controls (1) because they learned of the presence or significance of the lever, (2) because they learned of the presence of food, (3) because their observational experience allowed them to form an association between the depressed lever and the appearance of food, or (4) because they formed

an association between the action of pressing the lever and the appearance of food. In addition, there are other nonsocial learning processes that could result in an elevation in the observers' performance, for example, if observation changes the observer's motivational state.

Such experiments have engendered a fissioning of social learning phenomena, resulting in a multitude of terms to describe the different processes that can result in social learning and spawning a host of classificatory schemes. At present, there is only partial consensus over the terminology. One reason for the lack of progress in consensus is that early classifications stemmed from the goal of providing a list of mechanisms to be ruled out by researchers if they are to claim 'true' imitation (that is matching of specific motor patterns between observer and demonstrator). Thus, until recently, relatively little attention was paid to distinguishing among the many potential nonimitative processes. In addition, some terms are defined using criteria that are not directly observable (i.e., processes inside individuals such as 'goals' or 'intentionality') or that are not amenable to experimental manipulation in the laboratory.

Galef (1988) identified seven processes that could result in social learning, whereas in the most recent classification scheme by Hoppitt and Laland (2008), nine social learning processes are identified. Yet, other researchers may well add to this number of processes through further delineations. For simplicity sake, the scheme of Whiten and Ham (1992) is outlined in which four classes of social learning are distinguished, which can be mapped onto asocial learning mechanisms (Heyes, 1994). 'Stimulus enhancement' occurs when the observer learns from the demonstrator to which object or location its behavior should be oriented. Thus, observation of another individual exposes the observer to a single stimulus (S-learning) leading to a change in the observer's behavior. For instance, British birds learned to peck open milk bottles, regardless of their location, and drink the cream by having their attention drawn to milk bottles by other feeding birds (Hinde and Fisher, 1951). (The term 'local enhancement' is sometimes used to describe similar phenomena, whereby observers become more likely to visit or interact with objects at a particular location.) 'Observational conditioning' occurs when the observer learns to what circumstances a behavior should be a response. Thus, another individual's behavior exposes the observer to a relationship between two external stimuli (S–S learning), leading to a change in the observer's behavior. For example, infant rhesus monkeys learned to respond to the presence of a snake with fear and avoidance behavior through observations of the screams and avoidance behavior of adults when exposed to snakes (Mineka and Cook, 1988).

Third and fourth in Whiten and Ham's (1992) scheme were 'imitation' and 'goal emulation' which, unless functioning in disregard of the outcome for the demonstrator, fall under the broader term of observational learning. Observational learning occurs when the observer learns of a relationship between a behavior and its outcome (response–reinforcer R–S learning). Thus, another individual's behavior facilitates stimulus–response (S–R) learning resulting in a change in the observer's subsequent behavior. Within this class of social learning processes is imitation, which was said to occur when the observer learns from the demonstrator to perform a specific motor pattern in a particular context. For instance, orangutans had been reported to go through the motions of washing clothes following observation of human caretakers who were doing the laundry. More recently, we have seen further delineations such as 'contextual imitation' and 'production imitation' (Hoppitt and Laland, 2008), which take into account the novelty of the behavior copied, the accuracy of motor pattern replication, and sequence learning.

Whiten and Ham's (1992) final process was goal emulation, which referred to instances where the observer learns from the demonstrator the goal to pursue or the result that can be achieved. This definition is problematic in that is refers to the observer's understanding of the demonstration, which is unobservable to the experimenter. For example, juvenile chimpanzees that observed a stick being used to rake in out-of-reach food appeared to learn that the goal of the action was to gain food using a stick but developed their own means of doing so (Tomasello et al., 1987). More recently, the term goal emulation has been subsumed under the broader term 'emulation,' which generally refers to instances where the observer copies the results of a demonstrator's behavior rather than the actions themselves (Hoppitt and Laland, 2008). It is recognized that there are many ways in which this may be achieved including 'object movement reenactment' whereby observers recreate the movements of the objects manipulated by the demonstrator. For example, pigeons, chimpanzees, and children are able to solve a simple task when observing manipulandi that move as if by a ghost, with no agent visible (Hopper, 2010). Finally, 'affordance learning' whereby observers learn what can be done with an object/manipulandi is variously defined as emulation (Hopper, 2010) or observational conditioning (Hoppitt and Laland, 2008). It should be noted that more than one of these social learning processes, as well as asocial learning, may come into play in any single learning event. (See Hoppitt and Laland (2008) for additional information regarding evidence for the outlined social learning processes and several that have been omitted such as social enhancement of food preferences, response facilitation, and social facilitation.)

In general, there remains a lack of consensus as to whether any of the current classification schemes are comprehensive or whether each of a given classification's categories is mutually exclusive. The plethora of competing terms and usage reflect an ignorance of the biological basis of social learning. At present, there is little understanding of the genetic, neural, or physiological processes that underlie the processing of sensory information, the associationist or transformational organization of this information, and the production of motor patterns, during social learning. There has, however, been an increased understanding of the neural circuitry of imitation in recent years. The neural correlates of vocal imitation in song birds have been extensively studied for many years, but it was not until the discovery of 'mirror neurons' in macaque monkeys (Rizzolatti and Craighero, 2004) that the neural mechanisms of motor imitation received attention. These neurons are found in the ventral premotor cortex (area F5) and in the inferior parietal lobule (area PF). Both of these brain areas are interconnected and also connect with the superior temporal sulcus, which contains visual neurons that respond to seeing the actions of others. The mirror neurons are active

when monkeys perform goal-directed actions themselves and when they observe someone else performing that action. This circuitry, which enables the coding and mapping of observed actions onto the motor repertoire of the observer, would seem important for imitation. Indeed, in humans, areas 44 and 45 of the left inferior frontal gyrus, a region subsumed by Broca's area and the homolog of monkey F5, is specialized for imitation according to functional magnetic resonance imaging studies. As further detailed by Iacoboni (2005) this 'core circuitry' communicates with other neural systems depending upon the type of imitation performed (imitative learning or social mirroring); mirror neurons in monkeys and humans code the likely future actions of others enabling the anticipation of others' intentions; and mirror neurons in monkeys can change their activation responses according to observational experience, potentially enabling the learning of novel actions as required in imitation.

There is little doubt that the laboratory experimental approach has greatly enhanced the understanding of the processes that can result in social learning and their underlying mechanisms, and developed the methodologies and conceptual tools to further that understanding.

Studies of Animal Culture

The question of whether nonhuman animals exhibit culture, and the importance of this for understanding human behavior, depends fundamentally on definitions of culture. Definitions range from those that deem a species cultural if it merely exhibits traditions transmitted via social learning to those requiring that transmission involve teaching, imitation, or the generation of group-specific norms and ethnic markers. One point of consensus is the necessity, if not sufficiency, of social learning in underpinning culture. Thus, studies of social learning in nonlaboratory animals, in their natural habitat, become paramount. One of the earliest such studies was carried out in a population of Japanese macaque monkeys, where a young female began washing sweet potatoes provisioned on a beach before eating them. This previously unobserved behavior was subsequently seen to increase in frequency in the troop, with 90% of individuals eventually adopting the potato-washing behavior, presumably by learning from each other. There are hundreds, perhaps thousands, of such reports of the spread of novel behavior patterns in populations of animals. Well-known examples include the transmission of dietary preferences by rats attending to cues on each other's breath, the transmission of predator recognition in birds, mammals, and fish, the learning of vocal dialects in song birds, feeding on pine cones among black rats, and tool-using traditions in primates (see Bonner (1980) and Laland and Galef (2009) for readable overviews).

Probably, all vertebrates and many invertebrates will eventually be found to exhibit some form of social learning. Of note, these reports are observations of acquired behavior patterns spreading through single populations. In other cases, distinct cultural variations may characterize the behavioral repertoires of different populations of the same species. For instance, much birdsong is thought to be learned from older males, and frequently regional song dialects characterize neighboring populations. Undoubtedly, the most impressive case of local cultural traditions is provided by the different tool-using repertoires of populations of common chimpanzees. For instance, ant-dipping (using a stalk to fish for ants) is reported in some but not all chimpanzee populations and is thought to be maintained within populations through social learning processes. A now famous study documented 39 different behavior patterns, including tool usage, grooming, and courtship behaviors, that were habitual in some chimpanzee communities and absent in others, where ecological and genetic explanations had been discounted or deemed implausible (Whiten et al., 1999). Subsequently, there have been a number of high-profile, and oft-debated, reports of 'culture' in wild animals using this method, including in bonobos, macaques, capuchin monkeys, cetaceans, and orangutans.

These field reports are of great value in documenting the breadth of animal cultures and throwing light on the contexts in which social learning occurs. Indeed, without an understanding of social learning in the context in which it evolved, we cannot hope to elucidate the interaction between biological and cultural evolution. However, it is difficult in the field to establish reliably whether or not a behavior genuinely is socially learned (as opposed to learned independently by each animal), to determine whether local traditions reflect different cultural repertoires or varying ecological resources, or to investigate the psychological and social processes underlying the diffusion (Laland et al., 2009). The result has been a heated debate, which some have gone so far as to describe as 'war' (McGrew, 2002), about the reality of purported cultural behaviors in animals. Take, for example, the aforementioned case of chimpanzee ant-dipping. At Gombe in Tanzania, a long wand is inserted into nests, withdrawn as the ants swarm up it, and then run through the hand forming the ants into a ball that is then placed in the mouth. In contrast, in the Tai forest of the Ivory Coast a short stick is used that is then pulled directly through the mouth. The question arises as to why the Tai chimps use this seemingly less efficient technique when sticks of varying lengths are available at both sites. Is the difference purely cultural? Researchers studying at a third site, Bossou, discovered that the technique used was correlated with the species of ant being captured (Humle and Matsuzawa, 2002). They discovered, through using themselves as guinea pigs, that black ants swarm up sticks more aggressively and give more painful bites than red ants. Not surprisingly then, chimpanzees at Bossou use long sticks for black ants and short ones for red ants. These findings suggest caution is required when inferring a cultural explanation, but they do not indicate whether dipping techniques reflect differing ant prey at Tai and Gombe. Indeed, a recent study, comparing ant-dipping at 13 African sites, suggested that variation in techniques cannot be explained by the ants present (Schoening et al., 2008). For some, whether or not ant-dipping varies according to prey ferocity is beside the point: Byrne (2006) believes that the hallmark of culture is intricate complexity of behavior, and as the skill of ant-dipping is intricate and unlikely to be learned in its entirety by a solitary chimpanzee, there is an implied role for social learning.

Thus, hard evidence that putative animal traditions are socially learned is required. Researchers have invested their energies in two complementary approaches to remedy this situation: captive transmission chain and diffusion studies, and

statistical methods that may be applied to studies of wild animals. Attempts to investigate animal cultural traditions experimentally have moved away from the demonstrator–observer dyadic approach toward group-level studies under controlled conditions in captivity. 'Diffusion studies' are characterized by the introduction of a trained demonstrator to a population and a careful monitoring of the increase in the number of individuals displaying the target behavior over time (Lefebvre and Palameta, 1988). 'Transmission chain studies' involve the gradual replacement of a founder population of trained demonstrators with untrained animals (replacements representing the turnover of pseudogenerations) and explore whether the target behavior persists in the population when the trained founders have been removed (Laland et al., 1993). For example, studies of captive chimpanzees, involving the use of novel extractive foraging tasks (or 'artificial fruits'), have shown they can sustain foraging traditions, across pseudogenerations, through high-fidelity social learning (ensuring the behavior remains unaltered over time) and even transmit traditions from one group to another (Whiten et al., 2007).

Many such studies have been performed with a variety of species (reviewed by Whiten and Mesoudi, 2008) and have been valuable in determining how processes of social interaction impede, or propitiate, the diffusion of information. It became apparent that the structure of animal populations (dominance relations, familiarity of individuals, spatial positioning, sex, age, and size differences) strongly influences the pathways by which learned information spreads, a phenomenon known as 'directed social learning' (Coussi-Korbel and Fragaszy, 1995). For instance, novel skills often spread at different rates among juveniles and adults, among the two sexes, or within and between family groups (see Section The Adaptive Value of Animal Social Learning). However, although these findings add credence to the supposition that social learning is behind intersite variation in behavior, and the factors that may influence transmission, they do not tell us whether social learning can be implicated in natural behavioral variation. This led to researchers identifying social learning more directly in the wild using innovative experiments (reviewed in Reader and Biro, 2010), quantitative analyses (Sargeant and Mann, 2009; Whitehead, 2009), and proxy measures of social learning, such as patterns of affiliation and observation. For instance, both in capuchin monkeys and chimpanzees, there is evidence that individuals selectively watch others from whom they can learn skills they have yet to master (Biro et al., 2003; Perry et al., 2003).

More recently, researchers have begun to develop statistical methods for assessing the likelihood of asocial learning being responsible for an observed pattern of behavior, permitting inference of social learning when asocial learning has been deemed improbable (Kendal et al., 2009b; Franz and Nunn, 2009; reviewed in Kendal et al., 2010). Others have fit mathematical models to phylogenetic data to tease apart the role of genetics and culture in population-level behavioral variation (Lycett, 2010). These methods have been applied to captive populations as well as to wild lemurs, capuchin monkeys, chimpanzees, cetaceans, song birds, meerkats, and other species. Although the field has come far in recent years in being able to identify social learning in the wild, we still do not know what social learning processes are actually involved (but see Kendal et al., 2007; Hoppitt et al., 2012) and the role of factors such as directed social learning. However, as many of these new methods take advantage of social network theory, they will facilitate investigation of directed social learning and/or transmission biases (see Section The Adaptive Value of Animal Social Learning) in the spread of information within and between populations.

Typically, the new behavior pattern found in an animal culture will have been invented by a single individual. There has been comparatively little investigation of the process of innovation. It is not clear whether innovation should be characterized as a personality trait (associated with clever, creative, neophilic, or nonconformist individuals), a state-dependent variable (e.g., hunger may drive foraging innovation), or resulting from environmental changes (such as a drought) (Reader and Laland, 2003). Studies in fish (Laland and Reader, 1999), meerkats (Thornton and Samson, 2012), and primates (Reader and Laland, 2001) suggest that the adage 'necessity is the mother of invention' may be appropriate, with innovation appearing at elevated frequencies in hungry, low status, and small individuals. In contrast, a study of Callitrichid monkeys found that while younger individuals were more inquisitive, it was the older individuals who had the competence to innovate (Kendal et al., 2005a). Across primate species, incidence of innovation covaries with incidence of social learning (also with tool use and deception), suggesting that innovation and social learning may be manifestations of a general capacity for behavioral plasticity (Reader and Laland, 2002). Likewise, studies of birds indicate that innovation, or behavioral plasticity, is correlated with invasion success of species and thus their ability to overcome anthropogenic change to environments (Sol et al., 2002). However, within a given species it is not known whether innovative individuals tend to be good, average, or poor social learners. One recent study has, however, shown that in pigeons, at least, social and asocial learning abilities are positively correlated such that the two abilities are not traded off within individuals (Bouchard et al., 2007).

The Adaptive Value of Animal Social Learning

Theoretical analyses have explored the circumstances under which natural selection will favor reliance on social learning, as opposed to asocial learning or evolved nonlearned behavior (see Laland et al. (1996), for an overview of this literature). A consensus has been reached that the issue hangs, in part, on patterns of variability in the environment. In an environment that is changing comparatively slowly, or that exhibits relatively little spatial heterogeneity in resources, populations are able to evolve appropriate behavior patterns through natural selection, and learning is of little adaptive value. In contrast, in rapidly changing or highly variable environments, asocial learning pays. Here, natural selection cannot track environmental fluctuations quickly enough, while social learning is unreliable because it may lead individuals to acquire outdated or locally inappropriate behaviors. It is intermediate rates of change and patterns of spatial heterogeneity that favor social learning, and within this window vertical transmission of information (social learning by offspring from parents) is thought to be an adaptation to slower rates of change than horizontal

transmission (social learning among unrelated individuals of the same cohort).

Although these findings are theoretically robust, it is only recently that they have begun to be validated with empirical studies of animals (reviewed in Kendal et al., 2005b, 2009). For instance, we did not know whether contemporary populations of animals capable of social learning actually do switch between social and asocial learning depending on the pattern of environmental variability they experience in pertinent resources. Providing a partial answer to this question is a study of wild-caught sticklebacks where the fish ignored social information provided by other foraging conspecifics when their personal foraging information was highly reliable (representing a nonvariable environment), but were more likely to copy others as their own information became outdated and thus potentially unreliable (van Bergen et al., 2004).

More generally, attention has recently turned to what other factors affect reliance on social learning. For instance, is social learning more likely when individuals are confronted with an unfamiliar problem, or a threatening environment, or do individuals preferentially acquire social information from others with certain characteristics such as older, higher ranked, or successful individuals? A plethora of studies have been carried out with a variety of species investigating these 'cultural transmission biases' (Boyd and Richerson, 1985), which have also been termed 'social learning strategies' indicating when, what, and whom individuals should copy (Laland, 2004). For example, in the most recent collation of such studies (Rendell et al., 2011), there is evidence of 'context-dependent biases' regarding (1) the state of the observer (e.g., copy when uncertain or in a reproductive state), (2) the characteristics of the model (e.g., copy old, large, high status, or familiar individuals), and (3) the frequency of the to-be-copied behavioral trait in the population (e.g., copy if rare or in the majority). As many of these studies are conducted in highly controlled laboratory-type studies, we do not know, however, the extent to which these transmission biases function independently or in concert in freely living animals and how they influence the cultural patterns observed in wild animals. Preliminary evidence with chimpanzees indicates that multiple biases (including copy when uncertain and of low status, copy high-ranked and expert individuals) may be in play in the transmission of a single simple foraging tactic. Moreover, it is possible that transmission biases such as copy highly ranked individuals or 'conformity' (an exaggerated tendency to copy the majority) may explain intersite variation in chimpanzee cultural repertoires despite the sharing of ecology and migrants (Kendal et al., in preparation).

Social Learning and Social Intelligence

Social learning has been central to a number of social intelligence hypotheses, which suggest that complex social behavior drove brain evolution and intelligence. Allan Wilson's (1985) concept of 'behavioral drive' explicitly emphasizes social learning. Wilson noted a positive relationship between relative brain size and rate of evolution in vertebrates. He explained this by arguing that big-brained species exhibit greater behavioral plasticity, are more innovative, and are more capable of social learning than small-brained species. Cultural traditions, he argued, allow populations of animals to open up new niches and exploit novel resources, thereby exposing themselves to new selection pressures (known as 'niche construction,' Odling-Smee et al., 2003; Sterelny, 2007) and accelerating the rate of fixation of mutations. Wilson's hypothesis is supported by recent evidence that there is a significant correlation between relative brain size and incidence of both innovation and social learning in primates (Reader and Laland, 2002) and birds (Nicolakakis et al., 2003) and the intraspecific correlation of social and asocial learning in pigeons (Bouchard et al., 2007). Likewise, the enhanced cognition of 'enculturated' apes (e.g., Kanzi's language performance), raised by humans who emphasized socially guided exploration opportunities (Tomasello and Call, 2004), may be taken as consistent with this hypothesis.

More recently, researchers have put forth the 'Cultural Intelligence Hypothesis' as opposed to the 'General Intelligence Hypothesis.' For Herrmann et al. (2007), who gave a battery of tests to orangutans, chimpanzees, and 2.5-year-old children, this is evidenced in the fact that all species performed similarly in the domain of physical cognition but the children outstripped the other species when it came to social cognition. Thus, they propose that humans have a set of social cognitive skills (theory of mind, social learning, communication), emerging early in ontogeny, for participating and exchanging knowledge in cultural groups. In a study of what underlies the apparently human unique capability of cumulative culture (the increase in complexity or efficiency of cultural artifacts across generations), Dean et al. (2012), when comparing capuchin monkeys, chimpanzees, and children, also isolated sociocognitive factors including imitation, complex communication, and prosociality. Similarly, Whiten and van Schaik (2007) note a cross-species correlation of possession of complex culture (defined as a repertoire of multiple traditions) and social intelligence supporting culture (e.g., imitation), namely in the great apes. They admit, however, the paradox of robust findings of imitation in birds yet little evidence for possession of cultural repertoires, and vice versa for monkeys.

(For further details regarding the input of comparative studies of cultural cognition and material culture artifacts on understanding the evolutionary origins of humanity's socio-cognitive niche, see Whiten and Erdal, 2012 and Haslam et al., 2009, respectively.)

Implications for Conservation and Animal Welfare

Despite any controversies within the study of animal social learning and culture, the field has important implications for ethics, philosophy, and politics. Continuity of the human and the nonhuman animal mind affects ethical considerations in animal welfare and influences how we characterize humans' place in nature. Similarly, an understanding of animal cultures can be used to enhance the conservation and management of species facing extinction, as well as influencing conservation priorities through increasing awareness of the existence of not only genetic, but also cultural diversity in animals (for a review see Whitehead, 2010).

See also: Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Primate Behavior: Significance for Understanding Humans.

Bibliography

Biro, D., Inoue-Nakamura, N., Tonooka, R., et al., 2003. Cultural innovation and transmission of tool use in wild chimpanzees. Animal Cognition 6, 213–223.

Boakes, R., 1984. From Darwin to Behaviourism. Cambridge University Press, Cambridge, UK.

Bonner, J.T., 1980. The Evolution of Culture in Animals. Princeton Science Library, Princeton, NJ.

Bouchard, J., Goodyer, W., Lefebvre, L., 2007. Social learning and innovation are positively correlated in pigeons (*Columba livia*). Animal Cognition 10, 259–266.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. Chicago University Press.

Byrne, R.W., 2006. Culture in great apes. Philosophical Transactions of the Royal Society of London, Series B 362, 577–585.

Byrne, R.W., Whiten, A., 1988. Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans. Oxford University Press, Oxford, UK.

Coussi-Korbel, S., Fragaszy, D.M., 1995. On the relation between social dynamics and social learning. Animal Behaviour 50, 1441–1453.

Dean, L., Kendal, R.L., Schapiro, S., Lambeth, S., Thierry, B., Laland, K.N., 2012. Identification of the social and cognitive processes underlying human cumulative culture. Science 335, 1114–1118.

Franz, M., Nunn, C.L., 2009. Network-based diffusion analysis: a new method for detecting social learning. Proceedings of the Royal Society B 276, 1829–1836.

Galef Jr., B.G., 1988. Imitation in animals: history, definition, and interpretation of data from the psychological laboratory. In: Zentall, T.R., Galef Jr, B.G. (Eds.), Social Learning. Psychological and Biological Perspectives. Erlbaum, Hillsdale, NJ, pp. 3–28.

Giraldeau, L.-A., Valone, T.J., Templeton, J.J., 2002. Potential disadvantages of using socially acquired information. Philosophical Transactions of the Royal Society of London B 357, 1559–1566.

Haslam, M., et al., 2009. Primate archaeology. Nature 460, 339–344.

Herrmann, E., Call, J., Hernandez-Lloreda, M.V., Hare, B., Tomasello, M., 2007. Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. Science 317, 1360–1366.

Heyes, C.M., 1994. Social learning in animals: categories and mechanisms. Biological Reviews 69, 207–231.

Heyes, C.M., 2011. What's social about social learning? Journal of Comparative Psychology 126, 193–202.

Heyes, C.M., Galef Jr., B.G., 1996. Social Learning in Animals. The Roots of Culture. Academic Press, San Diego, CA.

Hinde, R.A., Fisher, J., 1951. Further observations on the opening of milk bottles by birds. British Birds 44, 393–396.

Hopper, L.M., 2010. 'Ghost' experiments and the dissection of social learning in humans and animals. Biological Reviews 85, 685–701.

Hoppitt, W., Laland, K.N., 2008. Social processes influencing learning in animals: a review of the evidence. Advances in the Study of Behavior 38, 105–166.

Hoppitt, W., Brown, G.R., Kendal, R.L., Rendell, L., Thornton, A., Webster, M., Laland, K.N., 2008. Lessons from animal teaching. Trends in Ecology and Evolution 23, 486–493.

Hoppitt, W., Samson, J., Laland, K.N., Thornton, A., 2012. Identification of learning mechanisms in a wild meerkat population. PLoS One 7 (8), e42044.

Humle, T., Matsuzawa, T., 2002. Ant-dipping among the chimpanzees of Bossou, Guinea, and comparisons with other sites. American Journal of Primatology 25, 551–581.

Iacoboni, M., 2005. Neural mechanisms of imitation. Current Opinion in Neurobiology 15, 632–637.

Kendal, J.R., Kendal, R.L., Laland, K.N., 2007. Quantifying and modelling social learning processes in monkey populations. International Journal of Psychology & Psychological Therapy 7, 123–138.

Kendal, R.L., Coe, R.L., Laland, K.N., 2005a. Age differences in neophilia, exploration and innovation in family groups of Callitrichid monkeys. American Journal of Primatology 66, 167–188.

Kendal, R.L., Coolen, I., van Bergen, Y., Laland, K.N., 2005b. Trade-offs in the adaptive use of social and asocial learning. Advances in the Study of Behaviour 35, 333–379.

Kendal, R.L., Coolen, I., Laland, K.N., 2009a. Adaptive trade-offs in the use of social and personal information. In: Dukas, R., Ratcliffe, J. (Eds.), Cognitive Ecology: The Evolutionary Ecology of Learning, Memory and Information Use. University of Chicago Press, pp. 249–271.

Kendal, R.L., Kendal, J.R., Hoppitt, W., Laland, K.N., 2009b. Identifying social learning in animal populations: a new "optionbias" method. PLoS One 4, e6541.

Kendal, R.L., Galef, B.G., van Schaik, C.P., 2010. Social learning research outside the laboratory: how and why? Learning & Behavior 38, 187–194.

Kendal, R.L., Hopper, L.M., Laland, K.N., Whiten, A., Brosnan, S.F., Lambeth, S.P., Schapiro, S.J., Hoppitt, W.H. Chimpanzees copy dominant and expert individuals, in preparation.

Laland, K.N., 2004. Social learning strategies. Learning & Behavior 32, 4–14.

Laland, K.N., Galef, B.G., 2009. The Question of Animal Culture. Harvard University Press, Cambridge, MA.

Laland, K.N., Reader, S.M., 1999. Foraging innovation in the guppy. Animal Behaviour 57, 331–340.

Laland, K.N., Richerson, P.J., Boyd, R., 1993. Animal social learning: towards a new theoretical approach. Perspectives in Ethology 10.

Laland, K.N., Richerson, P.J., Boyd, R., 1996. Developing a theory of animal social learning. In: Heyes, C.M., Galef Jr, B.G. (Eds.), Social Learning in Animals. The Roots of Culture. Academic Press, San Diego, CA, pp. 129–154.

Laland, K.N., Kendal, J.R., Kendal, R.L., 2009. Animal culture: problems and solutions. In: Laland, K.N., Galef, B.G. (Eds.), The Question of Animal Culture. Harvard University Press, Cambridge, MA, pp. 174–197.

Lefebvre, L., Palameta, B., 1988. Mechanisms, ecology, and population diffusion of socially learned, food-finding behavior in feral pigeons. In: Zentall, T.R., Galef Jr, B.G. (Eds.), Social Learning. Psychological and Biological Perspectives. Erlbaum, Hillsdale, NJ, pp. 141–163.

Lycett, S.J., 2010. The importance of history in definitions of "culture": implications from phylogenetic approaches to the study of social learning in chimpanzees. Learning & Behavior 38, 252–264.

McGrew, W.C., 2002. Ten dispatches from the chimpanzee culture wars. In: de Waal, F.B.M., Tyack, P.L. (Eds.), Animal Social Complexity. Harvard University Press, Cambridge, MA, pp. 41–61.

Mineka, S., Cook, M., 1988. Social learning and the acquisition of snake fear in monkeys. In: Galef, B.G., Zentall, T.R. (Eds.), Social Learning: Psychological and Biological Perspectives. Lawrence Erlbaum, Hillsdale, NY, pp. 51–73.

Nicolakakis, N., Sol, D., Lefebvre, L., 2003. Behavioural flexibility predicts species richness in birds, but not extinction risk. Animal Behaviour 65, 445–452.

Odling-Smee, F.J., Laland, K.N., Feldman, M.W., 2003. Niche Construction: The Neglected Evolutionary Process. Princeton University Press.

Perry, S., Baker, M., Fedigan, L., et al., 2003. Social conventions in wild white-faced capuchin monkeys. Current Anthropology 44, 241–268.

Reader, S.M., Laland, K.N., 2001. Primate innovation: sex, age and social rank differences. International Journal of Primatology 22, 787–805.

Reader, S.M., Laland, K.N., 2002. Social intelligence, innovation and enhanced brain size in primates. Proceedings of the National Academy of Sciences of the United States of America 99, 4436–4441.

Reader, S.M., Laland, K.N., 2003. Animal innovation: an introduction. In: Reader, S.M., Laland, K.N. (Eds.), Animal Innovation. Oxford University Press, Oxford, pp. 3–38.

Reader, S.M., Biro, D., 2010. Experimental identification of social learning in wild animals. Learning & Behavior 38, 265–283.

Rendell, L., Fogarty, L., Hoppitt, W.J.E., Morgan, T.J.H., Webster, M.M., Laland, K.N., 2011. Cognitive culture: theoretical and empirical insights into social learning strategies. Trends in Cognitive Science 15, 68–76.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neuroscience 27, 169–192.

Sargeant, B.L., Mann, J., 2009. From social learning to culture: intrapopulation variation in bottlenose dolphins. In: Laland, K.N., Galef, B.G. (Eds.), The Question of Animal Culture. Harvard University Press, Cambridge, MA, pp. 152–173.

Schoening, C., Humle, T., Möbius, Y., McGrew, W.C., 2008. The nature of culture: technological variation in chimpanzee predation on army ants revisted. Journal of Human Evolution 55, 48–59.

Sol, D., Timmermans, S., Lefebvre, L., 2002. Behavioural flexibility and invasion success in birds. Animal Behaviour 63, 495–502.

Sterelny, K., 2007. Social intelligence, human intelligence and niche construction. Philosophical Transactions of the Royal Society, B: Biological Sciences 362, 719–730.

Thornton, A., Samson, J., 2012. Innovative problem solving in wild meerkats. Animal Behaviour 83, 1459–1468.

Tomasello, M., Davis-Dasilva, M., Camak, L., Bard, K., 1987. Observational learning of tool-use by young chimpanzees. Human Evolution 2, 175–183.

Tomasello, M., Call, J., 2004. The role of humans in the cognitive development of apes revisited. Animal Cognition 7, 213–215. http://dx.doi.org/10.1007/s10071-004-0227-x.

van Bergen, Y., Coolen, I., Laland, K.N., 2004. Nine-spined sticklebacks exploit the most reliable source when public and private information conflict. Proceedings of the Royal Society of London, Series B: Biological Sciences 271, 957–962.

Whitehead, H., 2009. How might we study culture? A perspective from the ocean. In: Laland, K.N., Galef, B.G. (Eds.), The Question of Animal Culture. Harvard University Press, Cambridge, MA, pp. 125–151.

Whitehead, H., 2010. Conserving and managing animals that learn socially and share cultures. Learning & Behavior 38, 329–336.

Whiten, A., Ham, R., 1992. On the nature and evolution of imitation in the animal kingdom: reappraisal of a century of research. Advances in the Study of Behavior 21, 239–283.

Whiten, A., Goodall, J., McGrew, W.C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C.E.G., Wrangham, R.W., Boesch, C., 1999. Cultures in chimpanzees. Nature 399, 682–685.

Whiten, A., van Schaik, C.P., 2007. The evolution of animal 'cultures' and social intelligence. Philosophical Transactions of the Royal Society, B: Biological Sciences 362, 603–620.

Whiten, A., Spiteri, A., Horner, V., Bonnie, K.E., Lambeth, S.P., Schapiro, S.J., de Waal, F.B.M., 2007. Transmission of multiple traditions within and between chimpanzee groups. Current Biology 17, 1038–1043.

Whiten, A., Mesoudi, A., 2008. Establishing an experimental science of culture: animal social diffusion experiments. Philosophical Transactions of the Royal Society B 363, 3477–3488.

Whiten, A., Erdal, D., 2012. The human sociocognitive niche and its evolutionary origins. Philosophical Transactions of the Royal Society B 367, 2119–2129.

Wilson, A.C., 1985. Scientific American 253, 148–157.

Zentall, T.R., Galef Jr., B.G., 1988. Social Learning: Psychological and Biological Perspectives. Erlbaum, Hillsdale, NJ.

Social Learning Theory and Crime

Jeffrey T Ward and Chelsea N Brown, University of Texas at San Antonio, San Antonio, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social learning theory (SLT) is a leading explanation of criminal behavior which maintains that crime is learned and more likely to occur when individuals differentially associate with people who are criminally involved, experience greater exposure to delinquent models, anticipate or actually receive more rewards and fewer punishments for crime, and have a greater number of definitions favorable to crime. Empirical tests have garnered moderate to strong support for the theory. SLT serves as the foundation for many delinquency prevention and offender treatment programs, and has recently been merged with social structural concepts into a social structure–social learning model.

Social learning theory (SLT) (Akers, 1973) is a logically consistent and empirically testable theory formulated to explain the full spectrum of criminal and delinquent behaviors. The theory is capable of accounting for both generalization and specialization in offending. SLT maintains that criminal behavior is learned in the same manner that any human behavior is acquired. The fundamental ideas of SLT stem from Sutherland's theory of differential association and the principles of operant conditioning, which were first integrated in a multidisciplinary effort by Robert Burgess and Ronald Akers in 1966. In the decade to follow, Akers continued his work with the theory and advanced the 'social learning' terminology. In its current form, SLT maintains there are four central concepts that include differential association, imitation, definitions, and differential reinforcement. The main idea of SLT is that crime is more likely when individuals interact with and are exposed to people who are criminally involved and support delinquent conduct (differential association), experience greater exposure to delinquent models (imitation), anticipate or actually receive more rewards and fewer punishments for delinquent behavior (differential reinforcement), and have a greater number of definitions favorable to committing delinquency (definitions). Studies testing the key propositions of the theory have yielded strong empirical support, especially for the differential association and definitions variables (see Pratt et al., 2010). In addition to being one of the leading accounts of criminal behavior, SLT has important practical implications for crime control strategies and serves as the foundation for several successful offender treatment programs.

The Roots of SLT

SLT is the product of theoretical ideas, which cut across the disciplinary boundaries of sociology and psychology. The general notion that learning plays a pivotal role in the development of criminal and delinquent behavior can most directly be traced back to the works of sociologist Edward Sutherland, whereas the details of the learning process can be found in the writings of behavioral psychologists, most notably of which include B.F. Skinner. In a deductive theoretical integration effort in 1966, Robert Burgess and Ronald Akers took on the task of folding the central ideas of differential association theory into the broader principles of operant conditioning; their efforts resulted in the formation of a new theory known as differential association-reinforcement theory (see Burgess and Akers, 1966).

The foundations of differential association-reinforcement theory (and ultimately SLT) begin with Sutherland's theory of differential association which contains nine propositions as follows:

1. Criminal behavior is learned.
2. Criminal behavior is learned in interaction with other persons in a process of communication.
3. The principal part of the learning of criminal behavior occurs within intimate personal groups.
4. When criminal behavior is learned, the learning includes
   a. techniques of committing the crime, which are sometimes very complicated and sometimes very simple, and
   b. the specific direction of motives, drives, rationalizations, and attitudes.
5. The specific direction of motives and drives is learned from definitions of the legal codes as favorable or unfavorable.
6. A person becomes delinquent because of an excess of definitions favorable to violation of law over definitions unfavorable to violation of law.
7. Differential associations may vary in frequency, duration, priority, and intensity.
8. The process of learning criminal behavior by association with criminal and anticriminal patterns involves all the mechanisms that are involved in any other learning.
9. Although criminal behavior is an expression of general needs and values, it is not explained by those general needs and values, because noncriminal behavior is an expression of the same needs and values (Sutherland, 1947: pp. 6–7).

In sum, these nine propositions make clear that learning is fundamental to the origins of criminal behavior and draws attention to the importance of attitudes (definitions) in determining which individuals will be law violators. However, differential association theory falls short of providing a comprehensive explanation of how individuals learn criminal behavior.

In 1965, Ronald Akers was hired as an assistant professor at the University of Washington and worked alongside Robert Burgess, who was also an assistant professor and extensively

trained in the principles of behavior modification. Akers' vast expertise in criminological theory and Burgess' knowledge of fundamental behavioral learning principles resulted in a collaborative effort to restate and reformulate differential association theory in terms of the principles of operant conditioning. In explicating their theory, Burgess and Akers omitted (proposition 9), merged (propositions 1 and 8), and further specified (propositions 2 through 7) the original propositions contained in Sutherland's theory. Specifically, the reformulated propositions found in differential association-reinforcement theory are as follows:

1. Criminal behavior is learned according to the principles of operant conditioning.
2. Criminal behavior is learned both in nonsocial situations that are reinforcing or discriminative and through that social interaction in which the behavior of other persons is reinforcing or discriminative for criminal behavior.
3. The principal part of the learning of criminal behavior occurs in those groups which comprise the individual's major source of reinforcements.
4. The learning of criminal behavior, including specific techniques, attitudes, and avoidance procedures, is a function of the effective and available reinforcers and the existing reinforcement contingencies.
5. The specific class of behaviors which are learned and their frequency of occurrence are a function of the reinforcers which are effective and available and the rules or norms by which these reinforcers are applied.
6. Criminal behavior is a function of norms which are discriminative for criminal behavior, the learning of which takes place when such behavior is more highly reinforced than noncriminal behavior.
7. The strength of criminal behavior is a direct function of the amount, frequency, and probability of reinforcement (Burgess and Akers, 1966: p. 146).

In addition to maintaining an emphasis on social interaction and definitions as critical to the learning process, differential association-reinforcement theory clearly draws attention to the importance of reinforcement for learning criminal behavior. In fact, their sixth proposition claims that the degree to which individuals learn definitions favorable to law violations stems from levels of reinforcement. By using the principles of operant conditioning to reformulate the theory of differential association, Burgess and Akers were able to generate propositions that were more specific and testable. In the years that followed, Burgess turned his attention to other scholarly pursuits while Akers began a long career working with the theory that included further specification and the adoption of the 'social learning' nomenclature (e.g., see Akers, 1973), comprehensive empirical testing (e.g., see Akers et al., 1979), and the development of a cross-level theoretical integration of SLT with social structural explanations of delinquency (e.g., see Akers, 2009).

Akers' SLT

Akers' SLT holds that criminal behavior is learned, maintained, and modified in the same manner as any human behavior. According to the theory, what differentiates delinquent behavior from conforming behavior is the learning content not the learning process itself. While SLT is intended to be very general and encompasses all facets of the learning process, four central concepts form the cornerstone of the theory that include *differential association*, *imitation*, *definitions*, and *differential reinforcement*. Generally speaking, SLT suggests that individuals will be more likely to engage in criminal behavior when they have greater associations with people who engage in criminal behavior, greater exposure to delinquent models, more definitions favorable to criminal involvement, and greater actual or perceived rewards and benefits associated with criminal behavior.

Differential Association

Differential association provides the context in which learning occurs. According to the theory, the most important contexts for learning criminal behavior include peer groups and family units, though varying entities such as schools, neighbors, and media also provide alternative settings where some learning of criminal behavior may ensue. There are two fundamental aspects of differential association, namely the behavioral-interactional and normative dimensions. The behavioral-interactional dimension refers to the direct and indirect interactions and associations with individuals that engage in delinquent activities. The normative dimension refers to the values which one is exposed to as a result of one's associations and interactions with others. It is worth noting that the normative dimension was inherent in Sutherland's original theory; Akers' formulation of SLT keeps the normative dimension but adds the behavioral-interactional dimension to the overarching differential association construct. As differential associations set the learning context, SLT anticipates that the learning mechanisms including imitation, definitions, and differential reinforcement are in large part influenced by one's differential associations. For instance, the theory suggests that individuals with a greater proportion of their peers who engage in criminal behavior will be exposed to more delinquent models, will be subjected to a greater number of definitions favorable to delinquency, and will perceive more rewards and benefits associated with criminal conduct. In turn, these individuals will be more likely to engage in criminal behavior. In other words, the effect of differential association on criminal behavior is hypothesized to, at least in large part, be mediated by the imitation, definitions, and differential reinforcement variables.

Imitation

Imitation, or reproducing the actions of others, is a fundamental learning mechanism influencing behavior of all types across various developmental periods including childhood, adolescence, and adulthood. Imitation in the context of crime occurs when an individual commits a criminal act following the observation of someone else engaging in a similar crime. Observation most typically involves viewing someone's behavior directly in person, but viewing an individual's behavior indirectly through media or receiving verbal or written instructions specifying how to engage in criminal

behavior are alternative forms of observation. Imitation is thought to be most relevant when learning new criminal activities, though it can continue to exert some influence even on previously witnessed behaviors. Importantly though, individuals tend to not mindlessly reproduce the behavior that they observe; rather, characteristics of both the criminal model and the behavior itself weigh into the decision of whether the behavior will be imitated. In addition, beliefs about whether a behavior will be met with rewards as opposed to punishments are also highly relevant in determining if a behavior is likely to be imitated. This means that the learning mechanisms in SLT do not act to influence criminal behavior in isolation. Thus, individuals can technically learn how to commit criminal behavior through observation but elect not to actually commit any criminal acts due to lack of reinforcement and/or acquired definitions unfavorable for crime. It is worth noting that principles of 'observational' learning have been extensively developed elsewhere (see Bandura, 1977). According to Bandura, imitation requires individuals to pay attention to the model, to retain knowledge of what they witnessed, to be physically able to reproduce the behavior, and to be motivated to imitate the behavior. While Akers' theory encompasses these nuances, the proposition at a basic level states that individuals who are exposed to more criminal models will be more likely to engage in crime; and, exposure to criminal behavior that can be imitated is largely a function of one's differential associations.

Definitions

Definitions refer to one's attitudes or beliefs concerning the moral appropriateness or justifiability of criminal behavior. Definitions operate at both a specific and a general level. In other words, people may have general definitions unfavorable to crime but nevertheless engage in a particular type of criminal behavior because they have specific definitions favorable to this form of behavior. For example, a white-collar criminal may believe ordinary street crime to be morally reprehensible but feels that insider trading is justifiable in the realm of business. Another important consideration in SLT is that definitions favorable toward crime can either fully embrace the moral appropriateness of crime (positive definitions) or rationalize or excuse the deviant behavior (neutralizing definitions). In some subcultures, for example, violent behavior is embraced and a 'code of the street' guides comportment in everyday life (see Anderson, 1999). Even more radical are the definitions favorable to violence held by terrorists. While extreme values held by subcultures and extremist groups can generate high levels of criminal behavior, delinquency commonly occurs among those not possessing such positive definitions favorable to crime. Loosely held conventional values themselves open up the possibility for criminal learning, and neutralizations serve to rationalize or justify criminal behavior under certain circumstances.

Individuals form definitions favorable or unfavorable to delinquency through exposure to other's attitudes through their differential associations with delinquents and through the other learning mechanisms including imitation and differential reinforcement. Thus, attitudes toward specific or general acts of crime may be strengthened when one witnesses the behavior of others and believes or finds the criminal behavior to be rewarding. Once acquired, definitions work to influence behavior by making crime an approved or acceptable behavioral option, thereby making an individual more willing to take advantage of a criminal opportunity in a specific type of situation or, perhaps, in situations more generally.

Differential Reinforcement

Differential reinforcement concerns behavioral conduct in response to actual and perceived rewards and punishments associated with engaging in behavior. In general, SLT suggests that individuals who experience and perceive more benefits and fewer costs will have a greater likelihood of engaging in criminal behavior. In other words, criminal behavior is more likely when reinforcement toward criminal behavior is stronger and punishment is weaker. According to the logic of operant conditioning from which SLT heavily draws, reinforcement and punishment can each take two forms – positive and negative – thereby creating four fundamental learning scenarios that modify human behavior. Positive reinforcement involves receiving actual rewards or anticipating receiving rewards (e.g., obtaining goods) for engaging in crime. Negative reinforcement involves escaping punishment or detection. Either way, both positive and negative reinforcement are expected to result in a strengthening of criminal behavior. Conversely, both positive and negative punishment are expected to result in a weakening of behavior according to fundamental learning principles. Positive punishment involves the experience of formal or informal consequences as a result of engaging in the behavior, whereas negative punishment involves the failure to obtain rewards or positive results.

The rewards and punishments associated with crime can be social or nonsocial in nature. Social reinforcers and punishers of criminal behavior can take many forms such as peer approval or acquired goods. Importantly, however, something that might be a social reinforcer for one individual may be a social punisher for another depending on one's differential associations. Hence, peer disapproval would be the likely response to criminal involvement among individuals who associate with prosocial peers; whereas, peer approval would be the expected response to deviancy if one's peers are delinquent. Nonsocial reinforcers (and punishers) can also have important influences on behavior such as the physiological effects of drugs. According to SLT, nonsocial reinforcers tend to interact with the other learning mechanisms. For example, individuals' experiences from the physiological effects of drugs are argued to be colored by the definitions they are exposed to and the reinforcement they receive from their differential associations. Further, reinforcers and punishers that influence behavior need not be tangible or even directly received. Some religious extremists believe jihadist actions (i.e., criminal behavior) will result in rewards following death, whereas more conventional religious believers might expect these same behaviors to result in eternal damnation. In sum, criminal behavior is a direct product of the amount, frequency, and probability of reinforcement, which has been shown to be a complex construct that can take a variety of forms. Overall, the crux of the differential reinforcement concept in SLT suggests that individuals will be more likely to engage in

criminal behavior as positive (and negative) reinforcement is increased and positive (and negative) punishment is decreased.

Empirical Tests of SLT

The collective body of empirical research testing the core hypotheses of SLT provides moderate to strong support for the theory (Pratt and Cullen, 2000; Pratt et al., 2010). Studies assessing social learning examine the effects of one or more of the learning variables (i.e., differential association, imitation, definitions, and differential reinforcement) on criminal outcomes. In a classic first empirical test of SLT, Akers et al. (1979) used data that they collected from 3065 Midwestern middle and high school students to examine the individual and combined effects of all four social learning variables on alcohol and marijuana use. The social learning variables explained a very high proportion of the variance in these antisocial behaviors. Full models including all the four social learning variables explained 54% of the variance in alcohol use and 68% of the variance in marijuana use. Further, additional analyses revealed the relative importance of each of the social learning constructs; the rank ordering from most to least influential included differential association, definitions, differential reinforcement (social/nonsocial), differential reinforcement (social), and imitation. Other studies using these same data from the Boys Town study on alcohol and drug use demonstrate that the relative influence of social learning variables is greater than variables from other theoretical perspectives including social control and strain (Krohn et al., 1984; Akers and Cochran, 1985; Akers and Lee, 1999).

In addition to the Boys Town study, Akers was involved in other major projects including a panel study on adolescent smoking behavior (see Lauer et al., 1982) and another longitudinal study on the drinking behavior of elders (see Akers et al., 1989). In short, several empirical studies stemming from these alternative data sets also find support for the independent and collective effects of social learning variables on delinquency. Outside of Akers' own work, SLT is not typically tested in a comprehensive fashion and the details of the learning 'process' are infrequently the subject of explicit empirical investigation in criminological studies. Put differently, most studies examining SLT tend not to include measures of all the learning variables within the same study. In fact, many studies include only measures of differential association (e.g., peer delinquency) which is highly problematic because it becomes impossible to determine the mechanisms by which others' delinquency impacts a respondent's delinquency. By including only the concept of differential association in empirical tests of social learning, all of the effort by Akers to clarify Sutherland's vague and incomplete statements concerning the process of criminal learning is snubbed. Understanding how differential associations come to influence behavior is of critical theoretical importance (Warr, 2002). There are many potential mechanisms by which groups can influence behavior including, for example, loyalty, fear of ridicule, status, collective processes, and social learning (see Warr, 2002). Not surprisingly, Akers argues these mediating variables can be subsumed into the four existing variables of SLT.

Pratt et al. (2010) recently elucidated the empirical status of SLT through a meta-analysis of roughly 700 effect sizes from 133 studies that assessed the effects of one or more of the social learning variables. Differential association, definitions, differential reinforcement, and imitation represented approximately 55, 20, 20, and 5% of the effects examined in the meta-analysis respectively; hence, differential association is tested quite frequently in social learning models, whereas imitation is hardly ever included as a predictor of delinquent outcomes in criminological research. It is worth noting that SLT is mostly tested examining the influences of delinquent peer associations on delinquent and criminal behavior. The differential association effects of peers tend to be stronger than the differential association effects of other individuals including parents. For the learning variables overall, the mean effect size estimates demonstrate the strongest and most consistent support for differential association (0.225) and definitions (0.218). For these two variables specifically, tests for the effect estimates were statistically significant at least 70% of the time. Comparatively, differential reinforcement effects were statistically significant 56% of the time, whereas imitation effects were significant only 47% of the time. Moreover, the mean effect estimates for differential reinforcement (0.097) and imitation (0.103) were much lower. Taken together, the differential reinforcement and imitation variables have both weaker and less consistent influences on criminal outcomes than the differential association and definitions variables.

Overall, the key hypotheses of SLT receive empirical support, especially those concerning the effects of differential association and definitions on crime and delinquency. Before accepting this conclusion, however, a highly controversial measurement issue having important implications for the empirical validity of differential association warrants mention. There are two primary strategies to assess whether one is in differential association with delinquents. First, indirect measures of peer delinquency involve asking respondents about the behaviors of their peers (as well as about their own behaviors). Second, direct measures of peer delinquency involve asking respondents and their peers to each self-report their own behaviors. Thus, in the direct measurement approach, information about the peer's behavior comes directly from the peer himself or herself. The central concern with perceptual peer delinquency (indirect) measures is that respondents may 'project' their own delinquent tendencies into their perceptions of their peers, thereby artificially inflating the correlation between peer and respondent behaviors (Gottfredson and Hirschi, 1990). Indeed, indirect measures of peer delinquency reveal stronger effects on delinquency than do direct measures (Boman and Gibson, 2011; Meldrum et al., 2009; Weerman and Smeenk, 2005). To skirt this potential for projection bias, several peer delinquency researchers recommend using the direct measures of peer delinquency (peer self-reports), potentially captured via social network data. Akers vehemently maintains that it is the perception of the peer's behavior, not the actual behavior of the peer, which matters most in learning. This epistemological debate concerning the correct measurement strategy is unlikely to be reconciled through empirical means since attempts to construct peer delinquency measures that are perceived accurately enough to avoid potential bias have been unsuccessful (see Boman et al.,

2012a,b). Nevertheless, what is important for SLT's legacy is while the effects of differential association are attenuated when employing direct measures of peer delinquency they are often still of statistical and practical significance.

Practical Applications of SLT

As an explanation of crime, SLT suggests that criminal behavior is learned and is the product of one's differential associations, imitations, definitions, and differential reinforcement. Fortunately, learned behavior can be unlearned too. That is, SLT maintains it is possible for criminals to learn conforming behavior if their differential associations, definitions, reinforcement contingencies, and models are altered to promote prosocial influence. As a more proactive approach, at-risk youth can be targeted for delinquency prevention as well. Systematic reviews suggest that applied criminology that incorporates the fundamental ideas of learning, including cognitive behavioral therapy, is among the most effective approaches to delinquency prevention and offender treatment (Lipsey et al., 2007).

Applications of learning principles to delinquency prevention efforts and offender treatment programs are commonplace. Sometimes learning principles only loosely guide justice policy as is the case with the Gang Resistance Education and Training (G.R.E.A.T.) initiatives (see Akers and Sellers, 2009). The G.R.E.A.T. program consists of law enforcement-guided classroom instruction in schools, taking a preventative approach to the problems of youth delinquency, and gang involvement. Classroom instruction is intended to generate attitudes unfavorable to gang membership and to enhance police–citizen relationships which, in turn, should help to reduce community problems associated with gangs. Perhaps due to weak adherence to the learning principles, the empirical results of G.R.E.A.T. were not so great (see Esbensen et al., 2001). While some program efficacy measures were indicative of success including the development of more negative definitions of gang membership and greater association with prosocial peers, there were no statistically significant effects of the program on the fundamental outcomes of interest. Specifically, participants in the program were not less likely to join a gang and did not engage in lower levels of delinquency than nonparticipants. As a result of these findings, the program has undergone a comprehensive makeover (G.R.E.A.T. II) but preliminary findings of this modified program are only slightly more promising; short-term gang involvement, but not general delinquency, appears to be reduced by the program (Esbensen et al., 2011).

Other programs are more cleanly influenced by social learning concepts such as the Adolescent Transition Program (ATP) from the Oregon Social Learning Center. Focusing on parents and family dynamics, ATP involves 12 therapist-led interactive parent group sessions and 4 individual family meetings designed to enhance parental management skills including monitoring, socialization, and appropriate discipline. The logic of the program is that parents themselves need to learn appropriate techniques to effectively socialize their children. Evaluations of the program have shown that parents do learn positive parenting skills and, as a result, their children's delinquency decreases (Dishion et al., 1992). It is worth mentioning that other programs out of the Oregon Social Learning Center, including the Multidimensional Treatment Foster Care and Interests of Families and Teachers initiatives, also serve as exemplary models of the application of social learning principles to crime prevention and offender treatment effects.

The Social Structure–Social Learning Model

In 1998, Akers published the first edition of his book *Social Learning and Social Structure: A General Theory of Crime and Deviance*, which summarizes SLT and his research endeavors and, importantly, presents a model that incorporates macro- and mesolevel structural factors with SLT. The integrative approach taken this time was not a deductive one but rather sequential or 'end-to-end' one. According to the social structure and social learning model (SSSL) (Akers, 2009), the four social learning variables maintain their proximate effects on behavior as in the original theory. New to the model, however, are four social structural dimensions (i.e., differential social organization; differential location in the social structure; theoretically defined structural variables; and differential social location in primary, secondary, and reference groups). These constructs are hypothesized to influence individuals' behavior through their effects on the social learning variables. Importantly, the model argues that community crime rates are not directly related to social structural forces. Rather, Akers (2009) maintains that differential association, imitation, definitions, and differential reinforcement mediate the effects of social structure on crime rate. In short, social structure sets the context in which individuals learn criminal (or conforming) behavior and crime rates simply reflect how much criminal learning has occurred at the community or societal level.

Differential social organization refers broadly to the social structural correlates of crime including factors such as population density, urbanicity, and region. These features organize communities and societies toward or against crime. *Differential location in the social structure* refers to sociocultural and sociodemographic differences in communities and societies across class, gender, race, age, and religion. At the collective level these variables such as youth concentration or proportion Black serve to identify the locations or statuses of groups in the larger social structure. The *theoretically defined structural variables* concept is a catch-all dimension that encompasses fundamental structural perspectives about the sources of crime, the most important of which include anomie, conflict, and social disorganization. Finally, *differential social location in primary*, *secondary*, and *reference groups* refers to peoples' positions in peer, work, and family groups. These groups serve as the immediate context through which the other social structural dimensions impact individuals and have the most direct link to the social learning variables. Specifically, this construct in particular is closely intertwined with the differential association concept.

In sum, the four dimensions of social structure organize groups, neighborhoods, and societies, thereby setting the context for learning either law abiding or criminal behavior. Importantly, these social structural factors are hypothesized to influence individual behavior through the learning

mechanisms outlined in SLT, meaning that group crime rates signify higher levels of aggregate criminal learning in communities. Overall, post hoc interpretations of previous research not intended as explicit tests of the SSSL model lend support for its basic propositions. However, the empirical status of the model will remain unclear until comprehensive tests of the model are conducted (Akers, 2009).

Conclusion

Standard metrics that judge the quality of theories include parsimony, scope, logical consistency, testability, empirical validity, and practical utility. When SLT is viewed in light of these criteria, the theory holds up well. SLT contains only four key components, can explain a wide range of criminal and delinquent behaviors, has clear concepts, makes straightforward hypotheses, receives empirical support, and is one of the more useful theories for criminal justice policy. Of the four social learning variables – differential association, imitation, definitions, and differential reinforcement – it is the concepts of differential association and definitions that receive the strongest empirical support in the literature. This is quite intriguing and suggests that it is the normative aspects of differential association that are most responsible for antisocial conduct, which are the aspects of the theory that pay direct homage to Sutherland's differential association theory. At the same time, SLT is clear in anticipating interdependence between the learning mechanisms. While the theory also suggests there are some feedback effects of the learning mechanisms, it is clear in its suggestion that there is generally a sequence or directional process at play (Akers, 2009). In light of these considerations, it can be argued that the social learning process has been incompletely studied and there is still much to learn about the empirical status of SLT. This is truer still when considering the recent cross-level theoretical integration with macro- and mesostructural social forces, namely SSSL theory (Akers, 2009). At this juncture, it is safe to conclude that learning plays a nontrivial role in the development of criminal and delinquent behavior. Future research should continue to examine the specifics of learning process as well as the social structural conditions that set the contexts for social learning.

See also: Crime, Sociology of; Criminology: Development and Current Status of the Discipline; Delinquency, Sociology of; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Learning and Memory, Neural Basis of; Prevention of Crime and Delinquency.

Bibliography

Akers, R.L., 1973. Deviant Behavior: A Social Learning Approach. Wadsworth, Belmont, CA.

Akers, R.L., 2009. Social Learning and Social Structure. A General Theory of Crime and Deviance, second ed. Transaction, New Brunswick, NJ.

Akers, R.L., Cochran, J.K., 1985. Adolescent marijuana use: A test of three theories of deviant behavior. Deviant Behavior 6, 323–346.

Akers, R.L., Lee, G., 1999. Age, social learning, and social bonding in adolescent substance use. Deviant Behavior 19, 1–25.

Akers, R.L., La Greca, A.J., Cochran, J.K., Sellers, C.S., 1989. Social learning theory and alcohol behavior among the elderly. Sociological Quarterly 30, 625–638.

Akers, R.L., Krohn, M.D., Lanza-Kaduce, L., Radosevich, M., 1979. Social learning and deviant behavior: a specific test of a general theory. American Sociological Review 44, 636–655.

Akers, R.L., Sellers, C.S., 2009. Criminological Theories: Introduction, Evaluation, and Application. Oxford University Press, New York.

Anderson, E., 1999. Code of the Street: Decency, Violence, and the Moral Life of Inner City. Norton, New York.

Bandura, A., 1977. Social Learning Theory. Prentice Hall, Englewood Cliffs, NJ.

Boman IV, J.H., Gibson, C., 2011. Does the measurement of peer deviance change the relationship between self-control and deviant behavior? An analysis of friendship pairs. Journal of Criminal Justice 39, 521–530.

Boman IV, J.H., Stogner, J.M., Miller, B.L., Griffin III, O.H., Krohn, M.D., 2012a. On the operational validity of perceptual peer delinquency: Exploring projection and elements contained in perceptions. Journal of Research in Crime and Delinquency 49, 601–621.

Boman IV, J.H., Ward, J.T., Gibson, C.L., Leite, W.L., 2012b. Can a perceptual peer deviance measure accurately measure a peer's self-reported deviance? Journal of Criminal Justice 40, 463–471.

Burgess, R.L., Akers, R.L., 1966. A differential association-reinforcement theory of criminal behavior. Social Problems 14, 128–147.

Dishion, T.J., Patterson, G.R., Kavanagh, K.A., 1992. An experimental test of the coercion model: linking theory, measurement and intervention. In: McCord, J., Tremblay, R.E. (Eds.), Preventing Antisocial Behavior: Interventions from Birth through Adolescence. Guilford, New York, pp. 253–282.

Esbensen, F.-A., Peterson, D., Taylor, T.J., Freng, A., Osgood, D.W., Carson, D.C., Matsuda, K.N., 2011. Evaluation and evolution of the gang resistance education and training (G.R.E.A.T.) Program. Journal of School Violence 10, 53–70.

Esbensen, F.-A., Osgood, D.W., Taylor, T.J., Peterson, D., Freng, A., 2001. How great is G.R.E.A.T.?: results from a quasi-experimental design. Criminology & Public Policy 1, 87–118.

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.

Krohn, M.D., Lanza-Kaduce, L., Akers, R.L., 1984. Community context and theories of deviant behavior: an examination of social learning and social bonding theories. Sociological Quarterly 25, 353–371.

Lauer, R.M., Akers, R.L., Massey, J., Clarke, W., 1982. The evaluation of cigarette smoking among adolescents: the Muscatine study. Prevention Medicine 11, 417–428.

Lipsey, M.W., Landenberger, N.A., Wilson, S.J., 2007. Effects of Cognitive–Behavioral Programs for Criminal Offenders. Campbell Systemic Reviews.

Meldrum, R.C., Young, J.T.N., Weerman, F.M., 2009. Reconsidering the effect of self-control and delinquent peers: Implications of measurement for theoretical significance. Journal of Research in Crime and Delinquency 46, 353–376.

Pratt, T.C., Cullen, F.T., 2000. The empirical status of Gottfredson and Hirschi's General Theory of Crime: a meta-analysis. Criminology 38, 931–964.

Pratt, T.C., Cullen, F.T., Sellers, C.S., Windfree Jr., L.T., Madensen, T.D., Daigle, L.E., Fearn, N.E., Gau, J.G., 2010. The empirical status of social learning theory: a meta-analysis. Justice Quarterly 27, 765–802.

Sutherland, E.H., 1947. Principles of Criminology, fourth ed. Lippincott, Philadelphia, PA.

Warr, M., 2002. Companions in Crime: The Social Aspects of Criminal Conduct. Cambridge University Press, New York.

Weerman, F.M., Smeenk, W.H., 2005. Peer similarity in delinquency for different types of friends: a comparison using two measurement methods. Criminology 43, 499–524.

Social Mechanism

Petri Ylikoski, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The basic idea behind mechanism-based explanation is simple at its core, it implies that proper explanations should detail the 'cogs and wheels' of the causal process through which the outcome to be explained was brought about. This idea has become increasingly popular in both the social sciences and philosophy of science over the last two decades. At the core of the mechanistic approach is a criticism of widely held views about social scientific explanation, causation, and the nature of social scientific theories. However, it also has interesting methodological implications for the social sciences.

The Rise of Social Mechanisms

The idea of mechanism-based explanation in the social sciences does not have a single origin. The idea has become popular in recent philosophy of science (Craver, 2007), and has begun to have an influence on many social scientists. However, the rise of mechanisms in the social sciences predates the wide adoption of the idea by philosophers of science. In fact, a number of authors seem to have adopted the idea independently of each other and have utilized it to criticize explanatory practices in the social sciences. For these authors, the idea of mechanism-based explanation has provided a useful platform for criticizing existing research practices and views about the nature of the social scientific enterprise. Due to these diverse origins, there are many different definitions of social mechanisms (Mahoney, 2001). This diversity is understandable, as the idea of a mechanism has been used to support quite different methodological and theoretical positions (Hedström and Ylikoski, 2010). In a short review such as this, it is only possible to mention some of the more prominent advocates of mechanistic ideas.

Jon Elster (1989, 2007) is probably the most influential advocate of mechanisms in the social sciences and his many books are full of examples of mechanism-based thinking in action. For him, the core of this thinking is the idea that mechanistic explanations open up black boxes and show the cogs and wheels of the internal machinery underlying the phenomenon. While much of Elster's work has focused more on psychological rather than social mechanisms, his work has had a huge impact on how social scientists think about explanation.

The critical realism movement (e.g., Sayer, 1992; Archer et al., 1998) has been highly influential with respect to thinking in terms of mechanisms. The original source of its ideas about causation and scientific explanation is Rom Hárre's (1970) pioneering work, but the principal inspiration for critical realists has been Roy Bhaskar (1978, 1979). His critical realism is a tight philosophical package of ontological and epistemological views in which the idea of mechanistic explanation is the most intuitive part.

Within philosophy of the social sciences, both Daniel Little and Mario Bunge have been important spokesmen for mechanistic thinking. Little's textbook (Little, 1991) and articles (Little, 1998) show in an accessible way how thinking in terms of causal mechanisms helps to solve various methodological problems which characterize other approaches to the social sciences. Little also strongly emphasizes the important role of microfoundations in mechanistic explanations. Bunge's (1997, 2004) account of mechanistic explanation is part of his more general systemistic philosophy of science. Like Bhaskar, Bunge's ideas about mechanistic explanations are part of a much larger philosophical system. However, his ideas about mechanisms are quite compatible with what others have said about social mechanisms.

In political science, causal mechanisms have played an important role in debates about research methodology and causal inference (e.g., Mahoney, 2001; George and Bennett, 2005; Gerring, 2008). Many authors have presented mechanistic thinking as alternative to statistical methodologies and as a way to provide foundations for case studies and other small-N studies. As Alexander George and Andrew Bennett (2005) suggest, the idea of process tracing is useful for both development and testing of mechanistic explanations, and it is not the case that only statistical evidence is relevant when assessing causal claims. In general, the debate in political science has been quite similar to that in sociology: the advocates of mechanisms have criticized the simplistic empiricist uses of statistical methodology for ignoring the importance of causal process assumptions in causal inference.

The volume Social Mechanisms (1998) edited by Peter Hedström and Richard Swedberg is comprised of works by many social scientists who have been important in the development of the mechanistic perspective in social theory. The chapters by Raymond Boudon, Jon Elster, Diego Gambetta, Peter Hedström, Gudmund Hernes, Thomas Schelling, and Aage Sørensen represent what has become known as the analytical approach to social mechanisms. This approach has subsequently developed into analytical sociology movement (Hedström, 2005; Hedström and Bearman, 2009; Hedström and Ylikoski, 2010). While it is still developing and important differences exists between its advocates, the approach currently represents probably the most developed version of the mechanistic approach to social theory.

Mechanism-Based Explanation

The idea of mechanism-based explanation has been developed independently among social scientists and philosophers of

biology (Hedström and Ylikoski, 2010). In the social sciences, the idea of causal mechanism has been used mainly as a tool for exposing the unsatisfactory nature of existing theories and explanatory practices, while in the philosophy of biology the aim has been to find a descriptively adequate account of biological explanation. Despite these separate origins and motivations, both traditions are clearly building on similar ideas about scientific explanation. For example, they share the same dissatisfaction with the covering-law account of explanation (Hedström, 2005; Craver, 2007).

In contrast to the covering-law account of explanation (Hempel, 1965), the mechanism-based accounts are explicitly causal. Because of this, they can avoid many problems that have plagued the covering-law account. For example, unlike the covering-law account, the causal theories of explanation do not allow the length of the flagpole to be explained by the length of its shadow. In other words, the causal theories respect the asymmetry of the explanatory relation. The causal requirement also helps with another set of important counterexamples to covering-law account that is based on explanatorily irrelevant information. Thus mechanical approach has a better chance to deal with cases like men eating contraceptive pills and hexed salt that has proved to be embarrassing for the covering-law account (Salmon, 1989; Woodward, 2003). Similarly, the scarcity of real covering laws in sciences like biology, psychology, and the social sciences is not a problem for mechanism-based accounts of explanation. Observable regularities (often called effects) are things to be explained (Cummins, 2000), not the things that are responsible for explanatory understanding. Finally, mechanism-based accounts clearly distinguish between diagnostic and explanatory reasoning. While supporters of the covering-law account propose that the principal difference between explanation and prediction is that in the case of explanation the conclusion of the argument is already know to be true, the mechanism-based account severs the connection: diagnostic effects might allow for prediction, but they are not explanatory, and the sciences provide plentiful examples of explanatory theories that do not allow for precise predictions (e.g., evolutionary theory).

No consensus exists on the right definition of a causal mechanism. Although some theorists find such a situation frustrating, it is not clear that this is a real problem. The entities and processes studied by different sciences are quite heterogeneous, and it might be impossible to propose a definition of mechanisms that would be both informative and cover all the prominent examples of mechanisms. Some disciplines, such as cell biology (Bechtel, 2006) and the neurosciences (Craver, 2007), study highly integrated systems, whereas others, such as evolutionary biology and the social sciences, study more dispersed phenomena (Kuorikoski, 2009), so it is more plausible to think that informative characterizations of mechanisms are field specific. However, it is possible to give some general characteristics of mechanisms.

First, a mechanism is always a mechanism for something; it is identified by the kind of effect or phenomenon it produces. For this reason the characterization of the effect often requires some care. A standard roulette wheel, for example, does not have different mechanisms for distributing the ball to pockets 16 or 17; rather, the same mechanism produces all 37 possible outcomes. However, the effect does not uniquely identify the mechanism: often there are multiple alternative mechanisms for producing similar effects.

Second, a mechanism is an irreducibly causal notion. It refers to the entities of a causal process that produce the effect of interest. Some definitions of a mechanism make unnecessarily strong assumptions, however. For example, it is sometimes suggested the mechanism is sufficient for the effect. This is an all too strong requirement: a mechanism can involve irreducibly stochastic elements and thus affect only the probability of a given effect. Similarly, the requirement that a mechanism should be unobservable (Mahoney, 2001) is a hasty generalization. There is nothing in the notion of a mechanism that would imply that it is by definition unobservable. Most of the mechanisms constituting an automobile's engine, for example, are quite visible when one opens up the hood (Mayntz, 2004). Of course, when one appeals to mechanisms to make sense of statistical associations, one is referring to things that are not visible in the data, but this is different from them being unobservable in principle.

Third, a mechanism has a structure. When a mechanism-based explanation opens a black box, it makes visible how the participating entities and their properties, activities, and relations produce the effect of interest. More metaphorically, it makes the black box transparent. For this reason, the common suggestion that a mechanism is just an intervening variable misses an important point. The focus on mechanisms breaks up the original explanation-seeking why-question into a series of smaller questions about the causal process: Which entities are participating, and what are their relevant properties; how are the interactions of these entities organized (both spatially and temporally); and what changes in the system or in its environment would prevent or modify the outcome?

Fourth, there is a hierarchy of mechanisms. While a mechanism at one level presupposes or takes for granted the existence of certain entities with characteristic properties and activities, it is expected that lower-level mechanisms will explain them. In other words, the explanations employed by one field always bottom out somewhere (Craver, 2007). However, this fundamental status of certain entities, properties, and activities of a given mechanism is only relative, as these are legitimate targets of mechanistic explanation in another field. It is an inherent feature of the mechanistic view that the entities and mechanisms of various sciences are ultimately related to each other. The entities and activities that one theory postulates should be mechanistically explainable by another theory. Although it is sometimes (Kincaid, 1996) suggested that the idea of mechanistic explanation leads to an infinite regress, this is not the case. For a mechanism to be explanatory, it is not required that the entities, properties, and activities that it appeals to are themselves explained. The only requirement is that such entities, properties, and activities really exist; their explanation is a separate question (Ylikoski, 2012).

Finally, mechanisms seem to be combinable. Two or more mechanisms can be combined to form a more complicated one. For this reason, we can distinguish between atomic and molecular mechanisms (Elster, 1999, 2007): atomic mechanisms are the simplest mechanisms analyzed within a scientific field, whereas molecular mechanisms combine two or more

atomic mechanisms. Another way to conceive of the combinatorial nature of mechanisms is to distinguish between causal mechanism schemes and causal scenarios (Ylikoski and Aydinonat, 2014). This distinction captures the fact that there are two different ways to talk about mechanisms. Causal mechanism schemes are theoretically motivated accounts of how certain kind of effect could be produced. These schemes are often rather abstract and they are not assumed to be faithful descriptions of any particular causal process. They basically provide a how-possible explanation; it tells us how an effect of a certain kind could in principle be produced. In contrast, the causal scenarios are meant to be faithful representations of the processes that produce the effect in a particular case. They are often modified adaptations and combinations of more general mechanism schemes. While the fact that both mechanism schemes and causal scenarios are called mechanisms might cause some confusion, their different roles guarantee that neither way of talking about mechanisms is inherently incoherent.

While the notion of mechanism is important in understanding explanatory reasoning in science, one should not assume that it is the ultimate solution to all problems in the theory of explanation. Quite the contrary, the mechanistic theory presupposes accounts of explanatory relevance, causation, and the nature of generalizations that provide the basis for mechanisms. The notion of mechanism should not be treated like a black box. For example, a mechanism-based explanation describes the causal process selectively. It does not aim at an exhaustive account of all details but seeks to capture the crucial elements of the process by abstracting away details that are irrelevant. The idea of counterfactual difference making provides a handle on this process. The relevance of entities, their properties, and their interactions are determined by their ability to make a relevant difference to the outcome of interest. If the presence of an entity or of changes in its properties or activities truly makes no difference to the effect to be explained, it can be ignored. A natural way to understand these causal counterfactuals is to understand them as claims about the consequences of ideal causal interventions (Woodward, 2003, 2008).

Mechanisms and Causation

The idea of mechanism is also often associated with the analysis of causation. Some authors (e.g., Glennan, 1996) have even attempted to define causation in terms of mechanisms. However, it is prudent to keep these two things apart because most characterizations of mechanisms employ causal notions. A mechanistic definition of causation would be circular, and it would face a thorny question about the causal powers of fundamental (physical) entities. If causal relations at the fundamental level are not mechanical, the definition is false, and if they are mechanical, we end up with an infinite regress, which is regarded by many as an unsatisfactory consequence; with the infinite regress there would be no real causation.

The metaphysics of causation is still hotly debated among philosophers, so it is a strength that the mechanistic account of explanation is compatible with a number of different accounts of causation (Hedström and Ylikoski, 2010). However, the mechanistic perspective sets some important constraints. For example, the idea of productive causal activities associated with it implies a commitment to the locality of causal processes; whether A is a cause of B depends on facts about the spatio-temporally restricted causal process, not on what would happen in other similar situations. This means that theories which attempt to define causality in terms of regularities (like Hume's constant conjunction theory, and many probabilistic theories of causation) are not compatible with mechanistic theories. This obviously does not mean that supporters of the mechanistic perspective would have to disregard regularities as an important source of evidence about causal relations.

The relevance of mechanism-based reasoning is not restricted to explanation. Mechanisms also play an important role in the justification of causal claims. Especially in nonexperimental contexts, mechanisms have often play a crucial role in distinguishing true causal relations from spurious correlations. Mechanisms are useful in causal inference in two ways. The knowledge that there is a mechanism through which X influences Y supports the inference that X is a cause of Y. In addition, the absence of a plausible mechanism linking X to Y gives us good reason to be suspicious of the relation being a causal one (Hedström, 2005). Understanding mechanisms is also important in the extrapolation of causal findings. Here the assumption about the similarity of causal mechanisms is crucial for making reliable inferences from one setting or population to another (Steel, 2008).

While it would be too strong to say that mechanisms are necessary for causal inference, a fully satisfactory social scientific explanation requires that the causal mechanisms be specified. In this context, it makes sense to distinguish between simple causal claims and deeper explanatory claims that are based on an understanding of mechanisms. A simple causal claim tells us about counterfactual dependence: it tells us what would have happened if the cause had been different. In contrast, the mechanism tells us why the counterfactual dependence holds by showing what kinds of entities and processes underlie it. One could say that an account of causal mechanism integrates an isolated piece of causal knowledge into a much larger body of knowledge, and helps us to answer many natural follow-up questions about the conditions under which the causal dependence holds.

It is important, however, to emphasize that mechanisms are not a magic bullet for causal inference. In the social sciences, often the problem is not the absence of possible mechanisms, but how to discriminate between a number of potential mechanisms. To avoid lazy mechanistic storytelling, the mechanism scheme must be made both explicit and detailed, and its assumptions must be supported by relevant empirical evidence.

Mechanism-Based Theorizing

The idea of causal mechanisms is also related to broader ideas about the growth and organization of scientific knowledge. According to an old but still influential empiricist view, general scientific knowledge consists of empirical generalizations and more abstract theoretical principles from which these generalizations can (ideally) be deduced. This vision is challenged by

the mechanistic account of knowledge with its emphasis on scientific knowledge as embedded in mechanism schemes and not in empirical generalizations.

According to this mechanism-based view, scientific knowledge expands by adding to, or improving upon, items already present in the toolbox of possible causal mechanisms. Understanding accumulates when the knowledge of mechanisms becomes more detailed and the number of known mechanisms increases. The unification of scientific knowledge, then, consists of developing a comprehensive and consistent understanding of the causal structure of the world. Further, the world studied by the social sciences is complex, so the idea of a unified foundational theory should be treated with great suspicion. It is better to focus on theories of a more limited scope that are compatible with knowledge produced by other fields. Rather than dreaming about grand unified theories, social scientists should develop theories about social mechanisms that are mutually compatible and which form an integrated web where mechanisms at lower levels of organization explain the mechanisms that higher-level disciplines take for granted. For example, psychology explains (and corrects) the assumptions that social scientists make about human cognitive processes.

The above mechanistic vision of knowledge has much in common with Robert K. Merton's idea of sociological theories of the middle range (Merton, 1968). These theories are not just low-level empirical working hypotheses, or all-inclusive attempts to fashion comprehensive accounts of society. Rather, they are focused theories about certain aspects of the social world that can also be applied to other cases. They do not provide a comprehensive description of the empirical phenomenon, but instead capture something essential about the factors that explain it. While Merton's own discussion of theories of the middle range can be ambiguous, the idea of a mechanism-based explanation can be used to make it more precise (Hedström and Udehn, 2009). In the present account, the middle-range theories consist of the causal mechanism schemes discussed above. These schemes are more or less general in the sense that they can be employed and adapted to particular situations and explanatory tasks.

The idea of mechanism-based theories of the middle-range is important in the highly specialized and fragmented social sciences. Although empirical data, research methods, and substantial theories differ from one subfield to another, the general ideas concerning possible causal mechanisms are something these fields could share and thereby mutually benefit from. In this scenario, social theory provides a set of explanatory tools that can be employed in and adapted to particular situations and explanatory tasks. The mechanisms are general in the sense that most are not limited to a particular application. For example, the same type of mechanisms can be used for explaining cumulative advantage in scientific careers, success in cultural markets, and other contexts that are characterized by surprising inequalities in outcomes.

Individualism and Social Mechanisms

A controversial issue concerning social mechanisms has been its relation to the doctrine of methodological individualism. This is no surprise, as some of the most visible advocates of mechanism-based explanations (e.g., Elster, Boudon) have also advocated methodological individualism. Furthermore, as mechanism-based explanations open the black boxes underlying macro variables, they are bound to end up with smaller-scale entities like individuals and their activities. However, as Renate Mayntz (2004) has pointed out, there is no reason to assume that mechanistic thinking automatically leads to explanatory factors that are about the intentional actions of individual agents. Mechanisms that cite supra-individual entities or properties are certainly possible. For example, various filtering mechanisms that are analogical to natural selection are difficult to understand other than as population-wide processes, and when the selected units are organizations (for example, firms), it is natural to conceive of the mechanism as supra-individual. Similarly, it is hard to see how explanations that employ the structure of a social network as a crucial explanatory variable are compatible with strictures of methodological individualism. It, therefore, seems that mechanism-based explanations can refer to a variety of macro facts and processes that are not about individuals and their attributes. More generally, once the idea of there being a fixed and unique 'micro level' in the social sciences is given up, the association between mechanism-based explanation and methodological individualism vanishes. While a satisfactory mechanism-based explanation in the social sciences might refer to individuals, their relationships, and their actions, it is not guaranteed that they would only refer to the individualistically acceptable properties.

Furthermore, the arguments for mechanism-based explanation do not support the assumption that individual intentional explanations have some kind of explanatory privilege. Thus, while Raymond Boudon is expressing a widely felt intuition among methodological individualists when he writes: "When a sociological phenomenon is made the outcome of individual reasons, one does not need to ask further questions" (Boudon, 1998: 177), he is not spelling out an implication of the mechanistic account but rather articulating a doctrine that is in conflict with it. This "intentional fundamentalism" (Ylikoski, 2012) can take various forms, but it is often related to rational choice theory. However, the assumption that human deliberation is a black box that should not be opened is more in line with nineteenth-century hermeneutic romanticism than with explanatorily oriented social science. A mechanistic explanation appeals to micro-level processes, but nothing in the notion of mechanistic explanation implies that these micro things would always be facts about the intentional actions of individuals. Apart from being supra-individual, as argued above, some explanatory mechanisms might also be sub-individual. For example, various biases and heuristics of human cognition (Kahneman, 2011) might have many effects of sociological interest. It would therefore be foolish to limit the explanatory factors available to social scientists to those that can be reconstructed using simple belief-desire psychology. This argument should not be regarded as challenging the claim that there should exist a division of labor between the social sciences and the sciences of cognition. Rather, it points to the boundaries of this division of labor actually being adjustable and not fixed.

The above argument does not imply a denial of the fact that intentional explanations are, and will be, an indispensable part of the social scientific explanatory repertoire. From the mechanistic point of view, intentional explanations are legitimate causal explanations. Furthermore, the intentional attitudes of individuals play an important role in most mechanism-based explanations of social phenomena. The idea of social mechanism in itself, however, does not tell us how to conceptualize human action. Rather than relying on some preconceived ideas about human motivation or cognitive processing, the mechanistic perspective suggests that social scientific accounts of human agency should be based on findings and theories from the psychological and cognitive sciences. So while the idea of social mechanisms is quite often associated with rational choice theory (Gross, 2009), from a philosophical and general sociological point of view the connection between the two is rather weak. There is nothing in the idea of a mechanistic explanation that would require the explanation to be articulated in terms of rational choice theory. In fact, the requirement that mechanistic explanations cite the actual causes of the phenomenon to be explained often makes rational choice explanations unacceptable as they are built upon implausible psychological and sociological assumptions (Hedström and Ylikoski, 2010).

The key idea of mechanistic explanation is that explanation should represent the essential features of the actual causal structure that produces the observed phenomena. Thus it is not enough that the model 'saves the phenomena.' For unrealistic psychological and social assumptions to be acceptable, they must be simplifying idealizations that help the modeling but do not affect the central explanatory relationships in a significant way. This is not often the case in the applications of rational choice theory. However, as gaining an understanding of complex phenomena is only possible in a piecemeal manner, social scientific explanations must abstract away from many details of human mental life. Therefore, only those aspects of cognition that are relevant to the explanatory task at hand should be included in the explanation. In this sense, the explanatory task determines how rich the psychological assumptions must be. So although the mechanistic approach emphasizes the importance of intentional action in the explanation of social phenomena, it cannot subscribe to an axiomatic vision according to which a specific action theory should be used for all purposes. For many social scientific purposes, a relatively simple desire-belief-opportunity model would be sufficient (Hedström, 2005). For other purposes, some version of the pragmatist theory of action (Gross, 2009) could be fruitful.

Future Challenges

The recent work on social mechanisms has been motivated by two broad considerations. First, social scientists and philosophers have attempted to replace outdated philosophical ideas about social scientific explanation with more plausible ones. Thus the criticism of the covering-law account of explanation and associated ideas about causation have received much attention. A consequence of this is that the mechanists have been much clearer about what they oppose than what they advocate. Second, and more importantly, the idea of mechanism-based explanation has been used to advocate improvement in the explanatory practices of the social sciences. For example, analytical sociologists have emphasized the intellectual virtues of clarity and explicitness (Hedström, 2005). Humans are susceptible to illusion of depth of understanding; we tend to overestimate the detail, coherence, and depth of our understanding (Keil, 2003). With its implicit and vague explanatory standards, largely verbal mode of theorizing, and highly complex object of study, social theory is probably a showcase of this failure of metacognition.

The extensive debate about social mechanisms shows that while the idea of mechanistic explanation has been helping social scientists to avoid certain philosophical pitfalls, the mere adoption of loose mechanistic parlance will not fulfill its more important promise of improving social scientific explanatory practices. It has turned out that many things people call mechanisms are basically effects; they describe a certain kind of outcome but do not provide much insight into the process that produces it. For example, when Charles Tilly (2001) talks about democratization as a mechanism, it is quite clear that he is describing something he is interested in rather than giving an account of the cogs and wheels the drive that process. However, the problem is much more general: cumulative advantage, invisible hand, segregation, and self-fulfilling prophesy are all processes that are identified by their outcomes rather than their internal workings.

The future advancement of the mechanistic agenda in the social sciences requires that social scientists develop an increasingly sophisticated theoretical discussion about causal mechanism schemes. In this endeavor, an abstract definition of social mechanisms will be of little help. What is needed is a cumulative work on the exemplars of social mechanisms. Another challenge is development of a methodology for evaluating mechanism-based causal scenarios. Abstract causal mechanism schemes provide the material for alluring explanatory narratives that might be compelling but lack proper empirical support. The basic problem is that in the social sciences, developing sketchy mechanistic scenarios is rather easy; much more difficult is to articulate these sketches into full-blown explanations and finding evidence that discriminates between competing scenarios. Therefore, much depends on the way in which mechanistic ideas are put to use, otherwise social scientists will end up with mere mechanistic storytelling that lacks both theoretical rigor and empirical relevance.

See also: Analytical Sociology; Explanation: Conceptions in the Social Sciences; Functional Explanation: Philosophical Aspects; Mechanism-Based Causal Analysis; Methodological Individualism in Sociology; Rational Choice Explanation: Philosophical Aspects; Rational Choice Theory in Sociology; Scientific Explanation; Sociological Theory.

Bibliography

Archer, M.S., Bhaskar, R., Collier, A., Lawson, T., Norrie, A. (Eds.), 1998. Critical Realism: Essential Readings. Routledge, London.

Bechtel, W., 2006. Discovering Cell Mechanisms. The Creation of Modern Cell Biology. Cambridge University Press, New York.

Bhaskar, R., 1978. A Realist Theory of Science, second ed. The Harvester Press, Brighton.

Bhaskar, R., 1979. The Possibility of Naturalism. The Harvester Press, Brighton.

Boudon, R., 1998. Social mechanisms without black boxes. In: Hedström, Peter, Swedberg, Richard (Eds.), Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge, pp. 172–203.

Bunge, M., 1997. Mechanism and explanation. Philosophy of the Social Sciences 27, 410–465.

Bunge, M., 2004. How does it work? The search for explanatory mechanisms. Philosophy of the Social Sciences 34, 182–210.

Cummins, R., 2000. "How does it work?" versus "What are the laws?": two conceptions of psychological explanation. In: Keil, F.C., Wilson, R.A. (Eds.), Explanation and Cognition. MIT Press, Cambridge, MA, pp. 117–144.

Craver, C., 2007. Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience. Oxford University Press, Oxford.

Elster, J., 1989. Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge, UK.

Elster, J., 1999. Alchemies of the Mind: Rationality and the Emotions. Cambridge University Press, Cambridge, UK.

Elster, J., 2007. Explaining Social Behavior. More Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge, UK.

George, A.L., Bennett, A., 2005. Case Studies and Theory Development in the Social Sciences. MIT Press, Cambridge, MA.

Gerring, J., 2008. The mechanismic worldview: thinking inside the box. British Journal of Political Science 38, 161–179.

Glennan, S., 1996. Mechanisms and the nature of causation. Erkenntnis 44, 49–71.

Gross, N., 2009. A pragmatist theory of social mechanisms. American Sociological Review 74, 358–379.

Harré, R., 1970. The Principles of Scientific Thinking. Macmillan, London.

Hedström, P., 2005. Dissecting the Social. On the Principles of Analytical Sociology. Cambridge University Press, Cambridge.

Hedström, P., Bearman, P. (Eds.), 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.

Hedström, P., Swedberg, R. (Eds.), 1998. Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge.

Hedström, P., Ylikoski, P., 2010. Causal mechanisms in the social sciences. Annual Review of Sociology 36, 49–67.

Hedström, P., Udehn, L., 2009. In: Hedström, Bearman (Eds.), Analytical Sociology and Theories of the Middle Range, pp. 25–49.

Hempel, C., 1965. Aspects of Scientific Explanation. The Free Press, New York.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux.

Keil, F.C., 2003. Folkscience: coarse interpretations of a complex reality. Trends in Cognitive Sciences 7, 368–373.

Kincaid, H., 1996. Philosophical Foundations of the Social Sciences. Analyzing Controversies in Social Research. Cambridge University Press, Cambridge, UK.

Kuorikoski, J., 2009. Two concepts of mechanism: componential causal system and abstract form of interaction. International Studies in the Philosophy of Science 23, 143–160.

Little, D., 1991. Varieties of Social Explanation: An Introduction to the Philosophy of Social Sciences. Westview Press, Boulder, CO.

Little, D., 1998. Microfoundations, Method, and Causation. On the Philosophy of the Social Science. Transaction Publishers, New Brunswick.

Merton, R.K., 1968. Social Theory and Social Structure. Free Press, New York.

Mahoney, J., 2001. Beyond correlational analysis: recent innovations in theory and method. Sociological Forum 16 (3), 575–593.

Mayntz, R., 2004. Mechanisms in the analysis of social macro-phenomena. Philosophy of the Social Sciences 34 (2), 237–259.

Norkus, Z., 2005. Mechanisms as miracle makers? The rise and inconsistencies of the "mechanismic approach" in social science and history. History and Theory 44 (3), 348–372.

Salmon, W., 1989. Four Decades of Scientific Explanation. University of Minnesota Press, Minneapolis.

Steel, D., 2008. Across the Boundaries. Extrapolation in Biology and Social Sciences. Oxford University Press, Oxford.

Sayer, A., 1992. Method in Social Science. A Realist Approach, second ed. Routledge, London.

Tilly, C., 2001. Mechanisms in political processes. Annual Review of Political Science 4 (1), 21–41.

Woodward, J., 2003. Making Things Happen. A Theory of Causal Explanation. Oxford University Press, Oxford.

Woodward, J., 2008. Mental causation and neural mechanisms. In: Hohwy, J., Kallestrup, J. (Eds.), Being Reduced. Essays on Reduction, Explanation, and Causation. Oxford University Press, Oxford, pp. 218–262.

Ylikoski, P., 2012. Micro, macro, and mechanisms. In: Kincaid, H. (Ed.), The Oxford Handbook of Philosophy of the Social Sciences. Oxford University Press, pp. 21–45.

Ylikoski, P., Aydinonat, E., 2014. Understanding with theoretical models. Journal of Economic Methodology 21, 19–36.

Social Media

Bernie Hogan and Joshua Melville, Oxford Internet Institute, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social media represent a set of communication practices that can typically be described as 'many-to-many.' In contrast to broadcast media, consumers are typically also producers. In contrast to in-person communication, audiences are often ambiguous or underspecified. These practices have ushered in new forms of political organization and new ways to consider social identity as curated. They have also led to new insights about social networks broadly. Some researchers suggest social media lead to increases in social capital and well-being, while others argue this is at the expense of privacy.

Introduction: Defining Social Media

In some senses all media can be thought of as 'social.' All effective communication requires a sender and a receiver. Recently however, a class of communication practices have been given the moniker 'social media' to demarcate them from other, more traditional, types of media. This term first emerged in popular press (rather than in academic studies) following the dot com crash of 2001, and was initially tied to technologist and publisher Tim O'Reilly's pronouncement of 'Web 2.0' media as being participatory and collaborative. The term subsequently bound itself to almost all online practices involving multiple contributors. Sites that are considered social media range from collaborative encyclopedias such as Wikipedia, to social network sites (SNSs) Facebook and Twitter, photo-sharing sites Instagram, and social news site Reddit. Beyond this, technologies such as instant messengers, and to some even e-mail are considered social media.

With such a variety of media all falling under a single rubric it is difficult to completely map the contours of 'social media.' We follow Hogan and Quan-Haase (2010) in suggesting that there are two key elements of social media: 'many-to-many' communication and nonaddressed audiences.

The notion of many-to-many communication predates social media by a decade, but is conceptually well aligned. This is where people are both producers and consumers of content alongside multiple others. Many-to-many audiences can be contrasted with one-way audiences, where communication flows from content producers to consumers (as in broadcast media) or from large populations to a central authority (as in voting and petitions).

Nonaddressed audiences is a second and significant aspect of social media that distinguishes it from other forms. To be nonaddressed means that the message is not being sent to specific other people, but to some class of people, such as all Facebook friends, Twitter followers, readers of a bulletin board, and so forth. Although the boundaries are fuzzy, one might then consider that an e-mail that is sent to many people with each one on the cc line is not an exemplary form of social media, whereas posting a status update to others is.

We wish to consider these two features as sufficient rather than necessary for identifying social media. But more importantly, these features help to clarify many of the emerging issues and insights from the contemporary social scientific analysis of social media. In general, these issues tend to circulate around idea propagation, identity performance, political mobilization, and large-scale statistical regularities.

The technological infrastructure required to coordinate many-to-many communication between large masses of people tends to be very complex. As such, there is a greater precedent for broadcast and participatory media than social media. And yet we may think of some modest precedents for many-to-many media, such as a town bulletin board or parish announcement board. Given these precedents, we believe that the rise of social media highlights a particular arrangement of communication practices, and not particular gadgets, Web sites, or technologies.

Our focus on the communication practices of social media follows in the vein of medium theorists, such as Marshall McLuhan and Joshua Meyerowitz. However, there have been alternative and often inductive classifications of social media. Kaplan and Haenlein (2012), for example, suggest that each instance of a social media can be classified as one of SNSs, collaborative projects, content communities, blogs, virtual game worlds, and virtual social worlds. Although we find this categorization somewhat retroactive and fragile, it helps to denote that a variety of online media can be considered social media.

In the past decade there has been a great interest in social media from the public. One noted example is the Oscar-winning film, "The Social Network," a semifictionalized account of the emergence of Facebook. But social media has also generated massive interest within academia. This interest can of course be found in the obvious context of ICT-related journals (such as the Journal of Computer-Mediated Communication and Cyberpsychology and Behavior (recently renamed Cyberpsychology, Behavior and Social Networking)), but it can also be found in traditional sociological journals and even within statistics, physics, and maths, where social media generate compelling and challenging behavioral data that are pushing both social and hard sciences forward. Intense academic interest in social media also owes much to the fact that a large proportion of contemporary communication takes place through it, with no sense that this will decline in the near future. Illustrations of the scale of this communication are both staggering and indicative of the embedded, 'everyday,' character of much social media practice (Raine and Wellman, 2012). For illustration, as of this writing Facebook claims

over 901 million monthly active users, and a remarkable 126 billion friend connections. Though far smaller in comparison, Twitter currently claims over 140 million active users and 340 million daily tweets. Research findings have come from both within and across disciplinary boundaries, reflecting the efficacy of multidisciplinary approaches to studying social media.

Below we summarize some of the key debates within social media and indicate some of the foundational advances that have come through the analysis of social media data. In particular, we focus on identity construction, political polarization, the meaning (and uses) of friendship, and the 'laws of the Web' as seen through social media. We conclude with speculation about what enduring future directions for research into social media might look like.

Social Media and Social Identities

Identity Performance and Impression Management

A foundational issue in social media is how to understand self-presentation and online identity. This issue predates social media. As early as 1984, Kiesler identified many of the now-common concerns about flaming, mixed messages, and interpersonal ambiguity now prevalent on social media (Kiesler et al., 1984). Yet, studies of self-presentation go back even further if we broaden our focus to include off-line self-presentation. The symbolic interactionist tradition has been exploring the meaning of identity and the self for almost 100 years, arguably starting with George Herbert Mead. In this tradition (or some might say alongside this tradition), one figure has been consistently applied to identity and social media: Erving Goffman.

Although he had many contemporaries in identity theory, Erving Goffman's work has most successfully endured. Although his work on interaction rituals, stigma, and frame analysis have all been applied to social media, many scholars of social media tend to draw on his early notion of the dramaturgical approach. This approach, introduced in *The Presentation of Self in Everyday Life* (1959) suggests that individuals selectively reveal aspects about themselves in order to live up to some idealized notion of what is expected of them in a situation. Goffman's notion is particularly obvious when people are asked to 'create an identity' on any number of social media platforms.

Using a dramaturgical (or stage play) metaphor, Goffman saw identities as performances whereby people selectively draw upon their own histories as well as their expectations of the current situation and their specific role to play in this situation. Like a stage play, Goffman asserts that interaction has a front stage and a back stage. The front is the public face in any situation and is the site of impression management. The back stage, which is often conflated with private spaces, is where the front is readied.

Goffman's approach lends itself particularly well to the study of user profiles on social media services. Social media's many-to-many quality suggests that individuals are now aware that they have an audience for their content. One is not simply speaking to a specific set of people, but to an audience that consumes one's content alongside the content from everyone else on a platform. A string of theoretical and empirical papers have applied these ideas. Of these, danah boyd's work has been particularly notable for unpacking the idea of 'the audience' of social media content.

The work of boyd primarily focuses on the pressures facing teens of constructing and managing online and off-line identities. Through this focus she has unpacked the many social pressures that come from friending behavior, such as the need to add friends as a signaling mechanism, and the challenge of regulating which friends view which content (boyd, 2007). She reminds us that in contrast to everyday interactions, there is no simple way of articulating the audience for an online interaction on social media. While the audience of a Facebook profile might consist principally of a user's friends, it is not clear in a given moment which friends are seeing (or not seeing) which content. Moreover, since a friend online is merely a signifier for access to content, 'friends' can actually be parents, teachers, bullies, celebrities, workmates, neighbors, and a host of other social roles all in the same space.

In extending, Goffman, boyd, and colleagues have explored the key issue of 'collapsed contexts' in social media. The collapsed contexts of social media, where many disparate roles all have access to the same content lead to both significant challenges for determining what sort of content is appropriate, but also to the continual migration of people between SNSs, and between general SNSs, such as Facebook, and niche sites for subcultures, such as Tumblr and Reddit.

Psychological Approaches to Social Media

To the uninitiated, it may be tempting to think that social media can be separated out into 'good guys' such as Facebook and Twitter, who facilitate Egyptian revolutions and reunite families, and 'bad guys' such as anonymous blogs and image-boards that bully and harass while promoting crypto-anarchism. This is the simplified moral panic of a complex problem related to the tension between collapsed contexts and real-names/pseudonyms. It is based on the folk intuition that without one's off-line identity markers, people are unaccountable and therefore prone to communication that is nasty, brutish, and short.

Early work in computer-mediated communication noted that the absence of the external cues in off-line social situations does tend to foster interaction that is somewhat different than what one might expect. In psychology and social psychology, several theories of behavior have emerged to understand how people adapt to the paucity of social cues online. One, situational deindividuation (SIDE) theory, suggests that while online people experience a sense of deindividuation, and begin to take on qualities and behaviors of the group. Together with several correlates in off-line behavior (such as the 'bystander effect'), SIDE proposes that people really do 'lose' part of themselves online, in response to cues found on the specific site they are using (Joinson, 2002).

The second social psychological theory is the online 'hyperpersonal' relationship (Walther, 2007). This theory suggests that the absence of social cues tend foster a great deal of emotion as people read into whatever scant text is available. Rather than feeling like online text is cold and distant, participants tend to feel a groundswell of emotions over simple short sentences. These emotions can be both positive and negative.

A single nasty comment on a Web site could come from a bitter but harmless teenager or an obsessive stalker. In the absence of other cues, the imagination can be a powerful force in making sense of the bitter comment. The same phenomenon can also lead to strong feelings of community among nonlocal interest groups, where participants may have only said short sentences to each other and never even met off-line. One example is the social news site, Reddit, which now hosts the world's largest gift exchange, has in the past raised millions for charity and holds question and answer sessions with celebrities and politicians alike. President Barack Obama held such a forum ('Ask me Anything') during his 2012 reelection bid.

These rebuttals demonstrate that social media does not need to equate digital participation with off-line participation or online friends with 'real' friends. Individuals can engage in collective action and feel strong emotions solely from their interactions in any given medium. In doing so, they are not merely consumers (or even 'prosumers'), but real political actors in a highly connected world. This has led to one of social media's most active sites of inquiry: the political sphere.

Politics

Among political scientists, there is a growing feeling that old modes and methods of political communication – from elites to the masses (top down) and from the masses to elites (bottom up) – are being rapidly displaced and reconfigured by social media. Activists are building new movements and protests using social media that are increasingly successful at altering mainstream political agendas. In the United States, Members of Congress now routinely turn to Twitter to spar with partisan opponents or broadcast messages to constituents, and use YouTube to post campaign advertisements. In many countries grassroots movements have used social media to affect real policy change in ways that even direct action has failed to achieve. The power of social media in politics is not isolated to the United States, or indeed Western democracies. Commentators have suggested that social media have played a significant part in the organizing of protests and the protection of dissident voices in the popular uprisings associated with the Arab Spring (primarily through Facebook) and the Iranian Green Revolution (primarily through Twitter). The threats that social media presented to regimes have led to efforts of many countries to either crack down on social media, censor content or actors, or introduce surveillance during or after protests. At present, it is not clear precisely what role these media have had – whether it is symbolic, organizational, or merely helpful for the public expression of grievances.

The presence of social media in political action is an active area of inquiry and has generated a host of new concepts to address politics in a network society. One prominent concept is 'the filter bubble' that suggests individuals will have their content curated in such a way to promote like-minded opinions (Pariser, 2011). This sort of feedback loop reinforces much of the political polarization that has been observed in blogs and on Twitter. A second concept is the notion of slacktivism, or the idea that individuals who make symbolic gestures online are in fact engaging in political discourse. Or more specifically, that liking a cause on a social media site can be equated with political action. The most salient case of slacktivism in 2012 was the Kony2012 campaign that sought to oust an Ugandan dictator who was using children in his political conflict. Although it was widely tweeted, viewed online, and shared, little was actually accomplished. This was in part because Joseph Kony did not even live in Uganda anymore and that the problems facing the African nation were vastly more complex than the emotionally tinged video that was widely distributed.

A third concept, crowdsourcing, is one part direct democracy and one part mob rule. Crowdsourcing is sometimes considered a technique for new business models where large masses of people each contribute small amounts to a larger cause, either through voting, expertise, or labor. In politics, crowdsourcing has been used in government consultations as well as investigations into government fraud. It has been used for online petitions, which in the UK at least, has been institutionally approved through Number10.gov.uk. At that site, any petition that gets at least 50 000 signatures will at least get an official response. Also in the UK, crowdsourcing was used to pore over the expenses of members of Parliament to assess whether their expenses were legitimate or not. The newspaper *The Guardian* facilitated this by publishing a quarter of a million documents online to be examined. Social media power these efforts by driving traffic and interest, and providing the coordination to sustain these crowdsourcing efforts.

Each one of these topics has had significant academic and popular treatment, but few have provided strong conclusions about the influence of social media on political movements worldwide. The positive scholars thus focus on the 'social' aspect and the negative tend to focus on the 'media' aspect.

From the social aspect, scholars reflect on the collective power of millions of people adding their voice to a chorus of civic will or dissatisfaction. Anyone can help inform journalism or be a journalist in the right circumstances, and it has never been easier to add one's voice to a political campaign. Cha et al. point this out in *The Million Follower Fallacy* (2010) where they indicate that on globally relevant apolitical topics, celebrities and pundits tend to be the most regarded sources of information. However, in a political event such as the Iranian Green Movement after the contentious election, it was Iranian locals who held the most sway over information. In a more recent piece on the Spanish protest equivalent to Occupy Wall St., González-Bailón et al. (2011) indicate that recruitment into the participatory networks of tweeting did not occur because of a few well-placed celebrities, but due to the slow steady diffusion of many individuals with their own modest sized networks.

When the emphasis is placed on the 'media' portion of social media, critical scholars are interested in pointing out how the infrastructure that facilitates such sociality is not neutral. Instead, the technology itself can be an active agent in filtering, redirecting, and managing the flow of content. Specific to social media, Twitter deliberately curates its top trending topics, with some political topics (as well as pop sensation Justin Bieber) sometimes being excluded. Morozov has written at length about how even when these terms are not filtered, they can still be monitored by repressive regimes who come between a sender and a receiver (Morozov, 2012). This can lead to both active repression of high-profile social media pundits, informers, and scholars as well as passive self-censorship by those who fear someone tracking them. To note, this is not

simply happening in authoritarian countries, but in Western democracies as well. Canadian authorities have used social media to track down individuals who rioted in Vancouver after a Hockey final and the UK authorities have done the same after the high-profile London riots of summer 2011.

Perhaps more pernicious than direct surveillance and monitoring is the passive filtering of individuals into ideologically specific interest groups, with little opportunity afforded for contact between groups. This sort of polarization can happen without any intervention, as was shown in the now iconic visualization of the 2004 American political blogosphere by Adamic and Glance (2005). The image shows two clusters of nodes (blogs), with many links within each cluster (based on who 'reblogs' who) and few links in between. Adamic and Glance sought to demonstrate that there may be completely different blogging cultures online. This situation is vastly different from the world of broadcast news where presenters assumed that their audience would consist of those on both the left and the right. It is more pernicious because this sort of ideological segmentation can be reinforced by algorithms under the guise of 'personalization.' Facebook friend feeds, Twitter lists, and Google results can now be filtered to suit an individual's personal politics even without their explicit consent. On a simplistic level, these algorithms function by noting when a user clicks on more links from their right-leaning friends than their left-leaning ones.

Network Science

Social media are extremely data driven. Outside of antiquated bulletin board systems, social media are very heavily dependent on databases. These large tables of data are searched, sorted, and queried continually. Every now and then, researchers gain access to them for academic purposes. What has come of this is a series of remarkable mathematical insights at the nexus of computer and social sciences. Indeed, much of the findings discussed above come not only from experimental and survey research, but also the use of new methods based on trace data within the emerging field of 'Computational Social Science' (in America) and 'Digital Social Research'/'eSocial Science' (in Europe). Two of these insights in particular have proven robust to the point where their mention in a formal academic context is almost anachronistic. They are 'long-tail distributions' and 'small worlds.'

The importance of long-tail distributions is a substantial challenge to existing normal science. The twentieth century was characterized by the 'normal distribution' – most personalities, heights, weights, etc., tend toward an average or 'mean' and increasingly fewer end up either above or below this mean. The most controversial normal distribution may be intelligence, where it is asserted that most people are of average intelligence with few outliers at the top and the bottom. The Web, on the other hand, is anything but normal. Over 95% of edits to Wikipedia are made by less than 1% of users of the site. Most Web sites receive extremely few visitors, while a small number get millions per day. We must therefore ask ourselves what it means to talk about the 'average' Web site, the average number of edits to Wikipedia, or the average number of sales on Amazon.com. The distribution of these figures is in many cases described as a 'power law' (Barabasi, 2003), but despite this stylish name and rigorous statistical definition it is often more accurately considered as a long-tailed distribution. Though most people communicate with their local friends, there are still celebrities, A-list bloggers, obsessive Wikipedia editors, and so forth who either do far more communicating than everyone else, or communicate with far more people.

In the broadcast era, some people were on television and most were not. But on social media, celebrities are using the exact same platforms as their audience and often in similar ways. It is more the case that, at least in terms of audience, they benefit from a 'rich-get-richer' effect. Those on the top 10 (or the first page) tend to stay on that first page. Hindman (2009) has drawn on this point to suggest that it is not local town papers that are under the most threat from social media, but the middle papers – the state, provincial, and midsize city papers that are being squeezed online by the extremely big players, while the really small players may still find a niche in the long tail.

The second deep insight about social media is that its users are extremely well connected in social network terms. This does not mean everyone has a lot of friends. It means that the whole system can work very efficiently even if people mainly link to their close clusters of friends, as long as there are a few people here and there that link across dense clusters. This property, referred to as the 'small world' phenomenon, is one of the reasons that social media can be remarkably effective at diffusing information (including rumors and scares). The small world phenomenon was originally characterized by Milgram and Travers using a clever experiment involving forwarding letters to an unknown person using friends of friends. Their article popularized the notion that there are only 'six degrees of separation' between people around the world. Watts and Strogatz formalized this property in the late 1990s and early 2000s and found it in many network systems, including social media (Watts, 2002). More recently, Backstrom and colleagues have come up with a rather notable finding. Using the complete Facebook social graph, they have demonstrated that in terms of friendship links on the site, the average number of hops required for one person to know another is not six, but 3.75 or so (Backstrom et al., 2012). That is to say, the average Facebook user can reach any of the 900 million other users on the site using less than four hops *on average*, despite most people on Facebook linking only to off-line and typically local friends.

These patterns characterizing long tails and small worlds are part of the building blocks of burgeoning body of research on how social media diffuse information, cement inequalities, turn on a dime (as people migrate from Friendster to MySpace to Facebook to Twitter to the yet-undetermined-next-big-thing), and simultaneously reinforce the feeling that communication now is both extremely global and extremely local (or what Joshua Meyrowitz and later Barry Wellman termed 'glocalization').

Social Capital, Social Media, So What?

Social media, and SNSs in particular, tend to elicit strong opinions about small design issues. Users criticize Facebook for having a 'like' button, but no dislike button, and Twitter for limiting tweets to a seemingly trivial 140 characters, thereby

dooming them to banality. But one design decision above all has generated seemingly endless editorials, punditry and academic research – the use of the word 'friend' to signify a point of connection. How can an individual have many hundreds or thousands of friends? Do users even know their online friends in 'real life'? In the last half decade researchers have turned this question on its head by suggesting that instead of trying to define what a friend *is*, we should instead look to what social connections *do*.

Increasingly, evidence seems to be that these connections provide support, information, entertainment, and positive self-worth, as well as peer pressure, norms, and expectations of reciprocity. In short, they provide 'social capital.' So, do people feel that social capital exists in social media? A string of studies, but especially those of Nicole Ellison, Cliff Lampe, and their colleagues have demonstrated that in many contexts heavy users of social media do in fact report more social capital (Ellison et al., 2011). Moreover, they have also shown that levels of social capital tend to increase over time for users of SNSs. Other work, such as that of Moira Burke and her colleagues within the Facebook data team, has demonstrated that not all Facebook use is equal. Passive consumption may actually increase feelings of loneliness and exclusion. However, those who actively engage the site, 'like' things, tag, comment, ask questions, and wish others happy birthday tend to report more social capital (2011).

Future Directions for Research

This article has not focused on specific statistics on social media diffusion. These statistics are bound to change, and are difficult to validate in the first instance. We know that most Facebook friends are known off-line, and that the site is the biggest SNS in most countries in the world (with Russia and China being the notable exceptions at present). However, we are not interested in making a story about Facebook exclusively. The key research of the future will not give a better point estimate of how often these media are used, but how eff*ectively* they are used and to what end. A new crop of researchers have emerged to meet these challenges by noting that social media can be a decisive domain for investigating influence, diffusion, social connections, norm formation, and governance alongside a host of recurring and foundational issues around social identity.

Some of the strongest work ought to come from analyses that merge together multiple networks. For example, what is more influential, one's Twitter network, one's Facebook network, or one's e-mail network (bearing in mind the fluid and partial overlap of all three)? But one might also think of networks as bounded by ideology or language. How do different languages influence each other? For example, what sort of content is shared mainly on Weibo (in China) that is not shared on Twitter in English? Additionally, Haythornthwaite's media multiplexity hypothesis (namely that the more media people share the more close they are) could be updated when considering the massive diffusion of potential social media (2005). One might ask: How could we unfilter the filter bubble? Or: How can one rebuild segmented social roles after the context collapse of social media? Finally, if content is persistent and death inevitable, what should become of all the misbegotten profiles?

These questions are relevant whether people are using a desktop, an iPhone, or a tablet. Whether the next gadget is a wristband, heads up display, or a device tapping directly into the brain, it will do little to alter the prevailing notion of social media as a many-to-many practice. The practices themselves (filtering, identity management, curation, polarization, power laws, and small worlds, among many more) will persist. Perhaps in time our understanding will catch up with our technology. In the meantime, social scientists are fully justified in using existing theories of communication. Notions such as medium theory, hyperpersonal identity, impression management, and homophily can be honed, scoped, and elaborated, as researchers compare platforms, users, and practices across the continually unfolding social media ecology.

See also: Human–Computer Interfaces; Identity Offline and Online; New Media and Social Capital; Social Protest and New Media.

Bibliography

Adamic, Lada A., Glance, Natalie, 2005. The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 36–43.

Barabasi, Alberto.-Lazlo, 2003. Linked. The Penguin Group, New York, NY.

boyd, danah, 2007. Why youth (heart) social network sites: the role of networked publics in teenage social life. In: Buckingham, D. (Ed.), MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media Volume. MIT Press, Cambridge, MA, pp. 119–142.

Burke, Moira, Kraut, Robert, Marlow, Cameron, 2011. Social capital on Facebook: differentiating uses and users. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 571–580.

Backstrom, Lars, Boldi, Paolo, Rosa, Marco, Ugander, Johan, Vigna, Sebastiano, 2012. Four degrees of separation. In: Proc. 4th ACM Int'l Conf. On Web Science (WebSci).

Cha, Meeyoung, Haddadi, Hamed, Benevenuto, Fabrìco, Gummadi, Krishna P., 2010. Measuring user influence in twitter: the million follower fallacy. In: 4th International AAAI Conference on Weblogs and Social Media (ICWSM).

Ellison, Nicole B., Steinfield, Charles, Lampe, Cliff, 2011. Connection strategies: social capital implications of Facebook-enabled communication practices. New Media & Society 13 (6), 873–892.

Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Anchor Books, New York, NY.

González-Bailón, Sandra, Borge-Holthoefer, Javier, Rivero, Alejandro, Moreno, Yamir, 2011. The dynamics of protest recruitment through an online network. Scientific Reports 1.

Haythornthwaite, Caroline, 2005. Social networks and Internet connectivity effects. Information, Communication & Society 8 (2), 125–147.

Hindman, Matthew S., 2009. The Myth of Digital Democracy. Princeton University Press, Princeton.

Hogan, Bernie, Quan-Haase, Anabel, 2010. Persistence and change in social media. Bulletin of Science, Technology & Society 30 (5), 309–315.

Joinson, Adam N., 2002. Understanding the Psychology of Internet Behaviour: Virtual Worlds, Real Lives. Palgrave Macmillan, New York.

Kaplan, Andreas M., Haenlein, Michael, 2012. Social media: back to the roots and back to the future. Journal of Systems and Information Technology 14 (2), 101–104.

Kiesler, Sara, Siegel, Jane, McGuire, Timothy W., 1984. Social psychological aspects of computer-mediated communication. American Psychologist 39 (10), 1123–1134.

Morozov, Evgeny, 2012. The Net Delusion: The Dark Side of Internet Freedom. Penguin, New York, NY.

Pariser, Eli, 2011. The Filter Bubble: What the Internet Is Hiding from You. The Penguin Press HC, New York, NY.

Raine, Lee, Wellman, Barry, 2012. Networked: The New Social Operating System. MIT Press, Cambridge, MA.

Watts, Duncan, 2002. Six Degrees: The Science of a Connected Age. W.W. Norton, New York.

Walther, Joseph B., 2007. Selective self-presentation in computer-mediated communication: hyperpersonal dimensions of technology, language, and cognition. Computers in Human Behavior 23, 2538–2557.

Social Mobility, History of

Hartmut Kaelble, Institute for History, Humboldt University Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article on the history of social mobility presents important special terms and discusses the major sources used in this field. It traces the major debates in its history, i.e., the debate on the long-term change of social mobility; the comparative debate on societies seen as particularly fluid, such as the United States; the debates on the change of the gender contrast in social opportunities; and on the change of contrasts in social opportunities between ethnic groups. It finally makes proposals for further research.

General Questions

Social mobility was an attractive theme in the field of social history research during its early heydays in the 1960s and 1970s. Historians explored social mobility with two intentions. First, they wanted to cover the general history of the equality of opportunities in modern societies. Historians were interested in the question of whether the chances for social ascent improved or worsened in modern societies. The increasing openness or growing exclusiveness of the elites, the rise or decline in opportunities for social ascent from the lower middle or lower classes, the developing opportunities for immigrants and ethnic groups, changes in access to channels of social ascent such as education, wealth, family networks and marriage pools, and careers in business, the public services, politics, churches, culture, or entertainment were central topics. Modern societies could appear as particularly fluid, but also as blocked societies. Secondly, historians discussed the history of social mobility in Western societies in a comparative perspective. European and American historians examined the myth of promising opportunities for social ascent and advancement in America as compared with Europe. Historians also investigated the myth of the unblocked social mobility in Communist countries, i.e., in the years after the 1917 Revolution in Russia and also in the years after World War II in the imposed Communist regimes in Eastern and Central Europe. Social mobility was an example of an interdisciplinary historical field, to which historians as well as sociologists and political scientists contributed alike.

In more recent years, other new and basic questions have been added to these initial inquiries. European historians work more on the social mobility of ethnic groups and immigrants, because they became aware that Europe has experienced a major upheaval from being a continent of outmigration to become a continent of immigration. How immigrants succeeded and how their success differed depending on the values and resources of different ethnic groups became an important topic of study for European historians. Moreover the social mobility of women gradually became a somewhat more frequently explored theme, after having been largely neglected in earlier research. These investigations required different methods of research, since occupational careers were less central to women's social mobility for a long time. Finally, international comparisons were less limited to industrial societies in Europe and America, such as Britain, France, Germany, Sweden, or the Eastern United States; Eastern, Central, and Southern European countries as well as non-Western societies were examined more frequently.

Definitions and Sources

Like all fields of research, social mobility research has implicit definitions. Three definitions in the field are most important.

Historians usually investigate social mobility by looking at individuals. Under the heading of social mobility, they usually do not explore the grading up or grading down of entire social groups, i.e., ethnic groups, professions, or entire social classes. The rise or decline of entire families is not examined under this heading either. The Buddenbrook family would normally not be seen as an example of social mobility.

Moreover the research on social mobility focuses on mobility in social hierarchies rather than on spatial mobility. Geographical mobility, i.e., migration, commuting, traveling, and living abroad, is considered a different topic. This does not mean that historians are not aware of the strong influence of geographical mobility on social mobility. Historians explored the relationship between social ascent and migration and the effects of involuntary geographical mobility, such as expulsion or exile, on downward social mobility.

Social mobility research also concentrates on occupational mobility. When historians investigate career mobility (or intragenerational mobility), they usually trace the career paths of individuals, who either moved between different occupations and occupational positions or persistently remained in one single occupation. Historians examining the mobility between generations (or intergenerational mobility) including marital mobility, normally compare the occupation of an individual with the occupation of the father or less frequently with the occupation of the mother and the grandparents. Occupation is usually seen as the crucial and most easily available indicator of an individual in a historical society, though historians are aware that the situation of an individual also depends on other social and economic aspects such as income and wealth, housing and the place of residence, education, and social networks. The historical position and the change of an individual in the social hierarchy is usually explored via occupation. Hence, establishing an approximate

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62022-6

hierarchical classification of occupations is of major importance to historians dealing with social mobility. Upward and downward social mobility is usually measured by these classifications. Still, historians are also fully aware of the problem that the activity of an individual in history, more often in past than the present, might comprise various occupations at the same time, including professions still in the making.

The use of sources in historical research on social mobility also follows a particular approach. Historians link various historical sources and trace historical individuals through sources such as marital records, census materials, tax files and last wills, records of churches and public administrations, and autobiographical sources. The competence of historians in linking these various sources has grown substantially and this makes social mobility studies a highly sophisticated but also time-consuming field of research. Two other trends can be observed in the handling of sources. On the one hand, the secondhand use of sociological surveys by historians has become more frequent and efficient since the latter part of the twentieth century, because access to personal registers in public archives has been restricted due to data protection laws. Quantitative cohort analysis, i.e., research on the differences between the older and younger cohorts in surveys, allows approximate estimations with qualifications about historical trends in social mobility. Important international comparisons are based on cohort analysis. On the other hand, during recent years, social mobility studies have become highly sophisticated. Individual careers are more frequently explored in micro studies with as many details as possible. These studies investigate a number of richly documented individual cases rather than researching all members of a local society through quantitative methods. Micro studies often use autobiographical materials that are difficult to analyze quantitatively and do not allow generalizations. This type of micro study rarely concentrates exclusively on social mobility, but covers a broad variety of aspects.

Four Main Debates

A first debate on social mobility by historians revolves around the increase of social mobility during the last two centuries. There is an optimistic view and a skeptical view regarding long-term changes in social mobility. The optimistic advocates of the rise of social mobility usually argue that not only occupational mobility in general, but also upward social mobility has increased since industrialization. The industrial revolution is not seen as the major dramatic upheaval in the history of social mobility. The more we approach the present, the higher the recorded rates of social mobility. It is also argued that various major social changes have led to more social mobility and also to more opportunities for social ascent: The fundamental changes of the active population from the predominance of agricultural work until the nineteenth and early twentieth centuries to the predominance of industrial work, and the subsequent shift to service sector employment since the 1970s, generated substantial social mobility between occupations. The distinct increase of the sheer number of occupations in all modern societies since the industrial revolution also must have led to more social mobility. The general changes in professional mentalities, a weakening emotional identification with specific professions, social milieus, and also with specific local milieus, and an increasing readiness for job mobility and lifelong training, further enlarged the number of socially mobile persons. The rapid expansion of secondary and higher education especially since the end of the nineteenth century not only had a profound effect on the chances for better training, but also for entry into better paid job positions. The fundamental transformation of societies with very few academically trained persons and a massive majority of elementary school leavers or even analphabets, into societies with various levels of high qualifications and a small minority of elementary school leavers and illiterates, made social mobility and social ascent much easier. The rise of the welfare state, the mitigation of individual life crises, and the guarantee of individual social security clearly improved the chances for further training and for the purposeful use of occupational opportunities. The rapid increase in geographical mobility since the second half of the nineteenth century led to a widening of the labor market and to a greater variety of new chances. The general decline of the fertility rate during the late nineteenth and early twentieth centuries made it possible for parents not only to invest more in the individual support and education of their children, but also to further advance their own professional careers – not only in the case of men, but finally for women also. Changing social differentials of fertility rates also had an impact on social mobility. Equal opportunity policies, i.e., specific government policies aimed at improving educational and occupational opportunities – for lower classes, for ethnic and religious minorities, and later on also for women and for immigrants – also must have had an impact on social mobility. In sum, a substantial list of factors in favor of an increase in social mobility over the last two centuries can be put forward.

The advocates of stagnation or even of a decline in social mobility are a more heterogeneous group. Their arguments stem from very different ideas of social developments. Some scholars who share in the skeptical view, argue that the fundamental upheaval of modern societies during industrialization led to an unprecedented rise in social mobility, (both in upward and downward mobility); and that modern societies thereafter became more closed, since the generation of business pioneers ended and many occupational careers became more formalized and more dependent on formal education; as modern bureaucracies emerged; and as mentalities adapted to the modern, highly regulated job markets. Others hold the view that industrialization in the nineteenth and early twentieth centuries not only led to rising numbers and fundamental change in occupations, but also led to a class society in which the major social classes, the middle class, the lower middle class, the working class, the peasants, and in some societies also the aristocracy, tended to reinforce the dividing lines between social classes and, hence, to reduce rather than to enlarge, the number of mobile persons. Still other scholars discuss the stagnation of social mobility rates in a different and much more narrow sense: They argue that the long-term change in social mobility from the industrial revolution until the present was mostly structural, that is, it depended almost exclusively upon the redefinition of the active population and the professions, rather than on the reduction of social, cultural, and political barriers. According to this view, social mobility

stagnated, apart from the changes simply induced by transformations in occupational structures. Still other advocates of the long-term stagnation of social mobility argue that the educational and occupational chances of lower classes, of women, and of minorities might have improved to some degree, but inequality of opportunities remained, because the middle and upper classes, the male population, and the upper levels of the ethnic majorities did not experience loss of opportunities.

Ever since the beginning of quantitative research on social mobility after World War II, this debate has produced a large number of historical studies of social mobility with a wide range of results. Three main results ought to be mentioned in summary. First, a clear decline in social mobility rates can only be found in rare cases. Most studies show either steady or increasing rates of social mobility, depending on the type of community and country and on the generation and the period under investigation. Secondly, changes in overall social mobility rates do, in fact, depend to a large degree on changes in occupational and educational structures. So in fact, modern societies became more mobile to a large degree because of the expansion of education and because occupational change became so frequent and 'normal.' Finally, there is much evidence that the increase in educational and social mobility of the lower classes and women, did not increase to the detriment of the educational and occupational chances of the middle and upper classes and of men. Except for the Eastern European countries in some specific periods, social mobility was usually not a zero-sum game.

A second debate focuses on the issue of advanced social mobility in the United States and in Communist countries, respectively. The old debate on the lead of the United States dates back to at least the early nineteenth century, when the French social scientist Alexis de Tocqueville argued that American society offered more opportunities for social ascent than Europe.

Various pieces of evidence were presented indicating the American lead in this field. Some empirical studies by sociologists demonstrated that in some crucial aspects a clear American lead could be shown. This was especially true for mobility into the professions. Higher education in the United States was more widespread and offered more chances than in Europe. Hence, the social ascent of people from a lower class background into the professions requiring higher education was clearly more frequent in the United States than in Europe. Furthermore, comparative historical studies on late nineteenth and early twentieth centuries American and European cities showed that in particular cases, America was in a modest lead during this period: Unskilled workers in fact moved up into white-collar positions in American cities somewhat more frequently than in European cities. Finally, historians demonstrated that the important difference between American and European societies could be found in the American creed that the United States offered greater opportunities than the more closed European societies.

Other recent comparative studies tend to argue that American society is no longer in the lead in comparison with Europe. According to this view, industrialization and modernization led to similar opportunities everywhere. Various international comparisons of social mobility rates support this view and demonstrate that the overall rates were not distinctly higher in the United States than in older societies such as Western Europe or Japan. As these comparisons deal almost exclusively with the second half of the twentieth century, the results might be due to the fundamental social changes in Europe and Japan since World War II.

The opportunities for social advancement in Communist countries, especially in the USSR, during the 1920s and 1930s and in the Eastern European Communist countries in the late 1940s and 1950s are at the center of a less intense but important debate. Some historical studies of social mobility demonstrate that during these periods, rates of upward social mobility into the higher ranks of the social hierarchy in these countries were substantial compared with Western European societies. This was true partly because of the rapid expansion of higher education, partly also because of the Communist abolition or even physical extermination of the urban and rural middle class and partly because of far-reaching changes in the employment structure as a result of rapid industrialization enforced from above. However, the rise of social opportunities in Communist countries was, if it rose at all, largely limited to the period of their early beginnings. Most comparative studies of the 1970s and 1980s show that rates of social mobility were not distinctly higher in Eastern Europe than in Western Europe. This was partly because the Communist political and administrative elite became exclusive and gentrified, partly because social change slowed down, and partly because the expansion of higher education was reduced in several Communist countries.

Only a small number of historical studies have examined gender-related differences in social mobility. A comprehensive debate among historians has yet to develop. But gender differences might add new important aspects to the general debate about long-term trends in social mobility. Four conclusions can be drawn from the few studies that exist. First, in a more radical sense than in the study of male mobility, female mobility raises the question whether social mobility should in fact be centered around occupational mobility or whether other factors – such as marriage and unpaid or partially paid work in emerging professions – deserve much more attention. Secondly, the question of whether there was an increase in downward mobility during the transition to modern society, (i.e., during the rise of female activity outside the family sphere) is to be explored. A study of female social mobility in twentieth century Berlin demonstrated a high rate of intergenerational downward mobility among active women during the early part of the century. Thirdly, the studies on the social mobility of women clearly demonstrate the effects of economic crises and fundamental transitions on social mobility. Opportunities for women seem to have depended strongly on economic prosperity and on long-term social stability. In periods of economic crisis and rapid transitions such as the upheaval of 1989/91, more women than men were among the losers. Here again the study of female mobility might draw the attention of the historians to a more general aspect of mobility which requires further investigation. Finally, social mobility of women also demonstrates that definite changes of social opportunities were only reached in a long-term perspective. It was shown that even when important channels of upward social mobility such as education offered equal chances to women, this did not immediately lead to better

and more occupational opportunities for women. Besides the study of institutions and policies, the historical study of the experience and perceptions of social mobility is crucial.

The debate over long-term changes and international comparisons of the opportunities of ethnic groups in the nineteenth and twentieth centuries has also not yet fully started in general terms, in spite of numerous studies on individual ethnic groups. Whether the chances of social ascent for the Italian, Polish, Jewish, Irish, or Finish immigrants in the nineteenth and early twentieth centuries in France, Germany, Switzerland, Britain, and Sweden compared with the indigenous lower classes, were in fact clearly worse than for the Italian, Spanish, Portuguese, Yugoslavian, Greek, Turkish, and Maghreb immigrants in the industrialized West European countries since the 1950s in comparison to the new lower classes, and also worse than for the new immigrants since the 1990s into Southern and Western Europe, has not been examined so far. It is also not clear whether the chances for well-educated middle class immigrants to enter the various national middle classes in business, the sciences, entertainment, and culture in Europe, improved in the late twentieth and early twenty-first centuries, compared with the nineteenth and early twentieth centuries. Did a German middle class immigrant in liberal nineteenth-century Britain actually have less opportunities than a French middle class immigrant into the highly regulated late twentieth-century Germany? We implicitly assume that European societies gradually offer better chances for immigrants, since this has increasingly become a major issue in politics, in the media, and in scientific research since the 1970s. But precise comparisons in time and space have yet to be made.

Future Historical Research

Since the 1980s, social mobility has been less frequently studied by historians in Europe and the United States, with the exception of Dutch historians. The topic became less attractive, because historical change appeared to be less spectacular than had originally been assumed and also because international differences did not seem very intriguing. New research on social mobility did not look promising. Moreover, the standards for historical research on social mobility became so high and the access to archival material became so restricted because of data protection, that historians were discouraged.

Still, the history of social mobility offers a range of new questions besides the topics that have already been mentioned. International social mobility, i.e., the rising international paths of careers (especially of highly qualified persons), is one of these new topics. Working abroad for a certain period or one's entire life, often combined with an international marriage and careers supported by international families or in international organizations, is a rather recent development. But international social mobility always existed in history in large organizations such as the churches, in the European courts, among European artisans, and in outmigration from Europe during European expansion. The long-term historical continuities and the major upheaval of international social mobility pose a new challenge for historians. International historical comparisons of social mobility, which have always been of major interest but strictly limited to the Western sphere, are a second challenge. Expanding international comparisons by including other regions of the world, especially the new economic powers in East Asia, India, Latin America, and Russia and the study of the effects of the various globalizations will give social mobility research a new attraction. Moreover, a crucial element of the new poverty in industrial societies (the poverty of single-parent families, immigrants, drug addicts, broken careers) is seen in the almost total lack of opportunity to leave this new class of the precarious poor and achieve at least lower middle class status. Study of the extremely limited opportunities of this new class of the poor and its resemblance to older historical poverties in respect of social mobility, is a further challenge for historical research. Finally, historians are perhaps more open and more prepared to investigate change in representations of social mobility in government policies, in elites, among those who promote the channels of social mobility, and also among the mobile individuals themselves. This is a topic on which historians have a chance to be read attentively, also in other disciplines.

See also: Education, History of; Equality of Opportunity; Poverty in History; Social Classification; Social History; Social Stratification; White-Collar Workers, History of; Working Classes, History of.

Bibliography

Berger, Peter A., 2001. Soziale Mobilität. In: Schäfers, Bernhard, Zapf, Wolfgang (Eds.), Handwörterbuch zur Gesellschaft Deutschlands. Leske und Budrich, Opladen, pp. 595–605.

Bourdieu, Pierre, 1984. Distinction: A Social Critique of the Judgement of Taste. Routledge, London.

Breen, Richard, December 2004. Social Mobility in Europe. Oxford University Press, Oxford.

Breen, Richard, Luijkx, Ruud, Müller, Walter, Pollak, Reinhard, 2009. Nonpersistent inequality in educational attainment: evidence from eight European countries. American Journal of Sociology 114 (5), 1475–1521.

Erickson, Robert, Goldthorpe, John, 1992. The Constant Flux. A Study of Class Mobility in Industrial Societies. Clarendon Press, Oxford.

Federspiel, Ruth, 1999. Soziale Mobilität im Berlin des zwanzigsten Jahrhunderts. Frauen und Männer in Berlin-Neukölln 1905–1957. de Gruyter, Berlin.

Fitzpatrick, Sheila, 1979. Education and Social Mobility in the Soviet Union, 1921–1934. Cambridge University Press, Cambridge.

Ganzeboom, H.G., Luijkx, R., Treiman, D.J., 1989. Intergenerational class mobility in comparative perspective. Research in Social Stratification and Mobility 8, 3–84.

Groß, Martin, Klassen, Schichten, 2008. Mobilität: Eine Einführung. Verlag für Sozialwissenschaften, Wiesbaden.

Haller, Max (Ed.), 1990. Class Structure in Europe: New Findings from East-West Comparisons of Social Structure and Mobility. Sharpe, Armonk, NY.

Kaelble, Hartmut, 1985. Social Mobility in the 19th and 20th Centuries. Europe and America in Comparative Perspective. Leamington Spa. Berg Publishers, Dover, NH.

van Leeuwen, Marco H.D., Lesger, Cié, 2010. Soziale Mobilität, Enzyklopaedie der Neuzeit, Band 12. Verlag J.B. Metzler, Stuttgart/Weimar, 250–256.

van Leeuwen, Marco H.D., Maas, Ineke. Historical studies of social mobility and stratification. In: Annual Review of Sociology, vol. 36, pp. 429–451.

Lipset, Seymour Martin, Bendix, Reinhard, 1959. Social Mobility in Industrial Society. University of California Press, Berkeley.

Mayer, Karl Ulrich, 2009. New directions in life course research. The Annual Review of Sociological Research 35, S.413–S.433.

Sewell, William H., 1985. Structure and Mobility. The Men and Women of Marseille, 1820–1870. Cambridge University Press, Cambridge.

Sorokin, Pitirim A., 1959. Social and Cultural Mobility. Free Press, Glencoe IL.

Vincent, David, Miles, Andrew (Eds.), 1993. Building European Society. Occupational Change and Social Mobility in Europe, 1840–1940. Manchester University Press, Manchester.

Social Movements and Political Violence

Lorenzo Bosi, European University Institute, Fiesole, FI, Italy
Chares Demetriou, National Research University, Moscow, Russian Federation

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to provide an overview of the ways in which political violence and social movements connect, as well as of the ways in which these connections are studied in academia. It delineates and expands on three specific topics: the emergence of political violence, the disengagement from it, and the consequences of it. The reviews of violence emergence and disengagement are broken down by the level of analysis into macro, meso, and micro; the review of the consequences of violence is broken down into the political, economic, social, and biographical areas. At the conclusion of this article, we offer brief remarks on the field's future directions.

There are two facets to the issue of social movements and political violence. The one facet regards the phenomena to be studied, namely phenomena of social movement-related violence. These phenomena include violence by social movement members as well as violence against them. They are therefore wide-ranging, covering anything from protest by throwing food at politicians to protest by categorical killing, and from police control of street protest to state-approved torture, to mass killing of movement constituencies by militaries. The other facet regards the academic fields that study these phenomena. One field that does so is, of course, the field of social movement studies. But there are other fields, too, and among them the most visible – indeed much more visible and influential than the field of social movement studies – is the field of terrorism studies.

Thus the subject matters of the fields of social movement studies and of terrorism studies often overlap. (The former field predominately studies social movements outside the context of violence; the latter field on principle covers also violence which is unrelated to social movements.) As the two fields approach their common subject matter, however, two differences stand out. The one is theoretical. Social movement studies bring to their investigations of violence the field's standard theories and concepts, rather than developing separate ones particular to violence. In general, the social movement approach has evolved from a structural orientation, characteristic of the 1970s and 1980s, to the so-called political process approach dominant from the 1990s onward. Reflecting the influence of American sociology, the political process approach privileges organizational and relational analysis. It therefore locates political violence within the context of social, political, and cultural conflicts, and considers the turn to violence as being part of broader processes of political contention (Bosi et al., 2014). On its part, the field of terrorism studies has been mainly dominated by security questions – security of the public and/or the state – and has often produced policy-oriented knowledge (Goodwin, 2004). Stemming from a long tradition in political science concerned with a residual category of political violence resultant from the exclusion of interstate warfare, this field has privileged the theories and analytical perspectives dominant in political science. These perspectives, though varied, have in general eschewed the literature stemming from the field of social movement studies, though this has been less the case in recent years. The second outstanding difference between the two fields is conceptual, or more precisely, one about labels. Social movement studies avoid the usage of the label 'terrorism' (as well as the related 'terrorist' and 'terror') while obviously the field of terrorism studies thrives on it. For social movement scholars, this label is too political – a tool which sides in a contention employ to degrade their opponents (Tilly, 2004; Porta, 2013). While all labels (and terms) are value-laden, this particular label, the argument goes, is so heavily political that it is impossible to use it in neutral terms. While differences on theory and labels separate the two fields, however, it must be noted that a relatively small body of scholarship claims a position in the middle. This is the scholarship that utilizes both the term 'terrorism' and the analytical tools and theories of the social movement field of study.

The exposition that follows focuses on three areas of inquiry: the emergence of social movement-related violence, the disengagement from it, and the effects of it. This exposition is predominately concerned with the field of social movement studies, and this is done mostly out of reasons of economy. The literature that directly or indirectly refers to social movement-related violence is immense, including not just terrorism studies but also approaches to social conflict that range so widely that practically touch on all divisions in the social sciences. Only with regard to the first area of inquiry, on the emergence of violence, do we engage in a more comprehensive discussion. Given that this is the most researched area with respect to social movement-related violence, we juxtaposes insight from the field of social movement studies with insight from the field of terrorism studies. The other two areas of inquiry are far less researched and so we devote less space to them below.

The Emergence of Political Violence

Most of the attention of the field of social movement studies concerned with questions of political violence has been on its emergence. This attention covers both violence by social movements and violence by the state, in contrasts with the field

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96040-9

of terrorism studies, which tends to ignore state violence (unless it is counterterrorism).

Macro Level

A tradition of explanations of the emergence of social movement-related violence exists among mainstream terrorism studies at the macro level of analysis. In this line of studies, attention is placed on conditions that are external to the perpetrators of violence, which is to say, environmental conditions. These are considered to be root causes or permissive factors, setting the stage for political violence to emerge. Accordingly, such causes or factors are sought in the international system, the political culture, or the economy and other material conditions. While these causes/factors are often treated as conditions – hence structural elements affecting the likelihood for violence – processes are sometimes evoked too. In these instances, the processes typically regard modernization, sometimes its advance but more often the interruption of its advance. Either as structural conditions or as processes, these explanations betray a bias of focus, as they eschew state violence. Typically, attention is placed squarely on the populations outside the state that use or may use violence against the state. Hence environmental conditions create 'populations at risk,' as per an often-used label: the international system is said to create dominated and indignant populations, the economy populations with grievances, material factors deprived populations, political culture belligerent populations, and so on.

Guided by the positivistic epistemology that characterizes much of political science, especially of earlier years, these conditions are often presented as building-blocks to theory by way of discovery of necessary (sometimes even sufficient) conditions. While thought-provoking in their attempts to single out some root causes, however, neither one nor any combination of these conditions can be recognized as necessary (let alone sufficient). The evidence does not allow this; political violence clearly occurs in rich as well as in poor regions, in modern industrialized countries as well as in less developed ones, and so on. Indeed, despite the long line of research in macrolevel terrorism studies, no general theory has emerged to claim wide scholarly adherence. In fact, if one gives this scholarship a careless reading, one can easily be misled. For as Tore Bjorgo writes, the "limitation of the 'root cause' approach is that it may give the impression that terrorists are merely passive objects of social, economic and psychological forces: puppets obeying what these causes compel them to do" (2005: 3). Still, considered through nonpositive epistemology, this scholarship can be recognized to have enriched the understanding of the environments fostering violence.

The social movement approach has offered only a partial alternative to the search for macrolevel conditions characterizing some terrorism studies. It is a partial alternative because usually – though not always – it is not exclusively a macrolevel approach but rather an approach combining the macro and meso levels, and sometimes the micro level as well. The early collective action models that paid attention to the macro level were structural – hence hardly alternatives when epistemology is concerned. In particular, much of their attention was on the investigation of political opportunity structures, usually understood in terms of power configuration. But as structural models gave way to models of political process, attention turned to dynamics that connect different levels of analysis. Charles Tilly is one of the protagonists in this development, advocating and applying a systematic and multilevel exploration at the intersection of political violence and collective action. His approach, like much of the contemporary social movements approach, is organizational and relational, which means that chains of relations are treated as the 'glue' holding the various levels of analysis together (Tilly, 1998, 2003). In this approach, the causal dynamics are expected to be varied and even heterogeneous, though typically relational. At the macro level, they generate such figurations as those defining political fault lines and identities as well as power-asymmetries, which is to say figurations which are also articulated at the organizational meso level.

Just as with its epistemological scope, the social movements approach has a broader empirical scope than terrorism studies. Specifically, social movement scholars have been more ready than terrorism scholars to include state violence in their studies. Some, like Tilly, advocate the application of the same analytical tools to the study of both state and antistate violence. Others fix their focus on empirical studies of state violence. From these studies, in fact, a growing tradition of macroscopic research on state repression has developed (Porta, 1995). Thus, for example, one brunch in this tradition is concerned with historical patterns of policing protest, and another with the institutionalization of police autonomy.

Meso Level

Mainstream terrorism studies have developed a tradition in mesolevel analyses as well, focusing at the level of the group waging violence. Two distinct approaches have dominated this tradition: the ideological perspective and the instrumental perspective. The first approach tends to define armed groups as ideology-driven conspiracies against the political system. The members of these groups are considered to be attached to ideological precommitments, and their action is explained in terms of clashes of ideology or through ideology's role in recruitment, popular support, activism, selection of tactics, and so on. The second approach considers the members of armed groups to be instrumental actors, hence recognizing them the same capacity for rational choice as it is presumed to characterize all other political actors. The action of these groups is therefore explained through political and tactical calculations connecting means to ends. Whereas the first approach implies that devout adherence to ideology erases or reframes possible scruples regarding the use of violence, the latter approach assumes flexible end-game preferences of which violence is dispassionate calculated means.

These two traditions in terrorism studies are often considered contradictory to each other. In a sense, however, they are similar. For in both of them the primary unit of analysis relates to the realm of mental, cognitive operations. By contrast, the epistemology of the social movements approach tends to be both broader and more attentive to organizational analysis. This approach, therefore, may be open to the analysis of cognitive operations – typically through frame analysis or other cultural analysis – but includes organizational analysis as well.

Social movement strategy formation is a central subject matter that is usually placed under comprehensive analytical frameworks. Organizational analysis, on its part, often distinguishes three areas: the social movement organization, the social movement, and the social movement context. Regarding the social movement organization, the analysis can focus on the structure of hierarchy, the form of decision making, organizational resources, and so on. Also of interest in this area of analysis are issues relating to the clandestine character typical of violent groups; this is a differentiating characteristic vis-à-vis nonviolent movement organizations and can therefore be a point of comparison between violent and nonviolent contention (Zwerman et al., 2000). Regarding the social movement, the analysis typically pertains to the relations among the social movement organizations comprising the movement. These are often competitive relationships – over resources, popular allegiance, division of labor, and more broadly power (Alimi et al., 2012; Demetriou, 2007). In terms of violence, and in particular the switch from peaceful to violent tactics, analysis in these two areas may reveal consequential organizational splintering. The creation of the Provisional Irish Republican Army (PIRA) out of the ranks of the Irish Republican Army is an example of splintering at the level of social movement organization, and the creation of the Red Brigades out of the Italian Left movement is an example of splintering at the level of the social movement; in either case, the splintered organization proved particularly violent. Regarding the social movement context, too, the analysis can be, as it often has been, fruitful. For here one can examine the resources, opportunities, and constraints facing the movement, as well as the relations between the movement and interested publics or the state authorities. To appreciate how pregnant with consequences these relations are, just consider the last one – the relation between the movement and the authorities. This relation may take the form of repression by the authorities which research has shown to be an important factor in instigating political violence. Especially when the repressive measures appear to be hard, illegal, or confrontational, they may delegitimize the state and encourage antistate hostility and sympathy for armed groups (Porta, 1995; Wood, 2003). For example, the activities of the nationalist insurgency in Ireland in 1919–21 were often met with brutal reprisals by the state security forces, aimed at populations perceived as hostile; given that the main goal of the insurgency was the delegitimization of the state, the state through its tactic of reprisals was in effect doing the insurgents' bit (Demetriou, 2012).

In short, the meso level of analysis is the quintessential level of organizational analysis and the one through which social movement studies have thrived. To be sure, not all studies on violent social movements pursue as comprehensive organizational analyses as suggested above, but the developing trend is toward comprehensiveness (e.g., Alimi et al., 2012).

Micro Level

At the micro level of analysis, the main quest of terrorism studies is toward the understanding of the motivations of the activists engaging in violence. This follows a more general tendency, common outside social scientific academia and within the circles of policy-makers interested in counterterrorism, to offer motivation-type answers to the question "Why action happens?" A rich tradition therefore exists offering psychological explanations of the resort to violence. Reflecting an interest with rational choice theory, some scholars in this tradition bestow the motivation of greed on the activists whom they study. Combined with the activist's instrumentality, greed is therefore said to explain the choices in careers in violence. Other scholars in this tradition hold grievances to be the prime motor of motivation, while still others expand the pool of motivations to cover emotions, including emotive reactions to material conditions. Still other scholars develop violence-prone personality theories, variously privileging such psychological patterns as dependency, circular reaction, or the seeking of identity. Implicitly or explicitly, much work in this latter research agenda treats political violence as psychopathology.

Within social movement studies, the tendency is not to engage in microlevel analysis in isolation from broader analysis (Bosi and Porta, 2012). Here exists as well, however, a long-standing interest in the study of activists' motivations. A body of research covering activists' narratives exists in particular, though given the practical difficulties associated with interviewing active perpetrators of violence, these tend to be distant *post factum* accounts by ex-activists. But where the scholarship goes beyond the mere representation of the activists' voice, as it often does, then broader analysis is engaged in order to draw conclusions on collective behavior. Thus such standard analytical modalities as the examination of opportunity structures, of in-group and ramified relationships, and of framing are broad into the equation. Indeed, social movement research repeatedly demonstrates that the individual choice to join an armed group is better explained as collective choice. Thus, it is demonstrated that preexisting social relationships tend to influence greatly individual decisions to join armed groups. Also, it is shown that small networks of people with dense ties with each other (family relations, group of friends, school, local community, and so on) tend to progressively justify the need to use violence and through this they feel less responsible as single individuals (Bosi, 2012; Viterna, 2006; White, 1989).

The Disengagement from Political Violence

What disengagement from social movement-related political violence may mean exactly is open to interpretation. Phenomena of disengagement from violence have affinity with phenomena of political deradicalization and of cycles of violence – after all, it is often the passage of time that tells definite disengagement from temporary disengagement. Thus social movements scholars can, as they often do, study these phenomena indiscriminately (Fillieule, 2010). Nevertheless, a literature on disengagement from political violence does exist as such, and it tends to focus on marked behavioral modification on the part of militants or armed organizations. In this sense, disengagement does not imply necessarily change in values or ideals, but clear change in the means of contention (Bosi, 2013).

Macro Level

Given their clandestine character, armed groups and activists tend to progressively isolate themselves from the world around them. This may lead them to increasingly supercilious interpretations of their role in politics, highlighting the wisdom of their actions and the follies of their opponents. Still, these activists are never totally disconnected from the world. Armed groups are always and inevitably immersed in the conditions of their environment. There is therefore reason to study the environmental conditions that affect the likelihood for armed groups' disengagement from violence. In general, such conditions may relate to the closing of political opportunity for the groups or for the movements in which the groups are part, and to the falling supply of the elements that are vital to the survival of the group (membership, armaments, funding, and so on) or the movement (public support, alliances, and so on). What needs to be underlined regarding this course of macrolevel inquiry is that the focus is not so much on the activists' attribution of environmental changes (which may elude research, if it is done in real time) as on an objective assessment of these changes.

State repression is particularly important for the evolution of armed groups, since it can present a barometer of sorts for the political opportunities at hand. However, the available body of research does not offer strong evidence to suggest that repressive policies on their own lead to disengagement from violence. But then again, any isolated environmental condition carries limited efficacy in explaining (or forecasting) disengagement. Combinations of macro factors are more efficacious explanans. But more efficacious still are macrolevel analyses that are combined with mesolevel analyses. This is so because macrolevel analyses may account for the likelihood but not the timing of actual disengagement, where timing is typically analyzed through meso lenses. Thus a repressive policy needs to be analyzed in connection with its timing, which is to say, as a matter of conjuncture more than structure.

Meso Level

At the meso level of analysis, research can turn to an array of factors, emulating research done with regard to the emergence of political violence. Indeed, one may expect to find that the reversal of mechanisms precipitating violence may precipitate disengagement from violence. As noted at the introduction, however, the body of research on disengagement is rather limited, at least within the social movements literature. To the extent that one can talk about directions forming within this body or research, one can note an emerging attention to armed groups' strategy.

It is not a forgone conclusion that the use of political violence is met with public criticism. In legitimate polities, however, the likelihood that this may happen vis-à-vis antistate violence is high. In these situations, armed groups may find themselves in crossroads necessitating the adjustment of their strategy. Thus, on the one hand, they may opt to further distance themselves from public, including potentially their own constituency, by 'purifying' their stand in order to sustain activists' commitment in the face of increased adversity. On the other hand, they may opt to become more receptive to public opinion and the pull toward moderation which public opinion may represent, especially if this opinion comes from the movement constituency. Whereas the former outcome would lead to the increase of violence, the latter may lead to a progression away from violence. Facing choices such as these, armed groups are constrained by both external dynamics (political opportunities and public support) and internal dynamics (resources and intragroup dissent), all of which are complicated. The pathways toward or away from disengagement are therefore open-ended. For example, groups weak in material resources may react to opposition from their environment by further radicalizing their tactics, since this may shore up their organizational integrity. This may then accentuate group isolation to the point of the group becoming a sect, out of touch with political reality, and provoke revulsion among its actual or potential constituency. Such developments, in turn, may lead to defections from the group, which is to say, individual disengagement from political violence. This, in its turn, may lead to group disengagement form violence, further radicalization of tactics by the group, or group split. Relevant in this sort of dynamics, moreover, is the state counterpolicy, including counterviolence. The course of the movement–state interaction, too, is open-ended. Heavy-handed state policies, for example, certainly influence the course of contention, but not in a preordained direction: they may weaken or strengthen the resolve to antistate violence, reduce or enhance popular support, and so on.

Micro Level

To the extent that microlevel analysis is considered to pivot on psychology, disengagement from violence implies turning points in personal histories whereby commitments, emotions, and rationales change. A psychological barrier needs to be crossed. Typical of the social movements approach in general, however, psychological factors are here examined in relation to collectivities and their environments. Hence the attention is not so much on personalities and other attributes of individuals, but on the interactions among individuals, such as the breaking up of strong ties of friendship, loyalty, and trust. Projection on the future is relevant, too, and so the risk of social isolation can be a pivotal factor in 'changes of heart.' In other words, according to this line of analysis, maintaining or abandoning commitment to political violence depends on the emergent social psychological price entailed in the alternatives.

As with dynamics at the mesolevel of analysis, however, dynamics at the microlevel of analysis are not considered deterministic. Thus, for instance, if activists develop the view that old-standing violent tactics are no longer justified in the face of new political developments, this does not mean that they are able, or even willing, to disengage. Feelings of belongingness, the incapacity to see another possible future apart from that in the armed group, and the fear to live with the bad consciousness of having believed in wrong ideas, all may prolong activists' participation in armed groups. Indeed, perhaps the single most efficacious predictor of individual behavior is collective behavior – individuals would more likely commit to drastic change in their behavior if they do so in company.

The Consequences of Political Violence

The research within social movement studies on the consequences of violence is lacking behind the research on disengagement – let alone that on the emergence of political violence. This scarcity of research is striking if we consider that the purpose of the use of violence against the state, or by the state against social movements, is typically to elicit or prevent change. It is all the more striking given that the range of targets is wide. Antistate violence tends to target agents of the state, and violence by the state tends to target agents of social movements. But the two sides also aim at proxy targets – businesses, the media, constituencies – either because they are symbols, knots of power, or simply available. As with the literature examined above, the available social movement literature on the consequences of violence is skewed toward the study of antistate violence. Within the limits of this bias, however, the literature brunches out toward four areas – the areas of the polity, the economy, the society, and biography (Bosi and Giugni, 2012).

Political Area

While political violence typically affects policy, the magnitude of policy change may range greatly. On the one extreme, the change may entail the transformation of the state itself. Historically, this has been especially the case with regard to separatist movements, and the wave of decolonization in the 1960s offers many acute examples. Even separatist movements that fail their ultimate goal may precipitate the limited transformation of the state. For example, partly as a consequence of the campaign by the group Basque Homeland and Freedom, a considerable measure of autonomy was granted to the Basque region (Herrera, 2002). On the other hand, violent political contention may produce what from the claimants' point of view are negative, regressive effects. The massive and essentially permanent dislocation of Kurdish populations in the eastern part of Turkey by the government during the 1980s, designed as a response to the campaign of the Kurdistan Workers' Party, offers a prominent example of this.

More modestly, the effects of violent political contention may regard anything from public order policy (censorship, the banning of organizations, etc.) to legislative outputs (legislation as well as change of legislative procedures), to executive outputs directed at various targets including the armed groups' constituencies. These policies may outlast the campaigns of violence. Public order policy, typically aiming at treating the effects of violent political contention, may remain in effect permanently. Forms of policy aiming at the perceived causes of such contention may as well have long-lasting effects. For example, from 1972 onward and through the 1980s, the UK government made efforts to redress some of the grievances of the nationalist community in northern Ireland with the hope of undermining the campaign by the PIRA. But as it turned, the general direction of this policy, which included fairer hiring practices in the public sector and electoral reforms, has been the same as that which the Good Friday Agreement of 1998, ending the campaigns of violence, has promoted (Bean, 2007).

Economic Area

Violence obviously affects negatively the economy of the relevant state. It can damage property, prevent the use of capital and movement of goods, cause inflation and unemployment, and more generally inhibit economic growth or bring about outright depression. These effects can have negative regional or global economic ramifications, though the reverse might also be true – e.g., the loss in the tourist industry of a conflict-ridden state may be the gain in the tourist industry of a neighboring state. Analysis can distinguish between direct effects in the short run – including economic effects relating to increased medical provisions as well as to the direct loss of property and the like – to less direct effects in the long run. The latter sort of effects, typical of protracted violent political contention, may relate to anything from the deterioration of the health and educational systems to the loss of workforce through changes in migration patterns. In general, it can be said that the effects of political violence on the economy are negative but not all sectors are affected equally, while some sectors – notably those providing security services, infrastructure, consultancy, or news coverage and commentary – may benefit.

Social Area

The scope of social change resultant from the use of social movement-related violence is very wide indeed, while the direction of change is not preordained. Obviously the intensity, spread, and duration of violence are factors making a difference, and certain changes are expected to follow certain types of violence – e.g., emigration to follow intense violence against a population. Still, the direction of social change can vary. For instance, the levels of social trust or social integration may be reduced in the face of the dangers associated with the use of violence. Yet, at the same time, opposite phenomena may transpire vis-à-vis segments of a population. Thus the use of political violence may increase communal identification and improve the morale of discriminated communities. This seems to be the case, in fact, for many Muslims in Europe, especially of younger generations, who feel empowered by radical Islam, even when they condemn political violence in the name of Islam (Roy, 2004).

Nor are the patterns of public opinion transformation clear. Amid political violence, the prevailing change in the short run may well be the polarization of opinion. But in protracted conflicts, this direction of change may not hold. For example, some of the long-term effects of the Palestinian armed campaigns in the occupied areas in recent decades (the Intifadas, the Tanzim movement, and the campaigns by Hamas) point to the direction of moderation. According to one study, these campaigns contributed to the evolution of the political orientations of local populations by increasing their willingness to accept a Palestinian state and make territorial concessions, and by decreasing right-wing identification (Gould and Klor, 2009).

Biographical Area

The effects of violence on those affected by violence is, strictly speaking, a subset of what has just been discussed – i.e., effects

in the social area. But these effects also present an area of study offering particular insight on broader issues of social movement-related violence, not the least the question of disengagement. On the one hand, under study can be former activists, that is, individuals who had perpetrated violence. Even at low levels of activism, these individuals may face life-long consequences. An elementary yet important question is whether they honor or denounce their old commitments. Postconflict settings show that both sets of individuals are likely to be present and to complicate even more the political tangle associated with efforts at 'truth and reconciliation' – not just issues pertaining to the revision of the meaning of one's life, but also ones pertaining to culpability, punishment, and social closure.

Further, the victims of violence and those related to them may likely face life-long consequences. Much of what said above regarding postconflict contexts applies here too, including the broader question of how the personal experience and the collective experience are linked. Trauma, for example, has both a personal and a collective dimension, becoming in interconnected fashion a point of reference for both the victimized individuals and the communities in which they belong. The social movement literature only made limited inroads in this direction of research, despite the fact that this literature's rich insight on the interconnection between individual and collective action offers a spring board. Of course, the broader literature concerned with questions of victimhood and conflict – in psychology, social psychology, collective memory studies, and so on – is rich and also offers examples for the social movement scholars to emulate.

Future Directions

It is seen that the field of social movement studies has made gains in the area of violence, especially regarding the emergence of violence. At the same time, it is clear that biases and lacunae also exist in the field – i.e., more attention on social movement violence than on state violence, and insufficient attention on the disengagement from violence and on the effects of violence. Moreover, it is seen that the inquiries in the field are also pursued by other fields, especially terrorism studies, with divergent results.

Two opposing cases can be made for the future of the investigation of social movement-related violence. The one is in favor of pluralism. According to this position, the existing pluralism among the different fields in terms of theories and analytical frameworks is healthy and must be safeguarded. The other case is the opposite. That is, this pluralism is not healthy because it is the result of lack of knowledge across the fields and the academic disciplines which, to some extent or another, align with the fields; the pluralism must therefore be replaced by more unifying sets of theories.

Current trends exist in both directions. The present pluralism is sustained not so much by the intension of the participating scholars as by factors having to do with the general organization of academia; the overall trend is therefore toward the inadvertent perpetuation of pluralism. At the opposite end, various scholars promote broad analytic frameworks intended for research in different traditional fields. The 'Contentious Politics Approach' is the most prominent result from this kind of efforts. Promoted by scholars in the field of social movement studies – Charles Tilly, Sidney Tarrow, and Doug McAdam, among others – the approach argues that the delineation of research according to the concept 'social movement' (as well as 'terrorism') must give way to delineation according to the concept 'contentious politics' (McAdam et al., 2001). This concept is meant to denote any contention outside the formal channels of government. It therefore covers a broader category of phenomena than the category 'social movements,' including episodic contention by way of labor strikes and ethnic conflict. The advocates of this approach hold that such a wide range of forms of contention can be analyzed through a coherent analytical framework, thus allowing comparisons among phenomena which traditionally have been considered incomparable.

Further work is needed before this and other similar approaches become mainstream and replace more specialized fields; indeed, the level of the future success of such approaches might be measured by the extent that they entice members of the field of terrorism, since this is not only a large field but one which operates on peculiar sets of incentives. For the moment, it appears that the intellectual movement toward the comprehensive study of violence is gaining ground. If this is the case, however, we believe that a new challenge will emerge. To the extent that an enlarged pool of phenomena covering various forms of violent and nonviolent actions is treated as comparable, the need to arrive at new conceptual distinctions will arise. The challenge ahead might be, then, to produce conceptual differentiation that, while analytically integrating, allows for superior classification of phenomena.

See also: Civil Wars; Conflict and Consensus; Insurgency; Peace Movements; Revolutions, Sociology of; Securitization; Social Protest; Terrorism; Urban Conflicts; Violence, History of.

Bibliography

Alimi, E., Bosi, L., Demetriou, C., 2012. Processes of radicalization, a comparative approach. Mobilization 16 (4), 7–26.

Bean, K., 2007. The New Politics of Sinn Féin. Liverpool University Press, Liverpool.

Bjørgo, T., 2005. Root Causes of Terrorism: Myths, Reality and Ways Forward. Routledge, New York.

Bosi, L., 2012. Explaining pathways to armed activism in the provisional IRA, 1969–1972. Social Science History 36 (3), 347–390.

Bosi, L., 2013. Safe territories and political violence: the persistence and disengagement of violent political organizations. Nationalism and Ethnic Politics 19 (1), 80–101.

Bosi, L., Porta Donatella, D., 2012. Micro-mobilization into armed groups: the ideological, instrumental and solidaristic paths. Qualitative Sociology 35, 361–383.

Bosi, L., Demetriou, C., Malthaner, S., 2014. Dynamics of Political Violence. Ashgate, Farnham.

Bosi, L., Giugni, M., 2012. Political violence outcomes: a contentious politics approach. Mobilization 18 (1), 85–98.

Demetriou, C., 2012. Political radicalization and anti-colonial violence in Palestine, Ireland, and Cyprus. Journal of Social Science History 36 (3), 391–420.

Demetriou, C., 2007. Political violence and legitimation: the episode of colonial Cyprus. Qualitative Sociology 30, 171–193.

Fillieule, O., 2010. Some elements of an interactionist approach to political disengagement. Social Movement Studies 9 (1), 1–15.

Goodwin, J., 2004. Review essays: what must we explain to explain terrorism? Social Movement Studies 3, 259–265.

Gould, E., Klor, E., 2009. Does Terrorism Work? Unpublished Paper (No. 7589), Centre for Economic Policy Research.
Herrera, E., 2002. Nationalist extremism and outcomes of the state public policies in the Basque country, 1969–2001. International Journal on Multicultural Societies 4 (1), 17–41.
McAdam, D., Tarrow, S., Tilly, C., 2001. Dynamics of Contention. Cambridge University Press, New York.
Oberschall, A., 2004. Explaining terrorism: the contribution of collective action theory. Sociological Theory 22 (1), 26–37.
Porta, Donatella, D., 1995. Political Violence and the State. Cambridge University Press, New York.
Porta, Donatella, D., 2013. Clandestine Political Violence. Cambridge University Press, Cambridge.
Roy, O., 2004. Globalised Islam: The Search for a New Ummah. C. Hurst and Co, London.
Tilly, C., 2004. Terror, terrorism, terrorists. Sociological Theory 22 (1), 5–13.
Tilly, C., 2003. The Politics of Collective Violence. Cambridge University Press, Cambridge.
Tilly, C., 1998. Micro, macro, or megrim? In: Mikrogeschihte, Makrogeschichte: komplmentar oer inkommensurabel? Mit Beitragen von Maurizio Gribaudi, Giovanni Levi und Charles Tilly. Wallstein Verlag.
Viterna, J., 2006. Pulled, pushed, and persuaded: explaining women's mobilization into Salvadoran guerrilla army. American Journal of Sociology 112 (1), 1–45.
White, R., 1989. From peaceful protest to guerrilla war: micromobilization of the provisional Irish republican army. The American Journal of Sociology 94 (6), 1277–1302.
Wood, E., 2003. Insurgent Collective Action and Civil War in El Salvador. Cambridge University Press, Cambridge.
Zwerman, G., Steinhoff, P., Porta, Donatella D., 2000. Disappearing social movements: clandestinity in the cycle of new left protest in the U.S., Japan, Germany, and Italy. Mobilization 5, 83–100.

Relevant Websites

http://cosmos.eui.eu/Projects/cosmos/Home.aspx – COSMOS.
http://www.mobilization.sdsu.edu/index.html – Mobilization.
http://mobilizingideas.wordpress.com/ – Mobilizing Ideas.
http://polviolence.net/ – Political Violence Standing Group (ECPR).
http://www.tandfonline.com/toc/csms20/current – Social Movement Studies.

Social Movements, Geography of

Byron Miller, University of Calgary, Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social movements, like all social phenomena, are geographically constituted. Although social scientists increasingly stress the importance of the context of social action, the geographical constitution of social movements has received comparatively little attention. Increasingly, however, there are exceptions to this generalization. Contemporary approaches to the study of the geography of social movements proceed from the position that the social and the spatial are inseparable, and that social struggle is simultaneously geographical struggle. The resources and barriers that social movements draw upon and encounter – e.g., money, skills, collective identities, social networks, political opportunities, grievances – have their own geographies, giving rise to a geographically differentiated landscape of protest. Likewise, the social relations that social movements attempt to change are geographically specific; change in those relations often necessitates change in the geographic constitution of those relations. Social movements, explicitly or implicitly, must formulate geographically specific tactics and strategies that affect the likelihood of their success or failure.

Social movements, like all social phenomena, are geographically constituted. Although social scientists increasingly stress the importance of the context of social action, the geographical constitution of social movements has received comparatively little attention. However, there are important exceptions to this generalization. For instance, a number of scholars across the social sciences have engaged in projects to map the incidence and spatial diffusion of social movements. Such projects provide compelling evidence of the geographical specificity of the causes of social movement mobilization. These mapping projects, nonetheless, tend to treat the geographies of social movements simply as the outcomes of aspatial mobilization processes or as the 'containers' in which mobilization processes occur.

More contemporary approaches to the study of the geography of social movements proceed from the position that the social and the spatial are inseparable and that social struggle is simultaneously geographical struggle (Soja, 1980; Giddens, 1984). The resources and barriers that social movements draw upon and encounter – e.g., money, skills, collective identities, social networks, political opportunities, grievances – have their own geographies, giving rise to a geographically differentiated landscape of protest. Likewise, the social relations that social movements attempt to change are geographically specific; change in those relations often necessitates change in the geographic constitution of those relations. Social movements, explicitly or implicitly, must formulate geographically specific tactics and strategies that affect the likelihood of their success or failure.

What one means by 'social movements' is open to a wide range of interpretations. For present purposes social movements are defined broadly as collective mobilizations of people seeking social or political change. While social movements necessarily take place somewhere, their geography involves considerably more than an inventory of their locations. Geography is not simply a backcloth onto which aspatial social processes are mapped. Rather, it is a fundamental aspect of the constitution of all social processes, affecting their evolution and outcome directly. It is part of the explanation of social processes, as well as an outcome (Massey and Allen, 1984).

While a variety of geographic concepts can be employed in the analysis of social movements, four core geographic concepts are particularly important: space, place, scale, and networks. Each of these concepts has its own evolution, which can be seen in the ways in which analyses of the geography of social movements have changed over time.

Evolving Geographic Concepts and the Analysis of Social Movements

Spatial Analysis School of Social Movements

The spatial-analytic paradigm that dominated geographical analysis in the 1960s and 1970s worked from a conception of space that was based, by and large, on measures of distance. In this paradigm, social phenomena are described and explained in terms of absolute (e.g., with reference to a fixed grid such as latitude and longitude) or relative (e.g., transportation costs or travel time) location. Virtually all social phenomena have been shown to possess a 'distance decay' function as interaction costs increase with distance from the activity center. The spatial diffusion of information, as well as barriers to such diffusion, became a common theme in much spatial-analytic research, including that addressing social movements (*see* Diffusion: Geographical Aspects). The growth and mobilization of social movements is clearly dependent upon flows of information, as demonstrated by Sharp's (1973) analysis of the spatial diffusion of US postal strikes, and Hedstrom's (1994) work on the spatial diffusion of Swedish trade unions. Most spatial-analytic work on social movements, however, examines the copresence of specific social conditions within clearly delimited places. Adams (1973), for instance, examines the colocation of factors promoting urban protests, while Cutter et al. (1986) consider how characteristics measured at the state level are associated with support for nuclear weapons freeze referenda (*see* Peace Movements, History of). In the 1970s, attempts to explain the occurrence of social phenomena in terms of distance or

colocation came under heavy criticism for ignoring power relations and, in particular, the relation between 'the social' and 'the spatial.'

Contemporary Conceptions of Space and the Analysis of Social Movements

Influenced by the work of Henri Lefebvre and others, contemporary conceptions of space emphasize the mutual constitution of 'the social' and 'the spatial.' Space is seen increasingly as a necessary component of the constitution of power relations. Equally significant, space is no longer conceptualized primarily in terms of the physical location of material objects. Instead, both material spatial practices, as well as symbolic or representational spaces, are examined.

As political economy approaches swept through the social sciences in the 1970s and 1980s the geography of social movements was explicitly linked to the spatial constitution of material interests. The relationship between the geography of political and economic conditions and the incidence of protest is demonstrated in detail in Charlesworth's (1983) atlas of rural protest in the UK. Bennett and Earle (1983) show that the failure of socialism in the United States can be linked directly to geographic patterns of wage and skill differentiation; where differentiation was substantial, working class solidarity was undermined (*see* Labor Movements, History of). Similarly, Massey (1984) argues that the spatial division of labor directly shapes the material interests that underlie political movements. Markusen (1989) makes a somewhat different point: a region's stage in the profit cycle shapes the types of movements that are likely to emerge in it. Environmental movements are more likely to emerge in economically booming regions where rapid growth places stress on the environment, while progrowth politics are common in regions of decline (*see* Social Movements: Environmental Movements).

Geographically differentiated state systems also shape the geography of social movements. Kitschelt's (1986) comparison of the political structures of France, Sweden, West Germany, and the United States shows that differences in central state structures are related to national-level variations in antinuclear mobilization. Other studies link local-level variation in social movement mobilization and success to differences in political opportunity structures among local states (Duncan et al., 1988; McAdam, 1996; Tarrow, 1996). In the United States, many social movements began as highly localized movements in politically opportune places, then spread nationally (*see* Citizen Participation).

Symbolic or representational spaces affect the geography of social movements as well. The symbolic codings of spaces, i.e., the often subtle and unacknowledged signs that shape people's reactions to spaces, may indicate belonging for some groups and exclusion for others, empowerment for some groups and repression for others, safety for some groups and danger for others. Such codings can foster or hinder social movement mobilization. The creation of symbolic space is intertwined with the process of collective identity formation, as Marston (1988) shows for ethnic identity, Rose (1993) demonstrates for gendered identity, and Ruddick (1996) illustrates for interlocking 'race,' class, and gender identities (*see* Social Movements and Gender). Such identities are crucial to the formation of solidarity necessary for movement mobilization (*see* Social Movements, History of: General). The dominant symbolic codings of space are open to contestation, and counter-representations are often constructed by insurgent groups (Cresswell, 1996). Representational spaces, and their internalization, relate directly to the ways in which social movement messages resonate in some places, and not in others.

Place and the Analysis of Social Movements

Many geographic analyses emphasize the place-based constitution of social processes. Place is best understood as a multidimensional concept entailing 'locale,' i.e., the geographical setting of everyday social relations, 'location,' i.e., geographical position vis-à-vis wider scale interactions; and 'sense of place,' i.e., the esthetic and emotional bonds to locale (Agnew, 1987). The concept of place stresses the unique ways in which social processes may interact in specific milieus, calling our attention to the 'situatedness' of all human action and institutions. At the same time, place is not to be equated with 'the local.' Place implies the articulation of localized processes, with other processes operating at broader scales.

Much of the geographical literature on social movements stresses the ways in which identities and allegiances are shaped in place. Thrift and Williams (1987) analyze the ways in which consciousness and identity are produced through place-specific social practices and the time–space paths of individuals' everyday lives. Katznelson's (1981) work can be interpreted as a place-sensitive analysis of US workplace and community politics, stressing how distinct places shape identity in distinct ways. Savage's (1987) analysis of British working-class politics demonstrates how the interactions of workplace and lifeworld spaces have effects on identity that vary from place to place. Recent work on identity construction stresses the place-specific interplay of spaces of production, reproduction, and the state that produce multiple, fragmented, and sometimes contradictory identities, all with direct implications for social movement mobilization (Ruddick, 1996; Cope, 1996; Brown, 1997). Place itself can be a primary basis for collective identity construction. The defense of places associated with specific collective identities is a common theme of many new social movements (Melucci, 1994; Martin, 2003). Increasingly, cities find themselves in competition with each other for capital investment; in response, place-based identification and coalition building has often taken precedence over class and other axes of social identity (Hudson and Sadler, 1986; Swyndegouw, 1989).

Identity is not all that is shaped in place. Agnew (1987: 59) observes that "different places produce different degrees and orientations of organizational capacity." Routledge (1993), for example, shows that social and environmental movements in India faced opportunities for, and constraints to, mobilization that relate to place-specific social networks, systems of meaning, and patterns of political repression. Miller (2000) shows how a variety of place-specific characteristics, including political opportunity structures, available resources, material interests, and attachments to place, shaped the strategies and effectiveness of peace-organizing differentially in several Boston area municipalities.

Scale and the Analysis of Social Movements

The concept of place clearly carries within it issues of geographic scale. Geographic scale has commonly been thought of as a fairly straightforward technical issue relating to the telescoping frame of geographic analysis (from local to global), but such a notion of scale as pregiven or unproblematic is strongly challenged, especially in political geography. While many analyses of social movements may assume that the appropriate scale of analysis is the nation-state, such an assumption overlooks the often critical subnational, as well as international, constitution of social movement processes. Moreover, rather than being seen as an 'external fact awaiting discovery,' scale should be viewed as a "way of framing [and acting on] conceptions of reality" (Delaney and Leitner, 1997: 94–95) that is employed by social movement organizations and their opponents as a way of gaining political advantage. As Smith (1993: 90) explains:

> The construction of scale is not simply a spatial solidification of contested social forces and processes; the corollary also holds. Scale is an active progenitor of specific social processes. In a literal as much as metaphorical way, scale both contains social activity and at the same time provides an already partitioned geography within which social activity takes place. Scale demarcates the sites of social contest, the object as well as the resolution of contest.

Tarrow and McAdam (2005) point to two mechanisms by which social movements shift geographical scales: 'relational diffusion' addresses the territorial broadening of social movements through preexisting social networks, while 'brokerage' refers to the building of new network connections beyond the initial geographical territory of a struggle. While Tarrow and McAdam focus on 'horizontal' scalar relations of social movement agents, other analyses of scale focus on 'vertical' state relations by which the capacities and responsibilities of local, regional, national, and transnational institutions of the state are constituted, challenged, and altered (Brenner, 2004), in turn altering the capacities – and relevance – of the state institutions that social movements engage. Social movements adopt scale-specific strategies in response to the presence of political opportunities, as well as potential allies, at different scales of the state (Cox, 1997; Miller, 2000; Tarrow and McAdam, 2005; Nicholls et al., 2013).

Networks and the Analysis of Social Movements

Since the early 2000s much of the geographic research on social movements has focused on the formation of social movement networks. The principal theoretical foundation for much of this research has been Bruno Latour's actor-network theory (ANT), which focuses on "uncovering and tracing the many connections and relations among a variety of actors (human, nonhuman, material, discursive) that allow particular actors, events and processes to become what they are" (Bosco, 2006: 136). ANT differs considerably from the predominantly structural sociological approach to actor-networks. Instead of mapping the structure of network connections, ANT takes a poststructural approach to understanding the fleeting associations of heterogeneous human actors and nonhuman actants, attempting to break down the boundaries between the human and nonhuman worlds, and between structure and agency. Dissolving boundaries and tracing the formation of relationships is at the heart of this topological approach to network formation, an approach typically considered as 'relational' and counterposed to territorial notions of place and space. ANT approaches blossomed with the rise of antiglobalization movements that involving transnational organizing and, indeed, ANT-inspired approaches have yielded important insights into how antiglobalization movements organize and build identities and solidarities across national borders. Routledge (2003), Featherstone (2008), and Routledge et al. (2013) show how shared identities and solidarities are dynamically constructed and reinforced through transnational mobilization processes and that geographical notions of identity formation must consider the extralocal networks that are critical components of place-based social relations. Nonetheless, networks do not stand outside constructions of territory, scale, and place, which are themselves relationally constructed.

Despite the substantial insights yielded by each of the geographically oriented bodies of social movements research, there has been considerable debate over the ostensible superiority of one approach or another. But as Nicholls (2009: 78) argues,

> What has been interesting about these debates is not that one position is more persuasive than the other but that these various positions have contributed to revealing the mechanisms and dynamics associated with the various spatialities of human relations. The central analytical task at hand is therefore not to show how one form of spatiality is more important than another, but rather to show how these spatialities articulate with one another in actually existing social movements.

Geographic Tactics and Strategies of Social Movements

Given that all social movements are geographically constituted, all social movement organizations must make decisions regarding the geographic tactics and strategies they employ. Sometimes these decisions are explicitly considered, often they are not.

Social movements frame their messages in particular ways that they hope will resonate with particular groups (Martin, 2003). Given highly differentiated geographies of culture, symbolization, and interests, the resonance of any given frame will depend in substantial measure on the geographic targeting of the message. Similarly, the recruitment of social movement members is inextricably intertwined with the geography of mobilization activities (Harvey, 1985; Miller, 2000; Nicholls et al., 2013).

While identities constructed in place and place-based identities frequently are central to social movement solidarity, the absence of such bases of solidarity can be rectified through efforts to construct new places of group interaction and identity. Castells (1983), for example, details how the growth of gay bars, businesses, and gathering places was crucial to the construction of shared identity in the San Francisco gay

community. Castells also illustrates how, in the case of the Madrid citizens' movement, the construction of a community center in the Orcasitas shantytown was essential to breaking down social walls and building broad, place-based bonds around which the movement organized. Moyersoen and Swyngedouw (2013), going a step further, show how the creation of an innovative community center can actually serve to 'broker scales,' opening up entirely new opportunities for the creation of broader scale social interactions, negotiation, and bonding. Indeed, places – and especially cities – can be thought of as 'relational incubators' through which both strong and weak bonds are formed even as the processes that shape these places stretch around the globe (Miller and Nicholls, 2013).

The tactics and strategies that social movements employ to challenge the existing social order are rooted in an often-unexamined normative geography. Cresswell (1996) makes a compelling case that spatial structures and systems of places provide ideologically loaded schemes of perception. Dominant conceptions of what is 'normal' and 'appropriate' is defined "to a significant degree, geographically, and deviance from this normality is also shot through with geographic assumptions concerning what and who belong where" (p. 27). He illustrates his argument with several examples, including that of the Greenham Common women's peace camp that challenged British participation in the nuclear arms race through 'out of place' behavior that "denaturalized the dominant order [and] showed people that what seemed natural, could, in fact, be otherwise" (p. 125). Similar challenges to the authority of the state and of finance capital occurred, respectively, through the occupation of public squares across North Africa during the Arab Spring and of Zuccotti Park during the Occupy Movement.

Indeed, social movements can be mobilized and social conditions can be changed through a resymbolization of existing spaces or through a reterritorialization of existing symbolization (Pile and Keith, 1997). Harvey (1989: 233) makes a case for the resymbolization of space when he argues:

> If workers can be persuaded … that space is an open field of play for capital but a closed terrain for themselves, then a crucial advantage accrues to the capitalists … power in the realms of representation may end up being as important as power over the materiality of spatial organization itself.

An example of the reterritorialization of existing symbolization can be found in the work that has been done in several US metropolitan areas to recast the ongoing political battles between central cities and suburbs. In several metropolitan areas, this common conception of territorial interests is being replaced with one that aligns the central cities with the inner ring suburbs, creating a far more powerful territorial alliance behind the movement for a more equitable metropolitan distribution of affordable housing, school funding, and tax revenue (Orfield, 1997).

Many of the efforts of social movements to redefine the spaces of struggle involve attempts to 'jump' to a more favorable geographic scale (Smith, 1993). In addition to changing symbolic geography, altering the scale of material processes can be central to social movement strategy. Examples include shifting from regional to national-scale labor bargaining (Herod, 1998), and seeking out more favorable political opportunity structures at a different level of the state, e.g., local rather than central (Miller, 2000; Nicholls et al., 2013).

The geographic tactics and strategies adopted by social movements and their opponents play a powerful role in the nature of social movement mobilization and the degree of success or failure social movements achieve.

See also: Interest Groups, History of; Peace Movements; Political Protest and Civil Disobedience.

Bibliography

Adams, J., 1973. The geography of riots and civil disorders in the 1960s. In: Albaum, M. (Ed.), Geography and Contemporary Issues. Wiley, NewYork, pp. 542–564.

Agnew, J.A., 1987. Place and Politics. Allen & Unwin, Boston, MA.

Bennett, S., Earle, C., 1983. Socialism in America: a geographical interpretation of its failure. Political Geography Quarterly 2 (1), 31–55.

Bosco, F., 2006. Actor-network theory, networks, and relational approaches in human geography. In: Aitken, S., Valentine, G. (Eds.), Approaches to Human Geography. Sage, London and Thousand Oaks, pp. 136–146.

Brenner, N., 2004. New State Spaces: Urban Governance and the Rescaling of Statehood. Oxford University Press, Oxford.

Brown, M., 1997. Radical politics out of place? In: Pile, S., Keith, M. (Eds.), Geographies of Resistance. Routledge, New York, pp. 152–167.

Castells, M., 1983. The City and the Grassroots. University of California Press, Berkeley, CA.

Charlesworth, A. (Ed.), 1983. An Atlas of Rural Protest in Britain. University of Pennsylvania Press, Philadelphia, PA, pp. 1548–1900.

Cope, M., 1996. Weaving the everyday: identity, space, and power in Lawrence, Massachusetts, 1920–1939. Urban Geography 17 (2), I79–204.

Cox, K., 1997. Spaces of dependence, spaces of engagement and the politics of scale: or, looking for local politics. Political Geography 17 (1), 1–23.

Cresswell, T., 1996. In Place/Out of Place. University of Minnesota Press, Minneapolis, MN.

Cutter, S.L., Holcomb, H.B., Shatin, D., 1986. Spatial patterns of support for a nuclear weapons freeze. The Professional Geographer 38 (1), 42–52.

Delaney, D., Leitner, H., 1997. The political construction of scale. Political Geography 16 (2), 93–97.

Duncan, S., Goodwin, M., Halford, S., 1988. Policy variations in local states: uneven development and local social relations. International Journal of Urban and Regional Research 12 (1), 107–128.

Featherstone, D., 2008. Resistance, Space and Political Identities: The Making of Counter-Global Networks. Wiley-Blackwell, Malden and Oxford.

Giddens, A., 1984. The Constitution of Society. University of California Press, Berkeley and Los Angeles.

Harvey, D., 1985. Consciousness and the Urban Experience. Johns Hopkins University Press, Baltimore.

Harvey, D., 1989. The Condition of Postmodernity. Blackwell, Cambridge, MA.

Hedstrom, P., 1994. Contagious collectivities: on the spatial diffusion of Swedish trade unions, 1890–1940. American Journal of Sociology 99 (5), 1157–1179.

Herod, A. (Ed.), 1998. Organizing the Landscape: Geographical Perspectives on Labor Unionism. University of Minnesota Press, Minneapolis, MN.

Hudson, R., Sadler, D., 1986. Contesting work closures in Western Europe's old industrial regions: defending place or betraying class? In: Scott, A.J., Storper, M. (Eds.), Production, Work, Territory. Allen & Unwin, Boston, pp. 172–194.

Katznelson, I., 1981. City Trenches. University of Chicago Press, Chicago.

Kitschelt, H., 1986. Political opportunity structures and political protest: anti-nuclear movements in four democracies. American Sociological Review 49, 770–783.

Markusen, A., 1989. Industrial restructuring and regional politics. In: Beauregard, R. (Ed.), Economic Restructuring and Regional Politics. Sage, Newbury Park, CA, pp. 115–148.

Marston, S.A., 1988. Neighborhood and politics: Irish ethnicity in nineteenth century Lowell, Massachusetts. Annals of the Association of American Geographers 78, 414–432.

Martin, D., 2003. "Place-framing" as place-making: constituting a neighborhood for organizing and activism. Annals of the Association of American Geographers 93 (3), 730–750.

Massey, D., 1984. Spatial Divisions of Labor. Methuen, New York.

Massey, D., Allen, J., 1984. Geography Matters. Cambridge University Press, New York.

McAdam, D., 1996. Conceptual origins, current problems, future directions. In: McAdam, D., McCarthy, J.D., Zald, M.N. (Eds.), Comparative Perspectives on Social Movements. Cambridge University Press, New York, pp. 23–40.

Melucci, A., 1994. A strange kind of newness: what's 'new' in new social movements? In: Larana, E., Johnston, H., Gusfield, J.R. (Eds.), New Social Movements: From Ideology to Identity. Temple University Press, Philadelphia, pp. 101–130.

Miller, B., 2000. Geography and Social Movements. University of Minnesota Press, Minneapolis, MN.

Miller, B., Nicholls, W., 2013. Social movements in urban society: the city as a space of politicization. Urban Geography 34 (6), 1–22.

Moyersoen, J., Swyngedouw, E., 2013. LimiteLimite: cracks in the city, brokering scales, and pioneering a new urbanity. In: Nicholls, W., Miller, B., Beaumont, J. (Eds.), Spaces of Contention: Spatialities and Social Movements. Ashgate, Aldershot, pp. 164–187.

Nicholls, W., 2009. Place, networks, space: theorising the geographies of social movements. Transactions of the Institute of British Geographers 34, 78–93.

Nicholls, W., Miller, B., Beaumont, J. (Eds.), 2013. Spaces of Contention: Spatialities and Social Movements. Ashgate, Aldershot.

Orfield, M., 1997. Metropolitics: A Regional Agenda for Community and Stability. The Brookings Institution, Washington, DC.

Pile, S., Keith, M., 1997. Geographies of Resistance. Routledge, New York.

Rose, G., 1993. Feminism and Geography. University of Minnesota Press, Minneapolis, MN.

Routledge, P., 1993. Terrains of Resistance: Nonviolent Social Movements and the Contestation of Place in India. Praeger, Weatport, CT.

Routledge, P., 2003. Convergence space: process geographies of grassroots globalization networks. Transactions of the Institute of British Geographers 28 (3), 333–349.

Routledge, P., Cumbers, A., Nativel, C., 2013. Global Justice networks: operational logics, imagineers and grassrooting vectors. In: Nicholls, W., Miller, B., Beaumont, J. (Eds.), Spaces of Contention: Spatialities and Social Movements. Ashgate, Aldershot, pp. 298–324.

Ruddick, S., 1996. Constructing difference in public spaces: race, class, and gender as interlocking systems. Urban Geography 17 (2), 132–151.

Savage, M., 1987. The Dynamics of Working Class Politics: The Labour Movement in Preston, 1880–1940. Cambridge University Press, Cambridge, UK.

Sharp, V., 1973. The 1970 postal strikes: the behavioral element in spatial diffusion. In: Albaum, M. (Ed.), Geography and Contemporary Issues. Wiley, New York, pp. 523–532.

Smith, N., 1993. Homeless/global: scaling places. In: Bird, J., Curtis, B., Putnam, T., Robertson, G., Tickner, L. (Eds.), Mapping the Futures: Local Cultures, Global Change. Routledge, New York, pp. 87–119.

Soja, E., 1980. The socio-spatial dialectic. Annals of the Association of American Geographers 70 (2), 207–225.

Swyndegouw, E., 1989. The heart of the place: the resurrection of locality in an age of hyperspace. Geografiska Annaler 71B, 31–42.

Tarrow, S., 1996. States and opportunities: the political structuring of social movements. In: McAdam, D., McCarthy, J.D., Zald, M. (Eds.), Comparative Perspectives on Social Movements. Cambridge University Press, New York, pp. 41–61.

Tarrow, S., McAdam, D., 2005. Scale shift in transnational contention. In: Donatella Della Porta, D., Tarrow, S. (Eds.), Transnational Protest and Global Activism. Rowman and Littlefield Publishers, Lanham, MD.

Thrift, N., Williams, P., 1987. The geography of class formation. In: Thrift, N., Williams, P. (Eds.), Class and Space: The Making of Urban Society. Routledge & Kegan Paul, New York, pp. 1–22.

Social Movements, History of: General

L Tilly, New School for Social Research, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14360–14365, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The chief focus of this article is the history of social movements in the West that have attracted most scholarly attention; it draws particularly on studies done in the years 1975–2000 – a period in which the concept of social movement has been redefined and developed – for its theoretical framework and empirical base. The British movement for the abolition of slavery – an emblematic, and by some criteria the first social movement – serves as a concrete example of an early crusade that embodied many of the central features of the phenomenon. The article goes on to examine the movement for women's suffrage in the United States, Italian fascism, and transnational social movements such as proecology/environment organizations.

Social movements, according to sociologist Sidney Tarrow's definition, are "collective challenges, based on common purposes and social solidarities in sustained interaction with elites, opponents and authorities" (Tarrow, 1998: p. 4). This definition, derived from studies of both historical and contemporary social movements, locates the earliest movements in the late eighteenth and early nineteenth centuries. The emergence of social movements was contemporaneous with (1) the increased centralization of national states that accompanied war and revolution in Western Europe and North America; (2) the rise of the popular press and the circulation of notions of natural rights and liberty put forward by Enlightenment and revolutionary thinkers; and (3) particularly in England and the English-speaking colonies of North America, the rise of evangelical and nonconforming religions in which literacy and reading were cultivated and principles of human rights formulated. From that period and place, social movements have proliferated to the world.

The chief focus of this article is the history of social movements in the West that have attracted most scholarly attention; it draws particularly on studies done in the years 1975–2000 – a period in which the concept of social movement has been redefined and developed – for its theoretical framework and empirical base. The British movement for the abolition of slavery – an emblematic, and by some criteria the first social movement – serves as a concrete example of an early crusade that embodied many of the central features of the phenomenon.

The Abolition of Slavery in Great Britain

The eighteenth- and nineteenth-century British abolitionist movement was distinctive in that it was aimed at an institution that had only a minor presence in Great Britain, whereas most social movements have been concerned with issues close to hand. Slavery nevertheless involved many Britons; British ships sailing from Liverpool were active in the slave trade carrying slaves to Africa and slaves to the Americas, and the 'peculiar institution' was common in Britain's North American and universal in its Caribbean colonies. The question of the legality of slavery in Britain was raised as early as the 1720s, but it was not until 1772 that Lord Chief Justice Mansfield ruled that slavery was illegal under English law. English members of the Society of Friends had developed the earliest critique of slavery as contrary to Christian teaching, and local Quaker meetings first excluded slave traders in 1761 and later slave owners from membership.

Quakers then turned to political action to end British involvement in the slave trade, establishing in 1787 the Society for the Abolition of the Slave Trade, commonly known as the London Committee. This organization welcomed non-Friends as members, and recruited both middle- and working-class adherents. The political technique utilized by the Committee in this and later campaigns was mass petitions to Parliament. Thomas Clarkson, a gifted organizer, and other activists worked outward from London and local Friends meetings to form regional organizations, recruit members, distribute leaflets, and collect signatures on petitions to Parliament calling for the end of the slave trade. (Clarkson and his associates began their crusade with an exclusive focus on the slave trade, as less contentious than abolition.) Visual images portraying the horrors of the slave trade such as a print of the cross section of a slave ship with its tightly packed human cargo and Josiah Wedgwood's medallion of a kneeling chained slave, inscribed with the motto "Am I not a Man and Brother?" were highly effective in attracting supporters.

There was an outpouring of support, with about 60 000 male signers on the first petitions, including 11 000 men from Manchester, the center of the burgeoning cotton spinning industry. Parliamentary sympathy for the cause seemed strong up to 1792, when 519 petitions with up to 400 000 signers were collected. However, legislative support dropped with the slave revolt in Haiti, the contemporaneous radicalization of the French Revolution, and the outbreak of war against France. In 1794, the House of Lords decisively rejected the abolitionist resolutions presented to it.

Political opportunities for action on abolition reappeared in 1804, after Britain's victories in the Napoleonic Wars and its seizure of the French Caribbean colonies. Clarkson and other organizers again canvassed the country seeking signatures; he

also sought support from members of Parliament, in particular William Wilberforce, an Evangelical MP from Yorkshire who spoke eloquently about the evils of the trade and Samuel Whitbread, a nonconformist Protestant already known as a social reformer. Clarkson's powers of persuasion and his Parliamentary allies combined to produce in 1807 a majority vote in Parliament to forbid British ships' participation in the slave trade to both British possessions and foreign slave-holding colonies. The African Institution, founded in the same year, took as its task ascertaining the effectiveness of the British legislation.

Although British diplomats signed agreements eschewing the slave trade with other European powers, and indeed paid compensation when they did, the results were uneven. British activists resolved to continue their efforts, founding the Society for the Amelioration and Gradual Abolition of Slavery (known familiarly as the Anti-Slavery Society) in 1823 and the Agency Committee (1831), which demanded immediate and unconditional slave emancipation. The Agency's professional organizers traveled around England, promoting local petition drives and pressure on candidates in the Parliamentary elections to declare their support of full abolition. Members of the House of Commons elected under the 1832 Reform Act were ready to vote for the abolition of slavery in British possessions, but the House of Lords was not. The outcome was a compromise Emancipation Act of 1833 that set 1 August 1834, as the date on which slavery would end, but stipulated a complex gradual process to complete freedom.

The former slaves would become 'apprentices,' working for pay three-quarters of their time for their ex-masters; one-quarter of the ex-slaves' time was to be reserved for them to cultivate their newly granted plots. Slave owners would be compensated by the state for the loss of their 'property.' The story did not end there, however. Disagreements about strategy between the Anti-Slavery Society (which had accepted the compromise Emancipation Act) and the Agency (which had not) continued, and the latter established yet another organization (the Central Negro Emancipation Committee) and sought to persuade Parliament that oppressive conditions still prevailed. In the end, the complex system based on the 1833 Act proved difficult to manage, and the planters themselves ended apprenticeship in 1838.

The international abolition movement matured in the 1830s, and began to hold annual congresses (which some American abolitionists attended) in 1840. The United States Civil War (1861–65) was the outcome of a series of legal and political disputes about slavery and state's rights. In 1863, President Abraham Lincoln's Emancipation Proclamation ended slavery in the Confederacy, and during the postwar Reconstruction the Thirteenth, Fourteenth, and Fifteenth Constitutional amendments (which completed abolition, made the ex-slaves citizens, and guaranteed their right to vote) were ratified. The abolition of slavery in Cuba (1886) and Brazil (1888) ended legal slavery in the West.

The Social Character of Movements

Over the period 1975–2000, sociologists have begun to define 'contentious politics' – confrontations of ordinary people with authorities and/or elites – as the building blocks of social movements. Sidney Tarrow defines the conditions under which contentious politics is triggered:

> When changing political opportunities and constraints create incentives for social actors who lack resources on their own, they contend through known repertoires of contention and expand them by creating innovations at their margins. When backed by dense social networks and galvanized by culturally resonant, action-oriented symbols, contentious politics leads to sustained interaction with opponents. The result is the social movement.
>
> Tarrow, 1998: p. 2

The case of the British movement for the abolition of the slave trade and slavery illustrates this definition:

1. the movement was collective and interactive, between authorities and the populace; it was conducted early on through organizations established for the purpose and spread by movement professionals;
2. it was aimed at political authorities (in this case Parliament) that could take action to achieve the movement's goal;
3. each phase of activism on abolition began with an opening of political opportunity and ended when new political constraints closed off forward movement;
4. the network of Quaker meetings was the first activated in the cause, but other church groups and local branches of the antislavery societies joined the abolitionist network over the period;
5. interaction was contentious and drew on symbols that helped promote action; and
6. participation in the movement came to be perceived by activists as constitutive of their identities.

Woman's Suffrage in the United States

The formation of the United States woman's suffrage movement was closely connected to the social groups and organizations that promoted the abolition of slavery there. For example, American woman's suffrage activists Lucretia Mott and Elizabeth Cady Stanton first met each other at the London World Anti-Slavery Convention of 1840, where they discussed the possibility of forming an organization to promote women's rights. Mott, a New Englander then living in Philadelphia and the senior of the two, had founded the first Female Anti-Slavery Society in the United States; her home was a stop on the 'underground railroad' through which slaves escaped to Canada and freedom. The well-educated daughter of a judge, Stanton was born in western New York State, married an abolitionist leader, and eventually settled with her family in Seneca Falls, New York. Stanton and Mott renewed their acquaintance in 1848, when they and three like-minded women issued a call to a women's rights conference in the summer. To their surprise, some 300 women and men came, applauded Stanton's speech echoing the Declaration of Independence and passed a resolution on women's suffrage: "Resolved, that it is the duty of the women of this country to secure to themselves their sacred right to the elective franchise."

The movement developed slowly through annual conferences, spreading primarily to the Middle West and East coast. In 1851, Stanton met Susan B. Anthony, then a paid organizer in the temperance movement. Anthony brought her organizational skills to women's rights, beginning a drive in 1854 based on organizers in each county of the state of New York who recruited local women. These in turn collected signatures for a petition calling for women's control of their own earnings, mothers' guardianship of children in divorce, and women's suffrage. The first concrete accomplishment – a New York state law granting women property rights – was passed in 1860.

With the outbreak of the Civil War in April 1861, organizing efforts were temporarily suspended. Not for long, however, as Anthony and Stanton reactivated their abolitionist and women's rights networks in 1863, issuing a call to a mobilization meeting in New York. The outcome was the National Women's Loyal League that set to work gathering close to 400 000 signatures on a petition for the Thirteenth constitutional amendment that would free the slaves. Women also served on the Civil War Sanitary Commission, which raised funds and organized and administered nursing and support services for the military hospitals and convalescent centers.

The campaign for ratification of the Fourteenth and Fifteenth amendments (the first made male ex-slaves citizens and the second granted them the vote) split the former abolitionist movement colleagues between those who supported rights for the ex-slaves as a first priority, and those who supported universal suffrage. In 1869, two new organizations dedicated to women's suffrage emerged from the split: the National Woman Suffrage Association founded by Anthony and Stanton, which objected to the Fourteenth and Fifteenth Amendments because they specified male ex-slaves; and the American Woman Suffrage Association, founded by Lucy Stone, her husband Henry B. Blackwell, and Mary Livermore. These associations pursued the cause by different methods and separately for 20 years during which the westward movement was completed, the country industrialized, and new women's associations such as the Women's Christian Temperance Union (1874) and the National Association of Colored Women (1895) were founded. In 1890, the National and the American Associations merged to form the National American Woman Suffrage Association, with Elizabeth Cady Stanton as its first president; she was succeeded 2 years later by Susan B. Anthony who served to 1900.

Led after 1900 by Carrie Chapman Catt, a gifted organizer, the National American Association continued its strategy: seeking constitutional change state by state. Catt resigned in 1904 to play an active role in founding the International Woman Suffrage Alliance launched that year. There was little activity around the woman suffrage ('Susan B. Anthony') amendment to the United States Constitution for over 10 years. New tactics appeared in 1913, devised by the more militant new movement recruits Alice Paul and Lucy Burns who were appointed to the National American Suffrage Association's Congressional Committee. They promptly organized a demonstration of some 5000 women who marched in Washington in support of a constitutional amendment the day before Woodrow Wilson's inauguration for a second term. Despite the organization's permit for the parade, the police were unprepared for the crowd reaction and the women were mobbed and jeered. Later that year, Paul and Burns established the Congressional Union, and their organization eventually became independent of the National American, which continued to pursue its state-by-state strategy, with only slow progress.

In 1915, Carrie Chapman Catt agreed to lead the campaign for women's suffrage in New York State, and once again became president of the National American Suffrage Association with a newly energetic board committed to her plan for a renewed campaign for a national constitutional amendment. (The Congressional Union and its successor the Woman's Party did not join the effort. The Woman's Party nevertheless contributed to the process by courting arrest through disruptive and provocative action, which made the National American's actions look reasonable.) At its convention in late 1917, the latter warned that if Congress did not approve the suffrage amendment before the next election, it would adopt the British suffragette method: opposing the election of Congressional candidates who opposed the federal amendment, no matter what their party. The amendment passed in the House of Representatives by the required two-thirds majority in January 1918, the Senate in 1919, and ratification by the requisite proportion of states was completed in 1920.

Twentieth-Century Social Movements: Italian Fascism

In the early twentieth century, several movements demonstrated unequivocally the adaptability of the social movement to regressive ends. The Italian Fascist movement serves as an example.

After the end of World War One late in 1918, Italy was swept by waves of contentious politics: demobilized troops demanded jobs; peasants and agricultural laborers agitated for land; consumers called for price controls; and workers struck for higher wages. Among the groups engaging in collective action were the *fasci di combattimento* founded by Benito Mussolini (formerly a revolutionary socialist) who convened the first *fascio* in Milan on 21 March 1919. At a public rally 2 days later, the national program was only vaguely described (it was not fully formulated and publicized until months later). The program called for integration of all Italian 'national' territory into the nation, and pledged to oppose all neutralist politics no matter the party involved; other radical social and political demands included universal (male) suffrage, a constituent assembly, a republic with local autonomy, the 8-h day, worker participation in factory management, land to the peasants, elimination of the Senate and all noble titles, and an end to conscription, expropriative taxes, speculation, and monopoly.

The Fascists' first public action (15 April 1919) was a counterdemonstration together with students, World War veterans, and revolutionary syndicalists, which attacked an anarchist/socialist march protesting repression in Milan. It ended with the sacking and burning of the headquarters of the socialist newspaper *Avanti*! In the next 2 years, most Italian popular contention was dominated by socialists and organized social Catholics (represented by the newly founded Popular Party) who both sought to control and lead the strikes in northern industries and large-scale agriculture in the Po Valley, the land occupations, and eventually in fall, 1920, the factory occupations. These struggles all occurred in northern Italy,

where socialist and Popolari strength was concentrated. The Fascist counterattack from its urban bases (again in the north) began in the same period: physically assaulting unionized farm workers on strike, the headquarters of peasant leagues, cooperatives, and unions, burning left wing newspapers' offices and printing plants, and breaking up meetings of the socialist communal councils as well.

Socialist and Popular Party organizations crumbled as Fascist membership and attacks both increased. The Fascists had become a political force to be reckoned with and their success went hand in hand with the demobilization of the left. In August 1921, Mussolini signed the so-called Pacification Pact with the socialists and Popolari, promising to eschew violence, but his followers forced him to back away from the agreement and he publicly renounced it later in the year. He also announced the formation of a political party, which he described as a rival to the state, outside traditional politics. Violent clashes between socialists and Fascists continued into 1922, when the Fascist threat to seize the state forced the cancellation of a general strike planned by an ad hoc group of unions; by then the entire Po Valley, and much of North and Central Italy was in Fascist hands.

As the constitutional parties called for a new government, Mussolini threatened to march on Rome to seize power. The march never happened; instead, Fascist Black Shirts occupied provincial government and party headquarters, post offices, and police stations (again in the north, where socialist power had been concentrated) sometimes without resistance. Fascist threats and his own concerns led the king to refuse to sign the declaration of martial law that the prime minister had drawn up, and Mussolini's first government (based on a coalition of center and right parties) was formally constituted on 31 October 1922. Mussolini used nationalism wreathed in patriotism as a weapon against the left, and he continued to be a master of language and image once in power.

Although there had been social movements in the European colonies (and in China, as well) before the outbreak of World War I in 1914, there were many more non-Western movements in the twentieth century; they ranged in principal concerns from nationalism to widow burning and female circumcision (in India and Kenya, respectively). Nationalists sometimes rejected Western notions of modernity and connected customs that seemed unacceptable to Westerners to colonial cultural authenticity. Two examples are Mahatma Gandhi in his representation of himself with a spinning wheel and in his advocacy of *khadi* (handspun and woven cotton cloth, which he himself wore in his later years), and Kenyan nationalist Jomo Kenyatta who defended female circumcision on behalf of the Kikuyu Central Association, a preindependence nationalist organization. More recently, televised news reports and rapid communication by telephone, fax, and e-mail have made possible the almost instantaneous transmission of news not only in print, but of visual images as well. To what extent can we speak of transnational social movements?

Transnational Social Movements

The above discussion of antislavery and women's rights as social movements notes the international dimensions that were integral to social movements even in the late eighteenth century; similarities and connections of social movements across borders are not new. Sidney Tarrow offers a more restrictive definition of a transnational social movement:

> sustained contentious interactions with opponents—national or nonnational—by connected networks of challengers organized across national boundaries … What is important … is that the challengers themselves be both rooted in domestic social networks and connected to one another more than episodically through common ways of seeing the world, or through informal or organizational ties, and that their challenges be contentious in deed as well as in word.
>
> 1998: p. 184

Tarrow discusses three movements that fit his definition: the contemporary ecological and environmental movement, the European and American peace movement of the 1980s, and Islamic fundamentalism.

The ecological/environmental movement, which mounted coordinated worldwide Earth Day demonstrations in the late 1970s and early 1980s, provides the example discussed here. Some of its social movement organizations and their coordinated activities across borders do qualify as transnational organizations, as Tarrow demonstrates. Greenpeace, for example, has been a genuinely transnational social movement organization, with professional organizers and offices in many countries, and supporters who share a common world view and are willing to support its advocacy. Founded in Vancouver, Canada, according to its 1992–93 report it had over 5 million supporters (contributors), 43 offices, and over a thousand paid employees. It developed out of domestic movements that had similar goals, organized advocacy networks, and conducted contentious politics that cross national boundaries. Its members and sympathizers demonstrated against French nuclear testing in the Pacific and against abusive fishing for whales by Japanese and Soviet whalers.

Nevertheless, most international connections are less sustained and less professionalized than Greenpeace. Not all issues can attract support across the varied conditions of world polities, and most political, social, and economic international movements continue to work within states and national-level private institutions like large business organizations or labor unions rather than through transnational voluntary associations that represent individuals' world views directly.

Overview: Social Movements from the Eighteenth to the Twentieth Century

Sidney Tarrow's definition of the social movement (quoted in the opening sentence of this article) posits its applicability to the entire period since the first appearance of social movements in the late eighteenth century up to the present. His definition is based on the premise that, although the issues and tactics of social movements have changed over the two centuries in which they have proliferated, the defining form of social movements – their collective nature, common purposes, and shared base in social solidarities with roots in the social

structure in which they occur – have remained relatively unchanged. Not all social movement scholars agree with this premise; among European sociologists and political scientists in particular the concept of 'new social movements' has been developed since 1985.

Kriesi et al. (1995) argue that a new type of social movements, which address similar issues yet vary significantly in form according to the national political context, has emerged in Western Europe since 1968. Kriesi and his colleagues emphasize contemporary cross-national variation in domestic politics and corresponding differences in the timing and other characteristics of social movements. The examples that they discuss under the rubric of new social movements include "the ecology movement (with its antinuclear energy branch), the solidarity movement (with the Third World), the contemporary women's movement, the squatters" movement, as well as various other movements mobilizing for the rights of discriminated-against minorities' such as gays (1995, p. xii). The new type of social movements discussed by Kriesi et al. (1995) is classified according to the chief constituencies of the movements and the connections among them. Tarrow's definition of social movements leaves open the characteristics of their members and their goals and stresses the movements' long-standing form. Two alternatives present themselves: in terms of their political concerns and context then, recent social movements may herald changes in some aspects of movement politics and require redefinition of the concept. Alternatively, the broad definition with which this article begins argues for stability in the form of social movements even as the issues that bring them into the public arena change. Whether one of these alternatives will contribute to understand contemporary and future social movements better is an empirical question that can only be determined with a longer span of observations.

Bibliography

Anderson, B., 1991. Imagined Communities. Verso, London.

Chan, J., 2008. Another Japan Is Possible: New Social Movements and Global Citizenship Education. Stanford University Press, Stanford, CA.

Clawson, D., 2003. The Next Upsurge: Labor and the New Social Movements. ILR Press, Ithaca, NY.

Costain, A.R., McFarland, A.S. (Eds.), 1991. Social Movements and American Political Institutions. Rowman and Littlefield, Boulder, CO.

Eschle, C., 2001. Global Democracy, Social Movements, and Feminism. Westview Press, Boulder, CO.

Flesher, F.C., Cox, L., 2013. Understanding European Movements: New Social Movements, Global Justice Struggles, Anti-austerity Protest. Routledge, London.

Gibson, N.C., 2006. Challenging Hegemony: Social Movements and the Quest for a New Humanism in Post-apartheid South Africa. Africa World Press, Trenton, NJ.

Götting, A.C., 2010. New Social Movements. GRIN Verlag GmbH, München.

Grodsky, B.K., 2012. Social Movements and the New State: The Fate of Pro-democracy Organizations when Democracy Is Won. Stanford University Press, Stanford, CA.

Harter, J.-H., 2011. New Social Movements, Class and the Environment: A Case Study of Greenpeace Canada. Cambridge Scholars, Newcastle upon Tyne, UK.

Keck, M.E., Sikkink, K., 1998. Activists beyond Borders: Advocacy Networks in International Politics. Cornell University Press, Ithaca, NY.

Kelly, C.A., 2001. Tangled up in Red, White, and Blue: New Social Movements in America. Rowman & Littlefield, Lanham, MD.

Kriesi, H., Koopmans, R., Duyvendak, J.W., Giugni, M.G. (Eds.), 1995. New Social Movements in Western Europe: A Comparative Analysis. University of Minnesota Press, Minneapolis, MN.

Lee, S.H., 2007. Debating New Social Movements: Culture, Identity, and Social Fragmentation. University Press of America, Lanham, MD.

Marable, M., Mullings, L., 2009. New Social Movements in the African Diaspora: Challenging Global Apartheid. Palgrave Macmillan, New York.

McAdam, D., McCarthy, J.D., Zald, M.N. (Eds.), 1996. Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings. Cambridge University Press, Cambridge, UK.

Melucci, A., 1985. The symbolic challenge of contemporary movements. Social Research 52 (4), 789–816.

Offe, C., 1985. New social movements: challenging the boundaries of institutional politics. Social Research 52 (4), 817–867.

Petras, J.F., 2003. The New Development Politics: The Age of Empire Building and New Social Movements. Ashgate, Aldershot, Hants, England.

Tarrow, S., 1998. Power in Movement: Social Movements and Contentious Politics, second rev. ed. Cambridge University Press, Cambridge, UK.

Tilly, C., 1995. Popular Contention in Great Britain, 1758–1834. Harvard University Press, Cambridge, MA.

Todd, M.J., Taylor, G., 2004. Democracy and Participation: Popular Protest and New Social Movements. Merlin, London.

Touraine, A., 1985. An introduction to the study of social movements. Social Research 52 (4), 749–787.

Waterman, P., 2001. Globalization, Social Movements and the New Internationalisms. Continuum, London.

Waters, S., 2003. Social Movements in France: Towards a New Citizenship. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire.

Social Movements: A Social Psychological Perspective

John Drury, School of Psychology, University of Sussex, Falmer, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article analyses the treatment of social psychological issues in the study of social movements. It traces historically the development of theory in the study of social movements, from accounts of 'societal breakdown' and 'collective behavior,' through resource mobilization theory, and the 'new social movements' approach. More recent 'social constructionist' approaches to social movements will be outlined, which explicitly foreground the topic of identity and its creation. Finally, there is a discussion of the concept of 'interaction' in social movement theories, which in different forms has been understood as the process whereby new identities are created in social movement activity.

Introduction

Social movements have been defined as collective challenges by people who have a shared aim and solidarity, who act in a sustained way against opponents, particularly those in power (Tarrow, 1998: p. 9). While much research on the activity of social movements – i.e., collective action – takes place within social psychology, the study of social movements is dominated today by sociology and political science. Yet social psychology has always provided perspectives on the topic of social movements. And many sociological theories of social movements make psychological claims or use psychological concepts. Indeed, by their nature, social movements raise many social psychological questions. Thus, one point that many of those who study social movements today agree on is that, as well as potentially contributing to material social change, social movement activity can create *identities* for social movement actors. However, social movement theories have not always acknowledged this creativity. Indeed, for many years, social movements were conceptualized such that the question of identity creation could not arise.

This article will trace historically the development of theories of social movements, beginning with the earliest accounts of 'societal breakdown' and 'collective behavior,' and focusing on the social psychological question of identity construction. The article will show how later dominant perspectives – in particular resource mobilization (RM) theory and the 'new social movements' (NSMs) approach – in different ways each pointed to the necessity of a concept of identity and hence raised social psychological questions. In the second half of the article, more recent 'social constructionist' approaches to social movements will be outlined, which explicitly foreground the matter of identity and its construction.

A number of social movement theories, from different perspectives, highlight the role of *interaction* in the identity construction process. However, from a social psychological perspective, this concept is not always fully explicated. In particular, research on *social identity* suggests that the concept of 'interaction' can refer to a range of different types of process, and that these different processes might themselves 'interact' in important ways within social movement activity. The final sections of the article discuss concepts of interaction in social movement theories and social psychology.

The Occlusion of Rationality in Social Movement Theory: 'Societal Breakdown' and 'Collective Behavior'

Theories of social movements find their origins in the late nineteenth and early twentieth centuries in the work of such names as Tarde, Le Bon, Marx, Weber, and Durkheim. This was a period of revolutionary upheaval across Europe, and social and psychological theorists were often concerned with issues of immense social significance. Thus, Marx was interested in the practical question of whether particular working class movements were capable of overthrowing capitalism; Marx hoped they would. As a sociologist, Weber was interested in this kind of question too, but from the opposite side of the barricades, as it were. Weber tended to view crowds and masses with trepidation; they were seen as aspects of transition, perhaps necessary, but temporary, he hoped. Weber's (1978/1922: pp. 241–245) studies of charismatic leadership link it with irrationality and therefore point more to the dangers of collective behavior rather than to any positive potential; Weber preferred the 'mature' politics of the organized trade unions to the 'mass' politics of social movements (Eyerman and Jamison, 1991: p. 15).

At this time, theories of crowd behavior emerged that focused almost entirely on these supposedly inherent dangers of collective activity. These theories were developed to explain the phenomenon of the (proletarian) crowd in order to combat it and to justify effective methods of suppression (*see* Collective Behavior, Social Psychology of, this volume). However, of the many European bourgeois intellectuals proposing theories of the crowd during this period, only one has retained a place within modern social psychology – Gustave Le Bon. Le Bon (1947/1895) provided a popular synthesis of the ideas of many of the other crowd theorists of the time, according to which the 'primitive' psychology of the crowd to some extent generalized to all collective forms. This assumption influenced subsequent accounts of collective behavior for decades to come. As a consequence, early accounts of social movements tended to reproduce the notion that social movement activity was linked to irrationality.

In the twentieth century, the approaches of Herbert Blumer and Talcott Parsons informed a perspective on social movements that dominated the field at least until the late 1960s. Their approaches shared two assumptions: that individuals

behaved differently in groups and that individual deprivations, breakdowns of the social order, and homogenizing ideologies were important preconditions for the emergence of social movements (Mayer, 1991). The basic Parsonian model of social movement emergence was that of a breakdown in norms due to failures in the organs of the social system responsible for social control or social integration, which led to 'strains' and thence to collective behavior (e.g., Parsons, 1969: pp. 85, 110). In the Blumer account (e.g., 1951), which criticized yet was clearly influenced by that of Le Bon, collective behavior was said to follow a life cycle of stages, starting with structureless crowds and ending with the formation of institutionalized social movements. Movements were said to emerge through simple processes of rumor, contagion, and imitation. Blumer also pointed to some positive aspects of collective behavior, however, such as the potential social creativity that might accompany the breaking of the routines of normal, institutionalized behavior.

The most elaborated Parsonian structural–functional 'strain' theory of social movements is that of Smelser (1962). Smelser's 'master proposition' is that "[p]eople under strain mobilize to reconstitute the social order in the name of a generalized belief" (Smelser, 1962: p. 385). Smelser's model identifies various kinds of generalized belief (e.g., hysteria, wish fulfillment, hostility, etc.), and the conditions (such as panic) under which people develop and act on such beliefs. In Smelser's 'value-added' structural model, determinants of collective behavior at a higher level of the explanatory hierarchy (e.g., structural conduciveness and strain) act as preconditions for determinants at the next stage. Like all theories assuming that society is essentially a functional system, what was of interest to Smelser about social movements was first, the types of strains within the system that would give rise to such movements; second, the way such movements become co-opted back into the conventional order (this was Smelser's measure of their 'success'); and third, the way movements develop from an 'undirected' to a 'directed' (organized, coherent) stage (Banks, 1972).

Followers of Blumer developed symbolic interactionist accounts of collective behavior. Within this perspective, Turner and Killian's emergent norm theory (Turner and Killian, 1957) has been influential. It departs from the tenets of the Le Bonian tradition in two important respects. First, it claims that collective behavior is structured by norms – i.e., shared definitions of right and wrong that supply justification and coordinate action between people in particular situations. Second, it argues that the uniformity observed in crowd behavior does not emerge quickly through simple 'contagion' processes but rather develops only gradually. Collective behavior is said to emerge from a background within which norms derived from tradition and institutionalization are not available. Since the established social organization ceases to provide direction, crowd members seek new norms in order to define the situation and act within it. The norms that come to shape collective behavior are developed by crowd members themselves, through interaction and hence symbolic exchange (*see* Collective Behavior, Social Psychology of, this volume).

Despite differences in emphasis, the various types of models prevalent during this period all assumed that social movements did not have the capacity to influence societal development or policy outcomes (Mayer, 1991). In functionalist sociology, collectivities in conflict with 'dominant value systems' were categorized as 'deviant.' Being pathologized in this way, social movements by definition could not be *creative* of anything, whether of new improved forms of society or of themselves in meaningful ways. Hence an analytic division was established between social movements and social change (Banks, 1972). Although Turner and Killian (1957) recognized and attempted to theorize the ability of collectives to create norms, they were still committed to the assumption of the 'collective behavior' tradition that there are 'lower' (normless, irrational) forms of collective phenomena (such as riots, lynchings, panics, and crazes) that could develop into 'higher' ones, which have norms (such as social movements). Again, only collectivities that were seen as having become integrated as part of conventional modes of life – such as the modern trade union movement – were studied outside this broad 'pathology' paradigm.

The Turn to Rationalism in Social Movement Theory

In the 1960s, 'new' movements, such as the civil rights campaigns, anti-Vietnam protests, and the countercultural hippy movement all forced social movement theorists to face issues of power, conflict, and change. The student demonstrations, which wracked both the United States and France, were particularly intractable to the dominant theories and were the most shocking for sociologists (Eyerman and Jamison, 1991). In many cases, the social scientists sided with activists and debated issues of strategy and tactics within protest movements. Their own experience made it difficult for social scientists to maintain the irrationalist assumptions of the collective behavior approach (Zald, 1992). Therefore a new tendency developed within social movement theory. Instead of separating collective action from the political as a distinct and irrational 'other,' social movements became seen as *part* of politics; as with the rest of the political process, social movements were seen to be pursuing political goals with means appropriate to their situation. Moreover, the events of France in 1968 in particular led to a new widespread concern among sociologists with 'cultural crisis' (cf Touraine et al., 1983: p. 6). However, Europe and the United States produced differing theoretical versions of these new 'rationalist' viewpoints in their responses to the new circumstances.

The Rise and Crisis of Resource Mobilization Theory

In the revolt against functionalism, North American social movement theorists sought to go beyond a purely macro-level focus. These social movement theorists were interested in the question why some grievances and not others developed into social movements. The RM approach arose as an answer to this question. What was distinctive about the RM approach was its attempt to theorize social movements in terms of the utilitarian, instrumental tradition – the complete obverse of the irrationalist assumptions of the old 'collective behavior' approach.

Some had been influenced by Olsen's (1965) rational choice theory of collective action (Mueller, 1992; e.g., Oberschall, 1973: pp. 114–117). In the collective behavior approach social strains had borne the explanatory weight for the emergence of collective action, whereas Olsen referred to measures of costs and benefits among individual members of the collective. Paralleling this, RM theorists argued that the extent of material and organizational resources available to a group were a better predictor of social movement mobilization than structural strains and subjective grievances. For example, Native Americans, blacks, and women mobilized only in the 1960s and the 1970s, even though strain and discontent among them had been relatively constant for many years (Mayer, 1991).

The logic of RM meant that the questions shaping social movement research became: what are the resources available for a movement, how are they organized, how does the state facilitate or impede collective mobilization, and what are the results of collective action (Mueller, 1992)? But many other questions that might be asked of social movements were necessarily excluded, in particular questions of identity and of psychological processes within the collective constituting that identity. The assumption of instrumental rationality led RM theorists to concern themselves only with the way preferences could be satisfied; the question of how such preferences might arise was not examined. Therefore, the discovery empirically of preferences being created or changed through collective action itself was awkward for RM (Scott, 1990). Analogously, Piven and Cloward (1991) raise difficulties for RM by pointing to examples of organization resulting *from* rather than *causing* collective action; for instance, in the 1930s, organizing committees were created among craft workers only after, rather than before, a wave of wildcat strikes. Piven and Cloward's argument is that the logical consequence of the attempt to counter the irrationalist approach to social movements by emphasizing instead the rational use of resources is that where RM theorists see a group acting *without* 'sufficient organization' (and this tends to refer in particular to the more oppressed or poorer social groups), this is defined as irrational and apolitical; hence the approval of many RM theorists for social movement leaders' attempts to quell rioting (Piven and Cloward, 1991). The same dichotomy posited by irrationalist theories between rational organization and irrational primitive groups is therefore reproduced; only the boundary dividing the two types has been moved.

The links that the 'collective behavior' and functionalist traditions suggested existed between psychological processes and irrationality had prompted RM theorists to react by developing an account of social movements without any social psychology at all. But the necessity of a social psychology might be said to be built into rationalism's own key problematic, that of the 'free rider': as a group member, the individual will necessarily gain from group action without having personally to take part in it; why then should anyone take part? The absence of compulsions or individual incentives in many instances of collective mobilization meant that the logic of RM was vulnerable to rupture by alien concepts such as identity and, latterly, meaning construction (Mueller, 1992).

Even from the start of its dominance, therefore, RM was under attack from critics for its overrationalized conception of social movements (Morris and Mueller, 1992). Instead of conceptualizing social movements purely as a function of organizational resources, critics argued that such (social psychological) concepts as grievances, ideologies, and identities had to be taken into account. Later developments from within the RM tradition therefore departed from the original framework in at least two ways (Mueller, 1992). First there was the claim that the actor viewed as an isolated *homo economicus* (following Olsen's metatheory), rationally calculating her action on the basis of givens in the social environment, should be reconceived in terms of her embeddedness in group identities of class, gender, nationality, ethnicity, and so on. Second, it was claimed that social locations overlap, providing not only resources and opportunities but also cultural material that actors may draw upon (together) to construe a sense of grievance. In contrast to both strain and RM, therefore, the new theorists wanted to stress that both grievances and resources are mediated by meanings; the meaning of social reality taken for granted by both these frameworks needed to be understood instead as contestable.

The 'New Social Movements'

While North American sociology had reflected the dominance of empiricism, among Europeans there was a greater willingness to embrace philosophical issues such as the political implications of theory (Eyerman and Jamison, 1991). Hence at the same time that Blumer's microsociology and the anodyne 'consensus' model of Parsons flourished on one side of the Atlantic, on the other side, the class conflict accounts of Weber and Marx remained foundational. Therefore in Europe, the emergence of the movements of the 1960s prompted a fresh engagement with these fundamental frameworks: did the nature of the new developments refute or confirm the traditional models of conflict?

Herbert Marcuse argued that a fundamental shift had occurred in capitalism, according to which NSMs were now at the center of social change, in place of the industrial proletariat (e.g., Marcuse, 1964; see Scott, 1990). Marcuse himself linked collective action with psychological change, by claiming that social transcendence required a transformation of the inner psyche (Flesher Fominaya, 2014).

As Scott (1990) argues, although most of the NSM theorists rejected Marx, their own frameworks echo the Marxist-influenced Marcusian account in which a substitute is sought for the industrial working class as an antagonistic and transformational agent. There is no full agreement on the referent of the term 'new social movements,' although there seems little disagreement that what distinguishes all those so described (for example, the women's, peace, and environmental movements) is their difference from the 'old' industrialized working class movements. This difference is what is important in the NSM perspective.

An example of an NSM study is Touraine et al.'s (1983) analysis of the French antinuclear movement. He notes that the movement began as a defensive reaction to fear of accident and ecological disaster. The movement ultimately collapsed, and Touraine's study was an attempt to understand why it did not develop into a movement with a wider

constituency. Touraine locates the movement's failure in its inability to clarify its own position in relation to technology and to define a clear enemy; the movement was unable to go beyond making a general critique and did not specify other modes of development. In this account, then, the failure or success of a movement to re-create itself and society in its own image stems from the failures of its ideology. The implication is that, given correct self-understanding and clear articulation of grievances, the enactment of change would flow automatically.

In Habermas's (e.g., 1978/1969) account, social movements are indicators of potential crises in late capitalism (Scott, 1990). What Habermas sees as the evolutionary process of 'modernity' entails the extension of rationality into all spheres of life. In the political and cultural realms, this satisfies the modern demand that policies be rationally presented and justified, but it conflicts with the role of the sociocultural system as a supplier of motivations. The NSMs are understood as an attempt to resist the 'inner colonization' of technical rationality while at the same time demanding increased levels of rational justification. Although Habermas's account includes the social transformational and self-changing potential of social movements, its emphasis is more on the wider social conditions that give rise to such movements than the processes within movements through which identity might emerge and change.

'Social Construction'

Klandermans (1992) suggests that, though the dominant social movement theory frameworks of previous years were in a sense complementary (RM explaining the *how*, and NSM examining the *why* of collective mobilization), in fact they shared a common deficit: neither of them explained what makes people define their situation in such a way that participation in a social movement seems appropriate. The turn to broadly social psychological issues of identity and meaning in social movement theory occurred on both sides of the Atlantic. Thus the concern with 'social construction' is common to a range of theories rather than a single theoretical perspective. Some key examples will serve to illustrate common themes and issues.

Although concerned with structural affordances in a similar way to RM theory, McAdam's (1982) account of the role of political processes also argues that such conditions and political opportunities can only explain movement insurgency in combination with the mediating role of processes of meaning attribution. McAdam explains changes in attribution in directions favorable to insurgency ('cognitive liberation') in terms of the altered responses by the members of a political system to a particular challenger, which serve to communicate to insurgents that the political system is becoming increasingly vulnerable to challenges. Similar to Piven and Cloward's (1977) three dimensions of transformation, McAdam suggests three types of belief for such a change: (1) the system is seen to lose legitimacy, (2) the system seems more capable of being changed, and (3) there is the sense that the person can change things. However, these changes are not simply 'read off' from the situation between insurgents and the political system; rather, groups of people jointly create the meanings that they read into current and anticipated events. Hence strong social integration within groups is a precondition of 'cognitive liberation' (p. 50).

Another influential social constructionist approach is Eyerman and Jamison's (1991) 'cognitive praxis' account, which conceptualizes social movements as forms of activity by which individuals create new kinds of knowledge and identities, both for themselves and for the societies in which they are situated. In a study of the new environmental movements in three countries in Europe, Eyerman and Jamison point to three dimensions of cognitive praxis: cosmological (the basic assumptions or beliefs about ecology shared by movement participants), technological (the specific topics of environmental protest, as well as alternative technologies such as renewable energy), and organizational (e.g., antielitism). They claim that environmentalism can only take the form of a social movement to the extent that the three dimensions are actively combined in the creation of 'new conceptual space' (pp. 68, 77); the European environmental activists managed this, and gave the issue known as 'the environment' new political and economic significance. In explaining this knowledge-producing and meaning-constituting process, Eyerman and Jamison suggest that, though all participants in a movement can contribute, some do so more than others; 'movement intellectuals' are those within movements that help create both their own roles and the very movements they participate in (p. 98). These individuals are able to publicly articulate the movement's identity through speeches, books, articles, and media appearances. They are not completely free to choose their moments and their arguments, however, but must do so in the context of counterarguments and in relation to the power of the media to recognize them. Both established intellectuals and those 'home-grown' within the movements become transformed through their involvement in these debates and arguments (p. 119).

In contrast to the early focus on crowds, Eyerman and Jamison's (1991) emphasis on movement participants' ability to define themselves and the world around them intellectually appears to render actual face-to-face collective participation explanatorily redundant in processes of meaning construction. Hence demonstrations, direct actions, and other crowd events are barely mentioned in their account (effects rather than causes of change?). This is a consequence of their focus; although the knowledge-producing activities of the movements are said to shape their own collective identities, the emphasis is very much on societal values rather than the particular ideas of movement participants. Identities change as a function of – or at least through the same process as – societal values: through the interplay between intellectuals and the wider public.

Melucci's (1989, 1995) account is a more explicit example of the social constructionist approach in the NSM tradition, and is said by some to be the most systematic and influential theory of collective identity in social movement theory (Flesher Fominaya, 2010a). Melucci argues that collective action should be understood as a process through which actors produce meanings, communicate, and negotiate; social movements are therefore themselves social constructions. Collective action is made possible through creating a 'collective identity' – a shared definition of the social world, shared goals, and a shared conception of the possibilities and constraints of collective action; collective identity is thus "the process of 'constructing'

an action system" (1995: p. 44). Groups can be more or less successful in creating a collective identity; without a collective identity, a group cannot accomplish collective action. Melucci therefore recommends that sociologists attend to the ways that social movements succeed or fail in becoming a social actor (1995): how are ends and means interpreted by different parts of the movement and what kinds of conflicts, tensions, and negotiations can be observed during the process of construction and maintenance of the collective identity? Collective identity as a process should therefore be a topic of social movement research rather than an assumption.

Unpacking 'Interaction' in Social Movement Theories

A common theme in a number of contemporary social movement accounts of meaning and identity construction is that of 'interaction.' But what does this mean in social psychological terms? Klandermans (1997) points out that concepts like cognitive liberation (McAdam, 1982), frame alignment (Snow et al., 1986), collective identity formation (Melucci, 1989), and others have all been proposed to account for meaning construction in mobilization but that they refer to different levels of meaning construction without being clear about this. He suggests that there is a neglect of the detail of face-to-face activity in collective action as part of these meaning construction processes (Klandermans, 1992). He links this absence with what he says is the failure of models such as those of Melucci and McAdam to combine the collective and the 'individual' (intragroup) levels of action and change. Implicitly at least, they depict 'flat' collectives and identities; the movements they describe seem untroubled by variation, and participants are therefore understood to change together consensually (Klandermans, 1992). Klandermans argues that transformation and change in collective beliefs should be linked theoretically to the possibility of individuals within the collective arguing against the dominant interpretation within their own group. Beliefs are contested, refuted, reformulated, and defended within as well as between groups; indeed, Klandermans suggests that (intragroup) deviation is inevitable.

Klandermans's (1992) own account of 'multiorganizational fields' is proposed as a way of grasping the relation between inter- and intragroup constructive processes in social movements through the role of structural vehicles. A movement's multiorganizational field refers to the total possible number of groups and organizations with which the movement might establish specific linkages, whether supportive or oppositional; these might include such things as friends, schools, churches, leisure clubs, and unions, and might be social category based (e.g., 'race,' gender, or class). Multiorganizational fields are always changing: coalitions collapse, controversies are resolved, organizations that were once competitors decide to join forces, new groups appear on the scene, and so on. It is through a community's multiorganizational field that grievances are interpreted, means and opportunities defined, opponents identified, strategies developed, and outcomes evaluated. Klandermans suggests that, in collective situations, the discussions that take place through social links enable people to clarify their relations with each other and with those they may be in conflict with: hence the 'us–them' polarization that occurs between groups in conflict (p. 98). Both within the group and outside it, different arguments vie for supremacy: "in the clashes and confrontations between competing or opposing schemes, meaning is constructed" (p. 100).

Klandermans (1992) sees studies by Hirsch (1990) and Fantasia (1988) as exceptions to the tendency to ignore face-to-face interaction in (sociological) social movements research, and their examples are used by him to illustrate the multiorganizational fields model. Both studies found increased solidarity during collective events (a student demonstration and wildcat strike, respectively) and a dramatic change in the worldview of many of the participants. In each case, participants were radicalized by the experience of collective resistance: they were more critical of the authorities, more willing to engage in the action as time went on, and more confident in the efficacy of such activities. Based on these kinds of findings, in Klandermans's account, 'discussion and argument' (e.g., p. 86) and 'interpersonal interaction' (e.g., p. 99) are assigned a central role, both within the collective itself and between participants and spectators. However, viewing this account from the perspective of mainstream theories of group processes and intergroup relations in social psychology, it is not clear that referring to these concepts resolves all the problems raised with other accounts of meaning construction. Thus, as a merely descriptive term, 'interaction' might refer to people acting both as relatively unique individuals or in their capacity as members of groups. As a social psychological explanatory concept, however, 'interpersonal' interaction is distinguishable in important ways from 'intergroup' interaction (Brown and Turner, 1981), since in the first people act on the basis of personal identities and in the second people act on the basis of social identities (Turner et al., 1987). It is an empirical question whether or not those 'interacting' in particular social movement contexts define themselves as unique individuals or group members, and what the psychological effect is of each type of interaction. To the extent that 'interpersonal interaction' is left unexplicated as a social psychological process, it still says little about what it is about movement participation that leads to identity change. While some more recent sociological social movement studies of the role of interaction have acknowledged the distinction between (inter) group versus personal levels of psychological functioning (Flesher Fominaya, 2010b), this has not been a widespread development in social movement theory.

The nature of 'discussion and argument' is similarly undertheorized. In his account of 'persuasive communications' in social movements, Klandermans (1992) observes that people are more likely to be influenced by members of their own group than by those they are opposed to. But what is the theoretical grounding? *Why* should some (group members') arguments and not others prevail, or even be listened to? Without specifying how and why particular social channels or conflicts enable or disable particular arguments among particular groups, an appeal to 'discussion and argument' does not go far beyond the accounts of Melucci (1989) and McAdam (1982), which, as Klandermans says, seem to refer merely to 'ideas' rather than social movement action.

The role of different types of interactions in transforming collective identities was examined in an ethnographic study of social movement development at an antiroads campaign (Drury et al., 2003). Here, it was only under certain conditions that 'discussion and argument' served to redefine collective identity boundaries such that participants in a 'local' protest came to see themselves as part of a 'global' social movement. Specifically, analysis of participants' accounts before and after involvement in the eviction of a tree suggested the role of conflict with the police in producing an oppositional definition of the collective identity, facilitating links to other groups in resistance to illegitimate authority. The analysis thus showed how intragroup interaction (among different protesters) could be put together with intergroup interaction (between protesters and police) to make sense of this process of identity creation. Particular 'within-group' arguments and speakers only became influential when relations *between* groups had changed. The aim to stop the road building from 'destroying Wanstead' was common to 'activists' and many 'locals' from the beginning of the campaign. But, to the extent that the 'activists' were defined as different from self, their arguments for the national and even global significance of the campaign were less easily accepted. The perceived violent eviction by police served to redefine *all* those involved – 'locals' and 'activists' alike – as sharing a common identity in relation to the police. Within this changed context, according to which the road building was clearly 'political,' the 'political' links made by protesters (to the national roads program and to state- and business-sponsored environmental destruction more widely) came to make more sense and the parochial construction of the issue was transcended.

A final point on 'interaction' is that in a number of social movement accounts it is said to transform collective beliefs through "necessarily a *gradual* process" Klandermans (1992: p. 85). For example, the collective identity necessary for people to be able to act together is said to be "constructed and negotiated through a repeated activation of the relationships that link individuals (or groups)" (Melucci, 1995: p. 44). Gradual or sustained psychological change over time is certainly important in the life cycle of social movements (Taylor, 1989). But, as with the extended processes of rumor and milling that Turner and Killian (1957) suggest account for the emergence of norms out of extraordinary situations' 'normlessness,' this way of understanding psychological change cannot account for the sudden transformations observed in many crowd events (Reicher, 1984).

Conclusions

A trend toward focusing on identity construction is a key example of the return of the social psychological in social movement theory, but it is not the only one. Another important development is increased interest in the role of emotion following years of neglect in response to the 'collective behavior' tradition (e.g., Goodwin et al., 2000). The focus on emotions is sometimes connected to identities and interaction (e.g., Jasper, 2010). This linkage parallels developments in social psychology, where intergroup accounts of emotion have developed within the social identity approach to explain intergroup behaviors (Mackie and Smith, 1998).

A second social psychological trend in social movement theory has been the development of 'impact' accounts, reflecting the empirical question of the durability of psychological and other biographical changes among social movement participants (e.g., Kiecolt, 2000). This has been studied in terms of participants' life history prior to involvement (e.g., Della Porta, 1992) and measured in terms of their subsequent opinions and activities (e.g., Fendrich and Lovoy, 1988). A common finding in these latter surveys is that, even 10 or 20 years after participation in social movements, former activists still differ from nonactivists on a variety of measures, including continuing to espouse left-wing views, remaining active in contemporary movements, and being involved in liberal professions, all impacts attributed to their participation in social movements.

Together, these developments and those outlined in the rest of this article show the continued relevance of social psychological issues in the study of social movements. This is the case whether the social psychological perspective has been applied directly (from social psychological theories) or indirectly (from other social science disciplines addressing topics that are broadly social psychological) and whether the processes of interest (identity construction, interaction, or emotion) are understood as part of the preconditions for social movement activity, a correlate of that activity, or its result.

See also: Collective Action; Collective Behavior, Social Psychology of; Deindividuation, Psychology of; Identity and Identification, Social Psychology of; Intergroup Relations; Political Psychology; Self-Categorization Theory; Social Identity in Social Psychology; Social Psychology.

Bibliography

Banks, J.A., 1972. The Sociology of Social Movements. Macmillan, London.

Blumer, H., 1951. Social movements. In: Lee, A.M. (Ed.), New Outline of the Principles of Sociology. Barnes and Noble, New York, pp. 199–220.

Brown, R.J., Turner, J.C., 1981. Interpersonal and intergroup behaviour. In: Turner, J.C., Giles, H. (Eds.), Intergroup Behaviour. Blackwell, Oxford, UK, pp. 33–65.

Della Porta, D., 1992. Life histories in the analysis of social movement activists. In: Diani, M., Eyerman, R. (Eds.), Studying Collective Action. Sage, London, pp. 168–193.

Drury, J., Reicher, S., Stott, C., 2003. Transforming the boundaries of collective identity: from the 'local' anti-road campaign to 'global' resistance? Social Movement Studies 2, 191–212.

Eyerman, R., Jamison, A., 1991. Social Movements: A Cognitive Approach. Polity, Cambridge.

Fantasia, R., 1988. Cultures of Solidarity: Consciousness, Action, and Contemporary American Workers. University of California Press, Berkeley.

Fendrich, J.M., Lovoy, K.L., 1988. Back to the future: adult political behavior of former student activists. American Sociological Review 53, 780–784.

Flesher Fominaya, C., 2010a. Collective identity in social movements: central concepts and debates. Sociology Compass 4, 393–404. http://dx.doi.org/10.1111/j.1751-9020.2010.00287.x.

Flesher Fominaya, C., 2010b. Creating cohesion from diversity: the challenge of collective identity formation in the global justice movement. Sociological Inquiry 80, 377–404. http://dx.doi.org/10.1111/j.1475-682X.2010.00339.x.

Flesher Fominaya, C., 2014. Social Movements and Globalization: How Protests, Occupations and Uprisings Are Changing the World. Palgrave Macmillan, Basingstoke, UK.

Gamson, W.A., 1992. The social psychology of collective action. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. 53–76.
Goodwin, J., Jasper, J.M., Polletta, F., 2000. The return of the repressed: the fall and rise of emotions in social movement theory. Mobilization: An International Journal 5, 65–83.
Habermas, J., 1978. Toward a Rational Society: Student Protest, Science and Politics. Beacon, Boston (Originally published 1969).
Hirsch, E.L., 1990. Sacrifice for the cause: group processes, recruitment, and commitment in a student social movement. American Sociological Review 55, 243–254.
Jasper, J.M., 2010. Social movement theory today: toward a theory of action? Sociology Compass 4, 965–976. http://dx.doi.org/10.1111/j.1751-9020.2010.00329.x.
Kiecolt, K.J., 2000. Self-change in social movements. In: Stryker, S., Owens, T.J., White, R.W. (Eds.), Self, Identity and Social Movements. University of Minnesota Press, Minneapolis, pp. 110–131.
Klandermans, B., 1992. The social construction of protest and multi-organizational fields. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. 77–103.
Klandermans, B., 1997. The Social Psychology of Protest. Blackwell, Oxford, UK.
Le Bon, G., 1947. The Crowd: A Study of the Popular Mind. Ernest Benn, London (Originally published 1895).
Mackie, D.M., Smith, E.R., 1998. Intergroup relations: insights from a theoretically integrative approach. Psychological Review 105, 499–529.
Marcuse, H., 1964. One-dimensional Man. Sphere, London.
Mayer, M., 1991. Social-movement research in the United States: a European perspective. International Journal of Politics, Culture, and Society 4, 459–480.
McAdam, D., 1982. Political Process and the Development of Black Insurgency 1930–1970. University of Chicago Press, Chicago.
Melucci, A., 1989. Nomads of the Present: Social Movements and Individual Needs in Contemporary Society. Hutchinson Radius, London.
Melucci, A., 1995. The process of collective identity. In: Johnston, H., Klandermans, B. (Eds.), Social Movements and Culture. UCL Press, London, pp. 41–63.
Merton, R.K., 1968. Social Theory and Social Structure. The Free Press, New York.
Morris, A.D., Mueller, C.M., 1992. Preface. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. ix–x.
Mueller, C.M., 1992. Building social movement theory. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. 3–25.
Oberschall, A., 1973. Social Conflict and Social Movements. Prentice Hall, Englewood Cliff, NJ.
Olsen, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, Mass.
Parsons, T., 1969. Politics and Social Structure. Free Press, New York.
Piven, F.F., Cloward, R.A., 1977. Poor People's Movements: Why They Succeed, How They Fail. Vintage, New York.
Piven, F.F., Cloward, R.A., 1991. Collective protest: a critique of resource-mobilization theory. International Journal of Politics, Culture, and Society 4, 435–458.
Reicher, S.D., 1984. The St Pauls' riot: an explanation of the limits of crowd action in terms of a social identity model. European Journal of Social Psychology 14, 1–21.
Scott, A., 1990. Ideology and the New Social Movements. Unwin Hyman, London.
Smelser, N.J., 1962. Theory of Collective Behaviour. Routledge & Kegan Paul, London.
Snow, D.A., Rochford Jr., E.B., Worden, S.K., Benford, R.D., 1986. Frame alignment processes, micromobilization, and movement participation. American Sociological Review 51, 464–481.
Touraine, A., Hegedus, Z., Dubet, F., Wieviorka, M., 1983. Anti-nuclear Protest: The Opposition to Nuclear Energy in France (P. Fawcett, Trans.). Cambridge University Press, Cambridge.
Tarrow, S., 1998. Power in Movement: Social Movements and Contentious Politics. Cambridge University Press, Cambridge, UK.
Taylor, V., 1989. Social movement continuity: the women's movement in abeyance. American Sociological Review 54, 761–775.
Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford, UK.
Turner, R.H., Killian, L.M., 1957. Collective Behavior. Prentice-Hall, Englewood Cliffs, NJ.
Weber, M., 1978. Economy and Society: An Outline of Interpretive Sociology, vol. 1. University of California Press, Berkeley (Originally published 1922).
Zald, M.N., 1992. Looking backward to look forward: reflections on the past and future of the resource mobilization research program. In: Morris, A.D., Mueller, C.M. (Eds.), Frontiers in Social Movement Theory. Yale University Press, New Haven, pp. 326–348.

Social Network Algorithms and Software

Ulrik Brandes, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Graph-theoretic concepts feature prominently in network analysis because they facilitate the exploration and assessment of structural properties in intuitive terms. However, they also pose algorithmic challenges not normally encountered in population statistics and attribute-based data analysis. We review efficient combinatorial algorithms for core problems of social network analysis and discuss important aspects of software for these tasks.

Introduction

Social network analysis is the application of network science on social networks, i.e., social phenomena are represented and studied by data on overlapping dyads as the units of observation (Brandes et al., 2013c). Consequently, graphs are a straightforward and convenient mathematical representation that will be the basis of this article.

The aspects by which to analyze graphs are very different from those of attribute tables and, thus, also require different approaches. It is inaccurate but instructive to think of this as a distinction between numerical or algebraic computations on the one side and combinatorial algorithms on the other side. Determining graph properties typically requires listing, counting, traversing, and constructing or decomposing structures instead of evaluating formulas or solving equations.

Combinatorial problems often have inefficient, brute-force solutions, and they can be a challenge to design, let alone implement, more efficient ones. Consequently, software for social network analysis may differ substantially in utility by the algorithms used and the types of analyses supported.

Here we review a number of core problems in the analysis of social networks and point out efficient algorithms to solve them. The focus is on problems that discriminate social network analysis from other forms of social data analysis and statistics. Hence, we do not cover problems such as dimensionality reduction, clustering, or simulation. They are treated extensively in other articles of this encyclopedia.

In a second, shorter part, we review a number of issues pertaining to software tools that facilitate the application of network-analytic techniques.

We start by making precise the input to subsequently discussed algorithms.

Graph Representations

Algorithms operate on precisely defined, symbolically represented input. We, therefore, need to define formally the objects that represent feasible input to the algorithms in subsequent sections.

A *graph* is a pair $G = (V, E)$, where V is a set of *vertices* (sometimes called nodes or points) and E is a set of *edges* (sometimes called links or lines). In a *directed* graph, $E \subseteq V \times V$ consists of ordered vertex pairs. In an *undirected* graph, the order does not matter, so that $E \subseteq \binom{V}{2}$ consists of two-element sets of vertices. If $\{v, w\} \in E$, vertices v and w are said to be *adjacent*. The set of all vertices $N(v) = \{w : \{v, w\} \in E\}$ adjacent to v is its *neighborhood*, and we define $N[v] = N(v) \cup \{v\}$. Moreover, let $N(S) = \cup_{v \in S} N(v)$ for any set $S \subseteq V$. Throughout this article, we denote with $n = |V|$ the number of vertices and with $m = |E|$ the number of edges.

There are more general classes of graphs, e.g., with multiple edges connecting the same pair of vertices (*multigraphs*) or more than two vertices in an edge (*hypergraphs*).

To concentrate on the main ideas and simplify the exposition, however, we will assume simple undirected graphs (i.e., undirected graphs without multiple edges or loops) throughout. While generalization to directed multigraphs is straightforward in most cases, valued graphs (i.e., graphs with vertex or edge labels) may require more sophisticated adaptation. In fact, most network-analytic concepts admit distinct generalizations to the valued case.

The two main criteria for the evaluation of algorithms are running time and memory requirements, and often there is a trade-off between the two. Both are expressed relative to the size of the problem instance, and, to eliminate the dependency on the actual implementation and hardware, both are specified in terms of their asymptotic growth rather than precise timings or operation counts.

Two standard data structures are the *adjacency matrix*, a two-dimensional array indexed by the vertices, and the *adjacency list*, an array containing a list of neighbors for each vertex. If a graph $G = (V, E)$ is represented in an adjacency matrix $A = (a_{vw})_{v,\, w \in V}$, then

$$a_{vw} = \begin{cases} 1 & \text{if } \{v, w\} \in E \\ 0 & \text{otherwise} \end{cases}$$

for all $v, w \in V$. Hence, we can test adjacency in constant time by inspecting the corresponding entry in the adjacency matrix. In an adjacency list representation, on the other hand, we have to scan the list of neighbors to test whether a vertex has a particular neighbor.

The *density* of a graph is the ratio of actual edges and potential edges. For an undirected graph, this is $\frac{2m}{n(n-1)}$, as there are $\binom{n}{2} = \frac{n(n-1)}{2}$ vertex pairs. Graphs representing social networks are often sparse, i.e., $m \in \mathcal{O}$ or $m \in \mathcal{O}(n \log n)$, so that the vast majority of entries in the adjacency matrix are zero and

the representation is not space efficient. Note that the percentage of entries representing edges tends to zero for growing n. Hence, at least in general, the adjacency list representation is preferred.

The following toy problem is used to illustrate the benefits of the sparse adjacency list representation and to introduce an important structural insight. A *two-mode network* represents pairwise relationships between elements of two disjoint groups such as the affiliation of persons with organizations. These are represented in *bipartite* graphs, where the vertex set is $V = U \uplus W$ and all the edges $E \subseteq \{\{u, w\} : u \in U, w \in W\}$ are required to be incident to vertices in both of the two disjoint sets.

A *one-mode projection* is the graph obtained by restriction to one of the two vertex sets and considering two vertices adjacent if and only if they have a common neighbor. In the above example, this could be a pair of persons affiliated with the same organization.

This operation is conveniently expressed in matrix notation. Using an adjacency matrix representation $A = (a_{uw})_{u \in U,\ w \in W}$, the two one-mode projections are obtained as AA^T and A^TA. Note that the product matrices actually contain the numbers of common neighbors and hence represent a valued graph; this can of course be binarized if desired. Straightforward matrix multiplication requires a cubic number of steps, but the results contain edges only where two nonzero entries are multiplied. With an adjacency list representation, these can be enumerated directly and exclusively.

An important observation for the evaluation of running times of algorithms that iterate over the neighbors of each vertex is that this is proportional to the number of edges and hence the number of nonzero entries in the adjacency matrix. Let $\deg(v)$ denote the *degree* of vertex v, i.e., the number of vertices adjacent to v.

Lemma 2.1 (Handshake Lemma) $\sum_{v \in V} \deg(v) = 2m$

We will use this relationship frequently yet tacitly to bound the running time of algorithms in simple terms. A straightforward example is determining all vertex degrees by enumerating the edges each vertex is incident to. This shows that, no matter whether vertex degrees are represented in the graph data structure, we can assume them to be given for; otherwise, we can always determine them in time linear in the size of the graph.

Algorithm 1: Sort vertices by degree (small to large)

```
input  : simple undirected graph G = (V, E)
output: array vert (vertices ordered by degree)
        vertex array pos (position in array)
        array bucket (Δ(G) + 1 buckets, initially 0)

// count vertices of same degree
for v ∈ V do bucket[deg(v)] ← bucket[deg(v)] + 1

// compute prefix sums (last index of each bucket)
for d = 1, ..., Δ(G) do bucket[d] ← bucket[d − 1] + bucket[d]

// place vertices in bucket intervals (back to front)
for v ∈ V do
    bucket[deg(v)] ← bucket[deg(v)] − 1
    pos[v] ← bucket[deg(v)]
    vert[pos[v]] ← v
```

Another observation that turns out to be very useful is that sorting the vertices of a graph by degree can be done faster than in the general case. While general comparison-based sorting algorithms require $\Omega(nlogn)$ time in the worst case, we can exploit the discrete and linearly bounded range of vertex degrees to sort in linear time. Let $\delta(G)$ and $\Delta(G)$ be the minimum and maximum degree of a graph G, then $0 \leq \delta(G) \leq \Delta(G) \leq n - 1$ and we can sort vertices by degree as shown in Algorithm 1. Note that the first position in an array is assumed to be 0, not 1.

Cliques and Cores

One of the earliest sociometric concepts is that of a *clique* (Luce and Perry, 1949). It corresponds to a subgraph of a simple, undirected graph in which every vertex is adjacent to every other vertex. A clique is the idealized version of a cohesive group.

Finding a clique with the maximum number of vertices in a graph is an $\mathcal{NP}$-hard problem (see glossary), but due to the absence of large cliques in empirical social networks even enumerating all inclusion-maximal cliques can be feasible. A clique is inclusion-maximal, if it is not contained in a larger one; hence, an inclusion-maximal clique cannot be extended even though, in general, it is not of maximum size. Before giving an enumeration algorithm, however, we consider one of the many relaxations that have been proposed because it can actually be determined efficiently, always.

An inclusion-maximal set of vertices is called a *k-core* (Seidman, 1983), if the minimum degree in its induced subgraph is at least k. It is easy to see that the k-core is uniquely determined and contained in the l-core for all $l < k$. Note, however, that k-cores need not be cohesive and, in fact, can even be disconnected. The maximum k for which a vertex v is contained in the k-core is called its *core number*, $core(v)$, and the maximum core number of any vertex is the core number of the graph, $core(G)$.

Algorithm 2: Core numbers

```
input  : simple undirected graph G = (V, E)
output: vertex array c (core numbers)

sort vertices by degree using Algorithm 1

for v ∈ V do c[v] ← deg(v)          // initial guess (upper bound)

for i = 0, ..., n − 1 do
    v ← vert[i]
    foreach w ∈ N(v) do
        if c[w] > c[v] then
            // get first entry in bucket
            u ← vert[bucket[c[w]]]
            if u ≠ w then
                // swap
                pos[u] ← pos[w];   pos[w] ← bucket[c[w]]
                vert[pos[u]] ← u;   vert[pos[w]] ← w
            bucket[c[w]] ← bucket[c[w]] + 1          // adjust boundary
            c[w] ← c[w] − 1                          // reduce estimate
```

Batagelj and Zaveršnik (2011) describe a simple algorithm to determine the core number of each vertex, which we adapt in Algorithm 2. After sorting vertices by degree, we can eliminate all vertices of minimum degree since their core number corresponds to that degree. Their elimination reduces the degrees of all neighbors and we can remove those as well if their

degrees fall below the previous minimum. Since each removal reduces only the degrees of neighbors and this reduction is by exactly one, the degree-based ordering can be maintained by exchanging such a vertex with the first vertex of the same degree; incrementing the last position of the preceding bucket, the vertex is effectively moved one bucket to the front. When there is no more vertex with a degree not larger than the previous minimum, we repeat the entire process with the new minimum degree as the core number for the next batch of vertices until the graph is empty.

Algorithm 2 requires linear time and space as well as scales to very large networks, independent of their structural characteristics. Note that when the algorithm terminates, the vertices are ordered by core number. Among all possible vertex orderings this is the one that minimizes the number of neighbors following a vertex in the ordering. This property is advantageous in several contexts, and we will make use of it in the algorithms for clique and triangle enumeration below.

Algorithm 3: Clique enumeration

```
input  : simple undirected graph G = (V, E)
output: all maximal cliques
begin
    // grow cliques from single vertices
    for v_i = v_1, ..., v_n ordered by core number do
        grow(N(v_i) ∩ {v_1, ..., v_{i-1}}, {v_i}, N(v_i) ∩ {v_{i+1}, ..., v_n})
end
grow(X, C, P) begin
    if P = ∅ = X then
        list C                                   // no more neighbors
    else
        choose any u ∈ P ∪ X                     // pivoting
        for v ∈ P \ N(u) do
            grow(X ∩ N(v), C ∪ {v}, P ∩ N(v))    // add v
            X ← X ∪ {v};   P ← P \ {v}
end
```

Let us now return to the problem of listing all maximal cliques. Algorithm 3 does so by recursively growing a clique C from each vertex v_i. A new clique member must be adjacent to all present members, and the same clique must not be listed twice. Therefore, a partition $N(C) = P \uplus X$ of the neighbors of C is maintained, where P contains vertices that are potential new members and X contains vertices that are excluded (such as those tried previously). Every time a vertex $v \in P$ is added to C, the candidates remaining in P are restricted to their intersection with $N(v)$. We have found a maximal clique C for the first time, if there are no more candidates ($P = \emptyset$, maximality) and no excluded neighbors ($X = \emptyset$, non-redundancy).

This general approach was proposed by Bron and Kerbosch (1973), but Algorithm 3 features two improvements that result in significant speedup. The first one is due to Tomita et al. (2006) and is called *pivoting*. It is based on the observation that, given some fixed $u \in N(C)$, no extension with a vertex from $N(C) \cap N(u)$ can yield a maximal clique without also including u or some other vertex from $N(C) \setminus N(u)$. This reduces the number of recursive calls. While it may seem that further improvements are obtained by selecting $u \in P \cup X$ such that $|N(u) \cap P|$ is maximum, our own experiments suggest that the required overhead is excessive because the following improvement already achieves some of the advantages implicitly.

This second improvement is due to Eppstein et al. (2010) and fixes the order in which vertices are used to seed the recursion. By starting from singleton cliques in order of nondecreasing core number, the cardinality of P, i.e., the number of potential extension, is never more than $core(G)$.

Triad Census and Clustering Coefficient

The *triad census* is a vector containing the number of occurrences of nonisomorphic subgraphs on exactly three vertices. For an undirected graph, these correspond to four graphs shown in **Figure 1**. Note that, by combination of different configurations in nonnull dyads, these are refined to 16 such subgraphs in the directed case. The triad census was proposed as a means to assess local structure (Holland and Leinhardt, 1970, 1976) and has been in wide use since.

Straightforward computation of the triad census requires classification of all $\binom{n}{3}$ combinations of three vertices and thus takes $\Theta(n^3)$ time. The fastest known approach, on the other hand, is based on an algorithm of Chiba and Nishizeki (1985), which lists all triangles (cliques of three vertices, T_3 in **Figure 1**) of an undirected graph G in time $\mathcal{O}(core(G)m)$. Since core number and average degree are generally low in social networks, this is near linear time.

The idea of Algorithm 4 is to orient the edges according to an ordering $v_1,\ldots,v_n$ of the vertices, so that every triangle becomes a transitive triad with a uniquely determined last vertex v_i. The sets of preceding and succeeding neighbors $v \in V$ are referred to by $N^-(v)$ and $N^+(v)$, respectively.

For each oriented edge (u, v_i), the other outgoing edges (u, w) of its source u are considered, and it is checked whether w is also a predecessor of v_i, so that u, w, v_i form a triangle. By marking and unmarking all predecessors u of v_i before and after, this test requires only constant time. Using the elimination ordering of Algorithm 2, each edge (u, v_i) is compared with at most $core(u)$ many others, giving the desired bound on the running time.

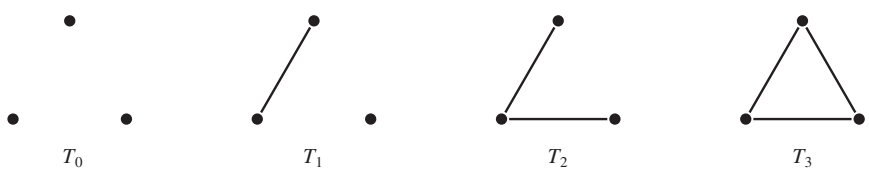

Figure 1 The four nonisomorphic graphs on three vertices.

Algorithm 4: Triad census

```
input  : simple undirected graph G = (V, E)
output: triad census T[0...3]
begin
    for i = 0, ..., 3 do T[i] ← 0
    for v_i = v_1, ..., v_n ordered by core number do
        T[2] ← T[2] + C(deg(v_i), 2)
        for u ∈ N^-(v_i) do mark u
        for u ∈ N^-(v_i) do
            for w ∈ N^+(u) do
                if w is marked then T[3] ← T[3] + 1        // triangle
        for u ∈ N^-(v_i) do unmark u
    T[2] ← T[2] − 3 · T[3]
    T[1] ← m(n − 2) − 2 · T[2] − 3 · T[3]
    T[0] ← C(n, 3) − T[1] − T[2] − T[3]
end
```

This triangle counting strategy is easily extended to yield the full census as follows. Every triad T_2 with exactly two edges has a unique vertex v_i of degree two. Pairing all $\deg(v_i)$ edges incident to v_i overshoots by counting each triangle three times (once each time v_i is contained in it). Since every edge forms a triad with each of the $n-2$ nonincident vertices, $m(n-2)$ counts every triad T_i once for each of the i edges it contains. Hence, the number of triads T_1 is $m(n-2)$ minus twice the number of triads T_2 and thrice the number of triads T_3. Clearly, the null-triads T_0 are all those that are not yet counted.

With a few additional case distinctions, Algorithm 4 applies to directed graphs as well. It can also be adapted straightforwardly to determine transitivity (triangle density), clustering coefficients (density of vertex neighborhoods), or constraint (Burt, 1992). Note that repeated counting of dyadic and triadic configurations is the computational bottleneck in probabilistic network models such as exponential random graph models (Lusher et al., 2013) or stochastic actor-oriented models (Snijders et al., 2010).

Role Equivalence

In contrast to Section Cliques and Cores, we now consider groups of actors defined by common structural roles rather than cohesion. See Borgatti and Everett (1992) and Lerner (2005) for background and an overview.

An equivalence relation $\sim \subseteq V \times V$ defined on the vertices of a simple undirected graph $G = (V, E)$ is called *structural equivalence* (Lorrain and White, 1971), if

$$v \sim w \Rightarrow N(v) = N(w)$$

for all $v, w \in V$. In other words, if two vertices are considered structurally equivalent, they have exactly the same neighbors. Any equivalence relation corresponds to a partition (into equivalence classes) and can also be represented as a *role assignment* r: $V \to \mathrm{IN}$, i.e., a function that assigns a role label to each vertex.

Algorithm 5 determines a role assignment based on structural equivalence that is maximum in the sense that every other partition into classes of structurally equivalent vertices is a refinement of this one. The algorithm is based on the straightforward observation that every vertex v is a witness for the nonequivalence of $w \in N(v)$ and $x \notin N(v)$. We therefore

Algorithm 5: Maximum structural equivalence

```
input  : simple undirected graph G = (V, E)
data   : integer array s (substitute role of neighbors)
         integer newrole (unused number)
output: vertex array r (role assignment)
newrole ← 1
for v ∈ V do r[v] ← newrole
s[newrole] ← 0
for v ∈ V do
    // assign neighbors to new role
    for w ∈ N(v) do
        if s[r[w]] = 0 then
            newrole ← newrole + 1
            s[r[w]] ← newrole;   s[newrole] ← r[w]
        r[w] ← s[r[w]]
    // clear substitutions
    for w ∈ N(v) do
        s[s[r[w]]] ← 0
        s[r[w]] ← 0
```

move neighbors w of v that had the same role i into the same newly created role $s[i]$, thus separating them from the non-neighbors. By linking the two role indices via s, we are able to reset this association in a second scan of the neighborhood of v. Note that after termination, the role indices that actually occur are in general not consecutive. By Lemma 2.1, running time is $\mathcal{O}(n+m)$.

The following is a concept of role that is based on relations to vertices that are equivalent rather than identical. An equivalence relation $\approx \subseteq V \times V$ defined on the vertices of a simple undirected graph $G = (V, E)$ is called *regular equivalence* (White and Reitz, 1983), if

$$v \approx w \Rightarrow N(v) \approx N(w)$$

for all $v, w \in V$, where two sets of vertices are equivalent if every member of one set has at least one equivalent member in the other set.

Every graph $G = (V, E)$ without isolated vertices has two trivial regular equivalences, $\approx\ = V \times V$ and $\approx\ = \{(v, v) : v \in V\}$, so that the maximum equivalence is not interesting. However, every partition of the vertices has a unique coarsest refinement that is regular, its *regular interior* (Borgatti and Everett, 1989).

Algorithm 6: Regular equivalence

```
input  : simple undirected graph G = (V, E), partition P of V
output: regular interior R of P
Q ← split(V, P)
R ← V
while R ≠ Q do
    find a block S ∈ R such that S ∉ Q
    find a block B ∈ Q such that B ⊂ S with |B| ≤ 1/2 |S|
    R ← (R \ S) ∪ {S \ B, B}
    Q ← split(S \ B, split(B, Q))
```

In other contexts, the regular interior is known as the *coarsest relational partition* and an efficient algorithm to determine it is given in Paige and Tarjan (1987). Similar to Algorithm 5,

Algorithm 6 is based on repeatedly *splitting* sets, which contain vertices with different roles in their neighborhoods. Given a partition $P = \{P_1,\ldots, P_k\}$ and a subset $S \subseteq V$ we define

$$spilt(S, P) = \{P_i \cap N(S), P_i \setminus N(S) : P_i \in P\}.$$

Note that $split(S, P) \preccurlyeq P$, i.e., splitting yields a refinement. It is important to note that the refinement obtained from a sequence of splits is independent of their order.

The sets P_i of a partition $P = \{P_1,\ldots, P_k\}$ are called *blocks*. A partition is *stable w.r.t. a set S*, if *split* $(S, P) = P$, and it is simply called *stable*, if it is stable w.r.t. each of its blocks. A stable partition corresponds to a regular equivalence and finding the regular interior thus corresponds to finding the coarsest stable partition.

A simple approach to find a stable refinement would be to choose any union of blocks as the next splitter until the partition is stable. Such an algorithm is correct because it maintains the invariant that the coarsest refinement of P is also a refinement of Q. The key to make the algorithm efficient is therefore the way in which blocks are chosen for splitting.

A second partition R is maintained that consists of unions of blocks of the current refinement Q. Starting from the entire set V, R is refined by splitting off blocks of Q, and the two subsets are tried as splitters of Q.

The algorithm maintains the invariant that Q is stable w.r.t. all blocks of R. This is asserted before the while loop, and whenever R is refined the two resulting blocks are used as splitters. Since $Q = R$ at termination, this implies that they are stable and that the algorithm is a specialization of the simple approach described above.

For efficiency, we maintain a list of the blocks in R that are unions of more than one block in Q and therefore qualify for the choice of S. Each of these *compound* blocks maintains a list of its blocks in Q. For every block $B \in Q$, a list of its member vertices and cardinality is maintained.

Observe that $split(S\backslash B, split(B, Q))$ results only in a three-way split of each block Q_i of Q. Since Q_i is stable w.r.t. S, the new blocks are $Q_i\backslash N(B)$, $(Q_i \cap N(B))\backslash N(S\backslash B)$, and $Q_i \cap N(B) \cap N(S\backslash B)$. It can thus be implemented by taking all neighbors of vertices in B out of Qi and classifying them according to whether all their neighbors in S are actually in B (this is what the cardinality counter is for). The entire split operation therefore takes only $\mathcal{O}(|B| + \sum_{v\in B}\deg(v))$ time.

Since B is always chosen as a smaller part of S, every vertex can be a member of a splitter B only $\mathcal{O}(\log n)$ times. The overall running time is therefore in $\mathcal{O}(m \log n)$, and the space requirement is linear.

Shortest-Path Centralities

Actor centrality indices in social networks have been classified along two dimensions (Borgatti, 2005): a diffusion mechanism (shortest paths, random walks, etc.) by which something (information, disease, etc.) spreads in a social network, and the nature of positions along the resulting pathways (beginning, center, etc.).

A *walk* in an undirected graph $G = (V, E)$ is an alternating sequence of vertices and edges such that edges appear only between their two end vertices. If no edge appears twice, the walk is called a *path*. A path starting from a source $s \in V$ and ending at a target $t \in V$ is called an (s, t)-path, and it is a *shortest* (s, t)-path, if it has the least number of edges among all (s, t)-paths. We use $d(s, t)$ to denote the length of a shortest (s, t)-path and refer to it as the *shortest-path distance* of s and t.

Two shortest-path centrality indices of common interest are

$$c_C(v) = \frac{1}{\sum_{t\in V} d(v, t)} \quad \text{closeness centrality}$$

$$c_B(v) = \sum_{s\neq v\neq t} \frac{\sigma(s, t|v)}{\sigma(s, t)} \quad \text{betweenness centrality}$$

where $\sigma(s, t)$ is the number of different shortest (s, t)-paths and $\sigma(s, t|v)$ is the number of shortest (s, t)-paths restricted to those passing through v.

Algorithm 7: Breadth-first search template

```
input  : simple undirected graph G = (V, E), source s ∈ V
data   : queue Q
output: vertex array dist (distance from root)
        vertex array pred (list of incoming edges on shortest paths)
for v ∈ V do dist[v] ← ∞
dist[s] ← 0
append Q ← s
⟶ init(s)
while Q not empty do
    extract v ← Q
    ⟶ settle(v)
    for e = {v, w} ∈ E do
        if dist[w] = ∞ then
            dist[w] ← dist[v] + 1
            append Q ← w
            pred[w] ← ∅
        if dist[w] = dist[v] + 1 then
            append pred[w] ← v
            ⟶ relax(v, e, w)
⟶ finish(s)
```

To compute these indices, we need to determine shortest paths and shortest-path distances first.

In simple undirected graphs, shortest-path distances are easily computed from a single source to all targets by breadth-first search (BFS). By itself, BFS is actually only a traversal strategy, and we have marked hooks in Algorithm 7 at which additional computations are typically performed. It should be clear how to determine closeness centrality of the source s by providing specializations of *settle* to add distances and of *finish* to invert the total.

Betweenness is a bit more challenging, but can still be obtained from the same template by introducing the derived relation of (*one-sided*) *dependency* $\delta(s|v) = \sum_{t\in V}\delta(s, t|v)$ with $\delta(s, t|v) = \dfrac{\sigma(s, t|v)}{\sigma(s, t)}$. Dependencies can be computed by an instantiation of BFS shown in Algorithm 8 and only need to be summed up to obtain betweenness centrality (Brandes, 2001).

In both cases, the running time is dominated by the n BFS traversals (one from each source) and thus in $\mathcal{O}(nm)$. Since the different BFS traversals do not depend on each other, overall space requirement is linear.

Algorithm 8: Source dependencies (specialized BFS)

```
data   : vertex array σ (number of shortest paths from source s)
         stack S
output: vertex array δ (dependencies of source s)
init(vertex s) begin
    for v ∈ V do
        σ[v] ← 0
        δ[v] ← 0
    σ[s] ← 1
end
settle(vertex v) begin
    S.push(v)
end
relax(vertex v, edge e, vertex w) begin
    if length(pred[w]) = 1 then          // one predecessor only?
        σ[w] ← σ[v]
    else
        σ[w] ← σ[w] + σ[v]
end
finish(vertex s) begin
    while S ≠ ∅ do
        w ← S.pop()
        for v ∈ pred[w] do
            δ[v] ← δ[v] + (σ[v]/σ[w]) · (1 + δ[w])
end
```

Software

While it is not unusual that analyses are scripted in general statistical software, few social network analysts will be implementing graph algorithms such as the above.

A variety of dedicated software tools is available in which the basic tasks of data management, analysis, visualization, and interactive exploration are addressed to varying degrees. Quite frequently, however, the algorithms implemented are not the most efficient.

In addition to fairly comprehensive tools with a graphical user interface, library implementations provide selected algorithms and functionality for adaption in one's own programing project. In some cases, useful functionality might also be available in tools that have been created for different purposes such as the analysis of molecular structures.

We are not going to discuss specific software in this article because a short discussion cannot do justice to the complexity of such tools, and is quickly outdated anyway. Instead, we review some general criteria on how to evaluate the utility of software.

Data

The generality of network representations and diversity of application scenarios have led to the coexistence of a huge variety of data formats, many of which are associated with one particular software. These formats differ not only in their syntax, but quite severely in their expressiveness. For software that does not support handling of certain network representations, such as longitudinal networks or collections of personal networks, it is likely that the associated format is not capable of representing them either. This is often a deliberate choice in order to ease the requirements for parsing the format.

In software for the analysis of social networks, data management is usually aligned with a default network representation. These are almost exclusively either matrices or graphs, so that most common formats for storing network data are comma-separated values tables for adjacency lists, adjacency matrices, and attribute tables, and an XML-based format for attributed graphs and hypergraphs such as GraphML (Brandes et al., 2013a). There are, however, numerous other, often proprietary, formats so that the transfer of data from one tool to another may require (possibly lossy) conversion.

As a general rule, it is good to have data available in tables representing adjacency matrices, edge lists, or attributes for archival purposes and data exchange. This includes computed indices, layout coordinates, and other information that may be of relevance for documentation and exchange. An expressive format that allows to represent various types of networks including longitudinal networks, hypergraphs, and network ensembles along with data attributes and graphical properties will ease data management and documentation substantially. Due to the lack of a common specification for graphical appearances, tools nevertheless have to support mapping of data attributes to graphical features.

An important consideration beyond supported data formats is data transformation capabilities. Most software tools will support, say, one-mode projections of two-mode networks. For exploratory analysis, however, even simple operations such as reorientation of edges, filtering, thresholding, hiding, copy & paste, or transformation of attribute values may be important. Although most of them are easily performed outside of a network-analytic software tool, this may require more technical expertise and is almost certainly less convenient.

Analysis

Quite a few software tools offer implementations of methods for the full range of analytic tasks except statistical modeling, for which dedicated software exists.

They differ substantially, though, in the variety of methods for the same task.

While most tools, e.g., allow for degree, closeness, and betweenness centrality computation, other common (e.g., eigenvector centrality, Katz's status, PageRank) or less common (e.g., β-centrality, random-walk betweenness, information centrality) indices may be available in selected stand-alone applications or libraries only. This applies even more to methods for cohesive groups and role equivalences.

In addition to comprehensiveness, efficiency may be an important criterion. As this article demonstrated, many network-analytic problems have simple, easy-to-implement solutions that are highly inefficient, whereas the design and implementation of efficient algorithms may require more effort and expertise. When the practical running times of a tool are too slow for repetitive or large-scale analyses, it is not necessarily the case that this applies to all other tools as well. A seemingly out-of-scope analysis may indeed be practical in another tool.

A similar concern is the level of generality at which methods are made available. Most procedures are parameterized, but the degree to which parameters can be controlled may vary. Similarly, generalizations of procedures beyond their original definition, say for connected binary (i.e., unweighted) networks, may or may not be available.

Finally, a rarely mentioned concern is accuracy. Combinations of data and algorithms can yield tricky cases or numerical computations that lead to flawed or imprecise results. Since there are no established procedures for assessing the accuracy of tools for network analysis, it is wise to cross-validate important and surprising results with multiple tools – although it should be clear that it is not necessarily the majority that is right on a specific task.

Visualization and Interaction

Social network analysis benefits tremendously from informative and attractive visual representations. This applies to both the discovery and communication of results.

Depending on the research question, available data, and usage scenario, different representations may be preferable (Brandes et al., 2013b). Software with strong analytic capacity, however, is most often limited to conventional representations with pointlike nodes connected by linelike edges. Even matrix representations are rarely supported, and only selected tools offer dedicated representations, for instance, for hierarchical clustering, two-mode networks, or event networks.

An underestimated aspect in network visualization is the effect of a network layout. Even though crisscrossing lines convey an air of detailedness and complexity, their effect on accurate and unbiased information display is mostly deteriorating. Due to the subtleties involved in the design and implementation of proper layout algorithms, however, many tool designers choose to revert to simple, established, and accepted algorithms, most notably the spring embedder of Fruchterman and Reingold (1991).

Social network studies differ in so many respects that it is implausible that some standardized representation and layout algorithm ignoring the analytic perspective are capable of providing the most appropriate visualizations. The survey of Brandes et al. (2013b) may serve as an inspiration regarding possibilities, including some concrete software. Researchers should not look for fancy graphics but analyze the suitability of visualization designs for their purposes.

Compared to other areas, the provision of interaction techniques for network data analysis and exploration is limited. While the possibility to change node positions and edge routing, label placement, or colors, sizes, and shapes is relatively common, little support is provided for view transformations such as focus and context or selective and linked views. Similarly, data transformation is initiated mostly via menu items or control panels.

Given the diversity of social network studies, it is unlikely that a single tool will suffice for the majority of purposes. As of now, the individual requirements on data management, analysis, and visualization may be served best by a specific combination of tools.

See also: Algorithm; Algorithmic Complexity; Clustering and Cohesion in Networks: Concepts and Measures; Network Centrality, Measures of; Structural Equivalence: Meaning and Measures.

Bibliography

Batagelj, V., Zaveršnik, M., 2011. Fast algorithms for determining (generalized) core groups in social networks. Advances in Data Analysis and Classification 5 (2), 129–145.

Borgatti, S.P., 2005. Centrality and network flow. Social Networks 27 (1), 55–71.

Borgatti, S.P., Everett, M.G., 1989. The class of all regular equivalences: algebraic structure and computation. Social Networks 11 (1), 65–88.

Borgatti, S.P., Everett, M.G., 1992. Notions of position in social network analysis. Sociological Methodology 22, 1–35.

Brandes, U., 2001. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25 (2), 163–177.

Brandes, U., Eiglsperger, M., Lerner, J., Pich, C., 2013a. GraphML. In: Tamassia, R. (Ed.), Handbook of Graph Drawing and Visualization. CRC Press, pp. 517–541.

Brandes, U., Freeman, L.C., Wagner, D., 2013b. Social networks. In: Tamassia, R. (Ed.), Handbook of Graph Drawing and Visualization. CRC Press, pp. 805–839.

Brandes, U., Robins, G., McCranie, A., Wasserman, S., 2013c. What is network science? Network Science 1 (1), 1–15.

Bron, C., Kerbosch, J.A.G.M., 1973. Algorithm 457: finding all cliques of an undirected graph. Journal of the ACM 16 (9), 575–577.

Burt, R.S., 1992. Structural Holes. Harvard University Press, Cambridge, MA.

Chiba, N., Nishizeki, T., 1985. Arboricity and subgraph listing algorithms. SIAM Journal on Computing 14 (1), 210–223.

Eppstein, D., Löffler, M., Strash, D., 2010. Listing all maximal cliques in sparse graphs in near-optimal time. In: Proceedings of the 21st International Symposium on Algorithms and Computation (ISAAC 2010). Lecture Notes in Computer Science, vol. 6506. Springer-Verlag, pp. 403–414.

Fruchterman, T.M.J., Reingold, E.M., 1991. Graph drawing by force-directed placement. Software – Practice and Experience 21 (11), 1129–1164.

Holland, P.W., Leinhardt, S., 1970. A method for detecting structure in sociometric data. American Journal of Sociology 76 (3), 492–513.

Holland, P.W., Leinhardt, S., 1976. Local structure in social networks. Sociological Methodology 7, 1–45.

Lerner, J., 2005. Role assignments. In: Brandes, U., Erlebach, T. (Eds.), Network Analysis: Methodological Foundations, Lecture Notes in Computer Science, vol. 3418. Springer-Verlag, pp. 216–252.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1 (1), 49–80.

Luce, R.D., Perry, A., 1949. A method of matrix analysis of group structure. Psychometrika 14, 95–116.

Lusher, D., Koskinen, J., Robins, G. (Eds.), 2013. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications. Cambridge University Press, New York.

Paige, R., Tarjan, R.E., 1987. Three partition refinement algorithms. SIAM Journal on Computing 16 (6), 973–989.

Seidman, S.B., 1983. Network structure and minimum degree. Social Networks 5, 269–287.

Snijders, T.A., van de Bunt, G.G., Steglich, C.E., 2010. Introduction to stochastic actor-based models for network dynamics. Social Networks 32 (1), 44–60.

Tomita, E., Tanaka, A., Takahashi, H., 2006. The worst-case time complexity for generating all maximal cliques and computational experiments. Theoretical Computer Science 363 (1), 28–42.

White, D.R., Reitz, K.P., 1983. Graph and semigroup homomorphisms on networks of relations. Social Networks 5, 193–234.

Social Network Analysis

Michael Steketee, Westat, Jenison, MI, USA
Atsushi Miyaoka and Maura Spiegelman, Westat, Rockville, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article we describe the analytic method of social network analysis (SNA). Specifically, we outline the foundations of SNA, key concepts and metrics, and current applications of SNA in areas of applied social and behavioral sciences. These areas include health studies, education research, organization research, team science, and information dissemination. As such, this article acts as an entry point into this method and a potential reference resource for individuals interested in exploring a deeper understanding of SNA.

Social network analysis (SNA) is used for measuring and analyzing the structural properties of networks of interdependent dyadic relationships. Such relationships can be interpersonal relationships like advice seeking, friendship, or trust that characterizes interactions between individuals; or interorganizational networks that characterize the relationships between organizations as a whole – including relationships, such as joint collaborations, resource exchange, information exchange, or even membership in common organizations (e.g., trade associations). One of the core assumptions of SNA is that the patterns of these relationships can have important effects on individual and organizational behavior, constraining or enabling access to resources, and exposure to information and behavior.

That assumption comes from the fact that society is composed of groups of individuals interacting with each other in social relationships. Through the accrual and persistence of such relationships social norms, culture, and institutions take shape. Critical to this process and, as such, to society, are the patterns and structures of social relationships, termed social networks, which provide constraints and opportunities for individuals to engage and influence their social world (Simmel, 1950[1908]; Frisby and Sayer, 1986). For example, the relationships among hospital staff (e.g., physicians and nurses) shape the provision and quality of care to patients. In schools, the relationships among principals, teachers, and parents are vital to the quality of education of students. And in scientific research, we increasingly see the importance of research teams working collaboratively in networks of social relationships for the production of rigorous research (Falk-Krzesinski et al., 2010).

SNA is used to study the structures of relationships such as those listed above by linking individuals (or other social units, such as organizations) and modeling interdependencies in behavior or attitudes related to configurations of social relations (O'Malley and Marsden, 2008). As such SNA can be a powerful technique for researching social phenomena such as the flow of information in a social network and the coordination, cooperation, or trust among groups of people. Just like any other social research procedure, we can use SNA to discover vital descriptive and inferential information about such social phenomena.

The import of SNA for applied social and behavioral sciences is also evidenced in the above examples. Education research, health research, and team science are three of many areas of study in applied social science. SNA allows researchers to understand the interrelationships that are crucial to such fields, for example, allowing us to identify the key actors related to professional development in a school system, to describe the amount of communication among medical professionals at a hospital, or better understand the collaborations created by a research initiative.

Social networks are at the core of our existence and due to the inherent interdependency of observations in social relationships; social networks require a specialized method of study. SNA is that method for describing, analyzing, and understanding social relationships. In this article we will review the foundations of the method, key concepts and metrics, and current directions in the use of SNA in areas of applied social and behavioral sciences.

Background

The concept of a social network can be traced back to the work of Simmel (1950[1908]) whose ideas on social distance, groups, dyads, and triads represent the foundation of SNA. These ideas were extended and operationalized by Moreno (1953[1934]) using graph theory and relational algebra to create the first social network graph and analysis. Advances in SNA have since produced a vast array of concepts, measures, and techniques. For a detailed look at such concepts and measures see Wasserman and Faust (1994), Scott (2000), Carrington et al. (2005), and Hanneman and Riddle (2005). In the following subsection, we briefly cover the basic concepts and measures of SNA needed to get newcomers started in the method.

Concepts and Measures

To begin, every social network includes *nodes*, *relationship*(*s*), and *ties*. (In the social network literature, nodes have also been termed actors and vertices, while ties have also been termed connections and edges. All are excepted terminology

referring to the same respective concepts.) A node represents the network members which can be individual persons or collective groups (e.g., boards of directors, social groups, clubs, or businesses). Relationships in a network are the basis on which nodes in a network are connected (e.g., friendship, collaboration, trading goods or services, or passing information). And a tie represents a connection between two nodes on the basis of a given relationship. Note that a relationship, and as such ties, can be *directed* (e.g., passing information) or *undirected* (e.g., joint collaboration). Also, there may be multiple types of relationships existing between the same set of nodes (e.g., there may exist, simultaneously, friendship, romantic, and collaboration networks among a set of coworkers at a business). The smallest social network possible is the *dyad*, a network containing two nodes with a single tie between them (see Figure 1). A dyad is not necessarily very interesting, but does represent the building block of larger networks. If a third node is added to the network and ties are made between each pair in the group, a triad is formed (see Figure 2). From there, larger networks of expanding size and structural complexity are formed (see Figure 3). Such larger network structures are typical of those studied in SNA and upon which SNA measures are utilized to describe and evaluate.

The unit of analysis in SNA is the tie and the data for a given social network are the ties that exist among a set of nodes. As such, social network data are interdependent – the connections for one node are directly related to the connections of other nodes in a social network. Two consequences of this fact are that (1) data must be entered into a *matrix* rather than a traditional format, and (2) traditional statistical analyses are not necessarily applicable to network data. To exemplify the first point, consider the collaboration network that has been created for the writing of this article – Mike, Atsushi, and Maura are the nodes in this collaboration and they are tied together based on their joint work. Table 1 demonstrates the data matrix for such a network and Figure 4 is the graphical representation of the network.

The example above is an undirected network. If we change the data in the matrix and thereby the network graph, we can see what a directed network looks like – in Table 2 and Figure 5 the relationship might be information sharing with ties representing incidents of sharing information.

Network data can be *egocentric* or *sociocentric*. Egocentric network data are collected from the perspective of a focal node or set of nodes. For example, survey respondents could be asked to list their friends. In this way a friendship network is constructed, however, the friends listed would not have an opportunity to confirm their tie to a respondent or list other ties they have. Sociocentric network data are a complete record of all ties among a bounded set of nodes. Clearly, more complete data are better, particularly for social networks. However, often complete network data are too difficult to obtain and we must rely on egocentric data. Another consideration is tie value, whereby ties in a network may be coded simply as binary (present or absent) or valued based on some scale. And lastly, there may be multiple *modes* in a social network, which occurs when there are multiple sets of nodes in the same network (e.g., persons and organizations).

Figure 1 Dyad social network. Note: graph created with NodeXL.

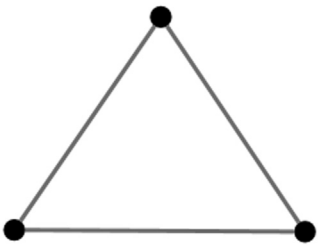

Figure 2 Triad social network. Note: graph created with NodeXL.

These pieces of information are the basic elements required for SNA. From here we can briefly discuss some social network measures which can be at the node level or the network level of analysis. Node-level measures are utilized to describe and examine the nodes of a network. Here characteristics of the nodes can be layered into the analysis as node-level measures. Another node-level measure, and perhaps the key one, is centrality. Centrality can be calculated a variety of ways and is used to identify key or central nodes in a network. *Degree centrality* is determined by the number of direct connections to a given node. *Betweenness centrality* finds that nodes fall between other nodes most often and is useful in finding places where a network would break off if a node with high betweenness was removed. And *closeness centrality* measures the reach of a given node or rather how close a given node is to others in a network. Closeness can be particularly meaningful in directed networks concerned with flows or disseminations of information, material, or disease. Lastly, with regard to node-level analysis, *distance* or *path* is a measure of the minimal number of ties between two nodes or the degrees of separation between two nodes. Distance is foundational to all node-level measures and can be used at the network level of analysis to describe the overall size of a network.

Network-level measures are utilized to describe and examine the overall structure of a social network and are often referred to as measures of network cohesion. *Density* is perhaps the most utilized network-level measure. Density is the proportion of the actual number of ties in a network to the number of possible ties in a network. Higher densities mean more interconnectedness among the nodes in a network. *Reachability* simply measures whether a node can reach all others in a network, directly or indirectly. *Distance*, as noted above, can be used at the network level such that the largest distance in a network represents its diameter and the average distance in a network gives us an overall sense of how separated most nodes are from each other. Similarly, average centrality can be calculated for a network to determine how centralized a network is, which determines if most ties belong to a small set of nodes or are divided more evenly across nodes. Furthermore, *subgroup* analyses can be conducted at the network level to determine if there are meaningful *clusters* of nodes separated throughout a network, which, like closeness, can be important regarding flows in directed networks. Lastly, note that calculations and interpretations of these SNA measures (e.g., centrality, distance, density, and subgroup clustering) will vary if ties are valued or directed.

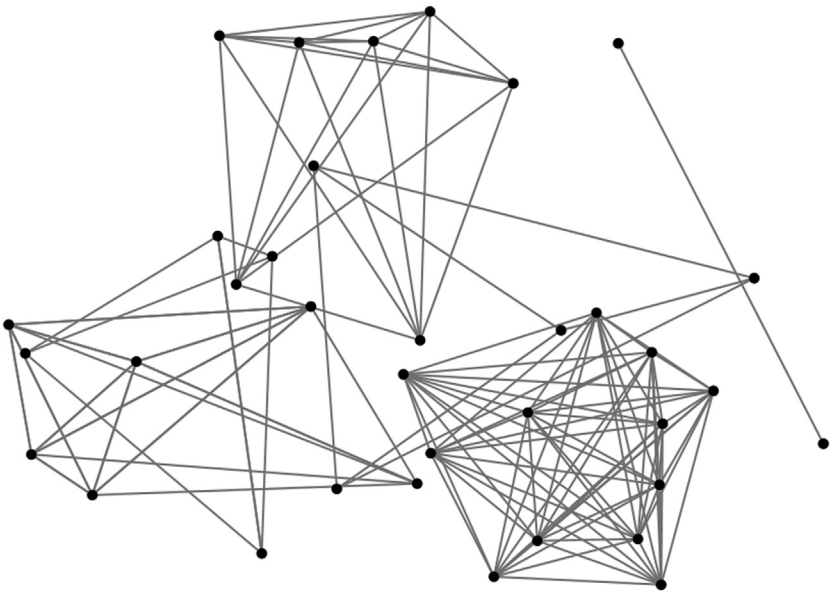

Figure 3 Example social network. Note: graph created with NodeXL.

In this section of the article we have briefly reviewed the basics of SNA. Next we turn to applications of SNA, dividing our review of applications by the types of analysis possible – node level, network level, and flow analysis.

Current Applications of SNA

In this section we describe some current applications of SNA to demonstrate the utility of SNA across areas of applied social and behavioral sciences. These areas include the study of team science, health research, education research, and organizational research. The section is broken out by types of network analysis to assist readers in understanding how SNA was utilized in each example. The focus in each subsection is on examples of applications rather than methodological details. Again, for more detailed information on the methods behind SNA measures and metrics, see Wasserman and Faust (1994), Scott (2000), Carrington et al. (2005), and Hanneman and Riddle (2005).

Node-Level Analysis

Node-level SNA focuses on the individual actors or nodes in a network and to an extent the small groups those nodes form in the larger network. Measures and metrics for node-level analysis center on the characteristics of nodes within a network, distance to determine the 'degrees of separation' between two nodes in a network, and centrality measures (e.g., degree, closeness, betweenness) to locate key nodes in a network. Node characteristics can be any categorization of the individuals within a network, such as persons' race, educational attainment; status as smoker or nonsmoker; role as a leader or follower; or condition as having a disorder or disease. Relational ties in a social network can be related to such node characteristics.

For example, Valente et al. (2005) examined the friendship patterns and smoking behavior of sixth and seventh graders across middle schools in southern California. They found popularity, as determined by network centrality, to be associated with increased susceptibility to smoke and smoking. Popular students were more likely to become smokers compared to less popular students. Such results point to the importance of understanding social networks and SNA for smoking prevention initiatives. Similarly, in examining the membership network of the National Network Consortium on Tobacco Control in Priority Populations, Luke et al. (2010) found that organizations' geographic location and status as a lead agency impacted the likelihood of connections among members. Specifically, individuals were more likely to connect if one was from a lead agency and more likely to connect if individuals were geographically close to one another.

Table 1 Social network data matrix

	Mike	*Atsushi*	*Maura*
Mike	0	1	1
Atsushi	1	0	1
Maura	1	1	0

Note: 1 = presence of tie; 0 = no tie.

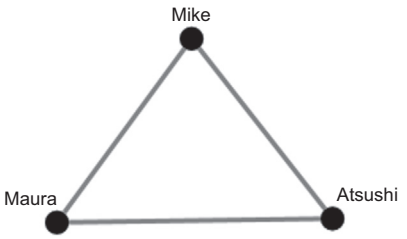

Figure 4 Collaboration network. Note: graph created with NodeXL.

Table 2 Social network data matrix for directed network

	Mike	*Atsushi*	*Maura*
Mike	0	1	0
Atsushi	0	0	1
Maura	1	0	0

Note: 1 = presence of tie; 0 = no tie.

Exploring organizational support for nutrition assistance programs at farmers markets, Dixit-Joshi et al. (2013) used organization type (e.g., federal government, state government, foundation, etc.) to describe and understand which organizations tend to support nutrition assistance programs offered by farmers markets across the United States. In another example, Liu et al. (2010) studied the relationship between social influence and rising autism diagnoses. Specifically, they found that children living very near to a child with autism are more likely to also be diagnosed with autism. They explain this result as the consequence of parents' interactions in social networks – one parent has a child diagnosed with autism, then shares their experience with others, who then are able to recognize similar autism markers in their own children. And Christakis and Fowler (2007) found that a person's chances of becoming obese were higher if they were friends with another person who became obese over a given period. Suggesting obesity, to an extent, spreads through social networks.

Centrality measures, such as degree centrality, betweenness centrality, and closeness centrality are utilized at the node level to identify important individuals in a network. Doing so can have important implications for future research, policy making, or planning. For example, evaluating the knowledge capacity of the Hydrogen Initiative Program at the Idaho National Laboratory, Birk (2005) utilized SNA to explore the network of advice and information seeking relationships in the program. She revealed that there existed a handful of individuals with high-degree centrality that the rest of the network relied on for information and advice. As such, Birk suggests that the loss of these individuals could create a void in the knowledge capacity of the program.

Additionally regarding centrality, Provan et al. (2010) examined the collaboration network among organizations in the Arizona Cancer Coalition. These organizations could be involved in any combination of three types of research networks (discovery, development, and/or delivery). The researchers found that high-degree centrality in the discovery and development networks were positively correlated to the importance of cancer research to the organization. Similarly, Leischow et al. (2012) found that among quitline organizations in the North American Quitline Consortium (NAQC), betweenness centrality was a critical measure of an organization's importance related to the sharing of best practices.

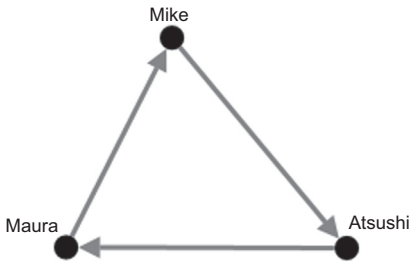

Figure 5 Directed collaboration network. Note: graph created with NodeXL.

Fredericks (2005) evaluated a developmentally disabled demonstration program designed by a state agency to optimize services and supports (e.g., early childhood development centers, day care, and universal pre-K) to consumers using existing sites. She found a small group of five organizations were highly central, with the highest degree and closeness centrality, in the information sharing network and represented a 'core' group integral to the function of the program.

In evaluating the implementation of a collaborative learning environment program for a Master in Public Administration (MPA) Program, Kapucu et al. (2010) used SNA to examine degree and closeness centrality measures for class friendships and shared class work among students pre- and postimplementation. They found that the program increased degree and closeness centrality across students in both the friendship network and class work network, suggesting the benefit of the collaborative learning program. And Hawe and Ghali (2008) studied the use of network analysis to model the social structure of high school staff and teachers in order to facilitate a school-wide health promotion intervention. They found betweenness centrality to be the most important measure for identifying individuals that can help expand the reach of their health promotion intervention. Specifically, individuals with high betweenness act as bridges or gatekeepers to the periphery of the network, while individuals with only high-degree centrality may not be able to reach others on the periphery of the network.

And Malinick et al. (2013) examined the relationship between social movements and media coverage using centrality in a two-mode network analysis. In this study, researchers employed a two-mode network strategy to evaluate the relationship between environmental activists' individual attributes and their centrality in the environmental movement's communication network. They found that being in a leadership position had very little to do with their centrality in the movement's communication network. However, they found clear evidence between centrality and media coverage. Based on these findings, they concluded that in the social movement, the strategically placed individuals, rather than the leadership, carry the responsibility of communicating with the media to spread and draw critically needed attention to their cause.

Regarding distance in SNA, this measure can be viewed as a node-level and a network-level measure. The distance between two nodes in SNA, as indicated earlier, is represented by the minimal number of connections that exist between them (i.e., six degrees of separation as popularized in media). At the network level, the maximum distance found in a social network represents the diameter of that network. At the node level, distance is a simple measure of how far apart the nodes are in a network. For example, Newman (2001) studied the scientific collaboration networks formed via coauthorships in scholarly publications. Newman posited that such networks are 'small worlds' (Milgram, 1967; Travers and Milgram, 1969) in which collaborators are part of tightly knit smaller groups and yet are

individually only a short distance from any other member of the network. He used average distance for the network to partially address this hypothesis, finding that scientific collaborators were indeed, on average, short distances from one another.

In this subsection we have reviewed recent applications of node-level measures in SNA, particularly node characteristics, centrality, and distance. Conveniently, the measure of distance brings us to the next subsection of this article – network-level measures in SNA.

Network-Level Analysis

In this subsection, we review applications of SNA measures at the network level of analysis. As mentioned previously, in using SNA we can apply a variety of measures to a network of interest. At the network-level of analysis, we attempt to uncover and describe the overall distribution of relations among nodes, instead of focusing on examining the attributes of individual nodes. For example, we might examine the density of the network, the size of the network, the number of subgroups in a network, or the average degree, betweenness, and closeness centrality of a network. Additionally, we can look at how the structure of a network changes over time or how the network structures of two groups or organizations of similar size compare to each other.

Examples of the use of density and average centrality in applications of SNA include the examination of friendships in classrooms (Kapucu et al., 2010) mentioned previously, the assessment of the role of networks in obesity prevention among day care centers (Marks et al., 2013), and the previously mentioned study of collaborative network involvement among 18 organizations within the Arizona Cancer Coalition (Provan et al., 2010). In examining the effect of collaborative learning in MPA program at the University of Central Florida, Kapucu et al. (2010) assessed the friendship network at the beginning and end of the school term. The average degree and closeness centrality measures were used to describe the significant increase in friendships and class work partnership, concluding that this increase would stimulate collaborative learning.

In assessing whether relational structures among child care workers play a role in obesity prevention, Marks et al. (2013) compared average centrality measures from network data that identified who provided dietary, physical activity, and policy information, who was involved in dietary decision-making with whom, and who consulted with whom on information for dietary and physical activity planning. The analysis showed staff interrelationships about physical activity, dietary, and policy information played key roles in each day care center. Additionally, they found that the network density varied for each day care center, suggesting that the SNA enabled the identification of unique center social dynamics. And as such, these differences could potentially be used to build future team capacity within the centers.

In studying collaboration in a cancer research network, Provan et al. (2010) used network data to describe the structure of network relations among cancer research organizations. They modeled network relations in three dimensions of research activities and measured the extent to which collaboration occurs within or across these three dimensions of research activity. They utilized overall network centralization and density to describe and compare the distribution of ties in the network and the overall connectedness of the network.

Another commonly used analysis at the network level is the identification and examination of network subgroups or substructures, sometimes referred to as clusters or clustering. Clusters refer to interconnected subgroups in a network and often can be displayed visually as regions in a network with relatively high local density and relatively few links to other clusters (Hoppe and Reinelt, 2010). Uncovering such subgroups in a network can be highly useful in applied fields. For example, Girvan and Newman (2002) used betweenness centrality to identify clusters of scientists in the larger collaboration network at an interdisciplinary research center at the Santa Fe Institute. These 'communities' centered on scientists' research topics or methodologies.

Hypothetically, in a leadership network, we could examine if the structures of network connectivity enable efficient sharing of information sufficiently and if the network effectively bridges clusters. Density and betweenness centrality are useful measures to detect and recognize the clusters. These measures offer important information about the connectivity among actors as well as clusters. For example, in the previously mentioned study of the NAQC, Leischow et al. (2012) discovered clear clusters representing the United States and Canada, and found a few quitline service providers had particularly high betweenness centrality and acted as leaders in sharing information. These findings suggested that although considerable amounts of communications and information sharing were occurring across the quitline service providers, the communication and information sharing patterns were not occurring equally – functioning as a type of coalition or community. They concluded that how the quitlines share information to improve practice will become increasingly important as quitline use expands.

Lastly, returning to Newman's (2001) study of scientific collaboration networks, he utilized the total number of collaborators and ties in the network, the maximum distance between collaborators of the network, and the average distance between collaborators to demonstrate the small world character of the network. These collaboration networks were huge in size, but there were relatively small average distances between collaborators. These patterns suggest that the network was characterized by significant clustering with a smaller number of bridges joining clusters of collaborators and conclude that such features are crucial for the development and functioning of a scientific community.

In this subsection, we discussed SNA measures at the network level. Specifically, SNA has been used in applied research to evaluate the networks of specific relational content. SNA is employed to study the networks of relationships among individuals, organizations, or social events and measures such as average centrality, density, and distance are used to describe and make inference about the larger network structure. In the next section, we discuss how SNA can be utilized to study flows of information, physical items, diseases, and ideas through networks.

Flow Analysis

In addition to describing static property, SNA has been used to model and explore how information, attitudes, or physical

items spread through network connections. All of these possibilities involve directional ties, as discussed in the background section of this article, but flow analysis additionally depends on the properties of actors, ties, and the ideas or items being transmitted. Considerations include (Borgatti, 2005):

- Can an actor be part of a path exactly once or more than once? A disease or infection that leaves the host immune cannot reach an actor more than once, while a physical item may travel to the same actor multiple times.
- Can a path be traveled exactly once or repeated? Some physical items, such as a loaned book, follow a path exactly once, while others, such as currency, can repeat a path.
- If a directional path connects actors A and B, can a directional path also connect actors B and A? A physical item can be returned, traveling from actor A to B and back to A, while a piece of information would not reverse its path.
- Do transmissions connect one actor to only one other actor, or can one actor connect to multiple actors? A physical item cannot travel from actor A to both actors B and C simultaneously, while a piece of information can duplicate itself and reach both B and C.
- If duplication occurs, are copies identical or merely similar to the original item? In a chain of e-mail forwards, simultaneous recipients read the same, replicated information. The spread of a virus reaches multiple recipients, but mutations and variations occur in the process.

Findings of network-level analysis may differ depending on the rules that govern flow. Simulations show that measures such as centrality, closeness, and betweenness may yield different results across different path restriction rules (Borgatti, 2005). As such the use of these measures should include consideration of whether closeness, for example, is truly capturing the distance or time by which goods or information travel from actor A to actor B.

In order to describe or analyze flow through a network, analyses also depend on the type of data collected. Researchers and evaluators may collect information directly from actors in a network by conducting a survey, or they may rely on secondary data or network information, if such data is tracked and accessible. Electronic communications provide an excellent opportunity to track the transfer of messages or information. Studies have used secondary data regarding online communities or message boards (e.g., Usenet, special interest groups), social networking Web sites (e.g., Twitter, Flickr), and e-mail messages (e.g., internal messages within a company or university). For example, Kumar et al. (2010) collected data from Usenet groups, focusing on message boards on a variety of subject matter areas, to track how conversations spread among their members. Members create original posts or reply to existing posts, creating branching conversation threads. Research focused on comparing conversation trees within networks (e.g., the size and depth of conversation trees), as well as how these features varied with the presence or absence of prolific, highly connected posters.

In order to explore how quickly and widely information spreads over a social networking Web site, as well as the role of friendship ties in this dissemination, Cha et al. (2009) used data from the Web site Flickr. This Web site allows users to upload photos, link to, or comment upon other users' photos, and create ties with other users. Rather than analyzing the entire network, the size of which could be classified as prohibitively large, analysts used a snowball sample in which they randomly selected users, and then branched out their initial sample by including other users to which their sample was connected, repeating this process over the data collection period. Similarly, Wang et al. (2011) used internal corporate e-mails over 2 years to explore the ways in which employees forwarded messages to their colleagues. Analysis of the flow of e-mails focused both on individual actors, for example, how quickly an actor spread information and to whom it was sent, as well as the entire network, for example, the structural properties of the flow network and the best models with which to explain the flow process.

When secondary data records are unavailable, either due to limited access to such data or its lack of existence, survey, and other self-reported information can help to describe the flow of information through a network. For example, in a study measuring communication within a lab group, Haythornthwaite and Wellman (1998), sought to explore how the relationships, both professional and personal, related to the frequency and modes of contact between its members. While some of this information might be available from secondary sources, for example, e-mail exchanges within a professional setting, researchers would be unlikely to have access to records detailing e-mail exchanges through external accounts, telephone exchanges, or in-person exchanges. In addition, capturing e-mail conversations only measures whether a message was sent from actor A to actor B, but the type of exchange, such as whether the message is a directive, a collaborative conversation, or a social discussion, is rarely available to researchers. Instead, network members were asked to report the number of times within a time period on which they received an assignment from another member, gave an assignment, and other types of transitive interactions based on their own judgments and recollections.

The considerations and studies laid out above have important implications for applied research on flow networks. For example, Burt (1987) studied the flow of medical innovation, specifically prescription of new drugs, through a network of physicians. He found that the dominate factor in the flow of medical innovation in this network was personal preferences among physicians, not their network position. While Valente and Davis (1999) used simulations of a communication network of physicians in an Illinois community to suggest that opinion leaders (identified via centrality) are crucial for the diffusion and flow of medical innovations in the network. For a wider review of other studies on flows and diffusion, see Strang and Soule's (1998) review focusing on the structural and cultural influences on diffusion processes.

Conclusion

In this article we gave a brief overview of the foundations of SNA and the basic concepts and measures used in applying the method. More advanced concepts and measures exist such as small worlds, preferential attachment, core-periphery structures, and estimated random graph models. Such advanced concepts and measures are deserving of attention,

but fall outside the scope and space of this article. We also covered a variety of examples of applications of SNA across applied social and behavioral science fields. These examples are certainly not exhaustive, but indicative of how SNA measures for node-level analysis, network-level analysis, and flow analysis can be used. In presenting this overview and these examples we have demonstrated both the conceptual and empirical import of SNA for addressing relational issues and phenomena in applied social and behavioral sciences.

The potential for applying SNA is vast, but not without challenges. The collection of social network data can be problematic, particularly when complete data on a network are needed to have a proper understanding of network structure. Once such challenges are overcome, SNA can be: employed to understand and progress communications in organizations, workgroups, schools, and health care systems; used to analyze and improve the dissemination of research findings and products by researchers and research initiatives; applied to investigate and find weaknesses in criminal and terrorist networks; utilized to evaluate the collaborations fostered by a new grant program or research initiative; and can be used to study the spread of communicable diseases. All of these examples are issues of a relational nature and as such require SNA to be addressed. The concepts, measures, and examples laid out in this article demonstrate how to begin to do so.

See also: Fertility and Social Networks; Network Analysis; Networks, Geography of; Social Network Algorithms and Software; Social Network Models: Statistical; Social Networks.

Bibliography

Birk, S., 2005. Application of network analysis in evaluating knowledge capacity. New Directions for Evaluation 107, 69–79.

Borgatti, S., 2005. Centrality and network flow. Social Networks 27, 55–71.

Burt, R., 1987. Social contagion and innovation: cohesion versus structural equivalence. American Journal of Sociology 92 (6), 1287–1335.

Carrington, P., Scott, J., Wasserman, S., 2005. Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge, MA.

Cha, M., Mislove, A., Gummadi, K., 2009. A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings from the 18th International World Wide Web Conference. ACM, Madrid, Spain, pp. 721–730.

Christakis, N., Fowler, J., 2007. The spread of obesity in a large social network over 32 Years. New England Journal of Medicine 357 (4), 370–379.

Dixit-Joshi, S., Burke, J., Das, B., Steketee, M., 2013. Nutrition Assistance in Farmers Markets: Understanding Current Operations. Report for U.S. Department of Agriculture, Food and Nutrition Service, Office of Research and Analysis. Project Officer, Williams, E.S., Alexandria, VA.

Falk-Krzesinski, H., Borner, K., Contractor, N., Cummings, J., Fiore, S., Hall, K., Keyton, J., Spring, B., Stokols, D., Trochim, W., Uzzi, B., 2010. Advancing the Science of Team Science. Clinical and Translational Science 3 (5), 263–266.

Fredericks, K., 2005. Network analysis of a demonstration program for the developmentally disabled. New Directions for Evaluation 107, 55–68.

Frisby, D., Sayer, D., 1986. Society. Ellis Horwood.

Girvan, M., Newman, M., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99 (12), 7821–7826.

Hanneman, R., Riddle, M., 2005. Introduction to Social Network Methods. University of California, Riverside, Riverside, CA. http://faculty.ucr.edu/~hanneman/.

Hawe, P., Ghali, L., 2008. Use of social network analysis to map the social relationships of staff and teachers at school. Health Education Research 23 (1), 62–69.

Haythornthwaite, C., Wellman, B., 1998. Work, friendship, and media use for information exchange in a networked organization. Journal of the American Society for Information Science 49, 1101–1114.

Hoppe, B., Reinelt, C., 2010. Social network analysis and the evaluation of leadership networks. The Leadership Quarterly 21, 600–619.

Kapucu, N., Yuldashev, F., Demiroz, F., Arslan, T., 2010. Social network analysis (SNA) applications in evaluating MPA classes. Journal of Public Affairs Education 16 (4), 541–563.

Kumar, R., Mahdian, M., McGlohon, M., 2010. Dynamics of conversations. In: Proceedings from 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, pp. 553–562.

Leischow, S., Provan, K., Beagles, J., Bonito, J., Ruppel, E., Moor, G., Saul, J., 2012. Mapping tobacco quitlines in North America: signaling pathways to improve treatment. American Journal of Public Health 102 (11), 2123–2128.

Liu, K., King, M., Bearman, P., 2010. Social influence and the autism epidemic. American Journal of Sociology 115 (5), 1387–1434.

Luke, D., Harris, J., Shelton, S., Allen, P., Carothers, B., Mueller, N., 2010. Systems analysis of collaboration in 5 national tobacco control networks. American Journal of Public Health 100 (7), 1290–1297.

Malinick, T., Tindall, D., Diani, M., 2013. Network centrality and social movement media coverage: a two-mod network analytic approach. Social Networks 35, 148–158.

Marks, J., Barnett, L., Foulkes, C., Hawe, P., Allender, S., 2013. Using social network analysis to identify key child care center staff for obesity prevention interventions. Journal of Obesity. http://dx.doi.org/10.1155/2013/919287.

Milgram, S., 1967. The small-world problem. Psychology Today 1, 61–67.

Moreno, J., 1953[1934]. Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama. Beacon House, Beacon, NY.

Newman, M., 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98 (2), 404–409.

O'Malley, A., Marsden, P., 2008. The analysis of social networks. Health Services and Outcomes Research Methods 8 (4), 222–269.

Provan, K., Leischow, S., Keagy, J., Nodora, J., 2010. Research collaboration in the discovery, development, and delivery networks of a statewide cancer coalition. Evaluation and Program Planning 33, 349–355.

Scott, J., 2000. Social Network Analysis: A Handbook. Sage Publications, London.

Simmel, G., 1950[1908]. In: Wolff, K. (Ed. and Trans.), The Sociology of Georg Simmel. Free Press, New York.

Strang, D., Soule, S., 1998. Diffusion in organizations and social movements: from hybrid corn to poison pills. Annual Review of Sociology 24, 265–290.

Travers, J., Milgram, S., 1969. An experimental study of the small world problem. Sociometry 32 (4), 425–443.

Valente, T., Davis, R., 1999. Accelerating the diffusion of innovations using opinion leaders. The Annals of the American Academy of Political and Social Science 566, 55–67.

Valente, T., Unger, J., Johnson, C.A., 2005. Do popular students smoke? The association between popularity and smoking among middle school students. Journal of Adolescent Health 37, 323–329.

Wang, D., Wen, Z., Tong, H., Lin, C., Song, C., Barabási, A., 2011. Information spreading in context. In: Proceedings from the 20th International World Wide Web Conference. ACM, Hyderabad, India, pp. 735–744.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, MA.

Relevant Websites

https://gephi.org/ – Gephi: The Open Graph Viz Platform.
http://www.insna.org/ – International Network for Social Network Analysis.
http://faculty.ucr.edu/~hanneman/nettext/ – Introduction to Social Network Methods.
http://nodexl.codeplex.com/ – NodeXL: Network Overview, Discovery and Exploration for Excel.
http://pajek.imfm.si/doku.php – Pajek: Social Network Analysis Software.
http://www.r-project.org/ – The R Project for Statistical Computing.
http://www.analytictech.com/ – UCINET: Social Network Analysis Software.

Social Network Models: Statistical

Philippa Pattison, University of Melbourne, Parkville, VIC, Australia
Stanley Wasserman, University of Illinois, Urbana, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 21, pp. 14375–14380, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Recent developments in statistical models for social networks reflect an increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. As a result, a growing importance has been accorded the problem of modeling the dynamic and complex interdependencies among network ties and the actions of the individuals whom they link. The early focus of statistical network modeling on the mathematical and statistical properties of Bernoulli and dyad-independent random graph distributions is reviewed. These early efforts have now been replaced by distributions that consist of theoretically and empirically plausible parametric models for structural network phenomena. These new models, discussed in detail here, allow complicated dependencies among the network ties, and have many possible special cases, reflecting the complexity of the social network paradigm.

Late twentieth-century developments in statistical models for social networks reflect an increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. As a result, a growing importance has been accorded the problem of modeling the dynamic and complex interdependencies among network ties and the actions of the individuals whom they link. The early focus of statistical network modeling on the mathematical and statistical properties of Bernoulli and dyad-independent random graph distributions has been replaced by efforts to construct theoretically and empirically plausible parametric models for structural network phenomena and their changes over time. These models in this article are based on graphs where the nodes are individual actors, which are 'linked' by lines representing relational ties (*see* Graphical Markov Models: Overview).

Statistical Models for Social Networks

Notation

In most of the literature on statistical models for networks, the set $N = \{1, 2,\ldots, n\}$ of network nodes is regarded as fixed and the edges, or ties, between nodes are assumed to be random. The tie linking node i to node j ($i, j \in N$) may be denoted by the random variable X_{ij}. In the simplest case of binary-valued random variables, X_{ij} takes the value 1 if the tie is present and 0 otherwise. Other cases of interest include:

1. valued networks where X_{ij} is assumed to take values in the set $\{0, 1,\ldots, C-1\}$;
2. multiple relational or multivariate networks, where the variable X_{ijk} represents the possible tie of type k from node i to node j (with $k \in R = \{1,2,\ldots, r\}$, a fixed set of types of tie); and
3. time-dependent networks, where X_{ijt} represents the tie from node i to node j at time t.

In each case, network ties may be *directed* (e.g., X_{ij} and X_{ji} are distinct random variables) or *nondirected* (e.g., X_{ij} and X_{ji} are not distinguished). The $n \times n$ array $\boldsymbol{X} = [X_{ij}]$ of random variables can be regarded as the adjacency matrix of a random (directed) graph on N; the state space of all possible realizations of these arrays may be denoted by Ω_n. The array $\boldsymbol{x} = [x_{ij}]$ denotes a realization of $\boldsymbol{X}$, with $x_{ij} = 1$ if there is an observed tie from node i to node j, and $x_{ij} = 0$ otherwise. In some cases, models may also refer to variables measured on actor attributes: for instance, the mth attribute $Z_1^{[m]}$ of actor i gives rise to a random vector $\boldsymbol{Z}^{[m]}$ with realization $\boldsymbol{z}^{[m]}$.

The adjacency matrix of an illustrative nondirected binary network is shown in **Table 1**. The data are from a study of interorganizational networks amongst mental health agencies conducted by Morrisey et al. (1994). The nodes of the network are 37 mental health agencies in one of the US cities in the Morrissey study; an edge links two agencies if they engage in regular and mutual exchange of information. The graphical representation of these data is shown in **Figure 1**. We use this network to illustrate several of the distributions and models described below.

Bernoulli Graphs

A simple statistical model for a (directed) graph assumes a Bernoulli distribution, in which each edge, or tie, is statistically independent of all others and governed by a theoretical probability P_{ij}. In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the probabilities to depend on structural parameters (Frank and Nowicki, 1993). These distributions have often been used as models (Bollobas, 1985), but are of questionable utility due to the independence assumption. Bernoulli graph distributions are described at length by Erdos and Renyi (1960) and Frank (1981).

Dyadic Structure in Networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. In one important series of

Table 1 Adjacency matrix for a nondirected binary network.

000000010110000001000000001000001110
000000000010000000000000001000000000
000000000000000000010000000000000000
000000000000000000000000000000000000
000000000000000010010000000010000010
000000000000000000000000000000000000
000000000000000000001011111000000000
100000000010000101000000000000000100
000000000000000000000000000000000000
100000000010100100000000000000000100
110000010100001000010000000110000000
000000000000100000000000000000000000
000000000101000000000100000000001000
000000000000000000000100100100100000
000000000010000000000011010010001000
000000010100000000000000000000000000
000010000000000000010000000000010000
100000010000000000010000000000000100
000000000000000000000000000000001000
001010000010000011000000101000000100
000000100000000000000010101000000000
000000000000111000000011111000000100
000000100000000100000110010110100011 00
000000100000000000000100101101000 1000
000000100000011000011111011100000 1100
000000100000000000000100101100000 0000
000000100000000000001111111010000 01100
110000100010011000000011111000000000
000000000010000000000000000000000000
000010000000000010000011000000000000
000000000000000000000000000000000000
000000000000011000000000000000000000
000000000000000000000000000000000000
100000000000100000100011101000000000
100000010100000001010110101000000000
100010000000000000000000000000000000
000000000000000000000000000000000000

developments, the p_1 model introduced by Holland and Leinhardt (1981), and developed by Fienberg and Wasserman (1981), recognized the theoretical and empirical importance of dyadic structure in social networks, that is, of the interdependence of the variables X_{ij} and X_{ji}. This class of Bernoulli dyad distributions and their generalization to valued, multivariate, and time-dependent forms, gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time. The model assumes that each dyad (X_{ij}, X_{ji}) is independent of every other and, in a commonly constrained form, specifies:

$$P(X=x)=\prod_{i<j}\exp\left\{\lambda_{ij}+\theta\left(\sum_{i\neq j}x_{ij}\right)+\rho\left(\sum_{i<j}x_{ij}x_{ji}\right)+\alpha_i\left(\sum_{j}x_{ij}\right)+\beta_j\left(\sum_{i}x_{ij}\right)\right\} \quad [1]$$

where θ is a density parameter; ρ is a reciprocity parameter; the parameters α_i and β reflect individual differences in expansiveness and popularity; and λ_{ij} ensures that probabilities for each dyad sum to 1. This model is log-linear (*see* Multivariate Analysis: Discrete Variables (Loglinear Models)).

An important elaboration of the p_1 model permits the dependence of density, popularity and expansiveness parameters on a partition of the nodes into subsets that can either be specified *a priori* (e.g., from hypotheses about social position) or can be identified *a posteriori* from observed patterns of node interrelationships (e.g., Wasserman and Faust, 1994). The resulting stochastic blockmodels represent hypotheses about the interdependence of social positions and the patterning of network ties (as originally elaborated in the notion of blockmodel by White et al., 1976). A potentially valuable variant of these models regards the allocation of nodes to blocks as a latent classification; Snijders and Nowicki (1997) have provided Bayesian estimation methods for the case of two latent classes. Another valuable extension of the p_1 model is the so-called p_2 model (Lazega and van Duijn, 1997). Like the p_1 model, the p_2 model assumes dyad independence, but it adds the possibility of arbitrary node and dyad covariates for the parameters of the p_1 model, and models the unexplained parts of the popularity and expansiveness parameters as random rather than fixed effects.

Null Models for Networks

Despite the useful parameterization of dyadic effects associated with the p_1 model and its generalizations, the assumption of dyadic independence has attracted sustained criticism. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. Following a general approach exemplified by early work of Rapoport (e.g., Rapoport, 1949), a number of conditional uniform random graph distributions were introduced as null models for exploring the structural features of social networks (see Wasserman and Faust, 1994 for a historical account). These distributions, denoted by $U|Q$, are defined over subsets Q of the state space Ω_n of directed graphs, and assign equal probability to each member of Q. The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric, and null dyads, as in the $U|MAN$ distribution of Holland and Leinhardt, 1975). When Q is equal to Ω_n the distribution is referred to as the uniform (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the members of Q and simulation of $U|Q$ is often straightforward (e.g., see Wasserman and Pattison, 2002), although certain cases, such as the distribution that is conditional on the indegree $(\sum_j x_{ji})$ and outdegreee $(\sum_j x_{ij})$, of each node i in the network, require more complicated approaches (Snijders, 1991; Wasserman, 1977).

A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features (e.g., Holland and Leinhardt, 1975). For example, the network displayed in **Table 1** has 84 edges, and 333 and 69 triads with exactly 2 and 3 edges, respectively. If Q denotes the subset of complete Ω_{37} with 84 edges, the expected number of triads of each kind is 250.1 and 50.8, and the observed values lie at the 0.003 and 0.005 percentiles of the $u|x_{++} = 84$ distribution,

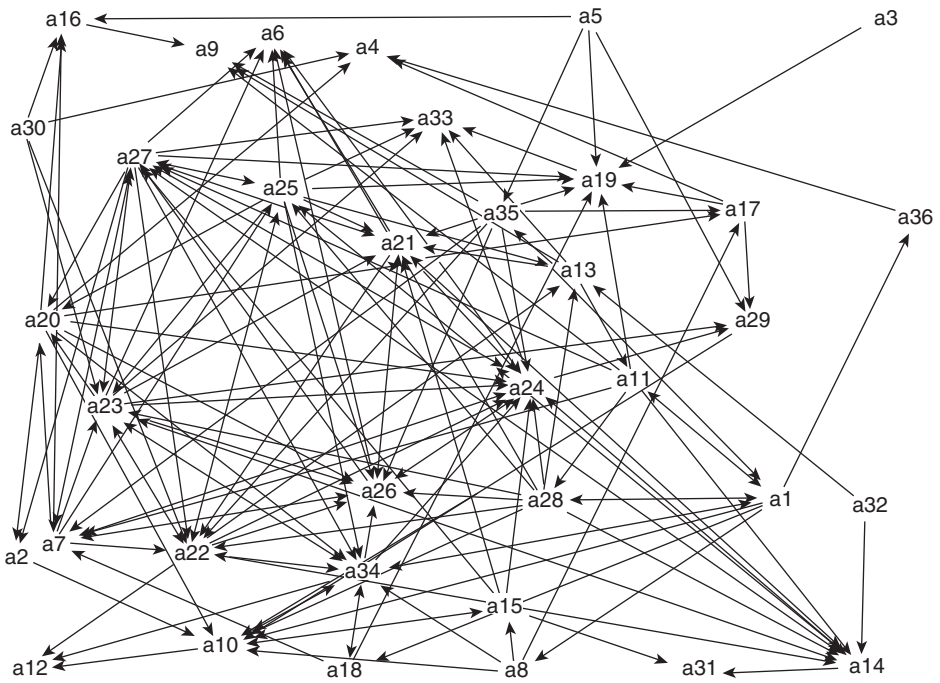

Figure 1 Graphical representation of interorganizational networks amongst mental health agencies in one US city (data from Morrisey, J.P., Calloway, M., 1994. Local mental health authorities and service system change: Evidence from the Robert Wood Johnson Foundation Program on chronic mental illness. The Milbank Quarterly 72 (1), 49–80.).

respectively. This general approach has also been developed for the analysis of multivariate social networks. The best-known example of this approach is probably the quadratic assignment procedure or model (*QAP*) for networks (Hubert and Baker, 1978). In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multivariate graph distribution that is conditional on the unlabelled graph structure of each univariate graph. In a more general framework for the evaluation of structural properties in multivariate graphs, Pattison et al. (2001) also introduced various multigraph distributions associated with stochastic blockmodel hypotheses.

Extra-Dyadic Local Structure in Networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss (1986) with the introduction of the class of Markov random graphs. This class of models permitted the parameterization of extra-dyadic local structural forms and so allowed a more explicit link between some important network conceptualizations and statistical network models. Frank and Strauss (1986) adapted research originally developed for modeling spatial dependencies (*see* Spatial Statistical Methods) among observations to graphs. They observed that the Hammersley-Clifford theorem provides a general probability distribution for $\boldsymbol{X}$ from a specification of which pairs (X_{ij}, X_{kl}) of edge random variables are conditionally dependent, given the values of all other random variables (*see* Graphical Markov Models: Overview).

Define a dependence graph $\mathbf{D}$ with node set $N(D) = \{(X_{ij} : i, j \in N, i \neq j\}$ and edge set $E(\mathbf{D}) = \{(X_{ij}, X_{kl})$: X_{ij} and X_{kl} are assumed to be conditionally dependent, given the rest of $\boldsymbol{X}\}$. Frank and Strauss used $\mathbf{D}$ to obtain a model for $\Pr(\boldsymbol{X} = \boldsymbol{x})$, denoted p^* by Wasserman and Pattison (1996), in terms of parameters and substructures corresponding to cliques of $\mathbf{D}$. The model has the form

$$\Pr(X = x) = p * (x) = (1/c)\exp\left\{\sum_{P \subset N(D)} \alpha_P Z_P(x\}\right\} \qquad [2]$$

where:

a. the summation is over all cliques P of $\mathbf{D}$ (with a clique of $\mathbf{D}$ defined as a nonempty subset P of $N(\mathbf{D})$ such that $|P| = 1$ or $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ for all $X_{ij}, X_{kl} \in P$);

b. $Z_p(x) = \prod_{x_{ij} \in p}^{x_{ij}} x_{ij}$ is the (observed) network statistic corresponding to the clique ρ of $\mathbf{D}$; and

c. $c = \sum_x \exp\{\sum_p \alpha_p z_p(\boldsymbol{x})$ is a normalizing quantity.

specification of a dependence graph raises a deep theoretical issue for network analysis: what constitutes an appropriate set of dependence assumptions? Frank and Strauss (1986) proposed a Markov assumption, in which $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ whenever $\{i, j\} \cap \{k, l\} \neq \varphi$. This assumption implies that the occurrence of a network tie from one node to another is conditionally dependent on the presence or absence of other ties in a local neighborhood of the tie. A Markovian local neighborhood for X_{ij} comprises all possible ties involving i and/or j. Any network tie is seen both as dependent on its neighboring ties and as participating in the neighborhood of

other ties, so the resulting model embodies the assumption that social networks have a local self-organizing quality. In the case of Markovian neighborhoods, cliques in the dependence graph correspond to triadic and star-like network configurations. A homogeneity assumption, that model parameters do not depend on the specific identities of nodes but only on the network configurations to which they correspond, yields a model that expresses the probability of the network in terms of propensities for subgraphs of triadic and star-like structures to occur.

For example, a homogeneous Markov model for the network of **Table 1** with parameters corresponding to edges, 2-stars, and triangles appears to be a substantial improvement in fit over a Bernoulli model with just an edge parameter, since the mean absolute residual for fitted edge probabilities drops from .220 to .175. Pseudolikelihood estimates for model parameters (Wasserman and Pattison, 1996) are −3.31, 0.056, and 0.616 for edges, 2-stars, and triangles, respectively, suggesting that graphs with a large number of triangles are relatively more likely in the assumed homogeneous Markov p^* distribution than those with few.

p^* random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity, local clustering, degree variability, and centralization. Valued (Robins et al., 1999) and multivariate (Pattison and Wasserman, 1999) generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability, generalized transitivity and exchange, and the strength of weak ties.

Prospects for Structural Models

The p^* family of random graph models poses many interesting questions, both substantive and statistical. A major empirical question is the sufficiency of these local structural descriptions: do they suffice, or do they need to accommodate other characterizations of local dependence, more global considerations, other types of constraints (e.g., temporal and spatial), or variability in local structural dependencies? On the statistical side, major questions are raised by the complexity and rich parameterization of models that are deemed to be theoretically plausible. For instance, does a Markov chain Monte Carlo (*see* Markov Chain Monte Carlo Methods) maximum likelihood approach to parameter estimation offer a viable alternative to approximate methods such as pseudolikelihood estimation (Strauss and Ikeda, 1990)? How can these distributions best be simulated, and what are the convergence properties of various approaches, such as the Metropolis-Hastings algorithm?

Dynamic Models

Empirically and theoretically defensible parametric models for networks constitute one challenge for statistical modeling, but an arguably more significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision-making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of longer-standing relational ties. Early attempts to model the evolution of networks in either discrete or continuous time assumed dyad independence and Markov processes in time.

A step towards continuous time Markov chain models for network evolution that relaxes the assumption of dyad independence has been taken by Snijders (1996). These models incorporate both random change and change arising from individuals' attempts to optimize features of their local network environment (such as the extent to which reciprocity and balance prevail) and use the computer package *SIENA* for fitting. A tension function is proposed to represent the function that actors wish to minimize and Robbins-Munro procedures are used to estimate free parameters. The approach illustrates the potentially valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such methods provide a promising focus for future development.

The value of simulation has also been demonstrated by Watts and Strogatz (1998) and Watts (1999) in their analysis of the so-called small-world phenomenon (the tendency for individuals to be linked by short acquaintanceship network paths). They considered a probabilistic model for network change in which the probability of a new tie was determined, in part, by existing patterns of neighboring ties (a clustering component) and, in part, by a small random component. They used simulations to establish that only a small random component was needed for the simulated network to be likely to possess the small-world property.

Models Bridging Analytic Levels

A recurring theme among social theorists is the need for models that integrate the dynamic, interdependent, interpersonal, and cultural contexts in which social action occurs. The network modeling challenge posed by such framing of the nature of social processes is the development of models that represent interdependencies across several analytic categories (e.g., individuals, relational ties, settings, groups).

Snijder's (1996) model for network change represents one step in this direction; the class of social influence models represents another (Friedkin, 1998). These models are exemplified by the network effect model for the mth attribute variable:

$$Z^{[m]} = \alpha x z^{[m]} + \gamma\beta + \varepsilon \tag{3}$$

where α is a network effects parameter, γ is a matrix of exogenous attribute variables with corresponding parameter vector β, and ε is a vector of residuals.

A further step in the direction of modeling interdependent node attributes and network ties has been to generalize the dependence graph approach to p^* network models (Section Extra-dyadic Local Structure in Networks) so as to include both node attribute variables ($Z^{[m]}$) and relational variables (X). This step is facilitated if directed dependencies are permitted in the dependence graph (i.e., dependencies in which one random variable is assumed to be conditionally dependent on a second, but the second not on the first). Robins et al. (2001) drew on graphical modeling techniques to construct models with directed dependencies among

network ties and individual attributes. If the dependencies are directed from network ties to node attributes, models for social influence are obtained; with dependencies oriented from node attributes to network ties, social selection models result. Indeed, this generalization introduces substantial flexibility in model construction, with potential applications to joint influence and selection models, discrete time models for network change, network path models, and group-level attribute models.

The dependence approach of Section Extra-dyadic Local Structure in Networks has also been used to construct principled models for more general multiway relational data structures representing interdependent analytic categories. Examples to date have included bipartite (Faust and Skvoretz, 1999) and *k*-partite (Pattison et al., 1998) graphs and illustrate how this general distribution has the capacity for a systematic and coherent approach to a variety of major network modeling challenges.

Conclusion

In addition to some of the specific future prospects already noted, a more general expectation is that the current classes of models are likely to be the foundation of substantial empirical and model-building activity in the early twenty-first century. From their early beginnings, statistical approaches to network modeling have now reached a point where sustained interaction is likely to be empirical research that allows the development of greatly improved models for processes occurring on networks, with important applications in many fields of social and behavioral science, including epidemiology, organizational studies, social psychology, and politics.

For more information on these approaches to social networks analysis, please see Wasserman and Faust (1994) and Wasserman and Pattison (1996, 2002).

See also: Graphical Markov Models: Overview; Multivariate Analysis: Discrete Variables (Loglinear Models); Network Analysis; Network Models of Tasks; Social Network Analysis; Spatial Statistical Methods.

Bibliography

Bollobas, B., 1985. Random Graphs. Academic Press, London.

Erdos, P., Renyi, A., 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61.

Faust, K., Skvoretz, J., 1999. Logit models for affiliation networks. In: Becker, M., Sobel, M. (Eds.), Sociological Methodology. Blackwell, New York, pp. 253–280.

Fienberg, S., Wasserman, S., 1981. Categorical data analysis of single sociometric relations. In: Leinhardt, S. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 156–192.

Frank, O., Nowicki, K., 1993. Exploratory statistical analysis of networks. In: Gimbel, J., Kennedy, J.W., Quintas, L.V. (Eds.), Quo Vadis Graph Theory? A Source Book for Challenges and Directions North-Holland, Amsterdam.

Frank, O., Strauss, D., 1986. Markov graphs. Journal of the American Statistical Association 81, 832–842.

Frank, O., 1981. A survey of statistical methods for graph analysis. In: Leinhardt, S. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 110–155.

Friedkin, N., 1998. A Structural Theory of Social Influence. Cambridge University Press, New York.

Holland, P.W., Leinhardt, S., 1975. The statistical analysis of local structure in social networks. In: Heise, D.R. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 1–45.

Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association 76, 33–50.

Hubert, L.J., Baker, F.B., 1978. Evaluating the conformity of sociometric measurements. Psychometrika 43, 31–41.

Lazega, E., van Duijn, M., 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. Social Networks 19, 375–397.

Morrisey, J.P., Calloway, M., 1994. Local mental health authorities and service system change: Evidence from the Robert Wood Johnson Foundation Program on chronic mental illness. The Milbank Quarterly 72 (1), 49–80.

Pattison, P.E., Wasserman, S., 1999. Logit models and logistic regressions for social networks, II. Multivariate relations. British Journal of Mathematical and Statistical Psychology 52, 169–194.

Pattison, P., Mische, A., Robins, G.L., 1998. The plurality of social relations: k-partite representations of interdependent social forms. In: Conference on Ordinal and Symbolic Data Analysis, Amherst, MA: Keynote address Sept. 28–30.

Pattison, P., Wasserman, S., Robins, G.L., Kanfer, A.M., 2000. Statistical evaluation of algebraic constraints for social networks. Journal of Mathematical Psychology 44, 536–568.

Rapoport, A., 1949. Outline of a probabilistic approach to animal sociology. I. Bulletin of Mathematical Biophysics 11, 183–196.

Robins, G.L., Pattison, P., Wasserman, S., 1999. Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika 64, 371–394.

Robins, G.L., Pattison, P., Elliott, P., 2001. Network models for social influence processes. Psychometrika 66, 161–190.

Snijders, T.A.B., Nowicki, K., 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification 14, 75–100.

Snijders, T.A.B., 1991. Enumeration and simulation methods for 0–1 matrices with given marginals. Psychometrika 56, 397–417.

Snijders, T.A.B., 1996. Stochastic actor-oriented models for network change. Journal of Mathematical Sociology 21, 149–172.

Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. Journal of the American Statistical Association 85, 204–212.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Application. Cambridge University Press, New York.

Wasserman, S., Pattison, P.E., 1996. Logit models and logistic regressions for social networks, I. An introduction to Markov random graphs and p*. Psychometrika 60, 401–425.

Wasserman, S., 1977. Random directed graph distributions and the triad census in social networks. Journal of Mathematical Sociology 5, 61–86.

Wasserman, S., Pattison, P.E., 2002. Multivariate Random Graph Distributions. Springer Lecture Note Series in Statistics, Springer-Verlag, New York.

Watts, D., Strogatz, S., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

Watts, D., 1999. Networks, dynamics, and the small-world phenomenon. American Journal of Sociology 105, 493–527.

White, H.C., Boorman, S., Breigen, R.L., 1976. Social structure from multiple networks, I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–780.

Social Networks

John Scott, University of Plymouth, Plymouth, UK; and University of Copenhagen, Copenhagen, Denmark
Frans N Stokman, University of Groningen, Groningen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous article by F.N. Stokman, volume 15, pp. 10509–10514, © 2001, Elsevier Ltd.

Abstract

This article reviews key ideas in social network analysis methods. Key concepts for describing the connections of individuals and groups into networks of social relations are introduced and defined. The article identifies the terminology of points and lines for understanding sociograms and reviews concepts of adjacency, degree, centrality, and forms of cohesion and subrouping. Developments in statistical methodology of hypothesis testing are examined and views of the evolution of social networks over time are considered. It is argued that social network analysis is to be considered as a methodological approach rather than a substantive body of theory, although it is shown that there are close links between social networks analysis and theories of social structure and social capital. It is argued that social network analysis is an essential adjunct to any approach to relational sociology.

Social network analysis is the study of the patterns of social relations that comprise social structures, treating these relations as networks of connections among the individuals and groups that enter into them. While these relations may be formed by particular individuals, social network analysis is not limited to microlevel interactions. Individuals form social relations as the occupants of institutionally defined positions in social organizations, and the social relations enter into the constitution of patterns of macrolevel relations that may equally be treated as social networks. A large number of concepts have been developed to characterize, measure, and compare network structures and positions in networks. These include the relative centrality of individuals, groups, and positions within networks; their clustering into subgroups; the overall cohesion or density of a network; and the centralization of networks around focal points. A number of statistical methods have been developed to estimate their values and to assess the significance of these measures for observed outcomes.

There are a number of introductions and overviews of social network analysis. A nontechnical introduction is Scott (2013), and there are a number of introductory surveys (Degenne and Forsé, 1994; Knoke and Song, 2008; Prell, 2012; Kadushin, 2012; Scott, 2012). The best advanced text is that of Wasserman and Faust (1994). A comprehensive collection of discussions and applications has been brought together by Carrington and Scott (Carrington and Scott, 2011).

Concepts and Techniques of Network Analysis

In social network analysis, social structures are represented as patterns of points (or vertices) and lines (or edges). The points represent individuals, organizations, or positions, while the lines represent the relations that connect them. Any type of dyadic relationship can be represented, including communication, friendship choices, advice, trust, influence, and exchange relationships. The lines depicting the relationship can be characterized by various attributes. Relations may first be described as directed or undirected. A line can be said to be directed from one point to another, for example, when one person chooses another as a friend but this choice is not reciprocated. The direction assigned to the line represents the likely flow of information or resources between the two participants. These directed lines are often termed 'arcs.' Undirected lines are those that represent reciprocated links or common memberships, as in the case of two people who meet as members of the same group or through participation at the same event. Lines can also be distinguished by signs attached to them to indicate the type of relationship. Positive or negative signs on directed lines, for example, enable the representation of positive and negative relations such as cooperation and conflict. The strength of a relation may be represented by a number, either an actual number (such as the size of a shareholding) or a scaled representation of strength. Such lines are said to be 'valued.'

This language of points and lines, with directions, signs, and values, provides an intuitive form of representation that can help in the drawing of simple sociograms. The sociogram, largely invented by Jacob Moreno (1934), is a simple visual mapping of the pattern of connections and is most easily used in the study of small groups. Indeed, the area of group dynamics (Cartwright and Zander, 1953) largely developed as a reflection on the consequences of different communication patterns in small groups. An example of this kind of work might be Heider's development of balance theory to study social influence within groups. Heider represented cognitive balance as a signed graph with three points. An individual under study can have a positive or negative relationship to another person. Both persons can have a positive or negative attitude toward another object (e.g., a third person or a certain activity). The cognitive system of the individual under study is in balance if and only if all three relationships are positive or two of them are negative. Harary et al. (1965) generalized the idea of balance to a whole social system.

Developments beyond simple representations in sociograms have involved the use of matrix algebra and graph theory (see particularly Harary et al., 1965). Major computational possibilities became possible by the representation of a social network in a matrix. The rows and columns represent the points, and the cells the relationships from the row to the column point. In an adjacency matrix, only ones (for a directed line from the row point to the column point) or

zeros (no such line) are used. In a signed graph, positive ties are represented by +1 and negative ones by −1. In a valued graph, the values of the ties can be given. Graphs with different types of relationships (denoted multigraphs) can be represented by stacked matrices. Graph theory is a branch of formal mathematics that has been used in the modeling of networks – or 'graphs' – in such diverse areas as electrical networks, computer networks, and networks of rivers. The use of graph theory allows the formal properties of large networks to be explored and summarized in numerical form. The basic concepts used to describe a pattern of points and lines in graph theory are adjacency, neighborhood, and distance, and these can be used to describe more complex relational structures (Bonacich and Lu, 2012).

Adjacency refers to the simple fact of connection: two connected points are said to be adjacent to each other. The 'degree' of a point is the number of other points to which it is adjacent, and the neighborhood of a point comprises the set of other points to which it is adjacent. When account is taken of the direction assigned to a line, it is possible to distinguish the 'indegree' and the 'outdegree' of a point and so to differentiate its neighborhood. Points in a network are connected by 'paths' – sequences of lines – of varying length, the number of lines in a path. The length of the shortest path indicates the distance between two points. Adjacent points are connected by a path comprising a single line and are said to be connected at a distance of one. Indirect links through, for example, a common friend are at a distance of two, and so on. The famous research of Stanley Milgram (1967) showed the significance of six degrees of separation: the fact that randomly chosen people, even if located in different continents, are typically connected to each other through intermediaries by a path of maximally length six.

These basic concepts allow the construction of measures of the relative centrality of points within a network. Freeman (1978) associated the many centrality measures with three important centrality dimensions in communication networks. These are local centrality, global centrality, and betweenness. Local centrality is measured by the degree, the number of points with which a point is directly connected, which is assumed to indicate the communication activity of a point. It is a measure of how well connected a point is in its local environment. In directed networks, centrality in terms of outdegree and indegree should usually be distinguished. In friendship choice networks, for example, the number of choices received (indegree) generally indicates centrality, which may be seen as a measure of likability. In influence networks, local centrality may be based on the number of outgoing relationships (outdegree). When the overall network is considered, rather than its local structure, one is dealing with global centrality. This involves the use of distance-based measures to indicate the relative proximity of points to other points in the network and the extent to which a point can communicate with other points independently of others. In the literature, various statistical measures on the distances of a point to other points are used to measure this dimension of centrality, such as the mean distance to other points or the maximal distance to any of the other points. Betweenness, or rush, is the third type of centrality and measures how important a point is for the transmission of information between other points. Betweenness measures assume that information is mainly transmitted through the shortest paths. A point with a high betweenness might be regarded as a key 'broker' or intermediary in a network.

The measures considered so far can all be considered to be attributes of particular points within a network, based on their relations with others. They are often referred to as measures for 'egonets.' However, it is also possible to consider the global structure of the network itself and to characterize its overall pattern. The most important are density, centralization, and clustering. Density is the number of distance-one relationships that actually exist relative to the total number that might possibly exist. It measures how 'complete' a web of connections may be. Dense networks are more important for control and sanctioning than for information. They also tend to generate a lot of redundant information and so result in many constraints and are inefficient for creative new solutions (Burt, 1992). Comparing densities of networks of quite different sizes is difficult, as large networks tend to be sparse and the measure of density varies directly with the typical neighborhood size (which increases much more slowly than the size of the whole network). The centralization of a network is a measure of the extent to which its connections coalesce around a small subset of globally central points. In connected networks, high centralization corresponds with a high variance of the degrees of the points. Snijders (1981) derived the maximal possible values of the variance, given the number of points or the number of points and lines. He also derived the expected variance for different null models under the same two conditions. This makes a good comparison of centralization possible in networks of different sizes and densities. The third global measure is the clustering or segmentation of the network. A simple summary measure of segmentation is the number of pairs of points at distance two or higher divided by the number of pairs of points at distance three or higher (Baerveldt and Snijders, 1994). More structural measures, however, employ forms of cluster analysis and subdivision to identify clusters, cliques, and components within a network. Such measures of the social circles within which individuals and groups interact provide a picture of the overall shape or landscape of a network. Structurally equivalent points do not have the same connections, but they do have the same types of connections. This kind of analysis has been seen as identifying typical roles or positions within networks.

A large number of other network studies examine the effects of the network structure on the behavior and attributes of the network members. In these studies, the network is considered as given and constant, and the ways in which this network influences processes or individuals in the network is examined. In many of these studies, only relationships of the individuals under investigation are collected (at most combined with their perception of the relationships among their network members). These ego-centered network studies examine the effects of differences in size and composition of the personal networks and the multiplexity of an individual's personal relations (Wellman and Berkowitz, 1997). Examples of effects studied are an individual's social well-being, social support, health, labor market position, and career.

Many of the central properties of social networks studied by graph theory have been shown to depend on what have been

called small-world conditions (Watts, 1999, 2003). That is, they apply when the pattern of connections is such that Milgram's suggestion of six degrees of separation holds. Fundamentally different processes will be in operation in networks where the number of connections is so low that this does not happen or where the connections are so high that almost all individuals are closely connected. Thus, a network that changes incrementally may reach a point at which the pattern of connections changes catastrophically and with radical consequences for the members of the network.

The bulk of the measures in social network analysis are based on so-called one-mode data. That is, when individuals are related to each other by common organizational membership, analysis has focused on either the individuals or the organizations. Recent developments have undertaken systematic simultaneous analyses of both individuals and organizations, representing them all as points within the same network. Measures of centrality, for example, are difficult to conceptualize and understand for two-mode data, but much work is progressing to clarify this. An early usage of two-mode data was block modeling, in which both sets of points were simultaneously partitioned into subsets that could be treated as 'equivalent' to each other (Everett and Borgatti, 1994).

An important area of development in social network analysis has been the use of multidimensional scaling, principal components analysis, and similar techniques to present graphical visualizations of social networks. These techniques allow a more rigorous extension of the principle of the sociogram and embed images of social networks in a multidimensional space in which distances and directions are depicted as they would be in a geographical map. Thus, distance in this social space differs from the path distance of graph theory as it takes account of the relative closeness of all points within the network. These developments have led to an impressive range of visualization techniques, many of which are accessible through the Pajek software (de Nooy et al., 2005). General network concepts can be usefully explored through the UCINET software developed by Freeman, Borgatti, and Everett (http://www.analytictech.com/).

Statistical Models for Social Network Analysis

Much research in social network analysis has been descriptive and has prompted questions about the significance of the structures described for actual behavioral outcomes. Recent developments have produced a range of statistical techniques that can allow the use of social network analysis in explanatory studies by assessing the statistical significance of the results.

Standard statistical models are more difficult to apply to social networks because the relationships between the points cannot be treated as independent observations. The most useful alternative to have been developed involves exponential random graph models in which observed patterns are compared with those found in large numbers of random computer simulations given certain characteristics of the overall network to assess the probability that the observations are the result of chance (Wasserman and Pattison, 1996; Snijders et al., 2006; Robins et al., 2007).

Evolution and Network Dynamics

Along with the move from descriptive to explanatory concerns, there has been a move from static to dynamic analyses. Existing concepts have been largely concerned with the static, cross-sectional features of social networks and fail to grasp the complexities of social change. Recent work has taken more seriously the fact that social networks change over time and that these patterns and processes of change must be explained. Like all social structures, social networks are produced and reproduced through individual and collective action, but their structures are generally the unanticipated and unintended consequences of these actions. Agent-based computational models of action (Axelrod, 1997) have been particularly useful in understanding this process and in allowing connections to be made with Markov models of stochastic change.

Holland and Leinhardt (1977) introduced Markov processes as the general framework for stochastic models of network evolution. The basic idea of Markov models is to conceive the social network structure as changing from one state into another over time. The parameters that govern the process concern the likelihood of transition from one of these four states into another. The original Markov models assume that the parameters are stationary over the whole process and that the population is homogeneous. Recent models have considerably increased the analytic possibilities of Markov models by eliminating these strongly limiting assumptions and recognizing that change parameters may well be dependent on the stage of network development and will be different for pairs within and between subgroups (Leenders, 1996). Snijders powerfully integrated Markov models with random utility models to allow a much stronger link between theory and statistical testing (Snijders, 1996), where actors are seen as making and breaking links on the basis of local knowledge and decision rules that generate unplanned but structured long-term changes. Many of the general issues in this area are considered in Doreian and Stokman (1997).

One of the most important theoretical topics in social network analysis concerns the relative importance of selection and influence processes: do, for example, friends seek and select friends with similar characteristics or become friends more similar because of their repeated interactions? As both networks and characteristics like attitudes change over time, the relative importance of the two processes is one of the key questions social networks researchers seek to answer. Snijders' dynamic approach to network analysis (Snijders, 2011) makes it possible to estimate the relative importance of the two processes in designs where we have data at three or more time points.

Method or Theory?

Social network analysis is primarily to be understood as a method of analysis and not as a substantive theoretical approach. It is, however, more compatible with certain theoretical approaches and can be considered as an essential adjunct of any relational sociology (Emirbayer, 1997). Early

arguments emphasized the links between social network analysis and exchange theory.

The fact that social networks create conditions for cooperation through information and sanctioning is due to exchange processes that create win–win situations (Homans, 1950; Blau, 1964). A fruitful and very promising approach is the study of the effects of social networks in noncooperative game theory (Raub and Weesie, 1990; Flache and Macy, 1996; Bienenstock and Bonacich, 1992). Network exchange theory specifically investigates the effects of network structures on the choice between alternative exchanges and on exchange rates (Willer, 1999). Major effects are particularly due to possibilities for social actors to exclude others. Exchange network theory illustrates again that effects of network structures are context sensitive and cannot be generalized without taking the context and substance into account. The integration of exchange theory and social networks has also been proved to be very successful in the field of policy networks. Most of these models build on Coleman's social exchange model and confine exchanges to influence network relationships (Laumann et al., 1987). These models make the step from microbehavior to macroeffects explicit and are able to predict outcomes of decisions, to derive the power of social actors and the value of decisions. Later models try to solve a number of remaining theoretical issues (Stokman and Van den Bos, 1992; Pappi and Henning, 1998). Stokman et al. (2013) give a more general theory of bargaining in which three fundamental bargaining processes (persuasion, logrolling, and enforcement) are related to three types of social networks (information, exchange, and hierarchical power networks). They argue that in any bargaining situation all three are present, but only one is likely to be dominant.

Social network analysis has recently, however, been shown to be especially fruitful for discussions of social capital. Social capital gives individuals access to resources of others that can be exploited for the realization of their goals and so constitutes an opportunity structure generated by social relationships (Lin, 1982; Coleman, 1990; Burt, 1992). The amount of social capital depends on the amount of these resources, their value for the goal realization of the individual, and the willingness of others to mobilize them (Flap, 1999). The value of the resources for the individual strongly depends on functional interdependence, the willingness of others to mobilize resources for the individual on their perception of the interdependence (i.e., the cognitive dependence). Social capital emphasizes the relationship between social capital and success of actors, whereas Lindenberg and Foss (2011) change the perspective to success of group production and derive the conditions for joint production motivation. Important ideas developed by Putnam (2000) on bridging and bonding capital can be directly modeled in network terms. Large and dense networks create shared information, high visibility, and common norms in a community. Burt (1992), on the other hand, stresses the importance of unique and nonoverlapping relationships for acquiring unique information in organizations, giving individuals a better chance to find creative solutions for problems and thus providing them better opportunities for career. Vedres and Stark (2010) disputed Burt's emphasis on broker positions for performance and introduced the concept of structural folds, where success is related to overlapping memberships in different groups rather than isolated positions between groups.

Functional and cognitive interdependencies differ particularly between what Granovetter (1973) has termed strong and weak ties. Strong ties are valued in themselves. The ties are not primarily instrumental for the attainment of other goals. Their value is based on the other individual as a person and the quality of the relationship with that individual. Family and friendship ties are typical examples of such relationships. Strong ties tend to be reciprocal, transitive, and clustered. Strong ties give a sense of belonging to a group and the group often has priority above the individual and individual relationships. Sharing is often based on need and norms tend to promote equality (Lindenberg, 1998). Creating negative attitudes toward other groups often helps to strengthen the predominance of the group, which may give strong negative externalities for society as a whole (think of gangs and other criminal organizations). Weak ties are valuable long-term relationships but their value is primarily instrumental, related to higher ordered goals, goals not primarily located in the relationship or individuals themselves. Weak ties tend to be less clustered; the group is less dominant and often only vaguely delineated. Reciprocity based on equity norms prevails.

Conclusion

Since the early 1970s, there has been impressive cumulative progress in social network analysis and research. From a rather isolated field, strongly oriented to descriptive structural and static analysis, social network analysis has grown into a well-embedded field, widely accepted as highly important for solving central theoretical problems of cooperation and coordination. With the growing importance of social networks in the information society with virtual communities developing in many segments of society, its importance and contributions to theoretical solutions can only grow. Issues of scope are challenging. Objects of study will vary from small group networks to social networks of billions of points. New techniques to visualize networks are also challenging. These allow certain structural characteristics to become visible and the effects of changes on these characteristics to become transparent. These developments can be followed by linking with the international virtual community of the International Network for Social Network Analysis and the links to be found there (http://www.insna.org/).

See also: Action, Theories of Social; Exchange: Social; Rational Choice Theory in Sociology; Social Structure; Status and Role: Structural Aspects.

Bibliography

Axelrod, R., 1997. The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton University Press, Princeton, NJ.

Baerveldt, C., Snijders, T.A.B., 1994. Influences on and from the segmentation of networks: hypotheses and tests. Social Networks 16, 213–232.

Bienenstock, E.J., Bonacich, P., 1992. The core as a solution to exclusionary networks. Social Networks 14, 231–244.

Blau, P., 1964. Exchange and Power in Social Life. Wiley, New York.

Bonacich, P., Lu, P., 2012. Introduction to Mathematical Sociology. Princeton University Press, Princeton and Oxford.
Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.
Carrington, P.C., Scott, J. (Eds.), 2011. Sage Handbook of Social Network Analysis. Sage, London.
Cartwright, D., Zander, A. (Eds.), 1953. Group Dynamics. Tavistock, London.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press, Cambridge.
de Nooy, W., Mrvar, A., Batagelj, V., 2005. Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York.
Degenne, A., Forsé, M., 1994. Introducing Social Networks. English translation. Sage Publications, Beverley Hills, 1999.
Doreian, P., Stokman, F.N. (Eds.), 1997. Evolution of Social Networks. Gordon and Breach, Amsterdam.
Emirbayer, M., 1997. Manifesto for a relational sociology. American Journal of Sociology 103 (2), 281–317.
Everett, M.G., Borgatti, S.P., 1994. Regular equivalence: general theory. Journal of Mathematical Sociology 19, 29–52.
Flache, A., Macy, M.W., 1996. The weakness of strong ties: collective action failure in highly cohesive groups. Journal of Mathematical Sociology 21, 3–28.
Flap, H.D., 1999. Creation and returns of social capital. A new research program. La Revue Tocqueville 20, 1–22.
Freeman, L.C., 1978. Centrality in social networks—conceptual clarification. Social Networks 1, 215–239.
Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.
Harary, F., Norman, R.Z., Cartwright, D., 1965. Structural Models. An Introduction to the Theory of Directed Graphs. Wiley, New York.
Holland, P.W., Leinhardt, S., 1977. A dynamic model for social networks. Journal of Mathematical Sociology 5, 5–20.
Homans, G., 1950. The Human Group. Harcourt Brace, New York.
Kadushin, C., 2012. Understanding Social Networks: Theories, Concepts, and Findings. Oxford University Press, New York.
Knoke, D., Song, Y., 2008. Social Network Analysis, second ed. Sage, Beverley Hills, CA.
Laumann, E.O., Knoke, D., Kim, Y.H., 1987. Event Outcomes. In: Laumann, E.O., Knoke, D. (Eds.), The Organizational State. Social Choice in National Policy Domains. The University of Wisconsin Press, Madison, WI.
Leenders, R.T.A.J., 1996. Evolution of friendship and best friendship choices. Journal of Mathematical Sociology 21, 133–148.
Lin, N., 1982. Social resources and instrumental action. In: Marsden, P.V., Lin, N. (Eds.), Social Structure and Network Analysis. Sage, London.
Lindenberg, S., 1998. Solidarity: its microfoundations and marco dependence: a framing approach. In: Doreian, P., Fararo, T.J. (Eds.), The Problem of Solidarity: Theories and Models. Gordon and Breach, Amsterdam, pp. 61–112.
Lindenberg, S., Foss, N.J., 2011. Managing joint production motivation: the role of goal-framing and governance mechanisms. Academy of Management Review 36, 500–525.
Milgram, S., 1967. The small world problem. Psychology Today 2, 60–67.
Moreno, J.L., 1934. Who Shall Survive? Beacon Press, New York.
Pappi, F.U., Henning, C.H.C.A., 1998. Policy networks: more than a metaphor? Journal of Theoretical Politics 10, 553–575.
Prell, C., 2012. Social Network Analysis: History, Theory and Methodology. Sage, London.
Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York.
Raub, W., Weesie, J., 1990. Reputation and efficiency in social interactions: an example of network effects. American Journal of Sociology 96, 626–654.
Robins, G., Snijders, T.A.B., Wang, P., Handcock, M.S., Pattison, P., 2007. Recent developments in exponential random graph (p*) models for Social Networks. Social Networks 29, 192–215.
Scott, J., 2012. Social Network Analysis, third ed. Sage, London.
Scott, J., 2013. What Is Social Network Analysis? Bloomsbury, London.
Snijders, T.A.B., 1981. The degree variance: an index of graph heterogeneity. Social Networks 3, 163–174.
Snijders, T.A.B., 1996. Stochastic actor-oriented models for network change. Journal of Mathematical Sociology 21, 149–172.
Snijders, T.A.B., 2011. Network dynamics. In: Carrington, P.C., Scott, J. (Eds.), Sage Handbook of Social Network Analysis. Sage, London.
Snijders, T.A.B., Pattison, P., Robins, G., Handcock, M.S., 2006. New specifications for exponential random graph models. In: Stolzenberg, R.M. (Ed.), Sociological Methodology, vol. 36. Basil Blackwell, Oxford.
Stokman, F.N., Van den Bos, J.M.M., 1992. A two-stage model of policy making: with an empirical test in the U.S. energy policy domain. In: Moore, G., Whitt, J.A. (Eds.), Research and Society, The Political Consequences of Social Networks, vol. 4. JAI Press, Greenwich, CT, pp. 219–253.
Vedres, B., Stark, D., 2010. Structural folds: generative disruption in overlapping groups. American Journal of Sociology 115, 1150–1190.
Wasserman, S., Faust, K., 1994. Social Network Analysis. Methods and Applications. Cambridge University Press, Cambridge, UK.
Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p*. Psychometrika 60, 401–426.
Stokman, F.N., Van der Knoop, J, Van Oosten, R.C.H., 2013. Modeling collective decision making. In: Nee, V., Snijders, T.A.B., Wittek, R (Eds.), Handbook of Rational Choice Social Research. Stanford University Press, Stanford, CA, pp. 151–182.
Watts, D., 1999. Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, Princeton.
Watts, D., 2003. Six Degrees. The Science of a Connected Age. W.W. Norton, New York.
Wellman, B., Berkowitz, S.D. (Eds.), 1997. Social Structures: A Network Approach. JAI Press, Greenwich, CT.
Willer, D. (Ed.), 1999. Network Exchange Theory. Praeger, Westport, CT.

Social Networks, Algebra of

Philippa Pattison, School of Psychological Sciences, University of Melbourne, Parkville, VIC, Australia
Lucia Falzon and Nectarios Kontoleon, Defence Science and Technology Organisation, Edinburgh, SA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We summarize algebraic representations for multiple social networks and review recent steps that have been taken to extend these representations to the case of dynamic multiple networks. The primary purpose of these algebraic constructions is to establish the foundations for some useful computations in multiple networks and to develop the basis for structural analysis and comparison of multiple network structures.

Introduction

Since social networks are fundamentally relational structures, it is not surprising that algebraic approaches to understanding relational structure have been explored in the context of social networks (for example, Bonacich and McConaghy, 1979; Boorman and White, 1976; Boyd, 1989, 1990; Kim and Roush, 1984; Luce, 1956; Boyd et al., 1989; Lorrain, 1975; Lorrain and White, 1971; Pattison, 1993; see also Pattison, 2009, 2011; Pattison and Bartlett, 1982; White, 1963). In this article, we summarize algebraic representations for multiple social networks and review recent steps that have been taken to extend these representations to the case of dynamic multiple networks. As we describe below, the primary purposes of these algebraic constructions are as follows: first, to establish the foundations for some useful computations in multiple networks and, second, to develop the basis for structural analysis and comparison of multiple network structures.

We begin with a fixed primitive set of *relational types*, such as friendship, collaboration, acquaintanceship, and so on. Each relational type is taken to be associated with a *network* on a fixed set N of actors or nodes. The network can be represented interchangeably as a *binary relation* (a set of ordered pairs of elements of N), a *directed graph* (a set N of nodes and a set of directed edges, each of which is an ordered pair from N), or an *adjacency matrix* on the node set N (that is, a matrix whose (i,j) entry is 1 if and only if (i,j) is an element in the binary relation for the network or, equivalently, an edge in its directed graph). For a relation of type A, we write $a(i,j) = 1$ if i is connected to j in the network for the relation, and $a(i,j) = 0$, otherwise. In other words, we represent the relation of type A as an $n \times n$ matrix $a = [a(i,j)]$, and the entries of the matrix a have values from the set $\{0,1\}$. We refer to unit entries of $a(i,j)$ as *links*, or *ties*, in the network a.

In the case of some relationships, such as parenthood, the direction of the network connection between two nodes is a vital consideration, whereas in others it may be assumed to be unimportant as, for example, in relationships of collaboration. Where the directionality of network ties is assumed a priori not to matter, the network is defined to be symmetric; that is, $a(i,j) = a(j,i)$ for all pairs i, j of nodes in N.

For a given set $\{A_1, A_2, \ldots, A_r\}$ of relational types, the set of networks $\alpha = \{a_1, a_2, \ldots, a_r\}$ on N constitutes a *multiple social network* on N. Below, we refer to the network elements of α as *generator* networks. To illustrate the constructions presented here, we use the hypothetical small example of a multiple social network $\alpha = \{a,b\}$ presented in **Table 1**.

In the first part of this article, we focus on the algebraic structures to which labeled walks and their sums in a multiple social network gives rise. In the second part, we consider *dynamic multiple social networks*, in which the set $\{0,1\}$ of possible relational values of $a(i,j)$ is replaced by a set of time intervals over which the relation or tie from node i to node j is taken to exist. Our focus then turns to time-ordered labeled walks (Moody, 2002) and the algebraic structures to which such walks and their sums give rise.

The Algebraic Structure of a Multiple Social Network

It is possible to define a variety of algebraic structures from a multiple social network (Pattison, 2009, 2011). The variations arise, in part, through the choice of elements over which the algebra is defined and, in part, through choice of the *operations* on those elements. As far as the choice of elements is concerned, most approaches have treated the constituent networks in a multiple social network as a set of generator elements (e.g., Boorman and White, 1976; Boyd, 1969, 1990; Boyd et al., 1989; Luce, 1956; Pattison, 1993; White, 1963), though some have restricted attention to the immediate *local* neighborhood of a selected node in the network (Breiger and Pattison, 1986; Mandel, 1983; Pattison, 1989; Winship and Mandel, 1983; Wu, 1983). The choice of operations has focused primarily on the binary relations of relational composition, Boolean addition and intersection and the unary converse or closure operations (e.g., Pattison, 2011). A natural partial order on networks has often also been incorporated (e.g., Pattison, 1993). Here, we focus principally on relational composition and Boolean sums as operations of primary interest, and also on the natural partial order relation, as explained below.

Table 1 A small hypothetical multiple social network $\alpha = \{a,b\}$ on node set $N = \{1,2,3,4\}$

	a				b			
1	0	1	0	0	0	0	1	0
2	0	0	0	0	0	0	0	1
3	0	0	0	1	1	0	0	0
4	0	0	0	0	0	1	0	0

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43116-8

The binary operation of *relational composition* is defined for arbitrary networks A and B on a node set N as the network corresponding to the Boolean product of the arrays a and b:

$$ab(i,j) = \sum_k a(i,k)b(k,j)$$

Note that both addition and multiplication are Boolean in this expression. The matrix $ab = [ab(i,j)]$ is said to have type AB and describes the presence or absence of labeled walks of type AB among pairs of members of the node set. In general, a *labeled walk* of type $A_1A_2\cdots A_p$ from a node i to a node j is a sequence of nodes $i = k_0, k_1, \ldots, k_p = j$ in which each pair (k_{m-1}, k_m) of adjacent nodes in the sequence is an edge in the network a_m $(m = 1, 2, \ldots, p)$. The length of the walk is p. Labeled walks in a multiple social network are important because they describe indirect relations implied by the generator networks and serve as potential conduits for social processes. For instance, if A refers to friendship and B refers to collaboration, then labeled walks of type AB describe pairs (i,j) of actors for whom j is a collaborator of a friend of i. If the generator set of networks comprises a single network a of type A, then aa, $(aa)a$, and so on, indicate pairs of nodes that are linked by walks in the network a of length 2, 3, and so on.

The *(Boolean) sum* of two arbitrary networks A and B on the same node set N is defined pairwise; that is,

$$(a+b)(i,j) = a(i,j) + b(i,j)$$

The pair of actors (i,j) are linked by the network $a + b$ if and only if i is linked to j by the network a and/or the network b. We refer to the type of the network $a + b$ as $A + B$. If A and B refer to friendship and collaboration, for example, then two nodes are linked by a relation of type $A + B$ whenever they are linked by a friendship and/or a collaboration tie.

We can also define a *natural partial order* on networks: for any two networks a and b on the same node set N, define $a \leq b$ if and only if

$$a(i,j) \leq b(i,j) \quad \text{for all } i,j \in N$$

Definition 1

A binary relation $\leq$ on a set S is a *partial order* if it has the properties:

1. *reflexivity*: $a \leq a$ for all $a \in S$;
2. *transitivity*: for any $a, b, c \in S$, if $a \leq b$ and $b \leq c$, then $a \leq c$; and
3. *antisymmetry*: for any $a, b \in S$, if $a \leq b$ and $b \leq a$, then $a = b$.

The natural partial order on networks with the same node set N is clearly a partial order.

The operations of relational composition and Boolean addition and the natural partial order relation have some well-known algebraic properties that give rise to some well-described algebraic structures on networks.

Definition 2

A (finite) *semigroup* $(S, \circ)$ is a (finite) set S endowed with a binary operation satisfying:

1. *closure*: for all $a, b \in S$, $a \circ b \in S$; and
2. *associativity*: for all $a, b, c \in S$, $(a \circ b) \circ c = a \circ (b \circ c)$.

If, in addition, $(S, \circ)$ satisfies:

3. *identity*: there is an *identity element*, $e \in S$ satisfying, for all $a \in S$, $a \circ e = e \circ a = a$,

then $(S, \circ)$ is termed a *monoid*.

Definition 3

A *partially ordered* semigroup (monoid) is a semigroup (monoid) $(S, \circ)$ together with a partial order $\leq$ on S that satisfies

$$\textit{isotonicity}: \text{if } a \leq b \text{ and } c \leq d, \text{ then } a \circ c \leq b \circ d \quad \text{for any } a, b, c, d \in S$$

It can readily be shown that the set of all binary relations on a set N forms a monoid under relational composition. The identity element is the *identity network* defined by

$$e(i,j) = 1 \quad \text{if } i = j; e(i,j) = 0, \text{ otherwise.}$$

Definition 4

For a given set of networks $\alpha = \{a_1, a_2, \ldots, a_r\}$ on node set N, we can define the *closure* α^* of α under relational composition as the set α^* of networks satisfying

1. *closure*: for all $a, b \in \alpha^*$, $a \circ b \in \alpha^*$;
2. *inclusive of generators*: $\alpha \subseteq \alpha^*$; and
3. *minimal*: for any set β that also satisfies closure and includes the generator set (that is, $\alpha \subseteq \beta$), then $\alpha^* \subseteq \beta$.

For any set of generator networks $\alpha = \{a_1, a_2, \ldots, a_r\}$ on N, α^* is a semigroup under relational composition, and $\alpha^* \cup e$ is a monoid (e.g., Boorman and White, 1976). Boorman and White referred to the semigroup α^* as the *role structure* on N generated by α, since it summarizes relationships of equality among all possible direct and indirect relations associated with the generator networks in α. It is worth emphasizing that implicit in the creation of α^* is the notion that any two networks in α^* are regarded as equal, if they link precisely the same pairs of actors in N. Boorman and White referred to this assumption as the *Axiom of Quality*, observing that it captures equations among direct and indirect relations on N. In other words, the semigroup α^* captures equations among all possible labeled walks that can be constructed from the generator networks $\{a_1, a_2, \ldots, a_r\}$.

Pattison (1993) argued that the partial order on networks in α^* also carries valuable information for analysis of the semigroup, and α^* and $\alpha^* \cup \{e\}$ are partially ordered semigroups and monoids, respectively, for any network generator set α.

In practice, the semigroup α^* of a multiple social network $\alpha = \{a_1, a_2, \ldots, a_r\}$ may be constructed as follows:

S1. Let $w_1 = \{a_1, a_2, \ldots, a_r\}$ and set $i = 1$.
S2. Construct the networks ab for each $a \in w_i$ and each $b \in w_1$. Place ab in the set w_{i+1} if $ab \neq c$, for any $c \in w_j$, $j = 1, 2, \ldots, i+1$; otherwise, $ab = c$ for some $c \in w_j$ $(j = 1, 2, \ldots, i+1)$ is an *equation* in the semigroup *on* α^*.
S3. If $w_{i+1} = \varnothing$, compute the partial order among networks in $w_1 \cup w_2 \cup \cdots \cup w_i$ and stop; otherwise, set i to $i+1$ and return to step S2.

The equations $ab = c$ may be recorded in a *right multiplication table* in which c appears as the table entry corresponding to the row labeled a and the column labeled b. If $n\alpha$ is the number of distinct networks in $\alpha^* = w_1 \cup w_2 \cup \cdots \cup w_i$, then the partial order can be written as an $n\alpha \times n\alpha$ matrix p in which $p(i,j) = 1$ if the jth element in α^* is less than the ith element in α^* and $p(i,j) = 0$, otherwise.

Definition 5

A *semiring* $(S,\circ,+)$ is a set S equipped with two binary operations $\circ$ and $+$ satisfying

1. $(S,\circ)$ is a monoid with identity e;
2. $(S,+)$ is a monoid with identity o;
3. *commutativity* of $+$: for all $a, b \in S$, $a + b = b + a$;
4. *distributivity* of $\circ$ over $+$: for all $a, b, c \in S$, $a\circ(b + c) = (a \circ b)+(a \circ c)$ and $(a + b)\circ c = (a \circ c)+(b \circ c)$; and
5. *absorption*: for all $a \in S$, $a \circ o = o \circ a = o$.

If $\leq$ is a partial order on S, then $(S,\circ,+,\leq)$ is a *partially ordered semiring* if, in addition,

6. *isotonicity* with respect to $\circ$: if $a \leq b$ and $c \leq d$, then $a \circ c \leq b \circ d$ for any $a, b, c, d \in$ S; and
7. *isotonicity* with respect to $+$: if $a \leq b$ and $c \leq d$, then $a + c \leq b + d$ for any $a, b, c, d \in$ S.

The set of all binary relations on a set N forms a partially ordered semiring, with relational composition and the Boolean sum taking the roles of the operations $\circ$ and $+$ and the natural partial order on networks on N. The identity element for the $+$ operation is the *null network* o defined by

$$o(i,j) = 0 \quad \text{for all } i,j \in N$$

We can also define the closure of the generator set α of networks under both relational composition and Boolean addition. In this case, the closure α^* is the minimal superset of α for which

1. *closure*: for all $a, b \in \alpha^*$, $a \circ b \in \alpha^*$, and $a + b \in \alpha^*$.

We will use α^* to refer to the closure of α, noting that it will be clear from the context over which operations closure applies.

The set of relations $\alpha^* \cup \{e,o\}$ closed under relational composition and Boolean addition and augmented by the identity and null networks form a semiring (e.g., Pattison, 2009). The semiring is partially ordered by the natural partial order on networks. The semiring may be constructed using a modified version of the algorithm above. The modification is to add an additional step between steps (S2) and (S3) of the algorithm above:

- compute $a + b$ for all $a \in w_j$, $j = 1,2,\ldots,i$; $b \in w_i$;
- place $a + b$ in w_{i+1} if $a + b \neq c$, for any $c \in w_j$, $j = 1,2,\ldots,i + 1$;
- otherwise, $a + b = c$ for some $c \in w_j$ $(j = 1,2,\ldots,i + 1)$ is an *equation* in the semiring.

In the special case where $\alpha = \{a\}$ comprises a single network generator, a number of mathematical results describe the relationship between network properties of a and algebraic properties of α^* (Kim, 1982; Pattison, 2009).

To illustrate these constructions, the multiple social network in **Table 1** gives rise to the partially ordered monoid whose distinct elements are listed in **Table 2** and whose right multiplication table and partial order are presented in **Table 3**. The same multiple social network gives rise to the partially ordered semiring with right multiplication table and partial order shown in **Table 4**. The addition table for the semiring is not shown but can be computed in a straightforward way using the Boolean addition operation on the elements of α^*.

Three primary applications have been made of these (partially ordered) semigroup, monoid, and semiring representations of multiple social networks in the literature on social network analysis.

The first is that many measures of interest in single and multiple social networks depend on walks in the network and hence formal algebraic representations of the type just defined establish the foundations for computation of those measures. It is well known, for example, that the *reachability* of node j from node i in a walk of length m or less in network a is captured by the (i,j) entry of the network $e \cup a \cup a^2 \cup \cdots \cup a^m$. Efficient algorithms for computation of walk-based measures can be developed given the underpinning algebraic properties (Brandes, 2001; Harary et al., 1965).

The second application takes advantage of the fact that these algebraic structures represent the relationships among labeled walks (and, in the case of semirings, their Boolean sums) in a way that does not depend directly on nodes and ties in the network constituents of α^*. As a result, structural comparisons across multiple social networks become possible, either through the identification of shared *homomorphic images* or shared sets of algebraic equations (Boorman and White, 1976; Bonacich, 1980; Bonacich and McConaghy, 1979; Boyd, 1969; Breiger and Pattison, 1978; McConaghy, 1981; Pattison, 1982, 1993).

Definition 6

If $\{a_1, a_2, \ldots, a_r\}$ and $\{a'_1, a'_2, \ldots, a'_r\}$ are two multiple social networks with a common set of relational types $\{A_1, A_2, \ldots, A_r\}$,

Table 2 Distinct elements in the monoid generated by the multiple social network $\alpha = \{a,b\}$ of **Table 1**

	a				b				o				ab				e			
1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 3 The partially ordered monoid α^* generated by $\alpha = \{a,b\}$ of **Table 1**: right multiplication table and partial order

Element	a	b	a	b	o	ab	e
a	*o*	*ab*	1	0	1	0	0
b	*ab*	*e*	0	1	1	0	0
o	*o*	*o*	0	0	1	0	0
ab	*o*	*a*	0	0	1	1	0
e	*a*	*b*	0	0	1	0	1

Table 4 The semiring α^* generated by $\alpha = \{a,b\}$ of **Table 1**: right multiplication table and partial order

Element	a	b	*Partial order*
a	*o*	*ab*	1010000000000000
b	*ab*	*e*	0110000000000000
o	*o*	*o*	0010000000000000
ab	*o*	*a*	0011000000000000
e	*a*	*b*	0010100000000000
$a + b$	*ab*	$ab + e$	1110010000000000
$a + ab$	*o*	$a + ab$	1011001000000000
$a + e$	*a*	$b + ab$	1010100100000000
$b + ab$	*ab*	$a + e$	0111000010000000
$b + e$	$a + ab$	$b + e$	0110100001000000
$ab + e$	*a*	$a + b$	0011100000100000
$a + b + ab$	*ab*	$a + ab + e$	1111011010010000
$a + b + e$	$a + ab$	$b + ab + e$	1110110101001000
$a + ab + e$	*a*	$a + b + ab$	1011101100100100
$b + ab + e$	$a + ab$	$a + b + e$	0111100011100010
$a + b + ab + e$	$a + ab$	$a + b + ab + e$	1111111111111111

then the semigroup $(\alpha')^*$ is a *homomorphic image* of α^*, if there is a mapping φ from α^* to $(\alpha')^*$ such that

1. for each $b \in (\alpha')^*$, there is some $a \in \alpha^*$ such that $\varphi(a) = b$; and
2. $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in \alpha^*$.

In this case, the mapping φ is said to be a *homomorphism*.

Interest is generally focused on those homomorphisms that are *generator-preserving*, that is, for which $\varphi(a_m) = a'_m$ for each generator a_m. If one semigroup is a (generator-preserving) homomorphic image of a second, it has the same structure as a partition of the second, and can therefore be seen as having a similar but 'coarser' structure than the first. Boorman and White (1976) advocated the use of the joint homomorphic image of the semigroups of two multiple social networks with the same generator types as a useful descriptor of shared structure, whereas Bonacich (1980) and Bonacich and McConaghy (1979) proposed the semigroup of the disjoint union of the two multiple social networks (for further discussion of the distinctive rationales of these two approaches, see Pattison, 1981).

A third application is to use structural analysis at the algebraic level to determine analysis at the network level, an approach pursued first by Boyd for group representations of kinship systems and later by others for more general semigroup representations of social networks (Bonacich, 1983; Boyd, 1969; Kim and Roush, 1984; Pattison and Bartlett, 1982; Pattison, 1982, 1993, 2009; White and Reitz, 1983). Structural analysis at the algebraic level generally entails identification of homomorphic image algebras in these approaches, and Pattison and Bartlett (1982; also Pattison, 1993) proposed analysis of an algebra into a set of maximally independent image algebras, termed *factors*. Pattison (1993) proposed a lattice-based algorithm for constructing the factors of a semigroup and an associated reduction of the underlying multiple social network. In many of these cases, use is also made of the natural partial order on α^*.

The Algebraic Structure of a Dynamic Multiple Social Network

The semigroup, monoid, and semiring representations described so far have entailed operations over networks in which ties take values from the set $\{0,1\}$, that is, in which ties are simply present or absent. In many social network applications of interest, however, network ties change over time. For example, two actors may be collaborators for a period of time, but then cease to collaborate; another two actors may have an antagonistic relationship, which later becomes friendship. Our second goal in this article is, therefore, to describe algebraic constructions for what we term *dynamic* multiple social networks. Following Kontoleon et al. (2013), we build on Moody's (2002) important insight that labeled walks of interest in dynamic multiple social networks must be *time-ordered* or *time-respecting* (see also Holme and Saramäki, 2011).

As for multiple social networks, we begin with a fixed primitive set of *relational types*, and a *network* associated with each relational type on a fixed set N of actors. Each network is now regarded as *dynamic* and the value of the tie $a(i,j)$ from actor i to actor j is a finite set of $n(a,i,j)$ strictly disjoint time intervals $\{[t_1,u_1], [t_2,u_2],\ldots, [t_{n(a,i,j)},u_{n(a,i,j)}]\}$ lying within some designated time period of interest, assumed without loss of generality, to be $[0,\omega]$. The time intervals $[t,u]$ are such that $t \leq u$, and if the tie is null, we define $n(a,i,j) = 0$ and write $a(i,j) = \varnothing$. (We choose to allow $t \leq u$ in order to accommodate discrete-time observation schemes.)

We note that for any arbitrary finite set of time intervals contained within $[0,\omega]$, there is a unique expression for the set of intervals in a form in which each pair of intervals is *strictly disjoint* (that is, for two time intervals $[t,u]$ and $[v,w]$, either $t \leq u < v \leq w$ or $v \leq w < t \leq u$); we call this form *canonical* and assume its use throughout. We refer to $a = [a(i,j)]$ as a *dynamic social network* and to $\{a_1, a_2, \ldots, a_r\}$ as a *dynamic multiple social network*.

As in the static case, we regard $\{a_1, a_2, \ldots, a_r\}$ as a set of network generators and consider dynamic analogues of the operations of composition and Boolean addition. An example of a dynamic multiple network is presented in **Table 5**. In this example, $n(a,i,j) \leq 1$, and $n(b,i,j) \leq 1$ for each pair (i,j) of actors.

Following Moody (2002), we define a composition operation for two dynamic networks a and b, which ensures that the relation ab comprises intervals for a pair (i,j) of actors in which it would be possible for some entity (e.g., an idea, a disease, or resources) to originate at i and reach j by flowing from i to some third party k along a tie labeled A and then from k to j along a tie labeled B. This definition of composition gives rise to what

Table 5 A small hypothetical dynamic multiple social network $\alpha = \{a,b\}$ on node set $N = \{1,2,3,4\}$ and time interval [0,5]

	a				b			
1	∅	[4,5]	∅	∅	∅	∅	[0,1]	∅
2	∅	∅	∅	∅	∅	∅	∅	[0,1]
3	∅	∅	∅	[4,5]	[0,1]	∅	∅	∅
4	∅	∅	∅	∅	∅	[0,1]	∅	∅

Moody termed *time-ordered* paths or walks in the dynamic multiple social network. In a recent review of temporal networks, Holme and Saramäki (2011) observed that the concept of a *time-respecting path* has emerged in a number of fields. In fact, Moody's definition for a walk labeled *AB* from *i* to *j* yields the broadest time interval over which the flow of some entity originating at actor *i* can reach actor *j* by traveling along a walk of type *A* to an intermediary *k* and then a walk of type *B* from *k* to *j*.

Specifically, if we write

$$a(i,k) = \left\{[t_1,u_1],[t_2,u_2],\ldots,\left[t_{n(a,i,k)},u_{n(a,i,k)}\right]\right\}$$

and

$$b(k,j) = \left\{[v_1,w_1],[v_2,w_2],\ldots,\left[v_{n(b,k,j)},w_{n(b,k,j)}\right]\right\}$$

then the composition *ab* of the dynamic networks *a* and *b* is defined to have ties $ab(i,j)$ in which

$$ab(i,j) = \cup_{k\in N}\left(\cup_{p=1,n(a,i,k)}\cup_{q=1,n(b,k,j)}\left[\max(t_p,v_q),w_q\right]\right)$$
$$\text{if } a(i,k)\neq\varnothing \text{ and } b(k,j)\neq\varnothing \quad \text{for some } k,$$

and

$$\begin{aligned} ab(i,j) &= \varnothing \quad \text{if } a(i,k) = \varnothing \text{ or } b(k,j) \\ &= \varnothing \quad \text{for each } k\in\{1,2,\ldots,n\} \end{aligned}$$

The logic of this expression is that for each potential third party *k*, and for each of the constituent intervals $[t_p,u_p]$ for which *i* is linked to *k* by a relation of type *A* and $[v_q,w_q]$ for which *k* is linked by a relation of type *B* to *j*, there is a possibility of flow through actor *k* to actor *j* of some entity arising at actor *i* during the period $[\max(t_p,v_q),w_q]$. We need to consider this possibility not only for all third parties *k* but also for all pairs of intervals making up $a(i,k)$ and $b(k,j)$, hence the union over both *p* and *q* as well as *k* in the expression above. As noted above, we assume following computation that $ab(i,j)$ is reexpressed in canonical form.

If, for example, $a(i,k) = \{[t,u]\}$ and $a(k,j) = \{[v,w]\}$, then the broadest time period over which flow could occur from actor *i* to actor *j* via *k* on a path of ties labeled *A* is $[\max(t,v),w]$. Importantly, this interval is smaller than or equal to $[v,w]$ and is the segment of the time period over which actor *k* is linked to actor *j*, which follows some segment of the period during which actor *i* is connected to actor *k*.

It should be noted that this form of composition is appropriate for the consideration of labeled walks as potential conduits of entities that endure in time. It is arguably less appropriate for the consideration of potential flows that are subject to various forms of dissipation (e.g., through forgetting or loss of relevance, in the case of an idea; conversion to a nonusable form in the case of resources; or cure in the case of a disease). As a result, Kontoleon et al. (2013) also consider a more general form of composition for dynamic networks in which the potential longevity of any flows along labeled walks can be reflected in the composition operation. The approach of Kontoleon et al. reflects the need articulated by Borgatti (2005) for a variety of different approaches to representing and analyzing flows and potential flows in social networks.

It is not possible to define an identity relation for the composition operation in dynamic networks just defined, but a left identity can be constructed. Define a (maximal) *left identity* dynamic network for composition as the network *l* with entries:

$$\begin{aligned} I(i,j) &= \{[0,\omega]\} \quad \text{if } i = j; \\ I(i,j) &= \varnothing, \quad \text{otherwise.} \end{aligned}$$

It follows from the definition that $la = a$ holds for all dynamic networks *a*.

A natural addition operation can also be defined for any dynamic networks *a* and *b* with

$$a(i,j) = \left\{[t_1,u_1],[t_2,u_2],\ldots,\left[t_{n(a,i,j)},u_{n(a,i,j)}\right]\right\}$$

and

$$b(i,j) = \left\{[v_1,w_1],[v_2,w_2],\ldots,\left[v_{n(b,i,j)},w_{n(b,i,j)}\right]\right\}.$$

We define

$$\begin{aligned} (a+b)(i,j) = &\left\{[t_1,u_1],[t_2,u_2],\ldots,\left[t_{n(a,i,j)},u_{n(a,i,j)}\right]\right\} \\ &\cup\left\{[v_1,w_1],[v_2,w_2],\ldots,\left[v_{n(b,i,j)},w_{n(b,i,j)}\right]\right\} \end{aligned}$$

and note that the latter can be reexpressed in canonical form as

$$(a+b)(i,j) = \left\{[x_1,y_1],[x_2,y_2],\ldots,\left[x_{n(a+b,i,j)},y_{n(a+b,i,j)}\right]\right\}.$$

We also define a null dynamic network *o* on node set *N*, which serves as an identity for the union operation on dynamic networks on *N*. The dynamic network *o* with entries

$$o(i,j) = \varnothing \quad \text{for all } i,j\in N$$

is termed the *null* dynamic network.

A natural partial order can also be defined for dynamic networks. For dynamic networks *a* and *b* with

$$\begin{aligned} a(i,j) &= \left\{[t_1,u_1],[t_2,u_2],\ldots,\left[t_{n(a,i,j)},u_{n(a,i,j)}\right]\right\} \text{ and} \\ b(i,j) &= \left\{[v_1,w_1],[v_2,w_2],\ldots,\left[v_{n(b,i,j)},w_{n(b,i,j)}\right]\right\} \end{aligned}$$

define

$$a \le b$$

if and only if, for all *i* and *j*, and for each $p \in \{1,2,\ldots,n(a,i,j)\}$, there is some $q \in \{1,2,\ldots,n(b,i,j)\}$ such that

$$v_q \le t_p \le u_p \le W_q$$

As for the static case, the operations just defined give rise to algebraic structures on dynamic multiple networks.

Definition 7

A *left dioid* is a set S endowed with binary operations, $\circ$ and $+$, and a partial order $\leq$, satisfying the following axioms:

1. *commutativity of* $+$: $a + b = b + a$, for all $a, b \in S$;
2. monoid structure under $+$: $(S,+)$ is a monoid;
3. *semigroup structure under* $\circ$ with left identity: $(S,\circ)$ is a semigroup, with left identity l such that $la = a$ for all $a \in S$;
4. *partial order*: $\leq$ is a partial order on S;
5. *absorption*: $a \circ o = o \circ a = o$, for all $a \in S$; and
6. *left distributivity of* $\circ$ *over* $+$: $a \circ (b + c) = a \circ b + a \circ c$, for all $a, b, c \in$ S.

Left dioids are described in more detail in Gondran and Minoux (2008). It can readily be demonstrated that the closure of a dynamic multiple network is a left dioid under the composition and union operations defined above.

By way of illustration, we consider the hypothetical dynamic network $\alpha = \{a,b\}$ displayed in Table 5 and construct in Table 6 the distinct dynamic networks generated by α under time-ordered relational composition. Table 7 contains the partially ordered semigroup α^* generated by time-ordered composition. It can be seen that although the pattern of nonnull entries in the dynamic multiple network of Table 5 mirrors the pattern of zero entries in the multiple network of Table 1, the structure of the monoid generated by the multiple dynamic network in Table 7 is quite different to that in Table 3. For example, $ab = ba$ is an equation in the monoid generated by the multiple social network in Table 1, but ab and ba are distinct in the partially ordered semigroup generated by the dynamic multiple social network of Table 5. This difference reflects the time-ordered nature of composition in the dynamic network and illustrates the importance of the temporal properties of the generator networks in understanding the structure of labeled walks to which they give rise.

Kontoleon et al. (2013) show how the algebraic structure of dioids may be exploited to assist in computing time-ordered paths and labeled walks in a dynamic multiple network. They construct a dioid of endomorphisms acting on the set of time intervals that allows very general composition operations. They describe application of the generalized Jacobi and Gauss–Seidel algorithms (e.g., Gondran and Minoux, 2008) to compute reachability in a dynamic social network and adapt Brandes' (2001) betweenness algorithm to a dynamic context. As they argue, the capacity to use time-ordered walks in a dynamic network is important because it allows for the construction of measures that are sensitive to the full dynamic structure of the network, moving beyond those based on an ad hoc time-slicing of the network.

Table 6 Distinct elements other than *a* and *b* in the partially ordered semigroup generated by the dynamic multiple social network $\alpha = \{a,b\}$ of Table 6

	o				ba				bb			
1	∅	∅	∅	∅	∅	∅	∅	[4,5]	[0,1]	∅	∅	∅
2	∅	∅	∅	∅	∅	∅	∅	∅	∅	[0,1]	∅	∅
3	∅	∅	∅	∅	∅	[4,5]	∅	∅	∅	∅	[0,1]	∅
4	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	[0,1]

Table 7 The partially ordered semigroup α^* generated by $\alpha = \{a,b\}$ of Table 6: right multiplication table and partial order

Element	a	b	a	b	o	ba	bb
a	*o*	*o*	1	0	1	0	0
b	*ba*	*bb*	0	1	1	0	0
o	*o*	*o*	0	0	1	0	0
ba	*o*	*o*	0	0	1	1	0
bb	*a*	*b*	0	0	1	0	1

Conclusion

The algebraic structure of a multiple social network describes the relationships among labeled walks and their unions. In an analogous way, the algebraic structure of a multiple dynamic social network describes the relationships among labeled time-ordered walks and their unions. In both cases, the algebraic structure can serve one of several key purposes: (1) as the foundation for algorithm development and hence for the development of efficient computational approaches for measures in static or dynamic networks that depend on walks in those networks (e.g., Harary et al., 1965; Gondran and Minoux, 2008; Kontoleon et al., 2013); and (2) as a description of the relationship among labeled (and, where relevant, time-ordered) walks in the network, and hence as the basis for structural analysis and comparison of multiple networks (e.g., Pattison, 1993, 2009).

See also: Bayesian Graphical Models and Networks; Clustering and Cohesion in Networks: Concepts and Measures; Network Centrality, Measures of; Social Networks; Statistical Clustering.

Bibliography

Bonacich, P., 1980. The "common structure semigroup", a replacement for the Boorman and White "joint reduction". American Journal of Sociology 86, 159–166.
Bonacich, P., 1983. Representations for homomorphisms. Social Networks 5, 173–192.
Bonacich, P., McConaghy, M., 1979. The algebra of blockmodeling. In: Schuesler, K.F. (Ed.), Sociological Methodology 1980. Jossey-Bass, San Francisco, pp. 489–532.
Boorman, S.A., White, H.C., 1976. Social structures from multiple networks. II. Role structures. American Journal of Sociology 81, 1384–1446.
Borgatti, S., 2005. Centrality and network flow. Social Networks 27, 55–71.
Boyd, J.P., 1969. The algebra of group kinship. Journal of Mathematical Psychology 6, 139–167.
Boyd, J.P., 1989. Social semigroups and green relations. In: Freeman, L.C., White, D.R., Romney, A.K. (Eds.), Research Methods in Social Network Analysis. George Mason University Press, Fairfax, VA, pp. 215–254.
Boyd, J.P., 1990. Social Semigroups: A Unified Theory of Scaling and Blockmodelling as Applied to Social Networks. George Mason University Press, Fairfax, VA.
Boyd, J.P., Haehl, J.H., Sailer, L.D., 1989. Kinship systems and inverse semigroups. Journal of Mathematical Sociology 2, 37–61.
Brandes, U., 2001. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25, 163–177.
Breiger, R.L., Pattison, P.E., 1978. The joint role structure of two communities' elites. Sociological Methods and Research 7, 213–226.
Breiger, R.L., Pattison, P.E., 1986. Cumulated social roles: the duality of persons and their algebras. Social Networks 8, 215–256.

Gondran, M., Minoux, M., 2008. Graphs, Dioids and Semirings: New Models and Algorithms, Operations Research/Computer Science Interfaces. Springer, New York.

Harary, F., Norman, R., Cartwright, D., 1965. Structural Models: An Introduction to the Theory of Directed Graphs. Wiley, New York.

Holme, P., Saramäki, J., 8 August 2011. Temporal Networks. E-print: arXiv:1108.1780v1.

Kim, K.H., 1982. Boolean Matrix Theory and Applications. Marcel Dekker, New York.

Kim, K.H., Roush, F.W., 1984. Group relationships and homomorphisms of Boolean matrix semigroups. Journal of Mathematical Psychology 28, 448–452.

Kontoleon, N., Falzon, L., Pattison, P., 2013. Algebraic Structures for Dynamic Networks. Defence Science and Technology Organisation, Edinburgh, South Australia.

Lorrain, F., 1975. Reseaux sociaux et classifications sociales. Hermann, Paris.

Lorrain, F., White, H.C., 1971. Structural equivalence of individuals in social networks. Journal of Mathematical Sociology 1, 49–80.

Luce, R.D., 1956. A note on Boolean matrix theory. Proceedings of the Americal Mathematical Society 3, 382–388.

McConaghy, M., 1981. The common role structure: improved blockmodelling methods applied to two communities' elites. Sociological Methods and Research 9, 267–285.

Mandel, M., 1983. Local roles and social networks. American Sociological Review 48, 376–386.

Moody, J., 2002. The importance of relationship timing for diffusion. Social Forces 81, 25–56.

Pattison, P.E., 1981. Equating the 'joint reduction' with blockmodel common role structure: a reply to McConaghy. Sociological Methods and Research 9, 286–302.

Pattison, P.E., 1982. The analysis of semigroups of multirelational systems. Journal of Mathematical Psychology 25, 87–118.

Pattison, P.E., 1989. Mathematical models for local social networks. In: Keats, J., Taft, R., Heath, R., Lovibond, S. (Eds.), Mathematical and Theoretical Systems. North Holland, Amsterdam.

Pattison, P.E., 1993. Algebraic Models for Social Networks. Cambridge University Press, New York.

Pattison, P.E., 2009. Algebraic models for social networks. Entry 00199. In: Myers, R.A. (Ed.), Encyclopaedia of Complexity and System Science. Springer.

Pattison, P.E., 2011. Relation algebras and social networks. In: Scott, J., Carrington, P. (Eds.), The Sage Handbook of Social Network Analysis. Sage, London, pp. 447–458.

Pattison, P.E., Bartlett, W.K., 1982. A factorization procedure for finite algebras. Journal of Mathematical Psychology 25, 51–81.

White, D.R., Reitz, K.P., 1983. Graph and semigroup homomorphisms on networks of relations. Social Networks 5, 193–234.

White, H.C., 1963. An Anatomy of Kinship. Prentice-Hall, Englewood Cliffs, NJ.

White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks: I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–780.

Winship, C., Mandel, M., 1983. Roles and positions: a critique and extension of the blockmodelling approach. In: Leinhardt, S. (Ed.), Sociological Methodology 1983–1984. Jossey-Bass, San Francisco, pp. 314–344.

Wu, L., 1983. Local blockmodel algebras for analyzing social networks. In: Leinhardt, S. (Ed.), Sociological Methodology 1983–1984. Jossey-Bass, San Francisco, pp. 272–313.

Social Neuroscience of the Twenty-First Century

Elizabeth A Necka, Stephanie Cacioppo, and John T Cacioppo, The University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social species are so characterized because they form organizations that extend beyond the individual. The goal of social neuroscience is to investigate the biological mechanisms that underlie these social structures, processes, and behaviors and the influences between social and neural structures and processes. Such an endeavor is challenging because it necessitates the integration of multiple levels. Mapping across systems and levels (from genome to social groups and cultures) requires interdisciplinary expertise, comparative studies, innovative methods, and integrative conceptual analysis.

Social Neuroscience

Social neuroscience brings together a broad variety of scientists, disciplines, and methodologies dedicated to investigating the biological mechanisms of social interaction, and thus represents an interdisciplinary scientific field that dives into the hard problem of the mutual influence of biology and social mechanisms (Cacioppo and Berntson, 1992, 2001) – a problem described by Frith and Wolpert (2004) as one of the major problems for the neurosciences to address in the twenty-first century. Based on the premise that all social behaviors are implemented biologically, this exciting and mature academic field seeks to understand (1) how biological systems implement social processes and behaviors; (2) how social structures that range from dyads, families, neighborhoods, and groups to cities, civilizations, and international alliances impact the brain and biology; and (3) how social and neural structures interact to produce these outcomes. In other words, social neuroscience investigates the interaction between social factors and the central nervous system, their manifestations, and potentially reciprocal influences across levels, from molecules to societies.

Social neuroscience and cognitive neuroscience both focus on the relationship between behavior and brain networks, but the basic perspective of each is distinct and complementary. Cognitive neuroscience views the human brain from the perspective of a solitary computing device capable of complex analyses and computations. Accordingly, attention, perception, memory, and language as an internal representational system are among the topics of interest. Social neuroscience, in contrast, views the human brain as a mobile, broadband computing device connected to others. Among the topics of interest are social learning, hierarchies, imitation, conformity, attraction, empathy, and language as a system for communication with others. According to the social brain hypothesis, the social complexities and demands of primate species have contributed to the rapid increase in neocortical connectivity and intelligence. Consistent with this hypothesis, the evolution of large and metabolically expensive brains in primates is more closely associated with social than ecological complexity (Dunbar and Shultz, 2007; Ott and Rogers, 2010), and the volume of some brain areas, such as the amygdala, correlate with the size and complexity of a person's social networks (Bickart et al., 2011).

Anecdotal evidence for the importance of both social and biological factors on outcomes such as health and well-being has existed for centuries, as the influence of social support on health was apparent even to the ancient Greeks. Hippocrates (460–370 BC) frequently prescribed 'association with friends' as a way to restore the body to its natural state of harmony and cure illness (Hothersall, 2004). In 1848, the case report of Phineas Gage, an American railroad construction foreman, documented his putative changes in personality and social behavior following damage to the ventromedial prefrontal cortex – changes that were sufficiently dramatic that his friends and acquaintances observed he was 'no longer Gage' (Damasio, 1996; Harlow, 1848; Macmillan, 2008).

Empirical evidence for the value of integrating social and biological levels of analysis accrued in animal (e.g., Harlow et al., 1965) and human (e.g., Cacioppo and Petty, 1979; Cacioppo et al., 1982) research, but the emphasis on disciplinary differentiation and specialization presented an active barrier to working across these levels. Most social scientists regarded animal and biological research as having little if anything to contribute to an understanding of suffering and problems that plagued industrialized societies (e.g., Allport, 1947; for a review, see Cacioppo, 2002). On the other hand, most biological scientists believed that social factors were irrelevant to understanding basic biological structures or functions and, even if relevant, were too complicated to be understood in the foreseeable future (e.g., Scott, 1991). It was in this context that Cacioppo and Berntson (1992) proposed the doctrine of multilevel analyses (see below) and the perspective of 'social neuroscience' as a multilevel, integrative field.

In research that followed, social species were found to fare poorly when forced to live solitary lives (Cacioppo et al., 2011). Humans, born to the longest period of abject dependency of any species and dependent on conspecifics across the lifespan to survive and prosper, do not fare well whether they are living solitary lives or whether they simply perceive that they live in isolation (e.g., Luo et al., 2012; Perissinotto et al., 2012). In an illustrative study, Caspi et al. (2006) found that perceived social isolation in adolescence and young adulthood predicted how many cardiovascular risk factors (e.g., body mass index, waist circumference, blood pressure, and cholesterol) were elevated in young adulthood, and that the number of developmental occasions (i.e., childhood,

adolescence, and young adulthood) at which participants were lonely predicted the number of elevated risk factors in young adulthood. Perceived isolation has also been associated with the progression of Alzheimer's disease (Wilson et al., 2007), obesity (Lauder et al., 2006), increased vascular resistance (Cacioppo et al., 2002a), elevated blood pressure (Cacioppo et al., 2002a; Hawkley et al., 2006), increased hypothalamic pituitary adrenocortical activity (Adam et al., 2006; Steptoe et al., 2004), less salubrious sleep (Cacioppo et al., 2002b; Pressman et al., 2005), diminished immunity (Kiecolt-Glaser et al., 1984; Pressman et al., 2005), reduction in independent living (Russell et al., 1997; Tilvis et al., 2000), alcoholism (Akerlind and Hörnquist, 1992), depressive symptomatology (Cacioppo et al., 2006; Heikkinen and Kauppinen, 2004), suicidal ideation and behavior (Rudatsikira et al., 2007), and altered gene expression including the underexpression of genes bearing anti-inflammatory glucocorticoid response elements (GREs) and overexpression of genes bearing response elements for proinflammatory nuclear factor κB/Rel transcription factors (Cole et al., 2007, 2011). Nonhuman social species from fruit flies to apes also show similar deleterious biological effects of social isolation (e.g., Makinodan et al., 2012; Ruan and Wu, 2008).

The questions posed by the social neuroscience perspective, methodological, and quantitative developments, and the promise of more comprehensive theories inspired a new generation of scientists to pursue a more complex picture of behavior and biological functioning. By its 20th anniversary, social neuroscience had become an active field of research across the globe, fueled by the establishment of societies and journals to advance and foster scientific research, education, and clinical applications.

Doctrine of Multilevel Analyses

Social species are so characterized because they form organizations that extend beyond the individual. The goal of social neuroscience is to investigate the biological mechanisms that underlie these social structures, processes, and behaviors and the influences between social and neural structures and processes. Such an endeavor is challenging because it necessitates the integration of multiple levels. Mapping across systems and levels (from genome to social groups and cultures) requires interdisciplinary expertise, comparative studies, innovative methods, and integrative conceptual analysis (Cacioppo and Decety, 2011). Cacioppo and Berntson's (1992) doctrine of multilevel analyses emphasized the importance of using multilevel analyses to test hypotheses across their different levels of organization and specified three illustrative principles one can use to investigate key questions in social neuroscience along the continuum of organizational levels.

The principle of multiple determinism specifies that behaviors can have multiple antecedents within or across levels of organization. For instance, although immune response was once thought to reflect only physiological responses to pathogens or tissue damage, a more complete understanding of immunity has led to demonstrations of how a person's perceptions of his or her close personal relationships may impact inflammation and immunity (e.g., Kiecolt-Glaser et al., 2010). Psychosocial stress, operating through the brain's perception of the meaning of events, can also increase proinflammatory cytokine production in the absence of infection or injury. Animal research has revealed related findings in mice: exposure of mice to 2 weeks of isolation enhances tumor liver metastasis in part via its suppressive effect on the immune system of the host (Wu et al., 2000). One implication of this principle is that comprehensive theories of social phenomena require a consideration of multiple factors from various levels of organization – e.g., from the biological and individual's level to the social level (Cacioppo and Ortigue, 2011). A second implication is that many-to-many mappings between elements across proximal levels of organization become increasingly complex as the number of intervening levels of organization increases. Accordingly, the articulation of these mappings and mechanisms underlying a given multilevel observation may be simpler when working across proximal rather than distal levels of organization.

The principle of nonadditive determinism specifies that properties of the whole are not always readily predictable by the simple sum of the (initially recognized) properties of the parts (Cacioppo and Berntson, 1992). For instance, the behavior of nonhuman primates was examined following the administration of amphetamine or placebo (Haber and Barchas, 1983). No clear pattern emerged until each primate's position in the social hierarchy was considered. When this social factor was taken into account, amphetamines were found to increase dominant behavior in primates high in the social hierarchy and to increase submissive behavior in primates low in the social hierarchy. A strictly biological (or social) analysis, regardless of the sophistication of the measurement technology, may not have unraveled the orderly relationship that existed.

Finally, the principle of reciprocal determinism specifies that there can be mutual influences among biological and social factors in determining behavior (Cacioppo and Berntson, 1992). For example, maternal behavior can alter expression of genes through a process of DNA methylation, and genes altered in this way then can affect subsequent maternal behavior (Zhang and Meaney, 2010). One important implication is that comprehensive accounts of human behavior cannot be achieved if the biological, cognitive, or social levels of organization are considered unnecessary or irrelevant.

Social neuroscience also flourishes because of an increasing number of sophisticated methods and techniques that have been developed in the past two decades. This point is discussed next.

Techniques Used in Social Neuroscience

To investigate the mutual influence of biological and social mechanisms, social neuroscientists, ranging from physicists to psychologists, epidemiologists to neurologists, philosophers to neurobiologists, and entomologists to zoologists, have begun to work together in interdisciplinary scientific teams using animal models, patients studies, and research on healthy individuals. These interdisciplinary collaborations have capitalized on a variety of methods and techniques ranging from

behavioral studies of implicit processes in lesion and split-brain patients to volumetric and neuroimaging studies across scales of neural organization in chimpanzees or healthy humans to cellular and molecular techniques in genetics and epigenetics. Even developed techniques such as meta-analyses and electroencephalography have seen upgrades that, for instance, permit investigations of the source and chronoarchitecture of the neural substrates of social processes (e.g., He et al., 2009; Ito et al., 2004; Ortigue et al., 2010). Importantly, the development of experimental manipulations of neural processes in humans through, for instance, the use of transcranial magnetic stimulation and neurotransmitter agonists and antagonists has also helped determine the causal significance of specific neural regions in social cognition, emotion, and behavior. Finally, increases in computational speed and capacities are increasingly simplifying the problem of addressing questions across levels of organization that involve large datasets and/or previously computationally prohibitive simulations or analyses.

Prospects

Interdisciplinary collaborations are a cornerstone of social neuroscience endeavors. A decade ago, the field was characterized by some as correlating social and cognitive functions with regions of brain activation using functional neuroimaging. The field has always had a broader foundation, however, and social neuroscience is now recognized as drawing on research on animals and humans, with phenomena at multiple levels of organization measured and/or manipulated to go beyond the specification of association across levels to specify the mechanisms operating at and between each of the levels of organization pertinent to an association. The manifold disciplinary expertise for such multilevel investigations is increasingly beyond the reach of individual scholars, so interdisciplinary teams of scientists are increasingly common.

Furthermore, the twenty-first century presents its own unique array of questions. For instance, the development and accessibility of the Internet have transformed major aspects of the social environment, including how and where people study, meet, shop, and interact. Understanding how online interactions are similar to and different from face-to-face social interactions and how they might best be performed to benefit users across the lifespan is of particular relevance to a world growing more and more dependent upon the Internet and social networking sites for social interactions. And the century is still young. What the short history of social neuroscience has shown is that the number of social neuroscience papers that appeared in multidisciplinary science journals nearly doubled in the first decade of the field's existence and that the field has attracted some of the best and the brightest scientists across a wide range of scientific disciplines, with high rates of citation of social neuroscience articles in their dominant disciplinary journals (Matsuall et al., 2011). By bridging the gap between animal and human studies, social neuroscience has contributed to the understanding of the mechanisms by which the social world (and its disorders) impacts health, well-being, and longevity. The successes that social neuroscience has already met in its first two decades of existence suggest a promising future as it opens a critical avenue for a better understanding of the biology of connected minds. A multilevel integration of the social, biological, and cognitive factors known to determine behavior could also provide scientists and practitioners with new therapeutic interventions to address acute and chronic individual, communicative, and social disorders such as autism, psychopathy, and social phobias (e.g., see reviews by Cacioppo et al., 2007; Cacioppo and Ortigue, 2011). The road ahead is not only replete with conceptual challenges and methodological issues, but also promises exciting scientific discoveries. In short, the twenty-first century is an exciting time in which to be a social neuroscientist.

See also: Bayesian Models in Neuroscience; Cerebral Cortex; Cognitive Neuroscience; Electroencephalography: Basic Principles and Applications; Embodiment Theory; Emotion, Neural Basis of; Empathy: A Social Neuroscience Perspective; Functional Brain Imaging of Language Processes; Neural Foundation of Morality; Neuroeconomics; Neuroesthetics; Neuroethics: Ethics of Science and Science of Ethics; Neuromarketing; Neuroscience of Education; Placebo Effect; Psychopathy; Self and Brain.

Bibliography

Adam, E.K., Hawkley, L.C., Kudielka, B.M., Cacioppo, J.T., 2006. Day-to-day dynamics of experience – cortisol associations in a population-based sample of older adults. Proceedings of the National Academy of Sciences of the United States of America 103, 17058–17063.

Akerlind, I., Hörnquist, J.O., 1992. Loneliness and alcohol abuse: a review of evidences of an interplay. Social Science & Medicine 34, 405–414.

Allport, G., 1947. Scientific models and human morals. Psychological Review 54, 182–192.

Bickart, K.C., Wright, C.I., Dautoff, R.J., Dickerson, B.C., Feldman Barrett, L., 2011. Amygdala volume and social network size in humans. Nature Neuroscience 14, 163–164.

Cacioppo, J.T., 2002. Social neuroscience: understanding the pieces fosters understanding the whole and vice versa. American Psychologist 57, 819–830.

Cacioppo, J.T., Amaral, D.G., Blanchard, J.J., Cameron, J.L., Carter, C.S., Crews, D., Fiske, S., Heatherton, T., Johnson, M.K., Kozak, M.J., Levenson, R.W., Lord, C., Miller, E.K., Ochsner, K., Raichle, M.E., Shea, M.T., Taylor, S.E., Young, L.J., Quinn, K.J., 2007. Social neuroscience: progress and implications for mental health. Perspectives on Psychological Science 2, 99–123.

Cacioppo, J.T., Berntson, G.G., 1992. Social psychological contributions to the decade of the brain: doctrine of multilevel analysis. American Psychologist 47, 1019–1028.

Cacioppo, J.T., Berntson, G.G., 2001. Social neuroscience. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Pergamon Press, Oxford, pp. 14388–14391.

Cacioppo, J.T., Decety, J., 2011. Social neuroscience: challenges and opportunities in the study of complex behavior. Annals of the New York Academy of Sciences 1224, 162–173.

Cacioppo, J.T., Hawkley, L.C., Crawford, L.E., Ernst, J.M., Burleson, M.H., Kowalewski, R.B., Malarkey, W.B., Van Cauter, E., Berntson, G.G., 2002a. Loneliness and health: potential mechanisms. Psychosomatic Medicine 64, 407–417.

Cacioppo, J.T., Hawkley, L.C., Berntson, G.G., Ernst, J.M., Gibbs, A.C., Stickgold, R., Hobson, J.A., 2002b. Lonely days invade the nights: social modulation of sleep efficiency. Psychological Science 13, 384–387.

Cacioppo, J.T., Hawkley, L.C., Norman, G.J., Berntson, G.G., 2011. Social isolation. Annals of the New York Academy of Sciences 1231, 17–22.

Cacioppo, J.T., Hughes, M.E., Waite, L.J., Hawkley, L.C., Thisted, R.A., 2006. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychology and Aging 21, 140–151.

Cacioppo, J.T., Ortigue, S., 2011. Social neuroscience: how a multidisciplinary field is uncovering the biology of human interactions. Cerebrum. http://www.dana.org/news/cerebrum (accessed 17.12.12).

Cacioppo, J.T., Petty, R.E., 1979. Attitudes and cognitive response: an electrophysiological approach. Journal of Personality and Social Psychology 37, 2181–2199.

Cacioppo, J.T., Petty, R.E., Quintanar, L.R., 1982. Individual differences in relative hemispheric alpha abundance and cognitive responses to persuasive communications. Journal of Personality and Social Psychology 43, 623–636.

Caspi, A., Harrington, H., Moffitt, T.E., Milne, B.J., Poulton, R., 2006. Socially isolated children 20 years later: risk of cardiovascular disease. Archives of Pediatrics and Adolescent Medicine 160, 805–811.

Cole, S.W., Hawkley, L.C., Arevalo, J.M.G., Cacioppo, J.T., 2011. Transcript origin analysis identifies antigen presenting cells as primary targets of socially regulated leukocyte gene expression. Proceedings of the National Academy of Sciences of the United States of America 108, 3080–3085.

Cole, S.W., Hawkley, L.C., Arevalo, J.M., Sung, C.Y., Rose, R.M., Cacioppo, J.T., 2007. Social regulation of gene expression in human leukocytes. Genome Biology 8, R189.

Damasio, A.R., 1996. The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society B: Biological Sciences 351, 1413–1420.

Dunbar, R.I.M., Shultz, S., 2007. Evolution in the social brain. Science 317, 1344–1347.

Fritz, C.D., Wolpert, D., 2004. The Neuroscience of Social Interaction: Decoding, Influencing, and Imitating the Actions of Others. Oxford University Press, Oxford.

Haber, S.N., Barchas, P.R., 1983. The regulatory effect of social rank on behavior after amphetamine administration. In: Barchas, P.R. (Ed.), Social Hierarchies: Essays toward a Socio-physiological Perspective. Greenwood, Westport, CT, pp. 119–132.

Harlow, H.F., Dodsworth, R.O., Harlow, M.K., 1965. Total social isolation in monkeys. Proceedings of the National Academy of Sciences of the United States of America 54, 90–97.

Harlow, J.M., 1848. Passage of an iron bar through the head. Boston Medical and Surgical Journal 39, 389–393.

Hawkley, L.C., Masi, C.M., Berry, J.D., Cacioppo, J.T., 2006. Loneliness is a unique predictor of age-related differences in systolic blood pressure. Psychology and Aging 21, 152–164.

He, Y., Johnson, M.K., Dovidio, J.F., McCarthy, G., 2009. The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. Social Neuroscience 4, 426–442.

Heikkinen, R., Kauppinen, M., 2004. Depressive symptoms in late life: a 10-year follow-up. Archives of Gerontology and Geriatrics 38, 239–250.

Hothersall, D., 2004. History of Psychology, fourth ed. McGraw Hill, Boston.

Ito, T., Thompson, E., Cacioppo, J.T., 2004. Tracking the time course of social perception: the effects of racial cues on event-related brain potentials. Personality and Social Psychology Bulletin 30, 1267–1280.

Kiecolt-Glaser, J.K., Garner, W., Speicher, C.E., Penn, G.M., Holliday, J., Glaser, R., 1984. Psychosocial modifiers of immunocompetence in medical students. Psychosomatic Medicine 46, 7–14.

Kiecolt-Glaser, J.K., Gouin, J.P., Hantsoo, L.V., 2010. Close relationships, inflammation, and health. Neuroscience and Biobehavioral Reviews 35, 33–38.

Lauder, W., Mummery, K., Jones, M., Caperchione, C., 2006. A comparison of health behaviors in lonely and non-lonely populations. Psychology, Health, and Medicine 11, 233–245.

Luo, Y., Hawkley, L.C., Waite, L., Cacioppo, J.T., 2012. Loneliness, health, and mortality in old age: a national longitudinal study. Social Science and Medicine 74, 907–914.

Makinodan, M., Rosen, K.M., Ito, S., Corfas, G., 2012. A critical period for social experience–dependent oligodendrocyte maturation and myelination. Science 337, 1357–1360.

Matsuall, S., Christen, M., Kaufman, I., 2011. Historical perspectives on social neuroscience. In: Decety, J., Cacioppo, J. (Eds.), The Oxford Handbook of Social Neuroscience. Oxford University Press, Oxford, pp. 9–27.

Macmillan, M., 2008. Phineas Gage: unravelling the myth. The Psychologist 21, 828–839.

Ortigue, S., Bianchi-Demichelli, F., Patel, N., Frum, C., Lewis, J.W., 2010. Neuroimaging of love: fMRI meta-analysis evidence toward new perspectives in sexual medicine. The Journal of Sexual Medicine 7, 3541–3552.

Ott, S.R., Rogers, S.M., 2010. Gregarious desert locusts have substantially larger brains with altered proportions compared with the solitarious phase. Proceedings of the Royal Society B: Biological Sciences 277, 3087–3096.

Perissinotto, C.M., Stijacic Cenzer, I., Covinsky, K.E., 2012. Loneliness in older persons: a predictor of functional decline and death. Archives of Internal Medicine 172, 1078–1083.

Pressman, S.D., Cohen, S., Miller, G.E., Barkin, A., Rabin, B.S., Treanor, J.J., 2005. Loneliness, social network size, and immune response to influenza vaccination in college freshmen. Health Psychology 24, 297–306.

Ruan, H., Wu, C.F., 2008. Social interaction-mediated lifespan extension of *Drosophila* Cu/Zn superoxide dismutase mutants. Proceedings of the National Academy of Sciences of the United States of America 105, 7506–7510.

Rudatsikira, E., Muula, A.S., Siziya, S., Twa-Twa, J., 2007. Suicidal ideation and associated factors among school-aged adolescents in rural Uganda. BMC Psychiatry 7, 67.

Russell, D.W., Cutrona, C.E., de la Mora, A., Wallace, R.B., 1997. Loneliness and nursing home admission among rural older adults. Psychology and Aging 12, 574–589.

Steptoe, A., Owen, N., Kunz-Ebrecht, S.R., Brydon, L., 2004. Loneliness and neuroendocrine, cardiovascular, and inflammatory stress responses in middle-aged men and women. Psychoneuroendocrinology 29, 593–611.

Scott, T.R., 1991. A personal view on the future of psychology departments. American Psychologist 46, 975–976.

Tilvis, R.S., Pitkala, K.H., Jolkkonen, J., Strandberg, T.E., 2000. Social networks and dementia. Lancet 356, 77–78.

Wilson, R.S., Krueger, K.R., Arnold, S.E., Schneider, J.A., Kelly, J.F., Barnes, L.L., Tang, Y., Bennett, D.A., 2007. Loneliness and risk of Alzheimer's disease. Archives of General Psychiatry 64, 234–240.

Wu, W., Yamamura, T., Murakami, K., Murata, J., Matsumoto, K., Watanabe, H., Saiki, I., 2000. Social isolation stress enhanced liver metastasis of murine colon 26-L5 carcinoma cells by suppressing immune responses in mice. Life Sciences 66, 1827–1838.

Zhang, T.Y., Meaney, M.J., 2010. Epigenetics and the environmental regulation of the genome and its function. Annual Review of Psychology 61, 439–466.

Relevant Websites

http://psychology.uchicago.edu/people/faculty/cacioppo – The University of Chicago, Department of Psychology.

http://www.s4sn.org/ – Society for Social Neuroscience.

https://hpenlaboratory.uchicago.edu – The University of Chicago High Performance Electrical NeuroImaging Laboratory.

Social Ontology

John Latsis, University of Reading, Reading, UK; and University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social ontology is the study of the nature of social being; those characteristics of reality that are necessarily dependent on human social interaction. There are two general approaches to social ontology in the literature. The first focuses on the ontological elaboration of theoretical concepts or assumptions of a particular author, theory, or approach. The second articulates a general conception of social ontology in which to anchor current or future theoretical and empirical research programs. This article introduces some major contributions to both approaches and discusses differences between them.

What is Social Ontology?

Ontology is the study of the nature of being. Social ontology is the study of the nature of social being; those characteristics of reality that are necessarily dependent on human social interaction. One prominent justification for the study of social ontology is that the social sciences have regularly imported methods of analysis from the natural sciences or mathematics without considering whether such methods are well suited to their object of study. A deeper analysis of that object has the potential to inform both methodological reflection and theory development within the social sciences.

Use of the term social ontology has increased markedly since the turn of the twenty-first century across the social sciences. In some disciplines including sociology, organization studies, and international relations, social ontology is sufficiently prominent to be invoked in key theoretical and meta-theoretical debates. However, the proliferation of the jargon associated with social ontology does not automatically indicate a coherent research paradigm: there is diversity both within and between the dominant positions in the field.

Even the term 'social ontology' contains a fundamental ambiguity which requires further clarification. Not only does it refer to the study of the nature of social reality, but it also refers to a specific conception of social reality. Thus a particular author or school of thought can be said to articulate, and defend or criticize 'a social ontology.' This ambiguity has led to a distinction within the literature between work that focuses on the ontological elaboration of theoretical concepts or assumptions of a particular author, theory or approach, and scholarship whose goal is to articulate a general conception of social ontology in which to anchor current or future theoretical and empirical research programs. In the remainder of this contribution, I shall refer to the first approach as 'scientific ontology' and the second as 'philosophical ontology,' though there are several alternative classifications in the literature. These two approaches will be discussed in greater depth in Sections Scientific Ontology and Philosophical Ontology respectively.

As suggested by the widespread use of philosophical jargon and the investigation of explicitly philosophical themes by social ontologists, there is a clear connection between both of these approaches and the philosophical literature. In particular, both approaches to ontological research can be seen as an outgrowth of scientific realism in analytic philosophy. Starting in the 1970s, realists provided a critical impetus by challenging logical positivism and empiricism in the philosophy of science, by implication threatening the dominance of positivist-inspired philosophy of social science. More importantly perhaps, realists also provided a general philosophical approach that retained a commitment to causal explanation as the cornerstone of the scientific endeavor – a position that hermeneutic approaches to the social sciences frequently rejected.

It is nevertheless entirely inaccurate to view social ontology as the top-down imposition of philosophical concepts and categories onto the social sciences. In fact, the current literature is dominated by practicing social scientists rather than professional philosophers, and contributions typically reflect the preoccupations of the fields or disciplines to which they contribute. So, for example, the economist Tony Lawson (1997) uncovers the ontological presuppositions implicit in the mathematization of economics, whilst the sociologist Margaret Archer (1995) addresses those of the structure–agency debate in social theory. Indeed, social ontology has recently been described as 'philosophy for social science,' rather than the philosophy of social science (Lawson et al., 2007: pp. 1–14).

Scientific Ontology

The main objective of the scientific ontology literature is to identify and elaborate the ontological presuppositions of social scientific concepts and categories. In so doing, social ontologists echo Willard Van Orman Quine's view that theory construction necessarily involves ontic commitments and that the analysis of these commitments is a legitimate domain of philosophical enquiry (Quine, 1948). Scientific ontologists need not, however, commit themselves to any particular worldview from which to judge or criticize their targets.

Scholars engaged in scientific ontology are usually not system builders, rather they see themselves as contributing to a better understanding of the basic categories of social analysis. There are, however, at least two different ways in which scientific ontology can engage with these categories:

1. By teasing out the implicit ontological commitments of a particular author, model, theory, or school of thought.

2. By engaging in the ontological elaboration or exploration of a social scientific concept or category, using extant theory as a reference or starting point.

The first strand is principally descriptive, whilst the second engages more directly in theoretical debates within the social sciences. Some influential contributors to the field clearly privilege the first strand, stressing their role as critically reflective interpreters of disciplinary discourse rather than direct participants in theoretical debates (Maki, 2001: pp. 1–14). Others argue more forcefully for the importance of the second strand (Pratten, 2007).

The purpose of teasing out the implicit ontological commitments of an author's contribution is particularly important in social scientific disciplines where those authors are the figureheads of schools of thought. Studies of this first type can also aim to elaborate or complete an ongoing exercise in the history of ideas by focusing attention on the ontological aspects of a well-known and influential contribution (Fleetwood, 2002). Nevertheless, trying to identify the ontological commitments of an (sometimes long dead) author is an extremely difficult task that will almost inevitably produce partial and tentative conclusions.

By comparison, analysis of the presuppositions of specific theories or models promises more decisive results, so long as the target theory is well articulated. The general strategy is to ask some variant of the following question: "What would the world have to look like for theory x to be true?" This is the strategy adopted by Lawson in his discussion of modeling in mainstream economics (Lawson, 1997: pp. 3–25). He begins by noting that mainstream economics endorses a conception of science that sees its objective as the production of lawlike generalizations of the form "whenever event x then event y." He goes on to demonstrate how the methods used by economists presuppose a world populated by isolated atoms interacting in what he calls a closed system (a system in which the empirical regularities required for these laws to be true exist). This ontological analysis of the presuppositions of current mainstream modeling practice is the foundation of his subsequent critique of the discipline. Scientific ontology can also be used to facilitate internal critique. This is the strategy adopted by Archer in her analysis of Anthony Giddens' 'structuration theory' (Archer, 1995: pp. 93–134). She argues that Giddens' elision of the categories of structure and agency – itself an attempt to avoid traditional reified conceptions of social structure – leads him to adopt a flat ontology of social practices. This ontology can no longer accommodate an account of the interaction between structural factors and individual action, which is seen as a central concern of social theory.

The second strand of scientific ontology usually starts with extant social scientific theory as well, but its objectives are somewhat different. Rather than uncovering the presuppositions of an author or theory, this strand of ontological research seeks to elaborate important categories and concepts. Thus, social ontologists have contributed to discussions of the nature of institutions (Lawson, 2003), rules and conventions (Latsis, 2005; Al-Amoudi, 2010) and technology (Faulkner and Runde, 2009; Lawson, 2009) amongst many other topics. Contributors to these debates have engaged directly with social–scientific research by pointing out weaknesses, tensions, or incoherence implicit in the use of these concepts and attempting to resolve them, thus providing the potential foundations for new theories.

The two strands of scientific ontology, as I have described them, overlap in many contributions but there is a major difference between them. A premise of the first strand of scientific ontology is that the practices of scientists in their target field or discipline are worthy of detailed scrutiny, though in some cases this scrutiny may be justified by institutional and sociological factors rather than epistemological ones. There is therefore a tendency in some parts of the literature to take the fact of knowledge production within the social sciences at face value. In the second strand, reference to the social scientific literature acts as the entry point into a more sustained analysis of social scientific concepts, which attempts to go beyond current disciplinary strictures and reorient social theory. Contributors to the literature on philosophical ontology have even more comprehensive and radical ambitions.

Philosophical Ontology

The project of philosophical ontology may be contingently related to scientific ontology by the fact that leading scholars usually engage with the work of others prior to the development of their own framework, but authors engaged in what I have called philosophical ontology are ultimately best thought of as system-builders. In other words, the intention of contributors to this literature is to develop and articulate conceptions of social being that constitute better, or more complete accounts of the social world than incumbent or competing views. This is a more ambitious project than the research that I have described under the heading of scientific ontology in Section Scientific Ontology and it is generally regarded as being more controversial. Moreover, proponents of philosophical ontology have readily engaged in critical debates within the social scientific disciplines, using their ontological perspective to inform and shape their critiques of extant theoretical contributions. Thus philosophical ontology is a critical, revisionary, and interactive project within the social sciences.

The contributions of Roy Bhaskar and John Searle stand out for their level of detailed elaboration and subsequent influence within the social sciences. Both perspectives are informed by philosophy, but one (Bhaskar's) emanates directly from the problems of the social sciences, whereas the other (Searle's) is principally motivated by problems in the philosophy of mind and language. I shall now discuss each briefly in turn before briefly describing a potentially significant difference between them in Section Naturalism and Emergence.

Starting from insights that were consistent with developments in contemporary philosophy of science, Bhaskar (1975) argued for a realist conception of science and developed the position known as critical realism in the social sciences (check cross-reference for CR). His general conception is notable for its reliance on a transcendental argument to move from the observation of regularities in events (surface phenomena) to positing the existence of the deeper, unobservable mechanisms that cause them. According to Bhaskar, in the

domain of the natural sciences, the mechanisms under investigation are relatively enduring and do not depend on the activities of humans for their continuing existence.

Bhaskar extended his argument to social mechanisms that depend on human activities in his seminal book *The Possibility of Naturalism* (Bhaskar, 1979). Whilst in the case of the natural sciences Bhaskar starts with experimental activity and moves, through a process of abduction, to their conditions of possibility, he takes a different tack in the case of the social sciences. Rather than beginning with the practices of social scientists – whose results and methods are deemed to be less credible than their peers in physics or biology – Bhaskar begins with everyday social practices themselves. From these, he infers that social ontology is necessarily 'relational' in the sense that all social entities are defined by and exist by virtue of their relations to other entities. This relational ontology coupled with the existence of intentionality (understood as the 'directedness' or 'aboutness' of human mental states), which differentiates the social from the natural realm, allows Bhaskar to develop the transformational model of social activity (TMSA).

The TMSA is the most influential articulation of a philosophical ontology currently available to social scientists. Amongst others, it has been used by leading scholars in sociology (Margaret Archer, Andrew Sayer), law (Alan Norrie), economics (Tony Lawson), organization studies (Mike Reed), political economy (Bob Jessop), and international relations (Colin Wight). The TMSA conceives of the social world as an open system populated by active intentional actors. Individuals are not isolated atoms, but evolve within an institutional environment characterized by emergent social structures that are inherited from prior generations. The temporal and processual nature of social life is recognized in this social ontology as each new generation of actors adopts, rejects, or transforms inherited structures from its predecessors.

John Searle's influential contribution to the philosophy of the social sciences 'The Construction of Social Reality,' constitutes an attempt to build a realist account of social ontology from a mixture of commonsense and natural scientific insights (Searle, 1995). Like Bhaskar, he does not regard the efforts of social scientists as a legitimate basis for the articulation of his ontological perspective. Instead, Searle takes the findings of the natural sciences and in particular commitment to both physicalism and the findings of evolutionary biology as the starting point of his analysis. Yet he nevertheless sees a distinct place for the elaboration of a social ontology based on three characteristics of human collectivities that distinguish them from the physical systems analyzed by the natural sciences. His approach is founded upon these three key characteristics, which are the building blocks of his social ontology. When combined, collective intentionality, the assignment of function and constitutive rules allow him to build a comprehensive picture of the social domain.

The first building block, collective intentionality, refers to the human ability to share intentional states with others, such that one can legitimately say 'we believe,' 'we desire,' or 'we intend' (check cross-reference for CI). In invoking the second building block of his philosophical ontology, Searle emphasizes the ability to collectively assign functions to objects by adopting shared attitudes toward them. He claims that even the 'natural' function of the heart (to pump blood) is collectively attributed against a background teleology that values life and survival (Searle, 2006). His third building block is the notion of a constitutive rule. The existence of constitutive rules allows the attribution of a status to an object that is not dependent on or connected to its physical structure. Money can be a coin, a piece of paper, a plastic card or an electronic signal, so long as we collectively agree to treat it as money. Searle sums up his position in the following terms: "I have tried to state the logical form of the assignment of a status function when it becomes regular, and thus a matter of a rule, as that of the constitutive rule of the form, X counts as Y, or more commonly, *X counts as Y in context C*. Thus, such and such counts as a $20 bill in our society" (2006: p. 18).

Naturalism and Emergence

I end this article on a note of divergence between the two influential approaches to philosophical ontology that I have just outlined. A natural question for any critically minded reader of the literature on social ontology is to ask oneself: 'Why do we need a specifically 'social' ontology?' What is so different about social being that it requires a different conceptual toolkit to the one provided by the natural sciences?

One response might be to invoke the existence of a social substance that is somehow separate from and different to the substance studied by the natural sciences. But this response is rejected by the vast majority of secular philosophers and social scientists, including most social ontologists. There is a widespread desire to adhere to some form of ontological naturalism; in other words, to accept that whatever substance or material undergirds our experience of the natural and social worlds is common to both. Searle explicitly states his desire to present a systematic ontology of the social world that is consistent with the findings of physics and biology. Yet both Searle's and Bhaskar's approaches require some account of the emergent properties of organized (social or biological) systems, which may be difficult to reconcile with physicalism (check cross-reference on emergence).

Searle's framework is ultimately reducible to the mental properties of individual agents, even though these contain irreducible social concepts such as 'we-intentions.' His ontological individualism (check cross-reference on Meth Individualism) rests on a complex neurophysiological substratum to which individual mental states cannot be reduced. For Searle, therefore, it is the emergence of human consciousness in the form of mental states which can be shared and employed collectively that necessitates the development of a specifically social ontology. For Bhaskar, however, social relations, rules, and institutions are seen as emergent properties of human collective activity that cannot be reduced to the mental states of participants. This difference has further implications for where the mechanisms investigated by social scientists are located. A Searlian social ontology does not recognize the causal powers of supra-individual entities, whereas the TMSA does.

See also: Causation: Physical, Mental, and Social; Conventions and Norms: Philosophical Aspects; Shared Belief: Philosophical Aspects.

Bibliography

Al-Amoudi, I., 2010. Immanent non-algorithmic rules: an ontological study of social rules. Journal for the Theory of Social Behaviour 40 (3), 289–313.

Archer, M.S., 1995. Realist Social Theory: The Morphogenetic Approach. Cambridge University Press, Cambridge.

Faulkner, P., Runde, J., 2009. On the identity of technological objects and user innovations in function. Academy of Management Review 34 (3), 442–462.

Fleetwood, S., 2002. Hayek's Political Economy: The Socio-economics of Order. Psychology Press, London.

Latsis, J.S., 2005. Is there redemption for conventions? Cambridge Journal of Economics 29 (5), 709–727.

Lawson, T., 1997. Economics and Reality. Routledge, London and New York.

Lawson, T., 2003. Institutionalism: on the need to firm up notions of social structure and the human subject. Journal of Economic Issues 37 (1), 175–207.

Lawson, C., 2009. An ontology of technology. Techné: Research in Philosophy and Technology 12 (1), 48–64.

Lawson, C., Latsis, J.S., Martins, N. (Eds.), 2007. Contributions to Social Ontology. Routledge, London.

Maki, U., 2001. Economic ontology: what? why? how?. In: The Economic World View. Cambridge University Press, Cambridge, pp. 1–14.

Pratten, S., 2007. The scope of ontological theorising. Foundations of Science 12 (3), 235–256.

Quine, W.V.O., 1948. On what there is. The Review of Metaphysics, 21–38.

Searle, J.R., 1995. The Construction of Social Reality. Simon and Schuster, New York.

Searle, J.R., 2006. Social ontology: some basic principles. Anthropological Theory 6 (1), 12–29.

Social Phobia across the Lifespan

Stefan G Hofmann, Hannah Boettcher, and Jade Q Wu, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social anxiety disorder is characterized by fear and avoidance of social situations in which one may be evaluated by others. The disorder results from a complex interplay among genetic factors (e.g., family history of the disorder); temperamental factors (e.g., behavioral inhibition and anxiety sensitivity); biological contributors; parenting style and other environmental considerations; and cognitive factors, such as *interpretation biases* and heightened self-focused attention. In this article, we discuss the contribution of these factors for the etiology and maintenance of the disorder throughout the lifespan.

Introduction

Social anxiety disorder (SAD), also referred to as Social Phobia, is a chronic and debilitating disorder (Kessler et al., 2005) that differs in prevalence and manifestation among different cultures (Heinrichs et al., 2006; Hofmann et al., 2010). According to the current Diagnostic and Statistical Manual of Mental Disorders (DSM-V; American Psychiatric Association, 2013), SAD is characterized by persistent fear of negative evaluation, typically presenting as severe distress in and/or avoidance of situations where the individual could be scrutinized by others. These include initiating or maintaining conversations, attending social events, dating, interacting with authority figures, eating or writing in public, and particularly public speaking, the most common fear in individuals with SAD (Pollard and Henderson, 1988). Such symptoms cause distress and often lead to impairments in social relationships, as well as educational and professional development (American Psychiatric Association, 2013; Stein and Kean, 2000).

In the previous edition of the DSM (i.e., DSM-IV-TR; American Psychiatric Association, 2000), clinicians were instructed to consider the generalized subtype if the individual feared or avoided most or all social situations. Otherwise, the person met criteria for the less pervasive nongeneralized subtype of SAD. There was debate as to whether these subtypes of SAD represented qualitatively different syndromes, distinct subgroups within a syndrome, or a continuum of severity within one group (Hofmann et al., 2004a,b; McNeil, 2001). The distinction is further complicated by high comorbidity between the generalized subtype and Avoidant Personality Disorder (APD; Schneier et al., 1991), which has several overlapping symptoms with SAD. This has prompted the suggestion that APD and subtypes of SAD may lie along a spectrum of severity (Hofmann et al., 2004a,b; McNeil, 2001). However, there is evidence that differentiates APD, the generalized subtype of SAD, and nongeneralized subtypes. These include different ages of onset (Mannuzza et al., 1995), physiological versus subjective anxiety profiles in response to behavioral challenge (Heimberg et al., 1990; Hofmann et al., 1995b; Boone et al., 1999), as well as cognitive processing (Hofmann, 2007). Nevertheless, the most recent edition of the DSM (i.e., DSM-V; American Psychiatric Association, 2013) no longer recognizes the generalized subtype. Instead, it has a 'performance only' specifier to capture cases in which an individual only has symptoms during public speaking/performance situations. However, it remains to be seen whether this change is empirically justified and clinically useful. A recent study examined the rates and correlates of this new DSM-V specifier in a sample of 204 treatment-seeking children and adolescents with SAD (Kerns et al., 2013). When applying the DSM-IV subtype classification, 64.2% of the sample was classified as having a generalized subtype of SAD, with the remaining 35.2% classifying as having a nongeneralized subtype SAD. Participants with generalized SAD were older, had more clinically severe SAD, showed greater depressive symptoms, and were more likely to have a comorbid depressive disorder than those with nongeneralized SAD. In contrast, no children in this sample endorsed discrete fear in performance situations with an absence of fear in other social situations, questioning the meaningfulness of the new DSM-V specifiers in treatment-seeking youth with SAD.

SAD can occur at all stages in life and is shaped by developmental factors (Ollendick and Hirshfeld-Becker, 2002). Consistent with the transactional model (Sameroff, 2009), which posits change as continuous reciprocal interactions between the individual and his or her environment, SAD develops over time as a product of genetic, temperamental, physiological, cognitive, and various environmental factors. Shaped by interactions among these factors, the individual is both an active agent and a recipient of external influences. In a pathological context, the integration of factors contributing to development can fail and lead to poor adaptation, which, in the case of SAD, is the failure to comfortably navigate social situations. Within this framework, we will summarize the developmental course of SAD and the developmental factors in its etiology and maintenance, including genetic, temperamental, physiological, parental, environmental, and cognitive factors.

Developmental Course of SAD

Prevalence throughout the Lifespan

Recent estimates suggest that between 5 and 12% of people will meet DSM-IV diagnostic criteria for SAD at some point in their lives (Grant et al., 2005; Merikangas et al., 2010; Kessler et al., 2005). Most adults with generalized SAD report having experienced social anxiety for as long as they can remember (Bourdon et al., 1988).

Because individuals with social anxiety often fear being negatively evaluated by others, it is reasonable to imagine that the prevalence of SAD in childhood and adolescence might increase with age, as individuals become increasingly aware of others' capacity to judge them. Indeed, the prevalence of SAD in preadolescent children has consistently been found to be very low, usually well below 1% (Cartwright et al., 2006). This is consistent with research documenting the mean age of onset for symptoms of SAD at 13–15 years old (Kessler et al., 2005; Grant et al., 2005). Both healthy and anxious individuals experience increased anxiety related to social situations over the course of childhood and adolescence (Ollendick et al., 1989; in Ollendick and Hirshfeld-Becker, 2002), including modest increases from mid to late adolescence (Merikangas et al., 2010). It is rather uncommon for SAD to develop after young adulthood (Beesdo et al., 2009) or during old age, with typical estimates around 2–4% (e.g., Karlsson et al., 2009). Taken together, it seems clear that SAD is not displayed equivalently throughout the lifespan.

Age-Dependent Presentation

Given the important role of situational factors in SAD, it is unsurprising that SAD often manifests itself differently across age groups. The same fear of negative evaluation may prompt anxious children to miss school, young adults to avoid dating, and older adults to decline friends' invitations to keep in touch. More specifically, behaviors that characterize SAD in young childhood (e.g., crying, refusing to leave a caregiver) may be replaced by other behaviors in older individuals (e.g., avoidance of social situations, which is impossible for very young children; Ollendick and Hirshfeld-Becker, 2002). Furthermore, the subjective experience of social anxiety is different in young children compared with older children and adults. Young children are more likely to feel overwhelmed by an intense rush of distracting thoughts in social situations, while older children and adults report a greater frequency of specific negative beliefs about the self and fears of negative evaluation. Older individuals are more likely to see that their fears are excessive (Ollendick and Hirshfeld-Becker, 2002), although this recognition does not necessarily relieve suffering.

Just as SAD may present itself differently at different ages, functional impairment also varies with age, although high levels of life interference are common across most sufferers. In school-age children, missing school and poor academic performance (e.g., skipping presentations or failing to participate in class) are of significant concern. In young and middle adulthood, difficulties may shift toward finding and keeping a job and getting married (Hofmann and DiBartolo, 2010). SAD decreases work productivity (Wittchen et al., 2000), earning potential (Schneier et al, 1992) and career-promoting behavior (Phillips and Bruch, 1988). While older adults may be less concerned with finding a mate or keeping a job, functional impairment is no less serious in later life given the negative health effects associated with a weak social support network.

Treatment Considerations

When left untreated, SAD is unlikely to remit spontaneously. In fact, research consistently demonstrates that without treatment, it follows a chronic and often pervasive course (Davidson et al., 1993; Liebowitz et al., 1985; Schneier et al., 1992, 1994; Stein and Kean, 2000; Stein et al., 1996, 2000). Fortunately, a number of effective treatments exist, most notably cognitive behavioral therapy (e.g., Hofmann and Otto, 2008). Such treatments have to take into account the unique advantages and limitations of the individual's life stage. For example, opportunities for social interaction tend to decline with age, making treatment of SAD a challenge in older adults. A danger of misdiagnosis exists for young children with SAD. Comorbidity is high between SAD and selective mutism as well as depression (Ollendick and Hirshfeld-Becker, 2002). Some socially anxious children behave in ways that lead to erroneous diagnoses, as when avoidant children are misclassified as displaying oppositional defiance (Ollendick and Hirshfeld-Becker, 2002).

Genetic Factors

There is clear evidence for the notion that SAD is familial (Stein et al., 1998). Stein et al. (1998) found that individuals with a first-degree relative with SAD are 10 times more likely to have SAD than those without socially anxious relatives. There is consensus that this family liability is, at least in part, due to heritable genes (Hettema et al., 2001; Stein and Gelernter, 2010).

In an attempt to identify which specific genes are associated with SAD, several molecular genetic studies have been conducted (see Norrholm and Ressler, 2009; Domschke and Dannlowski, 2010). These studies have been scarce, mainly focusing on the SERT gene (SLC6A4), which codes for serotonin transporter protein (Stein and Gelernter, 2010). For example, Furmark et al. (2004) found that 5HTTLPR, a polymorphism of the SERT genetic locus often associated with psychiatric disorders, was associated with symptom severity in individuals with SAD, as well as enhanced activity in their amygdala when anxiety was induced. Another study has also found some evidence for the association of circadian genes (e.g., ARNTL2 and DRD2) with SAD (Sipilä et al., 2010). However, it must be cautioned that the existing studies do not provide definitive proof linking these genes to SAD, given lack of replication, low power, and the continuing debate regarding the delineation and validity of the phenotype.

Investigators have attempted to overcome these difficulties, which are relevant for any psychiatric genetics research, by studying endophenotypes (i.e., intermediate phenotypes) of SAD. That is, instead of studying samples with SAD, these studies focus on meaningful and presumably more homogenous subsamples of individuals with particular traits important in SAD. For example, it has been found that 5HTTLPR is associated with shyness (Arbelle et al., 2003; Furmark et al., 2004), and that the association between 5HTTLPR and introversion, another trait linked to SAD, is moderated by variants of a gene that codes for brain-derived neurotropic factor (BDNF), demonstrating a gene by gene interaction (Terracciano et al., 2008, 2010). Low extraversion has also been associated with the gene-encoding regulator of G protein signaling 2 (RGS2; Smoller et al., 2008) and the dopamine D4 receptor gene (DRD4). The related trait of neuroticism has been associated with MAMDC1 (van den Oord et al., 2008) and SNAP25 (Terracciano et al., 2008). The trait with the strongest genetic liability in relation to SAD is behavioral inhibition, as shown in various studies (e.g., Stein and Gelernter, 2010). This temperament has been associated with the glutamic acid decarboxylase gene that encodes a GABA synthesis enzyme (Smoller et al., 2001), as well as the CRH and RGS2 loci (Smoller et al., 2003, 2008). We will discuss this trait in more detail in the following paragraphs.

Temperamental Factors

Behavioral Inhibition

Behavioral inhibition (BI) is a childhood temperament, first described by Kagan et al. (1984) that occurs in approximately 10–15% of white American children. It is associated with fear, avoidance, restraint and long latencies to approach in the face of unfamiliar persons, objects, or situations. Although not all children high in BI will develop SAD and not all those with SAD have high levels of BI (Ollendick and Hirshfield-Becker, 2002), this trait has been strongly linked to SAD. For example, it has been found that highly inhibited 6-year-old children had significantly higher rates of SAD (Biederman et al., 2001), and BI is correlated with social anxiety symptoms in young adolescents (Muris et al., 1999) and college students (Mick and Telch, 1998). Using retrospective self- and mothers' reports (Hayward et al., 1998; Chronis-Tuscano et al., 2009), as well as prospective methods (Schwartz et al., 1999; Hirshfeld-Becker et al., 2007), investigators have found that high levels of BI earlier in life predicted the development of SAD later. Young children with high BI, compared with those with low BI, had a 3.15 times increased odds of developing SAD within 5 years after the temperament was detected (Hirshfeld-Becker et al., 2007). There is further evidence that BI is associated with a decreased risk for disruptive behavior disorders, and that there is no association between BI and other anxiety disorders (Biederman et al., 2001; Schwartz et al, 1999). However, other studies suggest that BI may be associated with longitudinal, concurrent, or parental risk for other disorders, such as specific phobias, avoidant disorder, panic disorder, other anxiety disorders, and depression (Biederman et al., 1990, 1993, 2001).

Anxiety Sensitivity

Anxiety sensitivity (AS) is another commonly researched tempermental factor and possible endophenotype candidate. AS refers to heightened sensitivity to the physiological experience of anxiety (Reiss et al., 1986). When individuals high in AS experience sensations such as heart racing, blushing, or sweaty palms, these sensations are noticed more quickly and perceived as more distressing, compared with individuals with low AS. People with SAD and other anxiety disorders tend to display greater AS, which may contribute to both the etiology and maintenance of SAD. For example, someone with high AS may interpret their own nonpathological reactions to early social experiences (e.g., nervousness on the first day of school) as more unpleasant, leading to discomfort in social situations. AS also contributes to avoidance of anxiety-provoking situations, which has been consistently identified as a maintaining factor in SAD (e.g., Rapee and Heimberg, 1997).

Parental Influence

Children are more likely to develop SAD if they have anxious or depressed parents (Lieb et al., 2000; Mancini et al., 1996), and parental influence on the etiology of SAD is not limited to genetics. Parents may contribute to the development of SAD in a variety of ways, and a number of trends emerge in the ways socially anxious people describe their parents (Ollendick and Hirshfield-Becker, 2002). For example, parents of socially anxious individuals tend to display lower affection and are more likely to be both overprotective and rejecting of their children. They are also more likely to use shame as a punishment technique. Parents who provide negative evaluation of social performance, especially when it is coupled with punishment, teach children that social mishaps and social situations in general are associated with risk. Some researchers have hypothesized that early life experiences of repeated criticism and rejection by parents may sensitize children to worries about negative social evaluation (Bruch et al., 1989; Ollendick and Hirshfield-Becker, 2002).

Parents, caregivers, and teachers may also increase a child's risk for SAD by overemphasizing the importance of gaining others' approval and maintaining very high standards for social interaction – values underlying cognitive models of SAD (e.g., Hofmann, 2007). In addition, parents who themselves have substantial social-evaluative fear are more likely to display avoidant or anxious behavior around socializing, allowing their children to pick up maladaptive habits through modeling.

What remains unclear in discussions of parental influence is whether parenting styles create anxious children, whether anxious children elicit certain types of responses from caregivers, or whether children and caregivers mutually influence each other. Well-meaning parents may at some

point worsen the cycle of avoidance in socially anxious children by accommodating their behavior. For example, a sympathetic parent may let a child miss school due to anxiety, or speak up on his behalf when he is tongue-tied. A desire to alleviate children's social distress may be even higher in parents who themselves have experience with SAD.

Sensitizing Events

SAD may be catalyzed by chronic or acute sensitizing events. In fact, some studies suggest that about half of all socially anxious people describe their anxiety as traceable to a specific highly embarrassing situation (Öst, 1985; Stemberger et al., 1995). In contrast, another study found that just 15% of socially anxious participants believed their anxiety came from a specific event, and many people experienced humiliating events without developing SAD (Hofmann et al., 1995a). Thus, further research is required to identify the contextual factors responsible for determining whether or not an aversive social event will precipitate SAD.

In the case of chronic sensitizing events, an individual may experience repeated negative social feedback (e.g., bullying, criticism from parents when speaking in public) such that some or most social situations become associated with negative outcomes (Beidel and Turner, 2007). Alternatively, an individual may endure a single humiliating social event that is experienced as highly aversive, leading to fear and avoidance of a similar outcome in the future. It is less common, but also possible to develop SAD after witnessing another person undergo social humiliation (Öst, 1985), known as vicarious conditioning.

Unfortunately, it may only take one or two sensitizing events to contribute to a self-perpetuating cycle of social rejection, withdrawal, and anxiety. Children who have negative social experiences are likely to withdraw, and withdrawn children are more likely to be rejected by their peers, leading to a greater incidence of further negative social experiences.

Cognitive Factors

In recent years, research has focused on identifying and treating cognitive biases characteristic of anxiety disorders. Among the most important cognitive biases displayed by individuals with SAD are *attention bias* and *interpretation bias* (Heinrichs and Hofmann, 2001). *Attention bias* refers to the fact that socially anxious individuals are likely to notice negative social information (e.g., an audience member yawning during a talk) compared with positive social information (e.g., the same person smiling later). *Interpretation bias* reflects the tendency of socially anxious individuals to interpret ambiguous social cues as negative (e.g., believing that someone laughed at a comment out of derision rather than amusement).

Interpretation bias has been documented in young socially anxious children, who interpret ambiguous stories in a negative light (Bögels and Zigterman, 2000). In the specific example of *interpretation bias*, socially anxious children consistently rate their own social behavior more poorly than do other observers (Rapee, 1995). Cognitive biases in adults with SAD are well documented (Foa et al., 1996; Hofmann, 2004, 2007; Heinrichs and Hofmann, 2001) and recent clinical trials have attempted to modify cognitive biases as a treatment aim for SAD, with some success (e.g., Beard and Amir, 2008; see also Hallion and Ruscio, 2011; Beard et al., 2012). More generally, SAD may be maintained by *confirmation bias*, which is common in healthy populations and refers to people's tendency to seek out, notice, and remember information that is consistent with preexisting expectations (e.g., that one is unlikeable). An important question for future research is whether cognitive biases are a cause of SAD, a consequence of it, or both, and whether the specific biases change throughout life.

Adults with SAD show heightened self-focused attention in social situations and typically experience negative self-perception as a social object (Clark and Wells, 1995; Hofmann, 2007; Rapee and Heimberg, 1997). Adult SAD patients typically focus on their own thoughts, physical sensations, and behavior, and this is accompanied by self-scrutiny (Hirsch et al., 2003) and unfavorable images of the self that are interpreted as accurate (Hofmann and Heinrichs, 2003). When considering social situations, socially anxious individuals are more likely to picture themselves from the point of view of an observer (Hackmann et al., 1998). Directing attention away from the self, a challenge for people with SAD, has been shown to decrease both negative self-evaluation and anxiety (Wells and Papageorgiou, 1998; Hofmann et al., 2004a,b).

Negative self-perception as a social object, a critical aspect in contemporary theories of SAD may be more relevant in adult patients with SAD than in younger children who do not yet have a matured sense of self. Therefore, it is possible that temperamental and biological factors are more predominant in the expression of SAD in younger individuals, whereas self-perception, which is shaped by autobiographical, societal, sociodemographic, and cultural factors, may be relatively more predominant in adult individuals with SAD.

Summary and Conclusions

SAD is a prevalent and debilitating disorder that impairs an individual's functioning in social relationships, occupational performance, and other important areas of life. SAD can be expressed as a developmental disorder with varying prevalence and presentations across the lifespan. Different factors influence the expression of SAD from birth to adulthood. These factors include: genetic liability as demonstrated by family and molecular genetics studies; temperamental factors like behavioral inhibition, a childhood trait

characterized by avoidance and fear of novelty; physiological factors like AS, which contribute to the maintenance of anxiety symptoms by facilitating avoidance; parental influences, which can include parents' own psychopathology and/or enabling behaviors; sensitizing events such as particularly humiliating social experiences; and cognitive factors like *attention* and *interpretation biases*, which not only heighten the likelihood of feeling fear in social situations, but also perpetuate avoidance thereof. In sum, an individual is both a recipient and an agent in his or her interaction with factors associated with SAD. The resulting complex presentation varies among individuals. Taking a developmental perspective when confronted with each case would provide the clinician a well-rounded understanding of its idiosyncratic etiology and benefit treatment planning. Nevertheless, further work is needed to improve the diagnosis and classification of SAD. Future classification systems should integrate and utilize physiological, genetic, and cognitive factors contributing to SAD and related disorders.

See also: Anxiety and Anxiety Disorders; Panic Disorder and Agoraphobia Across the Lifespan; Phobias Across the Lifespan.

Bibliography

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., text revision. American Psychiatric Association, Washington, D.C.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Association, Washington, DC.

Arbelle, S., Benjamin, J., Golin, M., Kremer, I., Belmaker, R.H., Ebstein, R.P., 2003. Relation of shyness in grade school children to the genotype for the long form of the serotonin transporter promoter region polymorphism. American Journal of Psychiatry 160, 671–676.

Beard, C., Amir, N., 2008. A multi-session interpretation modification program: changes in interpretation and social anxiety symptoms. Behaviour Research and Therapy 46 (10), 1135.

Beard, C., Sawyer, A.T., Hofmann, S.G., 2012. Efficacy of attention bias modification using threat and appetitive stimuli: a meta-analytic review. Behavior Therapy 43, 724–740.

Beesdo, K., Knappe, S., Pine, D.S., 2009. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V. The Psychiatric Clinics of North America 32 (3), 483.

Beidel, D.C., Turner, S.M., 2007. Shy Children, Phobic Adults: Nature and Treatment of Social Anxiety Disorder. American Psychological Association, Washington, DC.

Biederman, J., Rosenbaum, J.F., Hirshfeld, D.R., Faraone, S.V., Bolduc, E.A., Gersten, M., Reznick, J.S., 1990. Psychiatric correlates of behavioral inhibition in young children of parents with and without psychiatric disorders. Archives of General Psychiatry 47 (1), 21.

Biederman, J., Rosenbaum, J.F., Bolduc-Murphy, E.A., Faraone, S.V., Chaloff, J., Hirshfeld, D.R., Kagan, J., 1993. A 3-year follow-up of children with and without behavioral inhibition. Journal of the American Academy of Child & Adolescent Psychiatry 32 (4), 814–821.

Biederman, J., Hirshfeld-Becker, D.R., Rosenbaum, J.F., Hérot, C., Friedman, D., Snidman, N., Faraone, S.V., 2001. Further evidence of association between behavioral inhibition and social anxiety in children. American Journal of Psychiatry 158 (10), 1673–1679.

Bögels, S.M., Zigterman, D., 2000. Dysfunctional cognitions in children with social phobia, separation anxiety disorder, and generalized anxiety disorder. Journal of Abnormal Child Psychology 28 (2), 205–211.

Boone, M.L., McNeil, D.W., Masia, C.L., Turk, C.L., Carter, L.E., Ries, B.J., Lewin, M.R., 1999. Multimodal comparisons of social phobia subtypes and avoidant personality disorder. Journal of Anxiety Disorders 13, 271–292.

Bourdon, K.H., Boyd, J.H., Rae, D.S., Burns, B.J., Thompson, J.W., Locke, B.Z., 1988. Gender differences in phobia: results of the ECA community survey. Journal of Anxiety Disorders 2, 227–241.

Bruch, M.A., Heimberg, R.G., Berger, P., Collins, T.M., 1989. Social phobia and perceptions of early parental and personal characteristics. Anxiety Research 2 (1), 57–65.

Cartwright-Hatton, S., McNicol, K., Doubleday, E., 2006. Anxiety in a neglected population: prevalence of anxiety disorders in pre- adolescent children. Clinical Psychology Review 26 (7), 817–833.

Chronis-Tuscano, A., Degnan, K.A., Pine, D.S., Perez-Edgar, K., Henderson, H.A., Diaz, Y., Fox, N.A., 2009. Stable early maternal report of behavioral inhibition predicts lifetime social anxiety disorder in adolescence. Journal of the American Academy of Child and Adolescent Psychiatry 48 (9), 928.

Clark, D.M., Wells, A., 1995. A cognitive model of social phobia. In: Heimberg, R.G., Liebowitz, M.R., Hope, D.A., Schneier, F.R. (Eds.), Social Phobia: Diagnosis, Assessment and Treatment. Guilford Press, New York, pp. 69–93.

Davidson, J.R.T., Hughes, D.L., George, L.K., Blazer, D.G., 1993. The epidemiology of social phobia: findings from the Duke epidemiological catchment area study. Psychological Medicine 23, 709–718.

Domschke, K., Dannlowski, U., 2010. Imaging genetics of anxiety disorders. NeuroImage 53 (3), 822.

Foa, E.B., Franklin, M.E., Perry, K.J., Herbert, J.D., 1996. Cognitive biases in generalized social phobia. Journal of Abnormal Psychology 105 (3), 433.

Furmark, T., Tillfors, M., Garpenstrand, H., Marteinsdottir, I., Långström, B., Oreland, L., Fredrikson, M., 2004. Serotonin transporter polymorphism related to amygdala excitability and symptom severity in patients with social phobia. Neuroscience Letters 362 (3), 189–192.

Grant, B.F., Hasin, D.S., Blanco, C., Stinson, F.S., Chou, S.P., Goldstein, R.B., Huang, B., 2005. The epidemiology of social anxiety disorder in the United States: results from the National Epidemiologic survey on alcohol and related conditions. Journal of Clinical Psychiatry 66 (11), 1351–1361.

Hackmann, A., Surawy, C., Clark, D.M., 1998. Seeing yourself through others' eyes: a study of spontaneously occurring images in social phobia. Behavioural and Cognitive Psychotherapy 26, 3–12.

Hallion, L.S., Ruscio, A.M., 2011. A meta-analysis of the effect of cognitive bias modification on anxiety and depression. Psychological Bulletin 137 (6), 940.

Hayward, C., Killen, J.D., Kraemer, H.C., Taylor, C., 1998. Linking self-reported childhood behavioral inhibition to adolescent social phobia. Journal of the American Academy of Child & Adolescent Psychiatry 37 (12), 1308–1316.

Heimberg, R.G., Hope, D.A., Dodge, C.S., Becker, R.E., 1990. DSM-III-R subtypes of social phobia – comparison of generalized social phobics and public speaking phobics. The Journal of Nervous and Mental Disease 176, 172–179.

Heinrichs, N., Hofmann, S.G., 2001. Information processing in social phobia: a critical review. Clinical Psychology Review 21, 751–770.

Heinrichs, N., Rapee, R.M., Alden, L.A., Bögels, S., Hofmann, S.G., Ja Oh, K., Sakano, Y., 2006. Cultural differences in perceived social norms and social anxiety. Behaviour Research and Therapy 44 (8), 1187–1197.

Hettema, J.M., Neale, M.C., Kendler, K.S., 2001. A review and meta-analysis of the genetic epidemiology of anxiety disorders. American Journal of Psychiatry 158 (10), 1568–1578.

Hirsch, C.R., Clark, D.M., Mathews, A., Williams, R., 2003. Self-images play a causal role in social phobia. Behaviour Research and Therapy 41, 901–921.

Hirshfeld-Becker, D.R., Biederman, J., Henin, A., Faraone, S.V., Davis, S., Harrington, K., Rosenbaum, J.F., 2007. Behavioral inhibition in preschool children at risk is a specific predictor of middle childhood social anxiety: a five-year follow-up. Journal of Developmental & Behavioral Pediatrics 28 (3), 225–233.

Hofmann, S.G., 2004. Cognitive mediation of treatment change in social phobia. Journal of Consulting and Clinical Psychology 72, 392–399.
Hofmann, S.G., 2007. Cognitive factors that maintain social anxiety disorder: a comprehensive model and its treatment implications. Cognitive Behaviour Therapy 36, 195–209.
Hofmann, S.G., DiBartolo, P.M. (Eds.), 2010. Social Anxiety: Clinical, Developmental, and Social Perspectives, second ed. Elsevier/Academic Press, Amsterdam, The Netherlands.
Hofmann, S.G., Heinrichs, N., 2003. Differential effect of mirror manipulation on self-perception in social phobia subtypes. Cognitive Therapy and Research 27, 131–142.
Hofmann, S.G., Otto, M.W., 2008. Cognitive-behavior Therapy of Social Anxiety Disorder: Evidence-based and Disorder Specific Treatment Techniques. Routledge, New York, NY.
Hofmann, S.G., Ehlers, A., Roth, W.T., 1995a. Conditioning theory: a model for the etiology of public speaking anxiety? Behaviour Research and Therapy 33, 567–571.
Hofmann, S.G., Newman, M.G., Ehlers, A., Roth, W.T., 1995b. Psychophysiological differences between subgroups of social phobia. Journal of Abnormal Psychology 104, 224–231.
Hofmann, S.G., Heinrichs, N., Moscovitch, D.A., 2004a. The nature and expression of social phobia: toward a new classification. Clinical Psychology Review 24, 769–797.
Hofmann, S.G., Moscovitch, D.A., Kim, H.-J., Taylor, A.N., 2004b. Changes in self-perception during treatment of social phobia. Journal of Consulting and Clinical Psychology 72, 588–596.
Hofmann, S.G., Asnaani, A., Hinton, D.E., 2010. Cultural aspects in social anxiety and social anxiety disorder. Depression and Anxiety 27, 1117–1127. http://dx.doi.org/10.1002/da.20759.
Kagan, J., Reznick, J.S., Clarke, C., Snidman, N., Garcia-Coll, C., 1984. Behavioral inhibition to the unfamiliar. Child Development, 2212–2225.
Karlsson, B., Klenfeldt, I.F., Sigström, R., Waern, M., Östling, S., Gustafson, D., Skoog, I., 2009. Prevalence of social phobia in non-demented elderly from a Swedish population study. American Journal of Geriatric Psychology 17 (2), 127–135.
Kerns, C.E., Comer, J.S., Pincus, D.B., Hofmann, S.G., 2013. Evaluation of the proposed social anxiety disorder specifier change for DSM-5 is a treatment-seeking sample of anxious youth. Depression and Anxiety 30 (8), 709–715.
Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62 (6), 617.
Lieb, R., Wittchen, H.U., Hofler, M., Fuetsch, M., Stein, M.B., Merikangas, K.R., 2000. Parental psychopathology, parenting styles, and the risk of social phobia in offspring: a prospective-longitudinal community study. Archives of General Psychiatry 57 (9), 859.
Liebowitz, M.R., Gorman, J.M., Fyer, A.J., Klein, D.F., 1985. Social phobia: review of a neglected anxiety disorder. Archives of General Psychiatry 42, 729–736.
Mancini, C., van Ameringen, M., Szatmari, P., Fugere, C., Boyle, M., 1996. A high-risk pilot study of the children of adults with social phobia. Journal of the American Academy of Child & Adolescent Psychiatry 35 (11), 1511–1517.
Mannuzza, S., Schneier, F.R., Chapman, T.F., Liebowitz, M.R., Klein, D.F., Fyer, A.J., 1995. Generalized social phobia. Reliability and validity. Archives of General Psychiatry 52, 230–237.
McNeil, D.W., 2001. Terminology and evolution of the constructs. In: Hofmann, S.G., DiBartolo, P.M. (Eds.), From Social Anxiety to Social Phobia: Multiple Perspectives. Allyn & Bacon, Needham Heights, MA, pp. 8–19.
Merikangas, K.R., He, M.J.P., Burstein, M., Swanson, M.S.A., Avenevoli, S., Cui, M.L., Swendsen, J., 2010. Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Study-Adolescent Supplement (NCS-A). Journal of the American Academy of Child and Adolescent Psychiatry 49 (10), 980.
Mick, M.A., Telch, M.J., 1998. Social anxiety and history of behavioral inhibition in young adults. Journal of Anxiety Disorders 12 (1), 1–20.
Muris, P., Merckelbach, H., Wessel, I., Van de Ven, M., 1999. Psychopathological correlates of self-reported behavioural inhibition in normal children. Behaviour Research and Therapy 37 (6), 575–584.
Norrholm, S.D., Ressler, K.J., 2009. Genetics of anxiety and trauma-related disorders. Neuroscience 164 (1), 272.
Ollendick, T.H., Hirshfeld-Becker, D.R., 2002. The developmental psychopathology of social anxiety disorder. Biological Psychiatry 51 (1), 44–58.
Ollendick, T.H., King, N.J., Frary, R.B., 1989. Fears in children and adolescents: reliability and generalizability across gender, age and nationality. Behaviour Research and Therapy 27 (1), 19–26.
Öst, L.G., 1985. Ways of acquiring phobias and outcome of behavioral treatments. Behaviour Research and Therapy 23 (6), 683–689.
Phillips, S.D., Bruch, M.A., 1988. Shyness and dysfunction in career development. Journal of Counseling Psychology 35 (2), 159.
Pollard, C.A., Henderson, J.G., 1988. Four types of social phobia in a community sample. Journal of Nervous and Mental Disease 176, 440–445.
Rapee, R.M., 1995. Descriptive psychopathology of social phobia. In: Heimberg, R.G. (Ed.), Social Phobia: Diagnosis, Assessment, and Treatment. Guilford Press, New York, pp. 41–66.
Rapee, R.M., Heimberg, R.G., 1997. A cognitive-behavioral model of anxiety in social phobia. Behaviour Research and Therapy 35 (8), 741–756.
Reiss, S., Peterson, R.A., Gursky, D.M., McNally, R.J., 1986. Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. Behaviour Research and Therapy 24 (1), 1–8.
Sameroff, A.E., 2009. The Transactional Model of Development: How Children and Contexts Shape Each Other. American Psychological Association, Washington, DC.
Schneier, F.R., Spitzer, R.L., Gibbon, M., Fyer, A.J., Liebowitz, M.R., 1991. The relationship of social phobia subtypes and avoidant personality disorder. Comprehensive Psychiatry 32 (6), 496–502.
Schneier, F.R., Johnson, J., Hornig, C.D., Liebowitz, M.R., Weissman, M.M., 1992. Social phobia: comorbidity and morbidity in an epidemiological sample. Archives of General Psychiatry 49, 282–288.
Schneier, F.R., Heckelman, L.R., Garfinkel, R., Campeas, R., Fallon, B., Gitow, A., Street, L., DelBene, D., Liebowitz, M.R., 1994. Functional impairment in social phobia. Journal of Clinical Psychiatry 55, 322–331.
Schwartz, C.E., Snidman, N., Kagan, J., 1999. Adolescent social anxiety as an outcome of inhibited temperament in childhood. Journal of the American Academy of Child & Adolescent Psychiatry 38 (8), 1008–1015.
Sipilä, T., Kananen, L., Greco, D., Donner, J., Silander, K., Terwilliger, J.D., Hovatta, I., 2010. An association analysis of circadian genes in anxiety disorders. Biological Psychiatry 67 (12), 1163–1170.
Smoller, J.W., Rosenbaum, J.F., Biederman, J., Susswein, L.S., Kennedy, J., Kagan, J., Slaugenhaupt, S.A., 2001. Genetic association analysis of behavioral inhibition using candidate loci from mouse models. American Journal of Medical Genetics 105 (3), 226–235.
Smoller, J.W., Rosenbaum, J.F., Biederman, J., Kennedy, J., Dai, D., Racette, S.R., Slaugenhaupt, S.A., 2003. Association of a genetic marker at the corticotropin-releasing hormone locus with behavioral inhibition. Biological Psychiatry 54 (12), 1376–1381.
Smoller, J.W., Gardner-Schuster, E., Covino, J., 2008. The genetic basis of panic and phobic anxiety disorders. American Journal of Medical Genetics Part C: Seminars in Medical Genetics 148 (2), 118–126.
Stein, M.B., Gelernter, J., 2010. Genetic basis of social anxiety disorder. In: Hofmann, S.G., DiBartolo, P.M. (Eds.), Social Anxiety: Clinical, Developmental and Social Perspectives. Academic Press, Burlington, MA.
Stein, M.B., Kean, Y.M., 2000. Disability and quality of life in social phobia: epidemiologic findings. American Journal of Psychiatry 157, 1606–1613.
Stein, M.B., Walker, J.R., Forde, D.R., 1996. Public speaking fears in a community sample. Prevalence, impact on functioning, and diagnostic classification. Archives of General Psychiatry 53, 169–174.
Stein, M.B., Chartier, M.J., Hazen, A.L., Kozak, M.V., Tancer, M.E., Lander, S., Walker, J.R., 1998. A direct-interview family study of generalized social phobia. American Journal of Psychiatry 155 (1), 90–97.
Stein, M.B., Torgrud, L.J., Walker, J., 2000. Social phobia symptoms, subtypes, and severity. Archives of General Psychiatry 57, 1046–1052.
Stemberger, R.T., Turner, S.M., Beidel, D.C., Calhoun, K.S., 1995. Social phobia: an analysis of possible developmental factors. Journal of Abnormal Psychology 104 (3), 526.
Terracciano, A., Sanna, S., Uda, M., Deiana, B., Usala, G., Busonero, F., Villafuerte, S., 2008. Genome-wide association scan for five major dimensions of personality. Molecular Psychiatry 15 (6), 647–656.
Terracciano, A., Martin, B., Ansari, D., Tanaka, T., Ferrucci, L., Maudsley, S., Costa, P.T., 2010. Plasma BDNF concentration, Val66Met genetic variant and

depression-related personality traits. Genes, Brain and Behavior 9 (5), 512–518.

van den Oord, E.J., Kuo, P.H., Hartmann, A.M., Webb, B.T., Moller, H.J., Hettema, J.M., Rujescu, D., 2008. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. Archives of General Psychiatry 65 (9), 1062.

Wells, A., Papageorgiou, C., 1998. Social phobia: effects of external attention on anxiety, negative beliefs, and perspective taking. Behavior Therapy 29, 357–370.

Wittchen, H.U., Fuetsch, M., Sonntag, H., Müller, N., Liebowitz, M., 2000. Disability and quality of life in pure and comorbid social phobia. Findings from a controlled study. European Psychiatry 15 (1), 46–58.

Social Properties (Facts and Entities): Philosophical Aspects

David-Hillel Ruben, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social ontology is the study of what kinds of things discourse and thought about society is committed to, in order to be true. The article considers various candidates: social entities, social properties (and hence social actions), social facts, and evaluates the case for the existence of things of each kind.

Several philosophical strategies are described that might permit escape from such commitment: elimination, replacement, and reduction. The likely success of these strategies is also evaluated. The article argues that the many problems besetting these strategies, especially that of reduction, strongly suggest the irreducible existence of things belonging to these kinds.

One area of philosophical concern, dating from the very beginning of philosophy itself, is ontology: what is there? or what kinds of things are there? Ontology is the philosophical study of what exists. Many questions about existence are straightforwardly empirical, and hence of no interest to the ontologist. Are there unicorns, wombats, or planets circling stars other than our Sun? Other existence questions are not empirical, and these are in the province of the ontologist. Are there any physical objects, ideas, forms, numbers, properties, sets, propositions, minds, concepts, a Deity, or social wholes? To answer these questions is to engage in producing a list of the metaphysical kinds of things that one believes exist.

'Existence' in what sense? Some philosophers have held that each metaphysically different kind of thing exists in its own, special sense of 'existence,' or they may distinguish between being, existence, subsistence, and so on. A more plausible view is that 'existence' is univocal, that whatever exists, exists in the same sense, and that differences in the 'mode' of existence, say, of mind and body (if such there be), arise not because existence differs in the two cases but because of the differences between minds and bodies themselves. Bodies have a spatial 'mode' of existence, it might be thought, and minds do not, but, if so, that is because of the differences between them, not because they exist in different senses of that word. Besides, if existence had a special sense for each metaphysical kind of thing, ontology could become somewhat trivial. To the question of whether, say, propositions really exist, it would be anodyne simply to say that they do exist, but only in the special proposition sense of existence; that minds exist, but only in the special mental sense of existence; or that possibilia subsist rather than exist (since that is what possible things do), whereas actual things exist rather than subsist. Ontology would entail no honest labor, and therefore no real rewards.

All thought and all discourse presuppose the existence of certain things in order for at least some part of that thought or discourse to be true. When we speak, indeed even whenever we simply think, we typically take ourselves to be speaking or thinking about various things and ascribing certain properties to them. These entities can be concrete objects, properties, events, processes, or abstract objects (to name but a few candidates). It is not easy to capture this notion of aboutness (see Goodman, 1961), but I rely on the readers' intuitive grasp of the idea. Let us call whatever I think or speak about 'entities,' and whatever I ascribe to them 'properties.' Introduced in this way, these are very broad notions, especially that of a property, but there is no reason to limit the idea of a property more carefully in this context.

Are there, ontologically speaking, properties as well as entities? There are at least two ways in which to handle this question. First, there is the austere point of view that says that ascribing properties is not sufficient to be ontologically committed to them. There are properties, on the austere view, only if we talk or think about them, refer to them, and such talk or thought cannot be eliminated or replaced by other talk or thought that makes no such ontological commitment (see Quine, 1963). On this austere view, in the sentence 'The table is white,' we refer to the table but not to the color white or to whiteness, so no ontological commitment to whiteness is made by such a sentence.

There are, however, cases in which we do appear to refer to properties, even on the austere view: for exmaple (1) 'The color crimson is a shade of the color red.' It is easy to show that (2) 'The set of crimson things is a subset of the set of red things' will not do as a replacement for (1). If no such replacement or elimination is possible, then there are some properties, on the austere view.

Second, there is the more relaxed approach, which says that we are ontologically committed to properties if they are actually instantiated (or perhaps even if only well defined), whether or not we refer to them or talk about them. Ascription alone, without reference, may still involve us in ontological commitment on the relaxed view. If so, then even 'The table is white,' which ascribes the property of being white to the table, but does not refer to that property, commits us to the property of being white as well as to the table, since the sentence, if true, insures that whiteness has at least one actual instance.

Facts are of a more recent provenance. There appear to be singular facts, like the fact that some entity has some property, general facts, like the fact that whatever has one property also has another, and facts of identity (this list is not necessarily exhaustive). Are there really such things as facts? If there are facts, are they merely true propositions or are they anchored in the world itself? Philosophical views about the nature of facts

range from the wordy (or conceptual) to the worldly about their nature, and from the skeptical to the realist about their existence, without there being any philosophical consensus on either their existence or their nature. Facts have been enlisted for many purposes, one such being a correspondence theory of truth, and another being causal analysis. Are sentences true when they correspond to the facts? Is it the fact that the match was struck that caused it to light? These two particular uses of facts have come in for some serious criticism. Even so, there may well be facts, even if they were useless in theories of truth and for analyses of causality.

Ontological Commitments and Three Ways to Wriggle Out of Them

Sometimes, the apparent ontological commitments we have when we speak or think about things turn out to be only apparent. We do not always have to take our apparent ontological commitments seriously. This can happen in several different ways, three of which I want to describe.

First, I can make genuine mistakes, for example, when I take myself to be thinking about something that does not in fact exist or attributing to something a property it does not in fact have: Atlantis, Vulcan, the ether, or being a witch. These are cases in which we really are intending to speak about something, or say something about it, but fail. Let us call these sorts of cases, 'cases of elimination.' There simply are no such things or nothing really has the property in question (there is such a property as the property of being a witch, but nothing really has that property), and if we have at first placed them on our list of what there is, or what things are like, on finding that they do not exist or are not like that, we remove or eliminate them from the ontological catalog. Some philosophers have said this about minds, or mental states or mental properties. In cases of elimination, nothing can be salvaged from talk about such things as it stands. Such talk is plain false (see Lycan and Pappas, 1972).

Second, there are cases that, on first appearance, seem to be ones in which I am speaking or thinking about something, but closer inspection dispels the appearance. Not all cases of apparently thinking about something turn out, on analysis, really to be such. I may not be seriously intending to talk about the thing in question at all. Fictional discourse and metaphor provide two examples of this. If Romeo said that Juliet was the sun, it would be churlish to accuse him of uttering a falsehood on the grounds that she is not a heavenly body. (Of course Romeo and Juliet do not exist anyway, so are fictional, as they are only characters in a play.) A different example, more relevant for social science, for example, is talk about the average family, or the stakes being high in deciding some question, but such talk or thought is not really about what it might seem to be about. There is no such thing or things as the average family or as high stakes.

Average-family talk or thought is merely shorthand for talking about the number of people in the country divided by the number of households (or something like this anyway). Talk about the stakes being high is merely a way of talking about the important consequences of some decision. There is no such thing as the average family, or as high stakes, nothing that should be taken ontologically seriously, beyond taking seriously households and important consequences. These expressions are, as it were, just a façon de parler.

Let us call these last two examples, 'cases of replacement.' (How one is to deal with fiction or metaphor is far trickier.) The original talk, r_1 ('the average family,' 'high stakes'), superficially has ontological commitments of some sort. The replacing talk, r_2 ('households' and 'important consequences'), does not have those same apparent ontological commitments (although of course it has some of its own). The burden of proof, in such cases, falls on the person wishing to show that some of the ontological commitments of the original talk are apparent but not real. He or she must indicate the replacing talk that lacks them. In these cases, our original talk gets salvaged, because it is true, if the replacing sentence without the same apparent ontological commitments is true.

Third, there are cases in which our talk or thought really and not just apparently is about some entities, but they turn out not to be sui generis, since they can be reductively identified as or with another thing. These are, appropriately, cases of reductive identification. These might be examples of the sort of reductive identification that claims so much of philosophers' energies. There are, it may be said, physical objects, but they are really only sets of sense data; there are mental states, but they are identical with brain states; there are numbers, but they are merely the same as sets of sets; there are properties, but they are nothing more than sets of similar objects. Facts might just be true propositions. Such attempts at reductive identification allow the philosophical cake to be both had and eaten. Along with common sense, it can be asserted that there are such things. However, along with some ideal of philosophical acceptability, they can be reductively identified with another thing after all.

Insofar as philosophical reductive identification succeeds, it allows us to shorten our ontological list, while not denying the claims of common sense. It is for this reason that the term 'reductive' is appropriate; these identifications permit a reduction in the number of basic ontological commitments. On a physicalist reductive analysis, for example, there are mental states, or mental properties, but they do not have to be taken ontologically seriously, beyond taking bodily or brain states or properties seriously. According to Berkeleyian idealism, there are tables, chairs, and so on, but they are only ideas in the mind. Once ideas in the mind are on the ontological list, adding tables and chairs would be superfluous. Tables and chairs are already on the list, if ideas in the mind are and that is 'all' they are.

Again, in cases of reductive identification, our talk gets salvaged, since the original talk is true if the analysis containing the reductive identifier is (*see* Rational Choice Explanation: Philosophical Aspects).

Some Issues in Social Ontology

In this article, we are interested not in ontology in general but in some issues of social ontology in particular. Thought and discourse about society, whether lay or by the social scientist, raises these ontological questions too. In social thought and discourse, we appear to be speaking at least sometimes about social objects, social events, social process, social properties,

and social facts. However, as we saw for the general case above, it is one thing to admit that there are apparently such entities as France, the United Nations, and the Icelandic working class; social properties such as the property of being a mayor, being a charter member of the United Nations, or being an alter; or events like World War II or the Great Depression, but quite another to insist that they comprise a world of sui generis social objects, events, or properties, or whatever, and that they cannot be either eliminated or replaced or reductively identified with something more physically or psychologically respectable.

Often, the positions described in the relevant literature about social ontology are posed in terms of individualism and holism, sometimes with the qualifying adjective 'methodological' added. Sometimes, the positions are expressed in terms of facts; at other times, in terms of entities. Supervenience claims for the social ('the social supervenes on the nonsocial') are sometimes posed in terms of facts, at other times in terms of properties. It might be claimed that there are no irreducible social facts, that there are no irreducible social entities, or that there are no irreducible social properties. These claims are not equivalent. Rarely, if at all, is the distinction drawn in the relevant literature between the existence of social entities, social properties, and social facts. Does individualism in the social sciences eschew social entities but admit irreducible social properties? Should it eschew both? What should it say about facts? Might there be distinct versions of individualism, so that it comes in stronger and weaker forms?

I have already given or will give examples of social objects (Quinton, 1975–6), properties (Ruben, 1985), and facts (Gilbert, 1989). We are committed to these, unless we can show that these turn out to be cases in which we can, without loss, eliminate discourse about them entirely or replace such talk with other talk that makes no similar ontological commitment, or for which there is some reductive identification available.

What does 'social' connote? The basic application of the term relates to properties. What makes a property a social property? This is, I think, a difficult question. Some entity's having a property may have a social cause or social effect, without making the property itself social. For example, the soil's being eroded may be caused by inept bureaucracy, and may result in a stock market plunge, but none of that makes the property of being eroded a social property. Whether a property is social depends on something about the property itself, and not on the causes or effects of its being instantiated.

A property is social (roughly, without some refinements unnecessary here) if and only if from the fact that it is instantiated, it follows that at least two people exist and have an interlocking system of beliefs and expectations about one another's thoughts or actions (Ruben, 1985). Intuitively, I think this gets our idea of sociality about right, although one might just think that the requisite number should be greater than two. For example, the property of jointly carrying a stone is not social on this account, and rightly so, because two people might jointly carry the stone without either being aware of what the other was doing, oblivious to each other's very existence. Jointly carrying requires two people, but makes no necessary demands on their beliefs and expectations about one another. Purchasing a stone is, on the other hand, a social property on this account, and rightly so. No two people could be engaged in such a transaction without having a whole host of beliefs and expectations about the actions of the other.

Social entities, if such there be, are entities such that they have at least one social property essentially. France, for example, is essentially a nation; the Icelandic working class is essentially a class, a working class, and Icelandic.

No discussion of social ontology could be complete unless it brought to the fore actions and social events (including processes and states) as well, and in particular social actions (*see* Joint Action and Collective Action). Once the idea of a social property is to hand, we can use it to introduce the idea of sociality as applied to these other social ontological categories as well. For example, armed with a preexisting concept of change, a social event can be thought of as a change of any sort occurring to a social entity (e.g., France's increase in population) or a change in which a nonsocial object looses or gains a social property (e.g., Fiorello LaGuardia becoming mayor of New York). In a similar manner, a social fact could be either the fact that some social entity has, loses, or gains some property or the fact that some nonsocial object has, loses, or gains some social property (and we can add to this for whatever other kinds of facts there might be). So our idea of the social fans outwards, as it were, from our idea of a social property to other social ontological categories.

Social entities, if such there be, are never concrete, in the way in which earlier examples of entities were. Concrete entities, like Socrates and the tallest building in New York, are exclusive space occupiers, no two of which can occupy the same space at the same time.

Social entities, in general, are space occupiers; France occupies a certain region of physical space, and the Icelandic working class is physically in Iceland. However, social entities are not exclusive space occupiers. Suppose a diocese of the Catholic Church, call it 'Sancta Gallia,' occupies exactly the same geographical area as does France. It could even be that all and only French people were members of the diocese. If so, two distinct social entities, France and Sancta Gallia, would occupy the same physical space at the same time and have identical memberships. However, France $\neq$ Sancta Gallia.

What, then, of the three options we have available, for dealing with these apparent ontological commitments of our social discourse? Elimination of discourse about social entities would seem to involve an intellectual loss; we could not even in principle say things we would want to say without the discourse. (Imagine where we would be without sociology, history, anthropology, and a part of economics, for example.) The issue here is not just one of linguistic economy. It is not that we could not nearly as easily say what we want without social discourse but could only do so in an altogether more baroque fashion; rather, the claim is that we would have no way in principle of conveying what we want to say without that social discourse.

On the other hand, replacement does seem possible for some very simple and atypical examples. Another case that has been mentioned in the literature is 'The Jewish race is cohesive,' which appears to commit the speaker or thinker to an entity, the Jewish race. However, it is alleged, this is replaceable by 'Jews tend to marry other Jews,' and so on, and the latter makes no such ontological commitment to the existence of a race. 'The Jewish race is cohesive' behaves very much like 'The average

family has 3.2 members.' (See J.W.N. Watkins, 'Ideal Types and Historical Explanation' and 'Historical Explanation in the Social Sciences,' the first of which is reprinted in Ryan (1973), and both of which are reprinted in O'Neill (1973).) In this last example, the replacement does eliminate reference to such an entity as a race, but it is hardly a good example of the replacement of the social by the nonsocial more generally, since the property of being Jewish is (arguably) itself a social property.

I do not wish to claim that these replacements are never available; perhaps they do work for the case just mentioned. However, they are atypical, and it seems hard to see what replacement strategy would work for most other examples of social ontological commitment. What seems clear is that most talk or discourse about society is neither like talk about average families nor like talk of high stakes or like talk about the Jewish race, and hence are not candidates for a replacement strategy. Consider 'France is a charter member of the United Nations.' What sentence or sentences could replace this, yet convey the same information, at least in the sense of having the same truth value, and yet make no ontic commitments to a nation state and a world organization, or anyway to something else that is itself social? I can here only invite the reader to suggest candidates; I myself know of none. Replacement does not work as a general strategy for the social.

However, it is one thing to admit that there are such entities as France, the United Nations, and the Icelandic working class, or social properties like being a mayor or being an alter, and that they can neither be eliminated nor replaced, but quite another to insist that they comprise a world of sui generis social objects, events, or properties and cannot be reductively identified with something nonsocial, something more physically or psychologically respectable.

Many philosophers are drawn to the sort of reductive strategies that I mentioned above (Mellor, 1982). So various reductive strategies suggest themselves for social entities: France is (merely) a geographical area or a set of persons, and so on. No one can 'prove' that no reductive strategy could work. (It is not very easy to prove a negative existential claim.) What must be done, I think, is to evaluate each reductive strategy that suggests itself, one by one. My view is that no reductive strategy for social entities or social properties proves successful.

There are some general lessons one can draw, however, from the failures. The vague thought, 'Nations are identical with the people who make them up,' is ill-formed. In any reductive identification, names or definite descriptions must flank both sides of the identity sign. Therefore, for instance, 'France = ….' Well, identical to what? Candidates to occupy the space on the right-hand side of the identity sign will either be extensional or nonextensional entities, in the following sense. Sets and aggregates and landmasses are extensional, since, if a and b are sets or aggregates or landmasses, $a = b$ if and only if a and b have exactly the same members, components, or parts.

No extensional entity in this sense could be identical with France, for example, since their identity conditions will differ. France can remain numerically one and the same, in spite of variation in its citizenry, residents, or land occupancy. Even if one considered the set of all the actual citizens of France past, present, and future, or pieces of land that France ever occupied at any time, past, present, and future, there would be modal differences.

France could have occupied a different piece of land than it did, or had different citizens or residents than it did, but this could not be true of the set or aggregate of its citizens, residents, or lands. France is nonextensional; sets, aggregates, and landmasses are extensional, and no nonextensional entity could be identical with an extensional one. (In fact, this little argument that relies on modal differences assumes that 'France' and 'the set of …' both rigidly designate, since the identity required is identity across all possible worlds. I would argue that this assumption is true.)

If France is a nonextensional entity and no nonextensional entity can be identical with an extensional one, then what about some nonextensional entity for the reductive identification of France? Examples that come to mind are things like groups. Is France to be identified with the group of French people? The problem with this suggestion is that a group (in the relevant sense) seems itself to be a social entity, and hence hardly a candidate for the reductive identification of France with a nonsocial entity. What the determined reductivist of the social to the nonsocial needs to find is a nonextensional entity that is also nonsocial, for the reductive identification of France, and I am aware of no such obvious candidate.

What sorts of properties are ascribed to things when we think or talk socially? We can think or speak of either social entities or nonsocial ones like Socrates or the tallest building in New York and, in either case, ascribe social properties to them: 'France is a charter member of the United Nations' ascribes to a pair of social entities, France and the United Nations, the relational social property of being a charter member of; 'Socrates was a prisoner in Athens' ascribes to a nonsocial entity, Socrates, the relational social property of being a prisoner in.

The relational social property, being the mayor of, certainly figures in 'Fiorello LaGuardia was mayor of New York,' so on the relaxed view, we are committed to the existence of social properties. However, the sentence is not about that social property. It is only about Fiorello LaGuardia and New York. Therefore, on the austere criterion of ontological commitment to properties, the sentence commits us only to the existence of a person and a city, but not to the existence of a social property.

Even were we to adopt the austere view, there are a host of arguments, modeled on ones about properties generally, which do move us to talking about social properties, and so to the existence of social properties. Here are two such arguments: (1) if Iceland is capitalist, and if France is capitalist, then there is at least one social property, being capitalist, that Iceland and France share; (2) if being a mayor was a role LaGuardia enjoyed, then it follows that there is at least one social property, being a mayor. The argument instanced in (1) and (2) parallels one by Putnam (1980) on properties generally.

We talk and think about social properties, and I do not see how we could simply eliminate, or do without, such discourse. Nor do I know of any discourse that could replace discourse about social properties, and that lacked the ontological commitment to social properties but allowed us to convey the same information or messages. Therefore, on both the austere and relaxed views about the existence of properties, there are social properties, as well as nonsocial ones.

However, we can still ask, could each social property be reductively identified either with some specific nonsocial property or with some Boolean construction of such (suppose the set

of nonsocial properties was closed under disjunction, conjunction, complementation, and quantification)? I doubt whether this could be so. Let me give an indication of why I think this.

Some social properties are both nonvariable and weak, in the following senses. A social property *P* is nonvariable if and only if *P* is instantiated, then some specific set *s* of interlocking beliefs and expectations exists. For example, the social property of participating in the custom of drinking tea at breakfast is nonvariable, since the beliefs and expectations necessarily associated with it must be the set of beliefs and expectations about drinking tea at breakfast. A social property *P* is weak if and only if *P* is instantiated, then the interlocking beliefs and expectations are about types of events that are not themselves social. The social property of participating in the custom of drinking tea at breakfast is weak as well as nonvariable, since the beliefs and expectations associated with it concern drinking tea at breakfast, which is a nonsocial property. Since it does not follow from the fact that the property of tea drinking at breakfast is instantiated that atleast two people exist.

However, the most characteristic sorts of social property are neither nonvariable nor weak; they are variable and strong. Consider the property of being a mayor. If that property is instantiated, it is true that some set of interlocking beliefs and expectations must exist, but there is no one particular set that must exist. There is an indefinitely wide range of things that a mayor might be expected to do in various different societies or social settings. Therefore, the property is variable, in the sense of being indefinitely or infinitely variably realizable by many different sets of interlocking beliefs and expectations. (Whether a property could be reductively identified with a property that was itself an infinite disjunction of properties is an open philosophical question.) Moreover, the property is strong, since what a mayor is expected to do involves events and activities like receiving the ceremonial key to the city, opening formal events, representing the city at certain social or political functions, and so on, and these things are themselves social events or activities.

It seems to me that these two features of typically social properties make the reductive identification of social properties with nonsocial ones exceedingly implausible. One might have thought that a social property could be reductively identified with the nested and interlocking set of beliefs and expectations with which it is associated. However, since there is an indefinitely long list of such sets associated with each typically social property, and the contents of these beliefs and expectations themselves require the existence of further social events or activities, the reductive identification appears to fail.

Thus far, I have thought of reductive identification as being the obvious strategy for the individualist about the social, who rejects elimination and replacement, but two additional strategies should be mentioned. First, supervenience was once thought to offer an attractive alternative to reductive identification for the individualist. A set of *S* (say, social) properties was said to supervene on a set of *P* (say, nonsocial) properties just in case, for any two objects, there could be no difference in their *S* properties unless there was a difference in their *P* properties. Sameness of two things in respect of their *P* properties entails sameness of the two things in respect of their *S* properties. The thought was that supervenience would tie the social down to the nonsocial in a nonreductive way that would still appeal to an individualist about the social.

A version of supervenience called either 'global supervenience' or 'regional supervenience' was thought to be especially suitable for the case of social properties (Currie, 1984; Pettit and List, 2011). The property of being a coin worth £1 did not supervene just on that single coin's physical properties but would have to include the physical properties of many other things as well, for example, whatever the physical properties were of the banks and other institutions necessary for there being coinage, monetary value, and so on.

The idea was meant to be that supervenience could provide an alternative to reductive identity, which was thought to be too strong a demand to place on the relation of the social to the individual, but would still provide a strong enough relation to 'tie' the social down to the individual and hence could count as some legitimate form of individualism. That hope has proved to be elusive. The notion of supervenience described above is too weak; it entails at most the covariance of social and nonsocial properties and goes no way in capturing the individualist's intuition that the social depends on the nonsocial in some way in which the nonsocial does not depend equally on the social. Accounts of supervenience strong enough to capture the idea of the dependence of the social on the nonsocial and that tie the social down to the individual seem to collapse back into reductive identity; accounts of supervenience weak enough to escape the collapse into reductive identity are not strong enough to tie the social down to the individual in the desired way. Supervenience seems not to be able to do what was expected of it, an account of nonreductive individualism in the social sciences (see the essays in Savellos and Yalçin, 1995).

Finally, the idea that social entities are wholes, whose parts are individual persons, has often appealed to those with an individualist agenda. It is not obvious why this should count as a type of individualistic doctrine, as it is purported to be, unless one adds the thought that a whole is no more than the sum of its parts. The plausibility of this doctrine relies, in any case, on assuming that the standard logic of wholes and parts, mereology, would carry over to our understanding of social wholes and their parts. The analogy is specious. The relation of an individual to the social 'whole' of which it is said to be a part is quite unlike that of the relation of a part to a genuinely mereological whole. Here is one difference. In the case of a part and the mereological whole to which it belongs, the idea of a part is transitive: if *a* is a part of *b* and *b* is a part of *c*, then it follows that *a* is a part of *c*. The social case is not transitive: if a person is a member of a club and the club is a member of an organization of clubs, it does not follow that the person is also a member of the organization of clubs. The mereological relation of parthood and the social relation of membership behave quite differently logically and the clarity and precision of the former cannot be used to elucidate the latter. If we cannot use the mereological ideas of whole and part for the social case, it is quite unclear what the meanings of 'whole' and 'part' are in their purportedly social usage.

The overall lesson in this discussion, I submit, is that these terms, 'holism' and 'individualism', do not name unambiguous positions unless and until various distinctions are drawn, and the ontological claims of each position are clearly specified. (See S. Lukes, 'Methodological Individualism Reconsidered,' reprinted in Ryan (1973).)

See also: Atomism and Holism: Philosophical Aspects; Causation: Physical, Mental, and Social; Causes and Laws: Philosophical Aspects; Identity and Identification; Individualism versus Collectivism: Philosophical Aspects.

Bibliography

Currie, G., 1984. Individualism and global supervenience. British Journal for the Philosophy of Science 35, 345–358.

Gilbert, M., 1989. On Social Facts. Routledge, London.

Goodman, N., 1961. About. Mind 70, 1–24.

Lycan, W., Pappas, G., 1972. What is eliminative materialism? Australasian Journal of Philosophy 50, 149–159.

Mellor, D.H., 1982. The reduction of society. Philosophy 57, 51–75.

O'Neill, J. (Ed.), 1973. Modes of Individualism and Collectivism. Heinemann Educational Books, London.

Pettit, P., List, C., 2011. Group Agency. Oxford University Press, Oxford, UK.

Putnam, H., 1980. On Properties. Reprinted in Putnam H Philosophical Papers, vol. 1. Cambridge University Press, Cambridge, UK.

Quine, W.V.O., 1963. On what there is. In: From a Logical Point of View. Harper and Row, New York, pp. 1–19.

Quinton, A., 1975–6. Social objects. Proceedings of the Aristotelian Society 76, 1–27.

Ruben, D.-H., 1985. The Metaphysics of the Social World. Routledge and Kegan Paul, London.

Ryan, A. (Ed.), 1973. The Philosophy of Social Explanation. Oxford University Press, London.

Savellos, E., Yalçin, Umit (Eds.), 1995. Supervenience: New Essays. Cambridge University Press, Cambridge.

Social Protest

Abby Peterson, University of Gothenburg, Göteborg, Sweden

© 2015 Elsevier Ltd. All rights reserved.

This article is revision of the previous edition article by L. Tilly, volume 21, pp. 14397–14402, © 2001, Elsevier Ltd.

Abstract

Ordinary people have often confronted more powerful opponents expressing their grievances and demands for restitution. The forms or types of collective actions taken vary widely, ranging from more conventional strategies of political persuasion such as petitioning and legal demonstrations, to more disruptive tactics such as strikes, sit-ins, and occupations, to more violent acts that inflict material and economic damage and loss of life. Just as societies have transformed over time the forms of social protest have changed over time. Social protest from our earliest history to today is a potent and effective motor for social change.

From the year 1040 to well into the 1100s, women in western Sweden fought against the Cistercian Order's Christian evangelizing efforts. Women carried out their social protests employing imaginative strategies of disobedience, entering Varnhem's Abbey dancing and defecating on the altar. Their persistent social protests generated considerable correspondence between the monks and the sitting Popes, lamenting over the difficulties they faced in their encounters with the rebellious Scandinavian women. On 16 December 1773, after officials in Boston refused to return three shiploads of taxed tea to Britain, a group of colonists boarded the ships and destroyed the tea by throwing it into Boston Harbor, what came to be known later as the Boston Tea Party. The social protest by the 'Sons of Liberty' has become an iconic event in American history. These examples of social protest remind us that social protest has been with us since the dawn of history. Ordinary people have often confronted more powerful opponents expressing their grievances and demands for restitution. Acts of social protest are acts of contention by people who lack the resources on their own to bring about social changes. In short, social protest is contentious collective action. In the nineteenth century social protests became increasingly organized and coordinated. The first international May Day demonstrations in 1890 were widely celebrated across Europe and continue to do so to this day. May Day demonstrations were the 1-day each-year events globally when organized labor could present its goals to the public; an 8-h workday was the first concrete demand, later in the 1890s to be supplemented by improved working conditions, universal suffrage, peace among nations, and international solidarity. The latter decades of the twentieth century witnessed a resurgence of social protests and widespread civil unrest. In August 1965, a questionable arrest of a black man in Los Angeles, California, triggered a 6-day civil disturbance. What has become commonly known as the "Watts Riot" resulted in 34 deaths, 1032 injuries, 3438 arrests, and over $40 million in property damages. Researchers found that a significant percentage of the white population surveyed in Los Angeles after the event defined the riot as an act of social protest against prejudice and discrimination against the black population (Jeffries et al., 1971). Social protests in Paris during May 1968 commenced with a series of student occupations and strikes that broke out at a number of universities and lyceés in Paris, following confrontations with university administrators and the police. The de Gaulle administration's attempts to quell those strikes by police action only inflamed the situation further, leading to street battles with the police – barricades were erected, cobblestones and Molotov cocktails flew through the air. The student protests were followed by the first wildcat general strike ever in France, which involved 11 000 000 workers, over 22% of the total population. The social protests reached such an extent of scope and intensity that government leaders feared civil war or revolution. However, when new elections were held in June, the Gaullist party emerged even stronger than before. This was the revolution that never was, in contrast to the Arab Spring. In 2011, massive social protests against authoritarian regimes swept across North Africa and the Middle East. The unexpected revolutionary events of the Arab Spring led to the overthrow of the regimes in Tunisia, Egypt, and Libya. After the so-called silent revolutions in Eastern Europe and the Soviet Union during 1989–91 coordinated social protests in authoritarian and semiauthoritarian regimes have emerged in former Soviet states, and most recently in North Africa and the Middle East. From the 1990s onward, transnational social protests with widened scopes of contention have increasingly appeared. More and more groups have engaged in social protest, not least we saw a dramatic increase in far-right antiimmigrant protests. A German neo-Nazi group calling themselves 'Die Unsterblichen' (The Immortals) orchestrated flash mob protests in cities all over Germany during spring 2011. In a social protest demonstration on 1 May in Bauzen, Germany, hundreds of black-clad figures with white masks suddenly converged in a street carrying torches and signs with extreme nationalist slogans, then dissipated before a major police presence could arrive. On 1 August 2012 the British Defense League and the Swedish Defense League joined forces in an act of social protest in Stockholm against what they called the 'islamization of western societies'. Their rally was met with an even larger number of counterprotesters in an act of social protest against the presence of right-wing extremists in the streets of Stockholm.

Social protests are collective acts of defiance. Social protest can be either violent or peaceful, dramatic or more routine. Social protest can include even what many would call unsavory demands, for example, the expulsion of immigrants. Acts of social protest can be more or less spontaneous and short-lived and acts of social protest can also be coordinated and sustained over time. This small selection of examples of social protest highlights that we are confronted with definitional and

situational problems. What is the historical legacy of social protest? How can we delineate social protest from social movements; social protest in relation to political protest; social protest in relation to major civil unrest or riots; and last, social protest in relation to revolutions?

Social Protest Defined

Ralph Turner (1969) defined an act of social protest as including all of the following elements:

> [t]he action expresses a grievance, a conviction of wrong or injustice; the protesters are unable to correct the condition directly by their own efforts; the action is intended to draw attention to the grievances; the action is further meant to provoke ameliorative steps by some target group; and the protesters depend upon some combination of sympathy and fear to move the target group in their behalf.
> (p. 816)

Turner calls our attention to social protest as an act of expressing grievances. Social protest is collective action that is meant to both call attention to grievances and seeks to confront authorities to take steps to address their grievances. If we break down the constituents included in the act of social protest we can delineate four elements. First, social protest is intentional and purposeful collective action. Second, social protest is a resource employed to signal protesters' concerns or grievances. Third, social protest is a form through which groups challenge authorities (Gamson, 1975), as such social protest is political performance. And last, social protests are the means toward achieving redressal of grievances.

The 'Political' in Social Protest

William A. Gamson (1975) in his classic work *Strategy of Social Protest* defined the groups that engage in social protest as an excluded group "seeking the mobilization of an unmobilized constituency … against an antagonist (lying) outside of its constituency" (pp. 16–17). Gamson's definition emphasizes that it is the disenfranchised that engage in social protest. Similarly, Francis Fox Piven and Richard Cloward (1977) argue that social protest is rooted in political exclusion and attempts to renegotiate the terms of political access in order to address grievances. Subsequently, social protest is inherently a political act, which seeks to implicitly lay its claim for political representation in a situation where the state regulates the political environment. While social protest is implicitly a claim for political representation, the collective action is enacted outside of the conventional channels of political representation. This is the element of protest in the conceptual duality of social protest. Social protest challenges state authorities to take political action on behalf of the protesters – social protest is political protest.

Who Protests

Which are those that engage in social protest? Classically, social protest was the sole means for subjugated groups to voice their concerns and demand restitution. Historically, it was slaves, peasants, laborers and crofters, women, the poor and disadvantaged, Black people in America, unemployed, religious minorities – the list is long – who, often in alliance with more powerful groups, engaged in acts of social protest. However, this story must be updated. David Meyer and Sidney Tarrow (1998) have proclaimed that in Western societies social protest has become so common and sponsored by increasing numbers and so many types of organizations that we can call them 'movement societies'. According to these authors, since the 1970s protest on the streets is no longer the sole prerogative of left parties or marginalized groups. The people who protest today are not necessarily the disenfranchised. On the contrary, particularly among those engaged in so-called new social movement politics, that is, the peace movement, environmental movement, women's movement, human rights movement, etc., we find a disproportionate number of well-educated, middle class professionals (Wallace and Jenkins, 1995). Furthermore, protesters today are not necessarily alienated from political institutions. Those who are more active in conventional politics are also more active in social protests. This development is not limited to Western democracies. For example, those who bore the brunt of social protests that brought down the Mubarak regime in Egypt 2011 were for the most part university-educated young people. Whereas in the past social protest was the means for the politically 'resource poor' to influence the polity, increasingly, the politically 'resource rich' are supplementing their political power potential with acts of social protest. Wide segments of the population are now engaging in acts of social protest.

Why Do People Protest

Social scientists at the *fin de siècle* regarded all social protest as deviant behavior, as irrational actions to be feared. This paradigm lived on well into the twentieth century where social protests were viewed as indicative of breakdowns or 'cracks' in the social fabric, and were alarming signs of felt grievances and discontent. By the 1970s, the theoretical focus had shifted from why people are aggrieved to why aggrieved people take part in social protest. Rather than explaining the irrational behavior of discontent, researchers now began to study the rational and instrumental aspects of why people protest. Why do people take the time and effort, and sometimes even the considerable risks, to participate in social protests?

In a classical study quantitative study of political protest in five Western democracies Barnes et al. (1979) that the among the standard social variables associated with a predisposition to protest, age, gender, and education mattered the most. The young, the better educated, and males were more likely to protest than the elderly, less educated and women.

Social psychologists investigate the motives behind protest participation and demonstrate that instrumental reasoning, identification, emotions, and ideological factors all come together to form a motivational constellation in response to the question as to why people protest (e.g., Klandermans, 1997). Bert Klandermans highlights three sets of motives, which inspire people to protest. First, we have the instrumental motive. People take part in a social protest in order to exert

influence, that is, to demand change for a perceived dissatisfaction or grievance. Second, a strong identification with a group enhances an individual's motivation to protest together with that group. Third, ideological motives for engaging in social protest include individual's search for meaning and most importantly their desire to publically express their views. Ideological motives are most often augmented by emotions, the affective component of ideology, for example, feelings of moral outrage over injustice. People are often led to social protest because they are angry and social protest provides them with the opportunity to voice to others those feelings. Klandermans emphasizes that all three sets of motives are more or less germane in understanding an individual's motives to participate in protest given that in some instances of protest one or two of the sets may be more important. For example, feelings of moral outrage can compensate for a situation where there is little hope of influencing change. However, he argues that we must consider all three sets of motives in some sort of juxtaposition in order to understand why people protest. Going beyond social psychological theories of protest, rational choice theorist Karl-Dieter Opp (2009) offers a model for accounting for social protests, which he calls a 'structural-cognitive model' in which he argues that structural factors and individual incentives mutually reinforce one another.

The 'Social' in Social Protest

Social protest brings people together for a common cause in concerted action, as such social protest is intrinsically social. The social processes of coming together, interacting to form common bonds, talking and assigning meaning to their actions are integral to social protest. The underlying precondition for social protest, as well as a consequence of protesting together with others, is the formation of a sense of 'we', a sense of being part of a collectivity in opposition to a jointly designated adversary. In order to understand how individuals and groups make sense of their protest actions we must grasp the processes through which a collective becomes a collective. Scholars within the so-called collective identity paradigm focus these social processes of collective protest. The late Alberto Melucci (1996), the leading theorist in this constructionist approach to social protest analysis, focused on how protesters define in cognitive terms the field of possibilities and limits they perceive while at the same time activating their relationships so as to give sense to their 'being together' in protest and to the goals they pursue. Melucci emphasized the emotional investments which enable individuals to feel like part of a common unity, a sense of we; desires and feelings, love and hate, trust and fear are all part of a body acting collectively. This is the underlying essence of the 'social' component in social protest.

The Forms of Social Protest

Demonstrations, strikes, occupations, vigils, mock funerals, marches and rallies, blockades, barricades, petitions, boycotts, cross-burnings, picket lines, canvassing, hunger strikes, riots, and guerilla theater, all come under the category of forms or types of social protest and are unconventional methods of political participation. The forms or types of collective actions taken vary widely, ranging from more conventional strategies of political persuasion such as petitioning and legal demonstrations, to more disruptive tactics such as strikes, sit-ins, and occupations, to more violent acts that inflict material and economic damage and loss of life. Sidney Tarrow (1998) differentiates three forms of protest actions: conventional, disruptive, and violent, acknowledging that contentious politics frequently combines all three elements. For example, a faction within a major peaceful demonstration can break out and occupy a square, while another faction can seek violent confrontation with the police. This is a common characteristic of large demonstrations today – their unpredictability. For example, the social protest demonstrations against austerity measures in Athens, Greece during 2011 most often led to violent riots between police and protesters.

We can also divide the forms of social protest between 'low-risk' actions and 'high-risk' actions. The latter are forms of protest that are illegal and the activists risk arrest, imprisonment, and/or bodily harm. High-risk forms of social protest run the gauntlet between nonviolent forms of civil disobedience and violent forms of protest, for example, riots, direct confrontations with adversaries, etc. While the vast majority of the forms of social protest are so-called low-risk actions, contemporary social protest often includes the more dramatic high-risk militant actions of small highly dedicated protest groups (Peterson, 2001).

Scholars use the concept of repertoires of contention to refer to the recurrent, predictable, and fairly narrow 'toolkit' of specific protest tactics used by a set of collective actors in a particular social protest (Tilly, 1995). However, even if tactical repertoires evolve slowly and protesters are more likely to select tactics with which they are familiar, innovations do come onto the scene. Sometimes, it is the novelty of the form of social protest, which can contribute to the success of the protest. For example, Doug McAdam (1988) shows that tactical innovations on the part of US civil rights activists such as sit-ins and freedom rides were effective because they caught authorities off guard. Innovations in the forms of social protest do not only catch authorities off guard, they also capture media attention spreading the messages of the protesters. Innovations are in short newsworthy.

The Historical Transformation of the Forms of Social Protest

The work of the late Charles Tilly (e.g., 1995) has made path-breaking contributions toward our understanding of the historical development of the forms of social protest. Tilly has introduced the notion of "repertoires of contention" which are historically situated "ways that people act together in pursuit of shared interests" (1995: p. 41). Sidney Tarrow (1998) emphasizes that repertoire is both a structural and a cultural concept, "involving not only what people *do* when they are engaged in conflict with others but what they *know how to do* and what others *expect* them to do" (p. 30, emphasis in original). Hence, repertoires are historically and contextually situated collections of forms of social protest. Just as societies have transformed over time the forms of social protest change over time.

Prior to the nineteenth century, social protests took the form of bread riots and grain seizures, peasant revolts that took the form of peasants brandishing pitchforks marching to 'occupy' land usurped by powerful landowners, religious rebellions, funeral riots, and so on. Social protest focused local issues and the forms of social protest were direct, often violent, and inspired by their claims. The traditional repertoire of contention was parochial and episodic in that the protest actions converged in a single community, the protesters took direct action that took particular forms that varied according to the issue pursued, the locality where the protest took place, and the group of people protesting (cf Tilly, 1995). For example, a group of hungry citizens could seize the grain from a merchant whom they thought was hoarding. The death of someone in the community could unleash a riot at the funeral. American colonists could parade with effigies of British authorities, and then burn the effigies.

During the nineteenth century, alongside of the older traditional forms of social protest new forms of social protest emerged that were more organized and engaged broader segments of the population, and that were more general and wider in scale. Compared with the traditional repertoire the new repertoire of contention, according to Tilly (1995), was cosmopolitan, modular, and autonomous, what he calls "modular collective action". The new forms of social protest were:

> *cosmopolitan* in often referring to interests and issues that spanned many localities or affected centers of power whose actions touched many localities. They were *modular* in being easily transferable from one setting or circumstance to another. … They were *autonomous* in beginning on the claimants' own initiative and establishing direct contact between claimants and nationally significant centers of power.
> (p. 46; emphasis in original)

The nineteenth century ushered in demonstrations, rallies, boycotts, strikes, and barricades and the nation-state became increasingly the target of the social protest. Social protest engaged wider segments of society, held together by new emerging forms of association and informal social networks. According to Sidney Tarrow (1998: pp. 43ff), informal networks underpinned the new emerging protest organizations and these networks were crucial for spreading the new forms of social protest. The expansion of the print media, together with growing literacy in Western societies, further helped diffuse the new forms of social protest.

During the nineteenth century, the rise of the modern state was supplemented with a new invention for social protest – the social movement. When social protests against powerful opponents are initiated, coordinated, and sustained we are confronted with this new invention of modernity. Not all social protests are part of a social movement. Tarrow (1998: p. 2) reserves the notion of social movement for sequences of social protest acts that are "based on underlying social networks and resonant collective action frames, and which develop the capacity to maintain sustained challenges against powerful opponents". Tarrow emphasizes that it is the capacity to sustain episodes of social protest that distinguishes social movements from earlier forms of contention. During the nineteenth century nationalist movements, labor movements, abolitionist movements, women's movements, temperance movements, and so forth, erupted on the national scenes throughout the Western world. Social protests became progressively organized in new ways that offered social protest stability and enhanced bases of power.

Social Movements

During the last 30 years, research on social movements has expanded prolifically and so also have the definitions of social movements legion. Snow et al. (2004, see for an excellent overview of social movement research perspectives) offer the following broad definition:

> social movements can be thought of as collectivities acting with some degree of organization and continuity outside of institutional or organizational channels for the purpose of challenging or defending extant authority, whether it is institutionally or culturally based, in the group, organization, society, culture, or world order of which they are a part.
> (p. 11)

Social movements share some of the key characteristics of social protest, that is, collective action outside of traditional institutional political channels that challenges authorities to take some action on behalf of the group. However, we cannot collapse the two binary concepts. Isolated incidents of social protest are not social movements. Only by sustaining collective action does an actor turn a contentious episode – a series of related and coordinated social protests – into a social movement. The key distinction between social movements and social protest is that the former is collective action sustained over time. Social protest is nevertheless a key, if not the key, element in social movements.

Contemporary Transnational Social Protest

We have seen a transformation of social protest from highly localized sites of protest to modernity's national scenes of protest. In this latter era, social protest was even coordinated and organized internationally. Abolitionists, women's suffragettes, peace activists, and not least the early labor movement activists met with their counterparts from other nations. The latter half of the nineteenth century saw a flora of international conferences that provided inspiration for the protest activists that participated. However, their social protest actions remained directed to their respective nation-states; they brought their grievances to their nation-states. We are witnessing today a new development in the scale and scope of social protest – social protest has gone increasingly global (see della Porta and Tarrow, 2005, for an excellent overview). This new form of transnational social protest is underpinned by the development of electronic media. The media spreads the influence of social movement actions and disseminates their messages within a mediated publicness where to paraphrase Todd Gitlin (1980), 'the whole world is watching' – simultaneously. Protesters today are more and more staging their actions as media events, the aesthetization of politics, in order to occupy this space of visibility which is both opening

and expanding in our historical epoch of what we can designate as media modernity (Peterson, 2001). The struggle to make oneself heard or seen is not a peripheral aspect of the social and political upheavals of the modern world; on the contrary, it is central to them. The struggle to capture the attention of media audiences is essential for an understanding of contemporary social protest. Alberto Melucci (1996) argues that collective action itself becomes a new medium and intervenes in the public discourse, the domain in which the words of protesters can be heard, by its interaction with the media system.

The global or planetary nature of mediated publicness in media modernity is revealed in the ability of media images and messages to act as a catalyst for action that spills well beyond the immediate locales in which events occur. Events taking place on the other side of the globe can serve as a rallying cry for the action mobilization of individuals and groups far removed from the spatial-temporal context in which these events initially occurred. The social protest event staged by 'Pussy Riot' in Moscow 2012, the subsequent arrest and conviction of three of the young women, gave rise to actions of social protest across the Western world.

The transnational environmental movement, understood in its broadest sense, offers us perhaps the clearest case of the widening scope of social protest. The environmentalist discourse has become increasingly couched in universal claims. Environmental degradation, it is asserted, is a problem that humankind faces in common. And indeed, thanks to industrialization and the spread of Western technologies across the globe, the physical threats to the environment are becoming more similar the world over. People throughout the world are finding themselves 'in the same boat', and this 'boat' is limited in that we have reached, and some would argue, overextended, the limits of our planetary system. Hence, green political thinkers have developed a rhetoric of globalism in order to promote and engender a process of planetary identification, challenging people to identify themselves as 'citizens of planet Earth' or as 'astronauts on spaceship Earth' rather than seeing themselves as citizens of some country, members of an ethnic group, comrades within a class. From the 1990s onward, environmental activists and governments alike see environmental issues as more often as not transcending national borders. Hence a new politics of environmentalism has been ushered in which transnational social protests direct their demands for redressal to both transnational political bodies and nation-states.

World Social Forums, the global environmental movement mentioned above, International Campaign to Ban Landmines (ICBL), are just some examples of social protest that are engaging protesters across the globe. While not neglecting their national contexts, these protesters are also challenging transnational authorities. Major demonstrations have been held in conjunction with, for example, economic summits, EU summit meetings, and global environmental summits. The global parameters of social protest mobilizations today have encouraged disparate groups and individuals to participate in mass demonstrations. Social protest groups, action networks, and individuals are increasingly pooling their resources in order to make their challenges heard and to reach their goals. Diversity is a major characteristic of contemporary social protest demonstrations, which are often described as multigenerational, multiethnic, multiclass, and multiissue. These temporary collective action mobilizations represent a broad spectrum of causes and goals. Participants represent a variety of issues and a wide diversity of targeted adversaries. The combination of groups and participants coming together creates a powerful impression and an impact, which is perhaps out of proportion with their individual strengths. The melding together of the various groups into one large body implies power, and attracts attention and publicity, which, in turn, draws more and more participants. Today, a single demonstration can (and most often does) include a cacophony of grievances expressed by a wide range of social protesters.

Social Protest and Revolutions

The cases of regime change in Eastern Europe in 1989 and the Soviet Union in 1991 were moments when weak states faced weak societies. While these countries more or less lacked civil societies from which social protest is traditionally mobilized, nevertheless massive social protests were launched against communist regimes in economic and political crisis. With the exception of Romania, nonviolent social protest toppled the regimes in Poland, Hungary, Czechoslovakia, and East Germany. Albert O. Hirschman (1993) employed his famous concepts 'exit and voice' to explain that while for nearly four decades exit had undermined voice in the German Democratic Republic the tumultuous events in 1989 led to a situation when many East Germans found the road back from exit and apathy to voice and public action (see also Opp et al., 1995). After a broad popular protest mobilization, on 11 March 1990 Lithuania became the first of the Soviet republics to declare its independence; Moscow first recognized this proclamation on September 1991 (following the abortive coup in Moscow). The remaining republics would take the Lithuanian example and one by one they declared their independence from the Soviet Union.

Color revolutions or Flower revolutions are the names that have been given to the protest mobilizations, which developed in postcommunist societies in Central and Eastern Europe after the year 2000. The mobilizations have used (mostly) nonviolent means to protest against governments seen as entrenched and authoritarian. In four successive countries, massive street protests followed disputed elections and led to the resignation or overthrow of leaders considered by their opponents to be repressive. They have adopted a color or flower as their symbol, which has embodied their nonviolent strategy.

While the peaceful revolution in Yugoslavia in 2000 did not adopt a color or specific symbol (aside from the clinched black fist), it is often referred to as a starting point for the color revolutions to come. The youth movement Otpar (Resistance) initiated a series of whistle blowing mass demonstrations in Belgrade, which led to the overthrow of Slobodan Miloseviç. Members of Opar have inspired and trained activists in related youth movements. The Centre for Applied Nonviolent Action and Strategies was founded in 2001 by some former Otpor members. This international nongovernmental organization has disseminated the lessons learned from their successful nonviolent social protests through scores of trainings and workshops for prodemocracy activists and others around the

world, including Georgia, Ukraine, Krygzstan, the Maldives, and Egypt. Forms of social protest can be transferred across borders, not only through media but also by direct contact with fellow protest activists.

The commonalities between these mobilizations of social protest were striking. In all of the countries, massive social protests followed disputed elections, which were the result of extensive grassroots campaigning and coalition building among the opposition. In all of the countries, youth resistance networks played a central brokering role as initiators and coordinators for the protests. Their participants used mostly nonviolent strategies to challenge the incumbent leaders. Another characteristic of the social protests was their simple straightforward message: the incumbent regime was evil and must be replaced.

Massive social protests, even in authoritarian or semi-authoritarian countries, can under favorable political circumstances give rise to revolutions. However, revolutions are not merely aggregates of social protest actions. Multiple centers of political authority are created transforming the conflict between challengers and members of the polity into a struggle for power. The ensuing conflict can then (at best) lead to a revolutionary outcome, which is the transfer of state power to new sets of actors. The examples of social protests mentioned above in some cases led to effective transfers of power, in others only partly successful revolutionary outcomes. However, in all of these cases of social protests demanding democracy and measures to come to terms with what was perceived as entrenched governmental corruption, the outcomes of the social protests were regime changes. Social protest is collective action that challenges authorities to redress their grievances; it is not by itself a bid for state power. Social protest in democratic contexts seeks most often policy reforms or group recognition. However, in authoritarian and semiauthoritarian contexts social protests can lead to revolutionary outcomes – political upheaval and the overthrow of regimes.

Conclusions

Social protest, the use of unconventional means of political persuasion, is a fundamental feature of democratic societies. Politics 'in the streets' is a necessary complement to parliamentary politics. And today – between elections – people are increasingly using this channel of political participation in order to make their voices heard. In nondemocratic countries, social protest emerges in the windows of opportunity that appear in even these societies. Activists seize these opportunities and stage social protest actions that have proven to be powerful enough to overthrow regimes. Social protest from our earliest history to today is a potent and effective motor for social change. Social protest witnesses the power of outsiders to influence the course of history.

See also: Civil Rights Movements; Civil Wars; Collective Behavior, Sociology of; Collective Identity; Crowds in History; Cuban Revolution, The; Environmental Movements; Ethnic, Racial, and Nationalist Movements; French Revolution, The; Gay and Lesbian Movements; Identity Movements; Lesbian and Gay Activism in Latin America; National Liberation Movements; New Media, Political Mobilization, and the Arab Spring; Peace Movements; Political Protest and Civil Disobedience; Political Repression; Queer Anarchism; Revolution, Theories of; Revolution; Revolutions of 1989–90 in Eastern Central Europe; Revolutions, History of; Revolutions, Sociology of; Social Movements and Political Violence; Social Movements, Geography of; Social Movements, History of: General; Social Movements: A Social Psychological Perspective; Social Protest and New Media; Strikes; Terrorism; The American Conservative Movement; Urban Conflicts; Youth Movements.

Bibliography

Barnes, S.H., Kaase, M., et al., 1979. Political Action: Mass Participation in Five Western Democracies. Sage, London.

Gamson, W.A., 1975. The Strategy of Social Protest. Dorsey Press, Homewood, IL.

Gitlin, T., 1980. The Whole World is Watching. University of California Press, Berkeley.

Hirschman, Albert O., 1993. Exit, voice, and the fate of the German Democratic Republic: an essay in conceptual history. World Politics 45 (2), 173–202.

Jeffries, V., Turner, R., Morris, R., June 1971. The public perception of the Watts riot as social protest. American Sociological Review 36, 443–451.

Klandermans, B., 1997. The Social Psychology of Protest. Blackwell, Oxford.

McAdam, D., 1988. Freedom Summer. Oxford University Press, New York.

Melucci, A., 1996. Challenging Codes: Collective Action in the Information Age. Cambridge University Press, Cambridge/New York.

Meyer, D.S., Tarrow, S., 1998. A movement society: contentious politics in a new century. In: Meyer, D., Tarrow, S. (Eds.), The Social Movement Society: Contentious Politics for a New Century. Rowman & Littlefield, New York, pp. 1–28.

Opp, K.-D., Voss, P., Gern, C., 1995. Origins of a Spontaneous Revolution: East Germany, 1989. University of Michigan Press, Ann Arbor.

Opp, K.-D., 2009. Theories of Political Protest and Social Movements: A Multidisciplinary Introduction, Critique, and Synthesis. Routledge, New York.

Peterson, A., 2001. Contemporary Political Protest: Essays on Political Militancy. Ashgate Publishers, Aldershot.

Piven, F.F., Cloward, R., 1977. Poor People's Movements. Pantheon, New York.

della Porta, D., Tarrow, S., 2005. Transnational Protest and Global Activism. Rowman and Littlefield, Lanham, MD.

Snow, D.A., Soule, S.A., Kriesi, H., 2004. Mapping the terrain. In: Snow, D.A., Soule, S.A., Kriesi, H. (Eds.), The Blackwell Companion to Social Movements. Blackwell, Oxford, pp. 3–17.

Tarrow, S., 1998. Power in Movement. Social Movements and Contentious Politics, second ed. Cambridge University Press, Cambridge, UK.

Tilly, C., 1995. Popular Contention in Great Britain, 1758–1834. Harvard University Press, Cambridge, MA.

Turner, R., December 1969. The public perception of protest. American Sociological Review 34, 815–831.

Wallace, M., Jenkins, J.C., 1995. The new class, postindustrialism, and neo-corporatism: three images of social protest in the western democracies. In: Jenkins, J.C., Klandermans, B. (Eds.), The Politics of Social Protest. Comparative Perspectives on States and Social Movements. University of Minnesota Press, Minneapolis, MN, pp. 3–13.

Social Protest and New Media

Sandra González-Bailón, Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Digital media are helping people self-organize and coordinate massive protests in the absence of formal organizations. This is challenging old theories of collective action, and encouraging research on large-scale communication networks. This article gives an overview of digitally enabled protests, and assesses the evidence on how new media facilitate their spread and coordination.

Definition of Concepts

Protests are a form of collective action that emerge to influence the decisions of governments or organizations by means of direct action (Opp, 2009). Examples include demonstrations, petitions, boycotts, sit-ins, or street blockades. New media encompass a range of digital technologies that allow users to engage in social interactions on a global scale, and be exposed to information and behavior, often in real time. Digital technologies are facilitating new forms of protest like hacktivism, denial of service (DoS) attacks or flash mobs (Chadwick, 2006; Earl and Kimport, 2011). These protest tactics rely on decentralized networks that operate in the absence of hierarchical organizations, a feature that makes them more flexible and adaptable; however, the lack of formal organizational structures also means that protests are more transient and volatile. This article offers an overview of theories and findings about how social protests emerge in the digital age.

Changes in the Logic of Collective Action?

Some scholars argue that new media are changing the scale of protests, but not the mechanisms that make them grow. Digital technologies, the argument goes, allow protests to grow bigger and faster, as when petitions are signed by thousands of people in a matter of hours; but they still rely on the same building blocks, namely, individuals solving the trade-off of whether it is worthwhile to participate in the collective effort. An increasing number of researchers, however, suggest that new media are stirring deeper change and reconfiguring the logic of collective action; this, they argue, requires a change of theoretical model (Bennett and Segerberg, 2012; Bimber et al., 2005; Lupia and Sin, 2003). The following three sections offer an overview of the principles that define the classic approach to collective action (i.e., the individual-level assumptions on how people face the decision to join a protest); how individual decisions concatenate to generate the cascading behavior that often accompanies the emergence of protests (i.e., what underpins the critical mass associated to successful protest events); and how new media are changing behavior on these two levels (i.e., the level of individual decisions and the level of collective dynamics).

The Classic Approach to Collective Action

The starting point of all classic approaches to collective action is the social dilemma that often arises during the provision of public goods. Public goods are nonexcludable, that is, everybody can enjoy them once they exist, even if they did not contribute to their provision. The assumption of classic approaches is that actors are rational, that is, they behave in accordance to the costs and benefits of their actions. What goes in their best interest in the case of collective goods is to free-ride. However, if all actors think the same way, no public good is provided and everybody is worse off. What this dilemma captures, in brief, is that individual and collective interests often go in opposite directions. Classic approaches to collective action try to understand how this dilemma is solved.

The most influential approach in this classic tradition is *The Logic of Collective Action* by Mancur Olson (1965), published decades before the Internet and new media were so widespread and ubiquitous. Olson identified three key elements that help solve social dilemmas. The first two are sanctions and selective incentives, which are intended to modify the cost–benefit equation. Sanctions punish people who do not comply with the provision of public goods (i.e., participation in protests), and selective incentives are intended to stimulate rational individuals toward contribution. The third element is related to the size of the organizations providing those incentives: to enforce sanctions and apply selective incentives small groups function better than large groups because they make it easier to monitor how people behave and punish or reward them accordingly.

This approach is built on the assumption that individuals are driven by selfish motivations, and that we need to build a structure of incentives or punishments to redirect that selfish rationality toward collective goals. However, alternative approaches that relax these assumptions have also been proposed. Sociologists like Jon Elster emphasize normative behavior, or the role that social norms play as a motivational force that is irreducible to rationality (Elster, 1989). People feel compelled to cooperate because breaking that expectation of behavior generates feelings of shame. Norms of cooperation can actually emerge from long-term self-interest: social dilemmas are not isolated events, and selfish behavior is often penalized in future interactions, which invites actors to reconsider their free-riding inclination. These strategic and temporal

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95076-1

dynamics are modeled in another classic work, *The Evolution of Cooperation*, which shows that cooperation produces more benefits in the long run (Axelrod, 1984). These two alternative approaches highlight an important element that is missing in Olson's approach, namely that individual decisions are interdependent – people in the real world do not make their decisions in isolation, but placed in a social context where influence links their actions.

The Importance of Interdependent Decisions

Acknowledging the importance of interdependent decision-making has a crucial effect on how we conceptualize collective action. As with other collective phenomena, the patterns observed on the aggregate level (i.e., the number of people joining a protest) cannot be directly derived from the motivations at the individual level. In the book *Micromotives and Macrobehavior*, Schelling (1978) introduces the notion of a critical mass, that is, the point beyond which the chain reaction of people driving other people to join a protest becomes self-sustaining. Thermostats are often used as a metaphor of this type of process: they regulate the temperature in a room by registering changes and responding to them; in a similar fashion, the level of contributions to a social protest responds to the contributions registered, activating a feedback loop that – as with thermostats – can be positive (and bring the protests to success) or negative (and bring the protests to failure).

The interdependence of decisions creates networks of interaction that mediate the way in which feedback mechanisms unfold. Threshold models offer one of the most prominent theoretical approaches to these network dynamics (Granovetter, 1978; Granovetter and Soong, 1983). The notion of threshold refers to the intrinsic propensity of people to join a protest, and it measures the fraction of other people that need to be already involved for someone to decide in favor of joining as well. It can be measured on a global level, as when we take into account how many protesters are already in the streets; or it can be measured on a local level, that is, as when we count the number of activated people in our circles of friends and relatives. Networks define the smaller group of reference to which people pay attention. Compared to Olson's approach, thresholds introduce more heterogeneity in the population of rational actors: some will have low thresholds (they will be first-movers or leaders), whereas others will have high thresholds (and will tend to be followers).

Threshold models try to understand how the diversity of intrinsic inclinations to participate in a collective effort impact on aggregated outcomes (Watts and Dodds, 2010). The ultimate number of people mobilized actually depends on how thresholds distribute in the population. It also depends on how networks mediate the process: two actors with the same propensity to participate might join at different times if one is connected to a larger number of people or if the group of reference behaves differently. Attaining a critical mass on the collective level depends on being able to activate a sufficiently large number of low threshold actors that are also well connected in the overall network structure. Again, this approach to collective action adds important nuances to the picture drawn by Olson's classic approach: individuals' self-interest and the incentives designed to lure them into collective action are not enough to understand the emergence of protests; networks of interaction, and how they allow information and behavior to diffuse, are also important elements of this collective dynamic.

The mediating effect of networks is particularly relevant in this era of new media and digital technologies, when connections and interdependence are more prevalent than ever. Research on these network effects is also becoming easier to conduct: digital technologies give protesters better tools to organize their actions; but they are also providing researchers with better data to analyze more accurately how episodes of collective action emerge.

The Digital Approach to Collective Action

Most approaches to digitally enabled collective action can be summarized with the idea of "organizing without organizations" (Shirky, 2008). However, as threshold models have long tried to highlight, this form of self-organization has been observed in many empirical settings that precede the existence of new media. Massive protests can emerge in the absence of formal, hierarchical structures as long as there are networks of communication that facilitate the diffusion of information and behavior. Social networks have long channeled that sort of dynamic (Gould, 1991; Lohmann, 1994). What is intrinsic to new media is the speed and scale at which diffusion can happen. Online networks and mobile technologies make exposure to information constant; this means that the spontaneous emergence of protests takes significantly less time, making it easier to attract the attention of a larger number of people.

However, technologies never determine the outcomes of what are essentially human dynamics: that online networks can spread information more efficiently does not necessarily mean that they are always efficient. In fact, for every instance of successful massive mobilization, there are many examples that fail to take off. It all depends, in the end, on how technology is used and how its affordances are exploited (Earl and Kimport, 2011). Likewise, explaining the use of digital technologies is different from explaining the determinants of protest behavior – values, attitudes, political efficacy, and motivations are still the key theoretical elements that explain mass political mobilization (Opp, 2009; Anduiza et al., 2012). Finally, digital technologies do not necessarily challenge the logic of collective action: online petitions, for instance, can now accumulate a critical mass more quickly, but the mechanisms underpinning its success have not changed (Earl and Kimport, 2011). Only certain forms of protest make use of technological affordances that were not available before, like DoS attacks; these are the empirical instances of collective action that might require rethinking classic theories.

Olson's logic was built on the premise that the provision of a collective good requires interactions in small groups: this facilitates coercion and the application of selective incentives. However, digital technologies are reducing organizational costs and allowing new forms of incentive and norm creation in large groups. They are also blurring the boundary between the private and public domains: networks that are personal and not necessarily political in nature can now be swiftly mobilized for

political purposes (Bimber et al., 2005; Lupia and Sin, 2003). Many examples of digitally enabled protests (from the anti-globalization movement in the 1990s to the wave of political protests that started in early 2011) challenge the explanatory capacity of classic approaches to collective action – and they highlight the importance that networks have as the main structures of communication.

The Evolution of Digitally-Enabled Protests

Protests enabled by digital technologies can be differentiated in two stages: a first stage of Web-based activism, and a second stage of social media activism (González-Bailón, 2013). In addition to being chronological, what differentiates these stages is the nature of the technologies employed (or the affordances of the media) and the prevalence of those technologies among the general population, not just among those politically inclined. Early empirical examples that belong to the first stage include the 1994 Zapatista movement (indigenous protests that were projected globally from the remote territory of Chiapas in Southern Mexico); the 1999 Battle of Seattle (the first big mass demonstration that used Internet technologies to coordinate protesters against global financial institutions); or the international campaign to ban land mines (a coalition on nongovernmental organizations (NGO) that won the Nobel peace prize in 1997, having used online technologies to raise awareness and gain prominence). More recent examples, which make digital protests transition into the second stage, include online platforms like 38 Degrees and MoveOn (which help bring people together to coordinate protests and organize campaigns); the Anonymous group (who use hacktivism as a way of protest); and, most visibly, the protests that erupted around the world in 2011, including the Occupy Campaign.

What characterizes the first stage of digital protests is the emergence of transnational networks, or what some have called a 'global civil society.' This was made possible by the increasing availability of Internet and Web technologies. The Zapatista movement stands as one of the most prominent examples of these early stages: this movement illustrates how local struggles could grab the attention of an international public by leveling their digital presence. Organizations (NGOs and advocacy groups, mostly) were the crucial nodes in these transnational networks, and Web sites and electronic mailing lists offered the main dissemination tools. As Internet technologies evolved, those transnational networks matured and a global public domain started to consolidate: the Web carried information to which everybody (with the appropriate search) could be exposed, and this created a common ground that acted as a global soundboard.

The second stage of digital activism brought a significant change: it made transnational networks easy to reach for individuals. Citizens could now bypass organizations and harness the power of communication networks more easily, with the ability to exploit the domino effect of their connections without the support of powerful organizations orchestrating the action. In other words, common users became the crucial nodes in the network. Social media and mobile applications blurred the (up to then) well-defined distinction between private and public connections, a division that still shaped the first generation of Web technologies.

The First Wave of Digital Protests

During the first wave of digital protests, the Web was the main weapon. Organizations were instrumental in the dissemination of information, and online networks, in the form of Web sites and hyperlinks, created the communities delivering that information and disseminating calls for action. Research on the Web presence of the Zapatista movement relates their online centrality to connections with the Web sites of hundreds of global NGOs that offered an international network of solidarity (Garrido and Halavais, 2003): all these organizations opened the routes through which attention and support could travel to the Zapatistas' cause. Given the importance of hyperlinks to direct traffic (either directly by means of clicks, or indirectly by means of search algorithms) being central on the Web was crucial to gain public attention.

During this stage, the Web acted as an amplifier of protest messages: it facilitated the formation of links across causes that until then were scattered in off-line collaboration networks, forged by NGOs and advocacy groups but often invisible to the public. Web presence was not a sufficient condition for immediate success (for every local struggle projected to global visibility, the Web sank many more into oblivion, Bob, 2005) but it offered a springboard to which very few protest movements would have had access before.

In addition to the dissemination of information, the Web also offered the infrastructure to launch new actors to prominence. Particularly relevant among these are independent media sites like Indymedia. This platform gave visibility to a global network of journalists that stood as an alternative to corporate media. It was instrumental in the coverage and coordination of antiglobalization (or global justice) protests, starting in Seattle in 1999, during the World Trade Organization conference (Shepard and Hayduk, 2002). Protesters could use these online resources to coordinate their actions, but also to build their own story of the events, on the margins of traditional media narratives and supported by their own graphic documentation, digital and easy to publish online.

These Web networks were very powerful in reaching audiences and creating awareness. The International Campaign to Ban Landmines, awarded the Nobel Peace Prize in 1997, is often cited as an example of how the Internet increased the speed and efficiency of group mobilization: in this case, to coordinate the efforts of six founding NGOs to build a global network in dozens of countries for the elimination of anti-personnel land mines. Five years after being launched, the campaign attained an international treaty.

Online petitions are an important component of how the Web transformed political protests. Because of the increased connectivity, and the lowering costs of information diffusion (no need of paper, print, or stamps), the number of supporters and signatories could be scaled up in shorter time frames. Posting on a Web site the details of a petition, and encouraging subscription to electronic mailing lists facilitated the diffusion of calls. Research on the spreading dynamics of a petition to organize opposition to the war in Iraq in 2002 shows that the chain letter unfolded in a narrow but very deep treelike

pattern: it did not reach many people in a few steps but through many steps that reached a few people at a time (Liben-Nowell and Kleinberg, 2008). E-mail lists allowed, in other words, successful instances of large-scale diffusion which before would have been more difficult to attain or would have been more easily broken at any of the stages they need to unfold.

In all these instances, the Web created global circuits of information and some shortcuts to reach potentially wide audiences. But people, the audience, had to enter those circuits to be exposed to the information. As content producers, users also had to create their networks from scratch: publishing a Web site did nothing to raise the profile of a cause if no other Web sites linked back to it, granting visibility in an otherwise vast and unchartered sea of information. The ability to mobilize resources, in the form of solidarity networks, was still an issue for most organizations wanting to raise awareness – and the reason why successful instances of mobilization were more often than not an exception.

The Second Wave of Digital Protests

During the second stage of digital protests, social media arrived to highlight the power of personal connections. The evolution of Web technologies, and the growing number of Internet users, changed some of the mechanisms by which online networks help protests emerge. The Web ceased being about creating new networks; new applications started to emerge that capitalized on already existing social networks. Social media and networking sites (like Facebook or Twitter) make personal networks the backbone of their services, and the main channels for information diffusion. In this new information environment, users do not need to enter an online circuit to be exposed to specific types of information (as they had during the first stage of online technologies); now it is their contacts and groups of reference (formed by colleagues, friends, and relatives, but also acquaintances and public figures) who bring that information without request.

During the first stage of digital protests, people had to search for calls for action; now, these calls come via friends, or friends of friends, in the constant flow of information that defines social media. This change has a significant impact on information diffusion: we are more likely to pass on information if the source belongs to our social circles (Bakshy et al., 2012). Social media allow us to monitor those circles of friends and acquaintances with constancy difficult to replicate off-line. This is crucial to understand the way in which protest movements emerge today; but two other transformations are equally relevant: mobile connectivity, and the infiltration of mainstream media into personal networks.

Mobile connectivity allows people to be constantly exposed to streams of information, regardless of location. This accelerates their reaction times to calls for action. It also means that the technology is more flexible to organize and coordinate protests once they are in motion, especially if they involve the management of camps in public squares, as it happened in 2011 in Cairo, but also in Madrid or New York (Castells, 2012; Gerbaudo, 2012). The integration of mainstream media and professional journalists into personal networks is also important: it means that online networks are not as clear an alternative now to corporate media as they used to be during the first wave of digital activism. Users now do not write on the margins of mainstream media to provide alternative views (as it happened with platforms like Indymedia); instead they engage in direct information exchange with mainstream media. Twitter is the most prominent example of this two-way mode of communication across old and new actors.

The consequence of all these shifts is that the boundaries of the private and public spheres – as predicted by early criticisms of the classic logic of collective action (Bimber et al., 2005) – get fuzzy: the channels of communication for private and public talk are conflated into the same networks; that is, a significant portion of information shared online goes through networks that are not necessarily political, but that can be activated for political purposes as circumstances arise. The activation of preexisting networks for a political cause is more effective than having to unfold (or revive) ad hoc communication structures: it replicates recruitment dynamics through social networks with the same logic applied prior to the emergence of the Internet. The key is that it is more effective to have a friend telling you of a petition, or of the date and venue for the next protest demonstration, than an impersonal organization to which you might not even belong.

The composition and rhythm of online protest networks change in this second stage, but many of the characteristics that were already highlighted in the early Web manifestations remain: online networks are still decentralized structures that pull together local sources of information and create efficient channels for a (potentially global) diffusion. However, support networks built with Web sites and hyperlinks responded to strategic planning by organizations; personal networks in social media sites, on the other hand, emerge spontaneously out of individual choices: some users are better connected than others, and some have geographically more diverse networks, creating the bridges that might lead to global diffusion dynamics.

The wave of protests that emerged in 2011 (from the Arab Spring, to the Spanish 'Indignados,' or the Global Occupy Campaign) seems to reflect this global interdependence of localized, personal networks. However, systematic research in this area, uncovering the actual bridges and network mechanisms that facilitated diffusion across national borders, is still lacking. Also, this wave of protests cannot be understood without taking into account the pervasive role of mainstream media: in spite of being part of many of those online networks, mainstream media still have a gripping influence in the spread and framing of news from the top-down. Finally, protest movements are not caused by social media activity; social media offer just a vehicle (one of the many) to help a protest movement rise from the shadows; the crucial explanatory factors are still social and economic, if not historical. But as a vehicle, it is true that online networks offer avenues for protest mobilization that allow unprecedented reach and speed. The following section explains why.

Online Networks and Protest Diffusion

The success of protest mobilization through online networks is difficult to predict because it depends on the right alignment of

conditions on different levels: at the level of local interactions (where personal information contexts are shaped), and at the level of global activity (where diffusion branches collide). Much in the same way as chain letter trees exhibit a narrow and deep structure, requiring many nested layers of forwarding to reach a large number of people, so the chain reactions in protest networks follow intricate paths that are difficult to anticipate – among other things, because they might end or split at any stage.

Diffusion chains to join a protest might end because people stop forwarding the information through their personal networks; and so information ends up being trapped in their social circles. The reasons why people might not want to continue a protest chain vary: maybe they did not register enough activity in their local network to consider the protest relevant (and were averse to the idea of becoming one of the leaders); or they were against the protest and decided not to join, in spite of local pressure; they might also break the chain for a number of reasons that are independent of the protests: social media, after all, operates in a world that still constraints online activity and puts a limit to political engagement.

The sum of all those individual decisions is also not enough to anticipate the emergence of large protest events. The cumulative effects of individual behavior are not linear: as Section The Importance of Interdependent Decisions explained, networks activate feedback mechanisms that can propitiate a sudden change in the mobilization dynamics, like exponential growth in the number of protesters recruited. These mechanisms are at the heart of network complexity – and the unpredictability of their dynamics – but they can be dissected and reconstructed to uncover their logic.

An analysis of the protests that emerged in Spain in 2011 (the 'indignados' or 'outraged' movement) does precisely this: it reverse engineers the network mechanisms that helped the protests grow from nowhere (González-Bailón et al., 2011). It shows that recruitment initiatives started all over the network (i.e., at the center and at the periphery) but that information cascades were more easily triggered by users who were sending protest messages from the core. A follow-up study, though, showed that a few influential actors – common users before the protests – were catapulted to global network prominence in a matter of days (González-Bailón et al., 2013), which testifies in favor of social media as a tool to build visibility when there was none.

Most chain reactions in online networks burn before long, that is, before they roll over to reach a large number of people (Borge-Holthoefer et al., 2013). And whereas large cascades tend to start at the core of the network, network centrality is not a guarantee of success. Interdependence makes chain reactions more likely, but also more unpredictable and rare. What has changed with social media is that there are more attempts to trigger those reactions – often without the large-scale vision or motivation; more attempts means that there are more successful cases, as the buzz around viral content attests. And political content is effervescent by nature, which helps create the rush that accelerates cascading behavior. But there is nothing intrinsic in online networks guaranteeing the success of digital protests.

Future Lines of Research

The protest movements that have emerged in the last two decades provide evidence that Internet technologies are helping users bypass traditional modes of mobilization. But answering why and how new media facilitate self-organization and information diffusion requires a careful approach to the networks they help forge, and to both the successes and the (more abundant) failures. The Web and social media have altered the way in which people interact and come together, making formal organizations less necessary and audiences potentially wider; but at the core of the process there are still motivations and chains of influence that transcend the technical capabilities of those tools: the main action still depends on people's willingness to act.

The study of online networks is still in its infancy, and the implications for the study of protest mobilizations in the digital age are yet to be fully understood. But social media have provided – in addition to tools for protesters – valuable data that are helping researchers gain insights into how people come together, and how networks facilitate or hinder the emergence of protests. Research to date has pointed to three relevant dimensions: the dynamics of message propagation and recruitment; the geographical spread of social networks; and the scales of human mobility. All three aspects are relevant to understand why some attempts to organize protest mobilizations succeed, and most fail. Future work will have to shed more light on these, especially if protest organizations (or governments or NGOs) are to draw lessons to exploit the potential of new media.

The available evidence suggests that collective action is more transient now, and that spikes of attention and protest bursts do not easily crystallize into more stable trends or outcomes. Protests can never become, by definition, stable and fixed organizational structures – their power lies, precisely, in their exceptionality. Online networks have proved to be good at channeling that sort of fluid commitment. However, as Internet and Web technologies continue to mutate, so will the strategies that are used to irrupt the political scene. In this, technology is only one part of the story, a very volatile part: soon, Twitter accounts will look as outdated as the gadgets the Zapatistas used in the Chiapas jungle to project their message. An accurate picture of how digital technologies are changing protest movements has to focus less on the nature of the technologies and more on how they are used to join people together in a collective effort. This article has introduced the main concepts that can help develop that theoretical understanding.

See also: Information Society; Media and Social Movements; Mobile Communication; New Media and Democracy in the Arab World; Political Communication; Public Sphere and the Media; Social Media.

Bibliography

Anduiza, E., Jensen, M., Jorba, L. (Eds.), 2012. Digital Media and Political Engagement Worldwide. A Comparative Study. Cambridge University Press, Cambridge.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.
Bakshy, E., Rosenn, I., Marlow, C.A., Adamic, L., 2012. The Role of Social Networks in Information Diffusion. Paper presented at the ICWSM, Dublin.
Bennett, W.L., Segerberg, A., 2012. The logic of connective action. Information, Communication & Society 15 (5), 739–768. http://dx.doi.org/10.1080/1369118x.2012.670661.
Bimber, B., Flanagin, A., Sthol, C., 2005. Reconceptualizing collective action in the contemporary media environment. Communication Theory 15 (4), 365–388. http://dx.doi.org/10.1111/j.1468-2885.2005.tb00340.x.
Bob, C., 2005. The Marketing of Rebellion. Insurgents, Media, and International Activism. Cambridge University Press, Cambridge.
Borge-Holthoefer, J., Baños, R.A., González-Bailón, S., Moreno, Y., 2013. Cascading behaviour in complex socio-technical networks. Journal of Complex Networks. http://dx.doi.org/10.1093/comnet/cnt006.
Castells, M., 2012. Networks of Outrage and Hope. Social Movements in the Internet Age. Polity, Cambridge.
Chadwick, A., 2006. Internet Politics. States, Citizens, and New Communication Technologies. Oxford University Press, New York.
Earl, J., Kimport, K., 2011. Digitally Enabled Social Change: Activism in the Internet Age. MIT, Cambridge, MA.
Elster, J., 1989. The Cement of Society. CUP, Cambridge.
Garrido, M., Halavais, A., 2003. Mapping networks of support for the Zapatista movement: applying social-networks analysis to study contemporary social movements. In: McCaughey, M., Ayers, M.D. (Eds.), Cyberactivism: Online Activism in Theory and Practice. Routledge, London.
Gerbaudo, P., 2012. Tweets and the Streets. Social Media and Contemporary Activism. Pluto Books, London.
González-Bailón, S., 16 April 2013. From Chiapas to Tahrir: Networks and the Diffusion of Protest, World Politics Review. Retrieved from: http://www.worldpoliticsreview.com/articles/12872/from-chiapas-to-tahrir-networks-and-the-diffusion-of-protest.
González-Bailón, S., Borge-Holthoefer, J., Moreno, Y., 2013. Broadcasters and hidden influentials in online protest diffusion. American Behavioral Scientist 57 (7), 943–965. http://dx.doi.org/10.1177/0002764213479371.
González-Bailón, S., Borge-Holthoefer, J., Rivero, A., Moreno, Y., 2011. The dynamics of protest recruitment through an online network. Scientific Reports 1, 197. http://www.nature.com/srep/2011/111215/srep00197/abs/srep00197.html#supplementary-information.
Gould, R., 1991. Multiple networks and mobilization in the Paris commune, 1871. American Sociological Review 56 (6), 716–729.
Granovetter, M., 1978. Threshold models of collective behavior. American Journal of Sociology 83 (6), 1420–1443.
Granovetter, M., Soong, R., 1983. Threshold models of diffusion and collective behaviour. Journal of Mathematical Sociology 9, 165–179.
Liben-Nowell, D., Kleinberg, J.M., 2008. Tracing information flow on a global scale using internet chain-letter data. PNAS 105 (12), 4633–4638.
Lohmann, S., 1994. Dynamics of informational cascades: the Monday demonstrations in Leipzig, East Germany, 1989–1991. World Politics 47 (1), 42–101.
Lupia, A., Sin, G., 2003. Which public goods are endangered? How evolving communication technologies affect the logic of collective action. Public Choice 117, 315–331. http://dx.doi.org/10.1023/B: PUCH.0000003735.07840.c7.
Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.
Opp, K.-D., 2009. Theories of Political Protest and Social Movements. Routledge, London.
Schelling, T.C., 1978. Micromotives and Macrobehavior. Norton, London.
Shepard, B., Hayduk, R., 2002. From Act up to the WTO. Urban Protests and Community Building in the Era of Globalization. Verso, New York.
Shirky, C., 2008. Here Comes Everybody. The Power of Organizing without Organizations. Allen Lane, New York, NY.
Watts, D.J., Dodds, P.S., 2010. Threshold models of social influence. In: Bearman, P., Hedström, P. (Eds.), Handbook of Analytical Sociology. Oxford University Press, Oxford.

Social Protest Movements and Social Work Practice

Carolyn Noble, Victoria University, Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of social protest movements is well established in the sociological literature, but its influence on social work's more progressive practice has been less well documented and discussed. That is until the reemergence of a capitalist recession across Western democracies in the late twentieth century, which resulted in massive rises in unemployment, homelessness and other social manifestations of poverty, and structural and personal disadvantage, characteristic of postwar Europe. Recalling the activism of social workers during the 1970s and their links with the radical milieu of the time, many social work scholars are calling for a reengagement with contemporary social protest movements to speak out against the unchecked growth in inequality and social disadvantage that has resulted from the recession and compounded by the more recent global financial crisis. This article explores what a more serious engagement with the politics of social protest movements might have on social work's more progressive politics and practice, especially the more change-orientated focus of community work and its links with social action.

Introduction

The explosion of mass social protest movements in the 1960s, 1970s, and 1980s across Western democracies resulted in the study of this phenomena going forward at an unprecedented pace as these protests gave rise to new ways of thinking about and engaging in the world (Della Porta, 2008; Tilly, 2012). Although protest movements were not new, what was new was their breadth, depth, scope, visibility, and their global reach. The women's movement, the civil rights movement, the peace and environmental movements, trade union and labor movements, gay and lesbian and human and animal rights movements, cultural and indigenous land rights movements, prodemocracy and anticorporation movements became formidable forces for social change in the second half of the twentieth century reawakening the potential strength of political protests and consequentially giving new potency to social unrest as a mean of challenging the status quo (Della Porta and Diani, 2009; Tilly, 2012). Issues concerning gender, ethnicity and race, sexuality, ability, democracy, human rights, and social justice were to be considered in the creation of more egalitarian social relations between and among different populations (Dominelli, 2006). Radical social work of the 1970s shared many of the aims, philosophy, and social action strategies characteristic of these social protest movements (Bailey and Brake, 1975; Langan and Lee, 1989). The social worker as agitator was an important aspect of radical practice.

Activists who were fed up with their sense of alienation and the social, political, economic, and cultural oppression and human rights abuses in their countries went out into the streets to give notice to whoever was listening (sometimes peacefully, sometimes violently) that these repressions would no longer be tolerated. Fighting for their universal entitlement to democratic governance and political openness, as well as basic human and environmental rights, social justice outcomes in legal, policy, and planning processes alongside their innate right to safety in expressing differences, to seeking equal educational opportunities and rights to a productive life, these protesters signaled that these rights would be fought for with sustained resilience (Della Porta and Diani, 2009; Tilly, 2012). The more recent protests across the Middle East (also known as the Arab Spring), the M-15 movement in Spain and the United States (US) inspired Occupy Wall Street movement, which swept the world in 2011–12 as well as smaller uprisings in sub-Saharan Africa, the Maldives, the Solomon Islands, and the democratic movement in Africa makes the study of this phenomena a current issue. Especially for social workers concerned with fighting for social justice, democracy, equality, and freedom from oppression in a world where poverty, injustice, intolerance, discrimination, and violence seem to be escalating across the globe (Dominelli, 2008, 2012b; Lavalette, 2011; Beinin and Vairel, 2011).

This surge in social movement activity has also helped rearticulate a more radical social work 'Left' agenda and the importance of social work in taking a stance, once again, against the exploitative regimes of the capitalist class and its neoliberal economic order and the inequality and injustice it brings to large sections of the population (Gray and Webb, 2013). Challenging social injustice is to challenge the political construction of the oppressed and marginalized as defined by the state and the power elites. Linking social movements with the radical reawakening of a progressive politics in social work; a politics that regards social problems and human suffering as having its root cause in the large-scale social and political processes rather than as a consequence of individual pathology or deficiencies, is timely given the growing inequality between people, communities, and countries. Linking with contemporary social protest movements will ensure citizen service users are treated equally in the allocation of resources and the benefits of the society more generally in order to create new, nonoppressive social and political relations (Dominelli, 2002b, 2008; Ferguson and Woodward, 2009).

While questions about what motivates and unites people to form resistance against seemingly powerful adversaries such as the state or military–industrial complexes are still being explored what is certain it that social movement protests remain, more than ever, people's choice for mass resistance

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28085-9

when abuses in power pose threats to 'ordinary' peoples' day-to-day access to material needs and resources as well as their social, political, and cultural rights despite the very real repressive counteraction they face (Della Porta, 2006; Snow and Soule, 2007). They represented the collective voice of disaffected individuals acting defensively and offensively for themselves by demanding protective legislation for their rights and real, material solutions to fix the wrongs being perpetuated against them (Crossley, 2002; Thompson, 2010; Buechler, 2011). Highly visible, articulate, and self-confident these activists, working outside of and against the state and its interests have developed a low cost, relatively effective, decentralized method to challenge a centralized, extraordinarily well-organized, seemingly impenetrable, and well-defended adversary (Snow and Soule, 2007; Sklar and Dublin, 2011).

So what is it about social protest movements that has and continues to capture citizen action? And what role has social work played in response to these movements' activism. First, I will discuss social movements in general then I will turn my attention to social work's response to this phenomenon.

Social Movements

Social protest movements have been variously defined as the new politics of international action (Cohen and Kennedy, 2000), as organized efforts by significant numbers of people to promote social change (Crossley, 2002), in Marxian terms as signs of impending collapse of the existing capitalist order or in postindustrial terms as attempts by individuals to move concerns away from class politics to apprehensions with their own sense of belonging, self expression, cultural, and individual freedoms influenced by the lifestyle politics of the hippie counterculture of the time (Pakulski, 1991). Influentially Touraine (1988, 2007) describes social movements as conflicts about the social control of cultural patterns and expressions rather than (just) material interests or class politics. He does not deny the importance of class politics as a site for movement activism but regards the fight over the main cultural patterns of the society as *the* primary site for struggle and social challenge. The aim of this struggle, he argues, is to gain control over the compromises of the past settlements that determined class, gender, race and culture, and ethnic and sexual arrangements in order to generate new meanings in these settlements. Beck (1997) talks about antipolitical politics or mass protests, where disaffected individuals confront injustice and impersonal repressive power with the only thing at their disposal – their own humanity and Leonard (1997) regards new social movement activism as framing a politics of hope for the politics of the future. Feminists in making their personal experience of oppression a public issue contributed to the social transformation of all women, not just their own lives and demonstrated how social action can be informed by personal issues to influence social change (Langan and Day, 1992; Dominelli, 2002b, 2006).

Crossley (2002) demonstrates how social movement protests have always accompanied unpopular or exploitative acts of social, cultural, and political expansion especially if these expansions rendered significant groups of people powerless, or uprooted the original inhabitants from their lands or cultures and deprived them of fair and equitable access to their country's resources and riches. For example, the religious, trade and capital, and cultural movements that sought (and fought) for global dominance during the sixteenth, seventeenth, and eighteenth centuries encountered strong resistance from the original inhabitants. In broader terms, antiimperialist, anticolonial movements, international trade union solidarity, and international workers movements date back to late nineteenth century and cultural internationalism (the protest movement challenging the dominance of Western ideas about economics, culture, and intellectual traditions), the women's suffrage and more organized global antislavery movements also assumed greater significance after World War I (Crossley, 2002; Snow and Soule, 2007; Smith, 2008). Indigenous struggles to protect their land, culture, and language were always an important part of these resistances from early settlement days. Indigenous resistance such as the mass slave revolt in Haiti and the sustained resistance by indigenous peoples in North and South America, Australia, and Pacifica are examples of this resistance (Guadalupe Vargas Hernandez, 2005; Merlan, 2005).

Social Movements and Social Work Practice

Powell (2001) noted as do others (Dominelli, 2002a, 2012b; Ferguson and Woodward, 2009; Allan et al., 2009; Mullaly, 2010) that social work has always been influenced by the politics of the day, sometimes more conservatively such as the Charity Organisation Society while at other times more progressively such as the Settlement Movement in the UK, which influenced the more community-focused early social work development in the US in the late nineteenth century. In more contemporary times, social workers have worked alongside activists in bringing attention to the pressing social, cultural, and political concerns they encounter in their everyday work as frontline practitioners and social policy activists (Noble, 2007). We know from personal experience that many social workers were out in the streets politicking alongside their clients, other professionals, and student activists. Joining with the new wave of collective consciousness that gripped the global world social workers stood in solidarity with these protesters and experienced a new found power in the 'collective' and its use as an immediate and effective strategy to demonstrate against, now publicly 'named,' long-standing grievances against the state, global companies, and interests as well as other forms of authority and repressive power. Fighting against all forms of oppression including lifestyle choice and ability as well as against the inequality in income and wages, access to money, and the benefits it brings and other forms of institutional domination and oppression, social workers used social action to fight for their own sense of justice as well as for the service users in their care (Thompson, 2010; Dominelli, 2012a,b). Social workers and community workers also participated in the United Nations initiatives such as the Copenhagen Social Development Summit and the Beijing Action Platform for Women in 1995 and were very active in the formulation of the Millennium Developmental Goals in 2000 (Dominelli,

2006). As stated above, what is new about these more contemporary protests is the vastness of their reach, the impact of their actions and the success of their demands or the degree with which they can unsettle the status quo.

Social work's attention anti-oppressive practice (AOP) has quite a body of seminal texts to its name. In particular, Ferguson and Woodward (2009) Dominelli (2002a, 2006, 2008, 2012b), Baines (2007), Allan et al. (2009), and Mullaly (2010) give a comprehensive overview of social works more radical theory and practice and these authors continue to be strong advocates for social workers to be active in naming and then fighting for social, political, cultural, and more latterly environmental justice. This progressive practice links back to the early radical social works' explicit class analysis and its links with the trade union movement, which influenced many social workers to form collectives against the state and to critique the nature of work and capitalist production and its role in reinforcing unequal social and political relations (Bailey and Brake, 1975; Langan and Lee, 1989). Following a class analysis, feminist social work confronted the gendered nature of the social, political, and cultural order demanding women free themselves from male power and its pervasive dominance in oppressing women and children. Feminist campaigns from abortion issues, to reproductive rights and decriminalizing prostitution, eliminating domestic violence, and securing better health initiatives showed how practice can result in policy and service provisions to change women's lives for the better and is well documented in the literature (Langan and Day, 1992; Dominelli, 2002b, 2006). AOP extended an analysis of class and gender to include a critique of ethnicity, sexuality, ability and power, and social relations not in distant or abstract terms but as material conditions that reproduce social, political, and economic inequalities in individuals' lives (Baines, 2007; Allan et al., 2009). This progressive practice challenged the oppressive social arrangements that excluded people from accessing resources or participating in public policy debates and political decisions that affect their lives. Attention was drawn to incidents of mass human rights violations, abuses in political power, mass misappropriation of wealth and natural resources, widespread corruption, economic decline and subsequent rise in levels of unemployment and poverty, ethnic conflict, religious repression and gender violence in their communities, and beyond in order to bring about more socially just relations and outcomes (Noble, 2004, 2007). Sociological literature, especially, critical theory became a feature of social work discourse and scholarship replacing the emphasis on psychological and individual well-being. Linking with critical theory, a progressive social work argues that structural inequality and individual oppression are a result of the dominant sociopolitical, cultural, and economic arrangements (or settlements) not the result of individuals who receive the consequences of these defective arrangements (Leonard, 1997; Pease and Fook, 1999; Allan et al., 2009).

These political ideals shared the ideals of social protest movements fighting for and defending citizen's rights, including the right to work, to adequate income, to education, to health, to a healthy environment, and to the protection of all citizens' right to engage in political, social, and cultural life. In applying these ideals to practice social workers were encouraged to highlight the social and human consequences of violence, suicide, high rates of incarceration, mental and physical ill health, poverty, social, religious and cultural marginalization and political repression, trafficking of women and children, environmental degradation, the mass atrocities against indigenous peoples and the impact that torture and war has on individuals, their families, and communities (Annetts et al., 2009; Motta and Nilsen, 2011). Importantly feminists argued for social workers to make the political personal in their own lives. Many social workers took up this challenge, as cited above in securing the presence of a social justice, human rights agenda for women and children as well as for people with disabilities, people receiving psychiatric care, people with learning difficulties, children in care and survivors of abuse and family violence, older people, asylum seekers and people living with HIV/AIDS, and other users in social support and income maintenance services (Noble, 2007; Ife, 2010). The social worker as agitator also helped establish governmental inquiries into the abuses children, the differentially abled, Indigenous peoples, and asylum seekers experienced under the care of the state as well as breaking down the walls of mental health institutions. Currently social workers are continuing their activism against the new austerity measures of neoliberal philosophy, which has a strangle hold on most Western counties, post-Global Facial Crisis, and the environmental movements in protecting the world's natural resources against the rampant depletion by unchecked consumer consumption (Ferguson and Woodward, 2009; Lavalette, 2011) Dominelli (2012b) in promoting green social work extends social work activity into the realm of social ecology and environmental and cultural sustainability.

The 'Dark Age' – the Impact of Neoliberalism on Social Work

As we review this history with the wisdom of the present situation it must be acknowledged that despite the enthusiasm for change toward a more active politics, an engagement with service users for fairer access to service and resources and attempts to increase their political power, social works attempt to change policies toward a more social justice focus has been patchy at best and ineffectual at worst (Powell, 2001). The salient fact remains that even the more progressive form of social work needs government funds and policy support for its well-intentioned ideals. Governments are reluctant to finance radical politics, especially organizations seeking political upheaval or are critical of their policies and practices. Dominelli (2012a), Ferguson and Lavalette (2011), Mullaly (2010), and Noble (2007) in particular document how the mood and enthusiasm of the radical ideals of the early 1970s and 1980s was gradually extinguished by the pervasive influence of neoliberal philosophies. Its not that the radical scholarship was not evident (see Dominelli, 2002b, 2008) or that there were no more attempts at social change; it is just that the overall impact of neoliberalism placed its scholarship and activities in its shadow, especially since the late 1980s. Many

contemporary scholars have summarized the impact of the neoliberal agenda on the welfare system and services as having resulted in: the primacy of individual enterprises over state-funded services, of private-based services over publicly funded ones; self-motivational ideals over collective aspirations to care for the 'common good' and finally small government and reduced welfare services and provisions over a progressive welfare state and universal welfare provisions undermining the traditions of working-class politics and social democratic considerations (Powell, 2001; Ferguson and Woodward, 2009; Ferguson and Lavalette, 2011; Dominelli, 2012a). The dismantling of the welfare state and its welfare provisions was/is in progress in favor of the liberal/conservative manta of individual pursuits over the collective and economic austerity over social capital spending and infrastructure developments.

The dismantling of the welfare state is exacerbated, argue Ferguson and Woodward (2009), by the pervasiveness of case management as the new practice model, which "undermines the worker-client relationship, transforming it into a financial relationship, with the client as the reconstructed consumer" (p. 148.). Even in the voluntary sector, case management is the proffered model of care despite, for example, New Labor's attempt in the UK to strengthen civil society through encouraging active citizenship and addressing social exclusion by growing and resourcing the voluntary, community, faith-based and private networks, and organizations separate from the state (Ferguson and Woodward, 2009). The role of the social worker as a change agent is almost invisible, leaving the social control aspect intact. There is no denying the contradiction and resulting tension between individual and social change and the social control aspect of social work's state-based role that has plagued social work since its inception. So where to now?

As social work's main concern is to work with people, families, and communities that experience oppressive social arrangements. Oppressive arrangements result from the effects of long-term economic readjustments and the movement of capital and manufacturing offshore and under and unemployment of large sections of the community and the effects of systemic and personal violence, of poverty, and homelessness as well as under representation in politics, financial and governing institutions, the classroom and universities. The call for social work to reengage and reconnect with social movement protest and social work's more progressive politics especially evident in social action and community work still seems relevant and for this reason timely to revisit.

Social Justice, Human Rights, and Progressive Politics

Despite the hostile sociopolitical climate for progressive ideals and sentiments and the pervasiveness of the neoliberal philosophy, progressive social work practice, and its commitment to emancipatory ideals, although lacking visibility in practice, is still evident in its scholarship, mission, philosophy, and theory even if some of its shine and potency has been lost (Rogowski, 2010). Indeed an ongoing commitment to social justice, human rights, and progressive politics is increasingly evident in many of the current social work literature and in some cases has been strengthened, especially in the scholarship coming from scholars in Australia, Canada, and the UK (Pease and Fook, 1999; Powell, 2001; Allan et al., 2009; Dominelli, 2002a, 2012a,b; Mullaly, 2010). Despite the neoliberal agenda, a progressive form of social work can also be found in social work curricula across international programs. For example, social work programs in the US, UK, Aotearoa New Zealand, South Africa, Canada, People's Republic of China, Hong Kong, Vietnam, and Cambodia's (just to mention a few) demand (with varying emphasis) a commitment to social justice, human rights, and for theories of ethnicity, race and whiteness, diversity, racism, and power to be taught in conjunction with practice skills in their curricula (Noble et al., 2009). This is also explicitly evident when new and developing programs use the International Association of Schools of Social Work's (IASSW) global standards (http://www.iassw-aiets.org/global-standards-for-social-work-education-and-training) as a guide when designing their programs and the international definition of social work informs many of these program's philosophical underpinnings. This quote from the definition that social work

> promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilising theories of human behaviour and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work.
>
> http://www.iassw-aiets.org/international-definition-of-social-work

reminds us and keeps alive its more progressive commitment to social change and social justice.

An important development in reasserting social works philosophical commitment to social justice and social change is its own global movement beginning in 2008 when the three partner organizations – IASSW, International Council on Social Welfare, and the International Federation of Social Workers – led by their respective presidents came together to consider specific actions to address rising inequality and its challenges. The Global Agenda, as it is known, has an explicit aim to work as a collective across its constituent members and communities to influence the UN's forthcoming initiatives, in particular, the post-2015 development agenda as well as to work with groups and communities worldwide to promote social work's role to protect and promote social justice, human rights, and sustainable development as central to its emancipatory mission. A policy document will be produced for the Tripartite conference in Melbourne in 2014 following extensive consultation across the global social work community (http://www.globalsocialagenda.org). This initiative reaffirms social work as a profession working in the human services and supporting people to achieve the best possible levels of social and personal well-being and working tirelessly to achieve human rights and social justice through social development, social and systemic change,

advocacy, and the ethical conduct of research. The Global Agenda is a collective effort by all social workers to repoliticize social work and recapture the radical manifesto of the 1970s and 1980s and the subsequent social protest and agitation it engendered (Bailey and Brake, 1975; Allan et al., 2009; Ferguson and Lavalette, 2011; Dominelli, 2012b).

Lavalette (2011), Dominelli (2002b, 2008, 2012a), and Allan et al. (2009) call to critical practice or social activism are relevant here. They argue that if social work takes its commitment to social justice and human rights seriously then, for the sake of the profession and the service users, it needs to articulate another way of doing business than what has been evident during the last decades of the neoliberal agenda. That is to demonstrate another way other than servicing a shift to the private sector, of passing the responsibility for managing risk back to the very families and communities who need such help and to challenge the increased corporatization of the whole welfare sector and return a humanity into the social work discourse that the social movement protestors are using to inform their global activism (Powell, 2001). So how can this happen?

Community Work and Social Action

Mayo (2005) argued in the 1970s that the opportunity for radical practice through community action was a missed opportunity. However, she still maintains that fostering new alliances with service users, acting collectively with broader coalitions such as social protest groups that are fighting for rights to employment, and/or adequate income, clean food and water, democratic citizenship, end of poverty, war, torture, ecological and social sustainability, and other socially progressive ideals "will always reflect a powerful alliance for resistance" (Lavalette, 2011: p. 127) and a force, I would add that progressive social work should once again harness. Mullaly (2010) and Ife (2010) argue that social workers have an ongoing responsibility to educate themselves and others about the consequence of imbalances in power, resources, decision-making, and quality of life brought about by the neoliberal and globalizing agenda and form alliances focused on resisting their impositions and social manifestations as sited in the early radical critique. The argument here is to take note of the discourse of progressive politics and the richness of its possible trajectories and its long-term links with social action and social justice that is, fight to give individuals or groups their 'due' within society. Much like the early manifestations of AOP, social workers need once more to agitate for the equitable distribution of resources, access to public benefits and services, equality before the law, and social and environmental development. This can be achieved, argues (Mullaly, 2010), through advocacy (personal and political), consciousness raising, and active promotion of fair distribution of resources and social change toward human betterment irrespective of group membership or social positioning and provides the most obvious platform for action (Noble, 2007). Concerned with advocacy and empowerment community action aims to address and change the oppressive social arrangements that exclude people without power from accessing resources or participating in effective decision-making that relate to their lives so that they can be active in public policy debates and political decisions (Baines, 2007; Mullaly, 2010). It assumes both an individual and group willingness to redress and refigure power imbalances informed by a social justice agenda. Through self-reflection and collective action, individuals are encouraged to identify and name the ways in which social structures shaped their experience of disadvantage and the way that orthodox social work was complicit in the process (Noble, 2007). Weeks et al. (2003) provide a beginning framework for global community action for those social workers looking for a way of resisting the changes associated with the new economy and current economic globalizing activities. They argue that linking local issues to global and regional analyses; working with international human rights instruments; developing transnational civil society, regional networks, and coalitions for advocacy and action by maximizing cyberspace; and developing alternatives to globalization would give progressive practice a 'heads up' and connect their praxis clearly within the social protest movements discourse (Weeks et al., 2003: pp. 13–14). These community workers are also keen to see a reimaging of local communities and action toward democratization of local economic relations, a move in line with social ecology and sustainability activism (Mullaly, 2010; Dominelli, 2012b). Tesoriero (2010) and Dominelli (2012b) see social work's community action groups' involvement with the global justice movement, and the antiglobalization, the ecological and sustainability movements and their links with social justice and human rights strategies as representing a vision of a future society achievable through repolitization of social work practice and its links with social protest movements. Ferguson and Woodward (2009) argue for social workers to relink with the current service users movements (namely mental health, disability, and indigenous resistances to welfare interference) to further the ideals of social work in the public arena. It is social protest movements and their capacity to generate significant social change in a way that no longer seems possible for more conventional politics or for that matter the more conventional casework and case management approach that can help renew and refresh social works value base and inspire a new outrage at past and present injustices. In responding to long-standing government and other policy-making institutional grievances, social protest movements have the capacity to introduce new political actors into contemporary politics whose sole purpose is to alter the status quo as a way of resolving their grievances and creating an alternative political climate in its wake (Tilly, 2012). The vision of these movements, in particular, can provide both a source of inspiration and as a framework for interpreting and seeking change in a planned way rather than as a purely reactive and unorganized response. It is here that Bailey and Brake's (1975) seminal work has reactivated an earlier progressive politics based in a reinvigorated community action/social advocacy approach to practice as evident in the call for a new politics of the Left by Gray and Webb (2013) and Lavalette (2011).

Community action/social advocacy social work assumes that all people have the ability to name and confront their experiences of discrimination and oppression and with collective action work toward creating a system more reflective of their true needs and ideals. Through self-reflection and collective action, individuals are encouraged to identify and name the ways in which social structures shaped their experience of disadvantage and the way orthodox social work was complicit in the process. Community action provides a vehicle for service users to develop their collective voices to express and determine social arrangements that would meet their true needs and ideals (Weeks et al., 2003; Noble, 2007). That is, community work combined with social action strategies is an ongoing and complex process of dialogue, exchange, consciousness raising, and critical reflection, education, and targeted action aimed at developing a partnership with communities to help build their own version of community that meet their needs and aspirations (Tesoriero, 2010; Weeks et al., 2003). What workers can give to the community is their passion, commitment, enthusiasm for social change combined with a vision of a better society, and a belief that change is indeed possible (Tesoriero, 2010). Activist social workers require congruence between personal values and beliefs and work practices, which frequently involve political activity outside their professional work settings. However, while social action can vary between reformist and revolutionary activist practice there can be distilled some common features. These include

1. An understanding of historical processes and the cycle of change.
2. The prioritization of the social structure in the analysis of social problems.
3. Shift from the focus on individual pathology to concentration on oppression and its many forms and sites.
4. Adoption of practice strategies that recognize and challenge the structures of oppression.
5. Development of egalitarian practices with service users.
6. The role of the social worker in activist practice (points 1–5, Healy, 2000: p. 240).
7. Use of multimedia such as art, music, literature, poetry, radio, film, drama, and social-media as means of exposing injustices and offering an alternative vision.
8. Relaying and retelling stories of successful change and current and past struggles in social work education.
9. Developing scholarship grounded in progressive politics that includes a global and local analyses.
10. Seeking creative forms of practice that link the global and the local analyses across social work knowledge, values, and skills (Dominelli, 2002b; Ife, 2010).

Conclusion

Scholarship on community work and social action and its links with social protest movements is, I argue, the most effective and salient practice method for the advancement of a more progressive social work practice. That is, a practice that will encourage the energies, commitment, and shared interests of oppressed people to work collectively to initiate the processes of change toward their liberation and freedom to expose their concerns. Let us not waste this new opportunity to harness its impact and effectiveness; let us get organized again!

See also: Afrocentric Approaches to Social Work; Social Action and Social Justice.

Bibliography

Allan, J., Briskman, L., Pease, B. (Eds.), 2009. Critical Social Work: Theories and Practices for a Socially Just World, second ed. Allen & Unwin, NSW.
Annetts, J., Law, A., McNeish, W., 2009. Understanding Social Welfare Movements. Policy Press, Bristol.
Bailey, R., Brake, M., 1975. Radical Social Work. Edward Arnold, London.
Baines, D. (Ed.), 2007. Doing Anti-oppressive Practice: Building Transformative Politicized Social Work. Fernwood Pub., Halifax.
Beck, U., 1997. The Reinvention of Politics: Rethinking Modernity in the Global Social Order. Polity Press, Cambridge.
Beinin, J., Vairel, F., 2011. Social Movements, Mobilisation & Contestation in the Middle East and North Africa. Stanford University Press, USA.
Buechler, S., 2011. Understanding Social Movements: Theories from the Classical Era to the Present. Paradigm Pub., USA.
Cohen, R., Kennedy, P., 2000. Global Society. Macmillan, London.
Crossley, N., 2002. Making Sense of Social Movements. Open University Press, Buckingham, UK.
Della Porta, D., 2006. Social Movements, Political Violence and the State. Cambridge University Press, Cambridge.
Della Porta, D., 2008. Democracy in Social Movements. Palgrave Macmillan, UK.
Della Porta, D., Diani, M., 2009. Social Movements: An Introduction. Blackwell Pub., Chichester.
Dominelli, L., 2002a. Anti-oppressive Social Work Theory and Practice. Palgrave, UK.
Dominelli, L., 2002b. Feminist Social Work Theory and Practice. Palgrave, Basingstoke.
Dominelli, L., 2006. Women and Community Action, second ed. Policy Press, UK.
Dominelli, L., 2008. Anti-racist Social Work, third ed. Palgrave Macmillan, UK.
Dominelli, L., 2012a. Anti-oppressive practice. In: Gray, M., Midgley, J., Webb, S. (Eds.), The Sage Handbook of Social Work. Sage, London.
Dominelli, L., 2012b. Green Social Work: From Environmental Crises to Environmental Justice. The Polity Press, Cambridge, UK.
Ferguson, I., Lavalette, M., 2011. Globalisation and global justice: towards a social work resistance. International Social Work and the Radical Tradition 49 (3), 309–318.
Ferguson, I., Woodward, R., 2009. Radical Social Work in Practice: Making a Difference. The Policy Press, UK.
Gray, M., Webb, S.A., 2013. Towards a 'New Politics' of Social Work. Palgrave Macmillan, London, UK.
Guadalupe Vargas Hernandez, J., 2005. Social movements for the recognition of the indigenous movement and the indigenous political ecology. Ra Ximhai 1 (3), 453–470.
Healy, K., 2000. Social Work Practices: Contemporary Perspectives on Change. Sage, London.
International Association of School of Social Work (IASSW). Definition of Social Work. http://www.iassw-aiets.org/international-definition-of-social-work (accessed 03.05.13.).
International Association of Schools of Social Work, 2004. Global Standards. http://www.iassw-aiets.org/global-standards-for-social-work-education-and-training (accessed 03.05.13.).
Ife, J.W., 2010. Human Rights from below: Achieving Rights through Community Development. Cambridge University Press, Port Melbourne, VIC.
Langan, M., Day, L., 1992. Women, Oppression and Social Work: Issues in Anti-discriminatory Practice. Routledge, NY.
Langan, M., Lee, P. (Eds.), 1989. Radical Social Work Today. Sage, London.
Lavalette, M., 2011. Radical Social Work Today; Social Work at the Crossroads. The policy press, Bristol, UK.

Leonard, P., 1997. Postmodern Welfare: Reconstructing an Emancipatory Practice. Sage, London.
Mayo, M., 2005. Global Citizens, Social Movements and the Challenge of Globalisation. Zen Books, London.
Merlan, F., 2005. Indigenous movements in Australia. Annual Review of Anthropology 34, 473–494.
Motta, S., Nilsen, A., 2011. Social Movements in the Global South: Dispossession, Development & Resistance. Palgrave Macmillan, UK.
Mullaly, R., 2010. Challenging Oppression and Confronting Privilege: A Critical Social Work Approach. Oxford University Press, Don Mills, ON.
Noble, C., 2004. Postmodern thinking: where is it taking social work? Journal of Social Work 4 (3), 289–304.
Noble, C., 2007. Social work, collective action & social movements. In: Dominelli, L. (Ed.), Revitalising Communities in a Globalising World. Ashgate, Aldershot, pp. 95–104.
Noble, C., Henrickson, M., Han, I. (Eds.), 2009. Social Work Education: Voices from the Asia-Pacific. Vulgar press, Carlton North, VIC.
Pakulski, J., 1991. Social Movements: The Politics of Moral Protest. Longman Cheshire, Melbourne.
Pease, B., Fook, J. (Eds.), 1999. Transforming Social Work Practice: Postmodern Critical Perspectives. Allen & Unwin, NSW.
Powell, F., 2001. The Politics of Social Work. Sage, London.
Rogowski, S., 2010. Social Work: The Rise and Fall of the Profession. Policy Press, Bristol.
Smith, J., 2008. Social Movements for Global Democracy. John Hopkins University Press, US.
Snow, D., Soule, S., 2007. The Blackwell Companion to Social Movements. Wiley-Blackwell, NY.
Sklar, K., Dublin, T. (Eds.), 2011. Women and Social Movements, International – 1840 to Present Day. Alexander St. Press, Alexandria, VA.
Tesoriero, F., 2010. Community Development: Community-Based Alternatives in an Age of Globalisation. Pearson, NSW, Australia.
The Global Agenda. http://www.globalsocialagenda.org (accessed 04.05.13.).
Thompson, L., 2010. Citizenship and Social Movements. Policy Press, UK.
Tilly, C., 2012. Social Movements 1768–2012, third ed. Paradigm Pub., US.
Touraine, A., 1988. Return of the Actor – Social Theory in Post-Industrial Society. University of Minnesota Press, Minneapolis.
Touriane, A., 2007. A New Paradigm for Understanding Today's World. Policy Press, UK.
Weeks, W., Hoatson, L., Dixon, J. (Eds.), 2003. Community Practices in Australia. Pearson Education Australia, Sydney.

Social Psychological Theory, History of

Peter Hegarty, School of Psychology, University of Surrey, Guildford, UK
Sebastian E Bartos and Katherine Hubbard, School of Psychology, University of Surrey, Guildford, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S.T. Fiske, volume 21, pp. 14413–14421, © 2001, Elsevier Ltd.

Abstract

This article reviews the early history of thought linking mind and society and social psychological theories since the late nineteenth-century beginnings of the disciplines *psychology* and *sociology*. We emphasize the growth of thinking about populations in nineteenth-century nation-states, diverse theories of the early twentieth century, Lewinian field theory, growth in American social psychology after World war II, the 'crisis' of the late 1960s and early 1970s, attribution theory and social cognition, the separation of North American, European, and Asian social psychological theories, and contemporary dilemmas about embodiment and culture.

The Long Past of Theory in Social Psychology

Social psychology – like psychology *in toto* – has a long past and a short history. In the late nineteenth century, scholars began to imagine 'social psychology' as new disciplines *sociology* and *psychology* aimed for sciences of 'the mind' and 'society' respectively (Allport, 1954; Collier et al., 1991; Farr, 1996; Greenwood, 2004; Jahoda, 2007; Kruglanski and Stroebe, 2012). Carlo Cattaneo (1801–69) used the term 'social psychology' – *psicologia sociale* – in 1864. The first years of the twentieth century, several texts of 'social psychology' were published including Paolo Orano (1901), *Psicologia sociale*, and Carlos Octavio Bunge (1903), *Principios de psicologia individual y social*. But principled thought about the relationship between 'mind' and 'society' long precedes these events.

Around the *Spring and Autumn* period, Chinese spiritual leaders, notably Confucius (551–479 BCE), wrote enduring texts about the maintenance of social harmony and the ideal life. The enlightenment of Gautama Buddha (563–483 BCE) in the Himalayas initiated a spiritual tradition oriented toward enlightenment through the practice of virtue. In 'Pericles' century,' Socrates and the sophists refocused philosophy on individual morality, society, and language. Plato's (427–347 BCE) *Republic*, Aristotle's (384–322 BCE) description of the uniquely 'rational' human soul, and the theory of 'humours' originated by Hippocrates (460–370 BCE) all influenced medicine and politics in the ancient world, and the literary and scientific culture of the Abbasid Empire (750–1258). In the context of the scholastic philosophy of medieval Europe, framed by matters of Christian faith, William of Ockham (1287–1347) called for theoretical simplicity, a value that survives as 'Occam's razor.' The Italian renaissance prompted enduring reflections on 'society' as separate from such exercise of state power, such as Niccolò Machiavelli's *The Prince* (1531) and Baldassare Castiglione's *The Book of the Courtier* (1528). The former author's name also identifies a twentieth-century personality variable.

The invention of the printing press, Protestant reformations, and the Catholic Council of Trent led many more Europeans to individualized relationships with Christian faith. This rise of individualism figured heavily in social theories of the late nineteenth and early twentieth century. The English civil wars limited the power of monarchs and prompted political theories about the basis of stable society. Thomas Hobbes' (1588–1679) *Leviathan* argued for a strong single ruler, but John Locke's (1632–1704) *Two Treatises on Government* (1689) argued for a 'social contract' between free individuals to jointly agree to be subject to the rule of law. Locke's vision of the person as a 'tabula rasa' contrasted both with Hobbes' view of 'the state of nature' as inherently brutal, and with Gottfried Leibniz' (1646–1716) later description of the human mind as innately capable of autonomous reason. Accounts of psychology's prehistory as a Locke–Leibniz debate make nature–nurture questions central but tend to obscure such theorists as Giambattista Vico (1668–1744) who emphasized how minds are shaped by history (see also Billig, 2008).

Philosopher-historian Michel Foucault (1926–84) described the seventeenth and eighteenth centuries as the period when 'the human' became an object of the sciences. In Britain particularly, systematic theories of the economy began to be theorized. The trend toward rational calculation of human life influenced the *utilitarian ethics* of Jeremy Bentham (1748–1832) in which pain and pleasure – not kings and queens – became the 'sovereign masters' of human conduct. The most radical effects of placing the rights-bearing individual at the center of social thought were the republican revolutions that formed independent nations in North America in 1776 and France in 1789.

The Origins of Disciplines in the Nineteenth Century

Nation-states relate populations of individuals to areas of territory. Nations are characterized by conflicting commitments to individual freedom and government, through the exercise of law. Three approaches to 'population' shaped social theories that predate the academic disciplines of sociology and psychology; statistical thinking, theories of evolution, and Hegel's dialectic. First the 1790 American census defined 'the people' as countable individuals whose 'will' the government would express. In nation-states, not all humans matched the narrow category of *the people* to be politically represented, in

gender, age, religion, or increasingly, in 'race.' In the United States, disputes about counting slaves, Native Americans, and immigrants proliferated. An explosion of statistical thinking in the 1830s saw the rising Bourgeoisie pondering aggregate rates of individual crimes and suicides. Adolphe Quetelet (1796–1874) observed that measures of individual physical attributes formed Gaussian distributions in aggregate, and specified median values as 'norms' (such as the 'body mass index' which still informs health psychology). Auguste Comte's (1798–1857) *Positive Philosophy* repeatedly imagined how such a science of society could rationally address the many social problems in France after revolution.

Second, Thomas Malthus' (1766–1834) *An Essay on the Principle of Population* of 1798 described population growth as geometrically increasing, food supply as arithmetically increasing, and demographic catastrophe as mathematically inevitable. Malthus influenced psychologist Herbert Spencer (1766–1834), whose popular works were celebrated in laissez-faire industrial Britain and later influenced early twentieth century social psychological theory. Spencer's works are often the target of critiques of 'social Darwinism's' Machiavellian attitude to the suffering and exploitation of the poor. Malthus did influence Darwin's (1809–82) theory of nature as a competition between individual organisms for survival in *The Origin of Species* (1859). The theory of evolution specified that variation was the raw material upon which natural selection acted to generate new forms, and that men embodied this variability more than women did (Shields and Bhatia, 2009). Darwin's cousin Francis Galton's (1822–1911) eugenic philosophy built on this idea in its proposals for societal improvement by increased reproduction among valued people, like political leaders and scientists, and by curtailing reproduction among the degenerate.

Finally, G.W.F. Hegel (1770–1831)'s *Phenomenology of Spirit* of 1807 described a dialectical process involving 'triads' in which self-consciousness emerges from its recognition by another in a power struggle, as that between a master and a slave. This 'dialectical' process was partially inspired by Hegel's observations of Napoleon's conquest of the Germanic states. The Hegelian suspicion of the unity of dominant knowledge created a doubtful orientation to all types of dominant knowledge. Karl Marx' (1818–83) dialectical *materialism* aimed to analyze and supersede the class conflict between capitalists and the proletariat created by nineteenth century industrialization. *The German Ideology* (1846) critiqued the 'false consciousness' of the ruling class who made their class interests appear 'natural,' Friedrich Engels' *The Condition of the Working Class in England* (1845) described the lived realities obscured by such false consciousness. Marx and Engels' *Communist Manifesto* (1848) was a blueprint for the revolutions of the twentieth century based on class consciousness. Hegel also informed W.E.B. DuBois' (1868–1963) *The Souls of Black Folk* (1903) and Simone de Beauvoir's (1908–86) *The Second Sex* (1949).

Psychology developed as a science of the individual mind in some isolation from these developments (Fancher and Rutherford, 2012). Immanuel Kant (1724–1804) insisted that the native human mind necessarily perceived the world through the constructions of its intuitions and categories. In the fertile environment of the German university system, this theory outlived Kant's own doubts about any possible empirical science of mind. Hermann Helmholtz (1821–94) timed the speed of the nerve impulse, Gustav Fechner (1801–87) graphed psychophysical relationships between physical stimuli and the sensations they produced, and Wilhelm Wundt (1832–1920) built an Institute for Experimental Psychology at Leipzig in 1879 which trained a generation of researchers who replicated psychological training in multiple locations. Wundt demarcated experimental psychology from *Völkerpsychologie*, the study of culture. Cultural 'thought' was not amenable to study via the individual mind in the laboratory, because it followed a different 'psychic' causality that allowed for creativity and emergence. A related distinction lays in Wilhelm Dilthey's (1833–1911) distinction between *Naturwissenschaft* (natural science) and *Geisteswissenschaft* (social science).

Wundt was rivaled by William James (1842–1910) who taught a pragmatist psychology in the 1870 and 1880s at Harvard, influenced by Alexander Bain on the cultivation of positive habits, and C.S. Pierce's on clear scientific theory and sustained logical thinking. James' *Principles of Psychology* (1890) describes the self as multiple and social, and his account has remained relevant to successive generations of social psychologists' theories of the self. James theory suggested greater unity between mind and society than Wundt's. He and other psychologists disapproved when Wundt's student, Hugo Münsterberg (1863–1916) distinguished the 'ideographic' and the 'nomothetic' to separate historical and psychological kinds of explanations as limiting psychological theory prematurely.

Late nineteenth century French social psychology was informed by Jean Martin Charcot's (1825–93) theory of *grand hypnotisme* describing susceptibility to hypnotism as a symptom of hysteria. Charcot's experiments were shown to be effects of unintended *suggestion*, but 'suggestion' a popular explanation of social influence and 'hypnotism' a metaphor for mindless suggestibility. Gustav LeBon's (1841–1931) *The Crowd: The Study of the Popular Mind* (1895) analogized the collective actions of crowds to the suggestible person under hypnosis, explaining why crowds appear to autonomously follow leaders. Gabriel Tarde's (1843–1903) *The Laws of Imitation* (1890) also used hypnotism to analogize social influence. Tarde's laws included *descent* by which 'inferior' classes imitate superior ones, *geometrical progression* by which ideas disseminate rapidly, and *the internal before the exotic* by which in-group members inspire more imitation than out-group members.

French theories of suggestibility were matched by William James' and others interest in the 'ideomotor action' by which the idea of an act might give rise to the act itself. Tarde's work was translated into English by psychologist James Mark Baldwin who had begun to describe imitation as the means by which children enter society. American economist Edward Alsworth Ross (1866–1951) 1908 *Social Psychology: An Outline and Sourcebook* theorized social behavior as a consequence of 'imitation' or 'suggestion' along French lines, expressing a hope that an individual might rise above these influences, 'to become a person and not a parrot' (p. 4). Ross emphasized the influence of social customs and individual leaders or 'great men.' His claims about Whites' 'race suicide,' and views that some races were more suggestible than others, mirrored the

class anxieties of the French, and reached the ears of United States leaders, as the country was increasingly thinking about its own population in eugenic terms.

LeBon also analogized the crowd mind to that of the 'primitive.' Unusual individuals in 'civilized' societies – such as 'Victor,' a 'wild boy' in Aveyron found in 1800 – had long raised civilized minds. During the late nineteenth-century period of European and American colonial expansion, theories about 'primitive' minds and societies proliferated to legitimate this unprecedented extension of global domination by the few. Physical measurement or *anthropometry* – as in the racist 'American school' of the 1830s, or the 1898 Cambridge Anthropological Expedition to Torres Strait – had limited theoretical scope to ground the category of 'race.' Edward Tyler's (1832–1917) *Primitive Culture* (1871) likened the minds of 'primitives' and children, and described 'primitive' animist beliefs that souls can inhabit things and be separated from bodies (Richards, 2010).

Twentieth-Century Social Psychological Theories

Sexuality, like primitivism, prompted particular anxieties about 'degeneracy' in the late nineteenth century. Sigmund Freud (1856–1939) theorized hysteria as an effect of repressed sexuality, to be revealed by analysis of dreams and symptoms. Freud's psychoanalysis had vast scope and regularly engaged the relationship between minds and societies. In *Totem and Taboo* (1913), Freud merged Locke's contract theory and contemporary anthropology to explain how laws emerged when a 'band of brothers' killed their father and repressed memory of the fact in agreement to share sexual rights to women and property. Here as elsewhere, women were secondary in Freud's theory even if they populated his consulting room (Gilman, 1990). Freud conceptualized the death instinct after World War I revealed unconscious compulsions to repeat among 'shell-shocked' soldiers.

William McDougall's (1871–1938) *Social Psychology* (1909) proposed an influential instinct theory for social psychology. Critical of utilitarianism, McDougall described pain and pleasure as mere indications of whether more basic instinctual drives such as fear, disgust, curiosity, anger, pugnacity, self-basement, self-assertion or the parental instinct were satisfied. Instincts might generalize, as through 'the sympathetic induction of emotion' by which an emotion is felt because it is observed in another. Both McDougall and Freud's theories of instincts were popular, but both drew criticism for their circular reasoning from behavior to taxonomies of hidden instincts and back again.

Several other influential social sciences theories of this period had comparatively *less* influence on social psychology. Émile Durkheim's *La Suicide* (1897) presented statistics to demonstrate how 'society' affected this most profoundly individual choice. Thorstein Veblen's *Theory of the Leisure Class* (1899), and Max Weber's *The Protestant Work Ethic and the Spirit of Capitalism* (1904) continued to bring historical explanation to bear on subjectivity but never became social psychology 'classics.' The absence of an influence of Marxism on American social psychology is notable, but the empirical tradition of making observations of the English working class continued, particularly with Charles Booth's *Life and Labour in London* (1902).

Finally, theories that separated the individual and society, such as Wundt's, influenced the form of social psychology taught in the Chicago philosophy department by George Herbert Mead (1863–1931) after 1901. Mead looked to Darwinism to avoid the teleology inherent in notions of historical progress in German theories. Building also on James and others, Mead described self-consciousness as arising from taking the role of another with oneself. At Chicago, contextual empirical studies of urban realities flourished particularly in the 1920 and 1930s under the leadership of Robert Park (1864–1944). More progressive theories of group differences emerged early on in this context. The *Polish Peasant in Europe and America* published 1918–20 by W.I. Thomas and Florian Znaniecki was influenced by anthropologist Franz Boas arguments against popular eugenics and notions of racial inferiority. Boaz and Thomas and Znaniecki examined the disruptive effects of urbanization and peasants' diverse 'attitudes,' although they did not measure 'attitudes' as later theorists did (Danziger, 2000a).

Social Psychology Theories between the World Wars

Social psychology increasingly aspired to address social problems between the World Wars. The scientific status of psychology in the United States was solidified by the rise of testing, measurement, and application, and behaviorist theory. In the 1920s, the measurement of individual attitudes, values, personality traits, and stereotypes all became more ordinary practices, and competing theories of what these constructs might be were generally occluded as they began to be measured with greater frequency and enthusiasm. Floyd Allport's (1890–1979) 1924 volume *Social Psychology* explained effects of suggestion and association in the behaviorist language of 'conditioned reflexes.' Allport defined the 'social' of social psychology in individualist terms, ruling out of court the idea that there was a 'group' level of description; which he described as a 'fallacy' (Danziger, 2000b). Freud's work stimulated a public appetite for psychological theory as sexual mores changed. Social theories, minds, and societies entered more active feedback loops particularly when social scientists were called upon to address social problems created by the great depression.

Theories of 'culture' and 'personality' put mind and society into theoretical relationships in ways that remained closely tied to psychoanalytic theory and practice rather than measurement. In 1928, anthropologist Margaret Mead (1901–78) published *Coming of Age in Samoa*, using variation in gender roles as 'negative instances' to disprove theories about universal human nature. Along with German émigrés such as Karen Horney (1885–1952), and American psychiatrist Harry Stack Sullivan (1892–1949), Mead and other anthropologists formed alliances around New York known later as 'the culture and personality school' whose members attempted to theorize modern subjective experiences of neurosis, anxiety, and freedom at the boundaries of psychiatry and the social sciences. Adolf Hitler's rise to power stimulated a very different need for 'cultural' explanation of consciousness and behavior.

At Frankfurt's Institute for Social Research, scholars such as Erich Fromm (1900–80) used Marx to theorize the problems of adjustment revealed by psychoanalysis as specific to capitalist societies. Fromm also contributed to the fertile culture and personality school, and his 1941 *Escape from Freedom* typifies a Marxist-Freudian style of historically embedded theory.

Hitler's rise to power in 1933 created a new form of dictatorship that stimulated a new interest in explaining 'suggestion' because of Hitler's effective propaganda. Prior to this, Gestalt psychology had flourished in Berlin, and psychologist Kurt Lewin's (1890–1947) attention was drawn particularly to social motivation (Ash, 1995). His 1935 *A Dynamic Theory of Personality* summarized 'field theory' – the most enduringly influential theoretical synthesis in social psychology from this period. Lewin specified that behavior was a function of a person and phenomenological proximal environment. Expressing a belief in the scientific status of measurement, Lewin's central axiom was expressed by a formula $B = f(P, E)$, used topographical diagrams to show which behaviors *could* occur, and 'vector psychology' to predict which behaviors *would* occur (Gold, 1999). The practical utility of this general theory seemed demonstrated by Lewin's own work on the effects of authoritarian, democratic and laissez-faire leadership. Muzafer Sherif's (1906–88) 1936 studies of conformity to group norms also expanded the concept of the 'norm' beyond sociology, toward something that could be manipulated in a situationist experiment to effect the behavior of individuals and groups. John Dollard (1900–80) and colleagues' *Frustration and Aggression* (1939) was a similarly situationist theory that drew on behaviorism, engaging Freudian concepts to theorize aggressive behavior as a form of drive reduction. By linking patterns of lynching to low cotton prices in the South it went beyond classic Lewinian theory in examining historical change as well as the individual's immediate 'environment.'

American Growth after World War II

The resolution of World War II led to unprecedented international efforts to bring about enduring peace that looked to the 'social sciences' rejections of the notions of race differences and aggressive instincts. A 'cold war' between the United States and the Union of Soviet Socialist Republics described by Eisenhower as a 'battle for hearts and minds' drew military defense and psyche closer together. Psychological research was generously funded by military organizations, private foundations, and the National Science Foundation, and grew in the United States to a totally unrivaled degree. Textbook publication in social psychology accelerated accordingly.

Theories of mind and society became increasingly mathematical and statistical. Influential theories that aimed to put the social sciences on mathematical footing included Von Neumann and Morgenstern's game theory, Shannon and Weaver's theory of communication, and systems theory. Psychologists more frequently used Ronald Fisher's inferential tests to stabilize doubt about the similarity or difference between groups in experiments. The 'construct validity' of tests was described by psychologists Lee Cronbach and Paul Meehl in 1955, curtailing faith in many popular interpretive projective tests. 'Variables' were increasingly theorized as existing in people prior to being observed, unlike the earlier operational understanding of variables as things that psychologists would 'vary' to produce experiments.

In this context, smaller theories developed with less ambitious scope and a closer relationship to experimental test. An early example is Carl Hovland's (1912–61) theory of persuasion as a process that occurred between a *communicator*, *communication*, and *audience*. Theory was increasingly tested in increasingly theatrical experiments, requiring more frequent deception of participants, and decreasing numbers of reminders – such as Hovland's (1959) – that the laboratory and the real world might be different from each other. The now-classic textbook theories and experiments date particularly from the 25 years after WWII; Solomon Asch's experiments on conformity, Muzafer and Carolyn Sherif's theory of realistic group conflict, Leon Festinger's theory of social comparison. British social psychologist Michael Argyle (2001) recorded being impressed by the United States culture of ingenious experimentation but also unsatisfied by its artificiality. Historian Frances Cherry (1995) described how implicit theoretical commitments can be gleaned from reading research from this period with an eye for the authors' troubled attempts to exclude 'stubborn particulars' from abstract experimental designs.

Prejudice became a new phenomenon to explain. Kenneth and Mamie Clark's study showed a preference for White dolls among Black children, and influenced the 1954 *Brown vs Board of Education* United States Supreme Court decision to desegregate schools and other institutions. Earlier theories about race and culture differences ceded to theories that explained belief in race difference as a form of *prejudice*, driven by authoritarian personalities in Adorno, Frenkel-Brunswik, Levinson, and Sanford's *The Authoritarian Personality* (1950), or the intersection of personality with stereotyping in Gordon Allport's *The Nature of Prejudice* (1954).

Milgram's notorious demonstration of obedience to authority emerged from this context. In Milgrams' experiments, experimental variables were manipulated affecting the extent to which naïve subjects would be willing to obey an experimenter's authority and to deliver what appeared to be lethal electric shocks to a stranger in an adjacent room. Milgram (1974) presented his work as an explanation of the banality of evil shown during the Holocaust. In his account, it was not authoritarian personalities, or German culture that could explain this most unfathomable of historical events. Rather, forces in the immediate social situation, comprehensible as vectors in Lewinian fields, held explanatory theoretical value.

Theories of the constructive active person were suggested by Lewin's theory by virtue of its phenomenological roots. Social psychological theory of this period was responsible in part for the slow shift away from behavior toward cognition as the object of theories to be tested by experimental psychology. In the early 1960s, the theories of Jean Piaget, the first English translations of Lev Vygotsky, and Bandura and Walters 'socio-behaviorist' theory of how children learn through observation all suggested a more active child than the tabula rasa of John Watson's behaviorism. Theory was increasingly understood as developing through hypothetico-deduction rather than induction, as philosophers such as Karl Popper and Thomas Kuhn critiqued the logical positivist assumption that induction and observation moved scientific theory forward. Leon Festinger's

theory of cognitive dissonance stimulated enthusiasm for experiments to test competing theories. Festinger theorized 'cognitive dissonance' as an unpleasant arousal state brought on by 'dissonant' thoughts, or dissonance between a thought and an action. Early experiments demonstrated the failure of parsimonious behaviorist theory to explain changes in attitudes without recourse to a need for 'dissonance reduction.' For example, Festinger and Carlsmith showed that young men reported linking a boring task particularly after they were given an insufficient monetary reward for leading another student to engage in it. However, Daryl Bem's 1967 self-perception theory invoked Occam's razor against Festinger's theory. Bem showed that some experimental results explained by 'dissonance reduction' could be attributed to the more simple process by which people infer their own attitudes from their behavior as if they were observing that behavior in another person. Several attempts to construct critical experiments to test the two theories followed.

Crisis and the Rise of Attribution Theory

The unprecedented growth of experimental social psychology prompted interrelated 'crises' toward the end of the 1960s (Collier et al., 1991). The first issue concerned the construct validity of experimental methods. Drawing on hypnosis studies, and echoing the questions that stalked Charcot, Martin Orne theorized the 'demand characteristics' by which participants try to fulfill experimenter's expectations of how they should act, as in Milgram's experiment, for example. Rosenthal and Rosnow (1969) showed dramatic unintended effects of such expectations on the results of human and animal experiments.

Second, the ethics of experiments, such as Milgram's, that deceived participants about purposes and procedures divided researchers. Some were happy to design experiments that scripted the use of electric shocks and orders to eat cockroaches, but Kelman's 1967 critique described deceptive methods as unethical and undermining of trust in institutions. Third, the *relevance* of theories and the shrinkage of theoretical ambition prompted anxiety, as debates about dissonance and attitude change led researchers to conduct 'experiments about experiments' rather than experiments about phenomena. Finally, Kenneth Gergen (1973) convincingly argued that history could modify all attempts to find universal laws of social behavior at the Lewinian level of analysis.

During this growth period, specifically *sociological* social psychologies grew more slowly and did not participate in the same level of crisis. Under Herbert Blumer's influence 'the Chicago school of symbolic interaction' emphasized how people act on the basis of meanings arising during social interaction and modified through interpretation. Symbolic interactionists eschewed quantification for reflexivity; insisting that research was also a process of symbolic interaction that social scientists must approach via 'sensitizing concepts.' Psychologists were most influenced by the writings of Erving Goffman (1922–82) among sociologists of this period. More skeptical approaches to generalizable social theory such as Glaser and Strauss' grounded theory, Harold Garfinkel's ethnomethodology, or Berger and Luckman's social constructionism influenced only those psychologists most determined that the crisis in experimental social psychology was a insurmountable.

Within the United States, *attribution theory* emerged as the theory to test experimentally from the crisis, shifting attention from the behavior of the person in the situation, to the cognitive *attributions* for behavior to either the person or the situation. Kelly's theory of covariation analysis (1967), Jones and Davis' *correspondent inferences* (1965), Jones and Nisbett's actor-observer effect (1972), and Ross (1977) *fundamental attribution error* kept Lewin's formulations central to social psychology, and responded to all four aspects of the crisis. Experimental methods could now claim to access cognitive meaning-making just as symbolic interactionists might do. Pencil and paper measures of attribution raised fewer ethical risks of earlier experiments on dissonance, or obedience. By describing people's behavior as subject to attributional 'biases' away from rationality, the theory implicitly argued for the relevance of a scientific study of social behavior to everyday life. Indeed, psychologists picked up their statistical tools and made theories of them, most obviously in work on judgment and decision-making such as that gathered in the 1982 Kahneman, Slovic, and Tversky volume, *Judgment Under Uncertainty; Heuristics and Biases*. During the 1970s and 1980s, cognitive theories were applied to an ever-widening range of objects in American social psychology, particularly the self. Theories of self-schemas, self-presentation, self-discrepencies, self-affirmation, and possible selves flourished in the 1980s, often looking back to James, Baldwin, and Mead, but always imagining the self as an individual cognitive representation, however the contents of that representation might originate in some 'social' field. The self was increasingly theorized as relevant to health; optimism; and 'positive illusions' were healthy, as in James' pragmatism. Troubling claims about individualism were largely ignored by 'social cognition' theories (Pepitone, 1981), but growth in social psychology elsewhere troubled their foundations.

European independence from culture-bound individualist American theories was formally declared by Israel and Tajfel's *The Context of Social Psychology* (1972). European social psychological theories tended to reject the valorization of the individual over the collective mind. First, Serge Moscovici's *La Psychanalyse, son image, son public* (1961) had introduced 'social representations' which used Durkheim to theorize the relationship between individual and collective representations, focusing on how specialized knowledge becomes widespread and shared. Second, in 1971, Tajfel published, with others, 'Social categorization and intergroup behavior.' A rejoinder to Sherif's (1966) Realistic Group Conflict theory, Tajfel's 'minimal groups' experiment showed that people would sacrifice total net gain to express a discriminatory prejudice in favor of their in-group over an out-group. The minimal groups experiment anchored Tajfel's Social Identity Theory of how individual's behaved differently because of their status as group members (Brown, 2000). Third, Rom Harré's 'ethogenic' approach to social observation drew on Goffman and other sources to describe how meaning is made up in real social environments. Marsh, Rosser, and Harré's study of violent group behavior during British soccer matches; *The Rules of Disorder* (1978) described collective behavior as ordered in

ethogenic terms rather than as a 'deindividuated' crowd of LeBon's sort.

In recent decades, theorists have varied as to whether they seek to explain the age-old disparity between attitudes and behavior, or to deconstruct it. The theory of reasoned action resolved the problem of predicting behavior from attitudes by theorizing a moderating effects of subjective norms between the two, becoming a hegemonic unifying theory of health interventions. In contrast, discourse analysis and discursive psychology have tended to deconstruct the boundary between attitudes and behavior. Following Ludwig Wittgenstein, British ordinary language philosophers withdrew interest in the 'truth status' of language, describing words as forms of action. This understanding informed Jonathan Potter and Margaret Wetherell's argument in their 1987 book *Discourse and Social Psychology*, that attitudes were of interest to social psychologists only when expressed in discourse, and hence were actions or 'behavior' in their own right. In contrast to dissonance theory, discourse analysts insist that talk is riven with contradiction that people barely notice, and 'discursive psychology' extended an understanding of social psychology without variables to memory, self, gender, and other topics (Potter, 2011).

By the 1990s, theorizing became more global as Asian experimental social psychology expanded and classic North American theories were shown to be culturally particular as in Joan Miller's early study of attributions among American and Hindu Indian adults and children. 'Universal' theories of social psychology developed within Western Industrialized Educated Rich and Democratic (WEIRD) nations are actually culturally particular. Theorists now often strive for innovation by abandoning the privileged position that WEIRD nations have in theoretical formations, as Moghaddam (2013) has recently done with democracy. Contemporary attempts to use the distinction between 'individualist' and 'collectivist' countries as a unifying framework for cultural psychology have garnered significant critique.

Simultaneously, the embodied organism has returned to challenge social cognitive theories. Instinct theories returned through evolutionary psychology, following Leda Cosmides' (1989) argument that humans evolved to reason about Locke-style social contracts. Explanations of autistic people as lacking a 'theory of mind' linked social cognition and neuroscience, focused attention on the mirror neuron system as the embodied basis of social cognition that resemble James' ideomotor action, Baldwin's imitation, and McDougall's sympathetic induction. Currently experiments showing how the implicit meanings of stimuli are evident in embodied ways are favored.

Conclusion

We conclude that unity in social psychology theory is an animal often stalked but rarely captured. Among social scientists, social psychologists have been relatively untouched by dialectical theories that emphasize how all types of dominant knowledge come to self-consciousness through conflict with their 'others' (Liu and Liu, 1997; Plon, 1975). However, hegemonic theories from Lewin's field theory to Festinger's cognitive dissonance have recognized individuals as being moved by environments in dialectical ways to new self-understanding and conduct. We cannot conscientiously say what theories would be possible if this irony in the science of social psychology were made more explicit.

See also: Attitude Formation and Change; Authoritarian Personality; Cognitive Dissonance; Collective Behavior, Social Psychology of; Critical Psychology; Cross-Cultural Psychology; Cultural Psychology; Embodied Social Cognition; Heuristics in Social Cognition; Intergroup Relations; Levels of Analysis in Social Psychology; Obedience: Social Psychological Perspectives; Persuasion Theories; Racism: Social Psychological Perspectives; Social Cognition; Social Comparison, Psychology of; Social Constructionism; Social Psychology: Research Methods; Social Psychology.

Bibliography

Allport, Gordon W., 1954. The historical background of modern social psychology. In: Lindzey, Gardner (Ed.), Handbook of Social Psychology, vol. 1. Addison-Wesley, Reading, MA, pp. 3–56.

Argyle, Michael, 2001. The development of social psychology in Oxford. In: Bunn, Geoff C., Lovie, A.D., Richards, Graham D. (Eds.), Psychology in Britain: Historical Essays and Personal Reflections. BPS Blackwell, Leicester, UK, pp. 333–343.

Ash, Mitchell G., 1995. Gestalt Psychology in German Culture, 1890–1967. Cambridge University Press, Cambridge, UK.

Billig, Michael, 2008. The Hidden Roots of Critical Psychology: Understanding the Impact of Locke, Shaftsbury, and Reid. Sage, London.

Bem, D., 1967. Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychological Review 74, 183–200.

Brown, Rupert, 2000. Social identity theory: past achievements, current problems, and future challenges. British Journal of Social Psychology 30, 745–778.

Cherry, Frances, 1995. The 'Stubborn Particulars' of Social Psychology: Essays on the Research Process. Routledge, New York.

Collier, Graham, Minton, Henry L., Reynolds, Gary, 1991. Currents of Thought in American Social Psychology. Oxford University Press, New York.

Cosmides, L., 1989. The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition 31, 187–276.

Danziger, Kurt, 2000a. Making social psychology experimental: a conceptual history: 1920–1970. Journal of the History of the Behavioral Sciences 36, 329–347.

Danziger, Kurt, 2000b. Naming the Mind: How Psychology Found Its Language. Sage, London.

Fancher, Raymond E., Rutherford, Alexandra, 2012. Pioneers of Psychology, fourth ed. Norton, New York.

Farr, Robert, 1996. The Roots of Modern Social Psychology: 1872–1954. Blackwell, Oxford, UK.

Gergen, Kenneth J., 1973. Social psychology as history. Journal of Personality and Social Psychology 26, 309–320.

Gilman, Sander, 1990. Freud, Race, and Gender. Princeton University Press, Princeton, NJ.

Gold, Martin (Ed.), 1999. The Complete Social Scientist: A Kurt Lewin Reader. American Psychological Association, Washington, DC.

Greenwood, John D., 2004. The Disappearance of the Social in American Social Psychology. Cambridge University Press, Cambridge, UK.

Hacking, Ian, 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.

Hovland, C.I., 1959. Reconciling conflicting results derived from experimental and survey studies of attitudes change. American Psychologist 14, 8–17.

Jahoda, Gustav, 2007. A History of Social Psychology, from the 18th Century Enlightenment to the Second World War. Cambridge University Press, Cambridge, UK.

Kruglanski, Arie W., Stroebe, Wolfgang (Eds.), 2012. Handbook of the History of Social Psychology. Psychology Press, New York.

Kelman, H.C., 1967. Human use of human subjects: The problem of deception in social psychological experiments. Psychological Bullletin 67, 1–11.

Liu, James H., Liu, Shu-Hsien, 1997. Modernism, postmodernism, and neo-confucian thinking: a critical history of paradigm shifts and values in academic psychology. New Ideas in Psychology 15, 159–178.

McDougall, William, 1909. An Introduction to Social Psychology, second ed. Methuen, London (First published 1908).

Milgram, Stanley, 1974. Obedience to Authority: An Experimental View. Harper & Row, New York.
Moghaddam, F.M., 2013. The Psychology of Dictatorship. American Psychological Association, Washington DC.
Pepitone, Albert, 1981. Lessons from the history of social psychology. American Psychologist 36, 972–985.
Plon, M., 1975. On the meaning of the notion of conflict and its study in social psychology. European Journal of Social Psychology 4, 389–436.
Potter, Jonathan, 2011. Re-reading discourse and social psychology: transforming social psychology. British Journal of Social Psychology 51, 436–455.
Rosenthal, H.C., 1967. Human use of human subjects: the problem of deception in social psychological experiments. Psychological Bulletin 67, 1–11.
Rosenthal, R., Rosnow, R.L., 1969. Artifact in behavioural research. Academic Press, New York.
Richards, Graham, 2010. Putting Psychology in Its Place: Critical Historical Perspectives, third ed. Routledge, East Sussex.
Sherif, M., 1966. The psychology of social norms. England, Oxford.
Shields, Stephanie A., Bhatia, S., 2009. Darwin on race, gender, and culture. American Psychologist 64, 111–119.

Social Psychology

Xenia Chryssochoou, Panteion University of Social and Political Sciences, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The historical roots, conceptualizations of the discipline, research practices, and institutional support of social psychology are discussed herein. This article attempts to present the discipline and the particular way it looks at phenomena that are simultaneously individual and social. Social psychology is considered as an interdisciplinary science that challenges the distinction between the individual and the social, and is an empirical science with diverse methodology and important applications.

How to Define an Area of Theoretical and Empirical Interest?

Many social psychologists, when asked about their job, must have faced the awkward situation of having to define the discipline to nonpsychologists. People seem to more or less understand what psychology is about (although this representation may be limited to what clinical psychology does) and what sociology does. However, they have difficulty in figuring out what social psychology is. Moreover, in their attempt to explain what they do, social psychologists are often either too precise, describing their own area of expertise, or too broad, speaking about "phenomena relating to people in groups and social categories." This difficulty is not due to the vagueness of the discipline or to the incapability of social psychologists to give a comprehensive definition of their area. This difficulty is common to the definition of every abstract notion and depends very much on the way a community describes and prescribes the issues that concern it. This difficulty relates to the history of social psychology community, its traditions and practices, and the institutions that support its existence. In other words, it is also difficult to define the content of the category as being 'American' or 'French' or 'Greek.' For this we rely on the consensual and contested understandings of the community of people that define themselves as such, their common history, practices and traditions, and the legal and institutional ways that recognize their existence. Such definitions are both descriptive and prescriptive. Social psychology, therefore, can also be defined by the way social psychologists understand and approach their area of expertise, its historical roots, its empirical traditions and practices, and the way institutions recognize it. In this article these issues are briefly discussed in order to present what social psychology does and to connect it with the social psychological articles in this volume.

Historical Roots and the Development of a New Scientific Endeavor

Before considering the definitions that social psychologists give, it is important to talk about the historical roots of this discipline. Social psychological phenomena were experimentally researched at the end of the nineteenth century and in the early years of the twentieth century. Triplett (1898) in the United States and Ringelmann (1913) in France researched the effect of the presence of others on the outcomes of individual or collective work. Triplett, with the *social facilitation effect* (term attributed to the effect by Floyd Allport in 1924), demonstrated that the mere presence of others increased individual performance, whereas Ringelmann showed that individual members of a group become less productive as the size of their group increases. Both of these phenomena, showing the importance of the presence of others in individual behavior, were further researched by many social psychologists.

These experiments were carried out in a period in which there was a great deal of interest in developing and increasing people's performance at work. Social psychology is very much linked to societal questions and concerns that lead to empirical research and theoretical developments. In a period of industrial development where the focus was on performance at work, Mayo (1945) proposed that social motives could also increase people's performance, even under poor working conditions, if workers made sense of the situation as a demonstration of interest on the part of their employers (Hawthorne effect). Thus, the way people make sense of a particular situation impacts their behavior. These early experiments can be considered as the first social psychological investigations where the presence of others (real or symbolic) and the meaning given to a particular situation influenced individual behavior.

The late years of the nineteenth century were also characterized by the work of Le Bon (1895) (*see* Collective Behavior, Social Psychology of; Social Movements: A Social Psychological Perspective; Deindividuation, Psychology of) and Tarde (1890, 1898), who greatly influenced social psychological work on masses and social influence. In this period when the masses came to the forefront of history, the work of Le Bon characterized crowds and collective behavior as irrational in comparison to the rationality of individual behavior, and the work of Tarde restricted influence to imitation. The division between the individual and the collective was established and continues to haunt social psychology. Even so, both scholars claimed that there are phenomena that could not be approached in a similar way to the one used for individual behavior.

For many, Wundt's *Völkerpsychologie* (1900–1920), with its interest in customs, language, and practices, is also viewed as a precursor of modern social psychology (Doise, 1995; Farr, 1996). Wundt's work influenced many American psychologists who came to Europe between 1865 and 1914.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24095-6

Farr (1996: p. 36) observes: "the inheritance from Wundt was an experimental psychology that was not social and a social psychology that was not experimental. If psychology first became an experimental science in Germany, then it was social psychology that became an experimental science in America."

Modern social psychology owes much also to the work of French sociologist Emile Durkheim, who in the second edition of his book on the division of labor in society (1893) asks a profoundly social psychological question: "What explains the fact that while becoming more autonomous, the individual becomes more closely dependent on society?" (cited by Doise, 1995). From the Durkheimian concept of collective representations, Moscovici (1961/76, 2008 for the English edition), later, develops the concept of social representations that continues to attract social psychologists around the world (*see* Social Representations). Two textbooks of social psychology were published in 1908, in Britain by McDougall and in the United States by Ross, testifying that this new discipline was born. During this period in the United States, Floyd Allport dominated the field, expressing an individualist, positivist, and behaviorist approach to the study of social psychological phenomena.

What we can learn from this brief account of the very early days of social psychology is that, contrary to assumptions (Jones, 1985/1998), the discipline developed on both sides of the Atlantic with scholars trying to understand the relationship between mind and society, the individual and the social. These early days were characterized mainly by behaviorist influences dominant in psychology at the time.

This dominant behaviorist approach was challenged by social psychologists, many of whom fled Europe in between the two world wars and developed a social psychology influenced by the Gestalt. Indeed, in the 1930s many social psychologists moved to the United States and brought with them a tradition of research influenced by the Gestalt, where relationships were the focus of attention. Koffka's *Principles of Gestalt Psychology*, for example, introduced this approach to an American public. A very prominent figure for social psychology was Lewin (1936), who proposed that behavior was a function of the interaction between the person and the environment. He applied this equation in the analysis of group dynamics and was one of the first to point out that a group is not the mere sum of its members. This approach was echoed by other social psychologists like Muzafer Sherif (autokinetic effect, 1935), who showed that norms were developed by groups in response to uncertainty and that they were collective outcomes beyond individual members' decisions. This approach was also important in the development of Heider's theory of causal attribution, balance theory, and interpersonal relationships (1958) and Asch's Impression Formation (1946). The Gestalt theorists established social psychology as an empirical and experimental science, emphasized relationships, and moved the focus of attention from behavior to social perception, albeit "Individualizing the Social" according to Farr (1996). At the same period, George Herbert Mead's courses, later published (1934) in *Mind, Self and Society*, established symbolic interactionism as an important field in social psychology and introduced the self as a social mediator between the mind and society. In this prewar period, social relationships and interactions emerged and became the focus of attention of social psychological work. Social psychology oscillates between an emphasis on individual behavior and the importance of the social context in which actions and relationships take place. This is the time when a clear division appeared between what is called a 'psychological' social psychology and a 'sociological' social psychology. It is also the time when experimentation was linked to social psychology as a defining feature of the discipline.

During World War II, much work was devoted to the study of the attitudes, thoughts, feelings, and adjustment of the soldiers. The *American Soldier* (Stouffer et al., 1949) was an immense research project that sealed the link between social psychology and the military, who continued to fund research and was instrumental in the development of social psychology on both sides of the Atlantic (Moscovici and Markova, 2006). The war created political pressure for action research, and Lewin's work (1947) on reorienting food habits away from food shortages to help with the war efforts is considered as an application of social psychological work on group dynamics.

Social psychology developed rapidly in the early years after World War II. The world was horrified by the brutality of the war and wanted to understand how this was possible. In the 1950s, Hovland at Yale researched propaganda, and Asch conducted his famous experiments on majority influence (*see* Persuasion Theories; Attitudes and Attitude Change). Adorno et al. (1950) studied the development of the "authoritarian personality." Also in 1952, Asch published *Social Psychology*, in which he set the goal for scientific research in psychology "to establish functional relations between conditions in the surroundings and psychological processes and between one psychological process and another" (p. 8). The appearance of Heider's book on interpersonal relationships (1958) and the famous Robber's Cave experiments (showing that the type of interdependence between groups impacts on social perception and behaviors) by Sherif et al. (1954) kept alive the Gestalt tradition in social psychology. Festinger's cognitive dissonance theory (1957) surprised the field and reversed the attitude–behavior relationship. This is the time (1966) that the European Association of Experimental Social Psychology (EAESP) was created with the support of American funds and efforts (Moscovici and Markova, 2006).

In the 1970s and the early 1980s, social psychologists continued the postwar tradition and were still interested in understanding prejudice and discrimination. Moreover, the famous Milgram experiment on obedience (*see* Obedience: Social Psychological Perspectives) made social psychological research and practices well known to the general public. In the presence of movements such as for civil rights, new themes emerged relating to group behavior and social change. Intergroup relations research by Henri Tajfel's team in Bristol (UK) and minority influence research by Moscovici in the Ecole des Hautes Etudes en Sciences Sociales in Paris marked the field and still continue to inspire research (*see* Intergroup Relations; Social Influence, Social Psychology of). Tajfel and colleagues were interested in understanding social change that occurred when people identified with social groups and acted as their members. They developed the notion of social identity that mediates the relationship between the individual and the group and is responsible for collective behaviors (*see* Social Identity in Social Psychology; Identity and Identification). In

Europe, in addition to minority influence research, Moscovici developed the concept of social representations from the Durkheimian concept of collective representations in order to understand social change. He was interested in how science becomes common sense, and how social knowledge develops and allows communication, conflict, and consensus to emerge (*see* Social Representations). Even today, these schools of thought continue to inspire research not only in Europe and the United States but also in Australia and Latin America.

This period also saw the rise of social cognition. This approach soon became dominant and helped social psychology to overcome its behaviorist past. Undoubtedly, social cognition contributed greatly to the study of social perception (*see* Social Cognition; Stereotypes in Social Psychology; Heuristics in Social Cognition). However, although it allowed psychology to overpass its behaviorist past, soon, by following the metaphor of the computerized mind became too individualized and social psychology was accused of having lost its societal grounding. Social cognition was criticized as ideologically driven, presenting an individual thinker whose cognitions defined the world, a cognitive miser, biased, using shortcomings to perceive and decide (Sampson, 1981). In 1972, Israel and Tajfel edited *The Context of Social Psychology: A Critical Assessment*, in which many prominent scholars tried to explain the distance that separated social psychological work from the social context. The publication of *The Social Dimension* in 1984, edited by Tajfel, was another attempt to find the lost 'social' in social psychology. These debates are still alive today in the social psychological community. It is also in this context that social psychology was challenged by cross-cultural psychology, when, for example, in 1988 Jahoda "accused" social psychology of being blind to cultural context and influences (*see* Cross-Cultural Psychology). In the same period, social constructionist approaches (*see* Social Constructionism) challenged the epistemology and methodology of the dominant sociocognitive approach and developed a series of social psychological investigations using qualitative methodology (*see* Social Psychology: Research Methods).

Our discussion of the current era of social psychology is purposefully brief. It is impossible to do justice to all theoretical accounts and to mention all the scholars who do social psychological research. This can be seen in the different articles of this encyclopedia. Moreover, this brief historical narrative cannot be exhaustive. Interested readers can consult the many historical accounts (e.g., Doise, 1995; Farr, 1996; Papastamou, 2002), and the article Social Psychological Theories, History of. An interesting account of the genesis of social psychology is proposed by Moscovici and Markova (2006) who, along with the theoretical developments, present the institutional construction of social psychology based on archival data. The purpose of our brief presentation is to give the reader the opportunity to trace the steps of social psychology in order to understand its current status and to grasp its domain.

The Domain of Social Psychology

As we have seen, over the years social psychology developed as an empirical, mainly experimental, science, emphasizing individual behavior and cognitions in context. Two brothers were instrumental in defining the discipline: Floyd and Gordon Allport. In 1924, Floyd Allport published a book on social psychology where he gave the following famous definition of the field (p. 4):

> There is no psychology of groups which is not essentially and entirely a psychology of individuals. Social psychology must not be placed in contradiction to the psychology of individuals; it is a part of the psychology of the individual, whose behavior it studies in relation to that sector of his environment comprised by his fellows ... There is likewise no consciousness except that belonging to individuals. Psychology in all its branches is a science of the individual.

Floyd Allport was clearly suggesting that social psychology is not an interdisciplinary field but a subdiscipline of psychology. His definition was instrumental to individualistic accounts and to the ideologically driven denial of society as anything more than a collection of individuals. Thirty years later, his brother Gordon Allport (1954, p. 5) gave another definition that is perhaps still the most cited definition in textbooks of social psychology:

> With few exceptions, social psychologists regard their discipline as an attempt to understand and explain how the thought, feeling and behavior of individuals are influenced by the actual, imagined or implied presence of others.

This definition puts processes of social influence and social perception to the forefront of social psychology. However, the individual still remains preeminent and many collective processes of knowledge production and social action are not included in this account.

The emphasis on the individual can be overcome if one articulates different levels of explanation, as proposed by Doise (1986). He suggested that social psychologists use different levels of explanation to make sense of the phenomena they study: (1) an intraindividual level where phenomena are explained by internal individual processes (see, for instance, categorization); (2) an interindividual and situational level where the dynamics of interaction between individuals explain the phenomena; (3) a positional (or intergroup) level where the explanation focuses on the position and status of individuals as members of groups with asymmetrical relationships; and (4) an ideological level where the explanation is based on the systems of social beliefs, representations, evaluations, and norms. In a more recent account of the levels of explanation, Doise (2012) adds two further levels: one where the explanation concerns biological accounts such as those used in social neuroscience studies, and a societal level, which includes explanations based on norms that function at a societal level such as human rights.

More than a typology of explanations, Doise's contribution to the definition of the domain of social psychology is his claim to articulate the different levels of analysis and explanation in order to avoid reductionist accounts of the different phenomena (*see* Levels of Analysis in Social Psychology).

Until now we have tried to circumscribe the domain of social psychology by presenting an historical account, famous definitions, and different levels of explanation. The difficulty with social psychology is that it deals with phenomena that are simultaneously psychological and social. What makes them so is perhaps the particular way social psychologists look at them. In 1988, Moscovici suggested that what differentiates social psychology from other disciplines is not its object but its particular gaze to phenomena and to the relations among them. Its particularity is to substitute to the binary relationship between a subject and an object into a three-party relationship: individual subject (Ego), social subject (Alter), and object (see **Figure 1**). This corresponds to a constant mediation that the social subject (Alter) performs. This relationship between the two subjects (Ego and Alter) impacts their relationship to the object and can be a simple copresence (as in the social facilitation effect presented earlier) or an interaction (as in the case of social influence).

For example, the relationship of an individual to his/her performance could be mediated or moderated (interaction) by the presence of others (real or symbolic). Moreover, interestingly, in the place of Ego and Alter one can substitute the individual with a group. This three-party gaze proposed by Moscovici makes the phenomena simultaneously individual and social. In recent years social psychologists have used this three-party relationship to explain, for instance, minority influence (Mugny, 1982) and politicized collective identity (Simon and Klandermans, 2001). If people look carefully at the work of social psychologists, they will notice that many concepts (like identity) act as mediators and that social psychological hypotheses are often hypotheses of mediation or moderation (*see* Social Psychology: Research Methods). Moderation and mediation hypotheses always involve three factors. In the case of moderation/interaction, the level of the third factor (the moderator) changes the relationship between the other two factors (the relationship between ego and object). This is a hypothesis of interaction that is usually formulated to answer a 'When' question: *When* would individuals be more prejudiced toward out-group members? When their in-group is of high or of low status? A hypothesis of mediation implies that the relationship between two factors depends on (is created by) a third factor. This hypothesis replies to questions of 'how' and 'why': *Why* people are prejudiced toward out-group members; it is because they identify with their own group that they wish to favor. These are simple examples, but social psychological literature is full of such hypotheses and sophisticated statistical techniques allow their testing (Baron and Kenny, 1986).

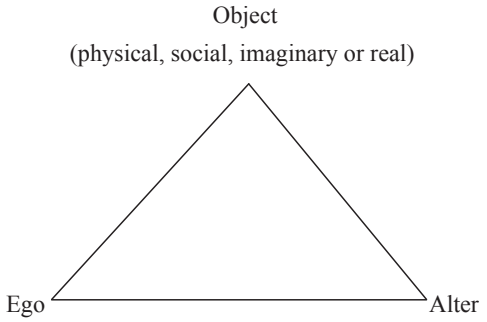

Figure 1 Moscovici's social psychological gaze.

From the previous discussion, we could say that the domain of social psychology is characterized by different levels of explanation whose articulation is aimed and by a particular gaze to phenomena that involve relationships of mediation or moderation. Within this framework, social psychological work aims to understand social perception, individual and collective behaviors, processes within groups, processes between groups, intraindividual and interpersonal processes, the production of social knowledge, ideologies and processes of communication, persuasion, and influence in order to understand how individuals and societies develop and change. To do so, social psychologists have elaborated concepts such as attitudes, stereotypes, social categories, beliefs, norms, identities, values, motives, collective memory, and social representations in order to approach different phenomena. There are articles in this encyclopedia concerning all these concepts.

Over the years social psychologists have worked on: intergroup relations, tyranny, sexism, racism, prejudice, leadership, decision making, risk, love and attraction, immigration and ethnicity, acculturation, prosocial behavior at an individual and intergroup level, aggression, conflict, xenophobia, dehumanization, stigma, obedience, collective behavior and action, social movements, mass killings and genocide, obedience, justice issues, and desegregation. For all these themes, there are specialized articles in this encyclopedia. There are also articles of theoretical interest to social psychologists such as about social cognition, social categorization, self-categorization theory, social attribution, essentialism, social constructionism, social representations theory, and cognitive dissonance.

Moreover, besides sociology and psychology that are part of its identity components, social psychology is closely linked to cross-cultural, indigenous, and political psychology. Culture and politics are two different fields within which social psychological phenomena evolve, and often social psychologists are interested in specializing in these areas. Thus, the interested reader can also find articles on cross-cultural and cultural psychology, indigenous psychology, culture and the self, culture and emotions, interpersonal relationships and culture, and political psychology.

Some questions that social psychologists try to answer are:

- How people understand and construct their social reality?
- How they feel in their environment?
- What motivates their actions?
- How and when they act together?
- The consequences of these actions?
- How people and groups influence each other?
- How people and groups produce knowledge norms and artifacts?

Social Psychological Practices, Ethics, and Institutional Recognition

From its very beginnings, social psychology was conceived as an empirical science. Initially, social psychology was mainly

experimental. This practice formed generations of social psychologists and characterized the discipline. There has been, however, criticism about the preponderance of experimentation (see also Gergen, 1973, 1995), and the development of social constructionist approaches opened the field to qualitative methods and epistemology. Moreover, survey methodology has been widely used and sophisticated statistical techniques have been developed to test social psychological hypotheses (for a comprehensive account, *see* Social Psychology: Research Methods). Nowadays, social psychologists use a variety of methods including techniques of neuroscience. Social psychological research is no longer characterized by a particular methodology. What does characterize the discipline is the importance that is given to empirical findings either to test a theoretical hypothesis or to help develop new theories.

The use of experimentation with human participants raises ethical issues. These concerns are raised in particular because experimental studies have sometimes used deception and not revealed initially to participants the real purpose of the study, in order to create the desired conditions of the social context. Famous studies such as the Milgram experiment and the Stanford Prison experiment by Zimbardo (1973) were criticized for the conditions to which they exposed their participants. Today, ethical concerns are very seriously taken into account, and ethical committees within universities review research projects and make sure that all ethical standards are kept. Important steps are taken to get participants' informed consent and to fully debrief them at the end of the study. Associations and academic societies have been involved in making public ethical codes of conduct (examples of these codes of conduct for research are the American Psychological Association standard 8: research and publication http://www.apa.org/ethics/code/index.aspx?item=11 and the British Psychological Society http://www.bps.org.uk/sites/default/files/documents/code_of_human_research_ethics.pdf see also the site of the European Union http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1433).

Nowadays, there are also debates about replication issues, and journals devote special editions looking to strengthen the discipline by assuring replication (see, for example, a 2012 special edition of Perspectives on Psychological Science). However, for some types of contexts and methodologies, replication is not the important issue. What matters is the understanding of the relationships and the phenomena occurring in a particular context. Social psychology as an alive community of research practice will continue to discuss these issues and enrich the scientific process.

At an institutional level, despite the interdisciplinary nature of social psychology, most scholars work in psychology departments and are trained through a psychology curriculum. Social psychology courses are considered core modules in degrees of psychology. There are many associations and academic societies of social psychology that organize conferences, academic activities, and summer schools for young researchers, such as: *Society for Personality and Social Psychology*, *Society of Experimental Social Psychology*, *European Association of Social Psychology*, *Society for the Psychological Study of Social Issues*, *Asian Association of Social Psychology*, *Society of Australasian Social Psychology*, *Canadian Psychological Association: Social and Personality Section*, *British Psychological Society: Social Psychology Section*, *Association pour la Diffusion de la Psychologie Sociale*, and many national associations. Moreover, social psychologists publish their work in a multitude of journals (for example, *Asian Journal of Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, *Canadian Journal of Behavioral Science*, *European Journal of Social Psychology*, *Group Processes and Intergroup Relations*, *Journal of Applied Social Psychology*, *Journal of Community and Applied Social Psychology*, *Journal of Experimental Social Psychology*, *Journal of Personality and Social Psychology*, *Journal of Social Psychology*, *Journal of Social and Political Psychology*, *Personality and Social Psychology Bulletin*, *Personality and Social Psychology Review*, *Social Psychological and Personality Science*). The multitude of learning societies and journals testify to the rich scientific environment of social psychology and the depth of institutional support it has.

Social Psychology: An Ongoing Scientific Endeavor

We set out to define social psychology using historical information, definitions, and conceptualizations of its domain by social psychologists, and by research practices and institutional recognition. What makes social psychology distinct is a combination of what it studies, how it studies, and the level of explanation it seeks. Social psychology is interested in understanding how people are transformed by society and how they transform society (Chryssochoou, 2004). It is interesting to study stability and change in societies through the study of processes that involve the interaction of the individual and the social groups. Hopefully, by the end of this article, the reader will understand that social psychology seeks to answer a fundamental question: why there is a conflict, a division between the individual and society. It, therefore, can be the science that challenges this division and as a consequence the ideology that it supports. It is the science whose studies are placed at a mesolevel of analysis that bridges the macrosocietal-cultural level with the micro-individual one. Through interactions, communication and influence processes, and through social actions, individuals and groups produce cognitions and representations that join the individual and the social. As Moscovici (2000, p. 114) puts it: *the field of social psychology consists of social subjects, that is groups and individuals who create their social reality (which is in fact the only reality), control each other and create their bonds of solidarity as well as their differences. Ideologies are their products, communication are their means of exchange and consumption and language is their currency*. Hopefully, the multitude of social psychological work at different levels of analysis will help to disentangle the processes involved and the contents produced in our complex societies.

See also: Attitude Formation and Change; Collective Behavior, Social Psychology of; Cross-Cultural Psychology; Deindividuation, Psychology of; Heuristics in Social Cognition; Identity and Identification, Social Psychology of; Intergroup Relations; Levels of Analysis in Social Psychology;

Obedience: Social Psychological Perspectives; Persuasion Theories; Representations, Social Psychology of; Social Cognition; Social Constructionism; Social Identity in Social Psychology; Social Influence, Social Psychology of; Social Movements: A Social Psychological Perspective; Social Psychological Theory, History of; Social Psychology: Research Methods; Stereotypes in Social Psychology.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.

Allport, F.H., 1924. Social Psychology. Houghton-Mifflin, Boston.

Allport, G.W., 1954. The historical background of modern social psychology. In: Lindsey, G. (Ed.), Handbook of Social Psychology, vol. 1. Addison-Wesley, Reading, MA, pp. 3–56.

Asch, S.E., 1946. Forming impressions of personality. Journal of Abnormal and Social Psychology 41, 258–290.

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Chryssochoou, X., 2004. Cultural Diversity. Its Social Psychology. Blackwell, Oxford.

Doise, W., 1986. Levels of Explanation in Social Psychology. Cambridge University Press/Maison des Sciences de l'Homme, Cambridge; Paris.

Doise, W., 1995. Organizing social-psychological explanations. In: McGarty, C., Haslam, S.A. (Eds.), The Message of Social Psychology. Blackwell, Oxford, pp. 63–76.

Doise, W., 2012. The homecoming of society in social psychology. In: Pires Valentim, J. (Ed.), Societal Approaches in Social Psychology. Peter Lang, Bern.

Durkheim, É., 1893. De la division du travail social: étude sur l'organisation des sociétés supérieures. Alcan, Paris.

Farr, R.M., 1996. The Roots of Modern Social Psychology. Blackwell, Oxford.

Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CA.

Gergen, K.J., 1973. Social psychology as history. Journal of Personality and Social Psychology 26, 309–320.

Gergen, K.J., 1995. Social psychology as social construction. The emerging vision. In: McGarty, C., Haslam, S.A. (Eds.), The Message of Social Psychology. Blackwell, Oxford, pp. 113–128.

Haney, C., Banks, W.C., Zimbardo, P.G., 1973. A study of prisoners and guards in a simulated prison. Naval Research Review 30, 4–17.

Heider, F., 1958. The psychology of interpersonal relations. Wiley, New York.

Jones, E.E., 1985/1998. Major developments in social psychology in the past five decades. In: Gilbert, D.T., Fiske, S.T., Lindsey, G. (Eds.), The Handbook of Social Psychology, fourth ed., vol. 1. McGraw-Hill, Boston.

Le Bon, G., 1895. La psychologie des foules. Alcan, Paris.

Lewin, K., 1936. Principles of Topological Psychology. McGraw-Hill, New York.

Lewin, K., 1947. Frontiers in Group Dynamic. Human Relations 1, 2–38.

Mayo, E., 1945. The Social Problems of an Industrial Civilization. Harvard University Press, Cambridge.

Mead, G.H., 1934. Mind, Self, and Society. In: Morris, Charles W. (Ed.), Chicago University Press, Chicago.

Moscovici, S., 1961/76. La Psychanalyse son Image et son Public. Presses Universitaires de France, Paris.

Moscovici, S., 1988. Le Domaine de la Psychologie Sociale. In: Moscovici, S. (Ed.), Psychologie Sociale. Presses Universitaires de France, Paris, pp. 5–22.

Moscovici, S., 2000. Society and theory in social psychology. In: Duveen, G. (Ed.), Social Representations. Explorations in Social Psychology. Polity, Oxford, pp. 78–119.

Moscovici, S., Markova, I., 2006. The Making of Modern Social Psychology. The Hidden Story of How an International Social Science Was Created. Polity, Oxford.

Moscovici, S., 2008. Psychoanalysis: Its Image and its Public. Polity, Cambridge.

Mugny, G., 1982. The Power of Minorities. Academic Press, London.

Papastamou, S., 2002. Why the "new review of social psychology". New Review of Social Psychology 1 (1), 30–53.

Ringelmann, M., 1913. Recherches sur les moteurs animes: Travail de l'homme, Annales de l'Institut National Agronomique, 2e serie, 12, 1–40.

Sampson, E.E., 1981. Cognitive psychology as ideology. American Psychologist 36 (7), 730–743.

Sherif, M., 1935. A study of some social factors in perception. Archives of Psychology 187.

Sherif, M., White, B.J., Hood, W.R., Sherif, C.W., 1954. Study of Positive and negative intergroup attitudes between experimentally produced groups: Robbers Cave Study. University of Oklahoma, Norman.

Sherif, M. (Ed.), 1966. Common Predicament: Social Psychology of Intergroup Conflict and Cooperation. Houghton-Mifflin, Boston.

Simon, B., Klandermans, B., 2001. Politicized collective identity. American Psychologist 56, 319–331.

Stouffer, S.A., Lumsdaine, A.A., Lumsdaine, Harper, M., Williams Jr., R.M., Smith, M.B., Janis, I.L., Star, S.A., Cottrell Jr., L.S., 1949. The American Soldier: Combat and its Aftermath. Princeton University Press, Princeton, NJ.

Tarde, G., 1890. Les lois de l'imitation: étude sociologique. Alcan, Paris.

Tarde, G., 1898. Etudes de psychologie sociale. Giard et Briere, Paris.

Triplett, N., 1898. The dynamogenic factors in pacemaking and competition. American Journal of Psychology 9, 507–533.

Wundt, W., 1900–1920. Völkerpsychologie, 10 vols.

Social Psychology, Sociological

David Rohall, Western Illinois University, Macomb, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Stryker, volume 21, pp. 14409–14413, © 2001, Elsevier Ltd.

Abstract

Sociological social psychology emphasizes the impacts of society on social psychological processes. There are three major perspectives in sociological social psychology: symbolic interaction, social structure and personality, and group processes. The symbolic interactionist perspective emphasizes the ways that individuals construct social reality through social interaction. Alternatively, the social structure and personality perspective emphasizes the relatively stable elements of social structure that impact interactions among people such as roles, statuses, and norms. Finally, the group processes perspective focuses on the ways that societal conditions are recreated in small group environments and how interactions in groups contribute to the maintenance of society. These perspectives are being extended to include the study of sociobiology, the intersection of social and biological conditions that impact social psychological process.

Three Faces of Sociological Social Psychology

Social psychology is the study of human behavior in a social context. Utilizing this definition, researchers from several fields may be considered 'social psychologists' but they approach the field in different ways and utilize different perspectives in their approach to the topic. Some psychologists, for instance, focus on the immediate social contexts of interactions in the study of social psychology while others focus on genetic or internal thought processes occurring during those interactions. They also employ specific psychological perspectives such as social learning theory. Sociologists employ different perspectives and methods in the study of human social behavior than psychologists and sociologists emphasize the societal conditions that impact social psychological dynamics. How, for instance, do people's position in society influence interactions in small groups? How do we learn social roles in society? Sociologists employ a cybernetic approach to social psychology in which individuals interact in ways that reproduce society and society impacts individual-level interactions through norms, roles, statuses, and other elements of society (**Figure 1**).

The three main perspectives in sociological social psychology include symbolic interaction, social structure and personality, and group processes (House, 1977, 1992). Symbolic interactionists traditionally employ qualitative methods to study the social construction of reality, the ways that individual interactions lead to the development of society (Berger and Luckmann, 1967). Social psychologists who employ the social structure and personality perspective typically utilize surveys and other quantitative research techniques to assess the impacts of society on individual-level thoughts, feelings, and behaviors once they have been produced. A third group of scholars in sociological social psychology emphasizes the study of societal processes during group interactions. In some cases, researchers examine how social conditions (e.g., status differences) impact interactions within the group. In other cases, these scholars emphasize the study of how status differences develop within groups. In this sense, groups represent a microcosm of larger society.

Figure 1 Linking levels of analysis in sociological social psychology.

The role of sociobiology in sociological social psychology has been growing over the last decade. Sociobiology is utilized both to understand the relative impacts of genetic dispositions on social psychological processes and to assess how social conditions may impact biological outcomes. Some social scientists have argued that characteristics such as intelligence quotients (IQ), for instance, are genetically determined (Herrnstein and Murray, 1994) while sociologists examine how social conditions impact those things or how humans may define those characteristics as more or less important in society.

Symbolic Interaction

Perhaps the earliest application of the sociological perspective in social psychology comes from the University of Chicago and the work of George Herbert Mead. Mead (1934) and other social psychologists utilizing this perspective employ the pragmatist school of thought from the Scottish moral philosophers such as David Hume and Adam Smith. This philosophy emphasizes reason and empiricism in the search for truth. As applied to the study of human interaction, symbolic interactionists focus on the study of day-to-day interactions rather than the discovery of abstract principles of large-scale social behavior.

Mead (1934) included elements of social learning theory in the development of symbolic interactionism in that humans learn social reality through observation and reinforcement. Children not only learn basic skills such as reading and writing through these processes but also roles and norms, the building blocks of society. Unlike the psychological applications of behavioralism, Mead emphasized the importance of agency, the ability to make independent decisions within the constraints of social conditions. From this perspective, humans are capable of continuity and change; people can choose to

learn and follow norms and roles of previous generations but they can also choose to modify those norms and roles or rebel against them altogether. In many cases, change and continuity are going on concurrently with some individuals (or groups of individuals) modifying existing conditions to various degrees.

While Mead is generally considered the father of the symbolic interaction perspective, the expression was coined by Herbert Blumer (1969). Blumer studied under Mead at the University of Chicago where he taught for 27 years (Rohall et al., 2013). He later took an appointment to the Department of Sociology at the University of California, Berkeley in 1952. Blumer argued that there are three principles of symbolic interaction. First, people create social reality through interaction among individuals. This principle simply states that there is no inherent meaning in any person, object, or behavior – names or labels and their value come from the negotiation of two or more people. Second, people utilize these names and labels to make decisions in life. It is in this sense that both social and physical elements of the world are socially constructed. Finally, Blumer argued that individuals engage in an interpretive process in which the meaning of a particular interaction may vary based on the background of the participants and other factors. If our social world is socially constructed, any number of factors may lead us to different assessments of a situation, including our unique socialization experiences as well as our application of agency.

Given the fluid nature of social reality, Blumer (1969) and other symbolic interactionists typically employ qualitative methods such as ethnographies and personal interviews to discover how individuals come to construct social reality and the processes that lead people to define and redefine important aspects of society (see, for example, Adler and Adler, 1994). Techniques such as the autoethnography give symbolic interactionists the ability to gain insights into meaning-making processes in everyday life (Anderson, 2008).

Just as we construct our social world, we also construct our sense of self through the same processes. Charles Horton Cooley's (1922) looking-glass self emphasizes how people develop their sense of self based on their perceptions of what other people think of them. Just like a mirror distorts images, however, we have a distorted sense of our self because we rely on our interpretations of people's response to us that may be more or less accurate. If we believe that others see us as good, we will believe that we are good and vice versa. Humans also have the ability to manipulate their self-image through a process called impression management (Goffman, 1959). Goffman argued that humans construct an image of themselves in different social arenas that people utilize to make decisions during those interactions. These images may or may not correspond with our sense of self but individuals generally accept self-presentations unless there is reason to believe otherwise. An individual at a party, for instance, may tell people that she is a doctor and people will generally believe her until there is evidence to show that she knows nothing about medicine, perhaps failing to answer a medical question.

Two significant theoretical strands employing the symbolic interaction perspective include identity and social identity theories and affect control theory. Identity theory focuses on the ways in which individuals construct their sense of self through interaction with other people (Burke and Stets, 1999; Stryker, 2002). Identity is defined as the sum total of an individual's roles, traits, and social categories. Individuals develop their sense of self based on their unique dispositions or set of traits – which may or may not be derived from our genetic makeup – but also statuses, roles, and social categories that derive from society. We develop our sense of self through the unique combination of roles and statuses but also by the ways in which we modify those roles and statuses utilizing agency. A similar theory, social identity theory, emphasizes the influences of group characteristics in identity formation (Hogg and Ridgeway, 2003; Tajfel, 1982). Once a group membership (e.g., race or ethnicity) is incorporated into one's sense of self, it initiates other social processes such as judgment of other social groups relative to one's own group. By accepting roles, statuses, and group affiliations, individuals develop a unique sense of self while maintaining important elements of society.

Affect control theory examines the role of emotions in identity processes. In this theory, people apply three sentiments to each person: relative goodness (evaluation), power (potency), and liveliness (activity) (EPA for short). Roles and statuses in society have differing EPA sentiments and serve as a way to evaluate our performance in those roles and statuses (Heise, 2007). An extensive Web site on affect control theory can be found at: http://www.indiana.edu/~socpsy/ACT/.

While most symbolic interactionists focus on individual-level interactions, some symbolic interactions have begun to apply the paradigm to study large-scale social processes such as cultural production and the use of culture once derived. Collective memory includes the shared beliefs about a person or event in society. They may include the memory of particular events that impact only a small number of people (e.g., something that happened at work) or large-scale social events such as the terrorist attacks in the United States on 9 September 2001. Symbolic interactionists examine how individuals in society come to define such events and categorize them (Fine and Beim, 2007). Examining the common narrative around the famous Lewis and Clarke expedition to the American frontier in the 1700s, Denzin (2007) argues that while most Americans categorize this journey as significant and positive in American history, the narrative of the journey can also be interpreted from the perspective of native Americans and others who may categorize the event as tragic, leading to the decimation of Native American life.

Gary Alan Fine has conducted numerous studies of the production of culture ranging from little league baseball teams to weather forecasting organizations (Fine, 2007, 1979). These studies show how individuals at the organizational level incorporate and manipulate larger cultural norms, values, and beliefs to meet the needs of the individuals within organizations. While some norms continue over time, as taught by existing members of an organization, new members introduce new norms while accepting or modifying some of the existing ones, helping to elaborate on larger processes of cultural change and continuity over time.

Social Structure and Personality

The second perspective in sociological social psychology is often referred to as social structure and personality (House, 1977,

1992). Scholars in the social structure and personality perspective generally concur with symbolic interaction principles but emphasize the relatively stable nature of society. Norms, for instance, may change over time but many of them remain over many years. As such, social psychologists can study their impacts on individuals utilizing quantitative methods like surveys and experiments.

House (1992) puts forth three major principles of the social structure and personality perspective. First, the components principle states that we must be able to identify the elements of society most likely to impact individuals' thoughts, feelings, and behaviors. These elements typically include statuses, our relative position in society, roles, expectations about how to act in those statuses, and norms, expectations about how to think, feel, or behave in a particular interaction. Other important components of social structure include social institutions and values. Social institutions are relatively enduring patterns of social behavior in society. Typical examples of institutions include family, work, government, education, and religion. Values are enduring sets of beliefs about the world. These are distinguished from attitudes or public opinion that may be more or less enduring. Many analysts, for instance, argue that Americans have a tendency to value independence and freedom to a greater degree than their peers in other countries (Bellah et al., 1996; Putnam, 2000). These values come from socialization and may impact interpersonal relationships above and beyond individual dispositions.

The proximity principle is based on the idea that people are most impacted by society through their immediate social surroundings (House, 1992). While large-scale events like an economic downturn may have consequences for everyone in society, its immediate impacts are not felt by each individual. Economic downsizing itself, for instance, may have no direct bearing on the mental health of individuals in society except for those individuals who are laid off as a result of the downturn. However, spouses and families of unemployed people may also experience some of its effects (see Dooley et al., 1988; Rohall et al., 2001). Society, from this perspective, primarily impacts social psychological dynamics when it has direct bearing on individuals or indirectly through social relationships with those people.

The psychology principle emphasizes the ways that individuals process proximal experiences. Self-esteem research, for instance, shows that reflected appraisals, our sense of how others view us, are as or more important to individuals' feelings of self-worth than self-perceptions, our observations of our own behavior and its consequences or social comparisons (Schwalbe and Staples, 1991). In this case, our interpretations of others' reactions to us are more important than social conditions themselves.

There are many theories that utilize the social structure and personality paradigm. Life course sociology, for instance, provides a framework for examining the impacts of large-scale social events on individuals in a scientific way (Elder, 1994). There are many facets to life course sociology but they can be distilled into four elements: historical and social contexts, timing, linked lives, and agency. Each birth cohort, people born around the same time period, is exposed to different sets of historical events. However, those events may not impact everyone in a cohort the same way. In his classic study of the Great Depression, Elder (1999, originally 1974) found that only families who were deprived (i.e., lost significant income) appeared to be impacted by the Depression. Further, most of those effects were felt through changing family dynamics. These findings reflect the proximity principle of the social structure and personality perspective in which we experience society – including historical events – through our immediate social environments.

The degree to which social and historical events impact individuals depends, in part, on their relative position in society and when it occurs in their lives. Rosenberg and Pearlin (1978) showed almost no relationship of socioeconomic status and self-esteem among youth (8–11 years of age) but the relationship of these variables became stronger with each successive age group (teenagers and adults). In this case, economic conditions were less relevant to the self-concepts of younger children in the study.

Events impact both people experiencing them and their families and close relationships. This relationship fits under the heading 'linked lives' because they connect or link us to the life events impacting other people. Gartner (2008), for instance, found that people who knew someone injured as a result of the 9/11 terrorist attacks in the US were more likely to disapprove of the president than people who did not know of any casualties.

Life course sociology also incorporates the concept of agency. People can react any number of ways in response to the same conditions albeit limited by those conditions. Deluca and Rosenbaum (2001) found, for instance, that students who worked harder generally did better in school compared to students who worked less hard. However, middle-class students appeared to gain more from their hard work than their peers from lower socioeconomic statuses. Agency may help to explain variations in how people respond to the same historical event.

Another theoretical application of the social structure and personality perspective is the stress process (Pearlin, 1989). The stress processes examine the relative impacts of stressors, resources, and social characteristics on mental health. This model tries to understand differential reactions to the same levels or types of stress (e.g., death of a loved one), representing the third principle of the perspective. In this case, people under distress may react differently to it based on access to things like social support or differing levels of mastery or sense of personal control. If someone with low mastery experiences a stressor, she may respond with greater distress (e.g., depression or substance abuse) than someone with higher levels of perceived control, ostensibly because she believes that she will overcome the problem. Similarly, people with greater access to support networks may respond better to stress than people with fewer supports (**Figure 2**).

The stress process also incorporates elements of the first principle of the social structure and personality perspective by highlighting the ways in which people with different background characteristics process stressful life events. For instance, people from lower socioeconomic backgrounds have both higher levels of stress (e.g., financial strains and poor living conditions) and poorer mental health (Aneshensel and Sucoff, 1996). There are also racial and gender differences in the stress process (Kessler and Neighbors, 1986; Robins et al., 1991).

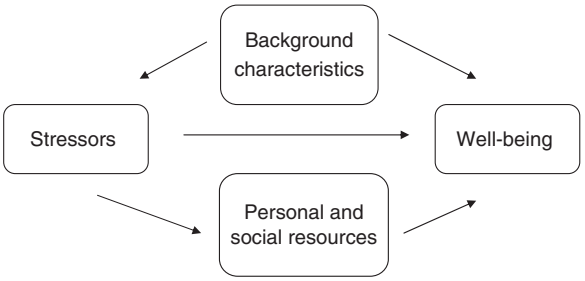

Figure 2 The stress process linking the societal conditions to individual outcomes.

Other applications of the social structure and personality perspective employ multilevel modeling to precisely assess the relative impact of macro- and micro-level conditions that impact individuals in society. These studies employ techniques that allow researchers to separate the relative impacts of actual community conditions from the perceptions of those conditions. Research shows that perceptions of neighborhood conditions (e.g., levels of vandalism, litter, vacant housing) are associated with outcomes like depression (Latkin and Curry, 2003); these findings may reflect the fact that people in those neighborhoods also report fewer social supports to buffer the negative effects of living conditions (Ross and Mirowsky, 2008). Studies that have tried to separate actual community conditions from perceived conditions suggest that most of the impacts of actual community conditions operate through perceived conditions: the impact of living in a poor community appears to operate completely through perceptions of those communities such that people in poor communities report worse perceptions of those communities, which lead to lower levels of well-being (Ross, 2000; Turner et al., 2013). These findings reflect all three principles of the social structure and personality perspective because individuals' positions (i.e., income levels) and proximal environments are processed internally, impacting some people more than others.

Group Processes

The study of group processes emphasizes the ways that basic social processes operate in group contexts. Groups include two or more people who have some level of dependence on each other and small groups are fewer than 20; small groups differ from larger groups in their ability to yield face-to-face interactions among its members. Dyads are two-person groups while triads are three-person groups (Simmel, 1950). This distinction is important because it shows that increasing the size of group exponentially increases the number of relationships while decreasing the levels of intimacy possible in a group. Sociologists distinguish several other types of groups that may impact interactions within those groups. Primary groups include family and close friends who interact on the basis of emotional attachments while secondary groups are those groups we affiliate with to achieve a specific goal or goals (Cooley, 1909). Work groups, teams, and volunteer organizations may provide some emotional bonding but affiliations are more instrumental in nature than in primary groups. An extension of these types of groups is reference groups. Reference groups are ones that may not include any physical interaction but people still employ them in making decisions. Individuals may pay annual dues to an online organization, symbolizing its importance to the person, but never attend meetings or interact with other members of that group.

Sociologists emphasize the ways in which inequality develops within group contexts and how outside statuses impact group dynamics. Four basic processes involved in the study of inequality are power, status, justice, and legitimacy. Power refers to the ability to obtain what we want despite resistance while status is our relative position in a group. Justice refers to perceptions of fairness when distributing things within a group while legitimacy is the sense that distributions are fair, regardless of how much people get in a group. These processes are typically studied under two major theories: social exchange theory and status characteristics theory, both utilizing social experiments.

Social exchange theory is based on the principle that people enter groups that provide them some benefit and leave groups when they perceive that they are not receiving a benefit from them. This theory is essential to explaining why people affiliate with other people in the first place but it also explains why people end relationships. There are four principles of social exchange theory (Molm and Cook, 1995). First, individuals in a group must have some form of dependence on each other – the theory does not apply to aggregates and acquaintance relationships. Second, within groups, individuals act in a way that will maximize personal benefits for themselves. The third principle, also known as the 'reciprocity principle', states that individuals will continue relationships as long as they receive about as much from a group as they put into it. If one or more members of a group believe that they are contributing more to the group good than they are receiving, they will seek alternative relationships outside of the group. Finally, the 'satiation principle' states that the value of what is exchanged may be reduced over time, thus explaining why people may end relationships even if they perceive them to be reciprocal.

Social exchanges may be direct, between two people, or indirect, exchanges among three or more individuals (Emerson, 1992; Homans, 1946; Molm and Cook, 1995). We may work for someone in order to receive pay but people may also help strangers with the understanding that they may require services from a stranger in the future. In this way, society can be seen as a network of direct and indirect exchanges. Status characteristics theory (sometimes called expectations states theory) employs elements of social exchange theory to understand how inequality develops in groups and how social statuses such as race and gender can impact exchange processes themselves. Robert Bales and other sociologists (Bales, 1965; Borgatta et al., 1954) analyzed interactions within small groups and determined that individuals who contribute most to group goals tend develop leadership status within those groups. Status appeared to be exchanged for service within the group.

Status characteristics theory focuses on the ways that outside characteristics influence status processes within groups (Blau, 1964; Berger et al., 1966, 1972; Lucas, 2003). Different status characteristics may influence the group, depending on the type of status. Specific status characteristics include those statuses that are important under specific conditions while diffuse status characteristics are those statuses that may influence the outcomes of all

exchange relationships. A mechanic may have high status among people who need their cars repaired but have no influence among a group of tennis players. Diffuse characteristics typically include individuals' sex and race that are believed to influence status dynamics in almost any group setting.

Sex status is one of the most widely studied areas of status characteristics theory. Research regularly shows that women receive less status than men in exchange relationships, even when they contribute most to the group (Lucas, 2003; Ridgeway and Diekema, 1989). Individuals develop different performance expectations of its members based on cultural characteristics, even if those characteristics have nothing to do with the group's task. Individuals with low- and high status develop a positive bias toward people who are believed to bring more resources into a group (Ridgeway et al., 1998). However, there are ways in which low-status members such as women can improve their chances of gaining status in groups by, for instance, legitimizing the position of women as leaders in a group setting (Lucas, 2003) or having female leaders emphasize how their work contributes to the good of the group (Ridgeway, 1982). Findings from the status characteristics approach help to elaborate the ways that larger social forces operate in groups and how group interactions help to maintain traditional stereotypes about people with different statuses in society.

Sociobiology

Sociobiology is the study of the relative impact of social and biological factors on individuals' thoughts, feelings, and behaviors. Sociologists also emphasize the ways that people construct the meaning of their biological world, giving more or less importance to certain traits or attributes (e.g., physical attraction) or/and how individuals come to define something as biologically inherited or not (e.g., sexuality).

Twin research is one major way that sociologists examine the relative impacts of genetic inheritance over social interaction in the production of social life. By comparing monozygotic (identical) twins who share 100% of their DNA with dizygotic (fraternal) twins, siblings, and nonrelated children, researchers can assess the degree to which nature impacts social psychological outcomes over social conditions because nonrelated family members share the same social conditions but not DNA, while identical twins share all of the same genetic traits. Using this technique, Nielson (2006) found shared genetic inheritance was more important for the transmission of verbal IQ and grade point average than shared family environments. However, social influences explained a significant amount of these outcomes and Horwitz et al. (2003) caution that twin studies can overstate genetic influences because identical twins both share a greater genetic as well as environmental background compared to both dizygotic and other siblings.

Gene–environment interactions represent another major way that sociologists and other social scientists assess the relative impacts of biology and social conditions on social psychological outcomes. Under this model, specific genes are thought to impact environmental outcomes and vice versa. Hence, while genes are thought to contribute to individuals' decision-making processes, environmental conditions interact with those genes and genes can also be impacted by social conditions. For instance, Guo et al. (2008) found that high levels of alcohol use among one's peers tend to bring about higher levels of genetic contribution to alcohol use. In other words, environmental conditions appeared to trigger the relationship between genetic disposition and alcohol use. In another study, genetic disposition explained variation in delinquency but these effects were mediated through levels of social control from family, schools, and peers (DeLisi et al., 2008).

Sociologists regularly test assumptions about the biological causes of social behaviors. Research generally shows differences in things like IQ vary among social groups and are thought to be inherited (Herrnstein and Murray, 1994). Girls, specifically, report lower math ability and are less likely to pursue careers in mathematics and the sciences. While some analysts believe that differences in math ability are genetically inherited, scholars in the group processes perspective have been able to influence both IQ scores and career aspirations simply by manipulating perceptions of abilities of subjects in experimental settings (see Correll, 2004; Lovaglia et al., 1998; Ridgeway et al., 2009; Rosenthal and Jacobson, 1968). While some emotional expressions such as fear and anger appear to have similar meaning across cultures, the conditions that arouse these and other emotions, their expressions and labels vary from place to place, suggesting both biological and social causes (Ekman, 2007; Kemper, 1987; Thoits, 1989). Sociological research also challenges the stereotypes of gender differences in emotional expression (Simon and Nath, 2004). Finally, while medical diagnoses may be associated with biological causes, their meanings and treatment change over time based on social influences such as the authority of medical professions, interest groups, and other groups like pharmaceutical companies who have a stake in the labeling and treatment of patients (Conrad, 2005).

Conclusions

Sociological social psychology emphasizes the ways that society shapes social psychological processes. From this perspective, society is a network of interaction among people (Blumer, 1969). Society is shaped by individuals' social interaction thereby creating a dialectic in which individuals are shaped by society. The symbolic interaction, social structure and personality, and group processes perspectives utilize this framework to examine the intersection of the individual and society. Even biological outcomes can be impacted by these social processes.

See also: Emotions, Sociology of; Exchange: Social; Interactionism, Symbolic; Macrosociology-Microsociology; Methodological Individualism in Sociology; Phenomenology in Sociology; Prejudice in Society: Psychological Perspectives; Social Identity in Sociology; Sociobiology and Sociology: A New Synthesis; Status and Role, Social Psychology of.

Bibliography

Adler, P., Adler, A., 1994. Social reproduction and the corporate other: the institutionalization of afterschool activities. The Sociological Quarterly 35, 309–328.

Anderson, L., 2008. Analytic autoethnography. Journal of Contemporary Ethnography 35, 373–395.

Aneshensel, C.S., Sucoff, C.A., 1996. The neighborhood context of adolescent mental health. Journal of Health and Social Behavior 37, 293–310.
Bales, R., 1965. The equilibrium problem in small groups. In: Hare, P., Borgotta, E.F., Bales, R.F. (Eds.), Small Groups: Studies in Social Interaction. Knopf, New York, pp. 424–456.
Bellah, R.N., Madsen, R., Sullivan, W.M., Swidler, A., Tipton, S.M., 1996. Habits of the Heart: Individualism and Commitment in American Life. Harper & Row, New York.
Berger, J., Cohen, B.P., Zelditch Jr., M., 1966. Status characteristics and expectation states. In: Berger, J., Zelditch Jr., M., Anderson, B. (Eds.), Sociological Theories in Progress, vol. 1. Houghton Mifflin, Boston, MA, pp. 29–46.
Berger, J., Cohen, B.P., Zelditch Jr., M., 1972. Status characteristics and social interaction. American Sociological Review 37, 241–255.
Berger, P.L., Luckmann, T., 1967. The Social Construction of Reality: A Treatise on the Sociology of Knowledge. Doubleday, Garden City, NY.
Blau, P.M., 1964. Exchange and Power in Social Life. John Wiley & Sons, New York.
Blumer, H., 1969. Symbolic Interactionism: Perspective and Method. University of California Press, Berkeley, Los Angeles, and London.
Borgatta, E.F., Bales, R.F., 1953. Interaction of individuals in reconstituted groups. Sociometry 14, 302–320.
Borgatta, E.F., Bales, R.F., Couch, A., 1954. Some findings relevant to the great man theory of leadership. American Sociological Review 19, 755–759.
Burke, P.J., Stets, J.E., 1999. Trust and commitment through self-verification. Social Psychology Quarterly 62, 347–366.
Conrad, P., 2005. The shifting engines of medicalization. Journal of Health and Social Behavior 46, 3–14.
Cooley, C.H., 1909. Social Organization: A Study of the Larger Social Mind. Charles Scribner's Sons, New York.
Cooley, C.H., 1922. Human Nature and the Social Order. Charles Scribner's Sons, New York.
Correll, S.J., 2004. Constraints into preference: gender, status, and emerging career aspirations. American Sociological Review 69, 93–113.
Deluca, S., Rosenbaum, J.E., 2001. Individual agency and the life course: do low-SES students get less long-term payoff for their school efforts? Sociological Focus 34, 357–376.
DeLisi, M., Beaver, K.M., Wright, J.P., Vaughn, M.G., 2008. The etiology of criminal onset: the enduring salience of nature and nature. Journal of Criminal Justice 36, 217–223.
Denzin, N.K., 2007. Memory: Lewis and Clark in Yellowstone, circa 2004. Symbolic Interaction 30, 297–321.
Dooley, D., Catalano, R., Rook, K., 1988. Personal and aggregate unemployment and psychological symptoms. The Journal of Social Issues 44, 107–123.
Ekman, P., 2007. Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life, second ed. Holt, New York.
Elder Jr., G.H., 1994. Time, human agency, and social change: perspectives on the life course. Social Psychology Quarterly 57, 4–15.
Elder Jr., G.H., 1999. Children of the Great Depression. Westview Press, Boulder, CO.
Emerson, R.M., 1992. Social exchange theory. In: Rosenberg, M., Turner, R.H. (Eds.), Social Psychology: Sociological Perspectives. Transaction Publishers, New Brunswick, NJ, pp. 30–65.
Fine, G.A., 1979. Small groups and culture creation: the idioculture of little league baseball teams. American Sociological Review 44, 733–745.
Fine, G.A., 2007. The construction of historical equivalence: weighing the red and brown scares. Symbolic Interaction 30, 27–40.
Fine, G.A., Beim, A., 2007. Introduction: interactionist approaches to collective memory. Symbolic Interaction 30, 1–5.
Gartner, S.S., 2008. Ties to the dead: connections to Iraq war and 9/11 casualties and disapproval of the president. American Sociological Review 73, 690–695.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, NY.
Guo, G., Roettger, M.E., Cai, T., 2008. The integration of genetic propensities into social-control models of delinquency and violence among male youths. American Sociological Review 73, 543–568.
Heise, D., 2007. Expressive order: confirming sentiments in social actions. Springer, New York.
Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press, New York.
Hogg, M.A., Ridgeway, C., 2003. Social identity: sociological and social psychological perspectives. Social Psychology Quarterly 66, 97–100.
Homans, G.C., 1946. The small warship. American Sociological Review 11, 294–300.
Horwitz, A.V., Videon, T.M., Schmitz, M.F., Davis, D., 2003. Rethinking twins and environments: possible social sources of assumed genetic influences in twin research. Journal of Health and Social Behavior 44, 1–129.
House, J.S., 1977. The three faces of social psychology. Sociometry 40, 161–177.
House, J.S., 1992. Social structure and personality. In: Rosenberg, M., Turner, R.H. (Eds.), Social Psychology: Sociological Perspectives. Transaction Publishers, New Brunswick, NJ, pp. 525–561.
Kemper, T.D., 1987. How many emotions are there? Wedding the social and autonomic components. American Journal of Sociology 93, 263–289.
Kessler, R.C., Neighbors, H.W., 1986. A new perspective on the relationships among race, social class, and psychological distress. Journal of Health and Social Behavior 27, 107–115.
Latkin, C.A., Curry, A.D., 2003. Stressful neighbourhoods and depression: a prospective study of the impact of neighbourhood disorder. Journal of Health and Social Behavior 44, 34–44.
Lovaglia, M.J., Houser, J.A., 1996. Emotional reactions and status in groups. American Sociological Review 61, 867–883.
Lovaglia, M., Lucas, J., Houser, J.A., Thye, S., Markovsky, B., 1998. Status processes and mental ability test scores. American Journal of Sociology 68, 464–480.
Lucas, J., 2003. Status processes and the institutionalization of women as leaders. American Sociological Review 68, 464–480.
Mead, G.H., 1934. Mind, Self, and Society from the Standpoint of a Social Behavioralist. Chicago University Press, Chicago, IL.
Molm, L.D., Cook, K.S., 1995. Social exchange and exchange networks. In: Cook, K.S., Fin, G.A., House, J.S. (Eds.), Sociological Perspectives in Social Psychology. Allyn & Bacon, Boston, MA, pp. 209–235.
Nielsen, F., 2006. Achievement and ascription in educational attainment: genetic and environmental influences on adolescent schooling. Social Forces 85, 193–216.
Pearlin, L., 1989. The sociological study of stress. Journal of Health and Social Behavior 30, 240–256.
Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.
Ridgeway, C.L., 1982. Status in groups: the importance of motivation. American Sociological Review 47, 76–88.
Ridgeway, C.L., Diekema, D., 1989. Dominance and collective hierarchy formation in male and female task groups. American Sociological Review 54, 79–93.
Ridgeway, C.L., Li, Y.E., Erickson, K.G., Backor, K., Tinkler, J.E., 2009. How easily does a social difference become a status distinction? Gender matters. American Sociological Review 74, 44–62.
Ridgeway, C.L., Kuipers, K.J., Boyle, E.H., Robinson, D.T., 1998. How do status beliefs develop? The role of resources and interactional experience. American Sociological Review 63, 331–350.
Robins, L.N., Locke, B.Z., Regier, D.A., 1991. An overview of psychiatric disorders in America. In: Robins, L.N., Regier, D.A. (Eds.), Psychiatric Disorders in America. Free Press, New York, pp. 328–366.
Rohall, D.E., Hamilton, V.L., Segal, D.R., Segal, M.W., 2001. Downsizing the Russian army: quality of life and mental health. Journal of Political and Military Sociology 29, 73–91.
Rohall, D.E., Milkie, M., Lucas, J., 2013. Social Psychology: Sociological Perspectives. Pearson, Boston, MA.
Rosenberg, M., Pearlin, L.I., 1978. Social class and self-esteem among children and adults. American Journal of Sociology 84, 53–78.
Rosenthal, R., Jacobson, L., 1968. Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development. Holt, Rinehart and Winston, New York.
Ross, C.E., 2000. Neighborhood disadvantage and adult depression. Journal of Health and Social Behavior 41, 177–187.
Ross, C.E., Mirowsky, J., 2008. Age and the balance of emotions. Social Science and Medicine 66, 2391–2400.
Schwalbe, M.L., Staples, C.L., 1991. Gender differences in sources of self-esteem. Social Psychology Quarterly 54, 158–168.
Simmel, G., 1950. The Sociology of Georg Simmel (K.H. Wolff, Trans.). Free Press, Glencoe, IL.
Simon, R.W., Nath, L.E., 2004. Gender and emotion in the United States: do men and women differ in self-reports of feelings and expressive behavior? American Journal of Sociology 109, 1137–1176.
Stryker, S., 2002. Symbolic Interactionism: A Social Structural Version. Blackburn, Caldwell, NJ.
Stryker, S., Vryan, K.D., 2003. The symbolic interactionist frame. In: DeLamater, J.D. (Ed.), Handbook of Social Psychology. Kluwer Academic/Plenum, New York, pp. 3–28.
Tajfel, H., 1982. The social psychology of interracial relations. Annual Review of Psychology 33, 1–39.
Thoits, P.A., 1989. The sociology of emotions. Annual Review of Sociology 15, 317–342.
Thoits, P.A., 1991. On merging identity theory and stress research. Social Psychology Quarterly 57, 101–112.
Turner, T.A., Shattuck, A., Hamby, S., Finkelhor, D., 2013. Community disorder, victimization exposure, and mental health in a national sample of youth. Journal of Health and Social Behavior 54, 258–275.

Social Psychology: Research Methods

Lia Figgou, Aristotle University of Thessaloniki, Thessaloniki, Greece
Vassilis Pavlopoulos, National and Kapodistrian University of Athens, Athens, Greece

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.M. Judd, volume 21, pp. 14405–14409, © 2001, Elsevier Ltd.

Abstract

This article presents some of the issues attached to the research methods used in Social Psychology. First it introduces the different methods used in social psychological research (quantitative and qualitative) and discusses the basic epistemological assumptions that lie on the basis of each of them. Then it presents methods of generating/collecting data and techniques of quantitative and qualitative data analysis. Finally, it briefly discusses issues related to validity and reliability of qualitative and quantitative research in Social Psychology and highlights the importance of addressing social psychological questions through the use and analysis of empirical data.

Over the years researchers identifying their research as social psychological have used a variety of methods to answer their research questions. These can be categorized as quantitative (experimental and nonexperimental) or qualitative (analytic techniques employed to analyze data generated through interviews and focus groups discussions, but also naturally occurring data). Although the division between qualitative and quantitative research methods coincides with debates about Social Psychology as a discipline, here we take the stance to present the different methods in a single article, highlighting their epistemological differences but assuming that both types of methods characterize the discipline. We will briefly present the methods, their epistemological foundations, and the questions they imply, and then we will present some techniques of generating and analyzing data. The article starts alphabetically from qualitative research methods and then moves to quantitative research.

Qualitative Research Methods in Social Psychology

The term 'qualitative research methods' is employed (1) to describe ways of generating and analyzing data which are not reducible to numbers (more often than not, texts and less frequently visual material) and (2) to refer to a broader framework within which certain epistemological and ontological assumptions prevail (Clarke and Braun, 2013). While in some cases qualitative research is used as a forerunner of quantitative research (analysis of interviews or focus group discussions often constitutes a first step toward the construction of a survey questionnaire) or as part of a mixed-methods research design, in other cases the choice of qualitative methods reflects a specific way of understanding social psychological phenomena and social psychological knowledge (*see* Social Constructionism).

The expansion of qualitative methods is, more often than not, related to the 'crisis' debates of the 1970s that included a critique on social psychological method(s) (*see* Critical Psychology). However, it is not until the late 1980s and 1990s that qualitative methods gained a foothold in Social Psychology, as a consequence of the development of post-modernist and post-structuralist perspectives and the so-called 'discursive turn' (*see* Social Psychological Theory, History of). Some basic principles of qualitative research – although different approaches do not share the same level of commitment to them – include the following: (1) a focus on meaning and interpretation, (2) a preference for inductive, theory-generating research, (3) sensitivity to the situated, context-specific nature of meaning, and (4) recognition that researchers' perspectives and subjectivities are intrinsically involved into the research process (something that in quantitative research is termed bias and it is treated as an unwelcome weakness). These qualitative methods are also used in other social science disciplines, especially in anthropology, cultural psychology, and cross-cultural psychology.

Methods of Collecting/Generating Qualitative Data

Interviews

Interviewing constitutes probably the most common and popular qualitative data collection technique. It normally involves a 'dialogue' with the researcher setting the agenda and asking questions and the interviewee being cast in the role of respondent. Nevertheless, interviews as a specific type of dialogue can be more or less structured. In structured interviews – rarely used in qualitative research – both the wording and the order of the questions are the same from one interview to another. In unstructured interviews, on the other hand, a free-flowing conversational style is adopted and respondents are encouraged to raise issues not originally included in the interview schedule. Biographical interviews which aim at the elicitation of research participants' personal stories with minimum researcher prompting constitute a paradigmatic example of unstructured interviews. Finally, in semistructured interviews, which are most commonly used in qualitative research, the researcher sets the agenda on the basis of their own interests and topics, but allows room for the participants' more spontaneous descriptions and narratives. Other distinctions are between one-to-one versus group interviewing, face-to-face versus

telephone interviewing or interviewing through the Internet (Madill and Gough, 2008).

Focus Groups

Focus groups constitute researcher/moderator-led group discussions designed to extract opinions about a topic. They have been originally developed in market research, but they gradually became a popular data generating method in academic research and especially in research projects that involve previously unexamined topics (Krueger and Casey, 2000). Focus groups provide a context which allows for the development of argumentation and counterargumentation and for the exploration of the interactional mechanisms involved in sense making. They are also considered a method appropriate to study groups whose voices are often marginalized within the larger society.

Interviews and focus group discussions are usually audio-recorded – researchers who intend to take into account nonverbal aspects of communication in their analyses tend to video record their interviews – and then transcribed. Transcription is a laborious task and demands prolonged practice. It is also an interpretative process that requires sensitivity on the part of the researcher to the nuances of oral speech and its differences to written language.

Naturally Occurring Data

This category includes a range of texts and interactions produced in the course of everyday life. The researcher is actually involved only in the sampling of the material. The virtues of using naturally occurring data are usually highlighted through their comparison to the artificiality of research interviews. According to critiques (Potter and Hepburn, 2005) an interview is carried out to serve the researcher's ends and agendas, which are external to the conversation itself and (potentially) irrelevant to the participants' interests. Naturally occurring data include archival documents (ranging from television programs and Internet materials to official/institutional archival data such as health records), naturally occurring conversations (therapy sessions, telephone calls recorded in the normal course by service providers) and – less frequently – visual material such as photographs or murals.

Observation

It has formed the basis for much qualitative research. In common with the category of naturally occurring data it is appropriate for the study of behavior that cannot be produced in an artificial environment for practical or even ethical reasons. Among its advantages is that it allows researchers to understand processes, to understand social life as involving interrelated series of events. Different types of observation are constructed on the basis of criteria such as the extent to which researchers intervene in the phenomenon of study or interact with research participants (Silverman, 1993). Structured observation refers to a situation where the researcher creates the context where a behavior can occur. Systematic observation involves a trained researcher who observes and codes the phenomena of study according to a prearranged set of criteria. Participant observation refers to a form of systematic observation whereby the observer interacts with the people being observed. Ethnography is a type of observational method used initially by cultural anthropologists and more recently by social psychologists to study and understand a group or culture. Ethnographic observation involves participation in a cultural-social context over a lengthy period of time.

Structured Methods of Data Collection

Sometimes qualitative research in Social Psychology may use more structured methods of data collection (also often used in quantitative research) such as open-ended questionnaires, Q-methodology and repertory grids, protocols (verbal or written records of observations or experience, obtained in response to a standardized question) designed, or vignettes (a scenario is provided for participants to consider and answer questions).

Methods of Qualitative Data Analysis

Qualitative Content Analysis

Content analysis is basically a quantitative method that involves establishing categories and counting the number of instances that these categories appear in a corpus of data. However, the method is also employed in qualitative research in which systematic classification procedures are used to categorize qualitative textual data into clusters of meaning (categories or themes). Qualitative content analysis is more interpretative in comparison with its quantitative counterpart and it is interested not only in the 'manifest' meaning of words or phrases but also in its 'latent' underlying meaning (Mayring, 2000). Coding systems consist of rules for assigning specified units into categories. Normally, categories are intended to be unidimensional, exhaustive, and mutually exclusive, but this is not always the case in qualitative content analysis (a piece of text may be relevant to more than one category). The implementation of content analysis has been drastically affected by the development of software packages, such as NVivo, Atlas ti, Dedoose.com, and the recently developed QCAmap. These tools can assist the researcher in handling and organizing large quantities of data but they cannot diminish the need for intellectual effort on the part of the analyst. Content analysis can be applied to a whole variety of data (including nonverbal data, such as pictures, drawings, gestures, etc.) and in relation to a variety of research questions. The fragmentation of texts and the decontextualization of data (as instances, pieces of text or answers are separated from their contexts) constitute potential disadvantages of the method.

Grounded Theory

A methodological approach was first developed by the sociologists Glaser and Strauss (1967). Glaser and Strauss criticized research derived from highly abstract theories that it fails to come up with explanations that are of relevance to those being studied. Hence, they developed a method aiming to generate a local, contextual theory, grounded on data. Therefore, the term grounded theory refers both to a method of inquiry, as well as to the theory produced by the implementation of method. The method includes an initial coding phase in which researchers work systematically through the data corpus generating codes (analytic labels) to describe both low-level concepts, as well as, more abstract categories. Throughout the process, they write memos elaborating their codes, identifying analytic gaps, and comparing analytic categories. Constant comparison and theoretical sampling constitute principal

analytic tasks of the method. Constant comparison involves continually comparing elements (analytic categories and theoretical presuppositions), while theoretical sampling involves the active sampling of new data that inform theoretical categories as the analysis proceeds. The emphasis of the original grounded theory method on *discovering* theory from data with its empiricist connotations was criticized by scholars who emphasized the constitutive nature of the analysis and developed more social constructionist revisions of the method (Charmaz, 2006).

Thematic Analysis

There is no agreement if it really constitutes a specific method in its own right. It is often rather unclear which its differences with qualitative content analysis are, and its use has been fairly inconsistent (see, however, Braun and Clarke, 2006; for a systematic introduction to the method). Thematic analysis also involves (usually inductive) coding of qualitative data into clusters of similar entities, or conceptual categories and the identification of consistent patterns and relationships between themes, so as to come up with a theoretical explanation of the phenomenon under study.

Interpretative Phenomenological Analysis

Its central focus is on the understanding of subjective experience and on the meaning attached by participants themselves to their lived experience. Unstructured or semistructured interview data are considered to be the ideal path to the participants' subjectivity. In terms of initial coding and the process of identification of emergent themes within the corpus of data Interpretative Phenomenological Analysis is also very close to some versions of thematic analysis and grounded theory (Smith et al., 2009).

Narrative Analysis

Narrative analysis refers to a cluster of analytic methods for interpreting texts or visual data that have a storied form. A common assumption of narrative methods is that people tell stories to help organize and make sense of their lives and their storied accounts are functional, and purposeful. Different approaches to narrative analysis are categorized on the basis of whether they focus on the narrative content or structure, with the thematic version interrogating what a story is about, while the structural version asks how a story is composed to achieve particular communicative aims. To this basic typology, according to Kohler Riessman (2008), one could also add the dialogic/performance narrative analysis, which focusses on the context and view of narratives as being multivoiced and coconstructed and the visual which links words and images in a coherent narrative.

Discursive Methods (Discourse Analysis/Discursive Psychology)

There are a variety of related – but also fundamentally distinct – methods grouped under the rubric 'discursive' or, most commonly, 'discourse analysis.' What is common between different methods is the recognition of the vital role of discourse in social life and an approach to language as social practice, instead of a pathway to inner cognitive entities (*see* Social Constructionism). Nevertheless there is little consensus between approaches on the ancestors and the epistemological basis of the analysis, on the analytic objectives, and on the conceptualization of the very notion of *discourse*. Potter and Wetherell (1987) – influenced by speech act theory, ethnomethodology, semiology, and post-structuralism – used the term 'discourse' to refer to virtually any language use and considered interpretative repertoires (recurrently used units of content, situated in certain – usually interview – contexts and oriented toward different interactional but also macro-social functions) as the units of analysis. On the other hand, scholars influenced by Foucault's work defined discourse as "a system of statements which constructs an object" (Parker, 1990: 191) and tended to fracture texts into discrete discourses which subjectify speakers and reproduce power relations.

The early 1990s witnessed also the emergence of discursive psychology (Edwards and Potter, 1992) which is informed by conversation analysis (CA) and prioritizes the use of naturalistic data. Its focus is on the role that descriptions of the world (including descriptions of psychological states) play in the management of speakers' accountability and in the formation of action. In the years to come, discursive psychologists became more deeply engaged with CA, working with conversational corpora from everyday and institutional settings, transcribed by the use of conventions developed by Gail Jefferson. Between the boundary lines among this strand of work, which affiliates with ethnomethodological traditions, and analytic perspectives which follow Foucaultian lines, other scholars proclaimed (Wetherell, 1998) an eclectic engagement with both CA and post-structuralism (and therefore engagement with both the interactional as well as the broader ideological functions of language use) as the most productive starting point for discursive methods in Social Psychology. This approach is known as critical discursive psychology.

Conversation Analysis

CA refers to a specific approach to the analysis of interaction that emerged in the 1960s in the work of Harvey Sacks (Sacks, 1995). CA is interested to understand social order by focusing analytically on the sequence of talk in interaction and on the ways participants organize mundane conversation. The recent engagement of discursive psychologists with CA blurs the boundaries between the two as analytic methods in Social Psychology.

Rhetorical Analysis

Interest in rhetoric in Social Psychology also arose as part of the discursive turn. Key text through which rhetoric was introduced in social psychological analyses constitutes Billig's (1987) '*Arguing and Thinking*.' Given his preference for scholarship, instead of methodology, Billig did not intend to introduce an analytic method in the narrow sense of the term. The text did not adhere to any specific methodology and neither included any methodological guidelines. Nevertheless, it served to deepen and enrich the understanding of how to approach analytically context and content in qualitative research, by advocating the need to consider the rhetorical relation between topics (as units of analysis). According to Billig (1987), in order to capture the meaning of any commonplace assumption, we should put it in

its argumentative context and examine it in relation to the commonplace that it aims to downgrade. Rhetorical analysis is also interested in relating meaning, or rather disputes over meaning, to the broader historical ideological context.

Validity and Reliability in Qualitative Research

There is considerable debate regarding the extent to which validity and reliability criteria – which have been defined in quantitative research and are imbued with its assumptions – are appropriate to evaluate the *quality* of qualitative research. According to some authors, qualitative researchers can and should incorporate criteria and techniques that deal with issues of reliability and validity and adjust them in their own epistemological framework. Silverman (1993), for example, maintained that triangulation techniques (of researchers or/and of data sources) can vitally contribute to the quality of research inquiry and can be compatible with the constructionist epistemology that underlies many qualitative research methods, if they are considered as highlighting the situated use of different accounts, instead of being treated as a means to prioritize one account vis-à-vis others.

Other researchers, however, prefer to replace validity and reliability with terms and criteria more compatible to the interpretivist background of qualitative research. Potter and Wetherell (1987) suggested *coherence* of the analytic framework (the potential of the analytic framework to give coherence in a body of data by accounting both for regularity and variability within it), *fruitfulness* (the extent to which the analytic scheme generates novel explanations), *new problems* (the identification of contradictions and exceptions from the explanatory scheme that necessitates new analytic questions and answers), and *participants' own orientation* (the way in which participants themselves see what analysts may consider as consistent, contradictory, etc.) as validation criteria of discourse analysis. Interest in understanding the phenomena from the participant's view as a validation criterion of the quality of the analysis has been developed also in other methods. Grounded theorists, for example, also maintained that the extent to which results may be recognizable and may be of relevance to those studied constitutes a vital criterion of research quality. Finally, a usually mentioned criterion of qualitative analysis's validation concerns the extent to which researchers' reflexivity is built into the analytic process.

Quantitative Research Methods in Social Psychology

Quantitative research is one that relies primarily on information of quantitative (i.e., numerical) nature. It conceptualizes reality in terms of variables, which measures and explores by applying statistical techniques in order to study relationships between them (Punch, 2014). Quantitative research is considered a 'top-down' approach in the sense that specific hypotheses deriving from a theory are tested through data analysis. Following the epistemological tradition of determinism, according to which events are accounted for by one or more causes, it assumes that behavior is explainable. Therefore, quantitative researchers try to make probabilistic predictions and generalizations by identifying cause-and-effect relationships. Supporting the ontological claim of objectivity, they strive to remain neutral and study the phenomena of interest 'from a distance,' which is expected to ensure validity of measurement and replicability of their findings (O'Dwyer and Bernauer, 2014).

As an empirical science, Social Psychology attempts to answer questions of quantitative nature about human behavior by testing hypotheses both in the laboratory and in the field. This corresponds to the distinction between experimental versus nonexperimental/correlational research designs. Although sometimes presented as essentially different or even incompatible, the difference between the two methods may be preferably understood in terms of the level of control exerted over the variables under study (Tajfel and Fraser, 1978).

Methods of Collecting/Generating Quantitative Data

Experimental

Although social psychologists employ a variety of methods to explore their research questions, the experimental paradigm has largely dominated the field, being characterized as the workhorse of Social Psychology (Wilson et al., 2010). An experiment involves measuring the effect of different conditions, intentionally manipulated by the researcher, on a behavioral outcome of interest. The former constitute the levels of an independent variable, while the latter is referred to as dependent variable. Not surprisingly, experiments are usually conducted in laboratory settings, which offer maximum control over the independent variable and minimal intrusion of third factors, thus allowing for testing hypotheses regarding causal effects.

The main advantage of an experimental design relies exactly on the ability to ensure all three preconditions for establishing causality, namely temporal precedence, covariation of the cause and effect, and exclusion of alternative plausible explanations (Trochim and Donnelly, 2007). This is not a simple task though: researchers must be careful to randomly assign individuals across the experimental conditions, to disentangle confounded variables, to avoid extreme conditions (that would lead to floor or ceiling effects), to exclude demand characteristics of experimental settings (that would lead to biased responses of participants), and to favor experimental versus mundane realism, i.e., 'true' psychological impact versus intrusive effect of everyday encounters (Hogg and Vaughan, 2010).

On the other hand, having enough control over the experimental conditions usually implies that the results cannot be generalized to everyday life, which means that there is a trade-off between internal and external validity. This can be dealt with by conducting field experiments, where experimental conditions are manipulated in real-life settings. However, some researchers warn that field experiments do not actually increase external validity as findings may or may not generalize to different real-life settings (Dipboye and Flanagan, 1979). Beyond the issue of validity, it may not be ethically acceptable or even possible for researchers to perform controlled experiments. In such occasions, nonexperimental methods are employed in order to collect quantitative data.

Nonexperimental

Quantitative nonexperimental methods refer to correlational studies, which focus on the naturally occurring associations among two or more factors. The use of advanced statistical software gave a boost in the number of variables, as well as in the complexity of the relationships, that can be examined simultaneously in a correlational study (see the next section). However, one should always bear in mind that any correlation-based technique, no matter how sophisticated, cannot determine causal effects. To compensate for reduced internal validity, the enhanced external validity associated with nonexperimental designs contributes to the development of social psychological theory by providing robustness (i.e., replicability of findings), representativeness (i.e., real-world processes), and social relevance (Brewer, 2000).

The most common method of data collection in a nonexperimental design is a survey study. Related developments – most notably, the widespread use of information technology – have broadened the range of choice of researchers to include techniques such as Internet research, diary methods, ambulatory assessment, and trace measures (see Reis and Gosling, 2010). These are briefly introduced below:

Survey

In quantitative studies, survey data are usually collected through the administration of self-report scales or questionnaires using the paper-and-pencil method. Surveys can obtain a large amount of information from a large number of participants in a short period of time. The psychometric properties (i.e., reliability, validity) of the measures, random sampling (in order to maximize representativeness of the general population), respecting the code of ethics (e.g., informed consent, anonymity, and confidentiality of responses), and – of course – implementing the appropriate statistical techniques are some key-issues to a successful survey. The numerous types of attitudes and values research are typical examples of survey studies.

Diary Methods

Diary studies adopt a within-subject approach by collecting repeated measures from the same number of participants at different time points. They are designed to capture life in real-world settings. Their use became more popular along with the development of multilevel statistical software, such as Hierarchical Linear Model (HLM) (see below). Topics of interest in diary studies are, among others, personality traits, emotions, social interaction, marital and family interaction, stress, and subjective well-being.

Ambulatory Assessment

This refers to the use of mechanical or electronic devices that record information about an individual's activity or state in natural, everyday settings. Some applications include ambulatory cardiovascular monitoring, electronic recording of the acoustic environment, activity monitoring, and location mapping (Reis and Gosling, 2010).

Internet Research

Initiated in the 1990s as a convenient means to increase the number of participants, complementary to the traditional paper-and-pencil method, the Internet soon became a new, dynamic, rapidly growing field of research. Of special attention are phenomena unique to the virtual environment, such as personal Web pages, forums, and the various forms of social networking (e.g., Facebook, Twitter, etc.). At the same time, a number of concerns have been raised against Internet studies, focusing on issues such as the characteristics of participants, the potentially fake identity of responders, or the reduced replicability and generalizability of findings compared to other methods. Although some of these issues have been addressed (e.g., Gosling et al., 2004), the physical distance between researcher and participant remains a major drawback and, at the same time, a distinctive advantage of Internet research.

Trace Measures

These studies are based on the fact that certain social behaviors, attitudes, cognitions, and emotions leave physical traces. Although Reis and Gosling (2010) accept that the environmental evidence of social psychological behaviors is largely untapped by researchers, they stress the potential value of trace measures, as the manifestations of an individual's thoughts, emotions, and actions go beyond physical environments.

Methods of Quantitative Data Analysis

Statistical methods are important tools for social psychologists in order to explore their research questions or examine specific hypotheses. These tasks typically correspond to the distinction between exploratory and confirmatory data analysis introduced by Tukey (1980), who wisely warned that confusing the two types of analysis can lead to systematic bias due to issues inherent in testing hypotheses suggested by the data. However, as Judd and Kenny (2010) point out, most data analysis in Social Psychology is essentially confirmatory, in the sense that researchers are guided by overtly or implicitly causal theories they seek to confirm. These models often include both direct and indirect effects; they may also expand to incorporate different levels of explanation, such as the intrapersonal, the interpersonal, the positional, and the ideological (Doise, 1986). In order to understand such complex relations in their data, researchers are supported by recent developments in advanced statistical techniques, such as structural equation modeling (SEM) and multilevel modeling (MLM), usually implemented by powerful software. In the following, these developments in quantitative data analysis will be briefly presented with respect to two issues of core importance in social psychological research, namely causal modeling and levels of analysis.

Causal Modeling

A causal model represents the relations between a given set of variables. It can take the form of a statistical equation and it is usually graphically depicted. SEM is in fact an extension of regression analysis for testing and estimating causal relations, which also allows for inclusion of latent constructs represented by a number of measured variables. The existing, constantly evolving SEM statistical software, such as AMOS, EQS, LISREL, and Mplus, providing powerful controls in what concerns model testing, modeling error, treatment of missing data, and testing invariance across multiple groups.

Causal models guide social psychological research by setting the ground for building and testing specific hypotheses.

Statistical handling of data collected within this framework can be quite complicated as it may include multiple variables and their relations are examined simultaneously in a number of combinations, which goes far beyond the traditional analysis of variance (ANOVA) and multiple regression techniques. Even in simple experimental designs as well as in nonexperimental correlational studies, research hypotheses may refer to indirect effects, which involve the inclusion of third variables explaining or modifying the relations between an independent factor and an outcome. Baron and Kenny's (1986) seminal work discussed in depth the conceptual, strategic, and statistical considerations of distinguishing between two types of third factors affecting a causal relation, i.e., the mediator and moderator variables.

Mediation

A simple mediation model suggests that the observed relation between an independent variable and an outcome can be explained by the effect of a third factor, known as mediator. According to Baron and Kenny (1986), three prerequisites are necessary in order to establish mediation: (1) the independent variable should be a significant predictor of the dependent variable; (2) similarly, the mediator should have a significant effect on the dependent variable; and (3) the mediator should be a significant predictor of the dependent variable while controlling for the effect of the independent variable. A relation is *fully mediated* when the direct path from the independent variable to the outcome becomes nonsignificant after accounting for the effect of the mediator. *Partial mediation* occurs when the mediator accounts for some, but not all, of the relation between the independent variable and the outcome. The amount of mediation is called the *indirect effect*.

A simple, common technique to identify mediation is Sobel's test (1982), which compares the difference in the relation between the independent variable and the outcome before and after inclusion of the mediator in the regression equation. More recently, Preacher and Hayes's (2008) nonparametric bootstrap method has become increasingly popular, which is recommended for small samples as it does not violate assumptions of normality. SEM programs or especially written statistical software macros can be applied in order to test for complex cases of mediation, such as multiple independent factors or outcomes, multiple mediators or latent variables used as mediators (see, for example, the Web pages of Andrew Hayes, http://www.afhayes.com/, and Kristopher Preacher, http://www.quantpsy.org/men.htm).

Beyond statistical complexity, it should be underlined that mediation is primarily a conceptual issue, i.e., the conclusions from a mediation analysis are valid only if the causal assumptions hold true (Judd and Kenny, 2010). Therefore, researchers should make sure that reverse causal effects (e.g., the mediator being caused by the outcome) or confounding variables (i.e., factors causing both the mediator and the dependent variable) are not present. Also, the mediator should be reliably measured in order to minimize bias.

Moderation

If mediation analysis attempts to explain 'how' or 'why' an effect occurs, moderation refers to the question of 'when' this effect holds, i.e., if the strength and/or direction of the relation between two variables is affected by the presence of a third variable, known as the moderator. Moderation suggests that the characteristics of the relation between an independent factor and an outcome vary at different levels of a moderator, while mediation suggests that the relation between an independent factor and an outcome is fully or partially explained by a mediator. Statistically speaking, moderation represents an interaction or a product in factorial ANOVA or in multiple regression, depending on whether the moderator is qualitatively or quantitatively measured, respectively (see Hayes and Matthes, 2009; for computational procedures of calculating interaction terms). From a conceptual point of view, hypotheses involving moderation address the question of invariance of causal relations between an independent variable and an outcome across different units, such as persons, situations, or cultures. In fact, any causal assumption implies some degree of invariance across persons. Judd and Kenny (2010) remind us of other sources of invariance including time points, situations, experimental stimuli (e.g., evaluative priming procedures), specific statistical techniques (e.g., meta-analysis, logistic regression), and cross-cultural comparisons. Invariance indicates the replicability of our findings and the limitations in generalizing our conclusions. In methodological terms, moderation analysis provides evidence of external validity as it examines to what extent a causal effect is considered universal.

Statistical handling of data in moderation analyses goes beyond simple identification of a significant partial effect of a product predictor in a linear model. Common problems include, among others, multicollinearity and power. Multicollinearity may yield coefficients with higher standard errors since the interaction term of two continuous variables (i.e., the moderator) inevitably correlates highly with the two main effects used to calculate it. Low statistical power indicates that replication of a significant moderation effect is difficult to achieve, especially with continuously measured independent variables, and calls for the use of large sample sizes.

In addition to the assumptions for applying specific statistical procedures, Judd and Kenny (2010) elaborate on some more perplexing issues of moderation testing. Not surprisingly, their discussion raises the problem of causality from a theoretical perspective once again. They show that the direction of causality between A and B (i.e., the decision on which variable will be treated as independent or outcome) matters a great deal in moderation testing, especially when the homogeneity of variance assumption is violated, so that A or B vary differently at high and low levels of the moderator. In such statistically ambiguous situations, a solid causal theory is necessary to indicate how moderation is assessed.

Mediated Moderation and Moderated Mediation

It is not rare that both mediator and moderator variables are integrated in a causal model. In these cases different hypotheses can be drawn, which are extensions of the simple mediation and moderation frameworks and they are defined accordingly. In *moderated mediation*, the mediating effect of a variable in the relation between an independent factor and an outcome varies across different levels of a moderator (Preacher et al., 2007). In *mediated moderation*, the effect of a variable moderating the relation between an independent factor and an outcome is accounted by the mediating process of another variable (Muller

et al., 2005). Therefore, moderated mediation and mediated moderation do not necessarily imply the existence of two different data sets, but rather two different causal models guiding research hypotheses, which can be based on the same data set.

Although mediation and moderation analyses are quite common in social psychological research literature, mediated moderation and moderated mediation hypotheses are not frequently examined, which would be particularly informative in extending the theoretical models used. This may be due to the statistical complexity of such questions, as computational routines to conduct these analyses are only recently being developed. Fairchild and MacKinnon (2009) provide illustrated examples of the equations required to simultaneously estimate mediation and moderation effects; they also underline the contribution of these techniques in applied settings, where the questions of 'how' and 'for whom' an intervention program is effective become crucial.

Modern SEM computer programs, with their user-friendly graphical interface and huge estimating capabilities, facilitate the conceptual and statistical handling of mediation and moderation analyses, provided that researchers are respectful with regard to consideration of assumptions (see Bentler and Chou, 1987; for an excellent overview). Testing of alternative models on the grounds of goodness-of-fit indices is particularly enlightening in terms of theory building. For example, longitudinal designs may benefit from examining four competing hypotheses referring to stability, causality, reversed causality, and reciprocal causality, respectively. Yet, it should be reminded that SEM is a confirmatory – as opposed to exploratory – technique, which means that it can be used to compare multiple theories that are specified a priori.

Levels of Analysis

Although many social psychological theories focus on individual processes, they are inherently placed at the interface between personal and group phenomena; therefore, hypotheses drawn on these theories may include multiple levels of analysis. The term *MLM* – which is a shortcut for *multilevel random coefficient modeling*, also known as *hierarchical linear modeling*, among others – is used to describe hierarchically structured data, i.e., observations at one level which are nested within observations at another level. For example, employees (L1) are nested within departments (L2) and organizations (L3); students (L1) are enrolled to different classes (L2) and schools (L3); in cross-cultural studies, individuals (L1) come from different countries/cultures (L2); in diary studies, emotions at different events or time intervals (L1) are nested within persons (L2). As Kreft and de Leeuw (1998) put it, once we realize hierarchies exist, we recognize them everywhere!

In statistical terms, the key question is whether observations are independent or not. In hierarchically structured data observations at L1 are not independent, which violates a fundamental assumption of traditional ordinary least squares (OLS) techniques, such as ANOVA and regression. Single-level analyses that ignore the hierarchical structure of data may provide misleading results because relations at different levels of analyses are in fact independent. Van de Vijver and Poortinga (2002) provide a taxonomy of multilevel fallacies occurring when different patterns of relations across levels are ignored. Additionally, in social psychological research inferences are made for a unit of analysis by studying random samples. MLM takes into account simultaneously the sampling error at different levels, which is not the case with traditional OLS. Therefore, MLM produces more accurate estimates than OLS because it considers the reliability of scores and the differences in sample sizes. These advantages are pronounced when hypotheses of interest concern within-unit relations, and when the data structure is irregular, for example, due to missing data (Nezlek, 2008).

MLM is appropriate when research questions involve more proximal (L1) and more distal (L2, L3…) factors influencing individual behavior, or simply when the data are hierarchically structured. Two issues of relevance need to be considered here: (1) the nature of the variables used; and (2) the nature of the relations between levels. Variables in a multilevel model may be intrinsic or derived (either aggregated or disaggregated; see Van de Vijver and Poortinga, 2002). The latter case raises concerns about equivalence, i.e., to what extent a construct measured in different groups has the same underlying structure. Structural equivalence is a precondition in order to perform group mean comparisons. It can be examined by calculating a congruence coefficient (e.g., Tucker's phi) through exploratory multilevel factor analysis procedures, or by applying multilevel confirmatory factor analysis with the use of SEM programs. Another question of interest is to what extent the same pattern of predictor–outcome relations holds across L2 units, e.g., across groups, cultures, or time points. This is typically referred to as a cross-level interaction. Conceptually similar to moderation, it is substantive for social psychologists who seek to explore the interplay between different levels of explanation (Doise, 1986) and can be examined through specialized statistical software (HLM).

Some Additional Comments

Error

Dealing with error in quantitative data analysis is almost as important as model testing. Although relevant assumptions (such as lack of measurement error and homogeneity of error variance) are embedded in OLS analyses, researchers sometimes tend to overlook them. Violation of these assumptions may lead to unfortunate conditions, such as measurement bias, attenuation of measures of association, reduced power in testing interactive effects, and underestimation of mediation effects, among others (Baron and Kenny, 1986). Therefore, obtaining reliable measures, identifying outliers, testing for nonnormality, and treating sources of error at different levels are necessary steps to be taken prior to any analysis involving specific hypothesis testing.

Variable Coding

In a regression equation, the parameter estimate of a predictor indicates the effect of the predictor on the criterion variable when all other predictors equal zero. Therefore, all simple effects are meaningless if zero is not a meaningful value for the predictors included in a regression model. This straightforward statement is not always fully understood by researchers, thus leading to a series of misinterpretations. When a categorical predictor is coded as 1 versus 2, then the slope of a continuous independent variable points to a nonexistent level of the

categorical predictor. An easy solution for this problem is to use a coding of 0 versus 1. Even in that case, the slope of a continuous independent variable is not a 'main effect,' as presented by many, but rather the simple effect of that independent variable when the categorical predictor equals zero.

In the same underlying rationale, when calculating the product for a regression interaction term it is advised that the two component variables are centered prior to computing their product, so that zero is an interpretable value for both predictors (Aiken and West, 1991). Of course the above is not necessary if zero is already a meaningful value, like for example, in Likert scales coded from −3 to +3, instead of 1–7. In multilevel models, variable centering can affect significance tests, but most dramatically it changes intercepts and error terms (Kreft and de Leeuw, 1998). HLM software provides two centering options (in addition to the uncentered solution), namely grand mean centering and group mean centering, depending on whether the slopes represent deviations from the overall mean or from each group's mean, respectively. With grand mean centering, estimates of slopes include between group differences in means of predictors, which are not included in group mean centering. At L2 (in two-level models), it is preferred to grand mean center, as this helps interpreting the intercept. At L1, group mean centering is closest to conducting within group regression analysis.

Missing Data

Missing data are quite underestimated by researchers. In any paper describing a quantitative statistical method there is a section devoted in how to treat missing values, often skipped by some readers. However, the presence of missing observations may have important implications, from restricting the generalizability of the findings to practically preventing an analysis from running at all. A typology describing missing data in terms of their underlying cause is owed to Rubin (1987): missing completely at random, missing at random, and missing not at random. The strategies for handling missing data include case (listwise) deletion, imputation by using a substitute value (such as the group mean or a multiple regression estimate), and multiple imputation (a method that estimates the sampling variance due to imputation). These strategies differ in terms of sophistication, which in turn attempts to compensate for their disadvantages. Therefore, they should be studied carefully before being applied in a given data set through modern powerful statistical software.

Judd and Kenny (2010) urge social psychologists to be more creative in how they deal with missing data, which they believe can lead to new insights about their research designs. Overall, perceiving statistical analyses as a challenging opportunity to discover rather than a routine of imposed restrictions will help researchers get the most out of their data and improve their theories.

Conclusion

In this article, we presented the methods used by social psychologists in order to set and address research questions, to generate and analyze empirical data, and to contribute to social psychological knowledge. We devoted space both to qualitative and quantitative research and we attempted to highlight not only the differences between them but also the diversity within each research tradition. Of course many supporters of each tradition would disagree about the extent to which a pluralistic stance toward method is preferable or even acceptable, although a *mixed-methods* research approach combining the quantitative and qualitative typologies is gaining ground in recent years (e.g., Bryman, 2006). In any case, it is hard to disagree that the diverse ways of generating and analyzing empirical data and the importance attributed to this process has given Social Psychology as a discipline a distinctive identity among the social sciences.

See also: Anthropology: Overview; Attitude Measurement; Critical Psychology; Cross-Cultural Psychology; Cultural Psychology; Implicit Social Cognition; Indigenous Psychology; Levels of Analysis in Social Psychology; Social Constructionism; Social Psychological Theory, History of; Social Psychology.

Bibliography

Aiken, L.S., West, S.G., 1991. Multiple Regression: Testing and Interpreting Interactions. Sage, Newbury Park, CA.

Baron, R.M., Kenny, D.A., 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Bentler, P.M., Chou, C.-P., 1987. Practical issues in structural modeling. Sociological Methods & Research 16, 78–117.

Bryman, A., 2006. Integrating quantitative and qualitative research: how is it done? Qualitative Research 6 (1), 97–113.

Billig, M., 1987. Arguing and Thinking: A Rhetorical Approach to Social Psychology. Cambridge University Press, Cambridge, UK.

Brewer, M.B., 2000. Research design and issues of validity. In: Reis, H.T., Judd, C. (Eds.), Handbook of Research Methods in Social Psychology. Cambridge University Press, New York, pp. 3–16.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3 (2), 77–101.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage, London.

Clarke, V., Braun, V., 2013. Successful Qualitative Research: A Practical Guide for Beginners. Sage, London.

Dipboye, R., Flanagan, M.F., 1979. Research settings in industrial and organizational psychology: are findings in the field more generalizable than the laboratory? American Psychologist 34 (2), 141–150.

Doise, W., 1986. Levels of Explanation in Social Psychology. Cambridge University Press, Cambridge, UK.

Edwards, D., Potter, J., 1992. Discursive Psychology. Sage, London.

Fairchild, A.J., MacKinnon, D.P., 2009. A general model for testing mediation and moderation effects. Preventive Science 10 (2), 87–99.

Glaser, B.G., Strauss, A., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Publishing Co, Chicago, IL.

Gosling, S.D., Vazire, S., Srivastava, S., John, O.P., 2004. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. American Psychologist 59, 93–104.

Hayes, A.F., Matthes, J., 2009. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. Behavior Research Methods 41, 924–936.

Hogg, M., Vaughan, G., 2010. Social Psychology, sixth ed. Prentice Hall, London.

Judd, C.M., Kenny, D.A., 2010. Data analysis in social psychology: recent and recurring issues. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, Hoboken, NJ, pp. 115–139.

Kohler Riessman, C., 2008. Narrative Methods for the Human Sciences. Sage, Thousand Oaks, CA.

Kreft, I.G.G., de Leeuw, J., 1998. Introducing Multilevel Modeling. Sage, Newbury Park, CA.

Krueger, R.A., Casey, M.A., 2000. Focus Groups: A Practical Guide for Applied Research, third ed. Sage, Thousand Oaks, CA.
Madill, A., Gough, B., 2008. Qualitative research and its place in psychological science. Psychological Methods 13 (3), 254–271.
Mayring, P., 2000. Qualitative content analysis. Forum: Qualitative Social Research 1 (2). Retrieved from: http://qualitative-research.net/fqs-texte/2-00/2-00mayring-e.htm.
Muller, D., Judd, C.M., Yzerbyt, V.Y., 2005. When moderation is mediated and mediation is moderated. Journal of Personality and Social Psychology 89, 852–863.
Nezlek, J., 2008. An introduction to multilevel modeling for social and personality psychology. Social and Personality Psychology Compass 2, 842–860.
O'Dwyer, L.M., Bernauer, J.A., 2014. Quantitative Research for the Qualitative Researcher. Sage, Tousand Oaks, CA.
Parker, I., 1990. Discourse: definitions and contradictions. Philosophical Psychology 3 (2), 189–203.
Potter, J., Hepburn, A., 2005. Qualitative interviews in psychology: problems and possibilities. Qualitative Research in Psychology 2 (4), 281–307.
Potter, J., Wetherell, M., 1987. Discourse and Social Psychology: Beyond Attitudes and Behaviour. Sage, London.
Preacher, K.J., Hayes, A.F., 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods 40, 879–891.
Preacher, K.J., Rucker, D.D., Hayes, A.F., 2007. Addressing moderated mediation hypotheses: theory, methods, and prescriptions. Multivariate Behavioral Research 42, 185–227.
Punch, K., 2014. Introduction to Social Research: Quantitative and Qualitative Approaches, third ed. Sage, London.
Reis, H.T., Gosling, S.D., 2010. Social psychological methods outside the laboratory. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, Hoboken, NJ, pp. 82–114.
Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
Sacks, H., 1995. Lectures on Conversation. Blackwell Publishing, Oxford, UK.
Silverman, D., 1993. Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction. Sage, London.
Smith, J.A., Flowers, P., Larkin, M., 2009. Interpretative Phenomenological Analysis: Theory, Method and Research. Sage, London.
Sobel, M.E., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology 13, 290–312.
Tajfel, H., Fraser, C., 1978. Social Psychology as social science. In: Tajfel, H., Fraser, C. (Eds.), Introducing Social Psychology. Penguin, Harmondsworth, UK, pp. 21–53.
Trochim, W., Donnelly, J.P., 2007. The Research Methods Knowledge Base, third ed. Thomson, Cincinnati, OH.
Tukey, J.W., 1980. We need both exploratory and confirmatory. American Statistician 34, 23–25.
Van de Vijver, F.J.R., Poortinga, Y.H., 2002. Structural equivalence in multilevel research. Journal of Cross-Cultural Psychology 33, 141–156.
Wetherell, M., 1998. Positioning and interpretative repertoires: conversation analysis and post-structuralism in dialogue. Discourse and Society 9 (3), 387–412.
Wilson, T.D., Aronson, E., Carlsmith, K., 2010. The art of laboratory experimentation. In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fifth ed. Wiley, Hoboken, NJ, pp. 51–81.

Social Question: Impact on Social Thought

Giovanna Procacci, University of Milan, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social question meant the rationality of poverty problems that took place with the rise of modern liberal societies. Within a social order based on the egalitarian premises emancipating individual rights and labor, poverty challenged both legal and economic emancipation, as well as their ability to organize a stable social order. Poverty therefore became a public concern about the dangers that inequalities produced by modernity, represented for the collective body. Searching for an autonomous definition of its scientific methods and hypotheses, at a distance from legal and economic sciences, early social science found its first field in the analysis of the social question. If social science was to analyze the organized group and its conditions, then poverty was both disruptive and crucial, because it could only be treated within a framework of obligations, more central to a collective, rather than a mere set of individual rights. Origins within poverty analysis deeply marked sociology foundation concepts making them organic, moral, and animated by a will to reform that gave the structure of a political program to scientific sociology. The social question therefore tied sociology to the development of an administrative state, reviving the 'intermediary groups' that liberalism had eliminated from the social body.

A modern social science not only originates within an intellectual project of widening rational knowledge, but also as a means to govern what Simmel (1908) called 'the material of social processes.' The making of what Comte (1848/1957) described as 'the preponderance of a social point of view' implied that there were problems that no established scientific approach could successfully address. Even more, the individualistic premises of modern legal and economic sciences themselves opened new problems from the vantagepoint of social organization. If the major trend of modernity has been toward the legal emancipation of the individual and the economic emancipation of labor from the burdens of traditional society, the emergence of social thought was to make evident that this could not suffice to organize a stable modern society. There was a need for models of organization and social integration, i.e., for a 'science of social order.' For such a science, the first object of analysis has been the 'social question,' that is, the mass pauperism growing alongside the rise of modern liberalism that social thought regarded as a danger for social stability. This became particularly clear in France where revolutions brought the problem of popular poverty to the forefront of the political arena.

The Experience of the Revolution

The idea that the French Revolution played an important role in the rise of sociology has been established by important sociologists such as Reinhardt Bendix and Robert Nisbet. The very term 'social science' became official during the French Revolution through authors like Condorcet (Baker, 1964; McDonald, 1993). Less common is acknowledging that its development was linked to the question of poverty and the new political importance it assumed. The Revolution was more than an intellectual advance in rationalism; it was a dramatic experience cutting bridges with most of what was known and abruptly revealing the depth of the social crisis. The new awareness of the crisis, more than the rationalist critique, was crucial to social thought.

Nisbet (1943), stressing the impact of the French Revolution over the making of sociological concepts, suggested that the experience of the crisis focused them on the 'social group,' as their own response to eighteenth century social theory centered on the opposition between the state and the individual. Social thought developed by establishing the group as an object of science, not only with an ethical implication, but also with a primary theoretical value in the study of man.

Why the group? Because the impact of the Revolution upon traditional social groups had driven the crisis until the destruction of the whole institutional frame, eroding the very bases of social belonging. The 'rational state,' as Nisbet put it, that the eighteenth century social theory had been regarded as the solution to the crisis of traditional bonds (religion, family, guilds, and associations) and the correlation to individuals' emancipation, no longer seemed able to nourish a real sense of social belonging. The group raised a new problem that could not be faced in economic or in juridical terms: to rebuild society required new foundations for the social bond. The experience of the crisis and the condition of insecurity due to the destabilization of order opened the way to a science that would fill the task of laying down the proper conditions for a new social order.

The Autonomy of 'Society'

The 'social question' was the most dramatic evidence for both the crisis and the need for a new social order. Poverty could no longer be regarded as a mere individual condition to be dealt with by charitable support and police repression, as in the Old Regime treatment of mendicancy and vagrancy. The political emancipation gained through the Revolution, as well as the economic emancipation guaranteed by a free labor, made the increase of poverty among the urban popular masses an issue of public concern (Simmel, 1908). How could the poor play a role in the new social order, be inscribed in the political arena of universal citizenship and in the liberalized market of productive labor? As Hannah Arendt (1963) stressed, the social

question broke the social unanimity around the revolutionary project – and from then onward, it has been the stumbling block of any revolution.

The modern idea of rights, making the poor just citizens the same as everyone else, and the modern economic imperative to include them into productive labor, did converge in giving a new visibility to the social question. In fact, poverty challenged both the achievement of individual autonomy as well as the generalization of free labor. It was therefore important to act against the social causes of poverty, and therefore to know them, what neither legal nor economic sciences could do. The importance of the poverty question grew together with the political awareness that a stable society demanded an end to the revolution. Both social and political problems contributed to a realization that social change could not be understood through individual actions, but rather as the result of relatively impersonal processes. Society appeared to have autonomous conditions to its own development and an autonomous ability to resist changing, both impossible to curb by law or through the market.

The social question has been the first object of application where the autonomy of social thought's premises, concepts, and methods was at stake. It was thus the arena for a gradual autonomization of social thought from economic and legal categories, founding the problem of order as specifically social (Procacci, 1993). In this way, it represented a crucial step toward the construction of the very idea of society as a theoretical concept and at the same time a practical field of investigation.

The Science of Social Order

The analysis of the social question had a crucial impact over the making of positive sociology. With Saint-Simon and Comte, the search for a new social order was firmly located within antisocial effects of industrial transformation: how was it possible to organize inequality in a way that would not prevent social stability? Their foundation concepts were marked by their source in the analysis of social inequality.

Firstly, they carried a vision of society as an organic entity, essentially concerned with the cohesion of all its parts – hence the importance of the metaphor of social organism. Most importantly, this located sociology on the side of the whole social body, the common interest, the collective being – apart from the opposition opened by liberalism between individual and collective points of view.

Secondly, all dangers for social order were interpreted in the perspective of social pathology, calling for some kind of therapeutics. Poverty was a pathological element within the social body – the task of social thought was to identify the means for restoring a normal society. Hence the importance of medical knowledge, with the normative character intrinsic in the idea of restoring a normal society. Methods of inquiry were also drawn out of medical practice: observation of symptoms, collection of cases, and statistical recurrences. This shaped concepts animated fundamentally by a will to reform. There was no longer a pure science; the program of positive sociology was a science able to found the criteria for acting in order to improve human conditions on positive knowledge.

Thirdly, the foundation concepts of positive sociology inherited the moral element that had characterized the philanthropic approach to poverty. Particularly with Comte, the problem of order became a question of moral consistence of society, impossible to solve through legal or economic reforms. His main theoretical questions run about the position of bourgeoisie and the working class in the new social order (Elias, 1978). For this reason, once the revolution exploded again in 1848 and the popular movement claimed political and economic reforms to solve the social question, he felt the need to engage into the public debate. His 'Positivist Society,' created a few days after the February insurrection, had the explicit task to oppose the positivist view of the social question to socialists. Insufficient labor did not depend upon capitalist economy, but upon an irrational organization that scientific development would overcome. The emphasis was rather on society's moral engagement to fight against poverty, a problem more difficult to solve: it required to put 'social sentiment' in the forefront, by promoting a sociability that the Comtian concept of 'duty' was to organize as the cement for new social bonds.

Democracy and Socialism

A different perspective is the one opened by Tocqueville's analysis of democracy. Elected in 1848 to the parliament by the first universal manhood suffrage, he was directly involved in writing the new Constitution which was to put an end to the revolution. As such, he politically intervened against the reforms claimed by the popular movement. Political organization of democracy above all demanded fighting its tendency to ally with a centralized power. The social question implied such a danger, strengthening the state that was expected to solve it; therefore nothing could be granted, except an exclusively moral engagement of society to reduce poverty. Distinguishing moral duties (*la morale*) from legal rights (*le droit*) was strategic in preventing the state from becoming either an economic agent providing jobs or a regulator reviving old protectionism. A political solution could only come from organising democracy: pushing people to act on collective issues and promoting public virtue instead of private freedom, were the only means for undermining the spectre of state despotism. The perspective of self-government led to rehabilitate associations as the arena where people would exert their public responsibility and resist the state.

The 1848 revolution was a crucial moment for social thought consolidation through its analysis of the social question. The work of Lorenz von Stein (1850/1921) provided important evidence and was a crucial source for Marx. Indeed it would be difficult not to mention the influence of social question over socialist social thought, as the revolution of 1848 exemplified. The experience of French revolutions was also central for Marx's historical concept of class, conceptualizing the fundamental inequality producing poverty in a capitalist economy. But class conflict was for Marx (Marx and Engels, 1848/1964) the key to development, and not a pathology to be cured; there could not be any other solution to it but the revolution. The strategical program of Marxian social thought deeply differed from the one of sociology: it was not a question of reanimating society through moral duties and intermediary

associations. Under this light, there is more continuity with Durkheim's theory of social solidarity.

In his study on socialism, Durkheim (1895/1971) stressed its affinity with sociology, given the central role played by organization and a set of social rules controlling the economic sphere. Since societies autonomously tend toward solidarity, the antagonism between capital and labor was a danger, and could only result from disorganization. The social question was a problem of socialization; even unions might work as a means of social integration. The concept of organic solidarity combined the characteristics of positive sociology and the attempt to enhance democratic strategies: it provided criteria for techniques reorganizing social bonds, such as the division of labor and the resurgence of intermediary groups.

In conclusion, the analysis of poverty contributed to strengthen a critique of economic and legal individualism toward a new science of social solidarity. The impact of the social question tied the fate of social science to the development of an administrative state, and engaged it into the political arena; thus promoting the re-establishment within the social body of those groups that liberalism had strategically eliminated.

See also: Civil Rights; Institutions; Liberalism: Political Doctrine and Impact on Social Science; Modernization, Sociological Theories of; Poverty, Sociology of; Revolutions, Sociology of; Socialism.

Bibliography

Arendt, H., 1963. On Revolution. Viking Press, New York.
Baker, K., 1964. The early history of the term 'social science'. Annals of Science 20, 211–225.
Comte, A., 1848/1957. The social aspect of positivism. In: Speller, R. (Ed.), A General View of Positivism (J.H. Bridges, Trans.). New York. Chap. 2.
Durkheim, E., 1895/1971. Le Socialisme. PUF, Paris.
Elias, N., 1978. What Is Sociology? (S. Mennell, G. Morrissey, Trans.) Hutchinson, London.
Foucault, M., 1972. L'Histoire de la Folie á l'Age Classique, Suivi de mon corps, ce papier, ce feu et la folie, l'absence d'oevre, second ed. Gallimard, Paris.
Marx, K., Engels, F., 1848/1964. The Communist Manifesto. Washington Square Press, New York.
McDonald, L., 1993. The Early Origins of the Social Sciences. McGill-Queen's University Press, Montreal.
Nisbet, R., 1943. The French revolution and the rise of sociology in France. American Journal of Sociology 49, 156–164.
Procacci, G., 1993. Gouverner la Misere. La Question Sociale en France. Seuil, Paris.
Simmel, G., 1908. Soziologie, Untersuchungen Uber die Formen der Vergesellschaftung. Germany, Leipzig.
von Stein, L., 1850/1921. In: Salomon, G. (Ed.), Geschichte der sozialen Bewegung in Frankreich. Germany, Munich.
Tocqueville, A., 1851/1978. Souvenirs [Recollections]. Gallimard, Paris.

Social Referencing during Infancy and Early Childhood across Cultures

Christine Fawcett, Uppsala University, Uppsala, Sweden
Ulf Liszkowski, University of Hamburg, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article reviews the state of knowledge on infants' social referencing. The developmental course of social referencing throughout infancy is described, as well as current debates on the motivation behind social referencing. Variation in social referencing depending on the type or information, information source, and characteristics of the infant are described. Finally, cultural variation and the presence of social referencing in nonhumans are discussed.

Classic Studies in Social Referencing

Social referencing describes the tendency to gather information from other individuals to regulate one's own behavior in ambiguous situations. Social referencing was first studied in adults (e.g., Asch, 1952; Schachter and Singer, 1962; Sherif, 1958), and until the 1980s, it was an open question whether infants would also show this behavior. Yet it seemed likely that they would, given that the literature on attachment, showing that infants see their caregivers as a secure base for exploration (e.g., Ainsworth et al., 1979) suggests that they could also see caregivers as a source of information about the world (Feinman and Lewis, 1983).

An additional motivation for studying social referencing in infants was to show that emotional expressions themselves were more than by-products of an emotional experience, but actually used as signals to others that could alter behavior (Sorce et al., 1985). If infants change their behavior based solely on the emotional expressions of others, then those expressions must have referential meaning. The emotion is about something in the world and thus can provide useful information for the infant.

In one of the earliest social referencing studies with infants, Feinman and Lewis (1983) examined 10-month-old infants' behavior toward a stranger when their mother gave positive or neutral signals about the person either directly to the infant or indirectly – when speaking to the stranger herself. They also measured infants' temperament and found that infants with an easy temperament who received direct positive information were more likely to approach the stranger. There were no other differences observed. However, it is possible that the situation was actually not ambiguous for most infants; thus the information from their caregiver did not lead them to adjust their behavior. That is, if infants know that a situation is safe or dangerous, they should not look for guidance from others.

The importance of uncertainty for social referencing was demonstrated by Sorce et al. (1985). Their study used a visual cliff with varying depths. That is, 12-month-old infants could cross from a normal-looking floor to a glass side with the floor seemingly close or far below them. When the floor was far, infants were unsure about whether crossing was safe and thus looked to their mother for information. If their mother's response was positive (i.e., a facial expression of happiness or interest), the majority of infants crossed the visual cliff. In contrast, if the mother's expression showed fear or anger, very few infants crossed. Their study also showed the importance of an ambiguous situation for social referencing to occur. If the floor of the visual cliff did not appear to drop off at all, infants crossed even when their mother's expression was of fear. In fact, infants rarely looked to her in this situation, and even those who did continued to cross, failing to change their behavior.

In another classic study, Hornik and Gunnar (1988) described the social referencing behavior of 12- and 18-month-old infants toward their caregiver in a more natural setting. This allowed them to examine the interplay of social referencing with other kinds of communication, such as affect sharing. Infants were presented with a rabbit in a cage and their behavior toward their mother was examined. The results showed that infants' referencing looks – those paired with a puzzled or wary facial expression – were as likely as affect-sharing looks when the rabbit was first introduced, and over time, referencing looks occurred much less often than sharing looks. This suggests that infants gather the information they need early on, but continue to want to share an experience with others over time. A second interesting finding from the study was that infants classified as initially wary or bold regarding the rabbit engaged in the same amount of referencing to the caregiver. This may show that social referencing can be used not only to seek valenced information (e.g., is it safe to approach?), but also instrumental information (e.g., what is this?) about an ambiguous situation.

This early work on social referencing in infancy served to establish several typical methodologies and to set the groundwork for future research that would more closely examine the process and development of social referencing.

The Developmental Course of Social Referencing

The first studies on social referencing focused on children around 1 year of age. How early do infants reference others for information when they are in ambiguous situations? A further question is whether referencing (i.e., seeking information) and behavioral regulation (i.e., acting on that information) develop at the same time.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23169-3

Earliest Evidence of Social Referencing

Using a cross-sectional design, Walden and Ogan (1988) explored the emergence of social referencing behaviors between 6 and 22 months of age. They found that at younger ages (6–9 months), infants were less discriminatory in their gaze to parents. When they looked at their parent, they did not always look to her face, suggesting that they may have only been checking that she was available, rather than looking for information. Younger infants also tended to look more to their parents when positive expressions were conveyed, whereas older infants (14–22 months) looked more to their parent's fearful expressions. Infants did not reliably regulate their behavior based on their parent's expression until between 10 and 13 months. At that age, infants were less likely to approach a toy after a fear reaction from the parent than a positive reaction. In addition, older infants were the only ones more likely to look to their parent before touching the toy. Interestingly, during a later free play trial after the parent was no longer giving emotional reactions, the older infants in their study (14–22 months) were actually more likely to touch the toy that had received a fear reaction, whereas the younger two groups were less likely to. In that situation, the toys were no longer moving or making noise and the experimenter invited infants to play with them. Thus, older infants may have inferred that the previously fearful toy was now safe. Together, this study suggests that social referencing behavior emerges around 1 year of age, but that it also continues to develop through the second year of life. Yet with a different experimental design, perhaps infants would be able to show social referencing at even earlier ages.

Until recently, the earliest evidence of behavioral regulation in response to other's emotional reactions was found at 8.5 months of age. Boccia and Campos (1989) found that infants were more wary of a stranger if their mother reacted with a stern rather than a cheerful greeting when the stranger entered the room. However, a recent study now shows that already by 5.5 months of age, infants are sensitive to emotion–object relations and regulate their own behavior based on those relations, but only for stimuli that had both facial and vocal emotional information (Vaillant-Molina and Bahrick, 2012). In a looking-time paradigm, infants demonstrated recognition of the specificity of another's emotional reactions by showing increased attention when the actor reversed her emotional reactions of happiness and fear to two different toys. Interestingly, they also preferred to touch the toy previously associated with a happy reaction. This suggests that they were also regulating their own behavior based on the actor's happy and fearful reactions, although they did not have to seek out the information as in typical social referencing paradigms. Another difference between this study and the standard ones is that it used habituation. Thus, the young infants were given as many repetitions of the emotional information as needed to fully process it. This extensive exposure could be one reason why infants were able to show social referencing ability at such a young age.

What Changes over Development?

Social referencing requires the ability to seek out information when needed, understand others' emotions, apply those emotions to objects or situations in the world, and to regulate behavior based on the information given. Infants' development in these areas can help give insight into when and how their social referencing develops and increases in sophistication.

Information-Seeking

Information-seeking may actually be one of the latest developing aspects of social referencing. Baldwin and Moses (1996) proposed that even though 12-month-old infants can make use of others' emotions to regulate their behavior, classic social referencing paradigms do not give clear evidence that infants actively seek out emotional information in an ambiguous situation. Instead, they may simply come across the information when they happen to look to an adult for other reasons, such as comfort-seeking or checking whether the caregiver is available. A likely prerequisite for information seeking is a concept of knowledge and ignorance. Infants must recognize that they lack information in an ambiguous situation and that another person may have useful information. It is an ongoing debate whether infants have stable concepts of knowledge and ignorance at 12 months of age (e.g., Liszkowski et al., 2008; Surian et al., 2007). Walden and Ogan (1988) found that only infants over 14 months of age are likely to check with their parent before approaching an ambiguous toy, even though younger infants' approach behavior is regulated by the caregiver's reaction.

A few studies have examined information-seeking, though not necessarily in a social-referencing context. Vaish et al. (2011) used an object-labeling paradigm and showed that 13- and 18-month-old infants were more likely to look to the person who labeled an object when it was necessary to disambiguate her label. That is, when two objects were available, infants checked her gaze to determine which one she was attending to, whereas when only one object was available, they were less likely to do so. Infants may also seek information when pointing at things (Baldwin and Moses, 1996; Liszkowski, 2005). Begus and Southgate (2012) looked at pointing behavior to determine whether infants seek information appropriately from others. They found that 16-month-olds were more likely to point to novel objects when interacting with a knowledgeable experimenter than an ignorant one, suggesting that they knew the knowledgeable person would be able to provide them with more information. While neither of these studies directly tested information-seeking in a social referencing paradigm, they do suggest that at least by early in the second year of life, infants can seek out information from others in a sophisticated way.

Emotion Recognition

Young infants are not yet able to approach and avoid situations in the same way as older infants can; however, sensitive measures of infants' attention can indicate some understanding of emotional reactions to objects. Hoehl et al. (2008) looked at whether very young infants' brain responses indicated that they were already sensitive to when another person reacted negatively to an object. They showed that infants' neural responses to objects differed if those objects had previously been shown in the context of a person gazing at the object with a negative emotional expression. This suggests that already at 3 months of age, infants can associate others' negative emotional reactions with objects and they

subsequently pay more attention to those objects, even when the emotional information is no longer visible.

The Hoehl et al. (2008) study only compared negative emotions to neutral expressions. Yet we also know that between 3 and 4 months of age, infants begin to discriminate between multimodal emotional expressions of happiness, anger, and sadness, as indicated by preferentially gazing at novel facial expressions in a habituation paradigm (Flom and Bahrick, 2007). By 7 months of age, infants also show increased attention for fearful facial expressions in that they have difficulty disengaging from a face with a fear expression (Peltola et al., 2008, 2009).

Studies on emotion differentiation cannot speak to whether infants recognize the meaning behind the expression they observe or whether infants' own emotional experience is affected in any way. Yet they show that infants have the first abilities necessary to begin learning more about the relevance of emotions in their world.

Referential Understanding

In addition to understanding what emotion another is expressing, social referencing requires that infants recognize the referent of the emotional reaction. Moses et al. (2001) showed that both 18- and 12-month-old infants appreciated the relevance of an emotional message when regulating their behavior toward an ambiguous toy. That is, an emotional reaction was delivered either by someone attending or not attending to the ambiguous toy. Infants were selective such that they differed in their approach to the toy based on positive and negative emotional reactions only for the toy to which the reaction was directed. For example, if the infant was examining one toy while the experimenter emoted toward a different toy, infants appropriately paired the reaction with the toy in the experimenter's view.

Repacholi (1998) also showed that 14- and 18-month-old infants have referential understanding of others' emotional reactions. Infants observed an adult gaze into each of two boxes and react positively to one and negatively to the other. Later, infants were more likely to explore the contents of the box that had been reacted to positively, even though they were equally likely to touch both boxes themselves. This suggests that infants accurately applied the emotional reactions of the adult to the contents of the box – even though they were not visible to the infant at the time – and not the box itself. Thus, early in the second year of life, infants seem to be able to accurately apply emotional messages to objects and situations.

Locomotion

A proposed factor influencing changes in social referencing around 8–9 months of age is the onset of crawling. Since crawling allows infants to be more independent from their caregivers, the need for social referencing and the distal communication it provides likely increases sharply once infants begin to crawl. Infants who have begun to crawl pay more attention to distant objects and check back with their caregiver more often than infants of the same age who are not yet crawling (Campos et al., 1992). Crawling infants also receive more prohibitions from their caregivers, particularly when the infant is far from the caregiver (Zumbahlen and Crawley, 1996). Together these findings from interview studies suggest that crawling may be a catalyst for social referencing behavior. In addition, infants with locomotor experience are more able to follow the referential gestures (gaze and pointing) of others, another important skill for social referencing (Campos et al., 2000).

There is also some evidence that the transition from crawling to walking could be influential in the development of social referencing. One study suggested that when infants begin to walk, they actually begin to socially reference less and engage in more social bidding than younger, crawling infants (Clearfield et al., 2008). However, it is possible that the novel toys used in that particular study may not have been sufficiently ambiguous to elicit social referencing behavior in the older, walking infants.

Can Mood Explain Social Referencing?

Given that evidence suggests that infants already understand the referential nature of emotion when they engage in social referencing, it is unlikely that mood can account for behavioral differences seen in social referencing situations. In addition, many studies have specifically ruled out mood as a confounding variable. For example, Hornik et al. (1987) found that caregivers' emotional reactions affected infants' behavior toward the particular toy that was the target of their reaction, but not to infants play behavior in general. Similarly, Feinman and Lewis (1983) accounted for mood by comparing infants' behavior toward a stranger (who had been reacted to either positively or negatively) to behavior toward their mother. Even when behavior toward their mother was subtracted, there was a difference in approach to the stranger based on the mother's expression, showing that mood and general willingness to approach another did not explain their findings. Walden and Ogan (1988) examined infants' own emotional expression during a social referencing task and again found no evidence that mood could be a confounding factor.

Finally, Stenberg and Hagekull (1997) directly tested the possibility of mood modification by having caregivers deliver either specific (directed to a particular object) or general positive or negative emotional messages. They found that while infants' behavior toward the toy was affected by the specific emotional messages, with more play following the positive than negative reaction, infants' overall activity was affected by the general emotional message. The findings suggest that when emotional information is specifically directed to an object, infants' own mood is not a factor in their behavior. In contrast, when general positive or negative emotional information is presented, infants do not have a clear target for that expression and react with overall differences in their exploratory behavior.

Whom Do Infants Prefer to Reference?

Do infants preferentially reference certain people over others? Research in this area has focused on infants' tendency to reference their caregivers compared to a familiar experimenter. The debate is rooted in the question of whether infants' looks

in ambiguous situations are information-seeking behaviors related to social referencing, or rather comfort-seeking behaviors related to attachment. In one such study, Zarbatany and Lamb (1985) presented infants with an unusual toy and either their mother or an experimenter was in the room posing fear or smiling expressions. Infants were found to reference both their mother and the experimenter, but their behavior toward the toy was only influenced by the reactions of the mother. When the experimenter was posing either expression, infants were unlikely to approach the toy. It was only when the mother showed a positive expression that they did so. A later study using a similar paradigm, but with both adults remaining in the room for the duration of the study, gave contrasting results. Infants both referenced the experimenter and regulated their behavior based on her reactions, with fewer and slower approaches to the toy following fear than smiling expressions (Klinnert et al., 1986). Thus, it is possible that infants in the Zarbatany and Lamb study were too uncomfortable in the room without their caregiver present that they were not able to approach the toy, even when the experimenter's expression was positive.

More recent studies have continued to find that infants tend to engage in more social referencing toward an experimenter than their caregiver. They do so not only when their parent is inattentive (Stenberg, 2003), but also when both parent and experimenter are providing information. Specifically, Walden and Kim (2005) presented 18- and 24-month-old infants with ambiguous toys and both caregiver and experimenter provided emotional reactions about the toy when infants looked toward them. They found that infants at both ages looked more toward the experimenter than their mother and that this difference was even greater for the 24-month-olds. Importantly, in a baseline looking measure, infants looked equally to their mother and the experimenter, ruling out overall interest in the experimenter as an explanation for the difference in referencing.

Stenberg (2009) found that infants' preference for referencing experimenters holds even when the caregiver is the one introducing the ambiguous toy and providing the emotional reaction about it. That study also looked specifically at 12-month-old infants' behavioral regulation following positive emotional reactions delivered by either the caregiver or experimenter. The findings showed that infants played longer with the toy following positive messages from the experimenter than the mother. However, since the study only included positive emotional reactions, it's not possible to say whether the pattern of findings would hold if the experimenter and caregiver were giving negative, rather than positive, messages.

Together, the findings suggest that infants may be relying on an inference of expertise when selecting whom to reference. That is, they look to the experimenter and trust her information more because they expect that the experimenter will know more about the ambiguous object or situation.

The Effect of Information Type on Social Referencing

What kind of information do infants use when socially referencing others? The modality of information – visual versus vocal, as well as the valence of the information being delivered could have differential effects on later behavioral regulation.

Modality

Multimodal emotional expressions – displayed with both facial and vocal information – are those earliest discriminated by infants. Flom and Bahrick (2007) showed that by 4 months of age, infants could discriminate multimodal expressions of happiness, sadness, and anger, and that they could only do so when temporal synchrony between the face and voice were intact. By 5 months, infants could discriminate the expressions with only vocal information and by 7 months they could do so with only visual information available. The same developmental trajectory for emotional expressions seems to be seen in social referencing. Twelve-month-old infants' crossing of a visual cliff was affected by multimodal and vocal emotional expressions, but not expressions of the face alone (Vaish and Striano, 2004).

In a more naturalistic study of the modality for social referencing messages, caregivers were asked to improvise their own emotional messages. Results showed that they tended to produce more gestures with messages of fear and more vocalizations with positive messages (Kim et al., 2010). Further, these differences were also what infants responded best to. They modified their behavior more based on multimodal and visual information in the case of fear and based on the amount of vocal information in the case of a positive message.

Even when the modality for an emotional message is video, infants regulate their behavior based on it. Specifically, Mumme and Fernald (2003) found that 12-month-old infants, but not 10-month-old infants, were less likely to touch a toy when they had seen someone on video react negatively toward it, compared to a toy that was reacted to positively or neutrally.

Valence

Are both positive and negative types of information valuable during social referencing? Vaish et al. (2008) propose that there is a negativity bias in social referencing, such that infants' behavior is more affected by negative than positive emotional reactions. When a neutral baseline is used as a comparison, most social referencing studies find little difference between infants' behavioral responses to positive versus neutral reactions from others, yet a significant difference between neutral and negative remains (e.g., Hornik et al., 1987; Mumme et al., 1996). This is important because it shows that social referencing may function primarily in one direction – encouraging infants to avoid objects and situations, but not to explore them.

Looking at valence from the caregiver's perspective, mothers naturally gave more intense positive reactions than negative ones when asked to improvise their own messages to infants (Kim et al., 2010). It could be that caregivers already know that a positive emotional message needs to be higher in intensity to influence their infant's behavior compared to a negative message.

Individual Differences in Social Referencing

Temperament

Findings on the relation between social referencing behavior and temperament have been somewhat mixed. Temperament refers to stable characteristics including emotional valence, fearfulness, inhibition, and ease of being comforted. It could be that infants who are more temperamentally fearful or less easily comforted would be more likely to reference others when encountering ambiguous situations or more likely to act on the information they receive during social referencing.

In one of the earliest studies on infant social referencing, Feinman and Lewis (1983) found that infants with easy, but not difficult temperaments increased proximity and offered more toys to a stranger following positive messages about the stranger from their mother. In contrast, Bradshaw et al. (1987) later found no relation between behavioral regulation during social referencing and positive/negative affect temperament, which corresponds to the easy/difficult dimension used by Feinman and Lewis. They did, however, find that infants with a more positive temperament made more social referencing looks to their caregiver.

Blackford and Walden (1998) investigated the role of temperamental fear on social referencing behavior. They found that low-fear 11- to 13-month-old infants were more likely to regulate their behavior based on positive or fearful messages from their caregiver regarding a novel toy than fearful infants who tended to avoid the toy altogether. Temperament did not have an effect on how often infants looked to their caregivers. Thus, the seeking of information remained the same, but the use of information differed for infants with different temperaments.

It is possible that the effect of temperament, at least fearful temperament, comes down to how ambiguous the experimental situation is for the infant. If the infant is temperamentally fearful, then he or she is already predisposed to avoid to the toy, despite a positive message from an adult. This is an important confound to keep in mind, since ambiguity is important for an accurate measure of social referencing.

Attachment Style

Social referencing and attachment are in some sense very closely intertwined. There is still some debate as to whether social referencing looks are actually comfort-seeking behavior, related to attachment, rather than information-seeking behavior (see Section Whom Do Infants Prefer to Reference? earlier in the article). However, even if social referencing is based on information-seeking, infants could differ in how likely they are to seek information from their caregiver depending on the quality of their attachment relationship. For example, in a study relating attachment security to referencing behavior, insecure-avoidant infants referenced their parent the least and insecure-resistant infants referenced the most, with securely attached infants in between (Dickstein et al., 1984). This is in line with the definitions of insecure attachment styles, in which avoidant infants tend to ignore their attachment figures, even in ambiguous situations, and resistant infants continue to seek comfort, but fail to actually be comforted by the attachment figure (Ainsworth et al., 1979).

Atypical Development

Children with autism spectrum disorder (ASD) tend to have significant deficits in social cognition, including social orienting, joint attention, and attention to others' distress (Dawson et al., 2004). Thus, it seems likely children with ASD would show impairments in social referencing. Sigman et al. (1992) found that 3- to 4-year-old children with ASD were less likely to respond to adults' emotional reactions than control participants, yet those who did respond regulated their behavior toward the toy accordingly. Their study could not examine referential looking, however, since the adult expressed an emotional reaction regardless of whether the child looked at her.

Since ASD is not typically diagnosed until 3 years of age, one recent study examined the social referencing abilities of infant siblings of children with ASD, who are themselves at high risk for developing the disorder (Cornew et al., 2012). For referential looking, infants who were later diagnosed as having autism were slower, and overall less likely, to spontaneously look to an adult when an ambigious toy was present, compared to typically developing infants. Interestingly, behavioral regulation based on others' emotional signals was found to be impaired in infant siblings of autistic children, regardless of whether they were later diagnosed with autism or not. Only typically developing control participants showed appropriate behavioral regulation in the study. The authors suggest that there may be a more subtle impairment present in the high-risk infants who do not develop ASD that accounts for their decreased social referencing.

Cultural Differences in Social Referencing

To our knowledge, no studies have specifically examined cross-cultural differences in social referencing. Even though cultures differ in their forms of communication, parenting practices, and social norms, some aspects of communication and emotion seem to be largely universal. For example, the expression and recognition of emotions (see Elfenbein and Ambady, 2002, for a review) and infants' reference to objects by pointing (Liszkowski et al., 2012) seem to be universal. In addition, an extensive study of early social cognitive skills in three cultures (Canada, India, and Peru) showed that gaze following and joint attention were not significantly different, despite widely varying parenting practices and social norms (Callaghan et al., 2011). Cultural differences have been reported in the frequency of mutual gaze in dyadic face-to-face interactions of parents and their young infants, with Nso parents engaging in less direct eye contact than German parents (Kärtner et al., 2010; Keller et al., 2011). However, it is unknown whether there is a relation between early mutual gaze and the later emergence of social referencing.

With little variation in the basic skills of emotion recognition, gaze following, and joint attention, it seems likely that social referencing would also be present cross-culturally. The primary difference in social referencing across cultures may rather be in the content of what children learn, rather than the use of referencing itself. That is, the mechanism of social

referencing may be universal, but the information gained through social referencing regarding what should be approached or avoided could depend on what culture the infant is a part of. A second possible difference could be in the individuals targeted during social referencing. For example, if children in some cultures spend more time with siblings than their parents (e.g., Harkness and Super, 1995), older children could be a more common source for social referencing information.

Social Referencing in Nonhumans

Social referencing does not seem to be an ability that is specific to human infants. Itakura (1995) observed mother–infant pairs of chimpanzees and found that infants were more likely to seek contact with their mother when novel objects were present, suggesting that they could be engaging in social referencing. In a more controlled experiment, similar to those used with infants, Russell et al. (1997) demonstrated that young chimpanzees made referential looks to a familiar human caregiver and modified their behavior toward novel toys based on the caregiver's reactions of happiness or fear.

Macaque monkeys were also found to develop social referencing behavior. Older (5- to 12-months), but not younger infant macaques made more looks to their mothers in the presence of a rubber snake than in a control condition with no snake present (Roberts et al., 2008). Yet when clear referential looks – looking to the mother directly before or after looking to the snake – were measured, only 13% of the macaques were found to engage in looking from the snake to their mother. Also, since the mother's emotional reaction could not be manipulated, it is not known whether the infant macaques also regulate their behavior based on her reaction. Yet this study does give some evidence that macaques look to their mothers when an ambiguous object is present, whether for information or comfort.

Domestic dogs were found to socially reference their owners regarding an ambiguous object (Merola et al., 2012a). All dogs in their study looked to their owner after looking at the ambiguous object, demonstrating robust referential looking. There was also some evidence for behavioral regulation in that dogs that received a negative message from their owners were more likely to stop moving around the room than those who received a positive message. Additionally, dogs were found to approach the object more if their owner also approached it than if their owner avoided it. In a later study, Merola et al. (2012b) found that dogs looked not only to their owner, but also to a stranger for social referencing. There was also clearer evidence for behavioral regulation in dogs, at least when the owner was providing the emotional message. Dogs that received a positive message were more likely to approach the object than those who received a negative message.

Together these studies show that several nonhuman species can engage in some aspects of social referencing. This is not altogether surprising given that primates and domestic dogs also show abilities related to social referencing, such as gaze following (e.g., Kaminski et al., 2009; Call et al., 1998). It remains to be seen whether other nonhuman species with sophisticated social cognitive skills, such as dolphins (Pack and Herman, 2006), also engage in social referencing.

Conclusions and Future Directions

Social referencing is an important part of how infants and young children learn about their world from social figures. From around the end of the first year, infants begin showing social referencing behavior – looking to others for guidance in ambiguous situations and regulating their behavior based on others' reactions. While social referencing is likely to be a universal skill, there are still individual, and possibly cultural variations. More work looking at these variations can help to give insight into the processes involved in developing social referencing and how it can be encouraged in children with autism, for example, who have impairments in this ability.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Autism Spectrum Disorders; Early Emotional Development and Cultural Variability; Emotion in Cognition; Emotion: History of the Concept; Emotional Development, Effects of Parenting and Family Structure on; Emotional Regulation; Infancy and Childhood: Emotional Development; Infancy and Human Development; Neonatal Temperament: From Phenomenology to Mechanism; Temperament and Human Development; Temperament.

Bibliography

Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S., 1979. Patterns of Attachment: a Psychological Study of the Strange Situation. Lawrence Erlbaum, Hillsdale, NJ.

Asch, R., 1952. Social Psychology. Prentice-Hall, Englewood Cliffs, NJ.

Baldwin, D.A., Moses, L.J., 1996. The ontogeny of social information gathering. Child Development 67 (5), 1915–1939.

Begus, K., Southgate, V., 2012. Infant pointing serves an interrogative function. Developmental Science 15 (5), 611–617.

Blackford, J.U., Walden, T.A., 1998. Individual differences in social referencing. Infant Behavior and Development 21 (1), 89–102.

Boccia, M., Campos, J.J., 1989. Maternal emotional signals, social referencing, and infants' reactions to strangers. New Directions for Child Development 44, 25–49.

Bradshaw, D.L., Goldsmith, H.H., Campos, J.J., 1987. Attachment, temperament, and social referencing: Interrelationships among three domains of infant affective behavior. Infant Behavior and Development 10 (2), 223–231. http://dx.doi.org/10.1016/0163-6383(87)90036-1.

Call, J., Hare, B., Tomasello, M., 1998. Chimpanzee gaze following in an object-choice task. Animal Cognition 1 (2), 89–99. http://dx.doi.org/10.1007/s100710050013.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., Tomasello, M., 2011. Early social cognition in three cultural contexts. Monographs of the Society for Research in Child Development 2, 76.

Campos, J.J., Anderson, D.I., Barbu-Roth, M.A., Hubbard, E.M., Hertenstein, M.J., Witherington, D., 2000. Travel broadens the mind. Infancy 1 (2), 149–219.

Campos, J.J., Kermoian, R., Zumbahlen, M.R., 1992. Socioemotional transformations in the family system following infant crawling onset. New Directions for Child and Adolescent Development 1992 (55), 25–40.

Clearfield, M.W., Osborne, C.N., Mullen, M., 2008. Learning by looking: infants' social looking behavior across the transition from crawling to walking. Journal of Experimental Child Psychology 100 (4), 297–307. http://dx.doi.org/10.1016/j.jecp.2008.03.005.

Cornew, L., Dobkins, K.R., Akshoomoff, N., McCleery, J.P., Carver, L.J., 2012. Atypical social referencing in infant siblings of children with autism Spectrum

disorders. Journal of Autism and Developmental Disorders. http://dx.doi.org/10.1007/s10803-012-1518-8.

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., Liaw, J., 2004. Early social attention impairments in autism: social orienting, joint attention, and attention to distress. Developmental Psychology 40 (2), 271.

Dickstein, S., Thompson, R.A., Estes, D., Malkin, C., Lamb, M.E., 1984. Social referencing and the security of attachment. Infant Behavior and Development 7 (4), 507–516.

Elfenbein, H.A., Ambady, N., 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128 (2), 203.

Feinman, S., Lewis, M., 1983. Social referencing at ten months: a second-order effect on infants' responses to strangers. Child Development 54 (4), 878–887. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/6617309.

Flom, R., Bahrick, L.E., 2007. The development of infant discrimination of affect in multimodal and unimodal stimulation: the role of intersensory redundancy. Developmental Psychology 43 (1), 238–252. http://dx.doi.org/10.1037/0012-1649.43.1.238.

Harkness, S., Super, C.M., 1995. Culture and parenting. In: Bornstein, M.H. (Ed.), Handbook of Parenting: Biology and Ecology of Parenting, vol. 2. Lawrence Erlbaum, Mahwah, NJ, pp. 211–234.

Hoehl, S., Wiese, L., Striano, T., 2008. Young infants' neural processing of objects is affected by eye gaze direction and emotional expression. PloS One 3 (6), e2389. http://dx.doi.org/10.1371/journal.pone.0002389.

Hornik, R., Gunnar, M.R., 1988. A descriptive analysis of infant social referencing. Child Development 59 (3), 626–634. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/3383671.

Hornik, R., Risenhoover, N., Gunnar, M., 1987. The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. Child Development, 937–944.

Itakura, S., 1995. An exploratory study of social referencing in chimpanzees. Folia Primatologica 64 (1–2), 44–48.

Kaminski, J., Brauer, J., Call, J., Tomasello, M., 2009. Domestic dogs are sensitive to a human's perspective. Behaviour 146 (7), 979–998.

Kärtner, J., Keller, H., Yovsi, R.D., 2010. Mother–infant interaction during the first 3 months: the emergence of culture-specific contingency patterns. Child Development 81 (2), 540–554.

Keller, H., Borke, J., Lamm, B., Lohaus, A., Yovsi, R.D., 2011. Developing patterns of parenting in two cultural communities. International Journal of Behavioral Development 35 (3), 233–245.

Kim, G., Walden, T.A., Knieps, L.J., 2010. Impact and characteristics of positive and fearful emotional messages during infant social referencing. Infant Behavior & Development 33 (2), 189–195. http://dx.doi.org/10.1016/j.infbeh.2009.12.009.

Klinnert, M.D., Emde, R.N., Butterfield, P., Campos, J.J., 1986. Social referencing: the infant's use of emotional signals from a friendly adult with mother present. Developmental Psychology 22 (4), 427–432. http://dx.doi.org/10.1037//0012-1649.22.4.427.

Liszkowski, U., 2005. Human twelve-month-olds point cooperatively to share interest with and helpfully provide information for a communicative partner. Gesture 5 (1–2), 1–2.

Liszkowski, U., Carpenter, M., Tomasello, M., 2008. Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. Cognition 108 (3), 732–739.

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., de Vos, C., 2012. A prelinguistic gestural universal of human communication. Cognitive Science.

Merola, I., Prato-Previde, E., Marshall-Pescini, S., 2012a. Social referencing in dog-owner dyads? Animal Cognition 15 (2), 175–185.

Merola, I., Prato-Previde, E., Marshall-Pescini, S., 2012b. Dogs' social referencing towards owners and strangers. PloS One 7 (10), e47653.

Moses, L.J., Baldwin, D.A., Rosicky, J.G., Tidball, G., 2001. Evidence for referential understanding in the emotions domain at twelve and eighteen months. Child Development 72 (3), 718–735.

Mumme, D.L., Fernald, A., 2003. The infant as onlooker: learning from emotional reactions observed in a television scenario. Child Development 74 (1), 221–237.

Mumme, D.L., Fernald, A., Herrera, C., 1996. Infants' responses to facial and vocal emotional signals in a social referencing paradigm. Child Development 67 (6), 3219–3237. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/9071778.

Pack, A.A., Herman, L.M., 2006. Dolphin social cognition and joint attention: our current understanding. Aquatic Mammals 32 (4), 443–460. http://dx.doi.org/10.1578/AM.32.4.2006.443.

Peltola, M.J., Leppänen, J.M., Palokangas, T., Hietanen, J.K., 2008. Fearful faces modulate looking duration and attention disengagement in 7-month-old infants. Developmental Science 11 (1), 60–68. http://dx.doi.org/10.1111/j.1467-7687.2007.00659.x.

Peltola, M.J., Leppänen, J.M., Mäki, S., Hietanen, J.K., 2009. Emergence of enhanced attention to fearful faces between 5 and 7 months of age. Social Cognitive and Affective Neuroscience 4 (2), 134–142.

Repacholi, B.M., 1998. Infants' use of attentional cues to identify the referent of another person's emotional expression. Developmental Psychology 34 (5), 1017.

Roberts, S.G.B., McComb, K., Ruffman, T., 2008. An experimental investigation of referential looking in free-ranging Barbary macaques (*Macaca sylvanus*). Journal of Comparative Psychology (Washington, D.C.: 1983) 122 (1), 94–99. http://dx.doi.org/10.1037/0735-7036.122.1.94.

Russell, C.L., Bard, K.A., Adamson, L.B., 1997. Social referencing by young chimpanzees (*Pan troglodytes*). Retrieved from: Journal of Comparative Psychology (Washington, D.C.: 1983) 111 (2), 185–193 http://www.ncbi.nlm.nih.gov/pubmed/9170283.

Schachter, S., Singer, J., 1962. Cognitive, social, and physiological determinants of emotional state. Psychological Review 69 (5), 379.

Sherif, M., 1958. Group influences upon the formation of norms and attitudes. Readings in Social Psychology 3, 219–232.

Sigman, M.D., Kasari, C., Kwon, J.H., Yirmiya, N., 1992. Responses to the negative emotions of others by autistic, mentally retarded, and normal children. Child Development 63 (4), 796–807.

Sorce, J.F., Emde, R.N., Campos, J.J., Klinnert, M.D., 1985. Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. Developmental Psychology 21 (1), 195–200. http://dx.doi.org/10.1037//0012-1649.21.1.195.

Stenberg, G., 2003. Effects of maternal inattentiveness on infant social referencing. Infant and Child Development 12, 399–419.

Stenberg, G., 2009. Selectivity in infant social referencing. Infancy 14 (4), 457–473. http://dx.doi.org/10.1080/15250000902994115.

Stenberg, G., Hagekull, B., 1997. Social referencing and mood modification in 1-year-olds. Infant Behavior and Development 20 (2), 209–217.

Surian, L., Caldi, S., Sperber, D., 2007. Attribution of beliefs by 13-month-old infants. Psychological Science 18 (7), 580–586.

Vaillant-Molina, M., Bahrick, L.E., 2012. The role of intersensory redundancy in the emergence of social referencing in $5^1/_2$ -month-old infants. Developmental Psychology 48 (1), 1–9. http://dx.doi.org/10.1037/a0025263.

Vaish, A., Striano, T., 2004. Is visual reference necessary? Contributions of facial versus vocal cues in 12-month-olds ' social referencing behavior Amrisha Vaish and Tricia Striano. Developmental Science 7 (3), 261–269.

Vaish, A., Grossmann, T., Woodward, A., 2008. Not all emotions are created equal: the negativity bias in social-emotional development. Psychological Bulletin 134 (3), 383–403. http://dx.doi.org/10.1037/0033-2909.134.3.383.

Vaish, A., Demir, Ö.E., Baldwin, D., 2011. Thirteen-and 18-month-old infants recognize when they need referential information. Social Development 20 (3), 431–449.

Walden, T., Kim, G., 2005. Infants' social looking toward mothers and strangers. International Journal of Behavioral Development 29 (5), 356–360. http://dx.doi.org/10.1080/01650250500166824.

Walden, T.A., Ogan, T.A., 1988. The development of social referencing. Child Development 59 (5), 1230–1240.

Zarbatany, L., Lamb, M.E., 1985. Social referencing as a function of information source: mothers versus strangers. Infant Behavior and Development 8 (1), 25–33.

Zumbahlen, M., Crawley, A., 1996. Infants' early referential behavior in prohibition contexts: the emergence of social referencing. Infant Behavior and Development 19, 244.

Social Relationships in Adulthood

Margaret S Clark, Trumbull College, Yale University, New Haven, CT, USA
Oriana R Aragon, Center for Scientific Teaching, Yale University, New Haven, CT, USA
Jennifer L Hirsch, Yale University, New Haven, CT, USA

Published by Elsevier Ltd.
This article is a revision of the previous edition article by M.S. Clark, volume 21, pp. 14423–14429, © 2001, Elsevier Ltd.

Abstract

The nature of adult close, communal, relationships is outlined. We define high quality, close, adult relationships in terms of the interpersonal processes that, ideally, characterize them rather than in terms of stability, satisfaction, or lack of conflict. We explain why it is desirable to define them in the former rather than the latter manner. Then, factors that result in people being attracted to one another and that motivate them to form such relationships are reviewed. Next, processes involved in initiating such relationships and the importance of, eventual, commitment to relational functioning are discussed. Placement of a communal relationship within a larger hierarchically arranged set of relationships is also discussed. Finally, we briefly touch on how a lack of trust in a potential or existing communal partner undermines the quality of these relationships.

What Is a Close Relationship?

Adults have many types of relationships: brief relationships with strangers, relationships with acquaintances, ones with people with whom they do business, as well as relationships that are commonly called *close* relationships. This article is about close adult relationships for it is these relationships, including friendships, romantic relationships, and family relationships, that have received the most research attention.

Whereas, people will readily agree that relationships with good friends, family members, and romantic relationships often exemplify close relationships; what does the term 'close' really mean? Researchers have suggested a variety of definitions.

One definition has to do with the degree to which each person is capable of influencing and does influence the other person's thoughts, feelings, and behavior (Berscheid et al., 1989; Rusbult and Van Lange, 2012). This definition picks up those relationships that most impact the nature of our hour-to-hour and day-to-day behavior, thoughts, emotions, and, in turn, well-being. This definition has proven useful for many purposes and, interestingly, it captures both our most important enemies, as well as most important friends (*see* Friendship during Adolescence and Cultural Variations; Friendship during Childhood and Cultural Variations).

Other researchers define closeness in terms of the extent to which each member assumes responsibility for promoting the welfare of the other member (Clark and Monin, 2006; Clark and Mills, 2012; Monin et al., 2008) or the, overlapping, extent to which people actually do understand, validate, and care for one another (Reis and Patrick, 1996; Reis and Shaver, 1988). Still others define closeness as the extent to which a partner's attributes, goals, and desires are considered to be a part of the self (Aron et al., 2004) or as the extent to which one is committed to a relationship defined as intending to stay in the relationship for whatever reason, be it satisfaction, lack of alternatives, past investments, or social or personal norms (Rusbult et al., 2006).

No one definition is 'correct.' The antecedents and consequences of each type of closeness differ and can (and should) be studied separately. All varieties of closeness likely contribute to a sense that a relationship is important.

For purposes of this article, successful closeness can be considered to consist in: (1) members being responsive to one another meaning that they support one another's welfare (Clark and Mills, 2012; Reis et al., 2004; Reis and Clark, 2013) and (2) members facilitating their own receipt of responsiveness from their partners by expecting, seeking, desiring, and accepting such responsiveness (Clark and Mills, 2012; Clark et al., 1998; Clark, 2011; *see* Welfare). Hallmarks of these relationships include support being given in response to partners' needs or desires and sometimes simply to signal care and, importantly, this support being given noncontingently. That is, the recipient does not owe the donor anything for responsiveness received (Clark and Mills, 1979). Needs are tracked and attended to (Clark et al., 1986) but just what has been actually given and received, ideally, is not (Clark, 1984).

These relationships are communal relationships (Clark and Mills, 2012). Among adults, they are typically mutual and symmetrical meaning that each person (friend, romantic partner) assumes about the same amount of responsibility for the other. Needs and desires as well as partner ability to meet them can differ so the actual giving and receipt of care may differ between partners.

Not All Communal Relationships Are Equally Strong

Successful communal relationships differ in the degree of responsibility members assume for partner welfare. These relationships vary in strength meaning that some close, communal relationships, entail members taking on extremely high levels of responsibility for one another's welfare (meaning that people expend extreme amounts of their own effort, time, and resources to enhance their partner's welfare) whereas the members of other, also communal, relationships take on more moderate or even low levels of such mutual responsibility but still treat those in a communal manner (Mills et al., 2004). Communal relationships of a variety of strengths can be successful for they can

operate on a communal basis (up to a level strength implicitly agreed and acted upon by both members).

Sometimes benefits that exceed that implicit level are given and accepted on a tit-for-tat or exchange basis. In emergency situations that upper bound may be violated.

Defining High-Quality Relationships in Terms of Good Interpersonal Processes

Trusting, noncontingent, and sustainable communal responsiveness can and does take on many forms: mutual tracking of needs and desires, expressions of one's emotions allowing partners to most easily infer needs and desires, efforts made to understand and validate partners, the giving and receiving care, and the inclusion of the partner and mutual engagement in satisfying activities. Whereas, much past work has suggested that high-quality close relationships can be defined in terms of their stability (persistence across time), high individual ratings of satisfaction, and/or absence of conflict, we suggest that use of such markers as indices of well-functioning communal relationships may often be flawed. First, not all stable relationships are happy ones. People often stay in poorly functioning relationships because they have poor alternatives or because they have invested much in the relationship that cannot be recouped (Rusbult, 1983). Second, people can be satisfied with a poor relationship because they have low standards (comparison levels) for the relationship given abusive past histories (Rusbult, 1983). Finally, conflict can be absent from a relationship simply because one person dominates over another who fears him or her. To evaluate relationships, therefore, it is best to directly observe whether they are characterized by trust, understanding, validation, and by acts taken to promote partner welfare given on a noncontingent basis along with an openness to reveal vulnerabilities and to accept help from one's partner.

Responsiveness Is a Key to the Existence of High-Quality Relationships, but What Is Responsiveness?

Communal care, often called responsiveness, takes many forms. A person can provide help to a partner in clear need by, say, helping him or her move his or her furniture into a new apartment. So too can care take the form of supporting one's partner in attaining his or her goals either actively by taking steps to support goal attainment such as helping a friend who wishes to get him or her license study for a driver's test or by giving a person time and space to pursue a good (Feeney, 2004; *see* Social Support, Psychology of). It can involve celebrating another person's successes, a process known as capitalization (Gable et al., 2004). It can involve satisfying a partner's need to belong by including him or her in conversations (Aron et al., 2000) or in games (Williams, 2007). Responsiveness also can be primarily symbolic. One partner, for instance, may tell a partner that he or she is a good friend, send a card, or simply smile at a partner's comments (Drigotas, 2002; Swann, 1987). Yet other forms of responsiveness involve putting up with a partner's bad behavior without reacting negatively, a process known as accommodation (Rusbult et al., 2005) or suppressing one's own needs when a partner has needs that surpass one's own.

Both giving and accepting care contribute to feelings of responsiveness in adult relationships. Another way to put this is to say that all behaviors that are intended to result in members of a relationship promoting one another's welfare contribute to responsiveness.

What Drives the Formation of Close Adult Relationships?

The existence of well-functioning communal relationships in a person's life contributes to that person's physical and mental well-being, a fact now well known and reported by psychologists and epidemiologists alike (e.g., Cacioppo et al., 2011).

From infancy on, we all have a need to belong, to connect with other, and to form supportive relationships (Mikulincer and Shaver, 2007; Baumeister and Leary, 1995; *see* Infancy and Childhood: Emotional Development). Indeed our own self-esteem appears to be based upon our acceptance by others (Leary and Baumeister, 2000). Thus, perhaps the most important factor motivating people to form new adult communal relationships is a current lack of a sufficient number of effectively functioning communal relationships resulting in feelings of loneliness. For instance, those who have lost a romantic partner will likely seek another. Those who leave one location to move to a new one will seek new, local, friends. A person who becomes estranged from one family member may turn toward initiating or, more likely, strengthening the communal relationship with another family member (*see* The Family as Institution). Supporting the claim that lack of sufficient number of communal relationships is a driving force toward forming new ones, researchers have found that reminding people of social rejections from a current partner, on average, increases their interest in meeting new people (Maner et al., 2007: studies 1 and 2; Williams and Sommer, 1997). Mirroring this, so too is it true commitment to current close relationship partners decreases both attention and attraction to new potential partners (Johnson and Rusbult, 1989; Lydon et al., 2003; Maner et al., 2009, 2008) and having a sufficient number of communal relationships with friends and family should reduce a drive to form new communal relationships.

Of course, once a person feels the drive to establish new communal relationships, that person must decide with whom to establish such a relationship. One important determinant of with whom adults form close, communal, relationships is kinship. People feel most responsiveness toward and expect most responsiveness from those with whom they have the closest biological ties (Monin et al., 2008). Attachment theorists suggest that parents and their infants are hardwired to form attachment bonds (which lead to noncontingent responsiveness caretakers) (Bowlby, 1969). Mothers experience surges in oxytocin at the birth of a child and also in conjunction with lactation that appears to assist them in bonding with their children (Feldman et al., 2007; *see* Prenatal and Infant Development: Overview, Current Trends, Future Directions). For their part, infants spontaneously cry when distressed eliciting care from caretakers. Moreover, their very cuteness (including

having large round heads and eyes) elicits both liking and care and precision in movement (Zebrowitz and Montepare, 1992; Sherman et al., 2009). These features, even when they occur on adult faces, elicit both judgments of physical attractiveness and responsiveness from others (Berry and McArthur, 1985). Culture also dictates high levels of responsiveness to kin as indicated in the saying, 'Blood runs thicker than water.'

Other factors also promote the formation of close relationships. Many have to do with qualities of the potential partner. People are attracted to physically attractive others (Eastwick and Finkel, 2008) an attribute that seems to consist in having features that are similar to the average of many faces (Langlois and Roggman, 1990; *see* Face Recognition, Psychological and Neural Aspects). This may be because averageness in this sense signals (or at least did signal in our evolutionary past) health and a lack of threat of contagion, and fertility (Langlois and Roggman, 1990: Thornhill and Gangestad, 1999; *see* Fertility Theory: Embodied-Capital Theory of Life History Evolution). Beyond this, researchers have found that people associate physical beauty with positive personality traits such as intelligence, popularity, and kindness (Dion et al., 1972). Something that may occur because people desire close relationships with attractive others and project that desire onto others (Lemay et al., 2010). So too may such inferences be based on the fact that, when people perceive partners to be attractive, they treat those partners in more positive ways which, in turn, brings out the best in those partners (Snyder et al., 1977). Finally, we may prefer attractive partners to others because they reflect positively on us when we are seen together (see Sigall and Landy, 1973 for men reflecting positively on women).

Perhaps the most mundane but simultaneously powerful determinant of becoming attracted toward others and, ultimately, forming close, responsive relationships with others is mere physical proximity. We are more likely to form close relationships with those who live quite close by to us than with others (Bossard, 1932; Festinger et al., 1950). Even very small differences in physical proximity can make in determining the formation of close relationships. We are more likely to form a relationship with a person in an apartment next door to us than merely two doors away (Festinger et al., 1950) and less likely to form relationships with someone seated next to us than one seat away or in our row of seats relative to another row (Back et al., 2008). Perhaps, this is because rejection hurts (MacDonald and Leary, 2005). Whereas it is natural and does not reveal one's interest in others or risk feeling rejected to talk with someone to whom one merely 'bumps into' whereas clearly making an effort to contact others reveals that one is trying to start a relationship and, therefore, makes one vulnerable to rejection. Protection from rejection may also be why another's liking for us is a potent determinant of our liking for that other (Lowe and Goldstein, 1970).

Other factors known to influence initial attraction include the other's similarity to us in attitudes, race, socioeconomic status, education, and activity preferences (Bryne, 1971; Newcomb, 1961; Klohnen and Luo, 2003; Singh et al., 2008); even similarity in birth date and name (Jones et al., 2004), familiarity (which may signal safety) (Moreland and Beach, 1992; Moreland and Zajonc, 1982), and attributes that remind us of people with whom we have had good (or bad) close relationships in the past (Andersen and Chen, 2002; Andersen et al., 1996).

How Do We Initiate, Then Maintain, Close Relationships?

Relationships with kin typically require little initiation and such relationships may be supported in ways that may have evolved (e.g., we may be hardwired to like physically similar people and to like familiar people) (*see* Partnership Formation and Dissolution in Western Societies). Other communal relationships, such as those with friends and romantic relationships, however, require complex dance of relationship initiation. After initial attraction has taken place, if we wish to establish a close, communal relationship we must promote ourselves as desirable relational partners while simultaneously evaluating the potential partner's communal attributes and protecting ourselves from the pain of potential rejection (Clark and Beck, 2011; Beck and Clark, 2010).

When it comes to self-promotion, recent research suggests that self-promotion featuring bragging about our own accomplishments or such things as our connections is not a good idea (Crocker and Canevello, 2008). Indeed, doing so to backfire makes people like the self-promoter less (Canevello and Crocker, 2010). Yet not that all self-promotion is bad; work on responsiveness suggests it just has to be the right sort of self-promotion. To initiate close, responsive relationships one must have compassionate goals. One must attend to opportunities to respond to the potential partner's needs, desires, and goals (Clark et al., 1986). Importantly, one must respond non-contingently (Clark and Mills, 1979). Then if the potential partner accepts one's responsiveness and shows signs of increased liking one can up one's responsiveness and, importantly begin seeking support as well. In fact, early in potential relationships people do seem especially likely to strategically present themselves as communally oriented. They are especially like offering help to others (relative to seeking it) (Beck and Clark, 2009b), and especially likely to bend over backward to appear not to be keeping track of just who has contributed what to a relationship lest responsiveness appear to be contingent (Clark, 1984).

Responsiveness typically drops off somewhat as relationships form, people come to trust in another's care, and they feel less need to strategically present themselves as communal in nature (Clark et al., 2010). In other words, they seem to relax a bit. After some work at promoting oneself as a caring person there comes a point at which one reveals vulnerabilities (as, for instance, by displaying emotion) and can observe how the other responds. This is known as entering a socially diagnostic situation (Beck and Clark, 2009a), and it is crucial to being able to evaluate a partner's interest in caring for the self. Of course entering or creating one's own socially diagnostic situations is also crucial to acquiring support and developing a communal relationship. Graham et al. (2008), for instance, have shown how revealing one's negative emotions indicative of needs (e.g., fear, sadness, anxiety) can elicit help and, in turn, promote the development of close, communal, relationships and as well as the development of intimacy in the closest of those

relationships. Both self-promotion and evaluating of a partner will, likely, be accompanied by a good dose of self-protection. Yet if all goes well, the person succeeds in convincing a partner that he or she is a good partner, decides that the other is a good partner as well and begins to drop the self-protection.

Making a Commitment Matters

A few close, typically kinship-based, communal relationships start with a commitment (as when a child is born and parents feel an instant commitment to care for that child across decades). However, most close, adult close relationships begin without a commitment even though people do quickly behave in communal ways in order to establish a relationship (Berg and Clark, 1986). Commitment, if it comes, comes later when higher levels of trust have been established. Commitments, when they are made, consist in decisions to stay in the relationship over the long term. Commitment is accompanied, ideally, with a drop in efforts to strategically self-present, the absence of continuing deliberations about whether to pursue the relationship or not, and letting go of self-protection (Beck and Clark, 2009a, 2010). The relationship then ideally moves to its members simply attending to one another's welfare and being appropriately responsive, conveying one's own state of welfare to the partner, and seeking and accepting appropriate responsiveness as needed. In addition and importantly, when neither person has pressing needs, mutually enjoyable and beneficial activities take place.

Ongoing Committed, Close, Communal Relationships

Ideally, communal relationships are characterized by trust, understanding, caring, support seeking, and engagement in mutually enjoyable activities. Members track each other's needs (Clark et al., 1986), understand and validate each other (Reis and Shaver, 1988), help each other, and feel good about doing so (Clark et al., 1987; Williamson and Clark, 1989, 1992) and bad when they fail to do so (Williamson et al., 1996). Members put up with partner anger without dropping liking and forgive partners (Finkel and Campbell, 2001; McCullough, 2000; Rusbult et al., 1991; Yoo et al., 2011). They are even biased toward seeing partner faults as virtues (Murray and Holmes, 1993) and hold positive illusions about one another which may become self-fulfilling prophecies (Murray and Holmes, 1997; Murray et al., 1996a,b; Murray et al., 2000). They also, seemingly automatically, protect their relationships by not being attentive to attractive alternatives to their relationship (Maner et al., 2009), and viewing their own relationship as being better than those of others (Johnson and Rusbult, 1989; Simpson et al., 1990; Van Lange and Rusbult, 1995). Such successful relationships are called, variously, communal relationships, responsive relationships, and secure relationships.

Adult Close Relationships Occur within Sets of Other Close Relationships

More frequent and appropriate responsiveness makes for better close relationships. Yet it is important to keep in mind that high-quality communal relationships can and do vary in strength (Mills et al., 2004) and that people tend to have more than one of these relationships. Very strong ones involve responsiveness that can be very costly to the responding partner in terms of time, money, and effort. These very strong communal relationships often include relationships with one's children, spouse, parents, and the closet of friends.

There is a limit to the number of close responsive relationships a person can have. People cannot be responsive to all other people's needs; they also do not need everyone to be responsive to their needs. People generally have very few, very strong communal relationships, more moderately strong ones, more casual communal relationships, and many very weak communal relationships (Clark and Mills, 2012).

The moderate and even weak communal ones characterized by expectations of more limited and less costly responsiveness but they are, nonetheless, of value and can be very successful. Members of such relationships simply must implicitly agree on the strength of the communal relationship they have or desire and their actions should fall in the appropriate bounds. For instance, casual friend ought not to give the partner an extravagant present. That would make the partner uncomfortable and threaten the relationship. Those who have weak or moderately strong or weak communal relationships also can and do understand when partners must attend to the needs of someone with whom that partner has a stronger communal relationship as, for instance, might happen when a friend says he or she cannot attend one's birthday party because his or her child is ill.

Importantly, people are responsible for their own needs and they implicitly place themselves in their own hierarchy of communal relationships. Well-adjusted people place themselves high in their hierarchies but often have one or a small number of others, for instance, a spouse or child, who they rank equally with themselves or even above themselves. It is these people for whom they are most likely to make large sacrifices or to forgive large transgressions (Clark and Mills, 2012b).

Abilities and Fortitudes Necessary to Pull off Closeness

We have emphasized success in adult relationships. Yet many people fail to achieve success in these relationships and a great deal of the scientific literature on close relationships has focused and continues to focus on impediments to forming and maintaining close relationships. There now is wide consensus that viewing the self and others positively and trusting that others can and will care for one noncontingently is a key to becoming involved in close relationships and to succeeding in them (Murray et al., 2006; Murray and Holmes, 2011). Absent confidence that others will be responsive to one and that they will accept one's responsive overtures, little support will be sought or given. Individuals who lack such confidence have been called, variously, rejection sensitive (Downey and Feldman, 1996), insecure in anxious and/or avoidant ways, or low in self-esteem (Murray et al., 2000), low in communal orientation (Clark et al., 1987). Such people appear to be chronically relationally (and defensively) self-focused meaning that they think about themselves and implications of partner for themselves constantly and have trouble allowing self-protection and focus to fade to the

background when situations call for focusing attention on partners or joint activities such as conversations, dancing, work projects that are best accomplished if people focus neither on self nor partner but on the joint activity itself (Clark et al., 2008). As a result they often neglect partners and activities and overreact to real or imagined slights, fail to enter or create situations in which they will find out that others care about them (Beck and Clark, 2009a), view others in negatively biased and often incorrect ways (Clark and Beck, 2011), and react to others' anxiety by withdrawing from them rather than providing support (Simpson et al., 1992). They may even strike out at the partners in defensive and relationship defeating ways (Murray et al., 2003). As this happens some partners may compensate striving to insure the insecure person's security (Lemay, in press). Yet this is likely at a cost to the self and the relationship and partners, at times, will withdraw.

A full discussion of how close relationships fail to be formed, fail to be adequately maintained and disintegrate is beyond the scope of this article (*see* Partnership Formation and Dissolution in Western Societies). The important point is that successful close relationships require not only knowing implicitly or explicitly what the norms of close relationships are and what is required (something most people likely know but may not see as unattainable), but also having trust in others' positive regard for the self that provides the fortitude one needs to venture forth into relationships, risk dependence, and follow those norms without fear (Murray and Holmes, 2011).

Summary

Close, communal, relationships consist in mutual intimacy (including understanding and validation of one another) and both giving and seeking responsiveness. They promote the mental and physical well-being of their members. Although poor relationships are often characterized by a lack of stability and conflict and low satisfaction, stability, lack of conflict and satisfaction are not necessarily indicative of high-quality close relationships. Rather researches the interpersonal processes involved in the provision and seeking of constructive responsiveness in these relationships (and the lack of destructive interpersonal processes) to assess whether a high quality close, communal, relationship exists.

See also: Face Recognition, Psychological and Neural Aspects; Family as Institution; Fertility Theory: Theory of Life History Evolution; Friendship During Adolescence and Cultural Variations; Friendship During Childhood and Cultural Variations; Infancy and Childhood: Emotional Development; Partnership Formation and Dissolution in Western Societies; Prenatal and Infant Development: Overview, Current Trends, Future Directions; Welfare.

Bibliography

Andersen, S.M., Chen, S., 2002. The relational self: an interpersonal social-cognitive theory. Psychological Review 109, 619–645.

Andersen, S.M., Reznik, I., Manzella, I.M., 1996. Eliciting facial affect, motivation, and expectancies in transference: significant other presentations in social relationships. Journal of Personality and Social Psychology 71, 1108–1129.

Aron, A., Aron, E.N., Tudor, M., Nelson, G., 2004. Close relationships as including the other in the self. Journal of Personality and Social Psychology. 60, 241–253.

Aron, A., Norman, C.C., Aron, E.N., McKenna, C., Heyman, R.E., 2000. Couples' shared participation in novel and arousing activities and experienced relationship quality. Journal of Personality and Social Psychology 78, 273–284.

Back, M.D., Schulkle, S.C., Egloff, B., 2008. Becoming friends by chance. Psychological Science 19, 439–440.

Baumeister, R.F., Leary, M.R., 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin 117, 497–529.

Beck, L.A., Clark, M.S., 2010. What constitutes a healthy communal marriage and why relationship stage matters. Journal of Family Theory and Review 2, 299–315.

Beck, L.A., Clark, M.S., 2009a. Choosing to enter or avoid diagnostic social situations. Psychological Science 20, 1175–1181.

Beck, L.A., Clark, M.S., 2009b. Offering more support than we seek. Journal of Experimental Social Psychology 45, 267–270.

Berg, J., Clark, M.S., 1986. Differences in social exchange between intimate and other relationships: gradually evolving or quickly apparent? In: Derlega, V., Winstead, B. (Eds.), Friendships. Springer-Verlag, New York, pp. 101–128.

Berry, D.S., McArthur, L.Z., 1985. Some components and consequences of a babyface. Journal of Personality and Social Psychology 48, 312–323.

Berscheid, E., Snyder, M., Omoto, A.M., 1989. The relationship closeness inventory: assessing the closeness of interpersonal relationships. Journal of Personality and Social Psychology 57, 792–807.

Bossard, J.H.S., 1932. Residential propinquity as a factor in marriage selection. American Journal of Sociology 38, 218–224.

Bowlby, J., 1969. Attachment and Loss: Vol. I Attachment. Basic Books, New York.

Bryne, D., 1971. The Attraction Paradigm. Academic Press, New York.

Canevello, A., Crocker, J., 2010. Creating good relationships: responsiveness, relationship quality, and interpersonal goals. Journal of Personality and Social Psychology 99, 78–106.

Cacioppo, J.T., Hawkley, L.C., Norman, G.J., Berntson, G.G., 2011. Social isolation. Annals of the New York Academy of Sciences 1231, 17–22.

Clark, M.S., 2011. Communal relationships can be selfish and give rise to exploitation. In: Arkin, R.M. (Ed.), Most Underappreciated: 50 Prominent Social Psychologists Describe Their Most Unloved Work. Oxford University Press, New York, pp. 77–81.

Clark, M.S., Lemay Jr., E.P., Graham, S.M., Pataki, S.P., Finkel, E.J., 2010. Ways of giving benefits in marriage: norm use, relationship satisfaction, and attachment-related variability. Psychological Science 21, 944–951.

Clark, M.S., Dubash, P., Mills, J., 1998. Interest in another's consideration of one's needs. Journal of Experimental Social Psychology 34, 246–264.

Clark, M.S., Mills, J.R., 2012a. A theory of communal (and exchange) relationships. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (Eds.), Handbook of Theories of Social Psychology, vol. 2. Sage Publications Ltd, Thousand Oaks, CA, pp. 232–250.

Clark, M.S., Monin, J.K., 2006. Giving and receiving communal responsiveness as love. In: Sternberg, R.J., Weis, K. (Eds.), The New Psychology of Love. Yale University Press, New Haven, CT, pp. 200–221.

Clark, M.S., 1984. Record keeping in two types of relationships. Journal of Personality and Social Psychology 47, 549–557.

Clark, M.S., Beck, L.A., 2011. Initiating and evaluating social relationships: a task central to emerging adults. In: Fincham, F., Cui, M. (Eds.), Romantic Relationships in Emerging Adulthood. Cambridge University Press, New York.

Clark, M.S., Graham, S.M., Williams, E., Lemay, E.P., 2008. Understanding relational focus of attention may help us to understand relational phenomena. In: Forgas, J., Fitness, J. (Eds.), Social Relationships: Cognitive, Affective and Motivational Processes. Psychology Press, New York, pp. 131–146.

Clark, M.S., Mills, J., 1979. Interpersonal attraction in exchange and communal relationships. Journal of Personality and Social Psychology 37, 12–24.

Clark, M.S., Mills, J., 2012b. Communal (and Exchange) relationships. In: Van Lange, P.A.M., Kruglanski, A.W., Higgins, E.T. (Eds.), Handbook of Theories of Social Psychology. Sage Publications, Thousand Oaks, CA, pp. 232–250.

Clark, M.S., Mills, J., Powell, M.C., 1986. Keeping track of needs in communal and exchange relationships. Journal of Personality and Social Psychology 51, 333–338.

Clark, M.S., Ouellette, R., Powell, M.C., Milberg, S., 1987. Recipient's mood, relationship type and helping. Journal of Personality and Social Psychology 53, 94–103.

Crocker, J., Canevello, A., 2008. Creating and undermining social support in communal relationships: the role of compassionate and self-image goals. Journal of Personality and Social Psychology 95, 555–575.

Dion, K., Berscheid, E., Walster, E., 1972. What is beautiful is good. Journal of Personality and Social Psychology 24, 285–290.

Downey, G., Feldman, S.I., 1996. Implications of rejection sensitivity for intimate relationships. Journal of Personality and Social Psychology 70, 1327–1343.

Drigotas, S.M., 2002. The michelangelo phenomenon and personal well-being. Journal of Personality 70, 59–77.

Eastwick, P.W., Finkel, E.J., 2008. Sex differences in mate preferences revisited. Do people know what they initially desire in a romantic partner? Journal of Personality and Social Psychology 94, 245–264.

Feeney, B.C., 2004. A secure base: responsive support of goal strivings and exploration in adult intimate relationships. Journal of Personality and Social Psychology 87, 631–648.

Feldman, R., Weller, A., Zagoory-Sharon, O., Levine, A., 2007. Evidence for a neuroendocrinological foundation of human affiliation: plasma oxytocin levels across pregnancy and the postpartum period predict mother-infant bonding. Psychological Science 18, 965–970.

Festinger, L., Schachter, S., Back, K., 1950. Social Processes in Informal Groups: A Study of Human Factors in Housing. Harper, Oxford.

Finkel, E.J., Campbell, W.K., 2001. Self-control and accommodation in close relationships. An interdependence analysis. Journal of Personality and Social Psychology 81, 263–277.

Gable, S.L., Reis, H.T., Impett, E.A., Asher, E.R., 2004. What do you do when things go right? The intrapersonal and interpersonal benefits of sharing positive events. Journal of Personality and Social Psychology 87, 228–245.

Graham, S.M., Huang, J.Y., Clark, M.S., Helgeson, V.S., 2008. The positives of negative emotions: willingness to express negative emotion promotes relationships. Personality and Social Psychology Bulletin 34, 394–406.

Johnson, D.J., Rusbult, C.E., 1989. Resisting temptation: devaluation of alternative partners as a means of maintaining commitment in close relationships. Journal of Personality and Social Psychology 57, 967–980.

Jones, J.T., Pelham, B.W., Carvallo, M., Mirenberg, M.C., 2004. How do I love thee? Let me count the is: implicit egoism and interpersonal attraction. Journal of Personality and Social Psychology 87, 665–683.

Klohnen, E.C., Luo, S., 2003. Interpersonal attraction and personality: what is attractive – self similarity, ideal similarity, complementarity or attachment security. Journal of Personality and Social Psychology 85, 709–722.

Langlois, J.H., Roggman, L.A., 1990. Attractive faces are only average. Psychological Science 1, 115–121.

Leary, M.R., Baumeister, R.F., 2000. The Nature and Function of Self-esteem: Sociometer Theory. Advances in Experimental Social Psychology. Elsevier, North Holland.

Lemay, E.P., Regulating relationship security of chronically insecure partners. In: Carroll, P.J., Arkin, R.M., Wichman, A. (Eds.), The Handbook of Personal Security. Psychology Press, Taylor Francis, NYL, NY, in press.

Lemay, E.P., Clark, M.S., Greenberg, A., 2010. What is beautiful is good because what is beautiful is desired. Physical attractiveness stereotyping as projection of interpersonal goals. Personality and Social Psychology Bulletin 36 (3), 339–353.

Lowe, C.A., Goldstein, J.W., 1970. Reciprocal liking and attributions of ability: mediating effects of perceived intent and personal involvement. Journal of Personality and Social Psychology 18, 201–207.

Lydon, J.E., Fitzsimmons, G.M., Naidoo, I., 2003. Devaluation versus enhancement of attractive alternatives: a critical test using the calibration paradigm. Personality and Social Psychology Bulletin 29, 349–359.

MacDonald, G., Leary, M.R., 2005. Why does social exclusion hurt? the relationship between social and physical pain. Psychological Bulletin 131, 202–223.

Maner, J.K., DeWall, C.N., Baumeister, R.F., Schaller, M., 2007. Does social exclusion motivate interpersonal reconnection? Resolving the "porcupine problem". Journal of Personality and Social Psychology 92, 42–55.

Maner, J.K., Gailliot, M.T., Miller, S.I., 2009. The implicit cognition of relationship maintenance: inattention to attractive alternatives. Journal of Experimental Social Psychology 92, 174–179.

Maner, J.K., Rouby, D.A., Gonzaga, G.C., 2008. Automatic inattention to attractive alternatives: the evolved psychology of relationship maintenance. Evolution and Human Behavior 29, 343–349.

McCullough, M.E., 2000. Forgiveness as human strength: theory, measurement, and links to well-being. Journal of Social and Clinical Psychology 19, 43–55.

Mikulincer, M., Shaver, P.R., 2007. Attachment in Adulthood: Structure, Dynamics, and Change. Guilford Press, New York.

Mills, J., Clark, M.S., Ford, T.E., Johnson, M., 2004. Measurement of communal strength. Personal Relationships 11, 213–230.

Monin, J.K., Clark, M.S., Lemay, E.P., 2008. Expecting more responsiveness from and feeling more responsiveness to female than to male family members. Sex Roles 95, 420–441.

Moreland, R., Beach, S.R., 1992. Exposure effects in the classroom: the development of affinity among students. Journal of Experimental Social Psychology 28, 255–276.

Moreland, R., Zajonc, R., 1982. Exposure effects in person perception: familiarity, similarity & attraction. Journal of Experimental Social Psychology 18, 395–415.

Murray, S.L., Bellavia, G.M., Rose, P., Griffin, D.W., 2003. Once hurt, twice hurtful: how perceived regard regulates daily marital interactions. Journal of Personality and Social Psychology 84, 126–147.

Murray, S.L., Holmes, J.G., 2011. Interdependent Minds: The Dynamics of Close Relationships. Guilford Press, New York.

Murray, S.L., Holmes, J.G., 1993. Seeing virtues in faults: negativity of the transformation of interpersonal narratives in close relationships. Journal of Personality and Social Psychology 65, 707–722.

Murray, S.L., Holmes, J.G., 1997. A leap of faith? Positive illusions in romantic relationships. Personality and Social Psychology Bulletin 23, 586–604.

Murray, S.L., Holmes, J.G., Collins, N., 2006. Optimizing assurance: the risk regulation system in relationships. Psychological Bulletin 132, 641–666.

Murray, S.L., Holmes, J.G., Dolderman, D., Griffin, D., 2000. What the motivated mind sees: comparing friends' perspectives to married partners' views of each other. Journal of Experimental Social Psychology.

Murray, S.L., Holmes, J.G., Griffin, D., 1996a. The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. Journal of Personality and Social Psychology 70, 79–98.

Murray, S.L., Holmes, J.G., Griffin, D., 1996b. The self-fulfilling nature of positive illusions in romantic relationships: love is not blind, but prescient. Journal of Personality and Social Psychology 71, 1155–1180.

Newcomb, T.M., 1961. The Acquaintance Process. Holt, Rinehart, & Winston, New York.

Reis, H.T., Clark, M.S., Holmes, J.G., 2004. Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness. In: Mashek, D.J., Aron, A.P. (Eds.), Handbook of Closeness and Intimacy. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 201–225.

Reis, H.T., Patrick, B.C., 1996. Attachment and intimacy: component processes. In: Higgins, E.T., Kruglanski, A.W. (Eds.), Social Psychology: Handbook of Basic Principles. Guilford Press, New York.

Reis, H.T., Shaver, P., 1988. Intimacy as an interpersonal process. In: Duck, W.W. (Ed.), Handbook of Personal Relationships. Wiley, New York, pp. 367–391.

Reis, H.T., Clark, M.S., 2013. Responsiveness. In: Simpson, J., Campbell, L. (Eds.), Handbook of Close Relationships. Oxford University Press, Oxford, England, pp. 400–423.

Rusbult, C.E., 1983. A longitudinal test of the investment model: the development (and deterioration) of satisfaction and commitment in heterosexual involvements. Journal of Personality and Social Psychology 45, 101–117.

Rusbult, C.E., Hannon, P.A., Stocker, S.L., Finkel, E.J., 2005. Forgiveness and relational repair. In: Worthington Jr., E.L. (Ed.), Handbook for Forgiveness. Brunner-Routledge, New York, pp. 185–206.

Rusbult, C.E., Van Lange, P.A.M., 2012. Interdependence theory. In: Kruglanski, A.W., Higgins, E.T., Van Lange, P.A.M. (Eds.), Handbook of Theories of Social Psychology. Sage Publications, Ltd, New York.

Rusbult, C.E., Coolsen, M.K., Kirchner, J.L., Clarke, J.A., 2006. In: Vangelisti, A., Perlman, D. (Eds.), The Cambridge Handbook of Personal Relationships. Cambridge University Press, New York, pp. 615–635.

Rusbult, C.E., Verette, J., Whitney, G., Slovik, L., Lipkus, I., 1991. Accommodation processes in close relationships: theory and preliminary empirical evidence. Journal of Personality and Social Psychology 60, 53–78.

Sherman, G.D., Haidt, J., Coan, J.A., 2009. Viewing cute images increases behavioral carefulness. Emotion 9, 282–286.

Sigall, H., Landy, D., 1973. Radiating beauty: effects of having a physically attractive partner on persona perception. Journal of Personality and Social Psychology 28, 218–224.

Simpson, J.A., Rholes, W.S., Nelligan, J.S., 1992. Support seeking and support giving within couples in an anxiety-provoking situation: the role of attachment styles. Journal of Personality and Social Psychology 62, 434–446.

Simpson, J.A., Gangestad, S.W., Lerma, M., 1990. Perception of physical attractiveness: mechanisms involved in the maintenance of romantic relationships. Journal of Personality and Social Psychology 59, 1192–1201.

Singh, R., Ng, R., Ong, E.L., Lin, P.K.F., 2008. Different mediators for the age, sex and attitude similarity effects in interpersonal attraction. Basic and Applied Social Psychology 30, 1–17.

Snyder, M., Tanke, E.O., Berscheid, E., 1977. Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology 35, 656–666.

Sprecher, S., Felmlee, D., 1992. The influence of parents and friends on the quality and stability of romantic relationships: a three-wave longitudinal investigation. Journal of Marriage and the Family 34, 888–900.

Swann, W.B., 1987. Identity negotiation: where two roads meet. Journal of Personality and Social Psychology 53, 1038–1051.

Thornhill, R., Gangestad, S.W., 1999. Facial attractiveness. Trends in Cognitive Sciences 3, 452–460.

Van Lange, P.A.M., Rusbult, C.E., 1995. My relationship is better than – and not as bad as – yours is: the perception of superiority in close relationships. Personality and Social Psychology Bulletin 21, 32–44.

Williams, K.D., 2007. Ostracism. Annual Review of Psychology 58, 425–452.

Williams, K.D., Sommer, K.L., 1997. Social ostracism by coworkers: does rejection lead to loafing or compensation? Personality and Social Psychology Bulletin 23, 693–706.

Williamson, G.M., Clark, M.S., 1989. Providing help and desired relationship type as determinants of changes in moods and self-evaluations. Journal of Personality and Social Psychology 56, 722–734.

Williamson, G.M., Clark, M.S., 1992. Impact of desired relationship type on affective reactions to choosing and being required to help. Personality and Social Psychology Bulletin 18, 10–18.

Williamson, G.M., Pegalis, I., Behan, A., Clark, M.S., 1996. Affective consequences of refusing to help in communal and exchange relationships. Personality and Social Psychology Bulletin 22, 34–47.

Yoo, S.H., Clark, M.S., Lemay, E.P., Salovey, P., Monin, J.K., 2011. Responding to partners' expression of anger: the role of communal motivation. Personality and Social Psychology Bulletin 37, 229–241.

Zebrowitz, L.A., Montepare, J.M., 1992. Impressions of babyfaced individuals across the life span. Developmental psychology 26, 1143–1152.

Social Resilience in Post-Tsunami Japan

Jin Sato, The University of Tokyo, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A disaster of unprecedented magnitude hit Japan in March 2011 when a massive tsunami slammed into the country's northeastern region. In the face of that tragedy, the people of Japan demonstrated remarkable resilience, and using the concept of 'resilience' as our guide, this article attempts to draw lessons from that experience. To do so, the article analyzes the structural conditions of certain affected regions that made them particularly vulnerable. It also proposes to suggest ways to encourage societies to help prepare for future disasters by learning to take advantage of less visible, often untapped potential within the community.

Diverse Disaster Experiences

Nearly 20 000 people were killed by the tsunami of 11 March 2011 and more than a year later 100 000 inhabitants were still displaced from their homes. Describing the magnitude of the disaster only in terms of quantifiable dimensions does not do justice to the many deep-rooted and hidden aspects of the tragedy: thousands of farmers losing their livestock and land, forcing them to abandon their once proud profession; whole villages relocated to resettlement sites where they faced discrimination; countless elders transplanted away from familiar surroundings of home and family, loneliness adding to their grief; families forced to leave friends and pets behind during evacuation. Even those who may be considered as 'lucky' for being allowed to return to their homes still face enormous challenges in attempts to restore their communities. The lack of employment opportunities is a major issue, as is the shortage of space particularly in Fukushima for storing decontaminated debris. Although outsiders tend to consider 'victims' and 'affected areas' as uniform entities, these regions themselves are characterized by enormous diversity with respect to culture, economy, and the damage induced by the tsunami.

Victims of the nuclear disaster in Fukushima, on the other hand, were concerned with totally different issues. Initially, the government failed to publicize the exact extent of nuclear contamination, and many children were exposed to high levels of radiation during the first few days; consequently, their future lives will always be tainted by the fear of possible health effects. Marriage for young women in Fukushima may face obstacles due to the unfounded rumor that their genes may be defective because of radiation, not to mention the worries of pregnant women. People are just beginning to comprehend the full extent of the uncertainties caused by the tsunami, and that they must confront these issues for at least the next few generations.

The flipside of this tragic prospect is the remarkable resilience of the people in their daily survival: "the capacity of a system to absorb disturbance and reorganize while undergoing change so as to still retain essentially the same function, structure, identity, and feedbacks" (Walker et al., 2004). In the context of a disaster, resilience signifies flexibility of the social system, the ability to 'resile from' or 'spring back' from a shock (UNISDR, 2009).

In the aftermath of the tsunami, the most important 'resource' was the network of people and the information conveyed through personal links and contacts with neighbors. This was particularly important in establishing communications to distant villages when electric and phone lines were down. For example, neighboring communities and municipalities, rather than the central or prefectural governments, were the first to provide food aid, prompting many victims to claim that having been without food for a few days, they would always remember the taste of the first rice ball (Tokyo Foundation, 2012). Local moving companies, able to take advantage of their knowledge of the roads and the availability of their vehicles, helped to transport emergency goods; furniture companies provided basic supplies to restart lives in the evacuation centers. Disaster of the magnitude of the 3/11 tsunami and the subsequent nuclear accident in Fukushima can be seen as a true test of how people, communities, and the state can pool their diverse resources, either as a planned or spontaneous effort, to support victims to carry on in the face of extreme shock.

The purpose of this article is to learn from this unprecedented disaster, applying resilience as our guiding concept. Let me hasten to add that identification of these lessons is, by definition, a political endeavor that involves the arbitrary decision of what constitutes important factors. My approach here is not to single out numerous individual lessons, classify them, and attach labels to each. Instead, I examine the context or background, which induced the particular events that helped to strengthen or weaken the resilience of the people.

The Making of Tohoku, a Disaster-Affected Region

Yamauchi (2012) argues that even more than a year after 3/11, most Japanese are yet to recognize that the country has lost a part of its capacity to supply such national essentials as energy and rice. While general consideration of resilience tends to focus on the actors and their behavior, I shall begin by taking note of the circumstances in one particular region, Tohoku, which gave rise to the type of livelihood that made the district very vulnerable to external shocks.

How was the region of Tohoku (literally meaning 'northeast') formed during the modernizing process of Japan? In August 1945 after the termination of the war against the Allied Powers, the immediate problem for democratized Japan was food supply, as the country had lost about 40% of its territory, including Korea and Taiwan, which had produced significant

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64044-8

amounts of rice for consumption on the mainland. Daily rations (consisting of rice, sugar, salt, tobacco, etc.) imposed by the government from 1946 to 1947 were usually less than half the minimum (2160 calories) considered to constitute an adequate diet (General Headquarters, 1949). With a considerable decline in the male population because of the war effort, and with productive capacity completely skewed, domestic capacity to produce food decreased roughly 40% compared to peace-time production levels (Yamauchi, 2011: p. 257).

A Harvard geographer, Edward Ackerman, who came to Japan as a technical advisor to the Supreme Commander of Allied Powers noted at the time (Ackerman, 1949: p. 7):

> On the basis of my travels and studies, I can say that even a "well-to-do" person in Japan today lacks most things a middle class American deems a necessity of life Heat in the winter is one. The lot of average man is one of continual hard labour, if a family is to be kept fed, clothed and housed in even a minimal manner. A house without heat, a diet without variety is the lot of the majority … He furnishes only the margin between hunger and life.

Expansion of rice fields in the inland periphery was encouraged, and Tohoku became the prime target for saving the nation from potential famine. This policy was promoted through such infrastructure development as irrigation systems and the introduction of improved fertilizers. Present-day Tohoku is not only a major supplier of rice (26%), but also the source of abundant marine products, accounting in 2008 for 14.5% of fishery products and 11.2% of aquaculture products (Ministry of Agriculture, Forestry and Fisheries, 2009), both of which make up the very foundation of the basic diet of Japanese people.

Food was not the only concern; the general mobilization period of the 1930s in preparation for war had focused on Tohoku to develop it as a major supply center for labor and energy (Oguma, 2011). It was only in 1966 that Japan was able to reach self-sufficiency in rice, at which time the government ironically began to suppress production in order to reduce oversupply. Labor was another important resource to be supplied by Tohoku, and students straight out of school flocked into its cities to meet the labor demand during the high economic growth period from the 1950s onwards.

It was exactly these characteristics that pushed Tohoku in a particular direction. The Japanese government has gradually 'colonized' Tohoku since the war to serve its objectives. Unlike external colonization, which involves a military invasion of an outside territory, 'internal colonization' is characterized by the region's eventual willingness to be integrated with the empire, i.e., the periphery willingly accepts its specific role within the political economy of a nation (Kainuma, 2011).

In fact, central authorities in Tokyo had long had the unfulfilled ambition to bring the countryside under its control. Despite efforts to send preappointed governors from the Home Ministry to each prefecture, internal colonization during the Meiji era had not succeeded, since there was no 'material' necessity or benefit for the localities to follow the direction set by Tokyo. Internal colonization, which had been triggered by introduction of the 'general mobilization' efforts for external war purposes in the 1930s, resulted in the structural delineation of Tohoku suppliers and the Tokyo consumers in electric power and food. Nuclear plants became the magical tool for creating employment *in situ*, the periphery no longer needed to rely on cities for employment. Furthermore, nuclear plants made it possible to counter-balance power relations: the countryside, being the 'center' of electricity, was prioritized over cities (Kainuma, 2011). With time, nuclear power plants became an indispensable part of the nation's economy that had been achieved through the 'automatic' subordination of the peripheries. Oguma (2011) concludes that the real reconstruction of Tohoku after 3/11 will require dismantling the legacy of center–periphery structure that dates back to the 1930s.

What should have been foreseen was the creation of a structure that triggered inequality and domination because of the country-side's importance in the production of food and energy. Yet only a few dared to address this problem, since the major constituents of the game, i.e., electricity companies and plant-recipient communities or 'villages' that were made up of comrades loyal to the system, gained from this relation and were thus happy. This structure of inequality was another, albeit hidden, element of stifled diversity.

Examples of Local Resilience

It may be useful to put the concept of 'resilience' back in its original place and discuss the ecological foundation of social resilience. As was already pointed out, the main characteristic of the region devastated by the 3/11 tsunami was its concentration of farmers and fishermen, i.e., people working in the primary industry. This had a double effect after 3/11. Some farmers, faced with considerable equipment losses and salt-water soaked rice fields, took this as an opportunity to quit their business, while others with help from outside organizations found innovative ways to recover. For example, in Onagawa town of the Miyagi prefecture, a 'welfare agriculture' project was developed. This involved labor-intensive vinyl greenhouses for growing tomatoes. Instead of aiming for profit and production, the program's main goal was to facilitate communications, encourage exercise, and to give the elderly and pensioners something to do. Unlike other communities facing difficulties in rebuilding themselves, civic movement in Onagawa is strong and innovative.

For many of the fishermen who had lost their boats and equipment, the tsunami-induced devastation was the impetus to give up, a thought they might have had for several years because the younger generation was unwilling to continue the family tradition of fishing, preferring to move to the cities. For others, despite the losses, the remarkable revival of water quality and marine life held promise and strengthened their determination to restart more sustainable fisheries.

Forests are Longing for the Sea, the Sea is Longing for the Forests

Perhaps one of the most famous innovations by the local people in the disaster-affected regions is the "forests are longing for the sea" movement, which is a tree-planting

project to regenerate the forests. Led by the charismatic Shigeatsu Hatakeyama, the project has been in effect in the Kesennuma area since 1989 when he and his colleagues began planting broadleaf trees upstream of the Okawa River to promote forest drainage. To date, the group has planted more than 30 000 trees. This has had a visible impact on the water quality downstream and the Kesennuma Bay. As Hatakeyama notes (2012) "The fishermen look ashore to the mountain forests as their guiding post when they return home from their trips". The activities under its catchy slogan – "forests are longing for the sea" – has accumulated considerable respect both domestically and internationally.

This local movement extends beyond the act of simply planting trees and reviving the weakening link between the forests and the sea. Oyster farmers, by nature of their profession, are considered over the purity of the river water, essential for growing oysters. As a result of economic growth and improvements in living standards, the coastal environment has deteriorated, with pollution entering the river from sewage, agricultural wastewater, and oil spreading from fishing processing facilities. This resulted in an outbreak of red tide plankton in the 1970s, causing the oyster meat to turn red in color, which decreased its market value. The successful campaign led by the fishermen resulted in halting the dam construction in the upstream as well cancelling regional development plans for large-scale forest cutting and land reclamation.

The impressive and dynamic efforts of the oyster farmers have also become an effective educational tool for understanding the vague and elusive relationship between resources and the environment (Oh, 2013). The project has served as an example for children to learn about the unity of nature and how forest deterioration and river pollution damage the fisheries. This enhanced mutual concern and communication between families living in the coastal and mountain areas. The movement, although originally implemented to restore the region's ecology, became an important symbol of public involvement during posttsunami reconstruction, particularly during protest against the government plan to install a high seawall to protect against future tsunamis. The concept of "the sea is longing for the forest" resonated well with the post-Fukushima shift in public awareness against hi-tech with more emphasis on making the best use of surrounding nature and environment. Now, many NGOs and civic organizations, both domestic and international, are supporting this locally initiated activity.

The Miracle of Kamaishi

After 3/11, foreign media praised the resilience of the people in Tohoku, and reported on how 'disciplined' the tsunami victims were. Reporters were particularly impressed by the people quietly lining up in cold weather, waiting for food and water without bickering or conflict. One of the most celebrated example of 'discipline' was the so-called the Miracle of Kamaishi.

For the last 7 years, schools in Kamaishi have taught the basics of evacuation by inviting experts on disaster management to speak to the students. The golden rules drilled into the children were "Don't make assumptions about disasters" and "Put yourself first and flee". The rather unconventional ethics of helping yourself before helping others was a reflection of a hard lesson learned from the repeated tsunami in the region, carried over from their ancestors. Although Kamaishi was struck by the catastrophic tsunami on 3/11, the majority of its 3000 primary and middle-school children survived, having decided on their own that it would be risky to take shelter at designated evacuation sites, choosing, instead, to flee to higher ground to escape the incoming wave. Many lives *elsewhere* were lost because people went out to search for their families instead of following this rule to flee. Fleeing on your own, and trusting that others are doing the same, requires courage. The students' actions were encouraged by their awareness of the dangers of the tsunami and how it behaves, but also by their successful application of this knowledge that had been a part of the school's routine curriculum.

Discipline and Security

Michel Foucault (2009) once made a useful distinction between 'discipline' and 'security'. Both of these concepts, which shall be explained below, are critical in understanding how we address disaster of abnormal scale, i.e., extreme events that from a statistical point of view are expected to occur only once in 1000 years.

The classical system of order implemented through legal codes was based on the concept of 'the permitted' and 'the forbidden': "don't do this, don't do that" in a binary form along with a certain gradation of punishment. According to Foucault, the modern system of governance, on the other hand, requires a series of supervisions, inspections, and various controls before a crime is even committed. This shift in the techniques of governance has allowed the penetration of power to individuals as 'disciplines'. 'Security', on the other hand, is concerned with the macro effects of individual behavior, and is summarized by such questions as the average crime rate, what factors might affect the increase or decrease of such rates, and perhaps most importantly, at what costs. The goal, therefore, is to identify "socially and economically acceptable limits and around an average that will be considered as optimal for a given social functioning" (Foucault, 2009: p. 5).

The distinction here is not just the perception of the source of the control and its intensity; rather it is the origin of the norm and how it could be internalized back into the hands of individuals. While discipline originates from a particular norm at the level of the individual and his soul, security comes from external probabilities, the sources of which that cannot be directly perceived: the probability of a traffic accident, typhoon, tsunami, etc.

What was the relation between security and discipline in the context of 3/11? I have argued that the basic foundation of this phenomenon was structured through Japan's strategy for economic development. Both security and discipline originate from the identical source of values. Whether individually or collectively, someone at the center of the decision-making progress will decide and deploy what is to be done, or more precisely, what is to be prohibited. In the case of the tsunami, decisions were related to which evacuation center was to be

used, where not to go, whose instructions should obeyed, etc. Many did comply with government instructions, perhaps too strictly, and subsequently became exposed to extreme danger. There is no way to verify how many lives could have been saved if people had relied on their own intuition. The point is that people had been disciplined *not to think* independently but to follow the collective 'trend'. The Miracle of Kamaishi was perhaps a rare case where traditional tsunami ethics of 'think independently' was successfully disciplined into children's behavior via collective education.

But what about the thousands of people who were misled by government information about wind direction and possible areas of nuclear contamination? While we can be taught certain safety procedures with regard to imminent danger from a tsunami, for instance, information on and attitudes toward such risks as nuclear pollution cannot be initiated by the locals alone. As with risks related to climate or epidemics, norms must be developed externally and communicated to individuals on the ground. Norms for external risks that are hard to detect (such as nuclear risk) must be determined externally, and this usually means the government must play the role. The critical point, thus, is how can this vertical communications flow be conducted to make certain that people eventually have sufficient information and knowledge for making the right decision at the right moment.

Sources of Spontaneous Resilience

Resilience constitutes perhaps more than just the strategic allocation of resources that enables populations to deal with disasters. It may also encompass a previously unrecognized 'space', apparently unoccupied, that could become the critical buffer when a society faces tragic choices.

The long tradition of top-down industrial policies has strengthened Japan's strong public reliance on central government and bureaucrats. The minimal development of civil organizations such as NGOs and NPOs independent from the government is one such indication (Sato, 2010). Ironically, Japan needed two disasters to realize the importance of responsible civic forces. The 1995 Great Hanshin earthquake gave birth for the first time to the nationwide culture of 'volunteerism', while the 3/11 tsunami and ensuring distrust of the government because of their poor response gave rise to various types of self-help movements. Mutual assistance at the communal level is not new in Japan. However, these national-scale disasters nurtured a cross-border (i.e., cross-prefectural) culture of organizational assistance, which resulted in an enriched informational environment, allowing people to seek independent sources of information, thus avoiding total reliance on government announcements.

For example, many of the Buddhist temples in tsunami-damaged regions took in evacuees even though they were not the officially designated evacuation centers. Many of the official centers had, in fact, been completely destroyed due to the overly optimistic predictions of tsunami penetration. During total collapse of the official radio system in the town of Ohtsuchi, many amateur radio operators helped to alert the community to evacuate (Asahi Shinbun, 2012). According to the statistics compiled by the Cabinet Office as of October 2012, there are currently more than 150 NPOs established *outside* the Fukushima prefecture to assist victims of Fukushima. These provide bus services at temporary housing areas, monitor radiation levels, set up children's playgrounds, and offer professional health counselling. This assistance tends to be smaller in scale than the official government-provided projects, but it is mobile and reaches those with specific needs in a timely manner. Naturally, these organizations are confronted with challenges of their own: now that 3/11 is increasingly fading away as an event of the past, donations to these NGOs are diminishing, which threatens the continuity of their activities.

The argument here is that it is not possible to be directly involved in nurturing resilience, since an important part of resilience constitutes an intangible capacity scattered within a society. What we could do, instead, is to offer an environment where this capacity can easily be summoned when necessary to play an effective role in minimizing external shocks. I propose that in order for a society to become resilient, it must embrace plurality, and value social change and development. A single criterion, such as efficiency, which has occupied an increasingly dogmatic position in all spheres of Japanese economy, should not downplay the other criteria for a livable society. Next I elaborate on some principles that are relevant in designing such an environment.

Building an Environment for Resilience

According to the UN, "resilience of a community in respect to potential hazard events is determined by the degree to which the community has the necessary resources and is capable of organizing itself both prior to and during times of need" (UNISDR, 2009: p. 24). I agree, but the point is how do we know what is 'necessary' and how can we invoke the right resources at the right moment? As we have seen, apparently unrelated factors to disasters turned out to be useful and often critical for survival. Planners must be aware of its own limitations.

In discussing how things could have been better in the face of the 3/11 tsunami, I wish to highlight three critical dimensions of developing a resilient society: scale, speed, and a sense of ownership.

Scale

Striking among the disaster-affected areas was the fact that almost no community could have survived without external assistance. This stark reality has been motivated by the economy of scale. While each primary-industry sector, such as forestry and fisheries, may define their own boundaries through quotas and rules so as to keep resource exploitation within an acceptable range, techniques to minimize transportation distances have encouraged the unlimited application of the scale principle, making it more economical, for example, to import timber from abroad. Geographical distance per se is no longer a significant factor in transporting a good from one place to another; it is the economic cost involved in such transaction that matters. No wonder forestry is not anymore a feasible industry in Japan, despite the country's 70% forest

cover. Economic consideration is important, but we must recognize other values and principles as well. In order to diversify values and decision-making authority accordingly, officials need to reconsider whether the scale of development is appropriate. The tsunami experience has highlighted once again the value of self-supporting, if not self-sufficient, local units that can utilize diverse resources from local origins. Decentralization should be encouraged not only from a political, administrative, and economic perspective, but also from the point of strengthening local capacities to sustain themselves.

Speed of Development

Unlike the question of scale that attracts considerable attention in all investment decisions, the issue of speed of development and social change is seldom on the agenda. How can a society cope with and pay for preparations for an event that might occur only once in 1000 years? The general timeline of human action is often constrained by our own life span, and cycle of planning. State projects are usually planned for 3–5 years in accordance with the expected life span of the institutions. Education plans usually have a longer vision, perhaps 10–20 years. Pension and social security issues require an even longer vision but still fall within the range of 2–3 generations.

Perhaps the longest projection within our material culture concerns natural resource planning, particularly forestry, where trees requiring 50–60 years to mature can be expected to be logged by the next generation. In fact in the field of natural resources, adjusting speed or duration has been actively sought after. There are restrictions on resource use in many communal resource management schemes that limit the types of tools permitted, or the length of time that resources can be collected, in order to give resources time to regenerate to avoid depletion. Thus, controlling the utilization rate of resources can be critical for the survival of a community that is heavily dependent on surrounding resources. For these reasons, builders of resilient society have good reasons to listen to people working in the renewable resource sectors.

Sense of Direction and Ownership

The third and perhaps the most vital criterion is an adequate sense of direction; in other words, a vision for the future. Should civilization be guided by the logic of capitalism and market economy? Or should some wise leader guide us toward a fair and equitable society? Again, this criterion should also embody a kind of self-assessment meter that would signal when the directives become extreme. Citizens, facing the tsunami, felt helplessness as individuals – not because of their lack of capacity or resources – but because the decision-making process had long been handed over to the central authority, as exemplified by the fact that the Japanese have no voice in selecting the electricity company they wish to use, regardless of an individual's preference with respect to how his or her electricity should be generated. The 3/11 tsunami has certainly made it clear that local people need, once again, to reclaim ownership for themselves. But this ownership encompasses responsibility and the deliberation of the sustainability and appropriateness of the methods and results in question. While the choice to abandon nuclear power does bring the decision-making process closer to the common people, ordinary citizens will need to spend more time deliberating over the types of energy available and the costs they are willing to pay.

People make mistakes. Thus any social system must include the build-in possibility of learning and making adjustments. The scope of adjustment or even reversal is another name for resilience. The experience of 3/11 imposed numerous lessons, ranging in scope from the role of mutual help at the community level, energy supply at the national level, to the very definition of well-being at the humanitarian level. But it must also be remembered that major disasters of similar magnitude have also occurred in the past. The prime question for us is to determine what lessons we have learned from these occurrences, and how that knowledge has been applied in practice. Perhaps one of the major tasks of social and behavior sciences is not so much to deliver new theories or discover innovation but simply to remember lessons of the past and to devise mechanisms for their modification if and when the need arises. In this context, the role of social and behavioral sciences is to work with the people towards creating a vision of 'things otherwise', i.e., going beyond the current system's narrow emphasis on maximizing cost-benefit ratios, in order to question the very foundation of such a system.

Resilience from Memories

Unlike the people in Asia during the great Sumatra earthquake, many residents of the Japan's northeastern coastal line had had some knowledge of these periodic natural catastrophes. It could almost be speculated that people, faced with these natural uncertainties, would have developed some type of unique coping mechanism or resilience. This is exactly the point that Terada Torahiko raised 90 years ago in connection with the 1933 tsunami. By deploring the forgetfulness of people who had failed to apply the lessons from the great tsunami of 1896, he investigates human nature in the following way:

> A tsunami hit the Pacific coast of northeastern Japan. The waves swept away every small city and village on the coast and took countless human lives and valuables. Being a natural phenomenon that has occurred so repeatedly, one would think that the residents of the affected regions would have long ago thought of some measures to prepare for it, and that they would have been able to prevent the disaster from happening.… That this is not quite the case, it seems, is the very human natural phenomenon of the human world. Having suffered from the tsunami, people in the beginning may move to higher places; however, over the next 5, 10, 15 or 20 years, the population most likely will move back to the lower lands (close to the shore). Thus, another fateful day quietly approaches them.

Whatever 'lessons' contemporaries may highlight, many of these will be forgotten in the long run and what tends to remain is the public's overconfident faith in institutions and technology. It was only after the disaster that many of the ancestral stones that had been inscribed by past generations to warn people about tsunamis were 'uncovered' and their message appreciated. Perhaps the most foundational resilience

is not how we design communities or infrastructure, but how we continue to talk about past disasters in order to retain fragile memories. One of the most powerful sources of resilience is, in fact, people's sense of belonging to their hometown and their determination to restore their lives.

Humans are not sufficiently rational or far-sighted enough to be able to completely avoid all future disasters. Disasters arrive unexpectedly, but so do their countermeasures. While it may sound unwise to rely on intangible resources that are usually hidden and scattered, it is these seemingly unimportant, untapped potentials that can turn out to be useful. This potential, when gauged by the principle of maximizing economic value, tends to be dismissed as causal or slack. Thus, there is reason to maintain an 'alternative', diverse space where a different, autonomous logic operates. Freedom that allows diversity to function when it is called for is what we should appreciate. The praise for decentralization and self-governance ultimately boils down to their value in allowing some degree of autonomy to survive, and not because self-determination has been eliminated for the sake of economic efficiency alone. This lesson was born not only from the 3/11 tragedy, but also from Japan's overall development policy of the past 50 years, a pattern that exacerbated the effects of the 3/11 tsunami.

See also: Community Resilience to Disasters; Demographic Impact of Disasters; Disasters in Social, Cultural and Political Context; Hazards and Disaster Research in Contemporary Anthropology; Mass Trauma: Psychopathological Effects across the Life Span; Military and Disaster Psychiatry; Natural Disasters: Health-Related Aspects; Social Work and Disasters.

Bibliography

Ackerman, E., 1949. A balance sheet for Japan. The University of Chicago Magazine 42 (3), 5–8, 20.

Asahi Shinbun, 5 November 2012. We have to do it ourselves.

Akasaka, N., 2012. Thinking about the "Shape of a Nation" from 3.11: Restructuring the Tohoku Studies. Shincho Sensho (in Japanese).

General Headquarters, 1949. Japanese Natural Resources (Natural Resources Section, General Headquarters, Supreme Commander for the Allied Powers).

Foucault, M., 2009. Security, Territory, and Population: Lectures at the Collège de France. Palgrave Macmillan, London.

Hatakeyama, 17 November 2012. Public lecture delivered at the International House.

Holling, C.S., 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics 4, 1–23.

Kainuma, H., 2011. The "Fukushim" Thesis: How the Nuclear Village Was Born. Seidosha (in Japanese).

Kosaka, R. (Ed.), 2011. Resilience of Locals: Learning from the Memories of Mega Disaster (Chiiki No Resiliense: Daisaigai No Kioku Ni Manabu). Shimizu Kohbundo Shobo.

Ministry of Agriculture, Forestry and Fisheries, 2009.

Ministry of Public Affairs, 1947. State of Nutrition of the People. Ministry of Public Affairs, Tokyo.

Nihei, N., 2012. The framework of "in-between disasters". In: Akasaka, N., Oguma, E. (Eds.), Beginning from the Periphery: A Study of Tokyo/Tohoku. Akashi Shoten (in Japanese), Tokyo.

Oguma, E., 2011. The hidden face of disaster: 3.11, the historical structure and future of Japan's northeast. The Asia-Pacific Journal 9 (31/6).

Oh, T., 2013. Why do fishermen plan trees. In: Sato, J. (Ed.), Governance of Natural Resources: Uncovering the Social Purpose of Materials in Nature. United Nations University Press, Tokyo, pp. 202–221.

Sato, J., 2010. Civil society engagement in Japan. In: Cheema, S.G., Popovski, V. (Eds.), Engaging Civil Society: Emerging Trends in Democratic Governance. United Nations University Press, Tokyo, pp. 232–245.

Terada, T., 1933. Tsunami and Men (Tsunami to Ningen). Iwanamishoten, Tokyo.

Tokyo Foundation and Kyozon no Mori Network, 2012. In: 101 Narrative Recording of the Disaster Affected Sits. The Tokyo Foundation Kyozon no Mori Network.

United Nations International Strategy for Disaster Reduction (UNISDR), 2009. UNISDR Terminology of Disaster Risk Reduction. UNISDR, New York.

Walker, B., Holling, C.S., Carpenter, S.R., Kinzig, A., 2004. Resilience, adaptability and transformability in social ecological systems. Ecology and Society 9 (2), 5. http://www.ecologyandsociety.org/vol9/iss2/art5/.

Walzer, M., 1983. Spheres of Justice: A Defense of Pluralism and Equality. Basic Books, New York.

Yamauchi, A., 2011. Tokyo/Tohoku over "famine". In: Akasaka, N., Oguma, E. (Eds.), Beginning from the Periphery: A Study of Tokyo/Tohoku. Akashi Shoten (in Japanese), Tokyo.

Yamauchi, A., 2012. Tokyo/Tohoku over famine. In: Akasaka, N., Oguma, E. (Eds.), Beginning from a Periphery. Akashi Shoten, Tokyo.

Social Science and Civil Society

Eddy Hogg, School of Social Policy, Sociology and Social Research, The University of Kent, Canterbury, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives an overview of key social science approaches to civil society. It addresses in turn civil society organizations, volunteering, and philanthropy. For each, it looks at how they are defined, their scope, how social science approaches the study of them, and how social policy has sought to encourage or coordinate them.

Introduction

Civil society refers to the range of organizations and activities which occur outside of the traditional economic binary of market and state. They comprise the activities of and which occur within those organizations variously called charities, not-for-profit organizations, NGOs, civil society organizations, and many more. The term civil society though is hardly less contested. In the United Kingdom, the Coalition Government quickly renamed the New Labour-founded *Office for the Third Sector* the *Office for Civil Society* (Hogg and Baines, 2011). This rejection of the term 'Third Sector' is indicative of the way in which how we refer to the sector is laden with political dimensions and sectional interests (Ridley-Duff and Bull, 2011).

Indeed the notion of seeing civil society organizations (CSOs) as a collective group of civil society organizations is, Haugh and Kitson (2007) argue, a relatively recent idea; prior to the 1978 Wolfenden report, the wide range of civil society organizations in the United Kingdom were considered to be parts of the various sectors – education, sport, social care – in which they operated. The legacy of this is that huge heterogeneity exists in the organizations that are considered to make up civil society (Halfpenny and Reid, 2002). Any term used to define such diverse organizations with diverse goals and diverse means of achieving them runs the distinct risk of excluding some types of organizations and overemphasizing the role of others (Alcock, 2010). The inherent weakness of any catch-all term for the sector is what led Kendall and Knapp (1995: 67) to describe it as a "loose and baggy monster."

This article looks at the three dimensions of civil society to which social science has paid the most attention: the nature of the voluntary sector, the volunteering which occurs within it, and the philanthropic donations which fund the work it does. It addresses these three items in turn, outlining what they are, giving an overview of their scope and scale and looking at how social policy approaches the study of them. In doing so, this paper outlines the social sciences' approach to civil society. While the data analyzed in this article is mostly concerned with the United Kingdom, much of it reflects wider patterns and themes, and where large discrepancies exist between nations this is noted.

Civil Society Organizations

What Are Civil Society Organizations?

Among CSOs exists enormous heterogeneity in terms of the work they do and the way they do it (Hardill and Baines, 2011). Organizations identifying as CSOs can include charities, cooperatives, social enterprises, infrastructure groups, and others (Hardill and Baines, 2011). Turner (2001) and Milligan and Fyfe (2005) note that what unites CSOs is that they promote opportunities for social participation, for reciprocity, and for local democratic involvement for their volunteers and service users. Kendall and Knapp (1995) propose that there are five different functions which CSOs may seek to fulfill:

- *Service providers*: CSOs which exist to provide a direct service to individuals or groups, or to provide information or support.
- *Mutual aid*: CSOs which engage in self-help and exchange around a common need or interest.
- *Pressure group*: CSOs may be involved in policy advocacy or organizing campaigns, to put pressure on decision makers.
- *Helping individuals*: CSOs which offer help on a case by case basis.
- *Resource coordinating*: CSOs which act as infrastructure groups, providing support, training, and contacts to other civil society organizations.

CSOs may have one or more of these functions, and similarly volunteers may engage in one or more of these areas (Rochester et al., 2012).

While CSOs fulfill a range of important roles in society, the services and support that they provide has also been criticized and the limitations of their work recognized. Salamon (1987) reflects on the shift from civil society provision to a state-organized welfare system and notes that this in part stemmed from a recognition of the inherent failures of civil society organizations. Salamon notes four main weaknesses in CSOs:

- *Particularism*: Giving to organizations is voluntary, so provision is greater for causes that more people support. Therefore, help is not universal but reflects the personal tastes of those with the resources to provide them.
- *Insufficiency*: As the next section shows (Mohan, 2011; Clifford, 2012), the geographical spread of CSOs is uneven. There is no guarantee that provision will be evenly spread or given where most needed.
- *Paternalism*: It is suggested that civil society support can create dependencies, whereby the recipient is at the mercy of the organization and its supporters. This power relationship has been criticized by scholars of historical British charity (Finlayson, 1990; Davis Smith, 1995) who argued that civil

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41061-5

society was often motivated by a desire to shape the working classes in the image of their masters.

- *Amateurism*: CSOs have been accused of being run amateurishly, a lack of paid professionals (despite many organizations having a number of paid staff) meaning that they are badly run.

These realizations of the weaknesses of CSOs, some more valid than others, have led to some CSOs being part of a mixed economy of welfare in which they work alongside public and private organizations to deliver services (Hogg and Baines, 2011). Others exist beyond what the state would normally deliver, and beyond what private organizations would see a profit in perusing.

Where They Are and What They Do

In the United Kingdom, the National Council for Voluntary Organisations estimates that there are over 162 000 active CSOs (NCVO, 2013). The majority of these (52%) are small organizations with annual incomes of under £10 000, no paid staff, and only a small number of volunteers (NCVO, 2013). Very few (under 3%) of the CSOs in the United Kingdom have an annual income of over a £1 000 000, but these organizations account for over three-quarters (78%) of voluntary sector income. This is what Milligan and Fyfe (2005) describe as a 'bifurcation' of civil society, with larger, and more corporatist organizations increasingly working with the state – and with state funds – to deliver public services. At the other end of the scale 'grass-roots' organizations, smaller in size but very much in the majority, continue to work in their local communities quite removed from any state subsidy or interaction (Hogg and Baines, 2011).

One of the voluntary sector failures that Salamon (1987) identifies is that the provision of CSOs is unevenly spatially distributed, and that as such CSOs cannot be relied on to cope with all of the social problems present in modern industrial societies. Certainly, the geographical distribution of formally constituted CSOs in the United Kingdom is uneven. There are over 2.5 times as many formal CSOs working at the neighborhood scale in the most prosperous areas of the United Kingdom compared within the most disadvantaged areas (Mohan, 2011; Clifford, 2012). This pattern is particularly the case for those CSOs which support cultural and leisure pursuits, which provide education and lifelong learning and which engage in general community activities. The relationship reverses – with more organizations in more deprived areas, for those CSOs which support economic well-being, training, and community cohesion (Clifford, 2012). CSOs in more deprived areas are more likely to be larger and more likely to be in receipt of public funding, while those in less deprived areas are likely to be smaller and to raise funds through philanthropic donations (Mohan, 2011; Clifford, 2012). There is also a rural–urban split in terms of the spread of CSOs in the United Kingdom, with rural areas having over 50% more CSOs than urban areas per 1000 people (NCVO, 2013).

CSOs also engage in a number of areas of work, pursuing a range of aims. The 2012 National Survey of Charities and Social Enterprises (Cabinet Office, 2010) found that the most common areas for organizations to work in the United Kingdom were education and lifelong learning (28% of organizations listing this as their main area of work), leisure (21%), health and well-being (17%), community development and mutual aid (17%), culture, arts, and music (13%), religious- and faith-based activities (13%), and in training (10%). CSOs differ not just in terms of what they do, but also in terms of the number of paid staff and volunteers they have. The majority of CSOs have no paid members of staff at all – they are run entirely by volunteer efforts – while only 10% of CSOs have more than 11 paid members of staff (Cabinet Office, 2010). From these figures, we can observe that 68% of CSOs in England in 2010 had fewer than 20 volunteers, while just 4% had more than 101 volunteers.

Social Policy and Civil Society Organizations

CSOs exist within a wider context of relations between the public and private sectors. In the United Kingdom, prior to the 1911 National Insurance Act, CSOs were for the most part wholly independent of the state, delivering services in fields in which the government made no provision (Hogg and Baines, 2011). Between the wars, CSOs and state institutions increasingly worked in partnership, with the National Council of Voluntary Organisations (NCVO) formed in 1919 (as the National Council of Social Services) with the aim of coordinating CSOs and developing their ties with the state (Davis Smith, 1995). The 1941 Beveridge Report proposed the creation of National Insurance and the National Health Service, but it also saw a key role for CSOs in healthcare and social welfare provision (Davis Smith, 1995; Alcock, 2010).

In the 1980s, governments on both sides of the Atlantic adopted the principles of reducing the role of the state and strengthening the role of the markets (Hogg and Baines, 2011). CSOs were central to this, with the contracting of work previously seen as the state's domain to CSOs, seeking to encourage, "innovative and entrepreneurial forms of service delivery" (Curtis, 2008: 227). In the 1990s, as the state contracted further still, CSOs became even more significant in the delivery of public services (Buckingham, 2009). Kendall (2000), though, fears that this account overplays CSOs' role during this period. His stance is consistent with the analysis that CSOs were more often safety net (Halfpenny and Reid, 2002; Haugh and Kitson, 2007).

In more recent years, CSOs have been mainstreamed in British policy making, though New Labour's Third Way and the Coalition Government's Big Society initiatives (Hogg and Baines, 2011). It should also be noted that while many CSOs have seen their relations with the state change significantly over time, others have continued to exist well away from the reach of the state.

Volunteering

What Is Volunteering?

Definitions of volunteering vary depending on the particular focus of a study, but there is broad consensus as to what a broad definition of volunteering might look like; tasks undertaken are unpaid (allowing for the payment of expenses), they aim to benefit individuals, groups, and communities other than, or in addition to, close relatives, and they are undertaken

through individuals' own free will (Hardill and Baines, 2011; Rochester et al., 2012). Within this broad framework, formal volunteering entails all of this diverse activity which occurs within an organizational setting. This includes volunteering done through groups, clubs, or organizations in the public, private, or voluntary sectors (Milligan and Fyfe, 2005; Rochester et al., 2012). It is distinct from informal volunteering, which involves volunteer help given to an individual, family, or informal group by an individual, family, or informal group, often on an ad hoc basis (Milligan and Fyfe, 2005; Rochester et al., 2012).

Who Volunteers?

A wide range of data are gathered regularly about who volunteers and what they do, across the globe. However, as Rochester et al. (2012) argue, the questions a data gatherer asks and the methodology they use all influence the results obtained and as such comparative work must be undertaken with caution. The Charities Aid Foundation's World Giving Index 2013 attempts to use the same methodology across the globe, and it cites Turkmenistan as the nation with the most regular volunteers, with 57% of citizens volunteering in a typical month. Sri Lanka is second with 46% volunteering each month, with the United States (45%), Myanmar and Philippines (both 43%), and Canada (42%) close behind.

In terms of who volunteers and what they do, robust cross-national data is in short supply, and the different methodological approaches in different national contexts (Rochester et al., 2012) makes comparing between different countries troublesome. In the United Kingdom, women consistently volunteer at higher rates than men, a pattern which a 1995 survey in Europe (Gaskin and Davis Smith, 1995) showed to also be the case in Belgium. However, in Denmark, Germany, Ireland, the Netherlands, and Sweden, men volunteered at a higher rate than women (Gaskin and Davis Smith, 1995).

In terms of what volunteers do, data from the United Kingdom suggest that volunteers are engaged across an extremely wide range of fields. Sport and exercise, hobby and social clubs, religious organizations, education organizations (including schools), health charities, and many more are all popular causes for formal volunteers (Cabinet Office, 2013). Across the wide range of CSOs, volunteers undertake a wide range of roles, from giving professional advice on a *pro bono* basis to undertaking manual work such as digging or cleaning (Low et al., 2007).

Why Do People Volunteer?

Much has been written about impulses to volunteer, much of it from psychology literature, and often using closed questions and quantitative methods to 'measure' reasons for engagement. Within social sciences, the realization that every volunteer's motivations are unique lead Rochester et al. (2012) to state that there are as many motivations as there are volunteers. A further problem is the gap between what volunteers say about what motivates them when asked and what they themselves think (Rochester et al., 2012). Nonetheless, there are some useful frameworks from social science which help us to understand why people volunteer.

Omoto and Snyder (2002) state that there are three stages of volunteering: antecedents, experience, and consequences, each of which exists at individual, organization, and community levels. Here, we are interested in the individual level, at which the three stages are as follows (Omoto and Snyder, 2002):

- *Antecedents*: This is what comes before decisions to volunteer and is formed of the prior experiences and events which shape an individual's pathway toward volunteering. Rooted in social experiences, antecedents include life events, available resources, identity, social expectations, and surrounding social support.
- *Experiences*: This refers to how volunteers experience their roles, and the satisfaction they derive from their volunteering. It is shaped by their relationship with the CSO they volunteer with, how they are supported by paid staff, and how they engage with other volunteers. From these experiences, volunteers derive differing levels of satisfaction.
- *Consequences*: These are the outcomes from volunteering, in terms of changes in what people think, know, and do. It concerns how volunteering affects self-identity, and how this helps build commitment to volunteering and to the CSO volunteered for. Consequences feed back into future experiences.

This framework embeds volunteers and their efforts within a community context, which influences the volunteer and the work they do (Omoto and Snyder, 2002). Musick and Wilson (2008) also place volunteering with a context of social relations, emphasizing the importance of being asked in decisions to volunteer.

Another way of understanding motivations to volunteer lies in the distinction between *philanthropy* and *mutual aid*. Philanthropic voluntary activities are said to be those, "where an individual or group offers services to others" (Finlayson, 1990: 184). Mutual aid concerns voluntary activity in which individuals or groups undertake volunteering for mutually beneficial reasons. This reciprocity need not be immediate or explicitly agreed, and may indeed be deferred over long periods, but it is integral to mutual volunteerism that there is an obligation and an expectation for reciprocation (Finlayson, 1990).

This binary between self-interested mutual aid and altruistically motivated philanthropy is rejected by Hardill et al. (2007), who argue that volunteering needs to be understood in the context of the individual's life, to move away from this divide. They propose instead four groups of volunteer impulse:

- *Giving alms*: These are volunteers who have, "… identified an unmet need and they want to make a difference" (Hardill et al., 2007: 405). Often these volunteers see themselves as helping people less fortunate than themselves. Such volunteering is, "… necessarily both self- and other-rewarding" (Hardill et al., 2007: 405).
- *Giving to each other*: These are volunteers who seek to help others with whom they perceive a shared issue, as a response to a problem they experience together.
- *Getting on*: These volunteers are seeking to develop "…skills and experience of value in labour market" (Hardill et al., 2007: 406) through undertaking voluntary work.

- *Getting by*: These are volunteers who, "… entered volunteering as a response to a milestone life event, and that it fills something missing or an emotional gap in life" (Hardill et al., 2007: 207).

Understanding philanthropy or mutual aid in isolation is therefore a fallacy; there is inevitable overlap between them and each volunteering decision draws upon a complex mix of these and other impulses, internal and external.

Social Policy and Volunteering

Volunteering exists within a wider policy context, and a number of public policy decisions have sought to increase engagement in volunteering. Rochester et al. (2012) note that it was under John Major's Conservative Government (1992–97) that public policy on volunteering first became joined up across government departments. Following this, New Labour engaged in a wide range of events, initiatives, and programs with the aim of encouraging volunteering. The Coalition Government also have a cross-government commitment to volunteering, in the form of the Big Society. Their Giving White Paper (Cabinet Office, 2010), which looks at the giving of both time and money – view volunteering as a form of giving. The Coalition's approach to volunteering and CSOs is far more 'soft-touch' than New Labour's, and as such direct programs or initiatives are far scarcer (Rochester et al., 2012).

Philanthropy

What Is Philanthropy?

Philanthropy, in its broadest sense, includes all voluntary action which is undertaken for the public good (Payton, 1988). Within this definition is both the giving of time, explored in the previous section, but more commonly the term is used to refer to the giving of money, which is explored here (Payton and Moody, 2008). Daly (2012) argues that the meaning of philanthropy has changed over time, and differs between different contexts. In arguing that philanthropy is an 'essentially contested concept,' Daly (2012) draws on Payton (1987) and Van Til (1990) to argue that the shear range of ways in which philanthropy is defined means that when we debate it, we differ in not just the meaning we attribute to it, but also in terms of the values and purposes we see it having.

When viewed through economic theory, the philanthropic giving of money to charity is illogical in extreme – homo economicus sees no utility maximization arising from giving his money away. So why do so many people give? Homo sociologicus, on the other hand, is a far more complex character, and as with sociological approaches to volunteering, when it comes to the study of the giving of money it is the social context which is key (Halfpenny, 1999). Andreoni (1990, 2006) suggests the 'warm glow theory' as a way of understanding why people freely give their resources to others – they do so in exchange for a warm feeling of having done something good. This theme of doing something good is echoed by Schervish's (1998) assertion that philanthropy as a social relation revolves around the moral value of *caritas*, of a love for others and their needs.

Social science approaches to philanthropy consider both the impact upon wider society and the impact upon the donor themselves, a tension which Frumpkin (2006) sees as involving an awkward balance between private interests and public needs. Veblen (1899) saw the act of giving as a conspicuous way in which donors show how much money they have, while Bourdieu (1984) conceptualized charitable support for artistic and cultural activities as a way in which donors can show off their cultural capital and exceptional taste. As such, Breeze (2012a) argues that acts of philanthropy are both an attempt by donors to shape and change the world around them, and a way in which those who give can change their own self-image and the way in which they are perceived (or hope they are perceived) by others.

Who Gives and To What Do They Give?

The practice of giving money to CSOs to spend on charitable causes is widespread across the world. In 24 countries, over half of the population give to charity in an average month, with Myanmar leading the way where 85% of citizens give each month (CAF, 2013). The United Kingdom comes in second globally, with 76% giving, while in the United States, Canada, The Netherlands, Australia, and New Zealand over 60% give in an average month. While all countries have some citizens who give regularly, it is not so widespread in a number of places – in 51 countries, the proportion giving is less than 20%, while in Georgia it is just 3% (CAF, 2013).

Data from both the United States (Andreoni, 2006) and the United Kingdom (NCVO/CAF, 2013) shows that people tend to give more with age, both in absolute terms and as a fraction of their income. The impact of gender is less marked, with Bekkers and Wiepking (2007) finding no significant differences between men and women's giving in the literature. Research from both sides of the Atlantic finds that those with more education tend to give more often and larger amounts than those with lower education (Andreoni, 2006; Low et al., 2007; Cabinet Office, 2013). In terms of the link between religion and giving of money, Bekkers and Wiepking (2007) find that almost every report on the link between religious attendance and philanthropic giving finds that those who attend regularly give more regularly and more often. Family impacts too – nearly every study which looks at the impact of marriage and having children finds that these increase the likelihood of giving and the amount given (Bekkers and Wiepking, 2007).

In terms of the causes people give to, religious causes dominate in the United States (Giving USA, 2013), while in the United Kingdom medical research and hospitals, children, and animals dominate year after year (NCVO/CAF, 2013). Differences in categories between the United States, the United Kingdom and other national contexts make comparisons difficult, but broadly speaking donors as a whole tend to support medical charities, children and animal charities, while major donors are more likely to support higher education institutions, arts and cultural organizations, and medical research (Breeze, 2013a).

Why Do People Give?

In their review of the literature on charitable giving, Bekkers and Wiepking (2007) identify eight mechanisms which have been identified as determining why individuals give to charity:

1. *Awareness of need*: Before they give, people have to become aware of a need for money to be given to that cause. This awareness can come about in a range of ways, including personal experience, charity advertising, news stories, or other ways.
2. *Solicitation*: A large proportion of donations are made because they are asked for – Bryant et al. (2003) find that 85% of donations were given following a request being made, while Bekkers (2005) finds that 86% of donations were made following a request.
3. *Costs and Benefits*: Giving costs money, and when this is made cheaper (e.g., through the Gift Aid tax program in the United Kingdom), then people are more inclined to give. Donating can also have tangible benefits, such as VIP access to events or areas not accessible to nondonors.
4. *Altruism*: People give to charity because they care about the work being done.
5. *Reputation*: This links back to Veblen and Bourdieu's ideas that giving can be used to improve an individual's social standing. Giving is seen as a good thing to do (mostly, although media portrayals in the United Kingdom would sometimes make you question this) and as such giving can help to improve a person's reputation among their peers.
6. *Psychological benefits*: Beyond the social benefits of giving, there are also psychological benefits, including boosts to self-image, producing positive moods, alleviating feelings of guilt, being a way of showing gratitude, and simply making one feel as though one is a good person.
7. *Values*: Donors give to organizations which reflect their own values, in the hope that they can make the world – or just the bit of the world they care about – a better place.
8. *Efficacy*: Donors want to feel as though they are making a difference to the cause(s) they support, and a feeling that their donation can make a difference is significant to giving decisions.

It should be stressed that for each act of giving, a range of different factors from across a number of these mechanisms may be in play, and that these will change for individuals over time.

Beyond looking at why individuals give to charity, Breeze (2013b) looks at how donors chose which charities to donate to. When deciding where to direct a charitable gift, donors base their decisions on four key areas:

- *Tastes*: Donors' own personal tastes are the most significant factor guiding where they direct their donations. As with the awareness of need and altruism identified by Bekkers and Wiepking (2007), Breeze (2013b) finds that donors support causes close to their hearts.
- *Experiences*: Each donor has their own personal history of experiences, on which they draw when making decisions where to donate. These may be based on childhood experiences, being touched by an illness or some other life event, and they change over the life course.
- *Perceptions of charity competence*: People want to give to charities which they perceive as competent, and where they can be confident that their donations will be used effectively.
- *Desire to have a personal impact*: Donors want to feel that their donation is making a difference to the work the charity does, hence often therefore explain what a donation of £50 or $100 can provide.

Again, for each donor the way in which they make these choices and negotiate these decisions is different, but Breeze's (2013b) heuristic offers a useful method of understanding how donors choose where to give.

Political Support for Philanthropy

Political support for philanthropy has in the United Kingdom been a feature of policy making since the 1980s, a period during which huge increases in private wealth occurred due to global capitalism, while simultaneously, as we have seen previously, the frontiers of the state were being rolled back (Pharoah, 2013). During this period, measures were introduced to stimulate giving, most significantly the Gift Aid tax relief scheme in 1990, enhanced in 2001 (Pharoah, 2013). The Coalition Government's support for giving continued, particular around their Big Society policy framework. However, governmental approaches to philanthropy are often, Breeze (2012b) argues, simplistic and underestimate the complexity of philanthropic decision making, as outlined earlier in this section. Instead, philanthropy is simplistically presented as a 'good thing,' the practicing of which will inevitably lead to stronger communities and better services. The qualities that philanthropy has, and the inherent weaknesses of civil society noted by Salamon (1987) mean that this easy acceptance of philanthropy as a good thing is potentially unhelpful.

Civil Society and Social Science

Historically and to the present day, CSOs and civil society actors have played crucial roles in funding and supporting social science research. Much of the initial growth of universities and research institutions, both secular and faith based, was organized and funded by civil society groups (Curti and Nash, 1965) and universities globally continue to receive significant philanthropic funding from alumni, foundations, and major donors (Breeze et al., 2011). Indeed, in the annual Million Dollar Donors (USA) and Million Pound Donors (UK) surveys, higher education institutions consistently emerge as the most common destination of major philanthropic donations (Breeze, 2013a; Coutts, 2013). Civil Society foundations and charitable organizations also fund social science in a number of other ways. Large funding bodies such as the Leverhulme Trust invest heavily in social science research, while charities invest in PhD researchers and in research projects and programs (Macmillan and Scott, 2003).

Civil Society has also often taken a lead in identifying areas for social science to explore and investigate. This is demonstrated by the pioneering campaign work of philanthropist criminal justice campaigner John Howard (1726–1790) and of the organization that takes his name, The Howard League for Penal Reform, formed in 1866, is the oldest penal reform organization in the world and played a significant role in the formation of the modern social science discipline of criminology (Morris, 1988). The experience that CSOs and those who work with them have of social issues feeds into social science research, a symbiotic relationship where research and civil society provision and campaigning mutually support each other.

Conclusion

This article has given an overview of key social science approaches to civil society. It has addressed in turn CSOs, volunteering, and philanthropy. For each, it has looked at how they are defined, their scope, how social science approaches the study of them, and how social policy has sought to encourage or coordinate them.

See also: Charity and Philanthropy, Economics of; Civil Society and the Public Sphere: History of the Concept; Civil Society, Concept and History of; Nongovernmental Organizations: Anthropological and Historical Aspects; Nonprofit Sector, Geography of; Voluntary Associations, Sociology of; Volunteerism, Psychology of.

Bibliography

Alcock, P., 2010. Big Society or Civil Society? a New Policy Environment for the Third Sector. http://www.tsrc.ac.uk/LinkClick.aspx?fileticket=PwhvBXnPGAU%3D&tabid=716 (accessed 03.03.13.).

Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. Economic Journal 100 (401), 464–477.

Andreoni, J., 2006. Philanthropy. University of Wisconsin, Madison, WI.

Bekkers, R., 2005. It's Not All in the Ask. Effects and Effectiveness of Recruitment Strategies Used by Nonprofits in The Netherlands. 34rd Annual ARNOVA Conference, Washington, DC, USA.

Bekkers, R., Wiepking, P., 2007. Generosity and Philanthropy: A Literature Review. Vrije University, Amsterdam.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgement of Bad Taste. Harvard University Press, Harvard.

Breeze, B., Wilkinson, I., Gouwenberg, B., Schuyt, T., 2011. Giving in Evidence: Fundraising from Philanthropy in European Universities. European Commission, Brussels.

Breeze, B., 2012a. Philanthropy. In: Southerton, D. (Ed.), Encyclopaedia of Consumer Culture. Routledge, London, pp. 980–983.

Breeze, B., 2012b. Donor and Governmental Perceptions of Philanthropy, in Philanthropy and a Better Society. Centre for Giving and Philanthropy, London, pp. 62–67.

Breeze, B., 2013a. Million Pound Donors Report. Coutts, London.

Breeze, B., 2013b. How donors choose charities: the role of personal taste and experiences in giving decisions. Voluntary Sector Review 4 (2), 165–183.

Bryant, W., Slaughter, H., Kang, H., Tax, A., 2003. Participating in philanthropic activities: donating money and time. Journal of Consumer Policy 26, 43–73.

Buckingham, H., 2009. Competition and contracts in the voluntary sector: exploring the implications for homelessness service providers in Southampton. Policy and Politics 37 (2), 235–254.

Cabinet Office, 2010. National Survey of Charities and Social Enterprises. Cabinet Office, London.

Cabinet Office, 2013. Giving of Time and Money: Findings from the 2012–13 Community Life Survey. Cabinet Office, London.

CAF, 2013. World Giving Index 2013. CAF, London.

Clifford, D., 2012. Voluntary sector organisations working at the neighbourhood level in England: patterns by local area deprivation. Environment and Planning A 44 (5), 1148–1164.

Coutts, 2013. The Million Dollar Donors Report. Coutts, London.

Curti, M., Nash, R., 1965. Philanthropy in the Shaping of American Higher Education. Rutgers University Press, New Brunswick.

Curtis, T., 2008. Finding the grit that makes a pearl. International Journal of Entrepreneurial Behaviour and Research 14 (5), 276–290.

Daly, S., 2012. Philanthropy as an essentially contested concept. Voluntas 23, 535–557.

Davis Smith, J., 1995. The voluntary tradition: philanthropy and self-help in Britain 1500–1945. In: Davis Smith, J., Rochester, C., Hedley, R. (Eds.), An Introduction to the Voluntary Sector. Routledge, London, pp. 9–39.

Finlayson, G., 1990. A moving frontier: voluntarism and the state in British social welfare 1911–1949. Twentieth Century British History 1 (2), 183–206.

Frumpkin, P., 2006. Strategic Giving: The Art and Science of Philanthropy. University of Chicago Press, Chicago.

Gaskin, K., Davis Smith, J., 1995. A New Civic Europe? A Study of the Extent and Role of Volunteering. The Volunteer Centre, Berkhamstead.

Giving USA, 2013. Giving USA 2013: The Annual Report on Philanthropy for the Year 2012. Giving USA, Indiana.

Halfpenny, P., 1999. Economic and sociological theories of individual charitable giving: complementary or contradictory? Voluntas 10 (3), 197–216.

Halfpenny, P., Reid, M., 2002. Research on the voluntary sector: an overview. Policy and Politics 30 (4), 533–550.

Hardill, I., Baines, S., P, 6., 2007. Volunteering for all? Explaining patterns of volunteering and identifying strategies to promote it. Policy and Politics 35 (3), 395–412.

Hardill, I., Baines, S., 2011. Enterprising Care: Unpaid Voluntary Action in the 21st Century. Policy Press, Bristol.

Haugh, H., Kitson, M., 2007. The third way and the third sector: new labour's economic policy and the social economy. Cambridge Journal of Economics 31 (6), 973–994.

Hogg, E., Baines, S., 2011. Changing responsibilities and roles of the voluntary and community sector in the welfare mix: a review. Social Policy and Society 10 (3), 341–352.

Kendall, J., 2000. The mainstreaming of the third sector into public policy in England in the late 1990s: whys and wherefores. Policy and Politics 8 (4), 541–562.

Kendall, J., Knapp, M., 1995. A loose and baggy monster: boundaries, definitions and typologies. In: Davis Smith, J., Rochester, C., Hedley, R. (Eds.), An Introduction to the Voluntary Sector. Routledge, London, pp. 66–95.

Low, N., Butt, S., Ellis Paine, A., Davis Smith, J., 2007. Helping Out: A National Survey of Volunteering and Charitable Giving. Cabinet Office, London.

Milligan, C., Fyfe, N., 2005. Preserving space for volunteers: exploring the links between voluntary welfare organisations, volunteering and citizenship. Urban Studies 42 (3), 417–433.

Mohan, J., 2011. Mapping the Big Society: Perspectives from the Third Sector Research Centre. TSRC Working Paper 62. http://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/working-papers/working-paper-62.pdf (accessed 03.03.14.).

Macmillan, R., Scott, A., 2003. On the case? Dilemmas of collaborative research. Area 35 (1), 101–105.

Morris, T., 1988. British criminology: 1935–48. British Journal of Criminology 28 (2), 20–44.

Musick, M., Wilson, J., 2008. Volunteers: A Social Profile. Indiana University Press, Indiana.

NCVO, 2013. UK Civil Society Almanac. NCVO, London.

NCVO/CAF, 2013. UK Giving in 2013. NCVO, London.

Omoto, A., Snyder, M., 2002. Considerations of community: the context and process of volunteerism. American Behavioural Scientist 45 (5), 846–867.

Payton, R., 1987. Philanthropy as a Concept. www.paytonpapers.org (accessed 03.03.14.).

Payton, R., 1988. Philanthropy: Voluntary Action for the Public Good. Macmillan, New York.

Payton, R., Moody, M., 2008. Understanding Philanthropy: Its Meaning and Mission. Indiana University Press, Indiana.

Pharoah, C., 2013. Private giving and philanthropy – their place in the big society. People, Place and Policy 5 (2), 65–75.
Ridley-Duff, R., Bull, M., 2011. Understanding Social Enterprise: Theory and Practice. Sage, London.
Rochester, C., Ellis Paine, A., Howlett, S., 2012. Volunteering and Society in the 21st Century. Palgrave Macmillan, Basingstoke.
Salamon, L., 1987. Of market failure, voluntary failure, and third-party government: toward a theory of government-nonprofit relations in the modern welfare state. Nonprofit and Voluntary Sector Quarterly 16 (1), 29–49.
Schervish, P., 1998. Philanthropy. In: Wuthnow, R. (Ed.), Encyclopedia of Politics and Religion. Congressional Quarterly, Washington, DC, pp. 600–603.
Turner, B., 2001. The erosion of citizenship. British Journal of Sociology 52 (2), 189–210.
Van Til, J., 1990. Defining philanthropy. In: Van Til, J. (Ed.), Critical Issues in American Philanthropy. Jossey Bass, San Francisco, pp. 10–25.
Veblen, T., 1899. The Theory of the Leisure Class. Macmillan, New York.

Relevant Websites

https://www.cafonline.org/ – Charities Aid Foundation.
http://www.philanthropy.iupui.edu/ – Indiana University School of Philanthropy.
http://data.ncvo.org.uk/ – NCVO Data Archive.
http://www.birmingham.ac.uk/generic/tsrc/index.aspx – Third Sector Research Centre.

Social Science and the Media

Andrea Römmele, Hertie School of Governance, Friedrichstrasse, Berlin, Germany
Barbara Pfetsch, Communication Theory and Media Effects, Freie Universität Berlin, Garystrasse, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Media is an integral part of the contemporary social world. As an object and subject of social science disciplines, media and mass media are defined, theorized, and approached in a multitude of ways depending on the discipline and question guiding the research. Mass media has a long history whose trajectory has made it a major factor in influencing and shaping society. The shape of media, and thus of society, is constituted by and reconstitutes the social and political contexts of any given temporal or social moment as it relates to the social systems and media infrastructure coinciding with it.

Social Sciences and the Media

Media as Institutions

Mass media refers to institutions that use technology to address large numbers of people. It is characterized by a one-way communication system between a single entity and a larger audience. Spoken or visual communication limits audiences to those physically present to see or hear it, making it difficult to reach a mass audience. The innovation of written language allowed communications to be kept in their original form, as well as to reach people who were not physically present at the time of their inception. The invention of the printing press allowed for the efficient (re)production of written content, making it possible for mass audiences to access written media. Media diffusion, although aided by the printing press, remained dependent on the spread of literacy.

The ability to print and produce media in large quantities using a printing press is commonly accepted as the first revolution of the mass media, but there have been further revolutions: the invention of radio, TV, and Internet Communications Technologies, also referred to as new media or digital communication. In contrast to theories of creative destruction (as suggested, e.g., by Joseph Schumpeter), the revolutions in mass media have added technologies to the mass media toolkit, rather than replace aging ones. Books were not abandoned with the dawn of radio, for example, nor has radio disappeared in the face of television, the digital revolution, and the Internet.

Upon examination of the waves of media innovations, it becomes clear that the addition of new revolutionary technologies is becoming more and more frequent; in other words, the development of technology over time is telescoping. The 1460s saw the spread of the use of the printing press across Europe; the mid-1800s witnessed the telegraph become widely used; film was becoming more accessible by the beginning of the 1900s, followed by radio gaining popularity in the 1920s. Television rose to prominence after WWII in the 1950s, and the 1970s saw the dawn of the personal computer. By the 1990s, people were surfing the World Wide Web in large numbers, but they did not stop reading books, as the success of online book vendors like Amazon demonstrates. However, the successful penetration of digital technology has changed the character of media fundamentally, as one-way communication from sender to audience was replaced by interactivity as the general mode of transmission.

Media as Infrastructure

Many theories and speculations have come along with the various technological revolutions of media. Marshall McLuhan (1911–1980), a prominent media theorist, argued that the medium of communication strongly influences the messages communicated through it. McLuhan's ideas represent a concept called technological determinism: a way of thinking that emphasizes the influence of technology on society's communication, rather than social processes. Foundations for technological determinism can be found in historical materialism, a theory initially developed by Karl Marx which posits that material conditions, including technology, are the primary influence on the organization of societies. Parallels to McLuhan's thinking are evident in Althusser's notions that discourse, and thus any expression of ideas through the use of language in media is constituted by the context in which it originates (Jorgensen and Phillips, 2002). A contrast to this way of thinking can be seen in Laclau and Mouffe's discourse theory, which asserts that language use (i.e., by the media) shapes our understanding of the world around us. Thus, the struggle to achieve power is the struggle to attach meaning and remove unwanted ambiguity from language (see Laclau and Mouffe, 1985). (For more on technological determinism and politics, see Hoff et al. (2000); Ong, 1982; or the work of Harold Innis or Marshall McLuhan.)

"Whatever we know about our society, or indeed about the world in which we live, we know through the mass media" (Luhmann, 2000: 1). This is the opening sentence of Niklas Luhmann's book *The Reality of the Mass Media*. If we accept Luhmann's idea, even in part, we accept the powerful role the media plays in shaping human society. This power intrigues social scientists and has led to large amounts of social scientific literature exploring the infrastructure of media as a tool for political, social, and economic power.

Media as Systems

Media and Power

Theories on the role of the media in society have a tendency to couple power and media, whether directly or indirectly. Michel Foucault (1980) links power and knowledge, asserting that power "forms knowledge [and] produces discourse." (119) Knowledge and discourse, if we refer back to Laclau and

Mouffe, or Luhmann, are inseparable from power and politics. Jürgen Habermas (1992) asserts that as media technologies grow and multiply, so does their propensity for being wielded as instruments of political power. "The public sphere, simultaneously pre-structured and dominated by the mass media, developed into an arena infiltrated by power in which, by means of topic selection and topical contributions, a battle is fought not only over influence, but over the control of communication flows that affect behavior while their strategic intentions are kept hidden as much as possible" (Habermas, 1992: 437). Pierre Bordieu and Habermas agree that media have experienced convergence with the economy, marketizing the arena of political power within media practices. The power of the media is thus constituted in the ability to determine views of the world through topic selection, or, as Entman (1993, 2007) refers to it, framing. Framing happens when media "select some aspects of a perceived reality and make them more salient … in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (Entman, 1993: 52), thereby promoting a certain interpretation. Media frames serve to set a political agenda, determining which problems become salient for the public and receive political attention (Entman, 2007: 164).

Media and the Modern State

With mass media securely bundled to politics and power (Mancini and Swanson, 1996: 11), social scientists have turned to examining the role the media plays in politics of the modern state. Hallin and Mancini, in their seminal work *Comparing Media Systems* (2004), assert that it is futile to try to understand the political system of a state without understanding the media, and vice versa. Voltmer (2012) notes that while the media and political systems are irreversibly intertwined, the "political system … ultimately has the power to make binding decisions and thus shapes the basic structure and functioning of the media system" (239). Media institutions are further linked to the state through the idea that nationhood and national identities are critical definers of the purviews that media institutions inhabit (Sparks, 2005: 37). Further work similarly emphasizes the crucial role media plays in any changes in political institutions (Coyne and Leeson, 2009; Schudson, 2002).

Media Systems and Contexts

Media is not an uniform idea, product, or set of actors. The role, output, goals, and impact of the media differ depending on their context. Hallin and Mancini (2004) undertook a major comparative study of media systems in the western world, creating three broad categories for an initial categorization. Although no country fits perfectly with any of the following models, Hallin and Mancini (2004) have provided the necessary guidelines and framework for considering the classification of media systems.

Important for appraisal of a media system are four dimensions:

- Media markets
- Political parallelism
- Journalistic professionalism
- State intervention.

These four dimensions "can be regarded as universally applicable tools for describing the dynamics of media and politics in different contexts: All media organizations have to generate revenues to produce their outputs; decide on whether or not to engage in alliances with social groups and, if so, which ones; internally agree on the rules of operation and the norms that set the standards for evaluating their outputs; and finally position themselves vis-à-vis the actors and institutions of political power" (Voltmer, 2012: 227).

Hallin and Mancini's (2004) use these four dimensions to categorize western media systems as follows: The Polarized Pluralist model is found predominantly in European Mediterranean states, and characterized by an elite orientation, significant parallelism between media outlets and political ideologies, and substantial state involvement in media. This contrasts with the horizontal orientation, corporatist alignment, and journalistic autonomy and professionalism of the Democratic Corporatist model, common in Northern and Central Europe. The third model is the Liberal model, predominant in the United States. This model is described as having a low degree of political and social parallelism, a media independent of state financing and regulation but subject to commercial pressures, and high journalistic professionalism. The Liberal model, often viewed as an information service, is commonly perceived as the dominant model of media system as globalizing tendencies proceed.

The study of Hallin and Mancini (2004) was criticized since their typology of media systems only accounts for Western Europe and North America. The authors reacted by editing a volume in which their categories are used to explore media systems: "Media Systems: Beyond the Western World" (Hallin and Mancini, 2012). This study demonstrates that in order to describe media systems in other regions of the world additional categories and models need to be included.

Media in Different Systems: Democracy, Autocracy, Dictatorship

The typology of media systems using Hallin and Mancini's framework is limited to national contexts which display the minimum tenets for democracy and is also rooted in western thought. "Like the Habermasian public, the Hallin-Mancini framework is predicated on bourgeois rationality and on assumptions and values that are not dominant within what might be an African model" (Hadland, 2012: 116). This is likely true to a greater or lesser extent for the rest of the world. (For more on specific critiques of Hallin and Mancini's model, see Hallin and Mancini, 2012.)

Further classifications of media types and systems have been published which apply to other regions; these also emphasize the connection between media and political systems. Ayish (2002), careful to describe his classifications as patterns rather than models to underline their dynamism (p. 152), identifies three categories for Arab Television media:

- Traditional Government Controlled
- Reformist Government Controlled
- Liberal Commercial.

Nisbet and Moehler (2005) similarly create categories for Sub-Saharan Africa, namely the following:

- Open Democratic
- Liberalized Democratic
- Liberalized Autocratic
- Repressed Autocratic
- Closed Autocratic.

These few examples (as well as others, see: Rugh, 1979; Curran and Park, 2000) draw upon some similar guidelines to those of Hallin and Mancini (2004), but also focus on other important dimensions such as the reproduction of the status quo, or support for change, which again stresses the idea that the media commonly functions as arena of power struggle.

Technological Change

Technological development very often brings about changes in media (options) and in the way people communicate. The more options (in terms of media) people have for communication, the greater the choice of a medium to fit the message. After all, politics is communication; messages will have different effects depending on the medium through which they are transferred, which means that with every new choice of medium, the complexity of communication and thus the complexity of the political arena increases.

In the past 50 years, two major technological developments have dramatically changed the way we communicate. The invention of television was an event that revolutionized the media landscape and impacted the way in which power struggles have taken place since. While literacy was a required prerequisite for consumption of media associated with the development of the printing press, television required only infrastructure (a television set, electricity, and a broadcaster) to reach mass audiences. In contexts where the infrastructure for television was present, the medium was adopted rapidly: In 1949, only 2.3% of US households had a television; by 1960, 87% of households had one (Proir, 2007: 59).

Television, with its possibilities for combining visual and audio inputs, quickly became a favorite among the spectrum of media options. With the rise of television, there was a sharp rise in consumption of news and other contents which were extended to less educated audiences who were less apt to read newspapers (Prior, 2007: 56). The inclusive and accessible format of television changed communication; not only well-organized content but also the visual presentation thereof became important. An oft-cited confirmation of the prevalence of television in political power struggles is the 1960 television debate between US-presidential candidates Nixon and Kennedy. A well-tanned, made-up, visually presentable Kennedy triumphed over the paler, sickly looking Nixon whose campaign had not grasped the magnitude importance of the event (Schroeder, 2008). Television allowed the media to focus more efficiently on the people who represented various interests in political debates (Mancini and Swanson, 1996: 11), making personal appeal and connection between political actors and their audiences an essential factor.

Until the turn of the twenty-first century, television had become the prime and most powerful medium in political communication: the ubiquity of TV not only changed the focus (as demonstrated with the Kennedy–Nixon debate) but also the context of communication. Television is a more expensive medium than print or radio; actors who vie for power via this medium can expect to reap rewards, but also to pay a handsome price. Thus, the development of television is seen to increase the role of money in political communication and debates (Mancini and Swanson, 1996) at the cost of rational debates.

'New' Media

The latest wave we were able to observe is the rise of new media. At the turn of the twenty-first century, many social scientists were hopeful that the introduction of digital media would reverse the long-term declines in civic and political engagement. The introduction of digital media changed the cost and structure of communication by increasing the speed with which information could be gathered, and growing the overall volume of information people could access (Bimber, 2003). It also created new opportunities for targeted, two-way communication across space and time. With an estimated 96% of the world population subscribing to mobile-cellular services, and 41% of households worldwide connected to the Internet (Estimated by the International Telecommunication Union, retrieved 22 January 2014 from www.itu.int.), the face of communication has changed.

As a result, cyber-optimists believe that digital media use will strengthen democracy by enhancing two-way, targeted communication between parties and citizens (Dertouzos, 1997; Grossman, 1995; Negroponte, 1995). However, research has emphasized that the outreach of political actors through new media has enhanced selectivity. Thus, political parties made little effort to include new or undecided groups. Instead, they continued to target their communication toward party members and partisans and used their online presence to strengthen ties with already strong supporters. Parties also made little use of the interactive features digital media offers. Instead of moving to a more democratic, horizontal style of interactive communication, political actors strengthened their control over the flow of political communication from the elites to the voters (Bimber and Davis, 2003; Gibson et al., 2003; Gibson & Ward, 1998; Jackson & Lilleker, 2009).

Unlike the Web 1.0 environment, the Web 2.0 environment blurs the boundaries between the user and the producer, in a way that any user can produce content by posting comments, uploading videos, and sharing information. Although the term 'Web 2.0' was coined in 2004 (O'Reilly, 2005), the shift from Web 1.0 to Web 2.0 was characterized by a rather gradual change from the ability to consume content to the ability to coproduce content (Cormode and Krishnamurthy, 2008). The result is a change in the way information is shared and produced. Moreover, this has caused optimism for a more inclusive and more democratic pattern of communication: "The democratic nature of Web 2.0 is exemplified by creations of large number of niche groups (collections of friends) who can exchange content of any kind (text, audio, video) and tag,

comment, and link to both intra–group and extra–group 'pages'" (Cormode and Krishnamurthy 2008: n.p.). The potentials of this form of communication open up the communication for political outsiders and challenges from civil society who have had little chances to access public and political debate under the conditions of traditional mass media.

The introduction of new social and participatory media in the mid-2000s has increased the democratic potential of political communication by lowering the barriers for participation and inaugurating new action repertoires of political mobilization. Hence, Bennett and Segerberg (2012) argue that mobilization takes new forms of personalized 'connective action' which complement the traditional modes of social movement activity and protest politics. The ability for individuals and groups to coproduce content and immediately access political debate via Internet and mobile communication challenged the gate-keeping role of traditional mass media and party elites to fully control communication as a top–down process of communication. Thus, the Internet revolution and the advent of Web 2.0 contribute to the emergence of new 'hybrid media systems' (Chadwick, 2013) consisting of traditional political actors like parties and governments, traditional media, and Internet activists who are entangled in extremely fast and dynamics of political agenda setting and framing.

Media Studied by Various Social Science Disciplines

While the current approach to mass media differs depending on the discipline, there exists a significant amount of cross-fertilization in the study of the media between the social scientific disciplines. Lazarsfeld/Bereleson/Gaudet's study entitled: *Voters Choice. How People Make up Their Mind in a Presidential Campaign* is a prime example of media and communication drawn on by all social science disciplines.

The strong link between media and politics (as emphasized above) commits political science to take an interest in media, and communication science to take an interest in the study of politics. Moreover, as politics is the struggle for influence (power) through communication, the mass media's functions have become an irreplaceable factor in achieving political ends. Once a polity has reached a size where it is no longer possible to include all citizens directly in the political process, messages must be produced and addressed to large numbers of unspecified recipients. In this way, politics takes place largely in the purview of the mass media, making the study of communication inseparable from the study of politics.

Political Science investigates media on three levels:

- macro
- meso
- micro.

At the macro level, agenda setting via the mass media is examined. The meso level of investigation includes the effects of media on the intermediary system. How do parties, interest groups, and NGOs communicate? What effect do the media have on these organizations (macro and meso levels are also closely related to sociological study of mass media)? The examination of voting behavior of individuals constitutes the micro level, where there is also significant overlap with psychology. The subfield of campaign communication developed around 1990 at the intersection of party research, electoral research, and communication and media research. An important field of study in political science pertains to the mass media's effect on political campaigns: do they matter (Farrell and Schmitt-Beck, 2002)? From the field of electoral research, the emphasis lies on changes in voting behavior as caused by campaigns.

While research in political science focuses on mass media and communication within the framework of political institutions and processes, communication sciences takes a broader perspective and goes beyond the framework of political decision making and governance. Though studying political communication as one subfield, it opens up the perspective by asking how flows of information and other forms of messages (like entertainment) are created in media organizations, media businesses and markets and journalism and media cultures (McQuail, 1983). Social scientists within the field of communication investigate the creation of media reality as an integral part of social reality and study the media as institutions that select and produce messages for public debate in society at large. Communication scholars ask how processes of communication influence are shaped by media practices and how they influence cultures, identities, and public sphere.

Psychology's line of inquiry focuses on the effects of the media on the individual in the paradigm of persuasive communication. The history of psychological thought on media can be divided into four major phases which have connections to the historical development of mass communication mediums. The first is the focus on media as propaganda, with strong effects on the individual. This phase is associated with the rise of radio and can be exemplified with the 1938 radio broadcast of the H. G. Wells' *The War of the Worlds*, which instigated fear and panic among many listeners who perceived the story to be reality. Psychology then played down the effects of the media on the individual in the second phase, a paradigm shift that was in part brought on by Lazarfeld et al. (1948) study, which challenged the idea of profound media effects on the individual and society, suggesting that individuals played an important role in the formation of each other's perspectives. The third phase coincided with the spread of television and suggested again that individuals were powerfully influenced by the media, which could influence what people think by influencing what they think about. The fourth phase is characterized by a plurality of models of thinking about the individual and media, including gratification, selectivity, and reception theories. At the same time, contemporary debate in media effects research has also started to question whether the traditional psychological paradigm of persuasive communication is still adequate to account for the impact of media under the new conditions of media environments, communication modes, and messages (Bennett and Iyengar, 2008).

Sociological approaches to media include elements from political and psychological theories, but the focus is on how media affects society. Lazarsfeld, for example, is seen as a sociologist, as are Habermas, Luhmann, and others. Sociology concerns itself with media's role in power struggles (as is the subject of political science), looking at how these come to be

and relate to one another. The public sphere (see Habermas, 1992) or public spheres are the arenas where mass media and other communication create consent, that is, normalize and disperse ideas as a basis for action. Sociologists also look at how media processes are formed. Thus, there is an emphasis on language use and meaning in the social world. A basic sociological question driving media research is how media contributes to societal change or reinforcement of the status quo.

Conclusion

Media, especially mass media, can determine our understanding of societies as well as society's understanding of itself. Power struggles, regardless of the configuration of social groups partaking of these struggles, take place largely through acts of communication. The extent to which media is therefore involved in social struggle renders it a strategic role as a power broker. This is the essence of the interest of the social sciences in studying media: understanding how societies come to be arranged the way they are, as well as to gain insight into past, present, and future changes in societal arrangements. Observing and analyzing changes in communication patterns, methods of communication, and the instrumentalization of media thus form an important part of the social scientific quest to understand, analyze, and evaluate society.

See also: Mass Media and Politics; Mass Media, Political Economy of; Media and History; Media and Social Movements; New Media and the Digital Divide; New Media, News Production and Consumption; New Media, Political Mobilization, and the Arab Spring.

Bibliography

Ayish, M., 2002. Political communication on Arab world television: evolving patterns. Political Communication 19 (2), 137–154.

Bennett, L.W., Segerberg, A., 2012. The logic of connective action. Information, Communication and Society 15 (5), 739–768.

Bennett, L.W., Iyengar, S., 2008. A new era of minimal effects? The changing foundations of political communication. Journal of Communication 58, 707–731.

Bimber, B.A., 2003. Information and American Democracy: Technology in the Evolution of Political Power. Cambridge University Press, Cambridge.

Bimber, B.A., Davis, R., 2003. Campaigning Online: The Internet in U.S. Elections. Oxford University Press, Oxford.

Chadwick, A., 2013. The Hybrid Media System. Politics and Power. Oxford University Press, Oxford.

Cormode, G., Krisnamurthy, B., 2008. Key differences between Web 1.0 and Web 2.0. First Monday 13 (6). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2125.

Coyne, C.J., Leeson, P.T., 2009. Media as a mechanism of institutional change and reinforcement. KYKLOS 62 (1), 1–14.

Curran, J., Park, M. (Eds.), 2000. Dewesternizing Media Studies. Routledge, London.

Dertouzos, M.L., 1997. What will be: How the New World of Information will Change Our Lives, first ed. HarperEdge, San Francisco.

Entman, R.M., 1993. Framing: toward clarification of a fractured paradigm. Journal of Communication 43, 51–58.

Entman, R.M., 2007. Framing bias: media in the distribution of power. Journal of Communication 57, 163–173.

Farrell, D.M., Schmitt-Beck, R. (Eds.), 2002. Do Political Campaigns Matter? Campaign Effects in Elections and Referendums. Routledge, London.

Foucault, M., 1980. Truth and power. In: Gordon, C. (Ed.), Power/Knowledge. Selected Interviews and Other Writings 1972–1977. Harvester Wheatsheaf, Hempstead.

Gibson, R., Ward, S.J., 1998. U.K. political parties and the Internet: "Politics as Usual" in the New Media? The Harvard International Journal of Press/Politics 3 (3), 14–38.

Gibson, R., Römmele, A., Ward, S., 2003. German parties and internet campaigning in the 2002 federal election. German Politics 12 (1), 79–108.

Grossman, L.K., 1995. the Transformation of American Democracy. The Electronic Republic. Viking, New York.

Habermas, J., 1992. Further reflections on the public sphere. In: Calhoun, C. (Ed.), Habermas and the Public Sphere. The MIT Press, Cambridge, Massachusetts, pp. 421–461.

Hadland, A., 2012. Africanizing the three models of media and politics. In: Hallin, D.C., Mancini, P. (Eds.), Communication, Society and Politics. Comparing Media Systems beyond the Western World. Cambridge University Press, Cambridge, UK, pp. 224–245.

Hallin, D.C., Mancini, P. (Eds.), 2012. Comparing Media Systems beyond the Western World. Cambridge University Press, Cambridge, UK.

Hallin, D.C., Mancini, P., 2004. Comparing Media Systems: Three Models of Media and Politics. Communication, Society and Politics. Cambridge University Press, New York.

Hoff, J., Horrocks, I., Tops, P. (Eds.), 2000. Democratic Governance and New Technology. Routledge, London.

Jackson, N.A., Lilleker, D.G., 2009. Building an architecture of participation? Political parties and Web 2.0 in Britain. Journal of Information Technology & Politics 6 (3-4), 232–250. http://dx.doi.org/10.1080/19331680903028438.

Jorgensen, M., Phillips, L.J., 2002. Discourse Analysis as Theory and Method. Sage, London.

Kovarik, B., 2011. Revolutions in Communication: Media History from Gutenberg to the Digital Age. Continuum International Publishing Group, New York.

Laclau, E., Mouffe, C., 1985. Hegemony and Socialist Strategy. Towards a Radical Democratic Politics. Verso, London.

Luhmann, N., 2000. The Reality of the Mass Media. Polity Press, Cambridge.

Lazarsfeld, P.F., Berelson, B., Gaudet, H., 1948. The people's choice: How the Voter Makes Up His Mind in a Presidential Campaign. Columbia University Press, New York.

Mancini, P., Swanson, D.L., 1996. Politics, media, and modern democracy: introduction. In: Swanson, D.L., Mancini, P. (Eds.), Politics, Media, and Modern Democracy. An International Study of Innovations in Electoral Campaigning and Their Consequences. Praeger Publishers, Westport, pp. 1–26.

McQuail, D., 1983. Mass Communication Theory, first ed. Sage, London.

Negroponte, N., 1995. Being digital, 1st ed. Knopf, New York.

Nisbet, E., Moehler, D., 2005. Emerging Political Communication Systems in Sub-Saharan Africa: Some Preliminary Models. In: Paper presented at the annual meeting of the American Political Science Association, 1 September, Washington, DC.

O'Reilly, T., 2005. What Is Web 2.0? Design patterns and business models for the next generation of software. Journal of Digital Economics 65, 17–37.

Ong, W.J., 1982. Orality and Literacy: The Technologizing of the Word. Routledge, New York.

Prior, M., 2007. Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections. Cambridge University Press, Cambridge.

Rugh, W., 1979. The Arab Press. Syracuse University Press, Syracuse.

Schroeder, A., 2008. Presidential Debates: Fifty Years of High-Risk TV. Columbia University Press, New York.

Schudson, M., 2002. The news media as political institutions. Annual Review of Political Science 5 (1), 249–269.

Sparks, C., 2005. Media and the global public sphere: an evaluative approach. In: de Jong, W., Shaw, M., Stammers, N. (Eds.), Global Activism, Global Media. Pluto Press, London, pp. 34–49.

Voltmer, K., 2012. How far can media systems travel? Applying Hallin and Mancini's comparative framework outside the western world. In: Hallin, D.C., Mancini, P. (Eds.), Communication, Society and Politics. Comparing Media Systems beyond the Western World. Cambridge University Press, Cambridge, UK, pp. 224–245.

Social Science and Universities

Sheldon Rothblatt, University of California Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although unrecognizable by their present names, the social sciences have been taught at universities from their founding. At medieval universities, they were tucked away in language or mathematical 'faculties' governed by the theories of Aristotle and his Jewish and Arabic commentators. Finding medieval approaches to knowledge constricting and insufficiently useful, critics and governments created alternative institutions: academies, technical and vocational schools, and specialized professional programs. These institutions varied by period and country. In the long run, however, universities proved to be the most adaptive and resilient, creating different structures for new specialties and new modes of teaching. The definition of an academic career changed to allow for greater disciplinary specialization. In the social sciences, special credit must be given to the Scottish universities of the eighteenth century for pioneering approaches to economics, political science, sociology, psychology, and history. From the nineteenth century onward, industrialism, urbanism, and global wars produced greater demands for pure and applied 'research' (regarded as a German innovation) and problem-solving methods. The digital revolution of the twentieth century has had similar consequences. Today, some social science activity takes place within governmental and private sectors. However, although the boundaries between institutions are more fluid than ever, the university remains the primary location for the fundamental advancement of all forms of knowledge.

Since they first appeared in the twelfth century, universities have been the main, if not exclusive setting for the development of the social sciences. No two universities are, or have ever been, alike, but all possess common features. A division of intellectual labor is foremost. The curriculum, composed of building blocks of knowledge called 'disciplines,' is organized into multiple teaching, training, or research structures called variously (depending upon time, place, and national language) 'faculties' (the oldest), 'institutes,' 'laboratories,' 'departments,' 'schools,' 'boards of studies,' 'colleges' (which can be freestanding), or, more recently, 'programs,' 'organized research units,' or 'centers.'

The Problem of Knowledge

Knowledge and understanding are acquired through classification and specialization. The absence of a universal measure or definition of 'the disciplines' does not alter the elementary cognitive principle that unclassified information is relatively useless. According to the philosopher and logician Alfred North Whitehead, specialism is the 'natural' path to meaning, which underlies all civilizations. Increasing specialism is a salient feature of modern life, an indispensable requirement for scientific investigation, intellectual discovery, and problem solving.

Institutionalizing Knowledge

Disciplinary specialization affects the primary internal structure of universities and their systems of teaching, research, funding, staffing, and even the arrangement of buildings and spaces. Specialties seek what some scholars call 'autonomy,' or control over their own destinies. Accordingly, specialties desire institutional guarantees of disciplinary self-regulation. But the opposite is equally true. The institutional forms themselves influence the content, development, and application of academic specialties. In turn, the missions of educational institutions such as universities are affected by outside pressures generated by governments, the judiciary, religious institutions, the professions, public opinion, the media, and, increasingly, the commercial demands of industrial or high-technology economies. The modern 'nation-state' has been an important element in the formation of all contemporary disciplines but particularly in the natural and social sciences. The relationship between the social sciences, their host institutions, and all other disciplines, as well as civil society and government, is consequently complex, unstable, and unpredictable. The rate of knowledge change and production is more rapid, however, in some historical periods more than others. The resulting disjunctions between the generation and absorption of knowledge can produce social consequences.

The Organization of Knowledge in the Medieval University

During the Middle Ages, university disciplines were incorporated into large divisions known as 'faculties.' Some faculties, as at Paris, were quasi-independent corporations within a single university, while others were less settled in their administration. 'Nations,' or houses for students from European regions were homes away from homes. They became subadministrative and teaching units with separate identities, as were colleges. The 'inferior' (or arts) faculty, with its broad divisions into language and mathematical studies, was a preparatory program for the 'superior' or 'higher faculties' in which the three great professional training programs were located – medicine, law (canon or civil), and theology.

The inferior faculty consisted of disciplines that contributed to the foundations of a general education. The seven liberal arts, inherited from Rome, were divided into two other general classifications. One was the language-based trivium – logic, grammar, and rhetoric. The second was the mathematically

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41064-0

based quadrivium – music, astronomy, geometry, and arithmetic.

The basic organization of knowledge in the arts and sciences teaching divisions was rendered more complicated by the teaching of the 'three philosophies' – natural, moral, and metaphysical – which were also derived from ancient classical sources. The analytical methods, as well as the substance of the philosophies, informed and interpenetrated all faculties and disciplines in a process of knowledge generation and debate that produced gradual and continuous revisions in the curriculum.

The Relationship of Disciplines to One Another

The disciplinary taxonomy does not by itself adequately explain the uses or content of disciplines. The study of literature, for example, of texts such as Virgil's Aeneid, was included within grammar and thereby subordinated to the teaching of syntax.

Architecture could be found within the quadrivium, allied to the mathematical sciences. Rhetoric was a particularly elusive discipline. In some universities (or periods), it was subservient to grammar or logic and often limited to letter writing and religious preaching, but in other circumstances it was an important ingredient in the training of lawyers and public servants. The character of political science (based initially on Aristotle's *Politics*, translated into Latin in the mid-thirteenth century) changed significantly when it was later associated with law and employed in the task of building bureaucratic central governments to replace medieval feudalities. Religious questions regarding the existence of the soul were not confined to the faculty of theology but were debated within logic, using such assigned texts as Cicero's or Aristotle's *Topics*.

Science, in the modern sense of the study of nature, and medicine were part of metaphysics, where various methods and subjects commingled. In a scribal-based and/or oral academic culture dominated by the writings of a relatively few classical authorities, with their Arabic, Jewish, and scholastic glossators and commentators, disciplinary crossover was readily achieved. The need to focus attention on a limited number of issues, notably those related to the spread and administrative growth of the universal Church of Rome or the fiscal and legal problems of princely and imperial courts, contributed to a unity of academic purpose.

Yet the principal differences of knowledge production and acquisition between the medieval and modern universities should not be exaggerated. Disciplines absorb concepts and methods from one another (in other words, the contents of disciplines overlap) and the direction of scientific investigation is influenced by the uses to which knowledge is put. The capacity of a discipline to describe its own trajectory is constrained (positively or negatively) by the faculty or disciplines (or professional associations) to which it might defer.

Today's universities have characteristics similar to earlier universities. The placement of contemporary disciplines like sociology, political science, ethical philosophy, or even cultural anthropology within a school of law determines the approaches and problems that receive primary attention. The study of economics, when situated within business schools, inclines more toward problem solving than toward the laws or principles underlying economic behavior. Historical inquiry, when located within departments of economics, political science, policy studies, music, or education, becomes the study of the evolution of those fields, or is oriented toward questions of interest to the supporting departments.

Anthropologists whose appointments lie within medical schools and hospitals address issues arising from the treatment and care of patients, as do ethical philosophers, psychologists, or social workers in the same environment. Architecture is affected differentially when associated with engineering, schools of urban planning, colleges of the environment, schools of design, or the liberal arts. The study of music in research universities favors composition, musicology, and the history of music more than performance. The examples of disciplinary plasticity and shifting allegiances are endless.

Continuities and Discontinuities in Knowledge Acquisition

The medieval university was not ideologically designed for intellectual discovery but for the dissemination of knowledge. Nevertheless, significant knowledge growth and development transpired, although often unrecognized by practitioners or credited by polemicists. Change was advanced by the normal process of argument, by the discovery of new texts, by the reinterpretation of older sources, by the evaluation of existing systems of logic and proof, by theological controversies such as the Great Schism of the fourteenth century, and by issues concerning ecclesiastical reform eventually culminating in the Protestant heresies.

One essential difference between medieval and modern universities is the higher degree of self-conscious disciplinary promotion in the modern university, as driven by modern professional and external imperatives. Another important difference is the greater amount of ambiguity and relativism tolerated by modern cultures.

Scholasticism and Humanism

The period termed the 'Renaissance' (now usually called the 'early modern' period) or 'rebirth' of classical learning, beginning with fifteenth-century Italy and spreading northward in Europe in the centuries following, has long occupied a central place in discussions concerning the differentiation of disciplines and the arrival of 'modern' conceptions of learning. Historians conventionally divide the learned into two groups: the medieval schoolmen and the early modern humanists. Three issues separated the two groups. The first was the type and variety of classical sources later available when the Greek Byzantine Empire collapsed in Asia Minor. Authors and texts unknown or known only through hearsay were brought into Western Europe. Because sources could now be compared, an interest in variorum texts and in the detection of forgeries grew. Historical criticism, whose place was hitherto minimal, developed as a cardinal feature of early modern scholarship and was applied relentlessly to canonical texts and the dogmatic foundations of Western Christianity. A second issue concerned the role of observation and experiment in the conduct of scientific

work. Humanists and practitioners of the 'new learning' charged that inductive reasoning was neglected in favor of deductive systems of logic derived from Aristotle and his commentators. A third issue was the purpose of the university itself. Detractors alleged that the medieval university was isolated from practical affairs, its members preoccupied with abstract knowledge and shopworn or irrelevant ideas. A corollary point involved personal style. Humanists stressed public speaking, appearance, and the *vita activa*, being interested in careers at the courts of state-building aristocracies in an age of conspicuous political symbols and ceremonies. They contrasted the philosopher-statesman style with the torpid and pedantic manner of university teachers.

Stereotypes, however plausible, are notoriously dangerous, especially when shaped by polemical intent. Not all humanists were outside the university system. Many remained within, attempting, with success, to alter scholastic habits. Some scientists – Galileo is the archetype – borrowed heavily from scholastic methods of interrogating nature and were not as dependent on experiment and observation as they, or encomiasts claimed. And some humanists, in their zeal to attract the patronage of rulers, became hired propagandists, sacrificing their integrity as scholars to hypocrisy and superficial elegance. Furthermore, whatever directions humanistic scholarship took, it is now evident that the world of new learning was first framed within the intellectual structures and methods created by the medieval universities.

Alternatives to Universities

Nevertheless, criticisms of the university as a hidebound, antiquated system of teaching were convincing enough to result in the introduction of different institutional forms and new sources of funding in which innovation, particularly of an applied nature, received a more immediate reception. Even as early as the fifteenth century, the teaching of the common law in England was given a home in vocational institutions designated 'inns of court,' leaving Oxford and Cambridge universities with the teaching of canon law, Roman law, and jurisprudence. In the following centuries, academies, botanical gardens, royal libraries, observatories, museums, military schools, and new colleges were established outside university systems. The Jesuit order, founded in the sixteenth century, was and remained active in educational work. In France in the course of the eighteenth century, the state's need for experts resulted in the creation of specialized engineering schools, starting with the *École des Ponts et Chaussées* in 1747. The Jacobins of the French Revolution adopted this form for their own instrumental purposes. Their *École Polytechnique* of the 1790s was followed in the next century by other *grandes écoles*, institutions of high quality whose mission was to train specialists for appointments in government, industry, engineering occupations, or education. At the same time, the French university system was dismantled by Napoleon as irrelevant to the tasks of a revolutionary and secular society. That system was only gradually rebuilt but never attained the prestige of the leading specialized schools. Only recently has the university acquired a few of the research functions typical of universities in other countries, high-level research in France being normally assigned to a separate system.

The French *grandes écoles* are unique institutions in their patterns of student selection, purpose, and prestige. They are also unique because their prestige is not based on a research mission as in other countries (although now several research functions have been added). Few educational institutions anywhere have equivalent elite connections to the government, bureaucracy, industry, and leading professions. However, other countries have attempted to emulate some aspects of the French engineering model, most notably in the establishment of upper level polytechnical establishments in European capitals such as London, Berlin, Vienna, Stockholm, and Copenhagen, and in US cities such as Cambridge, Massachusetts, and Pasadena, California. Many of these have been evolving toward a university form. Relevant social science departments, foreign language teaching, and, in some cases, humanities courses were added, especially in the twentieth century.

But a second influence is the German technical university as it evolved in the nineteenth century. Also in Germany and Austria, other specialized universities were created: *Wirtschaftsuniversitäten*, for example – institutions where disciplines could be grouped and united with a focus on economic problems. In the United Kingdom, regional universities and university colleges were established, combining applied with liberal arts studies with special attention to the economic or social problems of their regions. Most of these universities later evolved into full-scale research universities.

Medical schools and hospitals, freestanding or associated with universities, became umbrella or 'incubator' institutions for an astonishing range of clinical and scientific specialties. Also in the nineteenth century, geodetic surveys and sponsored explorations supported certain kinds of limited intellectual activity. Although professional education as contained in the 'superior faculties' was the featured curriculum of the medieval university, and especially law in the later centuries, the admission of other forms of vocational or occupational training was never guaranteed. Law, medicine, and theology were acceptable because they suited the needs of urban societies in which religion and secular life interpenetrated. Vocational latecomers had to fight for entry. Rivals for income and teaching appointments and for students, space, and a role in governance, they represented the interests of outsiders. Hence, it was that newer professionalizing occupations of the later centuries, especially the nineteenth, such as engineering, commercial studies, school teaching, public administration, and social work, were associated with separate schools where they might develop an applied mission according to their own rules and methods. Several occupations – notably architecture, accounting, and some legal and medical activities – were controlled by practitioners recruiting through guild apprenticeship methods. But guild and academic forms of education are not necessarily incompatible. The academic profession itself is the supreme example of guild-directed education, controlling staff appointments and curricula.

Early twenty-first century universities tend to be comprehensive. Subjects or emphases once deemed unacceptable to universities built on the German research model or influenced by it are now more likely to be welcome. New recruits to the university community are normally expected to conform to its prevailing styles of intellectual achievement. However, the rapidly increasing fragmentation of the knowledge base evident

everywhere, along with market discipline and government policy, suggests that universities today are less able than predecessors to impose uniform standards of work or a common scholarly ethos on their numerous departments, laboratories, and programs of instruction.

The Transformation of the University

Applied and experimental science owed more to medieval systems of logic and natural philosophy than critics granted. Yet there was indeed a 'scientific revolution' of the seventeenth century that transformed the medieval worldview, produced a materialist conception of the cosmos, departures in physiology, and eventually the treatment of disease, and revised theological speculation concerning human understanding and potential. Mathematical measurement developed as a formidable scientific tool; for despite the medieval quadrivium with its mathematical base, the acceptance of Aristotle's physics by the schoolmen had produced a science that was heavily qualitative and metaphysical in character. Scientific instruments such as the microscope and the telescope, and later the Leyden jar, greatly enhanced the scope of intellectual achievement. Sophisticated machines have continued to be indispensable for science (and certain fields of social science), and never more than in the early twenty-first century.

The 'Knowledge Revolution' of the Nineteenth Century

Historical criticism, the development of the laboratory, and new systems of biological classification underlay the eventual adoption of a research ethic that completely reoriented the pantheon of disciplines inherited from the past.

Teaching and professional training were the two educational missions of the medieval university. Now there was a third, to be called 'research.' The ideas that knowledge could be created, that discovery was an aim of all scholarship, and that the new was preferable to the old were revolutionary assumptions that profoundly affected all educational institutions, and especially universities. The research spirit, while commencing in the scientific revolution and noticeable here and there in later Enlightenment Scottish, German, Dutch, or Swedish universities, did not dominate university culture itself until the nineteenth century. Historically, it is most closely associated with the German university model.

The general acceptance of an academic culture oriented toward original inquiry gave enormous impetus to the establishment of new or subdisciplinary fields. Pressures for the autonomous representation of those fields within universities increased. But two lasting versions of a research ethic emerged: one advocating discovery as an end in itself, the other advocating the use of research results to solve social, economic, or political problems, or to serve the needs of a democratic, consumer society. Proponents of the first feared that applied research funded mainly from commercial sources or results-minded governments would transfer direct and certainly indirect control over discipline formation and knowledge production to agencies outside the universities. 'Pure' research would suffer.

Advocates of the second noted the disappearance of any useful distinction between 'pure' and 'applied' knowledge in the age of high technology. In any case, contemporary society, they maintain, requires applied knowledge as a condition of survival and prosperity. Arguments remain intense, and the available evidence suggests that both sides have a case.

The Place and Influence of the Social Sciences within Universities

As noted, the 'social sciences' were embedded in the scholastic arts curriculum. The bureaucratizing states of the early modern period found a place for some forms of applied social science, notably *Kameralwissenschaft*, within the law faculties. The humanists, however, strengthened belles lettres within the faculty of arts (later, in the German university, to be called the faculty of philosophy). Accordingly, the maturation of the academic social sciences was a slow, two-stage affair. First came new methodologies for the study of social change and process developed by thinkers both inside and outside universities. Second was the emancipation of social science from domination by arts and professional curricula. Later Enlightenment Scottish university historians and philosophers were especially important in providing new insights into the evolution of society and culture. The 'science of man,' the disciplines of 'conjectural history' economics, political science, anthropology, sociology, and psychology are lasting Scottish contributions. Nineteenth-century French intellectuals such as the St. Simonians and Auguste Comte developed some of these beginnings into an overarching scheme of human development based on the 'march of mind,' the progress from primitive religion through metaphysics to 'positive science.' While the separation of subjects such as sociology, anthropology, economics, or psychology from such parent fields as law, history, religion, and philosophy was a nineteenth-century development, the flowering of social science as a full-fledged set of university disciplines belonged to the twentieth century. The timing was not accidental. A focus on the comparative, normative (and deviant) behavior of groups, communities, and classes; on the functioning of institutions such as schools, bureaucracies, professions, religious organizations, trade unions, industrial firms, and family structures; and on the nature of belief and value systems of societies reflected the dominant features and problems of modern society with its mobile, urban, plural populations, and global reach.

Besides concentrating on issues and problems not conventionally incorporated into the inherited model of a university, the social science disciplines claimed the prestige gained by the physical sciences. As in the natural order, the human or social order could also be explained by scientific laws. And, although in time some of the rigor and promise of words such as 'scientific laws' would be mitigated and exchanged for softer terms such as 'patterns' or 'regularities,' the claim to scientific objectivity served as a major legitimizing strategy. It was also the basis for an instrumental approach to the resolution of social problems and the means by which the social sciences secured access to external funding sources.

The number of social science disciplines and the different strategies pursued by social scientists to become citizens of the republics of letters and science in varying national contexts is

a highly complex story. Yet some insight into the issues may be gained from a cursory look at the history of the London School of Economics and Political Science (LSE), founded in 1895 by socialists and left-wing liberals on the model of the *École Libre des Sciences Politiques* in Paris (the LSE was later incorporated into the federal structure of the University of London).

Two issues confronted the founders. The first was the independence of the LSE, solved initially by the creation of a wholly new establishment where the disciplines could function on their own terms. This conforms to the pattern indicated earlier, where alternative institutions provide the means of disciplinary entry into the knowledge universe. The second issue concerned the LSE's primary mission, and here the founders disagreed on basic aims and methods. Should the institution's mission be applied or scholarly? Should attention be focused mainly on metropolitan regions such as London or spread more widely? Were ideological commitments to social amelioration compatible with standards of objective research? These differences of purpose have never been completely resolved within the LSE's academic divisions.

Indeed, the differences cannot be resolved since they are inherent in the history of the social science subdisciplines. The structural flexibility of the modern research university has provided a practical solution. The applied social sciences have found homes in the professional schools, a number of which – social work, urban studies, education, public policy, and perhaps now journalism and media studies – are in fact based on the social sciences. But law, medicine, public health, and architecture all have room for social science subjects and methods. Elsewhere in the university, the 'autonomous' social sciences are separated into distinct disciplines according to the history of each. They are located in departments within 'colleges of letters and science' (the typical US case), traditional arts faculties, or new faculties of social science (in some European universities). They are also to be found in crossover and multidisciplinary programs and research units.

Ideas, concepts, and problem-solving methods developed by the social sciences are now commonplace elements of any university curriculum. Historical research is greatly influenced by psychology, anthropology, economics, sociology, statistics, and political science, but history, as a 'universal recipient,' has often if not invariably demonstrated intellectual flexibility. Yet literary studies have also become social scientific as, inter alia, critics incorporate interpretations derived from the sociology of class relationships, the psychology of the individual, and the anthropologist's interest in symbol and ritual.

In concert with the technical and physical sciences (and the revolutionary biological sciences), the social sciences have played a prominent role in tying the university to social or national objectives, strengthening its place in the formulation of policies affecting the distribution of wealth, educational opportunities, the provision for social services, and other polemical issues that are the concern of modern, mass democratic societies. The entanglements with powerful and influential outside interests are not to everyone's liking, as they compromise the university's decision-making independence. However, the university has never been a wholly autonomous institution, not even within free and liberal societies. Such relative independence as it enjoys is the result of negotiation or 'social contracts' that each generation of university leaders, teachers, and researchers is destined to conduct.

Controls over Disciplinary Formation

The organization of the disciplines varies in different national and university settings, affecting the kinds of disciplinary emphases receiving attention.

The Chair-Holder System

The medieval faculties were the ultimate internal determinants of the curriculum, with alternative institutions such as colleges providing some opportunities for independent innovations. Centuries later, with the evolution of the professorial or 'chair' system of academic leadership, best known in its German form, individual chair holders administered 'institutes' and laboratories and set the standard for teaching and research, especially at advanced levels. Although the institute structure was embedded in the traditional faculty structure, changes had occurred that made chair holders and deans the effective authorities, and they respected one another's independence.

This system, which became standard in continental Europe, Latin America, Israel, Japan, and wherever else the German model of a university took hold before 1939, led the way in world scholarship and science. One essential drawback, however, was that younger academics with newer specialties could not easily find room within the old academic structure. They had two options: to await chair vacancies at home or elsewhere, or to persuade institutions (and the government ministers who largely controlled financing, especially after the later nineteenth century) to establish new or different chairs in developing subdisciplines. Before unification in 1870, 'Germany' was a collection of independent *Länder* competing against one another for professorial talent. Thus, intra-German rivalry itself was a variable in the formation of new fields of inquiry, along with the adoption of radical critical methods (the 'higher criticism'), and philosophical conceptions concerning personal development (*Bildung*) and research ('Humboldtianism').

The Disciplinary Department

Many of the features of the organization of knowledge that had evolved in Europe also existed in the United States, whose higher education system, first established in the colonial period, reflected English and Scottish influences, then French, and finally German ones. The chair-holder system was much in evidence throughout the United States as well, but by the end of the nineteenth century it was losing its authority as the arbiter of the curriculum. It was gradually supplanted by the 'department.' A century later, its main features had been adopted in most university systems elsewhere. A department is based on a discipline rather than a collection of disciplines (as in a faculty structure) and is run by committees. The department head, elected or appointed usually in rotation, is *primus inter pares*. Although senior and influential, chair holders and their equivalents share authority with other teaching titles. The

relatively 'democratic' decision-making apparatus of the department is broadly in keeping with medieval legacies of guild self-government.

Another US innovation, a credit-unit system of instruction that combines teaching and examining in a single person, was a departure from British and continental European external examining practices, and underscored the intellectual freedom of students and individual instructors and researchers. In this context, subdisciplines were able to flourish. While it is still possible for an entire department to be built or rebuilt along narrow and exclusionary disciplinary lines, such solutions are generally short lived given the independence of individual members and the variety of available funding sources.

One other US innovation of the twentieth century deserves mention: the advent of the graduate school. The continental European chair-holder system was suited to advanced instruction and the award of advanced degrees because it relied on a selective system of elite schooling – the *lycée* in France, the *Gymnasium* in Germany and Scandinavia, their counterparts elsewhere – and gave the professoriate discretion in choosing the best candidates for creative work. The US university, which for historical reasons was not well supported by an excellent system of secondary school instruction, recruited undergraduates of mixed preparation. The solution to the problem of how to offer research-level instruction was found in the establishment of a wholly separate administrative division freed from the necessity of providing remedial and general teaching. This innovation contributed greatly to the building up of strong research traditions within the social sciences and other disciplines. Versions of the American graduate school appeared in Europe during the 1980s and 1990s in response or reaction to governmental policies encouraging the development of mass forms of secondary and higher education. However, these have not been as significant as in America.

Before and during the introduction of departmental and credit-unit systems of instruction, the US pattern of disciplinary formation was similar to the European one. Subjects that did not find an immediate reception within traditional colleges and universities, especially subjects stigmatized as 'practical' (which could include modern languages), sought a more hospitable environment in alternative institutions. These had been springing up ever since the American Revolution.

By the end of the nineteenth century, liberal arts colleges, comprehensive universities (combining professional and liberal education), specialized institutes (such as the West Point Military Academy), prestigious polytechnics (such as the Massachusetts Institute of Technology), a German-style 'research university' (the Johns Hopkins University), and state institutions with not only applied but also liberal education missions (the 'land grant' universities) were members of a differentiated system that had expanded in connection with national ambitions, market conditions, and a federal constitution.

Within the older US colleges and universities, internal reconfiguring took place alongside external differentiation. In the earlier nineteenth century, academic subjects such as the natural sciences or modern languages that were not necessarily applied but were regarded as inimical to prevailing teaching values were able to create alternate degree tracks, if not of equivalent prestige. Within the senior English universities of Oxford and Cambridge, subject-based 'schools' (Oxford) or 'boards of studies' (Cambridge) representing specialties effectively replaced the medieval faculties structure. As in the German case, professorial chair holders controlled the university system of lecturing. However, as Oxford and Cambridge were collegiate universities or federations of relatively autonomous societies (legally 'charitable' societies), the professorial system did not constrain the whole system of instruction. The collegiate structure with its small classes and tutorial arrangements in fact liberated the professors from routine or introductory teaching.

Criticisms of the Departmental Organization

The flexibility inherent in a departmental organization and its comparative openness to innovation reduce the temptation for subjects to migrate toward alternative institutional structures. Departments are therefore one of the factors that have brought the 'university' institutional form into the center of modern society as the principal source of knowledge generation and dissemination. Factors other than departmental organization – and they are outside the scope of this article – may include national imperatives, such as war and industrial competition, mass higher education requiring extensive private but especially public resources, professionalism, consumer economies, and the high division of labor so typical of contemporary society.

Yet the departmental organization has its critics. Discipline-based departments as collections of specialists are said to be protective of privileges and unreceptive to cross-disciplinary fields of study. A similar criticism but with a different purpose arises from the defenders of the US tradition of liberal education. Departments are attacked as the chief obstacles to the development of general education studies for undergraduates. In both instances, the response has been the creation of interdisciplinary or multidisciplinary departments, or cooperative programs of teaching, as well as special appointments in broad fields such as 'Humanities' or 'Social Sciences.' Such experiments are now typical of newer European universities.

But an opposite kind of blame has also been voiced, most notably by two famous US *enfants terribles*: the maverick Thorstein Veblen at the turn of the twentieth century and the more establishment 1930s critic Abraham Flexner. Both excoriated US departments for being far too open to innovation, fads, and market demand, and far too eager to attract the financial support of philistine parvenus. Flexner in particular admired the German university tradition with its high intellectual standards and *Kultur*. For him, the elite characteristics of the European higher educational system – the alliance between government bureaucrats who decided educational budgets and chair holders who controlled institutes – meant that demands for the teaching of subjects unworthy of incorporation into the noble inheritance called a 'university' could be successfully resisted.

In the twentieth century, the views of critics like Flexner proved to be shortsighted. Totalitarian governments in Nazi Germany and the communist states of Central and Eastern Europe, as well as their copies elsewhere, provided vivid lessons of how knowledge could be manipulated from above and turned from its natural channels of curiosity and negotiation to unimaginable and terrifying applications. The university

systems of those countries had to be rebuilt after World War II, and the task remained unfinished as the twenty-first century commenced.

Universities of the Twenty-First Century

While it is nearly an axiom of twenty-first century, higher education policy in all democratic societies that universities must be flexible in providing access, instruction, and research, disagreement exists on the institutional forms best suited to promote those ends and whether traditional educational values will prevail.

The Entrepreneurial University

Technologies associated with the arrival of massive open online courses employing the Internet as well as the possible competition of degree-granting 'corporate' universities have issued new challenges to all existing forms of higher education. One response is to adopt the new technologies and emphases and assimilate them to the existing fabric of the university. This in fact is being done on a significant scale, although critics fear that in the process, the inherited character of the university will suffer if the emerging teaching technologies and different kinds of research priorities are not accommodated to accepted structures and norms. The new research environment is heavily weighted toward market discipline, government demands, and problem solving in national and global markets since these are the dominant sources of funding. Nevertheless, it is hoped that even entrepreneurial outreach will be infused with less commercially inclined values regarding objectivity and curiosity in knowledge generation. The battles continue.

Fading Institutional Boundaries

Modern society is dynamic; its economy is a whirligig. All institutions are affected: families, churches, schools, labor unions, government agencies, philanthropies, and industrial organizations. The boundaries between them are permeable and 'transgressive.' University identity and academic loyalties are therefore facing formidable challenges. These have already succeeded in altering the attachments of academics to universities. For some, the university is no more than a *pied-à-terre*, a temporary convenience while larger research ambitions are pursued in a variety of other institutional directions. One body of analysts has noted that research projects are often cooperative undertakings by teams composed of university and nonuniversity representatives and that many joint projects actually originate outside universities in government laboratories or industrial firms. Universities are therefore not the only sources of innovation. Much research is not even open ended as in the traditional university but commissioned, 'commanded,' or 'strategic' (one carryover from the global wars of the twentieth century).

Another consequence of looser institutional boundaries is the greater flow of individuals from one institution to another as philanthropies, government agencies, the military, private or nonprofit think tanks, industries, financial bodies, and even the media exchange players, providing new colleagues and settings for intellectual activity. Such institutions are therefore tied together in mutually reinforcing networks. Rival research sites have existed for centuries, whereas, as noted earlier in the formation of alternative institutions, the circulation of those with academic qualifications has never been greater than today.

The flows of educated talent is greatest in the medical, biological, physical, and technical sciences, but the policy sciences have also profited, as well as social disciplines oriented toward problem solving in circumstances where city planners, environmental experts, health workers, education specialists, psychologists, economists, criminologists, and social workers are required. The opportunities for movement are less certain in the humanities and 'soft' social sciences. The differences between knowledge domains are accordingly sharpened, leaving certain of the humanities and social sciences as weakened guardians of the university's traditional boundaries and self-understanding.

A Primary Role for Universities

Nevertheless, even as great changes affect the nature of the research university, and those connected to it, the fact remains that all institutions depend upon university-based disciplines (whether housed in departments or in new 'centers') to educate and train the talent required for the success of a modern, professionalized nation. The university remains the primary home of the academic profession, or, to quote a social historian, the academic profession is the 'key profession' because it educates all the others.

See also: History, Professionalization of; Knowledge, Sociology of; Policy Knowledge: Universities; Science, History of; Science, Sociology of; Science: Constructivist Perspectives, History of; Scientific Knowledge, Sociology of; Universities and Science and Technology: Europe.

Bibliography

Ben-David, J., 1991. Scientific Growth. University of California Press, Berkeley, CA.
Bryson, G., 1945. Man and Society: The Scottish Inquiry of the Eighteenth Century. Princeton, NJ.
Clark, B.R., 1987. The Academic Life, Small Worlds, Different Worlds. Carnegie Foundation for the Advancement of Teaching, Princeton, NJ.
Clark, B.R., 1998. Creating Entrepreneurial Universities: Organizational Pathways of Transformation. Pergamon Press, Oxford, UK.
de Ridder-Symoens, H. (Ed.), 1992. A History of the University in Europe, vol. I. Cambridge University Press, Cambridge, UK.
de Ridder-Symoens, H. (Ed.), 1996. A History of the University in Europe, vol. II. Cambridge University Press, Cambridge, UK.
Flexner, A., 1994. Universities, American, English, German. Transaction Publishers, New Brunswick, NJ.
Geiger, R.L., 1986. To Advance Knowledge. The Growth of American Research Universities, 1900–1940. Oxford University Press, Oxford, UK.
Gibbons, M., et al., 1994. The New Production of Knowledge. Sage Publications, London, UK.
Henkel, M., 2005. Academic identity and autonomy in a changing policy environment. Higher Education 49, 155–176.
Kelley, D.R. (Ed.), 1997. History of the Disciplines. The Reclassification of Knowledge in Early Modern Europe. University of Rochester Press, Rochester, NY.
Kerr, C., 2001. The Uses of the University, fifth ed. Harvard University Press, Cambridge, MA.
Oleson, A., Voss, J. (Eds.), 1979. The Organization of Knowledge in Modern America, 1860–1920. Johns Hopkins University Press, Baltimore, MD.

Perkin, H.J., 1969. Key Profession, the History of the Association of University Teachers. A.M. Kelley, New York.
Rothblatt, S., 1997. The Modern University and Its Discontents. The Fate of Newman's Legacies in Britain and America. Cambridge University Press, Cambridge, UK.
Soffer, R.N., 1978. Ethics and Society in England: The Revolution in the Social Sciences, 1870–1914. University of California Press, Berkeley, CA.
Veblen, T., 1965. The Higher Learning in America. A Memorandum on the Conduct of Universities by Businessmen. A.M. Kelley, New York.
Wagner, P., Wittrock, B., Whitley, R. (Eds.), 1991. Discourses on Society. The Shaping of the Social Science Disciplines. Kluwer Academic, Dordrecht, The Netherlands.
Weisz, G., 1983. The Emergence of Modern Universities in France, 1863–1914. Princeton University Press, Princeton, NJ.

Social Science Infrastructure: North America (Research and Teaching)

Gisèle Yasmeen, Research – Social Sciences and Humanities Research Council of Canada, Ottawa, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition by D.L. Featherman, volume 11, pp. 7503–7511, © 2001, Elsevier Ltd.

Abstract

This article discusses two types of infrastructure: the first, pertaining to general scholarly infrastructure for the purposes of networking and the exchange of knowledge (i.e., scholarly infrastructure, also known as 'institutions'), and the second regarding infrastructure for the conducting of research itself (i.e., research infrastructure). The article provides a historical overview of the development of both types of infrastructure in both the United States and Canada. Finally, current and future challenges, particularly with respect to cyberinfrastructure are dealt with in the conclusion.

The preservation and advancement of knowledge presupposes an infrastructure that sustains research and exchange of new information, prepares requisite personnel, and advances cumulative knowledge of the empirical world and theoretical insights. This is no less true in the social sciences than in the natural sciences. While scholarship is increasingly international, with its own sustaining infrastructure, a base for advancing and preserving knowledge exists within each country. In the United States, the foundations of the existing infrastructure for the social and behavioral sciences (hereafter, social sciences) were laid in the early decades of the twentieth century and expanded after World War II (WWII). In Canada, some foundations were laid both before and after WWII and were further developed and consolidated beginning in the late 1970s. Certain limitations of this infrastructure are now becoming apparent, particularly as the full potential of the digital age calls for increasingly sophisticated cyberinfrastructure for research, and scholarly enterprise more generally. If not urgently addressed, these factors may constrain the development of the US and Canadian social science and may hinder the realization of an international science of society as the twenty-first century unfolds.

What Is Infrastructure?

There are two types of infrastructure required for the social sciences to flourish. While the two are related on an empirical level, they are kept analytically distinct for this article. One type – 'scholarly infrastructure' – is the organizations and institutions, such as national academies and scholarly associations, funding sources and review systems, libraries, databases, theoretical models, methodologies, journals and other communication media, and advanced training programs that separately and collectively sustain and promulgate the research enterprise. In the digital age, this has expanded to include "emerging topics such as linking publications to research data, combining licensing with open access or enhancing repository networks through 'services on top of repositories' such as usage statistics, text/data mining, research profiles etc" (see Association of European Research Libraries Web site).

The second type of infrastructure is required to conduct research itself. For most of science, 'research infrastructure' conjures up images of national laboratories, orbiting telescopes, polar weather stations, or supercolliders; in other words, research instruments are maintained at great expense. What this image misses are less tangible research resources, such as databases and data mining tools that sustain and advance the research enterprise. This is especially true for the social sciences for which the 'brick and mortar' components of infrastructure are fewer and less obvious (CBASSE, 1998).

Research infrastructure is defined as 'single-sited' or 'distributed' resources, including facilities, equipment, instruments, and related services used to conduct research. This includes "knowledge-based resources such as collections, archives, or structured scientific information; enabling Information and Communication Technology based e-Infrastructures such as Grid, computing, software, and communication networks" (see OSIRIS Web site).

Research infrastructure emphasizes shared and sustaining resources that not only make scientific research possible, but also to add value to research activity. 'Shared' refers to what is accessible as a public good for science, something whose value is not diminished by multiple uses or that produces economies of scale or multiplier effects. For example, a multimillion-dollar investment in building an instrument for magnetic imaging of the mind is much like a similar investment in a 10-year panel study of a nation's youth as they leave school and enter the workforce. Many investigators can use both 'instruments' for a great variety of studies, each of whose value is not diminished, indeed, perhaps even made greater by virtue of others using the same instrument for different projects. This is especially true when such projects – all conducted at low additional marginal costs because of the same base investment in the 'instrument' – allow for the aggregation of information and therefore newly emergent information not available from separate studies. For example, multiple time series of different topical domains might cross-reference brain images over time in cognitive and in emotional activity, or, from the same sampled cohort, might interrelate the economic outcomes of the school to work transition and of the trajectories of social development over the same period. The term 'sustaining' refers to necessary resources without which the advance of knowledge would be slowed or halted, or, would be very expensive to nearly impossible to reconstruct if left to deteriorate. For example, with respect to scholarly infrastructure, if the funding of the National Science Foundation (NSF) were cut in half for

a 4-year period because of a change in priorities, research universities might be forced to close some institutes, and some might even redirect their missions. From a research infrastructure standpoint, a 40-year time series on the American or Canadian electorate, broken for a decade, may never retrieve the value it would have had were it not interrupted.

Historical Development of Scholarly Infrastructure

In both the United States and Canada, the foundations of the infrastructure of the social sciences were laid in the two decades prior to WWII and just after its conclusion (Fisher, 1991; Prewitt, 1995; Bulmer, 2001).

The United States

Prior to WWII, a remarkable convergence of interests within the leadership of private philanthropy and of the burgeoning social science disciplines and their national associations yielded an extraordinary array of institutions and resources that were to both consolidate and accelerate the growth of social research.

Private philanthropy by industrial capitalists like Andrew Carnegie, the Rockefellers, and Russell Sage in the nineteenth century provided the initial means for social science's development in the early decades of the twentieth century (Bulmer and Bulmer, 1981; Prewitt, 1995). These private foundations created the first explicit scholarly infrastructure for social science. Whether this investment was motivated by the self-protective interests of a capitalist social class (i.e., Fisher, 1993) or reflected deeper religious and humanitarian concerns in the public interest (i.e., Bulmer, 2001; Lagemann, 1989) can be debated. In any case, the early infrastructure of social science emerged from a fortuitous convergence of the announced goals of philanthropic foundations and the coalescence, within the USA, of defined social science disciplines – history, political science, psychology, economics, anthropology, and sociology (Ross, 1991). An important additional element of that early infrastructure was the creation of graduate schools at a few universities, starting at Johns Hopkins in 1876 and then at Clark, the University of Chicago, Columbia, Harvard, and Yale.

As for the foundations, leaders such as psychologist Beardsley Ruml at the Laura Spelman Rockefeller Memorial drew inspiration from the natural and medical sciences as a means of understanding and, it was presumed, controlling the empirical world and set as a goal the creation of a corresponding science of society. This science was not an end in itself; it did not reflect so much a commitment to knowledge per se than to use rational analysis and scientific research to reveal the 'root causes' of the massive problems of industrial America and thereby to overcome them. Investment into research and the scientific enterprise, e.g., into the putative causes of urban poverty rather than exclusively into assisting poor families, reflected a broad sea change in the goals of private philanthropy in the early decades of the twentieth century (Prewitt, 1995). This change in philanthropy led to the creation of several research institutes in medicine and the natural sciences, e.g., the Rockefeller Institute for Medical Research in 1901. But it also coincided with the rapid emergence of several additional foundations and organizations that supported the development of social and behavioral research (see Table 1).

The early history of the Social Science Research Council (SSRC) amply illustrates the evolving close partnership between private philanthropy and university-based research of the social science disciplines after World War I (WWI). At the core of this history is a very effective network of leadership within the foundations and the small set of elite university research programs – leadership that circulated between these two worlds and galvanized common cause to promote a science of society. Charles F. Merriam, president of the American Political Science Association, spearheaded a cadre of counterpart economists, sociologists, and historians (and joined later by anthropologists and statisticians) with commitments to a scientific approach in their respective fields but also to multidisciplinary research inquiries. They founded the Social Research Council (SRC) in 1923 to use a scientific, interdisciplinary approach in analyzing the problems of society, to develop new methodological tools for this purpose, and to assist university-based academic programs in training a new generation of social scientists for this agenda. Merriam and colleagues considered but ultimately rejected the idea of affiliating with the American Council of Learned Societies – one of the first councils to span the social sciences and humanities. Instead, they pursued, unsuccessfully, incorporation within the National Research Council (NRC), the operational arm of the National Academy of Sciences, which at that time incorporated only the fields of psychology and anthropology, the disciplines bearing closest affinity with the natural sciences. The SRC might have failed were it not for Beardsley Ruml, newly appointed director of the Laura Spelman Rockefeller Memorial and recent PhD in psychology from the University of Chicago. By 1922, Ruml had $2 million per year to invest in developing a science of society, but he was uncertain that university-based disciplinary departments, or the associations per se, would progress rapidly in that direction. Ruml had endorsed Merriam's unsuccessful bid to be incorporated into the NRC and suggested that the SRC add 'Science' to its name, and thus, it became the SSRC. Subsequently, Ruml and the Memorial began a steady flow of financial resources to the SSRC, and through it to university-based faculties to develop their methodological capacities and the volume of science-inspired research into the burgeoning set of social problems of the era. During this era, the SSRC functioned as the social science counterpart to the NRC and at the same time served as adjunct to the Memorial as a 'pass-through' foundation. Between 1923 and 1929, the Memorial distributed about $40 million for social research and advanced (quantitative) methodological training, much of it distributed through the SSRC (Bulmer and Bulmer, 1981; Featherman, 1994).

As for the social science disciplines themselves, their progress as research fields can hardly be understood without acknowledging the catalytic importance of universities. This includes their gradual transformation as centers for academic research as well as for their original mission of educating undergraduates (Geiger, 1993). Here, too, the role of foundations and private investment in creating this element of infrastructure in the early twentieth century cannot be overlooked, even though their support of universities was quite selective and targeted (Bulmer, 2001). For example, the Rockefellers

Table 1 Milestones in the USA and Canada in the creation of scholarly infrastructure

USA	*Canada*
1865: Creation of the American Social Science Association (precursor to the American Economic Association and the National Conference of Charities and Corrections;	1882: Establishment of the Royal Society of Canada (RSC) by an Act of Parliament as the senior Canadian collegium of distinguished scholars, artists, and scientists. The RSC represents all fields of knowledge.
1876: Creation of graduate programs in the social sciences beginning at Johns Hopkins University, followed by Clark, University of Chicago, Columbia, Harvard, and Yale	
1903–27: Emergence of foundations and organizations supporting social and behavioral research – General Education Board (1903); Milbank Memorial Fund (1905); Russell Sage Foundation (1907); Carnegie Corporation of New York (1911); National Bureau of Economic Research (1921); Social Research Council (1923); Brookings Institution (1927); Social Science Research Council (1923)	Establishment of the National Research Council in 1916 (for the natural sciences and engineering). Establishment of ACFAS in 1923. The acronym originally stood for *Association canadienne-française pour l'avancement des sciences* and was renamed *Association francophone pour le savoir* with the original acronym conserved in 2001. The organization represents all fields of knowledge.
1929: President Herbert Hoover's appointment of a Blue Ribbon Panel to assess trends in American society, especially in the aftermath of the stock market crash.	1940: Creation of the Canadian Research Council in the Social Sciences Creation of the Royal Commission on National Development in the Arts, Letters, and Sciences (i.e., the 'Massey Commission') was officially appointed on 8 April 1949 by Prime Minister Louis St-Laurent and completed its work in 1951 1943: Creation of the Humanities Research Council of Canada (subsequently the Humanities Federation of Canada) 1951: The commission recommended significant investment in art, science, and culture and led to the creation of the National Library of Canada (now Library and Archives Canada) and the Canada Council for the Arts established in 1957.
1950: Creation of the National Science Foundation	1978: Creation of the Social Sciences and Humanities Research Council of Canada
1973: Creation of the National Institute of Education	
1990s: Full directorate status for the social and behavioral sciences at the National Science Foundation	1996: Merger of the Social Science Federation of Canada and the Canadian Humanities Federation
1995: National Institutes of Health (NIH), Office of Behavioral Research was opened with a direct reporting line to the Director	2000: Creation of the Canadian Institutes for Health Research, modeled on the NIH and replacing the Medical Research Council of Canada and expanding its mandate to include funding of all aspects of health research, including by social scientists. 2005: Creation of the Council of Canadian Academies

provided foundational financing for the University of Chicago and encouraged its evolution toward a research and teaching university. While the modern research university was not to emerge in the USA until after WWII, universities like Chicago and Columbia were notable incubators of empirical social scientific inquiry much earlier than 1940. At Columbia, for example, statistical research began in 1890 under the influence of Richmond Mayo-Smith and Franklin Giddings who were focused on local conditions and problems of urban life. Later, just prior to WWII, the Bureau of Applied Social Research under the leadership of Paul Lazarsfeld elevated Columbia's prominence in empirical social science and embraced national issues. This broader scope was a context for refining social theories and quantitative methodologies.

Elements of this new infrastructure included the University of Chicago and its celebrated Chicago School of Social Research, the SSRC in New York City, the National Bureau of Economic Research, and what later was to become the Brookings Institution. In the later period, with the rapid evolution of the federal administrative state following the great depression and WWII, the NSF and the National Institutes of Health (NIH) funded an expansionary phase of research. Most of this science-building activity occurred within the nation's research universities, themselves a product of the postwar era. The course of social sciences in the USA reflects the history of its infrastructure just as much as it does the intellectual biographies of its leaders and evolving corpus of scientific contributions (Deutsch et al., 1971).

The effectiveness of university–government partnership during the war prompted President Truman and Congress to establish the NSF in 1950. The size of the role for federal support of civilian research and development (R&D), especially for the support of 'basic' research during peacetime, was debated. And the place of the social sciences – viewed at best as 'applied' and not a 'basic' science – within the mission of the NSF, was heavily contested. Expenditures for these fields represented a mere 1.2% of the NSF budget in 1956 and only grew to 1.6% ($890,000) by the end of NSF's first decade (Larsen, 1992). Additional support for social and behavioral research emerged in an expanding NIH in the postwar years, although the NIH focused its resources on biomedical and other health-related science. Indeed the pattern of scattering federal support for science across agencies has sometimes impeded a consolidated and centralized federal science policy (and corresponding decentralized decisions about science funding) since the close of WWII. Funding for the social sciences came less from federal sources than from the foundations, but on a steadily declining basis after mid-twentieth century.

During the same postwar period, economists – most prominently among social scientists – were recruited into new elements of an expanding Executive Branch. The President's Council of Economic Advisors was established in 1946. But the most rapid expansion of the federal administrative state occurred in the 1960s during the Kennedy and Johnson presidential administrations; the federal government became one of the largest employers of social and behavioral scientists (Featherman and Vinovskis, 2001). This was a period of jarring racial turmoil and wide public perception of unsolved and long-standing social problems of crisis proportion, such as poverty amid plenty. The Great Society initiatives of the Johnson presidency, and large federal social welfare expenditures (over $60 billion in 1970, a 142% increase over the decade), drew upon social science research for technical guidance. At the NSF, a Social Science Division was created, somewhat expanding the legitimacy of these fields. Similarly, Congress created a National Institute of Education in 1973 to promote a broad agenda of research and evaluation of school reforms. Correspondingly, support for social research blossomed from $384 million in 1961–62 to $803 million in 1966–68, half coming from federal agencies (Featherman and Vinovskis, 2001). During this period, both academic and more policy-oriented social scientists joined faculties of programs and schools of public policy at a growing number of research universities. Thus, as during WWII, the so-called War on Poverty and associated aspects of the Great Society added to an expanding infrastructure, its funding, and its personnel base for social and behavioral sciences.

If the 1960s and early 1970s constituted something of a 'golden era' for social science, its infrastructure, and the role of its practitioners in policy circles, the era did not last long and ended for many in disillusionment. Large-scale quasi-experiments such as Project Head Start and the Negative Income Tax experiment, conceived and designed by social scientists in league with public policy officials, failed to demonstrate large or lasting impacts (White and Phillips, 2001; Gueron, 2001).

Solutions to social problems proved far fewer, more intractable, than promises; and social scientists argued openly among themselves about seemingly arcane (to politicians and the public) reasons for failure (Aaron, 1978). Politicians and the public gradually came to the view that public monies invested in social research were irrelevant if not wasteful. Meanwhile, another war, this one in Southeast Asia, deeply divided the American public during the Johnson and Nixon presidencies. Academics in general, and social scientists in particular, tended to oppose this war. From the perspective of the more politically conservative administrations of Presidents Nixon and Reagan, the Vietnam War had so politicized and radicalized the professoriate that scholars, especially in the social sciences, were no longer seen as providing impartial research and evaluations (Featherman and Vinovskis, 2001). By the mid-1980s, federal funding for the social sciences had been cut back draconically.

In the closing two decades of the twentieth century, a few events affected scholarly infrastructure of the social sciences. Following the Watergate episode and resignation of President Nixon, Americans grew more cynical of Washington, and government itself. Academic social scientists, perhaps with the exception of economists and lawyers, retreated to universities to undertake more theoretically inspired research. Those with a policy orientation either joined expanding faculties in schools of public policy or were recruited into a rapidly proliferating set of partisan think tanks (both liberal and conservative) that ringed the Capitol to provide politically inspired analysis and commentary to their respective constituencies.

Canada

In Canada, fledgling activities in the social sciences – as well as the humanities – went through similar soul-searching with respect to the relationship with government, the relationship between what were seen as 'purely' scholarly or scientific pursuits and those with a more 'applied' or 'problem-based' orientation. Canadian activities in the social sciences were initially supported by American foundations such as Carnegie and Rockefeller or specific activities such as conferences and publications (Fisher, 1991). This was, in many ways thanks to the prominence of and respect afforded to individuals such as Harold Innis who played instrumental roles in the development of social sciences both in Canada and on the world stage. The University of Toronto, McGill, and Queens were the three universities where social science research was the most developed at the time in English-speaking Canada. The SSRC of Canada – established in 1940 – eventually transitioned to become the Social Science Federation of Canada. Together with its sister organization, the Canadian Federation of the Humanities, the Council successfully advocated along with other scientific, cultural, and artistic interests for the establishment of the Royal Commission on National Development in the Arts, Letters, and Sciences (i.e., the 'Massey Commission') spurred by fears of the influence of the United States, the Commission was a large-scale consultative enquiry into the state of Canadian scientific and cultural institutions, the Commission led, ultimately to the creation of the Canada Council for the Arts and subsequently, the Social Sciences and Humanities Research Council. The Humanities and Social Science federations merged in 1996 for practical reasons but perhaps also, it might be argued, due to more blurred conceptualization of the social sciences and humanities in Canada where scientific and interpretive methods coexist. This may be due to the francophone presence in Canada and greater European influence where 'science' is often viewed more broadly as codified knowledge, as opposed to associated solely with the scientific method. Indeed, the francophone presence has had a significant impact on all aspects of scholarship in Canada including conditioning the formation of scholarly infrastructure (Gingras, 2006).

The Foundational Period of Research Infrastructure: Before WWII

While scholarly and research infrastructure are deeply intertwined, this section delves into the foundations of research infrastructure itself in both the USA and Canada, with a focus on the development of shared and sometimes capital intensive resources and tools, which have made the social research enterprise possible (**Table 2**).

Table 2 Examples of significant social science research infrastructure

The United States	*Canada*
The United States Census: Undertaken since 1790, it gradually developed into viable research infrastructure beginning in the nineteenth century.	Canadian Census: First conducted in New France in 1666 and, in British North America, beginning in 1871 2001 – The first Research Data Centers established 2006 – Compute Canada funded 2010 – Canadian Polar Data Network
Hull House Maps and Papers: Du Bois 'The Philadelphia Negro'; the Pittsburgh Survey	Ethnographies of indigenous peoples in Canada by anthropologists and missionaries
National Elections Studies	1965 to Present – Canadian Elections Study 2001–09 to present – Canadian Longitudinal Study on Aging
International: Luxembourg Income Study (comparative database)	

The United States

At the close of the nineteenth century, the roots of American social science drew their energy from investigative journalists, reformers in settlement houses, and statisticians seeking to count unemployment and develop reliable representations of labor markets in local areas (Bulmer, 2001). Statisticians such as Carroll Wright developed crude labor statistics for the state of Massachusetts after the civil war, mirroring the 'statistical movement' in Victorian Britain and the 'probabilistic revolution' pioneered by Quetelet and others in Europe during the early to mid-nineteenth century (Bulmer et al., 1991). What drove even this descriptive, social statistical enterprise was related to reformers and advocates for the poor like Jane Addams and W. E. B. Du Bois at the beginning of the twentieth century. Hence, social sciences at the time were fueled by a progressive social reform ideology, and extended to concerns about the fragility of democracy and threats to economic self-sufficiency in an America beset with racial and labor conflicts, massive numbers of immigrants, and unstable economic cycles with impoverishment for many.

Devoted to improving the lot of urban poor, the Settlement House movement, inspired and drew upon a series of mapping projects and urban surveys, e.g., Hull House Maps and Papers; Du Bois' The Philadelphia Negro, the Pittsburgh Survey. These protoscientific inquiries of the late nineteenth and early twentieth centuries, some commissioned and funded by the Russell Sage Foundation in the interests of improving social welfare, comprised one element of the Social Survey Movement. They, in turn, gave impetus to what was to become modern statistical sampling and survey research during and just after WWII (Converse, 1987; Bulmer et al., 1991).

Likewise, at the University of Chicago, the seminal empirical studies of W. I. Thomas on Polish immigrants to Chicago (published in five volumes with Florian W. Znaniecki from 1918 to 1920) gave rise to a 'Chicago School' of sociological research in 1915–35 under Robert E Park, Ernest Burgess, and William F Ogburn (Bulmer, 1984). It also led to a nascent science of American politics after 1920, with leadership from Charles Merriam and his students Harold Gosnell and Harold Lasswell.

Throughout the foundational period, however, the government of the USA played a rather insignificant role in creating the early infrastructure of the social sciences. Most government existed at state and local levels; the federal administrative state grew in size and importance after the great depression and WWII. While the USA Census dates from 1790, the US Bureau of the Census was created as a permanent federal agency in 1902. Other elements of a federal statistical system emerged subsequently.

Events such as wars and economic depressions, and governmental response to them, guided the evolution of the social sciences during the foundational period. For example, during WWI the mobilization of American men, (and their screening for fitness, and assignment) gave rise to nonexperimental, but nonetheless quantitative psychology, to the psychometric assessment of personality and mental abilities (drawing upon European precursors). This war effort spawned 'basic' research on individual differences, as well as an industry of psychological and educational testing (Carson, 2001).

William Ogburn, a Chicago sociologist, headed the SSRC research effort and filed its multivolume report, *Recent Social Trends in the United States*, in 1933 after President Hoover left office. The report took the view commonly accepted at the time: if facts and trends about households, towns, and institutions could be clearly discerned and precisely measured, then surely rational decision makers would use the information in the public interest. This view contrasted sharply with the orientation of the Social Survey Movement earlier in the century, when data were collected directly in the service of correcting urban problems (Bulmer, 2001). However, the report was widely criticized. Within elite academic circles, such as the leadership of the SSRC, reservations emerged about the potential distorting effects of association with political agenda setting, especially the compromise of scientific impartiality as a basis of intellectual inquiry. In the emerging federal policy world of the New Deal era, the report was dismissed for its lack of politically informed analysis and interpretative prescription: the facts could not and would not speak for themselves. What administrative decision makers of the emerging federal state sought was knowledge for action in a political arena and tested against the requirements of political decision making rather than against the requirements of academic social science professionals seeking legitimacy through impartiality (Lynn, 2001).

Canada

With respect to research infrastructure in Canada prior to WWII, it is important to keep in mind that – despite confederation and political independence from Britain in 1867 – a colonial mindset and structure in the world of academe prevailed. Hence, European and American influences were strong and the uniqueness of Canadian perspectives and research resources were not yet fully appreciated. Yet, there were periods of seminal transformation, which might be

considered the equivalents of some of the pathbreaking work seen in the USA in the same era. For example, Franz Boas, the father of modern anthropology, studied the indigenous peoples of the Pacific Northwest in British Columbia the Canadian High Arctic. In the nineteenth century, early Christian missionaries, particularly Roman Catholics such as Father Adrian-Gabriel Morice, but also Moravians and Methodists, documented languages, customs, and family structure, which later served as legal evidence in modern land claims.

Subsequently, the dominant Euro-Canadian society set out to complement the economically oriented theories developed by the likes of Harold Innis, W.A. MacIntosh, D.G. Creighton, and J.M.S. Careless. The publication of sociologist S.D. Clark's *The Social History of Canada* in 1942 might be seen as an important milestone in the development of a repertoire of knowledge at the time featuring a comprehensive approach (see The Canadian Encyclopedia Web site). Similar attention by both French- and English-speaking social scientists to the francophone presence is a distinguishing feature of the growth in research infrastructure at that time. H.B. Ames' pioneering sociological study of working-class Montréal, *The City Below the Hill* (1897; reproduced in 1972) is one example. Canadian social and economic analyses were influenced by both the Annales School as well as British social history. Geography also played a key role in the development of research resources such as mapping. Donald Fisher's The Social Sciences in Canada: 50 Years of National activity by the Social Science Federation of Canada (1991) provides copious detail on other similar tomes, which laid the foundations of Canadian social sciences prior to WWII.

Period of Punctuated Evolution: After WWII to the Early Twenty-First Century

In both the United States and Canada, the period after WWII represented a flourishing of research infrastructure, in part, due to the economic growth that distinguished the postwar era and the associated growth in universities and postsecondary training. This section points to a few key developments.

The United States

The great depression, New Deal, and subsequent American participation in WWII contributed to growth in the size and political scope of the federal administrative state. Scientists and engineers, as well as their social and behavioral counterparts, contributed to the war effort in a wide variety of federal agencies. Within the Department of Agriculture, for example, studies of the war's impact on the farm economy and households gave rise to more refined survey research and its technology, especially probability sampling. Because science and technology were so instrumental to the war effort, federal expenditures on R&D rose from $100 million in 1940 to $1.5 billion in 1945 (Featherman and Vinovskis, 2001). Thus, WWII created the funding infrastructure for a lasting partnership between the federal government and universities and stimulated the flourishing of the research university.

The abundant infrastructure of the golden era, especially in the form of repeated cross-sectional surveys (e.g., the Current Population Surveys; the National Election Studies), longitudinal studies (e.g., Panel Study of Income Dynamics; National Longitudinal Surveys of Youth), and quasi-experiments or intervention studies (i.e., Gueron, 2001) provided common databases for both academic- and policy-inspired research and for often fierce partisan debates over issues such as welfare reform (e.g., Danziger, 2001). Many of these surveys owe their existence to the Bureau of Applied Social Research at Columbia as well as the growth of the rival 'Michigan School' (Johnston, 2011).

Canada

In Canada, the aftermath of the Massey Commission and its legacy of public funding and institutions allowed for the development of systematic research infrastructure resources to nurture research in the social sciences. Examples include, the Canadian Election Study (CES), which is "among the world's oldest academically-based national election studies, with an almost unbroken record of fieldwork dating back to 1965" (Johnston and Blais, 2008) and Canadian contributions to comparative international data sets such as the Luxembourg Income Study. The CES may have unwittingly been a 'nation-building' exercise, as it was one of the first longitudinal data sets to involve cooperation by both English- and French-speaking social scientists. Indeed, in the development of survey research owe a debt to Canadian social scientists through advances in the use of repeated cross-sections (Johnston, 2011; Brady and Johnston, in press) and earlier pioneering work done by Canadian Saul Forbes Rae in collaboration with George Gallup (Gallup and Rae, 1940).

In the 1990s, Canada witnessed a significant turning point in the development of more capital-intensive research infrastructure, in part due to the establishment of the Canada Foundation for Innovation and its mission to support capital investment in research infrastructure, often in collaboration with federal funding agencies. A prime example is the establishment of the Research Data Centers in 1998, which make social statistics available in secure facilities across the country in a manner that does not contravene the Statistics Act.

The turn of the twenty-first century witnessed other exciting investments into Canadian social science research infrastructure. A prominent example is the multi-institutional Canadian Century Research Infrastructure initiative (2003) designed to digitize census data and make it accessible and interoperable. Another key development was the establishment of the Canadian Longitudinal Study on Aging (CLSA), "a large, national, long-term study that will follow approximately 50 000 men and women between the ages of 45 and 85 for at least 20 years." The development of the CLSA from 2001 to 2009 was facilitated by the creation of the Canadian Institutes of Health Research in 2000, modeled on the US NIH, and mandated to fund all aspects of health research, including from social science perspectives. In addition to collecting new data, the CLSA, also cleans and makes interoperable multitudes of existing data sets. Last, but certainly not least, the Canadian Polar Data Network is an outgrowth of the very successful data management undertaken supporting the context of the International Polar Year supported by the three national research granting agencies.

The Promise and Challenge of Big Data and Cyberinfrastructure

Now well into the second decade of the twenty-first century, the challenge and promise of cyberinfrastructure looms large. For decades, the European Science Foundation has led the way in developing successive science protocols, which take explicit account and fund cross-national projects and infrastructure elements such as data archives and advanced computational facilities on a shared, international basis (European Commission, 2010).

Nevertheless, forward strides have been taken in the past 10 years with respect to 'big data' and cyberinfrastructure in the USA and Canada. Cyberinfrastructure is defined as:

> the coordinated aggregate of software, hardware and other technologies, as well as human expertise required to support current and future discoveries in science and engineering. The challenge of cyberinfrastructure is to integrate relevant and often disparate resources to provide a useful, usable, and enabling framework for research and discovery characterized by broad access and "end-to-end" coordination.
>
> Berman, 2005

Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and repositories, visualization environments, and people linked together by software and high-performance networks to improve research productivity and enable breakthroughs not otherwise possible (see Indiana University UITS Web site). The USA and Canada have made considerable advancements in cyberinfrastructure in recent years.

In the USA, following a series of workshops in the early part of this century (see NSF Cyberinfrastructure Web site), the NSF created an Office of Cyberinfrastructure in July 2005 as a horizontal programing unit reporting directly to the Director of NSF. This resulted in creative funding for cyberscholarship across the US as well as the introduction of mandatory data management plans. By October 2012, the Office had been realigned as a unit within the Directorate for Computer and Information Science and Engineering (see University of Michigan Advanced Research Computing Web site).

In Canada, Genome Canada – which primarily funds 'laboratory' research but does provide support to social scientists working on the legal, economic, environmental, and social aspects of genomics – was a pioneer in data management and stewardship. Furthermore, the second edition of the Tri-Agency Policy Statement on Research Ethics provides best practices for the management of data. Canadian scholars have relished the opportunity to work creatively with their US and European counterparts for several editions of the Digging into Data challenge, which has also been a groundbreaking initiative in the field of big data and cyberinfrastructure.

In conclusion, both the USA and Canada have progressed substantially with respect to harnessing the potential of research infrastructure in the twenty-first century. By working together as well as with European, Asian, and collaborators from around the world, scholars in the social sciences – often in collaboration with the humanities or other fields of knowledge – are poised to make even more significant contributions to the preservation and advancement of knowledge.

See also: Associations, Scholarly; Data Bases and Statistical Systems: Sociology; E-Communications Platforms and E-Learning; Panel Surveys: Uses and Applications; Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations); Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Infrastructure: North America (Resourcing and Funding); Social Science Infrastructure: North America (Social Science Academies and Related Organizations).

Bibliography

Aaron, H.J., 1978. Politics and the Professors: The Great Society in Perspective. Brookings Institute, Washington, DC.

Berman, Francine, 2005. SBE/CISE Workshop on Cyberinfrastructure for the Social Sciences. Archived from the original on 11 September 2006. San Diego Supercomputer Center.

Brady, H.E., Johnston, R. Repeated cross-sections in survey data. In: Robert, S., Stephen K. (Eds.), Emerging Trends in the Social and Behavioral Sciences. John Wiley and Sons, Hoboken, NJ, in press.

Bulmer, M., 1984. The Chicago School of Sociology: Institutionalisation, Diversity and the Rise of Sociological Research. University of Chicago Press, Chicago.

Bulmer, M., 2001. Knowledge for the public good: the emergence of social sciences and social reform in late nineteenth and twentieth century America, 1880–1940. In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Bulmer, M., Bales, K., Sklar, K. (Eds.), 1991. The Social Survey in Historical Perspective, 1880–1940. Cambridge University Press, Cambridge, UK.

Bulmer, M., Bulmer, J., 1981. Philanthropy and social science in the 1920s: Beardsley Ruml and the Laura Spelman Rockefeller Memorial, 1922–1929. Minerva 19, 347–407.

Carson, J., 2001. A cultural history of intelligence. In: Porter, R., Ross, D. (Eds.), The Cambridge History of Science: Modern Social and Behavioral Sciences, vol. 7. Cambridge University Press, New York.

Commission on Behavioral and Social Sciences and Education (CBASSE), 1998. Investing in Research Infrastructure in the Behavioral and Social Sciences. National Academy Press, Washington, DC.

Converse, J., 1987. Survey Research in the United States: Roots and Emergence 1890–1960. University of California Press, Berkeley, CA.

Danziger, S., 2001. Welfare reform policy from Nixon to Clinton: what role for social science? In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Deutsch, K.W., Platt, J., Senghaas, D., 1971. Conditions favoring major advances in social science. Science 171, 450–459.

European Commission, 2010. Riding the Wave. How Europe Can Gain From the Rising Tide of Scientific Data. Final Report of the High Level Expert Group on Scientific Data. The European Commission. Available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

Featherman, D.L., 1994. SSRC, then and now: a commentary on a recent historical analysis. Items. Social Science Research Council 48 (March), 15–22.

Featherman, D.L., Vinovskis, M.A., 2001. Growth and use of social and behavioral sciences in the federal government since World War II. In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Fisher, D., 1993. Fundamental Development of the Social Sciences: Rockefeller Philanthropy and the United States Social Science Research Council. University of Michigan Press, Ann Arbor, MI.

Fisher, D., 1991. The Social Sciences in Canada: 50 Years of National Activity by the Social Science Federation of Canada. Wilfrid Laurier University Press, Waterloo.

Gallup, G., Rae, S.F., 1940. The Pulse of Democracy: The Public-Opinion Poll and How It Works. Simon and Shuster, New York.

Geiger, R.L., 1993. Research and Relevant Knowledge: American Research Universities since World War II. Oxford University Press, New York.

Gingras, Y., 2006. Why Canada Never Had an Association for the Advancement of Science. Physics in Canada, NovemberDecember Available at: http://www.chss.uqam.ca/Portals/0/docs/articles/Physics_in_Canada_07.pdf.

Gueron, J.M., 2001. Welfare reform at the state level: the role of social experiments and demonstrations. In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Johnston, R., 2011. Development of Survey Research. Oxford Bibliography Online. Oxford University Press. Available by subscription at: http://www.oxfordbibliographies.com/view/document/obo-9780199756223/obo-9780199756223-0019.xml.

Johnston, R., Blais, A., 2008. The Past and Future of the Canadian Election Study, Report to the Social Sciences and Humanities Research Council of Canada. Available at: http://www.politics.ubc.ca/fileadmin/user_upload/poli_sci/Faculty/johnston/JohnstonandBlais_May16__2__-_our_revisions.pdf.

Lagemann, E.C., 1989. The Politics of Knowledge: The Carnegie Corporation, Philanthropy and Public Policy. Wesleyen University Press, Middletown, CT.

Larsen, O., 1992. Milestones and Millstones: Social Science at the National Science Foundation, 1945–1991. Transaction Press, New Brunswick, Canada.

Lynn Jr., L.E., 2001. The making and analysis of public policy: a perspective on the role of social science. In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Prewitt, K., 1995. Social Sciences and Private Philanthropy: The Quest for Social Relevance. Essays in Philanthropy. Indiana University Center on Philanthropy. no. 15.

Ross, D., 1991. The Origin of American Social Science. Cambridge University Press, Cambridge, UK.

White, S.H., Phillips, D.A., 2001. Designing head start: roles played by developmental psychologists. In: Featherman, D.L., Vinovskis, M.A. (Eds.), Social Science and Policymaking: A Search for Relevance in the Twentieth Century. University of Michigan Press, Ann Arbor, MI.

Relevant Websites

http://www.libereurope.eu/committee/scholarly-research – Association of European Research Libraries, "Scholarly Communication & Research Infrastructures," Re-inventing the Library for the Future.

http://ccri.library.ualberta.ca/enindex.html – Canadian Century Research Infrastructure (CCRI).

http://www.thecanadianencyclopedia.com/articles/social-history – The Canadian Encyclopedia.

http://www.clsa-elcv.ca/ – Canadian Longitudinal Study on Aging (CLSA).

http://polardatanetwork.ca/?page_id=59 – Canadian Polar Data Network.

http://rt.uits.iu.edu/ci/index.php – Indiana University Cyberinfrastructure Web site.

https://www.nsf.gov/cise/aci/reports.jsp – National Science Foundation Cyberinfrastructure.

http://www.osiris-online.eu/riexplained.html – OSIRIS.

http://www.statcan.gc.ca/rdc-cdr/ – Research Data Centrers (RDC) Program.

http://arc.research.umich.edu/2012/09/08/nsf-realigns-office-of-cyberinfrastructure/ – University of Michigan Advanced Research Computing.

Social Science Infrastructure: Africa

Freek Cronjé, Johann van Wyk, and Jean D Losango, Bench Marks Centre for CSR, North West University, Potchefstroom, South Africa
Samuel Lule, Social Workers and Anthropologists (ESSSWA), Addis Ababa, Ethiopia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The institutional foundations for social sciences in Sub-Saharan Africa (SSA) experienced substantial transformations since the end of the colonial era. The region has seen a great mushrooming of new government-funded education and research institutions and currently has more than 700 universities, with the amount of private centers, institutes, networks, and nongovernmental organizations also expanding at a fast pace. Despite a significant increase in education and research infrastructure over the past half a century, the region also experiences an increasing backlog in capacity to accommodate an even faster increasing amount of social science students and scholars. The main challenge across the region is the lack of sufficient funding, particularly state resources, to keep pace with burgeoning populations and concomitant increases in demand for academic and research resources.

The purpose of this article is to provide an overview of the current state of the social sciences infrastructure in SSA, with specific reference to its funding, education, and research structures. One of the main findings is that the rapid growth in the number of students is proving to be a challenge to the sustainable financing of higher education. Africa is the only region in the world that has experienced a decrease in the volume of current public expenditure per student over the past two decades. This state of affairs is having an adverse impact on the quality, efficiency, and sustainability of education and research programs. As a result, and given that the state is the main source of funding for higher education in SSA, different forms of cost sharing to cover tuition and other types of fees (e.g., from household financing or student loan schemes) are increasingly being implemented in most African countries. Social sciences research in the region is predominantly practiced within the universities, but a few countries have research institutes or centers devoted to the social sciences. However, the most progress in terms of the development of social science structures appear to be at a pan-African level with networks such as Council for the Development of Social Science Research in Africa and Organization for Social Science Research in Eastern and Southern Africa making great progress in overcoming the challenges of fragmentation of social sciences activities and output in the region.

More specifically, this article also focuses on three case study countries in terms of different aspects of the social sciences; these countries are South Africa, the Democratic Republic of Congo, and Ethiopia.

Overview

Since the end of the colonial era, the institutional foundations for social science research, teaching, and related activities in Sub-Saharan Africa (SSA) have undergone significant transformation. Growing from a very small number half a century ago, there are now more than 700 universities in the SSA region (Ranking Web, 2014) and they are still increasing in number at a fast pace. New higher education institutions are being set up by both governments and private providers, while research centers, institutes, networks, and nongovernmental organizations (NGOs) are also burgeoning (Sall, 2010).

While many colleges and/or fully developed universities existed before independence across the continent, the development of social research and education of the social sciences is very much a postcolonial phenomenon (Mouton, 2010). However, the key concepts and theoretical frameworks with which most African scholars worked in the first independent decades were 'made in the West.' Subsequently, one of the main efforts of regional social science councils such as Council for the Development of Social Science Research in Africa (CODESRIA) and Organization for Social Science Research in Eastern and Southern Africa (OSSREA), and other professional associations of social science practitioners, has been to address challenges of autonomy by working toward building a networked, self-aware community of scholars. Some have been determined in creating a so-called 'African library' in order to replace what Mudimbe called the 'colonial library' (Sall, 2010).

However, one of the main challenges constraining growth in African-based social sciences research and education is funding. According to Tefera (2013), problems surrounding the financing of higher education institutions are worldwide; however, nowhere in the world is financing higher education more problematic than in SSA. The two main reasons for this challenge are the high and increasing unit (per student) cost of higher education; and the pressure for increasing enrollments due to rapid population growth including the fast increasing number of youth finishing secondary school with aspirations for tertiary education. There is a staggering need in Africa to raise enrollment numbers, which places severe pressure on regional governments. In many countries where a decade ago tens of thousands of students attended higher education institutions, the figure now amounts to hundreds of thousands. Given the strong link between higher education and development, it is now more crucial than ever for adequate resources to be invested into the sector in a sustainable and meaningful manner (Tefera, 2013).

Against the above, the purpose of this article is to provide an overview of the current state of the social sciences infrastructure

in SSA, with specific reference to its funding, education, and research structures. Three case studies – South Africa, the Democratic Republic of Congo (DRC), and Ethiopia – will also be put to the fore.

Social Sciences Infrastructure in SSA

Funding

The same as many countries in the world, the state is the main source of funding for higher education in SSA. However, the degree of dependence on the state is believed to be higher in SSA countries than internationally. Despite a lack of detailed data, it is reported that up to 90–95% of the total higher education operating budgets in the region is met out of government grants, with the remainder coming from student fees (Teferra and Altbach as cited in Tilak, 2011).

The rapid growth in the number of students is certainly a challenge to the sustainable financing of higher education. Africa has maintained its public investment in higher education between 1991 and 2006, allocating approximately 0.78% of its gross domestic product (GDP) and around 20% of its current public expenditure on education to higher education. However, during the same period, the total number of students pursuing higher education tripled, climbing from 2.7 million to 9.3 million at an annual average rate of 16%, while public resources allocated to current expenditure in that sector only doubled (an annual average rate of 6%). The decline in public expenditure per student is having an adverse impact on the quality and relevance of education programs. Africa is the only region in the world that has experienced a decrease in the volume of current public expenditure per student (30% over the last 15 years) (World Bank, 2010).

As a percentage of GDP, higher education has received a very small share in Africa, as the figures in **Table 1** suggest – far below 1%: namely 0.5% in Francophone Africa and 0.8% in Anglophone Africa (Tilak, 2011).

Apart from state funding, higher education in Africa also receives funding from the following sources (World Bank, 2010):

- *Private resources in higher education funding*: Faced with inadequate public financing, the share of private resources in higher education financing is expanding. The contribution from households accounts for approximately 25% of national expenditure (state and households) on higher education. It varies widely according to country, ranging from less than 10% in Mali, Chad, and the Republic of Congo to more than 50% in Uganda and Guinea-Bissau. However, household financing of higher education is relatively low when compared to household investment in other levels of education (30% of national expenditure in primary education and more than 45% in lower secondary education).
- *Cost sharing*: Different forms of cost sharing are being implemented in most African countries. As of 2009, at least 26 countries in Africa charge either tuition fees or other types of fees such as examination fees, registration fees, identity card fees, library fees, and management information system fees. Overall, higher education institutions in Africa generate about 30% of their income (from less than 5% in Madagascar and Zimbabwe to 56% in Uganda and 75% in Guinea-Bissau).
- *Financial aid programs*: To combine financial efficiency and equity goals, the introduction of tuition fees is accompanied by the development of financial aid programs and student loans. Financial assistance policies are critical components of cost-sharing policies in Africa. Most Francophone countries have privileged nearly universal financial assistance through free or subsidized social services (food, transportation, and housing) and scholarships for living expenses. Some have improved the targeting of public assistance and adopted means-tested policies. Student loans are operating in at least 13 African countries, mainly in Anglophone ones. Cost recovery remains the main challenge in most countries for student loans to be effective and sustainable.
- *External assistance to higher education*: International aid in support of higher education is on average US$600 million annually, or one-quarter of all international aid to the education sector in SSA. This relatively low share reflects the current emphasis given by most donors on the development of basic education and the achievement of Education for All. In addition to the small amounts of aid, two main factors limit the impact of aid. First, only 26% of aid to higher education goes directly to African universities and research centers. The remainder is provided through scholarships abroad or is accounted for by directly imputing student costs in the donors' universities. Second, aid is highly fragmented, owing partly to the lack of donor coordination. This imbalance in aid to higher education certainly limits its impact and makes its current allocation questionable.
- *Foundation support*: In addition to bilateral donors, foundations have played a growing role in financing higher education in Africa. Since 2000, seven American foundations (namely, the Ford, Carnegie, Rockefeller, MacArthur, Hewlett, Mellon, and Kresge foundations) plus the Partnership for Higher Education in Africa have made significant investments in African universities. Between 2000 and 2008, the partnership foundations contributed an aggregate of US$354 million toward higher education initiatives in seven countries (Ghana, Kenya, Mozambique, Nigeria,

Table 1 Public expenditure on education in Africa

	% of total government expenditure		*% of GDP*		
	Early 1990s	*Around 2003*	*Early 1990s*	*Around 2003*	*Higher education % of GDP, 2003*
Francophone Africa	22.9	17.6	3.3	2.7	0.53
Anglophone Africa	16.1	21.4	3.0	4.5	0.83
Other African countries	12.9	11.7	2.4	2.0	
Total Africa	19.3	18.2	3.1	3.3	0.63

Source: Brossard and Foko (cited in Tilak, J.B.G., 2011. Financing higher education in sub-Saharan Africa. Africanus, 41 (2), 4–31).

South Africa, Tanzania, and Uganda) and on an Africa-wide basis. The most significant focus of the partnership has been the development of universities' physical infrastructure and human and organizational capacity. Information technologies and connectivity to the internet are at the core of these efforts, with investments until 2010 of more than US$30 million (World Bank, 2010). Most European and Scandinavian countries also have donor programs and projects directed at SSA, e.g., Sweden (Swedish International Development Cooperation Agency (SIDA)) and the United Kingdom (Department for International Development (DFID)).

Education

Colleges, university colleges, and/or fully developed universities existed before independence in countries such as Sierra Leone, Ghana, Nigeria, Ethiopia, Uganda, Senegal, Rhodesia and Nyasaland, Egypt, Morocco, Algeria, Tunisia, and South Africa. However, the development of social research and of the teaching of the social sciences is very much a postcolonial phenomenon. In many African states, the postcolonial state built most of the research and training institutions (universities, institutes, and centers) during the first few decades after independence (Mouton, 2010). As mentioned earlier, there are now more than 700 universities in SSA. However, the development of educational institutions is not currently keeping pace with the rapidly expanding demand.

Ebrima Sall (cited in Mouton, 2010) identifies five factors that are particularly important in understanding the context of social science teaching and research in Africa:

1. economic change (from developmentalism to decline and liberalization);
2. social change (changes in the status of large sections of the elite from high to low, rapid urbanization, and a rise in mass poverty);
3. political change (authoritarianism, followed by liberal democratic reforms in some countries, conflicts and civil wars in others);
4. change in information and communication technologies; and
5. change at the level of the university and other social research institutions (expansion of the sector and diversification of types of institutions and governance systems).

Given these factors (among others), the SSA region finds it increasingly more difficult to ensure the supply of sufficient educational infrastructure. A common result of this inadequacy is high levels of 'out-migration' of social science students. In addition to the well-known flight of high-level human capital (academics and scholars) from the African continent over the past four decades, there has been another equally devastating 'secondary brain drain' – not at the level of scholars and scientists but at the level of postgraduate students. Many students in the region do not study in their home country because of the lack of adequate endogenous facilities, expertise, and very often simply (in the case of postgraduate students) because there are no master's or doctoral programs for them to enroll in.

The countries with the highest outbound mobility rates in SSA are Botswana (87%), Namibia (65%), Lesotho (53%), Swaziland (50%), Mauritius (41%), and Zimbabwe (31%). Students from Mozambique tend to go and study in Portugal; students from Mauritius prefer studying in France as first choice. In both cases, South Africa is the second preferred destination. For many other SSA countries, South Africa is one of the top three preferred educational destinations (together with the USA and UK). Many commentators have observed that an increasing number of universities in SSA are now reliant on South Africa and universities in the North for the training of their doctoral students in particular (Mouton, 2010).

If current trends continue apace, the total number of students for the entire African continent could reach between 18 million and 20 million by 2015. The investment required over the 2006–15 period to increase the capacity of current institutions (classrooms, libraries, laboratories, workshops, and lecture halls), establishing new institutions that are better distributed across the territory and improving administrative and teaching materials, is estimated at approximately US$45 billion (in 2006 dollars) for public higher education as a whole, of which US$20 billion is for low-income countries. However, very few countries in Africa have the leeway to increase the public financing of higher education (World Bank, 2010). Besides inadequate physical infrastructure (and lack of funds for sufficient development), the development of appropriate mechanisms for intergenerational learning is also lacking. As the first and second generations of African professors are retiring or nearing retirement, what have been called the third and fourth generations of scholars, including all those who are doing, just completing, or have just completed their PhD training, are likely to be faced with a shortage of qualified mentors within the universities (CODESRIA, 2005).

Research

Social sciences and humanities research within SSA is predominantly practiced within the universities. A few countries have government-funded research institutes devoted to the social sciences (as in the case of the Human Sciences Research Council (HSRC) in South Africa). In addition, there are also some social research institutes and research NGOs in many countries (such as the Institute for Basic Research in Kampala and again many examples in South Africa) that operate outside of the universities. Most of these research institutes and centers are funded through either international agencies or donor organizations with little if any government support. Given these exceptions, social science scholarship is conducted mostly within academic departments in the universities of the region (Mouton, 2010).

However, the inadequacy of funding has limited institutions' ability to offer adequate remuneration or to invest in infrastructure, research facilities, and equipment, thereby hindering overall research capacity. The consequences of this situation are evident when one considers the minimal contribution of African universities to international academic research (World Bank, 2010). It is well known that most African universities have gone through far-reaching changes over the past two decades, not least of which has been the huge increase in student enrollment numbers at the same

time as international and national funding of public universities has declined. This has left many universities and their capacity to undertake scientific research in a very precarious position. This is no less true of social sciences research (Mouton, 2010).

One of the major difficulties that the social sciences in the SSA region had, and still have, to face is fragmentation of activities and of the community of scholars as a whole. This fragmentation was largely, but not exclusively, due to the colonial partitioning of Africa into more than 50 states, most of which are small and economically dependent. Outside North Africa, where Arabization has been a major development in recent years, social science research is mostly conducted in European languages, particularly English, French, and Portuguese. The building of a 'networked community of scholars' therefore required efforts to transcend disciplinary, linguistic, gender, generational, regional, and ideological divisions (Sall, 2010).

The lack of government support for social science research in SSA translates into very little support for research institutes and centers dedicated to the social sciences and humanities whether based at universities or operating effectively as NGOs. Mouton (2010) compiled a list of research centers dedicated to the social sciences in 25 Sub-Saharan countries (excluding South Africa). The list indicates a total of 149 research institutes and centers. However, of these, only 79 (or 53%) had an active web site in June 2009. But having an active web site does not necessarily mean that the web site has current contents. According to Mouton (2010) only 65 (43% of overall total) of these web sites have contents that could be regarded (very charitably) as being recent: a web site was considered as being 'current' when it contained news or listed events at the center for the past 3 years (2007–09) (**Table 2**).

Table 2 Number of social science institutes by country and 'currentness' of web site (excluding South Africa)

Country	*No. of research institutes*	*Active web site June 2009*	*Current*
Kenya	18	13	12
Nigeria	13	11	6
Ghana	18	9	7
Burkino Faso	9	6	3
Namibia	6	6	4
Zimbabwe	10	6	2
Cote D'Ivoire	5	5	4
Senegal	8	5	5
Malawi	6	5	1
Botswana	4	4	3
Etiopia	8	4	4
Tanzania	6	3	3
Uganda	7	3	3
Benin	4	2	1
Mali	4	2	2
Lesotho	2	2	0
Madagascar	2	2	1
Cameroon	7	2	1
Gambia	3	1	0
Niger	1	1	1
Sierra Leone	1	1	1
Rwanda	1	1	0
Somalia	1	1	0
Chad	4	1	0
Gabon	1	1	1
	149	97	65

Source: Mouton, J., 2010. The State of Social Science in Sub-Saharan Africa. 2010 World Social Science Report: Knowledge Divides, Background Paper. International Social Sciences Council.

From a precursory scanning of these web sites, it became clear that the majority of these centers have a small staff (less than 10) and are invariably funded through international donor funding. This means that the majority of these centers are heavily dependent on so-called 'soft' money, which poses a constant threat to their long-term sustainability. With the 'decline' in the number of robust and vibrant university-based research centers, there has been an increase in transnational and regional research networks. One could argue that such networks are emerging as a direct result of the impact of forces of globalization, greater international collaboration, and increased access to the internet. But at the same time, such networks are also filling the 'void' left by strong national research centers (Mouton, 2010).

The following bodies could however be considered as the two most prominent regional social science research organizations on the continent:

- CODESRIA is headquartered in Dakar, Senegal. It was established in 1973 as an independent pan-African research organization primarily focusing on social sciences research in Africa. It is recognized not only as the pioneer African social research organization but also as the apex nongovernmental center of social knowledge production on the continent (Salih and Draman, 2012).

CODESRIA was established out of a desire to contribute to the growth of an independent, pan-African community of scholars engaged in the production of scientific knowledge of the highest quality and of high social relevance. The CODESRIA mandate is being fulfilled through the establishment of a number of research vehicles such as working groups and networks, and the designing and launching of research programs and projects. Some of these programs and projects are conceptualized and implemented as part of a variety of collaborative initiatives (CODESRIA, 2005).

CODESRIA has made considerable efforts to make its research accessible in Africa and internationally through the use of the internet which has been hailed as a success. The other strategy adopted has been to diversify its dissemination outlets through, for example, cooperation with other institutions for dissemination, such as Librairie Clair Afrique (Dakar, Senegal), University Bookshop Makerere (Kampala, Uganda), Editions Cl6 (Yaounde, Cameroon), Kalila wa Dimna (Casablanca, Morocco), and Mosuro/The Book Seller Ltd (lbadan/Lagos, Nigeria). These outlets are diverse in terms of national language (French in Dakar, Cameron, and Morocco and English in Nigeria and Uganda). These efforts have contributed to considerable improvement in dissemination.

Applicants to the research programs are increasingly linguistically diversified beyond Francophone and Anglophone, with increasing numbers of Arabic- and Portuguese-speaking participants in response to CODESRIA's pledge to increase the number of participants from these countries.

While CODESRIA is pan-African, it has ensured subregional (East, West, Central, South, and North) representation in its organs and activities (Secretariat, Executive Committee, Scientific Committee, research program coordinators, and participants) (Salih and Draman, 2012).

- OSSREA is a regional membership-based and donor-supported research and capacity-building organization whose mission is to promote dialogue and interaction between researchers and policy-makers in Eastern and Southern Africa with a view to enhancing the impact of research on policy-making and development planning. OSSREA's head office is located in the Main Campus of Addis Ababa University (AAU), in Addis Ababa, Ethiopia.

Membership is open to all individuals and institutions engaged in teaching and/or research. OSSREA not only engages in interdisciplinary research but also provides research grants to junior and senior scholars based in the region. It also offers training in research methodology. OSSREA operates major projects either independently or in collaboration with other institutions and produces a variety of scholarly publications ranging from its biannual journal to monographs, books, and research reports.

Its constituency is based on membership drawn from social scientists and institutions engaged in teaching and/or research in the social sciences in Eastern and Southern Africa. Currently, there are 21 member countries in the region where OSSREA has National Chapters and Liaison Offices. Institutional membership is open to all institutions, within or outside the region, engaged in the promotion of research in the social sciences in Africa (OSSREA, 2006).

The following section provides a general perspective on the state of social sciences in terms of funding, education, and research structures at country level, by presenting three 'case studies.'

Case Studies

South Africa

The Academy of Science of South Africa (ASSAf) published a report in 2011 on the state of humanities in South Africa. Some of the main findings in the report are: There is a crisis in the humanities reflected in declining student enrollments, falling graduations, and decreasing government funding (in real rands) within institutions of higher learning. The evolution and administration of government policy in the post-apartheid period has systematically benefited science, technology, engineering, and mathematics (the so-called STEM disciplines) to the exclusion and even detriment of the humanities disciplines in the country. The humanities within institutions of higher learning are in a state of intellectual stagnation and, singular innovations notwithstanding, have remained in this moribund condition for more than 15 years. The decline of the humanities has many causes that include government policy and funding, institutional choices and decision-making, school guidance and counseling, and parental and student preferences. The single most important threat to the growth of an intellectually vibrant scholarship in the humanities is the problem of the aging academic and research workforce, a factor that must be read alongside the evidence of a decline in doctoral graduates in the humanities fields (ASSAf, 2011).

Funding

The total income of the public higher education system is usually derived from three sources of funding, namely (1) government funds, which include subsidy (or block grant) amounts and earmarked funds for special purposes; (2) student tuition and all class fees, as well as accommodation or residence fees; and (3) private income such as donations, investment income, and income from contracts/grants for research or the delivery of other services from agencies other than the South African government.

The rates of growth (in nominal rands) of these income sources differed over a 9-year period from 2000 to 2008. Income from government sources increased at an average annual rate of 8.7%, student fees at an average annual rate of 12.2%, and private income at an average annual rate of 14.2%. Total income increased, in nominal rands, by R18.8 billion (or 137%) in the same period. The government funding share of the income total declined from 49% in 2000 to 40% in 2008, a decline that was absorbed equally by student fees and private income.

About 80% of the government higher education budget is allocated to block grant funds that are distributed through a mechanical subsidy formula. These block grants are described as 'council controlled funds' because they can be spent at the discretion of the council of each university. The remaining 20% of the government higher education budget is allocated as earmarked funding for special purposes. These earmarked amounts include at present the National Student Financial Aid Scheme of South Africa, funds for redeveloping the physical infrastructure of the higher education system, and special project funds such as support for foundation training and clinical training grants. Councils have no discretion as far as earmarked funds are concerned: They must ensure these amounts are used only for their designated purposes. The main drivers of the two government funding formulas that operated over the period 1996–2008 have been student enrollments and the classifications of education subject matter.

During the period up to 2004, the funding ratio between science and technology students and humanities students was 2.67 to 1. This ratio dropped when the new funding framework was introduced in full in 2006. In terms of the new framework, the humanities are divided into fields that have a weighting of 1.0 (law, librarianship, psychology, social work/public administration), those that have a weighting of 1.5 (communication, all languages, philosophy/religion, all social sciences), and those that have a weighting of 3.5 (fine and applied arts). The weightings of fields in the science and technology category are spread between those that have a weighting of 1.5 (computer science), those that have a weighting of 2.5 (architecture/planning, engineering, mathematical sciences), and those with a weighting of 3.5 (agriculture, health sciences, life and physical sciences). Because of these changes in weightings, the average funding ratio between science and technology students and humanities students has moved close to 2:1 (ASSAf, 2011).

Besides normal government subsidies, government funding for social sciences research projects is primarily obtained from

the National Research Foundation (NRF). The NRF was established through the NRF Act (Act No. 23 of 1998), following a system-wide review conducted for the Department of Arts, Culture, Science and Technology. The new entity incorporated the functions of the research funding agencies that were previously servicing various sections of the research community, namely the former Centre for Science Development of the HSRC and the former Foundation for Research Development that included several National Research Facilities.

As an independent government agency, the NRF promotes and supports research in all fields of knowledge. It also conducts research and provides access to National Research Facilities. The NRF receives its mandate from the NRF Act (Act No. 23 of 1998). According to Section 3 of the Act, the object of the NRF is to:

- promote and support research: through funding, human resource development, and the provision of the necessary facilities in order;
- facilitate: the creation of knowledge, innovation, and development in all fields of science and technology, including indigenous knowledge, and thereby to
- contribute to: the improvement of the quality of life of all the people of the Republic (NRF, 2014).

As already indicated under the Sub-Saharan section, external funding for social science centers/institutions conducting various research projects is also common and includes funders such as SIDA, DAAD (German Academic Exchange Service), DFID, USAID (United States Agency for International Development), SANPAD (South Africa-Netherlands Research Programme on Alternatives in Development), etc.

Education

Social sciences are primarily taught at the country's 25 universities and generally involve courses such as development studies, languages, philosophy, religion studies, communication, public governance, historical studies, health studies, political sciences, psychology, social work, sociology, and social anthropology.

South Africa began restructuring its higher education system in 2003 to widen access to tertiary education and reset the priorities of the old apartheid-based system. Smaller universities and technikons (polytechnics) were incorporated into larger institutions to form comprehensive universities (South Africa.info, 2012).

In 2011, two reports (The Academy of Science of South Africa (ASSAf)'s report (http://www.assaf.org.za/wp-content/uploads/2011/08/25-July-Final.pdf) and The Charter for the Humanities and Social Sciences Report (http://www.charterforhumanities.co.za/final-charter-report-released)) on the humanities and social sciences (HSS) in South Africa pointed toward the importance of humanities, while indicating certain challenges, including a decline of student enrollment and financial support since the advent of the country's democracy in 1994. Both of these reports called for a type of national structure that could improve the status and standing of the humanities in South Africa by enhancing scholarship, research, and ethical practice in the field, and advising government.

Subsequently, the National Institute for the Humanities and Social Sciences (NIHSS) was formally established in December 2013 as the result of extensive work toward drafting the 2011 Charter for Humanities and Social Sciences.

The overall mandate of the NIHSS is to develop and set up the institutional and logistical framework for the envisioned higher education institution for the HSS. The role of the institute is, broadly, to enhance and support the HSS in South Africa and beyond, and to advise government and civil society on HSS-related matters. It does so through its various programs, including the Catalytic Projects, the Doctoral Schools, and the African Pathways Programme, and through supporting the DHET in the implementation of the proposed corrective interventions.

Further functions of the NIHSS include:

- providing services to the higher education sphere within the HSS;
- advancing learning within its specific scope or application by ensuring collaboration or collaboration of the work of higher education institutions; and
- advising the minister on matters relating to its specific scope or application, or to higher education generally (NIHSS, 2014).

Research

The humanities (including the arts) and social sciences (including economic sciences) constitute approximately 37% of South Africa's total research output when measured in terms of papers in peer-reviewed journals. Two 'publication cultures' exist among South African scholars: Scholars in the humanities publish predominantly in South African journals, while scholars in the natural and health sciences publish much more in foreign journals. There is growing evidence of increased international collaboration resulting in increased 'visibility' of the country's humanities scholarship. The annual number of monographs submitted to DHET for subsidy initially declined but recently seems to have settled at an average of around 60 titles per year. The vast majority of monographs are in the humanities (82%). Scholars in the humanities fields benefit more from the DHET research subsidy scheme as there are many more journals in these fields (compared to the number of journals in the natural and health sciences) included in the list of accredited journals of the DHET. The humanities fields dominate the total number of doctoral graduates produced (47%); within the humanities three fields dominate in the production of the total number of doctorates: education (23.9%), religion (17.3%), and psychology (10.7%) (ASSAf, 2011).

South Africa hosts a few prominent institutes and research centers working in the social sciences fields – most of which are situated at the universities, but some also being independent of larger academic institutions, e.g., Africa Institute of South Africa, Public Affairs Institute, South African Institute of International Affairs, etc.

The most significant national research body focusing on social sciences research is the HSRC – which was established in 1968 as South Africa's statutory research agency and has grown to become the largest dedicated research institute in the social sciences and humanities on the African continent, doing cutting-edge research in areas that are crucial to development.

The council conducts large-scale, policy-relevant, social-scientific research for public sector users, nongovernmental organizations, and international development agencies. Research activities and structures are closely aligned with South Africa's national development priorities. The HSRC's integrated research programs provide single points of entry – complete with a critical mass of researchers – for interdisciplinary and problem-orientated research in the following areas:

- democracy, governance, and service delivery;
- economic performance and development;
- education and skills development;
- social aspects of HIV/AIDS, STIs, and TB;
- human and social development;
- population health, health systems, and innovation.

The organization is home to about 500 staff members, all dedicated to support the conducting of 150–200 projects per year on social science and humanities research. The HSRC has a strong footprint in other African countries through its affiliation with CODESRIA, participation in SADC and NEPAD, as well as hosting African Research Fellows in the organization. The HSRC also has a global presence and has been chosen to host the 2015 World Social Science Forum. The council's research outputs include reports for users, occasional papers, and scholarly articles in peer-reviewed journals, books, and a news magazine that allows us to improve the public understanding of science. These are disseminated in print through the HSRC Press, through policy briefs aimed at government and policy-makers, and through the media, including social media, the HSRC web site, conferences, and extensive research networks (HSRC, 2014).

Democratic Republic of Congo

The DRC faces immense challenges in delivering education services to all persons of school age in the country (OSISA, 2009). The Congolese government hardly delivers governance and public services, not even basic education services to all children of school age. Dictatorship, economic decline, more than a decade's civil war, and endemic corruption have unsurprisingly left an education system and research program that are catastrophically failing the nation's children, with consequences for the national economy. The most typical problem, however, is simply the absence of specific services (USAID, 2012).

Funding

To overcome the country's educational deficit, a massive increase in the education funding or budget will be necessary to pay for the expansion of teacher training, the development and distribution of teaching materials including books, the construction of new schools, and the improvement of the conditions of work for teachers and lecturers, especially those in remote areas of the country. But the Congolese government would have to take action to ensure that funds allocated to the Ministry of Education are being correctly spent and utilized.

Most efforts by the donor agencies have endeavored to solve one aspect of the challenges faced by the Congolese educational system and ignored the others. Whilst being well meaning, these projects such as the Education Management Information System (in French *Système d'information pour la gestion de l'éducation*) have provided little more than superficial gains and little long-lasting impact. (Recently, a computerized Education Management Information System in Congo was installed – within the Ministry of Education, set up as a project funded by the African Development Bank and UNESCO.) In the past there was also a real lack of communication between the various organizations such as UNICEF, UNESCO, and the African Development Bank involved in providing funds and assistance to the education sector of Congo. This lack of systematic long-term donor engagement and intervention has had an adverse impact on the sector since the millions of dollars allocated by the aid agencies have not been effectively or efficiently utilized. Future assistance projects must now find better ways of dealing with the root causes of the poor educational system in the DRC. The sector needs urgent reform to ensure that key role players provide incentives for the continuation of any assistance scheme whilst appealing to the interest of the learners. Only by adopting such an approach will the education sector in the DRC be able to contribute to the Millennium Development Goals of quality and good education for all.

Education

Much of public education, funding, and research development projects are simply provided through mixed arrangements (for instance, the state appoints the teachers, but it often falls to local communities to pool their resources to pay them, build schools, and provide supplies). In addition, the lack of means is a serious impediment to effective governance of the education system in Congo. Whether in the educational system or administration, insufficient funds prevent proper functioning. Things as mundane as office space, stationery, readings, books, desks, chalk, etc., as well as supplies, transportation, telephone units, and the like, are making effective administration and the delivery of a quality education for all often impossible. Therefore, more often than not, teachers and lecturers come to work but are then largely unable to properly fulfill their intended functions.

Furthermore, it must also be emphasized that staff that are entrusted with the management of bursaries and research funds are severely underpaid and often lack the proper qualifications and education necessary to adequately complete such tasks. Teachers and lecturers at all levels of the education system (from primary school to tertiary education) frequently demand bribes from students or pupils in order to pass a class. As women are the most vulnerable group in Congo, lecturers at public universities and teachers of the primary and secondary schools often propose to a learner a sex practice (whether she is brilliant or not), also known locally as 'Points Sexuellement Transmissives' – which can be translated as 'Marks sexually transmission.'

The qualifications of teachers also leave a lot to be desired. A legal framework is in place to govern the recruitment and promotion of workers in the education sector, including the law regulating the civil service generally, and ordinances establishing the education inspectorate, the administration service for pre-primary, primary, and secondary education, and the administrative staff at higher and university level ((1) la loi No. 81-003 du 3 Juillet 1981 portant Statut du Personnel de Carrière des Services de l'Etat; (2) l'ordonnance-loi No. 91-231

du 15 Aout 1991 portant règlement d'administration relatif au corps des inspecteurs de l'enseignement primaire, secondaire, et professionnel, etc.). These laws are generally in compliance with the relevant international standards and establish rules and regulations for nondiscriminatory and competitive recruitment, though they reflect their age in some respects: for example, by not establishing systems to guarantee representation of women among the administrative and teaching staff. More interestingly, the legal rules are practically unknown among those responsible for management of staff in the education sector, and the principles established are routinely ignored in practice.

The general public's lack of knowledge of the right to proper education and of the law in general, exacerbated by a rooted corruption system, constitutes a reason for the poor management of the education system in Congo, and thereafter a lack of implementation of laws and international standards, in line with the new ambition to establish a democratic dispensation in the country. This situation derives in part from the uncertainty of laws, but also from the limited education of many Congolese, including those who serve in official capacity, and the lack of widely available information and civic education.

Training centers exist in principle for both teaching and administrative staff, at school and university level. Examples include the Primary and Secondary Teacher Training Institute (*Institut de formation des cadres de l'enseignement primaire et secondaire*); the National Training Service (*Service national de formation*); in charge of in-service training for existing staff; the National Institute for Professional Training (*Institut national de préparation professionnelle*), responsible for training administrative staff; and the University Pedagogical Serve (*Service de pédagogie universitaire*). But since the withdrawal of donor supports in 1991, these training centers have been paralyzed due to lack of funds and do not fulfill their mandates for initial and continuing training for school teaching and inspectorate staff, a situation that may explain the large percentage of unqualified teachers in the system.

Teaching quality is further compromised by extremely low salaries, which are irregularly paid, subject to problematic provincial and functional disparities. Poor pay is also matched by poor conditions of work and a lack of resources in the schools and universities. Strikes are frequent in the sector as a result, especially at the outset of each academic year. Absenteeism is also a serious problem, with teachers simply failing to turn up to their posts on a regular basis (as a livelihood strategy, most of them in the provinces of mineral-rich areas find themselves as artisanal miners). In addition, it seems that there is minimal enforcement of ethical rules, sanctions are almost never applied even for gross misbehavior, let alone for poor performance, and the codes of conduct exist on paper only. As a result, most schools and universities are run by churches and private mining companies, and parents are expected to pay fees that are usually higher than their income.

Research

The DRC has some of the world's most outdated infrastructure for education, training, and research and development. A lack of detailed information about the distribution of teaching staff in Congolese schools is just one of the results of the generally poor system of information collection and management. However, it is fair to state that government expenditure on teaching and research is extremely low, and, as a result of widespread corruption, officials usually misuse the funds. Even though universities such as Universities of Kinshasa, Lubumbashi, and Kisangani have research centers (mainly in major cities), and schools exist in every city (see **Figure 1**), most of them are old and lack basic resources such as libraries, laboratories for research, and trained teachers.

Finally, the research funding process, appointment of teachers/lecturers, and granting of bursaries (if any available) that are intrinsic to the development/management of the education system are also vulnerable in that the employees of the Ministry of Education are easily manipulated by higher level government officials. Indeed, many of the employees working at the Ministry of Education are hired not on the basis of merit, but on their personal connection to relatives or friends in power. Because of the existing power dynamics, employees at the Ministry of Education also face an impossible situation where they cannot adequately follow normal procedure and regulations in appointing teachers or granting bursaries. Certain employees are paid by applicants under the table as part of an individual's effort to circumvent the process.

To sum up and as outlined above, the education and research development standard in Congo is low, and would require urgent fundamental reforms to spur the country's economic and social development. There is also a need for complete reform of the recruitment, training, and performance review systems for teachers and lecturers to restore a sense of professional pride among staff working in the sector, and ensure that those responsible for teaching the nation's children and youth are properly qualified for the job. Among measures needed in this regard is the revitalization and adequate funding of existing training colleges for teachers.

Ethiopia

Ethiopia is the second most populous country in Africa, next to Nigeria, hosting over 90 million people (CSA, 2007). According to 2011 DHS, Ethiopia has over 80 ethnic groups with diverse social, cultural, and linguistic identities. Although the service sector and industry have been supporting the economy recently, most of the population makes a living out of mixed agriculture. The total land surface is about 1.1 million square kilometers and shares borders with Djibouti, Eritrea, the Republic of the Sudan, the Republic of South Sudan, Kenya, and Somalia (CSA, 2011).

The history of modern formal education in Ethiopia is not more than a century old while informal education has been practiced in Ethiopian Christian religious institutions since the introduction of Christianity in the fourth century (Asgedom, 2005). Informal education in Islamic scholarship institutions has been provided in Ethiopia since 890 AD. In 1908 the first modern school, Menelik II primary school, was established and after 1925 education was recognized by the imperial government as a key tool for development and modernization. The traditional church and Islamic education institutions and scholarships are now competing with the modern educational establishments (Chaillot, 2009; Lule,

Figure 1 State of public schools in a remote area of the DRC – Oriental Province of Congo. Photo by Jean Didier Losango Nzinga, September 2012.

2013b). The government designed sector-specific program action plans in the education sector divided over 5 years in the past two decades (ESDP I–IV) with the intention of decentralization of management and sustaining educational finance (Yigezu, 2013).

Funding

In 1926, the imperial government issued a proclamation to collect 6% tax on all imports and exports reserved a budget to finance education in the country. In 1930, the government instituted the Ministry of Education and Fine Arts, with an additional budget of 2% of the national budget. Primary and secondary education remained free while a cost-sharing mechanism was introduced in 2003 for higher learning institutions.

The government of Ethiopia employs a historical-based budgeting on incremental bases of the previous years of performance, supplemented with the country's economic performance, government revenue, inflation rates, or institutional growth until 2011/12; a program-based budgeting was introduced (MOE, 2010; Yigezu, 2013).

One of the critical steps in financing education was increasing the ratio of the education sector finance against the national budget, which stood at 3.8% in 1995/96 and increased to 13.7% in 1999 under ESDP I. The Ministry of Education increased the budget allocation to over 22.3% of the nation's annual budget in 2012. ESDP IV encourages respective national higher education institutions to involve in generating resources by providing community services, consulting, and need-based training. Among these, higher learning education shared 24.5% (in 2009/10) and 21.4% (in 2012/13) of the education sector budget (MOE, 2005, 2010, 2013).

In terms of financing research from external sources, Ethiopia draws from multiple sources of funding, the same as South Africa, while other African countries pool resources from one, two, or three sources (Jones et al., 2007). The introduction of the long-awaited cost sharing of education intends to relieve the burden of public financing of education, while it partly motivates higher education institutions to have independent sources of funding education (Yigezu, 2013). For the students the costs of education have been free, as well as free nonacademic service, food, lodging, and health services.

Education

Higher learning institutions drive the basic strategic guidance from the nationally mandated Ministry of Education and its affiliate institutions such as the Higher Education Strategic Center and Higher Education Relevance Quality Agency. Higher education institutions in Ethiopia are intended to provide the nation's human capacity in education, research, and community services (FDRE, 2009; Yigezu, 2013).

The social and behavioral sciences in Ethiopia encompass different branches of fields of studies such as anthropology, archaeology, education, indigenous studies, language, political sciences and international relations, business and economics, gender studies, sociology, social work, history, geography, psychology, communication and journalism, philosophy, development studies, African studies, theatrical arts, and peace and security studies.

AAU is the oldest higher educational establishment in Ethiopia. AAU started its operation in 1950 under the name University College of Addis Ababa. It was renamed Haile Selassie I University in 1962 and then Addis Ababa University in 1975.

AAU runs Diploma, Bachelors, MD, DVM, Masters, Specialty Certificate, and PhD degree programs. It launched its first MSc programs in 1979 and its first PhD programs in 1987 (ESSSWA, 1998, 2014).

The expansion and development of higher education has been struggling with the political will as the state is the main financial source of funding for education, research, and community services (Yigezu, 2013). The rapid transformation in higher education in Ethiopia started in 2000 when it shifted from two public universities to 22 in 2010 and currently has 31 fully furnished public universities, with a total student population of over half a million (2012) from less than 5000 in 1996/97, and 31 997 in 2004/05 (MOE, 2010).

Since the turn of the millennium, higher learning institutions witnessed expansion and diversification to ensure equity and open access education in Ethiopia (Yigezu, 2013). The FDRE government in its proclamation 650/2009 reassured higher education institutions with the mandate to programs in regular, extension, distance, virtual, and continuing education. The proclamation further indicated that all the training at higher learning institutions use English as the medium of instruction. Teaching and learning at higher education institutions in Ethiopia is organized according to academic units of colleges, departments, faculties, schools, and institutes. These are independent units leading teaching and research, including developing the curriculum (FDRE, 2009).

In higher learning institutions, curriculum preparation rests on the academic units. The design is to enable citizens with scientific knowledge, independent thinking skills, communication skills, and professional values to also be allowed to receive training in becoming a competent professional. Curriculum development considers in its development the delivery mode to be research and study based, while keeping its content up to date (FDRE, 2009). The common trend in curriculum change or revision is after graduating the first cohort of students admitted to a program (Lule, 2013a).

In 2008/09, the government introduced a guideline on national university intake and preparatory education, especially on admissions to the science and technology, and social sciences and the humanities streams. ESDP IV has a section on determining the student population of higher education, cost sharing, and direction of tertiary education in Ethiopia, based on the recommendation of the ESDP III on relevance of the academic programs. The university intake nationally was regulated so that 30% of the students were placed in the social sciences and humanities while 70% were placed in the fields of science and technology. In 2008/09 the ratio was 52:48 and planned to be 70:30 in science and technology to social sciences and humanities. The ratio has been even more social and human sciences dominant in the earlier years than science and technology fields (MOE, 2005, 2010).

Private higher learning institutions are classified as major stakeholders in education and research. In 2010, the private sector colleges and universities enrolled about 18% of the student population enrolled to higher learning institutions. They are restricted on providing training for teaching staff as of 2005/6 as the government decided to do it through public institutions on grounds of quality. Most of the students in higher learning institutions in the private sector are in social and human sciences (Yigezu, 2013).

Research

Irrespective of the expansion in higher learning institutions in Ethiopia most of the publications until 2007 over a 20-year trend (1987–2007, 517 publications in the science citation index) came from AAU (Mouton, 2010).

In 2009/10 a massive move was made by the government as 13 new universities were inaugurated with a heavy investment in teaching and research facilities such as machinery, laboratories, chemicals, computers, classrooms and office facilities. All facilities highly focused on accessories for science and technology including simple items such as reference books. Among the 22 universities, five are intended to evolve as centers of science and technology (MOE, 2010; Tefera, 2007, 2013).

The current trend in ESDP IV, which will end in 2014/15, focused on consolidating reforms, on system expansion, and on change in governance. This will help improve quality and employability of university graduates, and sustainably develop the research capacity for knowledge creation while technology transfers will be harnessed in priority sectors. It prioritized on diversifying the funding base and mechanism through income-generating schemes and enhancing university industry linkages partnerships (MOE, 2010; Tefera, 2013; Yigezu, 2013).

As a pioneer institution in higher learning education, research, and community services in Ethiopia (social and behavioral sciences included), AAU has instituted a mandated office of the vice president for research and technology transfer (VPRTT). The department is one of the core pillars of the institution mainly in charge of initiating, planning, executing, following up, publishing, and disseminating research along the country's development priorities (AAU, 2013a). The VPRTT in 2013 reorganized institutionally, procedurally guided, and issued regular calls for research funding. These include:

1. national research priority areas funding (AAU, 2013b),
2. thematic research calls for funding (AAU, 2013c),
3. individual small grants research calls (AAU, 2013a),
4. sponsored research projects.

Unlike the earlier years, the VPRTT has conducted consultative meetings with stakeholders on identified research priorities and thematic research issues. Subsequently, the office has published its research manual, research priority areas, and thematic research guides. However, the involvement of the social and behavioral sciences in the thematic research, priority areas, and individual grants is very limited compared with other fields of studies (AAU, 2013b).

The Ministry of Education sets out procedures and strategic directions in education, research, and funding at all levels of education. The higher education institutions have vice presidents for research and technology transfer, who handle thematic research, priority identified research, individual research grants, and small grants for individual researchers based in their respective institutions (AAU, 2013a; JU, 2012).

External funding of research projects is also commonly secured by staff members of the respective institutions in teams and individually. There are stipulated incentive mechanisms for individual staff members who got grants from external sources of finance (JU, 2012).

Organizing annual research conferences for the dissemination of the respective university-wide research and publication

is trending at AAU, University of Gondar, Jimma University, Mekelle University, Bahir Dar University, and Dilla University. The respective academic units are establishing journals to publish their research work at college or university commonly university wide.

See also: Africa, Sociocultural Overviews: West Africa; Central Africa: Sociocultural Aspects; East Africa: Sociocultural Aspects; Education in Sub-Saharan Africa; Islam: Sub-Saharan Africa; Nationalism, Historical Aspects of: Africa; Southern Africa: Sociocultural Aspects; Urbanization in Africa.

Bibliography

AAU, 2013a. Research Manual. Addis Ababa University Press, Addis Ababa.
AAU, 2013b. Research Priority Areas. Addis Ababa University Press, Addis Ababa.
AAU, 2013c. Thematic Research Areas. Addis Ababa University Press, Addis Ababa.
Asgedom, A., 2005. Higher education in pre-revolution Ethiopia: relevance and academic freedom. The Ethiopian Journal of Higher Education 2 (2), 1–46.
ASSAf (Academy of Science of South Africa), 2011. Consensus Study on the State of the Humanities in South Africa: Status, Prospects and Strategies.
Chaillot, C., 2009. Traditional Teaching in the Ethiopian Orthodox Church: Yesterday, Today and Tomorrow.
CODESRIA (Council for the Development of Social Science Research in Africa), 2005. CODESRIA Research Policy.
CSA, 2007. Population and Housing Census. Addis Ababa.
CSA, 2011. Demographic and Health Survey. Addis Ababa.
ESSSWA, 1998. Ethiopian Society of Sociologists, Social Workers and Anthropologists Founding Bulletin.
ESSSWA, 2014. http://www.essswa.org.et/Links.html.
FDRE, 2009. Proclamation No. 650/2009. Higher Education Proclamation, Addis Ababa.
HSRC (Human Sciences Research Council), 06 September 2014. About. Web: http://www.hsrc.ac.za/en/about (accessed 06.09.14.).
JU, 2012. Guidelines and Procedures for Research. Office of the Senior Director for Research, Community Based Education and Graduate Studies. Jimma University.
Jones, N., Bailey, M., Lyytikainen, M., 2007. Research Capacity Strengthening in Africa: Trends, Gap and Opportunities. Overseas Development Studies, UK.
Lule, S., 2013a. A generation Sociological Training Propects and implications for Social Development in Ethiopia. (Under Review for ISA e-symposium).
Lule, S., 2013b. Professional Competence of Sociologists, Social Workers and Anthropologists in Ensuring Child Protection against Sexual Abuse and Exploitation.
MOE, 2005. Education Sector Development Program III-2005/6/–2010/2011 Program Action Plan. Addis Ababa.
MOE, 2010. Education Sector Development Program IV (ESDP IV). Addis Ababa.
MOE, 2013. Education Statstics Annual Abstract. Addis Ababa.
Mouton, J., 2010. The State of Social Science in Sub-Saharan Africa. 2010 World Social Science Report: Knowledge Divides, Background Paper. International Social Sciences Council.
NIHSS (National Institute for the Humanities and Social Sciences), September 06, 2014. Overview and Mandate. Web: http://www.hssi.org.za/content/overview-and-mandate (accessed 06.09.14.).
NRF (National Research Foundation), September 06, 2014. About NRF – Corporate Overview. Web: http://www.nrf.ac.za/about-nrf (accessed 06.09.14.).
OSSREA (Organization for Social Science Research in Eastern and Southern Africa), 2006. About OSSREA. Web: http://www.ossrea.net/index.php?option=com (accessed 05.09.14.).
The Open Society Initiative for Southern Africa (OSISA), 2009. The Democratic Republic of Congo: Effective Delivery of Public Services in the Education Sector. A Discussion Paper. OSISA, Johannesburg, SA.
Ranking Web, 07 September 2014. Webometrics Ranking of World Universities: Sub-Saharan Africa. Web: http://www.webometrics.info/en/Sub-Saharan (accessed 07.09.14.).
Salih, M., Draman, R., 2012. Evaluation of the Council for the Development of Social Science Research in Africa (CODESRIA).
Sall, E. 2010. Council for the development of social science research in Africa (CODESRIA). World Social Science Report: Knowledge Divides, Chapter 1: Social sciences facing the world. Co-publication commissioned by UNESCO from the International Social Science Council (ISSC).
South Africa.info, 2012. South Africa's Universities. Web: http://www.southafrica.info/about/education/universities.htm#.VAxj2mOKm_w (accessed 07.09.14.).
Tefera, D. (Ed.), 2013. Funding Higher Education in Sub Saharan Africa. Palgrave Macmillan, UK.
Tefera, D., 2007. Building Research Capacity in Ethiopian Universities the Realities and the Challenges. AA.
Tilak, J.B.G., 2011. Financing higher education in sub-Saharan Africa. Africanus 41 (2), 4–31.
The United States Agency for International Development (USAID), 2012. Democracy, Human Rights, and Governance Assessment of the Democratic Republic of Congo. Final Report Prepared by Tetra Tech ARD. USAID, Burlington, Vermont, USA.
The World Bank, 2010. Financing Higher Education in Africa. World Bank, Washington, DC.
Yigezu, M., 2013. Funding higher education in Ethiopia: modalities, challenges, opportunities and prospects. In: Tefera, D. (Ed.), Funding Higher Education in Sub Saharan Africa. Palgrave Macmillan, UK, pp. 38–70.

Social Science Infrastructure: Asia

Victor R Savage, National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While Asian governments are increasing their budgets for research and development (R&D), most financial outlays are mainly in applied areas in engineering, science, technology, and infrastructural fields. Although the 'soft' areas of R&D in social, cultural, economic, and political sciences have traditionally been relegated to university research centers, they are now gaining traction in private sector institutions and government-backed national think tanks. Given the many challenges Asian communities face (inequality of wealth, poverty, food insecurity, unemployment, domestic political instability, etc.) the importance of social and cultural R&D is likely to grow and become influential in shaping national and geopolitical policies in future.

Introduction

The main English daily in Singapore, *The Straits Times* featured a large advertisement inviting the public to a conference on 'Liveability in Singapore: Social and Behavioral Issues' in February 2014. The conference embodied three uniquely Asian trends in the social services: (1) such themed social conferences are generally rare in Asia compared to science, technology, engineering, and business conferences; (2) they are often organized by universities or national think tanks as the Singapore case (Singapore Management University) shows; and (3) the speakers are generally local nationals of high public visibility (namely Professor Tommy Koh, Laurence Lien, and Gerard Ee), as this conference demonstrates.

In Asia, academic institutions, public think tanks, private research organizations, and multinational corporations have had a relatively recent engagement with research and development (R&D). However, since the post–World War II period, R&D began to develop, and it expanded rapidly in the 1990s onwards. By and large, Asia has played a subservient role in research, science, and technology as compared to West. Over the past 500 years (sixteenth to twentieth centuries), Western science and research have dominated the global intellectual landscape (Gribbin, 2003). Despite earlier mathematical and scientific contributions from the Middle East, South Asia, and China (Arun, 2006), the main research breakthroughs in mathematics, science, and technology have emanated from the West – Europe and North America (Gribbin, 2003).

The overwhelming Western dominance of the global intellectual landscape was the major thrust of Harvard Professor Francis Fukuyama's (1992) thesis, "The End of History." Using a Hegelian historical paradigm, he argued that the West's dominance in all fundamental aspects of intellectual thought (science, technology, and ramifications in politics and social sciences) meant there was no challenge or antithesis, and hence the Western 'scientific thesis' has prevailed. Yet as Asian states rise to power (economically), one wonders whether this emergence will be buttressed by innovative, creative, and entrepreneurial intellectual endeavors in a sustained manner. In the book on *China's Century*, Brahm (2001) clearly missed the importance of intellectual and research contributions in assessing China's economic rise, political reform, and social engineering. In the last decade, however, Asian states have enhanced and expanded R&D. In 2011, Malaysia was home to 73 752 researchers, a threefold increase since 2006 (19 021 researchers), most of whom work in science and technology. Indonesia's Institute of Science (LIPI) consisted of 8000 researchers in 2013. Although focused on the hard sciences, the Institute also conducts research on social issues and history. Recent global studies of secondary school students' academic performance in science, mathematics, and problem solving have demonstrated that Asian students (South Korea, Singapore, Taiwan, Hong Kong, and Japan) are progressing much better than their counterparts in developed Western countries.

Given the quest for economic development in the developing countries through industrialization, R&D in Asia from the 1950s onwards was focused on technology, science, energy, transportation, and infrastructure hardware. Compared to science and technology, in the social and behavioral sciences, R&D assumed a marginal role in government plans and policies. Except for economics and literature, the bias toward the hard sciences is evident in the Nobel Prize categories as well as comparative data in prestigious journals, like *Nature* and *Science*. Yet, one must accept that science and technology is no one-size-fits-all solution to human problems. Feyerabend's (2013) critique of the tyranny of the ideology of science and Sheldrake's (2012) argument of the 'science delusion' are pertinent reminders of why diverse forms of R&D including the humanities and social sciences need to be given weightage for practical and efficient quotidian living. The rapid speed at which cultural changes, political movements, and social transformations are taking place in Asia underscores even more so an understanding of social and behavioral responses to change.

Given the 2012–13 Arab spring, the political turmoil in Thailand, political volatility in India, Bangladesh, and Pakistan, the challenges to hitherto stable ruling governments in Cambodia, Malaysia, and Singapore, and rising disparities of wealth and growing unemployment have all demanded the need for greater social science R&D. What is challenging for academics, politicians, bureaucrats, and corporate titans is trying to decipher, analyze, and comprehend political trends, cultural expectations, social explanations, and economic impacts. This can only be aided by greater investments in R&D in those areas. The Chinese Academy of Social Sciences is already demonstrating the value of such social research when it notes

declining levels of job satisfaction and the falling confidence in social welfare in China (Fenby, 2014: p. 82). In a world in which soft power and smart power (Nye, 2004, 2011) are required to maintain good national and international relationships, the importance of social and behavioral sciences, and the humanities becomes increasingly pertinent.

At a global level, there is also a bias in social science research toward developed countries. The nature of research is thus biased toward strong economies rather than toward developing, weak economies. For example, between 1985 and 2005, of the 76 000 papers in economics, over 36 000 papers focused on the United States – more than the combined papers on Europe, Asia, Latin America, the Middle East, and Africa (*The Economist*, 2014: p. 56).

Drivers of Social R&D

In Asian states, the nature of research funding and institutional R&D has been uneven due to three factors: (1) the general wealth of the country, (2) the level of development of tertiary education, and (3) the economic orientation of national development. Except for the more developed countries and political entities (Japan, South Korea, Taiwan, Hong Kong, and Singapore), the percentage of GDP allocated to research is relatively small: 0.08% in Indonesia, 0.25 in Thailand, and 0.63 in Malaysia. A relatively rich country such as Malaysia had 365 (2006) researchers per million inhabitants compared to Cambodia's 17 (2002). These percentages and figures are even more minuscule when dealing with social and behavioral research. Richer, more dynamic Asian countries (South Korea, Taiwan, China, and Singapore) have, however, increased the number of researchers from 16% in 2003 to 31% in 2007 (R&D Magazine, 2011: p. 30).

In Asian states, the drivers of social, cultural, and behavioral research have been an outcome of four factors and influences, as discussed in the following sections.

Colonial Research Societies: Sociocultural Statements

Historically speaking, colonialism in many ways sets the foundations and institutional base of social and behavioral R&D in Asia. Social science research in Asia was largely pioneered by colonial academic and quasi-academic societies. The British, Dutch, and French colonial powers initiated 'research societies,' which shed light on the culture, language, customs, behavior patterns, religious beliefs, folk medicines, and social practices of the peoples within their Asian colonial territories. Before Western colonies developed universities and higher educational institutions, colonial research societies, supported mainly by colonial civil servants and Western entrepreneurs, pioneered social and cultural research in Asia.

In Britain, the renowned Sanskrit scholar Henry Colebrooke established the *Royal Asiatic Society* (1823) to investigate subjects "for the encouragement of science, literature and the arts in relation to Asia." From its British origins, the Royal Asiatic societies then spread around into India (five subbranches), Sri Lanka, China (Hong Kong, Shanghai, Suzhou, and Beijing), Japan, Korea, and Malaysia. Like their parent society in Britain, each of these affiliated societies established their own journals, which later became the research journals for science, arts, and social issues across Asia. In Malaysia, *The Journal of the Malaysian Branch of the Royal Asiatic Society* (JMBRAS) still continues to publish quasi-academic articles of varied contributors, academics, bureaucrats, and the public. In some ways the development of colonial studies on indigenous cultures, religions, languages, and societies provided the Western academic canon in defining Asia and its cultural diversity. Through the Western cultural filters, Asia became for centuries a cultural antipode for the West (Steadman, 1969), best enshrined in Edward Said's (1994, 1985) 'ism' of 'Orientalism' and it's reverberations in Huntington's (1996) 'Clash of Civilizations,' and Buruma and Avishai's (2005) 'Occidentalism.' Said's thesis and Huntington's 'clash' of East–West cultures provided new impetus in understanding social, economic, and cultural issues in Asia both from Western perspectives and their Asian research centers.

Of the colonial powers, the Dutch were the most active in economic, social, and cultural research both at the 'national' and at the regional–provincial levels in Indonesia. Besides the flora and fauna of the islands, the plethora of Dutch studies on Indonesian societies is unparalleled in other Western colonies. Some of these Dutch scholarly works in social science however were published in English as books and journal articles (Bertram Schrieke; Jacob van Leur; W.F. Wertheim; Justus van der Kroef) (King, 2008). Since most American scholars could not read Dutch, a new wave of studies on Indonesian society was pioneered by American scholars like Clifford Geertz, Karl Pelzer, Margaret Mead, William Withington, Karl Jackson, and Benedict Anderson from the 1950s onwards, putting Indonesia's cultural landscape on the world stage.

National Narratives: Defining Community and State

After World War II, newly independent Asian states embarked on national self-assessments and developmental trajectories to boost nationalistic pride. After centuries of colonialism, many Asian countries desired postcolonial indigenous paths to development and modernization. Interests in indigenous knowledge, vernacular architecture, and folk science underscored postcolonial research agendas in many Asian states. While academics labeled these indigenous social and cultural developments under various other terms (e.g., postcolonial, postmodern, folk science, and indigenous knowledge), former Malaysian Deputy Prime Minister Anwar Ibrahim (1996) referred to these developments in the 1980s as the 'Asian Renaissance.'

Social science research, which underpinned university social science research agendas, was focused on defining nationalistic objectives of new states. However, many current Asian states were initiated and formulated under colonial rule. Benedict Anderson (1991) argued that 'three institutions' (census, the map, and the museum) were invented by the colonial government that shaped the colonial state; with independence all three 'institutions' continued. The taking of population censuses has continued every 10 years in most Asian states (India, Malaysia, and Singapore), albeit intermittently in others (Myanmar), as a means of collecting data on population growth, gender, age, and spatial distribution and the ethnic,

cultural, and religious profiles of citizens. The first comprehensive, Burmese census was conducted in 1931, and the next in 1983 (as an independent state). The Myanmar government in 2014 is conducting the census with the hope of shedding light on its 135 ethnicities (minus the Rohingyas) and the standard of living in the country.

In most countries, the Department of Population Census or National Statistics provides some basic analysis of the census data. However, in most cases, university faculty members make the most use of the data in social science research (economics, sociology, geography, psychology, social work, demography, and political science). In Singapore, the *National Population and Talent Division* is a key generator of population research. Its 2013 White Paper on "A Sustainable Population for a Dynamic Singapore" drew much public criticism for setting its population goal of 6.9 million by 2030. Given India's large population, the International Institute for Population Science does a lot of demography-related research. The Institute completed seven research projects in 2008–09 and has currently 19 ongoing research projects dealing with family health, child health, youth issues, and aging in both villages and urban areas. In Japan, the National Institute of Population and Social Security Research produces population projections and social security surveys especially with regard to Japan's aging population.

The colonial development of museums in the late nineteenth century became institutions defining the historical development of colonially administered societies, communities, and the states. Museums were meant to provide the institutionalized historical narrative of social entities, ethnic groups, and cultural communities, and remained important channels for defining and documenting colonial state histories. State museum narratives, however, provide subjective perspectives, especially when geopolitical conflicts and domestic incumbent governments want such narratives to suit their political agendas. Unlike research materials and publications by academics, museum exhibitions remain, in themselves, national, cultural, and social statements requiring study. In many Asian countries, museums are also sources of research publications (journals, newsletters, monographs, and books).

Maps were drawn to delineate state borders, natural resource locations, transport networks, and the physical geography of states. Mapping was politicized because maps demarcated territories and political jurisdiction. The colonial state utilized maps to define its territories and borders and thus institutionalized the colonial state. Thongchai (1994) provides the best testimony of how maps were used in the colonial period to define expanding British and French territories in Indochina at the expense of Thai national territory. Maps remain a highly contentious issue in defining Asia's many ongoing border wars (China, Japan, South Korea, and Southeast Asian states over the South China seas). Given the geopolitical significance of maps, such research has always been found in strategic institutions in government, defense ministries, and military establishments in India, China, Japan, Malaysia, Indonesia, and Singapore. Maps are visual–graphic expressions of social science data (political and geopolitical representations; economic, social, religious, and cultural issues), which are products of census data, national surveys, or academic research projects.

'Third World' Development Programs

Social science research continued in Asia after the colonial period, supported and abetted by Western institutions in Asia, Europe, and North America. Contributions of the Needham Institute in Cambridge (established in 1948) to understanding the developments of science and technology and social behaviors in China are undeniable (Temple, 1986). Under the vast academic umbrella of development studies, Western scholars expressed their interests in the cultural and behavioral aspects of Asian communities. Governance and labor management were important issues for manufacturing ventures and investments. But the hard sciences took precedence over the soft sciences in universities, polytechnic institutes, and state institutions as industrialization became an important economic driver.

Much of the social science research on Asian societies was carried out in universities and colleges. Universities in the Philippines, India, Indonesia, Malaysia, Myanmar, and Singapore formed faculties of Arts and Social Sciences. Research in sociology, psychology, economics, geography, and political science based on the Western approach was also established. In the social and behavioral sciences, however, Western social science theories, concepts, and models were Eurocentric but they continue to underpin academic thinking in many former Asian colonies.

Asian social research carried out in state and private universities often used Western benchmarks, models, and norms. The Western social systems were often considered the norms and Asian sociocultural representations were often seen as aberrations of the Western models. This was particularly evident in Asian social science research dealing with the pathways to development – where Weberian-type negative logic (e.g., cultural, religious, and social impediments; corruption; old-fashioned customs; and traditions) was often employed in explaining why 'Third World' Asian states could not develop (King, 2008). Western social science theories continued to explain why the developing world remained undeveloped due to psychological factors (McClelland, 1961). According to Austrian economist Erik Reinert (2007), it was David Ricardo's 'comparative advantage' economic concepts, institutionalized in Asia's colonial states, which set back development (weak industrialization and poor R&D) in Asia.

Enter Big Data Analysis

One of the major changes, which augured well for social science research was the breakthrough in the 1990s. Specifically, when IBM entered into statistical machine translation, soon social media like Google and other Internet portals created an explosion of social data. In Indonesia, 'big data' technology is important for large corporations (transnational corporations) where 70% of companies interviewed noted piloting the technology for 2014 (Sudev, 2014: p. 14). In an attempt to make sense of the politically and socially changing landscape, many Asian governments set up research centers to collect, document, organize, analyze, and store national data. With some of the biggest populated countries in the world (China, India, Indonesia, Pakistan, Bangladesh, and the Philippines), big data or "datafication" (Cukier and

Mayer-Schoenberger, 2013: p. 29) will make social science research an easier exercise in the decades to come.

Big data analysis provides three important contributions to social and behavioral R&D: (1) it helps to provide correct social readings of public sentiments, especially when there is evidence that cognitive biases delude us; (2) without big data analysis it is difficult to collect the 'right data' to make sense of public social media issues that tend to be massive, amorphous, uncontrolled, and difficult to read; and (3) in an information society, "data take center stage" (Cukier and Mayer-Schoenberger, 2013: p. 39). Big data allows for bigger samples of information on all aspects of human life. Digitization has transformed data into various purposes, social trends, and other behavioral patterns. In Asia's fledgling democracies, big data has mixed government responses – many governments try to contain Internet searches and muzzle e-mail chatter and blogs. But governments in a new digital age are also interested in establishing national identities, and the net benefits of connectivity provide a better quality of life, economic opportunities, enhanced creativity, and greater efficiency (Schmidt and Cohen, 2014).

The Asian R&D Centers

Traditionally, Asia has in many ways been engaged in social and humanistic research.

Tertiary Institutions

The rise of Asian economies such as Japan, China, India, South Korea, Hong Kong, Singapore, Taiwan, Thailand, Malaysia, Indonesia, and Vietnam since the 1960s has significantly changed the research landscape, enhancing tertiary education and state funding in R&D. Much of Asian research has emanated from tertiary institutions that have grown tremendously over the past three decades. For example, in 2008, Thailand's tertiary landscape included 78 public universities and 89 private higher education institutions with more than 2.2 million students.

The development of R&D in Asia depends in part on the development of tertiary education (universities, polytechnics, colleges, specialized institutions, etc.). Three features undergird tertiary institutions in Asia. Firstly, most Asian countries have universities or colleges that are (1) public (government) funded; (2) privately funded; and (3) overseas universities operating in Asian countries.

The earliest indigenous Asian university is the seventh century Nalanda University in Bihar. Given its historical importance, Nalanda is being reconstructed by India and other Asian countries (China, Thailand, and Singapore) to capture the rise of Asia in the twenty-first century. Under colonial rule, the earliest established university in the Philippines was under Spanish rule in the sixteenth century. Both the indigenous and colonial universities were based on subjects dealing essentially with the humanities, mathematics, languages, and religious studies.

Unlike the West and its emphasis on the 'individual,' in most Asian communities and societies greater attention is given to human nature harmony and social relationships in the family, clan, village, and community (Nisbett, 2003). Asian communities have a long tradition in trying to understand social and behavioral issues as expressed in their philosophies, religious tenets, and folk culture (Peterson, 2001).

One index of top research universities or institutes in Asia is the Nature Publishing Index. Asian research is clearly underrepresented globally – only two of the Asian research centers (Chinese Academy of Sciences and The University of Tokyo) ranked globally within the top 10 in 2013. Based on the number of publications in *Nature* in 2013, Japanese universities make up 50% (5), Chinese universities 30% (3), and Australia and Singapore 10% of the top 10 Asian research institutes or universities. Given that *Nature* publishes both hard and soft (social) science articles, Asian social science researchers can find a global voice; although this is currently lacking.

Asian universities still remain important behavioral and social research centers. Funding generally comes from state governments and in many of the more socialist countries (India, China, Myanmar, Singapore, Cambodia, Lao PDR, and Vietnam), state universities act as research arms of the government. The social and behavioral research areas vary from country to country and over time as well. In the rapidly aging countries and political entities in Asia (e.g., Japan, China, Taiwan, South Korea, Hong Kong, and Singapore) research on aging populations is of prime interest with wide ramifications in health, housing, economic dependence, and mobility. China has a whole research entity dealing with aging: China National Committee on Aging.

Within Asian Universities, several centers have contributed to social and behavioral research. One prevalent research arena lies in Area or Regional Studies (Asian and Southeast Asian Centers) found at the Centre for Asian Studies (University of Hong Kong), Center for Southeast Asian Studies (Kyoto University), Institute of Southeast Asian Studies (Singapore), and Institute of Southeast Asian Studies (Hanoi).

Many universities in Asia have major research centers in broad social, economic, political, and cultural issues. In Malaysia, Universiti Kebangsaan Malaysia (UKM) was established in 1995. The Institute for Malaysian and International Studies looks at policy issues in social transformation and the impacts of globalization on society. Likewise at Seoul National University, The Center for Social Studies undertakes interdisciplinary, social research in new areas on how societies adopt democratization, globalization, and reunification. Many Japanese Universities also have good social science research centers. One of the more recent research centers is Tohoku University's Center for the Study of Social Stratification and Inequality (CSSI) established in 2003, which tackles three key problems: (1) empirical evidence of the extent and types of disparity; (2) looking at theoretical problems and mechanisms of rising disparities; and (3) clarifying the social impact and providing policy recommendations for the impacts.

Private Research Organizations, Nongovernment Organizations

Many Asian research centers that deal with social, strategic, and political policies are either independent or linked to government-funded institutes. These include the Japan Institute of International Affairs (Tokyo, Japan), Atma Jaya Research Centre

(Jakarta, Indonesia), LIPI Jakarta, Centre for Policy Research (Penang, Malaysia), Institute of Philippine Culture (Quezon City, the Philippines), and Institute of Policy Studies (Singapore).

Many private research organizations are on their way to becoming nongovernment organizations (NGOs) with funding from the government and local private institutions or overseas agencies and NGOs. Unfortunately, in many poorer Asian states, government aided research centers, privately-based social and environmental research organizations, and NGOs cannot maintain objective social–cultural interventions.

Among the most important private organizations undertaking increasingly independent social, economic, and environmental research are the Asian Development Bank (Manila, the Philippines), the Mekong River Commission (located in Vientiane, Lao PDR), and the World Bank (Jakarta, Indonesia). In India, the Reserve Bank of India underwrites the Indira Gandhi Institute of Development Research (IGIDR). The IGIDR covers a wide gamut of research agendas in economics, finance, energy and the environment, urban planning, and the social and organizational issues of development. The United Nations also maintains several agencies in Asia, which undertake social and behavioral research. Of all the Asian states, Cambodia, the Philippines, and Thailand are home to the most international aid offices (American, Australian, Canadian, German, Japanese, and Swedish). Nearly all of the 27 United Nations specialized agencies are concentrated in a few Asian countries: Cambodia, India, the Philippines, Korea, Thailand, and Vietnam. These United Nations agencies – the United Nations Economic and Social Commission for Asia and the Pacific, the United Nations Population Fund, the United Nations Project Office on Governance, the United Nations High Commissioner for Refugees, UN-HABITAT, the United Nations Educational, Scientific and Cultural Organization, the United Nations Environment Programme, the United Nations Children's Fund, and the United Nations Institute for Training and Research – undertake national and transnational research programs on demography, food security, poverty reduction, urban issues, housing, and labor.

Government-Linked Research Centers

In many Asian countries, government research organizations undertake important social, cultural, and behavioral research. In Singapore, the Institute of Policy Studies (IPS) conducts research on demography, religion, environment, and surveys on pertinent sociocultural issues and the Institute of Southeast Asian Studies (ISEAS) focuses on political, economic, and sociocultural issues on Southeast Asia and industrial states. In Hong Kong, the Social Sciences Research Centre does considerable government commissioned work. In Vietnam, the National Center for Social Sciences and Humanities of Vietnam (NCSSH) is funded by the government and provides analytic proposals for state foreign and domestic policies. The Center is divided into 16 institutes and seven research centers with 930 researchers. In Indonesia, with its social and cultural complexity, private research organizations are addressing the country's many social challenges. One of the newest research centers founded in 2002 is the Bandung Fe Institute that focuses on interdisciplinary research on complex social science and policy issues.

Japan, India, and China all have well-established government-funded social science research centers that act as important barometers of government programs, thinking, and policies. In Japan, the government-funded Economic and Social Research Institute has a unit dealing with 'Well-Being Studies' that undertakes quality-of-life surveys and measurement of well-being in Japan and some Asia–Pacific countries.

In India, the Indian Council of Social Science Research (ICSSR) established in 1969, conducts a multiplicity of research activities. Besides conducting its own research it funds other research projects, provides technical training, organizes and sponsors workshops, and undertakes publications.

In China there are several research institutes, which concentrate on social and cultural research. The most prominent are the Chinese Academy of Social Sciences (CASS), the Center of Sociological Research and Development Studies of China (CSRDSC), the Center for China Urban and Regional Studies, the Research Center of Tourism Impact Studies, and China's Academy of Social Sciences (CASS). All these research centers operate as umbrella research organizations covering law, culture, sociology, economics, psychology, and even the humanities. CASS is a massive research center with 32 research institutes covering 300 disciplines with 3045 research staff members. In China however, political ideology and leadership visions drive the nature of research agendas: under Mao the four modernizations; under Prime Minister Wen Jiabao correcting and managing the four economic 'un's(unsustainable economy, uncoordinated economy, unbalanced economy, and unstable economy); and under President Xi Jinping the 'China Dream' (peace, development, cooperation, and win–win results).

Contributions of Asian Social and Behavioral Research

Scientific and technological research in Asian institutes and research centers has grown tremendously over the past two decades (since 1995) spurred by Asian governments' applied interests, industrialization policies, urbanization processes, and infrastructural developments. Many Asian countries (Japan, South Korea, Taiwan, China, Hong Kong, and Singapore) embark on emulating developed countries in reserving 2–3% of gross national product development for R&D. However, little of this money is used for social and behavioral research.

In Asia, social and behavioral research is not perceived as prestigious for several reasons. Firstly, scientific and technological research can be compared against international benchmarks for international standards of comparison. Secondly, applied and economic interests exist in science and engineering R&D, which are readily visible over shorter time frames. Thirdly, R&D in science and technology carries an ethos of national modernization and progress for many developing Asian states.

While Asian scholars in science, engineering, and medicine follow the beat of the Western academic drum, Asian intellectuals in the social sciences and humanities are often seen as political icons and social heroes within their own states. The Asian academic voices deal with grounded social research,

which is often critical of government policies. These national voices on social issues from varied universities and local think tanks include in Thailand: Kasian Tejapira, Pasuk Phongpaichit, and Suwinai Pharanawalai; in Indonesia: Budiman, Sjahrir, and Juwono Sudarsono; in Malaysia: K.S. Jomo, Johan Saravanamuttu, and Maznah Mohamad; and in the Philippines: Resil Mojares, Leonor Brionnes, Bernardo Villegas, and Francisco 'Dodong' Nemenzo.

In Asia, of all the social sciences, economics, and business studies seem to command the most academic attention. Given the importance of the economics to employment, Foreign Direct Investments, industrialization, entrepreneurship, productivity, trade, agricultural production, food security, poverty alleviation, and income disparities, economics and business studies remain important research agendas. Two of the best Asian economic icons from Delhi's School of Economics are Amartya Sen and India's former Prime Minister Manmohan Singh.

The massive infrastructural and technological changes are also impacting the social, cultural, political, economic, and environmental landscapes. If one is to sum up the major changes that have profound social, cultural, and behavioral impacts in Asia, five trends can be identified: globalization (information technology) and its local impacts; rapid urbanization; democratization of political systems; environmental issues (such as climate change); and demographic issues due to an aging population. In *The Age of Ecology* (Radkau, 2014), environmental and climate change challenges bring together both science and social science R&D – one needs to know the science behind ecological processes as much as there needs to be an understanding of the impacts of environmental hazards on human culture, environments, and lives. The 2014 Intergovernmental Panel on Climate Change (IPCC) has altered the global community to the challenges arising from sea level rises, droughts, floods, heat waves, and food shortages that urge countries and corporations to spend more finance on environmental and climate change R&D. Unfortunately, climate change impacts remain low priority on many government agendas in Asia (e.g., Bangladesh, Sri Lanka, Pakistan, Laos, Cambodia, Myanmar, the Philippines, and Vietnam) and hence research in these areas is sadly lacking.

All these challenges are, in Ferguson's (2012) view, underpinned by "institutional degeneration" – not only in the Western world but also in Asia as well. The developmental challenges in many Asian states require sound social science knowledge, policy applications, and political management. The Asian Development Bank (ADB) has been proactive in urban and environmental research in Asia and provides interesting research because they are comparative studies, which cut across several Asian countries and cities. In order to develop more grounded, realistic, and applied solutions, universities, corporation, and research centers need to engage in cross-cultural and trans-disciplinary dialogue not only in the social sciences but between the natural and social sciences. In short, there is not only a need for a dialogue of civilizations (Dallmayr, 2002) but also a dialogue between science and the social sciences. This would mean radically changing the impervious silos of disciplinary knowledge in Asian tertiary institutes, and enhancing interdisciplinary exchanges by encouraging social science and behavioral humanities voices to be heard in R&D programs.

See also: Asia, Sociocultural Overviews: Korea; Asia, Sociocultural Overviews: South Asia; Colonialism, Anthropology of; Culture Policy Regimes in China; East Asian Studies: Culture; East Asian Studies: Economics; Education: Anthropological Aspects; Japan: Sociocultural Aspects; Knowledge, Geography of; Postindustrial Societies; Poverty, Culture of; Southeast Asia: Sociocultural Aspects; Urban Planning in China; Urban Policies in Asia; Urbanization in China.

Bibliography

Anderson, B., 1991. Imagined Communities. Verso, London.
Anwar, I., 1996. The Asian Renaissance. Times International Books, Singapore.
Arun, B., 2006. The Dialogue of Civilizations in the Birth of Modern Science. Palgrave Macmillan, New York.
Brahm, L.J., 2001. China's Century. John Wiley & Sons (Asia) Pte Ltd, Singapore.
Buruma, I., Avishai, M., 2005. Occidentalism: A Short History of Anti-westernism. Atlantic Books, London.
Cukier, K., Mayer-Schoenberger, V., 2013. The rise of big data. Foreign Affairs 92 (3), 28–40.
Dallmayr, F., 2002. Dialogue among Civilizations. Palgrave Macmillan, New York.
The Economist, 2014. Academic research: the useful science. The Economist 410 (8868), 56.
Feyerabend, P., 2013. The Tyranny of Science. Polity, Cambridge, UK.
Fenby, J., 2014. Will China Dominate the 21st Century. Polity Press, Cambridge, UK.
Ferguson, N., 2012. The Great Degeneration. Allen Lane, London.
Fukuyama, F., 1992. The End of History and the Last Man. The Free Press, New York.
Gribbin, J., 2003. Science a History. Penguin Books, London.
Huntington, S., 1996. The Clash of Civilizations and the Remaking of the World Order. Simon and Schuster, New York.
King, V.T., 2008. The Sociology of Southeast Asia. Nias Press, Copenhagen.
McClelland, D., 1961. The Achieving Society. D. Van Nostrand Company, Inc., Princeton, NJ.
National Population and Talent Division, 2013. A Sustainable Population for a Dynamic Singapore. National Population and Talent Division, Singapore.
Nature Publishing Index, 2013. Asia-Pacific. Nature Publishing Group, London viewed 19 February 2014, from: http://www.natureasia.com/en/publishing-index/pdf/NPI2013_Asia-Pacific.pdf.
Nisbett, R.E., 2003. The Geography of Thought: How Asians and Westerners Think Differently…and Why. The Free Press, New York.
Nye, J.S., 2011. The Future of Power. Public Affairs, New York.
Nye, J.S., 2004. Soft Power: The Means to Success in World Politics. Public Affairs, New York.
Peterson, A.L., 2001. Being Human: Ethics, Environment, and Our Place in Nature. University of California Press, Berkeley.
Radkau, J., 2014. The Age of Ecology. Polity, Cambridge, UK.
Reinert, E.S., 2007. How Rich Countries Got Rich…and Why Poor Countries Stay Poor. Carroll and Graf Publishers, New York.
R & D Magazine, 2011. 2012 Global R & D Funding Forecast. R & D Magazine, p. 30.
Said, E., 1994. Culture and Imperialism. Vintage, London.
Said, E., 1985. Orientalism. Penguin, London.
Schmidt, E., Cohen, J., 2014. The New Digital Age. John Murray, London.
Sheldrake, R., 2012. The Science Delusion: Freeing the Spirit of Enquiry. Coronet, London.
Steadman, J.M., 1969. The Myth of Asia. A Clarion Book, New York.
Sudev, B., March 2014. Big data, big questions. Jakarta Post 24 (14).
Temple, R., 1986. The Genius of China. Simon and Schuster, New York.
Thongchai, W., 1994. Siam Mapped: A History of the Geo-Body of a Nation. University of Hawaii Press, Hawaii.

Social Science Infrastructure: Central and South America

Cecilia Mabragaña and Maria Beatriz Gavilán, Gerencia de Desarrollo Científico y Tecnológico, Buenos Aires, Argentina

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social science institutions in the region show an evolution from colonialism to e-education in the twenty-first century. Similarities can be traced to a shared historical background, strongly influenced by the Church and stronger or weaker ties to political regimes. During the twentieth century, the social sciences were mainly institutionalized in the national state universities. Thirty-three percent of all degrees and courses offered in the region are in the social sciences, of which the main disciplines include sociology, education, and economics.

Introduction

A broad analysis of the social science institutions in Latin America shows an evolution from colonialism to e-education in the twenty-first century. Similarities and differences can be traced in the historical background, the ties to the political regimes, the late institutionalization within universities, the creation of National Councils of R&T, and the new trends in evaluation and e-learning.

Historical Background

Colonialism

All Latin American countries possess strong institutions for research, education, and communication of the social sciences, but the main infrastructure is found in national universities. There are many similarities among the social science institutions in Latin America but also many differences due to the relatively late institutionalization of the social sciences within the universities, the influence of political regimes and military dictatorships, and differential trends across nations to follow European models. This article focuses more on the similarities than the differences.

The universities in Central America were all products of Spanish colonization and the introduction of the religious orders, especially the Jesuit and Dominican Orders. These Orders and their strong background in education provided the seeds for future universities, the opening of seminaries, and the establishment of religious institutions. The colonies were far from the Spanish motherland, its universities, and its prestigious Catholic seminaries. The young colonies needed trained civilians and ecclesiastic personnel to occupy important positions in local governments.

Many universities were founded during the colonial period, such as the Universidad Autónoma de Santo Domingo, created in 1535 by Papal Bull, and the University of San Carlos in Guatemala, created in 1676 by a Royal Decree of Carlos II of Spain (it is the fourth oldest university in America). These first universities were limited in scope but included social sciences and humanities in the curriculum.

Developments after Independence

The university system in Latin America flourished only after the wars of independence during the nineteenth century. The post-independence era witnessed the reestablishment of many universities whose activities had been interrupted during periods of political upheaval. The scholarship, academies, and universities of Europe and the United States greatly influenced this academic renaissance.

During the twentieth century, the social sciences were mainly institutionalized in national public universities where they could be intertwined with national political concerns and activities. During those years, many national Councils of Science were created to delineate policies and strategies for research and scientific infrastructure. In most Latin American countries, a science council is the principal state agency that funds research, grants scholarships, evaluates programs and projects, coordinates research institutions, and hosts scientific journal portals. CONICET in Argentina, CNPQ in Brazil, and CONACYT in Mexico have their own institutes, but they are closely linked to the national and state universities.

Institutionalization of the Social Sciences

Institutionalization across the Region

While a general trend of institutional consolidation of social science infrastructure took place in the second half of the twentieth century, the process varied according to unique sociopolitical contexts. The influence of democratic regimes versus military dictatorships differed significantly. From the 1930s through the 1960s, Brazil, Uruguay, and Chile consolidated the social sciences in a democratic environment, with political liberties and academic freedom. However, in Argentina, the introduction of social sciences took place only after Juan D. Perón was removed from power. The University of Buenos Aires, the biggest Argentine national university, was granted significant autonomy, and two of the leading representatives were the historian José Luis Romero and the sociologist Gino Germani. There are other examples of nondemocratic governments that did support social sciences, for example, the Mexican government, which although it had a single-party

regime, nevertheless sustained the development of the social sciences with public funds, especially anthropology and sociology.

Though the interaction between the social sciences and national government was very strong in the three largest Latin American countries, Mexico, Brazil, and Argentina, the nature of the interaction varied substantially. For example, military regimes in Uruguay, Argentina, and Brazil were suspicious and disruptive toward academic social science research and organizations in their respective countries, but this paradoxically strengthened independent research centers that received funding from external countries and independent sources. Although a one-party regime, Mexico did not improve its centers with external financing but rather through the migration of Latin American researchers leaving their own countries because of restrictions imposed by the military regimes. Mexico, in fact, received two waves of scientists: the first were Spaniards fleeing the Franco regime; the second were Latin Americans leaving their countries for political reasons. These political frameworks gave way to strong displacements of scientists and weakened research output. During those same periods, there was an intensive expansion of the number of students in social science disciplines, especially in countries that do not have entry quotas, such as Mexico and Argentina.

In these initial periods, there was a consistent influence of external professors, different approaches to social problems, and a very marked influence of political ideologies mixed into the sociological discourse. There was always a strong political connection between high-level social scientists and politicians. In Brazil, for example, two high-level scientists-turned-politicians were Fernando Henrique Cardoso and Ricardo Lagos. Post 1945, the social sciences were also greatly influenced by the Cold War, the Cuban revolution, the proliferation of leftist guerrillas, and the ideas of the Catholic Church, as delineated in the papal encyclical, Rerum Novarum.

Establishment of Chairs and Councils

Academic chairs in sociology and the social sciences were established in all countries by the early twentieth century. The first one was in the School of Philosophy and Letters at the University of Buenos Aires in 1898. This process continued at the Universidad de La Plata, Universidad de Córdoba, and later at the Universidad del Litoral, revealing a trend toward federalization. In 1940, an Institute of Sociology was established at the University of Buenos Aires. In this same year, Chile had established 50 social science chairs. Brazil entered more slowly into this process, but after the establishment of a chair of sociology in the Escola Livre de Sociologia in 1933, the institutionalization of social science expanded throughout Brazil. In Uruguay, the first chair of sociology was created in 1913 and later reestablished in the School of Law.

During this time frame, many interesting studies employed fieldwork methods introduced by independent scientists, and these provided another view of the academic vision of the region's social sciences. In Chile, the studies provided by economic structuralists working in the Comisión Económica para América Latina (CEPAL) were very influential throughout Latin America, and still are today. CEPAL provides Chilean social scientists with many programs, grants, and educational activities, such as Ward and later Durkheim, positivism, the German reconstruction of the spirit of the peoples, and variants of Marxism, which have proved greatly influential.

The Consejo Latinoamericano de Ciencias Sociales (CLACSO), based in Buenos Aires since 1967, had two important leaders in the early period of Latin American social sciences: the economist Aldo Ferrer, an important proponent of economic nationalism; and the engineer Enrique Oteiza, a founding member of CLACSO. CLACSO had (and still has) a strong pan-national role in Latin America and has connected research centers across Latin America. Currently, it is a network of 320 research centers and over 600 graduate programs in the social sciences and humanities.

The Facultad Latinoamericana de Ciencias Sociales (FLACSO) is an international, regional, autonomous organization created by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) in 1957. Its principal objective was and is to establish postgraduate programs in social sciences. It has seven academic units, eight programs, and official academic chairs in Argentina, Brazil, Chile, Costa Rica, Ecuador, El Salvador, Guatemala, Mexico, Panamá, Paraguay, Dominican Republic, and Uruguay. From 1957 to 1973, the only chair was located in Santiago de Chile; it was eventually terminated by the military regime of Augusto Pinochet. The School of Sociology of FLACSO-Chile had many well-known European professors. The first director was a Bolivian sociologist and philosopher, René Zavaleta. FLACSO has been a key influence on the 'Latino Americanization' of the social sciences (i.e., from creating a pan-national Latin American presence in the social sciences to a series of nationally focused social science faculties in various countries).

Internationalization

Institutionalization of the social sciences paralleled internationalization. Many educational missions, exchanges of academics, and visiting professorships were financed and sponsored by UNESCO and also by the Ford Foundation. The first specialized international social science associations were organized in Argentina: in 1950, the Latin American Sociology Association; then in 1960, the Italian sociologist Gino Germani founded the Asociación Argentina de Sociología.

In 1960, FLACSO signed agreements with the Ecole Pratique des Hautes Etudes in Paris, the University of North Carolina, and the University of Chicago. These agreements, along with links to various Jesuit universities, promoted external influence on the development of Latin American social science programs and were instrumental in establishing sociology programs in Brazil, Chile, and Argentina.

In the 1970s and 1980s, the Economic Commission for Latin America (ECLA/CEPAL) tackled two key subjects: development and underdevelopment in what was then called the 'Third World.' Seminal works on these topics were written by scholars such as Raul Prebisch, Celso Furtado, and Gino Germani in Argentina, Florestán Fernandez in Brazil, and Pablo Casanova in Mexico. During this time, the social sciences and many other institutions as well engaged in the discussion of social problems, experiencing a boom in certain

countries: Chile, Peru, Brazil, Argentina, Mexico, and Uruguay. The following is a summary of the institutionalization of the social sciences by country.

Social Sciences in Individual Countries

Mexico

In Mexico, the state and the social sciences have always been related, with weaker or stronger ties. Before the Spanish Conquest, Mexico was home to several major civilizations: Aztecs, Teotihuacanos, Mayas, and Mixteco-Zapotecos. The legacy of highly developed, indigenous civilizations has marked Mexican cultural identity. Though its roots are indigenous, Mexican cultural identity today is mestizo, a mix of local and Spanish influences. The first indigenous president, Benito Juárez, is a great symbol of the Mexican identity and the consolidation of Mexico as a modern nation, especially because he defeated the Hapsburg Empire.

The Mexican social sciences have their roots in the disciplines of history and anthropology. In 1917, the Instituto de Arqueología was founded; in 1929, the Universidad Nacional de Mexico became autonomous; and in 1939, the Instituto de Investigaciones Sociales and the Revista Mexicana de Sociología of Lucio Mendieta were created. Manuel Gamio was another pioneer of the social sciences; he carried out two major projects on the exploration of the archaeological site of Teotihuacán, and on the migration of Mexicans to the United States. At the beginning of the twentieth century, northward migrations were the focus of many sociological studies, and still are today.

The Mexican government founded the Instituto Nacional de Antropología e Historia (INAH) in 1939, as a key piece in the process of institutionalization. Today, the most important Mexican social science institutions are financed by the state. The Social Sciences Institute of the National Autonomous University of Mexico (ISSUNAM) was also founded in 1939, providing a framework for social research in Mexico. ISSUNAM facilitates research on a wide variety of subjects, such as agrarian labor, labor conflicts, indigenous marginalization, and politics. La Casa de España (later COLMEX), established in 1939, provided a space for Spanish intellectuals fleeing the Franco regime. It became a solid cultural center with full-time academics and many prominent intellectuals who published in the serial collection Las Jornadas de el Centro de Estudios Sociales de el Colegio de Mexico: José Medina, Octavio Paz, Raúl Prebisch, Kingsley Davis, among others. This Centro de Estudios Sociales (CES) was closed in 1946. In the late 1930s, Daniel Cosío launched the publishing house Fondo de Cultura Económica, which provided translations of all the fundamental works of Pareto, Durkheim, Weber, Keynes, Marx, and others. Today it depends on the federal government for support, but its editorial policy is totally independent.

Many exiled scientists stayed in Mexico, as happened with the Argentine exiles from the military regimes. The construction of the Universidad Nacional Autónoma de México (UNAM) campus, a big architectural undertaking funded by the state, began in 1947, bringing together all the schools, which today number more than 150 000 students and 30 000 professors. These three institutions, COLMEX, UNAM (ISSUNAM), and INAH, are the backbone of the social sciences in Mexico.

Professors began teaching sociology only in the 1960s. From 1965 to 1970, the publication of the book *La Democracia en Mexico* by Pablo González Casanova stimulated studies on the social problems of poverty, political power, and inequality affecting Mexican society. In 1967, the master's degree programs in sociology and political science opened in UNAM. After 1970, the increase in the number of PhDs was exponential. Between 1951 and 1979, the number of sociology students grew from 3 to 7000. This boom in registration was attributed to the commonly held belief that sociology would bring about (positive) structural changes. After this period of euphoria, in the 1990s, interest in these fields declined and students switched to economics and econometrics. The decentralization of COLMEX initiated the foundation of the Colegio de Michoacán and the Colegio de la Frontera Norte in Tijuana.

In the 1970s, Mexican students from all over the country that used to attend UNAM dispersed among 32 state universities. While the lack of professors was a problem, the new waves of exiles from Argentina and Chile contributed to university development. The 1971 creation of the Consejo Nacional de Ciencia y Tecnología (CONACYT) contributed to scholarship abroad and the overall supervision and support of all activity in research and technology.

As a discipline, anthropology did not receive the same rank or prestige as sociology and political science. Despite the fact that 10% of the Mexican population is of indigenous descent, anthropology has not been the focus of academic research and actually lost some of its importance in scientific terms.

Higher education in modern Mexico is a consequence of public policies and funds in spite of crises and political upheavals. Ten percent of the graduate students in Mexico opt for the social sciences.

Today, UNAM is the best-ranked university of Mexico (70) (Webometrics Ranking of World Universities – CSIC España 2013–14) and home to over 320 000 students.

Nicaragua

Nicaragua has over 50 universities, national and private. The oldest is the Universidad Nacional de Nicaragua (UNAN), founded in 1812 in the city of León. The Recinto Universitario Rubén Darío opened in 1969. Education, economics, and humanities departments were developed at the Managua campus. During the decade of the Sandinista revolution, the University of Managua grew in strength, number of students, and resources. In 1982, UNAN-Managua became fully independent from the León historical headquarters. Nicaragua is now a part of the Sistema Centroamericano de Evaluación y Acreditación de Educación Superior (SICEVAES).

Nicaragua offers a range of 194 degrees in the social sciences, all at the Instituto Nicaragüense de Investigaciones Económicas y Sociales (INIES). The university system offers several postdoctoral degrees: 10 in education and languages; 7 in business administration and economics; and 5 in the humanities, law, and social sciences. Two doctoral programs are offered jointly with the Universidad de Zulia, Venezuela, one in social sciences and one in economics.

Guatemala

The University San Carlos de Guatemala, founded in 1676 by a Royal Charter of Carlos III of Spain, is one of the oldest universities in Latin America. After Guatemala's independence from Spain, the university's activities were interrupted. Today this public university has 33 campuses, 10 schools, and other institutes, and offers eight postdoctoral programs.

Honduras

The most important university in Honduras is the Universidad Autónoma de Honduras (UNAH), which has campuses and schools in 11 provinces. There are also more than 30 private universities. The university system offers a total of 258 degrees, 81 in the social sciences. A doctorate in Comunicación y Desarrollo (Communication and Development) is offered jointly with the Universidad Complutense de Madrid. With the sponsorship of Andalucía, Spain, a special digital library was created for the study of migration and intercultural relations.

El Salvador

The first Salvadorian university was founded in 1841 in a struggle to gain independence from the prestigious Guatemalan University of San Carlos. Until 1967, it was the only university in the country. Lacking the professionals needed for its development, the university encouraged the professional characteristics of the Napoleonic model.

In 1944, the university changed its goals, giving priority to research, academic excellence, and library development. In 1972, military intervention ended this 'golden era,' but in 1992 an era of reconstruction began. Today there are 37 public and private universities in El Salvador, and more students attend private than public universities. Of the 258 degrees offered, 139 are in the social sciences. REDISAL is the network of Salvadoran researchers. The National Center of Research in the Social Sciences (CENICSH) was created in 2009. Universities are publically funded; the social sciences receive the largest proportion of funds (31%), followed by engineering and technology (28%), natural and exact sciences (17%), medicine and health care (12%), humanities (8%), and agronomy sciences (4%).

The Consejo Nacional de Ciencia y Tecnología (CONACYT) was created in 1992 as a decentralized, autonomous institution and the chief authority for all aspects of science and technology. It includes the registration of all Salvadorian researchers and has a Center for Technological Information. It promotes different types of scholarship cofinanced by UNESCO and the governments of Mexico, Japan (Monbukagakusho), and Germany (DAAD).

Costa Rica

The first university in Costa Rica goes back to the Casa de Enseñanza Santo Tomas founded in 1814. Political motivations closed the university in 1859, but the schools of law and agronomy still remained open. In 1940, the university reestablished itself as the Universidad Santo Tomas. Today it has seven regional campuses, 41 000 students, nearly 7000 full-time professors, 4000 postdoctoral students, and 600 doctoral students. The university has 13 schools of different disciplines; 30 research centers, of which six are of the social sciences; and 12 institutes, five of which are of the social sciences. The university is also home to the publication portals Latindex and SciELO.cr. Today, Costa Rica is home to 10 public and 51 private universities. The University of Costa Rica is the best-ranked university in Central America, especially in the social sciences, and it has schools of law, economics, education, and psychology. The university system offers 759 degrees, 222 of which are in the social sciences.

Panama

The Universidad de Panamá was created in 1935. The first campus was built in 1947, and 17 buildings were added between 1970 and 1989. Since then, many additional buildings have been added, and regional campuses were opened in different provinces. The Universidad de Panamá has 18 schools, seven of them in the social sciences. The governments of Mexico and France today cofinance Panamanian social science projects. Today Panama has three public and 19 private universities, and the Sistema de Internalización de la Educación Superior Centroamericana (SIESCA) fosters mobility and exchange of fellows and researchers. The total number of degrees is 819, with 287 of them in the social sciences.

Venezuela

Venezuela's best-ranked and oldest university is the Universidad de los Andes. It dates to 1785, when the Bishop of Mérida created an institute for young seminarians; King Carlos IV of Spain accredited it in 1806; and then in 1810, it was given the title of Real Universidad de San Buenaventura de Mérida de los Caballeros. The university was later completely destroyed by an earthquake, moved to Maracaibo, and then reinstalled in Mérida in 1824 as Universidad de Mérida. In 1905, after many closures and reopenings, its name was finally changed to the Universidad de los Andes. Currently, the Universidad de los Andes has two postdoctoral degrees in economics and statistics. Venezuela today has 25 public universities (some with regional campuses) and 27 private ones. The public universities are divided into autónomas (5) and experimental (20).

Dominican Republic

The first Latin American university, the Universidad Autónoma de Santo Domingo, was created by Pope Paul III per Papal Bull in 1538 and ruled by the Dominican Order. Following the Haitian invasion and recolonization period, in 1859 it was reformed along the lines of the European medieval academies, with four faculties (philosophy, law, medicine, and sacred letters). After many closures, reopenings, and 31 years of the Trujillo dictatorship, the university permanently reopened in 1964.

The proliferation of private universities and the opening of regional campuses can be explained by the aforementioned political crises. The best-ranked university is Universidad de Psicología Industrial Dominicana (UPID) and is a private university. The best-ranked public universities are Universidad

Autónoma de Santo Domingo, which has 19 campuses. One of the four public universities is the Facultad Latinoamericana de Ciencias Sociales (FLACSO).

Bolivia

The Universidad Mayor Real y Pontificia de San Francisco Xavier de Chuquisaca in Sucre was founded in 1624 and is the oldest Bolivian university. Founded by a Jesuit priest and expanded in 1775 to include the Academia Carolina, the university was initially dedicated to legal studies. Other universities were also created to train lawyers, and only after several decades and the expulsion of the Jesuit Order did the universities grow to include other disciplines.

The Academia Carolina was the most important university in colonial times and greatly influenced the regions of el Alto Peru and Río de La Plata. Eventually, the academy planted the seeds for the emancipation processes in Latin America by shaping leaders such as Mariano Moreno, Bernardo Monteagudo, and Juan José Castelli of Argentina; Manuel Rodriguez Quiroga of Ecuador; Mariano Alejo Alvarez of Peru; and Jaime de Zaldañez of Uruguay.

Two elements characterize Bolivian universities: university autonomy and the co-government of professors and students. In 1967, the structure and evaluation systems were modified, including the opening of private universities to reduce student overcrowding. Today Bolivia has 16 public universities and 33 private ones. The best-ranked university is the Universidad Mayor de San Simon. Future goals include the improvement of postdoctoral studies and development of research structures in all disciplines.

Peru

Peru is home to one of the first universities in Latin America, the Universidad Nacional de San Marcos. Founded in 1551 by the Dominican Order, the university was granted the support of King Charles V of Spain. It is the only university in Latin America with uninterrupted activity; since its creation, 223 deans have passed through its doors. Today it offers 62 degrees in 20 schools.

There are 35 state universities and 56 private ones. The best-ranked university is the Pontificia Universidad Católica del Perú (891), a private university in Lima established in 1917 with two faculties: school of law and school of letters. The Centro de Investigaciones Sociológicas, Económicas, Políticas y Antropológicas (CISEPA) is one of the university's strongest institutes, with 57 research projects and 237 researchers. The Universidad Nacional de San Marcos has 3360 professors, 672 of them in the social sciences. The Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) is the governing body of the national system of science, technology, and innovation; it is composed of the academy, the public research system, business organizations, and civil representatives.

Ecuador

Founded in 1946 by the Jesuit Order and the Archbishop of Quito, the Universidad Católica del Ecuador (PUCE) is Ecuador's best-ranked and oldest private university. Currently, Ecuador is home to 54 universities, of which 34 are state universities. The Council of Evaluation, Accreditation and Excellence of Higher Education evaluated and ranked the national universities in four categories according to academy, academic proficiency, research, and organization and infrastructure. Five universities are graded A, 23 graded B, 18 graded C, and 8 graded D. The future challenge for the authorities is to include Ecuadorian universities among the first 1000 universities in the international rankings.

Paraguay

Created in 1889, the Universidad Nacional de Asunción is Paraguay's oldest university. Also the best ranked (3034), the university consists of three schools: law and social sciences, medicine, and mathematics. From 1961 to 2013, 45 new private universities were founded. Today, la Universidad Nacional de Asunción has 12 schools and 2 institutes, offering 74 degrees to more than 40 000 students.

The Consejo Nacional de Ciencia y Tecnología (CONACYT) was created in 1997 as a public, autonomous institution. It is designed to plan, coordinate, and execute the priorities and policies of the National System of Science and Technology. Starting in 2006, the first international project with the support of Programa de las Naciones Unidas para el Desarrollo (PNUD) and Banco Interamericano de Desarrollo (BID) took place. It is steering the creation of a Researcher's Registry and supporting science (scholarships and programs) with a fixed percentage from the general national budget. Paraguay has 300 000 students, 81 percent of which are enrolled at private universities.

Colombia

Created in 1867, the Universidad Nacional de Colombia can trace its origins back to the independence era when its campuses spanned the colonial region of Gran Colombia, across Bogota, Caracas, and Quito. In 1867, the university consisted of six faculties: law, medicine, engineering, art, literature, and philosophy. University facilities also included the Astronomical Observatory, the National Museum, laboratories, the National Library, and two hospitals. By 1940, it offered 20 professional degrees, and between 1973 and 1986, many new graduate programs were created.

There are more than 32 state universities and over 100 private ones. The three best-ranked universities are the Universidad Nacional de Colombia, Universidad de los Andes Colombia, and Universidad de Antioquia. COLCIENCIAS is today the administrative Department of Science, Technology and Innovation. It coordinates the national system of Science and Technology and the Scienti Database (Colombian researchers). COLCIENCIAS also maintains PUBLINDEX for the categorization and evaluation of Colombian scientific journals.

Uruguay

The Universidad de la República was created in 1849. In its first years, it had four schools: medicine, law, natural sciences, and theology. During the first 40 years, there were successive conflicts with the government, and the influence of the Church

diminished. A second stage, which lasted until 1908, gave rise to the modern university, with different interventions and reforms according to the political party in power.

The Universidad de la República established the School of Humanities and Sciences and the Instituto de Economía in 1963. The influence of FLACSO in 1969 was crucial for the establishment of a bachelor's degree in sociology (1970). However, the military dictatorship prevented further development of the sociology department until after 1985.

The Faculty of Social Sciences was created in 1991, the Faculty of Psychology in 1994, and the Faculty of Communication and Information in 2005. There are six state universities and 21 private ones.

Uruguay has separate institutions supporting science and technology. Since 2006, the Agencia Nacional de Investigación e Innovación (ANII) has funded innovation in science and set strategic goals and programs in science policy: funds for research projects, incentives for innovation in business and commercial organizations, scholarships, the national registry of researchers, and the Portal Timbó (access to scientific publications). The Consejo Nacional de Innovación, Ciencia y Tecnología was also founded in 2006 as an expert body on science and technology for the executive government.

Chile

The oldest university is the Universidad de Chile, created in 1842; this was a landmark creation of the young republic. At first tied to the Catholic Church, its first dean was Dr. Andrés Bello, the Venezuelan humanist and aide to Simón Bolivar. Dr. Andrés Bello prepared the founding documents for the creation of the university as a state, secular, and public university. During the first decades of the twentieth century, the most prominent men of Chile passed attended the university. In 1950, the status of full-time professor was extended to all faculties. Today, the Universidad de Chile has about 27 000 students and 3000 professors.

FLACSO, CEPAL, and the Universidad Católica de Chile established the School of Sociology and the School of Economics in 1958. Two key subjects influenced and tackled by CEPAL in Chile during the 1970s and 1980s were development and underdevelopment and the concept of Third World countries.

The Comisión Nacional Científica y Tecnológica (CONICYT) was established in 1967 under the wing of the Ministry of Education. CONICYT supports research and promotes partnerships and scholarships; it hosts the SciELO.cl Program, the National Repository projects, and a network of researchers. Special emphasis has been given to women's studies.

Chile has 16 state and 49 private universities, and 10 research centers. The Universidad de Chile is the best-ranked university (297) together with the Pontificia Universidad Católica de Chile (608). CONICYT Chile will finance 22 new projects to attract Chilean researchers abroad (Programa de Atracción e Inserción de Capital Humano Avanzado PAI). These new projects are predominantly directed to Chilean researchers from the University of Pittsburgh and Johns Hopkins University in the United States, among other institutions.

Brazil

As in most Latin American countries, the origins of institutionalization of the social sciences date back to the legal and social programs of European positivism and the adoption of European and American standards. The study of sociology was initially fostered by the founding of the Scola Livre de Sociologia e Politica in 1934 and the Universidad de Sao Paulo (USP) in 1935, which provided an institutional framework for the social sciences and academic exchange with foreign professors from France, the United States, and Germany. In 1932, the Church turned its Instituto Católico de Estudos Superiores into the Faculdade de Filosofia das Faculdades Católicas, and then in 1946, founded the first Pontifícia Universidade Católica (PUC).

The Sao Paulo School of Sociology was established in 1955. The first master's program in the USP was offered in 1953. Between 1950 and 1960, a number of new institutions were created in Rio de Janeiro. Until the military coup in 1964, the hierarchy of disciplines was dominated by sociology. The importance of anthropology grew in Bahia, Recife, and cultural anthropology with Gilberto Freire in Parà. In addition, important results of anthropological research were displayed at the Goeldi Museum in Belém.

From 1965 to 1983 (until the end of the military dictatorship), social sciences expanded throughout the territory as a result of the establishment of universities, graduate programs, and research. There was also a strong internationalization and implementation of exchange programs with the support of government policies and funding agencies.

The USP model became the university standard and was reproduced in the federal system until 1960. A noteworthy change occurred in 1968 when the university law program modernized, inspired by the American model and influenced by the agreement between the Ministry of Education and Culture of Brazil and the US Agency for International Development (USAID). Scientific investigation was strongly concentrated in the universities and received a significant flow of resources from the government, including the social sciences. Founded in 1951, the Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPQ) is an important funding agency and supports exchange programs and other activities such as the platform LATTES that contains the biographies of all Brazilian researchers and professors.

The social sciences in Brazil's universities expanded dramatically in the 1970s and 1980s; their strength was the graduate programs: masters and doctoral programs doubled enrollment between 1974 and 1985, involving around 7000 scientists. The following are some of the most well-known themes, topics, and authors in the field of anthropology:

- Violence and Crime; Indigenous Ethnology: Alba Zaluar Vivarios de Castro, Carneiro da Cunha.
- Life style of the middle class, cultural habits, violence and drug abuse: Darcy Ribeiro, Cardoso de Oliveira, Amorim, Gilberto Velho.
- Comparative anthropological perspectives: Cardoso de Oliveira and Guillermo Ruben Peirano.

Another strong academic area is ethnicity, which was discussed mainly by Schwarcz, Carneiro da Cunha, and Pinto.

In sociology there are two classical fields of study: social classes and social mobility, pioneered by Guimaraes Valle Silva. Foreign institutions provided funding and the exchange of professors, particularly the Ford Foundation, which sponsored studies of gender, and race and ethnicity from 1960 to 1983. Currently, the Coordenaçao de Aperfeiçoamento de pessoal de Nível Superior (CAPES) has diversified participation in funding scholarships through its exchange programs with several countries, like the Santos Dumont agreement for Brazilian and French graduate programs.

Argentina

The University of Cordoba was the first university in Argentina, founded in 1631, which at that time was a province of the Virreynato del Perú, and later in 1776 was already part of the Virreynato del Rio de la Plata. All the historic, colonial universities were linked to the Jesuit and Franciscan Orders of the Catholic Church. For example, the governor of Buenos Aires founded the Universidad de Buenos Aires (UBA) in 1821. Its first dean was a priest, Antonio Saenz.

The process of institutionalization of the social sciences in Argentina has been marked by discontinuities due to political changes and military coups. These struggles also coincided with the establishment of the private universities. Political confrontations between students and professors and differences in disciplinary orientation were often the source of destructive conflicts.

Argentina's first university program in sociology was created in 1957, rooted in the Peronist movement and tied to the military dictatorship. The first chair of sociology in Latin America was established in 1889 in the Facultad de Filosofía y Letras of the Universidad de Buenos Aires (UBA). In 1918, a massive student strike at the University of San Carlos en Córdoba initiated the historic, autonomist movement across the Latin American continent: Manifiesto Liminar de la Juventud argentina a todos los Hombres Libres de Sudamérica. The chronology of the period can be divided into three periods with distinctive characteristics. The first period (1955–1966) began with a military coup against the Peronist government, which incited political repression and armed struggles. In the UBA there was an intense flux of foreign capital incentivizing research and academic activities. In 1955, the universities were taken over by the military government. Between 1957 and 1966, several undergraduate programs were established in different national and private universities with formal chairs of sociology, anthropology, and political science. The military coup in 1966 resulted in nearly 1000 professors resigning from the UBA and other national universities, especially in the social sciences. During those years, two important research centers were established: the Instituto de Desarrollo Económico (1958) and the Centro de Sociología Comparada (1963), as well as journals associated with them, Desarrollo Económico and the Revista Latinoamericana de Ciencias Sociales. Two publishing houses were important for the diffusion and legitimization of the social sciences: Paidós and Eudeba, the University Press of the University of Buenos Aires.

The Consejo Nacional de Investigaciones Científicas y Tecnicas (CONICET) was created in 1958 and governed by President Bernardo Houssay (Nobel Prize in Physiology or Medicine 1947). CONICET is an autarchic organization in charge of promoting science and technology across public and private sectors. It currently has more than 180 centers of research and its own degree researchers and professionals.

The period from 1966 to 1983 was marked by military coups, political repression, and social struggles. In 1979, an MA program was created with majors in sociology, political science, and education. This time period saw the establishment of independent training centers sponsored and supported by the Ford Foundation, the Swedish agency SAREC, and French agencies. Some of these centers are today part of CONICET's Institutes: Centro de Estudios e Investigaciones Laborales (CEIL) and Centro de Estudios Urbanos y Regionales (CEUR). These centers covered different thematic areas, like rural sociology, labor, and demographic issues. The creation of national universities in the interior of the country that included social science programs grew in number and scope. These new centers produced an interconnected system of researchers and ideas. CLACSO played a decisive role in the network of people and knowledge.

The third period (1983–2003) opened an era of constitutional governments and the reconstruction of universities. Although the economic situation had its ups and downs, questions and themes such as unemployment, poverty, crime, and corruption were and remain central problems and concerns for Argentine social scientists. A new school (Facultad) was created in the UBA that included the fields of sociology, political science, labor, and social work. Social science students represent 40% of the total enrollment. The Ministry of Education established a special agency for the evaluation of graduate courses and programs (CONEAU). The revitalization of the national universities paralleled the diminishing support and funding by foreign agencies.

The Universidad de Buenos Aires is the best-ranked university of Argentina, home to over 320 000 students. It is one of Latin America's largest universities, together with Mexico's UNAM. It has collaborative research projects and programs with other international universities and less developed ones with industry partners. There are no tuition fees, as in all public state universities in Argentina.

Argentina has suffered brain drain for the last 40 years, especially since 1966. After 1983, the brain drain was driven by financial and/or economic reasons. A large percentage of exiles have returned or kept close ties to the countries and institutions in which they worked, obtaining financial support for collaborative studies. The advent of information technologies and the Internet opened new possibilities for the development of networks and reestablishment in local research centers. The 820 researchers who returned to Argentina in 2012 represent 13% of Argentine researchers abroad (estimated at 6000).

FLACSO has proven a very important source for graduate programs in Mexico and Ecuador, and some foreign universities like de Università di Bologna and New York University have established branches in Argentina.

At present, CONICET has over 180 institutes and centers of research, of which 13% concentrate on the social sciences and humanities. From the total number of researchers in the different categories, 18% work in the social sciences.

Dissemination and Communication of Social Sciences in Latin America

Latin America has been a pioneer in organizing interdisciplinary, open access scientific journal portals. The first effort was a cooperative decentralized network of national scientific organizations: LATINDEX (Sistema Regional de Información en línea para revistas científicas de América Latina, Caribe, España y Portugal), which includes scientific journals of the region with approved common quality criteria hosted and run by UNAM, Mexico. Of the 22 295 journals, 10 567 belong to the social sciences.

Another important medium for dissemination is SciELO.org (Scientific Electronic Library Online). Established in 1998 in BIREME (Brazil), the journals included are selected by national scientific organizations. As of 2013, the library has 10 148 journals and over 10 000 000 citations; 226 are social science journals.

REDALYC (Red de Revistas Científicas de América Latina, el Caribe, España y Portugal) is a multidisciplinary open access journal portal, originally and predominantly dedicated to the social sciences and the humanities; it was developed in 2002 through a research program in UNAM, Mexico.

Distribution of Centers and Universities in Latin America

Latin America has 3745 centers of higher education, of which 2226 offer degrees in social sciences (Webometrics and QS Top University – Worldwide University Ranking 2013). The highest concentration is in Mexico (37%) followed by Brazil (13%) and Colombia (13%). Thus, Mexico, Brazil, and Colombia possess 65% of all centers of higher education in the region. These values correlate with the overall population structure. The second largest concentration of universities (25%) in the region can be found in Argentina (7%), Peru (7%), Chile (4%), and Venezuela (4%). Uruguay, Paraguay, Bolivia, Ecuador, El Salvador, Honduras, Nicaragua, Panama, and Dominican Republic account for the remaining 13%.

According to UNIVERSIA, the 1299 universities offering degrees in social science show a similar distribution. According to the QS Top University – Worldwide University Ranking 2013 selection, a ranking system for the 200 most important universities by discipline, only 14 Latin American universities are included, representing 7.5% in 38 subjects.

The analysis of degrees and doctoral courses is based on the ranking established by Q's in the table below. The higher quality and excellence is (1) Q numbers from 1 to 50; (2) Q 51 to 100 (median value); (3) Q 101 to 150; and (4) Q 151 to 200. The last quartile represents the lowest quality of the top 200 universities.

Brazil presents the following excellence structure and scale:

1. Q1/50 = 2 courses. Statistics (USP) and Education (USP)
2. Q51/100 = 7 courses. 2 Statistics (UNICAMP, UFRJ), 2 Economics (USP, FGV), 1 Education (UNICAMP), 1 Political Science (USP), 1 Law (USP)
3. Q101/150 = 13 courses. 6 Sociology (USP; UNICAMP, UFRG, UFRGS, UFB, UFMG), 1 Statistics (UFMG), 2 Education (UFRG, UFRGS), 2 Economics (UNICAMP, PUC RIO), 1 Political Science (FGV), 1 Communication and Media (UFRJ)
4. Q151/200 = 2 courses. 1 Political Science (PUC RIO), 1 Communication and Media (UFMG)

This distribution shows that values are concentrated in the median (7 + 13 = 20 courses out of 24)

Chile, with 4 universities and 7 courses, has the following structure and ranking:

1. Q1/50 = 1 course. Education (PUC-Chile)
2. Q51/100 = 3 courses. 1 Education (UCH), 1 Economics (PUC-Chile), 1 Law (UCH)
3. Q101/150 = 1 course. Political Science (PUC-Chile)
4. Q151/200 = 2 courses. Law (UDP, PUC-Valparaiso)

It is interesting to point out that Chile surpasses Argentina and Colombia in universities and courses. Argentina (UBA) has 2 courses in Q2 and Colombia (Universidad de los Andes) 1 course in Economics in Q3.

Mexico is represented with 1 university and 4 courses, 2 in Q1 and 2 in Q3.

Degrees and Postdoctoral Courses

Latin America is represented by 34 939 degrees and courses, of which 11 722 (33.5%) are in the social sciences (Data: December 2013).

Country	*All degrees*	*Social sciences*
Brazil	9205	3086
Argentina	4725	1534
Mexico	5070	1778
Chile	3340	899
Colombia	3480	1278
Uruguay	707	254
Peru	1829	660
Panamá	819	287
Bolivia	572	172
Costa Rica	757	222
Ecuador	865	365
El Salvador	415	139
Honduras	258	81
Nicaragua	462	194
Paraguay	350	121
Dominican Republic	587	213
Venezuela	1498	439
	34939	11722

Grouping the countries according to the prevalence of the social science degrees, Group 1 concentrates 67%, and is headed

by Brazil (27%) and followed by Mexico (15.5%), Argentina (13.3%), and Colombia (11.1%). The second group is headed by Chile (9.8%), followed by Peru (5.4%), Venezuela (3.8%), and Ecuador (3.2%). The median values (2.5%) are represented by Panama and Uruguay. The lowest group (1.2%) is represented in the following order: Dominican Republic, Nicaragua, Bolivia, Costa Rica, El Salvador, Paraguay, and Honduras.

Another important thing to look at is the importance of evaluation and accreditation of degrees and courses in higher education. The total number of degrees and courses of all levels in higher education – master's degrees, doctorates, and specializations – amount to 5691, representing 17% of all courses developed in Latin America.

The following table shows Brazil's concentration of evaluated courses and degrees (50%). CONACYT Mexico has played a significant role in the strengthening of the national evaluation system, including the National Register of Researchers. The National Register of Postgraduate Programs (PNP) has two categories: International Level and High Level.

Argentina's effort is especially visible in the evaluation of doctorate courses. Uruguay has recently incorporated evaluation of courses and degrees, performed by ARCO SUR.

Brazil, with 1405 evaluated higher education courses, together with Argentina's 842 courses, represents 80% of evaluated doctorate courses in the region.

During the last decades, the Sistema Centroamericano de Evaluación y Acreditación de Educación Superior (SICEVAES) has made a systematic effort to ameliorate all aspects of university performance in Central America. Different guides for self-evaluation and external evaluation are recommended for the countries involved.

The following table shows the percentage of students in the social sciences in comparison with other disciplines (UNIVERSIA, 2013).

Country	*All degrees and courses*	*All social sciences degrees*	*All degrees w/evaluation*	*All higher education degrees w/evaluation*	*All degrees social sciences w/evaluation*	*All social sciences higher education careers w/evaluation*
Brazil	9205	3086	2890	1405	669	244
Argentina	4725	1534	920	842	313	312
Chile	3340	891	636	253	110	54
Mexico	5070	1749	636	0	227	0
Colombia	3480	1261	569	340	159	109
Uruguay	707	254	10	2	0	2
Perú	1829	643	2	0	0	0
Bolivia	572	172	0	0	0	0
Costa Rica	757	222	0	0	0	0
Ecuador	865	365	0	0	0	0
El Salvador	415	139	0	0	0	0
Honduras	258	81	0	0	0	0
Nicaragua	462	194	0	0	0	0
Panamá	819	287	0	0	0	0
Paraguay	350	121	0	0	0	0
Dominican Republic	587	213	0	0	0	0
Venezuela	1498	439	0	0	0	0
Total	34939	11322	5608	2842	1472	715

New Technologies: E-courses and Social Sciences

As an indicator of the new Technology and Innovation Centers (TICS) introduced in universities, the number of online courses offered has been quantified by country, for all disciplines and for the specific participation of the social sciences.

The region has 1679 degrees and courses online, representing 4.6% of all courses offered. Brazil leads the trend with 528 courses (31.48%), Colombia (16.6%), Mexico (14.8%), Argentina (13.4%), and Chile (6.1%)

The following table has grouped the countries by quartiles.

PAÍS	*All degrees/ distance learning & online*	*Social sciences/ distance learning & online*	*% in region's total*	*% of social sciences*
Brazil	528	256	35%	48.5%
Colombia	279	105	14%	37.6%

Continued

PAÍS	*All degrees/ distance learning & online*	*Social sciences/ distance learning & online*	*% in region's total*	*% of social sciences*
Mexico	249	135	19%	54.2%
Argentina	225	98	13%	43.6%
Chile	103	13	2%	12.6%
Peru	79	46	6%	58.2%
Ecuador	51	6	1%	11.8%
Nicaragua	35	25	3%	71.4%
Venezuela	33	11	2%	33.3%
Uruguay	23	18	2%	78.3%
Paraguay	12	6	1%	50%
Bolivia	5	1	0%	20%
El Salvador	4	2	0%	50%
Panamá	3	2	0%	66.7%
Dominican Republic	2	2	0%	100%
Costa Rica	0	0	0%	0%
Honduras	0	0	0%	0%
Total	1631	726	100%	44.5%

From this general analysis we can draw the conclusion that the development of the institutions of social sciences in the region has not been homogeneous. The influence of political ideologies and context has been, and still is, very strong. There is a general trend to focus on evaluation and excellence of degrees and postdoctoral studies, highlighting Brazil's leadership. The diversification of degrees and courses in the social sciences shows an adjustment to the global trend. Overall, the prevalence of social sciences is 40%. The National Councils of Science and Technology represent a new focus on excellence and competitiveness, and the formulation of scientific policies.

See also: Agrarian Political Economy; Anthropology, History of; Anthropology: Overview; Applied Sociology; Catholicism; Colonialism, Anthropology of; Hegemony: Anthropological Aspects; Historical Sociology; Human Rights, Anthropology of; North America: Sociocultural Aspects; Political Science, History of; Poverty, Geography of; Race and Racism in the Twenty-First Century; Social Science Professions and Professionalization; Social Science and Universities; Sociology: Overview; South America, Archaeology of; Urbanization.

Bibliography

QS Top, 2013. QS Top University – World Wide University Ranking 2013.
UNIVERSIA, 2013–2014. University Network Fundación UNIVERSIA 2013–2014.
Webometrics, 2013–2014. Webometrics Ranking of World Universities: Cybermetrics Lab, (CSIC) Spain 2013–2014.

Social Science Infrastructure: East Asia and the Pacific (Research and Teaching)

Takashi Inoguchi, University of Niigata Prefecture, Niigata, Japan; and University of Tokyo, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the development of research and teaching infrastructure in social and behavioral sciences in East Asia and the Pacific. The first initiative in social science were made in Japan's before World War II. While bureaucratic authoritarianism and pervasive influence of Marxism impeded the development of research and teaching infrastructure in social and behavioral sciences, democratization under US-occupied Japan accelerated it. Second, infrastructure development in East Asia and the Pacific since 1945 is outlined. While foundations of opinion polling and survey taking were laid during colonial and authoritarian regimes, democratization and globalization have fostered and advanced social sciences in the academic world in East Asia and Pacific. Pan-Asian cooperation in social sciences is on the rise and research infrastructure is being developed that will facilitate cooperation and improving research infrastructures in the future.

Phases of Development in Japan after 1945

Modern social science originated outside Asia (Inoguchi, 1995). In Japan, modern social science initially meant, European-born Marxism planted onto Asian soil. Hence, the term social sciences described the 'opposition' science. *Oppositionswisssenschaft* meant the social theories of those opposing the regime. Opposition scientists wanted to reveal the oppressive and exploitative nature of the system. Marxist influence deeply permeated the international social sciences. Coupled with the overtly nationalistic undertones characteristic of Asia, Marxism hindered the diffusion of American behavioral and social sciences in the 1950s and 1960s. In other words, the dominance of Marxism in the social sciences in Asia posed higher barriers for American-style behavioral and social sciences, even in the 1950s and 1960s.

Furthermore, the dominance of state science, *Staatslehre*, the science that purported to be useful to the state's governance, had a lingering influence on the fledgling Japanese social sciences in the first half of the twentieth century and beyond. Instead of trying to come up with lawlike generalizations, state science was interested in supplying specific, concrete, and context-based knowledge. It encouraged historical and institutional descriptions. State science was interested in applied subjects like law and economics but not political science, sociology, or social psychology. State science was interested in history and geography, as far as they could be utilized in running the country and the colonies. Therefore, American social sciences were at first rejected because they were in stark contrast to opposition and state sciences.

Nevertheless, American-style social sciences started to permeate the rest of the world after 1945. The remarkable development of the American model of the professional academic market after the Second World War prepared the United States to excel in the social sciences in the latter half of the century. As carefully portrayed by Zunz (1998), the American model is based on a certain set of criteria within the academic community, characterized by competition and the loosely orchestrated concentration of government, industry, and the mass media to nurture and utilize academic resources and products. Furthermore, the influx of many European refugees stimulated American academics. The number of Nobel Prize winners from the United States was modest between 1901 and 1937, somewhat comparable to the figure of Japan after 1945. After 1938, however, the number of US Nobel Prize winners rose dramatically, achieving a dominant share.

Three major features of American-style behavioral and social sciences are: (1) a strong drive to come up with lawlike generalization; (2) a strong commitment to systematic empirical hypothesis testing; and (3) shared beliefs in anonymous referees producing good judgments and improved manuscripts. These features were not necessarily in harmony with the prevailing academic cultures of Japan in the 1940s and 1950s. Their interests lay in more historical and contextual descriptions, not necessarily in lawlike generalizations. Efforts to advance arguments were not necessarily made but by using systematic empirical hypothesis testing. Japanese scholars were extremely opposed to the idea of their work being anonymously refereed for publications. Thus, in the 1940s and 1950s, American-style behavioral and social sciences were not able to gain a foothold in Japan. Nevertheless, in the 1960s and 1970s, behavioral and social sciences as understood and practiced in the United States were slowly being adopted. This was also a period of extraordinary economic growth for Japan. In tandem with economic development, the academic infrastructure was consolidated, and academic orientations came of age in their national environments.

Institutional Structures of Social Sciences in Japan

The higher educational legacy of imperial Japan, which ended in 1945, contributed most directly to the shape of the academic infrastructure of the behavioral and social sciences in Japan after 1945. Its major features were:

1. an elitist structure of imperial universities with highly limited numbers of professors and enrolled students;
2. a curriculum serving the needs of the government in terms of supplying the elite cadres in applied areas of study such as medicine, agriculture, engineering, law, and economics; and

3. teaching two European languages, German and French, as well as English.

After 1945, the number of colleges and universities grew rapidly in US-occupied Japan, where each of the 47 prefectures built national universities. Furthermore, private universities grew in number in the 1960s and 1970s. The elitist nature of Japanese universities decreased considerably throughout these years. Rather, Japan registered a high admission rate of high-school graduates comparable to that of the United States, thanks to a very inclusive enrollment scheme. Curricula also changed together with the extraordinary growth of student enrollments and staff appointments from the 1960s to the 1990s. In the case of the University of Tokyo, for instance, the relative proportion of staff appointments remained the same. There have been more staff appointments in applied science areas, especially engineering, the life sciences, medicine, pharmacology, and agriculture. As of 2000, 80% of the annual budget allocation goes to those applied science areas. Only 20% goes to law, economics, literature, education, and other nonscience areas of study. The tendency of favoring applied sciences budgetwise further increased in the twenty-first century. There also seems to be an according increase in Nobel Prize winners from Japan. However, at most private universities, applied sciences have not been well prioritized in terms of faculty development and student enrollment, thanks to the general scarcity of funding as a whole.

However, this does not mean that a favorable national environment has been created for the development of the behavioral and social sciences. Within the schools of law, economics, and literature, behavioral and social sciences have not been given autonomous statuses in terms of staff appointment and budgetary allocation. The political science discipline has been included into the discipline of law (a legacy of state science); the school of economics has a curriculum giving very high importance to applied economics and history; and the school of literature has departments of sociology, social psychology, and anthropology. Moreover, the emphasis on European languages was replaced by the predominant use of English after 1945. This has facilitated the diffusion of American-style behavioral and social sciences.

Financial Setups of Behavioral and Social Sciences

The Funding of Social Science Research

Turning to the topic of institutional and financial setups, there are three major ways in which behavioral and social science research is conducted:

1. within the university system;
2. by nonuniversity institutes; and
3. project funding through national organizations and foundations.

Behavioral and social science research is predominantly conducted within the university system. Small-scale research is normally conducted by making best use of a small amount of research budget made available to each professor. At some universities, schemes for facilitating larger and more interdisciplinary research exist. However, the size of budgets tends to be small. Therefore, once research has higher funding needs, researchers can apply for more budgets through other schemes.

In addition to the university system, there are some research institutes primarily devoted to the planning, implementation, and dissemination of research on national and sometimes international spheres. They used to be entirely publicly funded but are now financed by a mix of private and public funding. For example, the National Museum for Ethnographic Research in Osaka focuses on anthropology. Thus, the institute organizes various projects with professors working at different universities. The Institute of Statistics and Mathematics in Tokyo is a similar organization that focuses on public opinion research. Some institutes work on public policy-related research topics like 'public policy priorities in aging societies', whereas others work on more purely academic subjects like 'life style, post-modern, post-Confucian, and post-Marxist'.

The system of project funding through national organizations and foundations is important in Japan. In Japan, the Japanese Society for the Promotion of Sciences plays a similar role to that of the National Science Foundation of the USA, giving financial support to various academic projects (the Japan Society for the Promotion of Sciences 2000). The National Institute for Research Advancement in Tokyo is an organization that funds mainly public policy-focused projects run by various think tanks in Japan and elsewhere. The Japan Foundation in Tokyo is a foundation financed by public funding to assist research financially and to disseminate academic, artistic, linguistic, and other professional projects globally. Many private foundations exist in Japan and South Korea in addition to those public organizations and foundations. Some focus on gender studies, academic and professional exchanges, human security, and international financial needs, whereas others give to more public policy-related projects. In Korea, the Sejong Institute is public and privately funded and works on a variety subjects in sociology, political science, and urban studies.

Financial Setups of Research Infrastructure Development

In terms of infrastructure development for the behavioral and social sciences, the university system is central and important. Most project-focused funding does not include infrastructure development. Support for developing academic infrastructures, such as joint centers for survey research, experimental laboratories, interuniversity consortiums for databases, electronic networks, or publishing houses, usually comes from public funding.

A joint university center for survey research does not exist in Japan. What exists is a 3- to 5-year partnership based on the project funding of scientific research. Within this coalition, researchers can assemble a research team and hire an opinion polling company for surveys. When the project ends, everything except the database and published volumes disappears. My own scientific research project is a recent example of this. It was funded by the Ministry of Education for over 4 years (1999–2002) and deals with 'Globalization and the Cultural Dimensions of Democracy in 18 Societies of Asia and Europe'. It has 1000 data points in each sample, collected by face-to-face interviewing (Blondel and Inoguchi, 2006; Inoguchi and Blondel, 2009; Inoguchi and Marsh, 2009). A number of universities such as the University of Tokyo and Keio University

in Tokyo have departments of social psychology and sociology, and research centers/institutes for research by mass communication. However, none of them encourage partnership/team building or undertake coordinating roles institutionally.

The same happens with experimental laboratories. Again, they must be built within project funding for scientific research and can be maintained and expanded if they continue for several years. One recent example is Toshio Yamagishi's experimental laboratory built to examine experimental and cross-national modes of trust (Yamagishi, 2000). With regard to social scientific databases, the University of Tokyo's Institute of Social Sciences has set up an interuniversity consortium for databases in a similar fashion as the University of Michigan's Interuniversity Consortium for Social and Political Research. Focusing on sociological, political, historical, and economic Japanese data, the consortium has been quite a success, albeit on a rather small scale, in organizing available databases and making them academically totally open and accessible to its members.

Election networks are funded by the Japanese Society of Promotion of Sciences and the Ministry of Education. With their funding, Ikuo Kabashima has conducted seven Japanese election panel surveys through the 1990s, producing six published volumes dealing with various aspects of Japanese electoral behavior (Kabashima, 1998–2000). Kabashima's database has largely expanded the database of the Journal called *Leviathan: The Japanese Journal of Political Science*. The database has been more narrowly focused on election data and has run on a much smaller scale.

Publishing houses dealing with behavioral and social science books do not necessarily thrive. Given the fact that orders by major university, libraries are irregular, academic publishers cannot anticipate amount of books that will be purchased. In this sense, the infrastructure for publishing research needs to be further consolidated. There is no meaningful contribution of English-language research to research publications. Kodansha International is the only commercially self-sustainable publisher in the English language. However, this publisher is not primarily focused on academic books. The United Nations University Press in Tokyo was quite successful in the late 1990s in continuously publishing good academic books on such topics as democracy, humanitarian interventions, and the global environment, albeit on a modest scale. Their publications have been distributed in North America by the Brookings Institution Press and have also been included in the Columbia International Affairs Online (CIAO) Internet service on new publications. University presses such as the University of Tokyo Press and Kyoto University Press have English-language publishing departments. However, they only publish sporadically and their published books are mostly based on translations of books originally written in Japanese.

Research and Teaching Infrastructure in Social and Behavioral Sciences in East Asia and the Pacific in the Twentieth and Twenty-First Centuries

The Foundations of Social Science under Colonial Rule

The development of research and teaching infrastructure in social and behavioral sciences in East Asia and the Pacific is related both to the governance under colonialism and after independence or in other words to authoritarianism and to democratization. The first steps toward setting up a research infrastructure were made under colonialism and post-independence authoritarian regimes. These regimes were interested to find out preferences and issues of citizens living under colonial or postindependence rule. For instance, in 1937, opinion polling was carried out for the first time in British Malaya because colonial authorities wanted to understand better what people perceived and preferred with regard to the kind of news items through Singapore Radio Station. However, opinion polling was only carried out sporadically under colonial rule. As Sen notes (1983), British India's colonial authorities were indifferent to poverty and famine, whereas the postindependent Indian government has been far more attentive to poverty with the result that the incidences of large-scale famine dramatically decreased. Sen credits Indian statisticians in developing infrastructure in gauging poverty and implementing rescue operations. Postindependence authoritarian governments did make use of opinion polling, but it was limited because for reasons of internal security and of constraints on freedom of expression, opinion polling never blossomed. Under authoritarianism questions tended to elicit progovernment voices unduly, whereas respondents tended to voluntarily suppress what might be taken as disturbing authoritarian governments.

The Fostering of Social Science Practice with Democratization

A more fully fledged development of research infrastructure came in tandem with democratization. It was in the 1980s during which opinion polling and other research and teaching infrastructure started to blossom. The advent of democratization in East and Southeast Asia meant that politics had to pay more attention to what and how people felt thought and decided on issues related to politics (Marsh et al., 1999; Blondel and Inoguchi, 2006; Inoguchi and Blondel, 2009; Inoguchi and Marsh, 2009; Chu and Wong, 2010).

In 1986, anti-Marcos people power party in the Philippines put an end to dictatorship. Consequently, opinion polling became more important. The Social Weather Stations (SWS) was founded in 1986 in order to regularly poll the people's opinions without government funding. In a similar vein, in Taiwan, opinion polling became a routine after Jiang Jinguo initiated the path toward democratization (Yu, 2012). In South Korea, democratization began under Roh Tae Woo, who was first democratically elected president. A democratic regime was consolidated under Kim Dae Jung who was the second democratically elected nonmilitary president from 1998 to 2003 and after (Cho, 2012). In Thailand, a constitutional revision took place in 1992, which paved the way toward democratization. Under Thaksin Shinawatra and his successor Ingluck Sinawatra, opinion polling became a common practice. Authoritarian regimes in Malaysia felt the need to loosen regulations by asking academics at the Universiti Kebansan Malaysia to carry out an opinion poll on voters and voter intentions 'like it was being done in the United States' in 1986 (Idid, 2012).

The Influence of Globalization on Research Infrastructures

Another important impetus in the 1990s and early 2000s was the steady tide of globalization. New advances in communication technology were driving forces that made the world 'flat', enabling a vast number of people all over the world to communicate and connect to each other with incredibly fast speed, at low costs and give access to massive content (Friedman, 2000). This had two effects on the infrastructure of social sciences. On the one hand, globalization enabled a massive number of people to participate in politics. Especially, the Internet and social media has supported people and interest groups in connecting to each other. These groups are often larger than the small governing elites. Globalization in this sense is contributing to democratization. About 300–500 million users are engaged in Chinese Social media, which emphasizes the possibility for China's vast population to connect and exchange thoughts and feelings (McDougal, 2011). Positively, people's voices are far more widely expressed, thanks to the ease of communication technology. To gauge the vox populi, opinion polling has advanced by leaps and bounds. On the other hand, the prompt and often massive responses to the regime's action or nonaction often results in the regime's sources of procrastination, wavering, or even termination to the initial action or nonaction itself.

Strengthening the Asian Research Base

Social science research in Asia in general and particularly in Japan and Korea has become well institutionalized in its respective national settings and organizational structures in the second half of the twentieth century. The two strong legacies of state science and opposition science that were prevalent in the first half of the twentieth century have weakened in Japan and Korea. Nevertheless, they provided a fairly solid foundation for social science research in the second half of that century (for such developments in political sciences in Japan, South Korea, and China, see Inoguchi, 2012a). Although its infrastructure needs to be improved in the future, the stage is clearly set for further consolidation of social sciences, especially in light of the two pronounced trends of international cooperation and cross-national comparisons.

In Japan and Korea, as well as in Asia in general, international cooperation began with the increasing collaboration with the United States. The United States have been a senior partner for the larger part of Asia's recent history. Through bilateral research and teaching with the United States, both Japan and Korea enhanced their national bases of social science research by the end of the twentieth century. By then, it had become obvious that international cooperation had become a trend in the United States, as well as in Asia. In much of the latter half of the twentieth century, regional cooperation did not flourish much partly because of diverging economics and politics. However, during the 1990s, Asian-Pacific countries, specifically Japan, Korea, China, and the 10 Association of Southeast Asian Nations members, came to interact among themselves.

It is not surprising that the new trend of enhanced intraregional research cooperation is in broad harmony with the increase in intraregional summit meetings. It may seem as if the newly found regional (i.e., Asian) identity supported this trend. However, the increase in intraregional research cooperation was primarily because of a growing interest in cross-national comparisons. After recording national achievements in terms of per capita income and literacy levels, it has become a norm to compare cross-national propositions. This has replaced the traditional approach of focusing on national patterns and developments. Nevertheless, there is still a long way to go until we can speak of an overall Asian social science. Yet, a number of such seeds have already been sown and the prospects are not bad. Globalization has changed the state–society relationship tremendously. An Asia–Europe survey comprising nine societies from Asia and nine societies from Europe was carried out in 2000 under the leadership of Takashi Inoguchi and Jean Blondel (Blondel and Inoguchi, 2006; Inoguchi and Blondel, 2009; Inoguchi and Marsh, 2009). The survey aimed to examine how citizens relate to the state by focusing on three variables: identity, trust, and satisfaction. This project represented a new method of establishing and enhancing intraregional and interregional cooperation. Academics in both regions benefit from working together: Asians learn from each other and build infrastructure for cooperation. Europeans discover more similarities than differences between Europe and Asia.

Another survey that was carried out was the General Social Survey in Japan, Taiwan, South Korea, and China, which used national resources to coordinate a questionnaire in order to conduct cross-national comparisons (Iwai et al., 2011). Its impetus was the global diffusion of US-led general social survey in East Asia.

Up until the late 1990s, the leadership for social science research cooperation in Asia–Pacific came primarily from the United States. Since then, there have been many local initiatives for intraregional cooperation, originating in Korea, Japan, Singapore, and Australia. As there has not been an all-inclusive regional social science research council taking conceptual leadership and funding facilitation, like the United States Social Science Research Council has long done, there are no definite outcomes at this point. Nevertheless, there are many budding initiatives and publications on subjects, such as national identity, globalization, democratization, human rights, and civil society. Universities such as the University of Tokyo, the University of Hong Kong, Yonsei University, Academia Sinca, Chularonkorn University, the National University of Singapore, and the University of Malaya as well as think tanks like the Institute of Southeast Asian Studies (Singapore), the Japanese Center for International Exchange (Tokyo), and the United Nations University (Tokyo) are major agents for more academic research. With the resulting, cumulatively enhanced networks within Asia–Pacific, one can hope with some cautious optimism that there is a development of pan-national social science in the region. Interregional social sciences would surely synthesize a diversity of views and angles that would add to empirically testable comparisons on major trends and patterns of democratization, digitalization, and globalization and their differentiated impacts on local societies.

Asian Barometers and Other Types of Scientific Research

After the first edition of The International Encyclopedia of Social and Behavioral Sciences was published in 2001, a tidal wave of development took place in terms of infrastructure buildup in East Asia and the Pacific. The American Political Science Section of Comparative Democratization's Newsletter published a special issue on regional barometers in which the leaders of key regional barometers summarized their aims, scopes, and findings. Along with the Afro barometer, the Arab barometer, the Latinobarometer, the Americas barometer representing each world region, the AsiaBarometer (Inoguchi, 2012b) and the Asian barometer are allotted barometers, representing Asia. The AsiaBarometer focuses on the quality of life (Shin and Inoguchi, 2009; Inoguchi and Fujii, 2012; Inoguchi, 2013), whereas the Asian barometer focuses on democratization (Dalton et al., 2008; Chu et al., 2010). The former has been carried out seven times, in almost every year of the 2000s (2003, 2004, 2005, 2006, 2007, 2008, and 2010), covering the entire Asian continent (East, Southeast, South, and Central). The Asian barometer has been carried out three times since 2002, mostly in East and Southeast Asia. The former is led by Takashi Inoguchi and his pan-Asian team, whereas the latter is led by Yun-han Chu and his hybrid team of East Asian and American academics. In this regard, the AsiaBarometer is distinguished from all the rest in that the rest are led by the combined American and regional leaders or they follow more or less standardized schemes of democratization adapted to each region even when leaders are from each region.

Besides opinion polling, other kinds of behavioral and social scientific research have blossomed recently. Social psychological experimental research has advanced especially in the fields of trust and cooperation (Yamagishi, 2000). Using prisoner's dilemma examples in game theory, Yamagishi and his team reveal cross-national similarities and differences among Japanese, American, and Chinese subjects. It highlights that Americans and Chinese are more proactive in taking cooperative initiatives, whereas the cautious Japanese tend to decline cooperative initiatives at first. Neurological behavioral research has flourished more recently focusing on neuroeconomic and neuropolitical behavior with regards to values and group formation, justice, and utilitarian calculus (Takahashi, 2011; Kato and Okada, 2011). It also highlights the incongruence between neurological response and verbal response in judging happiness, efficiency, and rights. Computer simulation is another area, registering behavioral research advances. On the basis of almost verbatim decisions of leaders leading to the Cuban missile crisis, Yamakage and his team have advanced the use of computer simulation one step forward. More system-founded computer simulations are also being developed to show insight into hegemonic change (Yamakage, 2011). Competing models of world leadership transition are shown with contrasting results.

Besides regional barometers, other infrastructural development should be noted. The SWS of the Philippines has carried out democratization and all the related subjects in the Philippines since 1986. Its impetus is Philippine democratization in 1986. The Center for the Study of Developing Societies, New Delhi, has a section of opinion polling in India, in which Yogendra Yadav and his team have carried out democratization studies on a subcontinental scope (Stepan et al., 2011). Public Opinion Program of the University of Hong Kong led by Robert Chung has invented a scheme of polling on citizens' political preference and some economic indicators. Its impetus is the return of Hong Kong from Britain to China in 1997 (Lee, 2012).

Looking Forward

Into the second decade of the twenty-first century, the prospect for progress in research and teaching infrastructure in social and behavioral sciences in East Asia and the Pacific is bright. Not only Japan and Korea but also Taiwan, Hong Kong, Singapore, China, Macao, Thailand, the Philippines, Malaysia, and Indonesia have been consolidating their infrastructure and research products based on such new infrastructure quite steadily and increasingly in concert in the region. I highlight three new developments in this regard. First, the region-wide associations and networks are sprouting and flourishing. The Asian Network of Public Opinion Research (ANPOR) was established with close ties with the World Association of Public Opinion Research and the American Association of Public Opinion Research. The first annual meeting will be held in Seoul in November 2013 some 100 members assembled from East Asia and the Pacific. Its journal, the Asian Journal of Public Opinion Research, has started publication in digital format in November 2013. The Asian Consortium for Political Research (ACPR) has been reconfigurating and consolidating itself with a new journal provisionally called Asian Political Science Review to be launched in 2015.

Second, research cooperation across national borders has started to flourish. Quality of life and democracy are two of the favorite subjects on which research cooperation has yielded most. The AsiaBarometer survey directed by Takashi Inoguchi was able to cover the entire Asian region (29 countries and areas) with focus on quality of life. The (East) Asian barometer headed by Yun-han Chu with the Asian and American team was able to cover some of East Asian and Southeast Asian countries with focus on democratization. More recently, the ANPOR has conducted its cross-national survey with focus on daily life satisfaction in Japan, Korea, Thailand, China, India, Malaysia, the Philippines, Macao, Taiwan, Hong Kong, Myanmar, and Pakistan. The Center for Empirical Political Science Research at the University of Niigata Prefecture has started to conduct e-mail survey research with focus on diverging varieties of democracy (direct democracy, procedural democracy, communitarian democracy, monitory democracy, and cosmopolitan democracy) through cooperation of political scientists in three academic associations: i.e., the American Political Association, the European Consortium for Political Research, and the ACPR.

Third, new forms of interdisciplinary research are on the rise. Utilizing insights from neuroscientific findings, Ken Suzuki (2013) has proposed the concept of divicracy. The notion of individual is tightly embedded with the modern

theory of representative democracy. Neuroscience has made it crystal clear that one person is divisible, i.e., one person can have many ideas and preferences without insisting on consistency and integration. Suzuki's mathematical formulation of divicracy based on this neuroscientific insight is iconoclastic of modern Western thought. Here, we find an infusion of neuroscience, physical science, and ethics. In East Asia and the Pacific, teaching often takes interdisciplinary forms and those barriers, which used to stand between humanities and social sciences on one hand and natural and medical sciences on the other seem to be reduced substantially in many schools. Although future developments are not quite clear at this moment, I am confident that the future is bright in the development of research and teaching infrastructure in social and behavioral sciences in East Asia and the Pacific.

See also: Comparative Studies: Method and Design; Data Bases and Statistical Systems: International Comparative Research; Data Bases and Statistical Systems: Political Science (General); East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Politics; Multidimensional Scaling I; Religions of East Asia; Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations).

Bibliography

Blondel, Jean, Inoguchi, Takashi, 2006. Political Cultures in Asia and Europe. Routledge, London.

Cho, Sung Kyum, 26–28 November 2012. The infrastructure of public opinion research in Korea. In: Asian Conference on Public Opinion Research 2012. University of the Thai Chamber of Commerce, Bangkok. Boromarajonani College of Nursing Saraburi, Thailand.

Chu, Yin-Wah, Wong, Siu-lun (Eds.), 2010. East Asia's New Democracies: Deepening, Reversal, Non-Libeeal Alternatives. Routledge, London.

Chu, Yun-han, Diamond, Larry, Nathan, Andrew, Shin, Doh Chull (Eds.), 2010. How East Asians View Democracy. Columbia University Press, New York.

Dalton, Russell, Shin, Doh Chull, Chu, Yun-han, 2008. Party Politics in East Asia: Citizens, Elections, and Democratic Development. Lynne Riener, Boulder.

Friedman, Thomas, 2000. The Lexus and the Olive Tree: Understanding Globalization. Anchor, reprint edition, New York.

Idid, Syed Arab, 26–28 November 2012. Public opinion research in Malaysia. In: Asian Public Opinion Research Conference, Bangkok.

Inoguchi, Takashi, 1995. Democracy and the development of political science in Japan. In: Easton, David, Gunnel, John (Eds.), Regime and Discipline: Democracy at the Development of Political Science. University of Michigan, Ann Arbor, pp. 269–293.

Inoguchi, Takashi, 2012a. Political science in three democracies, disaffected (Japan), third wave (Korea) and possibly fledgling (China). In: Trent, John, Stein, Michael (Eds.), The World of Political Science: A Critical Overview of the Development of Political Studies around the Globe: 1990–2012. Barbara Brudich, Opladen, pp. 11–40.

Inoguchi, Takashi, 2012b. AsiaBarometer's achievements, underutilized areas of the survey materials, and future prospects. American Political Science Association's Comparative Democratization Section Newsletter 10 (3), 2, 13–15.

Inoguchi, Takashi, 2013. Asia no joho bunseki dai jiten (Encyclopedia of AsiaBarometer's Data and Analysis). Nishimura Bookstore, Tokyo (in Japanese).

Inoguchi, Takashi, Blondel, Jean, 2009. Citizens and the State Attitudes in Western Europe and East and Southeast Asia. Routledge, London.

Inoguchi, Takashi, Fujii, Seiji, 2012. The Quality of Life in Asia: A Comparison of the Quality of Life in Asia. Springer, Dordrecht.

Inoguchi, Takashi, Marsh, Ian, 2009. Globalization, Public Opinion, and the State; Western Europe and East and Southeast Asia. Routledge, London.

Iwai, Noriko, et al., 2011. Cultures and Values in East Asia: East Asian General Social Surveys in Japan, Korea, China and Taiwan. Nakanishiya Publisher, Kyoto (in Japanese).

Kabashima, Ikuo (Ed.), 1998–2000. Studies in Japanese Electoral Behavior, 6 vol. Bokutakusha, Tokyo (in Japanese).

Kato, Junko, Okada, Kensuke, 2011. Simplification and shift in cognition of political difference: applying the geometric modeling to the analysis of semantic similarity judgment. PLoS One 6 (6), 1–9.

Keane, John, 2010. The Life and Death of Democracy. Simon and Schuster, London.

Lee, Winnie, 26–28 November 2012. Opinion polling in Hong Kong. In: Asia Conference on Public Opinion Research, Bangkok.

McDougal, Mathew, 2011. The Chinese Social Media Universe. Amazon Digital Service, Inc.

Sen, Amartya, 1983. Poverty and Famines: An Essay on Entitlement and Deprivation. Oxford University Press, Oxford.

Shaw, Kanyika, 26–28 November 2012. Public opinion research landscape in Thailand. In: Asian Public Opinion Research Conference, Bangkok.

Shin, Doh Chull, Inoguchi, Takashi (Eds.), 2009. The Quality of Life in Confucian Asia: From Physical Welfare to Subjective Well-Being. Springer, Dordrecht.

Stepan, Alfred, Linz, Juan, Yadav, Yagendra (Eds.), 2011. Crafting State-Nations: India and Other Multinational Democracies. Johns Hopkins University Press, Baltimore and London.

Suzuki, Ken, 2013. Namerakana syakai to sono teki (The Smooth Society and the Enemies). Keiso Shobo, Tokyo.

Takahashi, Taiki, 2011. Neuroeconomics of suicide. Neuroenddocrinology Letters 32 (4), 400–404.

Turkes, Shelly, 2012. Alone Together. MIT Press, Cambridge.

Yamagishi, Toshio, 2000. The Structure of Trust. University of Tokyo Press, Tokyo (in Japanese).

Yamakage, Susumu, 2011. Guidance to the Construction of Artificial Society. Shobo Hayama, Tokyo (in Japanese).

Yu, Eric Chenhua, 2012. The Role of Opinion Polls in Taiwan's Elections. Asian Public Opinion Conference, Bangkok, 26–28 November 2012.

Zunz, Olivier, 1998. Why the American Century? University of Chicago Press, Chicago.

Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations)

Gordon CK Cheung, School of Government & International Affairs, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

East Asia and the Pacific are economically developed. Yet, we know very little about the research, knowledge generation, and broader interconnections of universities and colleges in this region. After becoming industrially developed, what kind of problems and issues is this region faced with academically? Who are the institutions that can inform governments, and offer policy recommendations? Are they different from other research institutions in the Western world? This article introduces 10 social science research academies from the US, Australia, China, Singapore, Taiwan, Hong Kong, Japan, and Korea in order to understand their role, functions, and characteristics.

Academic Excellence in East Asia and the Pacific

After more than three decades of economic growth from the 1970s to the 1990s, the East Asian and Pacific region have entered a new age of technological advancement and innovation through research and development (Kim and Nelson, 2000). Universities, academies, and businesses (those elements which propelled US technologically in the developmental change in the 1970s and the 1980s) are now redefining the East Asian and Pacific countries' technological development and economic growth (Jorgenson, 2001). As Harvey Brooks and Lucien Randazzese mention: "Starting in the 1970s, and increasingly in the 1980s and 1990s, federal and state governments in the United States have established programs to stimulate formal university-industry relationships" (1998: 361). Having witnessed the success of the US in linking universities with industries, East Asia (China in particular) has invested heavily in research institutions and universities so they can compete with world class universities (Ma, 2007: 31–53). On the social sciences front, research and development in democratization, industrialization, regionalism, poverty reduction, economic development, social changes, political development, etc., regional governments require enormous support, insights, and policy recommendations from research institutions, centers, or academies. This is why social science academies in East Asia and the Pacific deserve detailed attention.

In a BBC report in February 2013, Prof Alan Smithers from Buckingham University admitted that "there was a group of countries and territories that generally performed better than England in all tables [education and tests], including Hong Kong, Japan, Singapore and Taiwan" (BBC news http://www.bbc.co.uk/news/education-21296635 (accessed 05.02.13.)). He went on to say that UK educational institutions should try to learn from these countries or territories to be more competitive in the world. In the *Index of Cognitive Skills and Educational Attainment 2012* (a global comparison of education prepared by Pearson and the Economist Intelligence Unit), four of the top five countries were in East Asia, in addition to Finland which ranked first: South Korea (2nd), Hong Kong (3rd), Japan (4th), and Singapore (5th). Sir Michael Barber, Pearson's chief education adviser, attributed the success of these East Asian economies of having a 'culture' of education (Sean Coughlan, "UK education sixth in global ranking" http://www.bbc.co.uk/news/education-20498356 (accessed 05.02.13.)). On the global higher education comparison, the *Times Higher Education World University Ranking of 2012* indicated that five universities from East Asia ranked among the top 50, between 23rd and 50th (Times Higher Education World University Ranking http://www.timeshighereducation.co.uk/world-university-rankings/2013-14/world-ranking/region/asia (accessed 01.11.13.)). Turning to the *QS World University Rankings 2013*, the picture has become more obvious for the rise of East Asia in global higher education competition. Among the world's top 50 universities, there were 10 universities from East Asia, ranked from 24th to 48th (QS World University Rankings http://www.topuniversities.com/university-rankings/world-university-rankings/2013#sorting=rank+region=71+country=+faculty=+stars=false+search= (accessed 01.11.13.)).

Kishore Mahbubani, Dean of the Lee Kuan Yew School of Public Policy at the National University of Singapore, contended that the shift of the global power from the West to the East is 'irresistible' (2008). Hong Kong, Taiwan, Singapore, and China also witnessed the extraordinary investment in tertiary education and the expansion of many research institutes, postgraduate programs, and overall higher education expansion in the 1990s and 2000s, which resulted in a golden era of higher education development in East Asia.

In the following sections, we are going to introduce 10 research institutes or academies in East Asia and the Pacific region to focus on their function, role, characteristics, development, etc. We begin with two institutes in the Pacific – the East-West Center (EWC) in Hawaii, US, and the Academy of the Social Sciences in Australia (ASSA). These two academies represent two western countries' investment in policy research, academic exchanges and cultural development in East Asia and the Pacific. Then, we look at four academies in the Greater China region; the Chinese Academy of Social Sciences (CASS) in the People's Republic of China, the Academia Sinica in Taiwan, the Universities Service Centre for China Studies (USCCS), and the Hong Kong Institute of Asia–Pacific studies (HKIAPS), both at the Chinese University of Hong Kong. CASS, under the State Council, is the largest social sciences academy in China. The Academia Sinica is the highest and most prestigious academy in Taiwan, and includes physical and social sciences. The USCCS is a university level research institute with

a long history. Previously known as USC, it is a well-known institute focusing on contemporary Chinese studies. The HKIAPS has 23 years of history and increasingly performs an important role in linking Hong Kong between China and the Asian Pacific region with its high quality academic research and policy-oriented studies.

In addition, we will examine the East Asian Institute (EAI) and the Institute of Southeast Asian Studies (ISEAS), both at the National University of Singapore. EAI is a research institute focusing on China and its relations with the East Asian region; while the ISEAS is one of the largest research institutes in the world focusing on Southeast Asian studies. Finally, we look at the Institute of Social Science (ISS) of the University of Tokyo in Japan and the East Asia Institute (EAI) in South Korea. The ISS has more than 50 years of history in applying social sciences to serve Japan's economy and society. The EAI is a renowned think tank looking at the international relations of the East Asian region. The choice of the academies reflects the aforementioned transformation of East Asia from industrial and manufacturing workshops to knowledge centers. This list is by no means comprehensive. The selected institutions merely aim to cover some major academies which have had significant impact on academia and policy-making.

The East-West Center (EWC), Hawaii (US)

The EWC was established by the US Congress in 1960 in order to promote "better relations and understanding among the people and nations of the United States, Asia, and the Pacific through cooperative study, research, and dialogue" (http://www.eastwestcenter.org/about-ewc/mission-and-organization (accessed 10.11.13.)). Its 21-acre Honolulu campus, designed by the famous Chinese architect, I.M. Pei, is next to the University of Hawaii, Manoa. The choice of Hawaii was unique in the sense that the geographical location is at a meeting point between the Asian Pacific region and the United States mainland. Established during the height of the Cold War, the EWC today provides many academic courses and offers a forum for free-thinking among the students, fellows, and visitors. In its 50 years history, it has already produced more than 57 000 alumni, and has more than 750 partner organizations.

The EWC is governed by an academic board of 18 members and currently chaired by Dr Charles Morrison. Its cultural, academic, research, and exchange missions are supported by nine related programs and bodies: the EWC Research Program, the EWC Education Program, the EWC Seminars Program, the EWC Pacific Islands Development Program, the EWC in Washington, the East-West Center, Office of External Affairs, the EWC Foundation, and the Office of Administration.

Through these nine fellowship programs, students, researchers, scholars, and academics can fully be engaged with the EWC. Apart from these fellowships, students can apply for the Asia Pacific Leadership Program Fellowship (the flagship program which has run for 13 years), the Asian Development Bank-Japan Scholarship Program (a link-up program with the University of Hawaii, Manoa), the East-West Center Graduate Degree Fellowship (a 24-month program supporting Master or PhD students), the Jefferson Fellowships Program (a 3-week short program), the Obuchi Student Scholarship, the Pohang Iron and Steel Company (POSCO) Visiting Fellowship Program (2-month program in relation to Korean studies), the ProSPER.Net Leadership Program (jointly developed with the United Nations University Institute of Advanced Studies [UNU-IAS]), and the United States–Timor-Leste (USTL) Scholarship (a newly developed program).

With the expert knowledge and specialization of fellows, visitors, and students, the EWC focuses on three major research areas, the Environment, Population, and Health; Innovation, Energy and Economic Growth; and Governance, Security and Justice. Research results, policy recommendations and reports, etc. are disseminated through seminars, conferences, and workshops, as well as through publications such as the *Asia Pacific Bulletin*, *AsiaPacific*, and other policy briefs and recommendations. Finally, the EWC in Washington allows more engagement and networking between policy makers and academic scholars through outreach activities, seminars, and publications.

Academy of the Social Sciences in Australia (ASSA)

As one of the oldest academies in Australia, the main objective of the ASSA is "… to advance the development of world-class social sciences in Australia and their contribution to local, national and global well-being" (http://www.assa.edu.au/about/plan (accessed 12.11.13.)). The ASSA was established on 7 July 1971 with 94 members on the Social Science Research Council of Australia as foundation fellows. In connection with the Australian Academy of the Humanities (AAH), Australian Academy of Science (AAS), and the Australian Academy of the Technological Sciences and Engineering (ATSE), these four academies provide interdisciplinary advice to the Australian government through the Australian Council of Learned Academies (ACOLA). According to the Strategic Intentions 2013–15, ASSA aims to achieve four missions:

- Promoting research, teaching, and advice in the social sciences
- Promoting national and international scholarly cooperation across disciplines and sectors, including collaboration with other academies
- Commenting on national needs and priorities in the social sciences
- Providing advice, especially on issues of national importance (http://www.assa.edu.au/about/plan (accessed 16.11.13.))

To facilitate these missions, the ASSA has developed five programs: An Annual Symposium, Public Lectures, Roundtables, State-based Fellows' Initiatives, and the Workshop Program. The Annual Symposium, the flagship program of the ASSA, is organized in November in Canberra. This one-day workshop, together with the concomitant Cunningham Lecture, addresses key issues in Social Sciences. Public Lectures, sponsored by the ASSA, are hosted by individual universities in order to help the outreach of ASSA. Academy Roundtables are forums to provide policy input and practical perspectives through interaction with high-ranking

government officials, leading academics and academic fellows. The State-based Fellow's Initiatives funds and helps fellows to organize their own workshops to deal with local issues. The Workshop Program performs a key function of the ASSA which is to tap into cutting edge knowledge in Social Sciences, especially in applying technology in lifting the productivity of Australia.

Currently, three funding and award schemes of ASSA are the major vehicles to promote academic excellence and global outreach. The International Program provides funding for established or young academics in Australia to travel overseas and build academic partnerships with other leading academics in the world. The Paul Bourke Award for Early Career Research (formerly Academy Early Career Award) is an annual award for outstanding academic work in Australia, chosen from four nominees in their respective social sciences categories. The Workshop Program grants academics for the organization of a 2-day workshop to promote the discussion of topical social science issues with policy significance. It generally funds six academic workshops annually. The dissemination of ASSA's research is done mainly through dialogs, papers, research publications, symposium publications, and workshop publications.

The Chinese Academy of Social Sciences (CASS)

The Chinese Academy of Social Sciences (CASS), previously the Department of Philosophy and Social Sciences under the Chinese government was established in 1977. It was almost completely destroyed during the Cultural Revolution (1966–76). The first president of the academy was Qiaomu Hu (1977–82). The current President is Weiguang Wang. It is the highest research organization under the State Council of the People's Republic of China. Included in it are more than 31 research institutes and 50 research centers, covering 260 research areas, such as law, economics, political science, sociology, geography, demography, linguistics, American studies, Japanese studies, European studies, Asian Pacific studies, Eastern European studies, Russian and Central Asian studies, African studies, Latin America studies, world economics and politics, and others. In total, CASS employs more than 3767 staffs, of which 2975 are researchers (http://bic.cass.cn/english/InfoShow/Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS (accessed 23.11.13.)).

CASS aims to promote theoretical as well as policy-oriented research on various disciplines of the social sciences. Among the various research activities are a program for visiting scholars, joint research collaboration, training programs, seminars, workshops, and public lectures. It has built up research connections with more than 200 global research institutions. While, CASS is serving a research purpose, the administrative arm of CASS is the Bureau of International Cooperation (BIC), which takes care of all the administrative works, including liaising and setting up agreements with foreign institutions, facilitating scholars and researchers in getting visas and passports for overseas research, and facilitating foreign scholars coming to China to carry out researches, etc. In the publication *2012 Think Tanks and Civil Societies Program*, CASS ranked 5th out of the 100 non-US think tanks and 1st among the 45 think tanks in the Northeast Asian region (2013: 42 and 57).

Academia Sinica, Taiwan

Academia Sinica was established in 1928 in the Republic of China (ROC) as the most prestigious academic institute researching physical and social sciences. After the ROC was moved from China to Taiwan in 1949, Academia Sinica gradually reestablished itself in Taipei. Particularly its former President, Yuan-Tseh Lee, (Nobel Prize winner in Chemistry 1986) made great strides in advancing Academia Sinica. There are now 24 research institutes and seven research centers covering three broad areas included in Academia Sinica: mathematics and physical sciences; life sciences; and humanities and social sciences (http://home.sinica.edu.tw/en/about/history_and_mission.html (accessed 24.11.13.)).

The highest academic institute of the ROC, Academia Sinica is directed by the Council of Academia Sinica, its President, and the Convocation. The Convocation is made up of 261 Academics, of whom 111 are active in Mathematics and Physical Sciences, 88 in Life Sciences, and 62 in Humanities and Social Sciences. The Convocation performs four functions:

- To elect academicians and honorary academicians
- To elect members to the Council of Academia Sinica
- To design policies on academic research
- To plan and carry out research at the government's request (https://db1n.sinica.edu.tw/textdb/academicians/index.php?lang=en (accessed 24.11.13.))

According to Chi Huey Wong, the current President, "… the foremost academic research institution in Taiwan, Academia Sinica has the mission to promote cutting-edge research in the humanities and sciences, cultivate academic talent and issue policy advisories to the government" (http://home.sinica.edu.tw/en/about/message.html (accessed 24.11.13.)).

Within Humanities and Social Sciences, the Academia Sinica has 12 research institutes. They are the (1) Institute of History and Philosophy, (2) the Institute of Ethnology, (3) the Institute of Modern History, (4) the Institute of Economics, (5) the Institute of European and American Studies, (6) the Institute of Chinese Literature and Philosophy, (7) the Institute of Taiwan History, (8) the Institute of Sociology, (9) Institute of Linguistics, (10) the Institute of Political Science, (11) the Institutum Iurisprudentiae (Institute of Law), and (12) the Research Center for Humanities and Social Sciences. Although they are under the umbrella of the Academia Sinica, each institute has its own administrative and research staff, and its own library and conference facilities. In fact, in order to qualify as an institute under the Academia Sinica, the institute has to go through a 10-year probation period, in which numerous assessments are used to evaluate its research strength and capacity.

Apart from research, the Academia Sinica actively promotes higher education, international academic exchange programs, organizing lecture series, providing international faculty services, etc. These and many other activities are facilitated by the International Affairs Office. The Academia Sinica also offers

degree programs and higher degree programs. Its PhD degree program, established in 2006, is facilitated through its Taiwan International Graduate Program (TIGP) and the partnership with other leading universities in Taiwan.

The Universities Service Centre for China Studies (USCCS)

Among senior scholars and academic workers involved in contemporary Chinese studies of the People's Republic of China (PRC), the importance of the University Services Centre (USC) is well-known. Its nickname, the 'Mecca for China Studies,' speaks for itself. Acquiring academic footage in the PRC during the Cold War and in particular during Mao's era was highly difficult, if not totally impossible. One of the key difficulties facing Western (even Chinese) scholars was the location of original sources, materials, and archives of the PRC. The idea of developing a center on contemporary Chinese studies was first conceived by Lucian Pye, professor at Massachusetts Institute of Technology (MIT), and Bill Marvel, president of Education and World Affairs in the US during the 1950s. (Another two centers on China studies were at the University of Washington and Harvard University.) They were thinking of setting up a center in Hong Kong, as a gateway between China and the West. The Social Science Research Committee and the American Council of Learned Societies helped the setting up of a scholars' committee which was funded by Carnegie–Mellon and the Ford Foundations (Vogel, 2005: 2–3).

The center was first established in an old building in 155 Argyle Street in Kowloon, Hong Kong. Its first director was Jerome Cohen. It moved to the Chinese University of Hong Kong in 1988 and was later renamed the Universities Service Centre for China Studies (USCCS). According to Ezra Vogel (Emeritus professor from Harvard University), "…several of the academics who conducted research at the Centre became themselves distinguished government servants: Stephen Fitzgerald became the first Australian ambassador to China, Richard Solomon became an assistant to Henry Kissinger in the early stages of the US–China relationship, Michael Oksenberg was central to normalizing relations between the United States and China, and David Wilson, the last civil servant to be governor of British Hong Kong, played a key role in negotiating China's resumption of sovereignty over the territory" (Vogel, 2005: 1).

USCCS helped nurture many academics with its abundant resources during the time when China was still closed to the outer world and research materials of the PRC were scarce. In terms of its research and archive, it owns more than 250 provincial and national newspapers from mainland China and over 1500 periodicals (Chinese mainly) from as early as the 1950s. As a center for Chinese studies, it houses one of the most comprehensive lists (over 1300) of Chinese regional and statistical yearbooks. It has its own library with more than 80 000 Chinese and English books on contemporary Greater China and the East Asian region. It is also equipped with online materials, electronic resources, microfilms, and other digitized scholarly papers on contemporary China, in both English and Chinese (http://www.usc.cuhk.edu.hk/Eng/Default.aspx (accessed 29.11.13.)). In terms of research dissemination, a weekly luncheon seminar where visiting scholars, researchers, and students all mingle to share research findings has long been a tradition of the USC, Some of the results will be published within the USC Seminar Series. Starting from 2005, the USC began to launch the Graduate Seminar on China, an annual event to promote Chinese studies across Asia and the world.

The Hong Kong Institute of Asia–Pacific Studies (HKIASP)

HKIAPS was established in 1990 as a result of its former Vice-Chancellor Professor Charles Kao's (Nobel Prize winner in Physics 2009) initiative to develop an institute of social studies supporting research on Hong Kong, China, and the Asian Pacific region. It is managed by a committee of 11 people. In its 20th Anniversary publication *HKIAPS at 20*, it laid out its mission to "promote multidisciplinary research on social, political, and economic development of Hong Kong, mainland China, and the Asia–Pacific region; to promote academic exchange; and to disseminate research findings" (2010: 3). According to *HKIAPS at 20*, HKIAPS has published 87 research monographs, 212 papers, two public policy forum papers and 18 USC seminar papers (2010: 90). Its current Director is Professor Fanny Cheung. According to its *Annual Report 2011–12*, HKIAPS had organized 96 conferences, symposia, seminars, public lectures touching on emerging issues on social, economic, and political areas as well as regional development (2012: 2).

After more than two decades of growth and development, HKIAPS has expanded to a staff of around 100 senior and junior researchers and administrative staff. Under the umbrella of the HKIAPS, it houses 10 research centers: (1) The Center for Chinese Family Studies; (2) the Center for Civil Society Studies; (3) the Center of Housing Innovations; (4) the Center for Quality of Life; (5) the Center for Social and Political Development Studies; (6) the Economic Research Center; (7) the Gender Research Center; (8) the International Affairs Research Center; (9) the Public Policy Research Center; and (10) the Research Center for Urban & Regional Development. In addition, it also includes seven strategic research programs: the Chinese Law Program; the Economic Policy Program; the Financial Markets Programs; the Program for Economic Education; the Social Indicators and the Social Development of Hong Kong Program; the South China Program; and the Trade and Development Program. The Telephone Survey Research Laboratory, which has a huge impact on the society, uses social survey to investigate the most current changes in the general public of Hong Kong toward the government and the society. Increasingly, the future research development of HKIAPS is closely evolved with the social and political changes in China as well as the rapid regional development in the Asian Pacific region.

The East Asian Institute (EAI), Singapore

The East Asian Institute (EAI) was established in 1997 to replace the Institute of East Asian Political Economy (IEAPE),

which was originally named East Asian Philosophies (IEAP) and established in 1983. In *Embracing the Past, Unfolding the Future*, the 10-year anniversary publication, the opening page spelt out EAI's mission was "…to promote both academic and policy-oriented research on East Asian development, particularly the political, economic and social development of contemporary China (including Hong Kong and Taiwan), and China's growing economic integration with the world and its political and security implications for the region" (EAI, 2007: 2). Its first director was Professor Gungwu Wang (the former vice-Chancellor of the University of Hong Kong from 1986 to 1995). Currently, its director is Yongnian Zheng who is the former Research Director of the China Policy Institute, at the University of Nottingham. As an autonomous organization, the EAI is governed by a management board of seven people: The Director; the Dean of Faculty of Arts and Social Sciences, the National University of Singapore (NUS); the Dean of the NUS Business school; the Dean of Faculty of Law; the Permanent Secretary of the Ministry of Defence; the Second Permanent Secretary of the Ministry of Foreign Affairs; and the Permanent Secretary of the Ministry of Trade and Industry (EAI, 2007: 8).

One of the key assets of the EAI as a research institute is that it is supported by its own endowment, making it more autonomous and flexible. It has its own library collections of Chinese and English materials on China and East Asia in all social sciences disciplines (the biggest in Southeast Asia). It owns more than 27 892 books including many Chinese yearbooks and statistics. It also subscribed to more than 359 Chinese and English journals and newspapers covering the PRC, Hong Kong, and Taiwan, as well as other microfilms. The EAI has more than 22 in-house researchers and regular visiting fellows and visiting senior fellows stopping by for 3 months to year-long research visits. Visiting fellows are provided with research space, computers, and funding. Among its research activities are weekly seminars, public lectures, and international conferences. In terms of publications, the EAI's *Background Briefs* are well received by Singapore's government as well as other leading research centers focused on policy-oriented analysis of China and East Asia, for instance the Fairbank Center at Harvard University. Other publications include working papers, occasional papers, and bulletins. It has two in-house journals: *China: An International Journal* is an academic outlet for peer reviewed articles, and *East Asian Policy* aims at policy-related papers. The institutes and its research staff regularly publish books and academic papers from their research results (http://www.eai.nus.edu.sg/Publications.htm (accessed 01.12.13.)).

The Institute of Southeast Asian Studies (ISEAS)

The Institute of Southeast Asian Studies (ISEAS) is one of the most celebrated research institutes focusing on Southeast Asia. With support from the National University of Singapore (NUS), the government of Singapore, other civic organizations, and most important of all, the Institute in Southeast Asia, ISEAS has played a key role in informing global academics and governments on every aspect of Southeast Asia, one of the most vibrant regions in the world. The ISEAS was established as an autonomous organization in 1968 by an act of Parliament. The *ISEAS Annual Report 2012–13* clearly points out that ISEAS is "A regional research centre dedicated to the study of sociopolitical, security, and economic trends and developments in Southeast Asia and its wider geostrategic and economic environment" (2013: 2).

Research in the ISEAS concentrates on three core areas: Regional Economic Studies (RES); Regional Social and Cultural Studies (RSCS), and Regional Strategic and Political Studies (RSPS). According to the aforementioned *Annual Report*, there were 22 (in-house and visiting) researchers working in RES, 69 in RSCS, and 37 in RSPC (2013: 10). There are also four units/centers under the umbrella of the ISEAS: The ASEAN Studies Centre (ASC); The Nalanda-Sriwijaya Centre (NSC); The Archaeology Unit (AU); and The Singapore APEC Study Centre. ISEAS is well-equipped with its own library system. As one of the most comprehensive libraries on Southeast Asian studies, ISEAS subscribes to 34 newspapers from different Southeast Asian countries, Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam. In total, it has more than 600 000 catalog records of books and research materials. The Southeast Asian Cultural Collection (SEACC) and Private Paper collections are both key components of the original archives of the library. The library is also undergoing a new development on digitization and automation (2013: 34–38). ISEAS is also featured with its flagship of the ISEAS Publishing House, which published social sciences, cultural, and geographical aspects of Southeast Asian studies. Four major journals under the ISEAS are the *ASEAN Economic Bulletin* (focusing on economic issues and policy analysis), *Contemporary Southeast Asia* (more than 30 years history of covering academic studies on Southeast Asian international relations), *SOJOURN* (covering cultural aspects on Southeast Asian studies), and the *Journal of Southeast Asian Economies* (academic articles on all the 10 Association of Southeast Asian Nations (ASEAN) economies).

The Institute of Social Science (ISS), Japan

The Institute of Social Science (ISS) is one of the oldest social science research institutes in Japan. The ISS was originally established in 1946 within the Tokyo Imperial University under Imperial Ordinance No. 349 (http://www.iss.u-tokyo.ac.jp/assesment/self/1-1.html (accessed 14.12.13.)). Nowadays, the ISS has four main areas of research: the Comparative Contemporary Law, the Comparative Contemporary Politics, the Comparative Contemporary Economics, and the Comparative Contemporary Societies. It houses 52 in-house and visiting researchers. Compared with other research institutes, ISS is different in the sense that it has a clear focus on Japan, by using social science to understand and to serve the Japanese society and economy.

The research projects and centers around the ISS are more centered on application and quantitative social sciences. For instance, the Center for Social Research and Data Archives provides comprehensive data on social science research in Japan. Other newly undertaken research projects aim at tackling the economic, social, and political problems that Japanese people are faced with today. The *Reconsidering Governance*

project, for example, studies corporate scandals and management efficiency. The *Gender Equality and Research Project* looks at work-life balance in modern-day Japan as a postindustrial economy. Other areas of research include the establishment of the Contemporary China Research Base in 2007 to study the economic growth and stability of China. The ISS hosts the flagship journal *Social Science Japan Journal (SSJJ)* and has a very strong library system for its research needs. Recently, the ISS announced that its future direction is to branch out to link up with other international, Asian and Pacific research institutes on various social sciences topics that can improve the study of Japanese society and economy. New collaboration research topics such as 'Contemporary Japanese Society' or 'The Twentieth-century Global System' are some new endeavors. Increasingly, the ISS is trying to establish more educational programs on graduate levels in order to train more social sciences practitioners for the country's need.

The East Asia Institute (EAI), Korea

The East Asia Institute (EAI) of Korea is a relatively new social sciences academy which was established in May 2002. According to the *2012 Think Tanks and Civil Societies Program*, the EAI is ranked 17th of the 45 think tanks in Northeast Asia (2013: 57). According to its own description, "… the EAI was established as an independent think-tank dedicated to developing ideas and formulating policy recommendations on the main challenges facing the region" (http://www.eai.or.kr/type/aboutGreeting.asp (accessed 10.11.13.)). The EAI has a more refined focus mainly on security and regional affairs. Second, it covers Northeast Asia only. By focusing on these two clearly defined areas, the EAI has more of a research focus, while activities and research dissemination can be more carefully tailored with specific impact from the user's points of view.

The scope of the research of the EAI can further be divided into four programs: (1) Peace and Security, (2) Public Opinion, (3) Democracy and Governance, and (4) Education and Human Development. The program of Peace and Security involves different seminars, conferences, and public lectures on regional security issues centring on China, Japan, and Korea's high politics. The Public Opinion Program mainly uses survey methods to come up with regular reports and assessment on global viewpoints, Korean politics, and Korean elections. The Program of Democracy and Governance is used to study Korea's democratic development with a focus on Korea's role in the globalized world. Among the others, the Education and Human Development program aims at broadening the participant base and offering various funding resources to scholars carrying out research with the EAI and its partner institutions such as the EAI Fellows Program, the EAI Global Academy, or the EAI Internship Program. The flagship EAI Fellows Program offers financial support (3 weeks) for established and budding scholars to disseminate original research in two of the partner institutes in Northeast Asia. Increasingly, it has become one of the most prestigious fellowships in East Asian studies. In terms of publications, the *Journal of East Asian Studies* (edited by Stephan Haggard from the University of California, San Diego) publishes peer-reviewed articles (sometimes from the EAI Fellows Program) on East Asian affairs and international relations. The *EAI Working Papers* provides a more policy-oriented East Asian focus on a more substantial scale. The *EAI Issue Briefings* offers a short yet precise assessment of East Asian politics and international relations.

See also: Social Science Infrastructure: Asia; Social Science Infrastructure: East Asia and the Pacific (Research and Teaching); Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Infrastructure: North America (Social Science Academies and Related Organizations).

Bibliography

Brooks, Harvey, Randazzese, Lucien P., 1998. University-Industry relations: the next four years and beyond. In: Branscomb, Lewis M., Keller, James H. (Eds.), Investing in Innovation: Creating a Research and Innovation Policy that Works. Mass: The MIT Press, pp. 361–399.

East Asian Institute (EAI), 2007. Embracing the Past, Unfolding the Future: Commemorating EAI's 10th Anniversary. EAI, Singapore.

Hong Kong Institute of Asia-Pacific Studies (HKIAPS), 2010. HKIAPS at 20. HKIAPS/Chinese University of Hong Kong, Hong Kong.

Hong Kong Institute of Asia-Pacific Studies (HKIAPS), 2012. Annual Report 2011–12. HKIAPS/Chinese University of Hong Kong, Hong Kong.

ISEAS Annual Report 2012–13, 2013. ISEAS, Singapore.

Jorgenson, Dale W., 2001. Information technology and the U.S. economy. The American Economic Review 91 (1), 1–32.

Kim, Linsu, Nelson, Richard R. (Eds.), 2000. Technology, Learning, and Innovation: Experiences of Newly Industrializing Economies. Cambridge University Press, Cambridge.

Ma, Wanhua, 2007. The flagship university and China's economic Reform. In: Altbach, Philip G., Balan, Jorge (Eds.), World Class Worldwide: Transforming Research Universities in Asia and Latin America. The Johns Hopkins University Press, Baltimore, pp. 31–53.

Mahbubani, Kishore, 2008. The New Asian Hemisphere: The Irresistible Shift of Global Power to the East. Public Affairs, New York.

2012 Think Tanks and Civil Societies Program, 2013. International Relations Program, University of Pennsylvania, Philadelphia, PA.

Vogel, Ezra F., 2005. Foreword: the first forty years of the universities service centre for China studies. The China Journal 53, 1–7. Celebrating the Universities Service Centre for China Studies (January).

Relevant Websites

http://www.sinica.edu.tw/main_e.shtml – Academia Sinica.
http://www.assa.edu.au/about/academy – Academy of the Social Sciences in Australia.
http://bic.cass.cn/english/infoShow/Arcitle_Show_Cass.asp?BigClassID=1&Title=CASS – Chinese Academy of Social Sciences.
http://www.eai.or.kr/english/ – East Asia Institute (EAI).
http://www.eai.nus.edu.sg/ – East Asian Institute (EAI).
http://www.eastwestcenter.org/ – East-West Center, Hawaii.
http://www.cuhk.edu.hk/hkiaps/index.html – Hong Kong Institute of Asia-Pacific Studies (HKIAPS).
http://www.iss.u-tokyo.ac.jp/index.html – Institute of Social Science (ISS).
http://www.iseas.edu.sg/ – Institute of Southeast Asian Studies (ISEAS).
http://www.usc.cuhk.edu.hk/Eng/Default.aspx – Universities Service Centre for China Studies (USCCS).

Social Science Infrastructure: European Union (Resourcing and Funding)

Thomas König and Helga Nowotny, Vienna Science and Technology Fund, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Over the past 15 years, European research funding has become increasingly important for social and behavioral research. This article provides an overview of the most important developments, funding opportunities, and ongoing debates related to the European Research Area.

Brief History of European Research Funding

The idea of research funding at European level has been discussed at least since the inception of the European Community. Several programs and instruments can even be traced back to before the 1970s. However, it was only around 1984 that the first Framework Programme for Research and Technological Development (FP) was set up and operationalized by the European Commission, with a moderate budget of 3.5 billion euros over a period of 4 years. The main political objective was to "strengthen the scientific and technological basis of European industry and encourage it to become more competitive at international level" (Andrée, 2009: p. 18). The consecutive Framework Programmes have grown considerably in budgetary size and scope to include the social sciences and even the humanities. The eighth multiannual Framework Programme, Horizon 2020 (from 2014 onward), has an estimated budget of almost 80 billion euros over a period of 7 years. Its ambition spans three 'pillars' of research funding that, together, are intended to boost innovation. The excellence pillar funds fundamental research through the European Research Council (ERC), the mobility program Marie (Skłodowska-) Curie Actions (MCA), and research infrastructures. The second pillar offers funding for industry- and small- and medium-sized enterprise-related participation while the third is devoted to seven societal challenges (for details, see below and Table 1). Though the Framework Programmes are not the only research and innovation funding instruments at European level, due to their broad size and scope, they dwarf all other programs. This finds its expression in the way other, formerly independent programs deal with the superior position of the European Commission's funding clout: They either succumbed, as in the case of the European Cooperation in Science and Technology (COST), which formally retains its independent structure, but fully depends on contributions through the Framework Programmes. Or they dropped out, as in the case of the

Table 1 Pillars and programs of FP8, Horizon 2020 (excluding Euratom)

Pillars	*Programs*	*Approximate budget (in billion euros)*	*SSH share*
Pillar 1: Excellent science	European Research Council[a]	13.095	Approximately 17%
	Future and Emerging Technologies	2.696	/
	Marie Skłodowska-Curie Actions[a]	6.162	Approximately 11%
	European Research Infrastructures (ESFRI and so on)[a]	2.488	Unresolved
Pillar 2: Industrial leadership	Leadership in Enabling and Industrial Technologies	13.557	/
	Access to Risk Finance	2.842	/
	Innovation in SMEs (small- and medium-sized enterprises)	0.616	/
Pillar 3: Societal challenges	Health, Demographic Change and Wellbeing[a]	7.472	Unresolved
	Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy[a]	3.851	Unresolved
	Secure, Clean and Efficient Energy[a]	5.931	Unresolved
	Smart, Green and Integrated Transport[a]	6.339	Unresolved
	Climate Action, Environment, Resource Efficiency and Raw Materials[a]	3.081	Unresolved
	Europe in a changing world – inclusive, innovative, and reflective societies[a]	1.309	100%
	Secure societies – protecting freedom and security of Europe and its citizens[a]	1.695	Unresolved
Additional programs	Science with and for Society[a]	0.462	Unresolved
	Spreading Excellence and Widening Participation[a]		Unresolved
	European Institute for Technology		/
	Non-nuclear Direct Actions of the Joint Research Centre		/

FP, Framework Programme for Research and Technological Development; ESFRI, European Strategy Forum on Research Infrastructures; SSH, Socioeconomic Sciences and Humanities.
[a]In principle relevant for social and behavioral sciences funding.
Sources: http://ec.europa.eu/programmes/horizon2020/h2020-sections; http://ec.europa.eu/research/horizon2020/pdf/press/fact_sheet_on_horizon2020_budget.pdf.

European Science Foundation (ESF), an association of national research funders and organizations that perform research, that is closing down its own funding capacities (e.g., Eurocores) after having been transformed into a purely coordinating body, Science Europe.

The Changing Role of Social Sciences in EU Research Funding

The European integration project is mostly driven by the establishment of a common European market. Research funding was thus initially perceived as a way to support the European economic base and its competitiveness by concentrating on science, engineering, and technology. However, immediately after the Maastricht Treaty, the fourth Framework Programme (1994–98) introduced the Targeted Socio-economic Research Programme (TSER). Under the fifth Framework Programme, it was renamed Key Action "Improving the Socio-economic Knowledge Base" (1998–2002) (Kastrinos, 2010: p. 300). According to the final report on TSER, the program was intended "to build up both the knowledge base and research networks for high quality, policy relevant, comparative European socio-economic research at both national and Community level, in order to support sustainable social and economic development" (European Commission, 1998: p. 5). Having funded 365 projects, and invested 267 million euros over the course of 10 years, the fourth and fifth Framework Programmes gave the social sciences a considerable share of the research budget for the first time (although, approximately only 1% of the budget for the two Framework Programmes went to social sciences projects) (Schögler, 2013: p. 59). The sixth Framework Programme widened its scope even further by bringing the social sciences and the humanities together for the first time under the funding stream "Citizens and Governance in a Knowledge-based Society" (Smith, 2003). The funding stream dedicated to research in the social sciences and humanities (SH) under FP7 was consequently called 'Socio-economic Sciences and the Humanities.'

The numbers depicted in Table 2 do not present a straightforward statistical picture of funding dedicated to social and behavioral research; the full extent of social and behavioral research within the FPs remains unclear. This is partly due to the fact that social scientists were invited to take part in and monitor scientific research (Luukkonen, 1998) not classified as 'social research' in early European Commission research funding schemes. In addition, as the FPs grew in size and scope, and as more and more funding instruments were developed, social researchers also participated in research projects in other parts of the FPs, either as collaborators for a specific research task, or (again) as aids in the dissemination of research results, and formulation of policy advice.

From FP6 onward, the situation became even more blurred, as funding for social research became more and more available outside the funding streams mentioned in Table 2. For example, while the mainstream dedicated to SH in FP6 was in priority 7 ('Citizens and Governance in a Knowledge-based Society'), some projects in priority 8 ('Scientific Support to Policies') were also flagged as social research. This trend continued under FP7, and with Horizon 2020, the diffusion or 'integration' of social scientific and humanities research into all topics has even become a political prerogative. Now the European Commission deliberately seeks to abolish the 'silos' of socioeconomic sciences and humanities (SSH) research funding inside its Framework Programme. Instead, SSH will be integrated into all of the societal challenges as 'cross-cutting priorities,' while only one of the seven will be devoted to 'Europe in a changing world: inclusive, innovative and reflective societies,' and hence primarily dedicated to funding social and behavioral research.

In addition to the dedicated funding for research, the MCA provided 4.5 billion euros in mobility grants under FP7. Out of eight panels, two were set up specifically to review proposals from the social sciences (one being called 'social sciences and humanities,' the other 'economic sciences'). Along several funding schemes for individuals and training networks, those two panels allocated more than 500 million euros to social scientists and humanists. Furthermore, within the legal framework of FP7, the ERC was founded. This funding body,

Table 2 European Commission funding in social research (Framework Programmes 4–8)[a]

Years	*FP*	*Total budget (in million euros)*	*SSH budget (in million euros)*	*%*	*Projects*	*Source*	*Name*
1994–98	FP4	13 120.8	112	0.85	162	ftp://ftp.cordis.europa.eu/pub/improving/docs/ser_tser_projects_all_calls.pdf	Targeted Socio-economic Research Program (TSER)
1999–2002	FP5	14 870.7	155	1.04	203	ftp://ftp.cordis.europa.eu/pub/citizens/docs/eur20635.pdf	Key action "Improving the Socio-economic Knowledge Base"
2003–06	FP6	19 256	266	1.38	169	ftp://ftp.cordis.europa.eu/pub/citizens/docs/synopses-all-calls-2006.pdf	Priority 7 "Citizens and Governance in a Knowledge-based Society"
2007–13	FP7	55 806.1	610	1.09	NA	ftp://ftp.cordis.europa.eu/pub/fp7/ssh/docs/synopsis_ssh_projects_fp7_low_definition_with_web_links.pdf	Theme 8 "Socio-economic Sciences and the Humanities"
2014–20	FP8[b]	80 000[c]	1253	1.57	NA	NA	Challenge 6 "Europe in a Changing World"

FP, Framework Programme for Research and Technological Development; SSH, Socioeconomic Sciences and Humanities.
[a]Compare with (Schögler, R., 2013. European Union Research Funding. Priority Setting in the Social Sciences and Humanities. University Graz, Graz, p. 67).
[b]Horizon 2020.
[c]Estimation in current prices (2013).

Table 3 European Research Council funding in social sciences and humanities domain 2007–13

Year	*Projects funded in SH domain*	*Budget of SH domain (in million euros)*	*Percentage of total annual budget*
2007	57	49 241	14.74
2008	56	95 392	17.34
2009	97	134 374	15.92
2010	126	186 744	15.95
2011	146	229 307	16.89
2012	164	256 611	17.00
2013	168	294 934	17.29
Sum	814	1 246 605	16.68

SH, Social Sciences and Humanities.
Source: http://erc.europa.eu/statistics-0.

strategically independent but legally bound to FP7, implied a radical shift of European research policy insofar as it followed an approach of scientific excellence as the only evaluation criterion without thematic priorities. The ERC dedicated approximately 17% of its annual budget to individual projects in the SH. As can be seen from Table 3, the ERC has funded 814 projects (2013) and dedicated more than 1.2 billion euros to the SH. Again, these numbers can be misleading since the ERC also funds interdisciplinary projects (König, forthcoming; Nowotny, 2009).

Different Funders, Different Programs, and Different Aims

Since acquiring European funding is mostly a question of 'EU-added value,' namely the funding of research whose value is higher than if were funded on a national level, the initial task of the Framework Programmes was to ask project applicants for collaborative research with project partners across Europe, thereby fostering networking activities (Pohoryles, 2002; Nedeva and Stampfer, 2012). In the meantime, however, the range of activities and programs has increased greatly.

Following their different intentions and aims, we can roughly distinguish four types of funding programs at European level: horizontal programs that support network activities of researchers with similar interests; programs funding research projects along given priorities and policy requests (top–down mode); programs that are open to research projects from all fields and disciplines, usually with individual principal investigators as applicants (bottom–up mode); and initiatives to build up and operate European-wide research infrastructures. In all the four types of programs, there is also funding for social and behavioral research foreseen, although greatly varying in size and visibility.

Beginning with programs supporting network activities, COST (see above) provides funding for European-wide workshops and trainings on specific topics. The oldest funding scheme at European level, COST is now also funded by the Framework Programme. In the same vein, the ESF ran several funding schemes for international workshops, conferences, and so on until recently. The ESF still operates Scientific Review Groups (previously Standing Committees), one of which is dedicated to the social sciences. The Review Group organizes research networking programs, and formulates recommendations related to research policy on social sciences in Europe. The latter function is likely to be overtaken by the newly founded organization, Science Europe. ESF's activities were potentially powerful through linking its policy recommendations with research meetings on related topics and financing preparatory meetings, which later developed into research schemes. Exactly such links may be missing in the new organization, which so far is conceptualized only as a policy and coordinating body.

The European Commission's Framework Programmes primarily fund research projects, not networking activities around research interests (although there are additional schemes for that, too). Under FP7, most grants were allocated under the two funding scheme formats called 'collaborative projects' and 'networks of excellence' (under the previous FPs, the formats had different names, but relied basically on the same concepts). This is also true for the program dedicated to the SH. Up until and including FP7, funding schemes usually required a certain rate of self-financing. Since Horizon 2020 (for details, see Table 1) sets out to simplify the entire European Commission funding administration, only one funding scheme exclusively dedicated to research funding has been set up. Called 'research and innovation actions,' this scheme does not require partial self-financing anymore. However, like its predecessors, it requires at least three different project partners from different parts of Europe to apply within one project proposal. As of April 2014, only the first calls have been published under the new Framework Programme; however, it seems as if most of the funding foreseen for social and behavioral research projects will be in the form of 'research and innovation actions,' too.

Within Horizon 2020, but distinct from the rest of the large project consortia, the mobility program, now called Marie Skłodowska-Curie Actions (MSCA), and the ERC will continue to provide substantial grants to researchers in the social and behavioral sciences. MSCA, while reducing its overall number of funding schemes, will keep its panel structure. Its overall budget will grow to 6.16 billion euros, provide mobility grants for students, PhDs, and postdocs – either to individual researchers or through funding training networks. The ERC will also see an increase in its budget to more than 13 billion euros by 2020. With three main funding schemes – the Starting Grant for young researchers up to 7 years after the PhD, the Consolidator Grant for those establishing their scientific career, and the Advanced Grant for established senior scholars – the ERC will continue to fund excellent researchers. Should the ERC's SH domain retain its budget share of 17%, more than 2.2 billion euros will be spent on basic research projects in the SH.

An important step toward European integration of research infrastructures has been made over the past years with the European Strategy Forum on Research Infrastructures that identifies and prioritizes important research infrastructures across the continent. Of particular interest to the social sciences are the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and the Council of European Social Science Data Archives (CESSDA). While the three initiatives are of crucial importance to social research in Europe, only SHARE has successfully transitioned into a European Research Infrastructure Consortium (ERIC), a new

legal format of an intergovernmental research entity. Becoming an ERIC guarantees continued funding from the participating members and, at the same time, allows members to raise additional research funding from outside sources. The future of the ESS and CESSDA remains unclear due to a lack of long-term commitment from both national governments and the European Commission.

The Impact of European Research Funding on the Social and Behavioral Sciences: Trends and Critiques

Over the past 20 years, European-level funding increased in importance for the social and behavioral sciences, leaving its mark on the structure of research collaboration, the way the sciences perceive themselves, and how scholars in those sciences feel appreciated. The EU has supported research since the 1970s (Banchoff, 2002: pp. 7–8). Even before the first Framework Programme, programs like COST and concerted enterprises like ESF populated the European research landscape. It is only with the skyrocketing budgets of FP6 and FP7 that the other European programs were dwarfed and some of them even incorporated (see above). The Lisbon Treaty then formally placed science policy on the EU agenda (TFEU, 2009: article 179).

On analyzing European Policy on research funding dedicated to the SH, three observations can be made (Kastrinos, 2010; Schögler, 2013). First, despite the European emphasis on priority-setting and thematic orientations, European research funding is moving more and more toward a 'diffusion-oriented model,' which emphasizes capacity building over a distinct mission. This is also the case with Horizon 2020, whose third pillar – societal challenges – targets the social sciences to build knowledge capacity on broadly defined topics. Second, the EU research programs have emerged as points of reference for the member states, both in terms of those topics (the specific challenges in the third pillar of Horizon 2020) and in the orientation toward diffusion instead of mission. It is already quite clear that the trend will increase under Horizon 2020, if only because this program is now so large that it cannot be ignored by social scientists and policy makers. Already, now we see funding schemes in European member states emerge that are substituting, or complementary to, the individual programs in the third pillar of Horizon 2020. Finally, despite its limited size within the overall FP budget and despite several national funding schemes targeting research in the SH, "in comparative terms this has been the largest targeted programme in Europe" for the social sciences and humanities (Kastrinos, 2010: pp. 301–304).

The European Commission research funding has been long criticized for its administrative burdens, as well as the unintended consequences of requiring research collaboration across national borders. In recent years, social scientists have also voiced misgivings about the limited size of available funding. This is somewhat ironical since, as previously discussed, funding opportunities for social and behavioral research have increased considerably at the European level. One reason for this concern is that funding dedicated to social and behavioral research is still quite small, in terms of absolute size within the overall FP budget (Table 2). At the same time, however, social and behavioral researchers in Europe have started to take European funding increasingly into account, not only because national funds are drying up in many member states, but also because it is recognized that national science policies are increasingly coordinated at, and influenced by the European level. Another reason lies in the fact that in a program that is largely dominated by top–down instruments, research topics are decided in political negotiations. In order to gain a footing, the social sciences would need a proper representational body at the European level. Thus, voicing public complaints has become the preferred mode of arguing on the part of the social science community, instead of proactive bargaining. An attempt to overcome this pattern was made at the conference 'Horizons for Social Sciences and Humanities,' where policy makers, among them high-ranking members of the European Commission's Directorate General for Research and Innovation, and representatives from the SH met and discussed the future of social science and humanities research (Mayer et al., 2013; Schögler, 2013: pp. 176–179). At the conference, the Vilnius Declaration was adopted and brought to the attention of European politicians, underscoring the 'values and benefits of integrating Social Sciences and Humanities' into Horizon 2020 (Hackett, 2014; Mayer et al., 2013: p. 25).

See also: Social Science Infrastructure: European Union (Research and Teaching); Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Infrastructure: North America (Resourcing and Funding).

Bibliography

Andrée, D., 2009. Priority-Setting in the European Research Framework Programmes (No. 2009-01925), Vinnova Analysis VA 2009:17. VINNOVA – Swedish Governmental Agency for Innovation Systems.

Banchoff, T., 2002. Institutions, inertia and European union research policy. Journal of Common Market Studies 40, 1–21. http://dx.doi.org/10.1111/1468-5965.00341.

European Commission, 1998. Targeted Socio-Economic Research Programme. Synopses of TSER Projects Funded as a Result of the Three Calls for Proposals 1995/1996–1997/1998. European Commission, Brussels.

Hackett, E.J., 2014. The vilnius declaration. Science, Technology, & Human Values 39, 3–5. http://dx.doi.org/10.1177/0162243913515373.

Kastrinos, N., 2010. Policies for co-ordination in the European research area: a view from the social sciences and humanities. Science Public Policy 37, 297–310. http://dx.doi.org/10.3152/030234210X496646.

König, T. Peer review in the social sciences and humanities at European level: the experiences of the European Research Council. In: Daniel, H-D. (Ed.), Research Quality in the Humanities, forthcoming.

Luukkonen, T., 1998. The difficulties in assessing the impact of EU framework programmes. Research Policy 27, 599–610.

Mayer, K., König, T., Nowotny, H., 2013. Horizons for Social Sciences and Humanities (Conference Report). Mykolas Romeris University, Vilnius.

Nedeva, M., Stampfer, M., 2012. From "Science in Europe" to "European science". Science 336, 982–983. http://dx.doi.org/10.1126/science.1216878.

Nowotny, H., 2009. Frontier Research in the Social Sciences and Humanities: What Does It Mean, What Can It Mean? Humanities and Social Sciences: Intersections and Joined Paradigms, ISSC-ICPHS workshop.

Pohoryles, R.J., 2002. The making of the European research area – a view from research networks. Innovation: The European Journal of Social Science 15, 325–340. http://dx.doi.org/10.1080/1351161022000042561.

Schögler, R., 2013. European Union Research Funding. Priority Setting in the Social Sciences and Humanities. University Graz, Graz.

Smith, J., 2003. Implementation of the European Research Area in the Social and Human Sciences (Discussion Paper No. EUR 20601), RTD Info. European Commission. Treaty on the Functioning of the European Union, 2009.

TFEU, 2009. Treaty on the functioning of the European Union. Official Journal of the European Union (2012/C 326/01).

Relevant Websites

http://cordis.europa.eu/improving/socio-economic/selected.htm – CORDIS Archive: Improving Human Potential & Socioeconomic research.

http://www.cost.eu/ – COST: European Cooperation in Science and Technology.

http://www.cessda.net/ – CESSDA AS: Consortium of European Social Science Data Archives.

http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges – HORIZON 2020 The EU Programme Programme for Research and Innovation: Societal Challenges.

http://www.esf.org/ – European Science Foundation.

http://www.europeansocialsurvey.org – European Social Survey.

http://ec.europa.eu/programmes/horizon2020/en/h2020-section/marie-sklodowska-curie-actions – HORIZON 2020 The EU Programme Programme for Research and Innovation: Marie Skłodowska-Curie actions.

http://erc.europa.eu/ – European Research Council.

http://horizons.mruni.eu – Horizons for Social Sciences and Humanities.

http://www.scienceeurope.org/ – Science Europe.

http://www.share-project.org/ – SHARE - Survey of Health, Ageing and Retirement in Europe.

Social Science Infrastructure: European Union (Social Science Academies and Related Organizations)

Thibaud Boncourt, European University Institute, Fiesole, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article aims at providing empirical and comparative information on the history of European social science organizations. It looks at two main questions: the factors that contributed to the creation of these organizations, and the differences between them in terms of their membership, organizational structure, and scientific activities. It shows that scientific, academic, and political rivalries played an important part in the creation of organizations and had long-lasting impacts on their development.

Introduction

The creation of European social science organizations is a relatively late phenomenon in the history of sciences in general and of social sciences in particular. When they were founded in the second half of the twentieth century, a wide array of scientific organizations already existed. National scientific academies, such as the Royal Society in England or the *Académie des sciences* in France, had indeed been created as early as the seventeenth century. International forms of scientific gatherings followed closely thereafter. They first took the shape of a few punctual congresses, mostly around the discipline of physics. After 1850, the number of such congresses began to grow at an exponential pace (**Figure 1**) as these gatherings became more diversified at the disciplinary level, with both natural (e.g., physicists, biologists) and social scientists (e.g., demographers, geographers) organizing such events. Several disciplines, such as statistics, set out to hold congresses on a regular basis (Brian, 1989), thereby paving the way for the creation of proper international scholarly associations.

Several international organizations were then created between the mid-nineteenth and late-twentieth century and by 1994, "there were approximately 350 active international science associations" (Schofer, 1999). The late-nineteenth and early twentieth centuries were periods of slow growth, as they saw the creation of about 30% of the associations. In the second half of the twentieth century, international science associations grew in number at an accelerating rate; over 70% of them were founded during that period. As **Table 1** and **Table 2** show, most international social science associations were created during that phase. More specifically, they were founded in the late 1940s or early 1950s, under the auspices of a UNESCO program that aimed at stimulating the expansion of social sciences worldwide (Platt, 1998; Boncourt, 2009). By contrast, it was not until the second half of the 1960s that social scientists set out to create European-wide professional organizations. There were important fluctuations across disciplines: while the European Association for Experimental Social Psychology was founded in 1966, the first European sociology organization (the European Consortium for Sociological Research) was created in 1991. The number of European organizations in a given discipline also varied across disciplines as some of them only numbered one (economics, anthropology, psychology) and others up to three (political science).

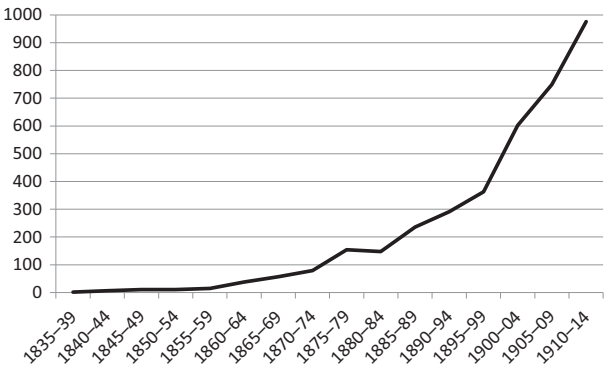

Figure 1 Number of international congresses held between 1835 and 1914. Source: Brian, E., 1989. Y a-t-il un objet Congrès? Le cas du Congrès international de statistique (1853–1876). Cahiers Georges Sorel 7, 9–22.

This article aims at providing empirical information on the creation and development of these organizations. It follows a comparative logic to identify common points and differences between organizations and disciplines. It looks at two main questions: the factors that contributed to the creation of European social science organizations, and the differences between these organizations in terms of their membership, organizational structure, and scientific activities. It uses qualitative and quantitative data gathered in the course of a doctoral work on the history of political science (Boncourt, 2011), as well as in the framework of the INTERCO-SSH research project. (The INTERCO-SSH ('International Cooperation in the Social Sciences and Humanities: Comparative Perspectives and Future Possibilities') is coordinated by Gisèle Sapiro and funded by the European Union Seventh Framework Program (FP7/2007–13) under Grant Agreement no 319974 (INTERCO-SSH).) It also draws on secondary sources and most notably from accounts of the history of these organizations written by their founders and past officers (e.g., Kuper, 2004; Moscovici and Markova, 2006).

The Creation of European Social Science Organizations

As **Table 3** shows, a first look at the main characteristics of European social science organizations enables us to outline

 http://dx.doi.org/10.1016/B978-0-08-097086-8.41056-1

Table 1 Creation of main international and European social science organizations

	Political science	*Sociology*	*Economics*	*Anthropology*	*Psychology*
1930–39			ES (30)		
1940–49					
	IPSA (49)	ISA (49)		IUAES (48)	
1950–59			IEA (50)		IUPsyS (50)
1960–69					
					EAESP (66)
1970–79	ECPR (70)				
1980–89			EEA (84)		
				EASA (89)	
1990–99		ECSR (91)			
	EpsNet (96)	ESA (92)			
2000–09					
2010–...	EPSA (10)		WEA (11)		

Legend	
ECPR (70)	European organisation
IPSA (49)	International organisation

Table 2 Names and acronyms of international and European organizations

Acronym	*Name*
EASA	European Association of Social Anthropologists
EAESP	European Association of Experimental Social Psychology
ECPR	European Consortium for Political Research
ECSR	European Consortium for Sociological Research
EEA	European Economic Association
EPSA	European Political Science Association
EpsNet	European Political Science Network
ES	Econometric Society
ESA	European Sociological Association
IIS	International Institute of Sociology
IPSA	International Political Science Association
ISA	International Sociological Association
IUAES	International Union of Anthropological and Ethnological Sciences
IUPsyS	International Union of Psychological Science
WEA	World Economics Association

several differences between them. First, they were created at different times, as the first of them (EAESP) was founded more than 40 years before the last (EPSA). Second, they differ in terms of the type of funding they benefited from: some associations received public support from national governments, international governmental organizations (such as the European Union) or universities, while others relied on private subsidies coming most often from philanthropic foundations (Ford, Royaumont, Volkswagen, Wenner-Grenn, and European Cultural Foundations). Third, organizations do not have the same geographical origins, as their first president could be based in the United Kingdom, France, Belgium, or Sweden. Finally, these associations have different intellectual orientations as they were conceived as a tool to either facilitate or impede the import of 'American' paradigms and methods into Europe. Following Benoit de l'Estoile's comparative work (L'Estoile, 2008), two types of organizations may indeed be identified: those of a hegemonic nature on the one hand, and those of a pluralistic kind on the other hand. Hegemonic internationalization is defined as "a result of the attraction by the more powerful centre on its own terms. The international idiom, linguistically and intellectually, is the one of the hegemonic centre." By contrast, pluralistic internationalization is characterized as an "endeavor with a strong utopian component: it advocates a meeting ground where in principle all can meet on their own terms. In essence, it strives at creating a kind of non-hegemonic internationalization" (L'Estoile, 2008: 120–121).

A closer look at the history of European social science organizations shows that their founders pursued different kinds of scientific, academic, and political strategies.

1. At the scientific level, most of them aimed at *changing the paradigmatic structure of their discipline* on the European stage by either promoting or resisting 'American' approaches. In

Table 3 Characteristics of main European social science organizations

Name	*Discipline*	*Date*	*Funding*	*First president*	*Orientation*
ECPR	Political science	1970	• Ford Foundation • University of Essex	United Kingdom	Hegemonic
EpsNet	Political science	1996	• European Union • Sciences Po	France	Pluralistic
EPSA	Political science	2010	• Founders' own funds	United Kingdom	Hegemonic
ECSR	Sociology	1991	• European Union	Sweden	Hegemonic
ESA	Sociology	1992	• Austrian Government	United Kingdom	Pluralistic
EEA	Economy	1984	• Central banks • Belgian Ministry of Economy • European Cultural Foundation	Belgium	Hegemonic
EAESP	Psychology	1966	• Royaumont, Volkswagen and Ford Foundations • Ecole Pratique des Hautes Etudes	France	Hegemonic
EASA	Anthropology	1989	• Wenner-Gren Foundation	United Kingdom	Pluralistic

the case of hegemonic organizations, founders strived to follow the American example by importing what they saw as its main strength, that is, its behavioral, statistical, and positivist factions. They criticized European approaches to social matters and claimed that it was necessary to 'catch up' on their American counterparts by importing 'modern' approaches into Europe and subverting dominant paradigmatic power relations in the continent. This was the case in political science, as founders of ECPR claimed that European political scientists did not know enough about 'sophisticated' approaches and methods to be able to interact, let alone compete with their American colleagues (Daalder, 1997; Budge, 2006; Boncourt, 2013). Forty years later, the founders of EPSA followed a similar logic when they felt that ECPR had become too open to sociological and historical approaches to politics, and that a new organization had to be created to showcase political science studies of a deductive and statistical kind. In case of sociology, founders of ECSR were driven by the urge to resist the rise of 'Marxist and phenomenological perspectives' in Europe that, in their view, impeded the development of a dialogue 'of a constructive and productive kind' between European and American sociologists. (Quotes are drawn from an interview with John Goldthorpe, a founding member of ECSR.) Similar logics were at work in the field of psychology, as EAESP was founded in close collaboration with American psychologists with a view of promoting experimental studies and "[improving] or [facilitating] communication between Americans and Europeans" (Moscovici and Markova, 2006). In the case of the European Economics Association, founding members all had close connections with American economists and econometrical techniques, and advocated the strengthening of the European discipline in order to compete with what was being done in the United States.

The creation of pluralistic organizations was guided by comparable mechanisms, as parts of American social sciences were also taken again as points of references. However, this time, American intellectual fashions were to be kept at bay in order for European scientific traditions and specific approaches to be further developed and to offer an alternative to hegemonic paradigms. Thus, the founding members of the European Association of Social Anthropology strongly criticized 'essentialist, idealist, and relativist' approaches that they saw as typical of American anthropology, and sought to "create an alternative space for theoretical debates, an institutional counterweight to American institutions" (Kuper, 2004). Similar dynamics led to the creation of the European Political Science Network whose founders aimed at promoting an alternative to the kind of Anglo-Saxon political science that they saw as typical of ECPR. Founders of the European Sociological Association pursued a similar objective, though in a less-confrontational way: without openly criticizing the nascent ECSR, they sought to create a forum where all kinds of European sociological approaches could find an echo. Thus, while the creation of European social science associations was in part driven by an ambition to provide European social scientists with specific organizational support, it was also linked to an aspiration to influence and even transform the paradigmatic structure of European social sciences, in an either 'hegemonic' or 'pluralistic' way.

2. The founding of European social science organizations was also meant to *play a role in academic competitions* by allowing certain universities, and even certain individuals, to obtain international resources and thus gain additional weight on the national and international stage. This is tangible in several cases. The creation of ECPR has to be understood in relation to the University of Essex's ambition to rival top social science faculties in the United Kingdom (such as the London School of Economics and Political Science, and the Universities of Oxford and Cambridge), in spite of its relatively recent creation (Boncourt, 2013). Similarly, the foundation of EAESP was a way for the Ecole Pratique des Hautes Etudes to establish its reputation in the field of French social psychology, which was at the time dominated by Sorbonne University. It also helped 'University lightweights' of the time to become more recognized in their discipline and academic fields (Moscovici and Markova, 2006). In the case of economics, the creation of EEA was part of a bid to enhance the international reputation of the University of Louvain-La-Neuve and its Center for Operations Research

and Econometrics (Drèze, 2008). In political science, Sciences Po pursued a similar agenda when it pushed for the creation of the European Political Science Network (EpsNet) in 1996, as a rival organization to the well-established ECPR (Bull, 2007; Boncourt, 2011). The founding of European social science organizations thus played a role both on the scientific and the academic stages.

3. Founding members of European social science organizations thus pursued both scientific and academic strategies. However, political agendas also played a role in their undertaking. As they sought to gather external support and funding for their projects, founders were indeed led to collaborate with political actors such as national governments and, more frequently, philanthropic foundations. In the context of the Cold War, these foundations indeed sought to influence European cultural developments, with a view of contributing to the strengthening of transatlantic ties and to the containment of Soviet influence. Studies of the 'cultural' or 'intellectual Cold War' thus show the extent to which these actors played a role in the development of literature, arts, sports, and sciences (e.g., Berghahn, 2002; Sirinelli and Soutou, 2008; Popa, 2011; Solovey and Cravens, 2012). Social sciences were a key part of this project, as foundations invested in the development of European economics, political science, social psychology, international relations and public administration, among other areas. By funding these disciplines, they sought to support the development of social knowledge in order to improve human welfare in the long run. Investing in Europe, rather than in the United States, was a way to help European social scientists catch up on their American counterparts, to deepen transatlantic connections and to strengthen European democracies (Carver, 1986; Saunier, 2003; Moscovici and Markova, 2006; Guilhot, 2011; Tournès, 2011). The Ford Foundation was the main actor in this process in the 1960s. At the time, the Foundation, which had been created in the 1930s with the ambition of enforcing peace and democracy around the world, encouraged the development of European social sciences, which were seen as a good vector for democracy (Magat, 1979; Gemelli, 1998).

Such logics were mostly at work in the case of hegemonic organizations, and philanthropic support took different shapes across disciplines and associations. In some cases, the Ford Foundation and other philanthropic organizations contributed to the structuring of transatlantic and transnational networks, which would later form the basis of European social science organizations. Several European social scientists thus held research fellowships in prestigious American universities such as Yale, Harvard, Berkeley, Princeton, or the Massachusetts Institute of Technology and built on their American connections to create European associations some years later. Most of the founders of ECPR, EEA and EAESP had thus benefited from such fellowships (Daalder, 1997; Moscovici and Markova, 2006). In other instances, the Ford Foundation provided more direct support by funding the new association itself, the University, or the research center that was hosting it. In the 1960s, its policy was indeed to send envoys on tours of some of the European universities and encourage the launching of transnational social science endeavors. As they sought to strengthen transatlantic ties, envoys mostly visited institutions with a track record of interconnections with American social sciences (Magat, 1979; Gemelli, 1998; Drèze, 2008). They notably visited the University of Essex, the University of Louvain, and the Ecole Pratique des Hautes Etudes, thus paving the way for providing financial backing to the creation of ECPR, EEA and EAESP (Moscovici and Markova, 2006; Drèze, 2008; Boncourt, 2013).

Aside of these philanthropic resources, some associations also received support from public authorities such as the European Union and national governments. The involvement of the EU comes at a later stage than that of philanthropic foundations and has implications for both hegemonic (such as ECSR) and pluralistic organizations (such as EpsNet). In the 1990s, the EU indeed sought to promote the structuring of a 'European Research Area.' Rather than pushing for the development of specific paradigms – be they of European or American origins – EU officials followed a geographical objective as they aimed at building scientific bridges between Eastern, Western, Southern, and Northern Europe. The main idea behind the funding of the European Political Science Network was thus to support the development of an organization that would be more closely connected to eastern and southern Europe than the existing ECPR was. In some other cases, national governments also offered financial support to the new organizations. The fact that these subsidies were provided by relatively 'small' countries to associations that held their founding meetings on their soil – ESA in Vienna and EEA in Brussels – could be taken as an indication of the fact that hosting European organizations also constituted a way, however marginal, for these countries to enhance their international profile. Just like philanthropic foundations and the European Union, national governments thus followed their own political agenda when providing funding for the creation of European social science organizations.

4. One last source of funding, albeit marginal, needs to be mentioned. The European Economics Association is indeed unique in having secured funding from central banks. This shows the extent to which the possibility for scholars to obtain financial support for their undertakings is also linked to the existence of strong connections between their discipline and nonacademic professional fields. The very nature of economic knowledge may thus have facilitated access to subsidies.

Several factors thus interacted to trigger the creation of European social science organizations. National and international power struggle and competitions came into play at the scientific level (competition between promoters of different paradigms and methods), the academic level (rivalries between universities) and the political level (competition between countries during and after the Cold War). The importance of these different factors varied according to the intellectual (dominant paradigms and methods) and organizational (existence of other European associations) structure of disciplines.

Table 4 Four most represented countries in each organization (2013)

EEA (economics)	*ECPR (political science)*	*EPSA (political science)*	*ECSR (sociology)*	*ESA (sociology)*	*EASP (psychology)*	*EASA (anthropology)*
Germany	UK	The United States	Germany	UK	The Netherlands	UK
UK	Germany	UK	Netherlands	Germany	UK	Germany
The United States	The United States	Germany	UK	Italy	Germany	France
Italy	Italy	Suisse	Norway	Russia	Italy	Italy

The Development of European Social Science Organizations: Membership, Structures and Activities

After adopting a historical perspective to describe the factors and strategies that led to the creation of European social science associations, this article follows a more sociological logic to describe and compare their current structure. It does so by looking, first, at the geographical scope of their membership. It then seeks to explain common points and differences between membership structures by looking at the organizational settings, activities, and intellectual orientations of associations.

The following tables and figures compile basic data on the current membership of eight European social science organizations (EASA, EASP, ECPR, ECSR, EEA, EPSA, and ESA), EpsNet having ceased to exist as an autonomous association after having been absorbed by ECPR in 2007. (As described below, the European Association of Experimental Social Psychology (EAESP) has been renamed European Association of Social Psychology (EASP).) Table 4 lists the four most represented countries in each association and allows us to highlight common points between disciplines and organizations. Some countries are indeed quantitatively dominant in most, if not all cases: Germany and the United Kingdom systematically count among the three most represented countries, and the United States and Italy also display high representation rates. 'Small' countries such as the Netherlands, Norway, or Switzerland also appear to be more represented than their demographical weight could lead one to expect. By contrast, demographically 'big' countries such as France and Spain are underrepresented. Following literature on the history of social sciences (e.g., Klingemann, 2008; L'Estoile, 2008; Gingras and Heilbron, 2009), one could take these numbers as linked to the linguistic position and the specific modes of internationalization of national fields. The strong presence of Germany and the United Kingdom and the relative overrepresentation of small northern European countries such as Scandinavian nations and the Netherlands could then be taken as a manifestation of their openness to the Anglo-Saxon world. Correlatively, the relative absence of France, Spain, and Portugal could be understood as a consequence of the existence of alternative international arenas where French, Spanish, or Portuguese are dominant languages.

By revolving around the properties of national fields, these interpretations enable us to make sense of the common points between the memberships of European social science organizations. However, they provide no satisfactory explanation for the important differences that divide these organizations. These differences are tangible at two levels. First, the number of countries represented in the memberships of organizations may vary by a factor of three, from 20 in the case of ECSR to 66 in the case of EEA (Figure 2). Organizations also differ at the level of the geographical repartition of their members. Figure 3 illustrates these fluctuations by classifying organizations according to the share of Western members (that is, western European and North American members together) that they gather. The case of EPSA, the most Western and American association in the sample (with 96.5% of Western members, against only 1.3% of eastern European members), thus contrasts sharply with that of ESA (67.7% of Western members against 27.7% of eastern Europeans).

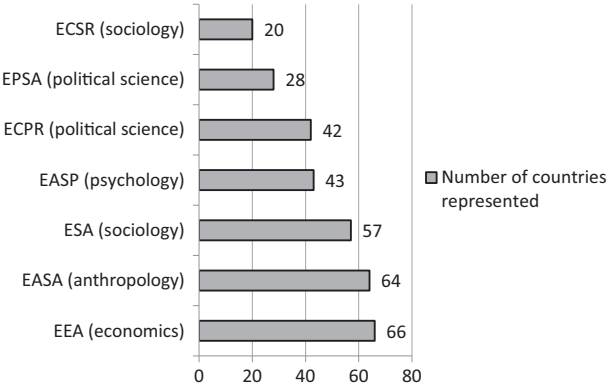

Figure 2 Number of countries represented in each organization.

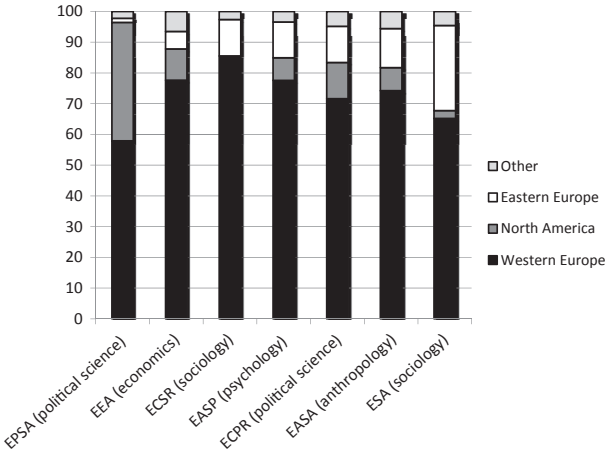

Figure 3 Geographical breakdown of the membership of European social science organizations (percentages).

These variations may be explained by looking at the specific characteristics of European social science organizations. As they face various constrains linked to the specific organizational, scientific, and political contexts they are embedded in, associations indeed follow different strategies to attract members and cumulate financial resources. These differences are tangible at a very early stage of their development and can be observed at three levels. The first is that of the type of members that associations gather, as most organizations work with individual members (EPSA, EEA, EASP, EASA, ESA) while a few are structured as consortiums of academic institutions (ECPR, ECSR). The second level is that of the geographical conditions that organizations set for membership: while some of them have always been potentially open to scholars based in any continent and country, others such as ECPR, ECSR, and EASP originally restricted their membership to European or even west-European academics. The third level is that of activities organized by associations: while most of them focus on organizing conferences and publishing a journal, some add to these classical activities by developing summer schools (ECPR, ECSR, EASP) and research workshops (ECPR).

This diversity of organizational forms plays a part in explaining variations in the geographical structure of organizations' memberships. Conditions for membership have an obvious impact, and other parameters have an influence as well. Consortium formats may indeed hamper geographical diversification, as non-Western academic institutions experience difficulties in paying their high membership fees. In the same way, costly activities such as summer schools may also limit geographical diversity. The fact that ECPR, ECSR, and EASP gather fewer countries than other organizations therefore appears logical.

However, differences in organizational settings should not be taken as the sole explanation for membership fluctuations. Indeed, as time went on, organizations grew and their structures evolved to become more and more similar. As the organizational environment became denser and more competitive (political science having for example up to three European organizations in the 2000s) and political contexts changed (notably in relation to the fall of communist regimes), associations became more acutely aware of the opportunities for growth and the risk of decline. (The case of EpsNet shows that these risks were more than just speculative as, after having been funded by the European Union, the organization failed to gather enough members to remain independent and was eventually absorbed by ECPR.) On the basis of strategic reviews produced by internal task forces, they adapted their rules and activities by importing from other organizations what they identified as 'best practices,' with the explicit objective of attracting new members. This organizational competition and isomorphism (DiMaggio and Powell, 1991) had two main consequences. On the one hand, it led to a softening of conditions for membership: some associations opened their membership to eastern Europe by reducing fees for scholars located in the region (ECPR, ECSR) and others created 'associate' or 'affiliate' membership categories to allow non-European, and especially American academics to join (ECPR, EASP). On the other hand, organizations diversified their activities by publishing new journals (the *European Journal of International Relations*, *European Political Science*, and the *European Political Science Review* in the case of ECPR; the *Journal of the European Economic Association* in the case of EEA; the *European Journal of Cultural and Political Sociology* in the case of ESA, etc.), organizing additional events and conferences (often dedicated to PhD students) and creating scientific prizes (the Hicks-Tinbergen medal in the case of EEA, the Jean Blondel prize for ECPR, etc.). The organizational landscape thus became increasingly uniform as membership options and activities became more numerous and homogeneous across associations.

Therefore, differences in the geographical scope of organizations must also be understood in relation to other factors. More specifically, the intellectual orientation of associations (hegemonic or pluralistic) also appears to have an impact on membership, as it may be more or less compatible with dominant national paradigms. Correlation coefficients between the representation rate of countries and the intellectual orientation of organizations indeed show that paradigmatic factors are linked to membership variations. The hegemonic character of organizations is first strongly and negatively correlated to the number of countries represented in each association (−0.5). Second, Germany and the United Kingdom are shown to be overrepresented in hegemonic associations, while eastern (Rumania, Poland, Russia), and southern European countries (France, Italy, Portugal, Greece) as well as Finland are significantly overrepresented in pluralistic organizations (Table 5). While detailed analyses of these results would go beyond the scope of this article as they would require careful consideration of national scientific histories, the existing literature tends to indicate that these tendencies are not random: the influence of Bourdieusian theories in France (Bouillaud, 2009) and in

Table 5 Correlation between the representation of countries and the intellectual orientation of each association

Country	*Coefficient*
Germany	Above +0.3
UK	
Belgium	Between −0.3 and +0.3
Hungary	
Norway	
Czech Republic	
Switzerland	
Denmark	
Sweden	
Austria	
Spain	
The Netherlands	
France	Below −0.3
Rumania	
Finland	
Italia	
Russia	
Greece	
Poland	
Portugal	

Finland (Rahkonen, 2008) and that of historical materialism in the former Soviet Union (Mespoulet, 2007) are indeed at odds with the paradigms promoted by the founding members of hegemonic organizations.

As the founders and officers of hegemonic associations perceived these paradigmatic orientations as having negative effects on their membership, they deployed strategies to try and soften the weight of this stigma. Some organizations therefore changed their names (such as EAESP giving up its 'experimental' label to become EASP) while others claimed to have an ecumenical approach (such as ECPR, a move that eventually led some of its members to opt out and found EPSA). While the impact of these strategies is difficult to evaluate, the very fact that they are deployed shows the extent to which organizations' officers themselves believe in the influence of paradigmatic labels on membership fluctuations.

Conclusion

This short account of the history of European social science organizations shows the value of applying two types of reasoning to the study of scientific internationalization. The first is relational, as the case of European organizations shows that scientific internationalization is linked to the interaction of scientific, academic, and political actors who all pursue specific agendas on the national and international stage. The second is confrontational as the creation of international organizations is here used a weapon to change or consolidate dominant scientific orders.

This account also leads to a call for further studies of the history of scientific organizations. While numerous researchers devote their work to sociological studies of sciences, few have sought to look at the creation, development and activities of local, national, continental, and global scientific organizations. As this short article shows, a look at the history of these associations may provide valuable insight into the changing intellectual and social structures of sciences over time.

See also: Cold War: Impact on Social and Behavioral Sciences; Social Science Learned Societies and Professional Associations.

Bibliography

Berghahn, V.R., 2002. America and the Intellectual Cold Wars in Europe, Princeton. Princeton University Press.

Boncourt, T., 2009. A history of the International Political Science Association. IPSA, Montreal, Canada.

Boncourt, T., 2011. L'internationalisation de la science politique: Une comparaison franco-britannique. PhD dissertation. Sciences Po Bordeaux, Bordeaux.

Boncourt, T., 2013. Acteurs multipositionnés et fabrique du transnational. La création du European Consortium for Political Research. Critique Internationale 59, 17–32.

Bouillaud, C., 2009. Réflexions sur les (non-)usages des "langages non naturels" en science politique en France au tournant du XXIe siècle. In: Delori, M., Deschaux-Beaume, D., Saurugger, S. (Eds.), Le choix rationnel en science politique. Débats critiques. Presses Universitaires de Rennes, Rennes, pp. 57–70.

Brian, E., 1989. Y a-t-il un objet Congrès? Le cas du Congrès international de statistique (1853–1876). Cahiers Georges Sorel 7, 9–22.

Budge, I., 2006. Jean Blondel and the development of European political science. European Political Science 5 (3), 315–327.

Bull, M., 2007. Is there a European political science and, if so, what are the challenges facing it? European Political Science 6 (4), 427–438.

Carver, E., 1986. Patronage and the directions of research in economics: the Rockefeller Foundation in Europe, 1924–1938. Minerva 24 (2–3), 205–222.

Daalder, H. (Ed.), 1997. Comparative European Politics: The Story of a Profession. Pinter, London.

DiMaggio, P., Powell, W., 1991. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. In: DiMaggio, P., Powell, W. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 63–84.

Drèze, J., 2008. The CORE Project from the Start. Center for Operations Research and Econometrics, pp. 2–3.

Gemelli, G. (Ed.), 1998. The Ford Foundation and Europe (1950's–1970's). Cross-fertilization of Learning in Social Science and Management. European Interuniversity Press, Brussels.

Gingras, Y., Heilbron, J., 2009. L'internationalisation de la recherche en sciences sociales et humaines en Europe (1980–2006). In: Sapiro, G. (Ed.), L'espace intellectuel en Europe. De la formation des Etats-nations à la mondialisation XIXe–XXIe siècle. La Découverte, Paris, pp. 359–379.

Guilhot, N. (Ed.), 2011. The Invention of International Relations Theory: Realism, the Rockefeller Foundation, and the 1954 Conference on Theory. Columbia University Press, New York.

Klingemann, H.-D., 2008. Capacities: political science in Europe. West European Politics 31 (1–2), 370–396.

Kuper, A., 2004. Existe-t-il une "école européenne" en anthropologie? Critique 680–681, 150–159.

L'Estoile, B.D., 2008. Hegemonic gravity and utopian pluralism: a comparative framework for analyzing the international space in anthropology. Journal of the World Anthropologies Network 3, 111–129.

Magat, R., 1979. The Ford Foundation at Work: Philanthropic Choices, Methods and Style. Plenum Press, New York and London.

Mespoulet, M., 2007. La "renaissance" de la sociologie en URSS (1958–1972). Une voie étroite entre matérialisme historique et "recherches sociologiques concrètes". Revue d'Histoire des Sciences Humaines 16, 57–86.

Moscovici, S., Markova, I., 2006. The Making of Modern Social Psychology. Polity Press, Cambridge.

Platt, J., 1998. A Brief History of the International Sociological Association: 1948–1997. International Sociological Association, Montreal.

Popa, I., 2011. La circulation transnationale du livre: un instrument de la Guerre froide culturelle. Histoire@Politique. Politique, Culture, Société 15.

Rahkonen, K., 2008. Bourdieu in Finland: an account of Bourdieu's influence on Finnish sociology. Sociologica 2 (2).

Saunier, P.-Y., 2003. Administrer le monde? Les fondations philanthropiques et la Public Administration aux Etats-Unis (1930–1960). Revue Française de Science Politique 53 (2), 237–255.

Schofer, E., 1999. Science associations in the international sphere, 1875–1990: the rationalization of science and the scientization of society. In: Boli, J., Thomas, G. (Eds.), Constructing World Culture: International Nongovernmental Organizations since 1875. Stanford University Press, Stanford, pp. 249–266.

Sirinelli, J.-F., Soutou, G.-H., 2008. Culture et Guerre froide. Presses de l'Université Paris Sorbonne, Paris.

Solovey, M., Cravens, H. (Eds.), 2012. Cold War Social Science. Knowledge Production, Liberal Democracy and Human Nature. Palgrave Macmillan, New York.

Tournès, L., 2011. Sciences de l'homme et politique. Les fondations philanthropiques américaines en France au XXe siècle. Classiques Garnier, Paris.

Social Science Infrastructure: North America (Resourcing and Funding)

Helmut K Anheier, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides a brief overview of the most important historical developments, current funding sources, and ongoing debates related to social and behavioral research in North America.

Introduction

The United States has developed a system of funding for the social and behavioral sciences that stands out in terms of scale and scope – scale when measured in monetary contributions for research investments and expenditures, and scope when taking into account the variety of funding sources and research fields. It is a system that involves hundreds of agencies, both public and private, issuing thousands of requests for proposals (RFPs) to universities, research institutions, and think tanks, even individuals. Furthermore, it is a competitive, professionalized system linking grant-seekers and grant-givers, and with nonprofit organizations and businesses operating as intermediaries to help with securing and managing funds. In some aspects it is also a highly individualistic system where philanthropists can decide on supporting some research projects rather than others of perhaps greater scientific merit based on their own preferences. Institutional diversity is the hallmark of the system.

The Canadian system of the social science funding shares many of the American features but is smaller in scale and shows less of an involvement of private foundations and individual philanthropy. Overall, however, the development of Canadian research universities and funding pattern has been more of a public matter than it was in the United States, with the main funding body, the Social Sciences and Humanities Research Council (SSHRC) (www.sshrc-crsh.gc.ca). What is more, the special status of Quebec means that this province has its own governmental funding body, the Fond de recherché de Quebec www.frq.gouv.qc.ca/en/. Nonetheless, the American and the Canadian funding systems show many connections in terms of RFPs open to applicants from either side of the border, joint programs, and common research teams.

The American social science funding system undoubtedly forms the world's largest and is unlikely to be rivaled in the short to medium term by the European system (even with the creation of the European Research Council and massive programs such as Horizon 2020), and the emerging Asian systems, with China being the most likely candidate. What is more, despite the fact that the great majority of funding in the US system remains within the country, the last decades have seen a rise in international activities. In other words, social science funding in North America is a system that operates not only at local, state, and federal levels but is also becoming increasingly international, given the various cooperation agreements signed, for example, between the National Science Foundation (NSF) (www.nsf.gov) and counterparts in Europe such as the German Research Council (www.dfg.de) or the Japan Society for the Promotion of Science (http://www.jsps.go.jp) in Asia.

Not surprisingly, the largest share by far of funding is for the natural and life sciences, including medicine. It is also here that the political–financial link to national defense requirement is clearest, and where the rationale of 'being in the national interest' is more easily made and communicated. In 2011 and 2012 social sciences and psychology represented approximately 5% of all higher education research expenditures (Britt, 2013). As we will see, this share has come under political pressure. We will return to this point in the concluding section after having addressed the historical development of the system as such.

A Historical Overview of Social Science Funding in North America

In the late nineteenth and early twentieth century, the modern American university system took root. This meant moving away from a system of typically small, denominational colleges to a diversified system of major research universities, both public and private, on the one hand, and great variety of colleges on the other. This development was primarily facilitated by both philanthropy and the land-grant movement. Private philanthropy played an especially important part in supporting the emerging social sciences, for example the Russell Sage Foundation which made a large contribution to the study of urban problems in New York City, or the creation of the University of Chicago by the Rockefeller Foundation (see Wheatley, 2010). Philanthropy also provided funding for institutional development such as the creation of the American Council of Learned Societies or the Social Science Research Council (SSRC) (see below). Foundations not only supported research projects, they also established academic departments, and generally helped with the development of individual disciplines such as economics and sociology. They established graduate schools and were instrumental in bringing to the United States highly qualified researchers from Europe, the then center of the world's social science community. One of the most influential US think tanks, Brookings, was established in 1927 through a merger of two sister institutions (Institute of Economics and Graduate School). The think tank, which carries out research on national and international topics, is today financed by a characteristic mix of government agencies, foundations, and leading corporations.

The federal government played only a small role until the 1930s and the inception of the New Deal; however, it began to emerge as a major actor during World War II. As part of the war effort, the government supported war-relevant research, often at a large scale, leading, for example, to the establishment of the RAND Corporation for developing the field of operations research (http://www.rand.org/). It is now next to Brookings, one of the world's leading think tanks. Sociology departments were commissioned to study the morale of combat troops (Stouffer et al., 1949), which resulted in theoretical developments (e.g., reference group theory) and methodology innovations (questionnaire design). Institutions carrying out work in the field of economics, such as the Department of Economics at Chicago, received resources to build and strengthen research capacities, in this case to develop a system of national income accounts to help plan macroeconomic policy in the aftermath of the great depression and in the face of shortages during World War II.

The 1940s and 1950s were also a time when corporations began funding the development of social science methods, especially audience research for radio stations (later also television) and advertising firms. Good example are the influential compilations of research methods by Paul Felix Lazarsfeld and Morris Rosenberg (1955), which owe much to support by media corporations and foundations alike, as did the creation of professional polling firms such as Gallup. In 1941 the National Opinion Research Center at the University of Chicago was established by Harry Field, who had previously worked for Gallup. The center was supported by George Gallup, Elmo Roper, and Archibald Crossley. George Gallup introduced Harry Field to Hadley Cantril (Princeton University), who in turn put him in touch with important social scientists of the time, such as Gordon Allport at Harvard University, to help secure necessary funding. The Marshall Field Foundation then agreed to help establish the Center with a 5-year grant totaling $250 000 (www.norc.org). In the 1960s, President Johnson pushed for the creation of another important US think tank, the Urban Institute, which was launched in 1968 to focus on applied social science research and policy applications. It is funded by the US government as well as private resources (in 2013: 54% federal funding, 39% private funding, and 7% other funding) (www.urban.org/).

In the decades to come, a distinct pattern emerged: different levels and agencies of government and philanthropy provided impulses for specific kinds of research and education and directed their funding streams accordingly, be it the Green Revolution or the Area Studies programs in the 1960s, race relations in the 1970s, health behavior research relating to the spread of HIV-AIDS in the 1980s, civic engagement in the 1990, or Big Data today.

Significantly, greater federal involvement did not replace the importance of local and state funding; on the contrary, it was often accompanied by parallel developments. The funding infrastructure of the state of California is largely a product of the postwar era when federal funding continued to grow. Nor did the establishment of the main governmental funding vehicle, the NSF in 1950, diminish the role of philanthropy. It mostly added to it and the rise of government funding encouraged philanthropy to seek new roles. For example, federal funding of students (loan programs, stipends such as Pell grants (www.studentaid.ed.gov/types/grants-scholarships/pell)) runs in parallel with support by thousands of foundations, often operating at the local level or targeting special applicant pools. Over time, a vast and diverse system of resourcing social science research emerged.

Funding Social Sciences: Portrait of the Main Funding Sources in the United States

Broadly speaking, the funding of the social sciences in the United States comes from four basic sources: government (at the federal and state level), the SSRC, foundations, and corporations. The following section sketches some larger agencies that fund the social and behavioral sciences. It concentrates on federal government agencies and foundations. The selection of the agencies and foundations is not representative and instead serves to illustrate the diversified landscape of funding.

Government

There are several key public agencies at the federal and state levels that sponsor and support the social sciences. Among the largest funders at the federal level are the NSF, the National Institutes of Health (NIH), and the Centers for Disease Control and Prevention (CDC).

National Science Foundation

The NSF was established in 1950 by Congress "to promote the progress of science; to advance the national health, prosperity and welfare; to secure the national defense …" (www.nsf.gov). With an annual budget of about $7 billion (2014), of which $6 billion are devoted to research, the NSF makes up around one-quarter of all federally supported basic research conducted in US colleges and universities.

The NSF is the largest funder of social science research in the United States. It makes up 63% of all federal support for basic research in the social sciences in US academic institutions. The NSF claims that in some fields, including anthropology, archaeology, political science, economics, sociology, and the social psychology, it is the predominant or exclusive source of federal support for basic research and infrastructure development. The Directorate for Social, Behavioral and Economic Sciences at NSF (SBE) supports research that builds fundamental knowledge of human behavior, interaction and social and economic systems, organizations and institutions (www.nsf.gov/dir/index.jsp?org=SBE). Current research and development priorities revolve around topics and issues such as homeland and national security, economic prosperity, integrating research and education and environmental quality. SBE emphasizes research that links different scientific fields, namely cooperations between the behavioral and social sciences and the engineering, physical, and life sciences fields. Moreover, it promotes international scientific cooperations (**Figure 1**).

As the two graphs and the chart below illustrate, funding for the social, behavioral, and economic sciences accounts for a relatively small portion of NSF's overall research budget. In 2013, social science funding totaled $216 million, which is around 4% of the agency's total research budget of $5.4 billion. Clearly, funding for mathematical and physical sciences,

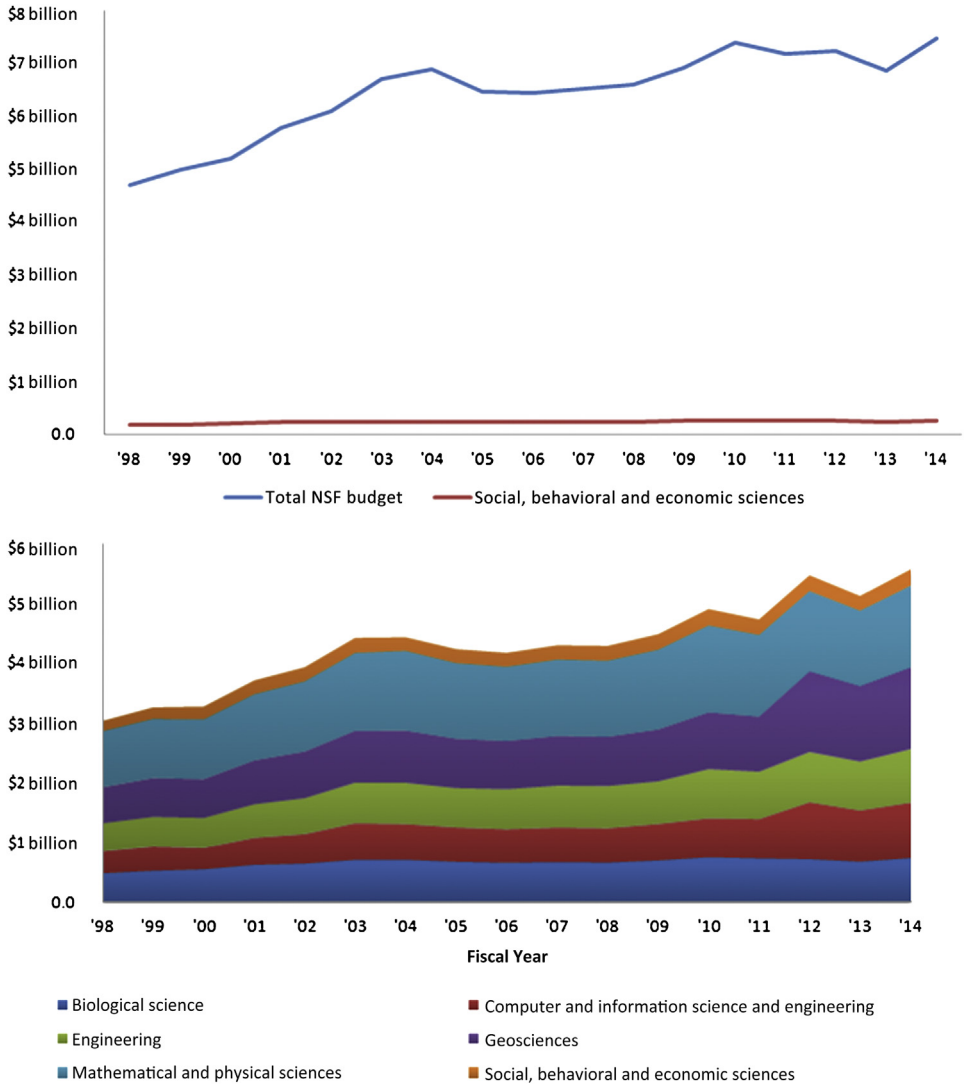

Figure 1 Trends in the NSF budget, 1998–2014. Source: http://www.socialsciencespace.com/2014/03/how-much-nsf-funding-goes-to-social-science/ based on NSF data.

geosciences, computer and information science, and engineering far surpasses the funding available for the social sciences. While the total funding for research and development has grown over the last decades, funding for the social and behavioral sciences has remained relatively stable. However, the 2014 bill to emerge from Congress stipulates a substantial cut in the agency's social science budget.

Generally speaking, the NSF funds research in nearly all fields of science and engineering. It does so through limited-term grants, cooperative agreements, purchase of research equipment for colleges, universities, and other public and private research institutions, and through fellowships for individuals or small groups of researchers. It also supports public and private cooperations at the national and international level. The NSF receives around 40 000 research proposals each year, of which 11 000 are funded. In 2013, approximately 4500 thereof came from the social and behavioral sciences, of which 900 received funding amounting to an annual median size of $45 000. Proposals are usually submitted in response to announcements on the NSF website but may also be submitted independent of such calls.

To fulfill its mission, which is to fund fundamental research in all fields and engineering – except for medical sciences – it uses a merit review process. Proposals are reviewed by a minimum of three selected and external researchers from a national pool of experts. Based on these inputs, NSF program officers will make an award or decline recommendation. The NSF consists of a director who oversees the NSF staff and management and a 24-member National Science Board that establish the overall policies. Each of them is appointed by the US President and confirmed by the US Senate.

National Institutes of Health

NIH is part of the US Department of Health and Human Services and is the primary US research agency in biomedical and health-related matters (www.nih.gov). NIH conducts its own research via its Intramural Research Program (IRP) and funds external medical research via its Extramural Research

Program. NIH with its IRP comprises 27 institutes and centers – among them the Library of National Medicine, an international source for researchers, and National Clinical Center, the world's largest hospital – and employs about 1200 principal investigators and 4000 postdoctoral fellows. This makes the NIH the world's largest research institution for biomedical science (irp.nih.gov/about-us/organization-and-leadership). NIH is also the largest source of funding for medical research in the world (www.nih.gov/about). Overall, the NIH spends about $30 billion annually on medical research. It devotes more than 80% thereof to more than 2500 universities and 300 000 researchers in the form of 50 000 competitive grants (www.nih.gov/about/budget.htm); 10% of the annual $30 billion budget is used to support research activities within NIH's own institutes and centers.

In recognition of the key role played by behavioral, social, and lifestyle factors in the cause, treatment, and prevention of illness and disease, the NIH created the Office of Behavioral and Social Sciences Research (OBSSR) (obssr.od.nih.gov/about_obssr/mission/mission.aspx). Since 1995 it advises and consults NIH scientists and others within and outside the federal agency. The office initiates and encourages studies to evaluate the contribution of behavioral, social, and lifestyle determinants in the development of public health-related problems and sponsors (inter-)national seminars, symposia, conferences, and workshops on state-of-the-art behavioral and social research. Although OBSSR does not have grant-making authority it has been active in organizing and funding a variety of trans-NIH research programs (through transfers to NIH Institutes and Centers). One of its instruments has been the NIH Basic Behavioral and Social Science Opportunity Network (oppnet.nih.gov). NIH-supported behavioral and social research includes research on behavioral and social processes, on interactions between biology, behavior, and social processes, and basic research on methodology and measurement in the behavioral and social sciences. At the individual level this involves studies on behavioral factors such as cognition, memory, and emotion. At the aggregate level it includes studies on the role of social variables such as the structure and dynamics of small groups, institutions, organizations, and larger structures (e.g., demographic, political, and economic factors) in public health-related matters. A list of funding opportunities offered through OPP Net and a number of other behavioral and social science funding opportunities announcements across the NIH can be found here: oppnet.nih.gov/funding-bssr-funding.asp.

Centers for Disease Control and Prevention

CDC is the US public health agency and part of the US Department of Health and Human Services. It is responsible for protecting public health and safety through monitoring, control, and prevention of infectious and noninfectious disease (such as diabetes or obesity), injury, and disability (www.cdc.gov). It does so increasingly with global reach, having established several regional centers (in China, Egypt, Guatemala, India, Kazakhstan, Kenya, South Africa, and Thailand) to detect outbreaks of infectious disease in different geographic areas and attempting to prevent said infectious diseases from being brought into the United States. CDC's activities include detecting and responding to health threats, providing specialized training in field investigations and laboratory methods, preparing and responding to pandemics, detecting and preventing zoonotic (animal–to–human) disease transmission, and providing emergency preparedness and risk communication training (www.cdc.gov/24-7/cdcfastfacts/nationhealth.html?s_cid=cdc_homepage_cdcinaction_001).

CDC implements its mission in collaboration with other US agencies at federal and local levels and other health-related agencies. The CDC thereby provides funding through a range of programs (e.g., Global Health Security Program, the Antibiotic Resistance Program, Food Safety Program, Prevention of Sexually Transmitted Diseases). CDC's annual budget revolves around $6.5 billion (FY 2015), of which it awards nearly 85% in competitive grants and contracts to research organizations and other health-related organizations that contribute to CDC's mission of health promotion through dissemination of health information, preparedness, prevention, research, and surveillance. CDC's mission is subject to congressional legislation.

Social and behavioral scientists help the agency to understand risk groups and the frequency, context, and determinants of risky behavior. They assist in developing, implementing, and assessing the agency's prevention programs (Snider and Satcher, 1997).

Social Sciences and Humanities Research Council

The SSHRC is Canada's federal research-funding agency which promotes and supports postsecondary-based research and training in the humanities and social sciences. Created in 1977, the SSHRC focuses on "developing talent, generating insights and forging connections across campuses and communities" (http://www.sshrc-crsh.gc.ca). SSHRC seeks to promote innovative, interdisciplinary, and international research on social, cultural, technological, environmental, and economic issues and is committed to the goal of ensuring "a better future for Canada and the world" (http://www.sshrc-crsh.gc.ca). SSHR is governed by a council appointed by the federal government to represent the interests of the academic, public, and private sectors. In 2010, following consultations with researchers, postsecondary institutions, and partner organizations, SSHRC made significant changes to its range of funding opportunities by reducing their number, increasing their flexibility, and aligning their objectives under three new umbrella programs: Talent, Insight, and Connection (http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/strategic_plan_2013-16-plan_strategique_2013-2016_e.pdf). While the SSHR has identified six key future challenges, it supports a broad range of topics and disciplines. Funding opportunities include grants, fellowships, and scholarships. In 2012–13 the SSHR-supported research totaled CAN$337 million. SSHR received more than 12 000 applications, of which 5000 were successful.

Social Science Research Council

The SSRC is a nonprofit organization initially founded through an initiative of the American Political Science Association in 1923 (www.ssrc.org). Its goal is to advance the social sciences by fostering innovative, interdisciplinary research, mobilizing different knowledge around important public issues and ethical values, such as justice, prosperity, and democracy, and

lastly to carve out niches in research programs. It seeks to innovate and identify emerging lines of research, so as to promote new fields of research and advancing social change. The SSRC particularly establishes and supports international research and international research collaborations that bring together academics, practitioners, and policy-makers. The SSRC's agenda, in terms of topics and activities, has changed over the course of time and has been responsive to changing problems and needs. In recent years it has focused on issues such as peace-building, human development, understanding the role of religion in the United States, and strengthening a global social science. The SSRC has an annual budget of $23 million. It is sourced from public and private institutions, that is from different (United States and foreign) governments, foundations, and corporations.

Foundations

According to the Foundation Center there are about 86 000 grant-making foundations in the United States (Foundation Center 2012). They bring a sizeable endowment to the table for financing development and research in general. There are several types of grant-making foundations, most of them (78 000 that is, 91%) being independent foundations. In 2012, independent foundations held assets worth $584 billion. Only about 2600 of the grant-making foundations are corporate foundations. They held about 6% of all assets (http://data.foundationcenter.org/#/foundations/all/nationwide/total/list/2012).

As outlined above, foundations have played a critical historical role in the setting up of the American university system. They have done so directly through the funding of specific projects and indirectly through the funding of higher education and teaching more in general. As such, they have also reinforced the idea of research and teaching (Wheatley, 2010, p. 74).

Although major funding projects by foundations can be found in the realm of the natural sciences, in 2011 there were nearly 300 foundations with a grant value of approximately $240 million financing social and behavioral science projects (http://data.foundationcenter.org/#/fc1000/subject:social_science/nationwide/top:foundations/list/2011).

Alfred P. Sloan Foundation

The foundation of the former President and Chief Executive Officer of General Motors Alfred O. Sloan has total assets of $1.7 billion, with around $20 million spent on social science/interdisciplinary projects in 2011. Driven by the idea that a "reasoned and systematic understanding of the forces of nature and society, when applied inventively and wisely, can lead to a better world for all," the Foundation focuses on science, technology, and economic institutions (www.sloan.org). It funds basic natural science research programs for a better understanding of digital information technology. Furthermore, it promotes access to higher scientific education and advances the public understanding of science, technology, and economics. The social science subprograms have a strong emphasis on economics and deal, amongst others, with the economic implications of the great recession and behavioral economics/household finance.

Andrew W. Mellon Foundation

The Mellon Foundation was established in 1940 as the Avalon Foundation. In 1953 it merged with Old Dominion Foundation and in 1969 changed to its current name (http://www.mellon.org). It focuses on higher education and scholarship, scholarly communications and information technology, art history, conservation, museums, and performing arts. It has total assets of $5 billion, with around $16 million spent in the social sciences. The Foundation supports a wide range of initiatives to strengthen the institutions that sustain scholarship in the humanities and 'humanistic' social sciences, primarily research universities but also a small number of centers for advanced study and independent research libraries. There are several fellowships like the Charles A. Ryskamp Research Fellowships or The John E. Sawyer Seminars on the Comparative Study of Cultures.

The Rockefeller Foundation

The Rockefeller Foundation was founded by John Rockefeller, his son John Rockefeller Jr., and Frederick Gates in 1913. The Rockefeller Foundation's historical mission is "to promote the well-being of mankind throughout the world" (www.rockefellerfoundation.org). One of its core objectives is to finance education, in particular higher education without discrimination by race, sex, or creed. Nowadays it primarily supports work that expands opportunity and strengthens resilience to social, economic, health, and environmental challenges. As an example, the Rockefeller Foundation Bellagio Center promotes interdisciplinary research through conferences and residency programs. The foundation has assets amounting to $3.6 billion. The value of grants made to the social sciences in 2011 was $8.8 million.

Ford Foundation

Founded by Edsel and Henry Ford, founder of the Ford Motor Company, the Ford Foundation began as a local philanthropic organization in 1936. It grew to be one of the largest and most influential foundations with 10 overseas offices, providing loans and grants of over $16 billion worldwide (http://www.fordfoundation.org/). Its mission is to "strengthen democratic values, reduce poverty and injustice, promote international cooperation and advance human achievement." With assets totaling $11.2 billion and a total of $7 million spent in the social sciences in 2011, it funds research that supports the work of different initiatives as well as public policy analysis that informs advocacy and public mobilization. Causes supported by the Ford Foundation range from 'democratic and accountable government' to 'gender, sexuality and reproductive justice' to sustainable development and human rights.

Bill & Melinda Gates Foundation

With total assets of around $37 billion, the Bill & Melinda Gates Foundation is one of the largest foundations worldwide.

Established in 1994 as the Wiliam H. Gates Foundation, it merged in 2000 with the Gates Learning Foundation to become the Bill & Melinda Gates Foundation. Driven by the understanding that every life is of equal value, it focuses on extreme poverty and poor health at the global level as well as failures in the US education system (www.gatesfoundation.org). In terms of North American social science funding, its focal point lies on research connected to education policies as well as migration.

Summary and Outlook

In the previous edition of the Encyclopedia for Social and Behavioral Studies, S.E. Cozzens came to a rather positive conclusion regarding the future of research in the United States. However, this prediction may have underestimated the tensions that are embedded in the American funding system, especially in the context of social and behavioral studies.

There is a fault line inherent in the US system, which is shaped by political, economic, and academic influences. The last few years have witnessed a fierce debate about governmental funding for research and development in the social science research area. Several Republican Members of Congress accused the NSF of misusing tax dollars for grants spent on social and behavioral studies research. The debates culminated in the bill on 'The Frontiers in Innovation, Research, Science and Technology (FIRST),' which was brought to the US House of Representatives in the 2013–14 session. This bill pronounces a clear priority of research and development in the 'basic research fields,' which are identified as being biology, chemistry, physics, computer science, engineering, and mathematics, and simultaneous cutbacks in funding for social and behavioral studies unless they are 'in the national interest' however defined. There are, though, politicians who fear a danger in the cutback of funding for social and behavioral studies, as the discipline covers important research areas such as terrorism, disaster management, global health, and others, and researchers in the field conduct significant long-term studies, for example on voting patterns.

What is more, there is also a major debate about the relevance and impact of the social sciences in both the United States and Great Britain, which is led, amongst others, by Patrick Dunleavy of the London School of Economics. Dunleavy strongly emphasizes the impact and contributions made by the social sciences to public policy issues, economic success, and to the awareness of the public in terms of understanding policy issues and social and economic changes.

It seems that the inherent tensions in the complex and diversified funding system have come to the forefront. The coming years will see if the system can accommodate the preferences of the various stakeholders (the academic community, government, political parties, philanthropy, and corporations) as well as it did in the past.

See also: Funding of Social Science, History of; Research Funding: Ethical Aspects; Social Science Infrastructure: European Union (Resourcing and Funding).

Bibliography

Anheier, H., Daly, S. (Eds.), 2007. The Politics of Foundation. Routledge, New York.
Anheier, H., Hammack, D. (Eds.), 2010. American Foundations. Brookings Institution Press, Washington, DC.
Britt, R., 2013. Higher Education R&D Expenditures Remain Flat in FY 2012. http://www.nsf.gov/statistics/infbrief/nsf14303/#fn1 (last accessed 05.09.14.).
Bastow, S., Tinkler, J., Dunleavy, P., 2014. The Impact of the Social Sciences: How Academics and Their Work Make a Difference. Sage Publications Ltd, London.
Cozzens, S., 2001. Autonomy and accountability for 21st century science. In: de la Mothe, J. (Ed.), Science, Technology and Governance. Pinter, London.
Lazarsfeld, P.F., Rosenberg, M., 1955. The Language of Social Science Research. A Reader in the Methodology of Social Research. The Free Press, New York.
Snider Jr., D.E., Satcher, D., 1997. Behavioral and social sciences at the Centers for Disease Control and Prevention. Critical disciplines for public health. American Psychologist 52 (2), 140–142.
Stouffer, Samuel A., Lumsdaine, Arthur A., Lumsdaine, Marion H., William, Robin M., Smith, M. Brewster, Janis, Irving L., Star, Shirley A., Cottrel, L., 1949. The American Soldier: Combat and Its Aftermath. Princeton University Press, Princeton, NJ.
Wheatley, S., 2010. In: Anheier, H., Hammack, D. (Eds.), American Foundations. Brookings Institution Press, Washington, DC, pp. 73–97.

Relevant Websites

http://data.foundationcenter.org/ – Foundation Center.
http://www.grants.gov – Grants.
www.nsf.gov – National Science Foundation.
http://dellweb.bfa.nsf.gov/starth.asp – Statistical and funding information on NSF budget.
http://www.sshrc-crsh.gc.ca/home-accueil-eng.aspx – Social Sciences and Humanities Research Council.

Social Science Infrastructure: North America (Social Science Academies and Related Organizations)

Steven Rathgeb Smith, American Political Science Association, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Academic associations expanded greatly in scope and diversity in the last 50 years, fueled by rising citizen demand for higher education, the increase in faculty positions and doctoral programs, and the growth of government funding. But growth is now much more difficult due to increased competition, fewer tenure track positions, the decrease in public funding for higher education, the shift in foundation priorities, and the digital revolution, which has given scholars a myriad of ways to network and access important information. Consequently, academic associations need to rethink their value proposition for membership, with important implications for their programming, advocacy and public engagement roles, and governance. This article focuses on academic associations in the social sciences and offers recommendations and next steps as they seek to remain relevant to the scholarly community, policymakers, and the citizenry.

Introduction

Academic organizations have played a crucial role in the growth of higher education in countries around the world in recent decades. These organizations include associations of scholars, coalitions representing higher education interests, research institutes, and think tanks. They often work in concert with government agencies to support the increase in the professoriate, and give citizens more opportunities in higher education. Associations grew in numbers and influence at the same time as faculty employment, student enrollment in higher education, and the number of doctoral programs were increasing. But academic organizations in the social sciences are facing a rapidly changing social and political environment, creating uncertainty and new challenges to their mission, strategic priorities, and role in public policy. This article focuses on academic associations in the social sciences, such as the American Political Science Association (APSA) and the International Studies Association (ISA) in the United States, although attention will also be devoted to other important academic organizations. The article will offer a historical perspective on academic associations, focusing on the social sciences, followed by an examination of the key shifts in the political and social environment of academic associations. The article then offers an analysis of the next steps for academic associations as they seek to remain relevant to the scholarly community, policymakers, and the citizenry.

Growth and Opportunities

Universities have existed for at least several 100 years with many universities in Europe founded in the Middle Ages. But academic associations are a phenomenon that is much younger, with many of the large disciplinary associations established in the late nineteenth and early twentieth century. For instance, the APSA was founded in 1903, reflecting the growth of universities in the United States and the trend toward greater professionalization in society at large. Nonetheless, these associations lacked substantial organizational formality and were often unincorporated groups of scholars who used the association to provide networking opportunities and professional advancement. Associations functioned like loose networks of scholars who came together for academic meetings and to discuss important disciplinary issues. For most of the twentieth century, the major disciplinary associations such as APSA, the American Sociological Association (ASA), the American Anthropological Association, and the American Economic Association (AEA) largely had a monopoly on membership in their particular field or discipline: academics needed to join one of these organizations in order to obtain access to key disciplinary journals, which were owned by the association. Moreover, associations offered an annual conference or summit, for which attendance was essential to upward mobility in many professions. Further, leadership in associations was critical to professional development and advancement especially at large elite universities. Indeed, the larger elite universities tended to dominate professional associations in terms of leadership, journal editorship, and committee participation. Associations were thus intermediaries between individual faculties, students, and important content (e.g., peer-reviewed articles). Moreover, they provided access to faculty job listings.

Social science associations also benefited from the general growth of higher education in the mid to late twentieth century. In the United States, federal funding of higher education, especially in the sciences, began to take off during World War II. This type of federal funding continued after the war and helped to build university infrastructure and fuel university expansion. But federal funding rose sharply in the 1960s with the overall growth of the federal role in public policy. A wide variety of federal funding programs flowed to universities: research grants in the social sciences, loans and grants for college attendance, and direct support for doctoral education. Overall, more extensive federal funds greatly increased employment opportunities in the social sciences. Consequently, academic associations grew accordingly: APSA membership, for example, rose from 7624 individual members in 1964 to 13 663 members in 1970. Other associations experienced a similar increase in membership (APSA, 1970).

Throughout this period, social science associations tended to pursue a traditionalist view of the proper function of

a learned society. The focus of association activities was on the annual conference, the publication of association journals, and overall support for research and teaching. Associations rarely engaged in advocacy; thus they did not actively strive to influence public policy, except for issues that directly related to the discipline such as government research funding. And even in these cases, the advocacy was often delegated to third-party coalitions and associations. Also, efforts to recruit and organize members were quite limited. Indeed, associations did not need to really devote substantial time and resources to member recruitment because membership rolls were rising without any effort on the part of the associations, thanks to the expansion of higher education institutions, fueled by increased federal funding, and demand for higher education on the part of the baby boom generation.

A Shifting Context for Academic Associations

In 1976, during the presidency of Jimmy Carter, the social and political environment for higher education began to change. The economy was plagued by slow growth, high inflation, and stagnant wages, leading to fiscal austerity by the government. Many grant programs established in the 1960s were cut quite sharply. Universities responded to cutbacks in public funding by reducing the number of faculty openings and, in some circumstances, eliminating smaller departments and faculties, especially in the social sciences and humanities. Membership and attendance at the annual conferences declined, albeit to varying extents depending on the association. Moreover, the Reagan administration in the 1980s initiated a long-term restructuring of the federal government, which caused a shift away from extensive federal support for higher education. Moreover, universities raised tuition fees and more directly competed with each other for students, as well as public and private funding. Yet, the number of individuals receiving doctoral degrees in the social sciences continued to rise during the 1980s and thereafter. But the number of applicants per job opening increased substantially as the growth of individuals with doctoral degrees outpaced the number of faculty openings. Membership growth in many associations stagnated as more individuals with doctoral degrees did not land a job in the academy.

Although membership growth in academic associations hit a plateau, associations still retained their monopoly on the university-based scholar, who needed to participate in academic associations for professional advancement. Despite the lack of growth in membership, the business model of academic associations continued to be sustainable: most of the association revenue was generated by membership and conference registration fees; special programs were externally funded by federal and foundation grants; and private philanthropy tended to be limited to special fund-raising campaigns rather than for operating funds. Formal development efforts were thus quite targeted. In short, associations had sufficient revenue to serve their members, albeit at a lower level than earlier periods.

The central position of the large disciplinary associations for professional social scientists began to erode with the establishment and growth of more specialized associations. Many of these younger associations were also interdisciplinary, attracting individuals from a wide variety of disciplines. Good examples include the ISA, the Association for Public Policy Analysis and Management, and the African Studies Association. But many regional associations were created as higher education expanded at the national and subnational level. These newer associations were important because they provided networking opportunities, as well as access to research and publications that had long been the province of the large disciplinary associations. The multiplying number of associations also reflected the increasing diversity of interests within academic associations. In earlier decades, associations were, as previously noted, relatively small, and the range of teaching and research interests was modest. But a variety of trends intersected to increase the diversity of the disciplines including globalization; more openness to new and different research approaches; and changes in association rules and procedures to promote greater representation of different interests. The expansion of higher education also provided more incentives for faculties to publish and network. Being more flexible, smaller associations offered an attractive alternative to large associations, especially for faculties from nonelite institutions.

Changing Landscapes for Associations

The Digital Revolution

The proliferation of smaller and more specialized associations was greatly abetted by the digital revolution and the advances in the World Wide Web. With the creation of email and other Web-based technological innovations, communications among like-minded scholars became easier, cheaper, and faster, reducing the transaction costs of creating a new association. Web-based communications also facilitated the sustainability of these newer associations.

Over time, new associations supplanted the large, disciplinary associations as the primary academic associations for some scholars.

The digital revolution has also fundamentally transformed the relationship between associations and their members. Scholars can now access association journals through their university library systems, effectively decoupling membership from journal access. Secondly, the Internet has made it far easier to network with colleagues, publicize member research, and learn about new research. Many political scientists, for example, are actively engaged in blogs where important substantive political analysis is posted and discussed. Scholars can now quickly disseminate their papers to colleagues around the world and conference attendance is no longer necessary to network. Thirdly, the digital revolution has made it possible to implement new evaluation tools for faculties. In the past, conference participation through a paper presentation or panel chair role was considered an important signal of your standing in the profession. Publications in university press books and peer-reviewed journals were, of course, also considered important for a scholar's prospects for promotion. But the advent of Google Scholar and journal impact factors that measure the utilization of published journal research have created new evaluation tools for faculties that reduce the value of conference participation as a signal of research quality and

scholarly prominence. Fourthly, the digital revolution has directly and indirectly encouraged the increase in new academic journals outside the orbit of academic associations. In earlier decades, larger disciplinary associations controlled the major journals in their field and received royalties from journals to support the association. By reducing operational and start-up costs, the digital revolution promoted the creation of new journals. One of the more notable examples is the new concept of open access journals that are entirely online, and allow any person to access the journal content. Many of these journals also publish articles on a rolling basis, meaning that accepted articles are placed online shortly after acceptance.

In its essence, the digital revolution has changed the production and dissemination of knowledge because scholars can easily access scholarly content that can help them in their own research and teaching; further, easy access to the Web also facilitates greater diversity in the production outlets of knowledge.

Public and Foundation Funding

The digital revolution of the last 20 years has coincided with an altered relationship between universities and their faculties, with important ramifications for academic associations including social science associations. During the mid to late twentieth century, the university infrastructure grew and faculties were rewarded with more positions, including assurances of tenure and decent salaries. But governments have encountered serious fiscal problems in recent years, leading to cutbacks in public support of higher education. As a percentage of the total number of faculty positions, the number of tenured positions has dwindled around the world, including the United States, albeit to varying degrees.

Support for research and conferencing has also become more constrained, especially at nonelite institutions. In the United States, private foundations provided substantial grants throughout the 1960s and thereafter to stimulate the expansion of the higher education infrastructure, including universities and academic associations. Associations received grants to hire faculties and launch new programs and departments. For example, a wave of public policy graduate programs was started in the 1970s with seed money from leading foundations. Similarly, many universities created geographically specific programs such as Russian studies or Latin American studies with federal and foundation grants.

Foundation funding was critical in launching and expanding the intermediary infrastructure of higher education as well. Starting in the 1960s and 1970s, many associations received funding to launch new initiatives on topics such as professional development and special conference programing. Moreover, leading national foundations supported the creation of new associations and interest groups to represent universities and higher education. In the field of nonprofit and philanthropic research, for instance, foundation funding directly led to the establishment or expansion of several intermediary organizations: the International Society for Third Sector Research and the Independent Sector, the Association for Research on Nonprofit Organizations and Voluntary Action, the Aspen Institute research fund, and the Urban Institute research program on nonprofit and philanthropic organizations. Over time, though, the expanded infrastructure produced greater competition among these organizations. Newly established, more specialized associations now competed with the traditional disciplinary associations for members and the new research programs at think tanks such as the Aspen Institute, the MDRC, the Urban Institute and Mathematica now competed for federal and private foundation contracts and grants.

This competition has intensified as foundation priorities have shifted in the last 10–15 years. After years of building the academic infrastructure, many leading foundations have refocused their funding priorities on more applied fields such as early childhood education, school reform, youth mentoring, poverty alleviation, and health prevention. The Gates Foundation, the largest foundation in the world, is a good example of this general tendency among contemporary foundations: a keen interest in directly addressing social problems with a relative disinterest in traditional academic research unless it has an applied component (such as direct implications for policy and practice.)

This shift in foundation grant practices has had important ramifications on the business plans of academic associations and other organizations, such as think tanks. Foundation funding helped to cross-subsidize operating expenses including the administrative overhead of associations. With the decline in foundation funding, academic associations such as APSA are, by necessity, forced to be more reliant on other sources of revenue including member fees, conference registration fees, and earned income such as building rent and corporate sponsorships. Creating a viable, sustainable business strategy is challenging, given the current competitive climate for members and grants. As discussed previously, large associations have a sizable membership base, which allow them to generate significant revenue from membership fees; yet, these fees have risen over time and many universities have reduced their funding for conference travel and association membership. Further, large associations face competition from smaller associations who typically have less expensive fee structures. Consequently, large associations cannot easily raise charges or conference registration fees to compensate for declines in external funding.

The change in foundation funding priorities has also had ripple effects on other types of academic organizations. Many intermediary organizations have been forced to restructure their operations and try to generate new sources of revenue. Think tanks and research firms have adapted by seeking new grant opportunities, partnering with other organizations, and broadening the scope of their research activities.

Smaller and younger membership associations face a similar dynamic as the large disciplinary associations such as APSA and ASA, but they are more reliant on their annual or biennial conference for revenue. The conference generates income through registration fees and encourages new memberships and renewals. Even in this scenario, conference fees may only generate enough income for a relatively small staff, complicating efforts to diversify the revenue base and recruit new members. Moreover the proliferation of associations and conferences has in turn meant that it is more difficult to use the conference for significant boosts in

revenue, without substantive changes in paper and panel acceptance policies.

Federal Policy

Changes in foundation priorities have occurred at the same time as important shifts in federal funding of research and teaching for higher education institutions and the social sciences in particular. Associations themselves have never been big recipients of federal grants, although they did receive funding for training, education and exchange programs in the post-1960s era. Federal funding supported the research of association members, especially in the social and behavioral sciences through the National Science Foundation (NSF), as well as other federal agencies. The opportunity to sit on federal grant review panels was a prestigious honor and the grant application process was very competitive. The annual conference of academic associations provided a vehicle for networking on possible grant applications; indeed, the annual conference often provided an opportunity to plan a grant application or showcase the results of federally funded research. With federal funding in decline, academic associations lose an incentive, both directly and intangibly, for membership and participation in associations. Reduced federal funding also means fewer faculty positions and fewer opportunities for scholars to be employed in think tanks and research organizations.

More broadly, the drop in federal grant funding for higher education has been accompanied by broad skepticism and even hostility toward social science research, peer review, and the effectiveness of higher education. Lawmakers have questioned the objectivity of the peer-review process for federal grants and the value of social science research for policy-making. In the case of political science, for instance, restrictions were placed on NSF funding to grants in national security or economic research, which was in the interest of the United States. These restrictions, the so-called Coburn amendment, were removed in 2014; however, skepticism in Congress toward social science research remains. Also, state lawmakers have increasingly demanded that public universities receiving state funds improve their productivity as measured by the time to graduation, student enrollments, and faculty class time. State funding of higher education has also been in decline throughout the country, albeit to varying degrees.

The decline in public funding and the overall competitiveness of the resource environment for higher education has squeezed the finances of many universities. One response by universities is to hire more adjunct and temporary faculty positions as a percentage of the total professoriate. Indeed, aspiring scholars are likely to encounter much more contingent employment prospects than earlier generations of scholars. In turn, this development has at least indirectly affected academic associations: with fewer full-time faculty positions, associations face recruitment and retention challenges for potential members.

The marked change in the political and academic context of academic associations also pushes associations and other academic infrastructure organizations to rethink their relationship to governments and the policy-making process. When government funding was rising, associations worked in partnership with government agencies; aggressive lobbying and advocacy for associational concerns was largely unnecessary. Indeed, academic associations largely avoided government relations staff, relying instead on third-party coalitions and networks to advocate their issues. Yet, the cutbacks have pressured associations to be more assertive in their advocacy and invest additional resources in government relations.

Further, associations now face new emergent policy issues related to the globalization of education. The international membership of large US-based academic associations such as APSA has been on the rise since the 1960s. It is much easier now, due to the Internet, for scholars to connect with each other around the world. Moreover, higher education has expanded throughout the world. This means that important questions related to academic freedom, such as the restrictions placed on scholars and their research in foreign countries, is arguably a more urgent concern of US-based associations. Also, advocacy on behalf of foreign scholars is much easier, due to the ability of associations and their members to rapidly disseminate information on academic freedom concerns throughout the world. For instance, the detention of a scholar in a foreign country could easily prompt an effort by academic associations to write letters to foreign embassies and post-information on the association Web site to protest the detention.

Advocacy such as trying to influence public officials and public engagement requires academic associations to be mindful of their tax status and the regulatory limitations on their ability to do lobbying. Most US associations are incorporated as 501 (c) 3 public charities; as a result, they face restrictions on their ability to do lobbying, since this is defined as direct influence on legislation by contacting lawmakers and members of congress by the federal Internal Revenue Service. Thus, academic associations incorporated under this provision of the tax law are prohibited from engaging in partisan activity and have a budget ceiling on lobbying expenses. But most academic associations rarely get close to their spending limits.

The Future of Academic Organizations in the Social Sciences

As discussed throughout this article, the more turbulent environment for academic organizations has required them to rethink their strategic priorities and relationship to members, government, and the citizenry more generally. Academic associations will need to create a different mix of incentives for their members to attract their initial enrollment and their continued participation. This effort requires associations to think differently about key aspects of their operations: annual conferences, journal publishing, member benefits, and their Web sites and social media presence. In essence, associations need to provide more value to their members in order to sustain their membership and participation and use current trends, such as the digital revolution, to their advantage. For example, the number of conferences has proliferated so greatly that each association needs to make their conference as distinctive as possible. Useful innovations in association conferences are live streaming of conference events, posting podcasts on the association Web site, extensive media coverage, greater use of social

media, and experimentation in panel and presentation format. These innovations also help ensure that the conference content lives beyond the 3 or 4 days of the actual summit. While the conference will always be a one-time event, the content itself can be made accessible to conference attendees, the wider academic and policy community, and interested citizens. Put another way, associations will need to think more strategically about their meetings, and their connection between the format and the content, as well as the association goals regarding membership and revenue. Even the location of the conference will be important, especially for smaller associations, since membership is often tied to conference attendance, and smaller associations are more dependent upon conference revenues to generate surplus revenue to support the administrative infrastructure.

Associations have also relied on royalties from their journals to support their operations. While journals were a powerful incentive for membership in the association, the movement to open access journals and the limitations on library budgets puts pressure on these types of revenues. Yet, open access journals and online journals offer associations more opportunities to create new and/or niche journals that might be of interest to members (or potential members). Again, larger associations are in a better position to provide these new publications.

Another important method to increase interaction and membership is through an active and engaging Web presence. Academic associations, like other nonprofit organizations, need an accessible and attractive Web site that provides a portal into the various benefits of membership. The Web site is also critical to networking, advocacy, public engagement, and increasingly to fund-raising. The Web site could be used for blogging, recruiting new members, and posting important news of the profession. Associations can also partner with other organizations and Web sites to link members with additional content and information.

The successful recruiting and retention of members will require academic associations to think creatively about the structure of membership benefits. The specific benefits will vary depending on the association and its own institutional history, but they should be linked to the strategic priorities of the association. Further, benefits should draw individuals into registering as members and maintaining their membership. Associations may also want to consider a tier membership structure that offers different levels of benefits depending on the amount of the membership fee, a departure from the current standard model in associations that provide all benefits for one price (although some associations have a graduated income structure).

The larger disciplinary associations have the advantage of already having a substantial membership base. The AEA, for instance, had over 18 000 members in 2013 (AEA, 2014), while the APSA had over 13 000 members (APSA, 2014). Smaller associations have less overhead, though, and in some cases they operate more informally. Regardless of the size, though, growth in membership is challenging given the current environment (see ACLS, 2014). Thus, all associations face incentives to generate nonmember dues revenue. Some associations, for example, own property and can thus generate income from tenant rents. An association may also build an endowment to support general program operations or specific initiatives or projects. Or the association may try to raise additional philanthropic support from individuals, but this task is quite difficult since most philanthropic support for academic associations is pledged for a specific purpose such as a scholarship award in the name of a prominent association member.

The successful strategic positioning of academic associations will also hinge on the concerted attention on the organizational governance structure. Importantly, academic associations have some unique features that distinguish their governance structure. Firstly, many board members are appointed due to their prominence in the discipline or subfields, rather than their particular skills (such as budget experience) or access to external resources. Secondly, membership of association boards and committees tends to be tied to disciplines, and often does not include relevant stakeholders such as foundation leaders, representatives from research institutions, or organizations representing higher education interests. Thirdly, the boards or councils of associations tend to meet relatively infrequently. Thus, it can be difficult for associations to respond quickly to emergent development, either in the organization or externally. Moreover, the small size of staff of some associations can complicate organizational governance. Considering these structural considerations, academic associations need to think very carefully about the composition of the council and committees, including the selection process, terms, and relationship to the executive staff, to ensure that the association's structure provides adequate oversight and the ability to be flexible in decision-making.

The altered context of associations places them in the spotlight as representatives of higher education and professoriates. Consequently, associations are expected to fully engage in complicated issues related to their own membership structure, as well as higher education more generally. Indeed, a host of complicated concerns now face higher education including the growing number of adjunct and contingent faculties, academic freedom for domestic and foreign scholars, decline in public funding, and rising demands for more accountability in higher education. In response, associations may invest more resources in government relations and advocacy, especially more staff resources. Nevertheless, this step is unrealistic for many small organizations that do not have substantial resources to devote to advocacy, government relations, and public engagement. As an alternative, associations can work with intermediary coalitions and other academic institutions to further their goals and priorities.

The policy and regulatory environment is complicated but it also provides ample opportunity for academic organizations to effectively advocate, if they devote resources and priority to it (also Berry, 2004; Pekkanen et al., 2014). Further, some national associations, which have focused on supporting conferences and service members, may need to be more forceful and visible in their advocacy on behalf of higher education. Increasing collaborations among the various actors in the association and coalition infrastructure is also essential, given the scarcity of new resources. Partnerships with organizations and interest groups involved in K-12 education could also be quite helpful for academic associations with a university-based membership or a focus on higher education.

Advocacy to influence public policy including grant program funding and regulations is also closely related to public engagement. In recent years, social scientists have been criticized for their disconnection from real-world problems and the inaccessibility of their research (Kristof, 2014). To address this critique, social scientists will need to do a better job in engaging with the citizenry and the policy process. This effort should include greater dissemination of academic research through social media, more focus on translating academic research for general audiences where appropriate, and establishing new outlets for social science research that go beyond traditional peer-reviewed academic journals and books. On this public engagement front, academic associations can provide key support and encouragement to individual members. Working with other partners such as academic publishers, universities, and coalitions, associations might also help to devise new venues and opportunities for greater public engagement by their members.

Given that there are multiple actors in the academic infrastructure at a time of shrinking, or stagnating, resources, it seems likely that organizational consolidation is likely to occur, especially among smaller coalitions and associations with overlapping missions. Larger research firms may also absorb smaller organizations that lack a diversified revenue base and more sustainable funding. Smaller academic organizations may also need to find a more financially stable partner (such as a university).

Conclusion

Academic associations expanded greatly in scope and diversity over the last 50 years, fueled by rising citizen demand for higher education, the increase in faculty positions and doctoral programs, and the growth of government funding. This article has discussed the changing infrastructure in which associations are operating. Yet, academic associations remain important for professional advancement. Doctoral students are socialized into the discipline by participation in meetings, and they can make valuable contacts and receive useful feedback on their research; they can also learn about job openings and the process of successfully landing a position. Young faculties with university positions may find participation in association conferences very helpful for raising their visibility within the profession. More senior faculties engaged in active research projects, especially at the elite institutions, will also continue to find conferences professionally valuable. Nonetheless, academic associations need to rethink their value proposition for memberships. In many associations, this process is already underway. Generally, associations need to restructure their governance in order to respond to their members more effectively, as well as to broaden their operations to include the wider public, educators, public and private funders, and policymakers.

Associations would also be well served by thinking of their organizations as part of a community of scholars and advocates concerned about higher education. The natural tendency of many nonprofit organizations is to focus on internal issues connected to organizational maintenance; but academic associations should be striving to connect important policy goals related to higher education, including pedagogy, academic freedom, and decent employment conditions, to their plans for sustainability. Moreover, the engagement of academic associations in these issues would also be an incentive for individual scholars to join and remain members of associations. Policy advocacy and participation in a shared community of organizations concerned with higher education would also elevate the profile of associations, help broaden the support for the organizations, and potentially increase the attractiveness of associations to donors. In the long run, associations need to actively strengthen the overall infrastructure of higher education including research enterprise and pedagogy, steps which will benefit their academic members, and greatly enhance organizational vitality and sustainability.

See also: Academic Careers in Comparative Perspective; Associations, Scholarly; Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography; Journals, Scholarly; Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations); Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Learned Societies and Professional Associations; Social Science and Universities.

Bibliography

American Council of Learned Societies (ACLS), 2014. ACLS Survey Data, unpublished.

American Economic Association (AEA), 2014. Table of Members and Subscribers. https://www.aeaweb.org/AboutAEA/table_members.php.

American Political Science Association, Report of the Executive Director, 1969–1970, PS, v. 3 (Summer): 535.

American Political Science Association (APSA), 2014. Membership Trend Data, unpublished.

Berry, Jeffrey M., 2003. A Voice for Nonprofits. Washington, DC: Brookings.

Kristof, Nicholas, 16 February 2014. Professors We Need You. The New York Times. http://www.nytimes.com/2014/02/16/opinion/sunday/kristof-professors-we-need-you.html.

Pekkanen, Robert, Smith, Steven Rathgeb, Yutaka, Tsujinaka (Eds.), 2014. Nonprofits and Advocacy: Engaging Community and Government in an Era of Retrenchment. The Johns Hopkins University Press, Baltimore.

Social Science Infrastructure: European Union (Research and Teaching)

Friedhelm Neidhardt and Georg Thurn, Wissenschaftszentrum Berlin für Sozialforschung, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 11, pp. 7511–7517, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

In recent years, social and behavioral sciences have become a standard element in higher education, in public debate and media coverage, in policy-making, administration, and corporate management. The increasing importance of social and behavioral sciences for the scientific community and beyond has affected the scope of research and teaching infrastructure considerably as well. In order to allow a better understanding of these development processes, the present article gives an overview on the most important past and current institutionalized social and behavioral science activities. In doing so, particular focus is put on cross-national research and teaching in the European social science.

Development after 1945

Modern social science, having originated in Europe, emerged as an institutionalized activity in research and teaching only after – and strongly influenced by – its powerful development in the United States. Its evolution in the twentieth century bears the imprint of the authoritarian and totalitarian regimes in continental Europe. While emigrant scholars contributed greatly to the development of the social sciences in the US, the scientific evolution in continental Europe became largely interrupted – more so in Germany and also in Mediterranean countries, and, to a lesser degree in France and Scandinavia. (The summary given in this article follows the "four phases of development in the social sciences" as delineated by Martinotti, 1999: pp. 87–91; for France, Italy, and Germany see also Wagner, 1990.)

By the end of the World War II, the social sciences hardly played any role in the European universities. For example, in sociology, there were only three chairs in France and none in Italy. In West Germany, political science chairs and a disciplinary association did not exist until 1950. To go to the US and study at an American university for sometime, therefore, became an almost necessary precondition for the further education and training of social scientists in Europe. Experiencing 'modern society' and learning the new skills of its systematic observation, made a particular impression on young scholars from countries like Germany, Italy, Spain, and Greece, which had been isolated from modern culture for decades. But it is also true of the Scandinavian countries, where the proximity of the Anglo-Saxon world favored cultural transfer.

As the postwar economic recovery progressed, it became generally understood that the social sciences were part of the modernization of Europe, seen as the expansion of the capitalist mode of production, as well as the democratization of society, both closely linked with the American model. While this caused conflicts and resistance from traditional cultural elites and the Marxist-oriented intelligentsia, there was an increasing diffusion of quantitative methods of data collection, particularly survey and primary data. While it is true to say that the European social science tradition had its own deep experience in quantitative analysis, one must also recognize that empirical research had mainly been based on the secondary analysis of 'process-produced data,' particularly statistical data collected by public authorities. What was missing in the continental tradition, and particularly in those countries that had been under despotic regimes, was the experience of field research and the collection of primary data (Martinotti, 1999: p. 89). Resistance to the introduction of these new instruments, particularly survey methods and sampling techniques, became expressed from different points of view of traditional European scholarship as well as from administrative agencies with their customary statistical procedures.

Under these circumstances, it was not before about 1960 that one can speak of a major expansion of research and acceleration of its academic institutionalization. European societies went through a period of economic 'upswing' that made large masses move to the metropolitan centers of industrial development. The postwar generation of young people, mobilized with the rapid expansion of educational systems, grew up as part of a mass culture, reinforced by the media, that provided a setting for the diffusion of ideas, lifestyles, and new organizational patterns without precedent in the European experience. The social sciences, particularly the younger and newly built-up ones, notably sociology, became highly attractive to the student generation for a number of reasons: the relatively young age of university personnel, compared with more established disciplines; the less precise definition of the academic institutionalization of the social sciences; and the proximity of their subject matter to the intellectual and existential preoccupation of the newly mobilized social groups.

From this background there emerged, in all European countries, a demand for theories of political relevance and of significance for diagnosing and overcoming societal problem constellations (see Dierkes and Wagner, 1992: pp. 612–615). This interest, with its commitment to change, stood in opposition to the program of a supposedly ahistorical and value-free explanation of social phenomena. Intellectual controversies and debates within the disciplines pitted the American social science tradition against a variety of Marxist approaches and

other 'critical' schools with their supposedly more 'enlightened' responses to the demand for a more encompassing understanding and a more 'committed' interpretation of social facts. With this orientation, major streams of the social sciences, with Paris and Frankfurt as major centers, acquired a somewhat militant role, linking up with new social movements and entering into debates about the engagement of intellectuals. The traditional social structures and actors, however, managed to retain their overall influence and to uphold traditional explanations for problems of economic and social change. Social science research in much of the 1960s and 1970s can therefore be characterized, on the one hand, by innovative work in many areas, e.g., political participation, labor organization, class structure, the 'capitalist state,' and problems of personal and sexual relations. On the other hand, theoretical innovation was limited and research perspectives remained concentrated on the nation-state – despite a growing awareness of conflicts posed by North–South differences, the role of multinational corporations, or global environmental problems.

It was in the later part of the 1970s and in the 1980s that the turmoil of debates, the imbalances between issues and approaches, the dividing lines between theoretical and empirical research gradually became superseded by an ongoing process of professionalization, of methodological consciousness, and of new theoretical directions. In retrospect, we can speak of a phase of consolidation in the social sciences in the period of economic recovery after the oil crisis. At a time when liberal social philosophies became dominant, with political sentiments tending to be antiwelfare, and economic policies being influenced by monetarism, the overriding mood seemed to run against the interest in social conditions and the study of society. While this indeed in some countries, notably the UK under Prime Minister Thatcher, made for severe cuts and structural changes, the science policy pendulum did not swing to similar extremes in continental Europe. Debates over the social sciences gradually shifted from theoretical controversies to discussions of their 'uses' in public affairs and policy-making. But by and large institutional growth continued, then slowing down to reach a steady state on advanced levels (see Martinotti, 1999: p. 90, Wagner, 1999: pp. 29, 39).

Toward the end of the 1980s, the social sciences had become well-established academic disciplines, and their presence in policy-making had been well advanced. The academic expansion had very much been related to, and benefited from, expectations with regard to practical uses of the social sciences in the process of societal modernization, which had become an important objective, actively pursued by parts of the political and economic elites. In the process, the role of intellectuals became redefined, turning from persons reflecting in general terms about the fate of humankind in history and society to experts for matters of steering society and the economy, and providing knowledge usable as a tool in reformist policy interventions. After the end of the high expectations about the 'scientification of politics' allegedly leading to the 'end of ideology,' social research had become a routine element of many policy processes and, in the late 1980s, of management in private companies as well (see Wagner, 1990: pp. 441ff). The increasing specialization of the applied social sciences, as well as the general trends of the professionalization of the disciplines and their institutionalization in universities and research centers, brought about an increasing fragmentation and specialization, which has sometimes been related to a tendency of 'deconstructivism' with critical attacks on large theoretical systems and linked to postmodernist movements. But these tendencies more often are understood as part of a more general search for theories capable of combining macrotrends with microsocial and individual dynamics, and, in so doing, aiming at overcoming the traditional juxtaposition of qualitative and quantitative methods, looking for integration in unified research designs (see Martinotti, 1999: p. 90).

Having developed to this advanced stage, the social sciences in the 1990s had to reorient themselves under two dramatic changes of focus: first, the end of the East–West opposition, and second, the process of globalization and internationalization in world regions, particularly the construction of the European entity.

There is an ever more strongly articulated need to gain a better understanding of and to contribute to problem-solving in the process of European integration. Developing a concept of European society has become a major challenge for the social sciences, which have traditionally evolved around the concept of a nation-state and of national societies under the stresses imposed by industrialization. In the last decade of the twentieth century, social scientific research the world over paid special attention to the post-socialist transition, especially in the countries of Central and Eastern Europe (CEE). Only slowly, but with increasing intensity in the first decade of the new century, has it become recognized that the tools and working conditions of social research need to be made adequate for the study of new problem constellations and supranational processes, and that social science policy-making and research programs have to be oriented toward cross-national comparison and international cooperation, particularly on the European level (see Dierkes and Biervert, 1992; Kaase et al., 1997; Erikson, 1999).

National Environments and Institutional Structures

At the beginning of the twenty-first century, it is fair to say that in the European Union (EU), the social sciences have become a standard element in higher education, in public debate and media coverage, in policy-making, administration, and corporate management. While, generally speaking, the knowledge of social reality and the diffusion of social facts have become vastly expanded in the previous five decades, the infrastructural basis for generating this knowledge has developed differently in the individual countries. The disciplinary boundaries of the social sciences, research approaches, problem orientations, and structural differentiations in the build-up of research facilities have become shaped by the respective national environments with their historical preoccupations, their academic-scientific 'cultures,' their political and administrative traditions. Expectations regarding the usefulness of social science knowledge for policy-making and the periodically strong demand for social science expertise have contributed to the institutional consolidation and at times – in varying degrees from country to country – to a push for new research structures reflecting characteristic orientations, problem constellations, and modes

of public discourse (see Wagner et al., 1991; Neidhardt and Smelser, 1992).

For example, the social sciences in Sweden may, to a considerable degree, be characterized by pointing, on the one hand, to a traditional demand for exact and reliable methods, going back to early studies of migration and social movements, and by referring, on the other hand, to the interaction with social policies in the Swedish welfare state model and thus to particular forms of contributing to problem-solving as well as, in the process, refining the theoretical outlook on society.

Contemporary social science orientations may often be explained in terms of specific historical evolutions. In the Netherlands, for example, there had been an early involvement of social science research in the practical problems associated with the territorial aspects of social planning, and the social sciences had become shaped in a continuous development by the specifically Dutch version of political conflict resolution, the *pacificatie*, and the corresponding development of institutional networks for interest mediation. While the forms of policy advisory work and research contracting in the Netherlands can thus be characterized in terms of pragmatism, tolerance, and depoliticization, it is quite a different picture that one would paint for Austria. Here, in the prewar period, the constellations of political conflict did not allow for the continuity of similar pragmatic approaches, and the modes of conflict resolution established after the war led to the involvement of social science approaches in the specifically Austrian type of corporatism, making for quite a different variant of a more general continental European phenomenon: the integration of social science research in institutional structures of political representation and in channels of communication between politics, public opinion, policy-making, and scientific expertise (see Wagner, 1990: pp. 460–463).

These examples, while meant to underline the importance of the diverse national settings, may also serve to illustrate a more general thesis explaining social science orientations with reference to historical paths of social and political development. In these terms, the social sciences emerged in Western European countries in a markedly different way from the way they evolved in England and the US, where the continuity of individualism and civil society made for a more pragmatic style of studying social and political problems empirically and with reformist perspectives. The social sciences on the European Continent, in turn, have been characterized by addressing more basically the foundations of a social and political order, which – given the traditional social conflicts, ideological cleavages, and propensities for radical change in Europe – have time and again been called into question. They have also shown a greater tendency to ascertain their basic categories and critically discuss their knowledge base.

Institutionally, in all European countries, the major part of social science research takes place within the university system. This refers first – and for most countries, foremost – to a multitude of small production units attached to professorial chairs in university departments and loosely organized faculties. Depending on the relative number of students and the weight of the professorial teaching load, university research consists mainly of small projects done with students and often linked to their theses.

At many universities we find, under similar names, institutes for social science research as organizational entities meant to provide better opportunities for cross-disciplinary orientation, for cooperating in empirical research and for contracting and administering larger projects with outside funding. But it is only at very few places that such additional forms of institutionalization, in combination with a considerable number of professorial and other staff positions in the respective departments, have led to larger and more consolidated university research centers. In Germany, such larger agglomerations of the social sciences exist at the universities of Mannheim, Cologne, Bielefeld, Berlin, and Munich (see Neidhardt, 1999).

By and large, however, it is outside the universities that more extensive research structures have become institutionalized. Extra-university institutes form a second element in the 'academic research system' in as much as they are to some degree publicly funded and not dependent in their agenda setting on short-term opportunities offered by the markets of policy analysis and corporate consulting, but rather able to integrate individual projects into larger programs. While indeed at times working on research tasks aimed at contributing to problem-solving, such establishments usually understand themselves as knowledge-oriented and 'science-driven,' and are – at different levels – connected to the universities (see Wagner, 1990: p. 441ff).

These institutes vary from country to country in number, size, and concept, depending on the respective national environment. In some cases, there is only one such extra-university institute as a central establishment, like the Greek National Centre for Social Science Research in Athens. The German case demonstrates a functional differentiation in two directions, which, for structural reasons, did not easily develop within the universities and thus called for extra-university institutionalization: on the one hand, the *Wissenschaftszentrum Berlin für Sozialforschung* (*WZB* – Social Science Research Centre, Berlin) and some institutes of the *Max-Planck-Gesellschaft* and of the *Leibniz-Gemeinschaft* (*WGL*), which undertake larger, long-term, internationally comparative, or problem-oriented projects in transdisciplinary research groups; on the other hand *GESIS-Leibniz-Institute for the Social Sciences*, a group of social science infrastructure establishments in Mannheim, Cologne, and Bonn, which provide professional services in support of social science research by documenting research projects and their results, by advising researchers on project design, by offering courses in advanced methodologies, and by archiving and giving access to empirical data for the purpose of secondary analysis.

Compared with the grand total of resources and activities in the universities, the 'extra-university' sector of the academic research system is still rather small in Germany and even smaller in Italy with only a few institutions of the *Consiglio Nazionale delle Ricerche* (*CNR* – National Research Council). In France, however, this sector is considerably larger, due to the historic separation of the universities from the centers of academic research. The dominant institution is the *Centre National de la Recherche Scientifique* (*CNRS*), which runs a large number of research institutes separately from or in cooperation with universities or *grandes écoles*; for sociology and ethnology, the *CNRS* with its division on the *Sciences de l'Homme et de la Societe* is far more important as a research institution than the

universities. Furthermore, there are the *Maison des Sciences de l'Homme* (*MSH*), the *Ecole des Hautes Etudes en Sciences Sociales* (*EHESS*), the *Fondation Nationale des Sciences Politiques*, the social science sections of the Paris Museums, and the *Collège de France* (see Machin in Lisle et al., 1984).

The basic institutional setup of the academic research system is – in varying forms of composition – being supported and complemented by a third institutional element: the system of project funding through national organizations and foundations. To take the German example again, the *Deutsche Forschungsgemeinschaft* (*DFG*), a publicly funded association of institutions of higher education and research, provides good opportunities for good projects – in whatever kind of research design – to be financed, supplementing the existing institutional resources and using peer review procedures to evaluate project applications. Moreover, the *DFG* and, with greater emphasis, other funding organizations (like the *Volkswagen Stiftung*) make attempts to establish researcher networks and cooperative arrangements within larger, more integrated research programs.

In addition to these institutes and programs that are constituted on scientific grounds with a primary interest in knowledge generation, we find in all the Western European countries other types of social science institutions that are – again in different forms and in varying degrees of development – oriented toward practical needs of government agencies and private organizations. These are, first, publicly financed research institutions directly operated by the state and subject to directives from the respective ministry interested in having a particular need covered by research and analysis. As a second type there are privately organized institutes, which may have some basic institutional budget, but receive the main part of their resources from contracts for work on projects commissioned by government agencies as well as corporations, unions, the media, etc. The 'state-administrative' part of this sector, often with longer traditions, comprises statistical and agronomic analyses, research on land use and regional development, on demography, public health, on education and vocational training, etc., usually with a strong empirical quantitative orientation. Such 'contractual' research became more developed in the 1960s and 1970s, relating to the policy issues of the day, and sometimes developing a marked expertise through repeated studies over time in a given area.

In some nations and at certain periods in their development, this sector may gain considerable – even dominating – weight in the overall spectrum of social science research (see Wagner, 1990: pp. 443–447). In the context of this article, however, there is no room for elaborating, e.g., on the importance of the government-directed nonacademic research sector in France, on the effects of administrative traditions and statistical-quantitative orientations, or on the channels for linking social science research to the system of 'planification' through, for example, the *Comite d'Organisation des Recherches Appliquées sur le Développement Economique et Social* (*CORDES*).

These different forms of applied research are indeed part of the institutional environment and the operating modes of social science research in the respective national contexts. However, while they provide career paths for social scientists and demonstrate the public use of social science knowledge, their infrastructural functions remain rather restricted in comparison with the academic research system with its opportunities for research experiences in larger contexts, for methodological training, for internationally comparative ventures, and for exchanges between approaches and orientations.

Strengthening the European Research Base

Social science research in Europe, in summary, has become institutionalized in a variety of national settings and organizational structures, depending on each country's specific historical background and cultural contexts. International exchange and cooperation, first with the United States, then increasingly across Europe, have served to develop the social sciences to higher degrees of professionalization, with a growing interest in cross-national comparison, but have by far not yet come near to what might be called a 'European' social science. At the beginning of the twenty-first century, the ongoing process of European integration had still not become adequately developed as a subject of research aiming at understanding the preconditions, the conduct, and the consequences of this process. In addition, the EU itself, its institutions and major policy areas require the increasing attention of the economic and social sciences, and it has become a matter of 'social science policy' concern to evaluate and redirect the structural conditions of social science research in the European countries and the perspectives of its development 'toward Europe.'

At the outset of such assessments, it has widely become recognized that the fragmentation of the European systems – in terms of social, political, and economic problems as well as science and research structures – should not only be understood as obstacle and retardation on the path to a truly European outlook. Turning disadvantage to benefit, one can see European social sciences, with their subject matter differing to some extent from country to country, as providing a densely developed empirical basis for generalizations across time and space. With its richness of constitutional, institutional, and cultural variations across (and also within) nation-states, Europe has come to be interpreted as a 'natural laboratory' for the social sciences, leading to strong recommendations concerning systematic empirical comparative research (see Kaase et al., 1997: pp. 3–4).

While the social science infrastructures in the different European countries have so far been mainly directed toward the study of national developments and problem-solving dimensions, the past decades, with still limited success, have put forward the notion of an increasing need for international cooperation and the corresponding infrastructures that make for exchanges, collaborative work, and joint projects. Network linkage has been advanced through the establishment of European journals in different disciplines, through a 'Network Scheme' by the *European Science Foundation* (*ESF*) and through European networks of research units. The most successful of these has been the European *Consortium of Political Research* (*ECPR*) with currently around 400 European institutional members and associate members in over 50 countries, from as far afield as New Zealand and Japan. The *ECPR* organizes annual joint sessions of workshops, which take place in

different countries and bring together about 500 European political scientists. A similar organization, the *European Consortium for Sociological Research*, exists for the neighboring discipline.

Meeting grounds for groups of researchers from different countries and opportunities for collaboration on joint projects over a period of time are being offered by institutes of advanced study, particularly the *Netherlands Institute for Advanced Study in the Humanities and Social Sciences* in Wassenaar and the *Swedish Collegium for Advanced Study in the Social*.

Sciences in Uppsala as well as – with a more general orientation reaching beyond the social sciences – the *Wissenschaftskolleg zu Berlin* or the *Central European University's Institute for Advanced Study (CEU IAS) (earlier under the umbrella of the Collegium Budapest, Institute for Advanced Study, whose operation has to be terminated in 2011 lacking of sufficient funding*). The latter, together with other centers of excellence in CEE countries, is meant to advance research capacities within its region by, not least, providing links with researchers from Western countries.

In the Framework Programmes (FP) of the EU the social sciences have not, until the 1990s, played an explicit role other than having been eligible for the funding of networks of younger researchers as part of general 'Human Capital and Mobility' or 'Training and Mobility of Researchers' programs. The Fourth FP (1994–98) did for the first time contains a separate program relating to the social sciences, 'Targeted Socio-Economic Research.' One part aimed at the evaluation of science and technology policy options in Europe. It included a part on research on education and training and one on research into social integration and social exclusion in Europe. The latter addressed processes of social exclusion and integration, causes of social exclusion, particularly unemployment, the impact of social integration policies, and research infrastructures. (For the EU research programs see http://cordis.europa.eu/home_en.html.)

The EU Fifth FP (1999–2003) contains, as a major instrument, 'Key Action: Improving the Socio-Economic Knowledge Base,' directed at research tasks on the management of societal change, on individual and collective strategies in a changing society, on employment and unemployment and on social cohesion in Europe, on new perspectives for learning, on governance, citizenship, and the dynamics of European integration, and on the challenge of EU enlargement. Under the Sixth FP (2002–06), this type of research found a place in the 'Citizens and Governance in a Knowledge-based Society' Programme. It mainly focused on both improving the generation, distribution, and use of knowledge and its impact on economic and social development, and identifying implications of European integration and enlargement for governance and the citizen. In the current programme, 'Socio-economic sciences and the humanities' (SSH), which is part of the European Research FP7 (2007–13), emphasis is given to eight areas, all of them aiming to study and offer answers to questions regarding the demographic change and quality of life; education and employment in view of the current economic trends; global interdependence and the transfer of knowledge; the well-being of democracies and political participation; cultural diversity and the values. (More details to the SSH as part of the FP7 see at http://cordis.europa.eu/fp7/ssh/about-ssh_en.html.)

In addition, the interest in social science research has become expressed in an ongoing increase of the budget available for social science. As a case in point, whereas the above mentioned SSH had a budget of € 610 million available of a total budget of around € 50 500 million (12%), a budget of € 29 700 million, 37% of the total € 80 000 million, is currently foreseen for research in societal challenges under the EU program Horizon 2020. This is an increase of nearly 25% in real terms compared with its predecessor, the FP7.

For the researchers, participation in the European research programs has meant, depending on the background of their experiences in the respective national systems of research funding, a reorientation toward new requirements and procedures, sometimes deviating considerably from established modes of competitive research application and evaluation. Following critical discussions and corresponding adjustments of program definition, of evaluation criteria and financing mechanisms, criticism, partly even opposition, centered on the bureaucratic procedures, on the professional status and the scientific excellence brought to bear on program design and project assessment, and on the policy orientation of EU research. (For a general discussion of EU research structures and policies see Max-Planck-Gesellschaft, 1994.) It remains to be seen whether the social science research task in the seventh and in subsequent EU FP such as the Horizon 2020 will reflect a convergence between competing principles and complementarity between the research programs of the EU, on the one hand, and the ESF with the different national research councils and scientific organizations, on the other.

One item of discussion that may lead to a clarification of perspectives and differentiation of roles is the desirable balance between basic and applied research in response to the 'European' dimensions of EU research tasks and their relatedness to EU policy areas. (As a position in the discussion, calling for more basic research, see Erikson, 1999: p. 32.) One example is the expression of the overall objective of the program, 'Improving the Socio-economic Knowledge Base' Key Action, namely:

> to make best use not only of the solid base of social science research that Europe has at its disposal, but also to harvest Europe's immense cultural diversity. It aims to improve our understanding of the structural changes that are taking place in European society, so that we can manage and direct the changes and involve Europe's citizens more in shaping their own futures. The Key Action will allow the results of various different types of work – empirical, comparative and prospective research – to be brought together within a cohesive framework. It will therefore be in a position to provide policy-makers with a sound knowledge of the challenges facing Europe, and of the policy options that are open to tackle them. It will focus on four areas.
>
> European Commission Publications; see http://ec.europa.eu/research/improving/leaflets/en/keyact01.html

Under this general headline, there has been a pronounced emphasis on four areas: (1) societal trends and structural changes, (2) technology, society, and employment, (3) governance and citizenship, and (4) new development models. The first will aim to clarify structural changes and other complex interactions between societal trends, changing lifestyles, economic evolution, labor market institutions, and cultural

patterns. The second will investigate the way different types of institutions – businesses, schools and universities, local authorities, governments, and consumers – interact to invent, produce, and market innovative goods and services. In addition, the area of the governance and citizenship aims to research the role that the various levels of governance (local, regional, national, and EU) play in Europe, and look at ways in which their interaction and accountability can be improved. Finally, the last area of the Key Action, the new development models, will investigate the dynamics of wealth creation and distribution in a global economy where intangible factors predominate. It will explore new sustainable development models that will square the circle by simultaneously fostering growth, job creation, environmental protection, equal opportunities, the reduction of inequalities, and improvement of the quality of life. (European Commission Publications; see http://ec.europa.eu/research/improving/leaflets/en/keyact01.html.)

The interest in 'Data integration' and 'research infrastructure' has become a common effort in other policy measures as well. It suffices to mention Eurostat, the statistical office of the EU. Even though it was originally established to meet the requirements of the Coal and Steel Community, over the years it has increasingly included social sciences in EU actions to support 'large-scale facilities' and by promoting 'access to infrastructures,' widening the notion of infrastructures from establishments for data processing and archiving to functions of opening up particular research experiences by coordinated international exchanges. Currently, its task has broadened to collect data for Economic and Monetary Union and developing statistical systems in candidate countries for EU membership so they can define, implement, and analyze Community policies more effectively (see Changing Role of Eurostat at their Web sites www.eurostat.ec.europa.eu).

This interest meets with a core concern of the ESF, particularly its Standing Committee for the Social Sciences, which has been active, albeit with limited resources, not only in promoting thematic projects in international cooperation, but also in the discussion and recommendation of social science policies and infrastructural development. In this context, it has prepared the proposal for a 'European Social Survey' which in a number of respect overlaps with, or runs parallel to projects partly financed by the European Commission, to develop a 'European System of Social Reporting and Welfare Measurement,' with about 20 national teams, including some Eastern European countries.

With the efforts to create a 'European System of Social Indicators,' emphasizing the scientific context and research dimensions of social reporting going beyond 'mere statistics,' an early interest of the European Community of the 1970s is being taken up again, very much in line with the corresponding activities of the Organization for Economic Co-operation and development (OECD), following the strong signal given by the so-called Brooks report on 'Science, Growth, and Society' of 1971 with its explicit inclusion of social science research in science policy concepts. In a report of 1997 for the *European Science and Technology Assembly*, an advisory body to the EU, a panel of leading social scientists completed a survey of the strengths, gaps, and weaknesses of the economic and social sciences in Europe. In this, they recommended measures for the support and advancement of interdisciplinarity for problem-driven research, particularly flexible organizational modes to link scholars from different, even distant, disciplines for a limited period of time; the comparability of data, especially longitudinal microdata, and the systematic linkages necessary for multilevel analyses; the research training of young scientists through graduate colleges, summer schools, etc.; incentives for independent work at earlier stages in academic careers; networks between funding organizations for granting money to comparative projects; national concentrations of methodological resources (similar to *ZUMA, the Zentrum für Umfragen, Methoden und Analysen*, in Mannheim) and links between them across Europe (see Kaase et al., 1997: pp. 4–10). The need for developing a common European research and training agenda was also pronounced by Federico Mayor, the former Director General of UNESCO, following a series of conferences concentrated on European perspectives on the social sciences (see Mayor, 1998).

These efforts from both the UNESCO and EU, together with the process of EU integration and EU enlargement, inter alia, have obviously stimulated the development of social research activities. A growing emphasis has been put on empirical data collection and analysis in recent years (see, e.g., European Commission, EU social indicators) aiming to monitor and systematically describe, inter alia, the current state of and changes in living conditions and the quality of life – not only at the supranational, but also at national and subnational level in all old and new Member States (see Noll, 2002: p. 2). Such developments can be regarded as fruitful in a wider sense: not only do they help to comparatively understand social, political, and economic structures in national contexts; but they also call for studies of intersocietal penetration and of arrangements that work toward international cooperation and integration between the EU members.

In addition to the increasing importance of cross-national research activities in the European social science, the teaching infrastructure in the EU has been also continuously extended at an international level. As a case in point, the European Commission has been increasingly put on emphasis on financial support of actions in the fields of Education (Erasmus, Leonardo da Vinci, Comenius, and Grundtvig), international Training (Erasmus Mundus, Tempus, Alfa, Edulink, the programme for cooperation with industrialized countries), Youth (the Youth in Action programme), and Sport (the new sport action). Finally, the latest EU Programme, Erasmus+ (2014–20), aims to bring all these seven programmes together. The programme gives opportunities to students, trainees, staff, and volunteers to spend a period abroad to increase their skills and employability. It supports organizations to work in transnational partnership and to share innovative practices in the fields of education, training, and youth. By increasing mobility between the EU and the Third Countries, in accordance with EU external policy objectives, it also contributes to the development of human resources and the international cooperation capacity of Higher education institutions in these countries (see European Commission 'Erasmus + 2014–20 Programme Guide').

In general, it can be stated that the EU has strengthened to some extent the role of social science within its borders. A pluralist and transnational social science in Europe should, however, keep looking for closer relations with social sciences in Asia, Africa, and Latin America, in addition to its profound ties with North America, in order to address and bring forward the need to link to global-wide concerns.

See also: Discipline Formation in the Social Sciences; Infrastructure Investment; Quantification in the History of the Social Sciences; Social Science Infrastructure: East Asia and the Pacific (Research and Teaching); Social Science Infrastructure: European Union (Resourcing and Funding); Social Science Infrastructure: North America (Research and Teaching); Social Science Infrastructure: North America (Resourcing and Funding).

Bibliography

Dierkes, M., Biervert, B. (Eds.), 1992. European Social Science in Transition: Assessment and Outlook. Campus/Westview, Frankfurt/Boulder, CO.

Dierkes, M., Wagner, P., 1992. Conclusions and recommendations. In: Dierkes, M., Biervert, B. (Eds.), European Social Science in Transition: Assessment and Outlook. Campus/Westview, Frankfurt/Boulder, CO, pp. 611–637.

Erikson, R., 1999. The prospects of European social science. In: The Social Sciences at a Turning Point? OECD Proceedings, Paris, pp. 31–35.

European Commission, 2014. EU Social Indicators. http://ec.europa.eu/social/main.jsp?catId=756&langId=en.

European Commission, 2013. Horizon 2020 – the EU's New Research and Innovation Programme. Website: http://europa.eu/rapid/press-release_MEMO-13-1085_en.htm.

Kaase, M., Baumert, J., Erikson, R., Hellwig, M., Martinotti, G., Meidinger, C., Menard, C., Newton, K., Semin, G., 1997. Survey of Strengths, Gaps and Weaknesses in European Science: Economic and Social Sciences. Report for the European Science and Technology Assembly (ESTA) and for the European Heads of Research Councils (EUROHORCs). Deutsche Forschungsgemeinschaft, Bonn, Germany.

Lisle, E., Machin, H., Yasin, S., 1984. Traversing the Crisis: The Social Sciences in Britain and France. Economic and Social Research Council, London.

Martinotti, G., 1999. The recovery of Western European social sciences since 1945. In: UNESCO World Social Science Report 1999. UNESCO Publishing/Elsevier, Paris, pp. 84–91.

Max-Planck-Gesellschaft, 1994. European Research Structures—Changes and Challenges: International Aspects of European Research Policy. Max-Planck-Gesellschaft, München.

Max-Planck-Gesellschaft, 2013. EU Research Funding at a Glance. Max-Planck-Gesellschaft, München. Website: http://eu.tuebingen.mpg.de/eu-research-funding-at-a-glance.html.

Mayor, F., 1998. European perspectives on the social sciences. In: Proceedings of the Second European Social Science Conference in Bratislava. UNESCO, Paris.

Neidhardt, F., 1999. Social science organization and policy issues in Germany. In: The Social Sciences at a Turning Point? OECD Proceedings, Paris, pp. 59–62.

Neidhardt, F., Smelser, N.J., 1992. Sociology. In: Dierkes, M., Biervert, B. (Eds.), European Social Science in Transition: Assessment and Outlook. Campus/Westview, Frankfurt/Boulder, CO, pp. 244–267.

Noll, Heinz-Herbert, 2002. Towards a European system of social indicators: theoretical framework and system architecture. In: Social Indicators Research, Special Issue, vol. 58, pp. 47–84, 1–3.

OECD, 1999. The Social Sciences at a Turning Point? OECD Proceedings, Paris.

UNESCO, 1999. World Social Science Report. UNESCO, Publishing and Elsevier, Paris.

Wagner, P., 1990. Sozialwissenschaften und Staat: Frankreich, Italien, Deutschland 1870–1980. Campus, Frankfurt, Germany.

Wagner, P., 1999. The twentieth century—the century of the social sciences?. In: UNESCO World Social Science Report 1999. UNESCO Publishing and Elsevier, Paris, pp. 16–41.

Wagner, P., Weiss, C.H., Wittrock, B., Wollmann, H. (Eds.), 1991. Social Sciences and Modern States: National Experiences and Theoretical Crossroads. Cambridge University Press, Cambridge, UK.

Social Science Learned Societies and Professional Associations

Jennifer Platt, University of Sussex, Falmer, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The historical emergence of social science associations, as part of the academic institutionalization of the social sciences, is sketched, and the membership and typical functions of learned societies – in particular their conferences, journals, and research sections – are described. The pattern of affiliations between disciplinary and cross-disciplinary, and national and international, associations is outlined, and it is suggested that these are sufficiently structured and comprehensive to have created what is in effect a world system.

What do we count as a social science association? In this context, it must be an organized body which figures in academic social science. However, the history of such bodies shows that some formal organizations have started much less formally, that the boundaries between academic and nonacademic have changed over time, that 'social science' may be more concerned with practical reform and social interventions than with social theory, and that some bodies straddle boundaries between social and physical science, or between social science and history or literature. The situation must be drawn more broadly to show the full picture.

The range of relevant associations has expanded enormously, starting with their development in the second half of the nineteenth century; before then, the main predecessors were on the one hand 'academies', usually autonomous national bodies representing wide fields of science and learning and with a socially elite membership, or informal 'invisible colleges', or on the other hand bodies concerned with social reform, such as the American Social Science Association (founded 1865). (Alter (1980) describes the emergence of an International Association of Academies, a German scientific initiative, and shows how what presented itself as learned collaboration transcending national boundaries was inevitably entangled with major political and economic changes and conceptions of the national interest.) There are different types of association, their difference traditionally labeled in English as that between the 'learned society', which concentrates on the advancement of the intellectual content of its sphere, and the 'professional association', largely concerned with the more material occupational interests of its members. The 'qualifying association', whose functions include the award and policing of the conditions of a qualification necessary to practitioners, could be seen as bridging that gap. Existing associations are certainly not all pure examples of these types, which should be treated as abstract concepts rather than empirical descriptions. This article focuses on the learned society bodies and functions.

Functions

Learned societies perform a wide range of functions. These commonly include the holding of conferences at which papers are given, the production of academic journals, and the representation of their constituency's knowledge and interests to the educational system, government, and the wider public; they may also include the award of prizes, the provision of training activities, the sponsorship of research, the production of teaching materials, the creation of sections for members working in specialized topic areas, and the establishment of ethical codes to guide professional behavior. Described in more theorized terms, these functions include establishment of the legitimacy of their field's knowledge claims, knowledge transfer, the maintenance of intellectual standards, disciplinary socialization and social control, and the encouragement of intellectual cooperation and social integration among their members. There is considerable variation among existing associations, depending on their size, history, and local traditions, but with a tendency for the same patterns of activity to diffuse across the field, even if similar functions are performed in different institutional ways. The larger associations have the resources to perform multiple functions.

Membership

Association members may be individuals or groups. Individual membership has more often in the past been confined to elite categories such as university professors, or holders of doctorates, and is still sometimes honorific and by invitation only, but in recent times the admission on application of at least other members of university faculty and of students has become normal. Where group membership exists it is commonly defined by discipline in cross-disciplinary national associations, and by national disciplinary association in international associations; the latter has been the standard pattern within the United Nations Educational, Scientific and Cultural Organization (UNESCO) sphere of institution-building (see later), though as the numbers of social scientists have grown the importance of individual membership alongside national representation has expanded considerably. Within this basic framework, room has often been found for cases which do not precisely correspond to the original constitution's concept. Thus, there is sometimes special provision for social scientists working outside academia, or for representation of teachers of a discipline in schools, whether within the structure of the national association or by the affiliation to it of a specialist association. Social scientists from the academic world's periphery (however, that is defined) often find reason to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03006-3

establish some affiliation with its perceived center, promoting social integration of the disciplines.

The System of Associations

An important contribution was made to the emergent pattern of associations by the policy of UNESCO which, following on from the pre-war League of Nations Committee on Intellectual Cooperation, in 1948 started to develop a plan to bring national bodies together and promote internationalism in the interests of world peace; it was hoped that cross-disciplinary and cross-national research would support that. Sometimes an earlier international association already existed that could be used as a basis; the formation of national associations to fill gaps, and affiliation to international bodies by those that already existed, was actively encouraged. (In the case of sociology, however, leading members of the Institut International de Sociologie (IIS) were felt to have been sufficiently close to fascist regimes for that to be unacceptable, so the foundation of the International Sociological Association (ISA) was set in motion (Platt, 1998: pp. 27–28).) An organizational hierarchy has thus in effect been created by UNESCO, making use of existing developments, in which national disciplinary associations are gathered together into international disciplinary associations, and those in turn are the members of cross-disciplinary associations such as the International Social Science Council (ISSC) and the International Council for Science (ICSU). (The latter retains the acronym from when it was known as the International Council of Scientific Unions.) These bodies have not exercised authority over their members, though their access to funds sometimes gives them influence, so they are at the top of the UNESCO hierarchy only in the sense that in principle they include all relevant associations, and their remit is at the highest level of geographical and intellectual generality. This hierarchy ignores the many independent special-interest smaller associations for subfields such as medical anthropology, history of economic thought, biographical methods, or figurational sociology. These are important, but in their intellectual activities very similar to sections, sometimes moving between or combining section and independent status, so are not discussed separately here. In this article, we start at the bottom of the hierarchy.

National Disciplinary Associations

The process of formation of national disciplinary associations is part of the establishment of a discipline as such, and of its professionalization. (The members of a national association are not necessarily citizens of that nation or working in it. For instance, in Britain 60% of individual members of the Royal Economic Society are recorded as living outside the UK; that is probably an extreme case (http://www.res.org.uk/view/index.html).) In many countries, there were no such associations by the 1950s – indeed little or no academic social science, though there might be more or less formal groups of intellectual notables, and what is now recognized as social science in contexts not given that name. National disciplinary associations are voluntary associations, not state bodies; they normally have elected officers serving fixed terms, often supported by an office staff, although much essential work is also done by post-holders without payment. In those countries which had developed some academic social science by then, mostly in Europe and North America, the process leading up to the modern situation started around the mid-nineteenth century. When the push to create national members for international associations came in the 1950s, many more emerged. Henderson (1996) gives detailed accounts of the emergence of German and British national economic associations, while Fourcade (2009) places French, British, and US economics associations in broader national contexts; Celebi (2002), Glatzer (1994), Patel (2002), and Platt (2003) deal with sociology in Turkey, Germany, India, and Britain. These examples are sufficient to indicate the diversity of the routes that may be followed, and the extent to which they may depend on current political concerns and the existing structure of associations – but still reach similar goals. Such bodies surely offer the commonest form of social scientific associational participation.

In countries with relatively small numbers of social scientists, whether as an effect of small total population size or the early stage of disciplinary development, ways have been found of creating viable associations based on larger constituencies. Thus in Canada for some time there was one national association for both anthropology and sociology, while in the antipodes there was a joint Australia/New Zealand sociological association (Crothers, 2005: pp. 73–75); both have now divided. In addition to such bodies, both the British Association (whose section on anthropology, for instance, started in 1884) and the American Association for the Advancement of Science were bodies, facing toward a broader general public, which covered a broad range of disciplines, including social sciences before there were single-discipline national associations. Larger regional associations such as CODESRIA (Council for the Development of Social Science Research in Africa) have been accepted by the international bodies into the same category as national members, and can offer more locally shared solidarity and frameworks of reference, as well as some international perspective, at lower cost. Cross-disciplinary national associations are less common; their appearance may be the start of the emergence of a new discipline. But when government policy affects broad areas of academia there have been cases, such as the USA's Consortium of Social Science Associations (COSSA), where associations have united in a new national body to defend their shared interests. Similarly, groups such as Sociologists for Women in Society cut across disciplinary boundaries as they address both the position of women in academia and the need for gender to be taken into account in its intellectual products.

There are also cases where national subdivisions, such as the English/French language cleavage in Canada, are important enough to lead to separate associations. Patel (2002) sees the Indian academic community as sharply hierarchical, with an important division between universities teaching in English and in local languages; the national association is mainly for the former group, while the latter have established a number of local associations. The USA used to have a Catholic association, now defunct, and the originally oppositional Society for the Study of Social Problems continues (while the American Sociological Association remains clearly dominant), but the country also has thriving regional groups such as the Southern

Sociological Association (Simpson, 1988: pp. 1–17) which provide contacts closer to home geographically.

International Disciplinary Associations

The main international disciplinary associations are in principle world-wide, although in practice the uneven diffusion of social science sets limits to their constituencies. Local differences in intellectual patterns contribute to the observed differences in the dates at which national associations have affiliated to the international associations of different disciplines, but each now has a very large number of national affiliates. (The total number of collective members available is subject to political events; thus the dissolution of the USSR increased the number markedly, while the reunification of Germany and division of Czechoslovakia halved and doubled their representation.) The collective members of these associations include some regional associations for Europe, Africa, and the Asia–Pacific area.

There have sometimes been frictions over the patterns of representation. The IIS for a while competed with the ISA to be regarded as the world association, organizing parallel world congresses, but eventually peace was made and IIS became formally affiliated to ISA. Although the headquarters of UNESCO were in Paris, and French was one of the ISA's official languages, US hegemony made English more salient; when the proposal of a French-speaking ISA section was rejected a separate francophone association, the Association Internationale des Sociologues de Langue Française was founded in 1958. That continues successfully to this day, but has also long made peace with ISA and become affiliated. A similar accommodation was reached between the International Union of Psychological Science (IUPsyS) (which also has a francophone affiliate) and the earlier International Association of Applied Psychology (IAAP); IAAP is now affiliated to IUPsyS, and their conferences alternate. By 2000 IUPsyS also had nine other affiliates, some of them regional groups (Rosenzweig et al., 2000: pp. 2–3).

One of the major problems that the international associations face is how to get sufficient funding to support a desirable range of activities; membership subscriptions are paid, but to make the essential membership from the global South (and students) practicable charges to them have to be low, and there may not be much practical scope for subsidising them from the North. Most associations may draw on a considerable input of voluntary labor, but that has its limits, and is particularly poorly suited to running a central office. The modern solution is to budget for profits from conferences and journals. The possible effects of the 'open access' movement on journal profits will only be seen in the longer term.

Sections

Disciplinary associations both national and international usually have 'sections' defined by their specialized research interests; how many are generated is influenced by the size of the membership, but there appear to be both some generally shared concerns, some of them responding to wide social movements such as feminism (Platt, 2010: pp. 109–110), and specifically national ones such as the Indian Sociological Association's 'Dalits and Backward Classes' or the Deutsche Gesellschaft für Soziologie's 'Ost- und Ostmittel-europa-Soziologie' (the latter now renamed 'Europasoziologie') which have arisen from local situations. (Some in practice are not confined to members of the discipline; lawyers appear in sections on deviance, clergy on religion, nurses on health and medicine.) Their size may vary considerably, which affects their programs, but a minimum level of activity is to organize meetings at world congresses and to circulate a newsletter; some will have more meetings, take part in joint and comparative research, and publish their own journal. For many individual members, sections have been their main locus of identification within the wider discipline, and the intellectual history of disciplines can in important ways be plotted through study of these groups.

Conferences

Organized international meetings of people with shared interests have commonly preceded the creation of international associations. For instance, international congresses of statistics were held regularly from 1853 to 1876 (Brian, 1989: p. 201); various Congresses of Anthropological Sciences were held before 1865 (International Union of Anthropological and Ethnological Sciences (IUAES) web site); the first of a series of geographical congresses was held in Antwerp in 1871 (International Geographical Union (IGU) website); and a one-off political science congress was held in Paris in 1900 (Coakley and Trent, 2000: p. 11). Thus some conferences led to the creation of associations, rather than being initially organized by them.

The organization of conferences is now a major and regular associational activity, and they have grown in size and expanded in the range of activities they offer. (Tourist offices send representatives to committee meetings to advertise their country's attractions, and firms exist whose business is to organize conferences for associations.) Plenary sessions are held, but the typical timetable of large conferences now provides many points at which choices between speakers, often classified by section, have to be made; special activities for students, or for tuition in the latest developments for faculty, are arranged; large publishers' exhibitions are included, there may be formal space for the conduct of job interviews, and a variety of committee meetings, social scientific tours, and social events are held. There is the opportunity to see the great in the flesh, get comments on your draft paper, develop cooperative work, meet old friends and new ideas, buy cut-price books, learn more about the country where the meeting is held, and maybe add an exotic holiday. Conferences may also serve broader social goals: they were important in the diffusion of early second-wave feminism, and have provided moral support and intellectual opportunities to members from countries where social science is restricted or discouraged.

Meanwhile, it is a recognized problem that many academics can get high costs of conference attendance paid by their department, but only if they are giving a paper; this leads to pressure to accept a large number of papers, so that the timetable becomes excessively crowded and running expenses rise. (Canada has an unusual solution to problems of distance: every year in May/June, a large number of learned societies in the humanities and social sciences hold their annual

conference in the same place and over the same period, maximizing contact possibilities and minimizing travel costs.)

Journals

A major contribution to the literature is made by the journals which associations produce. Crothers (2007) found that 51% of sociological journals listed in the ISI's Journal Citation Reports for 2006 were sponsored by associations; figures for other disciplines are not available, but may well be comparable. Many of these are run by commercial publishers on their behalf (Singleton, 1981), though the associations maintain control over decisions about intellectual content. The possession of an official journal is seen as important to the definition of associational identity (Hudson and Hudson, 2005); national associations normally have one, and may have several, while specialist networks have specialist journals. Long ago now, Lengyel (1967) reviewed the pattern of social science periodicals in the world. He found that different countries had different modes of journal sponsorship, depending on the structure of local institutions; correspondingly, the growth of international journals was related to the rise of international organizations. It is commonly assumed that the national association journals, in particular those with a nonspecialist remit, are at the top of the hierarchy of local intellectual prestige, and can be taken as representing the discipline. Whitley (1970: p. 255) suggests that "The study of journal operation is the study of the application of discipline norms to the judgment of research", and Vanderstraeten (2010: p. 561) sees journals as revealing "the collective character of the scientific practice". Coats' (1971: p. 32) finding that over the period 1886–1959 five general economics journals became increasingly similar in their contents could support this. In that sense, journals and learned societies are inextricably involved with each other as parts of the system of research and disciplinary reproduction.

International Cross-Disciplinary Associations

UNESCO has cross-disciplinary associations for natural science (ICSU), social science (ISSC), and the humanities (ICPHS) whose standard members are international disciplinary associations. Not every existing organization fits neatly into the preconceived categories of the system, but ways have usually been found to incorporate them. The question of where an association belongs has sometimes been answered by its having a dual affiliation; thus both IUPsyS and the International Geographical Union (IGU) are members of both ISSC and ICSU, and the IUAES also belongs to UNESCO's International Council for Philosophy and Humanistic Studies. Some associations involved in the system are not purely or mainly academic. Thus the World Federation for Mental Health, which belongs to the ISSC, has been a movement to promote mental health as much as to study it. Brody (1998: p. 165) notes that since it was not, like other affiliates, composed of academic researchers, it was less easy for it to follow ISSC initiatives up; although it took part in work such as that on human rights and new bio-technologies, it remained an association whose more significant contacts were elsewhere. Somewhat analogous is the International Peace Research Association, founded in 1964. Its main aim is to increase and improve research focused on world peace. Most of its members are not national or regional bodies, but individuals or research institutes. It has worked on topics such as international human rights, and gender and peace, which relate to UNESCO and ISSC agendas, but has not been prominent in ISSC affairs. Thus disciplinary patriotism has not prevented the growth of complex links, perhaps occasionally open to interpretation as disciplinary imperialism. Many ISSC activities have been on a project basis, and it has promoted research on specific topics seen as relevant to its international and cross-disciplinary remit; in recent years there have been major programs of work on poverty, global environmental issues, gender, and disaster risk. It also has responsibility for the *World Social Science Report*, and the biennial World Social Science Forum, a major conference on a topical theme related to ISSC interests.

The ISSC has changed considerably over time, in ways too complex to describe here (but see Platt, 2002). These changes reflect the intersection of the internal politics and policy goals of UNESCO, the funding sources available at different times, the agendas of some charismatic and interested individual social scientists, and the concerns of the growing set of associations involved. Its aim remains to advance the social sciences while attempting to solve global social problems.

Conclusion

The complicated pattern of associations that has developed uses a number of dimensions of classification:

National/regional/international
Disciplinary/cross-disciplinary/sub-disciplinary/topic-specific
Individual membership/group membership
Open individual membership/qualifications for admission/elite selection

These appear in different combinations. It is evident that complex needs are met, and there is considerable practical flexibility to allow new actors to be introduced and interrelationships to be created and maintained. As the social scientific population has increased, its members individual and collective have become increasingly differentiated and specialized, while at the same time arrangements to bring them together in various combinations have evolved. The result can reasonably be regarded as a system of interdependent parts.

See also: Discipline Formation in the Social Sciences; Ethical Questions in Social and Behavioral Sciences, History of.

Bibliography

Alter, Peter, 1980. The Royal Society and the International Association of Academies 1897–1919. Notes and Records of the Royal Society of London 34 (2), 241–264.

Brian, Eric, 1989. Statistique administrative et internationalisme statistique pendant la seconde moitié du XIXe siècle. Histoire de la Mesure IV (3/4), 201–224.

Brody, Eugene B., 1998. The Search for Mental Health: A History and Memoir of WFMH, 1948–1997. World Federation for Mental Health, Baltimore, MD.

Celebi, Nilgun, 2002. Sociology associations in Turkey: continuity behind discontinuity. International Sociology 17 (2), 253–267.

Coakley, John, Trent, John, 2000. History of the International Political Science Association 1949–1999. International Political Science Association, Dublin.
Coats, A.W., 1971. The role of scholarly journals in the history of economics. Journal of Economic Literature 9 (1), 29–44.
Crothers, Charles, 2005. History of sociology in New Zealand. In: Germov, John, McGee, Tara Renae (Eds.), Histories of Australian Sociology. Melbourne University Press, Melbourne, pp. 72–80.
Crothers, Charles, 2007. The changing profile of core publications in the discipline of sociology, unpublished paper.
Fourcade, Marion, 2009. Economists and Societies: Discipline and Profession in the United States, Britain and France, 1890s to 1990s. Princeton University Press, Princeton, NJ.
Glatzer, Wolfgang, 1994. Deutsche Gesellschaft für Soziologie [in English]. Soziologie, Special Edition 3, 216–231.
Henderson, James P., 1996. Emerging learned societies: economic ideas in context. Journal of the History of Economic Thought 18, 186–206.
Hudson, James R., Hudson, Patricia A., 2005. Associations and their journals: the search for an "official" voice. Sociological Perspectives 48.2, 271–289.
Lengyel, Peter, 1967. Introduction: social science periodicals in the world. International Social Science Journal 19 (2), 145–161.
Patel, Sujata, 2002. The profession and its association: five decades of the Indian Sociological Society. International Sociology 17 (2), 269–284.
Platt, Jennifer, 1998. A Brief History of the ISA, 1948–1997. International Sociological Association, Madrid.
Platt, Jennifer, 2002. Fifty Years of the International Social Science Council. ISSC, Paris.
Platt, Jennifer, 2003. The British Sociological Association: A Sociological History. Sociology Press, Durham.
Platt, Jennifer, 2010. Sociology. In: Roger, E. (Ed.), Backhouse and Philippe Fontaine, the History of the Social Sciences since 1945. Cambridge University Press, Cambridge.
Rosenzweig, Mark R., Holtzman, Wayne H., Sabourin, Michel, Bélanger, David, 2000. History of the International Union of Psychological Science (IUPsyS). Psychology Press.
Royal Economic Society. Home page. http://www.res.org.uk/view/index.html.
Simpson, Ida Harper, 1988. Fifty Years of the Southern Sociological Society. University of Georgia Press, Athens, GA.
Singleton, Alan, 1981. Learned societies and journal publishing. Journal of Information Science 3, 211–226.
Vanderstraeten, Raf, 2010. Scientific communication: sociology journals and publication practices. Sociology 44 (3), 559–576.
Whitley, Richard D., 1970. The operation of science journals: two case studies in British social science. The Sociological Review 18 (2), 241–258.

Social Science Professions and Professionalization

Felice J Levine and Nathan E Bell, American Educational Research Association, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article focuses on social science professions in the United States. It addresses the nature of these professions, the process of professionalization, degree production and graduate enrollment in social sciences fields, degree completion, and time-to-degree. Moreover, this article examines diversity and the social sciences doctoral degree, international comparisons in social sciences doctoral degree production, and the contours of the labor force in these fields.

Introduction to Social Science Professions and Professionalization

This article focuses on social science professions in the United States. It addresses the nature of these professions, the process of professionalization, patterns of enrollment and degree production, and the contours of the labor force in these fields. Professionalization refers to the processes through which individuals become professionals. Labor force refers to that portion of the population (in this case of social science professions) that is economically active, meaning either employed or, if unemployed, available to work. Although social science professionals are employed across work sectors, special attention is paid to the social science professoriate.

The Social Sciences as Disciplines and Professions

The social sciences refer to a large number of disciplines and interdisciplinary fields that in their own right constitute more than a single profession.[1] Disciplines like economics, political science, psychology, and sociology have their own specialized training, expertise, reward structures, associations, networks, and norms that reinforce distinctive professional identities. As scientific fields dedicated to the study of human social phenomena, they share common features. Yet, because of their arenas of expert knowledge, occupational cultures, collegial interactions, labor markets, and their distinctive identities as perceived by others, they comprise different professional groups. Moreover, these professions are not the same as the academic profession, although many persons with advanced social science degrees (especially with doctoral degrees) are academic professionals at universities and colleges.

The Process of Professionalization

What Is Professionalization

Professionalization involves the development of skills, identities, norms, and values associated with becoming part of a professional group. Through this process, individuals pursuing careers in specific social sciences acquire both substantive and methodological knowledge and develop understandings of their roles that permit them to function as professionals in their fields. By training new entrants and fostering social cohesion and identity, these professions seek to ensure that the work of their sciences will continue congruent with certain core principles and practices. Professionalization can commence during undergraduate education, is central to graduate training, and continues during early career stages.

Research on Social Science Professionalization

Research on professionalization dates back to the 1950s. Early research on professionalization focused on the medical (e.g., Becker et al., 1961; Merton et al., 1957) and legal (e.g., Lortie, 1959; Warkov and Zelan, 1965) professions. At the same time, there was also attention to higher education professionals, with initial research focused on faculty (e.g., Caplow and McGee, 1958; Parsons and Platt, 1968) rather than on socialization to academic or scientific professions (an exception is Berelson's study of graduate education published in 1960).

In the more than 50 years since, an important body of research and scholarship has evolved on the socialization to academic and scientific careers.[2] Much of this work focuses on graduate departments as the primary socialization site in the vital transition from student to professional (e.g., Austin, 2002; Gardner and Mendoza, 2010; Morrison et al., 2011; Weidman and Stein, 2003). In graduate departments, students learn what is expected and rewarded and, conversely, what constitutes unacceptable practices in their field or discipline. Graduate education aims to develop in students a level of scholarly proficiency through formal (e.g., courses, qualifying exams, dissertations and informal (e.g., attendance at scholarly meetings) training. In particular doctoral programs aim to initiate students progressively into more independent work and autonomous thinking that helps develop professionals who can both master knowledge and produce original work (see Sullivan, 1991). Research on doctoral education makes clear that professionalization depends not just on the quality of the substantive curriculum and research training but also on the quality of advising, mentoring, and peer relationships (see, e.g., Nettles and Millett, 2006).

Variation and Change in Social Science Professionalization

Beyond the shared norms of science that transcend any one discipline, scientific fields differ as to their intellectual tasks, their methods of inquiry, and how their work is done and rewarded (see Austin, 2002). For example, chemistry in comparison to sociology has a longstanding tradition of intensive collaboration on more tightly defined problems. Thus, chemistry students are more closely supervised, work in research teams, and are involved in joint publication. Sociology

students have more variable sets of exposure – from large-scale teams to research environments with intermittent contact with advisors, more independent work, and less engagement in collaboration. Sociology students, however, work and publish more collaboratively than do anthropology students, who are in a field more influenced by single investigator, ethnographic research (Babchuk et al., 1999).

Also, it is important to recognize that professionalization is a dynamic social process. As higher education and social science professions have changed over time, the aims and ambitions of professionalization have also transformed. Graduate programs are still primarily anchored on socializing the next generation of scientists and scholars who can contribute to knowledge production. This longstanding purpose, however, has expanded to include broader aspects of research training (e.g., exposure to peer review, considerations of ethical conduct, oral and written communication). Also, beyond an emphasis on research, there is increasing attention to other aspects of scientists' roles and responsibilities (e.g., teaching, service, human rights) as well as to professional opportunities in settings outside of the academy.

Enrollment and Degrees in the Social Sciences

Bachelor's Degrees

Undergraduate majors in social science fields constitute the potential pipeline of those who might be attracted to and eligible for becoming social scientists. They also are an important 'constituent' group because the number of majors in a discipline shapes the demand for trained faculty members in that area in the academic workplace. Even with only a bachelor's degree, they represent a sector of the future workforce that may ultimately supply jobs classified as social science occupations (ostensibly among those positions that do not require more advanced expertise or skills).

Compared to majors in mathematical, physical, and biological sciences, undergraduates typically declare majors in the social sciences later in their academic careers. This is primarily attributable to the scarce attention given to the social sciences as fields of science in the elementary and secondary school curricula. Therefore, many undergraduates have their first exposure to these fields as a part of fulfilling their general education requirements.

Despite later exposure to these fields, the social sciences confer a substantial number of bachelor's degrees. In 1980, even excluding psychology, bachelor's degree production was higher in the social sciences (93 119) than in the next highest field of science and engineering (63 942 biological and agricultural sciences bachelor's degrees). Thirty years later, in 2010, bachelor's degree production (again excluding psychology) continued to be higher in the social sciences (163 071) than in other science and engineering fields (see S&ED, 2013: Table 5).[3]

Bachelor's degree production varies considerably within the social sciences. As shown in **Table 1**, from 1980 to 2010, bachelor's degree production was substantially larger in psychology than in any other social science field, with political science also sizable. Furthermore, degree production increased over these three decades, but the rate of increase varied by discipline. While bachelor's degree production in economics increased by 47.4%, in psychology it increased by 129.9%.

The varying growth rates of disciplines within the social sciences have resulted in changes over time in the share of social science degrees awarded by individual disciplines. For example, bachelor's degrees in psychology accounted for 31.3% of all bachelor's degrees awarded in the social sciences in

Table 1 Degrees awarded in the social sciences by level, select years, 1980–2010

	Year						
	1980	*1985*	*1990*	*1995*	*2000*	*2005*	*2010*
Bachelor's degrees							
Economics	19 736	23 073	25 587	19 083	19 400	25 160	29 090
Political science	30 164	32 378	41 618	41 978	36 384	51 540	57 649
Psychology	42 513	40 237	54 018	72 601	74 656	86 031	97 746
Sociology	19 181	12 165	15 993	22 974	25 683	28 556	29 000
Other social sciences	24 038	17 180	22 152	28 676	32 030	39 315	47 332
Master's degrees							
Economics	2386	2532	2456	2838	2515	3504	3732
Political science	10 006	9481	9941	12 802	13 038	17 421	22 083
Psychology	7861	8481	9308	13 132	13 806	16 478	21 032
Sociology	1372	1045	1213	1790	2000	1500	1430
Other social sciences	5174	4090	4620	5829	5807	6990	8704
Doctoral degrees							
Economics	927	959	1007	1152	1086	1183	1190
Political science	730	666	734	894	986	990	1203
Psychology	3098	3117	3281	3429	3615	3323	3421
Sociology	600	461	428	540	617	536	638
Other social sciences	1115	1131	1162	1292	1466	1432	1782

Source: S&ED, 2013. Science and Engineering Degrees: 1966–2010. Detailed Statistical Tables, NSF 13-327. National Science Foundation, National Center for Science and Engineering Statistics, Washington, DC. Tables 5, 12, and 19 (accessed 04.09.13.).

Table 2 Graduate enrollment in science and engineering, select years, 1980–2011

	Year						
	1980	*1985*	*1990*	*1995*	*2000*	*2005*	*2011*
Engineering	74 335	96 018	107 658	107 201	104 112	120 565	146 501
Agricultural sciences	12 689	11 846	11 563	12 768	12 023	13 123	16 129
Biological sciences	47 261	45 709	49 602	58 344	56 282	68 479	75 423
Computer sciences	13 578	29 769	34 257	33 458	47 350	47 978	51 234
Earth, atmosphere, and ocean sciences	14 051	15 414	13 977	15 716	13 941	14 836	15 820
Mathematical sciences	15 311	17 563	19 774	18 504	15 650	20 210	23 801
Physical sciences	26 934	30 987	34 082	33 399	30 385	36 375	39 694
Social sciences	121 441	110 685	126 128	143 076	133 793	156 709	166 147
Social sciences	80 831	69 964	77 961	89 435	83 327	99 427	111 661
Psychology	40 610	40 721	48 167	53 641	50 466	57 282	54 486

Source: National Science Foundation, National Center for Science and Engineering Statistics, Graduate Students and Postdoctorates in Science and Engineering: Fall 2011 (NSF 13-331), Tables 2 and 3.

1980, but, by 2010, this share had increased to 37.4%. In contrast, economics bachelor's degrees decreased from 14.6% of all bachelor's degrees awarded in the social sciences in 1980 to 10.9% in 2010.

Graduate Enrollment

As can be seen in Table 2, graduate enrollment in the social sciences has grown steadily over the past three decades, fueled in part by increases in bachelor's degree production in these fields. At the graduate level, enrollment in the social sciences increased from 121 441 in 1980 to 166 147 in 2011 – up by 36.8%. Graduate enrollment is also larger in the social sciences than in all other science and engineering fields, with enrollees in social science fields accounting for 29.6% of all science and engineering graduate students. While the size of graduate student enrollment in the social sciences is substantial, the growth in enrollment of engineering graduate students over the 1980 to 2011 period approximately doubled, making it a close second.

Master's and Doctoral Degrees

As with bachelor's degree production, across all fields of science master's and doctoral degree production has visibly increased since the 1980s (see S&ED, 2013: Tables 12 and 19). In the social sciences, master's degree production overall increased from 1980 to 2010, largely attributable to substantial growth in political science and psychology (see Table 1). Master's degrees in engineering also grew substantially within this time frame.

At the doctorate level, degree production in the social science disciplines has grown only modestly, largely attributable to an increase in doctorate degrees in political science and in the broad category covering other social science fields. While in 1980 doctoral degree production in social sciences fields (6470) was larger than in other science and engineering fields, by 2010, degree production in social sciences (8234) was outpaced by production in biological and agricultural sciences (9036).

Within the social sciences, there were variations in doctoral degree production by discipline (see Table 1). Over the period from 1980 to 2010, the number of doctoral degrees awarded in psychology was substantially higher than in other social science disciplines. In 2010, psychology conferred 3421 doctoral degrees, five times more than sociology and nearly three times more than economics (although about 40% of the psychology doctorates awarded in 2010 were clinical doctorates). While doctoral degree production in psychology continues to outpace degree production in other social sciences fields, the rate of increase in psychology over the 30-year period (10.4%) is lower than in political science (64.8%), other social sciences (59.8%), and economics (28.4%).

Doctoral Degree Completion and Time-to-Degree

The duration of doctoral education remains a concern in higher education despite a decrease in time-to-degree over the last two decades. Based on the most recent data available from the National Science Foundation (see S&ED, 2010: Table 31), doctoral students in the social sciences tend to take longer to complete their degrees than students in other fields of science and engineering. In the social sciences, the median time-to-degree in 2010 was 7.7 years; this figure trailed the physical sciences (6.7 years), life sciences (6.8 years), and engineering (6.9 years). Doctoral students in the social sciences, however, typically complete their doctorates more rapidly than those in humanities, where the median time-to-degree in 2010 was 9.3 years.

Moreover, completion rates are a significant consideration (see Table 3). According to the Council of Graduate Schools (2008), approximately 56% of social science doctoral students and 55% of doctoral students in mathematics and the physical sciences complete their degrees within 10 years. These rates are visibly lower than the ones of doctoral students in engineering (64%) and the life sciences (63%). All fields of science and engineering, however, have a higher 10-year completion rate than in the humanities (approximately 49%).

Financial support for doctoral students is a relevant factor in considering time-to-degree and lower completion rates. On average, doctoral students in social sciences receive less non-loan financial support than students in other fields of science and engineering. According to 2010 data from the National Science Foundation (S&ED, 2010: Table 35), doctoral students

Table 3 Cumulative 10-year completion and attrition rates for students entering doctoral programs between 1992–93 and 1994–95, by broad field

	Rates by years since entering program							
Broad fields	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Completion rates								
Engineering	7.1	17.1	34.6	48.5	56.8	60.8	62.6	63.6
Life sciences	4.2	9.4	21.7	42.6	53.7	59.6	61.9	62.9
Math and physical sciences	2.5	8.9	23.4	39.3	48.2	52.2	53.9	54.7
Social sciences	6.7	11.5	20.5	31.0	40.9	47.5	52.7	55.9
Humanities	2.8	6.1	11.8	19.8	29.3	36.7	44.6	49.3
Attrition rates								
Engineering	20.2	23.1	24.1	25.0	25.8	26.4	26.6	26.7
Life sciences	17.1	21.0	22.3	23.7	24.9	25.6	26.0	26.2
Math and physical sciences	26.1	30.8	32.7	34.1	35.4	36.2	36.6	36.9
Social sciences	16.7	20.0	22.0	23.4	24.4	25.5	26.4	26.9
Humanities	14.9	18.4	20.7	23.8	26.7	29.1	30.6	31.7

Source: Council of Graduate Schools, Ph.D. Completion and Attrition: Analysis of Baseline Program Data from the Ph.D. Completion Project.

in the social sciences are more likely to primarily finance their doctoral education through their own resources (27.1%) compared to students in physical sciences (4.4%), engineering (4.5%), and life sciences (8.9%). As a result, 60.8% of social sciences doctorate recipients in 2010 graduated with debt, compared with fewer than half of the doctorate recipients in other fields of science and engineering (S&ED, 2010: Table 38).

Diversity and the Doctoral Degree

Gender

Since the 1980s, there has been a dramatic increase in women receiving doctoral degrees across all fields of science and engineering (see S&ED, 2013: Tables 19, 23, 25, 35, 40, 41, and 46). In 2010, 13 548 women received science and engineering doctoral degrees, a number which is more than three times higher than the 3961 women who earned such degrees in 1980. These numbers also represent a substantial difference in the shares of all science and engineering doctoral degrees awarded to women (40.9% of all science and engineering degrees in 2010 compared to 22.3% in 1980). Furthermore, fields vary considerably in the proportion of doctorates awarded to women, with the social sciences and the biological sciences conferring a much higher proportion of doctorates to women than the physical sciences and engineering.

There are notable differences within the social sciences as well (see S&ED, 2013: Tables 40, 42–45). In 1980, men earned the majority of the doctorates awarded in all social science fields, with psychology having the highest percentage of degrees earned by women (42.3%) and economics the lowest (12.5%). By 2010, women earned the majority of the doctorates awarded in psychology (69.8%), sociology (60.8%), and other social sciences (54.8%), but earned well under half of the degrees in political science (42.9%) and economics (34.5%).

Race/Ethnicity

The data on earned doctorates for racial and ethnic minorities who were US citizens or permanent residents were first reported in the 1970s. In 1980, across all fields of science and engineering, the number of persons of color receiving doctorates was quite small (814, or 6% of all doctoral degrees awarded to US citizens in science and engineering fields; see Hoffer et al., 2001: Table 8). The numbers of minorities earning doctorates in science and engineering and the proportion of science and engineering doctorates earned by minorities increased noticeably over the next three decades (S&ED, 2010: Table 23). By 1990, minorities earned 1963 science and engineering doctorates (or 12.3% of all science and engineering doctorates awarded to US citizens and permanent residents). By 2010, this number more than doubled to 4584 science and engineering doctorates (21.5%).

There are important differences across fields of science in the proportion of doctorates awarded to racial/ethnic minority groups (see S&ED, 2010: Table 23). Of the doctorates earned in engineering in 2010, 16.9% were Asians and 4.2% were black[4]; similarly, Asians received 10.5 and 11.1% of the doctorates, respectively, in the physical and life sciences, with blacks receiving only 3.0 and 5.1% of the doctorates in these fields. In contrast, greater balance across racial/ethnic minority groups is evident in social science fields. The percentage earning doctorates is more than 6%, respectively, for Asians, blacks, and Hispanics. The proportion of American Indians/Alaska Natives earning doctorate remains low (well under 1%) across all broad fields (see S&ED, 2010: Table 24).

Across the social sciences, racial/ethnic minorities differ in the percent of doctorates being earned (see S&ED, 2010: Table 24). For example, in 2010, economics and sociology awarded 22.7 and 22.3% of their doctorates, respectively, to racial/ethnic minorities, with psychology and political science somewhat lower at 19.2 and 18.3%, respectively. In contrast, anthropology was considerably lower – awarding only 14.7% of its doctorates to racial/ethic minorities. Asians earned higher proportions of doctorates in economics (11.8%) than in other social science fields, while blacks earned their highest share of doctoral degrees in sociology (8.9%).

Non-US Recipients

The numbers of non-US citizens on temporary visas earning doctorates in science and engineering almost doubled over the past three decades (see **Table 4**). In 2010, non-US citizens earned 33.4% of all science and engineering doctorates awarded in the US, up from 15.6% in 1980 (see S&ED, 2010: Tables 18 and 22). Across broad fields of science and engineering, there is considerable variability in the proportion of non-US recipients of doctorates – ranging from approximately 26% in the life sciences to 51% in engineering. In the social sciences, the shares of doctorates earned by non-US citizens are overall lower than in other fields of science. International students earned 8.9% of all social science doctorates in 1980 and 19.9% in 2010. Within the social sciences, however, there is also variability. Non-US citizens earned the highest share of doctoral degrees awarded in 2010 in economics/econometrics (51.8%) and the lowest shares in psychology (7.9%).

Table 4 Science and engineering doctorates earned by non-US citizens on temporary visas, 1980 and 2010

	1980			2010		
Broad field	*Total doctorates*	*Non-US citizen doctorates*	*Percent non-US citizen*	*Total doctorates*	*Non-US citizen doctorates*	*Percent non-US citizen*
Total science and engineering	18 088	2814	15.6	34 908	11 659	33.4
Life sciences	5501	725	13.2	11 266	2909	25.8
Physical sciences	4071	693	17.0	8319	3351	40.3
Social sciences	6037	535	8.9	7771	1548	19.9
Engineering	2479	861	34.7	7552	3851	51.0

Source: S&ED, 2010. Doctorate Recipients from US Universities, Science and Engineering Doctorates: Detailed Statistical Tables. National Science Foundation, National Center for Science and Engineering Statistics, Division of Science Resources Statistics, Washington, DC. http://www.nsf.gov/statistics/doctorates/ (accessed 06.11.13.), Table 18.

International Comparisons in Doctoral Degree Production

As is the case with doctoral degree production generally, doctoral production in the social sciences is higher in the US than in all other countries (see Table 5). In 2008, the latest year for which comparable data are available, over 8900 doctorates were awarded in social sciences in the US (National Science Board, 2012). Russia awarded the second largest number of social science doctorates (over 6800), followed by China (2270), South Korea (2240), and the United Kingdom (2230). In 2008, the US ranked second in terms of the share of all science and engineering doctorates that are in social science fields with 26.7%. First ranked was Russia, with 46.0% of science and engineering doctorates in the social sciences in 2008 (noticeably up from 27.8% in 2000). Also of note is that, while China has a sizable doctorate degree production in science and engineering (in 2008, 28 439 compared to 33 359 in the US), only 8% of the doctorates were in social science.

Table 5 International science and engineering doctorates, select countries, select years, 2000–08

	Year		
Country	*2000*	*2004*	*2008*
China			
All science and engineering	7766	14 858	28 439
Social sciences	514	1309	2274
Percent social sciences	6.6	8.8	8.0
Germany			
All science and engineering	11 306	9554	11 314
Social sciences	1106	1004	1417
Percent social sciences	9.8	10.5	12.5
Russia			
All science and engineering	15 745	16 003	14 915
Social sciences	4375	5910	6857
Percent social sciences	27.8	36.9	46.0
South Korea			
All science and engineering	2865	3531	3716
Social sciences	108	268	311
Percent social sciences	3.8	7.6	8.4
United Kingdom			
All science and engineering	7180	9150	9470
Social sciences	970	1950	2230
Percent social sciences	13.5	21.3	23.5
United States			
All science and engineering	25 423	26 573	33 359
Social sciences	8182	8275	8904
Percent social sciences	32.2	31.1	26.7

The countries shown in this table are large producers of science and engineering doctoral degrees for which longitudinal data were available.
Source: National Science Board, Science and Engineering Indicators 2012, Appendix Tables 2.37 and 2.38.

Current Patterns for New Doctorate Recipients

Across all fields, about 57% of new doctorate recipients in 2010 with postgraduation plans had plans for employment; the remaining 43% had plans for postdoctoral study (S&ED, 2010: Table 44). In the social sciences, a slightly smaller share (35%) had plans for postdoctoral study. In the social sciences, as in other fields, increasingly larger shares of newly minted doctorates are pursuing postdoctoral study over time. The shift has been substantial. For example, the share of new doctorate recipients in social sciences pursuing postdoctoral study grew from 15.7% in 1990 to 23.9% in 2000 and to the 35% noted for 2010.

Social science doctorate recipients are also more likely to enter the academy than doctorate recipients in other fields of science and engineering (S&ED, 2010: Table 46). Among new social science doctorate recipients in 2010, 59.9% took positions in the academy immediately after the receipt of the doctorate, compared with 52.6% of all new doctorate recipients. Also, while the share of all new doctorates entering the academy has been relatively stable over the past two decades, in the social sciences, the share of new doctorate recipients entering academic employment has increased, rising from 50.3% in 1990 to 59.9% in 2010.

The Social Science Labor Force

As noted in the introduction, the labor force includes those persons who are economically active – either employed, or, if unemployed, available for work. In 2010, according to the Current Population Survey conducted by the US Census Bureau, approximately 154 million persons were active in the civilian workforce of the US, including 15 million unemployed persons (Bureau of Labor Statistics, 2011). Social science professionals constitute only a miniscule segment of this overall labor force (less than one-half of 1%) if such professionals are defined as persons having an advanced degrees (a master's or doctoral degree) in a social science.

Employment Status of Social Scientists

The overall employment status of those with advanced degree training in the social sciences is quite favorable. According to the National Science Foundation's Scientists and Engineers Statistical Data System (SESTAT), there are about 1 050 000 social science professionals in the labor force (835 000 at the master's level and 218 100 at the doctoral level). Employment rates for these individuals are strong. In 2010, only 2.0% of doctorate-level social scientists and 5.5% of those with a master's degree were unemployed (see SESTAT, 2010[5]) – considerably below the overall 9.6% unemployment rate in the US in 2010 (Bureau of Labor Statistics, 2011). Among doctorate-trained social scientists, the vast majority (about 83%) was employed full-time (see SESTAT, 2003), and about 75% are employed in science and engineering occupations, primarily in social science occupations (see SESTAT, 2010).

Social Science Occupations and Professional Training

The labor force in the social sciences can be conceptualized as (1) those persons *trained* in the social sciences or (2) those persons *working* in social sciences *occupations*. In contrast to a focus on degree training per se, classification into occupational categories includes persons trained in other fields who are working in these areas but excludes persons with relevant training who are working in other occupational areas (see National Science Foundation, 1999). For example, social science occupations would include persons with doctorates in business who are teaching sociology in a university but would exclude persons with doctorates in sociology managing a business.

SESTAT classifies persons into occupations in terms of both work and education. Based on the SESTAT surveys, in 2010, approximately 518 100 persons with a bachelor's degree or higher were classified as being in a social science occupation. This occupational category includes jobs that vary widely as to the nature of the work and required expertise. Two-thirds (67.3%) of the individuals in social science occupations received their highest degree in a social science field. Furthermore, having an advanced degree beyond the bachelor's level also matters: the majority of persons working in social science occupations have an advanced degree (see SESTAT, 2010).

The value of an advanced degree can further be seen by examining the relationship between bachelor's, master's, or doctoral training in a field and working in an occupation classified as being in that same field (i.e., 'in-field' occupation). As shown in **Table 6**, for all fields of science, persons at successively higher degree levels are more likely to be working in occupations classified as 'in-field.' In the social sciences, 1.8% of those persons with bachelor's degrees were working in a social science occupation, compared to 17.2 and 64.5%, respectively, of individuals with master's and doctoral degrees. Moreover, doctoral degree training really matters in social science fields. In comparison to other fields of science and engineering, individuals with bachelor's and master's training have a very low presence in an in-field occupation. At the doctoral level, there is far less variability across fields of science and engineering in the proportion of persons working in 'in-field' occupations.

Sectors of Employment

Examination of doctorate-level professionals in the social sciences shows the centrality of the academic workplace in relation to other work sectors. Four-year colleges and universities are the dominant employer of doctorate-level scientists and engineers. Across all science and engineering fields in 2010, 45.0% of doctorate-level scientists and engineers worked at 4-year colleges and universities; in the social sciences, 46.9% did so. Business/industry also employed large numbers of doctorate-level scientists and engineers, but far less so in social science fields. Across all science and engineering fields, 41.9% were employed in business/industry, with only 38.1% of the social scientists employed in this work sector (see SESTAT, 2010).

Although specific disciplines within the social sciences followed this overall pattern, there are important differences. As shown in **Table 7**, in 2010, more doctorate-level sociologists and anthropologists (70.7%) were employed in 4-year colleges and universities than were social scientists in other fields. In contrast, greater proportions of psychologists (48.4%) and economists (39.9%) were employed in business/industry. For psychologists, this is ostensibly due to the sizable number of psychologists who are in clinical practice.

Wages in Social Science Professions

Overall, annual median wages for social scientists are in the same range as other science and engineering fields. In 2012, in US dollars, the median wages for social science professionals were $69 280 for psychologists, $74 960 for sociologists, $91 860 for economists, and $107 420 for political scientists. Other fields of science and engineering report such median 2012 wages as $66 260 for microbiologists, $80 580 for mechanical engineers, $101 360 for mathematicians, and $106 840 for physicists (see Bureau of Labor Statistics, 2014).

Table 6 Percentage of scientists with degree training in field of occupation by employment in field of occupation, 2010

	Computer/Math scientists		*Life scientists*		*Physical scientists*		*Social scientists*		*Engineers*	
Degree	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*
Bachelor's	48.0	1281	11.8	1260	19.1	397	1.8	3337	40.0	1885
Master's	54.9	536	38.9	258	34.8	135	17.2	789	48.2	637
Doctorate	74.3	69	59.6	246	55.2	160	64.5	214	56.9	157

Table derived from analyses of data in SESTAT, an online database. For information on the specific SESTAT variables used, please contact the principal author.
Source: National Science Foundation, Scientists and Engineers Statistical Data System (SESTAT).

Table 7 Doctoral social scientists by field and employment sector, 2010

	2-Year College or other school system		4-Year College or medical institution		Government		Business/Industry	
Degree	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*
Economics	1.9	528	47.8	13 400	10.4	2917	39.9	11 177
Political and related sciences	3.9	890	64.0	14 479	8.1	1843	23.9	5419
Psychology	6.3	7096	34.0	38 150	11.4	12 752	48.4	54 310
Sociology and anthropology	4.1	1224	70.7	21 107	7.0	2091	18.2	5427
Other social sciences	6.5	1355	63.0	13 192	5.8	1209	24.8	5189

Table derived from analyses of data in SESTAT, an online database. For information on the specific SESTAT variables used, please contact the principal author.
Source: National Science Foundation, Scientists and Engineers Statistical Data System (SESTAT).

These occupations include scientists at all education degree levels, as set forth in the 2014 Bureau of Labor Statistics (BLS) *Occupational Outlook Handbook*. For scientific and engineering fields, the entry-level education is typically at least a master's degree. These salaries are in line with salaries of other occupations with typical education requirements of a master's degree or a doctoral or professional degree (Lockard and Wolf, 2012: Table 6).[6]

Women and Minorities in the Social Science Labor Force

Gender

Of the 213 800 employed doctorate-level social scientists in 2010, 52.8% were male and 47.2% were female (see SESTAT, 2010). Employed female social scientists, however, are more likely than employed male social scientists to have received their doctoral degrees in recent years. For example, among social scientists who received their degrees between 1976 and 1980, approximately 70.7% were male. In contrast, of those who received degrees in 2006 or later, only 38.4% were male (see SESTAT, 2010). This pattern holds true in all social science disciplines, with women accounting for larger shares of more recent cohorts than earlier cohorts of employed doctorate-level professionals (see SESTAT, 2010).

Males and females with doctorates in the social sciences vary only modestly in terms of sectors of employment. Gender differences within disciplines are more striking than the overall trends. For example, a higher proportion of male than female economists was employed by 4-year colleges and universities (49.6 vs 42.1%), while a higher proportion of female than male sociologists and anthropologists were in business/industry (21.7 vs 15.0%) (see SESTAT, 2010).

Race/Ethnicity

Of the 213 800 employed doctorate-level social scientists in 2010, 81.2% were white, non-Hispanic; 11.8% were underrepresented minorities (African-American, Hispanic, or Native American), and 7.0% were Asian.[7] The data on doctorate-level social scientists of color in the workforce are consistent with the increase in the production of minority doctorates since the mid-1970s (see SESTAT, 2010). Of those in the workforce in 2010 who received their doctoral degrees between 1976 and 1980, 90.2% were white, non-Hispanic, 6.6% were underrepresented minorities, and 3.1% were Asian. In comparison, of those in the workforce receiving degrees since 2006, 72.6% were white, non-Hispanic, 17.5% were underrepresented minorities, and 9.96% were Asian (see SESTAT, 2010).

Employment sector data also point to important differences between minority and white doctorate-level social scientists (see SESTAT, 2010). In 2010, higher proportions of Asians (53.7%) and underrepresented minorities (51.4%) were employed at 4-year colleges and universities than white, non-Hispanics (45.7%). In contrast, a higher proportion of white, non-Hispanics were employed in business/industry (39.5%) than Asians (33.2%) and underrepresented minorities (31.6%). The higher proportions of minority doctorates employed in higher education compared to white doctorates may reflect efforts by the academic sector to diversify its workforce in the context of a low number of doctorates of color being trained.

Looking Ahead for the Social Science Professions

This article has provided an overview of basic patterns and trends in the preparation and employment of social science professionals. Enrollment and degree patterns suggest that the supply of new social scientists will remain strong at least for the foreseeable future. The current picture and projections also suggest that overall degree production will be met by employment opportunities in academic and nonacademic work sectors. Furthermore, among social science professionals at the doctorate level, higher education will continue to be the dominant employer for social scientists.

Psychology is the one exception. As has been the case for more than a decade, psychology (by far the largest of the social science professions) has a larger footprint in business/industry than in the academy, consonant with doctorate training in clinical, practice, and organizational specializations. The proportion has remained relatively stable for quite some time. In both 1997 and 2010, approximately 48% of psychologists were employed in business/industry and 34% were in 4-year colleges or universities. While for economists the academy still remains the primary employer at 47.8%, that number is down approximately 10% and business/industry is up approximately 10% since 1997 (see Survey of Doctorate Recipients, 1999: Table 17; compared to **Table 7** above).

Looking ahead, the BLS anticipates a relatively strong labor market for social scientists (Bureau of Labor Statistics, 2012). Between 2010 and 2020, BLS projects a 22% increase in new jobs for psychologists, an 18% increase for sociologists, an 8% increase for political scientists, and a 6% increase for economists. In addition, demand for postsecondary teachers, a major occupation for social scientists, is expected to increase by 17% between 2010 and 2020.

Employment opportunities and the salary picture also suggest that social science professions are among the occupational groups that will expand somewhat over the years ahead. The data indicate that the growth observed in the 2010–12 period is not merely growth in new jobs to recover jobs lost during the recession that hit especially hard in 2009. There was indeed overall growth in employment of 4.8% across all fields of life, physical, and social science occupations between 2006 and 2010 (see Lockard and Wolf: Table 1). This growth placed it in the top 3 of 22 major occupational groups – at a time when 12 of these groups experienced declines.

Because social science professionals at the doctorate level are primarily engaged in scientific research, their capacity to pursue the work that motivated entry into their professions is inextricably tied to the conditions of higher education and the availability of research funds. The current period is a time of uncertainty. Higher education institutions are seeking to provide accessible, affordable, and excellent education. Moreover, they are focused on modernizing their infrastructure and technology, and, at least for research-intensive institutions, serving as environments that can support scientific and scholarly productivity, creativity, and innovation under rapidly changing conditions of accountability, financing, and governance in the US and worldwide (see, e.g., Altbach, 2013; Arum and Roksa, 2011; Bok, 2013; Zumeta et al., 2012).

While higher education is undergoing a restructuring, it is also important to keep in mind that the academy is segmented into different types of institutions that employ social science professionals for missions that vary in function and purpose. The forms of work for social science professionals in diverse academic markets (e.g., research universities, liberal arts colleges, 2-year colleges) may change as higher education overall transforms and as sectors of the academy undergo specific changes. Increases in student–faculty ratios, more use of part-time or contractual appointments, or the increased inclusion of massive open online courses are just some of the contingencies that could shape the lives of social science professionals in the academy in the next decade.

Research funding also continues to be a salient consideration to the work of social science professionals. The US federal government provides substantial funding each year for basic and applied research (see National Science Board, 2012: Tables 4–35).[8] Over a 20-year period, however, funding obligated for social science research has been a small proportion of the total. In 1989, $20.77 billion was invested in research in science and engineering fields – with 4.68% allocated to psychology and the other social sciences. In 2009, the overall federal obligations grew to $63.7 billion, with the share to psychology/social science dropping slightly to 4.09%.

Within psychology/social sciences, psychology has over time received more federal funding than the other social science fields (accounting for 3.27% of the 4.09% in 2009). In comparison to the life sciences that receive over half of federal funds for basic and applied research or to the medical sciences or engineering that each receive better than 15%, the psychology investment is quite limited. Psychology, however, along with mathematics and computer sciences, has had one of the larger growth rates across scientific fields (8.3%) over the same 20-year period.

Despite areas of uncertainty, the overall picture for social science professions is favorable. The issues these fields examine and the theories and methods they bring to the study of human, social, and institutional processes hold promise of advancing knowledge that is valuable worldwide. Attention earlier in children's education to developing an appreciation for and savvy in these sciences would further attract talented undergraduates and also expand policy maker and public understanding of the contributions and significance of these fields.

End Notes

1. The term 'social science' as used in this article is intended to mean the social and behavioral sciences. These sciences include a range of disciplines and fields addressed to human and social processes. This article focuses primarily on four professional social science fields (economics, political science, psychology, and sociology) where sufficient national data are available in the United States to examine general trends and contrast these fields to other scientific professions.
2. Although, as noted earlier, scientific and academic professions are distinct, socialization for both are an integral part of the graduate education process. Students are exposed to the features, opportunities, and requirements of academic professions, even if their aspirations as social science professionals lie outside of the academy.
3. Psychology attracts substantial interest among undergraduates – many of whom are drawn to the possibility of clinical or counseling occupations. In 1980, there were 42 513 bachelor's degrees awarded in psychology; in 2010, there were 97 746 psychology bachelor's degrees.
4. The term 'black' as presented in the data table in S&ED (2010) refers to US citizens and permanent residents who are black/African American.
5. In prior releases of SESTAT, data were available in published tables. The 2010 data, however, are only available in a database where analyses need to be performed. For further information on the specific variables used in any calculations from SESTAT, please contact the principal author.
6. At the time of this writing only 2010 wage data are available for occupations by level of education. The median annual wage for a master's degree is $60 240 and for a doctoral or professional degree is $87 500.
7. These figures drawn from SESTAT (2010) include US citizens and permanent residents, as well as non-US citizens on temporary visas.
8. In 2009, the federal investment for basic and applied research was $63.7 billion – about 48% of the total federal investment in research and development (National Science Board, 2012: pp. 4–33).

See also: Discipline Formation in the Social Sciences; Social Science Infrastructure: North America (Research and Teaching); Social Science Infrastructure: North America (Resourcing and Funding); Social Science Infrastructure: North America (Social Science Academies and Related organizations); Social Science and Universities.

Bibliography

Altbach, P.G., 2013. The International Imperative in Higher Education. Global Perspective on Higher Education. Sense Publications, Dordrecht, Netherlands.

Arum, R., Roksa, J., 2011. Academically Adrift: Limited Learning on College Campuses. University of Chicago Press, Chicago.

Austin, A.E., 2002. Preparing the next generation of faculty: graduate school as socialization to the academic career. The Journal of Higher Education 73 (1), 94–122.

Babchuk, N., Bruce, K., Peters, G., 1999. Collaboration in sociology and other scientific disciplines. The American Sociologist 30 (3), 5–21.

Becker, H.S., Geer, B., Hughes, E.C., Strauss, A.L., 1961. Boys in White: Student Culture in Medical School. University of Chicago Press, Chicago.

Berelson, B., 1960. Graduate Education in the United States. McGraw-Hill, New York.

Bok, D., 2013. Higher Education in America. Princeton University Press, New Jersey.

Bureau of Labor Statistics, 2011. Labor Force Statistics from the Current Population Survey. http://www.bls.gov/cps/cpsatabs.htm (accessed 06.11.13.).

Bureau of Labor Statistics, 2012. The 2010–20 job outlook in brief. Occupational Outlook Quarterly 56 (1). http://www.bls.gov/opub/ooq/ (accessed 07.11.13.).

Bureau of Labor Statistics, 2014. Occupational Outlook Handbook, 2014–15 Edition. http://www.bls.gov/ooh/ (accessed 17.01.14.).

Caplow, T., McGee, R.J., 1958. The Academic Marketplace. Basic Books, New York.

Council of Graduate Schools, 2008. Ph.D. Completion and Attrition: Analysis of Baseline Program Data from the Ph.D. Completion Project. Council of Graduate Schools, Washington, DC.

Gardner, S.K., Mendoza, P., 2010. On Becoming a Scholar. Stylus, Sterling, VA.

Hoffer, T.B., et al., 2001. Doctorate Recipients from United States Universities: Summary Report 2000. National Opinion Research Center, Chicago.

Lockard, C.B., Wolf, M., 2012. Occupational employment projections to 2020. Monthly Labor Review 135, 84–108.

Lortie, D.C., 1959. Laymen to lawmen: law school, careers, and professional socialization. Harvard Educational Review 29, 352–369.

Merton, R.K., Reader, G.G., Kendall, P.L., 1957. The Student-Physician. Harvard University Press, Cambridge, MA.

Morrison, E., Rudd, E., Nerad, M., 2011. Early careers of recent US social science PhDs. Learning and Teaching 4 (2), 6–29.

National Science Board, 2012. Science and Engineering Indicators 2012. National Science Foundation, Arlington, VA. http://www.nsf.gov/statistics/seind12/ (accessed 19.01.14.).

National Science Foundation, May 3, 1999. Division of Science Resource Studies, 1999. Counting the S&E Workforce – It's Not that Easy. Issue Brief-NSF-99-344. National Science Foundation, Washington, DC. http://www.nsf.gov/statistics/issuebrf/sib99344.htm (accessed 22.01.14.).

National Science Foundation, 2013. National Center for Science and Engineering Statistics. Graduate Students and Postdoctorates in Science and Engineering: Fall 2011. Detailed Statistical Tables NSF 13-331. Arlington, VA. http://www.nsf.gov/statistics/sestat/ (accessed 22.08.13.).

Nettles, M.T., Millett, C.M., 2006. Three Magic Letters: Getting to Ph.D. Johns Hopkins University Press, Baltimore, MD.

Parsons, T., Platt, G.M., 1968. Considerations on the American academic system. Minerva 6 (4), 497–523.

S&ED, 2010. Doctorate Recipients from U.S. Universities, Science and Engineering Doctorates: Detailed Statistical Tables. National Science Foundation, Division of Science Resources Statistics, Washington, DC. http://www.nsf.gov/statistics/doctorates/ (accessed 06.11.13.).

S&ED, 2013. Science and Engineering Degrees: 1966–2010. Detailed Statistical Tables, NSF 13-327. National Science Foundation, National Center for Science and Engineering Statistics, Washington, DC (accessed 04.09.13.).

SESTAT, 2003. Scientists and Engineers Statistical Data System 2003. National Science Foundation, Division of Science Resources Statistics, Washington, DC. http://www.nsf.gov/statistics/sestat/ (accessed 22.08.13.).

SESTAT, 2010. Scientists and Engineers Statistical Data System 2010. National Science Foundation, Division of Science Resources Statistics, Washington, DC. http://www.nsf.gov/statistics/sestat/ (accessed 22.08.13.).

Sullivan, T.A., 1991. Making the graduate curriculum explicit. Teaching Sociology 19, 408–413.

Survey of Doctorate Recipients, 1999. Characteristics of Doctoral Scientists and Engineers in the United States: 1997, NSF 00-308. National Science Foundation, Arlington, VA.

Warkov, S., Zelan, J., 1965. Lawyers in the Making. Aldine Publishing Co., Chicago.

Weidman, J.C., Stein, E.L., 2003. Socialization of doctoral students to academic norms. Research in Higher Education 44 (6), 641–656.

Zumeta, W., Breneman, D.W., Callan, P.M., Finney, J.E., 2012. Financing American Higher Education in the Era of Globalization. Harvard Educational Press, Cambridge, MA.

Social Science, The Idea of

Fred D'Agostino, The University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Can human studies be scientific? The question is addressed by considering four others. What does it take for a subject to be scientific? Why does the issue matter? What are the impediments to human studies achieving standing as sciences? How are human studies to be classified if not as sciences? The potential impediments are the complexity of human phenomena, their instability over time, and the fact that they are always already 'pre-interpreted' by ordinary social agents. Alternative understandings of human studies are: literature, critique, and ideology. Crucially, the issue matters because of the institutional status which being a science confers.

Introduction

Can human studies, such as sociology and psychology, be properly scientific? This question is considered in four stages: (1) What does it take for a subject to be considered scientific? (2) Why does the issue matter? (3) What are the impediments to human studies achieving standing as sciences? (4) How are human studies to be classified if not as sciences? In relation to question (1), the importance of objectivity is stressed. In relation to question (3), three factors are identified: the complexity of human phenomena, their instability over time, and the fact that they are always already 'pre-interpreted' by ordinary social agents. In relation to question (4), human studies are compared to literature, to critique, and to ideology. And why does the issue matter? Because of the institutional status which being a science confers; because how human studies are classified influences where they look for models of their own activities; and because granting human studies the same cognitive authority as are enjoyed by some of the paradigm sciences may have morally repugnant implications.

Can human studies, e.g., sociology and psychology, be properly scientific? Since the mid-nineteenth century, philosophers and social scientists have debated this question more or less continuously. It has periodically assumed considerable urgency, indeed become a focus of attention within the disciplines, as new influences on these subjects were felt or as new models of science itself were developed. Indeed, the complexity and potential obscurity of the issue is already evident. There is an implied equation, "*human studies* are/are not *science*," in at least two variables. There has been considerable debate, at least since the 1850s, about what kinds of roles human studies might play. Likewise, what is required for a discipline to be scientific has also been a matter of continuing controversy. As Peter Manicas put it (1987: p. 168), "neither 'science' nor, a fortiori, the branches of the social sciences are 'natural kinds.'" Notice, furthermore, that the variables are not entirely independent of one another. What human studies practitioners have aspired to has itself been influenced by what they think about the desirability and the possibility of making their enterprises genuinely scientific. Similarly, though perhaps to a lesser extent, what has counted as scientific has reflected the actual practices of, inter alia, students of sociology and psychology. To approach this question, a suitably sophisticated taxonomy of possibilities is therefore required.

Some Taxonomic Distinctions and a Definition

If human studies are to count as scientific, then what it takes to make a science must first be considered.

Modality

First, *empirical* and *normative* approaches to this question need to be distinguished – roughly, "What are the sciences like?" and "What should the sciences be like?" Of course, these two questions cannot be treated in isolation, especially in view of the widely accepted maxim that "*Ought* implies *can*." This maxim means, in this instance, that those subjects whose scientific status is being assessed cannot sensibly be required to attain some excellence unless, of course, *paradigms* of scientific achievement have themselves attained this excellence.

There are, in fact, some further difficulties as the 'paradigm' sciences are not only not 'natural kinds,' they in fact exhibit some diversity among themselves (Whitley, 2000: pp. 121, 123) in relation to "the stability of problem formulations," "the extent to which work techniques are well understood and produce reliable results," and (Becher, 1989: pp. 87–88) the degree to which there is a diversity of niches into which individual enquirers separate themselves.

With these points in mind, the question about human studies can be understood in this way: Given what human studies *are* like, do they or can they conceivably measure up to whatever normative standards of excellence the best scientific disciplines *do* measure up to?

Focus

Of course, there are other issues about what it takes to make a science. The question of focus is perhaps the most important. Science and nonscience can be distinguished either in terms of their *products* or in terms of their *processes*. This fundamental distinction divides pre- and post-Kuhnian conceptions of science. Whereas *before* Thomas Kuhn's epochal work *The Structure of Scientific Revolutions* (1970),

science was distinguished from nonscience by the kinds of *statements* the two discourses produced, or, more importantly, by the status of these statements relative to epistemic standards of confirmation, generality, or objectivity, *after* this work, science was increasingly distinguished from nonscience in terms of the activities of its practitioners. Obviously, it will make a difference which of these criteria is deemed more appropriate. On the product conception, it might count against the scientific pretensions of sociology and psychology, for instance, that the more robust results in these subjects lack the kind of generality that robust results typically display in paradigm, 'hard' sciences such as physics and chemistry. But perhaps this lack of generality is irrelevant if a process criterion is adopted. Perhaps sociologists and psychologists obtain their less general findings by the same technical means and methods that natural scientists use in producing more general or more stable results.

Scope

Finally, there is the issue of scope. Roughly speaking, *wide* and *narrow* understandings of the sciences can be distinguished. On the narrow reading, it is the concrete and specific particularities of the paradigm sciences that provide criteria for judging other candidate sciences. Typically, the paradigm sciences are laboratory based, involve heavy reliance on statistical techniques of data analysis, and produce results in the form of highly confirmed differential equations relating measurable variables. On the narrow criterion, then, to count as a science, any other discipline must also exhibit these specific features. On the other hand, a wide approach to the issue requires of candidate sciences merely that they exhibit some abstract and high-level virtue which is shared by paradigm scientific achievements – that they permit the attainment of objectivity, for instance. A wide approach to the issue is probably indicated; at the concrete and specific level of technique and theory, there is enough variation *within* the paradigm sciences to rule out any narrow approach.

In considering whether human studies are scientific, an approach which is normative, process oriented, and wide in its scope should therefore be adopted. The question will be, in other words, whether it is possible that human studies be conducted in such a way that they might achieve whatever levels of objectivity in their procedures as are attainable in the paradigm sciences. (The issue is phrased in this way to permit the rather complicated judgment (likely to be favored by 'postmodernists') that human studies can achieve the same *degree* of objectivity as paradigm sciences, but that neither can achieve the *kind* of 'objectivity' that was previously thought to be characteristic of paradigm, natural sciences such as physics.)

Objectivity

To assess the scientific status of disciplines such as sociology and psychology, the objectivity of their procedures must, on this account, be assessed. But what is objectivity? Of course, there may be no more determinacy on this matter than on some others which have already been alluded to. Just as 'science' does not designate any natural kind, neither does 'objectivity'; it will be a matter of persistent debate what counts as objective. Some preliminary observations can nevertheless be offered.

First of all, transcendental and nontranscendental accounts of objectivity can be distinguished. In the transcendental account, objectivity in investigative activities has been attained when and only when all that is specific to the situation in which they are undertaken, including the personalities of the investigators, has been eliminated from them. Individuals who are otherwise dissimilar must, on this account, conduct themselves in some common way when they examine the objects of their investigations. (When scientists have managed to specify methods for the replication of results, some approximation to this ideal may have been achieved.) They know how to tell others who differ from them, how to negate the effects of these differences, and thus achieve the same results as they have. Objectivity, on this account, consists of transcending local circumstances by eliminating all factors that vary from one situation to another.

Of course, philosophers and others working under the rubric of 'postmodernism' have widely repudiated this ideal. True transcendence of local circumstances is not possible, it is not desirable, and, indeed, it is not even necessary for a form of objectivity (that is worthy of being pursued) to be attained. Objectivity requires not that local differences be eliminated, but, rather, that they be exploited in order to discover, where this is possible, a point of convergence of otherwise very different perspectives. According to this account, objectivity is attained when an individual is able to affirm from his or her own perspective a system of propositions that can likewise be affirmed from the otherwise different perspectives of such other individuals whose agreement matters (see Rorty, 1989).

Why the Issue Matters

Why should it matter whether human studies are or are not scientific, especially if there is little agreement about the concepts in terms of which the issue is defined. Discussing the so-called 'interpretive turn' in human studies, Geertz (1993: p. 8) put the point this way: "Whether this [development] is making the social sciences less scientific or humanistic study more so (or, as I believe, altering our view, never very stable anyway, of what counts as science) is not altogether clear and perhaps not altogether important." There are, in fact, at least three reasons why it matters whether human studies are considered scientific.

Institutionally

The scient*ism* of its public culture, i.e., the privilege that is granted to scientific disciplines and their methods and results, is one prominent characteristic of modernity. Science has immense, though perhaps no longer undisputed, cognitive authority, and plays an influential role in defining as well as creating the world that human beings inhabit. If human studies succeed in gaining recognition as scientific subjects, then they too could hope to exercise this kind of authority. Foucault, typically, puts the matter the other way around (1980: p. 85): "What types of knowledge do you want to disqualify in the very instant of your demand: 'Is it a science?' Which speaking,

discoursing subjects of experience and knowledge do you want to 'diminish' when you say: 'I who conduct this discourse am conducting a scientific discourse, and I am a scientist?' Which theoretical-political *avant garde* do you want to enthrone in order to isolate it from all the discontinuous forms of knowledge that circulate about it?"

Heuristically

Whether sociologists and psychologists see themselves and are seen by others as scientists will, of course, influence the conduct of their enquiries and the kinds of products they aspire to deliver. As Geertz (1993: p. 8) put it, the issue is "who borrows what from whom." Do human studies borrow techniques of observation, experimentation, and data manipulation from 'hard' sciences such as chemistry and biology, or do they, instead, look to other disciplines for their techniques or their standards of objectivity?

Indeed, this issue is part of a long-standing debate. In the late nineteenth century, the crucial questions were about the *Geisteswissenschaften* – are they concerned with understanding or with explanation? Can they articulate general psychic and/or social laws or must they, instead, confine themselves to the accurate description and contextualization of concrete and specific incidents (Windelband's *nomothetic* vs *idiographic* conceptions)? In the middle of the twentieth century, the key issue was the cognitive status of social scientific generalizations and, indeed, their very possibility in the face of innumerable obstacles and points of difference from paradigm 'hard' sciences. At the end of the twentieth century, the question is whether interpretation, with all its vicissitudes, constitutes a more central method for human studies than law-based explanation, and, indeed, whether interpretation itself is possible in a sufficiently 'objective' manner to permit the judgment that human studies are, perhaps in some extended sense, scientific. Another possibility, of course, is to acknowledge what is anyway already present in the hard sciences, namely, a kind of explanatory project that aims, not at the explanation of events, but, rather, at the explanation of patterns or regularities. So, for instance, statistical mechanics does not aim at the explanation of the trajectories of individual particles, but, rather, at the explanation, in terms of a small number of measureable variables of the patterns found in ensembles of these particles. Similarly, a scientific approach to social studies might aim at this kind of explanation, as, indeed, Hayek (1967) has advocated and as has been common, in the social sciences, *both* in the form of the explanation, via regression analysis, of trends in large corpora of data about individuals or prices or indeed any of the explananda of the social science *and*, more recently, in the form of agent-based modeling and, hence, the *generative* explanation of patterns and trends in multi-agent phenomena. (See for example Epstein, 2006.)

Morally

How human studies are conceptualized may have profound moral significance. This point is dependent on the previous two. Suppose human studies are modeled on the 'hard' sciences and that success in doing so is conspicuous enough so that these subjects gain the kind of cognitive authority that some of their practitioners have long aspired to. In this case, ideas about the human situation that are moral and cultural foundations of our longest-standing and most deeply embedded traditions are at risk of being undermined. Putnam (1978: p. 63) put it this way: "[S]uppose our functional organization became transparent to us. Suppose we had a theory of it, and we could actually *use* this theory in a significant class of cases. What would happen to us? … Would it be possible even to think of *oneself* as a *person*? … [T]he development of *that* sort of knowledge of ourselves and each other would modify our natures in ways that we cannot predict at all. Every institution we now have: arts, politics, religion – even science – would be changed beyond all recognition."

Some Positions

Whether human studies can be scientific is therefore too important a question simply to ignore. And, of course, it has not been ignored. Philosophers, social scientists, and others have, in fact, taken a number of positions on this question. Before proceeding with more substantive matters, a threefold primary distinction among these positions should be considered.

In the middle of the twentieth century, the most common position, though not so designated until later, was *naturalism*. According to this doctrine, there is nothing methodologically or theoretically distinctive about human studies in relation to paradigm, natural sciences. In particular, social and natural scientists legitimately share both aspirations (to discover and confirm general laws covering specific phenomena) and techniques (operationalization of variables, controlled experimentation, and statistical analysis of its results).

Both earlier and later than this period, however, a very common, perhaps even the most common position was *interpretivism*. According to this doctrine, human studies are distinguished from paradigm, 'hard' sciences both in their aspirations and in their techniques. Social scientists should not try to emulate the paradigm sciences narrowly construed (see Section Scope); they should strive, instead, to interpret human social and psychic phenomena, and should employ the techniques of hermeneutics, broadly construed, to achieve this.

Finally, toward the end of the twentieth century, though also but to lesser degrees at earlier periods, 'skepticism' about the social sciences has been an increasingly influential position. On this account, the laws of human being cannot be discovered, nor can properly objective interpretations of what human beings do be provided. Although the results of sociological or psychological research are often presented in the form of statements of knowledge, they are not entitled to this status, and must be understood in some other, less exhaulted way.

Skeptical positions may be either local or global. Global skepticism, often associated with the work of Jacques Derrida, repudiates all claims to determinate and methodologically adequate knowledge of human phenomena. Local skepticism adopts such a dismissive attitude only to certain claims, e.g., those that have been derived in some specific way or from some specific social position. (For instance, Marxists might be skeptical in this local sense about the claims produced by mainstream, and hence in their view 'bourgeois' sociology or psychology. Similarly, feminists are often skeptical about the

claims of mainstream and hence in their view 'patriarchal' social science.)

Impediments to Human Studies Being Scientific

That the legitimacy of human studies has so long been a matter of discussion is prima facie evidence that the investigation of human phenomena is not straightforward. What, then, are some of the reasons why it might not be possible to have a genuinely scientific approach to human social and psychic phenomena?

Pre-interpretation

Theorists otherwise as diverse as Peter Winch, R.G. Collingwood, and Alfred Schütz have pointed, in distinguishing human studies from science per se, to the fact that students of human phenomena work, whereas paradigm scientists do not work, with materials that already have meaning and significance (see, e.g., Winch, 1958: p. 87). There are strong and weak versions of this point. On the weak version, students of human phenomena must always rely, methodologically, on the objects of their research in ways in which paradigm scientists need not. Saying *what* is going on is preliminary to saying *why* it is happening, and hence to any scientific explanation. But what is going on, socially or psychically, must be understood, in Geertz's phrase, "from the native's point of view." How the subject understands what is going on determines the description of what is going on. On the strong version, once scientists have characterized "the native's point of view," they have discovered all that they need to or can know about the objects of their investigation. In either case, what paradigm scientists do and what human studies practitioners do is very different. The latter seem forced to rely on the subjective perspectives of the objects of their research. On the other hand, some sorts of scientific approaches to social phenomena are by no means ruled out even if we accept the reliance of social scientists on the conceptual apparatus of the objects of their enquiries. Again, there may be, within the particular conceptual world of the 'native,' discoverable regularities of behavior, whether at the individual or the statistically aggregated level. The scientificity of social enquiry is really only put under pressure by the strong thesis and this thesis is, in fact, itself put under pressure by, for example, the "hermeneutics of suspicion" (Ricoeur, 1970), which posits the unreliability of "the native's point of view" when it comes to understanding the forces and processes of social life.

Instability

There are three interrelated factors which make it unlikely that sociologists and psychologists will be able to develop stable, objective accounts of the phenomena they take an interest in.

First, there is the *historicity* of both the subjects and the objects of human studies research. When Gadamer (1975: p. 204) asks "Is not the fact that consciousness is historically conditioned inevitably an insuperable barrier to reaching perfect fulfilment in historical knowledge?" he articulates this fundamental difficulty: neither scientist nor subject 'stands still,' and hence what the one should say about the other changes as they themselves change. Paradigm, natural sciences, on the other hand, typically deal with statements of 'timeless' generality.

Of course, historicity and scientificity are not incompatible per se, and there is no reason, especially if the focus (see Section Scope) is on process rather than the product, to require the social studies to produce timeless generalizations.

Second, in human studies relations of representation are subject to *reflexivity* in a way that is not common in the sciences per se. What practitioners say about human phenomena influences how their subjects enact or react to these phenomena, and hence often undermines the claim, necessary if objectivity is sought, that scientists have accurately *represented* the behavior of their subjects. (Practitioners have, instead, caused or persuaded subjects to act so that their behavior conforms to the practitioners' descriptions of it.) To mid-twentieth century philosophers, this possibility was best known under the heading 'reflexive predictions' (self-defeating and self-confirming 'prophecies'), but the crucial points have since been generalized extensively. Of course, as Lyotard (1984: p. 57) rightly points out, reflexivity must not be understood in too one-sided a way. It is not just that sociologists and psychologists can create the objects of their investigations; it is, as well, that these objects can, *strategically*, resist the characterizations which practitioners attempt to fit them with. This too is an impediment to the achievement of objectivity as it is usually conceived.

Again, this point is hardly conclusive. Although not uncontroversial, ethically, the use of deception protocols in individual or small-group enquiries and the general invisibility to the population studies of many 'stand-off' observational techniques (e.g., the analysis of large corpora of data) mitigate the effects of reflexivity.

Finally, there is radical *indeterminacy* about the objects of investigation. This point is an ontological not an epistemological one. It is not that, in some or perhaps many cases, scientists cannot tell when they have characterized some phenomenon accurately. It is, rather, that there may be no sense, in these cases, to the very idea of accurate characterization. Typically, this means that there are, in the human domain, no fixed objects with well-defined characteristics for scientists to (try to) represent. In its more familiar forms, this point is, of course, an obvious development of Dilthey's observations about the 'hermeneutic circle.' It has been much emphasized by Derrida (1978) and others of his school. Specifically, if interpretation takes place against and in terms of a background that is not 'given' but must itself be an object of interpretation, then every interpretation has the form, not of an objective (if possibly incorrect) description, but, rather, of a (performative) proposal to see the situation in one of the many ways in which it could (as) legitimately be seen. There is on this account insufficient interpretive stability to warrant attributions of objectivity.

Complexity

The economist Hayek (1978: p. 26f.) had already emphasized a kind of complexity that human studies practitioners must try to deal with. Such 'organized complexity' resists reduction either to deterministic or to statistical models that might be appropriated from the 'hard' sciences. The behavior of a social aggregate often cannot, in other words, be reduced to the sum

or product of the behavior of its individual parts, nor can it be adequately understood by characterizing its statistical properties (as the behavior of an aggregate of gas molecules might be understood, for instance). Later theorists (see Mitchell, 2009 for a good summary) have emphasized a kind of complexity that arises when phenomena are subject to multiple causal factors which interact in a particular way, so that, for example, the *relationship* of two variables contributing to the phenomena itself varies according to the value of a third variable. This results (Kauffman, 1993) in a 'rugged landscape' of possibilities for understanding these phenomena … one which it is not easy to explore in a systematic way and in which it may not be possible to identify a best account of the phenomena.

Notice, furthermore, that there may be moral and prudential (as well as practical) reasons for refusing to reduce this complexity in ways standard in the 'hard' sciences, i.e., by 'experimentalizing' it. So, while scientists might be able to 'standardize' individual agents in an experimental setting so that their interactions *there* can be understood deterministically or statistically, there are moral objections to exporting these conditions to *extra*-experimental settings. Indeed, there may be prudential reasons for not using these techniques of standardization in 'naturalistic' settings – as is implied, of course, by Hayek's own analysis of the market. In particular, adequate functioning of a complexly organized whole may depend precisely on its individual elements *not* being standardized. In either case, there are thus limits to the objectivity (actually the 'ecological validity') of the results that might be obtained in laboratory settings.

Recent advances in 'generative social science' (Epstein, 2006) show, however, how complexity of this kind can be handled effectively, so long, of course, as the prediction or explanation of the behavior of individual social agents is not sought.

Interim Conclusion

All these impediments manifest themselves in an oft-expressed concern: There is often, especially compared with paradigm 'hard' sciences, a lack of progress in human studies. In particular, multiple and opposed paradigms persist in human studies, especially in any area of investigation which is likely to be of broadly 'political' interest and concern. As Geertz (1993: p. 4) has said, "calls for 'a general theory' of just about anything social sound increasingly hollow, and claims to have one megalomanic. Whether this is because it is too soon to hope for unified science or too late to believe in it is, I suppose, debatable. But it has never seemed further away, harder to imagine, or less certainly desirable than it does now."

Notwithstanding these concerns, progress in the admittedly 'local' understanding of trends and patterns has been noted and this is of some significance, both in relation to the aspirations to scientificity of the social studies disciplines and in relation to their efficacy in connection with policy development.

Alternative Conceptions of Human Studies

It may not make sense to compare sociology and psychology to the 'hard' sciences. How, then, could they otherwise be conceptualized … and hence institutionalized? There are a number of possibilities. (Manicas's cautionary note still applies, of course; if 'science' isn't a natural kind, then no more is 'literature' or any of the other disciplines with which human studies might be compared.)

Literature

If human studies cannot properly be compared to the 'hard' sciences, perhaps a comparison with literature and other forms of potentially edifying narrative or practical discourses is more appropriate. This, certainly, is the opinion of a wide range of late-twentieth century theorists, including Geertz, Charles Taylor, Gadamer, and others. Geertz, for instance, advocates (1993: p. 6) that we "turn from trying to explain social phenomena by weaving them into grand textures of cause and effect to trying to explain them by placing them in local frames of awareness …." Of course, it might seem, to practitioners, that such a heuristic reorientation (see Section Heuristically) would reduce the likelihood that human studies might have significant institutional influence (see Section Institutionally). However, it needs to be borne in mind that the institutional influence of sociology and psychology is rather patchy (this is made clear in utilization studies), and that literature still forms the basis for highly influential mass-media representations and analyses of the human situation.

Another related possibility is this. While sociology and psychology do not or cannot provide an objective account of human phenomena at the level of individual behavior, they can and do provide a variety of potentially important *perspectives* on these phenomena, and thus enable human beings to orient themselves more productively to them, especially by giving them frameworks for understanding these phenomena. In this vein, Rorty (1982: p. 247) draws a contrast between scientistic goals of prediction and control, on the one hand, and, on the other hand, practical goals of developing "descriptions which help one decide what to do." (With a particular political twist, this seems to have been Leo Strauss's approach (1989) too.) Here, of course, theories that explain or predict trends and patterns can play a significant role, as 'deciding what to do' is often a matter of deciding, not what an individual will do, but, rather, what a society ought to do about, precisely, trends and patterns – for example, the reduction, in the aggregate, of some undesirable behavior. Of course, since deciding what to do depends on values and other normative elements, and since there is widespread and apparently irreducible diversity in relation to these elements, the kind of 'practical knowledge' that nonscientistic forms of psychology and sociology might contribute to will not itself satisfy the most demanding requirements of objectivity, though this might well be on account of the difficulties of moral reasoning, not the difficulties of social studies enquiries.

Critique

There is a subtle difference, but an important one between the view that human studies offer edifying perspectives on or frameworks for the interpretation of human phenomena (see Section Literature) and the view that they offer a basis for critique and the advocacy of change. In the latter half of the

twentieth century, critical theorists (e.g., Habermas), genealogists (e.g., Foucault), feminists, and advocates for other marginalized groups (see Seidman, 1994) proposed and in some cases made progress in achieving a 'takeover' of human studies activities and institutions. Across this spectrum, techniques that have some claim to being 'antiscientific' (Foucault, 1980: p. 83) are widely employed. In particular, practitioners of critique shift the focus from showing how what is, socially, an intelligible consequence of its causal antecedents to showing how what is, socially, represents the highly contingent realization of practices and structures to which there are clear, achievable, and desirable alternatives. As Foucault says (Rabinow, 1984: p. 46): "[T]his critique will be genealogical in the sense that it will not deduce from the form of what we are, what is impossible for us to do and to know; but it will separate out, from the contingency that has made us what we are, the possibility of no longer being, doing, thinking, what we are, do, or think. It is not seeking to make possible a metaphysics that has finally become a science; it is seeking to give new impetus, as far and wide as possible, to the undefined work of freedom." It is not, then, a matter of objective representation, but, rather, of politically effective analysis of the possibilities for change. (This is, of course, an extension of Karl Marx's 11th thesis on Feuerbach: "Philosophers have only *interpreted* the world in various ways; the point is to *change* it.") Even in this context, of course, knowledge about the forces and factors sitting underneath trend and pattern data will be valuable.

Disciplines or Ideologies

There are (see Section Complexity) moral and prudential disincentives to investigating human phenomena experimentally and, a fortiori, to 'exporting' into 'naturalistic' settings those conditions in which objective, reproducible experimental effects have been obtained. This has not, of course, stopped such things from happening and, indeed, it is the main conclusion of Foucault and others of his school that human studies have played a key role in this process, and, in particular, that they have contributed to the production of a 'disciplinary society' in which individuals are governed via their consent, i.e., through the ways in which they are made into 'subjects.' Although Jürgen Habermas disputed much that Foucault wrote and did, he was at one with him on this point. Habermas (1994: p. 54) says, for instance: "It is no accident that these sciences, especially clinical psychology, but also pedagogy, sociology, political science, and cultural anthropology, can, as it were, frictionlessly intermesh in the overall technology of power that finds its architectural expression in the closed institution. They are translated into therapies and social techniques, and so form the most effective medium of the new, disciplinary violence that dominates modernity." (See Hoffmann, 2003 for the disciplinary use of social studies in the interwar Soviet Union.)

That human studies are 'disciplines' is one way of capturing their importance and their negative impact on the lives of ordinary people; that they have functioned ideologically is another way of making this point. Perhaps the most important issue here is the way in which human studies have been recruited by other powerful institutions to serve the specifically political interests of those institutions while seeming to provide a nonpolitical analysis, a value-neutral account of contemporary social conditions. The idea of expertise has played an important role in this development. Insofar as social and psychic problems can be modeled on purely technical problems, solutions can be offered and consumed on a purely technical basis, and the reality of their dependence on fundamentally political matters can remain inaccessible to consciousness.

Conclusion

Should human studies plausibly aspire to scientific status? This remains questionable, as it has long been. There are a variety of legitimate enterprises, including edification and critique which may be more important, for society, than explanation and prediction on an objective basis. Of course, these enterprises, whatever their status, may themselves be informed by properly conducted empirical enquiries. Can they realistically aspire to such a status? This depends very much on where the bar is set. If we look to social studies for the explanation and prediction of individual events in the social world, they will not, typically, satisfy the highest methodological demands that might be satisfied, at least under laboratory conditions, by the paradigm 'hard' sciences. On the other hand, there are other aspirations and, perhaps Rorty (see Section Literature) is right to suggest that human studies have aspired to too much, that their two main goals are disjoint and hence both cannot be achieved under the same methodological or institutional auspices. In this case, practitioners would perhaps be wise to disentangle their various objectives, and adjust their 'aspiration levels' (Simon, 1996: p. 30) to more realistic goals. Certainly, the analysis of trend and pattern data and its generative derivation using agent-based modeling techniques shows a middle way for those social studies which aspire to scientific status. There are methodologically sound 'local' patterns and trends that appear to be well understood, either through regression over variables or by computer simulations. Such an approach is, in fact, consistent with the even more radical adjustment of aspirations and disciplinary cultures. Toulmin calls for when he says (1992: p. 193) that "The task is not to build new, more comprehensive systems of theory with universal and timeless relevance, but to limit the scope of even the best-framed theories, and fight the intellectual reductionism that became entrenched during the ascendancy of rationalism … [thus reinstating] respect for the pragmatic methods appropriate in dealing with concrete human problems."

See also: Atomism and Holism: Philosophical Aspects; Causes and Laws: Philosophical Aspects; Economics, Philosophy of; Idealization, Abstraction, and Ideal Types; Individualism versus Collectivism: Philosophical Aspects; Interpretation in Anthropology; Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects.

Bibliography

Becher, T., 1989. Academic Tribes and Territories. Open University Press, Milton Keynes, UK.

Derrida, J., 1978. Writing and Difference. University of Chicago Press, Chicago.
Epstein, J.M., 2006. Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press, Princeton.
Foucault, M., 1980. Power/Knowledge. Pantheon, New York.
Gadamer, H.-G., 1975. Truth and Method. Seabury Press, New York.
Geertz, C., 1993. Local Knowledge. Fontana Press, London.
Habermas, J., 1994. The critique of reason as an unmasking of the human sciences. In: Kelly, M. (Ed.), Critique and Power. MIT Press, Cambridge, MA, pp. 47–78.
Hayek, F.A., 1967. Studies in Philosophy, Politics, Economics and the History of Ideas. Routledge and Kegan Paul, London.
Hayek, F.A., 1978. New Studies in Philosophy, Politics, Economics and the History of Ideas. Routledge and Kegan Paul, London.
Hoffmann, D., 2003. Stalinist Values. Cornell University Press, Ithaca, NY.
Kauffman, S., 1993. The Origins of Order. Oxford University Press, Oxford, UK.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Lyotard, J.-F., 1984. The Postmodern Condition. University of Minnesota Press, Minneapolis, MN.
Manicas, P.T., 1987. A History and Philosophy of the Social Sciences. Blackwell, Oxford, UK.
Mitchell, S., 2009. Unsimple Truths. University of Chicago Press, Chicago, IL.
Putnam, H., 1978. Meaning and the Moral Sciences. Routledge and Kegan Paul, London.
Rabinow, P., 1984. The Foucault Reader. Penguin, London.
Ricoeur, P., 1970. Freud and Philosophy (D. Savage, Trans.). Yale University Press, New Haven.
Rorty, R., 1982. Consequences of Pragmatism. University of Minnesota Press, Minneapolis, MN.
Rorty, R., 1989. Contingency, Irony, and Solidarity. Cambridge University Press, Cambridge, UK.
Seidman, S., 1994. Contested Knowledge. Blackwell, Oxford, UK.
Simon, H.A., 1996. The Sciences of the Artificial. MIT Press, Cambridge, MA.
Strauss, L., 1989. Social science and humanism. In: Pangle, T. (Ed.), The Rebirth of Classical Political Rationalism. University of Chicago Press, Chicago, IL.
Toulmin, S., 1992. Cosmopolis. University of Chicago Press, Chicago.
Whitley, R., 2000. The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford, UK.
Winch, P., 1958. The Idea of a Social Science. Routledge and Kegan Paul, London.

Social Scientists as Experts and Public Intellectuals

Stephen P Turner, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Experts and intellectuals in the social sciences have a long history of relating to the state and the public. These relations vary in kind from those based on technical knowledge applied to policy to cults to social scientists in organic relations to social movements to organized attempts to develop public policy guided by social science knowledge. The most successful early attempts were cameralism and official statistics, but intellectuals like John Stuart Mill also reached a wide public audience in the nineteenth century. In the late nineteenth century, social reform movements claimed expert knowledge. As the social sciences entered the university, however, the forms of influence changed. Under the influence of the Rockefeller philanthropies, social science became more 'realistic' and the emphasis shifted to creating professions dependent on academic knowledge and certification. Think tanks and other forms of knowledge intervention developed which relied on academic social science. Public intellectuals, however, speaking not to professions or bureaucrats, remained important.

People we now think of as social scientists and psychologists were 'experts' and sought to influence public policy as well as public discussion and personal behavior long before 'social science' was itself conceived of as an academic study. In many fields of the present social and behavioral sciences, practice preceded and formed the basis for the creation of a body of theoretical or systematic empirical knowledge. This is true to a significant extent of science generally, as writers on the origins of science in the 1930s, such as J.G. Crowther and Lancelot Hogben, emphasized. But the social sciences exhibit some distinctive patterns, and vary considerably depending especially on the way in which they relate to official bodies and publics. These relations have changed over time, especially with the disciplinarization of the social sciences, the support of foundations, and the present situation, in which there is a substantial governmental investment and use of social science. The changes also reflect changes in the state, in forms of social provision, and in the means of state economic intervention.

In understanding the role of social scientists as experts and public intellectuals, it is important to understand the ways in which this role has both developed and varied under these different circumstances. There are three major vectors relevant to understanding this role: the question of who pays for knowledge and how it is paid; the audience for knowledge; and the bureaucratic and political structures and traditions within which knowledge is applied. Audiences vary from the public to private professional practitioners (such as physicians and lawyers) to civil servants in expertized bureaucracies, the purposes that experts have in speaking to them, and the kind of knowledge at stake, varies as well. The purposes may range from influencing legislation or policy to influencing administrative decisions, to influencing personal behavior, to providing intellectual orientations relevant to political issues, and the content may vary from practical advice to abstract theory to petition signing in support of causes.

Expertise: A Typology

Experts may be categorized according to the way in which they relate to the public and public bodies. One kind of expertise, which other forms emulate, is modeled on technical knowledge, and specifically limits the role of the expert to technical advice based on empirical research. This kind of expertise is normally rooted in academic training and a community of self-defined and selected experts, and is made accessible to users, who support research related to their own aims. It gains its legitimacy from being publicly recognized as genuine expertise. Technical expertise of this kind is ordinarily self-limited to the specific domains of technical knowledge. In practice, this kind of knowledge is insufficient in itself to inform public policy or public opinion, but it can constrain the claims of other kinds of experts. However, the line between technical expertise and expertise beyond the narrow confines of what is established technically is often vague, and those who possess technical knowledge often claim more than is justified.

Expertise could be said to differ from scientific knowledge proper in that it represents the state of knowledge at a particular time, and is not limited to fully developed or tested theories or facts accepted as textbook knowledge by the academic community. In this respect, expertise resembles meta-analysis. Experts are normally expected to be aware of conflicting claims and the range of relevant knowledge and opinion on the topic in question, and to be able to aggregate this knowledge and make judgments about the methodological adequacy, significance, and relevance of various knowledge claims to present policy questions. But they are not expected to fully agree on their advice and conclusions, which may involve nontechnical elements, of values or policy considerations that are contested, often by other experts.

Within this very broad definition, several major variants may be distinguished. One model of expertise involves the direct appeal to ordinary citizens and persons on the basis of claims of special access to relevant knowledge with some sort of certification or sanction of the status of the expert in a community of experts or specialists that recognizes their expertise. The expert speaking directly to the public on the basis of specialized knowledge is, in contrast, speaking for specialists whose knowledge is being publicly represented by the expert, who is speaking for this community. This kind of expert appeal plays a role in all the social and behavioral sciences, but is far more developed in such fields as psychology and child

psychology than others. In this model, the audience of the expert is a body of consumers who use this knowledge for their own purposes, and judge it largely on grounds of its efficacy for them. In economics, the body of users is persons interested in issues of finance and policy who are themselves relatively sophisticated, though not expert in the same sense. 'Public intellectuals' are a variant of this category. They are taken to possess specialized knowledge, but they speak on issues that go beyond that specialized knowledge, and attempt to direct or influence public opinion in domains beyond the narrow domain in which they secured their professional reputation.

A second type of expertise is directed at official bodies from outside. An example of this is the field of international relations: here academic experts, with access to a body both of theory and empirical and historical material, attempt to influence the conduct of diplomats and government bodies. In this field, depending on the national traditions governing state bureaucracies, there is sometimes a 'revolving door' between academia and offices. However, strong civil service traditions, in general, tend to exclude 'outside' expertise. Though there are exceptions to this rule, there are significant local differences in the ways in which experts relate to bureaucracies.

An important particular form of this relation is expertized bureaucracies: cases in which academic qualifications or background become conditions for official appointment and in which advancement depends on continued engagement with professional associations. A complex variant of this relation was promoted by the Rockefeller philanthropies. The Rockefeller model, based on the experience of reforming American medicine, involved multiple elements: the professionalization of domains of activity, such as public administration in the United States, and involved the creation of degree programs, institutes, bodies of academic research, and intermediate institutions for continuing education and for the promotion of new expert knowledge to audiences of officials. This kind of expertise is not directed at 'the public' and indeed is claimed to be form of technical knowledge, but it is also created to be applied. The most developed social science application of this model is the field of public administration in the United States.

A variant of this use of expertise is associated with the New Public Management movement of the 1980s, which outsourced work requiring specialized expertise to consultants. This was motivated in part by efforts at reducing the size and personnel costs of permanent bureaucracies, but also was a response to long-standing issues that arose from the conflict between the hierarchical career structure and generalist character of civil service knowledge and the specialized technical knowledge relevant to specific decisions. In the British case, this conflict initially arose in connection with the siting of telegraph lines, and the anomaly of junior technical officials giving instructions to senior (but nontechnical generalist) officials.

Another type of expertise is recognized primarily by a specific audience, but not generally recognized or widely dismissed. The extreme form of this is the expertise recognized within a cult. This type of expertise shades into cases where there are groups, including specialists, that accept the claims of expertise, yet the claims are contested in the public domain. This type of contested expertise has a significant role in the history of the social sciences. Psychoanalysis and Marxism are examples, but there are many more: Auguste Comte's positivism was turned into a literal cult, and such movements as Reichian analysis were cultlike as well. But there is a similarity as well to conscious attempts to develop ideological or policy-oriented points of view on the basis of social science. Austrian economics and the Frankfurt School might be examples of this.

A third kind of expertise is associated with social movements and reform. The movements are independent of the state and aim at producing state action, normally legislative or regulatory change, and rely on expertise claims together with political pressure and organized public support. This kind of expert relationship has a large significance in the history of the social sciences in the predisciplinary period. During the 'disciplinary' period, this kind of relationship persisted, but conflicted with ideas about value-free science in fields like sociology. A new form of this relation appeared in the relation between women's studies and feminism, and is exemplified presently in the idea of 'public sociology' as involving an organic relation between sociologists and particular social movements.

This kind of expertise differs from the cult form in that cult forms generally demand exclusive loyalty and develop factions, whereas social reform movements are normally nonexclusive, allow for multiple involvement in different causes by the same persons, and tend to proliferate aims rather than factionalize, reflecting the fact that they are goal-oriented rather than ideological in character. The differences are not absolute, and in certain periods reform movements were colonized as fronts for ideological movements. Typically, however, reform movements have attempted to cooperate and coordinate with one another.

These different forms of expertise operate through distinctive institutional structures, and have distinctive products. Among the most important forms are what in science studies are known as 'boundary objects': reports or other accessible sources of knowledge that purport to represent the policy relevant facts on a particular topic. These are often produced by knowledge-intensive institutions, such as the World Bank, which employ large numbers of experts and seek to influence governments through the diffusion of expert consensuses. Expertise is disseminated in many other ways, however, from journalism and television appearances to textbooks directed at persons training for expertized bureaucracies. Among the characteristic institutional structures are think tanks, commissions, and panels.

Expertise and Public Intellectuals in the Predisciplinary Period

The cameralists, the political arithmeticians, and various political philosophers are acknowledged precursors and models for what became social science. These intellectuals prior to university 'social science' were paid in various ways, and their writings had public audiences. Niccolo Machiavelli and Thomas Hobbes, for example, were courtiers, advisors to powerful families whose public writings followed and were based on their advice and experience. The cameralists, who were administrators for the absolutist states of central Europe, in contrast, developed an intellectual tradition and interacted

with one another, but were not concerned with persuading the public. Instead, they represent a continuous tradition of administrative expertise based on a theory of statecraft: a doctrine of *Polizeiwissenschaft*, according to which the ruler of a principality enhances his own power by improving the economic and moral life of his subjects through mercantilist regulation and the paternalistic use of state administrative power through magistrates to improve morality. This was the beginning of academic social science: many of the cameralists were university professors who simultaneously held administrative positions, a pattern that persisted in the German-speaking world well into the nineteenth century.

These two patterns, the public and the bureaucratic, developed in different directions. In the eighteenth century, Hume and Adam Smith were supported by wealthy patrons, though they were not entirely dependent on them. Smith became a professor of moral science at the University of Edinburgh. Hume's history of England was a best seller. The Encyclopediasts, in France, also derived income from literary projects. This was the beginning of a period in which the educated public was a paying audience. 'Social science' at this stage was a part of a much larger body of successful literary publishing for this audience, and beneficiary of this growing and increasingly prosperous group.

It was not until the eighteenth century and the development of a significant book market that enlightenment thinkers such as Voltaire, Diderot, and Rousseau were able to support themselves through literature, but they could do so only barely and with aid. The famous Encyclopedia on which Diderot collaborated sold 25 000 copies; the income from this and his other writing was not enough to support him, but he was helped by Catherine of Russia, who bought his library and paid him a pension. Edward Gibbon and Hume were also supported by aristocrats.

In the nineteenth century, John Stuart Mill, who has been described as the greatest public intellectual of the last two centuries, worked for the East India company and derived income from his writings as well, especially his *System of Logic* (1882), which outlined his account of the methodology of social science, including his conclusion that social science, together with the utilitarian principle, would provide a complete policy science. He was part of a current of liberal thinkers, including Harriet Martineau, Walter Bagehot, and Herbert Spencer, who wrote for the market. The latter two got their start in financial journalism, which has ever since been a major conduit for communicating expertise, especially in economics, to the 'public,' and was the first domain in which experts could support themselves through writing.

These writers had a complex relation to the public and the literary market. Their most important intellectual works were not the source of their income. Most of them needed an inheritance to enable them to commit time to writing social science: Karl Marx was subsidized by Friedrich Engels, Comte by Mill. Spencer and Henry Buckle, author of the astonishingly successful *History of Civilization in England* (1864), inherited wealth that enabled them to turn to social science. It is a small irony that the great exponents of liberalism were unable to survive in the market. Nevertheless, the fact that they did derive income from the public literary market, and often through journalism, meant that they were subject of the discipline of communicating to this market, and this was reflected in their writings on social science, which were accessible rather than scholarly. For this reason, many of their works are still read today, in contrast to the works of the early academic social scientists.

The bureaucratic tradition, which was initially limited to the rationalizing absolutist states that were former parts of the Holy Roman Empire, developed into two strands: *Nationalökonomie*, a mercantilist state-oriented economic tradition, and statistics, which altered its original meaning of being concerned with the properties of states as part of *Staatswissenschaft* to the collection and governmental use of statistical data. The classic expression of the statistical office was the Royal Prussian Statistical Bureau, founded in 1805, which produced a large number of reports. Almost all of the important German contributors to statistics before 1860 were professors who had significant practical experience and held at some point, or at the same time, official positions in bureaus of statistics (Lindenfeld, 1997). In Britain, in contrast, social statistics was associated with reform movements and official bureaus and the statistical society grew out of local private statistical societies with a strong social reform interest, while academic statistics developed in connection with agricultural research. The leading figure of early social statistics, William Farr, who worked at the General Register Office for England and Wales, found it impossible to get academic legitimation for the field, but was nevertheless enormously influential as a sanitation reformer.

Statistics emerged as the best-organized social science by the middle of the nineteenth century, with international statistical congresses and a section of the British Association for the Advancement of Science. Statistics was understood not merely as a mathematical field, but as a science of humanity. The study of suicide was especially influential on public discussion, producing the 'suicide problem.' In the latter part of the century, the focus shifted to labor statistics, and national labor statistics bureaus, which rapidly developed during this period, exchanged publications and promoted information on the cooperative and profit-sharing movements. Statistical reports on wages and policy were widely discussed and debated, particularly in relation to the statistical phenomenon of unemployment and issues over wages.

Cultlike groupings played an important role among the early nineteenth century precursors of social science. The followers of Saint-Simon constituted a kind of lifestyle cult, which included ceremonies, hierarchies, a form of religious mysticism, and a kind of church. August Comte, who had been a secretary to Saint-Simon himself, created a similar structure on the basis of his sociology, a religion of humanity with churches (and a factionalized movement) that lasted almost a century. It is a matter of controversy whether one should include such movements as Marxism and psychoanalysis in this category, but they exhibit some common features, notably factionalization. Their claims to expertise were not universally accepted, and they did not become, except in the case of Marxism in the Communist world, academically normalized, but they influenced, as positivism did, many people outside the cult. Similarly, Saint-Simonianism was a major source of French feminism, and Marx's father was a member of a Saint-Simonian reading group.

These movements shade into another important category: reform movements with intellectual leaders, or, in the language introduced by Antonio Gramsci, organic intellectuals, that is to say intellectuals who identify with and intend to express the thoughts and aspirations of a particular social group, such as a class. The nineteenth century produced a vast array of these reform movements, organized to educate, to influence legislation, and often to provide services, including expert advice. The topics ranged from household management and child-rearing to social and economic legislation and prison reform. Among the most intellectually fruitful of these movements were cooperativism, Fabianism, and Henry George's single tax movement, which stimulated discussions of land reform. In the United States, these movements produced such celebrities as Jane Addams, in Britain Beatrice and Sidney Webb. These movements promoted the entry of the social sciences into the university, and for a while maintained a close relation to the newly academicized fields, they eventually came into conflict with them, for reasons that will become clear shortly.

Transitions to Disciplinarity

The academicization of social science was uneven between countries and between disciplines, but for the most part rested on the promise that the development of academic expertise would produce practical results. Much of it built directly on social reform movements. Movements for such things as prison reform and the special treatment of juvenile offenders as well as the regulation of child labor were important elements of the American Social Science Association of the nineteenth century. This body, which had a British analogue, was the precursor to the disciplinary social science bodies that were founded in the United States and elsewhere around the beginning of the twentieth century. These associations promoted the idea that expert knowledge was needed in the operation of state institutions as well as the creation of university social science, and were typically supported by wealthy donors as well as membership fees.

The London School of Economics was a project of the Fabian society, and was needed because of the reluctance of the traditional universities to accept the social sciences. In France, sociology entered the university under the academic heading of pedagogy. Psychology entered under this heading in the United States. Aside from *Staatswissenschaft*, the field with the greatest initial success in entering academic life was German *Nationalekonomie*, which presented itself not as a pure science but as providing the basis and ideology for national economic policy and development. German economists exercised their expertise in various ways, for example, through commissions. Max Sering, Adolph Wagner, and others represented the dominant 'Socialists of the Chair,' and were organized in the *Verein für Sozialpolitik*, which produced policy advice and studies. The problem of agricultural policy and rural life was common to all major countries, and produced a variety of responses in the form of commissions and research projects, as well as policy recommendations. These provided both a basis and transitional stage to the creation of academic policy-oriented fields of agricultural economic and rural sociology, and embedded this form of expertise in agricultural bureaucracies.

The transitional period in which there were a large number of nonacademic expert organizations and a relatively small academic presence of social science varied from country to country. In the United States, there was a vigorous reform community, especially in New York City, that paralleled the Fabian society with which academic social science cooperated, for example, by providing graduates and by participating in public educational efforts (Recchiuti, 2006). These organizations were oriented to the task of public education, and created many novel forms. By the early twentieth century, there were, for example, 'Social Museums' in Paris, Budapest, Munich, Bremen, Charlottenburg, and Frankfurt. In addition, these institutions typically held lecture series. In Britain, Patrick Geddes created 'Outlook Tower' in Edinburgh to educate about the social world; he also put on masques–costume dramas illustrating human progress. At LePlay House in London, the Sociological Society held lectures and exhibited the results of social surveys. In the United States, Harvard's Emerson Hall housed the Peabody Social Museum, and temporary museum-like displays were created for the major social surveys, notably the Pittsburgh and Springfield surveys. There were also social institutes of various kinds in Belgium, Denmark, and Sweden, and a large number of 'societies' with specific problem orientations in Germany.

There were also innovative expert organizational forms, such as the Municipal Research Bureaus in the United States, which supplied objective statistical information on issues related to urban governance, under an ideology of efficiency that had a large international following. But in the 1920s, the reform community was losing direction, having achieved many of its most popular practical goals, and was suffering from its association with prohibition and the temperance movement, one of the most powerful and visible reform movements, which had a vast international following.

The connections were broken in part because of the efforts of the Rockefeller philanthropies, which engaged in massive support for the social sciences. In Europe, the support was largely an effort to better establish the social sciences, but preference was given to work that was 'realistic.' In the United States, the aims were more precise. The aim of the foundations was to improve on the social reform movements of the nineteenth century and the efforts of such reformist foundations as the Russell Sage foundation, which provided reform expertise in a range of areas, by basing them on better data and greater objectivity. The strategy was to first promote more objective social science, then for it to be applied to social and economic problems. In this respect, they were largely disappointed: objective social science did not turn out to be a panacea. Nevertheless, such works as William F. Ogburn's massive *Recent Social Trends* project (1933) did establish objective social science in the public mind as a distinctive form of expert knowledge.

During the 1930s, Franklin Roosevelt assembled a team of expert advisors, known as the 'Brain Trust,' mostly from law schools, to help formulate policy. Among them, however, were academic social scientists, such as Rexford Tugwell, an agricultural economist and advocate of planning, which was a key issue of the decade. A similar role was played by John Maynard Keynes in Britain, who advised the cabinet directly, and in Italy by 'Mussolini's statistician,' Corrado Gini. These prominent figures legitimated the idea of academic expert advisors. In the United States, economic advice was formalized in the Council

of Economic Advisors, and academically trained economists dominated the boards of the Federal Reserve Banks.

The academic social sciences turned away from the reformers in the 1920s and 1930s in the United States, and a model of the role of the social scientific expert that limited the expert to the factual emerged out of this reorientation. One motivation was the experience of the First World War, in which academics in all countries provided intellectual support for their side. In the victorious countries, this produced brutal recriminations after the war, exemplified by Julien Benda's *The Treason of the Intellectuals* (1928[1927]). In the United States, the psychologist James McKeen Cattell was dismissed for his public opposition to the draft and became a hero in the cause of academic freedom, which led to resignations by other prominent Columbia faculty. The propaganda operations of the Office of War Information were retrospectively vilified, and a wave of historical revisionism challenged the assignment of war guilt to the Germans. In Germany itself, there was a broad debate, occasioned in part by Max Weber's address *Science as a Vocation* (2012[1919]), on whether professors should attempt to provide valid worldviews.

This was an important part of the background to developments in which social scientists, especially in the early-and midtwentieth century, sought to free themselves of the accusation that they were merely ideologists or apologists for the state or the established order by identifying a domain of 'science' or expertise that was entirely factual or at least neutral. This claim also facilitated the entry of social sciences into the university as scholarly disciplines, an entry which in many countries was very late and often controversial. Initially, this more limited conception was associated with fact-mongering and criticized as such, but later was given a more coherent defense by George Lundberg in *Can Science Save Us?* (1947) in terms of a version of the ideal of value-free social science, which associated democratic decision making with questions of ends, and made social science into the servant of decision making by supplying the factual basis on which to choose means – a way of making the distinction that recalled Mill and Weber, but could serve as a practical professional ideal for secure academic disciplines. This way of separating science from politics was criticized in such polemics as Robert S. Lynd's *Knowledge for What?* (1939) and Howard S. Becker's speech *Whose Side are We On?* (1967), and by 1960s radicals.

These self-limitations, however, were influential primarily in disciplines formerly associated with reformism, such as sociology, and were less important in such fields as economics, which recognized a normative as well as a positive aspect of disciplinary knowledge, and psychology, which had a clinical side. Management theory, which produced both a practical discipline and one of the most prominent public intellectuals, Peter Drucker, always had a normative or best practices component. International relations as a field has, especially in the United States, been associated with advising, and with powerful political positions in the diplomatic or national security system.

National Differences

Much of the literature on both expertise and public intellectuals focuses on national differences. In Raymond Aron's ironic phrase, France 'passes as the paradise of the intellectuals.' French intellectuals, since the Dreyfusards, are taken to be respected and influential, especially for their political pronouncements, and their collective approval or disapproval, classically in the form of petitions, has been the characteristic mode of action. The French intellectuals of the middle twentieth century are commonly taken as the high water mark of the influence of intellectuals, and this included some social scientists, such as Georges Friedmann and Aron himself, though it is normally philosophers and literary figures, such as Sartre and Camus, who are now regarded as paradigmatic public intellectuals.

This is quite different from possessing influence over ministries as experts, or of exercising control over the professional status and reputation of bureaucrats through professional organizations dominated by academics, or even of being party intellectuals with influence over party members. It has also been observed that the public role of French intellectuals reflects in part the limited possibilities of more direct impact on political and administrative decision making. When French intellectuals are driven to public commentary on politics, the commentary is especially on the morality of political decisions. This in turn points to the more general issue of the ways in which political and administrative structures as well as the receptivity of the relevant publics determines the possible roles of experts.

The case of international relations exemplifies the general principle that the role of social science experts is limited by the bureaucratic structures with which the social scientist interacts: In the United States, there is a 'revolving door' between government and academia and think tanks, while in Central Europe the civil servant tradition is strong, and there is normally very little direct influence on policy making by academic experts. However, there is a larger role of intellectuals in providing political orientation. Britain, which had a particularly insular and strong civil service tradition hostile to social science expertise, had two of the most influential of all social scientists: Keynes in economic policy, whose advice went directly to the cabinet, and R.H. Tawney, who laid the intellectual and moral foundations for the welfare state through his writings and also his engagement with the Christian social and workingmen's organizations. In Germany, intellectuals are widely discussed in newspapers, and are expected to have political viewpoints and to provide worldview-like intellectual orientation to political matters, though are not expected to engage in details of ministerial policy, and normally do not, as in France, engage directly in specifically 'moral' leadership.

Other national differences have significant effects for the way in which intellectuals in general and social scientists in particular interact with the public and with decision makers. The tax law governing foundations in the United States, together with a tradition of private sponsorship, creates a situation in which nominally nonpartisan organizations can seek to influence policy by studies and reports, or by creating an intellectual environment. 'Think tanks' represent a variation on this model. They are funded in order to provide intellectual substance to a point of view or policy orientation favored by the funders. They are normally not associated directly with a social movement, but do seek to influence policy and policy discussions. The private Mt Pelerin Society did this on an international level by developing promarket intellectual orientations to public policy issues in the postwar period, in

response to the intellectual dominance of planning-oriented intellectuals in the 1930s (Mirowski and Plehwe, 2009). Differing legal definitions of freedom of speech have effects as well. Political parties in Europe involve paid memberships, and generally exclude large donors, and political messages to the public are driven by parties, and think tanks normally receive some government support, whereas in the United States, they are normally privately funded.

See also: Comte, Auguste (1798–1857); Keynes, John Maynard (1883–1946); Keynesianism and the Question of State Interventionism; Positivism, History of; Positivism, Sociological; Science and Technology Studies: Experts and Expertise; Smith, Adam (1723–90); Vienna Circle: Logical Empiricism; Weber, Max (1864–1920); Weberian Social Thought, History of; von Hayek, Friedrich A. (1899–1992).

Bibliography

Becker, Howard S., 1967. Whose side are we on? Social Problems 14 (3), 239–247. http://www.jstor.org/stable/799147 (accessed 09.05.13.).

Benda, Julien, 1928[1927]. The Treason of the Intellectuals (Richard Aldington, Trans.). William Morrow & Company, New York.

Bliss, William D.P., 1897. The Encyclopedia of Social Reform. Funk & Wagnalls Company, New York & London. http://archive.org/details/encyclopediofsoc00blisrich (Internet Archive. Accessed 13.05.13.).

Buckle, Henry, 1864. History of Civilization in England, second ed. Appleton, New York.

Collini, Stefan, 2006. Absent Minds: Intellectuals in Britain. Oxford University Press, Oxford.

Debray, Régis, 1981. Teachers, Writers, Celebrities: The Intellectuals of Modern France (David Macey, Trans.). New Left Books, London.

Jacoby, Russell, 2000[1987]. The Last Intellectuals: American Culture in the Age of Academe. Basic Books, New York.

Judt, Tony, 1992. Past Imperfect: French Intellectuals, 1944–1956. The University of California Press, Berkeley, CA.

Lindenfeld, David, 1997. The Practical Imagination: The German Sciences of State in the Nineteenth Century. The University of Chicago Press, Chicago, IL.

Lundberg, George, 1947. Can Science Save Us? Longmans, Green & Company, New York.

Lynd, Robert, 1939. Knowledge for What? The Place of Social Science in American Culture. Princeton University Press, Princeton, NJ.

Mill, John S., 1882. A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, eighth ed. Harper & Brothers Publishers, New York. http://www.gutenberg.org/ebooks/27942 (accessed 09.05.13.).

Mirowski, Paul, Plehwe, Dieter (Eds.), 2009. The Road from Mont Pèlerin: The Making of the Neoliberal Thought Collective. Harvard University Press, Cambridge, MA.

Ogburn, William F., 1933. Recent Social Trends in the United States; Report of the President's Research Committee on Social Trends. McGraw Hill, New York & London. http://archive.org/stream/recentsocialtren01unitrich/recentsocialtren01unitrich_djvu.txt (Internet Archive. Accessed 13.05.13.).

Posner, Richard, 2001. Public Intellectuals: A Study of Decline. Harvard University Press, Cambridge, MA.

Recchiuti, John L., 2006. Civic Engagement: Social Science and Progressive-era Reform in New York City. University of Pennsylvania Press, Philadelphia, PA.

Turner, Stephen, 2001. What is the problem with experts? Social Studies of Science 31 (1), 123–149.

Weber, Max, 2012[1919]. Science as a profession and vocation (H.H. Bruun, Trans.). In: Bruun, Hans Henrik, Whimster, Sam (Eds.), Max Weber: Collected Methodological Writing. Routledge, London, pp. 335–353.

Social Simulation: Computational Models

Michael W Macy, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article compares computational and mathematical models and reviews three phases in social simulation - dynamical systems, microsimulation, and agent-based modeling. The latter can be used to identify possible micro-level mechanisms that might account for puzzling empirical patterns and to identify population-level implications of a set of behavioral and structural assumptions about individual interaction. Classic examples include Schelling's model of residential segregation and Axelrod's model of the evolution of cooperation. The article concludes with an assessment of important criticisms.

Computational social simulation involves the use of computer algorithms to model social processes. These algorithms require greater precision and logical rigor than natural language yet lack the generality of closed-form mathematical equations. The use of computer simulation by social scientists has increased in recent years, due to growing interest in modeling nonlinear dynamical processes using graphical interfaces that allow highly intuitive visual representations of the results. Social simulation has also undergone a qualitative change as the emphasis has shifted from prediction to explanation.

This article reviews three 'waves' of innovation: dynamical systems, *microsimulation*, and adaptive agent models (Gilbert and Troitzsch, 1999). These three approaches will be summarized as follows:

1. 1960s: Dynamical systems simulations were empirically grounded, inductive, holistic, and functionalistic.
2. 1970s: Microsimulation introduced the use of individuals as the units of analysis but retained the earlier emphasis on empirically based macrolevel prediction.
3. 1980s: Adaptive agent models revolutionized computational social science by simulating actors not factors (system attributes). These 'bottom-up' models explored interactions among purposive decision makers.

Although these three 'waves' overlap (e.g., economists continue to use holistic simulation of productive factors, microsimulation was invented in 1957, and nascent agent-based models appeared in the late 1960s), most of the recent excitement in computational modeling in the social and behavioral sciences has centered on agent-based simulation. So too has much of the controversy.

Dynamical Systems

The first wave of computer simulations in social science occurred in the 1960s (Cyert and March 1963). These studies used computers to simulate dynamical systems, such as control and feedback processes in organizations, industries, cities, and even global populations. The models typically consist of sets of differential equations that describe changes in system attributes as a function of other systemic changes. Applications included the flow of raw materials in a factory, inventory control in a warehouse, urban traffic, military supply lines, demographic changes in a world system, and ecological limits to growth (Forrester, 1971; Meadows et al., 1974).

Models of dynamical systems are functional and holistic (Gilbert and Troitzsch, 1999). *Functional* means that theoretical interest centers on system *equilibrium* or balance among interdependent input–output units. *Holistic* means that the system is modeled as an irreducible and indivisible entity. Although these models allow nested organizational levels, at any given level, they assume that sets of related attributes (such as economic growth, suburban migration, and traffic congestion) are causally linked at the aggregate level.

Microsimulation

The second wave of computational modeling in the social sciences developed in the late 1970s. Known as microsimulation, it represents the first step in the progression from factors to actors. In striking contrast to the earlier holistic approach,

> microsimulation is a 'bottom-up' strategy for modeling the interacting behavior of decision-makers (such as individuals, families, and firms) within a larger system. This modeling strategy utilizes data on representative samples of decision-makers, along with equations and algorithms representing behavioral processes, to simulate the evolution through time of each decision-maker, and hence of the entire population of decision-makers. (Caldwell, 1997)

For example, Caldwell and Keister use CORSIM, a large-scale dynamic microsimulation model of the US population of individuals and families, to integrate individual and family level wealth behavior with aggregate-level stratification outcomes. This microanalytic approach allows researchers to avoid a serious limitation in the earlier generation of simulations: population homogeneity. For example, the relationship between wealth and age may not be uniform across ethnic or religious subcultures, geographic regions, or birth cohorts. Complex interactions across subpopulations are lost when the focus is on the correlation among aggregate factors. This precludes the ability to predict the effects of policy changes that impact only certain groups.

Microsimulation solves this problem by modeling individual-level attributes as well as population-level vectors.

The database structure consists of individual records for a representative sample of a population, with a set of individual attributes measured at multiple points in time. These individuals are then 'aged' by updating their attributes, based on a set of empirically derived state-transition probabilities (e.g., an individual's move from work to retirement at age 65). These probabilities are then used to predict new states of each individual, and these are then aggregated to make population projections.

Agent-Based Models: The Third Wave

The third wave, agent-based models, attracted widespread interest beginning in the 1980s. These models extend the microanalytical approach by deriving global states from individual *behaviors*, not individual attributes. In microsimulation models, individuals are inert in two ways: (1) the state of an individual is a numerical function of other traits and (2) individuals do not interact (Gilbert and Troitzsch, 1999: p. 8). Agent-based models assume an individual has intentions or goals and makes choices that affect other agents whose choices in turn affect that individual. These models impose four key assumptions:

1. Agents interact with little or no central authority or direction. Global patterns emerge from the bottom up, determined not by a centralized authority but by local interactions among autonomous decision makers.
2. Decision makers are adaptive rather than optimizing, with decisions based on heuristics, not on calculations of the most efficient action (Holland, 1995: p. 43). These heuristics are implemented as logical rules and correspond to individual habits, social norms, organizational protocols, rituals, conventions, customs, and bureaucratic routines. They evolve at two levels, the individual and the population. Individual learning alters the probability distribution of rules competing for attention, through processes like reinforcement, Bayesian updating, or the backpropagation of error in artificial neural networks. Population learning alters the frequency distribution of rules competing for reproduction through processes of selection, imitation, and social influence (Latané, 1996). Genetic algorithms (Holland, 1995) are widely used to model adaptation at the population level.
3. Decision makers are strategically interdependent. Strategic interdependence means that the consequences of each agent's decisions depend in part on the choices of others. When strategically interdependent agents are also adaptive, the focal agent's decisions influence the behavior of other agents who in turn influence the focal agent, generating a complex adaptive system (Holland, 1995: p. 10).
4. Agents interact locally. Computational models of dynamical systems, as well as game-theoretic models of strategic interdependence, typically assume a population in which each individual is connected either to everyone or to a randomly distributed subset. Although there are few, if any, historical examples of social networks that are either random or fully connected, this assumption is nevertheless appealing because it allows analysts to derive closed form mathematical solutions. The few models that have incorporated local structure have assumed that social ties are exogenously determined and do not change over time.

Agent-based models have for the first time opened up applications of computational modeling to the study of social life embedded in dynamic networks.

Examples

Thomas Schelling's (1971) 'neighborhood segregation' model was one of the earliest agent-based social simulations. Schelling used an impressively simple and elegant model to address a perplexing empirical puzzle. The 1964 Fair Housing Act outlawed racial discrimination in the residential market, and surveys since that time show a progressive increase in tolerance of outgroup neighbors. Yet urban residential segregation has hardly changed. Schelling's model provides a compelling explanation. Consider a residential area that is highly segregated, such that the number of neighbors with different cultural markers (such as ethnicity) is at a minimum. If the aggregate pattern were assumed to reflect the attitudes of the constituent individuals, one might conclude from the distribution that the population was highly parochial and intolerant of diversity. Yet Schelling's simulation shows that this need not be the case. His 'tipping' model shows that highly segregated neighborhoods can form even in a population that tolerates diversity. The aggregate pattern of segregation is an emergent property that does not reflect the underlying attitudes of the constituent individuals.

Another classic is Axelrod's 1987 model of the evolution of cooperation, which addressed another empirical puzzle: how is it possible for cooperation to evolve, given the reproductive advantage to those who exploit the generosity of others? Axelrod used a simple agent-based model to study the viability of cooperation in a 'prisoner's dilemma,' a two-player game in which choices that are individually rational aggregate into collective outcomes that are suboptimal for both players. The results showed that strategies based on reciprocity (such as 'tit for tat') tend to be very successful. The secret of their success is that they perform well against copies of themselves without being vulnerable to predation. Axelrod showed how a simple rule of 'tit for tat' can gain a toehold even in a population of predators and go on to thrive.

Later Axelrod (1997) used computational models that couple homophily (likes attract) and social influence to explain the robustness of cultural diversity despite the relentless pressures of assimilation. If local similarity facilitates interaction, which in turn facilitates influence, it is indeed curious that the global outcome is diversity, not uniformity. The process resembles biological speciation in which initially similar organisms distributed over a large surface become sufficiently different from those in distant regions that they can no longer mate should they encounter one another through migration.

Although Axelrod characterized these models as 'evolutionary,' the agents were unable to learn or discover new strategies. Later models added the ability to learn new rules. The 'genetic algorithm' invented by John Holland is a simple

and elegant way to model strategies that can progressively improve performance by building on partial solutions. Each strategy consists of a string of symbols that code behavioral instructions, analogous to a chromosome containing multiple genes. The string's instructions affect the agent's reproductive fitness and hence the probability that the strategy will propagate. Propagation occurs when two or more mated strategies recombine. If different rules are each effective, but in different ways, recombination allows them to create an entirely new strategy that may integrate the best abilities of each 'parent' and thus eventually displace the parent rules in the population of strategies.

Agent-based models can implement adaptation at the individual level as well. While evolution alters the frequency distribution of strategies at the population level, learning alters the probability distribution of strategies within the choice set of an individual agent. For example, artificial neural networks are simple self-programmable devices that model agents who learn through reinforcement. Like genetic algorithms, neural nets have a biological analog, in this case, the nerve systems of living organisms. The device consists of a web of neuronlike units (or neurodes) that fire when triggered by impulses of sufficient strength, and in turn stimulate other units when fired. The magnitude of an impulse depends on the strength of the connection (or synapses) between the two neurodes. The network learns by modifying these path coefficients in response to environmental feedback about its performance. Neural nets have been used to study the evolution of religion (Bainbridge, 1995), kin altruism (Parisi et al., 1995), the emergence of status inequality (Vakas-Duong and Reilly, 1995), group dynamics (Nowak and Vallacher, 1997), and social deviance (Kitts et al., 1999).

Criticisms of Simulation

Agent-based simulations have been criticized as unrealistic, tautological, and unreliable.

1. Simulations, like mathematical models, tell us about a highly stylized and abstract world, not about the world in which we live. In particular, agent-based simulations rely on highly simplified models of rule-based human behavior that neglect the complexity of human cognition.
2. Like mathematical models, simulations cannot tell us anything that we do not already know, given the assumptions built into the model and the inability of a computer to do anything other than execute the instructions given to it by the program.
3. Unlike mathematical models, simulations are numerical and therefore cannot establish lawful regularities or generalizations.

The defense against these criticisms centers on the principles of complexity and emergence. How can simple agent-based models explain the rich complexity of social life? In complex systems, very simple rules of interaction can produce highly complex global patterns. "Human beings," Simon contends (1998: p. 53), "viewed as behaving systems, are quite simple." We follow rules, in the form of norms, conventions, protocols, moral and social habits, and heuristics. Although the rules may be quite simple, they can produce global patterns that may not be at all obvious and are very difficult to understand. Hence, "the apparent complexity of our behavior is largely a reflection of the complexity of the environment," including the complexity of interactions among strategically interdependent, adaptive agents. The simulation of artificial worlds allows us to explore the complexity of the social environment by removing the cognitive complexity (and idiosyncrasy) of constituent individuals.

The Exploration of Artificial Worlds

If Simon, Axelrod, and the complexity theorists are right, then the artificiality of agent-based models is a virtue, not a vice. When simulation is used to make predictions or for training personnel (e.g., flight simulators), the assumptions need to be highly realistic, which usually means they will also be highly complicated (Axelrod, 1997: p. 5). "But if the goal is to deepen our understanding of some fundamental process," Axelrod continues, "then simplicity of the assumptions is important and realistic representation of all the details of a particular setting is not." As such, the purpose of these models is to generate hypotheses, not to test them (Prietula et al., 1998: p. xv).

Holland (1995: p. 146) offers a classic example of the problem of building theories that too closely resemble actuality: Aristotle's mistaken conclusion that all bodies come to rest, based on an observation of a world that happens to be infested with friction. Had these observations been made in a frictionless world, Aristotle would have come to the same conclusion reached by Newton and formulated as the principle that bodies in motion persist in that motion unless perturbed. Newton avoided Aristotle's error by studying what happens in an artificial world in which friction had been assumed away. Ironically, "Aristotle's model, though closer to everyday observations, clouded studies of the natural world for almost two millennia" (Holland, 1995: p. 146). More generally, it is often necessary to step back from our world in order to see it more clearly.

For example, Schelling's neighborhood segregation model made no effort to imitate actuality or to resemble an observed city. Rather, the 'residents' live in a highly abstract cellular world without real streets, rivers, train tracks, zoning laws, redlining, housing markets, and so on. This artificial world shows that all neighborhoods in which residents are able to move have an underlying tendency toward segregation even when residents are moderately tolerant. This hypothesis can then be tested in observed neighborhoods, but even if observations do not confirm the predicted segregation, this does not detract from the value of the simulation in suggesting empirical conditions that were not present in the model that might account for the discrepancy with the observed outcome.

In short, agent-based models "have a role similar to mathematical theory, shearing away detail and illuminating crucial features in a rigorous context" (Holland, 1995: p. 100). Nevertheless, Holland reminds us that we must be careful. Unlike laboratory experiments, computational 'thought experiments' are not constrained by physical actuality and thus 'can be as fanciful as desired or accidentally permitted.'

Unwrapping and Emergence

A second criticism of agent-based modeling addresses a problem known as 'unwrapping.' Unwrapping occurs "when the 'solution' is explicitly built into the program [such that] the simulation reveals little that is new or unexpected" (Holland, 1995: p. 137). Critics charge that all computer simulations share this limitation, due to the inability of computers to act in any way other than how they were programmed to behave.

This criticism overlooks the distinction between the microlevel programming of the agents and the macrolevel patterns of interaction that emerge. These patterns are 'built in' from the outset yet the exercise is valuable because the logical implications of behavioral assumptions are not always readily apparent. Indeed, a properly designed computational experiment can yield highly counterintuitive and surprising results. This theoretical possibility is based on the principle of *emergence* in complex systems. Biological examples of emergence include life that emerges from nonliving organic compounds and intelligence and consciousness that emerge from dense networks of very simple switchlike neurons. Schelling's (1971) neighborhood model provides a compelling example from social life: segregation is an emergent property of social interaction that is not reducible to individual intolerance.

Rationality and Adaptive Behavior

A third criticism is that simulation models, unlike mathematical models, are numerical, not deductive, and therefore cannot be used to form generalizations. Worse still, simulations of evolutionary processes can be highly sensitive to the basin of attraction in which the parameters are arbitrarily initialized (Binmore, 1998). This problem is especially acute if the model is deterministic. For example, Axelrod's (1997) model of cultural assimilation appears to demonstrate global diversity, but later work showed that the outcome was a local equilibrium that could be disturbed by very small amount of random cultural mutation (Klemm et al., 2003).

However, game-theoretic mathematical models pay a high price for the ability to generate deductive conclusions: multiple equilibria that preclude a uniquely rational solution. Equilibrium selection requires constraints on the perfect rationality of the agents. Simon appreciates the paradox: "Game theory's most valuable contribution has been to show that rationality is effectively undefinable when competitive actors have unlimited computational capabilities for outguessing each other, but that problem does not arise as acutely in a world, like the real world, of *bounded rationality*" (1998: p. 38). But bounded rationality often makes analytical solutions mathematically intractable. "When the agents use adaptive rather than optimizing strategies, deducing the consequences is often impossible: simulation becomes necessary" (Axelrod, 1997: p. 4). In sum, simulation requires sensitivity analysis and tests of robustness, but properly implemented, it not only offers a solution to the problem of equilibrium selection but can also tell us something about how evolutionary systems move from one equilibrium to another.

Conclusion

From a narrowly positivist perspective, agent-based artificial worlds might seem unrealistic, unfalsifiable, and unreliable, especially when contrasted with earlier approaches to simulation that were data driven, predictive, and highly realistic (e.g., flight simulators). However, agent-based models may be essential for the study of open and complex adaptive systems, whether social, psychological, or biological. When used as an exploratory tool, 'bottom-up' simulation does not test theory against observation. Rather, these thought experiments suggest possible mechanisms that may generate puzzling empirical patterns.

A key theme that runs through these puzzles is that it is often very difficult to recognize the underlying causal mechanisms using conventional data-analytic methods. As Holland concludes (1995: p. 195), "I do not think we will understand morphogenesis, or the emergence of organizations like Adam Smith's pin factory, or the richness of interactions in a tropical forest, without the help of such models."

See also: Markov Decision Processes; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources.

Bibliography

Axelrod, R., 1997. The Complexity of Cooperation. Princeton University Press, Princeton, NJ.

Bainbridge, W., 1995. Neural network models of religious belief. Sociological Perspectives 38 (4), 483–496.

Binmore, K., 1998. Axelrod's The complexity of cooperation. The Journal of Artificial Societies and Social Simulation 1, 1.

Caldwell, S., 1997. Dynamic Microsimulation and the Corsim 3.0 Model. Strategic Forecasting, Ithaca, NY.

Carley, K., 1991. A theory of group stability. American Sociological Review 56, 331–354.

Cyert, R., March, J.G., 1963. A Behavioral Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Forrester, J.W., 1971. World Dynamics. MIT Press, Cambridge, MA.

Gilbert, N., Troitzsch, K., 1999. Simulation for the Social Scientist. Open University Press, Buckingham, UK.

Holland, J., 1995. Hidden Order: How Adaptation Builds Complexity. Perseus, Reading, MA.

Kitts, J., Macy, M., Flache, A., 1999. Structural learning: attraction and conformity in task-oriented groups. Computational and Mathematical Organization Theory 5, 129–145.

Klemm, K., Eguiluz, V., Toral, R., San Miguel, M., 2003. Role of dimensionality in Axelrod's model for the dissemination of culture. Physica A: Statistical Mechanics and Its Applications 327 (1–2), 1–5.

Latané, B., 1996. Dynamic social impact: Robust predictions from simple theory. In: Hegselmann, R., Mueller, U., Troitzsch, K. (Eds.), Modeling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Kluwer Dorderecht, Boston, MA, pp. 287–310.

Macy, M., 1995. Natural selection and social learning in prisoner's dilemma: co-adaptation with genetic algorithms and artificial neural networks. Sociological Methods and Research 25, 103–137.

Meadows, D.L., Behrens III, W.W., Meadows, D.H., Naill, R.F., Randers, J., Zahn, E.K., 1974. The Dynamics of Growth in a Finite World. MIT Press, Cambridge, MA.

Nowak, A., Vallacher, R., 1997. Computational social psychology: cellular automata and neural network models of interpersonal dynamics. In: Read, S., Miller, L. (Eds.), Connectionist Models of Social Reasoning and Social Behavior. Erlbaum, Mahwah, NJ.

Nowak, M., Sigmund, K., 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoners' dilemma game. Nature 364, 56–57.
Parisi, D., Cecconi, F., Cerini, A., 1995. Kin-directed altruism and attachment behaviour in an evoking population of neural networks. In: Gilbert, R.C., Coute (Eds.), Artificial Societies. University College London Press, London.
Prietula, M., Carley, K., Gasser, L., 1998. Simulating Organizations: Computational Models of Institutions and Groups. MIT Press, Cambridge, MA.
Schelling, T., 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1, 143–186.
Simon, H., 1998. The Sciences of the Artificial. MIT Press, Cambridge, MA.
Vakas-Duong, D., Reilly, K., 1995. A system of IAC neural networks as the basis for self-organization in a sociological dynamical system simulation. Behavioral Science 40, 275–303.

Social Stratification

David B Grusky, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In all known societies, the most important valued goods (e.g., income, power) are distributed unequally, with the privileged individuals and families enjoying a disproportionate share of them. The term 'stratification system' refers to the institutions that generate observed inequalities of this sort. The main components of such systems are (1) the institutional processes that define certain goods as valuable and desirable, (2) the rules of allocation that distribute these goods across the division of labor, and (3) the mobility mechanisms that link individuals to positions and thereby generate unequal control over valued resources.

The dispassionate observer of the last half-century of social stratification research might easily characterize the field as one of the great success stories of modern social science. The breakthroughs include, for example, (1) building a sophisticated infrastructure for monitoring key trends in poverty, segregation, income inequality, racial and gender inequality, and other labor market outcomes; (2) making real progress in addressing long-standing debates about how much labor market discrimination there is, whether residential segregation matters, how much social mobility there is, and whether social classes are disappearing; and (3) even making headway in understanding the deeper causes of poverty, mobility, and many types of inequality.

Although the progress is considerable, the field's shortcomings are also readily apparent. The most prominent failures have come about when scholars have turned to the idiographic task of understanding the dynamics of change. The spectacular takeoff in income inequality, one of the most consequential developments of our time, went largely unpredicted. The intransigence and resilience of school segregation was clearly not anticipated by many social scientists. The sudden 'stalling out' of long-standing declines in gender inequality remains poorly understood. The equally sudden evaporation of opposition to gay marriage and other type of gay rights would have seemed unimaginable just 5 years ago. The lightning-fast rise and fall of one of the first anti-inequality movements of contemporary US history (i.e., the Occupy movement) is likewise a puzzle.

The objective behind this article is to provide an introduction to the types of questions, both nomothetic and idiographic, that have been and continue to be taken on in the field of poverty and inequality. It would be impossible to provide some encyclopedic introduction to all that is known in this sprawling field. It is tractable, however, to present the *types* of questions that have been taken on, with a special emphasis on understanding how those questions have changed and evolved. Before turning to this task, it will be helpful to introduce some of the basic concepts used to describe the structure of inequality, the persistence of inequality, and the effects of inequality on social behavior.

Basic Concepts

The key components of stratification systems are (1) the institutional processes that define certain types of goods as valuable and desirable, (2) the rules of allocation that distribute these goods across various jobs or occupations in the division of labor (e.g., doctor, farmer, homemaker), and (3) the mobility mechanisms that link individuals to jobs and thereby generate unequal control over valued resources. It follows that inequality is produced by two types of matching processes: The social roles in society are first matched to 'reward packages' of unequal value, and individual members of society are then allocated to the positions so defined and rewarded. In all societies, there is a constant flux of occupational incumbents as newcomers enter the labor force and replace dying, retiring, or out-migrating workers, yet the positions themselves and the reward packages attached to them typically change only gradually. As Schumpeter (1953) puts it, the occupational structure can be seen as "a hotel … which is always occupied, but always by different persons."

What types of rewards are distributed via these two matching processes? It is increasingly fashionable to recognize that inequality is 'multidimensional,' that income is accordingly only one of many important resources, and that income redistribution in and of itself would not eliminate inequality (e.g., Sen, 1997). When a multidimensionalist approach is accordingly taken, one might usefully distinguish among the eight types of assets listed in **Table 1**, each understood as valuable in its own right rather than a mere investment item. The assets of **Table 1** may of course serve investment as well as consumption functions. For example, most economists regard schooling as an investment that generates future streams of income, while some sociologists likewise regard social networks as forms of capital that can be parlayed into educational credentials, income, and other valued goods. While most of the assets listed in **Table 1** are clearly convertible in this fashion, they are not necessarily regarded as investments by the individuals involved. In fact, many valuable assets can be secured at birth or through childhood socialization (e.g., the 'good manners' of the aristocracy), and they are therefore acquired without the beneficiaries explicitly weighing the costs of acquisition against the benefits of future returns.

The implicit claim underlying **Table 1** is that the listed assets exhaust all possible consumption goods and, as such, constitute the raw materials of stratification systems. The stratification field has developed a vocabulary that describes how these raw materials are distributed across members of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32137-7

Table 1 Types of assets and examples of advantaged and disadvantaged groups

Assets		*Examples*	
Asset group	*Examples of types*	*Advantaged*	*Disadvantaged*
1. Economic	Wealth	Billionaire	Bankrupt worker
	Income	Professional	Laborer
	Ownership	Capitalist	Worker (i.e., employed)
2. Power	Political power	Prime minister	Disenfranchised person
	Workplace authority	Manager	Subordinate worker
	Household authority	Head of household	Child
3. Cultural	Knowledge	Intelligentsia	Uneducated persons
	Digital culture	Silicon Valley residents	Residents of other places
	'Good' manners	Aristocracy	Commoners
4. Social	Social clubs	Country club member	Nonmember
	Workplace associations	Union member	Nonmember
	Informal networks	Washington A-list	Social unknown
5. Honorific	Occupational	Judge	Garbage collector
	Religious	Saint	Excommunicate
	Merit-based	Nobel Prize winner	Nonwinner
6. Civil	Right to work	Citizen	Unauthorized immigrant
	Due process	Citizen	Suspected terrorist
	Franchise	Citizen	Felon
7. Human	On-the-job training	Experienced worker	Inexperienced worker
	General schooling	College graduate	High school dropout
	Vocational training	Law school graduate	Unskilled worker
8. Physical	Mortality	Person with long life	A 'premature' death
	Physical disease	Healthy person	Person with AIDS, asthma
	Mental health	Healthy person	Depressed, alienated

population. The main parameters of interest pertain to (1) the overall amount of inequality, (2) the extent to which individuals are locked permanently into certain positions (i.e., rigidity), (3) the extent to which ascriptive traits fixed at birth are used for purposes of allocation, and (4) the degree to which the various dimensions of inequality cohere (or are 'crystallized'). These four types of parameters are reviewed below.

The overall *amount of inequality* in any given resource (e.g., income) may be defined as its dispersion or concentration among the individuals in the population. Although many scholars seek to characterize the overall level of societal inequality with a single parameter, such attempts will obviously be compromised insofar as some types of rewards are distributed more equally than others. This complexity clearly arises in the case of contemporary inequality regimes. The recent diffusion of 'citizenship rights,' for example, implies that civil goods have come to be widely dispersed across all citizens, whereas economic and political goods continue to be disproportionately controlled by a relatively small elite.

The *rigidity* of a stratification system is indexed by the continuity (over time) in the social standing of its members. The stratification system is highly rigid, for example, if the current wealth, power, or prestige of individuals can be accurately predicted on the basis of their prior statuses or those of their parents. It should again be emphasized that such rigidity (or 'social closure') will typically vary in degree across the different types of assets listed in Table 1.

The stratification system rests on *ascriptive processes* to the extent that conditions present at birth (e.g., parental wealth, sex, race) influence the subsequent social standing of individuals. If ascriptive processes of this sort are in operation, it is possible (but by no means guaranteed) that the underlying traits themselves will become bases for group formation and collective action (e.g., race riots, feminist movements). In contemporary societies, ascription of all kinds is usually seen as undesirable or discriminatory, and much governmental policy (e.g., antidiscrimination legislation) is therefore directed toward fashioning a stratification system in which individuals acquire resources solely by virtue of their achievements.

The degree of *crystallization* is indexed by the correlations among the various resources (e.g., income, wealth, education) that are socially valued. If these correlations are strong, then the same individuals (i.e., the 'upper class') will consistently appear at the top of all hierarchies, while other individuals (i.e., the 'lower class') will consistently appear at the bottom of the stratification system. By contrast, various types of status inconsistencies (e.g., a poorly educated millionaire) will emerge in stratification systems with weakly correlated hierarchies, and it is correspondingly difficult in such systems to define a unitary set of classes that have predictive power with respect to all resources.

Although the preceding terms are fundamental to the field, there are of course a great many others, some of which will be introduced below. The following sections lay out the nine subfields around which the field of inequality has been organized.

The Functions and Dysfunctions of Inequality

In economics and sociology alike, there is a long tradition of theorizing that emphasizes that some forms of inequality

have the positive effect of (1) inducing the most qualified workers to take on the important jobs and (2) incentivizing people to work hard and innovate. The famous trade-off thesis implies that a taste for equality is exercised at the cost of reducing output and potentially rendering all worse off. We cannot, as Okun (1975) so cleverly put it, "have our cake of market efficiency and share it equally." Although some scholars have taken the trade-off thesis as an article of faith (e.g., Davis and Moore, 1945), others treat it as testable by asking whether groups or societies with higher inequality are indeed more productive.

The trade-off thesis of course implies that inequality is functional only insofar as it delivers the proper incentives. There are many ways in which high-inequality regimes may not be incentivizing all people to work hard in ways that then increase total output. If, for example, inequality takes a simple winner-take-all form, it could be counterproductive and serve to reduce output. The relevant thought experiment here is that of the golf tournament in which all prize money is allocated exclusively to the winner (e.g., Freeman, 2012). In this setup, mediocre golfers who never had a chance of winning lack any incentive to exert themselves, with the implication that total output (as measured, say, by total strokes) will diminish. There are, in other words, certain forms of extreme inequality that may not succeed in creating the incentives that the trade-off thesis presumes. Likewise, insofar as inequality arises because of corruption and rent-seeking behavior, it is not contributing to overall output and cannot be justified in functionalist terms. There is presumably no collective benefit, for example, to allowing a chief executive officer to pack the board with cronies simply because they will then vote him or her a generous compensation package (see DiPrete et al., 2014). The current debate about the importance of rent and corruption in generating the takeoff in income inequality (e.g., Grusky, 2012) is motivated by precisely this interest in assessing whether extreme inequality of the sort observed today can be understood as efficient or functional.

Inequality in Comparative Perspective

If inequality is a ubiquitous feature of human society, it nonetheless has taken on many different forms. There is a long tradition of stratification research examining the historical and cross-national variability in the amount of inequality, the types of assets that underlie inequality (e.g., power, prestige, economic resources), the extent to which opportunities vary with social origins and other ascriptive traits, and the ideologies by which prevailing inequalities are legitimated. Although the differences across tribal, feudal, caste, socialist, and capitalist forms of inequality are especially stark on these and other dimensions (see Grusky and Weisshaar, 2014), even within contemporary capitalist societies the institutional variability is arguably impressive (see Esping-Andersen and Myles, 2008). In popular commentary, it is sometimes assumed that market economies are associated with a single characteristic form of inequality, but in fact the evidence clearly reveals the wide variety of ways that inequality has been and continues to be organized within market economies.

The same lesson is readily gleaned by examining how income inequality has evolved over the last century within market societies. As Atkinson et al. (2011) show, the trend in income inequality follows a U-shaped form in the English-speaking countries (i.e., the US, Canada, Ireland, the UK, Australia, and New Zealand), all of which run so-called liberal market economies. In each of these countries, we see inequality dropping precipitously in the late 1920s and during World War II, stabilizing at a comparatively low level over the next 30 years, and then taking off in the 1970s and ultimately returning to or approaching the extreme levels that prevailed in the 1920s. By contrast, a rather different pattern emerges for the Central European countries, such as France and Germany, that are running so-called coordinated market economies. Although these countries also underwent a real decline in inequality during the first half of the twentieth century, that decline is not followed by any subsequent rebound in inequality of the sort found in the US and the other English-speaking countries. Instead, one finds a rough stability in the amount of inequality or, in a few cases, even a continuing slight decline. Over the last quarter century, there has been much research on why some countries are experiencing a spectacular takeoff in income inequality, whereas other countries are experiencing a slow takeoff, a delayed takeoff, or none at all.

The Structure of Inequality

Because so much social science research is now focused on the takeoff in inequality, we have a growing body of knowledge on what is driving it (e.g., Goldin and Katz, 2008) and who benefited most from it (e.g., Hacker and Pierson, 2010). We know less, however, about the form that inequality takes. This is surprising because some of the most contentious debates within sociology are about the form rather than extent of inequality. Although most sociologists assume that inequality is organized around big social classes (e.g., capitalists, managers, craft workers) that affect the politics, interests, and life chances of their members, there is much debate about (1) exactly how big such classes are, (2) how best to draw the boundaries between them, and (3) what types of assets underlie them.

Because class models are discussed in detail in the article on social class, it is unnecessary to review them in much detail here. It is worth pointing out, however, that these debates about the form of inequality, although pursued rather obsessively in some quarters, do nonetheless speak to issues of some consequence. If the objective, for example, is to understand the experience of poverty, one should care whether those with low income tend also to have very little in the way of other resources, such as wealth or education. We should presumably be more troubled by a form of poverty in which many types of disadvantage (e.g., income, wealth, education) tend to come together. The form that inequality takes will also likely affect how it is experienced and whether it is well tolerated. In a highly crystallized system, an economic crisis might generate a resurgence of class-based rhetoric, with the sharpest vitriol aimed at a culturally homogenous upper class with distinctive lifestyles, attitudes, and behaviors. In a less crystallized system, the same economic crisis might generate a more fragmented

political discourse, perhaps equally contentious in its own way but with little coherence across issues and a weaker 'us versus them' rhetoric.

The main alternatives to class models are those that represent the stratification system in gradational terms. The gradational formulation, which was especially popular in the 1970s and 1980s, represents inequality either as a hierarchy of *individuals* differing in their income, wealth, or education or a hierarchy of *occupations* differing in their prestige or socioeconomic status. Although gradational models based on occupational prestige or socioeconomic status have largely disappeared from the day-to-day practice of inequality research, gradational models based on individual-level income, wealth, and education remain very popular. The demise of scales based on occupational prestige or socioeconomic status seems as much due to the inevitable half-life of any research tradition as to any hard empirical evidence that this form of gradationalism is any less important now than in the past.

The Ruling Class, Elites, and the Upper Class

There is a long tradition of stratification research exploring how power, privilege, and money may come together to form an elite at the top of the social hierarchy. In recent years, the 'one percent' has become the modern-day label for the elite, but few commentators would probably seriously believe that an arbitrary income threshold of that sort captures very well the structure of power and advantage at the top.

The key question within this research tradition is whether late industrial societies tend to generate a powerful and solidary upper class. The 'power elite' narrative (see Mills, 1956) has it that economic, political, and military elites form a unitary class by virtue of coresiding in the same neighborhoods, belonging to the same clubs, sharing a coherent political ideology, and interacting as approximate equals. This small elite occupies positions of considerable power and can therefore "mightily affect the everyday worlds of ordinary men and women" (Mills, 1956). As against this account, we might imagine that capitalists, by virtue of their economic power, have become the *de facto* 'ruling intellectual force' (Marx, 1970). Although top leadership positions are of course formally filled by political or military elites, it is hardly radical any more to suggest that they are acting largely at the behest of capitalists and corporations (who wield influence via lobbying, campaign funding, and the like).

More recently, Gouldner (1982) has suggested that a third claimant to power, the 'knowledge class,' is in ascendancy as expert knowledge comes to be understood as a source of value and a prerequisite for responsible and successful leadership (also Brooks, 2000). For new class theorists, the main struggle of the contemporary period is not between capitalists and workers (as workers long ago lost out), rather it is found in the ongoing 'civil war' (Gouldner, 1982) between capitalists with the power to buy influence and intellectuals who claim to have the knowledge to successfully wield it. The outcome of this civil war is far from obvious: If capitalists bring ever-growing economic power to the table, intellectuals rest their case on the efficiency of relying on results-driven scientific and expert knowledge.

Poverty and the Underclass

When stratification scholars have turned their attention to the bottom of the class structure, a rather different set of questions has tended to emerge. The scholarship on elites has typically emphasized the power struggles between competing factions (e.g., capitalists, intellectuals, political elites), whereas the scholarship on poverty has instead focused on the various *external forces* that have affected the amount and experience of poverty. These forces take the form of (1) impersonal economic factors (e.g., globalization) that have eliminated manufacturing jobs for low-skill workers (e.g., Wilson, 2011); (2) the relatively anemic redistributive programs within the US as a whole (e.g., Kenworthy, 2011) and the southern US more specifically (e.g., Newman and O'Brien, 2011); (3) the extremely high rates of incarceration among blacks (e.g., Western and Pettit, 2010); (4) the role of concentrated poverty in reducing access to jobs, good schools, and other neighborhood assets (e.g., Massey and Denton, 1993); and (5) the lasting effects of early childhood stress on cognitive development (e.g., Evans et al., 2011). The foregoing literature on the institutional sources of poverty is joined, of course, by much research on individual-level 'risk factors' that predispose one to poverty. It is well known, for example, that poverty rates are especially high among children, young adults, the less educated, blacks and Hispanics, and female heads of households (Danziger et al., 2012).

The institutional factors listed above combine to produce an unusually high poverty rate in the US. In 1973, the official poverty measure dropped to a historic low of 11.1%, but that level was never again achieved over the course of the following four decades. The recent recession, like most prior ones, has caused a surge in poverty. The official measure of poverty is currently near its highest level in the last half-century, and most scholars expect that poverty will remain high for many more years (Danziger et al., 2012).

Social Mobility

The language of stratification theory makes a sharp distinction between the distribution of social rewards (e.g., the income distribution) and the distribution of opportunities for securing these rewards. As sociologists have frequently noted, it is the latter distribution that governs popular judgments about the legitimacy of stratification: The typical American, for example, is quite willing to tolerate substantial inequalities in power, wealth, or prestige provided that the opportunities for securing these social goods are distributed equally. If the competition has been fairly run, most Americans are untroubled by a stratification system that amply rewards the winners. This line of reasoning has motivated a long tradition of research on social mobility examining whether the US is living up to this commitment to a fairly run race in which everyone, no matter how humble his or her origins, has ample opportunities to get ahead.

It is conventional to carry out such research with 'mobility tables' that represent the intergenerational flows between origin and destination class categories (e.g., Jonsson et al., 2011). These tables tell us, for example, how frequently

children born into the craft category move up to the professional or managerial sector, remain within the craft category, or move down to less skilled manual work. It is of course not possible on the basis of such raw observational data to make any straightforward assessment of the opportunities that have been made available to children born into each social class. This is partly because standard outcome-based measures of inequality are tainted by the confounding effect of differential tastes. For example, a worker with a well-developed taste for leisure may opt to work for relatively few hours, leading to low earnings but nonetheless optimal utility (by virtue of the high valuation placed on leisure).

Whenever differential outcomes are generated through the operation of differential tastes, some scholars would argue that the resulting differences in outcomes are unproblematic, as they simply express trade-offs that are being made between different types of valued goods (e.g., leisure and income). This argument implies that we should measure the distribution of opportunities that prevails before differential tastes can express themselves. The key objective under this formulation is to determine whether 'capabilities' (i.e., opportunities to secure rewards) are equally distributed, not whether rewards themselves, which reflect the operation of tastes, are equally distributed. It bears noting that, insofar as tastes themselves are endogenous (because, for example, children may cathect with their parents), it becomes yet more difficult to tease out the way in which social origins constrain destinations on the basis of conventional observational data.

This is all to suggest that one cannot read the mobility table as expressing straightforwardly the extent to which opportunities are unequally distributed. Even so, most sociologists and social scientists tend to hold aside all such complexities, and analyses of mobility tables are thus understood as exposing, if only imperfectly, the extent to which the liberal ideal of an open system has been realized. This objective underlies analyses of (1) the gross association between occupational origins and destinations, (2) the extent to which children garner a return from advantaged economic origins, (3) the extent to which educational investments are a key mediating variable between origins and destinations, (4) the role of social psychological variables, such as aspirations and expectations, in affecting the process of attainment, (5) the effects of networks on occupational and other outcomes, and (6) the effects of labor market institutions (e.g., unions) on remuneration and bargaining power.

Racial and Ethnic Stratification

When scholars study social mobility, they are of course exploring a particular form of ascription, a form that arises insofar as children born into well-off families are advantaged in the labor market, the marriage market, and in the competition for other valued resources. But of course class-based ascription is but one of many forms of ascription. Especially in the US, racial and ethnic inequalities are also very potent examples of ascription, and a large body of scholarship has accordingly developed to understand its causes and effects.

The obvious starting point in studying racial and ethnic inequality is to understand how the classification schemes themselves (e.g., 'black,' 'white,' 'Asian') come to be constructed and reconstructed over time. It is important, for example, to understand the processes by which such schemes come into being and are reproduced (see Omi and Winant, 1986) as well as the processes by which individuals are sorted (and sort themselves) into such schemes as are already in place (Saperstein and Penner, 2014). There is an increasingly large literature showing that racial and ethnic identities, far from being ascriptively fixed at birth, are in fact very malleable over the course of one's life.

As important as these processes are, there is yet more research on the dynamics by which entire immigrant groups, not just particular individuals, are sorted into the racial and ethnic system within the host country. This line of research examines, for example, whether new entrants are assimilating in 'straight-line' fashion (e.g., Alba and Nee, 2014), whether especially large and ongoing immigrant streams work to interrupt such straight-line assimilation (e.g., Jiménez, 2014), and whether ethnic enclaves work to hinder or impede the incorporation process (e.g., Portes and Manning, 1986).

There is also much research on the forces making for change and stability in racial and ethnic inequality. Because our various institutions have all been profoundly racialized, there are accordingly many institutional arenas in which change might occur, a point that Wilson (2011) nicely makes. At various points in US history, we have witnessed profound transformations of racial and ethnic inequality, transformations that appear to be linked to changes in residential segregation (e.g., Massey and Denton, 1993); changes in interracial assortative mating (Qian and Lichter, 2014); changes in laws, norms, and institutional forms (e.g., Sampson, 2012); and of course economic changes and transformations (e.g., Wilson, 2011).

Gender Inequality

It was once fashionable to argue that all forms of ascription are fated to wither away because they (1) rest on a caste-like allocation of labor that ignores considerations of merit and talent, and (2) are inconsistent with our moral commitment to running an equal-opportunity society (Parsons, 1970). By this logic, the ascription that persists to this day is but a residual holdover, with the twin forces of efficiency and egalitarianism slowly but gradually wearing it away.

It is of course by now clear that racial, ethnic, and class-based forms of ascription are deeply resistant to change even under the pressure of these forces. Although the same may be said of gender-based ascription, the gender form is nonetheless distinctive by virtue of playing out so prominently *within* families. Because we watch 'up close and personal' the gender division of labor in our families, we are under special pressure to believe essentialist stories justifying it, stories that represent women as especially suited for nurturing or service and men especially suited for cognitive or manual labor. These stories, so some have argued (see Charles and Grusky, 2011), take on a powerful life of their own, tend to legitimate and perpetuate an essentialist division of labor, and thereby impart a special staying power to gender-based ascription.

For this and many other reasons, we might expect gender inequality to evolve under a logic that may differ from that

governing racial, ethnic, or class inequalities. There is a long stream of research on gender exploring whether indeed gender is subject to a very distinctive logic of change. This research takes on such questions as whether women are 'opting out' of the workplace (e.g., Percheski, 2014), whether there is still much discrimination against women and mothers (e.g., Correll et al., 2008), whether there are 'glass ceilings' and other forms of workplace hypersegregation (e.g., Gorman and Kmec, 2014), and whether there is a recent 'stalling out' in the historic declines in gender inequality (e.g., England, 2010).

The Consequences of Inequality

The subfields discussed to this point have addressed the form and causes of inequality. By contrast, the classical literature on inequality, most notably that of Marx (1970), focused on the macrolevel *effects* of inequality, especially its implications for systemic change (e.g., revolution). This macrolevel tradition lives on today in the form of cross-national studies of inequality on total economic output, ethnic unrest, democracy, and much more. It also continues in the form of speculative commentary on the various ways in which present-day inequality regimes contain within them the kernel of future inequality regimes.

Although the macrolevel tradition is still in play, there is more research today on the comparatively modest question of how an individual's position within an inequality regime affects the course of her or his life. The scholars working within this microlevel tradition examine the effects of an individual's class position on (1) lifestyles, tastes, and consumption practices (e.g., Bourdieu, 1984); (2) childrearing and socialization practices (e.g., Lareau, 2003); (3) political attitudes and voting behavior (e.g., Frank, 2004); and (4) opportunities for a long and healthy life (e.g., Lutfey and Freese, 2014). It should come as no surprise that the effects of inequality are so wide-ranging. In contemporary societies, most everything is marketized and commodified, meaning that one's participation across a wide variety of domains is directly affected by the capacity to pay for that participation. There is still much debate over the appropriate reach of commodification, but that debate occurs within a more circumscribed range than it once did.

Conclusion

Is there any theme running through the many questions that sociologists of inequality have taken on? The field is of course sprawling and the questions of course diverse. If pressed nonetheless to identify a theme, all that can be noted is the characteristically modern tendency to problematize inequality, to ask how it came about, and to inquire as to the mechanisms generating it.

This problematizing orientation is distinctively modern in its underpinnings. For the greater part of human history, the stratification order was regarded as an immutable feature of society, and the objective of commentators was to explain or justify this order in terms of religious or quasi-religious doctrines (e.g., Bottomore, 1965; Tawney, 1931). It was only with the Enlightenment that a critical rhetoric of equality emerged in opposition to the civil and legal advantages of the aristocracy and other privileged status groupings. In its most radical form, this egalitarianism led to Marxist interpretations of human history, and it ultimately provided the intellectual underpinnings for socialist stratification systems. While much of stratification theory has been formulated in reaction and opposition to these early forms of Marxist scholarship, the field nonetheless shares with Marxism the distinctively Enlightenment premise that individuals are to be equally valued and respected regardless of differences in their economic or social standing (see Meyer, 2001). In part, the questions laid out here reflect this tension between the brute facts of extreme and sometimes growing inequality and the modern commitment to at least *some* forms of equality, at minimum an equality of respect and of opportunity.

See also: Elites: Sociological Aspects; Income Inequality; Inequality Analysis: Overview; Inequality, Social; Inequality: Comparative Aspects; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Poverty: Measurement and Analysis; Religion and Stratification; Social Inequality and Schooling; Social Inequality in Cultural Consumption Patterns; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Urban Poverty in Neighborhoods.

Bibliography

Alba, R., Nee, V., 2014. Assimilation theory for an era of unprecedented diversity. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 721–728.

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. Journal of Economic Literature 49 (1), 3–71.

Bottomore, T.B., 1965. Classes in Modern Society. Allen and Unwin, London.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge.

Brooks, D., 2000. Bobos in Paradise: The New Upper Class and How They Got There. Simon and Schuster, New York.

Charles, M., Grusky, D.B., 2011. Egalitarianism and gender inequality. In: Grusky, D.B. (Ed.), The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender, second ed. Westview Press, Boulder, pp. 389–404.

Correll, S.J., Benard, S., Paik, I., 2008. Getting a job: is there a motherhood penalty? In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, third ed. Westview Press, Boulder, pp. 759–770.

Danziger, S., Chavez, K., Cumberworth, E., 2012. Poverty and the great recession. In: Recession Trends Report. Russell Sage Foundation, New York, pp. 1–5.

Davis, K., Moore, W.E., 1945. Some principles of stratification. American Sociological Review 10, 242–249.

DiPrete, T.A., Eirich, G.M., Pittinsky, M., 2014. Leapfrogs and the surge in executive pay. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 104–113.

England, P., 2010. The gender revolution: uneven and stalled. Gender & Society 24 (2), 149–166.

Esping-Andersen, G., Myles, J., 2008. The Welfare State and Redistribution. http://dcpis.upf.edu/~gosta-esping-andersen/materials/welfare_state.pdf.

Evans, G.W., Brooks-Gunn, J., Klebanov, P.K., Winter 2011. Stressing out the poor: chronic psychological stress and the income-achievement gap. Pathways 16–21.

Frank, T., 2004. What's the Matter with Kansas? How Conservatives Won the Heart of America. Metropolitan Books, New York.

Freeman, R., 2012. (Some) inequality is good for you. In: Grusky, D.B., Kricheli-Katz, T. (Eds.), The New Gilded Age. Stanford University Press, Stanford, pp. 63–87.

Goldin, C., Katz, L.F., 2008. Race between Education and Technology. The Belknap Press of Harvard University Press, Cambridge.
Gorman, E.H., Kmec, J.A., 2014. Glass ceilings in corporate law firms. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 890–902.
Gouldner, A.W., 1982. The Future of Intellectuals and the Rise of the New Class. Oxford University Press, Oxford.
Grusky, D.B., March/April 2012. What to Do About Inequality? Boston Review. http://www.bostonreview.net/BR37.2/ndf_inequality.php.
Grusky, D.B., Weisshaar, K.R., 2014. A compressed history of inequality. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 44–52.
Hacker, J.S., Pierson, P., 2010. Winner-take-all politics: public policy, political organization, and the precipitous rise of top incomes in the United States. Politics & Society 38 (2), 152–204.
Jiménez, T.R., 2014. Why replenishment strengthens racial and ethnic boundaries. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 740–746.
Jonsson, J.O., Grusky, D.B., Di Carlo, M., Pollak, R., 2011. It's a decent bet that our children will be professors too. In: Grusky, D.B. (Ed.), The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender, second ed. Westview Press, Boulder, pp. 499–516.
Kenworthy, L., Fall 2011. How rich countries lift up the poor. Pathways 28–32.
Lareau, A., 2003. Unequal Childhoods: Class, Race, and Family Life. University of California Press, Oakland.
Lutfey, K., Freese, J., 2014. The fundamentals of fundamental causality. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 1050–1056.
Marx, K., 1970. Ideology and class. In: Arthur, C.J. (Ed.), The German Ideology. International Publishers, New York.
Massey, D.S., Denton, N.A., 1993. The missing link. In: American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge.
Meyer, J.W., 2001. The evolution of modern stratification systems. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, second ed. Westview Press, Boulder, pp. 881–890.
Mills, C.W., 1956. The Power Elite. Oxford University Press, Oxford.
Newman, K.S., O'Brien, R.L., Summer 2011. Taxing the poor: how some states make poverty worse. Pathways 22–26.
Okun, A.M., 1975. Equality and Efficiency: The Big Tradeoff. Brookings Institution Press, Washington, DC.
Omi, M., Winant, H., 1986. Racial Formations in the United States. Routledge, London.
Parsons, T., 1970. Equality or inequality in modern society, or social stratification revisited. Sociological Inquiry 40, 13–72.
Percheski, C., 2014. Opting out? In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 811–819.
Portes, A., Manning, R.D., 1986. The immigrant enclave: theory and empirical examples. In: Olzak, S., Nagel, J. (Eds.), Competitive Ethnic Relations. Elsevier, Amsterdam, pp. 47–68.
Qian, Z., Lichter, D.T., 2014. Are recent trends in intermarriage consistent with assimilation theory? In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 788–799.
Sampson, R., 2012. Legacies of inequality. In: Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago.
Saperstein, A., Penner, A.M., 2014. The dynamics of racial fluidity and inequality. In: Grusky, D.B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective, fourth ed. Westview Press, Boulder, pp. 687–695.
Schumpeter, J., 1953. Aufsätze zur Soziologie. Mohr/Siebeck, Tübingen.
Sen, A., 1997. From income inequality to economic inequality. Southern Economic Journal 64, 383–401.
Tawney, R.H., 1931. Equality. George Allen and Unwin, London.
Veblen, T., 2008. The theory of the leisure class. In: Grusky, David B. (Ed.), Social Stratification, Class, Race, and Gender in Sociological Perspective, third ed. Westview Press, Boulder, pp. 862–869.
Western, B., Pettit, B., 2010. Incarceration and social inequality. Daedalus 139 (3), 8–19.
Wilson, W.J., Spring 2011. Being poor, black, and American. American Educator 10–23.
Wilson, W.J., 1978. The Declining Significance of Race: Blacks and Changing American Institutions. University of Chicago Press, Chicago.
Wright, C.W., 1956. The Power Elite. Oxford University Press, Oxford.

Social Structure

John L Martin and Monica Lee, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Chazel, volume 22, pp. 15233–15237, © 2001, Elsevier Ltd.

Abstract

Social structure is about patterns of relations. Two main families of approaches conceive these patterns differently. One, a top down one, tends to see a global organization as stemming from a common shared culture. The second, a bottom up, pays more attention to agglomerative processes that begin with individuals who may or may not begin with shared cultural templates.

Social structure refers to patternings in social relations that have some sort of obduracy. Within this general definition, there are two primary families of more specific approaches. In the first, 'structure' may be used to refer on the macro level to the abstract organization of reciprocally defined social categories that are seen to comprise some social whole. In the second, the term can be used to refer to smaller scale 'social structures,' configurations of concrete relationships among individuals without reference to a notion of a larger societal totality. We organize our exposition accordingly. (We note that Porpora (1989) in addition gave as conceptions of social structure first, that of Anthony Giddens, which we treat here as an extreme form of the first understanding of structure, and second, relations between variables, but we have not seen any examples of people claiming this as a definition of structure, and we discuss this under the related heading of *social systems* below.)

Structure as Abstract Relations between Social Positions

Although all the approaches in this category link structure to some sort of organization of positions or types that anchor action, they differ as to the logic of the organization of the positions that may variously be taken to be social functions, roles, or classes.

Structure and Function

The idea of 'social structure' was first introduced by Herbert Spencer (e.g., 1896[1873]: pp. 56–60). At the time, the word 'structure' in biology referred to what we would now call 'organs,' sets of contiguous tissue that performed a specifiable function for the organism as a whole. Spencer argued that society had 'social structures' that carried out social functions.

Thus the root of the idea of social structure comes from the organismic metaphor applied to society. This metaphor is certainly an old one; in the Western tradition we often begin with Plato, who (we now say) suggested that the city might be understood as a 'man writ large,' and thus a convenient place for an anthropology. (In the *Republic* (II:368d) Plato argues that, given that those of us with imperfect vision have an easier time reading larger letters, we should find a place to study the nature of justice similarly writ large, and that is the city. From here, he uses our interdependence to derive the need for specific occupations, for specialization, and for trade (369–371).) Now indeed, Plato did suggest a mapping between characteristics of persons and those of the city. Most important, Plato made a distinction between (what might appear to us as) cognition, emotions, and instincts. Thus in the fourth book of the *Republic* (espec. §436), Plato had Socrates demonstrate that there are three parts of the soul that have different functions – that we "learn with one part of ourselves, feel anger with another, and with yet a third desire the pleasures of nutrition and generation." (The physical organs to which these were taken to correspond were for many centuries taken to be the head, the heart, and the liver, organized in a vertical hierarchy.) So too, he argued, the city has three classes, each of whom must do its part. Some make things (or 'make money,' corresponding to the appetitive), some make rules (corresponding to the *nous*, the intellect), and some make war. These correspond to three qualities of the good city, which should be wise (the wisdom of the city is the wisdom of its counselors), brave (the bravery of the city is the bravery of its fighters), and sober (*Republic*, §428–434). Now despite his famous emphasis on our interdependence and even on the division of labor, Plato did not propose an organismic model of the state, and it seems that no such developed analogies arose in Europe until the mass of differentiable urban occupations were no longer associated with servile status.

With the rise of materialist views of human beings and of society, organismic metaphors were extended into more developed allegories: thus Hobbes (1943[1651]: pp. 8, 171, 183–188, 193f., 246–257) proposed correspondences in the body politic to nerves, blood and joints, and could liken its states of illness to pleurisy, Siamese triplets, and constipation. But despite his categorical rejection of metaphor, Hobbes's use of the organismic language stemmed as much from his love of pursuing a simile as from his explanatory goals. (Thus not only is the distribution of goods analogous to nutrition, but these good ultimately come from either the land or the sea, '(the two breasts of our common Mother),' adds Hobbes (1943[1651]: p. 189).) Certainly, Hobbes treated the organismic predicate as one of metaphor and not as one of identity (e.g., systems "*may be* compared … to the similar parts of man's body" (emphasis added)), and alternated between it and others (the commonwealth as building, as Leviathan, or as 'Mortall God').

Although such organismic metaphors remained popular, we do not know of any attempt to produce a serious and detailed exposition until Spencer. Comte (1974[1842]: 510ff) too had advanced an organismic metaphor by applying it to the societal division of labor, examining the laws according to which the societal whole develops and how the whole regulates its constituent organs. But Comte's overriding theme of wholism (and his impatience to begin wholesale social reconstruction) led him to neglect a detailed consideration of the parts.

By contrast, Spencer argued that the functional needs of the whole not only were met by the parts (like Plato), nor that there were laws of development of the whole (like Comte), but that structures (and functions) have their own developmental tendencies that direct the social organism's course of development. ("Just as in the man there are structures and functions which make possible the doings his biographer tells of, so in the nation there are structures and functions which make possible the doings its historian tells of; and in both cases it is with these structures and functions, in their origin, development, and decline, that science is concerned" (Spencer, 1896[1873]: p. 52).) Further, despite his use of the organismic metaphor, Spencer (e.g., 1910: p. 244) always saw the functions that were met by social structures as being, at base, functional needs of the *individuals* who composed society (and not those of the social whole apart from the individuals). Human action is fundamentally driven to meet human needs and wants such as defending against enemies and obtaining food. Individuals cooperate to fulfill those ends, giving rise to structures. Thus 'social structure' referred to the internal organization of the social whole arising as a by-product of the functional interdependence of human life (a point that Malinowski was later to reiterate).

The organismic analogy developed in ways charted by Kaufmann (1908) and Levine (1995). What was crucial for sociology and anthropology was the duality of structure and function that was accepted by Durkheim (most notably in *The Division of Labor*) and elaborated by Radcliffe-Brown.

Like Spencer and Comte, Durkheim proposed that society is structured according to the functions that social groups fulfill in the division of labor. But while Spencer retained an emphasis on the ontological primacy of the individual, Durkheim unreservedly based his theory in the *sui generis* existence of social order. He examined how the functional interdependence upon which the division of labor is based served as society's source of social solidarity. The most fundamental 'function' in society is the division of labor's function of integrating and uniting the societal organism, thereby sustaining its life (1933[1893]: p. 11). While mechanical solidarity – solidarity based on individuals' similarities – was the dominant foundation for social bonds in earlier epochs, modern society, according to Durkheim, is bound primarily by 'organic' solidarity – solidarity that arises from individuals' interdependence in the societal division of labor. Differentiated social groups comprise a system of 'organs' (p. 132), each of which performs a specialized duty that sustains the whole. These organs are bound to each other by relationships of functional interdependence (pp. 83, 173), where the survival of the whole depends on each organ performing its specific duty. The more specialized the function of an organism, the more integrated and developed the society.

This perspective was influential in anthropology, largely mediated by the work of Radcliffe-Brown (1952: p. 179), who emphasized that a structure 'is thus to be defined as a set of relations between entities.' Further, while other theorists had focused on a few roles as key – kinship or occupation, most notably – Radcliffe-Brown (1952: 191f) argued that such structure had to be understood as involving *all* such relations. This suggested a more complete vision of an all-encompassing and mutually consistent social structure. Strangely, sociology tended to be influenced by earlier work in anthropology that drew more upon notions of kinship than social functions, perhaps because it had clearer implications.

Structure as Role

In particular, anthropologist Ralph Linton (1936) had suggested a second duality (after that of structure/function), namely that of status and role. A status, argued Linton (1936: p. 113) is a slot in social structure, or as he said (he used the idea of social 'systems,' not 'structures'), 'a position in a particular pattern.' Such a position has both rights and duties associated with it, and action in accordance with these is the performance of a *role*. Thus the role is the active counterpart to the status, the verb form as opposed to the noun. (One will note that in English, perhaps the bulk of early terms for common objects also have proper usage as verbs: to hammer is to use a hammer, to water is to put water on, and so on. So too a mother mothers, and this is a status 'roling.')

A number of sociologists have similarly examined social structure as an organized set of roles. This 'structure' was increasingly understood by sociologists as one-half, or one-third, of the realm of study at hand. Pitrim Sorokin (e.g., 1962) proposed a triad of *personality*, *society*, and *culture* (though he is more known for employing the latter two as a pair), and this triad was also adopted by both his younger colleague (and nemesis), Talcott Parsons and his graduate student, Robert Merton. Both saw the 'social' dimension as referring to something structural, and both paradigmatically took this to mean the organization of social roles, thus preserving the Lintonian duality. Parsons in particular emphasized the close mapping between structure and culture. In Parsons's (1961) words, "the structure of social systems … consists in institutionalized patterns of normative culture." Societal roles, in other words, are defined by specific norms that prescribe rules for behavior.

Robert Merton (1957) also explicitly began with this understanding of social structure, and indeed said "status and roles become concepts serving to connect culturally defined expectations with the patterned conduct and relationships which make up a social structure." (Parenthetically, we also note that Merton (e.g., 1976) maintained an interest in the emerging structuralist theories and believed them related to his own structural approach. We return to this branch below.)

In sum, this approach to social structure tended to emphasize culture, and see structure as, at least in the long-run equilibrium, a nonproblematic objectivation of cultural patterns. This priority tended to be reversed in most approaches that drew from Marx.

Structure and Class

Now we have said that Spencer was the first to use the term 'social structure,' and this is indeed true. However, Marx had

used similar language, and, with Engels in *The German Ideology* (Marx and Engels, 1976[1845–1846]: 35f), had written "The social structure and the state are continually evolving out of the life-process of definite individuals, however, of these individuals, not as they may appear in their own or other people's imagination, but as they *actually* are; i.e., as they act, produce materially, and hence as they work under definite material limits, presuppositions, and conditions independent of their will." The term translated as social structure, namely *gesellschaftliche Gliederung* (*Werke*, 3:25) carries more connotations of organization or arrangement, and somewhat fewer of a construction. Marx's conception in its most general is to emphasize what he called (in the 'Theses on Feuerbach') the 'ensemble of societal relations' (*Werke* 3:7).

The social relations of paramount interest to Marx were economic ones, more specifically, the social relations of production. These relations appear in juridical form as property relations, but in social form as the relations between classes. It is this set of relations that Marx argued constituted the 'economic structure' of society; the 'superstructure' erected on top of this includes legal, political, and ideological formations that have some degree of homology to the underlying economic structure (*see* Marxism). Thus as in the previous conception of social structure, we have a close duality between cultural understandings and patterns of relations. But while the first conception privileged the cultural realm as having explanatory priority, here it is the pattern of relations that must be understood first.

Structure as Culture as Structure

These conceptions tend to establish a duality between what one's position in structure is, and what one does; we have seen that in the Lintonian and Parsonian versions, this implies a duality between social structure and aspects of culture. This was taken to its logical vanishing point in the work of Anthony Giddens. Giddens decided to claim as 'structure' in social systems 'rules and resources,' arguing (e.g., 1979: pp. 64–66) that structure exists only as the structural – indeed, the structuring – properties of a social system. (Here he seems to have been influenced by the conception of structuralism discussed below.) Although intellectually pleasing, by placing a heap of social structure in the heads of the actors (in the form of rules), Giddens, wittingly or unwittingly, sent social theory careening down a hill in which structure and culture would end up being not simply dual, but indistinguishable. This was associated with further confusions.

For whether it is the expectation that the duties of the younger sister include looking after aging parents, or it is the class relation that the proletariat does not own the means of production but instead sells labor power, social structure was easily assimilated to constraint and restriction of individuals. As a result, especially in Britain, there arose a conviction that social action could be analyzed as a two-part beast, one part the structural constraint and the other, agency. Unlike the seemingly homologous division of the self by Mead into the 'me' (including, 'what they expect me to do') and the 'I' ('what I *actually* did'), here most conceptions made a differentiation between those actions that *reproduced* structure (and hence had no valid claim to agency) and those that challenged, transformed, or at least grumbled about the structure. But with the transition from (1) structure being defined as organized relations to (2) structure being defined by its pernicious effects, we leave the realm of sociology, and hence do not pursue this approach further. Instead, we briefly indicate how these top-down conceptions of social structure influenced and then were influenced by the development of two related concepts.

Continuities

Systems Theory

This conception of social structure influenced the development of two relatively coherent schools of thought, treated separately in this volume. The first is systems theory (*see* Systems Theory), which also started with the organismic metaphor (indeed, when Hobbes spoke of 'social systems,' he was using the then-current term 'system' for 'organ'). Modern systems theory, the rise of which is often assumed to begin with Ludwig von Bertalanffy's (1968) work in the 1940s, explores the regularities that emerge when sets of elements – whether physiological, environmental, or social – assume such a form that they become self-regulating. In other words, they maintain themselves within a range of possibilities (a state space) through self-correcting feedback mechanisms. In the context of social science, systems theory explores how societies remain stable and functioning as their constituent parts cooperate and adapt to one another. What is important for us is the effects that this had on notions of social structure.

Talcott Parsons first popularized a systems approach to the study of social structure in *The Social System* (1951). He went on to propose four systemic necessities – basic functions that systems must perform in order to maintain their existence – as key to understanding the organization and development of societies. Parsons tended to propose subsystems that were defined *wholly functionally* (and not structurally), and to examine the media that various subsystems used to order their relations or to interact with other subsystems. To the extent that such systems thinking was taken seriously, it tended to lessen interest in social structure as such (an emphasis preserved by Merton, who maintained an ambivalent relation to the idea of systems).

A further divergence of systems and structural thinking came with the work of Luhmann. Once mentored by Parsons, Niklas Luhmann developed a rival approach to systems theory whose core is communication rather than action. Humans are not coordinated into action systems but systems of observing and sense making. Actors confront an external environment in which massive amounts of information circulate. As they observe, select out, and make sense of, bits of information, the system is constituted (1996: p. 24). By observing the environment, the observer distinguishes between himself and the environment, and this act of distinction brings the observer into being. What is essential for us are three points: first, the components here are acts of communication, not social relationships. Second, there is a self-organization (*autopoiesis*) of this set of communications that may be largely decoupled from the needs of organic individuals. Third, such systems are more than a mere analytical tool that elucidates societal processes; rather, 'systems exist' (1995: p. 12). Thus although Luhmann did argue that the major social subsystems corresponded to well-understood realms (which he

successively explored: law, art, politics, religion, economy and education), we must be prepared for functional sub-sub-systems that are potentially divorced from needs of individuals and even from the role structure.

Structuralism

The second offshoot is the line of thought termed 'structuralism' (*see* Structuralism). This tradition in many ways branched off of Comte and Durkheim; although it is generally understood to have originated with the structural linguistics of Ferdinand de Saussure, Saussure himself was influenced by Durkheim's Comtean division between statics and dynamics. Saussure argued that words' meanings arise not from an intrinsic quality that links them to the phenomena they represent, but from the systematic distinctions that differentiate them (as signs or sounds) from one another in a linguistic system (Saussure 1983[1916]). (Words are "purely differential and defined not by their positive content but negatively by their relations with the other terms of the system. Their most precise characteristic is in being what the others are not" (Saussure [1916]1983: p. 651).) This approach was then elaborated in the science of phonology by Roman Jakobson, who examined how the organization of sounds into systems conveys linguistic meaning. In all cases, structuralism proposed two types of relations (syntax/system, metonymy/metaphor, affine/cognate) that, brought into the realm of social structure, suggested attention both to empirical relations of interaction (on the one hand) and relations of structural equivalence – functional substitutability (on the other).

Claude Levi-Strauss (1969[1949]) applied these structuralist principles to the study of social structure in *The Elementary Structures of Kinship*. Here Levi-Strauss examined the structural logics that organize complex patterns of kin relationships by understanding them as formal laws of transfers (as a marriage may be seen as the transfer of a daughter (or son) from one lineage to another). Levi-Strauss went on to focus on *cognitive* structures, but the encouraging results of this work on kinship led to a common enthusiasm for 'structure' in the social sense as well. There were relatively few rigorously structuralist approaches to social structure, however, though we discuss one below.

Structure as Patterns of Concrete Interaction

A second major approach divorces the idea of structure from the abstract notion of an organized society. Instead, it sees structure as concrete patterns of interaction among individual persons that may exhibit greater or lesser levels of organization. These social relationships are also attached to cultural meanings of various sorts – for instance, subjective orientations or cultural narratives. But these approaches tend to be more bottom up, and suggest a plurality of structural elements that may overlap with or without mutual reinforcement.

Structure as Forms of Interaction

This general approach was largely rooted in the work of Georg Simmel. Like Durkheim and Marx, Simmel saw society (or social life) as fundamentally about relations among persons. But he eschewed the structural functionalists' presupposition that society is divided into subparts organized into a (more or less) functional whole or system. Indeed, in his first major work, Simmel (1890: p. 13) challenges the organismic view of society, which sees society as 'an absolute unity.' Rather, this concept of 'society' is 'only secondary to the real interaction of its parts, only a result of it – as much in reality as in conceptualization.' 'The social organism,' speaking loosely, must be understood as but 'a summary expression of the interaction of the parts,' and it is only these parts – individuals – that have vitality. Rather than the process of differentiation leading to simpler and more unitarily defined individuals (as in Durkheim's conception), the true differentiation is internal to individuals who straddle multiple social roles.

As a result, Simmel decoupled the concept of structure from social groups' functional roles in society, and proposed a 'formal sociology' focused on patterns of interaction among individuals. These interactions – continuously broken off and remade – do not necessarily amount to 'actual organization' (1992[1908]: p. 33), nor must they conform to the routines of society's 'service organs' (Simmel, 1897). Thus, 'Society' for Simmel is social interaction among persons "continuously making connections and breaking them off and making them again, a perpetual flowing and pulsing that unites individuals, even when it does not amount to actual organization" (1992 [1908]: 33). Social life consists of the living interactions, constantly being made and remade, that pulse beneath and beyond their 'service organs' (Simmel, 1897). The regularity of these interactions comes from their being stamped from a limited set of forms.

This sociology of forms has been appealed to by some network researchers as an aegis for their own efforts. However, despite his focus on formal patterns like the dyad and the triad (actually terms supplied by his translator), Simmel's use of 'form' had a somewhat different meaning than a geometrical structure. Hoping to extend a neo-Kantian approach to social life, as he had previously attempted for the study of history, Simmel meant to uncover the 'a prioris' for social interaction, which he considered forms. Here he certainly meant to draw on Kant's idea of forms, quite possibly interpreted in the Platonic way of Hermann Cohen, as something structured in the subjectivity of actors – not a pattern of relationships itself. While purely geometric abstractions may be useful sociological tools, they cannot be separated from the complex subjective orientations from which they tend to emerge. These objective and subjective elements coalesce in ideal typical forms of 'sociation' (*Vergesellschaftung*): e.g., exchange, conflict, and super- and subordination.

Most important for the science of social structure was that Simmel, by not proposing a partition (or multiple partitions) of society into subparts bound by relations, was able to see structure as shifting patterns of interaction between individual persons and thereby frame structural analysis as a conceptual tool that could be applied to different types of connections. As he (1955[1922]) explained in *The Intersection of Social Circles*, an individual may stand simultaneously in multiple social groups that overlap with each other at the site of that individual. Society is a web of relationships that is not organized into a unified and harmonious set of structures but into multiple overlapping, often uncoordinated, and sometimes conflicting, ones.

Structure as Configurations of Individuals

This line of research was continued by Leopold von Wiese (1932), in his 'theory of relations' and his 'theory of social configurations (*soziale Gebilde*).' Von Wiese's work was freely translated by Howard Becker, who rendered *Gebilde* as 'plurality pattern.' Von Wiese's configurations or structures were ensembles of social relationships that had a particular form: he argued that there were three main types of configurations: masses, groups, and corporate bodies. These were structures, generally composed of other structures, that were somewhat farther from the individual than their component structures and consequently more abstracted. (Sorokin's (1962) own approach was similar in many ways, emphasizing the components of social organization as opposed to beginning with functional requirements.) The somewhat exciting aspect of von Wiese's work was the implication that one could begin with social relationships and scale upward, first to simple configurations composed of relationships, and then larger structures. This general approach has been used by Tilly (1998), Martin (2009), and Martin and Lee (2010).

Structure as Stories

This type of approach may seem to have a great deal in common with the structural approach of Harrison White (1992), who was concerned with the question of how social formations emerge. White's first book was a recasting of Levi-Straussian kinship structures using permutation matrices. He then applied this rigorous structuralist understanding of social structure to other types of social structure in a pair of important papers, the first of which (White et al., 1976) was very influential for social networks (see below); although the second (Boorman and White, 1976) was more theoretically central, it was more methodologically challenging and hence less influential. (This line of work was continued by Breiger and Pattison (e.g., 1986), among others.)

White went on to formulate a more general approach to social organization, and here he drew more heavily on S.F. Nadel's (1957) perspective with its cultural grounding, which became clear in White's magnum opus, where he posited a strict duality between structural form and cultural narratives. Structural ties are bound by common identities. Each tie exists both as a pair of individuals, as well as an identity – a story about the basis of the tie's solidarity that provides it continuity. Small structures like dyads concatenate to create larger ones as 'ties mesh with stories' (2008: p. 27). The conversations that take place in ties, the shared beliefs or understandings that bind them, link up to those of other ties, presenting the opportunity for those ties to concatenate into larger structures.

A somewhat related approach is seen in Ronald Burt's (1982) *Structural Theory of Action*, though the focus on action takes him more to the Parsonian legacy of a dialogue with economic theories of action. As with White, we start with a focus on the actor who sees social structure sometimes as if it were a jungle gym to be climbed, and sometimes as if it were an undergrowth to be hacked away with a machete. Even more than in the previous understandings, the duality of cultural and structural understandings is one that in no way requires an overarching order; rather, local ordering principles may emerge, but even when they do, skilled actors may choose to crack them.

Social Structure and Social Networks

As the discussion in the previous section makes clear, this approach to social structure often dovetailed, inspired, or turned into social networks research (*see* Social Networks). Scholarship on social networks focuses on the relationships between concrete actors rather than individuals, groups, or roles. Although the foundational theoretical works in social networks research (e.g., Barnes, 1954) envisioned the network as sprawling, connecting, and all-encompassing form, and hence diametrically opposed to the conventional ideas of well-nested social sub-sub-systems and/or groups, networks have often been understood as being a skeletal structure that should reproduce the outlines of social structure conventionally understood. This is, of course, an empirical question, but the challenges to mounting a full-scale social network survey of a population in order to determine the relation of networks to role structures remain daunting.

Conclusion

Social structure has proved one of the most confusing theoretical terms in our lexicon; it tends to be among the most abstract of our concepts, but has implications of solidity. It often seems to connote constraint of individual actions, but is itself nothing but a set of actions. It is regularly counterposed to culture, but just as regularly appears to be simply the second side of the same coin as culture. Perhaps the best that can be said is that it appears to be an idea we cannot do without.

See also: Action, Theories of Social; Class: Social; Exchange: Social; Institutions; Interactionism, Symbolic; Luhmann's Social Systems: Meaning, Autopoiesis, and Interpenetration; Macrosociology-Microsociology; Marxism in Contemporary Sociology; Power in Society; Social Networks; Sociological Theory; Status and Role: Structural Aspects; Structuralism.

Bibliography

Barnes, John, 1954. Class and committees in a Norwegian Island Parish. Human Relations 7, 39–58.

Boorman, Scott A., White, Harrison C., 1976. Social structure from multiple networks, II. Role structures. American Journal of Sociology 81, 1385–1446.

Breiger, Ronald L., Pattison, Phillipa E., 1986. Cumulated social roles: the duality of persons and their algebras. Social Networks 8, 215–256.

Burt, Ronald S., 1982. Toward a Structural Theory of Action. Academic Press, New York.

Comte, Auguste, 1974[1842]. The Positive Philosophy. Freely Condensed and Edited by Harriet Martineu. AMS Press, New York.

Durkheim, Emile, 1933[1893]. The Division of Labor in Society (G. Simpson, Trans.). Free Press, Glencoe, Illinois.

Durkheim, Emile, 1938[1895]. The Rules of Sociological Method (S. Solovay, J.H. Mueller, Trans.). The Free Press, Glencoe, Illinois.

Giddens, Anthony, 1979. Central Problems in Social Theory. University of California Press, Berkeley.

Hobbes, Thomas, 1943[1651]. Leviathan. The Clarendon Press, Oxford.

Kaufmann, Erich, 1908. Über den Begriff des Organismus in der Staatslehre des 19. Jahrhunderts. Carl Winter, Heidelberg.

Levine, Donald N., 1995. The organism metaphor in sociology. Social Research 62, 239–265.

Levi-Strauss, Claude, 1969[1949]. The Elementary Structures of Kinship (J.H. Bell, J.R. von Sturmer, R. Needham, Trans.). Beacon Press, Boston.
Linton, Ralph, 1936. The Study of Man. Appleton-Century-Crofts, New York.
Luhmann, Niklas, 1995[1984]. Social Systems. Stanford University Press, Stanford.
Luhmann, Niklas, 1996. The Reality of Mass Media. Polity Press, Cambridge.
Martin, John Levi, 2009. Social Structures. Princeton University Press, Princeton.
Martin, John Levi, Lee, Monica, 2010. Wie entstehen große sozialen Strukturen? In: Fuhse, Jan, Mützel, Sophie (Eds.), Relationale Soziologie: Zur kulturellen Wende der Netzwerkforschung. VS Verlag, Wiesbaden, pp. 117–136.
Marx, Karl, Engels, Frederick, 1976[1845–1846]. The German ideology. In: Collected Works, vol. 5. International Publishers, New York.
Merton, Robert K., 1957. The role-set: problems in sociological theory. The British Journal of Sociology 8, 106–120.
Merton, Robert K., 1976. Sociological Ambivalence and Other Essays. The Free Press, New York.
Nadel, S.F., 1957. The Theory of Social Structure. The Free Press, Glencoe, Illinois.
Parsons, Talcott, 1951. The Social System. The Free Press, New York.
Parsons, Talcott, 1961. An outline of the social system. In: Parsons, Talcott, Shils, Edward A., Naegle, Kaspar D., Pitts, Jesse R. (Eds.), Theories of Society, 2 vols. Simon & Schuster, The Free Press, New York, pp. 30–79.
Porpora, Douglas V., 1989. Four concepts of social structure. Journal for the Theory of Social Behaviour 19, 195–211.
Radcliffe-Brown, A.R., 1952. Structure and Function in Primitive Society. Cohen and West, London.
Saussure, Ferdinand de, 1983[1916]. Course in General Lingusitics. Open Court, Peru, Illinois.
Simmel, Georg, 1890. Über Soziale Differenzierung. Duncker & Humblot, Leipzig.
Simmel, Georg, 1897. The persistence of social groups: I. American Journal of Sociology 3, 662–698.
Simmel, Georg, 1950[1908]. Soziologie: Untersuchung uber die Formen der Vergesellschaftung. In: Wolff, Kurt H. (Ed.), The Sociology of Georg Simmel (K.H. Wolff, Trans.). The Free Press, Glencoe, Illinois.
Simmel, Georg, 1955[1922]. The web of group-affiliations. Translation of Die Kreuzung Sozialer Kreise, pp. 305–344 of Soziologie, by Reinhard Bendix, appearing as 127–195 in Conflict and the Web of Group Affiliations. Free Press, Glencoe, Illinois.
Simmel, Georg, 1992[1908]. Sociologie: Untersuchungen über die Formen der Vergesellschaftung. In: Rammstedt, O. (Ed.), Georg Simmel Gesamtausgabe, vol. 11. Suhrkamp, Frankfurt.
Sorokin, Pitrim A., 1962. Society, Culture and Personality. Cooper Square Publishers, New York.
Spencer, Herbert, 1896[1873]. The Study of Sociology. Appleton, New York.
Spencer, Herbert, 1910. The Principles of Sociology, vol. II. D. Appleton & Company, New York.
Spencer, Herbert, 1915. First Principles. Williams and Norgate, London.
Tilly, Charles, 1998. Durable Inequality. University of California Press, Berkeley.
von Bertalanffy, Ludwig, 1968. General Systems Theory. George Braziller, New York.
von Wiese, Leopold, 1932. Systematic Sociology on the Basis of the Beziehungslehre and Gebildelehre. Translated and Amplified by Howard Becker. John Wiley & Sons, New York.
White, Harrison C., 1963. An Anatomy of Kinship. Prentice Hall, Englewood Cliffs, New Jersey.
White, Harrison C., 1992. Identity and Control. Princeton University Press, Princeton.
White, Harrison C., Boorman, S.A., Breiger, Ronald L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–779.

Social Structure: History of the Concept

Charles Crothers, School of Social Sciences and Public Policy, AUT University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Although the palpably powerful nature of social structures and their effects seems obvious, it has taken social scientists – and particularly sociologists – a long time to develop adequate conceptions that both are philosophically defensible and also lead to empirical research. During the 'golden age' (in the late nineteenth century and for much of the twentieth century) of social structural analysis, rather underdeveloped ideas of social structure were, nonetheless, pressed into highly effective service in analyzing actual structures and this continues. Alongside these actual analyses, more sophisticated understandings have evolved although this area of study is still not adequately consolidated.

Introduction: Basics

Social structure is a concept and term used to capture the collective properties exhibited by social entities and to identify the characteristics of and specify the relationships among their component elements. It is widely considered to be one of the most important concepts/terms in the social sciences and a particularly central concept/term in sociology and important in other social sciences. However, writing on social structure has been scattered, piecemeal, and not particularly cumulative. Strangely, social theorists have been reticent in making their analyses of social structure more explicit (Callinicos, 2007; Crothers, 1996; López and Scott, 2000; Chew and Knottnerus, 2002).

There are at least four primordial images of social structure, which have to be somewhat separately attended to, although for adequate understanding each must be evoked:

Social Organization: Social structure is the concrete relations among concrete individuals and concrete groups, e.g., Networks.

Social Background Characteristics: Social structure is the relations among people sharing the same social background characteristics, or differing in terms of social background characteristics.

Institutional Structures: Social structure is the relations among people, categories, etc. laid down by ongoing organizations and cultural forms.

Underlying (Deep) Social Structures: Social structures exist (somewhat irrespective of the actions of those 'carrying' them) at deeper (and more abstract) levels.

With all these, they both constrain the possibilities for social activity and provide a structured range of opportunities.

The 'Language' of Social and Structure

According to the *Oxford English Dictionary*, Alex de Tocqueville was the first to use the term 'social structure,' followed by Herbert Spencer in 1872 and then references in the 1940s (Meyer Fortes), and 1960s (Claude Levi-Strauss). Perhaps it is a quite definitive indicator that the term really did not begin to be widespreadly deployed until well into the nineteenth century.

William Sewell Jr. (2005) cites Michel Foucault to suggest that 'structure' is a general metaphor and that "such usages originated … in seventeenth and eighteenth century botany, from which they spread to other natural and social sciences." However, Raymond Williams, in his *Keywords* (1983), provides a rather more lengthy account of the linguistic history. The earliest usages of 'structure' in the fifteenth century referred to the process of building. By the seventeenth century, it referred to both the whole product and the manner of construction – the meaning emphasized the inner structure: the arrangements among constituent parts. The term was used about the anatomy, as a contrast to the functioning (or performance), of a body part. In the 1870s, usages contrasted internal arrangements and (surface) decoration. In the twentieth century in linguistics, and then other social sciences, structure signaled an analytical or formal study rather than a comparative/historical or process approach (which had been more appropriate for the study of more familiar European social structures). There are similar complexities in the meanings of related terms such as form and system. He also notes that "much structuralist analysis is formalist in the sense of separating form and content and giving form priority" (Williams, 1983: p. 306).

There is also a somewhat separate history of the concept of 'the social' which is seen to have arisen as a distinctive order or domain of social reality through various eighteenth century discourses, which theorized and reflected on this order. 'The social' outflanked the more usual social categories up to that time: e.g., the polity, stratificational orders, etc. Its death is seen as arising as it is replaced by a virtual reality comprising signs and symbols seemingly detached from material reality. Social structure can be conceptualized as particularly embedded within this domain.

Phases in Development of the Concept:

Each of the successive 'cohorts' of social structural models is schematically outlined in **Table 1**.

The Prehistory of Concepts of Social Structure

There clearly is a 'prehistory' of concepts of social structure. Several writers have drawn attention to the collectivist aspects of social theory in Greek, Roman, Arab, Medieval, and Renaissance thought. After these periods of social thought, a strong individualism was built into the models of the classical political economists (such as the Scottish school) which clearly limited their interest in social structure. However, over the

Table 1 Phases in social structural thinking

Approach	*Time-period*
'Prehistory'	Pre-1890
Founding fathers	1890–1920
British structural–functional anthropology	1920–50
American structural–functional sociology	1950–65
Inter-regnum	1965–75
Structuralism	1975–85
'Process' reconceptualizations	1985–95
Poststructuralism/post-post-structuralism	1995–present

several centuries of the 'Enlightenment Project,' qualified holistic viewpoints of the social also evolved. Nevertheless, one of the major lessons in the history of social thought is that the extraction of conceptions, both which are *social* (i.e., not essentially political, or even economic) and which also involve *analytical*, *scientific* treatment of the collective properties of social entities, are hard won. Moreover, progress at developing working theoretical models of social structure was slow, in some considerable part because of the lack of development of institutionalized traditions of social theoretical work.

In Donald Levine's (1995) treatment of the Hellenic tradition in sociology, he analyses the approach used by Aristotle to study how the good life might be achieved. Although its moral purpose was paramount, nevertheless Aristotle's was a grounded inquiry. It involved examining the natures of these entities and the reasons why they arose. Aristotle identified the three forms of association:

- households (in which people join together for the sakes of reproduction and of self-preservation);
- villages; and
- city-states (to allow expression of discourse about bad and good, and as the culmination of the social life of its components) especially as represented by their constitutions.

Aristotle then analyzed variation among political communities in terms of their internal relationships between governing and subject strata.

"The constitutive elements of associational domains are their role sets and the resources available to them (e.g., husband–wife, parent–child, master–slave, master–inanimate possession relationships in the household, and ruler–ruled relationships in the political community). Responsibilities for each role–position are specified, together with the type of socialization he deemed necessary to cultivate the best qualities required in the carrying out of those responsibilities" (Levine, 1995: p. 118). Aristotle utilized his dataset of the constitutions of more than 150 communities to examine the best sources for regime stability, especially as appropriate to communities of different sizes and socioeconomic compositions.

Various scattered attempts at social structural analysis occurred thereafter. One of the more notable was that of the medieval Muslim philosopher Ibn Khaldun, who produced an interesting sociology of the cyclical variation of social forms in the Arab world. Charles-Louis Montesquieu in the mid-eighteenth century and then Jean-Jacques Rousseau continued the focus on political forms but began to show how these are embedded within broader societal, and particularly geographical, features. French social thinkers saw social structures as things, shaped by laws, arranged into hierarchies, which had development trajectories and were composed of interdependent cultural and social structures.

An historical sociology of the evolutionary development of societies was most developed by the eighteenth century Scottish *philosophers* in their 'Four Stages Theory' of progression of types of society – hunting, pasturage, agriculture, and commerce. This broad framework was extended and deepened by other writers.

Spencer provided the culmination of this framework in his work, based on considerable secondary analysis, the internal processes within societies, and their stages of evolution. Spencer visualized each stage as being denoted by (1) a given degree of differentiation *between* regulatory, operative, and distributive processes and (2) a level of differentiation *within* each process. He saw five basic stages in terms of their complexity: (1) simple without head (leadership), (2) simple with head (leadership), (3) compound, (4) double compound, and (5) treble compound.

Classical Social Structural Analysis

Some general notions about social structure – which C.W. Mills (1960) later termed the 'Classical Tradition' of sociology – gradually developed during the course of the nineteenth century as a by-product of historical studies, although it was also used in political analysis and social commentary, albeit not much connected up with social research in the narrower sense. This model included "… basic conceptions or theories of such matters as social stratification and political authority, of the nature of bureaucracy and of capitalism, of the scale and drift of modern life, of the ambiguity of rationality, of the malaise individual men so often feel" (1960: p. 5). Such classical sociological analyses focus on at least one social grouping set within a wider social formation. These social groupings often have a particular economic role, hold power, or are centrally involved with cultural development or cultural maintenance, such as social classes, estates, broad political movements, or the 'intelligentsia.' The historical trajectory of the grouping: its rise, continuance, and fall are examined together with the grouping's terminology (especially its own names for itself), internal positions, and external boundary maintenance; its economic base, political power, and cultural legitimacy; its demography; its social and ritual life; its relations to other social groupings, and the leadership, organizations, and networks through which it is organized and can mobilize to achieve any common goals. Tocqueville is an early structural analyst who suggested that through their voluntary associating, for joint interests, both publicly and privately, Americans are able to overcome selfish individualism and to create both an active political society and an active civil society independent from the state.

However, such structural analyses were not underpinned by a formal theory. None of the now-recognized Founding Fathers of Sociology had a well-developed and explicit model of social structure, but a semi-implicit model of social structure was

central within the work of each. For Marx and Weber in particular, this particularly focused on class structure.

Marx

The main thrust of Marx's later work lay in understanding the economic machinery of capitalism. His 'sociology' is both embedded within or envelops his 'political economy': Marx saw social life being lived within 'social formations,' composed of various economic, social, and political groupings: especially social classes which shaped the individuals involved with them.

At the center of each social formation is,

> The specific economic form, in which unpaid surplus-labor is pumped out of direct producers, determines the relationship of rulers and ruled, as it grows directly out of production itself and in turn reacts upon it as a determining element. Upon this, however, is founded the formation of the economic community which grows out of the production relations themselves, thereby determining simultaneously its specific political form. It is always the direct relationship of the owners of the conditions of production to the direct producers–a relationship always naturally corresponding with a definite stage in the development of the methods of labor and thereby its social productivity–which reveals the innermost secret, the hidden base of the entire social structure, and with it the political form of the relationship of sovereignty and dependence, in short, the corresponding specific form of the state.
>
> (Marx, 1981: 791)

The State (as the executive committee of the ruling or dominant class) ensured social cohesion and cloaked capitalist society in a blanket of *false consciousness*, which hid the way in which society was organized in the ruling classes' interests.

In addition, Marx depicted the dynamics of the whole formation which centered on the extent to which such social groupings and particularly social classes might cohere sufficiently to exert a social force in their own right. An important difference is that between a *class-in-itself* (not aware of the shared nature of its material interests) and a *class-for-itself* (which is aware of its shared interests and has some organizational capacity for realizing these). A critical element in building the conflict necessary for an eventual revolution is the dissipation of the cloying *false consciousness*, and its replacement with a *revolutionary consciousness*.

In turn, the dynamics of structural change are placed within a much wider historical framework, in which the successive developments of different modes of production are placed within a sequence. In a general way, ancient (slave), medieval (feudal), and modern (capitalist) modes are seen as part of an evolutionary sequence, with the capitalist mode (and perhaps deformed feudal peasant modes) leading on to the subsequent stage of a socialist mode of production. Societal change is seen as driven by the interaction between organizational arrangements in the economy (relations of production) and technology (productive forces).

Weber

Weber's views (see especially 1947) on social structure are more developed than those of Marx, and his theoretical reflections directly pertain to social structure. Nevertheless, Weber does not attempt to develop a full understanding of social structures. Rather, his tools of structural analysis are designed to assist in the analysis of those portions of historical experience, in which he had a particular interest. On the other hand, his work is cast within a very broad comparative and historical range: encompassing study of Western, Chinese, Indian, and Ancient civilizations. A central point is 'methodological individualism' requiring explanations in terms of the involvement of individuals and also need to meet the criteria of providing *objective* as well as *meaningful* explanation.

Weber's approach to social structure, including social classes, is much more flexible than Marx's. A class is seen as any social grouping which shares a similar position within a market (e.g., the labor market), with 'similar position' meaning that its members share similar life-chances. Rather than emphasizing only the economic/material sphere, Weber sees *several* dimensions around which social strata might form: class (concerned with economic matters), party (concerned with power), and status group (concerned with social prestige or honor).

Finally, Weber provided a particularly potent model of organizational analysis (and social structural analysis more generally) with his theoretical model of a bureaucracy. He depicted this as a 'machine' of hierarchically organized 'offices' filled by meritocratically appointed, salaried, careerist 'officials' who are constrained to efficiently tackle their collective tasks within a systematic framework of rules, records, and reporting.

Durkheim

Of the three (major) founding fathers of sociology, Durkheim was the one most seriously concerned with social structure. However, even for Durkheim, the precise term is used only in passing, and references are made more generally to society, and to the social bonds constituting society.

Durkheim was aggressively antireductionist and collectivist in approach. The concept of a 'social fact' was central, and social facts were seen as separate from, and overwhelming the individual.

Social facts are a framework of collectively shared obligations on the individual, involving either 'crystallized' (e.g., expressed in adages or writing) or more diffuse 'social currents.' Usually the collective shaping of a person is internalized and constant pressure is not needed. Such social pressure is felt most when a person resists obligations. This results in conformity to them because social groups or the State upholds them. At a more descriptive level, Durkheim developed a conception of *social morphology* involving accounts of the nature (e.g., size, functions), number, arrangement (e.g., spatial distribution), and interrelations (e.g., modes of communication, movement, and mutual obligation) among the social parts (individual or collective) of a society.

Durkheim clearly postulated that the central conceptual problem on analyzing social structures was its 'division of labor,' and the consequence for the structure's type and level of social integration. Whereas primitive social structures were seen as organized around similarity in their task structures, which was complemented by a tight 'repressive' social control, more complex societies had a division of labor characterized by complementary specialization, and this was then matched by a system of social control which emphasized 'restitution.'

Simmel

Of other sociologists in the 'founding' generation, the work of Simmel has proved to be of lasting significance. Simmel developed a 'formal' sociology (as opposed to sociological treatment of 'content'), in which properties of small groups and social situations were seen to affect the behavior. His approach could allow the compilation of systematic propositions about the operation of social structures. For example, Simmel wrote about the effects of differences in sizes of groups for interaction among their members. The conception of abstracting social content in order to reveal structural similarities across diverse situations also proved to be helpful to later analysts.

Diverse social groups may nevertheless show identical forms of behavior toward one another on the part of individual members – relationships such as superiority and subordination, competition, division of labor, formation of parties, representation, and inner solidarity coupled with exclusiveness toward the outside can be found in the state, in a religious community, in a band of conspirators, kin, in an economic association, in an art school, and in the family. Simmel's imagery of the way in which people in modern societies are complexly related was through cross-cutting social circles, as opposed to the concentric circles of more primordial situations.

Structuralist Approaches

British Structural–Functional Anthropology

Beginning in the 1910s, and maturing in the immediate post-World War I period, a self-conscious analytical approach to social structural analysis was developed by the British school of structural–functional anthropologists. This was begun by Malinowski, but predominantly shaped by Radcliffe-Brown (1957), and later summarized in a wider range of texts (e.g., Nadel's theoretical treatise: 1957). Earlier approaches in which emphasis was placed on the evolution and diffusion of cultural traits were rejected in favor of understanding the current operation of small-scale societies (and later, rather larger African ones). Any social practice was seen in terms of its 'role' within its societal context, which enabled 'Strange' tribal practices to be 'mapped onto' more commonly understood Western concepts. This approach was considerably inspired by Durkheim's views on social collectivities.

There were at least two main strands, with Malinowski endeavoring to reach down to the ways in which societies met the sociobiological needs of their members, while Radcliffe-Brown emphasized a more strictly social structural level of analysis. The latter approach came to dominate in structural analyses but the theme of how individuals relate to social structures reoccurs in later views and debates.

American Structural–Functional Sociology

During the 1950s and 1960s, American sociology was dominated by an approach generally labeled as 'structural-functionalist,' or (retrospectively) to use Mullins's (1973) term as 'Standard American Sociology' (SAS). To a considerable extent, this involved a translation of British structural–functional anthropology into the then-contemporary and larger scale American context. Although the functional mode of explanation (in terms of consequences, especially for environing social systems) continued, as with its anthropological predecessor, the inherent teleology has been later resisted, although functional analyses can still be seen as a useful heuristic approach for generating hypotheses.

There was a strong and consistent attention to the analysis of social structure through concepts such as 'role theory' and 'social system' and others. Another path lays in the development of research tools more appropriate to the investigation of the structures of large-scale societies, especially large-scale surveys. Theoretical and methodological advances often benefited each other. A range of standard social background variables was deployed in surveys, which were useful in predicting patterns of behaviors or attitudes. As Merton (1949) pointed out, these social background variables reflected the main social statuses which people might occupy.

While much of this sociology retained a strong interest in characteristics of individuals (or patterns among various of the properties of individuals) which the survey research methodology made possible, surveys were also employed to study the 'emergent' or 'structural' characteristics of various types of social unit such as groups, households/families, organizations, suburbs, regions, and nations. As well, considerable theoretical and methodological knowledge accumulated about how relations between individuals and various types of social entities might best be studied and explained. A useful example is the Lazarsfeld typology of the different types of individual and collective properties, in which individuals can influence other individuals or collectivities and in turn collectivities can shape individuals or other collectivities.

At a more macrosociological level, analyses of which social dimensions (social cleavages) were more salient in describing patterns of societal development and operation. For example, societies with cross-cutting cleavages were more likely not to be crippled by major conflicts since the widespread sharing of interests held in common by 'parties,' which stood in different positions in relation to a dispute, is likely to dampen its socially destructive effect. Comparative historical sociologists studied questions such as the degree of support for left-wing political movements, finding their explanations in terms of other features of the social structure: for example, lack of socialist development in USA was seen as flowing from the (supposedly) quite open-ended texture of the American social structure, and especially the possibility that upward mobility might be achieved by all.

The 'human ecology' aspects of the 1920s Chicago school of sociology were formalized by Duncan and Schnore into a 'POET' framework, which analyzed the interrelations among Population, Organization, Environment, and Technology (see also Hawley, 1986). This schema usefully pointed to the natural environment context of social structures and some of the key elements which shaped their long-term operations.

Role theory was an important conceptual foundation stone. Often the concept was used generally and descriptively to indicate that social life was socially patterned. In some theoretical approaches, emphasis was laid on the way roles and their occupants were shaped by social learning and monitoring by role-others, rather than as in the 'symbolic interactionist' formulation in which role-occupant's orientation to their role was stressed. Panoply of role-related concepts piled up

(e.g., Merton's status-and-role theory which allowed finer conceptual discriminations).

Parsons

The 'grand theoretical' work of Parsons endeavored to lay down a solid conceptual foundation and standard vocabulary for the social sciences, in particular, anthropology, sociology, and social psychology (e.g., 1973). The three interlinked master concepts – culture, social systems, and personality – are held to be separate, autonomous, and yet interpenetrating analytical systems. A social system is seen as the interaction of the plurality of persons and their relationships, with its central building block being that of the role. Roles are a sliver of someone's personality, but are more than this in that they also stand outside the personality. Roles constitute and govern the interaction between people – each in their respective roles – through providing complementary and mutually adjusting sets of expectations. Many roles are 'institutionalized' by being built into wider cultural patterns and controlled by social sanctions. However, the social system seems to be a residual concept and the cultural determinism is so overwhelming that there is little room left for autonomous social activity.

French Structuralism

During the late 1960s, two forms of structural analysis – both drawing inspiration from Marxism, and both emanating from Paris – captured much attention. Despite their relatively brief span of popularity, both marked out the extremes of a 'structural determinist' approach. For French anthropological theorist Claude Levi-Strauss, "The term 'social structure' has nothing to do with empirical reality but with models built up after it" (1963). Indeed, the drive of his analysis is to show how styles of 'native' thinking (in whichever culture is being studied) across all areas of cultural and social life are all embedded within deep semilinguistic patterns of thinking (possibly even mentalistic, being hard-wired into the digital structure of the brain). Thus, 'native' concepts of kinship patterns, religious ideology, and all else is subsumed under broad models of thinking that eventually reduce down, via transformation rules, into dualities dwelling in the mind. Levi-Straus's approach showed mind-stretching linkages, such as postulated themes running through the entire range of North American tribal cultures, but it seemed to falter in the hands of others.

The other structuralist development, which peaked at much the same time, was centered on the work of social philosopher Louis Althusser. In this approach of 'structural Marxism,' Marx was (re)interpreted to fit a highly determinist and scientific understanding of social development. Marx's model of the working of the capitalist economy was accepted as adequate and instead, attention focused on two consequent issues:

- How Marxism might be construed to offer satisfactory explanations of societies other than capitalist ones (and indeed to provide a theorized sequence of types of society)?
- How the whole of a society was to be seen as operating? A threefold distinction is assembled: society is seen as comprising the economy, the polity, and the ideological state apparatus. The functional role of the latter two sectors is to repress the population, by force and by imposing ideological consensus, into supporting the status quo set by the current mode of economic organization. A society can be described by which of the three structures is dominant; but no matter which is the dominant structure, it is always the economic structure that determines which is dominant: the economy is the determinant 'in the last instance.' In capitalism, the economy is the main institution, whereas earlier societies had been dominated by the family, or in feudal times by political and religious institutions.

Structuralism: 'Standard American Sociology' Supplemented

By the late 1960s and early 1970s, the conceptualizations used in SAS had not only begun to wane in popularity, but they were also directly attacked. Microsociologists were concerned that SAS required too formal and reified a view of people's relationship to the social structure: an 'oversocialized' conception of mankind. On the other hand, the Standard American social structural analysts were criticized for dealing with trivia, for ignoring major societal level phenomena outside the framework they were working within, for lack of any historical or comparative perspective, and for an implicit conservative support of the status quo. To overcome these limitations, the next period of development tended to emphasize structural approaches.

Conflict Theory

The critique of SAS as being overly concerned with 'consensus' was consolidated for a decade or so through the development of a supposedly polar opposite 'conflict' view: a 'soft-core' Marxism, or rather a 'left-Weberianism' whose proponents include Simmel as a fore-runner, Lewis Coser, Ralph Dahrendorf, and John Rex (e.g., Dahrendorf, 1968). This conflict, in contrast to the previously predominant emphasis on consensus, was useful for mobilizing alternative biases, but the distinction proved difficult to sustain over the longer term as it became realized that the two approaches could be seen as essentially complementary, sharing an underlying view of social structure.

The Social Construction of Reality

Berger and Luckmann (1967) developed in the late 1960s a sophisticated account of the 'social construction of reality' which outflanked and advanced the treatments then available. They emphasized the fragility of social reality and the involvement of people in constructing those very social structures, they then came to experience as solid and real, with mechanisms to cover-over any broaches in their facticity, including the sharing of experiential commonalities, the deployment of aphorisms and sayings, and then relating any remaining difficulties people had in fitting into their constructed reality, to broader underlying conceptualizations shared in that culture. This approach points to the major role of legitimating cultural structures, such as religion, which serve to secure the social structure.

Symbolic interactionists and kindred microsociologists tended to remain restricted to situational rather than structural

analyses, but some attempts were made to explore compatible links with more macrosociological topics purveying the general idea that larger social structures involve an often-fragile 'negotiated order,' in which the various groupings involved participate on an ongoing basis (e.g., see Maines, 1991).

Formal Sociology: Blau's Attempted Synopsis and Program

In 1975, Peter Blau organized a group of eminent social theorists to directly confront the concept of 'social structure,' including several involved in historically orientated or comparative studies, as well as those more inclined to survey and other quantitative work. Blau's definition is that "Social structure refers to the patterns discernible in social life, the regularities observed, the configurations detected."

Blau's own program of structural analysis is termed a 'primitive theory of social structure' which pursues a program of 'radical structuralism' which sticks to sheer objective structural features while culture is seen as irrelevant. (See also the extreme views of Mayhew (1980/81) concerning structure.)

Porpora (e.g., 1987) attacks this position by denigrating the deductivist philosophy of science at the heart of structuralist sociology. Rather, an alternative Marxian view of social structure is advanced which explores the relations between a realist and a social constructionist view.

Theoretical Reflections

In addition to these various alternative models of social structure, more specific conceptions were also explored.

- Network theory.
- Organizational Sociology.
- Social Movement theory.

The SAS approach has continued (and has continued to develop) within Sociology (especially in North American sociology), and many social science journals remain filled with studies carried out within its broad paradigm. A common rhetoric shared by many of these more sophisticated versions of SAS cleaves to the formalization, mathematical, and statistical aspects of social structure. For example, White (2008) develops an abstract and difficult conceptualization of social structure, which complements the treatment of network linkages with an extension into the cultural aspects which also shape them.

Kontopoulos (1993) developed a whole series of models of social structure in which their central logic is laid out cleanly to be inspected, leading to the argument that the model of 'heterarchy' is most appropriate for capturing the complexities of hierarchy and feedback so rife in social life.

Under the leadership of Alexander (and in the writing of several important Continental social theorists, including Habermas and Luhmann), a 'neofunctionalist' revival attempted to reclaim much of the theoretical territory earlier dominated by Parsons through the addition of a theory of group formation and conflict (Alexander, 1985, 1987).

David Lockwood (1992) also revisited Parsons's earlier work on the differences between integration at the systems level and integration at the social level. He carefully reviews the relevance of Marxian and Durkheimian theories, each with their dominant and supplementary contributions in providing explanations at both these levels, and indicates how they might be combined.

'Process' Reconceptualizations of Social Structure

By the late 1970s, the critique of ideas about social structure had turned in a much more strictly theoretical direction, although this assisted in the demise of interest in 'social structure' because it became detached from concern with social issues. One line of theoretical debate focuses on the relationship between the macro- and microlevels, while another angle on it concerns the relationship between agency and structure.

The 'figuration' word picture provided by Norbert Elias pointed to "The network of interdependencies among human beings is what binds them together." But two major theorists have brought out in-depth discussions of social structure: Bourdieu's 'social constructivism' and Giddens's 'structuration theory.'

Bourdieu has developed a sophisticated and appealing imagery of social structure and has connected this up with a wide range of empirical studies of particular social structures. His analysis works between the collective and individually based concepts, and between the subjective and the objective and in a correspondingly mediating position in terms of the other antinomies. Key concepts include 'habitus' (the mental schemas that people employ to order their world and which shapes their actions), 'capital,' and 'field': although these key terms are supplemented by many others.

In a series of theoretical works stretching from the mid-1970s, Anthony Giddens (e.g., 1984) has developed the most sophisticated generally available theory of social structure, usually advanced under the banner of the theory of 'structuration'. Giddens is less concerned with developing a conceptual framework, which would then allow the derivation of more specific theories, than with endeavoring to recover from the available stock of social theories, a vision of social structure which can be philosophically sustained.

In reaction to Giddens's somewhat mystical rendering of the linkage between agency and structure, Margaret Archer (e.g., 2003) has championed a 'morphogenetic' approach, drawn in part from earlier theorists and general systems theory, in which the various levels and phases and types of consequence are explicitly spelled out. In this approach, structure and action operate over different time periods, with structure leading to action which may have the consequence of structural elaboration. More recently, this approach has been developed by a book on the relations between culture and social structure, and another on social structure.

Sewell (2005) recognizes the untrammeled cognitive power of the concept of social structure, and sees it as afflicted by the three problems of appearing too hard, too impervious to change, and too open to a variety of meanings (in particular, the difference between the 'social structure' and 'cultural structure'). He seeks to clean up the concept from these three contaminants. He uses the work of Giddens and Bourdieu to endeavor to provide a sense of agency and thus overcome the too rigid immutability of social structure; explores the potential for providing an account of social change; and stresses the

involvement and interpenetration of both culture and resources within his concept of social structure.

Mouzelis (e.g., 1995) has been involved in a program of theoretical work to reclaim the centrality of sociological theory from the clutches of philosophy, doctrinaire Marxism, or culturalist idealism into which it too easily has recently fallen. Central is the hierarchy of social levels, and their myriad connectiveness. Giddens's concept of 'dualisms' is opened up when he points out that people vary in their relationship to the social structures they are in: sometimes being involved with merely reproducing the social structure while at other times standing off from it. He attempts to breakthrough the usual demarcation drawn between the micro- and the macrolevel by suggesting that some apparently microsituations have important macro-consequences. Analysis, he argues, should also concentrate as much on social integration as systems integration.

On the whole, Bourdieu's schema seems rather more satisfactory sociologically, than that of Giddens. Whereas Giddens's 'memory traces' seem quite psychological, Bourdieu's 'habitus' is clearly not only shared across each of the different groupings (e.g., classes), but further is an instrument of social differentiation in that different social groupings have different habituses: they are often complementary to each other within a wider system. Bourdieu's main drawback is that his analysis exudes a mysticism, which is not carefully exercised from his conceptual thinking. Other commentaries abound, but alternative theoretical programs developing these 'process reconceptualizations of social structure' have yet to gain any momentum (Elder-Vass, 2012; López, 2003; Parker, 2000).

More recently, some postmodernist writings (e.g., Baudrilliard, 1983) have challenged structural analysis as an explanatory approach, essentially by declaring that the subject-matter it is concerned with has evaporated under a deluge of simulacra and electronically borne imageries. The stronger version of postmodernist arguments can be that people's choices are so socially unconstrained that they are located merely in a 'statistical mass' (or even 'random mass') devoid of social control (although, an alternative argument is that the extent of cultural control through media and cultural images is overwhelming). The weaker version of this argument posits that the hold of modern social structures has loosened, and that the opportunity space of people's choices has so widened that the standard verities of sociology have lost their explanatory purchase. No clear empirical support for this viewpoint is adduced, and it is arguable whether this critique is able to be arbitrated by empirical research. In the absence of clearer empirical support, it seems reasonable to suggest that the 'end of the social' argument tends to have been painted in too brilliant hues. Indeed, the death of the social has been much exaggerated.

The most recent approach to analysis of social structure is that of 'analytical sociology' (e.g., Hedström and Bearman, 2009) which sees social structures as emerging from particularly social mechanisms that operate at an individual level, although buffeted by the social interaction contexts of the people involved. A discrete – but considerable – array of such mechanisms have been identified, although it remains unclear the extent to which this approach allows adequate analyses of social structures at the more aggregate or collective levels.

Conclusions

Conceptual debate concerning social structure has most importantly developed in sociology, together with attendant social philosophy but an interest in, and use of, the concept is spread across the other social sciences, but without raising further issues. Anthropology has a lively discourse, especially about the relationship between culture and social structure and Geography and Political science (but also economics and social psychology) all actively deploy the concept in many empirical analyses.

The task of pinning down the best way to conceptualize social structures continues to slowly advance. If the way ahead is not clear, certainly lessons can be derived from reflecting on the history of social structural analysis noting several common difficulties that any schema needs to be concerned with:

- It seems useful to quite clearly separate social structure both from its (external) causes and its effects.
- As well as seeking understanding of how a social structure operates, it is important to ground this understanding in the ongoing everyday life of members, and also to contextualize the social structure in terms of its long-term tendencies, and its relations with other social structures.
- Particular care is needed to link micro (e.g., the interactional order) and macro aspects of social structures, and it is particularly important to capture both the determined (structure) and the voluntaristic (agency) aspects of social life.
- Another distinction which may be helpful is that between culture and social structure, which certainly must always be addressed.

See also: AGIL, History of; Bourdieu, Pierre (1930–2002); Bureaucracy, Sociology of; Cooperation and Competition; Determinism: Social and Economic; Durkheim, Emile (1858–1917); Giddens, Anthony (1938–); Groups, Sociology of; Individualism versus Collectivism: Philosophical Aspects; Individuals in Society: History of the Concept; Integration, Social; Labor, Division of; Macrosociology-Microsociology; Merton, Robert K (1910–2003); Organizations, Sociology of; Parsons, Talcott (1902–79); Reference Group, History of; Sample Surveys: Cognitive Aspects of Survey Design; Simmel, Georg (1858–1918); Social Constructionism; Social Constructivism; Social Networks; Social Ontology; Social Properties (Facts and Entities): Philosophical Aspects; Social Structure; Social, The: History of the Concept; Society and People: History of the Concept; Solidarity, Sociology of; Status and Role: Structural Aspects; Weberian Social Thought, History of.

Bibliography

Alexander, Jeffrey (Ed.), 1985. Neo-functionalism. Sage, Beverley Hills.

Alexander, Jeffrey (Ed.), 1987. The Micro–Macro Link. University of California Press, Berkeley.

Archer, Margaret, 2003. Structure, Agency, and the Internal Conversation. Cambridge University Press, New York.

Baudrilliard, Jean, 1983. In the Shadow of the Silent Majorities or the End of the Social. Semiotext, New York.
Berger, Peter, Luckmann, Thomas, 1967. The Social Construction of Social Reality. Penguin, Harmondsworth.
Blau, Peter (Ed.), 1975. Approaches to the Study of Social Structure. Free Press, New York.
Callinicos, Alex, 2007. Social Theory: A Historical Introduction, second ed. Polity Press, Cambridge.
Chew, Sing, Knottnerus, J. David, 2002. Structure, Culture, and History: Recent Issues in Social Theory. Rowman & Littlefield Publishers, Lanham, MD.
Crothers, Charles, 1996. Social Structure. Routledge, London.
Dahrendorf, Ralph, 1968. Essays in the Theory of Society. Routledge and Kegan Paul, London.
Elder-Vass, Dave, 2012. The Reality of Social Construction. Cambridge University Press, Cambridge, New York.
Giddens, Anthony, 1984. The Constitution of Society. Polity Press, Oxford.
Hawley, Amos, 1986. Human Ecology: A Theoretical Essay. University of Chicago Press, Chicago.
Hedström, Peter, Bearman, Peter (Eds.), 2009. The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford.
Kontopoulos, Kyriakos, 1993. The Logics of Social Structure. Cambridge University Press, Cambridge.
Levi-Strauss, Claude, 1963. Structural Anthropology. Basic Books, New York.
Levine, Donald, 1995. Visions of the Sociological Tradition. University of Chicago Press.
Lockwood, David, 1992. Solidarity and Schism: 'The Problem of Disorder' in Durkheimian and Marxist Sociology. Clarendon Press, Oxford.
López, José, 2003. Society and Its Metaphors: Language, Social Theory and Social Structure. Continuum, New York; London.
López, José, Scott, J., 2000. Social Structure. Open University Press, Philadelphia, PA (Concepts in the social sciences).
Maines, David (Ed.), 1991. Social Organization and Social Process: Essays in Honor of Anselm Strauss. Aldine de Gruyter, New York.
Marx, Karl, 1981/1882. Capital: A Critique of Political Economy. Penguin, London.
Mayhew, Bruce, 1980/1981. Structuralism v individualism. Social Forces 59, 335–375, 627–648.
Merton, Robert K., 1949/1968. Social Theory and Social Structure. Free Press, Glencoe, IL.
Mills, C. Wright (Ed.), 1960. Images of Man: The Classical Tradition in Sociological Thinking. Braziller, New York.
Mouzelis, Nicos, 1995. Sociological Theory: What Went Wrong? Routledge, London.
Mullins, Nicholas, 1973. Theories and Theory-Groups in Contemporary American Sociology. Holt, Rinehart and Winston, New York.
Nadel, S.F., 1957. The Theory of Social Structure. Melbourne University Press, Melbourne.
Parker, John, 2000. Structuration. Open University, Phildelphia, PA.
Parsons, Talcott, 1973. Culture and social structure revisited. In: Schneider, L., Bonjean, C. (Eds.), The Idea of Culture in the Social Sciences. Cambridge University Press, Cambridge.
Porpora, Douglas, 1987. The Concept of Social Structure. Greenwood Press, Westport, CT.
Radcliffe-Brown, Alfred R., 1957. A Natural Science of Society. Free Press, Glencoe, IL.
Sewell, William, 2005. Logics of History: Social Theory and Social Transformation. University of Chicago Press, Chicago.
Weber, Max, 1947. The Theory of Social and Economic Organization. Free Press, Glencoe, IL.
White, Harrison, 2008. Identity and Control: How Social Formations Emerge, second ed. Princeton University Press, Princeton.
Williams, Raymond, 1983. Keywords: A Vocabulary of Culture and Society. Oxford University Press, Oxford.

Social Studies Education

Antonio J Castro and Ryan T Knowles, University of Missouri, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field of social studies education encompasses a variety of perspectives about the nature of social studies subject matter and its goals. This article traces twin aspects of social studies education: social studies as social science and social studies as social education. Scholars who see social studies as social science education view the field as incorporating the disciplinary concepts, knowledge-based, and strategies of a variety of unique disciplines (history, geography, economics, etc.). Those who advocate for social studies as social education place civic and democratic teaching as central to the goals and outcomes of social studies. Both approaches, however, serve to prepare youth to engage in an increasingly complex and diverse global world.

The term 'social studies education' originated in the United States as part of the 1916 report of the Committee on Social Studies for the Commission for the Reorganization of Secondary Education (Lybarger, 1990). Since its inception and the 1917 founding of the National Council for the Social Studies – an organization created to advocate for social studies education – tension and conflict continue to surround the meaning and purpose of 'social studies education' (Evans, 2006; Saxe, 1991). Jack Nelson (2001) explained, "Defining social studies is not an easy task; it is encumbered by a confounding history, conflicting conceptual ideas, and strong ideological divergence in both political and educational philosophy" (p. 15). For some, social studies education acts as an umbrella of various disciplines, such as history, geography, economics, and political science. For others, social studies education embodies a unique attention to the social and civic aspects of schooling and learning. Stephen Thornton (2008) labeled these two general approaches as social studies for social science and social studies for social education. In this article, we review concepts and ideas related to these two aspects of social studies education and suggest that these stances have the overlapping goals of preparing youth for later social and civic life.

Before we proceed further, we must address the international scope of social studies education. Keith Barton and Patricia Avery (in press) noted that internationally the term 'social studies' may have little significance to educators in different countries. In some countries, history and geography are taught independently of each other. In others, the terms *civics*, *social education*, and *social sciences* are used often to describe various coursework that could be considered 'social studies' in the American context. Furthermore, international research indicates that different purposes and meanings of history-centered teaching and/or social education occur given the context of each country and its political and educational school systems. For example, international researchers have found that history education and civic ideas are based on their unique national contexts (see Barton, 2001; Ho, 2010; Levstik and Groth, 2005; Pike, 2000). Regardless, the concepts discussed in this article will have relevance to a wide range of educators and researchers across the globe.

Social Studies as Social Science

The stance that social studies represent a variety of disciplines embodies a set of key assumptions. First, this approach grants each discipline (history, geography, economics, political science, etc.) its own status as an academic pursuit. John Braxton and Lowell Hargens (1996) described important characteristics of an academic discipline: codification or consolidation of knowledge into specific domains, paradigm development, and consensus thinking – both of the later describing common theoretical and disciplinary habits of mind. Second, this view holds that students in various levels of schooling can acquire foundational aspects of each discipline. American curriculum theorist, Jerome Bruner (1960), in his *The Process of Education*, proposed that "Any subject can be taught effectively in some intellectually honest form to any child at any stage of development" (p. 33). In order to accomplish this task, Bruner asserted that each discipline had a certain base structure, which included analytical skills and habits of mind or orientations used to construct and solve problems. Through the use of spiraled curriculum, children could engage in different aspects of the structure of the discipline without necessarily needing to focus on the accumulation of facts. Finally, this approach positions the 'social studies' as a loose label meant to group these disciplines around their general focus on understanding social life.

Here, we discuss the research and ideas related to three disciplines: history, geography, and economics. Given the assumptions listed above, the approach taken to the teaching and research of these areas relies first on identifying a set of domain-specific knowledge, techniques, or analytical skills. Second, scholars and educators then attempt to uncover the practices, challenges, and influences on the learning and teaching of the discipline.

History Education

Much of research in history education arose in the United Kingdom in the 1970s and 1980s at a time when curriculum reform efforts in history centered on teaching youth disciplinary concepts and skills related to history (Barton and Avery, in press). Later, work in the United States and Canada followed

suit in tracing the ways in which students interpreted historical events and thought about evidence (Barton, 2008).

In investigating history education research, scholars first established certain parameters of domain-specific knowledge. VanSledright and Limon (2006) described two kinds of knowledge related to history learning: substantive knowledge and procedural knowledge. Substantive knowledge includes both conceptual and narrative ideas about what happened in the past and overarching themes and trends that occur in historical accounts. Procedural knowledge relates to the knowledge of how to research and interpret the historical past, using a variety of analytical skills and tools.

In addition to knowledge, scholars also asserted a set of habits of mind employed by historians to solve problems of interpreting the past, called historical thinking. Levesque (2009) explained the difficulty of historical thinking,

> Historical thinking is, indeed, far more sophisticated and demanding than mastering substantive (content) knowledge in that it requires the acquisition of such knowledge to understand the procedures employed to investigate its aspects and conflicting meanings…To think historically is thus to understand how knowledge has been constructed and what it means. Without such sophisticated insight into ideas, peoples and actions, it becomes impossible to adjudicate between competing versions (or visions) of the past (p. 27).

Hence, historical thinking embodies the processes by which a historical account is constructed from various pieces of evidence. Scholars have identified different aspects of historical thinking, including historical evidence, continuity and change, progress and decline, historical empathy, and historical agency (Barton and Levstik, 2004; Levesque, 2009; VanSledright, 2011; Wineburg, 2001). Skills in evaluating historical evidence would encompass analytical techniques to determine the appropriateness of the source, its original purpose, its context, and any world assumptions or biases of the author of the source, as well as corroborating its information with other sources (Levesque, 2009). On the other hand, historical empathy implies the ability on the part of the historian to understand the context of the actors. According to Barton and Levstik (2004), historical empathy embodies a caring aspect as well as perspective recognition. In the latter, the historian must acknowledge a sense of normalcy in the everyday actions of those living in the past – even if these actions or mores seem inconsistent with the current context of the historian. Exhibiting historical empathy can be difficult for youth and students to grasp (Brooks, 2009). Historical thinking skills and specific domain knowledge have been at the base of research on the teaching and learning of history (Barton, 2008; Hicks et al., 2012; VanSledright and Limon, 2006).

Research in history education addresses a variety of different aspects of how youth come to understand and make sense of historical concepts and tools. For example, VanSledright (2002) documented how he taught a historical inquiry-based unit in a fifth grade classroom and how students responded to this instruction. Wineburg and colleagues (1997, 1991, 2001) investigated how different background knowledge affects the way teachers approach the teaching of history, differences between high school students' conceptions and use of historical interpretation as compared to that of a professional historian, and differences across gender in their construction of important historical events. Terrie Epstein (2009) captured how African-American high school students interpreted American history differently from their White counterparts. Likewise, Porat (2004) discovered that Israeli students' cultural memory (either conservative or more liberal) influenced their reading of history texts. These are but a few studies of a myriad of studies on the teaching and learning of history.

Geographic Education

Like research done in history education, scholars in geographic education investigate the teaching and learning of domain-specific skills and ways of thinking associated with the discipline of geography. In the United States and several countries, geography instruction has revolved around the teaching of the five themes of geography (Schoenfeldt, 2002). These themes include the concepts of location, region, place, human–environment interaction, and movement. The idea of location encompasses the absolute location of an area based on longitude and latitude lines and relative location, a more general description of where an area is situated and its surroundings. Region describes an area based on a particular physical characteristic (mountainous, river valley, tundra, etc.) and/or a human-made characteristic (political regions, cultural regions). The geographic idea of place describes both the human and physical aspects of an area that give it a distinctive feeling or atmosphere. The rise of place-based education in geographic education builds on the notion that if students interact with a place (e.g., visit a wildlife preserve or forest) they will be more environmentally conscious (Resor, 2010). Human–environment interaction refers to the ways in which human beings are affected by, adapt to, and change the environment. Finally, movement relates to the movement of people and goods through transportation and the movement of ideas through communication.

While the five themes capture a variety of ideas in geography, geographic literacy includes more specific processes. In the United States, the report of the Geographic Education National Standards Project (1994), *Geography for Life*, laid the foundations for geographic skills and ways of thinking that should inform the scope and sequence for geography education in the public schools. This report outlined a matrix of skills that include understanding the world in spatial ways, distinguishing places and regions, identifying unique physical and human systems, relating how environment and society interact with each other, and knowing the varied uses of geography (Hume and Boehm, 2001).

Barney Warf (1999) described the research in geographic education as a varied enterprise. He wrote, "Because education is an enormous, diverse, and complex set of tasks, ranging from instructing students with disabilities to curriculum design to outcomes assessment, the literature in geography inevitably mirrors the school of that project" (p. 587). Despite its wide scope, a few central concepts have dominated much of the research in geography education (VanSledright and Limon, 2006; Segall and Helfenbein, 2008). Research studies have focused on what children and youth know and understand

about geographic concepts. For example, Hickey and Bein (1996) identified concepts like location and directionality that are difficult for primary grade students, ideas like landforms and latitude versus longitude that pose challenges for intermediate grade students, issues of cultural geography and climate that middle grade students struggle with, and skills with decoding graphs representing climatic information and physical geography related to plate tectonics as challenging to high school students. Other studies traced how students come to understand the different themes of geography (Trygestad, 1997). One study by Thomas and Willinsky (1999) compared mental maps of the Pacific Ocean region drawn by students in 11 countries and two US states, finding that although students differed, students shared similarities in their mapping, suggesting that they believed that those living in the Pacific Ocean region had geographic, if not cultural, closeness with each other.

In addition to this research, several studies focus on the geographic knowledge and preparation of teachers. For example, British scholar Walford (1996) classified four ways that teachers viewed geography: *interactionists*, which stressed the interdepenendence of humans and their environment; *synthesizers* who utilize various disciplines in the study of the earth; *spatialists* who study the distribution, relations, processes, and consequences of geographic resources; and *placeists* who focus physical geography and the location of places and regions. Despite these orientations, however, Catling (2004) discovered that among the 218 elementary teachers surveyed, most participants defined geography as limited to the learning of facts, names, places, and physical aspects. Indeed, researchers have lamented over the incomplete preparation and knowledge-base of teachers in terms of geography (Segall and Helfenbein, 2008).

Finally, research in geography education also investigates the role of curricular policies and contexts on the teaching and learning of geography. Of recent concern, the move toward standardization and accountability poses threats to the teaching of geography, especially when geography subjects and social studies may not be tested (Segall and Helfenbein, 2008). Robert Bednarz and Sarah Witham Bednarz (2008) noted, "Geography runs the risk of becoming the 'subject left behind,' especially in today's politically conservative environment in which history is privileged among the social sciences" (p. 24). Other challenges, such as limited space in the curriculum and the lack of technology needed to explore tools such as Geographic Information Systems or computer-assisted cartography, continue to be pressing concerns for scholars (Segall and Helfenbein, 2008).

Economics Education

Like geography and history, economics education centers on a specific base of disciplinary concepts and skills. In 1995, the Joint Council on Economics Education in the United States published a final revision of the *Framework for Teaching Basic Concepts* in economics (Saunders and Gilliard, 1995). This framework included 21 concepts divided into four major categories. The first category, fundamental concepts, incorporated ideas such as scarcity, productivity, incentives, and opportunity costs. The next two categories, micro- and macroeconomics, included a variety of concepts that affect national and global economics, such as markets, supply and demand, exchange rates, interest rates, fiscal policy, and income distribution. Finally, international economics, the fourth category, introduced students to the topics of comparative advantage, barriers to trade, balance of trade, and economic growth. Although this concept-based approach appears relevant in the US contexts (Walstad, 2001), Miller and VanFossen (2008) noted that others scholars have provided different models to define economic literacy, or the ability to think like an economist. For example, Dahl (1998) characterized economic literacy as encompassing these six key ideas:

- There is no such thing as free lunch.
- Thinking incrementally.
- Markets coordinate consumption and production.
- Relative price changes guide decision-making.
- Trade promotes growth.
- Markets can fail.

Regardless of challenges to the concept-based model, the rise of standardized testing in the United States and the push for economics education as a stand-alone course has led to the dominance of the concept-based model to guide precollegiate economics education (Walstad, 2001). However, internationally, countries like the United Kingdom (Davies and Durden, 2010) and Korea (Hahn and Jang, 2010) see much less emphasis on economics as a subject for study among high school students. Instead, economic themes or principles are interwoven into an already crowded curriculum. So, variation across international educational contexts, governmental policies, and education reforms will greatly affect economics teaching and learning opportunities for students in schools.

Like geography education, the research in economics education appears varied and complex. Early foundational research into children's economic thinking began in Italy with the work of Berti and colleagues (Berti and Bombi, 1988; Berti, 1995). These studies explored children's conceptions of banking and monetary systems. Later studies of children's economic reasoning adopted a Piagetian stage development approach whereby children's understanding and misconceptions of economic principles become more sophisticated as they progress developmentally (Miller and VanFossen, 2008). Research among high school aged students explores various attitudes about economics (Phipps and Clark, 1993; Walstad and VanScoyc, 1990), students' background knowledge and competency (Becker et al., 1990; Walstad and Rebeck, 2001; Harris Interactive, 2005), and degrees in which students engaged in economic reasoning (Miller and VanFossen, 1994; VanFossen, 1995). A few studies have investigated the role of various instructional models, like problem-based learning, on economics education (see Son and VanSickle, 2000).

The stance that views social studies as social science has produced significant insights into the learning and teaching of various social studies subjects. Barton and Avery (in press) reminded scholars that any investigation, whether historical or geographical or economical, will always borrow from the key concepts of other disciplines. For example, how can a historian understand a civil war without understanding both the geographical and economic concepts of the time period?

Social Studies as Social Education

Unlike social studies as social science, social studies for social education views the purpose for teaching as preparing youth for their later participation in adult life and citizenship. Indeed, many, if not most, educators view the purpose of social studies as promoting the knowledge and skills necessary for civic participation (Barton and Avery, in press). Even with such strong support, education for citizenship can vary tremendously depending on the context of the school, teachers, and students. A leading researcher in international citizenship education, Carole Hahn (1998) asserted, "there is no one form of democracy and there is no one way of teaching democracy" (p. 236). As a result, many conflicts exist in relation to the role of social studies to promote citizenship.

Joel Westheimer and Joseph Kahne (2004) identified three kinds of citizens promoted in schools. First, the personally responsible citizen works and pays taxes, obeys laws, and is generally responsible within their community. In a crisis, these citizens may donate money or materials necessary to alleviate the problem. Secondly, a participatory citizen has an active orientation that works to organize their community and influence government policy toward the common good. During a crisis, these people are likely on the front lines organizing and collecting goods. Lastly, the social-justice-oriented citizens critically assess social, political, and economic structures to understand root causes of social problems. These people ultimately seek out avenues for change at the societal level. Westheimer and Kahne believed that democracy requires all three types of citizens; however, the participatory and social-justice orientations contribute the most to democratic societies.

Research on students focuses on views of citizenship and democracy. Large-scale international surveys demonstrate widespread student support for democratic values, such as free elections, strong civil society, women in politics, and freedom of speech (Torney-Purta et al., 2001; Schulz et al., 2010). In contrast to these findings, other research shows that students generally support 'thin' conceptions of citizenship similar to the personally responsible citizen described above. Students are very likely to anticipate voting, but less likely to anticipate taking part in more active political behaviors, such as collecting money, gathering signatures on a petition, protesting (legally or illegally), joining a political party, engaging in local politics, or expressing opinions in newspapers (Cleaver et al., 2005; Kennedy, 2007; Schulz et al., 2010). The discrepancy between support for democratic ideals and the lack of action to support such values necessitates education to encourage more participatory and social-justice citizenship orientations. Thus, social studies education attempts to address this disparity.

Conflicts in Citizenship Education

Within the United States, Diane Hess (2009) differentiated between civic education, which relates to *fitting in* to society as it currently operates, and democratic education, which emphasizes the dynamic and competitive nature of democracy. Walter Parker (2008) advocated for education for *Enlightened Democratic Engagement*, which incorporates active political engagement with both democratic values, such as rejecting unjust laws, fighting civic inequality, and cooperating with others regarding public policy. However, coming up with an ideal goal of citizenship education is much easier than developing schools that produce an ideal outcome. Not surprisingly, differing views on how to promote democratic citizenship education exist.

In practice, Barr et al. (1977) identified three civic outcomes related to the social studies. First, education for social transmission emphasizes fixed core knowledge and skills for students to learn and believe. Thus, education serves to reproduce society, as it exists, leading to passive civic behaviors and attitudes. Secondly, social studies as social science seeks to give students skills necessary to understand human behavior using social scientific methods inherent to anthropology, political science, historiography, and/or geography. The goal is for students to gain the ability to evaluate the world removed from subjective views based on biases or prejudices. Finally, education for reflective inquiry is based on students' self-selections of problems and working toward resolution. Through reflective inquiry, students gain the skills to explore issues relevant to them and ultimately address real social problems.

Criticisms of the three forms of citizenship education above focus on two key aspects. First, the frameworks discussed fail to address issues of group membership and identity, such as structural inequality, sexism, and LGBTQ issues. Secondly, the emphasis on the nation-state in traditional citizenship education neglects increasingly important global perspectives. Marianna Papastephanou (2008), a scholar in Cyprus, strongly argued that traditional notions of citizenship education, similar to the descriptions of Barr and colleagues, fail to provide views of society distinct from what already exist. Such views serve to promote the perpetuation of current political and social realities. Papastephanou called for an enlarged view of citizenship that goes beyond narrow considerations and allows for conceptions that are radically different from current society.

In response to the shortcomings of popular conceptions of citizenship education, various critical discourses promoting social justice have become more prominent (Bickmore, 2008; Tyson and Park, 2008; Knight Abowitz and Harnish, 2006). T.J. King (2012) asserted that education for social justice requires addressing authentic community needs, fostering democratic citizenship focusing on addressing current unjust conditions, attending to marginalization, humanizing the other, starting from lived experience, and cultivating critical attitudes. Such perspectives in education can serve to create avenues for addressing structural social, economic, and political inequalities, thus promoting a democratic society.

In recent decades, globalization has led to increased intercultural communication, transnational identities, and international environmental political issues (Merryfield and Duty, 2008). As a result, citizenship scholars work to incorporate global perspectives while maintaining local heterogeneity (Linklater, 1997; Pike, 2008). To address this tension, leading scholars in international education, Patricia Kubow et al. (1998), advocate multidimensional citizenship education. Such an approach involves four dimensions which include a *personal* dimension focusing on socially responsible habits of mind, heart, and action; a *social* dimension focusing on building civil society by placing importance on a commitment to action based on respect for the rights of others; a *spatial* dimension addressing overlapping identities (local, national,

regional, and global), while acknowledging the interplay between diversity and regional uniqueness; and finally a *temporal* dimension focusing on issues of citizenship regarding variations across time.

In short, social studies researchers emphasize the need for curriculum to move beyond the transmission of traditional civic behaviors to youth. Newer models of civic education focus more on active civic engagement with consideration of the multicultural and global realties of their society.

International Perspectives

International research on education for citizenship education centers on the development of *civic competencies*, or the role of education in developing political knowledge, skills, and attitudes (Hahn, 2010, 1998; Ehman, 1980). Early research based on theories of political socialization traces the influence of family, school, and media on the political views of youth. However, recent research takes a more constructivist stance and views students as active constructors of meaning, opposed to simply passive recipients of messages (Torney-Purta et al., 1999, 2001; Wilkenfeld et al., 2010; Beaumont, 2010). While a complete overview of international and comparative education is not feasible for this work, several major trends exist.

First, international researchers strive to build an understanding of a democratic school (Davies, 1995; Davies and Kirkpatrick, 2000; Davies et al., 2002, 2005; Trafford, 2003; Harber and Mncube, 2012). Clive Harber and Vusi Mncube (2012) summarized the discussion by stating that a democratic school should make a clear commitment to values of education for democracy, such as regular, free, but polite exchange of views and opinions. The structures and practices within a school involve significant sharing of power and decision-making between staff, students, and parents. Moreover, a democratic school culture should be characterized by relationships built on trust and respect, therefore avoiding corporal punishment along with any form of physical punishment and bullying. Breaking down these barriers allows students to experience a democratic atmosphere to build their capabilities for future civic engagement.

Second, researchers investigate classrooms that support an *open classroom climate*, which is commonly associated with fostering civic competencies. Hess and Avery (2008) referred to such climates as having a 'democratic ethos,' by allowing open disagreements and diverse opinions within the classroom. Avery and Simmons (2008), studying students from the Czech Republic, Lithuania, and Azerbaijan, discovered that using a technique called 'structured academic controversy' led students to feel more interested, engaged, and encouraged to express their own views. In South Africa, Finkel and Ernst (2005) concluded that revised curriculum focusing student and issue-centered pedagogy had a strong and positive relationship with student civic/political knowledge. Moreover, an open classroom climate has demonstrated a positive relationship with positive civic attitudes and potential for political participation in a variety of countries including the United States (Torney-Purta et al., 2001), Chile and Portugal (Torney-Purta and Amadeo, 2004), and Israel (Perlinger et al., 2006).

Finally, an emerging trend of international citizenship education research focuses on the role of Western democratic values in non-Western countries. A leading scholar on Asian citizenship, Kerry Kennedy (2004) called for an exploration of the compatibility of local and Western values with the hope of developing indigenous notions of citizenship. Cross-national research studies support the influence of traditional values on civic orientations. For example, East Asian notions of citizenship graft themselves to older values such as moral behavior in public life and welfare of the community over individual rights (Kennedy and Lee, 2014; Kennedy and Fairbrother, 2004: p. 294). These traditional values, inherent to Confucian tradition, may potentially negate harmful effects of rampant individualism common in Western democracy (Kennedy and Lee, 2008). Studying South African and Kenyan students, Kubow (2007) explored South African and Kenyan teachers' beliefs about democracy and concluded that their beliefs were informed by both global/Western rights-based civic conceptions and local/indigenous values of compassion, communalism, and concern for the interests of the community. Within Israel, Orit Ichlov (2005) noted tremendous differences between expected political participation between students in Hebrew state schools, Hebrew religious schools, and Arab schools. As such, conceptions of citizenship vary based on cultural contexts.

Social studies as social education attends to how schools can give students capabilities to support a democratic government and society. The research in this area addresses the interaction of local contexts and the influx of ideas through globalization and other forms of pluralism as they relate to civic and democratic education.

Teaching Practices for Social Studies Education

Several approaches exist for teaching in social studies to students and youth. These include strategies that incorporate problem-based learning, discussion of controversial events or situations, historical investigation, authentic pedagogy, simulations, and multimedia rich instruction. Here, we outline three strategies that are popular in the social studies research literature: inquiry-based learning, the use of Socratic seminars, and academic structured controversy.

Inquiry-Based Learning

Sometimes referred to as the historical inquiry method, this approach to teaching social studies content recently became even more popular in the United States with the publication of the *College, Career, and Civic Life (C3) Framework for Social Studies* in 2013, which promoted inquiry-based learning as the heart of social studies instruction. The process includes four basic steps. First, students with the help of teachers construct inquiry-based questions about a topic that is under investigation. These questions ought to be thematic or broad in scope. Second, students apply domain-specific strategies (e.g., historical thinking, geographic thinking, etc.) to investigate a variety of primary and secondary resource materials. Third, students interpret the data to construct responses to their original inquiry question. Finally, students communicate their findings with others through writing, advocacy, or activism. This final step enables students to

make their new knowledge base relevant to their classroom and/or local community. In the inquiry-based learning approach, teachers act as facilitators and provide useful resources to guide students' investigations.

Socratic Seminars

Social studies researchers advocate seminar discussions to focus on improving understandings of an issue at hand (Hess, 2009; Parker, 2008, 2010). In order to set up a seminar discussion a teacher must first establish classroom norms that work to encourage respectful discussion between students based on the prompt, while also deemphasizing the teacher as the centerpiece. Walter Parker (2004) suggested starting with a powerful prompt such as a text, film, artwork, or political cartoon. Subsequently a teacher asks an opening question such as, "What does this mean?" Students then express their opinions and respond to the opinions of peers, drawing ideas from evidence found in the documents. The ensuing discussion should work to provide greater clarity to the complexities of the topic. Finally, a debriefing session should take place to give context to the discussion. Debriefing may include a reflective writing or answering questions as a large group such as "What did we accomplish today?" Debriefing questions help students bring together various ideas and can be useful to encourage positive discussion norms for the future. Building understanding through seminar discussions can inform future projects that aim at solving problems, such as a structured academic controversy discussed below.

Structured Academic Controversy

In contrast to seminar discussions described above, a structured academic controversy works to uncover solutions to a problem. While Walter Parker's (2004) description of a structured academic controversy is commonly cited, David Johnson and Rodger Jonson (1985, 1988) developed the practice. Initially, a teacher or students should select a powerful issue and do research to gain a solid understanding beforehand. Background knowledge could come from independent research or through seminar discussions described above. Diane Hess (2009) advocated incorporating a controversial issue to encourage democratic problem solving. After selecting and researching issues to discuss, teachers should clearly articulate the purposes of the deliberation and encourage classroom norms of respect. Such norms emphasize that students listen to all sides equally, build on the ideas of others, back up opinions with clear reasons, and take turns speaking. Students start in small groups or pairs and are asked to research a specific stance on an issue. Next, students switch positions on the topic and advocate for the opposite side. Finally, students as a whole sort out their own position relative to the issue and brainstorm possible solutions. As a goal, students evaluate alternative courses of action related to a shared problem and ultimately address the question, "What should we do?" Following the group discussion, a large group debriefing can take place in order to bring together the ideas built in small groups. During this time, teachers encourage critiques in order to explore the complexities of the issue. Finally, follow-up writing assignments help students reflect on the issue.

Social studies practices are varied and may depend greatly on the school contexts, background of the students, and the expertise of the teacher.

The Future of Social Studies Education

Social studies education represents a complex and multifaceted field of study. While research and scholarship in the field will continue to investigate the learning of and teaching for social studies content, current trends occurring across the globe will also suggest new and relevant topics for social studies research. These topics may include the following:

- Determining the role of technology on the learning and teaching of social studies.
- Investigating the impact of globalization on social studies.
- Defining global citizenship and identity.
- Exploring immigrant education and teaching and learning social studies.
- Discovering ways to meet diverse learning and linguistic needs of students in social studies classrooms.
- Drawing distinctions and comparisons across varied international school contexts.
- Documenting the response to social studies education from individuals coming from marginalized communities.
- Tracing the influence of standardization on social studies teaching and learning.

Overall, the social studies field offers scholars ample opportunities to address important areas for research. As a field, social studies education is relatively young compared to other disciplines; therefore, traditions of inquiry are continually being formed and reformed.

See also: Economic Education; Education, History of; Educational Institutions, History of; Environmental Education and Education for Sustainable Development; Leadership Education: Theory and Practice; Science Education; Teaching Economics.

Bibliography

Avery, P.G., Simmons, A.M., March 25, 2008. Results from a three-year evaluation of teachers and students participating in the deliberating in the deliberating in a democracy Project. In: Paper Presented at the Annual Meeting of the American Educational Research Association, New York.

Barr, R.D., Barth, J.L., Shermis, S.S., 1977. Defining the Social Studies. Bulletin No. 51 of the National Council for the Social Studies. National Council for the Social Studies, Arlington, VA.

Barton, K.C., 2001. A sociocultural perspective on children's understanding of historical change: comparative findings from Northern Ireland & the United States. American Educational Research Journal 38, 881–913.

Barton, K.C., 2008. Research on students' ideas about history. In: Levstik, L.S., Tyson, C.A. (Eds.), Handbook of Research in Social Studies Education. Routledge, New York, pp. 239–258.

Barton, K.C., Avery, P.G. Developing a democratic knowledge base for social studies education. In: Gitomer, D.H., Bell, C.A. (Eds.), Handbook of Research on Teaching, fifth ed. American Educational Research Association, Washington, DC, in press

Barton, K.C., Levstik, L.S., 2004. Teaching History for the Common Good. Lawrence Erlbaum Associates, Mahwah, NJ.

Beaumont, E., 2010. Political agency and empowerment: pathways for developing a sense of political efficacy in young adults. In: Sherrod, L.R., Torney-Purta, J., Flanagan, C.A. (Eds.), Handbook of Research on Civic Engagement in Youth. John Wiley & Sons, Hoboken, NJ, pp. 525–558.

Becker, W., Greene, W., Rosen, S., 1990. Research on high school economic education. Journal of Economic Education 21 (3), 231–245.

Bednarz, R.S., Bednarz, S.W., 2008. Geography education: the glass is half full and it's getting fuller. The Professional Geographer 56 (1), 22–27.

Berti, A., Bombi, A., 1988. The Child's Construction of Economics. Cambridge University Press, Cambridge.

Berti, A., 1995. Knowledge restructuring in an economic subdomain: banking. In: Schnotz, W., Vosnaidou, S., Carretero, M. (Eds.), New Perspectives on Conceptual Change. Pergamon, New York, pp. 113–133.

Bickmore, K., 2008. Social justice and the social studies. In: Levstik, L.S., Tyson, C.A. (Eds.), Handbook of Research in Social Studies Education. Routledge, New York, NY, pp. 155–171.

Braxton, J.M., Hargen, L.L., 1996. Variations among academic disciplines: analytical frameworks and research. In: Smart, J.C. (Ed.), Higher Education: Handbook of Theory and Research, vol. 11. Agathon Press, New York, pp. 1–46.

Brooks, S., 2009. Historical empathy in the social studies classroom: a review of the literature. Journal of Social Studies Research 33 (2), 213–234.

Bruner, J., 1960. Process of Education. Havard University Press, Cambridge, MA.

Catling, S., 2004. An understanding of geography: the perspectives of English primary trainee teachers. Geojournal 60, 149–159.

Cleaver, E., Ireland, E., Kerr, D., Lopes, J., 2005. Listening to Young People: Citizenship Education in England (Citizenship Education Longitudinal Study: Second Cross Sectional Survey 2004). National Foundation for Educational Research, Department for Education and Skills, London, England.

Dahl, D., 1998. What Johnny can't choose: and what Johnny (and Jane) needs to know to understand the economy. The Region 12 (4), 5–11.

Davies, L., Kirkpatrick, G., 2000. The EURIDEM Project; a Review of Pupil Democracy in Europe. Children's Rights Alliance, London.

Davies, L., 1995. International indicators of democratic schools. In: Harber, C. (Ed.), Developing Democratic Education. Education Now, Ticknall.

Davies, L., Harber, C., Schweisfurth, M., 2002. Democracy through Teacher Education. CIER/CfBT, Birmingham.

Davis, L., Harber, C., Schweisfurth, M., 2005. Democratic Professional Development. CIER/CfBT, Birmingham.

Davies, P., Durden, G., 2010. Economic education in schools and universities in England. The Journal of Economic Education 41 (4), 413–424.

Ehman, L., 1980. The American school in the political socialization process. Review of Educational Research 50 (1), 99–119.

Epstein, T., 2009. Interpreting National History: Race, Identity, and Pedagogy in Classrooms and Communities. Routledge, New York.

Evans, R.W., 2006. The social studies wars, now and then. Social Education 70 (5), 317–321.

Finkel, S.E., Ernst, H.R., 2005. Civic education in post-apartheid South Africa: alternative paths to the development of political knowledge and democratic values. Political Psychology 26 (3), 333–364.

Fournier, J.E., Wineburg, S.S., 1997. Picturing the past: gender differences in the depiction of historical figures. American Journal of Education 105 (2), 160–185.

Geographic Education Standards Project, 1994. Geography for Life: National Geography Standards. National Geographic Research & Exploration, Washington, DC.

Hahn, C.L., 1998. Becoming Political: Comparative Perspectives on Citizenship Education. State University of New York Press, Albany.

Hahn, C.L., 2010. Comparative civic education research: what we know and what we need to know. Citizenship, Teaching and Learning 6 (1), 5–23.

Hahn, J., Jang, K., 2010. Economic education in Korea: current status and changes. The Journal of Economic Education 41 (4), 436–447.

Harber, C., Mncube, V., 2012. Education, Democracy and Development: Does Education Contribute to Democratization in Developing Countries? Symposium Books, Oxford, UK.

Harris Interactive, 2005. What American Teens and Adults Know about Economics. http://ncee.net/cel/whatamericansknowabouteconomics_042605-3.pdf.

Hess, D., Avery, P.G., 2008. Discussion of controversial issues as a form and goal of democratic education. In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy, pp. 506–518.

Hess, D.E., 2009. Controversy in the Classroom: The Democratic Power of Discussion. Routledge, New York, NY.

Hickey, G., Bein, F., 1996. Students' learning difficulties in geography and teachers' interventions: teaching cases from K-12 classrooms. Journal of Geography 95, 118–125.

Hicks, D., van Hover, S., Doolittle, P.E., VanFossen, P., 2012. Learning social studies: an evidence-based approach. In: Harris, K.R., Urdan, T. (Eds.), APA Educational Psychology Handbook, vol. 3. American Psychological Association, Washington, D.C., pp. 283–307.

Ho, L.-C., 2010. Don't worry, i'm not going to report you: education for citizenship in Singapore. Theory and Research in Social Education 38 (2), 217–247.

Hume, S.E., Boehm, R.G., 2001. A rationale and model for a scope and sequence in geographic education, grades K-12. The Social Studies 92 (1), 16–21.

Ichilov, O., 2005. Pride in one's country and citizenship orientations in a divided society. Comparative Education Review 49 (1), 44–61.

Johnson, D.W., Johnson, R.T., 1985. Classroom conflict: controversy vs. debate in learning groups. American Educational Research Journal 22, 237–256.

Johnson, D.W., Johnson, R.T., 1988. Critical thinking through structured controversy. Educational Leadership 45 (8), 58–64.

Kennedy, K.,J., 2004. Searching for citizenship values in an uncertain global environment. In: Lee, W.O., Grossman, D.L., Kennedy, K.J., Fairbrother, G.P. (Eds.), Citizenship Education in Asia and the Pacific, fourteenth ed. Kluwer Academic Publishers, Hong Kong, China, pp. 9–24.

Kennedy, K.J., 2007. Student constructions of "active citizenship": what does participation mean to students? British Journal of Educational Studies 55, 304–324. http://dx.doi.org/10.1111/j.1467-8527.2007.00381.x.

Kennedy, K., Fairbrother, G., 2004. Asian perspectives on citizenship education in review: postcolonial constructions or pre-colonial values. In: Lee, W.O., Grossman, D.L., Kennedy, K.J., Fairbrother, G.P. (Eds.), Citizenship Education in Asia and the Pacific, fourteenth ed. Kluwer Academic Publishers, Hong Kong, China, pp. 289–302.

Kennedy, K.J., Lee, H., 2014. Civic education in Asia. In: Sander, Wolfgang (Ed.), Handbuch politische Bildung. Wochenschau Verlag, Schwalbach/Ts, pp. 597–608.

Kennedy, K.,J., Lee, J.C., 2008. The Changing Role of Schools in Asian Societies: Schools for the Knowledge Society. Routledge, Abingdon.

King, J.T., 2012. Teaching for social justice across the curriculum: connecting theory and practice. In: Cogan, J.J., Grossman, D. (Eds.), Creating Socially Responsible Citizens: Cases from the Asia-Pacific Region. Information Age Publishing, Charlotte, North Carolina, pp. 13–34.

Klein, P.H., 1997. Knowing and Teaching Geography: A Qualitative Study of Pedagogical Content Knowledge and Schemata in Expert and Novice Teachers. Unpublished doctoral dissertation. Georgia State University.

Knight Abowitz, K., Harnish, J., 2006. Contemporary discourses of citizenship. Review of Educational Research 26 (4), 653–690.

Kubow, P.K., 2007. Teachers' constructions of democracy: intersections of western and indigenous knowledge in South Africa and Kenya. Comparative Education Review 51 (3), 307–328.

Kubow, P., Grossman, D., Ninomiya, A., 1998. Multidimensional citizenship: educational policy for the 21st century. In: Cogan, J.J., Derricott, R. (Eds.), Citizenship for the 21st Century: An International Perspective on Education. Kogan Page, London, pp. 115–133.

Levesque, S., 2009. Thinking Historically: Educating Students for the Twenty-first Century. University of Toronto Press, Toronto.

Levstik, L.S., Groth, J., 2005. "Ruled by our own people": Ghanaian adolescents' conceptions of citizenship. Teachers College Record 107 (4), 563–586.

Linklater, A., 1997. Cosmopolitan citizenship. Citizenship Studies 2 (1), 23–41.

Lybarger, M.B., 1990. The historiography of social studies: retrospect, circumspect, and prospect. In: Shaver, J. (Ed.), Handbook of Research on Social Studies Teaching and Learning. Macmillan, New York, pp. 3–15.

Merryfield, M.M., Duty, L., 2008. Globalization. In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy. Sage Publications Inc, pp. 80–92.

Miller, S.L., VanFossen, P.J., 1994. Assessing expertise in economic problem solving: a model. Theory and Research in Social Education 22 (3), 380–412.

Nelson, J.L., 2001. Defining social studies. In: Stanley, W.B. (Ed.), Critical Issues in Social Studies Research for the 21st Century. Information Age Publishing, Greenwich, CT, pp. 15–38.

Niemi, R., Junn, J., 1998. Civic Education: What Makes Students Learn. Yale University Press, New Haven.

Papastephanou, M., 2008. Philosophical presuppositions of citizenship education and political liberalism. In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy. Sage Publications Inc, pp. 40–56.

Parker, W., 2008. Knowing and doing in democratic citizenship education. In: Levstik, L.S., Tyson, C.A. (Eds.), Handbook of Research in Social Studies Education, pp. 65–80.

Parker, W., 2010. Listening to strangers: classroom discussion in democratic education. Teachers College Record 112 (11), 2815–2832.

Parker, W.C., 2004. Teaching Democracy: Unity and Diversity in Public Life. Teachers College Press, New York.

Perlinger, A., Canetti-Nisim, D., Pedahzur, A., 2006. Democratic attitudes among high school pupils: the role played by perceptions of class climates. School Effectiveness and School Improvement 17 (1), 119–140.

Phipps, B., Clark, J., 1993. Attitudes toward economics: uni- or multidimensional? Journal of Economic Education 24 (3), 195–212.

Pike, G., 2000. Global education and national identity: in pursuit of meaning. Theory into Practice 36 (2), 64–73.

Pike, G., 2008. Global education. In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy. Sage Publications Inc, pp. 468–480.

Porat, D.A., 2004. It's not here, but this is what happened: students' cultural comprehension of textbook narratives on the Israeli–Arab conflict. American Educational Research Journal 41, 963–996.

Resor, C.W., 2010. Place based education: what is its place in the social studies classroom? The Social Studies 101, 185–188.

Saunders, P., Gilliard, J., 1995. Framework for Teaching the Basic Economic Concepts. National Council on Economic Education, New York.

Saxe, D.W., 1991. Social Studies in Schools: A History of the Early Years. State University of New York Press, Albany, NY.

Schoenfeldt, M., 2002. Geographic literacy and young learners. The Educational Forum 66 (1), 26–31.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., Losito, B., 2010. ICSS 2009 International Report: Civic Knowledge, Attitudes, and Engagement among Lower-Secondary School Students in 38 Countries. International Association for the Evaluation of Educational Achievement, Amsterdam, The Netherlands.

Segall, A., Helfenbein, R.J., 2008. Research on K-12 geography education. In: Levstik, L.S., Tyson, C.A. (Eds.), The Handbook of Research in Social Studies Education. Routledge, New York, pp. 259–283.

Son, B., VanSickle, R., 2000. Problem-solving instruction and students' acquisition, retention, and structuring of economics knowledge. Journal of Research and Development in Education 33 (2), 95–105.

Thomas, L., Willinsky, J., 1999. Grounds for imagining a pacific community: mapping across boundaries and great divides. Journal of Geography 98 (1), 1–13.

Thornton, S.J., 2008. Continuity and change in social studies curriculum. In: Levstik, L.S., Tyson, C.A. (Eds.), Handbook of Research in Social Studies Education. Routledge, New York, pp. 15–32.

Torney-Purta, J., Amadeo, J.A., 2004. Strengthening Democracy in the Americas through Civic Education: An Empirical Analysis Highlighting the View of Students and Teachers. Organization of American States, Washington, D.C.

Torney-Purta, J., Lehmann, R., Oswald, H., Schulz, W., 2001. Citizenship and Education in Twenty-eight Countries: Civic Knowledge and Engagement at Age Fourteen. International Association for the Evaluation of Educational Achievement, Amsterdam, The Netherlands.

Torney-Purta, J., Schwille, J., Amadeo, J.A. (Eds.), 1999. Civic Education across Countries: Twenty-four National Case Studies Form the IEA Civic Education Project. The International Association for the Evaluation of Education Achievement, Amsterdam.

Trafford, B., 2003. School Councils, School Democracy and School Improvement. SHA, Leichester.

Trygestad, J., 1997. Students' conceptual thinking in geography. In: Paper Presented at the Annual Meeting of the American Educational Research Association, Chicago.

Tyson, C., Park, S.C., 2008. Civic education, social justice and critical race theory. In: Arthur, J., Davies, I., Hahn, C. (Eds.), The Sage Handbook of Education for Citizenship and Democracy, pp. 29–39.

VanFossen, P., 1995. Relevant indicators of expertise in economic problem-solving: a factor analysis. In: Paper Presented at the American Educational Research Association Annual Meeting. CA. Resources information Center, San Francisco. ERIC Document number: 388572.

VanFossen, P.J., Miller, S.L., 2008. Recent research on the teaching and learning of pre-collegiate economics. In: Levstik, L.S., Tyson, C.A. (Eds.), The Handbook of Research in Social Studies Education, pp. 284–306.

VanSledright, B., 2002. In Search of America's Past: Learning to Read History in Elementary School. Teachers College Press, New York.

VanSledright, B., 2011. The Challenge of Rethinking History Education. Routledge, New York.

VanSledright, B., Limon, M., 2006. Learning and teaching social studies: a review of cognitive research in history and geography. In: Alexander, P.A., Winne, P.H. (Eds.), Handbook of Educational Psychology, second ed. Lawrence Erlbaum, Mahwah, NJ, pp. 545–570.

Walford, R., 1996. "What is geography?" an analysis of definitions provided by prospective teachers of the subject. International Research in Geographical and Environmental Education 5, 69–76.

Walstad, W., 2001. Economic education in U. S. high schools. Journal of Economic Education 15 (3), 195–210.

Walstad, W., Rebeck, K., 2001. Assessing the economic understanding of U. S. high school students. The American Economic Review 91 (2), 452–457.

Walstad, W., VanScyoc, L., 1990. The effects of textbooks on economics understanding and attitudes in high school economics courses. Journal of Research and Development in Education 24 (1), 44–52.

Warf, B., 1999. Constructing a dialogue: geographic education and geographic research. Professional Geographer 51 (4), 581–591.

Westheimer, J., Kahne, J., 2004. What kind of citizen? The politics of educating for democracy. American Educational Research Journal 41 (2), 237–269.

Wilkenfeld, B., Lauckhardt, J., Torney-Purta, J., 2010. The relation between developmental theory and measures of civic engagement in research on adolescents. In: Sherrod, L.R., Torney-Purta, J., Flanagan, C.A. (Eds.), Handbook of Research on Civic Engagement in Youth. John Wiley & Sons, Hoboken, NJ, pp. 193–219.

Wineburg, S.S., 2001. Historical Thinking and Other Unnatural Acts: Charting the Future of Teaching the Past. Temple University Press, Philadelphia.

Wineburg, S., Wilson, S.M., 1991. Subject matter knowledge in the teaching of history. In: Brophy, J. (Ed.), Advances in Research on Teaching. JAI, Greenwich, CT, pp. 305–347.

Social Support and Recovery from Disease and Medical Procedures

Melika H Stephens and Keith J Petrie, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social support has been shown to moderate the negative effects of chronic illness and speed recovery from medical procedures. The effect of social support, however, is dependent to a large degree on the fit between the type of support offered and the needs of the patient. A mismatch in the type or intensity of support can lead to difficulties and hamper recovery. Patients low in social support are more likely to present with advanced illness or die while in hospital. While the evidence linking social support and health outcomes is strong, the mechanisms that are operating are still largely unknown.

Social support is usually defined in terms of the individual feeling valued and cared for by their social network as well as how well the person is embedded into a network of communication and social obligation. A large range of studies has found social support to have an important relationship in recovery from illness and medical procedures. Researchers generally conceptualize and measure social support in three main ways. First, in terms of the individuals' perceptions that they have people in their social network that they can turn to in times of need. Second, in terms of structural aspects of the individual's social network which include the number of personal ties and network density. This includes the number of people the individual has a connection with, as well as their membership of formal and informal social groups or organizations. Third, social support is defined in terms of functional support or the aid and encouragement provided to the individual by the social network. This can be further categorized into the type of functional support provided, in terms of being instrumental, financial, informational, appraisal, or emotional. Research shows that these different types of social support have differential effects on recovery from illness (Cohen, 2004; House et al., 1988).

Social Support and Recovery from Disease

Evidence examining the effect of social support on recovery from illness comes from a variety of medical illnesses including cardiovascular diseases, cancer, infectious disease, and other illnesses. Under most circumstances, social support has been shown to moderate the negative effects of illness and improve recovery, while low social support has been associated with a poorer prognosis in many clinical conditions. The effect of social support, however, is dependent to a large degree on the illness and the match between the type of support offered and the needs of the patient.

Cardiovascular Disease

In the cardiovascular area, there is strong evidence that poor social support and social conflict are significantly associated with increased risk of cardiovascular-related morbidity and mortality (Everson-Rose and Lewis, 2005). Reviews of epidemiological literature show low social support to be a risk factor for the development of cardiovascular disease in healthy individuals (Eng et al., 2002; Lett et al., 2005). In patients with established cardiovascular disease, low social support is also associated with a poor prognosis. A study of elderly patients with myocardial infarction (MI) found those with low levels of emotional social support were more likely to die in hospital within 6 months of their heart attack, even after controlling for severity of MI and other risk factors (Berkman et al., 1992). A recent systematic review and meta-analysis found that in patients with Coronary Heart Disease (CHD), low functional support is consistently associated with cardiac mortality, even adjusting for other medical and demographic variables (Barth et al., 2010).

Williams et al. (1992) followed up 1965 patients with coronary artery disease after angioplasty. After accounting for known medical prognostic factors such as cigarette smoking and family history of heart disease, the most important socio-economic variable in predicting time to death was the presence or absence of a spouse or confidant. This finding is consistent with other studies of heart attack patients that have found a spouse to play a protective role in outcome. Patients with CHD who have small social networks have been found to have an increased mortality risk, and this elevated risk was not attributable to confounding by disease severity, demographics, or psychological distress (Brummett et al., 2001). Consistent with the work in CHD mortality and worsening CHD progression, a study on the progression of coronary atherosclerosis found social isolation to be an important risk factor for accelerated progression of the disease (Wang et al., 2005). There is also some evidence that social support is more protective for cardiovascular health among men compared to women (Orth-Gomér, 2009).

An interesting avenue of new research is looking beyond the single assessment of social support and health outcomes following an acute MI and to assess whether changing levels of social support during early recovery influence later health status. In a study using this approach, Leifheit-Limson et al. (2012) found patients with worsening levels of social support 1 month following their heart attack had more angina, greater levels of depression, and lower disease-related quality of life at 1 year, compared to those patients who had continued high social support. This research demonstrates that changing levels of social support may have important effects on patient

recovery from illness and there may be benefit from assessing social support at more than one time point.

Cancer

Another important area investigating the effect on social support on recovery from chronic illness is the effect of social factors in the progression and prognosis of cancer. This research demonstrates how different social support dimensions can differentially affect illness recovery, and how the relationship between social support and cancer progression can vary with different types of illness. A systematic review by Nausheen et al. (2009) found the evidence for the relationship between social support and cancer progression is strong for breast cancer but consistently unconvincing for other types of cancer. This review also found that structural support was found to be more frequently associated with disease progression for breast cancer than functional support.

In terms of cancer survival, a recent meta-analysis (Pinquart and Duberstein, 2010) showed that in controlled studies, having high levels of perceived social support, larger social networks, and being married were associated with decreases in relative mortality risk of 25, 20, and 12%, respectively. However, the relationship varied by cancer site, with stronger associations of social support in patients with leukemia and lymphomas and stronger associations of network size observed in breast cancer patients. Moderator analyses suggested that having never been married is associated with higher rates of mortality from cancer than those who have been divorced or separated. This research is also consistent with work by Ell et al. (1992) which showed that the relationship between social support and survival may operate differently depending on cancer site and extent of the disease.

Larger social networks have been associated with lower breast cancer mortality. A prospective study of 2835 women with breast cancer found that those individuals who were socially isolated before diagnosis were twice as likely to die of their breast cancer compared to socially integrated women (Kroenke et al., 2006). Pinquart and Duberstein (2010) noted that in their large meta-analysis of social networks and mortality among cancer patients, associations between social network and mortality were stronger in young patients.

Researchers are beginning to appreciate that the impact of large social network on breast cancer survival may depend on the quality and burden of these networks. A recent study found that larger social networks predicted better prognosis following diagnosis, however, this association depended on the level of social support and burden within relationships (Kroenke et al., 2013). Women with both small social networks and low levels of support had significantly higher risk of mortality than women with large networks and high levels of support. However, women with small networks and high levels of support did not have a significantly higher mortality risk compared to women with large social networks and high support. These results indicate that it is the quality of the social support, and not just the size of an individual's social network, that is important for survival following breast cancer.

Social attachment, a type of social support reflecting connections with others, has also been associated with a survival advantage for patients with ovarian cancer (Lutgendorf et al., 2012). Patients in this study were followed up from the date of their surgery until their death. The median survival time for patients with low social attachment was 3.35 years, compared to 4.70 years for those with high social attachment. Another important finding from this prospective study was the nonsignificant relationship between instrumental social support and survival. The relationship between social support and cancer mortality has also been found in Japan. Social isolation and the perception of low support from family members were associated with higher risk of colorectal cancer mortality in Japanese men (Ikeda et al., 2013).

Infectious and Other Diseases

A well-controlled study looked at the susceptibility to developing the common cold in volunteer subjects given nasal drops containing one or two cold viruses (Cohen et al., 1997). The results showed that those volunteers with more diverse social networks had a fourfold lower susceptibility to developing upper respiratory illness than those with a smaller number of social ties. In this study, social contact per se was not as important as the diversity of the individual's social relationships. The more an individual had regular contact with a variety of social groups such as family, workmates, neighbors, and other social networks, the more resistant he or she was to developing a cold.

Another study examining the role of psychosocial factors on progression to AIDS found stress and social support to be significant predictors of HIV disease advancement (Leserman et al., 1999). In a follow-up period of 5.5 years, those participants who were more satisfied with the level of social support offered by their friends and families at study entry were twice as likely to be free from AIDS than participants who were dissatisfied with their available level of social support.

As well as infectious diseases, social support has been shown to have a positive effect on prognosis and recovery in a range of other medical conditions and illnesses. A study of low-income pregnant women found social support was associated with fewer complications during pregnancy and a higher infant birth weight. Research has found women who receive higher-quality social support prior to delivery give birth to healthier babies and experience less postpartum depression. In chronic illnesses, Penninx et al. (1998) found positive effects on depression for having a partner and having many close relationships, although the strength of the effect differed across illnesses. Other work has shown social support to be associated with lower levels of disability and depression in patients with diabetes and higher levels of compliance in patients on renal dialysis.

Social Support Interventions

A large number of interventions have been trialed in health care to improve the level of social support to patients. These interventions vary in terms of whether a group or an individual format is used and whether the focus is on providing additional social support to patients or bolstering the effectiveness of their current networks. Interventions also vary according to the source of the social support in terms of whether support comes from a peer with the same or similar illness, a spouse or family

member, or a health professional (Hogan et al., 2002). The range of trials of interventions makes it hard to draw firm conclusions about the effectiveness of social support interventions. The most common reported effect is a reduction of psychological distress and an improvement in functional activity. However, the evidence is not clear about what type of social support intervention works best in a particular patient group.

A popular source of social support for patients with chronic illness is informal support groups. Here, patients meet together to discuss their condition and the difficulties and successes they have experienced in coping with their illness. As well as sharing personal illness experiences, these groups offer the advantage of providing information about utilizing health services, normalizing experiences and concerns, and providing emotional support to group members. An analysis of social support groups found that the experience of physical illness was the most common reason for participation in such groups (Lieberman and Snowden, 1993). Even after excluding substance abuse groups, this American survey found coping with physical illness made up 42% of the self-help group population. This study also suggested that these groups were more effectively accessed by the middle-class populations who often already have good access to other support services.

The limit of such social support groups for patients with physical illness was that participants had to be well enough to be physically able to attend meetings. However, the Internet has changed this requirement. In the last 5 years, there has been a dramatic increase in the use of the Internet in most Western households. This has been accompanied by an increasing number of discussion and support groups for chronic illness. Now, without leaving home, chronic illness sufferers can read postings from other sufferers and participate in discussions about aspects of their condition.

New types of online support web sites have also emerged in recent years. PatientsLikeMe.com is a web site where patients connect with others who have the same illness. The web site is set up for patients to track progress of their illness and share data on symptoms, treatment, side effects, and functioning. The web site now covers over a thousand health conditions and also acts as a forum for patients to share their experiences of their illness with others. PatientsLikeMe also aggregates data from other patients with the same illness so patients can compare how they are doing in relation to others. Aggregated data from PatientsLikeMe.com have also been used for peer-reviewed scientific papers (Wicks et al., 2010).

While patients are using web-based communities to connect with patients with similar illnesses, sometimes the recipient does not always see actions by others designed to assist with medical problems as helpful. Negative reactions to intended helpful actions of others are quite common. Patients often report receiving support that they did not want or not getting assistance for a daily activity when it was needed. This mismatch between the need and provision of support has been the focus of study in the general social support area and is very relevant to the area of recovery from medical illness (Thoits, 2011). Here, the evidence suggests that different types of support are differentially valued depending on the relationship the provider has with the patient.

In an important study with cancer patients, Dakof and Taylor (1990) found emotional support was most likely to be valued when it came from intimate others, whereas informational support was more valued when it came from a physician or another cancer patient. This study found when there was a mismatch, such as when a patient needed emotional support from a spouse but received advice or information, then these actions were more likely to be seen as unhelpful. Unhelpful behaviors noted by patients were minimizing the problem, forced cheerfulness, medical care delivered without any emotional support, and being told not to worry. A later study using patients with the chronic nonlife-threatening illnesses of irritable bowel syndrome and chronic headache found these patients to put greater value on tangible and practical assistance and less on emotional support than cancer patients (Martin et al., 1994).

These studies highlight how the characteristics of the illness may strongly influence the type of social support needed by patients. In some life-threatening conditions, where patients are frightened or daunted by the demand of treatment and prognosis, emotional needs are more prominent. In other illnesses, where the illness forms a major hurdle to be met on a more routine daily basis without the threat of disfigurement or death, then more practical issues such as assistance with childcare and transport become more valued.

A disparity in the intensity of social support can also cause difficulties for patients. Sometimes an aversion or fear of physical illness can cause family to avoid contact with the patient. While emotional support is the type of support most received by patients, it is also the type of support likely to be perceived as most inadequate. However, overly intrusive offers of assistance and help can also cause a negative reaction and increased stress. These findings highlight the importance of finding a balance between the needs of the patient and the type and level of social support.

Social Support and Recovery from Medical Procedures

Although not as extensively studied as recovery from illness, there is considerable evidence that social support also has a positive influence on recovery from hospitalization and medical procedures. Hospitalization is an anxious and difficult experience for most patients. Research shows social support helps to moderate the negative effect of hospitalization and to speed recovery from surgery.

Often surgical procedures involve postoperative pain and functional impairment, as well as psychological distress. Krohne and Slangen (2005) investigated the influence of patients' social support on subjective and objective stress indicators before, during, and after surgery. Patients who scored high on social support showed less anxiety, received lower doses of narcotics, and had a shorter hospital stay than did patients with low support. There is also evidence that social support influences mortality following surgery. In a prospective study of 1290 patients undergoing coronary bypass surgery, Herlitz et al. (1998) found that patients who agreed with the statement 'I feel lonely' prior to surgery had more than double

the risk of dying after the operation, compared with patients who did not feel lonely.

There is good evidence that social support also influences an individual's length of stay in hospital. Contrada et al. (2008) investigated social support in length of stay following heart surgery. Psychosocial interviews were conducted 5 days prior to surgery. Low social support demonstrated an independent, prospective association with greater length of stay, supporting the role of social support in influencing adaption to heart surgery. Visits from family or friends during hospitalization have also been associated with reduced hospital stay following coronary artery bypass surgery (Kulik and Mahler, 1989). Married male veterans who received high support, defined as more spousal visits, recovered more quickly and took less pain medication compared to those who received low support. Married patients who had high social support were released from hospital an average 1.26 days earlier compared to those patients who received low support.

A recent study examined social connectedness in veterans recovering from major thoracic or abdominal operations (Mitchinson et al., 2008). Social networks were assessed by the number of friends and relatives an individual had and how frequently contact was made with members of their social network. Patients reporting larger social network had lower pain and opiate use and anxiety than patients with small or nonexistent social networks. Smaller social network size was associated with the likelihood of staying in hospital for 7 days or more following these major operations.

One fascinating aspect in the study of social support and recovery from medical procedures is the effect of the patient's hospital roommate on recovery. Research in this area demonstrates that whether a patient has a pre- or postoperative roommate can influence outcome. Patients about to undergo surgery who are also assigned to a preoperative roommate show greater levels of anxiety and a slower recovery than those assigned a roommate who has already been through surgery. As well as lower levels of anxiety prior to surgery, being placed with a roommate who has been through surgery resulted in getting out of bed more quickly following the operation and a fast discharge from hospital (Kulik et al., 1993). It seems likely then that the positive benefits resulting from a postoperative roommate are partly due to the natural relief of seeing someone who has survived the same operation that the preoperative patient is about to undergo, and by normalizing the feeling and problems following the operation so these are less likely to be viewed as unusual or abnormal by the patient. Thus a roommate who has been through the operation or procedure can serve as a practical role model for preparing patients for the postoperative pain, physical sensations, and sequence of recovery following the operation.

This research was extended in a quasiexperimental study among patients undergoing cosmetic surgery (Kulik et al., 2000). Patients were assigned to rooms that either facilitated sufficient postoperative contact with other patients, or they were assigned to rooms that were physically located so that they had little interpatient contact. The results indicated that patients in the high patient contact condition experienced less postoperative anxiety and reported higher overall satisfaction with their quality of care compared to patients who received low contact. Analyses of patient-reported postoperative patient interactions suggest several additional benefits from this contact with fellow patients, including additional emotional support, a reduction in uncertainty about what to expect, and the chance to compare progress and emotional reactions. These results are consistent with the idea that another patient can serve a beneficial adjunctive role, positively influencing a fellow patient's postoperative emotional well-being and satisfaction with care (Kulik et al., 2000).

An individual's marital status has been shown to be related to patient outcomes following medical procedures. In a large study of 40 820 patients admitted to a University Hospital in Ohio, Gordon and Rosenthal (1995) found that unmarried patients presented to hospital later than married patients. Unmarried surgical patients also had a higher risk of dying in hospital than married surgical patients, even after controlling for severity of illness, diagnosis, and other demographic factors. Furthermore, unmarried patients tended to stay longer than married patients. The differences found in this study were greater in patients who had never married when compared with patients who were widowed, divorced, or separated. This study suggests that being unmarried and in particular, never having married, increases your risk of negative outcomes during hospitalization. Hospitalization can also have a negative effect on the health of those in a close relationship with the patient. In a study examining mortality after the hospitalization of a spouse, Christakis and Allison (2006) examined over 500 000 couples who were enrolled in Medicare in 1993 and looked at hospitalizations and death over a 9-year follow-up period. The researchers found that hospitalization of a spouse was associated with an increased risk of death among elderly participants. This risk varied between different diagnoses but was generally higher for men than women.

Recent research from the area of psychoneuroimmunology has shown that conflict in a close relationship can negatively impact wound healing. Kiecolt-Glaser et al. (2005) investigated how hostile martial behaviors modulate blister wound healing. Couples were assessed twice in a crossover analysis. During the first admission, couples engaged in a structured social support interaction. During the second admission, couples discussed a marital disagreement. Blister wounds healed more slowly following couples' conflict discussions than after more supportive interchanges, and couples who were more hostile toward each other during both discussions had wounds that healed more slowly than couples whose interactions were less hostile. The overall differences related to hostility were substantial; small blister wounds in high-hostile couples healed at only 60% of the rate of low-hostile couples. Thus, wound healing appeared to be responsive to both the short-term stress of a conflict, as well as hostile behavior. Thus, while marriage may have beneficial effects of recovery, inadequate support or conflict could cancel out this effect.

Summary

The available evidence suggests social support has a positive influence on individuals' susceptibility to infectious disease and recovery and adjustment to a variety of chronic illnesses. In the cardiac area, there seem to be particularly strong benefits for

having a spouse or close confidant and these effects seem to be even stronger for men. With the rise of the Internet, electronic social support groups are providing an important source of information and peer support for patients with chronic illness. There is strong evidence that the lack of social support is a risk factor for negative outcomes following hospitalization. Patients low in social support are more likely to present with more advanced illness and to be more likely to die while in hospital. However, it should be noted that many of the studies in the area are correlational, and the mechanisms that link social support to recovery from illness are still largely unknown. Recovery from surgery is an example where action by others that conveys to patients that they are respected and valued has a positive effect on outcome. The type of roommate can also set up a more positive outcome from the operation by providing a constructive role model by minimizing catastrophizing and demonstrating effective coping strategies. A critical factor running through social support and the recovery from illness and medical procedures is the match between the needs of the patient and the type of social support provided. Different types of illness and procedures cause differing needs in terms of whether they require more emotional or tangible support. Emotional support is valued highest when it comes from a significant other. A mismatch in the type or intensity of support can lead to difficulties and hamper overall recovery.

See also: Cancer: Psychosocial Aspects; Cardiac Disease, Coping with; Childhood Cancer: Psychosocial Aspects; Coronary Heart Disease: Psychosocial Aspects; Health-Related Support and Self-help Mutual Aid Groups; Infectious Diseases: Psychosocial Aspects; Social Integration, Social Networks, and Health; Stress, Coping and Health.

Bibliography

Arora, N.K., Rutten, F., Gustafson, D.H., Moser, R., Hawkins, R.P., 2007. Perceived helpfulness and impact of social support provided by family, friends, and health care providers to women newly diagnosed with breast cancer. Psycho-Oncology 16, 474–486.

Barrera Jr., Glasgow, R.E., Garth McKay, H., Boles, S.M., Feil, E.G., 2002. Do internet-based support interventions change perceptions of social support?: an experimental trial approaches for supporting diabetes self-management. American Journal of Community Psychology 30, 637–654.

Barth, J., Schneider, S., Kanel, R.V., 2010. Lack of social support in the etiology and the prognosis of coronary heart disease: a systematic review and meta-analysis. Psychosomatic Medicine 72, 229–238.

Berkman, L.F., Leo-Summers, L., Horwitz, R.I., 1992. Emotional support and survival after myocardial infarction: a prospective, population-based study of the elderly. Annals of Internal Medicine 117, 1003–1009.

Brummett, B.H., Barefoot, J.C., Siegler, I.C., Clapp-Channing, N.E., Lytle, B.L., Bosworth, H.B., Mark, D.B., 2001. Characteristics of socially isolated patients with coronary artery disease who are at elevated risk for mortality. Psychosomatic Medicine 63, 267–272.

Christakis, N.A., Allison, P.D., 2006. Mortality after the hospitalization of a spouse. New England Journal of Medicine 354, 719–730.

Chung, J.E., 2013. Social networking in online support groups for health: how online social networking benefits patients. Journal of Health Communication: International Perspectives 19, 639–659.

Cohen, S., 2004. Social relationships and health. American Psychologist 59, 676–684.

Cohen, S., Doyle, W.J., Skoner, D.P., Rabin, B.S., Gwaltney, J.M., 1997. Social ties and susceptibility to the common cold. Journal of the American Medical Association 227, 1940–1944.

Contrada, R.J., Boulifard, D.A., Hekler, E.B., Idler, E., Spruill, T.M., Labouvie, E.W., Krause, T.J., 2008. Psychosocial factors in heart surgery: presurgical vulnerability and postsurgical recovery. Health Psychology 27, 309–319.

Dakof, G.A., Taylor, S.E., 1990. Victims' perceptions of social support: what is helpful from whom? Journal of Personality and Social Psychology 58, 80–89.

DiMatteo, M.R., 2004. Social support and patient adherence to medical treatment: a meta-analysis. Health Psychology 23, 207–218.

Ell, K., Nishimoto, R., Mediansky, L., Mantell, J., Hamovitch, M., 1992. Social relations, social support and survival among patients with cancer. Journal of Psychosomatic Research 36, 531–541.

Everson-Rose, S., Lewis, T., 2005. Psychosocial factors and cardiovascular diseases. Annual Review of Public Health 26, 469–500.

Eng, P.M., Rimm, E.B., Fitzmaurice, G., Kawachi, I., 2002. Social ties and change in social ties in relation to subsequent total and cause-Specific ortality and coronary heart disease incidence in men. American Journal of Epidemiology 155, 700–709.

Gordon, H.S., Rosenthal, G.E., 1995. Impact of marital status on outcomes in hospitalized patients: evidence from an academic medical center. Archives of Internal Medicine 155, 2465–2471.

Herlitz, J., Wiklund, K., Caidahl, M., Hartford, M., Haglid, M., Karlsson, B.W., Sjoland, H., Karlsson, T., 1998. The feeling of loneliness prior to coronary artery bypass grafting might be a predictor of short- and long-term postoperative mortality. European Journal of Vascular and Endovascular Surgery 16, 120–125.

Hogan, B.E., Linden, W., Najarian, B., 2002. Social support interventions: do they work? Clinical Psychology Review 22, 381–440.

House, J.S., Landis, K.R., Umberson, D., 1988. Social relationships and health. Science 241, 540–545.

Ikeda, A., Kawachi, I., Iso, H., Iwasaki, M., Inoue, M., Tsugane, S., 2013. Social support and cancer incidence and mortality: the JPHC study cohort II. Cancer Causes and Control 24, 847–860.

Kiecolt-Glaser, J.K., Loving, T.J., Stowell, J.R., Malarkey, W.B., Lemeshow, S., Dickinson, S.L., Glaser, R., 2005. Hostile marital interactions, proinflammatory cytokine production, and wound healing. Archives of General Psychiatry 62, 1377–1384.

Krohne, H.W., Slangen, K.E., 2005. Influence of social support on adaptation to surgery. Health Psychology 24, 101–105.

Kroenke, C., Kubzansky, L., Schernhammer, E., Holmes, M., Kawachi, I., 2006. Social networks, social support, and survival after breast cancer diagnosis. Journal of Clinical Oncology 24, 1105–1110.

Kroenke, C.H., Quesenberry, C., Kwan, M.L., Sweeny, C., Castillo, A., Caan, B.B., 2013. Social networks, social support, and burden in relationships, and mortality after breast cancer diagnosis in the life after breast cancer epidemiology (LACE) study. Breast Cancer Research and Treatment 137, 261–271.

Kulik, J.A., Mahler, H.I., 1989. Social support and recovery from surgery. Health Psychology 8, 221–238.

Kulik, J.A., Moore, P.J., Mahler, H.I.M., 1993. Stress and affiliation: hospital roommate effects on preoperative anxiety and social interaction. Health Psychology 12, 118–124.

Kulik, J., Shelby, D., Cooper, R., 2000. The effects of fellow patients on the emotional well-being and satisfaction with care of postoperative cosmetic surgery patients. Plastic and Reconstructive Surgery 106, 1407–1414.

Leifheit-Limson, E.C., Reid, K.J., Kasl, S.V., Lin, H., Buchanan, D.M., Jones, P.G., Peterson, P.N., Parashar, S., Spertus, J.A., Lichtman, J., 2012. Changes in social support within the early recovery period and outcomes after acute myocardial infarction. Journal of Psychosomatic Research 73, 35–41.

Leserman, J., Jackson, E.D., Petitto, J.M., Golden, R.N., Silva, S.G., Perkins, D.O., Cai, J., Folds, J.D., Evans, D.L., 1999. Progression to AIDS: the effects of stress, depressive symptoms and social support. Psychosomatic Medicine 61, 397–406.

Lett, H.S., Blumenthal, J.A., Babyak, M.A., Strauman, T.J., Robins, C., Sherwood, A., 2005. Social support and coronary disease: epidemiologic evidence and implications for treatment. Psychosomatic Medicine 67, 869–878.

Lieberman, M.A., Snowden, L., 1993. Problems in assessing prevalence and membership characteristics of self-help group participants. Journal of Applied Behaviour Science 29, 166–180.

Lutgendorf, S.K., De Geest, K., Bender, D., Ahmed, A., Goodheart, M.J., Dahmoush, Sood, A.K., 2012. Social influences on clinical outcomes of patients with ovarian cancer. Journal of Clinical Oncology 30, 2885–2890.

Martin, R., Davis, G.M., Baron, R.S., Suls, J., Blanchard, E.B., 1994. Specificity in social support: perceptions of helpful and unhelpful provider behaviors among irritable bowel, headache and cancer patients. Health Psychology 13, 432–439.

Mitchinson, A.R., Kim, H.M., Geisser, M., Rosenberg, J.M., Hinshaw, D.B., 2008. Social connectedness and patient recovery after major operations. Journal of the American College of Surgeons 206, 292–300.

Nausheen, B., Gidron, Y., Peveler, R., Moss-Morris, R., 2009. Social support and cancer progression: a systematic review. Journal of Psychosomatic Research 67, 403–415.

Orth-Gomér, K., 2009. Are social relations less health protective in women than in men? Social relations, gender, and cardiovascular health. Journal of Social and Personal Relationships 26, 63–71.

Penninx, B.W.J.H., van Tilburg, T., Boeke, A.J.P., Deeg, D.J.H., Kriegsman, D.M.W., van Eijk, J.T.M., 1998. Effects of social support and personal coping resources on depressive symptoms: different for different diseases? Health Psychology 17, 551–558.

Pinquart, M., Duberstein, P.R., 2010. Associations of social networks with cancer mortality: a meta-analysis. Critical Reviews in Oncology Hematology 75, 122–137.

Reynolds, J.S., Perrin, N., 2004. Mismatches in social support and psychosocial adjustment to breast cancer. Health Psychology 23, 425–430.

Thoits, P.A., 2011. Mechanisms linking social ties and support to physical and mental health. Journal of Health and Social Behavior (52), 145–161.

Uchino, B., 2006. Social support and health: a review of physiological processes potentially underlying links to disease outcomes. Journal of Behavioral Medicine 29, 377–387.

Wang, H.X., Mittleman, M.A., Orth-Gomer, K., 2005. Influence of social support on progression of coronary artery disease in women. Social Science & Medicine 60, 599–607.

Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., Heywood, J., 2010. Sharing health data for better outcomes on PatientsLikeMe. Journal of Medical Internet Research 12 (2), e19.

Williams, R.B., Barefoot, J.C., Califf, R.M., Haney, T.L., Saunders, W.B., Pryor, D.B., Hlatky, M.A., Siegler, I.C., Mark, D.B., 1992. Prognostic importance of social and economic resources among medically treated patients with angiographically documented coronary artery disease. Journal of American Medical Association 267, 520–524.

Social System

Milan Zafirovski, University of North Texas, Denton, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Fillieule, volume 23, pp. 15418–15423, © 2001, Elsevier Ltd.

Abstract

This article examines the treatment of the social system in the sociological and related literature. The social system is the main concept of sociological systems theory, which conceptualizes societies as well as groups within them as systems, spanning from classical to contemporary sociology and related to general systems theory in the physical and related sciences. The article addresses such issues as the defining properties of the social system, types of social systems, classifications of historical–empirical social systems, and structural elements of the social system.

The concept of 'social system' is one of the key concepts of sociology, especially its theoretical strand (Barber, 1993; Fararo, 1989; Luhmann, 1995; Parsons, 1951; Sorokin, 1957), and is used to describe and analyze society as well as various social groups within it. (Fararo (1989: p. 64) remarks that "thinking in terms of systems is an important element of theoretical sociology.") In this approach, society is conceptualized and analyzed as a general or total 'social system' and social groups as particular 'social systems' or subsystems that compose it. In particular, 'social system' forms the principal concept of sociological systems theory originating in early sociology (Collins, 1988; Fararo, 2001; Habermas, 2001), as well as applying and extending (Luhmann, 1995) general systems theory from natural science (Bailey, 1994; Berrien, 1968; Bertalanffy, 1975; Buckley, 1967) into the social and behavioral sciences.

Defining Properties of the Social System

The concept of social system is in the sociological and related literature defined by a number of essential properties that are typically interrelated and mutually reinforcing. First, the social system is defined by constitutive, structural properties (Giddens, 1984), thus as a set of constituents, i.e., parts, units, components, or structures. By virtue of these properties, the social system represents and operates as a structured and differentiated complex (Fararo, 1989; Luhmann, 1995) of elements of a social character. Among early sociologists, Comte (1983[1830–52]) states that every social or other system is composed of elements of the 'same nature with itself,' Pareto (1963[1916]) treats society and group as a 'social or an economic system' and its elements as constituting such a system, Simmel (1964[1908]) identifies system 'structure,' Weber (1968[1920]) refers to society as the 'general structures of human groups,' and Parsons (1951) regards the social system as having a 'structure.' Contemporary sociologists also characterize social systems as having 'structures or structural properties,' although not being structures in themselves (Giddens, 1979).

Second, the social system is defined by emergent properties, thus as an emergent phenomenon resulting from and typified with the combination, interrelations, interaction, association, and cooperation of elements. In terms of its emergent properties, the social system is considered the system of action and interaction between individual and especially collective actors. In early sociology, both Marx (1933[1849]) and Durkheim (1982[1895]) define the social system, i.e., society as more than an aggregate or sum of individuals and in terms of their 'relations' or 'association,' although the first emphasizes the 'relations of production in their totality' as constituting the 'social relations, society,' and the second focuses mostly on their noneconomic, nonmaterial forms like social institutions, while Spencer (1969[1876–96]) suggests that a necessary condition for the formation of society in the 'sociological sense' represents cooperation (spontaneous or consciously devised), beyond mere 'juxtaposition.' (Habermas (1975: p. 2) comments that "Marx developed for the first time a social-scientific concept of system" as well as system crisis in the form of capitalism. In turn, according to Collins (1988), "Spencer [was] the first general systems theorist.") In particular, Durkheim (1965a[1893]) proposes that the social system created by the 'association,' i.e., 'positive and durable relations' of a 'plurality of individuals' constitutes a 'specific reality' with its own distinct irreducible properties, thus an emergent phenomenon from their 'uniting together,' which forms the 'base' on which all social life is built. Similarly, Tönnies (1955 [1887]) characterizes society as a social system involving reciprocal relations or interactions, Pareto (1963[1916]) proposes that socially connoted elements compose the social system only through 'mutual correlations in social phenomena,' and Simmel (1964[1908]) considers 'large (social) systems' to be 'immediate (and constant) interactions' among individuals crystallizing as 'autonomous phenomena' and attaining their 'own existence and laws,' indicating emergent system properties. (However, Simmel (1964[1908]) relaxes, if not contradicts, the characterization of social systems by emergent properties stating that society "merely is the name for a number of individuals, connected by interaction", as also Smelser (1997: p. 30) comments.)

In later sociology, Sorokin (1957) defines social systems as 'systems of social relationships,' 'systems of interaction,' and the 'network of social relationships,' and Parsons (1951) as 'systems of social action' involving units and their 'structural interrelations,' a 'system of processes of interaction between actors,' thus the 'structure of the relations' between actors and

a 'network of interactive relationships.' In a similar vein, contemporary sociology characterizes social systems as 'systems of social interaction' with 'structures or structural properties' (Giddens, 1979) and 'social action systems' (Fararo, 2001).

Third, the social system is defined in the sociological and related literature by causal properties such as reciprocal causation or conditioning, codetermination, and interdependence. On account of such properties, the social system constitutes a causal, interdependent complex of interrelated elements generating various systemic outcomes during its own operation and in interplay with the environment (Luhmann, 1995). Spencer (1969[1876–96]) states that the units of the social system tend to be in the 'state of mutual dependence,' Durkheim (1982[1895]) posits that social systems are linked by 'bonds of interdependence,' Simmel (1964[1908]) asserts that society always involves connected individuals by 'mutual influence and determination,' and Mannheim (1956) proposes that a social system is characterized with 'causal sequences' operating recursively in its social structure. Also, contemporary sociologists conceive social systems as consisting of 'regularized relations' of interdependence between individuals or groups, i.e., social practices (Giddens, 1979), and as 'relations among elements' that are regulated through reciprocal 'conditioning' with 'successful conditionings' acting as systemic constraints (Luhmann, 1995).

Fourth, the social system is defined by unifying properties, and thus in terms of unity, unification, and integration. With regard to such properties, the social system represents or reaches a unity or union of interrelated and interdependent elements. Durkheim (1965b[1915]) argues that social systems have a 'certain unity'; Simmel (1955[1923]) assumes that they tend to reach their 'unity' in the form of their 'organic structure.' Contemporary sociological systems theorists condense these properties characterizing social systems in the terms of the 'unity of the whole' (a *unitas multiplex*) (Luhmann, 1995).

Fifth, the social system is defined by institutional, organizational, and functional properties, i.e., in terms of institutions and rules, organizations, and functions. In this respect, the social system constitutes an institutional, organizational, and functional complex or arrangement of interrelated, interdependent, and unified elements, i.e., a set of institutions and rules, organizations, and functions. Tönnies (1955[1887]) describes the modern social system (*Gesellschaft*) as an 'entirely artificial, man-made institution' permeated with a 'system of conventional rules,' Sorokin (1957) incorporates 'associations, organizations, institutions, and their interdependence' as well as entire societies or cultures into social systems, Parsons (1937) contends that the 'structure' of social systems entails especially 'institutionalized patterns' as well as 'a commonly understood system of cultural symbols,' and Merton (1968) registers the 'interdependence between social institutions' constituting a social system. Contemporary sociologists note that modern social systems theory concentrates on the 'functional rationality of large organizations' in Weber's tradition (Habermas, 2001), and that functionalism is a major 'subtype of system theory' (Collins, 1988). (Collins (1988: p. 80) adds that "classical Marxian theory is above all a system theory. It sets out all the elements of a social system and shows how they interrelate. The model is both static and dynamic.")

Sixth, the social system is defined by collective, objective, and complex properties. In this sense, the social system constitutes a supraindividual entity or collectivity of interrelated, interdependent, unified, and institutional elements characterized with objectivity and complexity. Durkheim (1966[1897]) postulates that the social system, above all society, represents inherently a 'system' of collective objective realities in being composed of other simpler societies or groups (the simplest being horde or clan), and similarly, Parsons (1951) considers both a society as a total social system and any 'complex partial social system' to be a 'network of collectivities.' Simmel (1955[1923]) recognizes that social systems are 'superindividual' entities, notably society being a 'structure' transcending individuals and the 'universal' but 'concretely alive,' such that the 'objective structure' of society forms a 'framework' of individuals and their characteristics and activities. Evoking Durkheim and Marx, contemporary sociologists depict social systems like societies as 'systems of objective relations' or 'objective potentialities' in the present reality, in particular, the 'material conditions of existence of a class condition' (Bourdieu, 1988), and suggest that the social system is an objective and external 'environment' for individual agency, thus causing structures to exist 'outside of actors' only within this systemic context (Alexander, 1998).

Seventh, the social system is defined by static, reproducing properties such as persistence or gradual change, self-reproduction, self-coordination, self-regulation, and self-evaluation or self-reflection. On this account, the social system constitutes a static, persistent or gradually changeable, self-reproducing, self-coordinating, and self-reflective or self-referential complex of interrelated, interdependent, unified, institutional, collective, and complex elements. Parsons (1951) defines a social system as a 'mode of organization of action elements' in connection with the 'persistence or ordered processes of change' of the system's 'interactive patterns.' Similarly, contemporary theorists stipulate that the particular elements constituting the social system have 'no duration' relative to it as a relatively enduring entity constantly reproducing them and thus expressing a 'peculiar constraint on autonomy' to the effect that the necessary condition for system existence is its endowing its 'momentary elements' with the 'capacity for connection' and thus their reproduction (Luhmann, 1995). In this context, social systems constitutes and function as systems only if referring to 'themselves' (Luhmann, 1995), thus becoming 'self-referential' (Collins, 1988).

Eighth, the social system is alternatively defined by non-static, dynamic properties such as growth, development, substantial or rapid transformation, and change, including crisis. In this regard, the social system constitutes an actually or potential dynamic, including crisis-prone, complex of interrelated, interdependent, unified, institutional, collective, complex, and reproducing elements. In early sociology, Comte (1983), Spencer (1969[1876–96]), Marx (1967[1867]), Tönnies (1955[1887]), Durkheim (1966[1897]), Simmel (1990[1900]), Pareto (1963[1916]), and Weber (1968 [1920]) observe, emphasize, and predict rapid and substantial social system change or dynamics, including crisis, especially recurring economic as well as political, religious, and moral crises in modern society. Also, some contemporary sociologists point to a 'system crisis,' particularly economic and

political crises, in the modern social system such as capitalist society (Habermas, 1975), and generally to the existence or possibility of 'dynamical social systems' (Fararo, 1989).

Ninth, the social system is defined in the sociological literature by time and space properties or coordinates. Owing to such properties, the social system exists and develops as an entity situated and delineated in time and space incorporating interrelated, interdependent, unified, institutional, collective, complex, reproducing, and dynamic elements. In such a characterization, by contrast to structures social systems exist, persist, and change in time–space, while being composed of concrete 'social practices' (Giddens, 1979, 1984).

Lastly, the social system is defined by bounded properties, i.e., boundaries or limits of their operation and validity, and thus degrees of relative closure and openness. Consequently, the social system constitutes a bounded or boundary-maintaining and relatively closed or open set of interrelated, interdependent, unified, institutional, collective, complex, reproducing, dynamic, and time–space elements. Parsons (1951) treats the social system as an instance of 'boundary-maintaining' systems, while critics observe that 'one of the prevalent (and 'unfortunate') 'connotations' of the term 'system' is its 'closedness' (Dahrendorf, 1968) and thus boundary maintenance. In this definition, they are boundary-maintaining systems by creating and maintaining 'boundaries' as a distinguishing trait of 'system' vs. 'structure' (Luhmann, 1995), although only modern societies such as nation-states are observed to have 'specifiable boundaries' (Giddens, 1984). (According to Giddens (1984: p. xxvi) the "degree of systemness in social systems is very variable and societies rarely have easily specifiable boundaries [until] the modern world of nation-states.")

Types of Social System

Certain (ideal or pure) types of social systems are distinguished in the sociological and related literature. In one typology, social systems are classified into simple and complex types, with intermediate cases, on the basis of the degree of simplicity or complexity of their structure or organization and functioning. Durkheim (1965b[1915]) seeks to identify a 'social system surpassing all others in simplicity' and contrasts it to more complex systems arising from various combinations and extensions of the simple, primitive type. In contemporary systems theory, what typifies complex social and other systems is the necessity of adaptation both to their 'environments,' i.e., other systems, and their own 'complexity' (Luhmann, 1995), with modern 'liberal society' being regarded as 'much more complex system' than portrayed by rational choice theory as the economic model of both market and noneconomic systems (Munch, 1994).

Another typology incorporates total and partial social systems. This is the probably most important typology in sociological terms as it distinguishes society as a total social system from its units such groups or collectivities ('societies') and domains within it as such partial systems. Durkheim (1965a[1893]) uses the term the 'total' social system with reference to society as a whole as distinguished from 'partial' systems or societies within it, including 'domestic society,' 'economic society,' 'religious society,' etc., and also Pareto (1963[1916]) understands the 'sociological system' in the meaning of societal totality. In particular, Tönnies (1955 [1887]) observes that modern capitalist 'Society' epitomizes a total social 'system' in the sense of being a 'totality' in respect of the 'limitless' scope of application of a 'system of conventional rules.' This elaborates on or evokes classical Marxian sociological theory in which the total social system analyzed is modern capitalism (Collins, 1988).

Elaborating on Durkheim and Pareto, Parsons (1951) explicitly distinguishes total and partial social systems as societal and subsocietal systems, respectively. In this typology, a total or 'empirically self-subsistent' social system, in the sense of fulfilling 'all the essential functional prerequisites of long-term persistence from its own resources,' represents a society, while any other social systems being 'partial.' Notably, Parsons (1951) proposes that society is a 'norm' for the social system such that any 'partial' social systems are placed and analyzed in its 'setting' as its parts. Sorokin (1957) defines a social system as 'unlimited, or totalitarian' in its 'extensity' to the extent of its embracing 'all the relationships of its members,' regulating their 'entire behavior and all their interrelations,' while distinguishing 'totalitarianism' from *laissez-faire* as the 'fluctuation of the quantitative aspect of the network of social relationships.' Sorokin's dichotomy of totalitarianism and *laissez-faire* is not identical to Durkheim-Pareto-Parsons' typology of total and partial social systems, so long as total social systems cum societies are not necessarily 'totalitarian,' while partial systems not invariably being *laissez-faire*. A contemporary distinction between the 'double meaning' of society as a 'bounded system' and as 'social association in general' (Giddens, 1984) implies a distinction between total and partial social systems cum societies and groups, respectively. Thus, modern national societies are considered the sole or main example of a 'bounded' social system. Also, a dichotomy between continuous and discontinuous social systems is implied in Simmel (1964 [1908]) observing that under certain conditions 'society would break up into a multitude of discontinuous systems.'

The probably most common typology incorporates closed and open social and other systems is based on the relative degree of system closure and openness and of rigidity and fluidity of boundaries. Generally, a closed social system attains (monopolistic) closure and maintains its rigid boundaries in relation to other systems as the environment (Parsons, 1951; also, Collins, 1988; Habermas, 2001; Luhmann, 1995), in contrast to an open system characterized with openness and fluid or permeable boundaries. Weber (1968[1920]) defines a social system or relationship as closed against 'outsiders' to the extent that their participation in it is 'excluded, limited, or subjected to conditions,' and open so long as the 'system of order' permits such participation to 'anyone' outside of it. In his view, a social system is closed or open depending on traditional, affective, and instrumental- and value-rational determinants, and the 'principal motives' for system closure involve "(a) the maintenance of quality; (b) the contraction of advantages in relation to consumption needs; (c) the growing scarcity of opportunities for acquisition" (Weber, 1968[1920]). An exemplar of this dichotomy is that between 'status society' and 'class society' (Weber, 1968[1920]) as closed and open social/stratification systems, which also

depicts or predicts a trajectory and development from system closure to greater openness – and to some extent renewed closure followed by a new opening, and so on – as do most classifications based on this criterion (Habermas, 2001; Luhmann, 1995). For example, Mannheim (1956) identifies 'different social systems of authoritarian and democratic character' in which the first embody closed 'order' and 'discipline,' and the second 'fluidity' and 'openness' as 'antithetical human ideals,' and which also represent successive, lower, and higher phases of system evolution in modern society.

Much of contemporary sociological and general systems theory treats closed social and other systems as a 'limit case' in the sense of systems attributing 'no significance' to the environment in contrast to open systems typified with such an attribution (Luhmann, 1995). Furthermore, some versions replace the closed–open system dichotomy with the distinction between 'self-referential closure' and 'openness' yielding the concept of self-referential, self-producing (autopoietic) social and other systems (Luhmann, 1995), which consider their own interrelations and goals 'part of the environment' (Collins, 1988). This is done on the ground that the dichotomy does not represent an 'opposition of types but a relationship of intensification' in that social and other systems may close and open at the 'same time' through intensifying and relaxing their boundaries as 'specific mechanisms' for the 'purpose of separating yet connecting (via) selection,' so the notion of a self-referential closed social system is not inconsistent with its 'openness to the environment' (Luhmann, 1995; also, Habermas, 2001).

Classification of Historical and Empirical Social Systems

Various classifications of historical–empirical social systems, first and foremost taxonomies of past and present societies, in the sociological and related literature (Lenski, 1994) can be divided into evolutionary and cyclical. Evolutionary (linear and nonlinear) and cognate classifications are prevalent in the literature, and treat social systems at certain historical–empirical stages as irreversible or nonrecursive over time in a continuous sequence from the lowest to the highest types as the supposed beginning and pinnacle of societal evolution.

In early sociology, Comte (1983) classifies historical social systems and the 'general systems of conceptions,' based on prevalent 'principal conceptions,' into the theological system, including the 'theocratic' and 'military' polity exemplified by feudalism, the metaphysical, and the positive, i.e., rationalistic, scientific, and industrial, as the assumed 'final' state of evolution. Spencer (1969[1876–96]) essentially adopts Comte's classification by classifying social systems, on the basis of 'unlikeness' between the kinds of predominant social activity and the 'resulting unlikeness' of organization, into 'militant' and 'industrial' types, the first being premised on the 'system of status' involving 'successive grades of subordination' and the principle of 'compulsory cooperation,' the second (in its 'ideal form') on the 'system of contract' and 'voluntary cooperation.' Spencer's (1969[1876–96]) full classification of historical social systems, in accordance with their 'degree of composition,' incorporates 'simple, compound, double-compound, trebly compound, and predominantly militant and predominantly industrial' societies. Marx (1933[1849]) classifies historical social systems into tribal, Asiatic, Ancient, feudal, and modern capitalist (or bourgeois) societies deemed 'totalities of relations of production' and different 'modes of production' as 'epochs' manifesting 'progress in the economic development of society' to be fully realized in a future postcapitalist society (communism). Similar to Comte and Spencer, his historical taxonomy of social systems assumes the main form of a dichotomy between 'feudal society' and the 'capitalist regime,' and thus an evolutionary transition from the first to the second system in treating the 'economic structure of capitalist society' (the 'bourgeois order') as emerging from 'the economic structure' of feudalism (Marx, 1967[1867]).

The probably best known and also most debatable sociological dichotomy of historical–empirical social systems is that of Tönnies (1955[1887]) dividing them into traditional Community (*Gemeinschaft*) as the 'lasting and genuine form of living' involving 'all intimate, private, and exclusive living together,' and modern capitalist Society (*Gesellschaft*) as instead 'transitory and superficial' typified by 'public life,' in a developmental sequence from the first to the second. However, Tönnies (1955[1887]) does not consider capitalist Society the final point of societal evolution or a desirable social system mostly on the account that 'all noncapitalists' within it are treated either 'like inanimate tools' forming the 'very essence of slavery' or as 'legally nonentities' denied the capacity for 'rational choice' consistent with Marx and contrary to Comte and especially Spencer. Specifically, Tönnies (1955[1887]) proposes that the 'whole movement' of historical–empirical social systems involves a 'transition' from 'original, simple, family-based communism' (Community) through 'absolutely detached' capitalism (Society) and ultimately to 'state-based and international socialism' resulting from the capitalist system.

Similar but identical to Tönnies, Durkheim (1965a[1893]) classifies total social systems into a society constituting a 'totality of beliefs and sentiments common to all members' ('collective conscience') and a 'system of different, special functions that definite relations unite,' i.e., societies with mechanical and with organic solidarity in an assumed evolutionary trajectory from the first to the second. Durkheim (1965a[1893]) states that an 'historical law' of societal evolution is that initially exclusive 'mechanical solidarity' tends to 'progressively' weaken, and 'organic solidarity' resulting from the highly developed division of labor to eventually prevail, thus 'two social types' corresponding to these 'two types of solidarity' indicated by repressive and restitutive law, respectively.

Simmel (1955[1923], 1990[1900]) provides various historical instances of social/stratification systems such as the Hindu caste system (also, Sorokin, 1957), the estate system in Europe, the class system, the merchantilist system, the ancient regime, the despotic regime, and the modern 'money economy' and culture as the latest and growingly important or prevalent example. Weber (1968[1920]) condenses various historical social-economic systems into traditional and modern

rationalist society/state/culture, notably economic traditionalism and rationalism epitomized by feudalism and modern capitalism, respectively, in an assumed (nonlinear) evolutionary trajectory from the first to the second (Parsons, 1937). (Parsons (1937: p. xv) describes Weber as "overwhelmingly the most important of the post-linear evolutionists.") Mannheim (1936) accepts Weber's dichotomy by distinguishing traditionalism and modernism in a developmental process from the first to the second, as does Parsons (1966) in his classification of concrete social systems into 'primitive, intermediate, and modern' societies, with the latter, specifically America, extolled as the last point and pinnacle of societal evolution (also, Giddens, 1984).

Cyclical and pseudocyclical classifications are also found in the sociological and related literature, and consider historical social systems recursive or recurring and reversible over time, as specific predictable elements or stages of an overall system cycle or rhythm alternating or fluctuating between them with some degree of regularity and so predictability, even if not necessarily periodicity (e.g., 10 or so years as in economic cycles prior to the Great Depression). Tönnies (1955[1887]) implies a cyclical system classification or conception registering 'a continuous process of simultaneous progress and decline' of some sociocultural systems (e.g., Germanic culture since the Roman Empire), and also observing that 'the power of Community, even in decline,' persists in the 'era of Society' qua capitalism and as the 'true reality of social life.' Moreover, the cyclical element is more explicit in Tönnies' (1955[1887]) observation and prediction that Community is or will be revived in modern 'socialism' after its decline in 'capitalism,' compatible with Marx's developmental trajectory from primitive to future 'communism' via the 'capitalist system' and other class-based social systems (Giddens, 1984) as an intermediate stage.

Among these classifications, the probably most explicit and prominent or notorious is Pareto's (1963[1916]) cyclical classification of past and present sociopolitical systems into aristocracy as elite rule, transient democracy as the '(un)stable reign by the masses,' and again aristocracy or oligarchy, and so on in the process of constant cycle or 'circulation of elites' versus the 'people.' Pareto (2000[1901]) contends that such cycles in societal rule recur in accordance with the 'great law of rhythm' in society, including 'waves' in ethics, religion, and politics, just as in economy (the 'business cycle'). Also, Sorokin (1957) presents a modified cyclical classification of historical social-cultural systems into 'ideational, sensate, idealistic, and mixed,' specifically systems of 'ascetic ideational,' 'active ideational,' 'active sensate,' 'passive sensate,' 'idealistic/mixed,' and 'cynical and pseudoideational' mentalities. Like Pareto, Sorokin (1957) posits what he denotes a 'rhythm of rise and decline of Ideational-Idealistic-Sensate' sociocultural systems, and even a 'super-rhythm' of such 'supersystems' of society/culture.

Structural Elements and Relations of the Social System

The structural elements of social system are denoted structures or subsystems, ranging from economic and political to cultural and others, and considered interconnected and interdependent.

Various divisions and conceptions of social-system structural elements or subsystems are found in classical and contemporary sociology and the related social sciences. In classical sociology, Comte (1983) incorporates in the social system such elements or subsystems as the 'social economy,' the polity, 'intellectual' and 'moral' phenomena, i.e., material, political, intellectual, and moral 'order' and 'power.' Marx (1970[1859]) distinguishes the 'economic structure of society' as the latter's 'real foundation' from its 'legal and political superstructure' and 'forms of social consciousness.' Spencer (1969[1876–96]) divides social-system elements into 'coordinating structures and functions' such as family, political and ecclesiastical structures and functions, and ceremonial institutions, and structures and functions 'coordinated' like 'regulative and operative industrial structures.' In his view, such structures and functions constitute the 'organization and life of each society' characterized with the 'interdependence of structures, functions, and products.' Also, Durkheim (1982[1895]) considers religious, moral, political, legal, and economic institutions – defined generally as 'all the beliefs and the modes of conduct instituted by the collectivity' – to be structural elements of the social system, including 'present-day social institutions.' For classical sociologists, especially Comte and Durkheim, the social system is logically and empirically both more comprehensive in scope and more complex in structures and functions than any of its structural elements or particular systems, including economy (except for Marx), as well as polity. Pareto (1963[1916]) condenses and articulates this position stating that the general 'sociological system' in the sense of society is invariably 'much more complicated' than the economic and any particular system.

Within later sociology, Sorokin (1957) and Parsons (1967) treat and denote social-system elements as 'subsystems' or 'structural components.' Parsons (1951) classifies the 'structural components of the social system' into the act as the unit of any 'system of action,' the status role as a 'higher order unit' in a 'macroscopic analysis of social systems' due to statuses and roles being 'not attributes' of actors, but the 'units of the social system,' the individual actor as a 'composite bundle' of statuses and roles vs. personality as a 'system of action,' and collectivity as 'actor and as object.' In particular, Merton (1968) focuses on the 'concept of social status' as a structural component or 'position in a social system' involving a 'distinctive array of designated rights and obligations,' and on the 'related concept of social role' referring to the 'behavior of status-occupants' in view of the 'patterned expectations of others.' In turn, Dahrendorf (1959) centers on the 'capitalist class system,' contrasted with the precapitalist chieftain system, in connection with the stratification 'system of social rewards,' as the crucial structural component of the modern industrial social system.

Similarly, contemporary sociologists emphasize 'stratification systems' (Collins, 1975) and different societal, including economic, political, and cultural-symbolic, 'domains' or 'fields,' 'power' or 'capital,' and 'subsystems' (Bourdieu, 1988) as 'social system parts' (Alexander, 1983)

or 'units' (Luhmann, 1995). For example, some divide 'constituents of social systems' into the 'exchange between social systems and their environment,' the 'goal values of social systems,' and the 'level of development of society' (Habermas, 1975), in which, however, 'social subsystems' by operating per their own logics or languages tend to be 'deaf' to their 'external effects,' as exemplified by the markets with its logic/language of prices (Habermas, 2001). According to another contemporary systems theory, social and other systems involve subsystems in the sense of 'internal system/environment relations within the system' and are disaggregated into 'elements and relations,' resulting in growing 'system differentiation' and 'system complexity' (Luhmann, 1995).

The structural elements or subsystems of a social system are typically characterized with interconnection and interdependence. Comte (1983) emphasizes the 'action and reaction of the different parts of the social system,' Marx and Engles (1947 [1847]) the 'connection of the social and political structure with production' forming the economic structure, and Simmel (1990[1900]) the interaction between an 'ideal structure' and an 'economic structure.' Sorokin (1957) and Parsons (1951) stress the interaction or interchange between various, including cultural, economic, political, and communal, subsystems in the total social system, and Merton (1968) the 'interdependence between social institutions' within the latter. In particular, Dahrendorf (1968) suggests that the 'social structure of power and authority' is linked with, by virtue of preceding and explaining, social stratification as a 'system of distributive status, i.e., a system of differential distribution of desired and scarce things,' as two structural components of the social system.

The Economic System and the Social System

One structural element or subsystem of the social system consists of the economy, i.e., the economic system. Sociological systems theory identifies and focuses on certain features, components, and effects of the economic system analyzed within the framework and operation of the total social system in contrast to conventional economics mostly analyzing economy as an isolated, independent, self-contained entity in relation to society.

First, these features of the economic system include cooperation and the division of labor. Comte (1983) states that 'in society the elementary economy' is characterized with the sentiment and activity of cooperation as 'preponderant' (and with 'sympathetic instinct' as secondary), thus the 'principle of cooperation' providing the basis for 'social statics' applied to economy, which anticipates 'comparative statics' in economics (Samuelson, 1983[1948]) and its extension in social systems theory (Fararo, 1989). Similarly, Spencer (1969[1876–96]) characterizes industrial society by the principle of voluntary and spontaneous cooperation implemented and expressed through the 'system' of commercial and other contracts. In particular, Comte (1983) sees the social division of labor, i.e., the 'distribution of functions' as the crucial feature of the modern economy, as do Spencer (1969[1876–96]) and especially Durkheim (1965a[1893]) who emphasizes the societal causes and effects of this process, as well as Marx (1967[1867]) analyzing its adverse results, and Adam Smith (1976[1776]) in economics focusing mostly on its internal operation and productivity within a production unit.

Sociological systems theory identifies another cardinal feature of the economic system in the production, distribution, and consumption of wealth or material goods as the standard definition of the economy and thus the subject matter of economics, especially its classical variant represented by Smith and his successors (Ricardo, Say, Mill, Marx as an economist, etc.). A variation on this feature is Marx's concepts of the forces and relations of production, the first comprising productive instruments combined with labor, and the second property relations in the sense of exclusive ownership and control of these resources in relation to nonowners. Marx (1970[1859]) uses the 'totality of these relations of production' to define and conceptualize the 'economic structure of society' considered the 'real foundation' for erecting a 'legal and political superstructure' and reflected in 'definite forms of social consciousness.' Durkheim (1982[1895]) notes that wealth is the 'object of political economy' as a social science in virtue of being deemed a defining and constitutive feature of the economic system, although rejecting economics' focus on the 'desire for wealth' as the supposed sole or main human motive. Durkheim (1994[1909]) includes among economic institutions those 'institutions relating to the production of wealth (serfdom, tenant farming, corporate organization, production in factories, in mills, at home)' and 'institutions relating to (its) distribution (rent, interest, and salaries).' Parsons and his associates consider wealth the main feature and outcome or output of the economic system for the 'total society' (Parsons and Smelser, 1956), and later sociologists suggest that economy solves the 'problem of production of economic resources or wealth for the social system' (Fararo, 2001), and note 'national systems of production' of wealth among modern societies (Habermas, 2001).

In sociological systems and other theories, like economics, additional and related attributes and elements of the economic system comprise exchange, money, prices, markets, contracts, entrepreneurship, and the like. Spencer (1969[1876–96]) defines 'industrial society' by the 'system of contract' in the economic and entire social system as a legal form of voluntary and spontaneous cooperation. Tönnies (1955[1887]) considers the 'market system' the essence and even equivalent of modern capitalist Society, and particularly emphasizes that, above all, within such a 'social system' money operates as 'power over all goods and pleasures' reducible to it as a 'common denominator' evoking Marx's (1967[1867]) theory of money. Also, Simmel (1990[1900]) considers and denotes the modern and in part earlier economic system as the 'money economy.' Weber (1968[1920]) depicts the market/price system as a 'struggle of man against man,' with money as its primary 'weapon' and prices as its 'expressions,' including a 'monetary system with the highest degree of formal rationality' in the form of financial accounting, as does in a similar way Veblen (1898). Weber (1968[1920]) implies that a defining feature and integral element of the modern capitalist economy is an 'economic establishment' constituted and operating as an 'organized system of continuous economic action,' thus what Mannheim (1956) calls the 'entrepreneurial system,' as Schumpeter (1949[1911]) especially analyzes and

emphasizes in economics. In turn, Durkheim (1994) exemplifies economic institutions and systems also by those institutions 'relating to exchange,' including 'commercial organization, markets, stock exchanges,' and as well as by 'monetary or financial systems,' and observes that contracts pervade economy while 'merely' expressing the 'respective state of economic forces.'

Also, sociological systems and related theory distinguishes various types of economic systems, as a particular variation of its typology of the total social system. Comte (1983), like Saint-Simon (1964[1816]) before, and Spencer (1969[1876–96]) identify and contrast primitive and modern economic systems, in particular feudalism and serfdom versus industrialism and capitalism. Marx (1933[1849]) identifies, within his central focus on capitalism, precapitalist, capitalist, and post-capitalist, including socialist and communist, economic systems as successive stages of the evolution of the economy, while predicting the 'inevitable destruction of the middle bourgeois classes and peasant estate under the present (economic) system' as the 'system of bourgeois economy.' As a variation on Marx's typology, Tönnies (1955[1887]) differentiates communal, capitalist-market, and socialist economic systems in an evolutionary sequence from the first through the second and to the third. Similarly, Simmel (1990[1900]) sharply contrasts the natural or self-subsistent economy versus the money economy, and Weber (1968[1920]) economic traditionalism and capitalism as exemplary economic modernism or rationalism.

Furthermore, sociological systems and related theory analyzes the relations between the economic system and the social system, unlike orthodox economics due to its treatment of the economy as an isolated, self-contained entity. Notably, it defines and specifies these as the relationship between economy as an element or subsystem and society as a system, i.e., as a part and a whole, in contrast to and differentiation from orthodox economics treating the first as the overarching system virtually outside and beyond society. Moreover, the relationship between the economic and social system, i.e., economy and society, has been a prevalent and continuing concern of sociological theory since Comte and Marx through Durkheim, Weber and Pareto through later (Parsons, 1937; Parsons and Smelser, 1956) and contemporary systems and other theorists (Fararo, 2001). (Fararo (2001: p. 177) suggests that "much of the tradition of general theoretical sociology has concerned the relationship between economy and society.")

Within this context, the prevailing and even consensual sociological premise and finding is that the economic system is a part or subsystem of and thus situated and implicated ('embedded') in and conditioned by the social system as the comprehensive whole, thus applying an holistic approach to the relations of economy and society. In a classic example, Comte (1983) posits that the 'various functions of the social economy' tend to be 'naturally implicated in relations of greater generality' constituting the social system, thus expressing the 'constant action of the whole upon the parts.' Echoing Comte, Pareto (1963[1916]) suggests that the economic system, constituted by logicorational actions motivated by 'tastes' and subjected to 'obstacles to the acquisition of economic values,' operates within the 'sociological system' deemed 'much more complicated' in virtue of being composed of both 'logical and nonlogical actions' induced by a wider set of 'residues (sentiments), derivations, interests, and proclivities.' Durkheim (1965a[1893]) proposes, in an influential proposition in economic sociology and social economics, that a commercial contract far from being 'sufficient unto itself' is 'possible' only due to a 'regulation' of an 'essentially social' character, because in such contracts not all elements are 'contractual,' thus suggesting the societal, notably institutional, bases or conditions of this and other market transactions. Evoking Pareto and Durkheim, for Parsons (1937) the economic system/concept only arises and operates, or is possible and understandable within 'social systems of action,' although being also applicable to the 'system of action of a particular individual' yielding 'Crusoe economics' as a fiction. In particular, Parsons (1951) characterizes wealth cum economic power as a 'lineally quantitative element or variable in the total equilibrium of social systems' in contrast to its political form as a nonlinear (zero-sum) variable.

Also, contemporary social systems theorists analyze the relationship between the economic and the social systems as that between a unit and a whole. Economy is considered the 'behavioral system of instrumental action' and 'technical control' (Habermas, 1971) within the total social system comprising also noninstrumental or noneconomic (political, communicative, and symbolic-cultural) actions (also, Alexander, 1998). For instance, the 'economic level of social systems' (Habermas, 2001) is exemplified by capitalism and communism during the Cold War, as well as varieties of capitalism during and after that period. In this respect, economy is characterized as both a particular 'social system' and a subsystem of the general social system (Luhmann, 1982). Based on its component of material resources and transactions, economy is deemed a 'closed self-referential system' as 'only' half of its equation and 'operational meaning' because it is also an open or opening subsystem of the social system in virtue of its functioning 'ultimately' taking place within or referring to the latter as its 'environment' (Luhmann, 1995). In this conception, the 'differentiation' of the economy as a particular system within the whole social system results from the 'interconnection of the conditions for closure and openness' because the 'unavoidable coupling of self-and-other-referential meaning references in all economic operations' necessitates definite social-systemic 'conditions' (Luhmann, 1995). Also, a contemporary Parsonian model treats economy as a 'functional subsystem' and thus with its own 'environment' consisting of the other subsystems of the total social system, while representing an 'ongoing solution to the problem of social adaptation' in the form of 'production of economic resources or wealth for the social system' (Fararo, 2001). In short, as leading contemporary economists admit, economy is an integral element, component unit of the social system (Arrow, 1994) within social systems theory (in contrast to 'pure' economics).

The relative significance of the economic system in the social system is determined and specified in two main alternative ways in sociological systems and related theory. In one rendition, economy represents the main subsystem and the prime causal factor of the social system as a whole, a position held by Marx, Tönnies, and most economists since Smith as well as 'rational choice' theorists. Marx argues that the 'social

structure and the State' evolve from of the economic 'life process of definite individuals' placed under 'definite material limits, presuppositions, and conditions independent of their will' (Marx and Engels, 1947[1847]). Arguably, the 'totality' of the relations of production making the 'economic structure of society' forms the 'real foundation' of polity ('legal and political superstructure') and culture ('social consciousness') (Marx, 1970[1859]). In general, for Marx's economic/conflict systems theory, orthodox economics and 'rational choice theory,' economy is the principal structural element and the explanatory variable of the social system.

Another alternative conception considers economy a relatively secondary subsystem and the causal effect or function of the social system, including political and cultural systems, a view characteristic of Comte, Simmel, Durkheim, Weber, Sorokin, Parsons, contemporary economic sociologists, social economists, etc. Comte (1983) and Durkheim (1965a[1893]) treat the 'social economy' as largely a secondary element of and determined by or implicated and embedded in the total social system, Simmel (1964[1908]) propounds that the 'form of economics, too' represents "merely a 'superstructure' (within) the purely sociological structure" contradicting Marx, while Weber (1968[1920]) states that an economic system crucially depends on 'protection by force' and thus on the 'political system.' Also, Sorokin (1957) and Parsons (1951) mostly regard economy as a subsystem of secondary or equal causal significance within the social system, especially in relation to the cultural system in the first theorist and the 'value-integration' system in the second. For instance, in Parsons' (1951, 1967) Adaptation, Goal-Attainment, Integration, Latent-Pattern-Maintenance (AGIL) model, economy (fulfilling the functional imperative of 'adaptation' A) is largely a secondary subsystem compared to the 'integrative system' (attaining system integration I) as the primary within the total social system and approximately of equal significance as polity (goal-attainment G) and culture (latent-pattern maintenance L). Generally, in Comte–Durkheim–Sorokin–Parsons' integrative and functional systems theory, economy represents mostly a subsystem of secondary relative significance and the dependent variable of the social system, in particular value/cultural systems.

Yet another alternative conception seeks to reconcile or integrate the previous two by treating economy as subject to historical and empirical variations in relative system significance and as both an explanatory factor and the dependent variable of the social system. Such a conception is especially characteristic for Weber (1968[1920]) in his main work *Economy and Society* and his followers, including to some degree Parsons in a book with the same title (Parsons and Smelser, 1956), as well as Simmel (1990[1900]) and Pareto (1963[1916]).

The Political System and the Social System

Another structural element or subsystem of the social system comprises polity, i.e., the political system. Sociological systems theory assumes and analyzes certain attributes, constituents, and consequences of polity analyzed, as economy, within the setting and functioning of the total social system in partial contrast to political science if the latter considers its subject as mostly a self-contained, autonomous structure vis-à-vis society.

An essential attribute or outcome of the political system is considered social regulation, direction, restraint, and order for public ends. Comte (1983) posits that the 'only object of any political system' its regulation of the 'spontaneous' social cooperation to 'expansion' in the function of directing such processes toward a 'determinate end,' especially government regulating the 'social economy' in its 'various functions' that are 'naturally implicated' in the social system of 'relations of greater generality.' Spencer (1969[1876–96]) proposes that political organization epitomized by government seeks to regulate 'affairs' in the social system through combining the 'actions of individuals' for common purposes (e.g., 'tribal or national offense and defense'), and restraining individual and groups 'in certain of their dealings with one another,' thus forming a specific 'part of social organization' performing 'directive and restraining functions for public ends.' In short, polity's main attribute or outcome in the social system is what Simmel (1990 [1900]) denotes the 'sociopolitical organization and order.'

Another crucial attribute or outcome of the political system is considered power or authority, although the latter is not necessarily limited to polity but generalized as social power or domination to the total social system, including economy. Comte (1983) suggests that 'every social power,' notably political 'authority,' derives from and is grounded on a 'corresponding assent of various individual wills' in connection with a collective, 'common action' rendering power its 'organ' initially and 'regulator' ultimately. Comte infers that all 'great power' develops only the 'prevalent disposition of the society,' conceivably the 'radical consensus proper to the social (system),' thus being legitimate domination in the sense of Weber. Weber (1968[1920]) famously classifies (ideal) types of political authority understood as legitimate power/domination into charismatic, traditional, and legal-rational, and also distinguishes between 'domination by virtue of a constellation of interests' (including monopoly) characterizing economy from 'domination by virtue of command' typifying polity. Parsons (1951) points out the 'double aspect' of social (and other action) systems, one being the 'problem of power relations,' another that of Hobbesian societal 'order,' both representing a 'well-marked set of emergent properties' of these systems in the form of 'political action elements.' More explicitly and systematically, contemporary conflict systems and related theory treats and depicts polity and all society as the 'system of organizational power' (Collins, 1975), i.e., of authority in 'imperatively coordinated associations,' including political as well as economic ones and others (Dahrendorf, 1959).

Yet another attribute and constituent of polity is considered law, i.e., the legal system. Tönnies (1955[1887]) emphasizes that a 'common and binding system of positive law, i.e., system of norms' represents an essential trait and component of the modern polity and (capitalist) Society overall such that the 'revolution in the social order' is followed by a 'gradual change of the law,' with contract becoming the 'basis of the entire system' sustained or modified by the conjunction of 'rational will' induced by 'interests' and the 'authoritative will of the state.' Weber (1968[1920]) defines law as consisting 'essentially in a consistent system of abstract rules' and treats the 'legal order' and the 'administrative system' (i.e., bureaucracy) as 'vital' to the rise and functioning of 'rational capitalism.'

Similarly, contemporary sociologists suggest that modern polity and society tends to create a 'subsystem' specializing in generating and enacting 'collectively binding decisions' in the form of an 'administrative state' (Habermas, 2001).

In sociological systems theory, types of political systems are contained or implied as particular cases in those of social system as general categories. Comte's (1983) typology of historical social systems contains that of political systems into the old 'theological-military' polity, including 'theocratic government,' as the 'conservative system' exerting 'retrograde tyranny,' and the new 'positive polity' as the liberal-democratic, with the intermediate case of the 'metaphysical polity' characterized with 'revolutionary anarchy.' In accordance with the 'fundamental law of social evolution,' Comte (1983) observes that the 'natural progress' of society (plus intelligence) tends to the 'complete elimination' of the theological-military and theocratic regime and predicts that the latter will 'never' again function as a 'basis for social order' instead based on the 'positive polity' that, while embracing 'all the essential aspects of the present state of civilization,' overcomes both 'retrograde tyranny' and 'revolutionary anarchy.' Mannheim (1956) divides political systems and ideologies into the non- or 'pre-emocratic authoritarian' system and mind and the modern 'democratic system,' specifically 'representative democracy' characterizing the Western 'governmental system,' and Dahrendorf (1959) describes the latter as the 'system of representation,' including 'electoral systems' and 'elaborate system of conflict regulation in the political sphere' as well as economy. Some contemporary sociologists (Habermas, 1973) retain the pre- or nondemocracy and democracy dichotomy suggesting that the 'positive value of the system of authority' can be realized only in the latter, namely, through the freedom of the 'polis' in the generic sense of a liberal-democratic political system.

As in the case of economy, sociological systems theory treats the relations between polity and the social system as those between a subsystem and a general system. Comte (1983) and Durkheim (1965a[1893]) consider polity, in particular the state, one of the integral elements (alongside economy) of the social system, thus a 'political society' (and 'economic society'). Marx and Engles (1947[1847]) treats the state, including systems of law, as the 'active, conscious, and official expression' of the 'structure of society,' above all its 'economic structure' constituted by the 'totality' of relations of production on the assumption that the latter 'in all ages' provides the 'basis of the State and the rest of the idealist superstructure.' In particular, he contends that the state represents the expression of class structure in that it arises and operates as the 'form' for asserting the 'common interests' of a 'ruling class' against other classes within the social system (Marx and Engles, 1947[1847]). Also, Pareto (1963[1916]) emphasizes that the 'system of government' in modern and premodern societies tends to be intimately linked with the 'character of governing class,' while both political and stratification systems being in a 'relationship of interdependence with all other social phenomena' within the 'sociological system.'

Contemporary systems theory continues treating polity or state as the subsystem of the 'whole social system' (Luhmann, 1995), while distinguishing the first from economy. According to such a distinction, the functional subsystem of politics shows 'no exact isomorphy' with that of economy for the reason that the 'communication medium of power' lacks the 'technical provision or integrative capacity of money' or wealth (Luhmann, 1995). In this account, while the 'use of power' not being strictly a 'political phenomenon' but extending in various ways and degrees into the whole social system, the 'concept of the state' primarily fulfills such a 'function' (Luhmann, 1995) through what Weber (1968[1920]) calls its 'monopoly' on the legitimate exercise of domination or force (Habermas, 2001; Smelser, 1997).

The Cultural System and the Social System

Culture also represents a structural element, subsystem of the social system, and in that sense a cultural system. Sociological systems theory generally characterizes culture as a system of ideas and thought arising and operating in and conditioned by the total social system. Saint-Simon (1964[1816]) registers 'systems of religions, politics, ethics, and education (and other) applications of the system of ideas' as essential characteristics and ingredients of culture also characterized by Comte (1983) as a 'system of ideas and conceptions,' including the 'educational system.' Similarly, Mannheim (1936) treats 'systems of ideas,' notably 'divergent' ideological or thought systems, including 'closed static systems of thought,' as fundamental cultural phenomena and the main subject matter of the sociology of culture and knowledge.

Relatedly, sociological systems theory regards culture as a system of values, ideals, and norms or rules. For Durkheim (1974[1924]), 'principal social phenomena,' notably culture, including religion, morality, and aesthetics, and even law and economy, constitute 'systems of values and hence ideals,' in particular a 'whole system of moral forces.' Also, Weber (1968 [1920]) points to a 'system of values' (e.g., an 'individualistic system of values') and a consequent 'system of norms and duties' with a binding force in a social system. Following Durkheim and Weber, Parsons (1937) adduces the 'existence of a common value system' and a resulting 'common system of normative rules' as the fundamental trait and element of the 'cultural system,' also composed of 'symbolic systems' premised on a 'cultural tradition' and values. Parsons (1951) particularly stresses the 'special relevance to the social system' of 'moral standards' thereby considered to be of the 'greatest direct importance' among 'value orientations.'

In sociological systems theory since Comte, Spencer, and Durkheim a particularly manifest and salient element or form of the cultural system is considered religion as a belief and value system, as Weber also emphasizes. Comte (1983) and Spencer (1969[1876–96]) analyze religious systems or institutions as the primary and original elements or forms of the cultural system, above all at its theological-military historical stage, while observing and predicting their ultimate demise and replacement by positive science due to scientific rationalism and progress and industrialization leading to their differentiation from political and other systems or institutions within the social system and so secularization. Elaborating on Comte and Spencer, Durkheim (1965b[1915]) analyzes the 'most primitive' religious system as that existing in a society with the least complex organization, suggesting that the 'first systems of representations' have 'religious origin.' Durkheim

(1965b[1915]) defines religion as 'a more or less complex system of myths, dogmas, rites, and ceremonies' with a certain unity, i.e., a 'system of cults' and a 'unified system of beliefs and practices' in relation to things, and emphasizing the 'immediate link between religious beliefs and the organization of society.' In turn, Weber (1968[1920]) focuses on analyzing comparative religious systems such as the 'great world religions' in interaction and interpenetration with the other elements of the social system, notably economy and polity. Like Durkheim and Weber, contemporary sociologists characterize religious and related belief systems as the 'idealized system of ends' (Alexander, 1998; Martin, 2002).

Another related pertinent element or form of the cultural system has considered morality as a system of ethical values and norms of human behavior in society. Durkheim (1974[1924]) defines morality as a 'system of rules of conduct' that is conditioned by and arises within the 'total (social) system' to the effect that changes in social structure necessitate a 'change in morals (as) a system of precepts.' Following Durkheim, Parsons (1951) considers 'moral standards' to be that dimension of 'value orientation' and by implication culture overall that is the 'greatest direct importance' because of their 'special relevance to the social system.'

Other elements or forms of the cultural system include systems of classification, symbolic systems, ideological systems, scientific-educational systems, etc. In particular, Durkheim and Mauss (1903) define classifications of categories as 'systems of hierarchized notions' ordered in 'fixed relations,' and grounded in and representing the 'classification of men' within the social system, so that the 'entire system of concepts' expresses the nature and structure of society and how it views the 'world,' thus 'logical hierarchy' being 'only another aspect of social hierarchy,' i.e., 'dominating and subordinated' groups. Similarly, Mannheim (1936) focuses on 'ideological systems' and their basis in and determination by the social system, especially 'collective action' and 'class stratification.' Contemporary sociologists also emphasize 'systems of classification' and 'symbolic systems' regarded as being 'always determined in several respects at once' (polysemy) by the social system, above 'material conditions' (Bourdieu, 1988), particularly stressing the 'prevailing symbolic system' and its social-systemic determinants and effects (Habermas, 1971). Likewise, systems theorists since Comte regard the educational system as the relevant component of the cultural system as well as that of social stratification (Dahrendorf, 1959; Bourdieu, 1988), and suggest that science and generally knowledge provides the 'solution of system problems in general' rather than the 'gratification of immediately empirical needs' (Habermas, 1971).

Like economic and political systems, types of cultural system generally correspond to and are implied in those of social system, with the classifications of the latter implying those of the former as the subsystem. Tönnies (1955[1887]) observes that in the modern social system the 'entire culture' has become a cultural system and 'civilization of state and of capitalism' premised on a moral 'system of conventional rules' in sharp contrast to morality as a 'purely ideal or mental system of norms for community life.' Simmel (1964[1908]) distinguishes a 'highly differentiated scientific culture,' 'intellectual culture,' and 'economic culture' as particular cultural systems or elements mostly within the modern social system.

Sociological systems theory typically considers the relations between the cultural system and the social system as those between a subsystem and a general system, although sometimes using the two terms interchangeably or as a composite term, e.g., 'sociocultural systems' (Sorokin, 1957), or even treating 'culture' as apart from and beyond the social system in action terms (Parsons, 1937). The general theoretical premise or empirical generalization is that only the social system either as a human society or group can and does create, have, and sustain culture as a system of ideas, values, norms, and symbols, including religious, moral, classification, symbolic, and scientific-educational systems, as especially Durkheimian sociology emphasizes (also, Mannheim, 1936; Merton, 1968; Sorokin, 1957).

In particular, system theorists analyze the relations of the cultural system to such other partial social systems as economy, stratification, polity, etc. One approach treats culture as mostly determined by the economic and class systems, and thus as the dependent variable. This is exemplified by Marx's (1970[1859]) proposition that the 'economic structure of society' determines 'cultural development,' in particular the class being the 'ruling material force of society' tends to be simultaneously its 'ruling intellectual force.' Also, contemporary conflict system theorists emphasize 'the objective relations between the system of economic production' and the 'system producing the producers,' including schools and family (Bourdieu, 1988). In this relationship 'material conditions' are the structuring and these cultural elements (or habitus) a 'structured' variable, and constituting a 'system of mechanisms' for the near-automatic 'reproduction of the established order by its own motion,' thus allowing the dominant class to assert its domination by simply letting the 'system' operate by 'its own course' (Bourdieu, 1988). In short, the outcome of the impact of economic and class structure on culture is a 'system of class culture' (Collins, 1975).

Another alternative approach instead regards culture (also) as a determinant of economic and other systems within the social system, and so an explanatory variable. This is exemplified by Simmel's (1990[1900]) assumption that money forms the 'pinnacle of a cultural historical series of developments' determining its 'direction' and thus that of corollary economic phenomena (commercial exchange, value, price, market, etc.). Notably, this approach is epitomized by Weber's (1968[1920], 1976[1905]) famous analysis of the influence of religion on economy, in particular of the world religions (above all, ascetic Protestantism) on the rise and development of capitalism.

The Civil System and the Social System

Yet another structural element, i.e., subsystem of the social system consists of civil society or societal community, also termed the (mostly) individual, private, and autonomous sphere of life, the 'lifeworld' (Alexander, 1998; Giddens, 1984; Habermas, 2001). Early sociological and philosophical definitions especially define and analyze civil society in counter-distinction and in part disjuncture from polity, specifically state as the public realm, but in association and even partial equation with economy, specifically private commercial and

other economic activity. Following Hegel, Marx and Engles (1947[1847]) defines civil society as the 'form of intercourse,' specifically the 'whole material intercourse of individuals' driven by 'private interest' and 'determined' as well as 'determining' the forces of production at a 'definite stage' of their development – i.e., only arising with 'bourgeoisie' – thus the 'social organization' developing from 'production and commerce,' so its 'anatomy' lies in 'political economy' as the 'basis of the modern State.' Following Marx, Tönnies (1955 [1887]) suggests that the basis of Civil Society (in the sense of modern bourgeois capitalism) is 'general commercial exchange.' Simmel (1964[1908]) conceptualizes civil society largely in social, particularly moral, terms, as the sphere of sociability emerging as a 'very peculiar sociological structure.' This anticipates the contemporary prevalent conception of civil society, and envisions that its 'minimum of ethical, peaceful behavior' as the necessary condition of its existence is not equal to but transcends the 'criminal code' (Simmel, 1964[1908]).

Most contemporary conceptions consider civil society a distinct social subsystem not only in relation to polity/state but also to economy/market, i.e., both a noneconomic and a nonpolitical sphere of life, thus qualifying Marx and following Simmel. Parsons (1967) treats civil society, also termed 'societal community,' as the 'integrative system' that is functionally differentiated from (although in interaction with) economy and polity, as well as culture, within the total social system (also, Fararo, 1989; Munch, 1990), particularly, as Merton (1968) also suggests, generating or affecting the 'motivation of individuals involved in a social system.' In this context, Parsons' societal community is regarded as introducing a 'fourth' subsystem or sphere within the social system indicating a 'world' featuring the 'subjectivity of the lifeworld,' joined with the 'abstraction of more highly rationalized systemic spheres' (Alexander, 1998). Other contemporary sociologists conceive civil society or the lifeworld as involving the 'normative structures (values and institutions) of society' (Habermas, 1975) and providing individuals with a 'shared store of cultural knowledge, socialization patterns, values, and norms' (Habermas, 2001), thus being almost equated or most closely related to culture within the social system. In sum, civil society is commonly defined as a 'private sphere' of life in separation from and in tension with – although historically the creation of and with origins in – the 'public sphere of the state' (Giddens, 1984), i.e., 'an informal social order' mostly nondominated by 'large-scale coercive' political as well as economic structures and instead creates and sustains through 'various forms of communication and reciprocity' (Alexander, 1998).

Consequently, civil society is considered to essentially differ from, even if related to, other social systems, especially economy and polity. In this view, the 'rationalization of the lifeworld' in modern society differs from 'rationalizing economic or administrative' systems and activities, instead encompassing 'cultural traditions, socialization of individuals and social integration' in contrast to economy and politics as 'functionally integrated' and 'self-regulating systems' regulated by 'money and administrative power' (Habermas, 2001). Moreover, it is argued that the 'balance' within the social system becomes negative when economic-political systems trespass into the 'lifeworld' in the sense that the 'invasion' of the latter by 'exchange relations and bureaucratic regulation' results in various 'social pathologies' (Habermas, 2001). In short, in the modern social system a strong civil society is deemed a necessary condition for 'effective democratic government' as well as a 'well-functioning market system' (Giddens, 2000). The preceding structural elements or subsystems of the social system are summarized in **Table 1**.

Table 1 Elements/subsystems of the social system

Economy	Polity
Division of labor, wealth, market	Public ends, legitimate power, law
Culture	Civil society
Ideas, values, norms	Private sphere, lifeworld

See also: Differentiation: Social; Economic Sociology; Institutions; Integration, Social; Societies, Types of; Sociological Theory.

Bibliography

Alexander, J., 1983. Theoretical Logic in Sociology. University of California Press, Berkeley.

Alexander, J., 1998. Neofunctionalism and after. Blackwell Publishers, Oxford.

Arrow, K., 1994. Methodological individualism and social knowledge. American Economic Review 84 (9).

Bailey, K., 1994. Sociology and the New Systems Theory. State University of New York Press, Albany.

Barber, B., 1993. Constructing the Social System. Transaction Publishers, New Brunswick.

Berrien, K., 1968. General and Social Systems. Rutgers University Press, New Brunswick.

Bertalanffy, L., 1975. General Systems Theory. George Braziller, New York.

Bourdieu, P., 1988. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Buckley, W., 1967. Sociology and Modern Systems Theory. Prentice-Hall, Englewood Cliffs.

Collins, R., 1975. Conflict Sociology. Academic Press, New York.

Collins, R., 1988. Theoretical Sociology. Harcourt Brace Jovanovich, San Diego.

Comte, A., 1983[1830–52]. Auguste Comte and Positivism. University of Chicago Press, Chicago.

Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford.

Dahrendorf, R., 1968. Essays in the Theory of Society. Stanford University Press, Stanford.

Durkheim, E., 1965a[1893]. The Division of Labor in Society. The Free Press, New York.

Durkheim, E., 1965b[1915]. The Elementary Forms of the Religious Life. The Free Press, New York.

Durkheim, E., 1966[1897]. Suicide. Free Press, New York.

Durkheim, E., 1974[1924]. Sociology and Philosophy. The Free Press, New York.

Durkheim, E., 1982[1895]. The Rules of Sociological Method. The Free Press, New York.

Durkheim, E., 1994[1909]. Emile Durkheim on Institutional Analysis. University of Chicago Press, Chicago.

Durkheim, É., Mauss, M., 1903. De quelques formes de classification - contribution à l'étude des représentations collectives. Année Sociologique 6, 3–42.

Fararo, T., 1989. The Meaning of General Theoretical Sociology. Cambridge University Press, Cambridge.

Fararo, T., 2001. Social Action Systems. Praeger, Westport.

Giddens, A., 1979. Central Problems in Social Theory. University of California Press, Berkeley.

Giddens, A., 1984. The Constitution of Society. University of California Press, Berkeley.

Giddens, A., 2000. The Third Way. Blackwell Publishers, Malden.

Habermas, J., 1971. Knowledge and Human Interests. Beacon Press, Boston.

Habermas, J., 1973. Theory and Practice. Beacon Press, Boston.
Habermas, J., 1975. Legitimation Crisis. Beacon Press, Boston.
Habermas, J., 2001. The Postnational Constellation. MIT Press, Cambridge.
Lenski, G., 1994. Societal taxonomies: mapping the social universe. Annual Review of Sociology 20, 1–26.
Luhmann, N., 1982. The Differentiation of Society. Columbia University Press, New York.
Luhmann, N., 1995. Social Systems. Stanford University Press, Stanford.
Mannheim, K., 1936. Ideology and Utopia. Harcourt, Brace & World, New York.
Mannheim, K., 1956. Essays on the Sociology of Culture. Routledge & Kegan Paul, London.
Martin, J., 2002. Power, authority, and the constraint of belief systems. American Journal of Sociology 107, 861–904.
Marx, K., 1933[1849]. Wage-Labour and Capital. International Publishers, New York.
Marx, K., 1967[1867]. Capital. International Publishers, New York.
Marx, K., 1970[1859]. A Contribution to the Critique of Political Economy. International Publishers, New York.
Marx, K., Engels, F., 1947[1847]. The German Ideology. International Publishers, New York.
Merton, R., 1968. Social Theory and Social Structure. The Free Press, New York.
Munch, R., 1990. Differentiation, rationalization, interpenetration–the emergence of modern society. In: Alexander, J., Colomy, P. (Eds.), Differentiation Theory and Social Change: Comparative and Historical Perspectives. Columbia University Press, New York, pp. 441–464.
Munch, R., 1994. Sociological Theory. Nelson-Hall Publishers, Chicago.
Pareto, V., 1963[1916]. The Mind and Society. Dover Publications, New York.
Pareto, V., 2000[1901]. The Rise and Fall of Elites. Transaction Publishers, New Brunswick.
Parsons, T., 1937. The Structure of Social Action. The Free Press, New York.
Parsons, T., 1951. The Social System. The Free Press, New York.
Parsons, T., 1966. Societies. Prentice Hall, Englewood Cliffs.
Parsons, T., 1967. Sociological Theory and Modern Society. The Free Press, New York.
Parsons, T., Smelser, N., 1956. Economy and Society. The Free Press, New York.
Saint-Simon, H., 1964[1816]. Social Organization, the Science of Man and Other Writings. Harper Torchbooks, New York.
Samuelson, P., 1983[1948]. Foundations of Economic Analysis. Harvard University Press, Cambridge.
Schumpeter, J., 1949[1911]. The Theory of Economic Development. Harvard University Press, Cambridge.
Simmel, G., 1955[1923]. Conflict. The Web of Group Affiliations. The Free Press, New York.
Simmel, G., 1964[1908]. The Sociology of Georg Simmel. The Free Press, New York.
Simmel, G., 1990[1900]. The Philosophy of Money. Routledge, London.
Sorokin, P., 1957. Social and Cultural Dynamics. Porter Sargent Publisher, Boston.
Smelser, N., 1997. Problematics of Sociology. University of California Press, Berkeley.
Smith, A., 1976[1776]. An Inquiry into the Nature and Causes of the Wealth of Nations. Clarendon Press, Oxford.
Spencer, H., 1969[1876–96]. Principles of Sociology. Archon Books, London.
Tönnies, F., 1955[1887]. Community and Association. Routledge and Kegan Paul, London.
Veblen, T., 1898. The instinct of workmanship and the irksomeness of labor. American Journal of Sociology 4, 187–201.
Weber, M., 1968[1920]. Economy and Society. Bedminster Press, New York.
Weber, M., 1976[1905]. The Protestant Ethic and the Spirit of Capitalism. Charles Scribner's Sons, New York.

Social Work and Disasters

Michael Cronin, Monmouth University, West Long Branch, NJ, USA
David N Jones, People Need People Consulting, Northampton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There has been an increase in the incidence and impact of both natural and human-induced disasters. Poverty always determines the relative severity of impact. Social workers' skills and experience are very relevant for disaster work, including concern for human rights, protection of children and vulnerable people, and social inclusion of those affected. Psychosocial and community recovery interventions are needed. The needs of workers are often similar to those they are supporting and must be addressed. Cultural competence must inform program delivery.

Introduction

Social workers' skills and experience are highly relevant for multidisciplinary disaster work. Of particular significance are their understanding of human behavior, emphasis on 'person-in-environment,' experience in community work and program planning, and recognition of the value of advocacy and social justice. With the apparent increase in and awareness of natural, technological, and human-induced disasters, the need for all social workers and other helping professionals to become aware of this work is increasing (Cronin and Ryan, 2010). The world social work community has recognized disaster response as a priority area for social work education and practice (International Federation of Social Workers et al., 2012).

What Are Disasters?

The term 'disaster' is variously defined by authors from different disciplines and humanitarian communities. The definitions all have political, ideological, cultural, and other biases. The more research is undertaken, the more definitions are devised. The United Nations defines disaster as "a serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources" (UN, 2009: p. 9).

Disasters are often described as resulting from (1) exposure to a hazard, (2) preexisting conditions of vulnerability, and (3) insufficient capacity or measures to reduce or cope with the potential negative consequences (UNISDR, 2009: p. 9).

$$\text{Disaster} = \text{hazard} \times \text{exposure} \times \frac{\text{vulnerability}}{\text{manageability}}$$

(Wisner, 2004).

The definition implies a sudden, overwhelming, and unforeseen event. At the community level, it can be a flood, fire, earthquake, loss of incomes, disease epidemic, or displacement through conflict. At the family level, a disaster could result in a major illness, death, economic, or social hardship. When occurring across a region, a large number of people can be affected.

Disasters can be categorized as either natural or human-induced occurrences. Natural disasters include weather phenomena (such as tropical storms, extreme cold or heat, high winds, drought), floods, avalanches, earthquakes, tsunamis, wild fires, landslides, and volcanic eruptions. Human-induced disasters include transportation incidents; industrial accidents; release of hazardous materials; building collapses; and extreme events such as famine, political conflict, and climate change. However, this distinction can be less clear-cut than it seems, with the severity of the impact of the disaster frequently contingent on human factors (e.g., socioeconomic status).

The Disaster Cycle

The characteristics of disaster work are often presented as a four-stage cycle: mitigation, preparedness, response, and recovery (National Governor's Association, 1979). Each phase requires distinct tools, strategies, and resources. Social work can be involved at each stage by contributing to strategy and policy development as well as by direct work with individuals and local communities.

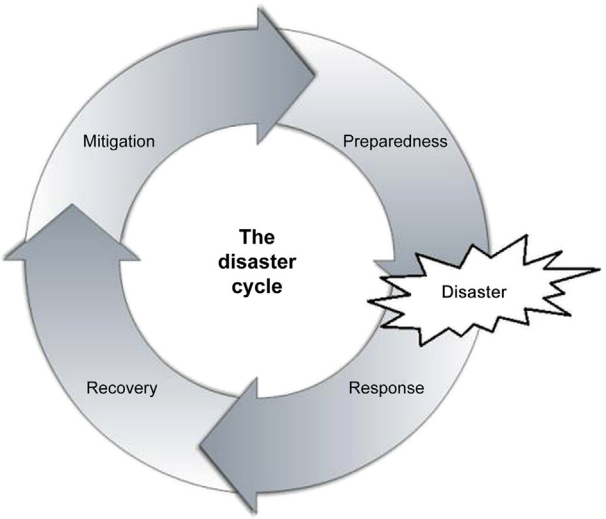

Source: unknown.

Mitigation: 'the lessening or limitation of the adverse impacts of hazards and related disasters.' Mitigation measures encompass engineering techniques and hazard-resistant construction as well as improved environmental policies and public awareness (UNISDR, 2009: p. 19). Mitigation involves steps to reduce vulnerability to disaster impacts, such as injuries and loss of life and property, and other efforts to make the community more resilient to catastrophic event, but the avoidance of emotional trauma and community disruption are also significant consequences.

Preparedness: "the knowledge and capacities developed by governments, professional response and recovery organizations, communities and individuals to anticipate, respond to and recover from the impacts of likely, imminent, or current hazard events or conditions" (UNISDR, 2009: p. 21). Preparedness focuses on understanding how a disaster might affect the community and how education, outreach, and training can build capacity to respond to and recover from a disaster. This may include engaging the business community and community groups and predisaster strategic planning.

Response: 'the provision of emergency services and public assistance during or immediately after a disaster to save lives, reduce health impacts, ensure public safety, and meet the basic subsistence needs of the people affected.' Disaster response or 'disaster relief' is predominantly focused on immediate and short-term needs (UNISDR, 2009: p. 24). Response addresses the immediate threats presented by the disaster, including saving lives, meeting humanitarian needs (food, shelter, clothing, public health, and safety), cleanup, damage assessment, and the start of resource distribution. As the response period progresses, focus shifts from immediate emergency issues to conducting repairs, restoring utilities, establishing operations for public services, and finishing the cleanup.

Disasters frequently reveal and exacerbate weak economic and social infrastructures and development, with significant gaps in organizational capacity, staff, and resources. For effective response and recovery, economic development agencies and stakeholders may need additional staff, capacity-building assistance, and social stability. This is illustrated in the contrast between the impact of the earthquakes in Chile and Haiti, the former making a rapid recovery thanks to effective planning and developed public services, while the latter had very weak social infrastructures and ineffective planning and is still struggling after several years of international aid (Arbour et al., 2011).

Recovery: 'the restoration and improvement of facilities, livelihoods, and living conditions of disaster-affected communities, including efforts to reduce disaster risk factors.' The recovery task of rehabilitation and reconstruction begins soon after the initial response phase, and is ideally based on preexisting strategies and policies that facilitate institutional responsibilities for recovery action and enable public participation (UNISDR, 2009: p. 23).

The recovery phase of disaster can be broken into two periods. The short-term phase typically lasts from 6 to 12 months and involves delivering immediate services. The long-term phase, which can be decades long, requires thoughtful strategic planning and action to address more serious or permanent impacts of a disaster. Investment in social and economic development capacity building becomes essential to foster economic diversification, attain new resources, build new partnerships, and implement effective recovery strategies and tactics. Communities must access and deploy a range of public and private resources to enable long-term recovery.

Human Impact of Disasters

Demographics

There is a strong interrelation between the severity of impact of disasters and poverty. Poorer people live in the most drought, flood, and extreme-weather prone regions and suffer most when catastrophe occurs (UNDP, 2012). Ninety-five percent of disaster fatalities occur in developing countries; the likelihood of death in a disaster is four times greater for people in lowest-income countries than those living in high-income countries (World Bank, 2001: p. 171). Disasters put hard-won development achievements at risk, reversing economic growth and progress toward the elimination of poverty, cause environmental damage, and result in human suffering. Disasters have killed more than 1.3 million people and affected an average of 220 million per year over the past two decades. In 2011, 106 million people were affected by floods, 60 million were affected by drought, mainly in the Horn of Africa, and almost 30 000 were killed. Disasters cost the world economy more than $380 billion in 2001 (UNDP, 2012).

Disasters have a differential impact on men and women, reflecting the distinct roles of the two sexes in political, economic, and social life (Pittaway et al., 2007; Zetter, 2012). In countries where gender discrimination is tolerated, women and girls are particularly vulnerable. Not only is the percentage of women and girls who die higher in these counties, but the incidence of gender-based violence – including rape, human trafficking, and domestic abuse – also increases exponentially during and after disasters. Most disasters place an undue burden on women and girls who are responsible for unpaid work such as providing care, water, and food for households (Zetter, 2012: p. 13). In the case of Hurricane Katrina in New Orleans, USA, most of those trapped were African-American women and their children, the poorest demographic group in that region (United Nations Development Program, 2010). Similarly, 32% of Nicaraguan farming households headed by females did not plant the following year, compared to 23% of male-headed farming households (Bradshaw, 2004: p. 26).

Social Implications of Disasters

On a long-term average, 184 deaths per day are recorded worldwide due to natural hazards. Most of these occur in countries that rank low on the Human Development Index (UNDP, 2011: p. 10). For every person killed, around 3000 more are affected (UNDP, 2011: p. 14) and become severely impoverished through the destruction of personal assets, damage to health and education infrastructure, and loss of

drinking water and sanitation. If the disaster is compounded by casualties from HIV/AIDS or armed conflict, social and economic structures can be destabilized. Natural disasters in poorer countries frequently result in major migrations or population displacement with far-reaching social and political consequences.

'Environmental refugee' became a more frequently used term in the 1990s, referring to people forced to leave homes due to natural disasters or environmental reasons. However, environmental problems never appear in isolation; their interconnection with other emergencies resulting from social or political factors results in the great number of refugees (Zetter, 2012: p. 20). The United Nations Environment Program estimates that there are 22–24 million environmental refugees (Biermann, 2001: p. 24). Most of these migrants move between developing countries. For example, each year, thousands of inhabitants of Bangladesh go to India due to heavy flooding (p. 24), often starting a new life in surroundings that are completely unprepared to accommodate them or are inappropriate for settlement. Similarly, the large number of environmental refugees in the Horn of Africa is not primarily a consequence of drought, which affected 60 million, but due to weak governments, political and military conflicts between competing factions, the destruction of traditional coping mechanisms, and inadequate infrastructures, thus resulting in serious political destabilization across the region (Zetter, 2012: pp. 21–23).

The political, social, and economic challenges facing host countries in coping with these refugees are often unmanageable (Zetter, 2012: pp. 19–21). Extreme poverty can result from disaster-related migration because the refugees lack every kind of social service, including health care, schools, and access to safe drinking water. Traditional family structures, once an important form of support during crisis, are destroyed. Groups of citizens that previously were not considered poor can fall into poverty due to forced migration and the political management of refugees.

Economic Implications of Disasters

Over the last 50 years, the global cost of natural disasters has increased 14-fold, with weather-related and natural disasters accounting for two-thirds of all losses (Guha-Sapir et al., 2004: p. 13). Annual economic losses due to natural disasters in the last decade reached an estimated $55 billion, with a high percentage of losses in infrastructure (German Federal Ministry for Economic Co-operation and Development (BMZ), 2001). While absolute levels of economic loss are greater in developed countries due to the far higher density and cost of infrastructure and production levels, less-developed countries suffer higher levels of relative loss when seen as the proportion of gross domestic product (UNDP, 2004: p. 20).

Developing countries are not only facing the potential of more frequent weather-related events. Rapid urbanization, the increased concentration of assets, environmental degradation, and other factors are also increasing the vulnerability of poor communities to disaster impacts. By 2025, roughly two-third of the world's inhabitants and the vast majority of wealth will be concentrated in urban centers. Many of the world's megacities (populations exceeding 10 million) are already situated in locations prone to major earthquakes and severe droughts, and along flood-prone coastlines, where the impacts of extreme climatic events and sea-level rises pose a greater risk of disasters. Today, more than half the global population resides in urban areas. Especially vulnerable to natural disasters are megacities like Jakarta, Istanbul, Buenos Aires, Athens, Saõ Paulo, and Mexico City (UNISDR, 2012). Urbanization in relatively smaller cities also increases risk – particularly in regions where infrastructures and institutions are ill equipped to cope with disasters. The vulnerability of this new generation of urbanites will become a defining theme of disaster risk in the coming decades (Zetter, 2012: p. 113).

Economic implications of disasters are conventionally categorized as direct, indirect, and secondary losses. Direct losses include loss of life and physical assets, damage to infrastructure, and loss of inventory and agricultural produce. Indirect losses include those caused by disruption of the flow of goods and services, like reduced opportunities for generating income. Examples of indirect costs include loss of earnings; unemployment; loss of productivity due to death, illness, and injuries; and increased public finance expenditure (Guha-Sapir et al., 2004: pp. 38–45).

The final category includes impact on macroeconomic variables such as economic growth, state budget levels, national indebtedness, and inflation. The secondary effects of disasters are important because these can have significant impact on long-term human economic development and social stability (International Debates, 2005).

Political Implications of Disasters

Disasters can evoke severe governance problems. The processes through which citizens articulate their interests and exercise their legal rights – an important aspect of the political dimension of disaster – can deteriorate, especially where political rights are poorly recognized. Inquiry into the economic aspect of governance reveals that financial decisions made during disaster and the reconstruction period have major implications for poverty and quality of life. Under time and financial pressures, politicians are likely to abandon carefully developed budget plans and overturn long-term development programs to release resources for immediate emergency aid. This money is often used for the recovery of one group at the expense of others. Economically poor groups with little access to decision making are likely to be left out. Gender and other fundamental social issues are often disregarded. The result can be the reaffirmation of existing inequalities (UNDP, 2004: p. 21).

Social Work and Disasters

Social work in the disaster process draws on methods and practice most of which are common to other members of the multidisciplinary team. These are described below. However, social work also contributes some distinct perspectives which, while not unique to social work, do have a particular emphasis arising from its core professional functions (Dominelli, 2010).

Concern for Human Rights

The response phase of disaster response is usually focused on meeting fundamental needs for safety, security, food, and shelter. In extreme situations the protection of life is the main concern. This can lead to expedient decisions under extreme pressure, which appear not to allow for consideration of other human rights, e.g., the right to family life, freedom from exploitation, and participation. Study of specific examples suggests that a lack of attention to human rights can result in unfair distribution of aid, exploitation of vulnerable people, and placing children and others at further risk (Klasing et al., 2011). Social workers have particular experience of situations where vulnerable people are at risk of losing their rights and of mediating in family and community situations where there are competing human rights (Banks and Nøhr, 2011: Case 2.1). These skills are important both in the immediate response and as a foundation for recovery (Zetter, 2012: pp. 47–71).

Populations at Risk – Protection of Children and Vulnerable People

Social work training and experience includes a focus on those who are most vulnerable and/or in need of protection. Social workers offer particular perspectives on disaster response involving vulnerable groups. Within a community affected by disaster, people can be seen as suffering primary, secondary, or nondirect impacts. The primary group comprises those who have experienced direct physical, material, and personal losses from the disaster. The secondary group involves those who have witnessed the destruction of the disaster, but have not directly experienced the specific impact. The nondirect category refers to those who were not directly affected by the disaster, but who suffer distress and uncertainty over the safety and well-being of family and friends.

Among the categories of those affected by disaster are vulnerable populations of concern: poor people; older persons; persons with disabilities; people who are isolated, institutionalized, or otherwise at risk; and all exposed children. These populations may require special attention during preparedness, immediate relief, and recovery activities (Congress et al., 2010; Dominelli, 2012; Smith et al., 2012).

The risks faced by those in disaster situations have been recognized in operational guidelines published by the United Nations (Inter-Agency Standing Committee, 2007). Risks can include the trafficking of vulnerable women and orphaned children on the pretext of 'rescuing' them, thereby resulting in sexual exploitation, virtual or actual slavery, or illegal adoption. Well-meaning, short-term 'rescue' activities can also cause long-term damage, for example, by misinterpreting the significance of symptoms of ill-health or removing children without parental knowledge or consent or preventing the reunification of families (Zetter, 2012: pp. 68–70). Social workers bring experience of handling such dilemmas.

Values and Ethics – Social Inclusion

Humanitarian workers share a humanitarian ethos or value system, based on respecting individual integrity, recognizing community order, and promoting human rights (e.g., International Federation of Red Cross and Red Crescent Societies, 1965; International Federation of Social Workers and International Association of Schools of Social Work, 2004). Social work places emphasis on self-determination and working 'with' people (rather than doing things to them). This approach encourages the active involvement ('social inclusion') of those involved in planning and deciding actions, something which UN agencies recognize has frequently been absent in the response to major disasters (Zervaas, 2007).

Psychosocial Interventions

Psychosocial interventions are immediate and delivered by a range of professionals, including social workers, and other mental health professionals. Community referrals are made for those with long-term needs. Interventions are comparable in small or large disasters. The micro practice interventions fall into the following categories: psychological first aid (PFA), crisis intervention, assessment, casualty support, advocacy, outreach, and staff support (Cronin and Ryan, 2010).

Psychosocial and mental health support can be both preventative, decreasing the risk of developing mental health issues, and therapeutic, helping individuals and communities to overcome and deal with psychosocial problems resulting from the disaster.

Psychological First Aid

PFA is the cornerstone of support. The concept of PFA for individuals exposed to highly traumatic events has been used in the field of crisis and disaster management for years (Reyes, 2006). Similar to medical first aid, PFA involves providing immediate care, by attending to the biopsychosocial needs of those affected. It is practical and proactive and should not be confused with psychological assessment or treatment, only provided by trained mental health professionals.

PFA includes provision of basic care, comfort, and support during and in the immediate aftermath of disasters. It is designed for delivery by mental health and other disaster response workers who provide early assistance to affected communities.

PFA includes four broad goals: (1) to relieve physical and emotional suffering, (2) to improve short-term functioning, (3) to help course of recovery, and (4) to provide linkage to critical resources (Halpern and Tramontin, 2007: pp. 203–204). PFA helps stabilize persons in distress by ensuring safety and providing assistance in meeting simple needs.

Helping people to notify family members, making arrangements for child care, and contacting social service agencies for needed assistance can help stabilize individuals in crisis. Validating thoughts and feelings regarding their experience and promoting efficacy can support steps toward healing. Aiding clients in connecting to their natural support systems can be a critical intervention. For those without any supports, disaster mental health workers will either stay with the client or check in often to establish a sense of connection so the client does not feel isolated.

Clients are often desperate for information from authorities. They often ask questions such as when they can

return, whether a loved one has been found or identified, or how to recover a pet that was left behind. The period of waiting for information can be agonizing, and although disaster mental health workers do not directly provide information that law enforcement and governmental agencies are responsible for, they do advocate for official briefing updates for clients.

Crisis Intervention

Crisis intervention is the provision of emergency psychological care to victims to assist them in returning to an adaptive level of functioning and to prevent or mitigate the potential negative impact of psychological trauma (Everly and Mitchell, 1999). Crisis intervention procedures have evolved from studies of grieving conducted by Erich Lindemann (1994) in the aftermath of a nightclub fire.

An individual is perceived to be in crisis if he or she has experienced a meaningful and threatening event; if he or she has increased fear, tension, or confusion; and if usual coping mechanisms are not effective in alleviating the stress symptoms. Crisis intervention aims to moderate the emotional state and help the individual find adaptive solutions (Roberts, 2000). This is best exemplified when a worker helps a client to process and manage emotionality that is interfering with being able to engage in disaster recovery planning and activity. Crisis intervention techniques involve the quick establishment of rapport, functional assessment, psychoeducation on stress symptoms and coping strategies, problem solving, the development of an action plan, and referral for follow-up mental health treatment as needed.

Assessment

Assessments play a critical role in determining what psychosocial or mental health interventions are needed. Before planning can begin, it is necessary to find out what has happened and how people have been affected. The goal of screening and assessment is to evaluate mental health functioning and identify the appropriate intervention or referral. The task is the usual informed assessment and decision making that one would make with any person – watching for symptoms, intervening as appropriate, and dealing with the most severe cases first (Cronin and Ryan, 2010). When large numbers of individuals are affected and only a limited number of responders are available, the ability to perform rapid triage becomes crucial. The greater the degree of exposure to the disaster or reaction intensity, the higher the priority for immediate disaster mental health contact.

For the majority of those affected by disaster, early assessment and intervention will be unnecessary. Studies have shown that, even in extreme disasters, the majority of people do not become incapable of functioning and usually make a full psychosocial recovery. Resilience and coping are probably the most common reactions seen following disaster (Ursano et al., 1996). For individuals experiencing symptoms beyond what is expected from extreme stress due to a disaster – such as psychosis, suicidal or homicidal ideation, or major depression – a prompt mental health assessment and referral for immediate acute care is warranted.

Casualty Support

All disaster work requires the act of compassionate presence, the ability to be present with those suffering, and to be of comfort without trying to fix or provide answers (Berliner et al., 2009). To stand with another human being experiencing their worst moments, so that they do not suffer alone, is a humble and sacred task. Providing compassionate presence is critical when mental health interventions include casualty support for the loved ones of victims missing or deceased or for those who have suffered serious physical injuries, along with their families.

While disaster mental health workers do not make death notifications, they may accompany law enforcement personnel or medical examiner or coroner for the purposes of providing support. Families and friends of those lost or feared lost in a disaster require special consideration. Mental health workers support these individuals in private areas, whenever possible. Compassionate presence while loved ones wait for confirmation of death creates an invaluable support system that continues as the mental health worker accompanies family to the death notification.

Advocacy and Outreach

Disaster mental health workers are often able to advocate for client needs at disaster service sites. Advocacy can include asking officials when the next client briefing will be, requesting specific meals for populations with dietary or religious restrictions, ensuring that the disaster response is culturally competent, and handling any special needs situations.

Disaster mental health workers also serve on outreach teams that access affected communities, going door-to-door to check on those who may not have been seen at service centers that open to meet the needs of individuals during disasters. Workers may also reach out to those in hospitals (Cronin and Ryan, 2010).

Community Engagement

People usually share the experience of disasters with others who may be family, friends, or neighbors – or frequently strangers. It is necessary to respond to both individual and group needs in all four stages of a disaster. Effective programs and services integrate and address multiple levels, including community-level interventions, drawing on local coping strategies, and community resources (von Peter, 2009). For many people affected by disaster, being able to contribute to the recovery provides an immediate path to personal stability and resilience. Communities in less-developed regions are often self-sustaining; local economies and skills can be identified and reenergized. Those in more developed situations can use networking skills and community capacity to engage in recovery. Social workers can play a facilitating and supportive

role, for example, identifying and strengthening community 'leaders' and helping those affected to develop recovery strategies.

Staff and Peer Support

Volunteers and staff are vulnerable to stress by virtue of their work and the circumstances in which they do it. Staff and volunteers are often poorly prepared for their own emotional reactions to providing care and relief to others. Volunteers may work long hours in challenging emergency environments, often putting aside their own needs. At the end of the day, they often feel inadequate to help beneficiaries with the tragedy they are facing. Additionally, as members of affected communities, staff and volunteers often work close to home and may experience the same losses and grief in their families and communities as the beneficiaries they are supporting. Basic first aid training or being part of a disaster response team is not enough to prepare volunteers for these experiences.

Most disaster workers experience mild, normal stress reactions. Several studies indicate that one out of every three disaster workers at some point experiences severe stress symptoms (Marmar et al., 1996; Gibson and Iwaniec, 2003; Fullerton et al., 2004; Reid et al., 2005; Young et al., 2007).

The well-known concept of stress was defined by Selye (1956) as the nonspecific response of the human body to any demand made upon it. Figley (1995) describes compassion stress as a natural outcome of knowing about trauma experienced by a client rather than as a pathological process. The symptoms of compassion stress can include helplessness, confusion, and isolation. When compassion stress is ongoing and severe it can become 'compassion fatigue,' defined as "a state of exhaustion and dysfunction, biologically, physiologically, and emotionally, as a result of prolonged exposure to compassion stress" (Figley, 1995: p. 11).

Disaster mental health workers have a key role in maintaining a stable workforce during prolonged disaster operations. Staff and volunteers can be supported at key points to reduce the likelihood of developing stress-related problems. The primary intervention consists of good, solid information about the tasks at hand and about stress and coping with emotional reactions to difficult situations. Such information prepares helpers to detect their own reactions and offers options for self-care and peer support. Appraisal and recognition from others that they and their work are valued are essential.

The needs of volunteers and staff are often similar to the needs of those they are supporting. A supportive environment is a crucial factor in minimizing stress. This can be achieved by providing accessible guidance and support from managers and peers, by creating an organizational culture where people can talk openly and share problems without fearing consequences, by arranging regular meetings which bring all staff and/or volunteers together and foster a feeling of belonging to a team, and by respecting the principle of confidentiality so that people feel it is safe to talk about stress and seek help.

Peer support means offering assistance to someone at the same level as the supporter, enabling staff and volunteers to share their knowledge, perspectives, and experiences for the benefit of one another. The principles are generally the same as for PFA and supportive communication. There is evidence that an active, supportive approach to stressful situations facilitates successful coping. Allowing someone to talk about reactions and feelings facilitates coping and helps the individual to deal with stressful situations. The peer supporter provides support only and does not become a counselor. Peer supporters provide short-term assistance and do not replace professional help (IFRC-CPS, 2009: pp. 96–97).

Cultural Implications

Social workers deal with the psychosocial impact of a disaster on individuals, families, and communities. Humans respond to the stress of a disaster in a variety of ways, influenced by many factors: beliefs and religion, age, gender, sexual orientation, ability, education, financial resources, communication experience among others. Some of those influences can be seen as cultural. Providing culturally competent services to those affected requires action before, during, and after a disaster. Creating culturally competent environments requires more than recruiting bilingual and bicultural workers, sponsoring diversity workshops, or hiring a 'token' racial or ethnic minority group representative (IFRC, 1994).

Culture is expressed through shared customs, values, social rules of behavior, rituals and traditions, and perceptions of human nature and natural events. Cultural competence must be a part of the program values, included in the program's mission statement, and encouraged in attitudes, policies, and practices at every level (e.g., Dominelli, 1996; Weaver, 1999; Council of Europe, 2011).

To develop culturally competent programs and services, planners must assess and understand the community's composition; identify culture-related needs of the community; be knowledgeable about formal and informal community institutions that can help meet diverse psychosocial needs; gather information from and establish working relationships with trusted organizations, service providers, and cultural group leaders; and anticipate and identify solutions to cultural problems that may arise (Myers, 1994).

Concluding Observations

There has been an increase in the incidence and impact of both natural and human-induced disasters. While more disasters and the highest financial costs are reported in the developed world, poverty always determines the relative severity of impact on individuals and communities. Investment in disaster preparedness has a social and economic benefit. Social workers' skills and experience are very relevant for disaster work, including concern for human rights, protection of children and vulnerable people, and social inclusion of those affected. Psychosocial and community recovery interventions are needed. The impact of disasters on workers is usually significant; their needs are often similar to those they are supporting and must be addressed. The nature of the response must take account of local culture and local social systems. Social workers, by training and experience, are

well placed to make a distinctive contribution at all stages of disaster work.

See also: Armed Conflict and Social Work: Significance and Implications; Community Resilience to Disasters; Demographic Impact of Disasters; Disasters in Social, Cultural and Political Context; Ecological Social Work: Shifting Paradigms in Environmental Practice; Green Social Work; Hazards and Disaster Research in Contemporary Anthropology; International Association of Schools of Social Work; International Council on Social Welfare; Mass Trauma: Psychopathological Effects across the Life Span, Military and Disaster Psychiatry; Natural Disasters: Health-Related Aspects; Social Work Theory.

Bibliography

Arbour, M.C., Murray, K., Arriet, F., Moraga, C., Vega, M.C., 2011. Lessons from the Chilean earthquake: how a human rights framework facilitates disaster response. Health and Human Rights: An International Journal 13 [Online]. http://hhrjournal.org/index.php/hhr/article/view/411/621 (accessed 31.12.12.).

Banks, S., Nøhr, K. (Eds.), 2011. Practising Social Work Ethics Around the World. Routledge, London.

Berliner, P., Ryan, D., Taylor, J., 2009. Making referrals: effective collaboration between mental health and spiritual care practitioners. In: Brenner, G.H., Bush, D., Moses, J. (Eds.), Creating Spiritual and Psychological Resilience: Integrating Care in Disaster Relief Work. Guilford Press, New York.

Biermann, F., 2001. Umweltflüchtlinge: Ursachen und lösungsansätze. Aus Politik und Zeitgeschichte 12, 24–29.

Bradshaw, S., 2004. Socioeconomic Impacts of Natural Disasters: A Gender Analysis. Economic Commission for Latin America and the Caribbean-ECLAC, Santiago de Chile.

Congress, E., Cronin, M., Wallace, M., 2010. Ageing, long-term disability, and disasters. In: Mayer, M.J., Demark, F.L. (Eds.), Issues of Ageing and Disability: International Perspectives. UN NGO Committee on Ageing, New York, pp. 5–12.

Council of Europe, 2011. Intercultural Competences in Social Services: Constructing an Inclusive Institutional Culture – Methodological Guide. Council of Europe, Strasbourg. http://issuu.com/councilofeurope/docs/issuu_guide_culture_institutionnelle_inclusive_e/27 (accessed 06.11.12.).

Cronin, M., Ryan, D., 2010. Practice perspectives of disaster work. In: Gillespie, D., Danso, K. (Eds.), Disaster Concepts and Issues: A Guide for Social Work Education and Practice. Council on Social Work Education, Alexandria, VA.

Dominelli, L., 1996. Deprofessionalizing social work: anti-oppressive practice, competencies and postmodernism. British Journal of Social Work 26, 153–175.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Dominelli, L., 2012. Green social Work: From Environmental Crises to Environmental Justice. Polity Press, Cambridge.

Everly, G., Mitchell, J.T., 1999. Critical Incident Stress Management: A New Era and Standard of Care in Crisis Intervention, second ed. Chevron, Elliott City, MD.

Figley, C.R. (Ed.), 1995. Compassion Fatigue: Coping with Secondary Traumatic Stress Disorder in Those Who Treat the Traumatized. Brunner/Mazel, New York.

Fullerton, C.S., Ursano, R.J., Wang, M.S., 2004. Acute stress disorder, post-traumatic stress disorder, and depression in disaster or rescue workers. American Journal of Psychiatry 161, 1370–1376.

German Federal Ministry for Economic Co-operation and Development (BMZ), 2001. Poverty Reduction: A Global Responsibility. Program of Action 2015. German Federal Ministry for Economic Co-operation and Development – BMZ, Bonn.

Gibson, M., Iwaniec, D., 2003. An empirical study into the psychosocial reactions of staff working as helpers to those affected in the aftermath of two traumatic incidents. British Journal of Social Work 33, 851–869.

Gillespie, D., Danso, K., 2010. Disaster Concepts and Issues: A Guide for Social Work Education and Practice. Council on Social Work Education Press, Alexandria, VA.

Guha-Sapir, D., Hargitt, D., Hoyois, P., 2004. Thirty Years of Natural Disasters 1974–2003: The Numbers. Centre for Research on the Epidemiology of Disasters, University of Louvain Press, Louvain.

Halpern, J., Tramontin, M., 2007. Disaster Mental Health: Theory and Practice. Thomson Brooks/Cole, Belmont, CA.

International Federation of Red Cross & Red Crescent Societies (IFRC). World Disasters Reports 1993–2012. IFRC, Geneva (Annual reports on people and disasters).

International Federation of Red Cross & Red Crescent Societies (IFRC) Center for Psychosocial Support, 2009. Psychosocial Interventions: A Handbook. IFRC CPS, Copenhagen. http://psp.drk.dk/sw40688.asp.

IFRC-CPS (International Federation of Red Cross & Red Crescent Societies Center for Psychosocial Support), 2009. Community-Based Psychosocial Support. IFRC-CPS, Copenhagen.

IFRC-CPS (International Federation of Red Cross and Red Crescent Societies-Center for Psychosocial Support), 2013. Caring for Volunteers: A Psychosocial Support Toolkit. IFRC CPS, Copenhagen. http://www.pscentre.org/wp-content/uploads/volunteers_EN.pdf (accessed 15.09.13.).

IFRC (International Federation of Red Cross and Red Crescent Societies), 1965. The Seven Fundamental Principles. Red Cross and Red Crescent, Geneva. http://www.ifrc.org (accessed 01.01.13.).

IFRC (International Federation of Red Cross and Red Crescent Societies), 1994. Code of Conduct. Red Cross and Red Crescent, Geneva. http://www.ifrc.org (accessed 01.01.13.).

IFSW and IASSW (International Federation of Social Workers & International Association of Schools of Social Work), 2004. Statement of Ethical Principles. International Federation of Social Workers. Bern. http://ifsw.org/policies (accessed 08.04.13.).

IFSW, IASSW and ICSW (International Federation of Social Workers, International Association of Schools of Social Work & International Council on Social Welfare), 2012. The Global Agenda for Social Work and Social Development: Commitment to Action. www.globalsocialagenda.org (accessed 20.03.12.).

Inter-Agency Standing Committee, 2007a. Protecting Persons Affected by Natural Disasters – IASC Operational Guidelines on Human Rights and Natural Disasters. Brookings-Bern Project on Internal Displacement, Washington, DC.

Inter-Agency Standing Committee-IASC, 2007b. IASC Guidelines on Mental Health and Psychosocial Support in Emergency Settings. IASC, Geneva. http://www.humanitarianinfo.org/iasc/pageloader.aspx.

International Debates, 2005. Climate change and poverty: socioeconomic costs of natural disasters. International Debates 3, 201–224.

Klasing, A.M., Moses, P.S., Satterthwaite, M.L., 2011. Measuring the way forward in Haiti: grounding disaster relief in the legal framework of human rights [Online] Health and Human Rights: An International Journal 13. http://hhrjournal.org/index.php/hhr/rt/captureCite/408/610 (accessed 31.12.12.).

Kondrat, M.E., 2011. Person-in-Environment. Oxford Bibliographies. OUP, Oxford.

Lindemann, E., 1994. Symptomatology and management of acute grief. American Journal of Psychiatry 101, 141–148.

Marmar, C.R., Weiss, D.S., Metzler, T.J., Delucchi, K., 1996. Characteristics of emergency services personnel related to peritraumatic dissociation during critical incident exposure. American Journal of Psychiatry 153, 94–102.

Miller, Joshua A., 2012. Psychosocial Capacity Building in Response to Disasters. Columbia University Press, New York.

Myers, D., 1994. Disaster Response and Recovery: A Handbook for Mental Health Professionals (SMA Publication No. 94-3010). U.S. Department of Health and Human Services, Washington, DC.

National Governor's Association, 1979. Emergency Preparedness Project Final Report. USA Government Printing Office, Washington, DC.

Pittaway, E., Bartolomei, L., Rees, S., 2007. Gendered dimensions of the 2004 tsunami and a potential social work response in post-disaster situations. International Social Work 50, 307–319.

von Peter, S., 2009. The concept of mental trauma and its transcultural application. Anthropology & Medicine 16, 13.

Reid, W.M., Ruzycki, S., Haney, M.L., Brown, L.M., Baggerly, J., Mescia, N., Hyer, K., 2005. Disaster mental health training in Florida and the response to the 2004 hurricanes. Journal of Public Health Management & Practice 11, 57–62.

Reyes, G., 2006. Psychological first aid: principles of community-based psychological support. In: Reyes, G., Jacobs, G. (Eds.), Handbook of International Disaster Psychology, Practice and Programs, vol. 2. Praeger, Westport, CT.

Roberts, A.R. (Ed.), 2000. Crisis Intervention Handbook: Assessment, Treatment and Research, second ed. Oxford University Press, New York.

Rosenfeld, Lawrence B., Caye, Joanne S., Lahad, Mooli, Gurwitch, Robin H., 2010. When Their World Falls Apart: Helping Families and Children Manage the Effects of Disasters. National Association of Social Workers Press, Washington, DC.

Selye, H., 1956. The Stress of Life. McGraw-Hill, New York.

Smith, F., Jolley, E., Schmidt, E., 2012. Disability and disasters: the importance of an inclusive approach to vulnerability and social capital. Sightsavers, Haywards Heath, UK.

UNISDR (United Nations International Strategy for Disaster Reduction), 2012. Making Cities Resilient Report 2012. United Nations, New York. http://www.unisdr.org (accessed 30.12.12.).

UNDP (United Nations Development Program), 2004. Reducing Disaster Risk: A Challenge for Development. UNDP, New York.

UNDP (United Nations Development Program), 2010. Gender and Disasters. UNDP, New York. http://www.undp.org (accessed 03.01.13.).

UNDP (United Nations Development Programme), 2012. Fast Facts: Disaster Risk Reduction and Recovery. UNDP, New York. http://www.undp.org (accessed 01.01.13.).

UNISDR (United Nations International Strategy for Disaster Reduction), 2009. UNISDR Terminology on Disaster Risk Reduction. United Nations, Geneva. http://www.unisdr.org/we/inform/terminology (accessed 15.09.13.).

United Nations Development Program, 2011. Human Development Report 2011: Sustainability and Equity. United Nations, New York. http://www.unhabitat.org/pmss/listitemdetails.aspx?publicationID=2917 (accessed 15.09.13.).

United Nations Human Settlements Program – UN-Habitat, 2010. State of the World's Cities Report-2010–2011. United Nations, Nairobi.

Ursano, R.J., Grieger, T., Mccarroll, J., 1996. Prevention of posttraumatic stress: consultation, training, and early treatment. In: Kol, B.V.D., Mcfarlane, A., Weisaeth, L. (Eds.), Traumatic Stress: The Effects of Overwhelming Experience on Mind, Body, and Society. Guilford Press, New York, pp. 441–462.

Weaver, H.N., 1999. Indigenous people and the social work profession: defining culturally competent services. Social Work 44, 217–225.

Wisner, B., 2004. Assessment of capability and vulnerability. In: Bankoff, G., Frerks, G., Hilhorst, D. (Eds.), Mapping Vulnerability: Disasters, Development and People. Earthscan, London, pp. 183–193.

World Bank, 2001. World Development Report 2000–2001: Attacking Poverty. World Bank, Washington, DC.

World Health Organization, 2011. Psychological First Aid: Guide for Field Workers. WHO, Geneva. Available at: http://whqlibdoc.who.int/publications/2011/9789241548205_eng.pdf.

Young, B.H., Ford, J.D., Watson, P.J., 2007. National Center for PTSD Fact Sheet: Disaster Rescue and Response Workers. US Department of Veteran Affairs, Washington, DC. http://www.ptsd.va.gov (accessed 31.12.12.).

Zervaas, D.P., 2007. Global Disaster Reduction Approaches: The Role of the United Nations International Strategy for Disaster Reduction (UN/ISDR). International Seminar on Disaster Planning, Management, and Relief: New Responsibilities for Social Work Education. CSWE, Barbados.

Zetter, R. (Ed.), 2012. World Disasters Report 2012: Focus on Forced Migration and Displacement. International Federation of Red Cross and Red Crescent Societies, Geneva.

Relevant Websites

http://www.redcross.org/ – American Red Cross: This organization is a part of the global Red Cross network, prevents and alleviates human suffering in the face of emergencies by mobilizing the power of volunteers and the generosity of donors.

http://www.redcross.org.uk/ – British Red Cross: This is a National Society of the international Red Cross movement, is a volunteer-led humanitarian organization that helps people in crisis, whoever and wherever they are.

http://www.cred.be – Center for Research on the Epidemiology of Disasters (CRED) promotes research, training, and information dissemination on disasters; keeps a running list of disasters that are or have occurred in the last week.

www.ready.gov – Disaster Preparedness: This U.S. government website has a variety of publications for download covering various topics about planning and preparation for disasters and emergencies.

http://www.udel.edu/DRC/ – Disaster Research Center-University of Delaware.

http://earthquake.usgs.gov/earthquakes/index.php – Earthquake Center, U.S. Geological Survey website.

http://www.fema.gov/ – Federal Emergency Management Agency (FEMA): The U.S. government website with information on disaster mitigation, preparedness, response, recovery, education, and references.

http://globalnetwork-dr.org/home.html – Global Network of Civil Society Organisations for Disaster Reduction: This is a major international network of nongovernmental and not-for-profit organizations committed to working together to improve the lives of people affected by disasters worldwide.

http://www.gdacs.org/index.asp – Global Disaster Alert and Coordination System: This system provides near real-time alerts about natural disasters around the world and tools to facilitate response coordination, including media monitoring, map catalogues and Virtual On-Site Operations Coordination Center.

http://hrrc.arch.tamu.edu/ – Hazards Reduction and Recovery Center-Texas A&M University.

http://www.ifrc.org/ – International Federation of Red Cross & Red Crescent Societies: The IFRC is the world's largest humanitarian network that reaches 150 million people in 187 national societies through the work of over 13 million volunteers.

http://psp.drk.dk/sw2955.asp – International Federation of Red Cross & Red Crescent Societies Psychosocial Support Centre: The center was established to promote, guide, and enhance psychosocial support initiatives carried out by Red Cross and Red Crescent National Societies globally.

http://www.humanitarianinfo.org/iasc/pageloader.aspx – Inter-Agency Standing Committee. The IASC is the primary mechanism for interagency coordination of humanitarian assistance. It is a unique forum involving the key UN and non-UN humanitarian partners. The *IASC Guidelines* on Mental Health and Psychosocial Support in Emergency Settings can be downloaded from this website.

http://www.cswe.org/CentersInitiatives/KAKI/KAKIResources/25187.aspxA – Kendall Institute Disaster Management Resources, a collection of resources addressing disaster management is available from the Council on Social Work Education in the USA.

http://www.nhc.noaa.gov/ – National Hurricane Center: A Web site of the U.S. National Weather Service.

http://www.preventionweb.net/ – Prevention Web, serves the information needs of the disaster reduction community.

http://www.reliefweb.int – ReliefWeb, an United Nations Web site for reliable humanitarian information on global crises and disasters. There are listing of events and locations where disaster relief is in operation.

http://www.unisdr.org/ – United Nations Office for Disaster Risk Reduction: This organization serves as the focal point in the United Nations system for the coordination of disaster risk reduction and to ensure synergies among disaster risk reduction activities.

Social Work and Psychodynamic Thought

Jerrold R Brandell, School of Social Work, Wayne State University, Detroit, MI, USA
Laura J George, Washington Center for Psychoanalysis, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews psychoanalytic theories of development and psychopathology, different models of psychodynamic treatment used in clinical social work practice, and several major critiques of psychodynamic thinking. Psychodynamic social work is defined as any psychoanalytically informed social work practice that makes selective use of psychoanalytic ideas about human development, pathogenesis, and treatment. The psychoanalysis of today, while preserving important traditions spanning over 100 years, is a far more sophisticated, complex, and scientifically elaborated body of knowledge. Its development argues convincingly for a reappraisal of what contemporary psychoanalytic theory and practice knowledge have to offer the clinical social work field.

Introduction

The first portion of this article reviews psychoanalytic theories of human development, psychopathology, and treatment used in clinical social work practice. Psychodynamic social work is here defined as any psychoanalytically informed social work practice that makes selective use of psychoanalytic ideas about human development, pathogenesis, and treatment. In the second portion, we consider five major critiques of psychodynamic thinking.

Psychodynamic Conceptions of Mind, Psychopathology, and Treatment

Conceptions of Mind in Classical Freudian Theory

Prior to the introduction of the structural model in 1923, Freud's understanding of human nature was that the unconscious mind contained instinctual drives (Freud identified two drives or instincts, sex (libido) and aggression, to which he ultimately attributed the most basic internal motivations for behavior), which are biologically based motivational systems seeking satisfaction in behaviors. According to this model, intrapsychic conflicts arise when unconscious desires press for representation in the conscious mind, but are prevented from achieving it by repression (a defense mechanism that excludes a psychologically painful or conflict-laden idea from conscious awareness. Repression is regarded as the prototype for defenses insofar as all defensive operations are designed to protect the self from conscious exposure to objectionable ideas, impulses, attitudes, and wishes. In Freud's early writings, repression was used more or less interchangeably and synonymously with the concept of defense, although as other defenses were elaborated, it came to be more narrowly defined). In this framework, accumulating tension that has not been permitted release, results in symptoms. Symptom relief occurs when disassociated mental contents are reconnected by breaking through the repression barrier, in effect making the unconscious conscious.

Freud also postulated four stages of psychosexual development based on erogenous zones. If there is too much trauma or frustration, a child cannot move to the next stage of development. The stages include the pleasurable sensations of the oral, anal, phallic, and genital phases. The phallic phase becomes important as the child moves toward resolution of the Oedipus complex, when there is simultaneous renunciation of the child's desire for sexual possession of the opposite-sex parent and for the same-sex parent's demise. The final resolution of this conflict permits individuals to establish a new internal structure, including a capacity to regulate desire and guilt. While working with soldiers returning from World War I, social workers found that Freudian theory was consonant with casework practice (Danto, 2009). However, a debate soon arose between the diagnostic school, which relied on Freudian theory with an emphasis on the unconscious, and the functional school, which stressed Otto Rank's ideas about the conscious will (Goldstein, 2002).

Conceptions of Mind in Ego Psychology

While early psychoanalysts gave weight to the drives, American theorists placed emphasis on having a strong ego that could cope with reality, impulses, and feelings. The id, the ego, and the superego were the three intrapsychic agencies in Freud's structural theory. In this model, the id is thought to be composed of aggressive or sexual wishes; the ego is considered the mediator between the id and the outside world, and the superego is, in effect, the individual's conscience. By calling upon a range of defense mechanisms, ego psychology proposes that the development of a strong ego helps individuals deal with conflicts. These ideas became the foundation for social work practice approaches such as the problem-solving model (Perlman, 1957).

Conceptions of Mind in Object Relations Theories

There are a number of theorists subsumed under the heading of object relations. The term object relations can refer to one's relationships with other individuals, to the internal mental representations of others, or to an internal image of oneself. The theory encompasses the idea that instead of being

motivated by their drives, individuals need to connect with others. More specifically, theorists propose that each individual bears the traces of early relationships, and that these internalized representations become part of each individual's relationship patterns. Though not generally credited, social work's early emphasis on the therapeutic relationship may be regarded as a forerunner of object relations theories' stress on the transference (a reflexive, unconscious repetition or revivification of varying combinations and patterns of ideas, fantasies, affects, attitudes, or behavior, originally experienced in relation to a significant figure from one's childhood past, that have been displaced onto an intercurrent interpersonal relationship).

In describing human development, the object relations school emphasizes how caregivers and infants attach and individuate (Bowlby, 1988), as well as how attachment influences brain development (Malekpour, 2007). Members of the women's movement were critical of the attachment theory's emphasis on the role of mothers, while minimizing the role of fathers (Hollway, 2006). Social workers were also critical, and thought material needs such as food and clothing were more crucial (Brandell and Ringel, 2007).

Conceptions of Mind in Self Psychology

Heinz Kohut, who founded self psychology (a treatment orientation within psychoanalysis that encourages the growth of an integrated and cohesive intrapsychic structure through empathic attunement and the internalization of particular functions initially proffered by the therapist), believed that a cohesively functioning self could not develop in isolation, but instead that beginning in childhood, individuals required selfobjects (a term devised by Kohut to refer to a particular kind of object relationship in which the object is actually experienced as an extension of the self, without psychological differentiation. According to the theory, there are distinctly different though equally necessary kinds of selfobject experiences: mirroring selfobjects, "who respond to and confirm the child's innate sense of vigor, greatness and perfection"; idealized selfobjects, "to whom the child can look up and with whom he or she can merge as an image of calmness, infallibility, and omnipotence" (Kohut and Wolf, 1978: p. 414); and alter ego or partnering selfobjects, who provide a range of experiences through which children acquire a sense of belonging and of essential alikeness within a community of others). Kohut defines selfobjects as objects experienced as being a part of or an extension of the subject's self. He identified three types of selfobject experiences: mirroring, idealization, and partnering. He thought mirroring selfobjects reflect back a child's greatness, idealizing selfobjects offer a child an infallible caregiver, and partnering selfobjects provide a child with the experience of someone who is like him or a sense of community. Related to the three types of selfobjects is an intrapsychic structure called the tripolar self (the intrapsychic structure linking particular kinds of self/selfobject relationships with corresponding poles of self-experience, constituting the grandiose-exhibitionistic self, the idealized parent imago, and the alter ego). The tripolar self consists of a grandiose exhibitionistic self associated with mirroring, an idealized self taken from experiences with the child's idealized selfobjects, and an alter ego that recognizes the reality that others bear an essential alikeness to the self. Kohut used the term transmuting internalization (the developmental process whereby a function formerly performed by another is taken into the self and becomes the enduring, unique psychological structure of that individual, occurring via optimal frustration and optimal gratification, and characterized by incremental accretion and consolidation) to refer to the gradual assimilation of functions (e.g., soothing) originally associated with the selfobjects. Although best known for his emphasis on the importance of empathetic attunement, Kohut also acknowledges that individuals need nontraumatic disappointments to develop (Kohut and Wolf, 1978). Due to social work's commitment to working with vulnerable populations who can be stymied by disappointments, there is potential for greater use of this framework in the creation of practice models.

Conceptions of Mind in Lacanian Theory

Jacques Lacan (2007) has made contributions not only to psychoanalysis, but also to linguistic and literary theory. His exploration of language involving Otherness, as it relates to race and class, offers psychoanalysts and social workers a point of convergence.

Lacan distinguishes three agencies in the psyche: the imaginary order, the symbolic order, and the Real. The imaginary order is characterized by primitive self-knowledge without language, fantasy, and images that equates with the infant's entry into society. The newborn is unable to distinguish himself from his mother or the Other and sees himself in fragments. Later, at 6–18 months, the infant recognizes his mirror image, his experience of himself as a unified individual separate from the Other. This is the first step toward the creation of the subject and it is the beginning of social interactions that help identify the self in terms of the Other (Sarup, 1993).

The symbolic order or the big Other involves linguistic collaboration characteristic of communities (Zizek, 2006). Lacan believes there is a set of signifiers that are more than words; they embody concepts with meaning that form narratives about the subject and create rules that govern behavior. Language eventually teaches one what to desire, and introduces children to loss. Lacan links desire to the oedipal crisis, which, in turn, leads to the child entering the symbolic stage. During a child's fifth year, when he or she needs to renounce the desire for the mother and suffers defeat in relation to the father, a symbolic castration takes place. There ensues a pact between the child and the parents in the social order that Lacan calls the Name-of-the-Father.

The structure of the Real is one in which the individual is free from all desires and not affected by worldly enticements. The Real resists symbolization but holds truth. It is impossible to attain; indeed, we only have partial access to the Real (e.g., in dreams) (Sarup, 1993). Note that Lacan's 'other' (lower case 'o') refers to a person with whom one has an imaginary relationship (someone similar to oneself) and 'Other' (upper case 'O') refers to a person or institution that serves a symbolic function (someone or something with authority) (Fink, 1997: p. 232).

Conceptions of Mind in Relational and Intersubjective Theories

Relational theorists argue that one develops as a result of one's early relationships with parents and other figures. Recurring relational patterns develop during infancy and childhood, and eventually unconsciously organize emotional and relational experiences, thus offering rules of what to expect in interactions. Both social work and relational theorists seem to emphasize a less hierarchical relationship model, as indicated in social workers' use of the word 'client' (vs 'patient') and in relational theorists' emphasis on the contributions of both therapist and client (Benjamin, 1988).

Psychopathology in Classical Freudian Theory

Freud thought symptoms or compromise formations occur when despite repression there is an inner conflict involving one's sexual and aggressive wishes with the external world. In his psychosexual development theory, Freud asserts that psychopathology occurs when there is a fixation of instincts and early developmental arrests that prevent individuals from being able to move forward in their development.

Psychopathology in Ego Psychology Theory

In ego psychology, psychopathology is directly related to how the ego functions as it mediates between internal forces and external demands. Having inadequate defenses or coping mechanisms lets in unacceptable sexual or aggressive wishes from the unconscious and superego disapproval that stymies growth. Ego functioning is also influenced by constitutional deficits, overstimulation, excess frustration, and defenses that were adaptive in childhood but are problematic in adulthood.

Psychopathology in Object Relations Theories

Object relations theorists propose that disturbances in the mother–infant relationship influence psychic structure formation and how one relates to others. Starting in infancy, problems can occur when the infant does not have a responsive caregiver who can offer him or her an internal representation of a relationship that brings with it a sense of well-being (Waddell, 2002). Object relations theorists are interested in severe psychopathologies and unyielding defenses (Kernberg, 1992). The clinical symptoms that suggest object relational problems include disturbances in an individual's capacity to develop mutually supportive relationships.

Psychopathology in Self Psychology Theory

Kohut essentially understands pathology in terms of selfobject disorders, which occur when there is a failure in cohesion or a fluctuation between feeling inferior and overestimating the self (narcissism). When individuals suffer from traumatic disruptions in ties with mirroring and idealized selfobjects, there is a corresponding failure of transmuting internalization or the taking in of functions (such as self-regulation) necessary for the development of self-cohesiveness. The result may lead to a range of symptoms, referred to as disintegration products, or an experience of anxiety referred to as disintegration anxiety (the fear of breaking up of the (nuclear) self, described by Kohut as the deepest form of anxiety an individual can experience, its origins traceable to the earliest preverbal stages of psychological development).

Psychopathology in Lacanian Theory

For Lacan, psychopathology arises not from conflicts involving ego, id, and superego, but from how an individual deals with truth and being symbolically castrated or dominated by society. Lacan formulated three categories of psychopathology – psychosis, neurosis, and perversion; neurosis is further subdivided into hysteria and obsession (sometimes including phobia). The ability to access the symbolic order (initiation into the Law or the Name-of-the-Father precipitated by the oedipal crisis) in neurosis is what differentiates it from psychosis, in which there is no access to the symbolic order. In other words, in psychosis, the unconscious is present but not functioning. The pervert, unlike the psychotic, is initiated in the symbolic order, but unlike the neurotic, does not fully understand the reasoning behind the Law.

Psychopathology in Relational and Intersubjective Theories

Because relational theories focus on interpersonal patterns, there are no extensive symptom listings that can be used as a diagnosis. In relational theories, problematic repetitive patterns of relating to others can reflect psychopathology. In addition, symptoms occur when a disruption of needs provokes anxiety in significant relationships within the client's life.

Classical Freudian Treatment

Freud's cure includes the undoing of repression, resulting in the freeing of energy and emotions. This undoing of repression makes the unconscious conscious, which eventually leads to the client gaining insight into the meaning of behaviors and symptoms. To facilitate a cure, psychoanalysis makes use of free association, interpretation (called the 'central therapeutic activity' of the psychoanalytic clinician, interpretation represents the clinician's understanding of a client's mental life. Some psychoanalysts link the concept almost exclusively to the psychoanalytic notion of resistance, believing that any interpretative comment made by the clinician should be expressly designed to draw a client's attention to a resistance. Others, however, have defined the term somewhat less narrowly, viewing an interpretation as a statement of new knowledge about a client. Several different basic kinds of interpretation have been described in the dynamic literature. These are genetic interpretations, dynamic interpretations, interpretations of

resistance, and transference interpretations), dream analysis, and the transference.

Ego Psychological Treatment

Clinically speaking, practitioners of ego psychology assist patients in strengthening their egos so that they can better cope with the pressures from their id, superego, and society in general. During treatment, the therapist helps individuals to develop mature defenses, and to use conflict-free adaptations when problems arise. An expansion of the person's capacity to accept unpleasant realities is a sign that ego strength is improving.

Object Relational Treatment

While there are certain conceptual dissimilarities among the various object relations theories, treatment usually focuses on the internalization (refers to those processes that contribute to the development of an internal world (sometimes referred to as the representational world) and to the consolidation of internal structure. Includes incorporation, introjection, and identification) of relationships, particularly the therapeutic relationship. Thus, object relations therapists often work in the transference while they examine the client's internal world. Practitioners with this orientation make interpretations, while monitoring their clients' as well as their own feelings, associations, and fantasies.

Self Psychological Treatment

Self psychological treatment, which is divided into the distinctive treatment phases of understanding and explaining, helps the client to develop missing functions that he or she failed to acquire earlier in life from selfobjects. Therapy is based on providing empathic experiences that might contribute to developing a cohesive self (the essential self-structure of a well-adapted, healthily functioning individual whose self-functioning evinces the harmonious interchange of ambitions, ideals, and talents with the events of everyday life). In addition to providing an attuned experience, working through ruptures helps the client achieve transmuting internalizations (i.e., taking in the therapist's functions, including affect regulation). Interpretations are used to help clients understand their overreaction to injury and explore how to restore cohesiveness after an empathic failure.

Lacanian Treatment

For Lacan, the goal of treatment is not to enhance an individual's well-being but to uncover the truth about the client's desire. Lacanian analysts believe that truths about a client's desire are present in discourse (even if discourse cannot articulate the whole truth about desire) (Qazi, 2011). In Lacanian treatment, the practitioner offers an interpretation that makes the unconscious desire conscious. In doing so, the therapist works with what Lacan calls signifiers (representing the subject) and the master-signifier (representing the subject for all signifiers), which can have traumatic meaning for the subject. Lacan believes therapy can help subjects modify their signifiers or find new signifiers, thus enabling them to reorder their sense of self and of their relations to others. This is achieved through a process called re-signification, which reduces signifiers' traumatic meaning by integrating them into self-understanding.

Relational and Intersubjective Treatments

Relational therapists are noted for emphasizing a collaborative interaction between client and therapist, their mutual construction of meaning in the therapeutic relationship, and their less adherent stance toward technical restrictions. Treatment in relational models delves into the pattern of the client's relationships with the therapist and others, and tries to ameliorate maladaptive patterns so a client can interact with others more securely.

Brief Psychodynamic Treatment

Due to managed care, concern about outcome measures, and a desire to meet their clients' needs, social workers often use brief treatment models. Since traditional psychodynamic therapies that explore relationship patterns often take time, clients with specific goals or limited time and financial resources frequently consider short-term therapy. Among the challenges implicit in brief treatments are the need to establish a working alliance rapidly and to focus on a specific objective that can be achieved within an agreed time. Therapists also find that while having a specific goal creates more structure, it limits free association.

Controversies Involving Freud's Legacy and Contemporary Critiques of Psychoanalytic Thought

Psychoanalytic thought has gradually been expropriated, absorbed, and popularized, pervading both our intellectual life and common culture. Signal psychoanalytic precepts such as dynamic motivation, unconscious fantasies, oedipal desire, separation anxiety, or the notion of intrapsychic conflict and defense have remained as disquietingly powerful in contemporary times as they were for Freud's detractors and adversaries in the last century. One example may serve as an illustration.

Jeffrey Masson (1994) argued that Freud's 'abandonment' of seduction theory was a deliberate, scientifically dishonest move calculated to end his isolation from the Viennese medical establishment due to his revolutionary ideas. Masson's controversial thesis soon led to a contentious debate, waged in the professional journals, as well as the popular media. Intellectual historians (e.g., Peter Gay, 1988) have deemed Masson's thesis 'preposterous' for two reasons. Not only did Freud, in subsequent years, continue to propose equally radical ideas, but he also acknowledged, though painfully, that real seduction and rape were indisputable facts in at least a few cases of which he had direct knowledge.

Various critiques have also been made against psychoanalysis and psychoanalytic thought. Below, we consider carefully five rather specific criticisms of psychoanalytic thinking (Portions of this section are adapted, with permission of the publisher, from Brandell's 2004 book, *Psychodynamic Social Work*.)

Psychoanalysis Is Not Empirically Valid

Opposition to psychoanalytic ideas on the basis that they are conjectural, untested, and unsubstantiated by empirical methods, or in more contemporary patois, not 'evidence based,' has constituted one of the greatest challenges facing psychodynamically oriented practitioners and academicians.

Psychoanalysis, while employing empirical methods, is not an experimental science. Freud, however, was very eager to link his clinical findings with the prevailing quantitative science of his day (Holzman, 1998), associated with scientists such as von Helmholtz, Herbart, Fechner, and Brucke (Sulloway, 1979). Freud's attempts to treat psychoanalytic data as analogous to findings from the biological sciences, as well as his efforts to create a psychological theory modeled after the physical sciences were less than successful (Galatzer-Levy et al., 2000).

Historically speaking, psychoanalytic investigations were, with some notable exceptions, principally not quantitative in nature. These exceptions included a handful of large-scale investigations, such as the Menninger Psychotherapy Research Project, the Boston Psychoanalytic Institute Studies, and Fonagy and Target's (Fonagy and Target, 1994, 1996; Target and Fonagy, 1994) work at the Anna Freud Centre. Much of the current research on psychodynamic and psychoanalytic treatments is summarized by Shedler (2010). There is arguably more research now – not simply employing qualitative but, increasingly, quantitative methodologies – than ever before. However, this body of research is still comparatively small, especially when one considers how extensively cognitive-behavioral therapies have been researched. Some have argued that the relative absence of experimental data in psychoanalytic research, the failure to present psychoanalytic data in an experimentally testable form equivalent, for example, to data derived from other, nondynamic forms of therapy, invalidates it as an empirical discipline.

However, such a stringent definition of empirical science also excludes sciences such as astronomy and geology (Holzman, 1998: p. 5). Regardless, in consequence of this predominantly qualitative, nonexperimental research base, exponents of psychoanalysis have been disadvantaged in an academic community that places disproportionate value on nomothetic research – investigations that search for general laws or traits, and that are consonant with the philosophy of positivism. This disadvantage has been magnified to some extent insofar as the effects of other therapeutic approaches (e.g., cognitive or behavioral treatment) are readily measured and their efficacy supported by quantitative methods. Although there is now greater consensus that the nature of psychoanalytic findings is indisputably empirical, the comparatively small body of evidence supporting claims of the therapeutic efficacy of psychoanalytic and psychodynamic treatments undoubtedly led to a general decline in their status in the 1980s and 1990s, a situation that has only recently captured the attention of psychoanalytic researchers.

Psychoanalysis Is Elitist and Irrelevant

Another common charge leveled at psychoanalysis is that it constitutes an elite form of treatment, an inappropriate and largely irrelevant body of knowledge for social work expressly designed for neurotic female patients, who possessed vast financial resources and were able to commit literally to years of intensive (four to five sessions per week) therapy. For these critics, psychodynamic thought is suspect, because it has the effect of derailing the profession from its historical mission – serving the poor and underprivileged. Perhaps the most prominent critique along these lines was advanced by Specht and Courtney (1997), though this argument has been made many times before and since.

The essential problem with this critique is that it is based on various misconceptions and falsehoods that nevertheless continue to influence social work academe. First, psychoanalysis even as practiced by Freud and his early adherents was never exclusively with a neurotic clientele. Many of Freud's patients might not even be considered neurotic today, but rather suffering from character disturbances, posttraumatic and dissociative disorders, and even psychotic-like problems. Furthermore, treatments averaged under 2 years, and according to well-documented research, Freud had equal numbers of female and male patients, some treated gratis. Even assuming such criticisms are not spurious, reflexively equating the psychoanalysis of Freud's time with the rich spectrum of contemporary psychoanalytic psychologies and treatments is both misleading and misinformed.

Are psychoanalytic theories irrelevant for contemporary social work practice? Continued reliance in social work education on the work of Erik Erikson (psychosexual/psychosocial epigenesis) and John Bowlby (human attachment theory), theorists deeply influenced by psychoanalytic tradition, argues convincingly against this supposition. While such developmental thinking continues to pervade the contemporary social work literature, its theoretical origins in psychoanalysis are often unacknowledged, or even disputed.

The most damaging assertion is that psychoanalysis was designed exclusively for rich people, and is unmindful of the very real problems of the poor, the disenfranchised, and the downtrodden. Once again, this represents an oft-repeated but patently false, revisionist notion of psychoanalytic history (Danto, 2005). Freud's expressions of social conscience, arising during an epoch favoring radical social democratic ideals, led to the creation of cooperative mental health clinics throughout Europe (Danto, 2005: pp. 1–2). These same rich traditions are represented in the work of leading social work authors (e.g., Redl and Wineman, 1951, 1952; Ekstein, 1966; Fraiberg, 1987; more recently, Marans, 2005; Bragin, 2010).

In the United Kingdom and elsewhere, a different, though not unrelated, critique emerging in the 1970s argued against a longer term psychodynamic approach for cases, particularly those under public agency auspices, where little behavioral or other symptomatic improvement could be demonstrated over time. In such instances, psychodynamic treatment was felt to

constitute a waste of scarce resources (L. Dominelli, personal communication, 31 May 2013).

Psychoanalysis Is Antifeminist

In discussing Freud's views of feminine development and female sexuality, it should be acknowledged that Freud, himself, seemed less than fully satisfied with his understanding of women, whose sexual life he once described as a "dark continent for psychology" (1925: p. 212). In that essay, he observes that his findings about female sexuality are only valuable to the degree that they "… turn out to have general validity and to be typical" (p. 258). Later, when he returns to this topic (1933), he cautions readers, "If you want to know more about femininity, enquire from your own experiences of life, or turn to the poets, or wait until science can give you better answers" (p. 135).

A number of psychoanalysts, many of them women, began in the 1920s to challenge Freud's conclusions about the nature of femininity as well as the biological determinism in which these suppositions were rooted. Karen Horney (1939), for example, articulated a model of female development at fundamental variance with Freud's classical schema. Believing that female sexuality represented a 'primal, biological principle,' she hypothesized a basic male envy of the female capacity for reproduction, and disputed Freud's claim that the wish for a baby represented a compromise solution to the girl's inability to acquire a penis. Motherhood, she asserted, represented the psychological bedrock, with a corresponding biological principle, the instinct for procreation.

Horney and others gave increasing emphasis to environmental forces in shaping the personality, and in particular, the course of female development. This body of work came to represent the culturalist tradition in psychoanalysis. Horney's final position on the topic was that femininity, as a product of culture, does not as such exist (Buhle, 1998). What Freud had termed 'feminine masochism' – feelings of inferiority, dependency, indeed neurosis itself – was not an immutable, biological given, but rather culture-specific.

Other feminists, both academicians and psychoanalysts (Sherfey, 1966; Shainess, 1975; Mitchell, 1974; Chodorow, 1978) have made significant contributions to understanding femininity and female psychosexual development. Arguably, the most influential among these is Jean Baker Miller (1976). Baker Miller is credited with launching the self-in-relation model, a radical new paradigm for understanding women's development, emphasizing the original mother–daughter relationship, and relational phenomena such as affiliativeness, nurturance, empathy (vicarious introspection or the feeling of oneself into the experience of another. The capacity for empathy is regarded as a crucial *sine qua non* for healthy development of the self; chronic and sustained lapses or failures in (parental) empathy are believed responsible for various kinds of structural pathology), and compassion. Baker Miller believed that empathy, as a central organizer or structure of the female personality, explained a great deal about both normality and dysfunction in women's lives. While Freud's Oedipus illuminated aspects of male development, advocates of self-in-relation theory, following Horney, argue that an entirely new model is required to understand female development.

Psychoanalysis Pathologizes Clients

The broad appeal of the 'strengths perspective' to contemporary social work has been underscored by its rapid incorporation into not only our practice and scholarly literature, but also our professional lexicon. Emphasizing empowerment, resilience, healing, and wholeness, this perspective is defined by its abiding focus on clients' personal narratives and workers' responses to them. The worker's capacity to listen specifically for the client's strengths in surmounting adversity and misfortune paves the way for 'transformation and growth.' The work is always collaborative, 'possibility' rather than 'problem-focused,' and unburdened by a priori assumptions about limitations imposed on the client by trauma and abuse, handicaps, illness, and other personal struggles.

Although not always named specifically in these discussions, psychoanalysis is implicitly assumed to place undue emphasis on client weaknesses and pathology, and is thus regarded as deficit-based. However, both historical and more contemporary developments argue against the legitimacy of such claims. The first development is the gradual ascendancy of the 'two-person' model in psychoanalytic thought. Indeed, the assumption that psychoanalysts can function as neutral observers, optimizing their position to make systematic interpretations of the client's transference, has been largely discarded. Instead, most contemporary psychodynamic clinicians assume a reciprocal, mutually influential relationship to exist between therapist and client, and further, that such mutuality and intersubjectivity are central to therapeutic progress.

Additionally, there is good reason to question the general assumption that psychoanalysis places undue emphasis on client weaknesses or abnormalities. For example, Heinz Hartmann (1939) long ago proposed a 'conflict-free ego sphere' composed of functions such as attention, concentration, intention, language, recall, and motor development, all of which subserve adaptation. This view influenced others, such as Robert White (1959), who examined concepts such as effectance and competence, and Erik Erikson (1959), the first psychoanalytic researcher to chart the human life cycle. Such contributions highlight normative development and normal intrapsychic mechanisms and structures. In fact, the field of resiliency, which is closely allied with the strengths perspective, is historically associated with several well-known psychoanalytic researchers (e.g., Coles, 1986; Anthony and Cohler, 1987; Lifton, 1993).

Psychoanalysis does place equal importance on psychopathology, but unlike biological psychiatry, attempts to extrapolate dynamic meaning from pathological signs and symptoms, rather than treating them merely in descriptive terms. Is this the equivalent of pathologizing the client? Particularly in light of the contemporary psychoanalytic emphasis on mutuality and reciprocity within the treatment relationship, and the gradual abandonment of the search for 'historical truths' in psychoanalytic treatments, such a critique seems now to lack substance.

Psychoanalysis Does Not Focus on the Environment

Another common charge leveled at psychoanalytic theories and psychodynamic clinicians is that psychoanalysis is not sufficiently sensitized to the client's environmental surround. Using examples principally tied to early classical psychoanalytic formulations, these critics note that Freud devalued the contribution of environment (or 'nurture'), instead positing that endogenous factors claimed far greater importance in healthy and pathological development. Although Freud continued to emphasize the role of constitution in favor of environmental forces in his developmental theory, by 1923, he had assigned a far greater role to object relations.

The contributions of culturalists (Horney, 1939; Fromm-Reichmann,1950; Sullivan, 1953), ego psychologists (Hartmann, 1939; Erikson, 1959), and 'middle traditionalists' (Fairbairn, 1952; Winnicott, 1965) highlight important shifts in the role imputed to the extramural environment. Others (e.g., Spitz, 1965; Mahler, 1968), whose clinical work and research focused on infant–mother interactions and pathology, also figured prominently in evolving psychoanalytic conceptions of environment between the 1940s and the 1960s.

Winnicott was among the first psychoanalytic theorists to devote serious attention to the role of environment. His view, which may have presaged ideas associated with constructivism and intersubjectivity theory, is notable for its designation of a transitional realm of experience linking objective, reality-based experiences to exclusively subjective ones.

Others, too, have shaped the contemporary psychoanalytic perspective on the environment, including John Bowlby (1988), Hans Loewald (1962), Heinz Kohut (1977), Daniel Stern (1985), and Robert Emde (1992). Using systematic observational methods, researchers such as Stern and Emde have approached environmental phenomena in ways that are far less encumbered by theory than, e.g., Mahler's early studies.

More recent psychoanalytic formulations appear to have made progress in elucidating the role of environment in development as well as in the process of treatment, and there is mounting evidence both of greater interest in the person-social environment matrix among many analytic writers as well as a willingness to consider different theoretical positions (e.g., social constructivism) in the quest for improved understanding of environmental phenomena.

Psychoanalysis and Clinical Social Work: a Reappraisal

Much has changed in the psychoanalytic world since Freud's time. Multiple psychoanalytic psychologies, each with its own vision of human development, psychopathology, and the treatment process, now stand alongside of Freud's classical formulations. Challenges to the empirical basis of psychoanalytic treatment and its formulations about development and dysfunction have spawned a new and more promising body of research. The shift from a one-person to a two-person psychology, the gradual ascendancy of relational ideas, interest and sensitivity in the social environment, and the emerging emphasis on neuropsychoanalysis have likewise been momentous developments.

These are among the more significant changes marking the beginning of the second century of psychoanalytic thought. Social work, too, has changed markedly since the earliest attempts to infuse psychoanalytic ideas into clinical theory and practice. Although psychoanalytic ideas once held a revered status in clinical social work, this has been altered, inter alia, by the introduction of newer theories of change, particularly nondynamic approaches, the shifting requirements of the agencies and clients we serve, and an overriding focus on short-term engagement.

The psychoanalysis of today, while preserving important traditions spanning over 100 years, is a far more sophisticated, complex, and scientifically elaborated body of knowledge. Indeed, we believe that these developments argue convincingly for a reappraisal of what contemporary psychoanalytic theory and practice knowledge have to offer the field of clinical social work.

See also: Afrocentric Approaches to Social Work; Classical Conditioning Methods in Psychotherapy; Constructivist Theory and Psychotherapy; Counseling and Psychotherapy: Ethnic and Cultural Differences; Diagnostic Processes in Clinical Psychology and Psychotherapy; Evidence-Based Psychotherapy; Experiential Psychotherapy; Feminist Psychotherapy; Group Psychotherapy, Clinical Psychology of; Imaging Methods in Psychotherapy; Interpersonal Psychotherapy; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Managed Care and Psychotherapy in the United States; Psychoanalysis in Sociology; Psychoanalysis: Current Status; Psychotherapy Integration; Psychotherapy Process Research; Psychotherapy and Pharmacotherapy; Virtual Reality in Psychotherapy; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Anthony, E., Cohler, B., 1987. The Invulnerable Child. Guilford, New York.
Benjamin, J., 1988. The Bonds of Love: Psychoanalysis, Feminism and the Problem of Domination. Pantheon, New York.
Bowlby, J., 1988. A Secure Base: Clinical Applications of Attachment Theory. Routledge, London.
Bragin, M., 2010. Clinical social work in situations of disaster and terrorism. In: Brandell, J. (Ed.), Theory and Practice in Clinical Social Work, second ed. Sage, Los Angeles, pp. 373–406.
Brandell, J., 2004. Psychoanalytic Social Work. Columbia University Press, New York.
Brandell, J., Ringel, S., 2007. Attachment and Dynamic Practice: An Integrative Guide for Social Workers and Other Clinicians. Columbia University Press, New York.
Buhle, M., 1998. Feminism and its Discontents: A Century of Struggle with Psychoanalysis. Harvard University Press, Cambridge, MA.
Chodorow, N., 1978. The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender. University of California Press, Berkeley.
Coles, R., 1986. The Political Life of Children. Houghton-Mifflin, Boston.
Danto, E., 2005. Freud's Free Clinics. Columbia, New York.
Danto, E., 2009. A new sort of salvation army: historical perspectives on the confluence of psychoanalysis and social work. Clinical Social Work Journal 37 (1), 165–184.
Ekstein, R., 1966. Children of Time and Space, of Action and Impulse: Clinical Studies on the Psychoanalytic Treatment of Severely Disturbed Children. Appleton-Century-Crofts, New York.
Emde, R., 1992. Social referencing research: uncertainty, self, and the search for meaning. In: Feinman, S. (Ed.), Social Referencing and the Construction of Reality in Infancy. Plenum, New York, pp. 79–94.

Erikson, E., 1959. Identity and the Life Cycle. In: Selected Papers, Psychological Issues, vol. 1. International Universities Press, New York.
Fairbairn, W., 1952. An Object Relations Theory of the Personality. Basic, New York.
Fink, B., 1997. A Clinical Introduction to Lacanian Psychoanalysis: Theory and Technique. Harvard University Press, Cambridge, MA.
Fonagy, P., Target, M., 1994. The efficacy of psychoanalysis for children with disruptive disorders. Journal of the American Academy of Child and Adolescent Psychiatry 33, 45–55.
Fonagy, P., Target, M., 1996. Prediction of the outcome of child psychoanalysis: a retrospective study of 763 cases at the Anna Freud Center. Journal of the American Psychoanalytic Association 44, 27–77.
Fraiberg, S., 1987. Collected Writings of Selma Fraiberg. Ohio State University Press, Columbus, OH.
Freud, S., 1925. Some psychical consequences of the anatomical distinction between the sexes. Standard Edition 19, 241–258.
Freud, S., 1933. New introductory lectures. Standard Edition 22, 1–182.
Fromm-Reichmann, F., 1950. Principles of Intensive Psychotherapy. University of Chicago Press, Chicago.
Galazter-Levy, R., Bachrach, H., Skolnikoff, A., Waldron, S., 2000. Does Psychoanalysis Work? Yale, New Haven.
Gay, P., 1988. Freud: A Life for Our Time. Norton, New York.
Goldstein, E., 2002. Psychoanalysis and social work: historical perspectives. Psychoanalytic Social Work 9, 33–40.
Hartmann, H., 1939. Ego Psychology and the Problem of Adaptation. International Universities Press, New York.
Hollway, W., 2006. The Capacity to Care: Gender and Ethical Subjectivity. Routledge, London.
Holzman, P., 1998. Psychoanalysis and Psychopathology. Aronson, Northvale, NJ.
Horney, K., 1939. New Ways in Psychoanalysis. Norton, New York.
Kernberg, O., 1992. Aggression in Personality Disorders and Perversion. Yale University Press, New Haven, CT.
Kohut, H., 1977. The Restoration of the Self. International Universities Press, New York.
Kohut, H., Wolf, E., 1978. The disorders of the self and their treatment: an outline. International Journal of Psychoanalysis 59, 413–425.
Lacan, J., 2007. Ecrits: The First Complete Edition in English. W.W. Norton & Company, New York.
Lifton, R., 1993. The Protean Self: Human Resilience in an Age of Fragmentation. University of Chicago Press, Chicago.
Loewald, H., 1962. Internalization, separation, mourning, and the superego. Psychoanalytic Quarterly 31, 483–504.
Mahler, M., 1968. On Human Symbiosis and the Vicissitudes of Individuation. International Universities Press, New York.
Malekpour, M., 2007. Effects of attachment on early and later development. The British Journal of Developmental Disabilities 53 (2), 81–95.
Marans, S., 2005. Listening to Fear: Helping Kids Cope from Nightmares to the Nightly News. Holt, New York.
Masson, J., 1994. The Assault on Truth: Freud's Suppression of Seduction Theory. Pocket, New York.
Miller, J. Baker, 1976. Toward a New Psychology of Women. Beacon, Boston.
Mitchell, J., 1974. Psychoanalysis and Feminism: Freud, Reich, Laing, and Women. Pantheon, New York.
Moore, B., Fine, B. (Eds.), 1990. Psychoanalytic Terms and Concepts. Yale, New Haven.
Perlman, H.H., 1957. Social Casework: A Problem-solving Process: A Fresh, Unifying Approach to Casework Practice. University of Chicago Press, Chicago.
Qazi, K.A., 2011. Lacanian concepts – their relevance to literary analysis and interpretation: a post structural reading. The Criterion: An International Journal in English 2 (4), 1–12 viewed 01.02.13. http://www.the-criterion.com/V2/n4/Qazi.pdf.
Redl, F., Wineman, D., 1951. Children Who Hate. Free Press, Glencoe, IL.
Redl, F., Wineman, D., 1952. Controls from within: Techniques for the Treatment of the Aggressive Child. Free Press, New York.
Sarup, M., 1993. An Introductory Guide to Post-structuralism and Postmodernism, second ed. The University of Georgia Press, Athens, GA.
Shainess, N., 1975. Authentic feminine orgasmic response. In: Adelson, E. (Ed.), Sexuality and Psychoanalysis. Brunner/Mazel, New York, pp. 145–160.
Shedler, J., 2010. The efficacy of psychodynamic psychotherapy. American Psychologist 65 (2), 98–109.
Sherfey, M., 1966. The nature and evolution of female sexuality in relation to psychoanalytic theory. Journal of the American Psychoanalytic Association 14, 28–128.
Specht, H., Courtney, M., 1997. Unfaithful Angels: How Social Work Has Abandoned its Mission. Simon & Schuster, New York.
Spitz, R., 1965. The First Year of Life. International Universities Press, New York.
Stern, D., 1985. The Interpersonal World of the Infant. Basic, New York.
Sullivan, H.S., 1953. The Interpersonal Theory of Psychiatry. Norton, New York.
Sulloway, F., 1979. Freud, Biologist of the Mind: Beyond the Psychoanalytic Legend. Basic, New York.
Target, M., Fonagy, P., 1994. Efficacy of psychoanalysis for children with emotional disorders. Journal of the American Academy of Child and Adolescent Psychiatry 33, 361–371.
Waddell, M., 2002. Inside Lives: Psychoanalysis and the Growth of the Personality. Karnac, London.
White, R., 1959. Motivation reconsidered: the concept of competence. Psychological Review 66, 297–333.
Winnicott, D.W., 1965. The Maturational Process and the Facilitating Environment. International Universities Press, New York.
Zizek, S., 2006. How to Read Lacan. W.W. Norton & Company, New York.

Relevant Websites

http://www.gaps.org.uk/articles/psychodynamic (accessed 10.06.13.) – gaps - What is a psychodynamic or psychosocial approach in social work?

http://www.google.com/search?q=interactive+websites+for+psychoanalytic+theory+in+social+work&rls=com.microsoft:en-us&ie=UTF-8&oe=UTF-8&startIndex=&startPage=1 (accessed 10.06.13.).

http://socialworkpodcast.blogspot.com/2012/06/psychodynamic-therapy-with-vulnerable.html (accessed 10.06.13.) – The Social Work Podcast - Psychodynamic therapy with vulnerable and oppressed populations: Interview with Joan Berzoff M.S.W., Ed.D.

Social Work and Refugees

Mary Nash, Massey University, Palmerston North, New Zealand; and Aotearoa New Zealand Association of Social Workers, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article uses recent literature and research to illuminate knowledge, values, and skills that social workers require to work with refugees, an emergent field of practice. Literature is drawn from books, journal articles, and online resources including relevant national policy documents. A change of tone is discerned in recent literature, in the form of discourse suggesting that refugees have become less deserving of the protection of their human rights. Their encounters with racism and xenophobia are noted. Dilemmas faced by social workers to fulfill their statutory functions while holding true to their professional principles of social justice are discussed.

Introduction

This article focuses on social work with refugees (rather than asylum seekers), using recent literature and research to illuminate the knowledge, values, and skills that social workers require when engaging in this emergent field of practice. The dilemmas faced by social workers to fulfill their statutory functions while holding true to their professional principles of social justice will be discussed.

Social work is operating in an increasingly international and global environment (Chathapuram and Link, 1999; Midgley, 2000; Powell, 2001). This is particularly apparent when working with immigrants, refugees, and asylum seekers. Literature on globalization and social work explicitly links social work practice to issues of citizenship, inclusion, and participation in civil society. Social workers are inevitably called upon to work in this context, either as agents of the state or in the voluntary sector as members of nongovernmental organizations or not-for-profit organizations. In nonstatutory settings, they are more likely to be able to express their professional principles of social justice. The International Federation of Social Workers (IFSW) reminds us that "The issues surrounding refugees are matters of justice and peace. Social workers in the vast majority of countries will be required to address the problems facing refugees in their practice" (IFSW, 2012: p. 1).

Potocky-Tripodi (2002: p. 3) argued that "Social work practice with refugees and immigrants requires specialized knowledge of the unique issues of these populations." Her comprehensive account of best practice in this new and challenging field is consistent with the IFSW position. She systematically explored the many factors, which practitioners need to recognize in the lives of this client group, using an analysis that reflects on the macro-, meso-, and microlevels of social work practice. Good practitioners, she observed, should be informed about human rights and social justice issues as well as international and local law surrounding immigrants and refugees. Nash and Trlin (2004: p. 2), reporting on findings from two postal surveys in New Zealand, found that practitioners need to be informed regarding service delivery systems and to be knowledgeable regarding key areas such as physical health, mental health, family dynamics, cultural diversity, language, education, and economic circumstances. Social workers also need to be culturally competent and have the requisite knowledge and skills to work appropriately with their immigrant and/or refugee clients. These points will be elaborated on throughout this article.

Definitions

People living away from their country of origin do so for a variety of reasons, sometimes voluntary and sometimes involuntary. When describing the latter, terms may be generic, such as involuntary migrant or more specific, from asylum seeker to convention refugee, or refugees with exceptional leave to remain, or even refugees in orbit (Joly et al., 1997) depending on country of origin and host country.

An immigrant is someone who moves in a planned way, from one country to another, in order to settle permanently in the new country. Immigrants know that they may, if they choose, return to their country of origin. This is not the case for refugees.

According to the United Nations 1951 Convention on the Status of Refugees, a refugee is a person who (http://www.unhcr.org/pages/49c3646c125.html 06.08.13):

> Owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country.

Asylum seekers, in contrast to immigrants and refugees, are people who enter a country without legal documents, or whose documents expire once they have arrived and who claim refugee status. An asylum seeker is a person who has asked for protection but has not received a decision on their application to become a refugee, or is waiting for the outcome of an appeal (SCIE, 2010). Asylum seekers have the right under the United Nations High Commissioner for Rights (UNHCR) Convention to seek a fair hearing to decide on their status as a refugee, but until that hearing, they may demand very little from the country they live in. It is essential that social workers should distinguish between the legal status of asylum seekers and refugees when working with migrants, because, in all countries, people will be entitled to different forms of public welfare support, depending on their legal immigration status. For this reason, using legal definitions

correctly is fundamental, and finding out how clients are placed will assist in understanding their help-seeking behavior and aspirations for settlement (Potocky-Tripodi, 2002: p. 13).

Useful information covering the demographics of refugees, asylum seekers, and displaced persons is available on the UNHCR Web site. The refugees of concern to UNHCR are spread around the world, with more than half in Asia and some 20% in Africa. They live in widely varying conditions, from well-established camps and collective centers to makeshift shelters or living in the open. It is estimated that there are about 19.5 million refugees in the world (http://www.unhcr.org/pages/49c3646c1d.html 07.03.13).

This article relates to social work with refugees, and this term is used with care to avoid confusion with social work with asylum seekers (see asylum seekers article) or migrants. Convention refugees are included here, but efforts have been made to refer to them in the context of a postasylum seeker category.

Literature Review, Introduction

Research and publications about refugees and social work have recently increased. As refugees from a variety of countries and ethnicities have faced displacement, trauma, and loss, there are resettlement opportunities for some. Resettlement calls for courage and resilience and may involve unforeseen difficulties such as stigma, unemployment or underemployment, unwelcome climatic change, new languages, strange housing, cultural differences, racism, and perhaps the most important of all, loss of family members.

Literature reviewed in this article is drawn largely from books, journal articles, and online resources such as relevant national policy documents. Journal articles date predominantly from the year 2000 to the present, and the journal *International Social Work* is one of the main resources. Material from the UNHCR Web site has also been helpful. This discussion of research themes and sources is designed to point up consensus and diversity, recognizing the inevitable ethical dilemmas in international social work and the range of practices within the parameters of different welfare states (Dominelli, 2004; Hugman et al., 2010). It begins with a brief historical approach, and records some of the political responses to refugee populations in nation-states which have, over time, developed very different welfare regimes (Sainsbury, 2006).

A change of tone can be discerned in the literature, in the form of a discourse suggesting that asylum seekers and refugees have become somehow less deserving of the protection of their human rights. Instead, they come under suspicion and their encounters with racism and xenophobia are noted in recent research (Masocha and Simpson, 2011; Refugee Council, 2004). The status of social work, its ethical values, and the expectations social workers face as they carry out their work are a particularly challenging context for practice.

Literature Review, Historical and Political Contexts

The 1951 Geneva Convention on the status of refugees defined refugees in relation to events that took place before 1951, which is why, in 1967, a protocol was added to the Convention to include people who were unable or unwilling to return to their country of origin post-1951 (Lacroix, 2006). Lacroix made a helpful observation, which underpins recent thinking about the uneasy relationship between universal human rights and the asylum/refugee responses instituted in individual nation-states. She writes:

> States which have adhered to the Convention have instituted their own refugee regimes, according to their own historical, political, ideological and social relationships to non-citizens. Hence the disparity among refugee determination systems across the western world and conflicting national views on what constitutes a refugee.
> Lacroix, 2006: p. 21

Sainsbury's (2006) comparative analysis of the social rights of immigrants in three countries exemplified three different types of welfare regime. She described the United States as having a liberal welfare regime, with an inclusive immigration policy in which rights are based on need and land of birth (*ius soli*). Germany, she described as having a conservative welfare regime, mediated by an exclusionary immigration policy such that rights are based on work and lineage (*ius sanguinis*). The third welfare category is exemplified by Sweden, with its social democratic welfare regime according to which rights are based on citizenship and residence (*ius domicilii*) (Sainsbury, 2006: p. 231). These three welfare regimes, each with its own criteria for negotiating rights, provide a useful schema for understanding ways in which different nation-states have interpreted universal rights legislation and developed and applied national policies for immigrants, especially asylum seekers and refugees. Both Hugman et al. (2010) and Pirjola (2009) pick up on this analysis to explore its implications for the uneasy fit between international and national human rights discourses. Hugman et al. apply it to the predicament in which social workers now find themselves, noting "the contradiction that lies between the universal rights of all human beings to live lives of dignity and the exclusionary nation-based notions of social citizenship" (Hugman et al., 2010: pp. 637–638).

The professionalization of social work cannot occur in isolation from increasing demands and expectations that social workers and colleagues face in their daily practice (Maidment, 2013). In 2001, the IFSW and International Association of Schools of Social Work jointly defined social work as "… social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work" (IFSW, 2012: p. 1). Qualified social workers are, therefore, expected to know about human rights legislation as it relates to refugees and to understand government policies, which provide the legal context for their practice in this field. They need to be aware of international agreements, conventions, and guiding statements that concern the status of refugees to become competent advocates for them.

Literature Review, Electronic Resources Informing Macro- and Micropractice

It makes sense when researching an article for an electronic encyclopedia, to identify relevant Web sites and electronic resources belonging to a range of agencies and organizations including the United Nations and other international

organizations, with selected statements and conventions on human rights, as well as examples of national legislation, government policies and social work practice standards, and refugee voices.

The IFSW Web site publicizes a range of policy statements for best practice including a standard setting policy statement on refugees (http://ifsw.org/policies/refugees/ retrieved 14.04.13). It emphasizes the need for a holistic approach when working with refugees, caring not simply for material, but also for psychosocial needs. Ethical values underpin this work, imbued with a belief in the intrinsic value of each human being and the rights of the individual to seek asylum and to be treated with respect for their human dignity. "… the social work profession accepts its share of responsibility for responding to the distress of refugees, and strives for the fullest possible involvement of refugees themselves in meeting their needs" (http://ifsw.org/policies/refugees/ retrieved 14.04.13: p. 1). Key areas of knowledge are outlined and a measure of the success of intervention is seen to be the provision of 'durable solutions' to the problems faced by refugees, together with "the achievement of self-sufficiency, economic independence, spiritual and intellectual fulfillment" (http://ifsw.org/policies/refugees/ retrieved 14.04.13: p. 1). The IFSW explicitly supports the work of the United Nations and believes in the value of working in partnership with other NGO organizations to lobby for social and political changes in this field.

The Office of the UNHCR was established on 14 December 1950 by the United Nations General Assembly. The agency is mandated to lead and coordinate international action to protect refugees and resolve refugee problems worldwide. Its primary purpose is to safeguard the rights and well-being of refugees (www.unhcr.org/). The site links directly to the Geneva Convention 1951, and includes a brief history of the UNHCR and describes the main areas of work it undertakes. The 'Figures at a glance' page is particularly helpful.

Human rights legislation is a fundamental resource for social workers when advocating for their clients. In the context of refugee work, as argued by Hugman et al. (2010), it provides a powerful and essential counterbalance to national or regional legislation and policies, which would seek to exclude or marginalize asylum seekers and refugees. In addition to the obvious sections on human rights, this site has good links to other topics including an interesting and innovative link to the UN Cyberschoolbus with its interactive declaration of human rights and moving stories by children whose rights have been abused (http://www.un.org/cyberschoolbus/humanrights/index.asp). Easily navigated, the Cyberschoolbus is definitely worth boarding and links one to the Declaration of the Rights of the Child and another to the Universal Declaration of Human Rights.

Two national organizations featured in this article are concerned either to set standards for good social work practice, or to see that social workers have resources for achieving the standards expected of them when working with refugees. The (American) Council on Social Work Education (CSWE) makes a pertinent, but generic statement that social workers

- understand the forms and mechanisms of oppression and discrimination;
- advocate for human rights and social and economic justice; and
- engage in practices that advance social and economic justice.

(CSWE Educational Policy and Accreditation standards, Educational Policy 2.1.5 www.cswe.org/). There is no direct mention of refugees as such here, but the standard is clearly relevant.

The (English) Social Care Institute for Excellence features a comprehensive guide to good practice with refugees and asylum seekers based on an extensive review of British research publications and a practice survey. Good practice occurs in a context of respect for the strengths and resilience of refugees and asylum seekers. Emphasizing the importance of good information and advocacy skills, this report is a very helpful resource for practitioners (SCIE, 2010).

Finally, a research project featuring dialogues with women refugees in Finland offers a politically aware, gendered analysis, with practical recommendations for ways of working with refugee women (Pittaway and Van Genderen Stort, 2011). The women participants' strongest message was that service provision aimed at assisting refugee integration has to be grounded in a deep understanding of the impact of these experiences on their lives. Many of the women commented that racist and xenophobic behavior and language served to keep the horror of these experiences alive (Pittaway and Van Genderen Stort, 2011: section 1, p. 7). This report models excellent and innovative research methods, provides many links to other research together with a set of excellent recommendations for social workers in this field.

The international human rights and social work agencies and organizations provide a global setting for principled social work practice with refugees. The few chosen to feature here have offered a range of information and links featuring practice standards, research, and a gendered approach which, in this field, is essential. The next section looks at core concepts and theoretical frameworks for practice.

Core Theories and Concepts

A number of theoretical frameworks and models exist to support the kinds of analysis that social workers will find useful when working in this field. Kunz (1981: pp. 49–51) categorized three types of refugee according to their reasons for leaving their countries of origin, reasons which he predicted would affect their resettlement experiences. The three different groups he recognized are 'majority-identified' refugees who identify with their homeland but not with its government; 'events-alienated' refugees, often religious or racial minorities who rarely entertain the hope of returning home; and 'self-alienated' refugees, who for various reasons no longer wish to identify with their nation.

George (2002) categorized models of settlement service delivery as theory based and practice based. Theory-based models include models of cultural competence, antiracist models, ecological models, and empowerment approaches. Practice-based models focus on the different stages or phases of immigrant settlement and adaptation.

Perhaps the most used theory-based model is the ecological model, which social workers are familiar with in many contexts. It ensures that the practitioner takes a holistic approach to their interventions, contextualizing refugees in international and national settings. The importance of a macrolevel analysis for working with refugees was eloquently

expressed by a respondent in a New Zealand research project who commented that:

> You have got to be able to have a structural analysis, you have got to be able to have a global analysis, you have to have political reasons … Because in this sector…if you only have case work skills, you only have … a micro-base approach to your work, it is not going to work, it will be very soul destroying.
> Nash et al., 2006: p. 350

Also within George's theory-based category, empowerment approaches can build on strengths-based practice (Saleebey, 2006) as well as on the social justice principles of solidarity and subsidiarity. Valtonen described social work in Finland as developing a new field of practice, which involves indirect social work and provides a 'pivotal link' (2001: p. 959) as social workers use their expertise to liaise skillfully between the clients' world, bureaucracies, and local community groups. Her study emphasized professional social work responsibilities for humanitarian and social justice work in this domain and identified key skills for social workers – in particular, cross-cultural skills.

George's (2002) second category of models of settlement service delivery, the practice-based model, promotes an understanding of the stages of migration and points up the challenges refugees have endured to arrive at a particular point. If social workers are to work successfully with them to facilitate settlement into a new country, this is also something they need to be aware of. George cited Cox's (1985, 1987) example of the stages approach involving premovement, transition, settlement, and integration, as well as a continuum model in which newcomers move from acclimatization, followed by adaptation and leading to integration (George and Fuller-Thomson, 1997). The stages approach has been taken up by social workers as a useful framework, which emphasizes matters of significance to the people they work with. Practice-based models also distinguish between services run by ethnic communities for their own people and services run by mainstream agencies for newcomers (George, 2002).

These frameworks are helpful in that they identify and clarify a variety of complex variables involved in settlement work with refugees. Many of the issues which these models and frameworks are designed to address are reflected in subsequent research by Nash and Trlin (2004) and Ager and Strang (2004). The latter developed an evaluation tool, 'Indicators of Integration,' based on research funded by the Home Office to develop a framework that would, without being prescriptive, facilitate the planning and evaluation of local projects designed to assist refugee settlement.

From the research, a framework for the indicators is structured into 10 domains, which represent the key factors involved in the process of integration work with refugees. These domains are categorized under four headings and within each domain there are further indicators that provide criteria for assessing integration pertaining to that particular domain (Ager and Strang, 2004: p. 5). The framework, while informed by research findings, is also intuitively perceptive for anyone who has worked in this field.

The four headings are as follows: The first, 'means and markers,' which refer to the four domains that are essential for settlement leading to integration. These are housing, employment, education, and health. These domains are self-explanatory, and without achievement within each of them, refugees cannot be said to have settled into their new community. The second heading is 'social connections' and it refers to social bonds, social bridges, and social links. These links are reminiscent of the ecological framework; in that, social bonds are the connections within a specific (microlevel) community, ethnic, religious, or other, while social bridges are the connections made with the members of external and different types of communities (mesolevel). Social links are (macrolevel) connections made with institutions at local or national levels and include state services. The social domains link well to the work of Valtonen (2001) referred to above.

The third heading 'facilitators' includes two domains, which again are predictably identified as 'language and cultural knowledge' and 'safety and stability.' Again, these domains are self-explanatory. The fourth heading 'foundation' as indicated by its name is a particularly important one, and contains a single domain, which is 'rights and citizenship.' The authors have included helpful examples of how the domains can be used at practice and policy levels, making them user-friendly for the busy practitioner or policy-maker.

This carefully structured framework can be used when presenting research into social work roles and practices. The author has found that recurring themes emerge from research into social work with refugees across time and place. Findings consistently indicate that refugees experience similar needs and requirements and that practitioners describe comparable levels of knowledge, resources, and interventions in approaching their work.

Social Work Roles and Practices

For example, George (2002) reported on research carried out in Canada to develop an appropriate model for settlement services designed to meet the needs identified by the service users themselves, as well as by social workers. She identified "full and equal participation of newcomers in Canadian society" (George, 2002: p. 468) as the aim of the Canadian Council for Refugees. As mentioned earlier, participation continues to rank high on the list of essential components of a successful settlement strategy. George (2002) distinguished two ways in which one can view settlement: first, as a two-way process in which both newcomers and the host society have to adapt to one another; and second, as the provision of settlement services, which can be seen as a duty based on human rights (George, 2002: p. 468). This fits well with Ager and Strang's (2004) 'rights and citizenship' domain and accords with the work of Lacroix (2006: p. 22), when she observes that "How we start linking the international context to local experiences is of particular concern for the social justice side of social work."

George (2002) noted that the Canadian Council for Refugees identified values such as inclusivity, empowerment, respect for newcomers, cultural sensitivity, collaboration, and reliability of services as important in resettlement work. These values reflect those expressed in the United Nations Summit for Social

Development – Ten Commitments (cited by Healy, 2001: p. 288). They are values that every refugee recognizes as essential for their well-being and dignity. A later research project, sponsored by the UNHCR and the University of New South Wales, Centre for Refugee Research, Australia (mentioned above), used a participatory approach to collect the stories and opinions of women refugees in Finland, concerning the stages of their lives as refugees. The research team set out to produce a model for government and civil society to use when working with refugees toward their settlement, with particular emphasis on women and children. The women "consistently recommended that they be involved in solution identification and implementation, and that their previous expertise and knowledge be utilized to assist their own communities" (Pittaway and Van Genderen Stort, 2011: p. 11).

This very useful research project has good detail and stresses some psychological needs for recognition by host workers and society, of traumas and hardships undergone before arrival in host country. Ten global core protection areas proposed by the UNHCR Gender Unit have been identified by UNHCR research using the Regional Dialogues method in India, Colombia, Jordan, Uganda, Zambia, and Thailand. Finland adds an industrial nation-state to the mix. In this set of dialogues, the women talked particularly of their experiences of seeking asylum and settling into Finland. The report compares the global core protection issues and those for Finland and shows a close comparison between them

Global protection issues	*Finnish protection issues*
Individual documentation	Individual documentation
Women's representation in leadership	Women in leadership positions
Sexual and gender-based violence	Sexual and gender-based violence
Shelter	Housing
Sanitary materials	Reception conditions
Health issues	Health issues
Legal remedies	Access to justice
Education	Education
Economic self-reliance	Economic self-reliance
General violence	Racism and xenophobia

Section 1.5, p. 9.

Particular issues of concern arising from the research were:

Language
Employment
Racism, xenophobia in host country
Increased domestic violence and reasons for this
Length of asylum application procedures
Lack of access to counseling to enable them to process the traumas they have experienced.

Section 1.6, p. 10 of 48.

They stressed also the importance of better policies for family reunification. Again, one notes the same themes coming through as from other research.

A small, supplementary selection of authors whose research is particularly relevant to social work intervention in this field includes Weaver and Burns (2001) on understanding the traumatic experiences faced by refugees and asylum seekers; Nash and Trlin (2004) who explored the experiences of generic social workers and NGO workers involved in settlement work with clients who were refugees or migrants in New Zealand; Hagelund, and Kavli (2009) on the importance of employment for refugees when they begin the arduous process of settlement in a new country; Al-Qdah and Lacroix (2011) on Iraqi refugees in Jordan with useful practical ideas; Kreitzer (2002) on the participation and planning of refugee camp programs by Liberian refugee women; and finally, not to leave out the significant but often neglected area of religion, Benson et al. (2012) have researched a study of newly resettled Hindu Bhutanese refugees in the United States.

Ethical Challenges

There remains one challenging matter, raised forcefully by Hugman et al. (2010) and Humphries (2004) who, in different ways, remind social workers of the dilemmas they face in upholding their professional ideals when working in this field. Humphries (2004: p. 93) argues that in relation to implementing immigration policy, "There is no clearer an example of the transformation of social work from a concern with welfare to a position of authoritarianism than in this field, paralleling a more general change in the profession towards a culture of blame and enforcement." In the context of international social work, Hugman et al. (2010: p. 633) ask: "Are social workers engaged in international social work exerting effective political pressure (within their own nation-states) behind their desire for a more equitable global society?" or "is a borderless social work likely to be egalitarian or another nightmare of (inappropriately) universalizing language?" Social workers engaged in settlement programs cannot avoid these questions, for they will inexorably be involved in advocacy, in family reunification issues, leading to international social work with refugees and asylum seekers.

Concluding Remarks

This article offers a necessarily brief guide to the terrain of social work with refugees. It began with defining terms, as a preliminary to looking at international conventions, and statements by human rights and global social work organizations. These are referenced using accessible Web sites, to enable readers to deepen their knowledge of the histories, politics, and values that inform the refugee world. With this foundation material, the author has used a macro-, meso-, microframework to organize a presentation of useful theoretical material at each of these levels, material which practitioners need to be aware of when working with refugees. Finally, research into aspects of refugee settlement in diverse countries has provided a selection of examples of social work practice and refugee experience.

Above all, this article is intended to provoke curiosity and to raise some of the ethical and professional issues that make social work so interesting.

See also: Green Social Work; Human Rights and Social Work; Refugees in Anthropology; Refugees: Geographical Aspects; Whiteness and White Supremacy.

Bibliography

Ager, A., Strang, A., 2004. Indicators of Refugee Integration: Final Report. Report to the Home Office Immigration Research and Statistics Service. http://www.homeoffice.gov.uk/rds/pdfs04/dpr28.pdf (accessed 19.04.13.).

Al-Qdah, T., Lacroix, M., 2011. Iraqi refugees in Jordan: lessons for practice with refugees globally. International Social Work 54, 521–534.

Benson, G.O., Sun, F., Hodge, D.R., Androff, D.K., 2012. Religious coping and acculturation stress among Hindu Bhutanese: a study of newly-resettled refugees in the United States. International Social Work 55, 538–553.

Cemlyn, S.J., Nye, M., 2012. Asylum seeker young people: social work value conflicts in negotiating age assessment in the UK. International Social Work 55 (5), 675–688.

Chathapuram, S.R., Link, R.J., 1999. All Our Futures: Principles and Resources for Social Work Practice in a Global Era. Wadsworth Publishing Company, Belmont.

Cox, D., 1985. Welfare services for migrants: can they be better planned? International Migration 23, 73–94.

Cox, D., 1987. Migration and Welfare: An Australian Perspective. Prentice Hall, New York.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge.

General Social Care Council, 2011. Codes of Practice for Social Care Workers. Available at: http://www.gscc.org.uk/cmsFiles/CodesofPracticeforSocialCareWorkers.pdf (accessed 04.11.11.).

George, U., 2002. A needs-based model for settlement service delivery for newcomers to Canada. International Social Work 45 (4), 465–480.

George, U., Fuller-Thomson, E., 1997. To stay or not to stay: characteristics associated with newcomers planning to remain in Canada. Canadian Journal of Regional Science 20 (1 and 2), 181–194.

Hagelund, A., Kavli, H., 2009. If work is out of sight. Activation and citizenship for new refugees. Journal of European Social Policy 19 (3), 259–270.

Healy, L., 2001. International Social Work. Professional Action in an Interdependent World. Oxford University Press, New York, Oxford.

Hugman, R., Moosa-Mitha, M., Moyo, O., 2010. Towards a borderless social work: reconsidering notions of international social work. Journal of International Social Work 53 (5), 629–643.

Humphries, B., 2004. An unacceptable role for social work: implementing immigration policy. British Journal of Social Work 34, 93–107.

IFSW, 2012. Refugees Policy Issue. IFSW General Meeting, Hong Kong, 21–23 July 1998. http://ifsw.org/policies/refugees/ (accessed 06.08.13.).

IFSW, 2013. Definition of Social Work. http://ifsw.org/policies/definition-of-social-work/ (accessed 06.08.13.).

Joly, D., Kelly, L., Nettleton, C., 1997. Refugees in Europe: The Hostile New Agenda. Minority Rights Group, London.

Kreitzer, L., 2002. Liberian refugee women: a qualitative study of the participation in planning camp programmes. International Social Work 45, 45–58.

Lacroix, M., 2006. Social work with asylum seekers in Canada: the case for social justice. International Social Work 49 (1), 19–28.

Maidment, J., 2013. ANZASW President's Report. Notice board Issue 10. anzasw.org.nz (accessed 15.04.13.).

Masocha, S., Simpson, M., 2011. Xenoracism: towards a critical understanding of the construction of asylum seekers and its implications for social work practice. Practice 23 (1), 5–18.

Midgley, J., July 2000. Social work and globalization. Social Work, 13–28.

Nash, M., Trlin, A., 2004. Social Work with Immigrants, Refugees and Asylum Seekers in New Zealand. New Settlers Programme School of Sociology. Social Policy and Social Work, Massey University, Palmerston North.

Nash, M., Trlin, A., 2006. A Survey of Non-government/not for Profit Agencies and Organizations Providing Social Services to Immigrants and Refugees in New Zealand. New Settlers Programme School of Sociology, Social Policy and Social Work. Massey University, Palmerston North.

Nash, M., Wong, J., Trlin, A., 2006. Civic and social integration: a new field of social work practice with immigrants, refugees and asylum seekers in New Zealand. International Social Work 49 (3), 345–363.

Office of the High Commissioner for Human Rights. http://www.ohchr.org/EN/UDHR/ (accessed 13.04.13.).

Pirjola, J., 2009. European asylum policy: inclusions and exclusions under the language of universal human rights language. European Journal of Migration and Law 11 (4), 347–366.

Pittaway, E., Van Genderen Stort, A., 2011. Protectors, Providers, Survivors. A Dialogue with Refugee Women in Finland. United Nations High Commissioner for Refugees, Centre for Refugee Research, University of New South Wales, Australia. www.unhcr.org/4ec3d7606.pdf (accessed 03.04.13.).

Potocky-Tripodi, M., 2002. Best Practices for Social Work with Refugees and Migrants. Columbia University Press, New York.

Powell, F.W., 2001. The Politics of Social Work. Sage Publications, London.

Refugee Council, 2004. Hungry and Homeless: The Impact of the Withdrawal of State Support on Asylum Seekers, Refugee Communities and the Voluntary Sector. Refugee Council, London.

Saleebey, D., 2006. The Strengths Perspective in Social Work Practice, fifth ed. Pearson Education, Australia.

Sainsbury, D., 2006. Immigrants' social rights in comparative perspective: welfare regimes, forms in immigration and immigration policy regimes. Journal of European Social Policy 16 (3), 229–244.

SCIE, June 2010. Guide 37. Good Practice in Social Care for Refugees and Asylum Seekers. http://www.scie.org.uk/publications/guides/guide37/index.asp (accessed 03.04.13.).

Social Care Institute for Excellence, 2010. Good Practice in Social Care for Refugees and Asylum Seekers. SCIE, London. http://search02.scie.org.uk/?q=refugees (accessed 15.04.13.).

Social Work Action Network, 2011. The Social Work Action Network. http://www.socialworkfuture.org/ (accessed 04.11.11.).

UNHCR, 20–22 June 2001. Respect our rights: partnership for equality. In: Report on the Dialogue with Refugee Women. Geneva, Switzerland. http://www.unhcr.org/refworld/docid/4649d5d72.html (accessed 15.04.13.).

UNHCR, Executive Committee, 2006. Conclusion on women and girls at risk. In: 56th Session, 6 October 2006, No. 105 (LVII) – 2006. http://www.unhcr.org/refworld/docid/45339d922.html (accessed 07.10.13.).

Valtonen, K., 2001. Social work with immigrants and refugees: developing a participation-based framework for anti-oppressive practice. British Journal of Social Work 31, 955–960.

Weaver, H., Burns, B., 2001. I shout with fear at night. Understanding the traumatic experiences of refugees and asylum seekers. Journal of Social Work 1 (2), 147–164.

Social Work Education

Malcolm Payne, Manchester Metropolitan University, Sutton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work education prepares people to practice as social workers, and also a social institution incorporating university and social agency interests. It creates practitioners' professional identities, bringing together academic learning with practice and life experience. Distinctive models of social work education exist in English-speaking, European and resource-poor countries. Attempts to create global standards for curricula and organization of social work education are contested. Practice experience in social agencies uses supervision linking administrative checking of actions, education, and personal support. Practice experience generates tensions between universities and social agencies about social work educations' aims.

Social Work Education and Social Work Identity

Social work education is the process of preparing people to practice as social workers, and developing their capacity to practice through learning and personal development. This process starts prior to professional qualification and continues during qualifying education and through lifelong professional development. Educational processes, therefore, help to create the professional and personal identities of social work practitioners.

As with all professional education, social work education also contributes to the formation of the identity of the profession, by its education curricula and the learning processes that are used. By virtue of their shared involvement, the personnel engaged in social work education in universities and in social agencies also develop a shared social identity, which constitutes an identifiable grouping within the social work profession. Thus, 'social work education,' the process, interacts with 'social work education,' the social institution, and both contribute to the identity of social work and contribute to social policy-making, higher education and social agencies.

University and Social Agency Contributions

Social work education is usually provided in universities or institutions providing education at a higher education level; 'universities' in this article includes other higher education institutions. Social work qualifications, therefore, lead to a degree-level qualification and in some countries, is also or mainly available at masters level.

Like other education for professions involved in helping human beings, social work education incorporates contributions from personnel in both educational institutions and social agencies. This is for three reasons. First, many people enter social work having been informal caregivers in domestic settings or paraprofessional or volunteer workers in social agencies, so their early experience as helpers or in agencies forms the basis of all later education. Low-skilled or volunteer jobs with little or no formal training is not regarded as social work, but is variously called paraprofessional work, social care practice, or more broadly 'care work' (Boddy et al., 2006). Paid caregivers employed by individuals as 'personal assistants' are a recent innovation because of 'cash for care' policies in many countries (Ungerson, 1999; Leece and Peace, 2010). Such policies provide grant aid for people to employ their own assistance, rather than staff employed by social agencies organizing services in people's own homes. If such staff or volunteers then decide to seek qualification as social workers, their experience and initial training contributes to the process of qualifying and postqualification social work education that they participate in, and may constrain or enhance social work education accordingly.

The second reason why social work education incorporates input from outside educational institutions is that most social work education includes work experience in social agencies. This generates continuing relationships between educational institutions and social agencies and creates a group of personnel concerned with such 'practice education' or 'practice teaching' or 'field education' or 'student supervision' – the designation of this part of social work education varies in different countries. These educational experiences, the specialized personnel involved, and the links between social agencies and educational institutions through which they are organized produce struggles for influence over curricula and learning processes and over policy-making about the balance of academic and social agencies.

The third reason for the links between educational institutions and social agencies is that, once qualified, practitioners continue in a process of personal professional development and lifelong education that sometimes involves engagement with educational institutions. Many universities and other educational institutions provide advanced postqualifying and doctoral education in social work as part of that engagement (Laot, 2001; Scourfield and Maxwell, 2008).

Models of Social Work Education

A number of models of social work education exist, and while there is increasing international mobility in the workforce, little research exists to understand differences (Spolander et al., 2011). Different forms of social work exist and while most social workers have some education in all, many countries emphasize one or permit specialization. The most widespread practice in developed countries is practical and psychological

help provided through interpersonal relationships with individuals and families, usually called casework until the 1970s; this is what most people understand by 'social work' practice. Groupwork is a development of this, bringing together individuals to meet in small groups, typically of between six and twelve. Groups may focus on changing behavior or bring together people with similar social difficulties to help and support each other, or engage in activities that have educational or personal development objectives (Papell and Rothman, 1966). Groupwork is also important in residential care, where people who are unable for various reasons to care for themselves live in communal settings. Community work involves helping people come together to identify issues of concern and take action to resolve them. This is increasingly called 'macro practice,' since it deals with larger populations than individual, family, or groupwork practice (Payne, 2014). In European countries, there is also an important distinction between social pedagogy or social education, and social work or social assistance (Boddy and Statham, 2009). Social pedagogy has a conception of social work as a positive educational process for individuals and groups with shared interests. It thus avoids the potentially stigmatizing connotations of Anglo-American social work traditions that see social work as a process for resolving problems and most clients of social workers as being socially excluded. An important source of social pedagogy is a group of German philosophers who aimed to develop those social aspects of education that particularly focus on poor people in societies (Lorenz, 1994: 91–97). Social pedagogy is used particularly with children in residential and day care and in community work in which individual and collective self-development and education interact.

Social development is the main form of social work in resource-poor countries where economic development is an important national priority and seeks to incorporate social progress alongside it. African, Asian, and Latin American social work education often contains a strong and sometimes overriding focus on social development (Midgley and Conley, 2010).

History of Social Work Education Structures

Social work education emerged from training courses for employees and volunteers provided by major charitable agencies in the US and European countries in the late-nineteenth and early-twentieth centuries. The charity organization societies operating in large cities at that time favored developing investigatory skills and appropriate moral attitudes among suitable middle-class women through supervision in the field. However, concerns about the quality of charity workers and a wish to claim expertise through establishing a scientific basis for casework methods and underlying political policies opposed to Fabian ideas of social reform led to the establishment of more academic courses (Jones, 1976). These built links with local universities and higher education institutes. The first courses were established in Amsterdam, London, New York, and Chicago in the period 1895–1910 (Kendall, 2000). During or after the first world war, as social work increasingly became a paid occupation, national associations of schools of social work, and international association of schools of social work (IASSW) emerged at the international social work fortnight in Paris in 1928 (Kendall, 1978). Such initiatives were extended to Australia, Canada, India, and a few French- and English-speaking African and Asian colonies over the next 10 years.

After the establishment of European welfare states following the Second World War, and the development of casework techniques, particularly in the US, social work education expanded both in universities and in specialized higher education institutions focusing on professional education in some countries. A network of universities and colleges in Africa, Asia, and South America began to develop courses similar to those in Europe and the US until the 1960s. This was also associated with the expansion of university social science teaching during that decade. There was continued differentiation as social work education in the developing world placed a greater emphasis on social development. These extensions of social work education were encouraged by the US Fulbright Fellowships, Canadian technical assistance, the Catholic International Union for Social Service, and United Nations' initiatives in family planning and eventually in community development (Younghusband, 1963).

Countries dominated by communist regimes de-emphasized the importance of social work in the period 1950–90, favoring instead social programs such as extensive day care and nursery provision for children and support for unemployed, sick, disabled, and retired people through workplace or trade union social support (Simpura, 1995). From the 1990s onward, former Soviet countries, Eastern and Central European countries, and China all developed extensive networks of social work education as a way of promoting social liberalization, improvements in large oppressive residential care institutions providing for people with mental illnesses and intellectual and physical disabilities and in training undereducated workforces hitherto employed in public welfare services (Payne, 2005: 227–241).

Global Standards for Curricula: Internationalist and Hegemonic Views

The interpretation of these developments is contested (Payne, 2005). The internationalist view (Kendall, 1978) proposes that there has been one stream of progress in the development of the profession and its education, based on a universal social science, leading to a coherent paradigm of social work that is relevant everywhere but adapted to fit particular national systems of welfare. This view is reinforced by the way in which former communist regimes adopted Western patterns of social work education in the 1990s. However, alternative views (Payne, 2005; Payne and Askeland, 2008) propose that the internationalist view is hegemonic, with the dominance of Western and particularly American cultural values and professional assumptions leading to the inappropriate importation of models of practice on the unproven assumption that these are universal. This has been a particular issue in relation to the transfer of American models of casework in resource-poor countries (Midgley, 1981), when social development practice was more relevant. In the hegemonic view, different national welfare regimes draw on local cultural and social philosophies, incorporating elements of models of practice in use in other

countries as these become available. There is evidence that this took place in the development of social work in the late nineteenth century, for example, in Japan, and in the differentiation of social work theory and education in the latter part of the twentieth century (Payne, 2005). The dominance of Western models often reflects colonial and postcolonial relationships between Western and resource-poor nations. The hegemonic view argues for understanding of the different cultural and service histories that inform local practice and for facilitating cultural translation between different national practices (Payne and Askeland, 2008).

A recent iteration of this debate has arisen in relation to the development of global standards of social work education, during the first decade of the twenty-first century, as international bodies in social work have sought to develop agreed standards (Sewpaul and Jones, 2005; Gray and Webb, 2008) and as communist and former Soviet countries have developed social work services.

The main area of debate is about Western cultural influence on the curriculum of social work courses, as against the assumed universality of research knowledge. Responding to concern about this issue, the global standards (International Federation of Social Workers, 2012) permit wide variation, allowing different national and cultural traditions to influence the curriculum in each country. The Standards propose that curricula should cover four main areas:

- the domain of the social work profession, for example, social structure and oppression, human behavior and development and the social environment, social work's origins and purposes, social welfare policies;
- the domain of the social work professional, for example, the development of the self-reflective practitioner, the interaction of personal values, and life experience and practice;
- methods of social work practice, for example, assessment, relationship-building and helping processes, social work values and ethics, research methods;
- the paradigm of the social work profession, for example, the dignity and worth of all human beings, capacity-building and empowerment, advocacy, building strengths, respect for diversity.

National curriculum statements cover similar ground in different ways.

The Social Work Education Process

Transformations in Identity

Professional education is a process through which an individual's personal identity is transformed so that it incorporates a professional identity as a social worker. Most elements of the global standards are less contested than the curriculum because they are focused on this process demonstrating a clear commitment to social work values, appropriate selection and learning processes, and fair assessment.

Three elements of the process move in parallel during a practitioner's career:

- practice experience,
- life experience, and
- academic learning.

Each social worker brings together these three elements to create a unique personal identity as a social work practitioner, with links to an academic and professional identity. Post-qualifying education often initiates or supports a practitioner in developing specialist emphases in their work, that further advances the formation of their personal academic and professional identities.

Each transformation in identity includes academic learning and the development of professional skills and values, but personal identity gained through life experience is crucial in social work education, prior to, during, and after any formal educational experience. This is because a social work practitioner incorporates their 'self,' for example, the experience that derives from their gender or ethnicity, as an aspect of their practice (Urdang, 2012). Practice involves modeling behavior and attitudes, using the practitioner's self in a relationship with clients with whom they work. The relationship enables them to influence people who may be excluded from or reject social relationships, and research confirms the importance of honesty and genuineness in relationships between practitioners and clients (Carkhuff and Berenson, 1977) to enable this influence to succeed. Therefore, social workers must be genuinely engaged with the interests of their clients, and their self must be developed to achieve this.

Life and Practice Experience

The importance of the preprofessional personal identity in the formation of a practitioner's professional identity means that even where it is not a formal requirement of entry, many applicants for admission to qualifying social work education find that personal life experiences or work experience motivate them to work and receive training in this field. Therefore, they bring into the education process experience of how people live and how social work is organized and practiced. Because students bring life or practice experience to, the education process requires social work education to use modes of adult learning (Knowles, 1984) that allow students to reflect on and incorporate this experience into the education process. This means that social work education is often problem based and collaborative rather than didactic, and also emphasizes equality, seeing teachers and learners and exploring knowledge and practice together. An experiment on a UK qualifying course developed 'enquiry and action learning' (Burgess, 1992), for example. It used a curriculum based on scenarios and problems encountered in practice, rather than on academic subjects, aiming to integrate academic teaching and practice learning, to build on the knowledge and skills students brought to the course, and to help them become self-directed lifelong learners, able to adapt to the changing demands that will be placed on them as professionals.

The importance of life experience in social work education also leads to a culture in which the personal suitability of applicants for admission to a course and of students during the course is an aspect of the assessment, which may lead to students being failed or excluded for personal rather than professional reasons.

Academic Learning

Academic learning in social work education draws on a wide range of research and scholarship. Pawson et al.'s (2003)

systematic review identifies five sources of knowledge, as follows:

- organizational knowledge,
- practitioner knowledge,
- policy community knowledge,
- research community knowledge, and
- user and caregiver knowledge.

This research indicates that conventional academic knowledge, based on journal and book publication, is not the only kind of knowledge that needs to be incorporated into social work education.

Moriarty's (2011) systematic literature review for the English Department of Health summarizes the available research, and cites available studies, including relevant international research.

- The social sciences, in particular, psychology, sociology, and social policy – although a background in economics and political science also offers contextual understanding for implementing social science in practice. Organizational and management studies also play a part in understanding the social context in which social workers practice.
- Law includes law that governs the responsibilities and legal powers of social workers and law in which they might have to advise clients, such as social security, housing, and other legal rights.
- Practice skills deriving from the social work literature and from micro-psychology, include the main elements of any practice, involving assessing clients, planning appropriate action, intervening in social situations to help them, and monitoring and evaluating the outcomes of the intervention. It also includes skills in interpersonal communication and influence in working with clients and colleagues, also drawing on micro-psychology.
- Cognate professional fields offer understanding that may be transferred to social work from medicine, nursing, teaching, psychology, and skills in interprofessional and partnership working.
- Ethics and values teaching draws on cognate professional literature, including medical ethics, and on philosophy.
- Research appreciation and research methods teaching appropriate to the level of the course sufficient to enable practitioners to understand and critically evaluate research relevant to their practice. This teaching also contributes to practitioner's ability to critically evaluate their practice either through personal research or by participating competently in agency and government quality assurance and evaluation research.
- Critical thinking and critical reflection skills enable practitioners to continue to be self-critical throughout their career (Fook and Gardner, 2007).

Academic learning in social work calls upon the full range of teaching methods commonplace in higher education, including lectures, seminars, and group discussions, together with individual and joint project work. At higher levels of education, this includes research projects, which may also have an integrative function. Practice skills are sometimes taught through skills laboratories, often assisted by video. These involve role-play scenarios to test out student skills and receive feedback, which aims to develop their competence. Agency staff, service users, and caregivers may also be involved in providing feedback to students on their performance in similar situations (Moriarty, 2011).

Practice and Field Education and Supervision

Moriarty (2011) documents a debate, however, about the effectiveness of laboratory teaching of interpersonal skills as compared with practice experience as a learning experience in social agency or field settings. There is a long history of research that suggests that social work students regard practice learning provided in agencies as the most important, enjoyable, or memorable element of their learning. Practice or field education enables students to observe the work of agencies and carry out research or organizational studies (Payne, 2004). It may also provide an important opportunity for international exchange (Payne and Askeland, 2008; Noble and Hendrickson, 2011).

In the Anglo-American model of social work education, a university typically (Bogo, 2010) organizes placements for its students in social agencies for substantial periods, sometimes amounting to half the course program. The most common arrangement is for an individual student to work in the agency, carrying out the same tasks as employees, and supervised by a social worker employed by the agency in doing so. A university staff member liaises with the agency supervisor. This experience allows students to apply and practice knowledge gained in the university setting, and gain experience in transferring academic knowledge to practice settings, and the supervisor assists them in doing so, eventually assessing their competence in practice. Other employees of the agency contribute to helping and assessing the student and provide experience in cooperating with colleagues. While the students undertake a full range of work, they do not carry a full workload, to facilitate the learning process.

Social work supervision, as it is provided both to practitioners and also to students, is a distinctive process containing three elements:

- Administration, checking that the social worker performs professional functions as required by the profession and the agency.
- Education, teaching the social worker knowledge, skills, practical techniques, and appropriate value orientations.
- Support, enabling the social worker to manage their personal reactions to their experiences in professional social work (Kadushin and Harkness, 2002).

This last element is necessary because experience of distressing circumstances, for example, events such as child abuse or bereavement, may be stressful and raise echoes of a practitioner's own life experience in ways that might adversely affect their capacity to function as a social worker or their judgment. Such supervision continues throughout a social worker's career, and experiencing it as part of social work education is an important socialization into the way in which practice is managed within the profession. The support and education elements are organized in an unusual way, compared with supervision in many health care professions, where they are often provided by a consultant independent of the practitioner's

employment. In social work, by contrast, it is typical for it to be provided by their manager. The student social work supervision arrangement mimics this professional convention (Carpenter et al., 2012; Hawkins and Shohet, 2012).

Practice or field education leads to an assessment of the student's competence to practice, again mimicking the social work professional convention that supervision is management as well as education and support. Professional qualification usually requires successful completion of both this element of the course and also the academic program. While there is evidence that the involvement of field or practice educators increases the rage of skills and experience available to social work courses, research also suggests that some find it difficult to fail students in this element of the course (Moriarty, 2011).

The engagement of agencies in field and practice education makes the organization of placements and supervision a site of tension over the influence of field agencies on the curriculum and assessment of social work education. Most of this disagreement reflects uncertainty about the objective of qualifying education (Moriarty, 2011). The areas of disagreement lie in the wish of agency managements for graduates of social work qualifying education to have a high degree of procedural knowledge, ability to assess risk in difficult situations, and decide on appropriate interventions, as against the view of universities that a broader curriculum with transferable skills that practitioners can then build upon in their early practice experience is a more appropriate objective of qualifying education.

As a result of this tension about aims, the position of field or practice educators as part of social education can be ambivalent. They may lean toward loyalty toward the employer conception of social work education or to their position within social work education. The employer view is often represented on behalf of employers or mediated by staff involved in the management and delivery of training, including the organization of field or practice education and student supervision. A related area of difficulty is the internationally increasing use of competence frameworks in practice assessment. These measure outcomes of professional education, in place of guidance on the curriculum to be provided in the teaching of (that is, input to) courses and offer a detailed statement of what students are able to do, tested in a work setting. This emphasis on detailed specification and on outcomes is an implementation of managerialist approaches in New Public Management (NPM), a new settlement in the 1980s of relationships between the citizen and the State and between management, political control and professional responsibilities (Clarke and Newman, 1997). It arose from the neoliberal analysis that comprehensive welfare states on the European social model were not economically sustainable. Consequently, new forms of management were required to control the tendency to extend the coverage of the state and to manage professional activities so that they were more financially efficient, in particular, using quantitative techniques, such as work study (Pollitt, 1993). NPM has a view of management as a generic expertise in which management has its own knowledge base and skills. The debate about competence-based education focuses on how social work education may be developed to be strongly applicable in the context of agency definitions of the aims and content of social work. There has been a greater trend to managerial control by the center, rather than discretion to a local professional service, greater use of legal and procedural processes rather than professional discretion, and greater use of managerialist techniques, such as control by specification of standards, training competences, and performance indicators, rather than permitting the flexibility of discretion. Professional knowledge and skill is subordinated to management and financial control. Work activity becomes deskilled as professional tasks are fragmented. Initially, this is done by codifying and dividing up tasks into smaller components, which practitioners should be able to carry out, called competences.

Competences required as the outcome of education are identified through the functional assessment of what social workers do. These are formalized in occupational standards, which workers should be able to meet, and competences are derived from these which specify workers' performance. Critics of this approach (for example, Dominelli, 1996) claim that organizing social work education through competence testing makes for a mechanistic approach to learning and assessment, which is inadequate for measuring professional knowledge and skills and professional judgment. Also, competence approaches minimize the importance of intellectual understanding and coherent analysis of the field of study; also that they reduce the importance of values within social work. However, competences can be written, and assessments carried out, to include underpinning knowledge and values. Another point is that focusing on knowledge, understanding, and values as a coherent whole makes it difficult for students to apply understanding in practical situations that requires selecting from and using knowledge in less comprehensive packages: competence approaches emphasize education based on practical ways in which the knowledge may be used. Other advantages are that competence frameworks define precisely what a student should be able to do, rather than relying on overall judgments of students' attainment in practice by academic or practice teachers. It also relies on statements about what social workers actually do, rather than theoretical accounts of practice, and is therefore likely to connect rather better with employers' requirements of their workers. Assessment of detailed competences as part of social work education permits a more explicit and defensible statement of practice assessment, but makes it more difficult to see students' attainments as part of their holistic development as practitioners (O'Hagan, 2007).

Lifelong Learning, Critical Reflection and the Learning Organization

Most professions have a process for qualified practitioners to engage in continuing professional development (CPD), a lifelong process in which professionals take responsibility for continuing to develop their skills and knowledge. They do this through reflecting on their work and developing and fulfilling learning objectives that will improve their practice (Megginson and Whitaker, 2007). A basic assumption of CPD is that individual professionals are responsible for and create their own program. Where they are employed professionals, however, the organization in which they work needs to become

a 'learning organization' that facilitates and supports its employees in achieving this (Gould and Baldwin, 2004).

Integral to CPD is that practitioners engage in critical reflection, which puts together three elements of social theory – reflection, reflexivity, and criticality – to describe a way of developing practice skills. Reflection helps professional practitioners plan and do their work with greater understanding of the situations, they are facing and to develop practice theories for professionals. They use the ideas that come out of reflecting on particular incidents to adapt existing theories or create new ones. Reflection is a circular process: you have an experience, reflect upon it and consequently change the way you act in the future. Reflection is crucial to learning from experience or from research, because it enables you to reconstruct or reorganize the experience so that you can use it (Redmond, 2006). Reflexivity derives from qualitative research methods, particularly as applied in feminist research (Finlay and Gough, 2003). It means a cyclical process, in which we study how what we observe affects our thinking and how that then affects what we do. The implication of this is that you incorporate your own and other people's reactions to an event, so that you do not take for granted your own perceptions. Reflection becomes critical when practitioners use critical social theory, which, in addition to reflexivity, does not take existing patterns of thinking and social organization for granted. This helps because it offers practitioners alternative ideas when shifting their own perceptions and those of clients and colleagues. Critical reflection has recently developed strongly in health and social care (Fook and Gardner, 2007) as both a training and a personal development device.

Through CPD, social work education continues throughout practitioners' careers, and includes programs of reading, organized personal experience, such as job swaps, shadowing a colleague with different responsibilities, short training courses, and advanced professional and academic qualifications. Advanced qualifications include advanced masters degrees and professional doctorates, which include specialized professional teaching alongside a research thesis. Social workers seeking academic roles, or wishing to undertake a substantial research project, will often complete a PhD instead of a professional doctorate.

Conclusion

Social work education is a process through which practitioners transform their personal identity by incorporating a professional identity as a social worker within it and continue to develop their identity and capacity within the profession during the whole of their career. It may be considered in three phases: prequalification training and experience, including life experience; qualifying education; and career-long CPD.

The institutional and organizational aspects of social work education brings together academic elements delivered within universities and practice elements delivered mainly in social agencies. In this way, social work education incorporates tensions between agency and university interests in controlling the curriculum and outcomes of qualifying education and CPD. The crucial issue in these tensions is the need to incorporate broader social science traditions of learning within a form of education that is also designed to train employees in the procedural and practical requirements of the role of social workers in agencies.

See also: Critical Social Work Practice; Research and Theory Building in Social Work; Social Work Theory.

Bibliography

Boddy, J., Statham, J., 2009. European Perspectives on Social Work: Models of education and professional roles. Thomas Coram Resarch Unit, London.

Boddy, J., Cameron, C., Moss, P. (Eds.), 2006. Care Work: Present and Future. Routledge, London.

Bogo, M., 2010. Achieving Competence in Social Work through Field Education. University of Toronto Press, Toronto.

Burgess, H., 1992. Problem-led Learning for Social Work: The Enquiry and Action Approach. Whiting & Birch, London.

Carkhuff, R.R., Berenson, B.C., 1977. Beyond Counseling and Therapy, second ed. Holt, Rinehart & Winston, New York.

Carpenter, J., Webb, C., Bostock, L., Coomber, C., 2012. Effective Supervision in Social Work and Social Care. Research Briefing 43. Social Care Institute for Excellence, London.

Clarke, J., Newman, J., 1997. The Managerial State: Power, Politics and Ideology in the Remaking of Social Welfare. Sage, London.

Dewey, J., 1910. How We Think. Heath, New York.

Dominelli, L., 1996. Deprofessionalizing social work: anti-oppressive practice, competencies and postmodernism. British Journal of Social Work 26, 153–175.

Finlay, L., Gough, B. (Eds.), 2003. Reflexivity: A Practical Guide for Researchers in Health and Social Sciences. Blackwell, Oxford.

Fook, J., Gardner, F., 2007. Practising Critical Reflection: A Resource Handbook. Open University Press, Maidenhead.

Gould, N., Baldwin, M. (Eds.), 2004. Social Work, Critical Reflection and the Learning Organization. Ashgate, Farnham.

Gray, M., Webb, S.A., 2008. The myth of global social work: double standards and the local-global divide. Journal of Progressive Human Services 19, 61–66.

O Hagan, K. (Ed.), 2007. Competence in Social Work Practice: A Practical Guide for Students and Professionals, second ed. Jessica Kingsley, London.

Hawkins, P., Shohet, R., 2012. Supervision in the Helping Professions, fourth ed. Open University Press, Maidenhead.

International Federation of Social Workers, 2012. Global Standards. http://ifsw.org/policies/global-standards/ (accessed 10.12.12.).

Jones, C., 1976. The Foundations of Social Work Education. University of Durham, Durham. Working Papers in Sociology 11.

Kadushin, A., Harkness, D., 2002. Supervision in Social Work, fourth ed. Columbia University Press, New York.

Kendall, K.A., 1978. The IASSW 1928–1978: a journey of remembrance. In: Kendall, K.A. (Ed.), Reflections of Social Work Education 1950–1978. International Association of Schools of Social Work, New York, pp. 170–191.

Kendall, K.A., 2000. Social Work Education: Its Origins in Europe. Council on Social Work Education, Alexandria, VA.

Knowles, M.S., 1984. Andragogy in Action: Applying Modern Principles of Adult Learning. Jossey-Bass, San Francisco, CA.

Laot, F.F., 2001. Doctoral Studies in Social Work: European Initiatives. Éditions de L École Nationale de la Santé Publique, Rennes.

Leece, J., Peace, S., 2010. Developing new understandings of independence and autonomy in the personalised relationship. British Journal of Social Work 40, 1847–1865.

Lorenz, W., 1994. Social Work in a Changing Europe. Routledge, London.

Megginson, D., Whitaker, V., 2007. Continuing Professional Development, second ed. Chartered Institute for Personnel and Development, London.

Midgley, J., 1981. Professional Imperialism: Social Work in the Third World. Heinemann, London.

Midgley, M., Conley, A. (Eds.), 2010. Social Work and Social Development: Theories and Skills for Developmental Social Work. Oxford University Press, New York.

Moriarty, J., 2011. Literature Review for the Curriculum Development Work Stream. Kings College London Social Care Workforce Research Unit, London. http://ec2-79-125-112-176.eu-wes.t-1.compute.amazonaws.com/sspp/kpi/scwru/pubs/2011/moriarty2011litreview.pdf (accessed 09.12.12.).

Noble, C., Hendrickson, M. (Eds.), 2011. Social Work Field Education and Supervision across the Asia Pacific. Sydney University Press, Sydney.

Papell, C.P., Rothman, B., 1966. Social group work models: possession and heritage. Journal of Education for Social Work 2, 66–73.

Pawson, R., Boaz, A., Grayson, L., Long, A., Barnes, C., 2003. Types and Quality of Knowledge in Social Care. Social Care Institute for Excellence, London.

Payne, M., 2004. Social work practice identities: an agency study of a hospice. Practice 16, 5–15.

Payne, M., 2005. The Origins of Social Work: Continuity and Change. Palgrave Macmillan, Basingstoke.

Payne, M., 2014. Modern Social Work Theory, fourth ed. Palgrave Macmillan, Basingstoke.

Payne, M., Askeland, G.A., 2008. Globalization and International Social Work: Postmodern Change and Challenge. Ashgate, Aldershot.

Pollitt, C., 1993. Managerialism and the Public Services: Cuts or Cultural Change in the 1990s? Blackwell, Oxford.

Redmond, B., 2006. Reflection in Action: Developing Reflective Practice in Health and Social Services. Ashgate, Farnham.

Scourfield, J., Maxwell, N., 2008. Social work doctoral students in the UK: a web-based survey and search of the index to theses. British Journal of Social Work 40, 548–566.

Sewpaul, V., Jones, D., 2005. Global standards for the education and training of the social work profession. International Journal of Social Welfare 14 (3), 218–230.

Simpura, J. (Ed.), 1995. Social Policy in Transition Societies: Experience from the Baltic Countries and Russia. STAKES, Helsinki.

Spolander, G., Pullen-Sansfacon, A., Brown, M., Engelbrecht, L., 2011. Social work education in Canada, England and South Africa: a critical comparison of undergraduate programmes. International Social Work 54, 816–831.

Ungerson, C., 1999. Personal assistants and disabled people: an examination of a hybrid form of work and care. Work, Employment & Society 13, 583–600.

Urdang, E., 2012. Awareness of self – a critical tool. Social Work Education 29, 523–538.

Younghusband, E., 1963. Tasks and trends in education for social work: an international appraisal. Social. Work (UK) 20, 4–11.

Social Work Ethics

Sarah Banks, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses professional ethics in social work, exploring key themes (core values, codes of ethics, ethical theories, ethical challenges, ethical decision making, and regulation of conduct) in current literature and practice. It considers whether dominant models of social work practice and ethics from the global North should be reconsidered in an international context.

Introduction

Ethics is about how human beings treat each other and the ecosystem. It covers matters of rights, responsibilities, well-being, and harm. This article explores social work ethics as a discipline-specific version of professional ethics, focusing on the conduct, character, and responsibilities of professional social workers. In the global North, 'social work ethics' is a distinct subject area in professional education and practice, marked by a growing body of specialist literature, its inclusion as a subject area in professional education, and the development of everlonger codes of ethics. While 'ethics' is less prominent in social work literature and educational curricula in the global South, interest is growing, as Anglo-American and European textbooks and models of practice are translated, adopted, and adapted worldwide.

Professional Ethics in Social Work

Professional ethics has a normative focus on how people in professional roles ought to be and behave. In early social work (casework) in late nineteenth-century Europe and North America, the focus was on the morality of service users, rather than the ethical conduct of social workers (Reamer, 2006). Concern with professional ethics grew slowly from the mid-twentieth century in the global North, as social work expanded and professionalized. The traditional model of professional ethics was premised on the idea that professionals derive considerable power from their expertise and status, and have potential to exploit, harm, or unduly influence service users. Hence, a shared set of values for social work was articulated and promoted, based on respect for the dignity of service users and their rights to make their own choices as far as possible, alongside a commitment from professionals to be trustworthy and reliable in selflessly serving fellow human beings.

The range and complexity of the subject matter of professional ethics, as reflected in professional and academic literature, has developed over time. Some themes have been remarkably consistent, although the main focus of attention may have changed. Table 1 shows key themes and a simplified interpretation of the changing content of social work ethics in the literature, which is predominantly from the global North.

Social Work Values

In discussions of social work ethics, the term 'values' is generally used to refer to ethical values – that is, particular types of beliefs about what is regarded as worthy or valuable for human welfare. For Reamer (1999: p. 3), "social work is among the most value based of all professions", grounded in "concepts such as justice and fairness" (p. 5). This relates to the purposes of social work, which in most societies include redistribution of resources and care to those in need, support for people to participate as fully as possible in society and control of people who are likely to harm themselves or others. These purposes presuppose values about human worth, people's collective responsibilities for each other, and the importance of maintaining social order. While the particular configurations of values underpinning social work and how these are expressed vary across different sources and countries, three clusters of complex values in the current social work literature can be identified (Banks, 2012a: p. 60):

1. *Respect for the dignity and worth of all human beings*: The obligation to respect each human being as an individual, treat all people as equally valuable, and respect and promote the human rights of individuals and groups to self-determination (particularly users of social worker services).
2. *Promotion of welfare or well-being*: The obligation to bring about benefits for service users and for society more generally, balancing benefits against risks of harm.
3. *Promotion of social justice*: The obligation to remove damaging inequalities between people and groups and promote the fair distribution of goods and services among people and groups.

These values may conflict, giving rise to ethical challenges for social workers. For example, respecting the rights of a 15-year-old young woman to confidentiality can conflict with the social worker's duty to promote her well-being, if her father is violent toward her. Such situations, which require sensitive decisions to be made about how to act, feature strongly in the social work ethics literature.

Codes of Ethics

Codes of ethics are key features of professions. Although what counts as a profession is contested, it is an occupation that

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28030-6

Table 1 Themes in social work ethics

Themes	*Traditional focus (1960s–80s)*	*Newer focus (1990s–2010s)*
Core values – the articulation, analysis, and interpretation of the core values of the profession	Promotion of the self-determination and welfare of service users	Promotion of empowerment of service users and social justice in society
Codes of ethics – the development, analysis, and application of codified statements of principles and standards of conduct in relation to service users, employers, and society, usually incorporating or based on the core values	Short oaths based on the commitments and character of social workers	Longer documents encompassing values, principles, and standards for social work
Ethical theories – the presentation of theories of ethics derived from the literature in moral philosophy and their use to frame, explain, examine, or critique the nature of ethics in social work	Deontology (focus on duties and the nature of right action)	Utilitarianism (focus on consequences of action), virtue ethics, care ethics
Ethical challenges – the identification and discussion of common ethical difficulties faced by social workers, often framed as dilemmas and exemplified through the use of case examples from practice	Recognizing and respecting service users' rights; social workers' paternalism vs service users' autonomy	Ensuring fair distribution of resources; employer regulation vs professional autonomy
Ethical decision making – the discussion of how to make decisions in ethically challenging situations, drawing on ethical theories and codes of ethics, including the presentation and application of decision-making models or frameworks, often applied to particular cases	Professional focused, stepwise, linear models	Dialogical, circular, iterative models
Regulation of conduct – ensuring social workers conduct themselves according to the accepted values and standards of the profession	Self-regulation according to a shared vision of good social work	Development of external standards, rules, and systems of professional licensing, registration, and misconduct hearings

makes claims for status in society. Professionalizing occupations generally develop professional associations to support, guide, and protect the interests of members of the occupation and safeguard service users. Professional associations for social workers exist in the majority of countries where social work is practised, usually membership is voluntary and most have ethical codes. These are written documents covering the core purpose of the profession, its values and ethical principles, and sometimes, rules or standards of practice. Their stated aim is to guide professionals and protect service users (by laying down expectations for ethical conduct). They also serve as 'mission statements' – summarizing the essence of good social work, contributing to defining occupational boundaries, and reinforcing professional identity and status. Codes of ethics may be used in disciplinary hearings considering complaints against social workers. This is increasingly common, as regulatory bodies are established to register and regulate social work according to professional codes of ethics or practice (Banks, 2012a: pp. 107–108).

Comparison of codes of ethics for social work from professional associations in different countries shows considerable variation in length and the extent to which they include long lists of standards of practice (Banks, 2012a: pp. 112–114). However, the statements of values and ethical principles are remarkably similar. This is partly because the statement of ethical principles of the International Federation of Social Workers (IFSW) and International Association of Schools of Social Work (IASSW) (2004) is often used to develop country-level ethical codes. Furthermore, some professional associations, when newly developing their ethical codes, have drawn on codes from other countries.

Many criticisms are leveled at codes of ethics, including that their statements are too general and open to interpretation. This applies particularly to shorter codes that remain at the level of values and general principles. One response to this criticism has been to develop more specific guidance on how to act in particular situations. This can produce the opposite critique, that longer codes comprising lists of ethical standards are too prescriptive and undermine professional judgment.

Ethical Theories

In textbooks and teaching on professional ethics, an overview of theories of ethics developed by moral philosophers is common. These theories cover matters such as what counts as the good life, right and wrong conduct, and good and bad qualities of character. Many ethical theories aim to be universal (applying across all places and times) and foundational (offering a single theory that identifies what counts as 'ethical' and prescribes ethical being/action). Ethical theorists often espouse a particular theory or theoretical approach. However, in professional ethics textbooks, a range of ethical theories may be offered to provide different ways of looking at ethical problems. These may be regarded as complementary rather than mutually exclusive. The following overview of ethical theories is adapted from Banks (2012b: pp. 7–10).

Principle-Based Ethics

Until recently, modern Western literature on professional ethics focused on identifying and describing general and universal principles to guide ethical conduct. These fall into two schools of thought: deontological (duty-based) ethics, associated with the eighteenth-century German philosopher, Kant (1785/1964); and teleological (consequentialist) ethics, associated with the nineteenth-century British utilitarians, Bentham and Mill (1863/1972). Deontological or Kantian ethics is based on the ultimate principle of respect for persons as rational and self-determining beings. Any action

that fails to accord respect to each individual person (such as lying) is wrong, regardless of whether it produces good consequences. Utilitarian or consequentialist ethics, in contrast, judges the rightness and wrongness of actions according to whether they produce a greater or lesser balance of beneficial over harmful consequences for the greatest number of people. According to utilitarianism, lying might be regarded as morally right, if it resulted in a good outcome (saving life or producing a lot of happiness).

These two schools of thought are in opposition, if aiming to develop a comprehensive ethical theory based on a key foundational ethical principle. However, in everyday life, and professional practice, principles that promote respect for individual choices and rights are as important as principles that promote good outcomes for individuals and society. Statements of ethical principles and codes of ethics for social work contain both types of principle, as can be seen in the international statement of ethical principles for social work (IFSW and IASSW, 2004). Arguably, some of the more challenging ethical dilemmas in social work lie in deciding when to compromise respect for an individual's right to freedom of choice and action to promote what is considered their greater good or the greater good of others or society in general.

According to principle-based approaches, ethical decision making is a rational process that involves applying general principles to particular cases. Decision makers should treat all similar cases in a like manner, as impartially and objectively as possible.

Character- and Relationship-Based Approaches to Ethics

There are alternative theoretical approaches to ethics that start with particular people and situations. Virtue ethics, for example, focuses on the qualities of character of the moral agent, and asks not "what should I do?" but "what kind of person should I be?" and "what would a good person do in this situation?" (Banks and Gallagher, 2009; Hursthouse, 1999; Swanton, 2003). A focus on the development of good qualities of character can be found in many ancient Eastern religious teachings, including the works of Confucius and Mencius, and Buddhist texts. According to Gyekye (2010), character also forms the basis of African ethics. In Western philosophy, virtue ethics is associated with the ancient Greek philosopher, Aristotle (350 BCE/1954), and later developments by Christian religious philosophers, Aquinas and Augustine. After declining in popularity, virtue ethics has recently undergone a revival in Western ethics, to complement or replace more abstract, principle-based approaches to ethics.

Other situated approaches to ethics include the ethics of care (Tronto, 1993; Held, 2006), which focuses on relationships between people and particular responsibilities inherent in special relationships (like mother and child); and the ethics of proximity, based on responsibilities experienced in face-to-face encounters between one person and another (Vetlesen, 1997; Levinas, 1989). The emphasis on relationships and responsibilities brings these approaches much closer to ways of thinking prevalent in the global South, where individuals are defined in relationship with others. Here, much less emphasis is placed on the individual or relationships between individuals per se, but rather the focus is on the community (communitarian ethics), seeking solidarity, harmony, and common good.

Narrative and Case-Based Ethics

Approaches to ethics that give primacy to character, relationships, and communities often use stories as a methodology. According to Nelson (1997), these stories: heighten moral perception and sensitivity; promote moral education; provide ethical justification; define one's moral identity; and make ethical evaluations through comparing stories. The term 'narrative ethics' refers to a cluster of methodologies that use stories, rather than a theoretical approach to ethics.

'Casuistry' or case-based ethical reasoning (Jonsen and Toulmin, 1988), sometimes grouped under the heading of 'narrative ethics,' is also a distinctive approach in its own right (a revival of a medieval Christian practice of providing moral guidance in particular situations). Rather than starting with ethical theory, casuistry begins with particular cases, taking into account the specific circumstances of each case in deciding what an ethically correct response might be. It works by taking a paradigm case, which is relatively straightforward and about which most people would agree in their ethical evaluations, and then compares the case at hand with the paradigm case to determine differences and similarities. This is analogous to the approach taken in legal reasoning, and requires skills in determining the morally relevant features of cases and creating taxonomies of types of cases and issues. Casuistry is not a normative theory (prescribing what is good or bad), but more like a method for making ethical assessments and decisions. In case-based ethics, 'moral reasoning' plays a crucial role. 'Reasoning' in this sense includes the use of moral intuition and practical wisdom and is not the same as rationality based on abstract principles.

If seeking an internationally relevant approach to social work ethics, this approach to ethical evaluation is helpful in that it starts with the case and advocates pursuing a detailed and careful analysis (Banks, 2012b). Sometimes people who espouse very different ethical and religious values may agree about what should be done in a particular case, by focusing on its details. Their differences emerge when they justify their ethical evaluations with reference to different values or theories.

Ethical Challenges

Social workers often play a role within welfare systems designed to care for, control, educate, or empower people. This inevitably leads to contradictory imperatives and ethical challenges (Dominelli, 2004, 2010). The key values and principles identified earlier may conflict and social workers have to make difficult choices regarding how to act and what roles to play. The literature frequently refers to ethical dilemmas and problems – difficult choices with no easy solution, where the rights, needs, and interests of different parties conflict. Yet it is also important to see social work as an occupation that has ethical issues embedded within everyday practice (not just when difficult choices have to be made). In considering the form of ethical challenges, a distinction can

be made among ethical issues, problems, and dilemmas (adapted from Banks, 2012a: p. 20):

- *Ethical issues* – pervade social work (including what appear to be 'legal' or 'technical' matters) in that practice frequently takes place in the context of state systems of welfare premised on principles of social justice and public good and social workers have professional power in the relationship with service users. For example, deciding whether to recommend home care services for a person with a disability in a straightforward case may not involve a social worker in agonizing over an ethical dilemma, but it is not devoid of ethical content (the service is premised on principles of care and distributive justice).
- *Ethical problems* – arise when a social worker sees a situation as involving a difficult ethical decision, but is clear about the right course of action. For example, the worker may decide to reject the application of a very needy person for home care services because this person does not fit the criteria.
- *Ethical dilemmas* – occur when a social worker sees himself or herself as facing a choice between two equally unwelcome alternatives, which may involve a conflict of ethical values, and it is not clear which choice will be the right one. For example, should the practitioner bend the rules for allocating home care services to help a very needy person, or follow the criteria and refuse the application? The worker is faced with a conflict between the interests of this individual and the public interest in having rules and criteria that apply to everyone.

Ethical Decision Making

Textbooks on social work ethics often include significant sections on deciding how to act in difficult situations (e.g., Bowles et al., 2006; Dolgoff et el., 2012; Reamer, 2006). The social worker is presented as an active moral agent, guided by ethical principles and making a critical interpretation of relevant ethical codes. The focus is on professionals and their internal reasoning processes, or on the reasoning processes of members of a professional team. Ethical difficulties are frequently framed as 'ethical dilemmas' and presented as cases, with readers invited to consider what decision they would make in such a situation, as the following short invented case illustrates:

Case 1: Deciding whether to accept a donation

The staff team running a domestic violence shelter has to decide whether to accept or refuse urgently needed funding from a private company. Accepting the money would contribute immediately to promoting and supporting the shelter's work on women's safety and welfare in a very violent neighborhood. Arguments for refusal include the fact that the company runs lap-dancing clubs (erotic dancing) and may gain credibility from being associated with the shelter. These links may undermine the shelter's values and integrity in the longer term.

To help students analyze cases and develop skills in ethical decision making, it is common to offer models, often leading the reader through a series of questions. There are numerous examples, ranging from those comprising a linear, stepwise process to more holistic, circular models (for a useful review, see McAuliffe and Chenoweth, 2007).

One commonly quoted linear approach is the five-step ETHIC model proposed by Congress (1999):

1. Examine personal, cultural, societal, client, agency, and professional values;
2. Think about the various UN declarations on rights and related covenants, codes of ethics, laws and agency regulations;
3. Hypothesize different courses of action based on varied decisions;
4. Identify who is most vulnerable, who will be harmed or helped; and
5. Consult with supervisors and colleagues.

These steps can encourage analysis of the issues at stake, although their simple structure masks the complexity of the range of different perspectives and factors, including emotional dimensions, and how to negotiate between them. It also does not take account of the importance of an iterative cycle of reflection and learning.

To counteract such limitations, McAuliffe and Chenoweth (2007) propose an 'inclusive' circular model, taking account of core social work values, based on four interlinked 'platforms' for good decision making (accountability, consultation, cultural sensitivity, and critical reflection), which includes the following steps:

- defining the ethical dilemma;
- mapping legitimacy (who has a legitimate stake in the decision-making process);
- gathering information;
- assessing alternative approaches and actions; and
- critical analysis and evaluation.

Decision-making models encourage systematic, comprehensive, and analytical thinking about the range of factors to be considered in making difficult ethical decisions and can be helpful in teaching and learning about ethical reasoning. However, in everyday practice, the many moment-to-moment small 'decisions' that are made about how to proceed may be quite intuitive and hardly recognized as decisions unless social workers are asked to justify or explain afterwards what they did and why. Hence, the focus of textbooks and teaching in professional ethics needs to extend beyond dilemmas and decision making.

Regulation of Conduct

The traditional emphasis in professional ethics on ethical decision making assumes that social workers are active moral agents engaged in making considered decisions based not only on regulations, laws, and codes, but also on their appraisal of the relevant professional values at stake and evaluations of people's rights, needs, and interests. In a climate of new managerialism and economic austerity, social workers are required to be more accountable for their actions and demonstrate effectiveness and

value for money, as well as ethical conduct. In many countries, especially in the global North, tighter regulations have been introduced to ensure efficiency and ethical conduct (Banks, 2011; Harris, 2003). In several countries, statutory regulatory bodies have been established for social work, the functions of which include maintaining a register of qualified professionals – with the power to remove people from the register if incompetence or unethical behavior is proven (Orme and Rennie, 2006). Regulatory bodies may protect the title, 'social worker' (only those registered can practise as social workers) and play a major role in licensing social workers and disciplining them on grounds of professional misconduct.

There is a slow trend toward statutory regulation of social work as an occupation with a protected title. Many in the profession welcome regulation as a contribution toward gaining occupational recognition and status. Others are skeptical of the benefits, recognizing that these carry the price of increasing state control and prescribing the conduct of individual social workers. The autonomy of the profession, and of professional associations, is reduced in the interests of maintaining and developing the credibility of the profession and standardizing practice. 'Ethics' comes to be associated with following predetermined rules or standards, rather than considered decision making and reflection (Banks, 2004, 2013a).

From Professional Ethics to Ethics in Professional Life

This article has focused largely on normative professional ethics – how social workers should be and act in work contexts. It is impossible to separate theoretical and practical prescriptions of what practitioners ought to do from accounts of what social workers actually do and say. While textbooks focus on normative professional ethics, there is an increasing research interest in 'empirical ethics' – studies of what people do and say in practice in relation to matters of rights, responsibilities, harms, and benefits. This may take the form of surveys to find out what people think about certain types of behavior (e.g., whether it is right to lend money to service users) and reports of actions they have taken (e.g., whether they have ever lent money to service users) or qualitative research studies involving interviews and/or observations relating to everyday practice. In ethnographic research, the ethical dimensions of social work are embedded in everyday practice and have to be identified and 'extracted' by research participants and/or researchers. This can lead to a different conception of ethics than found in normative textbook ethics, encouraging a view of ethics as involving emotion *and* rationality (e.g., empathy, fear, anger). It also encourages a view of ethics as embodied (a physical side to ethical being and acting); embedded (ethical norms are deeply contextualized in communities, organizations, and wider cultures and responsibilities relating to particular relationships); and evolving (created through dialog and relationships with others).

Approaching ethics as a situated practice resonates with theoretical approaches that focus on character, care, communities, and narratives, and is encapsulated in Margaret Urban Walker's (2007) 'expressive-collaborative' model of ethics. On this view, ethics is "a socially embodied medium of understanding and adjustment in which people account to each other for the identities, relationships, and values that define their responsibilities" (Walker, 2007: pp. 67–68). According to Walker, ethical life is a continuing negotiation among people. This contrasts with the dominant type of ethical theory, which presents a 'theoretical-juridical' model of ethics as a codifiable set of moral formulae to be applied by any agent in making a specific, justified action-guiding judgment (Walker, 2007: pp. 58–59). In the context of social work ethics, this implies a broadening of traditional normative professional ethics (focusing on abstract ethical principles, ethical dilemmas, ethical decision making, and, increasingly, ethical regulation) to what I have called elsewhere 'ethics in professional life' – seeing professional life as an ethical practice and ethical issues as embedded and embodied in everyday being and doing (Banks, 2010).

The Dangers of Imperialism in Ethics

In many countries in the Southern world, the 'expressive-collaborative' approach to ethics makes more sense than the 'theoretical-juridical' approach. The identification and abstraction of a dimension of practice that can be labeled 'ethical' may be quite challenging in some cultural contexts, especially if there is no specific word easily translatable as 'ethics' (as seems the case in many African languages). The growth of the 'ethics industry' in the North, including the production of textbooks, international codes of ethics, and specialized teaching modules, may entail exporting decontextualized models of ethics and sets of principles and rules that do not necessarily fit comfortably everywhere (Hugman, 2008; Ife, 2008). While the formulation of international principles that promotes respect for human dignity, human rights, and social justice is important and welcomed by social work practitioners in many countries, the theoretical-juridical model is just one approach to ethics. This presents a challenge, as it requires academics and practitioners to be alert to imperialism in ethics (transferring Northern values and approaches to the Southern world), while avoiding ethical and cultural relativism (uncritically accepting practices based on local norms). This challenge is exemplified in case 2 (summarized from Banks and Nøhr, 2012: pp. 91–94).

Case 2: A Dutch student in Vietnam

A Dutch social work student was undertaking fieldwork practice in a rehabilitation hospital in Vietnam. She accompanied two physiotherapists on a home visit to an 8-year-old boy, Trung, who had Duchenne Muscular Dystrophy (DMD, a severe form of muscular dystrophy characterised by rapid muscle degeneration). Trung was in a wheelchair, writing on a small table attached to the chair. His mother proudly talked about the school prize he had won for his handwriting and Trung talked about his favorite activities. His mother asked when there would be an improvement in Trung's physical condition. One of the physiotherapists said he did not know, and that Trung should keep doing his exercises.

On returning to the hospital the student asked the staff if any improvement was possible with DMD. They replied 'no'. There was a possibility that Trung did not have a long life ahead. However, it

was better that Trung and his parents did not know this, otherwise they would feel sad.

The student entered into a discussion, explaining that this was contrary to her social work ethics and that in the Netherlands professionals try to be open and honest about a patient's medical situation. By not telling the family the truth, the family did not have the chance to overcome the pain that comes with the prognosis. She recognised that this approach was culturally linked, but wondered if it was based on lack of knowledge, including not knowing how to deliver bad news. The physiotherapists were interested in the Dutch way of handling such situations, just as the student was interested in the physiotherapists' explanation that in Vietnam discussing a terminal illness with patients and families causes unnecessary misery.

For the Dutch student and Vietnamese physiotherapists (and all health and social care professionals) 'ethics' is an ongoing project. Referring to European ethical codes and textbooks may offer conceptual frameworks and point to issues to consider. However, these frameworks must be critically interpreted and located in the political and cultural context of specific countries and in relation to the aims, purposes, and ways of working of the relevant agencies. Being ethical in this case is not just about making decisions according to principles or following rules, but requires practitioners to engage in ongoing dialogs with themselves, their colleagues, and broader society – an 'expressive-collaborative' practice.

Concluding Comments

Social work ethics is about challenges and contradictions in everyday practice. There are no easy formulae for tackling challenges or resolving contradictions, although ethical codes and decision-making models may provide some guidance. Acting ethically and being ethical is an ongoing project that entails a process of critical reflexivity (being aware of the social and political contexts in which one is working and recognizing one's own position, role, and power). This can be described as 'ethics work' (Banks, 2013a). It involves emotions as well as reason; the development of qualities of character as well as decision-making skills; and an ability to live with and negotiate the inevitable ethical tensions in the work. Given social workers' roles at the interface of state welfare and social control, it requires moral courage for social workers to speak out against injustices and to resist becoming highly regulated administrators of evertighter welfare budgets. Social work is also inherently political (Banks, 2013b; Dominelli, 2010; Ferguson, 2008). The political dimensions of social work are inseparable from the ethical. This entails a collective commitment to working for human rights and social justice, and a conceptualization of social work as not only a socially useful and value-based profession, but also a worldwide social movement concerned to advocate alongside and for service users, and those who need services but are excluded.

Acknowledgments

I am grateful to: Palgrave Macmillan for permission to use extracts from Banks, S. (2012a) *Ethics and values in social work*; Routledge for materials from Banks, S. (2012b) Global ethics for social work? A case-based approach, in Banks, S. and Nøhr, K. (eds.), *Practising social work ethics around the world; and* the student who provided material for case 2.

See also: Social Work Education; Social Work Theory.

Bibliography

Aristotle, 350 BCE/1954. The Nichomachean Ethics of Aristotle (D. Ross, Trans.). Oxford University Press, London.

Banks, S., 2010. From professional ethics to ethics in professional life: reflections on learning and teaching in social work. In: Zaviršek, D., Rommelspacher, B., Staub-Bernasconi, S. (Eds.), Ethical Dilemmas in Social Work: International Perspective. Faculty of Social Work, University of Ljubljana, Ljubljana, pp. 119–132.

Banks, S., 2011. Ethics in an age of austerity: social work and the evolving new public management. Journal of Social Intervention: Theory and Practice 20 (2), 5–23.

Banks, S., 2012a. Ethics and Values in Social Work, fourth ed. Palgrave Macmillan, Basingstoke.

Banks, S., 2012b. Global ethics for social work? A case-based approach. In: Banks, S., Nøhr, K. (Eds.), Practising Social Work Ethics around the World: Cases and Commentaries. Routledge, Abingdon, pp. 1–31.

Banks, S., 2013a. Negotiating personal engagement and professional accountability: Professional wisdom and ethics work. European Journal of Social Work 16 (5), 587–604.

Banks, S., 2013b. Ethics. The Policy Press, Bristol.

Banks, S., Gallagher, A., 2009. Ethics in Professional Life: Virtues for Health and Social Care. Palgrave Macmillan, Basingstoke.

Banks, S., Nøhr, K. (Eds.), 2012. Practising Social Work Ethics Around the World: Cases and Commentaries. Abingdon, Oxon, Routledge, New York.

Bowles, W., Collingridge, M., Curry, S., Valentine, B., 2006. Ethical Practice in Social Work: An Applied Approach. Crow's Nest. Allen and Unwin, New South Wales.

Congress, E., 1999. Social Work Values and Ethics: Identifying and Resolving Professional Dilemmas. Nelson-Hall, Chicago, IL.

Dolgoff, R., Harrington, D., Loewenberg, F., 2012. Ethical Decisions for Social Work Practice, ninth ed. Brooks Cole, Belmont, CA.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Oxford.

Dominelli, L., 2010. Social Work in a Globalising World. Polity Press, Cambridge, MA.

Ferguson, I., 2008. Reclaiming Social Work: Challenging Neo-liberalism and Promoting Social Justice. Sage, London.

Gyekye, K., 2010. African ethics. In: Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/african-ethics/ (accessed December 2010).

Harris, J., 2003. The Social Work Business. Routledge, London.

Held, V., 2006. The Ethics of Care: Personal, Political, and Global. Oxford University Press, Oxford.

Hugman, R., 2008. Ethics in a world of difference. Ethics and Social Welfare 2 (2), 118–132.

Hursthouse, R., 1999. On Virtue Ethics. Oxford University Press, Oxford.

Ife, J., 2008. Human Rights and Social Work: Towards Rights-Based Practice, revised ed. Cambridge University Press, Cambridge.

International Federation of Social Workers & International Association of Schools of Social Work, 2004. Ethics in Social Work, Statement of Principles. IFSW and IASSW, Berne.

Jonsen, A., Toulmin, S., 1988. The Abuse of Casuistry: A History of Moral Reasoning. University of California Press, Berkeley, CA.

Kant, I., 1785/1964. Groundwork of the Metaphysics of Morals (H. Paton, Trans.). Harper & Row, New York.

Levinas, E., 1989. Ethics as first philosophy (S. Hand, Trans.). In: Hand, S. (Ed.), The Levinas Reader. Blackwell, Oxford, pp. 75–87.

McAuliffe, D., Chenoweth, L., 2007. Leave no stone unturned: the inclusive model of ethical decision-making. Ethics and Social Welfare 2 (1), 38–49.

Mill, J.S., 1863/1972. Utilitarianism. In: Utilitarianism, on Liberty, and Considerations on Representative Government. Dent, London.

Nelson, H.L. (Ed.), 1997. Introduction: How to Do Things with Stories. Stories and their limits: Narrative approaches to the bioethics. Routledge, New York and London, vii–xx.

Orme, J., Rennie, G., 2006. The role of registration in ensuring ethical practice. International Social Work 49 (3), 333–344.
Reamer, F., 1999. Social Work Values and Ethics, second ed. Columbia University Press, New York.
Reamer, F., 2006. Social Work Values and Ethics, third ed. Columbia University Press, New York.
Rossiter, A., 2011. Unsettled social work: the challenge of Levinas's ethics. British Journal of Social Work 41, 980–995.
Swanton, C., 2003. Virtue Ethics: A Pluralistic View. Oxford University Press, Oxford.
Tronto, J., 1993. Moral Boundaries: A Political Argument for an Ethic of Care. Routledge, London.
Vetlesen, A., 1997. Introducing an ethics of proximity. In: Jodalen, H., Vetleson, A. (Eds.), Closeness: An Ethics. Scandinavian University Press, Oslo, pp. 1–19.
Walker, M., 2007. Moral Understandings: A Feminist Study in Ethics, second ed. Oxford University Press, Oxford.

Social Work in China

Angelina WK Yuen-Tsang, Hong Kong Polytechnic University, Hunghom, Hong Kong
Wang Si-bin, Peking University, Beijing, People's Republic of China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work, first introduced to China in the 1920s, was eliminated as an academic discipline and profession in the early 1950s, and restored in the late 1980s to respond to a growing demand for social workers to resolve escalating social problems triggered by the open-door economic policy. In 2006, the Chinese government announced the formation of strong social work teams to develop a harmonious society, an initiative reinforced by national policies and plans supporting the profession's development countrywide. The authors highlight the evolution of social work post-1920: the background and events surrounding its reemergence; challenges; opportunities; and prospects for future development.

Introduction

Social work was first introduced in China in the 1920s, but was eliminated both as a university discipline and as a profession in the early 1950s. After a lapse of over 30 years, social work was restored to the university curriculum in 1987 because of the growing need for social work professionals to help in resolving escalating social problems triggered by the introduction of the open-door economic policy in 1979. At the same time, social work services were gradually reintroduced by the Ministry of Civil Affairs and other related government departments and NGOs. In 2006, it was announced during the Sixteenth Congress of the Chinese Communist Party that China needed to develop a strong team of social workers to serve the escalating social needs of the country. Subsequent to the announcement, social work as a profession has undergone phenomenal growth in China. The strategic plan jointly issued by 19 National Ministries and Committees of China, 'The Medium and Long Term Social Work Professional Talent Development Plan (2011–20),' has further reinforced this trend and consolidated the development of social work as a profession in China.

Stages of Development of Social Work

The Emergence of Social Work in the 1920s

In China, social work as philanthropy dated back to very early times. It included all kinds of social relief and social services provided by governments as well as private and religious organizations and had a strong flavor of paternalism. The beginning of professional social work could be linked to the introduction of the Young Men's Christian Association (YMCA) in the 1880s, when the first YMCA officers organized comprehensive religious activities and social services for young people in the cities of Tientsin and Beijing with the aim to alleviate social problems and promote social reform (Jun, 1996; Pang, 2010). In 1922, Yenching University (Peking University) established its Department of Sociology, and in 1925 it was renamed the Department of Sociology and Social Service, thus marking the official introduction of social work as a discipline to the university curriculum in China. Social work–related subjects such as casework, group work, social administration, social welfare, and mental health social work were taught in the university curriculum (Pang, 2010). Besides teaching social work subjects in the classroom, Yenching University established the Ch'ing Ho Fieldwork Base and provided a valuable arena for teachers and students to apply their knowledge and skills in social work practice (Xia and Guo, 2002). Many other social work education programs and social work subjects were subsequently introduced to other leading universities, including Fudan University, Furen University, Jinling University, Lingnan University, and Hu Jiang University. Graduates of these programs served in social services organizations such as the YMCA, YWCA, hospitals, schools for blind people and deaf people, mental health institutions, prisons, orphanages, relief organizations, and rural community service agencies (Pang, 2010).

Abolition of Social Work in the 1950s

Shortly after the establishment of the People's Republic of China (PRC) in 1949, China adopted the Soviet model of socialism and created a centralized planned economy with a view to eliminate class oppression, inequality, and social injustices in society. During the period of planned economy, social work as a profession was totally eliminated from the government establishment and the practical field of social welfare services. All social work programs were removed from universities during the 'restructuring of institutions of higher learning' in 1952 because it was claimed that social problems would no longer exist in a Socialist society and therefore there was no need for the continuation of social work education in the country (Lei and Shui, 1991).

In this stage of development, the government was the sole provider of social welfare services and social service functions were redistributed among the Ministries of Civil Affairs, Labor, and Health, as well as among state-sponsored national organizations such as the National Women's League, the National Labor Union, and the Communist Youth Corp (Wang, 1995). The Socialist 'from cradle to grave' welfare package provided the employees of state-owned enterprises with a wide range of social protection, including pensions, health care, housing, education, and other essential social services. Residual services

for vulnerable groups including elders, orphans, the handicapped, and mentally ill persons who did not have families or work units, were taken care of by the Ministry of Civil Affairs. During this period, the previously vibrant social work community virtually vanished from the scene and social workers trained in the earlier stage were dispersed to numerous work units, some related to social services and some with no direct relevance to social work (Yuen-Tsang and Wang, 2008).

Reemergence of Social Work in the Late 1980s

The introduction of China's 'open door' economic reform in 1978 by reformists within the Communist Party of China led by Deng Xiaoping steered China away from a welfare state model toward the neo-liberal welfare model, based mainly on economic efficiency and aimed at providing only basic welfare benefits to maintain social stability. It was anticipated that there would be a significant reemergence of social inequality and class divisions (Leung and Nann, 1995; Preston and Haacke, 2003). In view of the escalation of social problems, some leading academic and political leaders who were previously trained in social work and applied sociology such as Lei Jieqiong and Fei Xiaotong (Lei and Shui, 1991) fervently advocated the reintroduction of social work as a discipline into the university curriculum. In 1987, the Ministry of Civil Affairs, together with the State Education Committee discussed the restoration of social work education in China and eventually reached a joint decision that the Ministry of Civil Affairs was to invest 1 million RMB (renminbi) in the Department of Sociology of Peking University, so that Peking University could take the lead in reinstating social work education in China.

In the summer of 1988, the Asia Pacific Association of Social Work Education (APASWE) and Peking University co-organized a Seminar of the Asian-Pacific Region Social Work Education in Beijing which was attended by over 160 delegates from the region of whom over 100 were Chinese educators and officials interested in reintroducing social work education in China (Xia and Guo, 2002; Xiong and Wang, 2007). Following the admission of the first batch of social work degree and MPhil students to Peking University in 1989, many other universities and colleges such as Jilin University, Xiamen University, the China Civil Affairs Cadre College, the China College of Youth Politics, China Women's College, Yunnan University, Nankai University, and East China University of Science and Technology also started their social work programs in the early 1990s. The China Association of Social Work Education (CASWE) was formed in 1994 to strengthen the network among schools of social work in China and foster the advancement of social work education in China. During this period, social work schools in Hong Kong and Taiwan played a pivotal role in supporting the development of social work education in China through the provision of books and teaching materials, short-term 'training of trainers,' programmes, attachment opportunities, MPhil, and PhD scholarships, and the organization of regular seminars and conferences for exchange purposes (Yuen-Tsang, 1996).

The period also witnessed the reemergence of social work services in China. The gradual demise of the work units (*danweis*) together with the comprehensive welfare system during the reform era had called for the development of community support services for those without welfare coverage. The Ministry of Civil Affairs pioneered many community-based social services, such as meals-on-wheels, home-help services, daycare centers for elderly people; and developed long-term collaborative relationships with social service agencies in Hong Kong to enhance the quality of its services. In 1991, the Ministry of Civil Affairs established the China Association of Social Workers (CASW) with the purpose of fostering the professional development of social work in China, and in 1992, the CASW became an institutional member of the International Federation of Social Workers (IFSW) with the purpose of promoting connections with the international social work community (Wang, 2012).

Rapid Expansion and Professionalization of Social Work Since 1999

The development of social work since 1999 was phenomenal and was commonly regarded by social workers as the 'Spring of Social Work' in China. In 2003, Hu Jintao and Wen Jiabao took over as the fourth generation leadership of China as President and Premier, respectively, and shifted China's ruling ideology from one that only focused on economic growth to one that emphasized overall societal development. When they took office, they inherited a China wrought with internal social, political, and environmental challenges. One of the biggest problems was the increasing wealth disparity between poor and rich people in China, for which discontent and anger mounted to a degree that seriously challenged the leadership of the Chinese Communist Party (CCP). In response, the CCP proposed the 'Scientific Development Perspective' as the fundamental guiding principle in economic and social development. This emphasized a synergetic relationship between economic growth and social development; between urban and rural development; and between economic development and environmental protection (Wang, 2012; Yuen-Tsang and Wang, 2008). Consequently, the government developed a series of social policies to alleviate urban and rural poverty and provide basic social, economic, and medical protection for the masses. The abundance of social policy initiatives introduced was unprecedented and this period was regarded by some scholars as 'The Age of Social Policy' (Wang, 2004). In 2006, Hu released the document on 'The Decision Concerning Several Major Problems in Building a Socialist Harmonious Society' during the Sixteenth Central Committee of the CCP, spelling out the need for establishing 'a grand team of professional social workers to help in building a harmonious society,' and proposed the fundamental regulations of training, evaluating, recruiting, and stimulating professional social workers. This policy marked a new chapter in the development of social work in China and spearheaded the professionalization process for social work in China.

The higher education reform in 1999 accelerated the drastic expansion of social work education programs. While there were only four universities offering social work programs in 1988 and 24 in 1999, the number rose to 36 in 2001 and soared to 186 in 2006 and 262 in 2012, graduating over 10 000 students per year (China Association of Social Work Education, 2012). But the shortage of social work educators who had undergone

professional social work education was an issue of grave concern. In 2000, in response to the urgent need to develop a critical mass of social work educators who could effectively assume leadership roles in spearheading the future development of social work education in China, the Hong Kong Polytechnic University and Peking University developed the first Master of Social Work (MSW) Program in 2000 (Xiong and Wang, 2007). Between 2000 and 2012, the Programme nurtured over 200 graduates who undertook key leadership positions in higher education and social welfare organizations, and played influential roles in advancing social work education as well as social work development in China. In 2009, the Ministry of Education announced the launch of China's own MSW Programs and 33 renowned universities from different parts of China were granted permission to offer Masters programs. By 2012, a total of 61 universities were given approval by the Ministry of Education of China to offer MSW programs (Liu, 2012).

The development of professional social services during the period was also extremely rapid. In 1997, the Pudong district of Shanghai had taken the lead to launch pilot social work practice sites and established NGOs to provide professional social work services including school social work, youth outreach service, mental health service, and services for street children (Wang, 2012). Subsequently, many major cities such as Shanghai, Beijing, Tianjin, Wuhan, Changsha, and Guangzhou piloted various social service models and developed demonstration projects. The experience of these pilot projects were shared among members of the social work community through national seminars and publications, and had significant impact on the professionalization of social work in China (Wang, 2012).

In terms of policy support, the Chinese government was extremely progressive. In 2006, the Ministry of Civil Affairs and the Ministry of Personnel of China and other ministries jointly issued the 'Provisional Regulations on the Assessment of the Vocational Standards of Social Workers' and 'Measures for the Implementation of the Examination Regarding Professional Qualifications of Social Workers' with a view to establishing at the national level, a system for the assessment of the professional standards of social workers. These two documents signified that social work would be recognized as a profession in the Chinese mainland and that measures would be taken to develop a nationally recognized career ladder for social workers with clearly stipulated vocational levels and expectations (China Association of Social Work, 2009). In 2009, the Ministry of Civil Affairs announced the document 'Concerning Continuing Education Arrangements for Social Workers,' providing guidance and monitoring mechanisms for continuing education for social workers and stipulating that all those with Assistant Social Worker and Social Worker professional qualifications had to obtain at least 72 and 90 h (respectively) of continuing professional education during the three-year registration period. In 2009, the Ministry of Civil Affairs issued the document, 'Announcement Concerning the Fostering of the Development of Nongovernmental Social Work Organizations' which introduced measures to enhance the management, supervision, and systematic development of social NGOs in China. In 2012, the Ministry of Civil Affairs and the Ministry of Finance of China jointly issued a policy paper on the purchase of social work services, providing guidelines and measures for the use of public funding to purchase social services provided by NGOs to serve the needy and vulnerable sections of the population.

The most significant policy paper which is regarded as the watershed in the professionalization of social work in China is the '2011–20 Medium Long-term Strategic Plan on the Development of Social Work Professional Manpower Team' issued in 2012 by 19 Ministries and Committees in China. The policy paper confirmed the urgent need to develop professional labor power to support the development of a harmonious society in China and stipulated a target to train 500 000 social workers by 2015, and 1 450 000 social workers by 2020. The policy paper outlined the systematic strategies to be adopted by the government to train different levels of professional social work personnel and the measures to be taken to support, train, monitor, and evaluate the quality of social work manpower nurtured through this process. The policy paper was received with great excitement and joy by the social work community in China and it is expected that the social work profession will be greatly strengthened through the process and will be able to make significant impact on social betterment in China.

Salient Features of the Development of Social Work

Strong Leadership Role of the Government

The Chinese government has played a pivotal role in supporting the development of social work in the country. Despite the phenomenal economic achievements brought about by the introduction of market economy, social problems and social conflicts have accelerated. At the turn of the century, critical social issues including the widening gap between rich and poor people, disparity between urban and rural areas, corruption, environmental contamination, and ethnic conflicts had reached alarming levels. President Hu Jintao's 2004 announcement to adopt the 'harmonious society' as the guiding principle to develop the country, and the subsequent pronouncement in 2006 of the national policy to develop 'a grand team of professional social workers to help build a harmonious society' reflected the Chinese government's determination to address escalating social problems through different types of social work interventions (Liu, 2012; Xiao, 2009). Measures introduced by the government to enhance social work development include developing social work positions in different levels of government; supporting the establishment of NGOs; supporting the development of social work education programs at the post-secondary, degree, and MSW levels; and developing pilot projects on services for different target populations (Liu, 2012). The government sees its role as the main provider of social welfare resources with the aim of serving its people and gaining legitimacy for social governance. As such, it has been predicted that the Chinese government will remain the preeminent agent in setting social policies and designing and providing social services; and will also be the major employer of practitioners performing social work and social service functions in China (Tsang et al., 2008; Wang, 2004).

Social Work Education as a Catalyst for Social Change and Development

The active involvement of social work educators and the CASWE had been pivotal to the professionalization process of social work in China (Wang, 2012; Yuen-Tsang and Wang, 2008; Shi, 2012). Since the 1990s, key actors in CASWE have been extremely proactive in seizing all opportunities to promote social work and influencing government policies to recognize social work as a profession in China. The CASWE formed both regional and national networks among social work educators to facilitate communication; conducted training workshops to enhance professional competence among educators; developed practicum sites to provide quality fieldwork placements for students; arranged overseas exchange programs for educators and students; published social work textbooks to encourage the development of indigenous teaching materials; organized national and international conferences to encourage sharing and collaboration; conducted reviews on social work programs to assure quality; and provided consultation to the government on policies relating to social work and social work education. Moreover, in almost all the national task forces and commissions relating to social work and social work education, the government involved social work educators as consultants and experts and took their recommendations seriously in their deliberations. The role of social work educators in China as the catalyst steering the direction of the development of social work in China is evident.

Active Agency of the Social Work Community in Coconstructing the Social Work Profession

The strong partnership relationship among members of the social work community, including educators, government officials, nongovernmental organization leaders, and front-line practitioners, has also played a significant role in facilitating the development and coconstruction of a social work profession appropriate to the Chinese context. Key stakeholders of the social work community had interacted in a dynamic manner to promote and coconstruct the social work profession (Wang, 2012). They were close allies in promoting social work in the country and leverageding each other to develop policies and practices appropriate to the Chinese context. While the social work educators took active steps to generate social work practice theories and models through action research and pilot projects and influence policy formulation through consultation processes, many initiatives spearheaded by local governments and NGOs also contributed to the development of demonstration projects that impacted national service development. The active agency of the entire social work community in China throughout the past two decades has been instrumental in facilitating the professionalization process and the coconstruction of a social work profession appropriate to the sociocultural context of contemporary China (Yuen-Tsang and Wang, 2008).

Emphasis on the Indigenization of Social Work in the Chinese Context

During the early 1990s when social work was reintroduced in China, there was strong temptation for Chinese social work educators and social workers to borrow social work curricula and practice models without any change, from foreign countries. Some social work educators zealously advocated the adoption of a national social work curriculum to be modeled after the standard curriculum already developed in countries such as the United States and the United Kingdom. The tension and debates on whether to follow global trends and standards or develop indigenous theories and practice appropriate to the local context had been active in the profession since its reintroduction in China. While some believed that social work should follow international standards and could borrow directly from overseas models, those who advocated the use of the indigenous approach were convinced of the need to identify knowledge as understood and experienced by people within its own culture and felt that steps had to be taken to understand psychological and social phenomena as they existed within their cultural contexts (Bond and Hwang, 1986; Ge, 2004; Kim and Berry, 1993; Yang, 1999). Most leaders of the Chinese social work community appeared to have subscribed to the view that international experience should be integrated within the practical reality of China and that serious efforts should be made to indigenize social work practice in China's contemporary sociocultural context (Tsang and Yan, 2001; Wang, 1988). This discourse strongly influenced the subsequent development of social work in China and many experiments, such as the 'Shanghai model' and 'Shenzhen model,' were launched with the aim of discovering social work practice models appropriate to the cultural context of contemporary China.

Opportunities and Challenges Confronting the Future Development of Social Work

The development of social work in China seems to be extremely promising in view of the significant role that the profession is expected to play in national development. However, there are many critical issues confronting the profession which must be seriously tackled before the profession can achieve its expected goals and bring about the anticipated impact.

Regional Disparities in Social Work Development

The regional disparity between different regions in China is extremely wide. While the economic advancement in the coastal regions has been phenomenal since the introduction of the open-door economic policy, economic development in many parts of the Middle and Western regions of China remains seriously inadequate. These regional disparities have given rise to imbalanced development not only in the economic sphere but also in social development and welfare provisions (Shi, 2012). The widening gap between the income and living standards of urban and rural populations has been further aggravated by globalization processes (Zheng, 2010). The marginalized living experience of migrant workers in urban areas as a result of their lack of the official *hukou* (household registration) status in the cities and their consequent ineligibility to access government welfare and education provisions (Cheng and Selden, 1994; Fan, 2008; Wong et al., 2007) had

given rise to growing resentment and resistance among migrant workers (Chan and Pun, 2009; Pun et al., 2009; Pun and Lu, 2010). The magnitude of problems confronting local populations in the poorer regions has given rise to conflicts and discontent and may become sources of civil unrest. As such, the government is seriously contemplating the development of measures to redress such issues and ensure social harmony. The provision of welfare benefits and social work services is one of the viable options. But the relative lack of resources in these regions and the shortage of professional social work personnel have made it extremely difficult to develop the needed social work services in these localities. It is, therefore, a matter of urgency to develop a system of social work education and social services which is appropriate for the poorer regions of China to redress regional imbalances and provide needed assistance to help them alleviate poverty, address social problems, build local capacity, and enhance social harmony.

Tension between Meeting Needs of the Masses and the Demand for Professionalization

Another critical issue is the tension between the quantity and quality of social work education programs and social work services to be provided. Given the huge population of China and the complexity of its social problems, it is extremely difficult to provide adequate social work services for the majority of the population. The recently announced policy of the Chinese government on the '2011–20 Medium Long-term Strategic Plan on the Development of Social Work Professional Manpower Team' with the ambitious goal of training 500 000 professional social workers by 2015, and 1 450 000 social workers by 2020 is an indication of the urgent need for professional social workers to provide social services for the population. But given the lack of adequate qualified social work educators to provide training for the social work trainees; the lack of a system of well-established social services units and organizations to provide quality practicum sites; and the lack of a vibrant social work professional culture in China, it is doubtful whether the desired target could be reached in an effective manner. The need to satisfy the basic social service needs of the country within a short period of time will create tensions between quantity and quality and will require compromises.

Tension between Westernization and Indigenization

The constant tension between the tendency to borrow directly from the West and need to search for local practice models which are culturally appropriate has been a regular feature of the Chinese social work community since the early years of the profession's reintroduction in China. The tension seems to have resurfaced in recent years because of the sudden growth of social work education programs and social services in China following the government announcement of policies to professionalize social work. The high percentage of teachers of the newly approved MSW programs who have not gone through professional social work education themselves and their relative lack of direct supervised practice experience may negatively affect their ability to integrate theory with practice and to teach social work in a culturally sensitive manner.

Similar tensions are surfacing in the direct practice arena because of the sudden upsurge of government supported NGOs in China. As a response to the government's newly introduced 'purchased service policy,' many NGOs have been established during the past few years to provide services for needy service users. While the service workers of most of these newly developed NGOs are highly motivated, many of them have received little professional training and their organizations do not have well-established management structures and professional support services. As a result, the tendency to borrow directly from existing practice models developed in the West has become common practice. Being acutely aware of the problem, the government has been encouraging university social work departments to form their own NGOs or to actively engage with NGOs in order to provide placements for their students and at the same time provide services for specific target groups. This approach will hopefully encourage social work educators to engage in direct practice together with their students, and will help in generating local social work practice theories through the process.

Lack of a Strong Knowledge Base to Support the Rapid Development of Social Work

The reintroduction of social work to China only took place in the late 1980s and the profession was not given much recognition and support until very recently. During the past two decades, the social work academic community was very weak in terms of size, status, and social recognition as compared to other academic disciplines in the academy and social work was perceived as an applied discipline which did not require the support of rigorous research activities. Consequently, the resources devoted to supporting the development of social work were extremely limited and social work educators had encountered great difficulties in securing financial, personnel, and other resources to support their academic and scholarly activities. The relative lack of a strong knowledge base to support the rapid development of the profession negatively affected the quality of services to be provided and thus the inability of planned services to achieve intended objectives. Serious and systematic efforts are therefore urgently required to enable social work educators and practitioners to generate social work theories and practice models which are firmly grounded in solid practice research, with a view to providing strong theoretical support to guide the future development of professional social work practice in China.

Conclusion

The development of social work in China has gone through dramatic changes since its introduction in the early 1920s. Members of the social work community have been active agents in facilitating positive social change and development in society and have made significant impact on social work education, practice, and policy advocacy during different historical periods. They have not only responded to issues emerging in their particular historical and social contexts, but also contributed significantly to the professionalization of social work in the country through practice innovation, theory

building, and policy advocacy. Even though the social work profession in China is still in an early stage of development and there are numerous issues and challenges waiting to be addressed, it seems that the social work community is ready to take up the challenge with vision, determination, and high motivation. It is expected that the development of social work in China will be extremely promising not only in terms of quantity and scope, but also in terms of quality and social impact. Moreover, it is envisaged that social work development in China will not only contribute to social betterment within its own country, but may also have significant impact on and contribution to the development of social work worldwide.

See also: Asia, Sociocultural Aspects: China; Indigenous Psychology; Indigenous Social Work; Sexual Minorities in China; Social Work Education; Urbanization in China.

Bibliography

Bond, M.H., Hwang, K.K., 1986. The Social Psychology of Chinese People. Oxford University Press, New York.

Chan, C.K.C., Pun, N., 2009. The making of a new working class? A study of collective actions of migrant workers in South China. The China Quarterly 198 (1), 287–303.

Cheng, T., Selden, M., 1994. The origins and social consequences of China's hukou system. The China Quarterly 139, 644–668.

China Association of Social Work, 2009. Reports on Development of Social Work in China (1988–2008). Social Sciences Academic Press (China), Beijing.

Fan, C.C., 2008. China on the move: migration, the state, and the household. The China Quarterly 196, 924–956.

Ferguson, K.M., 2005. Beyond indigenization and reconceptualization: towards a global, multidirectional model of technology transfer. International Social Work. 48 (5), 519–535.

Gamble, S., Burgess, J., 1921. Peking: A Social Survey. Oxford University Press, Oxford.

Ge, L.J., 2004. The unique exploration of personality in chinese indigenous psychology. Journal of Central China Normal University (Humanities and Social Sciences).

Jun, X., 1996. Baptised in the Fire of Revolution: The American Social Gospel and the YMCA in China. 1919–1937. Associated Press, Inc., New York.

Kim, U.E., Berry, J.W., 1993. Indigenous Psychologies: Research and Experience in Cultural Context. Sage Publications, Inc, Newbury Park.

Lee, M.K., 2000. Chinese Occupational Welfare in Market Transition: Beyond the Iron Rice Bowl. St. Martin's Press, NY.

Lei, J.Z., Shui, Z.Z., 1991. The thirty years of the social work programme of Yenjing University. In: Chow, N., Yuen-Tsang, A.W.K., Wang, S.B., Zhang, H., Ma, F.Z., Li, Y.M. (Eds.), Status Quo, Challenge and Prospect: Collected Works of the Seminar of the Asian-Pacific Region Social Work Education. Peking University Press, Beijing, pp. 10–16.

Leung, J.C.B., 1994. Dismantling the 'iron rice bowl': welfare reforms in the People's Republic of China. Journal of Social Policy 23 (3), 341–361.

Leung, J.C.B., Nann, R., 1995. Authority and Benevolence: Social Welfare in China. Chinese University of Hong Kong Press, Hong Kong.

Liu, Z., 2012. Speech delivered at the opening ceremony of the China-US Collaborative on the Development of MSW Programmes held at the Peking University-Hong Kong Polytechnic University China Social Work Research Centre. Beijing, on 12 December, 2012.

Midgley, J., 1981. Professional Imperialism: Social Work in the Third World. Heinemann, London.

Pang, X.L., 2010. Safeguarding and Creation: Social Work in Contemporary China. Shijiazhuang, Hebei.

Preston, P.W., Haacke, J., 2003. Change in contemporary China. In: Preston, P.W., Haacke, J. (Eds.), Contemporary China: The Dynamics of Change at the Start of the New Millennium, Chapter 16. Routledge Curzon, London, pp. 3–24.

Pun, N., Lu, H.L., 2010. Unfinished proletarianization: self, anger, and class action among the second generation of peasant-workers in present-day China. Modern China 36 (5), 493–519.

Pun, N., Chan, C., Chan, J., 2009. The role of the state, labour policy and migrant workers' struggles in globalized China. Global Labour Journal 1 (1), 8.

Shi, B.N., 2012. The construction of a system of social work administration system with Chinese characteristics. Social Work Journal. Jianxi 6, 4–9.

Tsang, A.K.T., Yan, M.C., 2001. Chinese corpus, western application: the Chinese strategy of engagement with western social work discourse. International Social Work 44 (4), 433–454.

Tsang, A.K.T., Sin, R., Jia, C., Yan, M.C., March 2008. Another snapshot of social work in China: capturing multiple positioning and intersecting discourses in rapid movement. Australian Social Work 61 (1), 72–87.

Wang, S.B., 1995. The experience and development of social work in China. Social Sciences in China 2 (1), 97–106.

Wang, S.B., 1998. New force in Chinese social welfare: the development of social work in China. In: Wang, S.B., Leung, J.B., Mok, T.K. (Eds.), Social Welfare in China. Zhong Hua Publishing Company, Hong Kong, pp. 193–215.

Wang, S.B., 2004. The age of social policy and government capacity-building in social policy. Social Sciences in China 4 (2), 117–121.

Wang, Y., 2012. A Study of the Professionalization Process of Social Work in the Chinese Mainland: Interaction of the State, the Society and the Academic Community (1978–2006). PhD Thesis submitted to The Hong Kong Polytechnic University, Hong Kong.

Wong, L., 1996. The dilemmas of financing social welfare in China. In: Lo, T.W., Cheng, J.Y. (Eds.), Social Welfare Development in China: Opportunities and Challenges. IL: Imprint Publication Company, Chicago, pp. 69–84.

Wong, K., Fu, D., Li, C.Y., Song, H.X., 2007. Rural migrant workers in urban China: living a marginalised life. International Journal of Social Welfare 16 (1), 32–40.

Xia, X., Guo, J., 2002. Historical development and characteristics of social work in today's China. International Journal of Social Welfare 11 (3), 254–262.

Xiao, X.L., 2009. Social Harmony Concept and Public Policy Choices. MPhil Thesis submitted to the Nanchang University, Jianxi, China.

Xiong, Y.G., Wang, S.B., 2007. Development of social work education in China in the context of new policy initiatives: issues and challenges. Social Work Education 26 (6), 560–572.

Yang, K.S., 1999. Towards an indigenous Chinese psychology: a selective review of methodological, theoretical, and empirical accomplishments. Chinese Journal of Psychology 41 (2), 181–211.

Yuen-Tsang, A.W.K., 1996. Social work education in China: constraints, opportunities and challenges. In: Lo, T.W., Cheng, J.Y. (Eds.), Social Welfare Development in China: Opportunities and Challenges. IL: Imprint Publication Company, Chicago, pp. 85–100.

Yuen-Tsang, A.W.K., Wang, S.B., 2002. Tensions confronting the development of social work education in China: challenges and opportunities. International Social Work 45 (3), 375–388.

Yuen-Tsang, A.W.K., Wang, S.B., April 2008. Revitalization of social work in China: the significance of human agency in institutional transformation and structural change. China Journal of Social Work 1 (1), 5–22.

Zheng, H.Q., 2010. Rural social problems and rural social work in the context of globalization. In: Wang, S.B. (Ed.), New Starting, New Development. Annual Conference Proceedings of China Association for Social Work Education (2006–2010). Peking University Press, Beijing, pp. 563–578.

Social Work in Eastern Europe

Darja Zaviršek, University of Ljubljana, Ljubljana, Slovenia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Historical, economic, and social differences as well as common experiences of state socialism in Eastern Europe made social work a profession with many commonalities and differences across the region. With some exceptions, university-based social work education began after 1991. Modern social policies and social work practices have developed during the time of transition from communism to a free market, which produced economic hardship, ethnic wars, racism against ethnic minorities silenced under communism reemerging, increased social inequalities, large-scale migration, and a new understanding of individual needs and people's everyday desires, all reflected in social work practice.

Introduction

Eastern Europe politically and socially is part of Europe (although not always the case geographically) and encompassed countries embedded in state socialism as the dominant political system for approximately 50 years. These include former Soviet Union republics – Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russian Federation, Tajikistan, Turkmenistan, Ukraine, Uzbekistan, Albania, Bosnia and Herzegovina, Kosovo, Macedonia, Montenegro, Serbia, and countries that have become part of the European Union in 2004 (Czech Republic, Hungary, Poland, Slovenia, and Slovakia, including the three Baltic countries already mentioned), in 2007 (Bulgaria, Romania), and in 2013 (Croatia).

Their communist past has been an important marker that influences today's social work education, social welfare institutions, and social work knowledge and practice, but it is not the only signifier. Different historical powers (belonging to three different empires – Habsburg, Ottoman, and Russian Tsars), geographical and economic differences, large disparities between urban and rural areas, religious influences (Roman Catholic, Orthodox, Muslim), multiethnic territories, and relations between ethnic groups, influence everyday life among Eastern Europe's diverse populations.

After the change in political regimes, the 'iron continent' became fragmented through ethnic wars, massive economic crises, migration, and state-building ventures among the former Soviet Union and Yugoslav Republics alongside international political players (the World bank, International Monetary Fund, and European Union institutions), which influenced economic lives and social policy/legislation of the countries in the region. The area acquired different names according to geographic, economic, or social specificities or international interests (Central Europe, South East Europe, the Balkan region, the countries of the former Soviet Union, and Russia). Some names are historically loaded and negative. The pejorative word 'Balkanization' became used globally to describe phenomena prone to unpredictability, fragmentation, and conflict. Some Eastern European countries were defined as in need of 'Europianization' despite their geographical placement within Europe. Many countries were urged to stop human rights violations of minority peoples if they wanted to become part of the European Union (at least formally). This had some positive impact in democratizing societies and social work (Sasse, 2009). Across the region, a wide specter of different social work practices coexist, varying from very innovative, radical, and liberating ones for persons requiring services and support to oppressive, traditional, and narrow-minded pathology-based responses by professionals and social work institutions.

Social Work's Historical Development before 1991

Despite the belief that social work did not exist in Eastern Europe prior to 1991, there is considerable research showing that professional social work education was established in some countries in the region after World War I, but was disrupted and closed down at the end of the World War II, or soon afterward (Hering and Waaldijk, 2003; Schilde and Schulte, 2005; Hering and Waaldijk, 2006).

At the beginning of the twentieth century, Hungarian, Polish, Romanian, and Bulgarian founders of social work education periodically went hand in hand with the Western counterparts following the idea that social work has to become a profession based on proper education and training – Margit Slachta in Budapest, Helena Radlinska in Warsaw, Princess Ileana in Bucharest, Rayna Petkova in Sofia. They struggled against economic, social, and symbolic inequalities of populations during times of quick and brutal industrialization and deruralization of societies and some fought for the so-called 'woman's question' – women's right to vote, the right for civil marriage and divorce, and more gender equality.

During the period of state socialism, when existing social work training, schools, and institutes were closed down, some communist countries replaced them, e.g., 'social care assistance training' in Romania, 'social nurse' program in Czechoslovakia, and 'social service workers' training in Poland (Iarskaia Smirnova, 2013). For the communist leadership, social work was associated with capitalist ideology, seen as backward, and a counterideology compared with the socialist idea that universal employment would ensure that all other economic, social, and emotional needs of people were addressed by the system of universal social and health protection. Additionally, it was believed that the state as a 'good father'

would protect children, older people, and disabled people through a wide range of long-stay social institutions. Socialism was seemingly able to eradicate the need for social work interventions and ensure the well-being of every 'human being.' Additionally, the omnipotent communist leaders did not allow any innovations except those that came from the Communist Party. One of the first Yugoslav social work teachers who pleaded with the highest Communist Party members to open the school for social work in Slovenia in 1955 expressed their attitude:

> We could write about what happens in the West and about social work abroad, but always in such a way, that it was obvious that ours [Communist Party of Yugoslavia] had written or said that already before them [the Westerners]. We had to be always in front of anyone else (Zaviršek, 2005). (Katja Vodopivec (1917–2012) was a professor of law. In 1958, she wrote the first Yugoslav book on social work methods (in Slovenian language) which was not allowed to be used by social work students and teachers during communism) (Zaviršek, 2005a, 2012)

Yugoslavia was the only communist country that had established schools for social work after the communist take-over compared to 1952 in Croatia, 1955 in Slovenia, 1957 in Macedonia and Serbia, and 1958 in Bosnia and Herzegovina. Today, this is perceived as the unique development of social work education under communism (Zaviršek, 2005a,b, 2008, 2012). This top-down decision making has to be understood contextually, partially as the personal influence of some women Communist Party members who were commended for their commitment during the partisan struggle, were active in the area of welfare protection, and maintained Cold War equilibrium. When the Yugoslav Communists ended their close relationship with Stalin (Cominform split in 1948), they started to cooperate with the United States to avoid political isolation and economic vulnerability (Bilandžić, 1985). Soon afterward, American and Scandinavian advisers became involved in the processes of establishing the first social work curriculum in Croatia in the early 1950s, and social welfare institutions called 'centers for social work' throughout Yugoslavia in the early 1960s. Centers for social work have remained the major welfare institutions in all parts of former Yugoslavia until now.

Social work education also served as a useful way to carry out several goals of the Communist Revolution and partially change the gender order. Social work training ensured education for masses of women and was therefore potentially emancipating. It was simultaneously seen as suitable for retaining traditional femininity and women's 'natural vocation' intact. It created new jobs for women who became employed in large long-stay welfare institutions – nursing homes, boarding schools for disabled children, and asylums for people with disabilities, and was proof that the Communist Party cared about women, especially the peasantry and those from poor rural areas. Large, semiclosed places of welfare state institutions were meant to replace women's unpaid caring work at home. The institutionalization of social work training gave the Party control not only over welfare and gender regimes but also 'in the field,' where 'deviant' behavior such as alcoholism, work hatred, and prostitution could be carefully monitored.

The official definition of the socialist social work was that "social work carries out the goals of social policy" and social workers were primarily expected to work in the areas of social assistance delivery, family protection (which included foster care for children who had lost parents during the war, divorce, and single mothers issues), and to prevent 'deviancies.' Social work education was imposed from above to influence the everyday life of people from below.

From State Socialism toward Neoliberalism

The 'emergency welfare state' (the term used by Tomasz Inglot (2009)) processes after 1991 included new social policy legislation and the development of the mixed system of social services (governmental, nongovernmental organizations (NGOs), and private providers). The times of economic liberalization were harsh for most Eastern European countries and the majority of people (Jäppinen et al., 2011; Bessudnov et al., 2011). The existing industries and state enterprises collapsed and a period of enormous unemployment, among men, women, and ethnic minorities (especially Russians in the Baltic countries, and Roma people in the Czech Republic, Slovakia, Slovenia, Croatia, Romania, Hungary, and Bulgaria) started to dominate everyday life. Poverty expanded and remained the central fact of life. It affected most heavily the rural areas of the region, from where people migrated to capital cities, or from its more eastern parts toward the western parts, and from southern parts toward the north (from central Asia to Russia, from Ukraine to Poland; from Kosovo and Albania to Croatia and Slovenia). In Albania, rates of poverty in rural areas are almost 70% higher than in the capital (Ymeraj, 2007).

People faced many breaks in their everyday routines: privatization of publicly created goods during state socialism (kindergartens, local health centers, factories), flexibilization of labor and consequently unemployment after a long period of full employment, early retirements of a huge number of workers in some countries (Vanhuysse analyzed this phenomenon as the deliberate governmental strategy against the potential resistance toward the neoliberal reforms in Poland and the Czech Republic (2009).), ethnic conflicts and wars, economic migration and the transformation of extended family care systems and family life, the development of huge economic inequalities ('the new rich'), and the return of religious powers as the key political and social players in secularized societies. Eastern European countries moved from state socialism to neoliberalism. Many researchers revealed that international financial institutions heavily influenced national governments' decision making in the economic and social policy areas and worked toward the neoliberalization of the whole region (Deacon et al., 2007).

Violent ethnic conflicts (Croatia, Macedonia), ethnic wars (Armenian, Azerbaijan), and ethnic cleansing and mass killings (Bosnia and Herzegovina, Kosovo), which appeared almost immediately after the fall of Yugoslavia and Soviet Union, caused deaths, disability, forced migration, voluntary and compulsory remigration, increased poverty, transgenerational

loses, and trauma. In Bosnia and Herzegovina, 100 000–300 000 persons were killed and out of a population of 4.4 million, 1 million fled the country between 1992 and 1995 (Maglajlić Holiček and Rašidagić, 2007). Similarly, 850 000 persons fled from Kosovo in 1999 and 360 000 Kosovans sought refuge in neighboring Macedonia, comprising 17% of the whole population of Macedonia (Gerovska Mitev, 2007; Cocozzelli, 2007). After Kosovo proclaimed independence in 2008, social work students disputed whether they are ethnically 'Albanians,' 'Moslems,' or 'Kosovans' and remained divided into three different ethnic identities (personal notes, University of Pristine, 2008).

During state socialism, the nineteenth and twentieth centuries' primordial understandings of the blood–kin relationship based on singular ethnic belonging were replaced by the pragmatic communist ideology, which suppressed ethnic identities to construct a singular, but new identity of a 'socialist proletariat.' After 1991, ethnicity became a 'new' political and social identity for the vast majority of people. This identity was not formally recognized and had not been individually important for almost half the twentieth century. The 'imagined identities' were mobilized by political elites for a new political project of shifting national and state borders. In 1992, the new Slovenian government deprived more than 25 000 persons (official holders of Yugoslav passport who came to Slovenia as economic immigrants) of citizenship rights and erased them from the register of permanent residents of the Republic of Slovenia. They lost all social and political rights and were treated as 'illegal migrants' vulnerable to detention and deportation (Zorn, 2009). In Baltic countries, the exclusion of internal 'others,' e.g., the Russian ethnic minority, helped to construct seemingly ethnically homogeneous states.

Roma people, who are today the largest ethnic minority in Europe, have suffered severely during the past 20 years. There are 9–12 million Roma people in the region, with 2.5 million in Romania. The figures are estimates and include people who have defined their Roma ethnic belongings during National Censuses and those who have not (mostly because of racism). Roma men have lost their jobs due to closures of heavy industry and poor education. Additionally, they have lost alternative sources of earning (e.g., iron, collecting plastic bottles, simple craft, and music) and remained largely without income, and dependent on social assistance money. Many Roma fled ethnic wars. In Kosovo, for instance, Roma villages were burnt after the war, as the local population defined Roma as Serb collaborators. Many of them became asylum seekers, illegal migrants, and long-term homeless. Unemployment levels among Roma people in all countries of Eastern Europe are extremely high, currently ranging from 60 to 90%. Roma women are transgenerationally marginalized, living mostly in segregated settlements that they barely leave. They are unemployed, sometimes illiterate, but receive welfare assistance if they have children. Regardless of the size of Roma populations, all Eastern European countries articulate nationalist rhetoric against Roma peoples ranging from moderate to extreme, some ethnically motivated hate speech, and violent attacks against Roma settlements (Vermeersch and Ram, 2009). This ethnic hatred is the product of a long history of exclusion in the precommunist period, during communism (when Roma people lacked long-term industrial employment and were categorized in the lowest class among socialist citizens, called the lumpenproletariat), and today.

Economic crises and neoliberal transformations in national governments made thousands of people migrate to Western countries to seek jobs and support those who stayed in their country of origins with remittances. In Albania, of 3.1 million people, a quarter (mainly young men) left the country between 1990 and 2005 for Greece and Italy (Ymeraj, 2007). Thousands of women from Ukraine and Lithuania have become care workers, comprising part of the phenomena known as 'global motherhood' (Ehrenreich and Hochschield, 2003). They care for the family of their employers abroad while providing economic care for their families, especially their children, at home. In Germany alone, there are 200 000 female care workers from Eastern Europe with more than 30 000 from Lithuania. In the Ukraine, some estimates show, that in every third family at least one woman member is a migrant worker abroad (Tolstokorova, 2010).

Remittances from Eastern European emigrants have become either a crucial income source or the only family income that helps numerous families to survive. These are therefore desired by governmental officials. Only recently has large-scale migration by women become a governmental concern. In Lithuania, the mass media speak about the 'national epidemic,' while in Ukraine, social workers and other social service professionals speak about the 'Italian syndrome' (abandoned and lonely children receiving monthly incomes from their mothers working mainly in Italy). Some creative responses by social workers were developed to support children who are left alone, e.g., Lithuania introduced temporary guardianship for grandparents who care for children of migrant workers (Malinauskas, 2011). Not only relatives but also friends, neighbors, and school teachers provide social parenting for these children. Social parenthood has become a reality for a growing number of adults and children (Zaviršek, 2009b).

Nevertheless, many researchers express concern that in poorer Eastern European countries, the system of home care is entirely based on the unpaid care work of family members, mostly women, who are today leaving home to seek paid care work abroad (Prochazkova and Schmid, 2009). Older people who are left alone can afford to pay neither state-run nursing homes nor even more expensive private ones. In some countries, social workers are involved in poorly paid community-based home assistance services for the elderly.

Post-1991 Social Work Practice in Eastern European Countries

After 1991, new approaches in social work practice focused on individual and family counseling, the development of community-based services that aimed to replace large long-stay state institutions (especially regarding the care of 'social orphans'), foster care that aimed to replace the system of orphanages, international adoptions, violence against women and women's shelters, crises centers for children, day centers for people with disabilities, and activities linked to 'social inclusion.' For service users, 'social inclusion' was the consequence of early exclusion on the basis of

medical diagnosis seeing intellectual disability, mental health diagnosis, ethnic belonging (identity) as pathology, and early long-stay institutionalization (sending children into boarding school far away from home). Economic, social, and symbolic exclusions created by social workers, health professionals, and pedagogues have been followed by the welfare processes of 'social inclusion' involving the same professionals. Roma children, for instance, who had been labeled as intellectually disabled and sent to special schools across the region, have recently become welfare objects of 'social inclusion.' Not only the countries that became part of the European Union but also other states from the region have adopted modern legislation and policies that have foreseen social inclusion, community services, and deinstitutionalization, but lack implementation. The majority of countries signed and ratified the UN Convention of the Rights of People with Disabilities but in countries in Central Asia, activities for people with disabilities are almost entirely supported by international NGOs and done by the mothers of the persons with disabilities (Zaviršek, personal notes, September 2012, Bishkek; Rasell and Iarskaia-Smirnova, 2013).

The delivery of social assistance money to people who are unemployed, poor families, ethnic minorities, elders, and persons with disabilities is a main part of social workers' activities. In wealthier countries in the region with a wide range of long-stay governmental institutions and community-based welfare services (day centers for people with multiple diagnosis, sheltered workshops, crises centers for children in need, family helpers system, women refuges, and group homes for people with mental health problems and other disabilities) social workers provide placements to these institutions and organize payment contracts usually covered by local municipalities, the state, and the person and his or her relatives. Most of the welfare facilities are located in cities and people from rural areas have difficulties accessing them or must send relatives far away from home to long-stay institutions. The 2008 economic crises caused by international financial organizations forced radical cuts in social protection. Consequently, even more social work activities are focused on social transfer delivery, while community development activities have diminished.

Many types of welfare organizations, styles of work, and approaches have remained the same until the present, for example, long-stay institutions for disabled people despite pilot projects on deinstitutionalization, the denial of most legal and actual rights of children and adults with intellectual disabilities, the nonexistence of independent advocacy work, the denial of self-determination of service users, and the absence of social action approaches in deprived communities. Several conceptual fallacies persist. Many professionals see violence against women in the private sphere as individual pathology instead of structural inequality, welfare systems prioritize closed adoptions for the 'child's own good' instead of open ones, and many children and adults with disabilities are considered well protected in segregated boarding schools.

Social workers from governmental welfare services avoid any form of community action in Roma settlements, and the relationship between Roma and social workers remains highly conflictual (social workers take children into foster care, give conditional social transfer money, and blame mothers for not sending children to school). Local and international NGOs usually work in Roma settlements, mostly organizing different projects with children. Women's empowerment work and awareness-raising on how to protect themselves from violent men are sometimes undertaken by local and international NGOs, but the area remains highly contested. Discussions about child maltreatment and early marriages of Roma girls are avoided and silenced. A small number of *gojim* social workers have developed a strong trusting relationship with Roma men to tackle issues of violence against women and children. There are innovative projects in Bulgaria and Romania, where international NGOs together with social work students and local people set up Roma cooperatives in their settlements.

Local and international NGOs cover many different social work areas. Some social work innovations were only possible with the use of the international money, for example, violence against women, support for rape victims, the development of personal assistance projects and independent living for persons with severe disabilities, deinstitutionalization initiatives, cooperatives for people with long-term mental health problems, support services for mothers with children with impairments, counseling for stigmatized people living in the same-sex partnerships and transgendered people, medical aid and counseling for people affected with human immunodeficiency virus and acquired immunodeficiency syndrome. In Belarus, disability activists try to make the government sign the UN Convention of the Rights of People with Disabilities and a well-known organization for the prevention of women's slavery, *La Strada*, works against the trafficking of women and children across the borders of Russia, Belarus, and Ukraine.

The overrepresentation of the international NGOs had negative impacts too. A joke known across the region during the 1990s has addressed the phenomena: an Eastern European family consists of a mother, father, siblings, grandparents, aunts and uncles, and a local NGO! A common critique against international NGOs was that they dictated the topics of concern and ignored the context and needs of local populations, creating a 'project culture' instead of needs-based provisions (Arandarenko and Golicin, 2007). In some of the poorest countries in the region, international NGOs locally distributed food and other goods to people in need, which established them as the main players among people. A director of a state center for social work from Kosovo recalled: "They had the money to help people, and we didn't have anything to distribute. Then we lost even more credibility" (Zaviršek, personal interview, May 2009, Pristine).

Economic hardship has been experienced by social workers too, and poor salaries have pushed many out of the field (especially from local NGOs). There is no research about the impact of this situation on the quality of social work practice, but some personal stories of social workers show that they individually identify with their clients. A Kosovan social worker who has worked at the center for social work for 30 years has shared this personal story: "My son woke me up in the morning asking me for the money to buy bus tickets. I told him I couldn't give him the money as I didn't have it. Then I thought to myself that I became 'a social assistance case'. I made myself ready to go to work, went to the center for social work where

the clients already waited for me" (Zaviršek, personal interview, May 2008, Pristine.).

In the past decade, all countries of Eastern Europe started to promote market principles in social welfare, especially in the area of nursing homes for older people, sheltered workshops for people with intellectual disabilities, and individual counseling for private customers. The economic crises hit this area of welfare provision too. Private welfare businesses today seek opportunities in areas with cheap labor (establishment of older people's homes in eastern parts of the region for people from its western parts). Recently in Slovenia, older people and their relatives cannot afford to pay expensive residential care anymore, and long waiting lists to enter long-stay private institutions dwindled as facilities emptied. In one of the older people's home, staff went on strike because they had not received their salaries for months.

The Academization of Social Work and Social Work Networks

In 1988, the European Association of the Schools of Social Work conference was held in an Eastern European country for the first time. An anecdote from that conference manifests the *Zeitgeist* and differences between East and West:

> Coming from the West to the conference held by a socialist country, we expected to share with you our new enthusiasm about community work and collective action instead of focusing predominantly on the individual person. But the Yugoslav social work teachers were all against those ideas. We said, "The community work is the right thing to do!" And they said, "No, the work with individuals is what shall be done in social work." (personal communication, 2007)

For Yugoslav social work teachers who were forced to prioritize the community over the individual, the shift toward more individualistic approaches in social work was seen to be something radical and unwanted by the Communist regime. For the colleagues from the West, the turn toward the collective was a radical move, given their individualistically oriented societies that prioritized the individual above the collective, so community action was a more important shift for their social work theory and practice.

After 1991, all countries of Eastern Europe, with the exception of Turkmenistan, established social work education, mainly as a 4- or 5-year university program. The country with the largest number of social work undergraduate programs is the Russian Federation (175 social work departments), followed by Ukraine (50 social work programs), Kazakhstan (22 social work departments), Czech Republic (15 programs), Azerbaijan and Kyrgyzstan (seven universities where social work has been taught), and Romania (four universities) (Iarskaia Smirnova, 2013). The universities with social work programs that have been developed lately are in Montenegro (2003), Uzbekistan (2004), Moldavia (2007), and Kosovo (2012). After 2000, intensive development in postgraduate master-level studies appeared and intensive international collaboration in teaching and student exchanges took place. The processes of the academization of social work are seen in research and the establishment of doctoral studies in social work already in Armenia, the Czech Republic, Estonia, Georgia, Romania, and Slovenia. A cross-national collaboration is also shown in the first international doctoral program in social work and social policy (Indosow) from 2009, which is run by a consortium of the schools of social work from Eastern and Western Europe and was initiated by an Eastern European school of social work (Zaviršek, 2009a).

Today, most of the countries have national associations of social workers and some of them, like the Czech Republic, also have an association of social work educators (since 2009). Social work academics have been active in establishing a number of academic journals mostly in local languages (*The Journal of Social Policy Studies*, Saratov University; *Social Work*, University of Ljubljana, est. in 1962; *Social Policy and Social Work*, University of Belgrade, est. in 1971). Some of the journals contain a mixture of local and foreign languages (*Social Work Review*, University of Bucharest; *Social Work Yearbook*, University of Zagreb), while some are published in English only (*Social Work in Transition*, Kiev Mohyla and Sheffield Hallam Universities). (Only to mention few of them, although almost all countries in the region have academic social work and social policy journals.) In 2008, the schools of social work from the region have established the East European Sub-Regional Association of the Schools of Social Work, which is a large network of social work scholars from the region.

One disturbing characteristic of university-based social work education in some countries is the high number of social work students who enroll because they lack employment opportunities. The quantity of students threatens the quality of social work education. Some social work programs have become hybrid academic–welfare institutions that 'care' for young people and keep them 'in the system' until the age of 25 years or more. Many students seek motherly care instead of academic training from predominantly female members of the academic staff. The University of Sarajevo is a blatant example, but not the only one. Its Social Work Department enrolls 500 undergraduates yearly. They are taught and supervised by 11 academic staff employed in the department. One of the social work teachers bitterly complained: "The government is buying itself social peace while offering 500 social work students' placements at the university!" (Zaviršek, personal notes, March 2013, Sarajevo).

Social work programs across the region lack critical theorizing, historical perspectives, and an engaged community-based and advocacy-oriented practice, which would give students an understanding about differences, diversity, equality, and rights. The area of antioppressive practice is one of the most underdeveloped areas in social work across Eastern Europe. Even when social work students are taught to be critical about governmental social policy regimes and advocate for service users, when they become employed, most of their critical knowledge gets forgotten. This makes a huge gap between social work education and practice. Critical social workers often seek employment in NGOs (local or international), to avoid administrative social work in state welfare services. Lessening the divide between critical social work education and a grim social work practice as well as a thoroughly evaluated practice by the state are future challenges for social workers in the region. Nevertheless,

some social work teachers have worked hard to establish social work training as a social science discipline.

A parallel process alongside the academization of social work is practicing social work as a charity activity and without academic education following the revival of the religious social services. Church power has been reestablished economically after countries returned some (or as in Slovenia, even all) of its confiscated property from precommunist times. With increased neoliberalization, the state's abandonment of social assistance money, housing, and other forms of support for people in need, the church became involved in social welfare work by 'serving the poor.' At the same time, religious authorities became front-runners in morally driven debates against same-sex partnerships, condemning the equal rights of children from rainbow families and promoting increased births among ethnic majority women while enjoying the silent support of different governments. While professional social work has hardly established its moral authority as a human rights profession based on universal human rights documents, religious organizations have built their authority according to an ahistorical natural moral law and its ideological impact is growing.

In conclusion, these developments demand a strong critical social work presence in the region. Social workers who respect universal human rights for service users and support the individual needs of diverse populations to improve the lives of communities can help bring Eastern Europe from its grim history between the 'iron curtain' and neoliberalism toward a more democratic region where individual needs and rights are respected.

See also: Armed Conflict and Social Work: Significance and Implications; Marxist Approaches to Social Work; Neoliberalism; Roma People: Discrimination and Social Work Practice; Social Work Education; Social Work in Russia; Social Work with Racialized Groups: Frameworks for Practice.

Bibliography

Arandarenko, M., Golicin, P., 2007. Serbia. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 167–186.

Bessudnov, A., McKee, M., Stuckler, D., 2011. Inequalities in male mortality by occupational class, perceived status and education in Russia 1994–2006. European Journal of Public Health 22, 332–337.

Bilandžić, D., 1985. The History of the Socialist Fedarative Republic of Yugoslavia. Major Processes 1918–1985. Školska knjiga, Zagreb.

Cocozzelli, F., 2007. Kosovo. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 203–220.

Deacon, B., Lendvai, N., Stubbs, P., 2007. Conclusions. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 221–242.

Ehrenreich, B., Hochschield, A.R. (Eds.), 2003. Global Woman. Nannies, Maids and Sex Workers in the New Economy. Granta Books, London.

Gerovska Mitev, M., 2007. Macedonia. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 130–148.

Hering, S., Waaldijk, B. (Eds.), 2003. History of Social Work in Europe (1900–1960). Leske + Budrich, Opladen.

Hering, S., Waaldijk, B., 2006. Helfer der Armen – Hueter der Oeffentlichkeit. Die Wohlfahrtsgeschichte Osteuropas 1900–1960. Barbara Budrich, Opladen, Farmington Hills.

Iarskaia-Smirnova, E., 2013. Social work in post-socialist countries: divergence and common ground. In: Homfeldt, H.G., Bähr, C., Schröder, C., Schröer, W., Schweppe, C. (Eds.), Weltatlas Soziale Arbeit. BELTZ Juventa, Weinheim, München, pp. 353–378.

Inglot, T., 2009. Czech republic, Hungary, Poland and Slovakia: adaptation and reform of the post-communist 'emergency welfare states'. In: Cerami, A., Vanhuysse, P. (Eds.), Post-communist Welfare Pathways. Theorizing Social Policy Transformations in Central and Eastern Europe. Palgrave, Hampshire, New York.

Jäppinen, M., Kulmala, M., Saarinen, A. (Eds.), 2011. Gazing at Welfare, Gender and Agency in Post-socialist Countries. Cambridge Scholars Publishing, Newcastle upon Tyne.

Maglajlić Holiček, R., Rašidagić, E.K., 2007. Bosnia and Herzegovina. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 149–166.

Malinauskas, G., 2011. This Child Is Also Mine. A Narrative Approach to the Phenomenon of Atypical Custodial Grandparenthood. LUP, Lapland University Press, Rovaniemi.

Prochazkova, L., Schmid, T., 2009. Homecare aid: a challenge for social policy research. In: Ramon, S., Zaviršek, D. (Eds.), Critical Edge Issues in Social Work and Social Policy. FSD, Ljubljana, pp. 139–164.

Rasell, M., Iarskaia-Smirnova, E. (Eds.), 2013. Disability in Eastern Europe and the Former Soviet Union. Routledge, Abingdon.

Sasse, G., 2009. Tracing the construction and effect of EU conditionality. In: Rechel, B. (Ed.), Minority Rights in Central and Eastern Europe. Routledge, London, New York.

Schilde, K., Schulte, D. (Eds.), 2005. Need and Care. Glimpses into the Beginnings of Eastern Europe's Professional Welfare. Barbara Budrich, Opladen.

Sotiropoulos, D.A., Pop, L., 2007. Bulgaria and Romania. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 62–84.

Tolstokorova, A., 2010. Where have all the mothers gone? The gendered effect of labour migration and transnationalism on the Institution of Parentohho in Ukraine. Anthropology of the East Europe Review 28, 184–214.

Vanhuysse, P., 2009. Power, order and the politics of social policy in Central and Eastern Europe. In: Cerami, A., Vanhuysse, P. (Eds.), Post-communist Welfare Pathways. Theorizing Social Policy Transformations in Central and Eastern Europe. Palgrave, Hampshire, New York, pp. 53–72.

Vermeersch, P., Ram, M.H., 2009. The Roma. In: Rechel, B. (Ed.), Minority Rights in Central and Eastern Europe. Routledge, London, New York, pp. 61–75.

Ymeraj, A., 2007. Albania. In: Deacon, B., Stubbs, P. (Eds.), Social Policy and International Interventions in South East Europe. Edward Elgar, Cheltenham, Massachusetts, pp. 187–202.

Zaviršek, D., 2005a. "You will teach them some, socialism will do the rest.!" The history of social work education 1945–1960. In: Kurt, Schilde, Schulte, Dagmar (Eds.), Need and Care. Glimpses into the Beginnings of Eastern Europe's Professional Welfare. Barbara Budrich, Opladen, pp. 237–274.

Zaviršek, D., 2005b. Between unease and enthusiasm: the development of social work education in Yugoslavia. In: Hessle, S., Zaviršek, D. (Eds.), Sustainable Development in Social Work – The Case of a Regional Network in the Balkans. University of Stockholm, Stockholm, pp. 26–34.

Zaviršek, D., 2008. Engendering social work education under state socialism in Yugoslavia. British Journal of Social Work 38, 734–750.

Zaviršek, D., 2009a. Can development of the doctoral studies in social work resist the neo-liberalism within academia? Some comparisons. In: Ramon, S., Zaviršek, D. (Eds.), Critical Edge Issues in Social Work and Social Policy. Comparative Research Perspective. FSD, Ljubljana, pp. 219–237.

Zaviršek, D., 2009b. Between blood and care: social parenthood as the enlargement of the concept of parenthood in the current societies. Socialno Delo 48, 3–16.

Zaviršek, D., 2012. Women and social work in Central and Eastern Europe. In: Regulska, J., Smith, B.G. (Eds.), Women and Gender in Postwar Europe: From Cold War to European Union. Routledge, London, New York, pp. 52–70.

Zorn, J., 2009. A case for Slovene nationalism: initial citizenship rules and the erasure. Nations and Nationalism 15, 280–298.

Social Work in Organizations

Mark Lymbery, School of Sociology and Social Policy, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Much literature on social work focuses on practice with service users; relatively little attention is given to the need for practitioners to be able to function well in the organizational context. This article will explore the nature of organizations within which social workers are employed, taking the view that many problems experienced by practitioners can be traced back to failures of understanding their place in the world of employment. While social workers are accountable both to the profession and to the service user, the accountability that they have to employing organizations also must be accepted. Using examples drawn predominantly from the British context, the article will first focus on the changes that have occurred within organizations. From this understanding it will then discuss the various ways in which social workers might experience organizational difficulties, suggesting ways in which such problems can be managed. It will conclude by suggesting that effective social work is impossible without clarity and understanding about how the organizational context can best be managed.

Introduction

Most social work discourse focuses on the interaction between the practitioner and the service user, hence, relatively little attention has been given to the organizational context within which the social worker practices. Social workers are rarely autonomous; they work within organizations that impact strongly on each individual's work. In light of this, one cannot examine their practice without considering the context within which it is carried out (Howe, 1991). Indeed, it is the contention of this article that many social workers struggle to come to terms with the demands of the organization after qualifying (Bissell, 2012). In large measure, this can be attributed to a lack of understanding of how such bodies are structured and function, and what constraints this places on individual autonomy. Consequently, it is important for practitioners to reflect critically on the organizations within which they practice. After all, it is pointless to be an excellent practical social worker if the structures within which she or he operates is barely understood.

To address this lack of understanding, the author first charts the development of social work, as this is a central element of any debate about how the organization affects practice. From this, the article will explore the way that discussions about social work often fail to encompass the impact of its organizational location, preferring to present it as a freestanding activity without reference to either its purpose or its setting. This point will be illustrated with reference to the impact of specific changes that have been visited upon social workers in the British context. It will then argue that an understanding of these changes may help practitioners to manage the organizational context of their work more effectively. While the examples in this article are derived from Britain, they are intended to help make an argument that has resonance more widely: that the organizational arrangements that govern social work materially affect the work that practitioners are able to carry out.

The Development of Organizational Structures for Social Work

In Britain, as many histories make clear (e.g., Jordan, 1984), the origins of social work were tied into the sense of charity, and the obligations of the middle class toward people who were less economically fortunate (Harris, 2008). While this was expressed in different ways – the Charity Organization Society (COS) and the Settlement Movement espousing markedly different approaches to the 'problems' of late Victorian society (Harris, 2008) – it was animated by common core principles, Christianity and humanism being the most important. In addition, as the appellation of the COS makes clear, this period also represented a significant organizational formation for social work. The early growth of social work in Britain was later buttressed by university-level education, as both the COS and the Settlement Movement established close working relationships with universities (Jordan, 1984).

In this context, it is particularly significant that the founders of both the COS and the Settlement Movement shared a sense that social work had a 'mission' to fulfill within society – a term that connects to the religious underpinnings of the profession. Despite the differences between the two approaches to the profession, they shared a belief that society could be shaped through the activities of social workers. A core purpose of social work education was therefore to shape the characters and attitudes of those people charged with this responsibility (Dominelli, 1997). As social work progressed into the twentieth century, the sense of mission faded, marking a time when social work became less of a project focused on social change and more a function of society and the state (Payne, 2006). This process was intensified by the formation of statutory children's and welfare departments following legislation in the 1940s: the creation of unified local authority Social Services Departments following the Local Authority Social Services Act 1970 represented the apogee of the growth of British social work as a state-focused profession. Indeed, the substantial

increases in budgets during the early 1970s led directly to a growth in demand for qualified practitioners: consequently, this was a period of optimism for social work.

However, this only lasted for a few years, due to the desperate economic circumstances that consumed the British government during the mid-1970s. The worldwide recession created a particular economic climate for all publicly funded services; this clearly has similarities to the position of social work in the twenty-first century (Dominelli, 2010). If the 1970s is examined first, the decline of British industrial performance led to sharply increased unemployment. In turn, this brought about a substantial growth in public expenditure, balanced by markedly higher taxation. As a consequence of the problem that this created for the public finances, the Labor government was forced to seek a loan from the International Monetary Fund in 1976. In another echo of the disruption caused across the globe in recent years, the terms of the loan required the government to pursue deflationary policies and cut public expenditure (*see* Neoliberalism).

At this point, the modern organizational formation for social work in Britain can be said to shift in its balance and structure: subsequent implications have been both deep and long-lasting. There have been clear manifestations of this. For example, there has been an insertion of market principles into social services work, which has occurred in more countries than just Britain (Dominelli, 2010). This is particularly evident in adult social care following the implementation of Britain's National Health Service and Community Care Act 1990, which simultaneously combined market principles and a consumerist ideology alongside a valorization of the notion of competition (Rose, 1999). The influence of the Audit Commission on the delivery of services in Britain was substantial, emphasizing the disciplines of 'economy, efficiency, and effectiveness,' and securing the consent of local authorities to these particular determinants of their service (Kelly, 2003). This led to the breakdown of the old-established 'welfare bureaucracies' and the development of hybrid organizational forms in their wake (Billis, 2010). At the same time, budgets failed to keep pace with the increasing scale of needs, a problem that was exacerbated by the changing demographics of the British population.

Following the 2010 general election, there was another significant moment in the structure of politics in Britain. When no single party secured sufficient parliamentary seats to form a majority, a coalition government was formed between the Conservative and Liberal parties, with the Conservatives numerically dominant. In its response to the financial crisis that consumed the world economy at the time, the coalition announced a plan to slash the deficit by imposing substantial cuts to public spending, accompanied by some limited increases in taxation. The scale of these cuts was enormous and – in the eyes of some at least – unachievable (Taylor-Gooby, 2012). In addition, they have had a disproportionate impact upon local government, which is the site of much social care provision and the primary employer of social workers.

As can be seen in this brief survey of the organizational changes to have affected social work in Britain over the past few decades, numerous ideological and practical shifts have been observed. In a future section, the author will chart their significant impact upon social work. Before entering into these debates, however, the author will explore the failure to consider the influence of employing organizations in considerations of social work, which tends to promote the decontextualization of social work practice.

The Nature of Social Work

A convenient starting point for a consideration of this rests in the international definition:

> The social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilizing theories of human behaviour and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work.
>
> IFSW/IASSW, 2000

The core problem with any international definition is that it has to apply globally, irrespective of any differences between countries. Partly as a consequence, the scope of the definition is broad. Significantly, in relation to this argument, there is no mention of the organizational context of social work because this is left to local practitioners to define (Sewpaul and Jones, 2004). However, the final sentence contains an element that many social workers employed within statutory agencies in Britain and likely elsewhere would find problematic: the idea that their work is actually (apart from notionally) governed by principles of social justice. In addition, such practitioners might struggle to see their work as being about 'liberation.'

Debates about the nature of social work have tended to focus on its essential purpose. This has been apparent at a number of different stages in social work's development. A particularly clear account of this could be found in the 1950s, by comparing the positions held by Eileen Younghusband and Barbara Wootton (Harris, 2008; Wilson et al., 2011). Largely as a result of what Harris (2008) has termed as the 'psychiatric deluge,' social work theory – if not actual practice – was dominated by psychological approaches: it was this position that Younghusband represented. She held that social workers required deep knowledge of individuals in order to provide them with support, and that it was appropriate for them to look beneath the surface of the 'presenting problem.' This guidance was framed in an ambitious and arguably unrealistic fashion (Wilson et al., 2011), appearing to downplay the relatively small practical things that social workers could achieve, while also underestimating the impact of structurally caused social disadvantage (Dominelli, 1997).

This position was challenged by Wootton (1959), who advanced a more modest role for social workers, suggesting that they should rediscover the core focus of the early pioneers, the relief of poverty. From this perspective, she argued that practitioners could become as essential to the workings of society "as is lubrication to the running of an engine" (Wootton, 1959: p. 297). She also observed that much statutory social work practice was more in accordance with her ideas than those of Younghusband, and hence, that they fitted the organizational priorities of the time. Even in agencies that held more to the values and principles espoused by Younghusband, there later emerged a range of concerns expressed by service

users that seemed to confirm the arguments of Wootton: service users wanted their immediate problems to be resolved before deeper exploration ensued (Mayer and Timms, 1970). However, there remain limitations in her analysis. For example, unless *all* the problems encountered by individuals, families, and communities are held to be the product of poverty, understandings drawn from numerous disciplines including psychology and psychiatry would help people manage their lives (Wilson et al., 2011). There are some problems, e.g., mental health and child abuse, to which the depth of analysis espoused by Younghusband would be better suited.

However, the more modest interpretation of Wootton meant that the organizational location of practitioners became particularly significant. If social workers were to become an essential element in the structure of society, they needed to be operating in some way that linked them directly to the state. This occurred immediately after the Second World War when increasing numbers of British social workers were employed within local government and what they did emphasized the small-scale practical elements of practice highlighted by Wootton. The notion of 'state social work' (Harris, 2008) derives from this change, as the location of increasing numbers of social workers within statutory organizations affected the content of practice as well as its location.

The development of radical alternatives to social work sharpened the debate substantially. Numerous writers challenged the status quo, excoriating social work for its adherence to individual psychology and its parallel sequestration of structural issues (Bailey and Brake, 1975). They sought to develop a theory of social work based on a desire to change the social circumstances within which social workers and service users interacted, an explicitly 'transformational' (Payne, 2006) approach to practice. The core aim of this movement – as outlined in the Case Con Manifesto – highlights these broader goals: "The problems faced by social workers derive from the capitalist structure of society and the place of welfare within it: effective practice must stem from an understanding of how the welfare state has developed, and the pressures to which it is subject" (Case Con Manifesto, 1975: p. 144). Consequently, the radical social work movement saw its primary goal as changing the essential nature of the society within which social work was practiced, and social work had a role in its realization.

However, the movement struggled to overcome a core contradiction: it effectively pointed out the failings of the system, but could not define precisely what radical social workers should do in response to them. In particular, the organizational location of the majority of practitioners within the state created an overwhelming problem. While the radical social work agenda appeared to assume that practitioners had an unlimited capacity to challenge and change policy, this was highly unrealistic: employers were unlikely to encourage their staff to practice in overtly revolutionary ways. Due to the continued location of most British social workers within the state, this remains a core problem, and emphasizes the point that any consideration of the nature of social work has to consider the organizational context through which it is mediated. Radical social workers were aware of this dilemma and wrote *In and Against the State*, setting out the issues and potential responses (LEWERG, 1980), but did not materially change the situation.

This contradictory point at the heart of the profession is not changed by the more recent debates about the nature of social work, of which the exchange between Ferguson and Garrett is a particularly good example (Ferguson, 2001; Garrett, 2003). Here, the distinction is held to be between social work as an exercise in 'life politics' (Ferguson) or 'emancipatory politics' (Garrett). While its scope is ambitious, the idea of 'life politics' focuses attention closely on the individual and the social work role to help maximize her or his life chances. By contrast, the ideal of 'emancipatory politics' is to change the nature of society, and for individuals' lives to be improved through this collective action. Although the terms of reference of this debate have developed substantially, at base there remains a question about whether social work should be relatively modest in its ambitions, focused on improving the life chances of the individual, or whether or not it should have a broader vision and aim to do both (Dominelli, 2010).

As a result of this ongoing tension, Jordan's important question has a particular purchase (1984: p. 13):

> Is becoming a social worker primarily to be understood in terms of the 'helping', 'caring' or therapeutic content of the job, or according to the official, bureaucratic, legal and even potentially coercive powers and responsibilities it entails?

In relation to previous arguments, therefore, the nature of social work derives from a sense of what its purposes might be. Simply put, all the elements that Jordan cites are central to what a social worker might be required to do within the British context, and the organizational location of the practitioner is an important determinant of this. In Britain, many social workers operate in the field of child protection – performing a range of legally derived actions on behalf of organizations. While this should take place in the context of a caring approach, the organizationally defined requirements are clear, dictating the precise actions to be performed. Consequently, this substantiates Howe's point (Howe, 1991): the content and nature of social work needs to be understood by reference to the bodies within which practice occurs. As a result, any changes to organizational structures and priorities have a direct bearing on what practitioners have to do.

Organizational Change: The Impact upon Practitioners

It is important briefly to consider the major political and social changes to have affected the work of social workers in recent years. As before, the author makes specific reference to organizations within the British context, yet the arguments have an inevitably international focus, particularly where they address global phenomena. In this respect, the crisis of the mid-1970s exemplified this: a problem with the worldwide supply of oil and its pricing promoted a global recession. As often happens in such circumstances, public financing is directly affected. The contemporary response to the 2008 recession typifies this response: cuts in publicly funded services. However, striking differences with the 1970s were high levels of political turmoil; unemployment rose sharply and became a major factor in the

1979 election. Largely as a consequence of this, the Conservative Party was elected in 1979, and instituted a set of distinctive political prescriptions for the problems of society. These drew upon a combination of neoliberal economic ideas and neoconservative views about the role of the state and state institutions such as social work (Lymbery, 2004).

These changes had a direct effect on social welfare in numerous ways, particularly in reaction to neoliberal economics. Public welfare was held to be a burden on economic vitality and hence should be reduced in size; as a consequence, budgets failed to keep pace with needs – which were mushrooming as a result both of demographic change and the increase of social deprivation (Ferguson and Woodward, 2009). The supposed 'disciplines' of the market were applied wholesale to areas of welfare provision. For example, the principle of competition underpinned the implementation of community care in the 1990s and heralded the rapid growth of private (and, to a smaller extent, voluntary) sector organizations; this hastened the breakdown of the large bureaucracies that had characterized the growth of the welfare state (Dominelli, 2010). In addition, as has been previously noted, the establishment of the Audit Commission had a major effect on how local government (within which social work organizations were largely based) carried out its duties (Kelly, 2003).

When a Labor government was elected in 1997 it did not materially change the direction of policy. In order to impose a sense of order on these changes, they were collapsed into the single overarching concept of 'modernization,' even though there were often few linking factors to connect them (Harris and White, 2009). While there was a lack of clarity about what the changes meant for social work, there was also substantial continuity with past policies. In particular, many social care services continued to be opened to competition; consequently, there has been a substantial residualization of the social work role (Ferguson and Woodward, 2009). The breakdown of 'welfare bureaucracies' in England (although not in the other British countries) was hastened by the organizational separation of adult social care and children's services following the recommendations of the inquiry following the death of Victoria Climbié (Laming, 2003).

Society has also experienced many changes since 1979. There have been rapidly increasing levels of social inequality, with a consequent growth in the numbers of people experiencing inordinately high levels of poverty. This is significant for social work as it has long been acknowledged that "the most striking characteristic that clients of social services have in common are poverty and deprivation" (Schorr, 1992: p. 8). Therefore, any growth in the numbers of people experiencing poverty has increased the pressure on social services organizations. At the same time, population structures have been changing; a combination of stable birth rates and increased life expectancy will ensure a substantial number of older dependents by the middle of this century, leading to increased demands on social services, while budgets are already struggling to keep pace with need. Additionally, there has been an increased significance of social movements, acting on behalf of people in marginalized groups – e.g., black people, women, disabled people, gays and lesbians, that share a sense of social work being irrelevant to their needs (Lymbery, 2004).

While responding to needs with reduced resources has represented a difficult balance for social work to maintain, it was not effectively challenged throughout the early years of the twenty-first century, boosted by a buoyant economic performance. Where challenges have been presented (Ferguson and Woodward, 2009; Dominelli, 2010) they have foundered on a number of factors: the lack of a strong professional voice and the absence of political champions being the most important. However, the problems did not disappear, and were sharply magnified in 2008, as the risky behavior of various banks caused the entire Western economy to slide into recession. The British government responded in specific ways to this. First, enormous sums of money were found to 'bail out' the banking system, on the basis that the economic structure of Western world needed to be protected (Taylor-Gooby, 2012). At a stroke, this increased the level of the country's debt and the so-called 'structural deficit,' now presumed to have created the crisis (Fontana and Sawyer, 2011). Second, when the Conservative-led Coalition took office in 2010, it established a stringent austerity program to 'resolve' these economic problems (Taylor-Gooby, 2012). A major feature of its approach to the problem was an attempt to convince the country that the primary justification for the need for austerity was the profligate nature of the previous administration rather than the risky behavior of the banks (Clarke and Newman, 2012). In a direct connection with the past, it also perpetuated a sense that the poor were substantially to blame for the problems that transpired – a logic that was common in the nineteenth century (Harris, 2008). It also appealed to the prudent householder on the basis that such levels of debt were apparently outrageous – in effect applying microeconomic logic to a macroeconomic problem. In addition, the way of resolving the problem was by massive (and never before achieved) cuts in welfare expenditure (Taylor-Gooby, 2012); in accordance with the original austerity plan, it was proposed that £110.3 billion would be found to manage the problem of public debt within one parliament, of which 73% would be cuts to public expenditure and 27% to taxation (Taylor-Gooby, 2012).

Because the nature of social work is primarily to serve people who are both disadvantaged and poor, the implementation of this austerity package is likely to affect the poorest most deeply, despite the rhetorical commitment to fairness voiced by the coalition. Certainly, the proposed 27% cut in local government expenditure between 2010 and 2015 will have a particular impact on social care and social work. It is consequently important to develop a sense of how practitioners can be enabled to manage the organizational changes that are in train: this is the focus of the following section.

Managing the Organization

Given the problems identified in the previous section, social workers have two tasks to fulfill. The first problem is pragmatic in nature: how to survive and prosper in such difficult environments. The second is much more fundamental and perhaps impossible for a social worker to accomplish in isolation: how to change the nature of employing organizations such that they become more attuned to wider social upheavals and more

supportive as contexts within which to practice. This section will primarily engage with issues that connect to the first problem. Issues that derive from the second will be addressed in the conclusion.

There are several key consequences for the practice of social work that have to be acknowledged and worked with effectively. For example, the pressure on resources will lead a practitioner into difficult terrain as she or he will inevitably get into areas of conflict. To use Thomas's own terminology (1992) austerity measures are affecting the 'structural' context within which the 'processes' of conflict management are played out. Along with a colleague, Thomas developed a two-dimensional taxonomy of conflict handling modes, the Thomas–Kilmann grid (Thomas, 1992). In this taxonomy, the five modes of conflict management (competing, collaborating, compromising, avoiding, and accommodating) are classified in relation to two underlying dimensions, assertiveness and cooperativeness. As Sportsman and Hamilton (2007) have indicated, the 'behavioral predispositions' of most human services workers – except medics – tend to ensure that they focus on unassertive and often uncooperative approaches to conflict (avoidance). This is deeply problematic when the outcomes of conflict are so critical for the service user – social workers, therefore, need to have a much more rounded set of approaches to its management. The need for this is magnified when a worker has to balance the difficulty of reconciling the 'partisan' nature of optimizing the outcome for an individual with the ideal of affecting a 'systemic' change for the wider organization (Thomas, 1992). In many cases, these will be in opposition. For a practitioner, the imperative to manage conflict effectively on behalf of the individual or group may bring her or him into direct conflict with the organization, thus putting a premium on the capacity of the worker to operate effectively in conflict.

Another potential consequence of the austerity program lies in the likely increase in the levels of stress experienced by social workers. In social work, these are elevated in normal circumstances (Lloyd et al., 2002); if left to develop, they can contribute to negative outcomes – high levels of staff turnover, dissatisfaction with the job and the nature of practice, and potential burnout (Cherniss, 1995). There appears little doubt that changed economic circumstances have increased the problems of many service users: since it is social workers who operate closely with such people, it is likely that the gap between their espoused ideals and actual practice will widen, increasing the levels of stress that they might experience (Lloyd et al., 2002). The potential negative effects of stress can lead to profoundly damaging experiences, which can include problems in concentrating, depression, and increased levels of sickness – both real or imagined (Cherniss, 1995). The corrosive sense of futility that can develop can also contribute to high levels of emotional distress, leading to a downward spiral within which an individual's performance deteriorates and her or his sense of job satisfaction plummets.

That the job may be experienced as increasingly difficult does not mean that every practitioner will succumb to stress and burnout (Collins, 2008). There is the potential for workers to build appropriate coping strategies to create resilience against the pressures of the role. It is simultaneously important for any individual practitioner to feel that she or he is making a difference to people's lives, while relishing the manifold opportunities for change and variety that present themselves. At the same time, managers need to recognize that there are features of work that help to build a sense of strength and resilience, notably the fostering of appropriate levels of autonomy (Cherniss, 1995). There is clearly a responsibility for managers within organizations to think carefully about the factors that can enhance job satisfaction and reduce stress (Collins, 2008), however difficult that may be in the current climate.

Given these additional pressures Morrison (2007) has suggested that it is particularly important for social workers to be able to demonstrate active levels of 'emotional intelligence,' to be aware of the impact of external events, and to be proactive in managing them. The way in which organizations are structured clearly indicates that they require considerable flexibility on the part of practitioners to manage the tensions and emotions that are produced. Consequently, workers must be able to function effectively within their confines, despite the additional pressures that derive from straitened economic circumstances. This implies a high level both of self-knowledge and awareness of the needs of others, plus a sophisticated ability to manage the context within which the interaction takes place (Morrison, 2007).

Cherniss (1995) has indicated how important this is in relation to avoiding burnout. In his view, professionals – who are particularly prone to burnout – must move beyond the demonstration of 'technical proficiency' and develop both organizational and people skills that will enable them to negotiate through work settings successfully. He termed the aggregation of such abilities as 'organizational negotiation skills,' which include the capacity to avoid or quickly resolve interpersonal conflicts and 'hassles,' which can result in stress and burnout, alongside the ability of not to become distracted from the task at hand, and one's own work goals (Bissell, 2012). In relation to social work in conditions of austerity, the need for this – and its potential benefit – is obvious.

Conclusion

However, enabling practitioners to function more effectively within organizations is only part of the solution. The other rests in finding a way to make them a more sympathetic location for social work. If this is not accomplished, the consequences are potentially dire, but it remains a difficult task to achieve. The first point to assert is that any individual's attempt to do this from inside an organization is likely to end in frustration and failure – it is an activity that calls out for a more sophisticated united response (Ferguson and Woodward, 2009). The need for social workers to rediscover and build upon their collective identities is therefore strong.

The key element in this is the active gathering of support for a particular action or position. If, for example, a worker is concerned about the implications of a particular course of action, it will be more effective to question this in concert with others rather than expressing an individual view that can be readily set aside. Raising an issue collectively avoids the risk that the concern can be taken as one person's gripe and hence easily be marginalized. It is also important to recognize and

understand the nature of decision making within an organization, in order to understand where and how to intervene. If problems are being caused by inadequate or inappropriate decisions taken elsewhere in an organization, it is clearly appropriate to press for change within that context. However, if the problems are caused by decisions over which the organization has no control, effective forms of protest need to be directed toward the location of the specific problem.

The impact of the austerity program is a case in point; it is potentially unhelpful and unproductive to protest to an organization about the overall cutting of budgets where the body has no choice other than to make the cuts. Any protest would need to be constructed more broadly and aimed directly to secure a change in policy. However, decisions that relate to how the cuts are to be implemented are clearly appropriate for protest. If, for example, they result in a disproportionate impact on specific user groups it is right and proper – and arguably a moral obligation – for a social worker to protest against them; in such circumstances, it is also important to consider ways in which such objections can be most effective.

It is apparent, therefore, that consideration of the organizational context for social work practice is far from being an optional extra for practitioners. Rather, coming to grips with the organizational context is vital if they are to develop forms of reflexive and ethical practice (Hughes and Wearing, 2013). Being able to understand and creatively manage the organization is therefore an essential skill for a social worker (Bissell, 2012), without which the effectiveness of her or his practice will inevitably be diminished.

See also: Antisocial Behavior; Communication Skills in Social Work; Criminal Justice and Social Work Practice; Critical Reflexivity and Social Work Practice; Critical Social Work Practice; Ethics and Values; Gendered Care and Social Work Practice; Marxist Approaches to Social Work; Masculinity and Social Work Practice; Migration, Ethnicity, Aging and Social Work Practice; Neoliberalism; New Managerialism and New Public Sector Management; Professional Development; Radical Social Work; Roma People: Discrimination and Social Work Practice; Social Protest Movements and Social Work Practice; Social Work Theory.

Bibliography

Bailey, R., Brake, M. (Eds.), 1975. Radical Social Work. Edward Arnold, London.

Billis, D. (Ed.), 2010. Hybrid Organizations and the Third Sector: Challenges for Practice, Theory and Policy. Palgrave Macmillan, Basingstoke.

Bissell, G., 2012. Organisational Behaviour for Social Work. Policy Press, Bristol.

Case Con, 1975. Case Con manifesto. In: Bailey, R., Brake, M. (Eds.), Radical Social Work. Edward Arnold, London.

Cherniss, C., 1995. Beyond Burnout. Routledge, New York.

Clarke, J., Newman, J., 2012. The alchemy of austerity. Critical Social Policy 32 (3), 299–319.

Collins, S., 2008. Statutory social workers, stress, job satisfaction, coping, social support and individual differences. British Journal of Social Work 38 (6), 1173–1193.

Dominelli, L., 1997. Sociology for Social Work. Macmillan, London.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Ferguson, H., 2001. Social work, individualization and life politics. British Journal of Social Work 31 (1), 41–56.

Ferguson, I., Woodward, R., 2009. Radical Social Work in Practice: Making a Difference. Policy Press, Bristol.

Fontana, G., Sawyer, M., 2011. Financial austerity: lessons from recent events in the British Isles. Challenge 54 (2), 42–60.

Garrett, P.M., 2003. The trouble with Harry: why the "new agenda of life politics" fails to convince. British Journal of Social Work 33 (3), 381–397.

Harris, J., 2008. State social work: constructing the present from moments in the past. British Journal of Social Work 38 (4), 662–679.

Harris, J., White, V. (Eds.), 2009. Modernising Social Work: Critical Considerations. Policy Press, Bristol.

Howe, D., 1991. Knowledge, power and the shape of social work practice. In: Davies, M. (Ed.), The Sociology of Social Work. Routledge, London.

Hughes, M., Wearing, M., 2013. Management and Organisation in Social Work, second ed. Sage, London.

International Federation of Social Workers/International Association of Schools of Social Work (IFSW/IASSW), 2000. The Definition of Social Work. IFSW, Berne. http://www.ifsw.org/en/p38000208.html (accessed 29.10.12.).

Jordan, B., 1984. Invitation to Social Work. Martin Robertson, Oxford.

Kelly, J., 2003. The audit commission: guiding, steering and regulating local government. Public Administration 81 (3), 459–476.

LEWRG (London Edinburgh Weekend Return Group), 1980. In and against the State. Pluto Press, London.

Laming, Lord, 2003. The Victoria Climbie Inquiry. HMSO, Norwich.

Lloyd, C., King, R., Chenoweth, L., 2002. Social work stress and burnout: a review. Journal of Mental Health 11 (3), 255–265.

Lymbery, M., 2004. Responding to crisis: the changing nature of welfare organisations. In: Lymbery, M., Butler, S. (Eds.), Social Work Ideals and Practice Realities. Palgrave, Basingstoke, pp. 34–56.

Mayer, J.E., Timms, N., 1970. The Client Speaks: Working Class Impressions of Casework. Routledge & Kegan Paul, London.

Morrison, T., 2007. Emotional intelligence, emotion and social work: context, characteristics. Complications and contribution. British Journal of Social Work 37 (2), 245–263.

Payne, M., 2006. What is Professional Social Work? second ed. Policy Press, Bristol.

Rose, N., 1999. Powers of Freedom: Reframing Political Thought. Cambridge University Press, Cambridge.

Schorr, A.L., 1992. The Personal Social Services: An Outside View. Joseph Rowntree Foundation, York.

Sewpaul, V., Jones, D.N., 2004. Global standards for social work education and training. Social Work Education 23 (5), 493–513.

Sportsman, S., Hamilton, P., 2007. Conflict management styles in the health professions. Journal of Professional Nursing 23 (3), 157–166.

Taylor-Gooby, P., 2012. Root and branch restructuring to achieve major cuts: the social policy programme of the 2010 UK coalition government. Social Policy and Administration 46 (1), 61–82.

Thomas, K.W., 1992. Conflict and conflict management: reflections and update. Journal of Organizational Behavior 13 (3), 265–274.

Wilson, K., Ruch, G., Lymbery, M., Cooper, A., 2011. Social Work: An Introduction to Contemporary Practice, second ed. Pearson Education Ltd, Harlow.

Wootton, B., 1959. Social Science and Social Pathology. George Allen and Unwin, London.

Social Work in Russia

Elena Iarskaia-Smirnova, National Research University 'Higher School of Economics,' Moscow, Russia
Valentina Yarskaya, Saratov State Technical University, Saratov, Russia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines a range of the general and the unique features of the social work's development in Russia. It explores the historic context of Soviet rule and period of transition, describes the postsocialist development of social services as well as establishment of social work as an occupation in early 1990s, and concludes by tracing general trends in the development of social work.

Introduction

The article consists of five sections. Starting with outlining a brief history of social work in Russia, this article then turns to the main features of welfare developments and beginning of social work as an occupation during the transition from socialism to the democracy and market economy in 1990s and considers the contemporary development of social work as a profession and a training program in today's Russia. The article ends with outlining the core features of governance and professional status of social work. International influences and value clash are touched upon as well as the public understanding of the profession.

A Brief History of Social Work in Russia

Short historic background presents the picture of welfare policy and social work formations during socialism and under the transition. Various forms of social care and social service were developed under Tsarist Russia and during the socialism (Hering, 2009; Iarskaia-Smirnova and Romanov, 2009b; Iarskaia-Smirnova, 2011a; Waaldijk, 2011). As elsewhere, the preprofessional history of social work included religious and civic institutions and activities related with charity and social care. Orthodox, Catholic, Jewish, and Muslim traditions played important role in organization of social care and training for personnel in social services such as shelters for homeless, disabled, and impoverished people, institutions for orphans and poor children. Many such institutions were demolished or reestablished under the state socialism but on different ideological basis. The disciplines of sociology and psychology, as the most solid core for critical thinking and dissent, were eliminated from the university curricula or they were allowed only to follow the line issued in Moscow (Guzzetta, 1995: 192).

In its golden age, relating to the early Khrushchev and Brezhnev periods, the Soviet government built one of the most advanced systems of social assistance in the world, concerning access equality as well as the volume and quality of services (Schilde and Schulte, 2005). And although the right and duty for labor determined the access to many social services directly from the workplace, the segment of a universal welfare regime with typical disposition of domiciliary services available for all district residents was extended as well (Iarskaia-Smirnova and Romanov, 2009a).

As official state policy did not recognize the existence of social problems, a need for social work could not be articulated explicitly. The population not only passively enjoyed the benefits of socialist welfare, but many were actively seeking state welfare and demanded fairness, which led to repressions and elimination of protesters. The state and its various agents played a major role in carrying out the double-edged care-and-control task at all levels of social life. Because of limited resources of the state and following the socialist ideas of self-government and the socializing role of collective, voluntary children workforce was broadly used in providing help to people with disabilities, and the elderly; while issues of women and children, family issues have been dealt with by women's councils and trade union units. They performed the care-and-control functions that were in common with certain forms of professional social work service elsewhere (*see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'; Social Work in China). However, issues of disabilities, mental health, ethnic minorities, and violence against women and children were silenced (see: Waaldijk, 2011).

In the Soviet Union, social assistance and care were provided without the official presence of social work as a profession, which was introduced only in late 1980s–90s. During the 1990s, when Russia underwent the transition from socialism to market economy and democracy, crucial transformations occurred to the role of the state, the ideology of social support, and to the models and instruments of social policy and social services.

Social Welfare under Transition from Socialism

During the 1990s, many of the existing institutions and values were challenged and changed. Free education, public health care, and social benefits that had been a fact of life for decades in the Soviet Union became objects of deep nostalgia and frustration for many people, especially older people.

The economic decline, increase of levels of poverty and unemployment, and extreme lack of resources have tempered the welfare reform throughout the region. High levels of

poverty particularly affected vulnerable groups such as older people and people with disabilities (Standing, 1998).

Having embraced a different shape of economic and political development during socialism, Russia as well as other postsocialist countries (*see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'; Social Work in China) undertook big and various efforts to repair the former welfare states or build them anew (see Cook, 2007). The postsocialist transformation proceeded from 1990. The vector of social reforms has continued to change under the influence of changing national political agendas and international influences.

Russia inherited from the Soviet period a complex system of social security based on public institutions, without professional social work and with the small social transfers to a wide range social groups (people with disabilities, single mothers, veterans, altogether more than 150 categories of population), which were paid irregularly.

The government became aware of the need to develop a system of social services as a part of a modern and professional social welfare system and continued to develop, or started to establish, a new system of social services based on essentially new legislation. Social work was introduced in 1991 as a new occupation and simultaneously as a new university degree program. During the 1990s, many Russian universities established university programs in social work, and in some cases, in social pedagogy, and more recently, in youth work.

In Russia, the persons with a title 'social worker' stated on a badge, first became known to the public during *perestroika*, in late 1980s. It was a time of total shortage of food and goods, when people had been standing in huge queues in the shops in order to buy simple products. 'Social workers' had the right to buy things without waiting in a queue because they were assisting older people and people with disabilities with some domestic chores. A professional of a higher status received in 1991, a separate definition of 'specialist of social work.' Another title for a practitioner with similar qualifications, a 'social pedagogue' was introduced in early 1990s, adopted from the German and Danish models.

During the 1990s, a wide network of social services was established under the responsibility of the Ministry of Labour and Social Development (recently Ministry of Labour and Social Protection). This social services network has been expanding rapidly during the last 20 years. The workforce of social services provides care in many different settings, which according to the Social Service Federal Law (1995) includes home care, social assistance for family and children, social-rehabilitation services for children and youngsters under 18, helping children who are left without parents, shelters for children and adolescents, psychological–pedagogical help to the population, hotline services, night stay hostels, day care for lonely older people, residential social services (nursing homes for elderly and disabled, psycho-neurological nursing homes, nursing homes for children with mental retardation, nursing homes for children with disabilities), and older people's care. Residential care established in state-owned institutions and other state-based social care services are also found in Russia. Outreach work with youth delinquents, drug addicts, and homeless people is conducted mainly by NGOs including religious organization, which are active in big cities.

In the 1990s, a series of new social legislation have been adopted here. Russia as many other postsocialist countries became a member of the Council of Europe, the signatory to the Universal Declaration of Human Rights, and has ratified a number of other international human rights instruments as well as European Social Charter. New institutions had to be built to resume the place in the global international order (*see* Globalization, Welfare, and Social Work). The development of social welfare and social service legislation in Russia as in other postcommunist countries (*see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'; Social Work in China) has been modeled in close collaboration with international and supranational organizations including the World Bank, UNDP, USAID, UNESCO, EU, whose support was welcomed as an important economic and symbolic resource (USAID, 2008) (*see* International Social Work Student Exchanges).

In accordance with neoliberal ideology of reforms (*see* Neoliberalism; New Managerialism and New Public Sector Management), a universalistic approach was considered ineffective, and the emphasis in solving welfare problems was shifted in mid-2000s to a means-tested scheme. As in many other postcommunist countries, with the adaptation of a discourse of poverty, the needs of clients in Russia, too, were reduced to means-tested poverty relief (*see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'; Social Work in China). During the transition from socialism to the market, new surveillance techniques and disciplinary welfare practices were introduced here that pathologized and stigmatized the clients (Haney, 2002: 235).

Social Work in Postsocialist Russia

Under state socialism, the state provided residential care and material assistance for certain categories of population while many problems were to be solved by the families, teachers, police, and formal and informal collective networks. This approach has been challenged as narrow, but such an assumption is still widespread. Social work in the child protection area, deinstitutionalization, juvenile justice, work with disabled children and their families and also with young people are the new areas of research and training as well as practice models. The movement to protect and/or deinstitutionalize children in vulnerable situations is confronted by a weak child protection system and ideological conflict that might increase risks for children.

Assumptions on human rights as well as on role of the state, market, and civil society in social policy and social work can differ and the clash of ideologies is inevitable. One of the examples is presented by the clash between child-centered and family-centered child protection models in case of juvenile justice. Under socialism, juvenile justice did not exist and the policy on punishment has been rather sound (see McAuley and MacDonald, 2007). In 1999, juvenile justice was adapted from foreign experience under a French–Russian project intended to change the system toward preventive approaches and reducing imprisonment among young people. Within a few years, 20 Russian regions had set up juvenile courts with support of the UNDP, UNICEF, and foreign agencies and foundations from

France, Sweden, and other countries with involvement of the Supreme Court of Russian Federation (see Trygged, 2009). From 2005 to 2010, negative public opinion toward juvenile justice was accelerated by moral panics orchestrated by the mass media. In popular protest, various arguments have become mixed-up:

> Juveniles are often presented as a threat for society or as insecure victims who should be protected from careless parents, pedophiles, drugs, etc. The child protection agencies have been sometimes in public discourse named 'juvenile services' and criticized by public as the 'Western' idea of state control over the private life. These claims are formulated by the conservative media, nationalists groups, and by the Church but sometimes it is initiated or supported also by representatives of the government, discrediting juvenile justice. The critics claim that focusing on children's rights undermines an important role of family against Russian traditions. Some publics, social services, ombudsman, and NGOs support juvenile justice, but the national legislation reform on juvenile justice was blocked due to the stigmatization of the idea itself.

Public discourse that frames social work practice contains a mixture of various and often contradictory value positions concerning normative models of good citizens, men and women, families and children (*see* Feminist Research in Social Work). Old prejudices against certain groups have retained from the past (Guzzetta, 1995: 202) not only in Russia. The revival of neoconservative movements is known in many countries of the world, including 'old' and 'new' Europe and ex-Soviet republics (*see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'), where escalation of sexism, homophobia and racism, prejudices against Roma, people with disabilities, and other minorities are seen and experienced (*see* Critical Reflexivity and Social Work Practice; Roma Peoples, Discrimination and Social Work Practice; Critical Social Work Practice). Thus, the wider societal context of social work is constituted by the mixture of ideologies, where neoconservative voices sometimes are interwoven with social democratic and/or neoliberal agenda while "paternalist and authoritarian values of legislators, bureaucrats and even the public" (Freed, 1995: 40) are clashed with the agency of those service users, NGOs, and service providers who tries to criticize and challenge the networks of power (Iarskaia-Smirnova and Romanov, 2012) (*see* Critical Reflexivity and Social Work Practice; Choice, Empowerment, and Social Work).

The development of social work is framed by national and local policy contexts. One of such contextual elements is religion. In Russia, the role of religion in the lives of the people is taking on significant meaning, particularly the role of the Russian Orthodox Church and Islam in the revival of charity, and the role of the religious organization in service provision are gradually growing. It inherits historic traditions of mercy and charity, and volunteer help to the needy, which was widely spread in prerevolutionary Russian society and arranged around the churches and cloisters. During socialism, civil sector and religious organizations have been put under the strict state control. Volunteer movements started to rise again in late-Soviet times in 1980s and especially in postsocialist Russia since early 1990s. Various self-help and mutual aid groups, charity initiatives have been established on different value bases. Many religious organizations of different persuasions developed forms of aid including orphans in institutions, prisoners, homeless people and others in need. Some of them arrange seminars and courses for practitioners. Social work at the parish of Russian Orthodox Church has gradually become professionalized acquiring the common features of this profession. Since 2011, the Church administration has held annual distant learning courses on social service with the possibility of field practice facilitating the study of best practice examples in Moscow. The participants acquire theoretical foundations on casework and state social policy; learn about social work methods, social work organization and management, public relations, practical skills of working with homeless people, prisoners (see Romanov and Kononenko, 2014).

In order to mobilize resources of local communities to help vulnerable groups fully realize their social citizenship, social workers have collaborated with nongovernmental services, including women's organizations, trade unions, and human rights organizations. Some NGOs, which have grown out of service users' associations and grassroots movements, developed a strong emancipatory view in their work. The number of such organizations is limited and unstable due to the economic and political situation in Russia, where involvement of foreign donors is not encouraged while national funds to support nongovernmental activities are scarce (Iarskaia-Smirnova, 2011b).

Social Work Training

The revival of social thought in Russia, supported by national and international foundations, animated public discussion on matters of social inequality, social exclusion, and social problems and through the activities of university departments and independent research and training centers. While in several Central and Eastern European countries, the profession has acquired public recognition, and universities have been developing their programs for several or many years before the 1990s (e.g., Zaviršek, 2008; *see* Social Work in Eastern Europe: Surviving Neoliberalism after the 'Iron Curtain'), social work was a new occupation in Russia and other former Soviet republics.

The first four universities started to teach social work courses in 1991 and by 2011 the number of university programs of social work was about 175 and they covered the entire country. In addition, there are about 115 university programs in social pedagogy. Currently all universities are involved in a process of transformation toward the Bologna system (*see* Bologna Process and Social Work) switching to 4-year BA (established in 1992) and 2-year MA (established in 1996) levels, although some continue to offer traditional 5-year diploma programs of 'specialists in social work' (established in 1991). As an educational program, the Bologna Process was imported here from the West through the efforts of universities, individual scholars, and social work practitioners who became involved into various forms of collaboration, knowledge exchanges, academic mobility, and the establishment of joint degree programs (*see* International Social Work Student Exchanges; Social Work Education).

Opportunities for training and continuing education for practitioners and teachers became and still are available mainly

through international collaboration between foreign universities and practice agencies, the government, social services, and academia. Such schemes have been welcomed, and the projects provided resources for institutional and curriculum development. They also influenced teaching methods and widened opportunities for teachers to improve their qualifications through new cross-cultural communication experiences, and the acquisition of updated professional knowledge (see Iarskaia-Smirnova, 2011b) affected practice models in the projects that have involved governments and social service agencies (*see* International Social Work Student Exchanges; Social Work Education).

Due to the delayed professionalization of social work in Russia, there is a lack of professional knowledge and skills among practitioners. A number of collaborative international projects dealing with social work and welfare developments have been implemented since early 1990s, aiming to raise professional competencies in public authorities, managements, and among social workers in Russia (e.g., Gilbert, 1999; Eduards et al., 2009; Trygged, 2009) (*see* International Social Work Student Exchanges).

Much attention has been paid to adopting of new technologies, standards, and their application into work practice at a local level. Specialists have acquired new models of work; new approaches to social work have been adopted. Social service administrators and practitioners have very positive attitudes to opportunities for professional development whether it is locally based or has origins abroad. However, they have no autonomy to participate in any international or local projects without the permission of regional and/or national authority.

International collaboration induced multifaceted exchanges and influences. A number of strategic international partnerships on all levels of welfare governance and practice have been built (e.g., Trygged and Eriksson, 2009; Doel and Penn, 2007). Local academics have been actively engaged in such cooperative ventures, while governmental agencies and social service administrators have played an important role in opening the gates into the practice fields.

One of the most critical issues for the beginning of educational development in social work during the 1990s was the lack of teachers competent in the subject and lack of teaching materials. Due to the extensive and lengthy suppression of social sciences and psychology during Soviet times, social work textbooks have been inadequate in their critical appraisal of social problems, and preparing social workers for the reality they will face. The principles of nondiscriminatory or multicultural social work, empowerment, and social criticism are rarely present in the textbooks analyzed during 2005–10 (see Iarskaia-Smirnova and Romanov, 2008; and Critical Reflexivity and Social Work Practice; Choice, Empowerment, and Social Work). However, social work literature in Russian language is currently growing, a few translations are being produced, and many original works are published in journals and books by scholars and practitioners in the region.

The students are trained in research skills and given macro-level knowledge on welfare policy and legislation, as well as theoretical knowledge on psychology and theories of social work (Gilbert, 1999: 15). A practice component is still lacking in the curriculum. Formalized internship was, and partly still is, especially difficult (see Iarskaia-Smirnova and Rasell, 2014) to establish when appropriate field placements do not exist (Guzzetta, 1995: 205; see also Montague et al., 2008: 241). Although during 20 years of social work education, placements have been secured, field supervisors are still lacking in those settings where low salaries and the absence of career opportunities prevent employment of university graduates in social work agencies, while social workers with practical experience not necessarily have adequate qualifications (see Penn, 2007; and Social Work Education). The system also has employment problems in labor market for the graduates of such programs. Due to the low salaries offered to qualified social workers, young university graduates are choosing other jobs for themselves. Thus, the vast majority of social workers (or specialists in social work) in Russia still have not received formal education or professional training.

By setting up inadequate wage policies for social workers, the state has reinforced the societal assumption of this occupation as 'cheap women's labor' (Iarskaia-Smirnova and Romanov, 2008). These women "represent a *de facto* voluntarist service. Lacking a clear framework for career development coupled with the prospects of a reasonable standard of living, it is difficult to see how social work as a profession will attract" (Gilbert, 1999: 18) university graduates. The high fluctuation of personnel due to the poor wages, large caseloads, and insufficient training opportunities in many cases has hindered the professional development of social services in Russia.

Governance and Autonomy in Social Work

The top-down approach to solving social problems and regulating the profession is still in place as it was under state socialism. The state has the main jurisdiction over the new profession and provides it with financial and symbolic capital. There is special social services legislation in Russia defining social work in direct terms or as a social service provision. Here, as in some of the Eastern European countries, the concept of care or charity is a priority above the concepts of citizenship, entitlements, and civil rights in social service legislation (Krzyszkowski, 2004/2005). Social work as home assistance for older people is the most typical public understanding of the profession. Although a greater emphasis in education and legislation is placed on nonmonetary social services, "practising social workers find themselves means-testing clients for pensions, providing specialist services, and giving routine institutional and domiciliary care" (Gilbert, 1999: 18).

In the Law on the Foundations of Social Services (1995) the term 'social service workers' is used, while a 'social worker' is defined in the new draft of the Law on the Foundations of Social Services (2013, to be enacted in 2015) as a specialist with professional education relevant to the job functions, who works in any kind of organization that performs social services. The new law formulates the rights, responsibilities, and guarantees of a social worker, in particular, the rights to receive professional training and further qualification training, to form and build professional associations.

Since the beginning of the 1990s, the field of social work practice was developing rather separately from the field of

professional training, while the situation in the human resources of the social work services sector was characterized by low wages, labor shortages, high fluctuation of personnel, and insufficient opportunities of retraining. There is no regulative legislation that requires a formal education in social work for a person wishing to begin professional practice. National standards for social services specify that functions and activities of workers are gradually being adopted but their implementation is a slow process.

In the early 1990s, four professional associations were created: The Association of Social Workers, Association of Social Pedagogues and Social Workers, Association of Social Services Employees, and Association of Schools of Social Work. Also, several magazines on social work were also established. Since 2001, the Union of Social Pedagogues and Social Workers (the Union) has been acting as the national association with over 4500 members in 70 regions of the Russian Federation. It has been active in promoting the prestige of social work and building a national community of social work practitioners, administrators, and educators but was not in possession of a self-regulatory and licensing mandate. The Union publishes a periodical 'Sotsial'naia rabota' (social work). These professional associations have no power as regulatory bodies.

The most characteristic feature of the socialist legacies here is the persistence of the monopolized position of state-based public services and limited possibilities for creating a competitive environment (Romanov, 2008). In Russia, unlike in the USA or Australia, social workers are employed mostly by governmental agencies. Commercial sector is unlikely to develop here in the nearest future (Gilbert, 1999: 10) although the social services market in Russia is slowly developing, with a very few nongovernmental players, such as alternative nonstate providers. The government gradually increases budget allocations for the grants to be distributed among nongovernmental social services hoping that the third sector would grow.

In 2000s, the central government intended to make relationships between the citizens and the state more efficient and effective.

So far, welfare agencies are governed by a highly centralized bureaucracy and one can talk only about relative autonomy of social workers. However, there are some exemplary cases of social workers affecting structural changes (see Iarskaia-Smirnova and Romanov, 2013). For instance, a social worker from a northwest region of Russia has been working in a public agency and at the same time was a leader of the association of NGOs of people with disabilities of the region, taking part in the international projects on social work. In 2004, she supported a court appeal by a group of parents who sought preschool places for their children with disabilities because they had been rejected earlier by the city authority. Collaboration between the parents' movement, NGOs, and social workers with the engagement of the mass media has led to structural changes, such as establishment of personal assistants, reconstruction of facilities, and a new regional legislation, "On providing access to the public buildings and transportation for people with disabilities and other citizens with low mobility." In this case, a practitioner who promoted and was engaged with collective action, was able to enhance her professionalism through international exchange programs.

Since the early 2000s, a reform has gradually taken place by which public social services are being converted into semi-autonomous organizations. The idea is to make public services capable of operating in a quasi-market, without guaranteed financing and to compete for budgeting with other providers. This reform has been launched as an experiment in several regions in early 2000s and recently, the possibility to choose the legal frame (i.e., to choose to become semi-autonomous or continue to be based on a guaranteed budget scheme) has been opened for all social service organizations by Federal Law in 2010. It has made it possible for various types of organizations including nongovernmental services, to participate in market competition.

The competitive nature of a rising social services market is being fostered by the so-called 'contests of social projects' – a competition among various organizations for small grants arranged by the local governments. However, the procedure is characterized by a lack of transparency and lack of clear criteria for evaluation of the proposals and the outcomes. This process is limited by a lack of standards of social services, a weak knowledge base concerning the methods for working with clients, standard regulation in this field, a lack of skills in project management, and inadequate evaluation by many public and NGOs (*see* New Managerialism and New Public Sector Management). It is assumed that management of public services will gradually become more flexible, possibilities for commercial activity will grow, and the wages and motivation of workers will increase as well as opportunities for their professional development. It was expected by the authorities and now is shared by social service managers that the administrators and the staff of autonomous organizations would be more enterprising, acting in more innovatory and entrepreneurial way for the interests of service users and communities. A number of policy actors have been promoting their interests in shaping the legislative, socioeconomic, and political environment for social work, competing for legitimacy and status in policy processes. It is important for the government, national, and international foundations, as well as professional and academic community, to focus more on critical issues in social work both locally and globally (*see* Globalization, Welfare, and Social Work). It is worthwhile paying more attention to retraining programs and raising the level of skills of staff who already work for social service agencies, because the vast majority of staff in those agencies do not get basic training in the field. Education is called to contribute to the development of such reflexive professionals who would be able to recognize inequality on individual and institutional levels in direct contacts with clients or on a structural level during organizational, social, and political interaction and come up with innovative means of tackling these problems.

The importance of improving such training increases due to the intensive growth of social services and the demand for well-qualified personnel. There is a growing need for appropriate professional literature as well as for the popularization of civil society and human rights values by the mass media. An effective mechanism for the independent evaluation of social

services is also needed, to make it possible to target educational and fundraising activities.

Due to a lack of strong professional associations as well as communities of clients, it would be not easy task to define occupational standards and codes of ethics from below. Representatives of academia and the government take the lead in elaboration of the first versions of such documents. In 2013 occupational standards for social work have been elaborated and the code of ethics for the employees of social services has been drafted by the government. The professional associations took part in discussions of these documents. Gradually, social work will be developed and enhanced as a profession and this process would lead to a (re)formulation of qualification and values based on practical experience, national and international expectations, and knowledge in theory and methods of social work, human rights, and dignity. It is worthwhile to pay more attention to retraining programs and to raise the level of skills of staff already working for social service agencies who do not get basic training in the field. It is also important to conduct research of social policy reforms and social work practice in order to discuss the functions, competencies and values in professional community.

Acknowledgments

This study comprises the results from the project supported by the Russian Foundation for Humanities, grant No 14-03-00778 and Basic Research Programme of NRU HSE in 2014.

See also: Bologna Process; Critical Reflexivity and Social Work Practice; Critical Social Work Practice; Neoliberalism; New Managerialism and New Public Sector Management; Roma People: Discrimination and Social Work Practice; Social Work Education; Social Work Theory; Social Work in China; Social Work in Eastern Europe.

Bibliography

Cook, L.J., 2007. Postcommunist Welfare States: Reform Politics in Russia and Eastern Europe. Cornell University Press, Ithaca.

Doel, M., Penn, J., 2007. Technical assistance, neo-colonialism or mutual trade? The experience of an Anglo/Ukrainian/Russian social work practice learning project. European Journal of Social Work 10 (3), 367–381.

Eduards, K., Krivonos, M., Rylander, L., 2009. Lessons Learned from Development Cooperation with Russia. Swedish Support to the Reform Process 1991–2008. Main Report. Sida. http://www.sida.se/Documents/Import/pdf/Lessons-Learned-from-Development-Cooperation-with-Russia1.pdf.

Freed, A.O., 1995. Bulgarian social services and social work education. International Social Work 38, 39–51.

Gilbert, K., 1999. The Development of a System of Social Services in the Russian Federation. Working Paper Series September, Number WP 014/99. University of Wolverhampton. http://www.wlv.ac.uk/PDF/uwbs_WP014-99%20Gilbert.pdf.

Guzzetta, Ch, 1995. Central and Eastern Europe. In: Watts, Thomas D., Elliott, Doreen, Mayadas, Nazneen Sada (Eds.), International Handbook on Social Work Education. Greenwood Press, Westport, CT, pp. 191–210.

Haney, L., 2002. Inventing the Needy: Gender and the Politics of Welfare in Hungary. University of California Press, Berkeley.

Hering, S. (Ed.), 2009. Social Care under State Socialism: Ambitions, Ambiguities and Mismanagement. Verlag Barbara Budrich, Leverkusen.

Hussein, Sh, 2011. Social Work Qualifications and Regulation in European Economic Area (EEA). Final report. General Social Care Council and Skills for Care & Development. http://www.kcl.ac.uk/sspp/departments/sshm/scwru/pubs/2011/hussein2011eea.pdf.

Iarskaia-Smirnova, E., Romanov, P., 2012. Doing class in social welfare discourses: 'unfortunate families' in Russia. In: Salmenniemi, Suvi (Ed.), Rethinking Class in Russia. Ashgate, Farnham, pp. 85–105.

Iarskaia-Smirnova, E., Romanov, P., 2009b. The rhetoric and practice of modernisation: soviet social policy, 1917—1930s. In: Hauss, Gisela, Schulte, Dagmar (Eds.), Amid Social Contradictions. Towards a History of Social Work in Europe. Verlag Barbara Budrich, Opladen and Farmington Hills, pp. 149–164.

Iarskaia-Smirnova, E., 2011a. Gender and Class in Russian Welfare Policy: Soviet Legacies and Contemporary Challenges. University of Goteborg, Ineko, Goteborg.

Iarskaia-Smirnova, E., 2011b. Professional ideologies in Russian social work: challenges from inside and outside. In: Selwyn, Stanley (Ed.), Social Work Education in Countries of the East: Issues and Challenges. Nova Science Publishers, Inc., Hauppauge, New York, pp. 425–448.

Iarskaia-Smirnova, G., Rasell, M., 2014. Rasell integrating practice into Russian social work education: institutional logics and curriculum regulation. International Social Work 57 (3), 220–232.

Iarskaia-Smirnova, E., Romanov, P., 2009a. Multiplicity and discontinuity in the soviet welfare history (1940–1980). In: Hering, Sabine (Ed.), Social Care under State Socialism (1945–1989). Barbara Budrich Publishers, Opladen and Farmington Hills, pp. 213–225.

Iarskaia-Smirnova, E., Romanov, P., 2008. Gendering social work in Russia: towards anti-discriminatory practices. Equal Opportunities 27 (1), 64–76.

Iarskaia-Smirnova, E., Romanov, P., 2013. Social workers affecting social policy in Russia. In: Gal, J., Weiss-Gal, I. (Eds.), Social Workers Affecting Social Policy. An International Perspective. The Policy Press, London, pp. 101–119.

Krzyszkowski, J., 2004/2005. Social services system in Poland. IUC Journal of Social Work Theory and Practice (10). http://www.bemidjistate.edu/academics/publications/social_work_journal/issue10/articles/6_Poland.htm.

McAuley, M., MacDonald, I.K., 2007. Russia and youth crime. A comparative study of attitudes and their implications. British Journal of Criminology 47, 2–22.

Montague, J., Morgan, J., Sommerville, A., 2008. Travelling on an east-west social work journey: issues in trans-national collaborative working. In: Salustowicz, P. (Ed.), Social Policy and Social Work: From an International Development Perspective. Lit Verlag, Berlin, pp. 233–246.

Penn, J., 2007. The development of social work education in Russia since 1995. European Journal of Social Work 10 (4), 513–527.

Romanov, P., 2008. Quality evaluation in social services: challenges for new public management in Russia. In: Peters, G. (Ed.), Mixes, Matches, and Mistakes: New Public Management in Russia and the Former Soviet Republics. LGI, OSI, Budapest, pp. 9–53.

Romanov, P., Kononenko, R., 2014. Glocalization processes in Russian social work. International Social Work 57 (5).

Schilde, K., Schulte, D. (Eds.), 2005. Need and Care – Glimpses into the Beginnings of Eastern Europe's Professional Welfare. Barbara Budrich Publishers, Opladen and Bloomfield Hills.

Standing, G., 1998. Societal impoverishment: the challenge for Russian social policy. Journal of European Social Policy 8 (1), 23–42.

Trygged, S., Eriksson, B., 2009. Implementing Swedish models of social work in a russian context. Social Work and Society 7 (2). http://www.socwork.net/sws/article/view/78/337.

Trygged, S., 2009. Social work with vulnerable families and children in 11 Russian regions. European Journal of Social Work 12 (2), 201–220.

USAID, 2008. Social Work Education and the Practice Environment in Europe and Eurasia, vol. 27. The Rutgers University Center for International Social Work in Cooperation with Creative Associates International, Inc. and the Aguirre Division of JBS International, Inc. Available at: https://communities.usaidallnet.gov/st/node/6.

Waaldijk, B., 2011. Social work between oppression and emancipation. Histories of discomfort and inspiration in Europe. Social Work and Society 9 (2). http://www.socwork.net/sws/article/view/272/465.

Zaviršek, D., 2008. Engendering social work education under state socialism in Yugoslavia. British Journal of Social Work 38 (4), 734–750.

Social Work Theory

Barbra Teater, College of Staten Island, City University of New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social work theories serve as the foundation for social work practice by providing an understanding, explanation, and prediction of human behavior and social structures. Seven general categories of theories used in social work practice are reviewed and their relevance to the stages of social work practice of assessment, intervention, evaluation, and endings are discussed.

Introduction

Social work theories serve as the foundation for social work practice by providing an understanding, explanation, and prediction of human behavior, social structures, and social interactions. Theory informs each stage of the social work process from the initial stages of assessment, to selection and evaluation of interventions, to deciding when it is best to end services with clients. The profession of social work aims to promote social change and social justice by working with people and the environment in which they interact. Intervening within several different layers of society, including individuals, families, communities, and societal and political structures, is necessary to achieve this aim. Therefore, social workers must have a firm understanding of the operations of human behavior, relationships, social interactions, and social and political organizations and structures. Theories provide the necessary knowledge to begin to achieve this aim.

Social Work Theory

A theory is an organized set of assumptions, beliefs, or ideas about particular phenomena in the world. Theory is synonymous with hypothesis, presumption, speculation, belief, idea, and philosophy and is used to help explain or predict situations, actions, and consequences. The social work profession aims to intervene at the point where people meet their environment, which requires social workers to have an understanding of the operations and complexities of working with and within these different systems. Established theories serve as a basis on which to explain human behavior, growth and development, psychological and social functioning, the construction of social order, and the ideas of social justice.

Working within the different layers and multiple systems within society makes social work a unique profession. Social work is a relatively new profession. It originated in the nineteenth century to address the economic inequalities that became more prominent after the end of feudalism and the emergence of the industrial revolution. Social workers aimed to combat the effects of poverty and promote social justice. In doing so, social workers not only needed to work with and understand human behavior, but also understand societal and political structures and the interactions between society and individuals. This need to focus on individuals, and the environment in which they live and interact, required the profession to look to other academic disciplines where established theories of humans, society, and the interaction between the two were already established. Therefore, theories within social work are predominately drawn from other longer standing academic disciplines, such as psychology, sociology, and philosophy.

The Importance and Use of Theory within Social Work

The social work process involves stages of assessment, intervention, evaluation, and endings with an overall goal of promoting human growth and development and social justice (Teater, 2014; Turner, 2011). To initiate each of these stages, social workers must first have an understanding about how best to proceed with each stage given the particular situation, circumstances, and setting in which the practice is taking place. Social workers are working with vulnerable people in situations that are variable. A single approach to social work practice will not fit all. Social workers must tailor their practice to fit the specific needs of the client given her/his circumstances and a consideration of the environment in which the client is interacting. Theory is the critical factor in assisting social workers' understanding of a situation, hypothesizing about how to intervene, and speculating and predicting what might happen in the future (Teater, 2014; Turner, 2011).

Theory guides social work practice at every stage. During the assessment stage, theories help to explain what is happening with a client, what could have contributed to the presenting problem, and what is needed or required to alleviate the problem. A consideration of biological, psychological, and sociological theories during the assessment process assists in explaining the presenting problem and helps to structure and organize the social worker's thinking toward the next step of intervention. The intervention stage is based on theories of why the intervention will work given the particular situation and setting. Finally, theories will guide a social worker in the evaluation stage of the social work process by reassessing a client's biological, psychological, and sociological functioning and explaining the best time to end services.

Social workers assess for risk and needs and then recommend the best course of action for the client. The outcome is anticipated to be positive by alleviating or diminishing a presenting problem, however, it could equally lead to negative outcomes due to negligent practice. Therefore, social workers

should be held accountable for their work. One way to ensure accountability is to work within established theoretical frameworks, justifying the choice of theories and methods used based on evidence, and continually reflecting upon and evaluating the theories and methods used with the client.

The Relationship between Perspective, Theory, Method, and Model

Perspective, theory, method, and model are all important in understanding the role and significance of social work theory. A perspective is a particular value base or point of view that informs how one sees the world. A theory serves to explain or predict situations and provides a rationale as to the course of action that is needed to alleviate presenting problems. A method is the course of action taken to alleviate or diminish the presenting problem (method is often used interchangeably with intervention, practice, or approach). The choice of method is based on the particular theories that have been used to explain the current situation. A model is a structured and organized description of what generally happens in a particular type of situation.

Perspective, theory, method, and model are all equally important in social work practice as one concept depends, influences, or relates to another (Payne, 2014). Theories are generally developed based on a particular perspective. A perspective will influence the type of theories that are used in practice. Theories influence the methods that are selected, and models help to put theories into practice using specific methods. Therefore, social work theories can only be fully understood when considered alongside perspectives, methods, and models.

Types of Social Work Theory

Table 1 presents seven general types of theories that underpin social work practice, which are predominately developed within the professions of psychology and psychiatry and the academic disciplines of psychology, sociology, and philosophy. The following theories were selected as they serve as the basis for understanding human behavior, social interactions, and social constructions, which is required to conduct efficient and effective social work practice. The theories range from explanations of individual biological, psychological, and social development, to theories that explain psychological functioning, relationships, and social interactions, to theories that explain the societal influences on social order, construction of rules and norms, and explanations for inequalities and disadvantage. Table 1 includes the underlying perspectives and a selection of practice methods often employed when utilizing a particular theory.

Developmental Theories

Developmental theories explain the biological, psychological, social, and emotional development as stages over a portion (childhood, adolescence, adulthood, old age) or the whole of the life span. Many developmental theories focus on the development during childhood, such as Piaget's theory of cognitive development, which describes how a child (from birth to over 11 years) moves through discrete stages of cognitive development and intellectual growth to where she/he learns how to think and reason (Crain, 2011). Bowlby's (1988) attachment theory provides an explanation of how a child develops socially and emotionally based on her/his attachment to a primary caregiver, and subsequently the types of relationships and attachments that the child will possess as an adult.

Other developmental theories focus on individuals across the whole of the life span. Erikson's theory of psychosocial development consists of eight continuous stages of life that spans from birth to death and details how a person develops based on biological, psychological, and environmental factors (Crain, 2011). Each stage consists of a type of tension or crisis that an individual needs to work through to move to the next stage. Completing each stage successfully will result in more positive and healthy psychosocial development. Maslow's (1954) hierarchy of needs is a life span developmental theory that explains how individuals must have certain needs met before they can move to achieve a higher level of need. Maslow's theorized needs start with the basic physiological needs for survival and expands until the individual reaches the final need of self-actualization (a term often associated with humanistic theories). Additional theories focusing on the whole of the life span have examined faith, or spiritual, development (Fowler, 1981) and moral and ethical understanding and reasoning (Kohlberg, 1973). Although such theories influence social work practice, particularly through assessment, they have been criticized as being ethnocentric as they were developed based on Western (individualistic) societies, primarily with White men or middle-class individuals/families (Cianci and Gambrel, 2003).

Developmental theories influence social work practice by providing a basis for assessing and understanding a client's physiological, psychological, and emotional development. Such theories are primarily used in the assessment stage of social work practice where a social worker assesses the current level of development and functioning and uses this information to assist in explaining the client's situation, to determine the most appropriate form of intervention, and to hypothesize about future outcomes. Developmental theories are also useful during the evaluation and ending stages to determine any shifts or changes in levels of development.

Psychodynamic Theories

Psychodynamic theories focus on the psychological drives and forces within individuals that explain human behavior and personality. The theories originate from Sigmund Freud's psychoanalysis, which focused on the unconscious mind as the source of psychological distress and dysfunction. Psychoanalytic theory proposed the need for psychoanalytic therapy where the aim is to bring the unresolved issues, developed during childhood, or repressed trauma buried within the unconscious to the conscious mind in order for the client to begin to address these unresolved and underlying problems (Sharf, 2012).

Psychodynamic theories primarily deal with the unconscious motives that underpin an individual's personality and

Table 1 Theories used in social work practice

Theory	*Perspective*	*Practice method*
Developmental theories Includes: Cognitive development Attachment theories Life stage theories Theories of need Faith and moral reasoning	Developmental perspective	Assessment
Psychodynamic theories Includes: Psychoanalytic theory Ego psychology Object relations theory Self-psychology Crisis theory	Psychodynamic perspective	Psychodynamic therapy Psychoanalytic therapy Psychotherapy Jungian analysis Transactional analysis Crisis intervention Task-centered social work
Behavioral and social learning theories Includes: Behavioral theory Social learning theory Cognitive theory Social cognitive theory Group work theory	Social learning perspective	Behavioral therapy Cognitive therapy Cognitive behavioral therapy Cognitive-based transactional analysis Task-centered social work Group work
Humanistic theories Includes: Phenomenology Existentialism	Existentialism perspective; Phenomenological perspective; Developmental perspective; Humanistic perspective; Strengths perspective	Person-centered therapy Existential therapy Counseling Gestalt therapy Hypnosis and meditation Motivational interviewing Advocacy
Social constructivist theories Includes: Symbolic interactionism Communication theory Role theory	Social constructivist perspective; Strengths perspective	Narrative therapy Social constructivist approach Solution-focused brief therapy Antioppressive, antiracist, and antidiscriminatory practice Cultural competency Advocacy
Systems theories Includes: General systems theory Life model Family systems theory	Systems perspective; Ecological perspective; Ecosystems; Person-in-environment perspective; Strengths perspective	Assessment Couple and family therapy Family systems therapy Community development Community practice Task-centered social work
Critical theories Includes: Critical social work theory Conflict theory Empowerment theory Feminist theory Antioppressive, antiracist, and antidiscriminatory theory Environmental and green social work	Critical perspective; Conflict perspective; Radical perspective; Feminist perspective; Black perspective; Ecological perspective; Systems perspective	Advocacy Empowerment approach Antioppressive, antiracist, and antidiscriminatory practice Cultural competency Consciousness-raising Community practice Community development Task-centered social work Crisis intervention

behavior. Childhood experiences are seen as critical in the development of the personality, behavior, and psychological thinking of an individual in later life, particularly psychological distress and dysfunction. Freud's drive theory, involving the three states of being (id, ego, and superego), are seen as important in understanding the role of the unconscious. The id is the unconscious that seeks self-gratification and fuels instincts, the superego is the conscious moral reasoning based on one's moral values and society's values, and the ego is the mediator between the id and the superego and seeks to make decisions based on the id's instincts and need for self-gratification and the superego's call for decisions based on moral values (Sharf, 2012).

Defense mechanisms, transference, and countertransference are terms often used when considering psychodynamic theories. Defense mechanisms are the tools used by the

unconscious mind to prevent anxiety caused by unresolved issues and trauma. The mechanisms distort reality and are used to protect oneself by distancing from reality. Common defense mechanisms include denial, disassociation, regression, acting out, projection, or displacement (Sharf, 2012). Transference explains the act of a client unconsciously projecting thoughts, feeling and experiences of relationships, or interactions with previous significant figures onto a social worker. Countertransference is where the social worker's unconscious responds to signals received from the client and the social worker acts out a particular role (e.g., taking a parenting role) (Ruch et al., 2010).

Additional theorists have expanded on the ideas of the role of the unconscious and have shifted the attention in psychodynamic thought from one that focused on conflict, to one that focuses more on relationships. Jungian analysis and therapy explores the conscious and unconscious, but is equally interested in extroverted and introverted personality, archetypes, symbols, and dreams (Sharf, 2012). Adler's individual psychology, more commonly referred to as differential psychology, explores the ideas of inferiority, superiority, birth order, and individual differences (Sharf, 2012). Klein's object relations theory explores how relationships developed in infancy and childhood are embedded in the unconscious mind and form the focus of individuals' drives, views of themselves and others, influences their personality in adulthood, and dictates how they interact in interpersonal relationships (Sharf, 2012). Kohut's self-psychology expanded on object relations theory, aims to focus more on the self and the deficits within the self (Sharf, 2012).

Crisis theory is also classified as a psychodynamic theory as it explains how people cope with stressful situations and how they have the capability to grow, develop, and change based on the crisis. The theory holds that individuals reach a state of crisis when their existing coping skills are unable to deal with stressful or traumatic situations resulting in psychological and physiological distress (Caplan, 1964).

Psychodynamic theories are useful in social work assessments to explore a client's past experiences, hypothesize about how such experiences are contributing to the presenting problem and how to address the problem (or crisis). In this sense, psychodynamic theories influence social work assessments as well as the interventions employed, such as psychotherapy, crisis intervention, or transactional analysis. Acknowledgments and understandings of defense mechanisms can assist a social worker in explaining a client's behavior, interpersonal relationships, or reactions to information. Considering the role of transference and countertransference can assist the social worker in building the social work–client relationship (Ruch et al., 2010).

Behavioral, Cognitive, and Social Learning Theories

Behavioral, cognitive, and social learning theories explain how individuals come to learn how to think, feel, and behave. Behavioral and social learning theories were developed by theorists who disagreed with the psychodynamic theorists. Instead, they hypothesized that individuals learn how to behave by cognitively and behaviorally responding to cues received from the interactions with their social environment.

The development of behavioral and social learning theories originated through experiments conducted by psychologists during the early 1900s that explored how animals can be conditioned to respond in a particular way. The studies from Ivan Pavlov and John Watson lead to the ideas of classical and respondent conditioning, which proposed that behaviors are the result of prior learning and can be learned and unlearned. The studies from Edward Lee Thorndike and B. F. Skinner lead to the ideas of operant or instrumental conditioning, which proposed that providing positive or negative consequences to behaviors lead to learning. Thorndike and Skinner found that positive consequences for behavior, or positive reinforcers, will lead to an increase in the occurrence of the behavior and negative consequences, or negative reinforcers, will result in a decrease in the occurrence of the behavior (Teater, 2014). Therefore, the behavioral theories postulate that behaviors can be learned, unlearned, and modified based on the reinforcers received from the social environment.

Bandura's social learning theory (1977) extends on the behavioral theory by stating that not only do individuals learn to think, feel, and behave through the receipt of reinforcers, but they also learn through observing and then modeling the actions of others. Individuals must observe a model (e.g., parents, teachers, media characters), pay attention to the behaviors or actions of the model, imitate the behaviors or actions, and receive a reinforcement (positive or negative) for modeling the behavior or action, which will influence whether the behavior or action is imitated in the future.

Cognitive theory is also a learning theory, but rather than addressing the learning of behaviors that occur merely through reinforcers from the social environment, cognitive theory aims to explain how humans are thinking creatures, who will make choices about their behavior based on what makes the most sense to them. Cognitive theory, as expressed through Ellis' ABC model, states that individuals experience an activating event (A), they process this event through their cognition and belief system (B), which mediates the behavioral consequence (C) (Sharf, 2012).

Behavioral, cognitive, and social learning theories are useful to social work practice as they provide a basis for exploring how behaviors are learned and unlearned; how a client's social environment can be positively or negatively reinforcing particular thoughts, feelings, and behaviors; or how a client's cognition and belief system can be contributing to problematic behaviors. Such theories are primarily used in the intervention stage of social work practice where a social worker attempts to modify or eliminate a particular thought, feeling, or behavior. Such interventions often used in social work practice are task-centered social work, cognitive therapy, behavioral therapy, cognitive behavioral therapy, or group work.

Humanistic Theories

Humanistic theories used within social work practice originated from humanistic psychology, which developed in response to the psychodynamic and behavioral theories that focused on human behavior and personality being determined by the unconscious or through reinforcers from the social environment. Humanistic psychology saw that these two prominent theories viewed individuals in response to events that they were not able to fully control and failed to view the individual as meaning-making and purposeful. Humanistic

theories stress the ultimate good of humans and their potential, creativity, health, hope, meaning, connection, purpose, and ability to reach self-actualization, or for individuals to achieve their full potential (Crain, 2011).

Phenomenology and existentialism are two theoretical schools of thought that underpin humanistic theory. Phenomenology is the exploration of conscious perspectives and experiences of phenomena and the meanings one attributes to such phenomena. Existentialism focuses on individual existence and the meaning one gives to her/his life. The central aspect to both phenomenology and existentialism is the lived subjective and conscious experiences of individuals, how individuals experience and attribute meaning to the phenomena to which they encounter, how they make sense of life, and how they make sense of their place and meaning within the world (Sharf, 2012).

The five core values of humanistic theory include the following: (1) human beings supersede the sum of their parts; (2) human beings have their existence in a uniquely human context, as well as in a cosmic ecology; (3) human beings are conscious – they are aware and aware of being aware both of oneself and in the context of other people; (4) human beings have some choice and, thus, responsibility; and (5) human beings are intentional, aim at goals, are aware that they cause future events, and seek meaning, value, and creativity (Greening, 2006: p. 239).

Humanistic theory can be found to underpin aspects of developmental theories, such as Maslow's hierarchy of needs, where an individual aims to achieve self-actualization, and Erikson's later stage of old age, ego integrity versus despair, as well as many therapeutic approaches that aim to explore and respect the experiences of individuals as humans and the meanings they attribute to such phenomena. Humanistic theory underpins Rogers' person-centered approach, which highlights the importance of empathy, unconditional positive regard, and genuineness in developing a therapeutic relationship that will lead to personality and behavioral change within the client (Sharf, 2012).

Humanistic theories are useful to social work practice as they provide a theoretical basis for viewing individuals, their experiences in the world, and the meanings they attribute to such experiences. Such theories are primarily used in assessment and intervention stages of social work practice. Using humanistic theories in the assessment stage would involve a social worker being empathic, having unconditional positive regard, and being genuine when interacting with a client in order to establish a relationship and fully see the client as an individual when assessing her/his situation. The intervention stage might involve a social worker employing interventions that would explore the client's experiences, meanings, hopes, and aspirations. Such interventions often used in social work practice are person-centered approach, existential therapy, counseling, gestalt therapy, hypnosis, meditation, motivational interviewing, and advocacy.

Social Constructivist Theories

Social constructivist theories focus on the creation of reality and how individuals view the world. The theories are related to the humanistic theories, discussed above, in that they were influenced by phenomenology as well as philosophical ideas of what is real and what is socially constructed. The basis of the modern social constructivist theories stems from Berger and Luckman (1966), who explored reality creation and the influence of individual meaning based on life experiences, societal and cultural expectations, rules and norms, which they termed 'social constructionism.' Since then, several theorists have refined social constructionism into three more specific theories, which emphasize either individual or social forces in reality construction (Teater, 2014). Social constructionism holds that reality is constructed through the use of language in interactions with others and is primarily influenced by history, society, and culture. Constructivism holds that reality is constructed more through one's biological forces through developmental processes, cognitive structures, and the human mind. Social constructivism combines both social constructionism and constructivism by placing the emphasis of reality construction on both societal and biological factors. There is a joint focus on reality construction being influenced by both nature and nurture.

Social constructivist theories have underpinned several theories often used within social work practice. Symbolic interactionism focuses on three core principles of meaning, language, and thought. The theory proposed that people use symbols (words, rules, roles) to give meaning and to make sense of the world. The meanings are transmitted to others through language. People are believed to interact with others and society, and assign meaning and symbols to these types of interactions and relationships (DeLamater and Myers, 2011). Future interactions are, therefore, dependent on the types of meanings and symbols one has attributed to that relationship or situation, thus, individuals act on what they believe versus what is objectively true. Communication theory holds that "people cannot NOT communicate" and that "all behavior is communication and all communication affects behavior" (Watzlawick et al., 1967). Role theory examines how people play out socially defined roles (e.g., mother, sister, wife, manager, teacher) and their ability to adhere to society's expectations of acceptable and unacceptable forms of behavior for the particular role (DeLamater and Myers, 2011).

Social constructivist theories are useful to social work practice as they provide a theoretical basis for understanding how realities and views of the world are individual-specific and created through a combination of interactions with society, the societal structures of history, culture, rules, and norms, and the meanings that individuals attribute to such interactions. Such theories are primarily used in the assessment and intervention stages of social work practice. Using social constructivist theories in assessment involves a social worker taking a position of curiosity with the client, using language to attempt to understand the client's reality and view of the world, and acknowledging that no two clients have the same reality of view of the world despite living through similar experiences. The intervention stage involves a social worker employing interventions that would explore the client's experiences and meanings and attempt to reframe thoughts or views that are contributing to the presenting problem into ones that are more acceptable. Interventions also include the social worker challenging societal assumptions or social constructs that are preventing the client from growing and developing. Such interventions often

used in social work practice are narrative therapy; social constructivist approach; solution-focused brief therapy; anti-oppressive, antiracist and antidiscriminatory practices; cultural competency; and advocacy.

Systems Theories

Systems theories are based on the belief that individuals do not operate in isolation, but rather grow and develop in interaction with their physical and social environment. Systems theories derive from general systems theory, which explores the parts of a system that interconnect and interact to make a complete whole (Teater, 2014). Within social work, systems can constitute individuals, couples, families, communities, organizations, society, and the world. Systems theories hold that each system should be viewed as consisting of several elements that make the system a functional whole, and each system should be viewed in relation to the other systems that can cause a change or reaction within the main system. For example, when working with clients, social workers should consider the bio-psycho-social aspects of the client by looking at physical and psychological functioning, social relationships, and community or societal structures that impact on the client.

The life model (Gitterman and Germain, 2008) of social work practice was greatly influenced by system theories as well as the person-in-environment perspective (Karls and Wandrei, 1994), both of which examine how social work is a unique discipline, in that it focuses on the point where individuals interact with their environment. Such systems theories aim to move social work practice away from focusing solely on the individual, such as with development theories, psychodynamic theories, and behavioral theories, and instead focus holistically on the individual within her/his environment (often referred to as human behavior in the social environment). Consideration of the environment includes the physical space, the social context, and the individual's culture and history. The aim of systems theories is to create homeostasis, or a favorable person–environment fit, in that the individual interacts and responds to her/his environment where interactions and change are contributing to positive growth and development and social functioning.

Family systems theory adapted the main concepts of general systems theory in understanding and working with families. The family is viewed as a system with each family member playing a critical part. Family systems theory holds that a change in one part of the family system will create a change in other parts of the family system, yet this is often variable depending on the boundaries of the family, the patterns, messages and rules of the family, and the family's responsiveness to change (Sharf, 2012).

Systems theories are useful to social work practice as they provide a theoretical basis for assessing a client holistically by examining all the systems within her/his environment. Such theories are primarily used in assessment and intervention stages of social work practice where the social worker assesses the client holistically by considering psychological, biological, and social functioning, as well as assessing the interaction of other systems within the client's environment, particularly those that could be contributing to the presenting problem. Based on the assessment, underpinned by systems theory, the social worker determines which system needs the intervention. Although the client may be an individual, the social worker may deem the family system, community system, or even political systems as the focus for intervention. Interventions most commonly used in social work practice include couple and family therapy, family systems therapy, community development, and community practice.

Critical Theories

Critical theories in social work aim to examine and critique social and political structures and functioning and their effects on individuals, families, and communities (Payne, 2014). Such critiques and understanding of social and political structures will assist in social work's aims of tackling inequalities and disadvantage, and promoting social justice. Conflict theory is a form of critical theory, based on the works of Karl Marx and Max Weber, which focus on inequalities within society, such as wealth, power, class, and how such inequalities impact on individual life experiences and chances creating conflict between and within social groups (see Hier, 2005 for a review of critical theories). Numerous critical theories have been applied to social work practice in an attempt to acknowledge the inequalities and disadvantage that clients can and do experience within the current social and political structures and to encourage courses of action and interventions that will challenge such inequalities.

Empowerment theory acknowledges that oppression is present within society and affects individuals and communities, yet individuals and communities do have strengths and resources and can combat such societal oppression (Adams, 2008). Individuals need to have power and control over their lives in order to obtain and use the resources necessary for positive growth and development, which can be achieved through individual and collective strategies. Feminist theory explains the inequalities between women and men, the oppression of women, the development and consequences of socially defined gender roles and seeks to explain how development and life experiences differ for women and men (Enns, 2004; Hyde and Else-Quest, 2013). Antioppressive, anti-discriminatory, and antiracist theories acknowledge the structural oppression and discrimination that impacts on clients and seeks to promote social work practice that challenges such structures by working individually with clients and participating and joining in social movements that, in turn, challenge oppression and discrimination within larger structures (see Dalrymple and Burke, 2006; Dominelli, 2008; Robbins, 2011; Thompson, 2006). Environmental and green social work applies a social work perspective to the ways in which environmental issues and crises can lead to disadvantage and limit resources, and focuses on targeting the social–political forces that impact on individual, community, and global well-being in order to promote environmental and social justice (Dominelli, 2012; Gray et al., 2013).

Critical theories are useful to social work practice as they provide a theoretical basis for assessing a client within her/his environment by acknowledging how societal and political structures and processes can be oppressive or discriminatory. Such theories take the focus away from blaming the client and examining other factors much larger than the client that could

be contributing to the presenting problem. Such theories are primarily used in assessment and intervention stages of social work. Assessment strategies involve the social worker examining societal and political structures and how these are impacting on the client (e.g., individual, family, community). Intervention strategies involve the social worker providing direct work with the client or participating in social movements and advocacy to eliminate structural oppression and discrimination. Such interventions commonly used in social work practice include advocacy; the empowerment approach; antioppressive, antiracist, and antidiscriminatory practice; consciousness-raising; crisis intervention; community practice; and community development.

Conclusion: The Application of Social Work Theory

The seven broad categories of theories range from a focus on individual psychological development to the influence of societal and political structures on human and communal growth and development. The profession of social work seeks to promote human growth and development as well as social justice; therefore, knowledge of the complete range of theories is necessary for effective social work practice although the choice and application will vary.

The use of theoretical approaches within social work practice will vary depending on social, political, and cultural structures, agency and organizational structures, and individual social worker preference. Social workers practice under legal mandates, which prescribe when and how a social worker can intervene. Such legal mandates are influenced by social and cultural values and norms, which dictate what a socially acceptable behavior is. For example, a social worker will utilize different theories when working with mental illness based on the social and legal structures that define mental illness, theorize its cause, and hypothesize about ways in which it is to be treated or not. Agency and organizational structures also dictate the types of theories, social workers use in practice. Agencies may specialize on a specific aspect of social work practice, such as mental health or community development, and will, therefore, utilize theories that help to achieve the agency's aim. For example, one mental health agency may work from humanistic theories and focus on person-centered counseling or solution-focused brief therapy to alleviate a presenting problem, whereas another agency may work from psychodynamic theories and concentrate on psychoanalysis or psychotherapy that aims to bring the unconscious to the conscious. Finally, the application of social work theories is dependent on the social worker's values and preferences. For example, one social worker may believe that behavioral change is initiated through conditioning (behavioral theory) whereas another may believe behavioral change is initiated through individual motivation (humanistic theory). The theoretical stance of the social worker will influence the choice of intervention (cognitive behavioral therapy vs motivational interviewing). It is important to note that the influences of societal/political, agency/organization, and social worker preferences may not be congruent.

Social work practice does not necessarily adhere to one category of theories, but often social workers must choose from a range of theories to create an intervention package that is suitable for the client. For example, a child welfare social worker may use developmental theories to explain the current developmental stage of a child; systems theories to conduct a holistic assessment of the biological family; social constructivist theories to understand the child and family's perspective; behavioral, cognitive, and social learning theories to hypothesize about how thoughts, feelings, and behaviors are contributing to the presenting problem; and critical theories to understand the context in which the child and family operate and in which the social worker practices. In selecting theories for practice, social workers must acknowledge that not all theories will be applicable to all clients and that theories that appear to fit in one situation will not always fit in similar situations. Social workers must combine theoretical knowledge with research evidence and reflections on practice to promote most accurately human growth and development and social justice.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Behavior Therapy: Background, Basic Principles, and Early History; Bruner's Theory of Cognitive Development; Classical Conditioning Methods in Psychotherapy; Community Work; Concepts and Methods of Cognitive Therapies; Counseling Psychology; Counseling and Psychotherapy: Ethnic and Cultural Differences; Critical Social Work Practice; Ecological Social Work: Shifting Paradigms in Environmental Practice; Erikson's Theory of Psychosocial Development; Experiential Psychotherapy; Feminist Psychotherapy; Feminist Research and Social Work; Group Psychotherapy, Clinical Psychology of; Hypnosis; Integrated Attachment Theory; Interpersonal Psychotherapy; Marxist Approaches to Social Work; Moral Development, Theories of; Motivation: Life Course and Sociological Perspectives; Neo-Piagetian Theories of Cognitive Development; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Psychodynamic Therapy: Short-Term; Psychosocial Work; Psychotherapy Integration; Radical Social Work; Reinforcement, Principle of; Social Constructionism; Spirituality; Therapist–Patient Relationship; Vygotsky's Theory of Cognitive Development.

Bibliography

Adams, R., 2008. Empowerment, Participation and Social Work, fourth ed. Palgrave Macmillan, Basingstoke.

Bandura, A., 1977. Social Learning Theory. Prentice-Hall, Englewood Cliffs, NJ.

Berger, P.L., Luckman, T., 1966. The Social Construction of Reality. Doubleday, Garden City, NY.

Bowlby, J., 1988. A Secure Base: Clinical Application of Attachment Theory. Routledge, London.

Caplan, G., 1964. Principles of Preventative Psychiatry. Basic Books, New York.

Cianci, R., Gambrel, P.A., 2003. Maslow's hierarchy of needs: does it apply in a collectivist culture. Journal of Applied Management and Entrepreneurship 8, 143–161.

Crain, W., 2011. Theories of Development: Concepts and Applications, sixth ed. Pearson, Upper Saddle River, NJ.

Dalrymple, J., Burke, B., 2006. Anti-oppressive Practice: Social Care and the Law. Open University Press, Maidenhead.

DeLamater, J.D., Myers, D.J., 2011. Social Psychology, seventh ed. Wadsworth Cengage Learning, Belmont, CA.

Dominelli, L., 2008. Anti-racist Social Work, third ed. Palgrave Macmillan, Basingstoke.

Dominelli, L., 2012. Green Social Work: From Environmental Crisis to Environmental Justice. Polity Press, Cambridge, MA.
Enns, C.Z., 2004. Feminist Theories and Feminist Psychotherapies: Origins, Themes, and Diversity, second ed. Haworth, New York.
Fowler, J.W., 1981. Stages of Faith: The Psychology of Human Development and the Quest for Meaning. Harper & Row, New York.
Gitterman, A., Germain, C.B., 2008. The Life Model of Social Work Practice: Advances in Theory and Practice, third ed. Columbia University Press, New York.
Gray, M., Coates, J., Hetherington, T., 2013. Environmental Social Work. Routledge, Abingdon.
Greening, T., 2006. Five basic postulates of humanistic psychology. Journal of Humanistic Psychology 46, 239.
Hier, S.P., 2005. Contemporary Sociological Thought: Themes and Theories. Canadian Scholars' Press Inc., Toronto.
Hyde, J.S., Else-Quest, N., 2013. Half the Human Experience: The Psychology of Women, eighth ed. Wadsworth, Belmont, CA.
Karls, J.M., Wandrei, K.E., 1994. Person-in-Environment System: The PIE Classification System for Social Functioning Problems. NASW Press, Washington, DC.
Kohlberg, L., 1973. The claim to moral adequacy of a highest stage of moral judgment. Journal of Philosophy 70, 630–646.
Maslow, A., 1954. Motivation and Personality. Harper, New York.
Payne, M., 2014. Modern Social Work Theory, fourth ed. Palgrave Macmillan, Basingstoke.
Robbins, S.P., 2011. Oppression theory and social work treatment. In: Turner, F. (Ed.), Social Work Treatment: Interlocking Theoretical Approaches. Oxford University Press, Oxford, pp. 343–353.
Ruch, G., Turney, D., Ward, A., 2010. Relationship-Based Social Work: Getting to the Heart of Practice. Jessica Kingsley Publishers, London.
Sharf, R.S., 2012. Theories of Psychotherapy and Counseling: Concepts and Cases, fifth ed. Brooks/Cole, Belmont, CA.
Teater, B., 2014. An Introduction to Applying Social Work Theories and Methods, second ed. Open University Press, Maidenhead.
Thompson, N., 2006. Anti-discriminatory Practice, fourth ed. Palgrave Macmillan, Basingstoke.
Turner, F., 2011. Theory and social work treatment. In: Turner, F. (Ed.), Social Work Treatment: Interlocking Theoretical Approaches. Oxford University Press, Oxford, pp. 3–14.
Watzlawick, P., Beavin, J.H., Jackson, D.D., 1967. Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes. WW Norton, New York.

Social Work with Racialized Groups: Frameworks for Practice

Wanda T Bernard, Dalhousie University, Halifax, NS, Canada
Ellice Daniel, Ashura Counselling and Consulting, Surrey, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article considers the historical legacy of European dominance over individuals from racialized groups. It discusses the social construction of race as it relates to culture and the subsequent oppression of individuals from racialized groups. Practice frameworks for working with individuals from racialized groups are discussed and implications for social work practice are also included.

Introduction

Whilst racially sensitive social work practice has been used historically when teaching about practice with specific racial and ethnic groups (Devore and Schlesinger, 1981/1999), it has not maintained its currency, largely because it did not address issues of systemic racism and oppression. In this article, we consider the historical legacy of European dominance over individuals from racialized ethnic groups, their subsequent oppression, and consider frameworks for practice with these groups. It begins with a discussion of the social construction of race as it relates to culture, and then practice frameworks for working with individuals from racialized groups. It concludes with a discussion of implications for social work practice.

Race as a Social Construct

Understanding the social meanings attached to the word 'race' encourages an analysis of the social processes that give meaning to the word 'race.' The concept of racialization focuses on the process by which 'race' becomes meaningful in the context of unequal power relations rather than on 'race' itself. Racialization supersedes the 'race relations' paradigm in both the United Kingdom and the United States because it looks at how certain groups have come to be defined as groups of stratified 'races' and the various factors involved in the institutionalization of racism (Garner, 2010). "[R]acism exists throughout the globe … and provides a basis for social exclusion on the basis of race" (Dominelli, 1997: p. 7). Such exclusion is considered normal when dominant groups consider those who are different as the 'other.' Othering is the process of treating difference as deviant or pathological and therefore undesirable (Dominelli, 2003: p. 17). Othering reproduces privileges those that are deemed to belong, and disadvantages those considered outsiders.

Historically, interactions between those of European decent and individuals from racialized groups of people were dichotomous in nature with one group being the oppressor (European) and the other group being the oppressed. According to Vickers (2002), the term racialized refers to "the political and ideological processes by which populations are linked to physical characteristics" (p. 190) and socially constructed as racially distinct, primarily based on a manifestation of unequal power between groups (Baum et al., 2001).

The division between these two groups depicts one of the deepest and most pervasive histories relating to 'race' relations and racism. The concept of 'race' has been deeply involved in constructing and rationalizing imbalanced social relations since it was introduced in the seventeenth century (Baum et al., 2001). The primary purpose of this construction was to separate the European or dominant group from other groups of people and exclusion was the guiding principle for social order (Deters, 1997; Christian, 2000), as evidenced, for example, by rules about citizenship, immigration laws, and apartheid (Gillem et al., 2001). This view has also evolved out of a historical need to create a hierarchy that would maintain the status quo of European supremacy, dominance, and privilege.

Some social scientists view race as a social construct (Dominelli, 1997; Better, 2008; Kirton, 2001; Parker and Song, 2001) with harmful effects, while recognizing that racialized identities can be an important mobilizing force for those struggling against discrimination and disadvantage (Parker and Song, 2001). According to Baum et al. (2001), understanding race in terms of racialization avoids any reification of racial difference because it emphasizes the historically changing and politically contingent nature of the process of racialization. Therefore, according to the social constructionist view, race is a product, not a cause of racism (Kobayashi and Johnson, 2007).

Culture

According to Better (2008), a social construct is a culturally determined idea. Therefore, race must be understood within the context of culture. However, culture is not static and cannot be viewed as being uniformly distributed or having a uniform impact on all members (Korbin, 2002). Culture is "a complex whole that includes knowledge, values, beliefs and other capabilities and habits of people" (McGibbon and Etowa, 2009: p. 65).

Thus, a fixed definition reduces culture to stable groups of characteristics (Korbin, 2002) and encourages the perpetuation of stereotypes that do not recognize the ongoing changes that people and societies undergo (Mederos and Woldeguiorguis, 2003). Moreover, a fluid definition of culture recognizes change and individual differences while recognizing the lasting impact of certain experiences (Mederos and Woldeguiorguis,

2003; Yee and Dumbrill, 2003) and set the stage for more responsive service delivery (Mederos and Woldeguiorguis, 2003).

While maintaining the idea that culture is not static, it is necessary, "to unpack culture to understand how it works in all of its complexity" (Korbin, 2002: p. 638). The postmodernist understanding of culture is that all cultural identities are socially and historically constructed and, therefore, changeable in response to different internal and external contingencies (Williams, 2006). Accordingly, postmodern cultural 'competence' involves exploring the collection of identities and experiences that produce continually evolving, unique cultural experiences for clients (Dean, 2001).

A scan of the available research that discusses the dynamics of 'race,' culture, and ethnicity revealed terms such as cultural sensitivity, cross-cultural competence, cross-cultural expertise, cross-cultural effectiveness, cultural responsiveness, cultural awareness, culturally skilled, cultural competence, cultural safety, culturally relevant, and multicultural competence. Each of these terms include the term 'culture' which alludes to the idea that one can strive to become effective working with various cultures as opposed to analyzing the social processes that contribute to the racialization of groups of people.

Whaley (2008) found that cultural sensitivity and cultural competence were among the most popular terms used in the existing body of literature. Moreover, he found that the majority of literature encompassed cultural sensitivity while a smaller portion of the literature made reference to cultural competence. Although these two terms were noted to be the most popular in the existing body of literature, cultural sensitivity seemed to be more common in the existing body of literature (Whaley, 2008).

However, a focus on cultural competence and cultural sensitivity fails to recognize, and therefore does not challenge institutional racism (Dominelli, 1997).

Cultural Competence and Cultural Sensitivity

Modernist cultural competence has been a construct on which programs related to diversity have been based. Cultural competence is defined as "a set of congruent behaviors, attitudes, policies, and structures that come together in a system or agency or among professionals and enables the system, agency, or professionals to work effectively in cross-cultural situations" (Cross et al., 1989: p. 7). Cultural competence can be conceptualized as consisting of components such as cultural awareness, the affective dimension; cultural sensitivity, the attitudinal dimension; cultural knowledge, the cognitive dimension; and cultural skills, behavioral dimension (Rew et al., 2003).

Cultural competence may not be an appropriate construct because competence in a specific cultural group is likely not transferable between individuals, families, and communities (McGibbon and Etowa, 2009). The concept of cultural competence denotes working with cultural groups other than one's own (Korbin, 2002) and "the ability to transform knowledge and cultural awareness into health and/or psychosocial interventions that support and sustain healthy client-system functioning within the appropriate cultural context" (McPhatter, 1997: p. 261).

Moreover, some researchers (Pon, 2009; Sakamoto, 2007) feel it promotes an out-dated view of culture and is a form of racism. According to Pon (2009), cultural competence is similar to the processes involved in racism by means of "otherizing non-whites and by deploying modernist and absolutist views of culture to promote racialized and stereotypical views of cultural groups without ever having to use racialist language" (p. 64).

As an attempt to encourage cultural sensitivity, the perceived cultural differences may promote stereotypes and reinforce the concept of 'others' while perpetuating the idea that certain behaviors should be controlled and adjusted to resemble the norms of the dominant group (McGibbon and Etowa, 2009). For example, this approach assumes that cultural groups have stable identities, creating harmful stereotypes such as "Africans believe this … Mexicans believe that …" (McGibbon and Etowa, 2009: p. 76).

Consequently, these groups continue to be denied the realities of systemic racism, discrimination as well as other forms of social disparities and blamed for their difficulties (McGibbon and Etowa, 2009). Additionally, Dean (2001) views the concept of cultural competence as flawed. She argues that cultural competence assigns membership, into cultural categories, by particular aspects of identity such as race or ethnicity, which are viewed as fixed and sustained over time.

One of the challenges of using a cultural competence model is that it operates at an individual level, by applying specific techniques and skills based on a person's perceived knowledge about a particular culture. In a diverse society, people choose between various cultural options offered by a variety of cultural traditions. This approach does not capture the diverse, fluid, and complex nature of culture. Moreover, it may attempt to predict the beliefs and behaviors of those with racialized identities based on what is known about a person's 'race' or culture.

Since it is not possible to predict how a person from a particular culture self-identifies, a person's group membership cannot be assumed to indicate their culture. Moreover, stagnant definitions of culture fail to effectively address diversity, within membership to a cultural group, leaving minimal space for cultural change (Dean, 2001). Consequently, such an approach may inadvertently place blame on a person's culture. It may also emphasize cultural differences, which can obscure structural power imbalances, on an institutional level.

Cultural competence does not reach far enough in addressing institutionalized oppressions. It can leave those working from this approach, unequipped to deal with institutional racism and oppression, where it permeates, on structural and global levels. Moreover, it only reifies existing racial categories rather than deconstructing systemic factors that contribute to racism, discrimination, and oppression. Despite the widespread application of cultural competence, to promote organizational practices to meet the needs of diverse populations, it remains a vaguely defined goal, due to the elusive nature of its central construct, which is culture. Additionally, a failure of the cultural competency models is that they do not take into account the impact of colonization and the enduring legacy of oppression and domination, especially that imposed by the North on the South (Dominelli, 1997; Bernard and Moriah, 2007; Carniol, 2000; Dalrymple and Burke, 1995; Christensen, 1991; Fong and Gibbs, 1995).

A reliance on cultural competency frameworks can become what Dominelli (2003) calls a 'tool kit' approach to social work.

Social Work Ethics

Despite the elusive nature of culture, and the problematic of cultural competency models, ethical standards for social work practice in North America are based on both cultural competence and cultural sensitivity. Moreover, providing culturally competent practice has been at the core of ethical social work practice (Jackson and Samuels, 2011). According to the National Association of Social Workers (NASW),

> Culturally competent services are needed beyond race and ethnicity. Culturally competent social workers are also better able to address issues of gender and help persons with disabilities, older adults, gays, lesbians, bisexuals, and transgender people. A working knowledge of these groups' cultures and values helps social workers tailor care so it is effective and appropriate for their clients' needs. (NASW, 2008: p. 8)

Therefore, social workers are ethically required to understand cultural competence and social diversity. More specifically, "social workers should obtain education about and seek to understand the nature of social diversity and oppression with respect to race, ethnicity, national origin, color, sex, sexual orientation, gender identity or expression, age, marital status, political belief, religion, immigration status, and mental or physical disability" (NASW, 2008: p. 8).

Likewise, according to the Canadian Association of Social Workers (CASW)

> Social workers do not discriminate against any person on the basis of age, abilities, ethnic background, gender, language, marital status, national ancestry, political affiliation, race, religion, sexual orientation, or socio-economic status. (CASW, 2005: p. 4)

Specifically, social workers are required to demonstrate cultural awareness and sensitivity by striving to understand culture and its function in human behavior and society, recognizing the strengths that exist in all cultures, acknowledging the diversity within and among individuals, communities, and cultures. They should acknowledge and respect the impact that their own heritage, values, beliefs, and preferences can have on their practice and on clients whose background and values may be different from their own, as well as those perceived as similar. They should seek a working knowledge and understanding of a client's racial and cultural affiliations, identities, values, beliefs, and customs. A critical exploration of culture demands that social workers also examine the intersections of oppression, and to be aware of the diversity of experiences and realities within groups as well (Dominelli, 2003; Christensen, 1991; Zamparo et al., 2000)

Lastly, where possible, social workers should provide or secure social work services in the language chosen by the client (CASW, 2005: p. 4). Acknowledging the historical processes that have contributed to socially constructed ideas about race is crucial to attempt to gain an understanding of the oppression individuals from racialized groups faced historically and continue to experience presently.

Oppression

As outlined at the beginning of this article, oppression was historically based on coercive laws and rules to preserve the dominant European groups. According to Mullaly (2002), oppression is determined by a person being blocked from opportunities to self-development, being excluded from full participation in society, not having certain rights that the dominant group takes for granted, or being assigned a second-class citizenship. People with racialized identities have been discriminated against historically and are more likely to be oppressed on social, institutional, systemic levels and can result in internalized oppression (Mullaly, 2002). Oppression is complex, multidimensional, and fluid (Dominelli, 2003: p. 9) and is evident in social relations. Dominelli (2003) states that those in the dominant group deny and block access and agency to those in the target group.

Socially, members from racialized groups may be categorized and mistreated by virtue of their racialized identities. Once a social definition is constructed, various social practices and psychological responses contribute to institutional racism. Institutional racism "exists when the established policies and practices of a particular organization reflect and produce differential treatment towards, and outcomes for, various groups" (James et al., 2010: p. 89).

On an institutional level, this can be seen through laws, customs, and practices that exist to perpetuate inequities based on a person's racialized identity. Because individuals from racialized groups are excluded from the dominant privileged social positions and included in the subordinate group they may experience exploitation, marginalization, powerlessness, cultural imperialism, and violence. The processes involved in the institutionalization of racism illustrate how oppression is expressed and reproduced systemically by the actions of institutions and individuals working within these institutions.

Systemically, exploitation is realized through a structural relationship between the 'haves' and 'have not' groups. Also social rules about what work is, hierarchies in agencies, differences in wages all operate through a systematic process to produce and reproduce relations of power and inequality. This can be seen through the expectations that members from racialized groups will carry out menial labor tasks while management positions are held by members of the dominant groups. Accordingly, members from racialized groups are likely to be exploited and subjected to further marginalization (Young, 1990).

Marginalization primarily affects individuals from racialized groups who are, by virtue of social constructions, "confined to the margins of society because the labour market cannot or will not accommodate them" (Young, 1990: p. 43). These situations can contribute to these individuals feeling powerless. Young (1990) indicated that powerlessness affects primarily nonprofessional workers, but also individuals from racialized groups and is based on the social division of labor.

In this sense, the power and respect that accompany the privileged status of being professional also involve racism whereby, individuals from racialized groups who are professionals must continuously prove they deserve respect and that they belong (Young, 1990; Dominelli, 1997; Better, 2008; Bernard, 2006).

According to Young (1990), cultural imperialism is when the dominant group universalizes their experience and culture and uses them as the norm by projecting their experience and culture as representative of all humanity. The dominant group reinforces their position by measuring other groups according to dominant norms, to construct individuals from racialized groups as inferior and deviant (Young, 1990; Este, 1999).

Most oppressed groups suffer systematic violence simply because they are subordinate in the social pecking order (Young, 1990). This violence is not always physical. It can be manifested through harassment, ridicule, and intimidation, which serves the purpose of stigmatizing individuals belonging to subordinate groups.

Experiencing the combined effects of social, institutional systemic oppression can have a damaging impact on the psyche of individuals from racialized groups. Two psychological conditions brought about by oppression are a sense of inferiority and internalized oppression (Mullaly, 2002). A sense of inferiority denotes an acceptance of a second-class citizenship status in society by a member of an oppressed group and a belief that he or she and all members of that group are inferior to the dominant group (Mullaly, 2002). Internalized oppression not only includes the belief that one's self and one's social group are inferior, but also encompasses the behaviors that are self-harming and contribute to one's own oppression (Mullaly, 2002).

Accordingly, oppression occurs through the systemic constraints on members of racialized groups and "is more effective in achieving its apparent function of maintaining the privileged position of the dominant group when both victims and perpetrators are unaware of the dynamics of oppression" (Mullaly, 2002: p. 40). Subsequently, this lack of awareness is critical in worker–client interactions for both the worker as well as individuals with racialized identities.

Existing research highlights a variety of approaches aimed at increasing awareness in order to address oppressive policies, practices, and interactions between people. These approaches have centered on static ideas of culture that do not take into account that cultural identities are socially and historically constructed as well as the subsequent oppression of individuals from racialized groups.

Implications for Social Work Practice

Social workers practice in diverse societies increases the likelihood of working with members from racialized groups. According to Fook's approach to social work, as cited in Payne (2005), social work "concentrates on rethinking ideas about practice through knowledge, power, discourse, language and narrative, and identity and difference" (p. 243). Fook proposes a critical reflective approach to practice which challenges domination in external structures, social relations, and personal constructions and recognizes multiple and diverse constructions (Payne, 2005).

Social workers should be aware of the historical legacy of race relations in order to explore the social construction of race as it relates to oppression of individuals with racialized identities. Graham et al. (2010) found that alternative perspectives of the helping relationship need to be considered to allow for service delivery that meets the unique needs of individuals. To recognize the need for specific consideration of the cultural values and issues for those from racialized groups, perspectives need to be developed beyond the frontline level to more inclusive service delivery based on a structural analysis (Graham et al., 2010).

According to Graham et al. (2010), "social services in North America and beyond need to incorporate these ideals into practice and policy by rejecting cultural assumptions and stereotypes and realizing that different cultures face adversity in different ways" (p. 345). Their findings confirm that there is no all-inclusive way to work with those with racialized identities and agency policies as well as practices needed to evolve in dynamic ways and meet the unique needs of individuals. Similarly, Fook (1993, 2002), Payne (2005), Dominelli (1997, 2003), and Hugman (2010) call for a reimagining and rethinking of social work that addresses racism and oppression at each level; the individual, institutional, and cultural. Furthermore, such practices are envisioned as leading to a more socially just world.

Social service agencies and their workers working with individuals from racialized groups need to be aware of the historical processes that were legally sanctioned and contributed to the social construction of race. Because race as a social construction is a product, not a cause of racism (Kobayashi and Johnson, 2007), the exclusionary measures that were once overt are now subtler, but still seek to preserve the privileges of the dominant group. Social workers need to address these dynamics related to racism, discrimination, and oppression at micro-, mezzo-, and macrolevels to begin to redress the historical and present-day influences of discrimination and oppression.

On a microlevel, social workers must first begin to understand how the social construction of race and its perpetuation of oppression affect members from racialized groups. They must also recognize the reality of racism and be willing to challenge both the theorists' and the therapists' world with regard to the presence of racism (Deters, 1997). Successfully and sensitively working with people from racialized groups requires that the social worker has explored and continues to explore how she or he has internalized rules of racial categorization and any of the accompanying stereotypes. They must ask themselves questions such as what are my own assumptions about groups with racialized identities? What are my own internalized rules? How do my own internalized rules affect the way I work with my clients? How does my own racial identity or privilege impact the therapeutic relationship? This personal work by social workers is critical to creating a non-pathologizing environment (Dominelli, 1997; Christensen, 1991; Deters, 1997).

Furthermore, social workers must challenge socially constructed ideas about clients from racialized groups in order to encourage clients' stories about the oppression they have faced. As such, social workers must seek out, learn, and encourage the further development of theories that reflect the developmental

process of more than socially constructed ideas of racialized groups.

At a mezzo-level, social workers should consider focusing their intervention activities on the local institutions and organizations to which individuals from racialized groups belong. This will assist in the education and sensitization of agencies and agencies, personnel about the needs of individuals from racialized groups. For example, Ben Carniol (1995 quoting Moreau and Leonard) highlights the various empowerment practices that social workers engage in, within their practice, as well as in their communities. Similarly, Bernard (2006) shares stories from Black social workers in their fight for change in policy and practice in Nova Scotia. Specifically, a social worker can advocate for changes in the organizational policies and procedures for more effective responses to those identified needs. Dhooper (2003) suggests this can be accomplished through communicating and negotiating effectively with local institutions and organizations by involving communities of racialized groups to bring about significant institutional changes.

Finally on a macrolevel, social workers need to study their society to reveal the ways that forms of oppression can create problems out of socially constructed differences. This will allow social workers to build better communities with more trustworthy services and institutions, as well as better relationships and understanding with clients (Dean, 2001). According to Dhooper (2003), racism continues to divide North America and individuals with racialized identities still often find themselves subject to marginalization, discriminatory, as well as oppressive practices. Therefore, social workers should participate in policy-level discussions and search for solutions aimed at challenging social constructions that contribute to unjust racial hierarchies and safeguarding the interests of individuals with racialized identities.

Conclusion

Since 'race' is not fixed by nature, understanding the process of racialization brings the historical, social, and political processes to the forefront of the institutionalization of racism. Without an understanding of the concept of racialization, these processes can continue to make 'race' the constructed idea that resources are allocated, policies as well as programs are developed, and training is delivered. Is it not time to move beyond this notion of ethnic sensitive and culturally competent practice? As Hugman (2010: p. 157) asserts, "… cultural competence is not just about acquiring a set of skills, knowledge and values, but it is also about rethinking and reimagining social work. Rethinking and reimagining social work *must* embrace an anti-racist and anti-oppressive perspective to challenge structural inequality at its core," a vision outlined by Dominelli in her first edition of *Anti-Racist Social Work* and repeated in the 1997 second edition. Social work in the twenty-first century is well-positioned to be the "bridge between social work in a racist society and social work in a non-racist one" (Dominelli, 1997: p. 167), but to get there we must move beyond notions of ethnic sensitivity and cultural competence.

See also: Critical Race Theory; Oppression; Race Identity; Race and Racism in the Twenty-First Century; Race and Racism, Geography of; Racism, Sociology of.

Bibliography

Baum, A., Crase, S., Crase, K., 2001. Influences on the decision to become or not become a foster parent, Families in Society. The Journal of Contemporary Human Services 82 (2), 202–213 viewed 1 February 2010, Research Library.

Bernard, W.T. (Ed.), 2006. Fighting for Change: Black Social Workers in Nova Scotia. Nimbus Publishing, Halifax, NS, Canada.

Bernard, W.T., Moriah, J., 2007. Cultural competency: an individual or institutional responsibility? Canadian Social Work Review 24 (1), 81–92.

Better, S., 2008. Institutional Racism, second ed. Rowman & Littlefield, Lanham.

Canadian Association of Social Workers, 2005. Social Work Code of Ethics. Canadian Association of Social Workers (CASW), Ottawa, ON, Canada.

Carniol, B., 1995. Case Critical: Challenging Social Services in Canada, third ed. Between the Lines, Toronto, ON, Canada.

Carniol, B., 2000. Case Critical: Social Services & Social Justice in Canada, fourth ed. Between the Lines, Toronto, ON, Canada.

Christensen, C., 1991. Social Work Education at the Crossroads, The Challenge of Diversity. Canadian Association of Schools of Social Work, Ottawa, ON, Canada.

Christian, M., 2000. Multiracial Identity. MacMillan Press Ltd, London.

Cross, T., Bazron, B., Dennis, K., Isaacs, M., 1989. Towards a Culturally Competent System of Care, Volume I. Georgetown University Child Development Center, CASSP Technical Assistance Center, Washington, DC.

Dalrymple, J., Burke, B., 1995. Anti-Oppressive Practice, Social Care and the Law. Open University Press, London.

Dean, R.G., 2001. The myth of cross-cultural competence. Families in Society 82 (6), 623–630.

Deters, K., 1997. Belonging nowhere and everywhere: multiracial identity development. Bulletin of the Menninger Clinic 61 (3), 368–384.

Devore, W., Schlesinger, El, 1981/1999. Ethnic Sensitive Social Work Practice, first ed. Allyn and Bacon, Boston, MA.

Dhooper, S., 2003. Social work response to the needs of biracial Americans. Journal of Ethnic & Cultural Diversity in Social Work 12 (4), 19–47.

Dominelli, L., 1997. Anti-racist Social Work, first ed. MacMillan Press, London.

Dominelli, L., 2003. Anti-oppressive Social Work Theory and Practice. Palgrave Macmillan, Hampshire, England.

Este, D., 1999. Social work and cultural competence. In: Lie, Gwat-Yong, Este, David (Eds.), Professional Social Service Delivery in a Multicultural World. Canadian Scholar's Press, Toronto, ON, Canada, pp. 27–45.

Fook, J., 1993. Radical Casework: A Theory of Practice. Allen and Unwin, London.

Fook, J., 2002. Social Work: Critical Theory and Practice. Sage, London.

Fong, Lillian G.W., Gibbs, Jewelle Taylor, 1995. Facilitating services to multicultural communities in a dominant culture setting: an organizational perspective. Administration in Social Work 19 (2), 1–24.

Garner, S., 2010. Racisms: An Introduction. Sage Publications, Thousand Oaks, CA.

Gillem, A., Cohn, L., Thorne, C., 2001. Black identity in biracial black/white people: a comparison of jacqueline who refuses to be exclusively black and adolphus who wishes he were. Cultural Diversity and Ethnic Minority Psychology 7 (2), 182–196.

Graham, J., Bradshaw, C., Trew, J., 2010. Cultural considerations for social service agencies working with Muslim clients. Social Work 55 (4), 337–346 viewed 15 December 2012.

Hugman, R., 2010. Understanding International Social Work: A Critical Analysis. Palgrve MacMillan, London.

Jackson, K., Samuels, G., 2011. Multiracial competence in social work: recommendations for culturally attuned work with multiracial people. Social Work 56 (3), 235–245 viewed 15 December 2012.

James, C., Este, D., Thomas-Bernard, W., Benajamin, A., Lloyd, B., Turner, T., 2010. Race & Well-being: The Lives, Hopes, and Activism of Africans Canadians. Fernwood Publishing, Halifax, NS, Winnipeg, MB, Canada.

Kirton, D., 2001. Love and money: payment, motivation and the fostering task. Child & Family Social Work 6 (3), 99–208.

Kobayashi, A., Johnson, G.F. (Eds.), 2007. Introduction. In: Johnson, G.F., Enomoto, R. (Eds.), Race, Racialization and Antiracism in Canada and beyond. University of Toronto Press, Toronto, ON, Canada, pp. 3–16.

Korbin, J., June 2002. Culture and child maltreatment: cultural competence and beyond. Child Abuse & Neglect 26 (6–7), 637–644 viewed 15 December 2012.
McGibbon, E.A., Etowa, J.B., 2009. Anti-racist Health Care Practice. Canadian Scholars' Press Inc., ON, Canada.
McPhatter, A., 1997. Cultural competence in child welfare: what is it? How do we achieve it? what happens without it? Child Welfare 76 (1), 255–278.
Mederos, F., Woldeguiorguis, I., 2003. Beyond cultural competence: what child protection managers need to know and do? Child Welfare 82 (2), 125–142.
Mullaly, B., 2002. Challenging Oppression: A Critical Social Work Approach. Oxford University Press, ON, Canada.
National Association of Social Workers, 2008. Code of Ethics. National Association of Social Workers (NASW), Washington, DC.
Parker, D., Song, M. (Eds.), 2001. Rethinking 'Mixed Race'. Pluto Press, London.
Payne, M., 2005. Modern Social Work Theory. Lyceum Books, Inc., Chicago, IL.
Pon, G., 2009. Cultural competency as new racism: an ontology of forgetting. Journal of Progressive Human Services 20 (1), 59–71.
Rew, L., Becker, H., Cookston, J., Khosropour, S., Martinez, S., 2003. Measuring cultural awareness in nursing students. Journal of Nursing Education 42 (6), 249–257.
Sakamoto, I., 2007. An anti-oppressive approach to cultural competence. Canadian Social Work Review 24 (1), 105–114.
Vickers, J., 2002. The Politics of "race": Canada, Australia and the United States. Golden Dog Press, ON, Canada.
Whaley, A., 2008. Cultural sensitivity and cultural competence: toward clarity of definitions in cross-cultural counselling and psychotherapy. Counselling Psychology Quarterly 21 (3), 215–222.
Williams, C.C., 2006. The epistemology of cultural competence. Families in Society 87 (2), 209–220 viewed 15 December 2012.
Yee, J.Y., Dumbrill, G., 2003. Whiteout: looking for race in Canadian social work practice. In: Al-Krenawi, A., Graham, J.R. (Eds.), Multicultural Social Work in Canada. Oxford University Press, Don Mills, ON, Canada, pp. 98–121.
Young, I.M., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.
Zamparo, J., Bernard, W.T., Seebaran, R., Blum, E., Razack, N., DuBois, M., 2000. Anti-racist Training and Materials Final Report. Canadian Association of Schools of Social Work, Ottawa, ON, Canada.

Social, The: History of the Concept

Jean Terrier, Institut für Politikwissenschaft, Westfälische Wilhelms-Universität Münster, Münster, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article describes the history of the concept of 'the social' as a substantivated adjective. Despite the growing use of the terms 'society' and 'social' in the eighteenth and nineteenth century, it is only at the beginning of the twentieth century that 'the social' started being used as an independent concept. In its simplest and most general meaning, 'the social' is the ensemble of phenomena involving a relation between human beings, and thus distinct from the individual and the natural. Emile Durkheim and Marcel Mauss used the expression in this sense. In the 1950s, Hannah Arendt gave a more limited meaning to the expression: she spoke of 'the rise of the social' to describe and criticize the development of modern industrialized societies; in these societies, economic activities (the production and consumption of goods) have become the primary concern of individuals, and the distribution of benefits the central activity of governments, resulting in a loss of political freedom. In the last decades of the twentieth century, the importance of the concept of 'the social' has steadily grown, thereby becoming one of the fundamental notions of the social sciences, central to many social theorists. 'The social' has thereby acquired a meaning of its own, independent of 'society': a more abstract term applicable to the macrolevel, as well as to the microlevel and with less connotations of homogeneity and boundedness, it is even often described, instead of 'society,' as the proper object of the social sciences and social philosophy.

As a concept, 'the social' results from the substantivation of the corresponding adjective. Consequently, the history of its usage is closely related to the emergence and diffusion of the adjective 'social' and of the noun 'society.' The oldest roots of 'society' and 'social' are the Latin *socius*, a companion, which ultimately derives from *sequi*, to follow (Latour, 2005: 6). In current usage 'society' stands primarily for (1) a voluntary association of individuals with common goals; (2) more commonly and more abstractly, the most general context of human interrelations – a framework with identifiable boundaries within which human beings interact with a higher degree of intensity, and in which these interactions typically solidify in habitual practices, enduring representations, and institutions.

'Society' and 'Social' in the Eighteenth Century

Historically speaking, the first of these two meanings is older, but the second one has become more prevalent since the emergence of the social sciences. In the mid- to late eighteenth century, both meanings were equally present: 'society' shifted in meaning between 'partnership for a particular purpose' or 'companionship,' and "the basic form of collective human existence" (Baker, 2001: 86, 95). Quantitatively, the use of the word 'society' also grew, arguably in a reflection of two phenomena (Honneth, 2000: 14): first, the increasing visibility and role of associations, clubs, and salons; second, the development of geographic mobility and long-distance interrelations, especially of the commercial type. At the same time, society, in particular for the political philosophers of the time, was as much a future goal as it was a present reality. One aspect of the project of the thinkers of the social contract was, for instance, to transform the arbitrary communities inherited from the past into conscious associations of individuals founded on an explicit agreement, thereby turning 'society' in the second sense into a 'society' in the first sense (Sewell, 2005: 322, 323). A new entity that Thomas Hobbes called 'commonwealth' and John Locke 'civil society' would thus be brought into existence. In this line of reasoning, social relations are implicitly assumed to be a malleable, transformable stuff that could be shaped with relative ease by the action of governments (Terrier, 2011: xxiv, xxv; Nisbet, 1976: 241–251). This conception, however, was resisted by other equally influential political philosophers, such as Montesquieu, David Hume, or Adam Smith, who envisaged society as the spontaneous outcome of human interactions in specific historical and geographic contexts (Sebastiani, 2013). While politics could give some orientation to social developments, it could not completely redefine society in the second sense, which was perceived by these authors as a historical, not an artificial entity.

These debates on 'society' during the eighteenth century and the growing use of the term were accompanied by a parallel rise of the adjective 'social.' The quantitative diffusion of this adjective in France is well documented (Mintzker, 2008). While 'society' and 'sociable' were of common usage already in the seventeenth century, 'social' became prevalent only in the following one. Its original meaning was close to that of sociable, especially in expressions such as 'social virtues' (Mintzker, 2008: 505): it was used to qualify a person as polite, respectful, and well capable of human interaction. Later, 'social' was more and more used as the adjective of society in the second sense, with the meaning of 'pertaining to an organized group of human beings.' It is in this sense that Jean-Jacques Rousseau used the term in his *Social Contract* (1761), the first book to bear the adjective in its title, closely followed by Baron d'Holbach and his *Social System* (1771). 'Social science' and 'sociology' were late eighteenth-century coinages – both by Emmanuel Sieyès, a French revolutionary who insisted that the major task of the French National Assembly in 1789 was the creation of a politically and socially unified nation (Sewell, 1994; Guilhaumou, 2006). These expressions were popularized during the nineteenth century by Saint-Simon, Auguste Comte, and John Stuart Mill (Baker, 1964, 1989).

'Society' and 'Social' in Early Social Science

During the nineteenth century, two developments were of particular importance. To begin with, due to the growing influence of the labor movement, 'social' acquired a further meaning, namely 'concerning the poorer classes of society' as well as 'concerning the relationship between workers and employers.' Expressions such as 'the social problem' (used by Mill in 1851 in his *Autobiography*, cf Burchfield, 1986: 315) or 'the social question' (documented in French as early as 1834 – Rey, 1992: 1958) can serve here as examples. Second, by the end of the nineteenth century, the definition of 'society' as a large, bounded, and organized group of human beings became the default one. This was the result of the diffusion of the discourse and practice of the social sciences. Against the traditional emphasis of political philosophers on ideas, consciousness, and will, the social sciences were organized around the assumption of a 'thickness' and 'opaqueness' of human interrelations: they thematized the existence of collective habits and representations determined by historical contexts which could not be changed at will. The history of the social sciences is marked by a debate on the correct identification of the most important contexts to consider when trying to explain the various features of collective life. During the nineteenth century, naturalistic approaches were especially influential: many authors highlighted climate or race as major determinants of human behavior. Starting around 1890, this naturalism was increasingly challenged by classical sociology and social anthropology (Turner, 2006: 135). A common conviction of the 'Founding Fathers' of the social sciences (among others Emile Durkheim, Georg Simmel, Franz Boas, Max Weber, Marcel Mauss) was that, in Durkheim's formulation (1895), social facts should be explained by social facts, as opposed to natural ones. On the other hand, these authors were in clear disagreement concerning the nature of their object. Durkheim took it to be 'society' itself, an entity with distinct boundaries and a clear internal unity. By contrast, Weber emphasized 'social action' and Simmel emphasized 'interaction' (*Wechselwirkung*).

It is precisely here that the history of 'the social' as a substantive really starts. Before this time 'the social' either was unavailable as a noun (for instance, in Germany or in France, cf Rey, 1992: 1958) or it had, in English, the limited, concrete meaning of a 'small gathering,' a 'sociable meeting' (Poovey, 2002: 127). Without turning it into a major concept, Emile Durkheim and Marcel Mauss both made use of 'social' as a substantive, especially in contrast to other realms such as 'the psychological' or 'the individual' (Mauss, 1969b: 154; Durkheim, 1975: 25; Mauss, 2004: 138). They used the term to refer to the ensemble of phenomena displaying the characteristics of social facts. In the definition of Durkheim, the criterion of a social fact is that it exerts an external constraint on individual thought and action, as in the case of habitual practices, moral obligations, or legal norms. Mauss had a broader understanding of the social. He relativized the Durkheimian dichotomy of the individual and the social and tried to show how all collective representations and rules are reinterpreted by individuals, and therefore progressively transformed by them. He also offered a theory of the way in which individuals, when they are associated in political movements, can collectively modify the shape of the society (Mauss, 2013). Mauss defined the social as "that which is common only to the men who live in a clearly defined and organized single society," and contrasted it to the 'psychological in its pure form' (*le psychique pur*), which is common to all human beings (Mauss, 1969a: 227).

In Germany, in the same years, 'social' also started being used as a substantive by some authors. Even though 'the social' was not central in their work, both Weber and Schmitt used this expression and proposed explicit definitions. In his 1905 essay on "objectivity in the social sciences," Weber mentioned "the relationships among persons" as the vague but characteristic focus of a particular perspective on human beings which he called the "standpoint of the 'social'" (*Gesichtspunkt des Sozialen*) (Weber, 2011: 67, translation modified). In his *Concept of the Political*, Schmitt (1996a) argued that the enmity between groups is at the core of all political phenomena. 'The political' in this sense needed to be distinguished from 'the social' (*das Gesellschaftliche*), which is the realm of interaction between human beings for specific purposes (Schmitt, 1996b: 79), as opposed to the overarching, more fundamental purpose of group survival and the defense of collective identity.

Especially relevant to a history of the concept of the social are the skeptical views of both Simmel and Weber on the notion of 'society.' The two authors warned against the danger of hypostasizing society, i.e., of envisaging it as an actor in its own right, distinct from its members. When sociologists view society as a person (a tendency which was not absent from the work of Durkheim), they tend to project upon it all the various characteristics of an individual human being, such as a hierarchical form of organization, a strong internal homogeneity, and clear-cut boundaries. Even though societies, according to Simmel and Weber, may well take such a form in given circumstances, these elements should not be considered as necessary features of social life. To avoid these connotations of homogeneity and boundedness, both authors advised to speak of 'socialization' (*Vergesellschaftung*) as a process, as opposed to 'society' as an entity. This skeptical judgment on the concept of society became a recurrent theme in twentieth-century social science. The social scientists who introduced, during the twentieth century, new concepts that should serve as alternatives to 'society' (including 'the social') often found inspiration in their early criticism.

After World War II

In the decades following World War II, the idea of society as a bounded, tightly integrated totality became prominent, especially in American sociology (cf Urry, 2000: 11). Correspondingly, there was little conceptual space for a theorization of 'the social' as broader than, and thus distinct from, 'society.' Nevertheless, the concept of 'the social' remained in use, with markedly different meanings, in particular among those who were critical of the then dominant sociological theories.

An especially prominent example of a systematic use of 'the social' as an independent concept is Hannah Arendt (*The Human Condition*, 1998/1958). Her notion of the social was very idiosyncratic and can be only understood against the background of her entire political philosophy. For her, 'the social' is

that which pertains to the "life process when it has been channeled into the public realm" (Arendt, 1998: 45). Following Aristotle, she made a distinction between the public space of the *agora* and the private space of the *oikos* (household), in which economic activities take place. According to her, in Greek times the reproduction of material life was a private concern with little political relevance: "whatever was 'economic,' related to the life of the individual and the survival of the species, was a nonpolitical, household affair by definition" (Arendt, 1998: 29). 'The social' appeared during modernity as a consequence of a blurring of the distinction between the public and the private. With the rise of industrial society and its social problems, the question of material survival became a topic of collective interest. What was previously a private concern and a private endeavor was increasingly made public, thus creating a "curiously hybrid realm where private interests assume public significance" (Arendt, 1998: 35). This transformation, for Arendt, had disastrous consequences. While human beings possess different visions of freedom and the good, they all have similar physiological needs. When the political question of good government is replaced with the social question of the fulfillment of needs, speech, especially political dialogue, is no longer required. Governments can treat individuals as if they were identical and dispense with any serious democratic consultation of citizens. Politics turns into a mere allocation of resources, an administration of things. The public sphere where diverse individuals can confront their own visions of the good vanishes, and freedom inevitably declines. Totalitarian regimes are the monstrous, but not improbable, outcome of this process (Baehr, 2010). The antidote to the 'rise of the social' is a restoration of politics by way of a reopening of the public sphere of action and speech. Concretely, Arendt judged positively some of the experiments of democracy from below, such as the workers' councils of the Hungarian Revolution in 1956 (Arendt, 1958). They were an example of a reinvention of the political, understood as a distinct sphere of activity in which equal human beings, leaving violence and petty interests aside, present themselves to others in their singularity and freely act in concert to create a 'common world' (Arendt, 1998: 28).

Arendt's legacy is a complex one. She has been criticized for her elitist views of politics as an encounter between highly reflexive individuals unaffected by material difficulties (Wolin, 1983). Others have seen her as the most important inspiration behind the move toward a theory of participatory politics, represented by various authors of the second half of the twentieth century, from Castoriadis to Habermas (Honneth, 2000: 52, 53). At the conceptual level, interpreters have underlined the very peculiar nature of Arendt's definition of 'the social' (Pitkin, 1998: 201), which could be said to be closer to our understanding of 'the economic.' This may be the reason why the concept of 'the social' in the Arendtian sense was not taken up by other authors, and why even she later distanced herself from it. Nevertheless, Arendt's concept of 'the social' also encompasses, as one of its aspects, 'welfare state policies.' In this more limited meaning, 'the social' has become a relatively common term, both in the general public and in academic circles (e.g., Donzelot, 1984; Rose, 1996).

Another important critic of the dominant concept of society in the postwar era is Robert Nisbet. In many ways, his position was close to Arendt's, but his vocabulary was clearly a different one: in *The Twilight of Authority*, one of his later writings, he pleaded for a "liberation of the idea of the social," which he understood as voluntary association and the self-administration of local collectivities, "from the political," which he equated with the bureaucratic, militarized state of the postwar era (Nisbet, 1976: 241). Nisbet targeted especially the notion, inherited from the Enlightenment, that the state could single-handedly reshape society. This notion, according to him, was at the core of the project of the welfare state: its goal was to establish economic equality and political homogeneity by way of leveling social policies, unitary school programs, and political surveillance. The welfare state, however, made little case of local specificities and inherited traditions; it also overlooked the capacity and desire of individuals and groups to decide and act for themselves. Thus, like Arendt, Nisbet feared that the welfare state and its bureaucratic apparatus could breed mass conformity and undermine political liberty. He advocated a 'rediscovery of the social' (Nisbet, 1976: 241): in order to recover liberty and diversity, power should be "distributed into as many hands as possible," and this could be achieved only by way of a recognition of the "functional autonomy of [the] social units" that compose society as a whole, such as kinship groups or professional associations (Nisbet, 1976: 237).

'The Social' in the Late Twentieth and Early Twenty-First Century

In the decades 1980–2010, it is with the different, more general meaning of 'any phenomenon involving an interaction between distinct entities' that 'the social' became a common concept. In most cases, the entities in question are human beings, but some authors, such as Bruno Latour (2005, borrowing from Gabriel Tarde, 1999), argue that an association between animals, or between humans and things, also belongs to the realm of the social. The concept of the social has been at the center of many debates, including three especially prominent ones: the first one concerning the relevance of the concept of society (1); the second one concerning 'the social' in social history (2); the third one concerning the object of social philosophy (3).

1. Some practitioners of the social sciences have argued that 'the social,' and not society, should be seen as the proper object of their disciplines. The reason for this proposed replacement is a dissatisfaction with the assumptions typically built into the concept of society. 'Society,' its critics argue, is commonly taken to be an economically and culturally self-sufficient and stable entity with clearly delineated boundaries, a centralized political apparatus, and a homogeneous culture. Even though each of these assumptions has been challenged since the beginning of the social sciences, the critique of 'society' has become more prevalent since the 1970s. This is partly a reflection of new social developments. First, the rise of political dissent and social movements in these years (from the civil rights movement to the mobilizations that shook Eastern Europe) undermined the notion of society as a homogeneous, consensual block: new approaches in the

social sciences emphasized power, conflict, and violence (e.g., Foucault, 1975; Mann 1986–93; Tilly, 1990), or alternatively movement, creativity, and communication (e.g., Touraine, 1984; Joas, 1996). Second, the acceleration of the globalization process inspired many sociologists and political scientists to theorize the existence of social relations across societal boundaries, thereby challenging the assumption of a strong boundedness of society and the common equation of the social with the national (Chernilo, 2007). Already Frederik Barth (1969) had made the observation of continuous exchange across boundaries: this forced him to envisage social boundaries as porous and in permanent need of reinforcement and maintenance. With the increasing challenge to the nation-state, especially in Europe due to the Europeanization process, some authors started to advocate a 'cosmopolitan' or 'transnational' turn in the social sciences (cf Beck and Sznaider, 2006). These developments have created a fertile ground for conceptual change. Many authors have suggested new, alternative terms to avoid the problematic connotations of 'society' (cf Wagner, 1999; Turner, 2006). Without dropping 'society,' Bourdieu (1985, 1989) spoke of the 'the social space,' the 'social world,' and – occasionally – 'the social' (cf Calhoun, 2005). Mann (1986: 1) proposes an "organized power network." Some authors went further and emphasized 'the social' over 'society.' Raymond Boudon, the representative, in France, of rational action theory, published an introduction to sociology under the title *The Logic of the Social* (1979; English translation: 1981). The avoidance of the concept of 'society,' which Boudon, like many other methodological individualists, saw as the fetish of deterministic, holistic sociology, is revealing. At a different end of the sociological spectrum, John Urry (2000) argues that sociology should move beyond 'society' as the subject matter of sociology and focus instead on 'the social,' which he conceptualizes as the movement of persons, things, and signs within and across boundaries. In a similar vein, Latour also gives to the concept of 'the social' an especially prominent role. In *Re-Assembling the Social* (2005), he argues against the assumption that 'social causes' – in the form of ideologies, objective interests, or cultural representations – determine the action of individuals. He contrasts 'the sociology of the social' with his own 'sociology of associations.' The first one assumes that the action is caused by factors that are external to it, and views the social actor merely as a transmission belt of extant structures. Latour, following in the track of ethnomethodology, envisages the social as a permanent flux: competent actors constantly mobilize their intelligence to act in a complex environment: action is the result of a local process of negotiation with other individuals and things. In order to account for the enduring character of certain social phenomena, Latour suggests that face-to-face interactions (which Latour calls 'the social 3') are constrained by the presence of 'the social 2' – which is the material traces left by former actions, such as buildings, roads, words, and machines. More than society, it is these elements, as well as their associations in 'networks' and 'collectives,' that should be, according to Latour, at the heart of the sociological discipline.

2. A related line of critique of 'social explanations' and of 'society as a reified totality' (Sewell, 2005: 322) can be found in the work of critical historians such as Patrick Joyce (2002, 2010) and Michael Cabrera (2004, 2005). According to these authors, the dominant form of history-writing after 1945, social history, relied too heavily on social structure (especially the relations between classes) in its attempts at explaining social events: there had been a tendency to assume that a given configuration of the social automatically yields given historical events (for instance, the rise of the bourgeoisie causes a bourgeois revolution): the question of agency is discarded, and culture is perceived as a mere reflection of the social structure. Even after the 'cultural turn' of the 1970s and 1980s, many historians overemphasized the stability and homogeneity of collective representations and failed to include a consideration of the process of cultural production and transformation. Joyce and Cabrera advocate a serious engagement with the concepts of 'culture,' 'society,' and 'the social,' in order to avoid any reification of the realities they are taken to denote. Finding inspiration in the work of Michel Foucault and Charles Taylor, Cabrera suggests that historians should focus, not on society and the social, but on 'discourse' and the 'social imaginary.' 'Discourses' are impersonal, collective sets of representations which entail 'concepts, categories, and principles' and more generally 'meanings' with which to 'apprehend and conceptualize reality' (Cabrera, 2004: 22, 23). They play the role of a 'repertory' for individual and collective thought and action, thereby limiting their possibilities. At the same time, social actors actively select and combine elements from the discursive frame, so that they are not determined by it. Even though the abandonment of any notion of social causation and the rehabilitation of agency represents a shift away from classical social history, Joyce and Cabrera can be said to remain located within the paradigm of the social: as 'things' *shared* between various individuals, discourses and imaginaries pertain to 'the social,' in the sense of 'the complex interdependence' that characterizes 'human interrelationships' (Sewell, 2005: 365).

3. Among social philosophers, some authors have indicated 'the social' as that which should be the main focus of attention. This is the case of Axel Honneth. In his programmatic delineation of the identity of the discipline (2000; English translation: 1996; see also Honneth, 1995), he discusses a range of authors from Rousseau to Arendt, Habermas, and Castoriadis. He argues that one of the central themes of social philosophy has been the identification of some 'pathology' or dysfunction in human collectivities that prevents individuals from pursuing their own vision of the good. The 'healthy' social, by contrast, is a form of "normality tailored to conditions that enable human self-realization" (Honneth, 1996: 36). These 'conditions of self-realization' can be variously defined: historically speaking, philosophical anthropology has formulated different answers to this question, from Rousseau's emphasis on independence to Arendt's defense of collective speech and action (cf Jörke, 2005). Honneth himself, finding inspiration in Hegel, has developed a complex anthropology of recognition. In an argumentation that bears some resemblance with Arendt's, he argues that any society that structurally inhibits the recognition of the equal value of persons is bound to become bereft with various pathologies, including permanent individual suffering. In this line of reasoning, it is not society as a whole

which is at the center of social philosophy, but 'the social,' understood as the conditions under which individuals can live a decent life, in association with others.

Conclusion

Looking back, two developments of 'the social' can be highlighted in particular. First, 'the social' has acquired a meaning of its own, thereby becoming more independent of 'society': a more abstract term applicable to the macrolevel, as well as to the microlevel, and with less connotations of homogeneity and boundedness, it is even often described as an alternative to 'society.' If we take into account computing and information technology ('social networks,' the 'social Web'), it appears that 'the social' has recovered some of the connotations it had at the very beginning: connotations of companionship, collaboration, exchange, and freedom. Second, the importance of this concept has steadily grown, so that it has now become one of the central notions of social science. However, this did not happen without resistance. 'The social,' an 'exceptionally complex or polysemic concept' (Sewell, 2005: 324), has been at the heart of heated debates in the social sciences, with positions ranging from downright condemnation to very positive appreciation. Historically, the rise of 'the social' should be understood against the background of the impact of antiessentialist and antirealist approaches in the social sciences (cf Elder-Vass, 2012; Hacking, 1999). In the spirit of Simmel and Weber, these approaches describe social phenomena not as objective 'things' but as outcomes of social constitution processes; moreover, a widespread social belief in these entities is crucial to their emergence and stability. Thus researchers have introduced new concepts which indicate a shift of focus from clearly circumscribed things (such as the state, religion, or 'society') to more abstractly defined general phenomena (such as the political, the religious/the sacred, or 'the social').

See also: Community and Society: History of the Concepts; Individuals in Society: History of the Concept; Sociology, History of.

Bibliography

Arendt, Hannah, February 1958. Totalitarian imperialism: reflections on the Hungarian Revolution. The Journal of Politics 20 (1), 5–43.

Arendt, Hannah, 1998/1958. The Human Condition. The University of Chicago Press, Chicago.

Baehr, Peter, 2010. Hannah Arendt, Totalitarianism, and the Social Sciences. Stanford University Press, Stanford.

Baker, Keith Michael, 1964. The early history of the term social science. Annals of Science 20 (3), 211–226.

Baker, Keith Michael, 1989. Closing the French Revolution: Saint-Simon and Comte. In: Furet, François, Ozouf, Mona (Eds.), The French Revolution and the Creation of Modern Political Culture, Transformation of Political Culture, 1789–1848, vol. 3. Pergamon Press, Oxford.

Baker, Keith Michael, 2001. Enlightenment and the institution of society: notes for a conceptual history. In: Kaviraj, Sudipta, Khilnani, Sunil (Eds.), Civil Society. History and Possibilities. Cambridge University Press, Cambridge.

Barth, Frederik (Ed.), 1969. Ethnic Groups and Boundaries: The Social Organization of Cultural Difference. Universitetsforlaget, Oslo.

Beck, Ulrich, Sznaider, Natan, 2006. Unpacking cosmopolitanism for the social sciences: a research agenda. The British Journal of Sociology 57 (1), 1–23.

Boudon, Raymond, 1979. La logique du social. Hachette, Paris.

Boudon, Raymond, 1981. The Logic of Social Action: An Introduction to Sociological Analysis. Routledge & Kegan Paul, Boston.

Bourdieu, Pierre, November 1985. The social space and the genesis of groups. Theory and Society 14 (6), 723–744.

Bourdieu, Pierre, Spring 1989. Social space and symbolic power. Sociological theory 7 (1), 14–25.

Burchfield, R.W. (Ed.), 1986. A Supplement to the Oxford English Dictionary. Se-Z, vol. IV. Clarendon Press, Oxford.

Cabrera, Miguel A., 2004. Postsocial History: An Introduction. Lexington Books, Oxford.

Cabrera, Miguel A., 2005. The crisis of the social and post-social history. The European Legacy: Toward New Paradigms 10 (6), 611–620.

Calhoun, Craig, 2005. Centralité du social et possibilité de la politique. In: Durand, Pascal, Winkin, Yves (Eds.), Le Symbolique et le Social. La réception internationale de la pensée de Pierre Bourdieu. Liège: Editions de l'Université de Liège.

Chernilo, Daniel, 2007. A Social Theory of the Nation-State: The Political Forms of Modernity Beyond Methodological Nationalism. Routledge, London.

Donzelot, Jacques, 1984. L'invention du social. Essai sur le déclin des passions politiques. Fayard, Paris.

Durkheim, Emile, 1895. Les règles de la méthode sociologique. Alcan, Paris.

Durkheim, Emile, 1975/1913. Le problème religieux et la dualité de la nature humaine. In: Textes. Religion, morale, anomie, vol. 2. Éditions de Minuit, Paris.

Elder-Vass, Dave, 2012. The Reality of Social Construction. Cambridge University Press, Cambridge.

Foucault, Michel, 1975. Surveiller et punir. Naissance de la prison. Gallimard, Paris.

Guilhaumou, Jacques, 2006. Sieyès et le non-dit de la sociologie: du mot à la chose. Sciences Humaines – Revue d'histoire des sciences humaines 15 (2), 117–134.

Hacking, Ian, 1999. The Social Construction of What? Harvard University Press, Cambridge, MA.

Honneth, Axel, 1995. The Fragmented World of the Social: Essays in Social and Political Philosophy. State University of New York Press, Albany.

Honneth, Axel, 1996. Pathologies of the social. The past and present of social philosophy. In: Rasmussen, David M. (Ed.), The Handbook of Critical Theory. Blackwell, Oxford, pp. 369–399.

Honneth, Axel, 2000. Pathologien des Sozialen. Tradition und Akualität der Sozialphilosophie. In: Das Andere der Gerechtigkeit. Sozialphilosophisce Aufsätze. Suhrkamp, Frankfurt am.

Joas, Hans, 1996. The Creativity of Action. Polity Press, Cambridge.

Joyce, Patrick, 2010. What is the social in social history? Past and Present 206 (1), 213–248.

Joyce, Patrick (Ed.), 2002. The Social in Question. New Bearings in History and the Social Sciences. Routledge, London.

Jörke, Dirk, 2005. Politische Anthropologie: Eine Einführung. VS–Verlag für Sozialwissenschaften, Wiesbaden.

Latour, Bruno, 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, New York–Oxford.

Mann, Michael, 1986. The Sources of Social Power. In: A History of Power from the Beginning to AD 1760, vol. 1. Cambridge University Press, Cambridge.

Mann, Michael, 1993. The Sources of Social Power. In: The Rise of Classes and the Nation-States, 1760–1914, vol. 2. Cambridge University Press, Cambridge.

Mauss, Marcel, 1969a/1908. L'art et le mythe d'après M. Wundt. In: Œuvres. Représentations collectives et diversité des civilisations, vol. 2. Éditions de Minuit, Paris.

Mauss, Marcel, 1969b/1927. Divisions et proportions des divisions de la sociologie. In: Œuvres. Cohésion sociale et divisions de la sociologie, vol. 3. Éditions de Minuit, Paris.

Mauss, Marcel, 2004. Fait social et formation du caractère. Sociologie et sociétés 36 (2), 135–140.

Mauss, Marcel, 2013. La nation. Presses universitaires de France, Paris.

Mintzker, Yair, 2008. 'A word newly introduced into language:' the appearance and spread of 'social' in french enlightened thought, 1745–1765. History of European Ideas 34, 500–513.

Nisbet, Robert, 1976. Twilight of Authority. Heinemann, London.

Pitkin, Hanna F., 1998. The Attack of the Blob: Hannah Arendt's Concept of the Social. University Of Chicago Press, Chicago.

Poovey, Mary, 2002. The liberal civil subject and the social in eighteenth-century british moral philosophy. Public Culture 14 (1), 125–145.

Rey, Alain (Ed.), 1992. Dictionnaire historique de la langue française. M-Z, vol. 2. Le Robert, Paris.

Rose, Nikolas, August 1996. The death of the social? Re-figuring the territory of government. Economy and Society 25 (3), 327–356.

Schmitt, Carl, 1996a/1927. The Concept of the Political. Chicago University Press, Chicago.
Schmitt, Carl, 1996b/1931. Der Hüter der Verfassung. Duncker & Humblot, Berlin.
Sebastiani, Silvia, 2013. The Scottish Enlightenment. Race, Gender, and the Limits of Progress. Palgrave-Macmillan, London.
Sewell, William H., 1994. A Rhetoric of Bourgeois Revolution: The Abbé Sieyes and 'What Is the Third Estate?' Duke University Press, Durham.
Sewell, William H., 2005. Refiguring the 'social' in social science: an interpretivist manifesto. In: Logics of History: Social Theory and Social Transformation. University of Chicago Press, Chicago.
Tarde, Gabriel, 1999. Oeuvres. In: Monadologie et sociologie, vol. I. Synthélabo – Les empêcheurs de penser en rond, Paris.
Terrier, Jean, 2011. Visions of the Social. Society as a Political Project in France, 1750–1950 (International Studies in Sociology and Social Anthropology). Brill Academic Publishers, Leiden–Boston, MA.
Tilly, Charles, 1990. Coercion, Capital, and European States. AD 990–1990. Basil Blackwell, Cambridge, MA.
Touraine, Alain, 1984. Le retour de l'acteur. Essai de sociologie. Fayard, Paris.
Turner, Brian S., 2006. Classical sociology and cosmopolitanism: a critical defence of the social. The British Journal of Sociology 57 (1), 133–151.
Urry, John, 2000. Sociology beyond Societies: Mobilities for the Twenty-First Century. Routledge, London.
Wagner, Peter, 1999. 'An entirely new object of consciousness, of volition, of thought.' The birth and (almost) passing away of society as a scientific object. In: Daston,-Lorraine (Ed.), Biographies of Scientific Objects. Chicago University Press, Chicago.
Weber, Max, 2011/1905. 'Objectivity' in social science and social policy. In: Shils, Edward A., Finch, Henry A. (Eds.), Methodology of Social Sciences. Transaction Publishers, New Brunswick, NJ.
Wolin, Sheldon S., 1983. Hannah Arendt: Democracy and the Political Salmagundi 60, 3–19.

Socialism

Paul Thomas, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the literature of the Social Sciences, the word 'socialism' is often qualified with, or characterized by an adjective – 'utopian,' 'scientific,' 'democratic,' 'parliamentary,' 'Fabian,' 'evolutionary,' 'revolutionary,' 'Christian,' 'Guild,' 'actually existing,' and 'market.' (Additionally, one can read of socialist internationalism, realism, feminism, etc.) This article investigates the common ground or substratum these various characterizations of 'socialism' (or 'socialist') share by tracing out the historical emergence and provenance of the term.

Leszek Kolakowski's survey (1978) calls attention to "the surprising diversity of views expressed by Marxists in regard to Marx's so-called historical determinism." However, what is 'surprising' is not this diversity – which becomes much wider if we examine not Marxism in particular, but socialism in general. The surprising thing is Kolakowski's pontifical belief that faced with it he could "schematize with precision the trends of twentieth century Marxism" (1978, Vol. I, pp. 6–7). His failure to attain such precision serves as a salutary warning to anyone attempting to characterize the broader and much more 'diverse' terrain of socialism, of which Marxism – nineteenth- and twentieth-century Marxism – is itself but a 'main current.' In what follows, which will (in Wittgenstein's words) 'strive *not* after exactness, but after a synoptic view,' attention will be paid to Cole's admonition: that "the most that can be attempted is the discovery of some central core of meaning, present with varying additions in all or most of the manifold uses of the word (socialism)" (Cole, 1967: vol. I, p. 1). This entails sensitivity to the shifts in meaning that socialism has undergone since its inception; only beneath these shifts can a substratum of meaning emerge.

Capitalism, Socialism, and Democracy

The word 'socialism' first appeared in Robert Owen's *Co-operative Magazine* in 1827, *le socialisme* in the Saint-Simonian journal *Le Globe* in 1832. It meant a projected alternative sociopolitical system – this was still offered as a definition of 'socialism' in the 1933 OED – not a political movement or demand. Early socialist doctrine contained little about the proletariat, the class system, labor, or revolution. While John Stuart Mill's *Principles of Political Economy* accordingly distinguished socialism and communism as theories, not political movements, Marx and Engels's *Manifesto of the Communist Party* of the same year (1848) identified communism as a movement that would render socialism, so defined, beside the point because it was merely a theory – utopian, that is, because it was impracticable. It was communism, not socialism, that carried with it the idea of revolutionary struggle and human agency; a new and better society could not be wished or legislated into existence by benevolent doctrinaires, but had to await the advent of a politically conscious, class-based labor movement. This was a key shift in the coordinates, not just of communism, but of socialism. It too had now to be not just thinkable, but feasible. After the failure of the 1848 revolutions, communism ceased for many years to be practical politics, yet some time was to pass before Marx and Engels consented to having their cause described as "socialist" (Lichtheim, 1969: p. 209). They did so because socialism had become not just a doctrine but a movement. Thereafter, "Marxism created that distinctively German socialism which was soon to assume an ideological dominance over most of the continent, driving the older forms of socialism before it as chaff before the wind" (Cole, 1967: vol. I, p. 222).

The second key shift in the coordinates of European socialism occurred in 1917, when, against all the odds, communism in its Leninist form arose as a clear alternative to democratic socialism – the same democratic socialism that had itself, against all previous odds, survived World War One. Both authoritarian communism and democratic socialism laid claim to a lineage deriving from Marx; but they had little else in common. Twentieth-century communists showed no reluctance to clothe their particular doctrines in the broader mantle of socialism. The locutions, 'Union of Soviet Socialist Republics,' 'socialism in one country,' 'socialist realism,' 'socialist internationalism' (in its twentieth-century Soviet version, which meant subservience to the foreign policy of the USSR), the fact that after World War Two the East German Communist Party was officially called the 'Socialist Unity Party' – all these are instances of synecdoche, if not the sheerest verbal trickery. We should not allow them to addle us into thinking, languidly, that all twentieth-century communists were socialists, whereas not all socialists were communists. There were in the twentieth century not just noncommunist but also anticommunist socialists aplenty, socialists actively and vociferously opposed to communism, from the very moment when such opposition first became possible, at the Berne International in February 1919, when adherence to democracy was reasserted as a prerequisite for socialism (Cole, 1967: vol. IV, pp. 292–299). The socialism that survived World War One to become part of the European political landscape had had as its prerequisite not democracy but capitalism. Socialists had long been in the forefront of the struggle for democracy. (Indeed, socialism was weakest in those countries, the United States and the United Kingdom, where it was a less powerful force in the struggle for

democracy than it was on the European continent.) In Cole's words, Liebknecht, Bebel, and Kautsky in Germany, Guesde in France, indeed, most socialist stalwarts in Europe before World War One "felt that they had no *right* to make the revolution without the backing, or at least the assent, of the people. They thought of the proletariat as being, or as in the process of becoming, the majority of the people; and they envisaged the mass conversion of the proletariat to the socialist cause as a necessary preliminary to the revolution" (Cole, 1967: vol. III, pp. 946–947). This democratic perspective was not necessarily reformist or purely parliamentary; democratization was also meant to encompass work relations and trade union structures, and could scarcely have avoided doing so. What the Second International socialists rightly feared was war; the solitary success of Russian Bolshevism in 1917 was as little expected or foreseen as its eventual collapse was to be in 1989.

Marx, Marxism, and Socialism

Marx casts a longer shadow over the history of socialism than anyone else – not because every socialist agreed (or was even very familiar) with what he said, but because, all the same, his presence proved impossible to ignore. It is in this sense that Marxism can be called the *lingua franca* of socialism (Cole, 1967: vol. I, p. 379). Without Marx, we would still have had revolution as a word and as a concept, as the French and the Americans had put revolutionary change on the political agenda during the eighteenth century, and we would have had capitalism, socialism, communism, and (thanks to Babeuf) the proletariat too. The celebrated opening line of the *Manifesto of the Communist Party* was designed to call the reader's attention to 'the specter haunting Europe,' to invoke, not invent, communism. It is no doubt easier to imagine a world without Marx than a world without revolution, capitalism, communism, and socialism. But in the world we actually inhabit, all these still have to be seen through Marx. He may not have coined these terms, but he decisively set his seal on them. They still cannot properly be discussed without bringing him in. Marx contributed enormously to the vocabulary of socialism (class, class struggle, class war, class consciousness, ideology, alienation, the commodity, forces, relations, and modes of production, exploitation, contradiction, and crisis). Even though "socialist and communist movements (took) up some aspects of Marx's theory and discarded others, even this selectivity (was) carried out in Marx's name" (Gilbert, 1981: p. 270).

Yet much that was carried out 'in Marx's name' was attained without much knowledge of what he had written. The revisionism debate (at the turn of the twentieth century) that almost tore apart German Social Democracy and the Second International was about Marxist 'orthodoxy,' but was conducted in an astounding ignorance of Marx's writings. Although practicing politicians commonly regard themselves as having more immediately pressing concerns than questions of textual fidelity, these same questions were never ignored among Marxists – although the theoretical corpus to which socialist stalwarts appealed was not necessarily made up of Marx's writings. The problem was not that a canon of basic texts failed to emerge. (To the contrary, textual fidelity was adhered to tirelessly – and, at times, tiresomely.) It was that the canon as it developed was not so much derived from what Marx had written as it was constructed around what little of Marx was available in published form. Historical materialism – it still amazes people when they learn that this term was not Marx's own – had its basic texts long before the first appearance in any language (during the 1930s) of so basic a text as Marx and Engels's *The German Ideology*. These texts included Franz Mehring's *On Historical Materialism* (1893), G.V. Plekhanov's *The Development of the Monist View of History* (1895), Antonio Labriola's *Essays on the Materialist Interpretation of History* (1896), and Karl Kautsky's *The Materialist Interpretation of History* (1927). Even though Marx in some sense inspired these books and others like them, what he had to say eventually needed to be disentangled from their arguments. It could in no way be directly encountered through them or explained by them (Thomas, 1991: pp. 32–33).

The Varieties of Socialist Experience

Even if Marx was more known about than known, the fact remains that by the end of the nineteenth century a politically oriented, class-conscious European socialist movement had emerged in his name; and by this time the word socialist meant what it still means today, a social and political doctrine and a political movement or system. In view of this latitude of meaning, it is scarcely surprising that 'socialism' has come down to us in social science literature as a noun qualified by an adjective – as in 'utopian socialism,' 'scientific socialism,' 'state socialism,' 'revolutionary socialism,' 'evolutionary socialism,' or (until recently) as 'actually existing socialism,' or again (today) as 'market socialism.' Similarly, the adjective 'socialist' can be employed as a pendant qualifying or characterizing an abstract noun, as in 'socialist internationalism,' 'socialist economics,' 'socialist realism,' or 'socialist feminism.'

The above-listed collocations – the inventory is illustrative, not exhaustive – are not just convenient subdivisions, but serial attempts at spot-welding that can readily be arrayed in historical sequence. 'Utopian socialism,' as we have seen, dates from the 1830s and 1840s, 'Fabian' or 'evolutionary socialism' dates from the late nineteenth and early twentieth centuries, 'Guild socialism' dates from the period between the two World Wars, and 'actually existing socialism' dates from the latter days of Soviet communism. 'Socialist internationalism' peaked, then bottomed out dramatically, in the period before 1914; 'socialist realism' (like 'socialism in one country') belongs to the Stalinist period in the Soviet Union; 'socialist feminism' to Western scholarship of a later date. To array these composite terms historically in order to arrive at the meaning of their least common denominator, socialism, is not to pigeonhole them arbitrarily, nor yet to proceed reductively, if it is borne in mind throughout that each instance of spot-welding affected the coordinates of its successors.

More specifically, anarchism, syndicalism, and their offshoots (whether revolutionary or not) are kinds of socialism that were defined and elaborated against state socialism (Thomas, 1980); Engels in 1875 contrasted scientific with utopian socialism (even if the latter was largely a spent force,

the former was to enjoy some shelf-life well into the twentieth century); Lenin defined Bolshevism, then communism as he understood the term, against the parliamentary, reformist socialism of the pre-1914 German SPD (and by extension the Second International); this reformist socialism had itself been elaborated in contradistinction to the revolutionary socialism that appeared to have had its come-uppance at the time of the Paris Commune (1871). Lenin's successors proceeded to counterpose socialism in one country against the socialist internationalism the First and Second Internationals had espoused. Overall, these developments do not add up to a tidy, formulaic picture. Yet there is a subtle patterning in the carpet. Positions and tendencies throughout the history of socialism were debated, articulated, and defined in relation to other, pre-existing socialist positions and tendencies.

See also: Field Theory; Marx, Karl (1818–83); Marxism–Leninism: The Ideology of Twentieth-Century Communism; Social Democracy; Socialism: Historical Aspects; Welfare Policy Comparisons.

Bibliography

Bottomore, T.B., 1993. Socialism. In: Outhwaite, W., Bottomore, T.B. (Eds.), The Blackwell Dictionary of Twentieth Century Political Thought. Blackwell, Oxford, pp. 617–619.

Braunthal, J., 1966. In: Collins, H., Mitchell, K. (Eds.), History of the International 1864–1914, 2 vols. Nelson, London.

Cole, G.D.H., 1967. A History of Socialist Thought, 5 vols. Macmillan, London.

Gilbert, A., 1981. Marx's Politics. Communists and Citizens. Rutgers University Press, New Brunswick, NJ.

Kolakowski, L., 1978. P.S. Falla tr., Main Currents in Marxism, 3 vols. Clarendon Press, Oxford, UK.

Landauer, C., 1959. European Socialism: A History of Ideas and Movements from the Industrial Revolution to Hitler's Seizure of Power, 2 vols. University of California Press, Berkeley, CA.

Lichtheim, G., 1961. Marxism: An Historical and Critical Study. Praeger, New York.

Lichtheim, G., 1969. The Origins of Socialism. Praeger, New York.

Lichtheim, G., 1971. A Short History of Socialism. Praeger, New York.

Schorske, C.E., 1955. German Social Democracy 1905–1917. The Development of the Great Schism. Harvard University Press, Cambridge, MA.

Thomas, P., 1980. Karl Marx and the Anarchists. Routledge and Kegan Paul, London.

Thomas, P., 1991. Critical reception: Marx then and now. In: Carver, T. (Ed.), The Cambridge Companion to Marx. Cambridge University Press, Cambridge, UK.

Socialism: Historical Aspects

Logie Barrow, Universitaet Bremen, Bremen, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Andreucci, volume 21, pp. 14488–14492, © 2001, Elsevier Ltd.

Abstract

In the nineteenth century, socialism scared governments. By around 1900, it was the main anticapitalist current worldwide. It aimed at simultaneously humanizing and replacing capitalism. Confusions implicit in that simultaneity were aggravated by an illusion that socialists could prevent wars. That illusion exploded in 1914. Russia's Bolsheviks alone proved able to make peace, but this was founded on another illusion that socialists in developed countries would rescue them. Their resulting desperation nurtured their least democratic aspects, helping to polarize the world ideologically, until the Soviet Union collapsed in 1991. This collapse boosted doubts about further humanization, let alone the replacement of capitalism by socialism.

Forerunners and Pioneers

The word 'socialism' emerged in Britain around 1830 and in France during the 1830s, denoting a society that would not have individualism and private profit as its motive power: what it would remains harder to define. In Britain, it was first publicized by the humbly born Welshman, Robert Owen (1771–1858), who had made his fortune as an industrialist in Manchester, the 'shock city' of industrialization, before moving to New Lanark, near Glasgow. Here, he had made his name as a paternalist employer. He boosted profits by cutting the working day, thus refuting those economists who saw them as dependent on the 'last' hour of a very long one. He subjected employees to psychological pressures, more frequently positive than those applied in rival firms. He also housed them to high standards and, for him even more vital, offered them space for recreation, plus free schooling for all children and adults. Pedagogic methods were strictly Enlightenment, aiming to nurture Reason.

Mid-eighteenth-century Britain was at the forefront, both of what is nowadays called the industrial revolution and of 'laissez-faire,' i.e., free-market economics. Some writers, such as the multidimensional autodidact William Blake (1757–1827), denounced Northern England's 'dark Satanic mills,' and linked them to ruthless free-market contempt for the London poor. Similar unease grew in France, not only about being outpaced but also because similar processes were spreading there too, as also in the neighboring Prince-Bishopric of Liège.

But where were coherent alternatives? Equalization of incomes among early Christians was sometimes gestured to. Totally unknown, by contrast, was the practice of collective ownership among, say, sixth-century followers of the Iranian reformer Mazdak. European hindsight may recognize partial foreshadowings of socialism among the radical wing of the Hussites in fifteenth-century Bohemia, by the Anabaptists in 1530s' Munster and more coherently by the Diggers in England around 1650. These three movements were all more or less against private property, whatever their religious dissimilarites.

Fantasy was many centuries old about an ideal society which, in 1516, the Greek-speaking Sir Thomas More had playfully named *Utopia*, i.e., both 'good place' and 'nowhere.' It remained sufficiently lively in eighteenth-century Britain for satirical inversion. The Grand Jury of Middlesex testified clumsily to this by ordering the ceremonial burning of Bernard Mandeville's 1723 *Fable of the Bees*, thereby publicizing his claim that morality and what we would call capitalism were incompatible.

Nineteenth-century pioneers of 'socialism' defined themselves in relation to earlier writers and movements and, in particular, as furthering Enlightenment emphases on Reason against superstition. Some recent scholars (e.g., Margaret Jacob, Jonathan Israel) have distinguished 'Radical' and 'Moderate' Enlightenments. The latter, as typified by Voltaire's cultivation of enlightened despots such as Catherine of Russia or Frederick of Prussia, sought reform from above. Its 'radical' and often older (*c*.1700) rival sought much wider support, was more outspoken, was more censored (at least in the Netherlands), and is therefore, partly as a result, less visible. The 1797 'Conspiracy of Equals' was one late heir of that tendency: Francois Babeuf (who renamed himself 'Gracchus,' after two ancient Roman radicals) was guillotined along with most other members for conspiring against France's individualist Republic. The chasm between Voltaire and Babeuf, between from-above and (would be) from-below, can symbolize some of the key problems of definition, agency (i.e., who is seen as bringing what to whom), strategy, organization, and thus of democracy. Not least, if you feel driven to conspire to seize power before handing it over to the oppressed masses, what if they do not grasp your generosity?

For Charles Fourier (1772–1837), the Babouvistes' blunder had been to act politically at all. Fourier, a merchant's son who had grown to loathe everything mercantile, insisted on tackling what he called the 'social question' directly. He posited 12 'passions.' 'Passions,' David Hume had argued (1739), should have 'reason' as merely their 'slave.' But, from 1789 to 1790, conservatives had seen them as grotesquely freed by revolution: only Christianity might cage them again. Yet Fourier daringly made them into the motive power of a radically new society. Via their 'gravitation,' he derived – with, for him, the cool logic of a Newtonian scientist – 810 personality types or rather, given two 'sexes,' 1620 types: ideally, therefore, that was the number of members for each of his communities, or 'phalanxes.' These

should build 'phalansteries' as soon as possible. He thus sought to transpose utopia from libraries into practice. Two vital requirements were (1) to end marital rigidity and gender division of labor, thus allowing women as well as men to unfold themselves and (2) to use the latest technology for flexing all such divisions, including those between work and education. Respectability drove most followers to downplay the gender aspect. He had also embarrassed them with his aim of reforming Nature: his 'seas of lemonade' are still flowing. Yet he was not the first such Natural reformer: the Digger leader Gerrard Winstanley and, some decades earlier, the much-tortured Tommaso Campanella had preceded him. So the reception of Fourier's ideas was distortive in many ways. Later, though, socialists and others would applaud his characterization of laissez-faire as tending to monopoly, to mass impoverishment, and thus to what would later be called crises of overproduction.

Unlike Fourier, Owen embraced the frequent Enlightenment assumption that characters were 'formed by environment.' By reforming that, you reformed them. This, he felt, entitled him to hope that his 'Social System,' as he called it from 1813, could be introduced by existing power holders. Some indeed showed interest. But this evaporated when his 1819 *Report to the County of Lanark* showed him not to be Christian. Almost worse, he rejected marriage as cruel, not least for women. From the 1820s, he began to attract a wider following, mainly among members of trade unions (slowly emerging from a generation of illegality) and among self-educating poorer people. His predicament was sometimes odd: a big employer and believer in his own from-above superiority, serving as figurehead for organized workers, some of whom were moving toward ideas of replacing capitalism by means of a general strike (a strategy later dubbed 'syndicalism'). Others, by contrast, continued calling him their 'Social Father': uniquely superior to them.

Internationally, the 'Owenites' were hardly alone in falling into this prophetic dynamic. In France, the gap between leader and followers was often no less striking. From 1817, Claude-Henri de St Simon (1760–1825) was pamphleteering for rule by enlightened banker–industrialists. As with Owen, industrialization was not a disaster but the door to a new society. St Simon's word for the latter was, not socialism, but 'industrialism.' Here, scientists would replace priests as the elite. Thus, he was at least as top-down as Owen. But, like the Welshman, he and his successors gained followers among poorer groups whose approach could be far more bottom-up.

Nor did all such groups necessarily have to wait for better-off leadership. In Britain, one late example of Radical Enlightenment is the impoverished Newcastle (later London) street bookseller Thomas Spence (1750–1814). His repeated imprisonments on charges such as high treason helped him avoid invisibility. By the end of his life, his 'Spencean' followers, some of them trade unionists, were demanding socialization of a growing number of industries. From 1771, Spence had envisaged all land being owned by parish communities, which would rent it out to individual members. All members, women as well as men, would have one vote for the parish council and for the national senate. All those unable to work would be guaranteed free food and board. Abolishing private landownership would by definition abolish aristocracy, the House of Lords, and much else. Class would be further fought by a new phonetic alphabet: not only undermining accentual snobberies but also easing education and self-education. For Spence, all this would constitute 'the real rights of man,' a phrase he had used before Tom Paine's more readable but less collectivist two-volume defense of the French Revolution (*The Rights of Man*, 1791–92).

Owenites, like Fourieristes, sometimes tried to withdraw from political action of any kind within the Bad Old Society, founding exemplary communities. These would act as beacons of the New Social System, spreading it by example. Few did; most collapsed messily. Disillusion lasted into the next generation of socialists, typified by Karl Marx (1819–83) and his manufacturer–friend Friedrich Engels (1820–95) who scornfully labeled their predecessors 'utopian socialists.' The label spread to almost anyone who attempted to specify what the new society might look like. Sketchiness about aims is thus a frequent indicator of Marxist influence: apart from 'socialism' constructing 'communism,' one should not bind future comrades.

More positively, some poorer Owenites, particularly in industrial Northern England, moved toward erecting their own cooperative economy. In this, their nearest French equivalents would be the followers of Pierre-Joseph Proudhon (1809–65), a self-taught proofreader and sometime admirer of Fourier who, apart from coining the resounding motto 'Property is Theft' (1840), sketched a 'mutualist' system: workers' cooperation underpinned by low-interest banks.

As in England with some of Spence's followers, so in France, intellectual self-confidence propelled many artisan socialists into sympathy with radical conspiratorialists such as Auguste Blanqui (1805–81). Blanqui related resoundingly to a world of political clubs which, thanks to repression often fiercer than in Britain, met secretly. He himself was of respectable origins, but moved swiftly from what can be labeled social republicanism toward socialism: in British terms, from Paine to Spence. Not that his milieu lacked apostles of one or other utopian tendency. But many of them were themselves moving toward conspiracy, and Blanqui himself conspired tirelessly. He accumulated 35 years of detention and one death sentence. By around 1860, he was in effect strengthening anarchists such as the Russian Mikhail Bakunin (1814–76: possibly the first self-described 'anarchist') in their suspicion of mass organization and in their principle of 'Neither god nor boss.' His coup attempts could mobilize hundreds of conspirators. Into the early twentieth century, he attracted respect from some of those socialists who had been forced into illegality – as in Russia – including from the young Vladimir Ilyich Ulyanov, better known as Lenin (1870–1924). The majority of identifiable activists in the Paris Commune (1871) saw themselves as more or less Blanqui's followers: his election as the new regime's President was no surprise. That the rightwing government had detained him shortly before removed an instinctive military organizer while sparing him from massacres perpetrated by its own troops.

In France and by Blanqui's death, his followers and those of Marx were merging. They did so more tightly in 1898, by founding a Revolutionary Socialist Party. But this organization was hardly alone. Of greater long-term importance was the slogan that Blanqui had launched during 1870 against both the

rightwing Provisional Government and the advancing Prussians: 'la patrie en danger.' France, as in the early 1790s, remained the 'fatherland' of revolution. His cry was destined to smooth many French socialists' slide from internationalism to patriotism during the crucial days around the outbreak of war in 1914.

As such, a vibrant link between the early-middle and late nineteenth century, Blanqui had no equivalent in Britain – if we omit some Irish nationalists – whose stances on socialism varied. The Owenites had declined from the 1830s. The main political movement was for one-man-one-vote democracy, concisely formulated in the 'People's Charter.' Some 'Chartists' such as James Bronterre O'Brien (1805–64) were open to socialist ideas, although very few to most then-socialists' feminism. But few wished to spend time defining the root-and-branch reforms they all expected to bring about, once parliament had democratized itself: a stage never reached. Their differences over aim were anyway separate from those on strategy, thus guaranteeing further confusion. By around 1900, Europe's strongest socialist parties were to have their own press and engage constantly in mass petitionings and demonstrations, occasionally in political strikes, not least for the vote, and so had Chartists during the 1830s and the 1840s, sometimes massively, but without commitment to socialism.

Tighter trade unions of 'skilled' men were emerging in Britain from the 1840s. Their stability and moderation were a standing refutation of predictions in Marx and Engels' 1848 Communist Manifesto that intermediate groups would fall through a widening chasm between capitalists and an ever-swelling majority of the population, seen as 'proletarian' in Marx's sense of still powerlessly poor, but destined to overthrow capitalism because increasingly networked in and between large workplaces. Nevertheless, they may have felt flattered by another prediction: of class struggle as the motor of human history between the long-forgotten abandonment of 'Primitive Communism' and the coming attainment of full 'Communism.' When the manifesto's two authors organized an 'International' of social revolutionaries, they welcomed these union leaders as counterweights to those more revolutionary than themselves, particularly Bakunin. Those who did affiliate were moved partly by the hope of combating any import of strikebreakers from abroad. But any violent perspectives for Britain itself were normally anathema to most of them. Consequently, when Marx hailed the Paris Commune as an example for all proletarians anywhere, they mostly walked away in disgust: not for them a 'dictatorship of the proletariat,' with constant from-below democracy plus instant dispossession of the possessors. This walkout removed Marx's anti-anarchist counterweight. So he and Engels arranged for the International's headquarters to move to New York, then still sufficiently peripheral for quiet burials (1878).

A Social Democratic Party Cosmos Shatters

By the time its successor, seen as the 'Second International,' was formed on the hundredth Bastille Day (1889), much had happened. Not least in Germany. Here, the model emerged for which socialists worldwide long aspired: of a 'Social Democratic' Party (SPD), linked to a trade union movement and committed to replacing the capitalist system, even while working within it so as to improve social and working conditions in the here and now. True, there was some perspective that these improvements were themselves bringing socialism nearer. But tensions between the long and short terms were destined to complicate politics within any national capitalism. Worse, in 1914, they helped hollow the commitment of most socialists to internationalism: few gladiators supported demolishing their national arena. That revelation was to split socialists till nearly all 'communist' states collapsed in 1989–91. Twentieth-century socialism became an interaction between two socialisms: Communism and 'Western' Social Democracy.

So widely and lengthily can the SPD's incoherences be said to have reverberated. They can partly be traced back to the antics of one hyperactive polymath, Ferdinand Lassalle (1825–64). In 1863, he had helped founding a General German Workers' Association. In 1875, this had formed what soon became the SPD, jointly with followers of Marx. These, too, proved open to Lassallean hopes of using the existing – in Germany, highly illiberal – state for socialist purposes. Lassalle (who had irrelevantly perished in a duel) had even been prepared to ally the working class – publicly with poorer peasants, but privately with the aristocracy. Perhaps fewer SPDers would have remained under his posthumous influence, had they known more details of his contact with the Prussian (soon German) Chancellor, von Bismarck. The latter saw socialists among his main enemies, not least for their avowed internationalism. His strategy evolved into: granting universal male suffrage, but to a still weak parliament on a franchise, which disadvantaged working-class voters; social insurance and similar reforms as further sweeteners; and, from 1878 till he retired in 1890, a series of Anti-Socialist Laws. These were designed, not to outlaw the SPD directly, but to impede its workings and harass its leaders. Resoundingly, Bismarck's strategy boomeranged: repression boosted legitimacy so much that, by 1890, the Party enjoyed a vote larger than any rival and, by 1914, a membership over a million. Its 1891 program, crafted jointly by the fount of Marxist orthodoxy Karl Kautsky and his soon increasingly 'revisionist' comrade Eduard Bernstein, had by 1914 enjoyed translation into 16 languages.

It echoed themes heard internationally in 1889 Paris: mass mobilization for obtaining universal suffrage; support for all social reforms and any dispossession of private capitalists, thereby accumulating something that would somehow abolish capitalism; economic equality between women and men ('equal work, equal pay': a phrase slipperier than at first assumed); opposition to war. By 1914, the key tension between reforming capitalism and seeking its overthrow was obsessing many SPD intellectuals, with Kautsky defending existing principle, whereas Bernstein offering elegance too honest for most: "the aim [was] nothing; the movement everything." Only a handful, such as the internationally much-experienced Rosa Luxemburg, tried formulating revolutionary strategies such as the General Strike.

Such debates were hardly unique to Germany. But few countries had a comparable evolution. In Britain, long the most industrialized, avowed socialists had been few and weak since the collapse of Owenism, although by 1890 they were indulging in splits. Some, however, did settle down to allying

with trade unionists, irrespective of 'isms.' The result was a 'Labour Party' that, from early 1918, had a membership both individual and affiliated via a growing number of unions and, via the famous 'Clause 4' of its 'Constitution,' a commitment to "securing to the workers … the fruits of their labor, by means of the socialisation of production, distribution and exchange." 'Socialization' was deliberately ambiguous, dodging divisive clarities such as 'under workers' control.'

By contrast in Italy, little industrialized outside some Northern cities, the Socialist Party was by 1914 attracting 3 million votes, mainly from small-town artisans and, particularly in the landlord-ridden South, peasants. In France, another very different story, state repression after the 1871 Commune had given socialists many martyrs and exiles, but little unity or membership: low membership was to persist among trade unions and socialist parties into the twenty-first century, although punctuated by explosively widespread solidarity. In Iberia, actionisms such as anarchism fed on centuries of absolutist repression and intrusive Catholicism – all confirmed, at the start of the nineteenth century by a war of national liberation against an avowedly progressive invader (compare 1980s' Afghanistan?). But exceptionally, industrialized parts of Catalonia were able to produce a far richer range of tendencies.

Scandinavia and the cities of Bohemia and Austria produced labor and socialist movements comparable in local strength to the German, and usually modeled on it. Until Austrian Fascists crushed them in 1934, they inspired reformers, notably in Scandinavia and Britain, with achievements in housing, welfare, and education. Their intellectuals tried to upstage problems of nationality by uncoupling it from territory to link it to personal choice.

Further afield, socialism, anarchism, and (as in the 'Industrial Workers of the World') syndicalism in the continents of massive migration from Europe often remixed all the trends and tendencies in home countries, intensified during explosive strikes. Whatever the verbal commitments to internationalism, ethnic relations among people of European background varied. Invariably, 'nonwhites' were seen as inferior and, self-confirmingly, as wage cutters. 'Workers of the world unite and fight for a white South Africa' was already the strategy of militant Johannesburg strikers during 1913, 9 years before the slogan arrived with a more vast upheaval in 1922. Only during the 1920s was one self-taught Peruvian, J.C. Mariategui, to try reformulating Marxism so as to take account of non-European, here Andean, social forms. His early death (1930) guaranteed that every side in an increasingly split world movement would distort his legacy, till perhaps the 1990s.

By 1914, the main danger was of drift toward war. Socialists defined themselves as antiwar. When the International's 1904 congress coincided with a Russo-Japanese conflict, Georgii Plekhanov and Sen Katayama had publicly embraced. But machine guns do not embrace: what could socialists do? During 1905, Sweden's unions and Social Democrats had prevented their militarists from crushing Norwegian independence. In Belgium and elsewhere, general strikes had won men's right to vote. In many workplaces, 'syndicalism' was also lively. So the Second International proclaimed an 'international general strike' for the event of war. But the phrase remained nebulous. From July–August 1914, it contrasted cruelly with the penalties threatening male workers for resisting conscription (outside Britain till 1916, universal). Many reasoned that they were defending a less reactionary regime (e.g., French, German) against a more (respectively, German, Russian). Blanqui's 'Fatherland' remained 'in danger.'

At first, only tiny minorities resisted this logic. They were least tiny in the Russian Empire and, most bravely, in the already embattled Balkans where sectarian massacre and competing expansionisms had become traditional.

Russia proved strategic. Here, even though Marx himself had wavered, most Marxists had gravitated to his and Engels' own normal version of history: only when an old 'mode of production' had 'exhausted all its possibilities' would it be replaced. Thus, a society could not jump from 'feudalism' to 'socialism' over the 'capitalist' stage. Reality, though, complicated such stageism: Russia's large cities contained some world league factories but were lapped by oceans of peasants who had been freed from serfdom as recently as 1861. The question was how far should Marxists encourage the industrial minority among the exploited masses to go? Wars, by 1905 in the Far East and from 1914 along the Western and Caucasian borders, underlined that question: the Tsarist Autocracy was hopelessly hierarchical and worse. Stable answers had only come slowly from the various Marxist factions but, by 1914, the 'Mensheviks' (because in the 'minority' at one Social Democrats' conference) believed that a capitalist parliamentary regime was the maximum possibility, whereas Lenin's 'Bolsheviks' ('majority-ists') sometimes hoped for more, given the right circumstances. Both factions were fluid. Lev Bronstein ('Trotsky') was maneuvering in between, but had earned prestige during the 1905 revolution, presiding over the St Petersburg Workers' Council (Russian: 'Soviet'). In 1906, he had also launched the flexible phrase 'uneven and combined development' for those paradoxes of Russia's then unusual semiindustrialization.

February 1917 swept autocracy away. At this point, most Mensheviks were for continuing the already disastrous war so as to 'defend the Revolution.' Many Bolsheviks, including Joseph Djugashvili ('Stalin') remained confused. Only when Lenin returned from exile (April 1917, soon joined by Trotsky), did they begin evolving a policy that he finally sloganized as "Peace, Land [for the peasants], Bread [for the cities]." This was the program on which the mainly Bolshevik-led revolution of October 1917 was mounted, with the support of key soviets in Petrograd (St Petersburg's wartime name) and soon elsewhere. Socialists in industrial countries would rush revolutionarily to the rescue: that hope, which Trotsky had voiced as early as 1905, now proved short lived.

For 'Peace,' negotiations opened at Brest-Litovsk with Germany and its allies. The ensuing treaty cost Russia 1 million square miles, one-third of its population, and most sources of oil, coal, and iron. With such 'Peace' and such dwindling resources, the years to 1922 saw a deepening cycle of starvation, forced collection of grain from peasants triggering revolt, and further starvation. Revolts were often independent of 'white' (anti-Bolshevik) armies. A total of 21 foreign countries sent forces to help the whites. Southern Ukraine saw the largest peasant movement: no local novelty, unlike its leadership by an avowed anarchist, Nestor Makhno.

Rival Socialisms and Worldwide Diplomacy

After tens of millions of war deaths worldwide, socialists everywhere had to orient themselves more or less anew and, in relation to the Bolshevik regime, to work out their mutual identities. Lenin, increasingly incapacitated before his death in January 1924, became the fount of an emerging 'Bolshevik' orthodoxy. Under Tsarist repression, he had sometimes argued for building a tightly organized, hierarchical party. Contrastingly, during the revolutionary years of 1905–06 and 1917 when the Bolsheviks had, more slowly than Trotsky, recognized the potential of workers' and in 1917 also soldiers' soviets, he had favored flatter, more open structures, both for hastening revolution and apparently for afterward: "every cook can govern." With Civil War, workers' control of (plummeting) production lost substance, unevenly but inexorably. Just when the Bolsheviks were ascending to unprecedented influence among radical socialists, they were canonizing the least democratic version of their previous practice, while diverging from everything that most socialists had stood for. They embarked on what turned out to be the twentieth century's lengthiest and most widely delusive experiment in political monopoly. Not the bloodiest: that honor would go to Mao's China in the 1950s and the 1960s.

By the end of the Civil War (October 1922), uneven and combined collapse had replaced development. Politically, most of those industrial workers who had supported October had either died in the Civil War or had become bureaucrats. During March 1921, sailors based around the island of Kronstadt, off Petrograd, demanded complete reelection of all soviets. They were crushed as mutineers; many hundreds were executed. Trotsky commanded the forces crushing them. Two days later but deaf to irony, the shaken regime celebrated the fiftieth anniversary of the Paris Commune. The shock of Kronstadt speeded what the Bolsheviks saw as a retreat: the 'New Economic Policy' (NEP) restored a free market in grain. This eased the food situation, but only till the late 1920s. By then, the relationship between the Bolsheviks and reality at home and abroad had degraded yet further.

Abroad, however, that shock was muffled – except among anarchists, plus socialist mavericks such as the former suffragette, Sylvia Pankhurst. Partly, this muffling was due to ignorance. But what of those socialists who had seen parliamentary methods as too top-down? A few, such as Luxemburg, had warned against Lenin during his stricter phases as leading to Party 'dictatorship over the proletariat.' Her murder in 1919 hardly stopped 'Council Communists,' such as Anton Pannekoek, from arguing similarly through the decades. What pushed most socialists in the other direction, into long-term sympathy with the Bolsheviks, was a gnawing inferiority complex. Uniquely, the Bolsheviks had, apparently, 'done it': had made a revolution. True, French sailors and British dockers and soldiers had sometimes struck or mutinied against intervention in the Russian Civil War. But the very peripherality of these successes – let alone the failure of attempted or near revolutions in Germany, Hungary, Italy, and elsewhere during 1918–23 – only emphasized Russia's loneliness.

Parliamentary and other 'moderate' socialists varied from attacking Bolshevik repression, to a stance that lasted till the post-Bolshevik system collapsed. In this, you took property relations as straightforwardly defining a society's 'nature.' You noted, factually, that the means of production within Soviet society were (except during the NEP) overwhelmingly owned in the name of the toiling masses. You then deduced that this society was 'socialist' or at least 'building socialism.' True, you might lament the lack of democracy but, the louder you did so, the more you claimed to be 'doing it democratically.' Under governments such as Clement Attlee's in Britain (1945–51), 'socialization' involved no hint of democracy deeper than parliamentary.

Consensus in favor of top-down methods could also blur extremes of the socialist spectrum. Take two workaholic British intellectuals, Beatrice and Sidney Webb. Elitism and 'national efficiency' were the common threads between their pre-1918 contempt, and their post-1918 support, for a working-class party. In that year, Sidney formulated Clause 4 precisely so as to keep socialists happy. But in 1931, the two were shocked when Treasury civil servants allowed a Tory-led coalition more leeway on economic matters than, a few weeks earlier, a Labor-led one. "They never told us we could do this," Sidney protested – to his diary. He and Beatrice radicalized toward Stalinism. After their visits to Russia during 1935 and 1937, they published successive editions of their *Soviet Communism: a New Civilisation*? For 1937, they deleted the query. Thus, where Bolsheviks had impressed many previously libertarian socialists for actually doing 'it,' so their Stalinist successors appealed to Western despair about capitalism ever recovering from economic depression. In the upside-down world of Stalinist propaganda, Soviet industrialization was reaching unparalleled heights of efficiency. 'Planning' and 'efficiency' became keywords in most versions of socialism. Thus, in 1959 or roughly halfway through capitalism's longest-ever boom, British Labour's leader Hugh Gaitskell was postered as 'Man With a Plan,' and his more leftwing successor Harold Wilson enthralled the 1963 Party conference with eloquence about 'the white heat of the technological revolution.' In 1957, the Soviet Union renewed its claim on such rhetoric by launching Sputnik, the first artificial satellite. Growing excitement about computerization renewed the same claim a little later. Once again, propaganda inverted reality: Soviet mechanisms of economic control were as if designed to multiply the farces of GIGO ('Garbage In, Garbage Out'). Such was the fog of rival vacuities, that some optimists discerned the 'two systems' as 'converging.'

From the 1920s, support for what was now the Soviet Union meant uncritical acceptance of Moscow's changing definitions of world reality. These depended on power struggles at the top of the new structure. From the late 1920s, Stalin and the more or less fresh bureaucratic caste reacted to economic problems at home and further disappointments (to a varying extent, self-inflicted) in China and elsewhere by forcibly collectivizing agriculture for the swiftest possible industrialization. They presented this as building 'Socialism in One Country.' By now, commitment was irreversibly to the tightest of Lenin's prescriptions for party organization.

The same applied internationally. Foreign policy could confuse billions of people, although before Hiroshima it could not yet kill so many. The Communist International ran as an extension of Soviet foreign policy. Zigzags were the sole

consistent pattern. From 1927 to 1934, all socialists whom the Comintern could not control were treated as 'Left Fascists.' This 'zig' reduced many Communist Parties to sects. Worse, in Germany, it deepened the splits in the labor movement and thus helped bring Hitler to power (January 1933). Even then, change had to wait till 1934, after Vienna's Social Democrats had fought local Fascists physically and after Paris Communists and other antifascists had followed common sense, rather than party leaderships, to fight shoulder to shoulder.

So, from 1934 to 1939, policy 'zagged' toward broad alliance against Fascism. Stalin's foreign policy, however, aimed at allying against Hitler with Britain and France. The latter even had a Communist-supported 'Popular Front' government during 1936–37. But, at its head, the socialist Leon Blum refused to arm Spain's Republic against a clerico-Fascist coup: he feared risking war simultaneously along Rhine and Pyrenean frontiers. This gave a green light to Nazi and Fascist intervention, and made Stalin the Republic's sole supporter. His security apparatus exploited this to the full against any local political rivals, both Spanish and international. Partly as a result of this 'civil war within the Civil War,' the Spanish Fascists won (March 1939), redoubling France's strategic vacillations. The Munich agreement of autumn 1938, reached without officially informing Stalin, had neutralized the sole strategic obstacle near his borders to Hitler. Not surprisingly, he began looking for other allies.

But when he reached a nonaggression pact with Germany, the nature of the Comintern meant that international communists worldwide had to follow: their reality 'zigged.' Of course, once Hitler invaded on 22 June 1941, Stalin woke up from a 3-day panic to find himself in alliance with Britain: Hitler had 'zagged' for him. Little more than 3 years later, Stalin was ordering the two CPs emerging most strongly from the Resistance (the French and Italian) to go easy, while agreeing with Churchill on Soviet domination over Bulgaria and Rumania (where Communists were currently rare and very rare, respectively); Britain would recover control over Greece and would share influence in Yugoslavia '50/50.' Both Greece and Yugoslavia had major resistance movements, more or less Communist led. During 1944–49, the British, later the Americans, helped crush the Greek; in Yugoslavia, the Communist leader Tito was allowed to triumph. In 1948, he broke free of Stalin and, from the 1950s, installed much 'workers' self-management.' But he retained Stalinist structures: transparent in self-deception alone. Although, from the 1950s, he installed much 'workers' self-management,' everything collapsed into nationalism and massacre during the early 1990s.

Dissenters and Mass Murder

Analytically, dissenters from Stalinism had difficulty identifying when things had 'gone wrong.' Trotsky, understandably, emphasized his own defeat in the internal Party battle against Stalin, ending with his expulsion from the USSR in 1929. Till Stalin managed to have him murdered (1940), he expected Stalin's 'betrayal' of the revolution to be reversed, more or less from inside the apparatus. This was increasingly unrealistic, not least given massive purges. From the 1930s, Trotskyists around the globe were excommunicating each other over such questions as whether the Soviet Union was a 'degenerated' or 'bureaucratically deformed workers' state.' From 1944, that dogfight between past participles was even less relevant to countries whose regimes had arrived in Red Army tanks. But, into 1989–91, many Trotskyists in the West seem to have fantasized that a working class in one or more such states would resume the old song from whenever they thought Stalin had interrupted it. They derived hope from popular revolts against Soviet rule: from East Germany (1953) to Novocherkassk (1962), climaxing in the bloody drowning of Hungary's workers' councils (1956–57).

Strategically, Trotskyists suffered from having, all too often, to dispute the same part of the political spectrum as Stalinists. Admittedly, where the latter, true to that 'stageism,' which earlier Marxists had criticized, neglected a colony's working class to concentrate too narrowly on the peasantry, Trotskyists could make some impact in the labor movement. Examples would include Ceylon/Sri Lanka, where the LSSP and organizations derived from it remained important from the 1930s to the 1960s, and Vietnam. In Latin America, *c.*1960 Bolivia would be the leading example. Some of Argentina's were the weirdest, inflating their politics intergalactically to pin hopes on flying saucers from, of course, socialist planets: cargo cultism without contact or Fourier's lemonade.

In spring of 1945, when Vietnam's Trotskyists criticized the Stalinists for initially welcoming British forces to Saigon, the response was massacre. Physical 'liquidation' was already a Stalinist hallmark worldwide against rivals or perceived rivals: Spain had established a pattern. A smaller group of victims would be Polish Trotskyists from 1939: a minor target at coordination meetings of the Gestapo and Stalin's NKVD. In Paris, more than one Trotskyist would be eliminated amid the chaos of Liberation in 1944.

But of course loyal Stalinists in the USSR themselves lived perilously. Mass-murderous paranoia was routine from the early 1930s. A regime that had allowed about 6 million Ukrainians to starve to death for resisting forced collectivization during 1932–33 was also capable of more active measures. Repeatedly, fresh dangers were identified, or dreamed up, for liquidation. As if harmonizing with an officially now 'planned' society, the production of deportees and corpses was centrally decided and, at lower levels, quota driven. West Ukrainian Poles had the doubtful honor of being the first ethnic target, later that decade. Till Stalin's death in 1953, there followed Volga Germans, most Chechens and Crimea Tartars, many Balts, numerous members of other Muslim groups and, from around 1950, Jews. Only that supreme death may have saved the latter from mass deportation. Within overall targets, the sheer chanciness of purging is too forgettable – as when Stalin's last secret police chief, his fellow Georgian Lavrenti Beria, favored one Moscow football club by having some players from its local rival deported. Presumably, this was not why the first decision made by Stalin's other successors was to shoot Beria. Maybe there are similarly sick jokes around the perhaps 45 000 000 victims of Mao's 'Great Leap Forward' (his reworking of Stalin's forced industrialization) or of his ensuing 'Great Proletarian Cultural Revolution.'

Beria's murder soon came to symbolize the end of state terror. But was this enough to make the USSR desirably socialist after all? Not for all socialists. For some, the Kremlin's successive 'liberalizations' and 'thaws' were irrelevant to the system's obvious lack of any vestige of more than ritual control by the majority classes. Victor Serge, even while still inside Russia during the mid-1920s, was perhaps the first to transfer the Fascist (and exsocialist) Mussolini's self-description to Stalin: 'totalitarian.' During ensuing decades, others – e.g., Otto Ruehle, Pannekoek, C.L.R. James, Raya Dunayevskaya, and 'Tony Cliff' – talked of 'state capitalism,' so putting the Soviet regime on some continuum with capitalist societies of the 1940s to the 1980s, evidently closest to the most repressive of these. More writers played with the word 'bureaucracy,' thereby elevating symptom into system.

Other socialists reacted against the Soviet model more superficially. Thus the small British Communist Party lost one-third of its members, workers, as well as intellectuals, under the impact of Hungary and of Krushchev's blaming of everything bad on one man (both, 1956). In many countries, disillusion propelled many into a kaleidoscope, known as the 'New Left.' In Italy, 1956 and the reinvasion of Czechoslovakia in 1968 helped the Communist Party (PCI) become Social Democratic in all but name, at the regional level one of the world's most successful ones. This process was oiled by judicious use of the jottings of the party's founding intellectual, Antonio Gramsci (d. 1937): as he had to smuggle them past a Fascist prison governor, they were serviceably ambiguous.

Ironies of Cold War

But at the national level, even the PCI was excluded from power. This was because, from first to last, Soviet and Western leaders, including many socialists, had seen each other as rivals or enemies. This was played down most during Hitler's invasion of the Soviet Union from 1941 (except in Finland). But, as the Nazis proved even more self-defeatingly brutal than Stalin, that phase proved temporary. True, some individuals such as Armand Hammer, of Occidental Oil and much else, could do big capitalist business with the Kremlin from Lenin to Gorbachev. And political or other prisoners in at least China, East Germany, or North Korea have been used as strike-free labor for foreign firms. But ideologically, rivalry remained the main game.

This had at least six effects. First, people living under one regime tended to swallow the propagandistic self-image of the other. Second, the controllers of each feared such propaganda, thereby enhancing its plausibility. Third, just as imperial rivalries between the British and French Empires had been one precondition for even the bare survival of the (1791–) Haitian revolution, so twentieth-century anticolonial movements in Asia, Africa, and elsewhere could profit from the 'East–West' rivalry. Admittedly, there were also costs, as when Cold War pressures speeded Cuba's initially unorthodox revolution (1959) into the Soviet embrace, or as when the ANC's Eastern-bloc-trained security arm raped, tortured, and killed a number of young volunteers. But, once East–West rivalry ended, Cuba was high and dry, while the ANC had to compromise with those who had benefited from apartheid. It had of course been prepared for this somewhat by Soviet stageist prioritizing of 'national' over 'socialist' struggle against Western colonialisms. Unlike with 1990s South Africa, in 1927 the same logic had allowed China's nationalists to reward Comintern support by massacring communists out of the cities.

Fourth, Cold War merely complicated Western socialists' ambiguities over colonialism. Many members of Britain's 'Labour' Party had long been friendly to India's National Congress. Not so their leaders. True, even these recognized reality by abandoning the subcontinent during 1947. But they continued to see colonies as economically useful for metropolitan Britain, behaving imperialistically toward many anticolonial movements. During 1945–46, such internationalism extended to killing at least 10 000 in Surabaya (Java) alone as part of bringing Dutch colonialists back. French leaders, except during the short-lived premiership of Pierre Mendès France, supported successive bloodbaths in Indochina and Algeria (1945–62); in practice, their Communist competitors were seldom much better, unlike some conscripts and individual socialists. Such traditions boded ill for popular reaction to large-scale immigration from excolonies during and after the 1950s.

Fifth, from 1989 to 1991, the existence of a 'socialist alternative' to Western capitalism sometimes put the latter on its best behavior in matters social, economic and, with, for example, America's civil rights movement, 'racial.' This was partly because, sixth, the Nazis' mistake of 'going too far' had further diminished the attractivity of Fascism as an option for ruling groups in capitalism's heartlands such as Western Europe, although ambiguously in Southern Europe and never elsewhere. Even in 'heartlands,' the United States and Canada saw decades of political blacklisting against anyone seen as remotely deviant. Destruction grew physical when the Federal Bureau of Investigation killed some leading Black Panthers. Internationally, American-instigated crushing of anything remotely egalitarian in Guatemala (1954) and, above all, of Salvador Allende's socialist government in Chile (1973) reaffirmed the rules of ideological engagement, even while providing all wings of the worldwide socialist spectrum with plenty of ammunition for arguing that Allende had gone too far or, alternatively, not far enough. More homegrown reaffirmation had occurred with the erasure of at least half-a-million Indonesian Communists during 1965–66.

Fundamentally, Western socialists' greatest successes in policy and often in party politics came during capitalism's longest-ever boom, from 1945 till at least 1973. Conversely, the interwar depression, like the post-1980s neoliberalization of economy and society, was increasingly negative for socialists everywhere. The main exception strengthens this rule: the inauguration of decades of dominance by Sweden's Social Democrats during the 1930s. This had resulted, notneither merely from homegrown Keynesian economists, nor from the long-term compromise signed between the trade unions and the employers, nor even from a soon exemplary record in welfare-policy, but also from Nazi rearmament boosting demand for iron ore.

At any time, though, socialist leaders might be obliged to prove, at least as obsessively as nonsocialists, their 'fitness to

govern' the existing system. That three of Britain's five post-1945 devaluations of the pound sterling occurred under 'Labour,' underlines how difficult this could be. Finally, four successive lost elections impelled that party in 1995 to ditch even ritual commitment to Clause 4. More consistently than any predecessors, Tony Blair led his adjectivally flexed 'New Labour' into courting the 'center ground.' His adoption of Thatcherite economics was unoriginal: New Zealand's Labor government had preceded him. But his 1997 landslide victory made his airy talk about a 'Third Way' (what between, varied) into a success talisman for a number of socialist leaders. Back in 1959, Germany's SPD had dumped all its Marxism. The *c.*2000 SPD leader, Gerhard Schroeder, now followed Blair, if at increasing distance, in joining a rightward-moving consensus for privatizing the public sector and against many of the poorest, whom the Party had been founded to empower.

Summary and Outlook

Brevity requires abstraction. Roughly during the century from October 1917, socialists concentrated on either clambering out of the wreckage at various speeds (China's leaders the most capitalistically) or making capitalism less capitalist. Each preoccupation undermined itself, the latter all the more obviously now that neoliberalism is acceleratingly marketizing societies, let alone economies. Now that even procapitalist governments feel compelled to talk of international control, they are on socialist terrain. Whether socialists are heard there any more depends partly on how they see manual and, say, computer proletarians allying. Whatever happens, proletarianization itself is surely accelerating.

See also: Balance of Power, History of; Capitalism; Cuban Revolution, The; Dictatorship; Enlightenment; First World War, The; Imperialism; Industrialization; Labor History; Marx, Karl (1818–83); Modernization and Modernity in History; National Liberation Movements; Political History: History of Politics; Political Thought, History of; Revolutions of 1989–90 in Eastern Central Europe; Russian Revolution, The; Second World War, The; Social Movements: A Social Psychological Perspective; Welfare State, History of; Working Classes, History of.

Bibliography

Anderson, Andy, 2002. Hungary '56. AK, Edinburgh (Originally 1964).
Anweiler, Oskar, 1974. The Soviets: Russian Workers', Peasants' and Soldiers' Councils 1905–21. Pantheon Publishers, New York.
Bottomore, Tom, Goode, Patrick (Eds.), 1978. Austro-Marxism. Clarendon Press, Oxford.
Bullock, Ian, 2011. Romancing the Revolution: The Myth of Soviet Democracy and the British Left. Athabasca U.P., Edmonton.
Farber, Samuel, 1990. Before Stalinism: The Rise and Fall of Soviet Democracy. Polity Press, London.
Fourier, Charles, 1996. In: Jones, Gareth Stedman, Patterson, Ian (Eds.), Theory of the Four Movements. U.P., Cambridge.
Harrison, J.F.C., 1994. Robert Owen and the Owenites in Britain and America: The Quest for the New Moral World. Gregg Revivals, Aldershot.
Orwell, George. Homage to Catalonia. Secker and Warburg, London, successive reprints from 1938.
Sassoon, Donald, 1996 and 2010. One Hundred Years of Socialism: The West European Left in the Twentieth Century. I.B. Tauris, London.
Serge, Victor, 1963. Memoirs of a Revolutionary. English translation, Sedgwick Peter. Oxford U.P., Oxford. Later editions to 2012 from other publishers.

Socialist Law

András Sajó, Central European University, Budapest, Hungary
Mónika Ganczer, Hungarian Academy of Sciences, Centre for Social Sciences Institute for Legal Studies, Budapest, Hungary

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Sajó, volume 21, pp. 14493–14496, © 2001, Elsevier Ltd.

Abstract

Socialist law as a legal order was a major legal family with common characteristics. The concept of socialist law expresses the ideological legacy of Marxism–Leninism. It was understood by its proponents as the expression of the will of the working class, and an instrument of class struggle, as well as a means for building a communist society. The article examines the extent to which socialist law fits into any concept of law, in light of the facts that its roots were totalitarian and that it had a purely instrumental legitimacy. The authors also introduce the development, structure, and substantive and procedural institutions of socialist law, along with the socialist administration of justice and perception of international law.

Socialist law, as an actual legal order, was a major legal family whose members possessed similar characteristics. The common features of the legal systems of the Soviet Union and communist East-Central Europe resulted from the use of socialist law in maintaining communist party dominance over state and society. Yet the law in the various socialist states differed, too, not only because of modifications in the forms of domination but also because Soviet legal solutions were imposed on diverse cultures and societies.

The ideological legacy of Marxism–Leninism is expressed in the concept of socialist law. According to Marx, law was the will of the ruling class supported by statutory validity.

Socialist law as the positive legal system in communist countries was understood by its proponents as the expression of the will of the working class (or, when class struggle was less emphasized, of the 'toilers'), and it was an instrument of class struggle as well as a tool for building communist society. In the early years of the Bolshevik Revolution, workers' revolutionary legal consciousness was deemed the source of applicable law. Given the totalitarian roots of the legal system, to what extent the norms of the communist state fit into any concept of law remains a fundamental question. If law is a social phenomenon or institution that differs from its environment, is socialist law a law at all?

Stages in the Development of Socialist Law

Socialist law changed over time. It served totalitarian regimes whose efforts were not always successful. State socialism was totalitarian in the sense that it attempted to construct all-embracing state control over every sphere of social life in the service of a single goal. In the early years of the socialist revolution and during certain periods of Stalin's rule, physical control and repression through brute force played the crucial role in this attempt to achieve total social control. In the early phase of socialist legal theory, Pēteris Stuchka emphasized the state power as a "powerful weapon in the hands of the revolution." Stuchka promoted the revolutionary model of socialist legality, which as a continuation of revolution, understands the law not merely historically but dialectically, with a face toward the future. In this sense state power and proletarian dictatorship are essential for the law to develop in an organized manner. It followed that law as a tool of social management, or one that limits government action in accordance with its own specific logic, could not exist. Law was not a serious consideration for the revolutionary mindset. This disregard was facilitated by Marx's belief that there was no place for state and law in the classless society. But Marx also believed that the transition leading to the classless society would be long, and he expected that law, in socialism, would contribute to the extension of equality under the law. Lenin argued for a dictatorship of the proletariat 'based directly upon force and unrestricted by any laws.' In the postrevolutionary period, the communist party position was uncertain regarding the future of law. Yevgeny Pashukanis, the leading authority on communist legal theory, wrote in 1928 that law was a temporary system dictated by commodity relations, and thus the legal superstructure reflects the economic organization of the society. Once socialism had eradicated commodity relations, legal forms would also disappear.

In the 1930s, law (not only criminal law) was structured as an unrestrained authorization, and even a command, to intervene in anything and everything. Such calls to uncircumscribed, coercive intervention transgressed the limits of legal forms. The 1930 joint decree of the Central Executive Committee and the Council of People's Commissars of the USSR, which empowered local authorities to "take all necessary measures … to fight kulaks," would hardly satisfy any definition of law. But legal forms continued to exist even after the planned economy was fully developed. And even during the oppressive Stalinist regime, legal forms were employed at least for show trial and propaganda purposes. Law remained an important internal and external tool of legitimation as exemplified in the Soviet Constitution of 1936 (known as the Stalin Constitution). Law served as a force of total repression in the 1930s, in the sense that it reminded people that they could be brutally handled for no apparent reason and without any protection from other state institutions.

As communist party oppression neared anarchy, unlimited power turned against the elite and their supporters. Totalitarian disregard for the specific needs of the economy and other areas

of social activities drifted into dysfunction. Stalin himself 'brought law back in.' But the promised legal stability had 'nothing to do with the legal restraint on power, and it coincided with the period of greatest repression.' The system moved 'from lawless repression to repressive law' (Krygier, 1994). Officially, socialist law was proclaimed (by socialist leaders) to be the most developed form of law in history, as it served the most developed social formation (communism). It provided socialist legality, a qualitatively superior alternative to bourgeois rule of law. The typical differences between socialist law and bourgeois law, according to socialist lawyers, were the dominantly economic perception and educational function of socialist law as well as its aforementioned underlying purpose, the establishment of communism. From this point of view, bourgeois law is based on a defective capitalist economic system, which by itself precludes the very morality it seeks to create. In addition, because bourgeois law did not rest on the achievement of a grand objective, socialist lawyers considered the efforts of bourgeois law to create order to be a blind endeavor. In reality, however, socialist legality often served show trial and other propaganda purposes only, as exemplified by the Stalin Constitution.

Laws enabled the direct repression of Stalin's Great Terror, although the mass-scale extermination (the purges, the gulag, etc.) was based on action that presupposed a disregard of promulgated legal norms (trial procedure rules, evidence). This is not to claim that the 'disregard' was not officially sanctioned and ordered: disregard of the then-existing official law was a systemic component of the totalitarian use of power. The resulting destruction of social and individual resistance to communist authority, the irrevocable stamp of fear in almost every citizen, served in the remaining half of the twentieth century as the foundation of the repressive order, which could afford to make use of less coercive laws.

After Stalin's death, in 1953 the communist elite reverted to a certain level of legal formalism, by and large for their own protection. This was achieved by a duplication of sorts of the normative system. The legal authorities had strictly observed secret rules regarding actions affecting members of the communist *nomenklatura* (the system of patronage within politics and administration in Communist countries): to proceed against one of its members required authorization from the proper body within the *nomenklatura*. The *nomenklatura* followed its own 'rule of law,' which did not include the right to a hearing of the concerned. The legalistic turn had a certain spillover effect on the rest of the population. Increasingly, law became less directly coercive and punitive and, at least in some of the satellite states, in matters of secondary importance it no longer depended fully on politics. On the other hand, law was recognized as a system that provided structure to a bureaucratic state capable of economic centralization. (The copying of Soviet solutions resulted in less damage to legal forms in the East-Central European states than in the Soviet Union because the use of brute, Stalinist force lasted for shorter periods in these states and because Western legal traditions existed in some of them.)

Beginning in the 1970s, the relative autonomy of law was officially recognized. For reasons of regime legitimation, it was overemphasized to disguise a poorer reality of limited autonomy. Nevertheless, the formal characteristics of modern law (abstractness, general nature of commands, etc.) survived, although in simplified and distorted forms. The instrumental use of law served a bureaucratically ossified institutional politics. The official doctrine emphasized that the autonomy of law could be relative only vis-à-vis the economy, which determined all forms of social life. "Legislation, whether political or civil, never does more than proclaim, express in words, the will of economic relations" (Marx and Engels, 1976). In reality neither the economy nor other social relations were respected, and only specific concerns of power and domination mattered, as understood by the communist leadership. Law remained a tool of the planned economy, and it served the deliberate changes of the planning system.

Law was declared the engineering device that helped organize people in their building of communism. The composition of courts also promoted the educational function of socialist law, for citizens, as people's assessors, could experience the justness of both socialist law and judicial activity through direct participation in the administration of justice. Furthermore, the institution of people's assessors could be regarded as a step taken toward the ultimate disappearance of law. Revolutionary law, although an instrument of class struggle, had an element of internal legitimacy, as it served social (class) justice (through murderous revenge, expropriation, and redistribution). In the instrumental understanding of socialist law, moral considerations became irrelevant. Instrumentalism was also distorted, however, as 'relative autonomy' allowed only very limited respect for the characteristics of the tool. Legal dogmatics or consistency requirements were easily skirted. Yet there were actual elements of independence in law simply because the political system could not generate a unified will. There was competition within the communist political elite, which contributed to at least superficial respect for legal forms. Moreover, law-enforcement organizations had a vested interest in legalism, and to the limited extent that they played a role in party politics, law became a point of consideration. To the extent that the communists needed efficient state machinery, they needed law as an instrument of control over the state bureaucracy (*see* Administration of Justice). Law was important insofar as centralization required uniformity and unity of action.

The Structure of Socialist Law

Socialist law was understood as an ideological superstructure, and its provisions were subject to ideological scrutiny. Law (its provisions and understanding) had to be politically correct, as understood by whatever ideological line the communist party touted. The ideological roots and functions of socialist law led many commentators to the conclusion that it was a mutation of natural law. It was argued that it originated in a set of doctrines or expectations that were above the law. Even the vulgar Marxist claims that law as superstructure reflects economic realities might have been interpreted as being in the natural-law tradition: it was the mystical economy with its 'objective law' that determined law. Socialistic traditions of social justice also show structural elements of natural law. Max Weber criticized these socialistic tendencies to material justice as detrimental to the formal, rational qualities of law.

Socialist legal systems, as they existed in state-socialist regimes, were positivistic and functioned as positive legal orders. According to the prevailing definition, law was the hierarchically structured, sum total of state-made norms. The profound departure from promulgated norms and principles occurred not because of the observance of some kind of higher law. On the contrary, the departure was the result of changing secret orders, commands, and expectations emanating from the communist party leadership and secret services. Of course, once such external political intervention becomes routine or even institutionalized, one can argue that there is no more law in the sense of normative, predictable expectations; hence socialist law (at its worst) does not fit into any notion of law.

Socialist legal systems, in principle, were strictly hierarchic. Statutes (parliamentary enactments) were of absolute supremacy, while constitutions offered only guiding perspective and basic rules of state organization. Law had to originate from the people's representatives; therefore, judicial lawmaking was ruled out as ideologically unacceptable, after the early revolutionary period. Statutes were supreme because they were enacted by parliament, which was held to be the supreme state body. The rigid position on legal origins reflected the dominant nineteenth century German and French views and a simplification of Rousseau. Each socialist regime institutionalized some kind of small and reliable body within parliament (often called a presidium). The presidium had full power to act on behalf of parliament, which convened infrequently. The presidium had the power to amend existing laws. Decrees were subordinated to statutes. In theory, decrees were to serve exclusively for the execution of statutes. But in reality, substantive issues were often determined at the level of executive decrees. Officially courts were restricted to pure application of the law; they could not and did not engage in either judicial lawmaking or the adjustment of existing norms to the prevailing circumstances. Lenin described the socialist court as 'an organ of state power.' Judges were subject to the law and were not allowed to have an indifferent attitude toward the policy of the government. Courts had to endeavor to administer justice in conformity with the directives of the government, bearing in mind the principles of socialism and the socialist sense of justice. However, a number of practical issues were determined by guiding principles of supreme courts on the basis of 'the laws in force.' These guiding principles determined the application of particular enactments or at times filled the gaps in law. *Stare decisis* (the doctrine of precedent) and the binding force of precedent were unknown doctrines in this system; lower courts were, however, bound to follow interpretations of higher courts in the name of socialist legality. This was imperative given the informal control over the judiciary (*see* Administration of Justice).

Substantive and Procedural Institutions

The centralized socialist management of the economy and social life was suspicious of any expression of autonomy. The priority of 'social interests' (or 'society's interests') defined the legal system. No autonomy (in the sense of self-determination) was permitted, and legal structures enabling or protecting autonomy were pruned even from private law. The traditional continental divide between public and private law was refuted as bourgeois legality. As Lenin stressed: "We acknowledge nothing as 'private.' For us *everything* in the province of economics is in the domain of *public law* and not of private law …. Hence our task is … to broaden the right of the state to abrogate 'private' contracts." In practical terms, this meant not only the denial of privacy and sanctity of contract but also the primacy and special protection of state property. The other reason for giving public law prominence was that, as socialist law aimed to create a new social order, any violation of law affected, beyond the actual victim, the state as well as public interests.

Depending on the period, private property and private initiative were prosecuted, prohibited, or greatly restricted, at best. Private ownership was recast as personal ownership in order to allude to the fact that it must only be used for the satisfaction of personal needs and for the purpose for which it was intended, and this property could not be utilized for gaining profit. Property consisting of the 'means of production' was restricted even where it was permitted (as in the German Democratic Republic or, as far as land was concerned, in Poland). 'Social interests,' determined by the state, were favored above all other claims. Restrictions on personal freedom went beyond etatist concerns. Freedom of movement and residence were limited, and at least in the Soviet Union, where the czarist tradition survived, freedom of internal movement was subject to administrative approval.

Speech and press freedoms were permitted only in the interest of workers, which resulted in institutionalized or informal but total censorship. Personal choices and alternatives were limited or prohibited (including choices about education, health, profession, sexual identity, and the like). State control over the individual was extended to the intimate, private sphere. Decisions regarding family relations were subject to intervention by the *Prokuratura* (State Attorney's Office). Abortion was not an issue of self-determination but a matter of demographic or workforce politics, although certain social compromises were observed in this context. In Hungary, the nearly unlimited access to state-funded abortion was the result of an official repudiation of coerced demographic policies. In other countries, like the Soviet Union, abortion as the primary form of family planning derived from a total disregard for human dignity.

The legal restrictions imposed on the general population went hand in hand with the award of privileges for supporters of the regime, although only certain privileges were codified (e.g., access to higher education for family members of committed regime supporters). Privileges were not absolute. They were the reward for loyalty to the system. But the ordinary, loyal citizen (contrary to the *nomenklatura*) had no title or claim to privileges, which, when bestowed, were an act of grace and goodwill. The discretionary distributive powers of the public administration were used to favor the politically loyal (or at least to punish the suspect). This was the case, for instance, in the assignment of public housing. During the less dictatorial periods of state socialism, these discretionary methods of allocating goods and services were the primary tools of social control. Social control through the discretionary application of law put society in a condition of permanent dependence.

Administration of Justice

The contribution of the administration of justice to the restriction of liberty varied historically. During the Stalinist terror, courts served as theaters of repression and injustice. In less dramatic times, the administration of justice promoted etatism and protected the interests of the communist party. Most administrative decisions could not be appealed in court, although in Hungary and Poland, as a rule, there was the possibility of appeal against administrative decisions, except those involving the more sensitive issues of military and police administration.

As to judges, in the formative revolutionary years of socialist power, the legally trained personnel of the previous regime were replaced with incompetent but politically trustworthy laypeople. The basis of revolutionary justice was the revolutionary consciousness of worker-judges. After World War II, in the Soviet-dominated satellite states the 'democratic attitude' of lay assessors played a role in politically charged procedures. In more settled times, judges received formal legal training but worked in personal and organizational dependence on court presidents. Courts were technically under the control of the minister of justice. These courts were not structured to be independent, and the existing political control further increased the likelihood of 'telephone justice.'

Socialist legal theory emphasized the key role of the *Prokuratura* in the administration of justice. The organizational design of the Soviet *Prokuratura* was inherited from the czarist administration. It was a select, centralized body; procurators served in total subordination to their superiors, in what was a quasimilitary organization. They were much better paid than judges. The *Prokuratura* was in charge of criminal prosecution, representing the state in court, but it also routinely supervised the activities of the state administration. Lenin considered the *Prokuratura* the bulwark of socialist legality. He knew that a centralized state power could not afford any particularism, and all local, particular normative systems represented serious deviance. Normative particularism was especially dangerous if it originated in state bodies. As Lenin wrote to Stalin, "legality cannot be legality of Kaluga or the legality of Kazan but must be one single all-Russian legality." Thus the task of the Prokuratura was to "see to the establishment of a truly uniform understanding of legality." The supervisory powers of the *Prokuratura* included the authority to ask for ordinary and extraordinary review of court decisions, even when it was not a party in the case it sought to review. Ironically, the fear of localism resulted in minimal protection for the population: the central bodies had an interest in mobilizing local populations against local authorities, and they encouraged denunciation of local (nonpolitical) abuses to the extent that they deviated from centrally authorized injustice. The dependence of citizens on local services was attenuated in the name of central law. Rules of procedure aimed to protect the material and political interests of the state, and allowed for paternalistic intervention in private litigation. The procedures followed simplified continental models that had traditionally existed in the given country. The inquisitorial nature of the traditional continental criminal procedure was further aggravated by Soviet solutions regarding the admissibility of police-gathered evidence.

In civil matters, judges were required to establish 'objective justice' and were not bound by parties' motions. The paternalistic nature of continental civil procedure dramatically increased, resulting in further dependence, as far as the parties were concerned. On the other hand, the simplifications and limited access to court yielded cheap and relatively swift administration of justice (where courts were available at all). Civil litigation rates were stable, and the most common type of litigation was divorce related. Criminal litigation (except when criminal justice was used as a tool of class struggle, in the 1930s and 1950s) stabilized at a low level. Criminal behavior was limited because of the efficient social control exercised by the oppressive state mechanism.

The perception of the role of advocates was also peculiar, as they had to seek cooperation in the course of their activities with the judge and the prosecutor, and reveal the matter before the court in all its aspects, even at the expense of proving the guilt of their clients. Hence advocates were considered as servants of socialist legality, who contributed to justice and put the interests of society before that of the pleaded individual. In the Soviet Union, the objectiveness of advocates was fostered by a collective organization of advocates. Individuals had to contact such groups of lawyers, which in turn designated their advocate and determined the amount of fee to be paid.

Perception of International Law

A coherent theory on international law cannot be encountered in Marxist ideology. Economic considerations, however, made both the conducting of international trade and the development of international relations inevitable. The gaining of a public law character by socialist law also affected international law. In the wake of the realization that an international trade monopoly existed in socialist states, the state almost always appeared as a subject of treaties concluded with other states or their citizens or legal persons. Public and private international law intertwined, and the regulation of international economic relations became increasingly a matter of public international law. Similarly to domestic law, international law was considered class law. According to certain scholars, such as Evgeny A. Korovin, the international law of socialist states stood in opposition to the international law of bourgeois states. The doctrine of 'internationalist solidarity,' which was followed by socialist states and expressed the necessity of conformity to the Soviet Union, constituted the cornerstone of both the foreign policy making and regional cooperation of these states, and served to put into effect the policy formulated and directed by the Soviet Union.

This separate conception of international law, however, did not reckon with the fact that socialist states and capitalist states cooperated in their economic and other international relations and to maintain and ensure peace. Once the socialist bloc declared the peaceful coexistence of capitalist and socialist states, it had an impact on the development of international law. The principle continued to exist in the form of the principle of duty of states to cooperate, according to which states "have the duty to co-operate with one another, irrespective of the differences in their political, economic and social systems … ." (G.A. Res. 2625, 25 U.N. GAOR Supp. No. 28. (A/5217),

121 (1970)). Although socialist international law recognized the principle of sovereign equality and territorial integrity of states and the prohibition of intervention in the domestic affairs of states, in reality these principles were observed in their breach.

Socialist states generally considered both theories describing the relationship of international law and domestic law – that is, monism and dualism – equally acceptable. However, within the framework of monism they only recognized the primacy of domestic law over international law, and rejected the primacy of international law, claiming that this idea was contrary to the sovereignty of states.

The so-called Brezhnev doctrine, according to which socialist states must take into consideration the dictates of socialism and the fundamental interests of other socialist states in the course of making decisions, was likewise meant to restrict the political independence of members of the socialist bloc. The doctrine sought to reconcile its content with the respect for sovereignty by arguing that the sovereignty of individual socialist states could not be in opposition to the interests of the socialist world or the worldwide revolutionary movement, and as such, the observance of interests of other socialist states could not be labeled as a limitation of sovereignty.

The 'absolutization of sovereignty' (Kovács, 2007) can also be observed in the determination of matters falling in the exclusive domestic jurisdiction, or domaine réservé, of states. Individuals were generally deprived of their citizenship if they emigrated to a bourgeois state, mainly on the basis that they 'gravely violated the requirement of alliance of citizens.' At times the property of such individuals was confiscated in part or in whole, and they were not even allowed to return until the political transition, after which many of them were repatriated. The socialist conception of human rights, furthermore, demonstrates the socialist perception of the relationship of the individual and the state by differentiating human rights from the rights of the citizen, and by overemphasizing the duties of citizens.

In conclusion, although the prevailing definition of socialist law emphasized a hierarchically structured sum total of state-made norms, in reality a fundamental disregard of the promulgated law prevailed. Substantive and procedural institutions like the public prosecutor's office, political control over the courts, denial of access to court, and restrictions on fundamental rights, as well as institutionalized politically biased discretionary powers of the administration, undermined the rule of law.

See also: Field Theory; Law and Society: Development of the Field; Legal Systems, Classification of; Marxism–Leninism: The Ideology of Twentieth-Century Communism; Socialism: Historical Aspects; Socialism.

Bibliography

Collins, H., 1982. Marxism and Law. Oxford University Press, Oxford, UK.
David, R., Brierley, J.E.C., 1985. Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law. Stevens and Sons, London.
Eörsi, G., 1979. Comparative Civil (Private) Law. Akadémiai Kiadó, Budapest, Hungary.
Feifer, G., 1964. Justice in Moscow. Simon and Schuster, New York.
Hazard, J.N., 1969. Communists and Their Law. University of Chicago Press, Chicago.
Kovács, P., 2007. The effect of the change of political regime on the Hungarian doctrine of international law. In: Jakab, A., Takács, P., Tatham, A.F. (Eds.), The Transformation of the Hungarian Legal Order 1985-2005. Kluwer Law International, Alphen aan den Rijn, pp. 453–463.
Krygier, M., 1994. Marxism, ideology and law. In: Krygier, M. (Ed.), Marxism and Communism: Posthumous Reflections on Politics, Society, and Law. Rodopi, Amsterdam, pp. 135–168.
Marcuse, H., 1971. Soviet Marxism: A Critical Analysis. Penguin, Harmondsworth, UK.
Marx, K., Engels, F., 1976. The Poverty of Philosophy, Collected Works, vol. 6. Lawrence and Wishart, London.
Podgorecki, A., Olgiati, V. (Eds.), 1996. Totalitarian and Post-totalitarian Law. Dartmouth, Aldershot, UK.
Stuchka, P.I., Sharlet, R., Maggs, P.B., Beirne, P., 1988. P. I. Stuchka: Selected Writings on Soviet Law and Marxism. M. E. Sharpe, London, Armonk.
Sypnowich, Ch, 1990. The Concept of Socialist Law. Clarendon Press, Oxford, UK.
Szabó, I., Kovács, I., Péteri, Z., Bokor, H., et al., 1966. Socialist Concept of Human Rights. Akadémiai Kiadó, Budapest, Hungary.
Thompson, E.P., 1977. Whigs and Hunters: The Origin of the Black Act. Penguin, Harmondsworth, UK.
Vyshinskii, A.I.A. (Ed.), 1948. The Law of the Soviet State. Macmillan, New York.
Zweigert, K., Kötz, H., 1987. Introduction to Comparative Law. Clarendon Press, Oxford, UK.

Socialist Societies: Anthropological Aspects

Katherine Verdery, Graduate Center, City University of New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Socialist societies constitute a class of modern societies loosely sharing certain formal similarities in their political, property, and production regimes. They tended toward political monopoly by a Communist Party, a predominance of collective or state property forms, and centrally organized economic production toward redistributive goals. Over time, these formal features were modified by systemic reforms and everyday practices of both citizens and elites, including clientelism, kinship and ritual practices, gift-giving, and other self-forming behaviors. Socialist systems diminished many inequalities yet created new forms of hierarchy.

Socialist (also known as communist) societies constitute a class of modern societies loosely sharing two distinctive features: the political dominance of a revolutionary – usually a Communist – Party, and a property regime resting at first on state or collective/cooperative rather than private property forms, though subsequent reforms sometimes modified those. This definition excludes societies governed by socialist or social-democratic parties in multiparty systems, such as the Scandinavian welfare states. It has historically included, among others, the Soviet Union, the East European countries, the People's Republic of China, Mongolia, North Korea, Vietnam, South Yemen, Cuba, Nicaragua, Ethiopia, and Mozambique. These societies differed significantly from one another – as was apparent in their varied trajectories as socialism decomposed – but nonetheless share elective affinities that justify treating them together.

Socialist societies came into being with the 1917 Bolshevik Revolution and the establishment of the Soviet Union. Others founded later by various means (revolution, conquest, or annexation) all reflected the influence, both negative and positive, of the Soviet model created under Lenin and Stalin. All employed extensive political repression and centralized control, yet they also generally increased living standards as well as access to housing, employment, education, and medical care. Instituting a historic assault on social and gender inequality (see Gal and Kligman, 2000), they improved life chances for many disadvantaged people even while creating new forms of hierarchy.

Research on Socialist Societies

For much of the twentieth century, social-science understanding of these societies was politically circumscribed. Their own scholars were under Party control; nonetheless, invaluable knowledge came from dissidents (e.g., Casals, 1980; Konrád and Szelényi, 1979) and from scholars in countries with strong reform movements (e.g., Kornai, 1980; Wesolowski, 1979). In Western Europe and the United States, scholarship on socialist societies was hampered by impediments to on-site fieldwork and ideologically circumscribed by the concept of totalitarianism, which overstressed the importance of terror and top-down control. This situation improved beginning in the late 1960s, for Eastern Europe and the Soviet Union, as détente and the impetus for technology transfer produced an infrastructure for intellectual exchange and reciprocal influence (e.g., Bockman and Eyal, 2002), and in the 1980s, for China. On-site research led to analyses that discarded the idea of totalitarianism to investigate the inner workings of these societies in more detail, with greater attention to kinship, ritual, gender relations, the mobilization of labor, and the exchange of gifts and favors, as well as more subtle treatments of political economy (e.g., Hershatter, 2011; Hough, 1969; Humphrey, 1983; Walder, 1986; Yan, 1996).

Political Economy and Social Organization

From their beginnings in the nineteenth and early twentieth centuries, communist movements came to power in a variety of ways, sometimes on their own (e.g., Russia, China, Cuba, and Yugoslavia), and sometimes – especially during and after World War II – with the help of external communist parties (e.g., Vietnam, Romania, Hungary, and Poland). Although they varied widely in economic profile (from industrialized Czechoslovakia to agricultural China and Cuba to pastoral Mongolia), all socialist societies embarked on programs of state-led development through central planning that aimed at rapid economic growth by noncapitalist means. The following description begins with 'classic' socialism (see Kornai, 1992), prior to the reforms introduced at various times, and draws on research in both agricultural and industrial settings in China, the Soviet Union, and Eastern Europe.

Redistributive Bureaucracies and Reciprocal Exchanges

Of the three basic modes of exchange posited by Karl Polanyi – reciprocity, redistribution, and markets – socialist systems subordinated the operation of market principles to redistribution, which it made the chief exchange modality and the system's *raison d'être*: everyone would have the right to work, and the wealth they produced would be redistributed as welfare for all. Centralized redistribution required social ownership and control of productive resources, together with centralized appropriation of the product for reallocation along lines set by the Party. To orchestrate production and distribution, Party leaders developed central planning. Fulfilling plans depended

heavily, however, on relations of reciprocity, both clientelistic and egalitarian.

Although these were highly bureaucratized societies, they were not Max Weber's impersonal modern bureaucracies dedicated to procedural rationality. Dense social networks sustained reciprocal exchanges that, far from constituting resistance to bureaucratic control, were intrinsic to its functioning: horizontal exchanges enabled socialist managers to do their jobs, and clientelism facilitated controlling both labor and Party cadres in seemingly noncoercive ways. Personalistic ties extended far beyond the bureaucracy as well, linking ordinary citizens with one another as well as with bureaucrats. The warp and woof of socialist societies, then, consisted of vertical and horizontal relations of loyalty, patronage, and exchanges of goods, favors, and gifts (see, e.g., Ledeneva, 1998). Although the forms of these relations differed according to the various societies' histories and cultures (such as gift relations in China, *compadrazgo* in Cuba or Nicaragua, or specific kinship structures of Soviet minorities), they were widely prevalent everywhere. In every country, words indicating 'connections' (Russian *blat*, Chinese *guanxi*, Romanian *pile*, Hungarian *protekció*, Cuban *compañerismo*, and Polish *znajomości*) often appeared in daily speech.

Organized Shortage

Personal relations were integral to socialism in part because of certain pervasive features of planned economies: tension between Party planners and enterprise managers, bargaining and soft budget constraints, and barter amid shortage. Socialist property was organized in a 'hierarchy of estates' (Humphrey, 1983; Verdery, 2003), within which production occurred by disaggregating the ownership/control rights theoretically held at the center and devolving these downward to the lower level managers, where actual production took place. Because planning could never foresee all contingencies, these managers had considerable discretionary power. Expected to achieve Party directives with inadequate wage funds and to exceed their plan targets with insufficient raw materials, they bore responsibility far greater than the means allotted them for fulfilling it. Therefore, they strove constantly to expand their local power base and their effective capacities. In so doing, they posed a threat to their superiors, who thus were motivated to encourage subordinates' loyalty with favors and patronage.

In redistributive systems governed not by demand for products but by the Party's planned allocation, materials for production could not simply be bought on a market; their availability depended on the supplies budgeted in plans and on often-inefficient central distribution. Managers therefore requested more supplies than they needed, hoping to obtain enough materials to fulfill and exceed their targets: they bargained their plans. Because the planning mechanism required firms to produce regardless of profitability – they operated under soft budget constraints and were rescued rather than bankrupted if they lost money – local managers could overstate their needs for materials and investments and then hoard any excess, with impunity. They also strove to bargain their plan targets downward, making it easier to fill these and have goods left over. Comparable processes occurred in both industry and agriculture, as cadres everywhere manipulated information, underreported production, and engaged in illicit trade to benefit their firm or locale.

The result of bargaining and hoarding was endemic scarcity of the materials necessary for production; thus, classic socialist societies were *economies of shortage* (Kornai, 1980). Shortage caused not just competition among firms but also widespread exchanges, each manager supplying from his hoard today the materials needed by another who would return the favor tomorrow. Thus, firms hoarded materials not only to cover emergencies in their own production but also to backstop the supplies needed by others in their network. These arguments about shortage apply specifically to the process of production, but that process in turn had implications for the availability of consumer goods, generating the queues characteristic of many socialist societies. Although governments encouraged consumer practices, centralized provisioning and poor quality tended to produce alienated consumption, which undermined the governments' legitimacy (Fehérváry, 2013). At the same time, contrary to the prevailing image of socialism as 'wasteful,' limited goods and incomes generated complex systems for reusing and recycling resources (Gille, 2007).

Shortage also affected two other crucial resources: information and labor. Party planners could implement top-down planning only if they had detailed and accurate information. This happened rarely, for lower level cadres had many incentives to falsify information so as to build careers by appearing successful, evade responsibility for disasters, inflate their materials requests, provide cushions against production shortfalls, overfulfill their targets to receive bonuses, etc. From these discrepant needs came a perpetual struggle over information. Higher authorities hoarded it as a source of power and engaged in continual information gathering at all levels of society (Horváth and Szakolczai, 1992). Censorship and close management of information at the top provoked information hunger below, producing a culture of rumor, gossip, selective secrecy, and conspiratorial explanations of events, as well as a citizenry skilled at reading between the lines. Classic socialist societies were thus a special kind of information society (Szakolczai and Horváth, 1991).

Labor Shortage and Rights in People

An additional resource subject to shortage was *labor*. Especially in countries with large peasantries, such as Romania, China, or Vietnam, the absence of a proletariat made it difficult to mobilize skilled industrial workers, and relative economic underdevelopment precluded both attracting labor through high wages and substituting capital-intensive methods. As industrialization proceeded, labor became short in the countryside. Moreover, the shortage economy encouraged firms to employ excess workers so they could complete monthly plans once sufficient materials were on hand (in a frenzied effort known as 'storming'); managers therefore padded their workforce, exacerbating shortage. Labor shortage thus had both macro- and microdimensions. Means for stabilizing the labor supply included making workplaces the locus of benefits such as daycare, housing, vacation permits, pensions, and medical care. Certain cities were closed to immigration, and

construction of urban infrastructures was limited (Szelényi, 1983), thus compelling millions to become village-based 'peasant workers.'

Within each industrial and agricultural workplace, managers had to ensure a labor supply adequate to both normal production and periods of storming. Workers, for their part, might bargain for better conditions by withholding labor, through intentional slowdowns or time off for household tasks and moonlighting. Conflicting demands for labor put a premium on ways of accumulating it through gaining *rights in people*. For managers, these included extending patronage and favors (access to schooling, houses, or building sites), overlooking petty illegalities, and securing loose plans that enabled them to produce a surplus they might use to create labor-securing debt and exchange relations (cf. Humphrey, 1983). Both managers and others needing labor might appeal to kinship idioms, emphasize ethnic similarity, participate in special rituals, and expand networks of reciprocity through gift-giving. Thus, the quest for labor further encouraged personal ties and reinforced particularistic identities.

Second Economies and Socialist Reform

Believing they alone could best determine how social wealth should be redistributed, Party leaders initially opposed economic activity not encompassed within plans. The inability of planning to cover social needs at the given level of technological endowment, however, compelled officials to permit and even legalize some small-scale private effort, known as the 'second' (or informal, unofficial, or shadow) economy. Among its forms were food production on small household plots, after-hours repair work or construction, typing, tutoring, unofficial taxi services, etc. Because these activities overlapped with semilegal and illegal ones, their situation was everywhere precarious, with authorities persecuting them more in some times and places than in others. The second economy was largest in Hungary after 1968, for example, small in Cuba until Castro's 'Rectification' of 1986, harassed in Romania throughout the 1980s; it burgeoned in post-1978 China; extreme forms are reported for the Soviet Union, where entire factories ran illegal production after hours. Crucial to understanding the second economy is that it nearly always utilized materials from the first (or formal, official) economy; therefore, state production units subsidized the second economy's much-touted higher rates of productivity. The prevalence of second-economic activity both indicated popular rejection of the Party's definition of needs and helped to fill those needs by voluntarily lengthening the working day.

Pressure from the second economy was but one of many signs of difficulty in socialist planning that led to repeated efforts at reform, initiated from both within and outside the Party. Beginning with Lenin's New Economic Policy, every socialist society experienced cycles of reform and retrenchment, devolution and recentralization (cf. Skinner and Winckler, 1969; Nee and Stark, 1989). Most exhibited a trend toward less stringent planning and the introduction of market mechanisms, heightened material incentives, and mixed property forms. System-wide experimentation began after Khrushchev's 1956 'Secret Speech,' criticizing Stalin, and increased as each society moved from extensive development (mobilizing resources) to the intensive phase (heightened productivity). Hungary and Yugoslavia introduced the most durable early reforms; those in the Soviet Union ended in the collapse of the Soviet bloc, while comparable reforms continued in China, Vietnam, and Cuba. As they reformed, socialist societies increasingly diverged not only from the Stalinist model but also from one another, introducing path-dependent differences that became ever more marked after 1989/1991, into the twenty-first century.

Despite these differences, the reforms everywhere redefined basic units of activity (revitalizing villages or households, for example, at the expense of collective farms, teams, or brigades); altered gender relations (usually in favor of men and patriarchal authority) and increased other inequalities; affected networks of reciprocity (expanding horizontal over vertical connections); dismantled at least some socialist property forms (as with Hungary's specialist cooperatives and China's decollectivization, begun in 1978); shifted the locus of authority (usually downward, provoking resistance from higher-level bureaucrats); and entailed new, more intimate forms of state control (implied, for example, in Chinese rituals that no longer imagined gods or ancestors as inhabiting a nether world but found them immediately present; see Mueggler, 2001).

Social Engineering and Resistant Personhood

All Communist parties promoted massive projects of social engineering. They accorded a privileged place to knowledge and expertise, which the Party claimed to monopolize. Convinced of the creative power of language, Party leaders both used and modified language in commanding ways that resulted in a standardized authoritative discourse. They spoke of things that had yet to be created – the working class, the proletarian dictatorship, the 'new socialist man' – as if those things already existed, and they developed hieratic speech with reduced vocabularies, clusters of noun phrases, and few (often passive) verbs, creating a limited, static verbal world. Party cadres intervened exhaustively in all facets of life, aspiring to create new moralities and to control populations in myriad ways. They attacked religion as superstition, instituted new socialist rituals, and expanded educational access to create enlightened citizens indifferent to religious belief. They filled the air with moral exhortations aimed at instilling a new, puritan, socialist ethic. Toward better management of resources, they introduced massive population programs, from Romania's enforced pronatalism to China's stringent one-child policy, insinuating themselves into people's most intimate lives to often devastating effect (e.g., Kligman, 1998).

Dedicated to building an alternative to the world of bourgeois capitalism, communist leaders announced an assault on all forms of inequality, including those based in kinship status, gender, class, and ethnonational difference. Society would be homogenized, lineage-based kinship structures broken up, gender differences minimized and authority relations in families altered, classes and private property abolished, national minorities given equal chances. New social entities emerged – state and collective farms, teams, and brigades, for instance – to take the place of kin groups and village structures. When preexisting social relations proved

difficult to abolish, they would be coopted and turned to the Party's purposes (e.g., Kligman and Verdery, 2011). From all these efforts, leaders imagined, a 'new socialist man' would emerge, a new kind of person, along with the 'people as One,' a new kind of human community, bearing an exalted socialist consciousness.

The results of all these interventions were decidedly mixed. To begin with, the Party failed to speak with one voice, as ministries vied for resources rather than cooperating to remold society, and as cadres built their careers by under- or overexecuting central directives and engaging in rampant personalism. In addition, the annexation of preexisting social forms inevitably modified the purposes for which they had been coopted. Exhibiting greater capacity to formulate goals than to implement them, socialist governments were continually plagued by the unforeseen consequences of their policies. To solve conflicts among nationalities, they institutionalized ethnonational difference in ways that served to strengthen it; they embraced nationalist idioms that then overtook socialist ones (Hirsch, 2005; Verdery, 1991). Parties created working classes only to find themselves opposed by their creation, as with the Polish authorities and the Solidarity movement.

Beyond this, resistance to social engineering proliferated even as people adopted the new forms. Constant surveillance politicized behavior: small acts of joke telling became willful anarchy in the eyes of both authorities and joke tellers. Dissidents sought support internally and abroad for their critiques of the system, launching projects to build 'civil society' against the state. Less visibly, peasant refusal to accept starvation from China's Great Leap Forward or Eastern Europe's delivery quotas ultimately forced authorities to renounce these policies (Thaxton, 2008; Rév, 1987), while popular insistence on higher living standards meshed with the parties' own consumption goals to press for reform (Fehérváry, 2013).

Highly complex political effects centered on the making of persons and political subjects. Redistribution and socialist paternalism tended to create subjects disposed to dependency and passive expectation; early campaigns taught the poor to see themselves as victims of oppression (rather than simply of fate) and offered them the Party's instrument of vengeance, in the form of denunciations. Thus, the Party both aroused people's complicity and instilled self-denigration. At the same time, however, many adopted a sort of dual consciousness, in which compliance with socialist forms and values coexisted with the pursuit of one's own independent plans (see Yurchak, 2006). Sometimes but not necessarily conceived as outright resistance, this doubling might accompany a pervasive dichotomization of the world into 'Us' (the people) and 'Them' (the authorities). More significantly, it encouraged subtle forms of self-making in people's own terms: they sought out Western goods, created self-respect through diligent second-economy work while loafing on their formal jobs, participated in ethnic- or kin-based identities and rituals constitutive of self, and gave gifts not just to secure advantage but to confirm their sociality as persons and human beings (see, e.g., Dunn, 2004).

The collapse of Soviet-style regimes complicates a balanced assessment of socialism, which once represented great hope for countless people. Undeniably, socialist societies diminished certain forms of oppression, yet they created others. In many of these societies, an early élan for the expressed goals gave way to cynicism and a feeling of betrayal, as the means used to achieve those goals sabotaged their fulfillment. Admirable gains in education, medical care, and social well-being counted for less against the devastating purges, labor camps, Party-induced famines, environmental degradation, and abuses of human rights. Top-down engineering foundered as socialism's leaders resisted their citizens' efforts to educate them about social relationships, self and human feeling, the value of ritual practices, and the limits to imposed change. The lessons of these citizens' refusals, their silent revolutions from below, should guide future attempts to create alternatives to the world of global capitalism.

For further reading, see *Centrally Planned Economies, Cold War, Communism, Lenin, Mao Zedong, People's Republic of China, Socialism, Soviet Union, Stalin*.

See also: Asia, Sociocultural Aspects: China; Europe: Sociocultural Overview; Inequality: Comparative Aspects; Patron–Client Relations, Social and Anthropological Study of; Political Anthropology; Political Economy in Anthropology; Property: Anthropological Aspects; State: Anthropological Aspects.

Bibliography

Bockman, J., Eyal, G., 2002. Eastern Europe as a laboratory for economic knowledge: the transnational roots of neoliberalism. American Journal of Sociology 108 (2), 310–352.
Casals, F.G., 1980. The Syncretic Society. M. E. Sharpe, New York.
Dunn, E., 2004. Privatizing Poland: Baby Food, Big Business, and the Remaking of Labor. Cornell University Press, Ithaca.
Fehérváry, K., 2013. Politics in Color and Concrete: Socialist Materialities and Middle-class Citizenship in Hungary, 1950–2000. Indiana University Press, Bloomington.
Gal, S., Kligman, G., 2000. The Politics of Gender after Socialism. Princeton University Press, Princeton.
Gille, Zs, 2007. From the Cult of Waste to the Trash Heap of History: The Politics of Waste in Socialist and Postsocialist Hungary. Indiana University Press, Bloomington.
Hershatter, G., 2011. The Gender of Memory: Rural Women and China's Collective Past. University of California Press, Berkeley/Los Angeles.
Hirsch, F., 2005. Empire of Nations: Ethnographic Knowledge and the Making of the Soviet Union. Cornell University Press, Ithaca.
Horváth, A., Szakolczai, A., 1992. The Dissolution of Communist Power: The Case of Hungary. Routledge, New York.
Hough, J.F., 1969. The Soviet Prefects: The Local Party Organs in Industrial Decision-making. Harvard University Press, Cambridge.
Humphrey, C., 1983. Karl Marx Collective: Economy, Society and Religion in a Siberian Collective Farm. Cambridge University Press, Cambridge.
Kligman, G., 1998. The Politics of Duplicity: Controlling Reproduction in Ceauescu's Romania. University of California Press, Berkeley.
Kligman, G., Verdery, K., 2011. Peasants under Siege: The Collectivization of Romanian Agriculture, 1949–1962. Princeton University Press, Princeton.
Konrád, G., Szelényi, I., 1979. The Intellectuals on the Road to Class Power. Harvester Press, New York.
Kornai, J., 1980. Economics of Shortage. North-Holland, Amsterdam.
Kornai, J., 1992. The Socialist System. Princeton University Press, Princeton.
Ledeneva, A.V., 1998. Russia's Economy of Favours: Blat, Networking, and Informal Exchange. Cambridge University Press, Cambridge.
Mueggler, E., 2001. The Age of Wild Ghosts: Memory, Violence, and Place in Southwest China. University of California Press, Berkeley/Los Angeles.
Nee, V., Stark, D., 1989. Remaking the Economic Institutions of Socialism: China and Eastern Europe. Stanford University Press, Stanford.

Rév, I., 1987. The advantages of being atomized. Dissent 34, 335–350.
Skinner, G.W., Winckler, E.A., 1969. Compliance succession in rural communist China: a cyclical theory. In: Etzioni, A. (Ed.), A Sociological Reader on Complex Organizations. Holt, Rinehart and Winston, New York, pp. 410–438.
Szakolczai, A., Horváth, A., 1991. Information management in Bolshevik-type party states: a version of the information society. East European Politics and Societies 5 (2), 268–305.
Szelényi, I., 1983. Urban Inequalities under State Socialism. Oxford University Press, Oxford.
Thaxton, R., 2008. Catastrophe and Contention in Rural China: Mao's Great Leap Forward Famine and the Origins of Righteous Resistance in Da Fo Village. Cambridge University Press, Cambridge.
Verdery, K., 1991. National Ideology under Socialism: Identity and Cultural Politics in Ceauescu's Romania. University of California Press, Berkeley/Los Angeles.
Verdery, K., 2003. The Vanishing Hectare: Property and Value in Postsocialist Transylvania. Cornell University Press, Ithaca.
Walder, A., 1986. Communist Neo-traditionalism: Work and Authority in Chinese Industry. University of California Press, Berkeley.
Wesolowski, W., 1979 [1966]. Classes, Strata, and Power. Routledge and Kegan Paul, London.
Yan, Y., 1996. The Flow of Gifts: Reciprocity and Social Networks in a Chinese Village. Stanford University Press, Stanford.
Yurchak, A., 2006. Everything Was Forever, Until It Was No More: The Last Soviet Generation. Princeton University Press, Princeton.

Sociality In Anthropology

Nicholas J Long, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Carrithers, volume 21, pp. 14500–14504, © 2001, Elsevier Ltd.

Abstract

Although the concept of sociality was championed in the late 1980s as an alternative to 'culture' or 'social relations' as the principal object of anthropological inquiry, confusion often surrounds the term due to a lack of consistency in the way it is deployed. Following recent attempts to rethink the nature of 'the social,' human sociality can be best understood as the dynamic matrix of relations through which persons come into being, and which is navigated by an ethically imaginative and affectively receptive human subject. The challenge facing the discipline is now not only to trace and explain the diverse forms of human sociality but to decide how to incorporate nonhuman socialities into anthropological research and to develop richer accounts of the developmental bases for human intersubjectivity and ethical imagination.

As has long been recognized, human beings are 'social creatures,' whose interactions with each other stand out across the animal kingdom not only for their complexity but also for their sheer variety. It therefore follows that anthropologists, as social scientists who are profoundly concerned with documenting and understanding the ways in which human behavior unfolds in particular settings, should want to ground their inquiries in a theoretical model that can account for why humans exist with others in the world in the specific ways that they do. In other words, anthropology needs a theory of human sociality. To date, however, the concept of 'sociality' has been used in a confusing variety of ways in anthropological writings – a phenomenon that reflects internal disciplinary disagreements over how to theorize 'the social,' as well as the influence of disciplines such as analytic philosophy, developmental psychology, and behavioral ecology, whose established concepts of 'sociality' map incompletely on to the full range of issues with which anthropological research is concerned. Nevertheless, recent efforts to synthesize these different strands of sociality research into a coherent theoretical framework have offered productive new ways of conceptualizing 'sociality,' whilst also raising some difficult questions that the discipline will need to address in the years to come.

The Scope of Sociality

One of the problems confronting anthropologists seeking to engage with the concept of 'sociality' is that it has frequently been defined in ways that are heavily, and inappropriately, circumscribed. A helpful case through which to illustrate this problem is Kaushik Sunder Rajan's discussion of social movements in Mumbai. Taking workers' movements and the collective mobilization of experimental subjects as examples, he suggests that such formations represent moments in which a particular subjectivity (the identification of oneself as a worker, or as an experimental subject) is 'transformed into a sociality' – which is to say it becomes the basis of a shared social identification (Sunder Rajan, 2008: pp. 159, 171). Sociality, in this view, is linked to existence in a (politically motivated) associational community, leading Sunder Rajan to assert that 'sociality cannot be taken for granted – it never already exists but has to be brought into being' (2008: p. 159). Although he has clearly identified an important phenomenon, Sunder Rajan's decision to label it 'sociality' creates confusion for readers who have elsewhere encountered the more widespread anthropological view that sociality is a foundational dimension of the human condition. And it does so with good reason, for on a substantive level, in what sense is a state of shared social identification any more of a 'sociality' than the frustrating, but still highly social, experience of trying to express one's experience to others and yet never being fully understood? What if one were to conceal or suppress one's status as, say, an experimental subject, but nevertheless socially identify with others as a resident of Mumbai, or as a citizen of India? That, by Sunder Rajan's definition would also be 'a sociality.' And so what if, regardless of what other identifications one might make, one recognized (implicitly or explicitly) that one lived in a world shared with fellow beings whom one could identify as similar to oneself – as other humans? Surely that would also be a 'sociality'? Very quickly, we are left with a situation in which the idea of an 'associational community' can be both highly specific and very broad. The challenge this presents is not to draw a line that can decide where the category of 'sociality' begins or ends, but rather to account for how any given human being can participate with others in the world in multiple ways (some circumscribed, and others less so), and very often in multiple ways at the same time.

Sunder Rajanian definitions must therefore fall, but it is revealing to note that similar attempts by anthropologists to define sociality as an observable, bounded, and discrete phenomenon, rather than as an intrinsic feature of human existence, have also tended to equate the term with distinctive features of human life in groups, or with strongly motivated forms of association. Adam Chau (2006: p. 147), for example, has defined sociality as the 'condition of social copresence,' when humans (who may or may not be related or interact) gather in the same social space, such as a cinema or a crowd, while Peter Westoby (2009: p. 2) suggests that sociality is the purposeful and agentive process of creating new social relations, as opposed to the everyday processes of people simply 'relating together.' While such definitional

circumscriptions of the term may admittedly reflect the preoccupations of people who live in the various locations under study, it is concerning that the anthropologists writing about them are so quick to advance these as contributions to the discipline's analytic terminology. Their tendency to do so may partly reflect the influence of the ways in which the term 'sociality' has been deployed within anthropology's cognate disciplines. Behavioral ecologists have typically defined sociality as 'group living,' seeking to excavate the neurological and pheromonal mechanisms that facilitate this, and exploring the evolutionary advantages that such forms of living might convey. Within the discipline of philosophy, meanwhile, several figures have emphasized that although human beings are capable of thinking and acting as individuals – the perspective from which most philosophical arguments begin – it is also possible for them to act from a shared 'we-perspective.' In Raimo Tuomela's *The Philosophy of Sociality*, this 'we-perspective' equates to 'the group's point of view,' an emphasis he defends on the grounds that human beings do not only live but are adapted to living in groups, with the existence of mirror neurons pointing to group conformism as a coevolved psychological disposition (Tuomela, 2007: pp. 1, 240). For Margaret Gilbert (1997: pp. 22–23), 'we-ness' need not involve an entire group – it could simply be the relationship between two people – but it should reflect a sense of 'plural subjecthood' grounded in a specific type of mental connection between persons – that achieved by a joint commitment. Arguing that this is a "distinctive and consequential phenomenon that helps structure the life of all human beings," Gilbert (1997: pp. 17, 29) strongly advocates restricting the scope of the term sociality to such instances of plural subjecthood, fearing that they might be obscured in 'the blur of a broader notion of sociality.'

Prima facie, such definitions can seem appealing, and it is true that all of the authors discussed above have identified intriguing and important phenomena: potential forms and dimensions of human sociality that are worthy of sustained investigation. Such work, however, is unlikely to prove fruitful unless the research is grounded in a broader theory of sociality in the first place. Where group-oriented and plural subject-oriented definitions of sociality of the kind outlined above founder most fundamentally is in their mistaken assumption that sociality is not relevant to those situations in which one is acting as an individual, is not in a group, is not in relation, or is alone. However, such states of being are emergent through sociality, and themselves constitute a form of sociality. As Theodore Schatzki has argued in a discussion of the early Heidegger, to acknowledge that humans are social beings requires acknowledging that human existence is always 'being-with.' In whatever circumstances one finds oneself, there are always others who have a 'bearing' on one's life. Thus "being alone …, far from being an asocial condition, is a way someone who coexists can be, a particular relationship with others. In using a tool, moreover, one coexists with the merchant from whom it was bought, just [as] in opening a gift one coexists with the giver" (Schatzki, 2005: pp. 233–234). From this perspective, phenomena such as the participation in a political movement grounded on the common experience of being an experimental subject, or feeling a thrill in a crowded cinema, or experiencing a mental connection through a joint commitment with a partner, friend, or colleague are all ways in which other people could bear on one's life (and in which one could have a bearing on theirs). However, to crown any of them as quintessential sociality would not only be an arbitrary delimitation of human social experience, it would serve to distract from the many other important (if often less spectacular) ways in which living with others conditions human life. Moreover, it would ignore the fact that one is only ever in situations where one could join a political movement or establish a joint commitment, etc., as a result of previous ways of existing with others in the world. An anthropological theory of sociality thus needs to be able to account for these dramatic and remarkable ways of existing with others in the world, without restricting itself to them.

For these reasons, anthropology is better advised to adopt a broad and encompassing definition of sociality, recognizing it as the dynamic and continually emergent matrix of relations within which subjects are continually interacting and bearing upon each other in a process of coproduction. Such a position was first advanced by Marilyn Strathern and Christina Toren at the 1989 Manchester debate in anthropological theory, on the question of whether the concept of 'society' was theoretically obsolete. Arguing that it was, they instead emphasized the value of seeing the anthropological endeavor as an inquiry into sociality: "the relational matrix which constitutes the life of persons." In this view, humans are continually coming into being through dynamic social processes, rather than being 'socialized' into independent sets of rules, structures, customs, or meanings – and it is the documentation of such processes that is the proper task of anthropology (Strathern et al., 1996). Sociality itself, then, is not a dimension of the human experience divorceable from others. Rather it denotes the state of existence in a dynamic relational matrix, something that is intrinsic to being human.

This more dynamic and processual understanding of anthropology's principal object approach proved appealing to many in the discipline, who were growing increasingly dissatisfied with the static and totalizing notions of 'culture' that lay at the heart of much of the structuralist and interpretivist anthropologies produced during the 1960s, 1970s, and 1980s. It also proved a potent antidote to the emphasis on tracing formal patterns of 'social relations' that had lain at the heart of British structural functionalism for much of the twentieth century. Powerful as this initial formulation was, however, there are several points on which it has needed to be updated in the light of two related theoretical advances made during the 1990s and 2000s (Long and Moore, 2013).

The first is the actor-network theory of Bruno Latour (2007), which has highlighted that it is not just 'persons' or 'humans' that have a bearing upon each other – rather, social life is held together by a whole host of human and nonhuman 'actants.' An illustration of this point comes from the work of sociologist Vikki Bell (2001) in her discussion of the relationships that the children of separated parents have with their nonresident fathers. Although the experiences of the families she discusses are very diverse, in every case the father–child relationship is mediated by the technologies they use to keep in contact, principally the telephone. The fact that a phone can be handed to a child by his or her mother, sometimes accompanied by a wearied announcement of 'it's your

father,' that mobile phones are portable and can be used any time, and that phone calls interrupt the lives of the households they are calling and so need to be timed and spaced appropriately, have a considerable impact on how nonresident fathers conduct relations with their children and their former partners. One of the implications of Bell's analysis is that it is the phones themselves that lead to nonresident fathers becoming a different kind of father, and so we cannot speak of a changing relationship between the humans involved without taking into account the material technologies that mediate the relationship. In this sense, then, the families she discusses are assemblages of both human and nonhuman entities, with the nonhuman also being part of 'the social.' Clearly, this point needs to be incorporated into any definition of sociality, such that the relational matrix it speaks of is not simply one comprised of 'persons' but rather a much more diverse, heterogeneous mix.

A second crucial supplement to earlier formulations of sociality comes from recent discussions of 'affect theory' within cultural studies. These have emphasized that entities do not just bear upon each other through forms of 'social relation.' Phenomena as diverse as the rush of excitement that one can get from being in a crowd, an instant sense of attraction or disgust toward another person, and the state of being 'on edge' that a lone pedestrian may feel in a dangerous neighborhood all reflect the ways in which bodies and environments can transform how humans feel, perceive themselves, and act in the world despite there being no 'relations' to speak of between them. A variety of models have been proposed to account for how these bodily states arise (which range from the postulation of wholly autonomous affects circulating in the world to approaches which ground affective experience in social structures of power and meaning), with no resolution having yet been achieved within the discipline. Despite such ongoing debates, however, it is clear that – whatever their origins – such affective 'bearings' need to be included within contemporary understandings of sociality, rendering the concept yet more expansive in its scope.

However, once the concept has been revised to incorporate recent efforts to 'reassemble' the social, one might reasonably ask whether there is anything more to sociality than noting the mere fact of humans' existence in a relational world, and indeed whether anthropology even needs a concept of 'the social' or of 'sociality.' This is a question on which anthropologists are divided. Those who lean toward dispensing with such categories include Tim Ingold who in 1997 published a searing refutation of what he saw as the prevailing model of 'sociality' within the discipline of anthropology. Characterizing that position as the view that "the basis for society in the strict sense [comes from] the capacity, apparently unique to our species, to render the world meaningful in terms of representations that may be both assembled in the mind and communicated verbally to others," he argued that anthropologists needed to recognize that all human ideas about the world are "fashioned against the background of our active engagement with its diverse human and nonhuman constituents" (Ingold, 1997: p. 232). It thus made no sense, he argued, to privilege some relations as being more 'social' than others. "If some relations are social," he argued, "then all are, and all life would be social life. What need have we, then, for a concept of the social at all?" Of course, once the term 'social' had been abandoned, then the potential analytical utility of a term that means 'being social' – sociality – does not look great.

However, while most contemporary anthropologists would probably agree that Ingold is correct to oppose an exclusively anthropocentric notion of sociality, the recognition that all life is social life does not have to be as trivial as he supposes. If the concept of sociality denotes the state of existence in a dynamic relational matrix, then the simple fact of one's presence in the matrix is only one aspect of what is at stake in the term. The other half, neglected by Ingold, concerns the character of that existence, the ways in which entities behave within, and toward, the matrix of relations in which they are emplaced. In other words, although Latourian and Ingoldian interventions have made it possible to recognize all entities in the world as being social, and thus as embroiled in sociality, this does not mean that they all have the same socialities. The current challenge for anthropology is thus to identify what (if anything) is distinctive about the sociality of human beings, as well as to work out how the socialities of other entities with which human beings interact might be properly accounted for within anthropological analysis.

The Character of Human Sociality

Close your eyes and imagine a crowd of zombies. What do you see? As Shaka McGlotten (2011: p. 187) has argued, zombies are "frequently imagined together as a mass, a crowd, or a swarm, yet they remain alone even among others … . Their sociality differs in that they do not possess the reflective self-awareness or empathetic identification we take as the hallmarks of meaningful intimate connection with yourself and others. This … is part of what makes them inhuman."

These remarks offer some suggestive initial entry points for a discussion as to what might be distinctive about human sociality. This distinctiveness can partly be sketched in terms of the remarkable communicative and interactional capacities or abilities with which humans are endowed: capacities that certainly shape the character of human sociality but, as I shall explain shortly (and *contra* certain widely peddled definitions of 'sociality' as the 'capacity to be social') are not the sum total of it. The ability that is most highly developed in humans (but not in zombies) is intersubjectivity – something that lies at the heart of the process of human–human interaction. Although the term 'intersubjectivity' has, like sociality, been defined in a variety of ways, Alessandro Duranti (2010: p. 10) offers a particularly eloquent assessment of how the term might be understood within the anthropological project, suggesting that it represents the forms of awareness that one might have of "others who see the world, nature included, from their standpoint but, by manifesting behaviors that resemble our own, contribute to the making of one objective – shared or sharable – world possible." Particularly, appealing about Duranti's exposition of the concept is its recognition that intersubjectivity is a condition of possibility out of which many different potential 'worlds' might be emergent. As such, one of the most exciting and distinctive aspects of human interaction is its fundamentally open-ended and ever-emergent quality.

The possibility of intersubjectivity rests upon interactants having a 'theory of mind,' which allows them to attribute

beliefs, intentions, motivations, and even false beliefs to the people they are interacting with, such that they can then behave accordingly. This equips humans with a remarkable range of interactional possibilities, because it allows us to establish channels of communication with any human who also possesses a theory of mind, regardless of their lack of shared language, culture, or history. The processes by which such a theory of mind is acquired by humans, however, are still in the process of being fully delineated. Experimental evidence suggests that although even very young infants of around 12 months in age are capable of taking others' beliefs and intentions into account when acting, more complex aspects of theory of mind – such as the ability to recognize that an interlocutor might be acting under the influence of a belief that one knows to be false – are unlikely to emerge in children under the age of four years. What the precise developmental steps are, and how universally they are followed, remain frontiers of research.

In the absence of conclusive evidence, a number of theories have been advanced to account for how human capacities for interaction arise. The most fundamental point of disagreement is at what level these capacities are emergent and/or manipulable. One possibility, advanced by Enfield and Levinson (2006) is that humans are biologically endowed with an 'interaction engine' that incorporates theory of mind capabilities, biological constraints, and ethological proclivities (such as the mechanisms for conversational turn-taking or the correcting of errors and misunderstandings) and which is then 'fine-tuned' to a local cultural frame through the interactional matrix in which human development actually occurs. This view finds some support in the fact that even very young babies appear to participate in cooperative behavior (an apparent 'instinct' for cooperation) and in regularities of conversational dynamics across diverse settings, which supports the idea that they may have an ethological base (Schegloff, 2006).

Others are suspicious of such a view. Quite aside from the objection that the makeup of any given human's psychology and proclivities is always-already determined by interactional matrices – not least the one which led to their conception (Toren, 2012) – the fixed character of an 'interaction engine' seems at odds with the general behavioral flexibility that humans exhibit in most other areas of their lives. By that logic, one would expect the interactional matrix in which the child grows up to be the more significant factor for their capacities for interaction. Certainly, many sociocultural anthropologists would be naturally drawn toward the possibility that contextually variable representations of 'the mind' could have a significant effect on mental experience, and on the theories of mind that individuals hold (Astuti, 2012). Such a position, however, raises the question of why it is that despite widely differing practices in child-rearing, developmental outcomes appear to be largely the same across children around the world (Gaskins, 2006). One possibility is that the threshold for a 'typical' developmental outcome is set sufficiently low that the exposure to diverse experiences can allow broadly similar interactional capacities to develop in different settings. Another possibility is that the similarity of developmental trajectories has been overstated by researchers who have largely confined their work to occidental settings and/or whose methodologies have led them to overlook subtle but important variations in children's interactional skills.

What is beyond doubt, however, is that certain aspects of the interactional matrix can be definitive factors in determining people's capacities for social behavior. One of the most significant of these is exposure to language, and in particular linguistic terms related to theory of mind, such as the verbs 'to think' and 'to know.' The potential developmental role of language has been explored through a large number of studies with deaf people, including those who were only introduced to sign language in adulthood. Although current evidence suggests that deaf individuals do possess a theory of mind regardless of their degree of exposure to language, late language learners experience persistent difficulties with either reasoning about or expressing an understanding of the nature of ignorance, deception, and false beliefs in ways that threaten to foreclose certain possibilities for interaction (Morgan and Kegl, 2006).

Conversely, while individuals with autism were once (and sometimes still are) seen as quintessential examples of an 'asocial' state of human being, recent anthropological writings have recast autistic people's possibilities for social coordination as being promoted and impeded by the sociocultural ecologies they inhabit rather than as wholly determined by the neurological 'impairments' that seem to prevent them from developing a typical theory of mind. Thus, Ochs and Solomon (2010) suggest that one of the primary challenges for autistic people in establishing social relationships with others is a mismatch in their repertoires of social coordination. While nonautistic humans routinely interact face-to-face, this can lead to an experience of sensory overload for severely autistic children and prove less facilitative of social interaction than sitting side by side. Similarly, a slow, affect-laden 'baby talk' register of the kind commonly used to interact with small children appears to be counterproductive for autistic children, who respond better to moderately paced speech and restrained affect. In a separate paper, Solomon (2012) has explored how interactions with therapy dogs can be used to 'unlock' autistic children's capacities to play and become responsive social beings. Objects and activities from therapy sessions with the dogs (such as throwing a tennis ball) are then incorporated into other settings such as home and school, in the hope that this may facilitate responsiveness to and with other humans. What all these examples point to that it is the dynamic relational matrices in which autistic individuals are placed, and not just their neurology, that delimit the opportunities for social coordination available to them at any moment in time and going forward into the future.

This example from the anthropology of autism also serves to highlight a more general feature of human sociality, namely that dynamic matrices of relations can be subject to conceptualization, scrutiny, evaluation, and transformation by the very subjects who are involved in it. When parents and therapists (and anthropologists) attempt to understand the nature of autism and change the way they relate to the autistic children in their lives, they are doing so with a deliberate goal: to change the matrix of relations in which they are embedded, and to do so for the better. The ethnographic record is replete with accounts of similar endeavors, such as the creation of intentional communities or the instigation of revolutions. That is

not to say, however, that a matrix of relations always has to be transformed. Subjects might feel equally compelled to preserve or maintain it. But in each case, the trajectory that the relational matrix takes is determined by those embedded within it, through acts of what Henrietta Moore (2011: p. 16) has termed 'cultural invention.' What is significant about this observation is not merely the fact that sociality is capable of changing over time, but rather the possibility that it, and the impulses underpinning it, might be at least partly amenable to control.

The capacity for cultural invention is one of the most distinctive aspects of how humans inhabit matrices of relations. As far as we are aware, no other animal (and certainly no zombie) has the capacity to reflect in a sustained manner on their position in the world, to imagine a better way of being, and to then strive to implement it. The remarkable cognitive abilities that human beings possess, including their capacities for virtuality and imagination and their behavioral plasticity – all features associated with being biologically cultural – lie at the heart of this striking feature of human sociality. And yet there are several aspects of the process of cultural invention that warrant further explanation. The first is that we are 'culturally inventing' all of the time. This does not mean we are always busy planning how we will implement our visions of the good life. But every act we take – from deciding to phone a friend on a whim, or declining to strike up conversation with a stranger in a library, shapes the matrix of relations we inhabit. Moreover, it is precisely the fact that most of these choices are not planned or consciously thought-through but rather seem impulsive or instinctive that makes them so worthy of attention. This suggests that anthropological understandings of human sociality need to incorporate three distinct features. The first is the full extent of the dynamic matrix in which any given human being must of necessity inhere. The second is what has sometimes been glossed as 'the human capacity for social behavior' – but which might more precisely be understood in terms of the possibilities for and constraints upon social coordination that are afforded by both the biological makeup of particular humans (a point which includes species-level features and individual developmental trajectories) and the sociocultural ecologies they inhabit. To those, we now need to add a third dimension – one which can account for how a person's capacities to act within a particular matrix are deployed over their lifecourse. As Henrietta Moore and I have argued elsewhere (Long and Moore, 2013: p. 11), this requires a sophisticated account of agency, motivation, intentionality, and desire: in short, an understanding of the human subject as he or she exists in a given place and time, and more specifically *the ethical imagination*: "the forms and means … through which individuals imagine relationships to themselves and to others" (Moore, 2011: p. 16).

One of the reasons that Moore's conception of the ethical imagination proves particularly helpful for thinking about the character of human sociality is that it brings together diverse ways of thinking about what motivates the human subject, recognizing that all of them might have a role to play in accounting for how people navigate the relational matrices in which they are emplaced. Conscious reflective ethical practice of the kind attended to by Michel Foucault is clearly an important part of ethically imaginative endeavor – but so are the forms of identification and desire that proceed through fantasy, or through affective, embodied experiences. Subjects are not only motivated by their own perceptions of what is ethical, but also by fantasized imaginations of distant others, and of the kinds of person that they themselves will become when in relation with such others – as seen, for example, in the desire for 'Western' styles, goods, and partners in many parts of the non-Western world. Equally, practices of relating might be underpinned by more affective, embodied compulsions, as in McGlotten's (2011: p. 189) analysis of gay Americans who explained that they logged on to online cruising sites in order to experience 'a connection to the world' and thereby avoid 'feeling dead.' As these examples show, none of the components of the ethical imagination is free from the influence of broader social imaginaries, and so the ways in which sociality is navigated by particular humans will always reflect their historicity as a product of a long history of social relations in their past. Moreover, while every human being might be capable of ethical imagination, the capacity to act upon this is clearly unequally distributed, with some people in a dynamic matrix of relations being in a much stronger position to determine the conditions of possibility for others than vice versa. And since both ethical reasoning and fantasmic identifications can be profoundly affective, the question of how various dimensions of the ethical imagination interrelate remains a vibrant area of research. Nevertheless, and despite the wave of recent proclamations that anthropology has entered the era of the 'posthuman,' it appears that attending to the very distinctive qualities of *anthropos* and of human sociality will still be crucial for understanding the ways in which heterogeneous networks of entities are inhabited and navigated.

Pressing Agendas for Sociality Research in the Years to Come

Although the alleged distinctiveness of human–human interaction has been the topic attracting most attention from anthropologists concerned with sociality, the recognition that humans are always embedded in heterogeneous networks replete with nonhuman others has started to raise anthropological interest in the character of nonhuman socialities. For example, as Susanne Küchler (2013) has argued, recent innovations within materials science, notably the development of 'responsive' and 'intelligent' materials (such as self-cleaning glass, or containers that monitor the freshness of food) have made the socialities of materials increasingly complex and worthy of investigation in their own right, since the decentered cognition incorporated into their design gives them a capacity to recast and even navigate the networks of relations into which they are incorporated. However, although a growing body of anthropological work is taking note of materials in their specificity, the most vibrant current area of debate surrounds the question of how nonhuman animals interact with each other and with the humans in their lives.

The concept of affect – one of the central dimensions of Moore's conception of the ethical imagination – has proven particularly alluring for thinking about such relations. As noted above, affect points not to an interiorized emotional state but rather to a bodily state; a felt intensity. The significance of this is

that unlike the category of 'emotions,' which presupposes a certain and very human model of subjectivity and consciousness, the only prerequisite for experiencing affective states is a certain degree of sentience (the threshold of which, however, remains open to debate). This means that surges of affect can be shared by humans and nonhumans, affording the possibility of a form of intersubjectivity (or interaffectivity, if you will) that crosses species borders. As Rheana Parreñas (2012) has argued in a provocative recent ethnography of an orangutan rehabilitation center in Borneo, the affect of interspecies encounters might play a crucial role in shaping both human and nonhuman subjectivities. Yet as this self-same ethnography illustrates, salient differences across the species boundary remain. Human keepers are expected to be sensitive to the possibilities of affect, anticipating them and regulating them so as to keep orangutans 'wild.' This very injunction serves to illustrate the key difference between human and orangutan sociality: whilst humans are capable of virtuality, of planning and transforming the affectively laden relational matrices in which they act, orangutans are at the mercy of affect: as Parreñas (2012: p. 678) herself asks, "how can the orangutan not help but be enculturated by the affect produced in the copresence of humans?"

A good question, and one which reveals the depths of the challenges facing researchers currently striving to develop a multispecies anthropology sensitive to the often complex ways in which nonhumans navigate relational matrices of which humans are a part. For despite recent injunctions to engage with the 'alterworlds' of nonhuman species (Kirksey and Helmreich, 2010: p. 553), it remains very unclear quite how this could be done. Behavioral ethology offers points of departure for understanding the 'instincts' underpinning various forms of animal behavior, but remains oblivious to the forms of affective charge that anthropologists such as Parreñas are finding so analytically productive. This raises the question of whether anthropological theories of affect needs to be inserted into behavioral ecology, whether the capacity for experiencing affective surges might only be limited to (or highly developed within) certain species – for example, higher primates, or species that have been selectively bred to serve as companions to humans– or whether 'affect' is in fact only ever experienced by the human in the encounter, and denotes a distinctively human way of experiencing and acting in the world.

Finally, although the character of nonhuman sociality offers an intriguing and perhaps tantalizingly 'exotic' new frontier in anthropological inquiry, much about the character of human sociality remains unknown to us, including the precise developmental mechanisms that afford and foreclose possibilities for interaction, and the origins of the often complex desires and motivations that prompt us to manipulate our matrices of relations in particular ways. While decades of anthropological research has heightened our attention to the ways in which human behaviors and goals might be influenced by political and economic circumstances in which they exist, the precise psychodynamics that underpin such transformations remain open to more detailed investigation. Furthermore, new studies in the emerging subfield of the 'anthropology of precarious life' (Dave, 2012; Wool, 2012) are confronting us with the question of whether certain existential predicaments such as life in abject poverty, or recovery from shattering trauma, might engender distinctive registers of ethical imagination – such as those based on fleeting affections, rather than belief in durable attachments – and so point to important environmental influences on the character of sociality well into adulthood that warrant further theoretical excavation.

It is thus clear that much anthropological research into sociality remains to be done. The principal accomplishment of the past few decades has been to establish the relevance of a concept of 'sociality' – an ongoing, dynamic, and coconstitutive matrix of relations within which persons are forever in a state of becoming – as the relevant object of anthropological inquiry, displacing more its static predecessors such as the concept of 'cultural-' or structural functionalist models of norm-governed 'social relations.' Theoretical developments within anthropology and cognate disciplines such as the sociology of science and cultural studies have helped to demonstrate both the scope and the foundationality of the concept. The priority for anthropologists in the years ahead will be to collaborate with colleagues in related disciplines ranging from psychoanalysis and developmental psychology to behavioral ecology and materials science to understand in more detail the precise mechanisms by which humans and nonhumans alike come to navigate the relations that they share with others and thereby contribute to the maintenance or transformation of the forms of sociality that saturate our world.

See also: Affect and Emotion, Anthropology of; Intercultural Communication: An Anthropological Perspective; Mind, Theories of; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Astuti, R., 2012. Some after dinner thoughts on theory of mind. Anthropology of This Century. http://aotcpress.com/articles/dinner-thoughts-theory-mind/ (accessed 26.12.12).

Bell, V., 2001. The phone, the father, and other becomings: on households (and theories) that no longer hold. Cultural Values 5, 383–402.

Chau, A.Y., 2006. Miraculous Response: Doing Popular Religion in Contemporary China. Stanford University Press, Stanford.

Dave, N.N., 2012. Posthumanism in the postcolonial city. Paper read at AAA Annual Meetings, San Francisco, 14–18 November.

Duranti, A., 2010. Husserl, intersubjectivity and anthropology. Anthropological Theory 10, 1–20.

Enfield, N.J., Levinson, S.C., 2006. Introduction. In: Enfield, N.J., Levinson, S.C. (Eds.), Roots of Human Sociality: Culture, Cognition and Interaction. Berg, Oxford and New York, pp. 1–35.

Gaskins, S., 2006. Cultural perspectives on infant-caregiver interaction. In: Enfield, N.J., Levinson, S.C. (Eds.), Roots of Human Sociality: Culture, Cognition and Interaction. Berg, Oxford and New York, pp. 279–298.

Gilbert, M., 1997. Concerning sociality: the plural subject as paradigm. In: Greenwood, J.D. (Ed.), The Mark of the Social: Discovery or Invention? Rowman and Littlefield, Lanham, pp. 17–36.

Ingold, T., 1997. Life beyond the edge of nature? Or, the mirage of society. In: Greenwood, J.D. (Ed.), The Mark of the Social: Discovery or Invention? Rowman and Littlefield, Lanham, pp. 231–252.

Kirksey, S.E., Helmreich, S., 2010. The emergence of multispecies ethnography. Cultural Anthropology 25, 545–576.

Küchler, S., 2013. Materials and society. In: Long, N.J., Moore, H.L. (Eds.), Sociality: New Directions. Berghahn, Oxford and New York, pp. 177–189.

Latour, B., 2007. Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press, Oxford.

Long, N.J., Moore, H.L., 2013. Introduction: sociality's new directions. In: Long, N.J., Moore, H.L. (Eds.), Sociality: New Directions. Berghahn, Oxford and New York, pp. 1–24.

McGlotten, S., 2011. Dead and live life: zombies, queers, and online sociality. In: Boulk, S., Lenz, W. (Eds.), Generation Zombie: Essays on the Living Dead in Modern Culture. Jefferson, NC and London, pp. 182–193.

Moore, H.L., 2011. Still Life: Hopes, Desires and Satisfactions. Polity Press, Cambridge.

Morgan, G., Kegl, J., 2006. Nicaraguan sign language and theory of mind: the issue of critical periods and abilities. Journal of Child Psychology and Psychiatry 47, 811–819.

Ochs, E., Solomon, O., 2010. Autistic sociality. Ethos 38, 69–92.

Parreñas, R.S., 2012. Producing affect: transnational volunteerism in a Malaysian orangutan rehabilitation center. American Ethnologist 39, 673–687.

Schatzki, T.R., 2005. Early Heidegger on sociality. In: Dreyfus, H.L., Wrathall, M.A. (Eds.), A Companion to Heidegger. Blackwell, Malden, MA and Oxford, pp. 233–247.

Schegloff, E.A., 2006. Interaction: the infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In: Enfield, N.J., Levinson, S.C. (Eds.), Roots of Human Sociality: Culture, Cognition and Interaction. Berg, Oxford and New York, pp. 279–298.

Solomon, O., 2012. Doing, being and becoming: the sociality of children with autism in activities with therapy dogs and other people. Cambridge Anthropology 30, 109–126.

Strathern, M., Peel, J., Toren, C., Spencer, J., Ingold, T., 1996. The concept of society is theoretically obsolete. In: Ingold, T. (Ed.), Key Debates in Anthropology. Routledge, London, pp. 55–98.

Sunder Rajan, K., 2008. Biocapital as an emergent form of life: speculations on the figure of the experimental subject. In: Gibbon, S., Novas, C. (Eds.), Biosocialities, Genetics and the Social Sciences. Routledge, London and New York, pp. 157–187.

Toren, C., 2012. Imagining the world that warrants our imagination – the revelation of ontogeny. Cambridge Anthropology 30, 64–79.

Tuomela, R., 2007. The Philosophy of Sociality: The Shared Point of View. Oxford University Press, Oxford.

Westoby, P., 2009. The Sociality of Refugee Healing: In Dialogue with Southern Sudanese Refugees Resettling in Australia. Common Ground, Seaholme.

Wool, Z.H., 2012. Indurable sociality: precarious life in common and the temporal boundaries of the social. Paper read at AAA Annual Meetings, San Francisco, 14–18 November.

Socialization and Education: Theoretical Perspectives

Klaus A Schneewind, Ludwig-Maximilians-University Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, first, four perspectives in studying the processes of socialization and education, that is, chronological, topical, institutional, and cultural points of view, are briefly addressed. Then, three major groups of theories, namely biological theories (sociobiology and evolutionary psychology), psychological theories (psychoanalytic, attachment, epigenetic, and learning theories), and sociological theories (structural functionalism, symbolic interactionism, and conflict theories), are described in greater detail. Finally, an integrative framework for studying socialization and education is offered based on five aspects: units of environmental influence, characteristics of individuals, proximal processes, goals and values, and self-socialization.

Socialization refers to all processes by which an individual learns the ways, ideas, beliefs, values, and norms of his particular culture and adapts them as part of his own personality in order to function in his given culture. Education, on the other hand, deals explicitly with intentional attempts at influencing an individual's developmental course. In the following, some core anthropological assumptions guiding the thinking and theorizing about socialization and education are described. Second, major units of analysis, in particular phases of individual development, topics of socialization, socializing institutions, and cultural variations, are briefly addressed. Third, an overview of three groups of theoretical approaches to socialization and educational processes, that is, biological, psychological, and sociological theories, is presented. Finally, some elements of an integrative theoretical framework for studying socialization and educational processes are outlined.

Core Anthropological Assumptions

At the heart of all theorizing about socialization and education are central prototheoretical beliefs concerning the nature and development of human beings. In fact, as Montada (2008) has pointed out, there are at least four distinguishable models of individual development depending on whether the individual and/or his environment are conceived of as playing an active or passive part. Besides endogenistic, exogenistic, and self-creation theories exists a fourth group of theories that views both person and environment as active agents.

The latter group of theories rests on the common assumption that individual development is a dynamic interactional process that evolves in a sequence of reciprocally patterned person–environment transactions. Besides this basic assumption, there are at least six additional aspects that are central to a transactional conceptualization of human development, namely (1) learning capacity, (2) social dependence, (3) experience formation, (4) internal representations of experiences, (5) action competence, and (6) freedom of choice. Action competence and freedom of choice, in particular, are indispensable prerequisites for innovation and change concerning individual and sociocultural conditions of life.

Perspectives in Studying Socialization and Education

When looking at socialization and educational processes from an empirical point of view there are at least four different, albeit interrelated, approaches that can be distinguished, that is, a chronological, topical, institutional, and cultural approach.

In a *chronological perspective*, the course of individual development can be divided in different age-graded phases such as infancy, childhood, adolescence, and early, middle, and late adulthood. The sequence of phases and corresponding transitions is structured by individual developmental tasks (Havighurst, 1953) that, on the one hand, are based on physical and social cognitive maturation and, on the other, on normative expectations prevalent in a given society.

From a *topical perspective*, socialization processes have a strong impact on certain life domains such as (1) physical growth and health; (2) cognitive, social, moral, and spiritual development; (3) sex and gender differences; and (4) self and personality development.

From the vantage point of *institutions*, socialization takes place in different settings that, in some instances, are especially designed to advance the development of a broad spectrum of individual dispositions and competencies, either by formal or informal learning processes, that is, families, schools, peer groups, work settings, and so on.

Finally, from a *cultural* point of view, an individual's development depends a great deal on the values and beliefs that are typical of the culture to which he or she belongs. It has been shown, for example, that the prevalence of an independent-autonomous conception of the self within individualistic cultures versus an interdependent-relational self within collectivistic cultures is closely related to distinct socialization goals and practices stressing either autonomy or relatedness (Triandis, 1995).

Major Theoretical Approaches to Socialization and Education

Although in the present context it is not possible to give a full account of the vast array of theories dealing with the structural and dynamic aspects of socialization and education, three

major classes of theories, that is, biological, psychological, and sociological, are briefly reviewed (see the handbook edited by Grusec and Hastings, 2007, for detailed information).

Biological Theories

Sociobiology and Evolutionary Psychology

Based on Darwin's groundbreaking ideas concerning the phylogenetic evolution of the species, sociobiology and, somewhat later, evolutionary psychology contended that ultimate explanations of human behavior must take into account the superior goals of securing and expanding the successful reproduction of the human species. Especially mate selection and parental investment have been associated with universally evolved psychological mechanisms that are rooted in the human genotype as a consequence of phylogenetic adaptation (Johns and Belsky, 2008). For example, Belsky et al. (1991) have proposed and also presented some research evidence for an evolutionary theory of socialization relating different kinds of reproductive strategies to specific childhood experiences, which in turn entail specific outcomes of an individual's psychological and somatic development. More specifically, these authors argue for two distinct developmental pathways of reproductive strategy. One is characterized by stressful contextual and rearing environments, insecure attachments, early pubertal development, and onset of sexual activity within the context of short-term, unstable pair bonds, which in turn predispose to higher fertility rates and limited parental investment. In contrast, the other developmental trajectory shows the opposite pattern of rearing conditions and interpersonal relationships.

Behavior Genetics

While evolutionary theories focus on universal adaptive strategies, behavior genetic theories concentrate on explaining the variability of individual differences that can be accounted for by genetic and/or environmental influences (Plomin et al., 2008). Behavior geneticists, in particular, have shown that individual differences in cognitive and socioemotional dispositions are not only substantially influenced by the individuals' genotype but also by genetic components of the environment (e.g., via the parents' genotype impacting on their children's environment). Moreover, besides the conceptually and empirically useful distinction between shared and non shared environments, various kinds of genotype–environment covariation and interaction have been distinguished to account for genetically informed processes of individual development. In this vein, empirical analyses of cross-sectional and longitudinal data have led some authors to challenge the environmentalist position according to which, for example, familial influences such as parenting style or family climate have a strong impact on the development of individual differences in the filial generation (Moffitt and Caspi, 2007).

Psychological Theories

There is a wide range of psychological theories that have been applied to describing, explaining, and controlling socialization and educational processes. In fact, all theories dealing with personality and social development have much to contribute to this goal. In the following, four theoretical approaches are briefly described with a special emphasis on the processes of socialization and education.

Psychoanalytic Theories

In classical Freudian psychoanalytic theory the operation of defense mechanisms such as identification or sublimation plays a central role in socializing the growing individual's initially untamed libidinal and aggressive drives. Especially, coping with the oedipal situation via identification with the aggressor is supposed to be a developmental milestone, because it allows the young boy or girl to enormously expand his or her learning experiences, thus leading to a gradually more realistic and differentiated satisfaction of formerly more primitive expressions of his or her drive impulses. Later on, sublimation is the most elaborate and mature defense mechanism, which provides the individual with new ways in his private or professional life to follow through with his otherwise unsatisfied libidinal and aggressive needs at a socially acceptable level.

Fixation is another important psychoanalytic concept that has been invoked to explain character formation at different stages of an individual's psychosexual development, depending on either a lack, inconsistency, or abundance of gratifications at a particular developmental stage. Thus, distinct patterns of socialization practices at different psychosexual stages have been related to particular outcomes of children's personality development. Cultural anthropologists, in particular, have followed this line of reasoning in contrasting more lenient versus strict socialization practices in various cultures to demonstrate, albeit with only partial success, their impact on character formation (e.g., Whiting and Child, 1953).

While the classical Freudian psychoanalytic model is based on the transformation of primitive into more differentiated ways of need satisfaction, neoanalytic theories, especially object relations theories, approach the issue of socialization and personality development quite differently. In object relations theories the central dynamic principle is a person's need for interpersonal relatedness (Greenberg and Mitchell, 1983). In a developmental perspective, the individual acquires internal representations of the self, the other, and the relationship between self and other, which tend to become more and more differentiated. Thus, the quality of relationships that an individual is engaged in appears to be vitally important with respect to his internal representations of self and other (Kernberg, 1984). Empirical studies, however, are scarce and have mostly been limited to the pathology of object relations.

Attachment Theory

Although object relations theories bear some resemblance to attachment theory, the empirical status of the latter is much more elaborated. Attachment theory that combines concepts drawn from ethology, psychoanalysis, cognitive psychology, and systems theory has sparked off a multitude of theoretical, methodological, and empirical refinements (Cassidy and Shaver, 2008). Of particular importance is the discovery of various attachment types (i.e., secure, resistant, avoidant, and disorganized/disoriented attachment in infancy; secure, preoccupied, anxious, and dismissing attachment in adulthood) and corresponding internal working models. Moreover,

it has been shown that specific aspects of socialization are related to different attachment types in theoretically meaningful ways. For example, several aspects of caregiving that promote secure mother–infant attachments have been found. These include characteristics such as sensitivity, positive attitude, synchrony, mutuality, support, and stimulation. Similarly, a number of distinct interpersonal behaviors such as mutual need for intimacy, exchange of positive affect, and constructive conflict resolution have been observed in adult relationships belonging to the secure type.

Epigenetic theories

The common feature of epigenetic theories is that individual development is conceived as a sequence of qualitatively distinct phases concerning the structure of experience and behavior. A prominent example is Piaget's (1954) theory of cognitive development, comprising four major periods from the sensorimotor period to that of formal operations. These are associated with specific cognitive schemes that are built on the interplay of assimilation and accommodation processes. The epigenetic principle has influenced a series of theoretical conceptualizations concerning other developmental domains including moral reasoning, distributive justice, social perspective taking, or ego development (e.g., Erikson's 1950 psychosocial stages of development). While Piaget emphasized the epigenetic sequence of development as an inherent process of self-creation, Vygotsky (1978) put a stronger emphasis on the cultural influences on development. In particular, his concept of the zone of proximal development clearly stresses the importance of socialization and education in the progression of epigenetically patterned levels of development.

Learning Theories

This class of theories covers a wide range of theoretical approaches spanning from orthodox behaviorism characteristic of Watson and Skinner to social–cognitive and action–theoretic conceptualizations of learning. Although the principles of classical and operant conditioning are by no means outdated, the so-called cognitive revolution in psychology led to a paradigmatic shift in learning theories. One of the earliest contributions to this new type of theories was Rotter's (1972) social learning theory of personality which combines basic principles of learning within an expectancy-value framework. In particular, Rotter's concept of generalized expectancies for internal versus external control of reinforcement and the social conditions shaping these expectancies, especially within the family and in the school, stimulated a tremendous amount of research, presumably because of the affinity of this construct to other socially relevant concepts such as autonomy, personal control, helplessness, or alienation stemming from quite different theoretical positions.

More recently, and in a similar vein, Bandura's (2008) social cognitive learning theory of personality comprising – based on his model of triadic reciprocative relation between intrapersonal and environmental determinants – an agentic theory of the self has contributed much to the analysis and theory-guided application of specific procedures fostering the development of the self system by socialization and self-directed action. Besides his influential theorizing and research on the processes of observational learning and modeling, Bandura has focused on the construct of self-efficacy, that is, an individual's belief that he or she possesses the necessary competencies to attain specific goals. In addition, Bandura has shown that the concept of self-efficacy can also be extended to collective efficacy beliefs of groups or even societies.

Sociological Theories

Not unlike psychological theories there exist quite a few theoretical approaches that have been created within sociology to account for the processes of socialization and education. In the following, three are briefly described because of their historical and contemporary importance.

Structural Functionalism

The theory of structural functionalism is closely associated with the work of Talcott Parsons. The main question that Parsons addresses in his theory refers to processes of how individuals become members of a given society in order to guarantee the survival and maintenance of the social system (cf Parsons and Bales, 1955). According to Parsons, society is a global social system that is based on an integrated value system. The individual person participates in the social system by interacting with others in accordance with the various roles and positions he or she holds in that system. The global social system itself consists of hierarchically ordered subsystems that are characterized by corresponding institutionalized norms. On the one hand, these norms are supposed to be congruent with society's integrated value system and, on the other hand, they determine the expectations and rules attached to specific positions and roles. These are further specified with respect to a set of pattern variables comprising, for example, particularistic versus universalistic values. In this respect, the family plays a particularly salient role (Parsons and Bales, 1955). Drawing heavily on Freudian psychoanalytic theory, Parsons contends that via identification with their same-sex parents boys and girls become socialized into specific male and female role patterns. More precisely, gender-related role specialization is characterized by instrumental or task-oriented roles for males and expressive or person-oriented roles for females. Again, these gender-specific complementary roles are supposed to maintain the equilibrium of the family system as well as the social system as a whole.

Symbolic Interactionism

While Parsons' theory of structural functionalism focused on the conditions ensuring the conservation and stability of the social system, symbolic interactionism called into question whether the expectations and behaviors associated with the enactment of roles can be determined as precisely as Parsons suggested. Instead, Mead (1934), who is credited as the originator of the symbolic interaction framework, maintained that social roles are acquired by imparting and sharing the meaning of symbol systems, particularly that provided by language, in order to ensure an appropriate coordination of the actions of two or more interacting persons. The central vehicle by which this can be achieved is the process of communication which makes role taking a necessary prerequisite. In essence, the shared meaning of symbols is a co-constructive process which is an outgrowth of interpersonal communication. Therefore, the notion of socialization is crucial in symbolic interactionism although Mead

himself never used this term. Instead, he talked about the 'importation' of social symbols into a person's mind. For the developing child, the process of importation is marked by two stages, that is, a play and a game stage. In the play stage, the child is an actor with his own needs and interpretation of the situation (which in Mead's terms represents the self as subject or 'I'); while in the game stage, the child is an actor who is confronted with the needs and social expectations of other actors (which refers to Mead's self as object or 'Me'). In certain times and situations, however, balancing the 'I' and 'Me' aspects of the self may become quite conflictual, thus calling for appropriate means of intra- and interpersonal conflict resolution (Stryker and Vryan, 2003).

Conflict Theories

Generally speaking, conflict within and between social groups is inevitable because the needs and interests of individuals and ingroups are more or less incompatible with the needs and interests of other individuals or groups. At the core of conflict theories is the phenomenon of power, or more precisely the imbalance of power, in particular when the powerful determine the lives and resources of the underprivileged, and at the same time ignore to a great extent their needs and interests (Bohman, 2005). Conflict theories, as, for example, Marx' and Engels' dialectical theory of history, try to understand the conditions leading to social inequality and injustice. However, Marxism not only provides an explanatory framework for the understanding of social inequality and conflict, but also has an emancipatory element inasmuch as it urges political action to dissolve inequality by endowing the underprivileged with more power and control to allow for a more self-determined conduct of life. Marx and Engels used their theory to explain the differences and conflict between social classes. In principle, however, Marxist theorizing can be applied to all social constellations where inequality and lack of justice cause conflict as, for instance, in feminist theories that deal with social gender inequalities.

Unresolved conflict and the plea for emancipation call for appropriate means of conflict resolution which cover a wide range from physical aggression to negotiation between conflict parties. In order to ensure that basic human rights are not violated, socialization for negotiation and constructive conflict resolution is a vital prerequisite. At a philosophical level, Habermas (1984, 1987), a prominent contemporary proponent of critical theory, has suggested developing the art of 'communicative competence' to handle conflicts in a so-called 'ideal discourse situation.' This means that all participants in discourse are granted equal opportunity to express their own point of view and to comment critically on the contributions of their fellow participants, keeping in mind that the ultimate goal is to find a solution that all participants can live with. At a more practical level, specific procedures have been worked out to enhance constructive communication and conflict resolution skills, particularly for the family, school, and work context (Bercovitch and Jackson, 2009).

Elements of an Integrative Framework for Studying Socialization and Education

Based on the core anthropological assumptions and theoretical approaches described in the preceding sections, an integrative framework for the study of socialization and educational processes must at least cover the following five aspects.

Units of Environmental Influence

Since the developing person is a part of different and progressively expanding environments, the units of analysis should comprise all settings that may exert direct or indirect influences on the individual. Bronfenbrenner's (2005) ecological systems approach to human development provides an excellent framework for this purpose. By distinguishing between four systemically interrelated levels of developmental contexts, that is, the micro-, meso-, exo-, and macrosystems, the scope of the framework is sufficiently broad to encompass all possible socializing influences that an individual might have to deal with. More specifically, Super and Harkness (1994) have introduced the concept of 'a developmental niche' to describe the conditions that constitute the developmental context of a particular person, especially a particular child. According to these authors, there are three major components of the developmental niche, that is, (1) the physical and social contexts of everyday life, (2) culturally regulated customs of child care and child rearing, and (3) particular psychological characteristics of a child's parents. Together these three components constitute the proximal context that determines the child's experiential field. On the whole, broader ecological systems as well as more specific aspects of the developmental niche form the opportunity structures of the developing person.

Characteristics of the Individuals to be Socialized

Within systems of interacting persons the characteristics of the individuals to be socialized play an important role in shaping the processes and outcome of socialization as a result of their life history (Rhodewalt, 2008). Particularly salient individual characteristics that moderate and mediate socializing influences are (1) basic personal characteristics (e.g., genotype, health, and maturational status of cognitive, motivational, and temperamental dispositions); (2) characteristic adaptational strategies (e.g., information processing, coping and defense mechanisms, and volitional and emotion regulation processes); and (3) typical forms of experiencing the self and the world (e.g., self-concepts, experience of personal identity and coherence, and internal representations of relationships, situations, and environments).

Proximal Processes

To combine the characteristics of a person and his or her environment in a meaningful way, proximal processes, that is, reciprocal interactions between the human organism and the persons, objects, and symbols in its environment (Bronfenbrenner, 2005), are presumed to produce specific socialization outcomes. More specifically, with respect to personality development Caspi and Shiner (2006) have called attention to several kinds of processes explaining developmental diversity and continuity (e.g., environmental elicitation, construal, selection, and manipulation of person–environment transactions).

Goals and Values

In the behavioral and social sciences there is disagreement as to whether or not goals and values governing socialization and education should be incorporated as normative statements in corresponding theories, as is clearly the case in conflict theories of socialization. Although goals and values are undoubtedly objects of socialization theories and research, this controversy remains unresolved. However, even if the originators of socialization theories deliberately restrict themselves to descriptive and explanatory statements, such theories can nevertheless contribute to a great extent to the clarification of socialization goals, and to a better understanding of their antecedents and consequences. Thus, theories of socialization, and, particularly, corresponding research findings, provide valuable information that can be used to make informed choices with respect to one's own beliefs.

Self-socialization

Based on the anthropological assumptions presented above, socialization consists not only, as Mead (1934) contended, of the 'importation' of social symbols into an individual's mind, but deals also with the capacity to select and create personally and socially meaningful goals, due to the individual's progressively increasing action competence and freedom of choice. The latter process refers to the principle of self-socialization or self-cultivation within the limits of personal and sociocultural opportunity structures comprising informal and formal institutions (e.g., families, schools, vocational education, and training institutions). In this context, human beings are creating themselves and the culture in which they live. Moreover, self-socialization enables them to work on the lifelong project of designing their lives in a personally and communally rewarding way (Schneewind, 2004). Thus, theories of socialization and education are not only an outgrowth of humans' cultural activity but serve also as a critical guide for making well-informed choices to optimize the process of self-socialization.

See also: Behavioral Genetics: Psychological Perspectives; Educational Sociology; Learning Theories and Educational Paradigms; Learning and Instruction: Social-Cognitive Perspectives; Piaget's Theory of Human Development and Education; Social Inequality and Schooling.

Bibliography

Bandura, A., 2008. Toward an agentic theory of the self. In: Craven, R.G., McInerney, D.M. (Eds.), Advances in Self Research, vol. 3. Self-processes, learning, and enabling human potential. Information Age Publishing, Charlotte, NC.

Belsky, J., Steinberg, L., Draper, P., 1991. Childhood experience, interpersonal development, and reproductive strategy: an evolutionary theory of socialization. Child Development 62, 647–670.

Bercovitch, J., Jackson, R., 2009. Conflict Resolution in the Twenty First Century. Michigan University Press, Ann Arbor, MI.

Bohman, J., 2005. Critical Theory. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/critical-theory/ (access on 26.11.11).

Bronfenbrenner, U. (Ed.), 2005. Making Human Beings. Sage, Thousand Oaks, CA.

Bugental, D.B., Grusec, J.E., 2006. Socialization processes. In: Eisenberg, N. (Ed.), Handbook of Child Psychology: Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York.

Caspi, A., Shiner, R.L., 2006. Personality development. In: Eisenberg, N. (Ed.), Handbook of Child Psychology: Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York.

Cassidy, J., Shaver, R.P. (Eds.), 2008. Handbook of Attachment: Theory, Research, and Clinical Applications, second ed. Guilford Press, New York.

Erikson, E.H., 1950. Childhood and Society. W W Norton & Co., New York.

Greenberg, J.R., Mitchell, S.A., 1983. Object Relations in Psychoanalytic Theory. Harvard University Press, Cambridge, MA.

Grusec, J.F., Hastings, P.D. (Eds.), 2007. Handbook of Sozialization Theory and Resarch. The Guilford Press, New York.

Habermas, J., 1984. The Theory of Communicative Action, vol. 1. Beacon Press, Boston, MA.

Habermas, J., 1987. The Theory of Communicative Action, vol. 2. Beacon Press, Boston, MA.

Havighurst, R.J., 1953. Human Development and Education. Longmans Green, New York.

Johns, S.E., Belsky, J., 2008. Life transitions: becoming a parent. In: Salmon, C.A., Shackelford, T.K. (Eds.), Family Relationships. An Evolutionary Perspective. Oxford University Press, New York.

Kernberg, O.F., 1984. Object Relations Theory and Clinical Psychoanalysis. Jason Aronson, Northvale, NJ.

Mead, G.H., 1934. Mind, Self and Society. University of Chicago Press, Chicago.

Moffitt, T.E., Caspi, A., 2007. Evidence from behavioral genetics and environmental contributions to antisocial conduct. In: Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization Theory and Research. The Guilford Press, New York.

Montada, L., 2008. Fragen, Konzepte, Perspektiven. In: Oerter, R., Montada, L. (Eds.), Entwicklungspsychologie. Psychologie Verlags Union, Weinheim, 6. Aufl.

Parsons, T., Bales, R.F., 1955. Family Socialization and Interaction Process. Free Press, New York.

Piaget, J., 1954. The Construction of Reality in the Child. Basic, New York.

Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2008. Behavioral Genetics, fifth ed. Worth Publishers, New York.

Rhodewalt, F. (Ed.), 2008. Personality and Social Behavior. Psychology Press, New York.

Rotter, J.B., 1972. An introduction to social learning theory. In: Rotter, J.B., Chance, J.E., Phares, E.J. (Eds.), Application of a Social Learning Theory of Personality. Holt, Rinehart and Winston, New York.

Schneewind, K.A., 2004. Sechs Thesen zum Verhältnis von Sozialisationstheorie und Persönlichkeitsentwicklung oder (frei nach Heinrich von Kleist) Über die allmähliche Verfertigung der Persönlichkeit beim Leben. In: Geulen, D., Veith, H. (Eds.), Sozialisationstheorie Interdisziplinär. Aktuelle Perspektiven. Lucius & Lucius, Stuttgart.

Stryker, S., Vryan, K.D., 2003. The symbolic interactionist frame. In: Delmater, J. (Ed.), Handbook of Social Psychology. Kluwer Academic/Plenum Publishers, New York.

Super, C.M., Harkness, S., 1994. The developmental niche. In: Lonner, W.J., Malpass, R. (Eds.), Psychology and Culture. Allyn and Bacon, Boston, MA.

Triandis, H.C., 1995. Individualism and Collectivism. Westview, Boulder, CO.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Whiting, J.W.M., Child, I.L., 1953. Child Training and Personality: A Cross-cultural Study. Yale University Press, New Haven, CT.

Socialization in Infancy and Childhood

Melinda S Leidy and Ross D Parke, University of California, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.D. Parke and R. Buriel, volume 21, pp. 14516–14522, © 2001, Elsevier Ltd.

Abstract

This article discusses the role of family in the process of socialization in infancy and childhood. This takes several forms, including treating the family as a social system in which parent–child, marital, and sibling subsystems are recognized. The family socialization process is also recognized within an ecological perspective to appreciate how family environments shape and constrain their socialization practices. Specifically, the impact of social change, socioeconomic status, and ethnicity is discussed.

Socialization is the process by which children acquire the values, standards of behavior, attitudes, and skills that are viewed as appropriate to the culture in which the child resides. Many agents play a role in the socialization process including family, peers, schools, and the media. It is assumed that these various agents function together rather than independently. Families have been recognized as an early pervasive and highly influential context for socialization. Children are dependent on families for nurturance and support from an early age, which accounts, in part, for their prominence as a socialization agent.

Contemporary Perspectives on Family Socialization Theory

Several themes are evident in current theoretical approaches to socialization. First, the rise of systems theory (Minuchin, 1974; Parsons, 1951; Sameroff, 2009) has transformed the study of socialization from a parent–child focus to an emphasis on the family as a social system. To understand fully the nature of family relationships, it is necessary to recognize the interdependence among the roles and functions of all family members. It is being increasingly recognized that families are best viewed as social systems. Consequently, to understand the behavior of one member of a family, the complementary behaviors of other members also need to be recognized and assessed. For example, as men's roles in families shift, changes in women's roles in families must also be monitored.

A closely related theme involves the recognition of the importance of the historical time period in which the family interaction is taking place (Elder and Shanahan, 2006; *see* Life Course: Sociological Aspects). Historical time periods provide the social conditions for individual and family transitions: examples include the 1930s (the Great Depression), the 1980s (Farm Belt Depression), or late 2000s (The Great Recession). Across these historical time periods, family interaction may, in fact, be quite different due to the unique conditions of the particular era. In order to understand the nature of parent–child relationships within families, a multilevel and dynamic approach is required. Multiple levels of analysis are necessary in order to capture the individual, dyadic, and family unit aspects of operation within the family itself as well as to reflect the embeddedness of families within a variety of extrafamilial social systems (*see* Life Course and Event History Analysis). The dynamic quality reflects the multiple developmental trajectories that warrant consideration in understanding the nature of families in infancy. Distinctions among different developmental trajectories, as well as social change and historical period effects, are important because these different forms of changes do not always harmonize (Modell and Elder, 2002). For example, a family event such as the birth of a child may have more effect on a man who has just begun a career than on one who has advanced to a stable occupational position. Moreover, individual and family developmental trajectories are embedded within both the social conditions and the values of the historical time in which they exist. The role of parents as socialization agents is responsive to such fluctuations.

Consistent with a family systems viewpoint, recent research has focused on a variety of subsystems, including parent–child, marital, and sib–sib systems. We employ a tripartite model to understand the parent–child subsystem and the relation between this parent–child subsystem and children's social adaptation. The most common pathway is the parent–child relationship. Parents may influence their children through a second pathway namely in their role as direct instructor, educator, or consultant. In this role parents may explicitly set out to educate their children concerning appropriate norms, rules, and morals of the culture. In a third role, parents function as managers of their children's social lives and serve as regulators of opportunities for social contacts and cognitive experiences (Parke et al., 1994). In this article, we will focus on each of these pathways, as well as marital and sibling subsystems as contexts for socialization. Finally, we will examine the determinants of parental socialization strategies.

Quantitative and Qualitative Assessments of Mother and Father Involvement in Intact Families

In spite of current shifts in cultural attitudes concerning the appropriateness and desirability of shared roles and equal levels of participation in routine caregiving and interaction for mothers and fathers, the shifts toward parity are small but nonetheless real in the majority of intact families. Fathers spend less time with their infants, children, and adolescents than mothers not only in the United States but in other

countries such as Great Britain, Australia, France, and Belgium (Zuzanek, 2000). Mothers and fathers differ in the amount of time that they spend in actual interaction with their children. Pleck (2010) documents that fathers' involvement has increased slowly. Compared to the 1970s, proportional engagement (relative to mothers) was about 33%, whereas accessibility was approximately 50%. In contrast, recent estimates for the 1990s suggest that proportional engagement increased to approximately 70%, whereas accessibility was over 70%. Studies of African-American and Hispanic-American families confirm the pattern found for European Americans. Middle-class and lower-class African-American and Latino fathers were less involved in caregiving with their infants than mothers (Roopnarine, 2004; Roopnarine et al., 2005). Comparisons across ethnic groups (African, Hispanic, and European American) revealed few differences in the level of father involvement (Yeung et al., 2001).

Fathers participate less than mothers in caregiving but spend a greater percentage of the time available for interaction in play activities than mothers do. In North American families, fathers regardless of ethnicity (European American, African-American, or Hispanic-American) spent a greater percentage of their time with their infants in play than mothers, although in absolute terms mothers spent more time than fathers in play with their children (Yeung et al., 2001). The quality of play across mothers and fathers differs too. With infants and toddlers, fathers play more physically arousing games than mothers. In contrast, mothers play more conventional motor games or toy-mediated activities, and are more verbal and didactic (Parke, 1996, 2002). As children develop, fathers become more involved in physical/outdoor play interactions and fixing things around the house and garden, whereas mothers are more actively involved in caregiving and household tasks, in school work, reading, playing with toys, and helping with arts and crafts. In adolescence, the quality of maternal and paternal involvement continues to differ. Just as in earlier developmental periods mothers and fathers may complement each other and provide models that reflect the tasks of adolescence – connectedness and separateness. Recent evidence suggests that fathers may help adolescents develop their own sense of identity and autonomy by being more 'peer-like' and more playful (joking and teasing) which is likely to promote more equal and egalitarian exchanges (Larson and Richards, 1994). However, physical play is not such a central part of the father–child relationship in all cultures: mother–father differences in the amount of play are not evident in Sweden or Israeli Kibbutzim and several cultures (India, China, Africa), physical play is less common for either fathers or mothers (Shwalb et al., 2013).

Why do mothers and fathers play differently? Both biological and environmental factors probably play a role. Experience with infants, the amount of time spent with infants, the usual kinds of responsibilities that a parent assumes – all of these factors influence the parents' style of play. The fact that fathers spend less time with infants and children than mothers may contribute as well. Fathers may use their distinctive arousing style as a way to increase their salience in spite of more limited time. Biological factors cannot be ignored in light of the fact that male monkeys show the same rough and tumble physical style of play as American human fathers. Perhaps predisposing biological differences between males and females may play a role in the play patterns of mothers and fathers. At the same time, the cross-cultural data underscore the ways in which cultural and environmental contexts shape play patterns of mothers and fathers and remind us of the high degree of plasticity of human social behaviors.

Assessing Parent–Child Interaction: Two Approaches

Two approaches to this issue of the impact of parent–child interaction on children's socialization outcomes have been utilized: (1) a typological approach which focused on styles of child-rearing practices and (2) a social interaction approach which focused on the nature of the interchanges between parent and child.

The Typological Approach

The most influential typology has been offered by Baumrind (1973), who distinguished between three types of parental child-rearing typologies: authoritative, authoritarian, and permissive. Authoritative parents were not intrusive and did permit their children considerable freedom within reasonable limits, but were firm and willing to impose restrictions in areas in which they had greater knowledge or insight. In general, high warmth and moderate restrictiveness are associated with the development of self-esteem, adaptability, and social competence. In contrast, the authoritarian parents were rigid, power-assertive, harsh, and unresponsive to the children's needs. These children were often found to be unhappy, conflicted, and neurotic. For permissive parents, in spite of their reasonably affectionate relationship with their children, their excessively lax and inconsistent discipline and encouragement of the free expression of their children's impulses were associated with the development of uncontrolled, impulsive behavior in their children. Baumrind has followed these types of parents and their children from the preschool period through adolescence. She found that authoritative parenting continued to be associated with positive outcomes for adolescents as with younger children and that responsive, firm parent–child relationships were especially important in the development of competence in sons. Moreover, authoritarian child rearing had more negative long-term outcomes for boys than for girls (Baumrind, 1991).

Maccoby and Martin (1983) extended the Baumrind typology and included a fourth type of parenting style, called disengaged, which is characterized by neglect and lack of involvement. These parents do whatever is necessary to minimize the amount of time and effort of the interaction with their child (Maccoby and Martin, 1983). In infants, such a lack of parental involvement is associated with disruptions in attachment; in older children it is associated with impulsivity, aggression, noncompliance, and low self-esteem (Sturge-Apple et al., 2010).

The typological approach has been challenged on several fronts. First, questions remain concerning the processes that contribute to the relative effectiveness of these different styles. Second, it is unclear whether parenting styles are, in part, in response to the child's behavior. Placing the typology work in

a transactional framework would argue that children with certain temperaments and/or behavioral characteristics would determine the nature of the parental style. A third concern is the universality of the typological scheme. Studies have raised serious questions about the generalizability of these styles across either socioeconomic status (SES) or ethnic/cultural groups. In lower SES families, parents are more likely to use an authoritarian as opposed to an authoritative style, but this style is often an adaptation to the ecological conditions such as increased danger and threat that may characterize the lives of poor families (Furstenberg et al., 1999). Furthermore, studies find that the use of authoritarian strategies under these circumstances may be linked with more positive outcomes for children (Dodge et al., 2005). A second challenge to the presumed universal advantage of authoritative child-rearing styles comes from cross-ethnic studies. In Chinese families, authoritarian styles of child rearing are more common and some have argued that the application of these stylistic categories to Chinese parents may be 'ethnocentric and misleading' (Chao, 1994: p. 1111). Finally, in cultures where physical punishment is more normative (e.g., Thailand, the Philippines) the negative effects on children are lessened (Gershoff et al., 2010). Both contextual and cultural considerations need more attention in typological approaches to child rearing.

The Parent–Child Interactional Approach to Socialization

Research in this tradition is based on the assumption that face-to-face interaction with parents may provide the opportunity to learn, rehearse, and refine social skills that are common to successful social interaction with other social partners. The style of the interaction between parent and child is linked to a variety of social outcomes including aggression, achievement, and moral development. Parents who are responsive, warm, and engaging are more likely to have children who are more socially competent (Grimes et al., 2004). In contrast, parents who are hostile and controlling have children who experience more difficulty with age-mates. Moreover, these findings are evident in the preschool period, middle childhood, and adolescence (Harrist et al., 1994; Parke and Buriel, 2006). Although there is an overlap between mothers and fathers, evidence is emerging that fathers make a unique and an independent contribution to their children's social development. Studies (Hart et al., 1998; Isley et al., 1996) have shown that fathers continue to contribute to children's social behavior with peers even after accounting for the mothers' contribution. Although father involvement is quantitatively less than mother involvement, fathers have an important impact on their offspring's development. Quality rather than quantity of parent–child interaction is the important predictor of cognitive and social development.

Parental Instruction, Advice Giving, and Consultation

Learning about relationships through interaction with parents can be viewed as an indirect pathway since the goal is often not explicitly to influence children's social relationships with extrafamilial partners such as peers. In contrast, parents may influence children's relationships directly in their role as a direct instructor, educator, or advisor. In this role, parents may explicitly set out to educate their children concerning appropriate ways of initiating and maintaining social relationships and learning social and moral rules. Research suggests that young children in preschool and early elementary school gain competence with peers when parents supervise and facilitate their experiences, whereas among older children (middle school and beyond), greater parental supervision and guidance of children's peer relationships may function more as a remedial effort (Parke and O'Neil, 2000).

As children grow, the forms of management shift from direct involvement or supervision of the ongoing activities of children and their peers to a less public form of management, involving advice or consultation concerning appropriate ways of handling social problems. In a recent longitudinal study, McDowell and Parke (2009) found that parental advice giving predicted children's social competence, and, in turn, social acceptance 1 year later.

Parents as Managers of Children's Opportunities

Both fathers and mothers play an important role in the facilitation of their children's peer relationships by initiating informal contact between their own children and potential play partners, especially among young children (Reich and Vandell, 2011). However, from infancy through middle childhood, mothers are more likely to assume the managerial role than fathers (Parke, 2002). In infancy, this means setting boundaries for play, taking the child to the doctor, or arranging day care. In middle childhood, it was found that mothers continue to assume more managerial responsibility (e.g., directing the child to have a bath, to eat a meal, or to put away toys).

Parents make choices about neighborhoods and schools as well as the formal and informal activities in which their children can participate. In these ways, "parents act as designers when they seek to control or influence the settings in which children are likely to meet and interact with peers" (Ladd and Pettit, 2002: p. 286). These design decisions can influence children's social and academic outcomes. McDowell and Parke (2009) found that both mothers' and fathers' provision of opportunities for their fourth grade child was related to higher levels of positive social competence in fifth grade. Instead of viewing parents as acting alone in their designer roles, we prefer to view parents and children as codesigners in recognition of children's roles as shapers and negotiators across development. Many decisions are influenced by children's and parents' needs, wishes, and decisions.

Beyond the Parent–Child Dyad: The Marital and Sibling Subsystem as a Contributor to Children's Socialization

Children's experiences in families extend beyond their interactions with parents. Evidence suggests that children's

understanding of relationships is also shaped through their active participation in other family subsystems (e.g., child–sibling) as well as through exposure to the interactions of other dyadic subsystems (e.g., parent–parent).

Influence of Marital Discord on Child Outcomes

Considerable evidence indicates that marital functioning is related to children's short-term coping and long-term adjustment. Children exposed to marital discord are likely to have poorer quality of interpersonal relationships, including internalizing and externalizing behavior problems, and changes in cognitions, emotions, and physiology (see Cummings and Davies, 2010). Three perspectives have been offered to explain the possible links between marital relationships and children's social relationships. One theoretical framework, the indirect effects model, conceptualized marital discord as an indirect influence on children's adjustment that operated through its effect on family functioning and the quality of parenting. A second model, the direct effects model, focuses on the direct effects of witnessed marital conflict on children's outcomes rather than on the indirect effects. Cummings et al. (2010) have proposed a third model: the interrelations among marital quality, parents' psychological functioning, and children's outcomes. Labeled the 'parental mental health hypothesis,' this alternative focuses on the role of parental psychological functioning in accounting for the effects of marital conflict either as a risk or protective factor. All of these models have received empirical support (*see* Marital Interaction: Effects on Child Development).

The Sibling System as a Contributor to Children's Socialization

Siblings play a critical role in the socialization of children. Most children are likely to spend more time in direct interaction with siblings than parents and this array of interactions between siblings have been found to be typified by greater emotional intensity than the behavioral exchanges that characterize other relationships. Sibling relationships contribute to children's socialization in a number of significant ways (Howe et al., 2010). Through their interactions with siblings, children develop specific interaction patterns and social understanding skills that generalize to relationships with other children. Relationships with siblings may also provide a context in which children can practice the skills and interaction styles that have been learned from parents or others. Older siblings function as tutors, managers, or supervisors for their younger siblings and may function as gatekeepers who extend or limit opportunities to interact with other children outside the family (Zukow-Goldring, 2002). Also paralleling the indirect influence that the observation of parent–parent interaction has on children, a second avenue of influence on children's development is their observation of parents interacting with siblings. These interactions may serve as an important context in which children deal with issues of differential treatment and learn about complex social emotions such as rivalry and jealousy (Dunn, 1993).

Determinants of Family Socialization Strategies

One of the major advances in the field has been recognition of the importance of understanding the determinants of parenting behavior. Belsky (1984) proposed a three-domain model of the determinants of parenting, which included characteristics of the child, personal resources of the parents, and contextual sources of stress and support.

Child Characteristics

Child characteristics take two forms: universal predispositions that are shared by all children and individual differences in particular characteristics. Infants are biologically prepared for social, cognitive, and perceptual challenges and these prepared responses play a significant role in facilitating children's adaptation to their environment. Under the influence of recent advances in behavior genetics, there is increasing recognition of the role of individual differences in child temperament on parenting behavior (Rothbart, 2011). Although debates about the relative contributions of genetic and experiential factors to the emergence of individual differences in temperament continue (Reid et al., 2002), temperament clearly is a determinant of parental socialization tactics. Infants with difficult temperaments elicit more arousal and distress from caregivers than less difficult infants (Sanson et al., 2011). Children who are more difficult may elicit increasingly coercive strategies from parents (Reid et al., 2002). Alternatively, fearful children may respond optimally to subtle parental socialization strategies (Kochanska et al., 2007). Other characteristics, in addition to temperament, have been examined, including activity level, social responsiveness, and compliance level. In general, the more active, less responsive, and more noncompliant children elicit more negative parenting and more negative parental arousal and affect (Crouter and Booth, 2003). The impact of these individual differences on parental socialization behavior is not independent of environmental conditions. Crockenberg and Leerkes (2003) showed that the impact of a difficult infant temperament on the parent–infant attachment relationship varied as function of the degree of social support available to the mother, which underscores the potential modifiability of temperament-based influences.

Personal Resources

A variety of studies support the prediction that personal resources – conceptualized as knowledge, ability, and motivation to be a responsible caregiver – alter parenting behaviors (Belsky, 1984). Particularly striking are recent studies of how parental psychopathology such as depression will alter parenting behavior (Goodman and Gotlib, 2002). When interacting with their infants, depressed mothers show flat affect and provide less contingent stimulation than nondepressed mothers; in turn, their infants showed less attentiveness, more fussiness, and lower activity levels. Differences are particularly evident when depression is protracted and not merely transient (Campbell et al., 1995). These differences in interaction may place the infant at risk for later developmental problems. Infants of depressed mothers are more likely to develop insecure attachments

(Goodman and Gotlib, 2002). Other investigations have found links between severe and chronic depression and disorganized attachment behavior (Lyons-Ruth and Jacobvitz, 2008).

Social Support and Family Socialization Strategies

Recognition of the role of the community and community agents as modifiers of family interaction is necessary for an adequate theory of socialization. Social capital in a community can aid parents' socialization of their children through several pathways. First, when parents and children have community ties, more social support is available. Second, parental awareness of community services and their participation in shaping the institutions of the community promote the maintenance of values and norms that influence their children. Third, parental participation with their children enables closer supervision of children and reduces the time children spend with their own peers. The concept of social capital embodies the notion not only that parenting is a community enterprise (Elder and Conger, 2000) but also that children and adults are active players in the distribution of social capital. More attention needs to be given to children's role as active agents in this process.

SES as a Determinant of Family Socialization Strategies

There is a long history of research concerning the links between SES and/or social class and parenting beliefs and practices. In contrast to traditional assumptions that SES is a static state, SES is a dynamic concept. Moreover, most scholars agree that SES is multiply determined, and therefore the links with parenting are likely to be multiple as well. Over the course of childhood and adolescence, families change social class and change is greatest in the youngest ages. Over 50% of American children change social class prior to entering school (Featherman et al., 1988). In spite of the controversies surrounding this variable, there are SES differences in parental socialization practices and beliefs. Lower SES parents are more authoritarian and more punitive than higher SES families (Conger and Dogan, 2007; Hoff et al., 2002). Second, there are more SES differences on language measures than on nonverbal measures with higher SES mothers being more verbal than low SES mothers (Hoff et al., 2002). Some SES differences are independent of race and poverty. In China, where there are relatively small differences in income across groups who vary in terms of education, less educated parents used more imperatives with their toddlers than better educated mothers (Tardif, 1993). Similarly, studies of cognitive socialization found clear SES differences in African-American lower-class and middle-class families (Hess and Shipman, 1965).

Ethnicity as a Determinant of Family Socialization Strategies

Cross-cultural and intracultural theories have emphasized the importance of socialization goals, values, and beliefs as organizing principles for understanding cultural variations (Harkness and Super, 2002). In contrast to the older cultural deficit models of socialization, recent models emphasize how ecological demands shape socialization goals, values, and practices, and are viewed as adaptive strategies to meet the demands of the ecological settings. In the past, cultural deficit models were popular explanations for the socialization and child outcome differences observed between ethnic minorities and Euro-Americans. The focus on ethnic minority families has shifted away from majority–minority differences in developmental outcomes and more toward an understanding of the adaptive strategies ethnic minorities develop in response to both majority and minority cultural influences on their development.

The parents' individual history of interaction with the larger sociocultural context, including their awareness of their ethnic group's history within the larger society, affects the manner in which they socialize their children. An important dimension of socialization in ethnic minority families is teaching children how to interact effectively in dual cultural contexts; the context of their ethnic group and the context of the larger Euro-American society. Harrison et al. (1990) have adopted an ecological orientation to explain the diverse environmental influences that contribute to the socialization of ethnic minority children. They conceptualize the socialization of ethnic minority children in terms of the interconnectedness between the status of ethnic minority families, adaptive strategies, socialization goals, and child outcomes. Emerging out of the adaptive strategies of adults are the socialization goals that they endeavor to inculcate in children to help them meet the ecological challenges they will face as ethnic/racial minorities in a class- and race-conscious society. Ethnic/racial pride and interdependence are two important socialization goals that enable ethnic minority children to function competently as members of both their minority culture and the larger society (Harrison et al., 1990).

Research needs to take into account the acculturation level of parents and children in recent immigrant families and the effects it has on family processes and child outcomes. Intergenerational differences in acculturation can create role strains between parents and children that have implications for child-rearing styles, disciplinary practices, and overall parent–child relations. Together with acculturation, recognition of biculturalism as both an adaptation strategy and socialization goal is important. The effects of prejudice and discrimination on ethnic minorities, in such areas as social and emotional development, ethnic identity, and achievement motivation, deserve more attention. Language development research should also give greater attention to second-language acquisition (usually English) and bilingualism and its relation to cognitive development and school achievement. More attention must also be given to the role of ethnic minority fathers, grandparents, and extended family members in the socialization of children (*see* Parenting in Ethnic Minority Families: United States). Finally, observational studies with ethnic/racial minority families have the potential for yielding a contextual richness about the family environments of ethnic/racial minorities that can lead to more culturally relevant insights and theories about their socialization experiences.

The Impact of Social Change on Family Socialization

Families are not static but dynamic and are continuously confronted by challenges, changes, and opportunities.

A number of society-wide changes have produced a variety of shifts in the nature of family relationships. Fertility rates and family size have decreased, the percentage of women in the workforce has increased, the timing of onset of parenthood has shifted, divorce rates have risen, and the number of single parent families has increased (Parke, 2013). These social trends provide an opportunity to explore how families adapt and change in response to these shifting circumstances and represent 'natural experiments' in family coping and adaptation. Moreover, they challenge our traditional assumptions that families can be studied at a single point in historical time since the historical contexts are constantly shifting. Our task is to establish how socialization processes operate similarly or differently under varying historical circumstances. Some of these changes are scheduled or planned, such as reentry into the workforce or delaying the onset of parenthood; other changes, such as job loss or divorce, are unscheduled or nonnormative transitions. According to a life-course view, both scheduled and unscheduled transitions need to be examined to fully appreciate how these different types of changes alter family socialization strategies (Elder and Shanahan, 2006). These family transitions are adult-focused in contrast to child-focused transitions (e.g., entry to day care or junior high school) and underscore our assumption that adult developmental issues need to be directly addressed to understand how these transitions alter parental socialization beliefs and behaviors. At the same time, child developmental status plays a major role in determining how adults respond to these transitions. It is insufficient to focus on individual levels of analysis – either adult or child since individual, dyadic, triadic, and family units each follows their own developmental trajectory and the interplay among these separate developmental trajectories can produce a diverse set of effects on children. The role of these units (i.e., individual, dyad, or family) in modifying the impact of family transitions will vary as a result of these interlocking developmental curves. Both the timing and nature of family transitions and reactions to these alterations will be determined by the points at which particular individuals, dyads, triads, or families fall along their respective developmental life-course trajectories. Moreover, individual families can vary widely in the particular configuration of life-course trajectories. The particular configurations of these multiple sets of developmental trajectories need to be considered to understand the impact of societal change on families.

Conclusion

Socialization is a multiply determined process; while our focus has been on families as central socialization agents, it is critical to recognize the roles of other social agents such as schools, peers, and the mass media in this process. A fuller understanding of socialization will come from more attention to the interplay among these diverse agents and how cultural and ethnic background influence socialization practices and outcomes.

See also: Attachment During Infancy and Early Childhood: Cultural Variations, Development of; Divorce, Sociology of; Human Development, Theories of; Identity in Childhood and Adolescence; Life Course and Event History Analysis; Life Course: Sociological Aspects; Parenting Attitudes and Beliefs across Cultures; Personality Changes During Adolescence Across Cultures; Self-Development in Childhood and Adolescence; Sibling-Order Effects.

Bibliography

Baumrind, D., 1973. The development of instrumental competence through socialization. In: Pick, A.D. (Ed.), Minnesota Symposium on Child Psychology, vol. 7. University of Minnesota Press, Minneapolis, MN, pp. 3–46.

Baumrind, D., 1991. Parenting styles and adolescent development. In: Brooks-Gunn, J., Lerner, R.M., Peterson, A.C. (Eds.), The Encyclopedia of Adolescence. Garland, New York, pp. 746–758.

Belsky, J., 1984. The determinants of parenting: a process model. Child Development 55, 83–96.

Campbell, S.B., Cohn, J.F., Meyers, T., 1995. Depression in first time mothers: mother–infant interaction and depression chronicity. Developmental Psychology 31, 349–357.

Chao, R.K., 1994. Beyond parental control and authoritarian parenting style: understanding Chinese parenting through the cultural notion of training. Child Development 65, 1111–1119.

Conger, R.D., Dogan, S.J., 2007. Social class and socialization in families. In: .Grusec, J.E., Hastings, P.D. (Eds.), Handbook of Socialization: Theory and Research. Guilford Press, New York, pp. 433–460.

Crockenberg, S.C., Leerkes, E.M., 2003. Parental acceptance, postpartum depression, and maternal sensitivity: mediating and moderating processes. Journal of Family Psychology 17, 80–93.

Crouter, A.C., Booth, A. (Eds.), 2003. Children's Influence on Family Dynamics: The Neglected Side of Family Relationships. Erlbaum, Mahwah, NJ.

Cummings, E.M., Davies, P.T., 2010. Marital Conflict and Children: An Emotional Security Perspective. Guilford Press, New York, NY, US.

Cummings, E.M., Merrilees, C.E., George, M.W., 2010. Fathers, marriages, and families: revising and updating the framework for fathering in family context. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, fifth ed. John Wiley & Songs, Hoboken, NJ, pp. 154–176.

Dodge, K.A., McLoyd, V.C., Lansford, J.E., 2005. The cultural context of physically disciplining children. In: McLoyd, V.C., Hill, N.E., Dodge, K.A. (Eds.), African American Family Life. Guilford Press, New York, pp. 245–263.

Dunn, J., 1993. Young Children's Close Relationships: Beyond Attachment. Sage, Thousand Oaks, CA.

Elder, G.H., 1998. The life-course and human development. In: Damon, W. (Series Ed.), Lerner, R.M. (Vol. Ed.), Handbook of Child Psychology. Theoretical Models of Human Development, vol. I, fifth ed. Wiley, New York, pp. 939–991.

Elder Jr., G.H., Conger, R.D., 2000. Children of the land: adversity and success. In: Rural America. University of Chicago Press, Chicago.

Elder Jr., G.H., Shanahan, M.J., 2006. The life course and human development. In: Damon, W. (Series Ed.), Lerner, R.M. (Vol. Ed.), Handbook of Child Psychology. Theoretical Models of Human Development, vol. I., sixth ed. Wiley, Hoboken, New Jersey, pp. 665–715.

Featherman, D.L., Spenner, K.I., Tsunematsu, N., 1988. Class and the socialization of children: constancy, change, or irrelevance. In: Hetherington, E.M., Lerner, R.M., Perlmutter, M. (Eds.), Child Development in Life-Span Perspective. Erlbaum, Hillsdale, NJ, pp. 67–90.

Furstenberg Jr., F.F., Cook, T.D., Eccles, J., Elder Jr., G.H., Arnold Sameroff, 1999. Managing to Make It: Urban Families and Adolescent Success. University of Chicago Press, Chicago, IL.

Gershoff, E.T., Grogan-Kaylor, A., Lansford, J.E., Chang, L., Zelli, A., Deater-Deckard, K., Dodge, K.A., 2010. Parent discipline practices in six countries: frequency of use, associations with child behaviors, and moderation by cultural normativeness. Child Development 81, 480–495.

Goodman, S.M., Gotlib, I. (Eds.), 2002. Children of Depressed Parents: Mechanisms of Risk and Implications for Treatment. American Psychological Association, Washington, DC.

Grimes, C.L., Klein, T.R., Putallaz, M., 2004. Parents' relationships with their parents and peers: influences on children's social development. In: Kupersmidt, J.B., Dodge, K. (Eds.), Children's Peer Relations: From Development to Intervention. American Psychological Association, Washington, DC, pp. 141–158.

Harkness, S., Super, C.M., 2002. Culture and parenting. In: Bornstein, M. (Ed.), Handbook of Parenting, Biology and Ecology of Parenting, vol. 2. Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 253–280.

Harrison, A.O., Wilson, M.N., Pine, C.J., Chan, S.Q., Buriel, R., 1990. Family ecologies of ethnic minority children. Child Development 61, 347–362.

Harrist, A.W., Pettit, G.S., Dodge, K.A., Bates, J.E., 1994. Dyadic synchrony in mother-child interaction: relation with children's subsequent kindergarten adjustment. Family Relations: Interdisciplinary Journal of Applied Family Studies 43, 417–424.

Hart, C.H., Yang, C., Nelson, D.A., Jin, S., Bazarakaya, N., Nelson, L., et al., 1998. Peer contact patterns, parenting practices and preschoolers' social competence in China, Russia and the United States. In: Slee, P.T., Rigby, K. (Eds.), Children's Peer Relations. Routledge, London, pp. 3–30.

Hess, R.D., Shipman, V.C., 1965. Early experience and the socialization of cognitive modes in children. Child Development 36, 869–886.

Hoff, E., Laursen, B., Tardif, T., 2002. Socioeconomic status and parenting. In: .Bornstein, M.H. (Ed.), Handbook of Parenting, Biology and Ecology of Parenting, second ed., vol. 2. Erlbaum, Mahwah, NJ, pp. 231–252.

Howe, N., Ross, H.S., Recchia, H., 2010. Sibling relations in early and middle childhood. In: Smith, P.K., Hart, C.H. (Eds.), The Wiley-Blackwell Handbook of Childhood Social Development, second ed. Wiley-Blackwell, Oxford, UK, pp. 356–372.

Isley, S., O'Neil, R., Parke, R.D., 1996. The relation of parental affect and control behaviors to children's classroom acceptance: a concurrent and predictive analysis. Early Education and Development 7, 7–23.

Kochanska, G., Aksan, N., Joy, M.E., 2007. Children's fearfulness as a moderator of parenting in early socialization: two longitudinal studies. Developmental Psychology 43, 222–237.

Ladd, G.W., Pettit, G., 2002. Parenting and the development of children's peer relationships. In: Bornstein, M.H. (Ed.), Handbook of Parenting, Practical Issues in Parenting, second ed., vol. 5. Erlbaum, Mahwah, NJ, pp. 269–309.

Larson, R., Richards, M.H., 1994. Divergent Realities: The Emotional Lives of Mothers, Fathers and Adolescents. Basic Books, New York.

Lyons-Ruth, K., Jacobvitz, D., 2008. Attachment disorganization: genetic factors, parenting contexts, and developmental transformation from infancy to adulthood. In: Cassidy, J., Shaver, P.R., Phillip, R. (Eds.), Handbook of Attachment: Theory, Research, and Clinical Applications, second ed. Guilford Press, New York, NY, pp. 666–697.

Maccoby, E.E., Martin, J.A. (1983). Socialization in the context of the family: parent–child interaction. In: Mussen, P.H. (Series Ed.), Hetherington, E.M. (Vol. Ed.), Handbook of Child Psychology. Socialization, Personality, and Social Development, vol. 4, fourth ed. McGraw-Hill, New York, pp. 1–101.

McDowell, D.J., Parke, R.D., 2009. Parental correlates of children's peer relations: an empirical test of a tripartite model. Developmental Psychology 45, 224–235.

Modell, J., Elder, G.H., 2002. Children develop in history: so what's new. In: .Hartup, W.W., Weinberg, R.A. (Eds.), Minnesota Symposia on Child Psychology, Child Psychology in Retrospect and Prospect, vol. 32. Erlbaum, Mahwah, NJ, pp. 173–206.

Minuchin, S., 1974. Families and Family Therapy. Harvard University Press, Cambridge, MA.

Parke, R.D., 1996. Fatherhood. Harvard University Press, Cambridge, MA.

Parke, R.D., 2002. Fathers and families. In: Bornstein, M.H. (Ed.), Handbook of Parenting, Being and Becoming a Parent, second ed., vol. 3. Earlbaum, Mahwah, NJ, pp. 27–73.

Parke, R.D., 2013. Gender differences and similarities in parenting. In: Kline, K.K., Wilcox, W.B. (Eds.), Gender and Parenthood: Natural and Social Scientific Perspectives. Columbia University Press, New York.

Parke, R.D., Buriel, R., 2006. Socialization in the family: ecological and ethnic perspectives. In: Damon, W., Lerner, R.M. (Series Eds.), Eisenberg, N. (Vol. Ed.), Handbook of Child Psychology. Social, Emotional, and Personality Development, vol. 3, sixth ed. Wiley, New York, pp. 429–504.

Parke, R.D., Burks, V., Carson, J., Neville, B., Boyum, L., 1994. Family–peer relationships: a tripartite model. In: Parke, R.D., Kellam, S. (Eds.), Advances in Family Research, Exploring Family Relationships with Other Social Contexts, vol. 4. Erlbaum, Hillsdale, NJ, pp. 115–145.

Parke, R.D., O'Neil, R., 2000. The influence of significant others on learning about relationships: from family to friends. In: Mills, R., Duck, S. (Eds.), The Developmental Psychology of Personal Relationships. Wiley, London, pp. 15–47.

Parsons, T., 1951. The Social System. Free Press, Glencoe, IL.

Pleck, J., 2010. Paternal involvement: revised conceptualization and theoretical linkages with child outcomes. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, fifth ed. John Wiley & Sons Inc., Hoboken, NJ, pp. 58–93.

Reich, S.M., Vandell, D.L., 2011. The interplay between parents and peers as socializing influences in children's development. In: Smith, P.K., Hart, C. (Eds.), Wiley-Blackwell Handbook of Child Social Development. Wiley-Blackwell, New York.

Reid, J.B., Patterson, G.R., Snyder, J. (Eds.), 2002. Antisocial Behavior in Children and Adolescents: A Developmental Analysis and Model for Intervention. American Psychological Association, Washington, DC.

Roopnarine, J., 2004. African American and African Caribbean fathers: level, quality, and meaning involvement. In: Lamb, M.E. (Ed.), The Role of the Father in Child Development, fourth ed. John Wiley, Hoboken, NJ, pp. 58–97.

Roopnarine, J.L., Fouts, H.N., Lamb, M.E., Lewis-Elligan, T.Y., 2005. Mothers' and fathers' behaviors toward their 3- to 4-month old infants in lower, middle, and upper socioeconomic African American families. Developmental Psychology 41, 723–732.

Rothbart, M.K., 2011. Becoming who we are: temperament and personality. In: Development. Guilford Press, New York, NY.

Sameroff, A. (Ed.), 2009. The Transactional Model of Development: How Children and Contexts Shape Each Other. American Psychological Association, Washington, DC.

Sanson, A., Hempill, S.A., Yagmurli, B., McClowry, S., 2011. Temperament and behavior. In: Smith, P.K., Hart, C.H. (Eds.), Handbook of Social Development. Wiley-Blackwell, New York, pp. 227–245.

Shwalb, D.W., Shwalb, B.J., Lamb, M.E., 2013. Fathers in Cultural Context. Routledge, New York, NY.

Sturge-Apple, M.L., Davies, P.T., Cummings, E.M., 2010. Typologies of family functioning and children's adjustment during the early school years. Child Development 81, 1320–1335.

Tardif, T., 1993. Adult-to-Child Speech and Language Acquisition in Mandarin Chinese. Yale University Press, New Haven, CT.

Yeung, W.J., Sandberg, J.F., Davis-Kean, P.E., Hofferth, S.L., 2001. Children's time with fathers in intact families. Journal of Marriage and the Family 64, 780–792.

Zukow-Goldring, P., 2002. Sibling caregiving. In: Bornstein, M.H. (Ed.), Handbook of Parenting, Being and Becoming a Parent, vol. 3. Erlbaum, Mahwah, NJ, pp. 253–286.

Zuzanek, J., 2000. The Effects of Time Use and Time Pressure on Child Parent Relationships. Otium, Waterloo, Ontario, Canada.

Societal Impact Assessment

Reinhard Kreissl and Florian Fritz, Institut für Rechts- u. Kriminalsoziologie, Vienna, Austria
Lars Ostermeier, Zentrum für Technik und Gesellschaft, Technical University Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Societal impact assessment (SIA) refers to a set of guidelines and principles to identify the societal effects of new technologies, programs, and projects. Starting in the domain of environmental issues, SIA has expanded into other areas where new technologies start to shape the everyday life of citizens. With a strong focus on participatory and anticipatory strategies, SIA attempts to establish a broader understanding of technology taking into account the perspective of different stakeholders. This article briefly reviews key elements of SIA and then discusses its potential, using privacy impact assessment as a test case.

Introduction and Context

While traditionally social impact assessment (SIA) is about analyzing, monitoring, and managing the social consequences of development (Vanclay, 2003, 2006), this article proposes to broaden the perspective and conceptualize 'Societal Impact Assessment.' The term 'societal' encompasses the analysis of anything affecting human, natural, or artifactual systems. This allows for significant conceptual innovation: while hitherto SIA has worked under the assumption that actors, interests, and consequences of programs, projects, actions, and research can be identified, 'Societal Impact Assessment' acknowledges the potential of technology to alter society as such, rendering this altering almost irreversible (e.g., introduction of new energy sources, etc.). It thus allows to capture structural changes affecting society as a whole, leading to a new level of normality without stopping at the limits of a well-circumscribed development site or the identification of actors and vested interests. We therefore use the term 'Societal Impact Assessment' throughout this article.

The idea of conducting SIAs is a child of reflexive modernity (Beck, 1992). While science and technology for a long time were perceived as social forces of unambiguous progress liberating mankind from burdens and hazards of life and nature, the advent of modern technologies affecting society in unforeseen ways gradually reshaped the public image of science. Critical reappraisals of science and technology followed on spectacular scientific breakthroughs, paving the way for new technologies, like, e.g., nuclear weapons. Concerned scientists congregated at meetings like the *Pugwash* Conferences to critically reflect on the societal impact of their work. While the main focus of these debates was on the consequences of scientific knowledge, applied to create weapons of mass destruction, at the same time a critical stance toward science and technology emerged in the debate about environmental hazards. Rachel Carson's book *Silent Spring*, first published in 1962, is today perceived as the manifesto launching modern environmentalism (Carson, 1962). Carson critically examined the detrimental effects of pesticides, hitherto seen as benevolent chemical agents in agriculture. Concerned scientists from different disciplines started to take a critical look on their own achievements.

While the natural sciences came into the focus of societal critique in this period in the first decades after WW II, the social sciences had their heyday in the 1960s providing a scientific toolbox for planned social change and improvement. The Great Society Programme in the United States, initiated by American president Lyndon B. Johnson, heavily relied on the expertise of social scientists. Their task was to contribute to the analysis and definition of social problems to be addressed by socio-technical policies and to design the adequate measures to combat social inequality, crime, and poverty. But soon critique from within the academic community emerged, showing with the very same methods and approaches used to analyze and define social ills, that the cure might be worse than the disease. Social programs, designed to reduce poverty, so the critics claimed (see e.g., Moynihan, 1993), would rather stabilize a 'culture of poverty' than create greater social equality.

What can be demonstrated here is how scientific knowledge in different disciplines was transformed into a productive force, creating new technologies and changing society on a massive scale, thus putting the issue of social responsibility of science on the agenda. SIA was one strategy to develop an institutionalized and rational response from within the scientific community to this situation. In 1981 the International Association for Impact Assessment (IAIA) was established to provide a professional platform for the implementation of SIA. Membership of the IAIA was recruited from the academic community, private consultancies (who saw an emerging business case in SIA), government, and from industries. Other forms of institutionalization followed, such as the parliamentary technology assessment bodies (PTA) in different countries, advising Parliaments on the effects of new technologies. PTA established an international association at the European level with the foundation of the European Parliamentary Technology Assessment Network (EPTA).

The main focus of impact assessment activities was on environmental effects and sustainability. Green issues had a high political visibility and attracted substantial public attention. With the spread of different technologies into everyday life, new fields came into focus, starting from health and medicine, to nanotechnologies to energy and transport, and in recent times to issues of privacy and data protection with the spread of new communication and surveillance technologies. In many national jurisdictions conducting an impact assessment has become a mandatory element of implementing new technologies or projects.

Conceptually, SIA is a multidisciplinary professional enterprise. It requires technical expertise in different scientific fields and at the same time a sound understanding of broader societal issues and nontechnological effects of a given technology. In its broadest sense, SIA focuses on issues of human rights, social inequality, environmental issues, and public participation. These main topics can be spelt out in different directions depending on the problem at hand. At the same time SIA by definition is tied into a political field (Freudenburg, 1986), and thus establishing a position not contaminated by vested interests in the analysis and following a clear process of objective assessment is of utmost importance. In essence, there are two ways of framing the SIA: "In addition to being a field of research, SIA is conceived as being a methodological approach or framework" (Esteves et al., 2012: p. 34).

Different forms of SIA can be distinguished with regard to the scope of the approach taken. In a more narrow understanding SIA will investigate whether a given project develops within an existing regulatory framework, e.g., environmental standards, defined by law are met. In a broader sense it will include participatory elements involving different stakeholders and experts in a discursive exchange about a technology, project or program at hand (Harvey, 2011). Finally SIA can be designed in a more reflexive way, feeding back into the core processes of technology development by addressing the rationale and rationality of research and development from a societal impact perspective. This strategy is pursued predominantly in constructive technology assessment (CTA).

Methodologies and Approaches

Different methodologies and guidelines were developed to conduct SIA. Vanclay and Esteves (2011) define a number of consecutive steps to be followed when conducting SIA. Kemp (2011) lists different groups of actors or parties involved to be considered when assessing societal impact. Becker (2001) proposes a two-phase model, with the first, preparatory phase focusing on an analysis of the project under investigation, while the second or main phase makes use of scenario-building techniques to identify the range of future effects of a project. This forward-looking dimension is a key feature of SIA; but at the same time it poses one of the key problems, since the future-oriented logic of risk (see Luhmann, 1993) underlying the scenario technique can have a number of far-reaching consequences with regard to accountability and decision-making processes in the present.

The two main types of problems to be solved with an SIA can be classified as *cognitive* and *social* problems. With regard to cognitive problems, it is difficult to project future trajectories of technological development and impact, the more so when complex systems of systems are involved. Scenario techniques can help to address the problems of complexity, modeling interaction, and cascading effects. But at the same time understanding complex scenarios and discussing the probability of different options often requires detailed expert knowledge not always easily available to all groups of stakeholders involved.

Regarding social problems, ensuring an inclusive approach in SIA, i.e., giving all stakeholders and groups involved in an assessment exercise an adequate share of voice can be a quite challenging task. Community participation is sometimes used as a seemingly 'quick fix' for problems, without addressing the root of the issue (which often lies to unequal power distributions that are deeply engrained in polities, and sometimes in the very institutions of the community whose participation is sought). As a result, community participation sometimes takes the form of a tokenistic exercise; once 'affected communities' – or their leaders – have been heard, the respective box can be ticked off, and afterward the rule of previous power relations resumes (Peterman, 2004).

Similar concerns have been voiced about the emphasis on public participation (i.e., the participation of potentially everyone, not only those affected by a project or policy). The "reasons for seeking public participation [in impact assessment exercises are] highly diverse and not always mutually consistent," as O'Faircheallaigh (2010: p. 19) points out, and some projects specify the benefits of public participation only very poorly. The reason for this is that such benefits are treated as self-evident (for a typology of such benefits see O'Faircheallaigh, 2010). Others have argued that public participation only deserves its name when publics are actively involved in decision-making, and their stances have significant influence in the development of the project or policy in question (Bishop and Davis, 2002). To address the challenge, Prainsack (2014) has developed a grid to establish the degree of agency of members of the public vis-à-vis professional experts in a project. This grid aims to assess the strengths and weaknesses of a project from a participatory perspective.

A final problem emerging with public participation is of the NIMBY type. While some stakeholders may oppose to a certain project, because they are immediately affected, others may be supportive, as long as the immediate negative effects are not felt in their own backyards and they have access to the positive effects. An alternative option to the NIMBY position, popularized under the acronym NIABY (Not in Anyone's Backyard) is the wholesale refusal of a technology or project under discussion.

Beginning at an earlier stage of technology development processes, the approach labeled as constructive technology assessment (CTA) looks at the interaction between internal and external forces, developers, users, stakeholders, in the development and design of technologies. CTA's main value, it could be argued, is not about predicting possible future impacts of the technology as accurately as possible – rather, it lies in the dialogue and the interaction it requires the actors to engage in, while developing a technological solution. This dialogue and interaction contribute to enhancing a mutual understanding of each actor's needs and objectives. This is expected to improve problem-solving capacities even if negative impacts occur. This approach is characteristic for CTA. Moreover, CTA clearly sets itself apart from the 'deficit model' of public engagement, where publics are seen in need to be 'educated' about science in order to appropriately understand, and ultimately accept, new science and technology. CTA's approach to public engagement is strongly participatory and dialogical. In other words, the emphasis of CTA does not lie in predicting impacts and facilitating technological development and deployment, but it lies in understanding them (as well as the dynamics which facilitate or impede their occurrence).

It is precisely this understanding of the dynamics underlying the technology development process that opens up opportunities for intervention. These opportunities need to be identified, and then it needs to be specified in what ways such interventions can be most productive (Rip and Schot, 2002).

Understanding the dynamics of development allows one to identify opportunities of intervention, and specify how such interventions can be productive. A better understanding of impact anticipation and appropriate intervention can contribute to mitigating the complexity of the process.

There are three generic strategies CTA pursues (Schot and Rip, 1996): technology forcing, strategic niche management, and alignment. The authors argue that these strategies should be jointly deployed in the field in order to maximize reflexivity, anticipation, and learning processes.

Technology forcing is a term that originates in industry and signifies the setting of standards for industry that cannot be reached with current means and technologies; the standards thus 'force' industry into innovation. In CTA, 'technology forcing' describes the modulation of technology development from the demand side. This implies an inversion of traditional technology assessment (TA) approaches, as it is the formulation of requirements for impacts that drives the development of technology, rather than taking technology as a given.

The other two generic CTA strategies are *strategic niche management* and *alignment*. The former is about the creation of technological niches that, as protected spaces, allow for a safe experimental environment in which the interplay of technology, user practices, and frameworks such as the regulatory context can be looked at. The overarching aim here is to contribute to a more sustainable innovation journey by more rapid reaction to societal changes. *Alignment*, the third strategy within CTA, also looks at the interaction process of the assessment actors, aims to foster mutual understanding and dialogue and thus 'align' different perspectives.

Privacy and Surveillance Impact Assessment – A Test Case

One of the prominent cases for SIA in recent times has emerged out of the spread of surveillance technologies and social media. Electronically mediated communication, new forms of electronic commerce, and the widespread use of mobile phones not only created a technological boost in soft- and hardware development but at the same time opened hitherto unprecedented opportunities for the gathering and processing of person-related data by private enterprises and public authorities. This process of 'datafication' (Mayer-Schönberger and Cukier, 2013) has created a demand for a procedure termed 'privacy impact assessment' (PIA), a term that did not show up on the Social Science Research Network before 2012. Taking PIA as a showcase for SIA processes, problems and potentials of SIA can be neatly demonstrated.

PIA assesses both technologies and practices with regard to the risks they impose for infringements of privacy. The assessment process considers projects, technologies, products, services, policies, programs, and other initiatives (Wright et al., 2013). A PIA requires consultation with stakeholders in order to scan for measures that are necessary and appropriate to mitigate the identified negative impacts.

In 2009, the proposed European General Data Protection Regulation (GDPR) acknowledged the importance of PIA, holding that by engaging both industry and civil society stakeholders, EU Member States are to make sure that a framework for privacy and data protection impact assessments is put in place. A revised draft version of the regulation (European Commission, 2011) stresses the importance of concepts such as transparency, dialogue, and privacy by design. This proposed regulation also foresees, as part of the assessment procedure, a description of planned operations, an evaluation of risks and impacts on specific people and groups (and their freedoms and rights). This plan also has to envision appropriate protection and control measures.

A coherent methodological framework governing PIA has not yet emerged. Assessment requirements vary from case to case, both in scale and in scope, and also with regard to the actors involved. Wright and Wadhwa (2012) compiled a list of 16 steps applicable to both PIA and surveillance impact assessment:

1. Determine whether a PIA (or surveillance impact assessment) is necessary (threshold analysis).
2. Identify the PIA (or surveillance impact assessment) team and set the team's terms of reference, resources, and time frame.
3. Prepare a PIA (or surveillance impact assessment) plan.
4. Determine the budget for the PIA (or surveillance impact assessment).
5. Describe the proposed project to be assessed.
6. Identify stakeholders.
7. Analyze the information flows and other impacts.
8. Consult with stakeholders.
9. Determine whether the project complies with legislation.
10. Identify risks and possible solutions.
11. Formulate recommendations.
12. Prepare and publish the report, e.g., on the organization's Web site.
13. Implement the recommendations.
14. Ensure a third-party review and/or audit of the PIA (or surveillance impact assessment).
15. Update the PIA (or surveillance impact assessment) if there are changes in the project.
16. Embed privacy awareness throughout the organization and ensure accountability (cf Wright and Raab, 2012: p. 615).

This list is similar to others developed in SIA handbooks. It covers the major steps, actors, and topics to be addressed in an impact assessment.

What makes PIA a good case to demonstrate the workings and problems of an SIA is the involvement of complex interlinked technologies, the challenge to find a regulatory framework for a technology that easily escapes traditional juridical wisdom, the pervasiveness of the problem affecting practically the whole of the population in Western democracies, the role of the state as a relevant stakeholder with conflicting interests (guaranteeing by the same token security and privacy of the citizens), the a-territorial character of the domain to be regulated (i.e., cyberspace), and differential levels of awareness of the problem among the stakeholders.

The primary motivation for a PIA can flow from different sources: either the industry providing privacy affecting services is working under a regulatory regime, making data-protection mandatory or a group of civil society actors voices concern over the dangers of unrestricted data processing capabilities by industry or public authorities. With PIA it has often been the academic community of legal scholars or surveillance studies experts raising the issue of data protection and rallying for a PIA. As with many other technologies the immediate effects, potential hazards, and threats are not easily visible for the lay user despite the fact of being affected on a daily basis. Having a regulatory framework in place making SIA mandatory can be an asset in such a situation, since the force of law can stand in for the force of public concern. The problem though is that PIA recommendations often entail the element of an active user. In order to provide for adequate privacy protection citizens have to adhere to certain rules and practices and thus become activated in implementing PIA recommendations. At this point an intricate problem can be analyzed: when assessing the impact of a technology, program, or project participatory processes of stakeholder involvement often produce ambiguous results with regard to the evaluation of the impact as positive and desirable or negative. While in the context of a PIA, a high standard of privacy and data protection as such may be desirable the costs involved in maintaining such high levels have to be weighed against the gains. The textbook case for this is the use of loyalty cards. While such costumer convenience systems regularly have tremendous effects on the privacy of the users (collecting data about shopping habits and provide the basis for social sorting), they nonetheless are often seen as convenient gadgets from the consumers' perspective. Trading in privacy for convenience is a standard dilemma encountered in the context of PIA. A similar dilemma emerges when public authorities such as law enforcement agencies collect and process person-related data of citizens. Here the trade-off between privacy and security emerges. This points to another fundamental problem of impact assessment projects. Security as an impact defies any rational measurement when defined as the prevention of terrorist attacks. The threat of such attacks provides the pretext for severe privacy-infringing practices by law enforcement agencies. But any positive effects (i.e., nonevents) can hardly be measured, while the negative consequences (increased surveillance and control) are often quite obvious. Also it can be difficult to determine the scope of the phenomena to be assessed. When data and information are traveling at the speed of light around the globe, the problem of privacy and data protection at any given locale or societal context tends to become a global problem in the literal sense, much as global warming or global pollution. Binding legislation is limited to the nation-state and whatever happens beyond these limits is *de facto* beyond the jurisdiction of the state.

Taking PIA as a test case for the programmatic promises of SIA can demonstrate the limits of this approach while at the same time providing arguments for the urgency of adequate impact assessment measures. SIA needs to be developed further to handle problems of the type surfacing in PIA: ambiguous assessments, deterritorialization, limitations of traditional regulatory regimes, and the need of active involvement of stakeholders, not only in the assessment processes but beyond.

Future Perspective: Toward Balanced Risk Awareness

Living in a society soaked with many techno-social systems requires a clear understanding of the effects of technology. Against antitechnological attitudes reemerging in a kind of neo-Luddite move SIA can help to develop balanced risk awareness and open up new institutional spaces for the deliberation of developmental trajectories, using new and different technologies to overcome or curtail the global risks emerging from a naive use of technology. This at the same time requires a new techno-literacy and a democratic political culture of participation in decisions about future developments. SIA can help to foster both and thus will become more relevant.

See also: Awards, Public Recognition, and Evaluation; Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research; Ethical Questions in Social and Behavioral Sciences, History of; Needs Assessment; Participatory Action Research in Social Research; Policy Analysis; Prevention Research; Science and Technology Studies, History of; Social Network Analysis; Unintended Consequences: History of the Concept.

Bibliography

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, New Delhi.
Becker, H.A., 2001. Social impact assessment. European Journal of Operational Research 128, 311–321.
Bishop, P., Davis, G., 2002. Mapping public participation in policy choices. Australian Journal of Public Administration 61 (1), 14–29.
Carson, R., 1962. Silent Spring. Houghton Mifflin Co., Boston.
Esteves, A.M., Franks, D., Vanclay, F., 2012. Social impact assessment: the state of the art. Impact Assessment and Project Appraisal 30 (1), 34–42.
European Commission, 2011. Privacy and Data Protection Impact Assessment Framework for RFID Applications. Available at: http://cordis.europa.eu/fp7/ict/enet/documents/rfid-pia-framework-final.pdf (accessed 29.07.13.).
Freudenburg, W.R., 1986. Social impact assessment. Annual Review of Sociology 12, 451–478.
Harvey, B., 2011. Foreword: SIA from a resource developer's perspective. In: Vanclay, F., Esteves, A.M. (Eds.), New Directions in Social Impact Assessment: Conceptual and Methodological Assumptions. Edward Elgar, Cheltenham, pp. xxvii–xxxiii.
Kemp, D., 2011. Understanding the organizational context. In: Vanclay, F., Esteves, A.M. (Eds.), New Directions in Social Impact Assessment: Conceptual and Methodological Assumptions. Edward Elgar, Cheltenham, pp. 20–37.
Luhmann, N., 1993. Risk. A Sociological Theory. A.de Gruyter, New York.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data. A Revolution that Will Transform How We Live, Work and Think. Houghton Mifflin Harcourt, Boston, New York.
Moynihan, D.P., 1993. Defining deviancy down. American Scholar 62 (1), 17–33.
O'Faircheallaigh, C., 2010. Public participation and environmental impact assessment: purposes, implications, and lessons for public policy making. Environmental Impact Assessment Review 30, 19–27.
Peterman, W., 2004. Advocacy vs. collaboration: comparing inclusionary community planning models. Community Development Journal 39 (3), 266–276.
Prainsack, B., 2014. Understanding participation: the 'citizen science' of genetics. In: Prainsack, B., Schicktanz, S., Werner-Felmayer, G. (Eds.), Genetics as Social Practice. Ashgate, Farnham.
Rip, A., Schot, J.W., 2002. Identifying loci for influencing the dynamics of technological development. In: Sørensen, K.H., Williams, R. (Eds.), Shaping Technology, Guiding Policy: Concepts, Spaces and Tools. Edward Elgar, Cheltenham, pp. 155–172.
Schot, J., Rip, A., 1996. The past and future of constructive technology assessment. Technological Forecasting and Social Change 54 (2–3), 251–268.
Vanclay, F., 2003. International principles for social impact assessment. Impact Assessment and Project Appraisal 21 (5), 5–11.

Vanclay, F., 2006. Principles for social impact assessment: a critical comparison between the international and US documents. Environmental Impact Assessment Review 26 (1), 3–14.

Vanclay, F., Esteves, A.M., 2011. Current issues and trends in social impact assessment. In: Vanclay, F., Esteves, A.M. (Eds.), New Directions in Social Impact Assessment: Conceptual and Methodological Assumptions. Edward Elgar, Cheltenham, pp. 3–19.

Wright, D., Raab, C.D., 2012. Constructing a surveillance impact assessment. Computer Law & Security Review 28, 613–626.

Wright, D., Finn, R., Rodrigues, R., 2013. A comparative analysis of privacy impact assessment in six countries. Journal of Contemporary European Research 9 (1), 160–180.

Wright, D., Wadhwa, K., 2012. A step-by-step guide to privacy impact assessment. In: Empirical Research on Contextual Factors Affecting the Introduction of PIA Frameworks in EU Member States, Poland (April). Available at: http://www.piafproject.eu/ref/A_step-by-step_guide.pdf (last accessed 06.09.13.).

Societies, Types of

Guillermina Jasso, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Deliège, volume 21, pp. 14530–14536, © 2001, Elsevier Ltd.

Abstract

Societies may be classified along many dimensions – large or small, geographic location, resource base, form of government, kind of economy, and so on. This article describes two new approaches to classifying societies, both grounded in basic sociobehavioral theory, specifically, on Plato's and Aristotle's idea that the ways and means of seeking happiness generate distinctive types of societies. The first approach arises from the new unified theory (NUT) of sociobehavioral forces, which posits three forces from the Mertonian middle range and yields five types of societies (echoing Plato's hint). The second approach is based on four deeper engines of behavior, and yields 15 types of societies (four single-engine societies, six dual-engine societies, four triple-engine societies, and one quad-engine society) plus a null society, with no visible engine of social behavior – for a total of 2^n or 16 societies. In the language of happiness, the first approach is based on three middle-range carriers of happiness and the second on four deeper carriers of happiness. An as yet unanswered question is whether the middle-range carriers of happiness can be derived from the deeper carriers of happiness, or, equivalently, whether the NUT societies can be derived from the deeper-engine societies.

Introduction

Thinking about human behavioral and social phenomena leads in a natural way to classifications of societies. Societies may be classified along many dimensions – large or small, geographic location, resource base, form of government, kind of economy, and so on. Deliège (2001), writing in the first edition of this Encyclopedia, provides a useful overview of such classifications. Grusky and Ku (2008, pp. 13–18) provide a useful classification of stratification systems, conceptually closely related to types of societies, including a handy table. This article describes two new approaches to classifying societies, both grounded in basic sociobehavioral theory, specifically, on Plato's (*Republic*, Book VIII) and Aristotle's (*Politics*, Book 7, Chapter 8) idea that the ways and means of seeking happiness generate distinctive types of societies. The first approach, based on a recent synthesis of theories of three primordial sociobehavioral outcomes (PSOs) from the Mertonian middle range – justice (shorthand for justice and the other comparison processes), status, and power – yields five types of societies, echoing Plato's hint (Jasso, 2008). The second approach is based on four deeper engines of behavior, and yields 15 types of societies (four single-engine societies, six dual-engine societies, four triple-engine societies, and one quad-engine society) plus a null society, with no visible engine of social behavior – for a total of 2^n or 16 societies. Extending to status and power Rayo and Becker's (2007) evocative idea that comparison is a 'carrier of happiness,' the first approach is based on three middle-range carriers of happiness and the second on four deeper carriers of happiness. An as yet unanswered question is whether the middle-range carriers of happiness can be derived from the deeper carriers of happiness, or, equivalently, whether the NUT societies can be derived from the deeper-engine societies. The second approach has not received even a hundredth of the theoretical and empirical scrutiny of the first, and is included here for the sake of comprehensiveness and to honor the abiding challenge to probe ever more deeply.

Types of Societies, in the New Unified Theory of Sociobehavioral Forces

Brief Overview of the New Unified Theory

The new unified theory (NUT) of sociobehavioral forces integrates theories describing three sociobehavioral processes – status, power, and justice (and the other comparison processes) – generating identity and (partially) happiness and bringing under a single theoretical umbrella diverse mechanisms together with their effects across disparate domains and for both individuals and societies (Jasso, 2008). The integration is made possible by the remarkable similarity of the internal core of the theories.

The key foundational idea can be traced to Plato (*Republic*, Book VIII) and Aristotle, in Aristotle's (*Politics*, Book 7, Chapter 8) words: "Different men seek after happiness in different ways and by different means, and so make for themselves different modes of life and forms of government." The new theory formalizes the 'different ways and … different means' of seeking happiness by the operation of three sociobehavioral forces, in which the PSOs which give each force its name (status, power, and justice) are generated by distinctive mechanisms from personal quantitative characteristics (such as beauty and wealth) within groups formed by categories of personal qualitative characteristics (such as nativity, race, and gender). To illustrate, status may be generated from academic skill within a classroom, or self-esteem from tennis-playing skill on the court. Each bundle of elements, say, status–skill–classroom, simultaneously generates an identity and a magnitude of happiness. Different combinations of elements – e.g., power–wealth–club, status–horsemanship–regiment – generate distinctive identities and magnitudes of happiness.

A person's time series of bundles – or, equivalently, time series of identities – together with the associated configuration of elements, describes the person's inner sociobehavioral life in the pursuit of happiness. Some persons dwell on beauty, others on skill, others on possessions, and so on (an aspect of reality captured in Shakespeare's Sonnet 91); concomitantly, some

persons inhabit worlds bounded by race, nativity, or sex; and while some are moved by status, others fix on justice or power.

The configurations of elements in the sociobehavioral profiles of individuals in turn produce the societal configuration. Thus, both individuals and societies come to be seen as materialistic or status-conscious or racialist – or, alternatively, as even-keeled, absent any great preference for status over self-esteem, or wealth over intelligence, or nativity over gender. Like persons, societies are also dynamic, with a time series of configurations based on the profiles of individuals. And models based on the NUT yield a large number of testable predictions (both intuitive and counterintuitive, and including novel predictions) for a wide range of phenomena at both micro- and macrolevels and across disparate topical domains, revealing the thread from the inner sociobehavioral life to the myriad of individual sentiments and behaviors and finally to social phenomena at every scale, from dyads to small groups to nation-states and multinational entities.

A Closer Look at the NUT

The NUT begins with three sets of elements: (1) personal quantitative characteristics, (2) personal qualitative characteristics, and (3) PSOs. The three combine to provide a foundational relation: PSOs are generated by personal quantitative characteristics within groups formed by personal qualitative characteristics.

Personal quantitative characteristics are characteristics of which there can be more or less, or on which a person may rank higher or lower. The first kind – of which there can be more or less – is called *cardinal*; examples are wealth, land, head of cattle. The second kind – on which a person may rank higher or lower – is called *ordinal*; examples include, beauty, intelligence, musical skill.

Personal quantitative characteristics are called *goods* if more is preferred to less and *bads* if less is preferred to more. To illustrate, for most people, money is a good and time in prison is a bad. For simplicity, this article focuses on goods. However, extension to bads is straightforward.

Personal qualitative characteristics are unorderable, categorical personal characteristics. They may be dichotomous – like gender – or polytomous – like race, ethnicity, or religious affiliation. Personal qualitative characteristics give rise to groups and subgroups, with these groups and subgroups defined by the category of a personal qualitative characteristic. For example, the group 'residents of France' is defined in terms of a category of the personal qualitative characteristic 'country of residence,' and the subgroups 'women' and 'men' are defined in terms of categories of the personal qualitative characteristic 'gender.'

The basic idea that there are two kinds of personal characteristics, quantitative and qualitative, and that they differ in their sociobehavioral operation was pioneered by Blau (1974).

The PSOs are outcomes universally desired and deeply social. The sociological literature suggests that there are three – justice (shorthand for the comparison processes), status, and power. As noted, all three are generated by personal quantitative characteristics within groups formed by personal qualitative characteristics. For example, status may be produced by wealth within the population of a city, or by tennis-playing skill within the membership of a country club, or by horsemanship within a military regiment. Associated with each of the three PSOs is a distinctive mechanism and, in particular, a distinctive rate of change. Status and power arise as functions of a personal quantitative characteristic X; justice and the other comparison processes (like self-esteem) arise from the comparison of an actual magnitude of X to a magnitude envisioned, desired, or thought just, denoted X^* – or, as William James ([1891] 1952, p. 200) put it, the comparison of actualities to potentialities, or success to pretensions. Thus, in the basic equations for status and power, there is only one argument, X (or its relative rank), but in the basic equation for justice, there are two arguments, X and X^*.

What kinds of mechanisms connect the Xs and the PSOs? It is well-known that all the PSOs can be generated by the same quantitative characteristic; for example, beauty and wealth generate status, and they generate power, and they generate the comparison outcomes like self-esteem and the sense of justice. If all three PSOs are functions of the same quantitative characteristics, the question arises: What distinguishes them? Or are justice, status, and power merely different names for the same outcome? If we accept, at least provisionally, the possibility that the three PSOs are indeed distinct outcomes, then the challenge is to find the distinguishing feature. A strong contender is the *rate of change*.

In nature there are three rates of change – increasing, decreasing, and constant – and thus a natural and immediate conjecture is that the three PSOs each have a distinctive rate of change and thus the task is to match the three rates of change to the three PSOs. **Figure 1** illustrates the three rates of change in relations in which the PSO increases as X increases. For convenience, X is represented by a relative rank, assuming values from zero to one. But the graphs could as easily be drawn with cardinal magnitudes of X, and all the theoretical development below incorporates both magnitudes and relative ranks of X.

Which sociobehavioral outcome increases at an increasing rate? And which at a decreasing rate? And which at a constant rate?

The sociological literature, in work dating to the early 1950s, suggests that as the quantitative characteristic X increases, status increases at an *increasing* rate (Bales et al., 1951; Goode, 1978; Sørensen, 1979; Stephan, 1952; Stephan and Mishler, 1952) and the comparison outcomes at a *decreasing* rate (Blau, 1964; Jasso, 1978; Wagner and Berger, 1985). Because the literature does not provide a functional form for the relation between the Xs and power (Webster, 2006), the summary proposed by Jasso (2008) suggests that if power is indeed a basic and distinct sociobehavioral outcome – and not merely a synonym for justice or status – then it must vary at a constant rate with the personal quantitative characteristic.

To elaborate briefly, the convexity property of status is evident in the research on frequency of participation in small-group discussions conducted by Bales et al. (1951), Stephan (1952), and Stephan and Mishler (1952), as well as in the range of social situations analyzed by Goode (1978). Meanwhile, the justice and comparison literature notes that the negative second derivative is necessary and sufficient for the properties of deficiency aversion and loss aversion, which are considered essential in a justice evaluation function (Brown, 1986, p. 78; Jasso, 1978; Wagner and

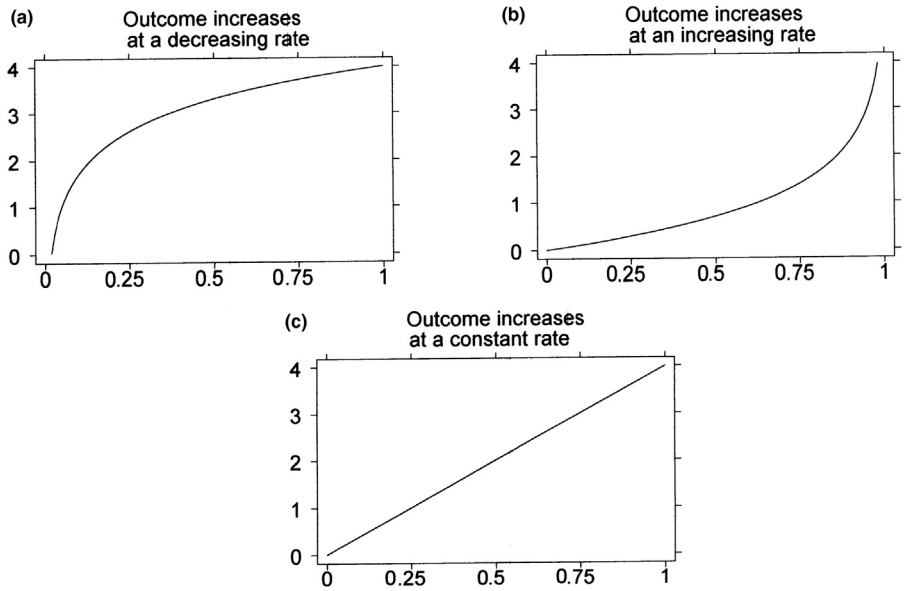

Figure 1 Illustration of three possible rates of change in an increasing function.

Berger, 1985, p. 719) and in comparison processes more generally (Jasso, 1990).

Of course, many functional forms display the increasing or decreasing rate of change proposed for status and comparison, respectively. The sociological literature goes beyond the rate of change and suggests specific functional forms for the status and comparison PSOs. For status, Sørensen (1979) proposed the following function, a function which embeds Goode's (1978) convexity condition:

$$S = \ln\left(\frac{1}{1 - r_X}\right), \qquad [1]$$

where S denotes status, X as before denotes the valued good (i.e., the valued attribute or possession), and r denotes the relative rank on the valued good. Although the valued good can be cardinal or ordinal, the status function notices only its relative rank.

In the case of comparison processes, Jasso (1978, 1990) proposed the following function:

$$Z = \theta \ln\left(\frac{X}{X^*}\right), \qquad [2]$$

where Z denotes any of the comparison outcomes, such as self-esteem or the justice evaluation, X is as above the valued attribute or possession (called in this literature the actual reward), X^* denotes the comparison referent (called the just reward in the special case of justice), and θ is the signature constant whose sign indicates whether the reward is viewed as a good or a bad and whose absolute value denotes expressiveness. When the actual reward equals the comparison referent, the outcome is zero (a neutral point which in the case of justice is called the point of perfect justice); when the actual reward exceeds the comparison referent, the outcome is positive (representing overreward), and when the actual reward is less than the comparison referent, the outcome is negative (representing underreward). Besides satisfying deficiency aversion and loss aversion, this function is also the only function which satisfies two additional conditions, scale invariance and additivity (Jasso, 1990). By additivity, the effect of X on Z is independent of the magnitude of X^*, and the effect of X^* on Z is independent of the magnitude of X. Scale invariance ensures that the outcome Z will be the same regardless of the units in which the reward is measured.

The expressiveness part of the signature constant θ plays an important part in empirical work but can be safely ignored in much theoretical work. The framing part of the signature constant is always important, but can safely be set as positive, given that results for bads are readily established from results on goods. Accordingly, in the rest of this article, the signature constant is fixed at $+1$.

If we provisionally accept the reasoning in the literature and if power is a sociobehavioral force in its own right, different from either justice or status, then we may posit a linear function:

$$P = a + bX \qquad [3]$$

where P denotes power and a and b are the intercept and slope, respectively. In the special case in which the intercept is zero and the slope is one, the function reduces to an identity function.

Power and the comparison processes are also generated by ordinal characteristics, such as beauty and skills of all kinds, or by the relative rank on a cardinal good. When the PSO is status or when the valued goods are ordinal (or the ordinal aspect of cardinal goods), personal qualitative characteristics provide the group or population within which the relative ranks are measured. As well, in the justice case, when the comparison referent is a summary characteristic of a distribution (such as the arithmetic mean, geometric mean, or median), personal qualitative characteristics provide the group or population within which the referent is measured. For example, those groups may be specified as a category of gender, nativity, citizenship, religion, or other qualitative characteristic.

The three PSOs (status, power, and justice) are thought to generate identity, emotion, and a wide range of behavioral and social phenomena. As well, extending to status and power the evocative language used by Rayo and Becker (2007) for comparison, the three PSOs are thought to operate as 'carriers of happiness.' In brief, during intervals when a particular PSO is active, it produces identity and generates happiness. Of course, happiness may also be generated by nonsocial phenomena, such as a beautiful sunset or the fruit of a tree.

Thus, an identity is a three-element bundle, consisting of the PSO, the quantitative characteristic, and the qualitative characteristic – for example, status-wealth-classroom or power-intelligence-dormitory. Persons may be thought of as collections of identities; they come to be characterized by their configuration of elements in the identities, termed *personality*. Similarly, societies are collections of persons; they come to be characterized by their configuration of elements in their members' identities, termed *culture*. Examples of personality in this technical theoretical sense are the status-obsessed, beauty-fixated, and gender-conscious; examples of culture include jock culture, materialistic culture, and racialist culture.

Finally, because status notices only relative ranks while justice and power respond to both amounts and relative ranks, there are five types of societies:

- justice-materialistic
- justice-nonmaterialistic
- status
- power-materialistic
- power-nonmaterialistic

Given that the PSOs generate happiness, this is a modern expression of Plato's idea that there are five types of societies, corresponding to five ways that people seek after happiness (Jasso, 2008).

An immediate natural question is: What do these societies look like? The PSO distribution provides a picture – symmetric or asymmetric, with varying proportions of happy and unhappy people, with extreme values that can range, or not, to depths of unhappiness and heights of happiness, with greater or lesser inequality.

In status, justice-nonmaterialistic, and power-nonmaterialistic societies, the valued good has an ordinal representation, and thus information about the PSO distribution is immediate. The status distribution is a negative exponential, the justice-nonmaterialistic is a positive exponential, and the power-nonmaterialistic is a linear transformation of the rectangular (corresponding to relative ranks). However, in the two materialistic societies, it is necessary to specify the distribution of the valued cardinal good in order to obtain the PSO distribution. Thus, there can be many subtypes of the justice-materialistic and power-materialistic societies, depending on the distributional form of the valued cardinal good.

As discussed above, in seeking happiness, individuals may cycle over different combinations of PSOs, quantitative characteristics, and qualitative characteristics. At each turn of the sociobehavioral wheel, so to speak, a new identity is generated and a new magnitude of happiness. Societies, too, exhibit this dynamic, changing type as their members change.

Note that the old idea that good things come in threes – *omne trium perfectum* (everything that comes in threes is perfect) – is manifested twice in the NUT: in the three elements (quantitative characteristic, qualitative characteristic, and PSO) and in the three PSOs (justice, status, and power).

Subtypes of Materialistic Societies in the NUT: Distributional Form of Valued Cardinal Goods

To model a valued cardinal good X, a useful set of distributions includes distributions with different combinations of lower and upper bounds, such as distributions with and without a safety net and distributions with and without a maximum amount of X. Here we use three two-parameter continuous univariate distributions whose properties are well-known and which are widely used to model income distributions – the lognormal, Pareto, and power-function (Johnson et al., 1994, 1995; Kleiber and Kotz, 2003). Together this set includes some desired combinations. The Pareto combines a safety net with no maximum. The lognormal lacks an upper bound, like the Pareto, but, unlike the Pareto, it has no safety net. The power-function has a maximum but no safety net.

A further useful feature of this set is that the power-function includes as a special case the continuous uniform which represents a distribution of relative ranks. Thus, the same analysis that establishes results for quantitative Xs also provides results for justice-nonmaterialistic and power-nonmaterialistic societies.

Perhaps an even more important feature of this set of three modeling distributions is that one of the two parameters governs all measures of inequality (Jasso and Kotz, 2008). Thus, in a theoretical analysis of inequality using probability distributions it is not necessary to examine several measures of inequality, as all are monotonic functions of this one parameter, called the general inequality parameter.

While all five types of societies may have subtypes based on the number and associations among valued goods (and in the case of justice societies, the association between the actual reward and the comparison referent), only the two materialistic societies – justice-materialistic and power-materialistic – have subtypes based on the distributional form of the cardinal goods.

Table 1 provides the PSO distribution corresponding to the five types of societies and three subtypes each for the two

Table 1 PSO distributions in the five types of societies in the new unified theory of sociobehavioral forces, with cardinal good distributions ascribed in the two materialistic societies

Type of society/Distributional form of valued good in materialistic society	*PSO distribution*
1. Justice-nonmaterialistic	Positive exponential
2. Justice-materialistic	
Lognormal	Normal
Pareto	Negative exponential
Power-function	Positive exponential
3. Status	Negative exponential
4. Power-nonmaterialistic	Rectangular
5. Power-materialistic	
Lognormal	Lognormal
Pareto	Pareto
Power-function	Power-function

Note: The justice societies are for the special case in which the comparison referent is constant.

materialistic societies. The justice societies shown are for the special case in which the comparison referent X^* is a constant. As shown, the parsimonious setup yields seven distinct societal configurations of the PSO. In more elaborate cases – for example, justice societies in which X^* varies and/or societies which value multiple goods and the goods may be independent, negatively associated, or positively associated, many additional PSO distributions are generated, including both classical distributions such as the equal, Laplace, logistic, and Erlang and also new distributions such as the mirror-exponential obtained for the case of a status society with multiple negatively associated goods (Jasso and Kotz, 2007).

Distributions provide pictures of societies. This approach makes plain that societies – and their pictures – differ by the carrier of happiness and the number, distributional form, and associations of valued goods. For visualization, **Figure 2** provides graphs of the quantile function for an ordinal good and three cardinal goods (modeled by the lognormal, Pareto, and power-function distributions), as well as for the basic types and subtypes of societies to which they lead (shown in **Table 1**).

Subgroups – PreExisting and Emergent

Subgroups defined in terms of the personal qualitative characteristics introduced above are called *preexisting subgroups* because they are already formed before there is any sociobehavioral operation. Operation of the PSOs generates new subgroups – called *emergent subgroups* – which crosscut the preexisting subgroups. For example, operation of the justice PSO produces three emergent subgroups – consisting of individuals who are underrewarded, justly rewarded, and overrewarded – and these crosscut preexisting subgroups (such as those formed by race or gender).

When the group or population has a preexisting subgroup structure generated by a qualitative characteristic, each individual has access not only to the identity generated by his or her magnitude or relative rank in the quantitative characteristic – now called the Personal Identity – but also to a new identity called the Subgroup Identity and defined as the average of the Personal Identities within the preexisting subgroup (Jasso, 2008). The Personal Identity and the Subgroup Identity are used in several models, as discussed below.

Identity and Intersubgroup Relations

Consider a society with two preexisting subgroups. In this section we discuss two models that have been formulated. In the first model, everyone thinks and acts as a member of the subgroup. In the second model, each person seeks to maximize own happiness.

First Model – Homogeneous PreExisting Subgroups

Suppose that each person attaches to the Subgroup Identity – a sociological application of the Newtonian principle that mass accumulates at a point. Then relations between the two preexisting subgroups can be understood as the relation between the two Subgroup Identities. In particular, the social distance between the two preexisting subgroups can be defined as the difference between the two Subgroup Identities.

For a priori theoretical analysis, suppose that the two preexisting subgroups are nonoverlapping in the valued good (and hence in the PSO). If wealth is the valued good, the poorest person in the top subgroup is richer than the richest person in the bottom subgroup. We can then address classic sociological questions, including questions about the effect of the proportion in the disadvantaged group and the effect of wealth inequality.

Theoretical analysis yields two main initial results. First, the effect of the relative size of the two preexisting subgroups on the social distance between them varies considerably across the different societies. Expressed in terms of the proportion in the bottom subgroup, the effect on social distance may be increasing, decreasing, or nonmonotonic, depending on the PSO and the distributional form of the valued good. Second, in the materialistic societies, social distance always increases with inequality in the good's distribution. **Table 2** summarizes these results, and **Figure 3** provides visual illustration.

Second Model – Heterogeneous PreExisting Subgroups

Suppose now that persons seek to enhance their identity – by maximizing their PSO score. Equivalently, given that the PSO also generates happiness, suppose that persons seek to maximize their happiness. Then they will compare the Personal Identity with the Subgroup Identity (as discussed in Jasso, 2001b, 2008). If Personal Identity is higher than Subgroup Identity, they will ignore their subgroup and attach to the idea of themselves and others as individuals, blind to their own and others' subgroup origins. But if Personal Identity is lower than Subgroup Identity, they will attach to the subgroup and ignore their Personal Identity, finding the happiness they seek within their preexisting subgroup.

In a society with two subgroups, this identity enhancing process will yield three new emergent subgroups – persons from preexisting subgroup A who attach to A, persons from preexisting subgroup B who attach to B, and persons from both preexisting subgroups who attach to self. Persons in the first two of the new subgroups are called Subgroupistas, persons in the third are called Selfistas.

To illustrate, **Figure 4** provides graphs of the status PSO in four hypothetical societies, each with two preexisting subgroups such that the preexisting subgroups are nonoverlapping in status (and in whatever whatever personal quantitative characteristic generated status). The four groups differ in the subgroup split, that is, the proportions in the two preexisting subgroups, ranging from a group in which the bottom preexisting subgroup has 25% of the population and the top preexisting subgroup has 75% to a group in which the bottom preexisting subgroup has 75% of the population and the top preexisting subgroup has 25%. A vertical line divides the two preexisting subgroups. The curve in each plot depicts Personal Identity – i.e., each individual's own status score. The long horizontal line represents the average status for the whole group, and the two smaller horizontal lines represent the average status for the two preexisting subgroups, that is, the Subgroup Identity. The intersection of the status curve and the Subgroup Identity marks the point at which Personal Identity equals Subgroup Identity. Regardless of the subgroup split, each preexisting subgroup has both Selfistas (the subset to the right of the intersection) and Subgroupistas (the subset to the left of the

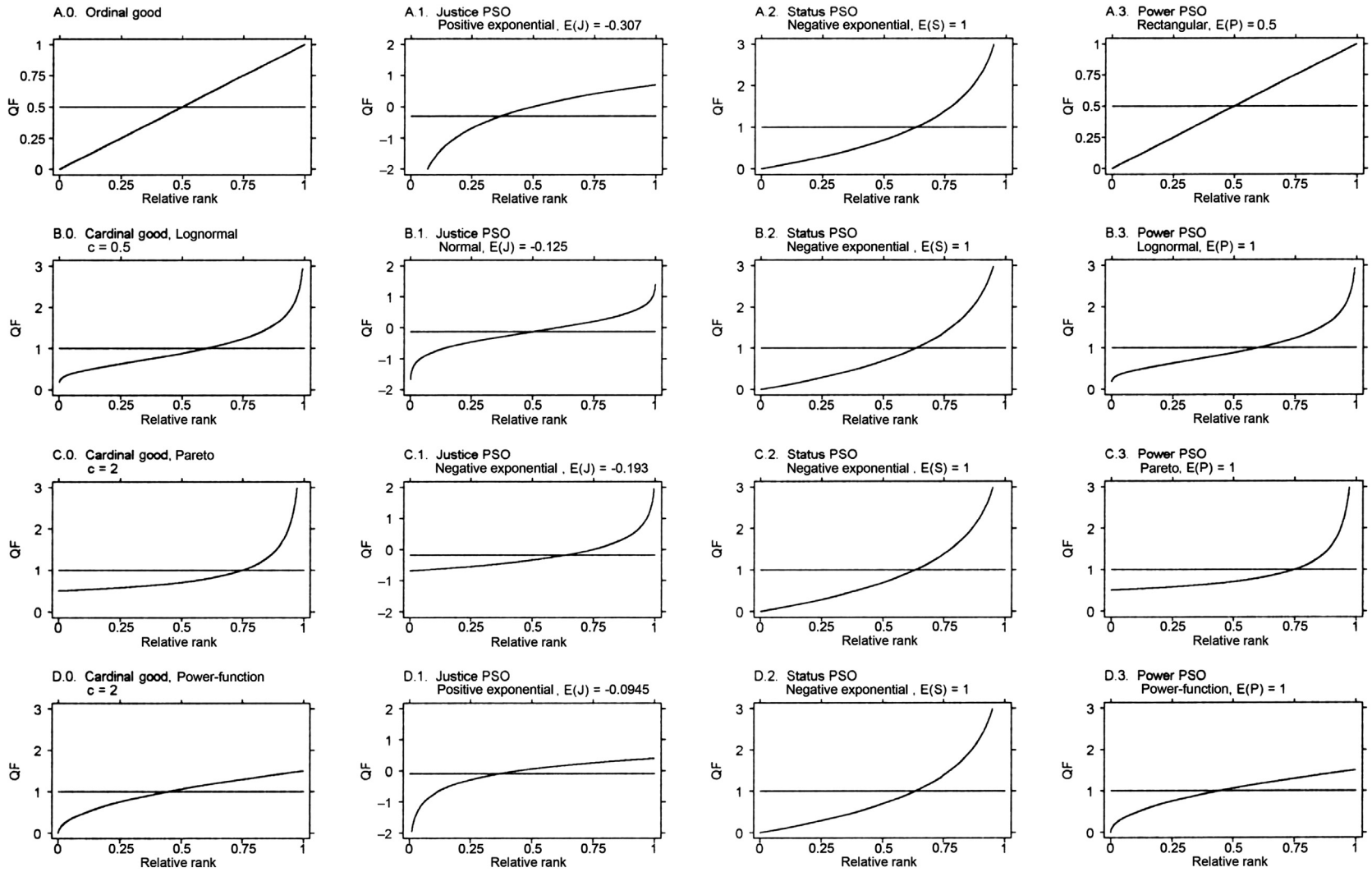

Figure 2 Quantile function of PSO distributions arising from valued goods, with cardinal goods modeled by three variates: single-good case. The distributions of the valued goods are in the .0 column and the PSO distributions in the .1, .2, and .3 columns. The justice PSO distribution is in the .1 column, the status PSO distribution in the .2 column, and the power PSO distribution in the .3 column. The justice PSO distributions pertain to the case where the comparison referent X^* is constant. All the graphs include a horizontal line at the arithmetic mean.

Table 2 Effect of proportion in the disadvantaged subgroup on social distance, in the five types of societies in the new unified theory of sociobehavioral forces, with cardinal good distributions ascribed in the two materialistic societies

Type of society/Distributional form of valued good in materialistic society	*As proportion in disadvantaged subgroup increases, social distance …*
1. Justice-nonmaterialistic	Decreases
2. Justice-materialistic	
Lognormal	Decreases, then increases, in symmetric U
Pareto	Increases
Power-function	Decreases
3. Status	Increases
4. Power-nonmaterialistic	Is constant
5. Power-materialistic	
Lognormal	Increases
Pareto	Increases
Power-function	Decreases

Notes: The justice societies are for the special case in which the comparison referent is constant. In the materialistic societies, social distance increases with the cardinal good's inequality.

intersection). Thus, it is always the case that the richest or most beautiful or most accomplished in each preexisting subgroup will be Selfistas, while the poorest or least beautiful or least accomplished will be Subgroupistas. This in turn suggests that 'the best and the brightest' in each preexisting subgroup cannot be trusted with important work for the preexisting subgroups.

In an application to residential segregation, this identity enhancement model predicts three emergent subgroups: (1) members of the bottom subgroup who are segregationist; (2) members of the top subgroup who are segregationist; and (3) members of both subgroups who are blind to their subgroups and thus integrationist. In the special case of blacks and whites, the model predicts three types of neighborhoods: (1) all-black; (2) all-white; and (3) mixed.

New quantities arise, adding to the proportions in the three main emergent subgroups the following: (1) proportions black and white among integrationists and among segregationists; and (2) proportions integrationist and segregationist among blacks and among whites.

It will be no surprise to learn that the NUT gives rise to a dazzling diversity of societies, with intricately differing configurations of all the quantities. The exact outcomes depend not only on the proportions in the preexisting subgroups, as in Figure 4, but also on the configuration of (1) the active PSO, (2) the number and associations among the valued goods, and (3) in the two materialistic societies, the distributional form of valued cardinal goods (Jasso, 2008). For example, in a status society the proportion integrationist is always less than half (both in the overall society and in each preexisting subgroup), but in a justice-nonmaterialistic society the proportion integrationist is always more than half (again both in the overall society and in each preexisting subgroup).

In sum, the new unified theory of sociobehavioral forces leads to five types of societies – consistent with Plato's intuition – two of these spawning subtypes. This parsimonious beginning generates a wide variety of outcomes. The seemingly mysterious effects of subgroup relative size on social distance,

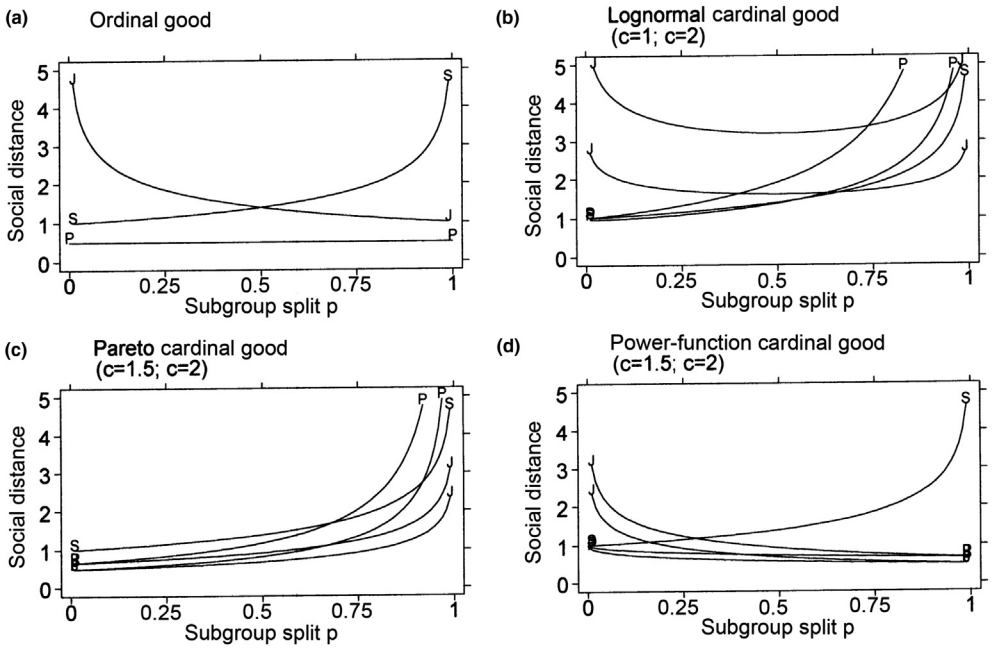

Figure 3 Social distance in justice, status, and power societies, with differing configurations of valued goods and goods' distributional forms, by subgroup split *p*. The letters J, S, and P denote the justice, status, and power societies, respectively. In the three cardinal-good societies, the greater the inequality in the good's distribution, the greater the social distance (and the higher the vertical placement of the curve). Source: Jasso, Guillermina., 2010. Linking individuals and societies. Journal of Mathematical Sociology 34, 40.

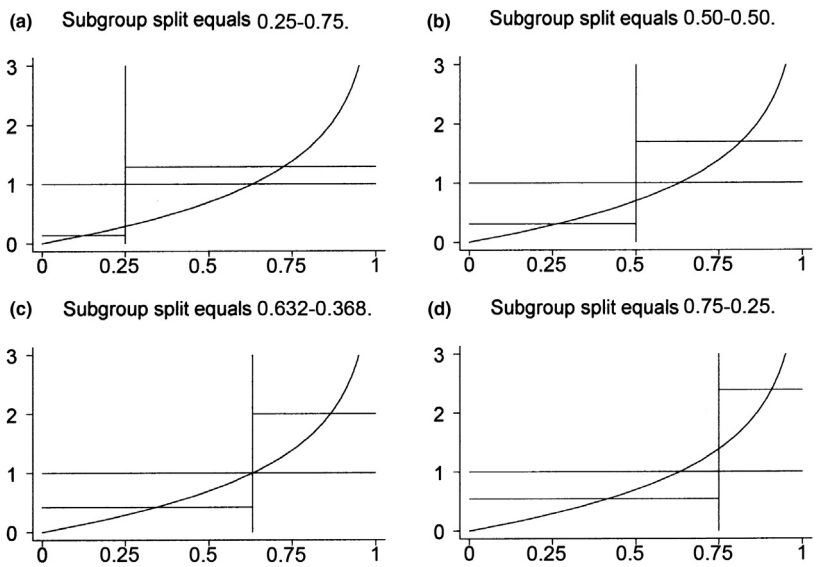

Figure 4 Selfistas and subgroupistas in four hypothetical groups, where status is the dominant PSO but the preexisting subgroup splits differ. Source: Jasso, Guillermina., 2008. A new unified theory of sociobehavioral forces. European Sociological Review 24, 422.

now increasing, now decreasing, the similarly mysterious configurations of segregationists and integrationists, all may be the orderly consequence of a few basic principles.

Types of Societies in the Deeper-Forces Model

Suppose that justice, status, and power – the engines of the NUT – are not the most fundamental forces generating human behavior. Suppose they are themselves generated by deeper forces, deeper carriers of happiness. This section briefly considers four candidates for deeper forces discussed in Jasso (2001a, pp. 350–352):

- to know the causes of things
- to judge the goodness of things
- to be perfect
- to be free

The first candidate-force is rooted in the individual's thirst for reliable knowledge about the nature of the world, a thirst for knowledge visible in the earliest pages of Genesis, in the history of philosophy and science, in a variety of 'descriptive' and 'positive' activities, and in people's everyday lives. For example, Virgil writes in the *Georgics* (II, 1, 490), "Felix cui potuit rerum cognoscere causas" ("Happy the person who could search out the causes of things"). It leads to what may be called a Science Society.

The second candidate-force pertains to the individual's drive to judge, to discern what is good and what is bad; it is visible in a variety of 'prescriptive' and 'normative' activities and, like the first candidate-force, in people's everyday lives. It leads to what may be called a Philosophy Society.

The third candidate-force reflects the individual's quest for perfection. As Aquinas (*Summa Theologica*, II, Part I, Q.5) put it, "Humans seek their own perfection." They desire "a well-disposed body" and "keenness of wit and beauty of body" (Aquinas, *ST* II, I, 4). Or as an American teenager describes her ballet classes at the Bolshoi, "I love how strict they are, how every movement is towards perfection" (Milzoff, 2014). Of course, it is difficult, often impossible, to alter the endowments dealt by nature and it may take long and arduous years to augment material possessions. Yet this seeking after perfection does not result in universal sorrow, for humans possess a faculty for achieving *the illusion of perfection*. There are two pathways: (1) find a combination of valued good and group such that the individual ranks at the top, or (2) acquire large holdings of cardinal goods. This quest for perfection, too, is visible in several literatures dating back to classical antiquity. Adkins (1972) provides valuable discussion of the human obsession with the competitive excellences, first documented among the Greeks. The quest for perfection becomes explicit in the *Confessions* of Augustine, the *Autobiography* and *Spiritual Exercises* of Ignatius of Loyola, and the *Way of Perfection* of Teresa of Avila. And it permeates Cervantes' *Don Quijote*. Cervantes, ever keen on human nature, comments often on this *issimus* phenomenon; for example, he has Don Quijote set out to become the best knight-errant and Sancho the best squire-to-a-knight-errant that ever lived. This candidate-force leads to what may be called a Perfection Society.

The fourth candidate-force – to be free – also appears since antiquity, for example, in Epictetus (1983). And it becomes a rallying cry in more recent centuries. It leads to what may be called a Freedom Society.

If each society had a single engine, this setup would give rise to four types of societies. But one can easily imagine societies with multiple engines – for example, societies which care about both Science and Philosophy, or both Science and Freedom. Having dual engines produces six new types of societies, triple engines spawns four more types, and finally there is one quad-engine society – for a total of 15. If one adds the null case – with none of the four engines – the total comes to 16 (that is, 2^n).

This four-engine model is in its infancy. Future theoretical development might explore the conditions under which one or another type of society would be realized, and it might derive testable predictions, including novel predictions. For now it is possible to suggest that the goods in the NUT reflect the perfection engine, that the comparison referent reflects the judgment engine. Empirically, factorial survey methods for studying beliefs and judgments provide ways to assess operation of the first and second candidate-forces, respectively.

Afterword

As Aristotle (*Nicomachean Ethics*, Book 1, Chapter 7) observed, "One swallow does not make a summer, nor does one day; and so too one day, or a short time, does not make a man blessed and happy." Societies may cycle over different types, as their members switch active carriers of happiness or valued goods or the goods change distributional form. The history of a group or population may be written in their members' time series of carriers of happiness and the ensuing time series of types of societies.

See also: Inequality Analysis: Overview; Social Stratification.

Bibliography

Adkins, A.W.H., 1972. Moral Values and Political Behaviour in Ancient, Greece: From Homer to the End of the Fifth Century. Norton, New York.

Aquinas, St. Thomas, [1265–1274] 1952. Summa Theologica (Translated by Fathers of the English Dominican Province. Britannica, Chicago.

Aristotle, [384–322 BCE] 1952. (W. D. Ross, Trans.). The Works of Aristotle. vol. 2. Britannica Press, Chicago.

Augustine, Saint, 1952. The Confessions (E.B. Pusey, Trans.). Britannica, Chicago. Original first published circa 400.

de Ávila, Santa, Teresa, 1982. Obras Completas. Annotated and with an, introduction by Luis Santullano. Based on the text edited and annotated by P. Silverio. Aguilar, Madrid. Items written between 1546 and 1582.

Bales, Robert F., Strodtbeck, Fred L., Mills, Theodore M., Roseborough, Mary E., 1951. Channels of communication in small groups. American Sociological Review 16, 461–468.

Blau, Peter M., 1964. Exchange and Power in Social Life. Wiley, New York, NY.

Blau, Peter M., 1974. Presidential address: parameters of social structure. American Sociological Review 39, 615–635.

Brown, Roger, 1986. Social Psychology: The Second Edition. Free Press, New York, NY.

Cervantes de Saavedra, Miguel, 1968. In: de Riquer, Martin (Ed.), Don Quijote de la Mancha. Juventud, Barcelona. Parts One and Two, first published in 1605 and 1615, respectively.

Deliège, R., 2001. Societies, Types of. S. 14530-14536. In: International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New York, NY.

Epictetus, 1983. In: White, Nicholas (Ed.), Handbook of Epictetus. Hackett, Indianapolis. Based on the lectures of Epictetus, who lived circa 50-130 A.D.

Goode, William J., 1978. The Celebration of Heroes: Prestige as a Control System. University of California Press, Berkeley, CA.

Grusky, David B., Ku, Manwai C., 2008. Gloom, doom, and inequality. In: Grusky, David B. (Ed.), Social Stratification: Class, Race, and Gender in Sociological Perspective. Westview, Boulder, CO, pp. 2–28.

Homans, George C., 1961. Social Behavior: Its Elementary Forms. Harcourt, Brace and Company, New York.

Ignacio de Loyola, San, 1963. In: Iparraguirre, I., de Dalmases, C. (Eds.), Obras Completas. Espasa-Calpe, Madrid. Items written between 1521, and 1556.

James, William, 1891[1952]. Principles of Psychology. Britannica, Chicago.

Jasso, Guillermina, 1978. On the justice of earnings: a new specification of the justice evaluation function. American Journal of Sociology 83, 1398–1419.

Jasso, Guillermina, 1990. Methods for the theoretical and empirical analysis of comparison processes. Sociological Methodology 20, 369–419.

Jasso, Guillermina, 2001a. Rule-finding about rule-making: comparison processes and the making of norms. In: Hechter, Michael, Opp, Karl-Dieter (Eds.), Social Norms. Russell Sage, New York, NY, pp. 348–393.

Jasso, Guillermina, 2001b. Studying status: an integrated framework. American Sociological Review 66, 96–124.

Jasso, Guillermina, 2008. A new unified theory of sociobehavioral forces. European Sociological Review 24, 411–434.

Jasso, Guillermina, 2010. Linking individuals and societies. Journal of Mathematical Sociology 34, 1–51.

Jasso, Guillermina, Kotz, Samuel, 2007. A new continuous distribution and two new families of distributions based on the exponential. Statistica Neerlandica 61, 305–328.

Jasso, Guillermina, Kotz, Samuel, 2008. Two types of inequality: inequality between persons and inequality between subgroups. Sociological Methods and Research 37, 31–74.

Johnson, Norman L., Kotz, Samuel, Balakrishnan, N., 1994 (1995) Continuous Univariate Distributions, second ed., vol. 1 (2). Wiley, New York, NY.

Kleiber, Christian, Kotz, Samuel, 2003. Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, Hoboken, NJ.

Milzoff, Rebecca, 31 August 2014. A Dancer Dreams across the World: An Alaskan Ballet Dancer Hones Her Craft at the Bolshoi. The New York Times.

Plato [circa 428-348/7 BCE] 1952. The Dialogues of Plato (B. Jowett, Trans.). Britannica, Chicago.

Popper, Karl R., 1963. Conjectures and Refutations: The Growth of Scientific Knowledge. Basic Books, New York.

Rawls, John, 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.

Rayo, Luis, Becker, Gary S., 2007. Evolutionary efficiency and happiness. Journal of Political Economy 115, 302–337.

Sørensen, Aage B., 1979. A model and a metric for the analysis of the intra-generational status attainment process. American Journal of Sociology 85, 361–384.

Stephan, Frederick F., 1952. The relative rate of communication between members of small groups. American Sociological Review 17, 482–486.

Stephan, Frederick F., Mishler, Elliot G., 1952. The distribution of participation in small groups: an exponential approximation. American Sociological Review 17, 598–608.

Wagner, David, Berger, Joseph, 1985. Do sociological theories grow? American Journal of Sociology 90, 697–728.

Webster Jr., Murray, August 2006. Status research since 2000. In: Paper Presented at the Annual Group Processes Conference, Montreal, Canada.

Society and People: History of the Concept

Lars Trägårdh, Ersta Sköndal University College, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'People' is a concept that has come to acquire two related, yet different, meanings. On the one hand, there is the democratic, political conception of 'the people' – the people as *demos* – and on the other there is the ethnic, cultural conception of 'a people' – the people as *ethnos*. In either sense of the word, the crystallization of the modern concept of the people can be traced back to the later part of the eighteenth century, when the idea of the 'sovereignty of the people' became central to democratic political theory and the nation-state emerged as the paradigmatic form of state institution, informed by the principle that ideally each state contained one nation/people, and that each nation/people constituted one state. In this article each meaning will be discussed, as well as their relationship to each other.

'People' is a concept that has come to acquire two related, yet different meanings. On the one hand, there is the democratic, political conception of 'the people' – the people as *demos* – and on the other there is the ethnic, cultural conception of 'a people' – the people as *ethnos*. In either sense of the word, the crystallization of the modern concept of the people can be traced back to the later part of the eighteenth century when the idea of the 'sovereignty of the people' became central to democratic political theory, and the nation-state emerged as the paradigmatic form of state institution, informed by the principle that ideally each state contained one nation/people, and that each nation/people constituted one state. In this article each meaning will be discussed, as well as their relationship to each other.

The 'People' and the Democratic and National Imaginaries

Prior to the eighteenth century the concept of the people was of relatively minor importance. Present from the outset were two basic meanings of the term, a general one, closely related to 'nation,' 'community,' and ' society,' and a particular one, which opposed the people to the elite. Take, for example, Samuel Johnson's *A Dictionary of the English Language* from 1755, which juxtaposes the following definitions: (1) "A Nation, those who compose a community," (2) "The Vulgar," and (3) "The Commonality; not the princes or nobles." Similarly, Abbé Fuertière's *Dictionaire universel* from 1690 lists both a positive, general meaning – "the mass of persons who live in one country, who compose a nation" – and a negative one, defined "by opposition to those who are noble, wealthy, and educated."

During the eighteenth century the concept, which until then had carried largely negative connotations, associated with the vulgar, the lowly, and the beastly – what the Romans referred to as *plebs* – underwent a dramatic change, and 'the people' was valorized and reformulated in distinctly positive terms, as the source of mortal virtue, political legitimacy, and cultural authenticity, a sentiment captured in the increasingly common saying, "the voice of the people is the voice of God." By around 1800 the 'people' was thus constituted as what Reinhart Koselleck calls a 'key-concept' (see Brunner et al. 1972–92). In dictionaries published after 1800 a veritable explosion of neologisms derived from the root words 'people,' '*Volk*,' '*peuple*,' '*narod*,' 'folk,' and so on, can be observed, reflecting what has been called the 'discovery of the people' (Burke, 1994).

This production of new words had two origins. On the one hand, there were the terms derived from the vocabulary of the English, French, and American revolutions, to which the principle of the 'sovereignty of the people' was the central tenet. In opposition to the Old Regime, with its feudal, hierarchical society with the sovereign king at the top of the social and legal pyramid, a new idiom was elaborated that spoke in terms of citizenship, civil liberties, and legal equality. In England the primacy of Parliament and especially the House of Commons gave expression to this new political reality, which also brought with it a new sense of common, popular English nationality. In France the Third Estate – which contained the now celebrated people – became synonymous with the nation, while the nobility was excluded from such membership, associated, as Abbé Sieyès (3 May 1748 to 20 June 1836) put it, with 'incurable idleness,' 'collapse of morality,' and a variety of exclusive and insulting social and political rights, privileges, and prerogatives. And in the United States, the Declaration of Independence spoke of the rights of 'we the people' to assert its sovereignty against the tyrannical and illegitimate powers of the English king.

On the other hand, there were Romantic, ethnographic terms inspired by the German philosopher Johann Gottfried von Herder (25 August 1744 to 18 December 1803) and his followers who ventured into the countryside to explore the very popular culture that hitherto had been the objects of scorn and contempt. In 1774 Herder coined the word *Volkslied*, or 'folksong,' as the title of a collection of songs he published, and a host of similar words pertaining to 'folklore' and popular culture soon became commonplace in most European languages. As in the realm of politics, 'the people' in this ethnographic sense was now to be rescued from oblivion and condescension and instead to be presented as the authentic bearer of national culture. True poetry, Herder suggested, gives expression to the 'soul of the people,' and according to Jacob Grimm (4 January 1785 to 20 September 1863) such authentic, national poems 'belong to the whole people.' For Herder, the

people was first and foremost a matter of history and culture, and he harbored an organic vision of the concept. Each people was thus conceived as a collective individual that possessed its own national character, its own language, its own mores, its own morality, and its own sense of self. In contrast to the political definition of the people as demos, Herder thus stood in the forefront on an elaboration of the people as ethnos. Indeed, Herder tended to perceive the emerging modern state with a great deal of suspicion, as a 'machine,' an 'artificial form of society' that threatens the organic, authentic culture of a people.

From a wider perspective these new terms belonged to a more general phenomenon of linguistic innovation during this period. As Lynn Hunt has observed in the context of the French revolution, 'words came in torrents' as the old political, social, and economic order broke down and attempts were made to semantically capture a new one (Hunt, 1984). As the rate of change accelerated, experience could no longer guide action in the present. A new interpretative language was required to make sense of the world. Thus attempts had to be made to formulate a vocabulary that no longer simply referred to what Koselleck calls 'the space of experience,' but instead reached into the future, creating a 'horizon of expectation,' terms through which prognostic political goals could be hatched (Koselleck, 1985). In other words, the concept of 'the people' acquired a utopian character. As such it came to occupy a central position in the 'democratic imaginary' (Joyce, 1994), as a concept that did not merely serve to describe an existing state of affairs, but also prefigured the future. By elaborating visions of popular sovereignty, a politically potent expectation was created that has come to define the context for politics ever since. Similarly, the people imagined as a form of community – the nation – was a vision that came to inform European and global politics as nationalism became a fundamental political force, erupting in the revolutions of 1848 – 'the springtime of the peoples' – in the unification of Italy and Germany, in the collapse of the Hapsburg Empire after World War I (WWI) in the shadow of President Woodrow Wilson's (28 December 1856–3 February 1924) ideology of the right to national self-determination, and in the politics of anticolonial national liberation movements during most of the twentieth century.

Structural Relations: Who Belongs to the People?

In structural terms, the difference between these two conceptions is one of how membership in 'the people' is delimited. In one case, the inclusion/exclusion is made in inside/outside terms. Here the emphasis is on setting limits against other peoples by defining the peculiar character of one's own people and culture in relation to those of others. One way is to define the people in terms of race and blood, and to determine citizenship according to the strictures of *jus sanguinis*. In its most extreme and lethal form, this logic underpinned the politics of national belonging practiced by the National Socialists in Hitler's Germany, but it has also been a basic legal principle for acquiring automatic citizenship in many other countries. Alternatively, membership may be determined through participation in what Jean Jacques Rousseau (28 June 1712–2 July 1778) called 'civil religion,' allegiance to a national constitution and to the law, a process by which individuals become active citizens, devoted to the nation, and to the practice of civic virtue. In this context, ethnicity tends to play a less defining role, and rights to citizenship may, as in France or the United States, be conferred automatically on anyone born in the country (*jus soli*) (Brubaker, 1992). In either case, citizenship and nationhood is by definition simultaneously a matter of inclusion and exclusion, however benign and civic the criteria for membership in the national community may be. Thus the imagined community of the people/nation is always both sovereign and limited, surrounded by and defined against other 'peoples' who likewise are bounded and sovereign within those boundaries (Anderson, 1983).

This emphasis on the people as a community built on affective ties of solidarity and comradeship, furthermore, requires a minimum of cultural if not ethnic homogeneity, and thus nation-states have consistently engaged in efforts at 'nationalizing the masses' through public institutions such as the schools and the military, creating a nationally minded people where there previously existed diverse and culturally heterogeneous populations, divided by all manner of differences in language, custom, and law, turning 'peasants into Frenchmen,' as the historian Eugen Weber memorably put it in the case of France (Hobsbawm and Ranger, 1983; Mosse, 1975; Weber, 1976).

In the other case, associated with the democratic definition of the people, the delimitation is made along a vertical axis where 'the people' is set against segments of the population at large that are defined as not belonging to the people. This can be done by excluding groups either at the top or the bottom. So, for example, the people can be defined as the entire population except the government, or the broad segments of the population versus the ruling elite. One expression of this democratic and 'from below' definition of the people found expression in the political movements called 'populism' that took form in, for example, the United States toward the end of the nineteenth century (Goodwyn, 1976). It has also been done by excluding groups at the bottom. Thus, to take the German case, the working class could be seen *vaterlandslose Gesellen* – as lacking a sense of 'fatherland' – whose primary identification was with the international proletariat and not with the German nation. The words *Pöbel* (from French *peuple*) and *Masse* were deployed in similar ways, to distinguish the 'real,' respectable *Volk* from the degenerate, urban underclass. Indeed, the delimitation could be made in both directions simultaneously; the Social Democrats, for example, could claim the noble title of the *Volk* defined against both the capitalist elites above and the *Lumpenproletariat* below.

In the socialist vocabulary, the concept of the people had to compete with related notions that figured prominently in the Marxist idiom, particularly 'class' and 'proletariat,' especially since it was often closely linked to the concept of the 'nation,' an idea that sat uneasily within Marxist theory. Eventually, however, 'the people' retained a largely positive status, as evidenced in the common usage of the concept of 'the people's republic' in referring to a variety of communist regimes. In the context of the liberal democratic political system of the West, the utopian impulse at the heart of the democratic definition of the people, one that envisions not only the sovereignty of the people but also that all individuals within a sovereign state

should belong to the people/*demos*, has led, at least in theory, to an increasing democratization of society and to the abolition of traditional obstacles to citizenship on the basis of categories such as class, creed, race, or gender.

National Variations: Neighboring Concepts and Political Ramifications

While the concept of the people has played a central role throughout the Western world and beyond, the impact of this key concept has varied considerably from country to country. This is true both when it comes to the specific political impact and how the 'people' came to be understood in relation to neighboring key concepts such as state, society, community, class, proletariat, and individual rights and liberties.

In the French case, the ideas of the people and the nation became strongly linked to the state and the political notion of the general will. Thus the French nineteenth-century historian Jules Michelet (21 August 1798 to 9 February 1874), in his influential book *Le Peuple* (1846), focused on the role of the Revolution and democratic politics as constitutive of the people/nation, and similarly Ernest Renan (28 February 1823 to 2 October 1892), in his essay, "What is a Nation" (1882), emphasized that the nation should not be seen as first and foremost a matter of blood, but as an expression of political will, 'a daily plebiscite.' As Renan put it, when it comes to nationality, "the people's wish is after all the only justifiable criterion, to which we must always come back." Here Renan was significantly arguing against fellow German historians over the status of Alsace-Lorraine, the provinces that lay in the borderland separating Germany from France, an area annexed by the Germans in the wake of the Franco-Prussian war of 1870.

By contrast, the German conception of the *Volk* and the nation was, pace Herder, at heart a cultural conception that at best saw the state as a protective shell safeguarding the Volk/nation, or, as in the case of Herder himself, saw the state as an artificial and even destructive form of society. At the same time, the French and the German conceptualizations of the people were similar in that they both were collectivist. *Le peuple* and *das Volk* were both formulations that suggested a collective, be it a national, ethnically/culturally defined community, or a collectivity that expressed the democratic, general will of the people.

In England the idea of the people came, as in France, to be strongly associated with political institutions and practices. However, whereas in France the idea of the people/nation tended to be associated with the collective, communal notion of the general will, which gave priority to the nation and to the social over the rights of the individual, in England it was more strongly associated with the assertion of parliamentary power, individual liberty, and the limitation of state power. Thus English national identity itself came to be linked to political as well as economic liberalism, and to a conception of the general good that took as the basic unit the freeborn individual Englishman.

These tendencies were even more fully realized in the United States, where the separation of powers and the protection of individual rights were enshrined in the constitution itself. The starting point was the stress on the American colonists' entitlement to "English liberties" and civil rights, rights that, it was claimed, had been trampled upon. The next step was to separate from this close association to Englishness and instead to refer, as was done in the Declaration of Independence, to the 'inalienable rights' ultimately grounded in universal natural law and 'the rights of man.' However, in spite of this emphasis on universality, the Constitution – and the rights and liberties it guaranteed – soon became the basis for the elaboration of an American civic nationalism, based on the idea that the American people, formed by its liberal democratic institutions, was exceptional in its commitment to individual liberty and equality. Thus in the United States, the idea of 'We the People' as the ultimate source of legitimacy and sovereignty was informed by a pervasive individualism and an enduring suspicion of centralized government. If in the French case the primacy of the people as a unified collectivity went hand in hand with strong central government, national standards, and a distaste for local difference, it was conversely true that in the US case the primacy of the people as a collection of free individuals informed an opposite tendency.

At the same time, Herderian notions of national character also came to infuse US self-understanding of the essential character of the American people, which revolved around the idea that America as a people and a nation was more mature and evolved than other peoples by virtue of its deep-seated capacity for democratic self-government. A similar fusion of an ethnic and a civic notion of the people occurred in Sweden, where leading Social Democrats after WWI came to embrace the idea that Swedes were endowed with a peculiar propensity for democracy and freedom, amounting to what one of them referred to as having 'democracy in the blood.' Citing a long, unbroken legacy of self-government that linked modern-day Swedes to the ancient yeoman peasants, the existence of a tradition was claimed, in which "democracy is rooted not merely in the constitution, but also in our traditions and in the disposition of the folk," as the Swedish prime minister Per Albin Hansson (28 October 1885 to 6 October 1946) put it in 1935.

On this basis, Swedish Social Democrats conjured up a vision of a *folkhem*, a people's home, which usefully can be contrasted with the contemporary German slogan, *Volksgemeinschaft*, or 'people's community,' championed by Hitler and the National Socialists. While both concepts celebrate national community, one prefigured the Nazi racial state, while the other became the organizing slogan for the Swedish welfare state. In the Swedish case, the concept of the *folk* connoted democratic sentiment rooted in a particular, common history; in the German one, the corresponding concept of the *Volk* referred to a racial community. This is particularly clear in the adjectival forms of *folk*/*Volk*, both equally difficult to translate into English. The German word, *völkisch*, means 'national' but with a distinct emphasis on race, and specifically with an anti-Semitic thrust. The Swedish word, *folklig*, conversely, is better translated as 'popular/populist/democratic,' with an association to the democratic legitimacy of claims made by those who credibly speak in the name of 'the people,' first and foremost the nonelite, that is, workers and peasants.

Thus the precise relationship between the concept of the people and such neighboring concepts as state, society, culture,

race, and the individual become absolutely central if we are to understand the political implications of the concept of the people. In one case, the proximity of *Volk*/nation/culture/race/collectivism is associated with racist nationalism; in another, the linkage between *peuple*/nation/state/society connotes a collectivist, if democratic, conception of popular sovereignty; in a third, the association with the people/individual liberty suggests a liberal and individual rights-based political order; while in a fourth one, the correlation of folk/home/solidarity/state informs the vision of a welfare state.

Cosmopolitanism, Human Rights, and the Contradictions of National Democracy

The tension between the idea of the people as nation (*ethnos*) and the people as the key concept in the democratic imaginary (*demos*) was, as we have seen, always present. Few, if any, actually existing nation-states have lived up to the ideal type that demanded, first, one people/nation in each state and, second, a state for each people/nation. During the nineteenth and early twentieth centuries these tensions were, however, partially masked by the emancipatory process of democratization within nation-states and the projects of national liberation in the remaining multinational empires, which overshadowed and obscured the continued and unresolved challenges when it came to reconciling nationalist and democratic claims, not least those related to the presence of ethnic minorities.

The embrace of the ideas of the people and the nation reached a high point with the outbreak of WWI and the subsequent postwar settlements. On the one hand, the 'Spirit of 1914' marked the victory of nationalist sentiments over class solidarity as German and French workers marched to war rather than joining hands under the red banner of the Second International. On the other hand, the American President Wilson's '14 points' and his related declaration of the 'right to national self-determination' at the end of WWI explicitly joined the projects of democratization and national liberation and unification under one grand vision. This resulted in a number of postwar readjustment of borders according to those principles as well as to the establishment of the League of Nations.

The idea of each people's right to self-determination would live on after World War II (WWII) during the era of national liberation movements in Africa and Asia. However, the horrors of the Holocaust and WWII produced in Europe a revulsion against ethnic nationalism that led to an embrace of trans- or postnational ideals, ranging from the establishment of the United Nations (UN) in 1945 and a number of institutions and declarations associated with the mission and ethos of the UN, such as the Declaration of Human Rights in 1948, to regional efforts to counternationalism and preempt future wars, most prominently the establishment of what over time became the European Union. In both cases, we can trace back the foundational ideas to the same historical period that spawned the concepts of the people and the birth of modern nationalism. While Herder and Rousseau can be seen as the founders of ethnic and civic nationalism, Immanuel Kant (22 April 1724 to 12 February 1804) can be viewed as the forefather of modern cosmopolitan ideas that reject the centrality of peoples and nations in favor of a vision that centers on the idea of one world and one humanity. In his essay "Toward Perpetual Peace," Kant thus proposes a *jus cosmopoliticum* (cosmopolitan law/right) with respect to "the right to the earth's surface which belongs to the human race in common," the realization of which could "finally bring the human race ever closer to a cosmopolitan constitution" (Kant, 1999: 8: 358).

In practice, the pitting of universal and cosmopolitan ideals against that of people and nation played itself out against the enduring presence of the nation-state in the context of increasing globalization. Globalization happened, first and foremost, in the economic arena, symbolized by institutions like the World Trade Organization, the World Bank, and the International Monetary Fund. But the more far-reaching Kantian ideals lived on as well, not least in the continued growth of human rights as field for both academic work and political activism, marked by, for example, the near-universal adoption of the UN Convention of the Rights of the Child in 1989 and the creation of the International Criminal Court in 2002 (Beck, 2006). However, the nation-state and the idea of the people and national identity and community did not disappear but instead entered a new phase in the early 1990s, fueled partly by populist reactions to elite cosmopolitanism and increased migration.

Thus the fluid movement between the two conceptions of the people again came to the fore during the famous Monday demonstrations in Leipzig, East Germany, during 1989–1990 as the Berlin Wall came down and the process leading to German unification unfolded. The crowds would almost seamlessly shift from a democratic chant "*Wir sind das Volk*" (we are the people) to the national slogan "*Wir sind ein Volk*" (we are one people). Already at that time, especially given recent German history, this blending of the national and ethnic with the democratic created anxiety, prompting political leaders like Helmut Kohl in Germany to call for Germans to imagine themselves as much as Europeans as Germans. But it soon became clear that the particular German inflection of this (con)fusing of the two concepts of the people was simply one example of the much broader problem of how to reconcile the ideals of national citizenship with cosmopolitan notions such as human rights. Neonationalism led to the breakup of not only the Soviet Union but also of Yugoslavia and Czechoslovakia. And by the late twentieth and early twenty-first centuries the tensions between demos and ethnos became harder and harder to ignore throughout Europe where European integration fueled more rapid migration within the European Union at the same time as globalization in the context of large gaps in wealth and stability exposed the European countries to an increasing inflow of political and economic refugees from the global south.

Since both national community and democratic citizenship historically were linked to the emerging global dominance of the nation-state, and fundamentally rested on boundaries that excluded at the very same time as they included, the increased volume and speed of migration have tended to expose what scholars refer to as the 'paradoxes' or 'contradictions' of (national) democracy. Thus Seyla Benhabib (2004) argues that national democracies are intrinsically exclusionary since they offer full rights and privileges only to members, i.e., citizens. This is so in spite of the fact that they internally speak in

a universalist language rooted in notions of equal democratic rights. However, this is a bounded universalism that distinguishes between national citizens and universal personhood (Bosniak, 2006).

This structural contradiction of national democracy comes to a head in the context of migration where it translates into differential treatment of persons on the basis of legal rules shot through with political and moral claims expressing the persistence of racialized hierarchies and notions of ethnic cultural membership (Barker, 2013). Migration, illegal as well as legal, and the attempts by the European Union and its member states to stem the flow and reassert boundaries has led both to concrete misery in the form of brutal border controls, detentions, and deportations of 'aliens,' and to a moral and political crisis as the contradictions between universal moral rules founded in notions of human rights and nationally bounded claims derived from the idea of citizenship in particular nation-states became ever more clear.

Conceptually, we are back, then, to competing moral and political ideals that had been formulated as early as the late eighteenth century. On the one hand, there is Immanuel Kant's cosmopolitan vision of perpetual peace, and on the other hand the twin conceptions of the primacy of the people in the form of national democracies in which all power emanated from the sovereign peoples in their hybrid ethnic/civic form. The Kantian conceptions have found new life after the murderous destruction of the WWII, expressed institutionally, morally, and legally in the UN, the international courts, and in declarations about human rights. At the same time, these dreams of transcending the bounded communities of the nation-states have run up against a resurgent neonationalism coming from both right-wing movements that emphasize ethnic national identity in the idioms of blood and belong, and from national democrats who embrace a welfare state nationalism that continues to speak in a universal civic language that extolls demos even as it reasserts the right of each nation-state to control its border. Here the neonationalist parties, particularly in welfare states like the Netherlands, the Nordic countries, France, and Britain, have been able to exploit a political vacuum created by the postnational turn of the established left-wing parties. These parties often acquire some of the rhetoric of national community that from the outset was part of the social democratic welfare state projects, resulting in a welfare state nationalism with more explicit ethnic overtones combined with a defense of values such as democracy, gender equality, and gay rights against the supposed threats of encroaching foreign values and practices associated with migrants, especially Muslims. Furthermore, this embrace of the key concept of 'people' and 'nation' by the neonationalist parties exploits a gap between cosmopolitan elites more focused on human rights and national-minded populations that seek refuge and protection in the name of national community. The old term 'populism,' earlier associated with popular and democratic aspirations, thus becomes a wholly negative term for these elites, who identify it with racist and xenophobic nationalism. This also exposes a fundamental challenge for neo-Kantian supporters of a cosmopolitan one-world vision: globalization so far has been dominated by a neoliberal approach that lacks the democratic and most particularly the solidaristic element of the politics of national community that produced the welfare states.

See also: Citizenship, Historical Development of; Citizenship, Sociological Aspects of; Civilization, Concept and History of; Civilizations; Community and Society: History of the Concepts; Cosmopolitanism; Democratic Theory; Ethnic Identity, Psychology of; Human Rights and Social Work; Identity Movements; Identity and Identification; Identity in Anthropology; Individuals in Society: History of the Concept; Integration and Multiculturalism: Focus on Western Europe; Nationalism, Sociology of; Nationalism: General; Nations and Nation-States in History; Personal Identity: Philosophical Aspects; Social History; Social Identity in Social Psychology; Social Identity in Sociology; Societies, Types of; State, History of; State, Sociology of the.

Bibliography

Anderson, B., 1983. Imagined Communities. Verso, New York.
Barker, V., 2013. Democracy and deportation: why membership matters most. In: Aas, K., Bosworth, M. (Eds.), The Borders of Punishment: Migration, Citizenship, and Social Exclusion. Oxford University Press, Oxford.
Beck, U., 2006. The Cosmopolitan Vision. Polity Press, Cambridge.
Benhabib, S., 2004. The Rights of Others: Aliens, Residents and Citizens. Cambridge University Press, Cambridge.
Bosniak, L., 2006. The Citizen and the Alien: Dilemmas of Contemporary Membership. Princeton University Press, Princeton.
Brubaker, R., 1992. Nationhood and Citizenship in Germany and France. Harvard University Press, Cambridge, MA.
Brunner, O., Conze, W., Koselleck, R. (Eds.), 1972–92. Geschichtlische Grundbegriffe, vols. 1–7. Klett, Stuttgart, Germany.
Burke, P., 1994. Popular Culture in Early Modern Europe. Scolar, Aldershot, UK.
Goodwyn, L., 1976. Democratic Promise: The Populist Moment in America. Oxford University Press, New York and London.
Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge, UK.
Hunt, L., 1984. Politics, Culture, and Class in the French Revolution. University of California Press, Berkeley, CA.
Joyce, P., 1994. Democratic Subjects. Cambridge University Press, Cambridge.
Kant, I., 1999. Practical Philosophy – Cambridge Edition of the Works of Immanuel Kant. Cambridge University Press, Cambridge.
Koselleck, R., 1985. Futures Past. MIT Press, Cambridge, MA.
Mosse, G., 1975. The Nationalization of the Masses. Cornell University Press, Ithaca, NY.
Weber, E., 1976. Peasants into Frenchmen. Stanford University Press, Stanford, CA.

Sociobiology and Sociology: A New Synthesis

Richard Machalek, University of Wyoming, Laramie, WY, USA
Michael W Martin, Adams State University, Alamosa, CO, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociobiology is the study of the biological basis of social behavior among animals and humans. In recent years, sociologists have begun incorporating key ideas and explanatory principles from sociobiology into sociology. This has resulted in criticism of the tabula rasa model of the human mind and its replacement by the adapted mind model. Evolutionary sociologists have proposed new explanations of the evolutionary origins of human sociality and societies, and they have applied evolutionary principles to the study of social behaviors including mating and reproduction, gender and gender inequality, social capital, and dominance and social inequality.

Although sociology was influenced by the Darwinian revolution, by the early to mid-twentieth century it had, for the most part, distanced itself from the life sciences, including evolutionary biology. In 1975, the evolutionary biologist and myrmecologist E.O. Wilson published *Sociobiology: The New Synthesis*, thereby setting the stage for an eventual reunion between sociology and evolutionary biology. By the 1990s, some sociologists had begun calling themselves evolutionary sociologists, and their work was informed by various branches of behavioral biology including, but not limited to, sociobiology, behavioral ecology, the neurosciences, population genetics, and ethology. In addition, the recent emergence of evolutionary psychology also influenced the development of the new evolutionary sociology. By 2004, the community of evolutionary sociologists in the United States had become sufficiently large to enable them to form a section of the American Sociological Association that is now called 'Evolution, Biology and Society.' This article reviews key aspects of sociobiology that contributed to the development of evolutionary sociology as it manifests itself in the early twenty-first century.

Evolution of Social Behavior by Natural Selection

Sociobiology is the study of the biological basis of social behavior among animals, including humans (Wilson, 1975). The key concepts and explanatory principles comprising sociobiological thought derive from neo-Darwinian evolutionary theory, the integration of Darwin's theory of evolution by natural selection and Mendelian genetics. Neo-Darwinian theory analyzes the traits of organisms as possible evolved adaptations. An adaptation is a phenotypic trait that evolves by natural selection and contributes to an organism's chances of survival and reproductive success. Evolved adaptations include morphological traits, physiological traits, and behavioral traits, including social behavior. Natural selection itself occurs as environmental forces such as predation, disease, climate change, and competition result in variable reproductive rates among individuals and thus, in differential gene distributions within populations over generational time. Thus, organisms are said to vary in fitness as their reproduction rates vary, thereby contributing to changes in gene (allele) frequencies over time. Evolutionary biologists distinguish between two types of fitness. Individual (or Darwinian) fitness refers to an individual organism's transmission of genes by producing offspring. Inclusive fitness refers to the combination of an organism's individual fitness and its influence on the fitness of biological kin other than its direct descendants. Simply put, there are two ways that an organism can gain genetic representation in future generations: (1) it can reproduce and (2) it can contribute to the reproductive success of close kin, a process called kin selection in evolutionary biology.

In addition to natural selection, sexual selection is an important evolutionary process, especially in terms of its influence on social behavior. Sexual selection occurs when evolved traits influence the ability of an organism to recruit a mate and produce offspring. For example, brightly colored plumage among some species of birds serves as an attractant to females, thereby enabling males that display the brightest coloration to reproduce more successfully than males with duller color patterns. Brightly colored plumage can be an indicator of a robust immune system, a trait that is likely to be favored (unconsciously) by females that are inspecting males for potential mates. Similarly, males may be adapted to express preferences for cues that signal potential high fertility among females, thereby courting such females more vigorously than other females.

An important topic in sociobiological inquiry is the problem of altruism. In evolutionary theory, altruism refers to any trait, the possession of which reduces the fitness of the individual bearing it while, simultaneously, increasing the fitness of another individual. As W.D. Hamilton (1964) explained, however, altruism can be more apparent than real. For example, if the sacrificial behavior of one individual reduces its fitness to zero (by sacrificing its life) and thereby saves the life of another, unrelated individual, an evolutionary biologist would identify such behavior as altruistic. However, if the recipient of the apparent altruist's beneficence was a biological relative, then the behavior would be viewed as phenotypically altruistic but genetically selfish. That is, by losing its life, the individual performing the sacrifice saved the life of a relative with whom it shared genes (alleles), thereby contributing to the survival of copies (replicas) of its own genes

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32010-4

present in the body of its relative. Thus, altruism can evolve by kin selection.

Natural selection can also produce another variant of altruism that is more apparent than real, and it is called reciprocal altruism (Trivers, 1971). If an individual behaves in a manner that reduces its own fitness and increases the fitness of a nonrelative, it is behaving altruistically. However, if in the future the recipient of the altruistic behavior reciprocates and behaves altruistically toward its benefactor, then the original act of altruism is better viewed as the first stage in a process of delayed reciprocity, which evolutionary biologists call reciprocal altruism. A recent review of research on reciprocal altruism revealed that, with a few exceptions, it is rare among nonhuman species, and that it appears that humans are without rival as the premier reciprocal altruists (Clutton-Brock, 2009).

In addition to altruism (and reciprocal altruism), sociobiologists investigate other forms of social interaction, including mutualism, commensalism, and social parasitism. In mutualism, both parties in a relationship receive benefits from the interaction. In commensalism, one party benefits, but the other neither receives a benefit nor incurs a cost. In social parasitism, one party extracts a benefit at the expense of the other party. These concepts have value for evolutionary sociological analysis. For example, Cohen and Machalek (1988) have adapted models of social parasitism that were developed for studying nonhuman social species in order to analyze patterns of expropriative crime among humans.

Until very recently, Hamilton's notions of kin selection and inclusive fitness constituted a largely unchallenged explanation of most patterns of cooperative social behavior among animals. The discovery that genetic relatedness is the bedrock on which many forms of cooperative group life are based was elucidated in Hamilton's analysis of the role of genetics in supporting highly cooperative societies of eusocial insects (ants, bees, and wasps, but not termites). Because ants, bees, and wasps feature a distinct form of genetics called haplodiploidy, the females are hyperrelated to each other, and it is to this hyperrelatedness that Hamilton and later others attributed the highly cooperative and cohesive social systems comprising colonies of ants, wasps, and bees. Specifically, males in these taxa develop from unfertilized eggs and inherit only one set of chromosomes (from their mother), while females inherit one compliment of chromosomes from their mother and a second compliment from their father. Consequently, full sisters are more closely related to each other ($r = 0.75$) than each is to her mother ($r = 0.5$). Hyperrelatedness among sisters means that they share, on average, 75% of their genes with each other, thereby creating a situation whereby there is a 25% greater coincidence of their genetic interests than there is among full siblings in diploid species, which share, on average, only 50% of their genes. In Hamilton's view (and according to what has essentially become orthodoxy in conventional sociobiology), the forces of kin selection are particularly strong among haplodiploid species, thereby explaining the great degree and complexity of social life among the eusocial insects.

Recently, however, E.O. Wilson and D.S. Wilson have challenged the sociobiological orthodoxy that kin selection and inclusive fitness explain eusociality among ants, bees, and wasps. Furthermore, they contend that kin selection cannot explain the highly evolved form of eusociality evident among termites, because termites are diploid. Thus, instead of attributing the basis of eusociality to the unusual genetics of haplodiploidy, Wilson and Wilson turn to an ecological explanation in which they argue that patterns of social life among ants, bees, wasps, and even termites more likely evolved as a response to a circumstance whereby members of a group occupy a defensible residence (such as a nest), live in reasonably close proximity to a food source, and provision their young (Wilson, 2008; Wilson and Wilson, 2007). They contend that the principles of kin selection and inclusive fitness have been overburdened in efforts to explain social behavior in general and eusociality in particular. Instead, they favor an ecological explanation and urge others to consider this approach as well. This approach entails the reintroduction of a version of group-selection reasoning into sociobiology, and a consensus among evolutionary biologists and sociobiologists has not yet developed on this issue.

The Adapted Mind

Supported by the work of anthropologists like Franz Boas and his students Margaret Mead and Ruth Benedict, the notion that the human brain and mind is a general, all-purpose information-processing machine came to dominate sociological thinking by the mid-twentieth century. In terms of information that might prescribe social behavior, the mind was viewed as content-free, until subjected to individual experience or exposed to the influence of culture. This point-of-view eventually became codified into what evolutionary social scientists now describe as the Standard Social Science Model (SSSM) of human nature and behavior. The key elements of the SSSM include the following: (1) Human social behavior is learned, not instinctive. (2) Only by endowing humans with the capacity for culture does the human brain influence social behavior. (3) The kaleidoscopic variability among human societies and cultures constitutes prima facie evidence that the common human genome does not determine human social behavior. (4) Culture, not biology, determines patterns of social behavior and society. (5) The complexity of human social behavior and societies is irreducible to biological factors, such as genes.

The SSSM dominated social scientific thought through most of the twentieth century by promoting oversimplified dualisms such as learning versus instinct, culture versus biology, and social versus psychological determinants of human behavior. The durability of the SSSM in sociological thinking was possible only because many social scientists remained ignorant of a number of key principles regarding biology and social behavior. First, the influence of genes on behavior is not direct. Rather, genes shape behavior indirectly, by guiding the ontogeny of the brain and central nervous system as well as the endocrine system and specifying their functional properties. Second, phenotypes, including social behavior, are not the product of genes alone. Rather, during the process of epigenesis, the environments in which genes reside contribute to the expression of the phenotype that is encoded in the genome. This means that the same genotype can generate a range of phenotypes under variable environmental conditions. The

extent of the range of phenotypes that can so develop is called the norm of reaction. Thus, phenotypic traits, including social behavior, are coproduced by genes and their environments. Third, the SSSM created a false opposition by contributing to the view that behavior must be either learned or instinctive when, in fact, learning itself is understood by evolutionary scientists to be instinctive (Pinker, 1997). The difference between the SSSM view of learning and that of sociobiologists is as follows: the SSSM construes the human brain as a general, all-purpose learning machine, which is equally prepared to learn one thing as well as another. Evolutionary-minded social and behavioral scientists, however, introduce the notion of prepared (or biased) learning, which means that natural selection has installed predispositions to learn and retain some types of information and information contents more readily than others. Specifically, the more adaptively relevant the information, the more likely it can be learned and retained. An example provided by psychologists is the Garcia effect (or Sauce-Béarnaise effect), which refers to ease and speed with which taste aversions can develop in response to potentially toxic foods. The notion of prepared learning has been applied to human social behaviors to suggest that the ability to detect cheating (nonreciprocity) in exchanges relations, identify prospective mates, and detect various forms of threat, for example, may be the product of prepared learning algorithms that evolved by natural selection.

The Phylogeny of Human Sociality

By reconsidering the nature and history of human sociality and societies in light of evolutionary theory, sociobiologists and evolutionary sociologists have begun to call into question long-standing orthodoxy about topics such as the fundamental building-blocks of human society. Specifically, work by Jonathan Turner and Alexandra Maryanski (2008) casts doubt on the assumption that the nuclear family is the primordial social unit from which larger units, including societies, have evolved. By conducting cladistics analyses of primates, especially chimpanzees, Maryanski concluded that humans are not a species that is predisposed to forming strong, durable social ties based on emotional attachments (1992). Instead, she (and later, Turner and Maryanski, 2008) contends that the earliest hominins lived in groups comprising weak social ties, and that social relations were highly ephemeral. This challenges the long-standing assumption that the basic building block of human society was the nuclear family, which was organized on the basis of a very strong pair bond between a male and a female who produced offspring. In fact, the likelihood that the earliest hominins formed social relationships comprising weak ties calls into question the notion commonly held by sociologists that primary groups have always been foundational to the evolution of human societies.

Recent work by Turner and Maryanski revives in new form the old idea that the primordial social units to which humans belong more close resembled a horde, within which social relationships formed and dissolved with considerable ease. By this way of reasoning, it was not until early hominins underwent a transition from an arboreal to a savanna lifestyle that natural selection favored a redesign of the brain that enhanced its emotional capabilities, thereby enabling the formation of strong social ties. The late Miocene Epoch featured a period of climate cooling, which contributed to the decline of forest habitats. As hominin populations were forced to move from forests into open plains and grasslands, they are likely to have confronted new hazards including greater concentrations of predators. The evolution of highly cohesive groups would have been adaptive in such high-risk environments. In response, selection would have favored the enhancement of hominin emotional capacities that could enable the formation and maintenance of strong social ties. While it is well-known that the cortical regions of the human brain are very large and evolved quickly, what is less commonly known is that subcortical regions of the brain are also much larger proportionately in humans than they are among other apes. The subcortical parts of the brain comprise structures such as the diencephalon, the amygdala, the septum, the hippocampus, and the transition cortices. Taken together, they constitute what is sometimes called the limbic systems, and they play a major role in producing and regulating human emotional life. The evolution of the capacity for strong, positive emotions like satisfaction-happiness would support the formation of strong social ties among individuals, and the evolution of negative emotions including shame and guilt would support systems of social control, thereby protecting cooperation from being disrupted by excessive selfishness. Thus, the evolution of an extensive array of emotional capabilities could enable early hominins to be transformed from a weak tie social species whose social relations were easily formed but also easily dissolved, into a strong tie social species with the capacity for forming and maintaining the type of cohesive and durable groups that sociologists call primary groups, including the nuclear family.

In conventional sociobiological thinking, societies are assembled primarily on the basis of common kinship according to the principles of kin selection and inclusive fitness. The population size of kin-based societies can be enlarged by other mechanisms, especially reciprocity among nonkin. In sociobiological theory, common or overlapping genetic interests among individuals provide an incentive for cooperative social interaction, because the fitness benefits of one individual's gains also serve the fitness interests of the other. Among biologically unrelated individuals, cooperation is favored by selection not because individuals share common genetic interests, but rather, because it yields somatic resources (food, protection) from which cooperators can benefit even if they are nonkin. In fact, behavioral biologists have shown that patterns of cooperation can develop even among members of different species, such as the widely discussed relationship between cleaner fish and host fish. Host fish suffer parasite loads that can compromise their health. However, by traveling to cleaning sites in areas such as coral reefs, they can make themselves available to smaller fish that remove the parasites from the bodies of the hosts, thereby gaining food. Though the host fish, a predator itself, could eat the cleaner, it refrains, thereby deriving the benefit of parasite removal by the cleaner. Both participants benefit from this form of cooperation, not because they are kin, but because their behaviors yield mutually rewarding consequences. The host–cleaner fish relationship provides an example of how cooperation can evolve in the absence of close biological kinship.

Trivers's 1971 seminal article on reciprocal altruism converged on a long-standing tradition in sociological thinking that emphasized the importance of reciprocity, exchange, and sharing as the basis of cooperation among groups and societies. Subsequent research on the possibility of reciprocal altruism among both humans and nonhumans elaborated the importance of understanding how social life could evolve on the basis of mechanisms other than kin selection. Much of that research was framed in terms of game theory, and within that tradition, researchers made extensive use of the 'prisoner's dilemma' model of cooperation. The problem of defection (cheating) in the prisoner's dilemma model becomes especially salient from an evolutionary perspective when the parties to cooperation are nonkin. If one individual is cheated by another to whom he/she is closely related, the victim still derives some benefit from the outcome because he/she shares genes with the defector. However, if the defector is a nonrelative, then the victim incurs a net loss in fitness. The importance of outcomes in social interactions comprising reciprocity and exchange has led animal behaviorists to conduct research about the possibility that nonhuman animals might possess capabilities resembling what could be called a sense of fairness. These researchers have adduced evidence among both chimpanzees and New World monkeys that indicates these animals appear to monitor the allocation of resources among members of their groups. This suggests that even some nonhuman primates may possess behavioral mechanisms built into their neuroanatomy for detecting and reacting to cheating.

The highly enhanced cognitive abilities of humans and the importance of reciprocity and exchange in human relationships implies that the ability to detect and react defensively to threats of cheating (nonreciprocity) is even more important in human society than in the social systems of nonhuman species. Experimental research by evolutionary psychologists supports the hypothesis that humans are likely to possess a cheating-detection mechanism that protects them from being victimized in social contracts that involve exchange relations (Cosmides and Tooby, 1992). Subsequent research about the role of emotions in exchange relations has led to the conclusion that emotions such as anger, moral outrage, and a desire for revenge function as a defense against cheating, thereby helping to inhibit the spread of defection within populations of cooperators.

Sociobiology and the derivative discipline of evolutionary psychology hypothesize that the species-typical features of the human brain (and mind) evolved during a time period that began approximately 2.5 million years ago with the appearance of the genus *Homo* and lasted through the hunting-gathering era, the end of which occurred about 10 000 years ago. Current estimates locate the appearance of modern humans about 100 000 to 200 000 years ago. Evolutionary psychologists contend that what they call the cognitive architecture of the human brain is the product of the environment of evolutionary adaptedness (EEA), a configuration of ecological, demographic, and technological environmental factors comprising the selection forces that designed the modern human brain. By this reasoning, the modern human brain and mind represent evolved adaptations to a hunting-gathering social world that is occupied by only a tiny fraction of contemporary humans. The vast majority of living humans reside in and must cope with social worlds that have succeeded the hunting-gathering era, and there is no reason to expect that the human brain is equipped with any specialized, evolved cognitive algorithms for coping with contemporary social environments. This has led scientists to posit a mismatch hypothesis that contends that the evolved properties of the human mind are adapted to a world that has largely vanished, and thus, the adaptability of the human mind may now be more important than the evolved adaptations with which the mind has been equipped by natural selection. The evolutionary sociologist Massey describes this situation by characterizing contemporary humans as strangers living in a strange land (2005).

The transition from hunting-gathering approximately 10 000 years ago to horticultural, agricultural, industrial, and postindustrial eras has provided evolutionary-minded sociologists with opportunities to reflect about familiar topics from a new point-of-view. Specifically, informed by sociobiology, evolutionary psychology, and other evolutionary disciplines, sociologists have posed new questions and posited new explanations of life in contemporary societies. For example, Massey has investigated the possibility that contemporary urban life presents daunting psychological as well as social challenges to its inhabitants (2005). Among such challenges are what can be called diseases of civilization, including mental health problems produced by prolonged exposure to high levels of stress in urban environments. Other health challenges that can be induced by chronic stress include hypertension, obesity, Type II diabetes, Type A personality disorders, immune system deficiencies, memory deficits, and various neurological disorders. Such health problems can intensify social stresses within families, work settings, and other social contexts, leading to social conflict and even violence.

Not all circumstances encountered by a Pleistocene mind living in a space-age world result in dire consequences, however. For example, Massey points to the resiliency with which humans can cope with the scale, anonymity, and impersonality of urban life by organizing themselves into small-scale, intimate social networks that closely approximate the hunter-gatherer band comprising 40–50 or so individuals. Such social networks, sometimes labeled urban villages, have been recognized by urban sociologists for decades and can be understood as manifestations of homophily, a preference to associate with individuals who are similar to oneself, thereby buffering the impersonal effects of large-scale urban environments. Yet, homophilous tendencies that yield closed social networks may also be implicated in the development of closed social networks that constitute social classes and segregated ethnic or racial enclaves, the effects of which can contribute to divisiveness within the larger communities in which they exist. The fate of a small group ape living in a post-EEA social world continues to stimulate research and inquiry by evolutionary sociologists.

Current Inquiry in Evolutionary Sociology

Inspired by sociobiology and related evolutionary sciences, the range of inquiry in which evolutionary sociologists have become engaged is rapidly expanding. The following provides an introduction to the sorts of research endeavors in which evolutionary sociologists are currently engaged. Though not

exhaustive, these areas of inquiry include mating and reproduction, gender and gender inequality, the role of social capital in patterns of social interaction, and dominance relations and patterns of social inequality.

Mating and Reproduction

In diploid species such as humans, males and females derive equal fitness returns from any offspring they produce. However, parental investment is asymmetrical, with females typically incurring greater costs in producing and nurturing offspring. The foundation of this inequality is anisogamy, the difference in size, number, and biological value of gametes. Sperm are small, numerous, and metabolically inexpensive, while eggs are large, few, and metabolically expensive. Furthermore, once fertilized, an egg imposes significant developmental costs on the mother. In a highly altricial (dependent) species like humans, the care and nurturing of offspring imposes much greater parental investment requirements on mothers than on fathers. The most significant form of such investment during the infant's early life is nursing, a cost incurred by females, because males do not lactate (Huber, 2007). A second type of reproductive asymmetry, paternity uncertainty, poses a threat to males but not to females. Barring recent medical developments like the transplantation of fertilized eggs, females are exempt from the experience of maternal uncertainty. However, paternity uncertainty is species-typical for humans, and it has resulted in a wide range of what sociobiologists call mate-guarding strategies developed and deployed by males. Sexual infidelity poses a threat to both males and females, but the nature of the threat and the sexual jealousy that it can activate differs between males and females. Female sexual infidelity threatens males because of paternity uncertainty, while male sexual infidelity threatens females with a potential loss of resources, including paternal investment in offspring. Thus, despite the fundamentally cooperative nature of mating, biological dimensions of reproductive relations also create a foundation for conflicts of interests that are built into human mating systems.

The division of reproductive labor has prompted evolutionary sociologists to look for male–female differences in mate preferences. Rosemary Hopcroft has found that females prefer commitment, social status, and resources in mates, while males cue on indicators of fertility, such as youth and beauty, and to a lesser extent than females, resources as well (2009). Furthermore, there is a stronger correlation between social status and fertility for men than for women. In an empirical test, Hopcroft reported results that were consistent with the Trivers–Willard hypothesis, which states that parental status influences differential parental investment in sons and daughters. Specifically, she found that boys in high-status families attain more years of formal education than do girls, but in low-status families, girls receive more support and attain more years of education than do boys. Sociobiologists explain such patterns in terms of differences between male and female lifetime potential fitness, with male potential fitness far exceeding that of females.

Gender and Gender Inequality

As discussed previously, an evolutionary sociologist has proposed that female lactation in humans imposes opportunity costs on mothers that are not incurred by males (Huber, 2007). In traditional societies, the birth of each child requires intensive maternal investment until the child is at least 4 years old. If the female bears several children in succession, this can deprive her of participation in areas of public life such as economics, politics, and military service, all of which provide males with status-attainment opportunities. Huber attributes the evolution of patriarchy among humans to this aspect of the reproductive division of labor in humans. Not until infant formula was developed was this aspect of maternal investment relaxed, making possible the development of more egalitarian relations in parenting.

Even before the emergence of sociological inquiry that is informed by sociobiological theory and research, social scientists had documented the influence of sex ratios on patterns of gender inequality (Guttentag and Secord, 1983). Specifically, high sex ratios create a scarcity of potential mates for males, thereby maximizing female dyadic power in choosing mates. Males commonly react to this power disadvantage by trying to limit female discretion in mate-selection interactions. In evolutionary terms, males try to restrict female choice and minimize male–male competition, both of which are important components of sexual selection. Restated in sociobiological terms, the Guttentag–Secord theory of sex ratios predicts that high sex ratios will activate male sexual jealously and maximize efforts by males at mate-guarding. Thus, demographic properties of a society can activate evolved psychological propensities for behaviors that represent efforts to respond adaptively to conflicts of interest.

Hopcroft attributes the roots of gender inequality and sex discrimination to evolved differences between male and female psychology (2009a,b, 2005, 2002). Like Huber, she uses evolutionary thinking to try to identify the phylogenetic origins of gender inequality in general, and patriarchy in particular. Unlike Huber, she explores the possibility of the existence of male–female differences in evolved psychology that could be implicated in the origins, development, and maintenance of patterns of gender inequality. Specifically, Hopcroft documents a pattern whereby adolescent girls tend to defer to males, but males of similar age do not defer to females. Conversely, adolescent females do not defer to other adolescent females. Hopcroft explains that young men should find deferential females more attractive as potential mates, because they pose less of a threat of paternity uncertainty and less of a challenge to mate-guarding efforts. Hopcroft notes that females from the ages of 15–18 typically experience lower self-esteem than do males of the same age, despite being more academically successful and less-likely to manifest behavioral problems. She interprets this anomaly using sexual selection theory, and speculates that the pattern of lower self-esteem among females in this age range may be an expression of evolved female deference to males. In her view, deferential female behavior to a male who is a prospective mate signals the promise of fidelity, thereby relieving males of the anxiety of paternity uncertainty in a future reproductive partner.

Hopcroft considers the alternative possibility that deference in young females toward males may be the product of culture and socialization alone. If so, she predicts that female deference to males should persist through all stages of the life

cycle, since the effects of patriarchal culture are not relaxed as women age. However, Hopcroft reports that female deference to males declines as women enter postchildbirth years, and the status of women in all societies rises as women pass menopause.

Social Capital

Social capital theory contends that social relationships are resources that can lead to the development and accumulation of human capital. For example, a stable family environment can support educational attainment and support the development of highly valued and rewarded skills and credentials. In evolutionary terms, social capital can be defined as any feature of a social relationship that yields reproductive benefits. According to Savage and Kanazawa (2002, 2004), humans have evolved preferences for companionship in general, and specific preferences for cues that signal higher levels of social capital. Since evolved preferences for certain types of social relationships should have been selected in the EEA, we can expect to observe gender differences that reflect the division of labor in foraging societies. For example, females can be expected to value and derive emotional satisfaction from membership in small social networks comprising close personal relationships constructed from strong social ties. These types of relationships would be especially adaptive for women by providing assistance in foraging and caring for children. Males would be expected to benefit more from membership in larger social networks constructed from weak ties that would constitute hunting groups, political alliances, and fighting parties. Males would benefit especially from forms of social capital that would confer resources and social status.

Savage and Kanazawa (2002) contend that reformulating sociological explanations of crime in terms of evolutionary theory can provide new insights about deterrents to crime. Specifically, choice and social control theories, cultural deviance theories, and strain theories all suggest that the threat of losing social bonds can be a powerful inhibitor of deviant or criminal behavior. Evolutionary theory invites predictions that social ties and relationships that were most closely related to reproductive success in the EEA are especially likely to produce a deterrent effect against deviance and crime. For women, behavior that would jeopardize close personal relationships would be particularly costly, thus, the threat of losing those relationships would deter such behavior. For men, social capital that yields material resources and social status is especially valuable, and the threat of losing those resources would be likely to inhibit deviant or criminal behavior. This type of reasoning is consistent with empirical findings that document male sensitivity to status loss or degradation.

Dominance and Social Inequality

Both sociologists and evolutionary behavioral scientists, including sociobiologists, study dominance relations and social inequality. Evolutionary sociologists look for the phylogenetic origins of social inequality in the evolved psychology of humans as well as the ecological contexts in which humans evolved and now live. Dominance hierarchies are established in face-to-face interactions among individuals and determine access to resources, including both reproductive resources (mates) and somatic resources (food, territory, and other material resources). In addition to dominance hierarchies, both official hierarchies (position rankings in formal organizations) and socioeconomic hierarchies (e.g., social class systems) exist in human societies. Evolutionary thinking has had more influence on the study of human dominance hierarchies than on official or socioeconomic hierarchies, though recently it has begun to be integrated into the study of phenomena such as class relations in urban areas (Massey, 2005).

Work by evolutionary sociologists like Mazur (2006, 2005) and Booth et al. (2006) has established a link between testosterone and rank in dominance hierarchies, but not aggression. In fact, Booth et al. (2006) have determined that the effect of testosterone on behavior is mediated by social relationships like parent–child relations, peer relations, and relations to siblings. Mazur (2006) has also determined that testosterone levels in males both influence and are influenced by dominance contests and their outcomes. Mazur also documented a relationship between testosterone levels and honor cultures which, in turn, influences levels of violence among young men.

Among nonhuman species, status (social rank) is positively correlated with reproductive success, but among humans, the correlation is equivocal (Hopcroft, 2006). In fact, some studies have found a negative relationship between social status and reproductive success. However, evidence suggests that the relationship differs between men and women, and that for men, higher social status is associated with fathering more children. Another area of research pertaining to evolution and social inequality is illustrated by the Trivers–Willard hypothesis, the prediction that wealthier families will invest more resources in support of sons, but poorer families invest more in support of daughters. While Freese and Powell (1999) failed to adduce evidence in support of Trivers–Willard, Kanazawa retested the hypothesis using sex-neutral measures and reported findings that indicate weak support for the hypothesis.

Finally, in reflecting on the implications of humans exhibiting an evolved psychology that is incongruous with the social and demographic characteristics of contemporary, urban-industrial societies, Massey (2005) discusses the possibility that an evolved propensity for homophily may help explain the development of social networks that constitute both ethnic enclaves and urban villages in large metropolitan areas. Evolutionary processes such as kin selection and reciprocal altruism may underpin the development of such networks. And though such networks can provide support for their members, they also can reinforce stratification systems by isolating groups of people from each other, creating both networks of privilege and networks of poverty. Massey further explains that because processes like kin selection and reciprocal altruism promote mutual assistance among, but not between, members of these networks, networks of poverty are often depleted of the resources they share among themselves, while cooperation among members of networks of privilege increase the resources that are available to members of those networks. Thus, an evolved psychology

designed for cooperation and sharing among members of societies in the EEA can, ironically, impose significant burdens on members of social networks comprising lower class positions in society.

Bibliography

Booth, A., Granger, D., Mazur, A., Kivlighan, K.T., 2006. Testosterone and social behavior. Social Forces 85, 167–191.

Clutton-Brock, T., 2009. Cooperation between non-kin in animal societies. Nature 462, 51–57.

Cohen, L.E., Machalek, R., 1988. A general theory of expropriative crime: an evolutionary ecological approach. American Journal of Sociology 94, 465–501.

Cosmides, L., Tooby, J., 1992. Cognitive adaptations for social exchange. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, pp. 163–228.

Freese, J., Powell, B., 1999. Sociobiology, status, and parental investment in sons and daughters: testing the Trivers-Willard hypothesis. American Journal of Sociology 106, 1704–1743.

Guttentag, M., Secord, P., 1983. Too Many Women? The Sex Ratio Question. Sage, Newbury Park, CA.

Hamilton, W.D., 1964. The genetical evolution of social behavior, I and II. Journal of Theoretical Biology 7, 1–52.

Hopcroft, R.L., 2002. The evolution of sex discrimination. Psychology, Evolution and Gender 4, 43–67.

Hopcroft, R.L., 2005. Parental status and differential investment in sons and daughters: Trivers-Willard revisited. Social Forces 83, 1111–1136.

Hopcroft, R.L., 2006. Status characteristics among older individuals: the diminished significance of gender. The Sociological Quarterly 47, 361–374.

Hopcroft, R.L., 2009a. Gender inequality in interaction: an evolutionary account. Social Forces 87, 1845–1872.

Hopcroft, R.L., 2009b. The evolved actor in sociology. Sociological Theory 27, 390–406.

Huber, J., 2007. On the Origins of Gender Inequality. Paradigm, Boulder, CO.

Kanazawa, K., 2001. Comment: why we love our children. American Journal of Sociology 106, 1761–1788.

Maryanski, A., 1992. The last ancestor: an ecological-network model on the origins of human sociality. Advances in Human Ecology 2, 1–32.

Massey, D.S., 2005. Strangers in a Strange Land: Humans in an Urbanizing World. W.W. Norton, New York.

Mazur, A., 2005. Biosociology of Dominance and Deference. Rowman and Littlefield, New York.

Mazur, A., 2006. The role of testosterone in male dominance contests that turn violent. Social Biology 53, 24–29.

Pinker, S., 1997. How the Mind Works. W.W. Norton, New York.

Savage, J., Kanazawa, S., 2002. Social capital, crime, and human nature. Journal of Contemporary Criminal Justice 18, 188–211.

Savage, J., Kanazawa, S., 2004. Social capital and the human psyche: why is social life "capital?". Sociological Theory 22, 504–524.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–37.

Turner, J.H., Maryanski, A., 2008. On the Origins of Societies by Natural Selection. Paradigm, Boulder, CO.

Wilson, E.O., 1975. Sociobiology: The New Synthesis. Belknap, Cambridge, MA.

Wilson, E.O., 2008. One giant leap: how insects achieved altruism and colonial life. Bioscience 58, 17–25.

Wilson, D.S., Wilson, E.O., 2007. Rethinking the theoretical foundation of sociobiology. Quarterly Review of Biology 82, 327–348.

Sociobiology, History of

Ullica Segerstrale, Illinois Institute of Technology, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The last third of the twentieth century saw a change in evolutionary reasoning from group selection to kin selection and from a focus on the individual to a focus on its genes instead. The theoretical breakthrough came with William D. Hamilton's 'inclusive fitness.' Although Edward O. Wilson's *Sociobiology: the New Synthesis* and Richard Dawkins' *The Selfish Gene* provided important scaffolding, the emergence of the actual research paradigm can be described as a collective process of mutual influences and distributed leadership. As research in 'sociobiology' under different names continued to expand and flourish, the first decade of the twenty-first century saw a challenge to the reigning paradigm, stimulating the further proliferation of research topics as well as integration of the field.

The Rise of the Sociobiological Paradigm

What's in a Name?

'Sociobiology' emerged as a new integrative discipline during the second part of the twentieth century. It transformed the study of animal social behavior into a field based on Neo-Darwinian reasoning, which interprets evolution as a change in gene frequencies in populations. This made it possible to develop mathematical models that could be subjected to empirical testing, turning traditional natural history into a more rigorous science.

The name 'sociobiology,' however, did not become the name in use for practicing biologists. Many scientists involved in what would have been counted as bona fide sociobiological research actively avoided the term, instead calling themselves as behavioral ecologists or functional ethologists. This was largely because of the political connotations of sociobiology caused by the debate around Harvard zoologist Edward O. Wilson's (1975) book *Sociobiology: The New Synthesis*. Those involved in what might have been called 'human sociobiology' were particularly sensitive, preferring to be called biosocial anthropologists, human ethologists, human behavioral ecologists, or Darwinian anthropologists. From the 1990s onward, much sociobiological research went under the new label 'evolutionary psychology.'

As it turned out, Wilson's particular, broadly conceived sociobiological research program identified in his book did not become the approach that was in fact used by practicing 'sociobiologists' (although they did not call themselves that). Instead, the actual sociobiological research paradigm-in-use consisted of a quite small but connected collection of theoretical insights introduced by Richard Dawkins' *The Selfish Gene*.

A Paradigm Shift

By the early 1970s a significant amount of both new theory and empirical studies of animal social behavior using an evolutionary perspective had accumulated. Social behavior was now regarded as undergoing evolution – just as morphological traits did – and treated in a quantitative way that allowed for predictions to be made. The excitement was so palpable that it caused some scientists to feel that they had experienced a personal 'paradigm shift.' For others again, the extension of an evolutionary perspective to human social behavior was both politically dangerous and scientifically unsound; it was 'bad science' that needed to be sharply criticized.

The core idea that revolutionized thinking was William Hamilton's concept of inclusive fitness, or alternatively, 'kin selection.' This central insight was followed by Robert Trivers' important papers on reciprocal altruism, parental investment, and parent – offspring conflict, and John Maynard Smith's concept of Evolutionarily Stable Strategy. The person who most strongly advocated for a change away from group selection was George Williams. His famous 'canon' effectively ruled out vague references to 'the good of the species' that could be found in textbooks of the time. Williams argued firmly that when formulating adaptive explanations, one should not involve higher levels of organization than absolutely necessary. Unless there existed compelling evidence to the contrary, one should stick with the simplest form of natural selection: that of alternative alleles in Mendelian populations.

What united the new theoretical approaches was a game-theoretical perspective, whereby genes could be imagined as strategic agents causing their carriers to behave in their (the genes') own interest, that is, to get propagated. Hamilton sometimes used this imagery in his papers, but it was Dawkins who took it to new heights in his popular book.

A Collective Effort

Many have noted a surge of books and articles published in the mid-1970s, summarizing and discussing the new ideas. Among these various 'syntheses' were textbooks by Jerram Brown and John Alcock, and Robin Fox' *Biosocial Anthropology* (based on a 1973 Oxford conference covering issues in human sociobiology) all published in 1975. Among seminal articles were Richard Alexander's long 1974 contribution 'The Evolution of Social Behavior' in *Annual Review of Ecology and Systematics*, and shorter articles by Alexander (1975), West-Eberhard (1975), and Maynard Smith (1976). Together, this makes for quite a concentration of intellectual, synthetic effort in the mid-1970s. What got particularly emphasized, however, were Wilson's *Sociobiology* and Dawkins' *The Selfish Gene*.

These two books became central for the early scaffolding of the emerging sociobiological paradigm. They were written independently of each other and were in fact quite different. Still, in a strange complementary way, Wilson and Dawkins together were able to claim the new field and identify its intellectual content. Wilson surveyed the territory, gave it a name, and handed it over to his animal behavior colleagues. Dawkins explicated the underlying logic of the field and identified the core theories.

At the same time there was an unusual amount of personal interaction and mutual reinforcement taking place, also across the Atlantic. Hamilton was influenced by the American, George Price, who had come to live in England. Price also collaborated with Maynard Smith. Young Hamilton was early on encouraged particularly by George Williams, finding in him a kindred soul. E.O. Wilson was an early supporter of Hamilton and a later supporter of Trivers, a doctoral student at Harvard, who worked closely with anthropologist Irven De Vore, who had been deeply inspired by Hamilton. Trivers' early papers, which extended a Hamiltonian type of reasoning, were an important inspiration for Dawkins (1989, p. 298). Dawkins asked Trivers to write the Preface to his book; Hamilton encouraged Trivers. Dawkins' and Wilson's books served as vehicles for spreading Hamilton's and Trivers' ideas, and Dawkins' book for the ideas of Williams, Maynard Smith, and so on.

There also was an important irritant – a book that many regarded as 'wrong': Vero Wynne-Edwards' (1962) group selectionist explanation that the mass gatherings and displays of grouse were means for population regulation and altruism for the good of the species. Williams and Maynard Smith were among those who felt challenged to prove him wrong.

One particular conference may well have been a crucial factor in the development of sociobiology, the lavish "Man and Beast" conference arranged at the Smithsonian Institute in Washington, DC, in May 1969. Entitled "Comparative Social Behavior," it gathered an abundance of eminent academics from the United States and elsewhere – largely biologists, including anthropologists and philosophers interested in evolutionary questions. Konrad Lorenz' (1966) book *On Aggression*, and popular ethological books championing a fundamentally brutish human nature had made biologists seem like relevant consultants on the problem of social violence, examples of which had been recently experienced in the US.

Among the invited were Edward Wilson, Richard Alexander, and young William Hamilton, as well as biosocial anthropologists Lionel Tiger and Robin Fox, Napoleon Chagnon, and William Irons. For Wilson, the conference was 'an early milestone in the development of sociobiology.' The conference and the resulting edited volume reinforced his belief that a common language was possible. 'The ingredients for Wilson's *Sociobiology* were all there!" as one of the organizers later noted.

Wilson's book had an ambivalent approach to ethology. At the beginning of his book he provocatively declared that ethology was going to be swallowed up by sociobiology by year 2000. Others commented that much of *Sociobiology* itself could as well have been found in a traditional ethology book (e.g., Crook, 1976). *Sociobiology* may in fact have stolen the show from similar summarizing efforts in ethology at the time, for example, Bateson and Hinde's *Growing Points in Ethology* (1976).

Such views illustrate the confusion about the actual meaning of 'sociobiology' at the time. Was it the name to be identified with Wilson's broad definition, the study of social behavior of all animals, or the more specific quantitative approach he had also recommended in his book – involving various population parameters and drivers of evolution – or with a set of particular new ideas of a handful of individuals?

From Group Selection to Kin Selection

Inclusive Fitness and Kin Selection – a Closer Look

If the shift from individual selection to group selection after World War II had been a 'stampede' (Hamilton's expression), then between the mid-1960s and mid-1970s we had yet another 'stampede,' this time from group selection to kin selection. According to some observers, if around 1965 you had to be a group selectionist to get published, 10 years later you had difficulty getting published if you were not a kin selectionist. As mentioned, some biologists experienced this paradigm shift as an intense personal revelation. In some departments, tensions arose between group selectionists and newly "converted" kin selectionists. Personal relationships were affected. 'Stampede' also meant that many rushed into kin selection without bothering to read Hamilton's (1964) paper very closely, or at all (see, e.g., Grafen, 1982).

The scientific breakthrough in sociobiology had come with Hamilton's mathematical demonstration of the conditions under which it would be possible for altruism as a genetic trait to evolve. (Altruism was here defined strictly as behavior that benefits the recipient, while incurring a cost for the donor.) 'Hamilton's rule' (Hamilton, 1963) postulates that altruism can evolve if the reduction in fitness (number of offspring) of the donor is more than made up for by the increased fitness (number of offspring) of the recipient. It is simply expressed in the formula $k > 1/r$, where k is the cost – benefit ratio of the behavior and r is the coefficient of relatedness between two individuals.

Hamilton developed his theoretical argument in more detail mathematically in a long two-part paper in the *Journal of Theoretical Biology 1964* (this became part of his PhD thesis which consisted of his first four published articles on aspects of social behavior). His basic ambition was to improve on R.A. Fisher's 'Fundamental Theorem of Natural Selection,' which mathematically expressed evolution as a change of gene frequencies in populations. Fisher in his path-breaking 1918 paper had shown how evolution could be interpreted mathematically – it could be regarded as the work of natural selection on the existing genetic variability in a population (assumed to be generated by constantly emerging small mutations). This is why Fisher equated fitness with the availability of such variation, that is, the statistical 'variance.'

Hamilton, a great admirer of Fisher, while always striving for the most general principle, realized that he could make Fisher's theorem even more fundamental by including the fact that individuals typically live in social groups and affect one another's life chances (especially number of offspring). The concept he coined was 'inclusive fitness.' Inclusive fitness added to Fisher's 'classical' individual fitness an additional component to express the effect of the individual on the fitnesses of its

relatives, as well as their effect on the individual's fitness (the effects could be both positive and negative).

Hamilton struggled a lot with the derivation of this formula and its mathematical formulation, and the paper is often said to be hard to read. It was also almost immediately misunderstood. What was missed, especially, was that 'individuals' here could be any types of organisms, for instance plants (which Hamilton discussed in the paper), and that individuals could share genes without being proper relatives – they could for instance identify each other by some similar phenotypic trait, or even preferences for similar habitats. Hamilton called this a 'superkinship' trait; Richard Dawkins in *The Selfish Gene* later suggested calling it a 'greenbeard trait.' (We see here an implicit recognition of the theoretical possibility of some kind of group selection.)

The paper that launched inclusive fitness was called "The Genetical Evolution of Social Behavior." It discussed in fact four types of social behavior: selfishness, cooperation, altruism, and spite. These could be organized in a 2 by 2 matrix as to the gain and loss to two interacting individuals: A gains, B loses; both gain; A loses, B gains; and both lose. (In 1970, Hamilton modeled spite in a new, useful way with the help of the Price equation, opening up a new field of research.)

Scientists soon started using Hamilton's rule as a convenient shortcut. This opened up a burgeoning new scientific research enterprise of testable models of animal social behavior. Likewise, 'kin selection' soon became the way in which inclusive fitness was understood in practice, that is, as natural selection acting on a group of relatives.

Hamilton and Maynard Smith

It was Maynard Smith (1964) who introduced the term 'kin selection' as an alternative to 'group selection' (a process of natural selection among groups rather than individuals, the idea being that a group of altruistic individuals would have an advantage over other groups; they did not need to be related, just possess the altruistic trait). Maynard Smith (like others, such as George Williams) believed group selection to be an unlikely phenomenon in nature, because even if groups with many altruists would have a clear advantage, the problem was to achieve the initial spread of altruism in a group of selfish individuals. How would individuals with altruistic genes ever be able to proliferate in a group before getting 'outreproduced' by selfish group members?

Hamilton never got over his belief that Maynard Smith, by launching the notion of kin selection, had tried to usurp his (Hamilton's) priority for the solution of the altruism problem. (Hamilton had realized that for altruism to evolve, the benefit of altruistic acts would need to fall on individuals likely to share genes with a donor. For relatives, this likelihood of shared genes, or 'coefficient of relatedness,' is ½ for parent – offspring and for siblings, and 1/8 for first cousins.) As he was struggling with the revision of his long 1964 paper for the *Journal of Theoretical Biology*, Maynard Smith had introduced the term 'kin selection' in a short piece in the highly prestigious scientific journal *Nature*. Also Maynard Smith's oft-repeated reference to J.B.S. Haldane as having 'said it first' triggered recurring bouts of suspicion in Hamilton. Haldane had allegedly said in a pub quip that he would lay down his life for (more than) two brothers or eight first cousins. (For a discussion of the Hamilton – Maynard Smith relationship, including a revealing letter exchange, see Segerstrale, 2013.)

It was Maynard Smith who had been the referee for Hamilton's long manuscript, and had asked for major revisions, which delayed publication. Maynard Smith, however, insisted that his suggestion to split it in two (with 'inclusive fitness' introducing part two) and to provide naturalist examples in fact saved the paper's readability. Also, according to Maynard Smith, the term 'kin selection' was born in a meeting with himself, David Lack, and Arthur Cain as they were together looking for an alternative to Wynne – Edwards' group selection approach.

There was a serious scientific problem as well. The term 'kin selection' did not capture the true, universal meaning of inclusive fitness (see below). It also misrepresented the relationship between kin selection and group selection. Hamilton himself did not see these two as opposed to each other, or kin selection as an alternative to group selection. This was already implicit in his 1964 paper, as seen above, and became clearer in 1970 and 1975 when he re-derived his original inclusive fitness formula from group selectionist premises with the help of the Price equation (which makes it possible to assess the total distribution of genetic variance inside and outside groups in a population).

Hamilton wanted 'inclusive fitness' to be a universal concept – broader than 'kin' selection, 'group selection,' or even 'reciprocal altruism.' He reasoned that many possibilities in fact existed: altruists may live with altruists simply because they have never parted, or because they are able to recognize fellow altruists (later Dawkins' 'greenbeard effect') or due to some pleiotropic effect of the altruistic gene on habitat preference. Therefore, the appropriate choice of terminology – kin selection, group selection, kin-group selection, low migration, assortation, etc. – would depend on the nature of the case in question (Hamilton, 1975).

Hamilton himself, however, did surprisingly little to publicize his later and "more elegant" Price-inspired inclusive fitness model, or even actively correct misunderstandings (others who did so were Dawkins, 1979 and Grafen, 1982; for Hamilton's motives see Segerstrale, 2013).

The Sociobiology Controversy – a Polarizing Episode

The Early Attacks

Wilson had defined sociobiology as 'the systematic study of the biological basis of all social behavior.' The idea was that, just like other features, behavior, too, was undergoing evolution. His book was intended as an encyclopedia of all the new information about animal social behavior, both theories and empirical studies that had accumulated over several decades. In this endeavor Wilson saw himself as continuing the work of the neo-Darwinian or Modern Synthesis – the grand fusion of Mendelism and Darwinism, and its extension to a number of biological fields (e.g., systematics, paleontology, and botany) – by incorporating also the field of behavior.

In his last chapter Wilson argued that a number of human behaviors, including sex roles, aggression, altruism, and even moral and religious beliefs could well have a biological basis.

To boost this argument, he drew parallels to the behavior of other primates and invoked existing research on selected traits from human behavioral genetics and twin studies. For a group of academics radicalized in the 1960s, however, Wilson's statements clearly supported a biological determinist view of humans, suggesting that social inequality was 'in our genes,' which would make social reforms futile.

The announcement of Wilson's book as news on the front page of *New York Times* in May 1975 triggered the spontaneous formation of the Sociobiology Study Group, among which could be found two of Wilson's Harvard colleagues, Richard Lewontin and Stephen J. Gould. The criticism of sociobiology soon escalated from a condemnatory letter in the *New York Review of Books* signed by Boston area academics to the famous incident at the 1978 meeting of the American Association for the Advancement of Science in Washington, DC, where some activists from the Committee Against Racism, rushing up from the audience to Wilson on the podium, poured a pitcher of ice-water in his neck.

The year 1975 was a time of liberal academic outlook and a lingering postwar taboo on biological explanations of human behavior. This was also the time of postwar 'culturism' promoted by people like Margaret Mead in anthropology and the behaviorist B.F. Skinner in psychology. But Wilson had violated the taboo and dared discuss biological underpinnings for human behaviors. The critics decided to forcefully denounce him as a 'bad' scientist, both morally and scientifically.

According to the Sociobiology Study Group, Wilson's real aim was to promote a conservative political agenda by defending social inequality as a natural state of affairs.

Wilson, dismissing his critics as 'tabula rasa Marxists,' steadfastly denied that he was considering these kinds of political consequences; rather, his primary goal was to provoke the social sciences into taking biology seriously. (It is true, though, that Wilson regarded himself as a taboo breaker and took pride in it.) Many may have quietly sided with the critics or been intimidated by the upheaval – very few of Wilson's colleagues came to his defense. The cultural and political climate was not on his side.

The critics' condemnation even extended to the name 'sociobiology.' Wilson himself explained that the logic of this kind of discipline had been apparent much earlier, and that he was simply following in the footsteps of J.P. Scott who had used this name within the Animal Behavior Society for a field located between biology and psychology and sociology (Wilson, 1980). For many political critics, though, the association was directly to 'social biology,' which in turn brought the thinking to 'Sozialbiologie' and its Nazi connections (e.g., Chorover, 1979). And after the scandal around Wilson's book, regular scientists, too, avoided calling themselves sociobiologists lest they be seen as accepting the field's alleged political goals. However, Patrick Bateson at King's College, Cambridge, who wanted a name for functional studies of social behavior that were not necessarily ecological, found the name 'sociobiology' handy. After much discussion with the participants in his new research seminar on sociobiology, and considering the political criticism, the resulting book bravely appeared as written by the King's College's Sociobiology Group and entitled *Current Problems in Sociobiology*. Others again resisted the label 'sociobiologist' because they felt that Wilson's view which embraced group selection, clashed with their own newly acquired gene selectionist view.

The Political Critique of 'Bad Science'

The big political concern of the critics (largely 1960s academic radicals) was that any biological explanation would automatically legitimize the social status quo and discourage social reform. This belief went hand in hand with the assumption that ideologically influenced bad science inevitably leads to bad social consequences. What to do? The critics simply appointed themselves 'weeders' of anything that seemed like bad science. What then was bad science? Anything involving modeling of hypothetical genes and other indirect approaches rather than experimental science – such fields as behavioral genetics, IQ research, and sociobiology (not surprisingly, many leading critics came from experimental fields). A critical industry was born, all set on revealing the underlying ideology of political conservatism, racism, sexism, and so on of 'bad science.' Meanwhile, those who defended sociobiology, behavioral genetics, and similar fields trusted the traditional 'linear' model of scientists producing useful results, which society later decided on in its democratic process (see Segerstrale, 2000a).

In relation to both *Sociobiology* and *The Selfish Gene*, the critics employed a particular reading strategy which I have called 'moral reading.' The aim was to imagine the worst possible social or moral consequences of selected sentences in the book, which were then held up to show the author's nefarious intent. Both Wilson's sometimes speculative language and Dawkins' colorful imagery of genetic strategies were clearly gold mines for the critical industry. Dawkins' particular nemesis was the British neurobiologist Steven Rose.

This move effectively served to obscure the profound difference between the Wilsonian and Dawkinsian approaches. Wilson's sociobiology was of a systemic, integrative kind. He was interested in bringing all social behavior under one big comparative umbrella, making sociobiology a comprehensive quantitative and predictive science. Dawkins, in contrast, wanted to explain to his readers a new way of thinking about evolution. Moreover, in his book, Dawkins was not including the human species. In fact, he relegated humans to the realm of culture – dominated by memes – transmittable information packages. This was largely ignored by the critics.

Centers for Education and Research

As the sociobiology controversy was raging, serious research and education was going on in a number of places. One of the most active leaders in the emerging field of sociobiology – although he did not call the field that – was Richard Alexander at the University of Michigan. Originally a specialist on crickets, his later claim to fame was to predict the existence and behavior of the first sociobiologically behaving mammal – the naked mole rat – before it was actually found. Alexander had retrained himself and also became a practicing human sociobiologist (or rather Darwinian anthropologist). Under his leadership Ann Arbor became an El Dorado for American graduate students of evolution, especially as he regularly invited important figures, such as John Maynard Smith, William Hamilton, or George Williams, to visit for a semester.

Hamilton arrived in 1977, and in 1978 Alexander together with Donald Tinkle was able to arrange a position of Museum professor for him. The many brilliant students from Hamilton's seminars later became eminent professors in the United States.

Meanwhile, George Williams ran an equally dynamic department at SUNY, Stony Brook, for a quarter century after his arrival in 1966, the year of publication of *Adaptation and Natural Selection*. Williams, an ichthyologist, gathered a large number of graduate students around him. Just like Hamilton, he was interested in basic principles, looking especially for seeming exceptions. Both Hamilton and Williams recognized that they were twin brothers of sorts, both in regard to their scientific taste and choice of topics (such as senescence, mate choice, and the evolution of sex), and even in personality features, despite their 10-year age difference. Williams was also the influential editor of the *Quarterly Journal of Biology* during the latter part of the twentieth century.

John Maynard Smith, meanwhile, had become the new Dean of Biology at the University of Sussex in 1965. Originally as an engineer and a student of JBS Haldane, he was a keen follower of biological developments and a great clarifier of issues, with a knack for coining useful concepts (such as kin selection and Evolutionarily Stable Strategy). He typically published lucid short articles on new ideas. As he once admitted, he did not always know where he got his ideas from – indeed, he processed a great amount of material, especially as a reviewer for the *Journal of Theoretical Biology*, *Nature*, and other journals. His collaborations with students and colleagues resulted in a large number of articles. His fair-mindedness gave him an unusual position in the sociobiology debate: although he sympathized with the political views of Lewontin and Gould, he refused to go along with their attack on gene-selectionist reasoning.

Other centers of activity were Oxford and Cambridge. At Oxford Richard Dawkins, John Krebs and others were busy from the early 1970s with various sociobiological game-theoretical scenarios, such as arms races. (However, they did not call the field sociobiology, but rather called it 'behavioral ecology' or 'functional ethology'.) In 1984 they were joined by Hamilton, who had left Ann Arbor to become a Royal Society Research professor at Oxford. By now, Hamilton had moved on to a different topic, the question of sexual reproduction and the role of parasites (pathogens) for its evolution and maintenance. Hamilton had a small number of graduate students of his own (he had no teaching duty), but was typically the informal adviser of visiting postdocs.

At Cambridge, Patrick Bateson, an integrating force, led an empirically oriented multiyear Sociobiology seminar of invited speakers, the King's College's Sociobiology Group, who together produced *Current Problems in Sociobiology*.

Coping with Culture

Sociobiology and Anthropology

In the last chapter of *Sociobiology* Wilson admitted that, compared with culture, evolution by natural selection in general was slow, and also that the genetic differences *between* different human populations were indeed small compared to the genetic variation *within* each population (here, he agreed with those who doubted the usefulness of 'race' as a biological category). Still, in a statement that alarmed his critics, he argued that the genes "maintain a certain amount of influence at least in the behavioral qualities that underlie variations between cultures" and that "even a small portion of this variance may predispose societies toward cultural differences." As examples he mentioned heritable psychological traits such as introversion and extroversion, or dominance and depression. More boldly, he suggested that potential "multiplier effects" might be able to speed up human evolution. What was needed, he suggested, was a new anthropological genetics (1975, pp. 550, 569–570).

Wilson discussed this idea in a short chapter in *Human Social Behavior: An Anthropological Perspective*, edited in 1979 by anthropologists Napoleon Chagnon and William Irons, then at Northwestern University. Biosocial anthropologists, as they called themselves, early on thought along sociobiological lines and there was cross-pollination of sorts. For instance, in *Sociobiology* Wilson quoted Lionel Tiger's idea of a human 'biogrammar.' Another biosocial anthropologist was Robin Fox, the convener of the 1973 Oxford conference on 'biosocial anthropology' and editor of the book with the same name. (Hamilton noted that had this conference been held 2 years later, it would have been classified as human sociobiology.) Fox and Tiger, alone or together, contributed a number of popular books in this broad biosocial or sociobiological spirit, such as *Men in Groups* and *The Imperial Animal*.

More work in the same spirit was pursued by the group at the University of Michigan, who called themselves 'Darwinian anthropologists,' which meant they applied kin-selectionist and other sociobiological models directly to such things as marriage arrangements, patterns of inheritance, or child abuse. Together some of these biosocial and Darwinian anthropologists, including Chagnon and Irons, and the evolutionary psychiatrist Randy Nesse, later took the initiative to the Human Behavior and Evolution Society, founded in 1988, with Hamilton as its first president.

Sociobiology Transforms Itself – or Does It?

While the Chagnon and Irons book contained Wilson's short piece arguing for a genetic anthropology, it also sported the more complex coevolutionary theory of anthropologist William Durham. It is not necessary to presume genetic predispositions for all adaptive behaviors, he argued. Humans might be maximizing their inclusive fitness for both cultural and biological reasons. A type of cultural selection may sometimes even replace natural selection: people may maintain a behavior which is no longer adaptive because of its continuing cultural association with high status.

In the early 1980s, Wilson decided to reconsider the relationship between genes and culture. In *Genes, Mind, and Culture* (1981) produced with Charles Lumsden, a young theoretical physicist, he demonstrated how the human mind, influenced by genes, tended to choose survival-promoting cultural elements. Here he could now respond to those who had criticized sociobiology for not addressing cultural differences. Wilson saw *Genes, Mind, and Culture* as a new version of sociobiology. Still, his and Lumsden's particular coevolutionary process was clearly aimed at 'keeping culture on a leash.' This

was achievable through 'the epigenetic rules' which connected genes with culture by affecting individuals' choices of cultural behaviors or artifacts, seen as isolated units ('culturgens') that could either be chosen or discarded. In *Promethean Fire* (1983), the authors discussed the gene – culture coevolution in a popular form.

Robert Boyd and Peter Richerson at this time (1985) worked with 'dual inheritance' models, which considered biological and cultural evolution at the same time, while Cavalli-Sforza and Feldman (1981) constructed Darwinian models of cultural evolution. These models afforded much greater opportunity for culture to affect genetic evolution and drive behavior to novel equilibria, departing from Lumsden's and Wilson's assumption that genes keep culture on a tight leash.

The idea of a coevolutionary feedback loop that included the mind stimulated other research at the time (while the particular mathematical models from *Genes, Mind, and Culture* were not used). Later, looking back, however, Wilson was disappointed with the lack of progress in the field of gene – culture coevolution, a field where the processes quickly became intractable and the models too complex. But there was also serious criticism. Social scientists strongly disagreed with the depiction of culture as a "sum of mental constructs and behaviors, including the construction and employment of artifacts" (e.g., Leach, 1981, p. 3). Maynard Smith and Warren (1982) questioned the mathematics.

Evolutionary Psychology – the Same or Different?

As Wilson in the 1980s moved away from sociobiology toward environmental activism (Wilson, 1984, 1992), an opportunity opened up for the emerging field of evolutionary psychology. This field was energetically promoted by Harvard-trained anthropologist John Tooby and psychologist Leda Cosmides, who were in turn supported by Harvard's Irven De Vore and other biosocially oriented anthropologists. Donald Symons in turn brought Tooby and Cosmides to Santa Barbara and helped found the Center for Research in Evolutionary Psychology.

Evolutionary psychology tried to actively distance itself from sociobiology and its political connotations. It was not dealing with genes but with universal features of the human mind, which was presented as once and for all formed in the Environment of Evolutionary Adaptedness (EEA). Evolutionary psychology also declared itself uninterested in such things as IQ and behavioral genetics. The mind is seen as a Swiss army knife, composed of a set of modules for solving typical problems encountered in the EEA and culture is based on a combination and extension of these. The 'bible' of this field, *The Adapted Mind* was published in 1992, with Tooby and Cosmides bringing in the established biosocial anthropologist, Jerome Barkow as a coeditor. Just as sociobiology had made the cover of *Time* magazine (1 August 1977), it was evolutionary psychology's turn to do the same (28 August 1995). It was also covered in popular books such as David Buss' *The Evolution of Desire* (1994) and Robert Wright's *The Moral Animal* (1994).

Evolutionary biology may well have seemed as an alternative paradigm to sociobiology. However, Wilson from the very beginning insisted that evolutionary psychology was the same as human sociobiology (Wilson, 1998). At the inaugural meeting of the Human Behavior and Evolution Society in Evanston, 1989, Wilson wondered why it did not call itself the Society for Human Sociobiology instead – the obvious name, according to him.

That was a meeting attended by all the leading names in 'sociobiology' – William Hamilton, E.O. Wilson, George Williams, Richard Alexander, Richard Dawkins, and Irenaeus Eibl-Eibesfeldt. What Wilson meant when he said that evolutionary psychology was 'the same' as sociobiology was most likely his revised view in *Genes, Mind, and Culture* where the epigenetic rules figure prominently. And indeed, the epigenetic rules were described as rules that guided the development of mind.

However, sociobiology is not typically seen as a science of the mind, and the name had already been saddled with political connotations. In 1996 a news notice reported that the Human Behavior and Evolution Society's journal was changing name from *Ethology and Sociobiology* to *Evolution and Human Behavior*. The report produced a protest from a number of HBES members who regarded themselves as practicing sociobiologists (Hrdy et al., 1996).

The Sociobiology Controversy Turns Scientific

Realism versus Adaptationism

As Wilson moved on to gene – culture coevolution, and later on to biodiversity and the environment, it was largely Dawkins who took up Wilson's mantle. In the early 1980s it was rather he and Gould who emerged as the most visible sparring partners in the sociobiology controversy. Meanwhile the controversy also moved away from strictly political issues and toward more philosophical issues underlying evolutionary theory (which were still seen as connected to political values).

While Dawkins wanted to explain more closely the mechanisms of evolution, for instance complex design as an accumulation of small changes over a long time (e.g., the eye, Dawkins, 1996), Gould insisted on accounting for the actual evolutionary forces and levels of selection in nature, paying attention to potential constraints on adaptation. As Dawkins tried to explain the underlying neo-Darwinian reasoning, Gould demanded total realism. A number of popular books came out of this strange symbiosis.

Under this demand for realism, evolutionary psychology emerging in the 1980s could be treated in largely the same way as sociobiology. 'Ultra-Darwinians' became a handy umbrella term for both sociobiologists and evolutionary psychologists (Gould, 1997). Ultra-Darwinians were presented as being interested only in those mistaken adaptive explanations, while the 'pluralists' (Gould) or 'naturalists' (Eldredge) correctly appreciated the true processes of evolution. But later the sociobiological camp, fortified by the philosopher Daniel Dennett, increasingly started examining and rebutting Gould's various claims (e.g., Dennett, 1995; Maynard Smith, 1995; Alcock, 1998, 2001).

Value-Laden Science

Underlying the critics' scientific objection to what they called 'adaptationism' there was their old moral/political concern. If

everything is optimally adapted, there is no point in changing society. But instead of adaptation if you emphasize discontinuity, contingency (historical coincidence), and chance, you indicate that in a radically new environment new types of individuals will flourish. It is not a question of the selection of the fittest, everybody gets his chance (see Gould and Lewontin, 1979 for early expressions of such views).

Later Gould (1989) emphasized chance and contingency even more, arguing, for instance, that in the Cambrian explosion many different phyla existed which later went extinct. Humankind may never have come about. Gould also protested against the idea that evolution has a direction (going, for instance, toward greater complexity). He suggested making the humble bacteria the biological model species. Gould and Lewontin's overall strategy, then, seemed to be to develop the kinds of scientific arguments which, if adopted as social belief and if acted upon, would conceivably have politically desirable rather than undesirable social implications!

Interestingly, Wilson, too, pursued this kind of value-connected agenda at least part of the time. In *Sociobiology* he played down the significance of IQ and dismissed race as a useful biological concept. Later (in *On Human Nature*, 1978) he argued for universal human rights, "because we are mammals," and observed that "the long-term consequence of inequity will always be visibly dangerous." He advocated a more liberal sexual morality and defended homosexuality as socially useful. The social implications Wilson himself derived from his own theories were always liberal.

In the American sociobiology debate both sides pursued science and values together: they criticized undesirable science, derived values from science, and even put science in the direct service of values by trying to produce 'politically correct' science. In the United Kingdom, however, participants tended to adopt the exact opposite strategy: a strictly objectivist one, keeping science and values apart (e.g., Dawkins and Maynard Smith). Wilson switched between an objectivist and a value-laden stance depending on the issue.

A Changing Social Climate

In the 1970s and 1980s it was still too early for biologists to speculate about human nature without triggering an avalanche of protests from fellow academics. In the strongly pro-'culturist' academic climate, Wilson was not able to override his critics' interpretation of *Sociobiology* and convince people that this book was not meant as a political contribution. But this was soon to change.

Toward the end of the 1980s, a notable climate change was already under way. The post – World War II taboo on biological explanation of human behavior appeared to be broken. The importance of biological factors was increasingly recognized in many fields. More dramatically, the development of biotechnology and particularly the Human Genome Project were making genetics into a household word. 'Genetics' and 'behavior' were no longer seen as incompatible when it came to humans. If anything, there was an indulgence in 'gene talk' and ever new reports of genes 'for' behaviors such as 'thrill seeking' having been found.

A particular blow to the culturist paradigm came with Derek Freeman's attacks on Margaret Mead's studies of life in Samoa (Mead, 1928; Freeman, 1983, 1998). New studies documented the actual existence of human cultural universals (Brown, 1991). New research on language, culture, and later even morality, presented animals as much more similar to humans, resulting in popular books such as Frans De Waal's *Good-Natured* (1996) and Robert Wright's *The Moral Animal* (1994). The role of nonverbal communication was rediscovered as a fundamental link between nature and nurture and an important foundation for social life in both animals and humans (e.g., Segerstrale and Molnar, 1997).

A significant background event was the collapse of Marxism in 1989. More serious for the critics, however, may have been the 1980s unexpected shift within the 'academic left' itself, moving from its 1960s political activism toward a new postmodern left. That new 'cultural left' in academia showed little interest in traditional left-wing concerns such as science and ideology, or 'good' and 'bad' science, being now preoccupied with postmodern theoretical analysis and 'standpoint epistemologies' instead (Segerstrale, 2000b).

One factor dampening any new uproar around biology could have been the upswing in religious fundamentalism in the US. Evolutionary arguments could not be radically dismissed any longer unless one wanted to make common cause with conservative creationism.

A Broader Role for Sociobiology?

In *Consilience: The Unity of Knowledge* (1998), partly at least as his answer to the so-called 'Science Wars' in the early 1990s, Wilson gave sociobiology, or rather its genes – mind – culture version, a new and expanded mission. He identified gene – culture coevolution as the solution to the quest for the unity of knowledge and overcoming the Two Cultures problem for the benefit of all. The epigenetic rules were the needed tools for connecting the natural sciences with the social sciences and the arts. By bringing together different fields of knowledge they made it possible for us humans to perceive the world consiliently as a 'seamless web of cause and effect.' Among the epigenetic rules already discovered were universal color terms, genetic control of many behavior-affecting medical conditions, and biological constraints on esthetic preferences. More such rules would be found as science progressed.

Conflict About Cooperation

The sociobiological core research paradigm under its various names has no doubt been fruitful, giving rise to blossoming scientific industries: earlier especially around Hamilton's rule and kin selection, reciprocity, ESS, and evolutionary game theory, and later especially around newer models based on parasite – host coevolution, such as Hamilton's Parasite Red Queen theory for the evolution and maintenance of sex, and the Hamilton – Zuk hypothesis of health (freedom from parasites) as a criterion for sexual selection. Inventing clever strategies for avoiding parasites presented a stimulating challenge for the biological imagination and gave rise to

a plethora of empirical studies. The established explanatory framework turned out to be able to accommodate also unexpected findings revealed by new research; for instance, the intricate mixture of cooperation and conflict in the life of social insects. Kin recognition got serious attention, and behaviors earlier believed to be rare, such as spite, attracted new interest.

The theoretical core paradigm that had emerged from the groundwork by the founding fathers – Hamilton, Williams, Trivers, and Maynard Smith – seemed strong. There did not seem to exist a natural phenomenon that it could not handle; if anything, well-met challenges might be regarded as proof of the robustness of sociobiological explanation.

The regrettable deaths of main figures – Hamilton in 2000, Maynard Smith in 2004, and Williams, in 2010, – did not radically affect the research paradigm; the founders' former students and followers, eminent professors themselves, upheld the legacy of their mentors as they trained their own students and extended the field.

A decade after William Hamilton's death, however, a challenge appeared. In an article in *Nature*, in 2010, three Harvard scientists strongly advocated the abandonment of kin selection, the ruling paradigm for 40 years. According to these scientists, kin selection (identified with inclusive fitness) should just be substituted with regular natural selection theory instead. They argued that inclusive fitness was difficult to calculate and, as an approach, had not generated empirical results, while the theory of kin selection was not holding up in the light of empirical evidence. It seems that they saw the theory as dependent on the validity of the so-called 'haplodiploidy hypothesis' of the origin of eusociality, a complex form of social organization involving high levels of cooperation. (Haplodiploidy is a particular genetic inheritance pattern in social insects which gives rise to sisters being more closely related to each other – and their queen sister – than to their own hypothetical offspring.) Inclusive fitness solved the puzzle of sterility in social insects and explained how it made evolutionary sense for the workers to work for the queen instead of reproducing themselves. However, the idea that eusociality had something to do with the special condition of haplodiploidy, was only a sideline suggestion in Hamilton's (1964) paper. He introduced the social insect example only as an illustration of the idea of inclusive fitness, which was always intended to be a universal theory (see Segerstrale, 2013: chapter 11 for details).

The authors' suggested substitution for kin selection was group selection (or rather multilevel selection). They noted they had shown mathematically that under particular ecological conditions (such as the need for common defense) the cooperation of unrelated insects could give rise to eusociality. Therefore, for eusociality to emerge under these conditions, it would only require the appearance of an allele for cooperation. Genetic relatedness was not needed. The group consisted of E O Wilson working together with evolutionary game theorist Martin Nowak and mathematician Corina Tarnita (Nowak et al., 2010). Wilson saw the issue of insect eusociality as directly connected to the idea of the *Superorganism*. That idea he had recently revived in books written with Bert Holldobler, especially *The Superorganism* (2009).

The strong attack on kin selection met with a mass response from the sociobiological community in the form of letters to the editor of *Nature*, together signed by some 150 scientists (see e.g., Abbott et al., 2011). Kin selection was not an alternative type of selection, it was natural selection. Kin selection and group selection were not opposites, they could be translated mathematically into one another (as Hamilton had shown, based on Price). The paradigm had in fact been fruitful and generated an abundance of research. It did not exclude ecological factors; these could be part of the cost and benefit calculation in Hamilton's rule. Finally, the letter writers noted, the critics with their alternative models for the evolution of cooperation had not demonstrated that relatedness was not important just by leaving it out. The critics did not change their mind.

The challenge, however, may have had a positive effect. The mass response could be seen as a collective display of practicing sociobiologists informing the world and each other about their diverse contributions (as shown in the letters' lists of references). In this way the challenge would have had an integrative function for the field, similar to the integration of subfields that E.O. Wilson originally intended with *Sociobiology*. At the same time, it stimulated the exploration of a broader spectrum of models of cooperation and eusociality, and in this way a further expansion of the field.

See also: Adaptation, Fitness, and Evolution; Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future; Cultural Evolution: Theory and Models; Evolution, History of; Gene–Culture Coevolution; Genetics and Society; Human Behavioral Ecology; Human Cooperation, Evolution of; Kin Selection; Sociobiology: Overview.

Bibliography

Abbott, P., et al., 24 March 2011. Inclusive fitness theory and eusociality. Nature 471, E1–E4 (response to Nowak, Tarnita and Wilson, 2010).
Alcock, J., 1975. Animal Behavior: An Evolutionary Approach. Sinauer Associates, Sunderland, MA.
Alcock, J., 1998. Unpunctuated equilibrium in the Natural History essays of Stephen Jay Gould. Evolution and Human Behaviour 19, 321–336.
Alcock, J., 2001. The Triumph of Sociobiology. Oxford University Press, Oxford.
Alexander, R., 1974. The evolution of social behavior. Annual Review of Ecology and Systematics 5, 325–383.
Alexander, R.D., 1975. The search for a general theory of behaviour. Behavioral Science 20, 77–100.
Alexander, R.D., 1979. Darwinism and Human Affairs. Pittman, London.
Allen, E., et al., November 13, 1975. Letter. The New York Review of Books, vol. 182, 184–186.
Barkow, J.H., Cosmides, L., Tooby, J., 1992. The Adapted Mind. Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.
Bateson, P.P.G., Hinde, R.A., 1976. Growing Points in Ethology. Oxford University Press, Oxford.
Boyd, R., Richerson, P., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.
Brown, J., 1975. The Evolution of Behavior. W.W. Norton, New York.
Brown, D.E., 1991. Human Universals. McGraw-Hill, New York.
Buss, D., 1994. The Evolution of Desire. Basic Books, New York.
Cavalli-Sforza, L.L., Feldman, M.W., 1981. Cultural Transmission and Evolution. Princeton University Press, Princeton, NJ.
Chagnon, N.A., Irons, W. (Eds.), 1979. Evolutionary Biology and Human Social Behavior: An Anthropological Perspective. Duxbury, North Scituate, MA.
Chorover, S., 1979. From Genesis to Genocide. Cambridge and London. MIT Press, Cambridge, MA.

Crook, J., 1976. Multiple reviews of Wilson's Sociobiology. Animal Behaviour 24, 703–704.
Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.
Dawkins, R., 1979. Twelve misunderstandings of kin selection. Zeitschrift fur Tierpsychologie 51, 184–200.
Dawkins, R., 1989. The Selfish Gene, Second edition. Oxford University Press, Oxford.
Dawkins, R., 1996. Climbing Mount Improbable. W.W. Norton, New York.
Dennett, D., 1995. Darwin's Dangerous Idea. Simon and Schuster, New York.
DeWaal, F., 1996. Good-natured. Harvard University Press, Cambridge, MA.
Durham, W.H., 1978. Toward a coevolutionary theory of human biology and culture. In: Caplan, Ad (Ed.), The Sociobiology Debate. Harper & Row, New York.
Durham, W.H., 1991. Coevolution: Genes, Culture and Human Diversity. Stanford University Press, Stanford, CA.
De Waal, F., 1996. Good-natured. Harvard University Press, Cambridge, MA.
Eisenberg, J., Dillon, W.S., 1971. Man and Beast: Comparative Social Behavior. Smithsonian, Washington, DC.
Fox, R. (Ed.), 1975. Biosocial Anthropology. John Wiley & Sons, New York.
Freeman, D., 1983. Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Harvard University Press, Cambridge, MA.
Freeman, D., 1988. The Fateful Hoaxing of Margaret Mead: Historical Analysis of Her Samoan Research. Westview Press, Boulder, CO.
Gould, S.J., 1989. Wonderful Life. W.W. Norton, New York.
Gould, S.J., 26 June 1997. Evolution: the pleasures of pluralism. The New York Review of Books 47–52.
Gould, S.J., Lewontin, R.D., 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London B 205, 581–598.
Grafen, A., 1982. How not to measure inclusive fitness. Nature 298, 425–426.
Hamilton, W.D., 1963. The evolution of altruistic behavior. The American Naturalist 97, 354–356.
Hamilton, W.D., 1964. The genetical theory of social behavior. I and II. Journal of Theoretical Biology 7 (1–16), 17–32.
Hamilton, W.D., 1970. Selfish and spiteful behaviour in an evolutionary model. Nature 228, 1218–1220.
Hamilton, W.D., 1975. Innate social aptitudes of man: an approach from evolutionary genetics. In: Fox, R. (Ed.), Biosocial Anthropology. John Wiley & Sons, New York, pp. 133–157.
Hamilton, W.D., 1980. Sex versus Nonsex versus parasites. Oikos 35, 282–290.
Hamilton, W.D., Zuk, Marlene, 1982. Heritable true fitness and bright birds – a role for parasites. Science 218, 384–387.
Hrdy, S., et al., 1996. Sociobiology's successes. Letters Science 274, 162–163.
King's College (University of Cambridge) Sociobiology Group, 1982. Current Problems in Sociobiology. Cambridge University Press, Cambridge, New York.
Leach, E., 1981. Biology and social science. Wedding or rape? Nature 291, 267–268.
Lorenz, K., 1966. On Aggression. Methuen, London.
Lumsden, C.L., Wilson, E.O., 1981. Genes, Mind and Culture: The Coevolutionary Process. Harvard University Press, Cambridge, MA.
Lumsden, C.L., Wilson, E.O., 1983. Promethean Fire. Harvard University Press, Cambridge, MA.
Maynard Smith, J., 1964. Group selection and kin selection. Nature 201, 1145–1147.
Maynard Smith, J., 1976. Group selection. Quarterly Review of Biology 51, 277–283.
Maynard Smith, J., 30 November 1995. Genes, memes, and minds. Review of D. Dennett: Darwin's Dangerous Idea: Evolution and the Meanings of Life. The New York Review of Books 46–48.
Maynard Smith, J., Warren, N., 1982. Models of cultural and genetic change. Evolution 36, 620–627.
Mead, M., 1928. Coming of Age in Samoa. Morrow, New York.
Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. Nature 466, 1057–1062.
Segerstrale, U., 2000a. Defenders of the Truth: The Battle for Science in the Sociobiology Dehate and beyond. Oxford University Press, Oxford.
Segerstrale, U. (Ed.), 2000b. Beyond the Science Wars. SUNY Press, Albany, NY.
Segerstrale, U., 2002. Neo-darwinism. In: Pagel, M. (Ed.), Encyclopedia of Evolution. Oxford University Press, Oxford, pp. 107–110.
Segerstrale, U., 2013. Nature's Oracle: The Life and Work of W.D. Hamilton. Oxford University Press, Oxford.
Segerstrale, U., Molnar, P. (Eds.), 1997. Nonverbal Communication: Where Nature Meets Culture. Lawrence Erlbaum, Mahwah, NJ.
West-Eberhard, M.J., 1975. The evolution of social behavior by kin selection. Quarterly Review of Biology 50, 1–33.
Williams, G.C., 1996. Adaptation and Natural Selection: A Critique of Some Contemporary Thought. Princeton University press, Princeton, NJ.
Wilson, E.O., 1971. The Insect Societies. Harvard University Press, Cambridge, MA.
Wilson, E.O., 1975. Sociobiology: The New Synthesis. Harvard University Press, Cambridge, MA.
Wilson, E.O., 1978. On Human Nature. Harvard University Press, Cambridge, MA.
Wilson, E.O., 1980. A consideration of the genetic foundation of human behavior. In: Barlow, G.B., Silverberg, J. (Eds.), Sociobiology: Beyond Nature/Nurture? AAAS Selected Symposium 35, 295–306. Westview Press, Boulder, CO, p. 295.
Wilson, E.O., 1984. Biophilia. Harvard University Press, Cambridge, MA.
Wilson, E.O., 1992. The Diversity of Life. Belknap Press of Harvard University Press, Cambridge, MA.
Wilson, E.O., 1994. Naturalist. Island Press, Washington, DC.
Wilson, E.O., 1998. Consilience: The Unity of Knowledge. Alfred Knopf, New York.
Wilson, E.O., Holldobler, B., 2009. The Superorganism. W.W. Norton, New York.
Wright, R., 1994. The Moral Animal. Random House, New York.
Wynne-Edwards, V.C., 1962. Animal Dispersion in Relation to Social Behavior. Oliver & Boyd, Edinburgh.

Sociobiology: Overview

Martin Daly, University of Missouri, Columbia, MO, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Daly and M. Wilson, volume 21, pp. 14539–14543, © 2001, Elsevier Ltd.

Abstract

'Sociobiology' refers to the study of social phenomena as products of Darwinian selection. It has been the dominant approach to the study of animal social behavior since the late 1960s, although the term has fallen from fashion. Its principal architects were G.C. Williams, who persuaded biologists that selection produces adaptive attributes that serve the interests of individuals, not groups, and W.D. Hamilton, who expanded Darwin's concept of reproductive 'fitness' to encompass the individual's impact on the replication of her genes in both descendant and collateral kin. Essentially the same conceptual approach is now more often called 'behavioral ecology.'

'Sociobiology' refers to the study of social phenomena within the conceptual framework of biological science. More specifically, it usually refers to an explicitly neo-Darwinian approach that grew out of the theoretical contributions of Hamilton (1964) and Williams (1966). The hallmark of this approach is the interpretation of social behavior, and its causal mechanisms and processes, as products of the Darwinian processes of natural and sexual selection (*see* Darwinism), and hence in terms of adaptive function or 'design' (Dawkins, 1986).

Scope and History of 'Sociobiology'

The term 'sociobiology' was apparently coined by J. P. Scott, a student of the genetics and development of mammalian social behavior, who began organizing paper sessions on 'animal behavior and sociobiology' at scientific meetings around 1950. In 1956, the Ecological Society of America established a new section of 'animal behavior and sociobiology,' the forerunner of today's Animal Behavior Society, with Scott as its chairman (Dewsbury, 2000). However, the word 'sociobiology' was not widely used before 1975, when Harvard entomologist Edward O. Wilson made it the title of an ambitious volume reviewing and synthesizing information and ideas about the evolution of social behavior across the animal kingdom.

Wilson (1975, p. 4) defined sociobiology as "the systematic study of the biological basis of all social behavior," but although this definition is frequently quoted, it does not convey the word's meaning in actual usage. In the broadest sense, since 'biology' is defined as "the study of living organisms, divided into many specialized fields that cover their morphology, physiology, anatomy, behavior, origin, and distribution" (*Oxford On-line Dictionary*), and since only living organisms engage in 'social behavior,' Wilson's definition could be construed as inclusive of all social science. However, 'sociobiology' is obviously not used as a synonym for 'social science' and most social scientists do not consider themselves biologists. A narrower construction of the meaning of 'biology' does not help Wilson's definition conform to usage: for most scientists, the phrase 'biological basis' connotes the genetic, endocrine, and brain mechanisms underlying social behavior, but these 'proximate causal' factors are not the primary focus of sociobiologists, and most work in these areas would not be considered sociobiological. In practice, both Wilson's tome and subsequent sociobiological work have been primarily concerned with applying theories from population biology, genetics, and ecology to the problem of explaining the evolution of social phenomena and their diversity both among and within animal species (*see* Comparative Method in Evolutionary Studies), in much the same way as one might endeavor to understand the evolution and diversity of anatomical structures.

Like anatomists and physiologists, whose research is guided by their interpretations of the component parts and subsystems of organisms as devices 'for' digestion, respiration, vision, circulation of the blood, and so forth, sociobiologists are 'adaptationists' (*see* Adaptation, Fitness, and Evolution). That is, sociobiologists assume that social phenomena – such as mating preferences, offspring sex ratios, warning calls, systematically discriminative treatment of young by parents, situationally contingent gregariousness, seasonality of territorial defense, and so on – have a functional significance for the behaving animal that can be illuminated by appropriate hypothesis testing. Moreover, sociobiologists are 'selectionists.' That is, their hypotheses about the adaptive functions of social phenomena generally entail the assumption that the social attributes under consideration have been shaped by a history of Darwinian selection to be effectively fitness promoting in ancestral environments (*see* Adaptation, Fitness, and Evolution). Indeed, effective contribution to fitness (the replicative success of the focal individual's genes relative to their alleles, and hence genetic posterity) is considered the superordinate criterion of adaptive function; more immediate aims, such as foraging efficiently, attracting mates, and protecting one's young from predators, are interpreted as having evolved as means to the end of fitness.

Sociobiological theory and research began to flourish only after the fallacy of naïve group-level adaptationism or 'greater-goodism' (Cronin, 1991) was dispelled. For over a 100 years, Darwin's (1859) concept of 'natural selection' was widely misunderstood, even by evolutionary biologists, who often uncritically interpreted evolved attributes as having been selected to promote 'the reproduction of the species'.

Williams (1966) dissected and demolished this fallacy, persuading a majority of evolutionists that the capacity of Darwinian selection to create adaptive organization resides primarily in the differential reproductive success of alternative 'designs' within species, that the attributes that evolve therefore tend to be functionally organized to help individual organisms to outreproduce conspecific rivals, and that any ensuing benefits or harms to a whole population or species are incidental byproducts of organismic designs rather than what the designs are organized to achieve. This critique inspired a view of organisms as 'evolved reproductive strategists' whose attributes must be understood as tactics for effective reproductive competition against others of their own kind. (It is important to note that the 'strategy' metaphor carries no implication of foresight or intent. Indeed, the concept applies to brainless organisms as fully as to those with complex cognition. Thus, a sociobiologist might refer to a plant's 'reproductive strategy,' and decompose it into special-purpose 'tactics,' such as germinating in response to some degree of abrasion of the seed husk, flowering in response to a threshold air temperature, avoiding self-fertilization by asynchronous maturation of male and female flower parts, and so forth.)

The second main conceptual development leading to the flowering of sociobiological theory and research in the final third of the twentieth century was Hamilton's (1964) formulation of 'inclusive fitness' or 'kin selection' theory (*see* Kin Selection). Classical Darwinism had construed the fitness that selection tends to maximize as personal reproductive success, but Hamilton noted that selection will favor any phenotype (*see* Genotype and Phenotype) that effectively promotes the replicative success of copies of the focal organism's genes relative to their alleles (alternative forms of the same gene), regardless of whether those genes reside in descendants or in collateral relatives. Thus, for example, the complex brood care and colony maintenance behaviors of sterile worker ants can be favorably selected if these acts promote the reproduction of a queen who is closely related to the workers. Hamilton proposed that the quantity that selection tends to maximize is not personal reproduction (sometimes called 'direct fitness'), but the focal individual's 'inclusive fitness effect,' which is its total impact on the proliferation of its alleles via both direct fitness and any effects that the focal individual may have on the reproductive success of collateral kin (indirect fitness). Hamilton's analysis implied that organisms are not simply evolved 'reproductive strategists,' but evolved 'nepotistic strategists.' In this context, 'nepotism' means directing benefits to one's genetic relatives, and the implication of Hamilton's revision of the fitness concept was, to quote Alexander (1979, p. 46), that living creatures "should have evolved to be exceedingly effective nepotists, and we should have *evolved* to be nothing else at all." (Note that Alexander, like most evolutionary biologists, uses the word 'nepotism' to encompass acts that promote both the direct and indirect components of fitness; confusingly, however, some authors use this term to refer only to acts that promote indirect fitness, excluding parental care from its domain. Even more confusing is the common misuse of the term 'kin selection' as a substitute for 'nepotism,' such that an individual who is helping relatives is said to be exhibiting or engaging in kin selection. Used correctly, 'kin selection' refers to a population genetical process whose expected phenotypic consequence, on an evolutionary time scale, is nepotism, which is often but not necessarily manifested as selectivity in social response. Many sociobiologists, however, including Hamilton, have eschewed this popular phrase, both because it has so often been misused and because even its correct usage is often misinterpreted as logically parallel to, and theoretically alternative to, 'individual selection' or 'group selection'.)

In the late twentieth century, the word 'sociobiology' fell from favor, with most practitioners instead referring to their field as 'behavioral ecology.' This label has a slightly broader scope since it encompasses the adaptationist, selectionist approach to not just social phenomena but asocial behavior as well, especially foraging. Thus, although one might suppose from its title that the journal *Behavioral Ecology and Sociobiology*, established in 1976, would cover a broader domain than *Behavioral Ecology*, established in 1990, the scope of the two journals is identical. Declining use of the word 'sociobiology' should not be mistaken for a decline in the paradigm's popularity. Indeed, during the 1980s, the sociobiological approach came to dominate the study of animal social behavior, as it still does (Alcock, 2001).

The 'Problem of Altruism'

A central concern of sociobiological theory and research has been the 'problem of altruism': if Darwinian selection is a process that favors and refines only those attributes that contribute to their bearers' fitness relative to the fitness of conspecific rivals, then why is nature not more consistently 'red in tooth and claw'? In other words, how can such prosocial phenomena as restraint in aggressive conflict or active cooperation have evolved? In this connection, sociobiologists define 'altruism' as action whose average consequence is a reduction in the actor's direct fitness (expected reproductive success) and an increase in the direct fitness of someone else. (It should be noted that although this definition encompasses much of what might be deemed 'altruistic' in an ordinary English sense, since its essential feature is self-sacrifice that benefits others, it differs from ordinary usage in the fact that other-regarding intent is irrelevant.) At first glance, it would appear that selection must always penalize action that is altruistic in this technical sense, and yet animals often perform acts that clearly qualify, such as working to help raise others' young, or paying some energetic or predation risk cost in order to warn neighbors of danger. How these phenomena came into existence and are able to persist in a Darwinian world was a mystery before the 1960s, although the unexamined greater-goodism of many biologists kept them from recognizing that there was something in need of explanation.

Hamilton's inclusive fitness theory remains the most frequently invoked and most successful answer. If the beneficiary of an altruistic act is a relative of the actor, and if that relative's gain is sufficiently large relative to the actor's sacrifice, then the net inclusive fitness effect for the actor can be positive and the altruistic tendency can increase in its relative frequency. More specifically, Hamilton theorized that the necessary condition for a heritable altruistic tendency to increase in prevalence under selection (often called 'Hamilton's rule') is $r * b > c$, where c is the cost (in units of expected direct fitness)

incurred by the actor, *b* is the benefit (in the same currency) enjoyed by the altruistic act's beneficiary, and *r* is a 'coefficient of relatedness' between the actor and the beneficiary, indicative of the extent to which the two share identical genes by virtue of descent from recent common ancestors.

A great deal of sociobiological research has been concerned with testing whether altruistic behavior in nature can be understood as nepotism in accordance with Hamilton's rule (*see* Kin Selection). Supportive evidence has mostly taken the form of demonstrations that animals behave differently as a function of their relatedness to their interactants or neighbors. Animals such as colonial ground squirrels and prairie dogs, for example, have been found to call out warnings of predator presence facultatively, according to the potential caller's degrees of relatedness to others in the vicinity; similarly, adult animals usually provide costly care only to their own young (Daly and Wilson, 1988), and where this is not the case, parental 'helpers' still tend to discriminate with respect to relatedness to themselves (Griffin and West, 2003). There has also been considerable interest in elucidating how animals 'recognize' kin when they behave nepotistically (Krupp et al., 2011); in some cases, they respond to cues that are under more or less direct genotypic control, such as certain odors, and in others to reliable circumstantial cues (e.g., those who share my mother's womb with me must be close kin), while in still other cases there is apparently no direct discrimination at all but a 'rule of thumb' that achieves *de facto* discrimination (e.g., behave cooperatively until I emigrate at maturity, at which time my average level of relatedness to those that I encounter will plummet).

Theoretical sociobiologists have also sought nonnepotistic solutions to the problem of altruism, with particular interest in the circumstances that permit reciprocity (you scratch my back and I'll scratch yours) to be evolutionarily stable. This issue has been addressed primarily by game theoretical models, which have highlighted the necessity that nonreciprocators (cheaters) be detected and denied the benefits provided to reciprocators, as well as the difficulty of establishing the requisite punishment of cheaters in multiparty games, when punishing is itself a sort of altruism (i.e., the provision of a 'public good'). Empirical investigations of nonnepotistic reciprocity in nonhuman animals have been few, however, and claims of such reciprocity are open to alternative explanations (Clutton-Brock, 2009), perhaps indicating that the complex social cognition that is necessary for all but the most limited forms of reciprocation-in-kind is restricted to human beings and our near relatives.

Since 2000, there has been a resurgence of interest in theoretical models of natural selection among social groups. It is generally accepted, however, that 'multi-level selection' models and 'kin selection' models are mathematically equivalent, providing alternative perspectives on the same phenomena (Queller, 1992). Despite several attempts to unseat Hamilton's theory, it remains the cornerstone of theories of social evolution (Gardner et al., 2011; Queller, 2011).

Sexual Selection and Sexual Conflict

Another major focus of sociobiological theory and research has been sexual reproduction. Problems include when and why organisms reproduce sexually (i.e., biparentally) rather than uniparentally, and, in the event of 'dioecy' (the existence of two sexes with distinct adaptations), exactly how and why females and males differ. In 1871, Darwin (1871) supplemented his 1859 theory of evolution under the influence of 'natural selection,' by proposing a distinct evolutionary force that he called 'sexual selection' (*see* Darwinism). Whereas natural selection referred to the nonrandom differential reproduction of types as a result of differential success in coping with such exigencies as finding food, avoiding predators, and enduring climatic factors, sexual selection referred to the nonrandom differential reproduction of types as a result of differential access to mates and their reproductive labors. Darwin distinguished the two because they could sometimes act in opposition, as for example when sexual selection favors the evolution of conspicuous courtship structures, such as the peacock's tail, that enhance mating success but apparently also increase mortality risk.

Sexual selection was the object of scant theoretical or empirical work for a century after Darwin proposed the concept. Williams (1966) and Trivers (1972) resurrected it as a focal topic in sociobiology, proposing that the extent to which sexual selection operates differently on females and males, and thus produces sex differences in anatomy, physiology, psychology, and behavior, was determined by the extent to which the reproductive efforts (especially parental care) of one sex constituted the principal resource limiting the fitness of the less investing sex (usually, but by no means always, males), and hence the object of competition among members of the less investing sex.

More recent work has focused on the criteria and consequences of mating preferences (Andersson, 1994), and on conflicts of interest between the sexes that may lead to both antagonistic male- and female-specific selection pressures at the same genetic loci and sex-specific adaptations 'designed' to nullify one another (Arnqvist and Rowe, 2005).

'Selfish Genes' and Intragenomic Conflict

An area of great contemporary interest concerns the evolutionary consequences of the fact that selection can operate differentially on different components of the genome. Because the male sex-determining Y chromosome is transmitted only from father to son in such animals as fruit flies and human beings, for example, mutations on the Y that have the effect of biasing a progeny toward male production have been observed to increase in prevalence over generations, even when they reduce the inclusive fitness of the host individual. Conversely, mitochondrial DNA and other cytoplasmic inheritance factors are transmitted only from mother to daughter, and mutations in these factors that bias sex allocation toward females may similarly proliferate despite negative inclusive fitness effects. The consequence is 'intragenomic conflict': simultaneous selection at different genetic loci, within the same population, in favor of alleles with antagonistic phenotypic effects (Burt and Trivers, 2006).

The recent discovery of 'genomic imprinting,' a process whereby certain genes are differentially activated according to whether they were inherited from one's father or mother, has greatly extended the known potential for such intragenomic

conflict, as for example when genes of paternal origin are selected to produce fetal and infantile phenotypes that effectively 'discount' the value of the mother's potential future offspring (who may have distinct paternity) more steeply than do genes of maternal origin, leading to the evolution of directly antagonistic phenotypic effects of maternally and paternally imprinted genes (Haig, 2002). In a sense, these discoveries challenge the orthodox sociobiological view of organisms as nepotistic strategists with an integrity of purpose subsidiary to the function of maximizing inclusive fitness, since selection clearly can and does favor the proliferation of genes that promote their own replication at the expense of 'their' organisms' inclusive fitness. However, the individual-level focus of sociobiological research continues to be productive (Alcock, 2001) and is unlikely to be abandoned soon.

See also: Adaptation, Fitness, and Evolution; Comparative Method in Evolutionary Studies; Darwinism; Genotype and Phenotype; Human Behavioral Ecology; Kin Selection.

Bibliography

Alcock, J., 2001. The Triumph of Sociobiology. Oxford University Press, New York.
Alexander, R.D., 1979. Darwinism and Human Affairs. University of Washington Press, Seattle, WA.
Andersson, M., 1994. Sexual Selection. Princeton University Press, Princeton, NJ.
Arnqvist, G., Rowe, L., 2005. Sexual Conflict. Princeton University Press, Princeton, NJ.
Burt, A., Trivers, R., 2006. Genes in Conflict: The Biology of Selfish Genetic Elements. Belknap Press, Cambridge, MA.
Clutton-Brock, T., 2009. Cooperation between non-kin in animal societies. Nature 462, 51–57.
Cronin, H., 1991. The Ant and the Peacock: Altruism and Sexual Selection from Darwin to Today. Cambridge University Press, Cambridge, UK.
Daly, M., Wilson, M.I., 1988. The Darwinian psychology of discriminative parental solicitude. Nebraska Symposium on Motivation 35, 91–144.
Darwin, C., 1859. On the Origin of Species by Means of Natural Selection. John Murray, London.
Darwin, C., 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.
Dawkins, R., 1986. The Blind Watchmaker. Longman, Harlow, UK.
Dewsbury, D., 2000. Obituary: John Paul Scott (1909–2000). Animal Behavior Society Newsletter 45 (2), 3–5.
Gardner, A., West, S.A., Wild, G., 2011. The genetical theory of kin selection. Journal of Evolutionary Biology 24, 1020–1043.
Griffin, A.S., West, S.A., 2003. Kin discrimination and the benefit of helping in cooperatively breeding vertebrates. Science 302, 634–636.
Haig, D., 2002. Genomic Imprinting and Kinship. Rutgers University Press, New Brunswick, NJ.
Hamilton, W.D., 1964. The genetical evolution of social behavior, I and II. Journal of Theoretical Biology 7, 1–16, 17–52.
Krupp, D.B., DeBruine, L.M., Jones, B.C., 2011. Cooperation and conflict in the light of kin recognition systems. In: Salmon, C.A., Shackleford, T.K. (Eds.), The Oxford Handbook of Evolutionary Family Psychology. Oxford, New York, pp. 345–364.
Queller, D.C., 1992. Quantitative genetics, inclusive fitness, and group selection. American Naturalist 139, 540–558.
Queller, D.C., 2011. Expanded social fitness and Hamilton's rule for kin, kith, and kind. Proceedings of the National Academy of Sciences USA 108, 10792–10799.
Trivers, R.L., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual Selection and the Descent of Man 1871–1971. Aldine, Chicago, pp. 136–179.
Williams, G.C., 1966. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press, Princeton, NJ.
Wilson, E.O., 1975. Sociobiology: The New Synthesis. Belknap Press, Cambridge, MA.

Sociocybernetics

Bernd R Hornung, Philipps University Marburg, Marburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Geyer, volume 21, pp. 14549–14554, © 2001, Elsevier Ltd.

Abstract

The article outlines the origin of sociocybernetics and its roots in systems science and cybernetics. It explains key concepts of first-order cybernetics, the cybernetics of observed systems, and second-order cybernetics, the cybernetics of observing systems. Central issues in the latter are organization, self-organization, and autopoiesis, all based on circular causality. The epistemological foundations in constructivism are exposed, which are closely related to the problem-oriented interdisciplinary approach inherent in systems science, cybernetics, and sociocybernetics. The article concludes with an argument for the relevance of sociocybernetics as a social science of complexity in the contemporary global and dynamic world.

The Concept and Its Origin

'Sociocybernetics' is a term belonging to a whole family of related terms like 'cybernetics', 'systems and cybernetics', 'systems theory', 'general system theory', 'systems approach', 'systems research', '(applied) systems analysis', 'systems methodology', 'systems thinking', 'systems philosophy', etc. Here, we cannot go into the details of these variations. Instead, we concentrate on 'systems' and 'cybernetics,' which are the key concepts. In general use, they are frequently interchangeable or appear in combination (e.g., in the names of WOSC – World Organization of Systems and Cybernetics and AFSCET – Association Française de Science et des Systèmes Cybernétiques, Cognitifs et Techniques). Hence they can be considered as synonyms like in this contribution, although the two terms come from different traditions and are not used uniformly in different languages and different national contexts.

Sociocybernetics is defined as 'systems science in sociology and other social sciences', whereby 'systems science' includes 'cybernetics'. The term systems science was chosen because 'science' is not limited to theory. It includes application, empirical research, methodology, axiology (i.e., ethics and value research), and epistemology (Geyer and van Dijkum, 1999: pp. 11–28). The term 'sociocybernetics' was first used as title of a book published by Felix Geyer and Johannes van der Zouwen (1978). The use of systems science and cybernetics in the social sciences dates back much further of course, even to the very roots of cybernetics like Norbert Wiener's book on *Cybernetics or Control and Communication in the Animal and the Machine* (1948 also 1956), whereby the 'animal' includes the human being and communication is not just a technical but also a very social phenomenon. A second root, on the systems side, is biology (Bertalanffy, 1968).

Under the name of 'sociocybernetics', however, systems science has made its way into the International Sociological Association (ISA) since 1980 and was firmly established as an ISA Research Committee at the 1998 World Congress of Sociology in Montreal (Geyer and van der Zouwen, 1986; Geyer, 1997; Marcuello Servós, 2006).

Systems, Cybernetics, and Circular Causality

Certain systemic or cybernetic aspects have been present in the social sciences since long. In a way, Marx can be considered a systems theorist just like Durkheim with regard to other aspects and, of course, Talcott Parsons. Sociocybernetics is different from these approaches as it strives to make full use of the complete set of theoretical and methodological tools systems science and cybernetics provide, which have flourished and spread since the Second World War (cf Buckley, 1968).

This means that sociocybernetics includes both so-called first-order cybernetics and second-order cybernetics. According to definition, cybernetics is the science of 'control and communication in the animal and the machine', including human beings and natural 'machines' (Wiener, 1948). This provides already a bridge between the natural sciences and the social sciences, which nowadays is broadened and developed, e.g., by human ecologists. Heinz von Foerster (von Foerster, 1981; von Foerster and Poerksen, 2002) went on to distinguish a first-order cybernetics, 'the study of observed systems', and a second-order cybernetics, 'the study of observing systems' or the 'cybernetics of cybernetics' (von Foerster, 1970). Second-order cybernetics is different from first-order cybernetics in two fundamental aspects. First, it is explicitly based on constructivist epistemology and second, one of its main concerns is self-reference and self-organization, including the autopoiesis of living systems. Second-order cybernetics pays particular attention to the observer dependence of knowledge, including scientific theories (cf Hornung, 1995, 2006). In the interdisciplinary spirit of systems science, sociology is clearly at the center of interest of sociocybernetics. The other social sciences, however, like psychology, anthropology, political science, etc., need to be included too, depending on the research question to be dealt with (Hornung, 2002, 2004, 2006).

This does not mean that first-order cybernetics and its basic concepts can be discarded and ignored by the social scientist. They are basic also to second-order cybernetics. Furthermore, systems science and cybernetics, and hence sociocybernetics, are intrinsically problem oriented. This means that it is a question of the problem at hand, whether for a given research

problem first- or second-order (socio)cybernetics is more appropriate.

A system can be defined as a whole consisting of interdependent components. Basic to any sociocybernetic approach, are the four key concepts which define a dynamic system (Laszlo, 1973):

1. The holon property: A system is more than the sum of its parts and has properties different from those of the individual parts. This also implies a boundary which distinguishes the system from its environment and which can be more or less closed. In the social sciences, it is often difficult to determine the boundary of such a 'unit' of research, because social systems, groups, e.g., often have very open boundaries which makes them difficult to define and to identify.
2. The systems hierarchy: This is not a hierarchy of command but a hierarchy of structure. Every system has parts, i.e., subsystems, and every system is embedded in a larger system. Whether the largest system, the universe, is indeed a system, can be disputed, because its boundary cannot be identified. At the other extreme, the search for the smallest system in the subnuclear sphere is still going on. Social systems are at a mesolevel between these extremes of natural systems. For analyzing systems and for acting in systems, this means that there is always a wider picture but also a narrower one.

These are structural aspects of systems. Two more basic properties refer to processes as follows:

3. Negative feedback: Here negative (or positive) does not imply any normative connotation but simply indicates the direction of an effect. Negative feedback in a chain of cause and effect means that an effect is produced which counteracts the initial cause, i.e., in the end diminishes the effect (negatively). Therefore, it is also called 'deviation reduction'. An example is the thermostat or in social systems negative sanctions, e.g., by legal authorities.
4. Positive feedback: A cause is producing an effect which in the following time step influences the cause in a way that it gets stronger. There is a deviation amplification. This, e.g., is the case of an avalanche in nature or of positive sanctions and reinforcement in social systems. In abstract terms, any case of exponential growth is technically a case of positive feedback, which from a normative point of view may not be positive at all.

The feedback implies circular causality of some sort and the concept of interdependence of parts or subsystems implies long causal chains with many bifurcations. Although the system boundary delimits the area of causes and effects, boundaries are usually not completely closed or they are closed only for certain exchanges. The controversy of open systems (Bertalanffy, 1968, Network) vs closed systems, which are professed in the social science, e.g., by Niklas Luhmann (1995), is by no means resolved. It may be considered another case for the problem-oriented approach. In a similar way, the systems hierarchy is a limiting factor for the spreading of effects. But also the hierarchy is not completely closed and permits upward and downward causality along with phenomena of emergence (Buckley, 1998).

With these concepts, sociocybernetics pushes the social scientist to become aware and to pay attention to long-term effects and side effects of events, actions, or measures taken. These effects include complex feedbacks along many pathways and across many bifurcations. In the end, these may lead to a dramatic positive feedback and amplification of the initial event or also to neutralizing and blocking any effect of even a major effort. Examples for both cases can easily be found in politics and its projects. Sociocybernetics thus provides useful theoretical tools to analyze such phenomena, even using 'only' first-order cybernetics.

Circular causality can be purely mechanical, a structural arrangement of causes and effects amplifying or neutralizing each other, but it also may involve information processing which, in a very simple way, is already present in the thermostat. Temperature is measured, 'observed,' the result is compared to a preset target value, and this comparison determines the reaction – increase or decrease temperature or no reaction. As simple as such information processing is in the thermostat, as complex it is in the human brain and psychological system or in modern societies with their complex institutions of parliaments, governments, administrations, etc.

At this point, second-order cybernetics enters the picture. The research object of the social scientist is not a (mechanical) object anymore investigated by a detached researcher. Instead, it is taken into account that the researcher influences his object of research being part of the (overall) system and not outside. Moreover, the object in its turn is an observing system. With regard to human beings with their psychological and cognitive capacities this is quite evident. In terms of a systems hierarchy, the latter can be seen as cognitive and psychological systems embedded in larger over-arching systems. Taking this into account, sociocybernetics leads away from a Durkheimian view of society as a large mechanism or organism, in which the individual is a small wheel, or from the Marxian view of society as an anonymous big 'system', which determines the life of its victims. Instead, it suggests an actor-oriented view (Kuhn, 1974; Etzioni, 1968; Burns et al., 1985). The social individual with his cognition is conceived as an observer and actor observing and interacting with other observers/actors. These are not necessarily other individuals. They can be social units like groups and organizations which have a capacity for (collective) information processing or also for knowledge and meaning processing. Knowledge can be defined as a network of interrelated information and meaning as knowledge related to goals and values (Bossel, 1998, 1996).

The basic concepts of systems and cybernetics still remain the same, but observation and information/knowledge/meaning processing are added. The recombination of all of this provides a highly complex picture of what are social systems and societies from a sociocybernetic point of view (cf Andreewsky and Delorme, 2006).

Systems, Actors, Steering, and Control

Cybernetics, much more than systems science, is closely associated or even identified with the concept of control, which in ordinary language has a mostly negative connotation. In systems science and cybernetics, however, it is used in a technical sense and as a complement to the concept of steering. Both steering and control, also called feedback and feedforward, are based on circular causality at a level which includes

information processing. They are basically the same process, but feedback reacts to an event that already happened. It is an attempt to change its consequence. Feedforward 'reacts' to an event which is only anticipated. It produces an effect that avoids an anticipated event. Steering a car or a ship – the term 'cybernetics' has its origin in the latter – is clearly based on anticipatory action. There is no point in steering anymore, once the car is after a curve in the field already.

Steering and control usually do not refer to the in-built 'natural' cycles of feedback in a system but to special steering or control units (cf Burns and Buckley, 1976). The main causal chains involved in the operation of such steering and control units – not to mention the (unintended) long term and side effects – are not necessarily limited to one particular subsystem. They can cut across several subsystems and several system levels, as can be seen easily in any large organization (cf Beer, 1972). This means that mechanisms and effects of steering and control in a system need to be clearly distinguished from feedback loops which kind of naturally result from the structure of a system and which frequently even operate behind the backs of the social actors. Hence sociocybernetics does not conceive of social systems and society as mechanisms enforcing their intrinsic laws against the individuals or as mere collections of individuals, i.e., social actors. Instead, they are seen as interactions of systemic structures and processes with social actors. The structures certainly confine and delimit but also enable social action, while the actors also change and modify the larger system which is their environment. Therefore, sociocybernetic theory is neither a pure 'system' theory at the macrolevel nor a pure actor theory but an actor–system theory or actor-oriented approach (Burns et al., 1985), whereby the actor himself is a particular kind of system too.

Organization, Self-Organization, and Autopoiesis

While steering and control are tools for organizing systems, the actor-oriented approach raises the question of the limits to (intentional) organization and self-organization. Organization, from a systems and cybernetic view, is the buildup of negative entropy (negentropy), i.e., of a more improbable state than the equal distribution of matter and energy. This concept comes from thermodynamics originally (Prigogine and Stengers, 1984) but can be used for symbolic and informational structures as well (Bailey, 1990; Stonier, 1990), which always depend on a physical medium, be it a brain, a computer or the paper of a book. Consequently, the concept of entropy and negentropy is also applicable to social systems and sociocultural systems, which are to a considerable extent symbolic and informational (or also conversation (Pask, 1975; Scott, 2011) and communication (Luhmann, 1995)) systems. Building up improbable states, i.e., structures, of matter/energy and information, however, is only one part of the story.

The other part is the persistence or sustainability of such structures (and processes), which, after all, constitute systems. The social sciences are usually dealing with dynamic systems. To be sustainable (or viable), they cannot exist with just any structure. Their structure needs to fulfill certain functions which permit survival in a dynamic environment (Bossel, 1996, 1998). This means to be able to react adequately (by steering and/or control) to influences from the environment which is apt to disturb the viability. This is possible only, if a system disposes of enough variety of possible reactions to the variety of influences, disturbances, or inputs from the environment – however, one wants to call them. This was formulated by Ross Ashby (1956), one of the fathers of cybernetics, in his 'Law of Requisite Variety'. In a somewhat different way, this was reformulated by Niklas Luhmann (1995: p. 26f) in the context of his complexity theory: Only (system) complexity can reduce (environmental) complexity.

This suggests a functional approach to social systems. Indeed, much of sociocybernetics and its predecessors is functionalist (e.g., Parsons, 1966, 1967, 1971; Luhmann, 1995). In applied systems analysis and applied sociocybernetics, it turns out, however, that a functional view is not enough. Also units have to be taken into account, which often have the same set of functions, e.g., communities, states, countries, etc. One of the few systems theorists who explicitly followed the unit approach was James G. Miller (1978). In applied and empirical sociocybernetics, leaving the theoretical level, the two perspectives need to be combined. This has barely been attempted on a theoretical basis. Kenneth D. Bailey (1994) proposed to integrate both views, of social units and of variables resp. functions, using Q–R analysis which comes from psychology.

In this framework, the units, in psychology usually the individuals, constitute one dimension (Q) and the variables/functions constitute a second dimension (R) which cuts across the individuals orthogonally. All of them have the same kinds of properties identified by means of variables/functions.

From this point of view a real-life system, e.g., an organization, is a unit incorporating a set of functions which permits it to survive and to be viable/sustainable in a changing environment. Organizing a system means to set up structures and processes which provide a greater variety of action and reaction – in terms of Ashby's Law of Requisite Variety – or, in Luhmann's terms, more system complexity to better cope with the environmental complexity.

While 'organizing' refers to an intentional effort of an actor to achieve this, 'self-organization' refers to a circular process of buildup of structures and processes which is not the result of the activity of one (central) actor but an unplanned and unplannable process. Such a process is determined by the interactions of several actors inside the system and interactions of those actors with the system structures and processes in which they are embedded.

Similar considerations like for social action hold for information processing (knowledge processing, meaning processing (cf Stonier, 1997)) at the symbolic level. This includes also communication. At the symbolic, cognitive, level, however, the concept of self-reference is used.

Self-organization and self-reference are based on circular causality just like a third concept, autopoiesis. This constitutes still another level of complexity. The theory of autopoiesis was originally developed by Humberto Maturana and Francisco Varela (1980), two Chilean biologists, seeking to explain what is life. They call the process autopoiesis by which a living system produces the very components of

which it is composed out of its own components. An organism uses its own cells to (re)produce the same kind of cells and to stay alive in this way. To do so, it needs matter/energy from outside, which it can transform into cells, i.e., there is a metabolism. At the level of information, however, there is no such import of information. According to Maturana and Varela living – autopoietic – organisms are informationally closed. They react to disturbances, but there is no inflow of something called information.

Niklas Luhmann modified the concept of autopoiesis to apply it to social systems. Whether this is legitimate and whether social systems are to be considered living systems remains disputable, however.

Problem Orientation Based on Epistemological Constructivism

The questions whether the concept and theory of autopoiesis is applicable to social systems, what is 'an actor', e.g., inside an organization, and what are the boundaries of the system under consideration (the 'reference system') show that the answers are not necessarily evident. Also the question can be raised, whether systems do exist indeed.

Systems theory and cybernetics have two answers to these problems.

Early systems theory and cybernetics tried to resolve such issues by resorting to an epistemological nominalism. This means that systems are not considered as entities existing in the empirical real world, but as scientific concepts which permit to consider all kinds of real-world phenomena as systems in order to do the corresponding research and analysis. This, not very satisfactory, epistemological approach was replaced by an epistemological constructivism. In the systems and cybernetics community, this was professed in particular by Heinz von Foerster (von Foerster, 1981; Müller and Müller, 2007). Constructivism, although not under this name, can be traced back in the philosophical tradition to Husserl, Schopenhauer, and Kant.

Epistemological constructivism asserts that 'things' are not ontologically given as such (which would be realism). However, it also asserts that they are not mere inventions of human thinking without any relation to an outside 'real' world (which would be idealism). Instead, constructivism asserts that something exists, which cannot be perceived, however. It is the perceiver or observer who names certain phenomena according to his convenience and who constructs his perceptions and cognitions. These constructions, however, are not arbitrary, as they have to prove their validity in practical tests of practical life. This view corresponds very much with the paradigms of Thomas S. Kuhn (1962), who says that a paradigm or also a scientific theory is neither true nor false but only more or less useful.

In a way, epistemological constructivism overcomes the Cartesian dualism. It can also be found in the work of the Mexican scientist Rolando Garcia (e.g., 2000, 2006), in the autopoietic theory of Maturana and Varela and, with certain reservations, in the sociological theories of Niklas Luhmann. The latter was strongly influenced by von Foerster, Maturana and Varela, and also by Edmund Husserl. A recent impulse to the development of constructivism was given by Arne Kjellman (2002, 2003) with his epistemological theory on subjectivism and the priverse.

Sociocybernetic Methodology

The fundamental constructivist approach is also reflected in the problem-oriented methodology frequently used in sociocybernetics. The basic idea is that the object of research is not determined by the givens of existing reality out there but by the research question or research problem. This determines which slice of empirical reality needs to be investigated to resolve such a reference problem and also which disciplines need to be involved. In other words, sociocybernetics is inherently interdisciplinary by its methodology.

Stating the reference problem and the corresponding reference system makes clear what are the relevant functions to be considered, which are the mechanisms or units fulfilling such functions, what is the environment, which are the subsystems, etc. Only then it can be meaningfully stated what is functional, dysfunctional, functionally equivalent, etc., also which structures may be functionally equivalent, thus avoiding the Parsonian conservatism. A systematic picture of this multidimensional theoretical apparatus is given by Hornung and Adilova (1997).

A problem-functionalist approach has a second advantage. It permits from the very beginning to include normative or axiological issues (cf Parra-Luna, 2000). After all, the concept of 'problem' itself has both a factual and a normative component. A problem is a state of affairs (a fact) which needs to be changed (a normative request).

To carry out systems and cybernetic research, computer modeling and simulation can be considered as the paradigmatic method in the sense of Thomas S. Kuhn (1962). With regard to sociocybernetics, however, it needs to be pointed out that social systems are barely suitable for computer modeling for quite a number of reasons which cannot be discussed here (cf van der Zouwen, 2006). Suffice to mention the problem of enough and consistent data to cover an entire model of a social system. However, modeling in general, from a sociocybernetic perspective, covers the whole range from mental or conceptual models to fully quantitative computer models. Conceptual models may be formulated in the form of texts or diagrams (cf Checkland, 1981; Checkland and Scholes, 1990), while qualitative models can be computer based (cf Bossel et al., 1989). Fully quantitative models can be of the Forrester type (systems dynamics models (Forrester, 1968, 1971)) or scenario models as originally developed by Meadows et al. (1972; Meadows and Robinson, 1985; Mesarovic and Pestel, 1975).

Beyond this, empirical and applied sociocybernetics can use any method of empirical research which is scientifically acceptable and suitable to obtain the data desired.

Sociocybernetics – Social Science of Complexity

Science is not just an end in itself but plays a major role in human life, from microbiology and genetics to globalization

and the warming of the planet. The social sciences might be expected to play an especially important role in the organization of social systems and the solving of their problems. Major problem lines of human society can be mentioned here briefly only. Some of them are old, some of them new, all of them are shifting, complexifying, and a few maybe dissolving. The ideology-based conflict between East and West with the Soviet Union has disappeared; the conflict North–South has been changing its character moving toward a more general poverty gap also inside the Western most modern countries. It has given way to what Huntington (2002) calls the 'clash of civilizations'. Overarching remains the conflict between man and nature – the ecological problems – and the problems resulting from globalization.

New kinds of problems are appearing – or at least become evident – at the individual level inside our social systems: psychological disease, depression, burnout, and sectarianism – the market of all kinds of religions. Exploitation, the traditional Marxist concern, is becoming replaced or rather complemented by self-exploitation. In other words, a new kind of alienation has appeared (for a sociocybernetic account of alienation, cf Geyer, 1980). At the root of this seems to be the loss of traditional certainties in combination with an increasing unmanageability both of social systems and of individual life. This in spite – or also because? – of all the progress in science and technology (cf Aguado et al., 2009).

In abstract terms, the ever increasing tremendous complexity of society and its technologies together with the ever accelerating speed of change – social and technological dynamics – risk to strain human capacities beyond their limits and to make the world inhabitable, both physically and psychologically.

Systems science and cybernetics in general and sociocybernetics in particular promise a framework which might be suitable to tackle these issues – at any rate, a better one is not in sight. While the theoretical apparatus seems adequate to build up enough theoretical complexity to cope with the world's complexity, the methodology, in the field of sociocybernetics, remains still insufficient. The question of data can be resolved only partially in specific cases and selected instances. Methodology and data, however, might not be problems of sociocybernetics but result from fundamental properties of the objects we are dealing with: social systems and human action at the level of second-order cybernetics.

In this respect, the ambitions and aspirations of sociocybernetics as a social science of complexity may be, as Johannes van der Zouwen (2006) expresses it in the subtitle of his article, indeed 'a bridge too far'.

See also: Action, Theories of Social; Alienation, Sociology of; Control, Social; Conversation Analysis: Sociological; Human Ecology; Innovation, Theory of; Luhmann's Social Systems: Meaning, Autopoiesis, and Interpenetration; Macrosociology-Microsociology; Methodological Individualism in Sociology; Population Dynamics: Mathematic Models of Population, Development, and Natural Resources; Social Structure; Social System; Sociological Theory; World Systems Theory.

Bibliography

Aguado, J.M., Scott, B., Buchinger, E. (Eds.), 2009. Technology and Social Complexity – Tecnología y Complejidad Social. Universidad de Murcia, Murcia.

Andreewsky, E., Delorme, R., 2006. Seconde Cybernétique et Complexité. Collection INGÈNIUM, L'Harmattan, Paris.

Ashby, W.R., 1956. An Introduction to Cybernetics. Chapman & Hall, London.

Bailey, K.D., 1990. Social Entropy Theory. State University of New York Press, Albany, NY.

Bailey, K.D., 1994. Sociology and the New Systems Theory. State University of New York Press, Albany, NY.

Beer, St, 1972. Brain of the Firm. John Wiley & Sons, Chichester West Sussex, New York.

von Bertalanffy, L., 1968. General System Theory. George Braziller Inc., New York.

Bossel, H., 1996. Ecosystems and society, implications for sustainable development. In: World Futures, vol. 47. OPA - Overseas Publishers Association, Amsterdam, pp. 143–213.

Bossel, H., 1998. Earth at a Crossroads. Cambridge University Press, Cambridge.

Bossel, H., et al., 1989. Wissensdynamik (Knowledge Dynamics). Vieweg Verlag, Braunschweig, Wiesbaden.

Buckley, W. (Ed.), 1968. Modern Systems Research for the Behavioral Scientist. Aldine Publishing Co, Chicago.

Buckley, W., 1998. Society - A Complex Adaptive System. Gordon & Breach Science Publishers, New York, London, Tokyo.

Burns, T.R., Baumgartner, Th, Deville, Ph, 1985. Man, Decisions, Society. The Theory of Actor-system Dynamics for Social Scientists. Gordon & Breach Science Publishers, New York, London, Tokyo.

Burns, T.R., Buckley, W. (Eds.), 1976. Power and Control. Sage Publications, London, New York.

Checkland, P.B., 1981. Systems Thinking, Systems Practice. John Wiley & Sons, London, New York.

Checkland, P.B., Scholes, J., 1990. Soft Systems Methodology in Action. Wiley, Chichester, London, New York.

Etzioni, A., 1968. The Active Society. Collier Macmillan, New York.

Forrester, J.W., 1968. Principles of Systems. Wright Allen Press, Cambridge, MA.

Forrester, J.W., 1971. World Dynamics. Wright-Allen Press, Cambridge, MA.

von Foerster, H., 1970. Cybernetics of Cybernetics. Paper delivered at 1970 annual meeting of the American Society for Cybernetics Washington, DC.

von Foerster, H., 1981. Observing Systems. Intersystems Publications, Seaside, CA.

von Foerster, H., Poerksen, B., 2002. Understanding Systems. Kluwer Academic Publishers/Plenum Publishers (Carl-Auer-Systeme Verlag), New York, Boston, Dordrecht, London, Moscow.

Garcia, R., 2000. El conocimiento en construcción. De las formulaciones de Jean Piaget a la teoría de sistemas complejos. Editorial Gedisa, Barcelona.

Garcia, R., 2006. Sistemas complejos. Conceptos, método y fundamentación epistemológica de la investigación interdisciplinaria. Editorial Gedisa, Barcelona.

Geyer, F. (Ed.), 1997. Sociocybernetics, complexity, dynamics and emergence in social science. Special Issue Kybernetes. The International Journal of Systems and Cybernetics 26 (6/7).

Geyer, F., van Dijkum, C. (Eds.), 1999. Newsletter 7. ISA - International Sociological Association. Research Committee 51 on Sociocybernetics (RC 51), vol. 4, no. 1. http://www.unizar.es/sociocybernetics.

Geyer, F., 1980. Alienation Theory. Pergamon Press, Oxford, New York.

Geyer, F., van der Zouwen, J. (Eds.), 1986. Sociocybernetic Paradoxes. Sage Publications, London, Beverly Hills, New Delhi.

Geyer, F., van der Zouwen, J. (Eds.), 1978. Sociocybernetics. Martinus Nijhoff, London.

Hornung, B.R., 1995. Sociocultural Evolution. Towards the Merging of Material and Informational Evolution. In: Association Internationale de Cybernétique (AIC) (Ed.), 14th International Congress on Cybernetics, Namur (Belgium), 21–25 August 1995. Proceedings. AIC, Namur, pp. 867–872.

Hornung, B.R., 2002. Minimal Conceptual Modelling (MINCOMOD). From Theory of Society to IT-Systems in Hospitals. International Review of Sociology 12 (2), 269–281.

Hornung, B.R., 2004. Sociocybernetic Problem-Functionalist Analysis of Community and Regional Development. NG -Nase Gospodarstvo/Our Economy. Review of Current Issues in Economics 50 (3-4), 17–23.

Hornung, B.R., 2006a. El paradigma sociocibernético. Conceptos para la investigación de sistemas sociales complejos, In: Marcuello Servós, Ch (Ed.), Sociocibernética. Lineamientos de un paradigma. Institución "Fernando el Católico" (CSIC), Zaragoza, pp. 41–79.

Hornung, B.R., 2006b. From Cultural Relativism to the Unity of Science by Means of Epistemological Constructivism. Foundations for Sociocybernetics. In: Marcuello

Servós, Ch, Fandos, J.L. (Eds.), Sociological Essays for a Global Society. Prensas Universitarias de Zaragoza, Zaragoza, pp. 81–120.
Hornung, B.R., Adilova, F.T., 1997. Conceptual modelling for technology assessment of it systems, smart cards and health information systems. Kybernetes 26 (6/7), 787–800 (special issue).
Huntington, S.P., 2002. The Clash of Civilizations and the Remaking of World Order. Simon and Schuster, UK.
Kjellman, A., 2002. The subject-oriented approach to knowledge and the role of human consciousness. International Review of Sociology 12 (2), 223–247.
Kjellman, A., 2003. Constructive Systems Science - the Only Remaining Alternative?. Report Series, no. 03–014, Stockholm University/Royal Institute of Technology, Stockholm.
Kuhn, A., 1974. The Logic of Social Systems. Jossey-Bass, San Francisco, CA.
Kuhn, Th S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Laszlo, E., 1973. Introduction to Systems Philosophy. Harper & Row, New York, London.
Luhmann, N., 1995. Social Systems. Stanford University Press, Stanford.
Marcuello Servós, Ch (Ed.), 2006. Sociocibernética. Lineamientos de un paradigma. Institución "Fernando el Católico" (CSIC), Zaragoza.
Maturana, H.R., Varela, F.J., 1980. Autopoiesis and Cognition. The Realization of the Living. D. Reidel Publishing Company, Dordrecht, Boston.
Meadows, D., Robinson, J.M., 1985. The Electronic Oracle. Computer Models and Social Decisions. John Wiley & Sons, Chichester West Sussex, London, New York, Singapore.
Meadows, D., et al., 1972. Limits to Growth. Signet, New York.
Mesarovic, M.D., Pestel, E., 1975. Mankind at the Turning Point. The Second Report to the Club of Rome. Hutchinson, London.
Miller, J.G., 1978. Living Systems. McGraw Hill, New York.
Müller, A., Müller, K.H. (Eds.), 2007. An Unfinished Revolution? Heinz von Foerster and the Biological Computer Laboratory/BCL 1958–1976. Edition Echoraum, Wien.
Parra-Luna, F. (Ed.), 2000. The Performance of Social Systems. Kluwer Academic Publishers/Plenum Publishers, New York, Boston, Dordrecht, London, Moscow.
Parsons, T., 1966. The Social System. The Free Press, Collier-Macmillan, London, New York.
Parsons, T., 1967. Sociological Theory and Modern Society. The Free Press, Collier-Macmillan, London, New York.
Parsons, T., 1971. The System of Modern Societies. Prentice Hall, Englewood Cliffs, NJ.
Pask, G., 1975. Conversation, Cognition, and Learning. Elsevier Publishers, Amsterdam, New York.
Prigogine, I., Stengers, I., 1984. Order Out of Chaos – Man's New Dialogue with Nature. Flamingo, London.
Scott, B. (Ed.), 2011. Gordon Pask. The Cybernetics of Self-organization, Learning and Evolution (Papers 1960–1972). Edition Echoraum, Wien.
Stonier, T., 1990. Information and the Internal Structure of the Universe. An Exploration into Information Physics. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.
Stonier, T., 1997. Information and Meaning. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.
Wiener, N., 1956. The Human Use of Human Beings – Cybernetics and Society. Doubleday, Garden City, NY.
Wiener, N., 1948. Cybernetics, or Control and Communication in the Animal and the Machine. MIT Press, Cambridge, MA.
van der Zouwen, J., 2006. Sociocybernetics and testability. A bridge too far? Festschrift for felix geyer. Kybernetes 35 (3/4), 426–432.

Sociodigital Revolution: Digital Natives vs Digital Immigrants

Kai Hakkarainen, University of Turku, Turku, Finland
Lauri Hietajärvi, University of Helsinki, Helsinki, Finland
Kimmo Alho, Helsinki Collegium for Advanced Studies and Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland
Kirsti Lonka, University of Helsinki, Helsinki, Finland
Katariina Salmela-Aro, Cicero Learning, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to review what engages *digital natives*, i.e., children and adolescents who have, from the beginning of their lives, been socialized to use sociodigital technologies. Surveying the research literature, we present findings as to what pursuits digital natives find interesting, motivating, and involving in both informal and formal learning environments drawing upon motivation theories of flow and engagement. Our conclusion, in this article, is that adolescents' ways of engaging in using digital technologies are heterogeneous; a minority of young persons has access to parental or peer support and facilitation that engagement in creative use of digital technologies together with their own motivation and efforts may require. Although sociodigital technologies facilitate creating and maintaining extended networks, cultivating technological fluency, and participating in passionate interest communities and networks, there are worrisome trends; these include 'addictive' use of technology, fragmented processing of information, and 'digital divides' between creative and educational use of sociodigital technologies.

Introduction

The purpose of this article is to review the pursuits that engage digital natives intellectually, emotionally, and socially (Prensky, 2001, 2012; Palfrey and Gasser, 2011). *Digital natives* are adolescents who have from the very beginning of their lives been socialized to intensive use of sociodigital technologies (i.e., integrated systems of technological tools, social media, and global information networks that support collaborative creation and sharing of activities, media, and knowledge). Prensky (2001) claimed that there is a profound discontinuity between engagement of the young people who have been immersed in sociodigital technologies from very early in their lives and the so-called *digital immigrants*, for instance, their parents and teachers, who have learned to use digital technologies in adulthood, if at all. At the age of 20, many Western adolescents have accumulated 10 000 h of computer or game console use, 10 000 h of mobile use and exchanges, and 200 000 received and sent e-mails and instant messages (Prensky, 2012: 78). Presumably because of extended socialization to use of sociodigital technologies, there appears to be a special relation between young people and various digital technologies that differentiates them from the older generations. Accordingly, adolescents are often very comfortable with various sociodigital tools and applications and able to fluently learn novel applications. They are thought to be used to more extensive multitasking and to be more effective in task switching than their parents and teachers while handling the intensive digital information and media flow (Palfrey and Gasser, 2011; Prensky, 2001). This possible generational gap should, however, be studied properly in controlled experimental setups. In any case, whenever young people have a problem, they seek help via the Internet whereas older participants still more often consult books and go to libraries. Rather than relying on traditional physical books and handwriting, digital adolescents constantly process virtual information on the screen. Being online is an integrated part of their daily life and they do not appear to experience a similar difference between online and off-line realities than older generations do (Ito et al., 2010). Further, at least a half of the young people are estimated to have taken part in such creative processes as creating media content and one-third of them have shared some personally produced content through the Internet (Jenkins et al., 2009; Lenhart and Madden, 2005).

According to Thomas (2011: 4), the discourse on digital natives typically involves three interrelated assumptions: Digital natives constitute a largely homogeneous generation and speak a different language vis-à-vis digital technologies from that of their parents, the digital immigrants; they learn differently from preceding generations of students; and they demand new ways of teaching and learning regarding technology. A number of investigators, however, have questioned one or more of these assumptions; indeed, the whole notion of digital natives has been said to be problematic; some have argued that the digital competencies of the young generations have been seriously misjudged (see e.g., Bennett et al., 2008). Most young people use sociodigital technologies only for everyday recreational purposes and do not have any of the claimed sophisticated competencies of alleged 'digital natives.' In many cases, differences between young people's digital competencies appear to be larger than differences between the younger and older generations (Jenkins et al., 2009; creative participation gap). There are considerably many representatives of older generations who have acquired high-level digital competencies (Bennett et al., 2008). Moreover, many educational researchers (see e.g., Jenkins et al., 2009) are more concerned about the differences among the digital natives and the underlying individual and contextual factors behind them. It is arguable that, so far, the discourse of digital natives has not been grounded on well-established empirical evidence (Bennett et al., 2008);

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26094-7

instead, one sees numerous problematic claims that have not been empirically verified.

We argue that the concept of the *digital native*, although controversial, is a useful working hypothesis for addressing the profound impact of sociodigital engagement on intellectual and socioemotional activities of young people. The evolutionary significance of the extended human childhood and adolescence is in allowing new generations of humans to adapt to a radically different environment from that of the older generations (Wexler, 2006). The earlier and the more intensively one adapts to the cognitive, social, and cultural 'prostheses' (Clark, 2003; Donald, 1991; Ritella and Hakkarainen, 2012) provided by the use of continuously present sociodigital technology, the stronger their impact on a young person's intellectual, emotional, behavioral, and social engagement is likely to be. To acknowledge developmental effects, we do not have to assume that adolescents' engagement in using sociodigital technologies is uniform; our approach is to examine, with a person-oriented approach, how various populations of youth use sociodigital technology and their associated motivations. Although the concept of the digital native is in many ways problematic, it may be taken as a provisional tool for addressing various issues that appear relevant to understanding adolescents' learning and engagement.

The purpose of this article is to review the literature concerning digital natives' engagement in their chosen pursuits, within the framework of theories of flow (Csikszentmihalyi, 1996; see chapter by Nakamura & Dubin) and academic emotions (see chapter by Pekrun; also Inkinen et al., 2014). In the first section we present findings on adolescents' engagement in using sociodigital technologies to 'hang out' with their friends. The second section focuses on findings concerning adolescent's engagement in using sociodigital technologies to pursue their interests. Section three compares the digital divides between adolescents who engage only in shallow use of sociodigital technologies and those who have access to instruments, practices, and support that creative engagement in using digital technologies requires. Finally, we will address limitations of our current understanding of digital natives as well as outline fruitful lines of deepening inquiry in the area.

How Hanging Out with Peers Engages Adolescents

Rather than assuming uniform generational differences, investigators would do well to examine patterns of engagement of various populations of adolescents in using sociodigital technologies (Helsper and Enyon, 2011; Palfrey and Gasser, 2011); we maintain that a person-oriented approach (Laursen and Hoff, 2006) is fruitful. Ito et al. (2010) carried out a year-long ethnographic investigation of digital youth's engagement in using sociodigital technologies, distinguishing three levels of engaging in technology-mediated activity, i.e., (1) friendship-driven use of sociodigital technologies for 'hanging out' with peers, (2) interest-driven 'messing around' with technology and media that is oriented toward developing associated skills and competencies, and (3) 'geeking out' for seriously cultivating expertise related to digital technologies or creative working with media. The proposed notion of adolescents being driven by either friendship- or interest-related aspects resonates strongly with the results of recent studies (Eynon and Malmberg, 2011; Kennedy et al., 2010) that have addressed the variation in adolescents' sociodigital activities. These investigations, despite some diversity in the methods used in them, share the key finding that the largest group of adolescents engages mostly in friendship-driven activities, and only a small minority participates frequently in more demanding, interest-driven activities.

More than 90% of adolescents in the Western countries have access to the Internet and recent study showed that 63% of them use sociodigital technologies once or several times a day for connecting with their friends and family members (Madden et al., 2013). They use mobile devices and social media applications to constantly keep up networking connections with their friends and to strengthen their face-to-face relations and sense of belonging with their peers (Ito et al., 2010). Three-quarters of adolescents are mobile Internet users in terms of using smart phones, tablets, or other mobile devices (Lenhart et al., 2010; Madden et al., 2013). This allows active participants to be in continuous connection with their friends through constant instant messaging, and lurking and commenting on friends' activities within social media. Virtual communication allows them to be together even when separated from one another temporally (e.g., evenings) or spatially (e.g., being out of their neighborhood when visiting relatives). Besides lurking and updating one's own status the friendship-driven activities rely on giving constant feedback (e.g., 'likes') to each other's updates and expecting social recognition in return. Sociodigital networks constitute an essential aspect of the social operating system of adolescents (Rainie and Wellman, 2012) determining the modality of connection, the nature of information exchanged, ways of collaborating, and so on. A decade ago, only a quarter of Finnish upper elementary school students reported using computers daily (Hakkarainen et al., 2000). Currently, many adolescents report having digital contact with their friends several times a day or having a continuous connection (e.g., 61.2% of elementary school sixth graders report *chatting* at least daily (36.6% several times a day or 'all the time') and 48.8% report using *social media services* at least daily (23.8% several times a day or 'all the time')). (The situation of young people's practices of using sociodigital technologies is rapidly changing. The numbers presented in this paragraph rely on data collected in 2013 from Finnish sixth grades across 33 schools of Helsinki, Finland.)

Beyond the intensity of keeping up networking contacts, the nature of adolescents' social networks is also changing (Gee and Hayes, 2011; Ito et al., 2010; Rainie and Wellman, 2012). Various applications of social media presently allow young people to keep up more extended and far-reaching networks than used to be the case. Adolescents who move from one school, neighborhood, city, or country to another, are able to keep up their friendship relations to an extent that was not possible before. Many adolescents have a very extensive network of 'friends' (these are connected persons, including acquaintances; according to Facebook 2012 statistics, an average Facebook user has 130 Facebook 'friends'; according to Madden et al. (2013) teens have an average of 300 'friends'). Currently, however, young people appear to be

moving from just building huge networks of designated 'friends' toward maintaining their social networks and web privacy in a new way and separating their public networks from their private (e.g., Madden et al., 2013). Some say this is due to the increasing presence of adults and family members in e.g., Facebook and the related difficulties of controlling their privacy.

Being in a constant networking connection is likely to make socioemotional regulation of activity become socially distributed. Many aspects of adolescents' lives are being collectivized by moving to the virtual network, including creation, keeping up, and breaking intimate relations (Ito et al., 2010). Young people often meet their intimate partners through the 'net' and publish their relationship status in social media. Beyond the sharing of positive emotions with peers, also negative emotions and frustration may spread through social networking connections. Simultaneously with providing a sense of relatedness, social media could make a young person to become obsessed with self-presentation in social media, compulsively following status updates of his or her contacts and managing impressions (Nadkarni and Hofmann, 2012). If participation in the virtual world replaces building of social relations and limits interpersonal encounters and interaction, it may lead to reduced well-being, disengagement, increased anxiety, and depression. The emergence of social media has made it clear that rather than leading to social isolation, intensive use of sociodigital technologies is usually embedded in a rich network of both online and off-line social interaction (Rainie and Wellman, 2012). The sociodigital technologies are, however, transforming so quickly that no one has clear understanding of the psychosocial impact and consequences of adolescents' engagement.

Interest-Driven Engagement with Sociodigital Technologies

When adolescents use sociodigital technologies to pursue their interests, they are 'messing around' or 'geeking out' with technology (Ito et al., 2010). The former is motivated by sharing interests, whereas the latter one involves enhancing competence and cultivating expertise; this has made some investigators enthusiastic about the positive impact of sociodigital technologies on adolescents' learning and development (Barron, 2006; Gee and Hayes, 2011; Ritella and Hakkarainen, 2012). Adolescents who become engaged in technology-mediated activities often start experimenting, with various technological tools and applications relevant to their interests. This 'messing around' represents, in turn, a transition zone between friendship-driven 'hanging out' and interest-driven 'geeking out' (Ito et al., 2010). It involves intensive engagement with digital technologies for sharing interests related to sociodigital technologies, for example, playing computer games or sharing and producing media. Different levels of digital participation may be distinguished, ranging from mere observation and follow-up to personal participation in creating and building media (Gee and Hayes, 2011; Ito et al., 2010; Jenkins et al., 2009).

Some adolescent are 'geeking out' (Ito et al., 2010) in terms of diving very deep into the sociodigital world and becoming experts in using sociodigital technologies, often being highly regarded among adolescents (Hakkarainen et al., 2000). They have appropriated sociodigital tools that allow, for example, modifying, remixing, and mashing up artifacts created by selecting, cutting, enhancing, filtering, and transforming media elements (Gee and Hayes, 2011; Ito et al., 2010); this requires new far-reaching network connections with persons who share the same interest. 'Geeking out' may thus involve a trajectory of actively cultivating technological competencies (Gee and Hayes, 2011), for instance, in terms of creating media objects (videos, pictures, music), modifying virtual environments ('modding' game environments), programming (creating new applications), or constructing technological systems (building robots).

Expert knowledge is now accessible to Internet users to an unforeseen extent, without requiring formal credentials (Gee and Hayes, 2011; Rheingold, 2012). Many adolescents cultivate their expertise by meeting more competent peers as well as with adult experts interested in similar issues on the Internet (Shirky, 2010; Rheingold, 2012). They may start functioning as 'passionate interest groups' (Gee and Hayes, 2011) with like-minded participants sharing knowledge and competence. Such learning is likely to emerge when adolescents "organize themselves in the real word and/or via Internet (or a virtual world) to learn something connected to a shared endeavor, interest, or passion" (Gee and Hayes, 2011: 69). Although everybody is allowed to participate in creating, building, and advancing the shared interest, more shallow participation is also acceptable. Social recognition of participants' accomplishments is determined by the depth of their participation rather than directly reflective of their age or social position. Interest groups have similar kinds of tensions as do other human communities, such as social conflicts, occasional 'flaming,' or 'hazing' of newcomers (Gee and Hayes, 2011: 71).

Although 'geeking out' involves positive developmental possibilities, there are also dangers involved. Middle adolescence is argued to be a period of especially heightened vulnerability to risky behavior (Steinberg) because sensation-seeking is high and self-regulation functions are still immature. In some cases, passions related to use of sociodigital technologies are harmonious in nature (Vallerand et al., 2007) in terms of being well-integrated with other life activities and corresponding to the participants' core values and self-concepts. Observations indicate, however, that many adolescents use sociodigital technologies excessively, find it hard to interrupt their activity, and pursue technology use until late in the night. When passions become obsessive in nature, the participants have a hard time controlling them and there emerge various conflicts with neglected areas of personal, social, and educational activities. In some cases, constant use of the sociodigital technologies becomes deleterious and leads to weakening of real-world networking relations, and, thereby, may hinder psychosocial development (Reich et al., 2012; Smahel et al., 2012). Engagement in using sociodigital technologies appears, consequently, to be a double-edged sword because of the benefit and possible harms of high-intensity participation.

Digital Divide: Creative Participation Gap

The above examination indicated that adolescents' engagement in using sociodigital technologies is likely to be extremely heterogeneous. Most adolescents use digital technologies for entertainment, consuming media, and other shallow purposes rather than engage in building their digital skills, creatively working with media, or adapting sociodigital technologies to support academic pursuits (Bennett et al., 2008; Eynon and Malmberg, 2011; Kennedy et al., 2010). Investigators, moreover, have expressed concerns about an emerging *Creative Participation Gap* (Jenkins et al., 2009), i.e., the unequal access to learning opportunities and formative experiences of mastering creative use sociodigital technologies; this is caused, for example, by social class, gender, or socioeconomic hardship. Adolescents coming from advantaged and disadvantaged homes have unequal opportunities of participating in sophisticated and creative use of sociodigital technologies. Because of socioeconomic and cultural factors, adolescents coming from less privileged families tend to have weaker possibilities for accessing sophisticated sociodigital technologies (programmable computers and relevant applications beyond mere game consoles) that would allow them to go beyond recreational use of sociodigital technologies and creatively engage with them (Ito et al., 2010; Eynon and Malmberg, 2011). Because of cultural stereotypes, the development of sophisticated skills of using sociodigital technologies is likely to be facilitated in the cases of male rather than female adolescents (Barron, 2004).

Adolescents need role models and concrete support for starting to use sociodigital technologies for creative purposes; in this regard, their parents are likely to have a very crucial role. Many adolescents do not have encouragement and parental guidance needed for taking part in pursuing challenging technology-mediated interests (Barron, 2004; Jenkins et al., 2009; Palfrey and Gasser, 2011). Simultaneously, however, an adolescent's own interests, motivation, and sustained efforts play a crucial role affecting the level and impact of parental support. Adolescents who are 'geeking' out may receive their initial inspiration from home, school, or the Internet and, subsequently, distribute their learning across many fields of activity (Barron, 2006). Participants may start sharing their interests or passions with their peers, join Internet-based interest groups, cultivate their interest in entrepreneurial activities (including summer jobs and internships), or utilize Internet-based information sources and network communities. In order to start cultivating sophisticated digital competencies, an adolescent needs to have someone who recognizes his or her interests, and spurs the development of his or her capabilities, in interaction with the adolescent's own agency and effort. From a developmental perspective, social recognition of one's human strengths and contributions to shared efforts is vital (Kindermann, 2007, Larson et al., 2004); digital environments are often very generous in socially recognizing one's achievements and efforts (Shirky, 2010). A special advantage of sociodigital technologies is that passive observation may more or less quickly phase-transform to active engagement in sharing one's experiences and contributions.

An essential aspect of the digital divide is the extent to which sociodigital technologies are used to support academic studies. Although young people may themselves be able to engage in using computers and the Internet to support their social activity, they are much more dependent on parental and teacher guidance when using sociodigital technologies to facilitate academic learning. Heikkilä and Lonka (2006) showed that even many first-year university students appear to have difficulties in self-regulatory skills. They are anxious about not being capable of systematic learning across subject domains beyond their immediate interests. It is crucial to learn to productive use Internet-based knowledge resources as well as various productivity tools, such as text processing, image processing, and presentation software as personal instruments for pursuing various aspects of academic studies. Educational use of sociodigital technologies is still concentrated among students coming from socioeconomically well-to-do families and families with good academic background. Systematic educational support is needed because students coming from disadvantaged homes are not likely to spontaneously cultivate associated skills and competencies.

The investigation of Laursen et al. (2010) indicates that it matters with whom one is 'hanging out.' 'Hanging out' with adolescents oriented toward school and academic activity can provide a protective belt, especially in the case of females, whereas hanging out with adolescents engaged in problem behaviors, such as drinking, is a risk (Kiuru et al., 2009). Networks may lead to either a positive (orientation toward school and academic development) or a negative (exclusion and reification from school) trajectory. Learning is facilitated by hanging out with other adolescents oriented toward studying; this is likely to reduce negative effects of risk behavior. More or less continuous real-time presence of peers when encountering challenges and frustration may make the adolescents overdependent on their peers' support. Adolescents' need, however, to have an ability to invest sustained intellectual efforts in pursuing their school studies, in spite of being temporarily isolated from their peers (Csikszentmihalyi et al., 1996).

Discussion

This article has presented evidence from research literature which indicates that adolescents' ways of using sociodigital technologies are very heterogeneous; although most Western adolescents intensively use sociodigital technologies, only some of them are adept technology users. Thus, we researchers should focus on studying *what* the engagements are of specific subgroups of digital natives rather than consider them as a homogeneous generational population. Social media have affected most strongly the second-generation digital natives born in 1995 or later; these cohorts of adolescents appear to use sociodigital technologies more intensively than earlier ones. Findings of Ito et al. (2010) indicate that most adolescents use sociodigital technologies for friendship-driven practices of 'hanging out,' in terms of being in a constant connection with their extended network of friends by using phone calls, instant messages, text

messages, and social networking services. Such hyperintensive interaction makes social networks a part of adolescents' social operating system (Rainie and Wellman, 2012) and enables them to coregulate their motivation an unforeseen extent. The sociodigital technologies provide affordances for sharing engagement and frustrations and appear to offer significant support to adolescents' activity and development. Adolescents who are 'hanging out' with peers oriented toward school and solving age-relevant developmental challenges have an advantage over those spending time with peers who exhibit various problem behaviors. Although social sharing of one's socioemotional experiences may be developmentally beneficial, adolescents also have to be able to show a certain level of emotional independence that concentrated academic efforts call for (Csikszentmihalyi et al., 1996).

We started this article by discussing the discontinuity between digital natives and digital immigrants, highlighting the distinctive orientation of the young generation to socially share their engagement. Digital adolescents appear to be motivated by activities that provide experiences of autonomy, a sense of competence, and a sense of belonging similar to those that previous adolescents have found outside of digital contexts (Deci and Ryan, 2004). A great deal of present-day creative activity takes place through the Internet, which allows peers to peripherally observe others' activities, to comment on and discuss them as well as to obtain social recognition. Such activity relies on an extended social network and provides multiple possibilities for engaging in more serious 'messing around' with sociodigital technologies, for instance, in making, sharing, and remixing media. Even if only a small percentage of adolescents take very active part in creative activities facilitated by sociodigital technologies, the sociodigital technologies provide affordances for persons to engage in such activities with a very low threshold of effort. Although sociodigital technologies provide all adolescents possibilities to pursue their interests, they are dependent on support of parents, teachers, and other adults (Barron, 2006). Adolescents coming from disadvantaged homes may not have access to sophisticated technology and parental facilitation of creative activity and guidance for learning how to use sociodigital technologies to support school work (Jenkins et al., 2009).

In spite of the radical transformation of adolescents' personal, social, and cultural environments by the sociodigital revolution, there is a lack of systematic longitudinal studies concerning the impact of sociodigital technologies on motivation, learning, and development. In order to overcome such limitations, the present authors' *Mind the Gap* project (www.wiredminds.fi) integrates research on educational psychology, developmental science, technology-mediated learning, and neuroscience to investigate the development of minds of digital natives at multiple levels (personal, family, peer, school class, school, neighborhood). It is possible that the digital natives' extensive technology use, including demands for multitasking and frequent task switching, improves their executive cognitive skills and the related brain functions (cf Patel et al., 2013). However, this issue has not yet been properly studied. In future studies, controlled experimental setups and evolved brain imaging techniques, such as functional magnetic resonance imaging are needed. These studies should be done urgently because it will be soon difficult to find participants for control groups who have still not had extensive exposure to sociodigital technologies.

See also: Computer Games and Academic Achievement; Computer Mediated Communication; Internet and Culture; Internet and Privacy; Internet and Social Media: Anthropological Aspects; New Media and the Digital Divide; Oversharing: The Eclipse of Privacy in the Internet Age; Sexuality and the Internet; Social Media; Teens, Gender, and Self-Presentation in Social Media.

Bibliography

Barron, B., 2004. Learning ecologies for technological fluency: gender and experience differences. Journal of Educational Computing Research 31, 1–36.

Barron, B., 2006. Interest and self-sustained learning as catalyst of development: a learning ecology perspective. Human Development 49, 193–224.

Bennett, S., Maton, K., Kervin, L., 2008. The 'digital natives' debate: a critical review of the evidence. British Journal of Educational Technology 39, 775–786.

Clark, A., 2003. Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press, Oxford.

Csikszentmihalyi, M., 1996. Creativity: Flow and the Psychology of Discovery and Invention. HarperCollins, New York.

Csikszentmihalyi, M., Rathunde, K., Whalen, S., 1996. Talented Teenagers: The Roots of Success and Failure. Cambridge University Press, Cambridge, MA.

Deci, E.L., Ryan, R.M., 2004. Handbook of Self-determination Research. The University of Rochester Press, Rochester, NY.

Donald, M., 1991. Origins of the Modern Mind. Harvard University Press, Cambridge, MA.

Eynon, R., Malmberg, L., 2011. A typology of young people's internet use: implications for education. Computers and Education 56, 585–595.

Gee, J.P., Hayes, E.R., 2011. Language and Learning in the Digital Age. Routledge, London.

Hakkarainen, K., Ilomäki, L., Lipponen, L., Muukkonen, H., Rahikainen, M., Tuominen, T., Lakkala, M., Lehtinen, E., 2000. Students' skills and practices of using ICT: results of a national assessment in Finland. Computers & Education 34, 103–117.

Heikkilä, A., Lonka, K., 2006. Studying in higher education: students' approaches to learning, self-regulation, and cognitive strategies. Studies in Higher Education 31, 99–117.

Helsper, E., Enyon, R., 2011. Digital natives: where is the evidence? British Educational Research Journal 36, 503–520.

Inkinen, M., Lonka, K., Hakkarainen K., Muukkonen, H., Litmanen, T., Salmela-Aro, K., 2014. The interface between core affects and the challenge-skill relationship. Journal of Happiness Studies. doi:10.1007/s10902-013-9455-6, 15, 891-913.

Ito, M., Baumer, S., Bittandi, M., Boud, D., Cody, R., et al., 2010. Hanging Out, Messing Around, and Geeking Out. The MIT Press, Cambridge, MA.

Jenkins, H., Clinton, K., Purushotma, R., Robison, A., Wiegel, M., 2009. Confronting the Challenges of Participatory Culture: Media Education for 21st Century. MacArthur Foundation. Retrieved 28.08.12 from www.newmedialiteracies.org/files/working/NMLWhitePaper.pdf.

Kennedy, G., Judd, T., Dalgarno, B., Waycott, J., 2010. Beyond natives and immigrants: exploring types of net generation students. Journal of Computer Assisted Learning 26, 332–343.

Kindermann, T.A., 2007. Effects of naturally-existing peer groups on changes in academic engagement in a cohort of sixth graders. Child Development 78, 1186–1203.

Kiuru, N., Nurmi, J.-E., Aunola, K., Salmela-Aro, K., 2009. Peer group homogeneity in adolescents' school adjustment varies according to peer group type and gender. International Journal of Behavioral Development 33, 65–76. http://dx.doi.org/10.1177/0165025408098014.

Larson, R., Eccles, J., Gootman, J.A., 2004. Features of positive developmental setting. The Prevention Researchers 11, 8–13.

Laursen, B.P., Hoff, E., 2006. Person-oriented and variable-oriented approaches to longitudinal data. Merrill-Palmer Quarterly 52, 377–389.

Laursen, B., Bukowski, W.M., Nurmi, J.-E., Marion, D., Salmela-Aro, K., Kiuru, N., 2010. Opposites detract: middle school peer group antipathies. Journal of Experimental Child Psychology 106, 240–256.

Lenhart, A., Madden, M., November 2, 2005. Teen Content Creators and Consumers. Pew Internet & American Life Project, Washington DC. Available online at: http://www.pewInternet.org/PPF/r/166/report_display.asp.

Lenhart, A., Purcell, K., Smith, A., Zickuhr, K., 2010. Social Media and Mobile Internet Use among Teens and Young Adults. Pew Internet & American Life Project. Available at: http://web.pewinternet.org/~/media/Files/Reports/2010/PIP_Social_Media_and_Young_Adults_Report_Final_with_toplines.pdf.

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., Beaton, M., 2013. Teens, Social Media, and Privacy. Pew Internet and American Life Project Report. Available at: http://www.pewinternet.org/~/media/Files/Reports/2013/PIP_TeensSocialMediaandPrivacy_ FINAL.pdf.

Nadkarni, A., Hofmann, S.G., 2012. Why do people use Facebook? Personality and Individual Differences 52, 243–249.

Palfrey, J., Gasser, U., 2011. Reclaiming an awkward term: what we might learn from "digital natives". In: Thomas, M. (Ed.), Deconstructing Digital Natives. Routledge, London, UK, pp. 186–204.

Patel, R., Spreng, R.N., Turner, G.R., 2013. Functional brain changes following cognitive and motor skills training: a quantitative meta-analysis. Neurorehabilitation and Neural Repair 27, 187–199.

Prensky, M., 2001. Digital natives, digital immigrants. On the New Horizon 9 (5), 1–6.

Prensky, M., 2012. From Digital Natives to Digital Wisdom: Hopeful Essays from 21st Century Learning. Sage (Corwin), Thousand Oaks, CA.

Rainie, L., Wellman, B., 2012. Networked: The New Social Operating System. The MIT Press, Cambridge, MA.

Reich, S.M., Subrahmayan, K., Espinosa, G., 2012. Friending, iming, and hanging out face-to-face: Overlap in adolescents' online and offline social networks. Developmental Psychology 48, 356–368.

Rheingold, H., 2012. Net Smart. The MIT Press, Cambridge, MA.

Ritella, G., Hakkarainen, K., 2012. Instrument genesis in technology mediated learning. International Journal of Computer-Supported Collaborative Learning 7, 239–258.

Shirky, C., 2010. Cognitive Surplus. Penguin, New York, NY.

Smahel, D., Brown, B.B., Blinka, L., 2012. Associations between online friendship and Internet addiction among adolescents and emerging adults. Developmental Psychology 48, 381–388.

Thomas, M., 2011. Technology, education, and the discourse of digital natives. In: Thomas, M. (Ed.), Deconstructing Digital Natives. Routledge, London, UK, pp. 1–11.

Vallerand, R.J., Salvy, S.-J., Mageau, G.A., Elliot, A.J., Denis, P.L., Grouzet, F.M., Blanchard, C., 2007. On the role of passion in performance. Journal of Personality 75, 505–533.

Wexler, B., 2006. Brain and Culture. The MIT Press, Cambridge: MA.

Socioeconomic Status and Academic Achievement

Kevin KH Chung, The Hong Kong Institute of Education, Tai Po, New Territories, Hong Kong SAR China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social economic status (SES), which is often measured as a combination of family income, parental education, and occupational status, has been identified as an important predictor of children's academic success in school. Even before entering school, children's academic performance can to some extent be predicted based on characteristics of their family SES backgrounds. Low-SES families are categorized as having low income which is often associated with the low educational standards of the parents, many of whom live on welfare or minimal wage. Such families are less likely to access social connections and educational resources at home and this may in turn place their children at risk for different academic problems. Initial academic skills seem to correlate with the home environment, where an environment of low literacy and numeracy possibly combined with health problems may negatively affect the development of children's academic skills. The fact that low SES families live together in communities further exacerbates the problem as schools in these communities tend to be under-resourced with high drop-out rates and generally less adequate educational standards. Thus not only do low SES children start school with academic disadvantages, the gap between them and middle SES children widens with time. This article reviews recent research examining the association between SES, child health, family characteristics, school environment, cognitive development, literacy and numeracy achievement, and country characteristics and student achievement. It may shed further light on cognitive development and academic achievement of children with low SES backgrounds and provide grounds for further work on research and practice.

Overview

Two hundred million children or more under the age of 5 years from both developing and developed countries may not fully achieve their developmental and academic potential primarily due to low socioeconomic status (SES) or poverty (Grantham-McGregor et al., 2007). SES factors, which are often measured in terms of family income, parent education, and occupational status, have been identified as an important predictor of children's outcomes. Even before entering formal schooling, children's social cognitive development and academic performance can to some extent be predicted based on characteristics of their family SES backgrounds. The development and achievement gap between children growing up in low SES families and those from well-off families begins in infancy, and widens with age (e.g., Heckman, 2006; Sirin, 2005). There is growing evidence from the US, the UK, and other developed countries that low SES status is correlated with multiple risk factors resulting in inequalities in children's early development and educational attainment. These risk factors are linked with poor child health, poor maternal education, increased maternal stress, poor parent and school support, and inadequate stimulation at home and in school. All these factors further link to poor school readiness skills, including children's physical, social, and emotional well-being, approaches to learning, cognition, and literacy and numeracy acquisition. Not only do low SES children start school with cognitive and academic disadvantages, the gap between them and middle SES widens with time. This article will review recent studies examining the association between SES, child health, family characteristics, home and school environment, cognitive development, literacy and numeracy achievement, and country characteristics during the school year. Understanding of the multiple risk factors can inform priorities for intervention and practices to promote early child development and learning. Broad implications for policies and practices reducing risk factors associated with low SES will therefore be highlighted briefly.

Child Health

Child health has been seen as a fundamental influence on child development and school progress. SES has been shown to be associated with children's health status, with the disparity of health status between low and middle SES children beginning at birth and widening over time (Rothstein, 2004). Frequently, low SES is implicated in multiple diseases including respiratory illnesses that tend to be associated with an increased likelihood of dental caries, higher blood lead levels, iron deficiency, stunting, and sensory impairment (Cohen, 1999). These outcomes may be due to conditions linking with low SES such as inadequate nutrition, exposure to tobacco smoke, failure to get recommended immunizations, and inadequate access to health care. Compared with children from middle SES families, children from low SES families are more likely to be born prematurely, to be low in birth weight, to be born with asphyxia, a birth defect, fetal alcohol syndrome, or AIDS and to have learning-related disabilities because their parents tend to live or work in hazardous and unhealthy environmental conditions, and to smoke, to drink alcohol, to use drugs during pregnancy, and to exercise less (e.g., Bradley and Corwyn, 2002). Such behavioral differences may also influence stress reactions and social affiliations, which are linked with poorer physical and mental health, which may in turn make it harder for low SES parents to provide warm, responsive care or to monitor their children properly.

As low SES children grow up, they generally experience more health problems than do their more affluent counterparts. For example, premature children living in poverty for the first 3 years of life tend to manifest problems in growth, health

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92141-X

status, intelligence, and behavior (Bradley et al., 1994). Moreover, low SES children may suffer from a variety of infections and nutrition-related diseases such as heart and kidney disease, asthma, respiratory infections, ear infections with hearing loss, and a high incidence of injuries. Malnutrition has been found to be associated with poor cognitive development resulting in cognitive delays, low IQ, and problems with motor development (e.g., Anthony et al., 2011). This can in turn cause problems with the ability to concentrate, to process information, and to focus on academic work. Children from low SES families are also less likely to have health insurance and access to health care, thereby leading to a greater risk of illness (Bloom et al., 2006). Health problems put low SES children at a greater risk of school absence than their middle SES peers. Research finds that low SES children are almost three times less likely to have regular school attendance compared to their middle SES peers, and student absences are strongly correlated with poor academic achievement (e.g., Ready, 2010). Because low SES children do not attend school regularly, they receive fewer hours of classroom instruction and consequently perform less well academically.

Family Characteristics

Why is there this correlation between SES and family characteristics with social and academic outcomes? The most important child development and learning takes place in the early years when the family unit provides nurture, support, and stimulation. How effective these are is influenced by the parent characteristics and the resources that families have to devote to child-raising, including their parenting practice and style, and their ability to provide a rich, responsive, and safe learning environment. The widely accepted indicators of SES, namely family income and parental education, are key factors influencing children's academic and social outcomes (Bradley and Corwyn, 2002). For example, cumulated studies (e.g., Riley et al., 2014) have found that economic disadvantage is likely to reduce parents' ability to provide warm and sensitive parenting, and lessen the probability that children will have access to cognitively stimulating materials (e.g., toys, books) and socially enriching experiences (e.g., cultural activities). Findings of these studies provide some support to the family stress model (FSM) and family investment model (FIM). As suggested by the FSM, economic hardship may intensify family stresses and has an adverse effect on parents' emotions, behaviors, and relationships, which in turn negatively impact parenting and parent–child relationships (e.g., Conger et al., 2010). According to the FIM, families with low income tend to put most resources into their basic family needs (e.g., adequate food and housing) and are less likely to make investment in the development of their children (e.g., learning materials and stimulation). Conversely, families with high income are capable of investing more resource in their child development thus leading to the academic and social success of their children (Bradley and Corwyn, 2002). Hence, SES affects child development and academic achievement.

In addition to family income, parent education is another indicator of SES which, too, may affect the child's academic achievement (e.g., Conger et al., 2010; Hampden-Thompson, 2013). For example, less well-educated parents may place a lower value on education than those who have a higher level, therefore investing less resource, both financially and psychologically, in their children's education. Perhaps parents with low level of education tend to be less able to arrange the home environment in a way that is conducive to learning, e.g., ensuring a quiet area for study. Such conditions are likely to affect the child, causing the internalizing of problems such as aggressiveness, opposition, and hyperactivity (Huisman et al., 2010).

Research is suggesting that it is the mother whose educational levels are most influential in low SES families (e.g., Hampden-Thompson, 2013). For example, low maternal education levels are often correlated with low knowledge about child-rearing and child development, leading to less supportive mothering (e.g., Simpkins et al., 2006). Perhaps low academic achievement of the mother results in less well-paid work and so the mother is likely to be overworked in order to help the family income leading to less time to give the child. Frequently the mother does not know what is needed to supply the enrichment opportunities (e.g., providing musical instruments, purchasing books, arranging trips to educational and cultural events) that help cognitive and social development, nor has she the finances to supply these. However, the effects of paternal education levels on fathering have been less examined. One study has found that less-educated fathers tend to have less positive engagement with their children compared to their more educated counterparts (Blair et al., 1994). Others have found paternal education to levels have little influence on such involvement (e.g., Castillo et al., 2011).

When we look at the influence of mothers on girls and fathers on boys, however, the pattern changes. Recent evidence (e.g., Frederickson and Petrides, 2008) has found that gender differences in language and math learning performance increase with lessening levels of maternal education; boys seem to perform less well than girls. However, at the same time, having highly educated mothers may offer more than less educated to the increase in learning performance of boys than girls. Similarly, Cook et al. (1996) suggest that a highly educated father may serve as a role model for boys to identify with academically. Therefore, highly educated fathers also apply greater influence on their boys than on their girls, thus enhancing boys' learning performance. As suggested previously, highly educated parents may be better at providing resources to their children. Possibly boys tend to be more susceptible to disadvantaged environment and could benefit from better socioeconomic environment (e.g., Zambrana et al., 2012). It seems that parental education levels is a protective factor in boys' learning performance and academic achievement.

Family economy and parental education levels may have differential influences on family and child outcomes (Duncan and Magnuson, 2003), and act through different pathways. Whereas family income may have a general influence on children's mental health, parental education levels may affect children's mental health more directly from the effect of parenting. Huisman et al. (2010) have reported that whereas low parental income may influence children's mental problems with both internalized problems such as anxiety and depression and externalized problems such as aggressiveness,

opposition, and hyperactivity, low parental education levels are likely to affect children's internalizing problems only.

Parenting practice also matters. It is known that parents influence their children's literacy skills by reading to them, by using complex language, and by providing appropriate linguistic input and scaffolding. Low SES parents may have difficulties providing this support for various reasons including a lack of skills, lack of time, and lack of awareness of the importance of this sort of parental input (Cabell et al., 2013). As a result, children from low SES backgrounds tend to receive less parental linguistic input and linguistic scaffolding, they fail to receive appropriate linguistic and communicative mediation from their surroundings, they are less actively involved as conversational partners in family get-together, and they engage less in joint book-reading and interactive writing with their parents (e.g., Aram and Levin, 2002). Moreover, low SES parents are less likely to be involved in their children's homework, to participate in school functions, and to volunteer activities relative to their middle SES counterparts. Lack of these practices, which are common in higher SES families, has a negative impact on the development of the skills of the lower SES children. In addition to the parenting practice, parenting style may shape child development and academic attainment. For example, there is evidence that poor parenting manifesting itself in a lack of warmth and involvement, often coupled with harsh and erratic discipline, is associated with aggressiveness, hostility and opposition, and symptoms of anxiety and depression in children (Ma et al., 2012). These parental practices are particularly associated with low SES families and many more low SES children therefore manifest symptoms of psychiatric disturbance and maladaptive social functioning than their middle SES counterparts. These low SES children tend to show internalizing symptoms such as anxiety and depression, and externalizing symptoms such as aggressiveness, opposition, and hyperactivity compared to their middle SES peers (Starfield et al., 2002).

Further evidence of the effect of low SES can be found in the studies of single parent and divorced parents (e.g., Carlson and Corcoran, 2001). The stress of poverty may also increase the possibility of family breakdown, separation, or divorce, which themselves have a detrimental effect on children and their ability to learn. For instance, growing up in single-parent families with low SES has been associated with poor parenting style, thus adversely affecting child cognitive, social, and emotional well-being (East et al., 2006). Compared to two-parent families, children growing up with single-parent families are more likely to have lower scores on cognitive tests and school achievement (Carlson and Corcoran, 2001). Furthermore, single-parent families may experience more stress and maternal depression, three times the level found in two-parent families primarily due to financial strain, poverty, discrimination, and social exclusion. This may, in turn, affect child-rearing behavior and parenting style (Petterson and Albers, 2001). It is possible that stressed and depressed parents are less involved, less sensitive and responsive to their children (e.g., answering child's questions, talking to the child, praising the child), less supportive and warm, and more negative when interacting with their children, leading to increased stress and child abuse. Consequently, living in single-parent families may lead to negative impacts on their child development and academic achievement.

School Environment

The influence of SES on the school environment including child care and preschool may also have an effect on later academic performance. Children from low SES families habitually reside in poor neighborhoods characterized by high levels of unemployment and crime, absence of adult supervision of children, and lack of opportunities for building social networks (e.g., Brooks-Gunn and Duncan, 1997). Living in poor neighborhoods may therefore lead to decreased positive opportunities and is highly correlated with social and emotional problems. At the same time, the majority of schools located in low SES neighborhoods appear to be underresourced and have a school climate less supportive of academic performance as well as of teacher–student relations. Because teachers spend a great portion time dealing with students' discipline problems and welfare, they often have lower involvement in pastoral care and lower expectation of their students' academic attainment. Low SES schools may also have less qualified teachers, which in turn leads to low teacher quality and less rigorous curriculum. The combination of low expectations and less parental involvement in school activities results in a negative school climate and increased risk of school violence, which may weaken children's sense of security and their ability to achieve academic success in schools (Suarez-Orozco et al., 2008).

Cognitive Development

Children's cognitive development and SES are frequently found to be highly correlated (e.g., Duncan et al., 2011), with children from low SES families tending to perform worse on measures of cognitive ability such as attention, working memory, and inhibitory or cognitive control than their more affluent counterparts. Low SES children are also found to be 1.3 times more likely to experience cognitive developmental delays than middle SES children (Brooks-Gunn and Duncan, 1997). The reason advanced for this is again that most low SES children live in a cognitively impoverished environment and receive poor parental nurture. This lack of parental sensitivity and poor cognitive stimulation may have an adverse influence on children's normal brain development, particularly in the prefrontal cortex (PFC) brain region thus posing an increased risk for learning. The PFC is responsible for executive control processes i.e., decision-making, planning, attention, and so forth and has a prolonged period of postnatal development. Research based on electroencephalographic recordings of brain activity (or event-related potentials) are in line with those behavioral studies analyzing attention, inhibitory control, and working memory where SES has consistently been found to correlate with working memory and inhibitory control. A series of research studies conducted by Noble, Farah, and colleagues (e.g., Hackman, et al., 2010; Noble et al., 2007) has shown significant disparities between low SES children and middle SES children in attention, working memory, and inhibitory control all of which are associated with and reflect activities in the PFC region of the brain. In particular, children from low SES families tend to have poor working memory and attention, and less inhibitory control compared to their peers from middle SES families, and these results suggest that low SES children are

more likely to reduce or disrupt their PFC than their middle SES counterparts. Taken together, low SES children may develop different neural systems from middle SES children, affecting the ability to attend and remember relevant information, and filter irrelevant information, and thus leading to cognitive developmental delays and low academic performance.

Literacy Achievement

SES has been shown to be strongly associated with literacy, the building block for all other academic learning: a failure to acquire adequate literacy skills means failure in other academic pursuits. Literacy skills have been recognized as crucial precursors to children's academic achievement throughout their school years. It is important that emergent literacy skills be nurtured and taught before children enter their formal schooling. Research has continually demonstrated that print motivation (i.e., appreciation of stories and books), oral language, vocabulary, concepts of print or print awareness, letter knowledge (i.e., identifying and naming letters), phonological awareness (i.e., identifying and manipulating individual sounds of language), narrative skills (i.e., describing events and retelling stories), shaping and writing skills are the most robust predictors of later reading and writing competence (e.g., Gettinger and Stoiber, 2007).

Although the emergent literacy skills are so important for children's literacy attainment, many children from low SES families may not acquire and develop these essential skills fully, and consequently they enter formal schooling around 2 years behind their peers from middle SES families in literacy development (Clark and Kragler, 2005). Research evidence shows that poor emergent literacy skills are also associated with SES as being linked to poor health, home and school environment, parenting style, and cognitive development (e.g., Anthony et al., 2011). Perhaps not surprisingly, then, low SES children usually lag behind their middle SES counterparts, with delayed development in language skills such as oral language, vocabulary knowledge, letter knowledge, phonological awareness, morphological awareness (i.e., awareness of and access to the meaning of words) (e.g., Schiff and Ravid, 2011). Thus, they often enter school with weakness in these language skills, which has cascading consequences on the development of reading and writing skills. Particularly, children of low SES show weakness in vocabulary, reading accuracy, reading comprehension, spelling, and writing. As a consequence, low SES children with poor emergent literacy skills fall further behind their middle SES peers.

Numeracy Achievement

Like literacy, numeracy is a critical factor in children's academic achievement, and here, too, SES plays a part. Research studies indicate that children's early numeracy skills before and during preschool or kindergarten are significant predictors of mathematics achievement in primary school and beyond (e.g., Jordan and Levine, 2009). The early numeracy skills such as concepts of counting, number competence (e.g., the ability to recognize changes in magnitude), quantity discrimination, numerical relationships (e.g., recognizing which of two numbers is smaller), and number naming have consistently been found to be strong predictors of mathematics achievement (Jordan et al., 2009). These numeracy domains may be important contributors to the development of formal mathematics skills such as addition, subtraction, ordinal sequencing, and mathematical word problems. Prior to starting school, children may use many mathematics concepts and skills in their play and to meet the general demands of daily lives at home.

Children from low SES families may exhibit a low average level of numeracy achievement in preschool or kindergarten, and they continue to trail behind peers from middle SES families throughout schooling. This can be demonstrated by children's use of their fingers to represent and manipulate the quantities represented by number words. In early grades this use of fingers seems to assist children to link with the nonverbal representations and conventional representations with number words and this in turn may lead to mental calculation and fact mastery. However if such finger strategies are prolonged and continuous, the practice may delay development in number competence. It has been found that many low SES children are less likely to use their fingers to represent and manipulate the quantities represented with number words compared to the middle SES children (Jordan and Levine, 2009). Again this may be linked to the quality of parental involvement with the child's very early development in that mothers who have time often spend hours playing counting games on their child's fingers and toes.

Because children from low SES families may have less stimulating home environments and fewer resources for number-related learning, they are less likely to receive support for mathematical development and preparation for school than their middle SES counterparts. Parents of low SES backgrounds may engage in fewer informal mathematics activities and when they do they tend to focus more on simple topics (i.e., counting and shape recognition) than on complex processes (i.e., numerical and geometric reasoning). Further, compared to middle SES parents, low SES parents spend less time teaching number skills to their children and play less informal learning activities such as playing board games with their children because these skills and activities may improve children's understanding of number competence, numerical relationships, and problem-solving skills. Thus, low SES children are more likely to perform less well on tasks related to numerical reasoning and relationships, numerical operations, ordinal sequencing, and mathematics story problems than their middle SES peers. In fact, low SES children have four times the difficulty in solving story problems than the middle SES children (Jordan and Levine, 2009). They are less likely to achieve academic performance equivalent to their middle SES peers, and the gap between low and middle SES children may widen over the course of the school year.

Country Characteristics and Student Achievement

Although evidence has showed that economic inequality widens the literacy and numeracy achievement gap between the privileged and disadvantaged children, countries or territories with greater economic inequality also have lower student achievement overall (Chiu and Khoo, 2005). The Program for

International Student Assessment (PISA), which is to measure the performance of 15-year-olds in reading and mathematics, has generally showed low academic achievement in poor countries with lower gross domestic product per capita. Indeed, the PISA 2012 results show that students in lower SES countries (e.g., Colombia and Peru) tend to have lower academic achievement with reading and math (OECD, 2013). In particular, girls are likely to perform less well on mathematics than boys. Possibly poorer countries have fewer resources at their disposal to invest in education and are likely to allocate their resources and wealth unequally, thus increasing inequality of opportunity and individual outcomes (Chiu and Chow, 2014). In addition, these countries often have fewer public resources (e.g., public libraries and museums) and less teacher quality and education (e.g., qualified teaching staff and certified English teachers). Conversely, students in richer countries may available themselves of these opportunities (e.g., reading more library books) to learn further given that richer countries have more public resources (Chiu and Chow, 2014). As has been discussed, school environment also plays an important role in student achievement. In countries with a high proportion of low SES families and communities, the problem of an inadequate school environment becomes national rather than local with high teacher turnover and low quality, inadequate resources, and increased student dropout rates, all having significant effects on low student performance (e.g., Chiu, 2010). Thus, greater inequality in countries increases the likelihood of lower student achievement.

Conclusion and Implications

This article provides an insight into the multiple risk factors related to SES that influence children's school readiness and in turn impact on children's cognitive, literacy and numeracy development, and their physical well-being and emotional competencies. Risk factors including poor child health, poor maternal education, increased maternal stress, poor parent and school support, and inadequate stimulation in the home and school settings, can negatively affect development and achievement. Understanding the multiple factors is vital when designing effective, evidence-based intervention programs and ameliorative policy actions. Many policies and intervention programs could be used to help children from low SES families overcome the academic disadvantages. For example, policies providing cash supplements and housing assistance might reduce maternal depression and stress. These problems arising from poor health of SES children could be addressed both with food assistance programs designed to improve general nutrition and reduce nutrition-based illnesses and with introducing children's health insurance so that health-related absences could be minimized. Incentives to attract high-performing teachers and principals to schools, and to provide free lunch programs and free textbooks located in low SES communities could be used and money put in to increase family and school resources to enhance the development of low SES children.

Because of the primacy of parenting skills in early childhood development, programs to teach these skills could be devised and implemented. If parents have the ability and time to give their children the same enriched environment as is found in middle SES families, then an improvement in school readiness and academic achievement should result. But before such programs could be implemented low SES families need to have financial assistance so that resources and time become available. Early literacy and numeracy intervention programs are designed to improve the academic achievement of low SES children and help their parents to acquire literacy and numeracy skills contributing to children's learning experience. In addition, such programs may help children connect learning and achievement with positive social interactions and attention with their parents. Furthermore, given strong association between unequal resource distribution, physical resource, and academic achievement, government can provide low SES children with physical resources and improve student learning through equal distribution policies and economic growth. Finally it must be acknowledged that poverty is a world problem of enormous complexity. It affects the academic achievement of the children who are born into it so that they too leave school with inadequate life skills and in turn become low SES parents. It is hoped that this article analyses many of the underlying causes of this lack of achievement and thus suggests areas where intervention could help to break the cycle of poverty.

Summary of the impacts of socioeconomic status

Domains/areas	*Consequences/outcomes*
Health/lifestyles	● Birth defects ● Childhood illnesses ● Inadequate nutrition ● Poor home learning environment
Family characteristics	● Less educated and low income parents ● Single-parent families particularly single mother ● Poor parenting practices and styles ● Poor child cognitive, language and social skills, and emotional well-being
Emotional and behavioral development	● Emotional and behavioral problem ● Mental health problem ● Internalizing and/or externalizing symptoms and behaviors
Cognitive/academic attainment	● Developmental delay ● Grade repetition ● High school dropout ● Learning disability ● Poor working memory and attention ● Less inhibitory control
Home and school environment	● Culturally impoverished home environment ● Lower SES school communities ● Inadequate educational resources ● Poor child development and achievement
Literacy and numeracy achievement	● Poor literacy and numeracy skills for low SES children ● Both benefit from enriched early home and school environment

Continued

Summary of the impacts of socioeconomic status	
Domains/areas	*Consequences/outcomes*
Country characteristics and achievement	• Poorer countries • Inequality in wealth and resource distribution • Poorer communities magnified

Acknowledgments

This research was supported by the General Research Fund of the Hong Kong Special Administrative Region Research Grants Council (HKIED: GRF 841212 and 840308, to Kevin K.H. Chung).

See also: Academic Achievement Motivation, Development of; Academic Performance, Effects of Socio-Economic Status on; Academic Self-Concept and Achievement; Ethnicity and Educational Achievement; Family and Schooling; Language Development: Influence of Socio-Economic Status; Literacy Education; School Achievement as Moderated by Sociodemographic Status: The US Case; School Achievement: Motivational Determinants and Processes; School Effectiveness Research; Social Inequality and Schooling; Socioeconomic Status and Health.

Bibliography

Anthony, E.K., King, B., Austin, M.J., 2011. Reducing child poverty by promoting child well-being: identifying best practices in a time of great need. Children and Youth Services Review 33, 1999–2009. http://dx.doi.org/10.1016/j.childyouth.2011.05.029.

Aram, D., Levin, I., 2002. Mother-child joint writing and storybook reading: relations with literacy among low SES kindergartners. Merrill Palmer Quarterly 48, 202–224.

Blair, S., Wenk, D., Hardesty, C., 1994. Marital quality and paternal involvement: interconnections of men's spousal and parental roles. Journal of Men's Studies 2, 221–237.

Bloom, B., Dey, A.N., Freeman, G., 2006. Summary Health Statistics for U.S. Children: National Health Interview Survey, 2005. National Center for Health Statistics, Washington, DC.

Bradley, R.H., Corwyn, R.F., 2002. Socioeconomic status and child development. Annual Review of Psychology 53, 371–399. http://dx.doi.org/10.1146/annurev.psych.53.100901.135233.

Bradley, R.H., Whiteside, L., Mundfrom, D.J., Casey, P.H., Kelleher, K.J., Pope, S.K., 1994. Early indications of resilience and their relation to experiences in the home environments of low birthweight, premature children living in poverty. Child Development 65, 346–360.

Brooks-Gunn, J., Duncan, G.J., 1997. The effects of poverty on children. The Future of Children 7, 55–71.

Cabell, S.Q., Justice, L.M., Logan, J.A., Konold, T.R., 2013. Emergent literacy profiles among prekindergarten children from low-SES backgrounds: longitudinal considerations. Early Childhood Research Quarterly 28, 608–620. http://dx.doi.org/10.1016/j.ecresq.2013.03.007.

Carlson, M.J., Corcoran, M.E., 2001. Family structure and children's behavioral and cognitive outcomes. Journal of Marriage & the Family 63, 779–792. http://dx.doi.org/10.1111/j.1741-3737.2001.00779.x.

Castillo, J., Welch, G.W., Sarver, C., 2011. Fathering: the relationship between fathers' residence, fathers' sociodemographic characteristics, and father involvement. Maternal and Child Health Journal 15, 1342–1349.

Clark, P., Kragler, S., 2005. The impact of including writing materials in early childhood classrooms on the early literacy development of children from low-income families. Early Child Development and Care 175, 285–301. http://dx.doi.org/10.1080/0300443042000266295.

Chiu, M.M., 2010. Inequality, family, school, and mathematics achievement. Social Forces 88, 1645–1676. http://dx.doi.org/10.1353/sof.2010.0019.

Chiu, M.M., Chow, B.W.-Y., 2 June 2014. Classmate characteristics and student achievement in 33 countries: classmates' past achievement, family socioeconomic status, educational resources, and attitudes toward reading. Journal of Educational Psychology. Advance online publication. http://dx.doi.org/10.1037/a0036897.

Chiu, M.M., Khoo, L., 2005. Effects of resources, inequality, and privilege bias on achievement. American Educational Research Journal 42, 575–603.

Cohen, S., 1999. Social status and susceptibility to respiratory infections. Annals of the New York Academy of Sciences 898, 246–253.

Cook, T.D., Church, M.B., Ajanaku, S., Shaish, W.R., Kim, J., Cohen, R., 1996. The development of occupational aspiration and expectations among inner-city boys. Child Development 67, 3368–3385.

Conger, R.D., Conger, K.J., Martin, M.J., 2010. Socioeconomic status, family processes, and individual development. Journal of Marriage and Family 72, 685–704.

Duncan, G., Magnuson, K.A., 2003. The promise of random-assignment social experiments for understanding well-being and behavior. Current Sociology 51, 529–541.

Duncan, G.J., Magnuson, K., Kalil, A., Ziol-Guest, K., 2011. The importance of early childhood poverty. Social Indicators Research. Advance online publication. http://dx.doi.org/10.1007/s11205-011-9867-9.

East, L., Jackson, D., O'Brien, L., 2006. Father absence and adolescent development: a review of the literature. Journal of Child Health Care 10, 283–295. http://dx.doi.org/10.1177/1367493506067869.

Frederickson, N., Petrides, K.V., 2008. Ethnic, gender, and socio-economic group differences in academic performance and secondary school selection: a longitudinal analysis. Learning and Individual Differences 18, 144–151.

Gettinger, M., Stoiber, K., 2007. Applying a response-to-intervention model for early literacy development in low-income children. Topics in Early Childhood Special Education 27, 198–213. http://dx.doi.org/10.1177/0271121407311238.

Grantham-McGregor, S., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., 2007. Child development in developing countries 1: developmental potential in the first 5 years for children in developing countries. The Lancet 369, 60–70. http://dx.doi.org/10.1016/S0140-6736(07)60032-4.

Heckman, J.J., 2006. Skill formation and the economics of investing in disadvantaged children. Science 312, 1900–1902. http://dx.doi.org/10.1126/science.1128898.

Hackman, D.A., Farah, M.J., Meaney, M.J., 2010. Socioeconomic status and the brain: mechanistic insights from human and animal research. Nature Reviews Neuroscience 11, 651–659. http://dx.doi.org/10.1038/nrn2897.

Hampden-Thompson, G., 2013. Family policy, family structure, and children's educational achievement. Social Science Research 42, 804–817.

Huisman, M., Araya, R., Lawlor, D.A., Ormel, J., Verhulst, F.C., Oldehinkel, A.J., 2010. Cognitive ability, parental socioeconomic position and internalising and externalising problems in adolescence: findings from two European cohort studies. European Journal of Epidemiology 25, 569–580.

Jordan, N.C., Kaplan, D., Ramineni, C., Locuniak, M.N., 2009. Early math matters: kindergarten number competence and later mathematics outcomes. Developmental Psychology 45, 850–867. http://dx.doi.org/10.1037/a0014939.

Jordan, N.C., Levine, S.C., 2009. Socioeconomic variation, number competence, and mathematics learning difficulties in young children. Developmental Disabilities Research Reviews 15, 60–68. http://dx.doi.org/10.1002/ddrr.46.

Ma, J., Horner, P., Han, Y., Grogan-Kaylor, A., Delva, J., Castillo, M., 2012. Predictors of discordance among Chilean families. International Journal of Child and Adolescent Health 5, 325–335.

Noble, K.G., McCandliss, B.D., Farah, M.J., 2007. Socioeconomic gradients predict individual differences in neurocognitive abilities. Developmental Science 10, 464–480. http://dx.doi.org/10.1111/j.1467-7687.2007.00600.x.

OECD, 2013. PISA 2012 Results in Focus: What 15-Year-Olds Know and What They Can Do with What They Know. OECD Publishing. http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf.

Petterson, S.M., Albers, A.B., 2001. Effects of poverty and maternal depression on early child development. Child Development 72, 1794–1813. http://dx.doi.org/10.1111/1467-8624.00379.

Ready, D.D., 2010. Socioeconomic disadvantage, school attendance, and early cognitive development: the differential effects of school exposure. Sociology of Education 83, 271–286. http://dx.doi.org/10.1177/0038040710383520.

Riley, M.R., Scaramella, L.V., McGoron, L., 2014. Disentangling the associations between contextual stress, sensitive parenting, and children's social development. Family Relations 63, 287–299. http://dx.doi.org/10.1111/fare.12063.

Rothstein, R., 2004. Class and Schools: Using Social, Economic, and Educational Reform to Close the Achievement Gap. Economic Policy Institute, Washington, DC.

Schiff, R., Ravid, D., 2011. Linguistic processing in Hebrew-speaking children from low and high SES backgrounds. Reading and Writing. Advance online publication. http://dx.doi.org/10.1007/s11145-011-9326-7.

Simpkins, S.D., Weiss, H.B., McCartney, K., Kreider, H.M., Dearing, E., 2006. Mother-child relationship as a moderator of the relation between family educational involvement and child achievement. Parenting: Science and Practice 6, 49–57.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. Review of Educational Research 75, 417–453.

Starfield, B., Riley, A.W., Witt, W.P., Robertson, J., 2002. Social class gradients in health during adolescence. Journal of Epidemiology and Community Health 56, 354–361.

Suárez-Orozco, C., Suárez-Orozco, M.M., Todorova, I., 2008. Learning a New Land: Immigrant Students in American Society. Harvard University Press, Cambridge, MA.

Zambrana, I.M., Ystrom, E., Pons, F., 2012. Impact of gender, maternal education, and birth order on the development of language comprehension: a longitudinal study from 18 to 36 months of age. Journal of Developmental and Behavioral Pediatrics 33, 146–155.

Socioeconomic Status and Health

Bjørgulf Claussen, Institute of Health and Society, University of Oslo, Blindern, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.A. Kaplan, volume 21, pp. 14554–14558, © 2001, Elsevier Ltd.

Abstract

Socioeconomic status is inversely associated with mortality and almost all measures of morbidity in terms of a social gradient that runs from the bottom to the top of society. The social inequalities are highest for drug-related diseases and injuries but are also high for cardiovascular diseases. The most clear-cut reason is social differences in lifestyle, which may explain about half of the health inequalities. The rest of the prevailing health inequalities in affluent societies may be explained by materialistic and psychosocial factors. The high and increasing inequalities in health in the most advanced welfare states are more difficult to explain. Inequalities in health between people in higher and lower social positions show that preventing many diseases and premature deaths is possible. In spite of governmental plans for reducing the divides in most affluent countries, social inequalities in health still seem to increase. A commission under the World Health Organization launched in 2008 a report proposing our leaders and popular movements to start a comprehensive move in policies to reduce social inequalities in the world's health.

In former times, and even today in poor countries, the inverse association between socioeconomic status (SES) and poor health has been obvious to all. Daniel Defoe in his *Journal of the Pleague Year* echoed such sentiments when he wrote "… the misery of that time lay upon the poor who, being infected, had neither food nor physic, neither physician nor apothecary to assist them, or nurse to attend." A change of view started when the Austrian physician Johan Peter Frank published his work in 1790 about the possibilities of reducing misery among poor people by public health measures. In 1848, the German physician and founder of modern pathology, Rudolf Virchow, showed the connection between typhus and poor living conditions in Upper Silesia, concluding with his famous sentence that "medicine is a social science, and politics is nothing else than medicine on a large scale." From the late 1800s, especially British statisticians explored the links between living conditions and health, and were able to demonstrate many new associations between health and SES.

This knowledge was summed up during the Second World War in the political program of the 'welfare state,' which, among others, had an aim to reduce the unfair social inequalities in health. The program was very successful in Northwestern Europe, giving the impression that such inequalities were about to disappear. Therefore, it was a shock when the British 'Black report' from 1980, named after the President of The Royal College of Physicians, Sir Douglas Black, showed that social inequalities in health were increasing in the United Kingdom and in many other Western countries. Since then, research in this field has steadily grown.

In 2005, the World Health Organization (WHO) established the Commission on Social Determinants of Health (2008), led by Sir Michael Marmot, to support countries and global health partners in addressing the social factors leading to ill health and health inequities. Their report concluded that social justice is a matter of life and death. It affects the ways people live, their chances of illness, and their risk of premature death. "We watch in wonder as life expectancy and good health continue to increase in parts of the world and in alarm as they fail to improve in others." Today, social inequalities in health are considered a well-established fact in research, policy, and the wider public.

The Relationship between Social Position and Health

Mortality is the core of health measures. In all affluent countries, it has been shown that mortality is inversely correlated with all available measures of SES – education, income, and occupation. Figure 1 shows the mean remaining years of life for all Canadians aged 25 years across deciles of income with 95% confidence intervals (CIs). The difference was 4.5 (CI 4.2–4.8) years between the lowest and highest deciles for women and 7.4 (CI 7.1–7.7) years for men. These differences are larger in the United States and the United Kingdom; in a study of mortality rates across income deciles in 11 Western countries, Sweden had the smallest difference, and the United States, the largest, with Canada as No. 4 and the United Kingdom as No. 8 (Leigh and Jencks, 2007).

The variance in mortality across SES is relatively high, showing that a lot of factors probably play a role. In a large European study, this was especially high in the low educational groups, largely driven by conditions causing death at younger ages, such as injuries and cancer (van Raalte et al., 2011).

Self-reported health is a good measure of general health. Figure 2 shows the prevalence of poor self-reported health for those with only primary education when the prevalence of those with postsecondary school education is indicated as 1.0 here across educational groups in 11 European countries (Cavelaars et al., 1998). A statistical method is used that adjusts for the size of the different groups, enabling comparability of the columns. Educational differences for this morbidity measure are largest in the Nordic countries and in the Netherlands. The reason may be that the best educated people in these countries enjoyed exceptionally good general health, while those with primary education reported about the same level of general health as in other European countries.

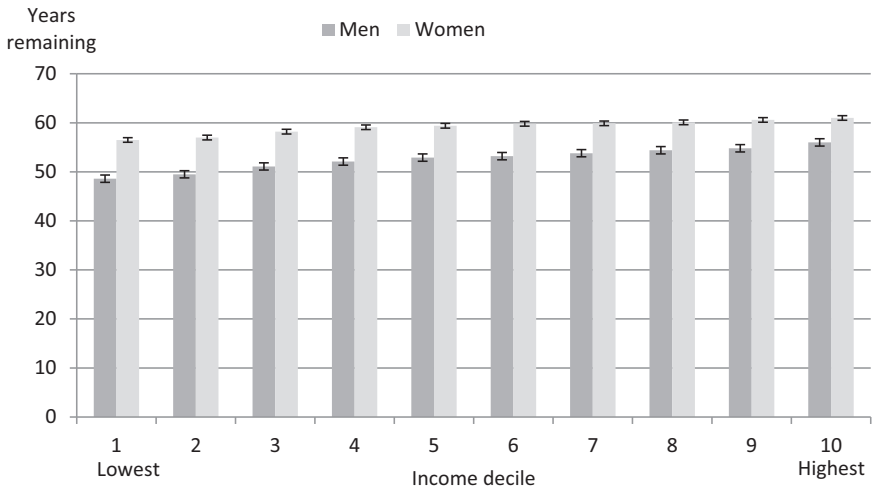

Figure 1 Remaining life expectancy at the age of 25 years after mortality rates in Canada during 1991–2001 by gender and income decile in 1991 with 95% CI. After McIntosh, C.N., Finès, P., Wilkins, R., et al., 2009. Income disparities in health-adjusted life expectancy for Canadian adults, 1991 to 2001. Statistics Canada. Component of Statistics Canada Catalogue No. 82-003-X Health Reports, Ottawa.

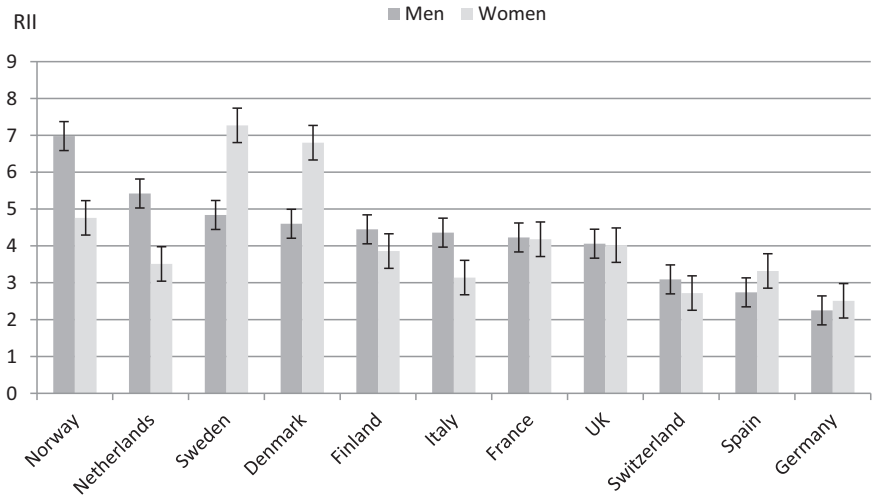

Figure 2 The prevalence rate of poor self-reported general health of those with primary education when the prevalence of those with university or other forms of higher education is like 1, expressed as relative index of inequalities (RII: i.e., a measure of inequalities taking into account the different number of people in each category across countries) with 95% CI. After Cavelaars, A.E.J.M., Kunst, A.E., Geurts, J.J.M., et al., 1998. Differences in self reported morbidity by educational level: a comparison of 11 Western European countries. Journal of Epidemiology and Community Health 52, 219–227.

Concerning Figures 1 and 2, it is important to remark that there are stepwise differences in mortality and morbidity across the whole range of social inequalities, such that a social gradient runs from the bottom to the top of society. This pattern is observed among many different diseases. The social inequalities are highest for drug-related diseases and next for injuries, which is understandable, but they are also high for cardiovascular diseases, a fact that is more difficult to understand (see below). Smoking-related cancers follow a steep social gradient, reflecting a high prevalence of smoking among lower SES population groups. Many other cancers demonstrate a similar but less pronounced pattern (Merletti et al., 2011). There are a few exceptions, however; breast cancer, for instance, has a positive association with SES. As age at first pregnancy is a risk factor, and as this age is higher among women with higher SES, this association seems to be explained (Strand et al., 2005). Yet, the problem is more complex because breast cancer incidence is also high among high-SES women without any pregnancy. Similarly, the positive association of testicular cancer with SES remains unexplained.

Age influences these patterns. Social health inequalities are largest for infants, children, and middle-aged people but are also remarkable for old people (Marmot, 2004). The differences are smallest for adolescents (Starfield et al., 2002) or may even be absent (West and Sweeting, 2004). Adolescents are on their way from the social status of parental life, and this may not adequately reflect their own social situation. Moreover, adolescence is a period when health-adverse behaviors are adapted and their adverse health effects become manifest at a later stage of their life course.

Gender is important. Women have all over less social inequalities in health than men (**Figures 1** and **2**). When occupational position is divided into five grades, the top-grade females usually have higher mortality than the second top grade, almost as high as for top-grade men. Some researchers have drawn the conclusion that these women at the top have a life more similar to men than their sisters below.

Lifecourse SES is important when assessing social inequalities in health from fetal life to old age (Graham and Power, 2004). A growing field of research has studied the impact of pregnancy factors on health. Living conditions of mothers are socially patterned, and risks of low birth weight and preterm birth are negatively associated with SES. Both conditions are linked to poor health in adult life (the Forsdahl–Barker hypothesis). Preterm birth is also important for infant health, being a complex cluster of behavioral, psychosocial, and genetic factors as well as contextual characteristics, including access to medical treatment. Many of these factors occur in combination, particularly in those who are socioeconomically disadvantaged or who are members of racial and ethnic minority groups. Childhood SES is especially important because of the strong association with adulthood health (Davey Smith et al., 1998).

Personality characteristics seem to influence both SES and health. In a large French firm, a sample of 14 445 employed persons answered personality tests and were followed up for nearly 13 years (Nabi et al., 2008). The association between risk of death and SES among men was attenuated with 28–29% for 'neurotic hostility,' 13–22% for 'antisocial' personality, and 13–16% for 'CHD-prone' personality (a personality type with increased risk of cardiovascular disease), depending on SES. Personality measures explained some of the mortality gradients across SES for men but had little explanatory power for women. It is not clear if personality represents a predictor or an outcome of social circumstances. A similar study, the British Whitehall study of health in the governmental offices in London, showed that cognitive ability accounted for 17–39% of health inequalities related to SES (Singh-Manoux et al., 2005). In a later US study, the French results of personality profiles and SES were replicated.

Lifestyle, like smoking, food habits, and training, explains a substantial part of the social gradients in health, which was demonstrated by the Whitehall studies (Marmot, 2004). This part is quantified as percentage of variance explained, and different reports observed a range varying between 30% and almost 90% (Stringhini et al., 2010).

Social capital is a concept for an index of amounts of social network, social trust, and norms for reciprocal help in a social group, according to the American sociologist Putnam. Ichiro Kawachi et al. (1997) have shown that social capital of the people in a region is inversely associated with mortality, analyzing state-level data from the United States. Similar associations between health and social capital in regions were later found in several Western countries and in some others, such as Columbia, Brazil, Japan, and South Korea (Murayama et al., 2012). These results have been interpreted as effects of psychosocial factors that influence health-related behavior and other individual-level mechanisms of importance. An example is the social capital factor at the workplace.

Job strain seems to be important for social inequalities in health as extensively shown by Johannes Siegrist and his group. In a study of 47 000 employees, job strain was associated with increased risk factors for cardiovascular disease (Nyberg et al., 2013). More specifically, exposure to an adverse psychosocial environment, defined by high demands and low control and/or by effort–reward imbalance, both more frequent in lower-SES groups, was associated with increased mortality risks (Siegrist and Marmot, 2004). Elevated heart rate and blood pressure and elevated morning secretion of the stress hormone cortisol were found when efforts spent at work were not in balance with rewards in terms of money, promotion prospects, job security, and esteem (Siegrist, 2010).

Area of residence seems to have an independent influence on health (Curtis, 2004). Small-area statistics show that mortality rates are inversely associated with mean income in the neighborhoods. Most of this effect is explained by the fact that people with low individual income are at elevated mortality risk (so-called 'compositional effect'). However, there is some indication that an unfavorable socioeconomic context, like low mean income in an area, exerts an additional influence ('contextual effect').

Social capital in the area, as mentioned above, may have such a contextual effect. Housing qualities are different across social position and very often cluster among people of similar position in separate neighborhoods, thus being both an individual and a contextual factor. Pollution, traffic, and related factors often make poor neighborhoods less healthy. Both the official infrastructure, like schools, sports facilities, parks, and roads, and the availability of private services, like shops for healthy food, are often socially patterned in a similar way.

Time trends are best studied in Europe. Social inequalities in mortality seemed to have reached a peak around 1920, and then decreased toward the 1960s among men and toward the 1970s among women, thereafter increasing again up to now (Mackenbach, 2013). European welfare states were built up after World War II, and one of the important aims was to reduce social inequalities in health. Today, it is clear that all its inhabitants enjoy better health and better economy than before, and yet, it is surprising that health inequalities do not diminish. To the contrary, some of their indicators seem to be still most pronounced in countries with high grants and universal health coverage, as **Figure 2** indicates.

Explanations of Social Inequalities in Health

Johan Mackenbach et al. (2012) has tried to explain the development of social inequalities in health in Western Europe, especially the paradox of the increasing time trend and the high inequalities in the most advanced welfare states. He proposes nine possible explanations:

1. Mathematical artifact
2. A theory of 'fundamental causes'
3. The life course perspective
4. Intragenerational social selection
5. Intergenerational social mobility
6. Psychosocial factors
7. Neomaterialistic explanations
8. Diffusion of innovations
9. Cultural capital

The first three explanations are general theories, which all would require more specific explanations of the 'welfare state paradox' mentioned above. This may be a *mathematical artifact* (1). As the overall levels of poor health decline and mortality differences become smaller, the relative inequalities may turn out to be magnified, or in other words, unlucky outfalls may not alleviate as much as lucky ones over time. A theory of *fundamental causes* (2) is proposed, claiming that fundamental conditions of social stratification ultimately underlay the observed health inequalities, rather than proximal risk factors like smoking, stress, etc. (Phelan et al., 2010). The *life course perspective* (3) maintains that biological programming in fetal and childhood development is the main cause of adult health inequalities, and many still unknown factors may be important in these vulnerable phases of life.

The next two theories focus on how the composition of the social layers is formed by health-related selection.

A theory of *social selection* (4) explains health inequalities by a high intragenerational social mobility in modern societies. It is probable that good health potentials will be associated with upward mobility, and vice versa. Direct social selection effects on mortality have been studied, and this research has demonstrated that intragenerational selection hardly plays any substantial role in explaining subsequent health inequalities (Blane et al., 1999). *Intergenerational social mobility* (5) has been growing in high-income countries in the postwar period (Osler et al., 2005). These slow processes are likely to increase in affluent societies over time, and to be larger in more advanced welfare states. Modern societies have become 'meritocracies' where social position depends on personal talents and efforts, and less on family background. These personal characteristics may partly arise from differences in childhood environments, but may also be due to selection from different social layers.

The two selection theories may explain to some extent how the increasingly smaller groups in the lower social strata tend to consist of people with increasingly smaller resources, thus resulting in a stronger impact of adverse conditions of living on their health. Four theories are focusing on the distribution of specific resources.

Psychosocial factors (6), like control over one's own living conditions, self-esteem, and feelings of success, were proposed to be relevant when Wilkinson and Pickett (2010) demonstrated that inequalities in health were higher in countries with high income differences, e.g., the United States, than in countries with a more equitable income distribution, e.g., Japan. Later, Wilkinson demonstrated that high income differences within distinct countries were associated with a high level of distrust, criminality, child neglect, anxiety, and depression across countries. The association between social capital and mortality is also in line with this theory.

Some years ago, the psychosocial theory was heavily disputed when Lynch et al. (2004) showed, based on more recent data from a higher number of affluent countries, that health inequalities were associated more closely with absolute than with relative income differences. Moreover, Wilkinson's results were restricted to the United States and the United Kingdom. These researchers concluded that health inequalities are more convincingly explained by what they termed a *neomaterialistic* (7) approach where factors like purchase power, housing, holidays, or the quality of infrastructure in different areas of residence matter. The dispute was quite vivid, but not really productive as psychosocial factors, like control and feelings of success, are often based on material conditions, e.g., on a society's willingness to distribute money. Therefore, differences between the psychosocial and the material approach may not be substantial.

Some researchers put the emphasis on *diffusion of innovations* (8), having observed that people in higher positions often tend to be early adapters of new behaviors, who are followed by those in a lower position with temporal delay, as is, for instance, the case with giving up smoking. This theory puts a heavy weight on the results from studies of *cultural capital* (9) and health inequalities. This notion is based on the widely known work of sociologist Pierre Bourdieu who studied social inequalities in consumption behavior. These cultural patterns are explained by differences in attitudes and competencies that are transmitted across generations. Cultural capital may be of growing importance for persons moving up in the social hierarchy, specifically in societies where signs of material prosperity are becoming less important and may exert favorable effects on health behavior.

Mackenbach assumes that 'neomaterialistic' and 'psychosocial' theories explain why social inequalities persist in societies where the majority has access to an 'affluent lifestyle,' while economic inequalities still persist. Yet, the two theories fail to explain the 'welfare state paradox' that inequalities are growing in welfare states. Here, factors like cultural capital and personal characteristics may help explaining the paradox because their effects become stronger under the more equalizing impact of welfare state regimes.

Biological Pathways

Death is a biological and individual occurrence. All nine explanations mentioned above concern phenomena located at the societal as well as the individual level. A crucial challenge for research concerns the question of how conditions associated with people's social position are translated into individual health, resulting ultimately in their death. These processes are still poorly understood, but some possible biological pathways have been elucidated.

Lifestyles, like smoking and diet, have well-known biological effects. But it is more difficult to unravel mechanisms linking psychosocial conditions with physical diseases, such as coronary heart disease with its underlying atherosclerosis. In this regard, Brunner (1997) discussed stress mechanisms as important factors of atherosclerotic development, referring to experiments with monkeys whose levels of the stress hormone cortisol and whose atherosclerotic lesions were increased as a function of disturbed social hierarchies. Social status is associated with stress physiology and health also in humans and other animals, according to Cavigelli and Chaudhry (2012). Psychological stress was shown to be inversely correlated with people's standing in their social hierarchy. Some studies have demonstrated elevated levels of stress hormones like cortisol among people with lower SES, but results are inconclusive. Work stress in terms of effort–reward imbalance was shown to be associated with high morning cortisol level (Siegrist, 2010), and the same holds true for a measure of low

job control (Siegrist and Marmot, 2004). These results can be interpreted as support for the psychosocial explanation of social inequalities (Marmot, 2004).

However, there are additional regulatory systems in the human organism where the central nervous system is intimately linked with endocrine and immune systems to adapt to external demands in differentiated and complex ways (Ulvestad, 2007). Immune function is related to SES in a similar way. The theoretical construct of allostasis (or allostatic load) was introduced to measure accumulated burden on the organism's regulatory systems, which in the long run results in functional loss and structural lesions (McEwen, 2012). These measures have contributed to the understanding of how constantly changing social and environmental factors influence physiological functioning and shape health inequalities, particularly along socioeconomic, gendered, racial, and ethnic lines. The literature demonstrates associations between allostatic load and physical and mental health as well as all-cause mortality (Beckie, 2012).

Nancy Krieger (2013) maintains that these relationships between social and bodily processes may change over time. She therefore suggests that diseases and their biomarkers may be thought of as *embodied histories*, that is, as emerging embodied phenotypes rather than as resulting from 'inborn biology.'

Political Implications and Conclusion

Reducing social equities in health has come to the political agenda in most Western countries. An example is governmental support to the Acheson Report (1998) in the United Kingdom. This report gave 39 principal recommendations in order to reduce poverty in women of childbearing age, pregnant women, young children, and older people. Recommendations included a reduction of income inequalities and an improvement of living standards of households among recipients of social security benefits. Moreover, deprived conditions of ethnic minority groups; unemployed, elderly, and disabled people; and families with children should be improved, and benefit levels and real living standards should be increased. Other recommendations concerned improvements of material conditions in schools, housing, environment, transport, and diet. The Labor government of 1997 tried to set much of this plan into policy but only some of the main aims were met (Mackenbach, 2010). Three main reasons of the failure to meet these goals were discussed. First, some entry points were not relevant for changing health and life expectancy; second, the proposed strategy was not able to use effective policies; and third, the strategy was not delivered at a large scale with sufficient impact.

Similar policy measures were proposed in the US context (Adler and Newman, 2002). An overview of distinct governmental programs in four countries, the United Kingdom, Sweden, Denmark, and Norway, showed that all strategies dealt with the challenge of increasing social inequalities in health, the English and Swedish strategies more so than the other ones, but there was no specific Nordic model in this field of welfare state politics (Valgårda, 2001).

Economic globalization may widen the gap between what we know about social health inequalities and health policy efforts in affluent societies (Siegrist, 2008). It is hoped that rapidly developing countries can learn from Western societies how important it is to identify, monitor, explain, and reduce social inequalities in health. Clearly, given the scope of this problem and the role of sociocultural differences across the world, substantial challenges remain for science as well as for policy.

Despite these challenges, the Commission on Social Determinants of Health (2008) of WHO demonstrated some promising ways of tackling inequalities in health. The commission stressed the importance of 'health in all' policies, emphasizing that action on the social determinants of health is needed, across the life course and in the wider social and economic spheres.

See also: Globalization and Health; Health Interventions, Community-based; Life Course Epidemiology; Public Health as Social Science; Social Epidemiology; Social Exclusion, Social Deprivation and Health; Socioeconomic Status and Health Care.

Bibliography

Acheson, D., 1998. Inequalities in Health. Report of an Independent Inquiry. HMSO, London.

Adler, N.E., Newman, K., 2002. Socioeconomic disparities in health: pathways and policies. Health Affairs (Millwood) 21, 60–76.

Beckie, T.M., 2012. A systematic review of allostatic load, health, and health disparities. Biological Research for Nursing 14, 311–346.

Blane, D., Harding, S., Rosato, M., 1999. Does social mobility affect the size of the socioeconomic mortality differential?: evidence from the Office for National Statistics Longitudinal Study. Journal of the Royal Statistical Society: Series A (Statistics in Society) 162, 59–70.

Brunner, E., 1997. Stress and the biology of inequality. British Medical Journal 314, 1472–1476.

Cavelaars, A.E.J.M., Kunst, A.E., Geurts, J.J.M., et al., 1998. Differences in self reported morbidity by educational level: a comparison of 11 Western European countries. Journal of Epidemiology and Community Health 52, 219–227.

Cavigelli, S.A., Chaudhry, H.S., 2012. Social status, glucocorticoids, immune function, and health: can animal studies help us understand human socioeconomic-status-related health disparities? Hormones and Behavior 62, 295–313.

Commision on Social Determinants of Health, 2008. Closing the Gap in a Generation. Health Equity through Action on the Social Determinants of Health. WHO, Geneva.

Curtis, S., 2004. Health and Inequalities. Geographical Perspectives. Sage publications, London.

Davey Smith, G., Hart, C., Blane, D., et al., 1998. Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study. British Medical Journal 316, 1631–1635.

Graham, H., Power, C., 2004. Childhood disadvantage and health inequalities: a framework for policy based on lifecourse research. Child Care Health and Development 30, 671–678.

Kawachi, I., Kennedy, B.P., Lochner, K., et al., 1997. Social capital, income inequality, and mortality. American Journal of Public Health 87, 1491–1498.

Krieger, N., 2013. History, biology, and health inequities: emergent embodied phenotypes and the illustrative case of the breast cancer estrogen receptor. American Journal of Public Health 103, 22–27.

Leigh, A., Jencks, C., 2007. Inequality and mortality: long-run evidence from a panel of countries. Journal of Health Economics 26, 1–24.

Lynch, J., Davey Smith, G., Harper, S., et al., 2004. Is income inequality a determinant of population health? Milbanks Quaterly 82, 5–99.

Mackenbach, J.P., 2010. The English strategy to reduce health inequalities. Lancet 376, 1986–1987.

Mackenbach, J.P., 2012. The persistence of health inequalities in modern welfare states: The explanation of a paradox. Social Science & Medicine 75, 761–769.

Mackenbach, J.P., 2013. Convergence and divergence of life expectancy in Europe: a centennial view. European Journal of Epidemiology 28, 229–240.

McEwen, B.S., 2012. Brain on stress: how the social environment gets under the skin. Proceedings of the National Academy of Sciences of the United States of America (Suppl. 2), 17180–17185.
McIntosh, C.N., Finès, P., Wilkins, R., et al., 2009. Income disparities in health-adjusted life expectancy for Canadian adults, 1991 to 2001. Statistics Canada. Component of Statistics Canada Catalogue No. 82-003-X Health Reports, Ottawa.
Marmot, M., 2004. Status Syndrome. Bloomsbury, London.
Merletti, F., Galassi, C., Spadea, T., 2011. The socioeconomic determinants of cancer. Environmental Health 10 (Suppl. 1), S7.
Murayama, H., Fujiwara, Y., Kawachi, I., 2012. Social capital and health: a review of prospective multilevel studies. Journal of Epidemiology 22, 179–187.
Nabi, H., Kivimäki, M., Marmot, M.G., et al., 2008. Does personality explain social inequalities in mortality? The French GAZEL cohort study. International Journal of Epidemiology 37, 591–602.
Nyberg, S.T., Fransson, E.I., Heikkilä, K., et al., June 20, 2013. Job strain and cardiovascular disease risk factors: meta-analysis of individual-participant data from 47,000 men and women. PLoS One 8, e67323.
Osler, M., Andersen, A.M., Batty, G.D., et al., 2005. Relation between early life socioeconomic position and all cause mortality in two generations. A longitudinal study of Danish men born in 1953 and their parents. Journal of Epidemiology and Community Health 59, 38–41.
Phelan, J.C., Link, B.G., Tehranifar, P., 2010. Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. Journal of Health and Social Behavior 51, S28–S40.
van Raalte, A.A., Kunst, A.E., Deboosere, P., et al., 2011. More variation in lifespan in lower educated groups: evidence from 10 European countries. International Journal of Epidemiology 40, 1703–1714.
Siegrist, J., Marmot, M., 2004. Health inequalities and the psychosocial environment: two scientific challenges. Social Science and Medicine 58, 1463–1473.
Siegrist, J., 2008. Social determinants of health – a cross-cultural perspective. International Journal of Public Health 53, 277–278.
Siegrist, J., 2010. Effort-reward imbalance at work and cardiovascular diseases. International Journal of Occupational Medicine and Environmental Health 23, 279–285.
Singh-Manoux, A., Ferrie, J.E., Lynch, J.W., et al., 2005. The role of cognitive ability (intelligence) in explaining the association between socioeconomic position and health: evidence from the Whitehall II prospective cohort study. American Journal of Epidemiology 161, 831–839.
Starfield, B., Riley, A.W., Witt, W.P., et al., 2002. Social class gradients in health during adolescence. Journal of Epidemiology and Community Health 56, 354–361.
Strand, B.H., Tverdal, A., Claussen, B., et al., 2005. Is birth history the key to highly educated women's higher breast cancer mortality? A follow-up study of 500,000 women aged 35–54. International Journal of Cancer 117, 1002–1006.
Stringhini, S., Sabina, S., Shipley, M., et al., 2010. Association of socioeconomic position with health behaviors and mortality. Journal of the American Medical Association 303, 1159–1166.
Stringhini, S., Dugravot, A., Shipley, M., et al., February 2011. Health behaviours, socioeconomic status, and mortality: further analyses of the British Whitehall II and the French GAZEL prospective cohorts. PLoS Medicine 8, e1000419.
Ulvestad, E., 2007. Defending Life: The Nature of Host-Parasite Relations. Springer, Dordrecht.
Vallgårda, S., 2001. Governing people's lives. Strategies for improving the health of the nations in England, Denmark, Norway and Sweden. European Journal of Public Health 11, 386–392.
West, P., Sweeting, H., 2004. Evidence on equalisation in health in youth from the West of Scotland. Social Science and Medicine 59, 13–27.
Wilkinson, R., Pickett, K., 2010. The Spirit Level: Why Equality is Better for Everyone. Penguin Books, London.

Relevant Websites

http://www.who.int/social_determinants/resources/mekn_report_10oct07.pdf – The Marmot Commission report.
http://www.who.int/social_determinants/thecommission/en/ – The Marmot Commission's video "Watch and learn about Social Determinants of Health".

Socioeconomic Status and Health Care

Marsha Regenstein and Paula M Lantz, The George Washington University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P.M. Lantz, volume 21, pp. 14558–14562, © 2001, Elsevier Ltd.

Abstract

There is a vast amount of evidence across countries that the use of health care services (including hospitalizations, physician services, and clinical preventive services) is positively associated with income, education, and other markers of socioeconomic position. In some analyses, lower socioeconomic status (SES) is associated with greater physician and hospital use, although it appears that these findings are primarily driven by higher rates of poor health status or medical need in socioeconomically disadvantaged populations. Three general sets of explanations have been investigated to explain SES differences in health care use: financial access or affordability differences (e.g., the poor use fewer services because they have fewer financial means); knowledge, attitudinal, and cultural differences in seeking care (e.g., the poor use fewer services because they value and/or understand the benefits of health care less than other people); and health care system factors (e.g., the poor use fewer services because of ways in which the health care system is organized). The empirical evidence from a number of countries suggests that health care system variables, especially nonfinancial components of access, offer the greatest explanatory power regarding SES differences in utilization. The policy implications of such findings are important, for they suggest that providing universal coverage or offering free health services to the poor is not a panacea for reducing social inequalities in health care utilization. Research across countries also suggests that, although SES differences in the use of health care services may contribute to some portion of observed SES disparities in health status, the contribution here is likely to be small. Social inequalities in health status are driven by myriad factors other than access to and use of personal health services.

Socioeconomic Status and Health Care

Although many factors determine the availability and use of health care, there is substantial evidence across countries that markers of socioeconomic status (SES), including income, education, and occupational status, are significantly associated with health care utilization. This includes health insurance coverage and the use of personal health services such as visits to primary care providers and specialists, hospitalizations, dental care, vision care, mental health services, prescription drug use, and home health care. Although there is ample evidence that people of higher socioeconomic position use more health care services and tend to receive higher quality care, the reasons behind these disparities are not fully understood. Later on, the relationship between SES and health services utilization is described, and explanations for this relationship are summarized.

The Relationship between SES and Health Care Utilization

Numerous studies have found a positive relationship between SES and most types of health services utilization, measured by whether or not a service type was used in the past year, or as the mean number of encounters/visits for the service during a specified time period (National Center for Health Statistics, 2012; Ross et al., 2006). Although there are some exceptions to these general findings, the bulk of the empirical findings to date suggests that the association is positive; that is, higher SES is associated with a higher rate of health care utilization (see **Table 1** for some examples from the United States). Exceptions occur in cases when certain types of health care utilization indicate poor access to health services that are considered essential to maintain health and well-being. An example of this is higher rates of emergency department (ED) use for lower socioeconomic levels, signaling potential barriers to accessing primary and specialty care services. Higher SES is also associated with lower rates of untreated disease and unmet medical needs. That SES has a positive impact on health care use has been observed for both men and women, and for children and adults.

While the association between SES and health care utilization is positive and graded in most studies, other patterns have been documented as well. An inverse association between SES and use of physician services was noted in Canada, along with the hypothesis that excess use by those of lower income and education levels is attributable to SES differences in health status (Kephart et al., 1998). Among children in five Nordic countries, few differences were found in the use of physician services for different socioeconomic groups (Halldorsson et al., 2002). In the United States, those without health insurance, those living in low-income areas, and racial/ethnic minorities are significantly more likely to receive potentially avoidable hospitalizations and medical procedures (Pappas et al., 1997; Blustein, 1998). In summary, people from lower economic and educational strata tend to have fewer physician visits for both primary and specialty care and to use fewer mental health, dental, vision, and preventive services. They also are less likely to have a usual source of care and have significantly higher rates of unmet health needs. In some analyses, lower SES is associated with greater physician and hospital use (especially for avoidable causes of injury or illness), although it appears that these findings are primarily driven by higher rates of poor health status or medical needs in disadvantaged populations.

Table 1 Health care utilization by poverty status for selected measures, United States, 2007–2010

Health service	*Poor*	*Near poor*	*Middle income*	*High income*
No health insurance				
Children (<18 years)	22.0%	22.8%	8.6%	4.2%
Adults (18–64 years)	44.1%	35.3%	13.8%	6.1%
No usual source of care				
Children (<18 years)	8.3%	7.5%	4.7%	2.1%
Adults (18–64 years)	33.8%	30.5%	20.5%	10.8%
Did not get care due to cost				
Adults (18–64 years)	23.4%	24.0%	15.2%	6.8%
Use of pap smears				
Women (>18 years)	65.1%	64.3%	71.3%	83.1%
Dental visit in past 12 months				
Adults (18–64 years)	41.0%	44.1%	59.6%	77.1%
ED visit in past 12 months				
Adults (18–64 years)	30.6%	25.6%	20.4%	17.0%
Persons with self-assessed fair or poor health				
All ages	20.9%	15.2%	8.3%	4.3%

Reproduced from Special Feature on Socioeconomic Status and Health, 2012. Hyattsville, MD.

SES and its proxies are not the only factors related to health care use. Utilization is also patterned by gender, geography, age, culture, and other factors. While it is difficult to sort out the independent effects of SES and race/ethnicity on health care, it appears that minority status has an important independent effect. In the United States, certain racial and ethnic minorities (most notably, African-Americans, Native Americans, and Hispanics) have been found to have lower rates of utilization in general, to receive less aggressive care for certain diseases, and to report less satisfaction with and more experiences with discrimination regarding, health care (Institute of Medicine, 2002; Fiscella et al., 2000). Even so, SES appears to be a major social predictor of health care use. The relationship is strong across different types of health services and is robust across a variety of study designs from a number of different countries.

Explanations for the Relationship between SES and Health Care Utilization

While the evidence for a relationship between SES and health care use is strong and compelling, there are many methodological challenges in measuring this association and its causes (Brown et al., 1998). First, the relationship is confounded by health status, which is related to both SES and the need for services. Thus, it is important to measure health services utilization within the context of medical need. Second, it is difficult to have all necessary variables in a single population-based data set, including valid and reliable measures of SES, health insurance, medical need, and other confounding factors. Third, there are many different types of health services, and relationships are not uniform in magnitude or patterns across these different types. Fourth, simple measures of utilization do not provide a very deep understanding of differences that may exist in terms of the quality, content, or continuity of care.

Despite the methodological challenges for this type of research, several factors have been investigated to explain why people of lower SES generally use fewer health services than those of higher social position and why their level of use does not appear to match their need. Dutton (1978) reported that explanations for this phenomenon generally fall into three categories: explanations that focus on financial access or affordability differences (e.g., the poor use fewer services because they have fewer financial means); explanations that focus on knowledge and attitudinal differences regarding care seeking (e.g., the poor use fewer services because they value and/or understand the benefits of health care less than other people); and explanations that focus on health care system factors (e.g., the poor use fewer services because of how the health care system is organized or responds to them). This set of explanations still reflects the primary reasons offered for the strong association between SES and health care utilization.

Financial or Affordability Explanations

Since being uninsured is strongly correlated with income and education, it is often assumed that the strong relationship between SES and health care use operates primarily through the more proximate variable of health insurance. People with health insurance generally have a usual source of care, a measure that serves as a proxy for access to stable and accessible primary and preventive health services (see **Table 2**). Those who are privately insured, who generally are employed in firms that offer health coverage as a benefit of employment, have greater access to health services and in general enjoy better health than individuals who are uninsured or are insured by Medicaid, the US health insurance program for certain categories of low-income and disabled individuals. Expanding Medicaid in certain US states has been associated with better access and greater use of health care services for low-income individuals (Sommers et al., 2012; Baicker et al., 2013). Major health reform initiatives in the United States through the Patient Protection and Affordable Care Act will result in millions of uninsured individuals obtaining health coverage through subsidized private health insurance markets for

Table 2 Health care utilization by insurance status and type of coverage for selected measures, United States, 2009–2010

Health service	*Private insurance*	*Medicaid*	*Uninsured*
No usual source of care			
Children (<18 years)	2.6%	3.4%	28.8%
Adults (18–64 years)	10.5%	12.5%	50.6%
ED visit in past 12 months			
Adults (18–64 years)	17.4%	40.2%	21.3%
Pap smear			
Women (>18 years)	84.2%	78.0%	61.9%

Reproduced from Special Feature on Socioeconomic Status and Health, 2012. Hyattsville, MD.

low- and middle-income individuals (known as health insurance exchanges) or state Medicaid expansions that enable more of the country's poor and near-poor to qualify for coverage (Rosenbaum, 2011). These policy changes recognize the strong relationship between coverage and access to health care and the effects of lack of health insurance coverage on health. They also hold promise for increasing the share of Americans who have a usual source of care, which can also yield benefits in terms of strengthening primary care and preventive services and decreasing use of unnecessary and costly care (Lillie-Blanton and Hoffman, 2005; Fiscella, 2011).

Nevertheless, even among people with health insurance, lower levels of income and education are associated with lower overall health care use (Collins et al., 2012). Although access to health care is often equated with financial access, access is in fact multidimensional, and other aspects of access (such as the availability, convenience, and acceptability of providers and services) are of equal importance (Gold, 1998). Thus, SES differences in health care use are not fully explained by economic or financial factors.

Knowledge, Attitudinal, and Cultural Explanations

Lower use of health services among disadvantaged populations may be explained in part by chronic stressors and challenges associated with poverty (Adler and Newman, 2002). These stresses may result in a lack of attention to health and underuse of health care services generally or lower use of care-seeking behaviors for specific symptoms or conditions. Increasingly, research on health services utilization and SES has focused on developing an understanding of patient preferences and individual health needs as they relate to health care. Patient centeredness has emerged as a core value of high-quality health care delivery that recognizes the need to respect individuals and families as they make health care decisions (Institute of Medicine, 2001). Patients' knowledge of the importance of health services is associated with racial and ethnic concordance between the patient and health care provider, the patient's preferred language, the effects of chronic health conditions or disabilities on the patient's health care experiences, and a health care provider's ability to provide health care that is culturally and linguistically appropriate (Agency for Healthcare Research and Quality, 2012a). African-Americans, Hispanics/Latinos, Asians, and other groups exhibit different use patterns that may reflect, in part, different knowledge, cultural values, and preferences that may also vary by SES. Dissatisfaction with health care, lower rates of compliance, poorer understanding of health conditions, and lower self-rated health status are better explained by health system-level deficiencies rather than previous constructs related to a culture of poverty (Niewczyk and Lwebuga-Mukasa, 2008; Baker and Newsom, 2003).

Health Care System Explanations

Measures of health care quality are lowest for poor and low-income individuals. Empirical evidence from a number of countries suggests that health care system factors, especially nonfinancial components of access, offer the greatest explanatory power regarding SES differences in utilization. Research has shown that SES differences in health care utilization are driven much more by structural elements of the health care system than by individual factors such as health insurance or other financial variables (Rosenbach et al., 1999). These system-level factors include cultural deterrents to care, the unwieldy nature of large medical bureaucracies, fragmentation in care, racism, and other forms of provider bias, transportation barriers, and nonfinancial dimensions of access.

Not only is SES related to health care utilization, it is also related to health care quality (Fiscella et al., 2000). In 2001, the Institute of Medicine of the US National Academies of Sciences defined high-quality health care as care that is safe, timely, effective, efficient, equitable, and patient centered. This work set the stage for further studies to describe racial, ethnic, and linguistic disparities in health care. Despite an emphasis on high-quality, equitable care, disparities associated with race, ethnicity, and income level persist, with key measures of quality showing little or no improvement over the last decade (Agency for Healthcare Research and Quality, 2012b). On many measures of quality, African-Americans and Hispanics in the United States continue to receive worse care compared to whites. Quality of care measures is lowest for poor and low-income individuals, as are most access to care measures. Even individuals with better access may not receive the full complement of services that are considered by experts to be within the recommended standard of care. Despite spending more than any other country on health care, the average adult in the United States receives only about 55% of recommended services and children receive even less (McGlynn et al., 2003; Mangione-Smith et al., 2007).

It is important to view these quality deficiencies within the broader context or system of how health care services are

organized, delivered, and financed. Payments for health services in the United States have traditionally been linked to volume or fee-for-service type arrangements, which leave little incentive for health professionals to improve the quality of care they provide. Efforts to monitor quality have also been hampered by limitations in data collection and analysis, a situation that may improve with the broad implementation of health information technology and electronic health records. Yet even with these advances, the policy implications suggest that providing universal coverage or offering free health services to the poor is not a panacea for reducing social inequalities in health care utilization. Changes to ways that health care is organized and delivered may prove critical to reducing health care disparities in health services utilization.

The Role of Health Care in Explaining SES Disparities in Health Status

There are many explanations for the strong relationship between SES and health status, as outlined in the accompanying article on this topic (*see* Socioeconomic Status and Health). This includes explanations that focus on different material resources, of which health care is viewed as an important type (MacIntyre, 1997). Research from a number of countries suggests that, for specific health conditions, the receipt of clinical preventive services and/or timely therapeutic care plays an important role in health outcomes, and thus may account for some portion of socioeconomic disparities in health status (Lantz et al., 1997). Therefore, it has been argued that efforts to increase the use of health services among the socioeconomically disadvantaged, primarily by reducing financial barriers to health care or health insurance, have great potential for reducing socioeconomic disparities in health status (Sommers et al., 2012; Andrulis, 1998).

Research at the end of the twentieth century, however, also suggests that the contribution of SES differences in health care use to overall SES disparities in health status is likely to be small (Adler et al., 1993; MacIntyre, 1997; Baicker et al., 2013). Evidence for this comes from a variety of sources, including, as described above, countries with universal health insurance. For example, the Black report demonstrated that socioeconomic inequalities in mortality in England and Wales increased rather than attenuated in the decades after the British National Health Services was established (Gray, 1982). It has also been shown that, in developed countries, there is no relationship between health indicators (such as life expectancy) and the proportion of economic resources spent on health care (Roberts and House, 2000). The United States, Canada, and Great Britain are all experiencing increasing disparities in life expectancy between higher- and lower income individuals as well as higher- and lower educated people, despite the availability of universal health coverage in the latter two countries (Congressional Budget Office, 2008). While it is important not to overgeneralize from these ecological studies, a wide variety of research strongly suggests that differential access to and use of personal health services is but one of myriad factors contributing to social inequalities in health status (Roberts and House, 2000; MacIntyre, 1997).

With respect to developing countries, it is clear that economic development, growth in formal education, and public health measures such as clean water and basic nutrition are of greater significance to population health than the delivery of health care. While the importance of personal health services (such as immunizations, HIV medications, or oral rehydration therapy) in less-developed countries should not be underestimated, broad public health and economic development advances are believed to afford the most potential for making significant improvements in health.

This is not to suggest that health care has no impact on individual health or social disparities in health status. There are indeed some diseases and conditions for which health services can significantly reduce morbidity and mortality, and improve physical functioning and quality of life, as demonstrated in clinical trials and other types of studies (Starfield, 2012; Kates, 2011; Andrulis, 1998). The bulk of the knowledge generated to date, however, suggests that while reducing SES disparities in health care may help to reduce socioeconomic and racial disparities for some specific diseases or health issues, significant inequalities in health status would be likely to remain. Environmental, community, socioeconomic, and psychosocial exposures to both health risks and health-promoting resources would still exist, creating disparities in the incidence and prevalence of a number of health problems. Furthermore, many health conditions and problems are not amenable to medical interventions or to prevention practices delivered through health care providers.

It is also important to emphasize that socioeconomic disparities in health care are not immutable (Fiscella et al., 2000). Thus, many researchers, clinicians, advocates, and policymakers alike believe that efforts should continue to be made across and within countries to reduce SES differentials in health care access and utilization by focusing on the financial, attitudinal, and most importantly, health system barriers that contribute to these significant disparities. The policy goal should be a health care delivery system in which all dimensions of access and the content and quality of care delivered are the same across social strata.

See also: Health Insurance: Economic and Risk Aspects; Social Epidemiology; Socioeconomic Status and Health; Unemployment and Mental Health.

Bibliography

Agency for Healthcare Research and Quality, 2012a. 2011 National Healthcare Disparities Report. US Department of Health and Human Services, Rockville, Maryland.

Agency for Healthcare Research and Quality, 2012b. 2011 National Healthcare Quality Report. US Department of Health and Human Services, Rockville, Maryland.

Adler, N.E., Newman, K., 2002. Socioeconomic disparities in health: pathways and policies. Health Affairs 21 (2), 60–76.

Adler, N.E., Boyce, W.T., Chesney, M.A., Folkman, S., Syme, L., 1993. Socioeconomic inequalities in health: no easy solution. Journal of the American Medical Association 269 (24), 3140–3145.

Andrulis, D.P., 1998. Access to care is the centerpiece in the elimination of socioeconomic disparities in health. Annals of Internal Medicine 129 (5), 412–416.

Baicker, K., Taubman, S.L., Allen, H.L., Bernstein, M., Gruber, J.H., Newhouse, J.P., Schneider, E.C., Wright, B.J., Zaslavsky, A.M., Finkelstein, A.N., 2013. The Oregon experiment – effects of Medicaid on clinical outcomes. New England Journal of Medicine 368 (18), 1713–1722.

Baker, G., Newsom, E., 2003. Socioeconomic status and dissatisfaction with health care among chronically ill African Americans. American Journal of Public Health 93 (5), 742–748.

Blustein, J., 1998. Preventable hospitals and socioeconomic status. Health Affairs 17 (2), 177–189.

Brown, M.E., Bindman, A.B., Lurie, N., 1998. Monitoring the consequences of uninsurance: a review of methodologies. Medical Care Research and Review 55 (2), 177–210.

Collins, S.R., Robertson, R., Garber, T., Doty, M.M., 2012. The Income Divide in Health Care: How the Affordable Care Act Will Help Restore Fairness to the US Health Care System. The Commonwealth Fund, New York.

Congressional Budget Office, 2008. Growing Disparities in Life Expectancy. Economic and Budget Issue Brief. Available from: http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/91xx/doc9104/04-17-lifeexpectancy_brief.pdf (12 June 2013).

Dutton, D.B., 1978. Explaining low use of health services by the poor: costs, attitudes, or delivery systems? American Sociological Review 43, 348–368.

Fiscella, K., 2011. Health care reform and equity: promise, pitfalls and prescriptions. Annals of Family Medicine 9 (1), 78–84.

Fiscella, K., Franks, P., Gold, M.R., Clancy, C.M., 2000. Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. Journal of the American Medical Association 283 (19), 2579–2584.

Gold, M., 1998. Beyond coverage and supply: measuring access to healthcare in today's market. Health Services Research 33 (3 part 2), 625–652.

Gray, A.M., 1982. Inequalities in health: the Black report: a summary and comment. International Journal of Health Services 12 (3), 349–380.

Halldorsson, M., Kunst, A., Kohler, L., Mackenback, J., 2002. Socioeconomic differences in children's use of physician services in the Nordic countries. Journal of Epidemiological Community Health 56 (3), 200–204.

Institute of Medicine, 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. The National Academies Press, Washington, DC.

Institute of Medicine, 2002. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. The National Academies Press, Washington, DC.

Kates, J., 2011. Medicaid and HIV: A National Analysis. The Henry J. Kaiser Family Foundation, Menlo Park, CA.

Kephart, G., Thomas, V.S., MacLean, D.R., 1998. Socioeconomic differences in the use of physician services in Nova Scotia. American Journal of Public Health 88 (5), 800–803.

Lantz, P.M., Weigers, M.E., House, J.S., 1997. Education and income differentials in breast and cervical cancer screening. Medical Care 35 (3), 219–236.

Lillie-Blanton, M., Hoffman, C., 2005. The role of health insurance coverage in reducing racial/ethnic disparities in health care. Health Affairs 24 (2), 398–408.

MacIntyre, S., 1997. The Black report and beyond: what are the issues? Social Science and Medicine 44 (6), 723–745.

Mangione-Smith, R., DeCristofar, A.H., Setodji, C.M., Keesey, J., Klein, D.J., Adams, J.L., Schuster, M.A., McGlynn, E.A., 2007. The quality of ambulatory care delivered to children in the United States. New England Journal of Medicine 357 (15), 1515–1523.

McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., Kerr, E.A., 2003. The quality of health care delivered to adults in the United States. New England Journal of Medicine 348 (26), 2635–2645.

National Center for Health Statistics, 2012. Health, United States, 2011, With Special Feature on Socioeconomic Status and Health. Hyattsville, Maryland.

Niewczyk, P.M., Lwebuga-Mukasa, J.S., 2008. Is poverty the main factor contributing to health care disparities? An investigation of individual level factors contributing to health care disparities. Journal of Equity in Health 1 (1), 22–40.

Pappas, G., Hadden, W.C., Kozak, L.J., Fischer, G.F., 1997. Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups. American Journal of Public Health 87 (5), 811–816.

Roberts, S.A., House, J.S., 2000. Socioeconomic inequalities in health: an enduring sociological problem. In: Bird, C.E., Conrad, P., Fremont, A.M. (Eds.), Handbook of Medical Sociology, fifth ed. Prentice Hall, Upper Saddle River, NJ, pp. 79–97.

Rosenbach, M.L., Irvin, C., Coulam, R.F., 1999. Access for low-income children: is health insurance enough? Pediatrics 103 (6 part 1), 1167–1174.

Rosenbaum, S., 2011. The Patient Protection and Affordable Care Act: implications for public health policy and practice. Public Health Reports 126 (1), 130–135.

Ross, J.S., Bradley, E.H., Busch, S.H., 2006. Use of health care services by lower-income and higher-income uninsured adults. Journal of the American Medical Association 295 (17), 2027–2036.

Sommers, B.D., Baicker, K., Epstein, A.M., 2012. Mortality and access to care among adults after state Medicaid expansions. New England Journal of Medicine 367 (11), 1025–1034.

Starfield, B., 2012. Primary care: an increasingly important contributor of effectiveness, equity, and efficiency of health services. SESPAS report 2012. Gaceta Sanitaria 26 (Suppl. 1), 20–26.

Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review

Beverly Gaudet, Faculty of Nursing, University of New Brunswick, Fredericton, NB, Canada
Nicole Letourneau, Faculties of Nursing and Medicine (Pediatrics and Psychiatry), University of Calgary, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While it is widely accepted that socioeconomic status (SES) has an impact on the development of children, it is critical that we define SES beyond a narrow focus on income. It is also imperative to understand its impact on a specific dimension of early childhood development, socioemotional development, beginning at an early age. Thus, the objective of this article was to conduct a systematic narrative review to examine the relationship between broader conceptions of SES and young children's social-emotional development. The findings suggest that SES begins to have a negative impact on social-emotional development early in life.

Childhood development and socioeconomic status (SES) are both identified by the Public Health Agency of Canada (2009) to be determinants of health that, individually or in combination, influence the health of individuals. Low SES is likely to negatively affect an individual's future prospects, opportunities, and quality of life (Hayes, 2008; Public Health Agency of Canada, 2009). According to Bradley and Corwyn (2002), the link between SES and social-emotional development in children has not been identified as consistently as the connection between SES and cognitive development in children, particularly in language and executive functions, which emerge in preschool and older childhood (Hackman and Farah, 2009). However, the period from birth to age of 3 years has been described as critical for physical, social-emotional, and cognitive development over the life span due to heightened sensitivity of the brain and other developing organ systems to environmental influences. Further, social-emotional development in infancy, specifically associated with brain development and mental health over the life span, may be the most important feature of development in early childhood (Center on the Developing Child at Harvard University, 2010; Siddiqi et al., 2007).

Cognitive development may in fact depend on social-emotional development, as managing emotions may be necessary for learning difficult tasks such as counting or reading (National Scientific Council on the Developing Child, 2011). In this way, the foundations of developmental health in all domains are linked to emotional well-being and have an impact on a child's future ability to maintain a job and work well with others and for practicing effective citizenship and contribute to a community (National Scientific Council on the Developing Child). Taken together, the sensitivity of young children's development to environmental influences is particularly worrisome when their families struggle with low SES. The early social environment is critical in influencing the formation of neural networks that regulate a child's social-emotional development, including response to stress and capacity for self-control (DiPietro, 2000; Hertzman, 1999; Irwin et al., 2007; Leventhal and Brooks-Gunn, 2000; McDonald, 2008; National Scientific Council on the Developing Child, 2011; Ridge, 2002).

Understanding the effects of SES on young children's social-emotional development has been complicated by a variety of factors, including (1) inconsistent measurement of SES across studies; (2) use of single indicators of SES (e.g., employment) rather than broader conceptions of SES; (3) inclusion of SES as a control variable rather than actually examining the impact of SES on social-emotional development; (4) low SES is an inclusion criteria for studies examining other risk factors such as maternal depression, maternal drug use, and quality of the home environment, so comparisons among SES groups are precluded; (5) little research has isolated SES as a specific indicator of social-emotional development in infants and children; and finally, (6) SES is often related to outcomes in older children such as those of school age. As a result, the link between SES and social-emotional development in young children remains unclear.

To more accurately and effectively affect policy and intervention programs, it is important to isolate at what point SES begins to impact social-emotional development in children. Poverty in early childhood may be more detrimental to long-term behavioral and achievement outcomes than if poverty is experienced during middle-childhood or adolescence (Duncan and Brooks-Gunn, 2000). In spite of the recognition that the first 3 years of life represent a period of particularly rapid development of cortical structure and function (DiPietro, 2000), very little work has rigorously explored the influence of poverty, let alone broader conceptions such as SES, on young children's social-emotional development. Thus, a review was undertaken to examine the influence of SES on young children's (from birth to 36 months of age) social-emotional development.

What Is Social-Emotional Development in Young Children? Why Is It Important?

Shonkoff and Phillips (2000) describe social-emotional development as the capacity to identify one's own feelings and the ability to constructively manage strong emotions. A child's ability to self-regulate in early development is thus interconnected with the child's relations with others.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23053-5

Social-emotional development also represents the capacity for children to develop and learn from relationships. For example, a toddler learns that the people he or she depends on for comfort will help when she or he is distressed; therefore, the toddler will be more likely to approach others with trust and expectations of empathy.

Social-emotional development is positively correlated with the growth of new skills and abilities in young children; thus, when children are unable to manage their feelings, their thinking can be impaired (National Scientific Council on the Developing Child, 2011). Birth to age of 3 years is a time to develop relationships, regulate and express emotions, form close and secure interpersonal relationships, and explore the environment and learn – all in the context of family, community, and cultural norms for young children (Fitzgerald et al., 2011; Siddiqi et al., 2007). In summary, infant mental health is directly related to healthy social and emotional development (DiPietro, 2000), and early social-emotional development provides the foundation for psychosocial well-being and mental health over the life span (Shonkoff and Phillips, 2000).

Social-emotional problems in infancy and young childhood reduce older children's capacity to forge healthy relationships with peers and adults, which then impedes fundamental cognitive, linguistic, and self-regulating abilities. In this way, early challenges with social-emotional development predict more serious mental health issues in later childhood, adolescence, and even adulthood (Bricker et al., 2004); as a result, by school age, many opportunities for intervention will have passed. In order to more accurately inform social policy and programs, it is critical to identify at what age low SES begins to have a negative impact on children's social-emotional development. Knitzer (2008) argues that early detection and treatment will prevent children and adolescents from becoming frequent users of the health-care system, thus reducing health-care spending over the long term. Bricker et al. (2004) posit that treating young children is less expensive and perhaps more effective than postponing intervention until middle-childhood or adolescence. Indeed, treatment becomes less effective when social-emotional problems become deeper rooted over time.

How Do We Define SES?

Poverty as a measure of SES has been criticized for its narrow focus on income (Bradley and Corwyn, 2002; Hackman and Farah, 2008; U.S. Department of Health and Human Services, 2000). SES includes not only income but also other socio-demographic variables that more fully characterize the social and economic environment for children. Moreover, a narrow focus on poverty as a measure of SES necessarily limits interventions to income support and precludes consideration of broader factors that could become the focus for interventions with low-SES families. As a result, increasingly SES is considered more broadly and includes not only income but also parental education, marital status, neighborhood characteristics, and child-care quality (Bradley and Corwyn, 2002). Much of the research done to date refers only to income (from all sources, including government subsidy) or income-to-needs ratio as a measure of SES. Stepping beyond income as the focus of SES gives a clearer illustration of someone's financial situation, as SES should include nonmaterial resources gained through social connections such as education (Bradley and Corwyn, 2002). Composite definitions of SES (using multiple indicators of SES) increase the validity and generalizability of findings. Thus, for the purposes of this article, rather than using a single indicator of SES such as income, SES was conceptualized more broadly. SES was defined using a composite measure of at least two of the following: parental education level, parental marital status, parental employment status, parental occupation prestige, and household income and eligibility for subsidy (Ensminger and Fotherill, 2003).

Theoretical Underpinnings of Review

Two theories guide the focus of this article. First, Erikson's theory of psychosocial development (1963) is based on qualities of the social environment that are required to meet a child's basic needs. Between birth and 1 year of age, children must establish trusting relationships with caregivers. If basic needs (e.g., for nourishment and comfort) are not met, then the young child learns to mistrust others with implications for mental health over the life span. Between the age of 24 and 36 months, young children must establish their autonomy in relationships with others. Children need to develop a sense of personal control over physical skills and gain a sense of independence; failure to achieve these skills results in feelings of shame and doubt with additional mental health implications over the life span. Second, Bronfenbrenner's ecological model of human development (1994) emphasizes the entire ecological system to understand human development. Early and throughout life, human development occurs through a series of complex reciprocal interactions between the individual and his or her immediate environment. Infants do not experience low SES directly, but only through their interactions with caregivers. Low-SES caregivers may be strained, stressed, and distressed in coping with the low-SES conditions, which may affect interactions with their infant. Further, young children may not receive adequate nutrition or stable housing from their low-SES caregivers, who are limited in what they can provide. The family environment provides children's first experiences and human contact; thus, families are most often the ones to mediate a child's contact with their larger environment (Kohen et al., 2002). Thus, Bronfenbrenner requires that we take a broader view of influences on children's development.

In summary, the purpose of this article was to conduct a systematic narrative review of the relationship between composite measures of SES and social-emotional development in children between ages of 0 to 36 months. The secondary purpose was to examine the influence of moderating variables.

Methods

Due to the complexity of issues surrounding the impact of SES on socioemotional development, a systematic narrative review was conducted to examine the relationship between SES and social-emotional development in young children. A systematic narrative review adheres to a set of scientific methods to prevent the limitations associated with bias by attempting to

assess and synthesize all relevant evidence to answer a clear, targeted question (Mays et al., 2005). Systematic narrative reviews seek to explore the diversity and variety of literature to be more comprehensive in identifying the strengths of evidence. The intent is to lend support to the research question or identify areas of uncertainty or gaps where more research is needed (Mays et al., 2005) in order to inform research, practice, and policy.

Search Strategy

Papers discussing the impact of SES on social-emotional development were searched electronically through CINAHL, MEDLINE, and PsycINFO – the databases most likely to yield relevant research. The following keywords were used: SES, socioeconomic factors, social class, family background, career mobility, poverty, employment, occupation prestige, household income, parental characteristics, parental education, marital status, emotional development, social development, social-emotional development, internalizing behaviors, depression, anxiety, infant development, and child development. Consistent with the characteristics of exemplary literature reviews, the ancestry approach was used to scan the reference lists from the retrieved articles to identify other relevant citations (Jackson, 1980).

Criteria

Inclusion of studies was limited to those published during the last 10 years (2002–12) to reflect a clearer, more contemporary understanding of SES and its associated implications. The search was further limited by including only those (1) written in English, (2) on human subjects, (3) on children 0–36 months old, (4) on children born healthy at full term, (5) on children with no organic health conditions or developmental disorders (e.g., autism, attention deficit hyperactivity disorder), and (6) on healthy parents. Children and parents with disorders (e.g., mental illness or developmental problems) were excluded from the search, as the review sought to examine social risks, not physical or biological risks. The final set of papers was achieved by consensus of the two authors. **Figure 1** summarizes the review process.

Results

A total of four studies were included; all were longitudinal birth cohort studies. The studies were conducted in Australia ($n = 1$), the United Kingdom ($n = 2$), and the United States ($n = 1$). The total sample size was $n = 31\ 381$. The sample was racially diverse; Flouri et al. (2010) described their sample as representative of families living in poverty in Northern Ireland, Scotland and Wales, and areas in England with populations of high ethnic minorities. The sample in the study done by Kiernan and Huerta (2008) came from areas with high proportions of ethnic minorities in the UK. Nicholson et al. (2012) described their samples as representative of Australia's population, including children of highly educated parents

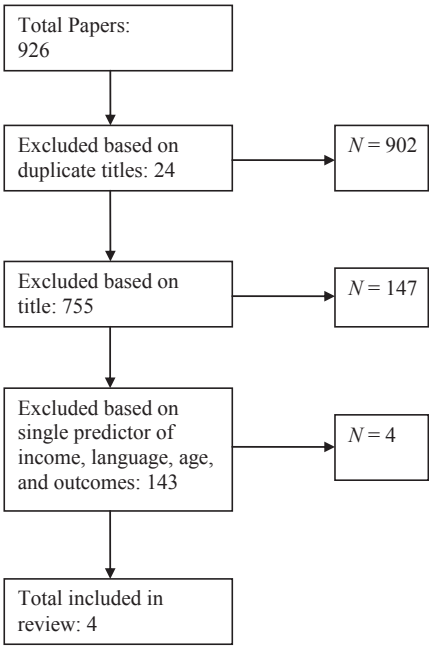

Figure 1 Search Results

overrepresented by 10%, and children who had single parents, were non-English speaking, and were living in rental properties were slightly underrepresented. Zilanawala and Pilkauskas's (2012) study contained an over sample of nonmarital births in large U.S. cities with populations over 200 000. As a result, their sample represented families who were more likely to be with low income, with absent fathers, and had mothers with a lower level of education than a sample representative of the nation. About one-third of mothers were black (35%), just under one-third of the sample were Hispanic mothers, and approximately one-quarter of the sample were white mothers.

The measurement of SES within these four studies varied. Flouri et al. (2010) used two composite measures of SES, including the Index of Multiple Deprivation (IMD) and family SES. The IMD assessed income deprivation, employment deprivation, health deprivation and disability, education, skills, and training deprivation, barriers to housing and services, living arrangements, and crime. The family SES measures examined overcrowding, lack of home ownership, receipt of income support, income poverty (below the poverty line), and lack of access to a car or van. Kiernan and Huerta (2008) used the measures of income poverty, financial difficulties, and housing tenure, i.e., owning or renting a home versus living in publicly subsidized housing. Nicholson et al. (2012) defined SES by the combination of parental income, education, and occupational prestige. Zilanawala and Pilkauskas (2012) defined SES as ability to pay utilities, food, and medical bills and housing instability. They also included a second measure of SES that included the household's income-to-needs ratio, education status, and employment status.

Of the four studies reviewed, all found a significant relationship between SES and internalizing behaviors in children from birth to 36 months of age. **Table 1** summarizes the findings of the studies. Two of the four studies found evidence

Table 1 Overview of search results

Author(s)/country	*Purpose*	*Design*	*Key measures*	*Findings and moderation*
Flouri et al. (2010) Children of 3 years of age in nine strata in the UK including areas with high ethnic minority populations as well as areas of child poverty	To determine the area and family contextual risk factors on young children's psychopathology	Longitudinal cohort study	• Area SES: income, employment, health, education, barriers to housing/services, living environment, and crime • Family SES: overcrowding, lack of home ownership, receipt of income support, income poverty, lack of access to a car • Emotional symptoms, peer problems, prosocial behavior *Child-level covariates*: age, sex, developmental milestones, and temperament (9 months)	• Family SES predicted prosocial behavior ($b = 0.043$, SE $= 0.021$) and peer problems ($b = 0.050$, SE $= 0.017$) • No moderation of relationship between SES and social-emotional development examined; however, the relationship between family SES and psychopathology more generally was moderated by child characteristics (e.g., age, sex) and family adversity
Kiernan, K.E. and Huerta, M.C. (2008) Children of 3 years of age from four countries in the UK, living in areas with high rates of child poverty and in areas with high proportions of ethnic minorities	To examine the extent to which SES in infancy and mother's mental health are associated with children's emotional development at 3 years of age	Longitudinal cohort study	• Economic deprivation: Income, financial difficulty, and housing tenure • Background controls: Marital status, maternal education level, maternal employment status at age of 9 months • Internalizing problems	• Findings revealed that children living in low-SES families have a higher risk of experiencing internalizing behavioral problems ($d = 0.18$, $p < .001$) • Children whose mothers are depressed tend to experience more internalizing behaviors ($d = 0.13$, $p < .001$). Total direct ($d = 0.10$, $p < .001$) and indirect ($d = 0.03$, $p < .001$) effect of SES on internalizing problems was explained through maternal depression and parenting practices. Affect sizes are small but statistically significant. A positive relationship between mother and child as well as parental disciplinary practices are related to fewer emotional difficulties in children ($d = 0.02$, $p < .001$ and $d = 0.01$, $p < .001$, respectively)
Nicholson et al. (2012) Children of 0 to 3 years of age broadly represent the Australian population	To determine at what age physical and developmental health outcomes are impacted when children experience early and persistent low SES	Longitudinal cohort study	• SES using a composite variable: Parental income, education, and occupational prestige • Developmental health indicators: Socioemotional competence, socioemotional difficulties (measured at 2–3 years of age)	• Identified socioeconomic groups 1 to 5 with 1 being the least disadvantaged and 5 being the most disadvantaged. Social-emotional difficulties measured at 2–3 years of age, OR 2.7 (CI 2.1–3.5) for group 5 compared to 1.3 (CI 1.0–1.8) for group 2 ($p < .001$) • No moderation of relationship between SES and social-emotional development examined
Zilanawala, A. and Pilkauskas, N.V. (2012) Children of 3 years of age in 20 large cities in the USA. Sample was racially diverse with mainly low-income mothers	Is material hardship associated with child socioemotional behavior? Are there differences by developmental timing? Are particular hardships more strongly associated with child behavior? Are there differences in the association between short- and long-term material hardship and child behavior?	Longitudinal cohort study	• Material hardship: Utilities, food and medical bills, housing instability • Covariates: Household income-to-needs ratio, mother's education, employment status, parental marital status • Internalizing behaviors	• Children experiencing material hardship at 3 years of age are associated with higher levels of internalizing behaviors such as anxiety, depression, and withdrawal with SD 6.04 ($p < .001$) • Mother's impulsivity and maternal depression are both strongly associated with higher internalizing scores for 3-year-old children ($b = 0.33$, $p < .001$ and $b = 0.15$, $p < .05$, respectively)

of moderation of the relationship between SES and social-emotional development. Flouri et al. (2010) utilized three measures of temperament for children at age of 9 months and again at age of 3 years. Broad and specific psychopathology was assessed with reporting by the child's main caregiver using the 25-item Strengths and Difficulties Questionnaire (Goodman, 1997). Three dimensions were measured in the Millennium Cohort Study using the Carey Infant Temperament Scale (Carey and McDevitt, 1978); these include mood, adaptability to different situations, and regularity of physiological functions. Findings revealed no relationship between the geographic-area IMD and prosocial behavior, emotional problems, or peer problems. However, the effect of family SES on social-emotional development (prosocial and peer problems) was significant. Their analysis revealed no moderation of the relationship between SES and social-emotional development; however, the relationship between family SES and psychopathology more generally was moderated by child characteristics (e.g., age, sex) and family adversity.

Kiernan and Huerta (2008) assessed behavioral adjustment at age of 3 years using the 25-item Strengths and Difficulties Questionnaire (Goodman, 1997). This questionnaire assesses behaviors associated with emotional symptoms or internalizing behaviors such as often unhappy or tearful, nervous or clingy in new situations, easily loses confidence, and easily scared. Findings revealed that children living in low-SES families have a higher risk of experiencing internalizing behavioral problems (0.18 *SD*; $p < .001$). Children whose mothers are depressed tend to experience more internalizing behaviors ($d = 0.13$, $p < .001$). Total direct ($d = 0.10$, $p < .001$) and indirect ($d = 0.03$, $p < .001$) effect of SES on internalizing problems was explained through maternal depression and parenting practices. Affect sizes are small but statistically significant. A positive relationship between mother and child as well as parental disciplinary practices are related to fewer emotional difficulties in children ($d = 0.02$, $p < .001$ and $d = 0.01$, $p < .001$, respectively).

Nicholson et al. (2012) used the Brief Infant-Toddler Social and Emotional Assessment Problems scale as a tool to identify children at risk for developing social-emotional problems and those who are currently experiencing social-emotional difficulties (Briggs-Gowan et al., 2004). This scale is based on parental report of a 31-item scale of internalizing behaviors (anxiety and depression) in children from 12 to 35 months and 30 days old. This scale includes items such as restlessness, looks for a parent when upset, and has less fun than other children. Nicholson et al. (2012) identified socioeconomic groups 1–5, with 1 being the least disadvantaged and 5 being the most disadvantaged. Social-emotional difficulties measured at ages 2 to 3 years indicated the odds ratio (OR) was 2.7 (CI 2.1–3.5) for group 5 compared to 1.3 (CI 1.0–1.8) for group 2 ($p < .001$). In other words, greater disadvantage was associated with higher odds of poor social-emotional development. Group 5 (most disadvantaged) showed an OR (odds ratio) of 2.7 (2.1–3.5) for 2–3 year olds compared to OR 1.3 (1.0–1.8) for the least disadvantaged group. There was no moderation of the relationship between SES and social-emotional development in young children.

Zilanawala and Pilkauskas (2012) conducted interviews at birth and when the children were 1 and 3 years old and an in-home study when the children were 3 years old. They utilized the Child Behavior Checklist (Achenbach, 1992) to assess anxious/depressed and withdrawn behaviors such as sadness, nervousness, worry, affection, and interest. Moreover, children's positive behaviors were assessed using the abbreviated Adaptive Social Behavior Inventory (Hogan et al., 1992) at 3 years of age, including social competence and prosocial skills with adults and peers. They reported that children whose mothers report any material hardship had significantly higher levels of internalizing behaviors. Children experiencing material hardship at 3 years of age are associated with higher levels of internalizing behaviors such as anxiety, depression, and withdrawal with *SD* 6.04 ($p < .001$). Mother's impulsivity and maternal depression are both strongly associated with higher internalizing scores for 3-year-old children ($b = 0.33$, $p < .001$ and $b = 0.15$, $p < .05$, respectively).

Discussion

Four longitudinal birth cohort studies focusing on the effects of SES on social-emotional development of young children fulfilled the search criteria of this article. While this provided a small set of papers for inclusion, the samples represented in the papers included over 31 000 children and only included studies with broader conceptions of SES than income. While all papers relied on parental report of children's social-emotional development with the addition of direct child assessment or teacher report by Nicholson et al. (2012), it is difficult to study children's development in isolation of parenting and the parent–child relationship (Shonkoff and Phillips, 2000). Low SES is directly related to stress, and stress invariably impacts parenting and the home environment (Linver et al., 2002; McLoyd, 1998). Evidence suggests that the emotional environment of the home often influenced by complexities such as marital dysfunction, negative parent–child relationships, coercion or abuse, or mental illness in the parent affects how children learn to interpret and express their emotions. As a result of these complexities, the parent may not be readily available to help children manage often confusing and overwhelming emotions they may experience (Bradley and Corwyn, 2002; National Scientific Council on the Developing Child, 2011). Consistent with these observations, this article revealed moderation of the relationship between SES and social-emotional development by maternal depression, maternal impulsivity, and qualities of the parenting environment such as parent–child relationships. Findings point to implications in two areas, including the need for policy development reflecting early intervention and income subsidy, both as means to limit the effects of low SES on the social-emotional development of young children. As well, the home environment must be considered as a source of influence on social-emotional development in young children.

Policy Development: Early Intervention and Income Subsidy

All four studies support the need for policy development in two critical areas: early intervention and income subsidy. Nicholson et al. (2012) advocated for effective early intervention

programs aimed at reducing inequalities in early childhood. Unintentional benefits may occur as a result of targeting even one area of childhood development. For example, a child may experience improvements in social development when the intervention is focused on literacy. Kiernan and Huerta (2008) emphasized that without early intervention, early childhood adversity is likely to have long-lasting effects on children's social-emotional development. Flouri et al. (2010) suggested their findings – young children who experience family adversity are at a higher risk of emotional difficulties – can be used to inform early interventions.

Nicholson et al. (2012) and Zilanawala and Pilkauskas (2012) support the notion that raising income or providing in-kind assistance alone will lead to beneficial social-emotional development outcomes in children. Zilanawala and Pilkauskas suggest there is potential for a return on public investments if material hardship for low-income families (such as food insecurity and housing insecurity) is improved through cash transfer programs when children are in the early stages of development.

Home Environment

The findings of Zilanawala and Pilkauskas (2012) suggest that while income alone does not significantly predict social-emotional development (anxiety and depressive characteristics), broader characteristics of the home setting such as food insecurity and housing insecurity are associated with poor social-emotional development in children. Flouri et al. (2010) also determined that family SES significantly predicted not only social-emotional development in young children but also child psychopathology. Moreover, family SES predicted development directly and independently of maternal education and family adversity; however, when child-level characteristics (such as gender and age) were added to models, the effect of family SES on emotional symptoms became nonsignificant. Adversity within the family was also found to be a stronger predictor of child psychopathology than family SES, even when child-level characteristics were added (Flouri et al., 2010). These findings suggest that early interventions could target characteristics of the home environment. Kiernan and Huerta (2008) reinforced the critical impact of the home environment, as their results indicated children living in low-SES families have a higher risk of experiencing internalizing behaviors; however, a positive relationship between mother and child and parental disciplinary practices may mediate emotional difficulties.

In conclusion, while it is well known that chronic exposure to adversity has a significant impact on all forms of development (Edwards et al., 2003; Ladd et al., 2000), this article focused on the earliest stage of social-emotional development, which underpins other forms of development. Clearly, findings revealed that SES (composite definition) begins to have a negative impact on social-emotional development early in life. From birth to age of 3 years is a time to develop relationships, regulate and express emotions, form close and secure interpersonal relationships, and explore the environment and learn – all in the context of family, community, and cultural expectations for young children (Fitzgerald et al., 2011). Embedded in the interpretation of SES and socioemotional health is a cultural context that varies across and within cultures that cannot be ignored. However, there are fundamental standards central to child development regardless of culture (Maggi et al., 2010). Infant mental health, which sets the stage for future mental health, is shaped by environmental exposures during this critical and sensitive period of development (DiPietro, 2000; Siddiqi et al., 2007).

The small set of papers uncovered in the review suggests a need for more study to examine the relationship between broader conceptions of SES and young children's social-emotional development within the context of cultural competence. Perhaps even more important is the utilization of randomized control trials and the examination of moderating variables that can provide guidance for interventions and policy to address the impact of low SES on children's social-emotional development globally.

See also: Academic Performance, Effects of Socio-Economic Status on; Behavioral Problems, Effects of Socio-Economic Status on; Childhood and Adolescence: Developmental Assets; Culture and Emotion; Early Emotional Development and Cultural Variability; Emotional Development, Effects of Parenting and Family Structure on; Emotions and Work; Language Development: Influence of Socio-Economic Status; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Social and Emotional Development in the Context of the Family; Socioeconomic Status and Academic Achievement; Socioeconomic Status and Health; Urban Poverty in Neighborhoods.

Bibliography

Achenbach, T.M., 1992. Manual for the Child Behavior Checklist/2–3 and 1992 Profile. University of Vermont Department of Psychiatry, Burlington, VT.

Bradley, R.H., Corwyn, R.F., 2002. Socioeconomic status and child development. Annual Review of Psychology 53, 371–399.

Bricker, D., Schoen Davis, M., Squires, J., 2004. Mental health screening in young children. Infants and Young Children 17, 129–144.

Briggs-Gowan, M.J., Carter, A.S., Irwin, J.R., Wachtel, K., Cicchetti, D.V., 2004. The brief infant-toddler social and emotional assessment: screening for social-emotional problems and delays in competence. Journal of Pediatric Psychology 29, 143–155.

Bronfenbrenner, U., 1994. Ecological models of human development. In: International Encyclopedia of Education, second ed., vol. 3. Elsevier, Oxford.

Carey, W., McDevitt, S., 1978. Revision of the infant temperament questionnaire. Pediatrics 61, 735–739.

Center on the Developing Child at Harvard University, 2010. The Foundations of Lifelong Health Are Built in Early Childhood. Harvard University, Boston, MA.

DiPietro, J.A., 2000. Baby and the brain: advances in child development. Annual Review of Public Health 21, 455–471.

Duncan, G.J., Brooks-Gunn, J., 2000. Family poverty, welfare reform, and child development. Child Development 71, 188–196.

Edwards, V., Holden, G., Felitti, V., Anda, R., 2003. Relationship between multiple forms of childhood maltreatment and adult mental health in community respondents: results from the adverse childhood experiences study. American Journal of Psychiatry 160, 1453–1460.

Ensminger, M., Fotherill, K., 2003. A decade of measuring SES: what it tells us and where to go from here. In: Bornstein, M., Bradley, R. (Eds.), SES, Parenting, and Child Development. Lawrence Erlbaum, Mahwah, NJ, pp. 13–28.

Erikson, E., 1963. Childhood and Society. W.W. Norton, New York.

Fitzgerald, H.E., Weatherston, D., Mann, T.L., 2011. Infant mental health: an interdisciplinary framework for early social and emotional development. Current Problems in Pediatric and Adolescent Health Care 41, 178–182.

Flouri, E., Mavroveli, S., Tzavidis, N., 2010. Modeling risks: effects of area deprivation, family socio-economic disadvantage and adverse life events on young children's psychopathology. Social Psychiatry and Epidemiology 45, 611–619.

Goodman, R., 1997. The strengths and difficulties questionnaire: a research note. Journal of Child Psychology and Psychiatry 38, 581–586.

Hackman, D.A., Farah, M.J., 2009. SES and the developing brain. Trends in Cognitive Sciences 13, 65–73.

Hayes, A., 2008. Social mobility as the engine of inclusion. Australian Institute of Family Studies Seminar Series. http://www.aifs.gov.au/institute/seminars/2008/ahayes.html (accessed 30.11.12.).

Hertzman, C., 1999. The biological embedding of early experience and its effects on health in adulthood. Annals of New York Academy of Science 896, 85–95.

Hogan, A.E., Scott, K.G., Bauer, C.R., 1992. The adaptive social behavior inventory (ASBI): a new assessment of social competence in high-risk five year olds. Journal of Psychoeducational Assessment 10, 230–239.

Irwin, L.G., Siddiqi, A., Hertzman, C., 2007. Early Child Development: A Powerful Equalizer. Final Report for the WHO's Commission on Social Determinants of Health. World Health Organization, Commission on Social Determinants of Health, Geneva, Switzerland.

Jackson, G.B., 1980. Methods for integrative reviews. Review of Educational Research 50, 438–460.

Kiernan, K.E., Huerta, M.C., 2008. Economic deprivation, maternal depression, parenting and children's cognitive and emotional development in early childhood. The British Journal of Sociology 59, 783–806.

Knitzer, J., 2008. Giving infants and toddlers a head start: getting policies in sync with knowledge. Infants and Young Children 21, 18–29.

Kohen, D.E., Brooks-Gunn, J., Leventhal, T., Hertzman, C., 2002. Neighbourhood income and physical and social disorder in Canada: associations with young children's competencies. Child Development 73, 1844–1860.

Ladd, C.O., Huot, R.L., Thrivikraman, K.V., Nemeroff, C.B., Meaney, M.J., Plotsky, P.M., 2000. Long-term behavioral and neuroendocrine adaptations to adverse early experience. Progress in Brain Research 122, 81–103.

Leventhal, T., Brooks-Gunn, J., 2000. The neighbourhoods they live in: the effect of the neighbourhood residence on child and adolescent outcomes. Psychology Bulletin 126, 309–337.

Linver, M., Brooks-Gunn, J., Kohen, D., 2002. Family processes as pathways from income to young children's development. Developmental Psychology 38, 719–734.

Maggi, S., Irwin, L.G., Siddiqi, A., Hertzman, C., 2010. The social determinants of early childhood, development: an overview. Journal of Paediatrics and Child Health 46, 627–635.

Mays, N., Pope, C., Popay, J., 2005. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. Journal of Health Services Research and Policy 10, 6–20.

McDonald, C., 2008. Children's lived experience of poverty: a review of the literature. Australian Research Alliance for Children and Youth (ARACY). http://www.healthinfonet.ecu.edu.au/key-resources/organisations?oid=1103 (accessed 01.11.12.).

McLoyd, V.C., 1998. Socioeconomic disadvantage and child development. American Psychologist 53, 185–204.

National Scientific Council on the Developing Child, 2011. Young Children Develop in an Environment of Relationships: Working Paper No. 1. www.developingchild.harvard.edu (accessed 01.11.12.).

Nicholson, J.M., Lucas, N., Berthelsen, D., Wake, M., 2012. Socioeconomic inequality profiles in physical and developmental health from 0–7 years: Australian national study. Journal of Epidemiology and Community Health 66, 81–87.

Public Health Agency of Canada, 2009. What Determines Health?. http://www.phac-aspc.gc.ca/ph-sp/determinants/index-eng.php (accessed 01.11.12.).

Ridge, T., 2002. Childhood Poverty and Social Exclusion: From a Child's Perspective. Policy Press, Bristol.

Shonkoff, J.P., Phillips, D.A., 2000. From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academic Press, Washington, DC.

Siddiqi, A., Irwin, L.G., Hertzman, C., 2007. Total Environment Assessment Model for Early Child Development: Evidence Report for the WHO's Commission on Social Determinants of Health. World Health Organization, Commission on Social Determinants of Health, Geneva, Switzerland.

U.S. Department of Health and Human Services, 2000. Healthy People 2010. Washington, DC, USA.

Zilanawala, A., Pilkauskas, N.V., 2012. Material hardship and child socioemotional behaviours: differences by types of hardship, timing and duration. Children and Youth Services Review 34, 814–825.

Sociolinguistics

Gerard Van Herk, Memorial University of Newfoundland, St. John's, NL, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Holmes, volume 21, pp. 14562–14569, © 2001, Elsevier Ltd.

Abstract

Sociolinguistics is the study of the relationship between language and society. Sociolinguists may investigate linguistic variation correlated with social categories, style, or attitudes; the use of language in interaction to affirm the relative status of participants, norms of interaction and audience, or the social construction of self; the development of languages in contact; or the efforts of society to regulate linguistic choice through education or language policy and planning.

Sociolinguistics is the study of the relationship between language and society. As in any hyphenated or hybrid subdiscipline, work described as 'sociolinguistics' ranges along a continuum between two poles, which at their most extreme serve more as idealized points of reference than as widespread types of practice. Toward the linguistic end, analysts investigate the linguistic variation that results from determinants of language change that may or may not be socially driven (and such variationist researchers may suggest that their work should be called linguistics, not sociolinguistics). Research in this area is sometimes termed social (or urban) dialectology, and is most often identified with the theoretical and methodological school developed by William Labov. Toward the socio end, analysts may investigate the social meaning and relative status of different language varieties, without necessarily dealing with the linguistic structure of those varieties. A well-known early researcher in this area is Joshua Fishman, who has labeled such work 'the sociology of language.' Somewhere in the middle of the socio/linguistic continuum we find research on how language is used in interactions, an area that includes work on style shifting, discourse, and pragmatics. Journals and conferences devoted to the field can be positioned along the continuum, but often include research from a wide range of research approaches.

This article will attempt a similar inclusive approach. The first section provides a brief description of language-focused variationist work; the second describes some research on language use in interactions; the third considers the role of language in the social and sociological realm.

Social Determinants of Language Choice: Change, Variation, Community

Although language researchers have approached the topics of speech communities and the effect of social categories on language behavior for a century, an identifiable subfield called sociolinguistics truly emerged in the middle of the twentieth century. Fischer (1958) observed class, gender, and stylistic correlates to schoolchildren's use of alveolar (ING) (e.g., *swimmin'* instead of *swiiming*). But it is the pioneering work of William Labov, from *c.*1963 onward, that establishes the research questions, methodology, and analytical techniques of (variationist) sociolinguistics. In a series of influential, large-scale quantitative studies in Martha's Vineyard and New York City, Labov (1963, 1972) demonstrated that linguistic variation was not random ('free') or peripheral noise, as linguists had often assumed, but rather systematically patterned. The choice between two or more linguistic *variants* (ways of saying the same thing, referentially speaking) was affected by speakers' relevant social categories (e.g., age, social class, gender, or ethnicity), the formality of their speech style, and their attitudes toward the social groups associated with each variant. Furthermore, speakers shared linguistic norms, in terms of both overt attitudes (toward variants and their users) and a less overt and possibly subconscious sense of which variant was appropriate for particular speakers in a particular situation. Labov proposed the term *speech community* to describe large groups whose members shared such norms.

Much early work was designed to illuminate the regularities of non-mainstream speech, and to develop models to explain historical developments by observing language change as it occurred ('using the present to explain the past'). As technological advances eased natural recording and large-scale multivariate analysis, Labov and others developed methodological and analytical techniques and principles that are still in use today. These include interview techniques designed to elicit the most natural speech possible and to vary participants' degree of attention to their own speech (by recording speakers reading word lists or reading passages, in addition to participation in interviews). One powerful concept has been the *apparent time hypothesis*, the idea that differences between generations at one point in time reflect a snapshot of ongoing community-wide change. Subsequent work by Lesley and James Milroy in working-class Belfast developed social network techniques ('friend of a friend') to find and record community members whose frequent interactions encouraged linguistic similarities. They defined networks in terms of their denseness and multiplexity, showing that people who were most integrated into their local community networks were also the most likely to use more vernacular forms, and to resist imported language changes (Milroy, 1980). Yet more recent work has borrowed concepts from education and anthropology to adopt a more local focus, that of the community of practice (Eckert and McConnell-Ginet, 1999), investigating how norms are created and reinforced through interactions.

Variationist studies across communities and languages have replicated many of the field's original findings. Generally,

women lead language change and are more likely to use standard forms in situations of stable variation. The language of higher-status social classes is generally seen as desirable, and speakers of all classes tend to move toward that end of their repertoire in situations of heightened attention to speech. However, the countervailing appeal of working-class forms (*covert prestige*) leads to the maintenance of such variants, as speakers benefit from their associations with authenticity (Trudgill, 1972). Real-time studies have shown that most speakers maintain consistencies in speech across their lifespan, supporting the premise underlying the apparent time hypothesis, although some movement in the direction of incoming variants is often found (Blondeau et al., 2002). The social categories that influence speech (gender, ethnicity, social class, age, etc.) often interact in complex ways that may vary depending on the community (e.g., the sociolinguistic meaning or performance of gender may vary by ethnicity, or that of ethnicity may vary by gender). In addition, specific social traits – everything from occupation to political affiliation – may be relevant in particular communities, or at particular times. This is particularly evident in community of practice work, most notably in Eckert's (2000) influential sociolinguistic description of a suburban Michigan high school, where use of incoming vowel changes was highly dependent on how speakers situated themselves with respect to two school-specific social groups, the jocks (who reflected or respected mainstream or middle-class social norms) and the burnouts (who did not).

Variationists study variables that differ in their degree of salience to community members (e.g., stereotypes versus subtler sociolinguistic markers, which speakers can control when they shift linguistic styles, and yet subtler sociolinguistic indicators, which are beyond or below speaker control, but whose rate of use varies from one speaker to another). Variables can also be drawn from a range of linguistic domains, from tiny changes in vowel position (Labov et al., 2006) to the choice between discrete verbal markers or even lexical or discourse variables such as incoming quotative *be like* (as in *I was like, Oh my God!*) (Buchstaller and van Alphen, 2012).

In addition to the community norms revealed by language use, sociolinguists have investigated norms revealed through attitudes toward variants, languages, and speakers. The vernacular speech of New York City and the American South, for example, is widely stigmatized, especially by speakers of Midwestern American varieties (Preston and Howe, 1987). And covert linguistic attitudes have been revealed through matched guise techniques (in which naïve judges believe they are hearing different speakers, when in fact they are hearing the same speaker adopting the features of different language varieties). Across communities, researchers find that traits ascribed to speakers of standardlike varieties form a cluster associated with status (employability, intelligence, competence), while speakers of nonstandard or lower-status varieties are ascribed traits associated with solidarity or approachability (friendliness, warmth, authenticity) (Lambert et al., 1960). There are profound real-world applications of this work, as speakers of particular ethnic varieties are consistently rated as less suited to higher-status employment, or are denied access to housing (Purnell et al., 1999).

Language in Interaction: Status, Solidarity, Style

Another broad branch of sociolinguistics investigates how language is used in interactions, reflecting the relative status of participants, the norms of interaction and audience, and the social construction of self through linguistic choices and changes.

Status and Solidarity

The status/solidarity distinction described in language attitude work (see Section Social Determinants of Language Choice: Change, Variation, Community) also holds in studies of language in interaction. Many languages feature distinct variants that are expected when addressing an interlocutor of higher status or social distance. These can include *V forms* of the 'you' pronoun (e.g., *vous, Sie*), or the use of formal language, or the interlocutor's last name and/or title (*Dr Van Herk*). Other features are expected when speaking to somebody of lower status, or to whom the speaker is socially close. These include *T forms* (e.g., *tu, du*), informal language, and first-naming (*Gerard, dude!*) or nicknames (*G-man!*). The choice of forms expresses the participants' understanding of the power imbalance and/or intimacy of the relationship. Related work on politeness in speech (Brown and Levinson, 1987) identifies two competing goals of participants: to respect the right of the interlocutor to operate freely or be left alone (negative politeness), and to express rapport, solidarity, or friendship (positive politeness). The relative importance placed on each politeness type can vary by culture, leading to misunderstandings (as when the negative politeness strategies of Korean-American shopkeepers are interpreted as hostile by positive-politeness-valuing African-American customers) (Bailey, 1996). Societies may also differ, of course, in deciding which social boundaries are to be privileged in valuing status and solidarity.

Power imbalances are also crucial to work in Critical Discourse Analysis, whereby researchers investigate how ideology and power are enacted and refined during everyday interactions and through public texts such as those found in the mass media (Fairclough, 1995). Discourse analysis methodology (critical or otherwise) usually involves a fine reading of short stretches of text, examining discourse-level features such as overlapping speech, turn-taking, interruption, or hedging (e.g., *perhaps, kinda*, or tag questions like *it's hot, isn't it?*). Much well-known work on such features grows from influential writings by Robin Lakoff (1975), who associated many such features with gender-based power imbalances. Sociolinguistic analysis of media texts and depictions is exemplified by Rosina Lippi-Green's (1997) close reading of the varieties of English used by Disney characters.

Interaction and Audience

The ability of speakers to adjust their speech to specific situations has been a research interest of linguistic anthropologists, who work near the boundaries of sociolinguistics and have influenced the discipline. Although speech may be used to reveal or perform power imbalances, conversations are also generally collaborative enterprises. By interpreting interactions

as jointly organized activity (Sacks, 1984), conversational analysis looks at how participants work to organize and ease exchanges, by choosing when and how to deploy resources like discourse pairs (question–answer), or a shared awareness of the roles and expectations of types of participants (e.g., customer, interviewer, teacher). Dell Hymes (1974) describes ways in which aspects of an interaction (such as setting, participants, emotional tone, or genre) can influence the language choices that speakers make. Work in this area also attempts to identify the components of specific speech events, or genre-related language choices. Work on genre has also recently become central to the highly quantitative field of corpus linguistics.

Sociolinguists, on the other hand, have tended to focus on the participants in an interaction, more than on other Hymesian factors (which may reflect in part the fact that much sociolinguistic data is drawn from a single speech type, the interview). An influential article by Bell (1984) relates intra-speaker differences (i.e., style shifting) to the speaker's understanding of audience: immediate audience members (addressees) have a greater effect on style shifting than more distant conversational participants, such as overhearers. In the terminology of Accommodation Theory (Giles, 1973), speakers may attempt to sound more like their interlocutors (i.e., to *converge*) through similar speech rate, vowel forms, or discourse markers to express solidarity. Or they may *diverge*, in order to differentiate themselves (for example, the working-class boys in Cheshire (1982) use even more nonstandard forms to their teachers than they do to their friends).

More recent work on interaction (and on sociolinguistic variation more generally) tends to focus on the agency of speakers, as they use language to build up or perform an identity, to decide how to present themselves to others. The social meaning of the linguistic/stylistic performance emerges from the performance itself, so that speakers can use language features for the social meaning they might carry, but at the same time each performance of identity subtly shifts the shared understanding of the meaning of the features used. In such a constructivist approach, speakers may choose to portray themselves differently in different situations, or for different audiences. This perspective problematizes long-held notions of authenticity, social categories, or even group consensus or membership. For example, speakers might use features associated with the language of a different ethnic group, in order to invoke the socio-symbolic associations of that group, a process generally referred to as 'crossing' (Rampton, 1995).

Code Switching

In interactions where all (or many) of the participants can use more than one language, we see processes similar to style shifting occurring across language boundaries, i.e., *code switching*. The term can be used for switches between varieties of a single language (for example, when a man talking about theater tickets with his wife on his cell phone suddenly switches into African-American English to maintain a masculine persona in front of other men), but it is generally associated with switches between (mutually incomprehensible) languages.

Sometimes, long stretches of conversation between bilinguals occur in a single language, if one or both participants evaluate the (Hymesian) requirements of the interaction as requiring that language. For example, a teacher might talk to a student in the language that their school uses for instruction, to maintain distance and to avoid claims based on shared social or linguistic group membership. This is usually referred to as *situational code-switching*. Other times, fully bilingual participants may shift rapidly and fluently between languages, even within a single sentence, based on the topic, or the impression they wish to convey, or even to exhibit linguistic skill or because such rapid switching has become the norm in their bilingual speech community. Such *metaphorical code-switching* requires a high degree of knowledge of the two (or more) languages involved, and seems to follow fairly strict linguistic rules on when switching can occur. Language mixing in bilingual communities often functions as a form of identity performance, establishing new hybrid cultures and cultural stances (e.g., in global hip-hop).

Language(s) in Society

Most of the world's humans are multilingual, as are almost all of the world's countries. In this section, I look at how sociolinguists study the consequences of languages in contact.

Societal Multilingualism

Through population movements and political decisions (e.g., the drawing of national boundaries across linguistic boundaries), most contemporary countries are populated by speakers of different languages. Such societal multilingualism can take many forms: in some countries, speakers of different languages occupy distinct regions (e.g., French and Flemish in Belgium). Other countries have large immigrant populations, who across generations are likely to move toward the host nation's language (e.g., Spanish and English in the United States). Still others have, for various reasons, maintained societal bilingualism across multiple generations, to the point where most of the inhabitants are fluent in two or more languages and use each one in different areas (domains) of their lives. Ferguson (1959) calls this *diglossia*, and describes how one language (the H or High language) tends to be used in the courts, government, schools, literature, and education, while the other (the L or Low language) is generally associated with the domains of home, the street, or perhaps humor. In Paraguay, for example, the local language Guaraní might be used in rural areas, or with intimates, while Spanish is more common in urban areas, in formal contexts (Rubin, 1968). Work since Ferguson has broadened the concept of diglossia to include more multilingual contexts where domain is significant, and nuanced the original framework to account for some fluidity. Even in diglossic situations, for example, speakers can use the 'wrong' language to lighten up a conversation, or to invoke borrowed connotations of authority or authenticity.

Diglossic situations may involve the maintenance of individual multilingualism over multiple generations, but it is more often the case (e.g., in the case of immigration) that generations will move from monolingualism in the immigrant language, to bilingualism, to eventual monolingualism in the target country language. This process of language shift

usually occurs across three generations, depending on the 'ethnolinguistic vitality' (Giles et al., 1977) of the community. Fishman (1991) describes a series of stages through which a language moves, usually domain by domain, from full vitality to eventual language death. In nonimmigrant language shift situations, smaller or endangered languages retreat to a single domain, the home, and even then eventually often for use only with or among older generations. Much sociolinguistic research on language revival has looked at how to slow the decline or disappearance of the world's less-spoken languages or to describe the 'linguistic imperialism' (Phillipson, 1992) by which widely spoken languages of power and wealth displace such languages.

Language Contact

A common consequence of societal and individual multilingualism is that the languages used begin to influence each other. Some words from one language are used by code-switching bilinguals often enough that they become part of the other language, eventually adapted to suit the pronunciation and linguistic rules of the new language. Among the words that are at different stages of entering English (and may or may not be considered 'real' English words by different speakers), we might consider *boondocks* (from Tagalog), *mensch* (from Yiddish), or *desi* (from Punjabi and other Indic languages). Such borrowings make up a good part of the word stock of colonializing languages like English. Speakers can situate themselves with respect to language communities through the use of such words (or longer constructions such as the loan translation *I'm after telling him that* in Irish-influenced English).

In contexts where speakers are forced to communicate across language barriers but have limited access to a shared language, it is possible that a pidgin language will emerge. In situations of trade, or on plantations, a lingua franca will be needed to allow basic communication. Pidgin languages tend to be structurally simple, with small vocabularies taken from an economically or politically dominant language, and are intended for specific restricted purposes. When pidgin languages start to be used as first languages and for broader purposes, they are elaborated so that they can perform all the linguistic and interactional functions of a full language. They then become creoles. Well-known creoles include Nigerian Pidgin English, Haitian Creole, and Tok Pisin, the official language of Papua New Guinea. Linguists continue to debate (often hotly) the details of the origins and development of creoles, but generally agree on the full functionality of such languages and on the degree to which they are often stigmatized by nonspeakers unaware of their history or richness.

Social and Political Organization of Language

Societies and the institutions that they develop are often actively involved in determining the fate of the languages of their citizens, and these processes provide sociolinguists with both research material and the opportunity to see sociolinguistic findings used to inform policy.

Education is a particularly intense area of sociolinguistic debate. The value of bilingual education as a means to either maintain linguistic and cultural diversity or speed the integration of minority language speakers tends to be a politically loaded question, implicating as it does immigration policy, minority rights, and questions of national identity. Often, sociolinguistic research findings on bilingual education in one societal context are inappropriately heralded to support only vaguely similar policies in other, very different contexts, as societal multilingualism varies widely in its details from place to place. When it comes to the value to education of different varieties of the same language, the debate is no less intense. Influential early work by sociologist Basil Bernstein (1961) suggested that working-class children's dense social networks led them to develop a highly context-specific language variety, which he called a 'restricted code'; middle-class children's speech, on the other hand, was said to rely more on linguistic rather than contextual cues, leading to an 'elaborated code' better suited to the school system. Later sociolinguistic work revealed the richness of working-class speech, and suggested that the in-school uncommunicativeness of working-class students had more to do with differing sociolinguistic norms and limited (meta)linguistic awareness on the part of educators. It remains difficult to empirically gauge the pedagogical value of integrating home speech into classroom situations, due to social and political resistance to such domain crossing.

Another area of societal organization of language of interest to sociolinguists is that of language policy and planning. Governments and other social actors (schools, churches, writers, pundits, editors) make conscious decisions about which language(s) will serve which roles in a society. Status planning can decide the official language of the country, or of domains like education, law, or government. Different policies will encourage assimilation of minority language groups, or linguistic pluralism, or vernacularization (the promotion of a local language to official or national status), or internationalization (the choice of a nonlocal language as a 'neutral' lingua franca). Corpus planning can decide how a language should be changed (or not) to meet all the needs of the society, for example by creating new words and technical terms or by discouraging the use of borrowings from other languages (e.g., the *Office Québécois de la Langue Français*, which encourages the use of *le mot-dièse* 'word + musical-sharp-sign' rather than *le hashtag*).

Conclusion

Readers are encouraged to investigate more detailed articles on many of the topics discussed here (see below), or to get a broader overview of current sociolinguistics by consulting recent introductory textbooks (e.g., Coulmas, 2013; Mesthrie et al., 2009; Van Herk, 2012).

Hyphenated disciplines like sociolinguistics are capable of developing into many new areas over the next generation or two. We are already seeing promising work in a direction suggested or predicted by Suzanne Romaine (2001: 14568) in the first edition of this work ("the steady integration of macrosociolinguistic ('sociology of language') approaches with variationist and microsociolinguistic (interactional/DA) approaches seems a likely direction of development."). We are also beginning to see productive disintegration of these areas, with quantitative sociolinguists adopting the large-scale approaches and intense

statistical rigor of corpus linguistics, while researchers at the qualitative end productively collaborate with anthropologists and gender studies researchers. Speakers' agentive behavior is becoming more and more central to discussions of variation and style. And we can hope that, again in Romaine's words, "sound descriptive sociolinguistic research will continue to provide the critical underpinning for language policy developments in areas such as minority groups, immigration, education, and second language teaching."

See also: Anthropology, Genomics, and Human Variation: National Roots; Bilingual Education: International Perspectives; Code Switching: Linguistic; Communities of Practice; Dialectology; Diglossia; Language Contact; Language Policy; Language and Ethnicity; Language and Gender; Language and Social Class; Language and Society; Multilingualism; Pidgin and Creole Languages; Politeness and Language; Social Identity in Sociology.

Bibliography

Bailey, B.H., 1996. Communication of Respect in Service Encounters between Immigrant Korean Retailers and African American Customers. Unpublished Master's thesis. University of California at Los Angeles, Los Angeles, CA.

Bell, A., 1984. Language style as audience design. Language in Society 13, 145–204.

Bernstein, B., 1961. Social class and linguistic development: a theory of social learning. In: Halsey, A.H., Floud, J., Anderson, C.A. (Eds.), Education, Economy, and Society. Free Press, New York, NY, pp. 288–314.

Blondeau, H., Nagy, N., Sankoff, G., Thibault, P., 2002. La couleur locale du français des Anglo-Montréalais. In: Dewaele, J.M., Mougeon, R. (Eds.), L'acquisition de la variation par les apprenants du français langue seconde [Acquisition et Interaction en Langue Étrangère 17]. DEPA-Université Paris VII, Paris, FR, pp. 73–100.

Brown, P., Levinson, S.C., 1987. Politeness: Some Universals in Language Usage. Cambridge University Press, Cambridge, UK.

Buchstaller, I., van Alphen, I. (Eds.), 2012. Quotatives Cross-linguistic and Cross-disciplinary Perspectives. Benjamins, Amsterdam, NL.

Cheshire, J., 1982. Variation in an English Dialect: A Sociolinguistic Study. Cambridge University Press, Cambridge, UK.

Coulmas, F., 2013. Sociolinguistics: The Study of Speakers' Choices, second ed. Cambridge University Press, Cambridge, UK.

Eckert, P., 2000. Linguistic Variation as Social Practice. Oxford UK, Blackwell.

Eckert, P., McConnell-Ginet, S., 1999. New generalisations and explanations in language and gender research. Language in Society 28 (2), 185–203.

Fairclough, N., 1995. Critical Discourse Analysis. Addison Wesley, Boston, MA.

Ferguson, C., 1959. Diglossia. Word 15, 325–340.

Fischer, J.L., 1958. Social influences on the choice of a linguistic variant. Word 14, 47–56.

Fishman, J., 1991. Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages. Multilingual Matters, Clevedon, UK.

Giles, H., 1973. Accent mobility: a model and some data. Anthropological Linguistics 15, 87–105.

Giles, H., Bourhis, R., Taylor, D., 1977. Toward a theory of language in ethnic group relations. In: Giles, H. (Ed.), Language Ethnicity and Intergroup Relations. Academic Press, London, UK, pp. 307–349.

Hymes, D., 1974. Foundations in Sociolinguistics: An Ethnographic Approach. University of Pennsylvania Press, Philadelphia, PA.

Labov, W., 1963. The social motivation of a sound change. Word 19, 273–309.

Labov, W., 1972. Language in the Inner City: Studies in the Black English Vernacular. University of Pennsylvania Press, Philadelphia, PA.

Labov, W., Ash, S., Boberg, C., 2006. Atlas of North American English: Phonetics, Phonology, and Sound Change. Mouton de Gruyter, New York, NY.

Lakoff, R., 1975. Language and Women's Place. Harper & Row, New York, NY.

Lambert, W.E., Hodgson, R.C., Gardner, R.C., Fillenbaum, S.J., 1960. Evaluative reactions to spoken languages. Journal of Abnormal and Social Psychology 60, 44–51.

Lippi-Green, R., 1997. English with an Accent. Routledge, London, UK.

Mesthrie, R., Swann, J., Deumert, A., Leap, W.L., 2009. Introducing Sociolinguistics, second ed. Edinburgh University Press, Edinburgh, UK.

Milroy, L., 1980. Language and Social Networks. University Park Press, Baltimore, MD.

Phillipson, R., 1992. Linguistic Imperialism. Oxford University Press, Oxford, UK.

Preston, D.R., Howe, G.M., 1987. Computerized generalizations of mental dialect maps. In: Denning, K.K. (Ed.), Variation in Language: NWAV-xv at Stanford Dept of Linguistics. Stanford University, Palo Alto, CA, pp. 361–378.

Purnell, T., Idsardi, W., Baugh, J., 1999. Perceptual and phonetic experiments on American English dialect identification. Journal of Language and Social Psychology 181, 10–31.

Rampton, B., 1995. Crossing: Language and Ethnicity Among Adolescents. Longman, London, UK.

Romaine, S., 2001. Sociolinguistics. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of Social & Behavioral Sciences, first ed. Elsevier, Philadelphia, PA, pp. 14562–14569.

Rubin, J., 1968. National Bilingualism in Paraguay. Mouton, The Hague, NL.

Sacks, H., 1984. Notes on methodology. In: Atkinson, J.M., Heritage, J. (Eds.), Structures of Social Action Studies in Conversation Analysis. Cambridge University Press, Cambridge, UK, pp. 21–27.

Trudgill, P., 1972. Sex, covert prestige, and linguistic change in the urban British English of Norwich. Language in Society 1 (2), 175–195.

Van Herk, G., 2012. What Is Sociolinguistics? Wiley-Blackwell, Boston, MA.

Sociological Theory

Jonathan H Turner, University of California, Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T.J. Fararo, volume 23, pp. 15656–15662, © 2001, Elsevier Ltd.

Abstract

Sociological theorizing has dramatically expanded its explanatory range over the last years. Yet, certain enduring problems still divide theoretical work, namely, the questions of whether or not sociology can be a science, whether or not it should be normative and critical or value-neutral, and whether the appropriate level of analysis is micro or macro. Nine basic theoretical perspectives have evolved over the last decades, each with many variants. These perspectives include functional, evolutionary, ecological, conflict, exchange, interactionist, structural, cultural, and critical theoretical approaches.

Points of Fundamental Disagreement

From its beginnings in Europe and the United States, sociology has been ambivalent about whether or not scientific theorizing like that in the natural sciences is possible, or even desirable. There are several points of intense disagreement.

Can Sociology Be a Science?

This is perhaps the most hotly debated issue because it immediately affects the kind of theorizing that is done in sociology. If there is a belief that sociology can be a natural science, theorizing will seek to isolate the generic phenomena that drive all behaviors, interactions, and patterns of social organization in all times and in all places. Theorists will attempt to develop abstract theoretical principles or 'laws' as well as formal models of generic social processes. Conversely, if sociologists do not believe that sociology can be a science, theorizing will be a variety of intellectual enterprises: history of ideas, descriptions of current or past events, concepts describing phenomena, ideological advocacy, philosophy, and meta-analysis of social thought. There are many justifications for this kind of activity but the most fundamental is that human behavior, interaction, and organization are seen as fundamentally different from the domains of the universe studied by the natural sciences. The most central of these is that humans possess agency and can change the fundamental nature of social organization, thereby obviating any theoretical principles or models that are presumed to be about universal phenomena.

Should Sociological Theorizing Be Normative?

Many sociological theorists reject the notion of value-free science. Sociology and its theories should address problems of social organization – most importantly inequality and oppression. But, the correlation between advocacy and science is not perfect. Many contemporary scientific theorists view the development of scientific theory as a necessary step in creating more just societies. Others see themselves as scientists but of a special kind: scientists devoted to understanding the dynamics of inequality and oppression with an eye to using this knowledge to destroy oppressive arrangements. In contrast to this position are those who seek to explain the operative dynamics driving the social universe, per se, rather than as part of a social activist and value-laden agenda.

What Is the Appropriate Level of Analysis

Whatever the position of theorists on the scientific prospects for sociology, there has always been a debate over the level of analysis that will give sociological theories the most explanatory power. Early sociology was decidedly macro in its emphasis, although microthinking penetrated almost all of the founding generation of sociological theorists at some point. But during the twentieth century, theorizing split into three main camps (Turner and Turner, 1990): macro, micro, and mesotheorizing. These levels of theorizing correspond to the way the social universe actually unfolds. There is a microlevel of face-to-face interaction, a macrolevel of societal, intersocietal, institutional, and stratification analysis, and a mesolevel that tries to explain those structures that constrain microinteractions, while at the same time being the building blocks of the macrouniverse.

One of the main issues in sociological theorizing has been linking theoretical explanations across all of these levels. For scientifically oriented sociologists, just like scientists in physics, biology, and economics, the theoretical link is the holy grail of explanation. Those critical of science often note that the failure of sociologists to make this theoretical linkage is 'proof' that sociology cannot be a science, conveniently ignoring the fact that no science has thus far achieved this lofty goal. In light of some recent theories in sociology (e.g., Lawler et al., 2009; Turner, 2010–2013), sociology has achieved considerable integration across levels; and in fact, it could be argued that sociology is further along in finding the holy grail than any other science. But, this issue will remain controversial.

Despite these controversies, sociological theorizing has made dramatic progress over the last 40 years. There are more theories with more explanatory power than ever before. The progress presents a problem in how to cover the range of theories in this delimited space. The best way, indeed the only way, is to divide up theoretical activity into a variety of perspectives or approaches, with each approach addressing the controversial issues outlined earlier as it examines a particular range of phenomena in the social universe.

General Theoretical Perspectives in Present-Day Sociology

Naturally, it is difficult to divide up theoretical activities into truly discrete approaches, but there are still at least nine broad rubrics that organize theoretical sociology today, each with many variants. Moreover, theories within a rubric can come into conflict with other, while many theories listed under different rubrics can overlap. Such is the somewhat chaotic nature of theorizing in sociology (Turner and Maryanski, 2008).

Functional Theorizing

Functional theorizing was sociology's first major approach, introduced by Auguste Comte and carried forth into the early twentieth century by Herbert Spencer ([1874–1894]1898), Emile Durkheim (1893[1933]), and to the midpoint of the twentieth century by anthropologists at which point functionalism became sociology's dominant theoretical approach in the 1950s and 1960s, only to seemingly disappear in the contemporary era. All functional theories begin with a basic presumption (Turner and Maryanski, 1979): social systems have functional needs or requisites that they must meet if they are to adapt to their respective environments. Theoretical explanation is achieved when 'the functions' of particular subparts of a system are discovered and analyzed for why they evolved and how they operate to meet, or fail to meet, universal functional requisites. Early functional theories focused on societies as the unit of analysis, and thus they posited needs or requisites that societies must meet, with the analysis of institutional systems (e.g., economy, polity, kinship, religion) in terms of the societal-level needs that they meet. Herbert Spencer and Emile Durkheim can be seen as the two templates for functional analysis. Spencer posited three basic needs, needs for operation (production and reproduction), for regulation (through power and unifying cultural symbols), and for distribution (of people, information, and resources across territories); Durkheim posited one overarching need for integration. Thus, by determining how institutional systems meet multiple or singular needs, theoretical explanation is achieved. When functionalism appeared to die for the first time in the early 1900s because of its association with evolutionary models (see next section), it was carried forth by anthropologists, particularly Bronislaw Malinowski (1944) and A.R. Radcliffe-Brown (1952).

When functionalism was reborn in sociology during the 1950s and 1960s, Talcott Parsons became the dominant figure, arguing like Spencer that all systems of the social universe reveal four basic needs (Parsons et al., 1953): adaptation (taking in needed resources from the environment, converting them for use, and then distributing them); goal attainment (establishing system goals and mobilizing resources to meet these goals); integration (coordinating differentiated parts of the system); and latency (or managing tensions within and among parts of the system as well as reproducing system parts). Other contemporary functionalists like Niklas Luhmann (1982) adopted a more Durkheimian strategy in positing one master need: the need to reduce complexity of systems in their temporal, spatial and material, and symbolic environments.

Functionalism declined and then virtually disappeared because of a number of problems: it seemed to many to be a cover for a conservative ideology emphasizing the functions of the status quo; explanations seemed to be tautological (system parts meet functions for survival and this is known because the system is surviving); and illegitimate teleologies in which the historical origins of system parts are caused by their functional outcomes. Yet, functionalism has never been extinguished, and there are scholars openly working with functionalist schemes, often downplaying the emphasis on system requisites in ways that mute the criticisms of this approach.

Evolutionary Theorizing

Revival of Stage Models

When evolutionary theorizing died in the first decade of the twentieth century, it did so because of the racism and ethnocentricism associated with viewing European societies as the most advanced compared to 'primitive,' preliterate peoples. Functional models of evolution, as well as conflict models like the one developed by Karl Marx, all saw societies as moving through a series of evolutionary stages from simple hunting and gathering through horticultural (plus pastoral and fishing variants) and agrarian stages, to the pinnacle of societal evolution: western industrial societies. When evolutionary theory reemerged in the 1960s (Lenski, 1966; Parsons, 1966, 1971), it came back in a much more sophisticated guise, with sensitivity not evident among nineteenth- and early twentieth-century theorists. More emphasis was placed on the mechanisms driving evolution, such as population growth, technological development, warfare, and ecological pressures. With the revival of stage model evolutionary theorizing, functional analysis crept in the back door with theorists emphasizing that societies become more complex as they confront selection pressures along a number of axes, such as population growth, environmental degradation, insufficient production, problems in reproduction, pressures to regulate conflict and complexity, and problems in distributing resources, information, and people across larger territories controlled by more complex sociocultural systems (Turner, 2010–2013).

The Darwinian Revival

At the same time these new stage-model evolutionary theories were coming back into vogue, new more explicitly Darwinian-inspired theories from outside sociology proper were being posited as an alternative way to explain sociocultural phenomena. Ironically, evolutionary theorizing had been rejected in the 1920s and 1930s because of the spread of Social Darwinism, but the revival of Darwinism was now purely biological, but nonetheless threatening to sociologists. These Darwinian approaches have been rejected by many sociological theorists, but a new field of theoretical sociology – loosely called 'evolutionary sociology' – has emerged to embrace and import models from biology, including sociobiology and evolutionary psychology, or to provide an alternative that is equally steeped in biology without the unfounded assumptions of sociobiology or evolutionary psychology (see Turner et al., 2014 for examples). This alternative is willing to consider the genetic

basis of human behavior as a result of selection pressures, but it adds many of the general ideas of contemporary stage-modeling evolutionary theories, such that pressures from genes are only one of many selection pressures. Equally important, many of these other selection pressures come from problems of organizing among humans' hominid ancestors in new ecological environments. There has often been considerable debate and even hostility among the various camps of more Darwinian evolutionary theories in sociology, but it is clear that biology is back in sociology and will continue to be an important line of theory and research.

Ecological Theorizing

In both Spencer's and Durkheim's functional and evolutionary theorizing we find another kind of biologically inspired approach: ecological theorizing that became firmly a part of sociological theory in the United States during the 1930s. Drawing from Darwin's key idea of natural selection, the basic arguments in all ecological theories are much the same: social units compete for resources within niches of resources that are necessary for their survival; and those that possess traits that enhance their capacity to secure resources and reproduce themselves will survive, while those that cannot, will die out or be forced to move to a new resource niche. Like Darwin and later Durkheim, ecological theories emphasize that as the number of units in a resource niche increases, density increases the level of competition; at some point, the niche cannot carry all of the units competing for resources. Thus whether considering the distribution of people and types of social units in urban areas (Niemela, 2011; Berry and Kasarda, 1977), where competition is institutionalized in the real-estate market, or the life and death of organizations competing for resource shares in dense niche (Hannan and Freeman, 1989), Darwinian processes are the driving force and can explain many of the dynamics of urban areas and populations of organizations. Some have even taken ecological theorizing back to the macrolevel by emphasizing that the relations among societies in their common environments can be understood in ecological terms (see, e.g., Hawley, 1986).

Conflict Theorizing

Conflict theorizing has always been part of sociological theory. For a time in the 1960s and 1970s, conflict theorizing became a self-labeled school of thought to use as a foil for attacking the weaknesses of Parsonsian functionalism. Indeed, during the period from 1960 to 1990, power and conflict were dominant topics in sociology, just as they had been for the founding masters in the classical tradition. Over the last few decades, however, the label conflict theory has receded as the basic substance of the approach has become a part of so many other theoretical traditions.

The central argument of all conflict theories is that inequalities in social systems generate tensions that, under specifiable conditions, lead subordinates in a system of inequality to mobilize for conflict against superordinates and, thereby, to change the structure and culture of a social system. Most conflict theories have been couched at the macro, societal level but in recent years, the approach has been extended to even more macrolevels in world systems theorizing (Wallerstein, 1974–1989; Chase-Dunn, 2001) and down to the microlevel of face-to-face interaction (Collins, 2008). In conflict theorizing, one also sees the diverse epistemological positions of sociology in play. Some are decidedly scientific and analytical; others are more historical and descriptive while being more abstract and analytical than most histories; still others are more normative and advocate particular forms of conflict to achieving value-laden end states; and a few are both. In all, the legacy of Max Weber and Karl Marx is prominent.

Analytical Conflict Theories

The first conflict theories to attack functionalism were analytical, often making Marx's model more abstract and adding qualifications of Max Weber and Georg Simmel (Dahrendorf, 1959; Turner, 1975, 2010–2013). These theories emphasized that conflict occurs in all social systems under specifiable conditions, including the level of inequality, abusive practices of superordinates, emergence of leaders, codification of grievances into ideologies and emotional arousal of subordinates, and mobilization of material and organizational resources to pursue conflict. These analytical theories converged with resource mobilization theories developed by those studying collective behavior (e.g., McCarthy and Zald, 2001; Zald and McCarthy, 1987; Tilly, 1993). Some analytical theories were more Weberian in origin and have sought to explain conflict at all levels of social organization in terms of variables central to Weber's theory, such as the multiple bases of inequality among different types of social units at varying levels of social organization (Collins, 1975, 2008; Turner, 2012).

Marxian Conflict Theories

Alongside these more scientifically oriented analytical theories can be found a variety of Marxist theories. These theories often became conflated with various types of 'critical theories' (see later sections) using Marxian ideas to criticize various socio-cultural conditions in modern societies and to advocate for particular types of change. Many downsized Marxian ideas to particular topics such as ethnic and gender inequalities (e.g., Bonacich, 1972; Chafetz, 1990) and developed critical analyses of these more specific forms of inequality. Still other approaches shifted the unit of analysis from societal social systems, where Marx's predictions failed, to the dynamics of intersocietal systems, where these predictions would seem more likely to occur as capitalism became global and the contradictions of capitalism would play themselves out and cause crises that might usher in socialism (Wallerstein, 1974–1989; Chase-Dunn, 2001). During this entire period from the 1960s to the present, Marxists have continued to wrestle with the question of why Marx's predictions about capitalism – polarization into two classes, mobilization for conflict by the proletariat, and the overthrow of capitalism – did not occur in any capitalist society. To address these failures in Marx's predictions, these Marxian approaches examined the problems of proliferating classes, the problem of conceptualizing exploitation, the contradictory class locations of persons and families, and other occurrences that went against Marx's

predictions (Wright, 1997), often culminating in advocating for a less Marxian view of socialism as the most realistic utopia (Wright, 2010).

Comparative-Historical Conflict Theories

During the entire period since the 1960s, when conflict theory began to attack functionalism, historical and comparative analysis of conflict, particularly revolutions, became highly prominent. These theories add new variables to the conflict model, such as the conditions causing the weakening of the state that makes it vulnerable to revolution and the demographic, ecological, geographical, and economic conditions in (mostly feudal) societies that increase revolutionary potential (Tilly, 1993; Skocpol, 1979; Goldstone, 1990). Much of this comparative-historical work also involved blending theories from collective behavior, especially from resource mobilization theories (McCarthy and Zald, 2001) with Marxian and Weberian conflict theories to particular historical cases in order to explain the mobilization of subordinates for conflict.

Microlevel Conflict Theories

In recent years, theorists like Collins (2008) have begun to explore the dynamics of interaction in conflict. These theories try to explain why and through what interpersonal processes individuals become violent or willing to be violent in various types of conflict – from crowd behaviors through combat units engaged in war to terrorist acts, often emphasizing the biosocial forces arousing various types of emotions and the interactional dynamics that lead to mobilization for conflict. These theories stress that acts of conflict, even when well organized in mobilized armies, do not assure that all combatants will actually engage in violence, thereby qualifying more meso- and macrolevel theories of conflict. Conflict mobilization and actual violence are ultimately interpersonal processes revolving around the arousal and then channeling of emotions (Collins, 2008).

Exchange Theorizing

For many decades in the nineteenth and twentieth centuries, sociologists were equally distrustful of utilitarian or neo-classical economics and psychological behaviorism, even though both contain the seeds for what became known as exchange theory in the 1960s. These exchange theories relaxed if not rejected many of the key extreme assumptions of neo-classical economics (e.g., actors seek to maximize utilities) and behaviorism (e.g., thinking and decision-making cannot be directly observed and, hence, theorized), while sustaining the basic idea that individuals and collective actors seek resources that bring rewards and reinforcement. In all sociological exchange theories, actors are seen to calculate the costs and investments in securing resources; and when the resources received exceed the costs required to procure them, individuals experience positive emotions and become willing to form more permanent exchange relations. Increasingly, exchange theories began to emphasize that, from these basic exchange dynamics, the meso- and macrolevel social structures of human societies are ultimately built up.

Exchange theories typically began with either behaviorist or utilitarian assumptions but, as they relaxed the extremes of these theoretical approaches, they converged into a set of theories emphasizing that both solidarity and conflict in social relations are related to exchanges of resources among individuals and collective actors. As these theories developed, additional dynamics were added to exchange, including inequalities in the relative power of actors (Blau, 1964; Emerson, 1962), calculation of not only costs and investments but also fairness and justice in exchanges, emotional arousal as an outcome of justice, commitments to exchanges (Cook and Emerson, 1978; Lawler and Yoon, 1996), and commitments to macrolevel structures (Lawler et al., 2009).

There is now a variety of general approaches to exchange theorizing that, in somewhat different ways, incorporate the above considerations into models and theories, all of which are committed to the epistemology of science. The most prominent are rational choice theorizing, which extends across the social sciences, and exchange-network theorizing, which remains more uniquely sociological.

Rational Choice Theories

This approach often retains the neo-classical notion that individuals seek to maximize their rewards, relative to costs and investments, but even with this rather extreme assumption, the approach attempts to explain sociological phenomena in terms of the underlying rationality of actors. One basic argument within sociology (Coleman, 1990) is that it is often rational for actors to give up certain rights in order to avoid what are sometimes termed negative externalities (or punishments and negative outcomes) by agreeing to abide by norms or to give authority to certain actors. It is rational to do so because, in avoiding negative externalities, the costs of giving up rights are less than the punishments that come with negative externalities. Thus, power and normative systems evolve in human social systems as rational actors seek to reduce negative externalities and, thereby, maximize the rewards that come with more stable patterns of social organization.

Social solidarity and group cohesion are also topics examined by rational choice theories (Hechter, 1987). Group solidarity, for example, can be explained by the degree to which individuals are dependent upon groups for receipt of intrinsic rewards or private goods created by interactions in the group; these are often very valuable and not easily secured elsewhere or only at high costs. Thus, it is rational for actors to monitor and sanction each other and to create normative systems that require individuals to contribute to group activities and, thereby, receive not only the rewards produced by this activity but also the additional rewards that come with a sense of group solidarity and other intrinsic reinforcers. Thus, seemingly normative processes can be explained by the rationality of actors. Conversely, when individuals participate in groups for extrinsic rewards alone, it is rational for them to free-ride or do as little work as possible, thereby forcing organizations to create more formal systems of monitoring and sanctioning, thus, reducing group solidarity but, nonetheless, reducing free-riding. Even though authority and normative systems impose costs on patterns of social organization, it is rational to create them in order to

sustain productivity of group members for whom it would otherwise be rational to free-ride.

Exchange-Network Theories

This theorizing blends network theory (see later text on structural theories) with exchange theory, while making power central to the exchange process. It also emphasizes that it is the nature of the relationship rather than the properties of actors that is to be subject of theorizing. Whether individuals or collective actors, the same basic exchange dynamics operate; and these revolve around the forms of relationships generated by power inequalities (Emerson, 1962). The more one actor controls resources that other actors value, the more dependent are the latter on the former, thus giving power to an actor who possesses valued resources. Thus, the power of actor A over actor B is a function of the dependence of B on A for a valued resource. Actor A will use this power and increasingly demand more resources from B; and this exploitation of B by A sets into motion balancing operations by B to shift the balance of power by several means: Offering new resources valued by actor A; joining in coalitions of Bs against A and/or cutting off alternative sources of the resources to A; doing without the resources offered by A; seeking alternative As or new sources for a valued resource; and other strategies that reduce dependence of Bs on As.

Also within this exchange network tradition, more interpersonal processes have been examined, in particular emotions, commitments to exchange relations, and commitments to macrostructures. The basic argument is that the more equal exchanges of resources are, the more positive are the emotions of actors; and over time positive emotions create commitments to a relationship and, ultimately, to macrostructures, especially when individuals are engaged in joint productive activities where each actor's contribution is difficult to determine, resulting in group solidarity that travels up networks of embeddedness among social structures that produce commitments to societies as a whole (Lawler, 2001; Lawler et al., 2009).

Convergence of Exchange and Conflict Theories

Some early exchange theories (e.g., Blau, 1964) made explicit the similarity between Marx-inspired conflict theories and exchange theories. In fact, conflict theory might be seen as a variant of exchange theories because conflict emerges with inequalities of power, in which the powerful exploit the less powerful and extract unfair and unjust rewards from subordinates. Such exploitive exchanges arouse negative emotions and set into motion conflict dynamics revolving around the ideological mobilization of subordinates to form conflict groups and redress their grievances through conflict.

Interactionist Theorizing

Early sociological theorizing was decidedly macro in its focus, but signs of a more microlevel orientation could be found in Spencer's ([1874–1894]1898) analysis of ceremonial institutions, Weber's (1968[1922]) later turn to verstehn analysis and types of action, Durkheim's (1963[1912]) increasing concern with emotion and rituals, Simmel's (1956 [1903]) analysis of self in groups, Alfred Schutz's (1967 [1932]) phenomenology, and most importantly Charles Horton Cooley' (1902) and George Herbert Mead's (1934) analysis of self and interaction processes. Present-day interactionist theorizing has now differentiated into several distinct lineages of theorizing.

Symbolic Interactionist Theories

Drawn from Cooley and Mead, symbolic interactionism now emphasizes the significance of self and identity processes. Self is increasingly conceptualized as a series of identities that individuals seek to verify in interactions with others; and depending upon whether or not verification occurs, persons will experience either positive or negative emotions (Burke and Stets, 2009). There are many variants of these identities theories, but most emphasize that there are several levels of identity, including core or person identities, social identities, group identities, and role identities. These are often conceptualized as a hierarchy with a persons' core feelings and cognitions about self being the most general, social identities attached to membership in categories (gender, ethnicity, age, etc.) as the next most general, group identities or attachments to corporate units being the next most general, and finally, role identities being the least general. To some extent, these various levels of identity are interconnected, with verification of an identity at one level having effects on confirming an identity at another level.

Some symbolic interactionist theories (e.g., Scheff, 1997; Turner, 2002, 2007) introduce psychoanalytic dynamics, emphasizing that individuals will often repress negative emotions like shame and guilt when their identities are not verified. With repression, the dynamics of emotion often disrupt interpersonal processes, while having potentially large effects on people's commitments to macrostructures in a society.

Dramaturgical Theories

This form of interactionist theorizing draws more from Durkheim's analysis of emotional-arousal rituals than from Cooley and Mead. The principle difference between symbolic interactionism and dramaturgy is that self and identity are viewed as more situational and transitory in dramaturgy than in symbolic interactionism. Self is something that individuals 'put on' for a particular presentation to others in a social situation rather than being a regulating gyroscope for behavior. Indeed, in dramaturgical theorizing, individuals often put on rather disingenuous selves for particular strategic purposes. Erving Goffman (1959) was the originator of this form of theorizing, where episodes of interaction are much like a theatrical presentation with the script inhering in norms and determining the broad contours of roles that actors interpret with some dramatic latitude in their presentations of self to an audience of others. Interactions are also normatively regulated by rituals (Goffman, 1967) that open, close, form, and repair episodes of interaction on the social stage and they are constrained by frames that actors impose to specify what is to be included and excluded from interaction (Goffman, 1974). During the course of interaction, individuals categorize situations in terms of the relative amount of work – practical, social, or ceremonial content

they are to contain. They also determine through rituals and frames the normatively appropriate forms of talk, as well as types and levels of emotions to be presented to others.

While interactions are very much constrained by the normative script, they also involve a great deal of strategic behavior in the presentation of self, reference to norms, imposition of frames, and use of rituals, often for nefarious purposes. Thus, interaction is not solely driven by people's efforts to verify identities, as more recent identities theories might emphasize; it is a dramatic presentation drawing from a cultural script containing norms that, in turn, specify the appropriate roles, frames rituals, talk, and feelings. These elements that guide interaction are often manipulated in self-presentations.

Dramaturgy has gone in many directions since Goffman's analysis. One is into the sociology of emotions, where actors experience negative emotions when encounters of interaction are breached. Actors also use emotions for manipulative purposes to gain favor, respect, or even sympathy form others (Clark, 1997). Another is the analysis of interaction rituals for how they arouse emotions and, thereby, build up solidarity and culture in groups (Collins, 2004). Yet another line of dramaturgical inquiry is the analysis of an emotion culture, which imposes feeling and display rules (in the cultural script) that individuals do not feel and must, thereby, work at or engage in alienating emotional labor to present to others (Hochschild, 1979, 1983). Yet another line of analysis focuses on deviance and the emotional strategies used to cope with pressures from multiple roles, subcultural marginality, ambiguities in norms, and at times rigid rules governing rituals (Thoits, 1990).

Role Theories

In recent decades, role theorizing has declined, although several sociologists continue the tradition (e.g., R. Turner, 2001; J.H. Turner, 2002). Mead introduced the notion of role taking to emphasize the process by which individuals read the gestures of others to determine their dispositions and likely courses of action. They can role-take not only with others but also with generalized others, thereby plugging themselves into the culture of a situation. For many decades, roles were theorized, emphasizing that roles are behaviors emitted by persons by virtue of normative requirements and status locations of individuals in a situation. Ralph H. Turner (1962, 2001) developed a more robust conception of roles by emphasizing the compliment of role-taking, the process of role-making whereby individuals emit both conscious and unconscious gestures as they seek to make and establish a role for themselves vis-à-vis other roles in a situation.

Status Theories

Status theorizing began at about the peak of role theorizing in the 1950s and 1960s (Berger, unpublished). In status theories, emphasis is on the expectations associated with status in situations, with status being defined along three dimensions: authority at locations in the division of labor of a corporate unit, membership in a social category (diffuse status characteristic), and prestige or honor that people can claim by virtue of their status locations or diffuse status characteristics. For any status, there are moral evaluations of the 'worth' of incumbents that are codified into broader status beliefs in the culture of a society. These status beliefs constrain the expectation states for individuals in a given status; that is, for any status there are expectation states for how people should behave. For those with higher authority in status locations, for those with more prestige, and for those with more valued diffuse status characteristics, expectation states give them advantages in interaction; they are allowed to initiate more interactions, to have their views followed, and to receive deference and obedience for their prestige and authority. If, however, individuals do not meet the expectation states on them, they will experience negative emotions, while others will often negatively sanction them. The same is also true of lower status individuals; if they violate the lowered expectations states on them by, for example, challenging higher status persons, they will often be sanctioned by their fellow lower status peers (Ridgeway, 1994; Milner and Walker, 2014). Thus, status beliefs and expectation states create a status order in all interactions, thereby giving interactions stability over time and place. This research theory tradition has continuously expanded its reach over the last decades (for a summary review, see Wagner and Turner, 2013).

Another line of status theories emphasizes the power of expectations. Individuals are conceptualized as seeking power and status in interactions, but their emotional reactions in situations are related to their expectations for gains or losses in status and whether or not they met or failed to meet these expectations (Kemper and Collins, 1990; Webster and Walker, 2014). The higher the expectations for maintaining or gaining in status, and the less these expectations are realized, then the more individuals will experience shame, whereas the more they expected to gain in power but in fact lost power, the more they will experience anxiety and lose confidence. Conversely, if expectations are for gains in either power or status which are then realized, individuals will experience positive emotions and gain in confidence.

Phenomenological Theories

German phenomenology emphasized the search for the origins of meaning, but in the hands of sociologists like Alfred Schutz (1967[1932]), it was put on a more sociological footing by theorizing how individuals go about constructing a sense of common meanings that lead to a sense of intersubjectivity. The most prominent extension of this phenomenological thrust came from a school of theorizing and research labeled ethnomethodology (Garfinkel, 1967; Heritage, 1984; Hayashi et al., 2013), which emphasized that there are hidden 'folk methods' that individuals use to construct a sense that they share a common world and that, for the purposes of interaction, they perceive this world in the same way as others. More important than actual intersubjectivity is the perception of intersubjectivity; this perception, even if false, allows encounters to proceed smoothly and avoid breaches.

Structural and Structuralist Theorizing

The analysis of social structure has always been at the heart of sociological theorizing; and yet, there is still relatively little

consensus over just what structure is in the first place and even less consensus over its operative dynamics. Early functional theories emphasized that social structure is a system of differentiated positions in diverse types of sociocultural formations, such as encounters and groups at the more micro level, organizations and social categories at the meso level, institutional and stratification systems at the macro level, and societies as a whole and intersocietal systems at an even more macro level. Conflict theories have also conceptualized societies as differentiated but the emphasis has been on the structure of stratification systems composed of social classes and on the institutional systems that generate the inequality that ultimately causes stratification. Interactionist and other microlevel theories have emphasized the interpersonal dynamics that ultimately are the building blocks of social structure and culture, while mesolevel theories have tended to focus on status dynamics in groups and organization, and on the operation of organizations in various institutional environments. Yet, all of these conceptions of social structure, while basically correct, have not led to a more unified and integrated view of social structure and its key dynamics.

There has been a whole intellectual movement lasting almost a century labeled structuralism that, unfortunately, did not clarify social structure – if anything, just the opposite. Drawing from Durkheim's view that social structure is the number and nature of parts linked together in some identifiable pattern through which resources flow, structuralism became conflated with, first of all, the notion of deep structures that dictate the formation of observable 'surface manifestations' of these deep structures. Later, the deep structures became the neurochemistry of the human brain as it directed particular types of cultural and social structural formations (Levi-Straus, 1979). In the end, structuralism has failed at providing sociology with a clear definition of social structure, much less its dynamics.

Some like Anthony Giddens (1984) have sought to integrate at least the imagery of structuralism with microtheories by viewing structure as a set of rules and resources organized into structural principles that govern institutional formations that constrain, while also being influenced by, the agency of actors as they seek to meet fundamental needs in interactions with others. Yet, while this kind of approach makes structuralism more sociological, it does not clarify the nature and dynamics of social structures.

At the opposite end of the continuum of clarity and precision is a broader theoretical and research tradition on networks that has sought to conceptualize social structures in very precise and highly abstract terms as nodes connected in various forms by linkages or ties among the nodes, with the linkages viewed as carriers of various types of resources – information, power, affection, commitments, material goods, and virtually anything that connects individuals and the systems in which they are located. The goal of such theorizing is to conceptualize key properties of networks – number of nodes, number and density of ties among nodes, centrality of ties through particular nodes, formation of cliques of subdensities of ties in a larger network, brokerage and bridging nodes connecting networks, and the like. Depending upon the configuration of a network – from a simple network of friendships or family ties through ties among organizations in markets to ties among nation states, the form of the network is presumed to have certain dynamic properties that can be theorized – although the actual amount of theorizing has been quite negligible compared to the computer algorithms that have been developed to describe networks (rather than explain their dynamics). There is a certain clarity achieved by more precise definitions of network forms and their measurements but thus far network theorizing has not fully realized its potential as an explanatory tool that can offer a more precise analysis of social structure. Thus, whatever the future of network theorizing, it can be safely stated that it has not realized its potential (but for works that do attempt to develop theory, see Friedkin and Johnsen, 2011; Burt, 1992).

Cultural Theorizing

Cultural theorizing has many of the same problems as does theorizing about social structure: a basic lack of clarity about the key properties of the phenomenon and their operative dynamics. Culture is defined as systems of symbols carrying meanings that provide a resource for actors while at the same time carrying directives that constrain the actors; and it has been conceptualized into various types of meaning systems – the most enduring being languages, texts, knowledge systems, values, ideologies, and norms. Although these are obvious simplifications, they have been used to good effect in much sociological analysis.

More recently, a debate over the connection of culture to social structures and interpersonal behaviors has emerged, with some arguing for a 'strong program' of studying culture, per se, independent of its connection to social structure and interactions. Instead, the emphasis should be on culture in its broader historical and social context, analyzing the unique elements of culture (Alexander and Smith, 2001). This approach derides 'the weak program' that prematurely connects and subordinates the dynamics of culture to social structural and interpersonal processes, while at the same time, underplaying the unique properties of culture in order to conceptualize culture in more abstract terms.

The Weak Program

Much cultural theorizing revolves around a weak program in which theorists isolate what are viewed as key elements of culture in all social contexts, with an eye to making generalizations about the dynamics of culture in general and then trying to connect culture to generic properties of social structures and microlevel interpersonal processes. Robert Wuthnow's (1987) theory of culture meanings views culture as a system of codes and the mobilization of resources produce and sustain these codes and the ritual practices directed at these codes. Codes must reveal moral objects and programs, direct attention to individuals' real self as distinct from roles played, and distinguish between forces that are out of people's control and those that are controllable by individuals' actions. Ideologies are the most important system of cultural codes, specifying what should be, and organizing rituals affirming the ideology. The social world is viewed in ecological terms as consisting of competing

ideologies whose adherents mobilize resources and thereby seek to recruit members to their cause.

Pierre Bourdieu's distinction among various forms of capital – cultural, symbolic, social, and economic – is woven into yet another weak program that attempts to generalize about the dynamics of culture. Bourdieu's (1984) analysis of stratification represents a good example because it challenges elements of the Marxian tradition with its emphasis on dominant (upper), intermediate (middle), and dominated (lower) classes. These classes differ by the relative amounts of economic, social, symbolic, and cultural capital possessed by their members, but what makes the argument more interesting is that within each class there are 'factions' revealing varying amounts of capital. The dominant faction in all three classes always has the most economic capital for a given class, the intermediate faction has less economic capital but reasonably high levels of symbolic, cultural, and social capital, while the dominated faction within a class has the least amount of economic and often relatively high levels of only cultural capital. It is often the case that even though individuals and families are members of different classes, their relative place within a class as measure by whether they are in the dominant, intermediate, or dominated faction within their respective classes leads to cultural equivalence and hence a common identity that often makes them more comfortable around members of the same faction from different classes than members of different factions within their own class. This homology among factions dramatically changes class dynamics in a society.

There are other examples of weak programs (e.g., Fine, 2012; Turner, 2010–2013; Collins, 1975). The key characteristic of all of these programs is that they seek to isolate universal elements of culture and theorize their dynamics. In so doing, they downplay what is unique about a culture at a particular time and place and, instead, focus on what is universal about elements of culture in all times and places.

The Strong Program

The strong program argues for an emphasis on culture in its unique context. However, when seeking to theorize cultural times, many strong programs are transformed into weak ones because of the effort to isolate generic properties of culture and theorize their dynamics. Jeffrey Alexander (2004), for instance, has been among a group of scholars advocating a strong program, but in his own theoretical statement on cultural pragmatics, Alexander converts his more descriptive works into a weak program, which may be necessary if theory rather than description of culture is the goal. His main idea is that as societies become more complex and differentiated, a disconnect emerges between the texts, background representations, and scripts of culture (as described in the strong program) and the ritual performances of actors in front of audiences; and this disconnection often makes actions and performances seem artificial and strained. Thus, successful performances by actors involve re-fusing performances so that they decode culture and arouse emotions in the audience, thereby pulling individuals into culture in its more robust forms.

Whether a weak or strong program, it is clear that cultural theorizing has accelerated over the last two decades in sociology. Partly as a reaction against conflict theories that often portrayed culture as a 'superstructure' or adjunct to material conditions and social structures, theorists have shifted to emphasizing culture as a force in its own right. Both the weak and strong programs emphasize that the dynamics of interpersonal behaviors and social structures cannot be understood unless theories on the properties and dynamics of culture as force in their own right are first developed.

Critical Theorizing

Since its beginnings, sociology has devoted much theorizing to the problems of modernity. Whether Marx's analysis of alienation and exploitation, Weber's negative portrayal of rationalization, Durkheim's analysis of anomie, Simmel's analysis of marginality and other pathologies of differentiation, and modernity, sociology has always been a critical discipline. Some theorists have sought to remain value-free even as they offered criticisms of problematic conditions; others have remained explicitly normative, advocating for changes in those patterns of social organization viewed as oppressive and harmful to humans.

Critical theories in the mid- to late twentieth century became not only critical of social conditions but also of the scientific pretensions that sociological knowledge can solve these problems; and thus critical theory is critical of both society and science, which is often seen as part of the problem.

The Frankfurt School

Early critical theorizing began in the 1930s in Europe among a group of scholars in the Frankfurt School, even though some were not German or even based in Frankfurt. The basic problem they confronted was that Marx's emancipatory dream seemed dead as fascism spread across Europe in the build up to World War II. Indeed, Weber's view of social life as increasingly caged in a steel enclosure of rational-legal domination by the state and spread of bureaucracy seemed to be a better description of trends, leading scholars to ask: How is emancipation from such rational-legal domination to be achieved? The answer to this question became highly philosophical but, in essence, a variety of scholars from Gyorgy Lukacs, Max Horkheimer, Theodor Adorno, Antonio Gramsci, Louis Althusser, or others emphasized that, while transformative change may not be possible in the present, it was important to expose and analyze patterns of oppression and to propose possible solutions, even if change itself was not possible at this point in history. Later, under the intellectual leadership of Jurgen Habermas (1981–1984), (1973[1976]), the Frankfurt School of critical theory began to emphasize, again in a highly philosophical way, that the public sphere of the eighteenth century where ideas were once debated without recourse to dogma needs to be reconstructed in the modern world. Habermas proposes that the goal must be to move toward what he saw as 'the idea speech act,' composed of three main elements or implicit assertions that acts are the most effective means to ends, are correct and proper in accordance with relevant norms, and are sincere as well as authentic. If social conditions can increase the production of such acts, the public sphere can

be recreated, and solutions to the problem of invasion of the lifeworld or culture by the material forces of production and state management can be overcome. A new kind of communicative rationality can emerge, which can weaken the means-end rationality that has dominated modern societies.

Postmodern Critical Theories

For two decades, many sociological theorists were positing a new stage of societal evolution: the postmodern. Like modernists theorists, they emphasized that the new postmodern condition reveals structures and cultural forms that are harmful and pathological, and at the same time, they have been equally hostile to science as presenting solutions to these pathologies. Science presumes to offer 'a grand narrative' about the dynamics of the social world, but this narrative should not enjoy a privileged position. Moreover, the era of grand narratives needs to come to an end, which is somewhat ironical because postmodernism has proposed its own grand narrative. This narrative has two variants: economic and cultural.

Economic postmodernism emphasizes the pathologies generated by overaccumulation of capital and its rapid movement around the globe (Harvey, 1989; Jameson, 1992; Lash and Urry, 1994). Such movement has a number of harmful effects, including reducing the price of labor and hence income from work all over the world; increased financial speculation in meta markets that often collapse and cause longer term economic dislocation; decreases in the power of national governments to control their money supply; and decreasing capacity of government to control capitalists and key economic actors bent on self-interest rather than collective interests of people. Coupled with cultural changes arising from the overproduction of technologies of reproduction (i.e., media images detached from empirical reality), many new pathologies for humans and societies emerge.

Cultural postmodernists emphasize the spread of technologies of reproduction through media images and the increasing detachment of culture from its origins as it is marketed as a commodity (Baudrillard, 1994; Gergen, 1991). When culture is detached from its material base and becomes a commodity, it is trivialized and becomes a free-floating bundle of symbols that can be tried on for appearance and then easily discarded. This process is accelerated by the prominence of cultural images, as they spread around the globe via many sources of media reproduction. The result is that symbols have ever-decreasing meaning and capacity to provide anchorage for self and identity. Images can be acquired via media and/or purchased as a commodity in markets, tried on for their transitory effect, and then abandoned. The result is that individuals become over-reflexive about self, while at the same time incapable of forming stable identities that give their actions coherence and meaning.

Postmodernist theory has been in decline now for about a decade because some of its claims seem excessive, without clear empirical support; and since postmodernists rejected the epistemology of science, they would not accept tests of their theoretical assertions as useful. Yet, there can be no doubt that both variants are denoting important sociocultural changes, but whether or not they represent something new remains open to debate.

American-Style Critical Theories

In the United States, sociology has always had a critical slant. The first American sociologists all pledged allegiance to science, but only a few had any training in science and most were concerned with amelioration in the post–Civil War era. Contemporary American critical theorizing is almost always normative, suspicious of the claims of science, and concerned with exposing and eradicating oppressive social conditions experienced by ethnic minorities, later with the oppression of women, and more recently with the question of gays, lesbians, and increasingly transsexuals. These issues all became part of civil rights movements and then the ideologies and concerns of these movements were absorbed into departments and programs in American academia where the ideological zeal remains but is supplemented by more academic analysis, most of which has been suspicious of science. The result has been a mix of historical description, philosophical discourse, outright advocacy, and activism. Much of this American style critical theorizing is not really very theoretical, but some of it is, in the broader sense of 'social theory.' Feminism is now the most active critical approach; although ethnic relations remain a strong theoretical concern, they are spread among both those committed to science in sociology departments and those more concerned with activism in various types of ethnic studies departments and programs. GLBT issues can also be found in women's studies departments, and even a few programs explicitly devoted to these questions, as well as older, established departments such as English and History programs in the United States (for summaries, see Turner and Turner, 2013).

The State of Sociological Theory

Today, sociological theorizing is a very broad enterprise, and the review here can only touch upon the range of theoretical approaches. There are many variants of theory within any of the main approaches outlined in this article, but the main contours of theory today are at least clear. Since theory in sociology must examine all of social reality, from its micro- through meso- to macrolevels of reality, it inevitably must be a very broad subject area. Many of the structures operating at various levels are, of course, subject areas in their own right, as can be seen by other articles in this encyclopedia, but it is theorizing that brings them together more than any other subdiscipline within sociology. Yet, the barriers created by specialization – indeed, overspecialization – in sociology represent one roadblock to more comprehensive theories. Another, more fundamental roadblock is that many who consider themselves theorists do not consider scientific theory as possible, or even desirable. Thus, this epistemological split assures that theory will remain rather spread out and chaotic. Such has been the nature of sociology from its beginnings, and if anything, the split between science and its skeptics is greater than ever in the second decade of the twentieth century.

See also: Action, Theories of Social; Critical Race Theory; Critical Theory: Contemporary; Critical Theory; Darwinism: Social; Dramaturgical Analysis; Durkheim, Emile (1858–1917); Ethnomethodology, General; Evolutionary Sociology; Evolutionary Theory, Structure of; Exchange: Social; Field Theory; Functionalism in Anthropology; Functionalism, History of; Game Theory; Interactionism, Symbolic; Marx, Karl (1818–83); Marxism in Contemporary Sociology; Phenomenology in Sociology; Queer Theory and Intersectionality; Rational Choice Theory in Sociology; Weber, Max (1864–1920); Weberian Social Thought, History of; World Systems Theory.

Bibliography

Alexander, J.C., 2004. Cultural pragmatics: social performances between ritual and strategy. Sociological Theory 22, 512–574.

Alexander, J.C., Smith, P., 2001. The strong program in cultural theory: elements of structural hermeneutics. In: Turner, J.H. (Ed.), Handbook of Sociological Theory. Kluwer Academic, New York, pp. 135–150.

Baudrillard, J., 1994. Simulacra and Simulation. University of Michigan Press, Ann Arbor, MI.

Berger, J., 1958. Relations between performance, rewards, and action-opportunities in small groups (unpublished Ph.D. dissertation).

Berry, B.J.L., Kasarda, J.D., 1977. Contemporary Urban Ecology. Macmillan, New York.

Blau, P.M., 1964. Exchange and Power in Social Life. Wiley, New York.

Bonacich, E., 1972. A theory of ethnic antagonism: the split labor market. American Sociological Review 37, 547–559.

Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge, MA.

Burke, P., Stets, J.E., 2009. Identity Theory. Oxford University Press, New York.

Burt, R.S., 1992. Structural Holes. Harvard University Press, Cambridge, MA.

Chafetz, J., 1990. Gender Equity: An Integrated Theory of Stability and Change. Sage, Newbury Park, CA.

Chase-Dunn, C., 2001. World systems theory. In: Turner, J.H. (Ed.), Handbook of Sociological Theory. Kluwer Academic, New York, pp. 589–612.

Clark, C., 1997. Misery Loves Company: Sympathy in Everyday Life. University of Chicago Press, Chicago, IL.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap, Cambridge, MA.

Collins, R., 1975. Conflict Sociology: Toward an Explanatory Social Science. Academic Press, New York.

Collins, R., 2004. Interaction Ritual Chains. Princeton University Press, Princeton, NJ.

Collins, R., 2008. Violence: A Micro-Sociological Theory. Princeton University Press, Princeton, NJ.

Cook, K.S., Emerson, R., 1978. Power, equity, and commitment in exchange networks. American Sociological Review 43, 721–739.

Cooley, C.H., 1902. Human Nature and the Social Order. Scribner's, New York.

Dahrendorf, R., 1959. Class and Class Conflict in Industrial Society. Stanford University Press, Stanford, CA.

Durkheim, E., 1893[1933]. The Division of Labor in Society. Free Press, New York.

Durkheim, E., 1912[1947]. Elementary Forms of the Religious Life. Macmillan, New York.

Emerson, R.M., 1962. Power-dependence relations. American Sociological Review 17, 31–41.

Fine, G.A., 2012. Tiny Publics: A Theory of Group Action and Culture. Russell Sage, New York.

Friedkin, N.E., Johnsen, E.C., 2011. Social Influence Network Theory: A Sociological Examination of Small Group Dynamics. Cambridge University Press, New York.

Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.

Gergen, K.J., 1991. The Saturated Self. Basic Books, New York.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. Polity Press, Cambridge.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Anchor Books, Garden City, NY.

Goffman, E., 1967. Interaction Ritual. Anchor Books, Garden City, NY.

Goffman, E., 1974. Frame Analysis: An Essay on the Organization of Experience. Harper and Row, New York.

Goldstone, J., 1990. Revolution and Rebellion in the Early Modern World, 1640–1840. University of California Press, Berkeley.

Habermas, J., 1973[1976]. Legitimation Crisis. Heinemann, London.

Habermas, J., 1981–1984. The Theory of Communicative Action, vol. 2. Beacon, Boston, MA.

Hannan, M.T., Freeman, J.H., 1989. Organizational Ecology. Harvard University Press, Cambridge, MA.

Harvey, D., 1989. The Conditions of Postmodernity: An Inquiry into the Origins of Cultural Change. Blackwell, Oxford.

Hawley, A., 1986. Human Ecology: A Theoretical Essay. University of Chicago Press, Chicago, IL.

Hayashi, M., Raymond, G., Sidnell, J., 2013. Conversational Repair and Human Understanding. University Cambridge Press, Cambridge.

Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.

Heritage, J., 1984. Garfinkel and Ehnomethodology. Polity Press, Cambridge.

Hochschild, A.R., 1979. Emotion work, feeling rules and social structure. American Journal of Sociology 85, 551–575.

Hochschild, A.R., 1983. The Managed Heart: Commercialization of Human Feeling. University of California Press, Berkeley, CA.

Jameson, F., 1992. The Postmodern Condition. University of Minnesota Press, Minneapolis, MN.

Kemper, T.D., Collins, R., 1990. Dimensions of micro-interaction. American Journal of Sociology 96, 32–68.

Lash, S., Urry, J., 1994. The End of Organized Capitalism. University of Wisconsin Press, Madison, WI.

Lawler, E.J., 2001. An affect theory of social exchange. American Journal of Sociology 107, 465–481.

Lawler, E.J., Yoon, J., 1996. Commitment in exchange relations: a test of a theory of relational cohesion. American Sociological Review 61, 89–108.

Lawler, E.J., Thye, S., Yoon, J., 2009. Social Commitments in a Depersonalized World. Russell Sage, New York.

Lenski, G., 1966. Power and Privilege. McGraw-Hill, New York.

Levi-Strauss, C., 1979. Myth and Meaning. Schocken, New York.

Luhmann, N., 1982. The Differentiation of Society (S. Holmes, C. Larmore, Trans.). Columbia University Press, New York.

Malinowski, B., 1944. A Scientific Theory of Culture and Other Essays. Oxford University Press, London.

McCarthy, J.D., Zald, M., 2001. The enduring vitality of the resource mobilization theory of social movements. In: Turner, J.H. (Ed.), Handbook of Sociological Theory. Kluwer Academic, New York, pp. 533–565.

Mead, G.H., 1934. Mind, Self, and Society. University of Chicago Press, Chicago.

Milner, M., Walker, L.S., 2014. Expectations, emotions and status. In: Stets, J.E., Turner, J.H. (Eds.), The Handbook of The Sociology of Emotions, vol. 1. Springer, New York.

Niemela, J., 2011. Urban Ecology: Patterns, Processes, and Applications. Oxford University Press, Oxford, UK.

Parsons, T., 1966. Societies: Evolutionary and Comparative Perspectives. Prentice-Hall, Englewood Cliffs, NJ.

Parsons, T., 1971. The System of Modern Societies. Prentice-Hall, Englewood Cliffs, NJ.

Parsons, T., Bales, R.F., Shils, A., 1953. Working Papers in the Theory of Action. Free Press, New York.

Radcliffe-Brown, A.R., 1952. Structure and Function in Primitive Society. Free Press, Glencoe, IL.

Ridgeway, C., 1994. Affect. In: Foschi, M., Lawler, E.J. (Eds.), Group Processes: Sociological Analysis. Nelson-Hall, Chicago, IL.

Scheff, T., 1997. Emotions, the Social Bond, and Human Reality. Cambridge University Press, New York.

Schutz, A., 1967[1932]. The Phenomenology of the Social World. Northwestern University Press, Evanston, IL.

Simmel, G., 1956[1903]. Conflict and the Web of Group Affiliations (K.H. Wolff, Trans.). Free Press, New York.

Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, New York.

Spencer, H., [1874–1894]1898. The Principles of Sociology, vol. 3. D. Appleton, New York.

Thoits, P.A., 1990. Emotional deviance: research agendas. In: Kemper, T.D. (Ed.), Research Agendas in the Sociology of Emotions. State University of New York, Albany, NY, pp. 180–203.

Tilly, C., 1993. From Mobilization to Revolution. Addison-Wesley, Reading, MA.

Turner, J.H., 1975. A strategy for reformulating the dialectical and functional theories of conflict. Social Forces 53, 433–444.
Turner, J.H. (Ed.), 2001. Handbook of Sociological Theory. Kluwer Academic/Plenum, New York.
Turner, J.H., 2002. Face-to-Face: Toward a Theory of Interpersonal Behavior. Stanford University Press, Stanford, CA.
Turner, J.H., 2007. Human Emotions: A Sociological Theory. Routledge, London.
Turner, J.H., 2010–2013. Theoretical Principles of Sociology, vol. 3. Springer, New York.
Turner, J.H., Maryanski, A., 1979. Functionalism. Benjamin-Cummings, Menlo Park, CA.
Turner, J.H., Maryanski, A., 2008. On the Origin of Societies by Natural Selection. Paradigm Press, Boulder, CO.
Turner, R.H., 1962. Role taking: process vs. conformity. In: Rose, A. (Ed.), Human Behavior and Social Processes. Houghton-Mifflin, Boston, MA.
Turner, R.H., 2001. Role theory. In: Turner, J.H. (Ed.), Handbook of Sociological Theory. Kluwer Academic, New York, pp. 233–254.
Turner, S.P., Turner, J.H., 1990. The Impossible Science: A Evolutionary History of American Sociology. Sage, Newbury Park, CA.
Turner, P.R., Turner, J.H., 2013. American-style critical theories. In: Turner, J.H. (Ed.), Contemporary Sociological Theory. Sage, Los Angeles, CA, pp. 701–725.
Turner, J.H., Machalek, R., Maryanski, A. (Eds.), 2014. Evolutionary Analysis in the Social Sciences. Paradigm, Boulder, CO.
Wagner, D.G., Turner, J.H., 2013. Status theories. In: Turner, J.H. (Ed.), Contemporary Sociological Theory. Sage, Los Angeles, CA, pp. 373–402.
Wallerstein, I.M., 1974–1989. The Modern World System. Academic Press, New York.
Weber, M., 1968[1922]. Economy and Society (G. Roth, C. Wittch, Trans.). University of California Press, Berkeley, CA.
Webster, M., Walker, L.S., 2014. Expectations, status, and emotions. In: Stets, J.E., Turner, J.H. (Eds.), Handbook of the Sociology of Emotions, vol. 2. Springer, New York.
Wright, E.O., 1997. Class Counts. Cambridge University Press, Cambridge.
Wright, E.O., 2010. Envisioning Real Utopias. Verso, London.
Wuthnow, R., 1987. Meaning and the Moral Order. University of California Press, Berkeley, CA.
Zald, M.N., McCarthy, J.D., 1987. Social Movements in an Organizational Society. Transaction Books, New Brunswick, NJ.

Sociological Traditions

Charles Crothers, AUT University, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.A. Tiryakian, volume 23, pp. 15824–15829, © 2001, Elsevier Ltd.

Abstract

Scholarly work in sociology in different places and at different times is guided (often simultaneously or sequentially) by a variety of traditions that take different forms (e.g., sometimes in quite formal 'schools') and emphasize different aspects of scholarly activity (metatheoretical, theoretical, methodological, ideological, etc.) and that rise and fall according to a range of societal, scholarly, and other conditions. Often, such traditions are national or regional, or at the very least are shaped by national or regional contexts. Apart from providing an analytical framework for understanding the role of traditions in sociology, this article also explores the broad course of historical developments of some major traditions.

The Importance of Traditions

Almost inevitably, sociologists' decisions are strongly, albeit often unconsciously, influenced at each point in their projects by received or developing traditions, paradigms, and lines of thought and are socially influenced by the 'social embedding' of such cognitive structures in 'schools,' 'theory groups,' 'research networks,' and other forms of intellectual social organization, some organized along national lines. Many sociologists would argue for the importance of traditions as providing the continuing backbone of sociological thinking, and are prepared to invest energies in their maintenance and communication, and often in debating the merits of alternatives. This is so particularly in the teaching activities of sociologists in which appropriate typologies are constructed and reconstructed.

As Levine (1995: p. 13) has remarked, sociology was officially born carrying with it a schema of its own history: for example, Comte coined the term and simultaneously laid out his version of its trajectory. Yet it seems strange that the very discipline whose business it is to pry into the affairs of other groups knows so little about itself that there is only very scattered data on its membership, their interests, and their theoretical and methodological positions. There is much debate about theory and types of sociology, but little knowledge of what views are held by the vast majority of workaday sociologists who presumably comprise the audience for these debates.

There is a considerable range of views about the nature and scope of traditions, which change over time and across national and other social contexts. Moreover, it is unclear what the extent of consciousness is among traditions, and the degree to which sociologists are able to follow through with proclaimed choices. Some sociologists clearly cleave to a particular approach, but at the other extreme, some deny being influenced by any specific tradition.

Much more has been written about national or regional sociologies. Several volumes have been collated, which include country-by-country accounts, although not usually country-by-country comparisons (these include several 'World Handbooks' with country chapters, e.g., Mohan and Wilke, 1994). There seem to be many claims to regional or country traditions, but not too many of these seem to survive closer examination. Many national sociologies follow similar trajectories. As with the overall trajectory of sociology, many countries did include in their prehistories of sociology particular figures who loomed over later sociological developments and who by projecting their particular scholarly idiosyncrasies bequeathed a particular national flavor to the sociology of their country. Such scholars may have been jurists, literary or philosophical theorists, or more generic social scientists. However, such national traditions tended to have short lives. In the postwar period, sociology worldwide was invaded by American sociology with its empiricism (symbolized by survey research and a 'scientific' approach) and structural–functional theory. However, by the 1970s, other traditions percolated out from the first world core of sociology to challenge the hegemony of the American approach: especially various macro- and microsociologies. This widening of perspectives, too, allowed room for the development of national traditions. But most national sociologies are probably best characterized as particular 'mixes' of the then current metropolitan (and therefore world) sociological traditions. Nevertheless, such local versions of the world tradition often have a local, national, or regional 'flavor.'

Conceptualizing Traditions

The term 'tradition' emphasizes rather too much the connotation of respect for past thinking. But 'school' implies rather too much a formal organization, leadership, and even intellectual control from a leader (or leaders). The ambiguities around Kuhn's (1970/1962) fecund term 'paradigm' have led to a major commentary industry. Here, the term 'national sociological tradition' refers to any cognitive formation organized along national lines that lends consistency to aspects of thinking among sociologists, whether the doctrine concerned is unique within sociology or is also shared by wider groupings of intellectuals or scientists.

Donald Levine (1995) has provided a useful imagery of what lies at the heart of any sustained and cumulative sociological discourse. He sees each tradition as a conversation, a dialogue between networks of sociologists who tend to share particular assumptions about social reality. Since, as with other largely text-based intellectual activities, time and space are limited as Barriers conversations can readily be held across as well as within generations and across countries (as well as within them). Levine defines traditions as inter- and intragenerational

conversations amongst scholars. The uniqueness of any tradition must usually be placed within the context of the other traditions then pertaining, and on the relations among these traditions. Usually, contemporaneous traditions are in competition, and sometimes in conflict, although they may also ignore each other.

There is a tendency among historians of social theory to restrict their concerns to traditions that are theoretical. But sociological work (and thus any fully formed tradition) must inevitably cover, at least, each of metamethodological, conceptual, methodological, data, and ideological aspects, including interrelating with the general population. In terms of their internal state, traditions may vary in terms of their degree of historical development, the tightness of their formal development, the tightness of fit of the linkages between the theory espoused and the facts considered important in the tradition, the closure of their boundaries, the degree of reflexive consciousness with which they are held by adherents, the moral tone that is pursued, etc. In his study of types of sociology, Menzies usefully distinguishes between 'theorist's theory' and 'researcher's theory,' which have different characteristics. Further, as Schumpeter pointed out, some of these schools are 'virtual' being merely groupings of writing, whereas some are real: based on ongoing social activities among its members (albeit at a distance). In addition, at least for socially real traditions, there are more social aspects to traditions that may also affect their trajectory: the degree of social cohesion, organizational contours, shared cultural assumptions of adherents, etc., and their fit with the environing culture and society.

The several aspects of a tradition can more formally be classified into the following:

Levels/ dimensions	*Substantive*	*Methodological*	*Moral*
Ideas	Concepts, theories, etc.	Methodology, assumptions, etc.	Moral vision
Facts	Findings, empirical generalizations, recommendations, etc.	Methods	Practical

Important writings in the (historical) sociology of sociology have drawn attention to each of these major dimensions, for example:

- Kuhn: the importance of underlying conceptual assumptions and models, and the intellectual community upholding the 'paradigm.'
- Gouldner: the importance of underlying conceptual assumptions and models, and the grounding of views in personal experiences.
- Freidrichs: the importance of the underlying moral views.
- Platt: importance of traditions of social research methodology.
- Levine: the importance of the type of 'organizing model' (narrative) about the trajectory of sociological development held by the historian.
- Burawoy: public, academic (professional), and critical approaches.

Classifications of theory are not only part and parcel of normal sociological argumentation but also are developed by those textbook writers whose contribution it is to provide classifications and histories of theories. Such classifications then tend to become built into classroom teaching, especially in theory (often taught as history of theory) and history of sociology courses, and begin to take on a life of their own. The validity of such classifications is often not tested.

Stephen Cole (1992) has drawn attention to the difference between the 'core' and the research front. He suggests that the cognitive structure of most sciences consists in a few theories and procedures over which there is complete consensus and which can be readily presented in textbooks, and a 'research front' marked by minimal consensus and much diversity of approaches, methods, and models. Another conception is that there is a 'world' of textbook sociology that has a life of its own, with often a minimal connectivity to the research front. Traditions may live rather different lives in each of these two different worlds.

As in other disciplines, there is international interest in how sociology is progressing in various parts of the world and several (edited) volumes have been constructed at fairly regular intervals that carry this out using a national frame of reference. However, neither the assumptions of national differences inherent in such an approach nor the reliability of judgments about national peculiarities drawn by observers is adequately tackled.

Traditions may have trajectories. Onto Kuhn's schema, which contrasted periods of 'normal science' and 'revolutionary periods,' philosophers of science have grafted other terms such as 'progressive and regressive shifts,' which spell out the progressions in more detail. Sociologists have also developed 'phase models' of the development of specialties – which might also apply to sociological traditions (cf Crane, 1972; Mullins, 1973; Rule, 1997). Traditions and specialties are often built on a slowly developed platform, and then break away into a fast-developing growth phase, before hitting a plateau and in some cases then declining. Institutionalization, and obtaining a secure supply of requisite resources and recruits, is necessary for a tradition to be sustained (cf Turner and Turner, 1990). Traditions often forge strong social ties among members, with master–apprenticeship relations being essential for their longer term growth (cf Mullins, 1973; Collins, 1994).

Alexander and Colomy (1992) have developed a terminology for charting the complex 'progress' of a tradition, which includes

- Elaboration (i.e., refinement), proliferation (extension of scope), and revisions (relating back to the core) as three lines of specification.
- Reconstruction, which involves acknowledging differences with the founder and openings to other traditions.
- Tradition creation: "The essence of tradition-creation is the synthesis of elements drawn from several existing and often competing intellectual paradigms, with the aim of generating the theoretical core of a new school" (p. 37). Alexander has argued that over time particular strategic cognitive patterns are likely to emerge, with disciples, for example, tending to soften the particularities that the tradition's masters tended to emphasize. Mulkay develops

a more radical argument whereby "Theoretical development is regarded as being neither continuous nor, in any direct way, cumulative. Instead, it is seen as arising from a number of discrete and intermittent theoretical reorganisations, which centre upon new strategies devised as replacements for the unsuccessful policies adopted by prior theory" (1971: p. 3).

Predisciplinary National Traditions of Social Theory

Levine argues that more light is thrown on the development of sociological thought by endeavoring to grasp the national channels in which it flowed for some centuries. Levine's schema builds on the (often binary) classifications of traditions set out by other writers, but pursues a more detailed examination. He "… define[s] them primarily as national traditions, for two reasons. The originative figures of modern sociology mainly cite fellow nationals, as, for example, Halbwachs is likely to cite Rousseau; von Weise, Simmel; Park, Sumner. More important, over the generations they reproduce what are palpably national characteristics. Moreover, when they engage in dialogue with parties from other national traditions, they do so, openly or by implication, in a more contrastive mode – as when Durkheim explicitly contrasts his French discipline with British and German traditions" (1995: pp. 99, 100). A national tradition then tends to submerge within a more universalistic discourse.

Each national culture tends to make similar ontological assumptions about the nature of social reality, and how it might be known (Levine, 1995). These assumptions underpin social science work in that country, and these views often have been articulated by important philosophers – who then act as something of a 'skirmish line' clearing the way for, and shaping, later sociological thought. In particular, cultural choices tend to be made between stressing the individual level as ontologically prior (as in the United Kingdom), or the collective level (France), and between an objective approach (as in both the United Kingdom and France) compared to a more subjective approach (as in Germany). Although there is a long-term consistency in the development of these approaches, there can be considerable variation within them, and they are often formed in part through a conversation with other theoretical traditions.

Levine identifies the various national traditions in sociology as follows:

- Aristotle, representing the Hellenic tradition, is concerned with the way in which different constitutional arrangements of societies (city states) shape the conditions for achieving human virtues, and thus human happiness, with the achievement of that potential arising from the deliberative judgments of the participants.
- The British tradition emphasizes the naturally sociable disposition of individual humans: for example, their preoccupation with the properties, rights, liberties, and utilities of individual actors, and then the conduct of the actors.
- The French tradition emphasized that "… society formed a supra-personal entity with properties and needs of its own, above all needs for normative regulation and solidaristic integration" and that it was a natural phenomenon.
- The German tradition emphasized human freedom as an essential element of moral decision making, and also affirmed the creative power of language as a distinctive feature of human action and morality.
- The transnational Marxian tradition emphasized elements elicited through critique of several of the other traditions in order to combine, in his final synthesis "… Kant's notion of self-determination, Hegel's notions of collective historical development, French notions of associated humanity and social classes and the British conception of the competitive pursuit of individual interests in the marketplace" (p. 222).
- The Italian tradition sought laws about societies as natural systems with the prominent features of these systems being the functioning of ruling classes or elites.
- American pragmatism stressed social activism disposed to solve problems on an experimental common sense basis.

Although the attention of historians of sociology is particularly drawn to examining theoretical traditions, there are also research traditions (or more broadly methodological traditions) that may not at all be strongly linked with the more theoretical traditions. The State, as the key institution in developing statistical information, is particularly important in shaping the types of empirical research and methodological developments that eventuate in a particular country.

Undoubtedly, the most important methodological breakthrough in the social sciences was the German historian's concerns with methods for validating the reliability of documents. This led to a more widespread tide of heightened methodological standards. In addition, Levine (1995: p. 276) suggests that "The empirical traditions also bore the mark of national dispositions, if not in such a pronounced from as the philosophical ones. England led the way with social surveys, systematic investigations of living and working conditions, mainly of members of the working class. France and Italy pursued the collection of national social statistics, work that enabled Durkheim to lead off so impressively with his analysis of divorce rates, educational levels, mental illness data, religious affiliation and the like in Suicide. Germany pioneered the experimental manipulation of subjects and also the systematic collection of ethnographic data in broadly defined culture areas. The US pioneered in producing census data and later in systematic gathering of information through personal documents and direct observation as well as interviews."

There has been relatively little recognition of these broader (philosophical) traditions of thinking that influence how sociological work is carried out. One theme has been that of positivism, which has been deployed by many more specific sociological traditions. Bryant usefully points out that positivism has cycled through several variants, including the French version of Comte and later, Durkheim, and the Austrian approach, before being developed with rather different emphases in its more modern form in the United States over the last century. Another theme that has attracted much discussion is the Marxian approach, which seemed to peak in the 1970s. This approach was often quite visible and self-conscious – even sometimes setting itself outside the normal boundaries of bourgeois sociology. Several writers (e.g., Giddens) have struggled to identify a postpositivist philosophical mood.

Midcentury US Sociological Traditions

However, these various European traditions tended not to be well institutionalized, with slender holds on university positions, and these faltered after World War 1 undermined several national groupings. The United States then became the setting in which earlier sociological traditions were selectively retained, and further sociological traditions unfolded.

In the mid-1930s, Talcott Parsons endeavored to establish a more solid cognitive base on which future social theory might be built through consideration of a mix of theorists (each perhaps representing different traditions, although it is not a point in Parsons's argumentation), which he argued shared common features critical of utilitarian doctrines and instead positing a more sophisticated 'theory of action': Marshall, Pareto, Durkheim, and Weber (with suppressed attention also to Simmel and to Marx). By the immediate post–World War 2 (WW2) period, this synthesis was broadly accepted, even if its particular manifestation in his writings was not. This gathering emphasis on theory was reinforced by many refugee social theorists who had migrated from Europe.

This theoretical synthesis extended into a sociological functionalism (or structural functionalism) is seen as the major approach that dominated sociological discourse in the 1950s and the 1960s, beginning earlier than this and certainly carrying on at least into the 1970s. Alongside this approach was the 'loyal opposition' of symbolic interactionists mainly concerned with microsociological processes. However, with the social upheavals and movements of the late 1960s and certainly during the 1970s, the hegemony of structuralism was seen as being challenged from below and above. Microsociologies became more fashionable, but so too did macrosociologies with (Weberian) comparative/historical and more radical Marxist approaches becoming more prominent (with a milder version lying in a postulated dichotomy between consensus and conflict sociologies).

On the methodological side, content analysis was developed, especially in the context of WW2 and the Cold War when direct access to totalitarian countries was denied and more indirect means of study were especially required. As consumer choices became more important and with the spread of advertising to media such as radio where the audience was not readily identified, survey research became increasing important as a method and was increasingly institutionalized. It enabled the scale of study to be extended from communities to nations (and later across the world). Surveys became inextricably linked for many sociologists as the landmark methodology of their discipline. The methodological emphasis of US social science, flowing from a broadly 'positivist' philosophical orientation, led to the very considerable systematization of social research methods there in the mid-nineteenth century and this led to the establishment of a tradition of social research methods texts that codified methodological knowledge. Moreover, this established American sociology then was actively exported around the world as a model for pursuing sociology.

However, closer examination of available content analyses of the published sociological literature shows that while functionalism was strongly established in the textbook literature, it failed to penetrate far into the research front of sociology. The many meanings of functionalism confuse the picture, and in particular diverted attention from the structuralism which underlay so much theoretical analysis and empirical research in this period and subsequently.

Contemporary Sociological Traditions

In the past quarter century, a much wider range of theoretical material has been incorporated into sociological theory, and the role of continental and British theorists has become far more prominent. However, the reception of such theories in other countries can be quite different to that accorded them in their home country: for example, some of the more successful French theorists have occupied marginal academic positions in France but have been accorded much emulation in Anglo–Saxon countries.

Sociological traditions were rewoven in the England-domiciled work of Giddens, a recent theorist sensitive to a wide range of traditions of sociology. Giddens began with a series of essays taking into account each of the received traditions mentioned earlier in this section, largely criticizing their deficiencies and also winnowing out the useful residue that might be reclaimed from their work: he began reviewing the received traditions (functionalism, materialism) but then examined a wide swathe of traditions in order to recruit appropriate ideas for sociology (including ethnomethodology, hermeneutics, and poststructuralism). The useful material he assembled was then incorporated in his own 'structuration theory.' In France, Bourdieu accomplished a similar synthesis with its own unique features and a better connection with social research work.

Consequently, by the 1980s, a considerable agenda of approaches was on the table: one stream emphasized Marxian approaches and more generally political economy, or conflict sociology. There was a rising tide of 'subjectivism' and focus on the individual social actor. In their useful summative presentation, Giddens and Turner suggest (problematically) that the changes were largely driven by a revised underlying philosophy of science imbibed by sociologists:

> Over the past two decades, however, a dramatic change has occurred. … developments in the philosophy of natural science have inevitably influenced thinking about the social sciences, while accelerating an increasing disillusionment with the dominant theories of 'mainstream social science.' The result of such changes has been a proliferation of approaches in theoretical thinking. Traditions of thought that previously had been either little known or ignored have become much more prominent: phenomenology, particularly associated with the writings of Alfred Schutz; hermeneutics, as developed in the work of such authors as Gadamer and Ricoeur; and critical theory, as represented by the works of Habermas. Moreover, older traditions of thought, such as symbolic interactionism in the United States and structuralism or post-structuralism in Europe, have more recently developed types of thinking, including ethnomethodology, structuration theory, and the 'theory of practise' associated in particular with Bourdieu. … There continues to be something of a 'mainstream', even if it is navigated by fewer than before. Parsonian structural-functionalism, for example, still exerts a strong appeal and, in fact, has undergone a considerable revival recently in the writings of Luhmann, Munch, Alexander, Hayes and others. (Giddens and Turner, 1988: pp. 2–3)

Just as this agenda of different traditions had become well established, especially within American sociology, a further wave of social thinkers came to the fore in European sociology. Whereas the earlier wave of groupings operated under doctrinal titles, albeit closely linked with particular key figures, the new round of thinkers were more individualistic. Some of the major figures clearly included Foucault, Derrida, Giddens, and Bourdieu. While some attempts have subsequently been made to name and classify the approaches adopted by these later-day theoretical saints, such classificatory bundlings have, however, been fiercely resisted by their protagonists. Over the past two decades, two rather more broad approaches also gained considerable notoriety and influence – although their demise has since begun – poststructuralism and postmodernism.

National Subtraditions

Besides social science–wide or discipline-wide traditions shared by sociologists, some traditions are specific to particular specialties. For example, addressing the specialty area of the sociology of science, Zuckerman (1988: p. 512) comments that there are not only different viewpoints but also that she sees (p. 513) these differences in views held within the specialty as flowing from national perspectives, especially between the United States as opposed to UK/European sociologists of science. With many exceptions, the North American approach tends to be more research orientated, functionalist, and 'positivist,' whereas the European approach is characterized as more comparative/historical and 'critical.' Other specialty areas offer other nationally specific patterns (e.g., see the chapters in Quah and Sales, 2000).

From time to time particular specialties seem of prime importance as pace setters at the cutting edge in setting traditions, and these may reflect national emphases. For example, in the 1960s, the 'new criminology' emphasized the application of symbolic interactionist and also Marxist approaches to the sociology of deviance, and this had broader implications for sociology as a whole. Another specialty area that has been of importance is that of social stratification, which has been a particularly contested area in which competing sociological traditions have been challenged to exhibit their causal efficacy.

Among genuine regional sociologies, the outstanding example is the dependency school of Latin American social scientists, although it is as much claimed by economics as by sociology.

For example, in the development period of functional analysis, one of its key contributions was in the area of stratification with the notion that stratificational orders were of functional consequence for their societies by motivating the filling of key societal roles by people with higher skills and ambitions since incumbents of such positions were rewarded by higher social and economic rewards.

Cross-National Integration of Traditions

Intratradition conversations usually allow fine-tuning of differences, whereas intertradition debate can be a major influence on the development of national traditions. Levine places particular emphasis, for example, on the way "the divergent postulates that underlie … persisting differences were honed and deepened in the course of centuries of mutual confrontations between British and French social theorists" (1995: p. 173). He then briefly sketches interactions involving Montesquieu against Hobbes, Rousseau against Hobbes, Smith against Quesnay, Comte against Smith, Mill against Comte, Spencer against Comte, Durkheim against Mill, and Durkheim against Spencer. More generally, Levine discusses links amongst German, French, Italian, British, and Marxian traditions. The structure of such intergeneration interactions is largely determined by the differential start times of each of the national traditions, which broadly has the pattern of British, French, German, Marxist, Italian, and American. Some patterns of 'alliance' can be seen in which some groupings of national traditions band together against others, while sharing internal differences. In particular, Levine sees "The formidable German defence of subject-orientated assumptions against Anglo-French support for naturalistic assumptions… [as originating] … one of the persistent fault lines in modern social science. Such dialogues can be effective in sharpening differences, as much as they lead to exploration of commonalities."

Transnational Contexts

There are several mechanisms through which national traditions are transcended: especially through international scientific associations and conferences. These can be at different levels of scientific intensity – merely diplomatic exchange to joint attention to scientific issues. International gatherings can be an arena for parading (and even exacerbating) differences or attempting to meld and conconstruct. The United Nations Educational, Scientific and Cultural Organization–supported International Sociological Association was a major sponsor of several national sociologies and perhaps this support carried with it the seeds of particular approaches to sociology.

However, forces conspire to keep traditions separate from each other. One mechanism is sheer mutual ignorance. The classic example undoubtedly is of the mutual unawareness of Durkheim and Weber – two giants of sociology working at exactly the same period and separated by only a few hundred kilometers of physical distance, although perhaps sheltered behind two only partly open national contexts. Another puzzle has been how long it took statistical methods appropriate to social data to emerge. Statistical methods were (famously) largely innovated in late Victorian England at a time when there was much social research, but measures such as the correlation coefficient were only pressed into sociological service (at Columbia University by Giddings in particular, see Camic and Xie, 1994) a bit later and on the other side of the Atlantic. In a long and almost despairing essay, Goldthorpe (2000) traces the failure of probabilistic statistical treatments of social phenomena to match with rational choice theory. In each of the main national traditions, the possibilities loomed but were never consummated. While organizational reasons did limit the linkage, intellectual barriers were more significant.

In some situations, rewards flow from establishing difference rather than trying to advance sociology cumulatively by carefully building on the work of others. In Lemert's portrayal of the French intellectual scene, for example, he argues that the intense competition for the spotlight encourages the celebration of difference, and underplays the constructive engagement between attentive sociological viewpoints to confront and perhaps reconcile overt differences. Indeed, Lemert argues, appropriate 'rivals' are not even explicitly named as audiences are sure to pick up subtle references.

Rivalry between traditions is often 'social': driven by interests in acquiring resources, recruiting bright students, and catching the attention of policy makers, funders, or the intelligentsia more generally. For example, the 1960s' rivalry between Chicago, Columbia, and Harvard was not necessarily combative but, nevertheless, was underlined by snideness and stereotyping.

The Shaping of Traditions: A Comparative/Historical Account of Factors Shaping Traditions

The rise and fall of (national and other) traditions are shaped by various cultural, ideological, political, institutional, cognitive, and social factors both generally and also in different ways in different national contexts. Such external influences may override some of the internal dynamics of traditions (which are often international).

The broadest influence flows from culture. Cultures stressing the importance, in both the natural and social realms, of the acquisition of rational explicit scientific principles and of empirical fact finding, and even more importantly the importance of developing systematic ways of interfacing the rational and the empirical, are much more likely to foster successful social science. An important influence on the development of social science is the model of natural scientific inquiry admired in that nation by its social scientists. Each of the main traditions defers to philosophers, and also scientists of more general importance, whose understandings of science were highly consequential for work in that tradition. For example, Newton's views on science had a particular resonance within subsequent British thinking.

Aspects of religious thinking in a country also affect the development of science and social science. Prior to the reformation and enlightenment, secular social thinking was often discouraged. Moreover, "In Protestant countries close relations developed between intellectuals and churches. Intellectuals were harnessed in the conflict with Catholic ideas and politics; there was more room for debate since (some) Protestant religions were not anchored by a central dogma; and since Protestant clergymen could raise families, intellectual dynasties could be more readily formed. Thus, in England, and also the Netherlands and Scandinavia, scientific innovation was linked with religious debate. However, in France secularisation of intellectual culture took place with support from the state and the court, and did not involve the development of scientific thought, since literary genres were dominant. Therefore scientization without secularisation in contrast to secularisation without scientization" (Heilbron, 1995: pp. 63–64).

Different cultures house much the same range of ideological perspectives, but some national consistencies can be found. French ideology more often stresses radical change, drawing on its rationalist heritage, whereas in the United Kingdom and the United States, emphasis is more on piecemeal reform, flowing from a strain toward empiricism. In Germany, ideology is often idealist, humanist, and antipositivist.

Each culture has somewhat different ways of portraying what is covered within the realm of the social sciences and what boundaries are set. In Germany, *sozialwissenschaften* is a broad conceptualization, whereas in the Anglo–US world, a sharper distinction is usually drawn between the social sciences and the humanities. France is more complex, with economics located within Faculties of Law, whereas the other human sciences were located within the broader humanistic framework of Faculties of Letters. Which particular disciplines are separately identified has also differed, with some continental university systems often presenting more 'policy-orientated' types of knowledge; for example, demography, criminology, and sociography in the Netherlands. More recently, the distribution of disciplines that developed in American universities has successfully diffused worldwide, and has become the norm against which the array of disciplines in other countries came to be measured. The main disciplines of social science, including sociology, tend now to have an American imprimatur. However, even if the general approach of the discipline has been set by its grounding in American institutional structures, the content of sociology has more recently been more often influenced by continental sources (and especially French-based social theory).

Each culture understands the relationships between its component 'social fields' rather differently. Although most modern societies have in common separate economic, political, and social realms, together with many minor arenas of social life (e.g., Legal systems), how each is constituted and what its relationships are to other spheres may differ considerably, and this has consequences for social science work in that society. The various disciplines tend to have a particular interactive relationship with one or other of the social realms in that society, with the social science discourse being in part 'constitutive' of that realm, and in part being shaped by it. These schema constitute 'deep structures' that may implicitly guide the development of particular national traditions over many generations. For example, in France, the state has long been a very central and powerful institution, although there has been a strong discouragement of scholarship in political science. In France, society ('the social') became distinct from church and politics, and then the economy, but in Germany, the distinction was resisted. The possibilities of the development of sociology were shaped by these differences.

There is also a link between national sociologies and their colonially linked sociologies: so that, for example, British or French influences shape the nascent sociologies of a range of other (excolony or influenced) countries.

Regional sociologies often stand behind national sociologies – either explicitly or not – and although geographical contiguity is often involved, common language can be even more powerful. Latin American sociologies have many similarities, and in contemporary Europe, considerable effort

is being put into constructing and critiquing the development of regionally based sociological viewpoints.

During the period of German university development from the late 1700s through the mid-1800s, literary and artistic intellectuals were not accommodated in universities. One result of this appears to have been "the independent non-university intellectuals became hostile to the new professorial form of knowledge production" (Heilbron, 1995: p. 24). This involved a Romanticist rejection of cold hard facts, which developed alongside, and in reaction to, the professionalization of knowledge, and provided an alternative and oppositional stock of intellectual resources, which has most recently been drawn on in the development of postmodernism.

The policy process is different in different states. In the United Kingdom, fact finding was institutionalized in the role of Royal Commissions, inspectorates, and social reform research associations, but these were weakly linked (at least as far as formal ties are concerned) with the policy making. Social network ties among various members of the elite could provide channels for information to percolate to policy makers. However, links between social researchers and policy makers have remained ever since generally at arm's length. In Germany, there was also considerable state involvement in statistics collection. On the other hand, in the Netherlands, a tradition of longer term policy making developed and this was able to articulate with a slew of empirically orientated research approaches such as sociography and social geography, which were particularly well developed there. The interest of states in social research and their capacity to promote it and then to utilize findings varies considerably.

Besides the state itself, social sciences traditions may be shaped through interaction with a range of coalition partners, including scholarly and/or professional organizations of social scientists, the social science–based semiprofessions, and a wider array of social movement organizations. In many countries, there has been a development of semiprofessions, at least partly based on sociological knowledge (social workers, planners, nurses, librarians, psychologists, economists, educators, and teachers), and their differential association with sociology has affected the development of national traditions.

A range of coalition partners, which differ in particular circumstances, may shore up the development of social science knowledge, especially where there is a cognitive and moral affinity. Such partners may include political parties, trade unions, probusiness groupings, welfare reform groups, or more generally, social movements. Often, some aspects of sociology relate to these in terms of some 'discursive affinity,' an overlapping of key concerns, and some basic similarity in cognitive assumptions and terms. In such circumstances, the sociological work provides some of the conceptual elaboration and/or the social information required to support the program of its ally, while the ally may assist in providing treatments of what issues are problematic, in providing empirical material (e.g., access to research sites), and in mobilizing resource support, plus providing a more general legitimacy. This relationship is often strengthened when the partner obtains parliamentary power or is in government. More recently, think tanks have been set up to mobilize social science knowledge for more specifically ideological purposes, especially in support of the doctrines of neoliberalism. Supporting social movements are often especially important in the internationally linking of national traditions; for example, Marx bequeathed his writings to the German socialist movement (which gained a parliamentary foothold in the 1880s and the 1890s), which harbored them to display for widespread trade union and academic utilization at the turn of century.

The effects of different institutional forms and the material basis of the provision of resources have been especially important for fueling the differential development of sociology. A range of institutional forms have been relevant in different times and places: coffee houses, salons, associations, university teaching departments, and research institutes (cf Coser, 1965). Salons and coffee houses can be significant for the flexible innovation of new ideas. Universities can be important for providing a more secure and longer staying environment, with systematization required for teaching and some degree of system rational allowed and pushed for. University settings propel specialization, and divisions of labor, especially in the form of formal development of separate disciplines. However, university teaching departments are not necessarily appropriate institutions to support larger scale research so that the tackling of larger topics, and also policy research, often requires the development of research centers where a specialized division of labor can be built up and resources for particular products mobilized.

Different national university systems have provided different contexts for the development of sociology. The German universities were reformed from the 1820s on and launched a range of more systematically based scientific work, especially in philology and then extending especially into history, which was placed on a far more scientific footing. The French system, which was not revivified until the 1870s was (and still is) highly centralized, which can mean the rapid institutionalization of a particular area of knowledge, although the centralization can prove intellectually stultifying. As with natural science, the American university system in which presidents have strong power to develop new areas, where there can be fierce competition for prestige among institutions, and where (at least in larger universities) the appointment of several full professors in each department fosters a democratic climate and a diversity of lines of research, seems to have been a particularly successful environment within which social science, including sociology, has flourished. The timing of reform to university systems seems also to have its own period effect: witness the intellectual outpourings follow in eras of educational reform from Germany/Prussia after the 1780s and France during the 1880s and after.

One particular design feature that supports innovation seems to have been the importance of role hybrids; those with one foot in practical concerns and the other in a setting allowing for systematization of ideas are often especially innovative situations. A particular sociological example was the early development of survey research in which academic sociologists played midwives to the more sophisticated methodological development of the research technique.

The institutions within which social science development takes place may also have an effect on the dominant cognitive style through which social knowledge is produced and debated.

German scholars often were ensconced in universities, whereas French scholars debated within scholars. As a result:

> Whereas the German intellectual was systematic, scholarly, even pedantic, the French intellectual tended to be orientated to science as well as to political controversy and to be brilliant and lucid as well as facile and flowery in exposition. (Collins, 1994: p. 14)

Cognitive properties may affect the development of national traditions, where there is reliance on subject matter that is strongly localized; for example, the study of languages seems to have nurtured a considerable degree of longevity of localized scholarship. All cultures are permeated with a reflexive 'folk knowledge,' but one of the earliest arenas for the development of social science knowledge were the many prototypical 'folk languages' in which 'native scholars' extracted some of the formal principles of their language in order usually to in turn use these principles in instruction and structuring of the language itself by guiding its development.

Conclusions

Since the long nation-based period of development of sociology throughout the nineteenth century, the national traditions have tended to become overwhelmed by more recent international developments in sociology. Nevertheless, distinct traces of the older national traditions can be found, and the legacies of particular sociological traditions can often been linked back to these national traditions.

During the period of the most active development in our current social science disciplines, innovative work was largely carried out in national traditions, albeit within a context of considerable international exchanges of views. Some of the key intellectual/cognitive features of these national traditions can be traced back through some centuries of intergenerational conversations. It can also be shown that differential institutional arrangements involving the state, other organizations, and the institutional features of knowledge-producing structures have also shaped social science work in particular national jurisdictions. Although the high period of the play of national traditions has largely passed and social science now seems to be relatively universalistic, national contexts still continue to shape the work of social science in each national community.

Sociology tends to a complexity that sustains separate traditions. As well as maintaining an *extensive* scanning range of research interests tendencies to differentiation are aggravated still further by sociology's theoretical and highly developed self-consciousness and reflexivity. A broad problem is that there are somewhat inadequate mechanisms for bringing traditions together. Part of this arises because of the essential empiricism of US sociology (albeit laced with middle-range theories), which is globally hegemonic, now abetted by poststructural approaches but which fails to be too concerned with theoretical synthesis. Another general barrier to more active synthesis is the condition under which continental social theory is produced, which also too often stresses on the idiosyncratic features of theories and theorists, rather than cumulation. This reflexivity about traditions has allowed sociologists to maintain some degree of control over their own intellectual concerns. But, the issue still remains – do traditions shape sociologists' work, are they merely epiphenomena resulting from hard and clear choices that sociologists make, or is there some interaction between the two?

See also: Associations, Scholarly; Discipline Formation in the Social Sciences; Scientific Disciplines, History of; Social Question: Impact on Social Thought; Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations); Social Science Infrastructure: European Union (Social Science Academies and Related Organizations); Social Science Infrastructure: North America (Social Science Academies and Related Organizations); Sociology, History of; Specialization and Recombination of Specialties in the Social Sciences.

Bibliography

Abend, Gabriel, 2006. Styles of sociological thought: sociologies, epistemologies, and the Mexican and U.S. quests for truth. Sociological Theory 24 (1), 1–41.
Alexander, Jeffrey, 1979. Paradigm revision and 'Parsonianism'. Canadian Journal of Sociology 4, 343–357.
Alexander, Jeffrey, Colomy, Paul, 1992. Traditions and competition: preface to a post-positivist approach to knowledge cumulation. In: Ritzer, George (Ed.), Metatheorizing. Sage, Newbury Park, pp. 27–54.
Baehr, Peter, O'Brien, Mike, 1994. Founders, classics and the concept of a canon. Current Sociology 42 (1), 1–151.
Ben-David, Joseph, 1973. The state of sociological theory and the sociological community. Comparative Studies in Society and History 15, 448–472.
Bottomore, Tom, Nisbet, Robert (Eds.), 1979. A History of Sociological Analysis. Heinemann, London.
Burawoy, Michael, et al. (Eds.), 2010. Facing an Unequal World: Challenges for a Global Sociology, 3 vols. ISA.
Camic, Charles, Xie, Yu, 1994. The Statistical Turn in American Social Science: Columbia University 1890–1915. American Sociological Review 59, 773–805.
Cole, Stephen, 1992. Making science: between nature and society. Harvard University Press, Cambridge, Mass.
Coser, L.A., 1965. Men of Ideas: A Sociologist's View. Free Press, New York.
Collins, Randall, 1994. Four Sociological Traditions. OUP, New York.
Connell, R.W., 1997. Why is classical theory classical? American Journal of Sociology 102 (6), 1511–1557.
Crane, D., 1972. Invisible Colleges. University of Chicago Press, Chicago.
Crothers, Charles, 2001. National traditions in the social sciences. In: Smelser, Neil (Ed.), International Encyclopaedia of Social and Behavioral Sciences, vol. 15. Elsevier Science, pp. 10323–10328.
Crothers, Charles, 2005. The diversity and insularity of sociological traditions. In: Calhoun, Craig, Rojek, Chris, Turner, Bryan (Eds.), Handbook of Sociology. Sage, London.
Eisenstadt, S.N., Curelaru, M., 1976. Forms of Sociology: Paradigms and Crises. John Wiley, NY.
Gerard, Delanty. (Ed.), 2006. The Handbook of Contemporary European Social Theory. Routledge, London; NY.
Giddens, A., Turner, J. (Eds.), 1988. Social Theory Today. Polity Press, Oxford.
Goldthorpe, John, 2000. On Sociology. Oxford University Press, Oxford.
Gouldner, A., 1970. The Coming Crisis of Western Sociology. Basic Books, NY.
Gouldner, Alvin, Sprehe, Timothy, 1965. Sociologists look at themselves. Transaction 2 (4), 42–44.
Heilbron, Johan, 1995. The Rise of Social Theory. Polity Press, Cambridge.
Heilbron, J., Guilhot, N., Jeanpierre, L., 2009. Vers une histoire transnationale des sciences sociales. Sociétés Contemporaines 1, 121–145.
Keen, Mike, Janusz, L. Mucha (Eds.), 2003. Sociology in Central and Eastern Europe: Transformation at the Dawn of a New Millennium. Praeger, Westport, CT.
Kuhn, Thomas, 1970/1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.

Lamont, Michele, 1987. How to become a dominant French philosopher: the case of Jacques Derrida. American Journal of Sociology 93, 584–622.
Lemert, Charles (Ed.), 1981. French Sociology: Rupture and Renewal since 1968. Columbia University Press, New York.
Levine, Donald, 1995. Visions of the Sociological Tradition. University of Chicago Press, Chicago.
Menzies, Ken, 1982. Sociological Theory in Use. RKP, London.
Merton, R.K., Riley, Matilda White (Eds.), 1980. Sociological Traditions from Generation to Generation. Ablex, Norwood, NJ.
Mohan, Raj, Wilke, Arthur S. (Eds.), 1994. International Handbook of Contemporary Developments in Sociology. Greenwood Press, Westport, CT.
Moran, Maria-Luz (Ed.). International Sociological Association Pre-Congress Volumes – Social Knowledge: Heritage, Challenges, Perspectives, ISA.
Mullins, N., 1973. Theory and Theory Groups in Contemporary American Sociology. Harper & Row, NY.
Munch, Richard, 1991. American and European social theory: cultural identities and social forms of theory production. Sociological Perspectives 34, 313–335.
Oberschall, Anthony (Ed.), 1972. The Establishment of Empirical Sociology. Harper & Row, NY.
Quah, Stella R., Sales, Arnaud (Eds.), 2000. The International Handbook of Sociology. Sage, London.
Rule, James, 1997. Theory and Progress in Social Science. Cambridge University Press, Cambridge, New York.
Steinmetz, George (Ed.), 2013. Sociology and Empire. Duke University Press.
Turner, S., Turner, J., 1990. The impossible science: an institutional analysis of American sociology. Sage, Newbury Park.
Wagner, Peter, Weiss, C.H., Wittrock, B., Wollman, H. (Eds.). Social Sciences and Modern States: National Experiences and Theoretical Crossroads. Cambridge University Press.
Wagner, P., Wittrock, B., Whitley, R. (Eds.), 1990. Discourses on Society XV. Kluwer Academic Publishers, The Netherlands.
Wells, Richard, Picou, Steven, 1981. American Sociology: Theoretical and Methodological Structures. University Press of America, Washington, DC.
Zuckermann, H., 1988. Sociology of science. In: Smelser, N. (Ed.), Handbook of Sociology. Sage, California.

Sociology and Psychiatry

Dirk Richter, Bern University of Applied Sciences, Bern, Switzerland
Heinz Katschnig, Ludwig Boltzmann Institute for Social Psychiatry, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sociology and psychiatry have a longstanding, albeit intermittently, conflicting relationship. During the mental health care reforms from the 1950s to the 1980s, mainstream sociology has supported psychiatry with both basic and applied research. Currently, the sociological perspective has become more critical. From a sociological point of view, several so-called mental disorders are nothing but medicalized behavioral features (e.g., grief reactions as depression). Further issues addressed are the current countermovement against medicalization ('recovery') as well as the ongoing stigmatization and social exclusion of persons who are diagnosed with a mental disorder.

Introduction

Sociology and psychiatry have a long-standing relationship. More than 100 years ago Emile Durkheim (1897), commonly regarded as one of the founding fathers of modern sociology, has in fact taken suicide as a starting point for developing sociological theory. Within psychiatry, this research is also regarded as the beginning of research on the social context of mental disorders. During the twentieth century and especially in the second half the relation became intensified.

On the one hand, psychiatry has profited from cooperating with sociologists by making use of sociological research methods. Until the last quarter of the twentieth century, sociology was at times even a leading scientific discipline for psychiatry, and at other times, still one of the major background theories. Robert Straus' distinction into 'sociology in medicine' and 'sociology of medicine' (Straus, 1957) could be applied to psychiatry and is named here as 'sociology in psychiatry.' Sociological research has had a deep impact on transforming mental health care systems after World War II all over the Western world. Among other things, the term 'total institution' was taken to develop the deinstitutionalization process during the mental health care reforms.

On the other hand, this collaborative relation has changed recently into a separation. Currently, there are only very few sociological authors who contribute substantially to psychiatry and mental health care. Reasons for this separation can be found on both sides of the separation. One reason seems to be that after many decades of sociopsychiatric research, the mainstream of psychiatry turned to neurobiologically informed disease models during the 1980s and the 1990s. Within psychiatry, the 1990s is known as the 'Decade of the Brain.' Another reason was the apparent lack of practical suggestions from sociology after most of the mental health care reforms were put into place. Many psychiatrists had the impression that the sociological literature on their profession would result in nothing more than criticism on psychiatry in general.

Nevertheless, in many areas of psychiatric research empirical methods of assessment that have their routes in social science research are being used. The authors do not discuss these issues here, since quite a few other articles in the second edition of this encyclopedia deal with such aspects. The authors do not deal therefore with topics such as the epidemiology, i.e., the social distribution of mental disorders, and, potential social antecedents (e.g., social class, gender, social stress), and also not with illness behavior, services research, and the like. Also, they do not discuss what is called social psychiatry, community psychiatry, public mental health, and similar population-related practice aspects of promoting mental health, preventing mental disorders, and treating persons suffering from mental disorders. In addition to the other articles in this edition of the IESBS, several recent handbooks (e.g., Morgan and Bhugra, 2010; Pilgrim and Rogers, 2010; Aneshensel et al., 2013; Richter, 2003) as well as scientific journals (*Social Psychiatry and Psychiatric Epidemiology*; *Epidemiology and Psychiatric Sciences*; *International Journal of Social Psychiatry*, *Society and Mental Health*) cover these and similar topics. However, many of these publications also deal with topics that can be subsumed under the authors' main focus, the 'sociology of psychiatry,' so they are also a source of information for the present article. Finally, since the boundaries between sociology and other social sciences, such as anthropology, history, and economics, are blurred, the authors refer occasionally to these neighboring social sciences.

Before illustrating – in the sections below – sociological thinking on psychiatry and its practice with several examples (power and labeling, medicalization, empowerment versus the sick role, social exclusion and stigma) a few further comments are necessary to better understand sociologists' views on psychiatry.

From a sociological perspective, psychiatry's everyday practice and its epistemology are closely linked. This notion of a strong nexus of practice and epistemology stems from the 1950s and the 1960s. In these decades, psychiatry received strong sociological criticism on both its morally and politically exclusive practices as well as on the more or less hidden ideology of addressing deviant behavior. Deviancy and its moral connotations became the major focus of sociological and sociohistorical research on psychiatry. While stigma research has received much attention in collaborative research between psychiatry and sociology, it has been criticized that psychiatry regards itself as part of the solution but not as part of the problem when complaining about the stigma of mental illness (Sayce, 1998).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27068-2

Sociologists have dealt not only with psychiatric concepts but also with the process of professionalization of psychiatry, which is closely linked both to conceptual issues and to the practice of psychiatry. The rise of the psychiatric profession during the nineteenth century has been regarded alternatively as having been due to their functional value for economic order and efficiency under capitalism (Scull, 1979) and to its role for maintaining the moral, rather than the economic order (Foucault, 1961). Today, the recruitment of doctors into the specialty of psychiatry is declining, often to such a degree that it is discussed whether psychiatrists are an endangered species (Katschnig, 2010), while related profession such as psychologists and psychotherapists are on the rise.

While sociology has thus been ambivalent toward psychiatry – oscillating between cooperation and criticism – a similar ambivalence can be stated for psychiatry's perspective on sociology. On the one hand, the application of sociological research methods to psychiatric problems have undoubtedly contributed to raise psychiatry's societal visibility and importance, by furthering epidemiological and social psychiatric research, which has shown that between one-quarter and up to nearly half of Western populations suffer from at least one diagnosable mental disorder during a person's life time. On the other hand, sociologists are disliked by psychiatrists since they have been and still are among the most radical critics of psychiatry, concerning both its concepts and its practice.

Post–World War II Sociological Theory on Psychiatry: Power and Labeling

While being the work of a philosopher, the French Michel Foucault's analysis on the constitution of mental health care in the seventeenth century and its moral connotations (Foucault, 1961) was very well received within sociology. From his perspective, psychiatry was one of the major societal agencies during the modernization process that led to an increasing disciplining of people who showed 'strange' or unwelcome behavior. At this time, the 'strange' behavior was not yet seen as a medical condition. The purpose of the disciplining strategies was mainly to ensure that people who were regarded as being able to work did their work in the end. In the context of enlightenment, this perspective changed into a medicalized notion of unwelcome behavior. Benevolent physicians were convinced of the treatability of these conditions while the patients were held as inmates in so-called asylums. Thus, the medical subdiscipline of psychiatry evolved as part of the modernization of power, which led to the emergence of the modern state.

This historical view was taken up in sociology after World War II, when psychiatry became the target of humanitarian and politically driven critique, which eventually triggered the psychiatric reforms of the 1960s and the 1970s. Although twentieth-century's psychiatric care was completely different from that of the eighteenth century, major principles appeared to have remained. Sociologists became critics of the impersonal way in which patients were processed. Goffman (1961), in his famous study *Asylums* (the most sold sociological book worldwide), based on participant observation, coined the concept of 'total institution' for the mental hospital and similar organizations that disregard the personal needs of their inmates. According to Goffman, the common feature of 'total institutions' such as mental hospitals, prisons, and even concentration camps was the nonseparation of the social spheres work, life, and leisure under a single authority with very strict rules for the inmates.

Goffman did not only cover the institutional part of the asylum but also followed the then major sociological perspective on mental disease (or 'insanity' at that time). Insanity was not seen as a trait of the inmate but as an ascribed feature that was forced upon a person and that had significant consequences such as admission to a mental hospital. In line with the then prominent theory of symbolic interactionism, the response by the social environment to a certain behavior constituted what was later called a 'label.' Labeling theory (e.g., Scheff, 1966) became a prominent theory, as it was able to demonstrate the power of forced-upon labels. Labeling occurs after an observed and as such classified rule violation. The advocates of 'labeling theory' claimed that the same deviant behavior may either go unnoticed or be labeled as mental illness, depending on the circumstances in which it occurs and is perceived by others. For instance, in a now famous study, Rosenhan (1973) observed that healthy test persons ('pseudopatients') who presented themselves to psychiatric hospitals and told the doctor that they were hearing voices (saying 'empty,' 'hallow,' and 'thud') were actually admitted and received a psychiatric diagnosis.

Once labeled, so the theory goes, the person's identity and social status is significantly altered. Part of the process is that the labeled person internalizes the new identity ascribed to him or her and becomes trapped in the new role of a mental patient. Even if one might assume something like an existing mental disorder, in the end, the label is seen as much more powerful than the 'basic disorder.' The power of the label has got to be acknowledged, however, contrary to the 'labeling' advocates; the forced-upon label will not be accepted freely by the persons in question. Rosenhan's test persons did not accept the label of being mentally ill and they did not become psychiatric patients eventually. Thus, being labeled does not exclusively lead to become (so-called) mentally ill.

Until today, psychiatry is regarded by sociologists as being a part of the power mechanisms that are enforced by the state on its citizens. Compulsory admission and compulsory treatment are intrinsic parts of psychiatric services. Since the beginning of the twentieth century, the legal professions have been increasingly involved in the process of admissions to mental institutions in order to counteract what has been regarded as mismanagement by the medical professions. This has prompted sociologists to analyze the social processes involved. There are basically two sociological positions: those seeing the legal and the medical systems as two competing institutions (Mechanic, 1978) and those who argue that, in practice, they join to become one actor, a 'treatment-control' system that 'takes care' of residual cases viewed as problematic for society, although the written law distinguishes between treatment and control (Dallaire et al., 2000). In the United Kingdom and other countries, a different dimension has been added to this issue with the concept of a 'community treatment order,' which would allow compulsory medication for patients living in the community and would let the pendulum swing back to medical control (Pilgrim and Rogers, 2010).

Recent Sociology: Mental Disorder as a Medicalized Social Construct

In line with recent theoretical developments in humanities (e.g., linguistic turn in philosophy) and with postmodern theories (Richter, 1994), the notion that a mental disorder exists as an objective entity is being contested within sociology until present. The roots of the early sociopsychiatric theories are still present today; however, theoretical reasoning has become far more sophisticated.

Current intrapsychiatry discussions on the introduction of the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) by the American Psychiatric Association have once again triggered research into how mental disorders are being constituted by the psychiatric community. The ever-changing terminologies, concepts, and boundaries of single disease entities (e.g., depression, autism) have made it clear that a biomedical model of mental disorders with clear etiologies and a broadly accepted nosology will neither currently nor in the near future be at hand. Methodologically speaking, recent diagnostic manuals have favored interrater reliability over validity. This is in line with the above-cited studies and theories that questioned that there is a common understanding of what a mental disorder is even within the psychiatric community (Horwitz and Wakefield, 2007).

The obvious arbitrariness of diagnostic entities that are used for categorizing diseases has highlighted once again the questions of where normality ends and where a disease begins. The so-called categorical disease entities suggest more or less clear boundaries that are often hard to see. Many theorists since long have been favoring dimensional models over categorical ones. From a practical medical point of view, however, a categorical nosology has significant advantages (see the debate in Shorter and van Praag, 2010). Medicine as a professional practice and as a research community as a whole deals with disease entities. The categorical disease model has certainly helped to advance psychiatry as an acknowledged medical subdiscipline and thus has boosted its image and its credibility in terms of mental health practice and research. A closer look on specific diseases reveals, however, that apparently well-known diseases or disorders such as depression or schizophrenia do not exist as proper entities.

This state of conceptual and theoretical confusion in terms of psychiatric diagnoses supports the critical stance taken by sociologists. The most prominent sociological theory in this regard is called the 'medicalization' theory. Peter Conrad, its original author, has defined medicalization as a "… process by which nonmedical problems become defined and treated as medical problems, usually in terms of illnesses and disorders" (Conrad, 1992: 209). By being medicalized, a certain behavior and condition is subject to the execution of medicolegal power. A prominent example was homosexuality, which was listed as a psychiatric condition until 1973, but was initially defined in the 1950s as a 'sociopathic personality disturbance.' Until the end of the 1960s there was a strong psychoanalytically inspired tendency to diagnose and treat homosexuals in order to cure them from what was regarded as a "'phobic' avoidance of the other sex caused by inadequate parenting" (Drescher, 2010: 429). The removal of this 'disorder' from DSM-IV was an important step to normalization and destigmatization of same sex relations. In this regard, it can be seen as a sort of demedicalization (Conrad and Slodden, 2013).

More recently defined disorders that are assumed to be medicalized social behavior include the attention-deficit/hyperactivity disorder, the social anxiety disorder, and grief and bereavement disorders. The last of these had triggered a long debate within the mental health treatment community before DSM-V was launched. The debate centered on the question of whether the emotional state of somebody who lost a loved one is a psychiatric condition that needs to be treated. Grief is a normal and expectable reaction to loss. However, its symptoms may not substantially differ from depressive reactions after losing a job or facing a divorce. Thus, current psychiatry is challenged by a certain dilemma, namely, to follow its philosophy of disregarding the etiology of disorders or to be charged with the accusation of more and more medicalizing 'ordinary' problems.

Related to this dilemma is a recent debate that concerns the boundary between 'normal' behavior and conditions and sickness. Allan Horwitz and Jerome Wakefield, two distinguished social scientists, have triggered this debate with several publications. One major book deals with the question of how sorrows that have been regarded as normal in earlier decades (e.g., reactions to losses, grief) are seen as depressive disorders nowadays (Horwitz and Wakefield, 2007). A similar change has occurred with the development of 'normal' fears into anxiety disorders (Horwitz and Wakefield, 2012). From a sociological perspective, these changes can be seen as examples of medicalization.

In contrast to several other genuine sociological approaches, Horwitz and Wakefield have made a case for the incorporation of biological factors into modern sociological thinking on mental health. Their 'harmful dysfunction' approach aims at distinguishing disorders from nondisorders. Evolutionary designed psychological functions that cannot be performed are the basis of a new concept of mental disorders, according to their work.

Countermovements against Medicalization: Empowerment and Recovery versus the Sick Role

Today, those receiving care from professionals start to threaten traditional professional identities and psychiatry in general. They are themselves becoming more powerful in forming self-help and advocacy groups and tend to use terms, such as 'users,' 'clients,' 'consumers,' and even 'survivors of psychiatry,' instead of the term 'patients.' These terms have different implications of how the traditional doctor–patient relationship is redefined. 'Client,' as opposed to 'patient,' implies partnership instead of submissivity and passivity. The term 'consumer' goes even further in claiming that the patient has certain entitlements whose fulfillment can be sued. The concept of 'survivor of psychiatry' denies the right of psychiatry to intervene into the person's problems – it is related to the antipsychiatry movement, which has been set up by former patients and therapists who reject the notion of mental disorders to exist. Pilgrim and Rogers (2010) give a detailed account of sociological thinking about these issues and how the situation has changed since Talcott Parsons (1951) has formulated the paradigm of the 'sick role.' According to Parsons, the sick role allows people to be temporarily exempted from social responsibilities, but, in exchange, they must accept that they cannot get well on their own, and that they have to look for professional help.

Another important development, which can, at least to some extent, counteract the medicalization of mental disorders is the Recovery and Empowerment movement (Amering and Schmolke, 2009; Slade, 2009). This interesting development might be able to reconcile, at least to some extent, mainstream mental health treatment with its critics. Under the labels of 'Empowerment' and 'Recovery' the experiences of persons with long-term mental health problems are included into the planning, delivery, and quality control of mental health services. Both concepts aim at enhancing the self-management of people who suffer from mental health problems. Recovery, additionally, supports users in developing an identity that goes far beyond the patient role. Finally, it aims at leading a meaningful life despite having had or still having mental health problems or disabilities. An important component of Recovery is providing support by peer users who are increasingly employed by psychiatric services.

Interestingly, the role of sociologists in supporting or evaluating this important shift is rather marginal. Only few sociological contributions on this topic have been published so far (Watson, 2012). This is astonishing because central sociological topics are addressed in these concepts. The empowerment of persons with mental health problems, and thus, the weakening to a certain extent of the role of 'traditional' professional, is one of the long-standing goals of sociological work in this field. A related issue to Recovery aims at addressing the social exclusion (see below) and enhancing the social inclusion of persons with mental health problems into society. However, only few sociologists nowadays conduct research into the entire field of mental health and into social exclusion problems of persons with mental health problems in particular.

The countermovement against medicalization and therapeutically induced despair could be a very interesting phenomenon from a sociological perspective. As sociologists have shown (Yanos et al., 2007), the marginalized group of socially excluded persons with mental health problems has fought a social fight against strong powers within the therapeutic and scientific communities. More and more, professions from the mental health treatment field, e.g., psychiatric nurses, regard the Recovery movement as a very helpful aid in achieving better treatment results. This, however, becomes a contested issue within the user movement as it is suspected that psychiatry uses patients to convince skeptical persons over to be treated.

Social Exclusion and Stigma

Persons with mental health problems are socially excluded from nearly every domain of the social world (Wright and Stickley, 2013). In economic terms, they are – on average – poorer than the healthy population and they suffer from a higher rate of unemployment. During the current economic recession, the unemployment risk for persons with mental health problems has risen further (Evans-Lacko et al., 2013): they are less often engaged in intimate partnerships, they are often deprived of legal and political rights, they are less often politically engaged, and they suffer from more physical health problems than their mentally healthy fellow citizens (Eikelmann et al., 2005). Another issue is the high rate of being victimized by aggressive acts.

This modern kind of social exclusion has taken over the longtime social exclusion of the 'insane' in asylums. By initiating the mental health reforms of the postwar decades, social and community psychiatry aimed at preventing any kind of social exclusion from which the persons concerned definitely suffer. Following the deinstitutionalization of mental health care, the former 'inmates' were supposed to live in the community (mainly in sheltered housing facilities) and to work in sheltered workplaces. It was assumed that many of those persons could be trained to lead an ordinary life and to be released from professional support.

However, from the current perspective, it becomes clear that the mental health care reforms did not succeed in terms of social inclusion. From a sociological point of view, it is clear that the initial ideas of social inclusion did not fit the conditions of modern societies. As the postwar societies in the Western world became more individualized and differentiated and as the social change happened more quickly than expected, many persons with mental health problems could not keep up with the challenges of modern life. Unfortunately, the postwar treatment system with its supporting and sheltering institutions had contributed unintentionally to the social exclusion by creating living conditions that were often not compatible with the conditions of the 'real' world.

It is currently debated whether – in the light of societal and political conditions that apparently are not supportive – individual support for reinclusion is an effective strategy (Clifton et al., 2013). To give an example, from a research perspective, supported employment is an effective strategy that aims at supplying persons concerned with a real-world job and with a job coach who is able to provide backup and encouragement when facing strains and problems. This strategy, however, cannot be successful when the general unemployment rate is high and persons with mental health problems are being discriminated against.

In addition to social exclusion, the social stigma of mental disorders, mentally ill persons, and mental health treatment facilities is persisting. Stigma research with regard to disability and illness is one of the most commendable contributions of sociology to health care in general and to psychiatry in particular (Goffman, 1963; Link and Phelan, 2013). Stigma research and the resulting antistigma initiatives have gained a lot of attention over the recent decades. As can be easily concluded, the stigma on mental disorders and on the persons concerned has a close connection to the social construction of mental health and illness. Bruce Link and Jo Phelan have, therefore, suggested to analyze stigmata by using a 'modified labeling' approach. The authors suggest that "… stigma exists when elements of labeling, stereotyping, separation, status loss, and discrimination occur in a power situation that allows them to unfold" (Link and Phelan, 2001: 367).

Being stigmatized can lead to severe consequences for the affected persons. Persons with mental health problems report about status loss and experiences of being discriminated against openly (Link and Phelan, 2013). Additionally, the social network tends to decrease in many cases after the diagnosis has become known to friends and acquaintances. From a therapeutic point of view, stigma management becomes an important issue. Recovery care includes dealing with questions such as to whom to disclose the 'illness,' when to talk about

mental health problems while setting up new relations, and also to discuss legal questions as if employers have to be informed about being treated for a mental disorder.

Bernice Pescosolido has recently summarized the relevant research activities as well as the knowledge from the US perspective (Pescosolido, 2013). Similar to experiences in Germany (Angermeyer et al., 2013), the attitudes of the general public have not changed much during the recent decades. However, the American public has become more sophisticated and probably more open when it comes to recognizing and disclosing mental health problems. This social change in relation to mental disorders is regarded as one factor that has led to steadily rising rates of mental health treatment utilization (Richter and Berger, 2013; Richter et al., 2008).

Conclusions

The relationship between sociology and psychiatry can be characterized as oscillating between cooperation and criticism of each. After a long time of close collaboration between sociology and psychiatry in the second half of the twentieth century (which was accompanied by the antipsychiatry movement), the current situation can be outlined not only as one of increasing interest in one another but also as being suspicious about each other's contributions. Given alone the persisting problem of the arbitrariness of the classification of mental disorders, sociological criticism of psychiatry does not seem to disappear. However, today's mental health care practice is increasingly including social aspects and moving toward psychosocial interventions once again, after the 'decade of the brain' has not succeeded in providing effective and efficient treatments for most mental health disorders. Psychiatrists might consider inviting sociologists to carry out more studies on these modern practices. It remains to be seen whether these new developments will help to bring sociology and psychiatry closer again.

See also: Chronic Illness, Psychosocial Coping with; Deinstitutionalization, Psychiatry of; Disability: Social and Psychological Aspects; Environmental Stress and Health; Ethics and Psychiatry; Medical Profession; Medical Sociology; Mental Disorders, Epidemiology of; Mental Health and Social Work; Mental Health, Mentalism and Sanism; Mental Health; Mental Illness, Etiology of; Mentally Ill, Public Attitudes towards; Psychiatric Social Work in the United Kingdom; Psychoanalysis in Sociology; Public Psychiatry and Public Health Mental Systems; Social Epidemiology; Suicide, Sociology of.

Bibliography

Amering, M., Schmolke, M., 2009. Recovery in Mental Health: Reshaping Scientific and Clinical Responsibilities. Wiley, Chichester.

Aneshensel, C.C., Phelan, J.C., Bierman, A. (Eds.), 2013. Handbook of the Sociology of Mental Health. Springer, New York.

Angermeyer, M.C., Matschinger, H., Schomerus, G., 2013. Attitudes towards psychiatric treatment and people with mental illness: changes over two decades. British Journal of Psychiatry 203, 146–151.

Clifton, A., Repper, J., Banks, D., Remnant, J., 2013. Co-producing social inclusion: the structure/agency conundrum. Journal of Psychiatric Mental Health Nursing 20, 514–524.

Conrad, P., 1992. Medicalisation and social control. Annual Review of Sociology 18, 209–232.

Conrad, P., Slodden, C., 2013. The medicalization of mental disorder. In: Aneshensel, C.C., Phelan, J.C., Bierman, A. (Eds.), Handbook of the Sociology of Mental Health, second ed. Springer, New York, pp. 61–74.

Dallaire, D., Mccubbin, M., Morin, P., Cohen, D., 2000. Civil commitment due to mental illness and dangerousness: the union of law and psychiatry within a treatment-control system. Sociology of Health and Illness 22, 679–699.

Drescher, J., 2010. Queer diagnoses: parallels and contrasts in the history of homosexuality, gender variance, and the diagnostic and statistical manual. Archives of Sexual Behavior 39, 427–460.

Durkheim, É., 1897. Le Suicide : Étude de sociologie. Felix Alcan, Paris.

Eikelmann, B., Reker, T., Richter, D., 2005. Social exclusion of the mentally ill; a critical review and outlook of community psychiatry at the beginning of the 21st century (in German). Fortschritte der Neurologie – Psychiatrie 73, 664–673.

Evans-Lacko, S., Knapp, M., Mccrone, P., Thornicroft, G., Mojtabai, R., 2013. The mental health consequences of the recession: economic hardship and employment of people with mental health problems in 27 European countries. PLoS One 8, e69792.

Foucault, M., 1961. Histoire de la folie. Librairie Plon, Paris.

Goffman, E., 1961. Asylums: Essays on the Social Situation of Mental Patients and Other Inmates. Doubleday, Garden City, NY.

Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Prentice-Hall, Englewoold Cliffs.

Horwitz, A.V., Wakefield, J.C., 2007. The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder. Oxford University Press, Oxford.

Horwitz, A.V., Wakefield, J.C., 2012. All We Have to Fear: Psychiatry's Transformation of Natural Anxieties into Mental Disorders. Oxford University Press, Oxford.

Katschnig, H., 2010. Are psychiatrists an endangered species? Observations on internal and external challenges to the profession. World Psychiatry 9, 21–28.

Link, B.G., Phelan, J.C., 2001. Conceptualizing stigma. Annual Review of Sociology 27, 363–385.

Link, B.G., Phelan, J.C., 2013. Labeling and stigma. In: Aneshensel, C.C., Phelan, J.C., Bierman, A. (Eds.), Handbook of the Sociology of Mental Health, second ed. Springer, New York, pp. 481–494.

Mechanic, D., 1978. Medical Sociology. Free Press, New York.

Morgan, C., Bhugra, D. (Eds.), 2010. Principles of Social Psychiatry. Wiley-Blackwell, Chichester.

Parsons, T., 1951. The Social System. Free Press, Glencoe, IL.

Pescosolido, B.A., 2013. The public stigma of mental illness: what do we think; what do we know; what can we prove? Journal of Health and Social Behavior 54, 1–21.

Pilgrim, D., Rogers, A., 2010. A Sociology of Mental Health and Illness. Open University Press, Buckingham.

Richter, D., 1994. Existentialism and postmodernism. Continuities, breaks, and some consequences for medical theory. Theoretical Medicine 15, 253–265.

Richter, D., 2003. Mental System and Social Context: Sociology of Mental Disorders in the Era of Biosciences (in German). Psychiatrie-Verlag, Bonn.

Richter, D., Berger, K., 2013. Are mental disorders increasing? Update of a systematic review on repeated cross-sectional studies (in German). Psychiatrische Praxis 40, 176–182.

Richter, D., Berger, K., Reker, T., 2008. Are mental disorders on the increase? A systematic review (in German). Psychiatrische Praxis 35, 321–330.

Rosenhan, D.L., 1973. On being sane in insane places. Science 179, 250–258.

Sayce, L., 1998. Stigma, discrimination and social exclusion: what's in a word? Journal of Mental Health 9, 331–348.

Scheff, T.L., 1966. Being Mentally Ill: A Sociological Theory. Aldine, Chicago.

Scull, A.T., 1979. Museums of Madness: The Social Organization of Insanity in Nineteenth Century England. St. Martin's Press, New York.

Shorter, E., van Praag, H.M., 2010. Disease versus dimension in diagnosis. Canadian Journal of Psychiatry 55, 59–64.

Slade, M., 2009. Personal Recovery and Mental Illness: A Guide for Mental Health Professionals. Cambridge University Press, Cambridge.

Straus, R., 1957. The nature and status of medical sociology. American Sociological Review 22, 200–204.

Watson, D.P., 2012. The evolving understanding of recovery: what does the sociology of mental health have to offer? Humanity and Society 36, 290–308.

Wright, N., Stickley, T., 2013. Concepts of social inclusion, exclusion and mental health: a review of the international literature. Journal of Psychiatric Mental Health Nursing 20, 71–81.

Yanos, P.T., Knight, E.L., Roe, D., 2007. Recognizing a role for structure and agency: integrating sociological perspectives into the study of recovery from severe mental illness. In: Avison, W.R., Mcleod, J.D., Pescosolido, B. (Eds.), Mental Health, Social Mirror. Springer, New York, pp. 407–433.

Sociology and the New Institutionalism

Victor Nee, Cornell University, Ithaca, NY, USA
Sonja Opper, School of Economics and Management, Lund University, Lund, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In sociology, new institutionalists led the revival in interest in institutions in organizational theory and economic sociology by shifting the focus of causal reasoning from agent-centric studies of economic and organizational actors to the relationship connecting the firm with its institutional environment. We suggest a multilevel causal model incorporating the connection between the subinstitutional domain of social action and concrete social relationships, and the meso- and macroinstitutional environment of customs, conventions, law, organizations, ideology, and the state as the key elements explaining the rise and demise of institutions. In this model, norms bridge the microworld of individual actors and social groups, and the broader institutional environment.

The new institutionalism and the focus on institutions for its foundational concept have gained ascendancy in economics, political science, and sociology. The claim that institutions matter is now a widely held assumption, which addresses a cross-disciplinary set of problems germane to organizational and economic performance. The interest in institutional analysis has given rise to a variety of new institutionalist approaches. Not since the behavioral revolution of the 1950s has there been a foundational concept that has generated so much interest, exchange, and debate in the social sciences. The variety of research programs inspired by the new institutionalism has produced robust findings that have cumulatively answered questions of basic interest to the social sciences. Although the new institutionalisms incorporate a broad variety of approaches with distinctive emphases, the common interest in institutional analysis involves a shared behavioral assumption of bounded rationality (Simon, 1957). The Coasian revival of economic interest in institutions through transaction cost reasoning provided a method that enables economics to 'study the world that exists,' which new institutional economists extended to a series of seminal work that explained a broad range of economic and organizational outcomes. In sociology, new institutionalists led the revival in interest in institutions in organizational theory and economic sociology by shifting the focus of causal reasoning from agent-centric studies of economic and organizational actors to the relationship connecting the firm with its institutional environment.

A broad consensus across the social sciences agrees that norms are mainly effective in the governance of exchange in close-knit communities of economic actors. These settings typically involve tribal and peasant societies, or in advanced industrial societies, highly specialized markets with a limited number of market players engaged in repeated exchange. Detailed case studies highlight the variability of motivations and mechanisms of private ordering applied in close-knit networks. Diamond traders, for example, systematically reject state-crafted law as insufficient for the specific business risks inherent in the diamond industry. Traders instead rely on a sophisticated extralegal contractual regime including arbitration proceedings enforced by trading clubs designed to disseminate information about reputation among members (Bernstein, 1992). A study of conflict resolution among cattle farmers in northern California sheds light on less-standardized procedures. Despite their differences, a common identity as ranchers of Shasta County sustains a live-and-let-live philosophy, which enables parties to practice mutual restraint. As long as accounts balanced along multiple dimensions of interpersonal relations, parties in disputes settled informally. In sum, private orders rely on norms to secure contracts through repeat transactions and community sanctions in close-knit communities and networks (Ellickson, 1991).

In their study of the rise of capitalism in China, Nee and Opper (2012) examine the broad-based robust entrepreneurial action supported by endogenous economic institutions arising from interacting individuals. They show that mutual dependence – on information flows, on access to scarce resources, on interfirm collaboration in technology development, and on privately organized supply networks autonomous from government-controlled sources – can provide the glue that endogenously reinforces long-term stability of contractual relationships, without reliance on third-party enforcement. Social norms can enable close-knit communities to successfully avoid the 'tragedy of the commons' problem in managing communal resources (Ostrom, 1990). Such governance structures rely on long-standing social relationships and community sanctions rather than on external authority to solve the collective action problem threatening the depletion of communal resources.

The Origin of Norms

Most norms evolve gradually through trial and error, with the behavior that brings about success reinforced. Benefits from norm-compliance, however, are not limited to tangible gains such as commodities. If the total gains from cooperation exceed the costs, private orders can provide a stable framework for economic actors to sustain repeat transactions. This is consistent with the causal sequence in the rise of political and economic institutions in the West. Norms often precede laws, and only become formalized, Robert Axelrod observes, once a "norm becomes firmer" and once "there is growing support to formalize

it through the promulgation of laws" (Axelrod, 1984). Civil liberties, he continues, the very foundation of democratic systems, exemplify this: "The legal system can only protect free speech if there is substantial support for it among a population willing to tolerate dissent and willing to protect those who exercise it" (p. 1106). Conformity between legal rules and social norms is therefore often the result of this particular sequence where laws simply support, maintain, and extend what is regarded as a just and appropriate social norm. As with political institutions, economic law often builds on established and widely accepted business norms. In Amsterdam, home of the world's first stock market and stock company (the East India Company, founded in 1602), formal shareholder rights were largely absent, and many of the financial transactions were even prohibited by law. Securities trading relied on self-interest and self-enforcement. It was not until 1851 that the Amsterdam Stock Exchange Association was founded to organize and regulate share trading. Also the London Stock Exchange did not receive its first codified rule book until 1812, while organized securities trading had started as early as 1698. The Brussels stock exchange worked for more than 100 years up to 1935 with minimal government regulation. Worldwide, the inception of securities markets was not usually preceded or even accompanied by formal state-mandated rules protecting shareholder rights. Company stocks were traded informally and transactions were treated as gentlemen's agreements, often conducted in local coffee houses or open marketplaces. Hence, new institutionalist economic sociology agrees with economic new institutionalists that good formal institutions have typically followed, rather than preceded, economic development (Glaeser et al., 2004).

In the new institutionalist economic sociology, institutions are self-reproducing social structures that provide a conduit for collective action by enabling, motivating, and guiding the interests of actors and enforcing principal agent relationships. Institutions comprise a system of interrelated informal and formal elements – customs, conventions, norms, beliefs, and rules – governing social relationships within which actors pursue and fix the limits of legitimate interests.

Norms and Formal Rules

Economic sociologists have long held the view of markets as self-reproducing social structures within which buyers and sellers transact across market interfaces. "However difficult it often is to account for all their detailed characteristics," market institutions "are at least the product of the very things economics disregards – the relatively permanent relationships between individuals or between groups, which form social structures" (Homans [1961]1974: p. 68). Economists concur that repeated exchange is commonplace even in markets characterized by the ease and reliability of impersonal exchange. In addressing how to characterize the market not as an abstraction but as a self-reproducing social structure, White proposed "embedding economists' neoclassical theory of the firm within a sociological view of the market" by integrating signaling theory into a sociological model of markets (White, 1981: p. 518; for signaling theory refer to Akerlof (1976) and Spence (1974)). In White's model, firms are like organisms seeking out niches in an ecology. Markets develop as endogenous institutional constructs as firms watch signals emitted by their competitors and position themselves to compete in a niche, "compatibly arrayed on the qualities which consumers see in them" (White, 1981: p. 519). The challenge to demystify markets calls for examining the relationship between norms embedded in informal social processes and the formal rules mandated and enforced by the state.

A multilevel causal model (see **Figure 1**) portrays the connection between the subinstitutional domain of social action and concrete social relationships, and the meso- and macroinstitutional environment of customs, conventions, law, organizations, ideology, and the state. In this model, norms bridge the microworld of individual actors and social groups, and the broader institutional environment (Nee, 1998). We assume that actors are rational in that they make decisions according to cost–benefit criteria. However, we do not see humans as the narrowly self-interested *homo oeconomicus* of neoclassical economics, possessing perfect information and unbounded cognitive capacity. Instead they use rule of thumb heuristics based on the information currently available to them (Marsh, 2002). The arrows pointing downward from the institutional environment indicate constraints placed on social action. The arrows pointing upward indicate that the macro environment is constituted and modified by social action. Economic actors can through collective action trigger changes in the institutional environment. Not only do informal norms gain in importance if many actors find it rational to decouple from existing formal rules, but once a critical mass is reached, the collective action of economic actors imposes pressure on

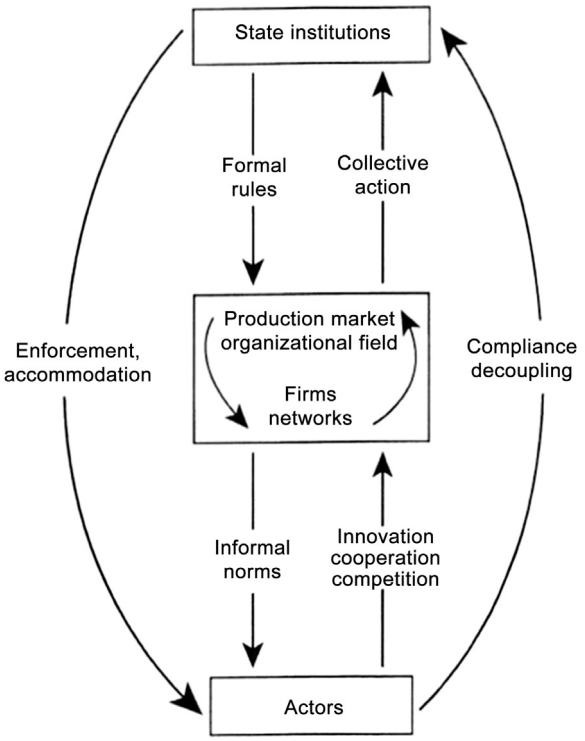

Figure 1 A multilevel causal model of institutions. Source: Nee, Victor, Opper, Sonja, 2012. Capitalism from below: Markets and Institutional Change in China. Harvard University Press, Cambridge, MA.

the state to respond by enforcing the existing legal and regulatory structures. Once this tipping point is reached, the state no longer can effectively enforce compliance in the face of a self-reinforcing collective action. This process of collective action can but need not involve formal types of lobbying.

The institutional mechanisms are distal, as opposed to the proximate mechanisms at the micro- and mesolevels of individuals and networks. Institutional mechanisms encompass the deeper causes because they shape the incentive structure for organizations and individuals, and thereby the contexts in which proximate mechanisms operate. The institutional-level mechanisms posited by economists and sociologists, despite differences in behavioral assumptions and conceptual language, are not as far apart as is commonly perceived (Nee, 2005). Market mechanisms schematically represented in the downward arrow from the institutional environment to organizational fields includes those embedded in labor markets, capital markets, raw material markets, and so on. The formal rules that make up the institutional environment and norms embedded in ongoing social relations, then jointly interact to shape economic behavior in the market place.

Actors are motivated by interests and preferences, often formed and sustained within close-knit groups. Groups of firms in turn are commonly arrayed in a variety of organizational fields, defined as a matrix of organizations that, "in the aggregate, constitute a recognized area of institutional life; key suppliers, resource and product consumers, regulatory agencies, and related organizations that produce similar services or products" (DiMaggio and Powell, 1983: p. 148). The production market is thus embedded in its organizational field as the core firms in a distinct niche. In the production market, firms compete and maneuver for advantage and status with peer firms in the same niche (White, 1981). Given uncertainties of upstream and downstream market conditions, they are guided by the signals they read from the operations of their peers. In competitive markets, pressures on firms stemming from Darwinian selection processes necessitate an interest-related logic of strategic action.

Within organizational fields, processes of conformity – *isomorphism* – motivate and guide organizations, endogenously giving rise to increasing within-field homogeneity. Paul DiMaggio and Walter Powell specify three mechanisms – coercive, normative, and mimetic – promoting isomorphism in organizational fields (DiMaggio and Powell, 1983). Coercive isomorphism integrates resource dependence theory into organizational analysis and assumes that organizations respond to societal claims and preferences; normative isomorphism specifies how professional associations influence organizational behavior under conditions of uncertainty, and mimetic isomorphism is about how facing uncertainty under high stakes, intendedly rational actors satisfice by looking for successful competitors and making plausible but incorrect guesses about the reasons for their success. Organizations tend to mimic other similar organizations that they perceive to be successful. Legitimacy-seeking is driven mainly by the firm's interest in its survival and profitability. Legitimacy can be viewed as a condition of fitness, which enables firms to enhance their survival chances and secure advantages in economic and political markets. For manufacturers especially, mimicking entails learning-by-doing in efforts to internalize the technological and organizational basis of success of leading competitors. As Kenneth Arrow showed, self-perfection and minor innovations gained by learning from experience in ordinary production drives endogenous economic growth (Arrow, 1962).

The social mechanism guiding the action of firms in an organizational field is not dissimilar from those influencing strategic action of individuals in close-knit groups. The process through which conformity is achieved has coercive, normative, and mimetic aspects (Homans [1961]1974; Emerson, 1962; Blau, 1964). Organizational actors are motivated by interests and preferences, often formed and sustained within the organizational field. Rationality is context-bound and embedded in interpersonal ties. Only if individual interests and preferences are well aligned with the incentives structured in the institutional environment, will they reinforce compliance with formal rules through self-monitoring and enforcement. Otherwise, if individual interests are not aligned with the structure of opportunity legitimized by the state, strategic interest may give rise to decoupling from institutionalized norms (Merton, 1968), which may in turn lead to the formation of opposition norms once a sufficient number of actors arrayed in organizational fields decide to decouple from the formal framework (as discussed below).

In this multilevel causal model, institutional change rests on four causal mechanisms, linking the subinstitutional domain of social action with the macroinstitutional environment as follows: First the market mechanism provides incentives and opportunities for economic actors to come up with bottom-up institutional arrangements to enable the growth of the private enterprise economy. Second, through random *mutations* institutional innovations emerge and through a process of trial and error, successful forms diffuse through the regional economy. Third, through *mimicking* of successful business ventures, swarms of followers pile into the niche leading to spatial geographical concentration. Fourth, *mutual monitoring and enforcement* in crosscutting networks of economic actors reinforce social norms.

Norms and Networks

What explains motivation for trustworthiness and abstention from opportunism in ongoing social relationships? Why is trustworthiness found more commonly in ongoing social relationships than in transactions between strangers? The answer is found in specifying the mechanisms intrinsic to social relationships that develop and maintain cooperative behavior, enabling actors to engage in collective action to achieve group ends. These mechanisms are rewards and punishment in social exchange and their use in upholding norms – shared beliefs and expectations. Numerous studies confirm the efficacy of social rewards and punishment in motivating trustworthy behavior and abstention from opportunism with respect to the norms of the group (Whyte, 1943; Barnard, 1968; Abolafia, 1996; Bewley, 1999). Informal norms arise in the course of social interaction as standards of expected behavior and are maintained when reward is expected to follow conformity and punishment, deviance. Members of a group routinely reward conformity to norms by conferring

social approval and status. Conversely, members punish failure to conform to norms through their social disapproval and, ultimately, through ostracism. Hence, the monitoring of norms is a spontaneous by-product of social interactions. In more complex exchanges, the same processes hold, but the pressure to conform also takes on a collective action dimension. Frequency of interaction, a characteristic feature of close-knit groups, lowers the cost of monitoring members, assuming they are in close enough contact with one another that information about members' conduct is common knowledge (Hardin, 1982; Macy, 1991; Gould, 1993). Reward and punishment in repeated exchanges – when actors take into account the weight of the future, as in ongoing relationships – motivates cooperative behavior. Community sanctions complement bilateral responses, if transactions are infrequent. In sum, trustworthiness and reliability as forms of cooperative behavior arise from rational action responding to social rewards and punishment in ongoing relationships.

In his detailed account of the interactions in a work group made up of a supervisor, 16 agents, and 1 clerk, Peter Blau provides an illustration of how self-interested action of individuals endogenously produces a pecking order through the exchange. Agents consulted fellow agents about the appropriate legal rules that applied to their case, rather than bring their questions to the attention of the supervisor who evaluated their work. The importance of social rewards and sanctions (e.g., status) in the normative regulation of a social structure are highlighted here. Routine social exchanges, like repeated transactions and community sanctions, comprise glue for the bottom-up institutional arrangements that endogenously arises to sustain long-term economic relationships. The behavior of economic actors frequently bears little resemblance to the legitimate courses of action stipulated by state-mandated formal rules. Instead, social groups based on personal connections serve to organize market-oriented economic behavior according to norms reflecting the private expectations and interests of individuals. They act often at odds with the goals formulated by politicians.

Norms operating in the shadows of state-mandated rules can both limit and facilitate economic action. On the one hand, a decoupling of norms from these rules can give rise to inefficient allocation of resources when individual actors collude to secure resources from government for their group, resulting in structural rigidities and economic stagnation (Olson, 1982). In Russia, mafia-like business networks operated to obstruct Boris Yeltsin's efforts at building a market economy. On the other hand, a decoupling of norms can also facilitate economic action and promote growth by providing a framework for trust and collective action. In China, informal privatization and local institutional arrangements have contributed to two decades of economic growth during the early stages of economic reform.

The Quest to Understand Institutional Change

What accounts for the difference in results? Given the variance of possible interactions between formal rules and social norms, it is a central task for theory development to better specify the nature of the relationship. Under what conditions will norms evolve into self-reinforcing opposition norms, which then undermine the effectiveness of formal rule? By opposition norms we refer to beliefs that enable, motivate, and guide collective action decoupled from and not legitimated by state-mandated rules. Opposition norms are commonplace in all societies and span a wide range of activities, be they those of alienated minority youths in America and Europe or the untaxed economic transactions between buyers and sellers in the informal economy of established market economies. Especially in the early stage of norm emergence, opposition norms often involve 'subterranean' activity below the radar of the state, and thus are not detected by law enforcers. (For an analysis of opposition norms in response to external sanctions, see Heckathorn (1990).) They can remain below the radar indefinitely, but at an inflection point, opposition norms can burst into the public arena to enable, motivate, and guide self-reinforcing collective action compelling politicians to respond. Political elites face a range of options in their response, from strengthening law enforcement and use of coercive force to adaptive institutional change.

To understand the link between opposition norms and endogenous institutional change, it is crucial to explore the role of private order allowing members of social subgroups to improve their welfare position, although the state is unwilling or unable to protect their economic transactions. Clearly, to the extent that social groups have interests and preferences independent of what politicians or organizational leaders want, the respective contents of social norms and organizational rules will emerge to 'bend the bars of the iron cage' imposed by formal rules. Norms are ideas that arise from the problem-solving activities of human beings in their strivings to improve their chances for success (the attainment of rewards) through cooperation. Individuals jointly produce and uphold norms to capture the gains of cooperation (Nee and Ingram, 1998). (Ellickson (1991: p. 167) argues in a similar vein, "Members of a close-knit group develop and maintain norms whose content serves to maximize the aggregate welfare that members obtain in their workaday affairs with one another.") Such norms arise through trial and error and are adopted by members of a group when they result in success. Hence, norm emergence involves a collective learning process in which members learn from experience gained through different strategies and gradually identify the one which seems to have an edge over others. (Experimental research confirms that subjects prefer to learn to cooperate than not cooperate in repeated public good games. See Isaac et al. (1994).) In this sense, established norms are focal points, which evolve through social learning processes (Young, 1996).

Whether decoupled norms operate only under distinct circumstances or more widely, developing into self-reinforcing opposition norms leading to endogenous institutional change, depends on whether a critical mass of societal participation is reached (Young, 1996, 2001; Nee and Opper, 2012). For instance, in all command economies it was commonplace that individuals and firms sought to alleviate shortages of the formal economy by resorting to black market activities and barter trading (Kornai, 1990). While this practice was against the formal rules, it never gave rise to endogenous institutional change. Although company managers hoarded goods and repair parts to prepare for unexpected shortages, these activities

remained limited in scope. In retrospect, the scattered occurrence of black markets and barter trading may have stabilized the system, as companies were better able to respond to temporary shortages. Similarly, black markets may have helped to contain social discontent.

Conclusion

In sum, the new institutionalism is a cross-disciplinary paradigm involving sociology, economics, and political science contributing to the understanding of the workings of institutions. Economic sociologists emphasize the role of informal institutional elements of norms and networks in enabling, motivating, and guiding economic action. In the sociological approach, the emphasis is on bottom-up sources of creativity and innovation in the emergence of economic institutions facilitating economic activity and organizational performance. In an effort to broaden understanding of phenomena of institutional innovation and change, economic, and organizational scholars employ a multilevel causal model where causal forces operate in both directions from institutional mechanisms embedded in macro structures to microlevel behavior, and from micro motives and behavior to the macro level. In the new institutionalist approach, the task of cumulative empirical effort is to identify both commonalities but also distinctive conditions shaping and facilitating bottom-up processes of institutional change.

See also: Collective Behavior, Social Psychology of; Cooperation: Sociological Aspects; Exchange: Social; Ideologies, Institutions, and the New Institutionalism; Institutionalism and Neo-institutionalism: History of the Concepts; Institutionalism; Institutions; Law and Economics; New Institutionalism in the Analysis of Complex Organizations; Norms; Old and New Institutionalism in Economics; Social Networks; Transaction Costs and Property Rights.

Bibliography

Abolafia, Mitchel, 1996. Marking Markets: Opportunism and Restraint on Wall Street. Harvard University Press, Cambridge, MA.

Akerlof, George, 1976. The economics of caste and of the rat race and other woeful tales. The Quarterly Journal of Economics 90, 599–617.

Arrow, Kenneth J., 1962. The economic implications of learning by doing. Review of Economic Studies 29, 155–173.

Axelrod, Robert, 1984. The Evolution of Cooperation. Basic Books, New York.

Barnard, Chester I., 1968. The Functions of the Executive. Harvard University Press, Cambridge, MA.

Bernstein, Lisa, 1992. Opting out the legal system: extralegal contractual relations in the diamond industry. Journal of Legal Studies 21, 115–157.

Bewley, Truman F., 1999. Why Wages Don't Fall during a Recession. Harvard University Press, Cambridge, MA.

Blau, Peter, 1964. Exchange and Power in Social Life. John Wiley and Sons, New York.

DiMaggio, Paul, Powell, Walter W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Journal of Sociology 48, 147–160.

Ellickson, Robert, 1991. Order without Law. Harvard University Press, Cambridge, MA.

Emerson, Richard, 1962. Power-dependence relations. American Sociological Review 22, 31–41.

Glaeser, Edward L., La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, 2004. Do institutions cause growth? Journal of Economic Growth 9, 271–303.

Gould, Roger V., 1993. Collective action and network structure. American Sociological Review 58, 182–196.

Hardin, Russell, 1982. Collective Action. John Hopkins University Press, Baltimore.

Heckathorn, Douglas D., 1990. Collective sanctions and compliance norms: a formal theory of group-mediated social control. American Sociological Review 55, 366–384.

Homans, George C., [1961]1974. Social Behavior: Its Elementary Forms. Harcourt Brace Jovanovich, New York.

Isaac, R. Mark, Walker, James, Williams, Arlington W., 1994. Group size and the voluntary provision of public goods: experimental evidence utilizing large groups. Journal of Public Economics 54, 1–36.

Kornai, Janos, 1990. The affinity between ownership forms and coordination mechanisms: the common experience of reform in socialist countries. The Journal of Economic Perspectives 4, 131–147.

Macy, Michael W., 1991. Chains of cooperation: threshold effects in collective action. American Sociological Review 56, 730–747.

Marsh, Barnaby, 2002. Heuristics as social tools. New Ideas in Psychology 20, 49–57.

Merton, Robert K., 1968. Social Theory and Social Structure. Free Press, New York.

Nee, Victor, 1998. Norms and networks in economic and organizational performance. The American Economic Review. Papers and Proceedings 88, 85–89.

Nee, Victor, 2005. The new institutionalisms in economics and sociology. In: Smelser, Neil J., Swedberg, Richard (Eds.), The Handbook of Economic Sociology. Russell Sage Foundation and Princeton University Press, New York and Princeton, pp. 49–74.

Nee, Victor, Ingram, Paul, 1998. Embeddedness and beyond: institutions, exchange and social structure. In: Brinton, Mary, Nee, Victor (Eds.), The New Institutionalism in Sociology. Russell Sage Foundation, New York, pp. 19–45.

Nee, Victor, Opper, Sonja, 2012. Capitalism from below: Markets and Institutional Change in China. Harvard University Press, Cambridge, MA.

Olson, Mancur, 1982. The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities. Yale University Press, New Haven, CT.

Ostrom, Elinor, 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.

Simon, Herbert, 1957. Models of Man: Social and Rational. John Wiley & Sons, New York.

Spence, Michael A., 1974. Market Signaling: Informational Transfer in Hiring and Related Screening Processes. Harvard University Press, Cambridge, MA.

White, Harrison C., 1981. Where do markets come from? American Journal of Sociology 87, 518.

Whyte, William F., 1943. Street Corner Society. University of Chicago, Chicago.

Young, H. Peyton, 2001. Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, Princeton.

Young, H. Peyton, 1996. The economics of convention. Journal of Economic Perspectives 10, 105–122.

Sociology, Epistemology of

Mario Bunge, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Three aspects of the epistemology (or philosophy) of sociology are examined: its ontology, epistemology in the narrow sense, and methodology. The matter of the nature of the social has two foci: the idealism–materialism dilemma and the individualism–holism–systemism trilemma. The main schools in the knowledge of the social are skepticism, apriorism, empiricism, and realism. And the discussions on the methodology of social research include empiricism (priority of data collection), interpretivism (or hermeneutics), and scientism (which combines data collection with theorizing). Finally, the sociology–epistemology connection and the sociology of epistemology are examined. The virtues and flaws of the various stands are pointed out, and reasons for favoring systemism and realism are proposed.

The word 'epistemology' will be taken here in its etymological acceptation, which is the one kept in the Romance languages, namely as philosophy of science. Up until the mid-twentieth century, physics and mathematics held the attention of most philosophers of science. The earliest anthology in the philosophy of social science appeared in 1968, and the *Philosophy of the Social Sciences* quarterly was born two years later. Since then, the discipline has been thriving. We will restrict our attention to the philosophy (or epistemology) of sociology. Three aspects of this discipline will be examined: its ontology, epistemology in its narrow sense (as the theory of knowledge), and methodology.

The matter of the nature of the social has two foci: the idealism–materialism dilemma and the individualism–holism–systemism trilemma. The main schools in the epistemology of the social are skepticism, apriorism, empiricism, and realism. And the discussions of the methodology of social research include empiricism (priority of data collection), interpretivism (or hermeneutics), and scientism (which combines data collection with theorizing). Finally, the sociology–epistemology connection and the sociology of epistemology will be examined. The virtues and flaws of the various stands are pointed out, and reasons for favoring systemism and realism are proposed.

Like any other special branch of the philosophy of science, the philosophy of sociology may be construed as having three main components: ontology (or metaphysics), epistemology in the narrow sense, and methodology (or normative epistemology). This order is not arbitrary, for every research strategy depends upon the type of knowledge being sought, which in turn depends on the nature of the thing studied. Thus, if the object of study is concrete, like a photon or a social group, and if what is sought is some objective knowledge of it, then the scientific method will be adopted. If, by contrast, the object of knowledge is assumed to be spiritual, and if things spiritual are deemed to be beyond the reach of science, then an alternative method will be preferred – such as revelation, intuition, interpretation (Verstehen), unchecked speculation, or narrative. In this article the ontology, epistemology, and methodology of sociology will be examined.

Ontology: Nature of the Social

The Idealism–Materialism Dilemma

Most sociologists adopt the vulgar view that the human being is a composite of body and mind (or spirit or soul), and do not ask how these two components are kept together. Hence, they do not characterize social facts as either material or spiritual, and do not worry about the place of sociology in the system of human knowledge: they are dualists. But some sociologists do worry about these questions, and argue for or against one of two monistic worldviews: idealism or materialism. Idealism is the view that everything is either ideal or ultimately ruled by ideas or symbols, whereas materialism holds that all existents are material or concrete.

In social studies, the idealism–materialism cleavage appears notably in the place and relative weight assigned to the so-called ideal (cultural and political) superstructure vis à vis the material (biological and economic) infrastructure. This very distinction betrays a dualistic stand, even if the preeminence or priority of either 'structure' (system) is postulated.

Consistent idealists will assert that everything social is spiritual (or cultural), whence sociology is a *Geisteswissenschaft* (science of the spiritual). This is the view of hermeneutics, symbolic interactionism, phenomenological sociology, and ethnomethodology. By contrast, materialists assert that persons and social groups are material, though not necessarily confined to the physical or even the biological level. And, unlike idealists, they claim that all cultural and political activities are influenced or even determined by environment, reproduction, and material production. Although it is undeniable that the reproductive patterns and means of subsistence condition culture and politics, it is no less true that the latter in turn help shape the former. In a systemic perspective, changes in the environment and in any of the subsystems of society (the biological, economic, political, and cultural) may initiate processes altering any of the others. People may work or fight, live or die, for King and country, no less than for food and family. There is no 'last analysis.'

Marxism is the oldest, and still the most influential, of the materialist schools in social science. However, given its proposed division between the material infrastructure and the ideal superstructure, it is questionable that Marxism is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32146-8

consistent. Moreover, philosophical Marxism is nowadays split into a large number of different schools, none of which is flourishing. By contrast, some of Marx's contributions to social science have survived, particularly his emphases on production, on the economic roots of the power elites, and on the economic causes of many an international conflict. Even some of Weber's studies are materialist, particularly those of the congruence between Hinduism and the Indian caste system, and on the causes of the decline of slavery in ancient Rome. Likewise the *Annales* historical school owes much to Marx.

Evolutionary psychology, the inheritor of sociobiology, is the newest and most popular materialist school in social studies. The evolutionary psychologists hold that everything social is ultimately biological, having evolved as an adaptation to the natural environment. This view overlooks all the types of self-destructive and antisocial behavior, from suicide to military aggression to environmental degradation to gullibility and ideological zealotry. And it ignores all social inventions (like formal education) and social revolutions (like the information revolution) – hence it is basically conservative. The evolutionary speculations on the origin of mental faculties are even more far-fetched and they lack empirical support. Still, something does remain of the attempt to reduce social science to biology, namely, the reminder that sociologists cannot afford to ignore human physiology and ecology. For example, they should not ignore that malnutrition stunts development and limits productivity; that plagues, foreign invasions, and wars are bound to eliminate some genes and change the immune system; and that a few more centuries of uncontrolled environmental degradation is likely to make our planet uninhabitable.

Evolutionary psychology, which is purely speculative as well as radically reductionist, should not be confused with either social cognitive neuroscience or neurosociology, which are brain-centered and abound in important empirical findings. One of the most striking recent findings is that the human brain contains plenty of 'mirror neurons' which are activated by the perception of other people's actions. The discoverer, Giacomo Rizzolatti, has suggested that our possession of such neurons in abundance explains social emotions, in particular empathy, and thus our spontaneous tendency to associate with others to constitute more or less permanent social systems.

The Individualism–Holism–Systemism Trilemma

There are essentially three consistent views on social structure: that it does not exist, that it is self-existing, and that it is a property of a system composed of individuals. The first, or individualism, identifies 'society' with 'crowd.' The second view, or holism, regards society as an organism, and the individual as wholly subordinated to it. The third, or systemism, views society as a system made up of individuals linked to one another by ties of different kinds and strengths. Individualism stresses agency at the expense of structure, and tends to explain action in terms of 'rational' (self-interested) decisions. Holism emphasizes collective traits, which it claims to be unexplainable, and it explains individuals by wholes. Lastly, systemism admits the novelty of systemic properties, as well as the possibility of explaining them in terms of interactions among individuals and systems.

Individualists rightly see in individual action the root of everything social. And even if they acknowledge the importance of the social context or situation, they resist the very idea of a social system (family, village, firm, etc.) Radical individualism cannot work in sociology if only because the idea of an individualist social science makes as much sense as Lebon's idea of crowd psychology. After all, sociology studies social facts, and although these result from individual actions, they have nonindividual traits. Families, businesses, schools, armies, religious congregations, and other social systems are structured groups, and a structure is a set of relations, among them the bonds holding the system together. Likewise, social processes, such as trade, governance, and war, are impersonal affairs. This is why there are no radical individualist sociologists, with the possible exception of Homans. Many who call themselves methodological individualists, from Weber and Pareto to Coleman and Boudon, admit the peculiarity of the social, and only attempt to explain it as emerging from individual actions; they may be said to be closet systemists.

Holists, from Aristotle to Hegel to Parsons to Luhmann, rightly emphasize the global (or emergent) properties, as well as the centrality of connections. But they exaggerate in claiming that social structure precedes and dominates individuals. This view, radical structuralism, is logically untenable, because there are neither wholes without parts, nor relations without relata, nor networks without nodes. In reality there are only interrelated individuals: relations, like their relata, are only the products of abstraction. For instance, there is no such thing as 'married' in itself, any more than there are unrelated spouses. Hence relations cannot act upon individuals, but interrelated individuals can alter or even sever some relationships – though never all of them at the same time except through suicide.

Systemism, first proposed by Holbach, and often confused with holism, is a sort of synthesis of the latter with individualism. It holds that, because individuals are interrelated, individual behavior can only be understood in terms of intentions as well as social constraints and stimuli. Thus, agency and structure are mutually inseparable. And social change, though brought about by individual action, consists in structural change – that is, change in the nature and strength of the bonds that hold social systems together. A social system may be characterized by its composition, environment (natural and social), structure, and mechanism. A change in any of these features may modify the other three. However, some features may resist changes in the others. For example, social order may survive demographic turnover and environmental catastrophe. Still, all four aspects coexist and, in the long run, important changes in any of the components are bound to alter the other three. Yet individualism ignores structure and underrates environment, whereas holism ignores composition and mechanism, and underrates the natural environment. Only systemism accounts for all four, and is thus likely to favor realistic models of social groups and effective social policies.

Epistemology: Knowledge of the Social

There are as many views on the nature of social studies as general views on the nature of knowledge: skepticism, apriorism, empiricism, and realism.

Skepticism

Skepticism comes in two strengths: moderate (or methodological) and radical (or systematic). The moderate skeptic doubts a thesis as long as no solid evidence for it is available. By contrast, the radical skeptic questions the very possibility of knowledge of some kind. Whereas moderate skepticism is part of the scientific attitude, radical skepticism is destructive and even self-destructive, because if everything is doubtful, why not skepticism itself? The variety of radical skepticism most common concerning the social is cognitive relativism, or the view that social facts cannot be known objectively. This view has been falsified by the growing pile of reliable social data and approximately true sociological hypotheses. Because every research project sets out to find out the truth about something, it presupposes the ability to find it.

Apriorism

Apriorism is speculation without reality checks. The 'grand social theories' of old were aprioristic. So are the evolutionary conjectures on the alleged genetic roots of all social traits, from the incest taboo to social stratification to ideals of human beauty. However, nowadays the most popular a priori theories in social studies are the rational choice models. All of these assume that we always act in self-interest and, moreover, that we proceed 'rationally,' that is, seeking to maximize our expected utilities. These theories are a priori because they are seldom if ever checked against empirical data. In particular, they postulate that agents have certain preference rankings (or utility functions), which are seldom checked. Moreover, some rational choice models are untestable. Thus, if agents behave in an obviously irrational fashion, they are said to be subjectively rational – as when they play the lottery. And if people are altruistic, it will be argued that altruism is nothing but enlightened self-interest. Being impregnable to empirical data, such models are unscientific. However, some rational choice hypotheses, notably the maximizing postulate, are testable, and as a matter of fact they have been falsified by some experimental economists.

Empiricism

Empiricism is the view that only empirical data count: that they are the source, content, and final judges of ideas. Note that the emphasis is on data (reports on observable facts), not on objective facts. So much so, that the logical positivists disallow all talk of objective facts as metaphysical, in going beyond experience. However, empiricism is not limited to positivist philosophers. Behaviorism is a prime example of empiricism, since it deliberately avoids the neural springs of behavior and it eschews theorizing. Ethnomethodology, although ostensibly tributary to the antiempiricist philosophies of Dilthey and Husserl, is another instance of empiricism, because all its practitioners do is to record overt everyday behavior.

Empiricism condemns students to the shallowness characteristic of common sense, because the most interesting facts of any kind are invisible, whence they must be guessed. For example, we can see a status indicator but not status, let alone class; and we can hear political talks, but not the interests that they manifest or disguise. Only serious sociological theory, such as that of Merton and Coleman, can suggest the search for social data that indicate social process, in particular the mechanisms behind appearances. By banning theory, or restricting it to correlating commonsense data, empiricism curtails experience. Moreover, its limitation has elicited some of the postmodern attacks upon science.

Scientific Realism

Scientific realism, or objectivism, has two components: the ontological thesis that the world outside the student exists on its own, and the epistemological thesis that it can be known. As applied to sociology, scientific realism is the view that – as Durkheim put it – social facts are just as real as physical facts, and that they can be known by blending theory with observation, measurement, and (whenever possible) experiment as well. Such knowledge is objective, but partial and gradual. In the case of social facts, research must include not only the facts themselves but also the way the actors perceive them – for example, the fact that people gauge social inequalities, and react to them, by reference to the adjoining social group.

Although scientific realism demands that theories be ultimately justified by data, it does not stifle the sociological imagination the way empiricism does. Indeed, a realist concedes that, just as some data trigger the curiosity of the theorist, at other times untested theories motivate the search for pertinent empirical data. And on still other occasions, advances come from the confrontation of either rival theories or mutually incompatible data. In short, according to scientific realism, theorizing and empirical research should go hand in hand, now matching, now challenging one another.

On the critical side, scientific realists will ask hermeneuticists (interpretivists) and other armchair students to exhibit empirical evidence for their 'interpretations' (hypotheses). They will ask rational choice theorists to specify and justify the utility functions they impute to people. They will remind Marxists and other 'economic imperialists' that real people engage in some activities that are economically unproductive or even counterproductive, and that many a social change – such as the diffusion of the computer and the revival of religious fundamentalism – has come about with little or no class struggle.

Methodology: Social Research Strategy

Methodology is the normative or prescriptive branch of epistemology: it attempts to lay down, justify, refine, or criticize research rules and procedures. Philosophers are interested only in general or cross-disciplinary strategic principles; they leave the special techniques to specialists. Here is a tiny random sample of the class of general methodological problems: How are causation and statistical correlation related? What is a social indicator? How does one infer intention from behavior? Can every social feature be quantified, or are there irreducibly qualitative properties? Is an ideographic (or particularistic) science possible? Here we will review only three rival research strategies.

Dataism: Collect Data

Social research is often equated with data hunting, gathering, and processing, whereas theory is regarded as mere data compression (as in curve fitting), or even as parasitic upon data collection. To be sure, there is no factual (or empirical) science without empirical data. But data alone do not comprise science, because they are what calls for explanation. To understand a bit of the world is to explain it. And explaining involves guessing (and then checking) mechanisms – causal, random, intentional, or hybrid. Hence, theorizing is just as important as data collecting, whether in sociology or in any other factual science. Moreover, only theory, by suggesting interesting variables and their interrelations, can spare us the barrenness and boredom of mindless data gathering.

Interpretivism: Impute Intention

The hemeneutic or interpretivist school holds, rightly, that we can explain an agent's behavior if we know the agent's intentions. But it overlooks the fact that we only have direct access to behavior, not to intentions. That is, the hemeneuticist is faced with the inverse problem of guessing intentions from actions. (Typically, inverse problems are insoluble or have multiple solutions. Think, for example, of the problem of decomposing a given even number into prime numbers, as opposed to the problem of adding prime numbers.) The hermeneuticist is entitled to guess anything, but not to present these guesses (hypotheses) as true unless they have passed suitable empirical tests. The imputation of intentions is not only dicey, it is also insufficient, because intentions, even if known, may explain success but not failure. The social constraints upon agency, and the reactions of other people to one's actions, also must be studied. That is, individuals must be placed in their systems.

Scientism: Gather Data and Craft Theory Consistent with Data

Scientism, first proposed by Condorcet and practiced by Durkheim, is the thesis that the scientific approach is the best of all approaches to cognitive problems of all kinds in all fields. Scientism should not be mistaken for naturalism, or the thesis that the social world is a fragment of nature, and consequently social research should be conducted in exactly the same manner as the study of nature. One may admit that everything social is made, while holding that, once made, a social item is out there, along with rivers and butterflies. That is, social reality is constructed, but not by the knower. And it is changeable, but not because the knowers change their viewpoint. Scientism does not involve the denial of the specificity and irreducibility of the social. It only holds that the scientific method is applicable across all of the disciplinary borders, and hence that it is advisable to trespass them whenever they are artificial. In the case of sociology, scientism favors its close relations with all its sister disciplines, particularly economics, political science, and historiography. Scientism also promotes the proliferation of social indicators, the production and statistical processing of data, and the crafting of empirically testable theories – in particular, 'middle-range' theories and mathematical models.

Hundreds of social indicators have been invented, measured, and used since the mid-1960s, and a whole journal devoted to them has been in existence since 1974. However, their very nature is still somewhat unclear, due to insufficient philosophical investigation. A social indicator bridges the observed to the unobserved, effects to causes, symptoms to mechanisms, and data to hypotheses. However, it does so ambiguously unless it can be shown that no other unobserved trait or event could possibly account for the data. For example, the popularity of low-cost merchandise may measure either consumer satisfaction or poverty; and frequent church attendance may measure either religiosity or the need for social integration (or for respectability). Ambiguity is typical of empirical indicators. By contrast, the indicators used in the more advanced sciences are unambiguous because they are backed by theories. For example, pH is a reliable measure of acidity because it is an index of proton concentration, and theory defines acids as proton donors. The moral for sociology is that the reliability of a social indicator can be increased either by grounding it in a theory or by using it jointly with other indicators of the same hidden trait.

Values and morals would seem to defy scientific realism. After all, Weber held that science ought to be value-free and morally neutral because he wished to preserve objectivity as well as the difference between science and ideology. Now, it is certainly correct to distinguish value judgments from factual statements. But not all values are subjective: some, such as peace, welfare, and truth, are not a matter of taste. Moreover, it is possible to investigate scientifically how values and moral norms emerge, and why some of them have become obsolete. It may even be argued that there are moral facts, such as misery and torture, and consequently that some moral judgments can be justified scientifically. Example: 'Anomie is wrong because it is personally and socially erosive.' It is also correct to distinguish science from ideology. Social scientists endeavor to understand social reality, not to alter it. However, some of their findings can be used by social activists, politicians, and managers to alter social conditions in a rational fashion rather than through costly improvisation. That is, a scientific ideology – one based on social science and empirically testable – is not out of the question.

The Sociology–Epistemology Connection

The traditional view of the relation between science and philosophy is that they are disjoint or even mutually incompatible. This view is mistaken, because science teems with philosophical concepts, such as those of reality, law, knowledge, and value, and because scientific research presupposes the reality and knowability of the world – a philosophical thesis. In short, philosophy and science, in particular social science, overlap partially. This overlap is only partial because only philosophers analyze philosophical constructs, and only scientists are equipped to study social facts.

However, not all philosophies are congenial with science. Whereas some philosophies are extraneous to science, others are relevant to it. For example, phenomenology is

extraneous to science because it rejects the scientific method and consists in egology, or the self-study of the self. Existentialism is even more alien, because it also rejects rationality and does not deal at all with problems of knowledge. The philosophies congenial with science examine, among other things, the problems, concepts, hypotheses, and methods common to science and epistemology. Hence, they can be judged to be pertinent and suitable to science, to the extent that they propose true accounts of actual scientific practice and help advance science. For example, Popper's philosophy is congenial with science because of its rationalism and objectivism. But his devaluation of confirmation is incompatible with the search for truth, since a hypothesis can only be said to be true to some extent if it has been well confirmed.

Sociology of Epistemology

The sociology of philosophy is only embryonic and mostly speculative. In particular, little is known about the social roots of the various epistemologies of sociology. For one thing, it is not true that they are motivated by class interests. Nor is it true that all idealists are reactionary and all materialists progressive. For example, Hegel's followers split into a left wing and a right wing. The second Comte inspired both the Mexican conservatives led by Porfirio Díaz and their contemporary Brazilian progressives.

What is true is that, once crafted for whatever reasons, some philosophical views favor certain social groups rather than others. For instance, radical skepticism, even while revolting against the establishment, objectively favors the latter in that it discourages scientific social research and therefore reasoned and well-organized opposition. Individualist philosophies favor anarchism and neoliberalism. By contrast, holistic philosophies jibe with communitarianism and totalitarianism. In short, some philosophies, particularly if fuzzy, can craftily be interpreted to suit anyone's political interests. But the more abstract philosophical ideas – in particular the logical and semantic ones – are socially neutral. Thus, neither the logical calculi nor the theories of (semantic) meaning and truth refer to social facts, whence they can neither support nor undermine any social ideology. However, they are useful to examine the cogency of any views whatsoever, and to this extent they have a subversive potential.

See also: Action, Theories of Social; Empiricism, History of; Grounded Theory: Methodology and Theory Construction; Interpretive Methods: Macromethods; Interpretive Methods: Micromethods; Macrosociology-Microsociology; Marx, Karl (1818–83); Methodological Individualism in Sociology; Methodological Individualism: Philosophical Aspects; Neutrality, Axiological; Phenomenology in Sociology; Positivism, Sociological; Social Structure; Social System; Sociobiology: Overview; Sociological Theory; Sociology, History of; Weber, Max (1864–1920).

Bibliography

Albert, H., 1994. Kritik der reinen Hermeneutik: Der Antirealismus und das Problem des Verstehens. Mohr, Tübingen.
Barnes, B., 1977. Interests and the Growth of Knowledge. Routledge & Kegan Paul, London.
Boudon, R., 1980. The Crisis in Sociology. Columbia University Press, New York.
Boudon, R., 1994. The Art of Self-persuasion. Polity, Cambridge.
Bunge, M.A., 1996. Finding Philosophy in Social Science. Yale University Press, New Haven, CT.
Bunge, M.A., 1998. Social Science under Debate. University of Toronto Press, Toronto.
Bunge, M.A., 1999. The Sociology-Philosophy Connection. Transaction Publishers, New Brunswick, NJ.
Buss, D.M., 2004. Evolutionary Psychology. Pearson, Boston.
Caplan, A.L., 1978. The Sociobiology Debate. Harper & Row, New York.
Dilthey, W., 1959. Einleitung in die Geisteswissenschaften. In: Gesammelte Schriften, vol. 1. Teubner, Stuttgart.
Durkheim, E., 1895. Les règles de la mèthode sociologique. Flammarion, Paris.
Fay, B., 1996. Contemporary Philosophy of Social Science. Blackwell, Oxford.
Fox, R., 1989. The Search for Society: Quest for Biosocial Science and Morality. Rutgers University Press, New Brunswick, NJ.
Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.
Gellner, E., 1973. Cause and Meaning in the Social Sciences. Routledge & Kegan Paul, London.
Gintis, H., Bowles, S., Boyd, R., Fehr, E. (Eds.), 2005. Moral Sentiments and Material Interests. MIT Press, Cambridge, MA.
Hedström, P., 2005. Dissecting the Social: On the Principles of Analytical Sociology. Cambridge University Press, Cambridge, UK.
Holbach, Thiry d', 1773/1969. Reprint Hildesheim. Le Système Social, vols. 3. Olms, Georg.
Luhmann, Niklas, 1995. Social Systems. Stanford University Press, Stanford, CA.
Menger, C., 1968. Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere. In: Gesammelte Werke, second ed., vol. 2. Mohr, Tübingen.
Merton, R.K., 1968. Social Theory and Social Structure, revised and enlarged ed. The Free Press, New York.
O'Neill, J. (Ed.), 1973. Modes of Individualism and Collectivism. Heinemann, London.
Pareto, V., 1963. A Treatise on General Sociology, vols. 4. Dover, New York.
Parsons, T., 1951. The Social System. The Free Press, New York.
Popper, K.R., 1945. The Open Society and Its Enemies, vols. 2. Routledge & Kegan Paul, London.
Popper, K.R., 1957. The Poverty of Historicism. Routledge & Kegan Paul, London.
Ramachandran, V.S., 2011. The Tell-Tale Brain. W. W. Norton, New York.
Wan, P.Y.-Z., 2011. Reframing the Social: Emergentist Systemism and Social Theory. Ashgate, Farnham, UK.
Weber, M., 1968. Gesammelte Aufsätze zur Wissenschaftslehre. Mohr, Tübingen.
Winch, P., 1858. The Idea of a Social Science. Routledge & Kegan Paul, London.

Sociology, History of

Johan Goudsblom, University of Amsterdam, Amsterdam, The Netherlands
Johan Heilbron, Centre de sociologie et de science politique de la Sorbonne (CNRS-EHESS), Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The development of sociology is rarely analyzed from a sociological point of view, using the insights, concepts, and methods of the discipline itself. The most common historical reconstructions are textbook histories, written for pedagogical purposes presenting the canonical figures in a more or less chronological order. In the absence of a standard sociological account, the development of the discipline may be conveniently regarded as a long-term process structured by five major stages: a long predisciplinary stage up to 1830, marked by what in retrospect may be recognized as proto-sociologies; the formation of an intellectual discipline, both as an idea and an ideal (1830–90); the formation of an academic discipline with diverging national traditions (1890–1930); the establishment of a fully fledged academic discipline (with autonomous degrees, departments, research facilities); and an emerging international hierarchy (1930–70); and a period of crisis and fragmentation, attempts at new synthesis, and globalization (1970–present).

The standard way in which surveys of the history of sociology are written and taught still lags behind actual developments in the field itself. The writing and teaching of the history of sociology continue to be strongly dominated by the tradition of the history of ideas. Only fairly recently have some major attempts been made to view the history of sociology from a sociological angle, using sociological concepts, methods, and theories.

In the absence of a standard *sociological* survey of the development of sociology, we shall in this outline mainly follow the familiar textbook trajectory. This approach inevitably implies distortions. It suggests a pattern of Whig history, in which the history of sociology is made into a neat succession of successes. To balance that bias, the reader should bear in mind that the development of sociology itself has been a social and cultural process, consisting of a multitude of short-term planned actions and interactions with aggregate results that in the long term have not been planned by any single individual. This process took place, moreover, as a part of wider social and cultural developments.

The context was primarily European and, since the second half of the nineteenth century, North American; yet, as Europe and North America belonged to a more encompassing global constellation, we should also consider other traditions of thought and the relevance of the processes of colonization and decolonization affecting nineteenth- and twentieth-century society and culture all over the world. Much of this work, however, for such a broader view of the discipline's development remains to be done.

Sociology, furthermore, was one of the social sciences amid other intellectual disciplines: academic disciplines ranging from physics and biology to history and philosophy, as well as non-academic disciplines including journalism, political debate, and literature. Sociology took on an articulate form in continuous dialogue with those other disciplines, sometimes borrowing ideas from them, sometimes opposing them. Most of the controversies within sociology itself reflected its relations to the larger field of intellectual activities. This would be true of the moral or political views informing sociological theories as well as of the preference for either quantitative, 'scientific' methods or a qualitative, 'hermeneutic' approach (Lepenies, 1988).

In addition to the differentiation into (1) theoretical and methodological orientations, sociology has also become differentiated according to (2) empirical specializations, and (3) national traditions. These various forms of differentiation are respectively related to (1) the prevailing orientation to other intellectual disciplines, including science, philosophy, literature, and journalism; (2) the general process of differentiation of functions in society at large; and (3) international relations. While the variety of empirical specializations and national traditions is easily visible, this is not so clearly the case with the way the national traditions are 'nested' in an evolving international constellation characterized by specific hierarchies, loyalties, and corresponding intellectual interests.

We shall not continuously refer to the more general conditions affecting sociology in the following account which, because of its encyclopedic format, is written in accordance with standard practice. The reader should realize, however, that these conditions are relevant to each of the five major stages into which the development of sociology can conveniently be divided:

1. A long 'predisciplinary' stage up to 1830, marked by what in retrospect may be recognized as 'proto-sociologies'.
2. The formation of an intellectual discipline, both as an idea and an ideal, 1830–90.
3. The formation of an actual academic discipline with diverging national traditions, 1890–1930.
4. The establishment as a fully fledged academic discipline with autonomous degrees, departments, and research facilities and an emerging international hierarchy, 1930–70.
5. A period of crisis and fragmentation, attempts at new synthesis, and globalization, 1970–present.

In discussing the successive stages, we shall have to observe a basic rule of 'phaseology' (Goudsblom et al., 1996: pp. 15–30): when a later phase begins, new elements are added to the previous constellation; but most of the elements which

were already present in the earlier phase will continue to exist, albeit in a modified and less conspicuous guise.

The Predisciplinary Stage

By a discipline we mean a "unit of teaching, research and professional organization" (Heilbron, 1995). There is a body of knowledge, recognized as such under a generally accepted name, laid down in textbooks and discussed in professional journals by practitioners.

All of these elements were absent from the field of sociology prior to the nineteenth century. Yet we can list many names of people who may, arguably, be regarded as predecessors, 'proto-sociologists': people who reflected about social phenomena, collected data about social life, or tried to find patterns in human history. Some individual contributors to these traditions have gained great renown in their own cultures. Thus, from ancient Greece, the names of such men as Herodotus, Plato, Thucydides, and Aristotle stand out. Although they have been 'appropriated' as founding fathers by the practitioners of the established academic disciplines of philosophy and history, many of the problems they addressed lay in the area of what was later to become sociology. Chinese, Indian, and Arab culture have similar intellectual traditions; here, too, a few individuals have become the famous exponents of those traditions (Collins, 1998).

In this brief survey, we shall be able to pursue only the Western line leading up to the emergence of sociology as an intellectual discipline in Western Europe and North America in the nineteenth and twentieth centuries. Since the Middle Ages, there was, on the one hand, a strong tradition of philosophizing about human social life, mainly in theological and moral-legal terms, and, on the other hand, an increasingly strong tradition of collecting facts, most often related to state matters and focused on knowledge of the population and its resources, both at home and abroad (as reported by explorers and travellers). By and large, the two traditions stood apart.

The transition from the predisciplinary stage to the actual conception of sociology as an intellectual discipline in its own right occurred in the eighteenth and nineteenth centuries. The shift consisted partly in conscious attempts to bridge the gap between theory and empirical inquiry, between philosophical reflection and the collection of facts about social life. The leading intellectual model, furthermore, was sought no longer in theology or metaphysics but in empirical science. Insofar as the focus had been either on the political realm or on more 'private' moral issues, it now moved to the structure and development of 'society' at large. This threefold shift in orientation was pioneered by the groups which have become known as the French *philosophes* and the Scottish moral philosophers. Their work was developed in different directions during the last decades of the eighteenth and the first decades of the nineteenth century.

Among the *philosophes*, Montesquieu was one of the first and most influential. He demonstrated that types of government are to be understood in relation to their moral and physical infrastructure. For an understanding of such connections, a recognition of the 'general spirit' of a nation is central. This 'spirit' or 'common character' results from the combined effect of various causes, including climate, religion, and commerce. Instead of drawing deductions from an original principle such as a political or social contract, Montesquieu thus started to unravel the complex interdependencies of human society.

While Montesquieu pioneered the study of what Auguste Comte later was to call 'social statics', his younger contemporary, the future statesman Turgot pointed to the possibility of also viewing human society from a developmental perspective, in other words, in terms of 'social dynamics'. That idea was further elaborated in theories of successive stages or phases by such writers as Condorcet and Henri de Saint-Simon.

In a kindred spirit, the Scottish moral philosopher David Hume proposed introducing "the experimental method of reasoning into moral subjects". Hume and his younger friend Adam Smith rejected theories based on an assumed state of nature. Such reasoning had characterized the natural law tradition – the predominant framework for early modern theories of government and moral obligation (Grotius, Hobbes, Locke, Pufendorf). For Hume, the state of nature was a hypothetical construct, incompatible with the precepts of empirical science. To consider states as the result of an original voluntary act of agreement was illusory and contrary to scientific procedure. Social institutions, in Adam Ferguson's famous phrase, "are the result of human actions, but not the execution of any human design". From this perspective, morality would have to be studied historically and comparatively, tracing its development through the successive stages of hunting, shepherding, agriculture, and commerce.

As the idea of a general social science gained ground, various attempts were made to connect that idea to existing scholarly disciplines. Some groups, particularly in France, pursued the ideal of a social science modeled after the natural sciences. Thus, by applying probability theory to voting procedures and judicial decision making, mathematicians such as Condorcet and Laplace proposed a 'social mathematics'.

British liberal thinkers such as Jeremy Bentham and James Mill somewhat similarly tried to build a social and political theory upon the idea of a moral arithmetic. Individuals and governments alike should promote the amount of happiness and reduce the amount of pain. Their felicific calculus, aimed at the greatest happiness of the greatest number, provided the means for assessing the utility of public institutions. Others, such as Saint-Simon, couched their analysis of early industrial society in a medical idiom advocating a 'social physiology'. Conservative social theorists such as Edmund Burke, Joseph de Maistre, and Louis de Bonald, by contrast, while recognizing the reality of 'social relations' and 'society', opposed the notion of a natural science of society, and rather sought inspiration in theology, philosophy, and history.

The Formation of an Intellectual Discipline (1830–90)

In the course of the nineteenth century, empirical social research became a regular and organized endeavor. Most of it was centered on the 'social question' and intimately linked with governmental agencies and reform movements. Besides a multitude of local initiatives, national associations arose for the coordination of these efforts: the French *Académie des sciences morales et politiques* (1832), the English *National*

Association for the Promotion of Social Science (1857), the *American Social Science Association* (1867), and the German *Verein für Socialpolitik* (1873). More theoretically informed work made use of the research findings, but was carried out mainly by individual scholars who either possessed private means or made a living by writing and lecturing. Thus, Auguste Comte, John Stuart Mill, and Herbert Spencer were for most of their professional lives publicists, men of letters with a strong interest in the sciences. All three contributed significantly to the formation of sociology as an emerging discipline, but none of them was a regular academic or was actively involved in social research.

Other writers, such as Alexis de Tocqueville, Karl Marx, and Adolphe Quetelet, shared similar interests, but consciously avoided the term sociology. Marx and Tocqueville have been incorporated in the canon of the discipline only after 1945. While Tocqueville called for a 'new political science for a new world', Marx' critique of political economy resulted in a theory of capitalism and class domination. The Belgian astronomer and statistician Quetelet continued the efforts of Condorcet and Laplace. He considered the ever growing body of statistics to be the proper basis for a 'social physics'.

When Comte publicly introduced the term sociology (the word had previously been used in an unpublished manuscript of the French revolutionary publicist Sieyès), he was likewise concerned with the idea of a fundamental science of human society. At the time, Comte had no interest in applied knowledge, rejected premature specialization, and wanted to construct an appropriate foundation for social science. The term sociology captured all of these ambitions quite well. It conveyed the idea of a positive science, suggesting both a high level of generality and a great degree of independence from the natural sciences. For Comte, sociology was to become an uncompromising positive science but should not be conceived of in terms of either physics or physiology. Instead, sociology was to biology what biology was to physics: a relatively autonomous science with its own subject matter and mode of conceptualization.

In the *Cours de philosophie positive* (1830–42), Comte explained that different methods prevail in the various sciences: the experimental method in physics, the comparative method in biology, and the historical method in sociology. Sociology had to be founded on a historical law. Since human beings specifically have the capacity to learn, the development of knowledge is the core of human development, and Comte considered his law of the three stages in the development of knowledge as the proper point of departure for sociology.

Comte's plea for sociology as the general social science was well received in England, first in the circle of John Stuart Mill, then by Herbert Spencer. Mill rejected Comte's criticisms of psychology and political economy, but admired his philosophy of science and embraced the idea of sociology. Yet, like the French followers of Comte, Mill never wrote a properly sociological study. One of the first sociological treatises appeared in the 1870s and was written by Spencer – the most widely read sociologist of the nineteenth century.

For Spencer, progressive change was the common denominator of all natural processes. From the maturation of an embryo to the evolution of life in general, all living things evolve from the simple to the complex through successive differentiation and integration. Evolution is the natural process of change from incoherent homogeneity to coherent heterogeneity. Human societies are no exception: they too develop through differentiation, a process steadily accompanied by higher levels of integration and coordination.

Spencer's view of evolution was thus broader in scope than both Comte's sociological and Darwin's biological theory. Combining a laissez-faire view borrowed from economics with an embryological model of growth, Spencer gave the evolutionary principle the status of a universal law and made it the core of his all-embracing system of synthetic philosophy. As stated in his influential essay 'The Social Organism' (1860), Spencer's sociology was cast in the organic idiom, while carefully maintaining an individualist perspective. In *Principles of Sociology* (1876–97) he elaborated his analysis, emphasizing the passage from 'militant' to 'industrial society'. In the latter, social integration is no longer imposed from a controlling center, but the spontaneous result of individuals who cooperate on the basis of a division of labor. Spencer strongly favored a laissez-faire stance, opposing state politics (in matters of welfare as well as in colonial expansion), and inspired what was – inaccurately – called social Darwinism.

His position as a freelance writer and lecturer allowed Spencer to make virtually unlimited claims for the scope of sociology. While he was most outspoken and influential in articulating the idea of sociology as an intellectual superstructure sustained by an evolutionary perspective and overarching all the other social sciences, that idea was shared by several other leading advocates of sociology in Europe and the United States.

The Formation of an Academic Discipline, 1890–1930

Sociology as an academic discipline sociology emerged between 1890 and 1930. The first courses and chairs were established, sociological journals appeared as the primary outlet for research; and sociological associations were founded for furthering intellectual exchange and professional interest. France, England, Germany, and the United States were the core countries, and in all of them distinct national traditions took shape.

Wherever sociologists managed to obtain a position in the academic world, they had to accommodate to the constraints of established institutional divisions. The realm of law was the domain of the legal profession. The study of the past was already successfully claimed by the historians. In a similar vein, psychologists succeeded in monopolizing individual development and behavior as their field of expertise, economists specialized in the economy, anthropologists in the structure and culture of nonliterate societies. Sociologists had little choice but to accept this division of academic labor as a social fact; this left a strong mark on the way they conceived of their own discipline.

Emile Durkheim was the pivotal figure in French sociology. He taught his first course in 1887, wrote classic works on the division of labor, suicide, and the rules of sociological method in the 1890s, and organized a network of highly productive scholars around his journal, the *Année sociologique* (1898–1912). By proclaiming that social facts had to be

explained by other social facts, Durkheim broke away from the predominant biologistic and psychological approaches. His explanation for variations in the suicide rate was a paradigmatic example. The social realm was a reality *sui generis* with its own structure and regularities. Systems of collective representations are an essential part of social life, not merely when studying religion or morality, but also for understanding the economy and political institutions. Following this line of argument, Durkheim and his collaborators (Marcel Mauss, Henri Hubert, Marcel Granet, Maurice Halbwachs, François Simiand) worked on a wide range of topics, contributing to many fields outside of sociology proper.

In England, sociology did not gain a proper entry into the university system until well after the Second World War. The rich tradition of factual inquiry was continued by government officials and wealthy individuals (Charles Booth, B.S. Rowntree), but remained almost entirely separated from academia. Despite the existence of a sociological journal and a professional association, the British academic establishment did not show much interest in sociology. The only chair for a long time had been established in 1907 for L.T. Hobhouse, who, in 1929, was succeeded by his pupil Morris Ginsburg. Both men maintained a cautious evolutionary perspective but neither one built up a network of productive scholars. Whereas British academics showed little interest in studying their own society, they had no such inhibition about studying the exotic regions of the British Empire. Social anthropology became a respected academic discipline and partly served as a substitute for sociological concerns.

Sociology in Germany emerged later than in France and England; a professional association was founded in 1910, a first chair in 1917, and a journal in 1921. The academic pioneers in Germany, men like Ferdinand Tönnies, Georg Simmel, and Max Weber, faced two powerful indigenous traditions. The oldest was that of the *Staatswissenschaften*. The proponents of these disciplines (law, economics, public administration) consistently subsumed civil society to the various functions of the state. This mode of thinking had its roots in eighteenth-century cameralism and was refashioned in the nineteenth century by idealist philosophers such as Hegel. After the founding of the German Empire, the members of the *Verein für Socialpolitik* maintained much of this conception of social science. They were opposed by liberal economists like Carl Menger during the first *Methodenstreit* in the 1880s and 1890s and by Max Weber and others during the second *Methodenstreit* in the decade prior to the First World War. An important aim in Weber's argument about the value-freedom of scientific inquiry was to create a greater distance from the practice of the state sciences.

The other tradition with which German sociologists grappled was the persistent opposition of the natural versus the cultural sciences: the former were supposed to be concerned with causal explanation, the latter with interpretative or 'hermeneutic' understanding. Early sociologists such as Comte and Spencer had often been rejected in Germany as representatives of an inadequately naturalistic conception of the human sciences. Reflecting on these epistemological issues, Max Weber proposed to transcend the dichotomy by ingeniously defining sociology as the science of social action, which by interpretative *understanding* seeks to *explain* its course and consequences.

Weber was nominated for a chair in sociology shortly before his death in 1920. His achievements and those of others of his generation (Simmel, Alfred Vierkandt, Werner Sombart, Franz Oppenheimer) contributed to the recognition of sociology as an academic discipline during the Weimar Republic (1919–33). Sociology expanded rapidly, opening opportunities for a generation of young scholars including Max Horkheimer, Theodor Adorno, Theodor Geiger, Karl Mannheim, and Norbert Elias. Although all these men had initially received a training in different academic fields, they were ready for brilliant careers in sociology. However, the seizure of state power by the National Socialist Party in 1933 made an abrupt end to the favorable outlooks for German sociology, forcing all these scholars along with many of their colleagues into exile.

Sociology's most rapid expansion took place in the United States. The first courses were taught in the 1870s and 1880s, while the expansion of higher education in the 1890s furthered a quick institutionalization. Within two decades, sociology was taught in most colleges and universities; textbooks standardized the concepts and methods of the new discipline.

American sociology developed primarily on the basis of empirical research into current social problems. The main topics were the issues which were widely debated by politicians, religious leaders, and social reformers: massive immigration, urbanization, industrialization, local communities, social (des)integration, and ethnic and race relations. Empirical studies proliferated, based either on statistical material or on actual fieldwork.

By the 1920s, the sociology department in Chicago, founded by Albion Small in 1893, had become the predominant center. Small also played a key role in the *American Journal of Sociology* (1895) and, together with George E. Vincent, wrote the most widely used textbook, *An Introduction to the Study of Society* (1894). The work of the Chicago School was founded on qualitative case studies, using observations, interviews, and life histories. Exemplary were William I. Thomas and Florian Znaniecki's *The Polish Peasant* (1918) and the collective volume on *The City* (1925).

Parallel to the entry of sociology into various national academic systems, the French sociologist René Worms founded an international association, the *Institut international de sociologie* (1893), and an associated journal, the *Revue internationale de sociologie* (1893). Worms brought together a considerable number of the best known European and North American sociologists; the major exception being his French competitors around Emile Durkheim (Fleck, 2011). The membership of his International institute, which held 13 international congresses between 1893 and 1937, included many nonacademic dignitaries, as well as numerous scholars from other disciplines, such as law and economics, and the focus of the meetings was more on the discussion of general themes than on the advancement of research. The limited significance of this incipient internationalization is well illustrated by the fact that the two most important sociologists, Durkheim and Weber, never met one another nor did they ever refer to each other's writings.

Establishment as an Academic Discipline, 1930–70

Between 1930 and 1970, sociology became established as an academic field at most universities in the Western world. The number of practitioners and students greatly increased, and

there was a proliferation of specializations in research and teaching. Far from being an even process of growth all over, however, these trends were marked by far-reaching structural changes: a severe break down in the development of sociology in Europe and a concomitant shift of intellectual leadership in the field from Europe to the United States.

The leaders of the first generation of American academic sociologists all acknowledged an intellectual debt to European masters. Even in the early 1920s, it was still not uncommon for a promising American social scientist to start his career with a European apprenticeship, as did Talcott Parsons who went to London and Heidelberg. At the time, however, the westward exodus of sociologists from Eastern Europe was already underway, as the communist regime in Russia and the fascist regime in Hungary tightened their grip, greatly restricting the margins for sociological inquiry and reflection. The year 1933 brought the decisive rupture: the seizure of power by the Nazis in Germany put an end to the flourishing sociology of the Weimar Republic and forced a great number of prominent and promising sociologists into exile, mostly to the United States (Fleming and Bailyn, 1969).

In contrast to Europe, sociology in the United States suffered no drastic interruptions in the years between 1933 and 1945. It remained a subject in most college curricula, with textbooks such as Park and Burgess's *Introduction to the Science of Society* (1924) continuing to supply an authoritative framework of concepts and ideas. At the same time, practically oriented research in areas of social problems and social policy was stimulated by increasingly generous funding from both government and private foundations (Platt, 1996). Empirical research received a powerful boost after 1941, when social scientists were recruited for the war effort and commissioned to study variables fostering the morale of the Armed Forces.

After 1945, the United States emerged as the major political, economic, and cultural power in the world. In the social sciences, including sociology, American examples were virtually unrivalled in setting an international standard of competence in method and theory. An impoverished Europe lacked the resources for challenging the American ascendency. The relatively autonomous sociological traditions in Germany and Italy were virtually destroyed. Sociology in England had only a minor position at the London School of Economics. In France, the Durkheimian tradition continued to exert considerable intellectual influence among anthropologists and historians, but the institutional basis for sociology as an academic discipline had seriously weakened.

Thus the scene was set for American dominance. The sociology departments at Harvard and Columbia University in particular gained international renown. At Harvard, Talcott Parsons developed a new synthesis of sociological theory, mainly derived from European predecessors, which he presented as a 'general theory of action' – although in its worldwide reception it became known as 'structural functionalism' or simply 'functionalism.' Still, Parsons' generally recognized status as the world's leading sociological theorist did not imply that his ideas put a mark on the entire discipline. There were some influential rival theorists, including his Harvard colleague George Homans who advocated a strictly scientistic 'exchange theory' inspired by the model of *homo economicus*. At Columbia, Parsons' former student Robert Merton propagated a more modest ideal of 'theories of the middle range,' which proved to be very well compatible with the work of his colleague Paul Lazarsfeld, an immigrant from Austria who specialized in the organizational and methodological aspects of social research.

The years between 1955 and 1975 in particular spelled a boom period for American sociology (Turner and Turner, 1990). The successful launching of Sputnik by the Soviets in 1957 inaugurated an era of unprecedentedly high government funding of scientific research in the United States, which also benefited sociology. In the 1960s, a combination of demographic trends (the postwar baby boom) and political discontent made for a dramatic increase in student enrollments. Professional self-confidence, which had been clearly mounting since the early 1940s, surged to new heights. The auspices were favorable for manifold activities in the realm of academic sociology: systematic elaboration of theoretical ideas in the Parsonian fashion; exploration of the possibilities of rivaling theories presented under such labels as exchange theory or symbolic interactionism; and empirical research, conducted in many forms from intensive 'qualitative' case studies to extensive quantitative surveys, and applied to a broad range of social phenomena. Sociologists in most other capitalist-democratic countries outside the United States adopted similarly diverse programs for teaching and research.

The expansion of the discipline was accompanied by more regular forms of international circulation. American foundations like the Rockefeller and the Ford Foundation supported empirical social science research, both inside and outside the US, and provided grants for studying at American universities. UNESCO initiated and funded international social science associations, among them the International Sociological Association (1949), which was founded on the basis of a small number of national organizations from the core countries. The growth of the ISA was initially assured by the increasing membership of national associations (Platt, 1998). In spite of the Cold War but in line with the UNESCO policy to promote international understanding, national associations from several Communist countries in Eastern Europe joined in the late 1950s. In the course of the 1960s and 1970s, membership increased further by allowing individuals and associate members to join, thus breaking away from the United Nations model of national representation. Among the associate members were regional organizations from Europe, Latin America, and Asia. As the meetings of the International Sociological Association testified, by the end of the 1960s sociology was a blooming pluralistic discipline, with a tacitly accepted hierarchy of prestige headed by American sociologists whose ideas set the agenda for the discipline at large. Unintendedly, the publication of the 17-volume *International Encyclopedia of the Social Sciences* in 1968 spelled the ending of that era.

Crisis and Fragmentation, Attempts at New Synthesis, and Globalization 1970–Present

If the years around 1970 formed a watershed in the development of sociology, the tendencies that then became manifest had been at work for some time. Mainstream academic

sociology as represented by the combination of functionalist theory and survey research rapidly lost its appeal to a new and vociferously critical generation. With much programmatic gusto, new 'paradigms' were launched, highly diverse in their implications, but all radiating the same radical spirit. Inspired by political protest movements, a culture of dissent arose in sociology, criticizing the intellectual canon and the professional stance of the discipline. The short-term result of this turmoil was the demise of functionalism as the leading theoretical perspective, and a flurry of controversies about a mixture of ideological, epistemological, and methodological issues. At a longer term, after a reaction to the surge of politicalization had taken place and the world economy was hit by a severe recession in the late 1970s, the manifest controversies receded and a state of fragmentation ensued to which most practitioners tacitly accommodated.

Fragmentation took many forms. The most obvious and least contested form was the proliferation of subdisciplines around an unsystematic array of specific research areas such as the sociology of the family, organizations, and culture. Then, equally noticeable but far more controversial, there was a plurality of theoretical orientations, often with corresponding methodological presumptions. Some theories were primarily attuned to specific forms of microsociology, marked by an affinity to various philosophic schools ranging from phenomenology (embraced by ethnomethodologists), through pragmatism (preferred by symbolic interactionists), to neo-positivism (for rational choice theorists). Network analysis emerged out of the study of microinteractions as well, but it developed into a set of techniques and ideas that could be applied at any level of aggregation. In the realm of macrosociology, functionalism as a perspective gave way to new approaches influenced by Marxism and emphasizing social conflict and long-term social transformations. Fragmentation was further enhanced by a weakening of the hierarchy in the international sociological community, growing geographical diversity, and an increased articulation of national traditions.

None of these changes came out of the blue; but together they produced an almost revolutionary shock to the world of sociology. Whereas the most famous sociologists of the previous generation had all been Americans, now European names again came to the fore: Michel Foucault, Pierre Bourdieu, Anthony Giddens, Jürgen Habermas, and somewhat later, Zygmunt Bauman, Ulrich Beck, and Manuel Castells. During the early 1970s, while many older sociologists deplored the patent end to the illusion of consensus in the discipline, there was also a widespread sense of excitement over new intellectual possibilities and career opportunities. Ideological debates around many current topics, from gender and race relations to the reform of psychiatry, were saturated with sociological references.

The later 1970s saw a dramatic reversal in this mood of confidence. General public interest turned away from sociology, student enrollments dropped rapidly, and research funds shrank – the only sediment of the earlier high tide that remained was internal fragmentation and pluralism. The general stagnation at first reinforced the tendencies toward fragmentation, as established subdisciplines tried to maintain their position in the face of diminishing institutional support for, and intellectual vigor of the sociological discipline as a whole. However, the 1980s and 1990s also brought new attempts toward a synthesis. Thus, in historical and comparative sociology the orthodox Marxist approach tended to be softened by combining it with a more liberal Weberian perspective. 'World systems' theory as originally developed by Andre Gunder Frank and Immanuel Wallerstein helped to break away from the implicit focus on the national state that characterized most sociological work during the preceding stage. Making the transnational order, conceived either as a world system, or alternatively as 'transnational society,' into the centrepiece for sociological study gave counterweight to the tendencies toward national parochialism. After the collapse of communism and parallel to the rapid development of computer technology and the internet, globalization became a key area of debate and research. It soon proved a challenge to the organization of the discipline as well. At the beginning of the twenty-first century, sociology is practiced in virtually every country and region around the world, levels of transnational communication and collaboration are rapidly rising, and English has become the language of global social science. While much of the production of articles and books remains concentrated in North America and Europe, where the most prestigious journals and departments continue to be located (UNESCO, 2010), sociology's more global scope has also provoked a vivid debate about Western hegemony, indigenous traditions of thought, and the diversity of globalizing sociology (Patel, 2010).

Other ventures in historical and comparative sociology have sought synthesis in a different direction. During the 1980s and 1990s, Randall Collins valiantly subjected a wide range of long-term cultural trends (including religion and philosophy as well as technology) to sociological analysis, while Jack Goldstone and Charles Tilly brought sophisticated techniques of statistical analysis to bear on huge sets of historical data. Meanwhile Norbert Elias's work, aiming at a thorough integration of sociology with history and psychoanalysis, was gaining influence inside and beyond the discipline.

These promising signs could not hide the fact, however, that after decades of rapid growth in the 1960s and 1970s, sociology had entered a period of less favorable circumstances. The welfare state came under siege in many Western countries, and neo-liberal policies in favor of deregulating and extending markets gained the upper hand. Even if the collapse of the Soviet empire and the termination of the Cold War have not spelled 'the end of history', these events did lessen interest in ideological matters and theories of society.

At the universities, the established social science disciplines (economics, sociology, anthropology, political science) have faced increasing competition from new multidisciplinary fields of study, which are more immediately focused on policy needs (organizational studies, management, business) or the demands of particular student groups ('Black studies', gender studies). The strains of the present situation are also conducive to a revival of the old polarity between 'scientistic' and 'humanistic' approaches. Provocative challenges to the social sciences have come from sociobiology and the neurosciences, while the scientistic posture gave a great deal of prestige to 'rational choice' models by means of which assumptions and

methods derived from economics are applied to a wide array of non-economic relations. Resisting the invasion of economics, other sociologists advocated an opposite move, to resubmit economic processes and institutions to sociological scrutiny and redevelop economic sociology. Ironically, the humanistic approach is posing a similar risk: embracing culturalistic and postmodern ideas, drawn mainly from current anthropology, literary studies, and hermeneutic philosophy, may also detract sociologists from the core of their discipline – the study of social processes and social structures.

The opposing social currents at work in the world today, as well as the more directly felt institutional pressures and intellectual challenges are confronting sociology with a variety of contradictory claims. Recognizing the many sources of confusion is a first requisite in order to arrive at a balanced response to this situation. Equally important for a more reflexive sociological practice is a continuous reconsideration of the history of the discipline.

See also: Comte, Auguste (1798–1857); Discipline Formation in the Social Sciences; Durkheim, Emile (1858–1917); Marx, Karl (1818–83); Social Science and Universities; Sociological Traditions; Sociology: Overview; Specialization and Recombination of Specialties in the Social Sciences; Spencer, Herbert (1820–1903); Weber, Max (1864–1920).

Bibliography

Boudon, R., Cherkaoui, M., Alexander, J. (Eds.), 1997. The Classical Tradition in Sociology, 8 vols. Sage, London.
Calhoun, C. (Ed.), 2007. Sociology in America. University of Chicago Press, Chicago.
Collins, R., 1998. The Sociology of Philosophies. Belknap Press, Cambridge, MA.
Fleck, C., 2011. A Transnational History of the Social Sciences. Robber Barons, the Third Reich and the Invention of Empirical Social Research. Bloomsbury, London.
Fleming, D., Bailyn, B. (Eds.), 1969. The Intellectual Migration: Europe and America 1930–1960. Harvard University Press, Cambridge, MA.
Goudsblom, J., 1977. Sociology in the Balance. Basil Blackwell, Oxford.
Goudsblom, J., Jones, E., Mennell, S., 1996. The Course of Human History. M.E. Sharpe, Armonk, NY.
Halsey, A.H., 2004. A History of Sociology in Britain. Oxford University Press, Oxford.
Heilbron, J., 1995. The Rise of Social Theory. Polity Press, Cambridge.
Kilminster, R., 1998. The Sociological Revolution. From the Enlightenment to the Golden Age. Routledge, London.
Lepenies, W., 1988. Between Literature and Science: The Rise of Sociology. Cambridge University Press, Cambridge.
Levine, D.N., 1995. Visions of the Sociological Tradition. Chicago University Press, Chicago.
Patel, S. (Ed.), 2010. The ISA International Handbook of Diverse Sociological Traditions. Sage, London.
Platt, J., 1996. A History of Sociological Research Methods in America 1920–1960. Cambridge University Press, Cambridge.
Platt, J., 1998. A Brief History of the International Sociological Association. ISA, Madrid.
Turner, S.P., Turner, J.H., 1990. The Impossible Science. An Institutional Analysis of American Sociology. Sage, Newbury Park, CA.
UNESCO, 2010. World Social Science Report. UNESCO/International Social Science Council, Paris.
Wallerstein, I., 2004. The Uncertainties of Knowledge. Temple University Press, Philadelphia, PA.

Sociology: Overview

Mitch Duneier, Princeton University, Princeton, NJ, USA

Published by Elsevier Ltd.
This article is a revision of the previous edition article by R. Boudon, volume 21, pp. 14581–14585, © 2001, Elsevier Ltd.

Abstract

Scholars of sociology, commonly defined as the scientific study of social relations, social institutions, and societies, conceive the objectives, uses, styles, and methods of their discipline in various ways. One type of sociologist grants it primarily an informative function, that of producing data and analyses to aid decision makers. A second type sees it as serving a critical function, that of identifying the various defects of societies. Finally, a third type sees sociology as a means of explaining social phenomena. These three orientations characterize contemporary as well as classical sociology. Sociologists also endorse a range of methodological and theoretical positions. While some oppose a positivist to an antipositivist approach, others oppose a holistic to an individualistic stance, a rational to an irrational theory of behavior, or a utilitarian to a nonutilitarian view of rationality. Max Weber's sociology is of particular value, for it offers a paradigm capable of transcending these alternatives. One may effectively refute the skepticism that has developed toward sociology and its variety of approaches by focusing on its genuine achievements and demonstrating how actual research has overcome theoretical oppositions in various cases.

The institutionalization of sociology followed a long and circuitous path. The actual term was coined by Auguste Comte and institutionalized in periodicals. The existence of sociology was also recognized implicitly, however. When, in the very first sentence of *The Old Regime and the Revolution*, Alexis de Tocqueville (1986[1856]) wrote "[t]his book is not a book of history," he meant that it dealt with what we would today call 'comparative sociology.' He did not use a word that was still uncommon and probably even considered Comte an eccentric figure. Weber too expressed doubts as to whether his work should be classified as sociology. The gradual institutionalization of the term and discipline owes much to Émile Durkheim and later, the Chicago school of sociology.

Sociology: A Unique or Multiple Discipline?

Despite being contemporaries and able to read each other's works, Durkheim never quoted Weber, nor Weber Durkheim. All the same, as Nisbet (1966) maintains, common inspiration can be detected in the works of Tocqueville, Karl Marx, Georg Simmel, Weber, Vilfredo Pareto, and Durkheim. All appear to agree that the rational utopian theory of society developed by the Enlightenment was unacceptable. Consequently all perceived the social functions of religion, the hierarchical distinctions between social ranks, the role of tradition, and the distinction between 'community' and 'society.'

Describing the identity of classical, and, even more, modern sociology can be more difficult than Nisbet proposes, for both classical and modern sociologists see very different uses for sociology and the ways in which it should be practiced.

Three Types of Objectives

The first type of sociologist sees the discipline as having an essentially *informative* function, however – that of producing data and analyses oriented implicitly and explicitly toward decision makers. Thus, in *Les ouvriers européens* (1855), Frédéric Le Play surveyed the conditions of life among European workers with the final objective of improving the social policies of the time. While the workers themselves would have known about their lives, this local knowledge was not directly available to societal decision makers. Collecting statistics related to crime is another example of sociology's informative function. Such data are crucial to the determination of crime policy. Gabriel Tarde, for example, became a sociologist because, as a judge and later as the head of the French Bureau of Statistics on Crime, he was concerned with an increase in crime rates. Joseph Schumpeter qualified the informative dimension of the social sciences as 'cameral' (*Kameralwissenschaften*). In the 1970s, Paul Lazarsfeld et al. (1967) offered an overview of the production of this 'cameral' sociology in the *Uses of Sociology*.

The second type of sociologist sees the function of sociology as *critical*, that of identifying the defects of actual societies and devising remedies to correct them. Henri de Saint-Simon and Pierre-Joseph Proudhon are considered pioneers in this type of sociology, while members of the Frankfurt School exemplify a modern version of 'critical' sociology.

The third type of sociologist sees sociology as having the same objective as other scientific disciplines, namely, that of explaining social phenomena that appear enigmatic at first glance. Weber, for example, wondered why peasants appeared more attracted to polytheistic than monotheistic religions, while Durkheim was interested in why all religions included the notion of 'soul' in one form or another. In such cases, the interests of sociologists are essentially of a *cognitive* character.

These three basic orientations can appear in combination. Marx was concerned both with explaining social mechanisms and changing society. Durkheim gathered data on suicide with the aim to explain variations in suicide rates, but also considered them indicators of the general state of modern society. All the same, the works of de Tocqueville, Weber, and Durkheim achieved the status of 'classics' essentially because they contained scientifically valid explanations for puzzling social phenomena, rather than for their diagnoses of the society of their time. Durkheim deemed sociology as socially useful only

if it was scientifically sound, meaning that good criticism could derive only from valid explanation. In contrast to Durkheim or Weber, the Frankfurt School is better known for its criticism of modern societies than for convincing explanations of puzzling phenomena.

These three orientations characterize modern, contemporary, and classical sociology. Marcel Foucault's *Discipline and Punish* is generally recognized more a 'critical' than a 'scientific' contribution to the field (Foucault, 1975). According to the 'labeling theory,' some features of individual A can be interpreted as a 'label' by observer B and can structure B's behavior and expectations toward A. This is common knowledge, however, so the theory does not introduce anything new. Its 'critical' function is more undebatable than its cognitive one. So, the 'labeling theory' is not a theory in the scientific sense. Yet as Foucault's theory, it has had a notable influence on our views and those of public officials as to the way in which prisoners should be considered and treated.

Several Types of Orientations

Sociology is often perceived as a discipline with a weak identity, not only because of the existence of its three basically different goals, but also because those sociologists who consider the main goal of their discipline to be the production of valid new knowledge on social phenomena endorse a variety of methodological and theoretical orientations. Weber's work, for example, is seen as illustrating a particular style of 'comprehensive' sociology that is currently opposed to a 'positivist' conception of the field. While the former proposes to analyze social phenomena as the outcome of comprehensible individual actions, the latter analyzes relations between social phenomena. Weber himself considered his version of 'comprehensive' sociology as being very different from that of others, and repeatedly spoke of 'comprehensive sociology in my sense (*in meinem Sinne*).' In the 1960s, 'quantitative' and 'qualitative' sociology were perceived as irreconcilable rivals; while the former deemed the collection of standardized data capable of being subjected to statistical analysis as the most effective research procedure, the latter insisted that such an approach produced biased findings and thus proposed more flexible methods. Robert Merton's 'middle range theory,' which aimed at explaining specific social phenomena, was seen as the proper alternative to the sort of general theory on 'society' as a whole proposed by Parsons, which C. Wright Mills polemically referred to as 'Grand Theory.' These notions illustrate some of the main more or less permanent and distinguishable conceptions of sociology.

Such diversity within the basic theoretical orientations of sociology is also one of its permanent features, observable in contemporary as well as classical sociology. In the early twentieth century, a 'methodological quarrel' set those who believed that sociology should follow the path of other scientific disciplines in opposition to those who, like Heinrich Rickert, maintained that while the goal of natural sciences was 'explanation,' the goal of the social sciences was 'interpretation.' In the 1960s, the quarrel was reopened with the 'positivists,' as their challengers called them, set against the 'hermeneuticians,' who once again saw sociology as charged with disentangling the 'meaning' and 'interpreting' social phenomena rather than identifying their causes, that is, 'explaining' them. Theodore Adorno and Karl Popper, were the main protagonists of the 'antipositivist' and 'positivist' groups, respectively (Adorno et al., 1972[1969]). This 'quarrel' has reappeared chronically in various forms. Thus, not long ago, ethnomethodologists claimed that the idea of producing an objective social knowledge was self-contradictory since social phenomena were understandable only from the inside of a social group: one could interpret, but not explain them.

Weber's sociology is particularly important for it defines a third way between 'positivism' and 'antipositivism.' To Weber, explaining a phenomenon means uncovering its causes. The causes of any social phenomenon reside in the set of individual actions, beliefs, or attitudes of the individuals responsible for it. The sociologist thus has to make these individual actions, beliefs, or attitudes 'understandable.' At the same time, however, understanding an individual action requires finding the action's meaning to the actor, in other words, the motivations and reasons behind it. For Weber this operation amounts to building a theory on the motivations of social actors that is subject to the ordinary criteria that every scientific theory must meet. Thus, if I see somebody cutting wood in his yard on a hot summer day, I would reject the theory that he is cutting wood because he is cold. On the whole, Weber holds that 'understanding' is a crucial moment in the 'explanation' of any social phenomenon.

Sociologists who are mainly interested in the 'cognitive' function of their discipline, namely, in its capacity to produce valid new knowledge, can thus be classified along a first principal dimension, according to whether they see it as capable of producing '*explanations*,' in the manner of other scientific disciplines, or merely generating '*interpretations*' of an unavoidably subjective character. Here, however, one needs to introduce a secondary distinction that opposes those who, like Durkheim in his *Rules*, believe that the subjectivity of social actors should be ignored by sociologists (Durkheim, 1895), and those who, like Weber (1951), consider that the subjectivity of social actors is essential to sociological knowledge and in no way incompatible with the discipline's scientific ambitions.

This last distinction correlates with another one. Some sociologists regard individuals as the atoms of sociological analysis, while others claim that sociology should ignore individuals and concern itself with analyzing associations between social phenomena and variables. While the former follow the principles of what Schumpeter called '*methodological individualism*,' the latter preach '*methodological holism*.' Weber explicitly declares that to him, individuals are the atoms of every sociological analysis and that sociology should be 'individualistic as far as its method is concerned' (Mommsen, 1965). Durkheim, by contrast, sees sociology as devoted to the analysis of the associations between observable social variables. He offers a theoretical defense of this viewpoint in *Rules* (1895) and a practical one in *Suicide* (Durkheim, 1897).

It suffices to evoke a few modern sociological schools to see that the above distinctions, according to which classical sociological works can be categorized, can also be used for modern studies. Thus, 'structuralism' falls on the side of positivism as it proposes to identify the stable relations that prevail

between social variables and claims to be hard science; on the other hand, it is holistic, since it treats individuals as the mere products or reflections of 'structures.' 'Rational choice theorists,' in turn, propose to analyze social phenomena as the effects of individual actions; they firmly believe that sociology can be as objective and scientific as any other discipline but on the condition that it respects the postulate that social phenomena are the outcome of individual actions motivated by individual interests. 'Ethnomethodologists' believe that sociology is an interpretative rather than an explanatory discipline, but are mainly concerned with the interpretations offered by individuals who are embedded in the social setting in question. Following the footsteps of Jakob Burckhardt or Fernard Braudel, holistic sociologists of history, such as Norbert Elias or Immanuel Wallerstein, focus on detecting diffuse orientations within the course of historical and social processes. They are more concerned with the interpretation than the explanation of social phenomena and are interested mainly in identifying global macroscopic trends.

The lasting relevance of the various conceptions of sociology emerging from the typologies discussed above explains why sociology appears as less of a unitary discipline than do many others. Furthermore, the relative popularity of each of these conceptions tends to follow an 'oscillatory' process, as Pareto would have termed it. In the 1960s, many circles considered structuralism the orientation that would make the social sciences genuinely scientific. Sharply criticized, however, it became obsolete. Viewed as 'positivist' and 'holistic,' it was replaced by antipositivist and individualistic currents such as ethnomethodology or 'phenomenology' in the sociological sense.

A Few Genuine Achievements

The best way to refute those skeptical of sociology's diversity is to examine the successes of its various typologies as outlined above.

Durkheim's *Suicide* (1897) is a classical example of the holistic-positivist type of sociology. The work offers systematic analysis of correlations between rates of suicides and a number of independent variables, such as gender, urban versus rural residence, religious affiliation (Protestant, Catholic, Jew), and so forth. Through his research, Durkheim discovered that rates of suicide increase with a rise in anomie. Other things being equal, professionals are exposed to more anomie than are peasants and are thus more prone to suicide. Anomie is also greater during economic booms when expectations tend to be too optimistic and open the way to disillusion, which eventually leads to higher rates of suicide. Durkheim's *Suicide* provides a methodological model that has inspired many writers. Thus, for example, Cusson (1990) has asked why rates of criminality have been characterized by similar trends in most Western countries with the exception of Japan and Switzerland. He attributes this exception to the fact that primary social control is more developed in these two countries, where, for historical reasons, networks of interpersonal relations typical of traditional societies have been maintained even in large cities.

It should be noted, however, that though such analyses are explicitly holistic, they implicitly consider individual motivations. Thus, in *Suicide*, Durkheim contends that periods of political crises can be associated with a decline in suicide rates, and that – other things being equal – they were effectively lower during major political upheavals in France during the last two decades of the nineteenth century, as well as during the crisis between Prussia and Austria in the mid-1860s. Like other theories in *Suicide*, this one sees an association between two variables defined at the societal level, and, in this sense, is holistic. However, it rests implicitly on the 'psychological' assumption that in periods of serious political crises, people are not exclusively concerned with their personal problems. The same can be said of most of Durkheim's analyses in *Suicide*. The activity of professionals, for example, is much less routinized than that of peasants; professionals cannot be content with following the same rules of behavior day after day, but instead are always devising new rules, objectives, procedures, or ideas. The greater uncertainty generated by such novelty breeds a higher level of anxiety. So, despite his official opposition to the psychological explanation of social facts and his recommendation that social facts be explained solely by social facts – as formulated in the *Rules* – Weber's sense of 'understanding' is implicit in Durkheim's *Suicide*. Thus, according to Durkheim, women in the nineteenth century had lower rates of suicide than did men because, being more exposed to social control, they were more guided by routinized rules and norms and thus less exposed to the anxiety generated by uncertainty.

The contrast between Durkheim's holistic antipsychological sociology and Weber's individualistic approach is therefore less marked than sometimes contended. Indeed, Weber is as interested as Durkheim in macroscopic relations. In his *Essays in the Sociology of Religion*, for example, he notes that Mithraism developed among Roman civil servants before reaching other strata (Weber, 1920–1921); it was more palatable to them than the old traditional Roman religion because they perceived it as a symbolic expression of the political philosophy that grounded Roman bureaucracy. Roman civil servants were hierarchized according to grades associated with specific functions; admission into the bureaucracy and promotion from one grade to the next was determined by impersonal procedures. Although the entire system was under the authority of the Emperor, he was regarded as the incarnation of the impersonal rules codified by Roman law. Now, Mithraism was a religion placed under the auspices of a God who too symbolized impersonal rules; as in the case of Freemasonry, so here, believers were admitted and graded according to a scale with the help of impersonal procedures. Weber therefore solved a macroscopic question – why are certain strata more receptive toward a particular religious doctrine? – in an explicitly 'individualistic' fashion, that is, by showing that Roman civil servants had strong reasons for perceiving Mithraism as being more attractive than the old Roman religion, which had originated in the context of a purely agrarian society. In Prussia, Freemasonry appealed more to civil servants than it did to other strata for similar reasons.

The Weberian vein is noticeable in many modern analyses. Merton's classic analysis of the anti-Black attitudes of White American workers during the Great Depression of the 1930s rests on the assumption that White unionists had strong reasons to believe that Black workers, who came from large paternalist farms in the South, were less prepared than White

workers from the South to be loyal to unions (Merton, 1949). Henri Bergeron (1999) wondered why methadone programs adopted in the Netherlands and Switzerland were rejected in France. His answer was that Lacanian theories, highly influential among French psychiatrists, had led them to diagnose drug addiction as the result of the disruption of social relations. French psychiatrists thus concluded that drug addicts could be cured by methodical restoration of their social relations. By contrast, Swiss and Dutch psychiatrists did not believe that there was a theory from which effective remedies against drug addiction could be drawn. As they did not feel capable of curing drug addiction, they believed that their main duty was to protect people, especially against the danger of AIDS, which could be contracted through drug addiction. The French had strong reasons for stressing that their role was to cure; the Dutch and Swiss, on the other hand, firmly believed that their job was to protect since they did not know how to cure. In an analysis such as this, one retrieves the highly flexible theory of rationality used by individualist sociologists who follow the Weberian tradition.

A Particular Version of the Individualistic Tradition: The Rational Choice Model

Contemporary sociology has developed a particular version of the individualistic tradition. Instead of the open, Weberian theory of rationality, it proposes to adopt the theory of rationality developed by economics, according to which social actors are allegedly moved by their interests and are capable of picking out the best means to serve their ends. This model, known as the 'rational choice model,' and first inspired by Jeremy Bentham, is useful. It has led to many constructive theories in several fields, most notably in the sociology of crime, education, and organizations. Its achievements have motivated some sociologists, such as James Coleman, to contend that in sociology it could serve the purpose of a general theory that is able to integrate heterogeneous findings. But this model cannot be deemed a realistic theory of behavior under all circumstances. Mancur Olson (1965) has stressed that his own theory of collective action, which applies the postulates of neoclassical economics, can explain why collective action does not occur as long as its axioms are believed to be realistic. He adds, however, that in many circumstances, it does not occur simply because people are moved and mobilized not only by their interests but also by their beliefs. Now, beliefs can seldom be satisfactorily explained by the interests of social actors alone. All in all, the 'rational choice model' has produced many convincing theories, but it cannot be regarded as a general theory, neither actually nor potentially.

What the styles of research illustrated by these examples hold in common are explanations of social phenomena that follow the criteria generally used in all scientific disciplines. In this sense, they belong to the 'positivist' category of the previous typology.

Interpretation versus Explanation

In his book on the methodology of the social sciences, *Problems of the Philosophy of History*, Simmel (1892) made the important point that history and the social sciences naturally raise two types of questions: one that can and one that cannot be given a unique, scientifically controlled answer. According to Rickert's vocabulary, some questions raised by sociologists aim at 'explaining' while others aim at 'interpreting' social phenomena. In his *Philosophy of Money*, Simmel (1900) himself proposed an *explanatory* theory for the decline of feudalism: from the moment that peasants in the Middle Ages could pay their rents in money rather than in kind, they were able to decide what to grow; placed in a position that allowed them to produce and sell surpluses, they became eager to participate in collective actions against landowners. Thus an apparently minor institutional innovation generated a chain reaction that eroded the feudal organization of society.

Other questions, however, call for '*interpretation*' (or '*description*') rather than 'explanation.' What are the main features of modernity? Will the twenty-first century be characterized by a religious revival? What does it mean to be poor in an affluent society? How do people live in poor areas? Such questions, of course, cannot be answered in one way. There are many ways of describing life conditions in a poor area. In short, thanks to the very nature of such questions, answers to them are doomed to be at least partly subjective. Actually, as Simmel states, questions of this sort can be answered in an interesting fashion only if analysts are able to inject their own sensibility into their description or interpretation. On the whole, answering these questions requires an ability that makes sociologists akin to artists. A good biographer, for instance, cannot simply follow the criteria that define a scientific work. He or she must also be able to give unity to the numerous observations collected from witnesses on the subject of the biography, on that person's behavior and decisions in innumerable circumstances. Many sociological works are faced with the same types of questions. Anthony Giddens's (1984) study of postmodern society, like Burckhardt's earlier study of the *Kultur der Renaissance in Italien* (1952[1860]), is an example of a work that attempts to characterize the main features of a historical era or identify broad historical trends. Other works dealing not with societies as wholes, but rather with local phenomena also belong to the interpretative–descriptive rather than the explanatory category; although they deal with social reality, they evoke the procedures of novelists. Oscar Lewis's *The Children of Sanchez* (1961) is a classic illustration of such 'literary' sociology. The anthropologist Clifford Geertz (1973) coined the expression 'thick description' to characterize the methodology of this type of work, while W.G. Runciman (1983–1997) developed methodological considerations that seem to elaborate on Simmel's insights.

A Complex Identity

In sum, all the types of sociology derived from the combination of these three principal ideal-typical objectives (explaining, informing, criticizing) and the four types generated by the combination of binary variables, 'holism' versus 'individualism' and 'positivism' versus 'hermeneutics' can be illustrated as easily through examples drawn from contemporary as from classical sociology. Perhaps this demonstrates that all these types are legitimate and fruitful. The price of this heterogeneity

lies in the difficulty of pinpointing the identity of sociology, while its benefit lies in sociology's ability to fill the gaps that other social sciences more devoted to a single paradigm, such as economics, are unable to fill.

See also: Action, Theories of Social; Critical Theory: Contemporary; Critical Theory; Durkheim, Emile (1858–1917); Hermeneutics; Historical Sociology; Marx, Karl (1818–83); Parsons, Talcott (1902–79); Phenomenology in Sociology; Positivism, Sociological; Rational Choice Theory in Sociology; Social System; Sociological Theory; Sociology, Epistemology of; Sociology, History of; Weber, Max (1864–1920).

Bibliography

Adorno, T.W., Dahrendorf, R., Albert, H., Popper, K.R., 1972[1969]. Der Positivismusstreit in der deutschen Soziologie. Luchterhand, Neuwied, Germany.

Bergeron, H., 1999. Soigner ou prendre soin des Toxicomanes: Anatomie d'une Croyance Collective. Presses Universitaires de France, Paris.

Burckhardt, J., 1952[1860]. Die Kultur der Renaissance in Italien. Kröner, Stuttgart, Germany.

Cusson, M., 1990. Croissance et Décroissance du Crime. Presses Universitaires de France, Paris.

Durkheim, E., 1895. Les Règles de la Méthode Sociologique. Alcan, Paris.

Durkheim, E., 1897. Le Suicide, Étude Sociologique. Alcan, Paris.

Foucault, M., 1975. Surveiller et Punir; Naissance de la Prison. Gallimard, Paris.

Geertz, C., 1973. The Interpretation of Cultures. Basic Books, New York.

Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge, UK.

Lazarsfeld, P.F., Sewell, W.H., Wilensky, H.L. (Eds.), 1967. The Uses of Sociology. Basic Books, New York.

Le Play, M.F., 1855. Les Ouvriers Européens, second ed. Imprimerìe Imperiale, Paris.

Lewis, O., 1961. The Children of Sanchez, Autobiography of a Mexican Family. Random House, New York.

Merton, R.K., 1949. Social Theory and Social Structure. Free Press, Glencoe, IL.

Mommsen, W., 1965. Max Weber's political sociology and his philosophy of world history. International Social Science Journal 17 (1), 44.

Nisbet, R.A., 1966. The Sociological Tradition. Basic Books, New York.

Olson Jr., M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Runciman, W.G., 1983–1997. A Treatise on Social Theory. Cambridge University Press, Cambridge, UK.

Simmel, G., 1892. Die Probleme der Geschichtsphilosophie. Duncker & Humblot, Leipzig, Germany.

Simmel, G., 1900. Philosophie des Geldes. Duncker & Humblot, Leipzig, Germany.

deTocqueville, A., 1986[1856]. L'Ancien Régime et la Révolution. Levy, Paris.

Weber, M., 1920–1921. Gesammelte Aufsätze zur Religionssoziologie. Mohr, Tübingen, Germany.

Weber, M., 1951. Aufsätze zur Wissenschaftslehre. Mohr, Tübingen, Germany.

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 23

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 23

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 23 *xliii*

VOLUME 23

Solidarity: History of the Concept 1
Patrick Cingolani

Solidarity, Sociology of 6
Michael Hechter

Somatoform Disorders 10
Jörg Angenendt and Martin Härter

Somatotherapy, History of 16
Santosh K Chaturvedi and Soumya Parameshwaran

Sorokin, Pitirim Alexandrovich (1889–1968) 22
Inna Ponomareva

Source Monitoring 28
Sarah E MacPherson

South America, Archaeology of 36
Dante Angelo and John H Walker

South America: Sociocultural Aspects 43
Norman E Whitten, Jr

South Asia, Archaeology of 50
Namita Sugandhi

South Asian Legal Traditions 58
Rohit De

Southeast Asia, Archaeology of 63
Miriam T Stark

Southeast Asia: Sociocultural Aspects 70
Hans-Dieter Evers

Southern Africa: Sociocultural Aspects 75
Robert Thornton

Space and Social Theory in Geography 82
Barney Warf

Space, Linguistic Expression of 89
Penelope Brown

Spatial Analysis in Geography 94
Manfred M Fischer

Spatial Association, Measures of 100
Arthur Getis

Spatial Autocorrelation 105
Robert Haining

Spatial Cognition 111
Daniel R Montello

Spatial Cognition during Infancy and Early Childhood Across Cultures, Development of 116
Jordy Kaufman and Mahla Cameron-Bradley

Spatial Data 126
Stephen C Guptill

Spatial Data Infrastructures 130
Max Craglia

Spatial Decision Support Systems 136
Robert G Pontius Jr and Kangping Si

Spatial Dynamics and Crime 142
Martin A Andresen

Spatial Equity 148
Richard Morrill

Spatial Interaction 152
Morton E O'Kelly

Spatial Mismatch 157
Sara McLafferty

Spatial Navigation 161
Thomas Wolbers

Spatial Optimization Models 172
Richard Church

Spatial Pattern, Analysis of 178
Lauren M Scott

Spatial Sampling 185
Robert Haining

Spatial Statistical Methods 191
L Mark Berliner

Spatial Thinking in the Social Sciences 198
Paul Claval

Spatial–Temporal Modeling 202
Emilio Casetti

Special Economic Zones and Cluster Dynamics: China 207
Lauretta Rubini, Marco R Di Tommaso, and Elisa Barbieri

Special Education 213
John D Molteni, Jule McCombes-Tolis, and Melissa Root

Special Education in the United States: Legal History 219
Mitchell L Yell

Specialization and Recombination of Specialties in the Social Sciences 225
Mattei Dogan

Speech Errors, Psychology of 229
Merrill F Garrett

Speech Perception 235
Dominic W Massaro

Speech Processing, The Cortical Organization of 243
Josef P Rauschecker

Speech Production, Neural Basis of 250
Juliana V Baldo and Nina F Dronkers

Speech Production, Psychology of 255
Herbert Schriefers and Gabriella Vigliocco

Speech Recognition and Production by Machines 259
Chin-Hui Lee

Spelling, Psychology of 264
Dominiek Sandra

Spencer, Herbert (1820–1903) 274
Jonathan H Turner

Sperry, Roger Walcott (1913–94) 279
Theodore J Voneida

Spirit Possession, Anthropology of 282
Ute Luig

Spirituality 285
Margaret Holloway

Split Brains and Human Consciousness 291
Colwyn Trevarthen

Sport, Anthropology of 299
Noel Dyck

Sport, Historiography of 307
Wray Vamplew

Sport, Sociology of 313
Peter G Craig

Sports, Economics of 318
Dejonghe Trudo

Standpoint Theory, in Science 324
Alison Wylie and Sergio Sismondo

Startle Reflex and Health 331
Terry D Blumenthal

State and Local Taxation 338
Lars P Feld and Friedrich Schneider

State and Society: History of the Concept 343
Gianfranco Poggi

State: Anthropological Aspects 347
John Borneman

State Formation, Theory of 353
Isebill V Gruhn

State, Geography of 356
David Storey

State, History of 362
Wolfgang Reinhard

State, Sociology of the 368
Hilton L Root

States and Civilizations, Archaeology of 371
EM Brumfiel

Statistical Analysis Packages 376
Hans-Jürgen Andreß

Statistical Analysis: Multilevel Methods 381
William M Mason

Statistical Analysis, Special Problems of: Outliers 387
Toby Lewis

Statistical Clustering 392
John A Hartigan

Statistical Distribution Models: Goodness of Fit Tests 397
John CW Rayner, Oliver Thas, and Donald J Best

Statistical Identification and Estimability 405
Gordon M Kaufman

Statistical Pattern Recognition 411
Richard M Golden

Statistical Sufficiency 418
Nancy Reid

Statistics, History of 423
Alain Desrosières

Status and Role, Social Psychology of 428
Lynn G Chin

Status and Role: Structural Aspects 435
John Scott

Stem Cell Research and Cloning 440
Catherine Waldby

Stereotypes in Social Psychology 446
Olivier Klein and Philippe Bernard

Stevens, Stanley Smith (1906–73) 453
Robert Teghtsoonian

Stigler, George Joseph (1911–91) 457
Milton Friedman

Stigma, Social Psychology of 461
Lauri Hyers, Eleanor Brown, and Jacqueline Sullivan

Stochastic Models 467
David J Bartholomew

Stochastic Optimization 473
Lauren A Hannah

Stock Market Predictability 482
Burton G Malkiel

Stockholders' Ownership and Control 490
Michael Useem and Gerald F Davis

Strain Theories and Crime 495
Sung Joon Jang and Robert Agnew

Strategic Adaptation 501
Torben J Andersen

Strategic Management 508
Robert A Burgelman

Strategic Planning for Public and Nonprofit Organizations 515
John M Bryson

Strategizing 522
J-C Spender

Strauss, Anselm (1916–96) 527
Isabelle Baszanger

Street Children: Cultural Concerns 533
Elizabeth A Swart and James D Wright

Stress and Cardiac Response 539
Jaime Vila, José L Mata, and Pedro Guerra

Stress at Work 546
Johannes Siegrist

Stress, Coping and Health 551
Kate Faasse and Keith J Petrie

Stress: Effects on Neonatal and Child Development 556
Vivette Glover

Stress in Adolescence: Effects on Development 562
Nada Krapić, Jasna Hudek-Knežević, and Igor Kardum

Stress in Childhood: Effects on Development 570
Steven G Little and Angeleque Akin-Little

Stress in Infancy and Early Childhood: Effects on Development 577
Jenalee R Doom and Megan R Gunnar

Stress: Psychological Perspectives 583
Scott M Monroe and Lori F Cummins

Strikes 588
John Kennan

Strong Program, Sociology of 592
David Bloor

Structural Adjustment in Africa: Implications for Social Work 598
Lengwe-Katembula J Mwansa

Structural Balance Theory: Recent Developments 605
Peter Abell

Structural Contingency Theory 609
Lex Donaldson

Structural Equation Modeling 615
Shawn Bauldry

Structural Equivalence: Meaning and Measures 621
Stephen P Borgatti and Travis J Grosser

Structuralism 626
Mauro W Barbosa de Almeida

Subaltern History 632
Partha Chatterjee

Subculture, Sociology of 637
Rebecca G Adams, Amy M Ernstes, and Kelly M Lucey

Subjective Wellbeing and Culture 643
Sim Jing Han, Sharon Koh, and Christie Napa Scollon

Subjective Wellbeing, Psychology of 648
Christina N Armenta, Peter M Ruberton, and Sonja Lyubomirsky

Sub-Saharan African Legal Traditions 654
Tinenenji Banda

Suburbanization and Suburbanism 660
Richard Harris

Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model 667
Frieder R Lang and Margund K Rohr

Suicide 673
Thomas Bronisch

Suicide, Sociology of 682
Julie A Phillips

Sumner, William (1840–1910) 689
Robert Bannister

Sun Exposure and Skin Cancer Prevention 696
Natalie Schüz and Michael Eid

Supervision, Abusive 701
Christine MY Kermond and John M Schaubroeck

Supply Chain Management 709
Lauren Xiaoyuan Lu and Jayashankar M Swaminathan

Suprasegmentals 714
Carlos Gussenhoven

Supreme Courts 722
Charles R Epp

Surveillance and Privacy, Geography of 727
Till F Paasche and Francisco R Klauser

Surveillance Studies 733
Gary T Marx

Survey and Excavation (Field Methods) in Archaeology 742
Martha Sharp Joukowsky

Survey Research Centers and Companies 747
Helmut K Anheier and Nikolas Scherer

Surveys and Polling: Ethical Aspects 752
Constance F Citro

Surveys: Extreme Response 758
Daniel M Bolt

Surveys: Question Wording and Response Categories 764
Nora C Schaeffer and Jennifer Dykema

Survival Analysis: Introduction 771
Niels Keiding and Per K Andersen

Survival Analysis: Overview 776
Ross L Prentice and John D Kalbfleisch

Survival Analysis: Whether, and If So, When? 783
Suzanne E Graham, John B Willett, and Judith D Singer

Sustainability: A Dialog of Values 794
John Blewitt

Sustainability and Sustainability Science 801
Robert W Kates

Sustainable Agriculture 807
Alison Heslin

Sustainable Development 812
Karl-Werner Brand

Sustainable Development and Social Work 817
Susan Buckingham and Victoria J Kina

Sustainable Development: An Economic Perspective 823
Edward B Barbier and Joanne C Burgess

Sustainable Lifestyles 828
Stewart Barr

Sustainable Practices 834
Michael Jonas and Beate Littig

Sustainable Production and Consumption 839
Daniel Welch

Sustainable Transportation 845
David L Greene

Symbolic Boundaries 850
Michèle Lamont, Sabrina Pendergrass, and Mark Pachucki

Symbolic Interaction: Methodology 856
Andrea Fontana

Symbolism in Anthropology 860
Janet Hoskins

Symptom Perception, Awareness and Interpretation 866
Omer Van den Bergh, Katleen Bogaerts, and Ilse Van Diest

Synaptic Mechanisms and Protein Synthesis in Memory 873
Fabio Benfenati and Flavia Valtorta

Syncretism 881
André Droogers

Syntax (General) 885
Masayoshi Shibatani

Syntax–Phonology Interface 895
Elisabeth Selkirk

Syntax–Semantics Interface 900
Erkki Luuk

Systematic Reviewing and Meta-Analysis 906
Jeffrey C Valentine, Therese D Pigott, and Timothy Lau

CONTRIBUTORS TO VOLUME 23

Peter Abell
London School of Economics, London, UK

Rebecca G. Adams
The University of North Carolina at Greensboro, Greensboro, NC, USA

Robert Agnew
Emory University, Atlanta, GA, USA

Angeleque Akin-Little
Akin-Little and Little Behavioral Psychology Consultants, Malone, NY, USA

Per K. Andersen
University of Copenhagen, Copenhagen K, Denmark

Torben J. Andersen
Center for Global Strategic Responsiveness, Copenhagen Business School, Frederiksberg, Denmark

Hans-Jürgen Andreß
Universität zu Köln, Köln, Germany

Martin A. Andresen
Simon Fraser University, Burnaby, BC, Canada

Dante Angelo
Universidad de Tarapaca, Aríca, Chile

Jörg Angenendt
University Medical Center, University Hospital Freiburg, Freiburg, Germany

Helmut K. Anheier
Hertie School of Governance, Berlin, Germany

Christina N. Armenta
University of California Riverside, Riverside, CA, USA

Juliana V. Baldo
Center for Aphasia and Related Disorders, VA Northern California Health Care System, Martinez, CA, USA

Tinenenji Banda
Cornell Law School, Ithaca, NY, USA

Robert Bannister
Swarthmore, PA, USA

Edward B. Barbier
University of Wyoming, Laramie, WY, USA

Elisa Barbieri
University of Udine, Udine, Italy

Mauro W. Barbosa de Almeida
Departamento de Antropologia, Universidade Estadual de Campinas, São Paulo, Brazil

Stewart Barr
College of Life and Environmental Sciences, University of Exeter, Exeter, UK

David J. Bartholomew
London School of Economics and Political Science, London, UK

Isabelle Baszanger
CNRS-CERMES, France

Shawn Bauldry
University of Alabama at Birmingham, Birmingham, AL, USA

Fabio Benfenati
University of Genova Medical School, Genova, Italy; and Fondazione Istituto Italiano di Tecnologia, Genova, Italy

L. Mark Berliner
The Ohio State University, Columbus, OH, USA

Philippe Bernard
Université Libre de Bruxelles, Bruxelles, Belgium

Donald J. Best
University of Newcastle, Callaghan, NSW, Australia

John Blewitt
Aston Business School, Aston University, Birmingham, UK

David Bloor
University of Edinburgh, Edinburgh, UK

Terry D. Blumenthal
Wake Forest University, Winston-Salem, NC, USA

Katleen Bogaerts
University of Leuven, Leuven, Belgium

Daniel M. Bolt
University of Wisconsin-Madison, Madison, WI, USA

Stephen P. Borgatti
LINKS Center for Social Network Analysis, Gatton College of Business and Economics, University of Kentucky, Lexington, KY, USA

John Borneman
Princeton University, Princeton, NJ, USA

Karl-Werner Brand
Technical University of Munich, Munich, Germany

Thomas Bronisch
Max-Planck-Institute of Psychiatry, Munich, Germany

Eleanor Brown
West Chester University, West Chester, PA, USA

Penelope Brown
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

E.M. Brumfiel
Albion College, Albion, MI, USA

John M. Bryson
University of Minnesota, Minneapolis, MN, USA

Susan Buckingham
Brunel University, Uxbridge, UK

Robert A. Burgelman
Stanford University, Stanford, CA, USA

Joanne C. Burgess
University of Wyoming, Laramie, WY, USA

Mahla Cameron-Bradley
Swinburne University of Technology, Hawthorn, VIC, Australia

Emilio Casetti
Ohio State University, Columbus, OH, USA

Partha Chatterjee
Columbia University, New York, NY, USA; and Centre for Studies in Social Sciences, Calcutta, India

Santosh K. Chaturvedi
National Institute of Mental Health and Neurosciences, Bangalore, India

Lynn G. Chin
Washington & Lee University, Lexington, VA, USA

Richard Church
University of California Santa Barbara, Santa Barbara, CA, USA

Patrick Cingolani
Bâtiment Olympe de Gouges, University-Paris Diderot, Paris Cédex 13, France

Constance F. Citro
Committee on National Statistics, National Research Council/National Academy of Sciences, Washington, DC, USA

Paul Claval
University of Paris-Sorbonne, Paris, France

Max Craglia
European Commission, Joint Research Centre, Ispra, Varese, Italy

Peter G. Craig
University of Bedfordshire, Bedford, UK

Lori F. Cummins
University of Notre Dame, Notre Dame, IN, USA

Gerald F. Davis
The University of Michigan, Ann Arbor, MI, USA

Rohit De
Department of History, Yale University, New Haven, CT, USA

Alain Desrosières
INSEE, Paris, France

Ilse Van Diest
University of Leuven, Leuven, Belgium

Marco R. Di Tommaso
University of Ferrara, Ferrara, Italy

Mattei Dogan[†]
National Center for Scientific Research, Paris, France

Lex Donaldson
School of Management, UNSW-Business School, University of New South Wales, Sydney, NSW, Australia

Jenalee R. Doom
University of Minnesota, Minneapolis, MN, USA

Nina F. Dronkers
Center for Aphasia and Related Disorders, VA Northern California Health Care System, Martinez, CA, USA; and University of California, Davis, CA, USA

[†] Deceased.

André Droogers
Vrije Universiteit, Amsterdam, The Netherlands

Noel Dyck
Simon Fraser University, Burnaby, BC, Canada

Jennifer Dykema
University of Wisconsin Survey Center, University of Wisconsin-Madison, Madison, WI, USA

Michael Eid
Educational Science and Psychology, Freie Universität Berlin, Berlin, Germany

Charles R. Epp
University of Kansas, Lawrence, KS, USA

Amy M. Ernstes
The University of North Carolina at Greensboro, Greensboro, NC, USA

Hans-Dieter Evers
Universiti Brunei Darussalam, Brunei Darussalam

Kate Faasse
Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand

Lars P. Feld
University of St. Gallen, St Gallen, Switzerland

Manfred M. Fischer
Vienna University of Economics and Business, Vienna, Austria

Andrea Fontana
University of Nevada Las Vegas, Las Vegas, NV, USA

Milton Friedman
Hoover Institution, Stanford University, CA, USA

Merrill F. Garrett
University of Arizona, Tucson, AZ, USA

Arthur Getis
San Diego State University, San Diego, CA, USA

Vivette Glover
Imperial College London, London, UK

Richard M. Golden
University of Texas at Dallas, Richardson, TX, USA

Suzanne E. Graham
Division of Educational Studies, University of New Hampshire, Durham, NH, USA

David L. Greene
Oak Ridge, National Laboratory, Oak Ridge, TN, USA

Travis J. Grosser
LINKS Center for Social Network Analysis, Gatton College of Business and Economics, University of Kentucky, Lexington, KY, USA

Isebill V. Gruhn
University of California, Santa Cruz, CA, USA

Pedro Guerra
Facultad de Psicología, University of Granada, Granada, Spain

Megan R. Gunnar
University of Minnesota, Minneapolis, MN, USA

Stephen C. Guptill
U.S. Geological Survey (Retired), Oakton, VA, USA

Carlos Gussenhoven
Radboud University Nijmegen, Nijmegen, The Netherlands

Robert Haining
Human Geography, University of Cambridge, Cambridge, UK

Sim Jing Han
Singapore Management University, Singapore, Singapore

Lauren A. Hannah
Columbia University, New York, NY, USA

Richard Harris
McMaster University, Hamilton, ON, Canada

Martin Härter
Center for Psychosocial Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

John A. Hartigan
Yale University, New Haven, CT, USA

Michael Hechter
Arizona State University, Tempe, AZ, USA

Alison Heslin
Emory University, Atlanta, GA, USA

Margaret Holloway
School of Social Sciences, University of Hull, Hull, UK

Janet Hoskins
University of Southern California, Los Angeles, CA, USA

Jasna Hudek-Knežević
University of Rijeka, Rijeka, Croatia

Lauri Hyers
West Chester University, West Chester, PA, USA

Sung Joon Jang
Baylor University, Waco, TX, USA

Michael Jonas
IHS, Vienna, Austria

Martha Sharp Joukowsky
Brown University, Providence, RI, USA

John D. Kalbfleisch
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Igor Kardum
University of Rijeka, Rijeka, Croatia

Robert W. Kates
Trenton, ME, USA

Gordon M. Kaufman
Massachusetts Institute of Technology, Cambridge, MA, USA

Jordy Kaufman
Swinburne University of Technology, Hawthorn, VIC, Australia

Niels Keiding
University of Copenhagen, Copenhagen K, Denmark

John Kennan
University of Wisconsin-Madison, Madison, WI, USA; NBER, Cambridge, MA, USA; and IZA, Bonn, Germany

Christine M.Y. Kermond
Michigan State University, East Lansing, MI, USA

Victoria J. Kina
University of Dundee, Dundee, UK

Francisco R. Klauser
Institut de Géographie, Université de Neuchâtel, Neuchâtel, Switzerland

Olivier Klein
Université Libre de Bruxelles, Bruxelles, Belgium

Sharon Koh
Singapore Management University, Singapore, Singapore

Nada Krapić
University of Rijeka, Rijeka, Croatia

Michèle Lamont
Harvard University, Cambridge, MA, USA

Frieder R. Lang
Institute of Psychogerontology, University of Erlangen-Nuremberg, Nürnberg, Germany

Timothy Lau
University of Louisville, Louisville, KY, USA

Chin-Hui Lee
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Toby Lewis
University of East Anglia, Norwich, UK

Beate Littig
IHS, Vienna, Austria

Steven G. Little
Walden University, Minneapolis, MN, USA

Lauren Xiaoyuan Lu
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Kelly M. Lucey
North Carolina State University, Raleigh, NC, USA

Ute Luig
Freie Universität Berlin, Berlin, Germany

Erkki Luuk
University of Tartu, Tartu, Estonia; and Stockholm University, Stockholm, Sweden

Sonja Lyubomirsky
University of California Riverside, Riverside, CA, USA

Sarah E. MacPherson
Centre for Cognitive Ageing and Cognitive Epidemiology, Human Cognitive Neuroscience, University of Edinburgh, Edinburgh, UK

Burton G. Malkiel
Princeton University, Princeton, NJ, USA

Gary T. Marx
Massachusetts Institute of Technology, Cambridge, MA, USA

William M. Mason
University of California, Santa Monica, CA, USA

Dominic W. Massaro
University of California Santa Cruz, Santa Cruz, CA, USA

José L. Mata
Facultad de Psicología, University of Granada, Granada, Spain

Jule McCombes-Tolis
University of Saint Joseph, West Hartford, CT, USA

Sara McLafferty
University of Illinois at Urbana-Champaign, Champaign, IL, USA

John D. Molteni
University of Saint Joseph, West Hartford, CT, USA

Scott M. Monroe
University of Notre Dame, Notre Dame, IN, USA

Daniel R. Montello
University of California, Santa Barbara, CA, USA

Richard Morrill
University of Washington, Seattle, WA, USA

Lengwe-Katembula J. Mwansa
University of Botswana, Gaborone, Botswana

Morton E. O'Kelly
The Ohio State University, Columbus, OH, USA

Till F. Paasche
Soran University, Soran City, Kurdish Region Iraq

Mark Pachucki
Harvard Medical School, Boston, MA, USA

Soumya Parameshwaran
Kasturba Medical College, Mangalore, India

Sabrina Pendergrass
University of Virginia, Charlottesville, VA, USA

Keith J. Petrie
Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand

Julie A. Phillips
Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

Therese D. Pigott
Loyola University Chicago, Chicago, IL, USA

Gianfranco Poggi
European University Institute, Fiesole, Italy

Inna Ponomareva
Kuban State University, Krasnodar, Russia

Robert G. Pontius
Clark University, Worcester, MA, USA

Ross L. Prentice
University of Waterloo, Waterloo, ON, Canada

Josef P. Rauschecker
Georgetown University Medical Center, Washington, DC, USA; and Institute for Advanced Study, TU München, München, Germany

John C.W. Rayner
University of Wollongong, Wollongong, NSW, Australia; and University of Newcastle, Callaghan, NSW, Australia

Nancy Reid
University of Toronto, Toronto, ON, Canada

Wolfgang Reinhard
Professor Emeritus of Modern History, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Margund K. Rohr
Institute of Psychogerontology, University of Erlangen-Nuremberg, Nürnberg, Germany; and University of Leipzig, Leipzig, Germany

Hilton L. Root
George Mason University, Arlington, VA, USA

Melissa Root
University of Saint Joseph, West Hartford, CT, USA

Peter M. Ruberton
University of California Riverside, Riverside, CA, USA

Lauretta Rubini
University of Ferrara, Ferrara, Italy

Dominiek Sandra
University of Antwerp, Antwerp, Belgium

Nora C. Schaeffer
University of Wisconsin Survey Center, University of Wisconsin-Madison, Madison, WI, USA

John M. Schaubroeck
The Eli Broad Graduate School of Management, Michigan State University, East Lansing, MI, USA

Nikolas Scherer
Hertie School of Governance, Berlin, Germany

Friedrich Schneider
Johannes Kepler University of Linz, Linz, Austria

Herbert Schriefers
Nijmegen University, Nijmegen, The Netherlands

Natalie Schüz
School of Medicine, University of Tasmania, Hobart TAS, Australia

Christie Napa Scollon
School of Social Sciences, Singapore Management University, Singapore, Singapore

John Scott
University of Copenhagen, Copenhagen, Denmark

Lauren M. Scott
Esri, Redlands, CA, USA

Elisabeth Selkirk
University of Massachusetts, Amherst, MA, USA

Masayoshi Shibatani
Rice University, Houston, TX, USA

Kangping Si
Clark University, Worcester, MA, USA

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany

Judith D. Singer
Harvard University Graduate School of Education, Harvard University, Cambridge, MA, USA

Sergio Sismondo
Queen's University, Kingston, ON, Canada

J.-C. Spender
Kozminski University, Warsaw, Poland

Miriam T. Stark
University of Hawai'i-Mānoa, Honolulu, HI, USA

David Storey
University of Worcester, Worcester, UK

Namita Sugandhi
Binghamton University, Binghamton, NY, USA

Jacqueline Sullivan
TW Ponessa & Associates, Exton, PA, USA

Jayashankar M. Swaminathan
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Elizabeth A. Swart
University of Southern California, Los Angeles, CA, USA

Robert Teghtsoonian
Smith College, Northampton, MA, USA

Oliver Thas
Ghent University, Gent, Belgium; and University of Wollongong, Wollongong, NSW, Australia

Robert Thornton
School of Social Science, University of the Witwatersrand, Johannesburg, South Africa

Colwyn Trevarthen
The University of Edinburgh, Edinburgh, UK

Dejonghe Trudo
University of Leuven (KUL), Leuven, Belgium

Jonathan H. Turner
University of California, Riverside, CA, USA

Michael Useem
University of Pennsylvania, Philadelphia, PA, USA

Jeffrey C. Valentine
University of Louisville, Louisville, KY, USA

Flavia Valtorta
San Raffaele Scientific Institute, University Vita-Salute S. Raffaele, Milano, Italy; and International School of Psychotherapy with Imaginative Procedures (SISPI), Milano, Italy

Wray Vamplew
University of Stirling, Stirling, UK

Omer Van den Bergh
University of Leuven, Leuven, Belgium

Gabriella Vigliocco
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Jaime Vila
Facultad de Psicología, University of Granada, Granada, Spain

Theodore J. Voneida
Northeastern Ohio Universities College of Medicine, Rootstown, OH, USA

Catherine Waldby
The University of Sydney, Sydney, NSW, Australia

John H. Walker
University of Central Florida, Orlando, FL, USA

Barney Warf
University of Kansas, Lawrence, KS, USA

Daniel Welch
The Sustainable Consumption Institute, The University of Manchester, Manchester, UK

Norman E. Whitten
University of Illinois at Urbana-Champaign, IL, USA

John B. Willett
Harvard University Graduate School of Education, Harvard University, Cambridge, MA, USA

Thomas Wolbers
Otto-von-Guericke University Magdeburg, Magdeburg, Germany

James D. Wright
University of Central Florida, Orlando, FL, USA

Alison Wylie
University of Washington, Seattle, WA, USA

Mitchell L. Yell
University of South Carolina, Columbia, SC, USA

Solidarity: History of the Concept

Patrick Cingolani, Bâtiment Olympe de Gouges, University-Paris Diderot, Paris Cédex 13, France

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Procacci, volume 21, pp. 14585–14588, © 2001, Elsevier Ltd.

Abstract

The modern sociological concept of solidarity originated with the French philosopher and politician Pierre Leroux and took on its definitive modern meaning during the French Third Republic as influenced by the progressive thinking of the eminent *homme de gouvernement* Leon Bourgeois. The term refers generically to unity or agreement of outlooks and behaviors among persons of like interests, as in 'working class' solidarity. In sociology, solidarity is understood to be essential in the success of any social movement. In this sense, the concept of solidarity came to be central to the thinking of many social commentators and critics, among them de Maistre, Comte, Proudhon, Marx, Bakunin, and others.

Introduction: Solidarity and Society

The history of *solidarity* is ridden with paradox: it is embedded in the ambivalent ties that bind the members of a community together in active or passive ways. In ancient Rome, the *gens* was held accountable for personal debts contracted by any of its members: it was expected to redeem those who were made prisoners, as well as to pay for the fines of the condemned. A form of solidarity was thus established by law between the person and the *corps* to which she or he belonged. The clan-based structure, with all its sacrificial logics or vendettas, is an institution of solidarity. Traces of this are enshrined in contemporary law: creditors may act 'in solidum' and groups may be jointly and severally liable for one of their parties' debts. Ambivalence also pervades the historic intellectual momentum that the word 'solidarity' was to acquire in the wake of the French Revolution and its successive readings. It resurfaces in the writing of the counterrevolutionary theocrat and traditionalist Joseph de Maistre, in connection with the notions of *solidarity in crime* and compensation for revolutionary executions, and also in the work of the Saint-Simonian philosopher and politician Pierre Leroux, who furthered its recognition as a theory and declared: "I was the first to borrow the notion of solidarity from the lawmaker (…) My wish is to replace Christian charity by Human solidarity" (Leroux, 1863). Focusing on sacrificial rituals, Maistre underscored 'human degradation' as evidence of human unity; Leroux, for his part, elaborated on the human need for friendship, on the grounds that one has only a half-life, "the other being part of humankind" (Maistre, 1854).

There may be doubts as to whether Leroux actually coined the term *socialism*, but one can hardly deny his unique contribution to the modern usage of the neologism *solidarity*. The word was extensively used in one way or another by a number of nineteenth- and twentieth-century social thinkers. It was eventually revisited not only by Fourierists and positivists, with Auguste Comte and Emile Littré, but also by mutualists, with Proudhon; internationalists engaged in a labor movement inspired largely by Proudhon, Marx, or Bakunin; and lastly, by solidarists. It was inscribed, with varying degrees of success, in the framework of the antinomy established by Leroux between the recognition of the social fact and the refusal to sacrifice the individual to the community. Yet it was mostly through the thought and policy of the Republican movement and the institutional project that led up to the creation of the French Third Republic that the concept was eventually to unfold as a sociological category on the public sphere, around Léon Bourgeois' notion of solidarism, which acknowledged our mutual interdependency even as it fully legitimized the individual.

Solidarity: From Archaism to Modernity

The sense of solidarity was essential not only for maintaining unity within the clan but also for ensuring harmony and pacified interaction between different clans. This is borne out by anthropological studies on gift exchange mechanisms, such as the *Potlatch* practiced by the *Kwakiutl* in Northwestern America, or the *Kula*, which involves significant intertribal exchanges throughout the Trobriand Islands on New Guinea's West Coast. Research conducted by Bronislaw Malinowski (1872–1942) and Franz Boas (1858–1942), and interpreted by Marcel Mauss (1872–1950), underscores the disinterested nature of such mechanisms. In cultural and religious frameworks, they formalize pledges and counterpledges to give, to receive, and to give back for the purpose of building and fostering networks of solidarity. What is exchanged is not only useful economically but also consists "above all *of courtesies, entertainment, rituals, military assistance, women, children, dances and feasts*" (Mauss, 1925b). The circularity of giving, by institutionalizing a commitment to solidarity, seems to have provided humankind with its earliest form of pacification. The archaic gesture of giving is visibly perpetuated in major religious traditions and civilizations, including ancient Greece, as described by Aristotle (384–322 BC). No less surprising are the popular practices and traditions observed in contemporary societies, which cultivate a strong relationship to spending as an expression of social interaction.

The unity of humankind is thus underpinned by a number of religions: the comments of Augustine (354–430) on the Old and New Testaments in the *City of God* remind us, for instance,

that all men were naturally brothers through Adam, even before they reached supernatural brotherhood in Christ. In fact, charity and hospitality are underscored in most religious texts, ranging from the holy scripture of ancient India, which offers a number of rules on hospitality and giving, to the book of Leviticus with its well-known commandment to "Love your neighbor as yourself." A sura in the Qu'ran instructs the faithful to "…obey and spend, it is better for your souls." The arguments and principles originally outlined by Aristotle for friendship in the Greek *polis* have permeated the West's political heritage as a whole. Affluent Greeks were subjected to mandatory standards of redistribution of wealth, mainly to support galleys for defense purposes, religious ceremonies, and associated theatrical events. But the conceptualization of such issues by Aristotle was aimed at taking interpersonal friendship to a collective, political level in order to ensure harmony within the community. To lawmakers, he explained, harmony seems to matter more than justice, for it is a determining political factor. Giving and redistribution establish bonds of mutual obligation among citizens, enabling the *polis* to find balance in a middle ground. Aristotle sees the practice of redistribution as a way to counter the extremes that challenge democratic societies, faced with balancing the demands of large impoverished populations with the excessive power of the wealthy few. Further, to the best interests of society, tributes paid to patrons and donors combine with honorable distribution of compensatory goods and services to the poor as symbolic rewards for those who contribute to ensuring economic balance and collective peace.

However, giving as mediation is not only found in the great theological and philosophical treatises on friendship but is also prevalent in our very midst, among the peoples of the world. Hospitality and in particular the feasts that we hold in the framework of alliances, the presents and the costly expenditures surrounding rites of passage, marriage, and birth are embodiments of the selfsame give and take through which humans have managed to engage with each other, generating the sort of precontractual inclination that giving provokes in the other, to create and to foster the bonds of solidarity. In his renowned essay *The Gift* (1925b), Emile Durkheim's nephew Marcel Mauss suggested we "return to the old and elemental," following a philosophical tradition in which solidarity and, more generally, mutual assistance, emerge as crucial wagers of modernity. "To return to the old and elemental" is to reclaim certain ancestral and universal practices of solidarity, and to adapt them to the needs of the rising twentieth century, by means of novel and thoughtful practices and modes of redistribution.

In many ways, however, it was the French Revolution which gave its momentum to the narrative and subsequently scholarly discourse on solidarity per se. To be sure, the French Revolution was viewed by some as a radical crisis or rupture of the social bond. To others, however, it appeared as an opportunity to rethink that selfsame bond in light of the 'in-solidarity' (the French neologism *insolidarité* is found in Fourier and Proudhon) that comes with free trade. Inspired by counter-revolutionary assessments, Auguste Comte (1798–1857) progressively regards the French Revolution as a source of 'Western anarchy.' Others, however, Leroux or Proudhon, for one, welcome it as a step toward a novel sociability, namely, an enhancement of justice. Comte sees historical continuity in the exhortation of a form of intergenerational solidarity that would, as such, run deeper than 'solidarity with contemporaries.' Opting for subordinating the present to the past and progress to order, he criticizes the 'uprising of the individual against the species' and suggests, in the long run, that duties should substitute rights. Sensing the hazards in authoritarian communism or Comte's Sociocracy, Leroux is careful to embed his solidarity in a dialectic between civil society and the State, and a full recognition of the individual. Proudhon, wary of any positivist tendency to 'absorb the individual into the agency,' tries for his part to reconcile, through mutualism, the social property and the workers who contribute to its development. "That is what our solidarity consists in," he declares in his final opus: "we are all clients to each other, servants all" (Proudhon, 1865). Heirs to Saint-Simonian critiques of private property's contingency on society's state of development, they each impugn, in their varied ways, what they see as stumbling blocks to general welfare and collective prosperity in the uneven forms of private property and inheritance. Following Proudhon, Marx, and the larger share of socialist and liberal thinkers, the labor movement denounces private ownership of the means of production and makes unity a major aim of its activism. It had every reason to embrace solidarity, as it needed to overcome divisive market laws and management practices as well as to forge coalitions among trade unions, not only for strikes but also for the broader struggle against capitalism. What emerges at first as *class solidarity* would ultimately lead to the movement's internationalist project and its renowned slogan: 'Workers of the world, unite!'

Solidarity and the Republic

Still, it took more than the labor movement and its theorists, all the more concerned with justice, equality, or social emancipation to bring the concept of *solidarity* to its ultimate consequences. It was more likely the development of *solidarism*, in close connection with the French Third Republic, which provided the needed intellectual stimulus for the concept to thrive under State jurisdiction. It was as though it had to take yet another historic turn in order to prove the full range of its coherence and practical effectiveness. It required a combination of intellectual, institutional, and political factors, which finally materialized at the founding of the Radical Party. To be sure, the Republican Party had directly inherited the sociological theories of positivism, owing mainly to the influence of Comte's disciple Emile Littré (1801–81) and of positivist-minded politicians such as Jules Ferry (1832–93) and Léon Gambetta (1838–82). And yet, it was the Radical Party which finally paved the way for solidarity to form as a concept, namely, around the personality and work of Léon Bourgeois (1851–1925), the theorist of radicalism and promoter of the notion of solidarism. A member of the French Parliament and a senator who served several terms as minster, Bourgeois is closely associated with the doctrine of solidarity. He revisited Comte's theory of continuity, but diverted it to a sort of individual personal development. From the positivist inheritance, he maintained the belief in the contingency of the past: no economic production, no scientific or technical innovation is

possible without the legacies of history. "It has rightly been said that humanity is made of more dead than living; our body, the products of our labor, our language, thoughts, Institutions, art, all are inherited, a treasure slowly accumulated by our ancestors" (Bourgeois, 1896). In this sense, we all have a social debt, or duty, to pay. This does not imply, as in Comte, however, a subjection to the Great Being. Through taxation and institutions that redress social inequalities, individuals repay their debt and recover their freedom. If the individual is obliged to society, however, society is no less responsible to the individual and thus for correcting social discrepancies that result from arbitrary contingencies such as individual birthrights and wealth. Whatever his or her social background, however, "the individual's responsibility is not absolute"; much to the contrary responsibility for the 'individual's debt' falls upon society, by virtue of the unequal birthrights and entitlements that are effective in its midst. It is society's duty to "make the effort" that enables individuals to rise above their conditions (Bourgeois, 1902). Solidarism gives rise to a full-fledged rationale for the legitimization of a range of institutions, from education to social security through progressive taxation, all seeking to settle the outstanding debt by contributing to individual personal development. Even though it respects private propriety, solidarism thus justifies any State intervention that strikes a balance between capital and labor, compensating, by force if need be, for the unfairness that results from capitalist excess. "Collectivization is not a goal of the system of solidarity (…) The doctrine of solidarity is to redeem freedom and individual propriety; it is the very justification for the freedom to own property" (Bourgeois, 1902). Underpinning the major premises and conclusions of the solidarist doctrine is a full range of theories on the subject of solidarity. For, even if Bourgeois is above all a politician and fully engaged in public life, numerous scholars have contributed to developing the concept of solidarity: a constellation of sociologists, economists, and philosophers led by Charles Gide (1847–1932), Emile Durkheim (1858–1917), and Alfred Fouillée (1838–1912).

Charles Gide, a lawyer and professor of social and political economics, at the Universities of Bordeaux and then of Montpellier (France), was a founder and an active member of the cooperative movement. His commitment and teachings contributed significantly to the elaboration, if not to the foundation, of the concept of solidarity. As early as 1890, his book *Quatre Ecoles d'Economie Sociale* laid down the groundwork for a new school of economics based on solidarity, which was further elaborated in his final lectures at the Collège de France, published in 1932. His work puts the concept in a historic perspective and engages a dialogue with various theorists – from Comte to Bourgeois. It is notable, for our purposes, that he draws a parallel between solidarity and charity, explaining, to that effect, that "the one who acts out of solidarity, such as the trade-union leader, the mutualist, the cooperator, may expect (…) in return (…) a higher economic regime in which s/he may find greater independence or security, or a higher standard of living. He would trade off a present good for a possible one; he would sacrifice a bit of his individual self to enhance his social self" (Gide, 1932, p. 199). In contrast, charity expects nothing in return. State and private institutions promote solidarity: for women and mothers, children and infants, childcare, as well as housing for the disadvantaged. Charity, for its part, is likened to a sentiment that is no doubt likewise concerned with justice, but constantly strives to increase it by expanding "the concentric circle of love to the same extent" (Gide, 1932, p. 214).

As for Emile Durkheim, who founded the French School of Sociology, he was deeply committed to the French Republic's political and institutional system. As professor of education and social sciences in Bordeaux, and later sociology at the Sorbonne, Paris, he developed his own theory of solidarity through constant dialogue with contemporaneous scholars. In *The Division of Labor in Society* (1893), he bases his sociological theory on a social philosophy that spans the entire nineteenth century, insisting that the whole is greater than the sum of its parts and that the social should prevail over the individual. Durkheim distinguishes the 'mechanical solidarity' of tribal groups from the 'organic solidarity' of contemporary societies: whereas the former is based on sameness and normatively contingent on the collective consciousness, the latter rests on division of labor, specialization, and recognition of the individual's freedom of opinion: its establishment appears to him to be 'anomic,' or conditioned by provisional circumstances. Several major commentators have noted, however, that he distanced himself from his earlier works and only returned to his theory of solidarity after further rethinking the issues of collective consciousness and moral authority. In fact, it was thanks to a later theory on spending, which was taken up by his nephew Marcel Mauss, that matters of giving and solidarity made their way back into his work. This is particularly evident when Durkheim shifts his focus from industrial to religious issues. Focusing on aboriginal Australian religious practices such as inebriation and collective spending, his *The Elementary Forms of Religious Life* (1912b) places 'the social fact' and its autonomy in a new light: "When there is close interaction and mutual influence among individual consciences, the resulting synthesis gives rise to a novel consciousness. Drawn by the collective mainstream, the individual rises above his/her own self-interest and dedicates him/herself entirely to the common purpose" (Durkheim, 1924, p. 133). Ultimately embodied in religious fervor, the social fact generates, by virtue of its desirability and capacity to overwhelm, such deep-seated feelings of solidarity and brotherhood that the Christian notion of "human equality and fraternity seems to us today to leave too large a place for unjust inequalities" (Durkheim, 1912b, p. 427). However, Durkheim contributed more to backing the notion of solidarity scientifically than to promoting its practical implementation. Like Charles Gide, he correlates and assesses solidarity, justice, and charity. Seeking to provide a scientific account of the conditions of collective generosity in a powerful and novel elaboration of the moral fact, he reaches the conclusion that charity has become, in early twentieth-century societies, an obligation that gives rise to the foundation of public institutions.

Alfred Fouillée, the third theorist, was unique not only for complementing the notion of solidarity with that of 'social property' but also for making a compelling argument on behalf of the specific domain of the State. He recalls the Saint-Simonian emphasis on the evolving nature of ownership and uses it to sound out the law. His aim is to uncover its scientific principles and to give it a practical dimension. Accumulated

productions condition the activities and development of those who are living and partaking in the vast process of transmission, which regulates the conditions for production and life in contemporary societies: "All the books that are read today are still printed by Gutenberg" (Fouillée, 1884, p. 21). In Fouillée, we find a third way between the individualism and socialism, privatization of the public good and collectivism, which is characteristic of radical republicanism. "Since the product of labor is common to the individual and to society, property theoretically contains both an individual and a social part" (Fouillée, 1884, p. v). The State has the right and the duty to intervene in the circulation of materials, goods, information, and people: "roads, postal and telegraphic services, currency – all of which really pertain to the public-service" (Fouillée, 1884, p. 39). The State acts in the name of justice to redress social discrepancies. "It plays a positive and brotherly role, striving to remedy the bad for the good" (Fouillée, 1880). The State must look after insurance plans: life, old age, health, and unemployment. But such plans, like primary education and other Republican Institutions, are not just entitlements. They are also duties with which all citizens must comply. 'Every sort of freedom entails a correlative duty.' In the course of his work, Fouillée provided pioneering insight into the instrumental role of the modern social State in the process of decommodification of social relationships, such as was recently developed by G. Esping-Andersen (Esping-Andersen, 1990). Such State mediation checks the subjugation of people to competitive market constraints, thus ensuring their personal development. For Fouillée, leaving the individual to cope with coalitions, corporations, and holdings is tantamount not only to depriving her or him of the means to defend her or his interests but also to denying him or her freedom from the pressures of the market, discrimination, and cronyism. State intervention aimed at preventing work hazards, ensuring sanitation and mediating social disputes in the workplace, as well as regulating pensions thus contributes to individual autonomy and emancipation from dominating forces. "State intervention in the workings of the economy is justified by law and the common interest (…) there is nothing socialistic about it." As Robert Castel (1933–2013) points out, Fouillée's notion of 'social property' prefigures the emergence of a 'labor society,' as the idea of a statute for workers was beginning to form a veritable social framework for solidarity (Castel, 1995).

Although solidarity was forcefully conceptualized and proclaimed around a French political movement, France was neither the first to implement social policy nor to establish a system of social security. To be sure, the French notion of national public assistance predates Beveridge's *Welfare State*. But the so-called 'social insurance' was implemented in other European countries prior to France. Its 1880 enactment in Bismarck's Germany was, nonetheless, expressly concerned with legitimizing an undemocratic State and instrumentalizing the social arena for political ends. Wiliam Beveridge (1879–1963), whose name is inextricably bound to the Welfare State, had started, for his part, by theorizing the notion of insurance, on the basis of Webb's social reform project. His 1942 proposal targeted the five giants that stood in the way of reconstruction: deprivation, illness, ignorance, lack of hygiene, and idleness (Beveridge, 1942).

Solidarity throughout the World

Solidarists did not just conceptualize and construct a pacifying social State at the national level. Many of their most prominent members, Léon Bourgeois for one, also militated for peace and envisioned a League of Nations that would promote a worldwide legal framework and socioeconomic harmony. Another member of the cooperative movement was Albert Thomas (1878–1932), who contributed to the creation of the International Labor Organization. He delighted in the idea that the League of Nations' Global Compact should foster 'genuine international solidarity.' The solidarists' awareness of interrelationships and attachments enabled them to grasp economic globalization processes as well as sanitation and communication issues. In 1925, Mauss had been alarmed by his visit to Bolshevik Russia and taken stock of a disastrous World War that had decimated Europe's youth. Reflecting on reconciliation and international peace in an era of rising communism and nationalism, he evokes the metaphor of King Arthur's Round Table: "Peoples, classes, families and individuals may become rich, but they will not achieve happiness until they can sit down like the knights around their common wealth. (…) Happiness is to be found in the prevalence of peace, in the rhythm of communal and private labor, in wealth amassed and redistributed, in the mutual respect and reciprocal generosity that education can impart" (Mauss, 1925b).

Twentieth-century communism is, nonetheless, unavoidable as a social movement that placed its full range of experience under the mantle of internationalism. Emblematically, the Soviet Union would identify its national anthem with the 'internationale' composed by French musician and poet Eugene Pottier. That internationalism, beyond its historical mutations, its betrayals in the name of State and, finally, its totalitarian deviations, kindled and constituted the ideal revered by millions of workers and peasants throughout the world. In the 1960s, it acquired renewed momentum through the support extended by Western youth in large urban centers to colonized peoples and the impugnation of imperialist practices in a number of the world's developing countries, notably in Africa, South East Asia, and Latin America. Outside from its institutional, republican, acceptations, the word solidarity had never been as visible as it was in the internationalist dimension of the 1960s. Far from any sentiment of compassion or charity, what emerged from this vision of unfairness and exploitation of 'Third World' resources and workforces was clearly a political sense of solidarity with the oppressed peoples of the world. This would eventually revive a certain degree of sympathy for the cause of compatriot laborers in metropolitan centers. The peoples of Asia, Africa, or Latin America, while struggling for their independence, had also joined the assembly lines of Europe and North America.

Since the nineteenth century, the promotion of solidarity has worked in a tension between an institutional pole, exemplified by solidarism, and an activist pole, which, albeit quite diversified, is somewhat emblematized by twentieth-century Marxism. This is still the case today, as solidarity is once again shored up by social movements. Solidarity has been reclaimed by various international social forums in their struggle against market hegemony and the alienating effects of the plutocratic forms of corruption challenging democracy today. But it is even

more spontaneous in the various collective demonstrations against financial speculation or the effects of the subprime crisis. A shared sense of indignation has emerged in the wake of the collective drama that has impoverished or indebted segments of the labor and middle classes. Pamphlets and books by *indignados* in Spain or actors of Occupy Wall Street in North America likewise proclaim the collective value of solidarity. "We are the 99" does not just reflect the increasingly gaping inequality between the large majority and the affluent few; it is also translates a common experience and a solidarity of experiences. The various 'occupations' have worked simultaneously as ways to participate transparently in public life and to live together, in solidarity, rejecting the constraints of money and domination. They are demonstrations of goodwill and solidarity in response to levels of exploitation and greed that were unparalleled in the 1960s. The Spanish pamphlet ¡*Democracia Real Ya!* stresses that "equality, progress, solidarity, free access to culture, environmental sustainability and development as well as people's welfare and happiness are the priories of every advanced society." Those who witnessed 'Occupy Wall Street' saw hundreds of people coming together and making decisions with no management structure, "driven only by principles and solidarity."

We will have to wait and see whether these trends are part of a sustainable dynamic and not just transitory jolts and short-lived rebellions. But chances are the globalized economy will put the ideal of solidarity in the political spotlight again, this time on an international scale. The dynamic of international exchanges tends to nourish a feeling that we are all part of a single world. It has intensified cooperation, bonded people together, and encouraged political rapprochement through solidarity against those who unduly profit from informational asymmetry and violate national and international laws and regulations. In addition, the lurking threats of climate change, evidence of our collective involvement in a range of ecological mechanisms with imminent dangers and planetary consequences, can only be checked through competent institutions operating on a world scale and in a spirit of solidarity. The conflict dynamic that is stirred by confiscation of knowledge and discoveries, or the abuse of public domains and collective property for the sole benefit of a privileged few, also continues to act as a lever for collective solidarity. The risk of ecological disaster summons the idea of humanity – its common history and fundamental responsibility to future generations – to the forefront of today's collective dynamics. One can only hope that this can result in an institutional framework for solidarity, to coordinate its endeavors, above and beyond particular aspirations and projects, denouncing and correcting persisting inequities among and within various segments of the world's population.

See also: Anthropology, History of; Attempted Utopias and Intentional Communities; Division of Labor: History of the Concept; Gift: History of the Concept; Interest, Psychology of; Interest: History of the Concept; Interests in Sociological Analysis; Intergenerational Solidarity; Labor, Division of; Literary and Philosophical Utopian Thought; Sociology, History of; Solidarity, Sociology of; State: Anthropological Aspects; Working Classes, History of.

Bibliography

Beveridge, W., 1942. Social Insurance and Allied Services. HMSO, London.
Boas, F., 1897. The Social Organization and the Secret Societies of the Kwakiutl Indians. Smithsonian Institution, US National Museum, Washington.
Bourgeois, L., 1902. Essai d'une philosophie de la solidarité (Conférences et discussions). Alcan, Paris.
Bourgeois, L., 1896. La solidarité, Paris. Armand Colin, Paris.
Castel, R., 1995. Les métamorphoses de la question sociale. Fayard.
Cingolani, P., 2007. Sociologie et propriété sociale. Corpus: Alfred Fouillée, n 53.
Cingolani, P., 2003. La république, les sociologues et la question politique. La Dispute.
Comte, A., 1851–1854. Système de politique positive ou traité de sociologie instituant la Religion de l'Humanité, 4 vols. Mathias, Paris.
Durkheim, E., 1893. De la division du travail social. Alcan, Paris.
Durkheim, E., 1924. Sociologie et philosophie. Alcan, Paris.
Durkheim, E., 1997. The Division of Labor in Society (L.A. Coser, Trans.). Free Press, New York.
Durkheim, E., 1912a. Les formes élémentaires de la vie religieuse. Alcan, Paris.
Durkheim, E., 1912b. The Elementary Forms of Religious Life (J.W. Swain, Trans.). George Allen and Unwin Ltd, London, 1964.
Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton.
Fouillée, A., 1884. La propriété sociale et la démocratie. Alcan, Paris.
Fouillée, A., 1880. La science sociale contemporaine, second éd. Hachette, Paris. 1885.
Gide, Ch, 1932. La solidarité – cours au Collège de France 1927–1928. Presses Universitaires de France, Paris.
Gide, Ch, 1890. L'école nouvelle. In: Quatre écoles d'économie sociale – Conférences, Genève, Stapelmohr. Fischbacher, Paris.
Leroux, P., tome 1, 1863. La grève de Samarez. Dentu, Paris.
Maistre, J. de, 1854. Soirées de Saint-Pétersbourg ou entretien sur le gouvernement temporel de la providence; suivies du Traité sur les sacrifices. J.B. Pélagaud et Cie, Lyon et Paris, 2 Tome 1 et 2. (1821).
Malinowski, B., 1922. Argonauts of the Western Pacific – an Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea.
Marx, K., Engels, F., February 1848. Manifesto of the Communist Party. London.
Mauss, M., 1925a. Essai sur le don – Forme et raison de l'échange dans les sociétés archaïques. In: Alcan, Félix (Ed.), Année sociologique, seconde série, 1923–1924. Paris.
Mauss, M., 1925b. The Gift. (I. Cunnison, Trans.). Cohen & West LTD, London, 1966.
Proudhon, P.J., 1865. De la capacité politique des classes ouvrières. Dentu, Paris.

Solidarity, Sociology of

Michael Hechter, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The discipline of sociology has taken the attainment of solidarity – in families, groups, organizations, and societies – as a principal outcome to be explained. This article compares normative, structural, and rational choice accounts of solidarity and discusses the likely paths of future research on this topic.

If economics is mostly about efficiency and growth, sociology is mostly about solidarity and social order. As economics suggests, competitive markets offer the best chance for an efficient allocation of productive resources. As sociology suggests, however, such markets can only arise in societies with sufficient order to produce and maintain stable property rights and other key public goods. Although both stability and growth are necessary to sustain any social formation over the long run, the disciplines of economics and sociology have largely developed in mutual ignorance, if not antipathy. This article suggests that progress has been made in the analysis of solidarity by integrating the two perspectives.

Solidarity – the binding of individuals into a cooperative collectivity on the basis of compliance with normative obligations – is an emergent attribute of groups that facilitates collective action and social order. The modern origins of the concept date from the beginnings of sociological theory. There are three leading theoretical perspectives on the emergence of solidarity. On the *normativist* view, solidarity is most likely to develop among individuals who share fundamental values, such as those promulgated in religious, ethnic, or national groups. On the *structuralist* view, solidarity arises from the sharing of common material interests as occurs, for example, among the members of social classes, ethnic groups, and genders. On the *rational choice* view, it owes to neither of these kinds of conditions; rather, it is a function of dependence and control mechanisms.

Since social order and collective action are the alpha and omega of classical sociological theory, solidarity has always been a central concept in sociology. The concept came to the fore following the changes wrought by the development of industry and the rise of market economies. Although they were unjustified in doing so, most classical social theorists regarded social order in agrarian communities as unproblematic. In their stylized view, such communities were technologically and demographically stagnant and uninvolved in long-distance trade. Given these assumptions and barring exceptional circumstances – such as those resulting from exogenous events such as wars or massive epidemics – social mobility would be minimal and most children would be destined to recapitulate the social lives and fates of their parents. As agrarian communities tended to offer relatively little scope for individualism, the attainment of solidarity was viewed as relatively unproblematic. According to Ferdinand Tönnies (1887), such communities (which he termed *gemeinschaften*) were breeding grounds for social relations based on deep emotional, quasifamilial commitments. Emile Durkheim (1893) referred to the solidarity characteristic of this kind of society as 'mechanical,' implying thereby that it had a certain automatic quality. Mechanical solidarity rests on a set of common values, internalized norms, and beliefs (a *conscience collective*) engendered by physical co-presence, a common focus of attention, and a sense of mutual awareness. This kind of solidarity is often expressed in a religious credo (Durkheim, 1912).

Mechanical solidarity appeared to be threatened, however, by the rise of industry and the expansion of market economies in Western Europe, North America, Australasia, and Japan beginning in the late eighteenth century. Among their many other effects, these transformations increased the size and scope of social networks, thereby affording many individuals greater options, ranging from marriage partners to occupations, in their daily lives. The growth of individualism and the concomitant decline of a *conscience collective* was the inevitable result (Simmel, 1922). It was far from evident how groups, communities, and societies could maintain mechanical solidarity in the wake of a burgeoning division of labor and its by-product, that is increasing individualism. Durkheim, the foremost early theorist of solidarity, initially argued that social order in industrial societies was not threatened by the division of labor; on the contrary, individuals in industrial society were held together by their mutual interdependence, a form of solidarity that he termed 'organic.' At the end of his life, however, he came to regard common values and norms (which he sometimes referred to as "the non-contractual basis of contract") as the basis for solidarity in all kinds of groups and societies (Durkheim, 1912).

Durkheim's insistence that solidarity requires compliance to a set of collective obligations based on particular values and norms has greatly influenced subsequent research. For example, a religious group is solidary only to the degree that its adherents comply fully with its obligations, both sacred (rites) and secular (tithing). Likewise, a working class is solidary only to the degree that individual workers regard themselves as members of a collectivity with its own obligations and refrain from engaging in activities such as strikebreaking that compromise or countermand these obligations.

Yet compliance to collective obligations can owe to both pecuniary and nonpecuniary causes. Consider the mobilization of labor in a profit-making firm. To attain maximum profit, firms demand that their employees perform to the best of their ability and refrain from shirking. Employees usually comply

with the firm's obligations not out of a sense of group solidarity but because they are individually compensated for doing so by receiving wages or salaries. In contrast, the members of a religious group or a working class are not given pecuniary compensation to comply with their corporate obligations. Hence, compliance in such groups must be due to other mechanisms.

Theoretical perspectives on solidarity may usefully be divided into three distinct types: *normative* perspectives suggest that solidarity arises from common values and norms, *structural* perspectives argue that it arises from common material interests, and *rational choice* perspectives hold that it arises from mutual interdependence.

The Normative Perspective

Normative theorists claim that solidarity is unlikely to result merely from the interaction of rational, self-regarding individuals (Parsons, 1951). This is because it is never in the interest of self-regarding individuals to adhere to collective obligations if they can profit by ignoring them. In a world of self-regarding agents, the only means of assuring the requisite compliance is by sanctioning. Normative theorists, however, hold that sanctioning is both too costly and unreliable to serve as a means of assuring high levels of compliance, especially among large numbers of individuals. In some cases, sanctioning breeds resistance. For normative theorists, the solution lies in values and internalized norms. These make individuals fit for social life by removing the conflict between the individual's interest and that of the relevant collectivity. If individuals willingly adhere to the values and norms that underlie corporate obligations, then they will act to promote collective ends without the need for carrots and sticks. According to normative theorists, socialization is the key to the attainment of solidarity. Solidarity therefore is most likely to emerge in socially homogeneous collectivities. It should be noted that the emergence of values and norms is exogenous to the theory, as are socialization processes. Moreover, although casual observation confirms that compliance with values and norms is never automatic, normative theory has little to say about the conditions under which the requisite compliance is likely to occur.

The Structural Perspective

Structural theorists typically, if implicitly, regard solidarity as an outcome of individually self-regarding action. In their view, individuals form solidary collectivities principally by virtue of sharing common (often material) interests. They discover that they share interests in the course of mutual interaction, and especially when they perceive themselves to be threatened by powerful antagonists (Marx, 1978 [1845–46], p. 179; Coser, 1956; Collins and Sanderson, 2009). When mutual interaction occurs repeatedly, this may set the stage for the development of sentiments, or affective feelings, that further strengthen compliance to collective obligations. However, the weaknesses of structural theories of solidarity are notable. Both commonality of interest and repeated interaction are insufficient causes of solidarity. To see why, note that solidarity varies widely among workers and married couples. Sometimes workers refuse to support striking colleagues, and the incidence of divorce has risen dramatically in nearly all industrial societies. Neither of these trends is easily accounted for by the explanatory apparatus of structural theories. Furthermore, if the individuals who share interests are self-regarding, as most structural theorists contend, then why would they ever join forces to pursue their common ends (a point that was already made by Parsons in 1937)?

The Rational Choice Perspective

This question is the point of departure for rational choice perspectives of solidarity, which attempt to reconcile sociological and economic perspectives. Truly self-regarding actors will *not* join a group to pursue common ends when, without participating, they can reap the benefit of other people's activity in obtaining them. Instead of contributing to the common weal, rational actors will free ride (Buchanan 1965; Olson, 1965) and behave opportunistically when it is expedient. The upshot is minimal solidarity despite an extensive commonality of interest. The rational choice theory of solidarity (Hechter, 1987) hinges on the group's capacity to preclude free riding.

In this theory, rational individuals form groups to obtain goods that they either cannot produce on their own or cannot attain as economically. The ultimate motivation for group formation is held to be the individual's desire to gain access to jointly produced goods. These goods can only be produced if prospective members comply with the rules and obligations that are necessary to ensure production of the given joint goods. The more dependent individuals are on the group for access to valued goods, the greater is its potential power over them. This dependence increases when the given goods are not readily available elsewhere, when members lack information about alternatives, when the costs of leaving the group are high, and when personal ties to other members are strong. Moreover, the greater the dependence of members on a group, the greater their willingness to comply with highly extensive obligations. Hence, dependence creates conditions not just for the emergence of corporate obligations, but for obligations that guide and regulate behavior to a high degree.

Yet dependence is an insufficient cause of compliance. Children tend to be highly dependent on their parents, but this is no guarantee that they will routinely perform their chores or study diligently in school. Dependence by itself is unlikely to resolve the free rider problem. Free riding can only be precluded when the given group has a high control capacity. This control capacity, in turn, is a function of monitoring and sanctioning. Monitoring is the process of detecting noncompliance to corporate obligations; sanctioning is the use of rewards and punishments to induce compliance. The theory therefore holds that compliance to corporate obligations – an indirect measure of solidarity in noncompensatory groups – is jointly a function of dependence and control mechanisms. That said, this theory remains incomplete because the sources of individual dependence are exogenous to it.

Since it can be attained as a result of compensation as well as normative obligation, the study of solidarity poses serious challenges to empirical researchers.

After a long period of dormancy, interest in solidarity has increased in several social science disciplines. The rational choice theory of solidarity has been used to explain variations in legislative party voting (Hechter, 1987), the membership rules of friendly societies and mutual benefit associations (Hechter, 1987), the survival of US intentional communities (Hall, 1988), cross-national crime rates (Hechter and Kanazawa, 1993), social order in Japan and the United States (Miller and Kanazawa, 2000), the prevalence of nationalism in history (Hechter, 2000), the survival of religions demanding high sacrifice from their adherents (Iannaconne, 1994; McBride, 2007), the causes of suicide bombing (Berman and Laitin, 2008), and the survival of individuals in prisoner of war camps during the American Civil War (Costa and Kahn, 2008). Whether group solidarity incites or inhibits intergroup conflict is now the subject of debate (Bhavnani and Backer, 2000; Fearon and Laitin, 1996; Gould, 1999). Legal theorists (Ellickson, 1991; Posner, 2000) have explored group solidarity as an institutional alternative to law as a means of dispute resolution. A variety of mathematical models of solidarity have been proposed by social and physical scientists (Doreian and Fararo, 1998). Finally, social psychologists have recently found that solidarity is more likely to develop among individuals who are engaged in intergroup conflict (Benard, 2012), on the one hand, and generalized as against reciprocal exchange networks, on the other (Willer et al., 2012).

Whereas the rational choice theory of solidarity emphasizes individual compliance with collective obligations, group solidarity evidently also entails the emergence of collective sentiments. The pioneering rational choice sociologist James Coleman (1990) argued that positive sanctions produced within groups may yield zealotry among some members. Other writers have sought to downplay the self-regarding behavioral assumptions of rational choice theory by stressing the role that emotions play in generating these sentiments of solidarity. On this view, Collins (2004) suggests that solidary groups are held together not so much by common interests as by positive emotional energy generated in the course of ritual interaction (Collins, 2004); this emotional energy is held to arise to the degree that people share a clarity of expectations (Turner, 2007). For Lawler et al. (2009), however, people form affective ties to groups if and when they attribute their individual feelings to membership in a group. They suggest that emotional energy derives not from ritual interaction, but from affiliation with given groups. Although an emphasis on the emotional basis of group solidarity is often regarded as an alternative to the instrumentalism of rational choice theory, it is unclear why positive emotional energy, whatever its source, cannot be modeled as a part – even quite an important part – of an individual's utility argument.

The empirical study of solidarity has been hamstrung by a lack of agreement on definitions of the concept. Defining solidarity as an intersubjective state (as normative theorists have typically done) posed a severe challenge for empirical research because there were no consensually accepted measures of individual values and internalized norms in social science. Direct measures of the sentiments or feelings that may hold within collectivities perforce were hard to come by.

Future Research

Future research on solidarity is likely to proceed along both methodological and substantive fronts. Methodologically, new measurement instruments have been developed that make it possible to gauge solidarity directly as an intersubjective state, rather than merely using behavioral compliance as a proxy. From Buchanan (1965) and Olson (1965) onward, rational choice models of solidarity have assumed that agents are purely self-regarding. Recent research findings challenge the adequacy of this assumption. For example, experimental games have been deployed in order to elicit variable individual motivations to contribute to public goods (Fehr and Gintis, 2007). These studies have demonstrated that individual motives, as revealed by strategies employed in playing public goods games, are mixed. Whereas some subjects indeed do consistently exhibit self-regarding motives, others seem to be more akin to the pro-social agents that are celebrated in classical sociological theory.

Substantively, there is much to learn about the interaction of solidary groups in wider social systems. Evidently, individual motivations are strongly affected by social contexts. Thus to the degree that small-scale societies engage in cooperative production, for example, self-regarding motivations are attenuated (Henrich et al., 2004). Likewise, individual participants in groups that employ generalized exchange identify with their groups – and appear to be more other-regarding – than members of groups that rely exclusively on direct exchange (Willer et al., 2012).

Since solidarity is built around a set of normative obligations and enforcement mechanisms, it follows that the normative content of any given solidary group may be more or less congruent with that of its global social environment. Thus, minority ethnic groups often espouse norms at variance with their host society. Adolescent peer groups evolve norms that are mutually inconsistent and generally inconsistent with school norms. Factory workers can embrace output restriction norms that are at variance with their employers' expectations, and inner city gangs and organized crime groups establish norms that are at variance with those of the society at large. The effects of solidarity in such countercultural groups are a matter of dispute among both social scientists and political decision makers (*see* Social Capital). Further research is needed to clarify the mechanisms that are responsible for the emergence of such countercultural groups and the consequences these groups have for the attainment of social order.

Along with sustained economic growth, solidarity is one of the two principal desiderata of polities and their rulers, not to speak of less encompassing collectivities. Typically, these two goals are considered to be mutually inconsistent: the policies that promote one ostensibly do so at the expense of the other. No doubt, this inconsistency has led economists and sociologists to focus their attention on different kinds of outcomes. This article suggests that much can be gained about the causes of solidarity, as well as economic growth, by combining the insights of each of these disciplines.

See also: Collective Action; Communities of Practice; Community Sociology; Community Studies: Anthropological; Community, Social Context of: The US Case; Durkheim, Emile (1858–1917); Integration, Social; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Societies, Types of; Sociological Theory.

Bibliography

Benard, S., 2012. Cohesion from conflict: does intragroup conflict motivate intragroup norm enforcement and support for centralized leadership? Social Psychology Quarterly 75 (2), 107–130.

Berman, E., Laitin, D.D., 2008. Religion, terrorism and public goods: testing the club model. Journal of Public Economics 92 (10–11), 1942–1967.

Bhavnani, R., Backer, D., 2000. Localized ethnic conflict and genocide: accounting for differences in Rwanda and Burundi. Journal of Conflict Resolution 44, 283–306.

Buchanan, J.M., 1965. An economic theory of clubs. Economica 32 (125), 1–14.

Coleman, J.S., 1990. Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge, MA.

Collins, R., 2004. Interaction Ritual Chains. Princeton University Press, Princeton, NJ.

Collins, R., Sanderson, S.K., 2009. Conflict Sociology: A Sociological Classic Updated. Paradigm Publishers, Boulder, CO.

Coser, L., 1956. The Functions of Social Conflict. Free Press, New York.

Costa, D.L., Kahn, M.E., 2008. Heroes & Cowards: The Social Face of War. Princeton University Press, Princeton, NJ.

Doreian, P., Fararo, T.J. (Eds.), 1998. The Problem of Solidarity: Theories and Models. Gordon and Breach, Amsterdam.

Durkheim, E., 1893. The Division of Labor in Society. Free Press, Glencoe, IL.

Durkheim, E., 1912. The Elementary Forms of Religious Life.

Durkheim, E., 1995. The Elementary Forms of Religious Life. Free Press, New York.

Ellickson, R.C., 1991. Order Without Law: How Neighbors Settle Disputes. Harvard University Press, Cambridge, MA.

Fearon, J.D., Laitin, D.D., 1996. Explaining interethnic cooperation. American Political Science Review 90, 715–735.

Fehr, Ernst, Gintis, Herbert, 2007. Human motivation and social cooperation: experimental and analytical foundations. Annual Review of Sociology 33 (1), 43–64.

Gould, R.V., 1999. Collective violence and group solidarity: evidence from a feuding society. American Sociological Review 64, 356–380.

Hall, J.R., 1988. Social-organization and pathways of commitment – types of communal groups, rational choice theory and the Kanter thesis. American Sociological Review 53 (5), 679–692.

Hechter, M., 1987. Principles of Group Solidarity. University of California Press, Berkeley, CA.

Hechter, M., 2000. Containing Nationalism. Oxford University Press, Oxford, UK.

Hechter, M., Kanazawa, S., 1993. Group solidarity and social order in Japan. Journal of Theoretical Politics 3, 455–493.

Henrich, J.P., Bowles, S., Boyd, R., Camerer, C., 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies. Oxford University Press, Oxford New York.

Iannaccone, L.R., 1994. Why strict churches are strong. American Journal of Sociology 99 (5), 1180–1211.

Lawler, E.J., Thye, S.R., Yoon, J., 2009. Social Commitments in a Depersonalized World. Russell Sage Foundation, New York.

Marx, K., 1978. The German ideology. In: Tucker, R. (Ed.), The Marx–Engels Reader, second ed. W. W. Norton, New York, pp. 146–200.

McBride, M., 2007. Club Mormon: Free-riders, monitoring, and exclusion in the LDS church. Rationality and Society 19 (4), 395–424.

Miller, A.S., Kanazawa, S., 2000. Order by Accident: The Origins and Consequences of Conformity in Contemporary Japan. Westview Press, Boulder, CO.

Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.

Parsons, T., 1937. The Structure of Social Action. McGraw-Hill, New York.

Parsons, T., 1951. The Social System. Free Press, Glencoe, IL.

Posner, E.A., 2000. Law and Social Norms. Harvard University Press, Cambridge, MA.

Simmel, G., 1922. The web of group affiliations. In: Simmel, G. (Ed.), Conflict and the Web of Group Affiliations. Free Press, New York, pp. 125–195.

Tönnies, F., 1887. Community and Society. Transaction Books, New Brunswick, NJ.

Turner, Jonathan H., 2007. Human Emotions: A Sociological Theory. Routledge, London; New York.

Willer, R., Flynn, F.J., Zak, S., 2012. Structure, identity, and solidarity: a comparative field study of generalized and direct exchange. Administrative Science Quarterly 57 (1), 119–155.

Somatoform Disorders

Jörg Angenendt, University Medical Center, University Hospital Freiburg, Freiburg, Germany
Martin Härter, Center for Psychosocial Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Somatoform disorders, or somatic symptom disorder, as it is denominated in *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, are a heterogeneous group of mental health problems that are characterized by enduring bodily complaints and symptoms that are not due to organic dysfunction or disease. These patients perceive a wide range of severe symptoms like pain, gastrointestinal, cardiovascular, sexual, and/or pseudoneurological symptoms, which cause permanent attention, worry, and distress. They excessively seek medical help and reassurance, but have difficulties accepting the non-pathological results in medical examinations. The concept of psychological and psychosocial influences and causes is rejected, and mental health services are usually not attended. Biological, psychological, and psychosocial factors interact as precipitating, triggering, and maintaining factors of psychopathology. The different subtypes (somatization, conversion, pain disorders, hypochondriasis, and dysmorphophobia) are described with their central features: clinical presentation, epidemiology, comorbidity, and course. Current etiological and pathogenic models are introduced as a clinical and theoretical basis for therapeutic interventions. Besides general psychotherapeutic principles more specific interventions addressing the modification of subjective health-beliefs, illness behaviors, and coping skills are described. The clinical significance of somatoform disorders stresses the need for training programs to provide and improve basic diagnostic and therapeutic skills, especially in primary care. Future directions for the management of these disorders are outlined.

Introduction

The history of somatization phenomena, the 'borderland between medicine and psychiatry' (Lipowski, 1988), or in terms of neuroscience the problem of the 'body–brain-interface' (Guggenheim, 2000), can be traced back to ancient Greek medicine. Here the term 'hysteria' was used to describe patients' somatic complaints without the presence of somatic disease. Hysteria (from hystera = womb) was understood as a psychological problem in women, whose intrapsychic conflicts, resulting from the unfulfilled wish to bear a child, lead to somatic complaints (*see* Hysteria: History and Critiques). Similar clinical observations by different European clinicians since the mid-nineteenth century (Briquet, Charcot, Janet, Breuer) were antecedents of the Freudian concept of conversion disorder. The psychodynamic etiology and psychoanalytic treatment approach to patients with conversion disorders, 'functional' or psychosomatic complaints, and hysteria was most influential to the 1970s. Since 1980 the term somatoform disorders had replaced the traditionally used, but not precisely defined above mentioned terms of hysteria, that were mainly confounded with etiological (psychoanalytical) assumptions. Currently in *Diagnostic and Statistical Manual of Mental Disorders* (fifth edition) (*DSM-V*; APA, 2013) the term was changed into 'somatic symptom disorder'. The concept of somatoform disorders encompasses a heterogeneous group of psychiatric disorders that are characterized by three core features: (1) patients suffer from enduring or recurrent bodily complaints, symptoms and/or pain (e.g., palpitations, headache, fatigue, and dizziness), which are experienced as indicators of somatic illness, (2) psychological and social factors – usually refused as major influencing factors by the patients – play a predominant role in the development and course of these disorders, and (3) patients experience a general preoccupation with bodily symptoms and health concerns. In *DSM-IV* and the tenth edition of *International Classification of Diseases* (ICD-10) the requirement, that somatic symptoms must be 'medically unexplained' was a central diagnostic prerequisite, which is eliminated in the current DSM version.

This article describes the various subtypes of somatoform disorders, their prevalence, and etiology. Critical aspects regarding diagnosis, treatment, and management of somatoform disorders in healthcare systems will be discussed.

Classification and Epidemiology

From a nosological perspective, somatoform disorders were grouped together in *DSM-III* in 1980 for the first time. However, this clustering was primarily based on clinical criteria and not on empirically derived or laboratory findings. The *ICD-10* includes most of the somatoform disorders of the *DSM-IV* (American Psychiatric Association, 1994), except for conversion disorders that are grouped with the dissociative disorders. The main feature of these disorders is the "repeated presentation of physical symptoms with persistent request for medical investigations, in spite of repeated negative findings and reassurances by doctors that these symptoms are not primarily due to somatic illness. If somatic disorders are present, they do not explain the nature and the extent of the symptoms or the distress and preoccupations of the patient" (WHO, 1991). For the diagnosis of all somatoform disorders the patients' symptoms have to be of sufficient severity to warrant clinical attention. The *ICD-10* differentiates between four main nosological entities (seven different diagnoses) which were characterized by different diagnostic criteria (see **Table 1**). Recurrent, multiple somatic complaints that are not associated with any physical disorder are the essential feature of somatization disorder. The diagnosis requires a history of many bodily symptoms of several years' duration with onset

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27069-4

Table 1 Somatoform disorders and their characteristics

Subtypes	*Characteristics and (selected) main criteria (*ICD-10*)*
Somatization disorder	● A history of at least 2 years of multiple and variable complaints which cannot be better explained by somatic factors or other psychiatric disorders (A) ● Preoccupation with the symptoms causes persistent distress and leads to seek repeated medical consultations (B) ● Persistent refusal to accept medical reassurance that there is no adequate physical cause (C) ● A total of six or more enduring autonomous symptoms from different groups (gastrointestinal, cardiovascular, genitourinary system, skin and pain symptoms) (D)
Hypochondriasis	● A persistent belief of the presence of a serious physical disease (A1) ● A persistent preoccupation with a presumed deformity or disfigurement (body dysmorphic disorder) (A2)
Somatoform autonomic dysfunction	● Symptoms of autonomic arousal attributed by the patient to a physical disorder (heart and cardiovascular system/upper and lower gastrointestinal tract/respiratory or genitourinary system) (A) ● Two or more autonomic symptoms like palpitations, sweating, dry mouth, flushing or blushing, epigastric discomfort etc. (B)
Persistent somatoform pain disorder	● Persistence of severe and distressing pain for at least 6 months which cannot be explained adequately by a physiological process or physical disorder (A)

before the age of 30. In somatoform autonomic dysfunctions symptoms of autonomic arousal (e.g., palpitations, sweating, and stomach discomfort) are the main complaints of the patients. Chronic pain syndromes (duration >6 months) like tension headache, lower back pain, and atypical facial pain are labeled as persistent somatoform pain disorder. Excessive and dispropriate thoughts, feelings, and behaviors can play either a major role within a general medical condition (e.g., rheumatoid arthritis) or may be the only determinants of pain (e.g., not otherwise specified lower back pain). The pain in one or more anatomical sites is the main focus of clinical presentation.

Conversion disorder is mostly a transient disturbance of physical functions (typically occurring in stressful situations) which do not conform to current concepts of the anatomy and physiology of the nervous system. Many disorders seem to be neurological disorders (e.g., motor disorders or disturbances of sensory function), but do not show the usual pathological signs. Hypochondriasis is not exclusively a secondary phenomenon within affective, anxiety, or psychotic disorders but a diagnostic subgroup of its own. Patients are preoccupied with their belief of already suffering from a severe illness or the fear of becoming ill with a life-threatening disease (e.g., cancer, stroke, or AIDS). The diagnostic congregation of body dysmorphic disorder – which is rarely seen as a full-blown syndrome – is different in the nosological systems. In the *ICD-10* it is a subgroup of hypochondriasis, with the preoccupation on a presumed deformity or body abnormality. Patients subjectively feel unattractive, ugly, and disfigured despite lacking evidence for major defects in appearance or stigmata. Positive feedback from significant others about normality in appearance has no corrective effects.

Somatoform disorders show high prevalence rates in the community, especially in the primary health care system. In industrialized societies up to 4–19% in the general population suffer from somatoform disorders (Hiller et al., 2006; Simon and von Korff, 1991). In medical outpatient and inpatient treatment, patients with the probable diagnosis of somatoform disorders are estimated at up to 25–30%. The onset is in adolescence or young adulthood, and the course in untreated cases is usually chronic with a waxing and waning of symptoms. Women, persons from lower social classes, and ethnic subgroups are more likely to develop somatoform disorder. Furthermore, somatization is more often seen after separation and divorce. Somatoform disorders are often associated with other psychiatric disorders, especially depression, anxiety disorders, substance abuse disorders, and personality disorders (*see* Depression; Anxiety and Anxiety Disorders; Personality Disorders). Patients with somatoform disorders are often not correctly recognized in primary care and therefore insufficiently treated. The latency between the onset of the disorder, the appropriate diagnosis, and the referral to adequate treatment is usually high. These factors lead to largely extended costs for medical and psychosocial services (14-fold costs of care than average costs of outpatients and sixfold costs in inpatients) (Barsky et al., 2005).

Etiology and Course

No single factor can explain the pathogenesis of somatoform disorders. Comparable with other psychiatric disorders, biological, psychological, and social factors interact in complex ways (*see* Mental Illness, Etiology of). Their impact probably varies over time within the same individual and between different patients resulting in a common final pathway. There is no general etiological model for somatization phenomena (Rief and Broadbent, 2007). However, interesting neurobiological studies demonstrate that patients with somatization disorder may have an abnormality in cortical and neuropsychological functioning, e.g., bifrontal impairment of the hemispheres, abnormal auditory-evoked potentials, and abnormal levels of cortisol (Rief et al., 2010; Tak and Rosmalen, 2010). Several etiological conditions play a prominent role, especially in maintaining the process and can be considered as empirically derived risk factors (see **Figure 1**).

1. Genetic risks, e.g., higher rates of somatoform disorders for monozygot than for dizygot twins, and in first-degree relatives.
2. Higher physiological reactivity which can lead to pervasive interoception of physiological signs. Unspecific bodily

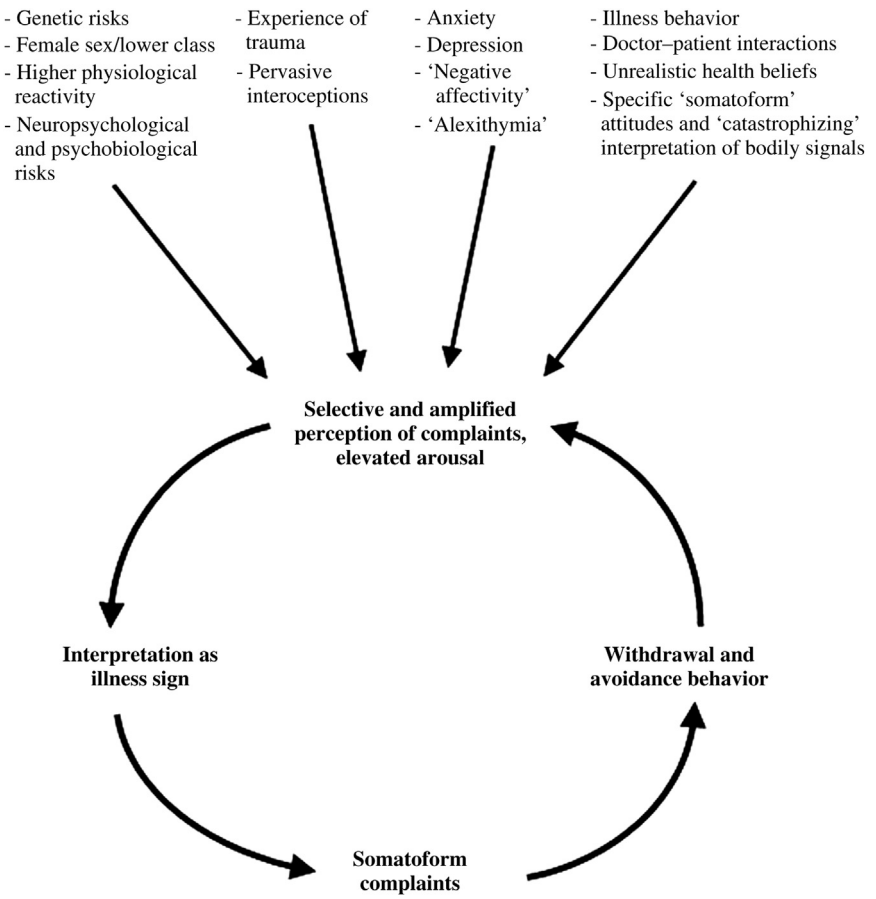

Figure 1 Risk model for somatoform disorders. Adapted from Rief, W., Hiller, W., 1998. Somatisierungsstörung und Hypochondrie (Somatization Disorder amd Hypochondriasis). Hogrefe, Göttingen.

sensations are perceived and interpreted as symptoms of disease. The consequence is the intensification of the complaints and elevated attention to them.

3. Biographic vulnerability, a high proportion of physical, sexual abuse (within the family), and traumatic experiences like war and natural catastrophes are overrepresented in these patients. Furthermore, chronically ill parents or siblings, and learning by model within the family may influence the localization and types of somatoform symptoms.
4. Personality traits, e.g., difficulties in the perception and expression of emotions (alexithymia) with an inverse relationship of reduced emotional expressions and heightened psychophysiological arousal in stressful situations.
5. Unrealistic health beliefs with a low tolerance for normal complaints, 'catastrophizing' ideas about physiological processes, and exaggerated expectations regarding effects of modern medicine to all kinds of complaints.
6. Chronification factors like multiple diagnostic procedures, the overvaluation of normal variants through the caring physician (iatrogen factors) and positive (e.g., attention from others) or negative reinforcement (e.g., defense mechanism for negative emotions, advantages of the sick-role, withdrawal from responsibilities).

Which of these potential factors are important, and how much impact they have in the individual patient have to be determined within the diagnostic and psychotherapeutic process (*see* Mental Illness, Etiology of).

Diagnostic Process

The diagnostic process follows a sequence starting with a careful monitoring of all presented bodily complaints and ending up with a reliable and valid diagnosis of a specific subtype of somatoform disorder and its severity (Schaefert et al., 2012). The knowledge of base rates for somatoform disorders and potential severe medical disorders which have to be ruled out, age of the patients, and their presentation style, history of other bodily complaints and excessive visits to other doctors might be hints for the assumption of the diagnosis.

The physician is confronted with the dilemma of possible false decisions: (1) they might diagnose somatoform disorder although the symptoms are better explained by a somatic disease which has to be treated by medical guidelines (type-1 error, 'false positive'). As a consequence patients might not receive appropriate medical treatment and might feel stigmatized as mentally ill. (2) The diagnosis of a somatic disease/dysfunction is assumed although the patient suffers from a somatoform disorder (type-2 error; 'false negative'). In most cases doctors and patients judge the risk of missing the correct

somatic diagnosis as more adverse and therefore try to minimize it. However, this can be one of the major factors maintaining somatoform disorders. A pragmatic rule for diagnosis is "to have somatic diagnostic efforts as much as necessary but as less as possible." Thorough, but not excessive or indefinite, medical examinations have to rule our severe and harmful organic disease and dysfunctions that might be causing the symptoms (e.g., hyperparathyroidism, multiple sclerosis, or brain tumor). It has to be supported by adequate and plausible procedures (e.g., physical and neurological examination, laboratory examination). Results and documents from other doctors should be revised in order to prevent repeated examination without relevance. A precise mental health examination for a complete overview of previous severe psychiatric disorders and minor complaints and the current psychopathological status is necessary. In single cases it can be difficult to decide whether the symptoms can better be explained by affective, anxiety, and personality disorders or whether they are real comorbid states. The probability of comorbid diagnoses positively correlates with symptom severity of somatoform disorders. The sequence over time and the severity of symptoms are helpful in finding the correct hierarchy. Symptoms, pathological illness behavior, course, and consequence of the complaints will allow one to find the adequate nosological diagnoses along with the operationalized diagnostic criteria. This diagnosis within the group of somatoform disorder has the status of a working hypothesis and is needed for concrete treatment planning. If the patient visits a mental health specialist he or she will gather specific information regarding nosological diagnosis. A sophisticated analysis of all relevant information concerning the individual development, learning, and predisposing factors, the onset, and further course of the symptoms, which is based on a life-chart interview, and information about actual psychosocial conditions, conflicts, and traumata is inevitable. Of course it partly depends on the theoretical assumptions and etiological models of the different psychotherapeutic schools (*see* Differential Diagnosis in Psychiatry).

Treatment

Because psychological factors are most influential in the development and course of somatoform disorders they have to be addressed in every effective treatment. Of course the setting, intensity, and content of psychological interventions can vary depending on factors like severity of symptoms (duration, intensity, functional impairments), comorbidity, insight, and motivation of the patient and availability of psychotherapy. In a step-by-step approach the primary care physician should start with short-term verbal interventions. Acute and isolated somatoform symptoms may respond to psychoeducation and counseling within the usual medical visits. Counseling and advice for general and easy-to-administer methods of symptom monitoring and management can be helpful. The physician should know the pitfalls in the treatment of patients with somatoform disorders and avoid procedures, which might reinforce the patients' pathological illness behaviors (e.g., no medical tests or laboratory examinations without clear indication). Information about and motivation for more specific forms of psychotherapy are essential if these basic interventions within the primary care system are not sufficient. For the more severe patients a referral to mental health experts is needed, for chronic and therapy-refractory cases even psychiatric or psychosomatic inpatient treatment may be necessary. Unspecific factors, which are important in any psychotherapy (e.g., therapeutic alliance, expectations of being helped, enhancing the sense of mastery, etc.) can be differentiated from specific therapeutic techniques and be adapted to relevant pathological processes in somatoform disorders (see below). Multimodal and multidisciplinary treatment approaches seem to be more effective than single treatment approaches (Mayou et al., 1995).

A validating, accepting, and (especially in the beginning) nonconfronting patient–therapist relationship is a prerequisite for active participation in therapy. It is also the basis for the development of an alternative, holistic model of illness, the motivation to change, and active participation in therapy. Self-monitoring diaries aim at a precise identification of symptoms in relation to external or interoceptive stimuli, events, and their positive and negative consequences in concrete situations. Psychoeducation and cognitive therapy are essential for a critical discussion and revision of inadequate concepts about health and disease and a better understanding of ongoing mechanisms. For example, the hypochondriac patient may realize the effects of a feedback loop leading to a vicious circle as is the case in increased autonomic arousal through anxious thoughts and worries about illness. In a collaborative way the therapist tries to convey a more realistic and helpful paradigm of body–mind interactions which aims at a better understanding and mastery of symptoms. Relaxation training, breathing exercises, or mindfulness-based interventions are introduced as methods for an active reduction of tension, arousal, and apprehension. The patient is motivated to reduce maladaptive illness behaviors like body scanning and checking, excessive medical visits, seeking reassurance, and social withdrawal. Cooperation with physicians and family members is warranted so that they can avoid inadequate reactions to the patients' illness behaviors. Alternative behaviors, e.g., exercises, sports, and positive activities have to replace the symptomatic ones and can help to re-experience bodily functions in a more positive way. Stress management, social competence, and problem-solving interventions can be used optionally to improve coping with stress, conflicts, and interpersonal problems. Furthermore, therapy has to prevent relapses and enhance a more balanced life style. As some of the symptoms might resist the described interventions or take a longer time to be effective, patients have to learn how to 'live with the residual complaints as best as possible' and to regain a better quality of life instead of waiting for complete remission.

Up to now there is no empirically validated form of specific pharmacological therapy. Clinical experience shows that in some cases tricyclic antidepressants and selective-serotonin-inhibitors might be helpful, especially in patients with comorbid depression, anxiety disorders, and pain. However, without concomitant psychotherapy the risk of relapse is usually high. Benzodiazepines and low-potency neuroleptics should not be given because of the risks of long-term intake, especially abuse and dependency in benzodiazepines and tardive dyskinesia in neuroleptics.

Compared to other psychiatric disorders the empirical evidence for effective treatments is still low in somatoform disorders. But there is a growing number of controlled outcome studies since the syndromes were clearly defined in 1980. These studies have to face the fact that most of these patients are not treated within mental health services but in the primary care system or in consultation and liaison services. Additional difficulties arise from the complexity of recent approaches, which can only be evaluated as 'total treatment packages' (*see* Antidepressant Drugs; Behavior Therapy: Psychiatric Aspects).

Summary and Outlook

The clearer operationalization of diagnoses within *DSM* and *ICD* classification systems stimulated a lot of empirical studies concerning the epidemiology, etiology, pathogenesis, and treatment of somatoform disorders in the last three decades. Medical and mental health services have realized the clinical relevance of these disorders and the need for early recognition, collaborative, stepped-care, and multidisciplinary intervention programs (Schaefert et al., 2012). Clinical observations and outcome studies have challenged the former statement, that patients with somatoform disorders in general have a poor prognosis. The acceptance of empirically based psychotherapeutic and psychiatric interventions can be enhanced if their positive effects are transparent to the patient. Convincing models for a better explanation and understanding of symptoms, the definition of limited and concrete treatment goals, and the provision of successful experiences in symptom reduction and improving coping abilities may have the function of a 'positive vicious circle' and convey a sense of mastery.

However, patients and their caregivers have to overcome a series of obstacles:

1. The first step is the correct recognition and early diagnosis by the primary care physician.
2. Divergent health belief systems of patient and physician may burden the relationship and cooperation between them, so that it is difficult to realize psychological interventions.
3. Although psychological treatment is indicated, physicians often avoid time-consuming verbal interventions or are not adequately trained for their application. However, it is not clear whether psychotherapeutic interventions, which are proven within a psychotherapeutic context, can be adopted to primary care without diminishing their efficacy.
4. If physicians try to refer to mental health experts a lot of patients refuse this referral (e.g., for reasons of stigma) or are not motivated for psychotherapy. Since empirical validation of treatment approaches for subtypes of somatoform disorders is still different, the selection of therapeutic methods often depends on local availability and personal preferences rather than on evidence-based criteria.

Each of these items is a challenge for clinical and research efforts as well as for a better dissemination of existing knowledge and competence to different health professionals who are involved in the treatment of somatoform patients. Since the first edition of this text a lot of empirical studies have contributed to our knowledge and are congregated in some *meta*-analyses and Cochrane reviews.

In the last decade a controversial debate concerning the nosological congregation in somatoform, related disorders and their diagnostic criteria (Rief and Hiller, 1999) was ongoing. Since the introduction of the modified concept of 'somatic symptom disorder' in *DSM-V* it is even more pronounced. Especially the elimination of the requirement that symptoms must be 'medically unexplained' is criticized. Prominent researchers call the *DSM-V* concept as poorly tested and as too 'loose'. In the case of 'false positive diagnosis' the patient has to bear the burden of severe or harmful consequences, especially if underlying medical causes are missed and not treated adequately. Additionally many patients maybe 'mislabeled as mentally ill' (Frances, 2013). High rates of mutual comorbidity between anxiety, affective, substance abuse, and somatoform disorders have challenged the usefulness of the concept of discrete and stable nosological entities and different treatment needs. Some clinicians had proposed a concept of spectrum disorders with more transient borderlines between different, but obviously linked, clusters of psychopathology and their courses in a long-term perspective. Specifically the relevance of subclinical forms of disorders as precursors and remaining psychopathological states after partial remission is stressed. Some recent candidates are thought to be potentially affiliated as psychiatric diseases with apparent parallels to somatoform disorders. Chronic fatigue syndrome, fibromyalgia, and multiple chemistry sensitivity, etc. are syndromes with bodily complaints of unknown etiology and pathogenesis. Their treatment might profit from adopting interventions, which have been successfully applied in somatoform disorders.

See also: Fear: Psychological and Neural Aspects; Freud, Sigmund (1856–1939); Hysteria: History and Critiques; Janet, Pierre (1859–1947); Normality and Mental Health; Psychosomatic Medicine; Stress, Coping and Health.

Bibliography

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV, fourth ed. American Psychiatric Association, Washington, DC.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders: DSM-V, fifth ed. American Psychiatric Association, Washington, DC.

Barsky, A.,J., Orav, E.J., Bates, D.W., 2005. Somatization increases medical utilization and costs independent of psychiatric and medical comorbidity. Archives of General Psychiatry 62, 903–910.

Bass, C.M., Murphy, M.R., 1990. Somatization disorder: critique of the concept and suggestions for future research. In: Bass, C. (Ed.), Somatization: Physical Symptoms and Psychological Illness. Blackwell, Oxford, UK, pp. 301–332.

Frances, A., 2013. The new somatic symptom disorder in DSM-5 risks mislabeling many people as mentally ill. British Medical Journal 346, f1580. http://dx.doi.org/10.1136/bmj.f1580 (Published 20 March 2013).

Guggenheim, F.G., 2000. Somatoform disorders. In: Saddock, B.J., Saddock, V.A. (Eds.), Comprehensive Textbook of Psychiatry. Lippincott Williams and Wilkins, Philadelphia, PA, pp. 1504–1532.

Hiller, W., Rief, W., Brähler, E., 2006. Somatization in the population: from mild bodily misperceptions to disabling symptoms. Social Psychiatry and Psychiatric Epidemiology 41, 704–712.

Kellner, R., 1985. Functional somatic symptoms and hypochondriasis. Archives of General Psychiatry 42, 821–833.

Lipowski, Z.J., 1988. Somatization. The concept and its clinical application. American Journal of Psychiatry 145, 1358–1368.

Mayou, R., Creed, F., Sharpe, M., 1995. Treatment of Functional Physical Symptoms. Oxford University Press, Oxford, UK.

Olde Hartmann, T.C., Borghuis, M.S., Lucassen, P.L., van der Laar, F.A., Speckens, A.E., van Weel, C., 2009. Medically unexplained symptoms, somatization disorder and hypochondriasis: course and prognosis. A systematic review. Journal of Psychosomatic 71, 325–334.

Pilowsky, I., 1997. Abnormal Illness Behaviour. Wiley, Chichester, UK.

Rief, W., Hiller, W., 1999. Toward empirically based criteria for the classification of somatoform disorders. Journal of Psychosomatic Research 46, 507–518.

Rief, W., Hiller, W., 1998. Somatisierungsstörung und Hypochondrie (Somatization Disorder amd Hypochondriasis). Hogrefe, Göttingen.

Rief, W., Nanke, A., 1999. Somatization disorder from a cognitive-psychobiological perspective. Current Opinion in Psychiatry 12, 733–738.

Rief, W., Broadbent, E., 2007. Explaining medically unexplained symptoms – models and mechanisms. Clinical Psychology Review 27, 821–841.

Rief, W., Hennings, A., Riemer, S., Euteneuer, F., 2010. Psychobiological differences between depression and somatization. Journal of Psychosomatic Research 68 (5), 495–502.

Schaefert, R., Hausteiner-Wiehle, C., Häuser, W., Ronel, J., Herrmann, M., Hennigsen, P., 2012. Clinical practice guideline: non-specific, functional and somatoform bodily complaints: Dtsch. Ärzteblatt International 109 (47), 803–813. http://dx.doi.org/10.3238/arztebl.2012.0803.

Simon, G.E., von Korff, M., 1991. Somatization and psychiatric disorder in the NIMH epidemiologic catchment area study. American Journal of Psychiatry 148, 1494–1500.

Tak, L.M., Rosmalen, J.G., 2010. Dysfunction of stress responsive systems as a risk factor for functional somatic syndromes. Journal of Psychosomatic Research 68, 461–468.

World Health Organization, 1991. International Classification of Diseases (ICD-10). WHO, Geneva.

Somatotherapy, History of

Santosh K Chaturvedi, National Institute of Mental Health and Neurosciences, Bangalore, India
Soumya Parameshwaran, Kasturba Medical College, Mangalore, India

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by T.A. Ban, volume 21, pp. 14600–14605, © 2001, Elsevier Ltd.

Abstract

The history of somatotherapy is as old as the identification of mental health problems, since ancient times. However, the modern history spans the past century. The history of somatotherapy for mental illness is also a reflection of the understanding of mental illnesses during those periods. This article traces the history from chemical substances to control aggressive behaviors (not modern psychopharmacological agents), to fever therapy and insulin coma treatments. On a parallel was the development of methods to induce seizures by chemical and electrical methods – electroconvulsive therapy – and by other methods of brain stimulation – repetitive transcranial magnetic stimulation, deep brain stimulations, and transcranial direct current stimulations. The future may see stereotactic surgeries and other methods of intracranial localized brain stimulations. Other physical methods of treatments, like vagus nerve stimulation, phototherapy, sleep-inducing therapies, and sleep deprivation therapies, had shorter lives.

Somatotherapy refers to the treatment of psychiatric disorders by physical, chemical, or surgical means. Ancient methods included the use of blood letting, blood-sucking leech bites, branding, cold packs, straight jackets, and many more rather cruel, traumatic and not-very-safe methods. The development of modern somatotherapy began in the nineteenth century with the introduction of morphine, bromides, and chloral hydrate for sedation and control of behavioral abnormalities. It continued in the first quarter of the twentieth century with the introduction of malaria-induced fever for syphilitic general paralysis, and prolonged sleep. It was during the 1930s that insulin coma therapy (ICT), electroconvulsive therapy (ECT), and psychosurgery were introduced. ICT and ECT were the first effective physical therapies for schizophrenia without considerable risk. It was during the 1940s that the first causal treatments in psychiatry, penicillin for syphilitic general paralysis and nicotinic acid for cerebral pellagra, were introduced. Introduction of therapeutically effective psychotropic drugs began with the introduction of lithium in 1949. By the year 2000, well over 20 neuroleptics, antidepressants, and anxiolytics were in clinical use. There are also three mood stabilizers for bipolar disease, and a couple of cholinesterase inhibitors for Alzheimer-type senile dementia. Introduction of the spectrophotofluorometer in the 1950s led to the development of neuropsychopharmacology. The development of receptor binding assays in the 1970s, followed by the separation of receptor subtypes in the 1980s and the introduction of a new genetic

Developments that led to pharmacotherapy

1806	Morphine	Serturner of Germany isolated morphine from opium	Morphine was used for controlling agitation and aggression.
1826	Bromine	Isolation by Balard in France	Bromine was separated from the ashes of the seaweed in France by Balard in 1826; potassium bromide was used for relieving restlessness and anxiety, and was introduced for clinical use in the late 1850s.
1832	Chloral hydrate	Chloral hydrate was synthesized in Germany by Liebig	Chloral hydrate was used for insomnia (Lehmann and Ban, 1970).
1880s	Hyoscine		Subcutaneously administered combinations of morphine (or apomorphine) and scopolamine (hyoscine) became the prevailing treatment modality of excitement and agitation in psychiatry by the dawn of the twentieth century (Ban, 1969).
1864	Barbiturates	Barbituric acid was synthesized by von Baeyer	The synthesis of malonylurea (barbituric acid) opened the path for the development of barbital (veronal), the first barbiturate hypnotic introduced in 1903. Barbital, together with many other long-, intermediate-, short-, and ultrashort-acting barbiturates, dominated the treatment of insomnia for about 50 years (Lehmann, 1993).
1930s	Inhalation of 30% carbon dioxide and 70% oxygen mixture		One of the targets was catatonic inhibition, which by the mid-1930s had been transiently removed in some patients by the inhalation of a 30% carbon dioxide and 70% oxygen mixture, and the administration of sodium amobarbital and apomorphine (Loevenhart et al., 1929).

Continued

—cont'd

Beyond behavioral control			
Mid-1890s	Fever therapy	Wagner-Jauregg (he received the Nobel Prize in 1927 for introducing the first effective treatment of neurosyphilis)	Wagner-Jauregg's fever therapy was based on his clinical observations that his patients, recovered from psychosis on developing high-grade fever, based on which he induced fever with tuberculin vaccine in several of his patients with general paresis and found favorable results. He replaced the vaccine in 1917 by blood from patients with tertian malarial attacks, because malaria could be terminated by quinine. Six of the first nine patients treated with 'fever therapy' improved; three fully recovered (Wagner-Jauregg, 1919). Though tried in other psychiatric disorders (e.g., depression, schizophrenia) and fever had been induced by means, e.g., typhoid vaccine, killed *Escherichia coli* suspension, sodium nucleinate, sulfur in oil, and hot bath, even in general paresis, only in one-third of the patients did it yield 'satisfactory' results (Pichot, 1983).
1900	Sleep therapy	Macleod	First introduced in Shanghai with the employment of bromides ('bromide sleep') by Macleod for mania in 1900. After 1 year, Wolff employed methylsulfonal ('Trionalkur'). Epifaneo in 1915 reintroduced 'prolonged sleep' ('*l'ipnosiprolongata*') with long-acting barbiturates, later adopted by Klaesi for schizophrenia as 'continuous narcosis' (*Dauernarkose*) (Klaesi, 1922). During the 1930, the 1940s, and the 1950s, 'sleep therapy' was used in conditions ranging from manic and catatonic excitement, through severe anxiety and depression, to various forms of schizophrenia. Various combinations used frequently – Somnifen (diethyl- and allyl-isopropyl-barbituric acid) used by Klaesi, Cloetta mixture (paraldehyde, amylhydrate, chloral hydrate, alcohol, isopropyl-allyl-barbituric acid, digalene, and ephedrine). Sleep therapy was of marginal benefit, but of considerable risk; there was 5% fatality from pneumonia and/or circulatory collapse. Nevertheless, 'sleep therapy' was lingering on, even after the introduction of new psychotropics – with the incorporation of drugs like chlorpromazine in some of the 'sleep mixtures' – before fading away in the 1960s.
1971	Sleep deprivation	Pflug and Tolle	Sleep deprivation was found to be therapeutic in endogenous depression (Pflug and Tolle, 1971). Since the early 1970s, sleep deprivation has been employed with success in the facilitation of the therapeutic effect of these drugs.
1923	Insulin		The first report on the successful treatment of symptomatic depression with insulin in diabetic patients was published in 1923. Insulin was introduced into psychiatry for stimulation of appetite and sedation in the late 1920s. Manfred Sakel's report in 1930 was one of the first on the therapeutic effect of insulin in morphine withdrawal.
1940s	ICT	Sakel	Sakel established the effectiveness of ICT in schizophrenia. He published his findings in a series of 13 papers in the *Wiener Medizinische Wochenschrift* between November 1934 and February 1935 (Sakel, 1935). ICT fared favorably to all other therapies used in the treatment of schizophrenia in the 1940s. ICT was short-lived in the Western world. In the Soviet Union and China, it remained one of the major forms of treatment of schizophrenia until the late 1980s.

technology in the 1990s, rendered the tailoring of psychotropic drugs possible.

Psychosurgery

Since antiquity, humans tried to treat mental illnesses surgically. Trephination of skulls dates back to the stone ages. Skull trephinations have been described to be performed in the Mesolithic and Neolithic eras; the procedure seemed to have taken place intra vitam since the healing process was obvious. Trephination was performed for religious reasons, as part of the treatment of somatic and psychic disorders, for segregation purposes, and also as a punishment.

Skull surgery was used for treatment of affective disorders started with the works of Gottlieb Burckhardt (1837–1907) and Ludvig Puusepp (1875–1942), at the dawn of the nineteenth century.

John Fulton (1896–1960) extensively studied the function of the cingulate cortex in primates. In his works with Carlyle Jacobsen, he reported their experience with two higher primates, which were lobectomized and became remarkably flaccid, though maintaining their ability to perform tasks.

The Portuguese neurologist Egas Moniz (1874–1955) along with Almeida Lima (1903–1985) introduced 'frontal leucotomy' for psychiatric illness, treating a woman suffering from 'mélancolie anxieuse' with satisfactory results. Similarly, Freeman and Watts and Tooth and Newton, working independently, performed prefrontal lobotomies for treating patients with mental disorders. However, the initial enthusiasm for this innovative treatment of medically intractable psychiatric disorders and chronic pain was soon tempered by reports of significant undesirable side effects.

Neurosurgeons started looking for alternative surgical procedures that would increase safety without reducing efficacy. Among them were less invasive and relatively more selective procedures such as cingulotomy, thalamotomy, amygdalotomy, and other stereotactically guided lesioning procedures, which targeted specific brain structures without damaging adjacent brain areas (Brotis et al., 2009).

Electroshock Therapy

The notion that convulsions may promote wellness has existed for centuries. During the sixteenth century, the Swiss alchemist Paracelsus gave camphor by mouth to induce convulsions and 'cure lunacy,' following which several cases of convulsions induced by chemical means, specifically camphor in oil, were documented in the eighteenth and nineteenth centuries (Payne and Prudic, 2009).

In 1934, Ladislas Joseph von Meduna, a Hungarian neuropsychiatrist, conducted experiments in treating patients with schizophrenia by inducing repeated seizures. Meduna's own neuropathology studies found that the glial cell concentrations in the brains of patients with epilepsy were much higher than normal, whereas glial cell concentrations in the brains of patients with schizophrenia were much lower. Meduna thus hypothesized that seizures may be 'protective' against psychosis and that inducing convulsions in patients with schizophrenia could reduce their symptoms (Fink, 2004). He demonstrated a significant decrease in psychotic symptoms in patients treated with a series of induced seizures with dramatic improvement especially noted in patients who appeared to have catatonia (Fink, 1984).

Meduna began with intramuscular injections of camphor, which was known to be a stimulant at lower doses. However, since camphor was painful and variable in its effect, he switched agents to pentylenetetrazol (Metrazol), a more refined synthetic substance. Although this method was effective in promoting seizures, it was still variable in its effects and produced a sense of panic in patients (Mankad et al., 2010).

In 1938, Italian scientists Cerletti and Bini succeeded in defining the parameters necessary for applying electricity

Psychosurgery	
Frontal leucotomy	Introduced in 1937 by Portuguese neurologist Egas Moniz, assisted by neurosurgeon Almeida Lima for the treatment of schizophrenia
	Based on the fact that, by creating lesions in the prefrontal cortex, abnormal activity, presumably originating in this region, would be prevented from spreading to other areas of the brain, which could worsen the symptoms of the disease
	Classical lobotomy causing extensive brain injury fell into disuse after the 1950s, being replaced by localized procedures
Anterior cingulotomy	Currently the most popular psychosurgical procedure performed in the United States, earlier it was performed as an open procedure; results using stereotactic cingulotomy have recently been described
Stereotactic subcaudate tractotomy	Introduced in 1964. Radioactive yttrium drops are implanted in the tissue to be damaged, the substantia innominata, and beneath the head of the caudate nucleus
Limbic leucotomy	Introduced in 1973 by Kelly and Mitchell-Heggs; a combination of bilateral cingulotomy and subcaudate tractotomy
Anterior capsulotomy	Lesion to the anterior limb of the internal capsule, containing radiating anterior thalamic fibers connecting the cortex to the pons
Amygdaloidotomy	Clinical experience with stereotactic lesions of the amygdala has shown that amygdaloidotomy in a large percentage of cases resulted in a marked reduction in antisocial behavior (Soares et al., 2013)

directly to the human scalp; they treated a 39-year-old man who was found delusional in a train station. He recovered fully after 11 treatments without adverse effects and thus ECT was born.

Electroshock therapy (EST) was administered through electrodes placed on both temples that Bini found safe in dogs. Cerletti and Bini reported their findings with EST to the Italian Academy of Medicine in Rome on 28 May 1938. The treatment was therapeutically effective and safe; the only drawbacks it had were fractures and memory disturbance. The risk for memory disturbance with EST was greatly reduced by the use of 'brief-pulse currents' and moving the path of the current in the brain away from the memory centers by locating the electrodes on one side, or over the very front of the head. An added advantage of bifrontal electrode placement is that seizures are elicited at near-threshold frequencies.

The risk for fractures was greatly reduced by pretreatment with a neuromuscular blocker. Bennett in the United States was the first in 1940 to employ curare to relax the muscle. Curare, a substance that competes with acetylcholine at the neuromuscular junction, was replaced by succinylcholine, a safer drug, which prevents the action of acetylcholine by depolarizing the muscle end plates.

To prevent the panic that is caused by not being able to breathe, succinylcholine administration is preceded by the administration of a short-acting barbiturate. ECT, administered under general anesthesia with continuous oxygenation and monitoring of the body's physiology during treatment, is the safest among the somatic therapies today. It offers specific advantages for the elderly, in pregnancy, and for patients with cardiac disease.

From that time through the 1950s, ECT was extensively used in the United States where it joined a cadre of somatic therapies, including psychosurgery and ICT, that were being used to treat severely ill patients for a broad variety of symptoms. Other somatic treatments were introduced during this period, but ECT was the only treatment that flourished.

The indications of ECT have been extended over the years from schizophrenia to include mania and depression. However, in mania and schizophrenia, it is used only in treatment failures or in patients with specific contraindications to drugs. In depression, ECT is also used as primary treatment in delusional psychotic patients and in patients perceived to have suicidal risks.

In the mid-1990s, a new technique, high-energy transcranial magnetic stimulation (TMS) of the brain, was introduced. Expectations were high that it could eventually replace ECT. This, however, was not to be the case. Compared with ECT, the physiological effects of repetitive transcranial magnetic stimulation (rTMS) are rather modest (Payne and Prudic, 2009).

Transcranial Magnetic Stimulation

The first reports documenting activation of human brain by an electromagnetic field appeared as early as 1896. Jacques-Arsèned'Arsonval and, later, Sylvanus P. Thompson (1910) demonstrated that placing the head near an electric coil generating a time-varying magnetic field induced the sensation of phosphenes (i.e., the subjective sensation of light) and vertigo. Subsequent efforts were pursued aimed at refining the technique; however, all these magnetic stimulation devices were technically difficult to operate, prone to extreme overheating, and had no mechanisms to address the interference generated by the stimulating fields.

In the late 1970s and the early 1980s, Anthony Barker and collaborators at the University of Sheffield, UK, overcame many of these earlier technical problems and developed a device capable of generating peak magnetic fields as high as 2 T. This work served as a precursor to the development of a stable and reliable device and, ultimately, the first commercially available magnetic stimulator. During the past 20 years, the application of TMS in clinical neurology and neuroscience research has grown steadily due to improvements in coil design and the development of stimulators capable of rapid discharge rates, its combined use with other neuroimaging modalities, and a greater understanding of the technique (George et al., 1999).

rTMS is a noninvasive neurophysiologic method to induce electrical currents within the brain. This technique can be used to transiently disrupt the function of a targeted brain area, functionally map cortical areas, assess cortical excitability, and even modulate its level of activity when delivered as a repetitive train of pulses. TMS is being widely used to investigate complex aspects of human brain function, such as learning and cognition, as well as sensory and motor function. Furthermore, its potential as a novel therapeutic treatment is being actively pursued in a variety of neurological and neuropsychiatric disorders (Merabet and Pascual-Leone, 2009).

In 1993, Hoflich et al. were the first to use TMS as a therapeutic modality in an open trial of TMS for depression (Hoflich et al., 1993). The best clinical gains have been observed on the use of rTMS for the treatment of major depression. In addition, therapeutic efficacy has been noted in mania, negative symptoms of schizophrenia, and potential anticraving in substance dependence (Mishra et al., 2011).

Vagal Nerve Stimulation

The background for studying whether vagal nerve stimulation (VNS) might have antidepressant effects, culminating in the first implant for this indication in July 1998 at the Medical University of South Carolina in Charleston (Rush et al., 2000) was a decade after the first human epilepsy implant (Penry and Dean, 1990). These hints were (1) mood effects of VNS observed in patients with epilepsy, (2) evidence by positron emission tomography scans that VNS affects the metabolism of important limbic structures, (3) the role of anticonvulsant medications in mood disorders, and (4) neurochemical studies in both animals and humans revealing that VNS alters concentrations of monoamines within the central nervous system.

During the clinical trials of VNS for epilepsy, several investigators noted mood improvements in their patients. Although decreased seizure frequency likely accounted for some of these improvements, the clinical impression was that mood improvement went beyond those attributable to improved seizure control alone. For example, some patients with minimal or no improvement in seizure frequency also reported substantial improvements in mood. Improved quality of life was also seen, and without seizure frequency changes in some patients (George et al., 2000).

Deep Brain Stimulation

Electrocorticography was further developed by many of the giants of neurosurgery: Keen, Krause, Sherrington, and Penfield. However, it took another 80 years before the technology advanced to the point that one could consider implantation of stimulating electrodes with which a patient could walk around.

Horsley had performed cortical ablations of the pyramidal system for movement disorders in 1890. Meyers first tested lesions of the basal ganglia to alleviate movement disorders in 1939; however, because these were open procedures before the advent of the operating microscope, the morbidity was significant with a 12% risk of mortality. This changed with the development of the stereotactic frame by Spiegel and colleagues.

In 1947, although Horsley and Clarke had designed an earlier stereotactic frame for use in animals using surface landmarks, it was not very accurate due to variability between subjects. Spiegel et al. used pneumoencephalograms to determine Cartesian coordinates of structures around the ventricles. In their first article describing the apparatus, they reported using the device for psychosurgery, pain (including phantom limb pain), pallidotomy for abnormal movements, electrocoagulation of the Gasserian ganglion for trigeminal neuralgia, and withdrawal of fluid from cystic tumors. They went on to develop a stereotactic atlas of the human brain to guide surgery. With these new tools, surgery could be performed with less than 1% mortality. However, even with the development of multiple frames, atlases, and lesioning techniques, there were issues of inaccurate targeting with stereotaxis. True microelectrode recording with low-impedance semimicroelectrodes was pioneered by Able-Fessard in 1961. These tools, as well as computer modeling, allowed for the development of neurophysiologic atlases of the human brain (Schwalb and Hamani, 2008).

There was an explosion of stereotactic functional neurosurgery in the late 1950s. Cingulotomy, anterior capsulotomy, and subcaudate tractotomy were used for pain and psychiatric disease as an alternative to the ice pick frontal leukotomies performed by Walter Freeman. Pallidotomy and thalamotomy were used for Parkinson disease and dystonia. The stage was set for electrical stimulation for movement disorders and psychiatric disease with the development of the first totally implantable pacemaker by William Chardack and colleagues for the heart in 1960.

Surgery for psychiatric disease had been decreasing since the introduction of chlorpromazine in 1954, and it effectively came to a halt, at least in most centers in the United States, in 1977, based on the recommendations of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (Kopell and Rezai, 2003).

Among the most promising new applications of deep brain stimulations (DBS) at present are psychiatric disorders. This field is clearly of great interest due to the epidemiology of these diseases, especially major depression out of which more than 35% are resistant to medications and ECT (Rush et al., 2006).

Phototherapy in Psychiatry

Phototherapy or light therapy in psychiatry is closely intertwined with the original description of the syndrome of seasonal affective disorder. Two decades ago, Rosenthal and colleagues described a series of patients with histories of recurrent depressions that developed in the fall or winter and spontaneously remitted during then following spring or summer (Golden et al., 2005). Light therapy has been most extensively applied to the syndrome of 'winter depression.' In winter, these patients complain of fatigue, hypersomnia, overeating, and feelings of sadness; in the spring, these symptoms typically remit spontaneously. Subsequent to the original study in 1980, over a 100 patients with winter depression or seasonal affective disorder have been treated with bright light (Lewy et al., 1982). The analysis of randomized controlled trials suggests that bright light treatment and dawn simulation for seasonal affective disorder and bright light for nonseasonal depression are efficacious, with effect sizes equivalent to those in most antidepressant pharmacotherapy trials (Golden et al., 2005).

Transcranial Direct Current Stimulation (tDCS)

Transcranial direct current stimulation (tDCS) is a noninvasive form of brain stimulation in which a relatively weak direct current is passed into the cerebral cortex through small scalp electrodes. This resulted in a modulation of cortical excitability and spontaneous neural activity, depending on the polarity of the stimulation. Because of its neuromodulatory effects, tDCS has been investigated as a treatment for neurological and psychiatric diseases (Kalu et al., 2012).

Attempts to explore the effects of tDCS on mood and depressive symptoms in humans have been made for several decades. Anodal stimulation was suggested to diminish depressive symptoms, while cathodal stimulation reduced manic symptoms. Unfortunately, these results were not replicated (Nitsche et al., 2008).

Some open-label studies and clinical observations reported positive effects of tDCS in depression. However, controlled trials failed to replicate these findings. These initial mixed findings, together with the development of psychotropic drugs, led to a decreased interest in tDCS as a treatment tool for major depression during the 1970s and the 1980s.

Methodological improvements and the promising results of rTMS in depression led to reevaluation of the therapeutic application of tDCS. Based on the hypothesis that the left dorsolateral prefrontal cortex is hypoactive in depression (Grimm et al., 2008), the common rationale of currently performed studies is to modify activity in the prefrontal cortex and reestablish the balance of left and right prefrontal cortex activation (Kalu et al., 2012).

Other physical therapies, used sparingly, and for very specific disorders are sleep deprivation therapy and aversive therapies used in behavioral medicine (most considered as a type of behavior therapy, rather than physical therapies).

See also: Bipolar (Manic-Depressive) Disorder; Deinstitutionalization, Psychiatry of; Obsessive-Compulsive Disorder; Schizophrenia, Treatment of.

Bibliography

Brotis, A.G., Kapsalaki, E.Z., Paterakis, K., Smith, J.R., Fountas, K.N., 2009. Historic evolution of open cingulectomy and stereotactic cingulotomy in the management of medically intractable psychiatric disorders, pain and drug addiction. Stereotactic and Functional Neurosurgery 87, 271–291.

Ban, T.A., 1969. Psychopharmacology. Williams & Wilkins, Baltimore, MD.

Fink, M., 1984. Meduna and the origins of convulsive therapy. American Journal of Psychiatry 141, 1034–1041.

Fink, M., 2004. Induced seizures as psychiatric therapy: Ladislas Meduna's contribution in modern neuroscience. Journal of ECT 20, 133–136.

George, M.S., Lisanby, S.H., Sackeim, H.A., 1999. Transcranial magnetic stimulation: applications in neuropsychiatry. Archives of General Psychiatry 56, 300–311.

George, M.S., Sackeim, H.A., Rush, A.J., Marangell, L.B., Nahas, Z., Husain, M.M., Lisanby, S., Burt, T., Goldman, J., Ballenger, J.C., 2000. Vagus nerve stimulation: a new tool for brain research and therapy. Biological Psychiatry 47, 287–295.

Golden, R.N., Gaynes, B.N., Ekstrom, R.D., Hamer, R.M., Jacobsen, F.M., Suppes, T., Wisner, K.L., Nemeroff, C.B., 2005. The efficacy of light therapy in the treatment of mood disorders: a review and meta-analysis of the evidence. American Journal of Psychiatry 162, 656–662.

Grimm, S., Beck, J., Schuepbach, D., Hell, D., Boesiger, P., Bermpohl, F., Niehaus, L., Boeker, H., Northoff, G., 2008. Imbalance between left and right dorsolateral prefrontal cortex in major depression is linked to negative emotional judgment : an fMRI study in severe major depressive disorder. Biological Psychiatry 63, 369–376.

Hoflich, G., Kasper, S., Hufnagel, A., Ruhrmann, S., Moller, H.J., 1993. Application of transcranial magnetic stimulation in the treatment of drug-resistant major depression. Human Psychopharmacology 8, 361–365.

Kalu, U.G., Sexton, C.E., Loo, C.K., Ebmeier, K.P., 2012. Transcranial direct current stimulation in the treatment of major depression: a meta-analysis. Psychological Medicine 42 (9), 1791–1800.

Klaesi, J., 1922. Ueber die therapeutische Anwendung der 'Dauernarkose' mittels Somnifens bei Schizophrenen. Zeitschrift fur die gesamte. Neurologie und Psychiatrie 74, 557–592.

Kopell, B.H., Rezai, A.R., 2003. Psychiatric neurosurgery: a historical perspective. In: Rezai, A.R., Rasmussen, S.A., Greenberg, B.D. (Eds.), Surgery for Psychiatric Disorders. W.B. Saunders Company, Philadelphia, pp. 181–198.

Lehmann, H.E., Ban, T.A., 1970. Pharmacotherapy of Tension and Anxiety. Thomas, Springfield, IL.

Lehmann, H.E., 1993. Before they called it psychopharmacology. Neuropsychopharmacology 8, 291–303.

Lewy, A.J., Kern, H.E., Rosenthal, N.E., Wehr, T.A., 1982. Bright artificial light treatment of a manic-depressive patient with a seasonal mood cycle. American Journal of Psychiatry 139, 1496–1498.

Loevenhart, A.S., Lorenz, W.F., Waters, R.M., 1929. Cerebral stimulation. Journal of the American Medical Association 92, 880–883.

Mankad, M.V., Beyer, J.L., Weiner, R.D., Krystal, A.D., 2010. Clinical Manual of Electroconvulsive Therapy, first ed. American Psychiatric Publication, Inc., Arlington.

Merabet, L.B., Pascual-Leone, A., 2009. Transcranial Magnetic Stimulation, 1055–1062.

Mishra, B.R., Sarkar, S., Praharaj, S.K., Mehta, V.S., Dwivedi, S., Nizamie, S.H., 2011. Repetitive transcranial magnetic stimulation in psychiatry. Annals of Indian Academy of Neurology 14, 245–251.

Nitsche, M.A., Cohen, L.G., Wassermann, E.M., Priori, A., Lang, N., Antal, A., Paulus, W., Hummel, F., Boggio, P.S., Fregni, F., Pascual-Leone, A., 2008. Transcranial direct current stimulation: state of the art 2008. Brain Stimulation 1 (3), 206–223.

Payne, N.A., Prudic, J., 2009. Electroconvulsive therapy: part I. A perspective on the evolution and current practice of ECT. Journal of Psychiatric Practice 15 (5), 346–368.

Penry, J.K., Dean, J.C., 1990. Prevention of intractable partial seizures by intermittent vagal stimulation in humans: preliminary results. Epilepsy 31 (Suppl.), S40–S43 (Abstract).

Pflug, B., Tolle, R., 1971. Disturbance of the 24-hour rhythm in endogenous depression and the treatment of endogenous depression by sleep deprivation. International Pharmacopsychiatry 6, 187–196.

Pichot, P., 1983. A Century of Psychiatry. Roger Dacosta, Paris.

Rush, A.J., George, M.S., Sackeim, H.A., Marangell, L.B., Husain, M.M., Giller, C., 2000. Vagus nerve stimulation (VNS) for treatment-resistant depressions: a multicenter study. Biological Psychiatry 47, 276–286.

Rush, A.J., Trivedi, M.H., Wisniewski, S.R., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. American Journal of Psychiatry 163, 1905–1917.

Sakel, M., 1935. Neue behandlung der Schizophrenie. Perles, Vienna.

Schwalb, J.M., Hamani, C., 2008. The history and future of deep brain stimulation. Neurotherapeutics 5, 3–13.

Soares, M.S., Palva, W.S., Guertzenstein, E.Z., Amorim, R.L., Bernardo, L.S., Pereira, J.F., 2013. Psychosurgery for schizophrenia: history and perspectives. Neuropsychiatric Disease and Treatment 9, 509–515.

Thompson, S.P., 1910. A physiological effect of an alternating magnetic field. Proceedings of the Royal Society B: Biological Sciences 82, 396–398.

Wagner-Jauregg, J., 1919. Ueber die Einwirkung der Malaria auf die progressive Paralyse. Psychiatrisch-Neurologische Wochenschrift 20, 251–255.

Sorokin, Pitirim Alexandrovich (1889–1968)

Inna Ponomareva, Kuban State University, Krasnodar, Russia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article deals with the biography and sociological heritage of an outstanding social scientist, Pitirim A. Sorokin, whose pioneering ideas on social behavior and motivation, social revolution and mass starvation, social stratification and mobility, sociocultural dynamics and fluctuations, and the crisis of modern culture and altruistic love reflected the problems exercising people's minds in the challenging twentieth century and indisputably shaped the mainstream Russian and American sociology of that time.

Biographical Facts

Russian–American sociologist Pitirim A. Sorokin was born in 1889 in the village of Turya located in the Vologda province in the rural north of Russia. His father was a craftsman and his mother was a peasant. Pitirim was only 3 years old when his mother died. After her death, his father took to drinking. Pitirim and his elder brother, Vasily, had to earn for their subsistence from their childhood (their youngest brother Prokopy lived with the family of a maternal aunt). Sorokin's childhood, spent among the Komi, was complicated, but enriched by a religious and moral background. Sorokin received his elementary education at a two-form school in the village of Gum, and later continued his education at the Khrenovo Teachers School. In 1906, he was arrested for revolutionary activities and jailed in Kineshma prison, where he spent 5 months (he lived through six imprisonments in Soviet Russia). In 1907, his thirst for knowledge led Sorokin to leave his native land, and he hitched a ride on a train to St Petersburg. In 1909, Sorokin joined the Department of Sociology at the Psycho-Neurological Institute, newly opened by EV de Roberty and MM Kovalevsky. However, he soon moved to the Law Faculty of St Petersburg University, where he specialized in law under the guidance of LI Petrazhitsky, whose ideas were later popularized by Sorokin among an English-speaking audience. As a student at the Law Faculty, Sorokin became interested in the problems of social behavior and after obtaining a first-hand experience in criminology during two more imprisonments he wrote his first monograph, *Prestuplenie i kara, podvig i nagrada* (Crime and Punishment, Heroic Service and Reward) (1914). Having graduated from St Petersburg University in 1914, Sorokin began his teaching career there. The period from 1914 to 1922 is characterized by Sorokin's aspiration to "combine intense academic work with gut-wrenching and dangerous political activities" (Johnston, 1998). He participated directly in the events of the Russian Revolution in 1917 that would be the turning point in Russian history, his opinions on which were being reflected in the columns he wrote for the political daily newspaper *Volya naroda* (The Will of the People) under the heading 'The notes of a sociologist.' Sorokin hailed the February Revolution for its overthrow of Czarism, but he criticized The October Revolution because of its radical programs and actions. He was a leading member of the Social Revolutionary Party and the Russian Constituent Assembly (he served as a secretary to its leader, Prime Minister Alexander Kerensky). Sorokin was imprisoned twice in 1918 for his anti-Bolshevist propaganda and vigorous ideological fight with Communist leaders such as Lenin and Trotsky (he was even condemned to death but finally was freed by Lenin's order). Two of Sorokin's brothers died at the time of the Revolution (one was executed and the other died in prison). Among these bleak events there was one festive occasion for Sorokin in 1917 – he was married to Dr Elena Baratynskaya, a cytologist-researcher, with whom he lived happily the whole of his life. In 1920, Sorokin founded the first Department of Sociology at Petrograd University. In 1922, he received the degree of doctor of sociology, the first in Russia's history, as a result of the public debate on his work *Sistema sotsiologii* (System of Sociology) (1920). However, after being arrested once again, in September 1922 Sorokin was exiled from Soviet Russia together with many other outstanding scientists and intellectuals. In 1923, after a short stay in Germany and Czechoslovakia, Sorokin was invited to USA to deliver a series of lectures on the Russian Revolution. First, Sorokin delivered a series of guest lectures at the Universities of Illinois and Wisconsin. From 1924 to 1930 he worked at the University of Minnesota. It was during the Minnesota years that Sorokin published many of his great works, which brought him to the forefront of American Sociology. Among these works were *The Sociology of Revolution* (1925), *Social Mobility* (1927), *and Contemporary Sociological Theories* (1928). From 1930 to 1959 Sorokin worked at the Harvard University. In 1930, he founded the first Harvard University Department of Sociology, which he headed until 1942 (in 1946 the department was reorganized by Talcott Parsons as the Department of Psychology and Social Relations). This period is characterized by a revision of his previous views, by the shift from behaviorism and empiricism to the problems of culture, spiritual life, altruism, and love. Sorokin's ideas finally took shape of his integral model, presented in the four-volume book *Social and Cultural Dynamics* (1937–41). In 1949, he established the Harvard Research Center in Creative Altruism, which he continued to direct until his full retirement in 1959. In 1965, Sorokin was elected the President of the American Sociological Association (http://www2.asanet.org/governance/sorokin.html). In 1968, following 2 years of serious illness, he died in his family house in Winchester, Massachusetts. His two sons, who were born soon after the Sorokin family settled in Massachusetts, became outstanding

American scientists. Peter P. Sorokin (born in 1931) is a physicist and coinventor of the dye laser. Sergei P. Sorokin (born in 1933) is a cytologist-researcher, professor at Harvard and Boston Universities, and a composer. They treasure in memory all the things connected with their father's life and try to preserve his great sociological legacy through the functioning of the Pitirim A. Sorokin Foundation established in 2011 in Winchester, Massachusetts (http://www.sorokinfoundation.org or http://www.cliffstreet.org; http://www.pitirimsorokin.org (in Russian)).

Major Works

Two periods of Pitirim Sorokin's scientific and creative work can be distinguished: the Russian period (until 1922) and the American period (1923–68). The key sociological issues investigated during his Russian period are social behavior and motivation, the positivistic system of sociology, and famine. The American period is characterized by the strong interest in revolution, social stratification and mobility, social and cultural dynamics, and altruistic love. However, despite his having explored such a diversity of social problems, Sorokin's sociological outlook is integralistic (Ponomareva, 2011).

The problem of social behavior and motivation is considered by Sorokin in his early work *Prestuplenie i kara, podvig i nagrada* (Crime and Punishment, Heroic Service and Reward) (1914). The book is written in the spirit of a neopositivistic and behavioristic synthesis. It represents a systematic work on criminology, but with a historical and sociological orientation. As a whole, the work is devoted to two interconnected problems: the problem of social behavior, revealing itself in acts, and the problem of social motivation, revealing itself in reactions to acts. For Sorokin, the aggregate of human behavior breaks down into a number of acts and reactions to them, correlating with each other as cause and effect. He singles out three main forms of acts and reactions: permitted legal acts, which refer to fulfilling one's rights and duties and correspond to standard demands, followed by proper reaction; forbidden unallowed acts, which contradict and violate the demands of proper behavior that are followed by punishment; recommended acts, representing desirable and voluntary actions, exceeding the necessary level of proper behavior, that are followed by reward. If proper behavior is accepted as a norm, forbidden and recommended behavior should be considered as a deviation from it. Forbidden acts in the form of crimes are deviations for the worse, and are followed by negative sanctions in the form of punishment. Recommended acts in the form of heroic deeds and services are deviations for the better, and are followed by positive sanctions in the form of rewards. Punishment and reward determine and transform the acts of human behavior.

The main work of Sorokin's Russian period of creativity is his *Sistema sotsiologii* (System of Sociology) (1920), dedicated to social analytics and written from the viewpoint of neopositivism. In *System of Sociology* he tries to define the subject and the boundaries of sociology. The subject of sociology is treated by Sorokin in the following manner: "Sociology studies the phenomena of human interactions, on the one hand, and the phenomena, which appear as a result of this process, on the other" (Sorokin, 2002). The structure of sociological knowledge includes theoretical and practical sociology. Theoretical sociology investigates the phenomena of human interaction from the perspective of their real state and includes three sections: social analytics, social mechanics, and social genetics. Social analytics studies the structure of social phenomena in their static condition. Its tasks are "defining society or social phenomena, describing its main features, analyzing the interaction process, formulating main social laws" (Sorokin, 1992a). Social mechanics (social physiology), which Sorokin considers the most important section of sociology, deals with the processes of human interaction and with the forces determining human behavior in conditions of social regrouping. It faces such tasks as the study of regularities, displayed in social phenomena; the decomposition of complex social phenomena into the simplest elements; the study of peculiar features of these elements and their effects on human behavior and social life; the classification of behavioral factors and the determination of individual or group behavior. Social genetics (social dynamics), representing the evolutionary theory of social life, studies historical tendencies and explains deviation from them in different epochs of human history and spheres of social interaction. Practical or applied sociology investigates the phenomena of interaction from the viewpoint of their proper state. It includes social policy and represents the discipline of scientific management of society. It is called on to help in the realization of rational social reforms and gives humankind the ability to manage social forces.

Having published two volumes of *System of Sociology*, Sorokin postponed the writing of the third volume in order to study famine, a phenomenon typical of revolutions and mass starvation. But many chapters of his book *Golod kak faktor* (Hunger as a Factor in Human Affairs) (1921) were mutilated and cut out by Soviet censors, and finally the whole edition was destroyed by the Soviet government. Later, after Sorokin's death, the book was translated by his wife Elena and published in the USA (Sorokin, 1975).

Among the first books published by Sorokin in English is *The Sociology of Revolution* (1925), analyzing the causes and features of revolutions and written "with great behavioristic inspiration" (Sogomonov 1992). Sorokin claims that there are three typical phases in the cycle of all great revolutions. The first – primary – phase is characterized by the joy of becoming free from the tyranny of the former regime and the expectation of promised reforms. This is the so-called 'prelude.' During this stage, the government is humane and gracious, its policy is gentle, indecisive, and often helpless. The second – destructive – phase, representing reaction, is characterized by the eradication of not only obsolete, but also viable institutions and values of society. This is the so-called 'flood.' During this stage, the government is merciless, tyrannical, and sometimes blood-thirsty. Its policy is destructive and violent. The third – constructive – phase is characterized by the efforts to create a new social and cultural regime, based, however, not only on new revolutionary ideals, but also on reanimated prerevolutionary institutions and values, temporarily destroyed during the second phase of revolution. The main causes of revolution are universal suppression of the basic instincts of the majority of population, the impossibility of satisfying them, and the disorganization of power and social control. Suppressed

instincts during the revolution include alimentary instinct, instinct of property, instinct of self-preservation, sexual instinct, instinct of freedom, and instinct of self-expression. Disorganization of power and social control is understood by Sorokin as the power base's ineffective resistance to the revolutionary rise of repressed masses and its incapacity to develop sufficient countermeasures against the pressure of repressed instincts, in order to reach a state of social balance and to redirect repressed impulses toward another, nonrevolutionary channel. The continuous suppression of instincts leads to the endless fatigue of the population, as a result of which the demand for boundless freedom is changed by the thirst for order. If society realizes that the only correct way of development is not through mortal hostility, but through the return to the roots of society, former institutions and customs, creative work, cooperation, mutual help and unity of its members and social groups, then revolution comes to its logical end. Characterizing revolutions, Sorokin calls them the great tragedy and "the worst way of improving material and spiritual conditions of masses' lives" (Sorokin, 1992b), because they are followed by violence and cruelty, the reduction of freedom, deformation of the social structure, the worsening of the economic and cultural situation of the working class, and the disintegration of society as a whole. Sorokin claims that thoughtful and consistently implemented reforms, not revolutions, can be the only right way to improve and reconstruct social life.

Sorokin's book *Social Mobility* (1927), innovative in its genre, still remains the classical work on the problem of social stratification and mobility. The terminology introduced by Sorokin is universally recognized. He describes the world as social space, filled with social ties and relations. In this space, representing a system of multiple coordinates, there are two main axes – axis X (i.e., horizontal measurements) and axis Y (i.e., vertical measurements). The stratification principle lies in the basis of the inner organization of social space. Sorokin suggests the following definition for social stratification: "Social stratification means the differentiation of a certain given totality of population into hierarchically superposed classes. It is displayed in the presence of the highest and lowest layers. Its basis and essence consist in uneven distribution of rights and privileges, duties and responsibilities, social welfare and hardship, social power and influence among the members of this or that society" (Sorokin, 2005). There are three main systems of social stratification, as a rule, closely connected to one another: the economic, the political, and the occupational. Sorokin states that social stratification is a constant characteristic of any organized society. Albeit changing in form, social stratification exists in all societies that declare equality of people. Social mobility is also a normal and natural state of society. It is understood by Sorokin as "any transition of an individual or social object or value from one social position to another" (Sorokin, 2005). In Sorokin's view, there are two different types of social mobility – horizontal and vertical. Horizontal mobility concerns transition from one social position to another, situated on the same level. Vertical mobility refers to transitions of people from one social stratum to one higher or lower in the social scale. According to the direction of the transition, two types of vertical mobility can be spoken about: ascending and descending. They imply social rise and social fall. Vertical mobility is considered by Sorokin to correspond to the three forms of social stratification: economic, political, and occupational. The most important channels of vertical mobility are social institutions such as the army, the church, schools, economic, political, and occupational organizations, family, and matrimony. These are peculiar 'social lifts,' as a rule, working like 'sieves,' which sift individuals, allowing some of them to go up and keeping others in the lower strata. Social stratification and mobility in society are predetermined by the fact that people are not equal, physically and mentally. Joint activity of individuals requires hierarchical organization.

In the work *Contemporary Sociological Theories* (1928), which set Sorokin apart from his colleagues with its encyclopedic sweep, he takes a review walk through 'the forest of sociological theory,' where we can find beautiful and healthy plants and flowers (which should be used in our work) alongside their barren counterparts (which should be avoided). A number of sociological schools (mechanistic, geographical, demographical, biopsychological, sociologism) are outlined in the book. Some of them have disappeared under modern conditions (e.g., socio-Darwinism) and some of them have been transformed by reconsidering their main principles (e.g., the mechanistic and biopsychological schools). Sorokin offered an operational definition of sociology, with different aspects of which all of the general schools concern themselves: "Sociology is a study, first, of the relationship and correlations between various classes of social phenomena (correlations between economic and religious; family and moral; juridical and economic; mobility and political phenomena and so on); second, that between the social and the non-social (geographic, biological, etc.) phenomena; third, the study of the general characteristics, common to all classes of social phenomena" (Sorokin, 1928). *Contemporary Sociological Theories* is the work which had been served as a textbook in many American universities for a long time.

Sorokin's ideas finally took the shape of his integral model, presented in the four-volume *Social and Cultural Dynamics* (1937–41). According to the integral model, any society can be understood only by studying its culture; that is, a system of meanings, norms, and values, emerging under the influence of the dual biosocial nature of the individual, and integrating into a united mental style. This style is characterized by the following key features: ideas about the nature of reality and the way of knowledge, by which they are obtained; the character of individuals' main goals and needs, the degree and manner of their satisfaction. According to Sorokin, there are three main sociocultural types: ideational, sensate, and idealistic. Each of them is characterized by its own forms of art, literature, truth and knowledge, legal consciousness, and ethical and social relations. The ideational sociocultural type is characterized by the following features: reality is represented as spiritual and eternal essence; the rational way of knowledge predominates; goals and needs are spiritual, and their satisfaction is the highest possible achievement; the manner of their satisfaction is characterized by curbing carnal and material needs up to the complete refusal of sensual temptation. Thus, the ideational sociocultural type is characterized by 'the beyond' orientation, revealing itself in the predominance of transcendental values and the acknowledgment of the earthly life as but a temporary refuge. The sensate sociocultural type is characterized by

opposite features: reality is represented as material and changeable essence; the sensual way of knowledge predominates; goals and needs are carnal and material, and to satisfy these is the highest possible achievement; the manner of their satisfaction "doesn't consist in the transformation or exploitation of individuals' spiritual world, but it consists in the transformation or exploration of the outer world" (Sorokin, 2000). The sensate sociocultural type is characterized by an 'earthly' orientation, revealing itself in the predominance of hedonistic values and the acknowledgment of earthly life as the highest reality. If there is an organic balance between spiritual and sensual elements, the intermediate sociocultural type, that is, the idealistic, appears. Reality is represented as both spiritual and material essence; the intuitive way of knowledge predominates; material needs and goals are subordinated to spiritual ones, the way of their satisfaction consists both in self-improvement and in transformation of the outer sensually perceived world. Historical development is a cyclic fluctuation of these three sociocultural types. Sociocultural fluctuations are the processes recurring from time to time in the history of social and cultural life.

Sorokin's interest in the problems of sociocultural dynamics is connected with the crisis of modern culture. It is investigated in works such as *The Crisis of Our Age* (1941), *Man and Society in Calamity* (1942), *Society, Culture and Personality* (1947), and *S.O.S. The Meaning of Our Crisis* (1951). The themes of these works were partly influenced by the Second World War. According to Sorokin, these are the three important trends of the modern epoch. The first trend consists of the shift of humankind's center of creative leadership from Europe to the wider sphere of the Pacific and Atlantic coasts, especially to America, Asia, and Africa. The second trend consists of the disintegration and the decline of the modern sensate sociocultural type. In Sorokin's opinion, the twentieth century is "the most bloody of all the previous 25 centuries of Greek-Roman and Western history" (Sorokin, 1997). In connection with the domination of sensual values and in the absence of a clear differentiating line between positive and negative phenomena, the number of interindividual and intergroup conflicts grows, the state of social anomie, which causes wars, revolutions, crime and violence, emerges. Hence, the third trend consists of the emergence and development of a new integral sociocultural type. Sorokin opposes the fruitless and destructive forces of the dying sensate formation, the main function of which is natural exploration, to the creative forces of the emerging integral formation, the main function of which is ennoblement and humanization of the individual.

Sorokin also foresaw the basic trends of the development of Russian–American relations in the twentieth century. In the comparative research piece *Russia and the United States* (1944), he noted numerous sociocultural similarities between these two countries. They are both characterized by unity in diversity as "melting pots of diverse racial, ethnic, national and cultural groups and peoples" (Sorokin, 2007). Their mentalities, moral standards, and the democratic structure of the main sociocultural institutions (family, legal and judicial systems, government, and religious institutions) are very much alike. That is why Russia and America can mutually influence each other. Russia can enrich American culture and the United States in turn can help Russia to put an end to violations of human rights. The book *Russia and the United States* was republished in 2007 with a new introduction by Edward A. Tiryakian who states that this work has cardinal meaning for the international order of the twenty-first century (Tiryakian, 2007). Tiryakian produces the following proof in confirmation of Sorokin's prophetic abilities: "he was right in the conclusion about the possibility of a firm peace between the two countries in spite of the ideological statements of the Cold War era. Fortunately, that war did not develop into the 'hot' war and the triumph of political democracy and the free market system have made Russia the mirror image of the US" (Tiryakian, 2007). Unfortunately, nowadays, in the beginning of the twenty-first century these two great countries stand on the cusp of a new ideological and informational war connected with the miscommunication and controversial views on the current situation in Ukraine and Crimea.

In his latest works connected with the functioning of the Harvard Research Center in Creative Altruism – *Altruistic Love* (1950), *S.O.S. The Meaning of Our Crisis* (1951), *The Ways and Power of Love* (1954), *American Sex Revolution* (1956), *Power and Morality: Who Shall Guard the Guardians?* (1959) – Sorokin states that the most important way to overcome the crisis and immorality of modern civilization is the moral reconstruction of humankind, in which altruistic love plays the most significant part. Despite that Sorokin's activity in the sphere of creative altruism did not take its proper place in the American academic environment (for instance, Coser regards it as Sorokin's 'later retreatism' (Coser, 1977, 2006)), he himself appraised it highly (as some American sociologists would also come to do later (Johnston, 1991; Nichols, 1999)). Sorokin spoke against the Cold War between the USSR and the United States and against the Vietnam and the Korean Wars.

Lasting Impact

Sorokin's ideas are required by the present scientific community due to the current realities, namely, the fight against the global financial crisis, international terrorism, and the spread of nuclear weapons.

The Pitirim A. Sorokin Foundation (http://www.sorokinfoundation.org or http://www.cliffstreet.org; http://www.pitirimsorokin.org (in Russian)) was established by his younger son, Dr Sergei P. Sorokin, in 2011 in Winchester, Massachusetts in order to promote and support interest in Sorokin's sociological legacy by carrying on different scientific and literary activities in the United States and in Russia. The work of the Foundation is superintended by its executive director, Dr Pavel P. Krotov.

The new section of the American Sociological Association on Altruism, Morality, and Social Solidarity (http://www.asanet.org/sections/altruism.cfm and https://sites.google.com/site/altruismmoralitysolidarity/) was also organized in 2011 by Vincent Jeffries, emeritus professor of sociology at the California State University, Northridge for running the tasks of the study, critical assessment, and advancement of ideas developed by Sorokin in the 1950s and 1960s. Its present chair is Matthew T. Lee, professor of sociology at the University of Akron. Modern American sociologists state that the study of altruism and social solidarity as an area of specialization in

sociology could make a great contribution to the discipline and to society as a whole (Jeffries et al., 2006). Many Russian sociologists are the current members of this ASA section.

In addition to the renewed interest in Sorokin's later work on altruism, his writings on civilization have also become a focus of social science studies and current rethinking and reinvigoration of the sociological tradition of civilizational analysis (Weber, Durkheim, Mauss, Sorokin, Elias, Nelson, Huntington, Eisenstadt), in Samuel P. Huntington's *Clash of Civilisations?* (1993) and in the research indicated in Said A. Arjomand and Edward A. Tiryakian's *Rethinking Civilizational Analysis* (2004). Such interest results from the changing demographic, political, and economic reality of the world, as it enters a new phase.

A lot of charitable, scientific, and educational events, taking place in Moscow, St Petersburg, and Syktyvkar, are evidence of the continuing renaissance of Sorokin which began in the 1990s in his historical homeland, Russia, after long-term neglect of his ideas in the USSR. The unique documents, photographic and film materials from the Sorokin family archive were displayed for the first time in Russia, in the documentary film *The Antithesis of Pitirim Sorokin* aired on the Culture Channel in April 2009. The Faculty of Sociology at Lomonosov Moscow State University (http://www.socio.msu.ru/) regularly hosts the International Conference 'Sorokin Readings' and owns the expert and analytical portal 'Pitirim Sorokin Foundation' (http://www.sorokinfond.ru/). Pitirim Sorokin – Nikolay Kondratieff International Institute (http://www.misk.newparadigm.ru/; http://www.eng.newparadigm.ru/index.html) founded in 1999 and headed by Professor Yu V. Yakovets carries on activities on publishing Sorokin's works, conducting international conferences, and developing long-time socioeconomic forecasts on the basis of the integral macroforecast methodology. Sorokin is widely studied nowadays by Russian students within the framework of such lecture courses as 'History of Russian sociology,' 'History of classical sociology,' and 'Current sociological theory.' The numerous articles by Sorokin and about him are constantly published in the prestigious Russian sociology journals *Sotsiologicheskie issledovaniya* (*Sociological Research*) and *Sotsiologiya* (*Sociology*).

See also: Culture, Sociology of; Inequality, Social; Mobility, Social; Revolutions, Sociology of; Russian Revolution, The; Social Stratification; Sociological Theory; Sociological Traditions; Sociology: Overview.

Bibliography

Arjomand, S.A., Tiryakian, E.A. (Eds.), 2004. Rethinking Civilizational Analysis. Sage, London.

Coser, L.A., 1977. Masters of Sociological Thought: Ideas in Historical and Social Context, second ed. Harcourt Brace, New York.

Coser, L.A., 2006. Mastera sotsiologicheskoy mysli. Idei v istoricheskom i sotsial'nom kontekste (Masters of Sociological Thought: Ideas in Historical and Social Context). Norma, Moscow.

Huntington, S.P., 1993. The clash of civilizations? Foreign Affairs 72 (3), 22–49.

Jeffries, V., Johnston, B.V., Nichols, L.T., Oliner, S.T., Tiryakian, E.A., Weinstein, J., 2006. Altruism and social solidarity: envisioning a field of specialization. The American Sociologist 37 (3), 67–83.

Johnston, B.V., 1991. Integralism and the reconstruction of society: the idea of ultimate reality and meaning in the work of Pitirim A. Sorokin. Ultimate Reality and Meaning 13, 96–108.

Johnston, B.V., 1998. Introduction. Pitirim Sorokin's science of sociology and social reconstruction: an intellectual journey. In: Johnston, B.V. (Ed.), Pitirim A. Sorokin: On the Practice of Sociology. The University of Chicago Press, Chicago, pp. 1–58.

Nichols, L.T., 1999. Science, politics, and moral activism: Sorokin's integralism reconsidered. Journal of the History of the Behavioral Sciences 35 (2), 139–155.

Ponomareva, I.V., 2011. Pitirim A Sorokin: the interconnection between his life and scientific work. International Sociology 26 (6), 878–904.

Sogomonov, A.Y., 1992. Primechaniya (Notes). In: Sogomonov, A.Y. (Ed.), Sorokin P.A. Chelovek. Tsivilizatsiya. Obshchestvo (P.A. Sorokin: Man. Civilization. Society). Politizdat, Moscow, pp. 536–539.

Sorokin, P.A., 1928. Contemporary Sociological Theories. Available at: library2.usask.ca/Sorokin/about/theories/method.

Sorokin, P.A., 1975. Hunger as S Factor in Human Affairs. Translated and with a Prologue by EP Sorokin. Edited and with an Introduction by T Lynn Smith. University Press of Florida, Gainesville.

Sorokin, P.A., 1992a. Sotsiologiya 1 (Sociology No. 1). In: Sogomonov, A.Y. (Ed.), P.A. Sorokin. Chelovek. Tsivilizatsiya. Obshchestvo (P.A. Sorokin. Man. Civilization. Society). Politizdat, Moscow, pp. 26–32.

Sorokin, P.A., 1992b. Sotsiologiya revolyutsii (Sociology of revolution). In: Sogomonov, A.Y. (Ed.), Sorokin P.A. Chelovek. Tsivilizatsiya. Obshchestvo (P.A. Sorokin. Man. Civilization. Society). Politizdat, Moscow, pp. 266–294.

Sorokin, P.A., 1997. Glavnye tendentsii nashego vremeni (The Basic Trends of Our Times). Nauka, Moscow.

Sorokin, P.A., 2000. Sotsial'naya i kul'turnaya dinamika (Social and Cultural Dynamics). RGKHI, St Petersburg.

Sorokin, P.A., 2002. Predmet sotsiologii i eyo otnoshenie k drugim naukam (The subject of sociology and its relation to other sciences). In: Ban'kovskaya, S.P. (Ed.), Teoreticheskaya sotsiologiya. Chast' 2 (Theoretical Sociology), vol. 2. Knizny dom 'Universitet', Moscow, pp. 105–135.

Sorokin, P.A., 2005. Sotsial'naya mobil'nost' (Social Mobility). Academia, Moscow.

Sorokin, P.A., 2007. Russia and the United States. Transaction Publishers, New Brunswick, NJ.

Sorokin, P.A., 1914. Prestuplenie i kara, podvig i nagrada (Crime and Punishment, Heroic Service and Reward). Izdatelstvo Dolbysheva, St Petersburg.

Sorokin, P.A., 1920. Sistema Sotsiologii (System of Sociology), 2 vols. Izdatelstvo Kolos, St Petersburg.

Sorokin, P.A., 1921. Golod kak faktor (Hunger as a Factor in Human Affairs). Izdatelstvo Kolos, St Petersburg (destroyed by the Soviet Government).

Sorokin, P.A., 1925. The Sociology of Revolution. J.B. Lippincott, Philadelphia, PA.

Sorokin, P.A., 1927. Social Mobility. Harper & Brothers, New York.

Sorokin, P.A., 1928. Contemporary Sociological Theories. Harper & Brothers, New York.

Sorokin, P.A., 1937–41. Social and Cultural Dynamics, 4 vols. American Book Company, Cincinnati.

Sorokin, P.A., 1941. The Crisis of Our Age. E.P. Dutton & Co., New York.

Sorokin, P.A., 1942. Man and Society in Calamity. E.P. Dutton & Co., New York.

Sorokin, P.A., 1944. Russia and the United States. E.P. Dutton & Co, New York.

Sorokin, P.A., 1947. Society, Culture and Personality. Harper & Brothers, New York.

Sorokin, P.A., 1948. The Reconstruction of Humanity. Beacon Press, Boston, MA.

Sorokin, P.A., 1950a. Altruistic Love. Beacon Press, Boston, MA.

Sorokin, P.A., 1950b. Leaves from a Russian Diary, Enlarged ed. Beacon Press, Boston, MA.

Sorokin, P.A., 1951. S.O.S.: The Meaning of Our Crisis. Beacon Press, Boston, MA.

Sorokin, P.A., 1954. The Ways and Power of Love. Beacon Press, Boston, MA.

Sorokin, P.A., 1956. Fads and Foibles in Modern Sociology and Related Sciences. Henry Regnery Co., Chicago, IL.

Sorokin, P.A., 1956. The American Sex Revolution. Porter Sargent Publishers, Boston, MA.

Sorokin, P.A., 1959. Social and Cultural Mobility. Free Press, Glencoe, IL.

Sorokin, P.A., Lunden, W.A., 1959. Power and Morality: Who Shall Guard the Guardians? Porter Sargent Publishers, Boston, MA.

Sorokin, P.A., 1963. A Long Journey: The Autobiography of Pitirim A. Sorokin. College and University Press, New Haven, CT.

Sorokin, P.A., 1964. The Basic Trends of Our Times. College and University Press, New Haven, CT.

Sorokin, P.A., 1965. Sociology of yesterday, today, and tomorrow. American Sociological Review 30 (6), 833–843.

Sorokin, P.A., 1966. Sociological Theories of Today. Harper & Row, New York.

Sorokin, P.A., 1975. Hunger as S Factor in Human Affairs. Translated and with a Prologue by EP Sorokin. Edited and with an Introduction by T Lynn Smith. University Press of Florida, Gainesville.

Tiryakian, E.A., 2007. Introduction to the transaction edition. In: Sorokin, P.A. (Ed.), Russia and the United States. Transaction Publishers, New Brunswick, NJ, pp. vii–xxviii.

Relevant Websites

http://www2.asanet.org/governance/sorokin.html – American Sociological Association page.

http://www.asanet.org/sections/altruism.cfm and https://sites.google.com/site/altruismmoralitysolidarity/ – Altruism, Morality and Social Solidarity; Section of the American Sociological Association page.

http://community.livejournal.com/pitirim_sorokin/ – Community in LiveJournal (in Russian).

http://www.sorokinfond.ru/ – The expert and analytical portal of the Faculty of Sociology at Lomonosov Moscow State University "Pitirim Sorokin Foundation" (in Russian).

http://www.socio.msu.ru/ – The Faculty of Sociology at Lomonosov Moscow State University page (in Russian).

http://www.misk.newparadigm.ru/; http://www.eng.newparadigm.ru/index.html – Pitirim Sorokin – Nikolay Kondratieff International Institute.

http://www.sorokinfoundation.org or http://www.cliffstreet.org; http://www.pitirimsorokin.org – The Pitirim A. Sorokin Foundation (USA) (in Russian).

http://sor-res-center.livejournal.com – The Pitirim A. Sorokin Foundation in LiveJournal (in Russian).

Source Monitoring

Sarah E MacPherson, Centre for Cognitive Ageing and Cognitive Epidemiology, Human Cognitive Neuroscience, University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Source monitoring is the ability to reconstruct and assess the amount and quality of contextual details that are to be remembered. Source monitoring can be classified in terms of internal, external, or reality monitoring judgments depending on the type of source judgments being made. Evidence from patients and neuroimaging studies, as well as developmental research, suggest source monitoring is associated with the frontal and medial temporal lobes. However, there still remains a debate in the literature whether these brain regions are necessary for source monitoring and whether distinct neural mechanisms underlie different source judgments.

Source monitoring refers to the ability to remember the conditions or context under which a memory was formed. Source monitoring can be classified into three groupings: internal, external, and reality source monitoring. *Internal source monitoring* requires individuals to discriminate between events generated from internal sources of information through their own actions or thoughts (e.g., "Did I actually switch off the oven or only imagine that I did?"), while *external source monitoring* refers to differentiating between memories externally produced by other people (e.g., "Did Sally or Jill switch off the oven?"). *Reality monitoring* is the ability to discriminate between sources of information generated internally and externally either in terms of self- versus other-generated information (e.g., "Did I switch off the oven or did Sally?") or implicit, internally derived aspects of memory (e.g., imagining the act of switching off the oven) versus more explicit, externally derived aspects of memory (e.g., perceiving someone else or switching off the oven yourself). Failures in source monitoring, such as an inability to differentiate between different sources of information (e.g., uncertainty about whether you actually switched off the oven or simply imagined that you did), can have serious consequences on everyday life. Source monitoring has been studied in the research areas of fundamental and applied memory including amnesia, neurological disorders, false memory and eye witness testimony, legal decisions, developmental memory, and cognitive aging.

Cognitive Theories of Source Monitoring

Source Monitoring Framework

The source monitoring framework proposes that item and source memory constitute the episodic memory system (see Johnson et al., 1993). While item memory relates to the ability to recall or recognize information that was presented earlier, source memory involves the recollection of the contextual details of a memory. In the late 1980s and early 1990s, research demonstrated that item memory and source memory were functionally dissociable. In certain situations, individuals who could consistently recall the content of information they learned, were not able to recollect the context associated with learning that information such as how they came to know a particular piece of information. The source monitoring framework proposes that individuals do not necessarily explicitly associate their memories with context. Instead, they retrieve and assess an association between the qualitative aspects of a memory (e.g., the perceptual, spatial, or temporal details, semantic concepts, emotional reactions, or the *cognitive operations* associated with elaborating and organizing information during encoding). In making source monitoring judgments, individuals evaluate the amount and quality of these details compared to an average or anticipated measure of detail for a particular memory type. Different types of source monitoring judgments vary in terms of these details depending on the task context, an individual's intentions, past experience, etc. External sources of information tend to have more perceptual, spatial, or temporal features (e.g., remembering the word was red) but fewer cognitive operations compared to internal sources of information (e.g., remembering that you thought about something red when the word was presented). Perceptually experienced situations may have more egocentric rather than allocentric spatial features compared to internal, imagined events. Source monitoring judgments are typically made using fast, automatic heuristic decision-making processes, although slower, controlled systematic processes can also be used. Both heuristic and controlled processes are influenced by prior knowledge and beliefs (e.g., the qualitative details a source memory of that type should have). In addition, the social and cultural circumstances in which memories are assessed influence what we remember, the frequency we do so, and what we consider to be fitting evidence that an event has actually been remembered. For example, a witness to a crime may focus on specific details of the perpetrator's appearance, where and when the crime took place and whether the details are credible. In contrast, a person discussing their wedding day might concentrate less on particular perceptual and temporal details and more on familiarity and emotional details. While the details in both situations may be encoded in the same way initially, distinct characteristics such as the perceptual and the emotional aspects may be weighted differently and the extent to which heuristic and systematic processing influence the memory differs depending on the required social situation. Furthermore, as source monitoring is typically based on judgment processes rather than on direct retrieval, motivational features may also influence the details that are first attended to

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51062-9

and then retrieved or given the most importance to after the event when later remembering.

Source Monitoring Assessment

Various paradigms have been devised to assess the different types of source monitoring judgments. To assess external source monitoring, individuals are generally presented with items that differ in terms of certain features during the study phase which might include location (e.g., above or below or to the left or the right of a cross in the center of a computer screen), time period (e.g., series one or series two of a study list), form (e.g., male vs female speakers or red vs blue fonts), or sensory modality (e.g., oral vs visual). At test, participants are then required to remember the contextual information associated with the to-be-remembered items. In internal source monitoring, individuals are required to discriminate between memories generated from internal sources (e.g., words generated from anagrams vs words imagined on a computer screen). In reality monitoring, individuals have to determine whether items were generated internally or externally in terms of the self (e.g., words they said or imagined) versus other (e.g., words someone else said) or implicitly (e.g., generated from an anagram or imagined) versus explicitly (e.g., words or pictures presented on a computer screen).

Although these source monitoring tasks explicitly focus on identifying the source of information, there are other tasks that are also thought to involve source monitoring processes such as remember/know judgments; context, relational, or associative memory tasks; memory binding tasks; inclusion/exclusion tasks; list discrimination; and any recall and recognition tasks, which require information from a particular source to be identified. The conditions for making source monitoring decisions vary and depend on the purpose of the source judgment at the time it is made.

Source Monitoring Errors

On occasions, individuals fail to differentiate between what actually took place and what they deduced or imagined, what they perceived and what was proposed to them, the actions of different individuals, and newly learned and previously known information. Difficulties with source monitoring occur because source judgments are based on evaluations of the details to be remembered rather than explicitly recalled memories. If the information is poorly encoded and, as a result, the information is incomplete or unclear, or there is disruption in the evaluation or retrieval of the memory, source monitoring errors can arise.

Source monitoring errors are also more likely to occur when there is a high degree of similarity between memories from two sources. Potential similarities might include two voices of the same gender, when words that are perceived or imagined are synonyms (e.g., jewel or gem) or line drawings that are perceived or imagined are visually similar. When individuals make these source errors, they are also inclined to borrow details from the encountered items including their color or their position on a computer screen. This means that associations between item memories can also influence source memory judgments. Other reasons for source errors might include a more relaxed criterion being used to evaluate memories, fewer qualitative aspects being used or details incorrectly evaluated, or source judgments being made under time pressure.

The Generation Effect

In memory research, it is generally agreed that asking individuals to generate the information they have to remember results in the information being better remembered than nongenerated information. This is known as the *generation effect*. Self-generation can be achieved through the generation of words from fragments, antonyms, or synonyms, keywords from paragraphs or pictures, or the answers to numerical problems. The process of generating items is thought to strengthen the encoding of the item-specific, semantic, or relational features of the stimulus, which results in better memory for those items (see Mulligan and Lozito, 2004 for a review).

Yet, whether this generation effect can also be applied to source monitoring judgments remains a topic of debate as generating words does not always lead to positive benefits for source memory. In some instances, an item-context trade-off is found where generating answers to questions benefits item memory, whereas reading statements benefits source memory. Therefore, item and source memory processes may be dissociable where generation provides benefits to item memory at the expense of source memory. Or it may be that different encoding strategies affect contextual details differently, as reading highlights the importance of the perceptual details of an item (e.g., color or font) whereas generation underlines the importance of conceptual or nonvisual information (e.g., location). A final consideration is that generation effects for source memory depend upon the type of source monitoring task adopted, as positive generation effects occur more often in reality-monitoring tasks, whereas negative generation effects occur more often in external source monitoring tasks.

Research has also focused on the contribution of cognitive operations to the generation effect. Early work in relation to the generation effect focused on cognitive effort rather than cognitive operations, where the more effort associated with generating items, the stronger or more characteristic (and therefore more easily accessible) the memory traces compared to nongenerated items. In addition, remembering the cognitive operations involved at study might benefit the retrieval of the generated information. The size of the generation effect (i.e., generated vs nongenerated items) was thought to be proportionate to the cognitive effort necessary to produce these generated items.

However, as the generation effect is not always found for more difficult items that require more cognitive effort or nonwords, cognitive effort cannot fully account for the generation effect. This meant that the literature moved away from focusing on cognitive effort, and toward a focus on cognitive operations. Cognitive operations might include processes that influence the reading of the items presented by the experimenter or influence the search and decision strategies

important for generating the materials (see Johnson et al., 1993). Harder encoding conditions involving more cognitive operations to generate items compared to easy encoding conditions (e.g., Was an anagram created or deciphered?) result in more successful source monitoring judgments. However, the same advantage is not found when individuals are asked to judge whether the anagrams are easy or difficult during the source monitoring test (see Foley and Foley, 2007). Overall, these findings suggest that both the encoding task and the cognitive operations involved in source monitoring tasks influence the strength and direction of the generation effect.

Source Monitoring Framework versus Dual-Process Theories

While dual-process theories of memory are discussed in the section on recognition memory, it is important at this juncture to discuss how such models differ from the source memory framework, as dual-process theories of memory are also thought to influence source monitoring (e.g., Eichenbaum et al., 2007; Wixted, 2007). Several dual-process models argue that episodic memory retrieval is dependent on two independent processes: recollection, which involves remembering certain details about the experience that the item was encountered and familiarity, which is a feeling that an item has already been presented but no additional information can be retrieved (Jacoby, 1991). Recollection is considered an all-or-none, threshold process where someone can either recollect information or not, whereas familiarity is a continuous, graded process that can vary by degrees. Successful source monitoring judgments are thought to rely on recollection where individuals make decisions about the contextual characteristics of a previously presented item (e.g., temporal, spatial, semantic). Both the dual-process and source monitoring framework models agree that recollection and familiarity can contribute to a memory judgment. However, while the dual-process models of memory consider recollection and familiarity as two separate processes that give rise to subjective experience, the source monitoring framework refers to recollection and familiarity as types of subjective experience and assumes that both are continuous, graded processes. Often the proportion of correct source monitoring judgments associated with recollection- and familiarity-based judgments does not differ, which suggests that the absence of recollection details does not prevent correct source attributions and familiarity also plays a role in source monitoring. Nonetheless, the dual-process models have mainly focused on situations where recollection is successful or not.

Dual-process theories of memory also differ from the source monitoring framework in terms of false memories. The source monitoring framework emphasizes that memories are based on features that are bound together during a particular event and the memories we produce are instilled with our knowledge, ideas, desires, and aspects of our imagination. According to the source monitoring framework, true and false memories can be explained using the same broad principles of memory (see Lindsay, 2008), and so recollection does not guarantee that accurate memories have been retrieved. Inaccurate source monitoring judgments involve source information that is generated by the self or from prior knowledge or information that is introduced from other items. In contrast, dual-process theories claim that inaccurate source monitoring judgments occur when the memory decision is based on familiarity and not recollection (e.g., Wais et al., 2008). As dual-process models tend to focus on the divergence between recollection and familiarity, the source monitoring framework focuses on explaining the details of true versus false recollection; this has led to some of the noticeable distinctions between the two theories.

Source Monitoring for Emotional Context

Another important issue associated with source monitoring is the influence of emotional information on source monitoring judgments. An individual who survives a life-threatening car accident might remember external contextual details directly associated with the emotional event (e.g., the other car), external details not directly associated with the emotional event (e.g., the road they were driving along), or internal details associated with their thoughts and feelings (e.g., feeling faint, worrying about damage to the car). Despite the fact that remembering emotional source information may often benefit individuals when they have to differentiate between their own and other people's memories, the processing of emotional information may also occur at the detriment of other information that provides better source cues. Indeed, while emotional items tend to be recalled and recognized better than neutral items, the influence of emotion on source memory varies depending on the type of source monitoring judgments being made. Source monitoring has been shown to improve when individuals have to remember the perceptual features of emotion words (e.g., font color) and yet decline for other contextual features (e.g., location or whether the information was seen or heard). It has also been shown that when participants emotionally focus on the self or another person when encoding emotional statements provided by one of the two speakers (e.g., how the participant feels about what two speakers are saying or how the speakers themselves are feeling), item memory is better for the self-focus condition compared to the other-focus condition, whereas source memory is poorer for the self-focus condition (e.g., Johnson et al., 1996). As individuals concentrate on the inner self and how they feel, emotional information is thought to draw attention away from external, perceptual information such as voice quality. This reduces the probability that the participant will bind the perceptual attributes of the speaker (e.g., voice sound, voice quality) to the content of statement made, resulting in a negative effect on source memory. Alternatively, if individuals focus on the other person's emotions while encoding information, the binding of the perceptual features associated with the speaker and the content of the statement allows for accurate source monitoring judgments (see Kensinger, 2007).

Patient Studies and Source Monitoring

Up until this point, the review has focused on source monitoring abilities in healthy individuals. However, to enhance our understanding of source monitoring processes, it is also

important to discuss the effects brain damage can have on source monitoring abilities. It was the early twentieth century when Claparède, the Swiss neurologist, first reported a case of alcoholic Korsakoff's syndrome who had a profound amnesia for recent events. More importantly, even when the patient was able to recall factual information, she did not believe that she had learned the information from Claparède himself and would adamantly dispute the fact that she had been read a story to recall. This inability to remember the origin of learned factual information and associating it with learning prior to, rather than during, the experiment is often referred to as *source amnesia*.

Since then, source errors have been reported in a number of single-case reports demonstrating dissociations between item memory and source memory. This led Schacter et al. (1984) to develop a paradigm for investigating source monitoring in patient groups. In this paradigm, individuals are asked questions regarding fictional characteristics about famous and unknown individuals by one of two experimenters and the correct answer is then provided. The question is then repeated again at a later point by one of the two experimenters and the participants not only have to provide the correct answer to the question, but also how they learned that information. This paradigm demonstrated that even if amnesic patients are able to recall learned fictitious factual knowledge, source recall remains low and patients tend to attribute learning knowledge to another situation rather than from one of the two experimenters within the experiment itself. However, not all amnesic patients demonstrate source amnesia, and even in those that do, its severity is not necessarily associated with the severity of their memory deficit (e.g., Shimamura and Squire, 1987). Therefore, amnesic patients can show a severe impairment in item memory in the absence of source monitoring deficits.

Medial Temporal Lobes

In terms of the neural basis of source monitoring, the medial temporal lobe, in particular the hippocampus, is thought to play an important role in episodic memory. However, research suggests that the role of the medial temporal lobe is to support both item and source memory to the same extent and the medial temporal lobe is not *necessary* for source monitoring performance. Although some medial temporal lobe patients demonstrate impaired source monitoring abilities when matched with healthy controls on item memory, there are also instances where patients with more focal hippocampal lesions do not necessarily differ from controls in terms of source monitoring performance. Furthermore, the lack of an association between item memory and source amnesia in amnesic patients led Shimamura and Squire (1987) to conclude that source monitoring relies on brain regions other than the medial temporal lobe.

Frontal Lobes

Frontal lobe lesions have also been associated with deficits in source monitoring abilities. A number of studies have demonstrated that item and source memory are dissociable in terms of frontal lobe involvement, as impairments in source memory can be found in the absence of item memory deficits. Luria (1976) first reported a frontal patient with source amnesia due to the rupture of an anterior communicating artery aneurysm. While the patient could remember picture-word paired associates 2 days and 1 week after learning them, he had no knowledge of how he came to learn these associations. Therefore, his memory for facts was intact but he had impaired memory for source information.

Luria's single-case report led to a number of studies investigating source monitoring in groups of frontal patients. Using a similar paradigm to that of Schacter et al. (1984), Janowsky et al. (1989) conducted one of the first source memory studies involving frontal patients. They demonstrated that although frontal patients could learn new information, they incorrectly attributed the source of that information more often than controls. These findings provided further support that patients with focal frontal lobe lesions are able to learn and remember new information but not remember the context in which the information was learned.

The role that the frontal lobes play in source monitoring decisions has been considered in terms of Johnson et al.'s (1993) source monitoring framework. Source monitoring is thought to require varying amounts of strategic search and judgment depending on how well the information is encoded. As frontal lobe lesions are thought to impair the ability to strategically search during memory retrieval, frontal-related deficits in source monitoring may be secondary to the strategic processing demands of the memory tasks involved. Indeed, when the strategic demands of the source monitoring tasks are reduced, frontal patients do not appear to be significantly impaired on item or source memory.

Furthermore, frontal lobe lesions do not always result in source monitoring deficits. There are also reports of amnesic patients whose source amnesia is more related to their amnesia severity rather than their performance on measures of frontal lobe function – although there are other studies that have reported significant correlations between source monitoring performance and measures of frontal lobe function. As the frontal lobes are a large, heterogeneous brain region, it may be that only certain subregions within the frontal lobes are associated with source monitoring. However, few patient studies have examined whether this is the case or not. Dorsolateral prefrontal lesions have resulted in difficulty remembering temporal contextual details compared to frontal patients with more medial lesions. Yet, ventromedial prefrontal lesions produce impairments in memory for perceptual but not semantic features. Therefore, different neural mechanisms may underlie distinct contextual features but further work is necessary to ascertain what these are.

Source Monitoring and Confabulation

Confabulation is a clinical syndrome, which is associated with orbitofrontal and medial frontal lobe lesions and typically results in source monitoring deficits. Confabulations are usually produced by amnesic patients with ruptured anterior communicating artery aneurysms, although they are also associated with other clinical syndromes such as Alzheimer's disease, frontotemporal dementia, Korsakoff's syndrome, neglect, and schizophrenia (for a review see Johnson et al.,

2000). Confabulations are classified as false accounts or statements about the self and/or the world that inadvertently occur due to an underlying pathological condition. Different accounts have been proposed to explain confabulation including deficits in memory control processes (strategic retrieval hypothesis: e.g., Burgess and Shallice, 1996) or confusion of the temporal context of events (temporality accounts: e.g., Schnider, 2003) or failure to distinguish between real and imagined events (source or reality monitoring accounts: e.g., Johnson et al., 1993). Those supporters of the reality monitoring account of confabulation state that confabulators cannot source their memories in terms of where or when the event took place and this leads to errors. Clinical descriptions of confabulating amnesiacs tend to be consistent with the notion that reality monitoring engages in both heuristic and systematic decision-making processes (Johnson, 1991). Reality monitoring failures associated with heuristic errors are produced by unusually striking mental experiences or a relaxed criterion being used to evaluate the retrieved memory, making it more likely the memory will be judged as real. Initially, the qualitative aspects of the memory may pass the heuristic reality-monitoring assessment, but then they are properly assessed by more systematic processes that test whether the memory is reliable, credible, and reasonable. Bizarre, implausible confabulations are thought to be the result of disturbances in systematic processing where vivid thoughts that have passed the heuristic reality monitoring examination are not identified.

Yet, the source monitoring account cannot sufficiently explain confabulation, as it cannot explain confabulations that only occur in episodic and not semantic memory. According to the source monitoring framework, confabulation should occur when any decision about the source or reality of a memory is needed. However, source monitoring framework theories have argued that there are differences in task difficulty in terms of well-encoded semantic information and more complex personal episodic memory and this is why some confabulators only show deficits on episodic memory tasks. Research has shown that confabulators have source monitoring deficits in terms of temporal order, spatial features, and distinguishing between the visual and auditory modality. Nonetheless, such source monitoring tasks do not differentiate confabulators from nonconfabulating amnesiacs and source monitoring deficits are found in other brain damaged groups who do not confabulate. Johnson and her colleagues point out that they are not suggesting that source monitoring deficits are the only impairments associated with confabulation but only one of a number of deficits that are necessary for confabulation to occur.

Source Monitoring and Neuroimaging

Neuroimaging studies involving functional magnetic resonance imaging provide another source of evidence for the mapping of source monitoring processes onto different neural mechanisms (see Mitchell and Johnson, 2009). Activation in the hippocampus (and parahippocampal region) increases when encoding and retrieving items for which the source is later successfully remembered compared to when the source is not successfully remembered. This activation is thought to be common to words, objects, and scenes, both at study and test, suggesting that these source memory effects are domain-general. In contrast, material-dependent source memory effects at study, but not test, have been reported for objects only in the left perirhinal cortex. Furthermore, the majority of neuroimaging studies provide evidence for medial temporal lobe activation during recollection-based judgments, but in the perirhinal cortex during familiarity-based judgments.

Frontal Lobes

In terms of frontal lobe activation, neuroimaging studies also suggest that the frontal lobes are associated more with source than item memory. More specifically, the dorsolateral prefrontal cortex is activated while encoding and retrieving source information (e.g., color, position) compared to item information. Work has also shown that this prefrontal activation is not limited to the type of features to be remembered, suggesting that the dorsolateral prefrontal cortex is involved in domain-general decision-making (e.g., Dobbins and Han, 2006).

While functional magnetic resonance imaging studies suggest that subregions within the frontal lobes play distinct roles in source monitoring, a conclusion to what these different roles are has yet to be reached. In some work, the left prefrontal areas have been associated with systematic source monitoring processes or strategically using contextual information to make a decision (i.e., recollection), while the right prefrontal areas have been associated with heuristic source monitoring processes or familiarity-based decisions where individuals have a decontextualized feeling that an item has previously been presented. Other research has suggested that the right dorsolateral prefrontal cortex is important for controlled, systematic postretrieval monitoring processes. Therefore, the question of whether it is the left or the right dorsolateral prefrontal cortex that is involved in systematic source monitoring processes remains unanswered.

In addition to dorsolateral prefrontal activation, the medial prefrontal cortex is activated during *origin* but not *context* reality monitoring. Origin reality monitoring is where self versus other judgments are made while, context reality monitoring is where perceived versus imagined source judgments are made. In particular, origin monitoring is thought to be associated with significantly more activation in the rostral prefrontal cortex (Brodmann area 10) compared to context monitoring (Simons et al., 2008). This is not surprising given that Brodmann area 10 is thought to be the gateway between processing perceptual or reflective information, which is likely to originate from the self.

Event-Related Potentials

Imaging methods measuring electrical brain activity have also been used to examine source monitoring processes where memory-related and event-related potentials can be discriminated between in terms of temporal onset (i.e., the time after test probe) and spatial location. One of the largest amplitude differences between target and distracter items in source event-related potential studies emerges approximately 800 ms after

the probe is presented in the right frontal region. This is referred to as the *right frontal effect*. This component is thought to reflect postretrieval processes or monitoring processes that are analogous with the decision processes proposed by the source monitoring framework. In addition, an earlier onset old/new event-related potential component at the left parietal electrode sites has also been reliably reported, which has a temporal onset approximately 600–800 ms after the probe is presented. This is referred to as the *parietal effect*. It is linked with activation in left inferior parietal cortex and is thought to reflect recollection. Finally, a third event-related potential component is reported in some source monitoring studies, which is referred to as the *late posterior negativity effect* and has a temporal onset of approximately 1000 ms. This late posterior negativity effect has a more negative amplitude for targets compared to distracters with the largest amplitude difference at posterior electrodes sites. The late posterior negativity effect is thought to reflect further examination of the combinations of features retrieved during demanding source judgments. However, although these three event-related potential components are associated with old/new source judgments, when they are examined in terms of correct versus incorrect source judgments, only the increase in the magnitude of the parietal effect has been associated with accurate source monitoring judgments.

While the majority of event-related potential studies have investigated the components of external source monitoring, there are also studies that have focused on reality monitoring. These studies have used paradigms that typically compare words generated versus heard, actions performed versus watched, and perceived versus imagined pictures. A larger parietal effect has been associated with internally generated sources compared to externally generated. However, it remains uncertain whether this parietal effect is because internally generated items encourage more vivid, strategic recollection or they involve more cognitive operations to create the mental representations for that event which individuals can direct themselves toward during remembering. Externally generated information and distracters do not have this targeted information.

In contrast, the right frontal effect is less consistently reported in reality monitoring studies, with only some event-related potential studies reporting bilateral frontal effects associated with making old/new judgments. It has been suggested that the lack of an old/new frontal effect might be because new items produce more positive event-related potentials during reality monitoring compared to externally generated items. Indeed, neuroimaging has reported left anterior ventrolateral prefrontal activation and medial anterior prefrontal activation when making internal source monitoring judgments. Further work is needed to determine whether the discrete frontal event-related potentials across studies are due to different qualitative features of internal sources or more general monitoring processes that are associated with distinct prefrontal regions.

Developmental Studies of Source Monitoring

Source Monitoring in Children

Further clues toward understanding source monitoring processes can be taken from research investigating how source monitoring abilities develop and change with age (see Johnson et al., 1993; Newcombe et al., 2007). A child-friendly adaptation of Schacter et al.'s (1984) paradigm has shown that 4-year olds produce a significantly greater number of extra-experimental errors than 6- and 8-year olds where they claimed they had learned the information outside the experimental situation (Drummey and Newcombe, 2002). As source amnesia is associated with frontal lobe dysfunction, these changes in source monitoring abilities between 4 and 6 years of age may reflect the maturation of function within the prefrontal cortex. Other developmental studies have compared different types of source monitoring abilities within the same group of young children and have shown that, in some paradigms, they are able to perform as well as adults. In studies asking children to perform, imagine, or perceive someone else performing an action, children do not make external monitoring errors (i.e., distinguishing between two individuals performing particular actions) or reality monitoring errors (i.e., judging whether they performed the action or someone else). However, they are more likely to make internal source monitoring errors where they mix up performing and imagining the action themselves (e.g., Foley and Johnson, 1985). This difficulty, children have with internal monitoring judgments, is likely to be due to the difficulty in distinguishing between similar sources (e.g., Lindsay et al., 1991).

Source Monitoring in Healthy Adult Aging

At the other end of the developmental spectrum, the majority of studies on healthy adult aging have demonstrated that age has a detrimental effect on source monitoring abilities (see Mitchell and Johnson, 2009). Although older adults are able to remember learned fictitious general knowledge facts, they have difficulty remembering whether these facts are learned during or prior to the experiment. Like young children, they are more likely to make source amnesia errors where they attribute information learned within the experiment as being learned outside the experiment. However, age-related source monitoring deficits are not reported in all cases and only apply to particular types of source monitoring. Indeed, significant age differences have been reported when older adults make internal and external source monitoring judgments but not reality monitoring judgments. It is thought that the cognitive operations associated with remembering that an item was produced rather than observed results in different cues, allowing both younger and older adults to successfully distinguish between these conditions. Contrastingly, the cognitive operations associated with remembering an item was produced or imagined are expected to be more alike, thus reducing their success as a cue in distinguishing between what one said from what one only imagined saying. In terms of differentiating between two voices, these judgments rely on the evaluation of perceptual information, as the quantity or type of cognitive operations would not provide a reliable cue for these types of source monitoring judgments. Therefore, in situations where cognitive operations are similar or not beneficial, older adults have to rely on perceptual information and this makes it more difficult to make these types of source monitoring distinctions.

Source memory deficits in older adults may potentially be explained by age-related changes in the brain regions that play

a role in source monitoring such as the medial temporal lobes, prefrontal cortex, and posterior parietal cortex (see Mitchell and Johnson, 2009). When source memory for various types of material is examined in younger and older adults, functional magnetic resonance imaging data have demonstrated that both younger and older adults recruit similar networks of brain regions regardless of the material type, and this includes the parietal cortex, prefrontal cortex, and the hippocampus. Age-related differences in material-independent activation are associated with the right lateral prefrontal cortex, and increases in activation in the perirhinal cortex are associated with improved performance in older adults.

Source Monitoring in Alzheimer's Disease

Research has also examined source monitoring impairments associated with pathological aging. The majority of studies have shown that Alzheimer's disease produces reality monitoring deficits where Alzheimer's disease patients, and to a lesser extent healthy older adults, have been reported to find remembering imagining another person performing an action significantly more difficult than self-performed, self-imagined, and other-performed conditions (e.g., Fairfield and Mammarella, 2009). It is less clear whether Alzheimer's disease affects other aspects of source monitoring. In some cases, Alzheimer's disease patients are not impaired in terms of external source monitoring where they might be asked to remember which one of the two people provided them with information. They are also not reported to show internal source monitoring deficits as they are able to distinguish between self-performed actions or self-imagined actions, or sentences that they themselves have completed or already completed sentences they have read. However, there are also instances where Alzheimer's disease patients are impaired in terms of external and internal source monitoring as they perform poorly on perceptual, spatial and temporal source monitoring tasks with the spatial source monitoring task being more sensitive to dementia. Therefore, while Alzheimer's disease patients are not impaired compared to healthy older adults on all aspects of source monitoring, the degree to which they are impaired on different tasks remains an open question.

In Summary

Research in the areas of basic and applied memory has provided, and will continue to provide, a wealth of experimental data attempting to understand the mechanisms fundamental to source monitoring. According to the source monitoring framework (Johnson et al., 1993), source monitoring relies on the reconstruction and evaluation of the qualitative details of a memory rather than explicitly associating memories with context. This can be done using fast, automatic heuristic decision-making processes, as well as slower, controlled decision-making processes. Lesions in both the medial temporal and frontal lobes have been found to result in source monitoring impairments, although it is less clear whether these brain regions are *necessary* for source monitoring. More recently, functional magnetic resonance imaging and event-related potential studies in healthy individuals have provided further evidence to support frontal lobe involvement in source monitoring judgments. Future work faces the difficult challenge of determining which subregions within the frontal lobes are associated with distinct source monitoring processes. It is only through combining the findings from neuropsychological, functional magnetic resonance imaging, and event-related potential studies that a better understanding of the neural mechanisms that underlie source monitoring might be achieved.

See also: Aging and Memory; Amnesia: General; Confabulation; Context Dependent Memory; Declarative Memory, Neural Basis of; Dysexecutive Amnesia; Episodic Memory; Eyewitness Testimony; False Memories; Flashbulb Memories; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Memory Binding; Post Traumatic Amnesia; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Recognition; Retrieval.

Bibliography

Burgess, P.W., Shallice, T., 1996. Confabulation and the control of recollection. Memory 4, 359–411.
Dobbins, I.G., Han, S., 2006. Cue-versus probe-dependent prefrontal cortex activity during contextual remembering. Journal of Cognitive Neuroscience 18, 1439–1452.
Drummey, A.B., Newcombe, N.S., 2002. Developmental changes in source memory. Developmental Science 5, 502–513.
Eichenbaum, H., Yonelinas, A.R., Ranganath, C., 2007. The medial temporal lobe and recognition memory. Annual Review of Neuroscience 30, 123–152.
Fairfield, B., Mammarella, N., 2009. The role of cognitive operations in reality monitoring: a study with healthy older adults and Alzheimer's-type dementia. The Journal of General Psychology 136, 21–39.
Foley, M.A., Foley, H.J., 2007. Source-monitoring judgments about anagrams and their solutions: evidence for the role of cognitive operations information in memory. Memory & Cognition 35 (2), 211–221.
Foley, M.A., Johnson, M.K., 1985. Confusion between memories for performed and imagined actions. Child Development 56, 1145–1155.
Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language 30, 513–541.
Janowsky, J.S., Shimamura, A.P., Squire, L., 1989. Memory and metamemory: comparisons between patients with frontal lobe lesions and amnesic patients. Psychobiology 17, 3–11.
Johnson, M.K., 1991. Reflection, reality monitoring, and the self. In: Kunzendorf, R. (Ed.), Mental Imagery. Plenum, New York, pp. 3–16.
Johnson, M.K., Hayes, S.M., D'Esposito, M., Raye, C.L., 2000. Confabulation. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, Memory and Its Disorders, vol. 2. Elsevier Science, Amsterdam, pp. 383–407.
Johnson, M.K., Hashtroudi, S., Lindsay, D.S., 1993. Source monitoring. Psychological Bulletin 114, 3–28.
Johnson, M.K., Nolde, S.F., De Leonardis, D.M., 1996. Emotional focus and source monitoring. Journal of Memory and Language 35, 135–156.
Kensinger, E.A., 2007. Negative emotion enhances memory accuracy: behavioral and neuroimaging evidence. Current Directions in Psychological Science 16 (4), 213–218.
Lindsay, D.S., 2008. Source monitoring. In: Byrne, J., Roediger III, H.L. (Eds.), Cognitive Psychology of Memory. Learning and Memory: A Comprehensive Reference, vol. 2. Elsevier, Oxford, pp. 325–348.
Lindsay, D.S., Johnson, M.K., Kwon, P., 1991. Developmental changes in memory source monitoring. Journal of Experimental Child Psychology 52, 297–318.
Luria, A.R., 1976. Cognitive Development, Its Cultural and Social Foundations. Harvard University Press, Cambridge.
Mitchell, K.J., Johnson, M.K., 2009. Source monitoring 15 years later: what have we learned from fMRI about the neural mechanisms of source memory? Psychological Bulletin 135 (4), 638–677.
Mulligan, N.W., Lozito, J.P., 2004. Self-generation and memory. In: Ross, B. (Ed.), Psychology of Learning and Motivation, vol. 45. Elsevier, San Diego, CA, pp. 175–214.

Newcombe, N.S., Lloyd, M.E., Ratliff, K.R., 2007. Development of episodic and autobiographical memory: a cognitive neuroscience perspective. In: Kail, R.V. (Ed.), Advances in Child Development and Behavior, vol. 35. Elsevier, San Diego, CA, pp. 37–85.

Schacter, D.L., Harbluck, J.L., McLachlan, D.R., 1984. Retrieval without recollection: an experimental analysis of source amnesia. Journal of Verbal Learning and Verbal Behavior 23, 593–611.

Schnider, A., 2003. Spontaneous confabulation and the adaptation of thought to ongoing reality. Nature Reviews Neuroscience 4, 662–671.

Shimamura, A.P., Squire, L.R., 1987. A neuropsychological study of fact memory and source amnesia. Journal of Experimental Psychology: Learning, Memory, & Cognition 13, 464–473.

Simons, J.S., Henson, R.N.A., Gilbert, S.J., Fletcher, P.C., 2008. Separable forms of reality monitoring supported by anterior prefrontal cortex. Journal of Cognitive Neuroscience 20, 447–457.

Wais, P.E., Mickes, L., Wixted, J.T., 2008. Remember/know judgments probe degree of recollection. Journal of Cognitive Neuroscience 20, 400–405.

Wixted, J.T., 2007. Dual-process theory and signal-detection theory of recognition memory. Psychological Review 114 (1), 152–176.

Relevant Websites

http://psychology.ucdavis.edu/labs/Yonelinas/PWT/index.cfm – Andrew Yonelinas: Human Memory Lab.

http://socrates.berkeley.edu/~shimlab/index.html – Arthur Shimamura Lab.

http://www.memlab.psychol.cam.ac.uk/ – Jon Simons Memory Lab.

http://memlab0.eng.yale.edu/ – Marcia K. Johnson: The Memory and Cognition Lab.

http://www.psychology.sunysb.edu/nfranklin-/?p=monitoring – Professor Nancy Franklin: Source Monitoring & False Memory.

http://web.uvic.ca/~dslind/?q=home – Steve Lindsay's Lab.

South America, Archaeology of

Dante Angelo, Universidad de Tarapaca, Aríca, Chile
John H Walker, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The archaeology of South America provides long-term context for problems in many disciplines, including history, anthropology, linguistics, ethnohistory, political science, biology, and ecology. Using a comparison between the Andes Mountains and the Amazon River, this article briefly reviews the history of the discipline, and outlines recent developments. The placement of South American archaeology is considered within scholarship both in South America and around the world.

Introduction

Archaeology is of interest to all scholars of South America because the continent's heritage extends long before historical records. Before the arrival of Europeans at the end of the fifteenth century CE, South American societies included both centralized states with organized political, administrative, military, religious, and economic institutions as well as groups that resisted those same states and/or avoided those institutions. Although it is possible to stretch and cut South American examples to fit universal models of how society or culture works, the archaeological record suggests that the historical particularities of those examples retain explanatory value (Silverman and Isbell, 2008).

The geography of South America can be startling to the outside observer: the Andes Mountains cover a distance equal to that between London and Mumbai, and the Amazon River is as long as a trip from New York to Berlin. These two features provide a simple structure for reviewing archaeological scholarship: Andean archaeology has a longer and richer history whereas Amazonian archaeology is much younger.

South American archaeology is useful to archaeologists from elsewhere in the world, as South America has long been a source of examples and evidence that confounds archaeological models, including classifications of societies and interpretations of process. Early monumental architecture along the Pacific coast is an excellent example, in which groups of people coalesced to erect stone temples as early as 6000 years ago, even in the absence of agriculture or pottery. Many attempts to classify and organize societies around lists of important traits founders on the archaeological record of South America. For example, civilization was once defined by lists of traits, including the use of a system of writing, but the Andean empire of Tawantinsuyu (or Inca) made this criterion irrelevant. The archaeological record is created and used in the present, with archaeologists as one of several kinds of interested parties creating and using it for their own reasons. This rhetorical stance of archaeology makes it more difficult for archaeologists to pronounce the unquestioned truth. The integration of various stakeholders into the conduct and presentation of archaeological research is a central theme in South American archaeology today (*see* Archaeology and Cultural Memory; Ethnic Identity and Ethnicity in Archaeology).

The Amazon and the Andes were central in the definition of cultural areas and regional subdivisions in the broader field of anthropology (Steward, 1946). Exploration of pre-Hispanic heritage began in the second half of the nineteenth century and seems to be tightly, although not always explicitly, linked to the formation of South American nation-states and the consolidation of the industrial era and growth of capitalist economies. On the one hand, having gained independence from Spain, the new nation-states sought to establish their legitimacy as descendants of glorious past cultures and societies. On the other hand, the explorations of these new territories derived from the need of European powers and a burgeoning United States to secure and allocate sources of raw materials for production to sustain thriving industrial economies.

Most nineteenth-century and early twentieth-century expeditions included aristocrats and members of European royal lineages – people wealthy enough to fund their own research. Others were sponsored by different emerging academic institutions and scientific organizations that attempted to include this vast territory within the limits of scientific knowledge. The first research of some of the most emblematic archaeological sites, like Tiwanaku in Bolivia as well as Pachacamac and Machu Picchu in Peru, resulted from these experiences. These expeditions established the foundations of Andean archaeology, and provided the first interpretations of cultural and social variability. Within a debate between evolutionism and historical particularism (*see* Archaeology, Theory in), the historical sequences and chronological frameworks for pre-Hispanic societies established the debates that have guided archaeological research since. Topics such as the peopling of the Americas, the origins of civilizations, relationships of exchange and interaction between peoples of different areas and regions, the processes that led to the domestication of plants and animals or to the development of various kinds of technology (architecture, metallurgy, agriculture, and others) – which therefore influence the development of these societies – remain among the main research questions in Andean archaeology.

The ecological diversity that characterizes the vast region, corresponding to variegated social and cultural historical as well

 http://dx.doi.org/10.1016/B978-0-08-097086-8.13002-8

as linguistic developments, promptly posed a challenge to scholars' explanation of the past. In 1931, Steward was commissioned by the Smithsonian's Bureau of American Ethnography to compile and edit a comprehensive ethnographic reference work on indigenous peoples of South America. The result of his work was an authoritative collection of essays written by renowned scholars that defined ethnographic and archaeological issues and was seminal in the definition of cultural areas. These areas were initially outlined as tentative (Steward, 1946–1950: 228), mostly following ecological criteria although resembling modern geopolitical divisions; this proved the importance of geophysical aspects in the partition of the old colonial territories as part of the process of nation building and explains archaeology's relevance for identity and nationalist enterprises. Later contributions, such as *La Arqueología de la América Andina* (Lumbreras and Batres, 1981), which resulted from an international workshop organized by UNESCO, were crucial in fine-tuning Steward's schema.

Dating the earliest presence of human populations in the Americas has been an ongoing concern of archaeologists. Early on, the debate about the peopling of the Americas was framed within a broader dispute about polygenetic and monogenetic theories of evolution during the last part of the nineteenth century. As a result, this controversy has fueled an intense, even passionate debate about the antiquity of South America's population (*see* North America, Archaeology of). Although debate about this subject continues today, important findings made at Monte Verde in southern Chile (Dillehay, 1997) and more recently on the Pacific coast and lowlands of Bolivian (Lombardo et al., 2013) have shown convincing evidence of earlier peopling of the continent as well as the diversity and complexity of the first South Americans. Most of the evidence found in such sites suggests that South America was populated in different migratory waves, colonizing different regions and habitats, and developing various social and technological strategies of subsistence. Adapting to the climatic changes that took place during the early Holocene (12 000–8000 BCE), groups of foragers sought resources in a shifting and ecologically diverse environment, ranging from the pampas and plains of Patagonia to the Andean mountain chain and its basins, and the lowlands of the Amazon Basin.

The exploitation and consumption of biotic resources, which include a wide range of marine resources as well as camelids and other land game, nuts, tubers, and other plants and vegetables, coincide with the warming of the climate and signal the change of strategies of subsistence, which, according to some authors, led to the specialization of certain groups. Whereas some groups of foragers had attained a specialized economy, like hunting or fishing, others had a more generalized economy, which was complemented with other logistical and opportunistic strategies in order to collect resources. Both of these groups, distinguished mainly by technological features in their cultural assemblages (and characterized mainly by their lithic industries), favored the population of certain areas, leading to the first settlements of the Archaic (10 000–4000 BCE). Predilection for certain areas, some authors suggest, derived in their population and subsequent emergence of sedentary occupations, privileging some parts of the Central Andes as well as certain sections of the Pacific coast.

The Andes

Given its large and diverse geography and social characteristics, it is difficult to frame the past of the Andes within a few specific topics without neglecting others. However, considering the variety of ecological and sociopolitical scenarios, the different forms of interaction between and within the peoples that populated this area need to be highlighted. Whereas conflict, warfare, and competition are crucial to many models of social interaction (*see* Conflict and War, Archaeology of: Behavior and Social Organization), the widespread networks established within and throughout the different regions of the Andes and further through time constitute the most important theme in South American archaeology. These concepts relate to a wide spectrum of social interaction among Andean societies as key elements in explaining the relationships between Andean societies; additionally, they present a powerful model incorporating aspects such as mobility and social organization of early archaic groups of the South Central part of the Andes, which prevailed until later periods (Murra, 1985; Nuñez and Dillehay, 1995). The connections and patterns of mobility between these different regions developed widespread networks of interaction and established the foundations for what would be called ecological complementarity (Murra, 1985).

Closely related to complementarity is the debate about the origins of Andean civilizations. This debate unfolded early in the twentieth century and involved the Harvard-educated Peruvian Julio C. Tello who would be known later as the father of Peruvian archaeology, whose autochthonous theory countered Max Uhle's theory of external origins. Tello proposed that Chavín de Huantar, an archaeological site located in the Peruvian *sierra*, was the 'mother culture' of Andean civilizations. He considered the evidence of monumental architecture and iconography at Chavín de Huantar to be related to a strong religious and sociopolitical structure of a society that began in the Amazon Basin, and evolved in the highlands to become the progenitor of civilizations. On the other hand, taking a diffusionist standpoint, Uhle postulated that the Andean civilizations were the result of migration from Mesoamerica and, hence, had their center in coastal areas. Not only was this debate related to the culture historical efforts of chronology building that characterized archaeology at the time, but it was also related to an ideological and theoretical controversy deeply inserted at the heart of South American archaeology.

Thus, for example, subsequent works in the Ucayali Basin (Lathrap, 1977) and at Chavín (Burger, 1992) were taken vis-á-vis Michael Moseley's *Maritime Foundations of Andean Civilizations* (1974). Both groups of theories took a materialistic point of view related to economic factors and means of subsistence. While those in favor of the origins in the highlands argued for the importance of agriculture and the domestication of animal species, mainly camelids, and the importance of interregional interaction between *sierras* and lowlands, those who favored the coastal origins highlighted the importance of marine resources, without precluding the practice of farming. In support of coastal origins, scholars cited evidence of cultigens such as beans, squash, and cotton that was found associated with artifacts of fishing technology in several excavations at sites on the Pacific littoral.

To the south, Wendell C. Bennett commenced research at Tiwanaku with sponsorship from the American Museum of Natural History. This monumental site in the Bolivian highlands near Lake Titicaca had been reported before by prehistorians like Bandelier, Stübel, and others who visited the region, following early colonial accounts. Uhle's earlier observations on stylistic sculptures and architecture provided him enough information to argue that Tiwanaku predated the Incas. Bennett's attention was centered in the monumental architecture of the site and other religious manifestations. He correlated and complemented his architectural data with an analysis of the style of ceramic decorations recovered from his excavations, establishing a chronological framework for the site. These chronological sequences and debates gave shape to the chronology of Andean pre-Hispanic societies.

Attempting to provide a chronological backbone accountable for stylistic and technological changes, John H. Rowe criticized previous chronological frameworks as inadequate to explain the origins and diffusion of archaeological traits. Based on his work in Ica Valley, Rowe proposed a pottery seriation sequence of horizons and periods, as units of time that could be used to correlate the most important events within different parts of the Andes (Rowe, 1962). Rowe's sequence was not initially based on absolute dating methods; it was Edward Lanning (1967) who later complemented and adjusted the initial sequence to the information provided by radiocarbon dates (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology). Although archaeological research in other areas and sites soon posed a series of challenges for these initial proposed sequences, the three main units of this sequence (the Early, Middle, and Late periods) remain the chronological backbone of Andean archaeology. Each unit is related to major cultural traits widespread through different regions of the Andes, implying evolutionary stages of social development.

The Early period is commonly linked to early hunter-gatherer societies, and is derived in large part from lithic technologies and a few examples of pristine architecture (Osorio et al., 2011; Shady et al., 2001). Production activities at the end of this period include permanent occupation and inhabitation of sites that led to less mobile lifestyles as well as technological innovations like ceramic and agricultural production, complementing pastoral and herding activities. Archaeologists have also scrutinized the processes of animal and plant domestication, which occurred as early as 8000 BCE in parts of the Andes, according to evidence of domesticated squash and maize at sites such as Las Vegas, Real Alto, and Loma Alta, in Ecuador. Although maize was adopted a bit later (5000 BCE), maize was ubiquitous at Real Alto by 2800–2400 BCE, along with achira, manioc, and some tubers. Similarly, evidence of arrowroot was reported in the Cauca Valley, Colombia. Parallel to developments in northern South America, the Andean highlands present solid evidence for the presence of contemporaneous plant domestication. In the Titicaca Basin, domestication of quinoa and other species had taken place by around 800 BCE (Bruno and Whitehead, 2003). Use of these plants was associated with an increased use of hoes, farming communities, and with the emergence of social complexity during the Formative period. Similarly, the domestication of llamas and alpacas around 5500 BCE contributed to the development of Andean society by providing a means of transporting goods as well as raw material for the manufacture of textiles, a significant commodity. New methodological procedures have built on earlier contributions on plant domestication and of incorporation of animals into consumption and ritual practices in earlier societies, providing wider frameworks of interpretation (DeFrance, 2006; Perry et al., 2007).

Whereas the domestication of animals and plants was an important aspect of this period, another has to do with gradual intensification related to agricultural production, technology, and political organization. Different kinds of earthworks were constructed to improve the conditions of or expand the capacity for food production and storage through different configurations of labor mobilization. Irrigation canals, for instance, were used to provide water to extensive areas of the northern Andes, especially in desert settings. Raised fields, generally built in swampy areas (Denevan, 2003), provided areas for cultivation and avoided seasonal or temporary floods by elevating agricultural surfaces. Other settings, like the steep and narrow valleys of the intermontane region, also show massive pre-Columbian terrace construction (Niles, 1999). These constructions provide evidence for a large and sustained modification of the natural landscape, and modified the climate by creating a humid warming effect to protect crops from low temperatures. Because of the magnitude and technological sophistication, most of these constructions – as was the case with raised fields and terraces – were associated with intensive agriculture and stronger economic strategies that involved investment in and administration of a massive work force. The estimated size of these work forces has been suggested as a correlate for social complexity (Kolata et al., 2000). However, drawing from experimental archaeology research in southern Peru, Erickson (1999) suggests that earthworks such as check dams, cross channels, sloping field terracing, and others could be the result of the work of domestic units without the need of hierarchically organized labor.

The Middle period includes both the so-called formative societies (Pucara, Chavín, Moche, others) and state societies (Tiwanaku, Chimu, Paracas, others). The former are characterized by the construction of public spaces, sunken temples, and incipient farming, the latter represent socially and politically consolidated societies. A significant feature of these state societies is a shared religious ideology that gradually became institutionalized. These shared ideologies were preceded by the Kotosh tradition in Chavín (Burger, 1992) and the Yaya Mama tradition in societies of the circum-Titicaca region. The importance of religious traditions has been proposed mainly from evidence found in emblematic sites such as Chavín de Huantar, Huari, Tiwanaku, Pachacamac, and various Moche sites. The monumental architecture found in these sites (Moore, 1996) can be related to former small villages of permanent residence and earlier religious centers. The importance placed on the religious and economic influence of societies such as Chavín, Tiwanaku, and Huari, which is manifested through a broad set of interregional networks connecting administrative and religious centers with hinterland regions, has privileged the South Central Andes as the cradle of Andean civilizations.

Evidence suggests that early settlements in other parts of the Andean coast, such as Real Alto and Las Vegas in Ecuador and

Las Haldas, Aspero, and Caral in Perú (Shady et al., 2001), could have had similar and parallel developments leading to early forms of urbanism. The consolidation of these processes gave rise to well-established societies characterized by complex social structures and institutions, and makes necessary an intricate scenario of interrelationships among regions in the Andes, the coast, and the Amazon.

One of the most representative characteristics of the Middle period is the strengthening of the exchange and interaction networks that articulated around some capital cities such as Tiwanaku (Kolata et al., 2000) located in the south-central Andes, the circum-Titicaca region or Pachacamac and Conchapata (Isbell, 2008) located in the northern Andes, which bound widespread regions. According to Isbell, these cities epitomized social formations with state government, well-defined hierarchical social structures, and an ample sphere of interactional dynamics. Characterized as a ritual, civic, and political center with the presence of monumental architecture, Tiwanaku is said to have covered an extensive area of the southern Andes, comprising the Bolivian *altiplano*, northern Chile, and northwestern Argentina, through its relationships of economic interaction. On the other hand, Huari, along with sites like Conchopata located at the *sierra*, is said to have expanded its domains to incorporate the coastal areas of present-day Peru. Architectural features like orthogonal compounds found in Pachacamac and other coastal sites suggest a strong, although not conclusive, relationship. A little further north of the Moche, another social entity seems to have been part of a complex relationship with the rest of the complex scenario during this period (Chapdelaine, 2011). Monumental sites such as Huaca de la Luna, Chan Chan, and others provide evidence of a complex development of social and political nature in this part of the Andes.

The Late period is characterized by the Inca state (D'Altroy, 2002) and its administrative control over a broad territory of the Andes sweeping from the present south of Colombia to the central Chilean valley of Aconcagua. According to traditional chronological schema, the Incas represent the most advanced stage of evolution of Andean societies prior to the arrival of the Spanish conquerors. Centered in Cuzco, the Incas had expanded their territory through a variety of strategies, annexing territory through violent domination or political persuasion. The diverse scenarios that the Inca occupation produced in the already complex sociopolitical landscape of the Andes characterize this period. The presence of the Incas left a distinctive imprint in the Andean landscape; many of their administrative centers are clearly recognizable by their architectural features, which follow the construction patterns of the capital city.

Archaeological evidence found in these administrative centers suggests that the Inca developed previously unseen systems of colonial control over the conquered territories. This was done through a complex and gradual process, mostly between AD 1000 and AD 1400, that involved such strategies as the uprooting of certain groups who served to establish new colonies of *mitmaes* or *mitmaquna* and were put in charge of conducting military, economic, and political or administrative tasks for the Inca ruler (Acuto, 2005). As the state attained more political power over conquered hinterland territories and grew demographically, it was necessary to reinforce previously existing production technologies like terraces and irrigation systems. These were sometimes combined with ideological and religious systems like the lines of shrines called *ceques* (Kolata, 2013). The Inca also developed mechanisms of social control, which included violent forms of suppression (Conrad and Demarest, 1990) as well as political enticement and alliances to attain voluntary submission. In many cases conflict characterized the relationships with local groups (Ogburn, 2007). By the early sixteenth century, when the Incas had achieved control over a vast territory, the Spanish conquest took place.

The Spanish conquest produced a cataclysm in the structure of Andean societies as the Spanish seized political and administrative power from the Incas and transformed Inca authority into what became the colonial system. Part of this new system included the *encomienda*: a concession of territory along with its population for the administration of the *encomendero* or feudal lord that resulted in quick amassing of fortune by the conquerors at the expense of the native population's peonage labor. Much of the debate concerning this period has been based on the chronicles and historic colonial documents; whereas, on the one hand, the information of colonial provenance has been a valuable source of information, on the other, it has been highly criticized for its ethnocentric perspective and bias against the colonized societies. Although some scholars have explored diverse aspects related to this period in the Andes such as mining, production, and consumption practices and architecture of early colonial cities, the colonial period has largely been overlooked by archaeologists and considered the domain of historians (Van Buren, 2010). This is in stark contrast with the Caribbean and Brazil, for example, where archaeological contributions to historical and colonial period research have delved in to issues like ethnicity and ethnic relations, slavery, and racial tensions in colonial societies (Funari and Britez, 2006). However, as different recent contributions emerging mainly in the last decades have shown, there seems to be an increasing interest in this period.

The Amazon

The Amazon has never seen the same intensity of archaeological research as the Andes. During the late nineteenth century, while explorers were documenting the cities, monuments, and arts of the Andes, at the mouth of the Amazon the spectacular art of the Marajoara was rediscovered (Schaan, 2012). However, the momentum of these initial efforts was not sustained and by the middle of the twentieth century, Amazonian archaeology was a very small field, dominated by fundamental disagreements about the archaeological record and its interpretation. Only in the early twenty-first century could the Amazon be thought of as having a mature and growing community of scholars, focusing on several problems from a variety of theoretical vantage points (Heckenberger and Neves, 2009). One recent summary (Stahl, 2004) called on archaeologists to raise their expectations with regard to the cultural achievements of pre-Columbian Amazonians. Archaeological study of pre-Columbian peoples changed our understanding of the ecology and natural history of the region. South America, and the Amazon in particular, represents the 'wilderness,' and the opposite of civilization. The Amazon

remains an important foil to the Old World of cities, post-industrial civilization, and modernity.

What archaeology adds is an increased and increasing understanding that this mythical wilderness has not only been inhabited for millennia (Erickson, 2006), but that the occupation changed the wilderness in lasting ways. Today, as forests are cleared, earthworks emerge in western Brazil, showing that the timeless forest has a definable history (Schaan, 2012). The connection between the artificial soils studied by soil scientists, and the occupation middens studied by archaeologists shows that large sedentary populations lined Amazonian rivers (Lehmann, 2003). To conserve the Amazon Basin because it is the last untouched wilderness denies the history and the humanity of the generations of Amazonians who farmed and fished, trekked, and hunted there. A more accurate understanding is to analyze the Amazon as a product of human society in relationships with water, fire, earth, trees, crops, weeds, fish, animals, and birds.

How can we use archaeology to better understand the Amazon without creating another myth? Keeping archaeology in conversation with many stakeholders, including indigenous Amazonians and other South Americans, should provide some protection. The more public and connected archaeological knowledge becomes, the harder it is for anyone to control. Today archaeologists are finding a wide range of evidence, which is changing our understanding of American archaeology more generally.

Ceramics are a powerful technology that demonstrates knowledge of materials and mastery of pyrotechnology. The earliest ceramics in the New World were created and used in the Amazon at Pedra Pintada, about 7500 years ago (Roosevelt et al., 1991). In the context of this significant development in technology, the Amazon does not suffer by comparison with other parts of the Americas. The choices potters make as they shape and decorate their pots express a huge range of cultural ideas. And after the pots are made, people use them to cook and conserve their food, bury their dead, and demonstrate and create their social status. Schaan (2012) analyzes ceramic 'tangas' as part of rituals associated with female initiation. The use of ceramic vessels as burial urns gives insight into indigenous ideas and behaviors about death. Ceramics are also at the center of modeling the movements of societies over the long term throughout the Amazon Basin (Neves, 2008; Hornborg, 2005).

Modification of the soil and control of water to improve conditions for agriculture was once studied as one of the prime movers of the rise of political complexity, the state, and civilization. Although this question no longer has the same urgency, the creation of landscapes of intensive agriculture is significant because it corresponds with the investment of labor in the landscape, increased agricultural potential, and the organization of that labor. Since the 1960s, it has become apparent that South America is literally covered with systems of pre-Columbian intensive agriculture (Denevan, 2003). Archaeologists are studying more of these systems, and new examples are emerging as satellite imagery becomes more widely available (Rostain, 2008). Not the least important implication of this part of the archaeological record is that some of these systems of intensive agriculture were sustained over at least several centuries. The archaeological record from Marajo Island documents the development of an art style and tradition on a par with those of more widely recognized societies in the Americas (Schaan, 2012). It also documents how nonstate societies controlled flows of water to engineer the environment.

One of the most important markers in the archaeological record is the city. Cities are the most complex artifacts created by humans. They are large, complex transformations of the environment which move humans away from their original connections to the environment into second and third order relationships. In these artificial places, new ideas about politics, economics, religion, art, and technology are incubated. Before the past few decades, no one would even have suggested that cities could have been part of the Amazonian archaeological record. Although it is a minority position, Heckenberger et al. (2008) describe a collection of large circular occupation sites in the Xingu as forming a network similar in scope and size to the cities that emerged in the early middle ages in Europe. These settlements were almost certainly not the products of a strong central power, but they could be the expression of a dynamic process that also included linguistic and cultural diversity, and novel forms of social organization. As a counterpoint to the precocious societies of the north coast of Peru, evidence of ceremonial architecture and sophisticated art from the fourth millennium BC has recently been discovered in the Ecuadorean Alta Amazonía (Valdez, 2013). Ceramics, domesticated plants, intensive agriculture, and urbanism are crucial in the deep history of civilization, and of Amazonia.

The archaeological record of the Amazon includes evidence regarding three points of general interest. First, it seems that the 'natural' environment of the Amazon is riddled with evidence that humans modified and manipulated plants, animals, and landscapes in ways that dissolve any easy contrast between nature and culture (Erickson and Balée, 2006). This is not to say that the Amazon is in every way an artificial construction, but instead something more fundamental. As is suggested by ethnographic accounts about how indigenous people think about nature and culture, these anthropogenic environments were built by communities that may have had very different ideas about how people relate to nature. For some Amazonians, trees and mounds are social beings, with whom one can and must interact, and this would suggest that different actions are appropriate (Dillehay, 2006). Animals are also in some cases thought of as ancestors. Pre-Columbian Amazonians did not separate nature and culture in the same way as Western science (Descola, 2013). Following the trajectory of such communities and such societies through deep history would seem to be of great utility as we make policy decisions about relationships between humans, technology, and resources.

Conclusion

One point of view says that South America is unique and independent of traditions in other parts of the world, one of the independent hearths of civilization. As civilization is defined more broadly and evidence of human achievement over very long periods of time accumulates, it would be foolish to imagine any place where humans had settled as being truly derivative from anywhere else. In South America, research has focused first on the Andes, and only in more recent decades on

the Amazon. The archaeological record suggests that pre-Columbian peoples did not observe such a strong division between these domains. The combination of the Andes Mountains with tropical latitude means that a wide range of environments are within easy reach of people on foot. Farmers, herders, and traders moved up and down the mountains, from coastal deserts and valleys up past the snowy peaks down to the forests, and savannas on the other side. The differences between these environments contributed to tremendous diversity in economy, culture, and politics. The movements of people, goods, and ideas up and down the mountains collapse any simple dichotomy between the highlands and the lowlands, between the Andes and the Amazon. The future of archaeology in the Amazon is sure to include interaction with Andean archaeology, but also closer integration with anthropology and tropical ecology, as the archaeological record becomes part of larger understandings of the Amazon Basin.

At the turn of the twenty-first century, archaeology in South America has undergone significant changes. In the Andes as well as in the Amazon, long-standing topics have remained central, such as early peopling, formation of state societies, and technological responses to environmental adaptation (Lombardo et al., 2013; Osorio et al., 2011; Roosevelt et al., 1991). A combination of residue analysis on artifacts and other techniques such as pollen and phytolith analysis has been instrumental in identifying and dating processes of domestication and use of different species of crops and plants in the past. The application of archaeometric technologies to fine tune chronological charts (Steele and Politis, 2009) and innovative nondestructive techniques have been employed to explore issues like compositions of metals and provenance of metal objects (Lechtman and Macfarlane, 2005) as well as obsidian and other resources commonly found in archaeological contexts (Yacobaccio and Morales, 2005). Many of these new techniques have been applied to address old questions (trade and exchange, ideological use of objects and raw materials, peopling of the Americas).

Theoretically, with the advent of changes in the field during the last decades of the twentieth century, Latin American archaeologists have also departed from traditional frameworks and incorporated different theoretical approaches to problem-oriented research (Politis, 2003). Although this is an ongoing process, recent work is also characterized by interest in landscape (Walker, 2012), agency, and the politics involved in archaeological knowledge production. The development of national and regional archaeologies, which once gave the discipline its foundational core, has been fruitfully criticized. Recent contributions from Latin American scholars have emphasized the need to scrutinize the relationships between archaeology and society, reflected mainly in the sociopolitical contexts of the different countries (Angelo, 2011; Benavides, 2004). As part of this, issues of representation and domination regarding identity, ethnicity, and nationalism have been examined along with issues of heritage and the situation of indigenous communities in relation to national cultural politics. Another new issue related to the recent past of the South American countries comes from the Dirty War period, developed mainly in Argentina, where archaeologists and forensic anthropologists have painstakingly recovered evidence to bring to justice those responsible for the case of the disappeared during the military dictatorship (Funari and Zarankin, 2006). As archaeology becomes more global, the themes of South American archaeology expand the theoretical, methodological, and political scope of our understanding of the past.

See also: Archaeology and Cultural Memory; Archaeology, Politics of; Civilizations; Environmental Archaeology; Ethnic Identity and Ethnicity in Archaeology; Ethnohistory; Food Production, Origins of; Language Families, Archaeology and History of; Mesoamerica, Archaeology of; Political Ecology: Anthropological Perspectives; Political Economy in Anthropology; Settlement and Landscape Archaeology; South America: Sociocultural Aspects; States and Civilizations, Archaeology of.

Bibliography

Acuto, F., 2005. The materiality of Inka domination: landscape, spectacle, memory, and ancestors. In: Funari, P., Zarankin, A., Stovel, E. (Eds.) Springer, New York, pp. 211–235.

Angelo, D., 2011. Turning tables in search of dialogue: the making of indiscrete spaces in latin american contexts. In: Gnecco, C., Ayala, P. (Eds.), Archaeology and Indigenous Peoples in Latin America. Left Coast Press, Walnut Creek, CA, pp. 89–108.

Benavides, H., 2004. Making Ecuadorian Histories: Four Centuries of Defining Power. University of Texas Press, Austin.

Bruno, M., Whitehead, W., 2003. Chenopodium cultivation and formative period agriculture at Chiripa, Bolivia. Latin American Antiquity 14, 339–355.

Burger, R., 1992. Chavín and the Origins of Andean Civilization. Thames and Hudson, New York.

Chapdelaine, C., 2011. Recent advances in moche archaeology. Journal of Archaeological Research 19 (2), 191–231.

Conrad, G.W., Demarest, A.A., 1990. Religión e imperio : dinámica del expansionismo azteca e inca. Alianza Editorial Mexicana : Consejo Nacional para la Cultura y las Artes, Mâexico, DF.

Cuellar, 2013. The archaeology of food and social inequality in the Andes. Journal of Archaeological Research 21 (2), 123–174.

D'Altroy, T., 2002. The Incas. Blackwell Publishers, Malden, Mass.

DeFrance, S., 2006. The sixth toe: the modern culinary role of the Guinea Pig in southern Peru. Food & Foodways 14 (1), 3–34.

Denevan, W.M., 2003. Domesticated Landscapes of Native Amazonia and the Andes. Oxford, New York.

Descola, Philippe, 2013. Beyond Nature and Culture. University of Chicago Press.

Dillehay, T., 2006. Monuments, Empire and Resistance. Cambridge University Press, Cambridge.

Dillehay, T., 1997. Monte Verde: A Late Pleistocene Settlement in Chile, vol. 2. Smithsonian Institution Press, Washington, DC.

Erickson, C., 1999. Neo-environmental determinism and agrarian "collapse" in Andean prehistory. Antiquity 73, 634–642.

Erickson, C., 2006. The domesticated landscapes of the bolivian amazon. In: Balée, W., Erickson, C. (Eds.), Time and Complexity in Historical Ecology: Studies in the Neotropical Lowlands. Columbia University Press, NY, pp. 235–278.

Erickson, C., Balée, W., 2006. The historical ecology of a complex landscape in Bolivia. In: Balée, W., Erickson, C. (Eds.), Time, Complexity and Historical Ecology. Columbia University Press, New York, pp. 187–233.

Funari, P., Britez, F. (Eds.), 2006. Arqueología Histórica en América Latina: Temas y Discusiones Recientes. Ediciones Suárez, Mar del Plata.

Funari, P.P.A., Zarankin, A. (Eds.), 2006. Arqueología de la represión y la resistencia en América Latina (1960–1980). Editorial Brujas. Gnecco and Ayala 2011.

Heckenberger, M., Neves, E.G., 2009. Amazonian archaeology. Annual Review of Anthropology 38, 251–266.

Heckenberger, M.J., Russell, J.C., Fausto, C., Toney, J.R., Schmidt, M.J., Pereira, E., Franchetto, B., Kuikuro, A., 2008. Pre-Columbian urbanism, anthropogenic landscapes, and the future of the Amazon. Science 321 (5893), 1214–1217.

Hornborg, A., 2005. Ethnogenesis, Regional Integration, and Ecology in Prehistoric Amazonia: Toward a System Perspective. Current Anthropology 4, 589–620.

Isbell, W., 2008. Wari and Tiwanaku: International Identities in the Central Andean Middle Horizon. In: Silverman, H., Isbell, W. (Eds.), Handbook of South American Archaeology. Springer, New York, pp. 731–759.
Kolata, A.L., Binford, M.W., Brenner, M., Janusek, J.W., Ortloff, C., 2000. Environmental thresholds and the empirical reality of state collapse: a response to Erickson (1999). Antiquity 74 (284), 424–426.
Kolata, A.L., 2013. Ancient Inca, vol. 12. Cambridge University Press, Cambridge.
Lanning, E.P., 1967. Peru before the Incas. Prentice-Hall, Englewood Cliffs, NJ.
Lathrap, D., 1977. Our father the Cayman, our mother the Gourd: Spinden revisited, or a unitary model for the emergence of agriculture in the new world. In: Reed, C. (Ed.), Origins of Agriculture. Mouton, The Hague, pp. 713–752.
Lombardo, U., Szabo, K., Capriles, J.M., May, J.-H., Amelung, W., Hutterer, R., et al., 2013. Early and middle holocene hunter-gatherer occupations in Western Amazonia: the hidden shell middens. PloS one 8 (8), e72746.
Lechtman, H.N., Macfarlane, A.W., 2005. La metalurgia del bronce en los Andes sur centrales: Tiwanaku y San Pedro de Atacama. Estudios atacameños 30, 7–27.
Lehmann, J. (Ed.), 2003. Amazonian Dark Earths: Origin, Properties, Management. Springer, Heidelberg.
Lumbreras, L.G., Batres, C.M., 1981. Arqueología de la América andina. Editorial Milla Batres, Lima.
Moore, J., 1996. Architecture and Power in the Prehispanic Andes: The Archaeology of Public Buildings. Cambridge University Press, Cambridge.
Moseley, M.E., 1974. The Maritime Foundations of Andean Civilization. Cummings Publishing Company, Menlo Park.
Murra, J.V., 1985. The limits and limitations of the "vertical archipelago" in the Andes. In: Masuda, S., Shimada, I., Morris, I. (Eds.), Ecological Complementarity, Andean Ecology and Civilization: An Interdisciplinary Perspective on Andean. University of Tokyo Press, Tokyo, pp. 15–20.
Neves, E.G., 2008. Ecology, ceramic chronology and distribution, long-term history, and political change in the Amazonian floodplain. In: Silverman, H., Isbell, W.H. (Eds.), The Handbook of South American Archaeology. Springer, New York, pp. 359–379.
Niles, S.A., 1999. The Shape of Inca History: Narrative and Architecture in an Andean Empire. University of Iowa Press, Iowa City.
Núñez Atencio, L., Dillehay, T., 1995. Movilidad giratoria, armonía social y desarrollo en los Andes Meridionales. Patrones de Tráfico e interacción económica, second ed. Universidad Católica del Norte, Chile.
Ogburn, D.E., 2007. Human Trophies in the Late Pre-Hispanic Andes: Striving for status and maintaining power among the Incas and other societies. In: Chacon, R.J., Dye, D.H. (Eds.). The Taking and Displaying of Human Body Parts as Trophies by Amerindians Springer, New York, pp. 505–522.
Osorio, D., Jackson, D., Ugalde, P.C., Latorre, C., Pol-Holz, D., Santoro, C.M., 2011. Hakenasa cave and its relevance for the peopling of the southern Andean Altiplano. Antiquity 85 (330), 1194–1208.
Perry, L., Dickau, R., Zarrillo, S., Holst, I., Pearsall, D.M., Piperno, D.R., Zeidler, J.A., 2007. Starch fossils and the domestication and dispersal of chili peppers (Capsicum spp. L.) in the Americas. Science 315 (5814), 986–988.
Politis, G.G., 2003. The theoretical landscape and the methodological development of archaeology in Latin America. American Antiquity 68 (2), 245–272.
Roosevelt, A.C., Housley, R.A., Da Silveira, M.I., Maranca, S., Johnson, R., 1991. Eighth millennium pottery from a prehistoric shell midden in the Brazilian Amazon. Science 254 (5038), 1621–1624.
Rowe, J.H., 1962. Chavin Art: An Inquiry into Its Form and Meaning. Museum of Primitive Art, distributed by University Publishers.
Rostain, S., 2008. Agricultural earthworks on the French Guiana Coast. In: Silverman, H., Isbell, W.H. (Eds.), Handbook of South American Archaeology. Springer, New York, pp. 217–234.
Schaan, Denise, 2012. Sacred Geographies of Ancient Amazonia. Left Coast Press, Walnut Creek, CA.
Shady, et al., 2001. Dating caral, a Preceramic site in the Supe Valley on the Central coast of Peru. Science 292 (5517), 723–726.
Silverman, H., Isbell, W. (Eds.), 2008. The Handbook of South American Archaeology. Springer, New York.
Stahl, P.W., 2004. Archaeology: greater expectations. Nature 432 (7017), 561–562.
Steele, J., Politis, G., 2009. AMS 14C dating of early human occupation of southern South America. Journal of Archaeological Science 36 (2), 419–429.
Steward, Julian (Ed.), 1946–1950. Handbook of South American Indians, Vol. 1–7. Smithsonian Institution - Bureau of American Ethnology, Washington, DC.
Steward, J., 1946. Handbook of South American Indians. Bureau of American Ethnology, Washington, DC.
Valdez, F., 2013. Primeras Sociedades de la Alta Amazónia: La cultura Mayo-Chinchipe Marañon. INPC, Quito.
Van Buren, M., 2010. The archaeological study of Spanish colonialism in the Americas. Journal of Archaeological Research 18 (2), 151–201.
Walker, J.H., 2012. Recent landscape archaeology in south america. Journal of Archaeological Research 20 (4), 309–355.
Yacobaccio, H.D., Morales, M., 2005. Mid-Holocene environment and human occupation of the Puna (Susques, Argentina). Quaternary International 132 (1), 5–14.

South America: Sociocultural Aspects

Norman E Whitten, Jr., University of Illinois at Urbana-Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Native South American cultures prior to 1492 represented a dynamic mosaic of languages, cultures, social systems, rituals, and power relationships. Long-distance trade – sometimes extending for thousands of miles – combined with bilingualism and multiculturality. By the creation of an empire through conquest, the Inca altered the cultural mosaic in the Andes in the fifteenth century. A new cultural system was violently created by the overseas Spanish and Portuguese empires after 1492 in their conquest and radical colonization of the continent. The Spanish and Portuguese brought black slaves from Africa and from Iberia with them. Many such people escaped to create free, fortified sites where Native American and African American people often collaborated, intermarried, and established political systems on the fringe of the dominant colonial territories. The Spanish initiated an ideology of ethnic blending, called *el mestizaje* ('the blending,' or miscegenation). In the late twentieth century, indigenous and black movements in South America worked against this false imagery and misleading ideology of ethnic blending in their respective nation-states and across national boundaries. Cultural diversification and nationalist centralization characterized the South American nation-states in the late twentieth century as they did during colonial and republican epochs. Such centripetal–centrifugal forces are seen in cultural performances and in social movements.

South America, the world's fourth largest continent, is composed of 12 nation-states and 1 territory. The significant differences among these 13 countries are matched by great internal sociocultural and ethnic diversity. Contemporary neoliberal and authoritarian governments strive to centralize and homogenize their populations. Internal cultural, social, and political movements heighten diversity and conflict. The military often plays a strong role in centralizing efforts. In the past decade, new forms of socialism have emerged in some nations that both challenge neoliberalism and at the same time seek to use it creatively for social reforms. The Roman Catholicism of European colonialism historically and culturally shapes the pan-national and internal cultural unity of the 10 Iberian-derived nations. Diversity is especially strengthened by the striking polarization between the small wealthy minority and the large poor majority; by indigenous and Afro-Latin American movements of cultural and ethnic self-affirmation; by the presence of significant ethnic blocs resulting from European migrations since the two World Wars; by the struggles among leftist, rightist, and populist political parties; by the challenge of the twentieth- and twenty-first-century spread of Protestantism; and by the liberal Catholicism of recent reforms.

Human and Topographic Diversity

South America's rapidly expanding population contains as wide a spectrum of phenotypic features as found anywhere in the world (e.g., Collier et al., 1985; Fuentes, 1992). Its modern nation-states are Venezuela, Colombia, Ecuador, Peru, Bolivia, Chile, Uruguay, Paraguay, and Argentina, where Spanish is the official language; Brazil, where Portuguese dominates; Guyana, where English is the national language; Suriname, where English, Dutch, and Sranan are prevalent; and Guyane, an administrative district (*département*) of France, where French is dominant. Quechua is also an official language of Bolivia, Peru, and Ecuador (Quichua), as is Guaraní in Paraguay. A myriad of other languages are spoken in most of these countries.

Western South America is dominated by the mighty Andes mountains, with snow-coned active volcanoes that rise to 20 000 feet. They extend southward from the Caribbean Island of Trinidad to terminate in southern Argentina at Tierra del Fuego. In this region, Venezuela, Colombia, and Ecuador are known as the northern Andes (and the former two as the Spanish Main); Peru and Bolivia constitute the central Andes; and Chile, Paraguay, Uruguay, and Argentina are called the 'southern cone.' Eastern South America features the great Amazon Basin, epitomized by the Amazon itself, the 'Ocean Sea.' Toward the headwaters of the huge feeder rivers of the Amazon lie the mountainous Guiana Shield in the north, the Brazilian Shield in the south, and the Andean foothills or *montaña*, rimmed by the *ceja* ('eyebrow') in the west. Northwest and northeast of the Amazon regions, and west of the Guiana Shield, are the plains (*llanos*) of Colombia and Venezuela, which phase into the Orinoco Basin, and west of the Brazilian Shield are the *llanos de Mojos* of Bolivia. A subregion that combines elements of Amazonian and Andean systems is the North Pacific coast. From northern Ecuador through western Panama, an Amazonian ecosystem prevails. From northern Peru southward, high aridity is characteristic (e.g., Collier et al., 1985).

Pre-Columbian People: Their Lands, Products, Societies, and Languages

Native South American cultures prior to 1492 represented a dynamic mosaic of languages, cultures, social systems, rituals, and power relationships (Denevan, 1976; Dillehay, 2000; Hornborg and Hill, 2011; Silverman and Isbell, 2008). Long-distance trade – sometimes extending for thousands of miles – combined with bilingualism, or multilingualism, and multiculturality. Such trade traversed some of the most rugged landscapes in the world, creating networks of native knowledge and imagery crosscutting great geographical differences. The antiquity of South American native peoples is highly controversial. In the east, and thence up the Amazon and Orinoco (the latter has its headwaters in the Parima and Pacarima

mountains of the Guiana Shield and flows west, then north, then east out into the Atlantic), evidence suggests antiquity of 12 000 years ago or earlier, but the routes of arrival are obscure (e.g., Dillehay, 2000). Neither transatlantic nor transpacific routes seemed credible to scholars in 1999, although sporadic contacts in South America by the first Americans with Asian-Pacific and Euro-African peoples remain subjects of possibility, discussed in the early twenty-first century. Plant domestication seems to have begun in Amazonia and in the Andes by about 8000 years ago. Interactions between Andean and Amazonian people are demonstrated by literature in the twenty-first century to be more common than previously thought (Hornborg and Hill, 2011; Silverman and Isbell, 2008).

The period beginning about 6000 years ago and continuing until the European conquest and colonization marks the florescence of fine ceramics in greater Amazonia, and the development of textiles, gold work, and silver work in the Andes and adjacent coastal regions. Pottery may have had its origin in the Upper Madeira basin near the border of Brazil and Bolivia; later, it is found in abundance at the mouth of the Amazon and subsequently along the banks of the Amazon. Ceramics manufactured on Marajó Island, an area the size of Connecticut in the United States, rivaled the quality of any ceramic wares found anywhere in the world in terms of their beauty of form and decoration (Hornborg and Hill, 2011).

By 4500 years ago, great westward migrations from the lower and central Amazon region were underway, and regional consolidation in the Andes was occurring. People from coastal, Andean, and Amazonian regions were in contact with one another, and with the people of the Caribbean Islands, Central America, and Mesoamerica. People who belong to three language families – Arawak, Carib, and later Tupi – spread from the eastern and central Amazon to the Orinoco drainage and on to the Caribbean (Arawak), and to the base along the rim of the *montaña*. Other Amazonian language families of prominence include Tukano, Tukuna, Jívaro, Pano, and Gê; these languages spread in a general arc from northwest of the Amazon around the central Amazonian rim to the southwest. Andean languages were also highly diverse. Quechua, a dialect of which was adopted by the Inca, has a known time depth of only 1000 years; scholars are still teasing out earlier linguistic diversity in the Central Andes.

The Inca Empire, radiating north and south out of its capital of Cuzco, in what is now Andean Peru, emerged in the fifteenth century to severely alter the cultural mosaic in the Andes, and set up political-economic and cultural shock waves that radiated eastward. This empire was in full conquest fervor and bureaucratic consolidation by the time that word arrived from the Caribbean and Central America of the pending Spanish advances down the Pacific Coast and up into the Andes. At the height of their expansion, the Inca controlled all Andean systems from central Chile to southern Colombia, and their north-south 3416-mile parallel cobblestone roadway (one coastal, one Andean) through this vast territory (385 000 square miles) amazed the conquering Spaniards. Although the Inca never dominated the adjacent indigenous people of Upper Amazonia, there is ample evidence of extensive trade and cultural exchange between Amazonia and the Inca-dominated Andes and coast (Richardson, 1994). A dialect of Peripheral Quechua (which is related to Quichua) was the lingua franca used by the Inca in its conquest, and Quechua became the language of millions of people throughout the empire and beyond (Whitten and Whitten, 2011). Other prominent Andean languages include Barbacoan and Chocoan in the northwest coastal region, Chibcha in the north Andes, Aymara in the central Andes, and Mapuche in the southern cone.

In the greater Amazonian areas east of the Andes, people depended on maize, manioc, a variety of root crops, and palm cultivation for caloric intake, and on intensive fishing and reptile and amphibian collecting for protein and amino acid balance. Hunting in the various ecosystems of the flooded forest, the upland forest, and the flood plains and riverine systems complemented the balance of fishing and horticulture. As time progressed, manioc (*Manihot esculenta* Crantz), which, like pottery, probably had its origins in the Upper Madeira basin (Hornborg and Hill, 2011), came to be the staple of the agricultural system, with maize sporadically and regionally grown and consumed. Use of palms was and is ubiquitous. Fishing in rivers, estuaries, streams, and coastal waters was of paramount importance, as were techniques for corralling turtles and other amphibians. Polities varied from well-developed chiefdoms and relatively small-scale states to more dispersed systems. Intermarriage between peoples at war with one another, and speaking different languages, seems fairly common.

In the Andean regions, the potato was the supreme crop, with beans, squash, maize, quinoa, and other grains providing a balanced diet when combined with animal protein. Approximately 240 species of potato were grown, harvested, and prepared in many ways. The llama (one of three Andean camilids), the guinea pig, and later the hairless dog were the primary animals raised for food, and the llama was also used as a beast of burden. Polities ranged from small-scale systems to states and the expansive Inca Empire (Richardson, 1994).

The European Conquest of Native Lands and People

A new cultural system based on conquest and capitalism was violently created by the overseas empires of the Spanish and Portuguese after 1492 as they sought huge profits in gold, silver, spices, and slaves. In the Treaty of Tordesillas, signed in 1494, the Spanish and the Portuguese divided up their colonies long before they knew the vast extent of their respective territorial 'holdings.' From the terms of this treaty, the Portuguese would lay claim to Brazil, and Spain would take the rest, including all of the Pacific regions from South America to the Philippines and Japan, all of the Andes, and a part of Amazonia, a territory itself nearly the size of the United States. Portugal's interests first lay in the route around the horn of Africa to India, and only a bit later in its vast new colonies (e.g., Goodman, 1992; Williams and Lewis, 1993). In 1499, while sailing (and/or writing) for Spain, Amerigo Vespucci signed his map of the 'New World,' thereby creating the currently favored name for the entire continent (America). During the early to mid-1500s, Spain set out to locate, conquer, and colonize its Amazonian, Andean, and coastal regions, and Brazil set out on the conquest of its Amazonian, coastal,

and mountainous regions. The Spanish first brought sugar to the Americas, and the Portuguese first brought the banana from India. Both of these overseas empires introduced cattle, sheep, goats, pigs, and chickens.

Everywhere they went, the conquerors encountered native people. They subdued them by enslavement, trickery, and trade, and they spread European and African diseases for which native resistance was low or nonexistent. Where possible they enlisted them in slaving activities, and they killed indigenous people with firearms, swords, spears, and arrows with steel points. In areas feasible for such warfare they sent their war horses and war dogs to maim and kill native people in their radical colonization of the continent (Hall, 2005; Varner and Varner, 1983). This brutal transformation resulted in 90% or greater depopulation of the native inhabitants throughout the continent (Denevan, 1976). Trade and warfare, now with an escalating colonial need for indigenous chattel slaves in the lowlands, and indigenous landed slaves in the highlands, created a persistent poverty that endures into the twenty-first century. Nonetheless, cultural creativity endured with local-level technical competence to promote subsistence and adaptations (e.g., Van Cott, 1995).

Blackness

The Spanish and Portuguese brought black slaves from Africa and from Iberia with them (Bryant et al., 2012). Many such people escaped to mountainous, forested, inaccessible areas, and created a myriad of free, fortified sites where Native American and African American people often collaborated, intermarried, and established powerful political systems on the fringe of the dominant Spanish and Portuguese territories. From these refuge sites, often called *palenques* in Spanish and *quilombos* in Portuguese, self-liberated residents established powerful political systems on the fringe of the dominant Iberian and colonial territories (Price, 1996; Redstall, 2005; Whitten and Torres, 1998). They threatened the Crown political economies. In some regions such as western Ecuador, southwestern Colombia, western Venezuela, and the north of Brazil, active trade between runaway people, often called *cimarrones*, and the colonial planters ensued. There was also movement among free black people between Africa, Iberia, and the Americas.

Mestizaje, Cultural Practices, and Ideologies of Racial Blending

The Spanish initiated an ideology of ethnic blending called *el mestizaje* ('the blending,' or miscegenation) that excluded both indigenous people and people of African descent. Spanish administrators from the Republic of Spaniards demarcated a Republic of Indians (or '*indios*'), so named by Christopher Columbus, who thought he was in India. There was no legal space for black people, except what they themselves created. Those excluded from the Spanish-Mestizaje realm were sometimes referred to as *la mancha*, 'the (racial) stain.' From this viewpoint, the dominant, European concept of hybridity prevailed in this hegemonic model of racial mixing: the white, civilizing genes and cultures were to mix with the savage or barbarian 'Indian' and 'black' to form such intermediate categories as *mestizo* and *mulato* in Spanish. In Portuguese, *mestiço* and *mulato* referred to white–black hybridity (Whitten and Torres, 1998; Whitten and Whitten, 2011).

Those people with hybridized 'racial' features, according to European and later Creole (New World born) ideologues, ranged up and down the scale from white civilized Iberian to dark civilizing Creole. People in the diverse *mestizo* category were known in many areas as constituting the *castas*. In some times in some places, the number of their racialized categories ranged from a dozen named 'types' to hundreds. Indeed, there could be five or more such 'types' among the sons and daughters of one couple. They were constituted as those people somewhere betwixt and between the savage and the civilized, and legal and colloquial labels for some of them included such racialized concepts as 'wolf,' 'coyote,' 'throw back,' 'up in the air,' and even 'there you are!' Translation of 'casta' as 'caste' is erroneous.

Black people (including slave, '*bozal*,' and free, '*ladino*') and native people (usually divided into tame, '*manso*,' and fierce or wild, '*bravo*') fell outside of the scheme of *castas*. Their own intermixture, often rendered in Spanish by the term *zambo*, was (and is) a category filled with racist ideas about danger and power, for here was a mixture without the civilizing effects of European-sponsored hybridity. Such people were sometimes called 'the Devil's mix' by the white and powerful (Redstall, 2005; Whitten and Whitten, 2011). An indigenous artist using European techniques in 1599 did the first oil painting in South America. His subject was the 'Zambos' of western Ecuador who, adorned in European princely clothes, wearing indigenous gold jewelry, and carrying spears with iron tips characteristic of Africa and Europe, walked 100 miles from their free Zambo Republic on the coast of Ecuador to Quito to swear alliance to the Spanish Crown in the face of English encroachments (Whitten and Whitten, 2011).

Before the rise of nation-states, movements of self-liberation by black people and indigenous people were common, and some were quite successful. Early revolts in the sixteenth century resulted in ethnic regions that endured into the era of nation-state emergence. In 1579, the Quijo and the Jívaro of the Ecuadorian *montaña* revolted in attempts to unite Amazonian and Andean people (Uzendoski and Calapucha-Tapuy, 2012; Whitten and Whitten, 2011). In Peru and Bolivia, the revolts of Tupac Amaru and Tupac Katari later did the same. In the forested and mountainous coastal zone of Alagoas, Pernambuco, Brazil, black and indigenous people formed the state of Palmares in the 1640s, and they fought Portuguese and Dutch soldiers for a century (Redstall, 2005). Black people in Suriname revolted in the mid-seventeenth century, and initiated a 100-year war against the Dutch and British, winning independence for six nations – Saramaka, Paramaka, Kwinti, Ndujka, Matawai, and Aluku – in the interior of this region (Price, 1996). In neighboring Venezuela, the Carib-speaking Yekuana expelled the Spanish in 1767, and maintained their complete independence for 100 years until a brutal nationalist regime militarized the area for the purposes of exploiting wild rubber (Guss, 1999).

The Transformation from Colonies to Nation-States

Wars of nationalist independence from Spanish and Portuguese colonial rule began in 1810 in areas of vast distances from one another, from what is now Mexico through northern South America to the southern cone. Those who achieved positions at the pinnacle of power in South America were *criollos*, 'Creoles,' who outnumbered those of Spanish descent. In turn, they were vastly outnumbered by their darker compatriots, especially those of African descent who fought in wars led by Simón Bolívar in the north, and by José San Martín in the south. As the Creole nationalist wars moved through slave territories, slavery was abolished, but blackness and indigenousness were kept under rigid control by fears that the revolution would become an 'ethnic' one. The ethnically marginalized continued to be excluded from dominions of power. Both Bolívar and San Martín envisioned a united Spanish-speaking Creole nation, and lightness of phenotype within the ideology of *el mestizaje* was clearly ascendant (Anderson, 1991; Stern, 1987; Whitten and Torres, 1998).

Such a continental nation did not materialize and, one by one, contemporary nation-states rapidly emerged from their previously liberated colonial regimes. By 1822, Spain had lost all of its territories in the Americas, and the Portuguese king lost his colony to a new Brazilian monarchy with its own New World Empire. By 1830, all of the republics of Spanish-speaking South America had emerged, only to begin border wars, geopolitical vestiges of which last to the present day. Peru, for example, invaded Ecuador in 1941, claiming half the latter's territory, and these countries fought a war over a very small sector of disputed territory as recently as 1981 and 1995, both of them in indigenous Jivaroan-occupied sectors long serving as refuge areas due to the rugged, mountainous, Upper Amazonian topography (Whitten and Whitten, 2011).

African chattel slavery continued in the new states side by side with black freedom, as various systems of oppression of indigenous people continued side by side with native revolts (Escobar, 2008; McKnight and Garofalo, 2009; Silverblatt, 2004). The last country to abolish slavery, Brazil, did not do so until 1888, and in some states illegal slavery existed side by side with manumission legalities. Not until 1966 (Guyana) and 1975 (Suriname) would two of the three Guianas gain nation status from the British and the Dutch, respectively.

The Fusion of Cultures

The powers of Spanish and Portuguese Catholicism came to South America in the heyday of the Spanish Inquisition (Silverblatt, 2004; Whitten and Torres, 1998). Powerful priests and others sought out indigenous and African systems of thought and knowledge to be stigmatized as idolatry and devil worship. An underlying fear of converted Jews permeated some middle ranks of conquest and colonialism. The Inquisition became known through history by different names and is now (2013) the Vatican's Congregation for the Doctrine of the Faith, once headed by Cardinal Joseph Ratzinger who went on to become Pope Benedict XVI.

Those of humble Iberian origin sought alternative sources of inspiration for religion and magic from indigenous and Black people, thereby heightening the symbolic powers of the oppressed. Fusions of Iberian, African, and indigenous belief systems and ways of knowing and 'seeing' became legion. All this occurred at microlevels in macrosystems of Iberian Catholic missionization that give almost all of South America its special 'Latin American' sociocultural character (Ades, 1989; Collier et al., 1985).

In the area of language, pidgins emerged out of African, indigenous, and European imperfect syntheses, to become internally 'perfect' Creole languages. In medicine, the humoral system from Mediterranean classical antiquity fused with African curing programs and indigenous shamanism. The result was hundreds, perhaps thousands, of highly regional systems of health and curing that, in turn, served as vehicles of communication with people in other regions who had experienced similar cultural fusions. Religious syncretisms became the *modus operandi* of colonizing curates and of indigenous and African specialists in esoteric and numinous thought and reflection. Even horrors of the conquest were syncretized into indigenous and Afro-Latin American cultures (e.g., Hill, 1988; Whitten and Whitten, 2011).

For example, the battle cry of the Spanish conquistadors, '*Santiago!*' came from Santiago de Matamoros (Saint James the Moor Killer) as the Moors were defeated during the Castilian reconquista that was concluded in 1492, the same year that the Jews were expelled from Spain. It was used to mark Spanish territories in the Americas. After the reading of the 'Requirement' document informing all 'Indians' of their loss of rights to the Spanish, the conquerors yelled 'Santiago!' and attacked the natives with firearms, war horses, and war dogs, often feeding parts of the dead natives to the dogs, and rendering fat from the same dead natives to rub on the wounds of dogs and horses (Varner and Varner, 1983). Santiago underwent a ritual reversal in the Central Andes, where this saint became the patron icon of indigenous, Quechua-speaking peoples, and was called forth to resist the Spanish. Also in this region, and in parts of Amazonia, indigenous people say that some outsiders, whom they call *pishtaco* in Peru and *ñaqaq* in Colombia, seek 'Indian fat' for nefarious purposes. This may well be a modern reference to practices of the Spanish conquerors – an indigenous dimension of historical consciousness – although it is usually treated as an item of strange and curious native lore.

Cultural Continuities

In the face of radical transformations and syncretisms, continuities were also abundant. Amazonian languages such as Tupi, Arawak, Tukano, Tukuna, Jívaro, and Pano continued in their respective indigenous areas, with trade between different people also continuing, in spite of the efforts of the Catholic curates, first the Jesuits and Franciscans, and then others, to 'reduce' the populations to nucleated areas. Tupi itself, as Neenghatú, became the Lengua Geral of rural Brazil until replaced by Portuguese in the late nineteenth century, and the same language, as Guaraní, became the national language of Paraguay, where citizens spoke it along with Spanish. Quechua (including pre-Incaic) spread in the Andes to Argentina and in

Upper Amazonia, where it may have served as a language of trade. Today, at least 13 million people speak one of many dialects of Central and Peripheral Quechua, making it by far the greatest indigenous language family in the Americas. Many languages that defy classification with other known indigenous languages continue to be spoken today, especially in the Andean nation-states with substantial Amazonian regions and in the eastern countries of Brazil and the Guianas (see, e.g., Collier et al., 1985).

Rich realms of storytelling, often called mythology, are legion among native peoples, Afro-Latin American peoples, and people of 'mixed' descent, among whom they constitute sources of spirituality and historicity embedded in more than 100 languages (see, e.g., Hill 1988). They convey rich information ranging from medicinal plants to healing procedures to ways of thinking about humanity and society that confront the models of hierarchically organized and exploitative nation-state bureaucracies. Such telling of pasts motivates symbolically the development of political movements such as that of Pachakutik in Ecuador in the mid-1990s, and the Ecuadorian indigenous uprisings in that nation during the same decade, continuing into 2012 (Stern, 1987; Whitten and Whitten, 2011).

Within the socioeconomic classes oriented toward national politics, which have long been articulated to the global economy, the figure of the *caudillo* has loomed large since the end of the colonial era. A *caudillo* is a powerful figure, usually a man but sometimes a woman, such as the late Eva Perón in Argentina, who emerges in a time of political-economic or sociocultural crisis and attracts followers from many sectors of national life, including radically opposed political parties and classes. One of the greatest *caudillos* of all time was the liberator Simón Bolívar. The *caudillo* rallies such followers to his or her own personal benefit, thereby often losing considerable influence, and even office, when the crises are over. It is to the benefit of the *caudillo*, then, to maintain an aura of crisis, which in South America today, and back through its tumultuous history, is not difficult to do. The *caudillo* is matched in the indigenous sector by the *cacique* (sometimes called by the Quechua term *curaca*), who shares the qualities just described. (*Cacique* is also used for leaders in organized crime.) Both *caudillos* and *caciques*, as leaders in peripheral nations, depend in large part on economic gains to come from industrialized, wealthy nations. The cultural form of dependence of South American countries is closely tied to the developmental policies of the center nations, whose funding for externally designed projects often supports *caudillismo* and *caciquismo* (e.g., Whitten and Whitten, 2011).

Modernities and Alternative Modernities

Modernity among the elite and middle classes of South America is tied to urbanity, whiteness, and Euro-North Americanized consumer culture. The doctrine of *el mestizaje* serves these interests in the sense that the identity of mestizo underscores the processes of social whitening (*blanqueamiento*). Similarly, the representations as mestizo of those below the elite or middle classes refers to the darker complexions and stigmatized and racialized vulgarity of the common people, who in many nations constitute the vast majority. For people of indigenous and Afro-Latin American descent, and for others classed by the elite as 'mestizos,' modernity consists of passing back and forth among different systems of thought and behavior. In the past 15 years, the idea of interculturality, which refers to honing the powers of passage from one cultural system to another, has gained considerable momentum in many regions of South America (Whitten and Whitten, 2011).

Passages themselves give rise to new dimensions of power politics in South American nations, just as they gave rise to the Creole imagined nationalities of the nineteenth century (Anderson, 1991). The alternative modernities in the Euro-American-dominated global economy that spawned earlier state nationalism are now sowing the seeds of ethnic nationalism, manifest in many countries by the self-conscious rise of new nationalities out of the symbolic transformations and continuities of syncretic formations. Public displays of collective selfhood by indigenous people tap significant financial resources through nongovernmental organizations of Europe and North America, giving such social formations and movements a new articulation to economic resources. Black selfhood, however, does not attract significant financial resources.

The nations of South America vary greatly from one another, and the salient characteristics of their respective modernities also vary. Twentieth-century migration has changed the complexion of many countries, as Europeans (Jewish, Catholic, and Protestant), Middle Easterners (Jewish, Christian, and Islamic), and Asians (Chinese, Japanese, Korean, and Vietnamese) have become prominent actors in various class sectors. In the southern cone nations of Argentina, Chile, and Paraguay live significant populations of German immigrants. The ex-president of Peru, Alberto Fujimori, is the son of Japanese immigrants, and the Ecuadorian ex-president, Jamil Mahuad Witt, is the son of Lebanese and German immigrants.

South American nationality is often expressed through public performances as crystallizations of 'authentic' local-level or regional cultures (Guss, 1999). Concepts of cultural hybridity and alternative modernity fold ethnic and cultural diversity into a unified national pride. Festive centralization is often matched by local-level and regional divergence, as people in many walks of life appropriate nationalist stereotypes about authenticity into their own self-presentations. For example, indigenous people of Andean Ecuador have long performed individualized constructions of the European festival of Corpus Christi. Cultural images of rainforest inhabitants are evoked in these performances. In response to both the politicized indigenous movements and the craving of foreign tourists for 'authentic' Incaic performative reproductions, some of these festivals have been renamed 'Inti Raymi,' solstice of the sun. Performance of them becomes transformed into a nationalist, 'mestizo' rendition of Incaic authenticity and also an indigenous rendition of their increasing cultural spaces in the modern nation-state (Whitten and Whitten, 2011).

In the late twentieth century, nationalist governments used their expanding military powers to stress centralization and homogenization, and pan-Roman Catholicism continues to be a centripetal force at the beginning of the twenty-first century. Nonetheless, indigenous and black movements of South America, together with labor movements, feminist movements,

and various cultural emphases promoting counterhegemonic diversity, work against the exclusive imagery of ethnic blending in their respective nation-states and across national boundaries. New and dynamic cultural mosaics are constantly created out of the interplay of secular-sacred centralization, ethnic-cultural diversity, and interculturality. The dynamic, transformative balance between centralization and diversity that began in early colonial times continues into the beginning of the twenty-first century, although again, interculturality tempers this balance (Silverblatt, 2004).

New Socialisms, New Internationalisms, and Enduring Struggles for Diversity

ALBA, the Bolivarian Alliance for the Peoples of Our America, was conceived and constructed by President Hugo Chávez of Venezuela in 2004 to strengthen his socialist revolution in that country. It began as an international organization with an alliance with President Fidel Castro of Cuba, as an exchange of Venezuelan petroleum for Cuban medical assistance. Bolivian President Evo Morales joined the alliance in 2006 to strengthen his Movement to Socialism, and President Rafael Correa of Ecuador joined in 2009 to underscore the 21st Century Socialism movement. Other nations of the Caribbean and Central America also joined. In 2008, UNASUR, the Union of South American Nations, formed, and in 2009 its headquarters became lodged in Quito, Ecuador, and its parliament in Cochabamba, Bolivia. By 2010, UNASUR comprised all 12 South American nations. One explicit platform espoused from time to time by spokespeople in UNASUR is the goal to replace some of the functions of the Organization of American States (OAS), seen by many to be dominated by the United States, as the focus for South American development. Accordingly, plans are underway for a new currency, improved communications systems, and a regional central bank. UNASUR seems to represent a South American analog to the European Union. More recently, another organization, the Community of Latin American and Caribbean States has emerged to explicitly replace the OAS (Whitten and Whitten, 2011).

Within these alliances, the three South American nations of ALBA profess, in various degrees, an outright hostility to aid from and alliance with the United States and seek instead 'lateral' agreements with other countries, especially China, Iran, and Russia to the east, and Canada to the North. Iranian President Mahmoud Ahmadinejad has been especially active in courting alliances with Brazil, Venezuela, and Ecuador, and some US writers say his aim is to acquire uranium from South America in exchange for expertise in developing nuclear power there. The interest in nuclear power in Venezuela and Ecuador certainly exists, but it appears that the latter country, at least, has no uranium to mine, let alone exchange (Whitten and Whitten, 2011).

Paradoxically, to many, the ALBA nations, and to some extent those of UNASUR, have adopted plans, policies, and practices of neoliberal development, particularly in the areas of petroleum exploitation, gold mining, and copper mining, to promote and develop a system of redistribution to the poor and the ethnically disenfranchised (the latter often being indigenous and Afro-Latin American peoples). One of the most contentious and paradoxical issues that has arisen in these three countries is the privatization of water by large foreign and domestic corporations as a socialist strategy for development and to counter poverty. A key to the infrastructure of ALBA is the alliance of *caudillos*, with Hugo Chávez (before his death) and Evo Morales the most forceful, and Ecuador's Rafael Correa somewhere between leftist rhetoric and (at times) moderate practices and policies. The presidents of Brazil, Chile, and Argentina play moderating roles in their UNASUR positions that nonetheless influence the ALBA alliance.

In spite of their very strong public proindigenous and Afro-Latin American ideological positions, Chávez (who claimed descent from indigenous and Afro-Venezuelan parentage), Correa, and even Morales (himself indigenous Aymara) have taken very heavy-handed moves to disenfranchise indigenous people. This is done by giving state contracts to foreign countries to exploit subsurface resources in indigenous and Afro-Latin American territories. In all three countries, indigenous and Afro-American people (and those of other ethnicities) have been granted autonomous use of their land. But the state (in the three socialist countries, as in other nation-states of the region) holds all rights to subsurface resources to be exploited to help development. The regions most affected by subsurface exploitation run down the Andean zone on the eastern and western slopes and lowlands, where a large number of indigenous people and Afro-Latin American people reside. In the Andes, where the majority of indigenous people live, large-scale mining efforts overwhelm any efforts at native territorial autonomous usufructuary activities.

The result in the past decade has been a series of seemingly unending clashes between growing and allying indigenous organizations and the states of Venezuela, Colombia, Ecuador, Peru, and Bolivia. Blood has been shed on both sides of the state–ethnic divide, and possible solutions to the conflicts often lead to even more conflict and stronger international and intraethnic alliances. Added to these clashes are two factors: the socialist states seek stronger and stronger suppression of the free press, while indigenous and to a lesser extent Afro-Latin American peoples have increasing access to media outlets, especially through direct e-mail and websites easily found by Google and other search engines.

Finally, it must be noted that the participation of indigenous people, and to a lesser extent black people, in national literary, bureaucratic, developmental, governmental, and entrepreneurial activities has increased and is increasing. Urbanity, higher education, foreign travel, bilingualism or multilingualism, and other traits erroneously reserved for nonindigenous people are quite salient among indigenous people, and manifest by congressional representatives, physicians, lawyers, and scholars. In short, the nation-state images of indigeneity, as confined to rural regions and in need of state-sponsored development and especially training in Spanish and Portuguese, are erroneous. They take us back to the colonial imagery of the foundation of a Republic of Spaniards and a Republic of Indians, with no colonial space other than slavery for African descendants. It is against this imagery that the struggle for international recognition in their own right by indigenous and Afro-Latin American peoples sporadically surges and endures.

See also: Americas, Sociocultural Overviews: Central America; Education in Latin America and the Caribbean: Systems and Research; Historical Thought and Historiography: Indigenous Cultures in the Americas.

Bibliography

Ades, D., 1989. Art in Latin America: The Modern Era, 1820–1980. Yale University Press, New Haven, CT.

Anderson, B., 1991. Imagined Communities, second ed. Verso, London (1983).

Bryant, S.K., O'Toole, R.S., Vinson III, B.V. (Eds.), 2012. Africans to Spanish America. University of Illinois Press, Urbana, IL.

Collier, S., Blakemore, H., Skidmore, T.E. (Eds.), 1985. The Cambridge Encyclopedia of Latin America and the Caribbean. Cambridge University Press, Cambridge, UK.

Denevan, W.M. (Ed.), 1976. The Native Population of the Americas in 1492. University of Wisconsin Press, Madison, WI.

Dillehay, T.D., 2000. The Settlement of the Americas: A New Prehistory. Basic Books, New York.

Escobar, A., 2008. Territories of Difference: Place, Movements, Life, Redes. Duke University Press, Durham, NC.

Fuentes, C., 1992. The Buried Mirror: Reflections on Spain and the New World. Houghton Mifflin, New York.

Goodman, E.J., 1992. The Explorers of South America. University of Oklahoma Press, Norman, OK (1972).

Guss, D.M., 1999. The Festive State: Race, Ethnicity, and Nationalism as Cultural Performance. University of California Press, Berkeley, CA.

Hall, G.D., 2005. Slavery and African Ethnicities in the Americas: Restoring the Links. University of North Carolina Press, Chapel Hill, NC.

Hill, J.D. (Ed.), 1988. Rethinking History and Myth: Indigenous South American Perspectives on the Past. University of Illinois Press, Urbana, IL.

Hornborg, A., Hill, J.D. (Eds.), 2011. Ethnicity in Ancient Amazonia: Reconstructing Past Identities from Archaeology, Linguistics, and Ethnohistory. University Press of Colorado, Boulder, CO.

McKnight, K.J., Garofalo, L.J. (Eds.), 2009. Afro-Latino Voices: Narratives from the Early Modern Ibero-Atlantic World. Hackett, Indianapolis, IN, pp. 1550–1812.

Price, R. (Ed.), 1996. Maroon Societies: Rebel Slave Communities in the Americas, third ed. Johns Hopkins University Press, Baltimore, MD.

Redstall, M. (Ed.), 2005. Beyond Black and Red: African-Native Relations in Colonial Latin America. University of New Mexico Press, Albuquerque, NM.

Richardson III, J.B., 1994. People of the Andes. St Remy Press, Montreal, Canada and Smithsonian Press, Washington, DC.

Silverblatt, I.M., 2004. Modern Inquisitions: Peru and the Colonial Origins of the Civilized World. Duke University Press, Durham, NC.

Silverman, H., Isbell, W.H., 2008. South American Archaeology. Springer, New York.

Stern, S.J. (Ed.), 1987. Resistance, Rebellion, and Consciousness in the Andean Peasant World. University of Wisconsin Press, Madison, WI.

Uzendoski, M.A., Calapucha-Tapuy, E.F., 2012. The Ecology of the Spoken Word: Amazonian Storytelling and Shamanism among the Napo Runa. University of Illinois Press, Urbana, IL.

Van Cott, D.L. (Ed.), 1995. Indigenous Peoples and Democracy in Latin America. St Martin's Press, New York.

Varner, J.G., Varner, J.J., 1983. Dogs of the Conquest. University of Oklahoma Press, Norman, OK.

Whitten Jr., N.E., Torres, A. (Eds.), 1998. Blackness in Latin America and the Caribbean: Social Dynamics and Cultural Transformations, vols 2. Indiana University Press, Bloomington, IN.

Whitten Jr., N.E., Whitten, D.S., 2011. Histories of the Present: People and Power in Ecuador. University of Illinois Press, Urbana, IL.

Williams, J.M., Lewis, R.E. (Eds.), 1993. Early Images of the Americas: Transfer and Invention. University of Arizona Press, Tucson, AZ.

South Asia, Archaeology of

Namita Sugandhi, Binghamton University, Binghamton, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.L. Possehl, volume 22, pp. 14612–14622, © 2001, Elsevier Ltd.

Abstract

This article outlines the major cultural periods that have been archaeologically documented in South Asia. Beginning with the Paleolithic and Mesolithic eras, the sequence of cultures from the Neolithic/Chalcolithic, Iron Age, Early Historic, and Medieval periods is presented, with some attempt to characterize the significant regional variation that is seen throughout time.

Nearly two and a half centuries of archaeological research in South Asia have uncovered rich and diverse traditions of prehistoric and early historic cultures spanning millennia. Data from excavation, survey, ethnoarchaeological work, and experimental research have provided a wealth of knowledge about the South Asian past that would be impossible to exhaustively detail in this short article. Although many have critiqued the practice of adopting a periodization scheme based on the European sequence or assumptions of social evolution, the chronology of South Asia is still conventionally divided into stages based on technology, typology, and subsistence regime – though absolute dating techniques also provide some anchors. This article will follow a roughly chronological sequence from prehistoric to modern times, and will outline major regional traditions associated with each period.

The Paleolithic

The Paleolithic is divided into Lower, Middle, and Upper phases, although other terms such as Pre-, Early, and Late Acheulean are also used. Stone tools associated with all three phases have been reported from across the Indian subcontinent. Many tools occur as surface finds but artifacts have also been recovered from stratified contexts. Many sites have extremely long sequences, though there is also discontinuity, indicating the transient and shifting nature of habitation during this time.

Early evidence for hominin activity across South Asia is sporadic until the early Middle Pleistocene, though there are claims that date back as early as 2.2 Ma at Riwat and the Pabbi Hills in the Siwalik deposits of northern Pakistan (Chauhan, 2009a). The earliest evidence consists largely of chopper and cleaver Oldowan or Mode 1 pebble tools that are found at several locations across the Indian subcontinent. The Lower Paleolithic is divided into core-flake and Acheulean industries. Early Acheulean assemblages dated prior to 400 kya are dominated by biface hand axes and choppers and distributed across the subcontinent, from northern Pakistan to Nepal, Karnataka, and in the Kortallayar valley in Tamil Nadu. Later Acheulean sites are more numerous, particularly in central and peninsular India, and are identified by a greater percentage of flake tools and use of prepared cores such as the Levallois technique (Ghosh, 1989).

The Late Acheulean–Middle Paleolithic transition is poorly understood as prepared core techniques persist into the terminal Pleistocene. Middle Paleolithic assemblages have been dated to between 125 kya and 40 kya, and are distinguished by a shift toward smaller tools and increasing use of siliceous raw materials such as chert, chalcedony, and jasper. Although only a few sites such as Nevasa in Maharashtra and Samnapur in the Narmada Basin are well stratified, Middle Paleolithic assemblages have been recovered from a wide range of ecological zones from the high-altitude environments of the Himalayan and sub-Himalayan regions at sites such as Sanghao cave near Peshawar in Pakistan, the Alakananda valley in Northern India, and Arjun 3 in Nepal, to the semiarid environment of the middle Krishna valley and Shorapur Doab in South India (Chauhan, 2009b).

The archaeology of the Paleolithic eras has mostly focused on the identification of stone tool types and understanding of subsistence. By the beginning of the Upper Paleolithic, evidence for more symbolic behavior is demonstrated by ornamental objects such as ostrich shell beads and circular stone platforms indicating possible ritual activity (Ghosh, 1989). Stone tool assemblages show a prevalence of parallel-sided blades and burins, and tools of other materials such as bone are also found. Important Upper Paleolithic sites include Baghor II in the Son Valley, Patne in Maharashtra, and Kurnool caves in Andhra Pradesh (Raju and Venkatasubbaiah, 2002). Hundreds of other Upper Paleolithic sites and finds have been documented across the Indian subcontinent, and show evidence for long-term but intermittent occupation. At the site of Bhimbetka, located in the Vindhyas, a series of occupations from the Lower Paleolithic onwards has been documented, and includes some of the earliest cave paintings in South Asia, dated to 30 000 BP (Chakrabarti, 1999).

The Mesolithic

The Pleistocene–Holocene transition is often associated with a 'Mesolithic' phase, which may be identified stratigraphically, or in terms of technology and subsistence. Like many transitional periods the South Asian Mesolithic is difficult to delineate and define (Chakrabarti, 1999). This phase is associated with the adoption of a wider spectrum of resources, and movement into a broader range of ecological zones than in

previous periods. Technologically, the Mesolithic period is characterized by small flake tools or 'microliths.' Although microliths are found during many preceding and succeeding phases of South Asian prehistory, their association with the Mesolithic remains strong, and the period is sometimes referred to as a 'microlithic phase' (Ghosh, 1989). Mesolithic remains have been reported across South Asia, often at sites with evidence for earlier Paleolithic occupation. However, only a few Mesolithic sites have been excavated including Bagor in Rajasthan; Chopani Mando in the Belan valley; and Sarai Nahar Rai, Mahadaha, and Damadama in the Ganga valley (Misra, 2002). In Sri Lanka, some of the earliest human remains have been discovered along with microlithic tools and artifacts made of bone and shell at sites such as Fa Hien Lena, Batadomba-lena, and Beli-lena, but sites dating prior to the Mesolithic are few and require additional verification (Deraniyagala, 1992).

The Neolithic and Chalcolithic Periods

As climatic conditions became more favorable during the Holocene, numerous communities across South Asia began the process of plant cultivation and animal domestication. This development is generally associated with the beginnings of the Neolithic period, but it is a process which began at different times across South Asia between the eighth and third millennia BCE. During this span of time, some regional cultures are characterized as 'Chalcolithic' based on the presence of copper artifacts. Copper begins to appear in limited quantities by the fifth millennium BCE, increasing in frequency over the next two millennia. Ground stone tools, early pottery, and evidence for permanent settlements are also associated with 'Neolithic' period sites across South Asia; however, there is much regional variability in Neolithic stone tool assemblages. Similarly, previous assumptions about the link between domestication, ceramics, and sedentism are now under question in many areas. Although there was a marked increase in sedentism between the eighth and third millennia BCE, various communities continued to practice more mobile lifestyles associated with pastoral or foraging subsistence strategies. This heterogeneity continues across South Asia to this day and must be factored into any consideration of past social dynamics.

Once believed to be a singular development, paleobotanical studies have now confirmed that agriculture was initiated at multiple locations across South Asia (Fuller, 2011). The earliest evidence comes from Baluchistan at the site of Mehrgarh in the Bolan valley, where the transition from seminomadic pastoralism to sedentary agriculture between the eighth and sixth millennia BCE may have been influenced by developments in Western Asia (Allchin and Allchin, 1982). Excavations at Mehrgarh yielded evidence for house structures, burials, and domesticates such as barley, wheat, and cotton. There is evidence for the gradual domestication of cattle, goats, and sheep, though wild game continues to form an important part of diet – this is a pattern seen across South Asia throughout time.

Several other sites in the northwest have provided evidence of small agricultural settlements between the sixth and fourth millennia BCE including Kili Gul Mohammad, Anjira, Mundigak, Gumla, and Sarai Khola (Ghosh, 1989; Singh, 2009). At these sites and many others, Neolithic levels are identified by features and artifacts such as mud brick structures, querns, pottery, and by paleobotanical and faunal evidence for domesticated species (Allchin and Allchin, 1982). Many sites and regions are associated with specific types of pottery traditions including Nal, Kulli, and Hakra ware sites. Copper objects, terracotta figurines, and beads made of semiprecious materials have also been recovered from Neolithic levels in many places, and excavated human burials attest to a variety of funerary customs during this time (Ghosh, 1989).

Elsewhere in South Asia, early Neolithic traditions have been documented in numerous regions, many having distinct traits marking them as potential loci of indigenous development. Rice is an early domesticate in the Ganga valley and surrounding plains, as well as in the northern fringes of the Vindhyas (Fuller, 2011). Though paleobotanical evidence for rice is documented from very early periods in these regions, understanding of its transition from a wild to cultivated crop remains ambiguous and is possibly dependent on the timing of hybridization of *indica* rice with *japonica* species introduced from China (Fuller, 2011). At sites such as Mahagara, Koldihwa, Chirand, and Senuar, Neolithic settlements have been dated from the fourth to third millennia BCE, though some earlier dates have also been suggested. In the Kashmir valley, Neolithic sites such as Burzahom and Gufkral are distinguished by the presence of plastered pit dwellings between the third and second millennia BCE. In the Northeast, Neolithic stone celts and axes have mainly been collected as surface finds. Neolithic levels at Sarutaru, Marakdola, Napchik, and Selbalriri have yielded some absolute dates in the range of the mid-second millennium BCE but there are suggestions that some of these sites may date to much later periods (Ghosh, 1989; Chakrabarti, 1999; Singh, 2009).

Further to the south, early agricultural village sites have been uncovered in Orissa and Bengal such as Kuchai, Gopalpur, and Pandu Rajar Dhibi, while in peninsular India, a number of Neolithic and Chalcolithic cultures have been documented, including the Ahar, Ganeshwar-Jodhpura, Malwa, and Jorwe complexes (Ghosh, 1989). These cultures have been dated from the third to late first millennia BCE and are contemporaneous with the Indus civilization. At sites such as Ahar, Gilund, Navdatoli, Daimabad, and Inamgaon, excavations have revealed evidence for copper artifacts in addition to stone implements. The presence of nonlocal materials such as shell, gold, ivory, and semiprecious stones indicates the existence of interregional networks of exchange across South Asia during this time.

In South India, the Neolithic traditions of the Southern Deccan are marked by unique patterns of material culture and paleobotanical evidence, suggesting this region was an independent center of early domestication. The cultivation of pulses and millets by the third millennium BCE at sites such as Hallur, Sanganakallu, Kurugodu, and Tekkalakota gradually intensified to include a wide range of local and introduced species (Fuller et al., 2004). A distinctive feature of the early 'Southern Neolithic,' confined mainly to Bellary and Raichur districts in Karnataka, are Neolithic ashmounds. These monumental accretions of vitrified dung indicate the importance of cattle in

the agro-pastoral economy and possibly represent the emergence of communal ritual activity (Paddayya, 1993; Johansen, 2004). Further south, Neolithic levels at Paiyampalli in Tamil Nadu have been radiocarbon dated to the mid-second millennium BCE. Very early dates have been suggested for the cultivation of barley, oats, and possibly rice in the Horton Plains (Premathilake, 2008), but overall the transition from the Mesolithic period to the better documented Iron Age remains poorly understood in Sri Lanka.

The Indus Civilization

As the small farming and pastoral communities of peninsular India began to develop, the earlier Neolithic societies of the Northwest were coalescing into a widespread tradition beginning in the fourth millennium BCE. The Indus or 'Indus Valley' civilization is also known as the 'Harappan' culture, in reference to the site of Harappa which, along with Mohenjo Daro, is one of the best known of the Indus urban sites. However, hundreds of other sites have now been documented across a vast region stretching from the foothills of the Himalayas to Iran and Afghanistan in the west, eastwards into the Ganga–Yamuna doab, and across Saurashtra to the Gulf of Khambhat (Kenoyer, 1998). These sites range in size from large urban centers greater than 100 ha to much smaller sites in the 1–5 ha range. Several periodization schemes have been developed to characterize the emergence and eventual decline of the Indus civilization including the Early (*c*.3200–2600 BCE), Mature (*c*.2600–1900 BCE), and Late (*c*.1900–1300 BCE) Harappan phases; a fourfold scheme of Early Food Producing (pre-6000–4000 BCE), Regionalization (4000–2500 BCE), Integration (2500–2000 BCE), and Localization (2100–1300 BCE) eras; and several more finely tuned regional and local chronologies based on pottery types or the stratigraphic sequence of a particular site (Ghosh, 1989; Possehl, 2002).

The earliest stages of the Indus civilization have been documented primarily along the Indus river system and in Kutch. This preurban phase is marked by the appearance of larger walled settlements and an overall intensification in agriculture, craft production, metallurgy, and trade. Many early sites such as Mohenjo Daro, Harappa, Rakhigarhi, Dholavira, and Ganweriwala grew into large cities by the mid-third millennium BCE (Kenoyer, 1998). In addition to this expansion, hundreds of new settlements were constructed across a vast and diverse landscape encompassing nearly one million square kilometers. Although much regional variation exists, this period is considered a time of standardization, stability, and general cultural efflorescence. With the exception of Saurashtra, where stone was frequently used, most Indus settlements were constructed of mud brick and baked brick architecture (Chakrabarti, 1999).

Hallmarks of the mature Indus tradition include the mass production of distinctive ceramic wares such as black on red Harappan ware and other large-scale manufacturing including shell working, bead making, and copper and bronze production. Thousands of seals and sealings have been recovered, many with common motifs such as the unicorn, humped bull, various wild animals, and a horned male figure sometimes referred to as a 'proto-Shiva' (see **Figure 1**). Many seals and other small objects bear short inscriptions in a logo-syllabic script that remains as yet undeciphered. Settlements of all sizes are laid out according to a walled grid plan, and have extensive water management systems for civic and domestic purposes (see **Figure 2**). Typical Indus craft items such as beads; shell ornaments; terracotta figurines; and objects of copper, bronze, silver, and gold have been recovered from many different types of sites, indicating the existence of strong intraregional networks linking villages, towns, and larger urban centers (Kenoyer, 1998; Possehl, 2002).

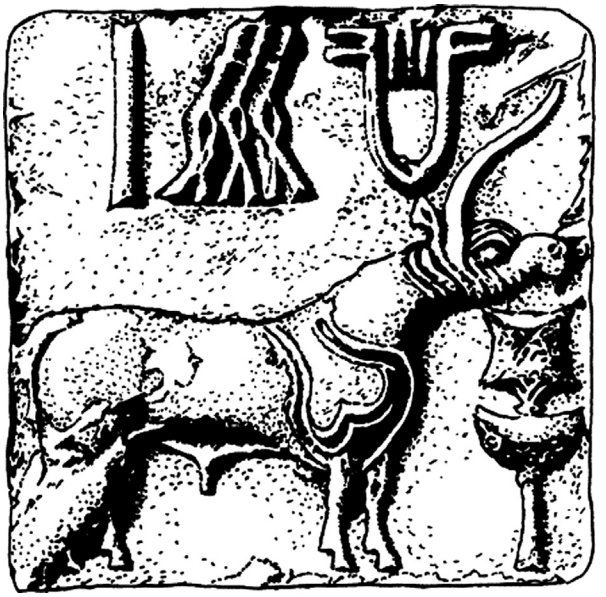

Figure 1 Harappan seal (from harappa.com).

Comparison of house sizes and human burials at many Indus sites suggests the existence of some social division and unequal access to wealth during this period. However, a distinctive absence of monumental architecture, opulent mortuary remains, or other forms of lavish elite display make it difficult to characterize the nature and structure of Indus society. Interpretations have ranged from ideas of centralized imperial control, to theories based on economic linkages (cf Ratnagar, 1991; Shaffer, 1982). Different interpretations of Indus religion have been put forward, including links to later Hindu practices. No Indus temples have ever been documented but evidence for ritual activity is materially marked by small objects and features such as figurines, ring stones, and brick-lined pits interpreted as 'fire altars.' Additionally, it has been suggested that the sophisticated hydraulic features found at most Indus sites indicate a ritual association with water or bathing (Possehl, 2002). Barring decipherment of the Indus script, or any more conclusive evidence, our understanding of religion during Indus times remains elusive.

Although our knowledge of political and social organization is still uncertain, the Indus civilization is typically classified as a 'state' level society, with features of complexity comparable to the contemporaneous Bronze Age societies of the Old World in Mesopotamia, Egypt, Western and Central Asia, and the Arabian Peninsula. Evidence for trade between

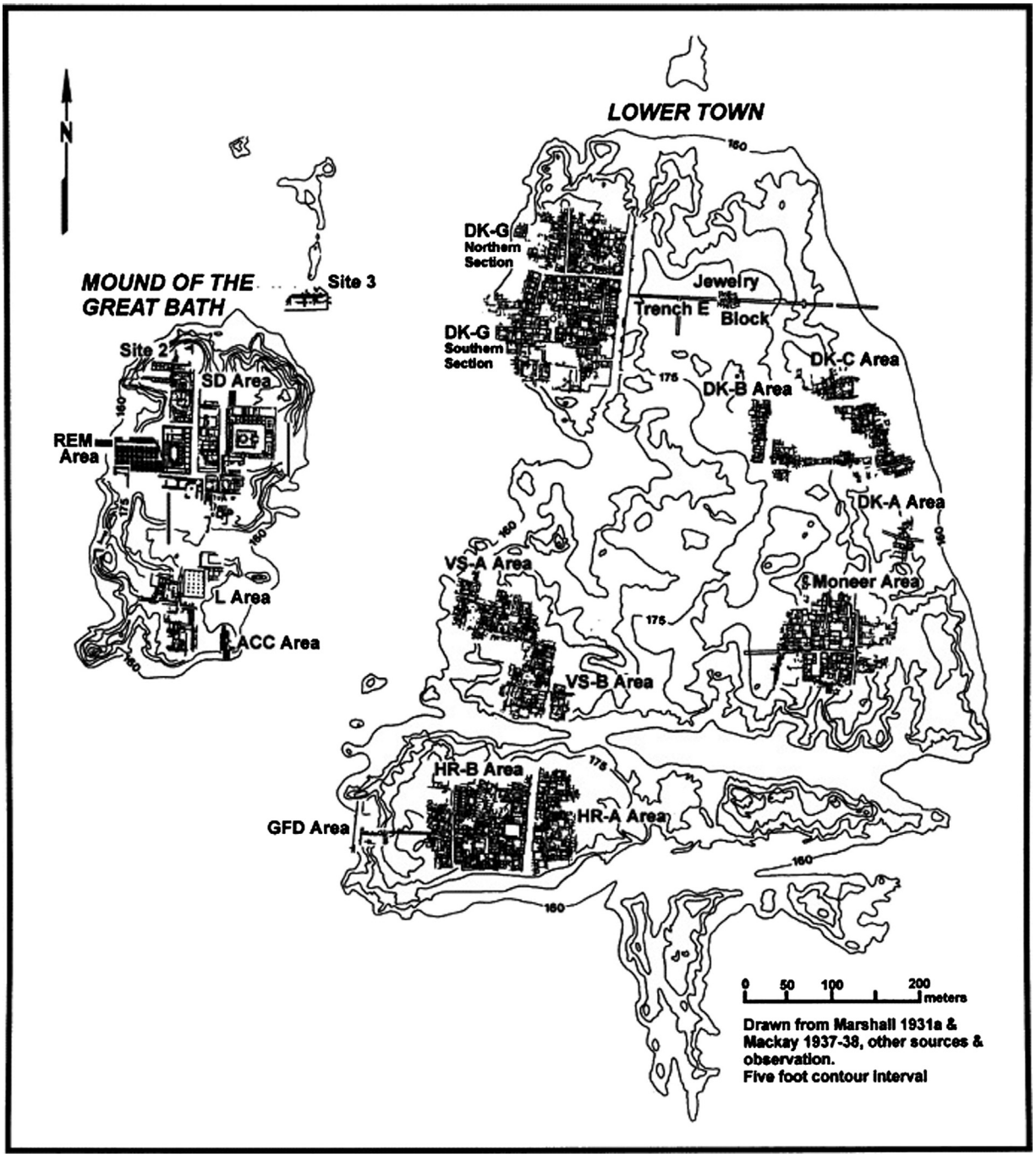

Figure 2 Mohenjo Daro (from harappa.com).

these societies has been documented by the sourcing of raw materials and artifacts, and by the Mesopotamian historical record. Harappan seals, pottery, and carnelian beads have been found at several sites in Mesopotamia, Western Asia, and the Persian Gulf, but few foreign artifacts have been recovered from Indus sites (Singh, 2009). Sumerian cuneiform texts referring to trade mention semiprecious stone; metals such as gold, silver, and copper; ivory; ebony; and various types of animals from 'Mehluhha' in exchange for fish, grain, and wool. Additionally, there is also evidence for exchange networks linking the Indus civilization to the Neolithic and Chalcolithic communities occupying other regions of South Asia (Possehl, 2002). The importance of trade and commerce to the Indus society is clear, and has been linked to both its

rise and fall. While trade may not have been a direct factor, it was certainly a significant activity in the Indus social and economic milieu and is an important avenue for further research.

The Iron Age

The second millennium BCE brought sweeping changes across much of South Asia, although the cause and nature of these transformations remain poorly understood. They have often been correlated with an overall contraction of Bronze Age societies in the Old World, but explanations of this broader decline remain speculative and the phenomenon is most likely due to a number of environmental and social factors. In the Northwest, the 'Localization' era of the late Indus civilization is characterized as a time of dissolving urban traditions and diminishing long-distance and interregional interaction. Settlements decrease in terms of both size and number, and there is a significant shift in material culture that varies by region.

Distinctive cultures such as the Jhukar and Cemetery H groups are identified by new forms of pottery and inhumation in the posturban layers of sites in Sindh and Punjab. To the north, the Gandharan Grave Culture has been dated from the early second millennium to the third century BCE and is characterized by a number of inhumation styles, objects of copper and bronze, and some evidence for iron and horses (Singh, 2009). The material cultures of these regions show some continuity with pre-Harappan and Harappan traditions but also have links with Central Asia. Theories that northern 'Aryan' invaders were responsible for the Indus decline have now been widely disproven but there is evidence for a gradual migration of Indo-European speaking groups into the Indian subcontinent during this period. These groups are difficult to trace archaeologically but linguistic studies have linked them to the origins of the Sanskrit language and the earliest layers of Vedic tradition. It is likely that the practices of these new groups merged with the indigenous traditions they encountered to form the ritual and social customs that developed in the post-Indus period. Although agriculture practices continued to diversify during this time, a shift toward pastoralism is often hypothesized.

The post-Indus period is conventionally depicted as illiterate and nonurban but there is strong evidence to suggest continued settlement in many regions, particularly to the east and south of the Indus valley. Sites such as Daimabad in the upper Godavari Valley and Rojdi in central Saurashtra show some evidence for expansion during later Harappan levels (Possehl, 2002). Archaeologically, the cultures of northern and central India are identified through a range of material forms such as Ochre Coloured Pottery, which has been discovered at sites across the Ganga–Yamuna doab, and copper hoards found primarily in the upper Ganga valley and eastern India (Ghosh, 1989). In addition to copper objects, more widespread evidence for iron use is documented across South Asia between the mid-second and mid-first millennia BCE and thus the period is increasingly referred to as the 'Iron Age,' though the use of 'Vedic period' as well as regional terms also persists. Iron yielding levels are associated with a ceramic type known as 'Painted Grey Ware' from the Indus to Ganga valleys.

In peninsular India, continuity is seen at many earlier Neolithic and Chalcolithic sites, where the Iron Age is associated with an intensification in agricultural practices, increasing settlement size, and the construction of monumental features known as megaliths. Megaliths have been documented in various contexts across South Asia and in many regions may have been built up until recent times (Moorti, 1994). Megaliths dated to the Iron Age are constructed in a range of forms such as dolmens, cairns, menhirs, and circular alignments. Some contain human inhumations and various objects such as iron implements, pottery, or faunal remains (Moorti, 1994; Brubaker, 2000–2001). In addition to mortuary purposes, megaliths were sometimes constructed to mark socially differentiated space and thus may serve as an indication of increasing political complexity in South India (Bauer, 2011). A ceramic type known as 'black and red ware' begins to appear more frequently, though it is seen during earlier and later periods as well. The Iron Age in Sri Lanka is also better documented than previous periods. At sites such as Anuradhapura and Ibbankatuva, early iron bearing levels have been dated as early as the eight and ninth centuries BCE (Deraniyagala, 1992).

The Early Historic Period

As with other cultural periods in South Asia, the Early Historic has varying start dates across different regions. By the sixth century BCE, many settlements begin to grow into cities once again, thus marking the beginning of what is known as the 'second urbanization' in northern India (Allchin, 1995). Initially centered in the Ganga valley, this urban expansion continued to spread across the northern and northwestern Indian subcontinent over the next two centuries and is associated with early literary traditions and the emergence of small regional states known as *janapadas*. Though there are no extant written records, several later texts have been assigned original composition dates of the mid-first millennium BCE. The political genealogies suggested by these texts can be cross-referenced with the numismatic record which also emerges around this time. Additionally, the development and spread of heterodox religions such as Buddhism and Jainism contribute their own traditions of literary and material culture.

In South India, the Early Historic period begins about three centuries later, and is based on the date of the Asokan edicts which are written in the Brahmi script, and also in Kharosthi. These inscriptions, which are distributed across the Indian subcontinent, are associated with the third king of the Mauryan Empire (*c*.320–180 BCE), and are the first extant inscriptions to be found in South Asia (see **Figure 3**). Though several possible origins for the Brahmi script have been proposed, its earliest development is still unclear. Research from Tamil Nadu and Sri Lanka has recovered graffiti-marked sherds from layers that

Figure 3 Asokan edict at Erragudi (personal photograph).

Figure 4 View of Sisupalgarh eastern rampart (courtesy of Monica Smith).

predate the Mauryan Empire (Deraniyagala, 1992; Mahadevan, 2003). These early dates may substantially revise our understanding of Brahmi, which is often attributed to foreign or north Indian origins.

With the advent of written records, the narrative of the South Asian past becomes less reliant on archaeology and the study of material remains. Nevertheless, historic period archaeology in South Asia has a long record of research, dating back to the historical geography of Colin Mackenzie in the late eighteenth century. Early archaeological research focused largely on monumental remains such as the early Buddhist stupas at Bharhut, Sanchi, and Sarnath. However, excavations in the early twentieth century at Pataliputra and Bhita in the Ganga Valley, and Taxila in the northwest gave some indication of the growth and character of the renewed urbanism that emerged during the Early Historic period (Chakrabarti, 1995). Although hundreds of cities and smaller settlements have been identified across South Asia, excavations have been limited as occupation at many sites has continued up until the present day. This significantly complicates efforts at more intensive exploration. Settlement evidence from survey and excavation research is often contrasted with details from Puranic and early Buddhist texts and with external references in Greek and Roman literature. More recently, remote-sensing techniques have been employed at sites such as Sisupalgarh and Kausambi in order to clarify additional details about city layout.

Layers associated with the Early Historic period are identified by diagnostic artifacts such as Northern Black Polished Ware, punch-marked coins of bronze and silver, beads made of semiprecious stones such as carnelian and lapis lazuli, and terracotta figurines (Ghosh, 1989). Many cities including Kausambi, Rajgir, Ahicchatra, Mathura, Atranjikhera, Chandraketugarh, and Sisupalgarh are surrounded by extensive ramparts and fortification walls which may have served both defensive and symbolic purposes (Allchin, 1995) (see **Figure 4**). Craft production areas have been identified at sites such as Ujjain (Ghosh, 1989), and more recently at Indor Khera (Menon and Varma, 2010), but understanding of settlement organization remains incomplete. Knowledge of Early Historic social life is still largely reliant on interpretations from the historical and epigraphic records. Inscriptions in Brahmi, Karoshti, and later regional scripts are found with increasing frequency from the third century BCE onwards and are carved in stone and later on copper plates and other objects. As both text and material artifact, the epigraphic record provides a wealth of detail on political, economic, and social forms during the Early Historic and later periods in South Asia.

In peninsular India, most settlements remained relatively small during the first part of the Early Historic period, though some sites such as Anuradhapura in Sri Lanka may have reached urban proportions quite early (Coningham, 1999). While the *janapadas* of the north and northwest were being absorbed into imperial polities such as the Mauryan Empire, the late Iron Age communities of southern India were steadily growing into larger settlements supported by diversifying subsistence regimes that included new cultigens such as rice. New styles of pottery such as Rouletted Ware and Russet-Coated Painted Ware are associated with Early Historic deposits (Allchin, 1995). Although conventional narratives claim that much of south India was incorporated into an almost pan-Indian Mauryan Empire by the third century BCE, recent research has proposed that interaction with northern imperial forces may have been much more limited (Ray, 2008; Sugandhi, 2013). Nevertheless, there is some evidence to suggest the deliberate articulation of northern forms of sovereignty and local traditions of authority during the Iron Age/Early Historic transition in the south. This begins the gradual incorporation of South India into the wider 'pan-Indic' customs characteristic of the latter half of the Early Historic period. This is first seen through the construction of Buddhist monuments, and later by the adoption of Brahmanical political and ritual traditions. Buddhist sites with initial dates of construction as early as the second century BCE have been documented at Kanaganahalli in Karnataka and Amaravati in Andhra Pradesh. Sri Lanka also has many early Buddhist structures that date to Early Historic times. According to historical sources, the first stupa at Anuradhapura was built in the mid-third century BCE (Coningham, 1999).

In contrast to earlier models of imperial conquest, recent narratives look to politically independent networks of commerce and religion as mechanisms of exchange and

integration between the cultures of the north and south (Ray, 2003). Evidence for large-scale craft production has been documented at sites such as Kodumanal in Tamil Nadu and Mantai in Sri Lanka, indicating participation in early economic networks. Trade items from the Mediterranean region have been identified at sites such as Arikamedu in Puducherry, Pattanam in Kerala, and at sites in Sri Lanka. Collectively, this evidence, along with written records, testifies to the early incorporation of South India and Sri Lanka in the interregional and long-distance relationships that traversed South Asia and connected it to the wider world (Ray, 2003). These connections certainly warrant further research in order to better understand the processes of integration that developed across the Indian subcontinent during the Early Historic period.

Understanding of religious networks has largely been inferred from the literary and epigraphic records but can also be understood from the distribution of monastery sites, which are often located at important points along trade routes. Early Historic ritual architecture is largely confined to Buddhist stone structures and the rock-cut caves of the Western and Eastern Ghats, which are mostly Buddhist and Jain during the earliest phases of construction. However, by the close of the Early Historic period, these traditions were on the decline, as the newly emerging *bhakti* (devotional) traditions of Hinduism began to take hold across the subcontinent. The earliest surviving Hindu temples are built of stone and have been dated to the first few centuries of the Common Era (Singh, 2009). Early temples were small structures that are sometimes difficult to identify, but over time they grew into larger complexes such as the sixth-century CE Vishnu temple at Deogarh in Bihar and the seventh-century CE Durga temple built at Aihole in Karnataka (Sugandhi and Morrison, 2011). They are most likely predated by wooden or brick structures, which often form the early core of later stone constructions.

From the limited archaeological evidence available, it is clear that the Early Historic period was an extremely dynamic time in South Asia which saw the emergence of a number of significant and enduring traditions, many of which continue up until the present day. Efforts to encapsulate these multiple developments within a single cultural period have proven frustrating for archaeologists attempting to characterize the regional variation observed in the material record. Nevertheless, understanding this variation remains an important goal in historic period archaeology of South Asia, and efforts to revise the chronology in order to reflect this are currently underway.

The Medieval and Modern Periods

Archaeological research on the subsequent Medieval and Modern periods (*c.*500–1500 CE, 1500 CE–present) is quite limited in contrast to the fields of art and architectural history. Studies of material culture have mostly focused on monumental architecture such as temples and forts, and on the numismatic and epigraphic records. Although many historic settlements across South Asia continued to be occupied into medieval times, little consideration has been given to these remains. An important exception is the research that has been conducted in the Tunghabhadra corridor in Karnataka around the city of Vijayanagara, capital of an imperial polity that held sway over much of South India between the fourteenth and sixteenth centuries CE (see **Figure 5**). Archaeological studies at Vijayanagara have examined issues of craft production, investigated the distribution of sites across the wider metropolitan area, and examined the effects of long-term human land use through paleobotanical studies (Sinopoli, 2003; Sinopoli and Morrison, 2006, 2009). With the start of the Modern period (*c.*1500 CE), the line between archaeology and the history of the discipline becomes blurred. Though a few projects have looked at the colonial use of space, archaeological studies of the Modern period are scarce.

Figure 5 Entrance to Vitthala temple, Vijayanagara (personal photograph).

Discussion

In closing this article of South Asian archaeology, a few final points should be noted. As the descriptions above have hopefully made clear, there is a tremendous amount of variation within and between the different cultural traditions that have been identified by archaeologists over time. As research continues to document this diversity, it is increasingly evident that the immense complexity of the South Asian past is only partially represented by existing culture historical classifications. Moreover, advanced techniques of analysis such as thermoluminescence and isotope studies continue to revise our understanding of chronology, social process, and transition across both space and time.

Although the continuous nature of occupation at many historic and prehistoric sites may provide valuable data on long-term human–environment interaction, it also raises the challenge of site destruction. Much archaeological research conducted in South Asia may be classified as 'rescue work' as modern processes of agriculture, construction, and mining threaten thousands of sites. Attempts at mitigating this destruction include community outreach programs and heritage management strategies adapted to fit the South Asian context. The development of increasingly efficient documentation methods also holds much value, as the pace of contemporary development rapidly outstrips preservation efforts.

Finally, the political use of archaeological knowledge must be mentioned and cautioned against. From the destruction of the sixth century BCE Bamiyan Buddhas in central Afghanistan to the court-ordered excavation of the Ram Janmabhoomi site at Ayodhya in northern India, recent events in South Asia have highlighted the significance that can be attached to archaeological remains and the way perceptions of the past may be used for contemporary political purposes. Archaeology is a science grounded in fact, but it is also an interpretive discipline conditioned by the biases of researchers and the public alike. Future studies must strike a balance between objective research goals and reflection on these more subjective factors as it continues to document various aspects of the South Asian past.

See also: Archaeology, Politics of; Asia, Sociocultural Overviews: South Asia; Buddhism; Food Production, Origins of; Historical Geography; Historical Thought and Historiography: South Asia; Language Families, Archaeology and History of; Religions of India; Third World.

Bibliography

Allchin, B., Allchin, R., 1982. The Rise of Civilization in India and Pakistan. Cambridge University Press, Cambridge.

Allchin, F.R., 1995. The Archaeology of Early Historic South Asia: The Emergence of Cities and States. Cambridge University Press, Cambridge.

Bauer, A.M., 2011. Producing the political landscape: monuments, labor, water and place in iron age Central Karnataka. In: Johansen, P.G., Bauer, A.M. (Eds.), The Archaeology of Politics: The Materiality of Political Practice and Action in the Past. Cambridge Scholars Publishing, New Castle upon Tyne, pp. 83–113.

Brubaker, R., 2000–2001. Aspects of mortuary variability in the South Indian iron age. Bulletin of the Deccan College Post-Graduate and Research Institute 60-61, 253–302.

Chakrabarti, D.K., 1995. The Archaeology of Ancient Indian Cities. Oxford University Press, Delhi.

Chakrabarti, D.K., 1999. India: An Archaeological History – Palaeolithic Beginnings to Early Historic Foundations. Oxford University Press, New Delhi.

Chauhan, P., 2009a. The lower paleolithic of the indian subcontinent. Evolutionary Anthropology 18, 62–78.

Chauhan, P., 2009b. The south asian paleolithic record and its potential for transitions studies. In: Camps, M., Chauhan, P. (Eds.), Sourcebook of Paleolithic Transitions: Methods, Theories, and Interpretations. Springer, New York, pp. 121–139.

Coningham, R., 1999. Anuradhapura. The British-Sri Lankan excavations at Anuradhapura. Salgaha Watta 2. BAR International Series 824, Oxford.

Deraniyagala, S.U., 1992. The Prehistory of Sri Lanka: An Ecological Perspective. Department of the Archaeological Survey, Government of Sri Lanka, Colombo.

Fuller, D.Q., 2011. Finding plant domestication in the Indian subcontinent. Current Anthropology 52 (Suppl. 4), S347–S362.

Fuller, D.Q., Korisettar, R., Venkatasubbaiah, P.C., Jones, M.K., 2004. Early plan domestication in Southern India: some preliminary archaeobotanical results. Vegetation History and Archaeobotany 13, 115–129.

Ghosh, A., 1989. Encyclopedia of Indian Archaoelogy, vols. 1 & 2. Indian Council of Historical Research, Munshiram Manoharlal Publishers, New Delhi.

Johansen, P., 2004. Landscape, monumental architecture and ritual: a reconsideration of the South Indian Ashmounds. Journal of Anthropological Archaeology 23 (3), 309–330.

Kenoyer, J.M., 1998. Ancient Cities of the Indus Valley Civilization. Oxford University Press and American Institute of Pakistan Studies, Karachi.

Mahadevan, I., 2003. Early tamil epigraphy: from the earliest times to the sixth century A.D. In: Harvard Oriental Series, vol. 62. Harvard University Press, Cambridge.

Menon, J., Varma, S., 2010. Reading archaeological evidence: ceramic and terracotta production at Indor Khera, Uttar Pradesh (200 BCE–500 CE). Indian Historical Review 37 (2), 187–216.

Misra, V.N., 2002. The mesolithic age in India. In: Settar, S., Korisettar, R. (Eds.), Indian Archaeology in Retrospect, Prehistory. Indian Council of Historical Research, vol. I. Manohar, New Delhi, pp. 111–125.

Moorti, U., 1994. Megalithic culture of South India: socio-economic perspectives. Ganga Kaveri Publishing House, Varanasi.

Morrison, K.D., 2009. Daroji Valley: Landscape History,Place and the Making of a Dryland Reservoir System. American Institute of Indian Studies, Manohar, New Dehi.

Paddayya, K., 1993. Ashmound investigations a Budihal, Gulbarga district. Man and Environment 18 (1), 57–88.

Possehl, G.L., 2002. The Indus Civilization: A Contemporary Perspective. AltaMira Press, Walnut Creek.

Premathilake, R., 2008. The emergence of early agriculture in the Horton plains, central Sri Lanka: linked to late Pleistocene and early Holocene climatic changes. Pragdhara 18, 91–104.

Raju, D.R., Venkatasubbaiah, P.C., 2002. The archaeology of the upper palaeolithic phase in India. In: Settar, D., Korisettar, R. (Eds.), Indian Archaeology in Retrospect, Prehistory. Indian Council of Historical Research, vol. I. Manohar, New Delhi, pp. 85–109.

Ratnagar, S., 1991. Enquiries into the Political Organization of Harappan Society. Ravish Publishers, Pune.

Ray, H.P., 2003. The Archaeology of Seafaring in Ancient South Asia. Cambridge University Press, Cambridge.

Ray, H.P., 2008. Interpreting the Mauryan empire: centralized state or multiple centers of control? In: Parker, G., Sinopoli, C. (Eds.), Ancient India in Its Wider World. University of Michigan Press, Ann Arbor, pp. 13–51.

Shaffer, J.G., 1982. Harappan commerce: an alternative perspective. In: Pastner, S., Flam, L. (Eds.), Anthropology in Pakistan: Recent Socio-Cultural and Archaeological Perspectives. Cornell University Press, Ithaca, pp. 166–210.

Singh, U., 2009. A History of Ancient and Early Medieval India: From the Stone Age to the 12th Century. Pearson, Delhi.

Sinopoli, C.M., 2003. The Political Economy of Craft Production: Crafting Empire in South India, c.1350–1650. Cambridge University Press, Cambridge.

Sinopoli, C.M., Morrison, K.D., 2006. Land use and settlement in the Vijayanagara metropolitan region: results of the Vijayanagara metropolitan survey. In: Shanmugam, P., Srinivasan, S. (Eds.), Recent Advances in Vijayanagara Studies: An Anthology. New Era, Madras, pp. 91–109.

Sugandhi, N.S., 2013. Conquests of dharma: network models and the study of ancient polities. In: VanValkenburgh, P., Osborne, J. (Eds.), Territoriality in Archaeology. Archaeological Papers of the American Anthropological Association, vol. 22 (1), pp. 145–163.

Sugandhi, N.S., Morrison, K.D., 2011. Archaeology of hinduism. In: Insoll, T. (Ed.), Oxford Handbook of the Archaeology of Ritual and Religion. Oxford University Press, Oxford, pp. 921–933.

South Asian Legal Traditions

Rohit De, Department of History, Yale University, New Haven, CT, USA

Published by Elsevier Ltd.

Abstract

South Asia (comprising the present states of India, Pakistan, Afghanistan, Nepal, Bhutan, Bangladesh, Sri Lanka, and the Maldives) is marked by the diversity of its legal traditions. The founders of comparative law, when dividing the globe into different legal systems, describe South Asian legal culture as Hindu law. However, South Asia is the site for Hindu, Muslim, Buddhist, tribal, British colonial and postcolonial legal regimes, along with the presence of Jain, Zoroastrian, French, Portuguese, and Roman-Dutch legal systems. Unique to South Asia is the fact that while these legal traditions appeared at different points of time, none was able to supplant the others completely and had to evolve a relationship of coexistence.

South Asia (comprising the present states of India, Pakistan, Afghanistan, Nepal, Bhutan, Bangladesh, Sri Lanka, and the Maldives) is marked by the diversity of its legal traditions. The founders of comparative law, when dividing the globe into different legal systems, describe South Asian legal culture as Hindu law. However, South Asia is the site for Hindu, Muslim, Buddhist, tribal, British colonial and postcolonial legal regimes, along with the presence of Jain, Zoroastrian, French, Portuguese, and Roman-Dutch legal systems. Unique to South Asia is the fact that while these legal traditions appeared at different points of time, none was able to supplant the others completely and had to evolve a relationship of coexistence.

Given the constraints of space, this article will focus primarily on the mainland subcontinent (India, Pakistan, and Bangladesh) with a few references to other traditions.

Ancient Indian Legal Traditions

The period from 600 BCE until about the twelfth century AD has been described as the heyday of Hindu law, evidenced by the proliferation of Sanskrit legal texts and commentaries. However, Hindu law stands as markedly different from other religious legal traditions. There is no definite starting point or any doctrinally central event and no central text or figure of a lawgiver (Menski, 2003a). Hindu law is better understood as a family of laws rather than a single unit, functioning at several different registers (Menski, 2003b). Orientalist scholars searching for a code of Hindu law would turn to the large corpus of Sanskrit texts (numbering in the hundreds) that are collectively known as the *Dharmasastra* (Kane, 1968). The Sanskrit word *dharma* has often been translated as law, though in reality it's a more complex concept evolving from the root *dhrt*, which means to 'hold, maintain, and keep.' Dharma signified appropriate behavior and encompassed duties, rights, moralities, conduct, virtues, and obligation. The Dharmasastra texts addressed three major topics: *achara* or appropriate behavior and rituals for different castes and different stages in one's life; *vyavahara*, the rules for resolving doubts about duties; and finally *prayaschitta* or rules about penances for violating dharma. While the earlier dharmasastras assumed compliance, the later texts emphasized punishment or deterrence to ensure that obligations are fulfilled. Rights, obligations, and punishments were all determined by caste, age, and status. Arguably the most famous of the *Dharmasastra* is the *Manusmriti*, or the Code of Manu, which was translated by the orientalist Sir William Jones in 1794. The *Manusmriti*, which laid down elaborate rules governing the behavior of castes and social groups, was celebrated as a compendium of laws because it was one of the first texts that drew authority from dharma itself rather than the gods. However, the primacy given to the Manusmriti has more to do with the limits of colonial imagination that was seeking a law code, rather than its actual reach and influence. Written at the time when organized Hinduism was being challenged by Buddhism and Jainism, the Manusmriti was seeking to shore up the authority of the caste system.

Despite the sophistication of the Dharmasastras or the prestige of their authors, it is unclear how much influence they had on governing everyday lives. These texts propose ideal models of social reality and there is limited evidence on the lived experiences of law. Historical accounts suggest that in most cases, local communities, caste groups, and villages retained considerable autonomy to decide disputes on custom or conventions that they could themselves evolve. Appeals from these did go to royal authority, but most state courts also decided the dispute relying on the community's practices.

Unlike its contemporary civilizations, there is nothing to suggest that there existed any class of persons versed in the Smriti whose profession was similar to that of modern lawyers (Kane, vol.3, 1946, 288). There are only isolated references to the term *niyoga* (or appointed in lieu of), which is the closest one can get to a professional class of lawyers; however, the main requisite here was a personal connection to the petitioner and not legal competence. Some classical scholars have argued that the class of lawyers was not required in Hindu law, because the administration of justice was the concern of the King, who had to ensure not only that the guilty were punished but also that the innocent were not prosecuted (Rocher, 1968).

Islamic Polities

While Muslims have been present in the subcontinent since the eighth century, it was in the eleventh century that the first Islamicate polities were established in the subcontinent following conquests in North India by Persianized Turks. During a period of turmoil in Central Asia, these rulers drew

legitimacy from the fact that they provided order and allowed Islamic life to flourish. Most courts patronized Sufi shaykhs and royal authority was closely tied to their shrines. A number of Muslim dynasties established control over large parts of north, west, and central India, the largest and most influential being the Mughal empire, which during its heyday (1526–1707) extended over most parts of contemporary South Asia.

Notionally, much of the subcontinent was governed by sharia law. However, in practice the sharia was not a rigid legal sacral system in South Asia. While Mughal commentaries on sharia law, such as the Fatwa-e-Alamgiri, explicitly state that the sharia was not applicable to non-Muslims in the subcontinent, historical evidence suggests that it came to be appropriated by all sections of society and formed the basis of a shared normative system. As Farhat Hasan shows, Hindu merchants frequently invoked provisions of Islamic law on the sale of property when dealing not just with Muslims but also with other Hindus. More strikingly, Hindu and Muslim women frequently appealed to sharia principles and moved state authorities to secure their claims to property and guardianship (Hasan, 2004). The sharia through the Mughal period was tied to local power relations, and became the arena of conflict of different social groups. Similar to findings in studies of Ottoman territories, women seem to have possessed a high degree of legal consciousness and were frequent actors in the courts. As Hasan has argued, several marginal groups resisted the powerful through "an insurrectionary awareness and acquisition of the dominant legal discursive system" (Hasan, 2004, p. 75).

It is important to note that even under Mughal rule, which saw a higher degree of centralization than earlier states, the state legal system barely permeated rural areas. The courts, the *qazi* (Islamic judge), the *faujdar*, (constable), and other legal institutions were primarily urban. Disputes and law enforcement in rural areas continue over the centuries to be dealt with by bodies called *panchayats* or local assemblies of notables. These could exercise jurisdiction over a territory like a village, a caste group, or a kin group.

While most litigants appearing before the courts represented themselves, by the seventeenth century there is the emergence of a class of men called *vakils*, who resembled modern lawyers. However, a vakil was usually invested with the authority to act for another but not necessarily plead for him before a court of law. European travelers commented both on the absence of lawyers from courtrooms and on the relative simplicity and efficacy of state courts. Under Emperor Aurangzeb's reign, the state began appointing vakils to look after various interests including to act as legal advisors to the poor (Calkins, 1969).

The Company Period (1601–1833)

In 1601, Queen Elizabeth I granted a charter to the English East India Company (EEIC) giving them the monopoly over trade with India. The EEIC successfully negotiated permission from Indian rulers to set up factories and port towns in Surat, Bombay, Calcutta, and Madras. The royal charters empowered the EEIC officials to administer justice in their territories. The charters required that all Company laws had to be 'reasonable' and not contrary to the laws and customs of the English realm. The charters had only granted the Company the jurisdiction over English subjects, but it soon found itself dealing with Indians who pursued claims in the Company courts as well. The Charter of 1726 established an 'Anglo-centric' legal order in Madras, Bombay, and Calcutta, which brought the Indian territories firmly into the British legal world. It explicitly provided that the law of the Company settlements in India was to be that of England, to the extent that circumstances would permit. It set up Mayor's courts headed by aldermen who were tenured for life, guaranteeing a degree of independence, and explicitly provided for appeals to the Privy Council. As recent scholarship has shown, judges and juries in these Company towns frequently protested deviations from English law and the arbitrary authority of the Company officials (Fraas, 2014).

However, both the nature of colonial rule and the legal system were transformed following the victories of the EEIC in the battles of Plassey and Buxar. From being an armed trading power, the Company found itself sovereign over territory greater than the size of Germany. Over the next 50 years, the Company state made three interventions that transformed South Asian legal cultures completely.

The colonial state, with its investment in creating a rule of law, displaced local authorities through new civil courts and formal procedures. Litigation became central to the creation of the colonial state. British legal procedure was based on the normative idea of equality, understood relationships in contractual terms, sought to arrive at a specific decision and saw a dispute as a discrete problem (Cohn, 1987). Indian society was not premised on equality or contract between individuals, but sought to locate them within a network of social relationships. Indigenous forms of dispute resolution instead postponed clear-cut decisions and aimed at compromise. Cohn argued that this made courts the site of a direct clash of the values between British and Indian society. Indians used the courts as a form of speculation or gambling. Given the cost and time involved, litigation was often a threat, rather than a method of resolving disputes. With the demilitarization of society, litigation became the new battleground and a form of public entertainment (Cohn, 1987, p. 573). This view was shared by colonial administrators who sought to police the 'lying, litigious native' by increasing court fees, punishing perjury, and changing evidentiary practices (Raman, 2012).

Secondly, despite the promise of the rule of law, it was marked by its absence in many aspects of colonial government. Scholars have argued that the rule of law was bound to fail in a colonial context, since colonial difference marked by race determined the relationship between the state and the individual. Legislation that sought to allow Indian judges to try Europeans was faced with widespread protests and had to be withdrawn. Executive authority was shielded from judicial oversight and the powers of the appellate courts were trimmed over successive years. Attempts to rationalize the administration and make it less arbitrary would constantly run up against the question of race and the relationship between the colonizer and the colonized. Indeed, the logic of the colonial state can be demonstrated by the way in which the provision for exercise of arbitrary power was incorporated into legal codes and normalized (Kolsky, 2005).

Finally, the colonial state created the domain of personal law. In an attempt to maintain the veneer of legitimacy granted by the Mughal state, the British administration had continued

to apply the law of Hindus and Muslims to matters of inheritance, marriage, and religious institutions (Kugle, 2001). Under the 1772 Regulations, Governor-General Warren Hastings stated that "inheritance and succession to lands, rents, and goods, and all matters of contract and dealing between parties, shall be determined in the case of Mahomedans by the laws of the Mahomedans, and in the case of Gentoos (Hindus) the laws and usages of the Gentoos" (Kugle, 2001). Hastings believed he was continuing the practice of the Mughals, who had left each non-Muslim community the right to administer its own law to its own members as long as they met certain conditions. However, the analogy was false since the Company state itself administered the law of each religious group. This magnanimity toward preserving indigenous law was selective. Initially, English judges were assisted by Hindu pundits and Muslim qazis in administering personal laws. However, the distrust of native judges led to their gradual removal from courts and the dependence of English judges on translations and volumes of precedence. The conversion of a relatively fluid localized system of Islamic law to a reified and static entity through "translations, codification, and adjudication" has been documented extensively by scholars (Anderson, 1993; Kugle, 2001). The British saw the sharia as analogous to canon law in Europe and attempted to discover an authentic code in an authoritative text. Thus, British courts narrowed down the number of precepts accepted as Muslim law while simultaneously giving finality to sharia, which it could not attain when authority was localized and distributed amongst the many jurisdictions and individuals. Similar interventions by the colonial government created a body of law often described as Anglo-Hindu law.

Buddhist and Jain legal traditions remain less well known in mainland South Asia, partly because of their underrepresentation in the colonial legal system. Scholars speculate that the British believed that the Jainas had no legal texts because the Jainas deliberated concealed their scriptures from British law to prevent the texts from being defiled. With the annexation of Burma a corpus of Buddhist legal texts were translated from palm leave manuscripts, but these largely dealt with rules for Buddhist monastic order and not for laypeople. What came to be defined as Buddhist law in Burma, consisted largely of customary practices of Burmese people who happened to be Buddhist. The Zoroastrian community or Parsis, became spectacularly successful legal entrepreneurs in harnessing the colonial legislature and judiciary to develop a Parsi law in the late nineteenth and early twentieth century. Recent scholarship suggests that Parsis enjoyed greater legal consciousness and were overrepresented both as litigants and as judges and lawyers within the colonial legal system (Sharafi, 2014).

Similar processes were observed when the colonial state sought to apply customary law, particularly areas inhabited by the thousands of tribal communities. However, by the process of recording, translating, and adjudicating custom for the colonial law courts, the nature of customary practices was transformed. Tribal groups remained ambivalent about the colonial readings of their customs, and on several occasions, asserted their customary rights over land and forests through demonstrations or open rebellion (Bhattacharya, 1996). In Afghanistan and the North Western province of Pakistan, informal institutions such as *shuras* (local councils of elders or religious scholars) and *jirgas* (councils of elders of the Pashtun tribes) continued to coexist alongside both colonial and postcolonial legal systems. In resolving disputes, they drew upon developed conceptions of morality and justice such as the Pashtunwali (the Pathan Code of honor). While overwhelmingly Muslim, these bodies frequently deviated from sharia practices on questions like inheritance rights for women.

These developments were mirrored in Sri Lanka, which had first come under European control by the Portuguese and then by the Dutch. On gaining control of the coasts, the Dutch introduced Roman-Dutch law. Dutch courts, like English ones in India, also administered customary and religious laws of the local Sinhalese and Tamils. The Dutch officials sought to codify these bodies of law and produced the *Thesawalamai*, or a treatise on the laws and customs of the Tamils of North Sri Lanka, in 1794. With the British conquest of Sri Lanka, the existing framework of laws was continued and Roman-Dutch law coexisted with English common law and customary law. The British conquered Kandy, the last autonomous state in Sri Lanka, and applied Kandyian law to the people from the region (Cooray, 1974). Like the Indian system of personal laws, the Sri Lankan laws were also applicable to people based on their origins. So a person who is originally from Kandy will continue to be governed by Kandyian law even if she resided elsewhere.

As the above processes demonstrate, the colonial state had transformed localized legal systems into overarching, textual bodies of law through a process of translation, codification, and adjudication. However, forum shopping continued to be a way through which litigants continued to maneuver the system. Groups resisted categorization into religious communities by claiming customary practices that differed; for instance, the Khoja Muslims of Bombay were successfully able to claim that they were governed by Hindu family law. In the absence of divorce provisions, women would convert to other religions in order to dissolve their marriages. Businesses and individuals would take advantage of the semiautonomous princely states that were not directly governed by the British to escape onerous legislation (Sharafi, 2010; De, 2010).

Portuguese-controlled territories were an exception to this religious patchwork and governed all subjects through the Portuguese Civil Code of 1867. In independent India, former Portuguese territories like Goa continue to have a uniform civil code in force. In these maneuvers the litigants were helped by the rapid growth of the modern legal profession in India. Indians had staffed the lower level legal positions from the earliest days of the Company rules; by the end of the nineteenth century they came to dominate the bar and rise to the bench. The explosion in the number of lawyers led to complaints about brieflessness and unemployment. As figures familiar with the colonial state, and one of the few groups not dependent on the state for their income, lawyers also began to dominate the Indian national movement. India today with over 1.2 million lawyers has one of the largest pools of legal professionals after the United States and in contrast to other peers like China and South Africa. However, on a per capita or per GDP basis, India ranks a lot lower. The legal profession in India, as in its neighboring countries, remains highly segregated with a small tier of highly paid lawyers and a large mass of struggling low-income professionals (Ganz, 2011).

High Colonial State and the Codification Projects

While religious laws continued to govern areas like marriage and inheritance, Islamic civil, criminal, and commercial laws were rapidly replaced by British legislation, given that law and order and commerce were both critical aims of the state. However, it was in the mid nineteenth century that utilitarian-minded bureaucrats were able to start a mammoth project of codifying civil and criminal law. Successive law commissions headed by Lord Macaulay, James Fitzjames Stephens ,and Henry Maine, enacted the Indian Penal Code, the Codes of Civil and Criminal Procedure, and the Evidence Act. The codification of procedural laws could encompass the utilitarian desires of unity, precision, and simplicity without intervening with the substantive content of the law (Metcalf, 1997). As utilitarian attempts to codify the law in the United Kingdom had been fiercely resisted in parliament, therefore South Asia came to serve as a laboratory. These codes that were originally written for India were exported to other British colonies in Asia and Africa. Entrepreneurial Indian lawyers who had gained experience with these legislations would set up practice in colonies like Fiji, Malaysia, or Kenya. Many of these colonies also had growing populations of South Asian immigrants, either in the form of merchants or as coolie labor. M.K Gandhi was one of the most prominent of these diasporic lawyers. Thus, South Asian legal culture, including the system of personal laws, was also exported through imperial networks.

Autonomous states like Nepal and the semiautonomous princely kingdoms in India often pioneered new legal reforms. The state of Baroda introduced a divorce law for Hindus in the late nineteenth century, Bhopal granted more rights to Muslim women and the state of Mysore was the first in the world to introduce affirmative action for lower castes in employment, education, and legislatures. Processes of centralization were more prevalent in these 'semiautonomous' territories, for instance the Nepali state through the nineteenth century broadened its powers to intervene in questions related to caste and jati rules, a matter that in British-controlled territories were left to caste groups themselves.

Postcolonial Legal Orders

India and Pakistan became independent in 1947, Sri Lanka in 1948. At various dates, each state adopted a republican constitution, including a judicially enforceable bill rights. Unlike the colonial state, the new states explicitly granted courts the power of judicial review and the ability to strike down legislation and administrative orders as unconstitutional. Unsurprisingly, in both democratic India and increasingly authoritarian Pakistan, tensions arose between the judiciary and the executive over the ability of the state to amend fundamental rights or the legality of marital rule (Austin, 2003; Newburg, 2004). While creating a powerful set of constitutional remedies, the new constitutional orders retained several techniques of colonial governance. For instance, the colonial institutions of the police, the army, and the bureaucracy remained structurally largely unchanged. The constitution recognized provisions where the executive could declare emergency and suspend fundamental rights. Even the fundamental rights that were granted were subject to certain textually inscribed limitations, for instance, freedom of speech was subject to limitations on grounds like public order, decency and morality, incitement to offense, sovereignty and integrity of the state, etc. Unlike the U.S. constitution which sought to protect its citizens from the government, South Asian (particularly the Indian) constitutions sought to empower the state. This is partly because the authors of the constitution envisioned it as a document that would bring about social transformation and economic redistribution.

Finally, the constitutions for the first time in South Asian history declared all citizens to be equal before law. The most drastic change from the precolonial system was the displacement of a normative inequality as the ground for conceptions of rights in India. Rights under colonial government were determined by caste, class, and gender and drew upon 'custom' (Guha, 2000). The Constitution however heralded a formal equality between all citizens, and created a common source of rights for all. The Indian constitution specifically abolished untouchability, and also provided for a wide range of affirmative action programs for certain castes and tribes, as well as women. The states of Pakistan and Bangladesh would also come to recognize the principle of substantive equality and provide for affirmative action for women in legislatures.

All South Asian states have tried to reconcile the existence of personal laws (which are often highly discriminatory) with constitutional equality. It has been suggested that the codification and control over the interpretation of the sharia mirrored the process of vesting political power away from Muslims. Not surprisingly, a significant component in the growth of Muslim nationalism was the attempt to regain control over this system. The early legislatures had attempted to reform family law in the 1930s, often to address questions of gender justice. In 1939, the Dissolution of Muslim Marriages Act provided Muslim women limited grounds for divorce (De, 2009). After independence, the Indian government in face of much opposition from Hindu conservatives was able to push through the Hindu Code bill, which replaced Hindu family law with statutory law. While codifying elements of Hindu law, the code provided for divorce and gave certain inheritance rights to women. Similarly, in Pakistan, the Family Law Ordinance of 1961 attempted to temper sharia law by making it institutionally difficult to practice polygamy, restricting the man's unilateral right to divorce, and improving the woman's rights to maintenance and control over her dower. Ironically, while each state was able to reform the law of its majority community, for reasons of legitimacy it left the laws governing minorities largely untouched.

In the 1970s, following the independence of Bangladesh, the Pakistani state came to rely more on Islam as a source of legitimacy. General Zia's government introduced the Hudood ordinances, which for the first time in two centuries introduced criminal law based on the shariat. The most controversial provisions dealt with rape, which required four adult male eyewitnesses to establish the charge. If a female victim admitted to sexual intercourse, but was unable to prove that it was rape, she was liable to be sentenced for the offense of *zina* or illicit sex. Despite vociferous opposition from women's groups and human rights organizations the ordinances were not finally repealed until 2006 (Lau, 2007).

The 1970s in India saw the growth of public interest litigation urged on by the Supreme Court. Public interest litigation was marked by the lowering of requirements of *locus standi* allowing various public-spirited persons to appear on behalf of those who could not reach court, the creation of innovative legal remedies including the establishment of monitoring commissions by the court and the *suo moto* taking up of issues by the court. The courts interpreted the constitutional right to life broadly to include the right to food, the right to education, the right to information, the right to health care, and the right to clean environment, creating the basis for new legislation. In many of these cases the courts worked closely with NGOs and social movements, transforming the relationship between law and politics. Public interest litigation doctrines have been readily adopted in Pakistan, Bangladesh, Sri Lanka, and Nepal with the courts bolstering their public legitimacy. Unsurprisingly, judges have emerged as the main site of opposition to governments and the dismissed Chief Justices have led opposition movements in Pakistan and Sri Lanka. More recently, public interest litigation has come under scrutiny for advancing a middle class agenda at the expense of the poor, for instance in ordering slums to be cleared in order to protect a "right to environment" (Narrain and Suresh, 2014).

Conclusion

Divided into several states with political forms ranging from constitutional monarchies, republican democracies, and authoritarian military dictatorships, South Asia legal systems diverged in many ways over the twentieth century. However, certain features remain in common. All systems are marked by widespread legal pluralism and competing legal authorities, including a state-sanctioned system of separate personal laws. With the exception of Nepal, the legal architecture of South Asia, from the court systems to procedural law was largely forged under British colonial rule. The legal codes authored by the colonial state continue to remain in force, with a few postcolonial additions. The Supreme Courts and the higher judiciary of all South Asian states have greatly expanded their powers and authority over the last two decades, emerging as important sites of opposition to government.

The vibrancy of the South Asian constitutional courts stands in stark contrast to the rest of the judicial apparatus. Lower courts remain poorly staffed and marked by corruption and delays. India today has the largest backlog of cases in the world with 30 million cases still pending. The delays are in part due to the colonial origins of procedural laws. Given that the lower judiciary was staffed by Indians whose judgment was never completely trusted by the colonial state, procedural laws retained provisions for appeals at every stage of the trial. As lawyers benefit from the delayed litigation, there has been strong resistance from the legal profession to any attempt to change the procedural laws.

This contrast between higher and lower levels of the legal system leads to a unique situation, where public discourse is marked with the language of law, justice, and constitutionalism while citizens remain cautious about approaching the legal system to settle private disputes.

See also: East Asian Legal Traditions; Globalization and Religion; Islam: Asia; Islamic Law; Latin-American Legal Systems; Legal Traditions of Southeast Asia; Middle Eastern and North African Legal Traditions; Religions of India; Sub-Saharan African Legal Traditions.

Bibliography

Anderson, Michael, 1993. Islamic law and colonial encounter in british India. In: Arnold, David, Robb, Peter (Eds.), Institutions and Ideologies: A SOAS South Asia Reader 165–185.

Austin, Granville, 2003. Working a Democratic Constitution: A History of the Indian Experience. Oxford University Press, Delhi.

Bhattacharya, Neeladri, 1996. Remaking custom: the discourse and practice of colonial codification. Tradition, Dissent and Ideology: Essays in Honour of Romila Thapar 20–51.

Calkins, Philipp B., November 1968–February 1969. A note on lawyers in muslim India. Law and Society Review 3 (2/3), 403–406. Special Issue Devoted to Lawyers in Developing Societies with Particular Reference to India.

Cohn, Bernard S., 1987. Some notes on law and change in North India. An Anthropologist Among Historians and Other Essays. Oxford University Press, Delhi, pp. 554–574.

Cooray, L.J.M., 1974. The reception of roman-dutch law in Sri Lanka. The Comparative and International Law Journal of Southern Africa VII, 295–320.

De, Rohit, 2009. Mumtaz Bibi's broken heart the many lives of the dissolution of Muslim marriages act. Indian Economic & Social History Review 46 (1), 105–130.

De, Rohit, 2010. The two husbands of Vera Tiscenko: apostasy, conversion, and divorce in late colonial India. Law and History Review 28 (04), 1011–1041.

Fraas, Mitch, Spring 2014. Making Claims: Indian litigants and the expansion of the English legal world in the eighteenth century. Journal of Colonialism and Colonial History 15 (1).

Ganz, Kian, November 25, 2011. "Demystifying India's Legal Market", The Mint at: http://www.livemint.com/Opinion/7wgxY0tcuLyw2yG2G0HcvL/Demystifying-India8217s-legal-market.html (as visited on 14.08.14.).

Guha, Sumit, 2000. Wrongs and rights in Maratha Country: antiquity, custom and power in eighteenth Century India. In: Anderson, M.R., Guha, S. (Eds.), Changing Concepts of Rights and Justice in South Asia. Oxford University Press, New Delhi, pp. 14–29.

Hasan, Farhat, 2004. State and Locality in Mughal. India: Power. Relations in Western. India, C. 1572–1730. Cambridge University Press, Cambridge.

Kane, P.V., 1968. History of the Dharmasastras: Ancient and Mediaeval Religions and Civil Law in India, vols. I–V. Bhandarkar Oriental Research Institute, Pune.

Kolsky, Elizabeth, 2005. Codification and the rule of colonial difference: criminal procedure in British India. Law and History Review 23 (03), 631–683.

Kugle, Scott, 2001. "Framed, blamed and renamed: the recasting of islamic jurisprudence in colonial south asia. Modern Asian Studies 35 (2), 257–313.

Lau, Martin, 2007. Twenty-five years of hudood ordinances-a review. Washington & Lee Law Rev 64, 1291.

Metcalf, Thomas, 1997. Ideologies of the Raj. Cambridge University Press, Cambridge.

Menski, Werner, 2003a. Comparative Law in a Global Context. Cambridge University Press, Cambridge.

Menski, Werner, 2003b. Hindu Law: Beyond Tradition and Modernity. Oxford University Press, New Delhi.

Narrain, Sidharth, Suresh, Mayur, 2014. The Shifting Scales of Justice: The Supreme Court in Neo-Liberal India. Orient Blackswan, Delhi.

Newberg, Paula R., 2002. Judging the State: Courts and Constitutional Politics in Pakistan, vol. 59. Cambridge University Press, New Delhi.

Newburg, P.R., 2004. Judging the State: Courts and Constitutional Politics in Pakistan. Cambridge University Press, Cambridge.

Raman, Bhavani, 2012. Document Raj: Writing and Scribes in Early Colonial South India. University of Chicago Press, Chicago.

Rocher, Ludo, 1968. "Lawyers" in classical hindu law. Law and Society Review 383–402.

Sathe, S.P., 2002. Judicial Activism in India. Oxford University Press, New York.

Sharafi, Mitra, 2014. Law and Identity in Colonial South Asia: Parsi Legal Culture, 1772–1947. Cambridge University Press.

Sharafi, Mitra, 2010. The marital patchwork of colonial south asia: forum shopping from Britain to Baroda. Law and History Review 28 (04), 979–1009.

Southeast Asia, Archaeology of

Miriam T Stark, University of Hawai'i-Mānoa, Honolulu, HI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E.A. Bacus, volume 22, pp. 14656–14661, © 2001, Elsevier Ltd.

Abstract

Southeast Asian archaeology runs the gamut from Pleistocene hominin sites to sixteenth-century capitals of the region's first nation-states. This article reviews the origins of modern humans, plant and animal domestication, metallurgy, and urbanism and state society. It discusses theoretical tensions, political concerns, and the future of Southeast Asian archaeology.

Introduction

One might justifiably argue that Southeast Asia as a region is an historical accident that first emerged out of the Supreme Allied Command in Asia in the Second World War. It was first defined as a geopolitical, economic, and sociocultural area, as the 'Southeast Asia Command' in 1943 and included then Ceylon (now Sri Lanka). Today we consider Southeast Asia as a coherent political region that includes the nation-states of the Republic of the Union of Myanmar (Burma), the Kingdom of Thailand, Lao PDR, the Kingdom of Cambodia, the Socialist Republic of Vietnam, Malaysia, the Republic of Singapore, the Republic of Indonesia, East Timor, the Nation of Brunei, the Abode of Peace, and the Republic of the Philippines. Some scholars also include Formosa/Taiwan (Republic of China) and southeast China as part of the region. The region extends *c*.5000 km from its northwest (Myanmar) to its southeastern most point (eastern Indonesia). All of the world's major religions, and hundreds of languages, are spoken in Southeast Asia today, which remains a cultural and linguistic mosaic.

Southeast Asia has a deep archaeological sequence that is broadly divided into the prehistoric and historic periods. Its prehistoric period extends back at least 1.8 million years into the Early Pleistocene; anatomically modern *Homo sapiens* have been reported from sites that date almost as early as 70 000 YBP in both mainland Southeast Asia (Laos; Demeter et al., 2012) and island Southeast Asia (Philippines; Mijares et al., 2010). Early Holocene hunter–gatherers exploited the region's extensive forests and riverine areas; farming was adopted by some, but not all, of the region's inhabitants by the third millennium BCE. Experimentation with copper and alloy technologies began at some point after the mid-second millennium BCE; mainland Southeast Asian archaeologists describe an 'Iron Age' that emerged by the mid-first millennium BCE. Southeast Asia's 'historic' period is associated with the series of polities or city-states across mainland Southeast Asia that shared Indic-influenced ideological systems and cycled in political prominence. While most scholars associate these polities with second millennium CE states like Angkor, Sukhothai, Pagan, and Majapahit, the region's earliest states arose in the mid-first millennium CE in both mainland and island Southeast Asia. Chinese visitors to the region at that time described kingdoms and empires in most places they visited. Islam entered the region with Arab traders in the twelfth and thirteenth centuries. European contact and colonialism in Southeast Asia held elsewhere as a critical hinge point in the archaeological record, marks only the most recent in an unbroken series of cultural encounters that Southeast Asians have experienced.

Almost all of Southeast Asia (except northern Vietnam) lies in the humid and monsoon tropics and experiences abundant rainfall. Its insular (island) region is characterized by geological instability, located as it is on the Ring of Fire. This has created a rich record of seismic, volcanic, and tsunami activity that doubtless left its stamp on human dynamics through time. Mountain ranges and long meandering rivers (notably the Irrawaddy, Chao Phraya, and Mekong) demarcate cultural regions in mainland Southeast Asia: mid-Holocene farmers flocked to the alluvial plains, while upland swidden farmers colonized the uplands of its mountain ranges (Higham, 2002: 83–89). These rivers offered throughways for communication, transportation, and trade; so did the extensive mangrove-fringed coastal zones that flanked the region's seascapes (Indian Ocean, South China Sea, and Java Sea; Spriggs, 2011).

Development of Southeast Asian Archaeology

Southeast Asian archaeology's relatively deep history begins in its region's rulers, who justified their genealogical right to rule through inscriptions as early as the mid-eleventh century. Interest in the Southeast Asian past continued through efforts by mid-nineteenth-century European colonial civil servants who reported ruins and archaeological artifacts. Archaeological practice, in western terms, accompanied European expansion across the region (Higham, 2002: 16–22). During Java's brief British rule (1811–16), Lieutenant Governor-General Thomas Stafford Raffles ordered the investigation and cleaning of the early-ninth-century Borobudur, the world's largest Buddhist monument, which is located in central Java. In the late 1850s, J.W. Earle explored prehistoric coastal shell middens in Wellesley Province, Malaysia. After the tragic death of naturalist Henri Mouhot in 1861, the 1864 publication of his travels up the Mekong and Tonle Sap rivers sparked a European fascination with the vanished Angkorian civilization.

By the late nineteenth century, military men, amateur naturalists, and geologists explored the region's nooks and crannies, and the earliest archaeological institutions were founded. A research officer in archaeology, under the Indian Archaeology Department, was appointed in 1881 in Burma; around the same time, the Dutch East Indies government

appointed archaeologist Willem P. Groenveldt to launch the first systematic investigations of Borobudur. By the turn of the century, the French had established École Française d'Extrême Orient to study monuments across French Indochina; Thai royalty established the Siam Society in 1904 to study its past. Shortly after Emilio Aguinaldo declared the independence of Philippines from Spain in 1898, Americans embarked on their colonial journey and also began exploring the archaeological past and anthropological present of the Philippines.

Late-nineteenth- and early-twentieth-century Southeast Asian archaeology was largely a colonial enterprise; all of Southeast Asia but the Kingdom of Siam was at that time under European colonial rule. Museums and archaeological societies were established by the 1920s across the region, and Southeast Asia' professional archaeologists – including Pieter Vincent van Stein Callenfels, Madeleine Colani, Robert van Heine Geldern, Ivor H.N. Evans, and Otley Beyer – launched major research projects during this time. Geologically oriented scientists explored Southeast Asia's Paleolithic and Neolithic pasts from northern Vietnam to western Indonesia, while art historians, architects, and epigraphers concentrated on its second-millennium CE classical civilizations. Most work was undertaken by foreigners; few non-Anglophone colonials trained local descendants of these ancient cultures in archaeological techniques. The British, in contrast, established an Archaeological Survey of Burma and trained its staff. Americans in the Philippines, under the direction of H. Otley Beyer, established an archaeology training program in the country's first anthropology department at the National University of the Philippines and encouraged the construction of the National Museum. The Far Eastern Prehistory Association formed in Batavia (Jakarta), Java, in 1929 as an offshoot of the Pacific Science Congress (Groslier, 1957); its establishment ushered in the modern era of archaeology in Southeast Asia.

World War II brought Southeast Asian archaeology to a halt for several years, and also transformed Southeast Asia's geopolitics profoundly. One impact of Japan's failed efforts to build its Greater East Asian Co-Prosperity Sphere during the war years was the region's gradual de-colonization. Indonesia declared independence in 1945 and Indonesians took over the Archaeological Service; it took 4 years of struggle before the country gained its liberty. The Philippines was granted independence in 1946, and Burma soon followed. By the mid-1960s, only East Timor remained a colony of Portugal, and later of Indonesia. Archaeological fieldwork accelerated across the region in the 1950s and 1960s. It was not, however, until the 1970s that truly Southeast Asian/western collaborative archaeological research began. Southeast Asian departments of archaeology were established, and so was the region's first archaeological journal, *Asian Perspectives* (editor: Wilhelm Solheim), in 1957. Southeast Asian archaeologists like U Aung Thaw (Archaeological Survey of Burma) began to publish monographs on their research and assume leadership roles in the field. Some foreign archaeologists (like Wilhelm Solheim and Chester Gorman) worked closely with Thai colleagues on their field-based research. Their work and that of many others put Southeast Asian archaeology on the world's stage. The establishment of SEAMEO ARCAFA (Southeast Asian Ministers of Education Organization Applied Research Centre for Archaeology and Fine Arts) in 1971 and subsequent launch of SPAFA (SEAMEO Project in Archaeology and Fine Arts) in Bangkok signaled the first significant entry of Southeast Asians into the archaeological research world.

More than a decade into the twenty-first century, Southeast Asian archaeology hardly resembles its predecessor from the pre-World War II years. Trained professional archaeologists work in every Southeast Asian country on prehistoric and historic period sites; most countries have heritage management ministries or bureaus that conserve (largely architectural, but some prehistoric) archaeological sites and manage visitation; every country offers undergraduate training in archaeology; and at least three countries, the Philippines, Vietnam, and Malaysia, offer PhD degrees in the field. Western archaeologists work closely with their in-country colleagues, and international, multidisciplinary teams are a regular feature on the archaeological landscape of island and mainland Southeast Asia. The following section outlines general findings from the Pleistocene through late Holocene periods, which covers biological developments like the emergence of modern humans and political developments like the emergence of states and empires.

From Pleistocene Hominins to Early States

Early Hominin Occupation of Southeast Asia

Eugene Dubois, a colonial Dutch Army medical officer, first uncovered evidence for Southeast Asia's Pleistocene hominin record when his prisoner field crew uncovered remains of what became known as *Pithecanthropus erectus* (now known as *Homo erectus*) along the banks of the Solo River, Java, in 1891. His European colleagues received this news coolly during Dubois' postdiscovery tour to report this 'the missing link' between apes and humans. It took several decades for the scientific community to accept that the Javanese fossils belonged to a human ancestor, and we now believe the 'Trinil 2' cranial fragment dates before 800 000 YBP. One of many discoveries reported from the region, most scholars now agree that *H. erectus* populations roamed the region by the Middle Pleistocene, and some assert an initial occupation beginning by *c.*1.8 million YBP.

Southeast Asian evidence for Middle Pleistocene hominins and their behavior concentrates in Java, although scholars have reported material from both island and mainland of Southeast Asia. Predominant tool types in associated lithic assemblages (pebble and flake tools) contrast markedly with the Acheulian hand axe industries that characterize African, European, and now South Asian *H. erectus* populations. Whether this difference in patterning reflects raw material availability remains debated, but what distinguishes Southeast Asia's record is unequivocal evidence for maritime voyaging capability that enabled Middle Pleistocene hominins to reach Flores Island (and cross a deep-water trench) by *c.*880 000 years YBP. Crossing this deep sea channel that was not connected to Sundaland (the biogeographical region of Southeast Asia that included current-day Borneo, Sumatra, Java, and mainland Southeast Asia) at any point during the Pleistocene required sophisticated cognitive abilities. Just how they reached Southeast Asia initially also remains an issue of some debate:

a coastal route (from South Asia) and an East Vietnam route, originating in southern China, are both possibilities.

Consensus now holds that anatomically modern humans dispersed out of Africa during the Late Pleistocene, after 125 000 YBP, and ultimately into Southeast Asia. Deciding how Southeast Asian hominins fit into competing scenarios of modern human origins (i.e., whether modern humans evolved somewhat independently in multiple locations where Middle Pleistocene hominins have been found (the regional model) or whether modern humans originated in Africa and dispersed outward in the (the replacement model)) requires us to examine the earliest evidence for modern humans in the region (*c*.70 000 YBP (Mijares et al., 2010)) and factors that may have constrained or encouraged migrations (*see* Demographic Impact of Disasters; Migrations, Colonizations, and Diasporas in Archaeology). Southeast Asia's well-dated super-eruption near Lake Toba in Sumatra, Indonesia *c*.74 000 YBP exerted global climatic effects; whether Toba's eruption created a genetic bottleneck (in which most extant human populations perished) remains a matter of debate. Current research on the time period in South Asia suggests that Paleolithic hominins survived these catastrophic fluctuations, but a current paucity of Southeast Asian sites that date between 90 and 70 000 YBP hampers our understanding of developments.

Disagreement continues over the taxonomic status of *Homo floresiensis*, a small hominin that produced stone tools and has been recovered from Liang Bua cave on the Indonesian island of Flores. Researchers agree that the fossils date to *c*.18 000 YBP, but debate continues over whether the *Homo* fossils represent a distinct hominin species (ostensibly descended from *H. erectus* or another earlier hominin) or instead a pathological pygmy population of modern *H. sapiens*.

Late Pleistocene/Early Holocene Hunter–Gatherers

Scholars have now reported dates from many Pleistocene sites beyond Java that predate 40 000 YBP, but the strongest signature for modern humans in Southeast Asia begins at this time. Cave sites have been documented in southern Thailand (Tham Long Rongrien), the Philippines (Tabon Cave), Borneo Island (Niah Cave), and eastern Indonesia (Leang Burung 2). Foragers who occupied these sites are associated with either flake or pebble/cobble tool assemblages; they practiced broad-spectrum subsistence that included terrestrial fauna (bovids, land tortoises, pigs, non-human primates, rodents), mollusks (freshwater fishes and shellfish), and sessile (immobile) coastal resources like mussels (*see* Hunting and Gathering Societies, Anthropology).

The number of sites dating after 15 000 YBP is now robust. Early Holocene foragers occupied maritime and riverine settings in lowland Southeast Asia. They left shell middens along waterways, worked upland areas for mammalian resources (e.g., deer, pigs, and monkeys; Higham, 2002: 46–49), and may have managed tree nut crops (canarium). We know the most about Late Pleistocene/Early Holocene populations associated with the Hoabinhian tool tradition represented by 'sumatraliths': stone tools, worked on one side, that are found across mainland Southeast Asia and as far south as Sumatra. Hoabinhian occupation has been documented in open-air and cave sites (Forestier et al., 2013) and include residential and mortuary sites (Zeitoun et al., 2013). Pebble tools, bone points, and antler wedges were also part of this industry; grinding stones have been identified in the latest Hoabinhian occupations.

Early to mid-Holocene foragers buried their dead with ocher, stone tools, and animal bones (Zeitoun et al., 2013). Substantial research has focused on determining whether the 'Hoabinhian' were the original settlers of mainland Southeast Asia and deciphering their relationship to extant indigenous (Negrito) communities across the region today. Some physical anthropologists associate well-dated Hoabinhian skeletal remains with 'Australo-Melanesian' populations; some Vietnamese and Thai sites contain 'Neolithic' phases superimposed atop Hoabhinian phases, and the Da But culture of northern Vietnam (Nguyen, 2005) contains evidence for intensive plant and animal use that could represent a transition to agriculture by indigenous inhabitants.

Defining the nature of Southeast Asia's earliest plant and animal domestication remains a subject of great scholarly interest. Archaeologists disagree over whether migrants entered the region with a suite of agricultural knowledge and displaced the extant foragers (whose descendants are today's Negrito populations), or whether Late Holocene foragers gradually intensified resource use until they became reliant on plant and animal domestication for subsistence (Bellwood, 2006). At issue is how this mid-Holocene plant and animal management system articulated with extant human groups in the region, and whether it coincided with a rapid population infusion of Austroasiatic speakers into the densely populated south China (*see* Language Families, Archeology and History of; Migrations, Colonizations, and Diasporas in Archaeology). If the latter process occurred, it may have been considerably more complex than is currently envisioned (Spriggs, 2011); it also may have generated 'friction zones' that held both intrusive farmers and indigenous foragers (e.g., Higham et al., 2011). Some may have intermarried, others colonized the region, and still others continued their southward journey into the Pacific. Archaeologists also argue over the nature and tempo of the process of plant and animal domestication across a region whose Neolithic cultivation regimens included both agriculture (rice, millet) and vegeculture (at least taro, and perhaps yams; *see* Horticultural Societies; Food Production, Origins of). Reasonable consensus exists, however, that Southeast Asia, writ large enough to include areas south of the Yangzi River, was the hearth/origin area for several key plant domesticates (most notably rice, yams, taro, and spices like nutmeg and mace) and animal domesticates (most notably dogs, pigs, chickens, and water buffalo).

Archaeologists also agree that some, if not all, mainland Southeast Asians practiced grain-based agriculture no later than 4000 YBP, and that their island Southeast neighbors embraced cultivation regimens a millennium after that (Amano et al., 2013). Vietnam's archaeological record for this mid- to late-Holocene transition is the most robust in the region and dates to *c*.4000 BP or earlier. The recently reported An Son site of southern Vietnam offers an excellent example of Neolithic Southeast Asia, with incised and impressed earthenware pottery, groundstone adzes and axes, fishhooks, beads, and a faunal assemblage that suggests riverine farming economy that was supplemented by dog and pig domestication (Bellwood et al., 2011).

Early Farming Communities

Archaeologists who routinely describe Southeast Asia's 'Neolithic' struggle to reconcile its material signatures with that found in the Near East or Europe (e.g., Bellwood, 2006). Southeast Asians added cultivation and earthenware ceramic technology to their foraging regimens between 4500 and 3500 years ago, but only slim evidence has been found for a monolithic shift to the kind of sedentary, village-based tribal organization that characterizes the Neolithic in other Old World regions. Limited archaeological evidence suggests that sprawling alluvial plains across the region's mainland and the insular zones' narrow river valleys housed growing cultivator populations. So did coastal and upland regions. All used polished or groundstone technology (particularly stone adzes and chisels, and groundstone bangles), cleared forests, and engaged in broad-spectrum subsistence regimens that included wild and cultivated resources (see review in Higham, 2002: 83–111).

Thailand's central and northeastern regions have yielded Neolithic sites that offer a glimpse of village-based farming life. Some archaeologists contend that craft specialization (like pottery production and/or shell adze production) and intercommunity exchange characterized these early communities like Khok Phanom Di (Bentley et al., 2007; *see* Craft Production, Anthropology of). Few well-dated Neolithic sites contain substantial skeletal and mortuary data. Southeast Asian archaeologists continue to devote more energy to tracing potential migratory routes of humans and their *Oryza sativa japonica* rice into the region than to exploring the nature of settlement or subsistence during this crucial period of time. Recovery of obsidian from shared sources (Reepmeyer et al., 2011), similar or identical artifact types, and a broad incised pottery stylistic horizon suggests that these 'Neolithic' communities engaged in large-scale interactional networks that stretched from Formosa into Sulawesi.

The Southeast Asian 'agricultural package' of rice, pottery, and sedentism arrived substantially later in parts of island Southeast Asia, and assumed different forms as the knowledge moved east toward Papua New Guinea (*see* Australia and New Guinea, Archaeology of). Borneo's 'Neolithic' peoples may have relied on sago rather than rice (Barton, 2012); in eastern Indonesia, sago was the carbohydrate staple. Direct evidence of plant domestication is scarce. Paleoenvironmental studies suggest that Sumatra's forests were impacted as early as 6400 YBP, and permanently cleared after 3000 YBP. Cave and rock shelter sites across island Southeast Asia offer evidence of pottery and rice *c*.4000 YBP, and burials associated with stone adzes, bone points, shell bangles, and incised earthenware indicate agricultural practices.

The Earliest Metal Age

Geopolitics and historical accident has skewed western knowledge of the Southeast Asian Bronze Age archaeological record toward Thailand, with recent research incursions into Myanmar and Cambodia. Mainland Southeast Asians engaged in copper and bronze metallurgical production and distribution by the mid to late second millennium BCE. Scholars continue to debate origin areas and transmission routes for bronze metallurgical knowledge (White and Hamilton, 2009). Northern Vietnam's Phung Nguyen culture (*c*.4000–3500 YBP) displays the earliest well-dated bronze metallurgy in the region. Within a few 100 years, metal workers in central and NE Thailand melted crushed copper ore into crucibles and cast this into stone and ceramic molds to create vessels that moved along riverine and coastal distributional networks.

Archaeologists have studied Mainland Southeast Asia's Metal Age (i.e., the 'Bronze Age' (sensu Higham, 1996) and the 'Iron Age') more intensively than any other prehistoric period, and Thai and Vietnamese sites have gotten the most attention. Most Metal Age sites are mortuary. They have yielded copper ingots, socketed ax heads, chisels, arrow and spear points, fishhooks, and sickles, and bioarchaeological approaches dominate the literature (*see* Bioarchaeology). The limited attention to regional-scale Metal Age settlement organization, using Thai survey and excavation data, suggests both the development of village-based specialization and subregional variability in production and distribution of metal objects (but see Eyre, 2011; *see* Intensification and Specialization, Archaeology of). Island Southeast Asia entered the Metal Age as much as 1500 years after mainland Southeast Asia. Sites proliferated across this southern region after the appearance of 'Iron Age' manifestations on the mainland.

Southeast Asia's Early Complex Polities and the 'Iron Age'

The earliest incontrovertible trappings of 'social complexity' (i.e., institutionalized stratification, intercommunity competition and conflict, a prestige goods economy; *see* States and Civilizations, Archaeology of) appeared all over Southeast Asia *c*.2500 YBP, which in the mainland is associated with the 'Iron Age.' It is during this time that the first localized metallurgical traditions, elaborate mortuary rituals, settlement hierarchies, and widespread stylistic horizons reflecting far-flung elite exchange networks became well established. Out of densely populated northern Vietnam, and arguably SW China (Yunnan and Guizhou Provinces today), emanated a bronze drum tradition whose use in collective ceremonies (still seen in some ethnohistorical upland groups) bore the signatures of complex societies. Bronze drum iconography – particularly from the Lake Dian drums of SW China – reflect shared ideological traditions, collective action, and possibly conflict (Higham, 1996). Vietnamese archaeologists have explored Iron Age sites across the Bac Bo region rather thoroughly, documenting settlement and mortuary sites. Geochemical studies suggest that nephrite (jade) artifacts were manufactured and traded from Formosa throughout central Vietnam and the central Philippines (Hung et al., 2007). Dong son bronze drums from northern Vietnam have been found as far south as Java and across much of mainland Southeast Asia.

Most research on Southeast Asia's Iron Age concentrates on human remains from cemeteries that contain, to varying degrees, markers of interpersonal violence and are associated with abundant bronze and iron weapons (e.g., Domett et al., 2011); these sites signal the rise of competitive polities. Some Southeast Asian populations lived along the southern fringes of the Han imperial world: these 'southern barbarians' resisted conquest everywhere but northern Vietnam. Most participated

in, and many were affected by, maritime silk trade that sliced through the region to link China with South Asia (Bellina and Glover, 2004). Island Southeast Asia's Metal Age was also, temporally speaking, the Iron Age since Island Southeast Asia lacked a preceding Bronze Age. Mainland Southeast Asia's endpoint of the Iron Age lacks a clear temporal boundary, although most archaeologists concur that the period between 2000 and 1500 BP marked the transition to history, and the appearance of the region's first indigenous writing systems, influenced largely by South Asian predecessor scripts (*see* Historical Archaeology).

Early States

The earliest archaeologically visible states arose during that transition to history across the river valleys and along the coasts of Mainland Southeast Asia from northern and central Vietnam (Allard, 1998; Southworth, 2004) to Thailand and Malaysia (Allen, 1998); they also emerged along Sumatra's coasts. Described and assigned toponyms by visiting Chinese emissaries, these city-states (Manguin, 2004) and coastal chiefdoms (Junker, 2004; *see* Chiefdoms, Archaeology of) formed a complex network of deep-water ports and entrepôts for goods that moved through the South China Sea toward both South Asia and the Mediterranean, and in the reverse direction to China. The region's earliest states varied in size and geographic extent. Most bore the influence of Buddhist and Hindu ideological systems that entered the region with trade and that sculpted ritual landscapes of Indic-influenced shrines, temples, and nucleated communities (Stark, 2006; *see* States and Civilizations, Archaeology of).

Southeast Asians' involvement in this dynamic world-system brought them wealth and non-local goods that facilitated the rise of landed elites and the legitimation of centralized authority, virtually all within a new ideological system from South Asia with roots in Brahmanic belief systems. Intensified interregional interaction and movement across the Bay of Bengal also generated an increasingly complex regional organization. Large centers, ruled by 'Hinduized' elite, rose and fell through time within large regions. The Pyu 'city-states' arose along and beyond the Irrawaddy River through Myanmar's central Dry Zone to the west; Dvaravati settlements flanked the receding Bay of Bangkok in central Thailand, and snaked to the northeast along the Chao Phraya river and its larger tributaries. Hinduized pre-Angkorian polities populated the lower Mekong region, expanding out from the preceding period through the construction of ashrams, shrines, temples, and large centers. The pre-Angkorian epigraphic record offers rich insights on the states of the late first millennium CE: their elites garnered power through ritual sponsorship, channeled resources through their community temple foundations, and participated with their followers in ritual practice that circumscribed their agricultural cycles and shaped their political lives (Higham, 2002: 298–335; Stark, 2004).

Most of these late first millennium polities have clear archaeological antecedents that date to the period from 2500 to 1500 YBP (Stark, 2006). Some Cham sites in central Vietnam sit directly atop Iron Age Sa Huynh archaeological sites; pre-Angkorian sites have underlying 'Funan' deposits from the first through fifth centuries CE, and Myanmar's Pagan has Pyu finger-marked bricks from the second century BCE to the eighth century CE (Hudson et al., 2001). Today's Southeast Asian nation-states trace their roots back to the classical states associated with the ninth through fifteenth centuries CE: Bagan (in Myanmar), Sukhothai and Ayutthaya (in Thailand), Angkor (in Cambodia), and Champa and Dai Viet (in Vietnam). Archaeologists can now assert, with some certainty, that these classical states have roots that extend backward in time to the region's first states. Until rather recently, art historians and historians have dominated research agendas on the ninth through fifteenth century states, whose monuments are major tourist destinations that are managed by cultural heritage agencies. The last three decades has seen a surge of archaeological interest in these states. For historical and logistical reasons, more field investigations have concentrated at Cambodia's Greater Angkor region than any other classical capital in the region (e.g., Evans et al., 2007). Increased attention to other late capitals offers rich comparative data for archaeologists who study the collapse and regeneration of premodern states.

Theoretical Tensions and Political Concerns in Southeast Asian Archaeology

Theoretical tensions shape the research directions of Southeast Asian archaeology and affect its practitioners. Artificial divides within and beyond archaeology hinder forward progress: 'prehistoric' and 'historic' archaeologists rarely communicate, and the latter struggle to reconcile art historical and documentary evidence with archaeological patterns. Some archaeologists are theoretical purists, using approaches from western academe (*see* Archaeology, Theory in) while others practice some form of culture history. Most scholars bifurcate into 'endogenous' and 'exogenous' camps to explain most origins: of modern humans, of food production, of metallurgy, of urbanism, and of states. Staunch proponents of each extreme ignore empirical findings and bolster their claims; those in the middle struggle to reconcile archaeological patterning with preconceived models developed elsewhere, for other archaeological records.

Archaeologists apply increasingly sophisticated methods to Southeast Asia's archaeological record, but it continues to resist categorization. Modern humans entered the region, but coexisted with small hominins (or hobbits) until the late Pleistocene. Populations adopted rice from Southeast China, but did not shift to a reliance on agriculture until many centuries later, in the Bronze Age (King et al., 2014). Urbanism and state-level society emerged in concert with 'Hindicization' in mainland Southeast Asia, but most of Island Southeast Asia remained free of cities despite their embrace of Indic ideas. Rulers called themselves *cakravartin* (universal ruler (Sanskrit)), yet ruled at the pleasure of their followers as their ephemeral polities rose and collapsed in regular succession. Respecting Southeast Asia's archaeological record for its own sake, rather than forcing it into extant models, takes great effort.

Southeast Asian archaeologists increasingly acknowledge the effect of political concerns on archaeological practice. Until recently, western-born (or at least western-trained) archaeologists set the agenda. As growing numbers of Southeast Asian nationals become archaeologists, discussions grow over

whether to follow western intellectual paradigms or their indigenous archaeological frameworks. Some countries, like Malaysia, not only train their own archaeologists, they restrict foreigner access to their archaeological resources (*see* Archaeology, Politics of). Others (Thailand, Philippines, Cambodia, and, increasingly, Vietnam) welcome international collaboration that includes capacity building for the next generation. Achieving parity in the Southeast Asian archaeological community is challenging, since most trained Southeast Asian archaeologists work in heritage management rather than in academia (*see* Cultural Resource Management: Conservation of Cultural Heritage).

Future Research Directions for Southeast Asian Archaeology

More Southeast Asian archaeologists today practice their trade than at any time in the region's very deep history. They study vastly different time periods, embrace radically different epistemologies, investigate myriad topics, use cutting-edge technologies, and report their work in more than 10 languages. Southeast Asia's archaeologists are cataloging the cultural and ecological mosaic as rapidly as possible. Archaeologists remain one step ahead of economic transformations that threaten the region's precious archaeological record. Future directions for Southeast Asia's archaeologists involve every period under study, and every ecological zone in the region. Southeast Asia contains karst limestone caves with Pleistocene deposits and speleothems that warrant attention before they are destroyed for construction materials. Southeast Asia's shell middens and mid-Holocene sites must be studied before rising sea levels eradicate them. Work is needed on settlement organization and habitation patterns to complement the rich and informative mortuary database available on the Neolithic, Bronze, and Iron Ages. Researchers could maximize the information potential of extant museum collections and by studying geochemical patterning in trade wares that circumscribed some states like Angkor, and that linked others through maritime trade, like Sukhothai. The region's sixteenth through nineteenth century past, which holds insights on the Early Modern period, is undergoing urbanization. The world grows closer each decade, and the future holds great potential for international collaboration in Southeast Asian archaeology.

See also: Archaeology, Politics of; Archaeology, Theory in; Australia and New Guinea, Archaeology of; Bioarchaeology; Chiefdoms, Archaeology of; Cognitive Archaeology; Craft Production, Anthropology of; Cultural Resource Management: Conservation of Cultural Heritage; Demographic Impact of Disasters; Food Production, Origins of; Historical Archaeology; Horticultural Societies; Hunting and Gathering Societies: Anthropology; Intensification and Specialization, Archaeology of; Language Families, Archaeology and History of; Migration History; States and Civilizations, Archaeology of.

Bibliography

Allard, F., 1998. Stirrings at the periphery: history, archaeology, and the study of the Dian. International Journal of Historical Archaeology 2, 321–341.

Allen, J., 1998. History, archaeology, and the question of foreign control in early historic-period peninsular Malaysia. International Journal of Historical Archaeology 2, 261–289.

Amano, N., Piper, P.J., Hung, Hsiao-chun, Bellwood, P., 2013. Introduced domesticated animals in the neolithic and metal age of the Philippines: evidence from Nagsabaran, Northern Luzon. Journal of Island & Coastal Archaeology 8, 317–335.

Barton, H., 2012. The reversed fortunes of sago and rice, *Oryza sativa*, in the rainforests of Sarawak, Borneo. Quaternary International 249, 96–104.

Bellina, B., Glover, I.C., 2004. The archaeology of early contact with India and the Mediterranean world, from the fourth century BC to the fourth century AD. In: Glover, I., Bellwood, P. (Eds.), Southeast Asia: From Prehistory to History. RoutledgeCurzon, New York, pp. 282–313.

Bellwood, P., 2006. Asian farming diasporas? Agriculture, languages, and genes in China and Southeast Asia. In: Stark, M.T. (Ed.), Archaeology of Asia. Blackwell Publishing, Oxford, pp. 97–118.

Bellwood, P., Oxenham, M., Hoang, Bui Chi, Dzung, Nguyen Kim, Willis, A., Sarjeant, C., Piper, P., Matsumura, H., Tanaka, K., Beavan-Athvield, N., Higham, T., Nguyen, Quoc Manh, Kinh, Dang Ngoc, Kien, Nguyen Khanh Trung, Huong, Vo Thanh, Bich, Van Ngoc, Quy, Tran Thi Kim, Thao, Nguyen Phuong, Campos, F., Sato, Y.I., Cuong, Nguyen Lan, Amano, N., 2011. An Son and the neolithic of Southern Vietnam. Asian Perspectives 50, 144–175.

Bentley, R.A., Tayles, N., Higham, C., Macpherson, C., Atkinson, T.C., 2007. Shifting gender relations at Khok Phanom Di, Thailand. Current Anthropology 48, 301–314.

Demeter, F., Shackelford, L.L., Bacon, A.-M., Duringer, P., Westaway, K., Sayavongkhamdy, T., Braga, J., Sichanthongtip, P., Khamdalavong, P., Ponche, J.-L., Wang, H., Lundstrom, C., Patole-Edouma, E., Karpoff, A.-M., 2012. Anatomically modern human in Southeast Asia (Laos) by 46 ka. Proceedings of the National Academy of Sciences 109, 14375–14380.

Domett, K.M., O'Reilly, D.J.W., Buckley, H.R., 2011. Bioarchaeological evidence for conflict in Iron Age northwest Cambodia. Antiquity 85, 441–458.

Evans, D., Pottier, C., Fletcher, R.J., Hensley, S., Tapley, I., Milne, A., Barbetti, M., 2007. A comprehensive archaeological map of the world's largest preindustrial settlement complex at Angkor, Cambodia. Proceedings of the National Academy of Sciences 104, 14277–14282.

Eyre, C.O., 2011. Social variation and dynamics in metal age and protohistoric central Thailand: a regional perspective. Asian Perspectives 49, 43–84.

Forestier, H., Zeitoun, V., Winayalai, C., Métais, C., 2013. The open-air site of Huai Hin (Northwestern Thailand): chronological perspectives for the Hoabinhian. Comptes Rendus Palevol 12, 45–55.

Groslier, B.P., 1957. The Far-Eastern Prehistory Association 1953–1957. Asian Perspectives 1, v–ix.

Higham, C.F.W., 1996. The Bronze Age of Southeast Asia. Cambridge University Press, Cambridge.

Higham, C.F.W., 2002. Early Cultures of Mainland Southeast Asia. River Books, Bangkok.

Higham, C.F.W., Guangmao, Xie, Qiang, Lin, 2011. The prehistory of a friction zone: first farmers and hunter-gatherers in Southeast Asia. Antiquity 85, 529–543.

Hudson, B., Lwein, Nyein, Maung (Tanpawady), Win, 2001. The origins of Bagan: new dates and old inhabitants. Asian Perspectives 40, 48–74.

Hung, Hsiao-chun, Iizuka, Y., Bellwood, P., Dung Nguyen, Kim, Bellina, B., Silapanth, P., Dizon, E., Santiago, R., Datan, I., Manton, J.H., 2007. Ancient jades map 3,000 years of prehistoric exchange in Southeast Asia. Proceedings of the National Academy of Sciences 104, 19745–19750.

Junker, L.L., 2004. Political Economy in the Historic Period Chiefdoms and States of Southeast Asia. In: Feinman, G.M., Nicholas, L.M. (Eds.), Archaeological Perspectives on Political Economies. University of Utah Press, Salt Lake City, pp. 223–251.

King, C.L., Bentley, R.A., Higham, C., Tayles, N., Fioarsdottir, U.S., Layton, R., Macpherson, C.G., Nowell, G., 2014. Economic change after the agricultural revolution in Southeast Asia? Antiquity 88, 112–125.

Manguin, P.-Y., 2004. The archaeology of early maritime polities of Southeast Asia. In: Glover, I., Bellwood, P. (Eds.), Southeast Asia: From Prehistory to History. RoutledgeCurzon, New York, pp. 282–313.

Mijares, A.S., Détroit, F., Piper, P., Grün, R., Bellwood, P., Aubert, M., Champion, M., Cuevas, N., De Leon, A., Dizon, E., 2010. New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. Journal of Human Evolution 59, 123–132.

Mudar, K.M., Anderson, D., 2007. New evidence for Southeast Asian Pleistocene foraging economies: faunal remains from the early levels of Lang Rongrien Rockshelter, Krabi, Thailand. Asian Perspectives 46, 298–334.

Nguyen, V., 2005. The Da But culture: evidence for cultural development in Vietnam during the middle Holocene. Bulletin of the Indo-Pacific Prehistory Association 25, 89–93.

Reepmeyer, C., Spriggs, M., Anggraeni, Lape, P., Neri, L., Ronquillo, W.P., Simanjuntak, T., Summerhayes, G., Tanudrijo, D., Tiauzon, A., 2011. Obsidian sources and distribution systems in Island Southeast Asia: new results and implications from geochemical research using LA-ICPMS. Journal of Archaeological Science 38, 2995–3005.

Shoocongdej, R., 2007. The impact of colonialism and nationalism in the archaeology of Thailand. In: Kohl, P.L., Kozelsky, M., Ben-Yehuda, N. (Eds.), Selective Remembrances: Archaeology in the Construction, Commemoration, and Consecration of National Pasts. University of Chicago Press, Chicago, pp. 379–399.

Southworth, W.A., 2004. The coastal states of Champa. In: Glover, I., Bellwood, P. (Eds.), Southeast Asia: From Prehistory to History. RoutledgeCurzon, New York, pp. 209–233.

Spriggs, M., 2011. Archaeology and the Austronesian expansion: where are we now? Antiquity 85, 510–528.

Stark, M.T., 2004. Pre-Angkorian and Angkorian Cambodia. In: Glover, I., Bellwood, P. (Eds.), Southeast Asia: From Prehistory to History. RoutledgeCurzon, New York, pp. 89–119.

Stark, M.T., 2006. Early mainland Southeast Asian landscapes in the first millennium AD. Annual Review of Anthropology 35, 407–432.

White, J.C., Hamilton, E.G., 2009. The transmission of early bronze technology to Thailand: new perspectives. Journal of World Prehistory 22, 357–397.

Zeitoun, V., Auetrakulvit, P., Forestier, H., Zazzo, A., Davtian, G., Nakbunlung, S., Tiamtinkrit, C., 2013. Discovery of a Mesolithic burial near the painted rockshelter of Ban Tha Si (Lampang province, Northern Thailand): implications for regional mortuary practices. Comptes Rendus Palevol 12, 127–136.

Southeast Asia: Sociocultural Aspects

Hans-Dieter Evers, Universiti Brunei Darussalam, Brunei Darussalam

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Southeast Asia is a very diverse region in terms of societies, languages, and cultures. Pluralism, loose structure, and highly ritualized political systems have been identified as typically Southeast Asian. Boundaries between ethnic groups are blurred, and ethnic classifications vary. Religious diversity has been increased by external religious movements and fundamentalist Buddhism, Christianity, and Islam. Social class formation has produced an increasing gap between the rich and the poor, as well as ethnic clientelism and power struggles between strategic groups. Strategies to develop a knowledge-based economy are widely followed to achieve developed-nation status.

The Sociocultural Construction of Southeast Asia

Conceptualizing Southeast Asia

Southeast Asia includes the Association of Southeast Asian Nations (ASEAN) states of Brunei, Indonesia, Malaysia, the Philippines, Singapore, Myanmar, Thailand, Cambodia, Laos, Vietnam, and East Timor. Despite its extreme diversity in terms of population density, ethnic groupings, religion, politics, and economic development, the area has common characteristics, many of which are constructed by social scientists and politicians and are increasingly recognized (King, 2008, pp. 15–17). In addition to the nation-building efforts of individual governments, a common Southeast Asian identity has emerged, at least in the urban middle classes.

Southeast Asia has undergone a multitude of crises, economic, political, cultural, and social. Most of these have resulted in a fresh look at what courses Southeast Asian societies have taken, and what concepts and theories can be used to analyze new developments. As a matter of fact, some of the new concepts and paradigms have been exported to other world regions, utilized and generalized and entered into the realm of general theory, while their procreators and their region of origin have often been forgotten. Dualism, plural societies, involution, strategic groups, and loosely structured social systems are by now standard terms in social science research, although originally they were empirically based concepts used to describe Southeast Asian social conditions.

What we now call Southeast Asia has gone through a series of cycles of the social and political construction of the region. As Said (1978, p. 5) has rightly asserted, "geographical and cultural entities … are man-made." Localities are not given, but are socially constructed. To mention just a few terms that over time have signified different external and internal efforts to construct a more or less unified image of the area: the 'Golden Chersonese' of Ptolemy, Further-India, Hinterindien, Extrême Orient, Nusantara, the Greater Co-Prosperity Sphere, French Indo-China, Pacific Asia, Southeast Asia, and, more recently, ASEAN.

Indianization and Islamization

The term Southeast Asia became common usage only during and after World War II. Before, the area was seen as culturally colonized by its large neighbors, India and China. Coedès (1968, pp. 15–16) described the classical states as États Hindouisés, or 'Indianized states,' and further stated, "Indianization must be understood as the expansion of an organized culture that was founded upon the Indian conception of royalty, was characterized by Hindu or Buddhist cults, the mythology of the *Purānas*, the observance of the *Dharmaśāstras*, and that expressed itself in the Sanskrit language. Far from being destroyed by the conquerors, the native peoples of Southeast Asia found in Indian society, transplanted and modified, a framework within which their societies could be integrated and developed." This marriage of Austronesian and other local cultures with Indian patterns of sociocultural organization has remained a common background, on which later Chinese, Muslim, and European influences have been grafted. The Balinese and several smaller societies, such as the Tengger, maintain the Hindu tradition.

Islam came to Southeast Asia probably as early as the eighth century AD, but became a major force of social and political change only from the fourteenth century onward. The creation first of a trading and religious elite, then of a political elite through conversion, introduced a new element into the coastal states of Southeast Asia along the straits of Malacca, the northern coast of Java, Sulawesi, and the Moluccas. The process of Islamization is a still-ongoing process.

Religious Diversity and Religious Revivalism

In addition to animistic beliefs and related belief systems, most major world religions are found in Southeast Asia. Indonesia, Brunei, and Malaysia together have the largest Muslim population worldwide. The majority of the population of Myanmar, Thailand, Cambodia, and Laos adheres to Theravada Buddhism, the Philippines has a majority of Christians; and Singapore and Vietnam are characterized by a multitude of religious beliefs often cutting across family and kinship lines, although Confucian ideals are also widespread.

During the second half of the twentieth century, religious revivalist movements swept through most of Southeast Asia. In Thailand, middle class–based religious organizations have preached a return to Buddhist values (Tambiah, 1976). The Javanese mystical tradition has influenced Muslim brotherhoods or led to the foundation of mystical sects: *aliran kebatinan*,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12020-3

Subud, and others (Geels, 1997; Lombard, 1990). The Singapore government has propagated Confucian values, while Christian sects have proliferated and found new adherents among students and young upwardly mobile employees.

In Malaysia, a missionary movement (*dakwah*) has led to a resurgence of Islamic practices, especially in the eastern coastal states of Kelantan and Trengganu. Fundamentalist Islamic thought, backed by Saudi Arabian Wahhabi influences, has led to a growth of restrictive practices, especially for women; the introduction of sharia law in some states; and the growth of terrorist organizations. Islamic revivalism is strong in Malaysia and southern Thailand, and is gaining strength in Indonesia, which has become the largest Muslim country worldwide (Houben, 2003). Brunei Darussalam is still a traditional sultanate, following a state ideology of a Malay Muslim kingdom (Hussainmiya, 1995).

Ethnic Diversity

Boundaries between ethnic groups are blurred, and there are differences between self-interpretation and ethnic classifications imposed by others. In Singapore, the population is officially classified into Chinese, Indians, Malays, and others (usually identified as Eurasians and Caucasians), although there are many more self-ascribed categories. The Malaysian government distinguishes between *bumiputra* (sons of the soil) and immigrants, but there are also *Orang Asli* (aborigines), Indians, and Chinese. People of Arab descent are no longer seen officially as a separate group, but in Indonesia, the culture, legal system, and language of larger ethnic groups such as the Javanese, Balinese, Batak, Minangkabau, and many others are recognized and expressed in cultural, although not political, terms. In Thailand, Laos, Myanmar, and Vietnam, the lowland population is distinguished from 'hill tribes,' an ethnically and linguistically diverse population. The difference between a lowland, rice-growing peasant society and nonsedentary 'tribal groups' practicing shifting agriculture has been emphasized but is disappearing slowly due to assimilation or replacement.

The peoples of Southeast Asia have also been classified along linguistic lines. Despite recent criticism, the classification into Sino-Tibetan, Austro-Asiatic, Tai-Kadai, and Malayo-Polynesian is still widely used (LeBar et al., 1964, 1972). In most countries or regions, we can distinguish one dominant lowland society and several upland satellite ones. Anthropological and sociological research has enriched our knowledge of the major societies that have been politically dominant in the formation of pre- and postcolonial states, and also the large number of smaller societies from the hunter-and-gatherer societies, like the *Orang Asli* groups in Malaysia, to larger groups, like the Shan and Kachin in Myanmar (Leach, 1954); the various hill tribes of Thailand, Laos, and Vietnam (Kunstadter, 1967); or the matrilineal Minangkabau of west Sumatra (Kahn, 1993).

Plural Societies

The in-migration of Arab, Chinese, and South Asian migrants in the wake of the expansion of a colonial economy has created what Furnivall termed a plural society, "with different sections of the community living side by side, but separately, within the same political unit" (Furnivall, 1948, p. 304). This ethnic pluralism was transcribed into colonial and postcolonial administrative practice, giving rise to a classification of the population into Chinese, Malay, Indian, and others in colonial and contemporary Singapore or into indigenous Malays as 'sons of the soil' (*bumiputra*) and migrant communities in Malaysia. Thailand, in contrast, has followed a policy of assimilation, in which all minorities are classified as Thai, whereas in Indonesia the Javanese dominance is balanced precariously by allowing the construction of ethnic, but not political, pluralism in the form of the cultural symbolism of local art forms and dances. With the exception of Singapore, one common national language has been introduced and enforced with varying degrees of acceptance and use in all Southeast Asian states.

Moral Economy

Geertz and others have underestimated the dynamics of Southeast Asian peasant societies. Rebellions and peasant uprisings have continued well into the twenty-first century (e.g., in Aceh and East Timor). The "moral economy of the peasant," illustrated by Scott (1976) with reference to Vietnamese and Burmese societies, indicates that a subsistence ethic governs peasant behavior. Only if the self-defined subsistence level is threatened, have peasants rebelled.

The moral economy also governs petty trade in the informal sector of most of Southeast Asia. The moral obligation of sharing work (*gotong royong* in Indonesia) and profits puts a burden on small traders, mostly women, who find it difficult to accumulate profits to sustain and expand their trading enterprises. There are, however, solutions to this 'trader's dilemma' (Evers and Schrader, 1994): where indigenous traders fail, immigrant trading minorities like the Chinese take over the business. Otherwise, the indigenous traders themselves improve their moral status by forming or joining religious fundamentalist movements, like the Santri traders of Indonesia, to escape the pressure of moral obligations. Often, trade is carried out on a subsistence level, as in the case of female petty traders, where the obligation to share meager profits does not arise because no cash is accumulated.

The informal sector of small handicraft production, petty trade, and other services is widespread in Southeast Asia (Evers and Schrader, 1994) and has increased in Vietnam during the transformation from a planned socialist economy to a more market-oriented economy, but it has declined in Singapore and Malaysia. This sector is playing an important role in cushioning the effects of land shortage in Southeast Asia's rice-growing areas, of rural–urban migration, and of a general lack of social security provisions.

Big Men and Theater States

The further one proceeds southeast through Indonesia and the Philippines, the more evident the Austronesian-Melanesian-Polynesian content of Southeast Asian culture becomes. Wolters (1982) has drawn attention to the 'big man' phenomenon, which he locates in Polynesia. The current emphasis on personal prowess, the power of the oratory, and admiration for superiors and leaders is put into this context. The state and its aristocratic

bureaucracy are admired as much in Thailand as in Malaysia, Brunei, or Indonesia.

The state in these countries is highly ritualized and has been named a 'theater state,' following Geertz's analysis of the precolonial Balinese states (Geertz, 1980). In earlier times, the state was largely a performance with little administrative or military power. It is debatable whether this analysis withstands the test of historical data, but at least the theme of vagueness, looseness, involution, and theatricality in some Southeast Asian societies pervades the anthropological literature.

Loosely Structured Social Systems and Involution

John Embree once classified Thai society as 'loosely structured,' thus igniting a long-drawn controversy among social scientists (Evers, 1980, pp. 163–198). Thai society has a cognatic kinship system, a shallow recognition of descent, and a high degree of individualism, but a highly regulated 'bureaucratic polity' and a tendency to form cliques and patronage. Another, equally unconvincing attempt at using single cultural traits or selected ethnographic details has stressed dualism as a common characteristic of Southeast Asian societies. There was a greater acceptance of the work of Geertz, who classified Javanese society as "involuted," by "overdriving of an established form in such a way that it becomes rigid through inward over elaboration of detail" (Geertz, 1963, p. 82). This process applied mainly to wet rice agriculture and its tenancy system, but was also seen at work by other scholars in other regions like the Philippines or Thailand, or in the relatively slow process of Southeast Asian urbanization up to the middle of the twentieth century (Evers and Korff, 2000).

Contemporary Southeast Asian Studies

Diasporas and Business Networks

Migrant communities, like the Overseas Chinese originating from southern China, the South Indian Chettiar, or the Hadramaut Arabs, have been classified as trading minorities in view of their economic position in Southeast Asian societies. Others migrated to Southeast Asia in large numbers at the beginning of the twentieth century to work in plantations, mines, and the construction sector. Chinese business has been studied in great detail. Singapore has become the center of the Chinese business community (Chan and Chiang, 1994; Menkhoff and Gerke, 2000). Confucian ethics, patrilineal kinship–based networks, and the patriarchal organization of family firms have been held responsible for economic success and the predominance of the Nanyang Chinese. Prosecution, discrimination, and conflict have strained interethnic relations in Malaysia, Indonesia, and Vietnam.

Social Class and Strategic Group Formation

Colonial Southeast Asian societies were "moulded on racial principles: belonging to the dominant white upper caste provided one with prestige and power largely independent of one's personal capabilities. A strict ritual was introduced and maintained, by force when necessary, to preserve the white caste from contacts with Asiatics on a basis of equality and to maintain the former's prestige as the dominant group" (Wertheim, 1968, p. 432).

In the course of decolonization and modernization, Southeast Asian societies have undergone a process of class formation (Evers, 1966, 1973, pp. 108–131; King, 2008, pp. 98–128). The colonial upper classes were replaced by ethnically differentiated elites, in which indigenous aristocrats, Chinese and European business people, and the upper ranks of a fast-growing bureaucracy and military officers played a major role. The concept of a strategic group can be used to analyze this new situation more adequately. The pattern of strategic group formation differed from country to country. In Thailand, which had escaped the fate of colonialism, the growing bureaucracy, backed by sections of the military, revolted against the aristocracy and replaced the absolute monarchy in 1932 with a new regime, in which the strategic group of the military gained the upper hand, until a growing group of professionals, university lecturers, and students managed to establish a democratic government. Strategic group conflict continued, mediated by the electoral process, and put big business into power in 2001. Similar strategic group conflicts, rather than class struggle, signified the situation in most countries, except perhaps in the socialist states of Vietnam and Laos, where party cadres, another strategic group, maintained power after the communist victory in the Vietnam War (Evers and Benedikter, 2009). The partial reprivatization of land and the deregulation of the economy introduced, however, new social processes, from rural–urban migration to corruption of government officials and the formation of a group of newly rich businessmen.

Class formation moved ahead, and the growth of a middle class became a much-researched topic in Southeast Asia (Robison and Goodman, 1996; Gerke, 2000; Embong, 2002). The growth of a new middle class was seen as a basic requirement for democratization and the building of a civil society. Whereas in Thailand, the Philippines, and Indonesia, a civil society with a proliferation of nongovernment organizations developed and pressed successfully for reformation and the introduction of democracy, the prodemocracy movement met with some success from 2011 in Myanmar, although a strategic group of military officers remained in power.

A working class emerged first through the development of plantations and mining, and later through industrialization (Ibrahim, 1998; Evers, 1966). Throughout Southeast Asia, the fruits of rapid development became during the new millennium increasingly concentrated in a super-rich upper class of business tycoons and more or less corrupt government officials and land speculators.

The middle classes have developed a special lifestyle of their own that blends Western consumerism with elements of Asian modes of living. Hawker centers, food courts, and open-air markets exist alongside big department stores, pizza parlors, and hamburger outlets. The anthropology of Southeast Asian food has to be concerned with syncretism and lifestyling as much as with religious symbolism.

Along with the demand for consumer goods, the demand for knowledge has also increased. In fact, the ASEAN (Association of Southeast Asian Nations) region has undergone a tremendous expansion of higher education. In the core states of Thailand,

Malaysia, Singapore, the Philippines, and Indonesia, the number of students and teachers, and of secondary schools, colleges, and universities, have risen dramatically. Malaysia and Singapore, in particular, are on their way to becoming knowledge societies, where economic and social development is increasingly knowledge driven. Most ASEAN leaders and governments have developed visions of the future that target a knowledge society as a way to achieve parity with the Western world. This vision is but one in a long line of ideas on how Southeast Asia should be seen, interpreted, and eventually constructed.

Family and Gender Relations

The small nuclear family is widespread, or has become the norm more recently, as in Vietnam (Pham Van Bich, 1999) or among Singapore Chinese. The position of women has been regarded as strong throughout Southeast Asia (Watson-Andaya, 2007), in particular in matrilineal societies like the Minangkabau (Kahn, 1993). Demographic changes and the decline of fertility due to family planning have strengthened the position of women in Thailand, Vietnam, and Indonesia, giving them greater control over their own lives (Gerke, 1992; Barry, 1996). Industrialization and the employment of female factory workers, especially in the garment and electronics industries, have led not only to migration and chances for educational attainment but also to exploitation and harassment of women (Sen and Stivens, 1998).

Globalization, Urbanization, and Locality

Right into the 1980s, Southeast Asia, except for the city-state of Singapore, showed only relatively low rates of urbanization (Evers, 2011). Since then, urbanization has taken off, creating large urban conglomerates like the greater Bangkok area, Jabotabek (Jakarta, Bogor, Tanggerang, and Bekasi), the Manila metropolitan area, and the Klang Valley in Malaysia, which includes the federal territory of Kuala Lumpur, the new administrative city of Putrajaya, and the high-technology center Cyberjaya. Southeast Asian urbanism has its roots in the Chinese-influenced walled-city form, like Bangkok or Mandalay; the Malay or Muslim ports like Malacca, Banten, or Makassar; the court cities like Yogyakarta; and the colonial cities like Yangoon Singapore, or Manila (Evers and Korff, 2000). Next to the city centers with their culturally globalized population, we find localities of migrant groups, still maintaining a sociocultural flavor of their own. The various China towns, 'Little India' in Singapore (Siddique and Shotam, 1982), and localities in greater Manila (Berner, 1997) should not be seen as remnants of traditional culture, but as newly constructed forms of urban society.

Values and Knowledge

Southeast Asian cultures as systems of beliefs, values, symbols, and presentations are as diverse as the ethnic plurality of the local and migrant populations. There are, however, some underlying trends that have been used to construct national cultures or an ASEAN identity. The family and its values, reaching decisions through consensus rather than the majority vote, and a religiously grounded work ethic have been stressed in the debate on Asian values. Social scientists have taken up the cultural analysis of modernity. In his study of Thai, Indonesian, and Philippine middle-class culture, Mulder (1992) has stressed the reinterpretation of classical concepts for the construction of modernity.

The knowledge on Southeast Asian societies is increasing fast (King, 2008). Most of the relevant books and articles are still produced by foreign scholars, affiliated to universities or research institutions around the globe, but an increasing proportion of works is written by Southeast Asian nationals or by scholars attached to Southeast Asian universities (Shamsul, 2011; Evers and Gerke, 2012). An epistemic culture of local Southeast Asianists is emerging, contributing to the further construction of Southeast Asia as a sociocultural unit.

An increasing number of Southeast Asian scientists and knowledge workers contribute to the emergence of knowledge-based economies and societies (Menkhoff et al., 2011), with Singapore and Malaysia as the most advanced countries. Despite the cosmopolitan nature of these societies, ethnic differences still play an important role in politics and social relations.

See also: Asia, Sociocultural Overviews: Korea; Asia, Sociocultural Overviews: South Asia; Big Man, Anthropology of; Diaspora; East Asian Studies: Culture; East Asian Studies: Economics; East Asian Studies: Gender; East Asian Studies: Politics; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity: Anthropological Aspects; Japan: Sociocultural Aspects.

Bibliography

Barry, K. (Ed.), 1996. Vietnam's Women in Transition. St. Martin's Press, New York.
Berner, E., 1997. Defending a Place in the City. Localities and the Struggle for Urban Land in Metro Manila. Ateneo de Manila University Press, Quezon City.
Chan, K.B., Chiang, C., 1994. Stepping Out: The Making of Chinese Entrepreneurs. Prentice Hall, New York.
Coedès, G., 1968. The Indianized States of Southeast Asia. East-West Center Press, Honolulu, HI.
Embong, A.R., 2002. State-led Modernization and the New Middle Class in Malaysia. Palgrave, New York/Houndmills.
Evers, H.-D., 1966. The formation of a social class structure: urbanization, bureaucratization and social mobility in Thailand. American Sociological Review 31 (4), 480–488.
Evers, H.-D. (Ed.), 1973. Modernization in Southeast Asia. Oxford University Press, Singapore.
Evers, H.-D. (Ed.), 1980. Sociology of Southeast Asia. Oxford University Press, Singapore.
Evers, H.-D., 2011. Urban symbolism and the new urbanism of Indonesia. In: Nas, Peter J.M. (Ed.), Cities Full of Symbols: A Theory of Urban Space and Culture. Leiden University Press, Leiden, pp. 187–196.
Evers, H.-D., Benedikter, S., 2009. Hydraulic bureaucracy in a modern hydraulic society – strategic group formation in the Mekong Delta, Vietnam. Water Alternatives 2 (3), 416–439.
Evers, H.-D., Gerke, S., 2012. Globalisation of social science research on Southeast Asia (Chapter 5). In: Ibrahim, Zawawi (Ed.), Social Science and Knowledge in a Globalising World. Malaysian Social Science Association, Kajang, pp. 103–116.
Evers, H.-D., Korff, R., 2000. Southeast Asian Urbanism: The Meaning and Power of Social Space. LIT Verlag/St. Martin's Press, New York/Münster.
Evers, H.-D., Schrader, H. (Eds.), 1994. The Moral Economy of Trade. Routledge, London.
Furnivall, J.S., 1948. Colonial Policy and Practice. Cambridge University Press, London.
Geels, A., 1997. Subud and the Javanese Mystical Tradition. Curzon, Richmond, VA.
Geertz, C., 1963. Agricultural Involution. University of California Press, Berkeley, CA.
Geertz, C., 1980. Negara, the Theatre State in Nineteenth-Century Bali. Princeton University Press, Princeton, NJ.
Gerke, S., 1992. Social Change and Life Planning of Javanese Women. Breitenbach, Saarbrücken/Fort Lauderdale, TX.

Gerke, S., 2000. Global lifestyles under local conditions: the new Indonesian middle class. In: Chua, B.-H. (Ed.), Consumption in Asia. Lifestyles and Identities. Routledge, London/New York, pp. 135–158.

Gomez, E.T., 1999. Chinese Business in Malaysia. Curzon, Richmond, VA.

Houben, Vincent J.H., 2003. Southeast Asia and Islam. The Annals of the American Academy of Political and Social Science 588, 149–170.

Hussainmiya, B.A., 1995. Sultan Omar Ali Saifuddin III and Britain: The Making of Brunei Darussalam. Oxford University Press, Oxford/Singapore.

Kahn, J., 1993. Constituting the Minangkabau. Peasants, Culture, and Modernity in Colonial Indonesia. Berg, Oxford, UK.

King, V., 2008. The Sociology of Southeast Asia: Transformations in a Developing Region. NIAS Press, Copenhagen.

Leach, E., 1954. Political System of Highland Burma. London School of Economics, London.

LeBar, F.M., Gerald, H.G., Musgrave, J.K. (Eds.), 1964. Ethnic Groups of Mainland Southeast Asia. HRAF, New Haven, CT.

LeBar, F.M., Gerald, H.G., Musgrave, J.K. (Eds.), 1972. Ethnic Groups of Insular Southeast Asia. HRAF, New Haven, CT.

Lombard, D., 1990. Le Carrefour javanais: essai d'histoire globale, vols. 3. École des Hautes Études en Sciences Sociales, Paris.

Menkhoff, T., Gerke, S., 2000. Chinese Entrepreneurship and Asian Business Networks. Curzon, Richmond, VA.

Menkhoff, T., Evers, H.-D., Chay, Y.W., Pang, E.F. (Eds.), 2011. Beyond the Knowledge Trap: Developing Asia's Knowledge-based Economies. World Scientific, New Jersey, London, Singapore, Beijing.

Mulder, N., 1992. Inside Southeast Asia: Thai, Javanese and Filipino Interpretations of Everyday Life. Editions Duan Kamol, Bangkok.

Pham Van, B., 1999. The Vietnamese Family in Change: The Case of the Red River Delta. Curzon, Richmond, VA.

Robison, R., Goodman, D.S.G. (Eds.), 1996. The New Rich in Asia: Mobile Phones, McDonald's and Middle-class Revolution. Routledge, London.

Said, E.W., 1978. Orientalism. Pantheon, New York.

Scott, J., 1976. The Moral Economy of the Peasant. Yale University Press, New Haven, CT.

Sen, K., Stivens, M. (Eds.), 1998. Gender and Power in Affluent Asia. Routledge, London.

Shamsul, A.B., 2011. Producing knowledge of Southeast Asia: a Malaysian viewpoint. In: Yusoff, Anis Y. (Ed.), Shamsul A.B., His Observations, Analyses and Thoughts. Institute of Ethnic Studies (KITA), Universiti Kebangsaan Malaysia, Bangi, Selangor, pp. 94–122.

Siddique, S., Shotam, N.P., 1982. Singapore's Little India: Past, Present and Future. ISEAS, Singapore.

Tambiah, S.J., 1976. World Conqueror and World Renouncer: A Study of Buddhism and Polity in Thailand against a Historical Background. Cambridge University Press, Cambridge, UK.

Watson-Andaya, B., 2007. Studying women and gender in Southeast Asia. International Journal of Asian Studies 4, 113–136.

Wertheim, W.F., 1968. Asian society: Southeast Asia. In: Sills, D.L. (Ed.), 1968. International Encyclopedia of the Social Sciences, vol. 1. Macmillan/The Free Press, New York, pp. 423–438.

Wolters, O.W., 1982. History, Culture, and Region in Southeast Asian Perspectives. ISEAS, Singapore.

Southern Africa: Sociocultural Aspects

Robert Thornton, School of Social Science, University of the Witwatersrand, Johannesburg, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The southern African region comprises of South Africa, Lesotho, Swaziland, Botswana, Namibia, Angola, Mozambique, Malawi, Zimbabwe, and Zambia. The southern African region shares similar and interrelated colonial histories, postcolonial interactions, and the broad outlines of their cultural histories. The Southern African Development Community provides a forum for governmental consultation, and economic and cultural cooperation, but also includes Tanzania, and French-speaking Democratic Republic of Congo, and the Indian Ocean island states of Seychelles, Mauritius, and Madagascar. From the middle of the twentieth century the decolonization process and its aftermath dominated the history of the region, but all countries derive a great deal of their cultural traditions today from the colonial period, including their official, dominant languages, and religion.

The precolonial African cultural heritage is seen mainly in the rural areas, especially in the institutions of chiefship ('traditional authority') that continues to hold a variable degree of autonomy from the state, especially with respect to land rights and land tenure. Swaziland, for instance, remains an absolute monarchy. Mozambique formerly banned chiefship, while tolerating it in practice. Indigenous traditions persist in indigenous healing traditions, divination, funerary, and kinship rituals, and also in many indigenous, nominally Christian religious practices. The former Portuguese colonies, Angola and Mozambique, continue to use Portuguese as their official languages, while the former British colonies and dominions (South Africa) use English. All also have many indigenous languages, with no single African language dominant in any country or across the region. Since the nineteenth century, all countries have had Christian majorities divided into many hundreds of sects and denominations, many with roots in indigenous religious beliefs and practices. Pentecostal and Evangelical congregations are growing rapidly. Small Muslim and Hindu minorities, predominantly of Indian origin, exist mainly along the Indian Ocean coast in South African and Mozambique. South Africa has a significant Jewish population whose ancestors arrived in the nineteenth century, mainly from the Baltic States. Ethnic groups, languages, religious affiliations, and precolonial polities span national boundaries, and movement of populations across all boundaries is relatively uncontrolled.

With a few exceptions, countries have struggled to establish viable democratic political systems and bureaucracies, to overcome ethnic, racial, regional, and economic conflicts, and to diversify and stabilize economies often dominated by only one or two industries (such as mining in Zambia and Botswana, or tobacco in Zimbabwe). Accordingly, overall economic development is precarious, with large informal, untaxed, cash-only economies providing a large proportion of economic activity in all. Botswana and South Africa have the highest per capita incomes, comparable with Brazil, Thailand, or Turkey. Mozambique, Malawi, and Zimbabwe are the poorest, comparable only with other poor African countries such as the Democratic Republic of the Congo. HIV adult prevalence rates (percentage of people infected with HIV in the age range of 15–49 years) are the highest in the world, ranging between roughly 10% in Malawi and 27% in Lesotho, Botswana, and parts of South Africa. Largely because of extremely high HIV/AIDS prevalence, malaria, and TB, life expectancy ranges between 48 and 58 years compared with a world average life expectancy of 70 years.

The small countries of Lesotho, Botswana, and Swaziland are, for the most part, monoethnic states dependent on South Africa, but have significantly different internal political arrangements and economic fortunes. South Africa is the most economically and ethnically diverse country in the region, with the strongest state and institutions of civil society. It dominates the region's economy. All other countries, especially landlocked Lesotho, Swaziland, Zimbabwe, Zambia, and Malawi, depend to a large degree on South African transport services and harbors. South African investments and companies increasingly dominate the economies of its regional neighbors.

Regional trade volumes increase annually, but the largest trading partners still remain outside of the region, primarily in Western Europe, but increasingly in China and India. At the beginning of the twenty-first century, all countries had begun to seek increasing intraregional cooperation in order to reverse the effects of conflict, isolation, economic decline, and political weakness of state structures. Thabo Mbeki, President of South Africa from 1999 to 2008, articulated this as the search for an 'African Renaissance,' but the concept proved to have little resonance by the early twenty-first century, even as regional economies, cultures, and populations continued to converge.

These factors both constrain and help to shape the sociocultural aspects of the region.

History

Paleoanthropological and archaeological evidence shows that 'the cradle of mankind' included southern Africa, and extended north to Ethiopia and Somalia. The earliest prehumans and humans evolved in the region, but then moved north out of Africa, apparently disappearing entirely from southern Africa between 70 000 and 40 000 years ago. Although these early humans were the ancestors of all of humanity, there is no evidence that links them directly to contemporary southern African populations. Earliest ancestors of the San (Bushman)

and Khoe (also Khoikhoi, Khwe, Hottentot) appeared in the archaeological record again around 20 000 years ago, while ancestors of the dark-skinned people from Central Africa began to appear around 2000 years ago. People from Europe first settled in South Africa 500 years ago, and in Angola, Namibia, and Mozambique shortly after that. Trade with the Indian Ocean existed along the coast of Mozambique up to 2000 years ago, and with the interior perhaps 1000 years ago. Small populations of Indians arrived in South Africa and elsewhere in the region from the mid-nineteenth century. Even smaller population of Chinese arrived around the same time, but south Asian and Chinese populations in all countries began to increase rapidly in the early twenty-first century. Today, all towns and cities in South Africa, for instance, have significant recently arrived Chinese, Indian, and Pakistani populations.

All African languages in the region derive from only two of the six major language families of Africa (Greenberg, 1966; Heine and Nurse, 2000, 2011), but most named languages are in fact clusters of dialects. The speakers of the diverse Khoisan language group (comprising five clusters, that are at best distantly related to each other) have lived in the region, and only here, from at least 20 000 years ago. These languages, noted for their click sounds that occur nowhere else in the world, are all but extinct today. Beginning around 2000 years ago, people speaking languages of the Bantu family, and possibly also the Khoe speakers, moved into the region quite rapidly from the north. Bantu languages originated north of the Congo forest zone in the region of today's Cameroun, while the Khoe languages may have come from the northeast. At about the same time, technology to smelt iron and make durable fired pottery spread throughout the region. Cattle, goats, and a mixture of crops from the Sudanic zone (millets and sorghum) and from the Indian Ocean region (sweet potatoes, bananas) also began to appear. At some time a crop from the New World, maize, became central to the diet of the entire region well before European incursions.

Before the arrival of agriculture, sheep, goats, and cattle, and just before the first millennium CE, the Khoisan people sparsely populated most of southern Africa. The Khoisan language family includes the Bushman (San) languages and the KhoeKhoe (also Khoi, Kwe, Hottentot). Social and economic difference distinguished the two groupings. The Bushman were primarily hunters and gatherers, living in small mobile bands of no more than 35–40 people. The Khoe were cattle and sheepherders, living in larger tribal groupings. Early European settlers at the Cape of Good Hope began to trade with the Khoe peoples (then known as 'Hottentots') in the seventeenth century, but as trade relationships broke down and as European pastoralists and farmers ('Boers') began to move inland in the eighteenth century, most Khoe pastoralists were gradually absorbed as clients and laborers, and occasionally as wives, by the Boers. Bushman were killed in skirmishes, displaced to remote deserts, or incorporated as servants and wives by European people moving inland from the south, and by Bantu-speaking Africans moving southward. By the nineteenth century, Khoisan speakers remained only in the dry areas inland from the west coast of South Africa, Namibia, and Angola, and in the Kgalagadi (Kalahari) desert of Botswana. The relatively large Nama (Khoi)-speaking group in Namibia (11% of only 2 million people) give 'Namibia' its name ('land of the Nama'). In South Africa where 'colored' people preserve much of the San and Khoe genetic heritage, Bantu languages such as Xhosa and Zulu preserve the linguistic heritage in the form of some vocabulary borrowed from the Khoisan speakers, and the characteristic 'click' sounds that were adopted by the Nguni languages (Zulu, Xhosa, Swazi). Their art, painted on rocks throughout the region, gives testament of a rich culture.

People who speak languages of the Bantu languages family also share broadly similar cultures. With the exception of their spread to Indian Ocean islands such as the Comoros and Zanzibar, they are restricted to tropical and subtropical continental southern, central, and eastern Africa. The culture and political economy of early Bantu speakers, like the Khoe, were characterized by residence in small, dispersed agricultural and pastoral settlements. A leading man or 'chief' usually presided over a 'court' composed of senior or royal males. The court usually excluded women, but included all surrounding people who wished to acknowledge the court's authority. In the precolonial period most people moved in small groups, combining and splitting from one another over time. Settled polities began to emerge in the fifteenth to sixteenth centuries, as, for example, around the court known as Great Zimbabwe and a few other large towns, sacred sites, and senior courts. Loose alliances and trade networks were formed between settlements, kin groups, and courts throughout the region. As these people were gradually incorporated into colonial states, formal hierarchies of chiefs developed, and became incorporated into a hybrid political system. Missionaries named languages for the purpose of translating the bible, and to enable governments to establish schools, only in the nineteenth century. Political boundaries and ethnic identities as they exist today did not emerge until the late eighteenth century.

Precolonial ritual practices focused on communication with the ancestors, healing cults, and acts intended to bring rain, fertility, and protection from witches, evil, and disease. Individuals, families, clans, or larger tribal or village groupings conducted rituals regularly. While there were generally no 'priests' in these traditional ritual systems, all societies relied on diviners, healers, and prophets who acted either as mediums for spirits during healing rituals or occasionally interpreted the will of the spirits in a political context. Specialist healers, 'shamans,' spirit mediums, or *sangomas* achieved altered states of consciousness known as 'trance' during these occasions. The institution of religious prophecy continues to have significant political and social effects. The role of traditional prophets and spirit mediums in the *Chimurenga*, 'war of liberation,' fought in Rhodesia-Zimbabwe was especially important, but resistance and syncretic religious movements led by traditional prophets appeared in all southern African countries during colonial times, and are still important. Strong beliefs in the existence of witches – individuals who caused evil or illness often unbeknownst to themselves – and beliefs in the existence of animal familiars of witches or other kinds of spiritual powers inherent in plants and animals are also widespread. Today, majorities in all countries are at least nominally Christian, with small numbers of Muslims such as the Yao of Mozambique, and the Swahili speakers in Mozambique, Malawi, and eastern Zambia. Many aspects of

precolonial belief have been incorporated into indigenous Christian beliefs and practices, while minorities of people continue to practice traditional religion in preference to, and sometimes in combination with, Christianity or Islam.

The warrior king Shaka established the Zulu kingdom in 1816 as the culmination of a series of raids and wars that gradually incorporated large numbers of agricultural and pastoral peoples in the southeastern lowlands ('lowveld') of South Africa. Repercussions of the Zulu conquest were felt as far north as Malawi and eastern Zambia, southern Tanzania and Mozambique, where displaced groups of people under warrior chiefs established temporary hegemony. These people were known as the 'Ngoni' in the north, or Nguni and Ndebele (Matebele, 'people of the court') in the south. Shaka is credited with introducing a number of military innovations, and with the banning of circumcision rituals, a social innovation that allowed him to recruit males into armies under his command. They met the British colonial advance in the latter half of the nineteenth century. The battle between Zulu and British armies at Isandlwana in 1879 was a famous defeat for the British forces, but the Zulu were eventually overcome and incorporated into the Colony of Natal. The contemporary name of the South African province of KwaZulu-Natal province recognizes the continuing balance between 'Zulu' and English and Afrikaans languages and cultures today.

The rise of the Zulu kingdom, together with movements inland of the Portuguese from the east coast, and the Afrikaners (Boers) from the British colony of Cape of Good Hope, led to general disruption of settled life in south-central Africa. This, in turn, created opportunities for widespread incursions, first by African and Arab slave and ivory traders from the east coast, and later for European adventurers, hunters, traders, settlers, and missionaries. The general lack of strong centralized states or other large-scale political organizations in the region contributed to the advance of colonial rule and expropriation of land by Europeans for farms and mines.

European Christian missionaries established mission stations in Nyasaland (now Malawi) in response to the devastation and social dislocation caused by Ngoni/Nguni incursions and led the way for the British government to declare Nyasaland a Protectorate. Zulu offshoots established the Gaza kingdom in the southern central region of Portuguese Mozambique over the Tsonga-speaking people and settled in large numbers in southern Zimbabwe as the 'Ndebele' people. Tswana-speaking groups, originally deriving from ancestral peoples known in the nineteenth century as 'Hurutshe' (also Ro(t)ze, Rozvi, Rotse) in the central highlands region of South Africa, Botswana, and Zimbabwe fled north in response to Zulu incursions and established temporary hegemony over the Lozi/Rotsi people in western Zambia, known as Barotseland. They managed to impose much of their language and culture on the previously independent populations. In South Africa, the Zulu expansion and its repercussions had the effect of disrupting virtually all previous settlements of people during the *Mfecane* (also *Difeqane* in Sotho, 'the crushing'), a period of vast population movements and reorganization lasting through the middle of the nineteenth century. Out of this emerged several strongly integrated 'tribes' as followers of powerful chiefs: the Sotho of Basutoland (now Lesotho), the Swazi of Swaziland, and the Tswana tribes of Bechuanaland (now Botswana). It also created an opportunity for white Afrikaans-speaking farmers, the Boers, to move north out of the Cape Province, a British Colony since 1802. After the Great Trek of 1836–37 during which hundreds of Afrikaans farmers moved north out of the Cape Colony, they established several independent republics, including the Orange Free State and the Suid-Afrikaanse Republiek ('South African Republic,' known as 'The Transvaal'). These maintained independence until British colonial forces defeated them during the South African War (1899–1902), also known as the 'Boer War,' or the 'Second War of Independence' by the Afrikaans-speaking citizens of these republics. Some of the earliest instances of 'guerrilla warfare' (informal, hit-and-run) occurred after the Boer forces were defeated in conventional battles. The first 'concentration camps' were created in which many Boer families were interned and died. Resentment and political tension between the Afrikaans-speaking (Afrikaners) and the English-speaking white South Africans continued to define political events for most of the twentieth century, while tension between black and white South Africans in the same period eventually led to the fall of Apartheid and constitutional, universal-franchise democracy from 1994.

Colonialism and Independence Struggle

In the late nineteenth century, Swaziland, Basutoland (now Lesotho), and Bechuanaland (Botswana) were declared British Protectorates after the chiefs or kings of these semi-independent polities signed treaties. This was done partly in order to secure them as labor reserves to provide labor for European farmers and the mines, and partly to keep them from the hegemony of the Zulu on the one hand and the Afrikaner (Boer) states on the other. The Zulu, unwilling to enter into negotiations, were eventually defeated in 1879 by British forces. Their king was exiled briefly but eventually returned as king of the Zulu people under the authority of the Colony of Natal.

Mozambique was colonized by Portugal in the fifteenth century but only came under effective control in the late nineteenth century. Zambia and Zimbabwe (previously Northern and Southern Rhodesia) were given as concessions to the British South Africa Company under Cecil John Rhodes, becoming known as Rhodesia in 1895. In 1924, the British South Africa Company was dissolved, and Northern Rhodesia and Southern Rhodesia became Crown Colonies under local white settler control. Together with the Protectorate of Nyasaland, they joined together in the Central African Federation in 1951 until independence was achieved by Northern Rhodesia and Nyasaland in 1964. The colony of Northern Rhodesia was renamed Zambia, while Nyasaland became Malawi.

Military conflicts in the later decades of the twentieth century severely limited the pace of economic and political development throughout the region. Protracted guerrilla struggles between settler populations (European with some from India) and black African nationalist organizations in Mozambique, Zimbabwe (former Southern Rhodesia), Angola, and Namibia (former South-West Africa) spanned most of the last half of the twentieth century. These conflicts generated

large cross-border refugee movements that, together with labor migration and urbanization, have resulted in large and long-term shifts in populations of all countries. Urban growth has been greatest in South Africa and Angola, and least in Malawi and Lesotho.

Southern Rhodesia continued to be ruled by its white minority 'settler' population as Rhodesia, and declared its independence from Britain in 1965. Though unconstitutional and condemned by Britain and the United Nations, the largely white government under Prime Minister Ian Smith continued to rule Rhodesia until 1980, despite massive trade and political sanctions. After the Zimbabwe African National Union (ZANU), with an ethnic base among the Shona, and the largely Ndebele Zimbabwe African Peoples Union (ZAPU) succeeded in defeating the white settler regime, Rhodesia became Zimbabwe under Robert Mugabe, the leader of ZANU. The defeat of the Smith's Rhodesian government was made possible by close cooperation between ZANU and African nationalist forces in Mozambique. These had begun by the late 1960s to wage a war of liberation, and offered safe havens for ZANU forces.

Zimbabwe's independence was followed by a period of violent repression by ZANU against the Ndebele followers of ZAPU from the middle of the 1980s, though peace was eventually restored. Mugabe's rule became increasingly autocratic despite maintenance of some democratic features, including elections. Twenty years after independence, Mugabe's government expropriated much of the land that was under the control of the country's remaining farmers of European descent. This group still retained about half of the arable land, and many of these farms were transferred to politically connected members of Mugabe's inner circle, while other farms were occupied by subsistence farmers, or were neglected and returned to bush. At the turn of the millennium, Zimbabwe entered a period of economic and political collapse. Its currency inflated to bizarre levels, and was eventually abandoned in favor of the US dollar and the South African Rand.

Operating from Tanzania, the Mozambique Liberation Front (FRELIMO) succeeded in gaining control of northern Mozambique in the early 1970s. In 1974, the Portuguese colonial government fell, due in large measure to the cost of its colonial wars in Angola and Mozambique. Mozambique was granted independence in the following year but suffered 16 years of civil war between FRELIMO and the Mozambique National Resistance (RENAMO) after the Portuguese withdrawal. Namibia, the former South-West Africa, had been brought under German control as a result of the Partition of Africa at the Berlin Conference in 1884. It was seized by South African forces acting under British authority during World War I, and then ceded to South Africa as a Trust Territory under the terms of the Armistice, and in accordance with a mandate of the League of Nations. After a protracted liberation struggle, it became independent Namibia under the government of its liberation organization, the South West People's Organization. German and Afrikaans are still widely spoken in Namibia, although English is the official language, and the Ovambo constitute about half the population.

With its Portuguese colonial heritage, Mozambique is clearly part of the southern African region culturally and historically. The Tsonga-speaking people of southern Mozambique straddle the border with South Africa. Mozambicans have long been intensively involved in the South African economy, first as labor migrants to white farms from as early as the eighteenth century, and later as laborers on South African mines from the late nineteenth century until the present. South African capital has flowed strongly into Mozambique after the end of its protracted civil war.

In South Africa, the African National Congress (ANC) came to power in South Africa in 1994 after the release in 1990 of Nelson Mandela from 24 years of imprisonment. With the end of colonialism, in all countries except South Africa, large numbers of white people withdrew, either moving back to European countries, or Israel, or into South Africa. Today, well over 95% of the population of all countries except South Africa is black African although economically significant minorities of European, Indian, and 'mixed' descent also live in all countries.

States and Politics

With the exception of South Africa, state institutions in the region are all relatively weak. The colonial states of Mozambique and North and South Rhodesia were never very strong, focusing mainly on controlling black labor migrations and political movements, and on delivering security and services to white farmers, miners, and business interests. Partly autonomous chiefdoms in rural areas and the activity of liberation forces also removed large areas from effective government control. Parts of these countries remained effectively beyond the reach of the state, neither paying much in taxes nor receiving many services from the central authorities. At the end of the twentieth century, however, all except Swaziland had democratic governments elected under universal franchise. The benefits of the postcolonial transition to democratic politics have been limited, however, by long liberation struggles in some cases (Zimbabwe, Mozambique, Namibia), postcolonial civil war (Mozambique), rampant corruption, mismanagement (Zambia, Zimbabwe, Malawi), and, for most, increasing poverty and environmental damage.

Throughout the last half of the twentieth century and the early twenty-first century relatively few persons and parties dominated the political process. Twenty years after the first universal-suffrage elections in 1994, however, South Africa had very active civil society, with three presidents having been elected. There were 142 registered political parties, most of them small, with only 13 represented in the national parliament. The ANC dominated national power centers, while the Democratic Alliance governed one province. Zimbabwe's two parties, the Movement for Democratic Change and Zimbabwe African National Union–Patriotic Front governed by coalition from 2009 to 2013, but then reverted to one party rule under Robert Mugabe, president since 1980. All other countries have been ruled for long periods by one party and by a few long-term presidents. Free political expression has often been suppressed. In South Africa, the National Party, dominated by white Afrikaners, ruled from 1949 until 1994, one of the longest periods of rule by a single party anywhere in the postwar world. The National Party implemented a set of laws and practices known as Apartheid ('separateness') that attempted to maintain exclusive political and economic power by 'whites,'

that is, people primarily of European descent who spoke Afrikaans and English. Racism was institutionalized, and black opposition parties were banned. Their members, such as the ANC leader Nelson Mandela, were either jailed or forced into exile. Only whites were allowed to vote. Within these limits, however, the political process was orderly, with presidents stepping down after their terms of office had expired. Following this trend, Nelson Mandela, the first president of a free South Africa, stepped down after serving a 5-year term to make way for Thabo Mbeki, his elected successor, who was succeeded in turn by Jacob Zuma.

The problem of how to integrate 'traditional' African politics of the rural areas (based on territorial, tribal, and kin-based polities) with the modern, urban-centered political systems deriving from Europe (either liberal-democratic or centralist-Marxist) has persisted to the present. While Botswana, Lesotho, South Africa, and Zimbabwe maintained 'advisory' constitutional roles for 'traditional authorities' (that is, predominantly rural tribal 'chiefs'), others such as Mozambique formally abolished chiefship. Tribal groups such as the Himba in Namibia and the Lozi of Zambia maintain kings, but without recognition by the state's constitution. With the exception of Swaziland, all possessed more or less liberal-democratic parliamentary political systems by the end of the twentieth century. Swaziland's absolute monarchy is based partly on precolonial African traditions of governance and partly on modern bureaucracy and police power. Botswana institutionalized traditional rulers who have a role in government through the House of Traditional Authorities. Lesotho maintained a constitutional monarch in an otherwise democratic dispensation. In white-ruled Zimbabwe (before 1980) and South Africa (before 1992), rural chiefs became an integral element of government control and were subsequently seen as part of Apartheid's oppression. During the 1970–80s, the Apartheid government created internal 'homeland republics' under the institution of modernized, salaried chiefships provided with minor bureaucratic support. Surprisingly, the twenty-first century saw strong continuity of these institutions in South Africa, even as intense political conflict grew around the role and power of local chief's courts in the national judicial system. In Mozambique, remnants of precolonial traditional political systems tended to divide the country into a traditional north, represented by RENAMO, and a more modernizing south, represented by FRELIMO, resulting in protracted civil war from 1977 to 1992. Chiefs continue play public roles in countries of the region.

Livelihood

While South Africa began to develop industrial-scale mining and large industry from the late nineteenth century, all countries have small industrial sectors today. Angola possesses huge oil reserves, while Zambia's copper mines have been central to its economy. In the later decades of the twentieth century, all countries in the region suffered periods of isolation from world markets as the result of political sanctions (Rhodesia-Zimbabwe and South Africa), distance from markets (Lesotho, Malawi, Zambia, Swaziland), warfare (Mozambique, Angola), or declining market prices for their primary products (Zambia). Worldwide political and economic sanctions were applied first to Rhodesia (Zimbabwe) throughout the 1970s during the period of white settler rule, and then to South Africa during the final decades of the struggle against Apartheid.

A polarity between the 'modern' urban centers and the surrounding 'traditional' rural regions has characterized the broad politico-economic structure of the region since colonial times. Throughout the twentieth century migration of rural, mainly agricultural people to urban areas has skewed gender balances as males moved to urban areas in search of wage labor, especially in mines and industry, while women tended to stay in the rural areas. Over two-thirds of the South African population is urban, while only one-third of Mozambique's population is urban, with other countries falling into this range. By the beginning of the twenty-first century, over 95% of whites and Indians in South Africa were living in urban areas but Black South Africans were less urbanized at 62% over all. Damage to family structures, unemployment, urban crowding, severely limited police capacity, and poor urban planning have often led to high levels of social insecurity and criminal violence throughout the region. Rural–urban migration and gender balance had stabilized by the end of the millennium, but high levels of mobility between city and country remained, with families and kin groups often attempting to maintain bases in both.

In Zambia, urbanization proceeded rapidly during the colonial period, with miners and work seekers clustering in the 'Copper belt' in the north-central Zambia. With collapsing copper prices during the 1980s and mismanagement of mines, many people returned to rural areas. Rural subsistence agricultural areas are characterized by dispersed residence, small-scale agriculture, and pastoralism, usually producing no significant economic surplus for trade.

Urban areas throughout the region are weakly centralized and sprawling. All large urban areas derive from colonial patterns of segregated 'suburbs' loosely grouped around multiple trading and industrial centers. Single houses (or complexes of houses, outbuildings, and shacks) surrounded by walled 'gardens' or yards comprise the formal urban settlement pattern in all of the countries' urban areas, while large informal shack settlements surround them.

The colonial economy based on mining, large-scale estate farming, and labor migration contributed to disruption of family life and systems of social control, especially in rural areas of Zambia, Zimbabwe, and South Africa. While mining and agriculture now account for less than 10% of the GDP in South Africa, these two sectors dominate the economies of other southern African countries as they did in South Africa in the early twentieth century. Lesotho and Malawi, lacking natural resources except water, have been largely labor-exporting countries, relying on remittances sent home by miners who work mainly in South Africa, but also in Zambia and Zimbabwe. The bulk of Botswana's GDP is derived from diamond mines, while Zambia relies on exports of copper, cobalt, and other minerals. Namibia is mostly arid, and depends on mineral exports, especially uranium, for foreign exchange earnings. All countries, however, participate in sending labor to the mines located throughout the region but dominated by South African mining houses.

At the beginning of the twenty-first century, economic and social prospects are dimmed by the HIV/AIDS pandemic. Southern Africa emerged as the hardest hit region in the world. The rate of population growth is predicted to fall below replacement rates in the first decade of the twenty-first century.

Language and Communication

The process of urbanization has resulted in the emergence of distinctive urban cultures with small manufacturing, trade, mining, and transport industries initially attracting large numbers of formerly rural people to urban areas, where different languages and cultures came into contact. They blended to form new urban-based languages and cultures. Examples include Isicamtu, 'talking' or tsotsitaal, 'gangster language,' in South Africa, based on Zulu grammar but blending English, Afrikaans, and Sotho or Twana words; or Fanagalo, meaning 'Do-like-this,' a pidgin based on Zulu used on the mines and farms throughout southern Africa. With the exception of the isolated Khoisan languages, all languages in the region, including English and Afrikaans, frequently grade into each other and borrow vocabulary and new linguistic forms from each other. New forms of music, dance, and religion blending African traditions and Western forms also arose.

No indigenous African language has achieved status or acceptance as a single national language or lingua franca anywhere in southern Africa. South Africa has 11 official languages, of which those with the largest number of first-language speakers are Zulu, Xhosa, Afrikaans, Pedi ('Northern Sotho'), and English, followed by, Sotho ('Southern Sotho') and Tswana (often 'Sotho-Tswana'), with smaller linguistic communities of Tsonga ('Shangaan'), Venda, Ndebele, and Swati ('Swazi'). Other languages spoken include Gujarati, Urdu, Hindi, Tamil, and Bhojpuri from India. Portuguese, Greek, and German are spoken by smaller communities from Europe. Chinese is spoken by recent immigrants from China and Taiwan, although the small community of Chinese who arrived in the nineteenth century speak only English. French is spoken primarily by Central African refugees and migrants in South Africa.

The linguistic diversity is even greater in Zambia, where 41 languages are spoken, including Bemba, Tonga, Nyanja (Chewa), and Lozi. In Mozambique, 33 languages are spoken, including Makuwa (Makua) in the north and Tsonga in the south. In Namibia 28 languages are recognized. Although over 70% of the population of Botswana speak Tswana, there are many small, nearly extinct languages spoken by Khoisan people, and by refugees from surrounding countries.

There is some mutual intelligibility among some of these languages. For instance, Sotho-Tswana languages (Pedi, the Sotho language of Lesotho, and Tswana) form a set of dialects, and the 'Nguni' languages (Zulu, Xhosa and Swazi) are similar and mostly interintelligible. Several languages have become lingua francas within limited areas of these countries – for instance Zulu in South Africa, or Bemba and Nyanja/Chewa in Zambia and Malawi. Almost all people speak several regional languages. As a consequence of this diversity, English has increasingly become the public language of government and the media even in the Portuguese-speaking countries. It is also the language of trade and diplomacy.

While Mozambicans use Portuguese as the language of government and education, they are increasingly turning to English, too, as they develop deeper political and economic links with other nations. Nevertheless, the large majority of all people in the region continue to rely on local Bantu languages at home. Afrikaans is indigenous to South Africa, where it evolved from dialects of Dutch in the Dutch colonial settlements at the Cape of Good Hope in South Africa. Today it is spoken throughout South Africa and Namibia, with small pockets in other southern African countries.

Within the region, South African society is the most complex with respect to color, race, religion, language, dialect, region, and class. Ethnic, political, and other kinds of identities – such as preference for certain sports or foods – are based on these differences, and also help to create and sustain them. After independence in Angola and Mozambique, many whites either went to Portugal or migrated to South Africa. Similarly, large white and Indian populations in Zimbabwe and Zambia also moved to South Africa, or left the region. These migrations greatly reduced ethnic and racial diversity in all of these countries.

Despite the previously enforced divisions and centuries of conflict across many different cultures and language groups, South Africa exhibits a degree of cultural, economic, and political unity that is perhaps unique in Africa. Virtually all of its people manage to communicate with one another in one or several common languages, shop in the same stores, watch the same television channels, furnish their houses in similar ways, and – ironically – share the belief that they are much more different from one another than they actually are in terms of many objective measures of cultural or social difference.

None of these primary indicators of social difference, however, tend to define a consistent or bounded category of people. In other words, people of many different colors or race speak the same language or set of languages, while people of the same region or class vary widely in what religion they follow, or in their political allegiances. People sharing the same religion and language – which is often a combination that defines a fairly closely knit ethnic or national community in Europe or Asia – may span widely different colors, classes, and regions in southern Africa. For instance, the majority of South Africa's 'colored' people speak Afrikaans and belong, like most white Afrikaners, to some denomination of the Reformed Church. Surprisingly, while virtually all schools and public places were rigidly divided by Apartheid policies in South Africa, by the early twenty-first century, schools, shopping centers, and other public places had quietly integrated with no more than a handful of violent or other racial incidents. Urban areas in all countries are marked by wealthy enclaves, often separated by walls and advanced security systems that are often surrounded by much poorer 'suburbs' and large shack and informal settlements of the very poor.

See also: Africa, Sociocultural Overviews: West Africa; Central Africa: Sociocultural Aspects; Colonialism, Anthropology of; Colonization and Colonialism, History of; East Africa: Sociocultural Aspects; Historical Thought and Historiography: Sub-Saharan Africa; Imperialism, History of.

Bibliography

Barnard, A., 1992. Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan People. Cambridge University Press, Cambridge.

Colson, E., Gluckman, M. (Eds.), 1951. Seven Tribes of British Central Africa. Manchester University Press, Manchester, UK.

Ehret, C., 1998. An African Classical Age: Eastern and Southern Africa in World History, 1000 B.C. to A.D. 400. University Press of Virginia, Charlottesville, North Carolina. James Currey, Oxford.

Freund, B., 1984. The Making of Contemporary Africa: The Development of African Society since 1800. Indiana University Press, Bloomington, IN.

Greenberg, J., 1966. The Languages of Africa. Indiana University Press, Bloomington, IN.

Hall, M., 1990. Farmers, Kings, and Traders: The People of Southern Africa, 200–1860. The University of Chicago Press, Chicago.

Hamilton, C., Mbenga, B.K., Ross, Robert (Eds.), 2010. The Cambridge History of South Africa. University of Cambridge Press, Cambridge.

Hammond-Tooke, W.D., 1974. The Bantu-Speaking Peoples of Southern Africa. Routledge, London.

Hammond-Tooke, W.D., 1993. The Roots of Black South Africa. Jonathan Ball Publishers, Johannesburg.

Heine, B., Nurse, D. (Eds.), 2011. A Linguistic Geography of Africa. Cambridge University Press, Cambridge.

Heine, B., Nurse, D. (Eds.), 2000. African Languages: An Introduction. Cambridge University Press, Cambridge.

Huffman, T.N., 2007. Handbook to the Iron Age: The Archaeology of Pre-colonial Farming Societies in Southern Africa. University of KwaZulu-Natal Press, Scottsville, South Africa.

Janzen, J., 1992. Ngoma: Discourses of Healing in Central and Southern Africa. University of California Press, Berkeley.

Landau, P.S., 2010. Popular Politics in the History of South Africa, 1400–1948. Cambridge University Press, Cambridge.

Relevant Websites

With the exception of South Africa, the Internet resources for southern African countries are quite limited. Many of the South African sites listed here also provide some information on other southern African countries. All of the sites listed below contain links to other resources and sites that provide much additional information.

These sites are provided by governments, and include government documents, legislation, tourism information, historical sketches, and other material:

http://www.gov.za – South Africa.

http://www.gov.na – Namibia.

www.gov.bw – Botswana.

www.gov.ls – Lesotho.

www.gov.sz – Swaziland.

www.zim.gov.zw – Zimbabwe.

Zambia.co.zm – Zambia.

Malawi.gov.mw – Malawi.

http://www.sahistory.org.za – The South African History Online provides extensive historical materials and resources for teachers and younger students, including videos, speeches, original documents, government documents, and more. Short descriptions of other southern African countries are also available.

http://www.saha.org.za – The South African History Archive contains mainly twentieth-century political and government documents on human rights, justice, and social history for South Africa.

http://www.iziko.org.za – Information on South African museums, many with extensive digital collections. See also the extensive digital archive of southern African rock art dating from 20 000 years before present to recent centuries at: http://www.sarada.co.za.

http://www.google.com/culturalinstitute/home?view¼fullscreen&hl¼en – Google Cultural Institute provides a portal for some materials on southern African art, culture, history, and geography.

http://www.aluka.org – Aluka, African Cultural Heritage Sites and Landscapes Collection, part of JSTOR (www.jstor.org, a non-profit shared digital library), is an archive of visual, contextual, and spatial documentation of African heritage sites, including southern African countries, especially South Africa and Mozambique including the 'Arquivo Histórico de Moçambique.'

Space and Social Theory in Geography

Barney Warf, University of Kansas, Lawrence, KS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Social theory and the analysis of spatiality have a long and rich history. Building on earlier traditions such as Marxism, feminism, and humanistic geography, this article addresses the major lines of conceptual thought in human geography in the early twenty-first century. It encompasses the discipline's major theoretical constructs under the broad umbrella of social constructivism and political economy, largely based on the work of Foucault. These precepts serve as a guide to the role and impacts of various forms of social theory in the analysis of multiple topics, including cultural geography, critical political geography, economic geography, the geographies of cyberspace, critical cartography, and GIS, and the social construction of nature.

Over the last four decades, contemporary human geography has embraced social theory with a vengeance. Building on several earlier schools of thought that emerged in the wake of positivism, including various forms of Marxism, feminism, structuration theory, and postmodernism, the intellectual landscape of human geography in the twenty-first century resembles a palimpsest consisting of multiple, overlapping, and intersecting layers. There is, to be sure, enormous variation among and within each of these lines of inquiry, as well as many apt summaries (e.g., Gregory, 1994; Benko and Strohmayer, 1997; Peet, 1998; Warf, 2004). Many, for example, draw on the enormously influential works of Michel Foucault (Crampton and Elden, 2007), which decisively sutured together power, knowledge, and space.

Given the diversity of theoretical views within the discipline today, it would be foolish to impose an artificial homogeneity that claims to represent all such perspectives. However, if any theme may be said to run throughout the diverse perspectives that characterize the field, it is social constructivism. Many entities once taken as given – as somehow lacking in social origins and consequences – have been revealed as the products of social relations embedded in particular times and spaces. The body, gender, ethnicity, poverty, even nature itself, all have been revealed as the contingent, power-laden products of configurations of class, power, discourse, and other dimensions of social life. Social constructivism points to the origins of phenomena within circuits of power and knowledge, the interests that are served (or not) by particular discourses, their effects and consequences, and the contingent ways in which they unfold under different historical and geographical circumstances. Constructivist perspectives have played out in different ways in various domains of human geography, including the cultural turn in economic geography, the poststructuralist emphasis on networks, critical geopolitics, critical cartography and participatory GIS, political ecology and the social construction of nature, and the growing interest in hybridity. One result of geography's intellectual sophistication is that a field long characterized by insecurity about its position within the academic division of labor has won newfound respect. Indeed, across a remarkably diverse range of disciplines, ranging from the humanities to the natural sciences, space, place, and spatiality have decisively inserted themselves into contemporary forms of knowledge production, a phenomenon that has come to be widely seen as the spatial turn (Warf and Arias, 2008).

Much of human geography has been deeply affected by poststructuralism (Harrison, 2006; Murdoch, 2006), a loosely related array of perspectives that reflect social constructivism's intersections with critical philosophy and literary analysis. Like an earlier generation of postmodernists, poststructuralism emphasizes the plurality of social worlds in which people live, an abiding concern for otherness and difference (and the power relations that shape it), and a deep empathy for the socially and politically marginalized. It recast once-given universal categories (e.g., 'women,' 'African-Americans') in the plural, emphasizing the diversity within as well as among groups. In addition to a rejuvenated politics of identity, which informs much of critical political and cultural geography, poststructuralism opened the door for analytical approaches such as actor-network theory (Bosco, 2006; Latour, 2007). Moreover, the rejection of all dualities – nature/society, global/local, black/white, economic/cultural, even human/nonhuman – as artificial and misleading remnants of modernist thought (Cloke and Johnston, 2005) gave rise to an insistent focus on hybridity as the antidote to the essentialisms that such categories create (Whatmore, 1997, 2001; Kwan, 2004). For example, in following Latour and transcending the human/nonhuman dichotomy, animal geographies have become adept at exploring the interface between humans and nonhuman animals (Urbanik, 2012).

In rejecting fixed categories, essentialisms, and dualisms, emphasizing the plural, human geography came to theorize all relations, both social and spatial, in relational terms. Massey's (2005) notion of power geometries played a key role in this process, in which networks became the defining metaphor of geographic interaction. Those drawing on the Deleuze-Guattarian 'flat ontology' have come to reject scale itself as an unnecessary obfuscation (Marston et al., 2005), stimulating a lively debate about the meaning and relevance of scale in a relational world (Paasi, 2004; Jones, 2009; MacKinnon, 2010). The emphasis on geographies as fluid and ever-changing also led to renewed emphasis on movement and mobility (Cresswell, 2006; Larsen et al., 2006), in which place has been conceptually repositioned as a process rather than a thing.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72053-8

Cultural Geography

Now one of the discipline's most theoretically sophisticated fields, cultural geography, exhibits numerous lines of theoretical inquiry that address the role of culture at multiple spatial scales, ranging from the social construction of the body to geographies of sexuality and ethnicity to the lasting impacts of the colonial imaginary.

Bodies, once taken for granted as given and now firmly embedded in their cultural milieu, have become instrumental to works on the spatiality of emotions and affect (Thrift, 2004, 2008; Thien, 2005; Tolia-Kelly, 2006; Lorimer, 2008; Anderson and Harrison, 2010). Nonrepresentational theory, which has been highly contested, merges long-standing interest in geographies as embodied social processes with poststructuralist interrogations of language to portray human beings as more than discursive products but also as deeply shaped by forces that have long been marginalized by the Enlightenment obsession with rationality, as manifested in the long-standing reign of the Cartesian cogito. Emotions, for example, once dismissed as atavistic or insignificant, have assumed pride of place in contemporary theorizations of the subject and subjectivity (Pile, 2010). Likewise, following Judith Butler, social roles once assumed to exist externally to the individual and acquired through socialization are now largely viewed through the lens of performativity (Nash, 2000).

Human geography's embodiment of social and spatial relations has deepened its engagement with sexuality (Howell, 2007). Drawing heavily on Foucault, this focus has centered largely on sexual minorities, including the construction of queer spaces at multiple spatial scales ranging from the body to the gay rights movement to global patterns of discrimination (Brown and Knopp, 2003; Oswin, 2008). Other variants have delved into the spatiality of masculinities (Berg and Longhurst, 2003), largely displacing earlier notions of gender that equated it solely with women and femininity, as well as sex workers and the intersections of sexuality, class, gender, ethnicity, and location.

Similarly, ethnicity, which has decisively displaced older notions of 'race,' has acquired new attention (Phillips, 2007). By contextualizing race as an ideology – as a socially constructed relationship, not a biologically determined 'given,' this literature has illustrated the complex ways in which race is constructed, experienced, and negotiated in everyday existence, how it intersects with class and gender, and the manner in which racial relationships of domination and subordination play out over space and time. Like gender, ethnic identity is fundamentally a power relationship that typically reflects, reinforces, and naturalizes existing lines of inequality, although it may also challenge hegemonic norm relationship. Race may be illusory in that it is socially created, but, as Peake and Schein note (2000), its effects are nonetheless quite 'real': denaturalizing race thus unmasks the power relationships that construct it socially and discursively. Because racial and ethnic identities are important in social reproduction, individual and collective identity, and the allocation of resources over space and time, race and racism are simultaneously material and discursive processes that reflect and constitute social relationships and places. The geography of whiteness reveals that it too is socially constructed in ways that render it invisible, as the 'default option' against which 'ethnic' identities are assessed (Dwyer and Jones, 2000). The 'naturalness' of whiteness – the assumption that it is not an ethnicity, hence not constructed, a position that equates ethnicity with nonwhites – is a large part of the power of institutionalized racism. Pulido (2000) points out that white privilege takes many forms and that its existence is typically rendered invisible. Racism can be intentional or unintentional, although, as she notes, "by reducing racism to a hostile, discriminatory act, many researchers…miss the role of structural and hegemonic forms of racism in contributing to such inequalities" (2000, 12).

Cultural geography's examination of colonialism, both past and present, as a discursive and ideological force has had significant implications for the analysis of globalization, including the role of indigenous peoples (Nash, 2002; Clayton, 2003; Sharp, 2009). Inspired by Edward Said (1978), postcolonialism in many respects represents a fusion of political and cultural geography, one that grapples with the powerful material and discursive effects of Western hegemony. Gregory's (2004) fascination with the Middle East, particularly the injustices visited upon Muslim peoples by Western powers, has been central to this project. Postcolonialism continues to inform efforts to understand differential trajectories of regional development and the goals and strategies that inspire them (Radcliffe, 2005).

Critical Political Geography

Inspired by the critical geopolitics of the 1990s (especially Ó Tuathail, 1996), political geography has undergone a sustained engagement with issues of discourse, culture, and identity, including feminist thought (Hyndman, 2004). Following Foucault, power is conceived largely in terms of its linkages to knowledge and ideology, funneled through numerous channels, simultaneously discursive and material, contingently exercised, and invariably resisted. Its effects are seen throughout the rhythms of daily life and social reproduction, including the media, educational systems, and the geopolitics of the body. Issues of immigration and citizenship loom large in this respect (Staeheli, 2011). As a result, popular culture has also received pride of place in this school, including comics and religious fundamentalism (Dittmer, 2005, 2008, 2012) and the broader phenomenon of banal geopolitics (Leib, 2011).

In the same vein, drawing on the enormously influential work of Edward Said (1978), orientalism has been the topic of repeated inquiries in political geography, including its internal variants (Jansson, 2003; Johnson and Coleman, 2012). Many geographers drew on the work of Giorgio Agamben (2005) and his thesis of the 'state of exception' to examine the geopolitics of the war on terror, including, for example, the Guantanamo prison in Cuba (Gregory, 2006).

Globalization has had important repercussions for political geography, including the continued viability of the nation-state and the rapidly changing significance of borders (Diener and Hagen, 2009; Popescu, 2011). Viewed through the telescope of social constructivism, territory too is a product of political relations (Forsberg, 2003). Finally, as environmental issues have grown in severity across the planet, environmental geopolitics has gained mounting significance (O'Lear, 2010),

including struggles over water and the paramount issue of anthropogenic climate change and its effects (Raleigh and Urdal, 2007).

Economic Geography

Economic geography's embrace of social theory has led to widespread changes in how the spatiality of economic activity has been conceived and theorized. Perhaps most significant has been the erasure of the long-standing dichotomy between 'society' and 'economy,' following the pioneering efforts of Granovetter (1985). An important step in this process was Florida's (2004) celebrated and enormously influential notion of the creative class. In this reading, the skilled professionals so central to cognitive-cultural urban economies place a high value on amenities, recreational opportunities, and local social climates that value diversity, including ethnic differences and tolerance of alternative lifestyles (notably gays). Cities that are tolerant, Florida maintains, are economically healthy; those that are restrictive and xenophobic are unlikely to attract the talent they need to remain competitive. Critics of Florida (e.g., Scott, 2006) argue that the creative class thesis underemphasizes the locational dynamic of urban agglomeration and the intense pressure on firms to maximize profits. Nonetheless, Florida's argument has won wide acclaim.

Belatedly, economic geographers came to embrace culture (Barnes, 2001; Martin and Sunley, 2001; Ettlinger, 2003; Gibson and Kong, 2005; Yeung, 2005). In part, economic geography's long-standing detachment from culture reflected its earlier infatuation with neoclassical economics, which views the economy as a black box utterly detached from the rest of society, as if it were devoid of social, political, economic (and spatial) origins and consequences. Economic and cultural geography – two major parts of a discipline renowned for the diversity of its theoretical views – suffered for decades in relative isolation of one another, as if the economy and culture had little to do with one another. Fortunately, what have long been distinct subdisciplines, each with its own body of theory, vocabulary, methodology, and topics of inquiry, recently have increasingly blended into one another, an intersection fostered by their mutual interest in issues of social relations, power, and location. Barnes (2005) explicitly seeks to dissolve the boundaries between the cultural and the economic by replacing them with a hybrid, one that offers the possibility of escaping traditional dualisms. Thus, as he puts it (p. 73), "The task of a hybrid economic geography should not be simply to reverse the binary and make culture primary, and the economy secondary. Rather, it should be to dissolve the binary altogether." As a result, economic geographers came to emphasize culture as a complex set of relations every bit as important as putatively 'economic' factors in the structuring of economic landscapes.

Another arena in which the inseparability of the economic and the cultural is made profoundly evident is the geography of consumption. When rescued from neoclassical economics, consumption offers an ideal opportunity to view the intersections of social relations, culture and ideology, and space. Not surprisingly, a fecund geographical literature on consumption and shopping has sprouted (Mansvelt, 2005; Leslie, 2009; Goodman et al., 2010), much of it pointing to how it shapes consumers and spaces. Geographies of food play an especially important role in this regard (Wrigley, 2002). In particular when linked to commodity chains, consumption offers an ideal pedagogic vehicle for exploring the politics of everyday life, identity, and the body (Hartwick, 2000).

A major concern of theorizations of economic landscape is the formation and role of clusters of economic activity (Vorley, 2008; Martin and Sunley, 2003; Cumbers and MacKinnon, 2004; Benneworth and Henry, 2004). Clusters represent the age-old theme of agglomeration economies in yet another reincarnation. Inspired by the turn to post-Fordism and attending rounds of vertical disintegration, clusters represent a variety of concentrated economic activity ranging from small, localized groups of manufacturing firms to enormous pools of producer services in global cities. Much of the search for the keys to their success hinged on the role of face-to-face contacts and the exchange of tacit knowledge (Storper and Venables, 2004). The analysis of clusters led economic geographers to emphasize more qualitative, contextual approaches to competitiveness, including the seminal role played by knowledge spillovers and positive externalities in the formation of 'learning regions.'

Like cultural and political geographers, economic geography continues to grapple with the multitude of issues and processes grouped under the term globalization. Dicken's (2011) architectonic volume *Global Shift*, which remains, after several decades and multiple editions, still the discipline's foremost summary of the topic. By now, the interpenetration of the local and the global (or 'glocalization') has become common place. Following the move to view places as networks, this literature decisively portrayed regional development in terms of globalized commodity chains, in which spatiality is unevenly and contingently produced through an intertwined series of both local and worldwide linkages (Amin, 2002; Sheppard, 2002; Coe et al., 2004).

Finally, economic, political, and cultural geographers have all wrestled with the thorny question of neoliberalism and its massive impacts around the world. David Harvey (2005a,b, 2006, 2011), still the doyen of Marxist theorization, has worked assiduously to unpack the dynamics of neoliberalism as both an economic and political project, one that reflects capitalism's incessant drive to recreate landscapes and reinvent itself periodically through the generation of crises, including the worldwide Great Recession that began in 2007.

Geographies of Cyberspace

Digital technologies have been central not only to the reconstruction of global capitalism over the last four decades, but also increasingly to the rhythms and contours of daily life. In December 2011, more than 2.25 billion people used the Internet, making it a tool of communications, entertainment, and other applications accessed by roughly 32% of the world's population (www.internetworldstats.com/stats.htm). For most users, these uses extend well beyond e-mail, the most common Internet use, to include bill payments and electronic banking; job and housing searches; stock trading; 'e-tail' shopping; searching for health information and news; online classes;

digital gambling; online video games; Internet telephony; hotel and airline reservations; chat rooms; electronic tax payments; downloading television programs, movies, digital music, and pornography; and popular sites and services such as YouTube, Facebook, and Google. In all these ways, and more, cyberspace offers profound real and potential effects on social relations, everyday life, culture, politics, and other social activities. Indeed, for rapidly rising numbers of people around the world, the 'real' and the virtual have become thoroughly interpenetrated. In this light, access to cyberspace is no longer a luxury, but a necessity. As its applications have multiplied, the Internet is having enormous impacts across the globe, including interpersonal interactions and everyday life, identity formation, retail trade and commerce, governance, and is affecting the structure and form of cities, in the process generating round upon round of non-Euclidean geometries in the context of a massive global wave of time–space compression.

By now, digital reality and everyday life for hundreds of millions of people have become so thoroughly fused that it is difficult, if not impossible, to disentangle them. In this context, simple dichotomies such as 'off-line' and 'online' fail to do justice to the diverse ways in which the 'real' and virtual worlds for hundreds of millions are interpenetrated. Yet for many others – the familiar litany of the poor, the undereducated, the ethnic minorities, and the socially marginalized – the Internet remains a distant, ambiguous world. Denied regular access to cyberspace by the inability to purchase a personal computer, the technical skills necessary to log on, or public policies that assume their needs will be magically addressed by the market, the information have-nots living in the economically advanced world are deprived of many of the essential skills necessary for a successful or convenient life (Crang et al., 2006). While those with regular and reliable access to the Internet drown in a surplus of information – much of it superfluous, irrelevant, or unnecessary – those with limited access have difficulty comprehending the opportunities it offers, the savings in time and money it allows, and the sheer convenience, entertainment value, and ability to acquire data from bus schedules to recipes to global news.

For many social media users, the self is increasingly constructed through their online interactions. In contrast to the prevailing Western model of subjectivity of the autonomous individual, digital media allow for a new form of sociability that greatly enhances the importance of relational ties, leading, arguably, to a different form of the self than the historically accepted view of the self-contained, autonomous individual whose conception largely minimizes social interactions. The argument hinges on the notion of the networked self, beings so immersed in digital social media that their online identities, personas, and reputations are as central to their lives as their off-line counterparts. Networked selves tend to live at the center of webs of digital ties that bind space in complex and often unpredictable ways, so that conventional geographies of everyday life have limited relevance to understanding their outlook and behavior. Following Adams (2005), the networked self in many respects is 'boundless,' that is, able to escape the physical and temporal limits of the human body. Indeed, following Haraway's famous cyborgian model of the self, theorizations of cyberspace have plunged into how computer software and hardware generate nonhuman agencies in which the digital and the real are seamlessly fused (Thrift and French, 2002; Dodge and Kitchin, 2005a,b, 2007).

Critical Cartography and GIS

One symptom of social constructivism's success in human geography is its penetration into the long-asocial worlds of cartography and GIS. This undertaking, which began in earnest in the 1990s, was closely associated with a shift in understanding in the nature of spatial representations. Traditional cartography and GIS relied heavily on a strict separation of observer and observed, sign and referent, facts and values, the representation is mirror of objective world, the correspondence theory of truth, and, not coincidentally Cartesian notions of space as given, not produced. In this reading, there is a seamless transition from the map creator's intent to the product to the reader/viewer's understanding, and representation is only technical problem in which interpretation is unproblematic. This perspective has been heavily criticized for denying the existence of an observer, the social context in which the production of spatial knowledge takes place, and creating a false sense of objectivity.

Language and representation were, of course, one of the great problematics of twentieth century social thought, with deep roots in philosophy and linguistics. Under the impetus of postmodernism (e.g., Derrida), traditional views began to unravel as both the creation of knowledge and its interpretation were increasingly resituated in social terms. Under the impetus of Foucauldian thought, the notion of discourse came to summarize the intimate connections between knowledge and power and how worldviews shape subjects by simultaneously enabling and constraining their understanding of the world in historically specific ways. Critical cartographers such as Denis Wood (1992, 2010) initiated a thorough-going deconstruction of maps as discourses, i.e., not simply as things but as embodiments of social relations. Such a move called attention not only to the social contexts and power interests that lay behind the creation of maps, but their effects and consequences, notably the naturalization of spatial relations. Maps, like all discourses, exist in a simultaneously determinant relationship with their contexts, helping to create the world that brought them into being.

Similarly, social constructivism decisively invaded the politically antiseptic world of GIS. Critics asserted that the subdiscipline that revolved around the technology had long attempted to act as a Trojan Horse to resmuggle empiricism back into the discipline. Naïve views could lead to digital panopticons and geoslavery (Dobson and Fisher, 2007). Inspired by the work of Pickles (1995), this process exposed GIS as a socially embedded technology (Sheppard, 2005; Schurrman, 2000, 2006), a shift with far-reaching consequences (Warf and Sui, 2010). Far from constituting an ostensibly objective view of reality, GIS was been discursively repositioned as a particular mode of geographing that privileges the quantitative over the qualitative, the observable over the nonobservable, form over process, and site characteristics over relational situation ones. As a result, the recent evolution of GIS has unveiled a fantastic zoo of new species, including, to take but a few examples, critical GIS (Sheppard, 2005b), participatory GIS (Elwood, 2006), volunteered geography (Goodchild, 2007), affective/emotional GIS

(Aitken and Craine, 2009), qualitative GIS (Cope and Elwood, 2009), feminist GIS (Kwan, 2002; McLafferty, 2005), queer GIS (Brown and Knopp, 2008), digiplaces (Zook and Graham, 2007a,b), GIS invoking indigenous knowledge (Robbins, 2003), and humanistic GIScience (Sui, 2004; Bodenhamer et al., 2010). In many ways, these developments reflect a highly productive engagement between the GIS community and social theorists and humanities scholars over the last decade, which resulted in studied attempts at 'pixelizing the social and socializing the pixel.'

The Social Construction of Nature

One largely unintended consequence of social theory's penetration of different segments of the discipline of geography has been the initiation of a gradual rapprochement between its human and physical branches, helping to heal a long-standing schism that plagued the field for a century or more (Hanson, 1999; Harden, 2012). In rejecting the traditional duality of 'nature' and 'culture,' the intersection of social theory and environmental analysis has given birth to a view often labeled the social construction of nature, a view that takes as its point of departure the notion that 'nature' does not exist independently of human activities nor can it be understood outside of social frames of reference (Castree and Braun, 2001; Castree, 2005; Bridge, 2011). Thus nature and society recursively shape one another both materially and discursively, so that attempts to differentiate them are doomed. The production of human geographies is inevitably also the production of nature. This process unfolds at multiple spatial scales, from the molecular (e.g., genetic engineering of crops) to the global (e.g., anthropogenic climate change).

Such a perspective injects a badly needed dose of political relevance into the analysis of environmental issues. Conflicts long seen as 'people versus nature,' such as deforestation, are repositioned as struggles among social groups (particularly classes), and intimately linked to the global and regional dynamics of regional inequality. Likewise, 'natural' hazards are not seen as natural, but closely tied to the social processes that generate differential vulnerability and susceptibility to floods, earthquakes, and other disasters, almost invariably along class, gender, and racial lines. Mounting concerns over environmental justice have politicized the analysis of health-related issues such as air quality and exposure to toxic wastes (Pulido, 2000). Finally, as anthropogenic climate change becomes increasingly visible as one of the world's most pressing social and environmental issues, human geographers have explored the political dynamics swirling around this issue (Demeritt, 2001; Wainwright, 2010).

Concluding Thoughts

Given the enormous heterogeneity of topics and subjects included within the broad domain of social theory and space, are there significant themes and commonalities that underpin this body of work? Human geography's embrace of social theory, largely originating with Foucault and encompassing various forms of social constructivism, has highly politicized the field, giving it great sensitivity to issues of power and social justice, injected a deep sense of morality. Much of the field self-consciously identifies with the socially and politically marginalized, including long-standing concerns regarding class, ethnicity, and gender as well as relatively newer sympathies for sexual minorities, indigenous peoples, the handicapped, animals, and others. Social constructivism has spilled beyond the confines of traditionally 'human' topics to inject these concerns into cartography, the analysis of GIS, and the social construction of nature. Given the unrelenting assault by neoliberalism in many arenas across the planet, these issues are decidedly nontrivial. One response has been the rise of geographies of empathy, caring, and responsibility (Lawson, 2007), including the adoption of cosmopolitan ideologies (Harvey, 2009; Warf, 2012). If these strategies make human geography more humane, then so much the better.

See also: Critical Realism in Geography; Cultural Geography; Economic Geography; Geographic Information Systems; Geographic Thought, History of; Geopolitics; Globalization: Geographical Aspects; Information Society, Geography of; Marxist Geography; Nature and Society in Geography; Political Geography; Postcolonial Geography; Race and Racism, Geography of; Scale in Geography; Sexuality and Space; Social Geography; Spatial Thinking in the Social Sciences; Urban Geography.

Bibliography

Adams, P., 2005. The Boundless Self: Communication in Physical and Virtual Spaces. Syracuse University Press, Syracuse.
Agamben, G., 2005. State of Exception. University of Chicago Press, Chicago.
Aitken, S., Craine, J., 2009. Affective visual geographies and GIScience. In: Cope, M., Elwood, S. (Eds.), Qualitative GIS: A Mixed Methods Approach. Sage, London, pp. 139–155.
Amin, A., 2002. Spatialities of globalisation. Environment and Planning A 34, 385–399.
Anderson, B., Harrison, P. (Eds.), 2010. Taking-place: Non-representational Theories and Geography. Ashgate, Farnham, UK.
Barnes, T., 2001. Retheorizing economic geography: from the quantitative revolution to the cultural turn. Annals of the Association of American Geographers 91, 546–565.
Barnes, T., 2005. Culture: economy. In: Cloke, P., Johnston, R. (Eds.), Spaces of Geographical Thought: Deconstructing Human Geography's Binaries. Sage, London, pp. 61–80.
Benko, G., Strohmayer, U. (Eds.), 1997. Space and Social Theory: Interpreting Modernity and Postmodernity. Wiley-Blackwell, New York.
Berg, L., Longhurst, R., 2003. Placing masculinities and geography. Gender. Place and Culture 10, 351–360.
Bodenhamer, D., Corrigan, J., Harris, T., 2010. The Spatial Humanities: GIS and the Future of Humanities Scholarship. Indiana University Press, Indianopolis.
Bosco, F., 2006. Actor-network theory, networks, and relational approaches in human geography. In: Aitken, S., Valentine, G. (Eds.), Approaches to Human Geography. Sage, London.
Bridge, G., 2011. The economy of nature: from political ecology to the social construction of nature. In: Leyshon, A., Lee, R., McDowell, L., Sunley, P. (Eds.), The Sage Handbook of Economic Geography. Sage, Thousand Oaks, CA, pp. 217–230.
Brown, M., Knopp, L., 2003. Queer cultural geographies – we're here! We're queer! We're over there, too! In: Anderson, K., Domosh, M., Pile, S., Thrift, N. (Eds.), Handbook of Cultural Geography. Sage, London.
Brown, M., Knopp, L., 2008. Queering the map: the productive tensions of colliding epistemologies. Annals of the Association of American Geographers 98, 40–58.
Castree, N., 2005. Nature. Routledge, London.
Castree, N., Braun, B., 2001. Socionature: Theory, Practice, and Politics. Wiley-Blackwell, Oxford and New York.

Clayton, D., 2003. Critical imperial and colonial geographies. In: Anderson, K., Domosh, M., Pile, S., Thrift, N. (Eds.), Handbook of Cultural Geography. Sage, London.
Cloke, P., Johnston, R., 2005. Spaces of Geographical Thought. Deconstructing Human Geography's Binaries. Sage, Thousand Oaks, CA.
Coe, N., Hess, M., Yeung, H., Dicken, P., Henderson, J., 2004. 'Globalizing' regional development: a global production networks perspective. Transactions of the Institute of British Geographers 29, 468–484.
Cope, M., Elwood, S., 2009. Qualitative GIS: A Mixed Methods Approach. Sage, London.
Crampton, J., Elden, S. (Eds.), 2007. Space, Knowledge and Power: Foucault and Geography. Ashgate, London.
Crang, M., Crosbie, T., Graham, S., 2006. Variable geometries of connection: urban digital divides and the uses of information technology. Urban Studies 43 (13), 2551–2570.
Cresswell, T., 2006. On the Move: Mobility in the Modern Western World. Routledge, New York.
Cumbers, A., MacKinnon, D., 2004. Clusters in urban and regional development. Urban Studies 41 (5/6) 959–969.
Demeritt, 2001. The construction of global warming and the politics of science. Annals of the Association of American Geographers 91 (2), 307–337.
Dicken, P., 2011. Global Shift: The Internationalization of Economic Activity, sixth ed. Guilford Press, New York.
Diener, A., Hagen, J., 2009. Theorizing borders in a 'borderless world': globalization, territory and identity. Geography Compass 3 (3), 1196–1216.
Dittmer, J., 2005. Captain America's empire: reflections on identity, popular culture, and post-9/11 geopolitics. Annals of the Association of American Geographers 95 (3), 626–643.
Dittmer, J., 2008. The geographical pivot of (the end of) history: evangelical geopolitical imaginations and audience interpretation of left behind. Political Geography 27 (3), 280–300.
Dittmer, J., 2012. Captain America and the Nationalist Superhero: Metaphors, Narratives, and Geopolitics. Temple University Press, Philadelphia.
Dobson, J., Fisher, P., 2007. The panopticon's changing geography. Geographical Review 97 (3), 307–323.
Dodge, M., Kitchin, R., 2005a. Code and the transduction of space. Annals of the Association of American Geographers 95, 162–180.
Dodge, M., Kitchin, R., 2005b. Codes of life: identification codes and the machine-readable world. Environment and Planning D: Society and Space 23, 851–881.
Dodge, M., Kitchin, R., 2007. The automatic management of drivers and driving spaces. Geoforum 38, 264–275.
Dwyer, O., Jones III, J.P., 2000. White socio-spatial epistemology. Social and Cultural Geography 1 (1–2), 209–222.
Elwood, S., 2006. Critical issues in participatory GIS: deconstructions, reconstructions, and new research directions. Transactions in GIS 10, 693–708.
Ettlinger, N., 2003. Cultural economic geography and a relational and microspace approach to trusts, rationalities, networks and change in collaborative workplaces. Journal of Economic Geography 3, 145–171.
Florida, R., 2004. The Rise of the Creative Class. Basic Books, New York.
Forsberg, T., 2003. The ground without foundation: territory as a social construct. Geopolitics 8 (2), 7–24.
Foucault, M., 1974. The Order of Things: An Archaeology of the Human Sciences. Routledge, London.
Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Allen Lane, London.
Foucault, M., 1989. The Archaeology of Knowledge. Routledge, London.
Gibson, C., Kong, L., 2005. Cultural economy, a critical review. Progress in Human Geography 29 (5), 541–561.
Goodchild, M., 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. International Journal of Spatial Data Infrastructures Research 2, 24–32.
Goodman, M., Goodman, D., Redclift, M. (Eds.), 2010. Consuming Space. Placing Consumption in Perspective. Ashgate, Farnham and Burlington.
Granovetter, M., 1985. Economic action and social structure: the problem of embeddedness. American Journal of Sociology 91, 481–510.
Gregory, D., 1994. Geographical Imaginations. UK: Blackwell, Oxford.
Gregory, D., 2004. The Colonial Present: Afghanistan, Palestine, Iraq. Wiley-Blackwell, New York.
Gregory, D., 2006. The black flag: guantánamo bay and the space of exception. Geografiska Annaler 88 (4), 405–427.
Hanson, S., 1999. Isms and schisms: healing the rift between the nature-society and space-society traditions in human geography. Annals of the Association of American Geographers 89 (1), 133–143.
Harden, C., 2012. Framing and reframing questions of human–environment interactions. Annals of the Association of American Geographers 102 (4), 737–747.
Harley, J., 1989. Deconstructing the map. Cartographica 26, 1–20.
Harrison, P., 2006. Poststructuralist theories. In: Aitken, S., Valentine, G. (Eds.), Approaches to Human Geography. Sage, London.
Hartwick, E., 2000. Towards a geographical politics of consumption. Environment and Planning A 32, 1177–1192.
Harvey, D., 2005a. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Harvey, D., 2005b. The New Imperialism. Oxford University Press, Oxford.
Harvey, D., 2006. Neoliberalism as creative destruction. Geografiska Annaler 88B, 145–158.
Harvey, D., 2009. Cosmopolitanism and the Geographies of Freedom. Columbia University Press, New York.
Harvey, D., 2011. The Enigma of Capital and the Crises of Capitalism. Oxford University Press, Oxford.
Howell, P., 2007. Foucault, sexuality, geography. In: Crampton, J., Elden, S. (Eds.), Space, Knowledge and Power: Foucault and Geography. Ashgate, Aldershot, pp. 291–316.
Hyndman, J., 2004. Mind the gap: bridging feminist and political geography through geopolitics. Political Geography 23 (3), 307–322.
Jansson, D., 2003. Internal orientalism in America: W.J. Cash's the mind of the South and the spatial construction of American national identity. Political Geography 22 (3), 293–316.
Johnson, C., Coleman, A., 2012. The internal other: exploring the dialectical relationship between regional exclusion and the construction of national identity. Annals of the Association of American Geographers 102 (4), 863–880.
Jones, M., 2009. Phase space: geography, relational thinking, and beyond. Progress in Human Geography 33 (4), 487–506.
Kwan, M., 2002. Feminist visualization: re-envisioning GIS as a method in feminist geographic research. Annals of the Association of American Geographers 92, 645–661.
Kwan, M., 2004. Beyond difference: from canonical geography to hybrid geographies. Annals of the Association of American Geographers 94 (4), 756–763.
Larsen, J., Urry, J., Axhausen, K., 2006. Mobilities, Networks, Geographies. Ashgate, London.
Latour, B., 2007. Reassembling the Social: An Introduction to Actor-network Theory. Oxford University Press, Oxford.
Lawson, V., 2007. Geographies of care and responsibility. Annals of the Association of American Geographers 97 (1), 1–11.
Leib, J., 2011. Identity, banal nationalism, contestation, and North American license plates. Geographical Review 101 (1), 37–52.
Leslie, D., 2009. Consumption. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, Oxford, pp. 268–274.
Lorimer, H., 2008. Cultural geography: nonrepresentational conditions and concerns. Progress in Human Geography 32 (4), 551–559.
MacKinnon, D., 2010. Reconstructing scale: towards a new scalar politics. Progress in Human Geography 35 (1), 21–36.
Mansvelt, J., 2005. Geographies of Consumption. Sage, London.
Marston, S., Jones, J.P., Woodward, K., 2005. Human geography without scale. Transactions of the Institute of British Geographers 30, 416–432.
Martin, R., Sunley, P., 2001. Rethinking the "economic" in economic geography: broadening our vision or losing our focus? Antipode 33 (2), 148–161.
Martin, R., Sunley, P., 2003. Deconstructing clusters: Chaotic concept or policy panacea? Journal of Economic Geography 3, 5–35.
Massey, D., 2005. For Space. Sage, London.
McLafferty, S., 2005. Women and GIS: geospatial technologies and feminist geographies. Cartographica: International Journal for Geographic Information and Geovisualization 40, 37–45.
Murdoch, J., 2006. Post-structuralist Geography: A Guide to Relational Space. Sage, London.
Nash, C., 2002. Cultural geography: postcolonial cultural geographies. Progress in Human Geography 26 (2), 219–230.
Nash, H., 2000. Performativity in practice: some recent work in cultural geography. Progress in Human Geography 24 (4), 653–664.
Ó Tuathail, G., 1996. Critical Geopolitics. University of Minnesota Press, Minneapolis.
O'Lear, S., 2010. Environmental Politics: Scale and Power. Cambridge University Press, Cambridge.
Oswin, N., 2008. Critical geographies and the uses of sexuality: deconstructing queer space. Progress in Human Geography 32 (1), 89–103.
Paasi, A., 2004. Place and region: looking through the prism of scale. Progress in Human Geography 28, 536–546.
Peake, L., Schein, R., 2000. Racing geography into the new millennium: Studies of "race" and North American geographies. Social and Cultural Geography 1 (2), 133–142.

Peet, R., 1998. Modern Geographical Thought. Blackwell, Oxford.
Phillips, D., 2007. Ethnic and racial segregation: a critical perspective. Geography Compass 1 (5), 1138–1159.
Pickles, J., 1995. Ground Truth: The Social Implications of Geographic Information Systems. Guilford Press, New York.
Pile, S., 2010. Emotions and affect in recent human geography. Transactions of the Institute of British Geographers 35 (1), 5–20.
Popescu, G., 2011. Bordering and Ordering the Twenty-first Century: Understanding Borders. Rowman and Littlefield, Latham, MD.
Pulido, L., 2000. Rethinking environmental racism: White privilege and urban development in Southern California. Annals of the Association of American Geographers 90 (1), 12–40.
Radcliffe, S., 2005. Development and geography: towards a postcolonial development geography? Progress in Human Geography 29 (3), 291–298.
Raleigh, C., Urdal, H., 2007. Climate change, environmental degradation and armed conflict. Political Geography 26, 674–694.
Robbins, P., 2003. Beyond ground truth: GIS and the environmental knowledge of herders, professional foresters, and other traditional communities. Human Ecology 31, 233–253.
Said, E., 1978. Orientalism. Penguin, Harmondsworth.
Schuurman, N., 2000. Trouble in the heartland: GIS and its critics in the 1990s. Progress in Human Geography 24, 569–590.
Schuurman, N., 2006. Formalization matters: critical GIS and ontology research. Annals of the Association of American Geographers 96, 726–739.
Scott, A., 2006. Creative cities: conceptual issues and policy questions. Journal of Urban Affairs 28 (1), 1–17.
Sharp, J., 2009. Geographies of Postcolonialism. Sage, London.
Sheppard, E., 2002. The spaces and times of globalization: place, scale, networks, and positionality. Economic Geography 78, 307–330.
Sheppard, E., 2005. Knowledge production through GIS: genealogy and prospects. Cartographica 40 (4), 5–21.
Sheppard, E., 2005a. Knowledge production through critical GIS: genealogy and prospects. Cartographica: International Journal for Geographic Information and Geovisualization 40, 5–21.
Sheppard, E., 2005b. Cartography and power: the "post-structuralist turn," 1982–1991. Cartographica 40, 85–111.
Staeheli, L., 2011. Political geography: Where's citizenship? Progress in Human Geography 35 (3), 393–400.
Storper, M., Venables, A., 2004. Buzz: face to face contact and the urban economy. Journal of Economic Geography 4, 351–371.
Sui, D., 2004. GIS, cartography, and the "Third Culture": geographic imaginations in the computer age. Professional Geographer 56, 62–72.
Thien, D., 2005. After or beyond feeling? A consideration of affect and emotion in geography. Area 37 (4), 450–456.
Thrift, N., 2004. Intensities of feeling: towards a spatial politics of affect. Geografiska Annaler 86 (1), 57–78.
Thrift, N., 2008. Non-representational Theory: Space Politics Affect. Routledge, London.
Thrift, N., French, S., 2002. The automatic production of space. Transactions of the Institute of British Geographers 27, 309–335.
Tolia-Kelly, D., 2006. Affect: an ethnocentric encounter? Exploring the 'universalist' imperative of emotional/affectual geographies. Area 38 (2), 213–217.
Urbanik, J., 2012. Placing Animals: An Introduction to the Geography of Human-Animal Relations. Rowman and Littlefield, Latham, MD.
Vorley, T., 2008. The geographic cluster: a historical overview. Geography Compass 2 (3), 790–813.
Wainwright, J., 2010. Climate change, capitalism, and the challenge of transdisciplinarity. Annals of the Association of American Geographers 100 (4), 1–9.
Warf, B., 2004. Advancing human geography at the commencement de siecle. Professional Geographer 56, 44–52.
Warf, B., 2012. Nationalism, cosmopolitanism, and geographical imaginations. Geographical Review 102 (3), 271–292.
Warf, B., Sui, D., 2010. From GIS to neogeography: ontological implications and theories of truth. Annals of GIScience 26 (4), 197–209.
Warf, B., Arias, S. (Eds.), 2008. The Spatial Turn: Interdisciplinary Perspectives. Routledge, London.
Whatmore, S., 1997. Dissecting the autonomous self: hybrid cartographics for a relational ethics. Environment and Planning D: Society and Space 15, 37–53.
Whatmore, S., 2001. Hybrid Geographies: Natures Culture Spaces. Sage, London.
Wood, D., 1992. The Power of Maps. Guilford, New York.
Wood, D., 2010. Rethinking the Power of Maps. Guilford, New York.
Wrigley, N., 2002. Food deserts' in British cities: policy context and research priorities. Urban Studies 39 (11), 2029–2040.
Yeung, H.W.-C., 2005. Rethinking relational economic geography. Transactions of the Institute of British Geographers 30 (1), 37–51.
Zook, M., Graham, M., 2007a. The creative reconstruction of the internet: google and the privatization of cyberspace and digiPlace. Geoforum 38, 1322–1343.
Zook, M., Graham, M., 2007b. Mapping digiPlace: geocoded internet data and the representation of place. Environment and Planning B 34, 466–482.

Space, Linguistic Expression of

Penelope Brown, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.C. Levinson, volume 22, pp. 14749–14752, © 2001, Elsevier Ltd.

Abstract

Spatial cognition is central to *human thinking*, and *spatial language* is thus an important area of study, as it may reveal fundamental properties of human thought. Recent research has shown that spatial language is much more divergent across languages than had previously been thought, suggesting significant cultural patterning of spatial conceptualization. This article reviews spatial language cross-linguistically, sets out a typological framework for the language of space, and considers the relationship of spatial language to spatial cognition, in the context of extensive linguistic diversity in the spatial domain.

Spatial cognition is at the heart of much human thinking, as evidenced, for example, in the explanatory power of spatial metaphors and diagrams, and the explication of spatial concepts has played a key role in the development of Western mathematics and philosophy over two and a half millennia (see Jammer, 1954). It is therefore easy to assume that spatial notions form a robust, universal core of human cognition; this assumption has dominated the cognitive sciences (e.g., Miller and Johnson-Laird, 1976). But naive human spatial language turns out to vary significantly across languages, both in the way it is semantically organized and the way in which it is coded. This has major implications for how we should think about universals in human thought and language.

The Western intellectual tradition has provided an elaborate metalanguage for spatial concepts, but the conceptual underpinnings of naive human spatial language utilize only a small subset of these concepts. Moreover, there are many aspects of our spatial perception and motor control which utilize much richer and more precise representations of space than are coded in language. The study of the representation of space in language, and its relation to cognition, must therefore be treated as an empirical enterprise, and this article reports on the increasing body of cross-linguistic information about how different languages structure the spatial domain (see Levinson and Wilkins, 2006).

Space as a Semantic Domain

Languages tend not to treat space as a single, coherent semantic field in the same way that they treat, for example, color (q.v.), kinship (q.v.), or ethnobotany – rather, space is structured as a large semantic field with many distinct subdomains, each of which is closely structured (and often varies considerably across languages). Nearly all languages have 'where' questions that cover the entire field (i.e., with the same or morphologically related forms for 'where-to,' 'where-from,' and 'where-at') and describe the location of one object primarily with respect to another. But beyond these general properties, spatial expressions are structured in distinct subdomains, notably topology, frames of reference, and motion, with place names and spatial deixis as complementary areas. These subdomains cross-cut in certain ways (e.g., frames of reference may play a role in motion description), but because they have their own internal organization they need to be kept analytically distinct.

Although the philosophical and scientific literature gives us many different ways to talk about space, natural languages tend to structure spatial description more in accord with the ideas of Leibniz than of Newton. That is, location is thought about as the place of one thing relative to another (rather than, as in Newtonian thinking, as specifications in the coordinates of an infinite, abstract three-dimensional 'box'). We will call the object to be located the 'figure' (F), and the object with respect to which the location is specified the 'ground' (G); other equivalent terms in the literature are 'theme' versus 'relatum' or 'trajector' versus 'landmark.' As Talmy (1983) noted, a good ground or landmark object is larger than and less mobile than the figure object to be located. The problems of location specification then largely boil down to the following functional considerations:

1. Where figure F and ground G are contiguous or coincident, a static F may be said to be 'at G.' Where G is large, it can be subdivided, so that F can be said to be 'at the X-part of G.' This is, roughly, the subdomain of topology.
2. Where F and G are separated in space, the problem is to specify an angle or direction from G in which F can be found (the search domain). This involves a coordinate system of some kind – a 'frame of reference' – of which there are three main types found in languages.
3. Where F is in motion, two kinds of ground are especially relevant: the source and the goal of the motion. To indicate the progression of motion (i.e., the increase or decrease of distance from a ground), motion can be specified as 'F is moving toward goal G' or 'F is moving from source S.' In addition, motion can be thought of as bringing about, or destroying, a topological relation, as in 'F moves into/out of G.' But to indicate a direction or vector of motion, both S and G must be specified, or a coordinate system of the kind mentioned in (2) utilized. Of course, motion can also be specified as taking place within a location, in which case the systems in (1) and (2) can be utilized.

These functional distinctions give us a number of cleavages: stasis (or location) versus kinesis (or motion), nonangular topological relationships versus angular relations specified in

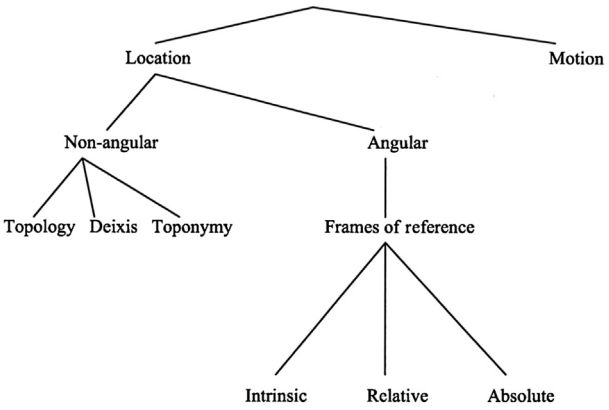

Figure 1 Crosscutting subdomains of spatial language.

frames of reference, with potentially cross-cutting domains as shown in Figure 1.

Some further details in Figure 1, like the three frames of reference, are described below. But the two subfields that have not been discussed further need to be mentioned here. One is *deixis*, which concerns the way in which the location of the speech participants can constitute a special ground or landmark (as in *here*), which then makes interpretation dependent on determining that location (as in *He used to come here*). Languages universally seem to provide a structured set of deictic elements with distinctions that are often spatial (demonstratives like 'this' vs 'that,' adverbs like 'here' vs 'there'). Deictic anchoring of spatial expressions is, however, often implicit or pragmatic: *the local pub* implies a pub close to some reference point, which unless specified, is taken to be the place of speaking.

A second subfield not treated further here is toponymy, or systems of place names. Place names have been much investigated from an historical point of view; since they are highly conservative they reveal a great deal about the locations of ancient languages and peoples. However, the theory and cross-linguistic typology of place names – e.g., why some are descriptive, others not – is hardly developed at all (see e.g., Hunn, 1993).

Universals and Variation in the Semantic Parameters in Spatial Subdomains

In this section, the semantic parameters involved in spatial description in two major subdomains are described to illustrate some of the different ways in which languages structure the spatial domain.

Topology

The subdomain of topology, as mentioned, concerns contiguous relations between figure and ground. The term is potentially misleading – topology proper is a branch of geometry concerned with constancies under continuous distortions such as stretching, but, following Piaget (Piaget and Inhelder, 1956), it is used in studies of spatial language to describe the sort of spatial relations covered by the English prepositions *on*, *in*, and *at* (Herskovits, 1986). These spatial terms are among the earliest learnt by Western children, and it has therefore been supposed that they form a universal innate core of spatial notions. However, cross-linguistic studies show that universal semantic concepts are only to be found at a more abstract level. Take, for example, the spatial relations covered by English *on*, as in *The cup is on the table*, *The picture is on the wall*, or *The fresco is on the ceiling* – clearly the term *on* covers a wide area which is already subdivided in the adpositions (prepositions or postpositions) of many other European languages. Some languages (like Japanese) conflate 'on' relations with 'over' relations (as in *The light is over the table*) in their adpositions; others (like Yukatek Maya) even include 'under' relations as well (as in *The ball is under the table*). A detailed analysis of a dozen languages from different families suggests that universal semantic concepts in this domain lie at a more atomic level (Levinson and Wilkins, 2006). If we think of the core prototype of English *on* as consisting of the notion of vertical superposition of figure over ground, with a horizontal ground supporting a nonattached figure, then the kind of notions that look universal are at the level of 'vertical superposition,' 'contact,' 'horizontal supporting surface,' and (non)'attachment.' Clearly, these more abstract atomic concepts can be rearranged in many different ways to constitute the semantic relations expressible in an individual language (Levinson and Meira, 2003).

Topological relations are coded in the European languages primarily in adpositions or case, but, as we shall see, these are not the only possibilities – some languages code topological relations entirely in nouns or verbs.

Frames of Reference

Linguists have presumed that there are two main frames of reference, which are often called 'deictic' and 'intrinsic,' and which can be illustrated by the ambiguity of the sentence *The boy is behind the bus* – either the bus is between the boy and us ('deictic') or he is at the rear of the bus ('intrinsic'). Psychologists have long known, however, that on the vertical dimension our perceptual systems use three frames of reference against which to measure uprightness: an object-centered frame, a viewer-centered frame, and an environmental frame. Cross-linguistic research shows that in fact these three frames of reference, which we will call the intrinsic, the relative, and the absolute frame, respectively, may be used on the horizontal plane too (Levinson, 1996). These three frames appear to be universally available, yet not all languages make systematic use of them all – some use the absolute frame, others the relative frame, while most use the intrinsic frame, and some use all three. The main universal constraint appears to be that the relative frame is dependent on the possession of the intrinsic frame.

The intrinsic frame of reference, as its name implies, has to do with the intrinsic parts of ground objects. Recollect that frames of reference are coordinate systems that specify angles or directions from a landmark. One way to achieve this is to designate a side or facet of an object, so that the figure can be said to be located on an axis drawn from the center of the object through the designated side. Most languages provide a vocabulary for parts of objects, but the ways in which they partition objects can be very variable. In English, for example, the criteria

for assigning the *front* of an object involve a complex mix of functional and orientational information: the front of a bus is the direction in which it canonically moves, the front of a building is the side in which it is normally entered, and the front of a television the side designed to be watched. In English the *top* of a bottle remains the top even when tipped over, but in other languages (like Zapotec) a more strictly orientational approach may assign the 'top' to whatever side is currently uppermost. Alternatively (as in Tzeltal), the assignment may rely entirely on the internal axial geometry of the object, so that, for example, the 'face' of an object may be a flat side on a secondary axis, even if normally this is undermost. What all these systems have in common is that they provide a system of designated sides to objects which is independent of the position of the viewer, and (apart from the vertical axis) independent of an environmental orientation. How they achieve this differs not only in detail but also in fundamentals.

Another way of obtaining angular distinctions on the horizontal plane is to use the body axes of the viewer. We can then say that something is, for example, to the 'left' of something else in the visual field. Full versions of such systems use a left/right axis and an orthogonal front/back axis, but again the way in which they map bodily axes onto the ground object is very variable. Although these systems are often called 'deictic' in the literature, this is misleading – they typically have a deictic center but they do not have to (as in *He kicked the ball to the left of the goal*), and the other two frames of reference can also have deictic centers (e.g.,'It is north of us,' 'It is at my left hand'). Relative frames of reference are, here, said to be 'relative' because they are relative to the body axes of some observer. Some languages translate the bodily coordinates onto the ground object without rotation; in such languages (like Hausa) *The boy is to the left of the tree* means just what it does in English, but *The boy is in front of the tree* means what in English is described as *The boy is behind the tree*. In contrast, some languages rotate the axes, so that the front of the tree is toward us, but the left of the tree is what in English we would call its right (this is probably a minority option, but is found in some dialects of Tamil). Some languages reflect the axes, as in English, so that left and right remain in the same directions as the viewer's body, but front and back are reversed or 'flipped' over.

Notice that terms like 'front' and 'back' and 'left' and 'right' also have intrinsic interpretations (as in *He's sitting at the chairman's left*), and for that reason ambiguities arise of the kind illustrated above with *The boy is behind the bus*. Relative systems are based on the intrinsic sides of humans, and probably arise as extensions of an intrinsic system to objects (like trees or balls) which resist the assignment of named sides. Hence the implicational relation: all relative systems are associated with intrinsic systems, but the possession of an intrinsic system does not necessarily imply the use of a relative one. Relative systems had been thought to be universal, but in fact it turns out that perhaps as many as a third of the world's languages do not employ them, instead making use of absolute systems.

Absolute systems rely on environmental gradients or 'geocentric' coordinates. Such systems use absolute directions, or fixed bearings, a bit like English *north*, but unlike in English, they are often used on all scales (as in *The cup to the north of yours is mine*, or even *There's a spider on your northern leg*), completely replacing the need for relative systems. There are many different varieties of absolute systems. Some of them (as in Australian Aboriginal languages) are fully abstract cardinal direction systems, with terms that may gloss 'north,' 'south,' etc., but which denote quadrants or arcs of specific angle, and are often based on bearings rotated from our cardinal directions. To use a system of this kind, speakers must always orient themselves precisely, using multiple environmental clues (wind directions, solar angles, sidereal ecliptics, landscape features, and the like). Other systems are based more directly on one major environmental cue, like prevailing winds (as in languages of the Torres Straits), river drainage (as in Alaska), a seaward/landward distinction (as in many Oceanic languages), or prevailing slopes in mountainous terrain (as in the Himalayas). For example, Tenejapan Tzeltal uses an 'uphill,' 'downhill,' 'across' system, based on the general lay of the land in the mountainous territory of Tenejapa (Chiapas, Mexico), but the system has been abstracted from the terrain to provide a fixed orientation, so that the bearings remain constant outside the territory. Many systems use fixed orthogonal axes, but not all; for example, the Austronesian languages in the Pacific islands typically use the monsoon winds as the basis for one axis, but an inland-sea opposition as the basis for another axis; as one moves around the island, the angles between the axes constantly change.

It is important to note that there are cognitive consequences of relying on an absolute frame of reference for routine spatial description on all scales. Communication using such an absolute system requires that speakers' mental computations are adapted to enable constant background tracking of the coordinate system(s) and 'dead-reckoning,' or updating of speakers' locations as they move through the landscape (Levinson, 2003). Various cultural practices will support this background tracking (e.g., absolutely oriented gestures, attention to movements of the sun, shadows, and distant landmarks). Studies on the use of pointing gestures in cultures where absolute systems dominate in everyday use have contributed greatly to our understanding of how spatial frames of reference are invoked, communicated, and switched in conversational interaction, and provided important insights into the cognitive background to absolute systems (e.g., Levinson, 1997; Haviland, 2000, 2003; Kita et al., 2001; Le Guen, 2011). They have shown that gesture is deeply integrated into the system of directional reference, making it clear that these are not simply linguistic systems but broader communicative ones.

This raises the question of how children learn to use such apparently cognitively demanding systems. There have been a number of studies of children's acquisition of absolute systems (de León, 1994; Brown, 2001; Brown and Levinson, 2000, 2009; Cablitz, 2002; Dasen and Mishra, 2010). These have generally found that children master the basics of their absolute system relatively early – by the age of 6 or 7 – compared with the acquisition of projective left/right in European children which is often not fully mastered till age 11 (Piaget and Inhelder, 1956). This result is consonant with other work on child spatial language showing that children's language use reveals very early sensitivity to language-specific semantic differences, adapting to the local system of spatial categories from the beginning (e.g., Bowerman, 1996; Bowerman and Choi, 2001). This is evidence that in this domain the child must *construct* the relevant categories; these are not given by innate endowment, contrary to what had been argued for the spatial domain.

How Spatial Distinctions Are Coded in Language Forms

Much of the literature gives the impression that spatial distinctions are primarily coded in adpositions, as in the prepositions of English. In fact, in most languages spatial information is distributed throughout the clause. Some languages have a score or more local cases (nominal suffixes) covering roughly the same distinctions as English prepositions (e.g., Finnish or the NE Caucasian language Avar). Other languages use a combination of case and adposition (as in Tamil), or case and spatial nominal (as in many Australian languages). Special classes of spatial nominals (e.g., 'front,' 'back,' and 'side') do a lot of the work in isolating languages like Chinese, and they play special roles in other languages (cf *front* in the English complex preposition *in front of*, and *north* in the adverbial use *He went north*). Many spatial oppositions are made constructionally, as when German prepositions like *auf* take the dative to indicate location but the accusative to indicate motion toward a goal.

Finally, verbs play a crucial role in spatial description. In the case of motion, there is often a special form class of motion verbs, and languages tend to conflate motion with manner (as in English *crawl*) or with direction or 'path' (as in Spanish *salir*: 'to go out'), but not generally both (Talmy, 1983, 2000). In many languages (e.g., Dutch), there is a small set of contrasting verbs (often derived from posture verbs) used to describe static location, encoding properties of the figure or ground, or both. Some languages have a larger set (as in Tzeltal), and in this case most of the semantic load is taken by the verbs, which may encode such specific details as 'be located in a bowl-shaped container' (Brown, 1994). Where the verbal component of locative statements is semantically specific in this way, it can impose a spatial categorization on objects in the world (e.g., long.thin.vertically.upright versus horizontally.extended.in.three.dimensions) that is depicted in gestures used when talking of putting and taking things; this provides a window on the possible cognitive effects of these categorizations (Ameka and Levinson, 2007). Even verbs without such spatial specificity – e.g., verbs of motion like 'go,' 'come,' 'enter,' and 'exit' (Wilkins and Hill, 1995; Kita, 1999), and of caused motion or placement verbs like 'put' and 'take' (Kopecka and Narasimhan, 2011; Slobin et al., 2011) display considerable cross-linguistic variability in their semantics. And even closely related languages may differ in their 'basic locative construction,' for example, in whether a preposition or a verb provides the unmarked construction for describing locative relations (as in Mayan languages; see Bohnemeyer and Brown, 2007).

In conclusion, it is clear that spatial concepts form a complex semantic field made up of structured linguistic subdomains. The semantic parameters involved may ultimately have a universal base, but the way in which the semantic subdomains are organized displays considerable cross-linguistic variation. There is also cross-linguistic variation in the way in which such distinctions are encoded in the form classes of languages.

There is now a great deal more information about spatial semantics in different languages than was available 30 years ago. This includes detailed descriptions of the language of space in different languages (e.g., de León and Levinson, 1992; Svorou, 1993; Dirven and Pütz, 1996; Danziger, 1998; Levinson and Wilkins, 2006), and areal surveys of spatial language (e.g., Levinson and Haviland, 1994 for Mayan languages; Senft, 1997 for Oceanic languages; and O'Meara and Pérez Báez, 2011 for Mesoamerican languages), as well as a number of cross-linguistic comparisons of spatial language in particular semantic domains, for example, landscape (Burenhult and Levinson, 2008), body parts (Majid et al., 2006), locative verbs (Ameka and Levinson, 2007), placement verbs (Kopecka and Narasimhan, 2011), and space/time metaphors (e.g., Majid et al., 2012). There is therefore much more visibility for the range of cross-linguistic variation that exists in the spatial domain and in the extension of spatial meanings to more abstract domains like time, kinship, and social relations. In addition, we now have some initial evidence for distinct patterns of neural activation in the brain for different frames of reference (Janzen et al., 2012).

This linguistic variation is especially interesting given the centrality of spatial concepts in human thinking – and indeed it has been shown that linguistic differences in frames of reference can make a difference to how humans memorize and reason nonlinguistically about space (Levinson, 1996, 1997, 2003; Pederson et al., 1998; Majid et al., 2004; Haun et al., 2011). There is robust evidence for several unrelated languages and cultures (ranging from Mexico to Australia to Africa to Nepal and Bali) that peoples' frame of reference usage in nonlinguistic cognition matches that of their language use. This linguistic variation and the associated interactional and cultural properties pose questions that are central to our understanding of how language and cognition are interrelated, how they are embedded in cultural practices, and how children's cognitive development may be shaped while they are learning a language.

See also: Human Spatial Orientation, Neural Basis of; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of; Spatial Cognition; Spatial Navigation; Virtual Reality and Spatial Cognition.

Bibliography

Ameka, F.K., Levinson, S.C. (Eds.), 2007. The typology and semantics of locative predication: posturals, positionals and other beasts. Linguistics 45 (5) (special issue).

Bloom, P., Peterson, M., Nadel, L., Garrett, M. (Eds.), 1996. Language and Space. MIT Press, Cambridge, MA.

Bohnemeyer, J., Brown, P., 2007. Standing divided: dispositional predicates and locative predications in two Mayan languages. Linguistics 45 (5–6), 1105–1151.

Bowerman, M., 1996. The origin of children's spatial semantic categories: cognitive vs. linguistic determinants. In: Gumperz, J.J., Levinson, S.C. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, UK, pp. 145–176.

Bowerman, M., Choi, S., 2001. Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In: Bowerman, M., Levinson, S.C. (Eds.), Language Acquisition and Conceptual Development. Cambridge University Press, Cambridge, pp. 475–511.

Brown, P., 1994. The INs and ONs of Tzeltal locative expressions: the semantics of static descriptions of location. Linguistics 32, 743–790.

Brown, P., 2001. Learning to talk about motion UP and DOWN in Tzeltal: is there a language-specific bias for verb learning? In: Bowerman, M., Levinson, S.C. (Eds.), Language Acquisition and Conceptual Development. Cambridge University Press, Cambridge, pp. 512–543.

Brown, P., Levinson, S.C., 2000. Frames of spatial reference and their acquisition; in Tenejapan Tzeltal. In: Nucci, L., Saxe, G., Turiel, E. (Eds.), Culture, Thought, and Development. Erlbaum, Mahwah, NJ, pp. 167–197.
Brown, P., Levinson, S.C., 2009. Language as mind tools: learning how to think through speaking. In: Guo, J., Lieven, E.V., Budwig, N., Ervin-Tripp, S., Nakamura, K., Ozcaliskan, S. (Eds.), Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin. Psychology Press, New York, pp. 451–464.
Burenhult, N., Levinson, S.C. (Eds.), 2008. Language and landscape: geographic ontology in cross-linguistic perspective. Language Sciences 30 (2/3) (special issue).
Cablitz, G.H., 2002. The acquisition of an absolute system: learning to talk about space in Marquesan (Oceanic, French Polynesia). In: Clark, E.V. (Ed.), Papers of the 2002 Stanford Child Language Research Forum. CLSI, Stanford, CA, pp. 40–49.
Danziger, E. (Ed.), 1998. Language, Space, and Culture, Special Issue of Ethos, vol. 26 (1).
Dasen, P.R., Mishra, R.C., 2010. Development of Geocentric Spatial Language and Cognition: An Eco-Cultural Perspective. Cambridge University Press, Cambridge, UK.
Dirven, R., Pütz, M. (Eds.), 1996. The Construal of Space in Language and Thought. De Gruyter, Berlin.
Haun, D.B.M., Rapold, C.J., Janzen, G., Levinson, S.C., 2011. Plasticity of human spatial memory: spatial language and cognition covary across cultures. Cognition 119, 70–80.
Haviland, J.B., 2000. Pointing, gesture spaces, and mental maps. In: McNeill, D. (Ed.), Language and Gesture. Cambridge University Press, Cambridge, pp. 13–46.
Haviland, J.B., 2003. How to point in Zinacantán. In: Kita, S. (Ed.), Pointing: Where Language, Culture, and Cognition Meet. Lawrence Erlbaum Associates, Mahwah, N.J./London, pp. 139–170.
Herskovits, A., 1986. Language and Spatial Cognition. Cambridge University Press, Cambridge, UK.
Hunn, E., 1993. Columbia Plateau Indian place names: what can they teach us? Journal of Linguistic Anthropology 6 (1), 3–26.
Jammer, M., 1954. Concepts of Space: The History of Theories of Space in Physics. Harvard University Press, Cambridge, MA.
Janzen, G., Haun, D.B.M., Levinson, S.C., 2012. Tracking down abstract linguistic meaning: neural correlates of spatial frame of reference ambiguities in language. PLoS One 7 (2), e30657.
Kita, S., 1999. Japanese enter/exit verbs without motion semantics. Studies in Language 23 (2), 307–330.
Kita, S., Danziger, E., Stolz, C., 2001. Cultural specificity of spatial schemas, as manifested in spontaneous gestures. In: Gattis, M. (Ed.), Spatial Schemas and Abstract Thought. MIT Press, Cambridge, MA, pp. 115–146.
Kopecka, A., Narasimhan, B. (Eds.), 2011. Events of Putting and Taking: A Crosslinguistic Perspective. Benjamins, Amsterdam.
Landau, B., Jackendoff, R., 1993. 'What' and 'where' in spatial language and spatial cognition. Behavioural and Brain Sciences 16 (2), 217–265.
Le Guen, Olivier, 2011. Modes of pointing to existing spaces and the use of frames of reference. Gesture 11 (3), 271–307.
Levinson, S.C., 1996. Frames of reference and Molyneux's question: crosslinguistic evidence. In: Bloom, P., Peterson, M., Nadel, L., Garrett, M. (Eds.), Language and Space. MIT Press, Cambridge, MA, pp. 109–170.
Levinson, S.C., 1997. Language and cognition: the cognitive consequences of spatial description in Guugu Yimithirr. Journal of Linguistic Anthropology 7 (1), 98–131.
Levinson, S.C., 2003. Space in Language and Cognition: Explorations in Linguistic Diversity. Cambridge University Press, Cambridge, UK.
Levinson, S.C., Haviland, J.B. (Eds.), 1994. Space in Mayan languages. Linguistics 32 (4/5) (special issue).
Levinson, S.C., Meira, S., 2003. Natural concepts' in the spatial topological domain – adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. Language 79 (3), 485–516.
Levinson, S.C., Wilkins, D. (Eds.), 2006. Grammars of Space: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge, UK.
Lyons, J., 1977. Semantics. Cambridge University Press, Cambridge, UK.
de León, L., 1994. Exploration in the acquisition of geocentric location by Tzotzil; children (Special Issue). In: Haviland, J., Levinson, S.C. (Eds.), Spatial Conceptualization in; Mayan Languages, Linguistics, vol. 32(4/5), pp. 857–884.
de León, L., Levinson, S.C. (Eds.), 1992. Spatial description in Mesoamerican languages. Special Edition of Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 45 (6).
Majid, A., Boroditsky, L., Gaby, A. (Eds.), 2012. Time in Terms of Space (Research Topic) (Special Issue). Frontiers in Cultural Psychology.
Majid, A., Bowerman, M., Kita, S., Haun, D., Levinson, S.C., 2004. Can language restructure cognition? The case for space. Trends in Cognitive Sciences 8 (3), 108–114.
Majid, A., Enfield, N.J., van Staden, M. (Eds.), 2006. Parts of the body: cross-linguistic categorization. Language Sciences 28 (2/3) (special issue).
Miller, G., Johnson-Laird, P., 1976. Language and Perception. Cambridge University Press, Cambridge, UK.
O'Meara, C., Pérez Báez, G., 2011. Spatial frames of reference in Mesoamerican languages. Language Sciences 33 (6), 837–852.
Pederson, E., Danziger, E., Wilkins, D., Levinson, S.C., Kita, S., Senft, G., 1998. Semantic typology and spatial conceptualization. Language 74, 557–589.
Piaget, J., Inhelder, B., 1956. The Child's Conception of Space. Routledge, London.
Senft, G. (Ed.), 1997. Referring to Space: Studies in Austronesian and Papuan Languages. Clarendon Press, Oxford.
Slobin, D.I., Bowerman, M., Brown, P., Eisenbeiss, S., Narasimhan, B., 2011. Putting things in places: developmental consequences of linguistic typology. In: Bohnemeyer, J., J.Pederson, E. (Eds.), Event Representation in Language and Cognition. Cambridge University Press, Cambridge, pp. 134–165.
Svorou, S., 1993. The Grammar of Space. Benjamins, Amsterdam.
Talmy, L., 1983. How language structures space. In: Pick, H., Acredolo, L. (Eds.), Spatial Orientation: Theory, Research and Application. Plenum Press, New York.
Wilkins, D., Hill, D., 1995. When GO means COME: questioning the basicness of basic motion verbs. Cognitive Linguistics 6 (2), 209–259.
Talmy, L., 2000. Toward a Cognitive Semantics, vols. 1, 2. MIT Press, Cambridge, MA.

Relevant Websites

http://www.acsu.buffalo.edu/~jb77/Mesospace.htm – Spatial language and cognition in Mesoamerica.
http://www.acsu.buffalo.edu/~jb77/Mesospace1b.html – Spatial language and cognition beyond Mesoamerica.
http://www.mpi.nl/news/news-archive/language-seems-crucial-to-think-about-space-1 – Max Planck Institute for Psycholinguistics.

Spatial Analysis in Geography

Manfred M Fischer, Vienna University of Economics and Business, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews some of the special properties of spatial data and the ways in which these properties have influenced developments in spatial data analysis. The discussion focuses on exploratory and model-driven (confirmatory) modes of analyzing spatial data, in particular spatial point patterns and area data, which occupy a prominent position in social science research. The article concludes with a brief discussion of some key areas in which the field of spatial data analysis has been growing in recent years.

The proliferation and dissemination of digital spatial databases, coupled with ever-wider geographic information systems (GISs) use, is stimulating increasing interest in spatial analysis from outside the spatial sciences. Scholars have begun to recognize that including the spatial dimension in social science research sometimes yields different and more meaningful results than analysis that ignores it.

Spatial analysis is a research paradigm that provides a unique set of techniques and methods for analyzing events – events in a very general sense – that are located in geographical space. Spatial analysis involves spatial modeling, which includes models of location allocation, spatial interaction, spatial choice and search, spatial optimization, and space–time (for an account of the diversity of approaches, see Fischer and Getis, 2010). Other articles in the encyclopedia take up these models (e.g., *see* Location Theory; Spatial Interaction Models; Spatial Optimization Models; Spatial–Temporal Modeling); this article concentrates on spatial data analysis, the heart of spatial analysis. The discussion here is on exploratory and model-driven (confirmatory) modes of analyzing and modeling spatial data (see **Table 1**).

Spatial Data and the Tyranny of Data

In essence, spatial data analysis focuses on detecting patterns and exploring and modeling relationships between such patterns in order to understand processes responsible for observed patterns. In this way, spatial data analysis emphasizes the role of space as a potentially important explicator of socioeconomic systems and attempts to enhance understanding of the working and representation of space, spatial patterns, and processes.

Spatial Data and Data Types

Empirical studies in the spatial sciences routinely employ data for which locational attributes are an important source of information. Such data characteristically consist of one or a few cross sections of observations for either micro-units such as individuals (households, firms) at specific points in space, or aggregate spatial entities such as census tracts, electoral districts, regions, provinces, or countries. Observations such as these, for which the absolute location and/or relative positioning (spatial arrangement) is explicitly taken into account, are termed 'spatial data.'

In the socioeconomic realm, points, lines, and areal units are the fundamental entities for representing spatial phenomena. This form of spatial referencing is also a salient feature of GIS (e.g., *see* Geographic Information Systems; Spatial Data Infrastructures). Three broad classes of spatial data can be distinguished:

1. *Object data*, where the objects are either points (spatial point patterns or locational data, that is, point locations at which

Table 1 Popular techniques and methods in spatial data analysis

	Exploratory spatial data analysis	*Model-driven spatial data analysis*
Object data		
Point pattern	Quadrat methods Kernel density estimation Nearest neighbor methods *K*-function analysis	Homogenous and heterogenous Poisson process models and multivariate extensions
Area data	Visualization methods and mapping such as linking and brushing views and cartograms Global and local spatial autocorrelation statistics	Spatial regression models: Spatial lag and spatial error models Spatial Durbin model
Field data	Choropleth maps Spatial cluster detection statistics	Trend surface models Semivariogram and covariance models
Spatial interaction data	Exploratory techniques for representing such data Techniques to uncover evidence of hierarchical structure in the data such as graph-theoretic and regionalization techniques	Poisson spatial interaction models Spatial filtering versions of spatial interaction models

events of interest have occurred) or areas (area or lattice data, defined as discrete variations of attributes over space).
2. *Field data* (also termed geostatistical or spatially continuous data), that is, observations associated with a continuous variation over space, given values at fixed sampling points.
3. *Spatial interaction data* (sometimes called link or flow data), consisting of measurements each of which is associated with a link or pair of locations representing points or areas.

The analysis of spatial interaction data has a long and distinguished history in the study of a wide range of human activities, such as transportation movements, migration, and the transmission of information (see Fischer and Wang, 2011, for a discussion). Field data play an important role in the environmental sciences, but are less important in the social sciences. This article, therefore, focuses on object data, the most appropriate perspective for spatial analysis applications in the social sciences. Object data include observations for micro-units at specific points in space, that is, spatial point patterns, and/or observations for aggregate spatial entities, that is, area data.

Of note is that point data can be converted to area data, and area data can be represented by point reference. Areas may be regularly shaped such as pixels in remote sensing or irregularly shaped such as statistical reporting units. When divorced from their spatial context, such data lose value and meaning. They may be viewed as single realizations of a spatial stochastic process, similar to the approach taken in the analysis of time series.

The Tyranny of Data

Analyzing and modeling spatial data present a series of problems; solutions to many of which are obvious, while solutions to others require extraordinary effort. Data exercise a power that can lead to misinterpretation and meaningless results; therein lies the tyranny of data.

Quantitative analysis crucially depends on *data quality*. Good data are reliable, contain few or no mistakes, and can be used with confidence. Unfortunately, nearly all spatial data are flawed to some degree (e.g., *see* Spatial Data). Errors may arise in measuring both the location and attribute properties but may also be associated with computerized processes responsible for storing, retrieving, and manipulating spatial data. The solution to the data quality problem is to take the necessary steps to avoid having faulty data determining research results.

The *particular form* (i.e., size, shape, and configuration) *of the spatial aggregates* can affect the results of the analysis to a varying, usually unknown, degree, as evidenced in various types of analysis (e.g., Openshaw and Taylor, 1979; Baumann et al., 1983). This problem generally has become recognized as the modifiable areal unit problem (MAUP), the term stemming from the fact that areal units are not 'natural' but usually arbitrary constructs. For reasons of confidentiality, social science data (e.g., census data) are not often released for the primary units of observation (individuals), but only for a set of rather arbitrary areal aggregations (enumeration districts or census tracts). The problem arises whenever area data are analyzed or modeled and involves two effects. One effect derives from selecting different areal boundaries while holding the overall size and the number of areal units constant (the zoning effect). The other effect derives from reducing the number but increasing the size of the areal units (the scale effect). There is no analytical solution to the MAUP, but questions of the following kind have to be considered in constructing an areal system for analysis: What are the basic spatial entities for defining areas? What theory guides the choice of spatial scale? Should the definition process follow strictly statistical criteria and merge basic spatial entities to form larger areas using some regionalization algorithms (see Wise et al., 1996)? These questions pose daunting challenges.

In addition, *boundary and frame effects* (i.e., the geometric structure of the study area) may affect spatial analysis and the interpretation of results. These problems are considerably more complex in time-series analyses. Although several techniques, such as refined *K*-function analysis, take the effect of boundaries into account, there is a need to study boundary effects more systematically.

An issue that has been receiving increasing attention relates to the suitability of data. If the data, for example, are available only at the level of spatial aggregates, but the research question is at the individual respondent level, then the *ecological fallacy* (ecological bias) problem arises. Using area-based data to draw inferences about underlying individual-level processes and relationships poses considerable risks. This problem relates to the MAUP through the concept of spatial autocorrelation.

Spatial autocorrelation (also referred to as spatial dependence or spatial association) in the data can be a serious problem (e.g., *see* Spatial Autocorrelation), rendering conventional statistical analysis unsafe and requiring specialized spatial analytical tools. This problem refers to situations where the observations are nonindependent over space. That is, nearby spatial units are associated in some way. Sometimes this association is due to a poor match between the spatial extent of the phenomenon of interest (e.g., a labor or housing market) and the administrative units for which data are available. Other times it is due to a spatial spillover effect. The complications are similar to those found in time-series analysis, but are exacerbated by the multidirectional, two-dimensional nature of dependence in space rather than the unidirectional nature in time. Avoiding the pitfalls arising from spatially correlated data is crucial to good spatial data analysis, whether exploratory or confirmatory. Several scholars even argue that the notion of spatial autocorrelation is at the core of spatial analysis (e.g., Tobler, 1979). No doubt much of current interest in spatial analysis is directly derived from the monograph of Cliff and Ord (1973) on spatial autocorrelation, which opened the door to modern spatial analysis.

Pattern Detection and Exploratory Analysis

Exploratory data analysis (EDA) is a collection of techniques for summarizing data properties, detecting patterns in data, identifying unusual or interesting properties in data, detecting errors, and formulating hypotheses from data. Many such techniques emphasize graphical views of the data that are designed to highlight particular features and allow the analyst

to detect patterns, relationships, outliers, and so forth. Exploratory spatial data analysis (ESDA), an extension of EDA (Haining, 1990; Cressie, 1993), is geared especially to dealing with the spatial aspects of data.

Exploratory Techniques for Spatial Point Patterns

Point patterns arise when the important variable to be analyzed is the location of events. At the most basic level, the data comprise only the spatial coordinates of events. They might represent a wide variety of spatial phenomena, such as cases of disease, crime incidents, pollution sources, or locations of stores (e.g., *see* Spatial Pattern, Analysis of). The visualization of spatial locations (spatial point patterns) can provide valuable insights into such data.

Research typically concentrates on whether the proximity of particular point events, their location in relation to each other, represents a significant (i.e., nonrandom) pattern. Exploratory spatial point pattern analysis is concerned with exploring the first- and second-order properties of spatial point pattern processes. First-order effects relate to variation in the mean value of the process (a large-scale trend), while second-order effects result from the spatial correlation structure or the spatial dependence in the process.

Three types of methods are important: quadrat methods, kernel estimation of the intensity of a point pattern, and distance methods. Quadrat methods involve collecting counts of the number of events in subsets of the study region. Traditionally, these subsets are rectangular (thus the name 'quadrat'), although any shape is possible. The reduction of complex point patterns to counts of the number of events in quadrats and to one-dimensional indices results in a considerable loss of information. There is no consideration of quadrat locations or of the relative positions of events within quadrats. Thus, most of the spatial information in the data is lost. Quadrat counts destroy spatial information, but they give a global idea of subregions with high or low numbers of events per area. For small quadrats, more spatial information is retained, but the picture degenerates into a mosaic with many empty quadrats.

Estimating the intensity of a spatial point pattern is very like estimating a bivariate probability density, and bivariate kernel estimation can be adapted easily to give an estimate of intensity. Choice of the specific functional form of the kernel presents little practical difficulty. For most reasonable choices of possible probability distributions, the kernel estimate will be very similar for a given bandwidth. The bandwidth determines the amount of smoothing. There are techniques that attempt to optimize the bandwidth given the observed pattern of event location.

A risk underlying the use of quadrats is that any spatial pattern detected may be dependent upon the size of the quadrat. In contrast, distance methods make use of precise information on the locations of events and have the advantage of not depending on arbitrary choices of quadrat size or shape. Nearest neighbor methods reduce point patterns to one-dimensional nearest neighbor summary statistics (Dacey, 1960; Getis, 1964). But only the smallest scales of patterns are considered. Information on larger scales of patterns is unavailable. These statistics indicate merely the direction of departure from complete spatial randomness (CSR). The empirical *K*-function, a reduced second-moment measure of the observed process, provides a vast improvement over nearest neighbor indices (Ripley, 1977; Getis, 1984). It uses the precise location of events and includes all event–event distances, not just nearest neighbor distances, in its estimation. Care must be taken to correct for edge effects. *K*-function analysis can be used not only to explore spatial dependence but also to suggest specific models to represent it and to estimate the parameters of such models. The concept of *K*-functions can be extended to the multivariate case of a marked point process (i.e., locations of events and associated measurements or marks) and to the time–space case.

Exploratory Analysis of Area Data

Exploratory analysis of area data is concerned with identifying and describing different forms of spatial variation in the data. Special attention is given to measuring spatial association between observations for one or several variables. Spatial association can be identified in a number of ways, rigorously by using an appropriate spatial autocorrelation statistic (Cliff and Ord, 1981), or more informally, for example, by using a scatter plot and plotting each value against the mean of neighboring areas (Haining, 1990).

In the rigorous approach to spatial autocorrelation, the overall pattern of dependence in the data is summarized in a single indicator, such as Moran's *I* and Geary's *c*. While Moran's *I* is based on cross-products to measure value association, Geary's *c* employs squared differences. Both require the choice of a spatial weight or contiguity matrix that represents the topology or spatial arrangement of the data and hence our understanding of spatial association. Getis (1991) has shown that these indicators are special cases of a general formulation (called 'gamma') defined by a matrix representing possible spatial associations (the spatial weight matrix) among all areal units, multiplied by a matrix representing some specified nonspatial association among the areas. The nonspatial association may be a social, economic, or other relationship. When the elements of these matrices are similar, high positive autocorrelation arises. Spatial association specified in terms of covariances leads to Moran's *I*, specified in terms of differences to Geary's *c*.

These global measures of spatial association can be used to assess spatial interaction in the data and can be visualized easily by means of a spatial variogram, a series of spatial autocorrelation measures for different orders of contiguity. A major drawback of global statistics of spatial autocorrelation is that they are based on the assumption of spatial stationarity, which implies *inter alia* a constant mean (no spatial drift) and constant variance (no outliers) across space. This was useful in the analysis of small data sets characteristic of pre-GIS times but is not very meaningful in the context of thousands or even millions of spatial units that characterize current, data-rich environments.

In view of increasingly data-rich environments, a focus on local patterns of association (hot spots) and an allowance for local instabilities in overall spatial association recently has been suggested as a more appropriate approach. Examples of techniques that reflect this perspective are the various geographical analysis machines developed by Openshaw and co-workers (e.g., Openshaw et al., 1990), the Moran scatter plot (Anselin, 1996), and the distance-based G_i and statistics of Getis and Ord (1993). This last has gained wide acceptance.

These G-indicators can be calculated for each location i in the data set as the ratio of the sum of values in neighboring locations (defined to be within a given distance or order of contiguity) to the sum over all the values. The two statistics differ with respect to the inclusion of the value observed at i in the calculation (included in, or not included in, G_i). They can be mapped easily and used in an exploratory analysis to detect the existence of pockets of local nonstationarity, to identify distances beyond which no discernible association arises, and to find the appropriate spatial scale for further analysis (e.g., *see* Spatial Association, Measures of).

No doubt, ESDA provides a useful means to generate insights into global and local patterns and associations in spatial data sets. The use of ESDA techniques, however, generally is restricted to expert users interacting with the data displays and statistical diagnostics to explore spatial information and to fairly simple, low-dimensional data sets. In view of these limitations, there is a need for novel exploration tools sufficiently automated and powerful to cope with the data richness–related complexity of exploratory analysis in spatial data environments.

Model-Driven Spatial Data Analysis

ESDA is a preliminary step in spatial analysis to more formal modeling approaches. Model-driven analysis of spatial data relies on testing hypotheses about patterns and relationships, utilizing a range of techniques and methodologies for hypothesis testing, the determination of confidence intervals, estimation of spatial models, simulation, prediction, and assessment of model fit. Upton and Fingleton (1985), Anselin (1988), Griffith (1988), Haining (1990, 2003), Cressie (1993), Bailey and Gatrell (1995), and more recently LeSage and Pace (2009) and Fischer and Wang (2011) have helped to make model-driven spatial data analysis accessible to a wide audience in the spatial sciences.

Modeling Spatial Point Patterns

Spatial point pattern analysis grew out of a hypothesis testing and not out of the pattern recognition tradition. The spatial pattern analyst tests hypotheses about the spatial characteristics of point patterns. Typically, CSR represents the null hypothesis against which to assess whether observed point patterns are regular, clustered, or random. The standard model for CSR is that events follow a homogenous Poisson process over the study region, that is, events are independently and uniformly distributed over space, equally likely to occur anywhere in the study region, and not interacting with each other.

Various statistics for testing CSR are available. Nearest neighbor tests have their place in distinguishing CSR from spatially regular or clustered patterns, but little is known from nearest neighbor tests about the nature of the regular or clustered patterns when CSR does not hold. The K-function may suggest a way of fitting alternative models. Correcting for edge effects, however, might provide some difficulty. The distribution theory for complicated functions of the data can be intractable even under the null hypothesis of CSR. Monte Carlo tests are a way around this problem.

If the null hypothesis of CSR is rejected, the next obvious step in model-driven spatial pattern analysis is to fit some alternative (parametric) model to the data. Departure from CSR is typically toward regularity or clustering of events. Clustering can be modeled through a heterogenous Poisson process, a doubly stochastic point process, or a Poisson cluster process arising from the explicit incorporation of a spatial clustering mechanism. Simple inhibition processes can be utilized to model regular point patterns. Markov point processes can incorporate both elements through large-scale clustering and small-scale regularity. After a model has been fitted (usually via maximum likelihood or least squares using the K-function), diagnostic tests have to be performed to assess its goodness of fit. Inference for the estimated parameters is often needed in response to a specific research question. The necessary distribution theory for the estimates can be difficult to obtain, in which case approximations may be necessary. If, for example, clustering is found, one may be interested in the question of whether particular spatial aggregations, or clusters, are associated with proximity to particular sources of some other factor. This leads to multivariate point pattern analysis, a special case of marked spatial point process analysis. For further details, see Cressie (1993).

Modeling Area Data

Linear regression models constitute the leading modeling approach for analyzing social and economic phenomena. But conventional regression analysis does not take into account problems associated with possible cross-sectional correlations among observational units caused by spatial dependence. Two forms of spatial dependence among observations may invalidate regression results: spatial error dependence and spatial lag dependence.

Spatial error dependence might follow from measurement errors such as a poor match between the spatial units of observation and the spatial scale of the phenomenon of interest. Presence of this form of spatial dependence does not cause ordinary least squares (OLS) estimates to be biased, but it affects their efficiency. The variance estimator is downward biased, thus inflating the R^2. It also affects the t- and F-statistics for tests of significance and a number of standard misspecification tests, such as tests for heteroskedasticity and structural stability. To protect against such difficulties, one should use diagnostic statistics to test for spatial dependence among error terms and, if necessary, take action to properly specify the spatially autocorrelated residuals. Typically, dependence in the error term is specified as a spatial autoregressive or as a spatial moving average process. Model parameters of such regression models are usually estimated by nonlinear maximum likelihood methods (Cliff and Ord, 1981; Anselin, 1988).

In the second form of spatial dependence, spatial autocorrelation is attributable to spatial interactions in data. This form may be caused, for example, by significant spatial externalities of a socioeconomic process under study. To specify a regression model involving spatial interaction, one must incorporate the spatial dependence into the covariance structure either explicitly or implicitly by means of an autoregressive and/or moving average interaction structure. This constitutes the model identification problem that is usually carried out using

the correlogram and partial correlogram. A number of spatial regression models, that is, regression models with spatially lagged dependent variables (spatial autoregressive models), have been developed that include one or more spatial weight matrices that describe the many spatial associations in the data. The models incorporate either a simple general stochastic autocorrelation parameter or a series of autocorrelation parameters, one for each order contiguity (see Cliff and Ord, 1981; Anselin, 1988; LeSage and Pace, 2009).

To avoid biased and inconsistent OLS parameter estimates, maximum likelihood procedures are fundamental, but data screening and filtering can simplify estimation. Tests and estimators are clearly sensitive not only to the MAUP but also to the specification of the spatial interaction structure represented by the spatial weight matrix.

There have been substantial advances in the analysis of area data in the past decade, illustrating that ignoring space by involving an inappropriate model specification will give wrong answers, and wrongly interpreting the parameters of an appropriate model is equally likely to mislead. An appropriate model in many circumstances turns out to be the spatial Durbin model (see LeSage and Fischer, 2008; LeSage and Pace, 2009) that is characterized by both a spatially lagged dependent variable and spatially lagged explanatory variables. The presence of lagged variables leads to nuances with respect to the interpretation of the effects of variables, since spillovers cause both direct and indirect effects that need to be accounted for.

Conclusions

The field of spatial data analysis has been growing in many directions in recent years. Four key areas for future growth are especially important to note. First, static (cross-sectional in time) spatial data analysis is restricted to analyzing and modeling some process at some specific point in time. A series of spatial analyses over time can shed light on change, but in other respects remains limited. The understanding that has been gained by the progress made in spatial statistics and spatial econometrics forms an essential element in the emergence of spatiotemporal data analysis.

A second key area relates to spatial data mining, the process of discovering interesting and potentially useful patterns in spatial databases (Miller and Han, 2009). Spatial data mining shares some of the objectives of ESDA, but is concerned with the development of automated procedures that can be applied to very large spatial databases for the purpose of detecting spatial clusters, spatial outliers and co-location and relationship patterns among different classes of point, line, and polygon (area) objects.

Third, neural networks have an important role to play as universal approximators in areas such as regression analysis, classification, and spatial interaction in application domains where solutions require knowledge that is difficult to specify a priori, but for which there are sufficient observations (see Fischer, 2009; Gopal and Fischer, 1996). For neural networks to find a place in spatial data analysis, they need to overcome some current limitations, mainly due to the relative absence of established procedures for model identification, comparable to those for spatial econometric modeling techniques. In particular, providing tests specifically designed to test the adequacy of neural network models is a research issue on its own right. Another important avenue for further research is the incorporation of spatial dependency in the network representation, which has received less attention in the past than it deserves.

Finally, as spatially referenced data sets become more prevalent in scientific applications, the desire for full inference and accurate assessment of model uncertainty has become increasingly important. Bayesian hierarchical models may be seen as a means for capturing different sources of uncertainty accurately, but such models present several computational challenges (LeSage et al., 2009). The special issue of *Computational Statistics and Data Analysis* – 'Spatial statistics: methods, models & computation' – has a number of articles that illustrate computationally feasible hierarchical models for analyzing challenging spatial data sets.

See also: Geographic Information Systems; Location Theory; Mathematical Models in Geography; Regional Science; Remote Sensing; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Data Infrastructures; Spatial Data; Spatial Interaction; Spatial Optimization Models; Spatial Pattern, Analysis of; Spatial Sampling; Spatial–Temporal Modeling; Urban Geography; Urban System in Geography.

Bibliography

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.
Anselin, L., 1996. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer, M.M., Scholten, H.J., Unwin, D. (Eds.), Spatial Analytical Perspectives on GIS. Taylor & Francis, London.
Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Longman, Harlow, UK.
Baumann, J.H., Fischer, M.M., Schubert, U., 1983. A multiregional labor supply model for Austria: the effects of different regionalizations in multiregional labour market modelling. Papers. Regional Science Association 52, 53–83.
Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.
Cliff, A.D., Ord, J.K., 1981. Spatial Processes, Models and Applications. Pion, London.
Cressie, N.A.C., 1993. Statistics for Spatial Data, revised ed. Wiley, New York.
Dacey, M.F., 1960. A note on the derivation of the nearest neighbour distances. Journal of Regional Science 2, 81–87.
Fischer, M.M., 2009. Neural networks for spatial data analysis. In: Fotheringham, A.S., Rogerson, P.A. (Eds.), The SAGE Handbook of Spatial Analysis. Sage Publications, London.
Fischer, M.M., Getis, A., 2010. Handbook of Applied Spatial Analysis. Software Tools, Methods and Applications. Springer, Berlin, Heidelberg, and New York.
Fischer, M.M., Wang, J., 2011. Spatial Data Analysis. Models, Methods and Techniques. Springer, Berlin, Heidelberg and New York (Springer Briefs in Regional Science).
Getis, A., 1964. Temporal land-use pattern analysis with the use of nearest neighbor and quadrat methods. Annals of the Association of American Geographers 54, 391–399.
Getis, A., 1984. Interaction modelling using second-order analysis. Environment and Planning A 16, 173–183.
Getis, A., 1991. Spatial interaction and spatial autocorrelation: a cross-product approach. Papers. Regional Science Association 69, 69–81.
Getis, A., Ord, J.K., 1993. The analysis of spatial association by use of distance statistics. Geographical Analysis 25, 189–206.
Gopal, S., Fischer, M.M., 1996. Learning in single hidden-layer feedforward network models: backpropagation in a spatial interaction modelling context. Geographical Analysis 28, 38–55.
Griffith, D.A., 1988. Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series. Kluwer Academic, Dordrecht.
Haining, R., 1990. Spatial Data Analysis in the Social Sciences. Cambridge University Press, Cambridge, MA.

Haining, R., 2003. Spatial Data Analysis. Cambridge University Press, Cambridge, UK.
LeSage, J.P., Fischer, M.M., 2008. Spatial growth models: model specification, estimation and interpretation. Spatial Economic Analysis 3, 275–303.
LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton.
LeSage, J.P., Banerjee, S., Fischer, M.M., Congdon, P., 2009. Spatial statistics: methods, models & computation. Computational Statistics and Data Analysis 53, 2781–2785.
Miller, H., Han, J., 2009. Geographic Data Mining and Knowledge Discovery. CRC Press, Boca Raton.
Openshaw, S., Taylor, P., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley, N. (Ed.), Statistical Applications in the Spatial Sciences. Pion, London.
Openshaw, S., Cross, A., Charlton, M., 1990. Building a prototype geographical correlates exploration machine. International Journal of Geographical Information Systems 4, 297–312.
Ripley, B.D., 1977. Modelling spatial patterns. Journal of the Royal Statistical Society 39, 172–212.
Tobler, W., 1979. Cellular geography. In: Gale, S., Olsson, G. (Eds.), Philosophy in Geography. Reidel, Dordrecht.
Upton, J.G., Fingleton, B., 1985. Spatial Data Analysis by Example. Wiley, New York.
Wise, S., Haining, R., Ma, J., 1996. Regionalisation tools for the exploratory spatial analysis of health data. In: Fischer, M.M., Getis, A. (Eds.), Recent Developments in Spatial Analysis – Spatial Statistics, Behavioural Modelling and Computational Intelligence. Springer, Berlin, Heidelberg, and New York.

Spatial Association, Measures of

Arthur Getis, San Diego State University, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial association means connectedness or relationship between and among variables over space (the Earth's surface). A single variable may be *spatially autocorrelated*; that is, values of the variable are somehow connected or related spatially. Many variables may be associated with one another at one or more sites. If there is spatial interaction, there is also spatial association. Maps can depict spatial association. A mathematical shorthand technique can be used to represent, in general, measures of spatial association. Scientists test or theorize about variables to determine whether spatial association, either observed or expected, actually can be confirmed. Statistical procedures that have been developed for identifying and measuring the existence of spatial association are outlined.

Introduction

Spatial association can characterize a single variable or many variables simultaneously. A single variable may be *spatially autocorrelated*; that is, values of the variable are somehow connected or related spatially (*see* Spatial Autocorrelation). Many variables – say, environmental variables such as air quality, water quality, certain species location, and human activities – may be associated one with another at one or more sites. In addition, spatial association and *spatial interaction* are related (*see* Spatial Interaction). If there is spatial interaction, for example, the movement of commuters to workplaces, there is also spatial association since home places are spatially connected to workplaces.

The scientist interested in the study of spatial association is motivated by the need to understand the spatial relationships between variables at specific sites or how variables of interest are spread across the Earth's surface. Many disciplines are particularly concerned with the study of spatial association. Among them are geography, regional science, ecology, epidemiology, regional and urban economics, geology, hydrology, and geomorphology. Since this is a broad and popular interest area, many methods and techniques of analysis have been created to confirm notions of spatial association between and among variables (Fischer and Getis, 2010). Scientists test or theorize about variables to determine whether spatial association, either observed or expected, actually can be confirmed. As a result of this interest, statistical procedures have been developed for identifying and measuring the existence of spatial association. After a short discussion of maps depicting spatial association, the remainder of this article will outline the characteristics of measurement procedures.

Maps Showing Spatial Association

Maps depicting spatial association fall into several categories: *in situ*, flows, time series, distance decay, and residuals. *In situ* maps reveal by means of overlays the possible association of two or more variables in a region (see Figure 1). Flow maps show the interaction between sites (see Figure 2). Time series usually are shown as a series of map views over time (days, years, etc.). These represent spatial relationships as they evolve or change over time (see Figure 3(a–c)). Distance-decay maps emphasize Tobler's first law of geography that "everything is related to everything else, [but] near things are more related than distant things." A distance-decay map might show that the effect of an earthquake declines with distance from its epicenter or that the influence of a retail store declines with distance from its site (see Figure 4). Residual maps show the remaining 'unexplained' variation after a model has been used to statistically explain a relationship between two or more variables

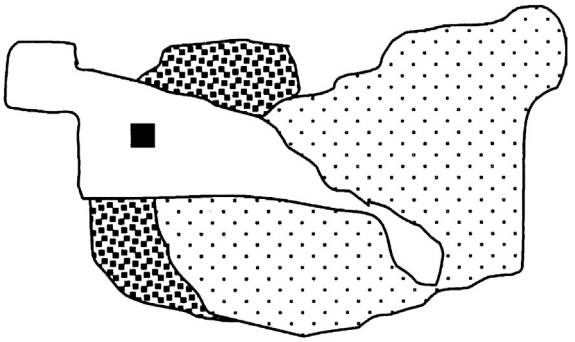

Figure 1 The study area represented by the black square indicates that there may be an *in situ* relationship among the three variables shown by different shading.

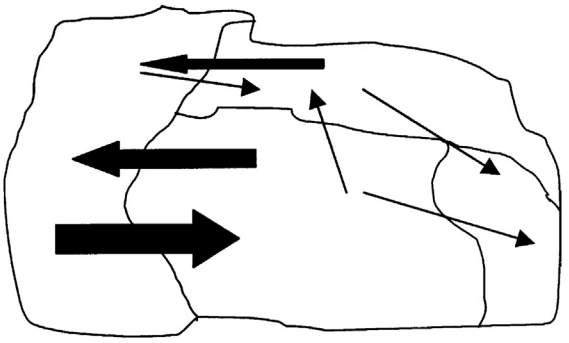

Figure 2 The width of the arrows indicates the magnitude of the interaction.

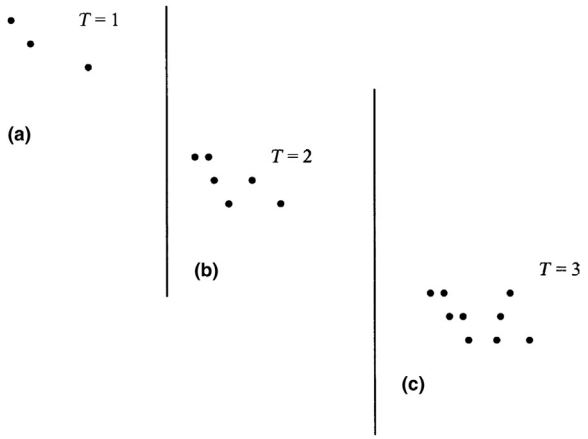

Figure 3 Over time {(a), (b), (c)}, there is an increase in the number of events at the same time as spacing between events is reduced.

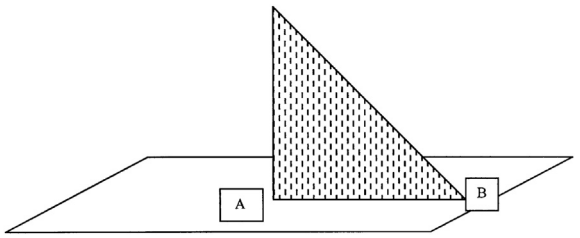

Figure 4 On the plane, as distance increases from A to B the effect on A of the variable indicated by the vertical triangle declines.

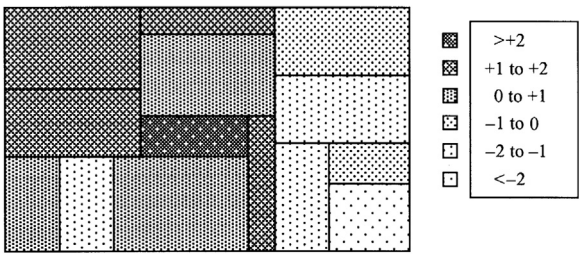

Figure 5 In standard deviational units, the map shows residuals from a regression model. It remains to be determined whether or not there is a spatially statistically significant association among the residuals.

(see Figure 5). The unexplained spatial variation represents that part of whatever total variation exists in a mathematical model that is not accounted for by the variables included in the model. Residual maps help researchers identify new variables that might belong in a more complete model, such as a transportation cost, a regional effect, or a previously unidentified physical or human anomaly.

Spatial Association in Matrix Terms

To answer scientific questions about spatial association, one can reduce the complexities of what one sees on a map to a mathematically manipulable set of numbers (Hubert et al., 1981; Getis, 1991). The following simple cross-product formula represents spatial association:

$$\Gamma = \sum_{i}^{n} \sum_{j}^{n} w_{ij} y_{ij} \qquad [1]$$

Γ is the measure of spatial association, and the w_{ij} are contained in a matrix, **W**, of values that represent the relationship of each location of interest to all other sites (see Figure 6(a) and (b)). Figure 6(a) is a hypothetical map of 10 regions. The numbers in parentheses represent the value that a hypothetical variable, Y, takes in each of the regions. Suppose, it is decided that regions that are contiguous to one another are associated. The **W** matrix (Figure 6(b)) indicates the spatial relationship of each region to every other. For Figure 6(b), the convention is adopted of placing a one in the matrix if a link (contiguity) exists and zero otherwise. The **Y** matrix (see Figure 6(c))

(a)

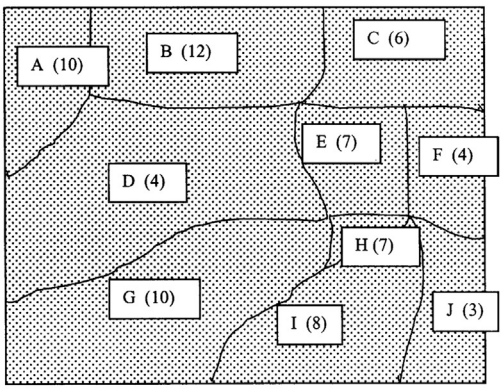

(b)

W =

	A	B	C	D	E	F	G	H	I	J
A	0	1	0	1	0	0	0	0	0	0
B	1	0	1	1	0	0	0	0	0	0
C	0	1	0	0	1	1	0	0	0	0
D	1	1	0	0	1	0	1	0	0	0
E	0	0	1	1	0	1	0	1	0	0
F	0	0	1	0	1	0	0	0	0	1
G	0	0	0	1	0	0	0	1	1	0
H	0	0	0	0	1	0	1	0	1	0
I	0	0	0	0	0	0	1	1	0	1
J	0	0	0	0	0	1	0	0	1	0

(c)

Y =

	A	B	C	D	E	F	G	H	I	J
A	0	2	4	1	3	6	0	3	2	7
B	2	0	6	3	5	8	2	5	4	9
C	4	6	0	3	1	2	4	1	2	3
D	1	3	3	0	2	5	1	2	1	6
E	3	5	1	2	0	3	3	0	1	4
F	6	8	2	5	3	0	6	3	4	1
G	0	2	4	1	3	6	0	3	2	7
H	3	5	1	2	0	3	3	0	1	4
I	2	4	2	1	1	4	2	1	0	5
J	7	9	3	6	4	1	7	4	5	0

Figure 6 The spatial pattern shown in (a) is reduced to a **W** matrix of spatial associations (b) and a **Y** matrix (c) made up of absolute differences between values associated with each region.

	A	B	C	D	E	F	G	H	I	J
A	0	**2**	0	**1**	0	0	0	0	0	0
B	**2**	0	**6**	**3**	0	0	0	0	0	0
C	0	**6**	0	0	**1**	**2**	0	0	0	0
D	**1**	**3**	0	0	**2**	0	**1**	0	0	0
E	0	0	**1**	**2**	0	**3**	0	**0**	0	0
F	0	0	**2**	0	**3**	0	0	0	0	**1**
G	0	0	0	**1**	0	0	0	**3**	**2**	0
H	0	0	0	0	**0**	0	**3**	0	**1**	0
I	0	0	0	0	0	0	**2**	**1**	0	**5**
J	0	0	0	0	0	**1**	0	0	**5**	0

$\Gamma = 66$

Figure 7 The value of Γ is made up of the sum of numbers shown in bold.

contains numbers that are the absolute difference between values for all combinations of the regions. For example, the absolute difference between values in regions C and E is shown as an element of the **Y** matrix in row C and column E, that is, $|6 - 7| = 1$.

The two matrixes, when summed as a cross-product as shown in eqn [1] (see **Figure 7**), give the value of Γ. If the values in the **Y** matrix are small for those same rows and columns that correspond to the $w_{ij} = 1$ values in the **W** matrix, the value of Γ will be small. For the example shown, Γ equals 66, an average of 2.2 for each of the 30 links between contiguous regions. If the number one appeared in all the **W** cells, the total Γ would have been 300 or a mean of 3.3 for each of the hypothetical 90 links. The implication here is that the contiguous neighbors that characterize **W** gives a smaller gamma value than one might expect if the values w_{ij} were randomly placed in the 10 regions. Γ indicates that there may be a spatial association among the Y values. We use the term 'may be' to indicate that we have not yet used a statistical test to confirm that Y is in fact a spatially associated variable.

The W Matrix

In the preceding paragraph, we outlined a simple form of a **W** matrix, one that was constructed based on contiguous spatial relationships between regions. The **W** matrix embodies our preconceived or derived understanding of spatial relationships. If we believe, or if theory tells us, that a particular spatial relationship is distance dependent, then the **W** matrix should reflect that supposition. For example, if it is assumed that a spatial relationship declines in strength as distance increases from any given site, then the **W** matrix will show that nearby areas are weighted more highly than sites that are far from one another. Various distance-decay formulations theorized or derived for such phenomena as travel behavior, economic interaction, or disease transmission would require the elements within the **W** matrix to reflect these effects. Thus, a typical **W** matrix might contain elements of the form $w_{ij} = 1/d_{ij}^{\alpha}$ (α greater than one). Thus, short distance relationships would be weighted more heavily than the long distances.

The Y Matrix

The nonspatial matrix, **Y**, gives us a view of how variables whose elements are georeferenced, that is, values of the variable are indexed at certain sites, *might* be associated with one another. The elements of the **Y** matrix, y_{ij}, represent the interaction of values of **Y**. Values can interact in a number of ways: addition, subtraction, multiplication, division, or combinations of these. An important type of interaction is that of covariance, that is, a value at one site may vary from the mean of all sites (the variation) and this difference is multiplied by the variation at another site. Taken over all sites, the *covariation* is

$$\sum_{i}^{n} \sum_{j}^{n} (y_i - \overline{y})(y_j - \overline{y})$$

and the *covariance* is this sum divided by the sample size, n.

Γ

When the matrix **W** and the matrix **Y** are similarly structured, one can conclude that **Y** values interact very much as the individual regions are spatially associated. Thus, Γ reveals the spatial association in the **Y** variable. In a previous example, we purposely created a **W** matrix that represented a relationship between contiguous regions. The fact that the values of the **Y** variable tend to be similar to one another when regions are contiguous allows us to conclude that the two matrices are similarly structured.

In the literature on spatial association, much attention has been given to the structure of **W** and **Y**. Each, in a sense, represents a related area of research. Those concerned with the **W** matrix have devised ever-increasingly complex representations of spatial relations among spatial units. From the simple contiguity structure presented earlier, empiricists and theorists have constructed **W** matrixes that represent supposed migration distances, disease transmission distances, assumed hierarchical spatial relations, political boundary effects, psychological distances, barriers to communication effects, economic effects of distance, social distances, and so on. In experimental work, the matrix **W** also has been used to represent concentric bands measured from various spatial foci or as representative of cumulative distances. These latter matrixes help researchers to identify 'critical' distances, that is, distances at which association reaches some peak.

Research on the structure of the **Y** matrix usually is a response to a particular type of possible spatial association. For example, the researcher may wonder whether the association is that of positive covariance. **Y** is structured in such a way that statistical tests can be carried out on the research hypothesis that statistically significant spatial association exists. Several of these formulations are given in the following section. Each is an alternative method of identifying spatial association.

Spatial Association as Covariance

Typical formula:

$$\sum_{i}^{n} \sum_{j}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y}) \Big/ \sum_{i}^{n} \sum_{j}^{n} w_{ij}(y_i - \overline{y})^2, \quad i \neq j \qquad [2]$$

This is the well-known *Moran's I statistic*, explicated by Cliff and Ord (1981), and used widely as a measure of spatial

association for spatial residuals from regression analysis. When the data used for testing purposes are georeferenced, the basic form of the linear regression model requires that the error term be spatially distributed randomly. Moran's I is used as a test for the randomness of residuals from regression. The elements of the **W** matrix are represented as w_{ij}. The **Y** matrix is the covariance and the remainder of the equation is a scaling device used to give the statistic useful testing properties.

Spatial Association as Difference

Formulas:

$$\gamma(h) = \sum_{i}^{n-h} \sum_{j=1+h}^{n} w_{ij}(y_i - \bar{y})^2 / 2n(h) \qquad [3]$$

$$c = n - 1 \sum_{i}^{n} \sum_{j}^{n} w_{ij}(y_i - y_j)^2 / 2 \sum_{i}^{n} \sum_{j}^{n} w_{ij}(y_i - \bar{y})^2, \quad i \neq j \qquad [4]$$

The better known of these formulas is the first, called the *semivariance* or *variogram*. Widely used among geoscientists, this measure helps to identify the decline in spatial association as distance (h) increases away from the sites in question. It is the backbone for the large research area that yields continuous map surfaces common in studies of meteorology, geography, and geology (Cressie, 1993). **W** is again represented by w_{ij}. **Y** is made up of $(y_i - y_j)^2$ and the remainder of the equation is used to scale the semivariance in terms of the variance. Equation [4] is known as Geary's c.

Spatial Association as Clusters

Formula:

$$k(d) = \left[A \sum_{i}^{n} \sum_{j}^{n} w_{ij}(d) / \pi n(n-1) \right]^{1/2} \qquad [5]$$

This measure, called the *K-function*, helps to identify *spatial clustering* (Ripley, 1981) within the region A. It is used primarily in ecological and epidemiological studies (Diggle, 1983). The **W** matrix is made up of the w_{ij} elements, represented as ones, up to an arbitrary distance d. The **Y** matrix is invisible but is assumed to have each element equal to one. In practice, K values are found for a series of distances (Aldstadt, 2010). Tests usually are carried out by simulating alternative patterns of points (*see* Spatial Pattern, Analysis of).

Focused Spatial Association

Formula:

$$G_i^*(d) = \sum_{j}^{n} w_{ij}(d) y_j / \sum_{j}^{n} y_j \qquad [6]$$

For each type of measure shown in eqn [6], there is a counterpart (usually with some variation) that allows one to determine whether or not spatial association exists within the proximal region surrounding a single georeferenced point. A popular measure of this type is the *Getis–Ord statistic*, $G_i^*(d)$, which for the **Y** matrix each georeferenced point i is weighted (Getis and Ord, 1992). As distance increases from the site in question, values of G_i^* increase if clustering is in evidence; otherwise, they decrease. Anselin's *Local Indicators of Spatial Association* or *LISA* statistics are designed for focusing Moran's and Geary's statistics (Anselin, 1995).

The focused measures are based on comparisons between local conditions and regional conditions. If local circumstances indicate greater or lesser association than in the region as a whole, tests on statistically significant spatial clusters of high or low y values are in order. The O statistic of Ord and Getis (2001) is designed as a test of local spatial association in the presence of regional spatial association. The study of spatial clustering can be carried further to measures of associated variance within clusters themselves (Ord and Getis, 2012).

Spatial Association as Interaction

The basic formula:

$$T_{ij} = y_i y_j / w_{ij}^{-\beta} \qquad [7]$$

There are a number of interaction models; eqn [7] is the simplest. These widely used models enable transportation and economic researchers to estimate the flow of automobiles, aircraft, migrants, goods and services, and so on, between places (Haynes and Fotheringham, 1984). The **W** matrix usually represents a distance-decay function, while **Y** is usually a multiplicative interaction matrix.

All of these models have the Γ form. All can be used to test hypotheses about or to measure the extent of spatial association. Summaries on the measurement of spatial association can be found in Getis (1966, 1999), Bailey and Gatrell (1995) and Fischer and Getis (2010).

See also: Mathematical Models in Geography; Spatial Analysis in Geography; Spatial Autocorrelation; Spatial Interaction.

Bibliography

Aldstadt, J., 2010. Spatial clustering. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis. Springer, Heidelberg, pp. 279–300.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geographical Analysis 27, 93–115.

Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Longman Scientific and Technical, Harlow, Essex, UK.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion Press, London.

Cressie, N.A.C., 1993. Statistics for Spatial Data, revised ed. Wiley, New York.

Diggle, P.J., 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, London.

Fischer, M.M., Getis, A. (Eds.), 2010. Handbook of Applied Spatial Analysis. Springer, Heidelberg.

Getis, A., 1991. Spatial interaction and spatial autocorrelation: a cross-product approach. Environment and Planning A 23, 1269–1277.
Getis, A., 1966. Local spatial statistics. In: Longley, P., Batty, M. (Eds.), Spatial Analysis: Modelling in a GIS Environment. Geoinformation International, Cambridge, UK.
Getis, A., 1999. Spatial statistics. In: Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (Eds.), Geographical Information Systems, second ed., vol. 1. Wiley, New York.
Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24, 189–206.
Haynes, K.E., Fotheringham, A.S., 1984. Gravity and Spatial Interaction Models. Sage, Beverly Hills, CA.
Hubert, L.J., Golledge, R.G., Costanzo, C.M., 1981. Generalized procedures for evaluating spatial autocorrelation. Geographical Analysis 13, 224–233.
Ord, J.K., Getis, A., 2001. Testing for local spatial autocorrelation in the presence of global autocorrelation. Journal of Regional Science 41, 411–432.
Ord, J.K., Getis, A., 2012. Local spatial heteroscedasticity (LOSH). Annals of Regional Science 48, 529–539.
Ripley, B.D., 1981. Spatial Statistics. Wiley, New York.

Spatial Autocorrelation

Robert Haining, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial autocorrelation is used to describe a particular form of spatial variation found in geographical data. Its presence is often indicative of something of interest in mapped data that requires explanation; may be used to tackle map interpolation problems; may raise statistical inferential problems because data are no longer independent. We discuss how spatial autocorrelation is measured and its presence is tested for; implications for spatial sampling, exploratory spatial data analysis, map interpolation, and confirmatory analysis including regression modeling.

Spatial autocorrelation, one of the 'special' properties of spatial data (Haining, 2009), is the term used to describe the presence of a particular form of spatial variation in a variable that is observed and measured at different locations across a geographical area. Spatially referenced variables may assume values (1) at any point on a continuous surface (such as land use at a point in time or annual precipitation levels across a region); or (2) at a set of discrete sites at fixed locations within a region (such as prices for the same commodity reported at the same time at a set of geographically scattered retail outlets); or (3) across a set of areas that subdivide a region (such as the count or proportion of households currently receiving one or more welfare benefits or who have been victims of burglary in the last 12 months across a set of Census tracts that partition an urban region). The term 'spatial autocorrelation' is highly suggestive of its meaning: 'auto' as in 'self'; 'correlation' as in the statistical use of that term to measure a 'relationship' – hence a measure of the relationship between the value of a variable at a location and the same variable but at another location separated by some specified distance (or, in the case of type (3) variables, some measure of 'lag' separation between areas). Positive spatial autocorrelation is the tendency for sites or areas separated by a specified distance or 'lag' to have similar values of the variable (i.e., both values are high or both are low). Negative spatial autocorrelation is the tendency for the values to be dissimilar (if one value tends to be high, the other tends to be low and vice versa).

These terms are often used with specific reference to values for areas or point sites that are close together in space. If values, recorded at nearby points or areas across a map tend to have similar values, it is often said that the map displays positive spatial autocorrelation; if they are dissimilar we say that the map displays negative spatial autocorrelation. Although not common, negative spatial autocorrelation can arise naturally in cases (2) and (3). (For example, in the case of type (3) variables, think of a chess board where the squares are the areas and hence the units of observation. The juxtaposition of black and white squares indicates negative spatial autocorrelation at the scale of the squares.) By contrast, negative spatial autocorrelation is impossible to imagine in the case of the continuous type of variable mentioned earlier (in (1)) as the spatial distance between any two sample points converges toward zero.

The presence of spatial autocorrelation in map data is often taken as indicative of something of interest in the map (at the distance or 'lag' scale at which it is identified) that calls for further investigation in order to understand the reasons behind the observed spatial autocorrelation (Cliff and Ord, 1981). If geographic phenomena display spatial autocorrelation at some scale, then this may provide the starting point for theorizing about the spatial outcomes of processes that may have their roots in, for example, economic or social science theory.

If spatial data display spatial autocorrelation, this has important implications for the application of statistical methods to such data. The term spatial autocorrelation (or some similar concept) arises in several different data analysis contexts. For example, it arises in the fields of spatial statistics (Ripley, 1981; Cressie, 1991), spatial econometrics (Anselin, 1988), and geostatistics (Cressie, 1991; Isaaks and Srivastava, 1989). If there is spatial autocorrelation in a data set, data values are not independent and each value provides some information on the values of the same variable at nearby sites or adjacent areas. This implies that within the spatial data set there is information redundancy (since data values are not independent) and for the purposes of hypothesis testing the size of the sample is not the same as the number of independent observations of the process. It is with the recognition of this property that statistical methods for analyzing spatial data part company with the methods of traditional statistics (where any sample is assumed to comprise a set of independent observations), sharing ground instead with the sorts of methods used to analyze time series data (which typically displays 'temporal autocorrelation').

The following sections will describe how spatial autocorrelation arises; how it is measured and its presence tested for; and the implications of spatial autocorrelation for spatial sampling, exploratory spatial data analysis, map interpolation, and confirmatory analysis including regression modeling. For an overview of current issues as well as detail on the history of research into spatial autocorrelation in geography, see Griffith (2009).

How Does Spatial Autocorrelation Arise?

Spatial autocorrelation can be the outcome of a deterministic process or a stochastic (random) process. In the case of

a deterministic process, there is conceptually only one set of values (those that are observed), and it is the arrangement of these values across the region that are defined as spatially autocorrelated or not. In the case of type (2) and (3) data, the reference framework for deciding whether such a distribution of values is autocorrelated is the geographic distribution of values that would have been obtained, had those same values been allocated randomly to the points or areas in the region. In the case of a stochastic process, the observed values are but one set of realized values from an infinite set of possible realizations (a 'superpopulation') that are generated by a two-dimensional probability model. The variable may display spatial autocorrelation because it is a realization of a model where the random variables are spatially dependent. A conceptually simple example of such a model is the following, which is called the simultaneous (spatial) autoregressive model (Cliff and Ord, 1981). Let,

$$y(i) = \rho \sum_j w_{i,j} y(j) + e(i) \qquad [1]$$

where $y(i)$ is the observed value of the variable Y for area i (so this is a type (3) variable), and ρ is a parameter. $w_{i,j}$ is nonzero if area j ($j \neq i$) is a neighbor of area i (for example, area j shares a common boundary with area i; that is, area j is said to be at 'lag order 1' from area i) and is zero otherwise, and $w_{i,i} = 0$. The set, $\{e(i)\}$ is independent and identically distributed (iid) drawn from a normal distribution with mean zero and variance σ^2 (iid $N(0,\sigma^2)$). It can be shown that if $\rho > 0$, then Y will be positively spatially autocorrelated (and hence any map of realized values) because Y at i is a function of Y in the neighboring areas. For other examples of these types of spatial models, see Besag (1974, 1975) and for reviews, see Haining (2003), Cressie (1991), and Richardson (1992). There are other generators for spatially dependent variables, and such models may derive from models that incorporate both space and time into their specification (Berliner). In the context of a stochastic model, the term spatial autocorrelation has a closely proscribed meaning and is associated with the second-order moment (correlation) properties of the random variable.

When thinking about real processes (as opposed to purely mathematical models), then spatial autocorrelation in a variable can arise in a number of different ways. Here are some examples:

1. Many geographic variables display spatial continuity (temperature can be recorded everywhere and generally varies smoothly) in some cases referring to 'space-filling' objects (e.g., patches of forest, farmland, water bodies, etc.). If the spatial framework used to record the variable is at a smaller scale than the scale at which the variable varies across the region, then values in adjacent areas of the spatial framework will tend to be similar. Remote sensing provides an illustration of this point. If land use remains the same over large tracts of land, but a grid of small pixels is used to record land use, then neighboring pixels will tend to record the same value of the variable. The pixel values are (positively) spatially autocorrelated, and the smaller the pixels used to record values the more slowly will spatial autocorrelation decay with increasing lag (pixel step) distance.
2. Even if the scale of the spatial framework used for recording data values is the same as the scale at which the phenomena varies across the landscape, measurement error can introduce spatial autocorrelation. This can occur with certain types of surveys where recorders are given blocks to survey so that any operator-specific biases in the recording process will be similar within blocks and may therefore induce a form of spatial autocorrelation in the errors (Milne, 1959). Spatial autocorrelation in errors also arises with remotely sensed data due to the nature of the recording device on the satellite. Due to light scattering, what is recorded by the sensor in one pixel, which supposedly refers to one parcel of land on the ground, is actually a function of terrain characteristics in neighboring parcels (Foster, 1980).
3. Variable values may show spatial autocorrelation because they depend on an underlying process in which spatial relationships affect the way the process unfolds. There are a number of examples. Diffusion processes are processes where some attribute, like an infectious disease, spreads through a fixed population and the probability of a new individual becoming infected depends on spatial proximity to an already infected individual (Cliff and Ord, 1981). Spillover processes are processes where the attribute at some location is a function not only of conditions at that location but also of conditions in neighboring areas, for example, house prices as a function of local neighborhood conditions (Anas and Eum, 1984), or city income levels as a function of income levels in neighboring cities and rural areas arising because of consumer expenditure patterns (Haining, 1987). Interaction processes are where actions at one location influence actions at another location, such as price setting at one retail outlet influencing the prices set at other competitor sites (Haining, 1984). Dispersal processes are where individuals themselves disperse (like seeds around a parent plant or new settlers colonizing a region) so that population counts are autocorrelated (Cliff and Ord, 1981). In all these cases, at any point in time or when the process reaches some steady state, then the variable may show evidence of spatial autocorrelation.
4. The measured variable may not itself be the outcome of any of the foregoing situations but may be causally linked to another variable that is. In this case, the spatial structure of the dependent variable inherits the spatial structure of the independent variable to which it is related. Spatial autocorrelation in a variable may also arise because of spatial autocorrelation in underlying parameters. The incidence of a noninfectious disease like a respiratory condition may be autocorrelated because the relative risk of the disease (which is a parameter and is itself a function of environmental characteristics such as air quality) is spatially autocorrelated.
5. Spatial autocorrelation can be a property of a derived set of values such as the residuals from a regression model. In this case, spatial autocorrelation may have arisen because the regression model failed to include an important covariate that had a spatially autocorrelated structure. The spatial structure of the omitted independent variable is then inherited by the regression residuals.

For further discussion, see, for example, Fortin and Dale (2009).

Measuring and Testing for Spatial Autocorrelation

Autocorrelation Tests

There are several statistics for measuring and testing spatial autocorrelation. In the case of nominal data, the join-count test developed by Krishna Iyer (1949) may be used. The statistic is based on counting the number of joins of a particular color or colors on a map. Suppose each area is colored either black (B) or white (W). The join-count test is based on counting the number of BB, WW, or BW adjacencies on the map. If a map is positively spatially autocorrelated, then the number of BB and WW joins should be 'large' and the number of BW joins 'small.' If a map is negatively spatially autocorrelated, then the number of BW joins should be 'large' and the number of BB and WW joins 'small' (Haining, 2003). The terms 'large' and 'small' are determined with reference to the sampling distributions of these counts that would be expected if the black and white colors were allocated at random across the areas (Dacey, 1965).

In the case of ordinal- and interval-level data, the two tests most commonly used are the Geary test and the Moran test (Geary, 1954; Moran, 1948). The Geary test is based on differences between the values of adjacent areas: $(y(i) - y(j))^2$, where $y(i)$ and $y(j)$ denote the values observed in areas i and j, respectively. If areas i and j are adjacent and if the map is positively autocorrelated, then $(y(i) - y(j))^2$ should be 'small.' The Geary statistic is based on summing these squared differences across the map:

$$\sum_i \sum_j \delta_{i,j}(y(i) - y(j))^2 \qquad [2]$$

where $\delta_{i,j}$ is 1 if areas i and j are adjacent neighbors and zero otherwise ($\delta_{i,i} = 0$). The statistic, [2], will be 'small' in the case of positive spatial autocorrelation relative to the case of a random distribution of values. On the other hand, [2] will be 'large' in the case of negative spatial autocorrelation.

The Moran coefficient is based on cross products for adjacent areas:

$$\sum_i \sum_j \delta_{i,j}(y(i) - \bar{y})(y(j) - \bar{y}) \qquad [3]$$

where $\bar{y}$ is the mean of the observations on Y. In the case of a positively autocorrelated map pattern, [3] will be positive since usually both terms in the cross product will be positive or negative. In the case of a negatively autocorrelated map pattern, the terms will be negative, since if one of the terms is positive the other will usually be negative.

These statistics are appropriate to regular lattices but less appropriate for situations where the areal framework is irregular. Both [2] and [3] are invariant under certain transformations of the areal system but for any given area i some adjacencies are likely to be more important than others and should arguably therefore have greater weight in the calculation of any spatial autocorrelation statistic. In this case, the binary weight ($\delta_{i,j}$) is replaced by the more general weight ($w_{i,j}$). The weights may be chosen to reflect the topology of the areas or some hypothesis about the degree of contact between areas (Cliff and Ord, 1981). For example, $w_{i,j}$ can be defined to reflect the length of the shared common boundary between the two areas i and j, or the distance between the geographic centers of the two areas, or to reflect some measure of the intensity of interaction between the two areas in terms of, say, traffic or trade flow. For an extended discussion of these issues, see Dubin (2009).

The inference theories for the Geary and Moran statistics have been developed under two different conceptualizations of the null hypothesis of randomness or no spatial autocorrelation. In one case (the randomization hypothesis), the sampling distributions of the statistics are obtained assuming that the observed values are distributed at random across the region. In the other case (the normality assumption), the distributions of the statistics are obtained assuming that the observed values are independent drawings from a normal probability distribution with the same mean and variance for each area. The assumption of equal variance is a strong assumption and often questioned in the case of irregular areal partitions where population counts vary substantially between areas and where the population count appears in the denominator of the variable (Waldhor, 1996). For further discussion of inference issues for statistics from geographical data, see Brunsdon (2009).

There are other, simpler techniques for exploring spatial autocorrelation that can be useful (Haining, 2003). Simply mapping the variable can be informative as can graphing the scatterplot of paired data values $\{y(i), W^*y(i)\}$. $y(i)$ is the value observed in area i and $W^*y(i)$ is the value in the ith position of the matrix product of W^* with y. W^* is the matrix of general weights $\{w_{i,j}\}$ where each row has been standardized to sum to 1. So $W^*y(i)$ is a weighted average of the local values of y around i. If there is positive spatial autocorrelation present, the scatterplot will have a distribution of points sloping upward from the bottom left of the graph. If there is negative spatial autocorrelation, the scatterplot will have a distribution of points sloping downward from the top left of the graph. If the distribution of values is random, the scatterplot will have no discernable trend. Cressie (1984) describes other graphical methods for exploring the spatial structure.

Spatial Correlograms and Semivariograms

The tests for spatial autocorrelation described above can be implemented at any number of spatial scales. That is, the neighbors of any area i need not be restricted to the adjacent neighbors but can be extended to include more remote neighbors. It may be informative to examine spatial structure at a range of distances simultaneously. Spatial correlograms and semivariograms are plots of spatial structure at different distances. The correlogram at distance g has similarities with the Moran statistic. It is based on the expression:

$$\sum_i \sum_j w_{i,j}(g)(y(i) - \bar{y})(y(j) - \bar{y})$$

where $w_{i,j}(g)$ identifies all areas (j) that are at a distance g from the area i. Distance g can be defined in terms of a distance metric or in terms of 'lag distance,' that is the shortest path on a graph that connects adjacent areas (Haining, 2003). The spatial correlogram plot (with $g = 1, 2, 3, \ldots, G$ on the horizontal axis of the plot) thus provides a description of how spatial autocorrelation varies with distance.

The empirical semivariogram has similarities with the Geary statistic as it is based on the expression:

$$\sum_i \sum_j w_{i,j}(g)(y(i) - y(j))^2$$

This, like the correlogram, provides a plot of spatial structure with distance thereby showing how spatial structure varies with distance (Isaaks and Srivastava, 1989). Both the correlogram and the semivariogram may be used to suggest models for spatial variation; for examples in human geography, see the special issue of *Geographical Analysis* edited by Kerry et al. (2010).

A Test for Local Autocorrelation

The statistics in the Autocorrelation Tests section are often referred to as 'global' or 'whole map' tests of spatial autocorrelation because they compute an average measure using all the paired observations on the variable that are separated by the specified distance. But spatial autocorrelation might be a property of just a subset of the data, or the spatial autocorrelation structure may be different on different parts of the map. There might be a local tendency for similar values to cluster together, but this may not be a 'general' or 'global' property of the data.

Local tests for spatial autocorrelation compute the probability of getting the observed collection of values in a subset of the region if all the data values had been distributed at random over the area. The local Moran statistic ($I(i)$) is based on (Anselin, 1995):

$$(y(i) - \bar{y}) \sum_j w_{i,j}(y(j) - \bar{y}) \quad i = 1, \ldots, n$$

where n is the number of areas so that there are as many statistics and hence tests as there are cases. A high positive value of $I(i)$ signals a local spatial clustering of similar values, whereas a high negative value signals a local spatial clustering of dissimilar values.

Spatial Autocorrelation and Spatial Data Analysis

The presence of spatial autocorrelation in a variable has implications for a number of different areas of data collection and data analysis. These include spatial sampling, map interpolation, exploratory spatial data analysis, and confirmatory analysis including modeling. Each of these areas is briefly discussed in turn.

Spatial Sampling

Spatial sampling is the process of collecting data according to some specified set of rules and then using the data to make inferences about the population from which the sample has been drawn. Consider the case where the purpose of sampling is to estimate the proportion of an area covered by a particular type of land use. In this case, we do not adopt a 'superpopulation' view of the problem but rather focus on the 'here and now' of the land use geography. A conventional estimator for the quantity is the sample mean:

$$\bar{y} = \frac{1}{n} \sum_i y(i)$$

where n is the size of the sample and $y(i)$ is 1 if sample site i is occupied by the particular land use, and zero otherwise. One sampling decision to be taken is whether to take a random sample, a stratified random sample, or some form of systematic sample. The spatial autocorrelation in the land use surface has an important influence on the relative efficiency of these three designs because, as noted above, the presence of spatial autocorrelation signals that data values recorded at one location carry information about data values at nearby locations. The stronger the spatial autocorrelation on the surface, the more important it is to select sample sites that minimize the duplication of information within the sample. Random sampling is the least efficient sample design because there is no guarantee that sample points will not be clustered in particular areas of the map and hence their attribute values strongly autocorrelated. (In addition random sampling may leave large unsampled areas). Stratified random sampling overcomes the worst features of random sampling by ensuring coverage, but sample sites can still be adjacent to one another on either side of a strata boundary. Systematic sampling ensures a minimum separation distance between sample sites (Dunn and Harrison, 1993).

Map Interpolation

A range of estimators have been constructed to estimate values at specified locations where no data have been collected. One of the more widely used methods, the kriging estimator, has the generic form:

$$\hat{y}(k) = \sum_j \lambda_{k,j} y(j)$$

where $\hat{y}(k)$ is the estimate at location k where currently no datum is available (Berliner). The values $\{y(j)\}$ are the sample data, and $\{\lambda_{k,j}\}$ are coefficients that are a function of the spatial autocorrelation both between locations k and j and among the sample values (Isaaks and Srivastava, 1989). In this case, the presence of spatial autocorrelation is used to construct estimates of missing or unobserved values; in addition, confidence intervals can be attached to the estimates. This confidence interval is a function of sample density in the vicinity of the site to be interpolated. Confidence intervals will be large if the site k is far from sample points for which information is available and small if there are sample data nearby.

Exploratory Spatial Data Analysis

Disease atlases are constructed to represent the geographic variability of a disease. Maps of standardized mortality ratios (SMRs), for example, display the geography of relative risk for a particular disease and may be used to provide an indication of where the disease risk might be highest – perhaps as a preliminary to exploring for particular risk factors. Maps constructed using small geographic areas suffer from a number of problems – the estimated ratios tend to be

unstable with extreme rates often associated with areas with the smallest populations. Bayes adjustment methods are often used to adjust such standardized ratios. These methods incorporate information from the rest of the map in order to reduce some of the random variation that is present in each area's SMR. A subset of these methods does not just use the SMRs from the rest of the map but places greatest weight on those estimates closest to the area to be adjusted. For this adjustment, spatial process models are used and incorporate assumptions about the spatial autocorrelation in the relative risks into the adjustment process (Mollie, 1996).

Other, simpler smoothing methods may also be employed to provide evidence on broad patterns in a spatial data set, for example, mean and median smoothers (Haining et al., 1998). These techniques all implicitly exploit some aspect of the concept of spatial autocorrelation – that places near to an area, i, provide information on the value at i.

Information Loss in Statistical Tests

Spatial autocorrelation can undermine the use of classical statistical techniques (based on the assumption of data independence). We consider one example. The correlation coefficient (r) is a common statistic for measuring the linear relationship between two variables (X and Y). The Pearson correlation coefficient varies between -1 and $+1$, with $+1$ signifying a perfect positive relationship between X and Y (as X increases, Y increases). The inference theory for the correlation coefficient is based on:

$$(n-2)^{1/2}\,|\widehat{r}|\,\left(1-\widehat{r}^2\right)^{-1/2} \qquad [4]$$

where n is the sample size and $\widehat{r}$ is the estimated correlation coefficient. Under the null hypothesis that the correlation in the population is zero, [4] is t distributed with $(n-2)$ degrees of freedom. However, the amount of statistical information carried by the n observations when the data are (positively) spatially autocorrelated is less than would be the case were the n observations to be independent. In the terminology of Clifford and Richardson (1985), the 'effective sample size' is less than the actual sample size n so the degrees of freedom for the test is less than n. The solution to the problem is to compute this effective sample size (N') which is obtained from the spatial correlogram (see Section Map Interpolation). Then in [4], N' replaces n and the statistic is t distributed with $(N'-2)$ degrees of freedom. For other examples of the effect of spatial autocorrelation on classical statistical techniques, see Haining (2003).

Modeling Spatial Variation

The classical normal regression model is defined as:

$$y(i) = \sum_{k=0}^{K} \beta_k X_k(i) + e(i) \quad i = 1, \ldots, n$$

where y denotes the dependent variable, n is the number of sample sites, X_0 is a vector of 1s, and the X_k ($k = 1, \ldots, K$) denote the set of K independent variables. The model parameters to be estimated are denoted β_0, β_1, …, β_K where β_0 denotes the intercept parameter and the rest denote the regression parameters. The $e(1)$, …, $e(n)$ are iid $N(0,\sigma^2)$ errors. One of the assumptions of normal regression is that the errors are independent. If this assumption is violated (as, for example, when the errors are in fact spatially autocorrelated), tests of significance on the parameters of the independent variables are affected (the risk of a type I error is increased) and the usual goodness of fit measure is inflated. This assumption has to be tested using the regression residuals (which are estimates of the errors). In the case of spatial data, an adaptation of the Moran test is often used, although other tests are available (Anselin, 1988, 2009). If the null hypothesis of no spatial autocorrelation is rejected, then remedial action is needed to avoid misspecification error. One approach is to seek out other independent variables that, because they were originally (and wrongly) excluded and spatially autocorrelated, caused the residual autocorrelation. Other forms of model specification may also need to be considered which include interaction terms, diffusion process terms, or spillover terms (see Section Measuring and Testing for Spatial Autocorrelation). Thus a respecified regression model might include spatially lagged terms of the dependent variable, one or more of the independent variables or a spatial model imposed on the errors. (Examples of these and other models are described in Ord, 1975; Anselin, 1988, 2009; Haining, 2003; Cressie, 1991.) Residual spatial autocorrelation undermines other forms of regression analysis including Poisson regression used to model small area count data (e.g., numbers of cases of a disease by area). A solution to this problem may entail the use of a spatial random effects term in the specification of the Poisson mean (Griffith and Haining, 2006; Law et al., 2006). A common underlying thread here is the problem of 'information loss,' associated with spatial autocorrelation, discussed in the Information Loss in Statistical Tests Section.

See also: Spatial Analysis in Geography; Spatial Association, Measures of; Spatial Data; Spatial Pattern, Analysis of; Spatial Sampling.

Bibliography

Anas, A., Eum, S., 1984. Hedonic analysis of a housing market in disequilibrium. Journal of Urban Economics 15, 87–106.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L., 1995. Local indicators of spatial association – LISA. Geographical Analysis 27, 93–115.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society B 36, 192–236.

Besag, J., 1975. Statistical analysis of nonlattice data. The Statistician 24, 179–196.

Brunsdon, C., 2009. Statistical inference for geographical processes. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 207–224.

Cliff, A., Ord, J., 1981. Spatial Processes: Models and Applications. Pion, London.

Clifford, P., Richardson, S., 1985. Testing the association between two spatial processes. Statistics and Decisions (Suppl. 2), 155–160.

Cressie, N., 1984. Towards resistant geostatistics. In: Verly, G., et al. (Eds.), Geostatistics for Natural Resources Characterization. Reidel, Dordrecht, The Netherlands, pp. 21–44.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Dacey, M., 1965. A review of measures of contiguity for two and k-color maps. In: Berry, B., Marble, D. (Eds.), Spatial Analysis: A Reader in Statistical Geography. Englewood Cliffs, NJ, pp. 479–495.

Dubin, R., 2009. Spatial weights. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 125–157.

Dunn, R., Harrison, A., 1993. Two dimensional systematic sampling of land use. Applied Statistics 42, 585–601.

Fortin, M.-J., Dale, M., 2009. Spatial autocorrelation. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 89–103.

Foster, B., 1980. Urban residential ground cover using Landsat digital data. Photogrammetric Engineering and Remote Sensing 46, 547–558.

Geary, R., 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician 5, 115–145.

Geostatistical methods in geography. Part I: applications in human geography. In: Kerry, R., Haining, R., Oliver, M. (Eds.), Geographical Analysis 42 (1), 5–117.

Griffith, D.A. (Ed.), 2009. A 40th anniversary celebration of A. Cliff and J. Ord, 1969, the problem of spatial autocorrelation. Geographical Analysis 41 (4), 343–363.

Griffith, D., Haining, R., 2006. Beyond mule kicks: the Poisson distribution in geographical analysis. Geographical Analysis 38 (2), 123–139.

Haining, R., 1984. Testing a spatial interacting-markets hypothesis. Review of Economics and Statistics 66, 576–583.

Haining, R., 1987. Small area aggregate income models: theory and methods with an application to urban and rural income data for Pennsylvania. Regional Studies 21, 519–530.

Haining, R., 2003. Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge, UK.

Haining, R., 2009. The special nature of spatial data. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 5–23.

Haining, R., Wise, S., Ma, J., 1998. Exploratory spatial data analysis in a GIS environment. The Statistician 47, 457–469.

Isaaks, E., Srivastava, R., 1989. An Introduction to Applied Geostatistics. Oxford University Press, Oxford, UK.

Krishna Iyer, P., 1949. The first and second moments of some probability distributions arising from points on a lattice, and their application. Biometrika 36, 135–141.

Law, J., Haining, R., Maheswaran, R., Pearson, T., 2006. Analyzing the relationship between smoking and coronary heart disease at the small area level: a Bayesian approach to spatial modeling. Geographical Analysis 38 (2), 140–159.

Milne, A., 1959. The centric systematic area sample treated as a random sample. Biometrics 15, 270–297.

Mollie, A., 1996. Bayesian mapping of disease. In: Gilks, G., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 359–379.

Moran, P., 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society, Series B 10, 243–251.

Ord, J., 1975. Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70, 120–126.

Richardson, S., 1992. Statistical methods for geographical correlation studies. In: Elliott, P., Cuzick, J., English, D., Stern, R. (Eds.), Geographical and Environmental Epidemiology: Methods for Small Area Studies. Oxford University Press, Oxford, UK, pp. 181–204.

Ripley, B., 1981. Spatial Statistics. Wiley, New York.

Waldhor, T., 1996. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. Statistics in Medicine 15, 887–892.

Spatial Cognition

Daniel R Montello, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial cognition concerns the study of knowledge and beliefs about spatial properties of objects and events in the world, and as such objects and events are represented in maps, language, and other symbol systems. Geographers study spatial cognition to address questions such as how spatial knowledge and beliefs are acquired and develop over time; the nature of spatial knowledge structures and processes; how people navigate and stay oriented in space; how people use language to communicate with each other about space; and how aspects of spatial knowledge and reasoning are similar or different among individuals or groups.

Spatial cognition concerns the study of knowledge and beliefs about spatial properties of individual objects and events, or collections of objects and events. Cognition is about knowledge: its acquisition, storage and retrieval, manipulation, and use by humans, nonhuman animals, and intelligent machines. Cognitive systems include sensation and perception, thinking, mental imagery, memory, learning, language, reasoning, and problem solving. Spatial properties include location, size, distance, direction, separation and connection, shape, pattern, movement trajectory, and more. These properties characterize the small and the large, but geographers focus on scales relevant to human activities on the earth's surface, such as temporary travel, migration, settlement, and economic activity. As such, geographers are mostly interested in the spaces of buildings, neighborhoods, parks, forests, mountain ranges, crop fields, cities, provinces, and countries. Geographers are also interested in the spaces of external symbolic representations of earth places and features, such as cartographic maps, language, and other geographic information displays. Spatial cognition plays a central role in many human activities and helps solve many human problems (Allen, 2007; Montello and Raubal, 2012), several of which geographers study. Cognition of and with spatiality is a fundamental component of human experience and is functional and relevant in many situations, both everyday and exceptional. The study of spatial cognition in geography sheds light on questions such as how spatial knowledge and beliefs are acquired and develop over time; the nature of spatial knowledge structures and processes; how people navigate and stay oriented in space; how people use language to communicate with each other about space and spatiality; and how aspects of spatial knowledge and reasoning are similar or different among individuals or groups. Among other issues, future research will concentrate on how understanding spatial cognition can improve the design and implementation of information technologies and on how technologies will change the cognition of space and place.

In humans, cognitive structures and processes are part of the mind, which emerges from a brain and nervous system inside of a human body that exists in a social and physical world. Within the last couple of decades, considerable advances have been made in understanding the brain bases for spatial cognition, including its neural and chemical underpinnings (Byrne et al., 2007; Maguire et al., 2000). Many of the advances in this area of research have resulted from new technologies, particularly functional magnetic resonance imaging, for scanning the ongoing real-time activities of the brains of healthy individuals engaged in cognitive acts, such as attempting to figure out where they are located in space. Spatial cognitive neuroscience continues, constituting some of the most important ongoing scientific research on spatial cognition today. Even geographers have started to explore these approaches to researching spatial cognition (Lobben et al., 2009). However, it is important to recognize that an increased understanding of the brain bases for spatial cognition does not eliminate the need to understand spatial cognition at functional, semantic, experiential, computational, environmental, and social levels.

The Geographic Study of Spatial Cognition

Geographers study spatial cognition for three reasons (Downs and Stea, 1973; Golledge, 1987). First and foremost, spatial cognition provides questions of geographic interest in their own right, given that cognition about space and place is an expression of human–environment or human–earth relations. Second, geographers have hoped that spatial cognition would improve their understanding of traditional phenomena of interest; for example, where people choose to shop should depend in part on their beliefs about distances and road connections. Finally, there has been an interest in improving the use and design of maps and other geographic information products which depend in part on human understanding of depicted spatial relations.

In addition to geographers and cartographers, researchers from many different disciplines study spatial cognition, including psychologists, linguists, biologists, computer scientists, architects and planners, anthropologists, and philosophers. Overviews of spatial cognition from various disciplinary perspectives, including geography and cartography, can be found in Allen (2004), Kitchin and Blades (2002), Lloyd (1997), McNamara et al. (2008), and Waller and Nadel (2012). While overlapping and interacting with these disciplines, the geographer's interest in spatial cognition is distinct in some ways. Geographers are interested in the earth

as the home of humanity. Their concern is with the cognition of humans, not the cognition of barn swallows or undersea robots. Geographers focus on spatial scales relevant to human activities on the earth's surface, such as daily travel, migration, settlement, and economic activities; a geographer wants to understand human conceptualizations of the layout of a city or a region, not of an atom, a desktop, or the solar system. Finally, the geographer ultimately desires to describe, predict, and explain external human–environment relationships, and focuses on internal mental and neuropsychological structures and processes only insofar as such internal structures and processes help in the understanding of external relationships.

History of Spatial Cognition in Geography

The notion that a subjective understanding of space partially mediates human relationships with an objective physical world is undoubtedly quite old. Within the modern discipline of geography, the study of spatial cognition first appeared with the work on geographic education early in the twentieth century. However, interest in spatial cognition flowered during the 1950s and 1960s. The subdiscipline of experimental cartography emerged with the insight that maps could be improved by understanding how humans perceive and think about information depicted in map symbols. In an attempt to explain human responses to floods and other natural hazards, the subdiscipline of environmental perception began to study people's subjective beliefs about the occurrence and consequences of hazard events. Most importantly for the study of spatial cognition, behavioral geographers began developing theories and models of the human reasoning and decision making involved in spatial behavior, for example, migration, vacationing, and trip scheduling. This approach gained insight from the work of Lynch (1960), a planner who argued that 'images' of cities guide people's behavior and experiences of those cities. All of these historical developments constituted a new focus on disaggregate or microscale approaches within human geography, approaches that focused on the geographic behavior of individuals as an accompaniment to traditional approaches that focused on human settlements and larger aggregations of people.

Topics in the Geographic Study of Spatial Cognition

The study of spatial cognition in geography sheds light on questions such as how spatial knowledge and beliefs are acquired and develop over time; the nature of spatial knowledge structures and processes; how people navigate and stay oriented in space; how people use language to communicate with each other about space; and how aspects of spatial knowledge and reasoning are similar or different among individuals or groups.

Acquisition and Development

Humans acquire spatial knowledge and beliefs directly as they move about the world via sensorimotor systems; they move their bodies in and through space, they see spatial properties and relations, they feel them with their hands, they hear them, and so on. People also acquire spatial knowledge indirectly via static and dynamic symbolic media such as maps and imagery, graphs, 3D models, animations, natural language, and virtual reality. Geographers are interested in how different media have consequences for the nature of acquired knowledge.

Spatial knowledge changes over time, through processes of learning and development. Changes in the child's spatial knowledge and reasoning (Newcombe and Huttenlocher, 2000) as well as changes an adult experiences while visiting a new place for the first time are of interest to geographers. Both physical maturation and experience influence development. A person's activity space – the set of paths, places, and regions traveled on a regular basis – is an important example of spatial experience that influences people's knowledge of space and place, no matter what their age is. Most people know the areas around their homes or work places most thoroughly, for example (as in the anchor-point theory of Couclelis et al., 1987).

A widely accepted model of spatial learning suggests that spatial knowledge of places develops in a sequence of three stages or elements (Siegel and White, 1975). First is landmark knowledge, based on the unique features or views that identify a place. Second is route knowledge, based on travel routines that connect ordered sequences of landmarks. The final element is survey knowledge, that is, quantitative knowledge of two-dimensional layouts that includes the simultaneous interrelations of locations; survey knowledge would support detouring and shortcutting. However, increasing evidence (Ishikawa and Montello, 2006) indicates that some people acquire survey knowledge surprisingly quickly, essentially as soon as they encounter a new place; at the same time, some people very slowly or never develop detailed or comprehensive survey knowledge of their surrounds.

Exposure to maps of places clearly advances the development of survey knowledge for most people. An interesting proposal by Uttal (2000) even suggests that we conceptualize earth-surface layouts as survey representations only because we are exposed so early and often to cartographic maps in our culture. Either way, geographers are interested in how people acquire spatial knowledge not only directly, but also indirectly, from maps, satellite imagery, globes, and other iconic symbol displays. Notably, people can learn spatial properties of very large or small spaces (continents, molecules) only via indirect sources such as maps, whereas in many other cases, people learn spatial properties via both direct and indirect sources (Thorndyke and Hayes-Roth, 1982). In other cases, people even acquire nonspatial knowledge by interpreting spatial displays that metaphorically represent nonspatial information (Skupin and Fabrikant, 2008).

Spatial Knowledge Structures and Processes

The metaphor of a cognitive map was coined by Tolman in a 1948 paper to refer to internally represented spatial models of the environment. The cognitive (or mental) map includes knowledge of landmarks, route connections, and distance

and direction relations; nonspatial attributes and emotional associations are stored as well (Downs and Stea, 1973). However, in many ways, the cognitive map is not like a cartographic 'map in the head' (Tversky, 2000). It is not a unitary integrated representation, but consists of stored discrete pieces including landmarks, route segments, and regions. The separate pieces are often partially linked or associated so as to represent hierarchies, such as the location of a place inside of a larger region (McNamara et al., 1989; Stevens and Coupe, 1978). Furthermore, spatial knowledge is not well modeled by metric geometries such as the Euclidean geometry of high school math. The ways in which spatial knowledge is internally represented in human minds lead to certain patterns of distortions in the way people answer spatial questions. For example, people often believe the distance from place A to B is different than from B to A. Turns are frequently adjusted in memory to be more like straight lines or right angles. At larger scales, most people have somewhat distorted ideas about the sizes and locations of continental land masses on the earth; for example, South America is thought to be due south of North America when it is actually rather far to the east as well. Friedman and Brown (2000) showed convincingly that people organize their understanding of spatial relations at global scales largely in terms of regions that simplify spatial content by exaggerating the similarity of latitudes of places within a region while exaggerating differences in latitudes of places between regions (Figure 1). Interestingly, these global regions in some cases correspond to countries or continents, and sometimes they do not.

Navigation and Orientation

Humans travel over the earth's surface to reach destinations. This typically requires planning and the ability to stay oriented while moving. Navigation is this coordinated and goal-directed travel through space. It can be understood to consist of two components, locomotion and wayfinding (Montello, 2005). Locomotion refers to guiding oneself through space in response to local sensorimotor information in the immediate surrounds, and includes such tasks as identifying surfaces of support, avoiding obstacles, and moving toward visible landmarks. Locomotion generally occurs without the need for an internal model or cognitive map of the environment. Wayfinding refers to the planning and decision making that allows one to reach a destination that is not in the immediate sensory field, and includes such tasks as choosing efficient routes, scheduling destination sequences, orientating to nonlocal features, and interpreting verbal route directions. Wayfinding tasks generally do require a cognitive or cartographic map of the environment. Various locomotion and wayfinding tasks vary greatly in their demand on attentional capacity. Following a familiar route is not very demanding on attention; finding one's way in a new city may be quite demanding.

Orientation is knowing "where you are," although the precision and comprehensiveness of this knowing varies greatly in different situations and for different people. Two fundamental processes are involved in orientation during navigation (Golledge, 1999). The first process is the recognition of external features or landmarks, also known as pilotage. In some cases,

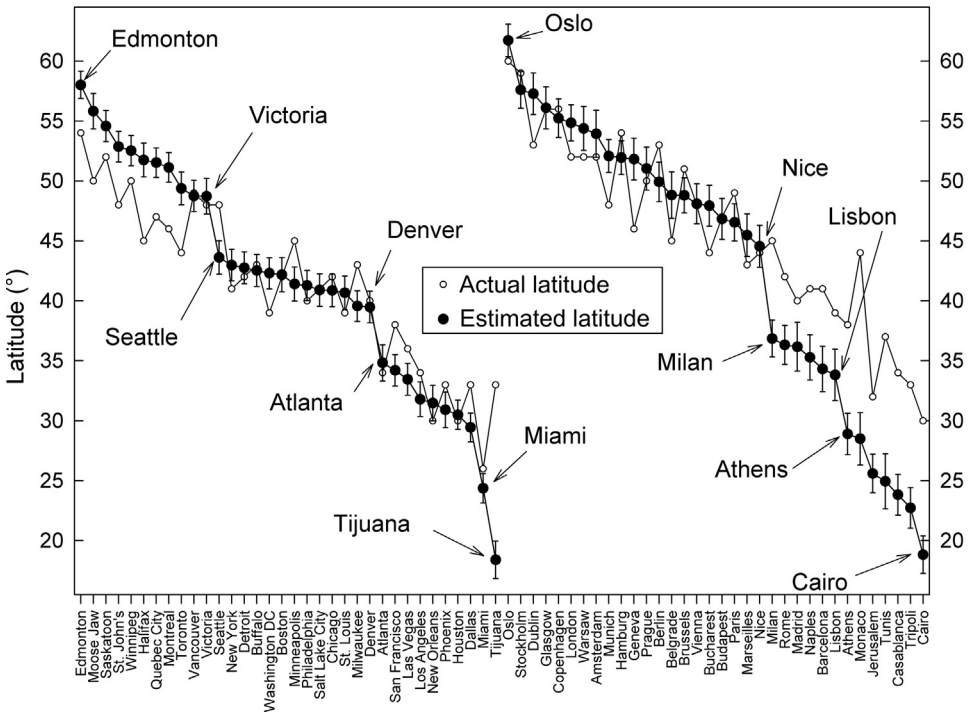

Figure 1 Patterns of latitude estimates for world cities indicate regional organization and reasoning for spatial knowledge about world geography. Cities on the abscissa are ordered according to estimated latitudes in the western and eastern hemispheres. Reproduced from Friedman, Alinda, Brown, Norman R., 2000. Reasoning about geography. Journal of Experimental Psychology: General 129, 193–219.

the recognized landmark is the goal, but more commonly the landmark aids orientation by serving as a key to the knowledge of spatial relations stored in an internal cognitive or external cartographic map. The second process is dead reckoning, updating a sense of orientation by integrating information about movement speed, direction, and/or acceleration without reference to recognized features.

People also use symbolic media such as maps to navigate and stay oriented. Maps used for navigation show robust alignment effects, a common confusion when using navigation maps with the top of the map not being the direction in the surrounds that one is facing when looking at the map. Alignment effects results from the fact that spatial perception is orientation dependent; objects and pictures have a subjective top and bottom, and forwards and backwards, in the perceptual field and in memory. In the case of road or trail maps, many people turn the map to bring about alignment. Digital navigation systems typically turn this way automatically. Unfortunately, this option is not available with ubiquitous 'You-Are-Here' maps, which are surprisingly often mounted in the environment in a misaligned way with respect to the viewer's perspective. The resulting alignment effect is the fact that most people interpret misaligned maps more slowly, with greater error, and/or with dissatisfaction (Levine et al., 1984). While an annoyance in many situations, misaligned emergency maps placed in public buildings are potentially disastrous.

Spatial Communication via Language

Spatial information is often communicated verbally, in natural language such as English or Chinese. People give and receive verbal route directions, read spatial descriptions contained in stories, and increasingly interact with computer systems via verbal queries. There are at least two notable characteristics of the way language expresses spatial information (Bloom et al., 1996; Mark and Frank, 1991). One is that language expresses mostly nonquantitative or imprecise ('fuzzy') quantitative information about space. Statements about connections and approximate location are more important than precise statements. For example, people say "turn left at the gas station" rather than "turn 80° after you go 0.6 miles." Great precision is typically unnecessary or even confusing. A second characteristic is that various aspects of the context of communication are critical in interpreting spatial language. Context is provided by knowledge of who is speaking, where they are, physical features in the situation, the previous topic of conversation, and so on. "The bicycle is near the church" is generally understood differently from "France is near England."

Individual and Group Similarities and Differences

No two individuals know exactly the same things or reason in exactly the same way about space and place. Some people are better at tasks such as wayfinding, learning spatial layouts, or reading maps. In some cases, there may be different ways to think about such problems, all of which may be effective. In such cases, one might speak of stylistic differences in spatial cognition rather than ability differences. Geographers and others are interested in measuring and explaining such individual differences. Traditional psychometric paper-and-pencil tests of spatial abilities tend not to measure these differences very well (Allen et al., 1996).

Many factors may contribute to variations in spatial cognition: body size, age, education, expertise, sex, social status, language, culture, and more (Levinson, 2003; Liben et al., 1981). A first goal for geographic research is to measure and document ways these factors might covary with spatial cognition. Females and males appear to differ in some ways in their spatial abilities and styles, for example, but not in others (Montello et al., 1999). In addition to describing such covariations, geographers have the research goal of distinguishing their underlying causes. For example, the two sexes likely differ in their spatial cognition in part because of different travel experiences afforded them as children by their socialized roles. Explanations for such covariations are generally quite difficult to determine, however, as one cannot readily do randomized experiments on person characteristics such as age, sex, culture, and activity preferences.

Technologies and the Future of Spatial Cognition Research

As stated earlier in this article, future research will continue examining the interactions of new information technologies and spatial cognition. Understanding spatial cognition should improve technology, increasing its ease of use, its effectiveness, its profitability, and its equitable distribution in society. Relevant technologies include mapping and route direction sites on the Web, geographic information system (GIS), global positioning system (GPS)-enabled navigation systems in automobiles and on cell phones, the computer-simulated worlds known as virtual reality or virtual environments, and location-based systems and services more generally (Montello and Raubal, 2012). A host of spatial-cognition questions are inspired by efforts to improve the effectiveness and efficiency of information systems (Hirtle, 2011). How can complex spatial and thematic information be depicted to promote comprehension and effective decision making, whether through maps, graphs, verbal descriptions, or animations? Although spatial cognition researchers believe in the practical usefulness of their research (Allen, 2007), there are not many clear demonstrations yet of its widespread application.

Furthermore, information technologies are likely to change spatial cognition, and our experiences and behaviors in space and place. Researchers will ask how exposure to new spatial information technologies will alter human ways of perceiving and thinking about the world. Will the regular use of GPS-enabled navigation systems diminish people's ability to maintain orientation by using perceptual-motor and cognitive systems? Will this be offset by the great reductions in lost time and lost lives resulting from using new technologies? The potential for technologies to enhance education is considerable (Committee on Support for Thinking Spatially: The Incorporation of Geographic Information Science Across the K–12 Curriculum, the National Research Council, 2006), but does the drive to integrate spatial technologies like GIS into educational settings run a risk of replacing thinking rather than

enhancing it? Ongoing and future research will shed light on these questions and more.

See also: Cartographic Visualization; Cognitive Maps; Environmental Cognition, Perception, and Attitudes; Hippocampus and Related Structures; Human Spatial Orientation, Neural Basis of; Mental Imagery: Visual Cognition; Navigation in Virtual Space: Psychological and Neural Aspects; Navigation; Neural Representations of Direction (Head Direction Cells); Space, Linguistic Expression of; Spatial Cognition During Infancy and Early Childhood Across Cultures, Development of; Spatial Decision Support Systems; Spatial Navigation; Spatial Thinking in the Social Sciences; Transportation, Psychology of; Virtual Reality and Spatial Cognition.

Bibliography

Allen, Gary L. (Ed.), 2004. Human Spatial Memory: Remembering Where. Lawrence Erlbaum Associates, Mahwah, NJ.

Allen, Gary L. (Ed.), 2007. Applied Spatial Cognition: From Research to Cognitive Technology. Lawrence Erlbaum Associates, Mahwah, NJ.

Allen, Gary L., Kirasic, Kathleen C., Dobson, Shannon H., Long, Richard G., Beck, Sharon, 1996. Predicting environmental learning from spatial abilities: an indirect route. Intelligence 22, 327–355.

Bloom, Paul, Peterson, Mary A., Nadel, Lynn, Garrett, Merrill F. (Eds.), 1996. Language and Space. The MIT Press, Cambridge, MA.

Byrne, Patrick, Becker, Suzanna, Burgess, Neil, 2007. Remembering the past and imagining the future: a neural model of spatial memory and imagery. Psychological Review 114, 340–375.

Committee on Support for Thinking Spatially: The Incorporation of Geographic Information Science Across the K–12 Curriculum, the National Research Council, 2006. Learning to Think Spatially. The National Academies Press, Washington, DC.

Couclelis, Helen, Golledge, Reginald G., Gale, Nathan, Tobler, Waldo, 1987. Exploring the anchor-point hypothesis of spatial cognition. Journal of Environmental Psychology 7, 99–122.

Downs, Roger M., Stea, David (Eds.), 1973. Image and Environment. Aldine, Chicago.

Friedman, Alinda, Brown, Norman R., 2000. Reasoning about geography. Journal of Experimental Psychology: General 129, 193–219.

Golledge, Reginald G., 1987. Environmental cognition. In: Stokols, Daniel, Altman, Irwin (Eds.), Handbook of Environmental Psychology. Wiley, New York, pp. 131–174.

Golledge, Reginald G., 1999. Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Johns Hopkins University Press, Baltimore, MD.

Hirtle, Stephen C., 2011. Geographical Design: Spatial Cognition and Geographical Information Science. Morgan & Claypool Publishers, San Rafael, CA.

Ishikawa, Toru, Montello, Daniel R., 2006. Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. Cognitive Psychology 52, 93–129.

Kitchin, Rob, Blades, Mark, 2002. The Cognition of Geographic Space. IB Tauris Publishers, New York and London.

Levine, Marvin, Marchon, Iris, Hanley, Gerard, 1984. The placement and misplacement of you-are-here maps. Environment and Behavior 16, 139–157.

Levinson, Stephen C., 2003. Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge, UK, New York.

Liben, Lynn S., Patterson, Arthur H., Newcombe, Nora, 1981. Spatial Representation and Behavior Across the Life Span. Academic Press, New York.

Lloyd, Robert, 1997. Spatial Cognition: Geographic Environments. Kluwer Academic, Dordrecht, The Netherlands.

Lobben, Amy, Lawrence, Megan, Olson, Judy M., 2009. fMRI and human subjects research in cartography. Cartographica: The International Journal for Geographic Information and Geovisualization 44, 159–169.

Lynch, Kevin, 1960. The Image of the City. MIT Press, Cambridge, MA.

Maguire, Eleanor A., Gadian, David G., Johnsrude, Ingrid S., Good, Catriona D., Ashburner, John, Frackowiak, Richard S.J., Frith, Christopher D., 2000. Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences of the United States of America 97, 4398–4403.

Mark, David M., Frank, Andrew U. (Eds.), 1991. Cognitive and Linguistic Aspects of Geographic Space. Kluwer Academic, Dordrecht, The Netherlands.

McNamara, Timothy P., Hardy, James K., Hirtle, Stephen C., 1989. Subjective hierarchies in spatial memory. Journal of Experimental Psychology: Learning, Memory and Cognition 15, 211–227.

McNamara, Timothy P., Sluzenski, Julia, Rump, Björn, 2008. Human spatial memory and navigation. In: Roediger, Henry L.I. (Ed.), Cognitive Psychology of Memory, vol. 2. Elsevier, Oxford, pp. 157–178.

Montello, Daniel R., 2005. Navigation. In: Shah, Priti, Miyake, Akira (Eds.), The Cambridge Handbook of Visuospatial Thinking. Cambridge University Press, Cambridge, pp. 257–294.

Montello, Daniel R., Lovelace, Kristin L., Golledge, Reginald G., Self, Carole M., 1999. Sex-related differences and similarities in geographic and environmental spatial abilities. Annals of the Association of American Geographers 89, 515–534.

Montello, Daniel R., Raubal, M., 2012. Functions and applications of spatial cognition. In: Waller, David, Nadel, Lynn (Eds.), Handbook of Spatial Cognition. American Psychological Association, Washington, DC, pp. 249–264.

Newcombe, Nora S., Huttenlocher, J., 2000. Making Space: The Development of Spatial Representation and Reasoning. The MIT Press, Cambridge, MA.

Siegel, Alexander W., White, Sheldon H., 1975. The development of spatial representations of large-scale environments. In: Reese, Hayne W. (Ed.), Advances in Child Development and Behavior, vol. 10. Academic, New York, pp. 9–55.

Skupin, André, Fabrikant, Sara Irina, 2008. Spatialization. In: Wilson, John P., Fotheringham, A. Stuart (Eds.), The Handbook of Geographic Information Science. Blackwell Publishing, Malden, MA, pp. 61–79.

Stevens, Albert, Coupe, Patty, 1978. Distortions in judged spatial relations. Cognitive Psychology 10, 422–437.

Thorndyke, Perry W., Hayes-Roth, Barbara, 1982. Differences in spatial knowledge acquired from maps and navigation. Cognitive Psychology 14, 560–589.

Tolman, Edwin C., 1948. Cognitive maps in rats and men. Psychological Review 55, 189–208.

Tversky, Barbara, 2000. Remembering spaces. In: Tulving, Endel, Craik, Fergus I.M. (Eds.), The Oxford Handbook of Memory. Oxford University Press, Oxford, pp. 363–378.

Uttal, David H., 2000. Seeing the big picture: map use and the development of spatial cognition. Developmental Science 3, 247–264.

Waller, David, Nadel, Lynn (Eds.), 2012. Handbook of Spatial Cognition. American Psychological Association, Washington, DC.

Spatial Cognition during Infancy and Early Childhood Across Cultures, Development of

Jordy Kaufman and Mahla Cameron-Bradley, Swinburne University of Technology, Hawthorn, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A fundamental question for developmental science is: how do young children represent the space around them. This question has been empirically approached from a number of angles. In this article, we describe some of these approaches and report on key findings regarding spatial cognition development in infants and young children. The specific areas covered here include developmental studies of spatial categorization in infants, spatial frames of reference available to infants and young children, geometric cue use by toddlers, and cultural variations in spatial thinking and spatial language.

Introduction

Spatial cognition involves the attainment, use, and adjustment of knowledge about spatial relationships and environments. The development of spatial cognition abilities is essential for a variety of skills essential for daily functioning (e.g., finding one's way around) and for higher-order processes (e.g., advanced mathematics). As such, an understanding of what young children understand about three-dimensional space and how this knowledge arises is of fundamental interest to developmental science.

Since space is universal to all organisms, one could hypothesize that knowledge of the metric properties of space should be innately available to infants. Alternatively, one could assume that space and all of the external world become represented as separate from the infant over time as a result of the infant's interaction with their world.

These questions have made the study of children's spatial representations an area of intense interest in the broader field of developmental science. In this article, we describe some of the major areas of study and key findings with respect to spatial cognition development in infants and young children. These areas include developmental studies of spatial categorization, spatial frames of reference, geometric cue use, and cultural variations.

Spatial Categorization in Infancy

'Spatial categorization' is a term that has been used to describe a variety of concepts. Here we focus on how the developing mind categorizes the relationships that are held between two objects. For example, adults often describe spatial relationships between objects as '*above/below*,' 'to the *right/left* of,' and '*inside/outside* of.' Developmental scientists have extensively studied the ability of preverbal infants to mentally classify such relationships.

Categorization of Simple Relationships

Beginning with the seminal work of Quinn (1994), the bulk of infant categorization studies have relied on experiments in which infants were habituated or familiarized with a particular spatial array and then were presented with a spatial array that differed from the original in some spatial dimension. In this series of experiments, Quinn (1994) habituated 3-month-old infants to a display where a dot was presented either above or below a horizontal line (see **Figure 1**). Following habituation, infants were found to recover interest when the dot had a new location in relation to the line (i.e., they looked longer at a dot above the line after habituating to the dot below the line). After conducting a series of controlled experiments to rule out explanations relating to encoding of the dot's actual position, it was concluded that infants can categorize and represent *above* and *below* relationships.

Recent research demonstrates that this ability varies across age depending on the particular spatial relationship under examination (Casasola, 2008). In fact, the earliest demonstration of spatial category is that of *left/right* relationships. Gava et al. (2009) found that under certain conditions even 3-day-old infants demonstrate discrimination between spatial configurations representing this relationship, suggesting that rudimentary spatial categorization is innate. Using a habituation paradigm, neonates were presented with various arrangements on a simple black and white display of moving or blinking dots on either side of a referent line. They found that infants were particularly sensitive to changes in the *left/right* relationship between two stimuli, rather than the absolute distance between the stimuli. It is worth noting that this very early ability is weak and highly constrained. The results were likely facilitated by the very simple testing

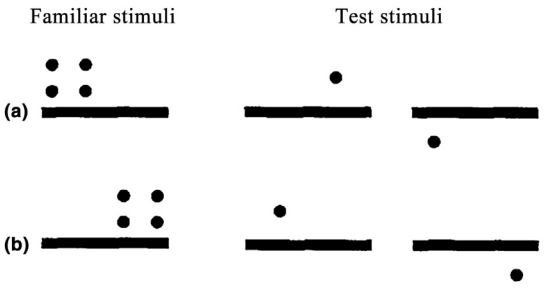

Figure 1

Table 1

Spatial category	Birth	1 month	2 months	3 months	4 months	5 months	6 months	7 months	8 months
Left/right	Initial (Gava et al., 2009)						Abstract		
Above/below				Initial (Gava et al., 2009)			Abstract		
Between							Initial		
Containment							Initial		
Support									

conditions used. Yet employing very simple, somewhat ecologically limited stimuli may be a necessary trade-off when assessing spatial categorization in newborns with very poor visual capacities.

Categorization in the vertical plane is generally not demonstrated until at least 3 months of age (Newcombe, 2002). While the *above/below* relationship is arguably comparable in perceptual complexity to *left/right* relations, the ability to classify the former develops second. This difference in onset may stem from infants having greater early capacity for eye movements in across the horizontal rather than the vertical plane (Gava et al., 2009; Quinn et al., 1996).

Even beyond a growing ability to differentiate in the vertical plane, spatial categorization development continues through the first year of infancy. For example, categorization at 3–4 months remains dependent on the presence of a central referent line, as infants do not seem to be able to internally generate a perceptual midline (Quinn, 1994). Additionally, until 6–7 months of age, this type of spatial discrimination falters when infants are prompted to categorize the locations of even slightly complex shapes. Specifically, infants at 3 months will readily discriminate *above/below* relationships, provided that the object has a simple shape, but not when it is more complex.

Categorization of More Complex Relationships

From 6 months of age, infants demonstrate a rudimentary recognition of the spatial categories of *between* and *containment* (Casasola et al., 2003). However, abstraction of these spatial categories does not appear until about 10 and 14 months of age (Casasola and Cohen, 2002). It is thought that this difference is due to the requisite two landmarks that must be referenced in order to establish a *between* relationship. This ability develops later, and is perceptually more complex than making reference to a single landmark (e.g., *above/below spatial categorization*; Lew et al., 2000).

With respect to understanding the *containment* relationship, different types of experimental paradigms reveal varying levels of spatial understanding in infants. When infants are prompted to manually place an object into a container, infants first demonstrate competence at around 14 months of age. Between 15 and 30 months of age, a great improvement is seen in infants' understandings of specific containment principles. For example, Shutts et al. (2009) observed that across this time period, infants took an increasing amount of care when selecting an object to insert into an aperture of a given size and shape. Prior to this point, infants demonstrate little selectivity. It has been suggested that a developing ability in mental rotation is a primary driver of developmental ability in such tasks (Levine et al., 1999).

In order to complete this seemingly simple object-into-aperture activity, it is also necessary to be able to relate the positive shape of the object to the negative space of the hole. This idea is perhaps more complex than it initially appears, as even adults tend to represent the shape of the negative space differently to the shape of the positive space, even when they are complementary (Bertamini and Croucher, 2003).

Categorization of the spatial relationship of *containment* highlights some of the inherent complexity in mapping development onto a timeline. As with other areas of infant cognition, findings of categorization ability depend largely on the type of experimental paradigm employed. For example, Hespos and Baillargeon (2001) report the initial demonstration of containment relationships at 3 months of age using looking times, much earlier than the previously reported 14 months. These very young infants were found to look longer at containment relations that are not physically possible (e.g., a large peg fitting apparently into a small cylinder) compared to possible containment relationships. These findings arguably reflect the cognitive/perceptual building blocks for later *containment* categorization ability.

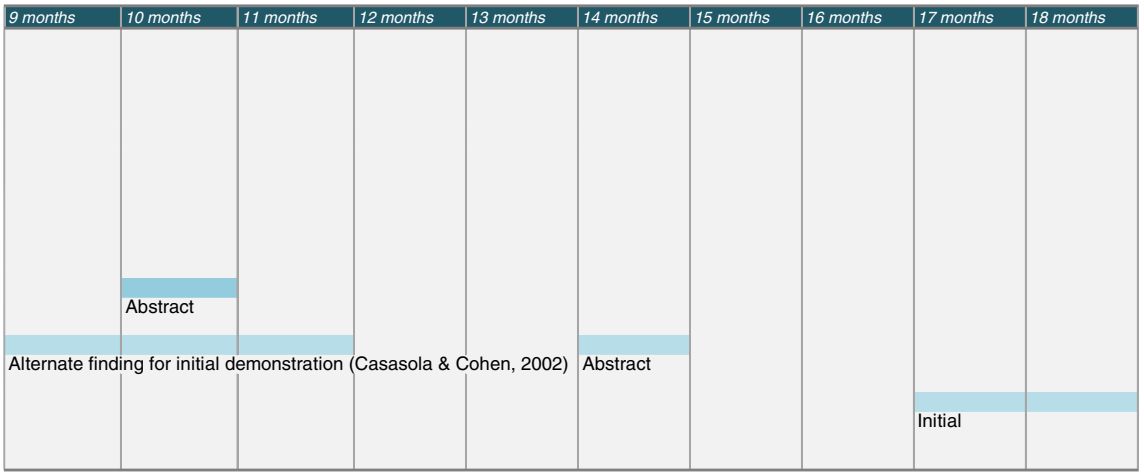

The understanding of *containment* and *support* as spatial categories coincides with a rise in the salience of spatial language. Specifically, the ability of an infant to discriminate between different types of *containment* or *support* may relate to the language with which they are familiar. For example, discrimination between tight- and loose-fit relationships for *containment* (such as a hand in a glove compared to an orange in a fruit bowl) and *support* (such as a building block pushed on to join another, compared to the blocks sitting on a table) is somewhat dependent on having the spatial vocabulary to describe (or at least reference) this difference (Casasola, 2008; see the last section of this article for reference to cultural variations arising from differences in the spatial language used).

Experience is another important factor in the formation of these more robust and complex categorical representations. Infants with more environmental experience are better able to determine if a space is occupied by one fused object or two separate objects (e.g., a cup and saucer; Needham, 2000). See **Table 1** for an overview of related studies.

Spatial Frames of Reference Used by Infants

So far, we have described the progression of infants' abilities in discriminating various relationships between objects. This kind of discriminative ability is necessary for certain successful object manipulations, but does not enable successful navigation through one's environment; nor does it enable proficient encoding and relocating of objects in one's environment. For this to be achieved, infants must possess and make use of a spatial 'frame of reference' (FoR).

With an effective FoR, one can represent their own position relative to their surroundings, as well as where items are located in relation to each other. There are a number of FoR types that are differentially suited to performing particular tasks. In the subsequent sections, we discuss commonly investigated FoRs with respect to infant spatial search ability.

Retinocentric versus Body-Centered FoRs

The *retinocentric* (eye-centered) FoR is a primitive spatial FoR and encodes locations as a series of eye movements. Such a FoR is best suited to learning repeating sequences of events while one is in a stationary position. For example, the locations of a bouncing ball could be encoded in this way: look up, then look down, then look up, etc. Retinocentric information can be extracted directly from sensory signals and involves comparing the point of interest with other points in the visual field in order to establish the object's position in space. Infant research on retinocentric frames has focused on how infants are able to progress to head or body-centric representations.

Initial evidence of a retinocentric limitation was demonstrated in a 'double-step' task with 3-month-old infants (Gilmore and Johnson, 1997). They found that after seeing two targets appear simultaneously (in the same horizontal plane) below a starting fixation point, infants were generally unable to compensate for their initial look to one first target when directing their gaze to the other target (see **Figure 2**). That is, they looked down to the first target correctly, but then looked down again to the second target when a lateral eye movement was necessary.

Kaufman et al. (2006) similarly found that 4-month-old infants who learned to follow a sequence of pictures appearing in opposite diagonal corners of a display showed retinocentric behavior when an unrelated, briefly presented stimulus appeared unexpectedly in the sequence. After the disruption, infants did not look back to a corner of the screen, but instead directed their gaze diagonally from the unrelated stimulus. Although complexity of the sequence and familiarity with the stimuli can influence retinocentric behavior, it is not until around 7 months of age that infants consistently account for intermediary eye movements and current eye position in sequence learning.

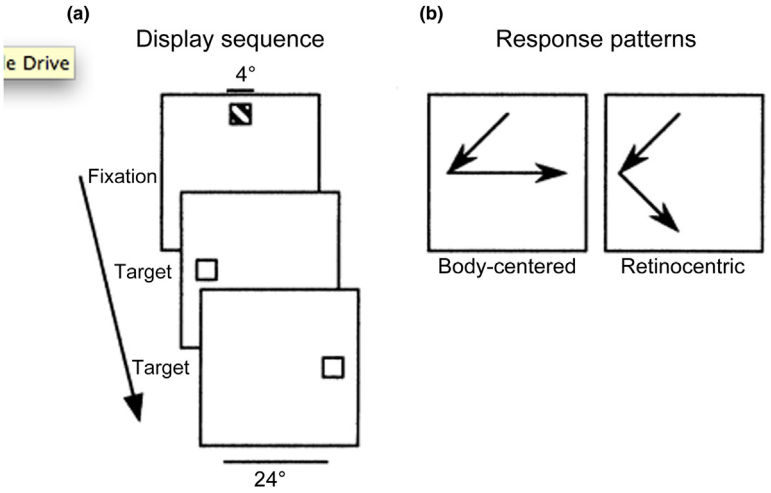

Figure 2

It is important to note that from the participant's perspective, a retinocentric FoR actually does work most of the time in these kinds of sequence learning tasks. Even if some of the stimuli are missed, generally when there is a consistent sequence, retinocentric-based eye movements will pay off with the child seeing the sought-after stimulus. But in tasks involving manual action or search, a retinocentric framework is almost entirely unsuitable. It is therefore not surprising that infants can and do employ body-centered representations under other circumstances.

Body-Centered versus Allocentric Representations

Early studies of infant FoRs were largely concerned with when infants are able to move from egocentric (or body-centered) representations to allocentric representations. Using a body-centered representation means that locations of objects are encoded with respect to the viewers' body – thus the effectiveness of this representational scheme fails once the viewer moves through space. In contrast, an allocentric representation entails representing locations with respect to an external framework (such as landmarks or other objects). To assess infants' ability to use these FoRs, researchers typically employ what can be called 'search-and-move' tasks. In such tasks, the infant initially searches (either manually or visually) for a hidden target in the experimental apparatus. Following this, the infant is moved around the apparatus to a new position (usually directly across from the starting position) whereupon he or she is prompted to search again for the desired target.

For example, Bremner (1978) used a version of this task where 8–9-month-old infants reached for a hidden toy in one of two hiding wells on a tabletop and were then passively moved around to the other side of the table where they were prompted to search again in a test trial. Generally, infants appeared to search 'egocentrically' or in a body-centered fashion. That is, if the toy was hidden in the hiding well on the left side of the table prior to the test trial, infants continued to reach to the left side following their passive movement in the test trial (although now the object is located on the right side from the infant's perspective). Similar findings were reported when infants were prompted to look through a right- or left-side window for a target rather than reach for a target (Acredolo, 1978).

While the results of this work have been interpreted as evidence for infants being limited to egocentric FoRs, subsequent work using variants of this task demonstrate greater representational flexibility even at very young ages. This research revealed numerous factors that can significantly reduce body-centered search in such tasks. These factors include the presence of salient, direct landmarks near the hiding position (e.g., Acredolo and Evans, 1980; Bremner, 1978), visible geometric cues (salient corners, walls, etc.) in the testing environment (Keating et al., 1986), and a familiarity with the testing environment (Acredolo, 1978).

More recent work suggests that the body-centered search tendencies of infants may stem not from a representational bottleneck, but rather from infants in these tasks not realizing that they were actually moved to a new position prior to the test trial. Kaufman and Needham (1999, 2011) have argued that because reaching for hidden objects is so challenging for young infants, they may be focusing on the process of reaching at the expense of attending to their passive movement around the room prior to the test trial. Using a series of visual habituation experiments, it was found that infants as young as 4.5 months appear to habituate not only to a body-centered view of an object, but also to its actual position in the room. Infants were found to dishabituate when examining the same object in a new position even if the object maintained its egocentric relationship to the child.

It seems likely then that the reason that self-produced locomotion (i.e., the onset of crawling) appears to trigger spatial competence (Lew et al., 2000; Bai and Bertenthal, 1992) is not that it allows new forms of representation. Rather, locomotion experience increases motivation and ability to track self-movement around the room. This is consistent with a number of research findings which show that increasing the complexity of infant motion decreases subsequent search performance (Kaufman and Needham, 2011; Landau and Spelke, 1988). See **Table 2** for an overview of related studies.

Table 2

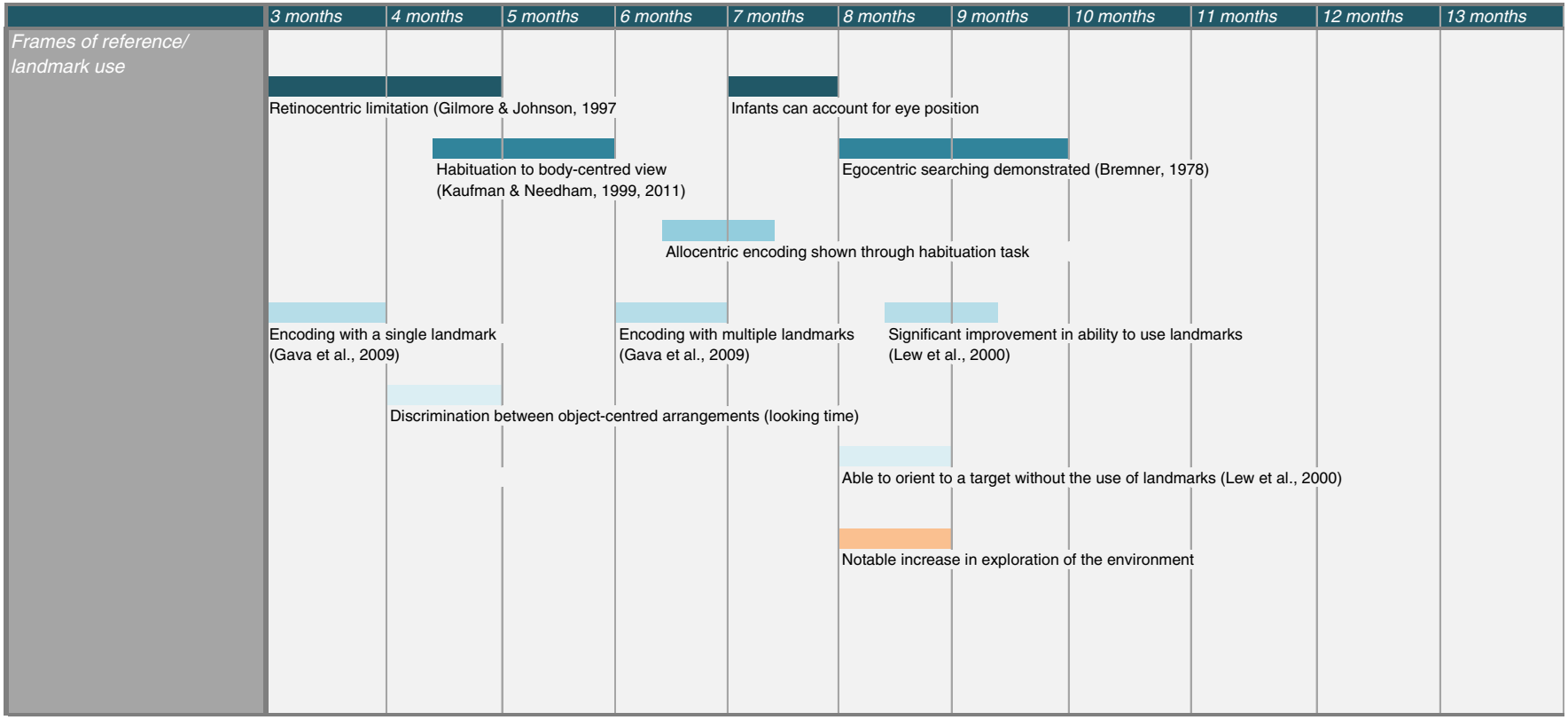
3 months
4 months
5 months
6 months
7 months
8 months
9 months
10 months
11 months
12 months
13 months
Frames of reference/ landmark use
Retinocentric limitation (Gilmore & Johnson, 1997
Infants can account for eye position
Habituation to body-centred view (Kaufman & Needham, 1999, 2011)
Egocentric searching demonstrated (Bremner, 1978)
Allocentric encoding shown through habituation task
Encoding with a single landmark (Gava et al., 2009)
Encoding with multiple landmarks (Gava et al., 2009)
Significant improvement in ability to use landmarks (Lew et al., 2000)
Discrimination between object-centred arrangements (looking time)
Able to orient to a target without the use of landmarks (Lew et al., 2000)
Notable increase in exploration of the environment

Table 3

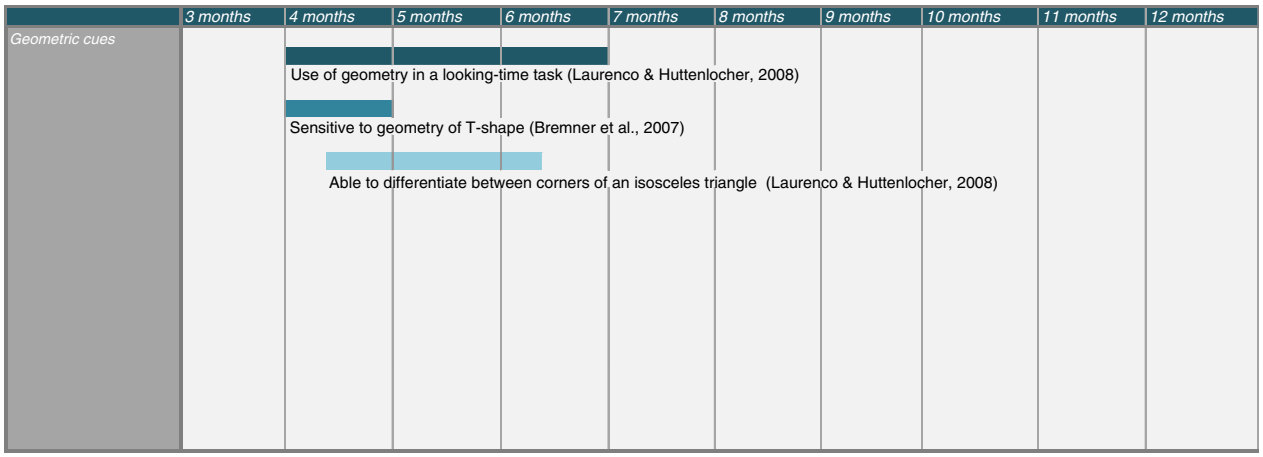

Geometric Cue Use in Infants and Young Children

In the previous section, it was noted that a failure to reorient could be misinterpreted as a limitation in representation ability. In many studies, infants were tested in environments with very sparse geometric cues, which it turns out, are often crucial to accurate reorientation. When experiments include informative geometric cues, they tend to reveal that infants perform markedly better at search-and-move tasks.

Processing of Geometric Information in Young Infants

As environmental geometric cues help performance in search-and-move tasks, it seems apparent that young infants have an ability to discriminate cues that differ in their geometry. Indeed, a number of studies have demonstrated this. For example, Lourenco and Huttenlocher (2008) used a looking-time procedure to examine whether 4.5–6.5 month olds were able to differentiate among the corners of a two-dimensional isosceles triangle presented on a screen. They found that when the orientation of the triangle was less variable, infants could discriminate between all corners, while for highly variable orientations, only the unique corner could be distinguished from the others.

Similarly, Bremner et al. (2007) have found that infants as young as 4 months of age are sensitive to an object's geometry. In this study, 4- and 8-month-old infants were familiarized with a T-shaped object with an internal light illuminating one of the T's three ends. When the object was rotated, infants showed a preference to look at the shape when the light's position was constant within the T-shape, regardless of the light's egocentric relationship with the infant. In order for this to happen, it is argued, infants must have formed a representation of the object's geometry.

Geometric Cue Use in Reorientation by Toddlers

While geometric cue use has been evaluated developmentally across a range of tasks, its role in 'reorientation' tasks has been an area of particular interest. Reorientation tasks involve disorienting the participant and determining what information is used to reorient within the testing apparatus. The reason for the intense interest in geometric cue use in child reorientation is that a special role for geometric processing has been suggested across species and across development in such tasks. Experiments with rodent species have indicated that geometric cues are often used preferentially over other landmark cues when reorienting and learning a new route – leading some researchers to posit the existence of a 'geometric module' in the brain which drives place learning (Cheng, 1986). In line with this, geometric cue use in reorientation has also been found in adult humans (Binet, 1894; Sandstrom et al., 1998). It is not surprising then that the possibility of preferential use of geometry over landmark cues has been extensively examined in young children as well.

Most of the developmental work on reorientation is based on a key series of experiments conducted by Hermer and Spelke (1994, 1996). In their experimental paradigm, toddlers (18–24 months) were disoriented (by being turned around in a circle with their eyes shut) in a novel environment that contained both geometric and nongeometric information. Geometric information included the large-scale (rectangular) shape of the room, and nongeometric information included the color of a wall, the patterning on a box, or the presence of a toy truck or bear (see **Figure 3**). They found that toddlers relied almost entirely on geometric information when searching for a hidden toy and were virtually unable to use nongeometric landmarks such as toys in the room or a colored wall. Instead, the participants divided their searches evenly between the diagonally opposite corners, uninfluenced by presence of nongeometric landmarks. As with nonhuman animal research in this area, the authors suggested that, in nature, geometric information is more stable and salient than local landmark information and thus is a more reliable cue for place learning and reorientation.

In a related study by Lee and Spelke (2008), 4-year-old children successfully reoriented using the shape of an area

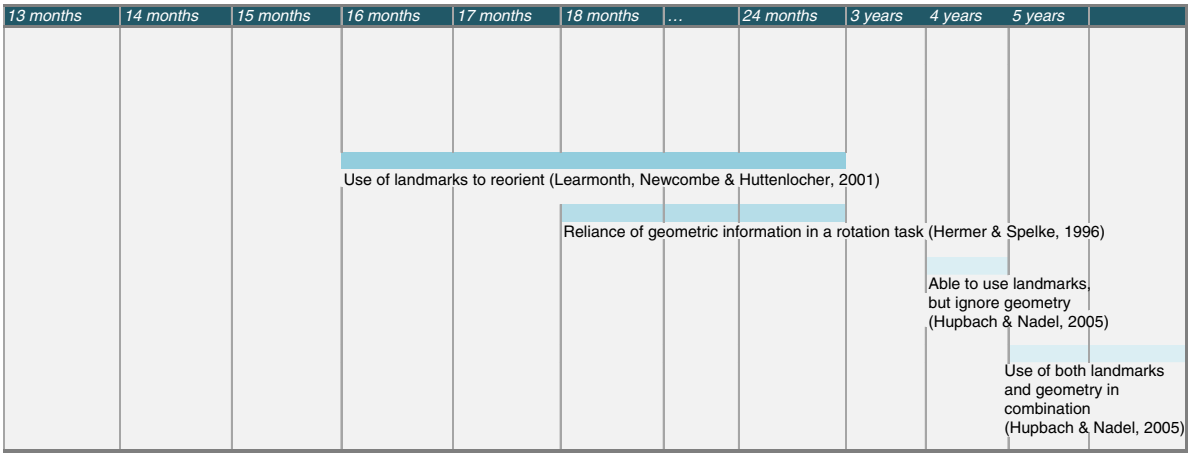

formed by a very low wall (short enough to step over), but not when the area was defined by large columns, or connected lines on the floor, even though they were able to detect the presence of all three types of boundary. They argued that in order to reorient effectively in a natural environment, one must rely on environmental information that is unique, stable, and enduring. As with color, surface markings (such as the lines on the floor) are unlikely to be stable. Additionally, although the columns were stable and enduring, like other large, fixed objects such as trees, they were not sufficiently distinguishable to specify which one was marking a particular direction or location. They argued that while landmark stability is important, it is not sufficient for reorientation. Rather, children rely on the geometry of extended three-dimensional layouts, such as the low wall in their experiment, or the overall landscape of a natural environment, that arguably provide sufficient stable, enduring, and unique geometric information for reorientation.

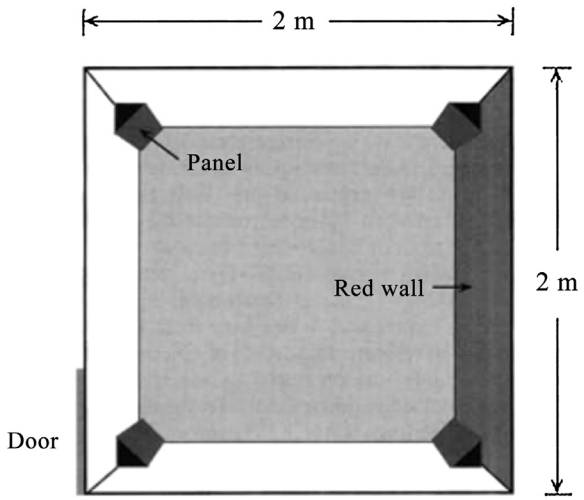

Figure 3

Nongeometric Cue Use by Toddlers

Recently, however, there has been a developing alternate view that young children do use nongeometric information to reorient, depending on the particular task. Learmonth et al. (2001) found that in contrast to Hermer and Spelke (1994, 1996), toddlers did reorient using nongeometric landmark cues. In the absence of landmarks, they found that children searched equally in the correct corner and the corner diagonally opposite it. However, when a colored wall was included in the apparatus, children did use it to search for a hidden toy, casting doubt on the idea that geometric information alone is used for reorientation. Learmonth et al. explained this difference in results by noting that Hermer and Spelke made use of a within-subjects paradigm, whereby participants were exposed to the enclosed space with and without the colored wall, and also allowed the children to play with the toys used as potential landmarks (to ensure they had noticed them). Learmonth et al. argue that these procedures likely limited the toddlers' perceptions of stability of the potential landmarks provided, making them unlikely to be used for navigation.

Nardini et al. (2008) support this finding and report greater landmark-based reorientation when toddlers are tested in an apparatus with multiple colored walls. Additionally, research has found that toddlers' tendency to reorient and search based on geometric cues is less consistent than originally reported. Specifically, geometric cue use apparently occurs in symmetrical but not asymmetrical testing environments (Lew et al., 2010).

Overall, the story of a geometric module in children as well as animal and adults continues to unfold. A main area of contention concerns what constitutes a geometric versus non-geometric landmark. For example, 3–5 year olds will make use of landmark information when tested in a large space, but rely on geometric information in small spaces (Learmonth et al., 2001, 2002). While this would seem to contradict a geometric module account, it could also be argued that landmark cues of sufficient size can be construed as a geometric cue. Further research will certainly address this issue. See Table 3 for an overview of related studies.

Cultural Differences

A common critique of studies in developmental psychology is that they predominantly make use of Western, educated, industrialized, and affluent samples that are not necessarily representative when it comes to many key domains, including spatial reasoning (Henrich et al., 2010). Studies that have examined spatial cognition across cultures have important cultural variations in spatial behavior, especially with the emergence of language. While little difference has been found between groups prior to this point, it is worth noting that cross-cultural comparisons prior to the development of language are few and far between (Haun et al., 2006).

Language and Spatial Categorization

The debate surrounding the issue of *linguistic relativity*, or direction of influence between language and cognition, is pertinent in cross-cultural developmental psychology research. While it is well understood that cognition influences language use, there is evidence to suggest that the reverse relationship also holds true. Within the context of spatial categorization, there is specific evidence for a bidirectional relationship between the acquisition of spatial categories and use of specific spatial language terms (Casasola et al., 2009).

As noted earlier in the present article, by 14 months of age, use of spatial language is increasingly influential in infants' use of different spatial categories. As noted by Choi (2006), infants tend to be less sensitive to spatial categories or relationships that are not illustrated in their native language. Moreover, a number of recent studies have examined the differences between the spatial categorization abilities of infants from English-speaking and Korean-speaking families in relation to the different systems of spatial organization between the two languages (e.g., Choi and Bowerman, 1991). English-learning toddlers distinguish between spatial categories based on the object's placement, such as *containment* (in) and *support* (on). On the other hand, Korean-learning toddlers distinguish between those that result in a tight-fit relation (*kkita*) for both *containment* and *support* events and those that result in loose-fit containment (*nehta*) or support (*nohta*) events.

Casasola and Cohen (2002) found that familiarizing English-learning toddlers with the general semantic category of *in* resulted in the formation of a broad category of containment that included, but did not differentiate between, the two specific subcategories. Similarly, they found that infants who were familiarized with only the specific tight- or loose-fit categories (but not the general *in* or *on*) could form spatial categories that included relations from both subgroups. Experience with spatial language may help infants to attend to the commonalities between these spatial relationships that are otherwise perceptually quite different (i.e., tight-fit containment compared to tight-fit support relationships).

Cultural Variation in FoR Use

In addition to the development of spatial categories, another key area of cross-cultural difference is in the use of FoR. Earlier in the article, we described young children's use of egocentric and allocentric frameworks. In addition to these FoRs, one can also encode spatial relationships with cardinal directions (i.e., north, south, east, west) which constitutes a geocentric FoR. Cross-cultural research has revealed that children in various parts of the world use a geocentric FoR to represent and describe locations (Haun et al., 2011; Majid et al., 2004). The use and development of this FoR, and the speed at which it is acquired, apparently varies as a function of culture (Henrich et al., 2010; Dasen and Mishra, 2010).

Numerous cases have been found globally where a geocentric FoR is the dominant style used, and develops much earlier than in Western groups, where it is typically learnt last and used least frequently. For example, it would be much more likely for a Western child to refer to a spoon *behind* a plate (relative to his/her own position), rather than *west* of it. However, the latter phrasing is common in a variety of languages, including those of the Yimithirr, Arrernte, and Warwa indigenous Australian peoples.

In a detailed compilation of studies conducted in Geneva and areas throughout Bali, Nepal, and India, Dasen and Mishra (2010) explore how a geocentric FoR develops in children as young as 4 years of age. These studies explore the conditions that favor the development of a geocentric FoR and its use in verbal and nonverbal spatial tasks. In the non-Westernized societies of Bali, Nepal, and India where tradition and language have been preserved, a geocentric FoR is preferred, even though these populations have the linguistic capability to use egocentric FoR. Children in the villages of Roopchandpur (India), Dolhaka (Nepal), and Sambangan (Bali) almost never used egocentric language when performing the research tasks. However, an egocentric FoR is sometimes chosen by younger children. This suggests that even though children brought up in these societies have the option of using an egocentric FoR, there is a marked preference for the geocentric referencing style.

In one of the studies referenced in the aforementioned book, Dasen and Wassmann (2004) compared Balinese and Swiss children on their uses of various FoRs. They found that the geocentric FoR was well established and used almost exclusively from 4 years of age by Balinese children, particularly those living in isolated locations and with less contact with the Indonesian language or media. Swiss children, in comparison, predominantly utilized an egocentric FoR when describing a scene, with no geocentric language at all being used in the task. Additionally, almost none of the Swiss children knew of the system of cardinal directions at 9–10 years of age.

Haun et al. (2011) explored cross-cultural use of FoRs in a series of nonlinguistic special memory tasks. They compared Dutch and Namibian children's ability to reconstruct spatial arrays of varying complexity from memory, according to their habitual (egocentric for the Dutch children and geocentric for the Namibian children) and nonhabitual linguistic FoRs. Children from both cultures were consistently impaired in their performance for their nonhabitual FoR, demonstrating that not only preference, but also competence in FoR use is linguistically aligned across cultures.

Unfortunately, as cross-cultural studies tend to examine children rather than infants, little is known regarding whether the differences outlined earlier are present prior to the development of language. Regardless, these cross-cultural

differences pose an exciting new frontier for developmental scientists to better understand spatial thinking and its development in young children.

Conclusion

In this article, we have described the key findings in four major areas of developmental research in spatial cognition. A recurring theme throughout this discussion is that the apparent ability to think spatially is heavily influenced by variations in the experimental paradigm chosen and in the particular culture studied. At one level of analysis, these discrepancies make it difficult to clearly state when young children develop or possess a certain ability to represent space. However, with a broader view one can see that the last five decades have given us a rich understanding of the specific factors that drive spatial competence across ages, paradigms, and cultures. As a result of this research, developmental scientists have generally moved away from simple models that impute representational limitations at specific ages and instead are focusing on factors that influence representational flexibility.

See also: Cognitive Maps; Human Development, Theories of; Infant Gaze Behaviors; Mental Imagery: Visual Cognition; Space, Linguistic Expression of; Visual Perception, Neural Basis of; Visual Space, Geometry of.

Bibliography

Acredolo, L.P., 1978. Development of spatial orientation in infancy. Developmental Psychology 14 (3), 224–234.

Acredolo, L.P., Evans, D., 1980. Developmental changes in the effects of landmarks on infant spatial behavior. Developmental Psychology 16 (4), 312.

Bai, D.L., Bertenthal, B.I., 1992. Locomotor status and the development of spatial search skills. Child Development 63 (1), 215–226.

Bertamini, M., Croucher, C.J., 2003. The shape of holes. Cognition 87 (1), 33–54.

Binet, M.A., 1894. Reverse illusions of orientation. The Psychological Review 1, 337–350.

Bremner, J.G., 1978. Egocentric versus allocentric spatial coding in nine-month-old infants: factors influencing the choice of code. Developmental Psychology 14 (4), 346–355.

Bremner, A.J., Bryant, P., Mareschal, D., 2005. Object-centred spatial reference in 4-month-old infants. Infant Behavior & Development 29, 1–10.

Bremner, A.J., Bryant, P., Mareschal, D., Volein, A., 2007. Recognition of complex object-centred spatial configurations in early infancy. Visual Cognition 15 (8), 896–926.

Casasola, M., 2008. The development of infants' spatial categories. Current Directions in Psychological Science 17 (1), 21–25.

Casasola, M., Cohen, L.B., 2002. Infant categorization of containment, support and tight-fit spatial relationships. Developmental Science 5 (2), 247–264.

Casasola, M., Cohen, L.B., Chiarello, E., 2003. Six-month-old infants' categorization of containment spatial relations. Child Development 74 (3), 679–693.

Casasola, M., Bhagwat, J., Burke, A.S., 2009. Learning to form a spatial category of tight-fit relations: how experience with a label can give a boost. Developmental Psychology 45 (3), 711.

Cheng, K., 1986. A purely geometric module in the rat's spatial representation. Cognition 23 (2), 149–178.

Choi, S., 2006. Influence of language-specific input on spatial cognition: categories of containment. First Language 26 (2), 207–232.

Choi, S., Bowerman, M., 1991. Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. Cognition 41 (1), 83–121.

Dasen, P.R., Mishra, R.C., 2010. Development of Geocentric Spatial Language and Cognition: An Ecocultural Perspective. Cambridge University Press, New York.

Dasen, P.R., Wassmann, J., 2004. A cross-cultural comparison of spatial language and encoding in Bali and Geneva. In: International Conference on Cognitive Science, Allahabad University, pp. 16–18.

Gava, L., Valenza, E., Turati, C., 2009. Newborns' perception of left–right spatial relations. Child Development 80 (6), 1797–1810.

Gilmore, R.O., Johnson, M.H., 1997. Body-centered representations for visually-guided action emerge during early infancy. Cognition 65 (1), B1–B9.

Haun, D., Call, J., Janzen, G., Levinson, S.C., 2006. Evolutionary psychology of spatial representations in the hominidae. Current Biology 16 (17), 1736–1740.

Haun, D.B.M., Rapold, C.K., Jenzen, G., Levinson, S.C., 2011. Plasticity of human spatial cognition: spatial language and cognition covary across cultures. Cognition 119, 70–80.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and Brain Sciences 33 (2-3), 61–83.

Hermer, L., Spelke, E.S., 1994. A geometric process for spatial reorientation in young children. Nature 370 (6484), 57–59.

Hermer, L., Spelke, E., 1996. Modularity and development: the case of spatial reorientation. Cognition 61, 195–232.

Hespos, S.J., Baillargeon, R., 2001. Infants' knowledge about occlusion and containment events: a surprising discrepancy. Psychological Science 12 (2), 141–147.

Kaufman, J., Needham, A., 1999. Objective spatial coding by 6.5-month-old infants in a visual dishabituation task. Developmental Science 2 (4), 432–441.

Kaufman, J., Needham, A., 2011. Spatial expectations of young human infants, following passive movement. Developmental Psychobiology 53 (1), 23–36.

Kaufman, J., Gilmore, R.O., Johnson, M.H., 2006. Frames of reference for anticipatory action in 4-month-old infants. Infant Behavior & Development 29, 322–333.

Keating, M.B., McKenzie, B.E., Day, R.H., 1986. Spatial localization in infancy: position constancy in a square and circular room with and without a landmark. Child Development, 115–124.

Landau, B., Spelke, E., 1988. Geometric complexity and object search in infancy. Developmental Psychology 24 (4), 512.

Laurenco, S.F., Huttenlocher, J., 2006. How do young children determine location? Evidence from disorientation tasks. Cognition 100, 511–529.

Laurenco, S.F., Huttenlocher, J., 2008. The representation of geometric cues in infancy. Infancy 13 (2), 103–127.

Laurenco, S.F., Huttenlocher, J., 2009. Location representation in enclosed spaces: what types of information afford young children an advantage? Journal of Experimental Child Psychology 104, 313–325.

Learmonth, A.E., Newcombe, N.S., Huttenlocher, J., 2001. Toddlers' use of metric information and landmarks to reorient. Journal of Experimental Child Psychology 80, 225–244.

Learmonth, A.E., Nadel, L., Newcombe, N.S., 2002. Children's use of landmarks: implications for modularity theory. Psychological Science 13 (4), 337–341.

Lee, S.A., Spelke, S., 2006. Reorientation and landmark-guided search by young infants: evidence for two systems. Psychological Science 17 (7), 577–582.

Lee, S.A., Spelke, S., 2008. Children's use of geometry for rotation. Developmental Science 11 (5), 743–749.

Levine, S.C., Huttenlocher, J., Taylor, A., Langrock, A., 1999. Early sex differences in spatial skill. Developmental Psychology 35 (4), 940.

Lew, A.R., Bremner, G., Lefkovitch, L.P., 2000. The development of relational landmark use in six-to-twelve-month-old infants in a spatial orientation task. Child Development 71 (5), 1179–1190.

Lew, A.R., Gibbons, B., Murphy, C., Gavin, B.J., 2010. Use of geometry for spatial reorientation in children applies only to symmetric spaces. Developmental Science 13 (3), 490–498.

Lourenco, S.F., Huttenlocher, J., 2008. The representation of geometric cues in infancy. Infancy 13 (2), 103–127.

Majid, A., Bowerman, M., Kita, S., Haun, Levinson, S.C., 2004. Can language restructure cognition? The case for space. Trends in Cognitive Sciences 8 (3), 108–114.

Mishra, R.C., Dasen, P.R., Niraula, S., 2003. Ecology, language, and performance on spatial cognitive tasks. International Journal of Psychology 38 (6), 366–383.

Nardini, M., Burgess, N., Breckenridge, K., Atkinson, J., 2006. Differential developmental trajectories for egocentric, environmental and intrinsic frames of reference in spatial memory. Cognition 101, 153–172.

Nardini, M., Braddick, O., Atkinson, J., Cowie, D.A., Ahmed, T., Reidy, H., 2008. Uneven integration for perception and action cues in children's working memory. Cognitive Neuropsychology 25 (7–8), 968–984.

Needham, A., 2000. Improvements in object exploration skills may facilitate the development of object segregation in early infancy. Journal of Cognition and Development 1 (2), 131–156.

Newcombe, N.S., 2002. The nativist-empiricist controversy in the context of recent research on spatial and quantitative development. Psychological Science 13 (5), 395–401.

Niraula, S., Mishra, R.C., Dasen, P.R., 2005. Linguistic relativity and spatial concept development in Nepal. Psychology and Developing Societies 16 (2), 99–124.

Quinn, P.C., 1994. The categorization of above and below spatial relations by young infants. Child Development 65, 58–69.

Quinn, P.C., Cummins, M., Kase, J., Martin, E., Weissman, S., 1996. Development of categorical representations for above and below spatial relations in 3- to 7- month-old infants. Developmental Psychology 32 (5), 942–950.

Sandstrom, N.J., Kaufman, J., Huettel, A.S., 1998. Males and females use different distal cues in a virtual environment navigation task. Cognitive Brain Research 6 (4), 351–360.

Shutts, K., Örnkloo, H., von Hofsten, C., Keen, R., Spelke, E.S., 2009. Young children's representations of spatial and functional relations between objects. Child Development 80 (6), 1612–1627.

Spelke, E.S., Gilmore, C.K., McCarthy, S., 2011. Kindergarten children's sensitivity to geometry in maps. Developmental Science 14 (4), 809–821.

Spatial Data

Stephen C Guptill, U.S. Geological Survey (Retired), Oakton, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial data are the sum of our interpretations of geographic phenomena. Taken literally, spatial data could refer to any piece of information that associates an object with a location – from stars in the heavens to tumors in the human body. Typically, however, when used in a geographical context, spatial data are restricted to describing phenomena on or near the earth's surface. In digital form, the data are the primary information needed by geographic information systems, the software tools for spatial-data analysis and by location-based services used by mobile devices.

What Are Spatial Data?

'Spatial data' are used as an all-encompassing term that includes general-purpose maps, remotely sensed images, and census-tract descriptions, as well as more specialized data sets such as seismic profiles, distribution of relics in an archeological site, or migration statistics. Information about buildings, bridges, roads, streams, grassland, and counties are other examples of types of spatial data. Broadly, spatial data refer to any piece of information that has an absolute or relative location. More commonly, spatial data are associated with information shown on maps and images of natural and man-made features that range in size from 1 m to 10 km. Such data are often referred to as geospatial data. Spatial data sets are collections of locational and nonlocational information about selected geographic phenomena. This collection of information forms the heart of a spatial database. Once the information is in the database, computers provide the capabilities to handle geographic data in ways that were previously impossible.

Data Models

A digital spatial data set represents a certain model of geographic reality. These models describe both the locational and nonlocational aspects of the phenomena of interest. For users to make sense of the data they are receiving, they must understand the conceptual model underlying the data.

The locational aspect of a spatial data set is usually described by one of two types of models, either grid-cell (sometimes called raster) data models or vector data models. Grid-cell data models use collections of fixed units of space, for example, cells of land 50 m on a side, to describe the location and extent of a geographic phenomenon. Digital remotely sensed images are collected in this form. Vector data models use graphic elements (points, lines, and polygons) to describe the location and extent of the phenomenon under study. Most digital map data are in this form.

The spatial data also must include all the nonlocational information used to describe the features in the database. In the case of digital image data, the amount of nonlocational data is relatively small, usually a set of values measuring characteristics such as the greenness of vegetation. With vector data, the amount of nonlocational, or attribute, data can be quite large, even exceeding the amount of spatial data. This includes information about the various features included in the database (e.g., the number of lanes of a highway) and their relationships to each other (e.g., that the railroad tracks lie between the stream and the highway). In the context of geographic information systems (GIS), features are the sum of the interpretations of geographic objects or phenomena. Buildings, bridges, roads, streams, grassland, and counties are examples of features.

Feature Data Structures

The real world of geographical variation is infinitely complex and often uncertain, but must be represented digitally in a discrete, deterministic manner. In one method of describing geographic reality, characterized by the information shown on maps, one thinks of the world as a space populated by features of various kinds – points, lines, and areas. This is commonly referred to as the vector or feature data model. Features have attributes that serve to distinguish them from each other. Any point in the space may be empty or occupied by one or more features. These features may exist in one combined database, or may be separated according to a theme or variable into a number of layers or classes. The terminology used to describe these features is given in Table 1.

The relationships between the nodes, arcs, and polygons are called topological relationships. In simple terms, the topological relationships are that arcs start and end only at nodes, and that polygons are enclosed by a series of arcs. Attention is given to neighboring geographical entities when coding arcs. For instance, the topological description of each arc includes not only the starting and ending bounding nodes but also the left and right designations of the cobounding regions. In this way, an arc is directed, that is, the first coordinate is that of the start node and the last coordinate that of the end node. Each

Table 1 Terminology for describing geographic features

Feature type	*Example*	*Digital database term (and synonyms)*
Point feature	Benchmark	Node (vertex, 0-cell)
Linear feature	Road centerline	Arc (edge, 1-cell)
Areal feature	Lake/pond	Polygon (face, 2-cell)

polygon record might also include the identification of any interior islands or, if the polygon is itself an interior island, the identification of the single exterior polygon. This data structure not only accommodates the natural pattern of mapped features efficiently, but also provides the information needed to facilitate automated editing and manipulation of the database.

Raster Data Structures

Instead of describing discrete geographic objects, geographical variation can be represented by recording the locational pattern of one variable over the study area. When the spatial interval of this sampling is regular (e.g., rectangular), the resulting matrix of observations is called a raster data structure. If there are n variables, then n separate items of information are available for each and every point in the area. Digital terrain model data and digital images are typically stored in this structure. Other types of information such as categories of land cover could also be encoded in a raster data structure. The cells of raster data (particularly raster data of images) are commonly referred to as pixels (derived from the term picture element).

Modeling Solids

So far, the discussions have dealt with data on a surface. However, certain types of spatial data require a more complete modeling of the third dimension. Solid bodies, such as geologic structures (ore bodies, petroleum reservoirs, etc.), need to be encoded not only with x, y locations, but also with z or depth information. The surface models mentioned earlier can encode z values, but only one z value is permitted for each location. This results in what is termed a 2.5-D model. True solid or 3-D modeling requires that multiple z values can exist at a given x, y location.

One mechanism to model such data is to extend the raster data model into an additional dimension. Thus, the 2-D squares become 3-D cubes. The cubes are often termed voxels (for volume elements). Each voxel is encoded with attribute data (such as rock type). The geolocation of any given voxel is implied by its location in the x, y, z array and the size of the cubes. The voxel representation allows one to model such things as hollow spheres, with a varying thickness of the shell, or solids with holes through them.

Another way to model solids is to use a technique called boundary representation. This is an extension of the vector data model to 3-D. A solid object is defined by its bounding surface. Just as 1-D lines bound 2-D areas, 2-D planar areas are located in 3-D space and are used to bound a 3-D solid. The various planar areas bound solids, much like the facets bound the shape of a jewel.

Describing Geographic Features

As noted earlier, spatial data models, in addition to providing locational data, must handle all the nonlocational information used to describe the features in a database. For users to understand the nature of the data, there must be a complete specification of the features being described. A comprehensive data specification includes the definition of the features as well as definitions of their attributes and attribute values. Other information required for a complete understanding includes delineation guidelines, data extraction rules, and representation rules. Understanding how the locational component of a feature is represented is of little use unless the user understands the nature of what is being described. What is really meant by terms such as 'forest,' 'marshlands,' 'residential land,' or 'bridge'? For example, one could declare a 'bridge' to be a feature. A bridge is one of a number of built-up structures on or near the surface of the earth. Its definition may be further refined to a structure erected to carry traffic over a depression or obstacle. Thus, the feature 'bridge' is an element of a set of phenomena (an 'erected structure'), with the common attributes of function ('to carry traffic') and location. It also has the common relationship of spanning another feature ('over a depression or obstacle'). The definitions and specifications of the data producers hopefully match to some degree the expectations of the user.

To help solve this data-definition problem, a feature data specification can be used to provide a detailed definition of each feature and how it is modeled. This technique is used by national mapping agencies in the United States, France, and Germany. The feature data specifications define the domain of features that may be represented in a spatial database. For each feature in the feature data specification, a domain of attributes is defined. Each attribute is assigned a value, taken from the domain of values specified by the attribute definition authority. In addition to the definitions, guidance is given on how a given feature is delineated, when it is collected, and how it is represented in the spatial data model.

Specific relationships, termed semantic relationships, are used to express interactions that occur between features that are otherwise difficult to represent with locational data, topological relationships, or feature descriptions. Such relationships can include the following: to flow, the relationship that occurs among the 2-D features composing a river system; to connect, the relationship that occurs (at junctions) among the 2-D transportation features composing a road network; to bound or to be bounded, the relationship that occurs between the perimeter of a city and its city limits; to be composed of, the relationship that occurs between a designated route of travel and the transportation features that compose it, or of the stream and pond features that comprise a watercourse; and to be above, the relationship that occurs between the features in an overpass or underpass of a freeway interchange. The intent of all these relationships is to model the interactions between features sufficiently to support GIS applications and automated cartographic-product-generation requirements.

Modeling Time

Just as maps change over time and must be revised, so do spatial databases. There are two basic ways to handle the temporal dimension associated with spatial data. The simplest way is to maintain multiple versions of a database, akin to keeping a historical map file. When changes occur, a new edition of the database is produced with the current data, and the previous edition is placed in a historical archive.

A more comprehensive, yet more complex, strategy is to maintain one database, where every database feature (and potentially its attributes) has time information associated with it. This is fairly straightforward for information in a raster form. For each cell, a variable-length list of attributes is maintained, noting any change in attribution and the time at which the change occurred. For feature-based vector data sets, the process is more complicated. Information is encoded into the database on when the feature was first incorporated into the database, if it presently exists, when it ceased to exist (e.g., a pond was filled in), and when it was recreated, perhaps in a different form (e.g., a road was rerouted). In this fashion, the existence and modification of any given feature can be traced.

Data Quality

The utility of spatial data in GIS depends directly on the quality of the spatial data used in the process. Geospatial data quality is measured along at least three different axes: space (location, places), time, and theme (content, characteristics). The space axis characterizes positional accuracy; the time axis indicates temporal accuracy and completeness; and the theme axis reflects semantic accuracy, attribute accuracy, and completeness. These qualities are not independent, however. Indeed, some data quality aspects, such as logical consistency, directly refer to the interrelationships of these data quality elements. Not all these quality elements have well-defined quality metrics associated with them. Metrics of semantic accuracy, for example, are not well understood. And some of the measures used for other quality elements may be inadequate.

Also needed are linkages between the quality information present in the data and the GIS-processing algorithms, so that software can produce error estimates on the end results. The results of GIS analyses have often been presented as definitive to decision makers (e.g., this is the best site for the municipal water well), with little or no qualification. One would expect that decision makers would insist on qualifying statements on any results presented to them. The methods for doing this are not well developed. For example, owing to the cumulative effects of error, it has been conjectured that as the number of data sets used in a GIS analysis increases, the accuracy of the result decreases. This conjecture runs counter to common practice, where multiple data sets are used with little regard for whether they are contributing more to the information content or to errors in the result.

Metadata

As technology provides faster and more efficient ways to transmit and process spatial data, data producers and users are realizing the potential value of exchanging and sharing spatial data. The ability to use existing spatial data is important to organizations that are trying to share data and to increase benefits from their data collections. This ability depends not only on being able to find data, but also on understanding the characteristics and quality of the data. Metadata, 'data about data,' provide information such as the characteristics of a data set, the history of a data set, and organizations to contact to obtain a copy of a data set. Standardized metadata elements provide a means to document data sets within an organization, to contribute to catalogs of data that help persons find and use existing data, and to help users understand the contents of data sets that they receive from others. Metadata support activities include determining existing data, determining the appropriate data set for a particular use, providing information to access data sets, and providing information to transfer data sets for processing and use.

While the benefits of metadata are manifest, several impediments hinder uniform adoption and use. These center on two topics: agreement on what metadata elements to collect, and the level of effort required to collect whatever metadata are required. Obviously, the two items are linked. From the user perspective, detailed metadata are highly desirable. More information will help the user determine the suitability of a given data set. From the data producer's perspective, metadata cost money; metadata elements should be constrained to a cost-effective set. Significantly more resources may be required to collect metadata about 'legacy' data holdings than data sets collected using techniques that capture the metadata as a part of the entire data-collection process.

Agreement on standardized terminology and definitions would not only provide benefits in the automation of the collection of metadata, but would also be exceedingly valuable in searching data catalogs. Several metadata 'standards' are in widespread use around the world. In general, they have data elements that have been selected to provide information about four topics:

Availability – data needed to determine the sets of data that exist for a geographic location;
Fitness for use – data needed to determine if a set of data meets a specified need;
Access – data needed to acquire an identified set of data; and
Transfer – data needed to process and use a set of data.

These elements form a continuum where a user cascades through a pyramid of choices when determining what data are available, evaluating the fitness of data for use, accessing the data, and transferring and receiving the data. The exact nature of these data elements continues to be the subject of much debate and discussion.

A collection of metadata records combined with data management and search tools forms a data catalog. The architecture of these data catalogs can vary from tightly held internal data indexes to widely accessible distributed catalogs spread across the Internet. Establishing a data catalog can be a major project. Among the items that need to be accomplished are to

- determine which programs or individuals produce, manage, or disseminate data that are to be included in the catalog;
- identify what geospatial data an organization holds, where data are located, and their condition;
- identify priorities for documenting previously collected or produced data based on demand, geographic extent, uniqueness, or other criteria to be determined by the organization;
- create the metadata for new spatial data;

- ensure that the metadata comply with applicable standards, to the extent possible for the organization;
- provide training and technical assistance to metadata producers, quality assurance personnel, and end users;
- update and maintain metadata as necessary;
- specify, acquire, and manage the hardware, software, and telecommunications services so that the metadata are electronically accessible;
- determine how to link or cross-reference spatial data and metadata to ensure consistency, maintainability, easy query or search, and easy access; and
- respond to questions about or orders for spatial data.

Clearly, the catalog function can go far beyond the basic listing of content. Extensions that allow for direct ordering or download of data, user comment or feedback, as well as requests for new data collections are all possible. Metadata and data catalogs are the means by which data users can more economically find, share, and use spatial data.

Ubiquitous Spatial Data

Prior to the twenty-first century, the creation and use of spatial data were associated, for the most part, with large, immobile, and often very expensive computers and specialized technology. The capability to exploit spatial data was limited to a few tens of millions of devices. The development of highly portable, locationally aware, computer devices (such as smartphones) has dramatically increased the role of spatial data in daily life. These mobile computing devices constantly utilize a variety of location-based services. Every smartphone is a potential user and creator of spatial data. The data may be as simple as the location of the phone at a given point in time. Or it could be a photograph tagged with a time and location. Virtually any observation that can be matched with a geographic location becomes spatial data. When networked to a cloud computing environment, the data created by an individual can be shared with over a billion people. As smartphones and the supporting technology spreads further, this number will probably increase to 2 billion by 2016.

Spatial data have gone from a controlled/centralized enterprise (dominated by government agencies and large commercial firms) to a decentralized activity, in which individuals can collect new spatial data, annotate existing data (e.g., with additional attributes or photos), and flag information that is out of date. Individuals can collect this information for their own use, or share it with the others in loosely collaborative ventures. Spatial data for entire countries previously void of freely available detailed data (e.g., North Korea) have been collected in this fashion. The product of this type of data collection and sharing has been commonly termed 'crowdsourced' or 'volunteered geographic information' (VGI).

The fact that the data are 'volunteered' causes VGI to be opportunistic in nature as opposed to being systematically collected. Some areas may have rich data collections with detailed depictions of natural and man-made features. Other areas close by may have relatively little data, simply because no one was interested in collecting data for that geographic region. The laissez-faire philosophy of VGI usually means that there are no standards or specifications for the content, character, or accuracy of the data being collected. This results in a very heterogeneous collection when data are assembled to cover a larger geographic region.

Where VGI can excel is in the near-real time collection of spatial data. VGI in support of disaster relief efforts, for example, has been invaluable. VGI can characterize ephemeral events such as floods or traffic congestion. The stream of geolocated data from these devices reflects the interests and motivation of the people who use them. In responding to an earthquake, for example, local citizens can map and describe damage with greater quickness and agility than a centralized organization.

A major difficulty is how to use VGI to enhance the general reference sets of spatial data that typically have been provided by government agencies or commercial firms (sometimes called base category data). VGI rarely has any associated metadata or other data quality information. Those responsible for maintaining the base category data must determine which sets of VGI should be used (if any) to modify their collections. The temptation is to ignore most of it. If data are of unknown and/or unknowable quality, then it cannot be used. It is the philosophy of some organizations that solicit VGI that the data ingested should be accepted as reported and changed by other users, over time, if it is not accurate. A common commercial approach is to use VGI as measures of change detection that are subsequently examined using the preferred techniques of the database providers. In essence, there are a variety of approaches to collecting, transmitting, and evaluating VGI. Finding viable solutions for exploiting the VGI spatial data streams will remain a significant challenge in the future.

See also: Citizen Mapping; Geographic Information Systems; Global Positioning System; Location, Absolute and Relative; Remote Sensing; Spatial Data Infrastructures.

Bibliography

Guptill, S.C., Morrison, J.L. (Eds.), 1997. Elements of Spatial Data Quality. Pergamon, Oxford, UK.

Langran, G., 1992. Time in Geographic Information Systems. Taylor & Francis, Philadelphia, PA.

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (Eds.), 1999. Geographical Information Systems, second ed. John Wiley, New York.

Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Kimlerling, A.J., Guptill, S.C., 1995. Elements of Cartography, sixth ed. John Wiley, New York.

Sui, D., Elwood, S., Goodchild, M. (Eds.), 2013. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, Heidelberg.

Spatial Data Infrastructures

Max Craglia, European Commission, Joint Research Centre, Ispra, Varese, Italy

© 2015 European Union. Published by Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Rhind, volume 22, pp. 14778–14785, © 2001, Elsevier Ltd.

Abstract

This article reviews the key developments in spatial data infrastructures (SDIs) over the last 10 years. The article raises three main issues: (1) SDIs are more complex than we first thought, and are more about building social networks than technologies and standards; (2) standards are helpful but when dealing with complex and multidisciplinary infrastructures they are not enough; and (3) the future is about integrating authoritative public sector data with information coming from citizens to have more participative environments in which to share knowledge.

Introduction

In the first edition of this Encyclopedia, Rhind (2001) introduced spatial data infrastructures (SDIs) as an extension of geographic information systems (GIS). Whereas in a GIS the emphasis is on the analysis of the data acquired through various means, in SDIs the emphasis is on the process of publishing and sharing data across multiple organizations and users. SDIs could be seen, in a sense, as the first and last step of an information flow that has GIS in the middle. First, users look in the catalog of one or more SDIs to see if the data they need are available, then they download the data to their GIS, do the analysis, and finally (often but not always) they publish the results in an SDI for others to share. This conceptualization of SDIs reflected the view of the time: in 2000–01, SDIs were emerging in many countries largely following the US model based on a set of key datasets (the foundation framework) underpinning many possible applications, a catalog of data available, and a clearinghouse where these data could be accessed. There was also a belief that adopting a few agreed standards would be sufficient to ensure the interoperability across organizations publishing and using the data.

In this article we revisit the concepts and operations of SDIs in the intervening decade and discuss three key issues: (1) SDIs are more complex than we first thought, and are more about building social networks and reaching agreements on the semantics of data than on technologies and standards; (2) standards are helpful, but when dealing with complex and multidisciplinary infrastructures, they are not enough: there will always be different standards and practices across different communities, and it is better to accept diversity than trying to impose uniformity that is unlikely to work; and (3) SDIs are still used primarily by public administrators and policy analysts, while a few large private sector companies have dominated the global market for the general public, enabling the dissemination and use of spatial information to an extent that could not be foreseen in the early 2000s. The way forward is for SDIs to articulate and demonstrate better their value proposition, open up to societal use by integrating data from citizens into a more participatory notion of data infrastructures. To address the first issue, I will make reference to the development of the Infrastructure for Spatial Information in Europe (INSPIRE), for the second, I will make reference to a Global Earth Observation System of Systems (GEOSS), and for the third, I will introduce the opportunities for harnessing citizen science.

SDIs as Social Networks: The INSPIRE Example

INSPIRE is a European legal act (Directive 2007/2/EC) setting up an infrastructure for spatial information in Europe based on infrastructures for spatial information established and operated by the 28 member states of the European Union (http://inspire.jrc.ec.europa.eu/). The EU is a union of 28 sovereign member states that agree through a series of international treaties (the latest being the Lisbon Treaty of 2009) the policy areas in which they wish to share responsibilities and resources, (e.g., agricultural, environmental, and regional policies) and those that remain instead the exclusive domain of the national governments (e.g., defense and immigration). The key decision-making bodies are therefore the national governments represented in the Council, each with a number of votes proportional to the size of the country, and the European Parliament that is elected by universal suffrage every 5 years. The European Commission is the civil service body of the EU and has the power of proposing legislation (to the Council and European Parliament), and then monitoring its implementation once approved. Not being a federal state also means that there is no equivalent to the US federal agencies collecting topographic or demographic data like the US Geological Survey and the Bureau of the Census. All data in the EU comes via the responsible organizations in the member states. As a result, setting up an EU-wide SDI can only be done in a decentralized way building on the SDIs and related activities established and maintained by member states.

The purpose of INSPIRE is to support environmental policy and overcome major barriers still affecting the availability and accessibility of relevant data. These barriers include:

- inconsistencies in spatial data collection: spatial data are often missing, incomplete, or alternately, the same data are collected twice by different organizations;
- lack or incomplete documentation of available spatial data;
- lack of compatibility among spatial datasets that cannot, therefore, be combined with others;

- incompatible SDI initiatives in a member state that often function only in isolation; and
- cultural, institutional, financial, and legal barriers preventing or delaying the sharing of existing spatial data.

The key elements of the INSPIRE directive to overcome these barriers include:

- metadata to describe existing information resources so that they can be more easily found and accessed;
- harmonization of key spatial data themes needed to support environmental policies in the Union;
- agreements on network services and technologies to allow discovery, view, download of information resources, and access to related services;
- policy agreements on sharing and access, including licensing and charging; and
- coordination and monitoring mechanisms.

INSPIRE addresses 34 key spatial data themes organized in three groups (or Annexes to the Directive) reflecting different levels of harmonization expected and a staged phasing (see Table 1).

Table 1 Key data themes addressed by INSPIRE

Annex I
Coordinate reference systems
Geographical grid systems
Geographical names
Administrative units
Addresses
Cadastral parcels
Transport networks
Hydrography
Protected sites
Annex II
Elevation
Land cover
Ortho-imagery
Geology
Annex III
Statistical units
Buildings
Soil
Land use
Human health and safety
Utility and governmental services
Environmental monitoring facilities
Production and industrial facilities
Agricultural and aquaculture facilities
Population distribution – demography
Area management/restriction/regulation zones and reporting units
Natural risk zones
Atmospheric conditions
Meteorological geographical features
Oceanographic geographical features
Sea regions
Biogeographical regions
Habitats and biotopes
Species distribution
Energy resources
Mineral resources

INSPIRE, Infrastructure for Spatial Information in Europe.

The legal framework of INSPIRE has two main levels. At the first, there is the INSPIRE Directive itself, which sets the objectives to be achieved and asks member states to pass their own national legislation establishing SDIs. This mechanism of European plus national legislation allows each country to define its own way of achieving the agreed-upon objective by taking into account its own institutional characteristics and history of development. As an example, Germany does not have a single SDI but a coordinated framework among each of its 16 states (Länder), plus one at the federal level (which also means that 17 different legal acts had to be passed to implement INSPIRE). Similarly, Belgium has three SDIs, one for each of its regions (Wallonia and Flanders) and one for Brussels. The INSPIRE directive also requires the establishment of an EU geoportal operated by the European Commission to which the infrastructures of the member states have to connect (see http://www.inspire-geoportal.eu).

To ease the difficulties of having multiple versions of INSPIRE, the directive envisages a second level of legislation, more stringent because it has to be implemented as is, and does not require follow-up national legislation. In European terminology this is called a Regulation. Therefore, INSPIRE envisages technical implementing rules in the form of formal legal acts (Regulations) for metadata, harmonization of spatial data and services, network services, data and service sharing policies, and monitoring and reporting measures to evaluate the extent of the directive's implementation and to assess its impact. Each of these Regulations needs the approval of the member states and of the European Parliament. The first implementing rules were approved in December 2008 for metadata, while the last set of implementing rules to specify the data models of Annex III data themes and for the interoperability of spatial data services were approved in 2013.

INSPIRE has some characteristics that make it particularly challenging. The most obvious is that it is an infrastructure built on those of 28 different countries in 24 languages. The requirements for multilingual services and interoperability among very different information systems and professional and cultural practices are, therefore, very demanding. This means, for example, that existing standards, where they exist, have to be tested in real distributed and multilingual settings. In the best scenario, all works well but for a European-wide implementation there is a need to translate the standards and related guidelines into the relevant languages (the standards by the International Standards Organisation (ISO) and the Open Geospatial Consortium (OGC) are typically in English only). In other instances, testing has demonstrated that the standards are not mature enough, or leave too much room for different interpretations, and thus they require further definition or individual bridges to make different system interoperate. In harder cases still, there are no standards available, and therefore they have to be created. This applies, for example, to the specifications required for the interoperability of spatial datasets and services, which is a central feature of INSPIRE.

Recall that each country in Europe has its own heritage and traditions that include different ways and methods for collecting environmental and geographic data, different traditions on how to analyze and visualize them, and different coordinate reference systems (sometimes more than one in each country), projections, and vertical reference systems. It is not enough,

therefore, for an SDI in Europe to help users find and access data; it is also necessary to understand the meaning of what one is accessing to make appropriate use of it. This means, in turn, that we need to develop not only translation tools to help overcome the 'natural' language barriers, but also common reference frameworks, classification systems and ontologies, data models, and schemas for each of the data themes shown in Table 1, against which the national data can be transformed or mapped. This is necessary because it is not possible to ask the member states and their national and local organizations to reengineer all their databases. Thus, the approach adopted is to develop agreed-upon European models and systems of transformation (on-the-fly or batch) so that the level of interoperability necessary for key European applications can be achieved. The approach sounds simple, but putting it into practice and sustaining it over a long period is complex, as it required 7 years of work to develop an accepted methodology (the Generic Conceptual Model) and tools, mobilize hundreds of experts in different domains, and deliver and test the specifications for the datasets falling in each of the themes identified in the Annexes of the directive. A visit to the INSPIRE Website (http://inspire.jrc.ec.europa.eu/index.cfm) in the data specifications' sections demonstrates the huge amount of work involved: Thousands of pages of specifications and tens of thousands of comments had to be addressed individually during the stakeholder consultations.

The organizational model put in place to develop INSPIRE is one of its more interesting features, drawing significant attention from outside Europe. In essence, it is a huge exercise in public participation, the like of which is most unusual in policy making. From the outset it was recognized that for INSPIRE to be successful and overcome the barriers to data access and use identified earlier, it was necessary for the legislator, implementers, and practitioners in the member states to come together and agree on a shared understanding of the problem and possible solutions. Therefore, an expert group with official representatives of all the member states was established at the beginning of the process in 2001, together with working groups of experts in the fields of environmental policy and geographic information to formulate options and forge consensus. The INSPIRE proposal was subject to an extended impact assessment (http://inspire.jrc.ec.europa.eu/reports/fds_report.pdf and http://inspire.jrc.ec.europa.eu/reports/inspire_extended_impact_assessment.pdf) to identify potential costs and benefits, before opening for public consultation. The revised proposal was then debated by the Council and European Parliament over a 3-year period before final adoption in 2007. This process in itself is a good example in democracy, but more interesting is the way in which interested stakeholders are continuing to participate in all the ongoing activities required to develop the INSPIRE implementing rules (i.e., the follow-up legal acts and detailed technical guidance documents).

To organize this process, two mechanisms have been put in place: the first is to engage the organizations at European, national, and subnational levels that already have a formal legal mandate for the coordination, production, or use of geographic and environmental information (the so-called Legally Mandated Organisations (LMOs)). The second mechanism aims to facilitate the self-organization of stakeholders, including spatial data providers and users from both the public and private sectors, in Spatial Data Interest Communities (SDICs) by region, societal sector, and thematic issue. The central roles that the SDICs play in the development of implementing rules include:

- to identify and describe user requirements (to be understood as acting in-line with environmental policy needs, as opposed to 'maximum' requirements beyond the scope of INSPIRE and beyond realistically available resources);
- to provide expertise to INSPIRE drafting teams;
- to participate in the review process of the draft Implementing Rules;
- to develop, operate, and evaluate the implementation pilots; and
- to develop initiatives for guidance, awareness raising, and training in relation to the INSPIRE implementation.

LMOs have similar functions but play also a central role in reviewing and testing the draft implementing rules and in assessing their potential impacts in respect to both costs and benefits.

Open calls for registration of interest by SDICs and LMOs were launched in 2005 and 2009 to put forward experts, paid by their own organization, and reference material supporting the preparation of the implementing rules. The response was immediately very good with thousands of experts and users able to contribute to this collective effort. Table 2 shows the extent of the community directly involved in shaping the policy and the technical documents.

Three aspects are particularity important in understanding the work and the challenges of the drafting teams: first, each expert represents a community of interest and therefore has the responsibility to bring to the table the expertise, expectations, and concerns of this community; second, each drafting team has to reach out to all the thematic communities that are addressed by INSPIRE. This is no small undertaking as the directive covers more than 30 different data themes (Table 1), which are the responsibility of multiple agencies at national regional and local levels across 28 countries. As a matter of comparison, it is worth recalling that the US National Spatial Data Infrastructure defined only seven framework themes (http://www.fgdc.gov/framework/handbook/index_html): geodetic control, ortho-imagery, elevation, transportation, hydrography, governmental units, and cadastral information, most of which have a federal agency taking the lead in data collection and management (see Rhind, 2001).

The implication for the drafting teams is that they have a much more difficult task in collecting and summarizing reference material, seeking common denominators and reference models, and developing recommendations, which satisfy user requirements without imposing an undue burden on those organizations that have day-to-day responsibility for data collection and management across Europe. Seeking compromise among different requirements and perspectives is crucial

Table 2 The INSPIRE community in 2012

- 491 Spatial Data Interest Communities
- 295 Legally Mandated Organisations
- 259 Experts (drafting teams and working groups)
- 3031 User organizations registered on the INSPIRE Website

to the work of each drafting team. Last it is important to note that the drafting teams retain ownership of their work. They make the recommendations and submit them to review to all the registered SDICs and LMOs, and the representatives of the member states. It is only after they have taken on board all the comments received that the European Commission takes ownership of the draft implementing rule and submits it to internal consultation. After revision and checking, the draft implementing rule goes through the final round of the democratic process before becoming a new legal act. This involves qualified majority voting by the representatives of the member states and the scrutiny of the European Parliament.

The complexity of this participatory approach is certainly innovative, not only in relation to the developments of SDIs but also more generally to the formulation of public policy at the European level. The outcome produces both consensus-based policy and the development and maintenance of a network of stakeholders that make it possible to implement more effectively this distributed European SDI. Participation is good, but it takes a lot of effort. In particular, maintaining the momentum of the process, commitment of hundreds of organizations and thousands of experts, and building a relationship of trust across multiple levels of governance have not been easy. It has nevertheless been essential. Without this social network, it would have not been possible to develop INSPIRE. What is interesting, in particular, is that now the network is used by the member states themselves to help them implement the technical infrastructure, sharing expertise, funding, and technical components across multiple member states. As an example, four neighboring countries, Belgium, the Netherlands, Germany, and the UK, organized in 2013 a conference to share experience in implementing INSPIRE with over 250 participants from multiple domains in both public and private sectors (www.poweredbyinspire.eu).

Another facet, which is even more important, is the organizational challenge in the member states to implement INSPIRE. The INSPIRE directive asking member states to establish and maintain their SDIs and set up appropriate coordinating mechanisms has given rise to a flurry of activity across Europe. In many countries, SDIs already existed and worked well at national or subnational levels. So the effort was more focused on agreeing to a division of responsibility than on setting up new structures. In other countries, INSPIRE offered an opportunity for the organizations that have been leading SDI developments for years to get their just recognition and acquire new status and legitimacy. In others still, the organizations and divisions of responsibility had to be set up from scratch. Of course, the difficult financial climate since 2008–09 has made it more challenging to invest in new infrastructures and ways of working. Hence, one of challenges in most countries has been to leverage resources available from different sources (European, national, international) and/or ensure strong synergy between the investment required by INSPIRE and those committed in related projects, for example, in the framework of e-government. In this sense, the work needed is critical, not only to align sources of funding but also to ensure that initiatives, standards, systems, and deployments are well coordinated and that they do not duplicate, or contradict, each other. This requires extending the networks of individuals and organizations built during the INSPIRE process to other domains both vertically (from national to local) and horizontally, across sector of activity – a nontrivial challenge. Moreover, as we try to work across domains we find another level of interoperability needed: from semantic to multidisciplinary interoperability.

Multidisciplinary Interoperability

A major challenge facing humanity at the beginning of the twenty-first century is to respond effectively to the global changes that are increasing pressure on the environment and on human society. In this context, the International Council for Science (ICSU) has called for

> 'global sustainability research'.
> (ICSU, 2010: p. 6).

ICSU identified five scientific priorities, or Grand Challenges, in global sustainability research through a broad consultation that was carried out in 2009–10 and involved over 1000 scientists from 85 countries:

1. Developing, enhancing, and integrating the observation systems that are required to manage global and regional environmental change.
2. Improving the usefulness of forecasts of future environmental conditions and their consequences for people.
3. Recognizing key thresholds or nonlinear changes to improve our ability to anticipate, recognize, avoid, and adapt to abrupt global environmental change.
4. Determining which institutional, economic, and behavioral responses can enable effective steps toward global sustainability.
5. Encouraging innovation (coupled with sound mechanisms for evaluation) in developing technological, policy, and social responses to achieve global sustainability.

To meet these challenges, ICSU argued for moving from research dominated by the natural sciences to research involving the full range of sciences and humanities, with a particular role for the social sciences to be more strongly involved in natural science research. It further argued that there is also an important need to involve users of science more closely in a way that is reminiscent of the participatory approach adopted by INSPIRE and described in Section SDIs as Social Networks: The INSPIRE Example:

> Research will often be most useful, and the results most readily accepted by users, if priorities are shaped with the active involvement of potential users of research results and if the research is carried out in the context of a bi-directional flow of information between scientists and users.
> (ICSU, 2010: p. 6)

Global sustainability research requires global data, and an infrastructure to deliver it. The Global SDI initiative described in Rhind (2001) has provided a useful forum to exchange experiences at the global level but never materialized into a technical infrastructure. Instead, the GEOSS, envisioned by the group

of eight most industrialized countries (G-8) in 2003 and currently two-thirds in its 10-year implementation plan, set out to provide the framework for integrating the earth observation efforts of the Group on Earth Observation (GEO), which includes 90 countries and more than 60 international organizations. GEOSS is a voluntary process on a best effort basis by the contributing organizations. It is therefore very different from INSPIRE which, being a legal framework, can impose standards and specifications. The GEOSS Work Programme is articulated in 26 tasks led by different organizations on a voluntary basis (see http://earthobservations.org/geoss_imp.php) and is underpinned by the GEOSS Common Infrastructure (GCI).

The GCI was launched in June 2008 and comprised three main sets of components: a Geo Web Portal (GWP) acting as entry points for the users, a set of registries (components and services, standards and interoperability, and best practice), and a clearinghouse. Data providers were encouraged to register their systems, data resources, and services in the Content and Service Registry (CSR), and the Clearinghouse harvested the metadata from the CSR, acting as a catalog to respond to user requests made through the GWP. During the initial phase of operations, there were three GWPs and three clearinghouses provided on a voluntary basis by organizations in both private and public sectors. In 2010, after a process of testing and review, it was decided to select the GWP operated by the European Space Agency, and the Clearinghouse operated by the US Geological Survey as the bases for the operations of the GCI.

By 2011, it became clear that there was a problem. Only a few hundred datasets and services had been registered by data providers and were available through the GCI, thus making the infrastructure of limited or no value. The approach to developing the GCI was copied from the US federal National Spatial Data Infrastructure, and similar efforts in Canada and Europe (including INSPIRE), which were considered best practice at the time. In fact the NSDI Cookbook (Niebert, 2004) has been for many years the template from which SDIs around the globe were developed. The problem is that this 'traditional' approach to SDI development is founded on the adoption of a limited set of standards that have emerged in the geospatial community (typically those by ISO and OGC), and on having the authority to enforce these standards on the contributing organizations. Already in the case of INSPIRE where the authority exists via the European legal framework, we have tested the limits of this approach when dealing with an infrastructure spanning multiple countries and languages. In the context of GEOSS, that is a global initiative, across many disciplines in the physical, environmental, and social sciences, and built on a voluntary basis (and no additional funding), the approach was faltering. Each discipline had developed over the years its own practices, standards, and protocols, and could see no benefit in making costly changes to adopt the standards requested by the geospatial community. Hence, only a low level of registration and interoperability had been achieved by the GCI by 2011.

A novel approach that was able to make a significant change was the introduction of the Brokering Framework (Nativi and Bigagli, 2009; Nativi et al., 2012, 2013). This framework builds the necessary bridges between different disciplinary infrastructures without asking them to make any changes to their normal way of operating. In more formal terms, brokering introduces a new middleware layer of service offerings in the SDI that realize the necessary mediation, adaptation, distribution, semantic mapping, and even quality checks required to address the complexity of a multidisciplinary infrastructure built by interconnecting disciplinary infrastructures. Brokers can do much more than just facilitate discovery and access of available resources: they may implement new capabilities for users and data providers such as advanced semantic discovery; resource tagging and clustering of results; quality control; data preprocessing and transformation; work flow management; and model interconnection. In essence, there are multiple brokers taking on different functions, but the overall new approach is to concentrate all the mediations in a separate middle layer of the infrastructure without asking each (disciplinary) infrastructure to make changes. The added value of the brokering approach was demonstrated in December 2011 at the plenary meeting of GEO: from a few hundred resources available in the GCI in March 2011, the introduction of brokering made it possible by December to discover some 28 million datasets or services, a two orders of magnitude increase in few months! For this reason, brokering was officially adopted as a central component of GEOSS in 2012 and is now the core of the GCI. Similar complex multidisciplinary infrastructures such as the National Science Foundation EarthCube (http://earthcube.ning.com/) are also considering the brokering approach (Nativi et al., 2011).

From Multidisciplinary to Participatory Science

As noted at the beginning of Section Multidisciplinary Interoperability, ICSU recognized the crucial importance of engaging more closely the users of scientific research, citizens, and decision makers, not just by communicating science more effectively but by fostering the active involvement of users in the scientific process itself. This form of participatory science is all the more necessary as the explosion of social media exposes scientists more than ever before to claims and counterclaims, and raises the need for transparency and openness of both data and processes on which scientific discoveries are made. This is the sense of the new Open Data initiatives in Europe and globally (see for example the Research Data Alliance http://rd-alliance.org). Social media provides also many new opportunities for participatory science, with citizens acting as producers and not just consumers of information.

There is long tradition of calling on volunteers to provide information relevant to science, as for example the Bird Count in the USA (http://www.birdsource.org/gbbc/) and Spring Watch in the UK (http://www.bbc.co.uk/nature/uk/). What has made a significant difference recently is the diffusion of the Internet-based and social networks that have exponentially increased the amount of geographically located information produced by individuals. In August 2006, geotagging facilities at Flickr started to operate; and by the year of 2007, more than 20 million geotagged photos had been uploaded to Flickr. In August 2011, Flickr announced its 6 billionth photo, with an increasing 20% year-over-year, over the last 5 years (http://blog.flickr.net/en/2011/08/04/6000000000/). Similarly, Twitter was born in 2006. The most impressive information is the increasing rate of messages. In 2010, the average number of Tweets sent per day was 50 million (http://blog.

twitter.com/2011/03/numbers.html), while in March 2012 it was 340 million (http://blog.twitter.com/2012/03/twitter-turns-six.html). Also in 2010, the geotagging feature of the Tweets was added. Although the amount of geo-enabled messages is around 1% of all messages, this still means 1 million geotagged information messages per day. The universe of Volunteered Geographic Information (VGI) (Goodchild, 2007), or neo-geography, is not uniform and Coleman et al. (2009) provide a useful categorization of contributors of VGI, while Craglia et al. (2012a) report on a large-scale experiment to assess the quality of VGI and its added value compared to official information in the case of forest fires.

The integration of SDIs, which are traditionally based on authoritative information from public administrations, with information from citizens and new networks of sensors promises a real step change: form static to dynamic information flows about the Earth, from a few users in the scientific or policy-making community to a participatory framework for all. Digital Earth is the term that encompasses this vision of a dynamic and participatory framework, the next frontier of SDIs (Goodchild et al., 2012; Craglia et al., 2012b). The next editions of this Encyclopedia will document how far promise has matched reality in implementing this vision.

Disclaimer

The views expressed in this article are of the author's alone and do not represent necessarily those of the Joint Research Centre or of the European Commission.

See also: Citizen Mapping; Citizen Science; Data Bases and Statistical Systems: Environment and Urban Studies; Geographic Information Systems; Information Society; Science, Public Engagement With; Spatial Analysis in Geography; Spatial Data; Spatial Thinking in the Social Sciences.

Bibliography

Coleman, D., Georgiadou, Y., Labonte, J., 2009. Volunteered geographic information: the nature and motivation of producers. International Journal of Spatial Data Infrastructures Research 3, 332–358.

Craglia, M., Ostermann, F., Spinsanti, L., 2012a. Digital Earth from vision to practice: making sense of citizen-generated content. International Journal of Digital Earth 5 (5), 398–416. http://dx.doi.org/10.1080/17538947.2012.712273.

Craglia, M., de Bie, K., Jackson, D., Pesaresi, M., Remetey-Fülopp, G., Wang, C., Annoni, A., Bian, L., Campbell, F., Ehlers, M., van Genderen, J., Goodchild, M., Guo, H., Lewis, A., Simpson, R., Skidmore, A., Woodgate, P., 2012b. Digital Earth 2020: towards the vision for the next decade. International Journal of Digital Earth 5 (1), 4–21.

Goodchild, M.F., 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. International Journal of Spatial Data Infrastructures Research 2, 24–32.

Goodchild, M.F., Guo, H., Annoni, A., et al., 2012. Next generation digital Earth. Proceedings of the National Academies of Science 109 (28), 11088–11094.

International Council for Science (ICSU), 2010. Grand Challenges in Global Sustainability Research: A Systems Approach to Research Priorities for the Decade. http://www.icsu.org/publications/reports-and-reviews/grand-challenges/GrandChallenges_Oct2010.pdf (accessed 27.04.13.).

Nativi, S., Bigagli, L., 2009. Discovery, mediation, and access services for earth observation data. Applied Earth Observations and Remote Sensing, IEEE Journal of 2 (4), 233–240.

Nativi, S., Craglia, M., Pearlman, J., 2012. The brokering approach for multidisciplinary interoperability: a position paper. International Journal of Spatial Data Infrastructures Research 7, 1–15.

Nativi, S., Craglia, M., Pearlman, J., 2013. Earth science infrastructures interoperability: the brokering approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6 (3), 1118–1129.

Nativi, S., Khalsa, S.J., Domenico, B., Craglia, M., Pearlman, J., Mazzetti, P., Rew, R., 2011. The Brokering Approach for Earth Science Cyberinfrastructure. EarthCube white paper. US NSF. http://semanticommunity.info/@api/deki/files/13798/=010_Domenico.pdf.

Niebert, D. (Ed.), 2004. SI Reference Manual: Cookbook. FGDC. http://www.fgdc.gov/international/sdi_cookbook.

Rhind, D., 2001. Spatial Data Infrastructures. International Encyclopedia of the Social & Behavioral Sciences. Elsevier, pp. 14778–14785.

Spatial Decision Support Systems

Robert G Pontius Jr. and Kangping Si, Clark University, Worcester, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Rushton, volume 22, pp. 14785–14788, © 2001, Elsevier Ltd.

Abstract

Spatial decision support systems (SDSS) combine storage, search, and retrieval capabilities of geographic information systems with decision models and optimizing algorithms to support decision making concerning spatial problems. These systems allow decision makers to use multiple spatial criteria to make locational choices by exploring alternatives, given spatial and attribute information. Characteristics of SDSS are becoming incorporated in social science theories and models that explain spatial decisions.

Introduction

Spatial decision support systems (SDSS) are computer-based systems that combine storage, search, and retrieval capabilities of geographic information systems (GISs) with decision models and optimizing algorithms to support decision making concerning spatial problems. Applications include political redistricting, route selection, facility location, and land use planning. SDSS are used increasingly in both private and public sectors. These systems allow decision makers to use multiple spatial criteria to make locational choices by exploring alternatives, given spatial and attribute information. In some SDSS, the users generate alternatives interactively. The capabilities of the SDSS manipulate the geographic information, report the characteristic attributes for each alternative, and allow users to participate in the choice evaluation process. Research in SDSS entails capturing and incorporating the knowledge of experts.

People frequently make choices that involve evaluating geographic context. Decision makers apply multiple spatial criteria to complex geographic data, such as information that relate to selecting routes, making service areas, or defining environmental hazardous areas. In social science, such complex decision-making problems are often ill-structured, in which case a set of theories, methods, and techniques must be developed. It has become common to link these together in computer software that operates on spatial data.

A number of challenges may occur when making decisions for ill-structured problems. First, it may be unclear how much of the attainment of one spatial criterion should be sacrificed in order to attain a given amount of another. The decision maker often wants to explore alternative solutions and reach a decision about a desirable solution after extensive exploration of the spatial context and the attribute information. That is, decision makers' preferences for a particular solution may be influenced by what can be attained in a particular context. Decision makers may have preferences for abstract entities, and the implementation of these preferences is mediated by knowing what can be attained. SSDS are designed to facilitate the exploration of these spaces, and thus to assist the search for acceptable solutions. A second challenge is the imperfection of spatial databases with respect to missing variables and erroneous data. Decision makers will sometimes consider their personal knowledge or search for additional facts before making a decision. A third challenge is that decision makers may choose not to accept a model's results due to concerns regarding the results' validity. All three challenges can lead decision makers to ask for computer-based, interactive, support systems that explore the problem and search for various solutions.

Solutions to SDSS problems normally have both spatial and nonspatial characteristics. Spatial characteristics concern the shapes of features and the distances among the features. Nonspatial characteristics are the attributes of those features, such as can be recorded in a table. Solutions to all SDSS problems exist in geographic space as well as in attribute information. As the geographic space of the solution to any problem changes, its attribute information changes too. For example, in making districts from which voters will elect representatives, the geographic solution space includes the map of the boundaries of the districts, the shapes of the districts, and the connectivity of census tracts that belong to the same district (Ricca and Simeone, 2008). The corresponding attribute information describes numbers of eligible voters in each district, voters' party affiliations, voters' past behaviors, and any attribute for which data are available and that a politician may view as relating to voters' likely future voting behavior. Such attributes might be voters' ethnicity, income, age, religion, etc. Any participant in a redistricting exercise may want to see how any change in the boundaries is accompanied by changes in any or all of these attributes. The redistricting process is to find solutions using multiple criteria concerning geographic space and attribute information. For example, typical spatial criteria for political redistricting are each district should be contiguous, districts should follow natural boundaries like bodies of water, districts should respect the integrity of communities, and districts should be compact to prevent the formation of odd-shaped districts that raise the suspicions of gerrymandering (Bozkaya et al., 2003). The nonspatial attribute information of the districts include criteria such as each district should have an approximately equal number of people, each district should have socioeconomic homogeneity to ensure representation of residents who share concerns, the collection of districts should allow representation of important minority groups, etc. Redistricting uses quantifiable criteria to compare one solution with another to find the better ones, and to exclude solutions that do not satisfy the criteria. A system for

simultaneously computing various regionalizations of territory constitutes an SDSS for electoral redistricting.

Components of SDSS

Figure 1 shows the components of a common SDSS as well as the flow of the decision-making process using SDSS. Besides the user interface and report generator, the SDSS consist mainly of three subsystems: a GIS, a model management system, and a group decision support system. After users define their spatial decision task, they interact with the three subsystems through the user interface of SDSS. GIS helps to visualize the data, analyze the data using spatial analysis tools, create supportive data layers, and store the data using the database management system. The graphical interface of SDSS also allows users to formalize their problems so that users can understand or construct their models using components that the model management system has. The model management system is then able to use the data managed by GIS in order to find a solution using existing algorithms. The group decision support subsystem provides cooperative work computing environments that can enhance SDSS by enabling group members to observe, manipulate, and evaluate solutions collectively, thus providing mechanisms for conflict resolutions. Finally, the SDSS generate reports so users can evaluate all the alternatives. Users may adopt the solution or go back to redefine the problem and redo the entire process to find a better solution.

Role of GIS

Various types of SDSS can be defined according to how geographic space and attribute information are generated. SDSS consist of tools for manipulating geographic information and reporting the attributes of the outcomes. The simplest SDSS incorporate the capacity to convey information from the geographic space to the attribute information or from the attribute information to the geographic space. These are key capabilities of a GIS.

A GIS is a system of hardware, software, and procedures that can support geographical decision making through the acquisition, management, manipulation, analysis, and display of spatially referenced information (Church and Murray, 2009). GIS provides decision makers the ability to visualize and analyze the input data as well as to model results from the decision-making process (Murray, 2010). Over the last several decades,

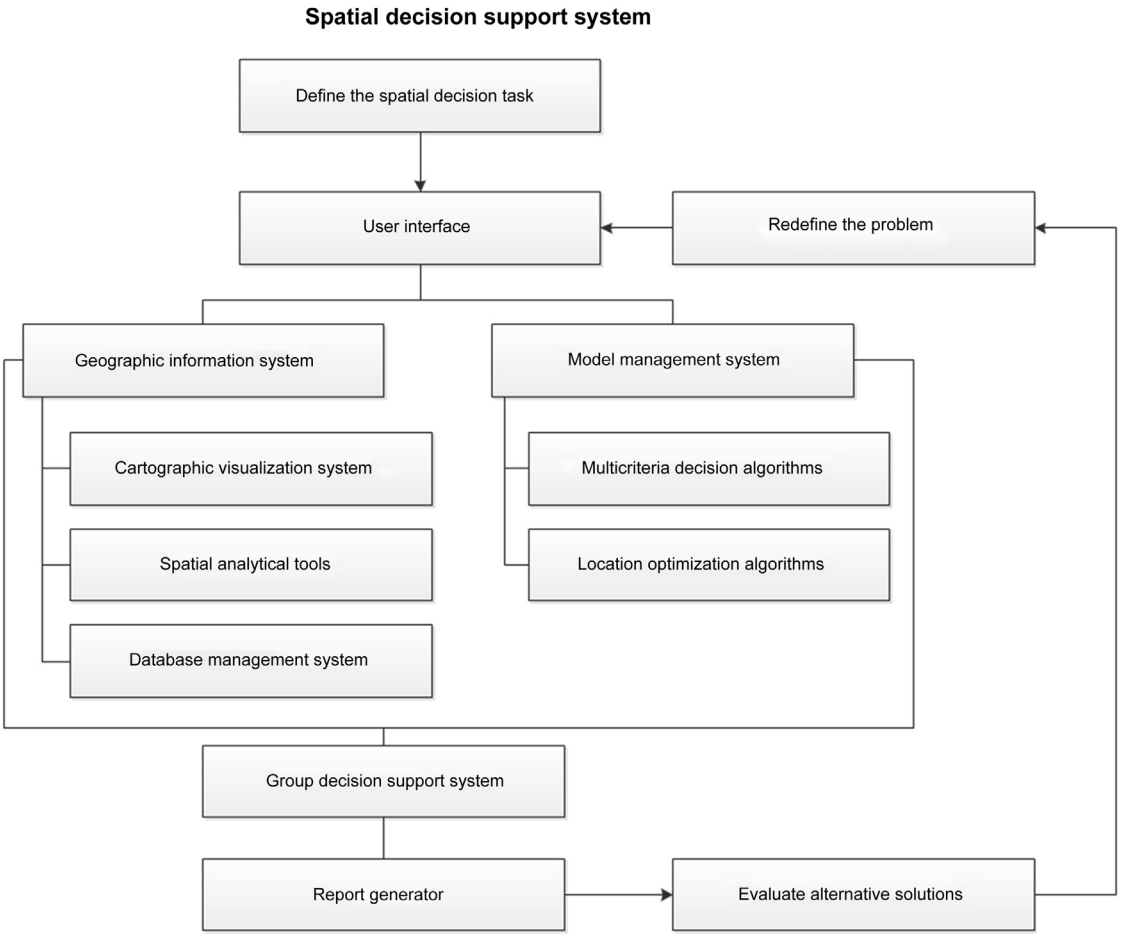

Figure 1 Flows among components of a spatial decision support System.

the fast growth of technologies for collecting spatial data, computing larger data sets, and using the Internet has aided the diffusion of GIS technology to a greater number of organizations. GIS is powerful because it has the ability to handle spatially explicit data as well as nonspatial data. Both types of data are managed using Relational Database Management Systems. GIS can be used for producing maps and integrating multisource data sets, such as data concerning population, topography, hydrology, climate, vegetation, transportation, etc.

In order to represent the real world in a digital environment, GIS uses two types of data: raster and vector. Raster data consist of pixels arranged on a rectangular grid, where each pixel in the grid has an individual value for each characteristic being represented. One type of spatial data manipulation for a GIS-based SDSS on raster data is map algebra. Map algebra involves the operations of addition, subtraction, multiplication, division, and other mathematical functions acting on individual pixels. Vector data consist of points, lines, and polygons. For vector data, the manipulation normally involves transforming and combining the basic geographic features. Spatial queries can also be conducted by defining the spatial characteristics such as within, outside, and intersect. Spatial queries extend traditional SQL functionalities to include geometric operations. For both raster and vector data, distance related operations such as buffer and proximity are very useful and widely used in some SDSS applications to create derived data layers with attributes in desired ranges. Longley et al. (2011) offer further elaboration and discussion of GIS.

The term SDSS is often reserved for systems that combine the functions of GIS with multicriteria decision analysis (MCDA) and optimizing models. These are developed for contexts that involve manipulating of information, generating of preferences, modeling of choices, and searching for locations that meet users' needs.

Two popular GIS-based SDSS softwares are Idrisi and ArcGIS. Idrisi is a commercial GIS software based mostly on raster data. Idrisi contains over 300 modules for analysis and display of spatial information. Idrisi has decision support modules including weighted linear combination (WLC), analytical hierarchy process (AHP), ordered weighted averaging (OWA), and multiobjective land allocation (MOLA), among others. Idrisi also has a wizard to assist in selection of appropriate decision techniques and a graphic modeling environment called Macro Modeler to design sequences of operations (Eastman, 2012). ESRI's ArcGIS suite of products provides the building blocks needed to implement WLC, including weighted overlay and map algebra, as well as a graphical modeling environment called Model Builder in which users can build customized SDSS.

Multicriteria Decision Analysis

Spatial decision problems typically involve a large set of possible alternatives, especially when there are multiple evaluation criteria. The alternatives are often evaluated by several people including decision makers, managers, stakeholders, and interest groups. MCDA analyzes potential actions based on multiple criteria, using decision rules to aggregate those criteria in order to rate or rank the alternatives (Eastman, 2012). Malczewski (2006) conducted a survey of literature on GIS-based MCDA. He remarks that criterion is a generic term that includes attributes and objectives, while MCDA is a blanket term that includes multiattribute and multiobjective decision making. During multiattribute decision making, users of SDSS apply various sets of weights and various methods to aggregate criteria in order to participate directly in the choice evaluation process. Widely used aggregation and weighting methods include Boolean overlay, WLC, AHP, OWA, which are described below:

1. Boolean Overlay is a method to create a new map by overlaying existing maps while using the Boolean logic functions, i.e., AND, OR, NOT. This process is easily understandable and useful for initial analysis. For example, suppose a local authority is trying to start new residential development, and the land they are looking for must be less than 30 miles from a highway, on flat slopes, and in residential zoning. Maps of parcels that satisfy each criterion could be created and then Boolean overlay could find parcels that meet all criteria.
2. WLC is a method of map overlay that aims to capture some of the value judgments and expert opinions within the spatial analysis procedure (Malczewski, 1999). Also known as simple additive weighting, WLC multiplies standardized criterion scores by relative criterion weights for each alternative. The steps are (1) define each criterion in terms of a map layer, (2) standardize each criterion map layer, (3) define a weight for each criterion, (4) construct the weighted standardized map layers, (5) generate an overall score by adding the weighted standardized map layers, and (6) rank the alternatives according to the overall performance score.
3. AHP is a priority-ranking technique that helps to break down complex problems into elements. AHP arranges the elements into a hierarchy using subjective judgments in order to assign numerical values based on the relative importance of these elements to the overall goals (Saaty, 1980). AHP generates weights for each criterion, which can then be used in a WLC process.
4. OWA extends WLC by using criteria-order weights to control the levels of criteria trade-off, allowing decision makers to place themselves along a continuous spectrum of risk tolerance (Yager, 1988). Based on fuzzy methods, OWA provides a consistent theoretical link between Boolean overlay and WLC, and thus opens the possibilities for aggregation of criteria (Jiang and Eastman, 2000).

Multiobjective decision making includes two types of methods: mathematical programing and heuristic methods. Objective and constraint are two terms that are commonly used in multiobjective decision making. An objective is a goal that guides the structuring of decision rules. A constraint serves to limit the alternatives under consideration. Mathematical programing methods formulate the objectives and constraints into a set of equations and inequalities concerning decision variables, and then solve them simultaneously. Mathematical programing methods consider all possibilities in the solution space and thus ensure an optimal solution. However, solving mathematical programing models can sometimes be computationally intensive and thus very time consuming. As an alternative to mathematical programing, heuristic methods

explore the solution space by identifying feasible solutions from which an acceptable one can be found (Tong and Murray, 2012). Although heuristic methods lack the capacity to ensure global optimality, heuristic methods are designed to generate acceptable solutions in a reasonable time frame. There exist some generic heuristic methods such as simulated annealing and evolution algorithms.

Simulated annealing randomly searches for solutions within the solution space. Initially, simulated annealing accepts worse solutions with a high probability value in order to avoid local optima. This probability is then gradually decreased to enhance the selectivity of the method toward improved global solutions, which is a process analogous to the cooling procedure for the control of crystallization during annealing in metallurgy (Kirkpatrick et al., 1983).

Evolution Algorithms follow the logic of how nature improves the survivability of a species through natural selection, reproduction over generations, and random genetic changes called mutation and crossover (Michalewicz and Fogel, 2004). A genetic algorithm (GA) is one type of evolution algorithm. In a GA, individual solutions to a problem are represented in an appropriate discrete structure, e.g., a sequence of bits that can be set according to the presence or absence of a specific characteristic. The possible solutions are initialized and each solution is assigned a fitness value, which is usually derived from an objective function. A GA is then executed iteratively where each iteration creates a new generation of solutions. In each generation, solutions with the highest fitness values have greater probabilities of being selected and reproduced to form a new generation. Evolutionary operations, e.g., mutation and crossover, are applied in each iteration so that the overall fitness of the population can be improved as the generations progress (Stewart et al., 2004).

MOLA is a heuristic method to solve land allocation problems (Eastman, 2012). MOLA determines a compromise solution that attempts to maximize the suitability of the overall solution given: a suitability map for each land use, a relative weight for each land use, and the amount of area to be assigned to each land use. MOLA usually allocates each pixel to the land use with the closest ideal point, but alternatively MOLA can allocate the pixel to an alternative land use that has a higher weight than the land use with the closest ideal point.

Greene et al. (2011) provide a summary of MCDA methods and a decision tree to aid the selection of appropriate methods. As an example, Ratick and Osleeb (2011) evaluated multiple MCDA methods to construct composite indices to measure the vulnerability of populations susceptible to lead contamination in the Dominican Republic.

Location Optimizing Algorithms

Many decision makers want to select locations to optimize some defined outcome. A large literature describes models for finding optimal locations, routes, etc.

Ratick et al. (2008) discussed an application of location models for locating storage facilities to supplement existing facilities in order to respond to the potential effects of natural and anthropogenic hazards. They used location modeling to determine the number of facilities and where to locate the facilities. The concept of backup facilities was considered to explore the interrelationships among hazards, vulnerability, and location.

Murray and Tong (2009) demonstrated an example where GIS-based SDSS are used in siting fire stations through human–computer interactions. Facilities were placed and relocated by users, while GIS was used for evaluating placement configurations.

Williams (2002) and Shirabe (2005) developed algorithms to search for the location of one site that is either most suitable for a certain land use or has the least cost to acquire. Both of their articles aimed to achieve contiguity of the site, meaning all the pixels contained by the site are connected. They both formed mathematical programing models but used different constraint sets to force the solution to be contiguous.

Williams et al. (2005) give a synthetic review of models and the spatial attributes considered for the optimal selection of nature reserve sites to protect a diversity of species or other conservation features. Spatial attributes such as the reserve shape and connectivity affect the persistence of species, the general ecological functioning of the landscape, and the ability to manage the reserve.

The traveling salesman problem is a route selection problem that asks the question: given a list of cities and the distances between each pair of cities, what is the shortest route that visits each city once and returns to the original city? It has applications in planning, logistics, as well as the manufacture of microchips. Both mathematical programing and heuristic algorithms have been developed to solve this problem efficiently (Applegate et al., 2011).

Network location problems occur when new facilities must be allocated on a network, e.g., a transportation network. These problems exist in the private sector, such as when a retail company would like to locate storage facilities. These problems exist also in the public sector, such as when a municipality would like to locate fire stations. Researchers have categorized some classic location problems in families such as the set covering problem, the *p*-center problem, and the *p*-median problem (Daskin, 2011). One example of a set covering problem is to minimize the number of fire stations so that all demand nodes are within a given response time, which is the time for a fire truck to travel to a demand site from the nearest fire station (Toregas et al., 1971). The *p*-center problem is to allocate a fixed number of fire stations on a network in order to minimize the maximum response time to any demand node (Tansel et al., 1983). If we use a storage facility as the example, then the *p*-median problem is to allocate a fixed number of facilities to serve demand nodes so that the transportation cost is minimized (Hakimi, 1964). The transportation cost considers the amount of product needed at each demand node and the distance between the demand node and the facility that serves the demand node. Those problems are well studied and there exists both mathematical programing and heuristic algorithms for the solutions.

Group Decision Support Systems

Spatial decisions are often made by collaborative efforts involving groups of people. Nyerges and Jankowski (1997)

discuss adaptive structuration theory in a GIS environment to assist groups who are involved in public land planning controversies, such as hazardous waste cleanup and habitat redevelopment.

Some SDSS incorporate methods for reconciling various views. Couclelis and Monmonier (1995) discuss this application and offer a computer-supported system to help promote the resolution of planning controversies. Their system uses GIS capabilities to present spatial details of the planning situation and uses a graphic interface to encourage reconciliation of opposing views. Armstrong (1993) describes how cooperative work computing environments can enhance SDSS by enabling group members to observe, manipulate, and evaluate alternative plans collectively, then to provide mechanisms for conflict resolution.

Research in SDSS seeks to capture and incorporate knowledge of experts (Armstrong et al., 1990; Leung et al., 1997). Some of the knowledge concerns the algorithms and their links to the decision task. Other knowledge is substantive knowledge about the processes that must be modeled.

Implementations of SDSS

The academic literature has many examples of prototype implementations of SDSS. Commercial SDSS applications include generic SDSS and domain-specific SDSS. Sugumaran and Degroote (2010) listed existing tools and softwares that fall into these two categories.

Generic SDSS allow users to address a wide range of problems. SDSS normally provide a rich set of decision support components such as spatial database management, analytical models, and a framework to allow rapid assembly for a specific application. Examples include systems such as Idrisi's Macro Modeller, ArcGIS's Model Builder, and Erdas' Imagine. These systems provide a graphical user interface that allows construction of models without any software development. Decision flows are represented by a graph, where modules and data are stored as nodes of the graph, while the connections between the nodes are stored as vertices. Graphical modeling tools give a clear layout of the decision process. Some tools provide additional control over the flow of the execution of components in the model, while other tools allow for the creation of iterative processes, so that batch processing of a large data set is possible.

Domain-specific SDSS are developed to support decision making for specific problems that are complex, yet occur with sufficient frequency to support the overhead cost of developing and testing. For example, Marxan is a freely available system developed for conservation planning. It is designed to solve reserve design problems by achieving a desired representation of biodiversity features for the least cost (Game and Grantham, 2008). Marxan uses simulated annealing to find near-optimal solutions. INDEX is a suite of planning tools that can be used for assessing community conditions, designing future scenarios, measuring a scenario's success, etc. INDEX also allows comparison of plan goals versus development proposals. CommunityViz is another widely used planning SDSS tool similar to INDEX. It is an extension to ArcGIS software and includes a large suite of visualization, analytical, and communication tools. Visualization capacities are emphasized in this tool that allows users to experiment with various scenarios and view results in a variety of formats including charts, tables, reports, maps, and three-dimensional models.

Looking Forward

SDSS represent the integration of ideas, methods, and techniques from many social, behavioral, and management sciences. The usefulness of SDSS in the policy and planning sciences is based on the degree to which SDSS successfully incorporate the findings and approaches of these sciences. SDSS have an important link with the fast growing field of geographic information science. The rapid development of national spatial data infrastructures provides a rich and expanding database that SDSS can use.

Challenges still remain for SDSS to address complex multidisciplinary issues that have spatial dimensions. Technological challenges concern the selection of software platforms, the development of user-friendly interfaces, the integration of spatial models, etc. Social, organizational, educational, and policy challenges must also be overcome (Sugumaran and Degroote, 2010). For example, the interaction with users during every stage of the development process is crucial for the success of SDSS, thus SDSS require support from all management levels in order to be incorporated into important decisions. Many individuals from disciplines that might benefit from the use of SDSS are unaware of the potential benefits. Spatial decision-making processes and tools, such as SDSS, must be included in university curricula. SDSS have become a science that builds upon the successful development of models concerning behavioral, social, and political processes. Theories and models of these processes must contribute to SDSS as the applications of SDSS expand. SDSS have become an important element in social science theories and models to explain decisions. The evolution from the use of social and behavioral models as exogenous units into SDSS to the models' incorporation as endogenous units into SDSS is an important development in the social sciences. This development marks the more explicit incorporation of geographic variables in social science theory and method.

Note: This is a substantially updated version of the previous article by Gerard Rushton published in the first edition of the IESBS (2001).

See also: Access: Geographical; Cartographic Visualization; Citizen Mapping; Environmental Planning; Geographic Information Systems; Location Theory; Mathematical Models in Geography; Organizational Decision Making; Planning Issues and Sustainable Development; Region: Geographical; Regional Planning: China; Regional Planning; Spatial Analysis in Geography; Spatial Data Infrastructures; Spatial Data; Spatial Optimization Models; Spatial–Temporal Modeling; Time-Space in Geography; Transportation, Supply, and Congestion.

Bibliography

Applegate, D.L., Bixby, R.E., Chvatal, V., Cook, W.J., 2011. The Traveling Salesman Problem: A Computational Study. Princeton University Press, Princeton, NJ.

Armstrong, M.P., 1993. Perspectives on the development of group decision support systems for locational problem-solving. Geographical Systems 1, 69–81.

Armstrong, M.P., De, S., Densham, P.J., Lolonis, P., Rushton, G., Tewari, V.K., 1990. A knowledge-based approach for supporting locational decisionmaking. Environment and Planning B: Planning and Design 17, 341–364.

Bozkaya, B., Erkut, E., Laporte, G., 2003. A tabu search heuristic and adaptive memory procedure for political districting. European Journal of Operational Research 144, 12–26.

Church, R.L., Murray, A.T., 2009. Business Site Selection, Location Analysis and GIS. Wiley, New York.

Couclelis, H., Monmonier, M., 1995. Using SUSS to resolve NIMBY: how spatial understanding support systems can help with the 'not in my backyard' syndrome. Geographical Systems 2, 83–101.

Daskin, M.S., 2011. Network and Discrete Location: Models, Algorithms, and Applications. John Wiley & Sons, New York.

Eastman, J.R., 2012. IDRISI Selva Manual. Clark Labs, Worcester, MA.

Game, E.T., Grantham, H.S., 2008. Marxan User Manual for Marxan Version 1.8.10. University of Queensland, St. Lucia, Queensland, Australia, and Pacific Marine Analysis and Research Association, Vancouver, British Columbia, Canada.

Greene, R., Devillers, R., Luther, J.E., Eddy, B.G., 2011. GIS-based multiple-criteria decision analysis. Geography Compass 5, 412–432.

Hakimi, S.L., 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research 12, 450–459.

Jiang, H., Eastman, J.R., 2000. Application of fuzzy measures in multi-criteria evaluation in GIS. International Journal of Geographical Information Science 14, 173–184.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Leung, Y., Leung, K.S., Shao, Z., Lau, C.K., 1997. A development shell for intelligent spatial decision support systems: 2. An application in flood simulation and damage assessment. Geographical Systems 4, 39–57.

Longley, P., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2011. Geographic Information Systems and Science. John Wiley & Sons, New York.

Malczewski, J., 1999. GIS and Multicriteria Decision Analysis. John Wiley & Sons, New York.

Malczewski, J., 2006. GIS-based multicriteria decision analysis: a survey of the literature. International Journal of Geographical Information Science 20, 703–726.

Michalewicz, Z., Fogel, D.B., 2004. How to Solve It: Modern Heuristics. Springer, New York.

Murray, A., 2010. Advances in location modeling: GIS linkages and contributions. Journal of Geographical Systems 12, 335–354.

Murray, A.T., Tong, D., 2009. GIS and spatial analysis in the media. Applied Geography 29, 250–259.

Nyerges, T.L., Jankowski, P., 1997. Enhanced adaptive structuration theory: a theory of GIS-supported collaborative decision making. Geographical Systems 4, 225–259.

Ratick, S., Meacham, B., Aoyama, Y., 2008. Locating backup facilities to enhance supply chain disaster resilience. Growth and Change 39, 642–666.

Ratick, S.J., Osleeb, J.P., 2011. Measuring the vulnerability of populations susceptible to lead contamination in the Dominican Republic: evaluating composite index construction methods. GeoJournal, 1–14.

Ricca, F., Simeone, B., 2008. Local search algorithms for political districting. European Journal of Operational Research 189, 1409–1426.

Saaty, T.L., 1980. The Analytical Hierarchy Process. McGraw-Hill, New York.

Shirabe, T., 2005. A model of contiguity for spatial unit allocation. Geographical Analysis 37, 2–16.

Stewart, T.J., Janssen, R., Herwijnen, M. v, 2004. A genetic algorithm approach to multiobjective land use planning. Computer & Operational Research 31, 2293–2313.

Sugumaran, R., Degroote, J., 2010. Spatial Decision Support Systems: Principles and Practices. CRC Press, Boca Raton, FL.

Tansel, B.C., Francis, R.L., Lowe, T.J., 1983. State of the art—location on networks: a survey. Part I: the p-center and p-median problems. Management Science 29, 482–497.

Tong, D., Murray, A.T., 2012. Spatial optimization in geography. Annals of the Association of American Geographers 102, 1290–1309.

Toregas, C., Swain, R., ReVelle, C., Bergman, L., 1971. The location of emergency service facilities. Operations Research 19, 1363–1373.

Williams, J.C., 2002. A zero-one programming model for contiguous land acquisition. Geographical Analysis 34, 330–349.

Williams, J.C., ReVelle, C.S., Levin, S.A., 2005. Spatial attributes and reserve design models: a review. Environmental Modeling and Assessment 10, 163–181.

Yager, R., 1988. On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man, and Cybernetics 8, 183–190.

Relevant Websites

http://www.tandfonline.com/doi/abs/10.1080/13658816.2013.763947#.UrNmM7QmSdw – Book on Spatial Decision Support Systems.

www.clarklabs.org – Clark Labs, Idrisi GIS Software.

http://placeways.com/communityviz/ – CommunityViz Software for Planning.

www.esri.com – ESRI, ArcGIS Software.

www.impetus.uni-koeln.de/en/project/concepts/sdss-basics.html – IMPETUS SDSS Project.

http://www.uq.edu.au/marxan/ – Marxan Software for Conservation Planning.

www.spatial.redlands.edu/redlandsinstitute/giscience/sds/ – Redlands Institute on SDSS.

Spatial Dynamics and Crime

Martin A Andresen, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Crime and its spatial dynamics have a long history dating back to the early 1800s in France. In this article, I review four subfields of research within spatial criminology that explicitly consider crime and its spatial dynamics: spatially referenced crime rates, location quotients, geographic profiling, and crime and place. The latest research from these subfields are reviewed with some discussion of their future directions in understanding spatial dynamics and crime.

Introduction

Crime and its spatial dynamics have been studied for close to 200 years since the work of André-Michel Guerry (1833) and Adolphe Quetelet (1842), French and Belgian, respectively. Their work was not only important for spatial criminology, but positivist criminology, more generally. Since this time, a large body of research has emerged that investigates the spatial dynamics of crime at a variety of scales and contexts.

There are many dimensions of spatial criminology that are of particular interest at this time, including: understanding the role of police foot patrol and car patrol in reducing crime, understanding the crime drop of the 1990s, furthering the understanding of relationships, such as unemployment and crime, and helping to understand the transition economies in central and eastern Europe. However, in this article, I will discuss four areas of research in spatial criminology that are explicitly concerned with the spatial dynamics of crime: spatially referenced crime rates, the location quotient, geographic profiling, and crime and place. Though these may not be the only aspects of spatial criminology that are explicitly concerned with the spatial dynamics of crime, they have all received a lot of interest within the field and continue to offer fruitful avenues of research in the future.

Spatially Referenced Crime Rates

In order to calculate a crime rate, there are two pieces of data that are necessary: crime data and population at risk data, shown in eqn [1]. The crime count is literally the count of criminal events in some defined geographic area. This count is most often based on official crime data, emerging from some aspect of the criminal justice system. The most well-known official crime data are the Uniform Crime Reporting (UCR) data. These data had its beginning in the United States in 1930 and now nearly all law enforcement agencies in the United States provide data to the Federal Bureau of Investigation for the UCR counts (Mosher et al., 2011; FBI, 2012). In Canada, the UCR began later, in 1962, and is gathered by the Canadian Centre for Justice Statistics – responding to the UCR survey in Canada is mandatory. In 1988, a new version of the UCR in Canada was started that added information regarding incidents, victims, and accused persons. This new UCR is referred to as UCR2; equivalent data in the United States are collected under the National Incident-Based Reporting System, that began in 1987 (FBI, 2012; Statistics Canada, 2012). Another form of official crime data is calls for service data from the police. Though often considered unofficial because these data are not dependent upon a criminal charge and often represent police activity, these are very useful data because a geographic location and time is most often included and the calls for service can be separated to only include the calls that the police identify as a criminal event (Sherman et al., 1989).

$$\left(\frac{\text{Crime count}}{\text{Population at risk}}\right) \times \text{Scalar} \qquad [1]$$

Despite the widespread availability of these official crime data, there are a number of limitations. First, crime reporting can vary from police detachment to police detachment. This may be due to factors such as the local policing and/or population culture considering factors such as (a lack of) tolerance on particular issues. Second, not all criminal events are reported to the police. In Canada, only 31% of criminal victimization was reported to the police in 2009, down from 34% in 2004 and 37% in 1999 (Perreault and Brennan, 2010). This is of critical importance in spatial criminology because if the spatial patterns of crime for official data are different from the spatial patterns of crime more generally, all inference from official data may be spurious. However, very little can be done regarding these first two limitations.

A third limitation is in regard to the need for population at risk data to calculate the crime rates. In spatial criminology, that often uses the census tract as its spatial unit of analysis, the most common source of population at risk data is the census – resident population. These data are very easy to obtain and high quality, but there are two issues with using these data. First, censuses are only undertaken every 5 (Canada) or 10 years (United States and United Kingdom) in most countries. Consequently, depending on the date of the crime data obtained there may be as many as 9 years discrepancy between crime count and population at risk data. If populations are stable over long periods of time this is not problematic, per se, but in cities that grow this could lead to significant overestimates of crime rates if crime counts are for a later year than the resident population.

The second issue with using the resident population from a census as the population is risk is whether or not it is an accurate representation of where people are. This is a critical question in spatial criminology because if the crime rate is to represent an overall measure of risk (Boggs, 1965) the variable representing the population at risk should measure where

 http://dx.doi.org/10.1016/B978-0-08-097086-8.45087-7

people actually are. Recent research on this issue has shown that the ambient population can be quite different from the resident population leading to changes in inference regarding spatially referenced crime rates (Andresen, 2006, 2013; Andresen and Jenion, 2010). In this work, data from the Oak Ridge National Laboratory was used to measure the ambient population that was defined as the number of people in a square kilometer averaged over a 24-h period for any typical day of the year – see Andresen (2006) for details of these data.

In an investigation of the similarity between maps of resident-based and ambient-based crime rates, Andresen and Jenion (2010) found that, despite the visual appearance of similar spatial patterns and a statistically significant relationship, resident-based crime rates were very poor predictors of ambient-based crime rates. Though we do not discuss this prediction, this is what we assume when we use the resident population: it may not be the best measure, but it is representative of the true population at risk. Moreover, places that attract populations throughout the day, measured using the ambient–resident population ratio, are commercial/shopping areas and major transportation routes.

In subsequent research, Andresen (2011, 2013) showed that ambient-based crime rates exhibited significant differences to resident-based crime rates in the context of local analysis, local Moran's *I*. In this work, not only did the spatial pattern of clustering change with some units of analysis changing statistical significance, but also when statistically significant in both analyses the classifications in some places would be completely different – most often, high crime areas remained high crime areas and low crime areas remained low crime areas, but the units of analysis contained within each of those classifications changed. In a spatial regression context, the similarity of results for resident- and ambient-based crime rates differed on the year of data and the size of the unit of analysis. In some cases, the results from the ambient-based crime rates outperformed the resident-based crime rate results in the context of: greater (pseudo) R^2 values, more reasonable parameter estimates, more statistically significant relationships, and a better concordance with theoretical expectations. In other cases, there were very few differences with limited consequences to the results.

Overall, this research shows that there are good theoretical reasons to question the use of the commonly used resident population in spatially referenced crime rates – crime rate calculations for what may be considered a 'closed system' of population movement such as a metropolitan area or a subnational unit (province or state) will most probably be unaffected. Though there may not necessarily be an impact on the results, such an impact is definitely possible and may alter any interpretations. Moreover, just because there is no impact for 1 year of data and 1 spatial unit of analysis, an impact may be present for a different year of data and/or a different spatial unit of analysis even in the same city.

A Spatial Alternative Measure of Crime: Location Quotient

As should be clear from the preceding section, the calculation of crime rates has two sources of measurement issues: the criminal event counts and the population at risk. There is an alternative to the crime rate that only relies upon criminal event data, however, that has been used in spatial criminology for 20 years but is arguably underutilized – the location quotient. The location quotient is a specifically geographic statistic that measures the over- or underrepresentation of some activity in a spatial unit of analysis relative to the entire study area. The location quotient is calculated as follows:

$$\mathrm{LQ}_{\mathrm{in}} = \frac{C_{\mathrm{in}}/C_{\mathrm{tn}}}{\sum_{n=1}^{N} C_{\mathrm{in}} \Big/ \sum_{n=1}^{N} C_{\mathrm{tn}}} \qquad [2]$$

where C_{in} is the count of crime i in subregion n, C_{tn} is the count of all crimes in subregion n, and N is the total number of subregions. The location quotient is a ratio of the percentage of a crime type in a subregion relative to the percentage of that crime type in the region as a whole; neighborhoods within a city, for example. If the location quotient is equal to one, the subregion has a proportional share of a particular crime; if the location quotient is greater than one, the subregion has a disproportionately larger share of a particular crime; and if the location quotient is less than one, the subregion has a disproportionately smaller share of a particular crime. For example, if a subregion has a location quotient of 1.20, that subregion has 20% more of that crime than expected given the percentage of that crime in the region as a whole. Consequently, it can be said that this subregion 'specializes' in that particular crime type.

First introduced to the criminological literature by Brantingham and Brantingham (1993) as a descriptive tool, the location quotient has been used by subsequent researchers to investigate a number of phenomena, such as drug markets, theory testing, shootings, and automotive theft markets – see Andresen (2013) for a review. The utility of the location quotient is that by using an alternative measure of criminal activity we are able to identify different patterns. For example, as outlined in Andresen (2013), the spatial patterns of crime rates versus location quotients are somewhat similar for a number of crime types, but notable differences emerge when considering specialization versus risk. In the case of assault in Vancouver, Canada, a crime rate map indicates that this crime type is highly concentrated in Skid Row. However, the east side of the city (and the north east side in particular) exhibits a lot of assault specialization. What does this mean? In order to answer this question, both the crime rate and the location quotient need to be considered. First, the risk of an assault is greatest in Vancouver's Skid Row. But, given that an individual is going to be a victim of crime, it is more likely to be an assault on the east side of the city relative to the west side of the city. Similar differences in the spatial patterns of crime emerge for other crime types as well, but not always in the same places: robbery, sexual assault, theft (more specialization of this crime type is present on the western more affluent side of the city), theft from vehicle, and theft of vehicle.

Perhaps the most interesting application of the location quotient in Andresen (2013) is burglary. The burglary crime rate map indicates that burglary has its greatest risk in and around the central business district and skid row – standard

social disorganization theory predictions. However, the location quotient map shows that the burglary specializes the most on the relatively affluent west side of the city where the burglary rates are the lowest. Of course, this does not mean that burglary is more frequent on the west side of the city than the crime rate map indicates, but it does mean that if an individual is going to be a victim of a crime on the west side of Vancouver it is most likely going to be a burglary. Though no research has been undertaken to support this claim, it has been hypothesized that this may explain some aspects of the fear of crime in low crime neighborhoods: though the burglary rate may be low, everyone whom an individual speaks to about criminal victimization has referred to a burglary of their home so other residents in the neighborhood begin to fear that crime.

Another interesting application of the location quotient was an attempt to understand the increasing crime rate from east to west in Canada that has been present, but not well understood, for at least 50 years (Andresen, 2009). For decades, the western provinces have had greater than national average crime rates across most crime types. However, no such pattern emerged for the location quotient. In the cases of violent crime, there were location quotients with moderate overrepresentation, but for provinces in western, central, and eastern Canada; in the case of property crime, no province had notable overrepresentation. In fact, the territories that are historically characterized by the greatest violent and property crime rates had the lowest location quotients for property crime – all three territories were overrepresented for violent crime, however. What this means is that the risk of criminal victimization is greatest in the west, but given that an individual is going to be a victim of crime, there is no spatial pattern that emerges to state that individual will be the victim of a particular crime type. In other words, the western provinces are not disproportionately violent than other areas in Canada; they simply have more of everything.

Lastly, Andresen (2013) also presented an inferential application of the location quotient (spatial regression) that produced some interesting results. Most important, the workhorse theories used in spatial criminology (social disorganization theory and routine activity theory) performed well in the context of crime specialization. However, because the dependent variable measured specialization not risk, the interpretations of the estimated parameters are different. In some circumstances these differences are subtle such that the estimated parameters have the same sign, but how one interprets the results changes; in other contexts, the sign is opposite from the expectation in a crime rate context (a negative relationship between burglary and the unemployment rate, for example) that requires a completely different, but theoretically justified, explanation.

In all of the applications of the location quotient, the researcher is forced to ask different questions with regard to the spatial pattern of crime, sometimes subtle and sometimes not. Asking these different questions, it is argued, allows the researcher to obtain a better understanding of the phenomena of crime. As such, the use of alternative measures of crime, particularly in a spatial context, should not be used to replace crime rates, but to supplement crime rates for a deeper understanding of the spatial dynamics of crime.

Geographic Profiling

One of the more well-known applications of the theories within spatial/environmental criminology is geographic profiling. Developed by D. Kim Rossmo, geographic profiling may be thought of as understanding the spatial dynamics of serial crime. Geographic profiling is an investigative methodology that uses the locations of a connected series of criminal events to determine the most probable area of the offender's residence – technically geographic profiling is used to identify the anchor point from which a serial offender bases his or her movements, but this is most often the residence. Because geographic profiling uses a connected series of criminal events, the most common uses of geographic profiling are serial murder, rape, arson, and robbery. However, geographic profiling may also be used for nonserial crimes that involve multiple locations. The major function of geographic profiling is to prioritize a list of potential offenders to aid in a corresponding investigation. Geographic profiling has proven to and continues to be a useful tool for the practice of investigative policing as well as the understanding of serial offenders from an academic perspective. Some of the more interesting recent geographic profiling applications have included: cellular telephone switch tower sites in kidnapping cases, the stores in which bomb components have been purchased, and a historical analysis of the locations of anti-Nazi propaganda postcards in left in the streets of Berlin, Germany, in the early 1940s (Rossmo, 2014). Additionally, there have been new applications of geographic profiling that have proven to be very instructive, but for different reasons: terrorism and biology/zoology and epidemiology (Rossmo, 2012).

As stated above, standard geographic profiling uses the related locations in a series of criminal events to help identify the offender's residence. At this time, geographic profiling has been applied to help military analysts identify the location of enemy military bases (Brown et al., 2005), but it can also be used to identify the targets of terrorist attacks. The study of terrorist activities has identified the existence of 'terrorist cell sites.' These terrorist cell sites may be meeting places, rented apartments, etc. that are used as part of the planning of a terrorist attack. As it turns out, there is a strong spatial component to these terrorist cell sites. The research undertaken by Rossmo and Harries (2011) goes through how geographic profiling may be used to identify terrorist targets using these terrorist cell sites. Effectively, this use of geographic profiling is undertaken by taking the method of geographic profiling and turning it on its head. Rather than attempting to help identify the offender, geographic profiling is used to help identify the victim, whether that victim is a place or a person. As it turns out, terrorist cell sites are set up along the same principles as other offending activities. A nonterrorist offender tends to search for criminal targets in known areas, close to home but not so close as to reveal his or her identify. A set of terrorist cell sites tend to be organized close to but not too close to the target of the attack. Consequently, if the terrorist cell sites are considered the 'connected series of criminal events' the geographic profile can then be interpreted as helping to identify the target of the terrorist attack rather than the home of the offender. This new application of geographic profiling has

obvious benefits to counterinsurgency and counterterrorism operations as another tool for identifying potential targets of terrorist attacks based on known terrorist cell site activities as well as the most probable locations of enemy military bases. This shows the reach and power of how the understanding of the spatial dynamics of crime may be useful in other domains of research and practice.

One of the other new applications of geographic profiling is in the realm of biology and zoology. In an analysis of two species of bats in Scotland, Le Comber et al. (2006) were able to show the utility of geographic profiling in analyzing the foraging patterns of the different species. Specifically, the authors were able to show that based on the use of geographic profiling the two different species of bats exhibited different patterns of foraging. Raine et al. (2009) were able to show that geographic profiling could be used to identify the likely locations for the nests of bumblebees, including the consideration of flower (target) densities. And Martin et al. (2009) were able to use geographic profiling in order to discover that the spatial patterns of shark predation were not random. The authors were able to identify the search base (anchor point) for these sharks in South Africa and understand the predatory strategies of these sharks considering factors, such as prey detection, capture rates, and competition from other sharks. Perhaps most interesting, Raine et al. (2009) found that smaller (younger) sharks exhibited significantly greater areas for their prey search patterns than larger (older) sharks. This suggests that as sharks age they learn, or at least refine, their hunting strategies. And lastly, geographic profiling has been applied to epidemiological research to identify the origins of infectious diseases, such as contaminated water sources for cholera and mosquito breeding grounds for malaria (Le Comber et al., 2011), and identifying the source locations for invasive species (Stevenson et al., 2012), both using the current locations of these phenomena.

Despite the general interest of the newer applications of geographic profiling, and their direct implications for the safety and security of different human populations, there is another significant implication of this more recent research. Though there are many serial criminal investigations that allow for more data to be available to refine the practice of geographic profiling, waiting for another set of serial homicides, for example, is not the most pleasant way of thinking of future data. If, and it does appear to be the case, the predatory behaviors of other animal species are similar to ours, this new research can be used to better understand the predatory behavior of humans. For example, if humans, similar to sharks off the South African coast, also reduce the size of their search area as they age, this information may be useful for the ranking of potential offenders in a criminal investigation; in such a situation, it may be possible to solve a serial criminal investigation sooner because of a better understanding of spatial predatory behavior, more generally.

Crime and Place

As mentioned above, spatial criminology as we know it today began almost 200 years ago with the work in France by Quetelet and Guerry. One of the trajectories of spatial criminology over these many years has been the identification of spatial heterogeneity within the common spatial units of analysis of the day. For many years, the standard spatial unit of analysis within this literature has been the neighborhood (however defined) or census tract. One of the great advantages of these units of analysis is the availability of census data for analysis such as with the crime rate and location quotient studies discussed above. Twenty-five years ago, the use of yet another smaller spatial unit of analysis began to emerge, called the microplace – this literature is often referred to as the crime and place literature. These microplaces could be addresses, street intersections, or street segments and became used for citywide analyses because it became apparent that neighborhoods were far from being spatially homogeneous with regard to criminal activity.

From a theoretical perspective in spatial criminology, the microplace makes a lot of sense. Within routine activity theory (Cohen and Felson, 1979), a criminal event occurs when a motivated offender and a suitable target converge in time and space. That convergence, of course, does not occur in a neighborhood, but at a discrete location. In the geometric theory of crime (Brantingham and Brantingham, 1981), criminal events dominantly occur at our activity nodes and the pathways between them, but there are also particular places within these activity nodes where the majority of criminal events will occur. And lastly, rational choice theory (Clarke and Cornish, 1985) posits that criminal events are the result of context-specific choices. That context can vary significantly within a 'neighborhood' such that the context of microplaces will be important for criminal decision making.

The first known research study to systematically investigate the microplace in a citywide analysis was Sherman et al.'s (1989) analysis of predatory crime in Minneapolis. In this study, Sherman et al. (1989) found that 50% of calls for police service were generated from 3% of street segments. This profound concentration of criminal events showed that even in the neighborhoods with the greatest criminal event levels had areas within them that were crime free. Smith et al. (2000) investigated the integration of routine activity theory and social disorganization theory considering the microplace rather than the more tradition census tract or other census-defined unit of analysis. They found that the integration of these two theories was far more successful using the microplace. Weisburd et al. (2004) and Groff et al. (2009) analyzed the microplace in Seattle using trajectory analysis. Similar to Sherman et al. (1989) these authors found that 50% of all criminal events were accounted for by approximately 5% of street segments, over a 14-year period. Moreover, Groff et al. (2009) found that street segments with the same trajectory tended to cluster together.

Some of the more recent research in this area shows the importance of the microplace in understanding the spatial dynamics of crime. Weisburd et al. (2013), in the context of crime prevention, showed that over a 16-year time period 1% of street segments in Seattle represented 23% of all criminal activity across the city. This is an incredible concentration of criminal events in any context. In another set of analyses in Canada, Andresen and Malleson (2011) and Andresen and Linning (2012) revealed how concentrated crime can be in two Canadian cities, Vancouver and Ottawa. In Vancouver,

Table 1 Concentration (within concentrations) of crime, Vancouver and Ottawa, Canada

	Percentage of street segments accounting of 50% of crime	*Percent of street segments that have any crime*	*Percent of street segments with crime that account for 50% of crime*
	Ottawa, 2006		
Commercial burglary	0.55	2.31	23.78
Residential burglary	1.36	4.86	27.95
Total burglary (aggregate)	1.67	6.61	25.30
Commercial robbery	0.01[a]	0.30	27.84
Individual robbery	0.31	0.65	47.66
Other robbery	0.30	0.82	35.93
Total robbery (aggregate)	0.38	1.46	26.10
Theft of vehicle	0.99	4.50	22.02
Total (without double counting)	1.70	9.52	17.87
	Vancouver, 2001		
Assault	1.62	18.75	8.64
Burglary	7.61	39.43	19.31
Robbery	0.84	5.32	15.87
Sexual assault	1.12	2.99	37.32
Theft	2.58	26.79	9.64
Theft of vehicle	5.97	27.11	22.01
Theft from vehicle	2.64	18.75	8.64
Total (without double counting)	5.02	61.42	8.18

[a]Actual value: 0.000823.

2001, 50% of the criminal events under study (listed in Table 1) were accounted for by just over 5% of street segments – an almost identical result compared to the work done at Seattle. This concentration varied by crime type from approximately 1% of street segments (sexual assault and robbery) to almost 8% (burglary). In Ottawa, 2006, the concentrations were even greater: 50% of the criminal events under study were accounted for by 1.7% of street segments, ranging from 0.01% of street segments (commercial robbery) to 1.67% of street segments (burglary) – the average is greater than the range because different crime types have their concentrations at different places.

The second column of Table 1 shows the percent of street segments that have any criminal events. In Vancouver, just over 60% of street segments have any criminal events, whereas less than 10% of street segments in Ottawa have any criminal events. This is an incredible concentration within Ottawa that may be due to the limited number of crime types under analysis, but is noteworthy nonetheless. The last column of Table 1 shows the percent of street segments with any criminal events that account for 50% of criminal events – this column uses column 2 as the base for the percentage, not all street segments in each city. These percentages represent concentrations within concentrations and indicate that there are clearly hot spots of crime within hot spots of crime and that spatial heterogeneity can even be present considering the microspatial unit of analysis.

Future Directions

Crime and its spatial dynamics is a dynamic field of research. Spatial criminology spans a wide range of research activities ranging from the practical analysis of police activities to the more abstract nature of theoretical testing – of course, better refined theories may serve better practical applications. The four subfields of spatial criminology reviewed above are still in their infancy, requiring a lot of future research to better understand the spatial dynamics of crime.

In the context of spatially referenced crime rates, further applications of the Oak Ridge National Laboratory data to investigate its utility in other contexts is in order. Additionally, there are other sources of data that may be used to estimate the ambient population for more appropriate estimates of the population at risk. For example, the SENSEable City Lab at MIT http://senseable.mit.edu uses cellular phone call data to estimate where people are. And with mobile technologies, more generally, using applications such as Facebook and Twitter with location data there are many possibilities to get a better understanding of where the population at risk may be. Moreover, considering data obtained from mobile devices, measurements of the population at risk may be generated for different days of the year and different times of the day. Research on the alternative measures of crime should involve a broader application of the location quotient in order to continue supplementing the more traditional crime rate. Additionally, the crime rate and location quotient are not the only available measures that can be applied to the study of crime and its spatial dynamics. Future research should investigate other alternatives to be found in other disciplines, such as economics or geography.

Future research in geographic profiling should continue to apply its method to counterinsurgency and counterterrorism in order to increase the safety and security of the affected populations. Additionally, the application of geographic profiling to other predatory species may continue to offer insight into the predatory behavior of human (criminal) activity. And lastly, continued research in crime and place will only serve to

provide a better understanding of the spatial dynamics of crime, particularly at the microspatial scale. Future research should continue to apply the microspatial unit of analysis to other contexts, but also use finer crime type classifications. Much of this crime and place research uses aggregate crime types, but the research that does use more detailed crime types has revealed variations along this dimension as well.

See also: Crime, Geography of; Decision and Choice: Bounded Rationality; Decision and Choice: Heuristics; Quantitative Methods and Crime; Routine Activities, Criminal Opportunities, Crime and Crime Prevention.

Bibliography

Andresen, M.A., 2006. Crime measures and the spatial analysis of criminal activity. British Journal of Criminology 46, 258–285.

Andresen, M.A., 2009. Crime specialization across the Canadian provinces. Canadian Journal of Criminology and Criminal Justice 51, 31–53.

Andresen, M.A., 2011. The ambient population and crime analysis. Professional Geographer 63, 193–212.

Andresen, M.A., 2013. The Science of Crime Measurement: Issues for Spatially-referenced Crime Data. Routledge, New York.

Andresen, M.A., Linning, S.J., 2012. The (in)appropriateness of aggregating across crime types. Applied Geography 35, 275–282.

Andresen, M.A., Malleson, N., 2011. Testing the stability of crime patterns: implications for theory and policy. Journal of Research in Crime and Delinquency 48, 58–82.

Andresen, M.A., Jenion, G.W., 2010. Ambient populations and the calculation of crime rates and risk. Security Journal 23, 114–133.

Boggs, S.L., 1965. Urban crime patterns. American Sociological Review 30, 899–908.

Brantingham, P.L., Brantingham, P.J., 1981. Notes of the geometry of crime. In: Brantingham, P.J., Brantingham, P.L. (Eds.), Environmental Criminology. Waveland Press, Prospect Heights, IL, pp. 27–54.

Brantingham, P.L., Brantingham, P.J., 1993. Location quotients and crime hot spots in the city. In: Block, C.R., Dabdoub, M. (Eds.), Workshop on Crime Analysis through Computer Mapping, Proceedings. Criminal Justice Information Authority, Chicago, IL, pp. 175–197.

Brown, R.O., Rossmo, D.K., Sisak, T., Trahern, R., Jarret, J., Hanson, J., 2005. Geographic Profiling Military Capabilities. Final Report Submitted to the Topographic Engineering Center. Department of the Army, Fort Belvoir, VA.

Clarke, R.V.G., Cornish, D.B., 1985. Modeling offenders' decisions: a framework for research and policy. Crime and Justice: An Annual Review of Research 6, 147–185.

Cohen, L.E., Felson, M., 1979. Social change and crime rate trends: a routine activity approach. American Sociological Review 44, 588–608.

Federal Bureau of Investigation, 2012. Uniform Crime Reports. http://www.fbi.gov/about-us/cjis/ucr/ucr (accessed 30.06.13.).

Groff, E., Weisburd, D., Morris, N.A., 2009. Where the action is at places: examining spatio-temporal patterns of juvenile crime at places using trajectory analysis. In: Weisburd, D., Bernasco, W., Bruinsma, G.J.N. (Eds.), Putting Crime in Its Place: Units of Analysis in Geographic Criminology. Springer, New York, pp. 61–86.

Guerry, A.M., 1833. Essai Sur la Statistique Morale de la France. Crochard, Paris.

Le Comber, S.C., Nicholls, B., Rossmo, D.K., Racey, P.A., 2006. Geographic profiling and animal foraging. Journal of Theoretical Biology 240, 233–240.

Le Comber, S.C., Rossmo, D.K., Hassan, A.N., Fuller, D.O., Beier, J.C., 2011. Geographic profiling as a novel spatial tool for targeting infectious disease control. International Journal of Health Geographics 10, 35–42.

Martin, R.A., Rossmo, D.K., Hammerschlag, N., 2009. Hunting patterns and geographic profiling of white shark predation. Journal of Zoology 279, 111–118.

Mosher, C.J., Miethe, T.D., Hart, T.C., 2011. The Mismeasure of Crime, second ed. Sage Publications, Los Angeles, CA.

Perreault, S., Brennan, S., 2010. Criminal Victimization in Canada, 2009. Statistics Canada, Ottawa, ON.

Quetelet, L.A.J., 1842. A Treatise on Man. William and Robert Chambers, Edinburgh.

Raine, N.E., Rossmo, D.K., Le Comber, S.C., 2009. Geographic profiling applied to testing models of bumble-bee foraging. Journal of the Royal Society Interface 6, 307–319.

Rossmo, D.K., 2012. Recent developments in geographic profiling. Policing: A Journal of Policy and Practice 6, 144–150.

Rossmo, D.K., 2014. Geographic profiling. In: Weisburd, D., Bruinsma, G.J.D. (Eds.), Encyclopedia of Criminology and Criminal Justice. Springer-Verlag, New York, pp. 1934–1942.

Rossmo, D.K., Harries, K.D., 2011. The geospatial structure of terrorist cells. Justice Quarterly 28, 221–248.

Sherman, L.W., Gartin, P., Buerger, M.E., 1989. Hot spots of predatory crime: routine activities and the criminology of place. Criminology 27, 27–55.

Smith, W.R., Frazee, S.G., Davison, E.L., 2000. Furthering the integration of routine activity and social disorganization theories: small units of analysis and the study of street robbery as a diffusion process. Criminology 38, 489–523.

Statistics Canada, 2012. Uniform Crime Reporting Survey (UCR). http://www23.statcan.gc.ca/imdb-bmdi/pub/indexU-eng.htm (accessed 30.06.13.).

Stevenson, M.D., Rossmo, D.K., Knell, R.J., Le Comber, S.C., 2012. Geographic profiling as a novel spatial tool for targeting the control of invasive species. Ecography 35, 704–715.

Weisburd, D., Bushway, S., Lum, C., Yang, S.-M., 2004. Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. Criminology 42, 283–321.

Weisburd, D., Groff, E.R., Yang, S.M., 2013. Understanding and controlling hot spots of crime: the importance of formal and informal social controls. Prevention Science, in press.

Spatial Equity

Richard Morrill, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial equity concerns the perceived unfairness of the well-being of individuals and of societies across space. The contested meaning of spatial equity is reviewed and discussed in six example arenas: social inequality (race, gender), economic inequality (income, wealth, opportunities), regional inequality, political representation, location of undesirable facilities, and urban inequality. Analysis of, and the varied reactions to, spatial inequity are discussed.

Spatial equity concerns the perceived unfairness of the well-being of individuals and of societies across space. Not only is the physical environment very uneven in its utility for human activities, the evolution of societies results in marked inequality in access to resources, rights, and rewards. If the degree of inequality is viewed as unfair and perhaps as imposed unfairly on persons and places, societies may adopt standards and institute programs to mitigate inequality, to reduce segregation, to improve mobility and access to resources, and to redistribute real income.

Although most social scientists may not know the term 'spatial equity,' almost everyone experiences it in everyday life. 'Spatial' implies that most of the experienced world, physical and social, takes up room; this has important social consequences. Elements of the natural and social world always have a location in space or a place, just as they do in time. Phenomena also have relationships of distance, spatial interaction, and accessibility with one another, and these attributes shape the causes and consequences of things. 'Space' has structure, that is, institutions and relationships that constrain what and who can be where – in other words, everything and everyone's place on the landscape.

'Equity' is a social construct by individuals, groups, or societies as to the degrees and kinds of inequalities that are acceptable or 'fair' or are unacceptable and 'unfair.' Thus the measure of equity is inescapably bound up with the individual or group position with respect to the basic human dilemma of individual and collective rights and responsibilities. There is no measure of where fair ends and unfair begins; society redefines the line as needed according to custom and law. Classic works include Perin's *Everything in its Place* (1977), Rawls' *A Theory of Justice* (1971) and Harvey's *Social Justice and the City* (1973). The latter was recently commemorated in Merrifield and Swyngedouw's *The Urbanization of Injustice* (1997). On Rawls, see also Clark (1985).

The Meaning of Spatial Equity

Webster defines equity as 'a legal system which supplements statutes or laws to correct, if necessary, its failure to account for what is morally just and fair.'

Societies may deal with questions of equity and fairness in the justice system or via regulations and legislated standards. Thus a 13-year-old may not claim an equity issue in not being allowed to drive, but a healthy 80-year-old can.

The very fact of space implies that the notion of spatial equality (that all parts of space are of the same quality or value) is without meaning; every bit of space is different from every other. For this reason, equity is of heightened importance and difficult to obtain. In economic theory an efficient equilibrium should be equitable as well. Structurally, monopoly and inheritance, for example, lead to inefficient and inequitable outcomes. But space itself may make a structurally efficient pattern geographically inequitable. Consider a common situation of an uneven distribution of people, far denser on one side of a jurisdiction than the other. The most efficient, travel-minimizing location for a public facility will be strongly pulled to the densest core, leaving those in the low-density edge in a highly unequal position (Morrill, 1973).

The physical environment is incredibly diverse, and utterly unequal from place to place and over time in its utility or value to human production, residence, exchange, or enjoyment. For physical, cultural, historic, and political reasons, individuals, groups, corporations, institutions, and societies are similarly diverse, and highly unequal in physical and mental capacity, experience, and mobilization of resources. They are profoundly unequal in their power to structure the landscape, to alter the environment, or to control more valuable locations.

Individuals, groups, and societies are continually aware of their unequal positions in space, which often leads to perceptions of unfairness on the part of the less territorially advantaged (those on the 'wrong side of the tracks'), and on the part of moralists, ethicists, and liberals offended by or concerned with territorial inequality.

Spatial Equity and Social Inequality

Inequality of power and of rewards itself may lead to resentment and a profound sense of unfairness of the social system. At the societal level, the equity issue is the social system itself, and the battles are more structural than spatial; for examples, calls for socialism or communism as the solution for unequal power and wealth, or battles over minimum wages and tax rates. An exception is perhaps the demand for political self-determination of colonies – struggles that were more territorial and cultural than structural.

But within a nation and social system, as in the United States, although many may want to change the system, the majority of U.S. citizens, for example, accept the competitive and individual system and understand that unequal power and wealth are inherent to the system, that the rich and powerful

can and will preempt the 'best places.' Thus a class-segregated society as such is not a spatial equity issue. After all, that poor person may win the lottery! Rather, issues of spatial equity arise when people feel unfairly treated, although they are playing within the rules of the game; when they feel that the rich and powerful manipulate the law itself to get more than the share that their competitiveness and wealth warrant. An example is the conviction of poverty-area residents that since they pay taxes, too, the potholes on their streets should have no lower priority than the potholes on the streets of the rich.

There are many examples of spatial inequality that raise equity issues. Starting at the global scale are historical colonialism and imperialism and contemporary dependency theory, hypothesizing that unfair power relations exploit and impoverish the Third World. Less technologically advanced cultures distrust and fear the power of Western and especially of U.S. culture. Smith (1977, 1994) provides extensive treatment of inequality from the global to the local scale. See also Knox (1975).

Within nations, prominent bases for raising issues of equity include (1) regional inequality, the sense of exploitation of the periphery by the core, or of poor regions by rich regions; (2) unequal access to public facilities and resources; (3) unequal or unfair treatment of areas because of their poverty, wealth, or power; (4) spatial manifestations of the sports analogy that a competitive society should have a 'level playing field'; and (5) spatial aspects of unequal treatment of categories of people because of race, ethnicity, language, religion, lifestyle, and gender.

Regional Inequality

Within the European Union, a dominant part of the budget and a large amount of research and rhetoric are devoted to the issue of lagging regions, built on the perception that it is somehow unfair that some regions are prosperous and growing while others are poor and languishing; that it shows the failure of the system rather than of the region (Kuntzmann, 1998). In U.S. economic history, class struggles often seemed secondary to the resentment of the South and West against the hegemony of the Northeast. In a generally private economy, the United States has a long tradition of regional redistribution via public programs; for example, the Reclamation Act subsidizing irrigation development in the West, federal defense outlays and massive highway programs, the Tennessee Valley Authority and the Bonneville Power Administration, and the Regional Development Commissions were all designed to improve the competitive position of disadvantaged parts of the country (Johnston, 1980). Harrington's classic *The Other America* (1962) raised the nation's consciousness of poverty and helped lead to the War on Poverty.

Efficient market decisions frequently take the form of the closure of factories and consequent community decline, leading to forced out-migration. In these cases, 'pure market' solutions implying the 'shipment' of the surplus labor to growing regions confront the immense nonmarket value of places to their long-term residents, and a strong sense of unfairness arises, that the decline is not the fault of local people (Clark, 1983). Societies often institute a variety of programs to reduce the social and economic costs, such as through subsidies, community purchase of the closed plant, tax forgiveness, or retraining and relocation assistance.

The same principle is applied at the local level through federal or local programs such as slum clearance, public housing, and 'enterprise or empowerment zones,' reflecting a political consensus that the landscape had become unfairly unequal, presumably because of forces beyond local control.

Voting and Representation

In the United States, perceptions of unfairness in representation led to a series of Supreme Court decisions outlawing malapportionment (unequal population per district) and racial gerrymandering that disenfranchised minorities (initial decisions ruled against 'diluting' and 'packing' minority voters, but later rulings forbade gerrymandering to maximize minority representation). Ironically, representation in the U.S. Senate intentionally violated the equal population principle, since every state has two senators, but this inequality was defended on grounds of avoiding unfair domination of smaller by larger states. But gerrymandering that is purely 'political' is allowed and widely practiced.

Overcoming the 'Unfairness' of Nature

As communities develop and compete with other places, perceptions of being treated unfairly by nature may arise. At the international scale, such perceptions may be used to justify invasions and wars. Within countries, governments at all levels spend huge sums to overcome natural limits, so they may compete on that more level playing field. Examples are bridges, tunnels, and artificial harbors built to improve the accessibility of a place to national and global markets.

Access to Public Facilities and Resources

This is the arena for which the word 'equity' is most likely to be invoked, and for which legal remedies may be sought on grounds of equal protection under law (Pinch, 1985). In the United States, until as recently as the 1960s it was the perception and probably the reality that weaker, poorer places among a set of jurisdictions, and weaker and poorer areas within a jurisdiction, were unequally and unfairly treated. Examples of inequitable access to public services and facilities entail the maintenance and quality of utilities and infrastructure, like roads; the location and quality of parks, playgrounds and pools, or of libraries; and the timeliness of police response calls, as if action were a function not of citizenship but of wealth – which it was. As a consequence of a series of court decisions over many years, cities have become far more attentive to more equal treatment across all parts of their jurisdictions, although huge historic inequities remain.

The construction and operation of public transportation systems have similarly been subject to scrutiny, as to the costs and benefits of subsidies to different parts of a region; poor, more transit-dependent districts no longer meekly accept inferior service and equipment. In most of the United States, school districts are a distinct set of governments. District populations are astoundingly unequal in income, and districts in tax bases. In many but not all states, courts have

interpreted state constitutions to mean that all students should have equal access to education, wherever they live. This has resulted in substantial redistribution of resources from richer to poorer districts, as well as to the subsidy of 'remote but necessary' small schools in very isolated communities.

The Location of Undesirable Facilities: Environmental Racism

A different issue of equity arises as to the differential ability of stronger and richer, or weaker and poorer, jurisdictions or parts of cities to ward off undesirable private and public activities. Less advantaged places widely perceive that they have long been and remain dumping grounds for unattractive and even damaging facilities – prisons, halfway houses, incinerators and landfills, and services to the 'less desirable' folks (drug addicts, alcoholics, the homeless); adult shops and theaters; and polluting and noisy factories. If the obnoxious facilities or activities are in segregated racial minority areas, charges of 'environmental racism' may be made (Bowen et al., 1995; Lester et al., 2001).

Poorer and minority areas also perceive also they are treated as tolerance zones for illicit activities and inadequate law enforcement. More generally, these areas are characterized by loss of jobs, retailing, services, and middle-class housing, leading to spatial mismatch between a declining inner city and more affluent suburbs. In more liberal cities and states, a consensus may arise for some degree of sharing the 'burden' of housing the less affluent, as through requirements for 'fair shares' of subsidized affordable housing, even in suburbs. In a very few metropolitan areas, where extreme political fragmentation results in highly inequitable resources for jurisdictions, a sense of injustice and unfairness has resulted in a degree of tax-base sharing, as for police and sewerage services.

Urban Inequality: Unequal Treatment of Categories of People

Many books focus on spatial inequality in U.S. and European cities, including Badcock (1984), Gottdiener (1985), and Katznelson (1981). An influential segment of the intellectual and planning community advocates fairly radical efforts to break down the high degree of racial and class segregation, as through stronger zoning control of settlement. The failed programs of mandatory busing for school desegregation were similarly based on a sense of injustice and on the premise that mixing in classrooms would lead to mixing in neighborhoods. As with schools, current efforts toward class mixing seem unlikely to be a match for the stronger societal forces for segregation. Although men and women are usually residentially integrated, the history of gender inequality resulted in unequal access to opportunities and resources, and in insecurely built environments. And because of this history, the very subject of inequitably gendered landscapes has only recently been studied (Pratt and Hanson, 1994).

Reactions to Spatial Inequity

As most of these examples show, the achievement of greater spatial equity seems limited to those aspects of everyday life in which public resources or facilities are involved. Social intervention in the form of redistribution of resources toward weaker places and areas seems to occur when a social consensus of 'market failure' arises – that the playing field has become too uneven, within the prevailing social order. Within the social system, it proves easier also to address inequities in some arenas of life than in others. Not surprisingly it is less threatening to those on top, usually the holders of wealth, to accept equalizing measures with respect to political participation and voting and social dimensions (race, language, and gender, although these remain highly contested), and even, if rarely, urban housing, than with respect to the truly key economic arena of income, taxes, and inheritance.

In the United States, issues of equity and redistribution were fought mainly at the national level for most of the period following World War II. Since the mid-1980s and national shifts toward the right – that is, toward greater individual than collective responsibility – the arena of action has moved to local and state levels, where coalitions of activist groups have sometimes been able to gain enough power to enact redistributive plans unimaginable at the national level, such as high minimum wages, rent controls, or mandates for affordable housing, in a process that might be called 'pluralist democracy' (*see* Postmodern Urbanism). In other areas and arenas, conservative coalitions stress faith in the market, privileging its efficiency over concerns for equity (Alexiadis and Eleftheriou, 2010). Indeed, the concepts of efficiency and equity are the very symbols of differing ideological stances toward inequality. The treatment of equity (including this discussion) is usually based on a liberal assumption of people's similarity in needs and values – ideas going back at least to John Stuart Mill. The new politics of identity and difference may complicate the meaning and evaluation of equity (Keith and Pile, 1993; Young, 1990).

The Analysis of Spatial Equity

Maps, and now geographic information systems (GIS), have proven highly effective in revealing the structure of inequality in the landscape – patterns of segregation, poverty, concentration of pollution, and unequal access – and have proven essential to convince legislative bodies and courts to address equity issues. A variety of both statistical and mathematical models have been used to measure inequality and its costs and to assess the impacts of policy changes. Awareness of the human costs of spatial inequality may be increased through more ethnographic studies of the situation of individual households.

Recent Developments

In the years since 2000, global capital has triumphed. Development disparities between countries have fallen, even while income inequality within countries has soared. This may have caused suffering and resentment, but not enough to alter the current social milieu. Indeed, social and political intervention with respect to social–spatial inequality (growing and declining, rich and poor regions or parts of cities) has almost disappeared, and even long-standing public enterprises, such as social housing, are being privatized. Thus the analysis of spatial equity has tended to narrow, focusing on environmental

injustice (Callewaert, 2002), transportation system planning (Monzon et al., 2011), and access to public facilities (Tsou et al., 2005). Additional efforts have been applied to measurement, including the importance of scale (Liao et al., 2010) and visualization (Talen, 2011). A few studies have addressed wider social processes, such as Abel et al. (2011) on the effects of gentrification on the poor, and studies of differential access of racial and income groups (Zhao, 2011; Talen, 2001). For health services, race was critical (see, for example, Kaukinen and Fulcher, 2004, on access to HIV services; and Zenk et al., 2006, on mammogram services). But in Ohio poverty was more important with respect to environmental justice (Bowen, 2001). These at least imply that spatial equity is something that should be addressed.

The complexity of the arena of spatial equity can be seen by the disparate reactions to the widespread implementation of 'smart growth' densification on environmental grounds, despite possibly adverse impacts on the less affluent (Pozdena, 2002). Indeed, while governments invoke planning to address social and spatial equity issues, the on-the-ground implementation of planning may often increase rather than reduce inequity.

See also: Access: Geographical; Location Theory; Location, Absolute and Relative; Postmodern Urbanism; Racism, Sociology of; Residential Segregation: Geographic Aspects; Spatial Interaction; Spatial Pattern, Analysis of; Urban Poverty in Neighborhoods.

Bibliography

Abel, T., Daniel, K., Jonah, W., 2011. Gentrifying inequities and skewed risk-scapes: the microgeography of environmental injustice in Seattle, AAG Annual Meeting.
Alexiadis, S., Eleftheriou, K., 2010. The trade-off between spatial equity and economic efficiency revisited. Evidence from the US states. In: Econ Papers (MPRA Paper from University Library of Munich).
Badcock, B., 1984. Unfairly Structured Cities. Basil Blackwell, Oxford, UK.
Bowen, W., 2001. Environmental Justice through Research-based Decision-making. Garland Publishing, New York.
Bowen, W., Salling, M., Haynes, K.E., Cyran, E., 1995. Toward environmental justice: spatial equity in Ohio and Cleveland. Annals AAG 85.
Callewaert, J., 2002. The importance of local history for understanding and addressing environmental injustice. Local Environment 7, 257–267.
Clark, G., 1983. Interregional Migration, National Policy and Social Justice. Rowan and Allenheld, Totowa, NJ.
Clark, G., 1985. Making moral landscapes: John Rawls's original position. Political Geography Quarterly 5, 147–162.
Gottdiener, M., 1985. The Social Production of Space. University of Texas Press, Austin, TX.
Harrington, M., 1962. The Other America. Macmillan, New York.
Harvey, D., 1973. Social Justice and the City. Johns Hopkins University Press, Baltimore, MD.
Johnston, R., 1980. Geography of Federal Spending in the United States. Wiley, Chichester, UK.
Katznelson, I., 1981. City Trenches. Pantheon, New York.
Kaukinen, C., Christopher, F., 2004. Visualizing the accessibility and spatial equity of HIV related services: A geographic analysis of Chicago neighborhoods. Population Association of America, Annual Meeting, 2004.
Keith, M., Pile, S. (Eds.), 1993. Place and the Politics of Identity. Routledge, London.
Knox, P., 1975. Social Well-being: A Spatial Perspective. Oxford University Press, London.
Kuntzmann, K., 1998. Planning for spatial equity in Europe. International Planning Studies 3.
Lester, J., Allen, D., Hill, K., 2001. Environmental Injustice in the United States. Myths and Realities. Westview Press, Boulder, CO.
Liao, C.-S., Hsueh-Sheng, C., Tsou, Ko-Wan, 2010. Explore the Spatial Equity of Urban Public Facility Allocation Based on Sustainable Development Viewpoint. Urban Planning. National Cheng Kung University, Taiwan.
Merrifield, A., Swyngedouw, E., 1997. Urbanization of Injustice. New York University Press, New York.
Monzon, A., Ortefa, E., Lopez, E., 2011. Efficiency and spatial equity impacts of high-speed rail extensions in urban areas. Universidad Politecnica de Madrid. Cities.
Morrill, R., 1973. Efficiency and equity aspects of optimum location models. Geographical Analysis 9, 215–226.
Perin, C., 1977. Everything in Its Place. Princeton University Press, Princeton, NJ.
Pinch, S., 1985. Cities and Services. Routledge, London.
Pozdena, Randall, 2002. Smart Growth and Its Effects on Housing Markets: The New Segregation. National Center for Public Policy Research (Quant Econ, Inc.).
Pratt, G., Hanson, S., 1994. Geography and the construction of difference. Gender. Place and Culture 1, 5–30.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Smith, D., 1977. Human Geography – A Welfare Approach. Edward Arnold, London.
Smith, D., 1994. Geography and Social Justice. Blackwell, Cambridge, UK.
Talen, E., 2001. School, community and spatial equity: an empirical investigation of access to elementary schools in West Virginia. Annals of AAG 81.
Talen, E., 2011. Geovisualization of spatial equity. In: The SAGE Handbook of GIS and Society. Sage (Chapter 24).
Tsou, Ko-Wan, Hung, Yu-ting, Chang, Yao-Lion, 2005. An accessibility based integrated measure of relative spatial equity in urban public facilities. Urban Planning, National Cheng Kung University Taiwan. Cities.
Walzer, M., 1983. Spheres of Justice. Basic Books, New York.
Young, I., 1990. Justice and the Politics of Difference. Princeton University Press, Princeton, NJ.
Zenk, S.N., Tarlov, E., Sun, J., 2006. Spatial equity in facilities providing low or no fee screening mammography in Chicago neighborhoods. Journal of Urban Health 83, 195–210.
Zhao, L., 2011. Spatial Equity Analysis in the Atlanta 10-county Region. Thesis, University of Alabama.

Spatial Interaction

Morton E O'Kelly, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by A.S. Fotheringham, volume 22, pp. 14794–14800, © 2001, Elsevier Ltd.

Abstract

Spatial interaction involves a wide range of flows between nodes: these include human movements (e.g., migration, commuting) and movement of materials (e.g., international trade). These flows are important aspects of urban and regional dynamics. At the regional level, interactions include many modes and types of flow, such as rail, air, highways, and even pipelines. In cities, spatial interaction takes a wide variety of forms and includes commuting to work and travel for shopping and other purposes. This article presents a brief guide to some of the more important aspects of the discoveries and applications in this field. Practical areas include aggregate and disaggregate models for flows.

Introduction

Spatial interaction deals with flows of people or material between places. The level and pattern of such interaction is a reflection of the need for movement (complementarity) and the difficulty or ease of overcoming intervening distance which serves as an impeding factor. In short, people and places matter (push and pull factors), distance matters, and spatial effects such as competing opportunities, clustering, and agglomeration matter. In examining spatial interaction, models can assist in answering some basic questions: How can we observe and measure these flows? Where are the flows coming from and going to? What problems arise in trying to match flows and infrastructure (e.g., congestion in surface transport, high costs for port installation, or environmental impacts of massive airports)?

Generally, the heaviest spatial interaction occurs along certain main lines of connection (trunk lines). Infrastructure is repeatedly overhauled and expanded to accommodate the *positive feedback* from ease of interaction to the demand for interaction between places (often nodes or places in cities). At the regional scale, the explosive growth of volume of flows between cities in the Northeastern corridor of the US dubbed 'Megalopolis' (Gottmann, 1957) has been accompanied by a dramatic drop in time needed for travel between cities (Janelle, 2005). Interaction can also reach critical levels of congestion: roads built for current capacity are congested soon after they open. Within cities, rents and density rise near transit stations. Spatial interaction is an important topic in the field of spatial analysis. Flows between units (zones, cities, nodes, counties) are hypothesized to relate to the sizes of the origins (sources) and destinations (sinks) and are inversely proportional to the intervening distance. This idea had its origins in simple *gravity models* but has progressed to provide a basic foundation for measurement of large-scale movement such as international trade, freight flows, and air passenger transportation. In addition, the topic also covers accessibility, commuting, trade areas, and many other elements involving movement and interaction in cities. Significant recent progress has been made in computing and mapping these measures. Basic texts have provided the elementary introduction to the notation and help to overcome the 'black box' nature of spatial interaction modeling techniques (Fotheringham and O'Kelly, 1989). The flows are also amenable to interactive spatial data analysis and to many types of models (e.g., log linear modeling; Upton and Fingleton, 1985). Additional resources on this topic include reference works on Isard (O'Kelly, 2004), Wilson (O'Kelly, 2010), and the online content of journals such as *Environment and Planning* and *Geographical Analysis*.

Modeling Spatial Interaction

SI models consider flows between origins (i) and destinations (j). Typically, the goal is to understand the flows or trips (T_{ij}) between i and j, and possibly to predict them in a future scenario. In a generic sense, the flows are a function of the spatial separation between i and j, as well as the characteristics of origins and destinations. Most SI models start with spatial interaction flows and treat them as dependent variables; these are modeled with independent effects including (1) a push factor representing the propulsiveness of each origin (e.g., the number of households in an origin zone); (2) a pull factor representing the attractiveness of each destination (e.g., size of a shopping center); and (3) a measurement of the spatial separation between origins and destinations (e.g., time, cost, or distance to overcome).

The end result depends on how we combine these elements and the precise role each element plays in particular specialized types of interaction. Spatial separation is usually measured using time or distance as a surrogate for the cost of overcoming the space between origins and destinations. Distance is used to denote the difficulty (perceived or real) of overcoming spatial separation; often distance is a proxy for the lack of information and decreased opportunity that occur when origins and destinations are widely separated. Distance is clearly not the sole determinant of behavior patterns as people often travel past the nearest available option in order to reach greater variety or better prices. Nevertheless, there has been a dominant concern in the literature with the principle of least effort and on the role of convenience. There is a strong expectation that interaction declines with increases in distance once other factors are controlled. In practice, the parameter used for the partial effect of distance is often written with a negative sign. To avoid

confusion, the reader should understand that, if $t = f(d)$, where t is number of trips or interactions and d is a measure of distance, there are two ways to mark up the distance decay effect. For example, $t = a + bd$ (where $b < 0$) or $t = a - bd$ (where b is a positive number entering with a negative sign). The beginner in spatial interaction models would do very well to understand that, while these two resultants are the same, the literature tends toward the latter means of notation. For example, Wilson's path-breaking work in spatial models frequently includes a function such as $t = \exp(-bd_{ij})$.

Another key factor is the extent to which data are treated as observed exogenous effects or are developed from within an interaction framework. To make this idea clear, suppose we *observe* a lot of interactions between counties. We could add up the observed interactions *leaving from* or *arriving at* each county and use these observed totals as controls on the values of any modeled estimate. If the estimates are arranged in a matrix with rows representing origins and columns representing destinations, we would logically require that the observed values control the marginal total of the *estimated* matrix. In other words, we could measure interaction and then take some care to make the model reproduce these data in aggregate, through a numerical balancing process called biproportional matrix adjustment. Another key departure is when we have some inconsistent measure of trip propulsion (from origins) and attraction (to destinations) from an independent set of proxy variables. The model framework can be set up to tolerate these inconsistencies quite well and will make estimates maximally consistent with the known data. In this case, the push and pull factors can be used as covariates that tip the interaction in one way or another but that are not strictly required to serve as controlling or constraining totals. This is a subtle but important distinction. When we use data as controls, we need parameters or balancing factors (fixed effects) to make sure our model adjusts to match these data. When we have exogenous surrogates that are trending in the data, we can use them as covariates (perhaps with associated signs that are calibrated if we can study these ideas using some observed data) but not as firm constraints.

To be clear, any time that an origin (by convention row i), or a destination (by convention column j), is unconstrained, the model is free to have such a variable act as a push or pull factor modifying the flows as a form of covariate. In fact, distance is such a covariate. It does not necessarily serve as a constraint on the interactions, but the higher and lower values of distance are expected to have a predictable role in the levels of these interactions. The actual strength of that effect is parameterized by a distance decay effect usually denoted with the Greek symbol beta.

These various types of SI models belong to a family (Wilson, 1971) (**Table 1**); instead of a series of ad hoc parameterizations of effects, the same models (with or without certain individual constraints) may be driven toward some special cases, often matched quite well to the substance and dynamics of the real world problem. For example, in the case of residential mobility, one could imagine a very short-term view where the places of residence and work are fixed and the only adjustment is the journey-to-work path. In the medium term, people are located in a fixed home, but could potentially readjust their place of employment residence. Finally, in the longer term, workers could consider both their residence and workplace locations as flexible.

In what follows, we present examples of three types of models: unconstrained (often used for migration 'M'), production constrained (often used for retail sales 'S'), and doubly constrained (often used for commuting trips 'T'). The decision to treat certain effects as fixed or as covariates needs to be considered carefully for each particular case. Nevertheless, this presentation gives a modern view of the ways that the models are linked without repeating well-known textbook treatments of constraints. There is a fourth model class which is highly similar in structure to the production constrained case (attraction case), but in the interest of a compact explanation, it is not reviewed here.

Migration

One way to write *The Unconstrained Model* is as

$$M_{ij} = f(V_i, W_j, d_{ij})$$

where M_{ij} is the number of migrants that begin at origin i and end at destination j, V_i represents an origin attribute that is a surrogate for the number of migrants beginning at i, W_j represents a destination attribute that is a surrogate for the number of migrants ending at j, and d_{ij} represents the spatial separation (assumed here to be measured by distance) between origin i and destination j. In fact, V_i and W_j can represent vectors with several key variables. **Table 2** (based on Fotheringham and O'Kelly, 1989) summarizes the types of directions in which these specific factors are expected to affect migration. In **Table 2**, V_i is used to represent six independent variables and W_j represents eight (as an example).

Within the origin zone, the number of people, cost of living, crime rate, and unemployment rate are all expected to be positively correlated with values of flows (push factors). On the other hand, the level of amenities and wage rate within the origin zone are expected to be negatively correlated with values of flows (negative push, i.e., retention factors). Typically, reversed patterns occur at the destination end. The hypothesized relationships between these attributes and the origin or destination values have generally been confirmed in many empirical studies. The positive sign relating number of people with both origins and destinations may seem contradictory since the same variable which explains the 'push' away from a zone will also explain the 'pull' of flows into a zone.

Table 1 Models for spatial interaction

	Production constrained	*Attraction constrained*	*Doubly constrained*
Origin	Model matches O_i	V_i as a covariate	Model matches O_i
Destination	W_j as a covariate	Model matches D_j	Model matches D_j
Balancing factors	A_i	B_j	A_i and B_j

Table 2 Typical empirical effects on migration (based on Fotheringham and O'Kelly, 1989)

Migration attribute of origin (push)	*Relationship* dM_{ij}/dV_i	*Migration attribute of destination (pull)*	*Relationship* dM_{ij}/dW_j
1. Number of people	+	1. Number of people	+
2. Cost of living	+	2. Cost of living	−
3. Level of amenities	−	3. Level of amenities	+
4. Crime rate	+	4. Crime rate	−
5. Wage rate	−	5. Wage rate	+
6. Unemployment rate	+	6. Unemployment rate	−
		7. Average house price	−
		8. Average temperature	+

Important early work on this topic at the Leeds School (Cordey-Hayes and Gleave, 2005) has confirmed that migrants in a city have a shorter duration of residence and may therefore be prone to another move or to return migration. These feedback effects are known to require a more elaborate estimation effort (Muth, 1971). The heavy attraction of a city for migrants also presumably pushes up the demand for housing and can make the attractive destination have a higher house price, an effect noted by Congdon (2010). Generally, his results confirm the statistical expectations in **Table 2**. Recent research on this problem continues, and scholars have emphasized the need to correct for biases that arise when key effects are omitted. Often what appear to be contrary results can be corrected and reconciled through the use of more elaborate specifications and more careful assumptions about process. Care is certainly required when the models are deployed in situations with added complexity. For example, regional migration within Mexico is clearly influenced by proximity to the US border (Peeters, 2012). These techniques have been deployed on some massive regional movements such as interregional movement and migration in China (Shen and Wang, 2012).

Retail

Spatial interaction to retail centers involves a fixed pool of demand at the origin and distributes that demand over a set of predefined retail locations. As such, the SI model is an ideal method for the estimation of probabilistically determined trade areas (as opposed to more deterministic methods such as Voronoi diagrams). The long tradition stemming from Reilly's law of markets allows the mapmaker to divide the territory between towns into trade regions or catchment areas. A particularly useful version of this problem is the weighted Voronoi diagram which combines the spatial separation due to distance decay and the attractive weight of the nodes (masses) (see O'Kelly and Miller, 1989).

Issues of spatial choice and sampling also intersect heavily with this topic. Because it is generally easier to collect a sample at the choice-based level (interviews at a shopping center, for example, rather than attempt to find shoppers for a destination in a random stratified sample), the issue becomes one of rescaling a choice-based sample to give a general idea of the probability of a respondent selecting option *j*. This is an exercise in Bayesian analysis. We would like to know how a consumer is attracted to the various options as they are spatially arranged – and whether this is influenced by the cluster or agglomeration effects (see Fotheringham's theory of competing destinations). Because of the relatively large range of choice that the consumer may face, including online purchases, and conversely the large range of origins from which a store draws demand, there is some technical challenge to visualizing the territories. It may be possible to show multiple attractions as a 3D graphic, or it may be easier to draw one map with isolines (contours). O'Kelly (2001, 2009) has provided an extensive review of retail location using spatial interaction tools.

In terms of the expected empirical effects, there are some common sense findings: The origin demand side is strongly related to income and family size (Bawa and Ghosh, 1999). Similar research has been performed by Fox et al. (2004), which divided the stores into different categories such as drug stores, grocery stores, and mass merchandisers. The results indicate that, as expected, income and family size have positive relationships to shopping expenditure. Given that pool of expenditure, the destination attractions are related by spatial models. Oppewal and Timmermans (1997) examine the effects of shopping center size and store variety on destination choice. This literature focuses on modeling the destination attributes such as price, product quality, service, and location accessibility. A logit model (a form of discrete choice model closely related to spatial interaction concepts (Anas, 1983)) is made for two categories of stores: food centers and clothing/shoes stores. The results quantify the way in which shopping expenditure is significantly influenced by selection attributes such as available variety and quality of goods, as well as the appearance and layout of the stores.

Commuting

Prior to innovative work by Wilson (1971) empirical analysis tended to focus on either employment or residential data and to view these two types of data as potentially inconsistent or, at the very least, difficult to reconcile. The spatial interaction representation of journey-to-work provides a means to *connect* these two aspects. The flows originating at zone *i* and destined for zone *j* are arranged in a table (T_{ij}) where the rows of the table point to origins and the columns represent destinations. In database terms, the spatial interaction matrix provides a pivot table (or a cross-walk) to link the origins and the destinations of the trip makers (O'Kelly et al., 2011).

Commuting and interaction involve a complex set of relationships among average journey length, the degree of randomness or organization in the pattern (entropy), and a distance decay effect typically measured by 'beta.' Peak travel during AM and PM rush hours accounts for the highest concentration of traffic on the highways around large metro areas. Significant variations in the lengths of these journeys by gender and by socioeconomic status have been noted (Sang et al., 2011). The data also are fraught with issues of openness of the system: for example, a large regional hospital might employ individuals from a multicounty region. The trips come from an area that far exceeds the immediate vicinity of the hospital. (See a case study concerning Rochester, Minnesota, home of the Mayo Clinic – Sang et al., 2011.) Such commuting patterns also have an important impact on the morphology of edge cities – where the ebb and flow of traffic required by the spatial separation of places of work and residence often stimulate congestion at places that are far from the traditional CBD core (O'Kelly and Mikelbank, 2002). Inevitably, spatial effects are influenced by the somewhat artificial placement of zonal boundaries (Putman and Chung, 1989). Cross-border effects arise as spatial interaction is an open system concept and intrinsically involves flow. It is difficult to put a line of demarcation on a map and say that all trips originating inside a region terminate inside the same region (they may in fact leak out).

Intensive interactions of this kind can place enormous stress on urban transport infrastructure (Hanson, 2004) especially in the absence of significant share using modes other than single occupancy vehicles. There is also a mismatch in the locations of places where entry-level jobs (which might be attractive to certain work force participants) in the logistically advantageous beltway locations and the more dense pockets of urban dwellings (old inner city suburbs). Lack of convenient public transit options linking these places is problematic. Such imbalance between work and home locations is referred to as a spatial mismatch and has raised a host of related issues (Kain, 1992). In addition to the concern for excess commuting where people may be driving across the city to reach suitable employment (O'Kelly and Niedzielski, 2009), such commuting could potentially require many hours of driving in congested traffic or riding public transit.

Calibration

A number of interrelated but separate tasks – and a suite of related techniques for these tasks – is generally referred to as calibration: (1) observe and measure current conditions (description); (2) understand flows as they are observed (how do interaction levels compare to benchmarks or expectations?); (3) produce a model from a limited number of relevant 'explanatory' variables (forecast); and (4) manipulate the expected flow in terms of forecasted changes in exogenous variables (sensitivity analysis). Plenty of challenges arise from the intrinsic unpredictability of behavior, the difficulty of forecasting exogenous variables, and the openness of the systems at hand. Experiments have continued to refine the design of empirical studies of spatial interaction; indeed, the richness of the problem is that it leaves open the possibility of new and improved approaches.

Spatial interaction, as discussed, has a variety of components that are treated as fixed effects (and require a multiplier or balancing factor for each case) or may include covariates that act as weights. There are, therefore, distinctive roles for computation. One aspect of the calibration of such models is to make sure that the balancing effects reproduce the required control totals within a high degree of accuracy. Other parameters are needed to make sure that the model replicates broad aspects of the data. Calibration is well suited to the general problem of fitting log linear models (Tiefelsdorf and Boots, 1995). A good amount of attention also has been given to this topic by Upton and Fingleton (1985). In simple terms, calibration is designed to make the model reproduce some features of the observed data. Often this can be framed as a maximum likelihood problem. Associated with the idea of maximizing likelihood is a set of conditions (first-order necessary conditions) which can be framed as a set of nonlinear equations. As is well known in numerical methods, the solution of a system of nonlinear equations requires that the equations be set up consistently and possess a feasible solution. Single-variable versions of these problems are amenable to graphic solution, but multivariate problems require some computerized implementation with a nonlinear equation solver. A purpose-built method for this is SIMODEL (see discussion in Fotheringham and O'Kelly, 1989), and more recently the data handling capability of MATLAB has allowed many analysts to code their calibration routines in that language.

The prevailing method has included a variety of biproportional matrix adjustments. Biproportional matrix adjustment is the idea that the balancing factors are swapped back and forth – first fixing one and solving for the other, then fixing that one and solving for the first one, and so on. This iterative method usually will yield a set of parameters that maintains the origin (row) and destination (column) totals quite well. The other covariates are bit more complex. One might imagine that the parameters could be adjusted to get a high R^2, but in fact, subsequent theoretical development with a broad range of numerical practice shows that the correct approach is to find maximum likelihood estimates. This means that the parameters are chosen (numerically) to make the model reproduce these observed features of the data. For example, if we observe a mean or a variance, there may be a parameter that will be required to adjust for this. Except in very special circumstances, these parameters cannot be simply computed from the data and need to be determined by numerical iteration. Other special cases are found by solving related optimization problems such as linear programs. Finally, recent research has opened the interpretation of the sensitivity of parameters to changes in the data (O'Kelly, 2012).

Accessibility (Spatial Choice)

Mapping accessibility has a long history in geography. By focusing attention on the ways in which distance and sizes of places confer an uneven surface of opportunity, accessibility maps convey a strong contrast between core and peripheral locations. The evolution of new forms of interaction and the construction of a massive amount of transport infrastructure have tied these regions together, but despite the time–space

convergence reported by Janelle (2005), it remains difficult to overcome the disadvantages of a peripheral location. Aggregate and microaccessibility are two different scaled approaches to the measurement of accessibility. At the macrolevel, O'Kelly and Horner (2003) show how to incorporate the level of access to a place on the basis of its surrounding populations. These maps are direct descendants of the foundational work by Harris (1954) and provide a panoramic view of a national level of accessibility. Such concepts have been adapted to map, for example, the level of access and spatial integration of regions to the Internet backbone infrastructure.

In a widely cited review, Kwan et al. (2003) consider recent advances in representational issues in applied methods for accessibility. Such a microapproach is useful in realistic consideration of spatial choice. For example, when assessing the relative merits of alternative destinations, or of alternative spatial options (such as apartments), one might take into account space and time accessibility to potentially important locations in the urban area. Such measures of potential path accessibility with the temporal and spatial domains combined have been a fruitful area of analysis for geographers interested in individual accessibility. To give one concrete example, a free community clinic with specific opening hours and a location on particular bus routes would be accessible to a range of places that can reach the clinic. On the other hand, it is relatively unhelpful to a neighborhood separated from the clinic by a freeway and without convenient transit options.

The idea of a choice set can be operationalized by emphasizing the places that are readily accessed by various modes of transport and by taking into account detailed information such as opening hours and specific services provided. Indeed, geographical databases have evolved greatly from the earlier work on these issues, and now it is common for published research in the area of spatial interaction to combine a great deal of detailed information about the destinations (the supply side), the origins (the consumer demographics), and the intensity of interaction. Such measures might be obtained from logs of GPS data or by volunteered information from monitoring of the subjects' individual movements. While the foundations for such ideas go back a long way in geographers' analyses of spatial interactions, the newest studies tap an unprecedented level of salient information about individuals. In this regard, many of the limitations of older studies that stemmed from the shortcomings of travel diaries, or lack of microdata, and indeed crude spatial representation have been overcome. In short, this is a time of opportunity for spatial analysts to combine data with many rich conceptual schemas.

See also: Access: Geographical; Mobility and Travel Activity Patterns; Spatial Analysis in Geography; Time-Space in Geography; Transportation Geography; Urban Geography.

Bibliography

Anas, A., 1983. Discrete choice theory, information theory and the multinomial logit and gravity models. Transportation Research Part B: Methodological 17 (1), 13–23.

Bawa, K., Ghosh, A., 1999. A model of household grocery shopping behavior. Marketing Letters 10 (2), 149–160.

Congdon, P., 2010. Random-effects models for migration attractivity and retentivity: a Bayesian methodology. Journal of the Royal Statistical Society: Series A (Statistics in Society) 173 (4), 755–774.

Cordey-Hayes, M., Gleave, D., 2005. Migration movements and the differential growth of city regions in England and Wales. Papers in Regional Science 33 (1), 99–123.

Fotheringham, A.S., O'Kelly, M.E., 1989. Spatial Interaction Models: Formulations and Applications. Kluwer Academic Publishers, Dordrecht; Boston.

Fox, E.J., Montgomery, A.L., Lodish, L.M., 2004. Consumer shopping and spending across retail formats. Journal of Business 77 (2), 25–60.

Gottmann, J., 1957. Megalopolis or the urbanization of the Northeastern Seaboard. Economic Geography 33 (3), 189–200.

Hanson, S., 2004. The context of urban travel. In: Hanson, S., Giuliano, G. (Eds.), The Geography of Urban Transportation. Guilford Press, New York, pp. 3–29.

Harris, C.D., 1954. The market as a factor in the localization of industry in the United States. Annals of the Association of American Geographers 44 (4), 315–348.

Janelle, D.G., 2005. Spatial reorganization: a model and concept. Annals of the Association of American Geographers 59 (2), 348–364.

Kain, J.F., 1992. The spatial mismatch hypothesis: three decades later. Housing Policy Debate 3 (2), 371–460.

Kwan, M.-P., Murray, A.T., O'Kelly, M.E., Tiefelsdorf, T., 2003. Recent advances in accessibility research: representation, methodology and applications. Journal of Geographical Systems 5 (1), 129–138.

Muth, R.F., 1971. Migration: chicken or egg? Southern Economic Journal 37 (3), 295–306.

O'Kelly, M.E., 2001. Retail market share and saturation. Journal of Retailing and Consumer Services 8 (1), 37–45.

O'Kelly, M.E., 2004. Isard's contributions to spatial interaction modeling. Journal of Geographical Systems 6, 43–54.

O'Kelly, M.E., 2009. Applied retail location models using spatial interaction tools. In: Fotheringham, A.S., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE Publications, Los Angeles; London, pp. 419–442.

O'Kelly, M.E., 2010. Entropy-based spatial interaction models for trip distribution. Geographical Analysis 42 (4), 472–487.

O'Kelly, M.E., 2012. Models for spatial interaction data: computation and interpretation of accessibility. Computational Science and Its Applications. Lecture Notes in Computer Science 7334, 249–262.

O'Kelly, M.E., Horner, M.W., 2003. Aggregate accessibility to population at the county level: US 1940–2000. Journal of Geographical Systems 5 (1), 5–23.

O'Kelly, M.E., Miller, H.J., 1989. A synthesis of some market area delimitation models. Growth and Change 20 (3), 14–33.

O'Kelly, M.E., Mikelbank, B.A., 2002. Social change and transportation in U.S. edge cities. In: Black, W.R., Nijkamp, P. (Eds.), Social Change and Sustainable Transport. Indiana University Press, Bloomington, pp. 93–98.

O'Kelly, M.E., Niedzielski, M.A., 2009. Are long commute distances inefficient and disorderly? Environment and Planning A 41 (11), 2741–2759.

O'Kelly, M.E., Niedzielski, M.A., Gleeson, J., 2011. Spatial interaction models from Irish commuting data: variations in trip length by occupation and gender. Journal of Geographical Systems 14 (4), 357–387.

Oppewal, H., Timmermans, H.J.P., 1997. Modelling the effects of shopping centre size and store variety on consumer choice behaviour. Environment and Planning A 29 (6), 1073–1090.

Peeters, L., 2012. Gravity and spatial structure: the case of interstate migration in Mexico. Journal of Regional Science 52 (5), 819–856.

Putman, S.H., Chung, S.H., 1989. Effects of spatial system design on spatial interaction models. 1: the spatial system definition problem. Environment and Planning A 21 (1), 27–46.

Sang, S., O'Kelly, M.E., Kwan, M.-P., 2011. Examining commuting patterns results from a journey-to-work model disaggregated by gender and occupation. Urban Studies 48 (5), 891–909.

Shen, J., Wang, G., 2012. Estimating the contributions of migration factors to interprovincial migration in China 1995–2000. Annals of GIS 18 (4), 257–266.

Tiefelsdorf, M., Boots, B.N., 1995. The specification of constrained interaction models using the SPSS loglinear procedure. Geographical Systems 2, 21–38.

Upton, G., Fingleton, B., 1985. Spatial Data Analysis by Example. Wiley, Chichester; New York.

Wilson, A.G., 1971. A family of spatial interaction models, and associated developments. Environment and Planning A 3 (1), 1–32.

Spatial Mismatch

Sara McLafferty, University of Illinois at Urbana-Champaign, Champaign, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In many American cities, residential segregation and the suburbanization of job opportunities combine to reduce geographical accessibility to employment for African-American, inner-city residents. This 'spatial mismatch' is generally thought to be an important cause of the high rates of poverty and unemployment in inner-city neighborhoods. This article examines how researchers evaluate spatial mismatch; the impacts of spatial mismatch on employment and earnings; the roles of transportation access, racial discrimination, and spatial job search in spatial mismatch; and the variation in mismatch across metropolitan areas. Strategies to ameliorate spatial mismatch, including housing, transportation, and economic development policies, are also discussed.

Spatial mismatch describes a broad set of geographical barriers to employment that stem from a disparity, or mismatch, between where people live and where appropriate job opportunities are located. A more specific form of spatial mismatch, the spatial mismatch hypothesis (SMH), argues that the combined effects of residential segregation and economic restructuring limit geographical access to employment opportunities for African-American, inner-city residents. This article emphasizes geographical dimensions of spatial mismatch, highlighting the changing spatial distributions of residences and job opportunities in cities; the measurement of geographical separation between the two; the social, economic, and spatial processes that underpin spatial mismatch; and policy implications.

Spatial mismatch is a historically specific concept that emerged in the 1960s amid widespread concern about racial discrimination and urban poverty. Although he never used the term 'spatial mismatch', John Kain (1968) introduced the concept in analyzing the combined effects of residential segregation in housing markets and the postwar suburbanization of jobs on black unemployment. Poor geographical access to jobs, especially jobs in manufacturing, was said "to reduce employment opportunities of Negroes who are already handicapped by employer discrimination and low levels of education" (p. 196). Although some subsequent researchers have defined spatial mismatch in narrow geographical terms, as long distances to jobs or small numbers of locally available jobs, Kain's research was notable in exploring the intersections between spatial and social barriers to employment. He considered how transportation access, employer discrimination, skills, and social networks affect workers' access to employment. These themes persist in the spatial mismatch literature.

Kain's hypothesis stimulated discussion and debate during the early 1970s. After that, the field languished for a decade until William Julius Wilson's research on the urban underclass brought spatial mismatch back to the forefront (Wilson, 1987). Since then, the SMH has attracted research attention in a wide range of social science disciplines. It has also been expanded to encompass geographical barriers to employment for ethnic minorities, immigrants, and other potentially vulnerable populations (Ellis et al., 2004; Liu and Painter, 2012).

Evidence of Spatial Mismatch

The great appeal of the SMH is that it connects two prominent features of postwar American cities: the strong segregation of residential areas by race and ethnicity and the movement of jobs, especially jobs in manufacturing, to the suburbs and overseas. But do these processes necessarily result in spatial mismatch? Many central cities are not only losing employment, but they are also losing population (Stoll, 1999). Thus, the balance of jobs to resident workers has remained relatively constant. Although the SMH emphasizes city–suburban differences, fine-grained patterns of residential and work location are much more relevant to job access. Cities and suburbs are not homogenous; within each, African-Americans and other ethnic minorities are concentrated in specific areas reflecting the persistence of racial and ethnic residential segregation.

A key measure of spatial mismatch is job availability – the number and proximity of jobs in relation to the resident workforce. Within central cities, black neighborhoods generally have a lower ratio of jobs to resident workers than do white neighborhoods, although the findings are highly sensitive to the sizes and boundaries of geographical areas used for measurement (Holzer, 1991; Stoll and Covington, 2012). Better metrics have been developed for estimating job availability, including the number of jobs within a reasonable commuting radius of residential zones, spatially defined jobs to worker ratios, and gravity-based measures that incorporate the frictional effect of distance on job access. The plethora of metrics has yielded diverse and often contradictory findings, with some studies uncovering evidence of spatial mismatch, as reflected in the lower availability of job opportunities in and near predominately black neighborhoods (Ihlanfeldt and Sjoquist, 1998), and other studies finding little evidence of racial disparity (Hellerstein et al., 2008).

Researchers have also combined the power of geographic information systems with detailed urban databases to compute new measures of employment accessibility (Houston, 2005). Transportation routes and physical and social barriers to movement are layered together with the uneven geographies of job opportunities and racially segregated residential areas, to describe the multifaceted influences on spatial access to employment. Recent applications also incorporate spatial and temporal elements including public transit availability, the

timing of transit schedules, and time–space constraints linked to everyday work and child care commitments (Rogalsky, 2010).

Commuting times and flows have also been widely used in evaluating spatial mismatch. African-American workers have longer commute times on average than do comparable white workers, and the differences persist after controlling for wages, mode of transportation, and personal and household characteristics (McLafferty and Preston, 1996). Black workers are also more likely to have 'constrained' work trips that involve long travel times to low-wage jobs (Johnston-Anumonwo, 1997). The propensity to work in one's local neighborhood – local working – is also low in black neighborhoods (Immergluck, 1998). Spatial mismatch, then, is evident in long work trips to jobs that provide little economic return.

Although the time and cost of commuting are important in their own right for workers' employment outcomes, they have several limitations as measures of spatial mismatch. By definition, the measures are only based on a selected group, namely, employed workers. This 'selection bias' implies that commuting indicators may not accurately represent geographical barriers to employment for the unemployed (Cooke and Ross, 1999). One way to address this problem is to analyze changes through time as people and jobs move away from central city locations. The impacts of such moves on workers are complex and context dependent. Case studies reveal that when firms move to the suburbs, black workers typically are more likely to quit, more likely to endure longer commute times, and less likely to relocate close to the firm's new site than are comparable white workers (Fernandez, 2008; Ihlanfeldt and Sjoquist, 1998). Black workers are more constrained in their residential locations, indicating the long-lasting effects of racial segregation and discrimination.

There are also problems in interpreting commuting patterns. Do long work trips reflect individual choices to travel further for high-wage, high-quality employment, or do geographical constraints compel long work trips? Many commuting decisions result from a complex mix of choice and constraints, as people juggle personal and family needs within the spatial and social bounds of daily life (Hanson and Pratt, 1995). These choices and constraints vary between men and women, highlighting the importance of gender in the SMH (Blumenberg, 2004).

Attention is turning to the role of place in spatial mismatch and the important variations in spatial mismatch within and among cities and countries (Houston, 2005). In the United States, spatial mismatch is primarily a large-city phenomenon that is most evident in cities like New York, Chicago, Atlanta, and Los Angeles. It is also more apparent in the 'rustbelt' cities of the Northeast and Midwest where economic restructuring and suburbanization are in full force (Ihlanfeldt and Sjoquist, 1998). Less has been written about spatial mismatch in racially and ethnically diverse cities outside the United States, although there are signs of mismatch in cities like London, Johannesburg, Guangzhou, and Berlin where ethnic minority populations cluster in areas of limited job opportunity.

Spatial Mismatch and Labor Market Processes

Spatial mismatch is rooted in the social, economic, and spatial processes that shape urban labor markets. These influence where people and firms locate, the channeling of workers into certain types of occupations, workers' ability to search for and commute to job opportunities, and firms' willingness to hire workers (Gobillon et al., 2007). Connecting spatial mismatch to employment outcomes requires an understanding of labor market processes, as well as an understanding of the changing geographical patterns of firms and people.

Urban labor markets are strongly differentiated according to gender, race, ethnicity, and class. This labor market segmentation concentrates blacks and Latinos in a narrow set of occupations, and the particular occupations vary with gender. Historically, black men worked as laborers, in transportation, manufacturing, and certain service occupations, whereas black women were concentrated in domestic service. Although these occupations have changed over time, there continues to be strong segmentation by race and gender. Moreover, many ethnic minority groups exhibit similarly high levels of occupational segmentation. Segmentation implies that groups of workers living in the same neighborhood who are socially diverse face different sets of potential job opportunities. An important theme in spatial mismatch research has been to tease out the geographical linkages and barriers between race-, gender-, and ethnically defined patterns of residential segregation and occupationally segmented employment locations (Wyly, 1996).

Some labor market segmentation results from differences in education and skills that coincide with racial and ethnic divisions. The 'skills mismatch' refers to the gap between the education and skill levels of inner-city residents and the requirements of available jobs. The transformation from a goods-producing to a service-producing economy has caused a loss of low-skill, manufacturing jobs from central cities, while the emerging 'information economy' demands a highly educated workforce. Thus, central cities may be rich in high-skill jobs, but poor in low-skill jobs that fit the educational backgrounds of disadvantaged central city residents. Poverty, poor-quality schools, and social dislocations limit educational achievement in central cities, resulting in the skills mismatch. These spatial and skills mismatches position workers and jobs in intersecting social and geographical spaces to limit employment possibilities.

Transportation is a crucial factor in spatial mismatch, influencing the ease and cost of overcoming distance in traveling to work. American cities are built around the automobile, the most flexible mode of transportation and the one that offers access to the widest geographical range of job opportunities. In contrast, travel by public transit is slow and limited to fixed routes and schedules. For inner-city residents, the hub and spoke design of most mass transit systems are poorly structured for travel to suburban job centers. Because blacks and Latinos have low rates of car ownership and licensing, they are much more dependent on public transit than are their white counterparts. Transit dependence affects all aspects of employment including job search and access to suburban jobs, and it is the single most important determinant of minority workers' long commute times. Some argue that this 'automobile mismatch' is a more serious problem than the spatial mismatch (Grengs, 2010; Taylor and Ong, 1995).

The process of spatial job search is also important to the SMH. People often search for jobs near their homes. As distance

from the home increases, information about nearby job opportunities decreases and the costs of job search rise. Research indicates that blacks and Latinos' job search patterns are centered upon their homes, and they tend to search in fewer areas and areas with weaker employment growth than do comparable whites (Stoll and Raphael, 2000). Social networks play a key part in spatial job search. Racial and ethnic minorities rely heavily on family and friends for finding employment. For low-income, low-skill workers, this may lead to job search in areas of low job growth and poor-quality jobs, thus aggravating the spatial mismatch.

Employer decisions about whom to hire may also contribute to spatial mismatch. Some employers refuse to hire workers who live far from the work site, fearing that the long commute will lead to lateness or absenteeism (Gobillon et al., 2007). Suburban firms' hiring decisions may be biased along racial and geographical lines, reflecting stereotypes about the inability of inner-city minority workers to interact effectively with suburban customers.

Impacts of Spatial Mismatch

Spatial mismatch affects peoples' lives in diverse ways. The SMH emphasizes the effects on poverty, unemployment, and low incomes, but the empirical record is mixed. Perhaps the strongest support for the SMH comes from research on minority youth. For this group, unemployment is strongly correlated with the average travel time to the types of jobs that youth normally hold and youths' proximity to areas of employment growth (Ihlanfeldt, 1992). Minority youth who face long travel times or distances to job opportunities are much more likely to be unemployed than those who have opportunities nearby. Looking beyond youth, research among vulnerable populations such as welfare recipients reveals mixed findings. Some evidence shows that unemployment is highly sensitive to the concentration of low-wage jobs nearby (Ong and Blumenberg, 1998), whereas other evidence points to a lack of association (Sanchez et al., 2004). Although the impacts of spatial mismatch on labor market outcomes clearly vary from one context to another, an extensive review found general support for the SMH, concluding that "the lack of geographical access to employment is an important factor in explaining labor market outcomes" (Ihlanfeldt and Sjoquist, 1998: p. 881).

Others follow Ellwood (1986) in arguing that "race, not space" (p. 181) is the main determinant of poor labor market outcomes. Racial discrimination permeates urban labor markets in the United States, reducing minority employment irrespective of geographical access to jobs. Racial discrimination is often rooted in place, as firms avoid hiring workers from inner-city, minority communities, or use far-flung, racially biased social networks in attracting employees (Kasinitz and Rosenberg, 1996). Critics of the SMH point to examples where high unemployment and dense job opportunities exist side by side. The challenge in examining 'race versus space' is that the two are deeply intertwined: Race is space in the United States. Kain (1968) recognized this in the original formulation of the SMH, noting the tightly interconnected impacts of racial discrimination by employers and discrimination in housing markets on geographical access to employment. Attempts at disentangling the effects of race and space find that both have large and significant effects on employment outcomes (Hellerstein et al., 2008).

The implications of spatial mismatch beyond labor markets are discussed in Wilson's path-breaking work on the urban underclass (Wilson, 1987, 1996). Wilson starts with spatial mismatch and economic restructuring and weaves a rich tapestry that includes the effects on male unemployment, marriage, household structure, and the implications for an array of social issues such as crime, migration, local institutions, and educational achievement. As jobs move away and incomes decline, the ties that bind communities, households, and institutions start to disintegrate. Spatial mismatch is not the only cause of these changes, but it is an important piece of the puzzle.

Spatial Mismatch and Urban Policy

Despite differences of opinion about the SMH, it has played an important role in urban policy formulation for several decades. Policies to address spatial mismatch involve efforts to bring people and jobs closer together or to improve the connections between them. Three general types of policies have been suggested. First, 'economic development' policies aim at expanding job opportunities in and near the areas where disadvantaged minority populations live. These place-based policies include spatially targeted public works programs, and the provision of subsidies, training programs, and infrastructure, to attract new industries and jobs to inner-city areas. The policies go by various names – for example, 'empowerment zones' and 'New Markets' initiatives – and they have had limited success. Many failed to generate sustained, long-term employment because they did not alter the profitability of inner-city locations and benefits leaked out from inner-city neighborhoods to other areas.

Second, 'dispersal' policies aim at facilitating the movement of people to job-rich suburban areas. These strategies focus on enabling mobility in the housing market. Policies include the vigorous enforcement of fair housing laws, the development of public housing in suburban areas, and the provision of subsidies to help inner-city residents move to the suburbs. A good example of the latter is the Gautreaux program in Chicago, which offered housing subsidies to residents of public housing who wanted to move to the suburbs. Careful studies of the impacts from Gautreaux and similar relocation programs reveal little or no improvement in employment outcomes (Chapple, 2007)

Third, 'mobility' strategies emphasize improving inner-city residents' access to suburban job opportunities, rather than changing the locations of jobs or residences. Included here are improvements in public transit to facilitate reverse commuting, creation of car pools and van pools, employer-sponsored transportation from inner-city areas to job sites, and efforts to improve car ownership among the poor. Such policies have immediate and obvious benefits; however, they have never been implemented on a large scale. Also, the most effective mix of policies varies among cities

depending on the density of urban development and the availability of transportation infrastructure.

After decades of experimentation, most analysts agree on the need for multipronged approaches that are sensitive to the particularities of place and context. Furthermore, there is general agreement that spatial mismatch cannot be isolated from the constellation of forces that affect poverty and unemployment. Breaking down the barriers of racial discrimination in job and housing markets, and improving health, housing, and education in disadvantaged communities, could make spatial mismatch irrelevant in the long run.

Looking ahead, the historical specificity of the SMH highlights the fact that spatial mismatch is not a characteristic of particular locations but of particular places where people are disadvantaged based on a combination of race, ethnicity, class, and poor geographical access to work. Changing patterns of urban development, including the gentrification of inner-city neighborhoods and the growth of poverty in the suburbs, coupled with evolving patterns of employment opportunity, mean that spatial mismatch is perpetually recast. Poor and minority suburban residents are more likely than others to live in jobs-poor suburban communities, highlighting a new form of spatial mismatch (Raphael and Stoll, 2010). Today's suburbs may evolve into tomorrow's 'inner cities' as uneven urban development redefines spatial mismatch.

See also: Environmental Determinism; Labor Supply; Local Economic Development; Mismatch on the Labor Market; Social Stratification; Spatial Interaction; Transportation Geography; Transportation Planning; Urban Geography; Urban Growth Models; Urban Poverty in Neighborhoods.

Bibliography

Blumenberg, E., 2004. En-gendering effective planning - spatial mismatch, low-income women, and transportation policy. Journal of the American Planning Association 70, 269–281.

Chapple, K., 2007. Overcoming mismatch: beyond dispersal, mobility and development strategies. Journal of the American Planning Association 72, 322–336.

Cooke, T.J., Ross, S.L., 1999. Sample selection bias in models of commuting time. Urban Studies 36, 1599–1613.

Ellis, M., Wright, R., Parks, V., 2004. Work together, live apart? Geographies of racial and ethnic segregation at home and at work. Annals of the Association of American Geographers 94, 620–637.

Ellwood, D., 1986. The spatial mismatch hypothesis: are there teenage jobs missing in the ghetto? In: Freeman, R., Holzer, H. (Eds.), The Black Youth Employment Crisis. University of Chicago Press, Chicago, pp. 147–185.

Fernandez, R.M., 2008. Race, spatial mismatch, and job accessibility: evidence from a plant relocation. Social Science Research 37, 953–975.

Gobillon, L., Selod, H., Zenou, Y., 2007. The mechanisms of spatial mismatch. Urban Studies 44, 2401–2427.

Grengs, J., 2010. Job accessibility and the modal mismatch in Detroit. Journal of Transport Geography 18, 42–54.

Hanson, S., Pratt, G., 1995. Gender, Work, and Space. Routledge, New York.

Hellerstein, J.K., Neumark, D., McInerney, M., 2008. Spatial mismatch or racial mismatch? Journal of Urban Economics 64, 464–479.

Holzer, H.J., 1991. The spatial mismatch hypothesis: what has the evidence shown? Urban Studies 28, 105–122.

Houston, D.S., 2005. Methods to test the spatial mismatch hypothesis. Economic Geography 81, 407–434.

Ihlanfeldt, K.R., 1992. Job Accessibility and the Employment and School Enrollment of Teenagers. W. E. Upjohn, Institute for Employment Research, Kalamazoo, MI.

Ihlanfeldt, K.R., Sjoquist, D.V., 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. Housing Policy Debate 9, 849–892.

Immergluck, D., 1998. Job proximity and the urban employment problem: do suitable nearby jobs improve neighborhood employment rates? Urban Studies 35, 7–23.

Johnston-Anumonwo, I., 1997. Gender, race and constrained work trips in Buffalo, 1990. Professional Geographer 49, 306–317.

Kain, J., 1968. Housing segregation, negro employment, and metropolitan decentralization. The Quarterly Journal of Economics 82, 175–197.

Kasinitz, P., Rosenberg, J., 1996. Missing the connection: social isolation and employment on the Brooklyn waterfront. Social Problems 43, 180–196.

Liu, C.Y., Painter, G., 2012. Immigrant settlement and employment suburbanization in the U.S.: is there a spatial mismatch? Urban Studies 49, 979–1002.

McLafferty, S., Preston, V., 1996. Spatial mismatch and employment in a decade of restructuring. Professional Geographer 48, 420–431.

Ong, P., Blumenberg, E., 1998. Job access, commute and travel burden among welfare recipients. Urban Studies 35, 77–93.

Raphael, S., Stoll, M.A., 2010. Job Sprawl and the Suburbanization of Poverty. Metropolitan Opportunity Series. Brookings Institution, Washington, DC.

Rogalsky, J., 2010. The working poor and what GIS reveals about the possibilities of public transit. Journal of Transport Geography 18, 226–237.

Sanchez, T.W., Shen, Q., Peng, Z., 2004. Transit mobility, jobs access and low-income labour participation in US metropolitan areas. Urban Studies 41, 1313–1331.

Stoll, M.A., 1999. Race, Space and Youth Labor Markets. Garland Publishing, New York.

Stoll, M.A., Covington, K., 2012. Explaining racial/ethnic gaps in spatial mismatch in the U.S.: the primacy of racial segregation. Urban Studies 49, 2501–2521.

Stoll, M.A., Raphael, S., 2000. Racial differences in spatial job search: exploring the causes and consequences. Economic Geography 76, 201–223.

Taylor, B.D., Ong, P.M., 1995. Spatial mismatch or automobile mismatch? an examination of race, residence and commuting in US metropolitan areas. Urban Studies 32, 1453–1473.

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Knopf, New York.

Wyly, E.K., 1996. Race, gender, and spatial segmentation in the Twin Cities. Professional Geographer 48, 431–445.

Spatial Navigation

Thomas Wolbers, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial navigation is a fascinating behavior that is essential for our everyday lives. It involves nearly all sensory systems, requires numerous parallel computations, and engages multiple memory systems. This article will present our current understanding of the cognitive processes required for successful navigation and the neuronal systems that support them. In addition, the final section will discuss how navigational abilities change as we get older, a particularly important topic for our aging societies.

Introduction

'The manner in which the sense of direction is sometimes suddenly disarranged in very old and feeble persons, and the feeling of strong distress which, as I know, has been experienced by persons when they have suddenly found out that they have been proceeding in a wholly unexpected and wrong direction, leads to the suspicion that some part of the brain is specialized for the function of direction' (Darwin, 1873).

Being able to maintain a sense of direction as we move about in our cluttered, everyday environments is a fundamental cognitive function that we often take for granted. Mammals rely on spatial cognitive processes for obtaining food, avoiding prey and finding mates. In humans, spatial navigation is indispensable for finding our way in complex environments, planning routes to distant locations, and returning to our hotel after a walk in a new city. As a consequence, when gradually progressing diseases (i.e., Alzheimer's diseases) or sudden lesions to the brain impair navigational abilities, patients often experience devastating effects on their everyday lives (Aguirre and D'Esposito, 1999). In contrast to Darwin's assumption of a circumscribed part of the brain, however, a widespread cortical and subcortical network supports navigational processes. This distributed instantiation is caused by the complexity of navigational functions (Wolbers and Hegarty, 2010, **Figure 1**):

- Spatial navigation is a multimodal process: spatial information relevant for navigation is usually derived in parallel from a variety of sensory cues, including cues provided by the environment (i.e., visual landmarks) and body-based self-motion cues. In addition, symbolic representations such as maps, signs, or verbal route descriptions can also convey spatial information.
- Spatial navigation involves parallel computations: incoming spatial information is not only used to extract various primitives such as position, facing direction, and the position of external objects, but these parameters are computed in multiple reference frames. In addition, executive mechanisms such as route planning, goal selection, etc. also contribute to guiding our moment-to-moment navigational decisions.
- Spatial navigation recruits parallel memory systems: online representations of position and facing direction, etc. interact with long-term memory representations of local views and the overall structure of an environment. In addition, more procedural memories in the form of habitual routes also guide navigation, in particular when we follow particular paths on a frequent basis.

These factors are highly interdependent (i.e., stored memories of an environment can affect sensory experience), and many navigational tasks bridge multiple levels (i.e., path integration involves perceiving self-motion cues and forming a spatial representation). In this article, I will discuss our current understanding of spatial navigation from a behavioral and from a neuroscientific perspective. As key navigational systems in the human brain show profound changes as we get older, the final section will provide a brief discussion of the decline in navigational abilities experienced by the elderly.

Navigation Based on Transient Spatial Knowledge

When we move about in unfamiliar environments such as a city we visit for the first time, avoiding getting lost is a key challenge. To address this challenge, our brain constantly computes information about our location and facing direction and about the position of important objects, even if they can no longer be seen (i.e., our hotel). While these computations are often based on cues provided by the environment (i.e., a salient building), our bodies also register information about the direction in which we have turned or a distance we have just walked. Collectively, these processes provide transient spatial knowledge that can be used for planning routes, even though they also support the formation of long-term spatial memories.

Path Integration and Spatial Updating

When we move about in an environment, one way to keep track of changes in position and orientation is to continuously monitor our body movements, a process known as path integration (Mittelstaedt and Mittelstaedt, 1980). Path integration has been studied extensively in various animal species, some of which can successfully navigate by path integration over large distances (Etienne and Jeffery, 2004). In humans, the most frequently used task to assess path integration abilities is the triangle completion task, which involves moving participants

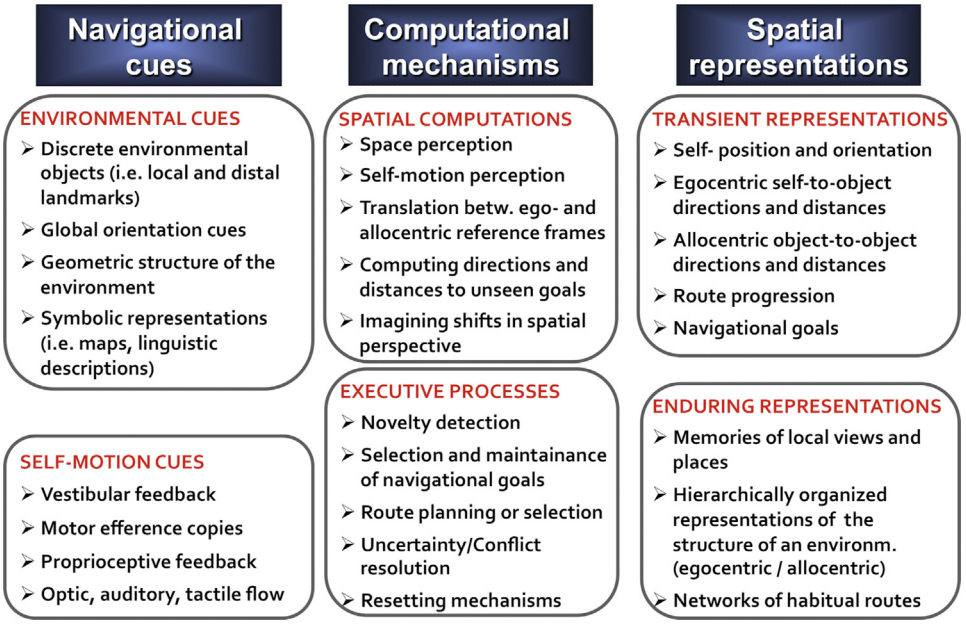

Figure 1 The complexity of spatial navigation. Schematic depiction of the sensory cues, computational mechanisms, and spatial representations that are involved in forming spatial knowledge and in using it to guide navigational behavior. Figure modified and reproduced, with permission, from Wolbers, T., Hegarty, M., 2010. What determines our navigational abilities? Trends in Cognitive Science 14, 138–146.

along the first two sides of a triangular pathway, and then asking them to return to their starting location on a direct path, thus completing the triangle (Figure 2). Participants are usually deprived of visual and auditory information, demonstrating that path integration is possible based on body-based cues only. These cues involve motor efference copies, proprioceptive feedback, and vestibular feedback about rotational and translational acceleration. However, path integration can also be based on visual self-motion cues as derived from optic flow patterns (Kearns et al., 2002).

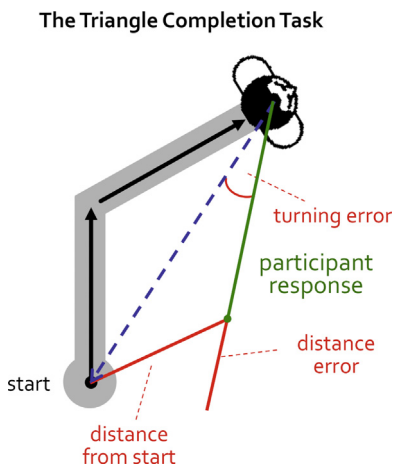

Figure 2 The triangle completion task. Participants are guided along two legs of an imaginary triangle before having to return to the start position – on a direct path – on their own. The dashed blue line shows the correct response, the green line an actual response of a participant. Various error measures can be used to quantify path integration performance, i.e., the absolute distance from the starting location, the angular deviation from the correct turning angle or the deviation of the covered distance from the correct distance.

In contrast to many animal species, path integration in humans is only reliable within shorter distances, because errors quickly accumulate, thus producing increasing uncertainty about one's location as the outbound path gets longer. What mechanisms could explain these path integration errors? First of all, seminal work by Loomis and coworkers (Fujita et al., 1993; Loomis et al., 1993) on blindfolded path integration showed that errors predominantly arise from an inaccurate encoding of the outbound path and not from imprecise execution of the homing trajectory. Secondly, path integration performance can be successfully captured by a leaky integration model (Lappe and Frenz, 2009; Lappe et al., 2007, 2011), in which homing errors are driven by two factors: inaccurate processing of self-motion cues (i.e., perceiving a displacement of 1 m as 90 cm) and an inaccurate updating of the positional estimate. Importantly, the latter is dependent on the current position – the larger the distance away from home, the more inaccurate the updating of the positional estimate will be. Thirdly, path integration errors are also driven by biases such as expectations about upcoming trajectories (Petzschner and Glasauer, 2011). An intriguing example of bias is that people perceive self-rotations as being closer to angles of 0°, 90°, 180° than they actually are (Saddala and Montello, 1989), which has been related to the organization of the body into forward–backward and left–right axes. And finally, humans can employ different strategies to solve path integration tasks, which results in differential patterns of error (Wiener et al., 2011): on the one hand, they might try to continuously update the homing vector during

the outbound path; on the other hand, they might store the outbound path as a series of translations and rotations, thus only computing the homing vector at the end of the trajectory. The latter strategy is less vulnerable to error accumulation as estimates for any given segment will be independent of all other segments, thus resulting on smaller homing errors (Wiener et al., 2011).

While path integration is a mechanism to keep track of one's position relative to a location an observer has physically occupied – i.e., the starting location of an outbound path – spatial updating refers to our ability to update the location of external objects. In other words, spatial updating operates on positional codes that have been derived from (visual) perception and that were not occupied by an observer. Constantly knowing where one is relative to surrounding objects is a central feature of situational awareness and readiness to act. For example, when objects go out of view, when people walk without vision in the dark or when, like pilots, firefighters, and deepwater divers, they operate in situations that are rendered visually unreliable and impoverished, efficient behavior crucially depends on accurate spatial updating.

Similar to path integration, spatial updating relies on the integration of body-based and external self-motion cues (Creem-Regehr, 2004). Humans can easily point to targets from a new position after making a turn while blindfolded, even if they have never seen the objects from this new perspective (Farrell and Robertson, 1998). Importantly, spatial updating occurs largely automatically both in real and in virtual environments (Farrell and Robertson, 1998; Riecke et al., 2007; and Wang, 2004). For example, when asked to ignore their movements and to point to the targets as if they had not moved, people show significant costs in response time and accuracy, which suggests that they have great difficulty in voluntarily preventing spatial updating from occurring.

Most experimental studies on spatial updating have used setups with relatively small numbers of objects, making a continuous updating of multiple object vectors a computationally tractable problem. However, additional strategies are likely to come into play when we have to deal with the complexity of our everyday environments. One such strategy is chunking, in which adjacent objects are grouped together and only one prototypical vector per chunk is updated during self-motion (Sargent et al., 2010). Alternatively, when moving about in a complex but familiar environment, humans tend to only update their own position and orientation – using path integration – before reconstructing object positions from the new vantage point (Waller and Hodgson, 2006).

Brain Systems Computing Transient Spatial Information

Our knowledge of the neural systems computing instantaneous spatial information has been fueled by several important discoveries from electrophysiological recordings in freely moving rodents. A key structure for self-localization is the hippocampus, whose principal neurons, the place cells, fire selectively in specific regions of the environment known as 'place fields' (O'keefe and Dostrovsky, 1971) (**Figure 3**). Place cells collectively provide a representation of position, and they receive information both about self-motion and stable environmental cues. For example, at least in rodents, proximal environmental boundaries primarily specify the location of place fields (O'keefe and Burgess, 1996), whereas distant landmarks are important for directional orientation (Jeffery et al., 1997).

Place cells, however, do not provide an absolute spatial signal, as their firing is also modulated by various nonspatial factors such as motivation, goals, etc. As a consequence, it was often speculated that positional information must be computed upstream of the hippocampus, and the discovery of grid cells in entorhinal cortex has confirmed this hypothesis (Hafting et al., 2005). Grid cells are directly afferent to hippocampal place cells, and they exhibit multiple, evenly spaced place fields. The precise position of an animal can be reconstructed from a population of grid cells, and many researchers assume that the unique place fields in the hippocampus emerge from the integration of grid cells with different spatial frequencies (Moser and Moser, 2008). Importantly, in monkeys, grid cell firing has been observed when the animal was merely looking at different spatial locations (Killian et al., 2012), suggesting that in primates – who are equipped with more developed visual systems – computing positional knowledge may not require physical locomotion. Finally, to be able to continuously update positional representations, grid cells must receive information about the direction in which an organism moves. This directional orientation is computed by the head direction system, a network of limbic structures whose neurons, the head direction cells, show directionally specific firing (Taube et al., 1990a) (**Figure 3**).

While the firing of place, grid, and head direction cells is strongly influenced by stable environmental cues such as boundaries and distal landmarks, they also support navigation based on path integration. For example, the periodic firing of grid cells and the directional firing of head direction cells persist even when all visual cues are removed and the animal navigates in total darkness (Hafting et al., 2005; Taube et al., 1990b). Moreover, neural activity in the head direction circuit is closely related to rats' homing direction (Van Der Meer et al., 2010; Valerio and Taube, 2012), and lesions to entorhinal cortex severely disrupts homing performance (Parron and Save, 2004). In humans, corresponding findings from lesion and neuroimaging studies also support the notion that structures in the medial temporal lobe, which contain grid, place, and head directions cells, are involved in path integration (Philbeck et al., 2004; Wolbers et al., 2007; Worsley et al., 2001).

In contrast to the wealth of studies on path integration, studies on the neural systems mediating spatial updating are scarce. Given that subregions of the posterior parietal cortex (PPC) code for spatial location in egocentric reference frames and receive multimodal self-motion cues, theoretical models have predicted spatial updating to be mediated by PPC structures such as the precuneus (Byrne et al., 2007). Lesions to the PPC, however, do not always impair spatial updating performance (Farrell and Robertson, 2000; Philbeck et al., 2000, 2001), but the conflicting findings of these studies may partially stem from small sample sizes and the largely variable lesion locations. More conclusive evidence has been provided by neuroimaging studies that required subjects to update object locations using optic flow (Wolbers et al., 2008

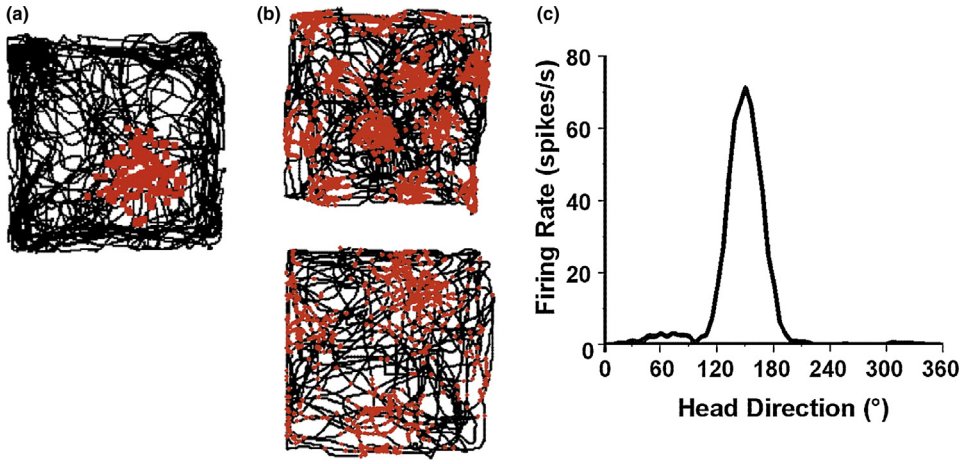

Figure 3 Place, grid, and head direction cells in rodents. (a) A classical place cell. The black lines show the path of a rat as it foraged in a 60 cm square box, and the red squares show action potentials from a single hippocampal neuron, superimposed on the position of the rat. A typical place cell will predominantly fire action potentials in one part of a small environment (the 'place field'). (b) Grid cells from dorsal (upper panel) and ventral (lower panel) dorsomedial entorhinal cortex. The figure shows the hexagonal array of firing fields typical of grid cells. Note the change in interpeak spacing between dorsal and ventral regions. (c) Data from a typical head direction cell, showing that firing rate (*y* axis) increases dramatically when the rat's head is within a restricted directional orientation (*x* axis). Figure reproduced, with permission, from Jeffery, K.J., 2007. Self-localization and the entorhinal-hippocampal system. Current Opinion in Neurobiology 17, 684–691.

(see also **Figure 1**); Jahn et al., 2012). These studies have consistently identified the precuneus as a structure that combines stored representations of external locations with incoming self-motion cues, thus providing moment-to-moment estimates of spatial locations. Importantly, the precuneus supports spatial updating independent of how we plan to interact with our environments, while the dorsal premotor cortex appears to use this information to plan upcoming motor actions (Wolbers et al., 2008) (**Figure 4**).

Navigation Based on Enduring Spatial Knowledge

Although path integration and spatial updating are often discussed as navigational mechanisms, both processes are also integral for the acquisition of enduring representations of space. When we become familiar with an environment, we can navigate to distant goals that involve the recognition of several places and a representation of spatial relations between objects outside the current sensory horizon. Collectively, these processes are often referred to as wayfinding behavior, a multifaceted construct that involves at least two different navigational mechanisms.

Route Knowledge and Cognitive Maps

Wayfinding behavior can be based on route knowledge and on cognitive maps. Route knowledge is often conceptualized as a goal-directed representation of a fixed sequence of landmarks, each of which is associated with a specific movement decision (Franz and Mallot, 2000; Mallot and Gillner, 2000). Navigation based on route knowledge proceeds by approaching environmental landmarks and by then executing the associated movement decisions (i.e., turn right at the church). Route knowledge has often been reduced to a mere sequence of stimulus–response pairs that, in effect, is 'empty' between landmarks (Siegel and White, 1975). It is often assumed to neither contain off-route information nor quantitative properties like Euclidean distances between landmarks (Golledge, 1999). However, because views of landmarks can prime upcoming landmarks, route knowledge also contains information about the temporal sequence of landmarks, but only in the direction in which the route was learned (Schweizer et al., 1998).

It is important to note that route knowledge not always provides associations between landmarks and movement decisions. Depending on the configuration of the environment, landmarks can also serve as beacons. A beacon specifies a goal location without being associated with a particular directional movement (i.e., go to the church or turn toward the post office). In this case, route knowledge is reduced to a mere sequence of beacons (i.e., first go to the church, then to the post office, etc.), which is a parsimonious knowledge structure. As a consequence, while routes are learned better when their landmarks serve as beacons than as associative cues, beacon-based route knowledge is less enduring and less likely to involve knowledge of directions in the environment than is the route knowledge formed from landmarks that serve as associative cues (Waller and Lippa, 2007).

Cognitive maps – often termed survey knowledge or configurational knowledge – constitute the most elaborate and flexible form of spatial representations. These representations embed the objects in an environment and their spatial relations into a common reference frame that is independent of the observer. Moreover, survey knowledge is thought to contain quantitative information about distances and angular directions. The term cognitive map was first coined by Tolman (1948) to express that rats in his experiments took novel shortcuts to a goal location, thus indicating an understanding of the spatial structure of the environment. The term suggests – and is often used to suggest – that we

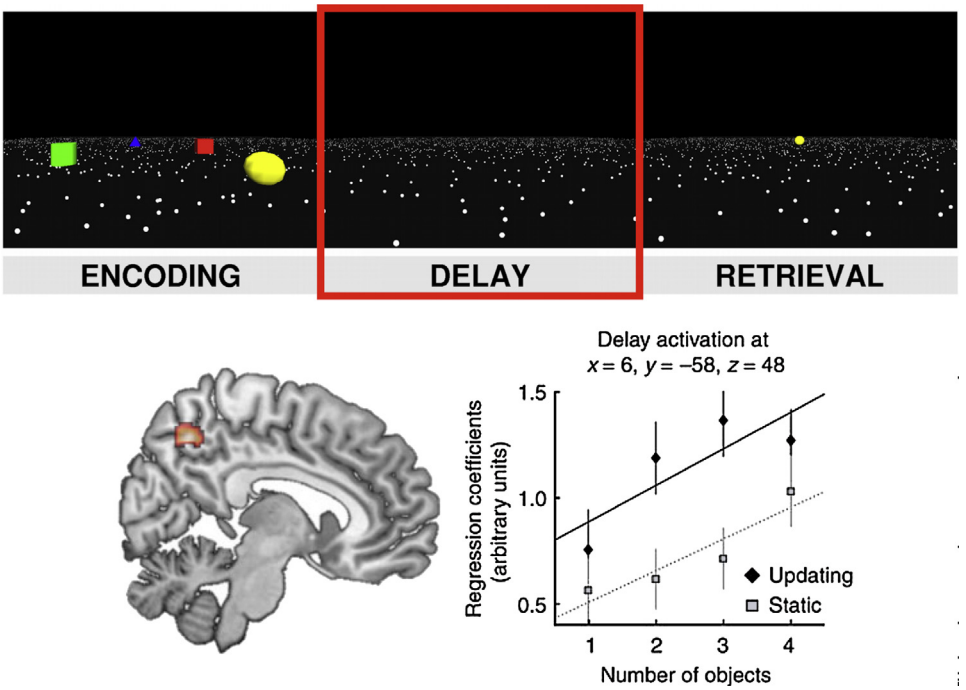

Figure 4 Neural mechanisms of spatial updating. Upper panels: in the study by Wolbers et al. (2008), participants underwent fMRI scanning while performing a spatial updating task in a virtual environment. The task required participants to first memorize the locations of 1–4 virtual objects that subsequently disappeared. In the ensuing delay phase, subjects either remained stationary (static trials) or they experienced a forward translation of a random distance. At the end of the trajectory, participants had to point with a joystick to one of the objects shown before. Lower panels: during the delay phase, only the precuneus showed activation profiles that were modulated (a) by the number of objects that had memorized and (b) by the presence of self-motion cues (updating). These data suggest that the precuneus uses incoming self-motion cues to continuously update egocentric object positions. Figure modified and reproduced, with permission, from Wolbers, T., Hegarty, M., Büchel, C., Loomis, J.M., 2008. Spatial updating: how the brain keeps track of changing object locations during observer motion. Nature Neuroscience 11, 1223–1230.

possess spatial representations similar to a real map, as if the world was represented from a bird's eye perspective. However, there is currently no direct evidence to support this claim, and many researchers (i.e., Thorndyke and Hayes-Roth, 1982) prefer to think of cognitive maps as 'see through representations' that preserve the ground-level perspective but in which distant locations can be accessed through intervening points (i.e., buildings).

A key feature that distinguishes cognitive maps from route-based navigation is that only the former enable flexible navigational behavior: whereas route-based navigation is only effective along the represented paths, survey knowledge can be used to find novel paths over unknown terrain and between unconnected routes, to take shortcuts and to point to unseen landmarks. According to influential models of spatial learning, both types of knowledge develop in a series of sequential qualitative transitions, from initial landmark knowledge to representations of routes and finally to flexible survey knowledge (Hart and Moore, 1973; Siegel and White, 1975). Landmarks are conceptualized as discrete objects or scenes that are stored in memory and recognized when perceived. Other than the local spatial information that is present in a scene, landmarks do not contain any spatial information. Once acquired, landmarks are then associated movement decisions and connected into chains that preserve the order experienced during navigation. These chains are the constituents of route knowledge, the second stage of spatial learning. The integration of different routes into a single representation along with metric information constitutes the final step, the formation of survey knowledge.

Recent evidence casts doubt on this strictly hierarchical conceptualization of environmental learning. For example, Ishikawa and Montello (2006) had subjects learn an environment while being driven in a car. The authors found that whereas some participants failed to develop survey knowledge even after 10 learning sessions, others had acquired at least partial metric knowledge already after the first exposure to the environment. More importantly, metric judgments about route distances were often better than when subjects had to rank order the route distances. This inferior ordinal accuracy suggests that forming a precise route representation is not a necessary precondition for the development of survey knowledge. In addition, the ability of humans and various animals to perform path integration tasks indicates that some metric knowledge about turning angles and the lengths of path segments is already present after single trial learning. As these empirical results are difficult to reconcile with hierarchical concepts, alternative models of spatial learning argue for a potential to acquire survey knowledge from the very beginning of a learning experience (Montello, 1998). In addition, the evolution of landmark, route, and survey knowledge might rather proceed

in parallel, although the latter seems to only emerge in subjects with a good sense of direction (Ishikawa and Montello, 2006).

Brain Systems Mediating Route Knowledge and Cognitive Mapping

Given that route knowledge is usually conceptualized as chains of associations between local views of an environment and specific motor responses, it comes as no surprise that brain structures classically associated with scene processing (parahippocampal place area (PPA)) on the one hand and with procedural learning (striatum) on the other hand have also been implicated in route learning. For example, objects encountered at intersections, which require a decision about where to go next, elicit larger activation in the PPA than objects encountered along the way (Janzen and Van Turennout, 2004). In addition, neuroimaging studies have shown that activation in the caudate nucleus is predictive of route retrieval accuracy (Hartley et al., 2003), and subjects who use a response-based strategy showed larger caudate activation than subjects who adopt a cognitive mapping strategy (Iaria et al., 2003). Interestingly, when the caudate is dysfunctional (i.e., as in Huntington's disease), compensatory activation in the hippocampus still allows for successful route negotiation (Voermans et al., 2004).

These studies, however, do not distinguish between routes learned as stimulus–response associations versus a series of beacons that do not need to be associated with specific movement decisions. Anatomically, the striatum is not a unitary system, but there exists a functional dissociation between the medial/ventral and the lateral/dorsal striatum (Voorn et al., 2004). The ventral and dorsomedial striatum receives extensive convergent input from multiple sensory and association areas of neocortex and the limbic system, while the dorsal and dorsolateral striatum has a distinct topographic pattern of connections with somatosensory and motor areas of neocortex (McGeorge and Faull, 1989). These findings could indicate that the dorsal and dorsolateral striatum – which comprises the caudate nucleus – is more important for motor response learning as required for stimulus–response associations, whereas the dorsomedial and ventral striatum could be more important for beacon strategies. Indeed, lesion studies in rodents implicate the dorsolateral striatum in stimulus response learning (Packard and McGaugh, 1996). Moreover, there is evidence that the dorsomedial/ventral striatum does not affect stimulus response learning to the same extent as dorsal/dorsolateral lesions (Yin and Knowlton, 2004). Finally, dorsomedial lesions affect the learning of a beacon strategy in the Morris Water Maze (MWM), whereas dorsolateral lesions do not (Devan and White, 1999).

Ever since the discovery of place cells, the hippocampus has been hypothesized to be the neural system that supports flexible, cognitive map based navigation. Specifically, as place cells receive input from grid cells about spatial position along with information about objects, they might store the locations of environmental objects in an observer-independent, allocentric reference system. Consistent with this assumption, rats with hippocampal lesions show profound deficits in the MWM when they need to swim to a location defined by external landmarks (Morris et al., 1982). Even though virtual renditions of the MWM have revealed similar deficits in humans with hippocampal lesions or degeneration (Hort et al., 2007; Goodrich-Hunsaker et al., 2010), it is important to note that the MWM only measures spatial knowledge of a local place. However, a key feature of cognitive maps is that they allow for inferring spatial relations between distant locations that often cannot be simultaneously perceived from a single vantage point, hence the MWM may only be of limited value for investigating large-scale cognitive mapping.

In humans, neuroimaging and lesion studies conducted in large virtual environments have consistently shown that the posterior hippocampus is implicated in retrieving knowledge about familiar environments, for example, when planning and navigating novel routes (Hartley et al., 2003; Iaria et al., 2007). In addition, hippocampal representations appear to contain metric knowledge about distances and directions, strongly suggesting the existence of quantitative cognitive maps in the place cell network (Morgan et al., 2011; Spiers and Maguire, 2007). In line with the plasticity of many cortical networks, the posterior hippocampus is enlarged in expert navigators such as London taxi drivers (Maguire et al., 2006). Crucially, this effect was not observed in London bus drivers who experience similar amounts of stress, self-motion, etc. but who always travel along habitual routes (Woollett and Maguire, 2011). The extensive use of elaborate cognitive maps for flexible route planning in the taxi drivers thus is likely to drive structural changes in the posterior hippocampus. However, given that patients with complete hippocampal lesions can retrieve cognitive maps from environments learned long ago (Teng and Squire, 1999), the hippocampus may only play a time-limited role in storing and retrieving cognitive maps.

The enlargement of the posterior hippocampi in London taxi drivers came at the expense of a reduced anterior hippocampal volume, which was paralleled by a reduced ability to learn novel spatial information. This findings point to a dissociation between the anterior and posterior hippocampus, with the former being mainly responsible for the learning process. To test this hypothesis, Wolbers et al. (2005) had subjects learn the layout of an unfamiliar environment from first-person navigation while undergoing fMRI scanning (**Figure 5**). The most important finding of that study was that activation in the anterior hippocampus predicted the amount of learning: sessions with strong anterior hippocampal activation led to substantial performance improvements, whereas in sessions with low activation, performance hardly changed. Together with similar findings from other groups (Iaria et al., 2007), these results indicate that the anterior hippocampus contributes to spatial navigation by integrating and consolidating novel spatial information such as the relative position of environmental landmarks into observer-independent cognitive maps.

Two further brain structures often implicated in cognitive mapping are the PPA and retrosplenial cortex (RSC). While the PPA has consistently been associated with processing the geometric structure of local scenes in a viewpoint-dependent fashion (Epstein, 2008), the role of RSC is less clear. In rodents, RSC contains head direction cells that provide a compasslike signal by continuously tracking an animal's

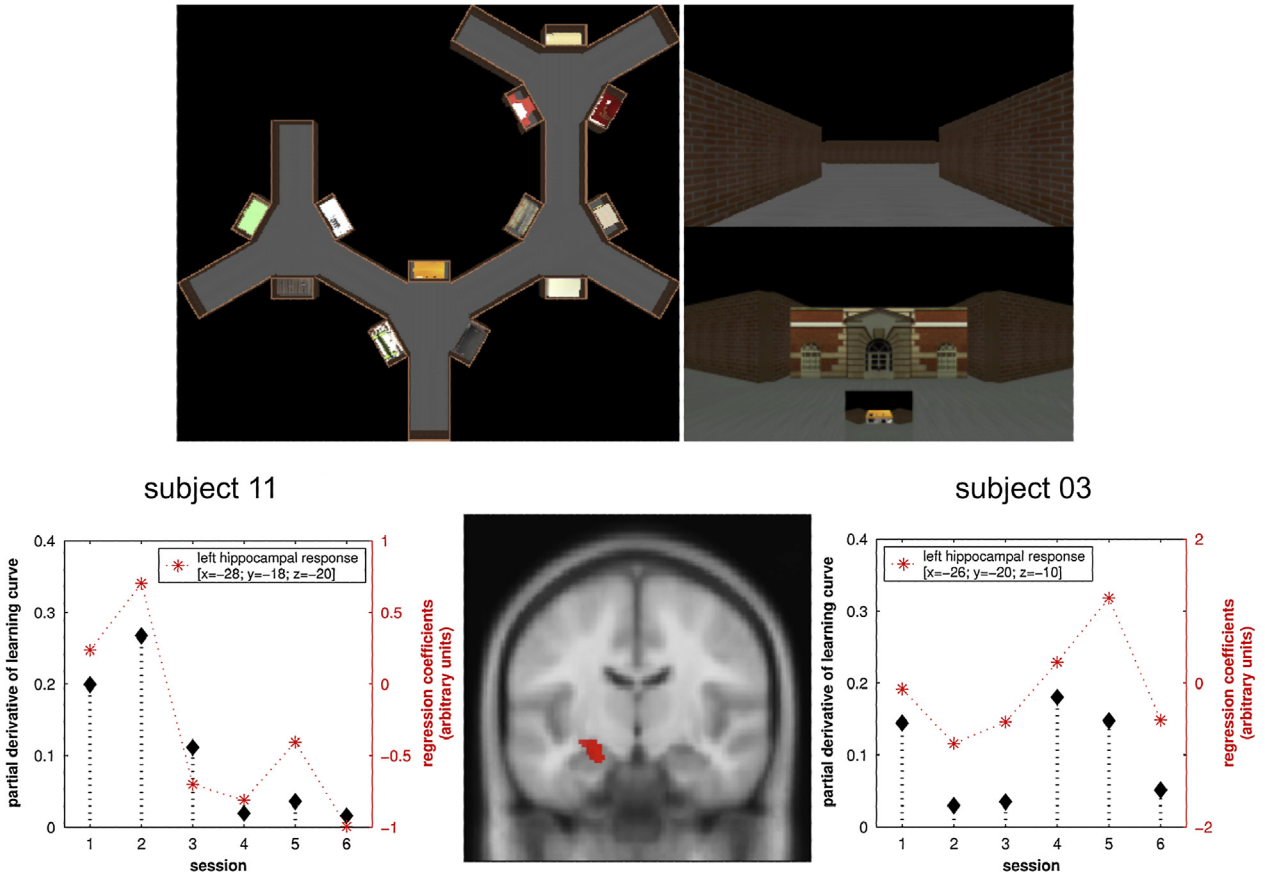

Figure 5 Learning cognitive maps. Upper panels: in the study of Wolbers and Büchel (2005), participants were repeatedly moved throughout a complex virtual environment, thereby encountering 12 distinct landmarks. Subjects were instructed to infer the spatial layout of the environment and the correct landmark locations, knowledge that was tested with a subsequent retrieval task. Lower panels: anterior hippocampal recruitment during cognitive map formation predicts the speed of learning. Hippocampal activation (shown in red) in subject 11 was strongest in the initial learning stage and decayed rapidly after performance had reached ceiling level in session 3. In contrast, the slower learning process in subject 03 was paralleled by stronger hippocampal activation in the second half of the experiment. As a consequence, anterior hippocampal activation appears to be most prominent whenever substantial performance improvements are observable. Figure reproduced, with permission, from Wolbers, T., Büchel, C., 2005. Dissociable retrosplenial and hippocampal contributions to successful formation of survey representations. Journal of Neuroscience 25, 3333–3340.

orientation (Cho and Sharp, 2001). Similarly, human patients with RSC lesions can identify familiar places and scenes, but they cannot use them for orientation (Ino et al., 2007; Maguire, 2001). However, a direct demonstration of head direction signals in human RSC is missing at present. Moreover, while some authors emphasize the RSC's role in forming memories of spatial information (Cooper et al., 2001), others assume that any computation that involves translation between ego- and allocentric reference frames requires RSC functioning (Byrne et al., 2007). Importantly, such computations are also needed to form enduring representations of complex environments as shown by neuroimaging studies (Wolbers and Büchel, 2005).

Aging of Navigational Abilities

The world's population is aging at an unprecedented rate. The number of people aged over 60 is set to rise from 784 million in 2011 to 2 billion by 2050 (United Nations, 2012), and a significant proportion will have to cope with normal or pathological cognitive decline. Despite huge efforts to understand the mechanisms of age-related cognitive decline, the number of studies looking at navigational abilities is surprisingly scarce. This is all the more surprising when we consider the following facts:

- in surveys, elderly people often report substantial declines in navigational abilities (Burns, 1999), for example, problems with finding their way in a novel environment,
- navigational difficulties can have a huge impact on the quality of life of the elderly, for example, by affecting social participation and physical activity,
- key brain structures supporting navigational functions such as entorhinal cortex, the hippocampus, and RSC show the earliest signs of age-related neurodegeneration, and
- age-related navigational impairments have been well characterized in nonhuman species, making it easy to generate testable predictions for studies with human participants.

In the aging brain, layer II of entorhinal cortex, which appears to contain the highest proportion of grid cells both in rodents and primates (Hafting et al., 2005; Killian et al., 2012),

is particularly sensitive to the deleterious consequences of aging (Stranahan and Mattson, 2010). Specifically, neurodegenerative changes such as neuron loss occur in the earliest stages of age-related cognitive decline and Alzheimer's disease; hence, disrupted grid cell computations could be among the earliest functional changes in the brain, which should have direct effects on path integration. Indeed, aging humans show profound deficits during walking without vision and in optic flow-based path integration tasks (Allen et al., 2004; Mahmood et al., 2009), and they are impaired at heading perception in virtual environments (Mapstone et al., 2006). Getting closer at the mechanisms that might explain such path integration deficits, older adults have been shown to exhibit a significantly reduced range of responses when asked to reproduce turns and distances and to complete triangular pathways (Harris and Wolbers, 2012). These results can be interpreted in the context of the leaky integrator model (Lappe et al., 2007), suggesting that due to an increased leak in the integration process, stored self-motion information decays more rapidly in old age.

In addition to path integration deficits, aging also affects the formation and use of cognitive maps. Aged rodents show deficits in tasks in which spatial relationships must be computed between environmental cues, while tasks requiring egocentric cues are less affected (Begega et al., 2001). Moreover, even when successfully trained on both place and response strategies, aged mice still show a spontaneous preference for egocentric response strategies, strongly suggesting an age-related increase in reliance on extrahippocampal systems to guide navigation (Nicolle et al., 2003). Such age-related navigational impairments are likely related to the age-related changes of the hippocampal place cell system. Specifically, the spatially localized firing of aged place cells has been shown to take longer to emerge, to be less stable across repeated visits to a familiar environment, and to be less capable of differentiating between familiar and novel environments (Lister and Barnes, 2009).

Similar to rodent findings, difficulties with employing cognitive map based strategies have been observed repeatedly in older humans (Moffat and Resnick, 2002; Moffat et al., 2007; Iaria et al., 2009; and Wiener et al., 2012). For example, older adults tend to disregard distal geometric information in an environment and focus more on local objects to guide navigation (Moffat and Resnick, 2002). Moreover, in the plus maze task, which can be solved with a striatum dependent egocentric response (i.e., go right) or a hippocampus dependent allocentric place (i.e., go east) strategy, older adults show specific deficits in switching to the place strategy (Harris et al., 2012, **Figure 6**). Consistent with these behavioral findings, fMRI studies with virtual renditions of the MWM have shown reduced activation of the medial temporal lobe in old age (Antonova et al., 2009; Moffat et al., 2006). Such findings, however, need to be cautiously interpreted, because an overall reduction of the Blood Oxygen Level Dependent (BOLD) signal may be due to a range of factors (i.e., hemodynamic changes in old age) that may or may not reflect a change in navigational processing.

Given the spontaneous preference for egocentric response strategies in aged rodents and humans, one may expect route-based navigation to be relatively preserved in old age. However, structural neuroimaging studies have shown that the caudate nucleus, a key structure for stimulus–response associations (see previous section), shows age-related neurodegeneration at similar rates compared to the hippocampus (Raz et al., 2005). As a consequence, it comes as no surprise that studies on route learning have revealed deficits in remembering associations between landmarks and directional information (Liu et al., 2011; Head and Isom, 2010; and Wilkniss et al., 1997). For example, Wilkniss et al. (1997) had participants navigate through the hallways of a hospital after being shown a map of the environment and the route they were required to follow. Overall, older adults took longer to navigate through the hospital than younger adults, and they made more frequent turning errors. In contrast, older adults were not impaired at remembering the landmarks encountered along the way, suggesting that simple landmark knowledge might still be useful for navigation.

As stated before, route knowledge can either be based on associations between landmarks and movement decisions, or it can involve a sequence of landmarks serving as beacons, thus not requiring the formation of links with particular movement decisions. Rodent studies suggest that the ventral striatum may be particularly important for beacon-based navigation, and anatomical data have demonstrated that the dorsal striatum may be more vulnerable to age-associated changes than its ventral counterpart (Tupala et al., 2003). Wiener and colleagues (2013) built on these findings to test whether aged humans show a spontaneous preference for beacon-based navigation strategies. Overall, older adults were impaired at route learning compared to a young control group, but even when underperforming older adults were removed from the data, a striking age-related bias remained: when these participants were asked to rejoin the familiar route from a novel vantage point, they overwhelmingly used landmarks as beacons and showed little evidence for forming associations with specific movement decisions. These finding strongly suggest that using landmarks as beacons may be a strategy least affected by aging, which could allow for successful route negotiation in old age whenever an environment provides salient beacons.

One consistent finding of many studies on cognitive aging – not only in the field of spatial navigation – is that performance variability between participants increases with advancing age. In other words, whereas many older participants show striking performance impairments, others perform at levels similar to young control groups. Such findings naturally raise questions about what differentiates people who experience rapid cognitive decline from those who are more resilient to age-related changes. Neuroanatomically, individual trajectories of structural brain changes vary greatly, which has been shown for some structures involved in spatial navigation (i.e., hippocampus, Raz et al., 2005). Whether or not such differences are related to predispositions such as brain size, intellectual abilities, level of formal education, etc. (Steffener and Stern, 2011; Coffey et al., 1999) is currently under intense debate. Moreover, given that navigational abilities are at least partially separate from other cognitive abilities (Hegarty et al., 2006), it will be important to determine whether an individual's general intellectual abilities or more specifically the level of

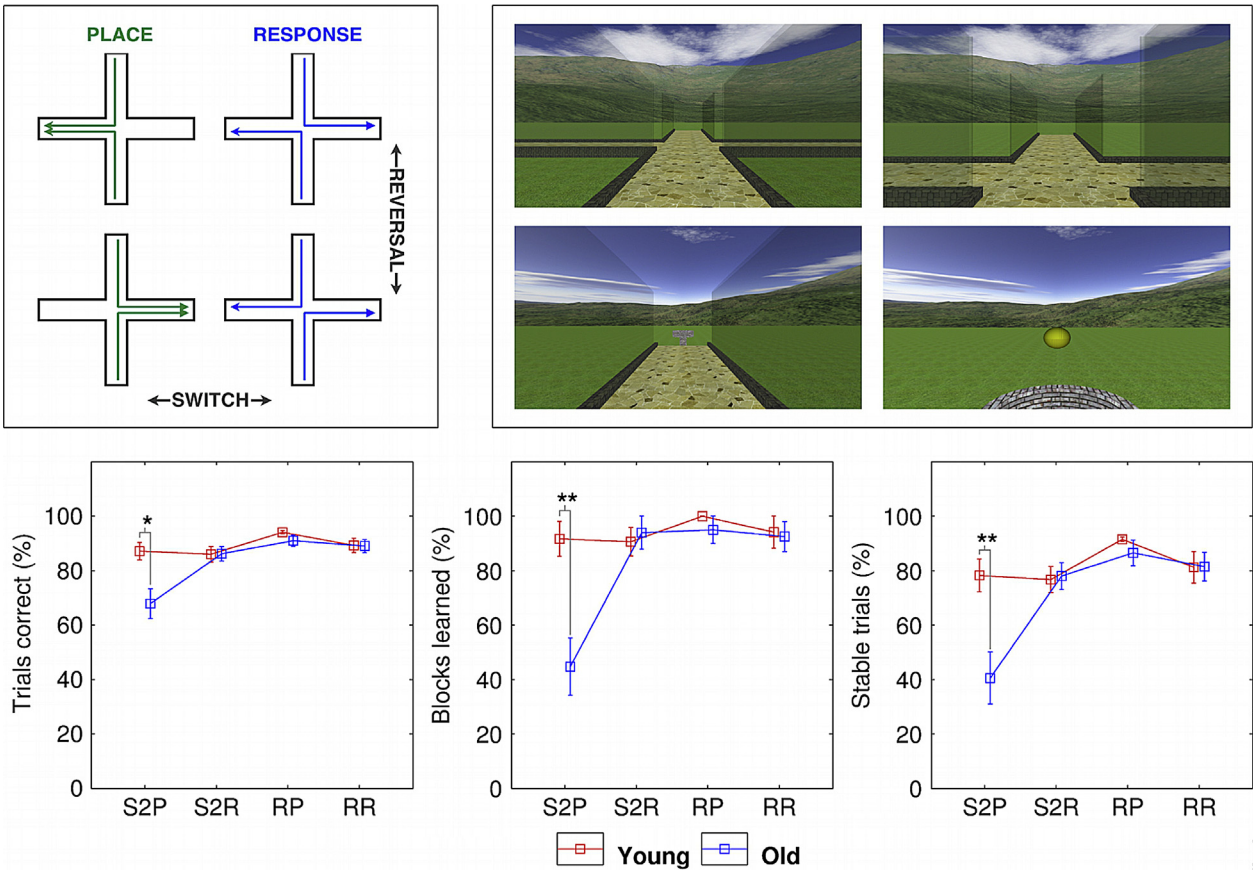

Figure 6 Impaired switching to an allocentric place strategy in older adults. Upper panels: left: diagrammatic representation of the plus maze task, illustrating place and response strategies, and switch and reversal change types. Right: screen images captured during the virtual plus maze task, at the start point, at the junction before turning, at the junction after turning, and at the goal (showing a reward). Upon reaching the junction, participants have to decide whether to go right or left. Participants need to learn the correct strategy from whether or not they receive a reward at the end of a trial. Strategy changes or reversals occur unbeknownst to participants. Lower panels: plus maze performance in terms of trials correct (left), blocks learned (middle), and stable trials (right) during blocks following switches to the place strategy (S2P), switches to the response strategy (S2R), reversals of place (RP), and reversals of response (RR) for young (red) and old (blue) participants. All three measures clearly show a selective deficit in older adults when required to switch from the response to the place strategy. Figure reproduced, with permission, from Harris, M.A., Wiener, J.M., Wolbers, T., 2012. Aging specifically impairs switching to an allocentric navigational strategy. Frontiers in Aging Neuroscience 4, 29.

navigational abilities may predict the rate of navigational decline experienced in old age.

Concluding Remarks

Ever since Darwin's visionary statement, much has been learned about the numerous cognitive computations that help us find our way in our everyday environments – and the underlying neuronal systems. Nevertheless, cognitive neuroscience research on spatial navigation still is in its early days, and many important challenges remain. For example, probably due to the complexity of navigational behavior, researchers have developed a huge variety of experimental tasks. These tasks are not always well understood in terms of the navigational processes they involve, which is particularly problematic for neuroimaging studies trying to assign specific functions to the activation differences they might observe between different experimental conditions. Secondly, relatively few studies have addressed questions of multisensory integration, even though animal research has clearly demonstrated that body-based cues are important for spatial computations carried out in circuits such as the place cell network. Thirdly, while humans differ in any cognitive ability (i.e., attention, memory), interindividual variability in navigational functioning is probably among the largest of all cognitive functions. Specifically, while some people find it easy to form a mental image of an environment even after just one or two visits, others get lost as soon as they have to leave a habitual route. At present, what differentiates these extremes – i.e., in terms of genetic predispositions or learned behaviors – is poorly understood. And finally, considering the existing animal research on the effects of aging on navigational processes, the low number of studies addressing similar questions in humans is surprising. Spatial navigation is an essential everyday function; hence a precise understanding of navigational impairments in old age is vital for improving the quality of life of our rapidly aging societies. In addition, some of the spatial computations affected by aging are likely to play a role for nonspatial functions as well, hence they could provide a novel window into general mechanisms

of age-related cognitive decline. Harvesting this unique opportunity shall be as fascinating as the topic of investigation – spatial navigation – itself.

See also: Aging and Memory; Learning and Memory, Neural Basis of; Non Declarative (Procedural) Memory.

Bibliography

Aguirre, G.K., D'Esposito, M., 1999. Topographical disorientation: a synthesis and taxonomy. Brain 122, 1613–1628.

Allen, G.L., Kirasic, K.C., Rashotte, M.A., Haun, D.B., 2004. Aging and path integration skill: kinesthetic and vestibular contributions to wayfinding. Perception and Psychophysics 66, 170–179.

Antonova, E., Parslow, D., Brammer, M., Dawson, G.R., Jackson, S.H., Morris, R.G., 2009. Age-related neural activity during allocentric spatial memory. Memory 17, 125–143.

Begega, A., Cienfuegos, S., Rubio, S., Santín, J.L., Miranda, R., Arias, J.L., 2001. Effects of ageing on allocentric and egocentric spatial strategies in the Wistar rat. Behavioral Processes 53, 75–85.

Burns, P.C., 1999. Navigation and the mobility of older drivers. Journals of Gerontology Series B, Psychological Sciences and Social Sciences 54, S49–S55.

Byrne, P., Becker, S., Burgess, N., 2007. Remembering the past and imagining the future: a neural model of spatial memory and imagery. Psychological Review 114, 340–375.

Cooper, B.G., Manka, T.F., Mizumori, S.J., 2001. Finding your way in the dark: the retrosplenial cortex contributes to spatial memory and navigation without visual cues. Behavioral Neuroscience 115, 1012–1028.

Cho, J., Sharp, P.E., 2001. Head direction, place, and movement correlates for cells in the rat retrosplenial cortex. Behavioral Neuroscience 115, 3–25.

Coffey, C.E., Saxton, J.A., Ratcliff, G., Bryan, R.N., Lucke, J.F., 1999. Relation of education to brain size in normal aging: implications for the reserve hypothesis. Neurology 53, 189–196.

Creem-Regehr, S.H., 2004. Remembering spatial locations: the role of physical movement in egocentric updating. In: Allen, G.L. (Ed.), Human Spatial Memory: Remembering Where. Lawrence Erlbaum Associates, Mahwah, NJ.

Darwin, C., 1873. Origin of certain instincts. Nature 7, 417–418.

Devan, B.D., White, N.M., 1999. Parallel information processing in the dorsal striatum: relation to hippocampal function. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 19, 2789–2798.

Epstein, R.A., 2008. Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in Cognitive Sciences 12, 388–396.

Etienne, A.S., Jeffery, K.J., 2004. Path integration in mammals. Hippocampus 14, 180–192.

Farrell, M.J., Robertson, I.H., 1998. Mental rotation and the automatic updating of body-centered spatial relationships. Journal of Experimental Psychology Learning Memory and Cognition 24, 227–233.

Farrell, M.J., Robertson, I.H., 2000. The automatic updating of egocentric spatial relationships and its impairment due to right posterior cortical lesions. Neuropsychologia 38, 585–595.

Franz, M.O., Mallot, H.A., 2000. Biomimetic robot navigation. Robotics and Autonomous Systems 30, 133–153.

Fujita, N., Klatzky, R.L., Loomis, J.M., Golledge, R., 1993. The encoding-error model of pathway completion without vision. Geographical Analysis 25, 295–314.

Golledge, R., 1999. Human wayfinding and cognitive maps. In: Golledge, R. (Ed.), Wayfinding Behavior. Johns Hopkins, Baltimore, MD.

Goodrich-Hunsaker, N.J., Livingstone, S.A., Skelton, R.W., Hopkins, R.O., 2010. Spatial deficits in a virtual water maze in amnesic participants with hippocampal damage. Hippocampus 20, 481–491.

Hafting, T., Fyhn, M., Molden, S., Moser, M.B., Moser, E.I., 2005. Microstructure of a spatial map in the entorhinal cortex. Nature 436, 801–806.

Harris, M.A., Wiener, J.M., Wolbers, T., 2012. Aging specifically impairs switching to an allocentric navigational strategy. Frontiers in Aging Neuroscience 4, 29.

Harris, M.A., Wolbers, T., 2012. Ageing effects on path integration and landmark navigation. Hippocampus 22, 1770–1780.

Hart, R.A., Moore, G.T., 1973. The development of spatial cognition: a review. In: Stead, D. (Ed.), Image and Environment. Aldine, Chicago, IL.

Hartley, T., Maguire, E.A., Spiers, H.J., Burgess, N., 2003. The well-worn route and the path less traveled. Distinct neural bases of route following and wayfinding in humans. Neuron 37, 877–888.

Head, D., Isom, M., 2010. Age effects on wayfinding and route learning skills. Behavioural Brain Research 209, 49–58.

Hegarty, M., Montello, D.R., Richardson, A., Ishikawa, T., Lovelace, K., 2006. Spatial abilities at different scales: individual differences in aptitude-test performance and spatial-layout learning. Intelligence 34, 151–176.

Hort, J., Laczo, J., Vyhnalek, M., Bojar, M., Bures, J., Vlcek, K., 2007. Spatial navigation deficit in amnestic mild cognitive impairment. Proceedings of the National Academy of Sciences of the United States of America 104, 4042–4047.

Iaria, G., Chen, J.K., Guariglia, C., Ptito, A., Petrides, M., 2007. Retrosplenial and hippocampal brain regions in human navigation: complementary functional contributions to the formation and use of cognitive maps. European Journal of Neuroscience 25, 890–899.

Iaria, G., Palermo, L., Committeri, G., Barton, J.J., 2009. Age differences in the formation and use of cognitive maps. Behavioral Brain Research 196, 187–191.

Iaria, G., Petrides, M., Dagher, A., Pike, B., Bohbot, V.D., 2003. Cognitive strategies dependent on the hippocampus and caudate nucleus in human navigation: variability and change with practice. Journal of Neuroscience 23, 5945–5952.

Ino, T., Doi, T., Hirose, S., Kimura, T., Ito, J., Fukuyama, H., 2007. Directional disorientation following left retrosplenial hemorrhage: a case report with fMRI studies. Cortex 43, 248–254.

Ishikawa, T., Montello, D.R., 2006. Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. Cognitive Psychology 52, 93–129.

Jahn, G., Wendt, J., Lotze, M., Papenmeier, F., Huff, M., 2012. Brain activation during spatial updating and attentive tracking of moving targets. Brain and Cognition 78, 105–113.

Janzen, G., Van Turennout, M., 2004. Selective neural representation of objects relevant for navigation. Nature Neuroscience 7, 673–677.

Jeffery, K.J., 2007. Self-localization and the entorhinal-hippocampal system. Current Opinion in Neurobiology 17, 684–691.

Jeffery, K.J., Donnett, J.G., Burgess, N., O'keefe, J.M., 1997. Directional control of hippocampal place fields. Experimental Brain Research 117, 131–142.

Kearns, M.J., Warren, W.H., Duchon, A.P., Tarr, M.J., 2002. Path integration from optic flow and body senses in a homing task. Perception 31, 349–374.

Killian, N.J., Jutras, M.J., Buffalo, E.A., 2012. A map of visual space in the primate entorhinal cortex. Nature 491 (7426), 761–764.

Lappe, M., Frenz, H., 2009. Visual estimation of travel distance during walking. Experimental Brain Research 199, 369–375.

Lappe, M., Jenkin, M., Harris, L.R., 2007. Travel distance estimation from visual motion by leaky path integration. Experimental Brain Research 180, 35–48.

Lappe, M., Stiels, M., Frenz, H., Loomis, J.M., 2011. Keeping track of the distance from home by leaky integration along veering paths. Experimental Brain Research 212, 81–89.

Lister, J.P., Barnes, C.A., 2009. Neurobiological changes in the hippocampus during normative aging. Archives of Neurology 66, 829–833.

Liu, I., Levy, R.M., Barton, J.J., Iaria, G., 2011. Age and gender differences in various topographical orientation strategies. Brain Research 1410, 112–119.

Loomis, J.M., Klatzky, R.L., Golledge, R.G., Cicinelli, J.G., Pellegrino, J.W., Fry, P.A., 1993. Nonvisual navigation by blind and sighted: assessment of path integration ability. Journal of Experimental Psychology General 122, 73–91.

Maguire, E.A., 2001. The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. Scandinavian Journal of Psychology 42, 225–238.

Maguire, E.A., Woollett, K., Spiers, H.J., 2006. London taxi drivers and bus drivers: a structural MRI and neuropsychological analysis. Hippocampus 16, 1091–1101.

Mahmood, O., Adamo, D., Briceno, E., Moffat, S.D., 2009. Age differences in visual path integration. Behavioural Brain Research 205, 88–95.

Mallot, H.A., Gillner, S., 2000. Route navigating without place recognition: what is recognised in recognition-triggered responses? Perception 29, 43–55.

Mapstone, M., Logan, D., Duffy, C.J., 2006. Cue integration for the perception and control of self-movement in ageing and Alzheimer's disease. Brain 129, 2931–2944.

McGeorge, A.J., Faull, R.L., 1989. The organization of the projection from the cerebral cortex to the striatum in the rat. Neuroscience 29, 503–537.

Mittelstaedt, M.L., Mittelstaedt, H., 1980. Homing by path integration in a mammal. Naturwissenschaften 67, 566–567.

Moffat, S.D., Elkins, W., Resnick, S.M., 2006. Age differences in the neural systems supporting human allocentric spatial navigation. Neurobiology of Aging 27, 965–972.

Moffat, S.D., Kennedy, K.M., Rodrigue, K.M., Raz, N., 2007. Extrahippocampal contributions to age differences in human spatial navigation. Cereb Cortex 17, 1274–1282.

Moffat, S.D., Resnick, S.M., 2002. Effects of age on virtual environment place navigation and allocentric cognitive mapping. Behavioral Neuroscience 116, 851–859.

Montello, D.R., 1998. A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In: Egenhofer, M.J., Golledge, R.G. (Eds.), Spatial and Temporal Reasoning in Geographic Information Systems. Oxford University Press, New York.

Morgan, L.K., Macevoy, S.P., Aguirre, G.K., Epstein, R.A., 2011. Distances between real-world locations are represented in the human hippocampus. Journal of Neuroscience 31, 1238–1245.

Morris, R.G., Garrud, P., Rawlins, J.N., O'keefe, J., 1982. Place navigation impaired in rats with hippocampal lesions. Nature 297, 681–683.

Moser, E.I., Moser, M.B., 2008. A metric for space. Hippocampus 18, 1142–1156.

Nicolle, M.M., Prescott, S., Bizon, J.L., 2003. Emergence of a cue strategy preference on the water maze task in aged C57B6 x SJL F1 hybrid mice. Learning Memory 10, 520–524.

O'keefe, J., Burgess, N., 1996. Geometric determinants of the place fields of hippocampal neurons. Nature 381, 425–428.

O'keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research 34, 171–175.

Packard, M.G., McGaugh, J.L., 1996. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiology of Learning and Memory 65, 65–72.

Parron, C., Save, E., 2004. Evidence for entorhinal and parietal cortices involvement in path integration in the rat. Experimental Brain Research 159, 349–359.

Petzschner, F.H., Glasauer, S., 2011. Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration. Journal of Neuroscience 31, 17220–17229.

Philbeck, J.W., Behrmann, M., Black, S.E., Ebert, P., 2000. Intact spatial updating during locomotion after right posterior parietal lesions. Neuropsychologia 38, 950–963.

Philbeck, J.W., Behrmann, M., Levy, L., Potolicchio, S.J., Caputy, A.J., 2004. Path integration deficits during linear locomotion after human medial temporal lobectomy. Journal of Cognitive Neuroscience 16, 510–520.

Philbeck, J.W., Behrmann, M., Loomis, J.M., 2001. Updating of locations during whole-body rotations in patients with hemispatial neglect. Cognitive, Affective & Behavioral Neuroscience 1, 330–343.

Raz, N., Lindenberger, U., Rodrigue, K.M., Kennedy, K.M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., Acker, J.D., 2005. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. Cerebral Cortex 15, 1676–1689.

Riecke, B.E., Cunningham, D.W., Bulthoff, H.H., 2007. Spatial updating in virtual reality: the sufficiency of visual information. Psychological Research 71, 298–313.

Saddala, E.K., Montello, D.R., 1989. Remembering changes in direction. Environment and Behavior 21, 346–363.

Sargent, J., Dopkins, S., Philbeck, J., Chichka, D., 2010. Chunking in spatial memory. Journal of Experimental Psychology Learning Memory and Cognition 36, 576–589.

Schweizer, K., Herrmann, T., Janzen, G., Katz, S., 1998. The route direction effect and its constraints. In: Freksa, C., Habel, C., Wender, K.F. (Eds.), Spatial Cognition, 1404 ed. Springer, Berlin.

Siegel, A.W., White, S.H., 1975. The development of spatial representations of large-scale environments. In: Reese, H.W. (Ed.), Advances in Child Development and Behaviour. Academic, New York.

Spiers, H.J., Maguire, E.A., 2007. A navigational guidance system in the human brain. Hippocampus 17, 618–626.

Steffener, J., Stern, Y., 2011. Exploring the neural basis of cognitive reserve in aging. Biochimica et Biophysica Acta 1822 (3), 467–473.

Stranahan, A.M., Mattson, M.P., 2010. Selective vulnerability of neurons in layer II of the entorhinal cortex during aging and Alzheimer's disease. Neural Plasticity 2010, 108190.

Taube, J.S., Muller, R.U., Ranck Jr., J.B., 1990a. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. Journal of Neuroscience 10, 420–435.

Taube, J.S., Muller, R.U., Ranck Jr., J.B., 1990b. Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. Journal of Neuroscience 10, 436–447.

Teng, E., Squire, L.R., 1999. Memory for places learned long ago is intact after hippocampal damage. Nature 400, 675–677.

Thorndyke, P.W., Hayes-Roth, B., 1982. Differences in spatial knowledge acquired from maps and navigation. Cognitive Psychology 14, 560–589.

Tolman, C.E., 1948. Cognitive maps in rats and man. Psychological Review 55, 189–208.

Tupala, E., Hall, H., Bergstrom, K., Mantere, T., Rasanen, P., Sarkioja, T., Hiltunen, J., Tiihonen, J., 2003. Different effect of age on dopamine transporters in the dorsal and ventral striatum of controls and alcoholics. Synapse 48, 205–211.

United Nations, D. O. E. A. S. A., Population Division, 2012. Population Ageing and Development 2012. United Nations, D. O. E. A. S. A., Population Division.

Valerio, S., Taube, J.S., 2012. Path integration: how the head direction signal maintains and corrects spatial orientation. Nature Neuroscience 15, 1445–1453.

Van Der Meer, M.A., Richmond, Z., Braga, R.M., Wood, E.R., Dudchenko, P.A., 2010. Evidence for the use of an internal sense of direction in homing. Behavioral Neuroscience 124, 164–169.

Voermans, N.C., Petersson, K.M., Daudey, L., Weber, B., Van Spaendonck, K.P., Kremer, H.P., Fernandez, G., 2004. Interaction between the human hippocampus and the caudate nucleus during route recognition. Neuron 43, 427–435.

Voorn, P., Vanderschuren, L.J., Groenewegen, H.J., Robbins, T.W., Pennartz, C.M., 2004. Putting a spin on the dorsal-ventral divide of the striatum. Trends in Neurosciences 27, 468–474.

Waller, D., Hodgson, E., 2006. Transient and enduring spatial representations under disorientation and self-rotation. Journal of Experimental Psychology Learning Memory and Cognition 32, 867–882.

Waller, D., Lippa, Y., 2007. Landmarks as beacons and associative cues: their role in route learning. Memory and Cognition 35, 910–924.

Wang, R.F., 2004. Between reality and imagination: when is spatial updating automatic? Perception and Psychophysics 66, 68–76.

Wiener, J.M., Berthoz, A., Wolbers, T., 2011. Dissociable cognitive mechanisms underlying human path integration. Experimental Brain Research 208, 61–71.

Wiener, J.M., Berthoz, A., Wolbers, T., 2012. Dissociable cognitive mechanisms underlying human path integration. Experimental Brain Research 208, 61–71.

Wiener, J.M., De Condoppa, O., Harris, M., Wolbers, T., 2013. Maladaptive bias for extrahippocampal navigation strategies in aging humans. Journal of Neuroscience.

Wilkniss, S.M., Jones, M.G., Korol, D.L., Gold, P.E., Manning, C.A., 1997. Age-related differences in an ecologically based study of route learning. Psychology and Aging 12, 372–375.

Wolbers, T., Büchel, C., 2005. Dissociable retrosplenial and hippocampal contributions to successful formation of survey representations. Journal of Neuroscience 25, 3333–3340.

Wolbers, T., Hegarty, M., 2010. What determines our navigational abilities? Trends in Cognitive Science 14, 138–146.

Wolbers, T., Hegarty, M., Büchel, C., Loomis, J.M., 2008. Spatial updating: how the brain keeps track of changing object locations during observer motion. Nature Neuroscience 11, 1223–1230.

Wolbers, T., Wiener, J.M., Mallot, H.A., Büchel, C., 2007. Differential recruitment of the hippocampus, medial prefrontal cortex, and the human motion complex during path integration in humans. Journal of Neuroscience 27, 9408–9416.

Woollett, K., Maguire, E.A., 2011. Acquiring "the Knowledge" of London's layout drives structural brain changes. Current Biology CB 21, 2109–2114.

Worsley, C.L., Recce, M., Spiers, H.J., Marley, J., Polkey, C.E., Morris, R.G., 2001. Path integration following temporal lobectomy in humans. Neuropsychologia 39, 452–464.

Yin, H.H., Knowlton, B.J., 2004. Contributions of striatal subregions to place and response learning. Learning Memory 11, 459–463.

Spatial Optimization Models

Richard Church, University of California Santa Barbara, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human activities are often arranged into zones or groups. For example, a town may have a central business district, residential areas, schools, medical clinics, etc. Activity patterns are determined in part by the basic layout. In contrast, a manufacturing plant is designed to have an assembly line or assembly areas, parts fabrication, inventory storage, etc. The layout of any facility helps determine the cost of manufacturing a product, just as the layout of a town helps determine the extent to which various people travel to accomplish daily tasks. The layout or location of all or a subset of such activities is called spatial optimization. This review article describes a number of different spatial optimization problems and concludes with directions for future research.

The concept of spatial optimization is quite simple. This topic involves identifying how land use and activities should be allocated and arranged spatially in order to optimize efficiency or some other measure of goodness. For example, how should fire protection resources be distributed across a city in order to maximize protection? How do we design a system of reserve lands that can ensure the survival of an endangered species? How should an area be divided into school attendance zones? These are but a few examples of spatial design problems. For each type of design problem there can be numerous if not an infinite number of alternative configurations or solutions. Optimization involves the attempt to identify the best of the alternatives. Simply put, spatial optimization is the science of optimal arrangement. The use of spatial optimization is growing owing to the existence of better spatial data, the emergence of powerful geographic information systems, decreasing costs of computation, and better computer graphics and solution algorithms.

At first we might think that such a topic is futuristic, cutting edge. Actually, spatial optimization is a geographical problem that is as old as mankind. For example, hunters and gatherers had to decide how to allocate their efforts in collecting foodstuffs, moving encampments, etc. Such societies needed to be efficient in order to survive. Anthropologists and geographers have used optimization to analyze the efficiency of these early societies (see, e.g., Van Reidhead, 1979). Even the ancient Chinese art of Feng Shui deals with the arrangement of activities in order to be in a harmonious state. Just as we are interested in how spatial arrangements evolved in the past, we are equally interested in how future patterns of activities should be designed and placed, especially to meet sustainability, health, justice, and resilience. For example, just where should that new store be located? Where should a dealer provide service facilities for their products? How should a network of warehouses be designed and operated in order to supply a number of retail locations? What network of supply routes and storage centers best responds with humanitarian aid in disasters such as an earthquake in Sichuan or a monsoon in Bangladesh? Such questions are but the tip of the 'iceberg' of possible spatial location/allocation problems.

As with all geographic problems, the temporal domain is important. For example, Broward County, FL operates a school system with a budget of US$3.1 billion (2013) involving 315 schools and 260 000 students. Every year, a 5-year forecast is made for enrollment at each school. Those schools with a projected enrollment of 95% or more of capacity are then subject to possible boundary adjustments as well as other enrollment management strategies (e.g., adding portables). One of the major objectives in planning new schools and creating attendance areas is to plan an expansion that involves the least disruption of students and school reassignments (Lemberg and Church, 2000). In a different example, to meet the fluctuations in daily emergency '911' call demand in San Diego County, CA, Rural Metro Medical Response operates between 2 and 24 ambulances during each day. Posting ambulances at various locations and dispatching them over time is a complicated logistical/spatial optimization problem. Rest assured that the patterns of deployment vary over the day in order to reduce response times, but just where should such posts be located to get the most service out of a limited number of vehicles? Since it costs over US$700 000 per year to keep an ambulance in operation, it is important to find an efficient service plan. As another example, in Los Angeles County, the County, the City of Los Angeles, and numerous other municipalities within the county operate approximately 392 fire stations. Altogether, this system is estimated to cost more than US$2 billion a year to operate. This means that there is an interest in relocating and reducing what is used as well as consolidating services across municipalities, without compromising on service quality, a problem that can be addressed as a spatial optimization problem.

Geographic scale is an equally important issue in spatial optimization. There are typically three basic scales in spatial optimization problems: micro, meso, and macro. The macro-scaled problem involves the highest level or area size (i.e., smallest map scale). At such a level, we might ask: 'in which cities should we have factories or warehouses?' We generally use no specific site information in making such a selection. Interactions between customers and facilities are modeled at an aggregated level. The objective is to identify those areas in which it makes the most sense to open a facility or activity. Such a decision might be considered 'strategic.' If we have a strategic plan to locate a regional warehouse in Reno, Nevada, the 'meso' level question is 'where in Reno should it be?' Such a problem would involve identifying the best available site, considering a wide range of criteria. This is a type of 'tactical'

decision making and is often performed to support a strategic decision. Once we have a site or a place, the micro-scaled problem involves carving up that space and allocating it into specific activities. For example, we may allocate store space to specific retail items or a shop floor to machine tools. Even malls are arranged spatially in order to take advantage of anchor stores and walking patterns (Bean et al., 1988). How people use such facilities and interact with their environment is an especially important aspect of this problem. In the industrial engineering literature, this type of problem is often referred to as a layout problem. Micro-scaled problems are akin to an 'operational level' problem.

Historical Roots of Spatial Optimization

Efficient land use activity was first discussed by Von Thunen in 1842, when he described how agricultural activities including wood lots should be optimally arranged around an isolated village. Primary industries such as forestry now use spatial optimization to identify how each stand (tree type and age class) is to be managed over time. These management plans include the building of roads, the provision of environmental standards, and the return on investment (Church et al., 1998). In Chile, forest companies optimize nearly every feature of their industry, including the scheduling and routing of vehicles for log transport (so that logs arrive when just needed at loading facilities such as ports), the placement of harvest towers and cable ways for steep slope harvesting, and the minimization of land devoted to unproductive elements (e.g., roads). The US Forest Service has utilized a spatial optimization model called FORPLAN (and its successor SPECTRUM) for the last three decades in developing management plans (Kent et al., 1991) for virtually every national forest in the United States.

Alfred Weber proposed one of the earliest industrial location models in 1908 (Wesolowsky, 1993). As a simple example, he described the location of a manufacturing point between two localized raw materials and a market. He considered that the three points could be plotted on a Cartesian plane as shown in Figure 1. Transport distances between the points were assumed to be Euclidean. Weber discussed how, under certain assumptions, that the ideal location was the one that minimized transport costs of the raw materials to the manufacturing point and the product to market. Georg Pick wrote an appendix to Weber's book that demonstrated how this model could be solved geometrically for certain problems. The first optimal solution algorithm for the more generalized Weber model involving many markets and raw materials was developed in 1937 by Weisfeld and reinvented by Kuhn and Kuenne and Cooper as well (Wesolowsky, 1993). Cooper subsequently proposed an extended model involving the location of multiple facilities (Wesolowsky, 1993).

In 1933 Walter Christaller wrote: "if there are now economic laws, then there must be also laws of the geography of settlements" (Christaller, 1966: p. 3). He asks the question: "are there laws which determine the number, sizes, and distribution of towns?" (Christaller, 1966: p. 1). Christaller attempted to find the laws that determined the number, size, and distribution of towns in a region of Southern Germany. Since that time geographers, archaeologists, and anthropologists have searched

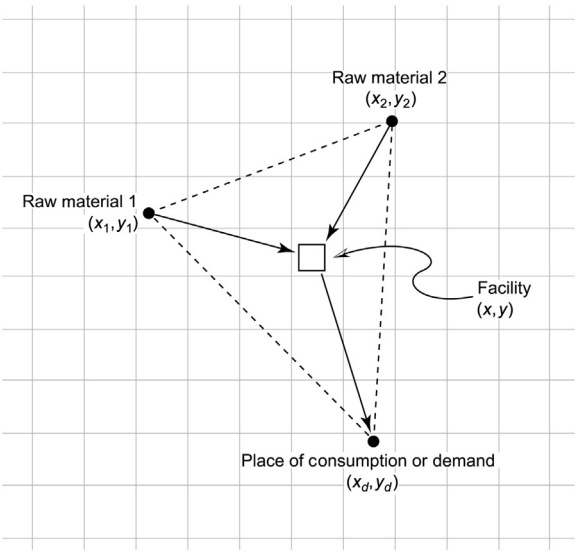

Figure 1 The classic Weber location triangle.

for Christaller's central place patterns in other regions and in other applications and have developed spatial optimization models that solve for the optimal spatial arrangement of towns and markets (Kuby, 1989). Virtually all early location theories have some type of spatial optimization counterpart today.

Selected Examples of Spatial Optimization

A large number of spatial design/arrangement/allocation problems have been approached using some form of optimization modeling. Techniques used to solve optimization models are typically found in the operations research (OR) literature. The field of OR traces its roots to World War II and the need to solve large-scale military logistics problems. One of the first OR models that was developed and can be solved optimally was the classical transportation problem. This problem involves a distributed set of supply points (called sources) and a distributed set of demand locations (called sinks). Each source has a finite capacity to supply a good and each destination has a specified amount of demand for that good that needs to be met. Between each unique pair of one source and one sink we can identify the most efficient route and the cost to ship one unit of that good between that source–sink pair. The objective of the classical transportation problem is to allocate the supplies to the demands so that the total transport cost is minimized and so that the demand at each sink is met. Garrison (1959) was the first geographer to discuss the utility of the transportation problem as a model to analyze spatial structure. Yeates (1963) and others have since used this spatial optimization model to analyze efficiency of an existing arrangement or to develop an optimal transportation plan.

Table 1 gives a representative sample of spatial optimization problems that can be found in the literature, including the classical transportation problem. The list includes vehicle

Table 1 Examples of spatial optimization problems

Problem name	*Problem definition*	*General reference*
Corridor location problem	Across a landscape, identify the optimal route for a right-of-way for a facility such as a pipeline, highway, or power transmission line that connects an origin and destination; across a landscape identify an efficient corridor for wildlife movement between two habitat cores.	Lombard and Church, 1993; Kemp and O'Hanley, 2010
Vehicle routing problem	Optimally route a vehicle from a starting point to visit a number of prespecified locations and then return to the starting point.	Weigel and Cao, 1999; Thill et al., 2011
Classical transportation problem	Allocate limited supplies at different locations to a number of spatially distributed demand points in such a manner as to minimize transportation costs of serving all demand.	Yeates, 1963; Camm et al., 1997
Simple and capacitated plant location problems	Locate a set of facilities and allocate their supplies to demand points in such a manner as to minimize facility and transportation costs.	Upchurch et al., 2009; Kuby et al., 2011
p-median location problem	Locate *p*-facilities so that the total weighted distance of serving all demand is minimized.	ReVelle and Swain, 1970; Densham and Rushton, 1992
Maximal covering location problem	Locate *p*-facilities so that the largest proportion of demand is covered or served within some maximal distance or time	Church and ReVelle, 1974; Tong et al., 2009
p-hub location problem	Locate *p*-transportation hubs and allocate demand to specific hubs in such a manner as to minimize total transportation costs.	O'Kelly and Miller, 1994
Site configuration/design problem	Identify the best set of parcels or land units that optimizes costs and meets shape, contiguity, size, and suitability constraints	Wright et al., 1983; Williams 2002
Districting, zonation, and region delineation problems	Partition spatial units into a set of districts or regions that optimize one or more criteria, examples school districts, political districts, etc.	Openshaw, 1997; Duque et al., 2012
Spatial allocation problem	Allocate specific activities to defined spatial units	Aerts et al., 2003; Ligmann-Zielinska et al., 2008
Disruption, fortification, and resilient design	Interdict a system in a manner to inflict the most harm, protect a system so that it escapes the most harm, design a system which is resilient to natural or artificial disasters	Grubesic et al., 2008; Church et al., 2004

routing, districting and zonation, transportation, facility location, site search, and space allocation. For each of the 11 problems listed, the name, definition, and a few references are given. For each of these generic problems, a number of models have been developed. The references cite the original work, a review, or a recent work that contains an up-to-date list of references. All but the last problem listed are classic problems that have been the subject of numerous papers in the geographic and regional science literature. The last problem area is relatively new and addresses issues of disruption, fortification, and resilience. Although there are classic spatial problems of risk such as delineating a flood plain or modeling a plume from a hazardous chemical spill, recent attention has been directed at the potential for system disruption and failure. Since the attack on the New York World Trade Center in 2001, concern for disruption, designing for resilience, and managing disasters have become a major research area. For example, Grubesic and Murray, (2006) examined the Internet backbone with the intent on finding those nodes which when removed would cause the greatest disruption to network topology (Grubesic et al., 2008). Church et al. (2004) developed an interdiction model to find those facilities of a system that are critical to efficient operation. It is logical to ask, how should protection resources be allocated to prevent disastrous outcomes (Church and Scaparra, 2007) and how can failures cascade through numerous interconnections and networks (Grubesic et al., 2008). Finally, shelter locations, evacuation routes, and resources for emergency services for making a quick response during hurricanes and other natural disasters have become an important area addressed by those working on spatial optimization problems (Alçada-Almeida et al., 2009).

All of the problems listed in **Table 1**, with the exception of the corridor location problem and the classical transportation problem, have a problem complexity that is called *Nondeterministic Polynomial* (*NP*). The class *NP* refers to a set of problems for which there exists virtually no chance that procedures will be found that can solve all problems optimally within a reasonable computational time on a computer. One of the reasons for this is that these problems require the use of integer programming, an area of optimization that can be both computationally intensive and limited in terms of the size of an individual application that can be solved optimally. Because of this fact, it is common to see that different techniques are used for solving specific models and that heuristics are needed for most moderate to large sized problems (see, e.g., Tong et al., 2009). Heuristics are specialized solution procedures that are designed to approach problem solving using either a set of rules or some search strategy. An example of a heuristic is given in the next section.

A Detailed View of a Spatial Optimization Problem

Hakimi (1965) proposed a network location model called the *p*-median problem. Consider the network depicted in **Figure 2**. Each node is connected to other nodes by arcs of given distances. Without any loss of generality, consider that each node represents a potential facility site as well as a point of demand. The measure of demand is represented as a demand

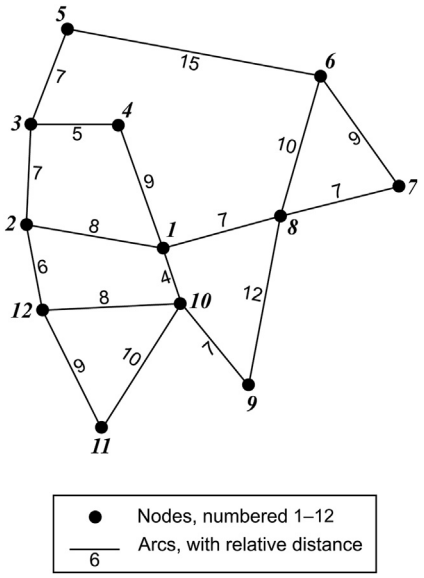

Figure 2 An example road network.

weight (in units of people, tons, trips, or some other measure of need to be supplied by a facility). Assume that each demand will be supplied or served by the closest facility. The transportation cost is assumed to be directly proportional to the weight of a demand multiplied by the distance to the closest facility (note: this is the same assumption used in the Weber model). The p-median problem involves selecting the locations of p-facilities so that the total weighted distance for all demand is minimized. Hakimi proved that at least one optimal solution to the p-median problem consists entirely of nodes of the network. Thus, the search for an optimal configuration can be limited to just the nodes of the network.

Searching for the best nodal configuration, and hence an optimal configuration, seems simple. Why not enumerate all of the possible facility patterns, and pick the configuration with the lowest weighted distance? Unfortunately, it is not that simple. Consider for example that we wish to locate 10 facilities on a network of 100 nodes. Such an enumeration would involve generating and evaluating the following number of combinations:

$$\binom{100}{10} = \frac{100!}{5!(100-5)!} = \frac{100*99*98*97*96*95*94*\ldots*1}{10*9*8*\ldots*1*90*89*88*\ldots*1}$$

or

$$\frac{100*99*98*97*96\ldots*91}{10*9*8*7*\ldots*1} \cong 17\,300\,000\,000\,000 \quad \text{configurations or alternatives.}$$

If a very fast computer could analyze 1000 solutions each second, it would take 548 such computers working simultaneously to finish the task in a year. Thus, for many problems enumeration is impossible. Since enumeration is not possible for many problems, it is reasonable to consider optimization. ReVelle and Swain (1970) proposed the first optimal approach to solving the p-median problem. The basis of their approach is to formulate the p-median model as an integer programming problem.

Consider a connected transportation network comprised of n nodes and at least $n-1$ arcs (note: it requires at least $n-1$ arcs to be connected). An example of a 12-node network is given in Figure 2. In this review we will consider the most general form of the p-median problem where each node represents both a place of demand and a potential facility site. Given any pair of nodes on the network, a shortest path exists between the pair of nodes, and the distance of that path is easily calculated by a variety of efficient techniques. Consider the following notation:

n = the number of nodes in the network
i,j = indices used to reference nodes, where the nodes are numbered from 1 to n
a_i = the amount of demand at node i
d_{ij} = the distance from node i to node j

$$x_{ij} = \begin{cases} 1, & \text{if demand at node } i \text{ assigns to a facility at node } j \\ 0, & \text{otherwise} \end{cases}$$

$$x_{jj} = \begin{cases} 1, & \text{if node } j \text{ self assigns indicating that a facility is located at node } j \\ 0, & \text{otherwise} \end{cases}$$

Using the above notation, we can define the following optimization model (ReVelle and Swain, 1970):

Objective function:	Minimize $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i d_{ij} x_{ij}$	
Subject to:		
1. Each demand must be served by either placing a facility at its node or by assigning to a facility elsewhere	$\sum_{j=1}^{n} x_{ij} = 1$	For each node i
2. Assignment is possible to only open facilities, where demand self-assigns	$x_{ij} \le x_{jj}$	For each node i and j where $i \neq j$
3. Exactly p-facilities are located	$\sum_{j=1}^{n} x_{jj} = p$	
4. Integer restrictions on variables	$x_{ij} = 0,1$	For each node i and j where $i \neq j$

The above model can be solved by the use of integer programming techniques, most notably, linear programming with branch and bound (LP/BB). In a relatively high percentage of problems, the optimal linear programming solution is integer optimal as well and the BB routine is not used. This property is called 'integer friendly.' Since BB algorithms may take considerable computational time to resolve noninteger solutions, finding an optimal integer optima without BB usually means relatively fast computational times for a given problem size. Unfortunately, the above model is large in terms of the number or variables and constraints (n^2 and n^2+1 respectively). Consequently there is a limit as to how large a problem can be solved by this specific model formulation and LP/BB.

It is possible to formulate the p-median problem with a slightly different constraint set. Consider the following form of the second constraint:

2. Assignments are possible to a facility, only if it has been selected	$\sum_{i=1}^{n} x_{ij} \leq (n-p)x_{jj}$	For each node j

The constraints that restrict demand assignments to only those nodes selected for a facility have been aggregated into one constraint for each facility. Consequently, this second model is more compact in that it contains only $2n + 1$ constraints rather than $n^2 + 1$ constraints. As models are typically easier to solve when they have fewer constraints, this would at first seem to be an improvement. Unfortunately, this second model form is not 'integer friendly' and consequently the use of LP/BB typically relies on the BB algorithm and resulting solution times can be very large. The lesson here is that there are always alternate ways in which to formulate any given spatial optimization model, and the difficulty in solving a given problem can be in part due to the mathematical form of the problem. The original problem formulation has met the test of time, as it is still the formulation of choice in describing the p-median problem. There are, however, new attempts at reformulating this problem so that it uses as few variables and constraints as possible. These models are particularly tuned to spatial properties (Church, 2008).

There are practical limits, however, as to the size of a p-median problem that can be solved optimally. For very large problems it is simply not possible to find an optimal solution with certainty. Because of this there has been a great reliance on heuristic solution procedures. Heuristics are often designed to exploit some type of search strategy. For example one of the best-known methods for solving the p-median problem is the 'swap' heuristic of Teitz and Bart (Densham and Rushton, 1992). The 'swap' heuristic starts with a randomly generated solution. Given this starting solution, each node is considered in turn as a possible replacement site for each of the currently selected nodes. If any swap involving this candidate node yields an improvement, the best swap with that candidate is made, creating a new and improved solution. Swaps are tested until no improvements can be made involving any node of the network. Repeating this procedure a few times has a good chance of finding a good, if not optimal solution. Other strategies are based upon 'survival of the fittest' genetic algorithms, simulated annealing and statistical mechanics, Tabu search, Monte Carlo sampling, etc. All heuristics are rated for their capabilities in finding good solutions.

The Problem of Aggregation and Domain Representation

Virtually all spatial problem representations contain some type of abstraction. For instance, it is impossible to completely characterize all elements of a coastal boundary without some form of line generalization. In the p-median problem described above, demand locations and possible facility locations were restricted to the nodes of a network. The basic problem in applying a spatial optimization model, like the p-median model, is in defining the scale at which sites and demand points are represented. For example, would one define each household as a demand point and vacant buildings and lots as potential facility locations or sites? Or, would it be more appropriate to aggregate demand to census blocks, census tracts, or some other spatial unit? The problem is that aggregation for spatial analysis and optimization, in and of itself, introduces the possibility of error, which was first recognized by Goodchild (1979). Goodchild described the problem of aggregating demand points into zones for the purpose of applying a location-allocation model and demonstrated that spatial differences in the final solution can be attributed to the process of problem aggregation.

The other issue of major importance is problem domain representation. Spatial optimization problems have been cast on continuous domains and discrete domains. For example, the p-median problem defined in the previous section was defined for a discrete domain of demand points and facility sites. However, the related Weber problem depicted in **Figure 1** is defined on a continuous Cartesian coordinate system where any point is a feasible facility site. Both continuous and discrete domains have their own attendant problems. One problem that has received considerable attention is the location of warning sirens across a bounded region, so that all points within the region are within a specified distance of a siren, where the objective is to minimize the number of sirens needed to cover the bounded region. In this case, the demand is described as a bounded polygonal region and the sites are defined as any point within the region (Tong and Murray, 2012). Whereas some problems are more easily approached as a discrete optimization, their counterpart continuous spatial representation may represent a more realistic application domain. Unfortunately, most discrete domain problems, although computationally difficult in themselves are more computationally tractable than their continuous domain counterparts.

Research Needs in Spatial Optimization

Research in spatial optimization has expanded over the years. This expansion is due in part to increasing computational power and availability of computers, better general purpose commercial products like Geographical Information Systems (GIS) and optimization software, fourth generation computer language tools like visual studio and C++, and high resolution displays. Some spatial optimization models are now a part of general purpose GIS software. There are, however, a number of research directions that must be taken in order for such modeling tools to be commonly available. To make this a reality, research is needed in the following areas:

1. develop better and more capable solution algorithms and heuristics. This is especially important in improving the capability to solve large problems of the simple plant location problem where the facilities have finite capacities;
2. expand existing models and develop more comprehensive spatial optimization models, especially in ecological reserve design and land use planning;
3. develop methods to better estimate the impact of data error and its propagation on model results. This is especially

important since all geographic data sets of realistic size contain error;

4. expand data model definitions and problem representations to better link spatial optimization models with GIS. See Tong and Murray (2012) for a good description of some of the nuances of spatial representations and model definitions. One application of considerable importance is the optimal location of cellular phone towers;
5. develop better visualization methods along with graphical interface approaches that can support collaborative decision making processes;
6. explore methods of data aggregation so that it may be possible to construct small problem representations that are easier to solve, but still retain all of the important problem characteristics.

Many industries and government agencies have been pushed to become more efficient. Increasing productivity has become a mantra of stockholder and citizen alike. In many settings new spatial optimization models are being proposed to help increase efficiency, promote equity, reduce environmental damages, and reduce energy needs (see, e.g., a carbon capture and sequester model by Kuby et al., 2011). For example, new spatial/temporal optimization models that are being used for emergency ambulance operations can create plans that improve service response without additional investment or keep service levels constant and cut costs. The economic incentive to improve services and reduce costs will expand the application base that is already very large.

See also: Applied Geography: A US Perspective; Geographic Information Systems; Industrial Geography; Mathematical Models in Geography; Scale in Geography.

Bibliography

Aerts, J.C.J.H., Eisinger, E., Heuvelink, G.B.M., Stewart, J.T., 2003. Using linear integer programming for multi-site land-use allocation. Geographical Analysis 35, 148–169.

Alçada-Almeida, L., Tralhão, L., Santos, L., Coutinho-Rodrigues, J., 2009. A multi-objective approach to locate emergency shelters and identify evacuation routes in urban areas. Geographical Analysis 41, 9–29.

Bean, J.C., Noon, C.E., Ryan, S.M., Salton, G.J., 1988. Selecting tenants in a shopping mall. Interfaces 18, 1–9.

Camm, J.D., Chorman, T.E., Dill, F.A., Evans, J.R., 1997. Blending OR/MS, judgment, and GIS: restructuring P&G's supply chain. Interfaces 27, 128–142.

Christaller, W., 1966. Central Places in Southern Germany (Carlisle W. Baskin, Trans.). Prentice-Hall, Englewood Cliffs, NJ.

Church, R.L., 2008. BEAMR: an exact and approximate model for the *p*-median problem. Computers & Operations Research 35, 417–426.

Church, R.L., Murray, A.T., Weintraub, A., 1998. Locational issues in forest management. Location Science 6, 137–153.

Church, R.L., ReVelle, C.S., 1974. The maximal covering location problem. Papers of the Regional Science Association 32, 101–118.

Church, R.L., Scaparra, M.P., 2007. Protecting critical assets: the r-interdiction problem with fortification. Geographical Analysis 39, 129–146.

Church, R.L., Scaparra, M.P., Middleton, R.S., 2004. Identifying critical infrastructure: the r-interdiction median problem and the r-interdiction covering problem. Annals of the Association of American Geographers 94, 491–502.

Densham, P.J., Rushton, G., 1992. A more efficient heuristic for solving large *p*-median problems. Papers in Regional Science 71, 307–329.

Duque, J.C., Anselin, L., Rey, S.J., 2012. The max-p-regions problem. Journal of Regional Science 52, 397–419.

Garrison, W.L., 1959. Spatial structure of the economy II. Annals of the Association of American Geographers 49, 471–482.

Goodchild, M.F., 1979. The aggregation problem in location-allocation. Geographical Analysis 11, 240–255.

Grubesic, T.H., Murray, A.T., 2006. Vital nodes, interconnected infrastructures and the geographies of network survivability. Annals of the Association of American Geographers 96, 64–83.

Grubesic, T.H., Matisziw, T.C., Murray, A.T., 2008. Comparative approaches for assessing network vulnerability. International Regional Science Review 31, 88–112.

Hakimi, S.L., 1965. Optimum distribution of switching centers in a communication network and some related group theoretic problems. Operations Research 13, 562–574.

Kemp, P.S., O'Hanley, J.R., 2010. Procedures for evaluating and prioritising the removal of fish barriers: a synthesis. Fisheries Management and Ecology 17, 297–322.

Kent, B., Bare, B.B., Field, R., Bradley, G., 1991. Natural resource land management planning using large-scale linear programming: the USDA forest service experience with FORPLAN. Operations Research 39, 13–27.

Kuby, M.J., 1989. A location-allocation model of Losch's central place theory-testing on a uniform lattice network. Geographical Analysis 21, 316–337.

Kuby, M.J., Bielicki, J.M., Middleton, R.S., 2011. Optimal spatial deployment of CO_2 capture and storage given a price on carbon. International Regional Science Review 34, 285–305.

Lemberg, D.L., Church, R.L., 2000. The school boundary stability problem over time. Socio-economic Planning Sciences 34, 159–176.

Ligmann-Zielinska, A., Church, R.L., Jankowski, P., 2008. Spatial optimization as a generative technique for sustainable multi-objective land-use allocation. International Journal of Geographical Information Science 22, 601–622.

Lombard, K., Church, R.L., 1993. The gateway shortest path problem: generating alternative routes for the corridor location problem. Geographical Systems 1, 22–45.

O'Kelly, M.E., Miller, H.J., 1994. The hub network design problem: a review and synthesis. Journal of Transport Geography 2, 31–40.

Openshaw, S., 1997. A geographical solution to the scale and aggregation problems of region-building, partitioning and spatial modeling. Transactions of the Institute of British Geographers New Series 2, 459–472.

ReVelle, C.S., Swain, R.W., 1970. Central facilities location. Geographical Analysis 2, 30–42.

Thill, J.C., Dao, T.H.D., Zhou, Y., 2011. Travelling in the three-dimensional city: applications in route planning, accessibility assessment, location analysis and beyond. Journal of Transport Geography 19, 405–421.

Tong, D., Murray, A.T., Xiao, N., 2009. Heuristics in spatial analysis: genetic algorithm for coverage maximization. Annals of the Association of American Geographers 99, 698–711.

Tong, D., Murray, A.T., 2012. Spatial optimization in geography. Annals of the Association of American Geographers 102, 1290–1309.

Upchurch, C., Kuby, M., Lim, S., 2009. A model for location of capacitated alternative-fuel stations. Geographical Analysis 41, 85–106.

Van Reidhead, V.A., 1979. Linear programming models in archaeology. Annual Review of Anthropology 8, 543–578.

Von Thunen, J.H., 1842. Der Isolierte Staat. Pergamon, London.

Weigel, D., Cao, B., 1999. Applying GIS and OR techniques to solve Sears technician-dispatching and home-delivery problems. Interfaces 29, 112–130.

Wesolowsky, G.O., 1993. The Weber problem: history and procedures. Location Science 1, 5–24.

Williams, J.C., 2002. A zero-one programming model for contiguous land acquisition. Geographical Analysis 34, 330–349.

Wright, J., ReVelle, C., Cohon, J., 1983. A multi-objective integer programming model for the land acquisition problem. Regional Science and Urban Economics 13, 31–53.

Yeates, M., 1963. Hinterland delimitation: a distance minimizing approach. Professional Geographer 15, 7–10.

Spatial Pattern, Analysis of

Lauren M Scott, Esri, Redlands, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Boots, volume 22, pp. 14818–14822, © 2001, Elsevier Ltd.

Abstract

Spatial pattern analysis methods provide insights about where things occur, how the distribution of incidents or the arrangement of data aligns with other features in the landscape, and what the patterns may reveal about potential connections and correlations. Focusing primarily on vector data (points, lines, and areas), this article introduces a sampling of descriptive and inferential exploratory spatial pattern analysis methods relevant to a variety of real-world problems and questions. Web links to additional resources for learning more about spatial pattern analysis are included in tables at the end of the article.

Spatial pattern analysis involves identifying, describing, and measuring the shape, arrangement, location, configuration, trend, or relationships in geographic data. Data are spatial or *geographic data* when they are collected with, or assigned, locations such as an X and Y coordinates (longitude and latitude) or an address. When data are associated with locations, they can be mapped, and mapping geographic data is an important first step in analyzing spatial patterns.

The next steps involve identifying, describing, and measuring the spatial characteristics of data. Is there a discernible pattern, order, or structure associated with a disease outbreak, a series of public protests, the results of state-wide test scores, or emergency call events? Researchers and analysts look for spatial patterns in data because often understanding *where* things occur provides important clues about *why* things occur. Understanding why (i.e., cause–effect relationships) has been described as the 'holy grail' of science (Cressie and Wikle, 2011), because it offers the potential for changing outcomes: remediation, intervention, alteration. When you understand the patterns associated with a series of crime events, for example, you have an important tool for predicting future crimes. Examining the patterns associated with consumer behavior can help focus marketing efforts to increase sales. Analyzing spatial variations in childhood obesity can help inform policies and programs aimed at reducing this growing epidemic.

Sometimes looking at a map is all that is required to identify and explain spatial patterns. It is important to recognize, however, that the cartographic decisions made when a map is created can have a big impact on what a map communicates (Monmonier, 1996). The maps below (**Figure 1**), for example, show standardized health care costs by hospital regions in the contiguous United States. All three maps are based on the same underlying data; the only difference is the cartographic display (class breaks based on equal intervals, quantiles, or standard deviations).

Consequently, when visual analysis is not sufficient to identify or to explain spatial patterns or when you need to quantify or to make inferences about the patterns in spatial data, you will want to adopt a more formal analytical approach. Focusing primarily on vector data (points, lines, and areas), this article presents a small sampling of the many spatial pattern analysis methods that can be used to move beyond visual analytics in order to describe, quantify, and effectively map spatial patterns. The first section covers simple descriptive pattern analysis methods. The second and third sections describe inferential methods for quantifying and mapping spatial patterns. The article concludes with some comments on future directions and links to web pages with more details, the math behind the methods, and the resources for learning more.

Describing Spatial Patterns

Descriptive spatial pattern analysis methods summarize the key characteristics and properties for a set of geographic features and their associated attribute values. **Table 1** provides examples of some geographic features, how they might be represented on a map, and potential attributes.

If asked to analyze a set of numbers such as household incomes, one might begin by computing the mean and the standard deviation. Similarly, analyzing spatial data often begins with computing spatial versions of the same statistics: the mean center and the standard deviational ellipse. Calculating the mean center helps understanding the central tendency of the features, and calculating the standard deviational ellipse highlights the spatial distribution and orientation of features around the mean center. Using these descriptive pattern analysis methods to summarize important spatial properties in data is an effective strategy for comparing groups of features across space (e.g., race/ethnicity integration or segregation) or highlighting changes in spatial patterns over time (e.g., urbanization trends).

Central Tendency

Points and Areas

Measures of central tendency are simple descriptive statistics that can help answer questions such as:

- Which resource is most accessible?
- Has the center of innovation changed over time?
- What is the primary direction of migration?

The most common measure of central tendency is the mean center statistic. It is computed as the average X coordinate and

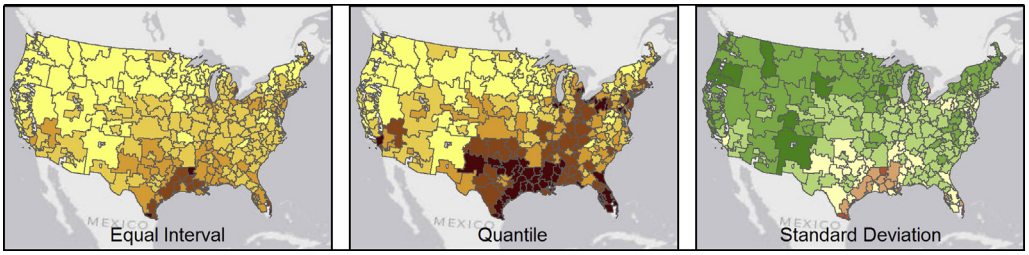

Figure 1 Geographic variations in per capita health care costs for Medicare recipients by hospital referral region, 2010. Data source: http://www.iom.edu/Activities/HealthServices/GeographicVariation/Data-Resources.aspx.

the average Y coordinate for all features in a data set. Although it is a simple statistic, it has powerful applications. Epidemiologists, for example, might compute the mean center for a set of health events in order to gather clues about the potential sources of a disease or to better understand mechanisms of disease transmission (Koch, 2005).

When the mean center is calculated using an attribute value as a weight, the effect is to pull the mean center toward the features with the largest attributes. A retail analyst looking for potential warehouse locations to serve a chain of grocery stores, for example, might compute a weighted mean center using the sales volume attribute for each store. By including this attribute in the mean center calculation, the X and Y coordinates for stores with the largest sales volumes will have the strongest impact on the resulting location of the weighted mean center (Mitchell, 2005).

In **Figure 2**, the weighted mean center for the population of California has been computed each decade between 1900 and 2000. In the first part of the century, the mean center is near San Francisco, perhaps reflecting the maritime trade and earlier population booms associated with the 1848 discovery of gold in the region. With each decade, however, the weighted mean center moves south owing to stronger population growth in southern California. This southern-moving trend is likely a reflection of the economic opportunities associated with the burgeoning movie and oil industries, as well as the increasing availability of air-conditioning (Scott and Janikas, 2010).

Lines

For line features, central tendency is computed using a circular statistic to summarize either the mean direction or the mean orientation. Given a set of lines representing animal movements, for example, calculating the mean direction will distinguish between the spring and fall migration pathways. Alternatively, calculating the mean orientation for data spanning all four seasons would be appropriate if the main focus is identifying the key corridors that migrating animals use. While a mean direction is often expressed as a line with one arrowhead, the mean orientation is general represented by a line with two arrowheads (**Figure 3**).

Other measures summarizing line features include the mean line length, mean center point, and circular variance. The *circular variance* is analogous to a standard deviation. It ranges from 0 to 1 and indicates how similar or how different the angles for a set of lines are from each other. Consider an analysis of lines at locations throughout southern Australia representing primary wind directions in January. If everywhere in the study area the wind blows consistently to the northeast, the circular variance will be near 0. If, on the other hand, the wind direction changes drastically from one location to the next so that it blows in many different directions depending on location within the study area, the circular variance will be near 1.

Applications for mean orientation include analyzing roads, urban spatial structure, or fault lines (Mardia and Jupp, 2000; Gaile and Burt, 1980). Mean direction applications include analyzing the consistency of hurricane pathways, migration routes for animals or people, or regional wind patterns. Mean direction has been used to analyze geographic variations in journey to work (Corcoran et al., 2009), to model journey to crime (Van Patten and Delhauer, 2007), and, more broadly, to describe patterns of transport, spatial interaction, flows, or diffusion (Gaile and Burt, 1980).

Spatial Distribution

While the mean center and linear mean pattern analysis methods provide information about central tendency,

Table 1 Examples of spatial features, geometric representation, and feature attributes

Features	*Typical geometry*	*Example attributes*
Cities	Points representing the center of a city or areas representing city boundaries	Unemployment rate, jobs, Hispanic population
Traffic accidents	Points	Severity, fatalities, date, time
Roads	Lines	Speed limit, traffic volume, road condition
Parks	Areas or points	Acreage, entry fee, estimated annual visitors
Counties	Areas	Population, average home values, median incomes

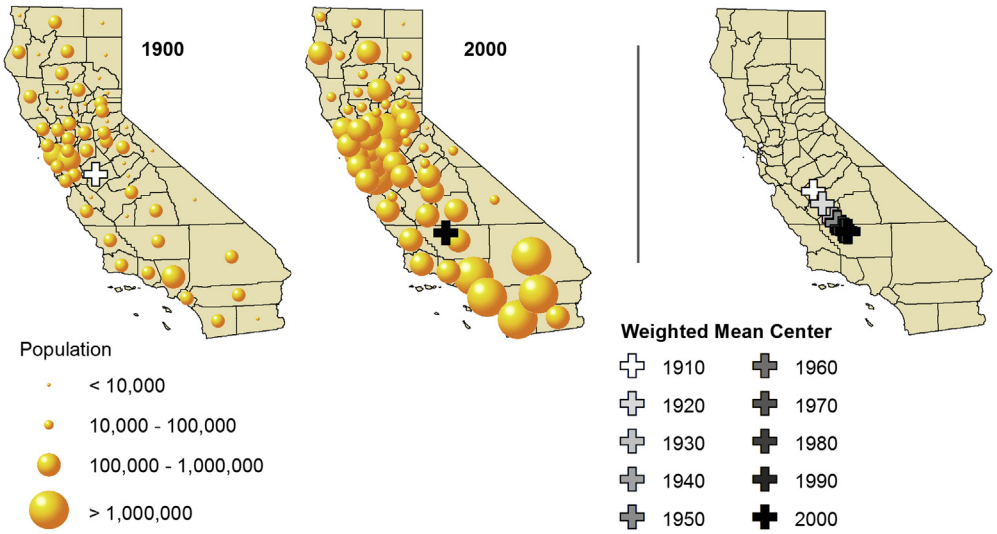

Figure 2 The center of population moves steadily south between 1900 and 2000 as growth in southern California outpaces population growth in northern California. Adapted with permission from Scott, Lauren M., Janikas, Mark V., 2010. Spatial statistics in ArcGIS. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin. Data source: http://www.census.gov/population/cencounts/ca190090.txt.

descriptive spatial statistics like the standard deviational ellipse summarize the distribution and orientation of features around the mean center. These pattern analysis tools help answer questions like:

- How quickly and widely is a disease spreading through the village?
- Where are the core areas for particular language dialects? How integrated are they and which have the broadest spatial extent?

Figure 4 shows an analysis of gang territories based on graffiti locations for cases where the gangs responsible could be identified. Each color reflects a different gang affiliation. Creating a standard deviational ellipse for each set of these gang-related incidents provides information about the core extent of gang territories and how the territories relate to each other and to various features in the landscape. Further, areas where the ellipses overlap are a good indication of locations with higher risk for gang-related conflict. This type of spatial pattern analysis can be very useful for making decisions about how best to allocate scarce police resources.

The size of the ellipse created is a function of the standard deviation of both the X and the Y feature coordinates, rotated to minimize the distance between the points and the axes of the ellipse. Analytics using standard deviational ellipses include creating a spatial index of segregation (Wong, 1999), targeting resource distribution for disease control (Eryando et al., 2012), and predicting crime (Kent and Leitner, 2007; Levine, 2002).

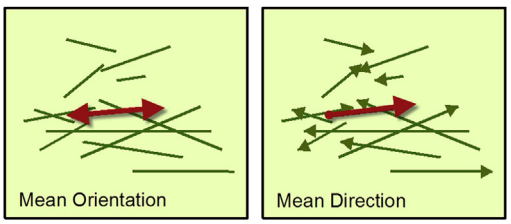

Figure 3 The linear mean statistic can assess overall direction or orientation.

Quantifying Spatial Patterns

Beyond the simple descriptive pattern analysis methods, a number of inferential statistical tests can effectively quantify

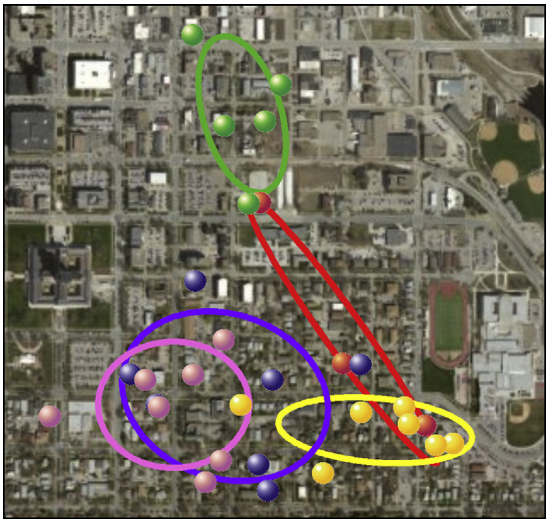

Figure 4 Standard deviational ellipses used to identify core gang territories. Adapted with permission from Scott, Lauren M., Janikas, Mark V., 2010. Spatial statistics in ArcGIS. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin. Data source: simulated based on expert knowledge.

spatial patterns. These methods compare the observed pattern of features, or feature attributes, against a theoretical spatial pattern. Most often, the theoretical pattern used as a baseline is complete spatial randomness (CSR) and, consequently, the null hypothesis for these statistical tests is that the observed pattern is the result of underlying spatial processes that are random in nature. These methods calculate the probability, expressed as a *p*-value, that the null hypothesis is correct. A small *p*-value (<0.05 or, more conservatively, <0.01) indicates statistically significant spatial patterns in the data, providing evidence that nonrandom spatial processes are at work. Only nonrandom spatial processes suggest cause–effect relationships and, consequently, indicate possible opportunities for remediation or intervention.

In addition to calculating *p*-values, these tools often report corresponding z-scores to indicate how the observed spatial patterns deviate from CSR. Spatial patterns that are statistically significant because they are more clustered than CSR generally reflect infectious, transmittable, or attracting types of underlying local spatial processes. For example, it is not uncommon for the retail, financial, and industrial activities in a city to cluster the result of zoning ordinances and/or to take advantage of economies of scale or economies of agglomeration. Spatial patterns that are statistically significant because they are more dispersed than CSR generally result from some form of competitive process where features repel each other in order to create distance. Only small trees can grow under big trees, for example. The small trees will be clustered as a result of seed transmission processes but will thin out over time, the result of competition for local resources.

A number of inferential pattern analysis methods focus specifically on the spatial pattern of point features. A common, decades-old method widely used in geography and ecology is the average nearest neighbor statistic. It works by computing the mean distance to each feature's closest neighbor and then comparing that distance to the distance that would be obtained for a random distribution of those same features (Mitchell, 2005). Another method, quadrat analysis, overlays point features with a fishnet mesh and counts the number of points occurring within each quadrat. A chi-square statistic is then used to assess how likely it is that the observed distribution of points among the quadrats might be the result of random spatial processes (Burt et al., 2009). Kernel density techniques extend the quadrat method by considering point intensity within a predefined radius, generally presenting results as a map-based surface (Murray, 2010). The K function is unique in that it looks at spatial clustering and dispersion for a range of distances, providing clues about the spatial scales at which the processes promoting clustering or dispersion are operating (Scott and Janikas, 2010; Csillag and Boots, 2005). These point pattern analysis methods can help answer questions like:

- Which types of businesses are most concentrated?
- Does the spatial pattern of a disease mirror the spatial pattern of the population at risk?

More common in the social and behavioral sciences are measures that consider both location and feature attributes. These methods typically operate on area features such as census tracts, parcels, or cities, where data have been aggregated to represent counts or rates. They provide a single summary statistic that quantifies *spatial autocorrelation* or the degree to which features nearby in space are similar with regard to a measured attribute (Getis, 2010). Consequently, global measures of spatial autocorrelation such as the Moran's I statistic are especially effective for understanding broad spatial patterns and trends (Mitchell, 2005). They can help answer questions like:

- Is there an unexpected spike in the number of riots or public protests?
- Is the new virus remaining geographically fixed or is it spreading to surrounding communities?

The thematic maps below (**Figure 5**) illustrate relative per capita income, by county, for the state of New York over several years. The darker areas reflect higher incomes. Unfortunately, a visual inspection of these maps is not sufficient to determine whether spatial segregation among the rich and poor counties is increasing or decreasing. It is difficult to discern the answer because doing so requires visual detection of both global and local differences among multiple locations over several years. Using the global Moran's I tool to quantify the degree of clustering for each year, however, and then creating a line graph of the z-score results, makes the overall trend clear (**Figure 5**). A higher z-score reflects more intense clustering. Consequently, the resulting sinking line of the z-score graph indicates that clustering is dissipating. This pattern analysis provides some evidence that spatial segregation among the New York counties has decreased overall between 1969 and 2002 (Scott and Janikas, 2010).

Similar to the K function, the global Moran's I statistic can be used to assess clustering across multiple spatial scales. By running the global Moran's I statistic for increasing distances and plotting the resultant z-scores, a picture of spatial autocorrelation emerges (**Figure 6**). Peak z-scores indicate distances where the processes promoting spatial clustering are most pronounced, and these distances often serve as appropriate scales of analysis for the local spatial autocorrelation methods discussed next.

Mapping Spatial Patterns

Global spatial autocorrelation measures, like the Moran's I statistic discussed in the previous section, help to quantify broad trends in data by computing a single indicator to summarize the statistical likelihood that the observed data reflect a random spatial pattern. They answer the question: is there a discernible spatial pattern or not? For geographic analyses, however, the more interesting question is often: where are the statistically significant spatial patterns located? Where are the hot spots, cold spots, and spatial anomalies? To address questions like these, local versions of the global spatial autocorrelation statistics, such as the Getis-Ord Gi* and the Anselin Local Moran's I, have been developed (Anselin, 1995; Ord and Getis, 1995; Getis, 2010). These pattern analysis methods consider each feature within the context of nearby neighboring features and compare local statistical properties of an attribute to global properties. If the local mean for a feature and its neighbors is unexpectedly large compared to

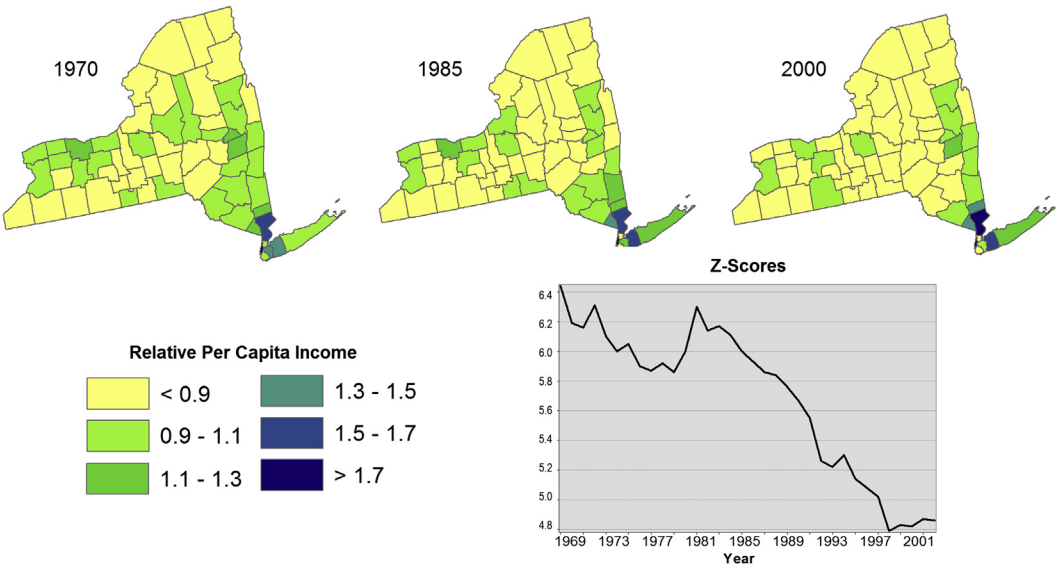

Figure 5 Relative per capita incomes by county for New York. Adapted with permission from Scott, Lauren M., Janikas, Mark V., 2010. Spatial statistics in ArcGIS. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin. Data source: Bureau of Economic Analysis, www.bea.gov/regional/.

the global mean (i.e., the mean value for all features in the data set), it is indicative of a *hot spot* (a spatial cluster of high values). Similarly, an unexpectedly small local mean suggests a *cold spot* (a spatial cluster of small values). An individual feature with a large attribute value surrounded by features with small values, or a feature with a small value surrounded by features with large values, reflects spatial anomalies. Accounting for the number of features and global variation in the attribute value being analyzed, these methods determine if the differences between the local and global properties are statistically significant, and return a result (a *p*-value and/or z-score) for every feature in the data set. Mapping these results effectively illustrates the spatial patterns (Aldstadt, 2010).

Local spatial autocorrelation statistics are used in a number of application areas including crime (Ratcliffe, 2010; Scott and Warmerdam, 2005) and health (Haque et al., 2012; Getis, 2004). In **Figure 7**, the Anselin Local Moran's I statistic is applied to data from the 2008 United States presidential election. The map shows statistically significant clusters of the Republican and the Democratic political party victories as well as spatial anomalies. A political strategist could use the results from this type of pattern analysis to help design more effective canvassing. Active campaigning in counties without strong allegiances to either party may be more effective than investing campaign money in counties where the opposing party has had strong wins. The spatial outlier (**Figure 7**) in the southern part of Texas reflects a Republican Party win surrounded by counties strongly supporting the Democratic Party. Understanding what makes the political leanings in this county so different from

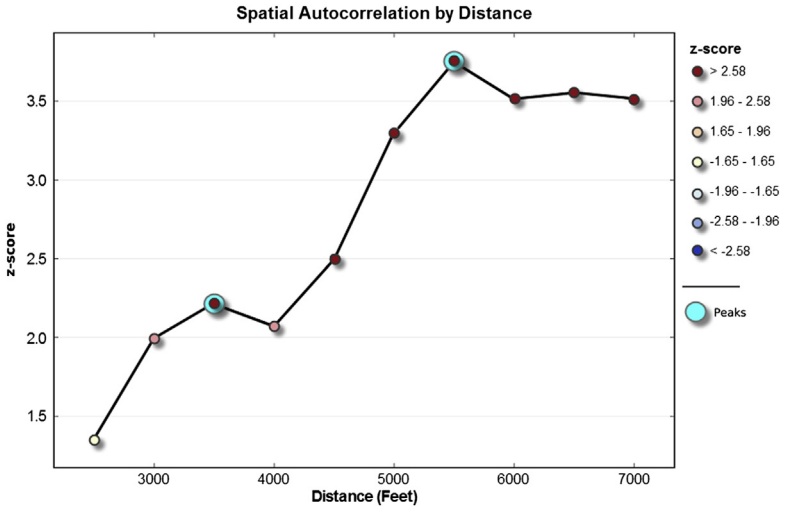

Figure 6 Analyzing the spatial pattern of crime, the peak z-scores at 3500 and 5500 feet reflect distances where the processes promoting clustering are most pronounced. Analysis is based on data obtained from https://data.cityofchicago.org/.

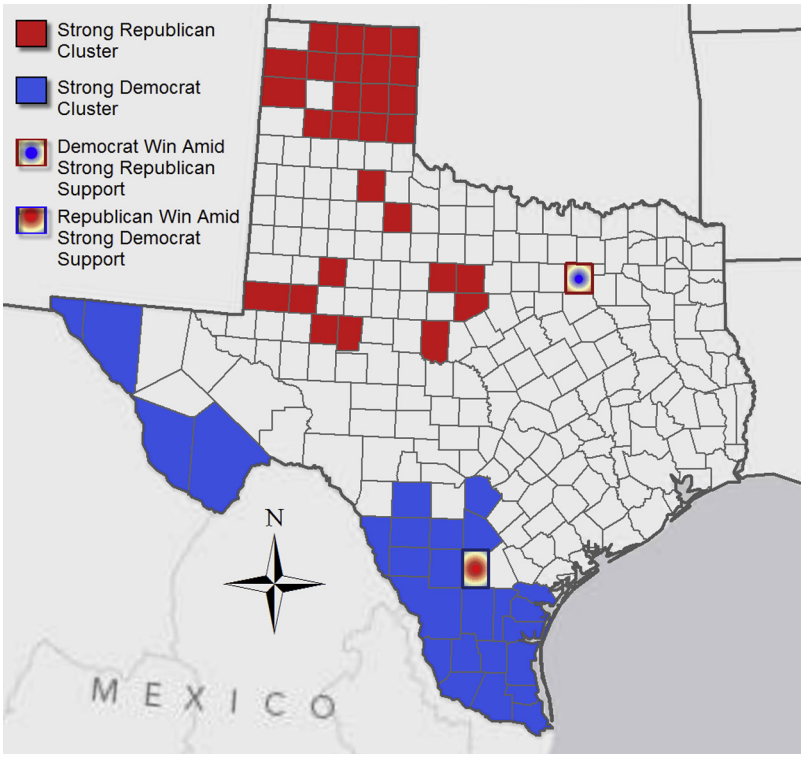

Figure 7 Hot spot analysis of the 2008 presidential election results by county for Texas. Data source: National Atlas of the United States.

its neighbors may help strategists design targeted messaging to win votes there.

Another approach to cluster analysis identifies groups of features based on inherent similarities for a set of specified attributes. The most commonly used methods rely on K-Means algorithms, but a number of other methods have also been developed (Murray, 2010; Duque et al., 2007), including methods that impose both spatial and temporal constraints on group membership (Esri, 2012). The goal of these methods is to classify features into natural groupings such that both the within-group similarity and the between-group differences are maximized. These methods have been used to study regional economics (Duque et al., 2006) and have been proposed for analyzing cancer rates (Guo and Wang, 2011). While these methods have been used most often in the physical sciences (e.g., climate zoning, landscape analysis) and for image classification, they have much broader application whenever there is a need to create homogeneous regions or to elucidate hidden patterns in spatial data (Miller, 2009; Guo and Wang, 2011). With increased availability of high-resolution spatial data, these types of analytics will be needed for understanding human–environment interactions and socioeconomic dynamics, and to address critical problems such as global climate change (Guo and Mennis, 2009).

Future Directions

Looking into the future of spatial pattern analysis methods and applications is both exciting and overwhelming. According to IBM (Selinger, 2013), 90% of the data today has been generated in the past 2 years, and new data are being created more quickly than are algorithms capable of processing them. Sources of these data include purchase transactions, sensors monitoring climate, and other aspects of the physical world, vehicle electronic transmissions (cars, planes, and ships, for example), utility meter readings, a host of data generated from social media sources (e.g., digital photographs, videos, blogs, tweets), and Global Positioning System-enabled mobile devices (Selinger, 2013). As a consequence, the ability to effectively describe, quantify, and map spatial patterns is becoming more important than ever. The challenge is not only dealing with this large volume of messy, dynamic, streaming data but also effectively analyzing its multidimensional characteristics.

Even prior to analytical considerations, however, methods for filtering data to eliminate noise, detect and remove errors, reduce redundancy, compensate for missing data, and deal with uncertainty will need to be refined. These filtering methods are a precursor to powerful analytics that must be capable of detecting and then predicting unexpected patterns, correlations, and trends. Effective pattern analysis techniques are not yet well developed for dealing with space, time, and mobility simultaneously. Methods that can detect coevolving spatial–temporal patterns – various event series that have similar tracks (similar spatial–temporal 'fingerprints') based on mutual locations, social networks, behaviors, opportunities and histories – also need to be developed (Shekhar et al., 2009).

Once these new pattern analysis methods are implemented, however, additional challenges surround how best to communicate the multidimensional results. An important goal

will be to extend descriptive cartographic visualizations (such as animations) by developing mathematical and statistical frameworks. These will be required to detect, for example, significant trends or anomalies from the increasing information overload in order to motivate effective action (Murray, 2010). In summary, development of new spatial pattern analysis methods, algorithms, and applications represents vital areas of research where scholars and analysts in the social and behavioral sciences can make enormous contributions.

See also: Access: Geographical; Applied Geography: A US Perspective; Diffusion: Geographical Aspects; Electoral Geography; Geographic Information Systems; Geography: Overview; Global Migration; Health, Geography of; Marketing Strategies; Mathematical Models in Geography; Population Geography; Race and Racism, Geography of; Regional Science; Residential Segregation: Geographic Aspects; Scale in Geography; Social Geography; Spatial Analysis in Geography; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Data; Spatial–Temporal Modeling; Time-Space in Geography; Urban Geography; Urbanization.

Bibliography

Aldstadt, Jared, 2010. Spatial clustering. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin, pp. 279–298.

Anselin, Luc, 1995. Local indicators of spatial association – LISA. Geographical Analysis 27 (2), 93–115.

Burt, James E., Barber, Gerald M., Rigby, David L., 2009. Elementary Statistics for Geographers, third ed. The Guilford Press, New York.

Corcoran, Jonathan, Chhetri, Prem, Stimson, Robert, 2009. Using circular statistics to explore the geography of the journey to work. Papers in Regional Science 88 (1), 119–132.

Cressie, Noel, Wikle, Christopher K., 2011. Statistics for Spatio-temporal Data. John Wiley & Sons, Inc., Hoboken, New Jersey.

Csillag, F., Boots, B., 2005. A framework for statistical inferential decisions in spatial pattern analysis. Canadian Geographer 49 (2), 172–179.

Duque, J.C., Artis, Manuel, Ramos, Raul, 2006. The ecological fallacy in a time series context: evidence from Spanish regional unemployment rates. Journal of Geographic Systems 8, 391–410.

Duque, J.C., Ramos, R., Surinach, J., 2007. Supervised regionalization methods: a survey. International Regional Science Review 30, 195–220.

Eryando, Tris, Dewi, Susanna, Pratiwi, Dian, Nugraha, Fajar, 2012. Malaria Journal 11 (1), 130.

Esri, 2012. How Grouping Analysis Works. http://resources.arcgis.com/en/help/main/10.1 (accessed 28.03.13).

Gaile, G.L., Burt, James E., 1980. Directional statistics. Concepts and Techniques in Modern Geography (25). Geo Abstracts, Norwich.

Getis, Arthur, 2004. A geographic approach to identifying disease clusters. In: Janelle, D.G., Warf, B., Hansen, K. (Eds.), Worldminds: Geographical Perspectives on 100 Problems. Springer, Dordrecht, pp. 81–86.

Getis, Arthur, 2010. Spatial autocorrelation. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin, pp. 255–278.

Guo, Diansheng, Mennis, Jeremy, 2009. Spatial data mining and geographic knowledge discovery – an introduction. Computers, Environment, and Urban Systems 33, 403–408.

Guo, Diansheng, Wang, Hu, 2011. Automatic region building for spatial analysis. Transactions in GIS 15 (1), 29–45.

Haque, Ubydul, Scott, Lauren M., Hashizume, Masahiro, Fisher, Emily, Haque, Rashidul, Yamamoto, Taro, Glass, Gregory E., 2012. Modeling malaria treatment practices in Bangladesh using spatial statistics. Malaria Journal 11 (1), 63. http://www.malariajournal.com/content/11/1/63 (accessed 28.03.13).

Kent, Joshua, Leitner, Michael, 2007. Efficacy of standard deviational ellipses in the application of criminal geographic profiling. Journal of Investigative Psychology and Offender Profiling 4, 147–165.

Koch, Tom, 2005. Cartographies of Disease: Maps, Mapping, and Medicine. ESRI Press, Redlands, CA.

Levine, Ned, 2002. CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v. 20). Ned Levine & Associates and the National Institute of Justice.

Mardia, Kanti V., Jupp, Peter E., 2000. Directional Statistics. John Wiley and Sons, Inc., New York.

Miller, Harvey, 2009. Geocomputation. In: Fotheringham, S.A., Rogerson, P.A. (Eds.), The SAGE Handbook of Spatial Analysis. Sage, Los Angeles, pp. 397–418.

Mitchell, Andy, 2005. The ESRI Guide to GIS Analysis. In: Spatial Measures and Statistics, vol. 2. ESRI Press, Redlands, CA.

Monmonier, Mark, 1996. How to Lie with Maps, second ed. University of Chicago Press, Chicago.

Murray, Alan T., 2010. Quantitative geography. Journal of Regional Science 50 (1), 143–163.

Ord, J. Keith, Getis, Arthur, 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis 27 (4), 287–306.

Ratcliffe, Jerry, 2010. Crime mapping: spatial and temporal challenges. In: Piquero, A.R., Weisburd, D. (Eds.), Handbook of Quantitative Criminology. Springer, New York, pp. 5–24.

Scott, Lauren M., Janikas, Mark V., 2010. Spatial statistics in ArcGIS. In: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer-Verlag, Berlin.

Scott, L.M., Warmerdam, N., 2005. Using spatial statistics for crime analysis with ArcGIS 9. ESRI ArcUser Magazine. http://www.esri.com/news/arcuser/0405/ss_crimestats1of2.html (accessed 28.03.13.).

Selinger, David, 2013. Big Data: Getting Ready for the 2013 Big Bang. www.forbes.com/sites/ciocentral/2013/01/15/big-data-get-ready-for-the-2013-big-bang/ (accessed 20.11.2013.).

Shekhar, Shashi, Gandhi, Vijay, Zhang, Pusheng, Vatsavai, Ranga Raju, 2009. Availability of spatial data mining techniques. In: Fotheringham, A.S., Rogerson, P.A. (Eds.), The SAGE Handbook of Spatial Analysis. Sage Publications, London.

Van Patten, Isaac T., Delhauer, Paul Q., 2007. Sexual homicide: a spatial analysis of 25 years of deaths in Los Angeles. Journal of Forensic Science 52 (5), 1129–1141.

Wong, David, 1999. Geostatistics as measures of spatial segregation. Urban Geography 20 (7), 635–647.

Relevant Websites

http://resources.arcgis.com/en/help/ – Math and additional details for the Local Moran's I statistic.

http://resources.arcgis.com/en/help/ – Math and additional details about spatially constrained cluster analysis.

www.esriurl.com/spatialstats – Analyzing the Spatial Pattern of Piracy.

http://www.census.gov/geo/www/cenpop/MeanCenter.html – US Census Bureau maps and animations tracing the center of population over time.

http://resources.arcgis.com/en/help/ – Math and additional details about the Global Moran's I statistic.

http://resources.arcgis.com/en/help/ – Math and additional details about the local K Function.

http://resources.arcgis.com/en/help/ – Math and additional details about the Linear Mean and Circular Variance statistics.

http://resources.arcgis.com/en/help/ – Math and additional details about the Mean Center and Weighted Mean Center statistic.

http://resources.arcgis.com/en/help/ – Math and additional details for the Standard Deviational Ellipse method.

www.esriurl.com/spatialstats, https://geodacenter.asu.edu/eslides – Spatial pattern analysis tutorials and free web seminars.

http://resources.arcgis.com/en/help/ – Additional information about p-values and z-scores.

Spatial Sampling

Robert Haining, Human Geography, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial sampling is undertaken to estimate an attribute of a spatially distributed population. The attribute might relate to a quantity (how much) or to the distribution of values (where). Two models underlie spatial sampling, the design-based and model-based approaches, and these are described. A key issue with spatial sampling is selection of the sample sites. We describe random, stratified random, and systematic spatial sampling and comment on their properties. Other forms of sampling, including clustered and nested sampling as well as progressive sampling including spatially adaptive sampling, are described. We discuss the implications of spatial heterogeneity and spatial autocorrelation for spatial sampling.

Sampling is undertaken in order to estimate one or more attributes of a population. Sampling is used instead of undertaking a census for many reasons. For example, a census may be impossible because the population is very large, perhaps even infinitely large, or it may be impractical because of cost. Some UK census data are made available only as a 10% sample partly to help preserve confidentiality but also because the data require manual coding, which is expensive. A census may be unnecessary because a value for the attribute is only required up to some predetermined level of precision and a sample size and estimator can be specified that will achieve the required level of precision. Sampling is also undertaken to improve the accuracy of databases because some methods of recording are known to produce undercounts (e.g., in the case of crimes, not all events are reported to the police) and some methods of enumeration miss certain groups in the population (e.g., the homeless in the case of a national census).

The term 'spatial sampling' is used to describe the activity of sampling when the population is spatially or geographically distributed. All of the individuals in the population have a geo-reference or geo-code that identifies their location with respect to some local or global frame of reference. The population might be a surface of land-use values or ground-level air pollution. In this case the population is the set of all points in the region and is infinitely large. The population might be the set of all households in a city, in which case the population, while not infinite, is nonetheless very large. The population might refer to spatial aggregates. For example, the population of households might be bundled into census units, so that the population is the set of such units for which aggregate quantities are of interest.

Spatial sampling is usually undertaken by recording data values at point locations such as positions on a line or surface, or recording some aggregate quantity for polygons (e.g., pixels, quadrats, or census tracts). In household sampling, a household may be treated as a point (is the head of a household unemployed?) or as an area-based quantity (number of persons per room).

Spatial sampling to estimate a population attribute involves (1) selecting the sample size (n); (2) specifying the sample locations ($s_1,\ldots,s_n$) according to some set of rules, often called the sampling plan; (3) choosing an estimator (for the population attribute); and (4) calculating the sampling error of the estimate. The sampled data $\boldsymbol{z} = \{z(s_1),\ldots,z(s_n)\}$ are used to estimate the attribute, which might be a 'nonspatial' attribute such as the mean value of the variable Z, or the proportion of the population with a given set of values on Z, or the maximum value of Z. These are sometimes referred to as 'how much' questions. The attribute to be estimated might be 'spatial': estimating the autocorrelation or semivariogram structure of the population, obtaining a map of the attribute of interest, or locating extreme or rare events. These are sometimes referred to as 'where' questions. In each case it is necessary to choose an appropriate estimator of the quantity and calculate the sampling error of the estimate.

There may be other issues that affect spatial sampling. There may be, for reasons associated with cost, a trade-off between on the one hand the greater levels of precision (smaller sampling error) obtainable by increasing the sample size versus the costs associated with collecting larger volumes of data. A sample survey that delivers acceptable levels of sampling error for a national survey will not achieve equivalent levels at smaller state or county-level scales without a larger sample and without stratification (by state or county). Hence there needs to be some assessment of the benefits of having information at the smaller scale set against the costs of a larger sample. One solution to this problem, which can be used where an estimate is available at the local level, is to adapt methods based on 'borrowing strength.' National surveys yield precise estimates of national rates, but these are biased estimators of local area rates because local areas have different rates. Estimators of local rates may be unbiased but because of small sample size (if many areas need to be surveyed) they are likely to have low precision. Estimators that 'borrow strength' combine these two sources of information by using weights that reflect the relative precision of the two estimators (Ghosh and Rao, 1994; Longford, 1999). Instead of borrowing strength from the national estimate, information can be borrowed just from geographically adjacent regions, a practice used, for example, in small-area disease-rate mapping (Elliott et al., 1992; Rao, 2005).

There may be particular problems of access to particular sample areas as well as the collection costs associated with adopting a dispersed distribution of sampling locations versus a more clustered distribution. Other issues may also need to be considered, such as whether sampling is required to undertake long-term monitoring and whether there are to be several

measured attributes recorded at each site so that some compromise is required between the different estimators (Haining, 2003). The form of the sampling problem may be to decide how best to add to, or conversely rationalize an existing sampling network where there may be costs associated with any reconfiguration (Arbia and Lafratta, 1997).

There are situations where some spatial object such as a polygon or a line is to be estimated, in which case the sample design is selected in order to estimate or represent characteristics of the object – such as the area of a polygon or the location of a line. In geographic information system databases, the representation and resolution of boundary lines between polygons is the outcome of a sampling process. The precision with which any boundary or shape can be represented is a function of the spacing of the sample points (Griffith, 1989).

Design- and Model-Based Approaches to Spatial Sampling

Suppose a region R is divided into a population of N units, where N might be very large. In the design-based approach it is assumed that $z(s_i)$ is the value of Z observed at the point s_i, which is one of the N points in R. The value of Z at each of the N units is considered fixed. We let $m(R)$ denote the (unknown) mean of the population of these N values. The problem is to estimate $m(R)$, and in the design-based approach Z is observed at n sample points where these sites have been selected according to some sampling plan where each unit has a known probability of being selected. Repeated sampling according to this scheme generates a distribution of values of the vector $\boldsymbol{z}$. This distribution together with an estimator of $m(R)$ defines a (design-based) sampling strategy. This estimator is used to make inferences about $m(R)$ because, as it is based on $\boldsymbol{z}$, it too will have a distribution of values under repeated sampling.

The model-based approach assumes an underlying random process with an assumed probability distribution that is generating the observed set of values in the N units of R (*see* Spatial Autocorrelation; Spatial Statistical Methods). So the value $z(s_i)$ is the value of a random variable – it is only one of a distribution of values that could have been realized at location s_i. It is usual to specify properties of the model generating the N data values including the first-order (mean, μ) and second-order (variance–covariance, $\sum$) properties. Now $m(R)$ is not a parameter but a random variable because it is a function of N random variables. To make clear the distinction, this mean will be labeled as $M(R)$ in the model-based case. Note that it could also be of interest to estimate parameters such as the mean, μ, of the underlying model.

In the model-based approach, the estimator of μ and the estimator of $M(R)$ both incorporate the spatial model generating the data explicitly in order to produce good estimates (e.g., Haining, 1988; Cressie, 1991). For example, the best linear unbiased predictor of $M(R)$ is of the form:

$$\sum_{i=1}^{n} a(i)z(s_i) \qquad [1]$$

where the $\{a(i)\}$ depend on the model and are chosen to minimize the mean square prediction error. In this approach the sampling plan is of secondary interest.

By contrast, in the design-based approach, the sample mean ($a(i) = 1/n$ in eqn [1]) is used to estimate $m(R)$, having been shown to have good properties (Bellhouse, 1977). Attention focuses now on the sampling plan in order to identify which one will give the smallest sampling variance or sampling error for the estimator. The advantage of this approach is that the estimator is easy to compute and it does not depend on specifying the underlying model. The model-based approach assumes a hypothetical superpopulation of possible (but unrealized) maps. The design-based approach is only concerned with the here and now of what has occurred. Perhaps more importantly from a conceptual perspective, the model-based approach can be used to estimate the parameters of the underlying signal (the parameters of the map generator, such as μ) rather than the particular (noisy) map. For example, in relative risk estimation (e.g., disease mapping), it is a quantity such as μ (the underlying mean relative risk) rather than $M(R)$ that is usually of interest (Elliott et al., 1992).

For a fuller discussion of the distinctions discussed in this section, see Brus and deGruijter (1997).

Sampling Plans

There are a number of different ways to select the sites to be sampled, and choice depends on the purpose of the sample and (possibly) the resources available for carrying out the survey. We first discuss random, stratified random, and systematic spatial sampling (Ripley, 1981) before turning to look at other sampling plans.

In the case of random sampling, the n sample sites are selected so that each member of the population has an equal and independent chance of selection. **Figure 1(a)** shows a random sampling plan on a surface where each point on the surface had an equal chance of being selected. **Figure 1(b)** shows a stratified random sampling plan on the same surface that has been divided into nine strata. Within each stratum, one sample site has been selected according to the rule for a random sampling plan, but more than one could be selected (see, for example, Delmelle, 2009: pp. 188–189). **Figure 1(c)** shows one type of systematic sampling – the centric systematic sampling plan. The region has been divided into square strata and the central point in the strata selected. This method of sampling contains no randomization. However, randomization can be introduced by randomizing on the location of the first sample point (the random start method). The other $(n-1)$ sample points occupy the same relative locations in the remaining $(n-1)$ strata. There are several variants of the systematic sampling scheme (Koop, 1990). Stratified random and systematic sampling both raise questions about the size and shape of the strata to use. The hexagon and the triangle are common choices. An understanding of the process generating the quantities of interest may help in deciding these parameters of the sampling plan (Cressie, 1991). For example, the size of the strata is often chosen so that the distances between sample points are large enough so that sample values are independent – in the language of geostatistics, intersample distances are equal to or greater than the range. A pilot survey may be important in helping to decide on

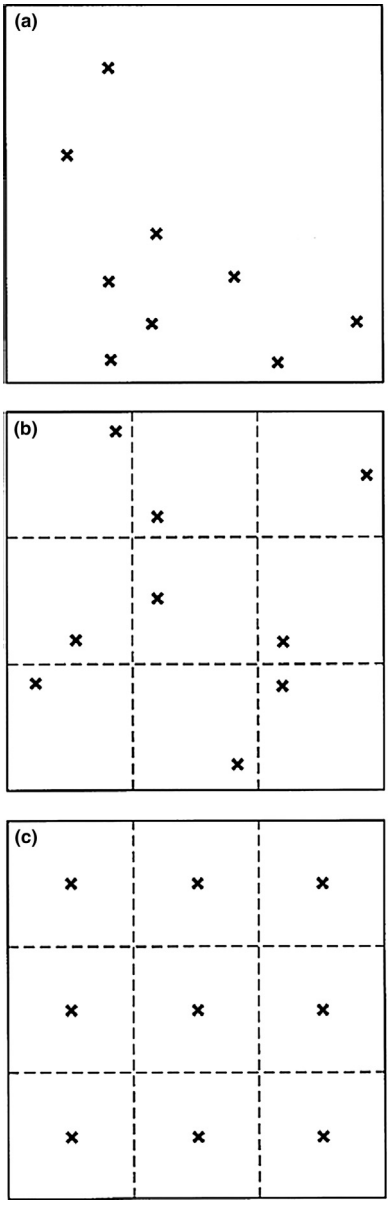

Figure 1 Spatial sampling. (a) Random. (b) Stratified random. (c) Centric systematic.

the size of the strata and hence what density of points is needed.

In addition to these sampling plans there are others that may be appropriate in particular circumstances, although there are sometimes problems associated with computing sampling variances and hence interpreting results. Cluster sampling involves grouping the population into clusters. These clusters are often defined so that subsets of individuals comprising the population are spatially close together within any one cluster. First the clusters are sampled and then the sample is drawn by selecting, usually at random, individuals from within the selected clusters. This sampling method is often cheaper and administratively easier to implement. Kahn and Sempos (1989) give the example of a health survey involving a large number of economically similar but geographically dispersed villages. Instead of drawing the sample of size n from all of the villages (which could involve visiting a large number, incurring considerable time and travel costs), a subsample of villages is drawn and the sample of size n is drawn from within this subset of villages.

Nested sampling may be used where variation in the quantity of interest is associated with different spatial scales – for example, different administrative tiers of a political system. Nested sampling may be employed, again to help keep down the time and cost of the sampling exercise. The area is subdivided at the first (highest) tier of the system and units are chosen at random (similar to cluster sampling). Then these selected areas are subdivided at the second tier and these in turn are sampled – and so on until the lowest tier is reached. Within the selected areas at the lowest tier, individuals are sampled (Figure 2(b)). As can be seen from Figure 2(b), this method does not control for distance between the sample points, so that if scales of variation vary continuously (rather than in terms of discrete hierarchical levels), as might be likely to arise with a surface of values, then this method of sampling may obscure the different scales of variation. Fixed-interval point sampling, as illustrated in Figure 2(a), avoids this problem. The distance between sample points is fixed at those intervals the investigator wishes to study closely (Oliver and Webster, 1986).

When sampling for the purpose of map construction, progressive spatial sampling may be useful. The sampler begins with a coarse grid of sample points across the study area, but in those subareas where spatial variation exceeds a threshold, sampling density is increased in order to better capture gradients (Figure 3). These ideas underlie adaptive spatial sampling, where the location of each new sample point depends on the findings from previously collected samples. Delmelle (2009: pp. 197–199) illustrates cluster adaptive sampling, where the aim is to delimit an area of excess risk such as contaminated land (Figure 4). After a first wave of sampling identifies locations where contamination levels exceed a specified threshold, additional samples are taken around these locations, with further samples beyond these until the level falls below the threshold (Cox, 1999). Haining (2003: pp. 109–111) reviews a Bayesian methodology suggested by Switzer to focus sampling in the search for rare archaeological artifacts.

Spatial Variation and Sampling Plans

Two characteristics of spatial variation have a particular impact on the choice of sampling plan: spatial heterogeneity and spatial autocorrelation. The spatial distribution of an attribute is said to be heterogenous if its properties (e.g., its mean or variance) vary across the study area. The presence of spatial heterogeneity often necessitates the partition of the study area into homogenous or quasihomogenous subareas (Wang et al., 2010). Spatial autocorrelation, on the other hand, and positive spatial autocorrelation in particular, which is the form most often encountered in practical situations, is the tendency for areas or sites that are geographically close together to have similar values. The presence of spatial autocorrelation in a population to be sampled has important implications for

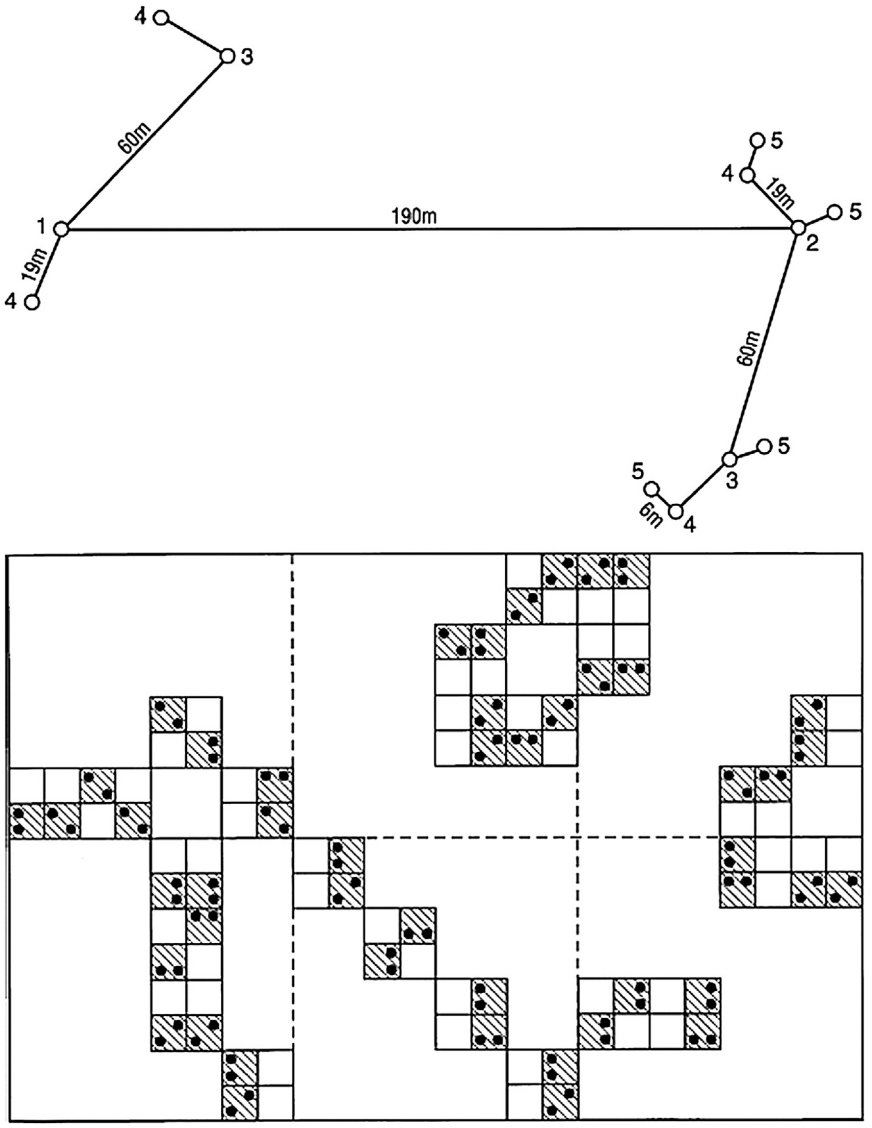

Figure 2 Spatial sampling. (a) Fixed interval. (b) Nested hierarchical.

the sampling variance of the estimator. Before discussing this it is necessary to distinguish between estimating surface properties in the case of the design-based approach (for example, estimating $m(R)$) and estimating underlying parameters in the case of the model-based approach (for example, estimating μ). We consider the design-based approach first.

If the purpose is to estimate $m(R)$, then the presence of positive spatial autocorrelation implies that each sample point carries information not only about its own site but also some information about neighboring site values. Consider the three sampling plans described in Figure 1. If values are positively spatially autocorrelated, one of the problems with random sampling is that because adjacent values will be similar there will be information redundancy in the case of those sample points that happen to be close together. It is also evident from Figure 1(a) that the random sampling plan leaves areas unsampled. Both of these problems with the random sampling plan are reduced to some extent by the introduction of strata, but as is clear in Figure 1(b) there are still some sample points that are close together, resulting in some information redundancy, and there are still some areas of the map not sampled – albeit smaller in extent. Systematic sampling, especially using the centric or the random start methods, reduces both effects. Intuitively we therefore expect that for a given sample size n, systematic sampling ought to produce estimates with lower sampling variances of a quantity such as $m(R)$. This, by and large, proves to be the case, because separating the sample sites in this way reduces possible redundancy in the information content of the sample (which occurs when sample observations are spatially autocorrelated). However, while the gains of moving from random to either stratified random or systematic sampling are considerable, the gains of systematic sampling over stratified random sampling are generally much smaller and seem to depend quite critically on the specific form of the underlying spatial autocorrelation (Ripley, 1981; Matern, 1986; Dunn and Harrison, 1993).

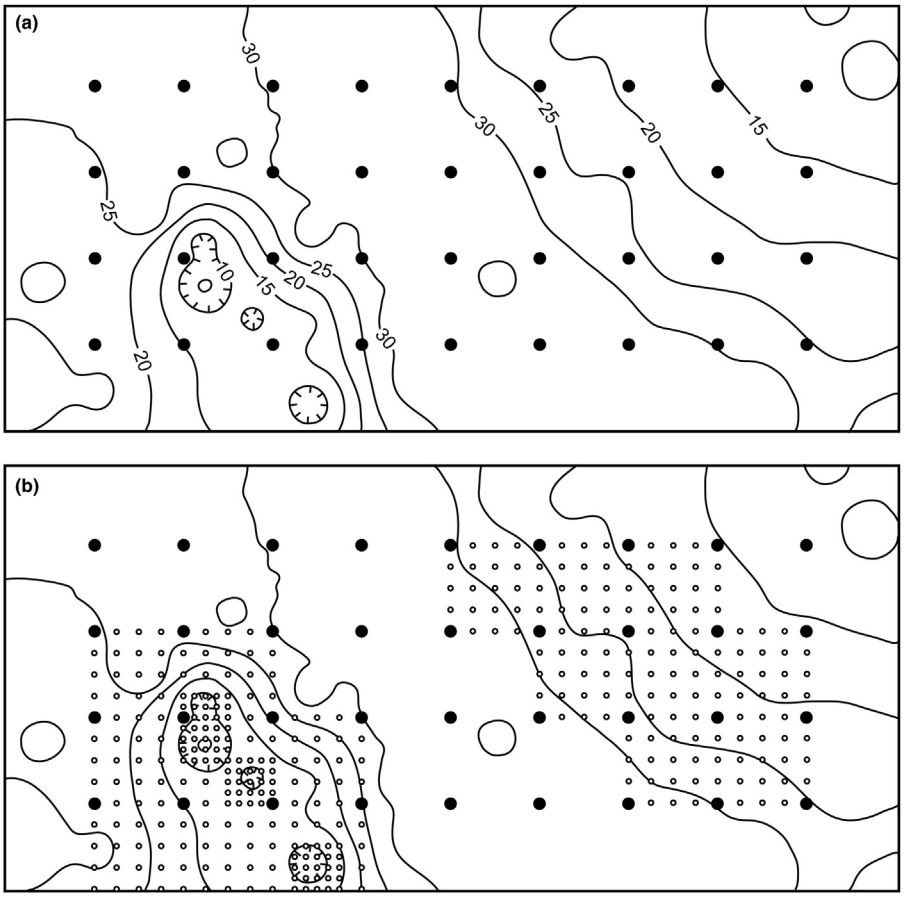

Figure 3 Progressive sampling. Systematic sampling initially with increased sampling density where stronger spatial variation is encountered. Reproduced with permission from Delmelle, E., 2009. Spatial sampling. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 183–206. (Figure 10.5.)

The sampling variance is estimated in the case of random sampling using:

$$\sum_{i=1}^{n}\left[z(s_i)-\overline{Z}\right]^2/(n-1) \qquad [2]$$

where

$$\overline{Z}=\frac{1}{n}\sum_{i=1}^{n}z(s_i)$$

Equation [2] provides an unbiased estimator in this case, and Ripley (1981) discusses related methods for estimating the sampling variance under stratified random and systematic sampling.

The presence of spatial autocorrelation has implications for other sampling plans. In the case of cluster sampling, a critical factor is between-cluster and within-cluster variation. If the clusters are just random assemblages of individuals in the population, then cluster sampling (with random selection of individuals from within each selected cluster) ought not to produce sampling variances any greater than, say, random sampling. If all the clusters differ greatly, then selecting a sample of them will not lead to good population estimates. If clusters have internal positive spatial autocorrelation (so that they are more homogeneous than the population as a whole), then cluster-sampling variance is likely to be bigger than random sampling. In the case of (hierarchical) nested sampling, if the population shows spatial autocorrelation, then again there will be information redundancy because some sample sites are close together. This effect may be somewhat less of a problem in the case of fixed-interval point sampling as a result of keeping sample sites apart.

In the model-based approach, if the purpose is to estimate the mean, μ, of a spatially autocorrelated model using the sample mean, then this estimator's sampling variance is inflated relative to its sampling variance under the assumption of a population that is independent. If eqn [2] is used as the estimator of the sampling variance, it can severely underestimate the true sampling variance of the sample mean. Although the sample mean is an unbiased estimator of μ, the best linear unbiased estimator (BLUE) is usually preferred, although it requires the correct specification of the underlying spatial model (see eqn [1]). As noted, the choice of sampling plan tends to be of secondary interest, but careful specification of the required sample size is needed if parameters are to be estimated to the desired level of precision (Haining, 1988).

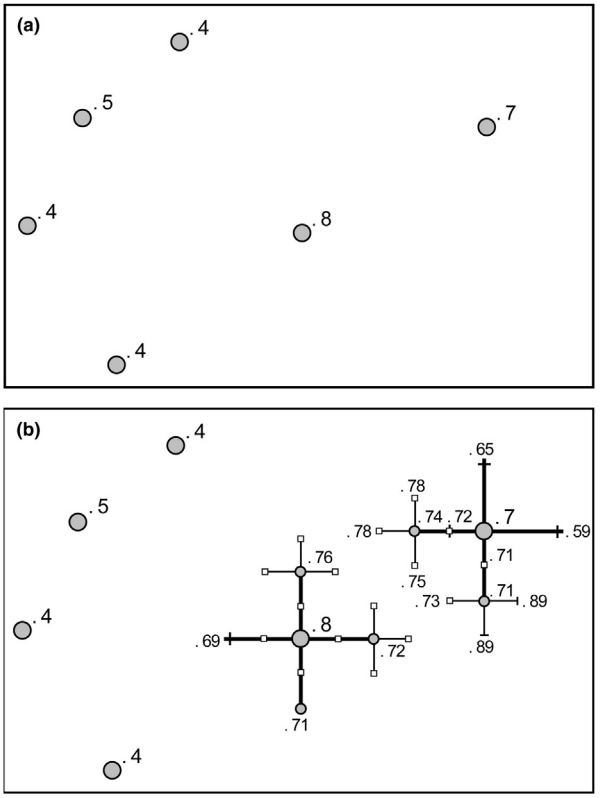

Figure 4 Cluster adaptive sampling. A site is fathomed (+) when its toxicity rate does not exceed the contamination value. Reproduced with permission from Delmelle, E., 2009. Spatial sampling. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 183–206. (Figure 10.6.)

In addition to undertaking sampling to estimate nonspatial properties of an attribute, sampling may be needed in order to estimate spatial properties (the correlogram or semi-variogram) or to construct maps based on optimal interpolation. Evidence suggests that particularly for optimal interpolation and for constructing contour maps for a variable, centric systematic sampling grids are best with a density chosen to achieve required levels of precision across the map and with a rectangular grid selected for practical reasons (Webster and Burgess, 1981). Interpolation error will vary across the map as a function of the density of sample points and the effect of the survey boundary. If anisotropy is present (directional differences in the autocorrelation structure) and/or heterogeneity, then both alignment and grid specification for the systematic sample will need to reflect these characteristics (McBratney et al., 1981).

See also: Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Sample Surveys: Model-Based Approaches; Spatial Analysis in Geography; Spatial Association, Measures of; Spatial Autocorrelation; Spatial Data Infrastructures; Spatial Data; Spatial Pattern, Analysis of; Spatial–Temporal Modeling.

Bibliography

Arbia, G., Lafratta, G., 1997. Evaluating and updating the sample design in repeated environmental surveys: monitoring air quality in Padua. Journal of Agricultural, Biological and Environmental Statistics 2, 451–466.

Bellhouse, D.R., 1977. Some optimal designs for sampling in two dimensions. Biometrika 64, 605–611.

Brus, D., deGruijter, J., 1997. Random sampling or geostatistical modeling?: choosing between design based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Cox, L., 1999. Adaptive spatial sampling of contaminated soil. Risk Analysis 19, 1059–1069.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Delmelle, E., 2009. Spatial sampling. In: Fotheringham, A., Rogerson, P. (Eds.), The SAGE Handbook of Spatial Analysis. SAGE, Los Angeles, pp. 183–206.

Dunn, R., Harrison, A.R., 1993. Two-dimensional systematic sampling of land use. Applied Statistics 42, 585–601.

Elliott, P., Cuzick, J., English, D., Stern, R., 1992. Geographical and Environmental Epidemiology: Methods for Small Area Studies. Oxford University Press, Oxford, UK.

Ghosh, M., Rao, J.N.K., 1994. Small area estimation: an appraisal. Statistical Science 9, 55–93.

Griffith, D.A., 1989. Distance calculations and errors in geographic databases. In: Goodchild, M., Gopal, S. (Eds.), Accuracy of Spatial Databases. Taylor and Francis, New York, pp. 81–90.

Haining, R.P., 1988. Estimating spatial means with an application to remotely sensed data. Communications in Statistics: Theory and Methods 17, 573–597.

Haining, R.P., 2003. Spatial Data Analysis: Theory and Practice. Cambridge University Press, Cambridge, UK.

Kahn, H.A., Sempos, C.T., 1989. Statistical Methods in Epidemiology. Oxford University Press, Oxford, UK.

Koop, J.C., 1990. Systematic sampling of two-dimensional surfaces and related problems. Communications in Statistics: Theory and Methods 19, 1701–1750.

Longford, N.T., 1999. Multivariate shrinkage estimation of small area means and proportions. Journal of the Royal Statistical Society, Series A 162, 227–245.

Matern, B., 1986. Spatial Variation, second ed. Springer Verlag, Berlin, Germany.

McBratney, A.B., Webster, R., Burgess, T.M., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – I theory and method. Computers and Geosciences 7, 331–334.

Oliver, M.A., Webster, R., 1986. Semi-variograms for modeling the spatial pattern of landform and soil properties. Earth Surface Processes and Landforms 11, 491–504.

Rao, J., 2005. Small Area Estimation. Wiley, New York.

Ripley, B., 1981. Spatial Statistics. Wiley, New York.

Wang, J., Haining, R., Cao, Z., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. International Journal of Geographical Information Science 24 (4), 523–543.

Webster, R., Burgess, T.M., 1981. Optimal interpolation and isarithmic mapping of soil properties: III changing drift and universal kriging. Journal of Soil Sciences 32, 505–524.

Spatial Statistical Methods

L Mark Berliner, The Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While adjustment for spatial dependence has long been mandated in different fields, spatial statistical models refer to modeling and data analysis of processes displaying spatial dependence. This article reviews the basics of spatial random processes, compares different aspects of spatial statistical models, and investigates several topics of spatial prediction (kriging, trend models, and multivariate methods). Bayesian spatial statistics and spatial point processes are also discussed. This article concludes with recent advances in spatial statistics, including spatial–temporal statistics, dimension reduction, and nonparametric methods.

Spatial statistics is characterized by modeling and data analysis for processes displaying spatially indexed dependence structures. Adjustment for spatial dependence has long been mandated in many disciplines. The very nature of *geography* is spatial. F. Stephen (1934) noted "… Data of geographic units are tied together… in dealing with social data, we know that by virtue of their very *social* character, persons, groups, and their characteristics are interrelated and not independent." In the context of agricultural field experiments, R.A. Fisher (1935) described the basic intuition underlying much of spatial statistics: "After choosing the area we usually have no guidance beyond the widely verifiable fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are far apart." Though Fisher developed randomized experimental designs to avoid explicitly modeling spatial dependence, agricultural applications are an important motivation for spatial statistics (Besag, 1991).

In the early 1950s the South African mining engineer D.G. Krige pursued methods for the spatial interpolation of data with the intent of predicting the distribution of ore (e.g., gold) in space. G. Matheron (France) extended that work, formalizing mining *geostatistics*, and coined the term 'kriging' to refer to a class of spatial prediction methods. In parallel, L.S. Gandin (Soviet Union) developed essentially the same procedures for applications in meteorology where they are known as *optimal interpolation* and *objective analysis*.

While basic geostatistical methods remain very useful, the methodology was limited in terms of (1) applicability, since the analyses were historically confined to *linear models* and (2) volume of data analyzed. Technological advances regarding both data collection methods and computation have led to expansion of applications and methods of spatial statistics. For example, *image analysis* is and will continue to be a critical research area. Problems include satellite *remote sensing* of the Earth's environment and ecology, including humans and their behavior and impacts; analysis of astronomical images from optical, radio wave, infrared, and related instruments; magnetic resonance imagery; and computer vision. Methods for *image restoration* (filtering out noise), *classification*, and *object identification* are sought. Similarly, the development of *geographic information systems* is a watershed of geographically referenced data, though the efficient extraction of information from such massive datasets offers new statistical challenges.

Spatial statistics also includes modeling and analysis of mapped *point processes*, in which the locations of observations are themselves random. For example, observations may include the locations and characteristics of trees in a forest (Stoyan and Penttinen, 2000) or positions and types of galaxies in space (Feigelson and Babu, 1992). For applications in archaeology, see Hodder and Orton (1976). Searching for patterns and *clusters* of data are key challenges. An introduction to spatial modeling of *random objects* can be found in Cressie (1993).

Perspectives and Comparisons

Let D denote a collection of *sites* or *locations*, denoted by s, distributed in space. (Formally, D is taken to be a subset of d-dimensional Euclidean space. This permits mathematical consideration of the differences and distances between sites.) In geostatistics, D is often a collection of sites on the surface of the Earth, indexed in two dimensions either by s = (latitude, longitude) or by Cartesian coordinates $s = (x, y)$, or in three dimensions with the addition of altitude or depth. In image analysis, s typically represents a *pixel* located on a discrete *lattice*. For each s in D, there exists some random quantity, $Z(s)$, of interest. The stochastic process $\{Z(s) : s \in D\}$ is a *random field* or spatial random process. Z could be a continuous variable (e.g., temperature); a discrete variable (e.g., a binary variate indicating the presence or absence of some characteristic, or a multivalued discrete variable of *gray levels*, especially in image analysis); a mixed variable (e.g., rainfall may be zero with positive probability at a location, yet continuous, if some rain occurs). Z may also be a multivariate variable.

Goals in spatial statistical analyses include distributional modeling of spatial random processes; estimation and inference for aspects of the spatial model based on observational data; prediction at sites where observations are not available; and filtering or noise reduction. Two approaches to spatial process modeling dominate practice. In the first, *jointly* or *simultaneously* specified models for the process $\{Z(s) : s \in D\}$ are developed. For example, in much of geostatistics, the variable $Z(s)$ is decomposed into 'mean' plus 'error' components. The

modeling challenge is the development of joint distributional information for the errors across space. A second strategy is to build *conditionally* specified models by modeling *local* dependence structures. This track directly builds on the intuition that variables in close spatial proximity tend to be related. This is quantified by formulating distributional features of $Z(\mathrm{s})$, *conditional* on the rest of the Z process but actually depending only on the values of Z at locations in a *neighborhood* of s. Practical modeling often focuses on specification of conditional means and variances. At least two important issues arise. First, to enable spatial prediction accompanied by meaningful measures of uncertainty, a conditionally specified model should correspond to a proper joint distribution. To that end conditionally specified *Markov random fields* are often used since there is a well-developed theory regarding the existence of implied joint models (e.g., Cressie, 1993: Chapter 6). Such models are readily analyzed by Markov chain Monte Carlo methods. The second issue involves dealing with variables located near the *boundary* of D, since this region typically dictates local dependence structures that are by nature different from those in the interior of D.

The primary goal in typical problems (e.g., geostatistics) is spatial prediction. Based on a vector of n observations, $\mathcal{Z} \equiv (Z(\mathrm{s}_1), \ldots, Z(\mathrm{s}_n))'$, we seek predictions of $Z(\mathrm{s}_0)$ where s_0 is an arbitrary site in D. Depending upon the level of assumptions made, two essential principles guide the predictive analysis. First, in analyses based on assumptions asserting only the trend and covariance structures, most predictions are derived as *best linear unbiased predictions* (BLUP). 'Best' refers to the property of having smallest prediction variance among all linear, unbiased predictors. In analyses which also include distributional assumptions, predictive analyses are based on the implied conditional distribution of $Z(\mathrm{s}_0)$ given the observed data $\mathcal{Z}$. For example, assuming mean squared prediction error as the criterion for assessing the quality of prediction, the optimal predictor is $\widehat{Z}(\mathrm{s}_0) = E(Z(\mathrm{s}_0)|\mathcal{Z})$; that is, the conditional mean of $Z(\mathrm{s}_0)$ given the data.

Both Bayesian and non-Bayesian viewpoints are applied in spatial statistics. Since most strategies view the primary variable of interest Z as *random*, the principle that spatial prediction is based on the conditional distribution, or at least its moments, of the unobserved variables given the data is clearly tied to the Bayesian view in both interpretation and computational techniques. This is particularly evident in image analysis, where a potentially nonrandom, 'true image' is modeled as a realization of a stochastic spatial process (e.g., Geman and Geman, 1984; Besag, 1989). On the other hand, despite similar stochastic modeling of the primary variables of interest, the origins of geostatistics and spatial point process analysis appear to be primarily non-Bayesian in language and spirit. In particular, early approaches to model fitting and parameter estimation were primarily non-Bayesian, though their interpretation as *empirical Bayesian* is quite natural. With the advent of powerful Bayesian computational strategies (i.e., Markov chain Monte Carlo), the Bayesian approach has emerged as a critical tool in applications of all areas of spatial statistics. It offers opportunities for effective analyses in complex spatial problems, including multivariate and inhomogeneous datasets and data types, as well as nonlinear and non-Gaussian datasets.

Spatial Random Fields

Assuming all quantities exist, let $\mu(\mathrm{s})$, $\mathrm{s} \in D$ be the mean (expected value) $\mu(\mathrm{s}) = E(Z(\mathrm{s}))$, and let $\mathrm{cov}(Z(\mathrm{s}), Z(\mathrm{s}'))$, defined for all pairs of sites s and s′ in D, be the covariance function of the process.

The process is (second-order or weakly) *stationary* if it has constant mean $\mu(\mathrm{s}) \equiv \mu$ and

$$\mathrm{cov}(Z(\mathrm{s}), Z(\mathrm{s}')) = C(\mathbf{h}), \quad \text{for all} \quad \mathrm{s} \in D, \mathrm{s}' \in D \qquad [1]$$

where $\mathbf{h} = \mathrm{s} - \mathrm{s}'$. That is, the covariance between $Z(\mathrm{s})$ and $Z(\mathrm{s}')$ depends on the locations only through their difference. Note that this implies that the variance of $Z(\mathrm{s})$, $\sigma^2(\mathrm{s}) \equiv \sigma^2$, is constant. The function $C(\mathbf{h})$ is known as the *covariogram*. *Strict* or *strong stationarity* involves similar requirements applied to the joint distributions of $\{Z(\mathrm{s}) : \mathrm{s} \in \mathcal{D}\}$ as opposed merely to first and second moments (Cressie, 1993: Section 2.3).

Alternatively, again assuming that $\mu(\mathrm{s}) \equiv \mu$, the process is *intrinsically stationary* if

$$\mathrm{var}(Z(\mathrm{s}) - Z(\mathrm{s}')) = 2\gamma(\mathbf{h}), \quad \text{for all} \quad \mathrm{s} \in D, \mathrm{s}' \in D \qquad [2]$$

The function $2\gamma(\mathbf{h})$ is known as the *variogram*; $\gamma(\mathbf{h})$ is the *semivariogram*.

If the process is stationary, then calculation implies

$$.5\ \mathrm{var}(Z(\mathrm{s}) - Z(\mathrm{s}')) = C(0) - C(\mathbf{h}) = \gamma(\mathbf{h}) \qquad [3]$$

Hence, stationarity implies intrinsic stationarity. However, variograms satisfying [2] may exist in some cases for which covariograms do not exist. For stationary processes, [3] allows one to view modeling covariograms and variograms as being equivalent, though some have argued that estimation based on data is enhanced for the variogram.

An intrinsically stationary process is *isotropic* if $\gamma(\mathrm{s} - \mathrm{s}') = \gamma(h)$ for all pairs of sites, where h is the *distance* between them. That is, the variogram depends only on the distance between sites (i.e., the length of the vector $\mathbf{h} = \mathrm{s} - \mathrm{s}'$), but not their relative directions. *Anisotropy* (the violation of isotropy) arises for processes that develop differentially in different directions of space (Cressie, 1993: Section 2.3). For example, temperature values, or variables depending upon temperature (e.g., type of agriculture, ecology, etc.), are roughly moderated by latitude. Hence, north–south covariance structures might well be different from their east–west counterparts. An example of an isotropic semivariogram is the *exponential model*:

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ \gamma_0 + \gamma_1\left(1 - \mathrm{e}^{-h/a}\right) & \text{if } h > 0 \end{cases} \qquad [4]$$

where $a > 0$. Note that as $h \to 0$, $\gamma(h) \to \gamma_0$, rather than 0. This discontinuous behavior, known as the *nugget effect*, occurs in a variety of commonly used semivariograms. Physically, γ_0 has been interpreted as arising from (1) very small-scale fluctuations in the Z process and/or (2) measurement error. However, if the process modeled is a sufficiently smooth function of space, the only mathematically valid explanation for the nugget effect is measurement error.

The exponential covariogram corresponding to [4] is

$$C(h) = \begin{cases} \gamma_0 + \gamma_1 & \text{if } h = 0 \\ \gamma_1 \mathrm{e}^{-h/a} & \text{if } h > 0 \end{cases} \qquad [5]$$

Another covariogram sometimes used is the *Gaussian model*:

$$C(h) = \begin{cases} \gamma_0 + \gamma_1 & \text{if } h = 0 \\ \gamma_1 e^{-h^2/b} & \text{if } h > 0 \end{cases} \quad [6]$$

where $b > 0$. (Use of the modifiers 'exponential' and 'Gaussian' accrues from similarities to the exponential and Gaussian probability density functions, rather than distributional assumptions for Z.) A very rich class of isotropic covariograms is the *Matérn class*, written here without a nugget for comparison:

$$C(h) = \frac{2\sigma^2}{\Gamma(\nu)}\left(h\sqrt{\nu}/a\right)^{\nu} K_{\nu}\left(2h\sqrt{\nu}/a\right) \quad [7]$$

where $\nu > 0$, σ^2 is the variance of $Z(\mathrm{s})$, $\Gamma(\cdot)$ is the usual gamma function, and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind.

Spatial Prediction

A standard spatial model takes the form

$$Z(\mathrm{s}) = \mu(\mathrm{s}) + \varepsilon(\mathrm{s}) \quad [8]$$

where $\mu(\mathrm{s})$ is a spatial mean model and $\varepsilon(\mathrm{s})$ is a zero-mean spatial error process. Critical issues involve the choices of models regarding both the mean or *trend* (also known as *drift*) $\mu(\mathrm{s})$ and the error process $\{\varepsilon(\mathrm{s}) : \mathrm{s} \in \mathcal{D}\}$. Such modeling is challenging in that both components explain aspects of the spatial structure present in Z. The conventional guideline is that the trend reflects large-scale spatial variation while the error process explains residual small-scale spatial dependence. Implementation of this guideline and interpretation of results vary with the context as well as the goals and knowledge of the analyst. As noted by Cressie (1993, p. 25), "What is one person's (spatial) covariance structure may be another person's mean structure." Also, note that in modern statistical approaches, trend models based on both spatial location and covariates are recommended. Trend models can also be incorporated in conditionally specified models, either via conditionally specified mean models, or with globally specified mean models and conditionally specified error models.

The Basics of Kriging

Ordinary kriging asserts a constant trend model. Linear predictors take the form

$$\widehat{Z}(\mathrm{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathrm{s}_i) \quad [9]$$

The unbiasedness condition, $E(\widehat{Z}(\mathrm{s}_0)) = \mu$, adds the side condition that $\sum_{i=1}^{n} \lambda_i = 1$. Minimization of the predictive variance (or mean squared prediction error)

$$E\left(Z(\mathrm{s}_0) - \widehat{Z}(\mathrm{s}_0)\right)^2$$

subject to the side condition yields simple formulas for the optimal values of the λ_i as appropriate functions of the covariance function. Alternatively, the formula can be written in terms of the variogram, enabling spatial prediction for processes that are intrinsically stationary (Cressie, 1993, p. 121–123). The achieved optimal value of the predictive variance is then used to measure the precision of the prediction.

The kriging formulas are derived under the assumption that the variogram or covariance function is known. Typically, data-based estimates of these functions are simply plugged into the kriging formulas. A variety of methods for estimation of the variogram (or covariogram) have been suggested. A simple procedure is to apply a method-of-moments notion: an empirical version of the semivariogram is

$$\widehat{\gamma}(\mathbf{h}) = \frac{\sum_{n(\mathbf{h})}\left(Z(\mathrm{s}_i) - Z(\mathrm{s}_{i'})\right)^2}{\#n(\mathbf{h})} \quad [10]$$

where the sum is taken over those data points satisfying $\mathrm{s}_i - \mathrm{s}_{i'} = \mathbf{h}$ and $\#n(\mathbf{h})$ is the number of such pairs of data points. A simple estimate of the covariogram is

$$\widehat{C}(\mathbf{h}) = \frac{\sum_{n(\mathbf{h})}\left(Z(\mathrm{s}_i) - \overline{Z}\right)\left(Z(\mathrm{s}_{i'}) - \overline{Z}\right)}{\#n(\mathbf{h})} \quad [11]$$

where $\overline{Z}$ is the simple mean of the observed values. However, since there are only n observations, these estimates cannot provide estimates for all $\mathbf{h}$. Practical estimates are typically constructed based on *binning* procedures. For example, under the additional assumption of isotropy, consider a collection of M nonoverlapping intervals

$$I_m = [h_m - \delta_m, h_m + \delta_m]; m = 1, \ldots, M$$

For each interval the above empirical sums and counts of pairs are taken over all pairs separated by distances in I_m. The resulting M values can then be smoothed to produce a function estimate.

The above estimates suffer from a variety of deficiencies, especially if the sample size n is small. A variety of alternatives have been suggested. One general approach is to *parameterize* the covariogram (or variogram), specifying it to be a member of some class (the Matérn class [7] is a popular choice). Methods for parameter estimation include least squares, weighted least squares, and maximum likelihood. A second avenue for estimation involves the use of nonparametric function estimation, especially in the context of estimating the spectral representation of the covariogram. Also, robust estimates with respect to outlying observations have been developed. See Cressie (1993) and Stein (1999) for further discussion of classes of variograms and estimation procedures. Sampson and Guttorp (1992) consider estimation in the nonstationary case.

Trend Models and Their Estimation

Universal kriging refers to spatial prediction based on trend models of the form

$$\mu(\mathrm{s}) = \sum_{j=0}^{p} f_j(\mathrm{s})\beta_j \quad [12]$$

where the $f_j(\cdot)$ are $p+1$ specified functions of space and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is a vector of $p+1$ unknown parameters. In vector notation, let $\mathbf{f}(\mathrm{s}) = (f_0(\mathrm{s}), \ldots, f_p(\mathrm{s}))'$, so the model is written compactly as

$$\mu(\mathrm{s}) = \mathbf{f}'(\mathrm{s})\boldsymbol{\beta}$$

Ordinary kriging is included by assuming $f_0(\mathrm{s}) = 1$ and $p = 0$.

A common example is a quadratic trend surface model: writing locations in Cartesian coordinates $\mathrm{s} = (x, y)$, assume that

$$\mu(\mathrm{s}) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$$

In addition to higher order polynomial trends, nonpolynomial models, such as *local polynomials* have been considered. The use of *splines* is highly related; see Wahba (1990) and Cressie (1993, p. 180–183).

Representation of the kriging model in matrix form provides a better understanding and enables parallels to familiar statistical linear regression modeling. The notation $\mathbf{X} \sim (\mathbf{m}, \mathbf{C})$ is read 'the random vector $\mathbf{X}$ has a multivariate distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.' The notation $\mathbf{X} \sim N(\mathbf{m}, \mathbf{C})$ adds the assumption that the distribution is multivariate normal. The standard universal kriging model is then

$$\begin{pmatrix} Z(\mathrm{s}_0) \\ \mathcal{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_0'\boldsymbol{\beta} \\ \mathbf{F}\boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \varepsilon(\mathrm{s}_0) \\ \boldsymbol{\varepsilon} \end{pmatrix} \qquad [13]$$

where

$$\begin{pmatrix} \varepsilon(\mathrm{s}_0) \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2(\mathrm{s}_0) & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix} \right) \qquad [14]$$

where $\boldsymbol{\sigma}$ is the $n \times 1$ vector of covariances between errors at site s_0 and the observation sites; $\mathbf{F}$ is the $n \times (p+1)$ matrix dictated by the functions $f_j(\cdot)$ evaluated at the observation locations; and $\boldsymbol{\varepsilon}$ is the vector of errors at observation sites. These errors are assumed to have zero means and covariance matrix $\boldsymbol{\Sigma}$ whose entries are the covariances $\mathrm{cov}(\varepsilon(\mathrm{s}_i), \varepsilon(\mathrm{s}_{i'}))$ for pairs of data points.

Formulas for the optimal linear predictor or BLUP are readily derived (Cressie, 1993, p. 172–180):

$$\widehat{Z}(\mathrm{s}_0) = \mathbf{f}'(\mathrm{s}_0)\boldsymbol{\beta} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\mathcal{Z} - \mathbf{F}\boldsymbol{\beta}) \qquad [15]$$

These formulas depend on appropriate values of the covariance function and $\boldsymbol{\beta}$, requiring estimation of both quantities. Note that use of [15] does not require spatial stationarity of the error process, though it is often assumed in estimation.

Estimation of $\boldsymbol{\beta}$ is typically accomplished via *generalized least squares*. Since the observations follow the model

$$\mathcal{Z} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad [16]$$

the generalized least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\beta} = \left(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}\boldsymbol{\Sigma}^{-1}\mathcal{Z} \qquad [17]$$

Estimation of the covariogram of the $\boldsymbol{\varepsilon}$ process proceeds as discussed earlier, though now the data used are the residuals from the fitted trend model: $\mathcal{Z} - \mathbf{F}\widehat{\beta}$. However, there is a clear difficulty. The very residuals needed to estimate the covariogram depend on the covariogram since $\widehat{\beta}$ is a function of $\boldsymbol{\Sigma}$. A variety of approaches to this problem have been suggested. One is to use some preliminary guess at $\boldsymbol{\Sigma}$ in [17], but base prediction formulas on the estimated covariogram of the resulting residuals. With the increase of readily available computation, the more modern approach *iterates* between generalized least squares estimation of β and covariogram estimation until some stability is achieved. Note that the extra uncertainty accruing from the use of data-based estimates of β and the covariance function is seldom accounted for in classical kriging.

Treatment of measurement error can be included in the modeling. Rather than observing $\mathcal{Z}$ directly, the observed data values, denoted by Y, may be modeled as $Y(\mathrm{s}_i) = Z(\mathrm{s}_i) + e(\mathrm{s}_i)$; $i = 1, \ldots, n$. Let $\mathcal{Y}$ denote the $n \times 1$ vector of observations. In matrix form we write

$$\mathcal{Y} \sim (\mathcal{Z}, \mathbf{D}_\gamma) \qquad [18]$$

where $\mathbf{D}_\gamma$ is the covariance matrix of the measurement errors $\mathbf{e} = (e(\mathrm{s}_1), \ldots, e(\mathrm{s}_n))'$. A common, but not unequivocal, assumption is that $\mathbf{D}_\gamma = \sigma_\gamma^2\mathbf{I}$ (i.e., the measurement errors are mutually uncorrelated). Also, assume $\mathbf{e}$ is uncorrelated with $\mathcal{Z}$. If Z is a smooth function of space, σ_γ^2 is a quantification of the nugget effect. To complete the formalization, interpret [18] as a specification *conditional* on $\mathcal{Z}$. The model [16] for $\mathcal{Z}$ implies that the *unconditional* model for $\mathcal{Y}$ is

$$\mathcal{Y} \sim \left(\mathbf{F}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_\gamma^2\mathbf{I}\right) \qquad [19]$$

(Compare this covariance matrix to the piecewise definitions of variograms in [5] and [6].) Without replicate observations at some of the observation locations, σ_γ^2 is not directly estimable in classical kriging. Typically, estimates are obtained by extrapolation of covariogram estimates for small distances.

Note that the goals of the analysis now include not only prediction of $Z(\mathrm{s}_0)$, but also prediction of $\mathcal{Z}$. Predicting $\mathcal{Z}$ based on $\mathcal{Y}$ provides *spatial smoothing*, as opposed to *spatial interpolation* when σ_γ^2 is assumed to be zero.

Multivariate Methods

In many applications, several variables may be of interest, suggesting the consideration of a random vector, $\mathbf{Z}(\mathrm{s}) = (Z_1(\mathrm{s}), \ldots, Z_K(\mathrm{s}))'$. In mining, the data might consist of measurements of amounts of K minerals at each observed site. In meteorology, data on temperature, pressure, and precipitation are often collected. In satellite imaging, multichannel sensors collect intensities in various bandwidths.

The classical geostatistical treatment of multivariate spatial data is known as *cokriging*. The modeling strategy is a direct generalization of the univariate case. For each coordinate $Z_k(\mathrm{s})$ of $\mathbf{Z}(\mathrm{s})$, consider a mean or trend model (see [12]),

$$\mu_k(\mathrm{s}) = \mathbf{f}_k'(\mathrm{s})\boldsymbol{\beta}_k, k = 1, \ldots, K \qquad [20]$$

Note that the spatial regression functions comprising the various f_k need not be the same for all k; indeed, the dimensions of the β_k may vary with k. The key second-order parameters are the covariance matrices,

$$\mathrm{cov}(\mathbf{Z}(\mathrm{s}), \mathbf{Z}(\mathrm{s}')) = \mathbf{C}(\mathrm{s}, \mathrm{s}'), \mathrm{s} \in D, \mathrm{s}' \in D \qquad [21]$$

The principles and basic formulas for best linear unbiased prediction remain essentially the same as for the univariate case. Since the notation is cumbersome, the reader is referred to Wackernagel (1995) for details.

Practical implementation of cokriging can be difficult. First, specification and subsequent estimation of $\mathbf{C}(\mathrm{s}, \mathrm{s}')$ is problematic, in that it involves specification not merely of K spatial

covariance functions (one for each variable) but also, $.5K(K-1)$ *cross-covariance functions*, one for each pair Z_k, $Z_{k'}$, $k \neq k'$. The number of observations critically limits the ability to estimate the required covariance functions. Indeed, sample size typically mandates the need for extremely simple, often implausible, covariance models and/or a reduction in K. Further, the sampling pattern (e.g., observations of only some coordinates of $\mathbf{Z}$ at some sites) can moderate the ability to estimate parameters and predict.

An alternative modeling strategy is to build multivariate models *hierarchically*; namely as a sequence of conditional models. A precursor to this view from the geostatistics literature is *kriging with external drift*. For example, suppose that spatial prediction of near-surface ozone is of interest. It is well accepted that temperature is predictive for ozone. Hence, spatial prediction of ozone could be enhanced if appropriate temperature data were available and included in the model. An elementary approach is to view temperature as a (fixed) covariate and to include it in the formulation of a universal kriging model for ozone. Next, suppose temperature observations are unavailable at sites at which ozone predictions are sought. One could essentially form a kriging model for temperature, and rely on temperature predictions to produce ozone predictions. These notions can be made rigorous statistically. Specification of a complete spatial model for ozone conditional on temperature combined with a marginal spatial model for temperature does imply a joint model for both, as might be formed in a cokriging approach. However, the parameterizations of the models can be different. In particular, rather than direct specification and estimation of the cross-covariance function for ozone and temperature, one specifies a conditional mean relationship between the two. For further discussion and examples see Royle and Berliner (1999).

Bayesian Spatial Statistics

Bayesian Geostatistics

Models for the Z process, such as the basic model [13] along with a measurement error stage as in [18] can readily be posed in the Bayesian framework. Process models for Z are viewed as *priors*; these priors may be simultaneously or conditionally specified. Specification of prior distributions on model parameters, such as the trend model coefficients $\boldsymbol{\beta}$, measurement error variance, and parameters of the modeled covariance function are also formulated. Predictive inference is then based on the posterior distribution of unknown quantities given the data. (A general review of Bayesian analysis is given in Bernardo and Smith, 1994.)

Direct analogs to conventional kriging often rely on multivariate normal distributions in both [13] and [18]. This imposition of distributional assumptions has been criticized. However, neither standard kriging nor Bayesian analyses based on normal distribution assumptions are recommended for highly non-normal data and processes. The Bayesian approach readily offers valid measures of uncertainty (e.g., Le and Zidek, 1992; Handcock and Stein, 1993) unlike those arising from the 'plug-in' of estimates procedures dominating non-Bayesian kriging.

With the advent of Bayesian computing advances, non-normal models can be analyzed. For example, settings in which the Z variable is discrete, such as in some imaging problems (Geman and Geman, 1984) or is count data modeled as via Poisson or related distributions (Kaiser and Cressie, 1997; Wolpert and Ickstadt, 1998), have been considered. Also, see Diggle et al. (1998).

Hierarchical Models

Bayesian hierarchical models for collections of variables are combinations of component models or *stages*. Specifically, a joint distribution for all uncertainty quantities is constructed as a product of conditional distributions. Motivations for hierarchical modeling include:

- individual stages often are formulated more easily than is the full joint model directly;
- scientific understanding of the relationships among variables is often readily available for components of the model;
- combining datasets consisting of very different types of data is often enhanced.

The first stage of a Bayesian hierarchical model is usually a *data model*; i.e., a measurement error model for the observations conditional on processes and parameters of interest (see [18]). In higher level stages, *process model priors* are formed to model and interrelate spatial processes, including latent processes and spatial random effects. By separating models of the data and the processes of interest, we often can incorporate science ranging from qualitative understanding to the use of underlying differential equations in modeling those processes; for examples of each notion see Royle and Berliner (2005) and Wikle and Berliner (2005), respectively. Finally, a *parameter model* specifies a prior probability distribution for all model parameters. See Banerjee et al. (2014) and Berliner (2000) for review and references.

The simultaneous developments of Bayesian hierarchical models and computational strategies to implement them have revolutionized the quality and range of problems we address in spatial and spatial–temporal statistics.

Spatial Point Processes

A spatial point process arises when the locations of events occur randomly in D. Basic data take the form $(s_1, \ldots, s_N)'$, where the number of locations, N, is a random variable.

The probability theory for point processes typically formulates a model for random points in $\mathcal{R}^2$. The model then implies a distribution for N, the number of events in the study region D, a bounded subset of $\mathcal{R}^2$, as well as the joint distribution of their locations. Generally, the treatment of edge effects then becomes a challenge (Ripley, 1988: Chapter 2).

A standard point process model is the homogeneous Poisson spatial process with intensity λ: for any k and any selection of disjoint subsets $A_1, \ldots, A_k$ of $\mathcal{R}^2$, assume that the numbers of events, $N(A_i)$, $i = 1, \ldots, k$, are independent random variables, each with probability

$$P(N(A_i) = n) = \frac{(\lambda|A_i|)^n}{n!}\exp(-\lambda|A_i|), n = 0, 1, \ldots, \quad [22]$$

where $|A_i|$ is the area of A_i, $\int_{A_i} \mathrm{ds}$. Note that for any set A, the mean $E(N(A)) = \lambda|A|$. The Poisson process is viewed as *completely spatially random*, since counts in disjoint intervals are independent and the probabilities in [22] depend only on A_i through its area and not its shape or position. This model is often used as a *null hypothesis*. Rejection of the Poisson model suggests that the process has some spatially dependent structure, or dependence among separated regions. Numerous generalizations have been considered. The Poisson *cluster process* allows cluster locations to be distributed according to one Poisson spatial process while events in clusters are distributed according to another Poisson model. Extensions to alternative measures on sets beyond simple area, including measures which vary in space leading to *inhomogeneous processes* have been considered. A *Cox process* is generated as a Poisson process, but the measures associated with sets are themselves random.

As in geostatistics and related areas, selected moments of a spatial process play roles in both theory and data analysis. Define the *intensity function* of a point process as

$$\lambda(\mathrm{s}) = \lim_{|\mathrm{ds}|\to 0} \frac{P(N(\mathrm{ds}) > 0)}{|\mathrm{ds}|} \quad [23]$$

where ds is a small square centered at s. If only one event may occur at a location, note that

$$\lambda(\mathrm{s}) = \lim_{|\mathrm{ds}|\to 0} \frac{E(N(\mathrm{ds}))}{|\mathrm{ds}|}$$

The second-order product density of a point process is

$$\lambda_2(\mathrm{s}, \mathrm{s}') = \lim_{|\mathrm{ds}|\to 0, |\mathrm{ds}'|\to 0} \frac{E(N(\mathrm{ds})N(\mathrm{ds}'))}{|\mathrm{ds}|\ |\mathrm{ds}|} \quad [24]$$

If $\lambda_2(\mathrm{s}, \mathrm{s}')$ depends only on $\mathbf{h} = \mathrm{s} - \mathrm{s}'$, the process is stationary; if $\lambda_2(\mathrm{s}, \mathrm{s}')$ depends only on the distance h between the points, it is isotropic. In the isotropic case Ripley (1977) developed a theory for the *K function*,

$$K(h) = \frac{E(\text{number of other points within distance } h \text{ of an event})}{\lambda} \quad [25]$$

and its relationship to $\lambda_2(h)$. A Poisson process is isotropic with $K(h) = \pi h^2$.

The K function plays a critical role in theory, because it can be readily computed for a variety of important point processes, and in practice, because it can be readily estimated from data. Simple data-based estimates take the form

$$\widehat{K}(h) = \frac{\sum_{i=1}^{N} \text{number of other points within distance } h \text{ of } \mathbf{s}_i}{\lambda N}$$

(λ is often estimated as $N/|D|$.) Note that this estimate is negatively biased since events outside the data collection domain D are not included. The intuition behind the K function relates generally to statistical *distance methods*, and specifically, to *nearest-neighbor methods*. Distance methods involve statistics which are computed from the set of all distances between events, rather than the actual locations. Substantial theory regarding the distributions of distance-based statistics is available.

Extensions of point processes include *marked point processes* for which a (possibly multivariate) variable Z is associated with each event, leading to data of the form $\mathcal{Z} \equiv (Z(\mathrm{s}_1), \ldots, Z(\mathrm{s}_N))'$. A related multivariate class of problems involves the modeling and data analysis for multiple point processes.

Recent Advances

Spatial–Temporal Statistics

Complex model structures typically arise when both spatial and temporal dependences and their interactions are present. Early attempts in this area often viewed time as if it were another spatial dimension, leading to relatively simple treatments of spatial–temporal data via spatial methods. However, the special role of time (i.e., the direction of its flow) often renders such approaches ineffective. Alternatively, the *Kalman filter* adapted to space–time problems has had wide application in the sciences. The basic method is analogous to Bayesian kriging in temporal settings. Spatial–temporal statistics is at the forefront of current statistical research. See Cressie and Wikle (2011) for a review and references. Many of advances in spatial statistics were developed in spatial–temporal settings. Such examples and references having a strong spatial component are included here.

Dimension Reduction

As the sizes of datasets grow, many traditional spatial analyses become difficult to implement. For example, as we saw in the section 'Spatial Prediction,' most traditional methods involve the modeling, estimation, and manipulation (e.g., inversion) of covariance matrices. In very large problems all three of these facets may become very difficult. Hence, development of techniques involving dimension reduction is a primary focus of current research.

Useful classes of dimension reductions involve modeling the mean and/or covariance functions as combinations of selected basis functions. For a vector of length n, consider [8] of the form

$$\mathcal{Z} = \sum_{j=1}^{p} \phi_j a_j + \varepsilon \quad [26]$$

where the ϕ_j are $p+1$ specified basis functions, the β_j are unknown parameters, and n is extremely large compared to p. For example, the ϕ_j may be the leading $p+1$ principal components (PCs) or empirical orthogonal functions, as they are known in the geosciences, of some selection of the covariance matrix of ε. When replicates of $\mathcal{Z}$ are available, their empirical covariance matrix is usually used. The singular value decomposition can be used to define the ϕ_j even if the number of replicates is less than n, though larger than $p+1$. Berliner (2000) provide an example where $n > 2000$ and replicates were defined across time. They used a small number of the leading PCs in the mean model and the next several PCs in defining the covariance matrix of ε.

Various approaches to the selection of basis vectors have been pursued. Methods such as those of the previous

paragraph fall in the general framework of Karhunen–Loeve expansions. Other examples include wavelets and neural nets. Wikle et al. (2001) use the normal modes of a governing partial differential equation as basis functions. Further, we need not require that the basis functions are orthogonal. This opens a wide range of *kernel* basis functions. The key idea is to model $Z(s)$ as the convolution of a weighting function and a stationary, typically Gaussian process. See Higdon (1998) and Wikle and Cressie (1999).

The above discussion focused on modeling the mean. Dimension reduction is also critical in managing high-dimensional covariances. One class of approaches is *sparse models*. Specifically, a large covariance matrix is modeled as the sum of an easily inverted covariance matrix (e.g., a scalar times the identity matrix) and a sparse matrix, i.e., most of its elements are equal to zero. The sparse matrix is used to capture local-in-space covariances.

Nonparametric Bayesian Methods

Nonparametric Bayesian methods involve modeling random quantities as being generated from some probability model which in turn is assumed to be generated from a probability model. Infinite random mixtures of distributions and Dirichlet processes are common models for generating a random probability distribution. In spatial statistics, such models have proven to be very useful, especially in settings in which clustering is present; see Gelfand et al. (2005).

Extensions

The techniques for modeling means and covariances associated with spatial statistics can be applied in settings. For example, building models for processes in space by some other variable, depending on the application, is the cornerstone many approached to the design and analysis of computer experiments. See Santner et al. (2003). Further, design of data collection in spatial settings is a critical and growing area in research. Finally, challenging areas are attracting numerous research efforts. For example, the study of spatially distributed networks is particularly interesting; see Barthelemy (2010).

See also: Bayesian Statistics; Hierarchical Models: Random and Fixed Effects.

Bibliography

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical Modeling and Analysis for Spatial Data 2/e. Chapman and Hall/CRC, Boca Raton, Florida.

Barthelemy, M., 2010. Spatial Networks arXiv:1010.0302v2 [cond-mat.stat-mech].

Besag, J.E., 1989. Towards Bayesian image analysis. Journal of Applied Statistics 16, 395–407.

Besag, J.E., 1991. Spatial statistics in the analysis of agricultural field experiments. In: Spatial Statistics and Digital Image Analysis. National Academy Press, Washington, DC.

Berliner, L.M., 2000. Hierarchical Bayesian modeling in the environmental sciences. Allgemeines Statistisches Archiv, Journal of the German Statistical Society 84, 141–153.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. John Wiley & Sons, Inc., New York.

Cliff, A., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London.

Cressie, N., 1993. Statistics for Spatial Data, revised ed. John Wiley & Sons, Inc., New York.

Cressie, N., 1995. Bayesian smoothing of rates in small geographic areas. Journal of Regional Science 35, 659–673.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. Wiley, Hoboken, New Jersey.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model based geostatistics (with discussion). Applied Statistics 47, 299–350.

Feigelson, E.D., Babu, G.J. (Eds.), 1992. Statistical Challenges in Modern Astronomy. Springer-Verlag, New York.

Gelfand, A., Kottas, A., MacEachern, S., 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association 100 (471), 1021–1035.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 6 721–741.

Handcock, M.S., Stein, M.L., 1993. A Bayesian analysis of kriging. Technometrics 35, 403–410.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environmental and Ecological Statistics 5, 173–190.

Hodder, I., Orton, C., 1976. Spatial Analysis in Archaeology. Cambridge University Press, London.

Kaiser, M.S., Cressie, N., 1997. Modeling Poisson variables with positive spatial dependence. Statistics and Probability Letters 35, 423–432.

Kyriakidis, P.C., Journel, A.G., 1999. Geostatistical space-time models: a review. Mathematical Geology 31, 651–684.

Le, N.D., Zidek, J.V., 1992. Interpolation with uncertain spatial covariance: a Bayesian alternative to kriging. Journal of Multivariate Analysis 43, 351–374.

Ripley, B.D., 1988. Statistical Inference for Spatial Processes. Cambridge University Press, Cambridge.

Royle, J.A., Berliner, L.M., 1999. A hierarchical approach to multivariate spatial modeling and prediction. Journal of Agricultural, Biological, and Environmental Statistics 4, 1–28.

Royle, J.A., Berliner, L.M., Wikle, C.K., Milliff, R., 1998. A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea. In: Gatsonis, C., et al. (Eds.), Case Studies in Bayesian Statistics IV. Springer-Verlag, pp. 367–382.

Sampson, P.D., Guttorp, P., 1992. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 87, 108–119.

Santner, T.J., Williams, B.J., Notz, W.I., 2003. The Design and Analysis of Computer Experiments. Springer, New York.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Stephan, F., 1934. Sampling errors and interpretations of social data ordered in time and space. In: Ross, F.A. (Ed.), Proceedings of the American Statistical Journal, New Series No. 185A. Journal of the American Statistical Association 29 (Suppl.), 165–166.

Stoyan, D., Penttinen, A., 2000. Recent applications of point process methods in forestry statistics. Statistical Science 15, 61–78.

Wackernagel, H., 1995. Multivariate Geostatistics. Springer-Verlag, New York.

Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.

Wikle, C.K., 2000. Hierarchical space-time dynamic models. In: Berliner, L.M., et al. (Eds.), Studies in the Atmospheric Sciences, Lecture Notes in Statistics, vol. 144. Springer-Verlag, New York, pp. 45–64.

Wikle, C.K., Berliner, L.M., Cressie, N., 1998. Hierarchical Bayesian space-time models. Journal of Environmental and Ecological Statistics 5, 117–154.

Wikle, C.K., Cressie, N., 1999. A dimension-reduced approach to space-time Kalman filtering. Biometrika 86 (4), 815–829.

Wikle, C.K., Milliff, R.F., Nychka, D., Berliner, L.M., 2001. Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds. Journal of the American Statistical Association 96, 382–397.

Wolpert, R.L., Ickstadt, K., 1998. Poisson/gamma random fields for spatial statistics. Biometrika 85, 251–267.

Spatial Thinking in the Social Sciences

Paul Claval, University of Paris-Sorbonne, Paris, France

Published by Elsevier Ltd.

Abstract

The interest of social scientists changed with time. Analyzing spatial distributions was a way to discover the relativity of human realities and their social nature. In the eighteenth century, the interest shifted toward the influence of environment on human behavior. Later, it focused on the role of distance on the way societies function. Since the 1970s, a spatial turn occurred: societies are always spatial realities; they build their environments, but depend on its natural dimension, which has to be preserved.

Space played a fundamental role in the birth of the social sciences in the eighteenth and early nineteenth centuries, since it allowed for comparative studies that were conducive to a search for explanation. Later, systematic social sciences neglected spatial problems, except for some functional approaches, which explored the role of space as environment in the evolutionist perspective of human geography or as distance as a key factor in the location of economic activity. Since the 1970s, the growing awareness of the fundamental diversity of societies, economies and polities was conducive to a rediscovery of the role of environment and distance, i.e., space, in social processes. A new interest in the sense of place – space as experience – developed also.

The objects upon which social sciences work, either people or artifacts, take up some room and are located in specific places. Cognition implies the building and mobilization of spatial skills, like orientation. Social scientists are generally more interested in particular spatial categories like direction, distance, place, area, landscape, village, town, and city than in space: the distribution of things and people, the behaviors and practices of everyday life, or the plans developed for the future deal with concrete geographical realities rather than the abstract idea of space. Even geography was traditionally more oriented toward places, regions, or zones than toward space: the idea of geography as a spatial discipline did not develop before the 1950s, and was soon criticized. Among the other social sciences, economics was the only one with a specialty branch in the field: spatial economics, which began to flourish in the 1820s. Among the cognitive sciences, the psychology of orientation developed before Jean Piaget described the stages of space construction by children. As a result, spatial thinking in the social sciences often appears split into different fields, *urban* studies, *rural* studies, *landscape* studies, the significance of the *regional approach* and of *place*, the idea of *environment*, the problem of *orientation*, etc.

Spatial thinking refers to five types of intellectual operations: (1) using geographical observations as a starting point for social analysis; (2) explaining, generally within a functional perspective, the role of different aspects of space in the life of social groups; (3) understanding the way people invest symbolic meanings in their environment; (4) wondering about the significance given to space in the definition of concepts like society, economics, culture, language, polity, etc.; or (5) deciphering the ways in which the human mind deals with space and builds and uses spatial categories. The first type was present from the earliest developments of the social sciences. The second type flourished mainly in the late nineteenth and most of the twentieth centuries. The third type is more characteristic of the rise of phenomenological approaches, since the 1930s or 1940s. The fourth type is mainly linked to the development of critical perspectives on the social sciences since the 1970s. The development of an interest for space by the cognitive sciences remained until recently independent from the other approaches.

Until the 1960s or 1970s, social sciences displayed more interest in time than in space: philosophies of history were central to Western thought. With the growing criticism of Western modernity, the situation changes. A new concern with space developed from the 1980s: a 'spatial turn.'

Spatial Distributions as a Point of Departure for Social Thinking

We are living in social systems and so deeply embedded into them that generally we think they are an expression of the natural order. In order to build a science of social realities, the social scientist has to introduce a distance between him and the object he studies. Travel narratives and geographical descriptions played a central role in this operation. Montaigne started to ponder about the relativity of social institutions when reading about the American cannibals. From Jean Bodin to Montesquieu, the diversity of political and social institutions nurtured the reflection on social and political life and was interpreted through a simple model, the influence of climate.

The first steps of economics were linked with geographic evidence: the role of transport in the distribution of economic activities; the complementarities between cities and rural areas; the economic circuit served as bases for the development of a more conceptual vision, which started with the analysis of market mechanisms by Turgot and Adam Smith's reflection on the division of labor.

With the development of travels and the improvement of their accounts, geographical comparisons became more sophisticated: they allowed for the development of new disciplines, like ethnography. The change of scenery could be provided by travel through time as well as space. Immanuel Kant emphasized the similarities between history and geography: spatial and temporal distributions were the first features to confront social scientists in their study of the world. He opposed spatiotemporal descriptions to systematic analyses.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72092-7

Partly under the influence of Kant, the idea that geography was essentially a description of spatial distributions developed and remained alive well into the twentieth century.

It was during the first decades of the nineteenth century that spatial comparisons played their most important role in the development of science. It was true for the natural sciences – Charles Darwin developed his evolutionary theory when traveling on the *Beagle*. In the social sciences, Alexis de Tocqueville built his analysis of democracy by comparing the United States, where it permeated all forms of social life, with Old Europe, where aristocrats were still influential. Karl Marx explored the opposition between medieval cities and rural areas, the Asiatic social and economic organization, and the industrial revolution in Britain before building the notion of mode of production.

By the middle of the nineteenth century, the cartographic model was increasingly replaced by models each with a logic dictated by the dynamic, immediate, interactive phenomena under study: the new specialized disciplines broke their earlier links with spatial thinking.

The Functional Approach to the Relations between Space, Social Life, and Thought

When social and behavioral sciences severed their former links with distributional approaches, they generally considered space as a neutral stand and ceased to use it as a significant factor for explaining what they observed. Many scholars thought, however, that space mattered: they had the idea that social life differed according to its physical setting, the fact that people lived in empty or overpopulated areas, and were more or less distant from each other. For Emile Durkheim the foundation of sociology lay in social morphology, but he was unable to provide a clear picture of it. The only way to take it into account was to adopt a functional perspective: space could be conceived of either as a collection of environments or a set of distances.

Space as Environment

In the environmental perspective, the emphasis was given to the local connections between people, artifacts, and place. The first environmental approaches were developed by the sensualist psychology of the eighteenth century. Man was a product of whatever his senses impressed upon his mind. His aptitudes and personality reflected his environment. If it was disorderly or unsound, social life had many reasons to be stressed. By the end of the eighteenth century, social scientists were well aware of the social and political implications of such a conception. Jeremy Bentham conceived his *Panopticon* in order to provide criminals with an environment, which would teach them safer forms of behavior. More generally, the idea that 'to reform society, you had to change its material setting' was central to the nineteenth-century utopian socialism and still present in the conception of urbanism developed by the International Movement of Modern Architecture in the 1930s.

Johann Gottfried Herder was more sensitive to the harmony, which developed between peoples and the country they lived in than to the sensual influence of these surroundings on their mind and behavior. The early development of geography in Germany was linked to the combined influence of Kant (evident in Alexander von Humboldt's conception of geographical description) and Herder (who was closer to Carl Ritter). Such a conception was important for all the social sciences in Germany, since it was conducive to a keener interest in peoples than in individuals or social classes. For many German geographers, the harmony between a people and the land they lived in was expressed through its landscapes.

The Herderian environmentalism was a literary and poetic one. Natural evolutionism had a more direct impact on the social sciences, since it was a scientific theory. The early forms of evolutionism as developed in the eighteenth and early nineteenth centuries had no direct impact on the then emerging social sciences. The publication of *The Origin of Species* occurred at a time when history and the idea of progress had become popular in the public opinion, and original forms of evolutionism had developed in sociology: hence the impact of its theses in the social sciences. For sensualist psychology or Herderian philosophy, the influence of things upon the human mind was a direct one. For scientific evolutionism, the relations were mediated through biological mechanisms.

Did environments shape completely the destiny of humankind and of any of its groups? Hot discussions developed between environmentalists and the scientists who believed in human liberty. In order to depict the relations between organisms and their environment, a new discipline, ecology, appeared. Named by Haeckel in 1866, its first scientific achievements occurred at the end of the nineteenth century; it became mature in the 1930s and 1940s. Geographers did not wait for the development of ecology to propose a new conception of their discipline: *Anthropogeographie*, or human geography, was named and shaped by Friedrich Ratzel in the 1880s and 1890s. For him and for the first human geographers in France, a strict environmentalism was an unbearable position, but physical constraints did exist and shaped at least partly human behavior and history: geographers pleaded for possibilism, which gave room to physical and biological constraints, but acknowledged the role of other variables.

By 1920, it had become evident that environmentalism was not a very fruitful way of conceiving the role of space in social functioning.

Space as Distance

Economics had developed another approach to analyze the role of space in the functioning of society. In the early nineteenth century, some economists were unhappy with the elimination of space from mainstream theory. They developed spatial economics as a sideline. It was centered on two problems: the uneven distribution of natural factors and the role of distance. The name of David Ricardo is associated with the first orientation and the ensuing theory of international specialization and trade. Johann Heinrich von Thünen explained how distances to the market determined farming specialization even in a perfectly uniform setting (1826–52). The progress of spatial economics was a slow one: people had to wait until Alfred Weber for a coherent theory of industrial location (1909), and Walter Christaller and August Lösch for an understanding of the location of service activities and the nature of urban networks (early 1930s).

At the end of the nineteenth and the beginning of the twentieth centuries, some anthropologists also considered distance as a major factor of cultural differentiation. It soon appeared that diffusionism could not serve as an overall explanatory theory. Diffusion was, however, a significant cultural process.

When geographers, tired from the hopelessly descriptive character of the possibilist approach, tried in the late 1950s and 1960s to develop a more systematic view of their scientific enterprise, many of them decided that geography should become a discipline of distances: they incorporated what spatial economics and anthropology could offer and went further through the emphasis they gave to communication processes. They used what a new scientific field, the study of communication, was developing.

The functional approach to social problems through the analysis of distance was fruitful. Its role in planning theory and practice was essential during the 1960s and 1970s. It allowed for the design of well-organized cities or suburbs. It did not prevent, however, the development of urban pathologies. In order to understand them, other approaches to the role of space in social life had to be developed.

Space as Experience

Functional approaches were specific to the social sciences. They did not offer many possibilities to the cognitive sciences. The situation is different for those who try to analyze space as experience.

Environmental Psychology

The curiosity of psychologists for spatial problems was initially focused on the problems of orientation and place recognition. Their interest changed with the genetic approach developed by Jean Piaget: his problem was to understand the way in which children explored their environment and built behavioral schemes and mental categories out of their experience and inborn capabilities.

The relations between man and environment are mediated through the senses: what are the specific contributions of sight, hearing, smell, touch, taste, and synesthesia to the individual experience of space? Michel Foucault, for instance, was fascinated by the place given to sight by Western civilizations and its role in social control.

Because of its emphasis on experimental methods, mainstream psychology ignored the social dimension of spatial learning and experience. Phenomenologists, Gaston Bachelard or Maurice Merleau-Ponty for instance, brought them back. Environmental psychology introduced the theme of proxemics.

The Psychological, Moral, and Religious Significance of Space

Through phenomenology and proxemics, space started to be conceived as a necessary component of communication systems: hence the contribution of linguistics and semiotics to spatial thinking. Landscapes ceased to be analyzed in a functional perspective. They were conceived of as supports for messages. People experience different forms of connivance with them.

A difference exists between analytical and symbolic communication: in the first case, the problem is to transfer complex sets of information from individuals to individuals; in the second one, it is to stir simultaneously many persons through a unique and short message. Symbolism appears in this way as a tool for mastering distance: hence the interest of geographers for this form of communication.

Space has other social meanings. Normative thinking relies on perspectives from the outside: people have to imagine 'Beyonds,' either within beings or things (when the way of immanence is chosen) or outside them, generally in another world (when the way of transcendence is selected). These choices are important since they introduce a qualitative differentiation of space: places where the Beyonds touch the surface of our world are loaded with sacredness, in contrast to the surrounding profane areas. The building of Utopias, the beliefs in the Golden Age, or in a Land without Evil are ways to use history or terrestrial space to build other types of Beyonds. The dreams they nurture are important in the shaping of rural or urban areas.

The new approaches to geography developed since the 1970s incorporate this interest in the human experience of space.

Social Experience as a Localized Experience

People cannot develop an experience of society outside specific places. The *Time Geography* developed by Torstein Hägerstrand stresses the fundamental diversity of spatial and social experiences developed by people along their life trajectories: it means that social experiences differ according to persons. In the field of communication, for instance, people participate in different intersubjectivity circles depending on the places they live in or had lived.

Such an observation is fundamental for all the social sciences: it reminds that societies, cultures, polities, languages do not exist in a social realm located out of space. They differ from place to place, because of the existence of *locales*, as expressed by Anthony Giddens. The social sciences of today can no more ignore spatial thinking: it is one of the main lessons of the postmodernist epistemological revolution.

The Spatial Turn in the Social Sciences

In the 1970s, many social scientists became dissatisfied with the prevailing conceptions of society and its surroundings. They criticized the static dimension inbuilt in structural theory and its elimination of human agency. They wished to better understand the way social space was built. They deplored the lack of interest in the natural dimension of human environments: hence the development of approaches more open to the spatial dimension of social realities.

The Structurationist Approach

How to give room to human agency? Structural theory envisioned societies as closed and coherent systems: it explained

their organization and functioning, but ignored human initiatives. History was rubbed out. Since the structures upon which emphasis was put were mainly social or economic, space was also eliminated.

Structuralism was built as a one-level theory. Structurationism was conceived as a two-level one: structures remained important for understanding the way social systems function at a given time, but human agency, which was analyzed at another scale, was able to modify the equilibrium of forces, which characterized them. It eventually was responsible for the rise of new structures, which reintroduced the time dimension into social explanation.

Anthony Giddens was interested in Hägerstrand's time geography, since it showed that society never exists as an overarching reality. It is made of a plurality of places. Things, which happen in them, are sometimes linked together. Societies are built as aggregates of *locales* where events are shaped by similar institutions and patterns of behavior. In this way, space was reintroduced into social approaches. The influence of Giddens' sociology was strong in the English-speaking countries.

The Production of Space and Postmodernity

Henri Lefebvre was a Marxist. As a sociologist, he had a keen interest in the specificities of rural and urban areas and was dissatisfied with the lack of interest of the French Communist Party in these problems. For him, Marx's principles of social and economic interpretation were right, but they were developed in another time. They have to be adapted to contemporary situations.

Nature has been transformed by human beings since their first appearance on Earth, but many of its aspects were preserved in traditional societies. With today's overall urbanization, the situation deeply changed. Men are creating a second nature. They are producing the space in which they live.

Fredric Jameson accepted Lefebvre's main conclusions. For him, postmodernity differed from modernity by the growing emphasis it put on spatial problems, and its lesser interest in history.

Environment and Society

The third component of the move back to space grew from the new concern with nature and environment preservation. Human beings do not exist in the ether. They are living organisms. They are part and parcel of nature. Their fate is linked with that of the Earth. With the growing human population and its increased consumption patterns, the pressure on natural environments is higher than ever. Many local or regional environments have lost their resilience, and global imbalances are threatening the future of nature and humanity.

Hence the necessity to pay more attention to the corporeal dimension of human beings, the way they insert in natural settings and their carbon footprint. Space ceases to be conceived of as a neutral support or an obstacle to movement: it becomes a living reality, which has to be preserved if humanity wishes to survive.

During the last 40 years, the interest of social sciences in space has considerably grown. This movement has three components: the first stresses the role of communication in the structuration of social life; the second focuses on the way societies generate the space in which they evolve; and in the third, space is the living environment, which ensures every organism – the human ones in particular – with the elements, air, water, nutrients they need.

See also: Anthropology, History of; Functionalism, History of; Geography: Overview; Social Question: Impact on Social Thought.

Bibliography

Bachelard, G., 1957. La Poétique de l'Espace. Presses Universitaires de France, Paris.
Dagognet, F., 1977. Une Epistémologie de l'Espace Concret. Vrin, Paris.
Eliade, M., 1957. Das Heilige und das Profane. Rohwolt, Hamburg.
Giddens, A., 1984. The Constitution of Society. Polity Press, Cambridge, UK.
Gould, P., White, R., 1974. Mental Maps. Penguin, Harmondsworth, UK.
Jameson, F., 1991. Postmodernism and the Cultural Logic of Late Capitalism. Verso, London.
Lee, T., 1976. Psychology and the Environment. Methuen, London.
Lefebvre, H., 1974. La Production de l'espace. Anthropos, Paris.
Livingstone, D.N., 1993. The Geographical Tradition. Blackwell, Oxford, UK.
Merleau-Ponty, M., 1945. Phénoménologie de la Perception. Gallimard, Paris.
Piaget, J., 1926. La Représentation du Monde Chez l'Enfant. Alcan, Paris.
Ponsard, C., 1984. History of Spatial Economic Theory. Springer-Verlag, Berlin.
Proshansky, H.M., Ittelson, W.H., Rivlin, L.G. (Eds.), 1970. Environmental Psychology. Holt, Rinehart and Wilson, New York.

Spatial–Temporal Modeling

Emilio Casetti, Ohio State University, Columbus, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spatial–temporal Models (STMs) formalize the variation of phenomena over time and across space. In the last 20 plus years, the scholarly trends most relevant to STMs have been the growth in importance, relevance, and reach of spatial econometrics, of panel econometrics, and of spatial panel econometrics. In response to these trends, four of the six core sections of this article define and illustrate by examples the following classes: cross-sectional STMs, cross-sectional STMs with spatial effects, panel STMs, and panel STMs with spatial effects. These core sections are preceded by an 'atypical STMs' section and are followed by a section dealing with the 'construction of STMs.' The conclusion touches upon a number of trends that are having an impact on the evolution of spatial–temporal modeling.

Introduction

Spatial–temporal models (STMs) formalize the variation of phenomena over time and across space. A model qualifies as an STM if at least some of its variables are spatially and temporally referenced. Spatial–temporal referencing means that the values of these variables are linked to points or intervals in time and simultaneously, to entities in geographical space such as points, areal units, or arcs or nodes of a network. STMs may also include purely spatial and/or purely temporal variables. Examples of purely spatial variables are distances, directions, spatial coordinates, and spatial dummies. Examples of purely temporal variables are time and functions of time, dummies associated with time intervals, and points in time.

STMs are primarily concerned with three classes of phenomena: spatial flows, spatial diffusion, and spatial–temporal change. The last of the three includes spatial variation of growth rates (for instance, of product per capita), qualitative spatial change (for instance, in land use), and change in spatial distributions (for instance, of crime events or of settlements). A systematic introduction to the spatial–temporal modeling of several classes of substantive phenomena can be found in Haggett et al. (1977).

In the social sciences, much of the conceptual work that precedes or is intertwined with mathematical modeling is based on notions from system theory. Books dealing with the nexus between spatial–temporal modeling and spatial systems include those by Huggett (1980), Klassen et al. (1979), and Banks (1994). STMs figure prominently in the edited collections by Anselin and Florax (1995), Fischer and Getis (1997), Griffith and Haining (1986), Wrigley (1979), as well as Fischer and Getis (2010).

In the last 20 plus years, the scholarly trends most relevant to STMs have been the growth in importance, relevance, and reach of spatial econometrics, panel econometrics, and spatial panel econometrics. In response to these trends, four of the six core sections of this article define and illustrate the following 'classes': cross-sectional STMs, cross-sectional STMs with spatial effects, panel STMs, and panel STMs with spatial effects. These core sections are preceded by an 'atypical STMs' section and are followed by a section dealing with the 'construction of STMs.'

Atypical STMs

The term atypical STMs is used here for models that are spatial–temporal, but do not fit precisely enough in any of the four classes of STMs discussed in the following four sections.

The example of atypical STM discussed in this section is from a paper by Casetti (2003). The paper concerns a pattern of international power dynamics that historically played an important role in the rise and fall of major powers. This pattern consists in the early attainment of power supremacy by comparatively small countries that developed earlier and subsequently lost their status to larger countries that developed later. Cases in point are the staggered rise in power of the United Kingdom, France, Germany, Italy, and Japan, and their subsequent loss of status to the United States and the USSR.

The STMs that formalize this dynamic are based on the following propositions:

- A country's aggregate product is an index of its power.
- A country's product per capita is an index of its development level.
- During development, the percentage rate of growth of product per capita and of population tends to increase at first and then to decline.
- Development latecomers tend to experience greater rates of growth of product per capita and greater rates of growth of population than development pioneers did at comparative development levels.

To produce the power STMs, some notation and definitions have to be established. Let Q, Y, and P indicate a country's aggregate product, product per capita, and population, respectively. Q is also a measure of power. Let

$$q = \ln(Q) \quad [1]$$

$$y = \ln(Y) \quad [2]$$

$$p = \ln(P) \quad [3]$$

Apostrophe superscripts indicate the derivatives of the superscripted variables with respect to time, so that

$$y' = \frac{\mathrm{d}y}{\mathrm{d}t} = \frac{Y'}{Y} = \frac{1}{Y}\frac{\mathrm{d}Y}{\mathrm{d}t} \quad [4]$$

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72066-6

$$p' = \frac{\mathrm{d}p}{\mathrm{d}t} = \frac{P'}{P} = \frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}t} \quad [5]$$

$$q' = \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{Q'}{Q} = \frac{1}{Q}\frac{\mathrm{d}Q}{\mathrm{d}t} \quad [6]$$

The y', p', and q' are instantaneous percentage rates of change of a country's aggregate product, product per capita, and population, respectively. By definition

$$Q = Y * P \quad [7]$$

$$q = y + p \quad [8]$$

$$q' = y' + p' \quad [9]$$

The power equations are [10] and [11]. The ts in these equations stand for time.

$$y' = \alpha_0 + \alpha_1 y + \alpha_2 y^2 + \alpha_3 t + \varepsilon_y \quad [10]$$

$$p' = \beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 t + \varepsilon_p \quad [11]$$

Given a sample of countries, and for each of these countries given the values of y', p', y, y^2, and t, Eqns [10] and [11] can be estimated, and after a few manipulations, the value of the measure of power Q can be obtained for any or all the countries in the sample, and for any value of t.

Cross-Sectional STMs

Cross-sectional STMs are models with spatially referenced rates of change over time as dependent variables, and in which at least some of the independent variables are spatial or spatially referenced. Cross-sectional STMs can be found everywhere in the social sciences. For instance, the cross-sectional STMs in which the dependent variable is the rate of change of variables such as income, unemployment, population, welfare rolls, and crime rates, for countries, counties, or regions, and in which the independent variables include indices of relevant characteristics of these geographical entities are indeed common. Also very common are the models in which rates of change are functions of distances or of spatial coordinates.

Cross-sectional STMs have had a strong presence in the empirical literature on the determinants of economic growth. A specific 'case' is the regression models in Barro (1991, 1997). They were estimated using a sample of 114 countries, had growth rates of gross domestic product (GDP) per capita as dependent variables, and GDP per capita, male schooling, fertility rates, democracy indices, inflation rates, and others as independent variables.

Cross-Sectional STMs with Spatial Effects

Cross-sectional and panel STMs with spatial effects use techniques and concepts from spatial econometrics. For more on spatial econometrics, see Anselin (2001, 2007), LeSage and Pace (2009), and Getis (2010).

The GeoDa program provides users with an exposure to the practicalities of spatial econometrics. Interested readers can download free-of-charge GeoDa and related documentation from the site https://geodacenter.asu.edu.

The following example of cross-sectional STMs with spatial effects draws upon Fingleton (2007), which analyses the relation between changes in wages and levels of wages during 'globalization.'

The data used are wages per worker in 1996 constant prices, for the years 1970 and 2000, and for a sample of 98 countries. The year 1970 is assumed to mark the approximate onset of globalization. Denote by 'w' the wage per worker in 1970, and by 'D' the difference of the wage per worker in 2000 minus the wage per worker in 1970.

The theory reviewed in Fingleton (2007: pp. 394–397) suggests that the wage difference, D, is low at low wage levels; accelerates to a maximum as wage levels rise; and then decelerates with further wage level increases. In other words, the theory suggests a quadratic model

$$\begin{aligned} D &= \alpha w^2 + \beta w + \gamma + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned} \quad [12]$$

Here and hereafter, the Greek letters indicate parameters, while their Roman counterparts stand for estimated parameters. Upon estimation, eqn [12] portrays a parabola with a maximum if the estimated parameters a and b are significant, and if $a < 0$ and $b > 0$. The ordinary least squares estimates of model [12] parameters given in Fingleton (2007: p. 399, Table 16.1) show that b is significant and greater than zero and that a is smaller than zero but is not significant. Fingleton (2007: p. 399) comments that "a quadratic function does not seem to describe the relation between … [D and w]." Also, on the basis of a Moran's I test, he remarks that "There is evidence of very significant residual spatial autocorrelation." In other words, the estimated quadratic is not congruent with theory.

The line of thought that Fingleton follows to render the model congruent with theory can be summarized in three propositions:

- Globalization is associated with a growing interdependence of national economies.
- These interdependencies translate in a plurality of 'shocks' that every country receives from its 'neighbors.'
- This interdependence of economies can be explicitly modeled via the spatial error model (Fingleton, 2007: p. 398).

On the basis of these propositions, eqn [12] was converted into its spatial error counterpart below:

$$\begin{aligned} D &= \alpha w^2 + \beta w + \gamma + u \\ u &= \rho W u + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned} \quad [13]$$

To estimate eqn [13], a specification of the spatial weight matrix W is required. W is a 98×98 matrix with its rows and-columns linked to the sample's countries. Indicate by W_{ij} the element of W at the intersection of the ith row and jth column. Equation [13]'s Maximum Likelihood (ML) parameter estimates reported in Fingleton (2007: p. 400, Table 16.2, column 3) is based on the following specification of W: $W_{ij} = 0$ if $d_{ij} > 10\,000$ miles; $W_{ij} = 1/d_{ij}$ if $d_{ij} < 10\,000$ miles, where d_{ij} is the great circle distance between the capital cities of countries i and j.

The ML estimated parameters of w and w^2 are both significant and are, respectively, greater and lesser than zero, as theory suggested that they should be. Actually, Fingleton experimented with several weight matrices, some 'not row standardized' as the one described above, and others 'row standardized,' and in all cases they produced parameter estimates consistent with theory. The message emanating from Fingleton's paper is clear: the spatial effects augmentation of an inadequate model may render it adequate. Note that here both the original model and its augmentation are based on solid theoretical foundations.

Panel STMs

In the past 20 plus years, spatial temporal modeling based on panel data has seen considerable growth. Evidence of this growth includes the number of panel STMs books and papers published in recent years, the variety of scientific fields in which panel STMs have been used, and the proliferation of panel STM linkages with other branches of econometrics. Prominent among these linkages is the one between panel STMs and spatial econometrics.

The fast spread of panel STMs has been credited to factors that include the following:

- Spatial panel data have become increasing available.
- The panel STMs proved effective in investigating issues that cannot be addressed by either cross-sectional or time series analyses alone (Hsiao, 2003; Baltagi, 2005).
- Panel STMs tend to use a larger number of data points, which increases the degrees of freedom and reduces collinearity (Hsiao, 2003: pp. 1–2).
- The panel literature dealing with, or related to, 'fixed effects' and 'random effects' proved capable of controlling the heterogeneity resulting from the absence of unavailable or unknown variables (Baltagi, 2007; Arellano, 2003: pp. 7–29; Baltagi, 2005: pp. 4–5 and 11–47).

The Baltagi and Pinnoi (1995) paper on the relationship between public infrastructural investments and private sector productivity will be used to illustrate an application of panel STMs. The panel data sets are time series of cross-sections. Here, each cross-section is made of 48 data vectors, each linked to one of the 48 contiguous US states. The panel data consists of 17 cross-sections, one for each of the 17 years in the time interval 1970–86. Consequently, the total number of data vectors is 816.

Each data vector contains the state name and the year to which the vector is linked, plus the values of eight variables for that state and year. The variables and their names are defined as follows:

1. Y = Gross state product
2. K = Private capital stock
3. L = Labor inputs (employment in nonagricultural pay rolls)
4. U = State unemployment rate
5. KH = Value of public capital in highways and streets
6. KW = Value of public capital in water and sewers facilities
7. KO = Value of public capital in other (public buildings and structures)
8. KG = Value of public capital in KH + KW + KO

All variables except U are in natural logarithms. The state unemployment rate, U, is entered in all models in order to attenuate the effects of temporal trends and of the business cycle. Y is the dependent variable in all the models in the paper. Let [14] and [15] denote two linear regression models defined by their dependent and independent variables:

$$Y = f(K, L, \mathrm{KG}, U) \qquad [14]$$

$$Y = g(K, L, \mathrm{KH}, \mathrm{KW}, \mathrm{KO}, U) \qquad [15]$$

These regression models are designed to determine whether the aggregate public capital KG in [14], or all its components KH, KW, KO in [15], contribute(s) significantly to the state's product. Both equations were estimated by the 'within estimator' of the 'fixed effects model' and by the 'between estimator' of the 'random effects model.' These models and estimators are discussed in chapters 2 and 3 of Baltagi (2005).

The estimates of [14] and [15] by 'between' regression can be found in columns 5 and 6 of Table 1 in Baltagi and Pinnoi (1995: p. 354), and the estimates of [14] and [15] by 'within' regression can be found in the same table in columns 8 and 9.

The interpretations of these four results are based on the well-established notion that 'between' and 'within' regressions indicate, respectively, long-run and short-run dynamics. For a review of this notion, see Baltagi (2005: pp. 200–205). The results of these four estimated regressions can be summarized as follows:

- The effects of the aggregate public capital, KG, in the short term are negative and nonsignificant, and in the long term are positive and significant.
- The effect of the three capital stocks, KH, KW, and KO, are not uniform. Namely, the effects of KH and KW are positive and significant in the short and long run. Instead, the effects of KO are negative and significant in the short run, and positive but not significant in the long run.
- The main conclusion is that the regression results from the [15] model give better insights into the complex interactions that shape the short-run and long-run effects of public investments on the private sector.

For a deeper analysis of these interesting results, see Baltagi and Pinnoi (1995: p. 355). Here, it will suffice to point out that their paper demonstrates very effectively how interesting and substantively important are the long-run short-run analyses that panel econometrics makes possible.

Panel STMs with Spatial Effects

The paper by Moreno and Lopez-Bazo (2007) will be used to illustrate an application of panel STMs with spatial effects. The two models in this demonstration are concerned with the effects of public investments on private sector productivity. However, the first model falls within the scope of panel econometrics, while the second is an implementation of the interface of panel and spatial econometrics.

In connection with model [16] the reader may find it useful to consult Baltagi (2005) and/or Hsiao (2003). In connection with model [17], in addition to these texts the

papers by Anselin et al. (2008) and Elhorst (2010) can provide a useful background.

The data panel used in this illustration consists of a time series of 17 cross-sections. Each cross-section is made of 50 data vectors associated with 50 Spanish provinces. The 17 cross-sections are linked to 2-year partitions of the 1975–97 time interval. Clearly the data vectors are spatially and temporally referenced, and are 850 in number.

Each data vector contains the values of seven variables. The variables are in natural logarithms, and are defined as follows:

1. Y = Manufacturing output
2. K = Capital inputs
3. L = Labor inputs
4. KL = Public sector 'local' capital stocks
5. KT = Public sector 'transportation' capital stocks
6. WKL = 'Local' spillovers
7. WKT = 'Transportation' spillovers

KL stands for capital stocks that produce effects locally, and KT stands for transportation capital stocks that produce effects both inside and outside the province in which they were created. A province's spillovers WKL and WKT are averages of the KL and KT values of the province's neighbors.

In terms of spillovers, infrastructural investments can produce effects directly in the province in which they are made, and indirectly as averages of the KL and KT values of this province's neighbors. Note that spillovers qualify as an example of the spatial econometrics' spatial lag variables.

The models reported below are discussed in Moreno and Lopez-Bazo (2007: p. 53). These models are

$$Y_{\mathrm{it}} = \beta_0 + \beta_1 K_{\mathrm{it}} + \beta_2 L_{\mathrm{it}} + \beta_3 \mathrm{KL}_{\mathrm{it}} + \beta_4 \mathrm{KT}_{\mathrm{it}} + m_{\mathrm{i}} + n_{\mathrm{t}} + \varepsilon_{\mathrm{it}} \quad [16]$$

$$\begin{aligned} Y_{\mathrm{it}} &= \beta_0 + \beta_1 K_{\mathrm{it}} + \beta_2 L_{\mathrm{it}} + \beta_3 \mathrm{KL}_{\mathrm{it}} \\ &= \beta_4 \mathrm{KT}_{\mathrm{it}} + b_5 \mathrm{WKL}_{\mathrm{it}} + \beta_6 \mathrm{WKT}_{\mathrm{it}} \\ &\quad + m_{\mathrm{i}} + n_{\mathrm{t}} + \varepsilon_{\mathrm{it}} \end{aligned} \quad [17]$$

The m_{i} and n_{t} terms are needed to include both region-specific effects and year-specific effects in the specification of the error term. The results from the estimation of models [16] and [17] can be found in Moreno and Lopez-Bazo (2007: p. 58, Table 2, columns 1 and 2).

In model [16], the estimated parameter of KL is significant, while the estimated parameter of KT is not. In model [17], the spillovers WKL and WKT have effects that are interesting and complex. First, the effects of KT in model [17] are stronger than in model [16], which suggests that its impact was underestimated in model [16]. Second, the WKL spillover is not significant, thus confirming that the effects of KL are local only. Third, the effect of WKT is significant and negative.

This third result is explained in Moreno and Lopez-Bazo (2007: p. 59) as the result of 'factor migration from regions with poor transport infrastructure endowment to regions with good endowments.' In other words, WKT reflects beneficial inflows of resources and skilled personnel into some provinces and damaging outflows of resources from other provinces. The negative and significant estimate of WKT's parameter indicates that in the balance the damages outweigh the benefits. This shows that here the integration of spatial econometrics with panel econometrics yields results that would have not been possible to obtain from either alone.

Constructing STMs

In many cases, STMs in the social sciences have been imported from non–social science disciplines. For instance, some of the diffusion models in geography and economics originated in physics and engineering (Banks, 1994). In general, the analogies between different phenomena may trigger interdisciplinary transfers of methodologies and models. The notions of systems and system dynamics and the schools of thought linked to them have played a major role in the construction of classes of STMs based on cross-disciplinary analogies.

However, in many circumstances the need may arise to construct STMs from purely spatial or purely temporal models, or to extend existing STMs to encompass relevant contexts that are outside their original scope. An approach to the incremental construction of mathematical models that can be easily applied to building STMs is the 'expansion method' (Casetti, 1997). In a nutshell, the expansion method involves widening the scope of a simpler 'initial model' by redefining some or all of its parameters into functions of expansion variables or variates that index substantively relevant contexts of the initial model. The terminal model thus produced encompasses both the initial model and a specification of its contextual variation.

To demonstrate the operation of the method, let us discuss its application to adding a time dimension to a purely spatial model of urban population densities. This demonstration is based on a variant of the model presented in Casetti (1973). The urban studies literature has identified two spatial patterns of urban population densities that are often referred to as density cones and density craters. In the cones, population densities decline exponentially with distance from city centers. Urban population densities that increase at first and then decline as the distance from the city center increases characterize density craters.

Let D, s, and ε denote, respectively, the natural logarithm of urban population density, distance from a city center, and an error term. Let

$$D = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \varepsilon \quad [18]$$

be an initial model subsuming both a density cone and a density crater as special cases. In fact, for $\alpha_1 < 0$ and $\alpha_2 = 0$ this model corresponds to a density cone, while for $\alpha_1 > 0$ and $\alpha_2 < 0$ it yields a density crater.

The literature has also identified distinctive patterns of temporal change in these spatial trends. They include the rising and flattening of the density cones, the widening of the density craters, and the gradual transformation of cones into craters. In order to extend the initial purely spatial model [18] so that it will encompass the spatial–temporal trends recognized in the literature, the α parameters are expanded by the stochastic expansion equations

$$\alpha_i = \alpha_{i0} + \alpha_{i1} t + \theta_i, \quad i = 0, 1, 2 \quad [19]$$

where the θs are error terms. The terminal model

$$D = \alpha_{00} + \alpha_{01}t + \alpha_{10}s + \alpha_{11}st + \alpha_{20}s^2 + \alpha_{21}s^2t + u \quad [20]$$

$$u = \varepsilon + \theta_0 + \theta_1 s + \theta_2 s^2$$

is obtained by replacing the right-hand sides of the expansion equations for the αs in the initial model. Upon estimation, this terminal model can be used to test for the occurrence of specific spatial–temporal trends and to portray these trends if they occur.

The model in this example is intended for estimation and testing statistical hypotheses. It should be noted though that the expansion method can be used to construct STMs of any kind. Any model that has parameters can be expanded, irrespective of whether it is deterministic, stochastic, or mixed, and irrespective of whether it is intended for solving, estimating, or optimizing.

Concluding Comments

Most STMs in the social sciences came into existence in the past 70-odd years, and have been influenced or shaped by trends that kept altering the landscape of the quantitative analytical social sciences during this time horizon. The comments that follow highlight some such trends and influences.

Mathematical models in general, and STMs in particular, have been acquiring an important role in government and private sector activities as aids to decision making, to planning, to policy development and evaluation, and for forecasting. In the process, they have been driven to address the complexities of the real world to an ever-increasing degree. The tendency toward more complex models is especially apparent in all kinds of STMs. Their proliferation is a reflection of their practical usefulness. The movement toward the computational counterparts of STMs that originated as purely mathematical formalisms is another aspect of these tendencies.

Increasingly complex STMs have also played an important role in the precise articulation of theories and conceptual frameworks that are important components of the social sciences. In the more theoretical applications, this growing complexity has been intertwined with conceptual and technical advances imported from mathematics and a variety of sciences. STMs related to nonlinear dynamics, and ones centering on optimal control techniques, are cases in point.

The entire evolution of STMs and of mathematical modeling at large has been permeated by the growth of stochastic modeling and by the penetration of stochastic elements into models born deterministic. For instance, deterministic cross-sectional STMs have been converted into 'mixed' or stochastic models by the addition of stochastic errors and/or of temporally or spatially autoregressive terms. An adequate stochastic specification and its attendant statistical econometric testing has increasingly become a must for empirical analyses to be credible.

See also: Mathematical Models in Geography; Regional Science; Spatial Analysis in Geography.

Bibliography

Anselin, L., Florax, R., 1995. New Directions in Spatial Econometrics. Springer, Berlin.

Anselin, L., 2001. Spatial econometrics. In: Baltagi, B. (Ed.), A Companion to Theoretical Econometrics. Blackwell, Oxford, pp. 310–330.

Anselin, L., 2007. Spatial econometrics. In: Mills, T., Patterson, K. (Eds.), Econometric Theory. Macmillan, New York, pp. 902–969.

Anselin, L., Le Gallo, J., Jayet, H., 2008. Spatial panel econometrics. In: Matyas, P., Sevestre, P. (Eds.), The Econometrics of Panel Data. Springer, Berlin, pp. 625–660.

Arellano, M., 2003. Panel Data Econometrics. Oxford University Press, New York.

Baltagi, B., Pinnoi, N., 1995. Public capital and state productivity growth: further evidence from an error component model. Empirical Economics 20, 351–359.

Baltagi, B., 2005. Econometric of Panel Data. Wiley, Chichester, England.

Baltagi, B., 2007. Panel data models. In: Mills, T., Patterson, K. (Eds.), Econometric Theory. Macmillan, New York, pp. 633–661.

Banks, R., 1994. Growth and Diffusion Phenomena: Mathematical Frameworks and Applications. Springer, Berlin.

Barro, R., 1991. Economic growth in a cross section of countries. The Quarterly Journal of Economics 106, 407–443.

Barro, R., 1997. Determinants of Economics Growth: A Cross-country Empirical Study. MIT Press, Cambridge.

Casetti, E., 1973. Testing for spatial trends: an application to urban population density trends using the expansion method. Canadian Geographer 17, 127–137.

Casetti, E., 1997. The expansion method, mathematical modeling, and spatial econometrics. International Regional Science Review 20, 9–33.

Casetti, E., 2003. Power shifts and economic development: when will China overtake the USA? Journal of Peace Research 40, 661–675.

Elhorst, P., 2010. Spatial panel data models. In: Fischer, M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis. Springer, Berlin, pp. 377–407.

Fingleton, B., 2007. Modeling globalization: a spatial econometric analysis. In: Cooper, R., Donaghy, K., Hewings, G. (Eds.), Globalization and Regional Economic Modeling. Springer, Berlin, pp. 393–416.

Fischer, M., Getis, A. (Eds.), 1997. Recent Developments in Spatial Analysis. Springer, Berlin.

Fischer, M., Getis, A. (Eds.), 2010. Handbook of Applied Spatial Analysis. Springer, Berlin.

Getis, A., 2010. Spatial autocorrelation. In: Fischer, M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis. Springer, Berlin, pp. 255–275.

Griffith, D., Haining, R., 1986. Transformations through Space and Time: An Analysis of Nonlinear Structures, Bifurcation Points and Autoregressive Dependency. Martinus Nijhoff Publishers, Dordrecht, The Netherlands.

Haggett, P., Cliff, S., Frey, A., 1977. Locational Analysis in Human Geography. Wiley, New York.

Hsiao, C., 2003. Analysis of Panel Data. Cambridge University Press, Cambridge, UK.

Huggett, R., 1980. System Analysis in Geography. Clarendon Press, Oxford, UK.

Klassen, L., Paelinck, J., Wagenaar, S., 1979. Spatial Systems. Saxon House, Westmead, UK.

LeSage, J., Pace, K., 2009. Introduction to Spatial Econometrics. CRC Press, Taylor Francis Group, Boca Raton, London, New York.

Moreno, R., Lopez-Bazo, E., 2007. Returns to local and transport infrastructure under regional spillovers. International Regional Science Review 30, 47–71.

Wrigley, N. (Ed.), 1979. Statistical Applications in the Spatial Sciences. Pion, London.

Special Economic Zones and Cluster Dynamics: China

Lauretta Rubini and Marco R Di Tommaso, University of Ferrara, Ferrara, Italy
Elisa Barbieri, University of Udine, Udine, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Clusters of specialized firms are now coming to be recognized as an important level of integration and organization in successful industrial economies. Such clusters, among which are industrial districts, were prominent in nineteenth-century European industrial development (in Italy, Britain, and Germany) and in the United States, and then in Japan, and are now viewed as being central to successful industrial development in the twenty-first century in China and India, where they are specifically promoted through ad hoc policies. In designating special economic areas, such as geographic clusters of firms or special economic zones (SEZs), national and regional authorities aim to attract and stimulate economic activities in specific, bounded areas of the country, where business-friendly rules can apply. After presenting the rationale and main features of SEZs, the article illustrates the experience of China, and in particular Guangdong Province and the Pearl River Delta, where SEZs based on Shenzhen and other urban centers have played a prominent role. A careful analysis of the policies adopted in the country to kick off and manage the transition process highlights a central concern of the government to manage the spatial patterns of industrial development. Specific tools have been developed in order to shape the geography of industry: cluster promotion has been pursued while also managing a number of SEZs with the specific aim to concentrate the early stages of industrial development in confined areas, learning from the prior experience of industrial districts in Europe, the United States, and Japan. In China, and especially in Guangdong, the development of SEZs and the promotion of clusters of firms have gone hand in hand to achieve long-term industrial development objectives.

Introduction

In spite of its prominence in neoclassical economic literature, 'the firm' is not the only institution relevant to production processes. Classical scholars such as Alfred Marshall in the nineteenth century understood clearly that what happens outside the firm, through external as well as internal economies, could be decisive to industrial success. Marshalls, pioneering studies of industrial districts in Britain and Germany have been taken up in the twentieth century in Italy by Becattini (1989), and in general terms by Coase (1937) and Williamson (1975), to cite just two of the distinguished authors who have tackled this subject. Supra-firm organizations emerged since the first English industrial revolution (Marshall, 1890/1920/1972), and they can hardly be considered exceptions. Clusters of firms, industrial districts, networks of companies, and business groups are all common manifestations of such supra-firm organizations.

Although each supra-firm structure has its own peculiarities, they also share some common features. In a sense, they all deal with the costs associated with market transactions without necessarily bearing the diseconomies inherent in large integrated organizations. Moreover, they all refer to a common idea that a higher degree of (collective) efficiency can be reached by independent single firms thanks to the exploitation of economies that these supra-firm organizations bring in their wake. Yet they can differ because of varying degrees of social embedding of the actors involved or because of the presence (or absence) of explicit collective actions.

In this scenario, the spatial dimension is an important additional distinguishing factor. In the case of industrial districts – or, more generally, in the case of clusters of specialized firms – spatial proximity allows firms to exploit Marshallian externalities (a constant market for skills, local specialized suppliers of goods and services, and access to specialized knowledge) as well as other external economies arising from explicit collective actions of neighboring firms.

For these reasons, development policies have tried to promote agglomerations of specialized firms in particular localities. Examples are the many European initiatives in favor of cluster development and consolidation (tracked, e.g., by the European Cluster Observatory, www.clusterobservatory.eu); the experience of the British biotechnology industry of Cambridge (Waters and Lawton-Smith, 2008); the interesting story of the high-tech pole of Sophie Antipolis in France; the 'one-city one-product' Chinese specialized towns (STs); the Brazilian shoe cluster in Synos Valley; the clothing cluster in Nairobi; the electronic engineering cluster in Bangalore, India; and many more.

There are three basic intuitive ideas behind these diffused kinds of policy interventions. First, clusters might offer unique advantages to firms because of geographical proximity within a specific area. Second, such proximity might also encourage important profitable collaborations among firms. These advantages make clusters 'special territories' for firms. Thus, policies in favor of cluster development can also be read as a way of offering firms a special site (zone, area, etc.) with unique favorable conditions. Finally, these territories are also special not only because in these sites firms cluster together but also because a variety of other relevant actors (public and private institutions, universities, laboratories, government agencies, etc.) agglomerate and firms are enabled to specialize (Mathews, 2012).

In this scenario, policies in favor of the establishment of special economic zones (SEZs) – especially in the form of high-tech development areas or dedicated industrial parks – can be understood from a similar perspective. They in fact display a strong geographical component where national and regional

authorities aim to attract and stimulate economic activities in specific, bounded areas of the territory. The aims to be achieved through such instruments are many, as will be briefly described in the section SEZs: Rationale and Organization. Here we wish to emphasize the implicit idea of these kinds of policies: attracting firms in a specific location will also create favorable conditions for the emergence of the above-illustrated cluster economies. Given that these economies are advantages that are available to firms localized in a specific area and usually specialized in the same economic sector (or a highly related one), it is not unusual to find SEZs or industrial parks that identify specific priority sectors.

For instance, in the case of Shenzhen, China, which is reviewed in the section Main Features for Successful (and Unsuccessful) SEZs, the concentration in the city of firms operating in the electronic sector, favored by the special privileges granted by SEZ status, has undoubtedly facilitated the flourishing of common services, the attraction of specialized workforce, and other sector-specific advantages that have multiplied the effects of the SEZ establishment per se. This setting has then reasonably encouraged joint actions and collaborations among a variety of relevant and specific actors, offering to the new clustered local economy unique competitive advantages.

The more general experience of China, and in particular that of Guangdong Province, seems to be particularly relevant in this context. The exceptional rate of growth that, starting from Guangdong Province, has characterized the whole country (Di Tommaso et al., 2012) calls for attention to the specific policy tools adopted to kick off and govern the transition process started at the end of the 1970s. A careful analysis of such policies highlights a main concern of the government toward the geography of production and the spatial patterns of industrial development. Specific tools throughout the different stages of growth have been developed in order to shape the geography of industry: cluster promotion has been pursued with ad hoc instruments while also managing a number of SEZs with the specific aim to concentrate the early stages of industrial development in certain designated areas. In China, and in Guangdong more than in other places, the development of SEZs and the promotion of clusters of firms have gone hand in hand in order to achieve long-term industrial development objectives.

SEZs: Rationale and Organization

The rationale behind SEZ establishment can be traced back to the eighteenth century, with free zones in Gibraltar (1704), Singapore (1819), Hong Kong (1848), Hamburg (1888), and Copenhagen (1891) (World Bank (WB), 2008). However, the first 'modern' SEZ was established in Ireland in 1959, followed by SEZs in lower-income countries, among them the Dominican Republic, India, and Taiwan. In the last 30 years, SEZs have become an increasingly important instrument of export promotion and economic development in both industrialized and industrializing countries. According to a database developed by the International Labour Organization (ILO), in 1975 there were 79 zones in 25 countries (all government owned), which grew to 3500 zones in 130 countries 20 years later (interestingly, more than half of which are private) (WB, 2008).

SEZs can be defined as 'demarcated geographic areas contained within a country's national boundaries where the rules of business are different from those that prevail in the national territory' (Farole, 2011, p. 23). This broad definition encompasses several typologies, which differ according to the kind of granted privileges, target firms, final market of destination, operational rules, and internal organization.

The main ones are the following (WB, 2008):

- Free trade zones (FTZs) (or commercial free zones). These are delimited duty-free areas offering all the facilities for trade, transshipment, and reexport. Usually in these areas, there are weaker customs controls.
- Export processing zones (EPZs). These industrial estates are for the location of export-oriented enterprises. They do not necessarily offer the same tax benefits as FTZs, but they do offer a functional advantage, especially for foreign exporters. There are also hybrid EPZs, divided into a separate area for export-oriented firms and a zone open to all industries independent of their export orientation. Normally, EPZs require that a specific percentage of production is exported. EPZs are a type of SEZ that has been studied more in the literature, and sometimes the two terms are used synonymously.
- Freeports. These are larger areas hosting all kinds of activities, from tourism to retail sales, and providing a wide set of incentives. For this reason, these areas tend to be more integrated in the host economy.
- Industrial parks (or enterprise zones or industrial development zones). These areas have the aim of revitalizing relatively underdeveloped urban or rural areas. Here the free trade component is less relevant, while special attention is given to infrastructures and services for industries. They are usually used in developed countries.
- Export processing factories (EPFs) (sometimes called enterprise zones or single factories). These are a case of ad hoc incentives granted to a specific firm. In this case, the firm operating under the privileges of the SEZ is not physically located within an industrial estate, but the boundaries of the zone coincide with the boundaries of the factory.
- Other specialized zones such as science and technology parks, logistic parks, airport-based zones, and so on.

The usual privileges from which firms benefit in SEZs are as follows (Farole and Akinci, 2011):

1. tax exemption and tax holiday periods, ranging normally from 5 to 20 years, although there are cases, such as Jamaica, where the tax holiday period is unlimited;
2. unlimited, duty-free imports of raw materials, equipment, machinery, and intermediate goods for the production of exports;
3. duty-free on exports;
4. liberal foreign exchange policies;
5. less governmental red tape and more flexibility with labor laws for the firms in the zones than for those outside;
6. above-average communication services and infrastructure such as roads, power supplies, and transport facilities together with, in some cases, subsidized utilities and rental rates;

7. no limitation on foreign ownership of the firms or on the repatriation of profits in most cases;
8. other incentives such as value-added tax (VAT) exemption and training subsidies.

The role assigned to SEZs differs from country to country. For instance, in industrializing economies such as South Korea, Malaysia, and Thailand, SEZs were established as part of a policy shift from an inward- to an outward-oriented economic policy. In these cases, SEZs, and in particular EPZs, were meant to evolve and integrate with the domestic economy until they would no longer be enclaves.

According to many theoretical contributions, these areas offer a liberal and free trade environment, while in many cases (first of all in China, which will be described in the section Main Features for Successful (and Unsuccessful) SEZs, but also elsewhere) SEZs have been a way to manage the transition, opening only specific areas while keeping the rest of the country closed. In other countries, such as the Philippines and Indonesia, SEZs were introduced as 'safety valves' in a context of inward-oriented economic policy in order to create at least some areas free from distortions. In this case, the appealing feature of SEZs is that the host country can enjoy several benefits in terms of job creation, improvement of the balance of payments, and absorption of technology and at the same time it can keep the domestic industry protected from foreign competition. More recently, SEZs have been used by countries where foreign direct investment (FDI) was particularly low. In these cases, they have been an instrument to compete with neighboring countries offering better incentives to foreign investors.

To summarize, and in line with the relevant literature (Warr, 1990; Johansson and Nilsson, 1997, Farole and Akinci, 2011; WB, 2008), the main objectives of SEZs have traditionally included:

- to provide foreign exchange earnings by promoting exports;
- to create jobs, in order to reduce unemployment;
- to attract FDI to the host country;
- to accelerate the industrialization process, thanks in part to desirable spillover effects on the local firms that are incentivized to sell goods and services to zone-based enterprises since these sales are exports and can benefit from export incentives;
- to accompany and support wider economic reform.

Main Features for Successful (and Unsuccessful) SEZs

Given that SEZs have had a variety of different policy goals, the literature has highlighted both successes and failures.

One of the principal challenges for SEZs has been the relationship between foreign companies and local producers (Johansson and Nilsson, 1997). In contemporary markets in both industrialized and industrializing countries, SEZs must be considered as platforms that are able to connect global and local actors. In other words, SEZs have to be 'glocal.' SEZs with only local producers separated from global producers or SEZs with only foreign producers isolated from the local economy could hardly be defined as successful. Effective SEZs are the ones that are able to play a crucial local–global binding role in the name of national and local industrial development policy targets. Multinational enterprises and FDI can – or cannot – exchange with domestic companies located within (or in proximity of) that specific area modern technology, export know-how, ideas, or knowledge on production, marketing, and distribution, thus stimulating and improving local production. One of the major obstacles to local industrial development is, in fact, the lack of 'capacity to package' the technical, marketing, and managerial know-how with domestic and external resources. If SEZ firms succeed in creating sufficient backward links with domestic firms, they can fill this gap (Johansson and Nilsson, 1997). The integration with the local economy is the most important and also the most difficult task for SEZs, and only a few countries have succeeded significantly in this respect.

A potential failure for policies based on the SEZ establishment is also connected to the negative effects that SEZs might have outside their boundaries. The existence of SEZs should not penalize the rest of the companies operating outside. The problem can be severe especially if inside-the-SEZ foreign companies compete (exploiting the advantages of preferential policies) with outside-the-SEZ domestic companies that cannot count on the same advantages. An interesting experience in this respect is that of the so-called equal footing policies (i.e., all the measures aim at granting the same benefits to SEZs' foreign exporting firms and to local exporting firms even if they are located outside the SEZ).

Other factors determining the success or failure of SEZs are indicated by the literature, such as the local economic environment or a good locational choice. Where SEZs are located can make the difference between success and failure. In other words, geography matters for SEZs.

Moreover, successful SEZs offer agglomeration economies that companies located in one specific locality might (or might not) exploit. SEZs are also agglomeration of firms, institutions, and people. Successful SEZs offer advantages connected to spatial proximity. And not only spatial proximity matters, but also specialization does. The international cases show that the more specialized SEZs are, the more likely they are to be successful in the medium to long run. This is probably due to the fact that besides the specific privileges granted by the SEZ establishment, there is also an incremental effect due to the exploitation of the above-mentioned cluster economies.

However, SEZs should be considered successful (or not) with reference to one specific national industrial policy framework. As has commonly been described, SEZs have been free-market and trade-opening tools. And in this field, in many circumstances, SEZs have been successful, playing an important pilot role. However, SEZs have also been an important instrument for governments interested in maintaining a role in guiding national industrial development and structural change. In other words, SEZs should also be evaluated by focusing on other long-run strategic-planning and policy design issues. SEZs not only have been tools for promoting 'free-market islands,' but also, in contrast, have been instruments utilized by interventionist governments involved in industrial issues. In the Chinese experience, it is possible to find both rationales in different historical phases and in different locations (Di Tommaso et al., 2012). In many other countries, the establishment of SEZs

has been functional to industrial policy strategic needs. Very often, and especially in recent times, SEZs have played a dynamic strategic role in responding to continuous changes in global industry. This evolutionary aspect of SEZs has also been underlined by Farole and Akinci (2011), who in their recent report suggest that the proliferation of SEZs has been favored by a globalization process that started in the 1970s, accelerating during the 1990s and 2000s, coupled with an increased fragmentation of manufacturing production leading to 'global production networks.'

Clusters and SEZs: The Case of China

Special Economic Zones: From Experimentation to Growth and Beyond

China's SEZ experience represents a unique case at the world level, where SEZs have been the main tool used to support the opening of the economy. They were designed as 'laboratories' in which to experiment with accelerated economic development and the controlled import of foreign technologies and capital, with the idea of expanding the process to the rest of the country if successful.

In particular, the main aims of the establishment of SEZs in China were (Ota, 2003):

1. To observe and understand capitalism;
2. To test different policies;
3. To acquire capital, modern technologies, and management methods;
4. To accelerate exports;
5. To create local job opportunities, especially for young people;
6. To generate foreign exchange to finance China's import needs;
7. To promote competition among regions;
8. To facilitate the future return process of Hong Kong, Macao, and Taiwan into China, by favoring a graduated reduction of income gaps between these places and mainland areas.

The first step of this process has been the promulgation in 1979 of the 'Law of the People's Republic of China on Chinese-Foreign Joint Ventures,' whose main objective was the attraction of foreign capital to support growth, while keeping control of the process. In the following year, three SEZs were established in Guangdong Province (in Shenzhen, Zhuhai, and Shantou) and one in Fujian Province (Xiamen), all close to strategic sources of foreign capital, Hong Kong (Shenzhen) and Macao (Zhuhai), and facing Taiwan (Shantou and Xiamen) (Ota, 2003). The central government allowed SEZs to make autonomous decisions with regard to relevant aspects for their development (excluding railways, postal and telecommunication services, banking, civil aviation, and national defense). They were encouraged to experiment with policy innovation, being certain that successful initiatives would have been fully recognized. To increase their autonomy and minimize the risks of political interference from the central government, they were purposefully located far from Beijing.

Foreign investors and other economic actors located in these areas benefited from this diverse institutional framework (Di Tommaso et al., 2012). Furthermore, they could count on the availability of a highly skilled labor force, a battalion of low-cost migrant workers coming from the rural areas of the province, and no risks of interference from trade unions.

Probably also because it was the first to be established, Shenzhen is surely the most emblematic case. The central government granted to the local administration a 10-year exemption from the submission of tax revenues to the central and provincial governments and, since 1981, the same political status as the capital. In 1992, the central government also granted it legislative power.

This special status allowed many revolutionary changes in regulation and administration to be tested and tried out. Among these were the selling of state land-use rights, the establishment of personnel system reforms, the abrogation of voucher quotas for food purchases in 1984 (10 years before the rest of China), the opening to foreign banks in 1982, the state-owned enterprises (SOEs) reform in 1986, and the establishment of a stock exchange in 1990 (for a comprehensive overview of the implemented reforms, see Di Tommaso et al., 2012).

This framework has allowed a very rapid growth of the city. In 1979 (i.e., only a year before being transformed into an SEZ), Shenzhen was an agriculture-based village, with an irrelevant contribution to China's GDP (less than 1%) and a per capita GDP below US$90 per year. There were only 20 000 residents with low education levels and poor living standards. The manufacturing sector was practically nonexistent. The only advantage the area had was its proximity to Hong Kong, Macao, and Taiwan and its low resistance to any kind of structural change because of its simple social structure and the absence of internal conflicts. Shenzhen over time demonstrated itself to be one of the fastest growing cities in China, with a yearly growth rate of 15% in 2008. Thirty years after the establishment of the SEZs, in 2010, *Forbes* magazine included Shenzhen among the top five most-crowded cities in the world.

Shenzhen now ranks second after Guangzhou in terms of per capita GDP: in 2009, it amounted to RMB 92 772 (approx. US$13 000). Its industrial output is the highest among all provincial cities (RMB 1541 billion – approx. US$217 billion), and industry accounts for a growing share of GDP formation and GDP growth. Shenzhen sees a huge concentration of activities on high-tech sectors, particularly electronics, which in 2008 accounted for 91% of the total output value of high-tech industries. This can be partly read as a consequence of the fact that at the initial phase of the opening processes, wholly foreign-owned enterprises were authorized only if they exported at least 50% of their output or produced high-tech goods. Therefore, several multinationals operating in technology-intensive sectors have established branches in the prefecture, such as IBM, Intel, Toshiba, Epson, Xerox, and Hitachi (Di Tommaso et al., 2012).

Great efforts have been made toward the reinforcement of the political independence of the area, while at the same time reducing the role played by the government in microeconomics. This has been done by means of the creation, in 2004, of government departments that are independent from national ministries.

As a consequence of the success of the first SEZs, SEZ status was extended to 14 urban coastal centers and to Hainan Island. Guangzhou was declared an 'open coastal city,' and in 1985 the Pearl River Delta was officially designated an 'open economic

region' (along with the Yangtze River Delta and the Golden Triangle Delta) with the aims of obtaining accelerated economic development and of knowing in advance what was happening on the international scene (Barbieri et al., 2012). In 1988, the central government authorized the establishment of two other SEZs – in the Pudong district (Shanghai) and in Hainan, and the latter was at the same time also separated from Guangdong and declared an autonomous province. Accelerated growth has, of course, created several problems that are now visible in Shenzhen. Congestion, traffic, pollution, difficulties with providing services to everyone in an overcrowded area, and so on are all issues with which policy makers are being challenged.

More recently, SEZs have been used by Chinese policy makers to help rebalance the development of the country. While the tool of 'special zones' remains relevant, it is employed for new targets and long-term objectives that place renewed emphasis on the cluster promotion aspects of policy.

An interesting example in this scenario is the Double Relocation Policy, launched in 2008 by Guangdong. The aim was to relocate the labor-intensive, resource-consuming manufacturing industries from the central area to less developed areas and the workforce from the agricultural sector to the secondary and tertiary sectors.

In order for the relocation to take place, the plan foresaw the ad hoc creation of 'industry-transfer demonstration parks,' aimed at hosting the sectors moving from their original locations. The main requirement is that all industries accepted by these parks be developed into one or more clusters, with one dominant sector and several supporting sectors (Wang and Yue, 2010, p. 194). By January 2011, the parks alone have absorbed about 426 000 people, more than 60% of which are coming from Guangdong. Firms located in the new industrial parks are encouraged to employ local people and the government will also provide free polytechnic education for the most rural poor (Di Tommaso et al., 2012).

Clusters: The Case of Specialized Towns in Guangdong Province

In the study of supra-firm structures, the Chinese case is particularly interesting not only for the peculiar use of SEZs that has just been described, but also because they have been used in tight connection with other policy instruments, such as in the case of Guangdong's STs (Di Tommaso et al., 2009).

Guangdong Province is among the most developed parts of China. In this area, the government played a strong role in industrial development, with great attention placed on how to influence the location of firms in the province. The aims were to manage the incredibly fast – and often disordered – development of some of its territories (with double-digit yearly growth figures since the beginning of the 1980s) while encouraging the development of the other, more backward areas. In particular, Guangdong government planners thought that favoring the agglomeration of specialized firms in a few specific locations could help attain the two aforesaid goals and more generally encourage industrial development, competitiveness, innovation, and technological advancement (Di Tommaso et al., 2012).

It has to be said that, similar to what happened in the rest of the world, in Guangdong Province the formation of several agglomerations of (similar or complementary) firms has been a spontaneous process. What makes Guangdong's (and China's) experience unique is the decision of policy makers to accelerate and guide this process by means of specific policy measures.

The origin of such measures dates back to the second half of the 1980s, when the Chinese government introduced the Spark Programme (Department of Science and Technology of Guangdong Government (DSTGG), 2003; Barbieri et al., 2010). The aim was to raise the technological innovation capacity in several rural areas of the country, particularly by establishing investment areas known as 'spark technology investment zones,' where firms could utilize subsidized funds coming from government and banks. These investment zones represent the seeds from which several Guangdong STs originated. In many cases, in fact, each area tended to specialize in a specific sector (one city, one product). Given that the town was the prevailing administrative unit among these agglomerations, they were called 'specialized towns' (Di Tommaso and Rubini, 2006; Di Tommaso et al., 2012).

This was the scenario when, in 1998, the Department of Science and Technology of the Guangdong government decided to launch the 'one city, one product' program to support STs. It officially acknowledges as STs those 'townships' (or, less frequently, 'counties' or 'urban districts') with a yearly industrial output beyond RMB 2 billion (approx. US$316 million, at the September 2012 exchange rate), with at least 30% of their industrial output (or employment) concentrated in one industry that is called the 'specialized sector.'

The officially recognized STs are then eligible to receive a subsidy from the provincial government, with additional cofinancing from the local government. These funds have to be mainly used to establish an innovation platform or center, driving firms to upgrade their production capabilities and develop new technologies. The scope of the program is to not just target the single firm but also enhance the reputation of the whole town, possibly supporting the development of a common and easily recognizable brand (DSTGG, 2003).

Since the launch of the program, the number of acknowledged STs has been constantly increasing: there were 71 in 2003, 277 in 2008, and 326 in 2012. Also, their weight in the provincial economy has been growing: in 2001, they represented 3.66% of the total provincial GDP, while in 2010 they accounted for 28.5%. The growth is equally striking in terms of the number of enterprises (from 14 800 in 2001 to more than 950 000 in 2010) and industrial output of the specialized sector (from RMB 42 billion in 2001 to more than RMB 132 billion 9 years later) (Di Tommaso et al., 2012).

Analyzing the evolution of the acknowledged STs over time, it is possible to identify specific trends. In particular, in the first years, the aim was mainly to strengthen the central-eastern area of the province (the so-called Pearl River Delta area). In these last few years, however, in line with the new targets of the government, the recognition of STs seems to have tried to rebalance the territorial development of the province (with more towns acknowledged outside the Pearl River Delta) while at the same time incentivizing specific sectors (such as services or tourism).

In this way, the promotion of clusters in Guangdong has become an explicit industrial policy model: a tool used by the

government to address more general industrial development aims (Barbieri et al., 2012).

Final Remarks

SEZs in their different forms (industrial development parks, high-tech development zones, EPZs, and the like) and industrial clusters share some important features. They are, first of all, 'special territories' geographically based in specific areas of a country. Second, they are places where the agglomeration of different actors (firms as well as institutions, research centers, universities, and so on) takes place, which is often actively encouraged by governments. Finally, they are locations where such agglomeration can give rise to clustering advantages, and this is the main reason why such special territories can be seen and have been used as policy instruments for industrial development.

In the case of China (and of Guangdong Province, in particular), these tools have been extensively utilized and coordinated by local, provincial, and national governments. The Chinese experience of developing and managing STs and SEZs shows some specific features that can stimulate the debate on the use of such instruments for industrial development:

1. Specialization (in the same sectors or in closely related ones). The continuous intertwining between SEZs and agglomerations of specialized activities is a distinctive feature in China. For example, many of the first Chinese STs originated from industrial parks. Again, Shenzhen, the first and most famous Chinese SEZ, shows a strong specialization in the electronic sector, despite not hosting any formally acknowledged ST.
2. Geography of production. The choice of the location of SEZs and industrial parks and the dynamic of official recognition of STs suggest that the identification of 'special territories' is guided by strategic objectives where the spatial and geographical dimensions of industrial development play a central role.
3. Adaptive capacity to changing conditions. The experiences of both SEZs and clusters in China show that they have been interpreted as tools of a long-term industrial development strategy where the aims and targets are being adjusted according to the changing conditions. STs are today increasingly encouraged in remote areas in order to stimulate territorial rebalancing, SEZs acquire more and more the form of relocation parks, and the first SEZs now concentrate on the development of high-value-added goods and services.

The extent to which each one of these instruments has been effective in favoring growth and industrial development has to be verified. What is certain is that SEZs and STs have played a central role in the industrial development strategy of China country.

Bibliography

Barbieri, E., Di Tommaso, M.R., Huang, M., 2010. Industrial policy and firm strategy in Southern China: government targets and firms' behaviour. European Planning Studies 18 (1), 83–105.

Barbieri, E., Di Tommaso, M.R., Bonnini, S., 2012. Industrial development policies and performances in Southern China: beyond the specialised industrial cluster program. China Economic Review 23 (3), 613–625.

Becattini, G., 1989. Sectors and/or districts: some remarks on the conceptual foundation of industrial economics. In: Goodman, E., Bamford, J., Saynor, P. (Eds.), Small Firms and Industrial Districts in Italy. Routledge, London.

Coase, R.H., November 1937. The nature of the firm. Economica, 386–405.

Department of Science and Technology of Guangdong Government (DSTGG), 2003. The Innovation of Industrial Clusters in Guangdong Province, Guangdong. Guangdong Provincial Department of Science and Technology (in Chinese).

Di Tommaso, M.R., Rubini, L., 2006. Cluster industriali e specialized towns nel Guangdong: la centralitá del Delta del Fiume delle Perle (Industrial clusters and specialized towns in Guangdong: the central role of the Pearl River Delta). In: Di Tommaso, M.R., Bellandi, M. (Eds.), Il Fiume delle Perle: La Dimensione Locale dello Sviluppo Industriale Cinese e il Confronto con l'Italia, The Pearl River Delta: the local dimension of the Chinese industrial development and the comparison with Italy. Rosenberg & Sellier, Turin.

Di Tommaso, M.R., Barbieri, E., Rubini, L., 2009. Industrial development policies in Southern China: the specialized towns programme. Economia e Politica Industriale, 179–198.

Di Tommaso, M.R., Rubini, L., Barbieri, E., 2012. Southern China: Industry, Development and Industrial Policy. Routledge, Abingdon.

Farole, T., 2011. Special Economic Zones in Africa: Comparing Performance and Learning from Global Experiences. World Bank, Washington, DC.

Farole, T., Akinci, G. (Eds.), 2011. Special Economic Zones. Progress, Emerging Challenges, and Future Directions. World Bank, Washington, DC.

Johansson, H., Nilsson, L., 1997. Export processing zones as catalysts. World Development 25 (12), 2115–2128.

Marshall, A., 1890/1920/1972. Principles of Economics, eighth ed. Macmillan, London. (Principi di Economia, a cura di A. Campolongo, UTET, Torino (traduzione dalla VIII edizione del 1920).

Mathews, J.A., 2012. Strategizing in industrial clusters and suprafirm structures: collective efficiency, increasing returns and higher order capabilities, 2010 Crafoord Lecture. In: Kalling, T. (Ed.), Strategy and Entrepreneurship. Institute of Economic Research, Lund Business Press, University of Lund.

Ota, T., 2003. The role of special economic zones in China's economic development as compared with Asian export processing zones: 1979–1995, Asia in Extenso, March, pp. 1–28.

Wang, J., Yue, F., 2010. Cluster development and the role of government: the case of Xiqiao textile cluster in Guangdong. In: Zeng, D.Z. (Ed.), Building Engines for Growth and Competitiveness in China: Experiences with Special Economic Zones and Industrial Clusters. The World Bank, Washington.

Warr, P., 1990. Export processing zones. In: Milner, C. (Ed.), Export Promotion Strategies. Wheatsheaf, New York.

Waters, R., Lawton-Smith, H., 2008. Social networks in high-technology local economies: the cases of Oxfordshire and Cambridgeshire. European Urban and Regional Studies 15 (1), 21–37.

Williamson, O.E., 1975. Market and Hierarchies. Free Press, New York.

World Bank (WB), 2008. Special Economic Zones: Performance, Lessons Learned, and Implications for Zone Development. World Bank, Washington, DC.

Special Education

John D Molteni, Jule McCombes-Tolis, and Melissa Root, University of Saint Joseph, West Hartford, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Special education as a discipline is a broad field whose focus is the education of students with exceptionalities that includes support for career and college readiness. International practices in educating students with disabilities are emerging in many countries with varied levels of inclusive practice and educational benefit. Increasingly, there is a climate of school accountability, for example, in the United States with Common Core State Standards, and outcome assessments have placed pressure on schools and school districts to evaluate the quality of their teaching by way of student outcomes. The role of the special educator within the school is one that has various permutations from a primary interventionist who delivers instruction to a collaborator/co-teacher within an inclusive setting. This article will review the history of special education, current trends, and future considerations for the field.

The Discipline of Special Education

Special education is a field in education/educational psychology that is focused on the effective education for children with exceptionalities. Students with exceptionalities may have learning, behavioral, emotional, or social difficulties or may be gifted and talented. Twice exceptional students are those who are both gifted in an academic area and have a disability in another area (e.g., behavior, emotional control, or socialization). Special educators are trained broadly to address the needs of these students in a variety of public, private, and nonpublic settings. Provisions for credentialing as a qualified special educator vary but can include broad training that covers all exceptionalities or may subdivide credentialing into specific practice areas. Examples of these areas include mild to moderate and moderate to severe disability, educators of the deaf and hard of hearing, gifted and talented or emotional and behavioral disorders. Additional subdivisions may occur within age ranges (e.g., early childhood, elementary, and secondary aged students).

The Council for Exceptional Children (CEC; www.cec.sped.org) serves as the international professional organization representing the interests of students with exceptionalities by addressing the needs of government policy, standards for training, and professional development. CEC provides standards for knowledge, skills, and ethical principles and practice standards for the initial and advanced preparation of special educators as well as specialty standards for informing preparation programs. Accreditation by the Council for the Accreditation of Educator Programs (caepnet.org) of teacher licensure programs bases their accreditation decisions on CEC standards. The CEC's Division of International Special Education and Services provides support for the dissemination of practices internationally and communication in international forums and professional development.

International organizations advocate for educational practices that improve outcomes for individuals with disabilities as well as broader education agendas for countries to support economic and social progress in countries. The United Nations, United Nations Education, Science and Cultural Organization (UNESCO), and United Nations International Children's Emergency Fund (UNICEF), among others, have put forth positions and policies regarding the rights of children with disabilities to educational programs and monitor these practices across nations. The United Nations (2006) outlined their position on inclusive environments for individuals with disabilities in the Convention on the Rights of Persons with Disabilities. With regard to education, Article 24 outlines the need to reduce discriminatory practices with regard to education, the inherent value and potential of individuals with disabilities, and inclusion of individuals with disabilities in the general education environment with reasonable educational accommodations. Additionally, the convention outlined the need to educate individuals with disabilities on life and social skills needed to interact with the larger community. UNESCO, in partnership with UNICEF, monitors and supports the efforts for inclusive educational practices outlined in Article 24.

UNESCO provides information regarding policies and practices that are necessary to support individuals with disabilities in inclusive programs. Through their database, Inclusive Education in Action (IEA, 2011) provides examples from around the world on inclusive practices and training programs for teachers who serve as educators for these children. Policy guidelines are also provided for member states to utilize in developing their approaches to inclusive education. Consideration of barriers to inclusive education including attitudes toward individuals with disabilities, cultural influences, and teacher training needs are central components to developing an inclusive educational program, the emphasis being that educating students with exceptionalities is a benefit to society socially and economically.

Flexible learning environments, child-friendly environments, and a developmentally appropriate curriculum that includes content beyond academic skills are a common thread in current special education practices. Additionally, evidence-based practices, those practices that have support in the scientific literature, are increasingly important to provide these core areas of an inclusive environment. This requires additional training and professional development in both preservice and in-service training.

The Role of the Special Educator

Special educators are tasked with providing instruction that produces optimal outcomes for students with exceptionalities across academic, social, behavioral, and adaptive skills. In many countries, they are required, by law, to develop educational programs that are individualized for students with exceptionalities that lead to educational benefit that is objectively measurable. Special educators, in addition to providing or supporting instruction of students with exceptionalities, collaborate with other members of the educational team. The members of this team include the parents/caregivers of the student, related services personnel (e.g., speech and language pathologists, classroom teachers, and occupational therapists), administrators, and others as determined by the needs of the student and the wishes of the parent/caregiver. Instruction may occur in the general education classroom, a resource room where the student goes to receive primary or supplemental instruction or a significantly separate setting where a student may spend the majority of the day. Wherever instruction is delivered, evidence-based instructional strategies, informed by formative and summative assessments, serve as the foundation for the special educator to meet the educational needs of their students.

In other countries, with the emergence of educating students with exceptionalities, the role of the educator is more broadly defined and training may be less specialized than in those with more established programming. There are emerging educational programs housed in universities that train preservice teachers in teaching methodology and inclusive environments. The approach to education varies by region and cultural influences and differs from traditional western approaches of education. This requires the training for teachers across settings to differ. Similarities in outcomes, that is, independent community involvement, are evident.

Training in Special Education

Models of training that exist to train special educators broadly fall under two categories, university-based programs and alternative certification (AC) programs. Both exist to train preservice teachers. Within the United States, shortages in special educators over the past two decades (Boe, 2006) have led to the development of AC programs. Questions have been raised about the quality of these forms of training programs. Quigney (2010) summarized research on AC programs and reported that traditional programs generally produce better trained educators. She also outlined a format for discussing the benefits and concerns regarding AC programs since they were responding to the shortages noted in special education. Teachers prepared within AC programs reported challenges with classroom management and content knowledge (Casey et al., 2011). The evidence is not definitive as authors have reported higher rates of employment for teachers trained through the AC programs although their tenure of employment was shorter (Robertson and Singleton, 2010).

Internationally, particularly in those with emerging special education programs, the shortages are due to several factors, including overcoming attitudes toward students with exceptionalities within schools and the broader community and a lack of expertise in these emerging programs. International experts may be brought in to support these efforts and support program development both within schools and training programs through universities. Systems change, particularly at a national level in many cases, is challenging and requires the support of the national government, particularly ministries of education, to move these programs forward.

Learning Theories Applied

Educators' beliefs concerning how it is that students learn impact, in part, how they design and deliver instruction. These beliefs may reflect one or more learning theories, which are theories that help to explain how individuals come to acquire, access, and apply knowledge and skills. While numerous learning theories have been proposed over the past century, three core frameworks have emerged as paramount: behaviorist learning theory frameworks, cognitivist learning theory frameworks, and constructivist learning theory frameworks.

Behaviorist learning theory frameworks are concerned with observable indicators of learning and the conditioning of these observable behaviors. An educator who ascribes to a behaviorist learning theory framework would regard a student's ability to provide a target response when presented with a given stimulus (e.g., problem type) as evidence of learning. S/he would likely utilize instructional cues when teaching, provide reinforcement to help shape students' behavior (e.g., utilize a points or reward system in instructional contexts), provide students with ample opportunities to practice the target behavior, and structure learning tasks to ensure that students master basic skills before moving onto more complex, integrated learning tasks. Curricula reflecting a behaviorist learning theory framework utilize a scope and sequence of instruction and emphasize mastery learning for all students.

While behaviorist learning theory frameworks focus on observable behaviors as evidence of learning, cognitivist learning theory frameworks emphasize the role that cognition plays in learning and behavior. Fundamental to cognitivist learning theory frameworks is the belief that learning processes (e.g., memory, attention, perception) impact observable behaviors and these processes can be improved and/or supported in order to maximize student learning and performance. An educator who ascribes to a cognitivist learning theory framework would support students in developing background knowledge and schemata and would make the relationships among information overt. In the classroom, s/he would likely utilize graphic organizers, think-alouds, and performance assessments to scaffold and assess the learning process. Curricula reflecting a cognitivist learning theory framework feature well-organized and clearly structured content, concern for how perceptual features of materials are organized and presented, and a tandem emphasis on strategy instruction and acquisition.

Constructivist learning theory emphasizes the role that the learner, including his/her prior experiences and frameworks, plays in constructing new knowledge. An educator who ascribes to a constructivist learning theory framework would regard knowledge as created rather than acquired, and would

assume the role of guide and facilitator in the classroom. S/he would likely utilize collaborative, problem-based learning opportunities rather than predetermined instructional sequences, and would emphasize reflective practice. Curricula reflecting a traditional constructivist learning theory framework would emphasize 'discovery learning' driven by students' intrinsic motivation to explore and experiment, with minimal teacher guidance.

In the field of special education, research primarily reflecting behaviorist and cognitivist learning theories has informed the development of specialized academic and behavioral intervention curricula, and the individualization of educational programs for students with disabilities. Terms and phrases such as 'direct,' 'explicit,' 'systematic,' 'multisensory,' 'strategy-based,' and 'data-driven' often signal that a curriculum is grounded in such theory. Constructivist learning theories may be reflected in more advanced, applied learning contexts, but even so, are most often coupled with instructional approaches that are informed by behaviorist and cognitivist underpinnings.

Evidence-Based Practice

The term 'evidence-based practice' has been used in the field of medicine to refer to treatment practices that consistently yielded positive outcomes when delivered under experimentally controlled conditions (Odom et al., 2005). Three decades later, the No Child Left Behind Act (NCLB) of 2001 institutionalized the concept of utilizing scientific research (e.g., experimental, quasiexperimental, single-subject design; replicable findings; published in peer-reviewed venues) to inform educational practice. NCLB's definition of scientifically based research is likewise reflected in the 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA), thereby affirming that scientific evidence should inform educational practice decisions in both general and special education contexts.

Defining 'evidence-based practice' for special education populations and settings has proven to be particularly challenging because of the complexity of the discipline, including the heterogeneity of disability presentations. This has often precluded researchers implementing true experimental inquiries (e.g., the random assignment of subjects to controlled and experimental conditions). Fortunately, Gersten et al. (2005) proposed a set of quality indicators for group experimental and quasiexperimental research in special education, and provided guidelines for how consumers could utilize these indicators to determine if a practice is evidence based. These quality indicators and guidelines continue to inform best practice in the field, and are referenced extensively in a CEC Division for Research White Paper titled *Thinking and Communicating Clearly About Evidence-based Practices in Special Education* (Cook and Cook, 2011).

Practitioners charged with selecting methods and materials for use with special student populations are also encouraged to consult the What Works Clearinghouse (WWC), a division of the United States Department of Education's Institute for Educational Sciences (IES). The WWC assesses the quality and findings of existing educational research and publishes their findings in the form of intervention reports. Here, it is possible to search for interventions by subject (e.g., reading, mathematics, and writing) and by name. While interventions included here may indeed have the backing of published research, it may be the case that the studies do not meet the evidence standards articulated by the WWC and, as such, no conclusion could be drawn about the efficacy of the intervention from an evidence-based standpoint.

This distinction highlights the difference between 'evidence-based practice' and 'research-based practice': 'research-based' programs often do not have the backing of published research that reflects the rigorous standards associated with the definition of scientific research proposed by NCLB, IDEA, and IES. This is the case, for example, when evaluating the efficacy of the Wilson Reading System® for students with learning disabilities. Of the 28 studies of the Wilson Reading System® employed with students with learning disabilities published between 1989 and 2009, none met the review protocol and evidence standards of the WWC. As such, given the standards of the IES, it could be argued that the Wilson Reading System® is best described as a 'research-based' versus 'evidence-based' intervention for students with learning disabilities.

CEC's Division for Research makes clear, however, that methods and materials that do not presently meet rigorous evidence-based practice standards for a given population of students should not summarily be disregarded. Rather, special educators are advised to utilize professional judgment when selecting interventions and to employ informed progress monitoring practices (see below) to monitor the efficacy of interventions provided.

Special Education Frameworks

All learners can benefit from different aspects of special education services, but most do not reach the level of legally requiring such services for success in school. The current principles of special education rely on the concept that delivering many of these services through the general education environment will help support more learners, regardless of their level of need. These principles are oftentimes delivered through the frameworks of Positive Behavioral Interventions and Supports (PBIS) and Response to Intervention (RtI). The goal of these frameworks is to allow all students the optimal opportunity to learn in the general education environment by using special education strategies to support the learning of all students. When more intensive strategies are required, these frameworks allow use of more directed interventions to assess whether a child with inimitable needs is responsive to such interventions. These systems are designed to elevate the number of students receiving effective interventions in the general education environment while lowering the number of students being referred to special education and thus removed from the general education environment.

Both PBIS and RtI are tiered frameworks that are designed to use differentiated instruction as soon as a problem is noted. PBIS uses this approach to address behavioral concerns while RtI is principally engaged in academic remediation. Both

systems are intended to address problems in a timely and effective manner by using quick response time and effective practices.

While PBIS was initially established as a framework in the United States, it is being implemented in over 15 countries worldwide. Several countries have national initiatives supporting the implementation. Countries implementing PBIS (or SW-PBS) include New Zealand, Australia, Canada, Netherlands, Sweden, Norway, Finland, Denmark, Portugal, Turkey, Spain, Italy, France, Germany, and Hungary (T.J. Lewis, personal communication, 5 February 2014; G. Sugai, personal communication, 4 February 2014).

Positive Behavioral Interventions and Supports

PBIS (also referred to as School-wide Positive Behavior Supports) is considered a framework upon which each individual school can scaffold its expectations for community behavior and respect. It is a primary prevention system designed for all students with additional supports provided for some students. Since PBIS is a framework and not a predetermined curriculum, the expected behaviors are developed from within the school community and are unique. PBIS is currently implemented in over 18 000 schools (www.pbis.org).

PBIS consists of three tiers of intensity. The first tier is implemented school-wide and consists of rules and expectations of behavior, typically both physical and social behavior. Rules and expectations are explicitly taught to all community members, with positive reinforcement systems ingrained in the program along with a continuum of consequences for problem behaviors. Inherent to this framework is the expectation that data be kept on multiple variables and assessed periodically for effectiveness.

Those students who do not respond well to the first tier of PBIS, based on data, begin to receive services through a second tier. This tier typically consists of about 10–15% of the school population. This tier is not implemented school-wide but is focused on smaller groups of students with similar behavioral concerns. These students are considered at-risk for serious behavior problems and the purpose is to prevent such development. Interventions in the second tier are often focused on student needs relating to peer relationships, low academic achievement, and challenging home environments (Ennis et al., 2011). These interventions are research based with data kept on their effectiveness.

The third tier is comprised of students with serious or chronic problem behaviors, which is typically about 5% of the school population. Students in tier 3 receive intensive, individualized research-based interventions to remediate their problem behaviors. Again, data are kept and analyzed to determine the effectiveness of each intervention implemented. Students who do not show positive response to tier 3 interventions are then referred for special education consideration.

Based on a literature review searching for evidence-based effective research practices, Simonsen et al. (2008) summarized the best classroom management practices as the following: "(a) maximize structure; (b) post-teach, review, monitor, and reinforce expectations; (c) actively engage students in observable ways; (d) use a continuum of strategies for responding to appropriate behaviors; and (e) use a continuum of strategies to respond to inappropriate behaviors" (Simonsen et al., 2008: 353). Within those overarching themes are effective tier 1, 2, and 3 practices such as physical classroom layout recommendations, rates of opportunities to respond, response cards, direct instruction techniques, guided notes, group contingencies, specific and contingent praise, etc. An outstanding source for educators on effective interventions is found at the WWC website (http://ies.ed.gov/ncee/wwc/).

Response to Intervention

RtI is a tiered framework that mirrors PBIS but through addressing academic problems in addition to behavior problems. RtI is a legal requirement in the United States under the latest iteration of IDEA and is thus expected to be implemented throughout that country's public schools. However, it is a visual and theoretical representation of the best practices of special education teaching that have existed worldwide for years (http://www.ldworldwide.org/pdf/rti/rti-series-1-what-is-rti.pdf). It is comprised of high-quality, differentiated instruction that is matched to a student's needs. The progress, or lack thereof, that the student makes is recorded and monitored frequently to assess effectiveness. Implementation of RtI is considered by some special education directors to be less complicated at the elementary level than at the secondary level (Sansosti et al., 2011), but is nonetheless an effective means of early intervention for students with academic needs.

An example of math interventions that can be implemented through the RtI tiered framework is given here to exemplify the system. A moderately effective tier 1 intervention includes screening students in the classroom who are at risk for math difficulties and providing interventions (Gersten et al., 2009). Some tier 2 and tier 3 interventions include using explicit and systematic instruction (providing models of useful problem solving, verbalizing effective thinking, guided practice, corrective feedback, and frequent review), use of visual representations of math concepts, and 10 min at every grade level devoted to review of basic math facts (Gersten et al., 2009). Through this tiered framework, at-risk students are identified early and appropriate interventions are administered and monitored to remediate difficulties prior to escalation of learning problems.

Universal Design for Learning

PBIS and RtI address behavior and academic problems through differentiated instruction and intervention. Another model of instruction that is currently in the field of special education similarly concentrates on success in the general education environment but does so through Universal Design for Instruction (UDI). Through this model, instruction and assessment in the classroom are designed specifically to allow optimum access to the materials for a wide variety of students, including those with physical and cognitive challenges (http://www.udlcenter.org). It is another means of supporting access to the general education curriculum for all students, through a flexible, positive framework of instruction.

UDI is an adaptation of the Universal Design concept from architecture. The concept is that the design of a building entrance, for example, be made usable by people with a variety of characteristics and needs. A traditional door at a building entrance is accessible to those with the ability to turn a knob, pull the door, step back from the door to allow it to swing, and hold it open to their side while they walk through. However, a door that senses someone's presence and automatically opens is also accessible to someone carrying a heavy box, pushing a stroller, ambulating in a wheelchair, and ambulating on crutches, as well as a small child (Burgstahler, 2009). This design is thus more universally usable, not just for those with disabilities. With UDI, instruction is made universally accessible in the same manner. In instruction, these are addressed through class climate (ensure a welcome and open environment where students can express their unique learning needs), interaction (offer interaction opportunities that are accessible to all, such as having a microphone to broadcast the speaker's voice), physical environments and products (develop a plan for safe egress for all students in an emergency), delivery methods (use multiple formats to deliver content), information resources and technology (post assignments early so students can begin projects early in the secondary school grades), feedback (provide prompt and useful feedback), assessment (use a variety of assessment types, such as short answer and multiple choice tests, papers, group and individual projects, etc.), and accommodation (be prepared to allow extra time to all test takers, if necessary, and be prepared for special accommodations for areas not met by UDI) (Burgstahler, 2009).

There are numerous research-based interventions that are effectively utilized through PBIS, RtI, and UDI. The keys to current best-practice international special education instruction are that it is based on proven interventions, data are kept and monitored to evaluate intervention effectiveness, and delivery of interventions begins immediately after problems are identified. Flexible and proactive delivery of instruction and expectations are also utilized to create environments of success for all students.

Future Directions and Issues in Special Education

Given the focus of special education and the needs for evidence-based practices for inclusion, instructional strategies and pedagogical knowledge is needed. While frameworks derived from special education research are now providing systemic approaches to addressing the needs of all students, more work is needed in developing instructional methodology that can be applied to students with varying abilities and exceptionalities that will lead to improved outcomes for all students. Continued research on universally applicable instructional methods is needed in this regard. Alternative teaching arrangements such as co-teaching arrangements that employ both a general and special education teacher in the classroom and other models will be supportive of these efforts. Increased collaboration between general and special educators in classroom settings will support these efforts (Friend et al., 2010; Solis et al., 2012).

High-quality teacher preparation for both general and special educators is a continuing effort and one that is central to developing competent, inclusive programs. Standards to general educators to be better informed on the needs of students with exceptionalities in their classrooms will support these students' participation in inclusive settings. Core competencies in classroom management, differentiated instruction, assessment, and tiered interventions for general educators will become increasingly necessary in the coming years and allow for a responsive system of instruction that will support inclusive programming. For special education preparation, efforts to identify competencies necessary to support and provide primary instruction to students are being undertaken.

Disproportional representation in special education is a concern in many countries. In established programs in special education, overrepresentation of nonnative speaking learners, students of immigrant status, and low-income students continues to be an issue in education (Algozzine, 2005) that requires redress with regard to eligibility, instruction, and educational outcomes. Efforts to involve community, families, and the school in complementary and collaborative efforts to improve outcomes including supplemental learning opportunities, family involvement, building in community supports, and outreach efforts are being evaluated for their effectiveness in bridging the achievement gap. It will be critical as inclusive school programs continue to develop internationally that these students are appropriately supported.

See also: Autism Spectrum Disorders; Education and Economic Growth; Educational Learning Theory; Learning Theories and Educational Paradigms; Reinforcement, Principle of; Teacher Education.

Bibliography

Algozzine, B., 2005. Restrictiveness and race in special education: facts that remain difficult to ignore anymore. Learning Disabilities: A Contemporary Journal 3 (1), 64–69. ERIC, EBSCO*host* (viewed 19.12.13.).

Boe, E.E., 2006. Long-term trends in the national demand, supply, and shortage of special education teachers. Journal of Special Education 40 (3), 138–150. ERIC, EBSCO*host* (viewed 19.12.13.).

Burgstahler, S., 2009. Universal design of instruction (UDI): definition, principles, guidelines and examples. DO-IT, University of Washington, Seattle.

Casey, P., Dunlap, K., Brister, H., Davidson, M., 2011. I only wish i'd known: voices of novice alternatively certified special education teachers. International Journal of Special Education 26 (1), 182–190. ERIC, EBSCO*host* (viewed 19.12.13.).

Cook, B.G., Cook, S.C., 2011. Thinking and communicating clearly about evidence-based practices in special education (White Paper). In: Council for Exceptional Children, Division for Research.

The Council for exceptional children's position on special education teacher evaluation, 2013, TEACHING Exceptional Children 45 (3), 73–76, ERIC, EBSCOhost (viewed 19.12.13.).

Elementary and Secondary Education Act, 1965 (USA).

Ennis, R., Swoszowski, N., Cain, N., 2011. The top 10 things to consider when implementing secondary-tier PBIS interventions. Beyond Behavior 20 (1), 42–44.

Ertmer, P.A., Newby, T.J., 2013. Behaviorism, cognitivism, constructivism: comparing critical features from an instructional design perspective. Performance Improvement Quarterly 6 (4), 50–72.

Friend, M., Cook, L., Hurley-Chamberlain, D., Shamberger, C., 2010. Co-teaching: an illustration of the complexity of collaboration in special education. Journal of Educational & Psychological Consultation 20 (1), 9–27. ERIC, EBSCO*host* (viewed 19.12.13.).

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J., 2009. Assisting Students Struggling with Mathematics: Response to Intervention (Rtl) for Elementary and Middle Schools. Available at: http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=2.

Gersten, R., Fuchs, L.S., Compton, D., Coyne, M., Greenwood, C., Innocenti, M.S., 2005. Quality indicators for group experimental and quasi-experimental research in special education. Exceptional Children 71, 149–164.
Individuals with Disabilities Education Act, 1975 (USA).
Kruse, J., 2009. Learning theories: pillars of teacher decision making. Iowa Science Teachers Journal 36 (2), 2–7.
National Center on Universal Design for Learning, 2012. What Is UDL?. Available at: http://www.udlcenter.org/aboutudl/whatisudl (accessed 19.12.13.).
No Child Left behind Act of, 2001 (USA).
Odom, S.L., Brantlinger, E., Gersten, R., Horner, R.H., Thompson, B., Harris, K.R., 2005. Research in special education: scientific methods and evidence-based practice. The Journal of Special Education 26, 371–385.
Office of Special Education Programs, 2013. Technical Assistance Center on Positive Behavioral Interventions and Supports. Available at: www.pbis.org (accessed 19.12.13.).
Quigney, T.A., 2010. Alternative teaching certification in special education: rationale, concerns, and recommendations. Issues In Teacher Education 19 (1), 41–58. ERIC, EBSCOhost (viewed 19.12.13.).
Robertson, J., Singleton, J., 2010. Comparison of traditional versus alternative preparation of special education teachers. Teacher Education and Special Education 33 (3), 213–224. ERIC, EBSCOhost (viewed 19.12.13.).
Sansosti, F., Goss, S., Noltemeyer, A., 2011. Perspectives of special education directors on response to intervention in secondary schools. Contemporary School Psychology 15, 9–20.
Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., Sugai, G., 2008. Evidence-based practices in classroom management: considerations for research to practice. Education and Treatment of Children 31, 351–380.
Solis, M., Vaughn, S., Swanson, E., Mcculley, L., 2012. Collaborative models of instruction: the empirical foundations of inclusion and co-teaching', psychology. The Schools 49 (5), 498–510. ERIC, EBSCOhost (viewed 19.12.13.).
United Nations, 2006. Convention on the Rights of Persons with Disabilities. http://www.un.org/disabilities/convention/conventionfull.shtml.
United Nations Education, Science and Cultural Organization, 2009. Policy Guidelines on Inclusion in Education. Retrieved from: http://www.inclusive-education-in-action.org/iea/dokumente/upload/72074_177849e.pdf.

Relevant Websites

Special Education
www.cec.sped.org – Council for Exceptional Children.
http://caepnet.org/ – Council for the Accreditation of Educator Preparation.
http://www2.ed.gov/about/offices/list/osers/osep/index.html – Office of Special Education Programs.
www.wrightslaw.com – Wrightslaw.

Response to Intervention
http://www.rti4success.org/ – National Center on Response to Intervention.
http://http://www.studentprogress.org – National Center on Student Progress Monitoring.
www.udlcenter.org – National Center on Universal Design.
www.pbis.org – Technical Assistance Center on Positive Behavioral Interventions and Supports.

Scientifically Based Research
http://www.centerforcsri.org/pubs/pg/index.htm – National Center for Comprehensive School Reform.
http://www.ed.gov/rschstat/research/pubs/rigorousvid/index.html – Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.
http://www.rti4success.org/index.php?option=com_content&task=view&id=1587 – Selecting Evidence Based Tools and Programs for Implementing Response to Intervention (webinar).
http://www.nifl.gov/partnershipforreading/publications/pdf/Standovich_Color.pdf – Using Research and Reason in Education: How Teachers Can Use Scientifically Based Research to Make Curricular & Instructional Decisions.
http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf – What Works Clearinghouse: Evidence Standards for Reviewing Studies.

Special Education in the United States: Legal History

Mitchell L Yell, University of South Carolina, Columbia, SC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.D. Smith, volume 22, pp. 14847–14851, © 2001, Elsevier Ltd.

Abstract

The development of special education in the United States has largely been led by developments that have occurred through federal legislation and litigation. The purpose of this article is to trace and briefly describe the legal history of special education. I trace the eight critical legal developments that have affected, and continue to greatly influence, the education of students with disabilities in public schools. I conclude by highlighting the many improvements in the lives of students with disabilities engendered by special education law.

The rights of children and youth with disabilities in the United States to receive special education services were gained largely through the tireless efforts of parents and advocacy groups in the courts and legislatures of this country. This legal history of special education can be characterized as movement from the exclusion of students with disabilities in public education to inclusion. In this article, I will trace the eight most critical historical events that have influenced the development of special education in the United States.

Critical Event #1: The Enactment of Compulsory Education Laws in the United States

Massachusetts was the first state to pass a compulsory education law in 1852 and Vermont passed the second in 1867. By 1918, compulsory education laws were in place in all states. Unfortunately, despite the enactment of these compulsory education laws, children with disabilities were often excluded from public schools. For example, in 1934, the Cuyahoga County Court of Appeals in Ohio ruled that the state statute mandating compulsory attendance for children ages 6 through 18 years gave the State Department of Education the authority to exclude certain students. The Ohio statute was an example of laws that mandated compulsory education, supposedly for all students, while simultaneously allowing school districts and states to exclude students with disabilities from public schools. It was widely believed that students with disabilities would not benefit from an education or that their presence may be disruptive to other students. In fact, as late as 1958, the Illinois Supreme Court ruled that compulsory attendance laws did not apply to children with 'mental impairments' and in 1969, it was a crime in North Carolina for parents to attempt to enroll their child with disabilities in public schools after the child had been excluded.

Critical Event #2: Parents Advocating for Their Children with Disabilities and the Formation of National Advocacy Groups

Because their children with disabilities were often excluded from public school, parents became advocates for ensuring the educational rights of their children. This advocacy paved the way for a national reexamination of the exclusion of students with disabilities from public schools. One of the first events addressing the exclusion of students with disabilities from public schools was the White House conference of 1910. A primary goal of this conference was to define and establish remedial programs for children with disabilities and special needs. This conference led to an increased interest in educating children with disabilities in public school settings rather than institutionalizing them.

In the years following the White House conference, the numbers of segregated classes and support services for students with disabilities in public schools increased significantly. Unfortunately, despite the increase in special education classrooms, many children and youth with disabilities were not identified for services or struggled in classrooms that were not appropriate for their needs. Moreover, in the 1930s, there was a decline in support for special education for students with disabilities and a decrease in the number of special education programs. This may have been due to the economic difficulties of these times. Additionally, the special education programs that did exist in public schools tended to be more restrictive and of a custodial, rather than educational, nature.

In response to the decreasing quality and numbers of special education programs in public schools, parents and educators began to come together to form groups that advocated for the educational rights of students with disabilities. For example, in 1933, the first such group formed in Cuyahoga County, Ohio. The Cuyahoga County Ohio Council for the Retarded Child consisted initially of five mothers of children with mental retardation who banded together to protest the exclusion of their children from school. The efforts of these parents resulted in the establishment of a special class for their children, sponsored by the parents themselves. Similar types of local groups were established throughout the nation during the 1930s and 1940s. In the 1950s, such groups came together to form advocacy groups on the national level. The National Association for Retarded Citizens (now the Association for Retarded Citizens) was organized in Minneapolis, Minnesota, in September 1950. Forty-two parents and concerned individuals from 13 local and state organizations met to establish what has become a powerful and significant organization of parents, families, and other persons with an interest in

improving services for persons with mental retardation. Other groups of parents and educators that became nationally prominent professional organizations include the Council for Exceptional Children, the Association for Persons with Severe Handicaps, the Autism Society of America, the United Cerebral Palsy Association, the National Association for Down Syndrome, and the Association of Children with Learning Disabilities. The advocacy movement was critical to the development of special education services. The activities of interest groups also provided information, stimulus, and support to the Congress when considering, developing, and acting on legislation. The formation of these national advocacy groups, therefore, was very important to the development of special education in the United States.

Critical Event #3: The US Supreme Court's Ruling in *Brown v. Board of Education*

In the 1950s and 1960s, the civil rights movement sought changes in society that would allow minorities, particularly African-Americans, equality of opportunity. These efforts led to litigation and changes in legislation in education. This legislation provided greater constitutional protection for minorities, and eventually for persons with disabilities. A landmark case, *Brown v. Board of Education* (1954), was a major victory for the civil rights movement and became an underpinning for further civil rights action. The *Brown* decision not only had a tremendous impact on societal rights for minorities, but also affected many aspects of educational law and procedure. Although it took time, the precedents set in *Brown* resulted in sweeping changes in schools' policies and approaches to students with disabilities.

State-mandated segregation of the races in the schools denied black students admission to schools attended by white students. The plaintiffs maintained that the practice of segregating schools was inherently damaging to the educational opportunities of minorities, that segregated public schools were not – and could not be made – equal, and that segregated public schools violated black students' constitutional rights under the 14th Amendment. As an extension of this argument, the Court maintained that state-required or state-sanctioned segregation solely on the basis of a person's unalterable characteristics (e.g., race) was unconstitutional. The high court also determined that segregation solely on the basis of race violated equal protection and denied children from minority backgrounds equal educational opportunity. This decision opened up a number of legal avenues for those seeking redress for students with disabilities. In *Brown*, the high court reasoned that because of the importance of education in our society, the stigmatizing effects of racial segregation, and the negative consequences of racial segregation on the education of those against whom segregation was practiced, segregated public schools denied students equal educational opportunities.

The ruling in *Brown v. Board of Education* has had a major influence on education law in the United States. The influence of this seminal decision was less direct in special education law, whereas it was, nonetheless, important in the history of special education. This is because advocates for improving educational opportunities for children and youth with disabilities saw the basic truths of *Brown v. Board* as equally applicable to the children for whom they advocated.

Critical Event #4: Federal District Court Rulings in the Educational Opportunity Cases

Sixteen years after the *Brown* decision, two seminal federal district court cases applied the concept of equal opportunity to children with disabilities. The two landmark decisions in which action was brought against state statutes and policies that excluded students with disabilities were *Pennsylvania Association for Retarded Citizens v. Pennsylvania* (1972) and *Mills v. Board of Education of the District of Columbia* (1972).

In January 1971, the Pennsylvania Association for Retarded Children (hereafter referred to as *PARC*) brought a class action suit against the Commonwealth of Pennsylvania in federal district court. The plaintiffs argued that students with mental retardation were not receiving publicly supported education because the state was delaying or ignoring its constitutional obligations to provide a publicly supported education for these students, thus violating the students' rights under both state statute and the Equal Protection clause of the 14th Amendment to the US Constitution. Witnesses for the plaintiffs established four critical points. The first was that all children with mental retardation are capable of benefiting from a program of education and training. Second, education cannot be defined as only the provision of academic experiences for children; it also includes experiences such as learning to clothe and feed oneself. A third point was that, having undertaken to provide all children in the Commonwealth of Pennsylvania with a free public education, the state could not deny students with mental retardation access to free public education and training. A final stipulation was that the earlier students with mental retardation were provided education, the greater the amount of learning that could be predicted, a point related to denying preschoolers with retardation access to preschool programs available to children without disabilities. *PARC* was resolved by consent agreement specifying that all children with mental retardation between the ages of 6 and 21 years must be provided a free public education, and that it was most desirable to educate children with mental retardation in a program most like the programs provided for their peers without disabilities. The decree, which was amended a year later, set the stage for continued developments regarding the educational rights of students with disabilities.

Soon after the *PARC* decision, a class action suit was filed in the federal district court for the District of Columbia. This suit, *Mills v. Board of Education* (1972; hereafter *Mills*), was filed against the District of Columbia's Board of Education on behalf of all out-of-school students with disabilities. The action was brought by the parents and guardians of seven children who presented a variety of disabilities, including behavior problems, hyperactivity, epilepsy, mental retardation, and physical impairments. These seven children were certified as a class, thereby representing more than 18 000 students who were denied or excluded from public education in Washington, D.C. The suit, which was based on the 14th Amendment, charged that the students were improperly excluded from school without due process of law. The court held that because

segregation in public education on the basis of race was unconstitutional, the total exclusion of students with disabilities was also unconstitutional. *Mills* resulted in a judgment against the defendant school board mandating that the board provide all children with disabilities a publicly supported education. In addition, the court ordered the district to provide due process safeguards.

The *PARC* and *Mills* decisions set precedent for similar cases to be filed across the country. In the 2 1/2 years following the *PARC* and *Mills* decisions, 46 right-to-education cases were filed on behalf of children with disabilities in 28 states. By the early 1970s most states had passed laws that required public schools to provide education to students with disabilities. The laws in the various states were quite varied, however, with some states having strong laws, some states having weak laws, and some states having no laws ensuring the educational rights of students with disabilities. Because of the unevenness of these laws, a federal standard was needed.

Critical Event #5: Early Involvement by the Federal Government in Education

The first significant federal involvement in the education of students with disabilities occurred in the late 1950s and early 1960s. Some of these early efforts included the Education of Mentally Retarded Children Act of 1958, in which Congress appropriated funds to train teachers of children with mental retardation, and the Training of Professional Personnel Act of 1959, which helped train leaders to educate children with mental retardation. In 1965, the Elementary and Secondary Education Act (ESEA) was passed and signed by President Lyndon Johnson as an important component of the war on poverty. The ESEA provided federal money to states to improve educational opportunities for disadvantaged children, including students with disabilities who attended state schools for the deaf, blind, and retarded. This law was the first time the federal government provided direct funding to the states to assist in educating certain groups of students. This law was a precursor to the laws that would provide direct assistance to states for educating students with disabilities.

In 1970, the Education of the Handicapped Act (EHA) became law. The EHA, which was to become the framework for legislation that followed, consolidated and expanded previous federal grant programs for educating students with disabilities and continued funding pilot projects at the state and local levels. The law also provided funding to institutions of higher education to develop programs to train teachers of students with disabilities and authorized funding for the development of regional resource centers to provide technical assistance to state and local school districts. The Education Amendments of 1974 authorized the creation of the National Advisory Council on Handicapped Children. The purpose of the amendments was to require that states receiving federal funding established goals for providing full educational opportunities for all students with disabilities. The law also provided funding for programs and specified due process procedures. The passage of these laws created many of the rights that would be included in later laws granting educational rights to students with disabilities. The early laws, however, had few enforcement mechanisms.

In 1973, Congress passed the Rehabilitation Act of 1973. Section 504, a short provision of this act, was the first federal civil rights law to protect the rights of persons with disabilities. According to the law, no handicapped person in the United States could, solely by reason of his or her handicap, be excluded from participation in, be denied the benefits of, or be discriminated against under any activity receiving federal financial assistance. The language of the law was very similar to the language in other federal civil rights laws (such as Title VII of the Civil Rights Act, which prohibited discrimination on the basis of race, and Title IX of the Education Amendments of 1972, which prohibited discrimination based on sex). Thus, the primary purpose of Section 504 was to prohibit discrimination against a person with a disability by any agency receiving federal funds, such as schools and universities. This includes students with disabilities attending preschool, elementary, secondary, or postsecondary education. Section 504 also required agencies covered by the law to provide assurances of compliance, to take corrective steps when violations are found, and to make individualized modifications and accommodations to provide services that are comparable to those offered to persons without disabilities.

Section 504 included a very broad definition of disability. The law protects from discrimination any person who has a physical or mental impairment that substantially limits one or more of the person's major life activities (e.g., learning, breathing, walking). Because of the expansive definitions of disability, Section 504 protects all students with disabilities, whether they receive special education services or not.

Critical Event #6: The Passage of the Education for All Handicapped Children Act

In 1975, President Ford signed the Education for All Handicapped Children Act (EAHCA) into law. The EAHCA, which was often referred to as P.L. 94-142 (the EAHCA was the 142nd Public Law passed by the 94th Congress), combined an educational bill of rights for students with disabilities with the promise of federal financial incentives to participating states. To receive funding, states had to develop policies ensuring that all qualified students with disabilities would receive a special education program meeting the requirements of the EAHCA. Funding would flow from the federal government to the state educational agencies and, finally, to local educational agencies (LEAs). To receive the funds, states had to submit plans that meet the federal requirements. Local school districts, in turn, had to have programs meeting the state requirements. The federal funding that states received would be based on an annual count of all children and youth served under the law. Thus, the EAHCA made the federal government a partner with the states in educating students with disabilities who were covered by the law.

In the EAHCA, the following categories of disability were eligible for special education services (some of these categories were added in later amendments to the EAHCA): autism, deaf-blindness, deafness, emotional disturbance, hearing impairment, intellectual disability (formerly known as mental retardation), multiple disabilities, orthopedic impairment, other health impairment, specific learning disability, speech or

language impairment, traumatic brain injury, and visual impairment (including blindness).

The purpose of the EAHCA was to ensure that all eligible students with disabilities were provided with a free appropriate public education (FAPE) that emphasized special education and related services designed to meet their unique needs and prepare them for further education, employment, and independent living. Thus, the focus of the EAHCA was to ensure that children and youth with disabilities have access to public education, and that this education be developed to meet their individual needs. To ensure that an eligible student's program of special education meets his or her unique needs, teams of persons from the student's school, which included the student's parents, would meet at least annually to collaboratively develop an individualized education program (IEP). A student's IEP is the blueprint of his or her special education program that must, according to the EAHCA, ensure that the student receives FAPE. Table 1 depicts the major principles of the EAHCA.

The EAHCA also included a system of procedural safeguards, which are legal checks and balances that protect the educational rights of children with disabilities and their parents. The procedural safeguards give a student's parents the right to contest in a due process hearing any aspect of their child's identification, evaluation, or program and even to file a suit in state or federal court to receive services. In fact, special education was soon to become a highly litigated area.

Table 1 The major principles of the EAHCA

Principle	*Description*
● Zero reject	● Local education agencies (LEAs) are required to locate, identify, and provide services to all eligible students with disabilities.
● Protection in evaluation	● LEAs must conduct a full, fair, and individualized evaluation to determine if a student has an IDEA-related disability and if the student needs special education services.
● Free appropriate public education	● LEAs are required to deliver an individualized program of special education services that confer meaningful educational benefit to a student.
● Individualized education program	● The vehicle by which school personnel and a student's parents develop a student's special education program. It is both a process for determining the program and a document outlining it.
● Least restrictive environment	● Students with disabilities must be educated to the maximum extent appropriate with students who are not disabled.
● Procedural safeguards	● A system of checks and balances to ensure that students with disabilities are afforded their rights under the EAHCA.
● Parental participation	● LEA personnel must collaborate with a student's parents in developing their child's special education program.

One of the important developments in the EAHCA was contained in a section of the law, now known as Part D, that required federal funds be invested in improving education for students with disabilities. Part D is intended to reform and advance personnel preparation and professional development in early intervention, special education, and transition services in order to improve results for children with disabilities. Through this section, the US Department of Education has invested millions of dollars in research, training, and technical assistance centers that actively support and advance effective practices for educational programming for students with disabilities. The law also funds parent training and information centers and community parent resource centers that support programs to educate and assist parents. Moreover, the law supports educators through grants to fund the preparation of special education teachers and their continuing education through professional development in research-based programming. The accomplishments of these investments are monumental and have led not only to improvements in special education, but also to meaningful improvements in the lives of all persons with disabilities.

In 1977, the US Department of Education issued the final regulations implementing the EAHCA. These regulations, which took effect at the start of the 1977–1978 school year, provided guidance for states and school districts developing special education programs in conformity with the EAHCA.

Critical Event #7: The US Supreme Court's Ruling in *Board of Education v. Rowley* (1982)

The US Supreme Court has heard a number of special education cases. Of these cases, arguably the most important was the first special education case heard by the court: *Board of Education of the Hendrick Hudson School District v. Rowley* (1982). In this case, the Court addressed the requirement in the EAHCA that students in special education receive FAPE.

Amy Rowley, a student at the Furnace Woods School in the Hendrick Hudson Central School District, was deaf and entitled to FAPE under the EAHCA. The year before Amy was to attend first grade at Furnace Woods, a meeting was held between her parents and school officials to determine future placement and special education services. The IEP team decided to place Amy in the regular kindergarten class to determine what supplemental services she might need. In addition to the special education services that were proposed, Amy's parents requested that a sign language interpreter be provided as part of her special education program. A sign language interpreter was placed in Amy's classroom for a trial period. But when Amy's IEP was developed for first grade the following year, a sign language interpreter was not provided.

Amy's parents then requested a due process hearing. In the ruling, the hearing officer agreed with the school district that federal law did not require the interpreter. The Rowleys appealed the decision to the New York Commissioner of Education. The Commissioner upheld the hearing officer's decision. Amy's parents then sued in federal district court, claiming that the district's refusal to provide a sign language interpreter had denied Amy FAPE.

Although Amy was doing better than the average child, the court ruled that Amy was not learning as much as she could without her handicap. Because of the disparity between Amy's actual achievement and her potential achievement, the district court ruled that Amy had been denied FAPE. The school district then appealed the decision to the US Court of Appeals for the Second Circuit. The appellate court affirmed the lower court's ruling. The school district then appealed the decision to the US Supreme Court. The high court granted certiorari (i.e., decided to hear the case). The high court considered two questions: (1) What is FAPE and (2) what is the role of state and federal courts in reviewing special education decisions?

In the majority opinion, Supreme Court Chief Justice Rehnquist wrote that FAPE consisted of educational instruction designed to meet the unique needs of a student with disabilities, supported by such services as needed to permit the student to benefit from instruction. The Court noted that the EAHCA required that these educational services be provided at public expense, meet state standards, and comport with the student's IEP. If the special education program allowed a child to benefit from educational services and was provided in conformity with the IEP and other requirements of the EAHCA, the student was receiving FAPE. Justice Rehnquist also wrote that any substantive standard prescribing the level of education to be accorded students with disabilities was conspicuously missing from the language of the EAHCA. Thus, the Court held that the EAHCA did not require the Hendrick Hudson school district to provide a sign language interpreter for Amy Rowley.

Critical Event #8: The Amendments to the EAHCA

The EAHCA has been reauthorized and amended a number of times since its original passage in 1975. Important amendments to the law were passed in 1986, 1990, 1997, and 2004.

The first amendments to the EAHCA, titled the Handicapped Children's Protection Act of 1986, authorized the award of reasonable attorney's fees to prevailing parties in special education lawsuits. The same year the EHA amendments required participating states to provide early intervention services to children with developmental disabilities from birth to their third birthday. Moreover, children from age 3 to 5 years were included under the federal law, thereby requiring schools to serve children with disabilities in special education programs.

In the 1990 amendments to the EAHCA, the name of the law was changed to the Individuals with Disabilities Education Act (IDEA). The law also (1) removed the term 'handicapped' and substituted the term 'disability'; (2) incorporated 'person-first' language, which puts the focus on the student rather than the label or category of disability (so instead of referring to a 'handicapped child,' the law used the phrase 'child with disabilities'); (3) added the categories of autism and traumatic brain injury to the list of eligible disabilities under the law; and (4) included a requirement to provide transition services for students by age 16 years.

The IDEA amendments of 1997 further strengthened the educational rights of students with disabilities. In fact, according to the House report on IDEA 1997, a major goal of IDEA 1997 was to improve and increase the educational achievement of students with disabilities. The IDEA amendments required that (1) students with disabilities would have access to the general curriculum and that the general educator would play a larger role in the process of developing a student's IEP, (2) IEP teams were to consider factors such as assistive technology and student behavior problems when crafting a student's IEP, (3) a student's parents were explicitly made a part of the team that makes eligibility and placement decisions for their child with disabilities, and (4) mediation should be used in resolving disputes between school personnel and a student's parents. The IDEA amendments of 1997 also addressed the issue of discipline, as Congress attempted to balance the need for discipline in schools with the requirement that special education students receive FAPE. To accomplish this, the law required that IEP teams consider positive behavior supports and interventions to address a student's problem behaviors in an effort to remediate and educate rather than merely punish.

In 2004, the Individuals with Disabilities Education Improvement Act (IDEIA) reauthorized and amended the IDEA. An important Congressional goal in passing IDEIA was to align IDEA with the ESEA, formerly referred to as No Child Left Behind, thereby increasing accountability for improving student performance. Thus, IDEIA included measures to increase academic results for students with disabilities such as requiring the use of scientifically based practices. The law also defined highly qualified special education teachers in line with the definition in the ESEA. Additionally, Congress sought to reduce the paperwork burden on teachers, expand options for parents, and reduce litigation.

The IDEIA also allowed school districts to spend up to 15% of their IDEA Part B funds on early intervening services (EIS) for students in kindergarten through grade 12, with an emphasis on students in kindergarten through grade 3, who had not been identified as needing special education or related services but who need additional academic and behavioral support to succeed in the general education environment. The purpose of this provision was to enable LEAs to identify young students who were at risk for developing academic and behavioral problems while they were still in general education settings, and then to address these problems by delivering interventions in a systematic manner by using research-based academic and behavioral interventions along with progress-monitoring systems. To implement EIS services, school district personnel would (1) provide school-wide academic and behavioral assessments for all students; (2) implement school-wide academic and behavioral services to students identified as needing additional services, including scientifically based reading instruction and positive behavior support systems; and (3) offer professional development activities to teachers and other school staff to enable them to deliver scientifically based academic and behavioral interventions.

Accomplishments Produced by Special Education Laws in the United States

In the early 1970s, efforts to ensure that students with disabilities received a public education that was appropriate to their individual needs began on the federal level. Before the 1970s, more than a million children with disabilities received no education whatsoever and many others received an

education that was not suited to their needs, often receiving such services in segregated settings. Additionally, many children with developmental disabilities were housed in institutions that focused on children's basic care needs rather than education. In 2014, almost 45 years after the original court cases, the battle to ensure access to public education for students with disabilities has been won. Today, access to public education is ensured for the approximately 12% of all public school students who have disabilities, over 6.5 million children, no matter how severe their disabilities may be. Moreover, programming now begins as soon a child is identified as having a developmental delay, even in infancy, which has led to many young children with disabilities and their families receiving high-quality early interventions that prevent or reduce the need for future services. Children with disabilities have gone from either being totally excluded from education or learning in segregated settings to learning in classrooms alongside their nondisabled peers in their neighborhood schools. More youth with disabilities now graduate from high school and over 30% enroll in postsecondary education. Additionally, trends in postschool employment of youth with disabilities have shown positive increases.

Clearly, special education laws in the United States have led to improvements in the lives of children and youth with disabilities. Much of the credit for this progress is due to parents of children and youth with disabilities who fought long and hard for their children's educational rights. These actions led to court cases and legislation at both the state and national levels, such as Section 504 of the Rehabilitation Act and the IDEA, which have been a major impetus for change. The protections of these laws have led states and school districts to focus on the education of students with disabilities and have led to a system of special education in the United States in which children and youth with disabilities are identified, their unique educational needs are assessed, meaningful programming is developed, and their progress in education is monitored and evaluated. Moreover, the research supported by the IDEA has increased the abilities of teachers to successfully work with students using evidence-based strategies and procedures. Special education laws in the United States have led to the creation of a system that has spread to many nations in the world.

Despite these impressive gains, the system can still be improved. For example, many students with disabilities still drop out of school, have difficulty finding jobs, and lead lives of dependency. To meet these challenges, special education laws will need to remain focused on ensuring the quality of educational experiences, fund research to improve programming, and stress accountability for results. These efforts, which began with the 1997 and 2004 amendments to the IDEA, will continue to ensure that students with disabilities receive a meaningful education, which in turn, will lead to an improved quality of life.

See also: Children and Law; Due Process in Law; Family Law in the United States; Juvenile Justice: International Law and Children's Rights; Law and People with Disabilities.

Bibliography

Ballard, J., Ramirez, B., Weintraub, F., 1982. Special Education in America: Its Legal and Governmental Foundations. Council for Exceptional Children, Reston, VA.

Board of Education v. Rowley, 458 U.S. § 176 (1982).

Brown V. Board of Education, 347 U.S. 483 (1954).

Education of the Handicapped Act of 1970, 20 U.S.C. § 1401 et seq.

Elementary and Secondary Education Act of 1965, 20 U.S.C. § 16301 et seq.

Huefner, D.S., Herr, C.M., 2012. Navigating Special Education Law and Policy. Attainment, Verona, WI.

Individuals with Disabilities Education Act Regulations, 34 C.F.R. § 300.1 et seq.

Individuals with Disabilities Education Act, 20 U.S.C. § 1401 et seq.

Katsiyannis, A., Yell, M.L., Bradley, M.R., 2001. Reflections on the 25th anniversary of the Individuals with Disabilities Education Act. Remedial and Special Education 22, 324–334.

Levine, E.L., Wexler, E.M., 1981. P.L. 94-142: An Act of Congress. Macmillan, New York.

Mills v. Board of Education of the District of Columbia, 348 F. Supp. 866 (D.D.C. 1972).

Pennsylvania Association for Retarded Citizens (PARC) v. Pennsylvania 343 F Supp. 279 (E.D. Pa. 1972).

Section 504 of the Rehabilitation Act of 1973, 29 U.S.C § 794 et seq.

Turnbull, H.R., Stowe, M.J., Huerta, N.E., 2007. Free Appropriate Public Education: The Law and Children with Disabilities, seventh ed. Love Publishing, Denver, CO.

U.S. Department of Education, n.d. Building the Legacy: IDEA 2004. IDEA.ed.gov (The most complete government Web site with information on the IDEA).

U.S. Department of Education, Office of Special Education and Rehabilitative Services, 2010. Thirty-Five Years of Progress in Educating Children with Disabilities through IDEA. Washington, D.C. PDF available at. http://www.edpubs.gov.

Winzer, M.A., 1993. History of Special Education from Isolation to Integration. Gallaudet Press, Washington, DC.

Yell, M.L., 2012. The Law and Special Education, third ed. Pearson, Upper Saddle River, NJ.

Specialization and Recombination of Specialties in the Social Sciences

Mattei Dogan[†], National Center for Scientific Research, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14851–14855, © 2001, Elsevier Ltd.

Abstract

The multiplication of scientific activities has brought a division of labor. The history of scientific advancement is a history of concatenated specialization. There is more communication between specialties belonging to different disciplines than between specialties within the same discipline. Specialization within disciplines, accompanied by their fragmentations, is the first process. Recombination of specialties across disciplinary borders is the second. Such a recombination has been called hybridization. A hybrid is a combination of two branches of knowledge. Many of the most creative specialties are hybrid specialties. Recombination of specialties appears in the borrowing of concepts, theories, and methods. The citation patterns in academic publications permit the analysis of the process of specialization and recombination by interferences of disciplines. A distinction is needed between interdisciplinary amalgamation and hybridization. Diminishing marginal returns have been perceived in the core of disciplines and a high productivity at their margins. This phenomenon has been called the paradox of density.

Specialization is an old phenomenon, observed as early as Auguste Comte. Recombination of specialties is also an old trend: Robert Merton noticed it in the early 1960s: "the interstices between specialties become gradually filled with new cross-disciplinary specialties" (Merton, 1963: p. 253). The multiplication of scientific activities and the growing complexity of knowledge have brought a division of labor, visible in increasing expertise in all domains. The history of scientific advancement is a history of concatenated specialization. In spite of this evidence, "even since the social sciences began to develop, scholars have repeatedly expressed apprehension about the increasing fragmentation of knowledge through specialization" (Smelser, 1967: p. 38).

Fragmentation and specialization are the two faces of the same coin. This double process is a condition of scientific advancement. The term 'discipline' is inherited from the vocabulary of the nineteenth century and is understood as a branch of instruction for the transmission of knowledge and as a convenient mapping of academic administration. In all universities, teaching, recruitment, promotion, peer review, and financing are still organized along disciplinary lines, but the great scientific laboratories and the most creative research centers are anchored on specialties, which in many cases cross the borders of formal disciplines, or are located at the interstices between disciplines. In the sociology of science the term 'discipline' is defined as a 'cluster of specialties' (Crane and Small, 1992: p. 198), the specialty being the main source of academic recognition and of professional legitimation (Zuckerman, 1988: p. 539).

In the natural sciences, specialization is celebrated as a mark of competence. Each contemporary natural science from astronomy to zoology is a conglomerate of dozens of specialties. Only in some social sciences, particularly in sociology, political science, psychology, and anthropology, does one hear complaints and lamentations about the fragmentation of the discipline into specialties. The word 'crisis' appears in the title of many books and articles about the identity and the contours of these disciplines (Horowitz, 1993; Turner and Turner, 1990). In reality, such 'crises' are symptoms of growth and expansion.

[†] Deceased.

Specialization

Analysis means breaking things into parts. All sciences have made progress, from the sixteenth century on, by internal differentiation and cross-stimulation among emergent specialties. Each specialty developed a patrimony of knowledge. With the growth of these patrimonies specialization became a necessity. Increasing specialization has led to the creation of subdisciplines. A specialty requires a high qualification in a delimited domain. The process of specialization has tended to disjoin activities which had previously been united, and to separate scholars belonging to the same formal discipline, but who are interested in different fields. Specialties are considered as parts of a larger discipline but "the modern scholar normally does not attempt to master his discipline in its entirety. Instead he limits his attention to one field, or perhaps two" (Somit and Tanenhaus, 1964: p. 49). Each formal discipline gradually becomes unknown in its entirety. No theory can any longer encompass the whole territory of a discipline. Talcott Parsons was the last sociologist to attempt such a theory. As a discipline grows, its practitioners become specialized and must neglect other areas of the discipline. It is for this reason that scissiparity, the amoeba-like division of a discipline into two, is a common process of fragmentation. Specialization provides researchers with methods so that they need not develop them anew. Refinements in methods are more easily transmitted among specialists. Specialization within disciplines is visible in national and international professional meetings. Everyone who has participated at a gathering of several thousand people has noticed the absence of coherence: 20–30 panels may run simultaneously, most of them mobilizing only a handful of people. The plenary sessions attract only a small minority, with most participants uninterested in issues encompassing the entire discipline.

Disciplines fragment along substantive, epistemological, methodological, theoretical, and ideological lines. To those in the field, the theoretical and ideological divisions are likely to seem more important than to others. Ralph Turner gives a description of this process in sociology, based on his experience as Editor of an important journal: "In the 1930s and

1940s the aspiration to be a general sociologist was still realistic. There was a sufficiently common body of core concepts and a small enough body of accumulated research in most fields of sociology that a scholar might make significant contributions to several, and speak authoritatively about the field in general. It is difficult to imagine the genius necessary for such accomplishments today" (Turner, 1991: p. 70).

The division of disciplines into specialties should be distinguished from their fragmentation into schools and sects. The term school refers to a group of scholars 'who stress a particular aspect.' The proponents of a school are emotionally committed to their approach. A school is simultaneously a subdivision of a field and a species of cult or sect (Smelser, 1967: pp. 6–7). While the multiplication of specialties is related to the maturation of the discipline, "the presence of numerous schools in a discipline generally be tokens a relative scientific immaturity" (Smelser, 1967: pp. 6–7). The same diagnosis is formulated by Gabriel Almond who, by crossing two dimensions, an ideological one and a methodological one, described four types of schools in political science (Almond, 1990: p. 15). These types are not determined by the process of specialization.

Various authors stress the importance of the patrimony's expansion for subsequent fragmentation. As disciplines grow, they fragment and most parts become the patrimonies of individual subfields; a rare few are passed down in the patrimony of several formal disciplines and become the classics. For Randall Collins, the increasing specialization in sociology could be explained by the growth in scale in the profession during the late twentieth century: "How does one make oneself visible when the sheer number of competitors increases? The field differentiates into various specialties. Instead of seeking recognition in the larger field, one opts for a smaller arena within which the discourse and the search for originality and for allies can go on" (Collins, 1986: p. 1340). A process of differentiation without integration is visible across disciplinary borders. Specialists "seem scarcely to recognize the names of the eminent practitioners in specialties other than their own" (Collins, 1986: p. 1340).

Another scholar who emphasizes the importance of specialties in the organization of scientific communities is Harriet Zuckerman: "Much historical and sociological evidence suggests that specialties have been prime audiences for many scientists: they are the explicit and tacit audiences – the reference groups – to which they address their work, just as they are the prime sources of wherewithal and rewards for that work" (Zuckerman, 1988: p. 539). As in some cathedrals, there are more ceremonies in the chapels than under the principal nave.

Specialties versus Disciplines

Analyzing the relationships between specialties *within* disciplines and between specialties *across* disciplines, two kinds of disciplines can be perceived: "restricted disciplines, such as most physical sciences, which would be expected to exhibit a high degree of linkage between different research areas within the discipline, but less linkage to other disciplines," and unrestricted sciences, such as most social sciences, which would be likely to exhibit relatively diffuse links among research areas both within and outside the disciplines (Crane and Small, 1992: p. 200).

Depending on the definition we adopt, there are in the social sciences between 12 and 15 formal disciplines, but dozens of specialties, sectors, fields, and subfields. In sociology, for instance, there are some 50 specialties, as indicated by the list of research committees of the International Sociological Association. There are as many in the International Political Science Association. Most of these groups cooperate to a limited extent within their respective discipline.

There is more communication between specialties belonging to different disciplines than between specialties within the same discipline. Sociometric studies show that many specialists are more in touch with colleagues who belong officially to other disciplines than with colleagues in their own discipline. The 'invisible college' described by Robert Merton and other sociologists of science is an eminently interdisciplinary institution because it ensures communication not only from one university to another and across all national borders, but also, and above all, between specialists attached administratively to different disciplines. The networks of cross-disciplinary influence are such that they are obliterating the old classification of the social sciences.

The division of disciplines into subfields tends to be institutionalized, as can be seen in the organization of large departments in many American and European universities. Another good indication of the fragmentation of the discipline is the increasing number of specialized journals. In the last two decades several hundred specialized journals in English relevant to social sciences have been launched. Most of these new journals cross the borders of two or three disciplines. Some other new journals have appeared in French and German. European unification has had an impact on the development of cross-national journals focusing on special fields. Specialization has consequences on the role of national professional associations and on the general journals. The size of specialties varies greatly: Egyptology and biosociology are small specialties; socialization and clientelism are very large. A segment which drifts away from the main body of the discipline has to reach a certain critical size before being recognized. Some of them enjoy a relative autonomy within the discipline.

A distinction has to be drawn between specialization within a formal discipline and specialization at the intersection of several monodisciplinary subfields. The latter can occur only after the former has become fully developed.

Recombination of Specialties across Disciplines

Specialization within disciplines, accompanied by their fragmentation, is the first process. Recombination of specialties across disciplinary borders is the second. In the history of science a twofold process can be seen: on the one hand, a fragmentation of formal disciplines and, on the other, a cross-disciplinary recombination of the specialties resulting from fragmentation. The new field may become independent, like social psychology, or may continue to claim a dual allegiance, like historical sociology. In the latter case, librarians are not

sure whether to place a work in the category of history or of sociology.

Such a recombination has been called hybridization (Dogan and Pahre, 1990: p. 63). A hybrid is a combination of two branches of knowledge. Jean Piaget suggested a different analogy: genetic recombination (Piaget, 1970: p. 524). The reasons for hybridization are clear enough: specialization leaves gaps between disciplines; specialties fill the gaps between two existing fields. Many of the new specialties are to a large degree hybrid specialties. The phenomenon is even more obvious in the natural sciences, e.g., in the awards of the Nobel prizes.

Some scholars recommend an 'interdisciplinary' approach. Such a recommendation is not realistic because it overlooks the process of specialization. Research enlisting several disciplines involves in fact a combination of segments of disciplines, of specialties, and not whole disciplines. The fruitful point of contact is established between sectors, and not along disciplinary boundaries. Considering the trends in the social sciences, the word 'interdisciplinarity' appears inadequate since it carries a hint of superficiality and dilettantism, and consequently should be avoided and replaced by hybridization of specialties. Different disciplines may proceed from different foci to examine the same phenomenon. This implies a division of territories between disciplines. In contrast, hybridization implies an overlapping of segments of disciplines, a recombination of knowledge in new specialized fields. Innovation in each discipline depends largely on exchanges with other fields belonging to other disciplines.

When old fields grow they accumulate such a mass of material in their patrimony that they split up. Each fragment of the discipline then confronts the fragments of other fields across disciplinary boundaries, losing contact with its siblings in the old discipline. A sociologist specialized in urbanization has less in common with a sociologist studying elite recruitment than with a geographer interested in the distribution of cities; the sociologist studying social stratification has more in common with his colleague in economics analyzing income inequality than he does with a fellow sociologist specialized in organizations; psychologists studying child development are much more likely to be interested in developmental physiology or in the literature on language acquisition than they are in other branches of psychology. A political scientist researching political socialization reads more sociological literature on the agents of socialization (family, church, school, street-corner society, cultural pluralism, etc.) than literature in political science on the Supreme Court, legislative processes, party leadership, or the recruitment of higher civil servants. Those working on the subfield of international relations in the nuclear age have little recourse to the classical literature in political science, but rather to economics, technology, military strategy, diplomatic history, game theory, nuclear physics, and engineering.

Most hybrid specialties and domains recognize their genealogical roots: political economy, political sociology, social geography, historical sociology, genetic demography, psycholinguistics, social anthropology, social ecology, biogeography, and many others. Some hybrid sciences do not indicate their filiation in their name: cognitive science, paleoarcheology. Eight social sciences also have roots in the natural sciences: demography, geography, psychology, anthropology, linguistics, archeology, criminology, and cognitive science. The hybrid specialties branch out in turn giving rise, at the second generation, to an even larger number of hybrids.

Recombination of specialties appears in the borrowing of concepts, theories, and methods (Dogan, 1996). Concepts are vehicles for communication between specialties belonging formally to different disciplines. Concepts are not equally important in all specialties. For instance, alienation is more frequently used in social psychology than in political economy, segregation more in urban studies than in comparative politics, and charisma more in political sociology than in social ecology.

There are very few general theories accepted by the majority of scholars in any discipline; the best known theories are specialized theories: they are what Robert Merton calls middle-range theories, which do not encompass an entire discipline, but only a part of it. A specialized theory can cross several disciplines with a minimal effort of adaptation. For instance 'rational choice' – considered by some as a theory and by others as a school or sect – is, except for some stylistic features, basically the same in political science, in sociology, and in economics. The theory of methodological individualism does not change on moving from one discipline to another. Two to three dozen theories peregrinate freely among formal disciplines. The high interdisciplinary mobility of specialized theories and of specialized concepts, and also of research methods, gives some centripetal orientation to the wide range of social sciences. The cross-disciplinary diffusion of concepts and theories is in fact a diffusion among specialties belonging to different disciplines.

Methodology divides some disciplines as much as ideologies, and at the same time builds bridges between specialties across the disciplinary borders. The scholars who practice factor analysis understand each other well no matter what their official disciplinary affiliation, but most of them encounter some difficulties in communicating, *within* their discipline, with their colleagues who do not practice quantitative methods. In some departments methodological controversies generate factions, which may or may not coincide with divisions between specialties.

The citation patterns in academic publications permit us to measure empirically the degree of coherence of a discipline, the relations between specialties within disciplines, and the interferences of disciplines. If specialists in a given subdiscipline tend to cite mostly, or exclusively, specialists of the same subdiscipline, and if relatively few authors cite across the borders of their subdiscipline, then the formal discipline presents a low degree of internal coherence. In this case the real loci of research are the specialties. If, on the contrary, a significant proportion of authors cross the borders of specialties, then the discipline as a whole appears more or less as an integrated territory (Dogan, 1997). Political scientists have borrowed in the 1930s from law, in the 1950s from sociology and law, and in the 1970s from sociology, philosophy, economics, history, and psychology. According to Craig Calhoun, "Sociology is neither the most open nor the most insular of social sciences" (Calhoun, 1992: p. 143). A study of patterns of citation during the period 1936–75 has found that sociologists cited 58% of the time articles in sociology journals, political scientists cited 41% of the time scholars from their own discipline, anthropologists referred 51% of the time to their colleagues, psychologists 73% to their own kin, and 79% of the

economists did the same (Rigney and Barnes, 1980: p. 117). In each social science a significant proportion of theoretical, methodological, and substantive communication has been with other disciplines, the most open being, in the late twentieth century, political science and the most autarchic economics.

In an analysis of journals identified as belonging to sociology and economics, a significant shift has been found from sociology to 'interdisciplinary sociology' and from economics to 'interdisciplinary economics,' between 1972 and 1987. The criteria for trade between disciplines was the proportion of citing references in journals of the respective discipline (Crane and Small, 1992: pp. 204–205). An analysis by the same authors in terms of clusters of references shows a clear increase in interdisciplinary relationships.

Most social scientists are specialists contributing primarily to their own subdiscipline. For instance, the *Handbook of Sociology* edited by Neil Smelser (1988) is divided into major specialties, without an overarching structure. Smelser is aware of this heterogeneity. Noting that Talcott Parsons had 'exaggerated the internal unity of the discipline,' which 'if applied today would appear almost ludicrous' (p. 12), he writes: "there appears to be no present evidence of an overarching effort at theoretical synthesis … and little reason to belief that such an effort is on the horizon," adding that "we have by now gone too far down the road of specialization and diversification" (p. 12). The ship of sociology is already built with watertight compartments. If the borders between specialties *within* disciplines are hermetic, in exchange, paradoxically enough, the frontiers between formal disciplines are open.

A distinction is needed between interdisciplinary amalgamation and hybridization by recombination of specialties belonging to different disciplines. A 'unified social science' was visible only in the early phase of the development of social sciences, when disciplines were not yet differentiated. Hybridization of specialties came much later, after the maturation of the process of internal fragmentation of disciplines. The word interdisciplinary is appropriate only for a certain phase of social sciences at the beginning of the twentieth century; it becomes misleading when it is used to describe contemporary trends, because at the end of the twentieth century only specialties are overlapping, not entire disciplines.

Some may confound recombination and synthesis. A synthesis brings a new interpretation, a personal or stylistic achievement. The difference is most clear in history. Arnold J. Toynbee's theory of history and Hobsbawn's *The Age of Empire* are syntheses. Fernand Braudel's *The Mediterranean World* is a recombination of segments of social sciences, largely history and geography. Perry Anderson's *Lineages of the Absolutist State* is largely synthesis, whereas Karl Wittfogel's *Oriental Despotism* is largely recombination. The recombinations are varied. Child development includes developmental physiology, language acquisition, and socialization. Indo-European studies encompass segments of historical linguistics, archeology, prehistory, and botany. Scholars in criminology come from subfields in law, sociology of deviance, social psychology, endocrinology, urban studies, social economics, and ethnopolitics. The study of artificial intelligence encompasses formal logic from philosophy, grammar and syntax from linguistics, and computer programing from computer science. Folklore studies include sectors from historical linguistics, cultural anthropology, social history, and comparative literature.

In one generation the number of social scientists has increased 10-fold. Their status in the profession depends mostly on their contribution to knowledge, and much less on their performance as teachers. As teachers they are generalists, as researchers, specialists. A strong feeling of diminishing marginal returns is perceived among the mass of social scientists. This phenomenon of saturation and overcrowding has been called the 'paradox of density': the tendency of densely populated disciplines to produce less innovation per head in spite of the greater effort applied in these overcrowded domains (Dogan and Pahre, 1990: p. 32). What is perceived by some scholars as 'dispersion' and 'disintegration' should be considered as manifestations of the growth of specialties, resulting in the weakening of the saturated disciplinary core, a process engendered by centrifugal and autonomous tendencies of the specialties across disciplines. The recombination of specialties across disciplines requires a remodeling of the intellectual map of the behavioral social sciences.

Bibliography

Almond, G.A., 1990. Discipline Divided, Schools and Sects in Political Science. Sage, Newbury Park, CA.

Calhoun, C., 1992. Sociology, other disciplines and the project of a general understanding of social life. In: Halliday, T.C., Janowitz, M. (Eds.), Sociology and Its Publics: The Forms and Fates of Disciplinary Organization. University of Chicago Press, Chicago, pp. 137–196.

Collins, R., 1986. Is 1980s sociology in the doldrums? American Journal of Sociology 91, 1336–1355.

Crane, D., Small, H., 1992. American sociology since the seventies: the emerging identity crisis in the discipline. In: Halliday, T.C., Janowitz, M. (Eds.), Sociology and Its Publics: The Forms and Fates of Disciplinary Organization. University of Chicago Press, Chicago, pp. 197–234.

Dogan, M., 1996. Political science and the other social sciences. In: Goodin, R., Klingemann, H.D. (Eds.), A New Handbook of Political Science. Oxford University Press, Oxford, UK, pp. 97–130.

Dogan, M., 1997. Les nouvelles sciences sociales: fractures des murailles disciplinaires. Revue Internationale des Sciences Sociales 153, 467–482.

Dogan, M., Pahre, R., 1990. Creative Marginality: Innovation at the Intersections of Social Sciences. Westview Press, Boulder, CO.

Horowitz, I.L., 1993. The Decomposition of Sociology. Oxford University Press, New York.

Merton, R., 1963. The mosaic of the behavioral sciences. In: Merton, R.K., Berelson, B. (Eds.), The Behavioral Sciences Today. Basic Books, New York, pp. 247–272.

Merton, R. (Ed.), 1968. On sociological theories of the middle range. Social Theory and Social Structure. Free Press, New York, pp. 39–53.

Piaget, J., 1970. General problems of interdisciplinary research in UNESCO. In: Piaget, J. (Ed.), Main Trends of Research in the Social and Human Sciences. Mouton, Paris, pp. 467–528.

Rigney, D., Barnes, D., 1980. Patterns of interdisciplinary citation in the social sciences. Social Science Quarterly 61, 114–127.

Smelser, N.J., 1967. Sociology and the other social sciences. In: Lazarsfeld, P., Sewell, W., Wilensky, H. (Eds.), The Uses of Sociology. Basic Books, New York, pp. 3–44.

Smelser, N.J. (Ed.), 1988. Handbook of Sociology. Sage, Newbury Park, CA.

Somit, A., Tanenhaus, J., 1964. American Political Science, a Profile of a Discipline. Atherton, New York.

Turner, R., 1991. The many faces of American sociology. A discipline in search of identity. In: Easton, D., Schelling, C.S. (Eds.), Divided Knowledge across Disciplines, across Cultures. Sage, London, pp. 59–85.

Turner, S.P., Turner, J.H., 1990. The Impossible Science: An Institutional Analysis of American Sociology. Sage, Newbury Park, CA.

Zuckerman, H., 1988. Sociology of science. In: Smelser, N.J. (Ed.), Handbook of Sociology. Sage, Newbury Park, CA, pp. 511–574.

Speech Errors, Psychology of

Merrill F Garrett, University of Arizona, Tucson, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Garrett, volume 22, pp. 14864–14870, © 2001, Elsevier Ltd.

Abstract

The use of error profiles in the study of language is reviewed, with specific attention to errors made in fluent speech by normal adult speakers. These are errors of performance, not errors of knowledge – i.e., speakers use an unintended sound or word, or place an intended element in an unintended location. Such phenomena are generally referred to under the label 'speech errors' in the psychological and linguistic literature The major results of error analysis indicate a staged architecture for sentence generation in which meaning, syntax, and sound representations are successively integrated. Speech error modeling converges in most major respects with the results of experimental studies of language generation.

The generation of spoken language is a complex task normally accomplished with high accuracy. Human beings are, on average, consummate 'talkers.' But, speakers do err occasionally, and in ways that reflect the mental processes underlying the act of speaking. That is the topic pursued here: What are the speech error patterns of normal language use, and what do they tell us about language processing?

Various departures from the ideal of perfectly articulated, fluent speech occur in ordinary conversation: hesitations, false starts, changes of mind in mid-sentence, and misarticulations. These have their own interest, but are not the focus of this discussion. The term 'speech error' has come to be used in a semitechnical fashion to pick out particular types of errors from the welter of everyday language. The targets are those fluently uttered forms in which some element or elements of the intended utterance are missing, mis-selected, or misordered. Speakers often, though not invariably, notice and correct these failures. Errors arising from ignorance of a word or of 'proper usage' and various nonfluencies are thereby distinguished from speech errors from this perspective.

Attention to speech errors has a long history ranging from the exploitation of their humorous potential ('Spoonerisms') to their recruitment for systematic scientific study. Many different investigators have contributed to this area, and the scientific purposes to which such data have been put are diverse, including both structural linguistic interpretation and psychoanalytic interpretations in the nineteenth and early twentieth century, and more recently, interpretation for modeling real-time language use. In this latter exercise, the collection and interpretation of speech errors strongly affected the initial framing of contemporary models of language production, and is systematically linked to the development of a number of experimental paradigms currently used to study production processing.

A Methodological Note

How is error data acquired? Several investigators have collected extensive records of errors heard during everyday language use. A classic such effort was that of Rudolf Meringer (Meringer and Meyer, 1895) in the late nineteenth century. Cutler and Fay (1978) describe his methods in their introductory essay to a reissue of Meringer's work. The practices they describe are characteristic of later diary studies. Typically, a written record of an error is made soon after observation (normally, a few seconds to a few minutes), with annotations regarding pronunciation and setting. Such observation has the strong appeal of face validity for processes of language formulation. However, influences of perceptual filtering and memory distortion must be carefully weighed. It is clear that minor phonetic deviations may be missed, and only changes clearly discernable with respect to the speaker's target are well represented in such error corpora. See Cutler (1982) for detailed discussion of these and related issues.

Diary studies exist for several languages. It may be said with some assurance that whether the target language is Chinese, Dutch, English, German, Italian, Spanish, Turkish, or other, the same general types of phenomena arise. Specific language features may afford options for error – e.g., Chinese tone errors – not available in English or related languages. Nevertheless, the general behavior of errors involves similar mechanisms across languages. Discussion here focuses on English, with some commentary on representative language differences. Error examples are taken from papers in the bibliography.

Types of Error

In the popular literature, 'speech error' is synonymous with 'Spoonerisms,' named after the Dean of New College, Oxford, Rev. W.A. Spooner (1844–1930), who was notorious for his frequent misplacements of speech sounds, sometimes to extraordinary effect. For example, in lecturing a wayward student, he is alleged to have said: "You have hissed all my mystery lectures. … and tasted two worms" ("missed all my history lectures … wasted two terms"). But Spoonerisms are special in several respects. They suggest that sound errors usually make words, and that they make a kind of wild and humorous sense. Neither condition is typical. Furthermore, Spoonerisms primarily exploit sound error, and the relevant phenomena are much more general. No account of Rev. Spooner's performance is given here; for comment, see, e.g., Robbins (1966).

Notwithstanding the exclusion of Rev. Spooner's legacy, there is no denying the charm of suitably chosen errors as a way to display the hidden structure in language. The errors are intricate, and they illustrate essential features of language processing. Examples (1)–(14) are representative:

1. '…sat down on a *s*ot *h*oddering iron' (…hot soldering…)
2. 'I won't hit the ba*rd* ha*ll*' (…ball hard)
3. 'I have a case of l*ay* gr*o*de flu' (…low grade…)
4. '…so I can start the s*tape* back up (…the tape…)
5. 'The curren*th* month is… (The current…)
6. '…an anguage *l*acquisition device… (…a language acquisition…)
7. 'Is your *purr*y *kitt*ing' (…kitty purring)
8. 'Make it so the *apple* has less *tree*s (…so the tree has less apples.)
9. 'She *sings* everything she *writes* (…writes everything she sings)
10. 'Bensinger takes to that left-field wall like *water* takes to *a duck* (…like a duck takes to water)
11. 'I've got whipped cream on my *mushroom*! (…on my mustache!)
12. 'You put your *nose*… your finger on it (…put your finger on it)
13. 'My! So that's the Porche *symblem* (symbol/emblem)
14. '*It doesn't care* (It doesn't matter/I don't care).

A rough task analysis sets the stage for interpretation of such errors. It is evident that systems for speaking will necessarily include mechanisms that select from tens of thousands of available words just those appropriate to an intended message ('lexical retrieval'), mechanisms that integrate the words into sentence structures ('syntactic construction'), and mechanisms that orchestrate the result of those operations into control structures ('phonological and prosodic integration') that drive motor outputs (articulation). Psychological explanation must reconstruct this complex of mental events.

Descriptively, in examples (1)–(7), sounds move around in various ways – they are exchanged anticipated, perseverated, and shifted. Important features of sound error include the following: Moved elements are *phonetically similar*, they correspond in *syllabic position*, and their immediate *phonetic environments are similar* (Boomer and Laver, 1968; Dell, 1984; Nooteboom, 1969; Caldognetto et al., 1987; del Viso et al., 1991; McKay, 1970; Vousden et al., 2000). The fact that error likelihood is so strongly influenced by similarity of structural environment suggests that errors arise, in part, out of mechanisms that associate independently specified processing elements (sounds, in this example) with a *planning framework* ('syllable or word frames,' in this instance). These errors and related phenomena indicate how sounds that make up words enter planning structures for sentences. In the task analysis, this reflects phonological and prosodic integration processes.

Examples (8)–(10) are 'word movement' errors. Some are of special interest because they reflect an analysis of words into stem and affix. In Example (8), 'apples' is decomposed into 'apple + s' – more abstractly ['apple' + plural marker]. The technical term in linguistic science for such a decomposition is morphemic analysis. Notice that the two types of morpheme (stem and affix) behave in very different ways. The stem moves, but the affix stays behind. One can also describe this error in terms of the notion of structural frame. By this hypothesis, the affix would be considered part of a phrasal frame that stem morphemes are assigned to as a sentence develops. This idea proves applicable to a wide range of error observations, and is strikingly manifest in languages with richly developed morphological structures. (Related facts are discussed in Section Generalizing the Observations about Processing Frames.) In the task analysis, these errors implicate syntactic integration.

Examples (11) and (12) provide another perspective on errors. Descriptively, these are *word substitutions*. Several points are relevant. First, (11) displays a phonetic similarity between target and intruding element; the phonetic similarities between 'mustache' and 'mushroom' are representative. Moreover, both words are nouns. Thus, the syntactic structure of the intended utterance is preserved. One may plausibly treat this as a failure – perhaps triggered by phonetic similarity – of mechanisms that retrieve elements from an inventory of word forms. But what about errors like (12)? Grammatical category is preserved, but phonetic similarity is not apparent. Quite unlike (11), an obvious meaning relation holds between target and intrusion. Section Lexical Selection Processes: Word Substitutions discusses this contrast. In the task analysis, these errors implicate lexical selection processes, and they encompass both form and meaning relations.

Examples (13) and (14) are *blend* errors. In word and phrase blends, two processing elements compete for a *single output slot*. Speakers are often aware of the competition and can report experiencing it. The competing expressions are, with rare exception, of the same grammatical class and are semantically equivalent for the intended message. This phenomenon is a prima facie demonstration of some parallelism in sentence production. Such errors implicate lexical selection processes and aspects of phrasal and phonological composition.

Interpretations of Error Phenomena

Speech errors have many variants. The challenge is to explain their similarities and differences in ways that link them to accounts of production mechanisms at the necessary levels of representation (phrases, words, morphemes, syllables, sounds). A robust generalization across the error types is relevant: Like elements interact, i.e., sounds interact with sounds, words with words, and phrases with phrases. These interactions indicate vocabularies of 'units,' and the error data shows they are linked in an architecture of processing stages discussed in Sections Accommodations in Error Processes–Generalizing the Observations about Processing Frames.

Accommodations in Error Processes

Errors move elements of the intended utterance to 'new' positions. Because languages display many local dependencies of form, such displacements could violate these. But, typically they do not. Example (6) illustrates: '*a* language acquisition device' became '*an* anguage lacquisition device.' The indefinite article changed from 'a' to 'an,' and so *accommodated* to the vowel onset ('anguage') created by the error. In Example (14), the bound morpheme derives 'easily' from 'easy' shifts. But, the resulting isolated stem 'easy' was correctly pronounced

as /izi/ rather than with a reduced final vowel (as would have been appropriate had the affix remained in place).

15. '… *easy enoughly* …' (easily enough)

Many such variations triggered by changes in phonetic context appear in accommodations. By these observations, processes that fix the surface forms of words are ordered later than the error-generating structures and associated operations: First, an abstract morphological and phonological representation is constructed and then regular processes flesh out its phonetic detail. This is a familiar notion in linguistics, and its expression in error phenomena is not greatly surprising (see Fromkin, 1971). But, bear in mind that it could have turned out otherwise. The speaker of Example (6) might have said 'a anguage' but did not; the speaker of Example (15) might have said 'easuh enoughly,' but did not; and so on. Error processes on this evidence are tightly bound to the structural systems that govern error-free speech.

Other Evidence for Ordering of Processes: Sound Exchanges and Word Exchanges

These error classes reveal syntactic and phonological planning structures. Sound interactions, and particularly exchanges as in (1)–(3), typically involve words in the same phrase. The relevant structure is prosodic (see Boomer and Laver, 1968; Fromkin, 1971; Garrett, 1980). Prosodic grouping reflects syntax, but is also motivated by intonation and the phonological structure of the words chosen to express a given syntactic form. Treating the domain of sound exchange as prosodic may also explain some cross-language differences. For example, sound exchanges in English are predominantly between words, but in Spanish, within words (del Viso et al., 1991). The greater morphological complexity and length of Spanish words, as compared to English, would require exchanges between words to cross prosodic boundaries more often for the former than the latter.

Word and phrase exchanges sharply differ from sound exchanges. As Examples (9), (10), and (16)–(18) suggest, they arise for elements in distinct phrases that correspond in grammatical category

16. '… I left my *briefcase* in the *cigar*' (… my cigar in the briefcase)
17. 'What we want to do is train *its tongue* to move *the cat*' (… the cat to move its tongue)
18. 'As you *reap*, Roger, so shall you *sow*' (As you sow … so shall you reap)

Nouns exchange with nouns, verbs with verbs, adjectives with adjectives (Bierwisch, 1970, Garrett, 1975; Nooteboom, 1969). On this evidence, operations at the point of error involve the grammatical roles of words (or minimal phrasal units). Bear in mind that sound exchanges differ from word exchanges in both respects. Recall also that sound exchanges implicate many detailed features of phonological structure, whereas word exchanges are minimally so affected (see Dell and Reich, 1981; del Viso et al., 1991). These differences in the role of phrasal domain, grammatical category, and phonological description suggest two processing systems. Syntactic constituents are built by a multiphrasal processor that does not represent the sound structure of its elements (i.e., of word stems, morphemes, simple phrases); failures yield word and phrase exchanges. A subsequent process then elaborates sound sequences for words within a phrase group; failures yield sound exchanges. Multiple phrasal units are present at the syntactic level, and a narrower planning window develops sound structures and pronunciation codes in prosodic groups.

Stranding Errors and Multiple Levels of Processing

The exchanges in Examples (7) and (8) decompose target words into their constituent morphemes and strand affixes. One finds stranding of a variety of productive structural markers, predominantly those of inflectional device (e.g., number, tense, and aspect markers, gerundives, comparatives). Evidence indicates that stranding arises at two points: within phrase (re prosodic groups for sound exchanges), and between phrase (re structures for word exchanges) (see Garrett, 1993). These ideas extend the earlier discussion of accommodation. That discussion described it as a phonologically conditioned response to error movements. But, consider Examples (19) and (20)

19. '…that I would *hear* one if I *knew* it' (…that I would know one if I heard it)
20. '…drei *Wochen* im *Tag* gearbeitet' (drei Tage in der Woche gearbeitet).

Here there is accommodation, but, strikingly, it involves irregular forms of inflected words. In Example (19), the appropriate irregular verb forms for 'know' and 'hear' appear. The German example (Bierwisch, 1970) shows several accommodations. Singular '*Woche*' takes plural marker 'en'; plural '*Tage*' has the singular form '*Tag*.' Further, German nouns bear grammatical gender: '*Woche*' is feminine, '*Tag*' is masculine. The articles required thus differ, and may interact with prepositions. Here, number and gender are indeed appropriate, including interaction with the preposition ('*im*' is a contracted form of '*in dem*').

What accounts for this impressive range of accommodations for these error types? They are lexically dependent and so not derivable from phonological environments (as described in Section Other Evidence for Ordering of Processes: Sound Exchanges and Word Exchanges). In fact, lexically dependent accommodation occurs predominantly for between phrase stranding exchanges for elements of corresponding grammatical category. In the account sketched in Section Other Evidence for Ordering of Processes: Sound Exchanges and Word Exchanges, such errors must occur earlier than sound errors – at the level of abstract phrasal construction, prior to the phonological interpretation of content words. Development of the irregular forms then follows from affiliation of the switched stems with stranded inflectional features.

Generalizing the Observations about Processing Frames

Stranding shows that inflectional elements are separately represented when stems are inserted into phrasal frames. This can be expressed by assuming that inflectional elements are parts of structural frameworks. If exchanges arise when lexical elements are misassigned to structural frames, it follows that parts of frames should not exchange. This expectation is borne out in

error corpora; inflectional exchanges are rare (Garrett, 1975; Stemberger, 1985). Another class of observations (morpheme shifts), not discussed here, may also be connected to this picture (see Garrett, 1980).

There is some basis for treating all functional elements both bound and free in this way. The claim is that functional elements (inflections, determiners, complementizers, etc. – the minor grammatical categories) are derivatives of structure and are thus recruited for incorporation into sentences by different mechanisms than those responsible for the selection of content items (nouns, verbs, prepositions, adjectives, adverbs – the major grammatical categories). This contrast has been discussed in the processing literature under the label of 'open class' and 'closed class' vocabularies. It might be construed as a distinction between those elements individually selected under conceptual control ('content elements') and those that are reflexes of the logic and syntax of the sentence (Levelt, 1989; Bock and Levelt, 1994). See Myers-Scotton and Jake (2001), for a related account based on 'code-switching' (a bilingual language practice that fluently combines elements from two languages into a single utterance). Note, however, that some accounts of speech error link profiles for these vocabulary types to their relative frequency of occurrence in the language (Stemberger and McWhinney, 1986; Dell, 1991). On this view, very high-frequency rates, rather than structural roles, govern error behavior of functional elements.

Lexical Selection Processes: Word Substitutions

Word substitutions provide a data class distinct from movement error. These reflect failures in lexical retrieval; Examples (21)–(25) illustrate

21. 'Are you left-handed? No, I'm amphibian, no, ambivalent… arghh, ambidextrous!'
22. 'If you can find a *garlic* around the house, I'll take it' (gargle)
23. 'All I want is something for my shoulders' (elbows)
24. 'He rode his bike to school *tomorrow*' (yesterday)
25. 'Yumm, are we having *lobsters* tonight?' (oysters)

A dissociation of errors with meaning similarities and those with form-based similarities is immediately apparent. Fay and Cutler (1977) in a seminal paper described these contrasts and proposed two selection processes, rooted, respectively, in content (semantic representations) and form (morphological and phonological representations). Depending on which process is compromised, errors should display quite different similarity metrics. Fay and Cutler's analysis, and subsequent studies, revealed multiple systematic form relations between target and error words that lacked a meaning relation – as Examples (11), (21), and (22), while errors with meaning relations, as in Examples (13), (23), and (24), lack such striking phonological resemblances.

These substitutions also interact with sentence structure: Grammatical category correspondence between target and intrusion powerfully constrains word substitutions whether semantically or form driven. This links to the movement error analysis in Sections Accommodations in Error Processes–Generalizing the Observations about Processing Frames. That analysis indicates staged processing in which phonologically uninterpreted elements are first associated with syntactic environments, followed by assignment of phonological representations to structural frameworks. Word substitution patterns of the sort just described therefore fit well into those structural processes. The distinction parallels the discussion of lemmas in Speech Production, Psychology of. Meaning-guided retrieval selects lexical representations (phonologically uninterpreted: lemmas) for input to the multiphrasal construction process that is the locus of word exchanges. Following this, word form retrieval (mapping from lemmas to word forms) provides inputs to the prosodic groups that are the locus of sound exchange errors.

The ordering is clear, but details of the sequencing less so. The differences between meaning and form-based substitutions have been a matter of sustained debate. A way to put this is to ask whether semantic selection is accomplished without feedback from form to meaning levels. Dell (1986) and others have argued that the process is to some degree interactive, and includes such feedback, yielding semantically related errors that also display a degree of form relation, as illustrated in Example (25). Such errors are a minority, but occur with greater than chance likelihood. See Dell et al. (1997) for a relevant review.

What Causes Error?

Several causal factors have been identified. Broadly speaking, they are either 'noise' or 'competition' based. See Butterworth (1982) for a helpful exposition. He distinguishes causes of error in terms of 'sentence internal' and 'sentence external' processes, with emphasis on 'competing plans' as error sources.

Fluctuations in efficiency of the neural systems that underlie language processes could clearly cause error. But such accounts require explicit linkage to production mechanisms. This step is well represented in work by Dell (1986) and others using network models to simulate error patterns; see also Levelt et al. (1999). Error caused by competing sentence formulations is also readily identifiable in blends – as in Examples (13) and (14). These give strong evidence for multiple encoding paths within a sentence. It is easy to generalize such conditions to rationalize certain word substitutions and sound errors as well.

'Sentence external' processes provide additional sources of competing expressions. Such may be discourse-relevant variations on sentence form or discourse external representations for an independent line of thought. This latter may also include interaction of linguistic processing with attentional mechanisms that filter automatic responses to the environment. A 'filtering breakdown' may lead to incorporation of environmental material into speech (see e.g., Harley, 1990).

Sigmund Freud (1901) is identified with accounts of errors motivated by thoughts related to unconscious wishes and desires. His account links those thoughts to the meaning of the words or phrases resulting from error. That requirement suggests that Freudian mechanisms are not a major factor in the generation of movement errors of the sort discussed here. A robust feature of virtually all such errors is their indifference to the meaning of error products (Bierwisch, 1970). An alternative focus for Freud's claims may lie in editorial

processes – i.e., accounts of self-monitoring and its relation to the mechanisms of sentence production (see Motley et al., 1982).

The Broader Picture: Building Models of the Language Production Process

Speech error analysis provides a rich source of information about the real-time processes that underlie human abilities to generate fluent, coherent, oral discourse. But, rich as it is, there are sharp limitations on its utility for certain questions about language production. Issues of semantic and discourse structure are awkwardly represented or absent from error data, and multiple questions about the temporal relations of lexical and phrasal integration are undetermined by the patterns of naturally occurring speech error. In particular, issues surrounding the potential interdependency of the staged processes described here are difficult to assess on error data alone. At precisely this point one turns to experimental studies of production performance.

Bock and Levelt (1994) provide a synthesis of several strands of production experimentation and points of integration with speech error patterns. A number of areas of experimental enquiry have afforded noteworthy elaborations of production theory. Bock (1987, 1990, 1995) and collaborators used picture description tasks and related techniques to uncover detailed relations between sentence generation and the underlying conceptual structures that sentences encode. Models of lexical retrieval have been extensively investigated via picture naming and related paradigms (*see* Speech Production, Psychology of). Experiments that focus on TOT (tip of the tongue) states provide complementary information on lexical staging for syntactic and phonological representations (see, e.g., Badecker et al., 1995; Vigliocco et al., 1997). Studies using error induction techniques also provide points of contact with spontaneous error data. See Shattuck-Hufnagel (1987) for a comprehensive series of error induction studies relating sound error to lexical and phrasal structure, including stress factors. Work by Baars et al. (1975) led to a series of investigations showing lexical influences on the probability of sound exchanges. Some aspects of the experimentation point to monitoring processes as the source of such effects (see Levelt et al., 1999 for review). An extensive experimental literature on induced agreement errors evolved from the original work by Bock and Miller (1991) on English number agreement (see Eberhard et al., 2005 for a synthetic review). Cross-linguistic study of agreement issues has been a significant component of this class of studies (see, e.g., Vigliocco et al., 1996). An expanding area of research activity in the last decade has focused on morphological processes. Such structure is extensively implicated in speech error patterns. For a systematic casting of speech error patterns that addresses morphological issues within a specific formal linguistic framework, see Pfau (2007). There is a substantial range of experimental work exploring effects of morphology in both comprehension and production. See Luttmann et al. (2011) for a recent experimental examination of lexical retrieval processes in production and their interaction with the semantic features of morphologically complex words.

See also: Aphasia; Speech Production, Neural Basis of; Speech Production, Psychology of.

Bibliography

Baars, B., Motley, M., McKay, D., 1975. Output editing for lexical status from artificially elicited slips of the tongue. Journal of Verbal Learning and Verbal Behavior 14, 382–391.

Badecker, W., Miozzo, M., Zanuttini, R., 1995. The two stage model of lexical retrieval: evidence from a case of anomia with selective preservation of grammatical gender. Cognition 57, 193–210.

Bierwisch, M., 1970/1982. Linguistics and language error. In: Cutler, A. (Ed.), Slips of the Tongue. Mouton, Amsterdam.

Bock, J.K., 1987. Coordinating words and syntax in speech plans. In: Ellis (Ed.), Progress in the Psychology of Language, vol. 3. Erlbaum, London, pp. 337–390.

Bock, J.K., 1990. Structure in language: creating form in talk. American Psychologist 45, 1221–1236.

Bock, K., Levelt, W.J.M., 1994. Language production: grammatical encoding. In: Gernsbacher, M.A. (Ed.), Handbook of Psycholinguistics. Academic Press, London, pp. 945–984.

Bock, J.K., Loebell, H., Morey, R., 1992. From conceptual roles to structural relations: bridging the syntactic cleft. Psychological Review 99, 150–171.

Bock, J.K., Miller, C.A., 1991. Broken agreement. Cognitive Psychology 23, 45–93.

Boomer, D., Laver, J., 1968. Slips of the tongue. British Journal of Disorders of Communication 3, 2–12.

Butterworth, B., 1982. Speech errors: old data in search of new theories. In: Cutler, A. (Ed.), Slips of the Tongue. Mouton, Amsterdam.

Bock, K., 1995. Sentence production: from mind to mouth. In: Miller, J.L., Eimas, P.D. (Eds.), Handbook of Perception and Cognition, Speech, Language, and Communication, Vol. 11. Academic Press, Orlando, Fl, pp. 181–216.

Caldognetto, E., Tonelli, L., Vagges, K., Cosi, P., 1987. The organization of constraints on phonological speech errors. In: Proceedings of the XIth International Congress of Phonetic Sciences, vol. 3. U.S.S.R, Tallinn, pp. 50.4.1–50.4.4.

Cutler, A., 1982. The reliability of speech error data. In: Cutler, A. (Ed.), Slips of the Tongue. Mouton, Amsterdam.

Cutler, A., Fay, D., 1978. Introduction (ix–xxxv). In: Versprechen und Verlesen: Eine Psycologisch-Linguistische Studie, New edn. John Benjamins, Amsterdam.

del Viso, S., Igoa, J.M., García-Albea, J.E., 1991. On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. Journal of Psycholinguistic Research 20, 161–185.

Dell, G., 1984. The representation of serial order in speech: evidence from the repeated phoneme effect in speech errors. Journal of Experimental Psychology, Learning Memory & Cognition 10, 222–233.

Dell, G., 1986. A spreading activation theory of retrieval in sentence production. Psychological Review 93, 283–321.

Dell, G., 1991. Effects of frequency and vocabulary type on phonological speech errors. Language and Cognitive Processes 5, 313–349.

Dell, G., Reich, P., 1981. Stages in sentence production: an analysis of speech error data. Journal of Verbal Learning and Verbal Behavior 20, 611–629.

Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., Gagnon, D.A., 1997. Lexical access in aphasic and nonaphasic speakers. Psychological Review 104 (4), 801–838.

Eberhard, K.M., Cutting, John C., Kathryn Bock, J., 2005. Making syntax of sense: number agreement in sentence production. Psychological Review 112, 531–559.

Fay, D., Cutler, A., 1977. Malapropisms and the structure of the mental lexicon. Linguistic Inquiry 8, 505–520.

Freud, S., 1901. The psychopathology of everyday life. Monatschrift fur Psychiatrie und Neurologie 10, 1–13.

Fromkin, V., 1971. The non-anomalous nature of anomalous utterances. Language 47, 27–52.

Garrett, M.F., 1975. The analysis of sentence production. In: Bower, G. (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 9. Academic Press, New York.

Garrett, M., 1980. Levels of processing in sentence production. In: Butterworth, B. (Ed.), Language Production, Speech and Talk, vol. 1. Academic Press, London, pp. 177–220.

Garrett, M., 1993. Errors and their relevance for models of language production. In: Dittman, J., Blanken, G. (Eds.), Handbooks of Linguistics and Communication Science: Linguistic Disorders and Pathologies. Walter De Gruyter, New York.

Harley, T., 1990. Environmental contamination of normal speech. Applied Psycholinguistics 11, 45–72.

Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Levelt, W.J.M., Roelofs, A., Meyer, A., 1999. A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1–75.

Luttmann, H., Zwitserlood, P., Bölte, J., 2011. Sharing morphemes without sharing meaning: production and comprehension of German verbs in the context of morphological relatives. Canadian Journal of Experimental Psychology 65 (3), 173–191.

McKay, D.G., 1970. Spoonerisms: the structure of error in the serial order of speech. Neuropsychologia 8, 323–350.

McKay, D.G., 1972. The structure of words and syllables: evidence from errors in speech. Cognitive Psychology 3, 210–227.

Meringer, R., Meyer, K., 1895/1978. In: Cutler, A., Fay, D. (Eds.), Versprechen und Verlesen: Eine Psychologisch Linguistische Studie. Goschensche Verlag, Stuttgart. John Benjamins, Amsterdam.

Myers-Scotton, C., Jake, J., 2001. Explaining aspects of code switching and their implications. In: Nicol, J. (Ed.), One Mind, Two Languages: Bilingual Language Processing. Blackwell, Oxford, pp. 91–125.

Motley, M., Camden, C., Baars, B., 1982. Covert formulation and editing of anomalies in speech production. Journal of Verbal Learning and Verbal Behavior 21, 578–594.

Nooteboom, S.G., 1969. The tongue slips into patterns. In: Sciarone, A., et al. (Eds.), Nomen, Leyden Studies in Linguistics and Phonetics. Mouton, The Hague, pp. 114–132.

Pfau, R., 2007. Cheap repairs: a distributed morphology toolkit for sentence construction. In: Schütze, C., Ferreira, V. (Eds.), The State of the Art in Speech Error Research. MIT Working Papers in Linguistics, #53.

Robbins, R.H., 1966. The warden's wordplay: toward a redefinition of the spoonerism. Dalhouse Review 46, 457–465.

Shattuck-Hufnagel, S., 1987. The role of word onset consonants in speech production planning: new evidence from speech error patterns. In: Kellar, E., Gopnik, M. (Eds.), Motor and Sensory Processing in Language. Erlbaum Press, Hillsdale, NJ.

Stemberger, J., 1985. An interactive activation model of language production. In: Ellis, A. (Ed.), Progress in the Psychology of Language, vol. 1. Erlbaum, London.

Stemberger, J., McWhinney, B., 1986. Form-oriented inflectional errors in language processing. Cognitive Psychology 18, 329–354.

Vigliocco, G., Antonini, T., Garrett, M., 1997. Grammatical gender is on the tip of Italian tongues. Psychological Science 8, 314–317.

Vigliocco, G., Hartsuiker, R.J., Jarema, G., Kolk, H.H.J., 1996. One or more labels on the bottles? Notional concord in Dutch and French. Language & Cognitive Processes 11, 407–442.

Vousden, J.I., Brown, G.D.A., Harley, T.A., 2000. Serial control of phonology in speech production: a hierarchical model. Cognitive Psychology 41, 101–175.

Speech Perception

Dominic W Massaro, University of California Santa Cruz, Santa Cruz, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This psychological account of speech perception includes a description of the stimulus variables that support it and the internal processes that account for it. Research and theory indicate that speech perception is a form of pattern recognition that is influenced by multiple sources of bottom-up and top-down information.

Speech Perception warrants an entry in this encyclopedia along with Motion Perception and Music Perception. One possible implication of a separate entry for Speech Perception is that it differs significantly from other types of perceptions. In fact, much of the six or seven decades of research and theory centered on this topic has assumed that speech is special. If speech is special, then how it is perceived must also be special. In many ways, the specialty assumption is congruent with the more prominent belief that language and how it is acquired is special.

Speech Is Special

A highly influential proposal by Noam Chomsky envisioned language ability as dependent on an independent language organ (or module), analogous to other organs such as our digestive system. This organ follows an independent course of development in the first years of life and allows the child to achieve a language competence that cannot be elucidated in terms of traditional perception and learning theory. This mental organ, responsible for the human language faculty and our language competence, matures and develops with experience, but the matured system does not simply mirror this experience. The language user arrives on this planet with rule systems of a highly specific structure capable of acquiring languages that necessarily have a structure corresponding to these rules. This innate knowledge is necessary to become an expert in using language, even though the typical child has a paucity of the language input. The data of language experience are so limited that no learning process of induction, abstraction, generalization, analogy, or association could account for our observed language competence. Somehow, the universal grammar given by our biological endowment allows us to learn to use language appropriately without learning many of the formal intricacies of the language. This theory is highly controversial and there are many reasons to remain skeptical. As counterevidence, psychologists are demonstrating that infants' language is highly influenced by typical perception and learning processes, and linguists are documenting that the child's language input is not as sparse as the nativists claim (Scholz and Pullum, 2006).

Luckily for our purposes, we can evaluate the claim that speech is special without resolving the language is special controversy. There are several aspects of speech that appeared to demand specialty status. Perception researchers typically expect an orderly relationship between a stimulus and its resulting percept. The loudness of a tone increases as it is increased in intensity, for example, and two different intensities would give different percepts. However, the speech percept appeared to be characterized by a non-invariant relationship between it and the eliciting stimulus. An historical example gives the surprising perception of consonant–vowel (CV) syllables /di/ as in deed and /du/ as in do. We hear the phoneme (Smallest segment of spoken language that is functional in communicating meaning.) /d/ in both cases, but the speech signal differs significantly in the two cases. Of course, not every scientist was convinced that phonemes did not have invariant signal properties in different vowel contexts, and research in the six or so decades since the original claim has continued to attempt to uncover these properties (e.g., Kapoor and Allen, 2012). Even so, no one has yet provided convincing evidence of invariance.

Faced with this conundrum, researchers looked for a solution outside of the normal framework of perception in which the goal is to describe the stimulus characteristics responsible for perceptual outcomes. An obvious alternative was to believe that a motor theory could reinstate the regularity between signal and percept. A motor theory of speech perception would assume that we somehow perceive the vocal tract gestures creating the sounds rather than the sounds themselves. As recently observed by Hickok, motor theories have a long history in behavioral science. Over 100 years ago, Walter B. Pillsbury (1911: p. 84) said "The [motor] theory is so simple and so easy to present that everyone is glad to believe it. The only question that any one cares to raise is how much of it will the known facts permit one to accept." It is sobering how little has changed in the intervening century given that many scientists have not been swayed by Pillsbury's caveat. Speech provided a natural domain for a motor theory account because speech perception and speech production appeared to be so tightly linked. The perceiver simply perceives the intended phonetic gestures of the speaker rather than the variable auditory speech. For our /di/-/du/ example, the perceiver simply recovers the interlocutor's intention to articulate /d/ in both syllables. For most of us, it is difficult to intuit how the motor theory accounts for speech perception and how it resolves the lack of invariance problem. How do we access the talker's intended articulations? (Coarticulation: Pronunciation (articulation) of a speech segment is influenced by its neighboring segments.) If we could see the talker, we would obtain a few hints, and in fact observing the talker dramatically influences what we hear. The normal view is not sufficient, however,

because much of the articulation is hidden inside the mouth. More telling is the common observation of successful phone conversations when the talkers are not visible.

What's most disappointing about motor theory is that its advocates do not offer specific testable hypotheses about how it accomplishes speech perception. Most of the explanations have taken the form of analysis by synthesis models. Given a speech event, the perceiver benefits from synthesizing possible alternatives during their processing of the speech input and somehow determining which of these possibilities is the best match. To accomplish this selection, however, the perceiver still requires access to the speech signal, and hence we have regressed back to the original challenge of understanding how the speech signal is processed.

In one of the few explicit accounts, Vihman (2002) describes how motor processes might work in speech acquisition. The infant practices canonical babbling and produces CV sequences at 6–8 months of age. This practice in production sensitizes the infant to similar input speech from their caregivers. The familiarity of CV sequences from babbling allows them to be easily recognized because they pop out of the acoustic stream. However, although these patterns would become familiar with practice and this increase in familiarity might facilitate perception, it does not mean that the motor processes involved in babbling were also functional during the infant's speech perception.

In fact, infant speech perception is much more sophisticated than what could be predicted from canonical babbling. We know that receptive language is acquired before productive language, so it is difficult to understand how motor behavior would contribute to speech perception. Motor theory does not solve the problem of speech perception. Infants understand speech well before they produce it intelligibly. Furthermore, infants are highly sensitive to the statistical properties of segmental speech input at 6–8 months (Saffran, 2003), which could not be anticipated by canonical babbling.

Roy (2011) documents 67 instances in which his son attempted to pronounce 'water' before he was able to pronounce it correctly. Many of these instances illustrate that he was able to perceive and understand the spoken word but was simply unable to produce it accurately. This scientific observation revives the anecdotal one in which a father is mimicking his son's mispronunciation of the word 'rabbit.' His son corrects his dad saying, "No dad, not rawwit, but rawwit." The son clearly could perceive the difference between 'rabbit' and 'rawwit' even though he was not able to produce the difference. Toddlers (13–15-month-olds) have few words in their productive vocabulary but can compute the relations in a spoken sentence (Hirsh-Pasek and Golinkoff, 1996). Hickok (2009) also reviews long-standing clinical findings that falsify motor theory. For example, persons with Broca's aphasia can perceive and understand speech but cannot produce it. Thus, research on language acquisition like the research and clinical findings with adults does not support motor theory.

The astute reader will notice that we have not solved the invariance problem that motivated a motor theory account. To solve the invariance problem between acoustic signal and phoneme, Massaro (1972) proposed the syllables V, CV, or VC as perceptual units in speech, where V is a vowel and C is a consonant or consonant cluster (unit of speech that is functional in speech perception). This might appear as too easy of a solution because one might argue that the bigger the unit the better. So why not assume a word, a phrase, or even a sentence as the perceptual unit? Luckily, there were other constraints on the size of the unit. The speech representing the perceptual unit must be maintained in a pre-perceptual auditory memory in order to be functional for recognition (Massaro, 1972). Backward masking experiments indicated that pre-perceptual auditory memory had a limit of roughly 250 ms, which can hold these syllables but not longer units (this assumption was built into the foundation of the model discussed in Section The Fuzzy Logical Model of Perception). Assuming that this larger segment is the perceptual unit that reinstates a significant amount of invariance between signal and percept, Massaro and Oden (1980: pp. 133–135) reviewed evidence that the major coarticulatory influences (that are responsible for the /di/-/du/ differences) on perception occur within these syllables, rather than between the syllables. Any remaining lack of invariance across these syllables could conceivably be disambiguated by additional sources of information in the speech stream.

Popular theories, such as the motor theory, will be reinvented and will evidently endure well beyond their many falsifications as we will discuss in the next section.

Mirror Neurons

The discovery of mirror neurons has apparently rejuvenated motor theories. A mirror neuron fires both when an animal performs an action and when the animal observes the same action performed by another animal (Rizzolatti and Craighero, 2004). Mirror neurons could serve as the basis for imitation and therefore for learning. Our understanding, however, is that mirror neurons cannot account for perception because they would overgeneralize. The macaque certainly experiences the difference between seeing a conspecific action and its own action, but the same mirror neurons are putatively activated by these very different events and experiences. If only the mirror neurons were responsible for speech perception, the interlocutor would not distinguish between her perception of the utterance from her production of the utterance. Thus, mirror neurons alone cannot account for perception (see Hickok, 2009; Lotto et al., 2009 for similar observations and additional critiques).

Psychophysics of Speech Perception

In any domain of perception, one goal is to determine the stimulus properties responsible for perception and recognition of the objects in that domain. The study of speech perception promises to be even more challenging than other domains of perception because there appears to be a discrepancy between the stimulus and the perceiver's experience of it. For speech, we perceive mostly a discrete auditory message composed of words, phrases, and sentences. The stimulus input for this experience, however, is a continuous stream of sound (and facial and gestural movements in face-to-face communication) produced by the speech production process. Somehow, this continuous

input is transformed into more or less a meaningful sequence of discrete events.

In a seminal paper, Miller and Nicely (1955) were interested in the acoustic properties of speech that were important for the recognition of English consonants. They asked participants to identify 16 initial consonants placed before the vowel /a/ as in father. Degradation of performance was measured under varying conditions of noise and filtering. The standard measure of performance is in terms of a confusion matrix which shows each of the responses to each of the test stimuli. One goal of this research is to determine which acoustic properties best explain the confusion matrix. The authors chose linguistic descriptors as rough indices of the acoustic properties. These included voicing (low frequency periodic energy), nasality (spectral changes caused by energy flow through the nasal tract), frication (high frequency energy), place of articulation (formant location and transitions), and duration. Although these properties were treated as binary (or trinary) linguistic dimensions, we now know that a more accurate description would be in terms of a continuous value (Massaro and Cohen, 1999). One disadvantage of these studies is that the speech signal is necessarily degraded in order to generate errors. The results may not generalize to typical situations in which the speech is not degraded.

Greenberg and his colleagues have pursued Miller and Nicely's original quest for understanding how the various acoustic properties representing linguistic features contribute to speech perception. They find that not all linguistic features are created equal. Based on systematic studies of confusion matrices in Danish, they find that the ease of resolving three features is ordered voicing, manner, and place, respectively. They also find that resolving place of articulation is highly correlated with correct consonant identification. They interpret this result to mean that resolving place of articulation is central to speech recognition. Allen (e.g., Kapoor and Allen, 2012) and his colleagues have provided some innovative techniques to assess which acoustic properties are important for speech recognition. They systematically manipulate the natural speech signal to emphasize or degrade an acoustic feature. They systematically study the burst of energy at the onset of stop consonants in the initial position. Based on their results, they conclude that the frequency range and relative intensity of the plosive burst and its temporal relationship to the sonorance of the following vowel are sufficient features for correct identification of the four stop consonants /da/, /ga/, /ta/, and /ka/. In an earlier study, they found that /ba/ is characterized by a broad band of energy in the burst. If this information is missing, the segment is confused for the consonant /v/.

The auditory speech signal is distributed across a frequency spectrum and it has been traditional to treat this spectrum as composed of independent frequency bands. These bands are putatively independent sources of information for intelligibility of the speech signal, which are important for the accepted measures, namely Articulation Index, Speech Intelligibility Index, and the Speech Transmission Index. However, this assumption is untenable as demonstrated nicely by the research of Christiansen and Greenberg (2012).

Although six decades later, there is still active controversy over the ecological properties of the speech input that are actually functional in speech perception. One issue, revived by recent findings, is whether the functional properties in the signal are static or dynamic (changing with time). Traditionally, static cues such as the location of formants (bands of energy in the acoustic signal related to vocal tract configuration), the distribution of spectral noise as in the onset of *saw* and *shawl*, (and the mouth shape at the onset of a segment) have been shown to be effective in influencing speech perception. Dynamic cues such as the transition of energy between a consonant and the following vowel have also been shown to be important. For example, research has shown that the second formant (F2) transition defined as the change between the F2 value at the onset of a CV transition and the F2 value in the middle of the following vowel is a reliable predictor of the place of articulation category (Sussman et al., 1998).

Controversy arises when research is carried out to argue for one type of cue rather than another. For example, investigators isolated short segments of the speech signal and reversed the order of the speech within each segment (Saberi and Perrott, 1999). In this procedure, a sentence is divided into a sequence of successive segments of a fixed duration such as 50 ms. Each segment is time reversed and these new segments are recombined in their original order, without smoothing the transition borders between the segments. In this fashion, the sentence could be described as globally contiguous but locally time-reversed. The authors claimed that the speech was still intelligible when the reversed segments were relatively short (1/20th to 1/15th of a second). Their conclusion was that our perception of speech was demonstrated to be primarily dependent on higher-order dynamic properties rather than the short static cues normally assumed by most current theories. This type of study and logic follows a tradition of attempting to find a single explanation or influence of some psychological phenomenon. However, most successful research in psychology is better framed within the more general framework of *ceteris paribus* (all other aspects neutral). There is good evidence that perceivers exploit many different cues in speech perception (e.g., Greenberg and Arai, 2004), and attempting to isolate a single functionally sufficient cue is futile.

There is now a large body of evidence indicating that multiple sources of information are available to support the perception, identification, and interpretation of spoken language. There is an ideal experimental paradigm that allows us to determine which of the many potentially functional cues are actually used by human observers, and how these cues are combined to achieve speech perception (Massaro, 1998). The systematic variation of the properties of the speech signal and quantitative tests of models of speech perception allow the investigator to interpret the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro, 1998). Thus, this research strategy addresses how different sources of information are evaluated and integrated, and can identify the sources of information that are actually used.

The Demise of Categorical Perception

Too often behavioral scientists allow their phenomenal impressions to spill over into their theoretical constructs. One

experience in speech perception is that of categorical perception (Situation in which a speech stimulus is perceived in terms of its category membership rather than in terms of its surface properties.) meaning that we perceive an utterance as a specific syllable, word, or phrase. This phenomenal outcome is easily interpreted as a perceptual process that is categorical. This means that the perceptual system only has access to categorical information about the speech event rather than access to continuous information about the degree to which various events are possible. My experience, particularly now with an increasing hearing loss of high frequencies, is just the opposite. In many instances, I am stumped with a percept that fits several possible alternatives rather than one categorically. This experience is to be expected, given a large body of empirical research that reveals that speech perception is not categorical.

Using synthetic speech, it is possible to make a continuum of different speech sounds varying in small steps between two alternatives. Categorical perception has been a central concept in the experimental and theoretical investigation of speech perception and has also spilled over into other domains such as face processing (Beale and Keil, 1995; Etcoff and McGee, 1992). Categorical perception was operationalized in terms of discrimination performance being limited by identification performance. In a very influential study, Liberman et al. (1957) used synthetic speech to generate a series of 14 CV syllables going from /be/ to /de/ to /ge/ (/e/ as in gate). The onset frequency of the second formant transition of the initial consonant was changed in equal steps to produce the continuum. In the identification task, observers identified random presentations of the sounds as /b/, /d/, or /g/. The discrimination task used the ABX paradigm. Three stimuli were presented in the order ABX; A and B always differed and X was identical to either A or B. Observers were instructed to indicate whether X was equal to A or B. This judgment was supposedly based on auditory discrimination in that observers were instructed to use whatever auditory differences they could perceive.

The experiment was designed to test the hypothesis that listeners can discriminate the syllables only to the extent that they can recognize them as different phoneme categories. Although this categorical perception hypothesis can be quantified in order to predict discrimination performance from the identification judgments (Massaro, 1987), only a qualitative analysis was carried out. A rough correspondence between identification and discrimination led the authors to incorrectly conclude that discrimination performance was fairly well-predicted by identification performance. Unfortunately, there are several notable limitations in these and many similar studies. Discrimination performance is consistently better than that predicted by identification; it is possible that participants are making their discrimination judgments on the basis of identification rather than on their auditory discrimination, and it is likely that any alternative theory would have described the results equally well (Massaro, 1987).

In many areas of inquiry, a new experimental paradigm enlightens our understanding by helping to resolve theoretical controversies. Following developments in signal detection theory, we used rating experiments to determine if perceivers indeed have information about the degree of category membership (Massaro and Cohen, 1983). Rather than being asked for categorical decisions, perceivers were asked to rate the stimulus along a continuum between categories. A detailed quantitative analysis of the results indicated that perceivers have reliable information about the degree of category membership. Although communication forces us to partition the inputs into discrete categories for understanding, this property in no way implies that speech perception is categorical. To retrieve a toy upon request, a child might have to decide between *ball* and *doll*; however, he/she can certainly have information about the degree to which each toy was requested.

Although categorical perception has been discredited, it is often reinvented under new guises. The hypothesis of a 'perceptual magnet' is that prototypical instances of a speech category function like magnets for surrounding sounds that are less prototypical members of that category, creating sharper boundaries between categories (Kuhl, 1991, 2004). The perceptual magnet putatively creates representations which produce good discriminability of speech segments near category boundaries and poor discriminability near the category prototype. (Ideal or central member of a category or concept: Robins are prototypical birds, whereas eagles and penguins are not). Thus, the perceptual magnet effect a reformulation of categorical perception. Evidence for the perceptual magnet requires an experiment to show how discrimination is directly predicted by a measure of category goodness. Lotto et al. (1998) found, however, that discriminability was *not* poorer for vowels with high category goodness, in contrast to the predictions of the perceptual magnet theory. The same authors also observed additional negative evidence that category goodness ratings were highly context-sensitive. If category goodness is functional in discrimination, it should be relatively stable across different contexts. (Prototype Description: A summary description of a category in terms of a representation for the ideal member of the category.)

Some proponents of categorical perception are also skeptical of the view that speech is special. As evidence, they cite several experimental studies that demonstrate that nonhuman animals also produce categorical perception. The belief is that these studies reveal that spoken language capitalized on natural auditory discontinuities for defining speech categories. Perhaps the most telling negative evidence is that the 6000 or so languages of the world have carved up the speech stimulus space in very different ways, precluding this possibility (Lindau and Ladefoged, 1983). My belief is that there are too few possible discontinuities for just a single language, let alone a plethora of languages. Most importantly, the 'categorical perception' results in nonhuman animals, like the corresponding results with humans, can be equally or better explained in terms of continuous perception.

It was common to attribute categorical perception to infants as well as adults (Eimas, 1985; Gleitman and Wanner, 1982). Although early studies appeared to find that infants noticed differences only between sounds from different speech categories and not between sounds from within the same speech category, follow-up studies quickly demonstrated that infants discriminate differences within, as well as between, categories (Massaro, 1987: p. 239; McMurray and Aslin, 2005). Other infant studies revealed how rapidly infants develop an affinity for their language. Infants have been shown to behave differently to two different speech segments at 4 months of age but not at 14 months, if the two segments are not informative

in their language (Werker and Tees, 1983). This ability to quickly learn the meaningful distinctions in their language is not unique to speech. Research in sign language has revealed exactly parallel results. Infants raised with American Sign Language show a sensitivity to different handshapes at 4 months of age but do not behave differently to them at 14 months, if they are not functional (Baker et al., 2006).

More generally, research with infants reveals that they discriminate the multiple dimensions of the auditory speech signal. However, the meaning of these differences in the language must be learned, and infants are not prewired to categorize the signals into innate phonetic categories. It is just as false to attribute categorical perception to the infant and child as it is to claim that fully developed adults are categorical perceivers (Massaro, 1987).

Although categorical perception is about as old as speech research itself, we must conclude that speech is perceived continuously. Speech decisions are necessarily categorical. Research reveals conclusively that audible and visible sources of bottom-up information and top-down sources of contextual information generate continuous representations. As discussed in Section The Fuzzy Logical Model of Perception, these representations are integrated to provide a usually robust interpretation of the language input (Massaro, 1987, 1998). Most importantly, the case for the specialization of speech is weakened considerably because of the central role that the assumption of categorical perception has played (Liberman and Mattingly, 1985). Finally, several neural network theories such as single-layer perceptrons, recurrent network models, and interactive activation have been developed to predict categorical perception (Damper, 1994): its nonexistence poses great problems for these models.

Speech Perception as Pattern Recognition

The study of speech perception has matured into an interdisciplinary endeavor, which involves a varied set of experimental and theoretical approaches. It includes the fundamental psychophysical question of which properties of spoken language are perceptually meaningful and how these properties signal the message. Independent variation of several properties, along with a quantitative theoretical analysis, is a productive paradigm to pursue not only the psychophysical question but also the issue of how the multiple cues are used together for perception and understanding. We have learned that spoken language understanding is influenced by multiple sources of information from several modalities and that perceivers derive continuous information from these many sources. In addition to these bottom-up sources, higher-order context is also used in speech understanding. Several productive theoretical approaches address the complex question of how we so easily understand another's utterances.

It is commonplace to have the impression that foreign languages are spoken much more rapidly than our own, and without silent periods between the words and sentences. Our own language, however, is perceived at a normal pace (or even too slowly at times) with clear periods of silence between the words and sentences. In fact, languages are spoken at rates that can differ by about a factor of two (Pellegrino et al., 2011). The reason for these different rates appears to be due to the information density of the speech: some languages are denser than others, and it appears that all languages are spoken to have about the same density. If your language is less dense, you speak more quickly so your listener stays attentive. Even with these differences, foreign languages give the impression that they are being spoken more quickly. Our troubling experience with a foreign language is simply that we do not know it. These experienced differences with known and unknown languages are solely due to the memory structures and psychological processes involved in speech perception.

We define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input. The empirical and theoretical investigation of speech perception has blossomed into an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics.

Given the existence of multiple sources of information in speech perception, each perceived continuously, a new type of theory is needed. The theory must describe how each of the many sources of information is evaluated, how these sources are combined or integrated, and how decisions are made. The development of a promising theory has evolved from sophisticated experimental designs and quantitative model testing to understand speech perception and pattern recognition more generally. A wide variety of results have been accurately described within the Fuzzy Logical Model of Perception (FLMP). (A logic based on the premise that propositions can be graded in truth value, rather than simply being true or false. Degree of truth is represented by a truth value between 0 (perfectly false) and 1 (perfectly true). Fuzzy logic also gives algorithms for computing the negation, conjunction, and disjunction of continuous truth values.)

The Fuzzy Logical Model of Perception

The three processes involved in perceptual recognition are illustrated in Figure 1 and include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: (1) Each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives; (2) the sources of information are evaluated independently of one another; (3) the sources are integrated to provide an overall continuous degree of support for each alternative; and (4) perceptual identification and interpretation follows the relative degree of support among the alternatives.

Given multiple sources of information, it is useful to have a common metric representing the degree of match of each feature. Two features that define a prototype can be related to

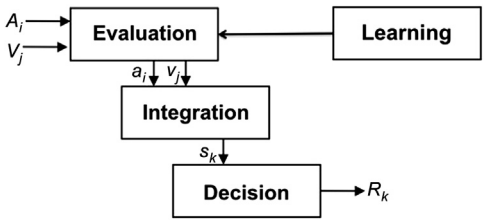

Figure 1 Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The different sources of information are represented by upper-case letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

one another more easily if they share a common currency. To serve this purpose, fuzzy-truth values (Goguen, 1969; Zadeh, 1965) are used because they provide a natural representation of the degree of match. Fuzzy-truth values lie between 0 and 1, corresponding to a proposition being completely false and completely true. The value 0.5 corresponds to a completely ambiguous situation, whereas 0.7 would be more true than false and so on. Fuzzy-truth values, therefore, not only can represent continuous rather than just categorical information, but also can represent different kinds of information. The truth values in speech perception would correspond to the likelihood of a speech category given the sensory information. The variability of the sensory information signaling a specific speech category would also influence the truth value (Bejjanki et al., 2011). (Mental Representation: The mental content of perceptions, ideas, images, beliefs, thoughts, memories, and hypotheses. These are symbols because they stand for something else. There is no reason that the representation or symbol has to be discrete).

Figure 1 also illustrates how learning is conceptualized within the model by specifying exactly how the feature values used at evaluation change with experience. Learning in the FLMP can be described by the following algorithm (Friedman et al., 1995; Kitzis et al., 1999). The initial feature value representing the support for an alternative is initially set to 0.5 (since 0.5 is neutral in fuzzy logic). A learning trial consists of a feature (such as closed lips at onset) occurring in a test item followed by informative feedback (such as the syllable /ba/). After each trial, the feature values would be updated according to the feedback, as illustrated in Figure 1. Thus, the perceiver uses the feedback to modify the prototype representations and these in turn will become better tuned to the informative characteristics of the patterns being identified. This algorithm is highly similar to many contemporary views of language acquisition (Best and McRoberts, 2003; Best et al., 2001; Werker and Logan, 1985)

In the course of our research, we have found the FLMP to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition. People are influenced by multiple sources of information in a diverse set of situations. In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation.

Multimodal Speech Perception

Speech perception has traditionally been viewed as a unimodal process, but in fact appears to be a prototypical case of multimodal perception. This is best seen in face-to-face communication. Experiments have revealed conclusively that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro, 1998). Although the results demonstrate that perceivers use both auditory and visible speech in perception, they do not indicate how the two sources are used together. There are many possible ways the two sources might be used. We first consider the predictions of the FLMP.

In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to the feature value. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$, is equal to.

$$P(/\mathrm{da}/) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)} \qquad [1]$$

In the previous work, the FLMP has been contrasted against several alternative models such as a weighted averaging model (WTAV), which is an inefficient algorithm for combining the auditory and visual sources. For bimodal trials, the predicted probability of a response, $P(/\mathrm{da}/)$ is equal to.

$$P(/\mathrm{da}/) = \frac{w_1 a_1 + w_2 v_j}{w_1 + w_2} = w a_i + (1 - w) v_j \qquad [2]$$

The WTAV predicts that two sources can never be more informative than one. In direct contrasts, the FLMP has consistently and significantly outperformed the WTAV (Massaro, 1998).

More generally, research has shown that the results are well-described by the FLMP, an optimal integration of the two sources of information (Massaro and Stork, 1998). A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro and Stork, 1998).

Recent findings show that speech reading, or the ability to obtain speech information from the face, is not compromised by oblique views, partial obstruction, or visual distance. Humans are fairly good at speech reading even if they are not

looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998).

Higher-Order or Top-Down Influences

William Chandler Bagley published perhaps the first experimental study of speech perception in his dissertation in 1890. In Bagley's experiment, naturally spoken words were recorded and played back to eight members of the Cornell University Psychology Department on Edison phonograph cylinders. The words were pronounced by deleting one of the consonant sounds in the word. Bagley evaluated the effects of context on word perception: The word was presented alone, with one or two related words, and at the beginning, middle, and end of complete sentences. This experiment may have been the first systematic distortion of the spoken message, although people doubtless have been playing with the words they utter since they began uttering them (witness Pig Latin, for example). The results demonstrated that subjects were often able to correctly recognize the distorted words and, in addition, the word's context word recognition. Correct word recognition was better if the word was placed in the middle of a sentence, for example, relative to being presented alone.

Consistent with the framework being developed, there is now a substantial body of research illustrating that speech perception is influenced by a variety of contextual sources of information. Bottom-up sources correspond to those sources that have a direct mapping between the sensory input and the representational unit in question. Top-down sources or contextual information come from constraints that are not directly mapped onto the unit in question. As an example, a bottom-up source would be the stimulus presentation of a test word after the presentation of a top-down source, a sentence context. A critical question for both integration and autonomous (modularity) models is how bottom-up and top-down sources of information work together to achieve word recognition. For example, an important question is how early can contextual information be integrated with acoustic-phonetic information. A large body of research shows that several bottom-up sources are evaluated in parallel and integrated to achieve recognition (Massaro, 1987, 1994). An important question is whether top-down and bottom-up sources are processed in the same manner. A critical characteristic of autonomous models might be described as the language user's 'inability' to integrate bottom-up and top-down information. An autonomous model must necessarily predict no perceptual integration of top-down with bottom-up information.

The model tests have established that perceivers integrate top-down and bottom-up information in language processing, as described by the FLMP. This result means that sensory information and context are integrated in the same manner as several sources of bottom-up information. These results pose problems for autonomous models of language processing, which assume that perceptual outcomes are impenetrable by context effects. Evidence of multiple sources of information and perceptual influences of context effects validates our critique of categorical perception. If the auditory input constrained the perception to a single speech category, then context could have no productive influence. When the context agreed with the speech category, there would be no benefit. When context disagreed with the categorization, the perceiver would be at a loss as to what to do. Having continuous information from both bottom-up and top-down sources allows for their productive integration.

It is important to mention that the FLMP have been proven to be mathematically equivalent to Bayes theorem, which is an optimal algorithm for combining two information sources (Massaro, 1987; Massaro and Friedman, 1990). However, we prefer the common metric to be fuzzy-truth values because they more naturally represent continuous information than do the probabilities of Bayes theorem. If language processing evolved across many millennia of language use, it should not be surprising that speech communication settled on an optimal situation. The many experimental and theoretical sources of evidence for the FLMP thus in turn supports a Bayesian analysis of speech perception, which invites it within an acceptable general framework of perception and learning.

Neurological Underpinnings of Speech Perception

This article is dedicated to a psychological level of description of speech perception. Of course, there is a neurological level that can be equally informative (Massaro, 1989). Many researchers are actively involved in discovering what brain mechanisms are involved in speech perception and how they work together to allow speech understanding and to support speech production. Speech perception or recognition is distinguished from speech production in a recently proposed dual-stream model (Hickok and Poeppel, 2007). A ventral stream supports perception and a dorsal stream interfaces with articulatory networks to allow speech production. A few assumptions of the model that find support at the neurological level are relevant to our review of speech perception. First, there are brain processes that allow the influence of multiple sources of information on speech perception. Second, the model rejects outright that motor behavior intervenes in speech perception (Hickok et al., 2009). Both these assumptions at the neurological level are consistent with our conclusions at the psychological level.

Bibliography

Baker, S.A., Michnick-Golinkoff, R., Petitto, L.A., 2006. New insights into old puzzles from infants' categorical discrimination of soundless phonetic units. Language Learning and Development 2, 147–162.

Beale, J.M., Keil, F.C., 1995. Categorical effects in the perception of faces. Cognition 57 (3), 217–239.

Bejjanki, V.R., Clayards, M., Knill, D.C., Aslin, R.N., 2011. Cue integration in categorical tasks: insights from audio-visual speech perception. PLoS One 6, e19812. http://dx.doi.org/10.1371/journal.pone.0019812.

Best, C., McRoberts, G., Goodell, E., 2001. American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology. Journal of the Acoustical Society of America 1097, 775–794.

Best, C.T., McRoberts, G.W., 2003. Infant perception of non-native consonant contrasts that adults assimilate in different ways. Language and Speech 46, 183–216.

Christiansen, T.U.C., Greenberg, S., 2012. Perceptual confusions among consonants, revisited – cross-spectral integration of phonetic-feature information and consonant recognition. IEEE Transactions on Audio, Speech, and Language Processing 20, 147–161.

Damper, R.I., 1994. Connectionist models of categorical perception of speech. In: Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks, vol. 1, pp. 101–104.

Etcoff, N.L., McGee, J.J., 1992. Categorical perception of facial expressions. Cognition 44, 227–240.

Eimas, P.D., 1985. The perception of speech in early infancy. Scientific American 252, 46–52.

Fitch, H.L., Halwes, T., Erickson, D.M., Liberman, A.M., 1980. Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception & Psychophysics 27, 343–350.

Friedman, D., Massaro, D.W., Kitzis, S.N., Cohen, M.M., 1995. A comparison of learning models. Journal of Mathematical Psychology 39, 164–178.

Ganong III, W.F., 1980. Phonetic categorization in auditory word recognition. Journal of Experimental Psychology: Human Perception and Performance 6, 110–125.

Gleitman, L.R., Wanner, E., 1982. Language acquisition: the state of the state of the art. In: Wanner, E., Gleitman, L.R. (Eds.), Language Acquisition: The State of the Art. Cambridge University Press, Cambridge, UK.

Goguen, J.A., 1969. The logic of inexact concepts. Synthese 19, 325–373.

Greenberg, S., Arai, T., 2004. What are the essential cues for understanding spoken language? IEICE Transactions on Information and Systems E87, 1059–1070.

Hickok, G., 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. Journal of Cognitive Neuroscience 21, 1229–1243.

Hickok, G., Holt, L.L., Lotto, A.J., 2009. Response to Wilson: what does motor cortex contribute to speech perception? Trends in Cognitive Sciences 13, 330–331.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402.

Hirsh-Pasek, K., Golinkoff, R.M., 1996. The Origins of Grammar. Evidence from early language comprehension. MIT Press, Cambridge, MA.

Kapoor, A., Allen, J.B., 2012. Perceptual effects of plosive featur modification. Journal of the Acoustical Society of America 131 (1), 478–491.

Kuhl, P.K., 1991. Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. Perception & Psychophysics 50, 93–107.

Kuhl, P.K., 2004. Early language acquisition: Cracking the speech code. Nature Reviews Neuroscience 5, 831–843.

Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C., 1957. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54, 358–368, 753–771.

Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.

Lindau, M., Ladefoged, P., 1983. Variability of feature specifications. In: Perkell, J., Klatt, D. (Eds.), Invariance and Variability in Speech Processes. Lawrence Erlbaum Associates, Inc., New Jersey, pp. 464–479.

Lotto, A.L., Hickok, G., Holt, L.L., 2009. Reflections on mirror neurons and speech perception. Trends in Cognitive Sciences 13, 110–114.

Lotto, A.J., Kluender, K.R., Holt, L.L., 1998. Depolarizing the perceptual magnet effect. Journal of the Acoustical Society of America 103, 3648–3655.

McMurray, B., Aslin, R.N., 2005. Infants are sensitive to within-category variation in speech perception. Cognition 95, B15–B26.

Massaro, D.W., 1972. Preperceptual images, processing time, and perceptual units in auditory perception. Psychological Review 79, 124–145.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Erlbaum Associates, Hillsdale, NJ.

Massaro, D.W., Friedman, D., 1990. Models of integration given multiple sources of information. Psychological Review 97, 225–252.

Massaro, D.W., 1994. Psychological aspects of speech perception: implications for research and theory. In: Gemsbacher, M. (Ed.), Handbook of Psycholingristics. Acadamic Press, New York, pp. 219–263.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press, Cambridge, MA.

Massaro, D.W., Cohen, M.M., 1999. Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. Journal of Speech, Language, and Hearing Research 42, 21–41.

Massaro, D.W., 2012. Speech perception and reading: two parallel modes of understanding language and implications for acquiring literacy naturally. American Journal of Psychology 125 (3), 307–320.

Massaro, D.W., Chen, T.H., 2008. The motor theory of speech perception revisited. Psychonomic Bulletin & Review 15 (2), 453–457.

Massaro, D.W., Cohen, M.M., 1983. Categorical or continuous speech perception: a new test. Speech Communication 2, 15–35.

Massaro, D.W., Oden, G.C., 1980. Speech perception: a framework for research and theory. In: Lass, N.J. (Ed.), Speech and Language: Advances in Basic Research and Practice, vol. 3. Academic Press, New York, pp. 129–165.

Massaro, D.W., Stork, D.G., 1998. Sensory integration and speechreading by humans and machines. American Scientist 86, 236–244.

McMurray, B., Aslin, R.N., 2005. Infants are sensitive to within-category variation in speech perception. Cognition 95, B15–B26.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America 27, 338–352.

Ouni, S., Cohen, M.M., Ishak, H., Massaro, D.W., 2007. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP Journal on Audio, Speech, and Music Processing 2007. http://dx.doi.org/10.1155/2007/47891.

Pellegrino, F., Coupé, C., Marsico, E., 2011. A cross-language perspective on speech information rate. Language 87 (3), 539–558.

Pillsbury, W.B., 1911. The Essentials of Psychology. Macmillan, New York.

Pitt, M.A., 1995. The locus of the lexical shift in phoneme identification. Journal of Experimental Psychology: Learning, Memory, & Cognition 21, 1037–1052.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annual Review of Neurosciemce 27, 169–192.

Roy, D., 2011. http://www.ted.com/talks/deb_roy_the_birth_of_a_word.html.

Saberi, K., Perrott, D.R., 1999. Cognitive restoration of reversed speech. Nature 398, 760.

Saffran, J.R., 2003. Statistical language learning: Mechanisms and constraints. Current Directions in Psychological Science 12, 110–114.

Scholz, B.C., Pullum, G.K., 2006. Irrational nativist exuberance. In: Stainton, R. (Ed.), Contemporary Debates in Cognitive Science. Basil Blackwell, Oxford, pp. 59–80.

Sussman, H.M., Fruchter, D., Hilbert, J., Sirosh, J., 1998. Linear correlates in the speech signal: the orderly output constraint. Behavioral & Brain Sciences 21 (2), 241–299.

Vihman, M.M., 2002. The role of mirror neurons in the ontogeny of speech. In: Stamenov, M., Gallese, V. (Eds.), Mirror Neurons and the Evolution of Brain and Language. Benjamins, Amsterdam, pp. 305–314.

Werker, J.F., Tees, R.C., 1983. Developmental changes across childhood in the perception of non-native speech sounds. Canadian Journal of Psychology 37, 278–286.

Werker, J.F., Logan, J.S., 1985. Cross-language evidence for three factors in speech perception. Perception & Psychophysics 37, 35–44.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 338–353.

Speech Processing, The Cortical Organization of

Josef P Rauschecker, Georgetown University Medical Center, Washington, DC, USA; and Institute for Advanced Study, TU München, München, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The cortical organization of speech encompasses two aspects: The decoding of the speech signal that, as for other complex sounds, begins in Heschl's gyrus and continues as a hierarchical process into anterior-lateral regions of superior temporal cortex. These nonprimary auditory areas form a ventral processing stream via the uncinate fascicle into inferior frontal cortex (ventrolateral prefrontal cortex). Neurophysiological studies in the macaque have demonstrated single-cell responses to subphonetic features like frequency sweeps and bandpassed noise bursts with specific bandwidths. Neuroimaging and direct recordings from auditory cortex confirm the existence of similar neural representations in humans. A second auditory processing stream projects from Heschl's gyrus posteriorly and dorsally through inferior parietal regions into premotor and dorsolateral prefrontal areas. This posterodorsal stream is important for the matching of articulatory features of speech with acoustic-phonetic features processed in the ventral stream.

Introduction: What is Speech?

In linguistics, speech processing is an essential part of spoken language. In the domain-general version of the word, speech is only one category of complex sounds, a category that is used to communicate with other members of the human species. The acoustic properties of this sound category are largely determined by the mechanics of the sound production machinery that humans have at their disposal. This explains why the phonetics of speech are usually described in terms of their production types and locations within the vocal tract (Ladefoged, 2001). Examples are ('bilabial,' 'velar'…), which refer to anatomical parts of the vocal apparatus, or ('fricative,' 'nasal'…), which characterize acoustic features of speech that are generated by a specific constellation of vocal tract elements. Speech sounds are also shaped by the resonance frequencies and filter properties of the vocal apparatus, thus creating 'formants.' Finally, speech throughout evolution must have been optimized in terms of transmission of information or saliency. Thus, frequency sweeps are part of virtually any human language or, for that matter, of virtually any acoustic communication system. Speech sounds obviously also have to be adapted to the human hearing range.

Processing of speech sounds occurs at every level of the auditory system. However, it is not until the level of auditory cortex that speech processing contributes to communication, and speech sounds get categorized and take on meaning. In this article, I will focus on the cortical organization of speech processing, first describing the hierarchical organization of the auditory ventral pathway, which enables the processing of speech and its constituent elements in the acoustic-phonetic domain, and second the contribution of the sensorimotor dorsal pathway in the categorization of speech sounds and matching of articulatory with acoustic-phonetic features. For the first part, we have appropriate animal models, including nonhuman primates, at our disposal that illustrate hierarchical processing of communication sounds, as examples of other complex sounds (or 'auditory objects'). The nonhuman primate model can be directly related to human imaging studies that show close parallels between these primate species. The second part on the relation between speech processing and articulation, also referred to as audio-motor coupling, is more difficult to tackle in an animal model, but I will nevertheless try.

Hierarchical Processing of Complex Sounds

The inner ear is basically built as a frequency analyzer. In a process similar to a Fourier analysis, it breaks down complex sounds into their frequency components and maps them along the basilar membrane of the cochlea (base to apex) from high to low frequencies (von Békésy, 1960). At the level of primary auditory cortex (A1), this frequency specificity is still preserved at the level of single neurons. In almost every neuron, one can find a 'characteristic' or 'best' frequency, for which the neuron shows its lowest threshold or responds with the highest firing rate, respectively. Similarly, auditory cortex maintains the neighborhood relationships between frequencies as a tonotopic map that is first established in the cochlea. At first sight, this makes primary auditory cortex appear as a cortical replication of the cochlea, just as primary visual cortex is sometimes referred to as a 'cortical retina,' because this is where perceptual processing really begins.

Indeed, auditory cortex seems to perform neural operations similar to visual cortex, which is underscored by the existence of similar cell types in both cortices: frequency modulated (FM) selective neurons have been described in A1 of cats (Mendelson and Cynader, 1985; Tian and Rauschecker, 1994) and monkeys (Tian and Rauschecker, 2004), which respond better to frequency sweeps than to constant-frequency tones, just as V1 neurons prefer moving bars over stationary spots of light (Hubel and Wiesel, 1962). Similarly, analogues of simple and complex cells have been described recently (Tian et al., 2013), which distinguish between on- and off-subregions within a neuron's response area or receptive field. This demonstrates that primary auditory cortex does in fact more than just replicate the cochlea: it begins to integrate spectrotemporal information over a small area. At the same time, these findings suggest that neocortex uses similar neural operations (a 'canonical code') across sensory modalities.

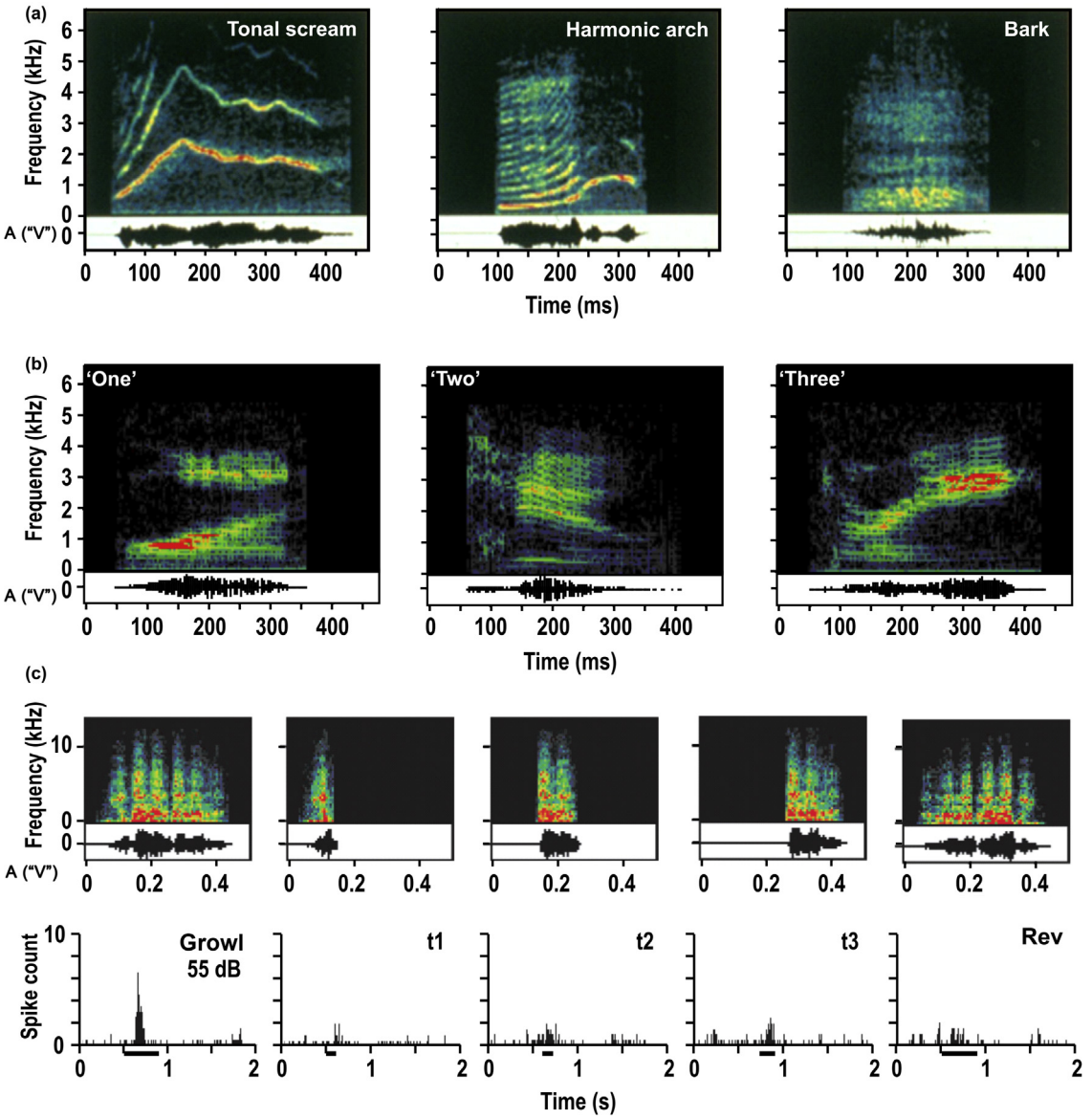

Figure 1 *Communication sounds and single neurons responding to them.* Spectrograms of conspecific vocalizations in (a) rhesus monkeys and (b) humans. Communication calls are commonly characterized by frequency-modulated (FM) sweeps and bandpassed noise (BPN) bursts. (c) Selectivity for specific vocalizations is generated by nonlinear combination of features in the spectral or temporal domain. An example of the latter is shown. A single neuron responds to the whole call, but not to its temporal components ('syllables') nor to its time-reversed version.

Cells at the next stage of auditory cortical processing, the 'belt' region situated on both sides (laterally and medially) of primary 'core' areas, integrate even more strongly across frequencies. Accordingly, belt neurons no longer respond best to pure tones but prefer bandpassed noise (BPN) bursts of a specific bandwidth and center frequency (Rauschecker et al., 1995). These frequencies are still organized in a tonotopic (or rather cochleotopic) fashion. Both BPN and FM-selective neurons are obviously suitable for playing an important role in the processing of vocalizations and human speech, as they are ubiquitous elements of communication sounds (Figure 1(a) and (b)). BPN bursts correspond to formants, and FM sweeps are analogous to formant transitions.

Parallel Processing Streams

Two of the lateral belt (LB) areas in secondary auditory cortex, the anterolateral area (AL) and the caudolateral area (CL), form the origins of two major pathways for the processing of auditory cortical information (Rauschecker and Tian, 2000). Both of these streams make an important contribution to speech processing, but in quite different ways.

Selectivity for Species-Specific Calls in the Auditory Ventral Stream

As noted earlier, neurons in the LB of the monkey generally respond more vigorously to time-variant FM sweeps than to

tones of constant frequency, and FM sweeps are also more effective than BPN bursts. When tested directly with whole monkey calls (MC; **Figure 1(a)**) or components thereof (**Figure 1(c)**), the neurons respond more vigorously to them than to pure tones. MC stimuli were also generally more effective than BPN bursts.

LB neurons responded differentially to different types of MC, especially in AL and areas anterior to it. Acoustic-phonetic response selectivity is based on specific combination, i.e., nonlinear summation, of features in the spectral and the temporal domains (**Figure 1(c)**) (Rauschecker et al., 1995). These spectral and temporal facilitation mechanisms can be compared to combination sensitivity, as described in other species (Suga et al., 1978; Margoliash and Fortune, 1992). In high-density direct cortical surface recordings in humans while they listened to natural, continuous speech, phonetic features were directly related to tuning for spectrotemporal acoustic cues, some of which were encoded in a nonlinear fashion (Mesgarani et al., 2014).

One of the most striking differences between core and belt areas identified so far is the difference in this ability to nonlinearly integrate information both in the spectral and time domain. While more than half of the neurons in LB show some form of nonlinear interaction, only ~10% (or less) of the neurons in core areas A1 or R display this form of behavior, which is one of the strongest arguments for the hierarchical organization of speech processing in auditory cortex. Areas further downstream in the auditory ventral stream show even more evidence for auditory object selectivity in neurophysiological recordings (Romanski and Goldman-Rakic, 2002; Cohen et al., 2009; Kikuchi et al., 2010, 2014).

In humans, detailed neuropsychological studies involving quantitative lesion analysis have recently confirmed that the anterior STG makes a more significant contribution to speech perception than the posterior STG (Dronkers et al., 2004). At the phonological level, phonemic discrimination between real words and nonsense syllables was little affected after lesions of posterior STG, whereas this ability was severely impaired in aphasics with lesions of anterior STG (Blumstein et al., 1977) but not in Broca aphasics, for which the anterior STG was spared. Generally, infarcts in bilateral anterior STG cause auditory agnosias, in which people fail to identify familiar sounds (Clarke et al., 2000; Vignolo, 1982), and the anterior temporal lobe is the location of atrophy in 'semantic dementia' (Mummery et al., 2000; Patterson et al., 2006). Functional neuroimaging studies also demonstrate anterior ST regions to be activated more than posterior regions when subjects listen to speech (Obleser et al., 2006; Scott et al., 2000; Binder et al., 2000) or to other complex sounds (Leaver and Rauschecker, 2010). Finally, in a meta-analysis by DeWitt and Rauschecker (2012) specific to syllables, words, and phrases, syllables were represented just lateral to Heschl's gyrus, words anterolateral to that, and short standard phrases even more anteriorly in the ST (**Figure 2**).

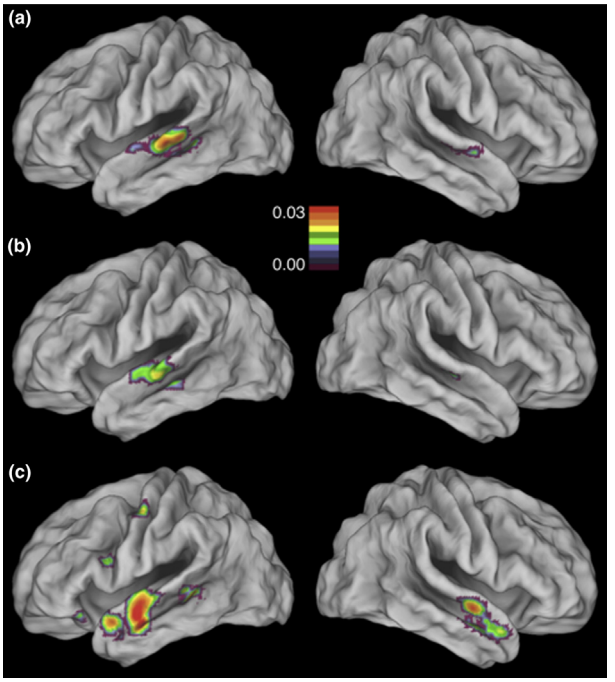

Figure 2 *Meta-analysis of speech sound representations in the human brain.* Results of neuroimaging studies are shown with foci meeting inclusion criteria for ALE-statistic maps for regions of significant concordance ($p < 10^{-3}$). Analyses show an anterior progression with phoneme-length studies showing greatest concordance in left mid-STG (a; $n_{exp} = 14$), word-length studies showing greatest concordance in left anterior STG (b; $n_{exp} = 16$), and phrase-length analyses showing greatest concordance in left anterior STS (c; $n_{exp} = 19$). n_{exp} denotes the number of experiments (i.e., the number of contrasts from independent samples) contributing to each analysis. From DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences of the United States of America 109 (8), E505–E514.

Spatial Selectivity and Sensorimotor Integration in the Auditory Dorsal Stream

The original dual-pathway model of auditory cortical processing (Rauschecker and Tian, 2000) assumes that two largely segregated processing streams originating in the LB subserve the two main functions of hearing: identification of sounds, including speech sounds; and localization of sounds in space. Evidence is strong, from work in both humans and nonhuman studies, that the anteroventral pathway (including anterior superior temporal cortex, aST) supports the former function (see above), whereas the posterodorsal stream (including posterior superior temporal cortex, pST, and inferior parietal lobule, IPL) supports the latter, i.e., processing of space and motion-in-space. Both findings (the assignment of speech sound recognition to aST and a role of pST and IPL in space and motion) would seem to be at odds with classical claims that assume a function in speech and language for pST and IPL in humans.

Is the Dorsal Pathway Really Involved in Speech Perception?

Many textbooks refer to pST and surrounding regions as 'Wernicke's area,' so it seems as if this view dates back to Carl Wernicke (1874), who described patients with lesions of the ST region having difficulties with various aspects of speech. Closer examination of Wernicke's case studies reveals, however,

that the pertinent lesions were not necessarily found in pST alone. A figure in one of his own textbooks (Wernicke, 1881) explicitly marked the whole ST region as speech-related, including its anterior aspects. To reserve the term 'Wernicke's area' for the posterior one-third of ST is, therefore, misleading. The role of the STG as a whole, including its anterior portions, in speech perception was all but forgotten until the linguist Harry Whitaker (1971) and the neurosurgeon Wilder Penfield (Penfield and Roberts, 1959) separately noted the shift from Wernicke's original work.

Wernicke did, however, make the insightful claim that auditory ST regions subserving the deciphering of speech sounds must be connected somehow with the motor speech area in the frontal cortex, which had been discovered by Broca (1861) about a decade earlier. Based on gross anatomical studies of aphasic stroke patients, later researchers assumed that this functional connectivity was provided by a fiber bundle that wound its way from the pST region to Broca's area, the 'arcuate fascicle' (Geschwind, 1965). Present-day work is being performed with high-resolution structural imaging techniques (Rilling et al., 2008; Bernal and Ardila, 2009; Keller et al., 2009). At least one of these studies has revealed that a direct connection from pST to Broca's area is sparse at best (Frey et al., 2008). Instead, most fibers projecting to Broca's area from ST originate in its anterior aspects and follow a ventral pathway via the extreme capsule and/or the uncinate fascicle (Ebeling and Von Cramon, 1992; Friederici et al., 2006). In fact, Wernicke himself suspected that the connection from ST to Broca's area went via the anterior insula, a region that has recently been found to play a role in communication sound processing of monkeys (Remedios et al., 2009). All this adds to the support for an anteroventral pathway in auditory speech processing and one might be tempted to reject the claim of a specific pST (and dorsal stream) involvement in speech processing altogether. However, this would be throwing the baby out with the bathwater.

In order to confirm a genuine role for the pST region in speech and language and to reconcile this role with the spatial functions of that region, one merely has to give up the claim that pST is involved in the 'perception' of speech, including the acoustic-phonetic decoding of speech sounds. Instead, one needs to analyze incidents under which pST and IPL are activated by sounds or tasks with other than spatial connotations.

Dorsal Stream: Representation of Action Sounds and Auditory Perception-Production Links in Voice and Speech

Various studies have demonstrated activation of left parietal regions while subjects were listening to sounds generated by actions, such as tool sounds (Lewis et al., 2005; Pizzamiglio et al., 2005; Engel et al., 2009). These activations often include posterior STS and STG regions, especially when contrasted with unrecognizable control sounds. One possibility is that these regions contain representations of 'doable' sounds (Rauschecker and Scott, 2009). In particular, the medial PT region has been suggested to contain templates of 'doable' articulations (not limited to speech sounds) against which incoming sounds are matched (Warren et al., 2005).

Monkey studies have shown that neurons in auditory cortex are suppressed during vocalization (Müller-Preuss and Ploog, 1981; Eliades and Wang, 2003). This finding is consistent with results from humans, which indicate that superior temporal areas are suppressed during speech production (Numminen et al., 1999; Curio et al., 2000; Houde et al., 2002). This suppression or attenuation of auditory cortex is found even with covert articulation and lipreading, suggesting the existence of an efference-copy pathway from premotor regions to auditory cortex (Kauramäki et al., 2010).

It has been argued that mechanisms of this kind may exist to help distinguish the effects of actions caused by oneself from those caused by the actions of others (Blakemore et al., 1998), specifically differentiating between one's own voice and the voices of others (Rauschecker and Scott, 2009). However, in nonhuman primate studies, auditory neurons that are suppressed during actual vocalizations are often more activated by distorted vocalizations (Eliades and Wang, 2008). This suggests a role for these neurons in the comparison of information from the auditory and motor systems during speech production (Guenther, 2006). Work in humans using distorted feedback of speech production has indeed shown enhanced bilateral activation in pST to distorted feedback, even if it is below the threshold for explicit awareness (Tourville et al., 2008).

There have also been persistent claims for a role of the IPL, i.e., the angular and supramarginal gyri, in phonology (Caplan et al., 1992), particularly an involvement in the 'phonological-articulatory loop' (Baddeley et al., 1984; Aboitiz et al., 2006). This has been confirmed in several functional imaging studies, though the precise localization of activity does vary with the type of task used (Gelfand and Bookheimer, 2003; Buchsbaum and D'esposito, 2008). What seems clear is that the IPL, like pST, is not driven by acoustic-phonetic factors in speech processing but is associated with more domain-general factors (Friederici et al., 2006; Rauschecker and Scott, 2009; Bornkessel-Schlesewsky and Schlesewsky, 2013).

Unified Function of the Dorsal Stream

As documented earlier, posterior ST regions and the IPL participate in the processing of auditory space and motion. At the same time, pST and IPL in humans are also involved in the processing and imagery of auditory sequences, including in speech and music. Both regions receive input from premotor areas in the dorsal and ventral premotor cortex (PMC). Thus, premotor areas may be responsible for assembling the motor patterns for the production of sound sequences. The sounds being produced activate neuronal assemblies in the auditory cortex, which in turn get matched with the corresponding premotor neurons that help produce the sounds. In this way, specific sensorimotor networks are established which, together, represent the sound sequences in a quasimotor code. During learning, subcortical structures like the basal ganglia and the cerebellum are active in binding the correct sets of sensory and motor neurons together.

The same process can be assumed to be at work during learning of speech and speech production. Once learned, listening to speech activates the same circuits as during

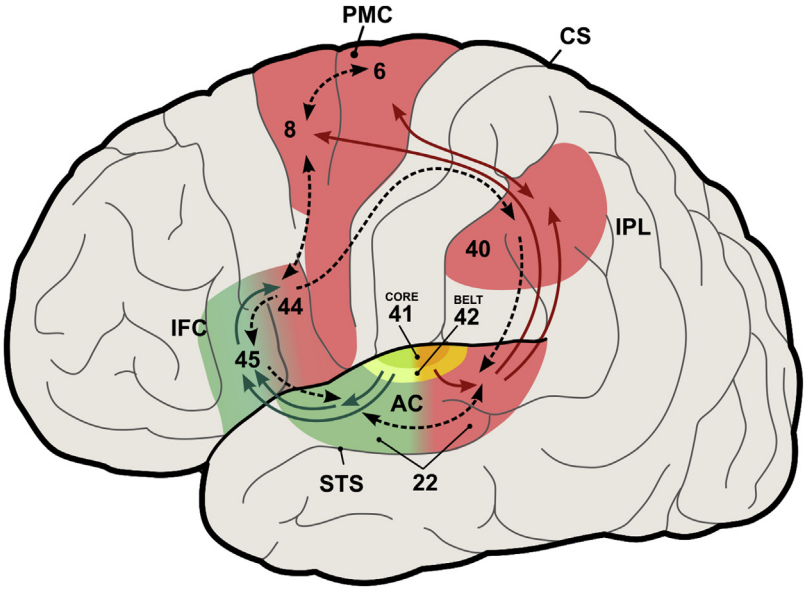

Figure 3 *Dual streams supporting speech processing in the human brain.* The anteroventral stream is shown in green and the posterodorsal stream in red. Cross-stream integration in the inferior frontal gyrus is depicted by the transition from red to green. Black dashed arrows denote feedback connections. Abbreviations: AC, auditory cortex; CS, central sulcus; IFC, inferior frontal cortex; IPL, inferior parietal lobule; PMC, premotor cortex. Numbers denote Brodmann areas. Figure is adapted from Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724.

speech production. While it may not strictly be accurate to talk about a 'motor code' for speech perception (Liberman et al., 1967), correct speech does require a closing of the loop between perception and production and will lead to activation of both networks. The connection between auditory areas in the ST and speech planning areas in the frontal cortex around 'Broca's region,' as postulated by Wernicke, runs through aST and inferior frontal cortex; the loop is closed through PMC via IPL and back to auditory cortex (**Figure 3**). Learning to produce new sounds in a foreign language should, therefore, lead to changes in both sensory and motor representations of the corresponding sounds.

This basic structure of the underlying frontoparietal-sensory loops responsible for sensorimotor planning and control is best described as that of an 'internal model' or 'emulator,' as it has been known in motor control theory and robotics (Rauschecker and Scott, 2009). Such models have been used to describe reaching movements or planning of movement trajectories (Kawato, 1999; Desmurget and Grafton, 2000; Sabes, 2000). More recently, these models have been used to model perception and imagery as well (Wolpert et al., 2003; Grush, 2004). The IPL appears to provide an ideal interface for feed-forward information from motor preparatory networks in the PFC and PMC to be matched with feedback signals from sensory areas. The goal of the internal model is to minimize the resulting error signal in this process.

The feed-forward projection from BA 44 and vPM can be considered the pathway carrying an 'efference copy' or 'corollary discharge' in the classical sense (Sperry, 1950; Von Holst and Mittelstaedt, 1950), informing the sensory system of planned motor articulations that are about to happen. This signal provides a predictive quality to activity running from frontal areas to the IPL, which therefore anticipates the sensory consequences of action. The feedback signal coming to the IPL from pST, on the other hand, can be considered an 'afference copy' (Hershberger, 1976) or reafference with relatively short latencies and high temporal precision (Jääskeläinen et al., 2004; Kauramäki et al., 2010). It can be thought of as a sparse but fast primal sketch of ongoing sensory events (Bar et al., 2006) that are compared with the predictive motor signal in the IPL in real time at every instance. In that sense, both spatial processing and real-time processing of speech (and music) make use of the same general internal model structures that enable the instantiation of smooth sequential motor behaviors, including visuospatial reaching as well as articulation of speech. At the same time, these sensorimotor loops also support the disambiguation of phonological information.

See also: Agnosia; Aphasia; Birdsong and Vocal Learning During Development; Cerebral Cortex; Computational Neuroscience; Cortical Areas Engaged in Movement: Neuroimaging Methods; Cortical Development: Transplantation and Rewiring Studies; Functional Brain Imaging of Language processes; Human Language: Evolutionary Precursors; Human Spatial Orientation, Neural Basis of; Language Processing, Functional Magnetic Resonance Imaging of; Motor Cortex; Multisensory Integration, Principles of; Neural Plasticity; Neural Representations of Objects; Neurolinguistics, History of; Occipital Lobe (Visual Cortex): Functional Aspects; Parietal Lobe; Phonology; Prefrontal Cortex; Premotor Cortex; Serial

Order, Representation of; Sign Language; Speech Perception; Speech Production, Neural Basis of; Speech Production, Psychology of; Temporal Lobe; Topographic Maps; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral; Visual Word Recognition, Neurocognitive Psychology of; Working Memory, Neural Basis of.

Bibliography

Aboitiz, F., Garcia, R.R., Bosman, C., Brunetti, E., 2006. Cortical memory mechanisms and language origins. Brain and Language 98, 40–56.

Baddeley, A., Lewis, V., Vallar, G., 1984. Exploring the articulatory loop. Quarterly Journal of Experimental Psychology 36 A 233–252.

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E., 2006. Top-down facilitation of visual recognition. Proceedings of the National Academy of Sciences of the United States of America 103, 449–454.

Bernal, B., Ardila, A., 2009. The role of the arcuate fasciculus in conduction aphasia. Brain 132, 2309–2316.

von Békésy, G., 1960. Experiments in Hearing. McGraw-Hill, New York.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 10 (5), 512–528.

Blakemore, S.J., Goodbody, S.J., Wolpert, D.M., 1998. Predicting the consequences of our own actions: the role of sensorimotor context estimation. Journal of Neuroscience 18, 7511–7518.

Blumstein, S.E., Baker, E., Goodglass, H., 1977. Phonological factors in auditory comprehension in aphasia. Neuropsychologia 15 (1), 19–30.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2013. Reconciling time, space and function: a new dorsal-ventral stream model of sentence comprehension. Brain and Language 125 (1), 60–76.

Broca, P., 1861. Remarques sur le siège de la faculté du language articulé: suivies d'une observation d'aphémie (perte de la parole). Bulletin de l'Association des Anatomistes 6, 330–357.

Buchsbaum, B.R., D'esposito, M., 2008. The search for the phonological store: from loop to convolution. Journal of Cognitive Neuroscience 20, 762–778.

Caplan, D., Rochon, E., Waters, G.S., 1992. Articulatory and phonological determinants of word length effects in span tasks. Quarterly Journal of Experimental Psychology 45, 177–192.

Clarke, S., Bellmann, A., Meuli, R.A., Assal, G., Steck, A.J., 2000. Auditory agnosia and auditory spatial deficits following left hemisphere lesions: evidence for distinct processing pathways. Neuropsychologia 38, 797–807.

Cohen, Y.E., Russ, B.E., Davis, S.J., Baker, A.E., Ackelson, A.L., Nitecki, R., 2009. A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. Proceedings of the National Academy of Sciences of the United States of America 106, 20045–20050.

Curio, G., Neuloh, G., Numminen, J., Jousmaki, V., Hari, R., 2000. Speaking modifies voice-evoked activity in the human auditory cortex. Human Brain Mapping 9, 183–191.

Desmurget, M., Grafton, S., 2000. Forward modeling allows feedback control for fast reaching movements. Trends in Cognitive Sciences 4, 423–431.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences of the United States of America 109 (8), E505–E514.

DeWitt, I., Rauschecker, J.P., 2013. Wernicke's area revisited: parallel streams and word processing. Brain and Language 127 (2), 181–191.

Dronkers, N.F., Wilkins, D.P., Van Valin Jr., R.D., Redfern, B.B., Jaeger, J.J., 2004. Lesion analysis of the brain areas involved in language comprehension. Cognition 92 (1–2), 145–177.

Ebeling, U., Von Cramon, D., 1992. Topography of the uncinate fascicle and adjacent temporal fiber tracts. Acta Neurochirurgica (Wien) 115, 143–148.

Eliades, S.J., Wang, X., 2003. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. Journal of Neurophysiology 89, 2194–2207.

Eliades, S.J., Wang, X., 2008. Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature 453, 1102–1106.

Engel, L.R., Frum, C., Puce, A., Walker, N.A., Lewis, J.W., 2009. Different categories of living and non-living sound-sources activate distinct cortical networks. NeuroImage 47, 1778–1791.

Frey, S., Campbell, J.S., Pike, G.B., Petrides, M., 2008. Dissociating the human language pathways with high angular resolution diffusion fiber tractography. Journal of Neuroscience 28, 11435–11444.

Friederici, A.D., Bahlmann, J., Heim, S., Schubotz, R.I., Anwander, A., 2006. The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proceedings of the National Academy of Sciences of the United States of America 103, 2458–2463.

Gelfand, J.R., Bookheimer, S.Y., 2003. Dissociating neural mechanisms of temporal sequencing and processing phonemes. Neuron 38, 831–842.

Geschwind, N., 1965. Disconnexion syndromes in animals and man. Brain 88, 237–294, 585–644.

Grush, R., 2004. The emulation theory of representation: motor control, imagery, and perception. Behavioral and Brain Sciences 27, 377–396 discussion 396–442.

Guenther, F.H., 2006. Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, 350–365.

Hershberger, W., 1976. Afference copy, the closed-loop analogue of von Holst's efference copy. Cybernetics Forum 8, 97–102.

Houde, J.F., Nagarajan, S.S., Sekihara, K., Merzenich, M.M., 2002. Modulation of the auditory cortex during speech: an MEG study. Journal of Cognitive Neuroscience 14, 1125–1138.

Hubel, D.H., Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology 160, 106–154.

Jääskeläinen, I.P., Ahveninen, J., Bonmassar, G., Dale, A.M., Ilmoniemi, R.J., Levänen, S., Lin, F.H., May, P., Melcher, J., Stufflebeam, S., Tiitinen, H., Belliveau, J.W., 2004. Human posterior auditory cortex gates novel sounds to consciousness. Proceedings of the National Academy of Sciences of the United States of America 101, 6809–6814.

Kauramäki, J., Jääskeläinen, I.P., Hari, R., Möttönen, R., Rauschecker, J.P., Sams, M., 2010. Transient adaptation of auditory cortex organization by lipreading and own speech production. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 30, 1314–1321.

Kawato, M., 1999. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology 9, 718–727.

Keller, S.S., Roberts, N., Hopkins, W., 2009. A comparative magnetic resonance imaging study of the anatomy, variability, and asymmetry of Broca's area in the human and chimpanzee brain. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 29, 14607–14616.

Kikuchi, Y., Horwitz, B., Mishkin, M., 2010. Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 30, 13021–13030.

Kikuchi, Y., Horwitz, B., Mishkin, M., Rauschecker, J.P., 2014. Processing of harmonics in the lateral belt of macaque auditory cortex. Frontiers in Neuroscience 8, 204.

Ladefoged, P., 2001. Vowels and Consonants: An Introduction to the Sounds of Languages (Online). Blackwell, Oxford.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. Journal of Neuroscience 30 (22), 7604–7612.

Lewis, J.W., Brefczynski, J.A., Phinney, R.E., Janik, J.J., Deyoe, E.A., 2005. Distinct cortical pathways for processing tool versus animal sounds. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 25, 5148–5158.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychological Review 74, 431–461.

Margoliash, D., Fortune, E.S., 1992. Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. Journal of Neuroscience 12, 4309–4326.

Mendelson, J.R., Cynader, M.S., 1985. Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation. Brain Research 327, 331–335.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343 (6174), 1006–1010.

Müller-Preuss, P., Ploog, D., 1981. Inhibition of auditory cortical neurons during phonation. Brain Research 215, 61–76.

Mummery, C.J., Patterson, K., Price, C.J., Ashburner, J., Frackowiak, R.S., Hodges, J.R., 2000. A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. Annals of Neurology 47 (1), 36–45.

Numminen, J., Salmelin, R., Hari, R., 1999. Subject's own speech reduces reactivity of the human auditory cortex. Neuroscience Letters 265, 119–122.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., Rauschecker, J.P., 2006. Vowel sound extraction in anterior superior temporal cortex. Human Brain Mapping 27 (7), 562–571.

Patterson, K., Lambon Ralph, M.A., Jefferies, E., Woollams, A., Jones, R., Hodges, J.R., Rogers, T.T., 2006. "Presemantic" cognition in semantic dementia: six deficits in search of an explanation. Journal of Cognitive Neuroscience 18 (2), 169–183.

Penfield, W., Roberts, L., 1959. Speech and Brain Mechanisms. Princeton University Press, Princeton, NJ.

Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'amico, S., Di Russo, F., 2005. Separate neural systems for processing action- or non-action-related sounds. Neurolmage 24, 852–861.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724.

Rauschecker, J.P., Tian, B., 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 97, 11800–11806.

Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268, 111–114.

Remedios, R., Logothetis, N.K., Kayser, C., 2009. An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 29, 1034–1045.

Rilling, J.K., Glasser, M.F., Preuss, T.M., Ma, X., Zhao, T., Hu, X., Behrens, T.E., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nature Neuroscience 11, 426–428.

Romanski, L.M., Goldman-Rakic, P.S., 2002. An auditory domain in primate prefrontal cortex. Nature Neuroscience 5, 15–16.

Sabes, P.N., 2000. The planning and control of reaching movements. Current Opinion in Neurobiology 10, 740–746.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 (Pt 12), 2400–2406.

Sperry, R.W., 1950. Neural basis of the spontaneous optokinetic response produced by visual inversion. Journal of Comparative and Physiological Psychology 43, 482–489.

Suga, N., O'neill, W.E., Manabe, T., 1978. Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustache bat. Science 200, 778–781.

Tian, B., Kusmierek, P., Rauschecker, J.P., 2013. Analogues of simple and complex cells in rhesus monkey auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 110, 7892–7897.

Tian, B., Rauschecker, J.P., 1994. Processing of frequency-modulated sounds in the cat's anterior auditory field. Journal of Neurophysiology 71, 1959–1975.

Tian, B., Rauschecker, J.P., 2004. Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. Journal of Neurophysiology 92, 2993–3013.

Tourville, J.A., Reilly, K.J., Guenther, F.H., 2008. Neural mechanisms underlying auditory feedback control of speech. Neurolmage 39, 1429–1443.

Vignolo, L.A., 1982. Auditory agnosia. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 298 (1089), 49–57.

Von Holst, E., Mittelstaedt, H., 1950. Das Reafferenzprinzip (Wechselwirkungen zwischen Zentralnervensystem und Peripherie). Naturwissenschaften 37, 464–476.

Warren, J.E., Wise, R.J., Warren, J.D., 2005. Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends in Neurosciences 28, 636–643.

Wernicke, C., 1874. Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis. Cohn & Weigert, Breslau.

Wernicke, C., 1881. Lehrbuch der Gehirnkrankheiten für Aerzte und Studirende. Verlag Theodor Fischer, Kassel, Berlin.

Whitaker, H.A., 1971. On the Representation of Language in the Human Brain; Problems in the Neurology of Language and the Linguistic Analysis of Aphasia. Linguistic Research, Edmonton, Alberta.

Wolpert, D.M., Doya, K., Kawato, M., 2003. A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358, 593–602.

Speech Production, Neural Basis of

Juliana V Baldo, Center for Aphasia and Related Disorders, VA Northern California Health Care System, Martinez, CA, USA
Nina F Dronkers, Center for Aphasia and Related Disorders, VA Northern California Health Care System, Martinez, CA, USA; and University of California, Davis, CA, USA

Abstract

Recent evidence has determined that numerous brain regions play a vital role in speech production. These regions include both cortical and subcortical areas, as well as white matter tracts. This article reviews research from studies in neurologic patients as well as functional neuroimaging studies with healthy individuals that delineate the neural networks underlying the planning and execution of articulatory movements.

Brain Regions Involved in Speech Production

In the nineteenth century, the French neurologist Paul Broca confirmed that speech mechanisms could be localized in the human brain. He did this by interviewing a patient with a severe speech production disorder whose output was limited to the recurring utterance, *tan*. Upon the patient's death, Broca examined the brain and concluded that the patient's inability to speak was due to a lesion in the inferior part of the frontal lobe (Broca, 1861b). A second patient also had severely reduced speech and was subsequently found to have a similar cortical lesion (Broca, 1861a).

Since the time of Broca, scientists have found that lesions to Broca's area alone are not enough to produce lasting speech deficits (e.g., Alexander et al., 1989; Dronkers et al., 2000; Mohr, 1976). Many have attempted to diagram and map additional areas that might also subserve speech and language functions. This article focuses on evidence related to brain mechanisms of speech production. Specifically, evidence will be reviewed from lesion studies, electrocortical stimulation, and functional neuroimaging, all of which are helping to elucidate the neural basis of the planning and execution of articulatory movements.

The involvement of these neural structures can best be understood in the framework of a working model that follows the utterance from its conception to its ultimate articulation. Numerous studies have confirmed that the temporal lobe is important for the translation of concepts into linguistic representations (Baldo et al., 2013; Damasio et al., 1996, 2004; DeLeon et al., 2007; Foundas et al., 1998; Raymer et al., 1997; Trebuchon-Da Fonseca et al., 2009). Once the utterance has been linguistically formulated, it must be transferred to the speech mechanisms that will carry out its production. Articulation itself requires planning, initiation, modification, and execution. In addition to Broca's area, it has become clear that many areas are involved in various aspects of speech production (Eickhoff et al., 2009; Guenther, 2006; Tourville and Guenther, 2011; Price, 2012). Each of these areas is discussed in the following sections in terms of the patient studies that have helped to identify them and the functional neuroimaging research that has corroborated these findings.

Motor Face and Premotor Cortex

The muscles of the vocal mechanism are innervated by cranial nerves V, VII, IX, X, XI, and XII, which receive impulses from the cortex via the corticobulbar tract. The motor face area of primary motor cortex, as well as premotor cortex, is believed to be the source of these projections (Duffy, 2012). Lesions to these areas result in a pure motor speech disorder with slow, effortful speech and impaired articulation but intact language (Alexander et al., 1989). Such lesions are presumed to disrupt the corticobulbar fibers that activate the cranial nerves necessary for speech.

Penfield and colleagues originally showed that a number of regions in primary motor and premotor cortex were involved in speech output, as electrocortical stimulation of these sites resulted in speech arrest (Penfield and Roberts, 1959). Similarly, Ojemann and Mateer (1979) showed that, in four cases, speech arrest followed stimulation of inferior motor and premotor regions in the left frontal cortex.

Activation of primary motor and premotor cortex is commonly reported in neuroimaging studies of speech production and is typically associated with end-point motor execution (Brown et al., 2009; Guenther, 2006; Peeva et al., 2010; Price, 2012; Riecker et al., 2005; Wildgruber et al., 1996, 2001; Wilson et al., 2004). Motor cortex activity associated with speech production, as opposed to simple movements of the articulators, is biased toward motor cortex in the left hemisphere (Terumitsu et al., 2006). There is relative consensus about the role that primary motor and premotor cortices play in end-stage speech production, but there is less agreement about the precise role and location of signals leading up to this stage, as described below.

Broca's Area

Traditionally, the literature has emphasized Broca's area as the critical brain structure for speech production. However, the function of Broca's area is continuously being re-evaluated. Data from stroke patients and from neurosurgical excision have revealed that removal or injury to Broca's area results in a transient mutism that resolves in 3–6 weeks (Dronkers et al., 2000; Penfield and Roberts, 1959). Furthermore, focal lesions to Broca's area do not result in a persisting Broca's

 http://dx.doi.org/10.1016/B978-0-08-097086-8.54018-5

aphasia, and Broca's aphasia may result from lesions outside of Broca's area (Dronkers et al., 2000). In the acute or hyperacute phase, poor perfusion and/or infarcts in Broca's area have been related to impaired speech production (Davis et al., 2008; Hillis et al., 2004); however, this association does not hold for chronic patients with Broca's aphasia (>6 months post-onset; Ochfeld et al., 2010).

With the increasing acknowledgment of the contribution of other brain areas to speech production (discussed in this article), Broca's area is currently viewed as functioning within a larger network of brain regions that support speech production (Hickok, 2012; Sörös et al., 2006). Indeed, a majority, though not all, neuroimaging studies of speech production have reported networks of activation that include Broca's area in tasks involving phonological encoding and articulation, such as picture naming, repetition, word generation, and reading (Eickhoff et al., 2009; Indefrey and Levelt, 2000; Meyer and Jancke, 2006; Moser et al., 2009; Papoutsi et al., 2009; Price et al., 2011). Fine-grained analyses have revealed distinct subregions within Broca's area that are involved in different aspects of speech production (Amunts and Zilles, 2006). For example, Papoutsi et al. (2009) compared fMRI activation in response to pseudowords made up of high versus low frequency sublexical units and found that more ventral portions of the inferior frontal gyrus were related to phonetic encoding while more dorsal regions were related to phonological encoding.

In addition, a number of non-articulatory processes have also been shown to be dependent on Broca's area (Fadiga et al., 2006). An interesting study by Emmorey et al. (2007) showed that the same portion of anterior Broca's area that was activated for overt naming with spoken production was also activated during naming with American Sign Language, suggesting that this area is *modality-independent … not strongly tied to oral-acoustic phonological features of spoken language* (p. 206).

Insular Cortex

The insula is a region of neocortex hidden beneath the intersection of the frontal, parietal, and temporal lobes. In earlier studies, the insula was implicated only vaguely in speech production (e.g., Ojemann and Whitaker, 1978; Tognola and Vignolo, 1980), but more recent studies have found a more direct association. Additional studies have shown that insular cortex is highly connected to other cortical as well as subcortical regions involved in speech production (Eickhoff et al., 2009).

Stimulation of insular cortex in patients undergoing surgery for refractory epilepsy results in speech arrest and a reduction in voice intensity (Afif et al., 2010). In stroke patients, a number of studies have linked left insular cortex to an inability to coordinate complex movements necessary for articulation. Dronkers (1996) reported that patients with apraxia of speech, a motor speech disorder, had lesions in the superior precentral gyrus of the left insula, while those without apraxia of speech did not. Dronkers concluded that this small portion of the insula is critical for coordinating articulatory movements. In a more recent set of analyses using voxel-based lesion symptom mapping, this region of the insula was most critical for complex articulatory movements when the stimuli were multisyllabic, had a high degree of travel between places of articulation, or began with a consonant cluster (Baldo et al., 2011; see also Ogar et al., 2006).

In keeping with these lesion findings, neuroimaging studies with healthy individuals have reported activation in the insula during generation of syllables and single words, word reading, and picture naming (Bohland and Guenther, 2006; Indefrey and Levelt, 2000; Kato et al., 2007; Wise et al., 1999), particularly when the tasks involve articulation of nonrepeated, novel or phonologically complex words relative to automatic or simple articulatory patterns (Moser et al., 2009; Murphy et al., 1997; Shuster, 2009; Riecker et al., 2008; but see Ackermann and Riecker, 2004; Shuster and Lemieux, 2005). An alternative hypothesis is that the anterior insula plays a role in speech production via its role in controlling breathing during speech (Ackermann and Riecker, 2010).

Superior Longitudinal and Arcuate Fasciculi

The arcuate fasciculus is a bundle of fibers that connects the temporal and frontal lobes. As it travels through the parietal lobe, it clusters with the superior longitudinal fasciculus, which incorporates frontoparietal fibers as well. The arcuate fasciculus was identified by Geschwind (1970) as the tract that could potentially connect Wernicke's language area to Broca's speech area as described in Wernicke's original model (Wernicke, 1874). The Wernicke–Geschwind model predicted that disruption of this tract would result in isolated repetition deficits. However, recently a more comprehensive production disorder has been reported. Dronkers and colleagues (Dronkers et al., 2000) found that lesions that sever the arcuate/superior longitudinal fasciculus complex result in a severe production deficit with a complete loss of propositional speech (see also Bates et al., 2003). Only automatized utterances remain. These patients have lost more than just repetition skills – they are unable to transfer any information from temporal lobe language areas to anterior speech mechanisms.

Broca's two original patients were also noted to have this same production disorder, which Broca termed 'aphemia' (Broca, 1861a,b). Dronkers et al. (2007) had the opportunity to perform MR imaging on the postmortem brains of Broca's two original patients. They found that deep lesions also existed in these brains, in addition to the lesions Broca noted in the inferior frontal lobe. These deep lesions involved the arcuate/superior longitudinal complex in both cases, suggesting that this lesion also played a role in these historic patients' production deficits.

More recent studies have confirmed a role for the arcuate/superior longitudinal complex in speech production. Using probabilistic fiber maps derived from diffusion tensor imaging of normal controls, Marchina et al. (2011) found that greater involvement of the left arcuate fasciculus predicted more profound speech production deficits. The role of the anterior portion of the left arcuate fasciculus in fluency was confirmed by Fridriksson et al. (2013) with a larger number of patients. Speech production errors have also been noted after electrostimulation of the arcuate or superior longitudinal fasciculus

during intraoperative mapping for tumor resection (Duffau et al., 2002; Maldonado et al., 2011).

Supplementary Motor Area

The supplementary motor area is located in the superior frontal gyrus and extends medially between the hemispheres. Penfield and others have shown that electrocortical stimulation of this region in patients undergoing surgical removal of epileptic tissue either causes patients to make involuntary vocalizations or interrupts their ability to speak (Penfield and Roberts, 1959).

The supplementary motor area is involved in the initiation of sequential, voluntary movements, including speech (Riecker et al., 2008; Ziegler et al., 1997). This function seems to be true for the supplementary motor area even in the right hemisphere, as demonstrated by several case studies as well as functional imaging data (e.g., Indefrey and Levelt, 2000; Laplane et al., 1977). Infarctions of the supplementary motor area often result in a transcortical motor aphasia (Rapcsak and Rubens, 1994): Spontaneous speech is initially dysfluent, but comprehension, repetition, and reading remain relatively intact. Similarly, complete excision of the supplementary motor area causes a transient aphasia with deficits in articulation that resolve quickly (Penfield and Roberts, 1959).

Functional neuroimaging tasks that elicit supplementary motor cortex activation include the control of breathing for speech and vocalization (Murphy et al., 1997) and automatic speech (e.g., reciting months of the year; Ackermann et al., 1998). Distinct portions of the supplementary motor area appear to be activated for different stages of speech production, including lexical selection, sequence encoding, and motor output (Alario et al., 2006; Riecker et al., 2008). Peeva et al. (2010) reported that a loop between the left supplementary motor area and the left basal ganglia was associated with speech production at the level of the phoneme, with specific phonemes having their own distinct representation in the supplementary motor area. One model of speech production suggests that the supplementary motor area, in conjunction with the basal ganglia, is responsible for gating the output motor signals to the articulators (Tourville and Guenther, 2011).

Basal Ganglia

The basal ganglia are subcortical nuclei comprised of the caudate, putamen, and globus pallidus that interact with motor cortex and with a number of other subcortical structures in a series of feedback loops that help to maintain motor activity (Webb and Adler, 2008). Lesions to the basal ganglia can result in several types of movement disorders including Parkinson's and Huntington's disease, which each have their own specific speech characteristics (Duffy, 2012). Patients with Parkinson's disease (which results from a lack of dopaminergic innervation of the caudate/putamen) exhibit hypokinetic speech with decreased intensity and little modulation of pitch or loudness. In contrast, patients with Huntington's disease (which results from a loss of GABAergic, inhibitory neurons in the caudate) exhibit hyperkinetic speech with erratic control of pitch and loudness. Persistent developmental stuttering has been associated with altered activity in the basal ganglia as well (Alm, 2004; Giraud et al., 2008). These data suggest a role for the basal ganglia in the modulation of articulation and phonation.

Similar disorders can be seen in patients with focal lesions in the basal ganglia. For example, Pickett et al. (1998) reported a case study of a patient with bilateral lesions in the putamen and caudate. Her speech was initially characterized as dysarthric and was very difficult to understand, with poor articulatory and phonatory control. Similarly, Fabbro and colleagues reported at least two cases in which lesions in the basal ganglia resulted in hypophonia and a reduction of speech initiation and output (Fabbro et al., 1996).

Functional neuroimaging studies have shown that regions within the basal ganglia are activated during speech production tasks, including covert speech (Nota and Honda, 2004; Riecker et al., 2005, 2008; Seghier and Price, 2010; Shuster, 2009; Wildgruber et al., 2001). Wildgruber et al. found that the basal ganglia were differentially activated depending on the rate of output and suggested that this brain region may be more critical at low frequencies. It is thought that the basal ganglia modulate motor speech output by way of inhibiting or controlling signals to motor regions (Peeva et al., 2010).

Cerebellum

The cerebellum has traditionally been thought to be involved in fine coordination of motor acts including speech. Evidence for this comes from patients with focal cerebellar lesions. Ataxic dysarthria following cerebellar injury is a motor speech disorder that appears to be due to a loss of coordination in muscular control of the speech apparatus, as opposed to loss of initiation or execution (Duffy, 2012). Mutism has also been associated with midline lesions in the cerebellum, in both children and adults (Coplin et al., 1997). However, this mutism is often transient and can resolve into ataxic dysarthric speech.

A broader theory of cerebellar function suggests that the cerebellum acts as a timing mechanism. That is, the cerebellum's role in motor coordination (for speech and other coordinated motor acts) is based on a more general mechanism that allows for the precise temporal control of such movements (Ackermann, 2008; Ivry and Keele, 1989). Consistent with this hypothesis, Gandour and Dardarananda (1984) and others have shown that cerebellar patients have abnormal voice onset time distributions. It is also argued that the cerebellum plays a role prior to production stages, as it is activated during subvocal rehearsal as well as covert production (Ackermann et al., 1998; 2007). The role of the cerebellum in speech is likely multifaceted, based on the functional heterogeneity of the cerebellum in motor control in general (Ackermann, 2008).

Conclusions

Much has been learned about the neural processes involved in speech based on studies that have examined the specific deficits associated with lesions in particular brain regions. While basic language mechanisms involved in transforming concepts

into linguistic representations cluster in posterior brain regions, speech production is generally supported by more anterior brain regions. New research has suggested a role for the insula in articulatory planning, the supplementary motor area for the initiation of sequential speech movements, the basal ganglia and the cerebellum for pitch, loudness, rate, and timing, and primary motor face cortex and premotor cortex in the execution of articulatory movements. The superior longitudinal fasciculus provides the mechanism by which language information is thought to be transferred to these anterior speech regions from posterior language areas. Functional neuroimaging studies in healthy individuals have corroborated these findings and have allowed the visualization of the networks involved in speech production online (Bohland and Guenther, 2006; Eickhoff et al., 2009; Sörös et al., 2006). Unlike early maps of the language system showing a unitary region (Broca's area) responsible for speech production, it is now clear that a large network of cortical and subcortical regions, as well as white matter pathways, work in concert at different levels of the speech output pathway (Eickhoff et al., 2009; Golfinopoulos et al., 2010; Guenther, 2006; Hickok, 2012; Kent et al., 2000). One model (Riecker et al., 2008) describes a close interaction between the left supplementary motor area, Broca's area, superior cerebellum, and insula to support the earlier stages of speech production (i.e., preparation and initiation), and a close interaction between motor cortex, the basal ganglia, and inferior cerebellum to support end-stages of speech production (i.e., execution). At the same time, much is still unknown about the precise role each of these areas plays and how they interact to support the seemingly seamless act of speech production (Tourville and Guenther, 2011). Continued progress in brain imaging methods and computational modeling will no doubt continue to provide a clearer picture in the years to come.

Acknowledgments

The contents reported/presented within do not represent the views of the Department of Veterans Affairs or the United States Government.

Bibliography

Ackermann, H., 2008. Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. Trends in Neurosciences 31 (6), 265–272.

Ackermann, H., Riecker, A., 2004. The contribution of the insula to motor aspects of speech production: a review and a hypothesis. Brain and Language 89 (2), 320–328.

Ackermann, H., Riecker, A., 2010. The contribution(s) of the insula to speech production: a review of the clinical and functional imaging literature. Brain Structure and Function 214 (5–6), 419–433.

Ackermann, H., Mathiak, K., Riecker, A., 2007. The contribution of the cerebellum to speech production and speech perception: clinical and functional imaging data. The Cerebellum 6 (3), 202–213.

Ackermann, H., Wildgruber, D., Daum, I., Grodd, W., 1998. Does the cerebellum contribute to cognitive aspects of speech production? A functional magnetic resonance imaging (fMRI) study in humans. Neuroscience Letters 247, 187–190.

Afif, A., Minotti, L., Kahane, P., Hoffmann, D., 2010. Middle short gyrus of the insula implicated in speech production: intracerebral electric stimulation of patients with epilepsy. Epilepsia 51 (2), 206–213.

Alario, F., Chainay, H., Lehericy, S., Cohen, L., 2006. The role of the supplementary motor area (SMA) in word production. Brain Research 1076 (1), 129–143.

Alexander, M.P., Benson, D.F., Stuss, D.T., 1989. Frontal lobes and language. Brain and Language 37, 656–691.

Alm, P., 2004. Stuttering and the basal ganglia circuits: a critical review of possible relations. Journal of Communication Disorders 37 (4), 325–369.

Amunts, K., Zilles, K., 2006. A multimodal analysis of structure and function in Broca's region. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's Region. Oxford University Press, New York.

Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F., 2003. Voxelbased lesion-symptom mapping. Nature Neuroscience 6 (5), 448–450.

Baldo, J.V., Wilkins, D.P., Ogar, J., Willock, S., Dronkers, N.F., 2011. Role of the precentral gyrus of the insula in complex articulation. Cortex 47 (7), 800–807.

Baldo, J.V., Arevalo, A., Patterson, J.P., Dronkers, N.F., 2013. Grey and white matter correlates of picture naming: evidence from a voxel-based lesion analysis of the Boston Naming Test. Cortex 49, 658–667.

Bohland, J.W., Guenther, F.H., 2006. An fMRI investigation of syllable sequence production. Neuroimage 32 (2), 821–841.

Broca, P., 1861a. Nouvelle observation d'aphémie produite par une lésion de la troisième circonvolution frontale (New observations of aphemia produced by a lesion of the third frontal convolution). Bulletins de la Société d'anatomie (Paris), 2e serie 6, 398–407.

Broca, P., 1861b. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole) (Remarks on the location of the faculty of articulate language, followed by an observation of aphemia (loss of speech)). Bulletins de la Société d'anatomie (Paris), 2e serie 6, 330–357.

Brown, S., Laird, A.R., Pfordresher, P.Q., Thelen, S.M., Turkeltaub, P., Liotti, M., 2009. The somatotopy of speech: phonation and articulation in the human motor cortex. Brain and Cognition 70 (1), 31–41.

Coplin, W.M., Kim, D.K., Kliot, M., Bird, T.D., 1997. Mutism in an adult following hypertensive cerebellar hemorrhage: nosological discussion and illustrative case. Brain and Language 59, 473–493.

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., Damasio, A., 2004. Neural systems behind word and concept retrieval. Cognition 92 (1–2), 179–229.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380 (6574), 499–505.

Davis, C., Kleinman, J.T., Newhart, M., Gingis, L., Pawlak, M., Hillis, A.E., 2008. Speech and language functions that require a functioning Broca's area. Brain and Language 105 (1), 50–58.

DeLeon, J., Gottesman, R.F., Kleinman, J.T., Newhart, M., Davis, C., Heidler-Gary, J., Lee, A., Hillis, A.E., 2007. Neural regions essential for distinct cognitive processes underlying picture naming. Brain 130 (Pt 5), 1408–1422.

Dronkers, N.F., 1996. A new brain region for coordinating speech articulation. Nature 384, 159–161.

Dronkers, N.F., Plaisant, O., Iba-Zizen, M.T., Cabanis, E.A., 2007. Magnetic Resonance Imaging of the Brains of Paul Broca's Historic Cases. Brain 130 (Pt 5), 1432–1441.

Dronkers, N.F., Redfern, B.B., Knight, R.T., 2000. The neural architecture of language disorders. In: Gazzaniga (Ed.), The New Cognitive Neurosciences. MIT Press, Cambridge, MA.

Duffau, H., Capelle, L., Sichez, N., Denvil, D., Lopes, M., Sichez, J.P., Bitar, A., Fohanno, D., 2002. Intraoperative mapping of the subcortical language pathways using direct stimulations An anatomo functional study. Brain 125 (1), 199–214.

Duffy, J.R., 2012. Motor Speech Disorders: Substrates, Differential Diagnosis and Management. Elsevier Health Sciences.

Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K., 2009. A systems perspective on the effective connectivity of overt speech production. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367 (1896), 2399–2421.

Emmorey, K., Mehta, S., Grabowski, T.J., 2007. The neural correlates of sign versus word production. Neuroimage 36 (1), 202–208.

Fabbro, F., Clarici, A., Bava, A., 1996. Effects of left basal ganglia lesions on language production. Perceptual and Motor Skills 82, 1291–1298.

Fadiga, L., Craighero, L., Roy, A., 2006. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's Region. Oxford University Press, New York.

Foundas, A.L., Daniels, S.K., Vasterling, J.J., 1998. Anomia: case studies with lesion localization. Neurocase 4 (1), 35–43.

Fridriksson, J., Guo, D., Fillmore, P., Holland, A., Rorden, C., 2013. Damage to the anterior arcuate fasciculus predicts non-fluent speech production in aphasia. Brain 136 (11), 3451–3460.

Gandour, J., Dardarananda, R., 1984. Voice onset time in aphasia: Thai. Brain and Language 23, 177–205.

Geschwind, N., 1970. The organization of language and the brain. Science 170, 940–944.
Giraud, A.L., Neumann, K., Bachoud-Levi, A.C., von Gudenberg, A.W., Euler, H.A., Lanfermann, H., Preibisch, C., 2008. Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. Brain and Language 104 (2), 190–199.
Golfinopoulos, E., Tourville, J.A., Guenther, F.H., 2010. The integration of large-scale neural network modeling and functional brain imaging in speech motor control. Neuroimage 52 (3), 862–874.
Guenther, F.H., 2006. Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39 (5), 350–365.
Hickok, G., 2012. Computational neuroanatomy of speech production. Nature Reviews Neuroscience 13 (2), 135–145.
Hillis, A.E., Work, M., Barker, P.B., Jacobs, M.A., Breese, E.L., Maurer, K., 2004. Re-examining the brain regions crucial for orchestrating speech articulation. Brain 127 (7), 1479–1487.
Indefrey, P., Levelt, W.J., 2000. The neural correlates of language production. In: Gazzaniga, M.S. (Ed.), The New Cognitive Neurosciences. MIT Press, Cambridge, MA.
Ivry, R.B., Fiez, J.A., 2000. Cerebellar contributions to cognition and imagery. In: Gazzaniga, M.S. (Ed.), The New Cognitive Neurosciences. MIT Press, Cambridge, MA.
Ivry, R.B., Keele, S.W., 1989. Timing functions of the cerebellum. Journal of Cognitive Neuroscience 1, 136–152.
Kato, Y., Muramatsu, T., Kato, M., Shintani, M., Kashima, H., 2007. Activation of right insular cortex during imaginary speech articulation. Neuroreport 18 (5), 505–509.
Kent, R.D., Kent, J.F., Weismer, G., Duffy, J.R., 2000. What dysarthrias can tell us about the neural control of speech. Journal of Phonetics 28 (3), 273–302.
Laplane, D., Talairach, J., Meininger, V., Bancaud, J., Orgogozo, J.M., 1977. Clinical consequences of corticectomies involving the supplementary motor area in man. Journal of the Neurological Sciences 34, 301–314.
Marchina, S., Zhu, L.L., Norton, A., Zipse, L., Wan, C.Y., Schlaug, G., 2011. Impairment of speech production predicted by lesion load of the left arcuate fasciculus. Stroke 42 (8), 2251–2256.
Maldonado, I.L., Moritz-Gasser, S., Duffau, H., 2011. Does the left superior longitudinal fascicle subserve language semantics? A brain electrostimulation study. Brain Structure and Function 216 (3), 263–274.
Meyer, M., Jancke, L., 2006. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's Region. Oxford University Press, New York.
Mohr, J.P., 1976. Broca's area and Broca's aphasia. In: Whitaker, H., Whitaker, H. (Eds.), Studies in Neurolinguistics, vol. 1. Academic Press, New York.
Moser, D., Fridriksson, J., Bonilha, L., Healy, E.W., Baylis, G., Baker, J.M., Rorden, C., 2009. Neural recruitment for the production of native and novel speech sounds. Neuroimage 46 (2), 549–557.
Murphy, K., Corfield, D.R., Guz, A., Fink, G.R., Wise, R.J., Harrison, J., Adams, L., 1997. Cerebral areas associated with motor control of speech in humans. Journal of Applied Physics 83, 1438–1447.
Nota, Y., Honda, K., 2004. Brain regions involved in motor control of speech. Acoustical Science and Technology 25 (4), 286–289.
Ochfeld, E., Newhart, M., Molitoris, J., Leigh, R., Cloutman, L., Davis, C., et al., 2010. Ischemia in Broca area is associated with Broca aphasia more reliably in acute than in chronic stroke. Stroke 41 (2), 325–330.
Ogar, J., Willock, S., Baldo, J., Wilkins, D., Ludy, C., Dronkers, N., 2006. Clinical and anatomical correlates of apraxia of speech. Brain and Language 97 (3), 343–350.
Ojemann, G., Mateer, C., 1979. Human language cortex: localization of memory, syntax, and sequential motor-phoneme identification systems. Science 205, 1401–1403.
Ojemann, G.A., Whitaker, H.A., 1978. Language localization and variability. Brain and Language 6, 239–260.
Papoutsi, M., de Zwart, J.A., Jansma, J.M., Pickering, M.J., Bednar, J.A., Horwitz, B., 2009. From phonemes to articulatory codes: an fMRI study of the role of Broca's area in speech production. Cerebral Cortex 19 (9), 2156–2165.
Peeva, M.G., Guenther, F.H., Tourville, J.A., Nieto-Castanon, A., Anton, J.L., Nazarian, B., Alario, F.X., 2010. Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. Neuroimage 50 (2), 626–638.
Penfield, W., Roberts, L., 1959. Speech and Brain Mechanisms. Princeton University Press, Princeton, NJ.
Pickett, E.R., Kuniholm, E., Protopapas, A., Friedman, J., Lieberman, P., 1998. Selective speech motor, syntax and cognitive deficits associated with bilateral damage to the putamen and the head of the caudate nucleus: a case study. Neuropsychologia 36, 173–188.
Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62 (2), 816–847.
Price, C., J Crinion, J.T., MacSweeney, M., 2011. A generative model of speech production in Broca's and Wernicke's areas. Frontiers in Psychology 2, 237.
Rapcsak, S.Z., Rubens, A.B., 1994. Localization of lesions in transcortical aphasia. In: Kertesz, A. (Ed.), Localization and Neuroimaging in Neuropsychology. Academic Press, San Diego, CA.
Raymer, A.M., Foundas, A.L., Maher, L.M., Greenwald, M.L., Morris, M., Rothi, L.J.G., Heilman, K.M., 1997. Cognitive neuropsychological analysis and neuroanatomic correlates in a case of acute anomia. Brain and Language 58 (1), 137–156.
Riecker, A., Mathiak, K., Wildgruber, D., Erb, M., Hertrich, I., Grodd, W., Ackermann, H., 2005. fMRI reveals two distinct cerebral networks subserving speech motor control. Neurology 64 (4), 700–706.
Riecker, A., Brendel, B., Ziegler, W., Erb, M., Ackermann, H., 2008. The influence of syllable onset complexity and syllable frequency on speech motor control. Brain and Language 107 (2), 102–113.
Seghier, M.L., Price, C.J., 2010. Reading aloud boosts connectivity through the putamen. Cerebral Cortex 20 (3), 570–582.
Shuster, L.I., Lemieux, S.K., 2005. An fMRI investigation of covertly and overtly produced mono-and multisyllabic words. Brain and Language 93 (1), 20–31.
Shuster, L.I., 2009. The effect of sublexical and lexical frequency on speech production: an fMRI investigation. Brain and Language 111 (1), 66–72.
Silveri, M.C., Di Betta, A.M., Filippini, V., Leggio, M.G., Molinari, M., 1998. Verbal short-term store-rehearsal system and the cerebellum: evidence from a patient with a right cerebellar lesion. Brain 121, 2175–2187.
Sörös, P., Sokoloff, L.G., Bose, A., McIntosh, A.R., Graham, S.J., Stuss, D.T., 2006. Clustered functional MRI of overt speech production. Neuroimage 32 (1), 376–387.
Terumitsu, M., Fujii, Y., Suzuki, K., Kwee, I.L., Nakada, T., 2006. Human primary motor cortex shows hemispheric specialization for speech. Neuroreport 17 (11), 1091–1095.
Tognola, G.R., Vignolo, L.A., 1980. Brain lesions associated with oral apraxia in stroke patients: a clini-neuro-radiological investigation with the CT scan. Neuropsychologia 18, 257–272.
Tourville, J.A., Guenther, F.H., 2011. The DIVA model: a neural theory of speech acquisition and production. Language and Cognitive Processes 26 (7), 952–981.
Trebuchon-Da Fonseca, A., Guedj, E., Alario, F.X., Laguitton, V., Mundler, O., Chauvel, P., Liegeois-Chauvel, C., 2009. Brain regions underlying word finding difficulties in temporal lobe epilepsy. Brain 132 (Pt 10), 2772–2784.
Webb, W., G Adler, R.K., 2008. Neurology for the Speech-Language Pathologist, fifth ed. Mosby/Elsevier, St. Louis, MO.
Wernicke, C., 1874. Der aphasische Symptomencomplex (The Aphasia Symptom Complex). Kohn and Weigert, Breslau.
Wildgruber, D., Ackermann, H., Klose, U., Kardatzki, B., Grodd, W., 1996. Functional lateralization of speech production at primary motor cortex: a fMRI study. NeuroReport 7, 2791–2795.
Wildgruber, D., Ackermann, H., Grodd, W., 2001. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. Neuroimage 13 (1), 101–109.
Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7 (7), 701–702.
Wise, R.J., Greene, J., Buchel, C., Scott, S.K., 1999. Brain regions involved in articulation. Lancet 353, 1057–1061.
Ziegler, W., Kilian, B., Deger, K., 1997. The role of the left mesial frontal cortex in fluent speech: evidence from a case of left supplementary motor area hemorrhage. Neuropsychologia 35, 1197–1208.

Speech Production, Psychology of

Herbert Schriefers, Nijmegen University, Nijmegen, The Netherlands
Gabriella Vigliocco, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14879–14882, © 2001, Elsevier Ltd.

Abstract

Research on speech production investigates the cognitive processes involved in transforming thoughts into speech. This article starts with a discussion of the methodological issues inherent to research in speech production that illustrates how empirical approaches to speech production must differ from related fields like language comprehension. Then, an overview of the processing stages engaged in production as generally agreed upon by the different theories in the field is presented. The processing stages are: conceptual preparation – turning a thought into a verbalizable message to be expressed; grammatical encoding – developing a syntactic frame for the to-be-uttered sentence; and phonological encoding – realizing the phonological content of the syntactic frame. For each processing stage, basic theoretical assumptions, representative findings, and open issues are discussed.

Speech production refers to the cognitive processes engaged in going from mind to mouth (Bock, 1995), that is, the processes transforming a nonlinguistic conceptual structure representing a communicative intention into a linguistically well-formed utterance. Within cognitive psychology, research concerning speech production has taken various forms such as: research concerning the communicative aspect of speaking; research concerning the phonetics of the produced speech; and research concerning the details of the cognitive processing machinery that translates conceptual structures into well-formed linguistic utterances. We focus on the latter.

Native adult speakers produce on average two to three words per second. These words are retrieved from a lexicon of approximately 30 000 (productively used) words. This is no small feat: producing connected speech not only entails retrieving words from memory, but further entails combining this information into well-formed sentences. Considering the complexity of all the encoding processes involved, it is impressive that we produce speech at such a fast rate while at the same time remaining highly accurate in our production. Bock (1991) estimated that slips of the tongue occur in speech approximately every 1000 words despite the ample opportunities for errors. In the following sections the cognitive processes subserving this ability will be discussed.

Methodological Issues

The prototypical empirical approach in cognitive psychology consists of systematically varying some properties of a stimulus (input), and to measure a corresponding behavior. Systematic relations between input and behavior are then used to infer the cognitive processes mediating between them. For example, researchers interested in language comprehension may systematically vary properties of a linguistic input such as the syntactic complexity of sentences or the frequency of words, and measure a corresponding behavior, such as reading times. Differences in the latter are taken as an indication of differences in the processes engaged in building an interpretation of a sentence (the output). Thus, the input is directly observable and completely accessible to systematic manipulation, while the output and the cognitive processes leading to it are inferred from the behavior.

The situation is different for research in speech production. The input (the conceptual structure to be expressed), is not directly observable and not readily accessible for experimental manipulation. By contrast, the output (i.e., the spoken utterance), is directly observable. Given this situation, research on speech production has started by focusing on properties of the behavior. Most prominent in this respect are analysis of speech errors, and of hesitations and pauses as they occur in spontaneous speech. These analyses motivated the general theoretical framework for speech production (**Figure 1**) which is described in detail below. More recently researchers have started to make speech production accessible to standard experimental approaches from cognitive psychology (see Bock, 1996).

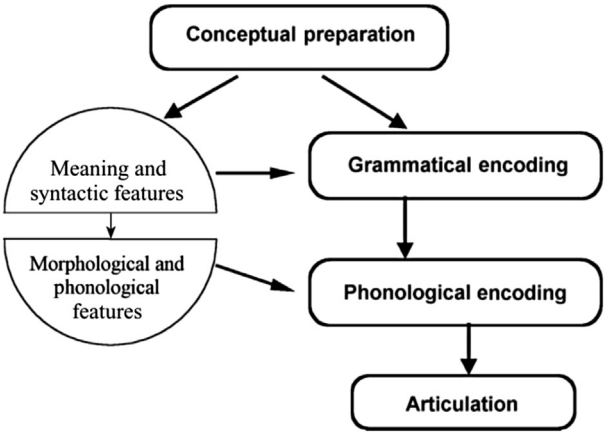

Figure 1 Levels of processing in speech production. Left part refers to stored representations, right part refers to processes integrating these representations into well-formed linguistic utterances.

Levels and Processes

Figure 1 provides an outline of the levels of processing involved in speech production. These levels are shared by most current psycholinguistic models (e.g., Dell, 1986; Garrett, 1988; Levelt, 1989). *Conceptual preparation* refers to the processes that transform thoughts into verbalizable messages. *Grammatical encoding* refers to the processes that turn the messages into syntactically well-formed skeletons for sentences. *Phonological encoding* refers to the processes that provide flesh to the syntactic skeletons, realizing the phonological content of the sentences. At these latter two processing levels, stored lexical information is retrieved from a long-term memory store, the mental lexicon. In particular, during grammatical encoding it is assumed that abstract lexical representations, specifying the syntactic properties of the words, but not their phonological content, are retrieved. During phonological encoding, word form information is retrieved. Finally, *articulation* refers to the processes that implement the phonetic content and turn it into motor commands for speech.

These processing levels were originally inferred from properties of speech errors (*see* Speech Errors, Psychology of). Speech errors can be analyzed with respect to the type of linguistic unit involved (words, morphemes, phonemes), and with respect to constraints that hold for errors concerning a given type of unit. For example, word exchanges (e.g., 'on the room to my door' instead of 'on the door to my room') are constrained by syntactic factors (like the syntactic category of the exchanged words) whereas phoneme exchanges (e.g., 'heft lemisphere' instead of 'left hemisphere') are constrained by phonological factors (like their position as onset, nucleus, or coda of a syllable). These and related results suggest a distinction between a processing level that is primarily concerned with syntactic processing (grammatical encoding), and another level primarily concerned with phonological processing (phonological encoding). The fact that phonemes are relevant units in speech errors also shows that the phonological form of a word is not stored and retrieved as a whole, but built from smaller units, namely the phonemes.

Planning of a sentence is not completed on one processing level and then sent to the subsequent level. For example, phonological encoding starts as soon as the syntactic structure of some part of a sentence has been determined, while the grammatical encoding of the remainder of the sentence continues. This piece-by-piece generation of a sentence is referred to as incremental production. Although there is no consensus on the size of the relevant planning units at the different processing levels, it is generally agreed upon that the size of the planning units becomes smaller as one moves from conceptual preparation through grammatical encoding to phonological encoding and articulation. Incrementality provides a natural account of fluency in speech production. Furthermore, incremental production reduces the burden of storage of intermediate results of the production process at each level.

Although there is general agreement on the levels of processing, theories differ with respect to how the flow of information between levels is characterized. According to modular views, each level of processing first produces a complete output representation for a given planning unit. Only then is this output representation passed onto the next level (e.g., Levelt, 1989). According to interactive views, every processing stage can pass partial results to the next processing level, before the output representation at the higher level has been completely specified, and lower levels of processing can provide feedback to higher levels of processing. As a consequence, processes at the phonological level can affect processes at the grammatical level (e.g., Dell, 1986). In the following sections, we discuss the different processing steps in more detail.

Conceptual Preparation

A first important goal of conceptual preparation is to establish which parts of the conceptually available information are going to be encoded, and in what order (see Levelt, 1989, for an overview). A second goal is to convert the conceptual information into a format that is a suitable for the linguistic formulation processes. One important open issue concerning conceptual preparation is how to characterize the mapping between conceptual and grammatical encoding. First is the question of whether conceptual preparation takes language specific properties into account (*see* Language and Thought: The Neo-Whorfian Hypothesis). Languages differ in which conceptual or formal properties need to be realized as a detail of the sentential form. For example, in English the word 'friend' does not carry information concerning the sex of the friend. In Spanish the corresponding word is differentially inflected for a man ('*amigo*') or a woman ('*amiga*'). In English, adjectives used as predicates (e.g., 'tall' in 'The friend of Luis is tall') do not agree in gender with the noun; in Spanish they do (e.g., '*El amigo de Luis es alto*' or '*La amiga de Luis es alta*'). Thus, in these two languages, conceptual information concerning natural gender must (Spanish) or need not (English) be conveyed by a sentence. Second is the question of whether conceptual information permeates the processes occurring during grammatical encoding beyond providing its input (Vigliocco and Franck, 1999).

Grammatical Encoding

Grammatical encoding refers to the processes involved in developing a syntactically well-formed sentence. It comprises first, those processes that map the relationships among the participants in a conceptual representation (e.g., agent, patient, etc.) onto functional syntactic relations between the words of a sentence (e.g., subject, direct object, etc.). Next, on the basis of the resulting hierarchically organized syntactic frame for the sentence, the words in the sentence are linearized in a manner allowed by the language being spoken. The first step is also referred to as *functional level processing*, the second as *positional level processing* (Garrett, 1988). Distinguishing between building hierarchical and linear frames provides a solution to an important problem that the language production system faces. Speech production has to be incremental to allow for fluent utterances, but at the same time the resulting utterance has to obey language specific constraints that force the use of only certain word orders. Assuming incremental conceptualization, however, the order in which parts of a conceptual message are processed do not necessarily correspond to a word order allowed in the speaker's language. This problem can be

solved by separating the construction of hierarchical structures from the serial ordering of the words. In this way, hierarchical structures can be built in an incremental manner as soon as lexical elements are available, these can then be mapped to permissible linearly ordered positions (Vigliocco and Nicol, 1998). Evidence compatible with such a separation comes from studies showing that the linear position of words can be primed by the previous presentation of the same linear order, even if the hierarchical structure differs (Hartsuiker et al., 1999).

An important open question with respect to grammatical encoding concerns whether the encoding at this level can be influenced by feedback from phonological encoding (Bock, 1986). Research on the process of computing number agreement between a sentence subject and a verb suggests that this grammatical process is sensitive to whether number information in the subject noun phrase is overtly realized in the phonological form of the noun, suggesting the possibility of feedback from phonological encoding to grammatical encoding (e.g., Vigliocco et al., 1995; but see Section Self-monitoring and Repair for an alternative interpretation of the results).

Phonological Encoding

Phonological encoding refers to the processes that are responsible for determining the phonological word forms and prosodic content of the sentence (e.g., Levelt, 1999). First, the phonemes of words are retrieved from the mental lexicon, together with a metrical frame which specifies stress pattern and number of syllables in the word. Following this retrieval process, the resulting sequence of phonemes is syllabified according to a language specific set of syllabification rules. The domain of syllabification is assumed to be the phonological word which can but need not coincide with a lexical word. It should be noted that in this view, the syllabic structure of an utterance is computed on-line. The reason for this assumption is that the actual syllabification of a word in running speech depends on the context in which it appears. For example, the word 'deceive' in isolation is syllabified as 'de-ceive,' but in the context of the utterance 'deceive us,' the syllable structure becomes 'de-cei-veus.' This observation also provides a functional reason why the phonological form of a word is not stored and retrieved as one entity: such a representation would have to be broken up into its constituent parts whenever the stored syllabification of a word does not agree with its syllabification in the context of running speech. The representation formed by phonological encoding, a syllabified phonological code, forms the input for the articulatory processes which realize this code as overt speech. It is an open issue whether the transition from phonological encoding to articulation involves accessing a 'syllabary,' i.e., memory representations specifying the motor tasks that have to be performed to generate each syllable (Levelt, 1999).

Lexical Access

Grammatical and phonological encoding make use of stored representations of words. Parallel to the distinction between grammatical encoding and phonological encoding, two levels of stored representation are distinguished. On one level, words are represented as abstract syntactic (and semantic) entities (technically also referred to as lemmas), and on the other level their phonological form is specified. Evidence for two levels of representation in the mental lexicon comes from tip-of-the-tongue (TOT) states. It has been demonstrated repeatedly that speakers in a TOT state can report grammatical properties of a word (e.g., the grammatical gender of a noun) above chance level even when they are unable to access the phonological form of the word.

It is unresolved whether, within the lexicon, the flow of information from meaning to form can be characterized as discrete, cascading, or fully interactive. Discrete serial models of lexical access (e.g., Levelt, 1999) assume that, first, on the basis of some fragment of a conceptual structure, a set of lemmas becomes activated. This set includes the target lemma as well as semantically related lemmas. Phonological encoding of the word is initiated only after the target lemma has been selected. Hence, the phonological forms of words which are semantically related to the target word should not become activated. Cascading models (e.g., Peterson and Savoy, 1998) assume that every activated lemma sends some activation to its corresponding word form, even before the actual target lemma has been selected. Thus, also the phonological forms of semantic competitors of the target word should receive some activation. Finally, interactive models (e.g., Dell, 1986) allow for feedback from the phonological level to the lemma level. Thus not only will the phonological forms of semantic competitors receive activation, but the phonological forms will also send activation back to all lemmas with similar phonological forms.

The question of whether the phonological forms of semantic competitors do receive activation, as predicted by cascading models and feedback models, or not, as predicted by discrete serial models, is still an empirically unresolved issue. Phonological activation of semantic competitors has been demonstrated for the case of near-synonyms like 'couch' and 'sofa' (e.g., Jescheniak and Schriefers, 1998; Peterson and Savoy, 1998) and has been interpreted in favor of cascading of activation. However, as the evidence is presently restricted to this specific semantic relation, it has been suggested that near-synonyms might be a special case (e.g., Levelt, 1999).

Concerning the question whether there is feedback from lower levels to higher levels in lexical processing, the lexical bias effect has been cited as evidence for feedback. The lexical bias effect concerns the observation that phoneme errors lead to existing words rather than nonwords more often than would be expected by chance. This finding can be explained by feedback from the level of phonological segments to some higher level representing words as one integral unit (e.g., Dell, 1986). According to discrete serial models, by contrast, the chance of incorrect selection of a phoneme should be independent of whether the resulting string of phonemes forms an existing word or not (but see Section Self-monitoring and Repair).

Self-Monitoring and Repair

Speakers not only produce speech, but also monitor their own speech output and this monitoring may occur on overt speech as well as on 'internal speech.' Assuming an internal monitor implies that not all misfunctions of the speech production system will become visible in the eventual output. Rather,

some will be discovered and repaired before they are actually articulated. Such covert repairs can play an important role in the interpretation of speech error results. For example, the lexical bias effect in phoneme errors has also been explained as the result of a pre-articulatory covert repair mechanism. This covert repair mechanism might be more likely to detect and correct phoneme errors leading to nonwords than those leading to words before the resulting error is actually produced. Hence, the lexical bias effect would not imply feedback of activation from lower to higher levels. It should be noted that related arguments have been put forward with respect to other empirical findings which appear to speak against discrete serial models, like the potential influence of phonology on grammatical encoding processes. It is a matter for future investigation to design empirical tests that allow us to differentiate between feedback and monitoring explanations of the relevant empirical phenomena.

See also: Aphasia; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Speech Errors, Psychology of; Speech Production, Neural Basis of.

Bibliography

Bock, J.K., 1986. Meaning, sound and syntax: lexical priming in sentence production. Journal of Experimental Psychology: Learning, Memory and Cognition 12, 575–586.

Bock, J.K., 1991. A sketchbook of production problems. Journal of Psycholinguistic Research 20, 141–160.

Bock, J.K., 1995. Sentence production. From mind to mouth. In: Miller, J.L., Eimas, P.D. (Eds.), Handbook of Perception and Cognition, Speech, Language, and Communication, second ed., vol. 11. Academic Press, San Diego, CA, pp. 181–216.

Bock, J.K., 1996. Language production: Methods and methodologies. Psychonomic Bulletin and Review 3, 395–421.

Dell, G.S., 1986. A spreading-activation model of retrieval in sentence production. Psychological Review 91, 283–321.

Garrett, M.F., 1988. Processes in language production. In: Nieumeyer, F.J. (Ed.), Linguistics: The Cambridge Survey, Biological and Psychological Aspects of Language, vol. 3. Cambridge University Press, Cambridge, UK, pp. 69–96.

Hartsuiker, R.J., Kolk, H.H.J., Huiskamp, P., 1999. Priming word order in sentence production. Quarterly Journal of Experimental Psychology 52A, 129–147.

Jescheniak, J.D., Schriefers, H., 1998. Serial discrete versus cascaded processing in lexical access in speech production: Further evidence from the co-activation of near-synonyms. Journal of Experimental Psychology: Learning, Memory, and Cognition 24, 1256–1274.

Levelt, W.J.M., 1989. Speaking. From intention to articulation. MIT Press, Cambridge, MA.

Levelt, W.J.M., 1999. Models of word production. Trends in Cognitive Science 3, 223–232.

Peterson, R.R., Savoy, P., 1998. Lexical selection and phonological encoding during language production: Evidence for cascaded processing. Journal of Experimental Psychology: Language, Memory, and Cognition 24, 539–557.

Vigliocco, G., Franck, J., 1999. When sex and syntax go hand in hand: Gender agreement in language production. Journal of Memory and Language 40, 455–478.

Vigliocco, G., Nicol, J., 1998. Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? Cognition 68, B13–B29.

Vigliocco, G., Butterworth, B., Semenza, C., 1995. Constructing subject–verb agreement in speech: The role of semantic and morphological factors. Journal of Memory and Language 34, 186–215.

Speech Recognition and Production by Machines

Chin-Hui Lee, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W. Ward, volume 22, pp. 14882–14887, © 2001, Elsevier Ltd.

Abstract

Speech synthesis and recognition by machines has been a fascinating topic for scientists and engineers for many years. Inspired by speaking and listening machines in science fiction, we have witnessed fast and steady technology advances in recent years. In this section, we first describe linear prediction speech analysis which is the most widely adopted model in digital speech communication and speech synthesis. Concatenative speech synthesis is then discussed for text-to-speech synthesis. It is followed by a discussion on hidden Markov modeling of fundamental speech units which is the most successful technique for both speech recognition and speech synthesis. Finally we conclude our discussion by contrasting speech recognition by humans and machines.

Speech is often regarded as the most natural means of communication among humans (Cherry 1968). The speech communication process, characterized as a *speech chain* (Denes and Pinston, 1993), consists of two parties: one speaking to the other while both listening, and vice versa. It involves human-based speech production for one party and speech recognition and understanding for the other party. Since speech is a human's unique capability, scientists and engineers have always been fascinated with the idea of replacing one of the two parties in speech communication that leads to voice user interfaces (Cohen et al., 2004). Due to advances in digital speech processing (Rabiner and Schafer, 2010) and computer technologies, the field of speech production (Fant, 1960; Flanagan, 1965) and recognition (Rabiner and Juang, 1993; Lee et al., 1996; Jelinek, 1997) by machines have made tremendous progresses resulting in a widespread deployment of speech synthesis and recognition products and services in different languages in the last two decades. We now describe the technologies behind speech production and recognition by machines in the following.

Speech Production by Machines

Mechanical devices made to produce human-like vowel sounds dated back to some acoustic resonators created by Kratzenstein in 1779 (Ohala, 2011). An illustration of a physical human speech production mechanism in a two-dimensional rendition is shown in the upper panel of **Figure 1** which displays all the human organs involved in generating the speech signal. The lungs acted as the source of excitation to pump air through trachea and vocal cords. The vocal and nasal tracts play the

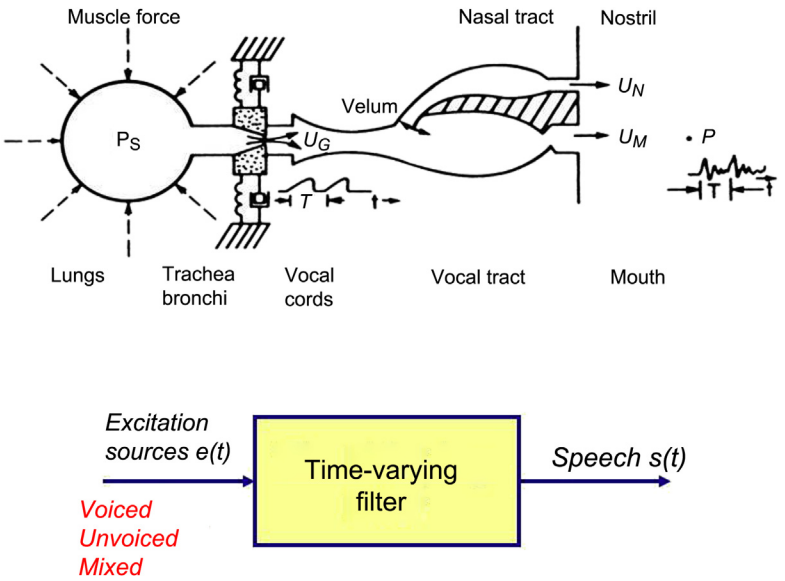

Figure 1 Upper panel: A physical model of the human speech production mechanism. Lower panel: A source-filter engineering model approximating the human speech production system.

roles of resonators to shape the energy concentration in certain frequencies to create the sensation of formants, while the glottis opens and closes in a periodic manner to shape a glottal pressure wave, U_G, and create the sensation of pitch for voice sounds. Finally the sound pressure waves, U_M and U_N, are released through the lip and nostrils, respectively, into the air to produce sound waves. Different sounds are generated in sequence with different configurations in the human speech production system (Flanagan, 1965).

The whole process can be simplified in a *source-filter model* (e.g., Fant, 1960; Rabiner and Schafer, 2010) shown in the bottom panel of **Figure 1** in which the changing voice spectra are simulated by a sequence of time-varying filters with emphasis on different formant combinations. In the meantime, the source excitation, $e(t)$, is modeled by a periodic pitch-like pulse train for voiced sounds, by noise-like signals for unvoiced sounds, and by their mixtures. The speech signal, $s(t)$, is then produced by passing the excitation, $e(t)$, through the time-varying filters and concatenating the signals in time. Although the engineering approximation system is simple-minded, it is now ubiquitously deployed in all the landline and wireless telephone communication systems and implemented in all telephone handsets, smartphones, and terminal devices equipped with speech communication capabilities.

Linear Prediction Model

Linear prediction is widely considered as the most adopted model for speech waveform analysis. Following *autoregressive moving average* (ARMA) modeling of time series, such as the speech signal, short-time speech spectra can be approximated as rational spectra with a small number of poles and zeroes (Atal and Schroeter, 1978). Moreover, poles are often regarded to be keys to speech spectral representation, and therefore autoregressive (AR), or linear predictive, modeling of speech has been studied extensively in the past four decades. Almost all the modern wireless speech communication standards are based on linear predictive coding (LPC) of speech. The classic tutorial (Makhoul, 1975) and book (Markel and Gray, 1976) provide rich sources of information about linear prediction of speech. Its theory, statistical properties, and usages for distance comparison and speech recognition had also been thoroughly investigated (e.g., Itakura, 1975; Rabiner and Schafer, 2010).

In essence, given a segment of speech signal, $\{s(t), t = 1, \ldots, T\}$, each speech sample, $s(t)$, can be approximated by a linear combination of its immediately previous Q samples with a prediction error, $e(t)$, as follows:

$$s(t) = \sum_{q=1}^{Q} \mathrm{a}(q)s(t-q) + e(t)$$

where Q is the order of the *linear prediction filter* which can be used as the time-varying filters in the bottom panel in **Figure 1**, and the set of *linear prediction coefficients*, $\{a(q), q = 1, \ldots, Q\}$, can be used to form an *all-pole* spectrum for the section of speech, $\{s(t), t = 1, \ldots, T\}$, being analyzed. Furthermore, the prediction error signal, $\{e(t), t = 1, \ldots, T\}$, serves as the excitation sequence in **Figure 1**. Given the two sets, $\{a(q), q = 1, \ldots, Q\}$ and $\{e(t), t = 1, \ldots, T\}$, the signal, $\{s(t), t = 1, \ldots, T\}$, can be exactly recovered. However in practical situations, such as in today's digital speech communication, these two sets of parameters are estimated, coded, and therefore approximated, from real speech. For telephone speech, Q is usually set at 10, and the coefficient vector or its alternative representation (Markel and Gray, 1976), at each short-time instant, is vector quantized with a small number of bits to facilitate efficient encoding for wireless speech communication.

Concatenative Speech Synthesis

When we are interested in speech synthesis from text, or TTS synthesis (Taylor, 2009; Sproat, 1998), production models, such as LPC, can be adopted for speech generation. Moreover, pronunciation models associated to text have to be established after some text analysis. This is typically accomplished by segmenting real speech examples into synthesis units, such as phonemes, diphones, syllables, words, phrases, and sentences. Synthesis parameters are then extracted from these units and then concatenated according to the pronunciation specification of the corresponding texts. Finally speech is produced, segment by segment, according to the speech synthesis parameters for each corresponding unit. This process is known as *concatenative speech synthesis*. Unit extraction is a key to improve the quality of synthesized speech. Hand-labeling is often used to warrant a precise unit selection. Signal processing algorithms are also needed to smoothen the speech transition at the concatenation boundaries. Larger units, such as phrases, often produce better quality for concatenative speech synthesis than short units, such as phonemes.

Recently, hidden Markov model of units for speech recognition have been successfully adopted as speech synthesis units as well (Tokuda et al., 2013). In this case, a sequence of HMM parameters can be used to model sound transitions more smoothly than waveform concatenation, and therefore HMM-based speech synthesis often produces smooth-sounding speech which sometime implies good speech quality. As we will show later in the speech recognition section that HMM parameters can be easily estimated from a large collection of speech data with only sentence transcription, or initialized with another existing set of HMMs. It makes unit segmentation and selection not as demanding as in the conventional TTS design. It also facilitates automation of most of the TTS system production. Therefore, we have seen a fast deployment of TTS for many languages by companies, such as Google and Microsoft, because they need to offer different languages in their products and services. Similar to speech recognition design, a good TTS system for a talker often requires recording a large number of utterance examples from the particular speaker, say about 1–10 h. When personalization is concerned it is often desirable to use only a small number of talker-specific speech examples to build a TTS system. HMM-based speaker adaptation (Lee and Huo, 2000) has been adopted for parameter conversion from a set of available TTS-based HMMs to the HMMs for the target speaker. As a result, we have observed a reasonable speech quality with only few minutes of speech examples from the intended speaker. However, the synthesized speech does not sound entirely similar to the intended target speaker. More research is needed.

Speech Recognition by Machines

In science fictions, we have often seen machines equipped with an amazing ability of performing human-centric tasks. For instance, *HAL9000*, in Arthur Clark's '2001: A Space Odyssey' is fluent in speech communication, singing, playing chess, and lip reading. Even after 30 years of fast technology progresses since HAL's conception, only some of these capabilities demonstrated by HAL can be realized in today's world as indicated in the commemorative book *HAL's Legacy* (Stork, 1997). Knowing the difficulty in automatic speech recognition and understanding (ASRU) by machines, and inspired by the unlimited possibilities offered by *intelligent* voice communicators, scientists and engineers have been working hard on this topic in the last 50 years. All the effort can be summarized in two focused tasks, namely, speech-based problem solving demonstrated by R2D2 and speech translation demonstrated by *C–3PO*, in George Lucas' space saga films *Star Wars*.

The first realization of ASR was a hardware for spoken digit recognition of a single person with telephone-quality speech (Davis et al., 1952). Speech analysis and pattern matching were performed by analog circuits. By examining the knowledge sources used by humans in speech understanding as shown in the upper panel in **Figure 2**, starting from the left-hand side, we see that knowledge in acoustic phonetics (Stevens, 2000; Olive et al., 1993) is used for acoustic models to recognize sound patterns, then in phonotactics to build lexical models to recognize feasible sound sequences, and subsequently in syntactics to form language models to recognize word sequences. These three pieces are the fundamental building blocks to establish an *automatic speech recognition* (ASR) system. The reader is referred to the textbook (Rabiner and Juang, 1993) for a comprehensive introduction of the ASR theory.

By adding a semantic model to the ASR module we can perform *spoken language understanding* (SLU). Finally by incorporating discourse and dialogue history into a pragmatic model, we can have *dialogue management* (DM), an essential component for speech-based problem solving. If we replace the DM module by a machine translation (MT) module, we can establish a speech translation system. Moreover, if a spoken response component is needed, then text responses in DM and MT can be converted into speech of the listener's language through TTS synthesis (e.g., Taylor, 2009; Sproat, 1998) as we have discussed earlier. We will now examine the ASR system implementation in detail in the following.

Shannon's Channel Decoding Paradigm

Assume that a talker formulates a sequence of words, W, to speak out and utter a speech signal, $\{s(t)\}$. This process can be considered as converting W into $\{s(t)\}$ through an unknown *channel*. Therefore an inversion process to infer W from $\{s(t)\}$ can then be formulated as a channel decoding problem (Shannon, 1948). Suppose the speech signal can be represented as a sequence of feature vectors, X, and the joint distribution of X and W is denoted as $P(X, W)$. Then we can simplify the problem by assuming the input to the channel is X. Condition on X, then the channel can be characterized by the distribution, $P(W|X)$. Now ASR can be formulated as a channel decoding problem to maximize the posterior probability, $P(W|X)$. By using *Bayes rule*, we have $P(W|X) = P(X|W)P(W)/P(X) = P(X, W)/P(X)$. Since $P(X)$ is the same for all

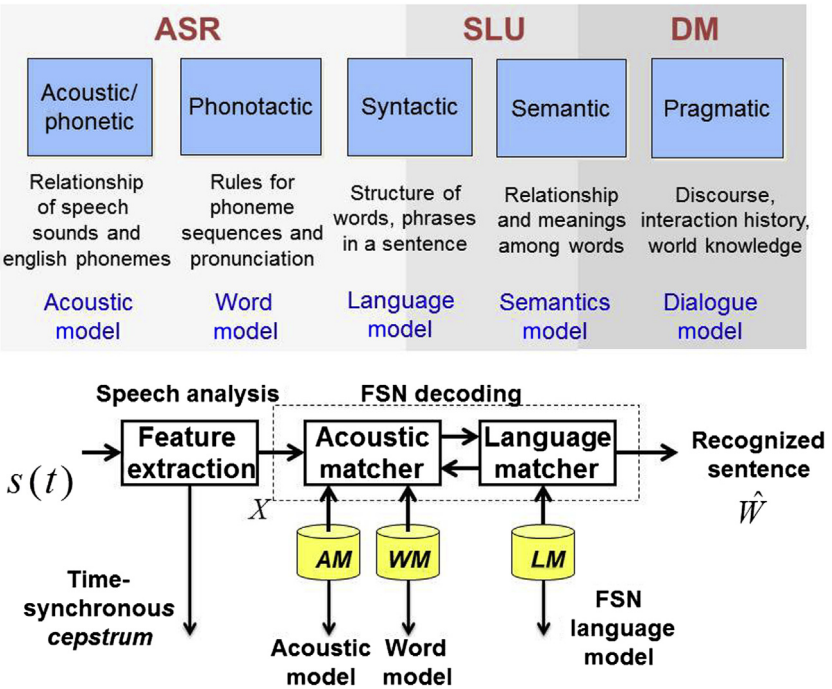

Figure 2 Upper panel: Knowledge sources used for automatic speech understanding (ASU). Lower panel: Block diagram of a state-of-the-art automatic speech recognition (ASR) system.

sentences W, in the search space Γ, and can be dropped, the optimization problem is equivalent to maximizing the joint probability, $P(X, W) = P(X|W)P(W)$, that is,

$$\widehat{W} = \text{argmax}_{W \in \Gamma} P(X|W)P(W)$$

where the optimal word sequence $\widehat{W}$ is obtained by searching all possible sentences in Γ. This is the well-known *minimum risk Bayes decision rule* (e.g., Jelinek, 1997) which achieves the smallest number of sentence errors when the distributions, $P(X|W)$ and $P(W)$, are given to the ASR system designers.

In the real world the two key distributions are not exactly known and their forms need to be assumed and their corresponding parameters need to be estimated from a large collection of speech and text examples, commonly known as *training data*. The estimated distributions can then be used to replace the unknown distributions in the above equation and it becomes the *plug-in maximum a posteriori* (PIMAP) decision rule (Lee and Huo, 2000) which is commonly adopted in all state-of-the-art ASR systems. Although the optimality of the minimum risk Bayes decision rule can no longer be preserved as the distributions for the acoustic and languages are not known, a good speech recognition performance has been observed in many practical ASR tasks.

Speech Recognition System Design

A block diagram of a typical speech recognition system is illustrated in the bottom panel of **Figure 2**, in which the feature extraction module on the left-hand side of the system converts the input signal into a sequence of feature vectors, X, which is then decoded by two modules, acoustic and language matchers, to come up with a recognized sentence. As in LPC-based speech analysis, discussed earlier, short-time Fourier transform (Rabiner and Schafer, 2010) is usually performed at every 10 ms with a 30-ms window to obtain the speech spectrum and then converted into short-time cepstrum (Rabiner and Schafer, 2010). The feature vector at every time instance, $x(t)$, is then formed by concatenating a 13-dimension cepstral vector with its first and second time derivatives to obtain a 39-dimension vector. The quantity X is simply represented by $\{x(t), t = 1, \ldots, T\}$, where T is the number of 10-ms frames in the speech utterance $\{s(t)\}$.

The two quantities needed to evaluate the product in the right-hand side of the above equation, $P(X|W)$ and $P(W)$, are obtained as follows: the acoustic matcher uses phoneme and word models to compute conditional acoustic probability, $P(X|W)$, for a given word sequence W, which can be represented by a sequence of phoneme sounds because each word w in W can also be characterized as a sequence of phonemes according to a word pronunciation dictionary. Furthermore, the language matcher uses task language models to compute word sequence probability, $P(W)$, which can be evaluated with word n-gram probabilities, trained from a large text corpus (Manning and Schutze, 1999). On the other hand the acoustic probability is usually evaluated with hidden Markov model (HMM) (Rabiner, 1989) of phonemes. We will describe this model in detail in the next subsection.

Both models not only give good ASR performances but also fit into the framework of *finite state network* (FSN) representations of *knowledge sources* so that the speech recognition problem can be solved as a *network search* problem over a complex network representation of speech and language with each node in the network characterized by spectral variation of different sounds while each directed arc in the network representing permissible transitions between sounds following both phonotactic and syntactic models imposed by lexical and grammatical constraints shown in **Figure 2**. For a 20 000-word *large vocabulary continuous speech recognition* (LVCSR) task with each word on an average composed of 5 phonemes and 3 states per phoneme HMM, the FSN will have about 300 000 nodes and a few billion arcs for a trigram grammar. Moreover, the network search needs to be carried out at every 10-ms frame and this poses a tremendous computational challenge. *Dynamic programming* (DP) algorithms and *beam search* techniques have been adopted (Ney and Ortmanns, 2000) to efficiently search this type of huge networks in real time using modern hardwares.

Hidden Markov Model

To this date, the most widely adopted modeling approach to ASR is to use a set of HMMs as the acoustic models of subword (e.g., phonemes or syllables) or whole-word units to approximate $P(X|W)$, and to use the statistical n-gram model as language models for words to approximate $P(W)$ (Rabiner, 1989). By using the plug-in MAP decision rule discussed earlier, it has been shown that given a large set of *representative* training speech and text data, good statistical models of speech and language can be constructed to achieve a high performance for a wide range of ASR tasks. Therefore, HMMs and n-grams seem to provide a good approximate parametric forms for $P(X|W)$ and $P(W)$, respectively. Although speech is not generated by HMMs, they are mathematically well-defined and capable of simultaneously modeling both the spectral and temporal variation in speech. The HMM and n-gram parameters are usually estimated with the *maximum likelihood* (ML) criterion (Rabiner, 1989; Jelinek, 1997). The first implementation of an HMM-based ASR system was demonstrated in the Dragon System (Baker, 1975). Progress in acoustic modeling with HMM has advanced a great deal in the last four decades (Lee and Huo, 2000).

The probability distribution function characterizing each HMM state can be as simple as a Gaussian density or as complicated as a Gaussian mixture model (GMM) (Rabiner and Juang, 1993) to model spectral variation of different sounds. It can also be approximated by a neural network (Bourlard and Morgan, 1994). More recently deep neural network (DNN), with multiple and wide hidden layers, has been revitalized to model HMM state observation densities and even for feature extraction due to a discovery of efficient learning architectures and algorithms (Hinton and Salakhutdinov, 2006). It has been shown to improve accuracies for many standardized ASR tasks and other applications.

Human versus Machine Speech Recognition

Following the ASR system design methodology discussed earlier, the state-of-the-art ASR systems often achieve a high recognition accuracy for well-formed utterances for resource-rich tasks in which a large collection of speech and text

examples is available for training. However, the ASR problem is still far from being solved because not all information available in the speech knowledge hierarchy can be easily integrated into the FSN representation of the task used for decoding. When compared with human speech recognition (HSR), ASR systems often deliver much more recognition errors on the same task, especially in adverse conditions (Lippmann, 1997).

It is believed that some of the current issues of integrating various knowledge sources in top-down integrated search (Ney and Ortmanns, 2000) can be partially addressed by processing techniques that can take advantage of the full set of acoustic and language information in speech. It has long been postulated that HSR determines the linguistic identity of a sound based on detected evidences or cues shown in the top panel of **Figure 2**. This calls for a bottom-up framework that links speech processing with information extraction, by spotting speech cues with a bank of attribute detectors, weighing and combining evidences to form cognitive hypotheses, and verifying these theories until a consistent recognition decision can be made. The recently proposed automatic speech attribute transcription (ASAT) (Lee and Siniscalchi, 2013) framework is an attempt to mimic some HSR capabilities with asynchronous speech event detection followed by bottom-up knowledge integration and verification. In the last few years it has demonstrated good potentials; again, more research is needed.

See also: Phonetics, Articulatory; Speech Production, Psychology of.

Bibliography

Atal, B.S., Schroeter, M.R., 1978. Linear prediction analysis of speech based on a pole-zero representation. Journal of the Acoustical Society of America 64 (5), 1310–1318.

Baker, J.K., February 1975. The DRAGON system – an Overview. IEEE Transactions on Acoustics, Speech and Signal Processing 23 (1), 24–29.

Bourlard, H.A., Morgan, N., 1994. Connectionist Speech Recognition: A Hybrid Approach. Springer.

Cherry, C., 1968. On Human Communications. MIT Press.

Cohen, M.H., Giangola, J.P., Balogh, J., 2004. Voice User Interface Design. Anderson-Wiley.

Davis, K.H., Biddulph, R., Balashek, S., 1952. Automatic recognition of spoken digits. Journal of the Acoustical Society of America 24 (6), 637–642.

Denes, P.E., Pinston, E.N., 1993. The Speech Chain: The Physics and Biology of Spoken Languages, second ed. W.H. Freeman and Company, Oxford. A film from the Bell Telephone Laboratories can be seen in. http://techchannel.att.com/play-video.cfm/2011/8/10/AT&T-Archives-The-Speech-Chain.

Fant, G., 1960. Acoustic Theory of Speech Production. Mount, The Hague.

Flanagan, J.L., 1965. Speech Analysis, Synthesis and Perception. Springer Verlag.

Hinton, G.E., Salakhutdinov, R.R., July 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Itakura, F., February 1975. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 23 (1), 67–72.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press.

Lee, C.-H., Soong, F.K., Paliwal, K.K. (Eds.), 1996. Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publisher.

Lee, C.-H., Huo, Q., August 2000. On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proceedings of the IEEE 88 (8), 1241–1269.

Lee, C.-H., Siniscalchi, S.M., May 2013. An information-extraction approach to speech processing: analysis, detection, verification and recognition. Proceedings of the IEEE 101 (5), 1089–1115.

Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Communication 22 (1), 1–15.

Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proceedings of the IEEE 63 (4), 561–580.

Markel, J.D., Gray Jr., A.H., 1976. Linear Prediction of Speech. Springer.

Ney, H., Ortmanns, S., August 2000. Progress in dynamic programming search for LVCSR. Proceedings of the IEEE 88 (8), 1224–1240.

Ohala, J.J., 2011. Christian Gottlieb Kratzenstein: pioneer in speech synthesis. In: Proc. ICPhS XVII.

Olive, J.P., Greenwood, A., Coleman, J., 1993. Acoustics of American English Speech: A Dynamic Approach. Springer Verlag.

Rabiner, L.R., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall.

Rabiner, L.R., Schafer, R.W., 2010. Theory and Applications of Digital Speech Processing. Prentice Hall.

Rabiner, L.R., February 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

Shannon, C.E., July, October 1948. A mathematical theory of communication. Two parts in The Bell System Technical Journal 27, 379–423, 623–656.

Sproat, R. (Ed.), 1998. Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer.

Stevens, K., 2000. Acoustic Phonetics. MIT Press.

Stork, D.G. (Ed.), 1997. HAL's Legacy: 2001's Computer as Dream and Reality. MIT Press.

Taylor, P., 2009. Text-to-Speech Synthesis. Cambridge University Press.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K., May 2013. Speech synthesis based on hidden Markov models. Proceedings of the IEEE 101 (5), 1234–1252.

Spelling, Psychology of

Dominiek Sandra, University of Antwerp, Antwerp, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In many alphabetic orthographies a written word encodes phonemes, morphemes, and several types of constraints. This paper reviews whether children are sensitive to these different types of encoded information and at what age. The results of experiments are related to so-called stage models, which claim that children progress through a series of well-delineated stages, each of which is characterized by the mastery of a new skill. Many experimental data are incompatible with stage models. Some important findings on the spelling performance of experienced spellers are reported. The paper ends with the question whether spellers rely on abstract rule representations or not.

The Rise of Current-Day Psycholinguistics

Although the study of language with the techniques of experimental psychology is certainly much older than many researchers think (see Levelt, 2012), the way in which the mental processes and representations that subserve language use are currently studied does not date back very far in time. As a matter of fact, current-day psycholinguistics is largely the result of the cognitive revolution that was initiated by the linguistic work of Noam Chomsky in the early sixties (Chomsky, 1965). One of his main ideas – that language could be considered a mental organ whose internal structure could be studied on the basis of systematic observation – paved the way for researchers who were interested in the nature of that 'mental organ': its design (structure) and the processes operating on that structure.

However, there is a considerable difference between the goals that Chomsky set himself and the goals of current-day psycholinguists. Chomsky's main goal was to identify the mental foundations behind language (essentially syntactic foundations). His method was to identify which sentences are well-formed (syntactically possible) and which ones are ill-formed in a language. In doing so he hoped to find universal constraints on any language, a set of syntactic restrictions that is innate, specific to the human species, and, hence, genetically encoded. It was a major departure from the prevailing behaviorist doctrine, whose proponents refused to speculate on mental structures and processes because such 'entities,' by definition, escape direct observation; and for that reason (according to behaviorist belief) do not lend themselves to scientific study. (Note that behaviorists never denied the existence of mental processes and representations; they only claimed that it is impossible to study them in a scientific way. Note also that Chomsky was never interested in mental processes but only in the basic syntactic design (the overall mental structural framework) of human languages.)

As the result of Chomsky's focus on the mental underpinnings of language – even though he was never interested in the mental processes and actual representations that are mobilized in actual language use, like speaking (which he referred to as performance, in which all sorts of processes distort a clear view of the hard-wired core of the innate language faculty) – he stood at the beginning of a new scientific era. In this era, psycholinguists, i.e., experimental psychologists with an interest in language or linguists with an interest in the techniques from experimental psychology, started to wonder how it is possible to engage so fluently and automatically in language activities like speaking, listening, reading, and writing. From the end of the sixties of the previous century, psycholinguists designed experiments in which they attempted to unravel the mental processes and representations that are involved in these highly automatized language activities. It follows that these researchers were interested in performance. As a matter of fact, language performance constituted their object of study.

The current contribution deals with the psychological processes and structures that are involved in spelling. As a matter of fact, many psycholinguists are not concerned with spelling. The bulk of experimental work in psycholinguistics deals with the processes that are involved in speech processing (production and perception) and in visual word recognition (reading). This is understandable: all these activities instantiate forms of overtrained behavior: the processes operating over the mental representations that are needed to perform these activities have become automatic and are executed so rapidly, that they even escape our ability to achieve conscious control over them. For instance, we have to make no effort to understand a person who speaks our language, nor can we fixate on a written word and refuse to recognize it. The latter fact is nicely demonstrated by the Stroop effect: people have a hard time saying that the word 'blue' is printed in yellow, for instance, due to a conflict between the recognition of the color and the automatic and fast recognition of the word, which cannot be brought under conscious control (Stroop, 1935). The automaticity that characterizes these language activities is the result of the fact that most people continually engage in these forms of language behavior, i.e., every single day of their lives. The issue that fascinates psycholinguists so much and drives their research questions is, which mental processes make it possible to speak, listen, and read so effortlessly, i.e., why these processes have become automatic and how this automaticity should be understood.

In contrast to these highly automatized language activities, writing is a great deal less automatic. It requires much more conscious attention and involves many more active decision processes than the other three forms of language use. Accordingly, it falls somewhat out of the central focus of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.53039-6

psycholinguists' work. Spelling, being the most peripheral component of the writing process (leading to the actual output of the word in the form of a sequence of letters), also often requires conscious deliberation about how to spell a word, especially when complex spelling rules are involved. Despite this, spelling is the most automatized aspect of the writing process. Moreover, even though spelling is a less automatized form of language activity than speaking, listening, and reading, it still deserves to be studied in its own right because (like the other forms of language activity) it requires complex mental processes and representations. Psycholinguists studying the spelling process are interested in the factors that affect these processes and representations.

What Does a Written Word Encode?

In what follows I will focus exclusively on insights that have been gained from the study of alphabetic writing systems. Space limitations make it impossible to cover the entire spectrum, including research on logographic and syllabic writing systems.

An alphabet is a system of written symbols (letters) that are used to represent the phonemes in the spoken language. Because many languages have more phonemes than the available number of letters in the alphabet they often take recourse to letter combinations to represent some phonemes (e.g., the 'ea' in 'head,' the 'ou' in 'cough'). Hence, phonemes constitute the first piece of information that is encoded in written words.

Even though alphabetic writing systems involve mappings between phonemes and graphemes, some of these systems are more complex than others. The so-called shallow alphabetic systems (e.g., Italian, Czech) come close to one-to-one mappings between phonemes and graphemes, i.e., any given phoneme is almost consistently represented by the same grapheme. At the other end of the continuum, we find the so-called deep alphabetic systems, which involve much more complex mappings between phonemes and graphemes, i.e., a single phoneme can be represented by different graphemes, depending on the word in which it occurs (e.g., the English phoneme 'i' can be represented by 'ee' as in 'tree'; 'ea' as in 'tea'; 'e' as in 'me'; 'y' as in 'theory').

However, many alphabetic writing systems do not only encode phonemes. Often they also orthographically represent morphemes. In such orthographies, the spelling pattern of a morpheme is kept constant across the set of words in which it occurs. In such cases, the orthographic pattern does not match the pattern that would be obtained by applying phoneme-to-grapheme mappings. Whenever the application of the phonological and morphological spelling principles predicts two different spelling patterns, the morphological principle overrides the phonological one.

There are two situations in which this occurs (see also Pacton and Deacon, 2008). The first situation is when the pronunciation and/or the stress pattern of the same morpheme may differ across morphologically related words. Examples are 'heal'-'health' and 'nation'-'nationality.' As can be seen, the spelling of the derived word does not represent its pronunciation (otherwise the spelling of 'health' would be 'helth') but its morphological structure. In English (and the orthographies of many languages), this morphological spelling principle also applies at the level of the suffix. For instance, the past tense suffix in the verb forms 'kicked,' 'called,' and 'committed' is always spelled as 'ed,' even though the suffix is pronounced as 't,' 'd,' and 'Id' respectively. Clearly, the variation in the suffix pronunciation is not orthographically represented and priority is given to the principle of keeping the spelling of the suffix constant.

The second situation is when application of the morphological principle results in the spelling of letters that cannot be heard in the pronunciation of the word, the so-called 'silent' letters. The French language contains lots of words containing such letters. The final letter in 'galop' ('gallop') and 'prend' ('take') is not pronounced but is yet spelled, because the sound corresponding to this letter ('p' and 'd' respectively) can be heard in the morphologically related French words 'galloper' ('to gallop') and 'prendre' ('to take'). A similar phenomenon can be observed in Dutch, where the third person singular, present tense always ends in the letter 't' (representing the 't' suffix), even when this suffix cannot be heard in the pronunciation of the verb form. Hence, the third person singular, present tense of the verb 'worden' is 'wordt' (stem + t,) even though the 't' cannot be heard, due to the fact that a stem-final 'd' is already pronounced as 't.' Hence, 'word' (stem and first person, singular, present tense) and 'wordt' are homophones. However, despite the fact that the 't' is a silent letter in 'wordt,' it is spelled because it can be heard in other third person singular, present tense forms like 'wast' ((wast), 'washes'). Note that the motivation for spelling the silent letters in the French examples above is due to derivational relationships between words, whereas it is due to an analogy between grammatically equivalent forms in the case of the Dutch examples (all of them being third person singular, present tenses). However, the overarching principle is the same: application of the morphological principle ('one morpheme, one spelling pattern').

Besides phonemes and morphemes, orthographies also use graphemes to reflect contextual constraints. Three types of such constraints can be distinguished: positional, graphemic, and phonemic ones. The spelling of the English phoneme 'k' is an example of a positional constraint on spelling. This phoneme can be spelled as 'ck' in word-final position but never in word-initial position: compare 'car,' 'kill' to 'truck,' 'clock.' The spelling of the English phoneme 's' is an example of a phonemic constraint: it is generally spelled as 'se' when following the vowel 'au' but as 'ce' when following the vowel 'ai': compare 'house,' 'mouse' to 'rice,' 'mice.' The phoneme 'z' illustrates what is meant by a graphemic constraint. In word-final position this phoneme is always spelled as 's' when preceded by a consonant but can be spelled in several ways when preceded by a vowel: compare 'dreams,' 'falls,' 'bears' to 'bees,' 'buzz,' 'please.' As these constraints are encoded in written words, spellers may become sensitive to them, even though in some cases the constraints only reflect tendencies and no absolute rules.

Now that we know the different types of information that are encoded in written words, we will see to what extent they have an impact on the spelling process. We will largely concentrate on what has been discovered about the spelling processes

of children, but will also devote a section to factors that affect the spelling process of experienced writers.

Which Factors Affect Children's Spelling?

First, a short description will be given of what has been considered the prototypical model of spelling development in children: the stage model. Then evidence will be presented on children's sensitivity to the different types of information that are encoded in written words. This evidence will be useful to assess the validity of stage models as a description of how children learn to spell.

Stage Models Spelling (and Reading)

Stage models (Ehri, 1986; Frith, 1980; Gentry, 1982) are models about children's progression as they learn to read and spell. Children's knowledge in a particular stage about written words is both applied for reading and spelling purposes. The most salient property of stage models is that they set out from the assumption that the child learns to master the spelling of words by progressing in a stepwise fashion, through a number of stages, which are clearly delineated from each other. Each stage is characterized by the child's sensitivity to a particular kind of information. For instance, skills that are mastered in the second stage are not available to the child in the first stage yet.

Let us take Frith's stage model as an example in case. In the first, so-called logographic phase the child focuses on salient letters in a word. At this stage it has access to only a small number of words (sight vocabulary). It can read these words and also makes attempts to spell them by relying on a number of salient features. For instance, the fact that the word 'yellow' has two sticks in the middle distinguishes this word from other words and is a property that assists the child in reading and spelling this word. However, a reading and spelling strategy based on this type of salient feature in written words cannot be successful in an alphabetic spelling system, for the simple reason that these features quickly lose their distinctiveness when more words are learnt. When a word like 'follow' or 'hollow' is learnt, the two sticks in the middle of the word are not helpful any more. The failure resulting from a reliance on salient properties in written words makes the child move into the alphabetic phase. In this phase the child learns that letters represent sounds, i.e., it achieves insight into the basic property of an alphabetic script: that in order to read or spell, one has to connect letters or letter combinations (graphemes) to sounds (phonemes) or vice versa. However, the application of grapheme-to-phoneme (GP) conversions or phoneme-to-grapheme (PG) conversions does not guarantee that a word is correctly read or spelled. For instance, applying these rules to read or spell a word like 'health' will fail: the vowel will be read as the 'ea' in 'heal' and the word will likely be spelled as 'helth.' In order to become a fluent reader and speller in the language the child, therefore, has to move on to the final phase: the orthographic phase. In this last phase it overcomes the shortcomings of mechanically applying GP and PG conversions and learns the conventional spellings of words.

This sounds like a neat story. It seems intuitively plausible that children acquire literacy in such a stepwise fashion. However, is this indeed the way they learn to spell? We will attempt to answer this question by reporting evidence pertaining to the kinds of information children use when spelling and the ages at which they can use this information. To anticipate the outcome of this exercise: the evidence does not confirm the idealized concept of stage models but rather overwhelmingly suggests that children are sensitive to different types of information at the same time. We will structure our analysis in terms of the three information types discussed in the section 'What Does a Written Word Encode?'

Children's Sensitivity to Phonological Information

Frith's second stage – the alphabetic stage – marks the onset of 'true' spelling. When children realize that letters are the same 'things' as sounds but in the visual modality, i.e., that they represent sounds, they have discovered the essence of an alphabetic script. Hence, it is reasonable to expect that when a child learns which letter represents which sound, it will be able to spell many words, including words that it has never seen before. A simple prediction follows: children who are quick in discovering this key of alphabetic scripts will become the best spellers (and readers).

Note that it can be anticipated that children who have to learn the relationships between phonemes and graphemes in a deep orthography, where the same grapheme can represent different phonemes or a single phoneme can be spelled in different ways, will have a harder time learning to spell (and read) than their peers in shallow orthographies. Any system based on one-to-many mappings is obviously more difficult to master than a system based on one-to-one mappings.

English-speaking children, indeed, experience more problems in learning GP and PG mappings. As an (important) aside, this has caused one of the biggest problems in the history of reading education in the United States: the adoption of the look-and-say method (also known as the whole-word method). This method explicitly bypasses GP mappings because many graphemes correspond to different phonemes, inevitably causing many errors in reading. To avoid these problems the method states that children should be trained in recognizing whole words, i.e., unanalyzed letter patterns, which they directly have to connect to their pronunciation. The results of using this method have been dramatic: many American children failed to learn to read and spell successfully. In an influential book with the significant title 'Why Johnny Can't Read – And What You Can Do About It' (Flesch, 1955), the author offers a simple explanation for the question that is implicit in the title: children who are dependent on the look-and-say method are denied the benefit of GP and PG mappings, which form the backbone of any alphabetic script. Note that even in a deep orthography as the English one, many letters are always sounded out in the same way, especially consonants (e.g., 'b,' 'd,' 'f,' 'k,' 'm,' 'p,' 's,' 't,' 'v,' …). A quarter of a century later the same author re-emphasized his message in a book whose title cannot be misinterpreted: 'Why Johnny Still Can't Read – A New Look at the Scandal of our Schools' (Flesch, 1981). The answer was again very simple: because educational programs still eschewed a phonics

approach and favored the look-and-say method – the latter being the ideal ingredient for creating bad readers (and spellers).

Flesch's emphasis on the importance of a phonics approach to teaching reading (and spelling), based on the observation that look-and-say programs dismally fail, suggests that a sensitivity to phonemes is an indispensable part of learning to become literate. This has, indeed, been confirmed in many experiments. For instance, the longitudinal research reported by Caravolas et al. (2001) showed that phoneme awareness and knowledge of letter–sound correspondences predicted English children's spelling success in the first year of formal schooling. Children with high scores produced the most phonologically acceptable (not necessarily correct) spellings. Importantly, other cognitive measures such as IQ and memory did not explain significant variance between children's performances. An additional finding was that the best spellers in the first year also performed best in producing the conventionally correct spellings in the second year, which sometimes deviated from the spellings yielded by applying PG conversions. This finding suggests that phonological awareness is a strong initial driving force when learning to spell, and that once the insight has dawned upon the child that graphemes represent phonemes, the way has been paved for a successful trajectory in learning to spell. Note that this research was done in English, a language with a deep orthography. Hence, even in such a language, phonological awareness and a good knowledge of PG mappings are crucial in learning to spell.

In this respect, the study by Caravolas et al. (2005a) is interesting. The results showed that phoneme awareness had the same impact on spelling performance in Czech, which has a highly consistent orthography, as in English, with its inconsistent orthography. In other words, mastery of PG mappings is crucial in learning to spell, irrespective of their consistency: "Phoneme awareness appears to be a core component skill of alphabetic literacy, which is equally important for learners of consistent and inconsistent orthographies" (p. 107). Flesch would be delighted with this outcome.

Obviously, a child's learning speed will be higher when the orthography has highly consistent PG mappings, i.e., children will make more errors in deep than in shallow orthographies, even when they are aware of the phonemes in the speech stream and of the PG mappings. However, this difference in learning to spell correctly does not discredit the most important issue: phonological awareness – the possibility to recognize phonemes in the seamless sound-stream and, hence, the ability to relate these phoneme units to orthographic units – is the engine of learning to spell in any alphabetic system. (For additional evidence, see Caravolas et al., 2005b; Cunningham, 2006.)

One might hypothesize that phonological decoding is a first step (the engine) in helping a child to become literate but that with increased experience, reliance on PG mappings diminishes. Although this idea does not sound unreasonable at first sight, it does not fit the facts. In a large regression study of French children from first through fifth grade, Lété et al. (2008) discovered that feed-forward consistency, i.e., the consistency with which phonemes are mapped onto graphemes, was a strong predictor of spelling performance in all grades and, moreover, did not vary across grades. Hence, phonological awareness and the associated use of PG mappings are not only crucial but also constant factors in spelling.

The assertion of stage models that at a particular time in their literacy development, children need to become consciously aware of phonemes in order to be able to map them onto the units of an alphabetic script (graphemes) matches the facts. Children who quickly learn GP and PG mappings become good readers and spellers. The ones lagging behind in this learning process become poor readers and spellers. The question is whether children can become aware of several units of information (see Section What Does a Written Word Encode?) at the same time. We will take up this question in the following sections.

Children's Sensitivity to Morphological Information

Is there any evidence in children's spellings that they are aware of the morphological structure of a spoken word – and that they take this awareness into account when spelling that word? The answer emanating from several experiments is: yes. The first evidence supporting this answer comes from the work reported by Rebecca Treiman and her colleagues. Treiman et al. (1994) studied five- to eight-year-old children and compared their spelling performance on words like 'duty' and 'dirty,' i.e., words in which the 't' is realized as a 'd' in American English. As the 'r' is inaudible, the pronunciation of the two words only differs in the vowels. Still, children made significantly more errors of the type 'dudy' than of the type 'didy.' The authors explain this finding by appealing to children's early sensitivity to morphemes (morphological awareness): they more often write the 't' in a form like 'dirty' rather than in a form like 'duty' because they are somehow aware that the former word is derived from the root 'dirt,' whereas such a derivational relationship is not present in the case of words like 'duty.' Treiman and Cassar (1996) arrived at a similar conclusion when comparing words like 'brand' and 'canned.' In both cases the spoken version of these words ends in the consonant cluster 'nd.' Young children have the tendency to leave out the first letter in consonant clusters, i.e., consonant clusters are difficult and for that reason often simplified. However, the number of such omission errors was much larger for monomorphemic words like 'brand' than for multimorphemic words like 'canned.' Again the authors suggest that this finding indicates that children are sensitive to the fact that 'brand' is not derived from another word whereas 'canned' is. They will leave out the 'n' in 'canned' less often because they realize that it belongs to the stem of the word: 'can' (i.e., they have morphological awareness).

Importantly, Deacon and Bryant (2006) countered the possible critique that superior performance on multimorphemic words might result from the fact that the embedded morphemes have a higher frequency of occurrence than the letter groups in monomorphemic words, such that the so-called morphological effect would, in fact, be an orthographic effect in disguise. They controlled for orthographic frequency by comparing children's spelling performance on orthographically matched words that only differed in their morphemic status: monomorphemic versus multimorphemic (inflected words). Example pairs are 'rocket'–'rocked,' 'turnip'–'turning.'

The spelling of the first segment ('rock,' 'turn') corresponding to the stem in the inflected words was better in the multimorphemic words. This finding indicates that the effect is a true effect of morphological sensitivity rather than the reflection of orthographic frequency.

These findings are not limited to English. In French, too, researchers have shown early effects of morphological awareness in the spelling of words. As mentioned earlier one salient characteristic of French is the omnipresence of silent letters, i.e., letters that are not pronounced in word-final position (e.g., 'tabac,' 'tabacco,' where the final consonant letter is not pronounced). Sénéchal (2000) and Sénéchal et al. (2006) showed that children who are good spellers use morphological strategies to determine whether a letter that is inaudible should be spelled nonetheless. For instance, when having to spell the noun 'galop' ('gallop'), whose 'p' is not pronounced, they relate it to the verb 'galopper' ('to galop'), where a 'p' sound is heard, and thus infer that a 'p' should also be spelled in the noun 'galop,' even though it cannot be heard in the word's pronunciation. The ability to spell these morphologically motivated, silent letters was predicted by a task that tapped into the children's morphological awareness: a morphologically based analogy task. Children were presented with a word pair like 'gris'–'grise,' which are respectively the masculine and feminine forms of the French adjective for 'gray,' and were asked to provide the second word for probes like 'blond' (answer: 'blonde'). Children obtaining high scores on this analogy task also obtained high scores on the spelling of words containing silent letters, indicating that good morphological awareness is a predictor of the ability to correctly spell morphologically motivated, silent letters. Casalis et al. (in press) showed that such a morphological spelling strategy sometimes results in overgeneralization errors: using the verb 'numéroter' ('assign a number') as a reference word, children will sometimes spell 'numérot' ('number'), based on the erroneous inference that an audible 't' in the verb reveals a silent letter 't' in the noun (the correct spelling is 'numéro').

The early development of morphological awareness and the fact that even very young children already make use of that awareness for the purpose of spelling words is bad news for stage models. This finding is incompatible with the assumption of distinct stages that succeed each other in a neat, developmental sequence, each being characterized by the development of a particular skill. The findings concerning morphological awareness, rather, suggest that multiple forms of awareness develop at the same time and are immediately used by children for spelling purposes.

Children's Sensitivity to Spelling Constraints

As discussed in the section 'What Does a Written Word Encode?,' a distinction can be made among three types of spelling constraints, besides the crucial determinants of phonological and morphological awareness: positional, phonemic, and graphemic constraints. Are children sensitive to these constraints and, if so, how early in their development?

Doublets have been used in several studies addressing the question whether young children make use of positional constraints. Doublets are reduplicated letters, like 'll' in 'mall.' The position of such doublets is restricted in many languages. For instance, in English they can occur in word-medial position ('muffin,' 'cellar') and word-final position ('mass,' 'stuff,') but never in word-initial position. In her 1993 paper, Treiman studied what children wrote in the classroom, i.e., she did not perform an experiment but collected authentic materials produced in a naturalistic setting. Interestingly, she found that children in Grade 1 already used consonant doublets and were sensitive to the positional constraint that exists in English: these children already used consonant doublets more often in medial and final word-positions than in word-initial positions. Importantly, they had caught on to this positional constraint without having been told about its existence: this type of spelling knowledge is not taught at schools. Hence, children's knowledge that consonant doublets are not used in word-initial position is the outcome of an implicit learning process.

In another study Treiman performed a controlled experiment (Cassar and Treiman, 1997) in which children in Grades 1 and 2 were studied with respect to their sensitivity to the position of consonant and vowel doublets. The authors used an orthographic constraints test, in which the child had to indicate the most wordlike item in a pair of two nonwords, e.g., 'baff' and 'bbaf.' The children more often rejected the nonword with a doublet in initial position, leading to the same conclusion as the naturalistic study: they had implicitly induced the positions at which English doublets can occur. The authors additionally observed that children had even more fine-grained knowledge about these doublets: they knew which vowels and consonants could appear in doublets and which could not. This was clear from the fact that they more often rejected nonexistent doublets than existing ones, when they both appeared in the same legal position (e.g., 'geed' was preferred over 'gaad,' 'yill' over 'yihh'). Again, they had implicitly derived this knowledge from the written material that they had been exposed to.

Again, similar findings were obtained in other languages, which obviously strengthens their reliability. In French, a consonant doublet can only occur in word-medial position ('ballon,' 'laisser'). Pacton et al. (2001) found that children in Grades 1 to 4 (note: even Grade 1) are already sensitive to this positional constraint: they judged nonwords as more wordlike when the doublet appeared in medial position than when it occurred in initial or final position (e.g., 'lammi' was preferred over 'llami'). They obtained the same result with frequent and infrequent or nonexisting doublets ('lammi'–'llami' vs 'xevvu'–'xxevu'), although children's performance was better on the frequent doublets. In the final section of this article, this will have serious implications for an important theoretical issue. At this point, it suffices to say that French children, like English children, apparently learn implicitly the positions in which consonant clusters can appear.

Space limitations make it impossible to cover more experiments addressing this issue. However, interested readers are referred to a study by Bernstein and Treiman (2001), in which the authors demonstrate that children prefer that an invented grapheme for a particular vowel sound be used in the same position as the one in which it was learnt. The study by Lehtonen (2006) on Finnish children's sensitivity to consonant doublets also deserves mention here. It leads to the same conclusion as all preceding experiments: very young children

(Grade 1) are already aware of the positional constraint regarding doublets in the written language.

Just as young children's early sensitivity to the morphological structure of words was a threat to stage models of spelling development, so is their early sensitivity to positional constraints to the use of doublets. In Grade 1, when they should be exclusively concerned with GP and PG mappings according to stage models, they are already sensitive to positional constraints as well. So, once more it appears that young children can make use of multiple information sources encoded in grapheme strings, at the same time.

Besides positional constraints there are phonemic and graphemic constraints. These will be treated together here. In a language like English there are dependencies between phonemes, such that one phoneme predicts the spelling of another. For instance, the sound 'k' is typically spelled as 'c' before many vowels, as in 'cost,' 'cat,' and 'cow,' but is spelled as 'k' before some vowels like 'e' and 'i' as in 'kettle' and 'kill.' Adults are sensitive to these dependencies. They rely on the vowel to spell word-final consonants or codas (Kessler and Treiman, 2004; Perry et al., 2002), but also to spell word-initial consonants or onsets (Kessler and Treiman, 2004). They also make use of consonants to spell vowels: Treiman et al. (2002) demonstrated that college students rely on the nature of codas and onsets when spelling vowels (see also Treiman and Kessler, 2006). Whereas it is interesting to observe that experienced spellers capitalize on dependencies between phonemes, the question at hand is whether young children are able to do the same.

Hayes et al. (2006) tackled this question by investigating how the performance of young children (Grade 2 (seven to eight years old), Grade 3, and Grade 5) in the spelling of codas and onsets related to the performance of college students. In the materials that they used, the coda's spellings were dependent on the preceding vowel: 'short' vowels being typically associated with what they called extended consonants ('pull'), and 'long' vowels with single consonants ('sail'). The materials focusing on the spelling of the onsets involved the spelling of the 'k' sound, which is typically spelled as 'k' before the sounds 'e' and '/i' ('kept,' 'kill') and as 'c' before other vowel sounds ('camp,' 'cool'). The tasks of the participants involved the spelling of pseudo-words and a forced-choice task between two pseudo-words, where the instruction was to indicate which item best represented a given pronunciation. The results indicated that Grade 2 children were already sensitive to the coda when determining the vowel spelling, although this sensitivity continued to increase with age. However, these same children were equally sensitive to the vowel as college students when the spelling of the onset was concerned.

Hayes et al.'s experiment had been designed to make it possible to disentangle the impact of the vowel's pronunciation (phonotactic effect) and its spelling (graphotactic effect). Hence, the data were analyzed to find out which factor was the determinant. The outcome was that the effects were graphotactic in nature. For instance, the spelling of the coda was not determined by vowel length (a phonological factor) but by the number of letters in the spelling of the vowel: "Each additional letter in the vowel spelling strongly decreased the likelihood that the coda spelling would be extended" (p. 36). The conclusion is that children as young as seven already make spontaneous use of graphotactic cues for determining the spelling of onsets and codas in a word. The authors explicitly point out that their finding is at odds with the predictions of stage models. According to such models these young children should only rely on GP and PG mappings. Once again, we come to the conclusion that young children are much more sensitive to other sources of information as well, and that stage models do not capture what is really going on during spelling development. (See also the study by Aaron et al. (1998) comparing the performance of hearing and deaf children on a memory test for pronounceable and non-pronounceable pseudo-words – 'doof,' 'kram' versus 'dfoo,' 'rmka' – which yielded the result that deaf children prefer the orthographically well-formed pseudo-words equally much as the hearing children, suggesting that these regularities are implicitly induced on the basis of sheer exposure.)

Finally, let us turn to children's sensitivity to the entire orthographic pattern of words: the full orthographic representation of the word. Martinet et al. (2004) report two experiments which clearly show that French children who had received literacy instruction for only three months were already sensitive to lexical knowledge, that is, the mental representation of a whole word's orthographic pattern. In one experiment the authors found that the children's spelling of irregular words was better when the words had a high-occurrence frequency in their reading materials than when their frequency was low. The critical irregular target grapheme was spelled more accurately in the high-frequency words, irrespective of whether the whole word was spelled correctly. The researchers found this pattern twice: after three and after nine months of literacy education. In another experiment, they showed that the children relied on stored knowledge of orthographic patterns when they had to spell French pseudo-words. Their spelling of pseudo-words like 'diRo,' a neighbor of the existing word 'siRo' ('sirop,' syrup), was more accurate than the spelling of control items like 'liko,' which has no neighbor. This finding illustrates that very young spellers draw on analogies with orthographic patterns of whole-word representations in their mental lexicon. These authors, too, present their findings as evidence against the predictions made by stage models: "Overall, the present results suggest that it is not appropriate to view spelling acquisition as a stage-like sequence where lexical knowledge establishes only in later stages of literacy acquisition once alphabetic skills are sufficiently well established. […] Such results are consistent with models of spelling development where lexical knowledge and alphabetic skills mutually influence and develop simultaneously from the very beginning of literacy acquisition" (p. B17).

Lété et al. (2008) arrive at a similar finding. In their regression study on the spelling performance of children in the First to Fifth Grades (see above), they also observed an effect of word frequency from the First Grade onwards. Even though the latter effect took a large jump between Grades 1 and 2, it was already present in First Grade children, just like in the Martinet et al. (2004) study. The word-frequency effect can be considered the fingerprint of lexical representations (representations in the mental lexicon,) and it appears that the ability to read and spell such representations is immediately established, even at the onset of learning.

When considering all study results pertaining to children's morphological sensitivity and their sensitivity to the different types of constraints that are encoded in written words, the only possible conclusion is that stage models do not offer a good description of their developmental trajectory when learning to spell. In contrast to what stage models proclaim, the evidence strongly suggests that children catch on to different types of information that are encoded in written words at a very young age (often from Grade 1 onwards) and use that information when they have to spell themselves – phonemes, morphemes, positional constraints, phonological and/or graphotactic constraints, and lexical knowledge (fully specified spelling patterns).

Factors Affecting the Spelling of Experienced Spellers

In this section, the focus will be on two types of errors that are still made by highly experienced writers: proximity errors and homophone intrusions. In order to discuss these error types the research that was performed in Fayol's research team (on French) and in my own team (on Dutch) will be discussed.

Fayol et al. (1994) set out from the observation that sometimes writers make spelling errors in which the spelling pattern of an inflected verb does not have the suffix spelling that it should have on the basis of the grammatical subject. Rather the suffix is spelled as if the immediately preceding noun were the grammatical subject. To make this less abstract, consider the sentence 'le chien des voisin arrive' ('the neighbor's dog arrives'). A typical spelling error for this sentence would be 'le chien des voisins arrivent,' in which the 'nt' spelling of the suffix (which marks the third person plural) seems to be induced by the preceding plural noun and has thus replaced the correct 'e' spelling (which marks the third person singular, as the grammatical subject – le voisin – is third person singular). Crucially, the two suffixes –'ent' and 'e' are homophonous in the language (both 'arrive' and 'arrivent' are pronounced (aRiv)) and thus do not facilitate spelling. Fayol et al. used a dual-task paradigm to induce such errors in a dictation task: subjects heard a sentence and simultaneously a series of clicks, when the audio fragment was played to them. They had to try to remember the sentence and at the same time count the number of clicks. Once the sentence was finished, they had to write it down and report the number of clicks they had heard. Using this paradigm, they could induce the type of errors illustrated above, which they called 'proximity errors': errors caused by the word in the immediate proximity of the target verb form. Hence, they could illustrate that when a computational procedure is called for in order to determine the spelling of a verb suffix, a situation of 'cognitive overload' can make it impossible to terminate this process. When this occurs, the spelling is determined by irrelevant information sources: in this case the preceding noun. The error is understandable: working-memory is overloaded so that the computational process cannot retrieve the grammatical subject in time, and the first noun that is encountered in working-memory is used for the purpose of determining the suffix spelling.

My colleagues and I observed a similar effect when studying errors in Dutch verb forms (Sandra et al., 1999). The spelling of these verb forms is very straightforward: a suffix is attached to the stem (concatenative morphology), the set of suffixes to choose from is very small, and children are taught these spelling rules since about age 10 (with continuing attention for errors on these forms throughout the educational system). Despite this, many errors are made when a verb form has a homophone in its inflectional paradigm, even by experienced spellers. For instance, the pronunciation of the third person singular, present tense form 'wordt' (becomes) is identical to the pronunciation of the first person singular, present tense form 'word' (become), because a stem-final 'd' is pronounced as 't.' This situation of homophony causes many intrusion errors. One of our findings was that more homophone intrusions were made when the distance (measured in number of words) between the grammatical subject and the inflected verb form increased. We made this observation in a speeded dictation task, administered to 18-year-old secondary school students. What is important for our current purpose is that this finding suggests the same conclusion as Fayol et al.'s proximity errors: when circumstances make it difficult for the computational process to determine the spelling of the suffix, this process cannot terminate, and produces an output such that the suffix spelling is determined on the basis of extraneous factors. In Fayol et al.'s study, the circumstances causing difficulties for the computational process were created by the dual-task paradigm. In our study these circumstances were created by the distance between the grammatical subject and the inflected verb form, which made the retrieval of the grammatical subject from working-memory more time-consuming than the time that was available for spelling the verb form. In Fayol et al.'s study, this situation of cognitive overload made it possible for the noun closest to the verb to determine the (wrong) suffix spelling, whereas in our own study the existence of a homophonic verb form made it possible to spell the wrong suffix.

The homophone intrusions mentioned above constituted an interesting error type in their own right, in the Sandra et al. experiments. More particularly, they were not randomly distributed but yielded a clear pattern: most homophone intrusions were instances where the higher-frequency form was spelled when the lower-frequency form was the actual target. Some Dutch verbs have a form ending in 'dt' (stem-final 'd' followed by the 't' suffix) that occurs with a higher frequency than its homophonous 'd' form (henceforth: DT dominant verbs), whereas other Dutch verbs have a 'd' form that occurs more frequently than its homophonous 'dt' form (henceforth: D dominant verbs). We found that DT dominant verbs cause more homophone intrusions when the 'd' form is the target form, whereas D dominant verbs cause more of these errors when the 'dt' form must be spelled. We interpreted this interaction effect between Dominance and Spelling Pattern as follows: whenever the computational process cannot determine the spelling of the inflectional suffix of a homophonous verb form (because it is a time-consuming process and writers often have insufficient time to retrieve the grammatical subject) a retrieval process accesses the two spelling patterns of the homophonic forms and gives priority to the higher-frequency form. Note that this means that fully regular, inflected verb forms are stored in the mental lexicon, even though this is not logically required, as their spelling patterns can be

computed on the basis of rules (see Sandra, 2010; Sandra and van Abbenyen, 2009 for more experiments addressing these issue; see Sandra, in press, for an overview of all experiments pertaining to these homophone intrusions).

Largy et al. (1996) reported similar findings in French. These authors presented spoken sentences to their subjects, e.g., 'Le chimiste prend des liquides. Il les filtre.' ('The chemist takes liquids. He filters them.') During the presentations of these sentences, subjects also heard a number of phonologically similar words and were told to remember both the sentence and the set of words (dual-task paradigm). The final verb form in the critical sentences was a singular form that was homophonous with a noun. The authors observed that many errors of the following type were made: 'Le chimiste prend les liquides. Il les filtres.' In these errors the last form ('filtres') is the plural of the homophonous noun 'filtre' ('filter'). This plural noun form is also homophonous with the verb form. These errors occurred particularly often, when the noun intruder had a higher frequency than the target verb form. Quite likely a double frequency effect is at work here: (1) the verb activates its homophonous noun, which is selected when it has a higher occurrence frequency than the verb (comparable to what happens in Dutch when two verb forms are homophonous, cf above), and (2) this homophonous noun is spelled in its plural form (remember that it has the same pronunciation in both singular and plural) because the previous word ('les') cannot only be used as a direct object in French (as in the target sentences) but also – and much more frequently – as the plural form of the definite article (i.e., 'the'). The joint operation of these two frequency-driven processes results in the errors of the above type. Importantly, we see that a similar error mechanism is at work in two different languages (Dutch and French): homophone intrusions are made in a regular inflectional domain as the result of the 'pressure' from the higher-frequency homophone.

As can be derived from the above, the spelling errors on regularly inflected verb forms in Dutch and French can be traced back to the same mental processes and representations: (1) the cognitive overload in working memory makes it impossible for the computational process to determine the suffix spelling in time, thus opening the door for extraneous influences (for instance, causing proximity errors), and (2) one of these extraneous influences is the operation of a retrieval mechanism when the regular verb form has a homophonic variant, a mechanism that provides access to two spelling patterns, gives priority to the higher-frequency one, and thus leads to more intrusions of the higher-frequency form than of the lower-frequency one (see Sandra and Fayol, 2003, for a systematic comparison between the two languages).

Abstract Rules or Statistical Learning?

A final question to be addressed is whether the findings in spelling research suggest that spellers rely on abstract representations of spelling rules or depend on a form of statistical learning. A favorite technique for studying this question is to present words or pseudo-words that (1) require the application of a spelling rule but (2) also contain a particular spelling cue, i.e., a cue that suggests a particular spelling, which need not correspond to the spelling required by the rule. The rationale behind this technique is that such spelling cues should be totally irrelevant, as a rule, by definition, makes total abstraction of spelling cues. Indeed, a rule applies to abstract categories (like noun, verb, etc.). Hence, any interaction between spelling rules and spelling cues would suggest that spellers do not rely on an abstract rule when spelling.

Kemp and Bryant (2003) studied the spelling of English plurals, using real words and pseudo-words. They focused on the spelling of the 'z' sound when it signals plurality (e.g., 'tears'). It is clear that the spelling rule for plural formation abstracts away from the fact that the final sound is 'z.' The English plural suffix is always spelled as 's.' Moreover the rule also abstracts away from the preceding graphemes. In order to find out whether spellers applied an abstract rule or were also affected by the orthographic context, two types of roots and pseudo-roots were studied: those ending in a voiced consonant and those ending in a vowel. Whereas the sound 'z' preceded by a voiced consonant must be spelled as 's' – English consonants cannot be followed by the letter 'z' (e.g., orthographic clusters like 'bz,' 'dz,' … do not exist in word-final position) – the sound 'z' preceded by a long vowel allows other spellings as well (e.g., 'bees,' 'buzz,' 'please,' 'breeze'). The results showed that English children, both good and poor spellers, made more spelling mistakes on the plurals of roots and pseudo-roots ending in a long vowel. This strongly suggests that when spelling a plural form, English children do not rely on an abstract rule. If they had done that in the experiment, they would have been insensitive to the preceding grapheme, as the spelling rule ('add s') applies to the abstract category 'noun' and is, hence, indifferent to graphotactic cues. It seems that multiple sources of information are unconsciously combined when spelling the plural of a (pseudo-) word.

Even more surpirsing is Deacon and Pacton's finding (2006, mentioned in Pacton and Deacon, 2008) that this effect is very persistent. They ran a similar experiment but used adults with tertiary education as their subjects this time. They presented pseudo-nouns in sentence contexts and in isolation. Here too a graphotactic effect emerged in the data: fewer 's' spellings occurred after a long vowel than after a consonant. Apparently (and surprisingly) even experienced spellers cannot avoid interference from irrelevant graphotactic cues.

Pacton et al. (2005) arrived at a similar conclusion in French. Using French children (ages: 7–11) and adults, they studied how the diminutive suffix, which sounds like [et], was spelled, depending on the preceding letter. The suffix should always be spelled as 'ette,' no matter which letter precedes it. However, Pacton et al. wondered whether the fact that this suffix occurs much more frequently after the letter 'r' than after the letter 'f' in the set of French diminutives would have an effect on subjects' spelling performance. Pseudo-words were used (e.g., 'vitar,' 'vitaf'). As in the English studies on the spelling of the plural suffix 's,' the findings indicated that the rule was not applied in an abstract way, i.e., by ignoring the irrelevant preceding letter. In all subject groups, including the group of adults, a sensitivity to the graphotactic context was found: the correct spelling 'ette' was produced more often in words ending in the letter 'r' ('virarette') than in words ending in the letter 'f' ('vitafette'). Hence,

spellers – even experienced ones – appear unable to ignore the preceding orthographic context and are more prepared to spell 'ette' after the letter 'r' (thus producing a high-frequency orthographic sequence in the set of diminutive words) than after the letter 'f.'

Pacton et al. (2001) provided particularly convincing evidence that spellers do not make use of abstract rules. When studying children's sensitivity to the positional constraints on the use of French consonant doublets (see above), the researchers not only found that children develop a very early sensitivity to the fact that such doublets can only occur in word-medial position; they also found that this positional-sensitivity dropped when consonant doublets were used that do not occur in French. Whereas such doublets obviously must have given rise to strange letter sequences, there should have been no drop in performance if the children had used an abstract rule. Indeed, such a rule, by definition, applies to any consonant doublet: familiar or novel. In other words, an abstract rule guarantees a seamless transfer to novel doublets – which was not the case.

The bottom line of these (and other) experiments of this type can be briefly summarized: spellers do not appeal to abstract spelling rules while ignoring all kinds of information that are irrelevant for rule application. Rather they seem to pool all information sources related to the spelling of a sound. One of the most surprising findings in this connection is that even experienced spellers do not succeed in 'switching off' these irrelevant spelling cues and make spelling errors as a result of it. To be honest, these findings do not reject the existence of abstract rule representations. That is only one possibility. Another is that spellers have such representations but cannot avoid other information sources competing with these rules, and such competition causes spelling errors. Hence, the question whether abstract rules are a chimera or not is still unsolved. However, it is already astonishing to observe, across different experiments, that even experienced spellers are unable to ignore irrelevant spelling cues when they have to apply a spelling rule.

See also: Case, Morphological: Linguistics; Psycholinguistics: Overview; Speech Errors, Psychology of; Writing Systems, Psychology of; Writing Systems; Writing, Evolution of.

Bibliography

Aaron, P.G., Wilczynski, S., Keetay, V., 1998. The anatomy of word-specific memory. In: Hulme, C., Joshi, R.M. (Eds.), Reading and Spelling: Development and Disorders. Erlbaum, Mahwah, NJ, pp. 405–419.

Bernstein, S.E., Treiman, R., 2001. Learning a novel grapheme: effects of positional and orthographic context on children's spelling. Journal of Experimental Child Psychology 79, 56–77.

Caravolas, M., Hulme, C., Snowling, M.J., 2001. The foundations of spelling ability: evidence from a 3-year longitudinal study. Journal of Memory & Language 45 (4), 751–774.

Caravolas, M., Kessler, B., Hulme, C., Snowling, M., 2005a. Effects of orthographic consistency, frequency, and letter knowledge on children's vowel spelling development. Journal of Experimental Child Psychology 92, 307–321.

Caravolas, M., Volin, J., Hulme, C., 2005b. Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: evidence from Czech and English children. Journal of Experimental Child Psychology 92 (2), 107–139.

Casalis, S., Deacon, S.H., Pacton, S. How specific is the connection between morphological awareness and spelling? A study of French children. Applied Psycholinguistics, in press.

Cassar, M., Treiman, R., 1997. The beginnings of orthographic knowledge: children's knowledge of double letters in words. Journal of Educational Psychology 89, 631–644.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, Massachusetts.

Cunningham, A.E., 2006. Accounting for children's orthographic learning while reading text: do children self-teach? Journal of Experimental Child Psychology 95, 56–77.

Deacon, S.H., Bryant, P., 2006. This turnip's not for turning: children's morphological awareness and their use of root morphemes in spelling. British Journal of Developmental Psychology 24, 567–575.

Deacon, S.H., Pacton, S., 2006. An empirical test of rules versus statistics: children's learning of spelling conventions (Abstract). Canadian Psychologist 47, 183.

Ehri, L.C., 1986. Sources of difficulty in learning to spell and read. Advances in Developmental & Behavioral Pediatrics 7, 121–195.

Fayol, M., Largy, P., Lemaire, P., 1994. When cognitive overload enhances subject-verb agreement errors. A study in French written language. Quarterly Journal of Experimental Psychology 47A, 437–464.

Flesch, R., 1955. Why Johnny Can't Read – and What You Can Do about It. Harper, New York.

Flesch, R., 1981. Why Johnny Still Can't Read: A New Look at the Scandal of Our Schools. Harper & Row, New York.

Frith, U., 1980. Cognitive Processes in Spelling. Academic Press, London.

Gentry, J.R., 1982. An analysis of developmental spelling in GNYS AT WRK. The Reading Teacher 36, 192–200.

Hayes, H., Treiman, R., Kessler, B., 2006. Children use vowels to help them spell consonants. Journal of Experimental Child Psychology 94, 27–42.

Kemp, N., Bryant, P., 2003. Do beez buzz? Rule-based and frequency-based knowledge in learning to spell plural -s. Child Development 74, 63–74.

Kessler, B., Treiman, R., June 2004. Sensitivity to Statistical Contextual Patterns When Spelling Consonants. Paper Presented at the Meeting of the Society for the Scientific Studies of Reading. Amsterdam.

Largy, P., Fayol, M., Lemaire, P., 1996. The homophone effect in written French: the case of verb-noun inflection errors. Language & Cognitive Processes 11, 217–255.

Lehtonen, A., 2006. Sources of information children use in learning to spell: the case of finnish geminates. In: Joshi, R.M., Aaron, P.G. (Eds.), Handbook of Orthography and Literacy. Lawrence Erlbaum, New Jersey, pp. 63–80.

Lété, B., Peereman, R., Fayol, M., 2008. Phoneme-to-grapheme consistency and word-frequency effects on spelling among first- to fifth-grade french children: a regression-based study. Journal of Memory & Language 58, 952–977.

Levelt, W., 2012. A History of Psycholinguistics: The Pre-Chomskyan Era. Oxford University Press, Oxford.

Martinet, C., Valdois, S., Fayol, M., 2004. Lexical orthographic knowledge develops from the beginning of reading acquisition. Cognition 91, B11–B22.

Pacton, S., Deacon, S.H., 2008. The timing and mechanisms of children's use of morphological information in spelling: a review of evidence from English and French. Cognitive Development 23, 339–359.

Pacton, S., Fayol, M., Perruchet, P., 2005. Children's implicit learning of graphotactic and morphological regularities. Child Development 76, 324–339.

Pacton, S., Perruchet, P., Fayol, M., Cleeremans, A., 2001. Implicit learning out of the lab: the case of orthographic regularities. Journal of Experimental Psychology: General 130, 401–426.

Perry, C., Ziegler, J.C., Coltheart, M., 2002. How predictable is spelling? Developing and testing metrics of phoneme–grapheme contingency. Quarterly Journal of Experimental Psychology A 55, 897–915.

Sandra, D., 2010. Homophone dominance at the whole-word and sub-word levels: spelling errors suggest full-form storage of regularly inflected verb forms. Language & Speech 53, 405–444.

Sandra, D. Persistent spelling errors shed light on how the human mind works during writing. In: Ravid, D. (Ed.), Proceedings of the Annual Meeting of the Hebrew Society for Linguistics, in press.

Sandra, D., van Abbenyen, L., 2009. Frequency and analogical effects in the spelling of full-form and sublexical homophonous patterns by 12 year-old children. The Mental Lexicon 4, 239–274.

Sandra, D., Fayol, M., 2003. Spelling errors with a view on the mental lexicon: frequency and proximity effects in misspelling homophonous regular verb forms in Dutch and French. In: Baayen, H., Schreuder, R. (Eds.), Morphological Structure in Language Processing (Trends in Linguistics). De Gruyter, The Hague, pp. 485–514.

Sandra, D., Frisson, S., Daems, F., 1999. Why simple verb forms can be so difficult to spell: the influence of homophone frequency and distance in Dutch. Brain & Language 68, 277–283.

Sénéchal, M., 2000. Morphological effects in children's spelling of French words. Canadian Journal of Experimental Psychology 54, 76–85.

Sénéchal, M., Basque, M.T., Leclaire, T., 2006. Morphological knowledge as revealed in children's spelling accuracy and reports of spelling strategies. Journal of Experimental Child Psychology 95, 231–254.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643–662.

Treiman, R., 1993. Beginning to Spell: A Study of First-grade Children. Oxford University Press, New York.

Treiman, R., Cassar, M., 1996. Effects of morphology on children's spelling of final consonant clusters. Journal of Experimental Child Psychology 63, 141–170.

Treiman, R., Cassar, M., Zukowski, A., 1994. What types of linguistic information do children use in spelling? The case of flaps. Child Development 65, 1318–1337.

Treiman, R., Kessler, B., 2006. Spelling as statistical learning: using consonantal context to spell vowels. Journal of Educational Psychology 98, 642–652.

Treiman, R., Kessler, B., Bick, S., 2002. Context sensitivity in the spelling of English vowels. Journal of Memory & Language 47, 448–468.

Spencer, Herbert (1820–1903)

Jonathan H Turner, University of California, Riverside, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14894–14898, © 2001, Elsevier Ltd.

Abstract

Herbert Spencer was the best-read sociologist of the late nineteenth and early twentieth centuries. He wrote major treatises on biology, psychology, sociology, and ethics as part of his lifetime project of forging a *Synthetic Philosophy* that would integrate science and ethics under a law of evolution that was loosely built from ideas in physics. His sociological work emphasized developing laws of human organization as societies differentiate in response to growth. He illustrated these laws with cross-cultural and comparative data. In developing his laws of societal evolution, Spencer founded functionalism, a school of thought that analyzes how basic needs for survival generate selection pressures for structures operating along four fundamental dimensions of societal existence: production, reproduction, regulation (coordination through power), and distribution (of information, people and resources). Spencer's ideas continue to inform modern sociology, but often the origin of these ideas as coming from Spencer is lost.

Herbert Spencer was born on 27 April 1820 in Derby, UK, and after a long and distinguished career as a major intellectual, he died on 8 December 1903 in Brighton, UK. During Spencer's childhood, he was privately tutored, first by his father and then, after the age of 13, by his uncle in Bath. Thus, Spencer received little formal education outside of his family, although his in-depth tutelage in mathematics and science was sufficient for him to contemplate as a relatively young man a grand Synthetic Philosophy in which a few general or 'first principles' (1862) could be used to explain the operative dynamics of physical science, biology (1864–67), psychology (1855), sociology (1874–96), and ethics (1892–98). He also was a moral crusader, writing important philosophical statements on such topics as education (1860) and social wellbeing (1850, 1892–98).

Spencer's sociology comes late in his career, beginning with a methodological treatise on sources of bias (Spencer, 1873), continuing with an effort to systematically array cross-cultural and comparative data (Spencer, 1873–34), and culminating with the use of data to illustrate the basic laws of sociology (Spencer, 1898b [1874–96]). In this article, the goal is to see how Spencer's grand scheme developed and how it evolved into a profoundly important sociology (for a comprehensive listing of primary works by Spencer and of works on Spencer, see Perrin (1993)).

Biographical Influences on Spencer's Sociology

Several important aspects of Spencer's biography are important in examining his impact on social science. First, he never held an academic position, and hence, he did not have any students to carry forth his work after his death, although the wealth inherited from his uncle and from his book sales did allow him to commission other academics to perform research and, even after his death, to continue to completion the volumes of his *Descriptive Sociology* (Spencer, 1873–34). Second, he lived an extremely spartan life, and yet at the same time, he interacted in London Clubs and various assemblies with the elite scientists and intellectuals of his time. Indeed, it is from these scholars – figures such as Thomas Huxley, Joseph Hooker, John Tyndall, and even Charles Darwin – that Spencer absorbed much of his knowledge about current discoveries in science. Third, Spencer was immensely popular in his time. His works generally came out in serial form in magazines and private subscriptions, later being packaged as whole books which sold more than 400 000 copies during Spencer's lifetime. Yet, despite his enormous popularity, Spencer was an enigma because he had no formal academic affiliation and because he lived a very private and, it appears, rather neurotic life punctuated by hypochondria. Yet, in all his works, several key influences are evident. Reading Thomas Malthus, Karl Erst von Baer, and Charles Darwin gave Spencer a view of the universe as evolving from simple to more complex through competition; extensive reading of the physical sciences gave Spencer a post-Newtonian conception of science, a view emphasizing that universal laws could explain all domains of the universe; and reading of Auguste Comte also gave Spencer a view of sociology as the search for general laws as well as a conception of social organisms as having structures devoted to meeting basic needs and of societal evolution as involving an increasing complexity of structures meeting these needs. Spencer (1968 [1864]) was, however, always rather defensive about perceptions of Comte's influence on him. Out of this mix of influences came Spencer's grand Synthetic Philosophy that was to unify physical science, psychology, biology, sociology, and ethics under first principles borrowed from the physics of Spencer's time.

The Stigma of Spencer's Moral Philosophy

Spencer's moral philosophy is one reason that his intellectual star began to fade in the early twentieth century. In his first major work, *Social Statics* (1850), almost a decade before Darwin published *On the Origin of Species*, Spencer coined the phrase 'the survival of the fittest' that was to haunt him for the rest of his career. The more general philosophy in this work simply argued that individuals should be free to pursue their happiness, as long as they do not abridge the rights of others to do the same, although there was a hard edge to this philosophy emphasizing that the less fit should not be supported by the state. Spencer (1904) complained bitterly in his autobiography

about how he had become type cast as a result of this first book, but it is hard to feel sorry for him because he made the same 'libertarian' (in today's vocabulary) statements in his last work, *The Principles of Ethics*. For a discipline such as sociology, which is liberal in the twentieth century as opposed to the nineteenth century sense where freedom of the individual and lack of governmental activism are emphasized, this libertarian-sounding ideology has always worked to obscure the important contributions of Spencer to sociology.

The 'First' or 'Cardinal' Principles

Spencer's use of the physics of his time to formulate his 'cardinal principles' is now rather dated. Spencer postulated three basic principles: (1) the indestructibility of matter, (2) the continuity of motion in a given direction, and (3) the persistence of force behind movement of matter. From these general principles, three corollaries were derived: (4) the transferability of force from one type of matter and motion to another, (5) the tendency of motion to pass along the line of least resistance, and (6) the rhythmic nature of motion. From these principles, Spencer (1880[1862], p. 286) formulated his famous law of evolution as 'change from a less coherent form to a more coherent form, consequently on the dissipation of motion and integration of matter.' This law was presumed to apply to the evolution of celestial bodies, chemical compounds, organisms, and superorganic bodies (organization of organisms). Spencer was, therefore, the world's first general systems theorist.

The first or cardinal principles are less interesting than what Spencer did with them as he applied them to organic and superorganic evolution. In *Principles of Biology*, Spencer used the first principles to formulate the cornerstone of his biology and sociology: increasing size of an organic or superorganic mass requires an exponential differentiation of the structure necessary to support the larger mass. This principle is still used in biology, although its origins to Spencer are long forgotten; the principle is employed in sociology to study differentiation at all levels of social organization, with only a vague sense that this idea comes from Spencer. In many ways, this theoretical cornerstone of Spencer's Synthetic Philosophy actually comes more from his training and brief practice as an engineer than it does from his first principles; for within engineering it is well established that larger structures require more complex systems of support.

Spencer's star faded long ago in biology, and there is only a hint of his influence in education and philosophy today. Where Spencer still remains a figure deserving of at least mention, often in a somewhat derisive way, is in sociology. In order to appreciate what Spencer's legacy is today, we must review his contributions to sociology, where he still lives on as a marginal figure in sociology's pantheon of great masters that includes such giants as Karl Marx, Max Weber, Émile Durkheim, and George Herbert Mead.

Spencer's Functionalism

Spencer is famous for his 'organismic analogy' in which he lists in a more formal manner than Comte before him the parallel principles of organization between organic and superorganic bodies. But this analogy is often given too much emphasis in commentaries on Spencer because it hides a rather sophisticated functionalism. For Spencer, all superorganic systems composed of organic units must meet three basic requisites if they are to survive: (1) operation or (a) production of life-sustaining substances and (b) reproduction of units; (2) regulation or coordination and control of units; and (3) distribution or movement of necessary information and substances among units. Populations which cannot meet these requisites will experience 'dissolution,' a notion first developed in *First Principles* (1880 [1862]). Thus, Spencer did not view evolution as an inevitable process of movement from simple to complex as populations sought to resolve problems of operation, regulation, and distribution; on the contrary, he recognized that societies often disintegrate. In this discussion of functional requisites, Spencer introduces two conceptions of 'selection.' On the one hand, Spencer argued that evolution of societies tends to increase complexity because, as societies come into conflict, the more complex and better organized generally will prevail, thereby ratcheting up the overall level of societal complexity. This argument is one of the first 'group selectionist' statements presented in either biology or the social sciences. On the other hand, there is another kind of selection which is at the core of all functionalist arguments. This selection occurs alongside Darwinian selection and revolves around problems and crises in production, reproduction, regulation, and distribution. When problems in any of these spheres emerge, they set into motion efforts to search for solutions by goal-seeking social units. Thus, social selection in superorganic systems is not blind because the very units subject to selection can change their structure and thereby avoid being selected out. Spencer argues that most selection in sociocultural systems is of this second type, where problems of adaptation activate efforts to resolve crises with respect to production, reproduction, regulation, and distribution.

Since these requisites are the essential problems of survival in all superorganic systems, including both human and nonhuman societies, differentiation of structure during the course of evolution will revolve around these four axes. First will come differentiation of productive structures, followed by differentiation of reproductive units. This increased differentiation along these axes will generate selection pressures for their coordination and control, leading to the elaboration of political systems. As differentiation of operative and regulatory units occurs, new systems of distribution become necessary if dissolution is to be avoided. As these new systems of distribution emerge, e.g., markets, communication, and transportation infrastructures, they create new selection pressures for their regulation while at the same time providing positive feedback for further differentiation of production and, later, reproduction. Thus, movement of superorganic systems from simple to complex forms revolves around selection pressures along the axes of operation, regulation, and distribution as well as around the mutual feedforward and feedback effects of differentiation along one axis on the other axes. This kind of analysis is far more sophisticated than the typical commentary on functionalism acknowledges.

Spencer's substantive sociology also offers some very important models of system dynamics. One element of these models follows from the basic law: as a population grows, the

increasing size generates escalated logistical loads for operation, regulation, and distribution which, in turn, set off selection pressures in the manner described above. Spencer's political sociology should be viewed in this context. For Spencer, once power is concentrated, this initial concentration will be used to gain even more regulatory control. As power becomes more concentrated, inequalities will escalate as those with power extract an ever-greater proportion of a population's resources. As inequalities increase, disintegrative pressures mount leading to the concentration of more power to suppress the potential conflict that always inheres in inequality; and as this process is iterated, system dissolution is likely. Geopolitics are also involved in these dynamics, in several senses. First, elites in the regulatory sector often engage in war to secure the resources of other populations which only increases the concentration of power and the scale and scope of inequalities. Second, such war-making efforts often distract subordinates in a system of inequality, at least for a time, from their true interests in changing the system of power generating inequality. Third, when power becomes too concentrated, resistance to this power inevitably increases, often leading to protest movements and other actions that lead to some deconcentration of power. Yet, the longer deconcentrated power lasts, problems of coordination and control escalate, creating selection pressures for reconsolidation of power. Societies thus cycle between concentrated profiles of power (what Spencer termed the 'militant') and deconcentrated power (what Spencer called 'industrial' profiles) – a line of analysis that predates Vilfredo Pareto's discussion of the circulation of elites.

The Data Behind Spencer's Sociology

These lines of argument have a very contemporary ring to them, and they can be found throughout Spencer's *Principles of Sociology*. Moreover, they are rather easily converted into laws or principles since this is how Spencer thought (see Turner, 1985, for an effort to extract the principles from Spencer's work). One of the criticisms of Spencer's functionalism is that it is so speculative; while this may be so, Spencer also amassed enormous amounts of data to support his theoretical speculations. Indeed, the reason that Spencer's *Principles of Sociology* is such a long work is that it is filled with data. In fact, many scholars complained at its first publication in serial form and later as a three-volume book that the analytical ideas get lost in all the data. Thus, Spencer assembled data far in excess of almost any other sociologist in the history of the discipline. He hired professional academics to gather historical and ethnographic information, ordering it into categories that followed from his general functionalist scheme. There are many tragedies in the contemporary view of Spencer, but none is greater than the loss of what Spencer began in the 1860s and first published in 1873 as the inaugural volume of his *Descriptive Sociology* (the last volume was published in 1934).

Spencer was the first to execute what was later done by those assembling the Human Relations Area Files (HRAF) (e.g., Murdock 1953). Spencer recognized that in order to make generalizations about human societies, a large comparative database was necessary. But these data had to be arrayed so that they could be compared, leading Spencer to develop a category system for recording the data being assembled by professional academics. The result was the multivolume *Descriptive Sociology* (1873–34) which arrays data historically for those societies that have a written history and cross-sectionally for those less developed societies without a written history. The publication of these oversized volumes has been completely lost to modern social science, but if Spencer's lead had been followed by others in the nineteenth century, all social science would have a much firmer empirical base than is provided by HRAF. At the very time Spencer was assembling available information on 'primitive' societies, these societies were being changed or destroyed by colonial forces, with the result that by the time anthropologists began to study them in detail in the 1930s, their pristine character had long been compromised. *Descriptive Sociology* could have been Spencer's greatest legacy, but like so much of his work, the contemporary world remains largely ignorant of his efforts to create a database for assessing the plausibility of his theories.

What Might have Been, and What Can Still Be

By the time of Spencer's death in 1903, his star had already begun to fade in intellectual circles. One reason for this decline can be attributed to Spencer as a lone scholar working outside of academia. Without students to carry on his work and without a school of thought institutionalized in academia, there was no one and no place to carry forward Spencerian sociology. Work still continued on *Descriptive Sociology* but this was much narrower in scope than Spencer's philosophy or even his sociology, and moreover, most of this work was contracted to scholars who were not committed to Spencer's grand scheme. So, when Talcott Parsons (1937) asked 'who now reads Spencer,' the answer was not very many scholars inside or outside of sociology. Spencer's star had fallen at the same time that evolutionary thinking came under attack in anthropology, and with no one to defend Spencer's sociology, he fell into obscurity. Yet, let us imagine for a moment what might have happened if Spencer had been defended by students or had established a school entrenched in academia like Durkheim's *année* school in France. In conducting this thought experiment, we can come to appreciate not only what might have been, but more fundamentally, what is still available in Spencer's legacy.

The first legacy would have been a Human Relations Area Files that came earlier, before colonialism completely destroyed indigenous populations. Moreover, sociology would have had a database that was comparative in several senses: the data would have been arrayed in common categories that are more sophisticated than George P. Murdock's (1953) HRAF; the data would have been both historical and ethnographic and would have allowed systematic comparisons of different types of societies at any point in time. Only in the last three decades has comparative-historical and ethnographic data been reintroduced back into sociology and, ironically, this reintroduction has been part of a movement to rekindle interest in the very evolutionary thinking that was used to discredit Spencer.

The second legacy is that the evolutionary paradigm would have survived and avoided the 50-year hiatus. Stage models have utility in sociological analysis, as has been demonstrated in the work of Gerhard Lenski (1966), Stephen K. Sanderson (1995), Jonathan Turner (1984, 1997), and others in recent

years. Moreover, adopting Spencer's views on evolution would have avoided the view of evolution as inexorable, since dissolution would be a prominent point of analysis as would dialectical cycles within either evolution or dissolution.

A third legacy would have been a sophisticated functionalism that, unlike mid-twentieth century approaches, sought to articulate theoretical principles about the dynamics of social processes rather than simply cross-tabulate lists of functional requisites with lists of social processes. Functionalism fell into disuse not so much because it was dead wrong but because it was not done well. A reading of Spencer would have led the way to a functionalism that does not build conceptual schemes as much as highlight fundamental forces whose operative dynamics can be captured in theoretical principles.

A fourth legacy would have been a sociology that did not lose sight of its status as a science in which the goal is to generate theoretical principles that explain the operative dynamics of the social universe. Spencer's methodological treatise and his articulation of actual laws of human organization would have sustained the discipline better than Marx, Weber, Simmel, or Durkheim could, as various antiscience and excessively relativistic movements have come and gone for the whole of the twentieth century in sociology.

A fifth legacy would have been the actual substance of Spencerian sociology. Imagine if sociologists had mined *The Principles of Sociology* for insights into human organization with the same intensity as we have delved into Marx, Durkheim, Weber, or Mead. Spencer provided not only highly abstract theoretical principles but also many principles about the operation of stratification systems, polity, kinship, religion, economy, markets, ceremonies, professions, organizations, education, and many other key structures, and he documented these generalizations with copious amounts of historical, ethnographic, and for his time, contemporary data. The sociology of major institutions (e.g., family, education, religion, etc.) would be less parochial and descriptive than it is today because it would examine the dynamics of the institution with ethnographic as well as historical-comparative data organized by basic theoretical principles.

A sixth legacy would have been the continued connection between biology and sociology. This connection has only recently been revived along all the fronts found in Spencer's work. There is the ecological approach emphasizing Darwinian dynamics that Spencer saw before Darwin and that are now used to examine the relations between resources niches and a wide variety of social processes. Indeed, there is the revival of selection arguments in general (e.g., Runciman, 1998), and moreover, an effort to move beyond Darwinian selection to considerations of group selection (e.g., Sober and Wilson, 1998) and Lamarckian processes so evident in Spencerian sociology (Turner, 1995).

A seventh legacy is a political sociology that connects stratification processes within society to geopolitical dynamics in world systems. At the core of Spencer's sociology is a concern with power that is the equal of that to be found in either Marx or Weber. One of Spencer's main explanatory variables is how the concentration of power alters all other social processes, especially stratification and institutional systems, and his analysis of how war and commerce in the world system are implicated in power dynamics would not only have supplemented Marx and Weber but would have led to the emergence of world systems theory at the beginning rather than at the three-quarters mark of the twentieth century (e.g., Wallerstein, 1974).

An eighth legacy is the emphasis on the fundamental relation between size, differentiation, and modes of integration. Much of this legacy was rediscovered in the mid-twentieth century in organizational studies on administrative intensity, but the basic dynamic operates at all levels of society. If sociologists had pursued Spencer's lead and focused in 1910 on this fundamental relationship in the social universe, not only would organizational studies been moved forward at an earlier point in history but also other social units might well have been analyzed for how these relations among size, differentiation, and integration play themselves out.

These eight lines of thinking that are part of Spencerian sociology were available to all sociologists by the early 1890s. In a sense we have squandered this incredible legacy, but it is not too late. The above thought experiment can be more than an intellectual exercise in the history of ideas. A reading of Spencer today can still inform sociology. There are gems of insight to be discovered and used to analyze societies in almost everything Spencer wrote. They simply sit there on library shelves, waiting to be used by scholars in the twenty-first century. There is probably little more to be gleaned from poring over the texts of others in sociology's early pantheon, but the neglect of Spencer for the whole of the twentieth century provides a rich resource for the twenty-first century.

See also: Conflict: Anthropological Aspects; Conflict: Organizational; Cross-Cultural Research Methods in Psychology; Cultural Evolution: Theory and Models; Development: Organizational; Development: Social-Anthropological Aspects; Elites, Anthropology of; Elites: Sociological Aspects; Ethics and Values; Evolution, History of; Evolution, Natural and Social: Philosophical Aspects; Groups, Sociology of; Hierarchies and Markets; Motivational Development, Systems Theory of; Religion, Family, and Kinship; Religion: Morality and Social Control; World Systems Theory.

Bibliography

Lenski, G.E., 1966. Power and Privilege: A Theory of Social Stratification. McGraw-Hill, New York.

Murdock, G.P., 1953. Social structure. In: Tax, S., Eisely, L., Rouse, I., Voegelin, C. (Eds.), An Appraisal of Anthropology Today. University of Chicago Press, Chicago.

Parsons, T., 1937. The Structure of Social Action. McGraw-Hill, New York.

Perrin, R.G., 1993. Herbert Spencer: A Primary and Secondary Bibliography. Garland Publishers, New York.

Runciman, W.G., 1998. The selectionist paradigm and its implications for sociology. Sociology 32, 163–188.

Sanderson, S.K., 1995. Social Transformations: A General History of Historical Development. Blackwell, Cambridge, MA.

Sober, E.D., Wilson, S., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

Spencer, H., 1873. The Study of Sociology. Kegan Paul, London.
Spencer, H., 1873–1934. Descriptive Sociology, or Groups of Sociological Facts. D. Appleton, New York.
Spencer, H., 1880. First Principles. A. L. Burt, New York [1862].
Spencer, H., 1887. The Principles of Biology. vols. 2. D. Appleton, New York. originally published 1864–67.
Spencer, H., 1888. Education: Intellectual, Moral, and Physical. D. Appleton, New York. originally published in 1861.
Spencer, H., 1892. Social Statics. D. Appleton, New York. originally published in 1850.
Spencer, H., 1892–98. The Principles of Ethics. D. Appleton, New York.
Spencer, H., 1898a. The Principles of Psychology. vols. 2. D. Appleton, New York. originally published in 1855.
Spencer, H., 1898b. The Principles of Sociology. vols. 3. D. Appleton, New York. originally published between 1874–96 [1874–96].
Spencer, H., 1904. An Autobiography. Vols. 2. Appleton, New York.
Spencer, H., 1968. [1864] Reasons for Dissenting from The Philosophy of M. Comte and Other Essays. Glendessary Press, Berkeley, CA. originally published in 1864.
Turner, J.H., 1984. Societal Stratification: A Theoretical Analysis. Columbia University Press, New York.
Turner, J.H., 1985. Herbert Spencer: A Renewed Appreciation. Sage, Beverly Hills, CA.
Turner, J.H., 1995. Macrodynamics: Toward a Theory on the Organization of Human Populations. Rutgers University Press, New Brunswick, NJ.
Turner, J.H., 1997. The Institutional Order. Longman, New York.
Wallerstein, I., 1974. The Modern World System. vol. 1. Academic Press, New York.

Sperry, Roger Walcott (1913–94)

Theodore J Voneida, Northeastern Ohio Universities College of Medicine, Rootstown, OH, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14899–14901, © 2001, Elsevier Ltd.

Abstract

This article reviews the major scientific contributions of Nobel Laureate Roger W. Sperry, whose collective studies revolutionized our understanding of brain function. His early work, beginning in the 1940s, culminated with the publication of his *Chemoaffinity Theory* in the mid-1960s. This theory holds that neural networks are determined by a genetically controlled chemical code. The 'hard-wired' nature of this network results in a specific, highly adaptive range of behavior, which is unalterable by training. His next series of studies related to the anatomic basis of neocortical information processing. He was able to show, in several simple, yet ingenious experiments, that perception is dependent upon vertically oriented cortical columns. The work, which resulted in the Nobel Prize (1981), relates to the function of connections between the 2 brain halves, and the lateralization of cognitive abilities in humans. These studies led to his final major contribution, which holds that consciousness is an 'emergent state' resulting from, but greater than, the sum of all the neuronal activity which produced it. This conscious state – the mind – continuously feeds back onto the brain by way of 'downward causation,' resulting in a continuous, highly dynamic process of change in neuronal activity from one moment to the next.

Roger W. Sperry, scientist, teacher, philosopher, and humanitarian was born in Hartford, Connecticut, USA on August 20, 1913. He was married to Norma Dupree in 1949; they had 2 children. He died on April 17, 1994.

The diversity of Sperry's interests is not only reflected in his distinguished academic career, but also in his avocations. In addition to his major contributions to our understanding of brain structure and function, he was an unusually talented artist, sculptor, and paleontologist. His multiple talents across many areas attracted a constant stream of graduate students, postdoctoral fellows, and visiting scientists to his Psychobiology Unit at Caltech for four decades.

Sperry's scientific contributions centered on four major areas, the first of–which began in the early 1940s, while he was still a graduate student in Paul Weiss' laboratory at the University of Chicago. Paul Weiss, whose views dominated the field of neural development at the time, held that the leading tips of growing neurons solely responded to local mechanical stimuli, and were in no way influenced by distant chemical cues. His numerous tissue culture studies had essentially closed the door on any possibility of chemical influences on neuronal growth. Weiss also held that the normal function of the nervous system was in no way dependent upon the precise patterns of neuronal connections. One example of this, according to him, was that when motor nerves were transplanted into any foreign muscle, normal movements would eventually occur after a period of training.

Sperry, as was to become his custom, challenged these ideas, and designed a typically simple, but ingenious experiment to test whether or not specific nerve connections are critical for normal function. Using rats, he transplanted motor nerves from leg muscles of one side to muscles of the opposite side, then tested their movements after the new nerve-muscle connections were established. Function was found to be severely impaired, and no amount of training could correct it (Sperry, 1945). These results also had important implications for humans. Neurosurgeons at the time were transplanting nerves in wounded war veterans in attempts to help them regain digit and limb movements, which had been lost as a result of war injuries. In spite of extensive post-surgical training, these transplants met with little success. During his military service with the Office of Scientific Research and Development (1942–45), Sperry, along with Weiss, showed that surgical transplantation of nerves to antagonistic muscle groups followed by intensive training simply did not work because transplanted nerves were inherently different from the nerves which normally innervated those muscles. This finding resulted in significant modifications of treatment procedures, thus saving a great deal of time, frustration, and money.

During this same period it was reported that if the optic nerve were sectioned in amphibians, optic axons would regenerate into the brain, and normal function would be restored. Sperry immediately recognized that this finding offered the perfect opportunity to test his ideas regarding functional specificity and the possible role of chemical influences on neuronal growth patterns. He first confirmed the above results (Sperry, 1950), then once again designed a brilliantly simple study – the now classic eye rotation experiment in the frog – to test his ideas. A hungry frog will accurately strike out with its tongue to obtain a fly held in front of it. Sperry sectioned the optic nerves of normal frogs, and rotated the eyeballs by 180 degrees. After regeneration had taken place and connections were reestablished in that part of the brain which is known to receive visual fiber input (an area called the optic tectum), he tested the frog's ability to capture a fly. The premise was that if the retinal axons had regrown into their normal sites of termination (in spite of the fact that the eye had been left/right reversed and turned upside-down), the frog should misdirect its strike by 180 degrees in any direction. If, however, the regenerating optic fibers disregarded their original target sites and established new central connections corresponding to the rotated eyes, the frog would be capable of striking accurately at the target fly. The answer was unequivocal. When the fly was held 90 degrees to the frog's right, it struck out 90 degrees to its left–180 degrees out of phase. The same result was obtained for all directions – left, right, up, down. The strike was always 180 degrees off, and no amount of retraining could correct it. The conclusion from both of the above lines of evidence was

clear: peripheral nerves cannot be 'forced' to establish connections with foreign muscles, and centrally regenerating nerves will reconnect to their original sites, in spite of radical alterations of their starting positions. A number of other rerouting studies were carried out, each of them supporting the basic premise, which Sperry termed 'the chemoaffinity hypothesis.' Briefly stated, the growth and development of a pre-patterned neural network is determined by a genetically controlled chemical code, and the 'hard-wired' nature of this network results in a specific, highly adaptive range of behavior which is unalterable by training. Other studies over the next 20 years, using neuroanatomical and neurophysiological approaches, provided confirmation of this hypothesis. Attardi and Sperry (1963) provided further, histological confirmation by showing that regenerating retinal axons in the goldfish would always regrow to their original terminal sites in the brain.

Two major theoretical papers were published by Sperry in the mid-1960s (see Voneida, 1997; Sperry, 1965a), in which he precisely described his chemoaffinity theory. In spite of numerous attempts over the next three decades to disprove the theory, it has prevailed with only slight variations and elaboration of the basic premise. It was primarily this work, begun as a predoctoral student in 1938 and pursued through the early 1960s that resulted in his election to the National Academy of Sciences.

A second major area of Sperry's work deals with cortical organization related to higher cognitive function in mammals. He joined the renowned psychobiologist Karl Lashley at Harvard for his postdoctoral work, where he began a new series of studies on the role of electrical fields in neocortical function. He was not comfortable with the idea that electrical fields or waves acting in a volume conductor (the brain) are critical for neocortical processing. His first study to challenge this was published in the late 1940s (see Sperry, 1947). Here he demonstrated that motor coordination in monkeys remained virtually unaffected after multiple transections of sensorimotor cortex. He later confirmed this by demonstrating that neither cortical cuts, the insertion of numerous short circuiting wires or insulating plates into the cortex had any adverse effect on cortical function. These studies demonstrated that perception is dependent upon vertically oriented afferent and efferent cortical axons, predating the discovery of vertically oriented cortical columns. Sperry's premise was based on his keen understanding of neuroanatomy and neurophysiology, including the work of Santiago Ramon y Cajal and Lorente' de No, both of whom had demonstrated the importance of radial cortical connections to cortical function.

Thus, in a few carefully conducted experiments, Sperry had once again upset two major theories of brain function, Gestalt electric field theory of perception and the reduplicated interference pattern hypothesis. Indeed, when the renowned neuroembryologist Viktor Hamburger presented Sperry with the Ralph Gerard Award from the Society of Neuroscience in 1979, he proclaimed, 'I know of nobody else who has disposed of cherished ideas of both his doctoral and his postdoctoral sponsor, both at that time the acknowledged leaders in their fields.'

Sperry's third major contribution relates to the function of neuronal connections between the two brain halves (called commissures), interhemispheric communication, and lateralization of cognitive abilities in humans. Most of this work was done at Caltech, where he served as Hixon Professor of Psychobiology from 1954 to 1994. It was these studies, carried out between the mid-1950s through the 1970s (see Sperry, 1981), which resulted in the Nobel Prize in Physiology and Medicine (shared with Hubel and Wiesel) in 1981. The early studies, carried out by Sperry and his students at Caltech from the mid-1950s through to the 1960s, elegantly elucidated some of the major functions of interhemispheric connections in memory transfer and eye-hand coordination. Restriction of sensory input to one brain hemisphere in animals having undergone surgical separation of the two brain halves was shown to limit the learning of various tasks to the 'input' hemisphere; the opposite side was capable of learning, but remained naive to those tasks until trained. Learning curves for each brain half were virtually identical; it was as if two separate brains were housed within a single cranial vault. It was during this period that he also began his work on the lateralization of cerebral function in humans. Neurosurgeons Philip Vogel and Joseph Bogen had performed a surgical section of the major pathway connecting the two brain halves on a World War II veteran who was suffering from progressively worsening epileptic seizures. Working in collaboration with Vogel and Bogen, Sperry and his students began a series of tests directed at understanding the effects of this surgical hemispheric separation on human perception, speech, and motor control. The work involved comparisons of cognitive abilities between the two separated brain halves, demonstrating differences theretofore unrecognized.

For the next 20 years, the work of Sperry and his collaborators revolutionized our understanding of brain function. Prior to this time, the right hemisphere had been virtually neglected by neuroscientists. At best, it was felt to play only a minor role in human cognition, with all major cognitive functions emanating from the left, or dominant half. As a result of Sperry's research, the right hemisphere gained its long overdue recognition as a full partner to the left, playing a major role in holistic, parallel, and spatial abilities. The left brain half was found to be superior to the right in tasks involving analytical, sequential, and linguistic processing. Perhaps most importantly, however, Sperry demonstrated that the combined effects of bihemispheric activity amounted to much more than the simple additive effects of the two separate hemispheres, and led him into the area of his fourth and final major contribution; namely, the relationship between brain and mind, science and values.

Sperry began thinking and writing about the subjective aspects of brain function as early as the late 1950s. In 1959, for example, as part of the discussion at the Josiah Macy Conference on the Central Nervous System and Behavior, he stated that he had never been entirely satisfied with the behaviorists' interpretation of brain function in purely objective terms, with no reference whatsoever to subjective experience. He pointed out that this does not mean that we should abandon the objective approach; it simply means that we should also include subjective experience in our working models of the brain. Again, in his 1965 paper, *Merging Mind, Brain, and Human Values* (see Sperry, 1965b), he reinforces this thesis in his declaration that conscious forces must be included in any model or description of brain function, stating that in this

model, '... conscious mental psychic forces are recognized to be the crowning achievement of some five hundred million years or more of evolution.'

Sperry continued to think and write about the mind, consciousness, and their relationship to human values right up to the end of his life. He postulated that the mind is an emergent property of brain function, and like most emergent properties, bears little resemblance to the hundreds of millions of neurons in the central nervous system. He emphasized that while the mind is not the *same* as the brain, it is *dependent* upon it. This point is an important one, for it separates Sperry's views from the philosophical concept of dualism. Thus brain and mind are by no means separate entities, but function rather as coexistant partners through the processes of emergence and downward causation, or feedback. Taken together, these two properties result in a continuous, highly dynamic process of change from one moment to the next, drawing constantly on previous experiences by way of learning and memory, whether they occurred in the distant past or only moments ago. He strongly argued against reductionism, which dominated psychological thinking during the first half of the twentieth century. He believed that reducing consciousness to its separate components obliterates the emergent properties of mind, with '... *all its great power and uniqueness*' (see Sperry, 1993).

Sperry then elevated his concept of mind as an emergent property of brain activity to a global level, stating that:

> The new paradigm affirms that the world we live in is driven not solely by mindless physical forces but more crucially, by subjective human values. Human values become the underlying key to world change (see Voneida 1997).

He urged his scientific colleagues to use the very unique and powerful forces of mind and consciousness toward improving and preserving the quality of life on our planet, rather than the reverse. His urgent appeal finally began to be heard when his long-term friend and colleague, Professor Rita Levi-Montalcini, herself a Nobel Laureate, convened the International Conference of Human Duties at the University of Trieste in November, 1992, to discuss these ideas in greater detail. Regrettably, Sperry's illness prevented him from attending the first meeting, which included 10 Nobel Laureates and numerous others, representing such widespread disciplines as neurobiology, chemistry, physics, economics, and theology. This conference led to the formation of The International Council of Human Duties, which formulated 'The Declaration of Human Duties' in 1994. The document has been published (Trieste Declaration, 1997) and forwarded to The United Nations, for its consideration as a corollary to the 'Declaration of Human Rights.'

Sperry's thinking about subjective experience, consciousness, the mind and human values were crystallized in his paper 'The Impact and Promise of the Cognitive Revolution' (Sperry, 1993), delivered at the American Psychological Association's centennial meeting in 1993. It expressed his great hope and sincere belief that if we humans could simply be persuaded to put our collective minds together and to use the enormous emergent power that they are capable of generating, we might very well begin the new millennium by not only improving the quality of life on the planet, but also by significantly increasing our chances for survival.

See also: Brain Asymmetry; Mind–Body Dualism.

Bibliography

Attardi, D.G., Sperry, R.W., 1963. Preferential selection of central pathways by regenerating optic fibers. Experimental Neurology 7, 46–64.

Sperry, R.W., 1945. The problem of central nervous reorganization after nerve regeneration and muscle transposition. Quarterly Review of Biology 20, 311–369.

Sperry, R.W., 1947. Cerebral regulation of motor coordination in monkeys following multiple transection of sensorimotor cortex. Journal of Neurophysiology 10, 275–294.

Sperry, R.W., 1950. Neural basis of the spontaneous optokinetic response produced by visual inversion. Journal of Comparative and Physiological Psychology 43, 482–489.

Sperry, R.W., 1952. Neurology and the mind-brain problem. American Scientist 40, 291–312.

Sperry, R.W., 1965a. Embryogenesis and behavioral nerve nets. In: Dehaan, R.L., Ursprung, H. (Eds.), Organogenesis. Holt, Rinehart and Winston, New York, pp. 161–185.

Sperry, R.W., 1965b. Mind, brain and humanist values. In: Platt, J.R. (Ed.), New Views of the Nature of Man. University of Chicago Press, Chicago, pp. 71–92.

Sperry, R.W., 1981. Some effects of disconnecting the cerebral hemispheres (Nobel lecture.). Almquist and Wiksell, Stockholm.

Sperry, R.W., 1983. Science and Moral Priority: Merging Mind, Brain, and Human Values. Columbia University Press, New York.

Sperry, R.W., 1993. The impact and promise of the cognitive revolution. American Psychologist 48 (3), 878–885.

Trieste Declaration of Human Duties, 1997. A Code of Ethics of Shared Responsibilities. Trieste University Press, Trieste, Italy.

Voneida, T.J., 1997. Roger Walcott Sperry, 1913–1994. The National Academy Press, Washington, DC.

Spirit Possession, Anthropology of

Ute Luig, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives a short overview of the main manifestations of spirit possession. It outlines previous theoretical approaches, like psychological, sociological, and hermeneutical interpretations, but focuses predominantly on newer developments. These are the important roles of spirit possession in coping with the consequences of war, especially in postsocialist countries as well as in the increasing exchange of knowledge and import of spirit performances between diaspora communities and their countries of origin.

Despite its prohibition by some colonial governments and processes of modernization and secularization, spirit possession remains a vital force in the religious landscape of African, Asian, and Latin American countries. It is at present on a high which refutes the once popular view of spirit possession as modernity's other, characterized as primitive and backward. Instead, spirit possession cults are creative agents of change mediating between persons, regions, and even continents and between space and time. Interesting forms of symbiosis and rivalry exist with all major world religions. Pentecostal churches especially have become severe competitors in the last 20 years due to the popularity of the Holy Spirit and a very extensive media marketing (Meyer, 2006).

Conceptualizing Possession

Despite many attempts to define spirit possession (Howard and Mageo, 2006; Dawson, 2011a), the wide varieties of beliefs, religious traditions, and new inventions make it difficult to define what possession is and what it means for different people. The most conclusive is still Boddy's view that possession "is a broad term referring to an integration of spirit and matter, force or power and corporeal reality, in a cosmos where the boundaries between an individual and her environment are acknowledged to be permeable, flexibly drawn, or at least negotiable" (Boddy, 1994: p. 422). The relationship between spirits and humans has been accentuated further by Lambek as the trajectory of the spirit from "the ethereal into the material world" (2010: p. 28). In this act, the spirit offers his presence to his host as a form of "poiesis, as something being made, brought into being, and thereby also bearing on the world into which it is born" (ibid., 2010: p. 31). The myriad embodiments of spirits as ancestor spirits, former kings, or slaves as well as of specific ethnic or social groups, as guardians of the land or animals of the bush to name but a few underline their great variety. They are thought of as distinct personalities with special tastes for food, clothes, jewelry, technical equipment, colors, rhythms, and music which have to be provided by their hosts. The disclosure of the spirits' identity is dependent on different forms of communications, like dreams, images, voices, or eventually glossolalia but also through a sudden attack of illness or misfortune. These signs are interpreted as the spirits' calling which either lead to their exorcism or to their integration into the self of the afflicted person. These encounters are acted out in ritual performances in which spirits, mediums, and the public interact through the recitation of words, music, and dancing. Trance or ecstasy mark the very moment of transformation in which spirit and host fuse their identities. The identification with the spirit opens new horizons for the possessed who are situated in the global flows of present and past images.

Discourses on Possession

The fascination with the subject on the part of scholars has generated a large body of literature. Whereas in the nineteenth and the beginning of the twentieth century European case studies played an important role in discussions of spirit possession, processes of secularization shifted the scholarly interest in spirit possession onto non-Western societies. From the beginning, theories of spirit possession reflected the mainstream discussions in anthropology and were thus heavily influenced by changing theoretical paradigms. In recent years, the once dominant psychological and sociological approaches have been replaced by a hermeneutical interpretation of spirit possession.

Psychological and Sociological Approaches

An important issue in early studies was the truth or falseness in the experience of possession which de Heusch (1962) classified as authentic or inauthentic possession. As a consequence, psychological interpretations (Bourguignon, 1973) loomed large associating possession with multiple personalities and an altered state of consciousness. It was described as a form of pathology due to personal stress or social change.

Firth's (1964: p. 247) distinction between spirit possession and spirit mediumship opened a sociological path for new interpretations. He conceived spirit possession as a form of control by an external force of an individual whereas in spirit mediumship the spirit communicates through the medium with other people. Arguing from a functionalist point of view, Beattie and Middleton (1969) as well as Lewis (1989) elaborated on these differentiations. They were concerned with the social and political functions of spirit possession and its role in social change. Lewis' emphasis on gender inequalities which was *the* topic of the time inspired a vast amount of literature which, however, was extremely critical of his

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12149-X

instrumentalist approach to spirit possession and his ignorance of its symbolic meanings as well as its theatrical dimension.

Hermeneutical Approaches

The sociological–functionalist approach was replaced by more sophisticated analyses which focused on the close link between possession, femininity, and the malleability of self. These studies concentrated on the opportunities for women and men to explore new identities, negotiate power relations between the sexes, and open up new spaces for social action through the deconstruction of male sexual hegemonies (Behrend and Luig, 1999: p. XVIIf). Boddy's (1989: p. 270) description of spirit possession as a 'metacommentary' was a benchmark since she drew attention to its counterhegemonic potential as well as its function as 'satirical allegory' (ibid.). The very ambivalence inherent in spirit possession ranging from subversion to confirmations of social order and traditional values was and is an ongoing discussion in the literature. Victor Turner (1967) had prepared the ground for such analyses by stressing the polysemic nature of symbols. The construction of symbolic worlds by mimetic ethnographies (Kramer, 1993) connected the past with the present. The spirits of powerful rulers, like the Sakalava kings in Madagascar (Sharp, 1999), displaced memory of slavery (see Wendel, 1999; Kenyon, 2011) or memory of the founder of the Kimbanguist church in a diasporic community in London (Garbin, 2011), are part of societies' cultural archives which are reinvigorated through performances of 'history from below' (Fabian, 1990) in spirit possession cults.

Such analyses are constitutive of the distancing gaze which is so characteristic of spirit possession performances. It was brought into near perfection by Jean Rouch' (1955) film about Hauka spirits which memorized and mocked the colonial past in Ghana. The possession with Hauka spirits opened an outer space for inner feelings which in everyday life had to be suppressed but were tolerated in spirit performances. Rouch's film initiated vehement discussions on ethnocentrism, on visual anthropology, but also on the relationship between spirit possession and colonialism (Stoller, 1995).

New Developments

During colonial times, many leaders of resistance movements were either spirit mediums or prophets, a tradition which was continued until the liberation wars of the 1980s in Southern and Eastern Africa (Behrend, 1999; Lan, 1985). Spirits were sometimes used for support and guidance by both sides during military confrontations like, for instance, between Renamo and Naprama fighters in Mozambique (Wilson, 1992). In African as well as Asian war-stricken societies, spirit mediums are also consulted to overcome the various consequences of war by retracing the whereabouts of missed bodies (Kwon, 2008; Perera, 2001) that were not given a proper burial out of political circumstances and are a constant threat for the living as long as the necessary rituals have not been performed. Spirit mediums, according to Perera, can also serve as an arena for negotiating feelings of aggression or revenge which are taboo in official Buddhism. They thus facilitate individual feelings of justice and alleviate processes of reconciliation which most states are unable to offer their citizens. They also give meanings to former inexplicable events and allow for the improvement of the karma of dead family members. Spirit mediums' compliance of these important social and political roles help to overcome what is called trauma or Post-traumatic stress disorder (PTSD) by medical doctors and psychologists.

The concept of trauma which is considered as a universally applicable medical category has come under severe attack (Kleinman and Desjaralais, 1996) because of its exclusion of different cultural and religious interpretations of pain. Kleinman and others shaped the sensitivity for contextual analysis for the attention of unequal power relations as well as the use of language and images when discussing 'social suffering' (Kleinman et al., 1997). In the same context, Laderman and Roseman (1996) demanded greater attentiveness toward music, dance, and rhetoric in the tradition of Stoller's work (1997) in Niger on the importance of the senses to understand the healing potential of spirit possession.

The fact that at present spirit possession takes central stage in the solution of conflicts be they related to war, to conditions of work (Ong, 1987), or other social problems led to a fundamental change in cult membership. Possession cults are no longer dominated by women, no more concentrated in rural areas, and are not even representative for the poor and downtrodden. The transcending of class boundaries is reported for several countries (Kapferer, 1991; Dawson, 2011b) and extremely common among members of Pentecostalism. But perhaps most striking is the revitalization of spirit possession in the megacities like London (Krause, 2008), New York (Brown, 1991), or Montreal (Drotbohm, 2008). Being exported by diaspora communities from the periphery to the center and vice versa, spirit possession cults are no longer confined to local places but have become nodes between several distinct worlds (Hüwelmeier and Krause, 2010). A case in point is the intensive exchange between spirit mediums in the Silicon Valley (US) and Vietnam (Fjelstad, 2010) in which commodities and therapies circulate with great intensity (Endres, 2010). Spirit mediums are thus able to create new cultural traditions, flourishing markets, and never-ending flows of ideas which are transported and embodied through traveling spirits.

Bibliography

Beattie, J., Middleton, J. (Eds.), 1969. Spirit Mediumship and Society in Africa. Routledge & Kegan Paul, London.

Behrend, H., 1999. Power to heal, power to kill. Spirit possession and war in Northern Uganda (1986–1994). In: Behrend, H., Luig, U. (Eds.), Spirit Possession, Modernity and Power. James Currey, Oxford, pp. 20–33.

Behrend, H., Luig, U. (Eds.), 1999. Spirit Possession, Power and Modernity in Africa. James Currey, Oxford.

Boddy, J., 1989. Wombs and Alien Spirits. Women, Men and the Zar-Cult in Northern Sudan. The University of Wisconsin Press, Madison, WI.

Boddy, J., 1994. Spirit possession revisited: beyond instrumentality. Annual Review of Anthropology 23, 407–434.

Bourguignon, E., 1973. Religion, Altered States of Consciousness and Social Change. Ohio State University, Columbus, OH.

Brown, K.M., 2001/1991. Mama Lola: A Vodou Priestess in Brooklyn, updated and expanded. University of California Press, Berkeley, Los Angeles.

Dawson, A., 2011a. Introduction: possession and invocation in contemporary context. In: Dawson, A. (Ed.), Summoning the Spirits. Possession and Invocation in Contemporary Religion. I.B.Tauris, London, pp. 1–20.

Dawson, A., 2011b. Spirit, self and society in the Brazilian new religion of Santo Daime. In: Dawson, A. (Ed.), Summoning the Spirits. Possession and Invocation in Contemporary Religion. I.B.Tauris, London, pp. 143–161.

de Heusch, L., 1962. Cultes de possession et religions initiatiques de salut en Afrique. Brussels. Annales du Centre d' Ėtudes des Religions 2.

Drotbohm, H., 2008. Of spirits and virgins: situating belonging in Haitian religious spaces in Montreal, Canada. Suomen Antropologi: Journal of the Finnish Anthropological Society 33 (1), 33–50.

Endres, K.W., 2010. 'Trading in spirits'? Transnational flows, entrepreneurship, and commodifications in Vietnamese spirit mediumship. In: Hüwelmeier, G., Krause, K. (Eds.), Traveling Spirits. Migrants, Markets and Mobilities. Routledge, London, pp. 118–132.

Fabian, J., 1990. Power and Performance: Ethnographic Explorations through Proverbial Wisdom and Theatre in Shaba, Zaire. University of Wisconsin Press, Madison, WI.

Firth, R., 1964. Essays on Social Organization and Values. University of London, London.

Fjelstad, K., 2010. The travels of Len Dong Spirits and their mediums. In: Hüwelmeier, G., Krause, K. (Eds.), Traveling Spirits. Migrants, Markets and Mobilities. Routledge, London, pp. 52–66.

Garbin, D., 2011. Embodied spirit(s) and charismatic power among Congolese migrants in London. In: Dawson, A. (Ed.), Summoning the Spirits: Possession and Invocation in Contemporary Religion. I.B. Tauris, New York, pp. 40–57.

Howard, A., Mageo, J.E., 2006. Introduction. In: Howard, A., Mageo, J.E. (Eds.), Spirits in Culture, History and Mind. Routledge, London, pp. 1–10.

Hüwelmeier, G., Krause, K. (Eds.), 2010. Traveling Spirits. Migrants, Markets and Mobilities. Routledge, London.

Kapferer, B., 1991. A Celebration of Demons. Berg Publishers, Oxford.

Kenyon, S., 2011. Spirits and slaves in the Sudan. In: Dawson, A. (Ed.), Summoning the Spirits. Possession and Invocation in Contemporary Religion. I.B.Tauris, London, pp. 5–73.

Kleinman, A., Desjaralais, R., 1996. Violence, culture and the politics of trauma. In: Kleinman, A. (Ed.), Writing at the Margins. Discourses between Anthropology and Medicine. University of California Press, Berkeley.

Kleinman, A., Das, V., Lock, M. (Eds.), 1997. Social Suffering. University of California Press, Berkeley, Los Angeles, London.

Kramer, F., 1993. The Red Fez. Art and Spirit Possession in Africa (Malcom R.Green, Trans.). Verso, London, New York.

Krause, K., 2008. Spiritual spaces in post-industrial places: transnational Pentecostal churches in North-East London. In: Smith, M.P., Eade, J. (Eds.), Transnational Ties: Cities, Identities, and Migrations. Transaction Publishers, New Brunswick, NJ, pp. 109–130.

Kwon, H., 2008. Ghosts of War in Vietnam. Cambridge University Press, Cambridge.

Laderman, C., Roseman, M. (Eds.), 1996. The Performance of Healing. Routledge, New York.

Lambek, M., 2010. Traveling spirits. Unconcealment and undisplacement. In: Hüwelmeier, G., Krause, K. (Eds.), Traveling Spirits. Migrants, Markets and Mobilities. Routledge, London, pp. 17–35.

Lan, D., 1985. Guns and Rain: Guerrillas and Spirit Mediums in Zimbabwe. James Currey, London.

Lewis, I.M., 1989. Ecstatic Religion. A Study of Shamanism and Spirit Possession, second ed. Routledge, London (first ed. 1971).

Meyer, B., 2006. Impossible representations: pentecostalism, vision and video technology in Ghana. In: Meyer, B., Moors, A. (Eds.), Religion, Media and the Public Sphere. Indiana University Press, Bloomington, pp. 290–312.

Ong, A., 1987. Spirits of Resistance and Capitalist Discipline: Factory Women in Malaysia. State University of New York Press, Albany, NY.

Perera, S., 2001. Spirit possession and avenging ghosts: stories of supernatural activity as narratives of terror and mechanism of coping and remembering. In: Das, V., Kleinman, A., Lock, M., Ramphele, M., Reynolds, P. (Eds.), Remaking a World. Violence, Social Suffering and Recovery. University of California Press, Berkeley, pp. 157–200.

Rouch, J., 1955. Les maîtres fous, DVD.

Sharp, L., 1999. The power of possession in Northwest Madagascar. In: Behrend, H., Luig, U. (Eds.), Spirit Possession, Power and Modernity in Africa. James Currey, Oxford, pp. 3–19.

Stoller, P., 1995. Embodying Colonial Memories. Spirit Possession, Power and the Hauka in West Africa. Routledge, London.

Stoller, P., 1997. Fusion of the Worlds: An Ethnography of Possession among the Songhay of Niger. University of Chicago Press, Chicago.

Turner, V., 1967. The Forest of Symbols. Aspects of Ndembu Ritual. Cornell University Press, Ithaca, London.

Wendel, T., 1999. Spirit possession, slavery and ritual consciousness. A case study of the Tchamba cult among the Mina in Togo. In: Behrend, H., Luig, U. (Eds.), Spirit Possession, Power and Modernity in Africa. James Currey, Oxford, pp. 111–123.

Wilson, K.B., 1992. Cults of Violence and Counter-Violence. Journal of Southern African Studies 18 (3), 527–582.

Spirituality

Margaret Holloway, School of Social Sciences, University of Hull, Hull, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Spirituality and spiritual care is a growing field in the caring professions. This development has been led by nursing; social work has remained cautious and unsure of its role although the current emphasis on well-being is contributing to a change in attitude. Key to this has been the development of the concept of humanistic spirituality, which is distinguished from religion and argued to belong to everyone. Despite this, and the multiple spiritual care models available, spirituality and spiritual care have remained contested and problematized. An exciting development, progressed by social work, has emerged from challenges to the Western paradigm from indigenous spiritualities and ecospirituality perspectives.

Origins and Development

Spirituality studies are a burgeoning field in the literature of the caring professions and applied disciplines, which has emerged since the mid-1980s. Although a range of disciplines has contributed to its development as a dimension of professional practice – including social work, education, palliative medicine, business studies, and chaplaincy – it is nursing which is credited with raising the profile of spiritual care within secular health care systems (McSherry and Ross, 2012). This movement began in the Western developed world, where the primary thrust has been the distinguishing of spirituality from religion and, in consequence, the opening up of spiritual care as not only relevant for people of no religious persuasion, but also belonging to the remit of secular professionals and organizations. A parallel development has occurred in Asia, Africa, the Pacific Rim, and practitioners working with the indigenous peoples of Australia, New Zealand, and North America, where the language of spirituality has been utilized to reaffirm traditional, religious, and cultural values and practices that had been neglected in the widespread adoption of Western models of education and practice. This latter development has been led by social work, where it belongs to the drive for culturally competent practice (Coates et al., 2006), and business, where it is applied to a critical reappraisal of workplace ethics, values, and culture (Smith and Rayment, 2010).

The interest in spirituality and spiritual care among the caring professions should be understood in the context of broader societal developments. Sociologists of religion and other scholars have pointed to a number of apparently opposing trends: the decline in traditional religious adherence among white indigenous populations (Voas and Crockett, 2005); swelling church congregations (particularly RC and Pentecostal) in areas with high immigrant and migrant populations (Brierley, 2000); stable or growing Muslim communities (Office for National Statistics, 2011); and a growth in the popularity of 'New Age' religions and spiritual practices (Heelas and Woodhead, 2005). Davie has also suggested two trends that link practicing Christians with nonpracticing members of the local community – "believing without belonging" and "vicarious religion" (Davie, 2007) – both operating as the means by which people, who for the most part loosely define themselves as agnostic, maintain a tenuous connection with organized religion through use of its social functions, such as weddings, christenings, and funerals, and a largely nostalgic attachment to core beliefs and festivals, such as Christmas.

Some of the earliest writings about spirituality in the caring professions make explicit links with this pluralism of belief in the multicultural contexts in which they operate, as a significant imperative for a critical reappraisal of the neglect of a spiritual dimension in their practice (Lloyd [Holloway], 1997). Others start from a critique of the technicist, procedurally driven services that have been a feature of most health and social care systems in the developed world since the 1990s, and accompanying calls for a return to relationship-based practice, to point to the basis for that practice being in the profession's Judeo-Christian origins. For example, nursing texts, particularly the earliest spirituality literature, point to the beginnings of the profession in Christian caring institutions (McSherry and Ross, 2012); social work writers point to the heavy reliance on both Judeo-Christian thought (Bowpitt, 1998) and humanistic psychology in the development of social work's underpinning values and principles (Payne, 2011). However, such accounts are careful to distinguish contemporary spirituality practice from religious practice; the notion of 'spiritual but not religious,' introduced in the early literature, remains one of the most frequently used by the caring professions. Market surveys and opinion polls also confirm that spirituality is seen by many as, "a phenomenon independent of organised religion" (Nolan and Holloway, 2013: p. 161).

Concept and Definitions

Defining this broader, more diffuse 'spirituality' has been, and remains, a preoccupation. Definitions are premised on the assertion that a spiritual dimension is relevant for everyone, irrespective of whether or not they identify as followers of a religion. Despite the intense focus on defining what is meant by the term, there remains no universally accepted definition. Rather, within the multitude of definitions can be seen frequently repeated themes and aspects, put together in different contributions with varying emphases. For critics, the inability to agree a common definition highlights the fundamental weakness of the concept, but others point to the breadth and nuance contained in the attempts to define

spirituality as its strength. Early suggestions that the lack of a common definition inhibits the development of spiritual care have been overtaken by interest in its "practical relevance and application" (McSherry and Ross, 2012: p. 212). A core theme found in possibly every definition is the centrality of *meaning and purpose*. This is most usually coupled with a process such as the 'pursuit of meaning' or 'meaning making.' A second theme concerns *relatedness and connectedness* (Swinton, 2012). Less common, but an important strand, nonetheless, suggests a link with moral behavior and ethics (Holloway, 2012). These three strands are themselves interrelated. For example, it may be through relationships with others, and for some, a divine 'Other,' that individuals find their greatest sources of meaning; it may be through their behavior toward others that people are able to express the connectedness of their shared humanity; for some professionals, spirituality is the most fundamental connection they make with the people they care for and their work is the source of deepest meaning in their life (Sinclair et al., 2006). Canda and Furman (2010), whose work has been in the vanguard of articulating a spiritual dimension for social work practice, which they further describe as being at the 'heart' of social work, combine these ideas in a complex definition that describes spirituality as:

> … the wholeness of what it is to be human … The spiritual relates to the person's search for a sense of meaning and morally fulfilling relationships between oneself, other people, the encompassing universe, and the ontological ground of existence, whether a person understands this in terms which are theistic, atheistic, nontheistic, or any combination of these.
>
> Canda and Furman, 1999 (first edition): pp. 43–44. See also second edition (2010) p. 75

One argument made by social work writers is that there is considerable overlap between these core elements of spirituality and the core values and principles of social work:

> The cluster of values and world-views to which the concept of spirituality points us have been around far longer than social work, yet in very significant ways they have always belonged to the heart of social work.
>
> Holloway and Moss, 2010: p. 2

Alongside these core elements, however, is identification of the phenomenon of *transcendence*, which some writers, including from social work, include in their definition. For example, Lindsay adds to the search for meaning and purpose, "a moral dimension which reflects a concern with relationships to others, the universe, *and to some transcendent being or force* (my italics)" (Lindsay, 2002: pp. 31–32). Kellehear, a sociologist who writes about palliative care, sees the inclusion (implicit or explicit), or not, of transcendence in conceptualizations of spirituality as the key distinguishing factor (Kellehear, 2000). This is not a familiar term for social work and it is not used in the earlier spirituality literature emanating from the caring professions, although it is embedded in religious and New Age spirituality discourses. More recently, however, caring professionals have discussed transcendence in the sense of 'going beyond' the limitations of immediate circumstance and a wider context for the self. This is variously expressed as "transcending the ordinariness of everyday life" (Kellehear, 2000); being able to "transcend the situation such that it no longer has a problematic, painful or oppressive impact" (Holloway and Moss, 2010: p. 101); "grounding a person in their own centre and taking them beyond themselves" (Hegarty, 2007: p. 43).

Some writers choose to describe what might be termed the *attributes* of spirituality, rather than attempt a definition. Frequently highlighted are love, joy, compassion, caring, awe, wonder and mystery, beauty and creativity. van Hooft (2012) also notes that these are experienced 'in the presence of,' and contribute to, experiences of transcendence. Culliford helpfully explores how such 'feelings' are more than emotions but central to what he terms the 'psychology of spirituality':

> Consider the phrases, 'to be filled with awe', 'To be ecstatic with joy', 'To feel choked right up', 'To be paralysed with fear', and 'To be racked with tears'. What do they signify? They mean above all that the person is alive and responding at a profound and unpremeditated level to events beyond their control, beyond (at that instant) even the grasp of their intellect … these phrases describe strong emotions that often arise at life-changing moments.
>
> Culliford, 2011: p. 52

The search for a definition of spirituality goes on. In 2009, the US National Consensus Project on Palliative Care proposed the following as a 'consensus definition':

> Spirituality is the aspect of humanity that refers to the way individuals seek and express meaning and purpose and the way they experience their connectedness to the moment, to self, to others, to nature, and to the significant or sacred.
>
> Puchalski et al., 2009: p. 887

Recognizing that all definitions are to some extent culturally situated, Nolan et al. (2011) nevertheless propose a remarkably similar definition for the European context of palliative care. This suggests that the dominant articulations of contemporary spirituality in fact emanate from the Western paradigm of health and social care practice. Highlighting the themes that predominate in this literature, Holloway et al. (2011) summarize:

> Spirituality is more basic than religiousness. It is a subjective experience that exists both within and outside traditional religious systems. Spirituality relates to the way in which people understand and live their lives in view of their core beliefs and values and their perception of ultimate meaning. Spirituality includes the need to find satisfactory answers to ultimate questions about the meaning of life, illness and death It can be seen as comprising elements of meaning, purpose, and connection to a higher power or something greater than self.
>
> p. 18.

It is apparent that these understandings of contemporary spirituality are essentially *humanistic* (Elkins et al., 1988) and the term 'humanistic spirituality' is increasingly used. While the origins of humanistic spirituality can be seen in humanistic and transpersonal psychologies and allied to atheistic beliefs, it has

evolved in more recent practice as an umbrella and bridging concept within which traditional religious, secular, and New Age religious spiritualities can find common ground in the notion of common humanity (Nolan and Holloway, 2013). This makes it attractive to professionals working in multicultural contexts characterized by belief pluralism, seeking common meeting points to engage meaningfully with the individuals, families, and communities that use their services (Holloway, 2006). A humanistic spirituality that defines 'the spiritual' as concerned with the essence of being human (Swinton, 2007) and asserts this over functionalist and materialistic positions, sits well with professional discourses which privilege the inherent value and dignity of the person (van Otterloo et al., 2012).

Spiritual Care

It follows that for those who accept that spirituality applies to everyone, *spiritual care* is potentially relevant for every service user and belongs to the remit of every worker. However, health care workers in general and social workers in particular have shown reluctance and resistance to engage with spirituality (Holloway, 2007). Even for those practitioners who recognize spirituality as a potentially important dimension for service users and acknowledge its neglect in their practice, confusion and uncertainty is expressed about how to offer spiritual care and ambivalence as to its appropriateness within their role (Moss, 2012). Empirical studies find both health care professionals and social workers reporting that little or no attention was given to spirituality in their initial training (Furman et al., 2004), combined with constant calls to redress this deficit at all career levels (Yardley et al., 2009).

The importance of spiritual care in palliative and end-of-life care and bereavement is, however, recognized across the practice disciplines, including social work (Gwyther et al., 2005) and has generated the most research. Spiritual care research and models in this field have also used spirituality measures generated from other contexts. A systematic review of the English language literature from 2000 to 2010 on spiritual care at the end of life found, "a vast amount of documentation relating to spirituality, spiritual care and to end of life care" (Holloway et al., 2011: p. 8). After strict exclusion criteria were applied, 248 sources from 17 different countries (of which 24 were from social work and a further 4 interdisciplinary and social science making reference to social work) were examined in detail. This literature gives some clue to the reasons underlying the difficulties reported by practitioners. Importantly, it was noted that in the gray literature, there is a plethora of systems-level guidance which recommends that spiritual needs should be assessed and which describes the system for implementing this, but there is considerably less specificity about practice at the front line. Likewise, there is a growing tendency to include spiritual care in professional standards (see Society for Spirituality and Social Work, Arizona State University) but little educational research including evaluation of teaching and training programs (Holloway et al., 2011).

The most researched aspect of spiritual care is the assessment of spirituality, religiosity, and spiritual need or distress. The UK Department of Health review (Holloway et al., 2011) found continuing evidence that most health care practitioners do not undertake spiritual assessments, and of those that do, there is a wide variation in what that constitutes. A plethora of reasons is giving for failing to assess spiritual need: patient (*sic*) and/or family reluctance; fearing that they might cause distress; not knowing how to approach the subject; feeling embarrassed or awkward; regarding the subject as too personal; trusting that the chaplain or social worker will do it; having no belief themselves in God; the patient (*sic*) being unable to communicate; the patient (*sic*) being short-term; the worker not having enough time. However, while there is considerable resistance to formally undertaking a spiritual assessment, practitioners identified a number of ways in which they felt they were engaging with spiritual care. These included: 'being there'; supportive care – 'holding a safe nurturing space'; empathy and compassion – entering into 'weak places'; sharing the journey; sustaining. Thus, there is an implication that palliative care practitioners are happier with spiritual care perceived as *being* rather than *doing*. The social work literature, although comparatively sparse, generally works from the premise that a more active intervention is implied as a consequence of a heightened spiritual awareness and recognition of spiritual need. This will be discussed in the ensuing section.

Practice Models

Despite calls from practitioners for more training and practice guidance, there are vast numbers of models for spiritual care available – indeed, the array of models is often cited as a source of confusion. Rather less is known, however, about use of these models by practitioners or evaluation of their effectiveness. The research has generated a large number of assessment models but considerably fewer which might be termed spiritual care interventions. Social work is more inclined than other professions (such as nursing) to distinguish between assessment and intervention, which tend to conflate the two under the banner of 'spiritual care.' It is possible that this multiplicity of spiritual assessment models parallels a growing emphasis in all professions on assessment and is the natural first step in taking forward the growing interest in spirituality (Pesut and Sawatzky, 2006).

Assessment models can be categorized under four approaches: generic, quantitative, narrative (qualitative), and domain-based (Draper, 2012). Generic models ask broad open-ended questions about sources of meaning, support, and satisfaction within the person's life, which may lead to a focus on spirituality. Quantitative instruments use standardized and validated scales to 'measure' the degree of spirituality and religiosity of the individual, often within a broader dimension such as well-being. Narrative models provide a framework for exploring the person's 'story,' employing biographical and qualitative tools with an emphasis on the spiritual journey. Domain-based approaches explore spirituality as one domain of the person's life alongside the physical, emotional, and social domains, or alternatively focus on the spiritual dimension within every other domain. The greatest number of spiritual assessment models come from the US in the form of validated scales but these may not translate well to other societies and appear to have limited use as practice models

compared to their employment as research tools (Holloway et al., 2011).

Despite a significant concentration on spiritual need in the literature, there has been surprisingly little evaluation of the different approaches to spiritual assessment. What there is, debates the nature and ethos of the particular professional practice and how spiritual assessment fits into its overall approach to assessing problems and needs. Medicine and nursing recognize the importance of evidence-based practice and construe this as using validated quantitative tools, but construct arguments for incorporating other forms of spiritual assessment around epistemology and practice process (Draper, 2012). Thus, the type of knowledge that the clinician values in practice may be less likely to arise from facts that can be demonstrated empirically, than from processes which they have learned to trust:

> Modern nurses who understand their practice in terms of the 'nursing process' are therefore likely to assume that … the point of assessment is to identify patients' needs and to represent them in numbers … or words.
>
> Draper, 2012: p. 7

Social work writers have developed their ideas on spiritual assessment through focusing on the core values, skills, and approaches of the profession and the ways in which attention to spiritual issues and needs should be an embedded part. For example, there is a marked leaning toward holistic practice (Gilbert, 2011); narrative review (Hodge, 2005), empowerment, and partnership (Holloway, 2007); pursuing social justice (Holloway and Moss, 2010); change and transformation (Canda and Furman, 2010); relationship-based and humanistic practice (Payne, 2011); and critical practice (Gardner, 2011). Perhaps as a consequence of social work's long-standing reluctance to embrace quantitative measures, there is little engagement with the use of standardized scales and 'measurement' approaches. Moreover in keeping with its history as a 'problem-based' professional activity, the social work and spirituality discourse is rather more developed in relation to assessing and addressing spiritual need and distress than identifying service users' spiritual resources. The marked exception to this has been driven by mental health service users (including professionals who also identify as service users) who have insisted that their spirituality should be taken account of as a positive resource in their lives and not (as they frequently experience from psychiatrists) as a dimension to be ignored or vilified as the source of the problem (Coyte et al., 2007).

Spiritual care interventions fall broadly into two categories: integrated models, which embed spiritual care within a holistic framework of care; and spiritual care models, which take spiritual need as the starting point and may treat spiritual care as a discrete aspect (Holloway et al., 2011). A well-known example of integrated models is Sulmasy's biopsychosocial–spiritual model, based on the concept of 'whole person synergy' (Sulmasy, 2002). Sulmasy outlines a multidisciplinary care framework, based on the concept of dynamic interaction between the different aspects of a person's life. It fits with both domain approaches to spiritual assessment and the application of systems theory in social work interventions.

> according to this model, this spiritual state may … be modulated by the biopsychosocial state of the person, and the spiritual state may also modulate the biopsychosocial state
>
> p. 7

An example of a discrete spiritual care model, which is popular in the UK, is the Marie Curie Cancer Care spiritual and religious competencies that identifies four levels of competency and the tasks and functions required at this level as well as the palliative care professionals (and hospice volunteers) who should offer spiritual care at each level.

Contributions from Social Work

Generic social work theory has paid scant regard to spirituality, Payne's humanistic approach (2011), and Dominelli's global perspectives on antioppressive practice (Dominelli, 2010) providing notable exceptions. Likewise, social work's contribution to the spirituality literature has been relatively light compared to other professions of which nursing has produced the bulk of the material. Only a handful of practice models have been produced and they reflect the core concerns of social work in general. Canda and Furman (2010) build their framework for 'spiritually sensitive social work' on a foundation of six ethical principles incorporating commitment to the values of service, social justice, the dignity and worth of the person, the importance of human relationships, integrity, and competence. Gardner (2011) reflects these same values in her nine principles for 'critical spirituality practice,' adding to them a critically reflective dimension. Hodge favors a biographical approach to exploring with the person their spiritual and religious narrative; his "integrated framework" (Hodge, 2001) combines 'guided self-narrative' with an interpretive framework exploring six psycho-social-moral domains: affect, behavior, cognition, communion, conscience, and intuition. Holloway emphasizes partnership, building her Fellow Traveller Model of joining (spiritual awareness), listening (spiritual sensitivity), understanding (spiritual empathy), and interpreting (spiritual exploration) around the notion of companions on a spiritual journey (Holloway and Moss, 2010) and "maintaining the spirit" (Lloyd [Holloway], 1997). Moss's 'Why/Wow' model relates spiritual care to the tension in social work of engaging with both the light and the dark sides of human experience (Holloway and Moss, 2010). Chan et al. (2006) draw on eastern religions and philosophy to propose her 'body–mind–spirit' model for all social work practice. All acknowledge the dimension of cultural competence and spiritual practice competencies and standards are emerging, some as stand-alone and others incorporated into generic statements (Gwyther et al., 2005; Hodge, 2007). Furness and Gilligan (2010) emphasize that social workers should also be competent in addressing the significance of religious belief in the lives of many service users and detail a religious competencies framework. Mathews (2009) and Crisp (2010), while not proposing a model as such, take a life course approach to spiritual development, drawing on familiar social work concepts and tools. Similarly, spirituality as a component of quality of life and well-being (Harvey and Silverman, 2007),

the concept of resilience and a strengths perspective (Ungar, 2008), the language and approaches of restorative justice (Cheers et al., 2006), and the social worker as change agent (Canda and Furman, 2010) have been adopted by writers on spirituality and social work.

There is a thread which runs through these models proposed by social work and which is distinctive to them. This is the understanding that spiritual awareness and recognition of spiritual need are not of themselves enough; it falls within the profession's remit to respond directly to need and therefore all models detail the process of assessment and subsequent intervention, frequently through an interdisciplinary approach in which the social worker refers to and/or works alongside a religious or spiritual care professional. This is important in offering tools to address the concerns and inhibitions reported by practitioners (see earlier section). However, social work has made a more radical and significant contribution in its challenging of the Western paradigm, both in terms of the conceptualizing of spirituality and the model of spiritual practice. This contribution stems from the diverse voices of the global social work community. Social workers in societies where religion is a way of life and secularization is not the assumed default position have challenged the language of 'spiritual but not religious' as unhelpful and meaningless (Wong and Vinsky, 2009). Social workers belonging to and/or working with the indigenous peoples of North America, Australia, and New Zealand have drawn on insights from a spiritual tradition that understands the relationship between human beings and their environments as sacred and which therefore implies a wholly different set of priorities and starting points for social work interventions (Coates, 2007). Social workers from Asia have challenged the individualism inherent in the contemporary personalized and internalized spirituality discourse, endorsing an alternative world view based on communitarian values and responsibility to others (Yip, 2005). Thus ecospirituality models (Nolan and Holloway, 2013) that enrich understanding of the connections between place, identity, and well-being have much to offer to the emerging interest in care environments and aging and dying 'in place.' Equally, the privileging of connectedness and relationship offers an important corrective to therapeutic models which emphasize individual self-actualization to the neglect of social context and nurturing of social capital. These perspectives are leading a potentially significant evolution of the original conceptualizations of spirituality and understandings of spiritual care in the caring professions, although their adoption outside of social work is as yet limited.

This insistence that valuing diversity and challenging social inequality and oppression is core to "spiritual practice" (Coates, 2007) is a theme possibly unique to the social work and spirituality literature, which at the same time recognizes that religious and spiritual traditions can be equally oppressive (Furness and Gilligan, 2010). Indeed, anecdotal evidence would suggest that both witnessing and experiencing religious cultures as oppressive, particularly toward certain groups such as women and nonheterosexuals, lies behind much of the (historical) antipathy in social work to inclusion of spirituality in their practice. This is approached on an individual therapeutic level through the exploration of harmful spiritual narratives and the cocreation of a new empowering narrative (Hodge, 2001, 2005). At the social structural level, writers draw on liberation spiritualities and theology to argue that spiritual practice unites individual, social, and political concerns (Holloway, 2007). Understanding spiritual concerns as well as spiritual and material resources as integral to community development is, according to Chile and Simpson, "the fire of social justice which brings about radical transformation of structures of society" (Chile and Simpson, 2004: p. 322). Gardner (2011) offers critical social theory as the means by which both sides of the spirituality empowerment-oppression see-saw may be held in balance, reminding that, "there are certain values we would want to affirm across spiritual and religious boundaries as well as values we wouldn't" (p. 48). The key to critical spiritual practice is thus making explicit the dynamics of power in whatever direction they are operating. This chimes with arguments from mental health service users that for too long their spirituality has been seen by psychiatrists as part of their problem, rather than as they experience it, a resource which helps them to make sense of their illness (Coyte et al., 2007). A growing band of psychiatrists, many coming from the transpersonal psychology tradition, support this position, utilizing spiritual perspectives in a person-centered approach to their practice (Culliford, 2011).

See also: Counseling Psychology; Culturally Appropriate Interventions in Social Work; Culture and the Self: Implications for Psychological Theory; Ecological Social Work: Shifting Paradigms in Environmental Practice; Health Social Work; Indigenous Social Work; Liberation Theology and Social Work: Anti-Oppressive Common Grounds; Mental Health and Social Work; Migration, Ethnicity, Aging, and Social Work Practice; Māori Social Work; Relational Social Work; Social Work Ethics; Social Work with Racialized Groups: Frameworks for Practice; Subjective Wellbeing, Psychology of.

Bibliography

Bowpitt, G., 1998. Evangelical Christianity, secular humanism, and the genesis of British social work. British Journal of Social Work 28, 675–693.

Brierley, P., 2000. Religion. In: Halsey, A., Webb, J. (Eds.), Twentieth Century British Social Trends. Macmillan, London.

Canda, E., Furman, L., 1999. Spiritual Diversity in Social Work Practice: The Heart of Helping. The Free Press, New York.

Canda, E., Furman, L., 2010. Spiritual Diversity in Social Work Practice: The Heart of Helping, second ed. The Free Press, New York.

Coates, J., Grey, M., Hetherington, T., 2006. An eco-spiritual perspective: finally, a place for indigenous approaches. British Journal of Social Work 36 (3), 381–399.

Coates, J., 2007. From ecology to spirituality and social justice. In: Coates, J., Graham, J.R., Swartzwebtruber, B., Ouellete, B. (Eds.), Spirituality and Social Work: Selected Canadian Readings. Canadian Scholars Press, Toronto, pp. 213–228.

Chan, C., Ng, S., Ho, R., Chow, A., 2006. East meets West: applying eastern spirituality in clinical practice. Journal of Clinical Nursing 15 (7), 822–832.

Cheers, B., Binell, M., Coleman, H., Gentle, I., Miller, G., Taylor, J., Weetra, C., 2006. Family violence: an Australian indigenous community tells its story. International Social Work 49 (1), 51–63.

Chile, L., Simpson, G., 2004. Spirituality and community development: exploring the link between the individual and the collective. Community Development Journal 39 (4), 318–331.

Coyte, M., Gilbert, P., Nicholls, V., 2007. Spirituality, Values and Mental Health: Jewels for the Journey. Jessica Kingsley, London.

Crisp, B., 2010. Spirituality and Social Work. Ashgate Publishing, Farnham.

Culliford, L., 2011. The Psychology of Spirituality: An Introduction. Jessica Kingsley Publishers, London.
Davie, G., 2007. The Sociology of Religion. Sage, London.
Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.
Draper, P., December 2012. An integrative review of spiritual assessment: implications for nursing management. Journal of Nursing Management 20 (8), 970–980. http://dx.doi.org/10.1111/jcnm.12005.
Elkins, D., Hedstrom, L., Hughes, L., Leaf, J., Saunders, C., 1988. Toward a humanistic-phenomenological spirituality: definition, description and measurement. Journal of Humanistic Psychology 28 (4), 5–18.
Furman, L.D., Benson, P.W., Grimwood, C., Canda, E., 2004. Religion and spirituality in social work education and direct practice at the millennium: a survey of UK social workers. British Journal of Social Work 34 (6), 767–792.
Furness, S., Gilligan, P., 2010. Religion, Belief and Social Work: Making a Difference. Policy Press, Bristol.
Gardner, F., 2011. Critical Spirituality. A Holistic Approach to Contemporary Practice. Ashgate Publishing, Farnham.
Gilbert, P. (Ed.), 2011. Spirituality and Mental Health. Pavilion Publishing, Brighton.
Gwyther, L., Altilio, T., Blacker, S., Christ, G., Csikai, E., Hooyman, N., Kramer, B., Linton, J., Raymer, M., Howe, J., 2005. Social work competencies in palliative and end-of-life care. Journal of Social Work in End-of-life and Palliative Care 1 (1), 87–120.
Harvey, I.S., Silverman, M., 2007. The role of spirituality in the self-management of chronic illness among older African and Whites. Journal of Cross Cultural Gerontology 22, 205–220.
Heelas, P., Woodhead, L., 2005. The Spiritual Revolution: Why Religion Is Giving Way to Spirituality. Blackwell, Oxford.
Hegarty, M., 2007. Care of the spirit that transcends religious, ideological and philosophical boundaries. Indian Journal of Palliative Care 13 (2), 42–47.
Hodge, D., 2001. Spiritual assessment: a review of major qualitative methods and a new framework for assessing spirituality. Social Work 49, 27–38.
Hodge, D., 2005. Spiritual lifemaps: a client-centered pictorial instrument for spiritual assessment, planning, and intervention. Social Work 50 (1), 77–87.
Hodge, D., 2007. The spiritual competencies scale. A new instrument for assessing spiritual competencies at the programmatic level. Research on Social Work Practice 17, 287–295.
Holloway, M., 2006. Death the great leveller? Towards a transcultural spirituality of dying and bereavement. Journal of Clinical Nursing Special Issue Spirituality 15 (7), 833–839.
Holloway, M., 2007. Spiritual need and the core business of social work. British Journal of Social Work 37 (2), 265–280.
Holloway, M., 2012. Social Work. In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), Oxford Textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 235–241.
Holloway, M., Moss, B., 2010. Spirituality and Social Work. Palgrave Macmillan, Basingstoke.
Holloway, M., Adamson, S., McSherry, W., Swinton, J., 2011. Spiritual Care at the End of Life. A Systematic Review of the Literature. Department of Health, London. Available online: www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_123812.
van Hooft, S., 2012. The western humanist tradition. In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), The Oxford Textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 43–48.
Kellehear, A., 2000. Spirituality and palliative care: a model of needs. Palliative Medicine 14, 149–155.
Lindsay, R., 2002. Recognizing Spirituality: The Interface between Faith and Social Work. University of Western Australia Press, Crawlet.
Lloyd, M., [Holloway, M.], 1997. Dying and bereavement, spirituality and social work in a market economy of welfare. British Journal of Social Work 27 (2), 175–190.
Mathews, I., 2009. Social Work and Spirituality. Learning Matters, Exeter.
McSherry, W., Ross, L., 2012. Nursing. In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), Oxford textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 211–218.
Moss, B., 2012. The pedagogical challenge of spirituality: a 'co-creative' response. Journal of Social Work 12 (6), 595–613.
Nolan, S., Saltmarsh, P., Leget, C., 2011. Spiritual care in palliative care: working towards an EAPC task force. European Journal of Palliative Care 18 (2), 86–89.
Nolan, S., Holloway, M., 2013. An A–Z of Spirituality for the Caring Professions. Palgrave Macmillan, Basingstoke.
Office for National Statistics, 2011. Religion in England and Wales. Part of 2011 Census. Crown, London.
van Otterloo, A., Aupers, S., Houtman, D., 2012. Trajectories to the New Age: the spiritual turn of the first generation of Dutch New Age teachers. Social Compass 59 (2), 239–256.
Payne, M., 2011. Humanistic Social Work: Core Principles in Practice. Palgrave Macmillan, Basingstoke.
Pesut, B., Sawatzky, R., 2006. To describe or prescribe: assumptions underlying a prescriptive nursing process approach to spiritual care. Nursing Inquiry 13, 127–134.
Puchalski, C., Ferell, B., Virani, R., Otis-Green, S., Baird, P., Bull, J., Chochinov, H., Handzo, G., Nelson-Becker, H., Prince-Paul, M., Pugliese, K., 2009. Improving the quality of spiritual care as a dimension of palliative care: the report of the consensus conference. Journal of Palliative Medicine 12 (10), 885–904.
Sinclair, S., Raffin, S., Pereira, J., Guebert, N., 2006. Collective soul: the spirituality of an interdisciplinary palliative care team. Palliative and Supportive Care 4, 13–24.
Smith, J., Rayment, J., 2010. Globally fit leadership: four steps forward. Journal of Global Responsibility 1 (1), 55–65.
Sulmasy, D.P., 2002. A biopsychosocial-spiritual model for the care of patients at the end of life. The Gerontologist 42 (Oct), 24–37.
Swinton, J., 2007. Forgetting whose we are: theological reflections on personhood, faith and dementia. Journal of Religion, Disability and Health 11 (1), 37–62.
Swinton, J., 2012. Healthcare spirituality: a question of knowledge. In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), The Oxford Textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 99–104.
Taylor, C., Walker, S., 2012. Compassion: luxury or necessity? In: Cobb, M., Puchalski, C., Rumbold, B. (Eds.), The Oxford Textbook of Spirituality in Healthcare. Oxford University Press, Oxford, pp. 235–241.
Ungar, M., 2008. Resilience across cultures. British Journal of Social Work 38 (2), 218–235.
Voas, D., Crockett, A., 2005. Religion in Britain: neither believing nor belonging. Sociology 39 (1), 11–28.
Wong, Y.-L.R., Vinsky, J., 2009. Speaking from the margins: a critical reflection on the 'spiritual but not religious' discourse in social work. British Journal of Social Work 39, 1343–1359.
Yardley, S.J., Walshe, C.E., Parr, A., 2009. Improving training in spiritual care: a qualitative study exploring patient perceptions of professional education requirements. Palliative Medicine 23, 601–607.
Yip, K., 2005. A dynamic Asian response to globalization in cross-cultural social work. International Social Work 48 (5), 593–607.

Relevant Websites

http://ssw.asu.edu/portal/research/spirituality – Arizona State University School of Social Work.
http://www.basspirituality.org.uk – British Association for the Study of Spirituality.
http://www.rish.ch/ – European Network of Research on "Religion, Spirituality and Health".
http://mariecurie.org.uk – Marie Curie Cancer Care.
www.rcpsych.ac.uk/expertadvice/treatments/spirituality – Royal College of Psychiatry Spirituality Special Interest Group.

Split Brains and Human Consciousness

Colwyn Trevarthen, The University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Research on the consequences of complete section of the forebrain commissures in cats and monkeys, and in human patients operated to control epilepsy, has clarified how the brain mediates perception, learning, memory, and the control of movements, and has raised many questions. No single process of consciousness and no simple localization of 'modules' of perception or reasoning have been found. However, consistent differences between the awareness of the left and right hemispheres in human beings, and regarding their capacity to integrate experience of objects with skilled action, to employ language in speech or writing, to imagine contexts for discrimination and decision making, to attribute qualia to objects, and to sympathize with the motives and feelings of other persons, cast light on how the human brain is uniquely adapted for cooperative cultural life. The findings of split brain research have importance for the whole range of scholarship that attempts to explain the human condition.

A History of Split-Brain Research

A 'split brain,' usually the brain of a mammal, is one in which all direct, 'one-neuron' connections between the two forebrain cerebral cortices have been cut. The largest interhemispheric bridge, or commissure, is the corpus callosum (Bogen, 1985; Innocenti, 1986) (see **Figure 1**). The first split-brain operation was performed on a cat in 1952, combining commissurotomy (sectioning of the corpus callosum and anterior commissure) with division of visual crossover fibers, the optic chiasm, at the base of the brain. Split-brain cats see and learn with each eye connected to a different hemisphere. Transfer of visual learning between the hemispheres is abolished (Myers and Sperry, 1953). In monkeys, commissurotomy separates two eye–hand coordination and learning systems, each hemisphere having predominant cortico-motorneuronal control for limb movements on the other side of the body and within the visual half field on that side (Sperry, 1961). An intention to move one hand determines which cortical visual system will perceive and learn (Trevarthen, 1962, 1965, 1974a, 1987). Perception around the body in the periphery of the visual field, including 'ambient' preattentive awareness of space and motion properties of objects, is not divided in the split brain, indicating that subhemispheric (brain stem) systems, which remain unified, can integrate core perceptuo-motor functions directed to the 'behavior space' surrounding the body (Trevarthen and Sperry, 1973; Trevarthen, 1987).

In the 1960s, Los Angeles neurologists and neurosurgeons Bogen et al. (1965), encouraged by the integrated functional state of split-brain animals, performed commissurotomy operations on human patients to control life-threatening epilepsy, with excellent results. Psychological tests of these patients, under the direction of Roger Sperry at the California Institute of Technology, demonstrated a divided consciousness and two different cognitive systems for focal conscious awareness and skilled movements (Sperry et al., 1969). Only one, in the left hemisphere, had normal control of speech, and it was superior in logical or 'propositional' thinking. Right-hemisphere consciousness understood language at a simple level, and it was superior in awareness of relationships and configurations of experience called 'appositional' (Bogen, 1969). Roger Sperry (1982), who initiated split-brain research in animals and supervised the tests of awareness in humans after commissurotomy, received

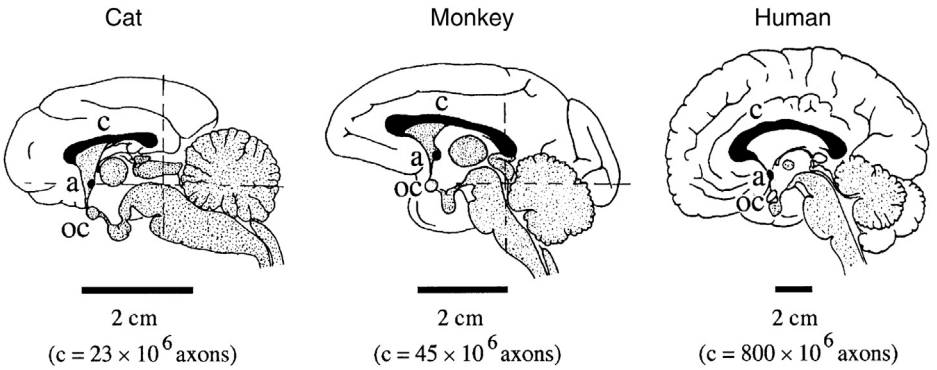

Figure 1 Anatomy of the cerebral commissures. Sagittal sections of the brains of cat, monkey (rhesus macaque), and human adult, showing the corpus callosum (c), anterior commissure (a), and optic chiasm (oc). The estimated numbers of nerve axons in the corpus callosum of each species are indicated. This illustrates the enormous expansion of cortical functions in human beings, and their bilateral integration through the corpus callosum.

a Nobel Prize for this work in 1981 (Sperry, 1982). Split-brain research has been enormously influential, clarifying intrinsic functions of consciousness in the brain, and the cerebral symmetry of human motives or cognitive strategies that guide different purposes in action. However, the experimental tests performed with individual subjects, concentrating on the visual sense, touch perception of objects by the hands, cognitive abilities for identifying objects, and especially language and rational skills, have led to restricted views. They do not relate the functions of consciousness to their meaning for active, imaginative, and emotion-charged life in the world at large, in human relationships, and in the social and cultural world. Now, with the aid of functional brain imaging of normal subjects engaged in a wider range of tasks, and with information from infancy research on the innate capacities of the uneducated human brain for sharing experience, a new interpretation of the effects of commissurotomy in humans, acknowledging the asymmetric substrates of motivation and attention in subhemispheric systems, is required. It will need to incorporate scientific understanding of how advanced mammalian brains organize motives and emotions to engage with and evaluate the physical environment with intelligent prospective consciousness, how they learn concepts and representations, and how they may collaborate by communicating their intentions, experiences, and affections by expressive body movements.

Perception, Learning, and Prospective Control of Movements

Before the 1940s, the functions of the corpus callosum, estimated to have about 1000 million axons, were unknown (Akelaitis, 1944). Early in the 1950s, Sperry at the University of Chicago, who was researching the anatomy of visual form perception in the cerebral cortex, proposed a surgical investigation that was to yield new insights into the brain systems of consciousness (Evarts, 1990: pp. xix–xxi). Ronald Myers, a research student with Sperry, demonstrated that a cat's visual experience was divided in two by cutting the corpus callosum, and proved that such experience could be transferred by the callosum between the hemispheres (Myers and Sperry, 1953).

In the 1950s and 1960s, Sperry's research group, now at the California Institute of Technology (Caltech), exploited what was now called the 'split brain' to study visual and tactile perceptions, memories, and motor control in cats or monkeys (Sperry, 1961). Studies with monkeys showed that, while ordinary pattern discrimination was completely split, less 'figurative' or focal visual experiences were not completely duplicated after split-brain surgery. They also proved the importance of 'motor sets' or 'intentions to move' that 'anticipate' sensory information for the whole body, and that invoke undivided systems of the brain stem and cerebellum. They confirmed an approach to conscious perception of objects not only as the processing of an input of information from the distance senses for identifying and manipulating objects, and for remembering them, but also as goals for coordinated movement, which must both anticipate consequences within the body, proprioceptively, and regulate the serial ordering of movements in hopeful projects of engagement of a coherent Self with the external world (Sherrington, 1906; Lashley, 1951; Sperry, 1952; Bernstein, 1967; Maturana and Varela, 1980) and then in companionship with other Self-aware subjects (Maturana and Varela, 1992; Bråten, 2009; Malloch and Trevarthen, 2009; Gallese, 2012).

In 1962, the effects of commissurotomy, undertaken to arrest life-threatening epilepsy, were investigated in a human patient (Bogen, 1985). Sperry directed psychological tests of consciousness, learning, and various skills after the operation (Sperry, 1968). Commissurotomy to treat multifocal epilepsy in other patients made it possible to replicate tests of rational consciousness in the left and right hemispheres; their similarities and differences; their convergent influence over the acts, ideas, and feelings of the whole person; and the role of language (Sperry, 1968, 1974, 1982; Sperry et al., 1969).

Commissurotomy research confirmed Sperry's theory of the causal potency of motor intentions (Sperry, 1952), and encouraged him to develop a theory of self-consistent conscious ideas and beliefs, and moral impulses (Sperry, 1977, 1983). It stimulated the neuropsychological analysis of hemispheric differences unique to humans, and, by clarifying the special modes of processing in cortical territories, supported the 'cognitive revolution' of the 1960s (Sperry, 1993).

Since the classical period of split-brain research in the 1960s and 1970s, interest in hemispheric differences has flourished, and the cognitive processes, language processes, and memory of commissurotomy patients have been studied in new, detailed ways. The effects of the operation on visual awareness, visual attention, and capacity to use language or make interpretations of configurations of items in the visual world have been fully confirmed (Holtzman et al., 1984; Benson and Zaidel, 1985; Nebes, 1990; Trevarthen, 1990; Corballis, 1991; Gazzaniga, 1995). Although commissurotomy was highly effective in limiting seizures, the operation carries risks, and, with improved means of controlling epilepsy, it is now rarely performed.

Split-Brain Experiments and Intelligent Movements of Cats and Monkeys

Myers applied the stereo-microscope surgery developed by Sperry to cut the crossover of visual nerves in the optic chiasm under a cat's brain, so each eye would lead to only one cerebral hemisphere (**Figures 1** and **2**). In 1953, he and Sperry reported the transfer of visual memory between the hemispheres in cats with the chiasm divided, and the abolition of this transfer when the corpus callosum was cut (Myers and Sperry, 1953). Outside carefully controlled test situations, the animals moved normally.

Monkeys also experience independent visual perception and memory in the two brain halves after the optic chiasm and forebrain commissures (corpus callosum and anterior commissure) are cut (**Figure 1**). Again, the operated animals retained normal coordination of movements and attention when free to orient, correcting for loss of awareness due to the surgery. Split-brain monkeys learn two, mutually contradictory,

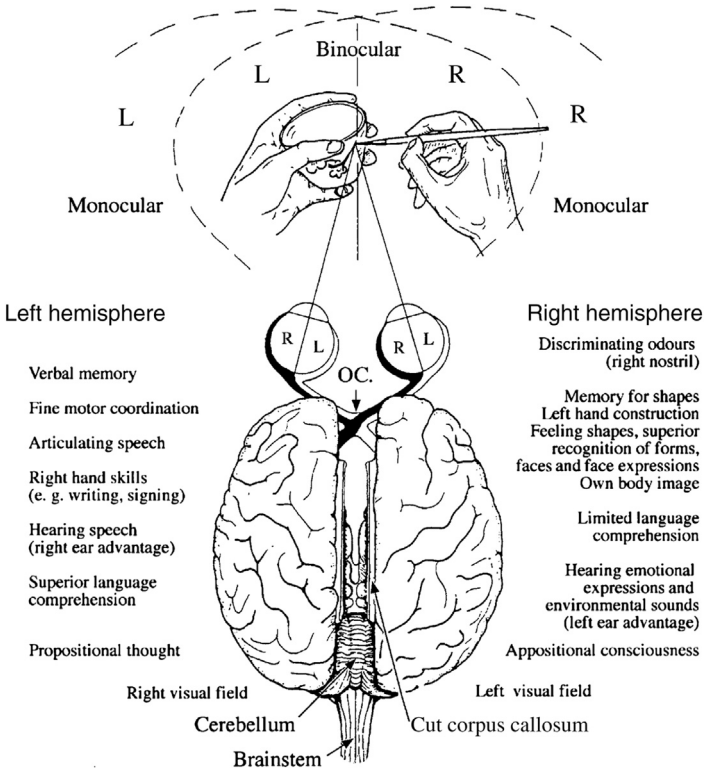

Figure 2 Human brain showing asymmetric functions in the cerebral hemispheres. The hemispheres are separated after commissurotomy (the 'split-brain' operation). Above: the two hands in the visual field, with the subject's attention fixated on the tip of the paintbrush. L = left half of visual field; R = right half of visual field; oc = optic chiasm.

visual discrimination habits in the right and left half-brains, showing that they possess double and separate realms of consciousness (Trevarthen, 1962; Sperry, 1977). In the cat, each disconnected hemisphere directed movements of the whole body, but with monkeys, a hand on the same side as one seeing hemisphere was, at least in the immediate postsurgical period, less willing to respond and, if forced into action, was clumsy, as if blind. Crossed pathways linking each cortex to the opposite hand were superior for guiding fine, purposeful movements of the fingers. Related split-brain projects demonstrated shifts of attention between the two separated halves of the cortex and the effects of preparatory sets to move in particular ways. The results revealed the global design of the mammalian brain for awareness, learning, and voluntary action, and the special adaptations for conscious control of the hands of primates (Sperry, 1961; Berlucchi, 1990).

What a chiasm–callosum sectioned (split-brain) monkey sees and remembers (i.e., in which hemisphere visual perception and learning occur when both hemispheres are presented stimuli) depends on which hand the animal chooses for response. This is evidence that intention can instruct experience and that activities of the neocortex are directed intrinsically by purposeful impulses generated in the brain as a whole, not just by processing the input of stimulus information at the cortical level (Trevarthen, 1987).

When a monkey learns a complex two-handed skill, the memory of how to move is located more in the hemisphere opposite the dominant hand (i.e., the hand that the individual spontaneously prefers to use for fine, consciously regulated manipulations) (Trevarthen, 1978a). A split-brain monkey's visual awareness can be turned on or off in one side of the brain by changing which hand the animal uses to manipulate the test apparatus to gain a reward. When using the right hand, the monkey becomes conscious on the left side of its cerebrum and is unable to see with the right cortex. This causes the monkey to be inattentive to things in the left half of its experience, and makes it likely to imagine things that did not exist there. This clarifies 'perceptual neglect' and 'illusory completion,' phenomena known from clinical neuropsychological research with consciousness of patients who have lost function in parts of their cerebral cortex.

Tests of vision in the divided brain led to the theory that the sensory systems are linked to a hierarchy of motor coordinative systems, which either gear body movements to the layout of the surrounding world, or direct the subject to select objects that have life-relevant qualities (Trevarthen, 1978b). Two active visual systems were distinguished: wide-field and part-conscious *ambient vision* informs whole-body locomotion and orientation; and *focal vision*, which employs the foveae to see within only a few degrees of visual angle and with precise intermittent fixation of gaze, inspects detail and identifies objects (Trevarthen, 1968, 1974a, 1987). The former, involving many subcortical centers as well as parts of the parietal cerebral cortex, acts as motivator and map reader for attention in the latter, which involves more conscious vision

of shapes, colors, and patterns. The split-brain operation divides focal vision because its essential processes are confined to the cortex, but it leaves ambient vision largely intact.

Human Commissurotomy Patients and the Cultural Mind

In 1960, Joseph Bogen, a Los Angeles neurosurgeon, saw a patient who was having progressive, life-threatening seizures as a result of head injuries sustained as a paratrooper. In the light of research with monkeys that showed that commissurotomy could leave essentially normal control of whole-body action and the distribution of awareness, it was decided that cutting the corpus callosum and other interhemispheric commissures, 'total commissurotomy,' would reduce the epilepsy. This operation was carried out in 1962, and the epilepsy improved significantly. Sperry and Bogen, with Michael Gazzaniga as research assistant, began systematic psychological tests at Caltech on the effects on perception, speech, and motor control. Instead of relying on standard neurological procedures, original tests were devised based on principles learned from split-brain animals (Sperry, 1968, 1974; Sperry et al., 1969). The first report, with clinical observations made by Bogen, described the resultant left-right division of human consciousness (Gazzaniga et al., 1962). Norman Geschwind and Edith Kaplan, in Boston, described comparable 'deconnection' phenomena in patients with lesions that interrupted cortico-cortical links (Geschwind and Kaplan, 1962). The main hemispheric differences revealed by the early work at Caltech with commissurotomy patients are summarized in Sperry's Nobel address (Sperry, 1982).

Tests confirmed that awareness of objects, either seen by half of the brain (in the left and right halves of the visual field) or felt in the left or right hand, was essentially normal, but, as in the earlier animal studies, when free orientation was prevented, the left and right mental realms were as separate as if in separate heads (**Figure 2**). Only when the subject shifted visual fixation or passed an object from one hand to the other did the one hemisphere know what the other perceived. As with the monkeys, human commissurotomy subjects experienced unified *ambient* visual functions but split *focal* vision (Trevarthen and Sperry, 1973; Trevarthen, 1990, 2004). This difference in functional level of control, defined by the motor activities of the subject in selective engagement with objects in the world, is of fundamental significance for interpretations of the meaning and value of knowledge gained by research on commissurotomy patients (see **Figure 2**).

Although both hemispheres of a commissurotomy patient proved to be conscious in all modalities of sense, only the left could speak or write. Surprisingly, however, given that left-hemisphere lesions can produce total aphasia, the right hemisphere, although speechless, could nevertheless comprehend speech (Gazzaniga and Sperry, 1967; Zaidel, 1990; Trevarthen, 2004). It picked up information quickly from utterances made by the left half-brain, could read common words flashed to the left visual field, and could follow verbal instructions. It was, however, conspicuously weak at calculating and in comprehending logical propositions. Jerre Levy, taking a lead from differences revealed by tests of intelligence, showed that in the nonspeaking right hemisphere, classically referred to as the 'minor' hemisphere, mental faculties that imagine spatial relations and transformations in space were more highly developed than in the left 'dominant' side of the brain. Thus, two different, complementary kinds of cognitive strategy, two kinds of consciousness, were demonstrated (Levi-Agresti and Sperry, 1968; **Figure 2**).

The surgically separated left hemisphere commands fluent speech and can generate and retain phonological images of speech sounds. It performs serial logical propositions such as are expressed in spoken, gestural, or written language, and it is adept at the kind of combinatorial and systematic logic that is required for formal mathematics employing symbols or score-based musical thought by skilled performers. The right hemisphere, in contrast, has superior mastery of the immediate 'appositional' experience (Bogen, 1969) – of forms perceived by sight, sound, or touch – and, according to recent research, of odors, which are detected and evaluated emotionally by the right nostril (which projects to the right hemisphere) but named by the left (Herz et al., 1999). The right hemisphere can remember affect-laden experiences such as these better, can distinguish and remember faces better, and has richer reactions to colors and to affective tones of the voice in song or speech and its prosodic quality or 'musicality' (Panksepp and Trevarthen, 2009; Turner and Ioannides, 2009). It perceives more easily resemblances in topological but not Euclidean geometry. Commissurotomy patients have a color anomia, an inability to name colors, because the disconnected right hemisphere has a superior but inarticulate ability to perceive and evaluate color (Levy and Trevarthen, 1981). There is evidence that an undivided brain-stem color vision defines a dichromatic distinction between blue space and the reddish-brown solids, part of the natural layering of the conscious mind to serve different forms of action in the world (Trevarthen and Sperry, 1973).

Each hemisphere separated from its partner by section of the corpus callosum can orient attention to left or right, and can experience illusions of experience in the ipsilateral field that has no clear vision, enabling perceptual completion, which is dependent on a relaxed motor state (Trevarthen, 1974a; Levy et al., 1972; Trevarthen, 1987, 2004). Jerre Levy and Colwyn Trevarthen used 'stimulus chimeras,' which are pictures showing different half objects on either side of the midline, to observe biases in cognitive activities of the hemispheres induced by what a commissurotomy patient intended to do, or by what he or she had been told to think about (Levy and Trevarthen, 1976). They called the motive sets that were capable of turning on or off one or other hemispheric cognitive mechanisms, sometimes independently of the different cognitive aptitudes on the two sides, 'metacontrol.' Similar habitual arousal biases in normal subjects may underlie individual differences in cognitive aptitudes and personalities. They indicate characteristics involving the core motivating mechanisms of the Self mediated by subcortical and brainstem integrative systems (Northoff and Panksepp, 2008), which have asymmetries derived from complementary 'ergotropic' and 'trophotropic' regulations of the autonomic system, concerned, respectively, with assertive action on the environment and with self-sustaining nurturance of vital

functions of the body (Hess, 1954; Trevarthen, 1996). Emotional asymmetries have received little attention in research with split-brain subjects.

Theoretical and Philosophical Implications and Problems: Seeking a Wider Scope for Enquiry

Iain McGilchrist (2009) presents a comprehensive review of evidence from neuropsychology, phenomenological psychology, and cross-cultural studies that the right hemisphere is adapted for vigilance, sustained alertness, and the emotional intensity of experience for novel events, while the left chooses and pays focused attention to elements, seeking recognition of what is known for rapid action. These complementary modes of awareness and attention to experience involve different motor activities and support different skills, for the individual and in communication. Both of the hemispheric modes of awareness are required for the mastery of cultural experience, and for the individual to become a contributor to elaborate and sophisticated life in society with its different employments.

Michael Gazzaniga has given special importance to the top-down 'interpretations' of focal interests and their sequencing in the left hemisphere. He concludes that human left-hemisphere intelligence, with language, overcomes the limited, more primitive, animal awareness of the right hemisphere by giving learned rational interpretations, in symbolic form, to what is perceived (Gazzaniga, 1985, 1998). This theory of human consciousness, privileging language, neglects the role of intuitive, implicit, emotional, and intersubjective contexts for the development of discriminating and rationalized awareness, and the importance of creative thinking for the capacity to imagine a less focused, wholistic view of elements in their context, and to imbue them with feeling (Bogen, 1997; Bogen and Bogen, 1969; Panksepp, 1998, 2005; McGilchrist, 2009). All of human cultural cooperation and production depend on intimate communication of this more passionate motivation, besides systematic rationality and its codification in symbols (Trevarthen, 2012). The lexicon of language is learned with awareness of whole-body expression and the intrinsic asymmetry of manual gesture (McNeill, 2000). If single subjects are required to make precise identification and naming of different forms of objects, with rational inferences to explain them, then the left-hemisphere abilities have an advantage. The special requirements of mutual, intersubjective intelligence, and how they may be mediated by special intentional and emotional capacities for shared experience that are constitutional in the human brain (Gallese, 2012), are not revealed.

Tests with commissurotomy patients support findings with brain-damaged and normal subjects showing the right half of the brain to be involved in emotional communication and the mediation of direct social interactions (McGilchrist, 2009). The development of asymmetry of behaviors in infancy, and the asymmetry of motives and emotions supporting maternal affection and care for young children, gives evidence that differences of function in disconnected hemispheres in adult humans arise from inherent differences in the motives and emotions of the left and right sides of the brain (Schore, 2000, 2012; Narvaez et al., 2012). Anatomical differences in the fetal body and brain (Trevarthen, 1974b, 1985, 1996, 2001) and fetal hearing of the mother's heartbeat and preferential recognition of her voice by the right hemisphere (DeCasper and Prescott, 2009) confirm that both emotional differences and predispositions for different learning abilities are established by morphogenesis. They form as systems that will play a part at all stages of the learning of language and cooperative skills (Trevarthen and Delafield-Butt, 2013).

Commissurotomy research has stimulated neuropsychological investigation of the effects of unilateral brain injury; tests of normal perceptual, cognitive, and motoric asymmetries; and research on the anatomical asymmetries of the cerebral cortex. The findings have been invoked to explain why children with similar opportunities for learning may be different in educational performance, and why adults show different intellectual aptitudes and job preferences. It is likely that hemispheric biases in cognition and intellectual differences between adults are both consequences of inherent styles in purposes and emotions, and of different experiences and opportunities for practice and for communication with others. There is good reason to believe that, for a species that depends upon sharing of elaborate cultural knowledge and skills that require many complementary talents and inclinations, the development of the brain relies upon both genetic and epigenetic production of individual differences, such as is evident in the variety of physical appearance; of skin color, hair, eyes, and facial configuration; and of manners of moving and voice, which are so evident in humans. In addition, the brains of modern humans are particularly susceptible to modifications by cultural experience and artistry, which enhance individualities (Han and Northoff, 2008; Trevarthen, 2012).

In the normal brain of an individual, the hemispheres work in tight partnership. The intuitive response to experience and the evocation of imagery linked to phenomenal reality by metaphor, on the one hand, and verbal analysis and prescription of esthetic judgments, on the other, are two natural brain systems that develop as complements in the making and communicating of language, technology, and art.

Findings with commissurotomy patients have been used to support theories of 'modular intelligence' derived from the neuropsychological and cognitive science fields, and attempts have been made to apply these theories to evolutionary changes in hominid behavior and the different kinds of cognitive processing that are inferred. However, such models of the mind are incomplete if they do not incorporate explanations of the motivations that generate both action and awareness, and the communicable emotions that evaluate them (Trevarthen, 2004). The skills, concepts, and memories of language are stored over wide areas. Parallel-distributed processes contribute to the comprehension of language, as for facial recognition, spatial awareness, and categorical understanding of objects, or for executive planning and motor sequencing. The development of the cortical elements and assemblies involved in all of these functions depends on associative cortico-cortical links, both within each hemisphere and through the corpus callosum. These supplement equally massive connections to and from cell populations beneath the cortex, which integrate intentional

self-awareness and give emotional appraisal to the expected consequences of agency (Panksepp and Panksepp, 2000). The human subcortex is capable of integrating a conscious motivation of movements of the whole body, including the expression of socially responsive emotions even in the absence of the neocortex (Merker, 2007).

A majority of humans are 'right-handed' and have been since Neolithic times, preferring to take conscious control in the right hand, especially for communicating by gesture, carefully directing implements, and finishing the details of their fabrication. About 15% prefer the left hand. Most of us tend to move the right half of our mouths more, or earlier, in speaking, and look right as we speak. These skills, elaborated by learning, originate in asymmetries implanted in subcortical systems before birth (Trevarthen, 1996).

It has been clear for more than 100 years that talk, and especially the perception and execution of the rapid articulations of mature speech, is more strongly represented in the left hemisphere. There is now evidence that the learned performative skills of singing and instrumental music have the same laterality to the left hemisphere as skilled speech, but the affective qualities of music are more strongly appreciated in the right hemisphere (Turner and Ioannides, 2009). Right-hemisphere lesions selectively impair object and event recognition by sight or hearing and may have devastating effects on the capacity to recognize who another person is and what their emotion may be. Right-posterior lesions can also weaken a person's discrimination of his or her own expression of feeling, and lower self-awareness generally. The cognitive and performance functions of the right half of the brain are clearly important to a person's cultural life in a community, and to the acquisition of many skills that language may find hard to describe, and they serve to support language-mediating functions in the left hemisphere.

One important defect of awareness found in commissurotomy patients, contrasting with the overall normality of their behaviors and experiences when free to orient their actions, is alexithymia, a difficulty in identifying and describing feelings (TenHouten et al., 1986). It appears to be a straightforward consequence of dissociation between the left hemisphere's abilities to identify and describe experiences and the right hemisphere's immediate affective response to a moving event. This dissociation interferes with the function of the corpus callosum to mediate between creative imagination and factual specification, which goes beyond the function of transferring perceptions and memories (Bogen and Bogen, 1983). Dahlia Zaidel found that disconnection of complementary aspects of experience also interferes with the formation of new memories following the separation of the hemispheres (Zaidel, 1990).

There has been much speculation about how recurring asymmetries in works of art might be related to the anatomy of the brain, and our common brain asymmetry does appear to have been unknowingly exploited by many painters to generate esthetic effects (Gardner, 1982; Levy, 1988; Trevarthen, 1995). There is a school of art teachers who draw on split-brain research to guide pupils to make stronger use of the right hemisphere, claiming that this generates more dynamic, coherent, and aesthetically pleasing forms, and some require right handers to sketch unthinkingly with their left hands (Edwards, 1979). We have to learn how to talk about what is beautiful in art, and it is inherently difficult to recall what we say. Language offers a powerful tool for introspecting, for planning actions, and for manipulating information in 'phenomenal absence' from its source (i.e., in memory), but the lexicon of speech is limited for representation of the quality of imagination, and it seeks aid from expressive gesture, intonation, and metaphor. Musical communication, which is universal among humans, serves to express affective relations and to establish a sense of belonging to a community of vital agents who share emotional appraisals of companionship in experience, from infancy (Malloch and Trevarthen, 2009).

Sperry's work with commissurotomy patients convinced him that coherent, Self-related brain-generated awareness and evaluation have a determining role over the experience obtained by learning. He concluded that science has to give consciousness an interactive, causal role in brain function. Subjective mental states explain conscious behavior and its changes, and they determine what a person is and does. This view gives the higher 'macro,' mental, vital, and social forces equal importance in scientific explanations of life, such as in physics and chemistry. However, it is a conception of consciousness as a top-down process, and there are good grounds for taking a rather different view that sees consciousness as a product of the emotions that constitute awareness of a 'core self,' mediated from the brain stem, below the elaborations of memory in the cerebral cortex (McGilchrist, 2009: pp. 220, 221). Sperry's thinking about human consciousness acknowledges the importance of values and emotion, as is summarized in the *Nobel Prize Conversations* (Sperry et al., 1985).

See also: Asymmetry of Body and Brain; Attention, Neural Basis of; Brain Asymmetry; Brain Development: Ontogenetic Neurobiology of; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuroscience; Consciousness, Neural Basis of; Cortical Areas Engaged in Movement: Neuroimaging Methods; Epilepsy; Face Recognition, Psychological and Neural Aspects; Gesture in Linguistics; Hemispatial Neglect, Neural Basis of; Human Spatial Orientation, Neural Basis of; Interhemispheric Interaction in Language and Cognitive Processes; Learning and Memory, Neural Basis of; Mental Imagery: Visual Cognition; Neural Basis of Language Development; Neural Representations of Objects; Neuropsychological Testing; Occipital Lobe (Visual Cortex): Functional Aspects; Perception of Action Space: Using Multiple Frames of Reference; Person and Self: Philosophical Aspects; Premotor Cortex; Prosopagnosia; Self and Brain; Sign Language; Speech Production, Neural Basis of; Visual Imagery; Visual Perception, Neural Basis of; Visual Streams: Dorsal and Ventral.

Bibliography

Akelaitis, A.J., 1944. A study of gnosis, praxis, and language following section of the corpus callosum and anterior commissure. Journal of Neurosurgery 1, 94–102.

Benson, D.F., Zaidel, E., 1985. The Dual Brain. Guilford Press, New York.

Berlucchi, G., 1990. Commissurotomy studies in animals. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier, Amsterdam, Oxford, New York, pp. 9–47.
Bernstein, N., 1967. Coordination and Regulation of Movements. Pergamon, New York.
Bogen, J.E., 1969. The other side of the brain, II. An appositional mind. Bulletin of the Los Angeles Neurological Society 34, 135–162.
Bogen, J.E., 1985. The callosal syndromes. In: Heilman, K.M., Valenstein, R. (Eds.), Clinical Neuropsychology, second ed. Oxford University Press, New York, pp. 295–338.
Bogen, J.E., 1997. Does cognition in the right hemisphere require right hemisphere language? Brain and Language 57, 12–21.
Bogen, J.E., Bogen, G.M., 1969. The other side of the brain III: the corpus callosum and creativity. Bulletin of the Los Angeles Neurological Society 34, 175–195.
Bogen, J.E., Bogen, G.M., 1983. Hemispheric specialization and cerebral duality. Behavioral and Brain Sciences 6 (3), 517–520.
Bogen, J.E., Fisher, E.D., Vogel, P.J., 1965. Cerebral commissurotomy: a second case report. Journal of the American Medical Association 194, 1328–1329.
Bråten, S., 2009. The Intersubjective Mirror in Infant Learning and Evolution of Speech. John Benjamins Publishing Company, Amsterdam/Philadelphia.
Corballis, M.C., 1991. The Lopsided Ape: Evolution of the Generative Mind. Oxford University Press, New York.
DeCasper, A.J., Prescott, P., 2009. Lateralized processes constrain auditory reinforcement in human newborns. Hearing Research 255, 135–141.
Edwards, B., 1979. Drawing on the Right Side of the Brain. J-.P. Tarcher, Los Angeles.
Evarts, E.V., 1990. Foreword: coordination of movement as a key to higher brain function: Roger W. Sperry's contributions from 1939 to 1952. In: Trevarthen, C. (Ed.), Brain Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry. Cambridge University Press, New York.
Gallese, V., 2012. Introduction. In: Ammaniti, M., Gallese, V. (Eds.), The Birth of Intersubjectivity. Norton, New York.
Gardner, H., 1982. Art, Mind and Brain. Basic Books, New York.
Gazzaniga, M.S., 1985. The Social Brain. Basic Books, New York.
Gazzaniga, M.S., 1995. Principles of human brain organization derived from split-brain studies. Neuron 124, 217–228.
Gazzaniga, M.S., 1998. Brain and conscious experience. Advances in Neurology 77, 181–192.
Gazzaniga, M.S., Sperry, R.W., 1967. Language after section of the cerebral commissures. Brain 90, 131–148.
Gazzaniga, M.S., Bogen, J.E., Sperry, R.W., 1962. Some functional effects of sectioning the cerebral commissures in man. Proceedings of the National Academy of Sciences United States of America 48, 1765–1769.
Geschwind, N., Kaplan, E., 1962. A human cerebral deconnection syndrome. Neurology 12, 675–685.
Han, S., Northoff, G., 2008. Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nature Reviews Neuroscience 9, 646–654.
Herz, R.S., McCall, C., Cahill, L., 1999. Hemispheric lateralization in the processing of odor pleasantness versus odor names. Chemical Senses 24 (6), 691–695.
Hess, W.R., 1954. Diencephalon: Autonomic and Extrapyramidal Functions. Grune and Stratton, Orlando, FL.
Holtzman, J.D., Volpe, B.T., Gazzaniga, M.S., 1984. Spatial orientation following commissural section. In: Parasuraman, R., Davies, D.R. (Eds.), Varieties of Attention. Academic Press, New York, pp. 375–394.
Innocenti, G.M., 1986. General organization of the callosal connections in the cerebral cortex. In: Jones, E.G., Peters, A. (Eds.), Cerebral Cortex, Sensory-Motor Areas and Aspects of Cortical Connectivity, vol. 5. Plenum, New York, pp. 291–353.
Lashley, K.S., 1951. The problems of serial order in behavior. In: Jeffress, L.A. (Ed.), Cerebral Mechanisms in Behavior. Wiley, New York, pp. 112–136.
Levi-Agresti, J., Sperry, R.W., 1968. Differential perceptual capacities in major and minor hemispheres. Proceedings of the National Academy of Sciences United States of America 61, 1151.
Levy, J., 1988. Cerebral asymmetry and aesthetic experience. In: Rentschler, I., Herzberger, B., Epstein, D. (Eds.), Beauty and the Brain. Biological Aspects of Aesthetics. Birkhauser Verlag, Berlin, pp. 219–242.
Levy, J., Trevarthen, C., 1976. Metacontrol of hemispheric function in human split-brain patients. Journal of Experimental Psychology: Human Perception and Performance 2, 299–312.
Levy, J., Trevarthen, C., 1981. Color-matching, color-naming and color-memory in split brain patients. Neuropsychologia 19 (4), 523–541.
Levy, J., Trevarthen, C., Sperry, R.W., 1972. Perception of bilateral chimeric figures following hemispheric deconnection. Brain 95, 61–78.
Malloch, S., Trevarthen, C., 2009. Communicative Musicality: Exploring the Basis of Human Companionship. Oxford University Press, Oxford.
Maturana, H., Varela, F., 1980. Autopoiesis and Cognition: The Realization of the Living. Reidel, Boston, MA.
Maturana, H., Varela, F., 1992. The Tree of Knowledge: The Biological Roots of Human Understanding. Shambhala, Boston, MA.
McGilchrist, I., 2009. The Master and His Emissary: The Divided Brain and the Making of the Western World. Yale University Press, Newhaven and London.
McNeill, D., 2000. Language and Gesture. Cambridge University Press, Cambridge, UK.
Merker, B., 2007. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. Behavioral and Brain Sciences 30, 63–134.
Myers, R.E., Sperry, R.W., 1953. Interocular transfer of a visual form discrimination habit in cats after section of the optic chiasm and corpus callosum. The Anatomical Record 115, 351–352.
Narvaez, D., Panksepp, J., Schore, A., Gleason, T. (Eds.), 2012. Human Nature, Early Experience and the Environment of Evolutionary Adaptedness. Oxford University Press, New York.
Nebes, R.D., 1990. The commissurotomised brain. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier, Amsterdam, Oxford, New York, pp. 3–168.
Northoff, G., Panksepp, J., 2008. The trans-species concept of self and the subcortical–cortical midline system. Trends in Cognitive Sciences 12 (7), 259–264.
Panksepp, J., 1998. Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press, New York.
Panksepp, J., 2005. Affective consciousness: core emotional feelings in animals and humans. Consciousness and Cognition 14, 19–69.
Panksepp, J., Panksepp, J.B., 2000. The seven sins of evolutionary psychology. Evolution and Cognition 6 (2), 108–131.
Panksepp, J., Trevarthen, C., 2009. The neuroscience of emotion in music. In: Malloch, S., Trevarthen, C. (Eds.), Communicative Musicality: Exploring the Basis of Human Companionship. Oxford University Press, Oxford, pp. 105–146.
Schore, A.N., 2000. The self-organization of the right brain and the neurobiology of emotional development. In: Lewis, M., Granic, I. (Eds.), Emotion, Development, and Self-Organization. Dynamic Systems Approaches to Emotional Development. Cambridge University Press, New York, pp. 155–185.
Schore, A.N., 2012. Emotion development in infancy. In: Narvaez, D., Panksepp, J., Schore, A., Gleason, T. (Eds.), Human Nature, Early Experience and the Environment of Evolutionary Adaptedness. Oxford University Press, New York.
Sherrington, C.S., 1906. The Integrative Action of the Nervous System. Yale University Press, New Haven.
Sperry, R.W., 1952. Neurology and the mind-brain problem. American Scientist 40, 291–312.
Sperry, R.W., 1961. Cerebral organization and behavior. Science 133, 1749–1757.
Sperry, R.W., 1968. Mental Unity Following Surgical Disconnection of the Cerebral Hemispheres: The Harvey Lectures, series 62. Academic Press, New York, pp. 293–323.
Sperry, R.W., 1974. Lateral specialization in the surgically separated hemispheres. Ch. I. In: Schmitt, F.O., Worden, F.G. (Eds.), The Neurosciences. Third Study Program, vol. 3. MIT Press, Cambridge, MA, pp. 5–19.
Sperry, R.W., 1977. Forebrain commissurotomy and conscious awareness. The Journal of Medicine and Philosophy 2 (2), 101–126.
Sperry, R.W., 1982. Some effects of disconnecting the cerebral hemispheres (Nobel lecture). Science 217, 1223–1226.
Sperry, R.W., 1983. Science and Moral Priority. Columbia University Press, New York.
Sperry, R.W., 1993. The impact and promise of the cognitive revolution. American Psychologist 48 (3), 878–885.
Sperry, R.W., Gazzaniga, M.S., Bogen, J.E., 1969. Interhemispheric relationships: the neocortical commissures; syndromes of hemisphere disconnection. In: Vinken, P.J., Bruyn, G.W. (Eds.), Handbook of Clinical Neurology, vol. 4. North Holland, Amsterdam, pp. 273–290.
Sperry, R.W., Eccles, J., Prigogine, I., Josephson, B., Cousins, N., 1985. Nobel Prize Conversations. Saybrook, Dallas, TX.
TenHouten, W., Hoppe, K., Bogen, J., Waiter, D., 1986. Alexithymia: an experimental study of cerebral commissurotomy patients and normal control subjects. American Journal of Psychiatry 143, 312–316.
Trevarthen, C., 1962. Double visual learning in split-brain monkeys. Science 136, 258–259.
Trevarthen, C., 1965. Functional interactions between the cerebral hemispheres of the split-brain monkey. In: Ettlinger, E.G. (Ed.), Functions of the Corpus Callosum. Ciba Foundation Study Group, No. 20. Churchill, London, pp. 24–40.
Trevarthen, C., 1968. Two mechanisms of vision in primates. Psychologische Forschung 31, 299–337.
Trevarthen, C., 1974a. Functional relations of disconnected hemispheres with the brain stem and with each other: monkey and man. In: Kinsbourne, M., Smith, W.L.

(Eds.), Hemispheric Disconnection and Cerebral Function. Charles C. Thomas, Springfield, IL, pp. 187–207.

Trevarthen, C., 1974b. Cerebral embryology and the split-brain. In: Kinsbourne, M., Smith, W.L. (Eds.), Hemispheric Disconnection and Cerebral Function. Charles C. Thomas, Springfield, IL, pp. 208–236.

Trevarthen, C., 1978a. Manipulative strategies of baboons and the origins of cerebral asymmetry. In: Kinsbourne, M. (Ed.), Asymmetrical Function of the Brain. Cambridge University Press, New York and London, pp. 329–391.

Trevarthen, C., 1978b. Modes of perceiving and modes of acting. In: Pick, J.H. (Ed.), Psychological Modes of Perceiving and Processing Information. Erlbaum, Hillsdale, NJ, pp. 99–136.

Trevarthen, C., 1985. Neuroembryology and the development of perceptual mechanisms. In: Falkner, F., Tanner, J.M. (Eds.), Human Growth, second ed. Plenum, New York, pp. 301–383.

Trevarthen, C., 1987. Sub-cortical influences on cortical processing in 'split' brains. Wenner-Gren International Symposium Series. In: Ottoson, D. (Ed.), Duality and Unity of the Brain, vol. 47. Macmillan: Basingstoke, Hampshire/Stockton Press: New York, pp. 382–415.

Trevarthen, C., 1990. Integrative functions of the cerebral commissures. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier, Amsterdam, pp. 49–83.

Trevarthen, C., 1995. Mother and baby – seeing artfully eye to eye. In: Gregory, R., Harris, J., Rose, D., Heard, P. (Eds.), The Artful Eye. Oxford University Press, Oxford, pp. 157–200.

Trevarthen, C., 1996. Lateral asymmetries in infancy: implications for the development of the hemispheres. Neuroscience and Biobehavioral Reviews 20 (4), 571–586.

Trevarthen, C., 2001. The neurobiology of early communication: intersubjective regulations in human brain development. In: Kalverboer, A.F., Gramsbergen, A. (Eds.), Handbook on Brain and Behavior in Human Development. Kluwer, Dordrecht, The Netherlands, pp. 841–882.

Trevarthen, C., 2004. Split brain and the mind. In: Gregory, R.L. (Ed.), Oxford Companion to the Mind. Oxford University Press, Oxford, pp. 872–878.

Trevarthen, C., 2012. Born for art, and the joyful companionship of fiction. In: Narvaez, D., Panksepp, J., Schore, A., Gleason, T. (Eds.), Human Nature, Early Experience and the Environment of Evolutionary Adaptedness. Oxford University Press, New York, pp. 202–218.

Trevarthen, C., Delafield-Butt, J., 2013. Biology of shared experience and language development: regulations for the inter-subjective life of narratives. In: Legerstee, M., Haley, D., Bornstein, M. (Eds.), The Infant Mind: Origins of the Social Brain. Guildford Press, New York, pp. 167–199.

Trevarthen, C., Sperry, R.W., 1973. Perceptual unity of the ambient visual field in human commissurotomy patients. Brain 96, 547–570.

Turner, R., Ioannides, A., 2009. Brain, music and musicality: inferences from neuroimaging. In: Malloch, S., Trevarthen, C. (Eds.), Communicative Musicality: Exploring the Basis of Human Companionship. Oxford University Press, Oxford, pp. 147–181.

Zaidel, D.W., 1990. Memory and spatial cognition following commissurotomy. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier, Amsterdam, Oxford, New York, pp. 151–166.

Zaidel, E., 1990. Language functions in the two hemispheres following complete commissurotomy and hemispherectomy. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 4. Elsevier, Amsterdam, Oxford, New York, pp. 115–150.

Sport, Anthropology of

Noel Dyck, Simon Fraser University, Burnaby, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E.P. Archetti, volume 22, pp. 14913–14916, © 2001, Elsevier Ltd.

Abstract

Although traditionally inclined to view sport as an insubstantial topic for investigation, anthropologists have in recent decades devoted increasing attention to the powerful attractions of games, sports, and embodied athletic practices, along with their effects within social life. Working within and beyond conventional sociological definitions of sport, anthropologists have applied ethnographic and comparative methods to good effect. Treating sport as embodied social practice that hinges around its playfulness and celebratory capacities, they are examining its varied meanings and uses, which range from the formulation of individual and cultural identities to the facilitation of domestic, economic, and state purposes.

The Anthropology of Sport

The positioning of anthropology in the nineteenth century as the science of the nonmodern and 'primitive' drew its practitioners away from the study of athletic games and contests, not least because recently codified Western sports were being proclaimed as exemplars of modernity (Archetti, 1998b). But the ensuing anthropological neglect of sport and sportlike activities was never absolute. Several early ethnographers and ethnologists crafted detailed descriptions of Native American games (e.g., Mooney, 1890; Culin, 1907). For the most part, however, games and athletic competitions that anthropologists stumbled across while conducting field research tended to be noted sporadically, if at all, and regarded as incidental to more definitively 'indigenous' social and cultural customs. Only occasionally did individual anthropologists (e.g., Firth, 1930; Fox, 1961) make a game or sport a focal part of an ethnographic report.

This pattern began to alter in the later decades of the twentieth century as part of a more general rethinking of the objectives and practices of sociocultural anthropology triggered by a rising tide of processual and performance-oriented theoretical approaches. This prompted, among other things, an easing of the inclination to rate games and sports as ephemeral and inconsequential pastimes. In rapid order there appeared a seminal essay on the Balinese cockfight (Geertz, 1972), an ethnographic film on Trobriand cricket as an 'ingenious response to colonialism' (Leach, 1976), and a monograph that traced the dynamics of Afghan political leadership through the equine sport of *buzkashi* (Azoy, 1982). These and other accounts raised novel ethnographic and analytic possibilities by looking into games and sporting activities played by peoples living in places that had figured prominently within the disciplinary literature. The sport practices and passions of persons and groups recounted in such works could not readily be dismissed as falling outside the bounds of anthropology.

Studies focusing upon the ritual forms, social relationships, cultural symbols, and performative identities generated through games or sports of various types also began to be mounted beyond the boundaries of traditional anthropological locales. For instance, essays about baseball in the United States (Gmelch, 1972), American football (Arens, 1975; Dundes, 1978), and wrestling in small-town Ontario (Freedman, 1983) guided anthropological readers into less usual ethnographic settings, as did MacAloon's pioneering study of the modern Olympic games (1981). These and other works cumulatively placed the investigation of sports, athletic competitions, indigenous games, and comparable forms of embodied practice within the ambit, if not necessarily at the center, of anthropological endeavor. By the end of the twentieth century, anthropologists had published analyses of, for example, traditional forms of wrestling (Alter, 1992) and cricket (Appadurai, 1995) in India, the rise of modern sports in post-Maoist China (Brownell, 1995), baseball in the Dominican Republic (Klein, 1991), capoeira in Brazil (Lewis, 1992), bodybuilding in the United States (Klein, 1993), the 1994 Winter Olympic games in Norway (Klausen, 1999a), and football, polo, and tango in Argentina (Archetti, 1999a). The capacities of sport to engage players and spectators as well as to mobilize involvement by sport organizers, politicians, and entrepreneurs featured in some anthropological conference papers and publications. What remained to be seen was whether the study of sport would progress beyond the status of a permissible niche enthusiasm to become a more widely read and revealing branch of inquiry within the discipline.

Anthropologists come to the study of sport by various routes. Taking cognizance of a sport played or an embodied practice conducted within a certain setting might be an unexpected outcome of examining other matters, as was the case in Manning's ethnographic encounter with cricket in the course of mapping the articulation between politics and festivals in the Caribbean (Manning, 1981). Delineating the attributes of a game or embodied practice might be driven by ethnographic curiosity as much as by an interest in sport per se. Reporting these findings in a subsequent publication might complete a one-off encounter with sport for one anthropologist or mark a more fundamental reorientation of intellectual interests for another. Either way, anthropologists whose attention is drawn to the study of a sport or sportlike practice must choose whether to conduct their inquiries primarily within the rubric of anthropology, locate these within the burgeoning interdisciplinary field of sport studies, or, seek to straddle, if not necessarily reconcile, these fields of study.

An underlying consideration in such decisions involves the contested manner in which sport has been envisioned and defined by scholars. Historians and sociologists have largely

tended to reserve the term 'sport' to refer to athletic games and practices formalized in Western industrial societies during the nineteenth and twentieth centuries. Thus, Guttmann's (1978) rendering of modern sports characterized these with regard to (1) their intrinsically secular nature; (2) the provision of equality of opportunity to compete and the creation of equal conditions of competition (both of which are vital to the establishment of the achievement principle); (3) the specialization of roles in modern sport activities; (4) the rationalization of modern sports through the adoption of standardized rules; (5) the bureaucratic organization of sports to determine rules and manage competitions; (6) the quantification of competition results to establish national and international standards of achievement; and (7) the concept of the record, which was a logical outcome of the quantification of sports. In Guttmann's view, 'primitive' sports were sometimes secular but shared none of the other six characteristics, while Greek, Roman, and Medieval European sports possessed some but not all of these indispensable properties of modern sports.

In contrast, Blanchard, a cultural anthropologist, offered a less historically and culturally fixed interpretation of sport, defining it as "a game-like activity having rules, a competitive element, and requiring some form of physical exertion" (Blanchard, 1995: 9). The broader and narrower uses to which the term sport might be put has been noted by Elias, a sociologist who warned that minimizing the differences and maximizing the similarities between modern sports and the game-contests of antiquity yields a distorted representation of the relationships between them (Elias, 1986: 129–132). Nevertheless, what is included or excluded from definitions of sport may, as Sutton-Smith suggests, tell us more about the purposes of those locked into such definitional exercises than the phenomena they purport to explain (Sutton-Smith, 1995: xii; 1997). Similarly, MacAloon (1987) has observed that the category of sport emerged historically and has varied over time. To insist upon reducing all of the variation this entails in order to shore up a general analytical definition of sport would, in his words, be to risk losing the object "in the act of trying to be sociological about it" (MacAloon, 1987: 106). Paraphrasing Levi-Strauss, he counsels that, in the study of sport, as in the study of anthropology, "the general patterns we seek lie neither in similarities nor differences, but in the *similarities among the differences*" (MacAloon, 1987: 112).

Scholarly writing on sport has, thus, tended to branch in two directions. One is to define the nature of sport and, in particular, 'modern sport' as tightly as possible in order to limit the range of the field and maintain the centrality and significance of certain definitive aspects of these activities. The other is to widen the range of activities to be considered, thereby opening possibilities for identifying previously unrecognized commonalities between different modes of sport and embodied athletic activities. Although some anthropologists work under the aegis of the first approach, which is the preferred one within the field of sport studies, others are drawn to the second. This more inclusive approach affords conceptual space for ethnographic investigation of forms of embodied practice that in some respects resemble and may even reference particular aspects of institutionalized conventional sport activities without needing to be formally labeled as such. Ruling on whether, for instance, practicing a martial art, running in a large-scale civic marathon race, or participating in a figure skating competition meets the threshold for what sport ought to entail is beside the point. When those who engage in a form of embodied practice reckon it to be even somewhat 'like a sport' and themselves to be performing 'like athletes,' it behoves anthropologists to take notice of how these practices, views, and beliefs are enacted and with what effects.

Techniques of the Body

Crucial to the study of sport and related forms of body practice are the ways in which, from one society to another, people not only use their bodies but also acquire embodied techniques that connect the realm of everyday body practices to the public world of shared performances (MacAloon, 1984). Identifying the body as a social as well as psychological and biological phenomenon, Marcel Mauss – a preeminent figure of early twentieth-century anthropology – nominated education (or training) as being dominant among all elements of what he termed the art of using the human body (Mauss, 1973: 73). Proceeding from the concrete to the abstract, Mauss retraced the steps by which he had come to recognize the body as "man's first and most natural instrument. Or more accurately, not to speak of instruments, man's first and most natural technical object, and at the same time technical means, is his body" (1973: 75). Summing up the social dimensions exhibited, for instance, by differing styles of walking, Mauss concluded that there was perhaps no 'natural' way for an adult to walk, but only one or another socially mediated and transmitted form of perambulation.

Mauss's formulation of techniques of the body prefigured foundational aspects of Foucault's (1977, 1978, 1985, 1986) writings on the body as a site of discipline and power as well as Bourdieu's (1984, 1990) commentaries on sport. Mauss selected the Latin word *habitus* as the appropriate one for referencing habits that differ not simply between individuals but especially between societies, educations, class proprieties and fashions: "In them we should see the techniques and work of collective and individual practical reason rather than in the ordinary way, merely the soul and its repetitive faculties" (1973: 73). What Mauss bequeathed to those curious about the social dynamics and meanings of body techniques – including those employed in running, swimming, and dancing – was an analytical framework within which ethnographic observations about the varying ways in which people use their bodies might be linked to larger social processes and purposes.

Augmenting one's capacities to acquire and master specific techniques remains part and parcel of embodied practice within any sport. This may be worked toward by myriad means, ranging from close observation and imitation of the actions of especially respected and accomplished performers to programs of formal instruction directed by coaches, trainers, or masters. Systems of exercise, diet, and rest, not to mention the use of sundry ostensibly performance-enhancing substances, may also be applied to extend the capabilities of athletes and other performers. But it is rarely just bodies that are being targeted for modification.

The regimes of discipline to which athletes explicitly or incidentally commit themselves by opting to play a given sport

or take up a specified form of embodied practice may also be enlisted to sculpt their character and dispositions, first and foremost through the structured training of their bodies. Shaping and preparing an athlete's body may, therefore, be treated as a means for improving not simply embodied performances but also, through these, a person's social behavior and moral worth. This premise underpinned the Victorian doctrine of muscular Christianity, which advocated the melding of Christian evangelism with vigorous and athletic masculinity, diffusing it across the British Empire and beyond (Alter, 2004, 2006; Foster, 2006; MacAloon, 2006). More than a century later, Wacquant (2004) shows how boxers on Chicago's South Side devote themselves to a demanding bodily craft that, nonetheless, distinguishes them locally as persons deserving respect by virtue of their demonstrated self-control. In similar manner, students and practitioners of the Japanese martial art of aikido transfer corporeal lessons and principles imparted during practice sessions to other parts of their lives (Kohn, 2003, 2008). Highly visible body modification projects of the type undertaken by bodybuilders can also contribute to refashioning one's social identity and sense of self (Bolin, 1992, 1997; Klein, 1993; Rapport, 2007).

The social and existential ends to which sport and other forms of embodied practice can be turned by participants and promoters are not readily separable from the techniques of the body through which these are sought. Alongside the possibility that sport might be employed to celebrate cultural ideals or to recast individual lives runs a stream of practical considerations and physical consequences. Activity-induced pain and injuries of various kinds (Howe, 2004) are by no means infrequent experiences for amateur and professional athletes, dancers, and those who 'work out' aerobically or undergo weight training routines. Attentiveness to the workings of one's body is required of those bent on taking part in athletic activities without jeopardizing either their ability to play or their general health. By the same token, addressing and bolstering the performance of bodies shaped by birth defects, serious accidents, or aging become compelling matters for those intent upon participating in sports or other forms of embodied practice in spite of these conditions (Howe and Jones, 2006; Howe, 2011; LeClair, 2009).

Anthropologists have by and large been reluctant to entertain approaches that might prove susceptible to biologically deterministic explanations of human and social development. As mentioned by Besnier and Brownell (2012), one of the challenges initially facing anthropological studies of sport was to unleash these from "essentialized concepts of national character anchored in biological bodies" (Besnier and Brownell, 2012: 444). Recently, however, embodiment has begun to be perceptively interrogated by anthropologists who are charting connections between the sociocultural, biophysical, and neurological facets of sport (e.g., Downey, 2005a,b, 2007, 2010; Sands, 2010). The most striking of these findings are reported in detailed and reflexive ethnographic accounts of what individual researchers have encountered while endeavoring to comprehend and acquire body techniques not previously familiar to them. The difficulties acknowledged by Bourdieu (1990: 166) in articulating "things we understand only with our bodies, outside conscious awareness" are being tackled through what Samudra (2008) has termed as 'thick participation.' Downey's meticulous analyses of the Brazilian martial art of capoeira lead him to conclude that since sensory capacities can in part be cultivated like a skill, practitioners of capoeira are doing more than simply acquiring techniques of the body: they are also tuning their perceptions and actions, thereby creating a culturally and neurologically distinctive way of being (2005a: 98–102, 146, 204–205). The upshot of this for the study of sport is that physical movements and skills are not merely culturally meaningful but also physiologically and neurologically consequential (Downey, 2010). Put another way, biology may be shaped in part by the types of cultural preferences manifested in athletic practices.

Sport as Social Practice

In addition to featuring distinctive techniques of the body, sport produces socially patterned activities engaged in for amusement, recreation, competitive achievement, or, in the case of sport professionals, livelihood. The more-or-less elaborately organized games, competitions, and athletic practices that constitute sport take assorted forms, ranging from individual or team sports, or some combination of the two; competitions staged on playing fields, ice rinks, in swimming pools, and other venues; to games played by recreational, amateur, or professional athletes. When conversing about sport, individuals' affinity for and experience with particular games or forms of athletic practice provide convenient templates. Enthusiasts of, for example, football or ice hockey may not bother to accommodate tennis or taekwondo within their notions of what sport comprises. In contrast, anthropologists, who intentionally expand the domain of games and activities to be considered as forms of athletic practice, introduce a comparative approach into the study of sport activities and relationships. Their inquiries attempt to recognize both the specificity and subtlety of differing types of games and sport forms as well as commonalities and disjunctions encountered across this vast and dispersed field of human endeavor.

An obvious starting point in examining sport as social practice involves disentangling the sets of roles performed in conjunction with athletic events. Along with players, athletes, or competitors, there can also be – depending upon the complexity of a game or sport and the level at which it is being performed – referees, judges, or umpires; coaches, managers, or instructors; club, league, and sport federation representatives, administrators, and event organizers; financial sponsors and sport business personnel; announcers, analysts, and sport media journalists; and, not least, spectators, fans, or viewers of televised sport events. The enactment and coordination of these roles is a central aspect of sport, but how this is accomplished may differ substantially (or not much at all) from one sport or level of competition to another.

The rules of a game or competition are similarly constitutive elements, although these may be stretched or overruled by unwritten conventions invoked by players, officials, or coaches. This happens both in informally organized games and at the highest levels of professional competition. Indeed, as Bourdieu has suggested, a sport, at a given moment,

> …is rather like a musical work: it is both the musical score (the rules of the game, etc.), and also the various competing interpretations (and a whole set of sedimented interpretations from the past); and each new interpreter is confronted by this…when he proposes 'his' interpretation.
> (1990: 163)

Despite differences in how a game or sport may be played by elite, recreational, or child athletes, even glaring disparities can be glossed over if establishing the likeness of these differing forms of a sport takes priority. A rudimentary athletic contest arranged on the spur of the moment can, therefore, be performed as if it was the imagined equivalent of a professional rendering of that sport. Fieldwork-based accounts are most illuminating when these explore the immediate circumstances as well as the broader sporting and social contexts within which particular games and sporting events take place.

Anthropologists have conducted detailed ethnographic studies of sport spectatorship in order to move beyond stereotyped depictions of fans' activities and behaviors. William W. Kelly (2004) reports in Japan not only that baseball fans deport themselves differently than fans of other sports, but also that varying styles of spectatorship can coexist within a single ballpark. In a similar vein, Swyers (2010) explains why merely 'going to a game' is not the same as 'being in the bleachers' at a Chicago Cub's game, where a self-described 'community' of regulars has fashioned perduring relationships and rituals that are demonstrably important to them as individuals. The differing styles and sentiments of self-selecting groups of football fans in England (Armstrong, 1998), Norway (Hognestad, 2003), and Mexico (Magazine, 2007) further establish that being a fan involves exercising choice and creativity in shaping this role to suit distinctive social and sporting tastes.

Sport, Play, and Celebration

No less important than the embodied and social nature of sport is the scope of its celebratory capacities. Because games and athletic competitions constitute forms of play, these stand temporally and experientially apart from mundane aspects of everyday life. Play, as Bateson (1973) argued, is essentially a communicative act. In his well-known example of monkeys engaging in play by pretending to fight, he contends that they share an implicit message: "this is not serious, this is play" (Bateson, 1973: 152ff). What permits such messages, says Csikszentmihayli, "is that play ultimately is a state of subjective experience" that can exist only "where there is an awareness of alternatives" (1981: 19). Distinguishing between playfulness and seriousness depends upon the contexts within which actions occur. The communication of which is which, which is preferred, and by whom, is of primary concern both for participants and analysts. What is marked and celebrated by athletes and spectators are forms of 'play with a purpose' that they choose to celebrate and contrast to the alternatives offered by everyday life.

Meanings generated by or attached to sport are refracted through the playfulness that underpins games and athletic practices. Thus, scoring more goals than one's opponents or defeating competitors in a race or test of physical prowess can be made to stand for larger matters than an ability to kick a ball into a net or to swim faster than others. What can also be brought into play is the celebration of more complicated types of identities and ideals. Sports furnish powerful means for incarnating otherwise abstract and elusive propositions. The stock activities of doing, watching, and debating athletic performances overlap and may be shared between on-field and off-field participants. Athletes are featured performers, but they also observe and discuss the forms of embodied practice within which they participate. Many fans or spectators also possess personal experience of having played the sports they watch as supporters. Among them are those who may have followed their favorite sport for more years than some of the journalists commenting upon a given match or competition. Fans are routinely enjoined to respect the spatial boundaries of events and not interfere with the flow of play, but they are also recognized as key participants in athletic performances. Without them, these would be merely instances of kicking a ball around, running, or skating. There is virtually no sport or form of athletic practice that does not in some manner anticipate its performances being watched and commented upon, whether by fellow athletes, paying fans or conscientious parents and family members. Even a solitary jogger or a group of children playing an impromptu ball game on a street can transfigure what they are experiencing into a performance of sorts by imagining an audience that watches their every move.

This mingling of play and imagination permits athletes, spectators, and event organizers to interweave their respective performances as improvised social drama. Together, they may revel in victories that take the form of prestigious championships, simple wins, or improved performances. But what may also be attended to, beyond the outcome of a given match or performance, is the assertion or transgression of social boundaries. Underlying the staging of any athletic event are potentially divisive questions about the identity and selection of participants: Who will take part? Who will not be allowed to participate? Who might I (or we) refuse to play with or against? Do we regard our opponents as worthy competitors or as persons unlike and beneath us? Matters of race, class, nationality, ethnicity, gender, sexuality, age, or disability can be readily accentuated or ignored within athletic activities, making these matters of intrinsic interest for anthropologists. Tuastad's (1997) essay on the political role of football for Palestinians in Jordan, Klein's (1991) study of the oscillating resistance to American cultural imperialism and dependence upon Major League Baseball (MLB) in the Dominican Republic, Bolin's (1997) account of competitive women bodybuilders' challenges to conventional notions of femininity, as well as Starn's (2011) examination of the golfer Tiger Wood's fall from grace in the world of professional golf and celebrity all show how sporting activities may be tied to extraathletic social distinctions and conflicts.

Ethnographic attention has also been directed to the cultivation and expression of sporting moralities that extend beyond the registering of wins or losses to address matters of aesthetics and cultural integrity. This, in turn, has stimulated analytically innovative conversations between ethnographers

of sport, dance, and other forms of embodied identities (Dyck and Archetti, 2003). For example, Archetti (1999a, 2003) has traced the plurality of male identities and moralities within the popular cultures of football and tango in Argentina, as well as the contrasting standards of heroic masculinity exhibited within cross-country and alpine skiing events in the Olympic Winter Games of Albertville (Archetti, 1999b). Hognestad's (2003) evocation of the endeavors of those Norwegians who opt to support English club teams rather than local football teams reveals this to be a considered stance that distances them symbolically from cloying aspects of Norwegian cultural nationalism. Armstrong's (1998) study of notions of honor and shame among English football supporters, Bromberger's (2007) account of the political restructuring of identities though football in contemporary Europe, and Magazine's (2007) delineation of the anticlientelistic standpoint of fans of the Puma football club in Mexico City illustrates how supporters may prefigure their sporting interests in terms of distinctive schemes of cultural and political preferences. Even within the confines of athletic competitions, elaborate codes and arguments may be applied to reckoning the relative morality of 'fair play,' deception, and breaking the rules, as Mewett (2002) and Nuhrat (2012) have demonstrated.

Sport can also serve as a referent for other social and political considerations. Krauss (2003) recounts how Germany's unexpected championship at the 1954 World Cup competition allowed a defeated people to experience and express pride in the exploits of the men representing them for the first time following the debacle of World War II. In the wake of the Cultural Revolution, it was the Chinese Women's Volleyball team's successive World Championship triumphs that prepared the way for the renewal of sporting and other relations with the world beyond (Brownell, 1995). In Cuba, success in baseball and Olympic sports has become a source of pride for the Cuban people as well as critical matters of state (Carter, 2008). Less exalted modes of participation by former child and youth soldiers in locally organized football leagues in war-torn Sierra Leone have, nonetheless, achieved invaluable social outcomes due to the irresistible attractions of play. The social rehabilitation of those who have been brutalized and who have in turn brutalized others is guided by recognition that the very possibility of games, leagues, and competitions depends upon their individual and collective ability to practice restraint (Richards, 1997).

Sport in Global Perspective

The codification and expansion of organized leisure sports in nineteenth-century Britain, northern Europe, and North America unfolded during a period of rapid industrialization, urbanization, and colonization. The subsequent diffusion of several of these sports to other parts of the world added transnational dimensions to what had initially been local and national forms of athletic practice. For instance, the creation in London in 1863 of the Football Association (FA) gave birth to the world's first national football organization. The launching of the FA Cup competition in 1872 and the Football League in 1888 expanded the structure of the sport in Britain and accommodated its professionalization. The creation of the Fédération Internationale de Football Association (FIFA) in 1904 laid the basis for future international competition, eventually leading to the staging of the first World Cup in 1930. By the late 1930s, more than 50 national football associations belonged to FIFA. Today, FIFA has more member states than the United Nations, making football a thoroughly global sport.

A second sport initiative destined to achieve global reach was inaugurated in 1894 with the creation of the International Olympic Committee (IOC), a nonprofit, nongovernmental organization based in Switzerland. The first Summer Olympics staged by the IOC was held in Athens in 1896, and the first Winter Olympics in Chamonix, France, in 1924. A third major athletic development, the spread of baseball across the United States during and after the Civil War, sparked widespread participation in the game and the founding of professional leagues. Ensuing struggles between team owners to consolidate and control a profitable baseball industry led to the establishing of MLB, a corporate entity that took for itself the power to determine what would be in the 'best interests' of the game. The first MLB championship series officially entitled the 'World Series' was played in 1903 by teams from Boston and Pittsburgh, notwithstanding the fact that baseball had been played in Cuba since 1874 and Japan since the 1880s.

Maneuvering within the specificities of sport histories and practices, anthropologists focus upon the diverse relationships, processes, and meanings that have shaped sport as "a connective tissue of modern life on imperial, international, and global scales" (William W. Kelly, 2007: 79). Bromberger identifies the fascination for football in north-western Mediterranean cities as resting upon its capacity to link the universal and the particular: "On the one hand it symbolizes the prominent values of the modern world: competition, achievement, solidarity, the part played by chance in individual or collective destinies…On the other hand, it offers an opportunity to assert one's collective identity" (Bromberger, 1994: 173). Archetti's (1997a,b, 1998a,b, 1999a) analyses of Argentinian football elucidate its importance in linking a Latin American country to Europe while permitting Argentinians to develop their own preferred style of playing the game. Armstrong and Mitchell's (2008) study of the interpenetration of local and global football in Malta, a small nation on the fringes of the European Union, notes that holding office in a local football club has traditionally served as a springboard into politics. While Malta is hardly a football power, the Maltese Football Association and local teams benefited materially after voting in support of Germany's application to host the 2006 World Cup (Armstrong and Mitchell, 2008: 140).

MacAloon's (1984) appraisal of the Olympics as an institution characterized it in terms of four central performative genres: spectacle, festival, ritual, and games. More recently, MacAloon has deftly dissected the discourse of 'legacy' by examining the ethnography of shifting patterns of Olympic speech among key administrative players in today's 'Olympic Family,' including IOC administrators, candidate city committees, and transnational consultants (MacAloon, 2008). Working outside the corridors of the IOC, a team of Norwegian anthropologists tracked a number of aspects of the planning and staging of the 1994 Winter Games in Lillehammer

(Klausen, 1999a). Pujik's (1999) narrative of the complicated dealings between the local organizing committee and IOC officials in scripting the opening ceremonies for the Lillehammer Games, Berkaak's (1999) description of the erasure of vernacular, local meanings in the completion of key competitive venues for the Games, and Klausen's (1999b) recounting of a fundamental dispute between the hosts and IOC over the organization of the Torch Relay that preceded and opened the Games all serve to illustrate the nature and extent of differences of purpose between these erstwhile partners. Brownell's (2008) study of why hosting the 2008 Games was so significant for China moves beyond prosaic calculations of the costs and benefits of taking on such a responsibility to locate this undertaking within the longer history of colonialism and sport in that country.

Once termed as 'America's Pastime,' baseball has been played intensively for more than a century in several countries and is currently being actively promoted by Major League Baseball International (MLBI) in other national settings for virtually the first time. Anthropological scholarship on transnationalism in baseball has examined the game as it is played in Japan (Kelly, W.W., 2004, 2006, 2009, 2011), Cuba (Carter, 2008; Eastman, 2008b), the Dominican Republic (Klein, 1991), Mexico (Klein, 1997), the United States (Gmelch, 2006), as well as Europe and South Africa (Klein, 2006). Following the IOC's 2005 decision to drop baseball from the Olympics, MLB moved to create its own international competition, the World Baseball Classic (WBC), which is modeled on the FIFA World Cup. The United States has not yet won the WBC on the field, but MLB firmly controls this competition. Klein (2006) and John D. Kelly (2006) have dissected the implications of corporate control through MLBI and the WBC of transnational baseball, linking these to broader patterns of American economic and political hegemony.

In addition to marketing sporting competitions, worldwide media coverage of these, as well as all manner of sport apparel and goods, the globalization of sport entails the migration of athletes – professionals as well as some who are nominally amateurs – across national boundaries. Indeed, 29% of all players listed on the 2005 opening day rosters of MLB teams were born outside the United States (Klein, 2006), while more than 40% of ballplayers with minor league contracts in the United States were foreign-born. An even greater reliance upon foreign-born players exists in the English Premier League, which employs players from across Europe, the United States, Africa, South America, and elsewhere. The transnational movement of athletically talented players to prestigious leagues and wealthy teams in a few focal countries has become a defining feature of a so-called emerging global sporting order. Within this model, embodied talent from more peripheral sporting locales migrates to powerful sporting centers while commoditized sport products and hegemonic sport cultures flow in the opposite direction.

Although attentive to the political economy of sport labor migration, anthropologists have tended to conduct their work in this field ethnographically. Armstrong and Mitchell (2008) show how African footballers may parlay playing for a Maltese team into obtaining Maltese (and, thus, also European) citizenship, thereby facilitating their access to more prestigious leagues in Europe. Besnier (2012) details the arrangements that lead Tongan rugby players and their families to Japan and New Zealand in numbers that may not seem large from the perspective of the receiving teams and nations, but that have deep impacts within Tonga. Dyck (2011) examines the paradoxical pursuit by young athletically talented middle-class men and women – usually with the full support of their parents and coaches – to win athletic scholarships that will take them to American colleges and universities, notwithstanding that most of these families could readily afford to enroll their sons and daughters in Canadian universities. Carter, who has conducted field research both in Cuba and in Northern Ireland, argues that mobility is not so much a personal quality possessed by individual athletes as a highly valued commodity that is produced in the social spaces between where these athletes are initially located and where they intend or wish to arrive (Carter, 2011: 25).

Anthropologists characteristically situate their studies of sport – including those touching on athletic labor migration – in particular settings that they seek to comprehend by taking account of individual and collective aspirations and experiences. By employing this approach, they retain a capacity to notice and look deeply into not only 'big' sports but also an array of decidedly 'nonhegemonic' but, nonetheless, locally preferred sporting practices that other disciplinary approaches are inclined to overlook, if not outright dismiss. Eriksen's (2007) assessment of the factors that recommend certain sports as candidates for 'globalization' also draws our attention to the other end of the spectrum: the considerations that prompt adherents of less recognized or celebrated sporting practices to steadfastly defend the small but viable niches within which these are played and embraced.

The Allure of Sport

The myriad attractions of sports and embodied athletic practices along with the crisscrossing uses to which these can be put make the study of sport a lucrative and intriguing undertaking for anthropologists. The roster of participants who engage with sport in one manner or another is extensive. Governments at all levels recognize the capacities of sport to connect leaders to citizens and, thereby, to promote projects that celebrate the nation (Appadurai, 1995; Brownell 2008), validate the state (Nkwi and Vidacs, 1997; Carter, 2008), or provide readily controllable venues for the expression of civic tensions (De Biasi and Lanfranchi, 1997; Ronsbo, 2003). Sport has also figured in attempts to sponsor the integration of immigrants into their new societies (Dyck, 2007), although it can also be marshaled as a means to marginalize immigrants (Grey, 1992). Sport organizers and institutions, ranging from bodies such as the IOC to MLB to local amateur sport leagues, may harbor wide-ranging if not always fully reconcilable objectives, ranging from public service to earning a livelihood.

Anthropological studies of child and youth sports in Denmark (Anderson, 2008), Canada (Dyck, 2012), Norway (Broch, 2003), and Sweden (Lithman, 2000) speak to the ways in which sport is employed by parents and communities as a means for attending not just to the health and developmental

needs of boys and girls but also to the stated and unstated concerns of the adults and agencies that support these activities. What these studies also reveal are the ways in which child and youth athletes endeavor to work with, against, or around initiatives that others have put into place to shape and manage their participation in sport.

The ways in which sport and athletic practices may be arranged to address, if not always fully satisfy, the varied desires of participants – whatever their age, gender, or level of physical abilities – are virtually infinite. The reasons that anthropologists might wish to look into this field of inquiry remain equally unbounded.

See also: Dance, Anthropology of; Play, Anthropology of; Sport, Sociology of.

Bibliography

Alter, J.S., 2006. Yoga at the Fin de Siecle: muscular Christianity with a 'Hindu' twist. In: MacAloon, J.J. (Ed.), Muscular Christianity in Colonial and Post-Colonial Worlds. Routledge, London, pp. 59–76.

Alter, J.S., 2004. Indian clubs and colonialism: Hindu masculinity and muscular christianity. Comparative Studies in Society and History 46, 497–534.

Alter, J.S., 1992. The Wrestler's Body: Identity and Ideology in North India. University of California Press, Berkeley, CA.

Anderson, S., 2008. Civil Sociality: Children, Sport, and Cultural Policy in Denmark. Information Age Publishing Inc., Charlotte, NC.

Appadurai, A., 1995. Playing with modernity: the decolonization of Indian cricket. In: Breckenridge, C.A. (Ed.), Consuming Modernity: Public Culture in a South Asian World. University of Minnesota Press, Minneapolis; and London, pp. 23–48.

Archetti, E.P., 2003. Playing football and dancing tango: embodying Argentina in movement, style and identity. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 217–229.

Archetti, E.P., 1999a. Masculinities: Football, Polo and the Tango in Argentina. Berg, Oxford; and New York.

Archetti, E.P., 1999b. The spectacle of heroic masculinity: Vegard Ulvang and Alberto Tomba in the Olympic Winter Games of Albertville. In: Klausen, A.M. (Ed.), Olympic Games as Performance and Event: The Case of the XVII Winter Olympic Games in Norway. Berghahn, New York, pp. 195–220.

Archetti, E.P., 1998a. The potero and the pibe: territory and belonging in the mythical account of Argentinean football. In: Lovell, N. (Ed.), Locality and Belonging. Routledge, London; and New York, pp. 189–210.

Archetti, E.P., 1998b. The meanings of sport in anthropology: a view from Latin America. European Review of Latin American and Caribbean Studies 65, 91–103.

Archetti, E.P., 1997a. The moralities of Argentinian football. In: Howell, S. (Ed.), The Ethnography of Moralities. Routledge, London; and New York, pp. 98–123.

Archetti, E.P., 1997b. 'And Give Joy to my Heart': ideology and emotions in the Argentinian cult of Maradona. In: Armstrong, G., Giulianotti, R. (Eds.), Entering the Field: New Perspectives on World Football. Berg, Oxford; and New York, pp. 31–51.

Arens, W., 1975. The Great American football ritual. Natural History 84, 72–81.

Armstrong, G., 1998. Football Hooligans: Knowing the Score. Berg, Oxford; and New York.

Armstrong, G., Mitchell, J.P., 2008. Global and Local Football: Politics and Europeanisation on the Fringes of the EU. Routledge, London; and New York.

Azoy, G.W., 1982. Buzkashi: Game and Power in Afghanistan. University of Pennsylvania Press, Philadelphia, PA.

Bateson, G., 1973. A theory of play and fantasy. In: Bateson, G. (Ed.), Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution and Epistemology. Paladin, St. Albans, pp. 150–167.

Berkaak, O.A., 1999. A place in the Sun: the vernacular landscape as Olympic venue. In: Klausen, A.M. (Ed.), Olympic Games as Performance and Event: The Case of the XVII Winter Olympic Games in Norway. Berghahn, New York, pp. 137–172.

Besnier, N., 2012. The athlete's body and the global condition: Tongan Rugby players in Japan. American Ethnologist 39 (3), 491–510.

Besnier, N., Brownell, S., 2012. Sport, modernity, and the body. Annual Review of Anthropology 41, 443–459.

Blanchard, K., 1995. The Anthropology of Sport: An Introduction, revised ed. Bergin and Garvey, Westport, CT; and London.

Bolin, A., 1997. Flex appeal, food, and fat: competitive bodybuilding, gender, and diet. In: Moore, P.L. (Ed.), Building Bodies. Rutger's University Press, New Brunswick, pp. 184–208.

Bolin, A., 1992. Vandalized vanity: feminine physiques betrayed and portrayed. In: Mascia-Lees, F.E., Sharpe, P. (Eds.), Tattoo, Torture, and Adornment. State University of New York Press, Albany, NY, pp. 79–99.

Bourdieu, P., 1990. Programme for a sociology of sport. In: Bourdieu, P. (Ed.), In Other Words: Essays towards a Reflexive Sociology. Stanford University Press, Stanford, pp. 156–167.

Bourdieu, P., 1984. Distinctions: A Social Critique of the Judgement of Taste. Cambridge University Press, Cambridge.

Broch, H.B., 2003. Embodied play and gender identities in competitive handball for children in Norway. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 75–93.

Bromberger, C., 2007. Through the looking glass of football. In: Demossier, M. (Ed.), The European Puzzle: The Political Structuring of Cultural Identities at a Time of Transition. Berghahn, New York, pp. 119–140.

Bromberger, C., 1994. Foreign footballers, cultural dreams and community identity in some north-western mediterranean cities. In: Bale, J., Maguire, J. (Eds.), The Global Sports Arena: Athletic Talent Migration in an Interdependent World. Frank Cass, London, pp. 171–182.

Brownell, S., 2008. Beijing's Games: What the Olympics Mean to China. Rowman & Littlefield, Lanham, MD.

Brownell, S., 1995. Training the Body for China: Sports in the Moral Order of the People's Republic. University of Chicago Press, Chicago.

Carter, T.F., 2011. In Foreign Fields: The Politics and Experiences of Transnational Sport Migration. Pluto, London.

Carter, T.F., 2008. The Quality of Home Runs: The Passion, Politics, and Language of Cuban Baseball. Duke University Press, Durham, NC; and London.

Csikszentmihalyi, M., 1981. Some paradoxes in the definition of play. In: Cheska, A.T. (Ed.), Play as Context: 1979 Proceedings of the Association for the Anthropological Study of Play. Leisure Press, West Point, NY, pp. 14–26.

Culin, S., 1907. Games of the North American Indians. In: Games of Skill, vol. 2. University of Nebraska Press, Lincoln, NB; and London. Reprinted 1992.

De Biasi, R., Lanfranchi, P., 1997. The importance of difference: football identities in Italy. In: Armstrong, G., Giulianotti, R. (Eds.), Entering the Field: New Perspectives on World Football. Berg, Oxford; and New York, pp. 87–104.

Downey, G., 2010. Throwing like a Brazilian: on ineptness and a skill-shaped body. In: Sands, R.R., Sands, L. (Eds.), The Anthropology of Sport and Human Movement: A Biocultural Perspective. Lexington Books, Lanham, MD, pp. 297–326.

Downey, G., 2007. Seeing with a 'Sideways Glance': visuomotor 'Knowing' and the plasticity of perception. In: Harris, M. (Ed.), Ways of Knowing: Anthropological Approaches to Crafting Experience and Knowledge. Berghahn, New York, pp. 222–241.

Downey, G., 2005a. Learning Capoeira: Lessons in Cunning from an Afro-Brazilian Art. Oxford University Press, Oxford.

Downey, G., 2005b. Educating the eyes: biocultural anthropology and physical education. Anthropology in Action 12 (2), 56–71.

Dundes, A., 1978. Into the endzone for a touchdown: a psychoanalytic consideration of American football. Western Folklore 37, 75–88.

Dyck, N., 2012. Fields of Play: An Ethnography of Children's Sports. University of Toronto Press, Toronto, ON.

Dyck, N., 2011. In pursuit of the 'Full Ride': American athletic scholarships and mobility, sport and childhood in Canada. Anthropologica 53 (1), 53–66.

Dyck, N., 2007. Playing like Canadians: improvising nation and identity through sport. In: Coleman, S., Kohn, T. (Eds.), The Discipline of Leisure: Embodying Cultures of 'Recreation'. Berghahn, Oxford; and New York, pp. 109–125.

Dyck, N., Archetti, E.P., 2003. Embodied identities: reshaping social life through sport and dance. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 1–19.

Eastman, B., 2008. Baseball in the breach: notes on defection, disaffection and transition in contemporary Cuba. In: Eastman, B., Ralph, M., Brown, S. (Eds.), America's Game(s): A Critical Anthropology of Sport. Routledge, London; and New York, pp. 126–157.

Elias, N., 1986. The genesis of sport as a sociological problem. In: Elias, N., Dunning, E. (Eds.), Quest for Excitement: Sport and Leisure in the Civilising Process. Basil Blackwell, Oxford, pp. 126–149.

Eriksen, T.H., 2007. Steps to an ecology of transnational sports. Global Networks 7 (2), 154–165.

Firth, R., 1930. A dart match in Tikopia: a study in the sociology of primitive sport. Oceania 1, 64–97.

Foster, R.J., 2006. From Trobriand cricket to rugby nation: the mission of sport in Papua New Guinea. International Journal of the History of Sport 23 (5), 739–758.
Foucault, M., 1986. The History of Sexuality, Volume III, the Care of the Self. Pantheon Books, New York.
Foucault, M., 1985. The History of Sexuality, Volume II, the Use of Pleasure. Pantheon Books, New York.
Foucault, M., 1978. The History of Sexuality, Volume I, an Introduction. Pantheon Books, New York.
Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Pantheon Books, New York.
Fox, R., 1961. Pueblo baseball: a new use for old witchcraft. Journal of American Folklore 74, 9–16.
Freedman, J., 1983. Will the Sheik use his blinding fireball? The ideology of professional wrestling. In: Manning, F.E. (Ed.), The Celebration of Society: Perspectives on Contemporary Cultural Performance. University of Western Ontario, Congress of Social and Humanistic Studies Bowling Green University Popular Press, Bowling Green, OH; and London, Canada, pp. 67–79.
Geertz, C., 1972. Deep play: notes on the Balinese cockfight. In: Geertz, C. (Ed.), The Interpretation of Cultures. Basic Books, New York, pp. 412–453.
Gmelch, G., 2006. Inside Pitch: Life in Professional Baseball, second ed. University of Nebraska Press, Lincoln, NE.
Gmelch, G., 1972. Magic in professional baseball. In: Stone, G.P. (Ed.), Games, Sports and Power. Dutton, New Brunswick, NJ, pp. 128–137.
Grey, M.A., 1992. Sports and immigrant, minority and anglo relations in Garden City (Kansas) High School. Sociology of Sport Journal 9, 255–270.
Guttmann, A., 1978. From Ritual to Record: The Nature of Modern Sports. Columbia University Press, New York.
Hognestad, H., 2003. Long-distance football support and liminal activities among norwegian fans. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 97–113.
Howe, P.D., 2011. Impairment: sporting bodies: sensuous, lived, and impaired. In: Mascia-Lees, F.E. (Ed.), A Companion to the Anthropology of the Body and Embodiment. Wiley-Blackwell, Oxford, pp. 276–291.
Howe, P.D., 2004. Sport, Professionalism and Pain: Ethnographies of Injury and Risk. Routledge, London; and New York.
Howe, P.D., Jones, C., 2006. Classification of disabled athletes: (dis)empowering the paralympic practice community. Sociology of Sport 23, 29–46.
Kelly, J.D., 2006. The American Game: Capitalism, Decolonization, World Domination, and Baseball. Prickly Paradigm, Chicago.
Kelly, W.W., 2011. Köshien Stadium: performing national virtues and regional rivalries in a 'theater of sport'. Sport in Society 14 (4), 481–493.
Kelly, W.W., 2009. Samurai baseball: the vicissitudes of a national sporting style. International Journal of the History of Sport 26 (3), 429–441.
Kelly, W.W., 2007. Is baseball a global sport? America's 'national pastime' as global field and international sport. In: Giulianotti, R., Robertson, R. (Eds.), Globalization and Sport. Blackwell, Oxford; and Malden, MA, pp. 79–93.
Kelly, W.W., 2006. Japan: the Hanshin Tigers and Japanese professional baseball. In: Gmelch, G. (Ed.), Baseball without Borders: The International Pastime. University of Nebraska Press, Lincoln, NB; and London, pp. 22–42.
Kelly, W.W., 2004. Sense and sensibility at the Ballpark: what fans make of professional baseball in Japan. In: Kelly, W.W. (Ed.), Fanning the Flames: Fans and Consumer Culture in Contemporary Japan. State University of New York Press, Albany, NY, pp. 79–105.
Klausen, A.M., 1999a. Introduction. In: Klausen, A.M. (Ed.), Olympic Games as Performance and Event: The Case of the XVII Winter Olympic Games in Norway. Berghahn Books, New York, pp. 1–8.
Klausen, A.M., 1999b. The torch relay: reinvention of tradition and conflict with the Greeks. In: Klausen, A.M. (Ed.), Olympic Games as Performance and Event: The Case of the XVII Winter Olympic Games in Norway. Berghahn Books, New York, pp. 75–96.
Klein, A.M., 2006. Growing the Game: The Globalization of Major League Baseball. Yale University Press, New Haven, CT.
Klein, A.M., 1997. Baseball on the Border: A Tale of Two Laredos. Princeton University Press, Princeton, NJ.
Klein, A.M., 1993. Little Big Men: Bodybuilding Subculture and Gender Construction. State University of New York Press, Albany.
Klein, A.M., 1991. Sugarball: The American Game, the Dominican Dream. Yale University Press, New Haven, CT; and London.
Kohn, T., 2008. Creatively sculpting the self through the discipline of martial arts. In: Dyck, N. (Ed.), Exploring Regimes of Discipline: The Dynamics of Restraint. Berghahn, Oxford; and New York, pp. 99–112.
Kohn, T., 2003. The aikido body: expressions of group identities and self-discovery in martial arts training. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 139–155.
Krauss, W., 2003. Football, nation and identity: German miracles in the postwar era. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 197–215.
Leach dir, J., 1976. Trobriand Cricket: An Ingenious Response to Colonialism, 54 min. Media Resources Center, University of California, Berkeley, CA.
LeClair, J., 2009. Sport and health: global challenges to biomedical definitions of disability. vis-à-vis: Explorations in Anthropology 9 (2), 203–219.
Lewis, J.L., 1992. Ring of Liberation: Deceptive Discourse in Brazilian Capoeira. University of Chicago Press, Chicago, IL.
Lithman, Y.G., 2000. Reflections on the social and cultural dimensions of children's elite sport in Sweden. In: Dyck, N. (Ed.), Games, Sports and Cultures. Berg, Oxford; and New York, pp. 163–181.
MacAloon, J.J., 2008. 'Legacy' as managerial/magical discourse in contemporary Olympic affairs. International Journal of the History of Sport 25 (14), 2060–2071.
MacAloon, J.J., 2006. Introduction: muscular christianity after 150 years. International Journal of the History of Sport 23 (5), 687–700.
MacAloon, J.J., 1987. An Observer's view of sport sociology. Sociology of Sport Journal 4, 103–115.
MacAloon, J.J., 1984. Olympic games and the theory of spectacle in modern societies. In: MacAloon, J.J. (Ed.), Rite, Drama, Festival, Spectacle: Rehearsals toward a Theory of Cultural Performance. Institute for the Study of Human Issues, Philadelphia, pp. 241–280.
MacAloon, J.J., 1981. This Great Symbol: Pierre de Coubertin and the Origins of the Modern Olympic Games. University of Chicago Press, Chicago, IL.
Magazine, R., 2007. Golden and Blue like My Heart: Masculinity, Youth and Power Among Soccer Fans in Mexico City. University of Arizona Press, Tucson, AZ.
Manning, F.E., 1981. Celebrating cricket: the symbolic construction of Caribbean politics. American Ethnologist 8 (3), 616–632.
Mauss, M., 1973. Techniques of the body. Economy and Society 2 (1), 70–88.
Mewett, P.G., 2002. Discourses of deception: cheating in professional running. Australian Journal of Anthropology 13 (3), 292–308.
Mooney, J., 1890. The Cherokee ball game. American Anthropologist 3, 105–132.
Nkwi, P.N., Vidacs, B., 1997. Football: politics and power in Cameroon. In: Armstrong, G., Giulianotti, R. (Eds.), Entering the Field: New Perspectives on World Football. Berg, Oxford; and New York, pp. 123–139.
Nuhrat, Y., 2012. Making fairness in Turkish football. Anthropology News, June.
Pujik, R., 1999. Producing Norwegian culture for domestic and foreign gazes: the Lillehammer Olympic opening ceremony. In: Klausen, A.M. (Ed.), Olympic Games as Performance and Event: The Case of the XVII Winter Olympic Games in Norway. Berghahn Books, New York, pp. 97–136.
Rapport, N., 2007. Bob, hospital bodybuilder: the integrity of the body, the transitiveness of 'work' and 'leisure'. In: Coleman, S., Kohn, T. (Eds.), The Discipline of Leisure: Embodying Cultures of 'Recreation'. Berghahn, New York; and Oxford, pp. 23–37.
Richards, P., 1997. Soccer and violence in war-torn Africa: soccer and social rehabilitation in Sierra Leone. In: Armstrong, G., Giulianotti, R. (Eds.), Entering the Field: New Perspectives on World Football. Berg, Oxford; and New York, pp. 141–157.
Ronsbo, H., 2003. The embodiment of male identities: alliances and cleavages in Salvadorean football. In: Dyck, N., Archetti, E.P. (Eds.), Sport, Dance and Embodied Identities. Berg, Oxford; and New York, pp. 157–175.
Samudra, J.K., 2008. Memory in our body: thick description and the translation of kinesthetic experience. American Ethnologist 35 (4), 665–681.
Sands, R.R., 2010. Anthropology revisits sports through human movement. In: Sands, R.R., Sands, L. (Eds.), The Anthropology of Sport and Human Movement: A Biocultural Perspective. Lexington Books, Lanham, MD, pp. 5–37.
Starn, O., 2011. The Passion of Tiger Woods: An Anthropologist Reports on Golf, Race, and Celebrity Scandal. Duke University Press, Durham, NC.
Sutton-Smith, B., 1995. Foreword. In: Blanchard, K. (Ed.), The Anthropology of Play: An Introduction, revised ed. Bergin and Garvey, Westport, CT; and London, pp. xi–xiii.
Swyers, H., 2010. Wrigley Regulars: Finding Community in the Bleachers. University of Illinois Press, Urbana, IL.
Tuastad, D., 1997. The political role of football for palestinians in Jordan. In: Armstrong, G., Giulianotti, R. (Eds.), Entering the Field: New Perspectives on World Football. Berg, Oxford; and New York, pp. 105–121.
Wacquant, L.J.D., 2004. Body & Soul: Notebooks of an Apprentice Boxer. Oxford University Press, Oxford; and New York.

Sport, Historiography of

Wray Vamplew, University of Stirling, Stirling, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Eisenberg, volume 22, pp. 14919–14924, © 2001, Elsevier Ltd.

Abstract

Sport history is the sports memory of a nation, but it is a contested terrain because of conflicts over the nature and validity of evidence and the application of theory. It has developed academic societies that publish journals, but there is also a strong tradition of amateur research. Suggestions are made for a research agenda and the direction in which the subject may progress.

Definition of Sports History

Sport history's major contribution to the study of sport is the time dimension. It can be considered the sport memory of a nation: without sport history, there is sporting amnesia. It can set straight the sporting record, but it can also explain why some things changed and why others continued unamended. History can provide the evidence to set events and incidents in their proper context and aid in explanation by giving an awareness of underlying forces. If we want to know where sport is heading, it is useful to know where it has been. It provides the benchmarks for measuring progress and change (or the lack of it). It can help us appreciate the difference between trend and fluctuation and realize that not everything seen as 'important' in sport need have a permanent influence or that everything in modern sport is new. Indeed, the sporting past has shaped the sporting present as all sports have some inheritance from the past, be it rules, governing bodies, styles of play, competitions, or equipment.

Yet it should be emphasized that historical knowledge is always provisional. Apart from 'sportifacts,' showing who won what, where, and by how many, there is no absolute truth in sport history. The sport historian attempts to make sense of the past by finding evidence, interpreting that evidence, and using it to come to a plausible conclusion. However, it should be recognized that sport history is a contested terrain and can be seen from different perspectives, involving diverse interrogations and interpretations of the source material.

Moreover, as Hill (2011: x) points out, 'history' cannot be separated from 'historians,' who have a personal relationship with the subject that can be influenced by upbringing, education, and political beliefs. In his seminal work, Booth (2005: 211) has called for greater reflexivity within the discipline: "an awareness that historians play creative roles in the production and presentation of history," and argues that historians should be more open with their value judgments and acknowledge how subjectivity affects their approach and narrative.

Development of the Subject

There are instances of early academic sport history pioneers who sought "to situate their work with larger societal dialogs and paradigms" (Pope, 1997: 1). Almost a century ago Paxson wrote on the rise of American sports, and the interwar decades saw German academics documenting sources on historical sporting incidents, persons, institutions, and events. However, it was not until the late 1960s that the subject really took off, in association with a general historical move into 'history from below,' which looked at the everyday lives of ordinary people. Almost everywhere two schools emerged, those trained via physical education and those with a background in history. The former focused more on the development of particular sports, whereas the latter was concerned with how sport fitted into the society in which it existed. It was these practitioners who developed sport history into a specialism to rank alongside other emerging 'new' history subdisciplines such as women's history, labor history, urban history, and black history. If chronology and diffusion were the watchwords of the human movement school, then themes and context were those of the historians; to the latter, sport history looked at more than the ludic and placed sporting issues in a wider social, political, economic, or cultural environment. There has been tension between the two schools of thought but generally they have coexisted, often appearing on the same platforms though professing different ideologies. These tensions are set to rise again as sport history consolidates itself in new geographical areas.

Avenues of Communication

Societies

The first major sport history society was the International Committee for the History of Physical Education and Sport (ICOSH) founded in 1967 under the umbrella of the International Council of Sport Science and Physical Education. This quickly became dominated by scholars from Eastern Europe, and ensuing political differences led to a breakaway second organization, the International Association for the History of Physical Education and Sport (HISPA) founded in 1973. Its aim was to establish links between sport historians and all major institutions in sport science – not, it should be noted, with mainstream history. Meanwhile, across the Atlantic, the North American Society of Sport History (NASSH) was established in 1972. In Britain, the British Society for Sports History (BSSH) emerged in 1981 out of a History of Physical Education Study Group, but at its inaugural conference in 1982 it formally adopted its current title. Down in the Antipodes, the Australian Society for Sports History (ASSH) was founded at the fourth biennial Sporting Traditions conferences. The numerical strength of these new organizations and their reluctance to

involve themselves in HISPA or ICOSH posed a threat to the European organizations, which, in a defensive move assisted by the thaw in the Cold War, merged in 1989 to form the International Society for the History of Physical Education and Sport (ISHPES), with less emphasis on the scientific side of sport. Nevertheless, although this society holds conferences around the globe, it has remained Eurocentric in its administration. Another European society was founded in 1995, emerging out of the European Network of Sport Science Institutes of Higher Education. First known as CESH, it formally became the European Committee for Sports History in 2006. Although open to all sport historians, it focuses its efforts on helping young scholars develop a European dimension in their work.

Journals

The first regularly published academic journal in the field was the *Canadian Journal of History of Sport*, which started in 1972 and became part of the Human Kinetics publishing empire in 1995 when it changed its name to *Sport History Review*. It was soon followed by the *Journal of Sport History*, which began as the official organ of NASSH in 1973. In Australia ASSH began to publish *Sporting Traditions* in 1984, which remains the Society's journal. Within Europe, significant sport history journals include *Stadion*, *Internationale Zeitschrift des Sports*, *Revue Internationale d'Histoire du Sport*, *Sportzeiten*, and *Ludica*, funded by the Benetton Foundation, which specializes in pre-twentieth-century material. More recent additions to the outlets for sport history include the CESH-sponsored *European Studies in Sport History* (2008), the Scandinavian-based *Idrottsforum* (2008), and the Brazilian *Recorde* (2008), the latter two both online journals. Two journals devoted to the Olympics, the *Journal of Olympic History* and *Olympika*, are essentially specialized sport history outlets.

By far the most truly global in its coverage is the *International Journal of the History of Sport*, which began life in 1984 as the commercially produced *British Journal of Sports History* but, due to internal political division and a failure to get the publisher to agree to limit subscriptions to members, was not associated with BSSH. In 1987, still under the foundation editorship of J. A. Mangan, it adopted its new title. It now publishes 18 issues a year, including regional ones for Asia, the Americas, Australasia and the Pacific, the Middle East, Africa, and Europe. It also sponsors workshops that lead to special issues of the journal. Initially published by Frank Cass, it is now under the aegis of Routledge. The BSSH's own journal was the *Sport Historian*, which emerged in 1994 out of a newsletter started in 1982. It was renamed *Sport in History* in 2003, around the time that it too began to be published by Routledge.

There are also several sociology of sport journals that open their pages to sport historians, among them the *Journal of Sport and Social Issues*, the *International Review of Sports Sociology*, the *Sociology of Sport Journal*, and a newcomer, *EAST*, which focuses on Asian sport.

Public History

Sport history does not appear only in academic books and articles. There is also what can be termed public sport history, which is generally popularizing, often propagandized, and sometimes entertaining. For a discussion of the topic by leading sports historians who have been involved in public history, see Booth et al. (2008). It has three major forms. First there is what can be termed populist public history, which uses media outlets, particularly television, film, and the Internet. Films and television programs are often criticized by academic historians as the historian's voice becomes mediated by producers and directors who sacrifice historical accuracy for a story line, but this is what attracts the viewers; and far more people watch sport history films or programs than ever read history books. The Internet is often the resort of amateur historians, amassing facts about their favorite team or player and offering constant corrections to Wikipedia entries. More recently, however, a few academic sport historians have opted to put their work online with open access. Whether it gets read by the public is conjectural. Second, we have sport museums, the public face of sport history. These can be the best places to replicate the performance, drama, romance, passion, and emotion of sport, and they have done much to educate through entertainment. Unfortunately, too often they have catered to the nostalgia market and, in doing so, perpetuated myths; lacked historical objectivity and subtlety of argument; failed to contextualize artifacts; eschewed the controversial; concentrated on sport that was competitive, adult, and a male-dominated activity; and had an obsession with winners and winning. A third form is the 'official' history, authorized or commissioned by a governing body or the like. Here the criticism is of both omission and commission: The funders are told what they want to hear, and a spin is often put on controversial issues.

Evidence and Theory

History is an empirically based, interpretive social science: Unless there is some evidence from the past, there can be no sport history. However, the traditional sources of evidence from archives and newspapers – prior to television perhaps "the great instrument of popular communication" (Hill, 2006: 121) – have been supplemented by new ones. Oral history can provide a personal perception of events and what they meant to particular people, giving life to dry historical evidence. Visual material enables us to appreciate what the past looked like. Both film and photography confirm the very existence of the past, with film having the added dimension of movement, the body in action being a central feature of sport. Fiction could be a valuable source as it is a cultural force that shapes how people understand the world around them, but sports historians have been reluctant to use it, viewing it as unreliable and subjective. Nevertheless, novels, particularly those written within the period being studied, can cast light on the context within which sport took place and give insights into matters on which conventional sources are opaque, in particular the role of sport in everyday life.

History is dependent on evidence, though it is important for sports historians to interrogate their sources so as to assess their authenticity and validity. Hill (2006), for one, has stressed that the press should be seen as a text to be interpreted rather than as a factual source to be accepted. Indeed, some aspects of reportage are on a par with inventing tradition: Adding

anecdotes, selecting facts, and forwarding opinion can help sell newspapers but tarnish them as reliable, straightforward source material. Few sport historians have fully embraced the postmodernist view of history being virtually creative writing, but there is a greater willingness to accept that evidence can be prejudiced and can privilege some information above others. What evidence is collected and what is saved can be functions of power in past and present society. Hence subordinate groups – usually people who do not keep diaries, were not on committees, and were not interviewed by reporters – do not always get their voices 'heard' in historical documents. Moreover, 'evidence' can be falsified or manipulated either deliberately with the doctoring or cropping of photographs or more subconsciously when oral history interviewees confuse memory with hindsight.

Booth (2005) takes sport historians to task for a failure to engage more extensively with theory and criticizes those who simply gather facts to tell a story. Yet more sport historians have used theory and theoretical concepts than he gives the discipline credit for. Some sport historians, particularly those from a social history background, have applied theory explicitly in their work, notably that from the European philosophers Antonio Gramsci, Michel Foucault, and Pierre Bourdieu, joining those who earlier had implicitly used the ideas of Karl Marx. Economic historians of sport had always thought in an economically theoretical framework. Many sport historians have indeed embraced theoretical frameworks such as modernization, materialism, hegemony, structuration, feminism, discourse, and textualism and thus adopted the precept of Howell (2007: 461) that "to write good history … requires us to engage critically with the theoretical discourses that shape our historical understanding and to contemplate intellectual trends that can cross disciplinary boundaries." Nevertheless, it is true that few sports historians explicitly discuss theoretical issues; however, many of them are aware of theoretical concepts and theories, which they use to inform their work and help them make their explanations and arguments. Class, status, and power have featured in sport history for a long time, and as new conceptual issues emerged on the sociopolitical agenda, they have been taken up by sport historians. For example, the linguistic turn has entered sport history and the application of cultural studies theory to historical sources is providing an avenue to explore what sport meant to players, officials, and fans. This has had two important influences on sport history. First, it helps us understand more what people believe happened in their sporting experiences, even if it didn't; and, second, it demonstrates how the historical text can influence what people believe happened in sport.

It ought to be noted that two overarching theories have been developed respectively by Guttmann (1978, 2004) and Szymanski (2008) to explain the development of sport. Guttmann postulated sports history's own version of 'modernisation' in which he argued that modern sport possesses seven structural characteristics. First, it was secular with no religious reasons for participation. Second, it should demonstrate equality: Theoretically, everyone should have an opportunity to compete, and conditions of competition should be the same for all contestants. Third, it introduced the idea of specialization: Everyone who wanted to could join in folk football, a sport in which there were no sharply defined roles, but the emphasis on achievement in modern sport brought in specialization both within a sport and between sports. Fourth came rationalization, in particular the development of rules, which in primitive societies were often considered 'divine instructions' (God-given rituals not to be tampered with by mere humans); in contrast nonsecular modern sports have been invented and have written rules. Even more rationalization came via the development of coaching and sports science. His fifth feature was bureaucratization. Almost every major modern sport has national and international organizations that have developed extensive bureaucracies to establish universal rules for their sport and oversee their implementation. These were not required when there were no written rules. Sixth was quantification by which modern sports transform every athletic feat into statistics. Following on from quantification is his seventh point, the modern emphasis on records. Like many models, Guttmann's was an ideal-type postulation that may never have all its conditions fully satisfied. However, it has stood the test of time, if not in its entirety then as a basis on which others have built. Some critics have suggested that the model requires more input on press publicity, marketing, commercialization, and professionalization. Others believe that Guttmann has stuck too rigidly to his original ideal-type model and has refused to revise it in the light of new research findings, though they do acknowledge he has integrated some of the critical perspectives into his own work.

In the early eighteenth century, a movement began in Britain that involved the formation of clubs for many purposes, not least sports such as cricket, golf, pugilism, and horseracing. They enabled people with a common purpose to come together, provided a basis for agreeing common rules and regulations, created a framework for competitive interaction, and secured a location for participation and sociability. Szymanski has argued that modern British sport emerged from these new forms of associativity, which developed autonomously in Britain following the retreat of the state from the control of associative activities. This was in contrast, he contends, to the situation in countries such as France and Germany, where club formation continued to require the explicit or implicit approval of the state. Here, modern sports developed in ways consistent with, or even in the service of, the objectives of the state, most notably the need to maintain military preparedness. His critics acknowledge the ambition of the analysis but suggest missing elements and alternative causal factors. They argue that more evidence is required to support the hypothesis, that he should have looked further back in time for his European material, that he failed adequately to address the issue of class, and that he understated the role of commercialization. More recently, Vamplew (2012) has argued for a study of sports clubs using theories of social, cultural, physical, human, and financial capital.

Finally, it should be emphasized that, until substantiated by evidence, theories are just competing hypotheses. They might aid our understanding, but they do not explain a situation completely. Empirical support is a necessary concomitant for accepting any hypothesis. Moreover, no theory is immutable. If the facts do not fit the theory, then the historian should check the facts again and, if still convinced that the facts are correct, then the theory should be modified. Historians must be prepared to not only use theory but also adapt it.

Research Agenda

The reader keen to know the state of play in the subject is directed to several recent publications. In a series of 'Presidential Reflections' published in the *Journal of Sport History* between 2007 and 2009, past presidents of ISHPES, BSSH, ASSH, and NASSH presented their views on the achievements (and failures) of researchers in their particular domains. Most of the established sport history journals also have issues or detailed articles dealing with the historiography of particular themes. For summaries of the existing knowledge in a variety of countries, readers should consult Pope and Naurright (2010), who provide a comprehensive guide to the international field of sport history as it has developed as an academic area of study. A later publication by Edelman and Wilson (2013) analyzes the state of current scholarship and summarizes ongoing debates but with a wider thematic approach.

Still, much remains to be discovered. Geographically, we know little about sport in Africa or South America. More work is required on the historical diffusion and cultural transmission of sport between nations. Imperialism has long been a major concern of sport historians, mainly focusing on how Britain exported its sports and games to the colonies and how the recipient countries adopted and adapted them. For those who have the languages, similar knowledge is available on French and German sporting imperialism. Outside formal empire, a project based at the University of Cambridge is examining the spread of sport and sports within Europe, with a focus on the various responses to the appearance of foreign sports, including aversion, resistance, adoption, adaptation, and reinterpretation (Tomlinson and Young, 2011). There is also emerging work on America's sporting empire, not merely the sociological emphasis on globalization. However, we need more intracontinental comparisons. The *Journal of Sport History*, the official outlet for NASSH, has yet to publish an article on Mexico, all comparative work being restricted to the United States and Canada. There is also the problem of sport history tending to be conducted within the parameters of the nation state with a reluctance (or an inability) to cross national boundaries. Language can be a barrier but, collaborative work can circumvent this as in the previously mentioned project trying to develop ways of researching a sport history of Europe. Themes such as associativity, gender relations, children in sport, and racial and ethnic participation could all lend themselves to cross-cultural comparisons.

On the economic side, our knowledge is deficient on the production of, and trade in, sporting goods. We know something about the racehorse, whose genealogy dates back to the 1600s, but where did the early cricket bats come from? The demand for a particular sporting good can be volatile, and there is a role here to look at fads and fashion in sporting equipment and apparel. We also know very little about sport sponsorship and advertising by which business firms tried to make potential customers aware of their products. Historically, it existed with sportsmen endorsing cigarettes and various liniments and in brewers painting their names across grandstand roofs, but there have been no studies on either its extent or effectiveness. There is also a great gap in our historical knowledge of sport promoters and entrepreneurs.

Sport history needs its own version of 'history from below' and should look more at the vast range of informal and unorganized sport practices. Much more is needed on noncompetitive sport and on nonelite sport. However, this involves detailed research work with scant academic reward, at least in the short term. To count the number of tennis clubs, defunct and existent, in Nottinghamshire as one amateur enthusiast is undertaking is a painstaking task and still not that useful without similar work for comparative purposes on another Nottinghamshire sport or another county's tennis courts. Although there is a plethora of individual club histories, the concept of the sports club itself, the voluntary organization within which most sport takes place, is worthy of further exploration. This ubiquitous institution took many forms including private sports clubs; work-based ones; those associated with the alcohol trade, religious faiths, politics, education, and the military; and those stemming from local authority intervention into the recreational sector.

This move away from elite sport could also include children's sport, which is relatively untouched. More needs to be done on the role of the family but even that other major socializing agency, the school, has not been adequately explored. Within Britain, most work has looked at the Victorian public schools, though it is still not clear whether the practice matched the rhetoric and how effective the socialization process was. State schools are practically virgin territory, though it is from the pupils of such educational institutions that most professional sportspersons emerged. Some children, such as the boy jockey and child caddy, were actually employed in professional sport. Here there is an opportunity to explore the concepts of child labor (exploitative) versus child work (positive) and the dichotomy between work activities that are socializing and those that are instrumental.

Two sets of concepts that have been researched extensively but still offer opportunities for further work are race and ethnicity and gender. Race and ethnicity have featured, particularly, in studies of the United States, Australia, and South Africa, which have outlined the historical stages of the recognition in white sporting circles of the black athlete. These have involved novelty value, exclusion, segregation, discrimination, and eventually acceptance, though often still carrying stereotyping based on perceptions of black physical ability. The place of women in sport has been extensively covered. For most of the nineteenth century, and well beyond in many countries, half the population faced restrictions on playing sport simply because they were female. Depending on the sport and the nation involved, the mechanisms used included social disapproval from men and women, expert scientific and medical opinion, and the ancillary and regulatory rules imposed by clubs, organizations, and competitions. However, little work has been done on other factors such as the time demands of employment and child-rearing, and preference for other leisure activities. Nor has much been done on women as spectators despite photographic evidence of their involvement. What is now occurring is a recognition that sport and gender is no longer synonymous with sport and women. Masculinities have been added to the gender agenda. Modern sport was established as a male province from its conception and has done much to construct masculinity in many societies. As a site of much male voluntary activity, sport has been a major method of gender fixing.

One failure by historians of sport is to fully grasp the emotion and physicality of sport. It has not been shown what it

was like to be a passionate football supporter in the 1880s or a 'manly' English batsman facing up to the hostile Australian fast bowling in the early cricket test matches. It would be poor history simply to transpose the modern experience of being a fan or a player into the past. Sport historians must search for new source material in match programs, the local press, diaries, and even novels of the time. Somehow we need to hear the voices of the period.

Another issue that ought to be considered is that, to be really representative of sporting participation, sport history ought to take more cognizance of losers, for losing is the most typical of sporting experiences: There can be only one champion, only one cup winner. There is a place for the history of the also-rans: The marathon runners who hit the wall and the Tour de France cyclists who fail to end a stage. Little, too, has been written about those who disliked sport and the reasons for their aversion.

It should be emphasized that there will never be the final word on any topic researched by sport historians, only the best prevailing opinion at a point in time. As Polley (2007: 4) puts it, the past is continually being rewritten "in the light of … current needs and interests": As the cultural climate changes, new ways of looking at historical issues are formulated. Indeed, Howell (2007: 460) has summarized how sport history has changed during his academic career:

> … the now hackneyed iterations of history from below; the flourishing and withering of social control analysis; the flexible cultural Marxism of the sixties and seventies espoused by Edward Thompson, Raymond Williams, and others; the discipline's fascination with Gramscian notions of hegemony; issues of agency; the emergence of subaltern studies; Foucaultian discourses about social technologies, technique and discipline; the excavation of representations and the imagined worlds of past and present; and the cultural and linguistic turn, … All of these have caused changes in the way in which sports historians have gone about their business.

The Future of Sport History

The sport history world is one of plurality with distinct variations in its origins, development, and current state across the world. This was very clear from the discussion at a forum during the 25th anniversary conference of the British Society of Sports History in 2007, at which the presidents of four major sport history organizations presented their views on the state of the discipline in their own bailiwicks. A paper by Hoffman and Kay at the NASSH Congress of 2011 that analyzed the content of sport history journals and conference programs in Germany, Britain, and the United States demonstrated the diversity of approach currently on display within these nations. They cautioned that as scholarly work and knowledge exchange become increasingly international, and sport historians regularly present their findings beyond their own national societies, we should be aware that we are not all singing the same songs at the same time to the same rhythm.

As Pope and Nauright (2010: 3) stated, "sports history is no longer an academic curiosity, but this does not mean that it has joined the mainstream." A major problem facing sport historians is that, although some mainstream historians might acknowledge that "sport is immensely important to any serious attempt to reconstruct a nation's collective life," few read their work (Johnes, 2008: 70). Whereas sport historians think they have much to offer to debates in globalization, race and ethnicity, and particularly identity since "sport helps shape our sense of who we are" (Hill, 2011: 73), mainstream historians, when they bother at all, see sport history as a source of interesting examples but contributing little to the central issues. One reason that they may be unaware of good work is that, till recently, no sport history journal appeared in the Arts and Humanities or Social Science citation indexes. This is changing. The *International Journal of the History of Sport* has just been accepted by Thomson Reuters, and the *Journal of Sport History* by Scopus. Of course, Google may be the salvation of sport history as more articles are scanned into its system. That said, some mainstream journals have published sport history articles. Among other moves toward closer relations is that, in Australia, ASSH has managed to secure a stream within the Australian Historical Association national conference, and the explicit intention of the forthcoming handbook edited by Edelman and Wilson (2013) is to engage effectively with an audience of mainstream historians to whom the work is primarily directed. Hopefully they can be persuaded that sport history offers a "dynamic accessible window into politics, economics, gender, race, class formation, ideology, religion and virtually any discipline" (Pope, 1997: 3).

Some sport historians believe that the way to be accepted by the mainstream is to move down a strongly theoretical avenue. Certainly, this could help establish a beachhead in progressive mainstream history, especially if theory was modified and developed, not just applied. However, the downside is the possibility that it would drastically curtail nonacademic readership who have no time for the theory of Foucault and Bourdieu or theoretical concepts such as hegemony and discourse. They would not necessarily accept that historical perspective is contested terrain with a plurality of meanings and that there are different versions of events depending from whose perspective the narrative is being constructed. Nor might they appreciate Booth's argument that all "facts are propositional statements about the nature of reality" and that all sources distort or filter the truth and all of them need interpretation (Booth, 2005: 30). Such nonacademic amateur sport historians should not be dismissed or abandoned as mere antiquarians. These fans with keyboards may not contextualize well, but they usually get their 'sportifacts' right and provide solid bricks with which academic historians can build their arguments. An exemplar here is Vamplew's (2012) study of the development of the British golf club using over 300 histories of individual clubs written mainly by amateur enthusiasts.

Another way forward would lead sport historians to encompass the broader field of leisure history, which might be progressed by connecting up with histories of taste, leisure, and consumption. Johnes (2008: 69) argues that sport cannot be regarded as special but is just another form of leisure. Sports are physical, but so too are dancing and gardening; sports could be competitive, but so could be baking and music; sport could be social or solitary just like many leisure activities. He adds that "there is surely no emotion or attraction involved in sport that

could not also be found in other pastimes." Moreover, the organization of leisure and sport were both based on clubs with rules and regulations. Some academics would even suggest that some sports might have more in common with other leisure activities than with other sports. Collins has asked what commonality is there between polo at Hurlingham and wrestling in Hull and suggested that it would "make more sense to study them in their own particular social contexts, as activities and manifestations of meaning for particular social groups, rather than as part of the continuum of sport" (Collins, 2007: 399).

Whatever path is taken, none is likely to dominate the subdiscipline. Currently, there are dichotomies between those who opt for quantification and those who prefer a qualitative approach; between those who seek information at the aggregate level (often the quantifiers) and those who look at the individual (mainly the nonstatistical historians); between those who apply theory and theoretical concepts, and those who are more empirically focused; and between those who pose modern questions in a historical setting and those who try to understand what mattered to those in the past. Different schools of sports history can coexist: They did so in the past, and they do so today, so why not in the future?

See also: Civil Society, Concept and History of; Mass Media and Sports; Sport, Sociology of.

Bibliography

Booth, Douglas, 2005. The Field: Truth and Fiction in Sport History. Routledge, Abingdon.

Booth, Douglas, Cronin, Mike, Phillips, Murray G., Brown, Douglas, 2008. "Our" voices in public history. Journal of Sport History 35, 371–427.

Collins, Tony, 2007. Review article: work, rest and play: recent trends in the history of sport and leisure. Journal of Contemporary History 42, 397–410.

Edelman, Robert, Wilson, Wayne (Eds.), 2013. Oxford Handbook of Sport History. Oxford University Press, Oxford.

Guttmann, Allen, 1978, 2004. From Ritual to Record. Columbia University Press, New York.

Hill, Jeffrey, 2006. Anecdotal evidence: sport, the newspaper press, and history. In: Phillips, M. (Ed.), Deconstructing Sport History. State University of New York Press, Albany, pp. 117–130.

Hill, Jeffrey, 2011. Sport in History: An Introduction. Palgrave Macmillan, Basingstoke.

Howell, Colin, 2007. Assessing sport history and the cultural and linguistic turn. Journal of Sport History 34, 459–465.

Johnes, Martin, 2008. British sports history: the present and the future. Journal of Sport History 35, 65–71.

Polley, Martin, 2007. Sports History: A Practical Guide. Palgrave Macmillan, Basingstoke.

Pope, Steven W. (Ed.), 1997. The New American Sport History. University of Illinois Press, Urbana.

Pope, Steven W., Nauright, John (Eds.), 2010. Routledge Companion to Sports History. Routledge, Abingdon.

Szymanski, Stefan, 2008. A theory of the evolution of modern sport. Journal of Sport History 35, 1–32.

Tomlinson, Alan, Young, Christopher, 2011. Towards a new history of European sport. European Review 19, 487–507.

Vamplew, Wray, Summer, 2012. Concepts of capital: an approach shot to the history of the British golf club before 1914. Journal of Sport History 39, 299–331.

Sport, Sociology of

Peter G Craig, University of Bedfordshire, Bedford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by T.W. Crosset, volume 22, pp. 14932–14935, © 2001, Elsevier Ltd.

Abstract

Since the 1960s sport sociology has been a growing and important subfield within sociology. Sport has a significant presence in the modern world that intersects with political, economic, social, cultural, international, national, and local structures and interests. As this brief overview details through its examination of these dynamic and complex relationships, sport sociologists have not established a dominant theoretical paradigm. Rather, they constructed a critical dialogue that has progressively drawn on a range of theories and concepts including: functionalism; Marxism and neo-Marxism; figurational sociology; cultural studies; feminism; postmodern and poststructural perspectives. The field continues to develop as it seeks to explain sport's connections to contemporary issues such as globalization, technological change especially in regard to media technologies, climate change, and sustainable lifestyles.

Introduction

Though the origins of sociological theories and concepts have a clear trajectory starting in the nineteenth century with the writings of Comte, Marx, Durkheim, and Weber, the intellectual history of sociology's interest in sport has more recent origins. There is some debate as to how these seminal theorists viewed the relevance of sporting practices, however, it is reasonable to contend that sport and leisure practices had, by the late nineteenth century, become arenas of sociological interest in that they were established and significant parts of the private and public life. This awareness did not, however, produce any significant levels of analysis and it was not until the 1960s, that a substantive body of sociological theory and research into sport began to be produced. Today sport sociology is a highly prolific field of research and theory constituting a distinctive subdiscipline within sociology. The field is supported by a number of well-established academic bodies (British Sociology of Sport Association, North American Sociology of Sport Association) and peer-reviewed journals (e.g., *The International Review for the Sociology of Sport; Sport and Social Issues; Sport, Education and Society*; etc.) that are also associated with a wide range of annual and quad annual academic international conferences. Across a global span of universities, sport sociology is an essential component of sport science departments and their delivery of undergraduate and postgraduate degrees in sport.

Since the 1960s (Kenyon and Loy, 1965), sociological interest in sport has become extremely wide-ranging and has been informed by the transformative economic, social, cultural, and technological issues impacting the social world. Some of the most influential of these focus on: the impact of globalization and the cultural pervasiveness of competitive sport within everyday social life; the economic power of a globalized (but mostly North American and European) network of intersecting 'sport industries' and transnational corporations; sport's interconnectedness to a growing consumer culture; the intersections between privatized sporting practices and contested power relations across and within a diversity of social groups (class, gender, sexuality, race/ethnicity, disability, age); sport and national identity.

Given this diversity of interests, it is hardly surprising that the field has sometimes been characterized by antagonistic and sometimes acrimonious exchanges between scholars (Giulianotti, 2004). Putting these 'theory wars' to one side, there are a number of concerns that broadly unite the field. One concern has been to analyze and comprehend the social linkages between sports institutional structures and social practices and modernity's evolving institutional and cultural character. Another has been its interest in agency/power dynamics in that as Sugden and Tomlinson (2002: p. 9) note "power relations are social relations."

Sport and Modernity

Social historians, such as Huizinga (1955), evidence the foundational connection between humanity's universal ludic element and the historical development of human civilizations. For Huizinga (1955), the social and cultural characteristics of modernity progressively challenged the playful character of sport as irrational and unproductive and inserted a new rationalism – the dominance of economic and labor relations over all other forms of life. Modernity's structural and ideological frameworks therefore had a powerful influence on modern sport's codification and governance. In the nineteenth century this was particularly exemplified by Victorian Britain that was becoming ever more industrialized, capitalist, urbanized, bureaucratically organized, and internationally interconnected. The confluence of the global influence of the British Empire and its internal processes of modernization created many of the sports that today have a dominant global presence. These sports include, boxing, varieties of football sports (soccer, codes of rugby-football, American football), bat and ball games (cricket, baseball, and tennis), and a family of athletic activities located under the title of track and field.

Though powerful social stratifications based on class, gender, and race were always evident, the growing cultural

significance of these sporting forms (allied to their use in educational settings) created a largely unquestioned process that established belief in the moral virtues and social benefits attainable from such sports (Craig and Mellor, 2010). Located in this realization, sociologists and historians began to analyze how the rationalized elements of modern sport were/are dialectically connected to modernity and how this distinguished it from manifestations of recreational play and games within previous historical epochs. Illuminated by these analyses, it was realized that sport was a highly significant and complex social and cultural construction existing in spheres of being that can be both: physical and abstract, competitive and playful, institutionalized and autonomous, commodified and part of the human commons. Founded on its codified, rationalized, and institutionalized constructions, modern sport's tangible social reality became ever more open to empirical academic analysis. In locating aspects of these analyses within perspectives drawn from Marx, Weber, and latterly Giddens and Bourdieu, sport sociologists have also determined that these social constructions do not and cannot exist other than through their active and reflexive production and reproduction through people's everyday actions. The powerful social bonds connecting fans, athletes, teams, and sporting localities routinely demonstrate that sport is rooted in the norms and obligations of social reciprocity and citizenship. The significance of sport therefore cannot be apprehended in the absence of a much broader understanding of the structure of the society as a whole (Bourdieu, 1986, 1988).

Sport – A Sociological Perspective

For the first half of the twentieth century, sport's nineteenth century genesis meant that in most Western countries sports were run as independent bureaucratic social organizations that were predominantly viewed as having little or no direct connection to the broader processes of democratic organization, governance, and economic prosperity. Regarded mostly as a matter of personal choice and self-interest, sport received minimal interest within the academic sociology of the time. This changed through the latter part of the twentieth century when social, economic, and political transformations and crises drew sport into a sharper sociological focus. These included the postwar rise of the welfare state, civil rights and feminist movements, postcolonial nationalist movements, East–West Cold War ideological and political confrontations. Given its ubiquity, sport could not remain immune from the social movements of the 1960s and 1970s. In the United States, manifestos such as the *Athletic Revolution* (Scott, 1971) and Edwards (1969), a critique of the racialized reality of many sports started to challenge the established consensus. In Europe and Canada, Jean-Marie Brohm (1978) and Rigauer (1981) lead a strong Marxist critique that identified sport as alienated and a bastion of capitalist class inequalities. In the UK, sport became located into social democratic ideas about the welfare state and was increasingly represented as a social right under the 'Sport for All' banner.

Drawing on these critical analyses the field has continued to produce an extensive and important body of social analysis that has its own distinctive genealogy. This genealogy is informed by a number of distinct approaches that parallel sociology's complex and dynamic body of theoretical perspectives and epistemological approaches. Though the major perspectives which have influenced the field remain open to debate (Giulianotti, 2005; Jarvie, 2006), there is a fairly well-accepted consensus that emphasizes the significance of the following perspectives: functionalism; Marxism and neo-Marxism and cultural studies; interpretative approaches; figurational sociology; feminism, postmodernism, and globalization theories. Though the field is often reduced to these perspectives, it is important to state two caveats. While many of sport's most influential theorists are associated with one perspective more than others, none has attained dominance within the field. This is even more apparent when distinctions are made between the epistemological traditions within different countries. Across university departments and academic journals there are distinct differences between the United States, Canada, the UK, and Europe. Moreover within much of the recent literature covering gender, race, sexuality, and sporting bodies, the boundaries between feminist, cultural studies, postmodern and poststructuralist analyses are often arbitrary and fluid, and many would argue that these separations are mostly habitualized academic conventions rather than an artifact of necessary epistemological or ontological coherence.

Given that numerous texts provide those interested in sport sociology with comprehensive discussions and critiques of these perspectives, it is not the intent of this article to provide yet another contribution to this already well-populated area. Instead, the intention is to highlight some of their most salient points within this work. The discussion is organized in three sections. The first reflects on sport's institutional and functional characteristics. The second examines the impact of the critical cultural turn within sociology and specifically how sociology has regarded sport as a cultural product and a set of social practices reflective of dominant and subordinate power relations. The third briefly details some of the emergent themes that are starting to have a presence within sport sociology.

The Institutional Analysis of Sport: *The Functionalist Turn*

Theorists such as Durkheim, Parsons, and Merton influenced much of the initial sociological analysis of sport. In adopting this structural functionalist perspective, this work broadly accepted the premise that the social world could be understood as a system of integrated and interconnected working parts. This work was primarily empirical and focused on establishing 'social facts' and was primarily adopted by American sociologists (Loy et al., 1978). Their goal was to establish if there was empirical proof for popular assumptions regarding sport's capacity to 'function' positively as a social institution and contribute to the Parsonian mandate to achieve necessary functional prerequisites. More specifically their interest was on how sport could integrate members of the society, provide an arena of positive goal setting and attainment of positive and character-building

attributes, make a contribution to the health and well-being of the nation, create community, and foster national identity.

The results were by and large inconclusive (Fine, 1987). By the late 1970s, concerns regarding processes of socialization started to take a more critical turn as the linkages between participation in sport, class and gender inequality, racism, and other socially problematic behaviors (cheating, violence, sexual abuse, etc.) became matters of concern. Contemporary empirical research continues to provide insights into a number of issues: the changing institutional structure of globalized sport (Maguire, 1999); the increasing rapidity and impact of technological change on the world of elite sport performance; the use and regulation of performance enhancing drug use and nontherapeutic surgical body enhancements; and the contradictory representations of the 'natural embodied athlete' (Cole, 2002). The ongoing penetration of computer-mediated technologies to all aspects of modern (or indeed postmodern) life has created a growing interest in the study of how social media is transforming sport (Wilson, 2007); the impact of new forms of control, surveillance, and conformity; the power of networked sport (Sugden, 2002; Castells, 2009).

The Cultural Analysis of Sport: *The Critical Turn*

From the 1970s, sociological research on sport took a critical and ethnographic turn, documenting structural and cultural inequalities within sport, and challenging many of its popular ideological representations. In the UK, this was led by authors, such as Hargreaves (1986), and allied scholars influenced by the Centre for Contemporary Cultural Studies based in Birmingham and in North America by Gruneau (1988, 1999) and others associated with the Massachusetts School. These approaches collectively attempted to re-center the sociological study of sport within a neo-Marxist perspective that retained an overriding concern with capitalist political economy. The theoretical grounding of much of this and other subsequent work (Sugden and Tomlinson, 2002) drew on the work of Gramsci's notion of hegemony. Analyzed through these more critical conceptual lenses, sports were no longer perceived as immutable and traditional entities but as contested social practices produced by athletes and organizers within the context of historically specific social constraints. Underpinned by the hegemony of bourgeois culture, modern sport became regarded as a powerful ideological tool naturalizing the celebration of competition and winning organized by the principles of capitalism (belief in the free market, commodification, Fordist managerialism, and division of labor). Although often hidden by its own institutionalized rhetoric, sport was recognized as a powerful ideological resource whose control was/is essential to the interests of dominant social groups. Also pivotal to this perspective was the idea of resistance. A counterpoint to sport's relative autonomy was that it could be harnessed by subordinate groups to challenge dominant ideologies and structures. Over three decades of research by a substantive international group of sport sociologists has drawn on these perspectives to explore issues of class, gender, race/ethnicity, sexuality, and disability that detail how hegemony is open to contestation and that sport is a powerful social arena in which meaning is not only constructed, but challenged and reconstructed (Craig and Beedie, 2010).

Paralleling the development of cultural studies, Eric Dunning, a student of Norbert Elias cultivated a somewhat different cultural orientation. Dunning incorporated Elias' grand theory of social development into a figurational approach to sport sociology (Elias and Dunning, 1986; Dunning, 1999). The figurational approach focuses attention on social processes and interdependencies that frame a diversity of issues ranging from hooliganism to athlete migrations and male bonding. The figurational approach distinguishes itself from the neo-Marxist and cultural studies perspectives through its comprehensive focus on social structures while taking into account a group processes such as emotions, class, gender and race distinctions, and communication patterns.

During the 1980s, feminist sport theorists such as Jennifer Hargreaves and Ann Hall incorporated aspects of the figurational approach with cultural studies and its emphasis on hegemony and resistance to move the analysis into the examination of gender relations within sport. The distinctive waves of feminist analysis (Hargreaves, 2004) have established a globally influential sociological field within which sport is theorized as a male preserve contributing to the oppression of women physically and sexually. Though this observation represents a fairly simplified summary, the wide-ranging and disparate analyses representing a feminist stance assert two notable principles: that sport offers women the potential for subverting male dominance by enabling women to experience their bodies as strong and powerful; that the dominant constructions of sport can be contested and restructured to align feminist social agendas.

Emerging from this feminist critique has been the necessity to construct a critical understanding of the physicality of sport. Most classical sociological theory either ignored or diminished the social significance of the body. Yet the empowerment and transformation of the body is at the heart of sport's transformative potential. The synthesis of Foucault and feminist theory has enabled sport scholars to problematize the body and establish it as a site of control and domination through which power is exercised and enjoyed (Giulianotti, 2005). While bodies can be empowered through sport, they are also constrained, shaped, consumed, and act as a powerful ideological resource.

A significant development within cultural studies was the inclusion of a range of cultural factors and artifacts as essential to a critical analysis of the social meanings of sporting events. A growing literature offers critical readings of the mediated discourses surrounding sporting events (Whannell,…) and global sports stars such as Michael Jordan and David Beckham within an ever-present celebrity culture (Kellner, 1996; Cashmore, 2014; Smart, 2005). In its expose of the discursive meanings of sport and its analysis of how these narratives do ideological work reproducing and justifying depictions of meaning and presence (plus absence), this still-growing body of literature represents some of the most influential work within the general field. As Cashmore (2013) details, images of black athletes, even those purporting to offer positive images that are intended to subvert racist

attitudes, still retain the potential to reinforce racial stereotypes associating black athletes with crime, drugs, and poverty. The same is also true for many of the normative images of women in sport, which consistently emphasize those athletes who are deemed to be marketable through their sexualized attractiveness and adherence to the norms of heterosexualized femininity while making invisible those who fail to fit this profile.

Partly informed by the feminist critique of sport, postmodern and poststructuralist perspectives started to have a significant influence on the study of sport from the late 1980s and 1990s. Some of the distinguishing features of these perspectives are their rejection of metatheoretical analyses and their 'grand narratives' (Lyotard, 1984) and their replacement by an emphasis on diversity and contingency. In relation to sport, it's proponents (Rail, 1998) recognized the dominance of transient consumer culture and its use of images and spectacles to enculturate individuals and their search for identity into an endless cycle of (meaningless) consumption and illusion (Baudrillard, 1995). Through their use of yearly cycles of professionalized sport and quad annual sporting mega events such as the Olympics and World Cup (Roche, 2000), a dominant and pervasive media culture implodes sport into realms of hyperreality, mass entertainment, commodification, and passive consumption (rather than active participation). At the pinnacle of these processes are the lifestyles and images of celebrity athletes albeit that these lives are often illusionary and fundamentally disconnected from the lives and everyday reality of their adoring fans. Although its analytic presence has now diminished, the principal value of the postmodern and poststructuralist stances was their questioning of the authority of sport sociology's more structuralist and realist approaches.

Although it has retained an interest in sport's location within popular culture one of the responses to the postmodern turn has seen a significant reexamination of the institutions of sport and their structural intersections with political and cultural power structures (Sugden and Tomlinson, 2002). These examinations often focused on the globalized sport industry and its associated global flows of people, ideas, culture, capital, commodities, and technology. Some of the concerns evident in this work have been the impacts on national identity and its coherence within a multicultural world; growing levels of inequality and social instability; the maintenance of sporting traditions and community; equitable sport development.

Sport in the Twenty-First Century: Emergent Sociological Perspectives

As we progress ever deeper into the twenty-first century, sport sociology has a number of emergent challenges that will become more persistent as time passes. The economic crisis of 2008 has created a number of challenges and contradictions facing the global capitalist system (Harvey, 2014; Stiglitz, 2013; Krugman, 2013). In respect to sport these concerns are (re) directing many sport sociologists to return to the issues of inequality and social and economic development. Though concerns about equity, especially in relations to gender and race, will remain highly significant within the field, there is a diversity of emergent transformative processes attracting attention within sport sociology. Some of the most prominent concerns that will frame significant research and analysis over the coming years are the links between the growth of interactive online gaming, social networking sites, and their relationship to the formation and reproduction of sedentary lifestyles; concerns regarding health and well-being especially in respect to the global explosion in levels of obesity; the impact of technological innovations that range across diverse areas of inquiry such as the media and the body – this latter issue is particularly evident through the growing presence of disability sport; public disquiet about sport governance in some of its undemocratic and unregulated power structures within globally important sport bodies such as FIFA and the IOC; finally, sports intersections between its often-abstract globalized structures and the lived realities of its performance within local everyday contexts of life will continue to provide insights into the production of social life in the twenty-first century.

See also: Bourdieu, Pierre (1930–2002); Castells, Manuel (1942–); Elias, Norbert (1897–1990); Environmental Sociology; Ethnography; Foucault, Michel (1926–84); Governmentality; Modernity; Multiculturalism; Popular Culture; Social Classification; Sport, Historiography of.

Bibliography

Baudrillard, J., 1995. The virtual illusion. In: Theory Culture and Society, vol. 12, pp. 97–107.
Bourdieu, P., 1986. Distinction. Routledge and Kegan Paul, London.
Bourdieu, P., 1988. Program for sport sociology. Sociology of Sport Journal 5 (2), 153–161.
Brohm, J.-M., 1978. Sport: A Prison of Measured Time. Ink Links, London.
Cashmore, E., 2013. Black Sportsmen. Routledge, London.
Cashmore, E., 2014. Celebrity Culture, second ed. Routledge, London.
Castells, M., 2009. The Rise of the Network Society: Information Age: Economy, Society, and Culture, second ed., vol. I. Wiley-Blackwell, London.
Coakley, J., 2009. Sports in Society: Issues and Controversies, tenth ed. McGraw-Hill, Boston.
Cole, C.L., 1994. Resisting the canon: feminist cultural studies, sport and technologies of the body. In: Birrell, S., Cole, C.L. (Eds.), Women, Sport and Culture. Human Kinetics, Champaign, pp. 5–29.
Cole, C.L., 2002. Body studies in the sociology of sport. In: Coakley, J., Dunning, E. (Eds.), Handbook of Sport Studies. Sage, London.
Craig, P., Beedie, P. (Eds.), 2010. Sport Sociology. Learning Matters, Exeter.
Craig, P., Mellor, G., 2010. Sport's organisation and governance. In: Craig, P., Beedie, P. (Eds.), Sport Sociology. Learning Matters, Exeter.
Dunning, E., 1999. Sport Matters: Sociological Studies of Sport, Violence and Civilization. Routledge, London.
Edwards, H., 1969. The Revolt of the Black Athlete. Free Press, New York.
Elias, N., Dunning, E., 1986. Quest for Excitement. Basil Blackwell, Oxford.
Fine, G.A., 1987. With the Boys: Little League Baseball and Preadolescent Culture. University of Chicago Press, Chicago.
Foley, D., 1990. The great American football ritual: reproducing race class and gender inequality. Sociology of Sport Journal 7 (2), 111–135.
Giddens, A., 1981. A Contemporary Critique of Historical Materialism. Macmillan, London.
Giulianotti, R., 2004. Introduction: sport and social theorists – a plurality of perspectives. In: Giulianotti, R. (Ed.), Sport and Modern Social Theorists. Palgrave, London.
Giulianotti, R., 2005. Sport: A Critical Sociology. Polity Press, Cambridge.
Gruneau, R., 1999. Class, Sport and Social Development. University of Massachusetts, Amherst, MA.

Hargreaves, J., 1986. Sport, Power and Culture: A Social and Historical Analysis of Popular Sports in Britain. Polity Press, Cambridge.
Hargreaves, J., 2004. Querying sport feminism: personal or political? In: Giulianotti, R. (Ed.), Sport and Modern Social Theorists. Palgrave, London.
Harvey, 2014. Seventeen Contradictions of the End of Capitalism. Profile Books, London.
Huizinga, J., 1955. Homo Ludens: A Study of the Play Element in Culture. Beacon Press, Boston.
Jarvie, G., 2006. Sport, Culture and Society: An Introduction. Routledge, London.
Jarvie, G., McGuire, J., 1984. Sport and Leisure in Social Thought. Routledge, London.
Kellner, D., 1996. Sport media culture and race—Some reflections on Michael Jordon. Sociology of Sport Journal 13 (4), 458–467.
Kenyon, G.S., Loy, L.W., 1965. Toward a sociology of sport: a plea for the study of physical activity as a sociological and social psychological phenomenon. Journal of Health, Physical Education and Recreation 38 (24), 68–69.
Krugman, P., 2013. End the Depression Now! Melrose Road Partners, New York.
Loy, J., McPherson, B., Keynon, G., 1978. Sport and Social Systems: A Guide to the Analysis, Problems and Literature. Addison-Wesley Publishing, London.
Lyotard, J.-F., 1984. The Postmodern Condition. Manchester University Press, Manchester.
Maguire, J., 1999. Global Sport: Identities, Societies, Civilizations. Polity Press, Cambridge.
Rail, G. (Ed.), 1998. Sport and Postmodern Times. SUNY Press, Albany, NY.
Rigauer, B., 1981. Sport and Work. Columbia University Press, New York.
Roche, M., 2000. Megaevents and Modernity: Olympics and Expos in the Growth of Global Culture. Routledge, London.
Scott, J., 1971. The Athletic Revolution. Free Press, New York.
Smart, B., 2005. The Sport Star: Modern Sport and the Cultural Economy of Sporting Celebrity. Sage, London.
Stigliz, J., 2013. The Price of Inequality. Norton and Co., New York.
Sugden, J., Tomlinson, A. (Eds.), 2002. Power Games: A Critical Sociology of Sport. Routledge, London.
Sugden, J., 2002. Network football. In: Sugden, J., Tomlinson, A. (Eds.), Power Games: A Critical Sociology of Sport. Routledge, London.
Wilson, B., December 2007. New Media, social movements, and global sport studies: a revolutionary moment and the sociology of sport. Sociology of Sport Journal 24 (4), 457–477.

Sports, Economics of

Dejonghe Trudo, University of Leuven (KUL), Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.W. Scully, volume 22, pp. 14938–14944, © 2001, Elsevier Ltd.

Abstract

Professional team sports have a peculiar economic dimension. One of the main differences between the organization of professional team sport and the market is that in a competitive market firm chooses its level of output to maximize profit. In contrast, sport clubs are limited in their output by the league, which determines the membership and number of clubs and the number of games played in a season. Another peculiar fact is that team owners in US professional team sports have a more profit-maximization goal and European club owners are more interested in win maximization, which results in big financial losses. This results in a different approach and use of measurements. In the closed leagues of the US is uncertainty of outcome (UOO) and competitive balance, the solid fundament of sport economic analyses, which resulted in the introduction of cross-subsidizations such as revenue sharing, salary caps, and drafting systems. In historical times was there also a reserve clause for transferring players between clubs, but this was forbidden by law in both continents and free agency was introduced. In the open leagues of Europe is UOO less important and are most of the fans focused on some major teams who have to compete for the championship. Cross-subsidization is not a main issue and in most countries is their only a collective sale of broadcasting rights.

Introduction

Sport economics is an exponential growing study topic in the academic world. It started in North America with a central and neoclassical focus on professional team sports and diffused first to the UK and then to the rest of Europe and finally to the world. The increasing cash flow in the sector combined with leisure time and purchasing power resulted in the emergence of a new economic industry. Sport as a leisure industry became a hot item and besides the organization of professional sport leagues came topics as the economic impact or feasibility of sport events, financing sport infrastructure, measuring return on investment, and more European items namely, innovation in sports industry, public funding of sport, sport participation in relation to health economics, and social impact of sport, a central topic in academic research.

We notice in American handbooks on sport economics such as *Sports Economics* of Fort (third edition in 2011) and in *The Economics of Sport* (fifth edition in 2014) by Leeds and von Allmen a total absence of recreational sport. On the other hand, we have the European-oriented handbook *Sports Economics: Theory, Evidence and Policy* of Downward et al. (2009) and *Sport en economie: samen in de spits* of Dejonghe, T. (2012) that has as much chapters on participation sport than on professional sport leagues and also contributes on the economics of sports events and infrastructure. The book of Downward et al. (2009) was more or less a successor of *The Economics of Professional Team Sports* by Downward and Dawson (2000), a more European approach of professional team sports. A neoclassical theoretical framework of European and American professional league models is Késenne's *The Economic Theory of Professional Team Sports: An Analytical Treatment* (2007).

In this article, we will only discuss the economics of professional team sports. The 'planned' organization of leagues in North America will be compared with the less rigid organization of Europe.

The Peculiar Economics of Professional Team Sports

Professional team sports can be analyzed from an economic perspective in which labor (mainly players), capital, and land (stadiums and other facilities) are combined by clubs that supply teams to produce a saleable product – the game or contest. To obtain revenue to pay for inputs it is necessary to exclude nonpaying spectators, hence the universal use of stadiums and a traditional suspicion of live broadcasting of sports. For teams to cover their costs each must be guaranteed a minimum number of home games for which tickets can be sold, which suggests 'league' or round-robin tournaments rather than knock-out competitions. As an organizational structure leagues comprise sets of individual clubs, which fall within the broader governance structures. Another difference between the organization of professional team sport and the market is that in a competitive market firm chooses its level of output to maximize profit. In contrast, sport clubs are limited in their output by the league, which determines the membership and number of clubs and the number of games played in a season. The league is a cartel with clubs acting as firms and can be seen as a violation of antitrust laws (Sherman Act (1890) and Clayton Act (1914)). Despite these legal frameworks, there has been little government intervention in sport markets.

There is potential division of interests between league and club because of the existence of the externalities in production. The first publication about economics of professional team sport were articles of Rottenberg (1956) and Neale (1964) about the players' labor market in baseball and the peculiar economics of professional sport in the US. In Europe, came a reaction of Sloane (1971) who noticed the existence of a different professional team sport model in Europe. Those articles created a massive spin-off of research and publication to understand and optimize the professional team sports and their leagues. In the US are most of the interesting questions on the economics of sport related to professional team sport. This results in a focus on the industrial organization of sports in the

major leagues (National Football League (NFL), Major League Baseball (MLB), National Basketball Association (NBA), and National Hockey League (NHL)) and government subsidies on professional sport.

US Profit-Maximizing Objectives versus European Win-Maximizing Objectives

Drawing upon the historical development of professional sports in Europe and the US, most US writers assume that profit maximization is the goal of team owners. In contrast, Sloane (1971) argued that European club owners might more accurately be viewed as utility maximizers (later on referred to as win maximization) where a major determinant of utility is playing success (winning), although profit is important for the club's long-run financial stability. Forgoing profits, and even underwriting losses, therefore suggests that club ownership is an act of consumption more than investment. The impetus to this assumption was the relegation and promotion of hierarchical sports systems in Europe, coupled with the historic desire in European football for success in international competition.

Sloane's (1971) proposed team utility function is

$$U = U\left\{P, A, X, \left(\prod - \prod^{\mathrm{MIN}}\right)\right\}$$

where 'U' is the utility level attained, which depends upon performance on the field 'P,' home attendance 'A,' health of the league 'X,' and on the excess of posttax profit 'Π' over a minimum acceptable level of posttax profit 'Π^{MIN}.' The notion of an acceptable minimum profit reflects factors such as the club's need to secure its long-run future. Any club that has outstanding debt has to display the ability to service that debt. Over and above servicing debt, the club may need to generate an acceptable level of profit; if it is a publicly quoted limited liability company, the directors need to produce enough dividends to keep the shareholders happy. If they cannot, the share price will fall and the company may become the subject of a hostile takeover. In contrast, if a club has access to a sympathetic and wealthy owner (Chelsea with Roman Abramovich, Manchester City with Sheikh Mansour, AS Monaco with Dimitry Rybolovlev, Paris Saint Germain with Nasser Al-Khelaifi, Juventus with the Agnelli (FIAT) family, AC Milan with Silvio Berlusconi, Shakthar Donetsk with Rinat Akhmetov, etc.) or to favorable treatment from local or national government (Real Madrid, Barcelona, AS Monaco) the minimum acceptable level could actually be negative.

Theoretically, one can expect that other things equal an increase in any one of the four factors will make the supporters, the owners, and the management happier. As Sloane notes, the more closely the four explanatory variables are correlated with each other the more probable it is that the predictions of the profit-maximizing and utility-maximizing models will resemble one another. Unless the predictions differ in some readily identifiable manner attempts to distinguish the objectives of teams will fail.

Uncertainty of Outcome and Competitive Balance

The first contributions to the economics of professional team sports stressed the importance of a key externality in professional team sport known as the uncertainty of outcome (UOO) hypothesis. This implies a more or less equal distribution of talent if there has to be UOO and that gate receipts and broadcasting rights depend crucially on UOO of the games played in the league and on the competitive balance in the league. The North American profit-maximization approach defines the UOO as the solid fundament of sport economic analyses and makes competitive balance as the most important item of sport leagues. If the top club always beats the next ranked club, and so on, the games would be exhibitions, not contests, and they would have little market value. Some degree of UOO is necessary in sports, and this uncertainty is greater the more equal the playing strengths of the competitors. Greater equality of playing strengths among the clubs is wealth maximizing. Attendance at games and viewership on television is higher the closer the standings of the clubs are over a season. The acceptance of UOO and competitive balance implies that the market of professional team sports differs significantly from most industries. Competitive balance depends mainly on the comparative financial strengths of the clubs. Teams earn revenue primarily from ticket sales and broadcast fees. The main cost is player payroll. Attendance is determined chiefly by market (city) size and the club's historical reputation. Since teams are located in cities of markedly different size, the prospect of fielding a competitive team depends in part on where the club is located, but also on how revenues are divided up among the clubs. This resulted in a system of cross-subsidization between the clubs in American leagues. In the US, leagues have intervened in sporting labor market or regulated the distribution of club revenues from broadcasting and in some cases of gate receipts. The arguments in the US are that with a distribution of clubs in unequal size markets and with a ban on unilateral relocation of clubs to the more lucrative markets, the more unequal the division of revenue among clubs the less will be the competitive balance within a league. The desire to promote long-run UOO or competitive balance is due to the fear that domination by a few teams may result in a loss of consumer interest, revenue, and profit. The problem is that there is no convincing body of evidence that sports leagues do evolve toward perfect balance. Clearly, if the bigger members of an industry systematically grow more rapidly than smaller ones, and if the process continues, the industry will become more concentrated. If the small firms grow faster, and if the process continues, the industry will tend to become less concentrated. Given enough time, the sizes of all firms (as measured by, say, revenue or assets) would converge on an 'equilibrium' value, although whether this occurs also depends on random factors. This means that convergence or divergence in the relative distribution of revenues would have an impact on competitive balance and could be a support for cross-subsidization.

In Europe, noticed Dobson et al. (2001) for the English Premier League that from the 1920s until the late 1950s there was a period of steady adjustment toward equilibrium. From the late 1950s until the late 1970s it is suggested that there was no tendency to revenue convergence but this occurred from the late 1970s until about 1990 and convergence seems also to have broken down in the 1990s. The authors conclude there may have been two periods of relative equilibrium, followed by two of major change. The causes of these major changes

occurred during a period of marked social change. In the period 1950–1970s, population drift toward southern England and also from the inner city, and increased access to private transport may have been important, as well as the media in the creation of national support for some teams. It also starts about the time that the maximum wage was abolished, forcing clubs to increase admission charges that had been virtually unaltered since the 1920s. The 1990s breakdown of convergence took place against the simultaneous effects of the Bosman Ruling (*see* The Sport Labor Market) that enhanced players' bargaining power, the influx of money from broadcasters, and the setting up of the Premier League. Add the transition to all-seat stadiums and the necessary rise in ticket prices and there are plenty of factors that would be expected to shake up the industry structure. However, in summary it seems is that the UOO hypothesis seems to be rather overworked as a phenomenon that affects attendances in as much that it not ubiquitously important despite the ingenuity with which it has been measured.

Methods of Cross-Subsidization

Another remarkable characteristic of American leagues is the limitation of the number of teams and the absence of promotion and relegation from or toward a hierarchal lower league. These closed leagues in US sport create, with exception of some major metropolitans, spatial monopolies and a latent demand for enlargements of number of clubs in the league. An advantage of this scarcity of supply is that it has been used to leverage public investments for sport facilities. The combination of the US acceptance of cartelization of the closed sport leagues and the 'need' for UOO and competitive balance results in the creation of specific peculiar measurements known as cross-subsidization. An economic rationale for cross-subsidization lies in the market structure of sporting leagues that can be viewed as a producer cartel within which clubs need to compete in sporting terms, but also to cooperate to ensure that the sports are managed effectively. To attain these ends, leagues have traditionally set the terms of sporting competition and have directly influenced the economic aspects of competition, influencing admission price structures, negotiating television deals and sponsorship arrangements, and last, but not least, controlling the terms upon which players may move between clubs.

The Sport Labor Market

While different in detail, leagues have intervened in sporting labor market or regulated the distribution of club revenues. Leagues have attempted to influence the distribution of talent by applying three major types of labor market policy instruments: drafting systems, salary caps, and reservation option arrangements.

The best-known example of *a drafting system* is the 'Rookie draft,' where the weakest teams have in some degree the first pick of the new playing talent at a price that is discounted on the player's market value. The intention of this measurement is that talent will be allocated between the clubs to improve competitive balance. For example, the source of recruitment to the NFL is college football, which comprises amateur players. The reverse draft ensures that every year the professional teams obtain bargaining access to the new crop of graduates in reverse order of how they finish in the professional league. The intention is to give last year's least successful outfits the first option to pick new recruits. Consequently, the weaker teams have the opportunity to enhance their stock of playing talent at a price, that is, probably discounted on the player's genuine market value. Over time the intention is that such a system reallocates talent between the teams to promote competitive balance. Of course, in hierarchical leagues, with promotion and relegation, the replacement of weaker teams with stronger teams reallocates talent in a similar way. There is, however, one potentially major flaw in the thinking behind drafts, or for that matter how promotion and relegation affects talent distribution. For example, in the draft system, despite the opportunity for weak teams to initially sign new, higher quality talent, there remains an economic incentive for the stronger teams to buy them from the weak ones and the weak teams to sell them to the stronger teams. Leaving aside the personal ambitions of players to wish to play for stronger teams, the clubs would have incentives to trade because the players will probably be offered a contract with a wage below their market value by the team allocated first pick in the draft. Consequently, the gap between what the player can earn from the weaker team and the true market value of the player becomes an economic rent over which teams can bargain. The stronger team can offer the weaker team more than they have paid for the player, while not paying the player their market value. It follows that both clubs can make a relative financial gain through the trade. The player may or may not receive some of this gain. This depends upon the contract of the player.

Over their histories, in some leagues, players have been restricted from moving between clubs within a league, the so-called *reserve clause*. Club owners have claimed that reservation of players or the imposition of transfer fees was necessary to maintain financial soundness of small clubs and to preserve competitive balance within leagues. Economists, since Rottenberg (1956) have been very skeptical of this claim and there is no empirical evidence to support it. While it is true that there is a tendency for the best players to migrate to the larger-city franchises, freedom of player movement or restrictions on player-initiated movement have no bearing on the distribution of playing talent within a league (El-Hodiri and Quirk, 1971). Under free agency, players will migrate to the best-paying clubs. The franchises that pay the most are those that expect the largest incremental revenue from those players' performances. Since an increment in the win percent raises revenue more in a large city than in a small city, the best players tend to migrate to the big-city franchises. At issue is whether in a market where player-initiated movement is forbidden (player reservation, transfer fees), would a small-city club keep a star player that it had acquired through a draft or an assignment? On the roster of that club, the player is worth, say, $1 million in incremental revenue. On a large-city franchise, that player's performance is worth, say, $3 million. Without a scrupulously enforced league ban on cash sales of player contracts, the franchise in the small city has an incentive to sell the player's contract to the big-city

club and capture some portion of that $2 million difference in incremental value. Therefore, players are allocated across clubs based on their highest incremental revenue. Whether the player or the club owns the contract (property right) is irrelevant to the allocation of playing talent within a league. That the efficient allocation of a resource is independent of who owns the property right is, of course, the famous Coase (1960) invariance theorem, but the essence of that theorem was discovered independently by Rottenberg (1956) in his earlier paper on the baseball players' market. While the distribution of playing talent is independent of the degree of freedom in the market for players, the distribution of the rents between owners and players is not. The players obtain a much higher share of the rents arising from their scarce talent under free agency than would be so under league restrictions on player-initiated transfers. Under player reservation, player pay is less than the players' contribution to club incremental revenue. Under free agency, the players, ultimately, receive all of the incremental revenue. It is shown that the intrateam salary structure, with its wide pay differential between starting and backup players is rational and wealth maximizing. Competitive balance is value maximizing, also in part, the appeal of professional team sport is the extraordinary playing ability of the athletes. This is demonstrated repeatedly on television and commented on widely. Such feats are value maximizing for the leagues, since they promote interest in the sport. Prior to the mid-1970s, baseball and basketball players could not negotiate freely with more than one team, and this restriction applied in football and in hockey, until more recent times. In European football (soccer), until the Bosman Ruling in 1995, player movement was restricted by the substantial transfer fee that had to be paid to the club owning the player's contract. The result of player reservation in professional team sport was that players were exploited economically by the owners, in the sense that they were paid less than the incremental revenue that their performance brought to the club (Scully, 1974). It is not the case that the athletes in professional sports were poor by any means, since they were paid on average several times a working man's wage. But, their pay was not equal to what they could have earned, if they were free to offer their talent to the highest bidding club. Free agency has led to an enormous explosion in player salaries and in the share of club revenue captured by the players. Concern about the heavy spending of a number of professional clubs across Europe resulted in the introduction in 2009 of UEFA Financial Fair Play measurements. It was hoped that the regulation would lead to a more 'level playing field' by preventing clubs with wealthy owners who make substantial cash gifts to their club from gaining an unfair advantage over other clubs that are run on a more substantial business model. The full implementation of Financial Fair Play was delayed until 2015 by the European Club Association, an independent body representing football clubs at European level.

In Europe, a general principle of EU law, better known as the Bosman Ruling, that discrimination on grounds of nationality regarding EU citizens was illegal and that all citizens are allowed to work in other EU countries under the same conditions, was the start of free agency in sport. The consequence of the Bosman Ruling was that all players at the end of their contract became agents who were free to move at an international labor market. This resulted consequently in the fact that clubs lost their monopsony power over players. On the contrary, market power went to those players who could sell their talents to the team that offered the highest wages. In combination with this power shift, European football faced the elimination of the limitation of the number of foreign players, which resulted in a substantial migration of player talent. The Bosman Ruling made factor mobility of labor possible and resulted in a reallocation of player talent. The arguments of the football federations and UEFA to defend the existing transfer system were that transfer fees are a form of revenue sharing and remuneration of assets and that without these financial compensations clubs would eradicate youth development. Parrish and McArdle (2004) notice that the European Court of Justice foresaw in its verdict alternative possibilities such as 'revenue sharing' and 'salary cap' to maintain or even increase competitive balance. The problem was that this kind of cross-subsidization would be difficult to implement in the structures of European football. This is because open league structures are historically embedded in European team sports. Moreover, UEFA and the national federations never took these alternatives into concern. The academic world, however, countered arguments on the probability of concentration of player talent and decreasing competitive balance. Késenne (1998), for example, indicated that North American sports economists, such as Scully (1989) and Quirk and Fort (1992), analyzed the elimination of the 'reserve clause' in the US, a system comparable to the transfer system before Bosman, in the MLB. Their conclusion, based on existing theoretical and empirical studies, was that professional leagues in the US would benefit if limitations on freedom of players movement were eliminated. Their central argument came from Rottenberg's (1956) article that anticipated on the already mentioned Coase theorem. Rottenberg argued that a limitation through a reserve clause would not prevent migration of talent toward major teams. In Coase's terms the distribution of resources, in this case player talent, would not be affected by the distribution of ownership rights. Translation of these theoretical findings to professional sports means that free movement of players will have no effect on their distribution and was defined as 'the invariance principle' (El-Hodiri and Quirk, 1971). This North American approach was used as an argument to demonstrate that the Bosman Ruling would have no impact on the distribution of talent in Europe. The problem was that the European professional team sports environment did not fulfill the main criteria of Rottenberg namely, that all firms had to be nearly equal if each is to prosper. A fundamental fact that the European court did not consider, at least not explicitly, is that football players do not move around in a common market as long as there are independent leagues. This implies that smaller national leagues in particular cannot afford to keep their most talented players in a free-agency market. Consequently, the transfer fee can be motivated as an instrument to stimulate development of talent in small-market teams. Haan et al. (2007) warned that free movement of players could be the death penalty for many European leagues. According to them, international differences will increase and international competition will become less exciting. Changing structures toward an open labor market in separated product markets resulted in a competitive disadvantage for smaller market leagues and their teams. The introduction of an open labor market resulted in a concentration of talent into the Big 5 (England,

Spain, Italy, Germany, and France) as well as more recently in Eastern European leagues with 'new money' (Russia and Ukraine). The main reason for this divergence is the opening of the labor market with migrating players in separated win-maximizing product markets with almost no revenue sharing. Some sports economists see a solution in a further Americanization of European football, resulting in a European Major League Football that resembles an American Major league, with one of more semiclosed divisions and forms of cross-subsidization such as salary cap and revenue sharing. From an economic viewpoint, creating a European Major League Football would be the natural outcome of the transformation of locally embedded football into a more business- and media-oriented entertainment. This idea, however, stands in opposite of European football culture with its strong relations with local communities, called 'topophilia,' and the existence of historical traditions and rivalries in every country. Dobson and Goddard (2011) also notice that those in favor of a total withdrawal of major teams from their domestic league underestimate the importance of domestic history and tradition as typical characteristics of football identity in Europe. Arnaud (2006) refers also to the tradition of the football fan who wants a 'twin-pillar' structure consistent of a national competition and a European competition. In sum, local embedment and identity are defining characteristics of European professional football. The EU recognized the specificity of sport in its 2007 voted White Paper on Sport where it takes account on the specific characteristics of sport. Hereby, they recognize the specificity of sporting activities, sporting rules, and sport structures. One of these specificities is the need to ensure uncertainty concerning outcomes and to preserve a competitive balance between clubs taken part in the same competition.

On the whole, the introduction of free agency has eliminated monopsonistic exploitation of players in professional team sports. Nevertheless, in the US leagues, pockets of exploitation exist, particularly among promising rookies, journeymen players, and first-round draft choices from the amateur ranks. This is so because there is not complete freedom in the players' market, such as would occur with teams competing for all players by offering 1year or multiyear contracts for their services. New entrants into the professional ranks, if they survive the intense competition for their positions, must wait several years before they are free to negotiate with other clubs.

Salary caps unlike drafting systems directly target the financial cost of talent. Salary caps can either involve fixing the cost of individual and thereby aggregate player costs by establishing a maximum wage or represent a maximum absolute amount that clubs can spend on players in total, or expressed as a percentage of team (or league) turnover. In the US, the NBA, hoping to strengthen weaker teams' finances, adopted a salary cap, after the alternative of gate-revenue sharing was blocked by the combined efforts of the strongest drawing teams. The frequently offered justification for the control of wages and/or wage bills is that, in principle, it makes the best talent affordable to all teams and which will help to produce competitive balance. A salary cap can be hard or soft. A hard cap means that the total wages in a team may not exceed a maximum amount. In a soft cap system may a certain maximum be exceeded in some circumstances but a luxury tax must be paid. On the other hand, there exists a salary floor. That is a minimum amount that must be spend by a team on wages. One mechanism, invented by basketball is the team salary cap, a device that in theory imposes a maximum salary as a percent of gross revenue (and, a minimum percentage that applies to small-city clubs). American football has a team salary cap, and baseball wants one, but has been blocked in its ambition by the players' union who only allowed a luxury tax. Can a salary cap equalize playing strengths within a league? Suppose that a binding salary cap is imposed that fixes player salaries at a certain percentage of league revenues. The marginal cost of playing talent falls under the salary cap from its level under free agency. If scrupulously enforced, all clubs have the same team salary costs, and if there are no information asymmetries about expected playing talent and no differences in coaching quality, all clubs will have 0.500 records. Like the solution of player reservation with a ban on cash sales of player contracts, the distribution of win percents brought about by a hard salary cap is stable, but it is not profit maximizing nor does it maximize league revenue. A large-city team can earn more profit by acquiring better playing talent and a small-market club can earn more by lowering, its level of playing talent.

In basketball, the team salary constraint is relaxed significantly by the so-called Larry Bird exemption. Under the exemption, a team may re-sign a veteran free agent at whatever cost, and not have that player's salary count in the salary cap. This has led to most clubs having player payrolls well in excess of the cap, and this transforms the team salary limit concept into a soft or nonbinding constraint. The Larry Bird exemption was the main point of dispute between the basketball players and the owners, prior to the 1998–99 season. Salary caps have led, also, to creative financial arrangements such as large player signing bonuses relative to annual salary, back-end-loaded long-term player salary contracts, deferred player compensation, and the shifting of revenue from team operations to other, nonteam entities. Sufficient erosion in the constraint of the salary cap, as with a player reservation system with player transfers for cash, can lead to the same distribution of playing talent as is attained with unrestricted free agency.

In Europe is a salary cap in soccer absent and occurs only in the more recent formed professional leagues in rugby and hockey.

Revenue Sharing

If the degree of competitiveness in the players' market does not affect competitive balance, will it be affected by the division of revenue among the clubs? Traditionally, of course, this has concerned the gate revenue – monies paid by spectators at the turnstiles. Because imbalance is created by teams with different revenue bases, which are then spent on playing talent, to redistribute some home-team revenue implies that a greater absolute amount of income is transferred from the larger club to the smaller club, than the other way around. This means that the larger club experiences a net reduction in revenue, while the smaller team a net gain in revenue. As in many other team sports, away-team shares of gate revenue have fallen steadily over the years, and, currently, a variety of arrangements exist in the US ensuring that the home team receives the largest share of revenue. The NFL operates an unusually generous 60/40

split in favor of the home club, while the NBA and NHL have no gate sharing. In European soccer exists currently only gate sharing arrangements in cup competitions.

As far as TV revenue is concerned, in the US local TV coverage provided no revenue for visiting teams although national TV revenue was shared. In European football exists in most leagues a collective bargaining of broadcasting rights and a part of the total sum is been distributed equally among teams. The other part is divided according to recent and/or historical performances and in Italy is a part divided according to the number of attendances and the inhabitants of the local market.

See also: Behavioral Economics; Mass Media and Sports; Sport, Historiography of; Sport, Sociology of.

Bibliography

Arnaud, J.-L., 2006. Independent European Sport Review. European Union, Brussels.
Coase, R., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.
Dejonghe, T., 2012. Sport en economie: samen in de spits. Arko Sports Media, Nieuwegein.
Dobson, S., Goddard, J., 2011. The Economics of Football, second ed. Cambridge University Press, Cambridge.
Dobson, S., Goddard, J., Ramlogan, C., 2001. Revenue convergence in the English soccer league. Journal of Sports Economics 2 (3), 257–274.
Downward, P., Dawson, A., 2000. The Economics of Professional Team Sports. Routledge, London.
Downward, P., Dawson, A., Dejonghe, T., 2009. Sports Economics: Theory, Evidence and Policy. Elsevier, Butterworth-Heinemann, Oxford.
El-Hodiri, M., Quirk, J., 1971. An economic model of a professional sports league. Journal of Political Economy 79, 1302–1319.
Fort, R., 2011. Sports Economics, third ed. Prentice Hall, Upper Saddle River.
Haan, M., Koning, R., Witteloostuijn, A., 2007. Competitive balance in national European soccer competitions. In: Albert, J., Koning, R. (Eds.), Statistical Thinking in Sport. Chapman & Hall/CRCPress, Boca Raton, pp. 63–75.
Késenne, S., 1998. De aangekondigde dood van het Belgische voetbal. Streven 65, 511–519.
Késenne, S., 2007. The Economic Theory of Professional Team Sports. Edward Elgar, Cheltenham, UK.
Leeds, M., von Allmen, P., 2014. The Economics of Sports, fifth ed. Pearson, London.
Neale, W., 1964. The peculiar economics of professional sport. Quarterly Journal of Economics 78 (1), 1–14.
Parrish, R., McArdle, D., 2004. Beyond Bosman: the European Union's influence upon professional athletes' freedom of movement. Soccer and Society 7, 403–419.
Porter, P.K., Scully, G.W., 1996. The distribution of earnings and the rules of the game. Southern Economic Journal 63, 149–162.
Quirk, J., Fort, R., 1992. Pay Dirt: The Business of Professional Team Sports. Princeton University Press, Princeton, NJ.
Rottenberg, S., 1956. The baseball players'-labor market. Journal of Political Economy 64, 242–258.
Scully, G., 1974. Pay and performance in major league baseball. American Economic Review 64, 915–930.
Scully, G., 1989. The Business of Major League Baseball. University of Chicago Press, Chicago.
Scully, G., 1995. The Market Structure of Sports. University of Chicago Press, Chicago.
Sloane, P., 1971. The economics of professional football: the football club as a utility maximiser. Scottish Journal of Political Economy 17 (2), 121–146.

Standpoint Theory, in Science

Alison Wylie, University of Washington, Seattle, WA, USA
Sergio Sismondo, Queen's University, Kingston, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Sismondo, volume 22, pp. 14952–14955, © 2001, Elsevier Ltd.

Abstract

Standpoint theory is based on the insight that those who are marginalized or oppressed have distinctive epistemic resources with which to understand social structures. Inasmuch as these structures shape our understanding of the natural and life worlds, standpoint theorists extend this principle to a range of biological and physical as well as social sciences. Standpoint theory has been articulated as a social epistemology and as an aligned methodological stance. It provides the rationale for 'starting research from the margins' and for expanding the diversity of backgrounds and experience represented in scientific communities.

Standpoint theory is a social epistemology and aligned methodological stance predicated on the understanding that, as recent work in science and technology studies (STS) has shown, scientific knowledge of all kinds is irreducibly social (*see* Science, Sociology of). It is characterized by a distinctive 'inversion' thesis according to which those who are marginalized or oppressed, and therefore often discounted as knowers, may be better positioned to understand social relations and a range of other phenomena than those who are comparatively privileged. This thesis underpins research programs within the social and behavioral sciences and offers a rationale for participatory action research (PAR) and community-based participatory research (CBPR) captured by critical race theorists, feminists, and others who advocate 'starting research from the margins.' It has also been extended to the biological and physical sciences by practitioners and STS scholars who are intent on understanding how the social relations that structure scientific inquiry shape research agendas and outcomes across the sciences.

Standpoint Theory's Central Argument

The central arguments for the epistemic advantage of those who are oppressed or marginalized stem from Hegel's 'master–slave' dialectic, Marx's discussions of ideology, and particularly from Georg Lukács' discussion of the privileged standpoint of the proletariat. In his 1922 essay 'Reification and the Consciousness of the Proletariat' (in Lukács, 1971), Lukács argues that the proletariat has the potential to develop a privileged standpoint from which to understand the rationalized structure of commodities and capital. Under capitalism, the bourgeois tend to accept their position as natural and just; their class privilege makes it possible to retain a rationalizing illusion of subjectivity with respect to their own agency. By contrast, Lukács argues, members of the proletariat routinely run up against arbitrary structures that oppress them and remove much of their subjectivity. In this they directly confront the constraining nature of the social relations that sustain capitalism, and the disconnect between these conditions of oppression and its rationalization. They thus have the resources to recognize that the system and its rationalization is a historical product, its contingency masked by legitimating ideology. Louis Althusser and Antonio Gramsci also offer key insights for understanding how the lived experience of oppression can give rise to the formation of a critical standpoint on oppressive social arrangements (Hennessy, 1993). In essence, then, these originating forms of standpoint theory offer an account of the blinding power of ideology and of how it can be overcome (*see* Ideology, Sociology of).

Later standpoint theorists make the case that these insights can be extended to a range of intersectionally defined social locations. Social differentiation along lines of gender and race, for example, also has the potential to put those disadvantaged by systemic inequality in a position to understand its mode of operation and effects with particular clarity. As developed by feminist social scientists and theorists through the 1980s – for example, by Nancy Hartsock (1983), Patricia Hill Collins (1986), Dorothy Smith (1987), and Sandra Harding (1991) – the cornerstone of feminist standpoint theory is a recognition that the rationalization of gender norms as natural reifies social relations, and the observation that direct experience of gender discrimination has been the catalyst for articulating a distinctive perspective on gender relations. Crucially, feminist theorists argue that situated knowledge of this kind does not, in itself, constitute a standpoint (Hennessy, 1993). It is through working against discrimination that it becomes possible to understand the contingency and the unfairness of ideologically rationalized conditions of life that go unrecognized or are unproblematic for those in positions of relative privilege. To the extent that these insights generate a critical standpoint on knowledge production – on the ways in which the institutions and products of science are inflected by systemic inequity and the ideologies that rationalize them – the situated knowledge of the dispossessed becomes relevant to doing and to understanding science.

Criticisms and Elaborations

The central and most controversial claim associated with standpoint theory is that particular social positions have the potential to produce privileged perspectives, not merely other perspectives. This has attracted three lines of criticism: that, to

sustain an attribution of epistemic privilege, standpoint theorists must be committed to an essentialist conception of the social locations and identities of oppressed groups; that they embrace an untenably strong formulation of the 'inversion' thesis, on which subordinate social location, in itself, warrants an automatic attribution of superior insight; and that these commitments entail a corrosive, self-undermining relativism. Although the central tenets of standpoint theory have been developed in diverse ways by its advocates (see the Signs symposium on Hekman (1997)), it is difficult to identify anyone who endorses essentialism or automatic privilege, much less the relativism assumed to follow from these theses. These views are more typically attributed to standpoint theorists, often in the context of the science wars and identity politics debates of the 1990s, than embraced or presupposed by them.

Where the charge of essentialism is concerned, standpoint theorists generally take as their point of departure established critiques of naturalizing (e.g., biological determinist) conceptions of gender and race difference, critiques that demonstrate the historical contingency of these social categories and of the systemic inequalities they underpin. Rather than simply inverting the terms of these constructs, standpoint theorists understand them to function as ideological constructs in terms of which normative gender and race relations are reified and rationalized. In this spirit the assumption that women must all have something in common – some distinctive shared attribute, experience, or situation – has been discredited by feminist researchers; intersectional analyses (studies of the intersections among disenfranchised groups) reveal a great deal of diversity among those who are structurally subordinate. By extension, the proliferation of feminist perspectives militates against any assumption that there is one correct feminist vantage point from which the true nature of gender relations will be revealed. The methodological formulations of standpoint theory associated with Smith and developed by Harding – the recommendations to ask how the 'everyday world' looks to those who primarily operate 'off-stage' (Smith, 1974) and to 'think from women's lives' (Harding, 1991) – signals an early commitment to intersectional analysis and suggests how diversity can be built into standpoint theory.

The converse worry is that if there is no gender-specific experience shared by all women, then there is nothing to ground a distinctive women's perspective or feminist standpoint. Similar concerns were raised in the context of the identity politics debates of the 1990s with respect to race and class identity, and a number of other subject positions that anchor standpoint theories. This objection only holds, however, if it can be demonstrated that systematic social differentiation makes no difference to the experience of those so identified. Socialist feminists like Hartsock and Smith, who developed the framework for feminist standpoint theory in the 1970s and early 1980s, made the case for recognizing the ramifying impact of sex/gender norms on the lives of those socialized as women, shaping the material conditions of their lives, their relations of production and reproduction, and thus their epistemic capacities, their access to epistemic resources, and their credibility as knowers. More recently, feminist and critical race theorists (e.g., Alcoff, 2006; Moya and Hames-García, 2000) have argued that collective identities do not have to be 'quasi-universal' to give rise to distinctive bodies of experience and perspective that can mobilize a critical standpoint. This resonates with Rosemary Hennessy's recommendation that a feminist standpoint be understood as a form of "critical practice"; its ground is not identification with an oppressed social group, but a "disidentifying collective subject of critique," a subject position constituted through systemic analysis of the social production of difference (1993: p. 30). The key point that emerges from this literature is that, although they are contingent historical constructs, systems of inequality structure our lives in ways that generate not just individual but systematic differences in what we are positioned to experience and to know, and in what resources we have for developing critical standpoints on systemic inequality and its effects (Wylie, 2012). Formulated in non-essentialist terms, this structural 'situated knowledge' thesis (Haraway, 1991) is the central tenet of contemporary standpoint theory and the basis for its claims about epistemic advantage on the margins.

The second line of criticism – that standpoint theorists accord automatic and comprehensive epistemic privilege to those who are socially marginal or subjugated – targets the 'inversion' thesis. The objection is that, in itself, the experience of oppression or marginality cannot be assumed to guarantee better, more reliable, and probative knowledge. Most obviously, the oppressed are often kept in their place in part by being denied access to key epistemic resources like literacy, education, and the information necessary to understand the systemic conditions responsible for their oppression (Narayan, 1988), while insider-outsiders may be loath to risk their tenuous status by challenging community norms and can be fierce defenders of the status quo. A strong reading of the 'inversion' thesis is, then, clearly untenable given the structural nature of the situated knowledge thesis that standpoint theorists take as their point of departure: comprehensive epistemic 'privilege' cannot be attributed automatically to those who are oppressed, just because they are oppressed, and standpoint theory cannot provide a cut-and-dried formula for identifying particular standpoints that can be counted on to produce more objective knowledge. It is something of an irony that this second criticism is leveled against standpoint theorists; the case standpoint theorists make for recognizing epistemic advantage on the margins presupposes a robust appreciation of all the ways in which systems of oppression impose significant epistemic deficits on those they disadvantage. Not surprisingly, they typically construe the metaphor of privilege in terms of possibilities for understanding; consequently, the social location of those who are systemically disadvantaged can also be a source of contingent and problem-specific epistemic advantage. As Uma Narayan (1988) develops this point, workers on a production line in a free trade zone have the experiential resources to understand much more clearly than plant owners and consumers exactly how profit is extracted from their labor, but they may not have the resources to grasp the structural conditions that determine the international flow of goods and capital, and keep their wages low.

The third objection, that relativism is inescapable, presupposes the first two objections and is presented as a *reductio ad absurdum* of standpoint theory as a whole. The criticism is that, unless standpoint theorists assume that social standing is, in itself, a legitimate basis for adjudicating knowledge claims,

they have no grounds for claiming that the insights of the oppressed should be taken seriously, much less preferred over those of the privileged. And if they do embrace a strong reading of the epistemic privilege thesis, an untenable relativism follows: standpoint theorists substitute an appeal to political values and interests for considerations of evidence and good reasons in the evaluation of knowledge claims, and by so doing they undermine ideals of objectivity. Moreover, as the response to charges of essentialism indicates, there are multiple forms and experiences of oppression, even within categories such as race, class, and gender, and thus multiple-related political values and interests.

This line of argument presupposes a conventional foundationalist epistemology according to which objectivity is equated with value-neutral, interest-free impartiality and it is assumed that if we give up the conviction that there are universal, trans-historical, and context-transcendent norms of justification, we will have no basis for comparative evaluation of divergent knowledge claims and a debilitating epistemic relativism will be inescapable. Standpoint theorists join a broad constituency of STS scholars and social epistemologists who have made a convincing case for various forms of contextualism (Knorr Cetina, 1981), and they share an appreciation that appeals to context-transcendent norms of justification not only are untenable but also obscure the contingency and interest-specificity of the norms we have come to accept as a measure of objectivity. The polarized opposition between an idealized 'view from nowhere' and self-undermining relativism by no means exhausts the possibilities for adjudicating epistemic credibility (Rolin, 2006). As critics and theorists of oppression, standpoint theorists clearly recognize that effective action aimed at changing oppressive social relations requires accurate, probative knowledge of these conditions, and they are acutely aware of the costs of being wrong. They are suspicious of illusions of certainty wherever they arise, but especially when such illusions mask the interests embodied in dominant modes of understanding. Rather than rejecting objectivity as a regulative ideal, standpoint theorists are more typically intent on reformulating it in terms that take into account the situated nature of scientific inquiry and its extant norms of justification. The self-identified radical social constructionist Donna Haraway joins standpoint theorists in emphasizing the need to take up the challenge of developing a "usable doctrine of objectivity," one that integrates both a recognition that all knowledges are situated and a "no-nonsense commitment to faithful accounts of a 'real' world" (1991: pp. 187–188).

In direct contrast to the relativism with which the theory is charged, standpoint theorists are notable for arguing that the resources of situated knowledge have the potential to significantly improve the reliability and integrity of the claims we ratify as knowledge; its partiality is not necessarily an epistemic liability. Harding, for example, advocates norms of 'strong objectivity' (1991, 1993), a requirement that the assessment of knowledge claims should take into account, not only the primary evidence and reasons presented in their support, but also the conditions under which they have been produced and ratified: the social, historical circumstances that have shaped the interests of practitioners, the institutional contexts in which they work, and the evolving norms of justification in terms of which these claims are warranted. The goal here is not wholesale rejection of authoritative sources of knowledge like the sciences but to calibrate our confidence in them and to determine how they might be improved, especially with respect to scope of applicability and broader impacts. In this spirit, standpoint theorists direct attention to the transformative impact that feminist, race- and class-based, and other such critiques have had on a range of sciences. They note that, time and again, it has been outsiders to the research establishment or 'insider-outsiders' (Collins, 1986) – practitioners who bring an underrepresented and sometimes an explicitly political angle of vision to bear – who have identified gaps and errors in research programs that privileged 'insiders' have persistently missed, raising the epistemic bar on a number of dimensions and opening up productive new lines of inquiry (Wylie, 2003). Although she identifies herself as a constructive empiricist, Helen Longino's proceduralist account of objectivity captures a number of the insights that are central to the lessons standpoint theorists draw from consideration of such cases (Intemann, 2010). Longino (2002) argues that the beliefs we should count as knowledge are those that have been subjected to critical scrutiny from 'multiple points of view'; mobilizing the resources of diversely situated knowers can counteract the effects of insularity and contribute to the breadth and robustness of scientific thinking.

As reformulated in response to critiques of essentialism, automatic privilege, and relativism, standpoint theory is best understood as a conceptual framework, an open-ended set of analytic resources, for addressing jointly descriptive and normative questions about the epistemic effects of oppression (Wylie, 2012). Its point of departure is a situated knowledge thesis, one that presumes the epistemic relevance of systemic conditions of inequality without specifying in advance what their relevance will be in a given context. It directs attention to asymmetries in epistemic circumstance that suggest that the oppressed may contingently accrue important epistemic advantages (not comprehensive or automatic epistemic privilege). In this it counters patterns of epistemic injustice that discount the distinctive experience, perspectives, and hermeneutic resources of those who are socially marginal (Fricker, 2007). And it provides a rationale for the thesis that genuinely inclusive scientific communities are more likely to recognize their own limitations and to produce more balanced and better warranted understandings of their subject domains.

The Domains of Standpoint Thinking in and about the Sciences

There are a number of contexts in which standpoint theory, conceived in these terms, is relevant to the practice of science and to understanding the sciences. We identify two broad domains in which feminist and other genres of standpoint theory have been or can be productively applied.

Content Critiques: From Problem Choice to Framework Assumptions

Contextual values and interests – ideology of various kinds – clearly affect the choice of topics given priority in the

sciences. For example, feminist critiques of sexism and androcentrism in the sciences, developed both by practitioners and STS scholars, have been instrumental in bringing gendered interests into sharp focus; interests that are reflected in and served by the research agendas pursued by a predominantly male research establishment in a wide range of social, life, and physical sciences. Standpoint theory figures both as a catalyst for these critiques and as a framework for understanding them.

Critiques of problem choice are best developed, and their standpoint theory implications most explicit, in the social sciences, where social relations are a direct subject of inquiry. In any number of contexts, researchers influenced by feminist, critical race, and postcolonial theory, as well as related perspectives, have drawn attention to ways in which research agendas in their fields take the attributes, activities and interests of privileged men as the norm, focusing inquiry on questions that concern them. In the case of feminist research programs that took shape in the 1960s and 1970s, critiques of androcentrism catalyzed what initially were 'remedial' interventions aimed at adding women and gender to existing research programs. The women entering anthropology in unprecedented numbers in the 1960s were quick to recognize the need for compensatory ethnography; the lack of attention to women as subjects meant that women's roles and relationships, and their distinctive subcultures and activities, were largely absent from ethnographic narratives. Early critics observed that this inattention reproduced 'dominant male systems of perception'; women's spheres and cultural understandings were considered only insofar as they figured in or were valued by the public languages and worlds of men (Ardener, 1975). In a classic example from history, Joan Kelly-Gadol (1977) objected that the fortunes of women had been left out of account in research on the Renaissance. Economists called into question estimates of national productivity that excluded consideration of the contributions made by women in the form of unpaid labor within the home or within family businesses (Waring, 1988). The standpoint theorist Patricia Hill Collins (1986, 2000[[1990]]) drew on her working-class Black experience to throw into relief the race and class, as well as gendered, presuppositions that pervasively structured sociological research on 'the family.' And, famously, the feminist psychologist Carol Gilligan (1982) challenged Lawrence Kohlberg's influential model of moral maturation on the grounds that it had been based entirely on interviews with boys. This commitment to redefine research priorities also was evident in a substantial shift of focus in primatology, where research had been structured by the assumption that "primate society revolves around males" (Strum and Fedigan, 2000: p. 5); here, a quite literal reorientation of field observation practice generated substantial new data on the roles and activities of female primates.

In all these contexts, critiques of problem choice quickly exposed gender bias in framework assumptions and methodological norms that proved to have much wider implications, calling into question not just localized cases of science gone wrong, but pervasive androcentrism in science as usual – indeed, in some of the best science on offer (Harding, 1986). The feminist anthropologists who turned their attention to the unstudied domestic spheres in which women operated soon called into question the underlying distinction between public and private; the problem with androcentric ethnographies was not just the absence of women but the imposition of ethnocentric categories presumed to be universal that did not begin to do justice to the diversity of gendered roles and relations they were encountering. The questions Kelly-Gadol raised about historical accounts of the Renaissance were intended from the start to challenge the periodization schemes in terms of which history had been written. The Renaissance, she argued, was anything but a period of cultural rebirth, viewed from the vantage point of women; moreover, the "pattern of relative loss for women" that she noted for this period was a direct consequence of its accomplishments and typical for periods of "so-called progressive change" (Kelly-Gadol, 1976: p. 810). Reflecting on women's history more generally, Peter Novick observes that "feminist perspectives in history were as relevant to such male activities as war and diplomacy as they were to realms in which women dominated" (Novick, 1988: p. 496). Likewise, historians, sociologists, and economists have all had to reassess the assumption that the 'household' could be taken for granted as a relatively fixed and uncontroversial unit of analysis. When they attended to gender roles and relations they found conflicts, differences in interest, and diverse social configurations that call into question not only what they had assumed about households, but about labor markets, 'human capital,' and the causes and effects of poverty (Collins, 2000[1990]). As controversial as Gilligan's positive account of a 'different voice' has been, her critique of Kohlberg was pivotal in destabilizing the Piagetian assumption that moral development must follow a single, universal trajectory. Finally, the new data from primatology have not only fundamentally changed our understanding of contemporary primates – bringing into focus the central role of females and the importance of social dynamics other than aggression, hierarchy, and competition – it has also reframed theorizing about human evolution.

This roster of examples could be expanded in any number of directions. As Novick argues for history, critical analysis arising from underrepresented, subdominant subject positions – as defined by race, gender, and class, colonial status and indigeneity, by ableist and heterosexist norms, to name a few – have to varying degrees transformed virtually every field of human, behavioral, and social sciences since the 1970s, throwing into sharp relief the interest and context-specificity of entrenched, seemingly neutral field-defining agendas and framework assumptions. In this respect, standpoint theorists join STS scholars in detailing the myriad ways in which the values that set the research agenda also permeate the research process. Given their central interest in conditions of oppression, standpoint theorists are centrally concerned to understand, as well, the role played by authoritative research in the construction and legitimation of the social world from which these values arise. They attend to what Ian Hacking (1999) refers to as 'looping' effects, by which scientists not only reproduce but participate in constituting and amplifying the social relations that they see in society; they ask not only whose interests are reflected in entrenched research programs, but who benefits and who is harmed.

Such critiques are by no means limited to the social sciences. Consider, for example, the objections raised by the

women's health movement of the 1960s and 1970s to the conventional focus of medical research on the ailments of men, and the presumption that male bodies and disease profiles could stand as the norm for the medical diagnosis and treatment of women. As Londa Schiebinger describes it, feminist activism was responsible for "a great awakening … of mainstream medicine to women's health concerns"; it directed attention, not only to neglected conditions that primarily afflict women, like breast cancer, but also put pressure on medical researchers to take account of ways in which women's health and disease profiles diverge from models based on studies of male subjects (Schiebinger, 1999: p. 113). Heart disease is one prominent example of a well-studied condition, the understanding of which was based until the late 1980s almost entirely on samples of men, even though it concerned the effects of hormones like estrogen. More generally, the practice of relying on all-male samples to test drug safety and efficacy has been a highly visible target of critique, resulting in regulations that now explicitly require the inclusion of female subjects. The AIDS/HIV activism that resulted in the redesign of antiretroviral drug trials is another well documented example in which outsiders to the research establishment played an instrumental role in exposing systematic bias in the direction of research effort (Epstein, 1996). In the domain of technology (*see* Gender and Technology: From Exclusion to Inclusion?), both intended and indirect effects have been the focus of critique. In fact, every field of inquiry has at least some instrumental uses, and many forms of applied knowledge have some relevance to the gendered, as well as race and class, structure of society. Topics within so-called 'pure' physics are connected to military uses, and are funded precisely because of those uses. Those who have the least opportunity to participate in benefits of high status of military engagement suffer disproportionately from the consequences of war.

Contemporary standpoint theory was in large part catalyzed by these interventions. Harding is explicit that the motivation for her early exploration of standpoint theory was the question, "why has the sex/gender system become visible only now?" (the title of Harding's, 1983, essay). She found it striking that on this and many other fronts the sciences had not proven to be self-correcting. Ingrained sexist and androcentric biases had persisted for decades, and the impetus to address these problems – in everything from social relations and sex difference to economic systems and biomedical conditions – was due to external stakeholders whose interests were not reflected in or served by conventional research programs, and to the situated, often explicitly political standpoints brought to bear by the growing number of women and other underrepresented groups who entered the academy in the 1960s and 1970s. But rather than compromising the integrity of the research process, the questions generated by their diverse perspectives often significantly raised the epistemic bar; they subjected entrenched assumptions to newly rigorous critical scrutiny and called for consideration of a wider range of interpretive or explanatory possibilities, new lines of evidence, and undervalued methodological strategies, in the process making substantial contributions to their disciplines' understandings of traditional as well as new subject matters.

This proliferation of standpoints poses a challenge to conventional norms of justification and conceptions of objectivity on at least two fronts: it throws into relief the partiality and interest-specificity of bodies of knowledge that had been presumed objective (in the value-neutral, interest-free sense); and it demonstrates the epistemic contributions that can arise from mobilizing explicitly situated perspectives. Standpoint theorists take up this challenge, articulating the conceptual framework for a research program in social epistemology and STS that responds to this crisis of confidence in conventional norms of justification. It is characterized by the articulation of a situated knowledge thesis that highlights the systemic conditions that configure our epistemic capabilities, by an emphasis on recognizing epistemic advantage on the margins, and by the proposal of alternative conceptions of objectivity. This last includes proceduralist accounts that focus attention on the jointly social and cognitive conditions that ensure the possibility of transformative criticism, and norms of strong objectivity that build an ongoing critical appraisal of the conditions of knowledge production into the adjudication of knowledge claims.

Equity Critiques and Collaborative Practice

To varying degrees, diverse forms of discrimination continue to be a normal part of the working environment in all of the sciences, embodied in everything from the allocation of funds and skewed citation patterns to patterns of workplace segregation, the structure of the tenure system, and the myriad effects of insider networks and hierarchical prestige systems. Equity activists, who, as standpoint theory predicts, are disproportionately members of underrepresented groups, vocally challenge these entrenched practices as unjust and unsupportable in an enterprise defined by meritocratic ideals. Increasingly, however, the impetus for documenting and counteracting discrimination within the sciences is not only a commitment to social justice but also, at an institutional and state level, a concern that the sciences cannot afford to squander trained talent by tolerating entrenched prejudice. Standpoint theorists extend this generic argument for equity, arguing that discrimination exacts more specific costs than a diminution of the pool of trained talent; it robs the sciences of key epistemic resources as well. A growing number of cases bring to light the costs of gender-based discrimination: feminist examples include key discoveries by women that were not recognized for decades, and patterns of workplace segregation and citation bias that conspire to marginalize the contributions from subfields or from disciplinary niches dominated by women. More generally, the recent history of content critiques outlined above draws attention to the pervasive risks of groupthink in socially homogenous, insular research communities. Standpoint theorists join equity activists in making the case for recruitment and training practices, and for reconfiguring reward structures that foster diversity, but they add an epistemic justification for these interventions. They make the case that, although the contributions of researchers drawn from diverse backgrounds cannot be specified in advance, they can be counted on to expand the range of experience – to develop novel heuristics and perspectives on which the scientific community can draw when subjecting its norms of justification and institutional practices, as well as its

best theories, to critical scrutiny. There is ample evidence coming from a cross-section of the sciences – especially the social and behavioral sciences – that this kind of diversity is necessary to ensure the integrity and reliability of the knowledge ratified as scientific.

On this account, standpoint theory is not only a critical practice but also a source of normative insight about strategies for improving scientific inquiry, making it more robust, relevant, and accountable. Dorothy Smith has long insisted that standpoint theory is, above all, a methodology predicated on a commitment to level the hierarchical relationship between researchers and subjects, recognizing the wisdom of those of us who are insiders to the contexts studied by social scientists. This is the principle that underpins the recommendation to "start research from the margins" (Smith, 1974; Harding, 1991). A similar case for recognizing relevant expertise held by nonscientists is also made in a range of biological, medical, and physical sciences. Brian Wynne (1996) details what sheep farmers know about the ecological impact of nuclear fallout that scientists miss altogether in the context of an STS analysis of expertise, and any number of citizen science initiatives, DIY-research communities, and the medical research activists discussed earlier bear witness to the expertise that lay practitioners have to offer fields as diverse as astronomy, genomics, infectious disease research, and field biology.

These principles of recognizing and engaging various forms of expertise that lie outside professional science are perhaps best exemplified by the longstanding traditions of CBPR and PAR. The ideal here is that researchers should work in partnership with nonscientific communities, collaboratively defining the problems they will address and drawing on the expertise of community partners at every stage of the research program. The sociologists Elizabeth Petras and Douglas Porpora (1993) make explicit the standpoint theory rationale for such practice, arguing that PAR practice is motivated not only by a political, ethical commitment to be responsive to and to treat local communities respectfully, but by a recognition that the poor and oppressed bring significant epistemic resources to the table. They are 'epistemically privileged' in ways that can be expected to 'generate knowledge and insights' on a number of dimensions – a "spectrum of theory, methods and substance" (Petras and Porpora, 1993: p. 107) – that represent significant contributions to disciplinary social science. The history of such research in its various forms is detailed with particular clarity in the context of recent debate about CBPR practice in development contexts (Hickey and Mohan, 2004).

Conclusions

Research in STS brings into sharp focus the local and contextual origins of scientific knowledge. However universal the decontextualized results of inquiry may seem to be, they carry with them parochial assumptions, distortions and lacunae that arise from a particular social order. Standpoint theorists take up a particular set of questions and challenges posed by these insights. They engage in a practice of science critique aimed at countering the erasure of the situated interests that shape the sciences, and they are particularly attentive to the role played by dominant systems of knowledge in reproducing and amplifying systems of oppression. At the same time, standpoint theorists are intent on improving scientific practice. They draw a number of critical and constructive lessons from critiques that demonstrate that insular scientific communities are not reliably self-correcting; it often takes the input of 'insider-outsiders' who bring to bear dissonant experience and a critical angle of vision to identify persistent errors and missed possibilities in inquiry. Standpoint theorists make the case for a modest 'inversion thesis,' arguing that a range of epistemic advantages may accrue to those who must navigate the constraints of an oppressive social order from a position of systemic disadvantage. And, on the basis of that thesis, they articulate an epistemic rationale for fostering diverse, representative, and democratic scientific communities; communities in which the integrity and applicability of scientific knowledge requires the input of diverse critical perspectives, including those of nonscientists. This framework for critique and constructive intervention is potentially relevant to all branches of science, inasmuch as scientific inquiry is a thoroughly social process and perspectives that bring into view the effects of social structure are thus valuable throughout the sciences.

See also: Constructivism and Constructionism: Methodology; Feminism in Anthropology; Feminist Research and Social Work; Gender and Technology: From Exclusion to Inclusion?; Gender and Women's Studies, Applied Research On; Gender, Feminism, and Sexuality in Archaeological Studies; Human Sciences, History of; Ideological Constraint: History and Current Status of the Concept; Ideology, Sociology of; Participatory Action Research in Environmental and Ecological Studies; Queer Theory and Intersectionality; Science, Sociology of; Situated Knowledge, Feminist and Science and Technology Studies Perspectives; Social Constructivism.

Bibliography

Alcoff, L.M., 2006. Visible Identities: Race, Gender, and the Self. Oxford University Press, Oxford.

Ardener, S. (Ed.), 1975. Perceiving Women. J.M. Dent & Sons, London.

Collins, P.H., 1986. Learning from the outsider within: the sociological significance of black feminist thought. Social Problems 33, 14–32.

Collins, P.H., 2000[1990]. Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge, New York.

Epstein, S.G., 1996. Impure Science: Aids, Activism, and the Politics of Knowledge. University of California Press, Berkeley, CA.

Fricker, M., 2007. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford.

Gilligan, C., 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Hacking, I., 1999. The Social Construction of What? Harvard University Press, Cambridge, MA.

Harding, S., 1983. Why has the sex/gender system become visible only now? In: Harding, S., Hintikka, M. (Eds.), Discovering Reality. Reidel, Dordrecht, pp. 311–324.

Harding, S., 1986. The Science Question in Feminism. Cornell University Press, Ithaca, NY.

Harding, S., 1991. Whose Science? Whose Knowledge?: Thinking from Women's Lives. Cornell University Press, Ithaca, NY.

Harding, S., 1993. Rethinking standpoint epistemology: what is "strong objectivity"? In: Alcoff, L.M., Potter, E. (Eds.), Feminist Epistemologies. Routledge, New York, pp. 49–82.

Haraway, D., 1991. Situated knowledges: the science question in feminism and the privilege of partial perspective. In: Haraway, D. (Ed.), Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York, pp. 193–202.

Hartsock, N.C.M., 1983. The feminist standpoint: developing a ground for a specifically feminist historical materialism. In: Harding, S., Hintikka, M. (Eds.), Discovering Reality. Reidel, Dordrecht, pp. 283–310.

Hekman, S., 1997. Truth and method: feminist standpoint theory revisited. Signs 22, 341–365 [with Comments and Reply, 366–402].

Hennessy, R., 1993. Women's lives/feminist knowledge: feminist standpoint as ideology critique. Hypatia 8 (1), 14–34.

Hickey, S., Mohan, G. (Eds.), 2004. Participation: From Tyranny to Transformation? Exploring New Approaches to Participation in Development. Zed Books, New York.

Intemann, K., 2010. 25 years of feminist empiricism and standpoint theory: where are we now? Hypatia 25, 778–796.

Kelly-Gadol, J., 1976. The social relation of the sexes: methodological implications of women's history. Signs 1, 809–823.

Kelly-Gadol, J., 1977. Did women have a renaissance? In: Gridenthal, R., Koonz, C. (Eds.), Becoming Visible: Women in European History. Houghton Mifflin, Boston, MA, pp. 137–164.

Knorr Cetina, K., 1981. The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science. Pergamon, Oxford, UK.

Longino, H., 2002. The Fate of Knowledge. Princeton University Press, Princeton, NJ.

Lukács, G., 1971. History and Class Consciousness; Studies in Marxist Dialectics. MIT Press, Cambridge, MA.

Moya, P.M.L., Hames-García, M.R., 2000. Reclaiming Identity: Realist Theory and the Predicament of Postmodernism. University of California Press, Berkeley, CA.

Narayan, U., 1988. Working together across difference. Hypatia 32, 31–48.

Novick, P., 1988. That Noble Dream: The "Objectivity Question" and the American Historical Profession. Cambridge University Press, Cambridge.

Petras, E.M., Porpora, D.V., 1993. Participatory research: three models and an analysis. The American Sociologist 23, 107–126.

Rolin, K., 2006. The bias paradox in feminist standpoint epistemology. Episteme 3, 125–136.

Schiebinger, L., 1999. Has Feminism Changed Science? Harvard University Press, Cambridge, MA.

Smith, D., 1974. Women's perspective as a radical critique of sociology. Sociological Inquiry 44, 7–13.

Smith, D., 1987. The Everyday World as Problematic: A Feminist Sociology. Northeastern University Press, Boston, MA.

Strum, S.C., Fedigan, L.M. (Eds.), 2000. Primate Encounters: Models of Science, Gender, and Society. University of Chicago Press, Chicago, IL.

Waring, M., 1988. If Women Counted: A New Feminist Economics. Harper & Row, San Francisco, CA.

Wynne, B., 1996. May the sheep safely graze: a reflexive view of the expert-lay knowledge divide. In: Lash, S., Szerszynski, B., Wynne, B. (Eds.), Risk, Environment and Modernity. Sage, London, pp. 44–83.

Wylie, A., 2003. Why standpoint matters. In: Figuero, R., Harding, S. (Eds.), Science and Other Cultures: Issues in Philosophies of Science and Technology. Routledge, New York, pp. 26–48.

Wylie, A., 2012. Feminist philosophy of science: standpoint matters. Proceedings and Addresses of the American Philosophical Association 86, 47–76.

Startle Reflex and Health

Terry D Blumenthal, Wake Forest University, Winston-Salem, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The startle reflex is a response to sudden and intense stimulation in auditory, visual, or tactile modalities and is characterized by a variety of muscular contractions and autonomic responses. The response is most often measured in humans with the motor potential (electromyography) of the eyeblink reflex. This startle eyeblink response is preserved across the life span and is analogous to whole-body startle responses in other animals. Startle is very sensitive to small variations in eliciting stimulus parameters; task situations; personality characteristics; clinical conditions; psychopharmacological agents and their history of use; conditioning; and a variety of other stimulus, person, and situational factors. Startle has been used to investigate many research questions dealing with information processing deficits in a range of health-related conditions.

Introduction

The startle response is most often considered to be a defensive reflex, with aspects of protection, escape, and interruption, elicited by sudden intense stimulation in several sensory modalities. Startle reactivity varies as a function of the eliciting stimulus parameters, but it also varies based on personality and clinical characteristics of the person, as well as task and situational demands. This makes the startle response a very sensitive measure of automatic and intentional processes at rest or during task performance in normal control participants or difficult-to-test populations. As such, it is useful in evaluating information processing in a wide range of situations. With regard to health, startle is most often used to investigate clinical or subclinical conditions involving stress, anxiety, or information processing deficits, although the application of startle methodology is advantageous in a wide range of conditions, as will be elaborated below. Since startle is affected by attention, emotion, arousal, pharmacological changes, and a host of other factors, it can be used to assess variations in those factors.

The Startle Response and Its Modification

The Ubiquity of Startle

Startle is a particularly useful research tool, since this response can be measured across species, across ages within species, and across states within individuals. The presentation of a sudden, intense stimulus will generally result in very rapid initiation of motor activity that interrupts ongoing behavior and serves to either facilitate escape or activate defensive responses. In mammals (including humans), the startle response often involves crouching and limb flexion to protect the back of the neck and rapid closure of the eyelids to protect the eye (Fendt et al., 2001). The same responses can be measured in humans and lower animals, allowing for the use of animal models to investigate human conditions and for the investigation of the physiology and biochemistry underlying these conditions (Eaton, 1984; Davis, 2006).

Startle is also found across the life span and can be measured in the same way from the newborn to the elderly. This allows researchers to use the same response measure to investigate developmental changes in those factors that affect startle responding such as emotion and task performance. Startle is also found in both awake and asleep individuals, increasing its usefulness in studying automatic processing in difficult-to-test populations such as some clinical patients, infants, etc. (Blumenthal et al., 1987).

Another advantage of startle is the fact that it is relatively automatic, thereby avoiding the problem of demand characteristics that can contaminate self-report measures. The very fast latency of this response (less than 50 ms for the eyeblink component) means that the response has already occurred by the time the participant is consciously aware that a startle stimulus has been presented.

Elicitation and Measurement of Startle

The most common way to elicit startle is by the presentation of a sudden and intense sound, although reliable startle can also be obtained with flashes of light, sudden vibration, puffs of air to the face, a tap to the forehead, or an electrical stimulation of the trigeminal nerve or peripheral nerves (Blumenthal et al., 2005). Although startle reactivity can be measured from skeletal muscles and muscles in the neck, the most widely used measure of human startle is the eyeblink response, a measure of the electromyographic (EMG) activity in orbicularis oculi, the muscle that causes the eyelids to close. Eyeblink EMG is the preferred response measure because it is most sensitive (available at lower stimulus intensities), it habituates more slowly and less completely than other motor components of startle, and it is one of the earliest components of startle to appear after stimulus onset. Another important consideration is the very high signal-to-noise ratio of this eyeblink EMG response; since orbicularis oculi does not have muscle spindles, it generates very low EMG at rest relative to the very large response elicited by a startle stimulus. This allows for the reliable measurement of startle responses on single trials, repeated many times in a session, with no need for averaging in order to see the response (Blumenthal et al., 2004). This is also an easy response to measure, requiring the placement of just two surface recording electrodes on the face below the eye, with a ground electrode anywhere else on the body. This electrode preparation can be used across the life span, from newborns to the elderly (Blumenthal et al., 2005).

The startle eyeblink response measured from orbicularis oculi is driven by the facial nerve (CN VII), with input to the motor nucleus of CN VII from the startle center in the brain

stem, the nucleus reticularis pontis caudalis (nRPC) (Davis et al., 1999). The nRPC receives input from auditory, visual, and tactile stimulation, as well as from many other areas of the brain. These inputs can modulate the startle response, which means that startle can be used to index activity in those various brain areas.

Prepulse Inhibition of Startle

In addition to startle being a sensitive measure in and of itself, it can also be used to investigate other processes that are ongoing when the startle stimulus is presented. For example, if the startle stimulus is presented within 30–800 ms after a weak stimulus, called a prepulse or lead stimulus, the startle response will often be smaller than that elicited without that weak preceding stimulus (Blumenthal, 1999; Graham, 1975). This reduction in startle is called prepulse inhibition (PPI) of startle and occurs for prepulses that are as weak as the psychophysical detection threshold and for prepulses that are in any sensory modality (Blumenthal and Gescheider, 1987; Hackley, 1999). This PPI is seen on the first trial on which the prepulse precedes the startle stimulus, even if the prepulse has been presented dozens of times alone before that first pairing, illustrating the fact that this is not a function of learning or warning (Blumenthal, 1997). PPI is seen with presentation of discrete prepulses, but also with gaps in an ongoing background sound (Fournier and Hebert, 2013) or with frequency shifts in a background sound. The degree of PPI is sensitive to various properties of the prepulses, such as intensity, duration, rise time, and bandwidth, which means that PPI might be useful in assessing the processing of these various prepulse properties (Blumenthal, 1999). PPI is also sensitive to a wide variety of individual differences, such as gender and menstrual cycle phase (Jovanovic et al., 2004), and a variety of clinical disorders and personality differences (see below).

Attentional Modulation of Startle

Just as PPI can indicate prepulse processing, as described above, startle reactivity can also indicate direction of attentional resources. PPI occurs when prepulses and startle stimuli are in the same or different modalities, within specific time windows (e.g., 30–800 ms). At longer time intervals, as occurs in some target detection tasks, a startle stimulus occurring in the sensory modality that is the current focus of attention, as in the presentation of an acoustic startle stimulus during an auditory vigilance task, will tend to elicit a larger-magnitude startle response relative to that in a no-task condition (Hackley, 1999). If the task stimuli and the startle stimuli are presented in different sensory modalities, the startle response tends to be smaller (Schicatano and Blumenthal, 1998). The fact that startle reactivity is affected by attentional direction has resulted in startle being useful in assessing the deployment of attentional resources (Hackley, 1999).

Modulation of Startle by Affectively Valenced Stimuli

Just as startle has been used to probe background cognitive processing and direction of attention, it has also been applied to the investigation of emotion. This most often involves a procedure that is referred to as affective modulation of startle (Lang and Bradley, 2013). Startle elicited during the processing of highly arousing unpleasant stimuli is often larger than when neutral or pleasant stimuli are being processed, since both the startle stimulus and the background valence 'match' (both are unpleasant). When the background is pleasant (a valence 'mismatch'), startle is often smaller than that in the neutral background condition. This valence match–mismatch hypothesis has led to a great deal of research in which the defensive startle response is used to index the degree of unpleasantness of whatever background stimulus is being processed when the startle stimulus is presented. This has most commonly involved processing of visual background stimuli (pictures, objects), but has also been demonstrated during the processing of unpleasant sounds and odors, as well as during the imagining of pleasant and unpleasant situations (Bradley et al., 1999).

The relationship between startle reactivity and emotion can also be considered in research on fear and anxiety (Davis, 2006). Pairing a neutral stimulus with a painful stimulus (e.g., electric shock) in a classical conditioning paradigm can result in that neutral stimulus becoming a conditional fear-inducing stimulus. Pairing that conditional stimulus with a startle stimulus can result in larger startle during the subsequent presentation of the conditional stimulus, a result referred to as fear-potentiated startle. The context in which the fear conditioning occurred can itself cause an increase in startle, an example of anxiety-potentiated startle (Grillon and Baas, 2003). These fear- and anxiety-potentiated startle techniques will be discussed in more detail below.

Clinical Conditions

There are several conditions in which abnormal startle reactivity may be an indication of some underlying problem. This abnormality may appear as either hyperreactivity or hyporeactivity to startle stimuli. Abnormalities of startle modulation can also be used to investigate a variety of disorders. We must also keep in mind the fact that 'abnormal startle' can be difficult to determine, since 'normal' startle reactivity, that seen in individuals selected for the absence of any neurological or psychiatric condition, can vary by as much as a factor of 40 across individuals within a study (Blumenthal et al., 2005). Therefore, the distributions of level of PPI in patients and controls, for example, often show considerable overlap.

Startle Syndromes

Hyperekplexia

The most extreme example of excessive startle reactivity is hyperekplexia, involving both larger-magnitude responses and shorter latencies of responding (see Bakker et al. (2006) for an excellent review of startle syndromes). This term is often used to refer to exaggerated startle, although it has been used more specifically to refer to exaggerated startle that is based on a hereditary disorder characterized by abnormal genes that code for a particular glycine receptor. Major hyperekplexia involves both excessive startle responding and muscular stiffness in newborns, whereas the minor form involves only

hyperexcitable startle responding, without neonatal stiffness, and appears later than the major form. Major hyperekplexia also often involves rapid retraction of the head when the nose or upper lip is lightly tapped, and this seems coherent with the defensive retraction aspects of the startle response. Less well-defined cases of hyperekplexia often fall into the category of either 'sporadic' or 'symptomatic' hyperekplexia (see Bakker et al., 2006). As with many neurological disorders, hyperekplexia is characterized by a wide range of symptoms that are more or less pronounced across individuals, appearing at various stages of development, all of which makes labeling or categorizing these disorders difficult. The one common symptom of the hyperekplexia-related conditions is exaggerated startle reactivity. The motor aspects of hyperekplexia are most effectively treated with clonazepam, a GABA receptor potentiator that compensates for the deficit caused by mutations in glycine receptors and dampens startle reactivity.

Culture-Specific Startle Disorders

Bakker et al. (2013) describe a number of hyperstartling conditions that appear to be specific to certain cultures or locations, the most well-known of which is found in Indonesia, in a hyperreactive startle condition referred to as Latah. In addition to larger startle responses, these conditions often involve a variety of other behaviors (other motor behaviors, vocalizations, echolalia, forced obedience, etc.), many of which are stereotypical and similar across cultures. Bakker et al. (2013) compared eyeblink EMG measures of startle in Latah patients with those of matched control participants and found that this measure of startle was larger and more probable in the patients, but no differences in onset latency or habituation were found. However, startle components in the neck and body muscles were more probable and faster in patients, and they habituated more slowly. Bakker et al. report that startle reactivity was greater than normal in these Latah patients but not as pronounced as would be seen in hyperekplexia. Part of the increased startle reactivity attributed to Latah may be due to the higher rate of anxiety disorders in the patients (see below).

Startle Stimulus-Induced Disorders

Bakker et al. (2006) summarize work describing several disorders that can be triggered by exposure to a startle-eliciting stimulus such as epileptic seizures, reflex myoclonus, and startle-induced stiffness. Reflex myoclonus is in some ways similar in appearance to hyperekplexia but tends to be localized to specific muscle groups. Contraposed to startle-induced stiffness, patients with cataplexy may lose muscle tone when exposed to startling stimuli. Two points to take away from this discussion are (1) that abnormal startle reactivity can be seen in a wide variety of disorders and (2) that sudden, intense stimuli, such as those used to elicit startle, can also elicit a variety of other changes in ongoing motor activity, some of which constitute the expected startle motor response, but others of which are less coherent with the traditional startle complex.

Schizophrenia

The human disorder to which startle research has been most often applied is schizophrenia. Specifically, there is strong evidence for a deficiency of PPI in schizophrenics, as well as in the first-degree relatives of schizophrenics and in people who are subclinical for schizophrenia-like symptomology (Takahashi et al., 2011). PPI is often considered to be an operational index of sensorimotor gating (Swerdlow et al., 2008). When patients are compared with normal controls, PPI is often, but not always, less pronounced in the patients, which is thought to reflect a failure of sensorimotor gating. This gating failure is believed to be related to, or indicative of, other failures of gating in these patients such as those that result in cognitive fragmentation and flooding (Braff et al., 2001).

Much recent research has shown that PPI deficits are not limited to schizophrenia but are found in a variety of other clinical conditions (e.g., Tourette's disorder, Huntington's disease, obsessive-compulsive disorder) as well as nonclinical individual differences (e.g., as a function of gender, menstrual cycle, personality characteristics, etc.; Swerdlow et al., 2008). This finding of PPI deficits in other disorders is in part based on the fact that similar symptoms (i.e., gating deficits) may appear across a variety of disorders. Orthogonally, not all schizophrenics show PPI deficits, suggesting that this indicator of gating deficiency may be specific to (and therefore informative of) a subset of this spectrum of disorders. Also, successful treatment of the clinical symptoms is not always paired with recovery of PPI. To summarize, abnormal PPI of startle cannot be used to diagnose any disorder in the schizophrenia spectrum, but this measure, both in humans and in animal models, may be useful in characterizing some of the neurological mechanisms underlying a variety of disorders and conditions.

Anxiety Disorders

With regard to anxiety disorders, startle may be useful in distinguishing between the concepts of fear and anxiety, with fear being more stimulus-specific, short-term, and signaling imminent danger, whereas anxiety involves a more tonic, contextually triggered state. Davis et al. (2010) summarize research that suggests that these two conditions may be based on somewhat different underlying neurological mechanisms, dissociable by measuring startle reactivity in specific circumstances (e.g., a light paired with footshock might come to elicit fear-potentiated startle upon subsequent presentations, but simply being placed in the testing chamber can elicit context-potentiated startle in the rat). Findings analogous to those from animal research investigating fear- and context-potentiated startle (Davis, 2006) have also been shown in human research, with increased startle magnitude in the presence of painful stimuli or conditional stimuli that have been paired with painful stimuli (Lang and McTeague, 2009). As an example of the fact that this parallel is functional but not literal, startle reactivity is potentiated when rats are exposed to light (Davis, 2006) and when humans are exposed to darkness (Grillon et al., 1997), situations that are anxiety-provoking in each species.

McTeague and Lang (2012) review literature dealing with startle reactivity in a variety of anxiety disorders. In the case of posttraumatic stress disorder (PTSD), for example, 'exaggerated startle' has been considered one of many potential diagnostic symptoms, although the definition of 'startle' in this context does not generally depend on the detailed measures of EMG activity described here. However, exaggerated startle is more likely to be seen shortly after a traumatic event, and it is more

likely in more severe cases of PTSD. Also, several researchers have shown that even when baseline startle is no greater in PTSD patients than controls, startle elicited in the presence of trauma-related ('trigger') stimuli is often significantly greater in patients (Grillon and Baas, 2003; McTeague and Lang, 2012). Analogous differences in startle in the presence of 'trigger' stimuli can be seen in patients with eating disorders or in abstaining drug addicts (see below). Also, PPI of startle has been found to be impaired in unmedicated panic patients (Ludewig et al., 2005), obsessive-compulsive disorder patients (Hoenig et al., 2005), and individuals with high trait anxiety (Franklin et al., 2009a).

The affective startle modulation paradigm described above has been used to illustrate the power of phobia-related stimuli to potentiate startle in phobic patients (McTeague et al., 2009). Unpleasant pictures tend to cause an elevation of startle in most people, but phobia-related pictures further increase startle only in people reporting the phobia that is depicted in the picture (animals, blood, etc.). This increased startle reactivity is subsequently decreased back to normal levels after behavioral therapy to treat the phobia (Vrana et al., 1992), suggesting that startle and phobic symptoms may be related and that both may be affected by some forms of therapy.

Stress

Acute stress causes an increase in the activity of the hypothalamus–pituitary–adrenal gland (HPA) axis, which results in a release of cortisol into the bloodstream. One action of cortisol is to dampen further HPA activity via a negative feedback loop. However, repeated exposure to cortisol can also damage neurons in the hippocampus, which may be at the base of some forms of anxiety. Roemer et al. (2009) and others have shown that increasing cortisol increases startle reactivity in humans.

Grillon et al. (2006) showed that a high dose of orally administered cortisol was related to larger fear-potentiated startle responding, whereas Grillon et al. (2011) showed that cortisol increased anxiety but not fear. Also, this exaggerated fear response is attenuated by pharmacological suppression of cortisol in PTSD patients (Jovanovic et al., 2011). However, Miller et al. (2011) report decreased, rather than increased, startle after oral administration of hydrocortisone, whereas Richter et al. (2011) reported rapid disruption of PPI, with no effect on control startle reactivity after injection of cortisol. These studies, among others, suggest that the impact of cortisol on startle and PPI is affected by the dose, the method of administration (injection, oral, intranasal), the characteristics of the participant, and the timing of the measurement of startle relative to administration of the cortisol. To the extent that cortisol is used to model stress, these situational and constitutional factors may also moderate the degree to which stress is reflected in variations in startle reactivity.

Huggenberger et al. (2013) showed that the development of PPI in young infants may be related to maternal stress. They found that the development of PPI of acoustic startle in the first 4 months after birth was attenuated in babies whose mothers reported higher levels of social stress. This may be because the chronic stress of the mother affected the development of the infant prenatally, with PPI changes from birth to 4 months of age indexing the early development of information processing systems.

Stress also has effects on cardiovascular functioning, and some of these relationships have been investigated with the startle response. Schulz et al. (2011) found that the arterial baroreflex, a regulatory feedback loop that relates heart rate to blood pressure via the vagus nerve, can attenuate startle reactivity if the timings of baroreflex activation and startle stimulus presentation are optimal. If stress alters peripheral vasoconstriction and shortens the preejection period (Schaechinger et al., 2001), these effects may be seen in the timing at which baroreflex activity modulates startle responding. In fact, exposing participants to cold pressor stress (by immersing the hand in ice water) has exactly this effect, shifting the optimal interval for baroreflex modulation of startle. This suggests that startle might be useful in identifying individuals who will be more strongly affected by particular stressors. Schulz et al. (2011) also suggest that the cardiac modulation of startle and the impact of stress on this system may be useful in the investigation of psychosomatic interoceptive symptoms, such as heartbeat detection, due to the involvement of common neurological mechanisms in both situations.

Depression

Affective blunting is common in unipolar depression, and these patients also show less-pronounced startle modulation by affectively valenced background pictures (Dichter and Tomarken, 2008). More specifically, Dichter and Tomarken found that when the startle stimulus followed picture onset by 3.5–4.5 s (the temporal relationship that has been shown to optimize affective startle modulation), startle potentiation by unpleasant background stimuli was weak in depressed individuals, and startle attenuation by pleasant pictures was absent in these patients. Dichter et al. (2004) found that the blunted affective modulation of startle is still present after successfully treating depressive symptoms with antidepressant medication. Another study found that the extent to which affective startle modulation is seen is inversely related to the severity of the depressive symptoms (Kaviani et al., 2004). The fact that startle modulation is less pronounced in more depressed patients, even after they have been treated with antidepressants, suggests that startle might be a marker for symptom severity that is not affected by successful treatment of those depressive symptoms. Finally, Vaidyanathan et al. (2014) found that affective modulation of startle was only abnormal in participants who had recurrent depression, being normal in those who had experienced only one depressive episode.

Eating Disorders

The affective startle modulation method has been used to evaluate the emotional impact of food-related stimuli in patients with eating disorders, as well as in nonclinical participants with tendencies toward such disorders. One might expect food cues to be perceived as positively valenced, suggesting that startle should be reduced in their presence. However, the impact of food cues on startle is affected by degree of food deprivation, eating disorder symptomology, and whether the food is perceived as being imminently

available or unavailable. Food deprivation has been found to result in increased startle in the presence of food-related pictures, perhaps due to the induction of an unpleasant frustrative nonreward condition (Drobes et al., 2001). When actual food items, rather than pictures of food, are presented, startle can be decreased in their presence (Hawk et al., 2004), unless the participant is told that the food item, though present in the testing room, would not be available for consumption until several hours later (Rejeski et al., 2010). Startle has also been used to investigate emotional reactivity to body image information, both in pictures (Spresser et al., 2012) and in words (Herbert et al., 2013).

Addiction

Many researchers have used the startle response and its modulation as a way to evaluate the impact of drugs of addiction and stimuli related to these substances. Startle and its modulation have been used to investigate the effects of withdrawal following abstinence, the emotional impact of drug-related cues, and the impact of drugs on task performance. These lines of research have utilized startle reactivity, PPI, and affective modulation of startle as tools for the investigation of drug-related effects associated with caffeine (Flaten and Blumenthal, 1999), tobacco (Greenstein and Kassel, 2010), alcohol (Rubio et al., 2013), heroin (Walter et al., 2011), and cannabis (Mathias et al., 2011). Many of these effects are complicated by such factors as previous level of use, duration of withdrawal, task demands, conditioning history, expectancy, and polydrug use.

Sensory Deficits

One of the great advantages of the startle response is the fact that it can be elicited by stimuli in several modalities. Inputs from visual, auditory, and tactile sensory systems all project to the pontine startle center; therefore, startle can be elicited by sudden and intense flashes of light, bursts of sound, puffs of air, electrical shocks to the face or arm, or taps to the face (Blumenthal et al., 2005). These various inputs interact, summing to yield larger startle when they are presented close together in time. The multimodal sensitivity of startle allows for its use even when one sensory system is impaired, for example, visual startle elicitation in hearing-impaired participants. The fact that startle is affected by a variety of stimulus parameters (e.g., intensity, duration, bandwidth) means that variations in startle might be useful as indicators of variations in sensory processing. This suggests that startle might be useful as a tool for identifying sensory deficits in nonverbal or difficult-to-test populations such as infants (Blumenthal et al., 1987). Even in compliant adults, methods more objective than self-report may be useful in diagnosing and delineating sensory processing deficits. For example, Fournier and Hebert (2013) report that tinnitus can be objectively identified by the application of PPI of startle methodology, wherein the 'prepulse' is a gap in a background sound. In control participants, such a gap inhibits the startle response to a subsequent intense stimulus, but this inhibition is weaker in tinnitus patients, presumably because the tinnitus 'fills in' the gap in the background sound.

Brain Damage

Honeycutt and Perreault (2012) found that startle is intact in patients who have had cortical strokes. In fact, startle responses are often slightly larger in stroke patients than in control participants, perhaps due to the removal of descending tonic inhibition of startle from the frontal lobes. Honeycutt and Perreault used startle stimuli to modulate movement in the dominant arm that was paralyzed due to the stroke. The startle response in this limb appeared as a flexion response, and it was suggested that startle-activated flexion might be useful to the patient in situations in which that flexion is the intended or appropriate movement. However, the startle-activated flexion tended to interfere with a planned extension of the limb in the same patients, although this interference decreased across trials as the startle flexion response habituated.

Williams and Wood (2012) report on affective modulation of startle in traumatic brain injury (TBI) patients and matched controls. Both groups showed the expected startle attenuation in the presence of pleasant pictures, but TBI patients failed to demonstrate the expected potentiation of startle when viewing unpleasant pictures. Finally, Ueki et al. (2006) report that patients with amnestic mild cognitive impairment showed more PPI of acoustically elicited eyeblink startle responses, whereas those diagnosed with Alzheimer's disease showed deficient PPI to the same stimuli.

Many studies have measured startle and its modulation by prepulses and affective stimuli in patients with a variety of neurological and cognitive impairments. The findings vary widely, perhaps based on the rostrocaudal level of the damage and the extent of that damage, differentially impacting the more-or-less or preattentive or attentive processing involved in the particular startle application used.

Nonclinical Individual Differences

Startle and its modulation have been used to investigate a wide range of personality characteristics (see Blumenthal and Franklin, 2009), only a few of which will be mentioned here. PPI has been found to be deficient in nonclinical individuals who show antisocial or borderline personality disorder symptomology (Franklin et al., 2009b), nonsuicidal self-injury (Franklin et al., 2010), and autism spectrum symptomology (Perry et al., 2007). Using the affective modulation of startle paradigm, Hazlett et al. (2007) found greater reactivity in the presence of unpleasant pictures in borderline personality disorder patients than controls.

Startle and Well-Being

Most startle research has focused on its ability to illustrate automatic processing in conditions of disorder or ill health. However, the fact that startle is a defensive response suggests that it might also be a useful measure in studies of adaptability and wellness. The general theme would be that a decrease in startle reactivity might indicate lower activation of emergency systems or reduced reaction to stress. This would be dependent on the reliability of startle in a particular context. Startle may also be useful in indexing the impact of various therapeutic

interventions: If startle decreases after intervention, then the intervention may be useful as a means to decrease stress, HPA axis activity, etc., with startle being an index of usefulness that is independent of self-report bias. A pre–post shift in startle reactivity would suggest that the technique has been effective at some level in decreasing defensive reactivity. Of course, not all successful therapeutic interventions will be paired with changes in startle or its modulation, but this is still a topic worthy of further investigation.

Emotion regulation can involve either suppressing or enhancing emotional processing and responding (Gross and Thompson, 2007). Enhancement of negative or unpleasant emotional responding would be expected to increase startle, especially if the emotion is fear or anxiety, as described above. Startle would also be increased following any manipulation that either increases arousal or focuses attention onto the startle stimulus. For example, Kiken and Shook (2011) report that mindfulness training can reduce negativity bias, a tendency to perceive danger in neutral situations. Therefore, we might expect mindfulness training to also reduce the startle response if arousal and attention can be equalized across conditions.

Conclusion

The startle response has proved to be a very sensitive, relatively noninvasive, and reliable way to quantify information processing, attention, arousal, and emotion, both across individuals and groups and within individuals across time. If startle varies as a function of some clinical condition, intervention method, personality factor, or other parameter, then startle might be useful in investigating that condition, method, factor, or parameter. This has been shown with clinical and subclinical conditions of ill health and may, in the future, prove to be just as true in cases of positive growth and well-being.

See also: Behavioral Medicine; Cardiovascular Reactivity and Recovery; Interoception; Mental Health; Personality and Health; Stress and Cardiac Response.

Bibliography

Bakker, M.J., van Dijk, J.G., van den Maagdenberg, A., Tijssen, M.A.J., 2006. Startle syndromes. Lancet Neurology 5, 513–524.

Bakker, M.J., van Dijk, J.G., Pramono, A., Sutarni, S., Tijssen, M.A.J., 2013. Latah: an Indonesian startle syndrome. Movement Disorders 28, 370–379.

Blumenthal, T.D., 1997. Prepulse inhibition decreases as startle reactivity habituates. Psychophysiology 34, 446–450.

Blumenthal, T.D., 1999. Short lead interval modification. In: Dawson, M.E., Schell, A.M., Bohmelt, A.H. (Eds.), Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science. Cambridge University Press, Cambridge, UK, pp. 51–71.

Blumenthal, T.D., Avendano, A., Berg, W.K., 1987. The startle response and auditory temporal summation in neonates. Journal of Experimental Child Psychology 44, 64–79.

Blumenthal, T.D., Cuthbert, B.N., Filion, D.L., Hackley, S.A., Lipp, O.V., van Boxtel, A., 2005. Committee report: guidelines for human startle eyeblink electromyographic studies. Psychophysiology 42, 1–15.

Blumenthal, T.D., Elden, A., Flaten, M.A., 2004. A comparison of several methods used to quantify prepulse inhibition of eyeblink responding. Psychophysiology 41, 326–332.

Blumenthal, T.D., Franklin, J.C., 2009. The startle eyeblink response. In: Harmon-Jones, E., Beer, J. (Eds.), Methods in Social Neuroscience. Guilford Press, New York, pp. 92–117.

Blumenthal, T.D., Gescheider, G.A., 1987. Modification of the acoustic startle response by a tactile prepulse: effects of stimulus onset asynchrony and prepulse intensity. Psychophysiology 24, 320–327.

Bradley, M.M., Cuthbert, B.N., Lang, P.J., 1999. Affect and the startle reflex. In: Dawson, M.E., Schell, A.M., Boehmelt, A.H. (Eds.), Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science. Cambridge University Press, Cambridge, UK, pp. 51–71.

Braff, D.L., Geyer, M.A., Swerdlow, N.R., 2001. Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies. Psychopharmacology 156, 234–258.

Davis, M., 2006. Neural systems involved in fear and anxiety measured with fear-potentiated startle. American Psychologist 61, 741–756.

Davis, M., Walker, D.L., Lee, Y., 1999. Neurophysiology and neuropharmacology of startle and its affective modulation. In: Dawson, M.E., Schell, A.M., Bohmelt, A.H. (Eds.), Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science. Cambridge University Press, Cambridge, UK, pp. 95–113.

Davis, M., Walker, D.L., Miles, L., Grillon, C., 2010. Phasic vs sustained fear in rats and humans: role of the extended amygdala in fear vs anxiety. Neuropsychopharmacology 35, 105–135.

Dichter, G.S., Tomarken, A.J., 2008. The chronometry of affective startle modulation in unipolar depression. Journal of Abnormal Psychology 117, 1–15.

Dichter, G.S., Tomarken, A.J., Shelton, R.C., Sutton, S.K., 2004. Early- and late-onset startle modulation in unipolar depression. Psychophysiology 41, 433–440.

Drobes, D.J., Miller, E.J., Hillman, C.H., Bradley, M.M., Cuthbert, B.N., Lang, P.J., 2001. Food deprivation and emotional reactions to food cues: implications for eating disorders. Biological Psychology 57, 153–177.

Eaton, R.C. (Ed.), 1984. Neural Mechanisms of Startle Behavior. Plenum, New York.

Fendt, M., Li, L., Yeomans, J.S., 2001. Brain stem circuits mediating prepulse inhibition of the startle reflex. Psychopharmacology 156 (2–3), 216–224.

Flaten, M.A., Blumenthal, T.D., 1999. Caffeine-associated stimuli elicit conditioned responses: an experimental model of the placebo effect. Psychopharmacology 145, 105–112.

Fournier, P., Hebert, S., 2013. Gap detection deficits in humans with tinnitus as assessed with the acoustic startle paradigm: does tinnitus fill in the gap? Hearing Research 295, 16–23.

Franklin, J.C., Bowker, K.B., Blumenthal, T.D., 2009a. Anxiety and prepulse inhibition of acoustic startle in a normative sample: the importance of signal-to-noise ratio. Personality and Individual Differences 46, 369–373.

Franklin, J.C., Heilbron, N., Guerry, J.D., Bowker, K.B., Blumenthal, T.D., 2009b. Antisocial and borderline personality disorder symptomatologies are associated with decreased prepulse inhibition: the importance of optimal experimental parameters. Personality and Individual Differences 47, 439–443.

Franklin, J.C., Hessel, E.T., Aaron, R.V., Heilbron, N., Prinstein, M.J., 2010. The functions of nonsuicidal self-injury: support for cognitive-affective regulation and opponent processes from a novel psychophysiological paradigm. Journal of Abnormal Psychology 119, 850–862.

Graham, F.K., 1975. The more or less startling effects of weak prestimulation. Psychophysiology 12 (3), 238–248.

Greenstein, J.E., Kassel, J.D., 2010. The effects of smoking on selective attention as measured by startle reflex, skin conductance, and heart rate responses to auditory startle stimuli. Psychophysiology 47, 15–24.

Grillon, C., Baas, J.M.P., 2003. A review of the modulation of the startle reflex by affective states and its application in psychiatry. Clinical Neurophysiology 114, 1557–1579.

Grillon, C., Bass, J.M.P., Pine, D.S., Lissek, S., Lawley, M., Ellis, V., Levine, J., 2006. The benzodiazepine alprazolam dissociates contextual fear from cued fear in humans as assessed by fear-potentiated startle. Biological Psychiatry 60, 760–766.

Grillon, C., Heller, R., Hirschhorn, E., Kling, M.A., Pine, D.S., Schulkin, J., Vythilingam, M., 2011. Acute hydrocortisone treatment increases anxiety but not fear in healthy volunteers: a fear-potentiated startle study. Biological Psychiatry 69, 549–555.

Grillon, C., Pellowski, M., Merikangas, K.R., Davis, M., 1997. Darkness facilitates the acoustic startle reflex in humans. Biological Psychiatry 42, 453–460.

Gross, J.J., Thompson, R.A., 2007. Emotion regulation: conceptual foundations. In: Gross, J.J. (Ed.), Handbook of Emotion Regulation. Guilford Press, New York, pp. 3–24.

Hackley, S.A., 1999. Implications of blink reflex research for theories of attention and consciousness. In: Dawson, M.E., Schell, A.M., Bohmelt, A.H. (Eds.), Startle Modification: Implications for Neuroscience, Cognitive Science, and Clinical Science. Cambridge University Press, Cambridge, UK, pp. 137–156.

Hawk Jr., L.W., Baschnagel, J.S., Ashare, R.L., Epstein, L.H., 2004. Craving and startle modification during in vivo exposure to food cues. Appetite 43, 285–294.

Hazlett, E.A., Speiser, L.J., Goodman, M., Roy, M., Carrizal, M., Wynn, J.K., Williams, W.C., Romero, M., Minzenberg, M.J., Siever, L.J., New, A.S., 2007. Exaggerated affect-modulated startle during unpleasant stimuli in borderline personality disorder. Biological Psychiatry 62, 250–255.

Herbert, C., Kubler, A., Vogele, C., 2013. Risk for eating disorders modulates startle-responses to body words. PLoS ONE 8, e53667.

Hoenig, K., Hochrein, A., Quednow, B., Maier, W., Wagner, M., 2005. Impaired prepulse inhibition of acoustic startle in obsessive-compulsive disorder. Biological Psychiatry 57, 1153–1158.

Honeycutt, C.F., Perreault, E.J., 2012. Planning of ballistic movement following stroke: insights from the startle reflex. PLoS ONE 7, e43097.

Huggenberger, H.J., Suter, S.E., Blumenthal, T.D., Schaechinger, H., 2013. Maternal social stress modulates the development of prepulse inhibition of startle in infants. Developmental Cognitive Neuroscience 3, 84–90.

Jovanovic, T., Phifer, J.E., Sicking, K., Weiss, T., Norrholm, S.D., Bradley, B., Ressler, K.J., 2011. Cortisol suppression by dexamethasone reduces exaggerated fear responses in posttraumatic stress disorder. Psychoneuroendocrinology 36, 1540–1552.

Jovanovic, T., Szilagyi, S., Chakravorty, S., Fiallos, A.M., Lewison, B.J., Parwani, A., Duncan, E.J., 2004. Menstrual cycle phase effects on prepulse inhibition of acoustic startle. Psychophysiology 41 (3), 401–406.

Kaviani, H., Gray, J.A., Checkley, S.A., Raven, P.W., Wilson, G.D., Kumari, V., 2004. Affective modulation of the startle response in depression: influence of the severity of depression, anhedonia, and anxiety. Journal of Affective Disorders 83, 21–31.

Kiken, L.G., Shook, N.J., 2011. Looking up: mindfulness increases positive judgments and reduces negativity bias. Social Psychological and Personality Science 2, 425–431.

Lang, P.J., Bradley, M.M., 2013. Appetitive and defensive motivation: goal-directed or goal-determined? Emotion Review 5, 230–234.

Lang, P.J., McTeague, L.M., 2009. The anxiety disorder spectrum: fear imagery, physiological reactivity, and differential diagnosis. Anxiety, Stress, and Coping 22, 5–25.

Ludewig, S., Geyer, M., Ramseier, M., Vollenweider, F., Rechsteiner, E., Cattapan-Ludewig, K., 2005. Information-processing deficits and cognitive dysfunction in panic disorder. Journal of Psychiatry & Neuroscience 30, 37–43.

Mathias, C.W., Blumenthal, T.D., Dawes, M.A., Liguori, A., Richard, D.M., Bray, B., Tong, W., Dougherty, D.M., 2011. Failure to sustain prepulse inhibition in adolescent marijuana users. Drug and Alcohol Dependence 116, 110–116.

McTeague, L.M., Lang, P.J., 2012. The anxiety spectrum and the reflex physiology of defense: from circumscribed fear to broad distress. Depression and Anxiety 29, 264–281.

McTeague, L.M., Lang, P.J., Laplante, M.C., Cuthbert, B.N., Strauss, C.C., Bradley, M.M., 2009. Fear imagery in social phobia: generalization, comorbidity, and physiological reactivity. Biological Psychiatry 65, 374–382.

Miller, M.W., McKinney, A.E., Kanter, F.S., Korte, K.J., Lovallo, W.R., 2011. Hydrocortisone suppression of the fear-potentiated startle response and post-traumatic stress disorder. Psychoneuroendocrinology 36, 970–980.

Perry, W., Minassian, A., Lopez, B., Maron, L., Lincoln, A., 2007. Sensorimotor gating deficits in adults with autism. Biological Psychiatry 61, 482–486.

Rejeski, W.J., Blumenthal, T.D., Miller, G.D., Lobe, M., Davis, C., Brown, L., 2010. State craving, food availability, and reactivity to preferred snack foods. Appetite 54, 77–83.

Richter, S., Schulz, A., Zech, C.M., Oitzl, M.S., Daskalakis, N.P., Blumenthal, T.D., Schaechinger, H., 2011. Cortisol rapidly disrupts prepulse inhibition in healthy men. Psychoneuroendocrinology 36, 109–114.

Roemer, S., Nees, F., Richter, S., Blumenthal, T.D., Schaechinger, H., 2009. Endogenous cortisol suppression with metyrapone enhances acoustic startle in healthy subjects. Hormones and Behavior 55, 314–318.

Rubio, G., Borrell, J., Jimenez, M., Jurado, R., Grusser, S.M., Heinz, A., 2013. Variables involved in the cue modulation of the startle reflex in alcohol-dependent patients. Addiction Biology 18, 170–180.

Schaechinger, H., Weinbacher, M., Kiss, A., Ritz, R., Langewitz, W., 2001. Cardiovascular indices of peripheral and central sympathetic activation. Psychosomatic Medicine 63, 788–796.

Schicatano, E.J., Blumenthal, T.D., 1998. The effects of caffeine and directed attention on acoustic startle habituation. Pharmacology, Biochemistry, and Behavior 59, 145–150.

Schulz, A., Plein, D.E., Richter, S., Blumenthal, T.D., Schaechinger, H., 2011. Cold pressor stress affects cardiac attenuation of startle. International Journal of Psychophysiology 79, 385–391.

Spresser, C.D., Keune, K.M., Filion, D.L., Lundgren, J.D., 2012. Self-report and startle-based measures of emotional reactions to body image cues as predictors of drive for thinness and body dissatisfaction in female college students. Body Image 9, 298–301.

Swerdlow, N.R., Weber, M., Qu, Y., Light, G.A., Braff, D.L., 2008. Realistic expectations of prepulse inhibition in translational models for schizophrenic research. Psychopharmacology 199 (3), 331–388.

Takahashi, H., Hashimoto, R., Iwase, M., Ishii, R., Kamio, Y., Takeda, M., 2011. Prepulse inhibition of startle response: recent advances in human studies of psychiatric disease. Clinical Psychopharmacology and Neuroscience 9, 102–110.

Ueki, A., Goto, K., Sato, N., Iso, H., Morita, Y., 2006. Prepulse inhibition of acoustic startle response in mild cognitive impairment and mild dementia of the alzheimer type. Psychioatry and Clinical Neurosciences 60, 55–62.

Vaidyanathan, U., Welo, E.J., Malone, S.M., Burwell, S.J., Iacono, W.G., 2014. The effects of recurrent episodes of depression on startle responses. Psychophysiology 51, 103–109.

Vrana, S.R., Constantine, J.A., Westman, J.S., 1992. Startle reflex modification as an outcome measure in the treatment of phobia: two case studies. Behavioral Assessment 14, 279–291.

Walter, M., Degen, B., Treugut, C., Albrich, J., Oppel, M., Schulz, A., Schaechinger, H., Durstler-McFarland, K.M., Wiesbeck, G.A., 2011. Affective reactivity in heroin-dependent patients with antisocial personality disorder. Psychiatry Research 187, 210–213.

Williams, C., Wood, R., 2012. Affective modulation of the startle reflex following traumatic brain injury. Journal of Clinical and Experimental Neuropsychology 34, 948–961.

State and Local Taxation

Lars P Feld, University of St. Gallen, St Gallen, Switzerland
Friedrich Schneider, Johannes Kepler University of Linz, Linz, Austria

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14955–14961, © 2001, Elsevier Ltd.

Abstract

State and local taxation comprises those taxes that are collected at the subfederal government levels in order to finance state and local public services, assigning some discretion in the determination of rates and bases of these taxes to subfederal governments. The existence of state and local taxes is explainable by the geographical incidence of public services within a federation. In order not to provide wrong incentives for free riding, each individual should pay for the public services consumed and decide on the level of provision. Thus, state and local taxes are imposed to finance state and local public services. Because federal, state, and local taxes already exist, guidelines for the assignment of different taxes to different government levels are necessary. These guidelines are derived by considering different forms of interjurisdictional externalities. Actual tax systems in OECD countries do not, however, necessarily follow these guidelines of optimal subfederal taxation. Since these tax systems are not merely the outcome of historical incidents, political and institutional factors add to the positive understanding which kinds of taxes are used by subfederal jurisdictions.

State and local taxation comprises those taxes that are collected at the subfederal government levels in order to finance state and local public services, assigning discretion in the determination of rates and bases of these taxes to subfederal governments. Consider a nation with different layers of government that are only serving administrative purposes. Each government level only distributes the public services to and collects the taxes from that nation's citizens upon which they agreed at the national level. In such a country, no state or local taxation is necessary since the level of public goods and services is decided upon uniformly for the whole country. This leaves citizens in some local jurisdictions, which prefer more public services than the national average, as much dissatisfied as citizens in local jurisdictions, which prefer less. It is possible to provide public services, the benefits of which are limited geographically and which are not national in scope, at different levels in different subfederal jurisdictions. Enabling citizens to consume publicly provided goods and services at different levels according to their preferences makes everyone better off without making someone worse off. In order to let people, who want to consume more, pay a higher tax price for those state and local services, taxes must be differentiated accordingly between states and local jurisdictions. In such a world with a differentiated supply of publicly provided goods, each individual can reside in a jurisdiction where a certain level of public services is provided to adequate tax prices. The art of state and local taxation is the assignment of different kinds of taxes to government levels such that an invisible hand properly guides a nation in an ideal world to an optimal multiunit fiscal system. This is not necessarily the case in the real world and some answers exist why fiscal systems are not optimal.

The Rationale of Tax Assignment to the State and Local Levels

The basic idea of a power to tax for states and local jurisdictions is that their public goods and services should be financed by those citizens benefiting from them and deciding on their quantity and quality. In this case, individuals choose their place of residence according to their preferences and their dislike for congestion. A set of fiscal jurisdictions consisting of equal taste communities then results from a process of 'voting by feet' (Tiebout, 1956). Olson (1969) calls the basic condition for such an outcome the 'principle of fiscal equivalence.' It requires assignment to which public goods and services should be provided at which level of government. The guideline for this decision is the geographical incidence of those benefits and the costs borne by citizens in order to finance them. As some public goods and services reasonably cannot be provided at the local level, some kinds of taxes may not be assigned to the local level for good reasons, too.

There are numerous reasons why the proper assignment of tasks and the power to tax to the subfederal levels are crucial to avoid suboptimal outcomes. For example, take the Scala in Milan, the famous Italian opera house. As it is partly financed by user charges and partly by subsidies paid from general taxes of the city, it will be congested because residents living in the suburban towns are able to visit it paying only the user charges and thus a lower price than the city's citizens. Nonresidents receive some of the benefits from the public service without paying an adequate tax price. This spillover is one of the externalities that Gordon (1983) summarized in his treatment of an optimal tax system for subfederal jurisdictions. Another externality exists if nonresidents pay some of the taxes of state or local jurisdictions. Taxes are exported to nonresidents and state and local governments thus have an incentive to provide public services at a higher level than their residents prefer. An example of tax exporting can be found in the taxation of tourist trade or of natural resources in some US states. For the year 1962, McLure (1967) estimated for the US states that overall tax export rates were between 19 and 28% in the short run. In the long run, only 10 states exported more than 24% of their taxes or less than 16%. The estimates are corroborated in a more recent study by Metcalf (1993). Mieszkowski (1983) estimated the 'total waste' in Alaska in the beginning of the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71026-9

1980s that accrued from excessive public revenues and expenditures that were possible because of high natural resource taxes to about $500 million or 13% of estimated petroleum revenues. Tax exporting is not limited to commodity taxes or natural resource taxes. It may also occur if individuals, firms, or capital are mobile.

If factor mobility is allowed for, the problems of subfederal taxation are even enhanced. If a state or community levies a higher personal or corporate income tax than its neighboring jurisdictions, its mobile citizens or firms emigrate (or move their capital there) in order to enjoy a lower tax burden. Such a situation is characterized as tax competition. In doing so, mobile factors reduce the tax burden of residents and firms in the state or community they move to and increase the tax burden in those jurisdictions they leave. These changes in tax burdens usually are not considered by public authorities in these jurisdictions when deciding on the level of public goods and services such that these are provided at a lower than 'optimal' level. Since there is a local loss from taxation that does not correspond to a social loss, the cost of public services is overstated and jurisdictions will tend to underspend on state and local public services. Buchanan and Goetz (1972) call this effect a fiscal externality.

The effects of tax competition among states and local jurisdictions are similar to those of international tax competition and they vary according to different degrees of mobility of individuals and firms, to the number of mobile factors, to the number of jurisdictions involved in that competition, on the size of the jurisdictions involved, and to the number of tax instruments available to tax authorities. The assessment of the usefulness of these effects becomes rather complicated if competition using tax instruments is complemented by competition using public expenditures, like subsidies or specific infrastructural services, but also tax expenditures like tax holidays to attract mobile production factors.

The impact of tax competition on the possibilities of state and local governments to redistribute income by tax-transfer schemes is relatively clear-cut. Stigler (1957) already stated that "redistribution is intrinsically a national policy." Suppose that a region adopts a progressive income tax program designed to achieve a significantly more egalitarian distribution of income than exists in neighboring jurisdictions. If rich and poor households are mobile, such a program would create strong incentives for the wealthy to move out to neighboring jurisdictions. In this case, local redistribution induces sorting of the population, with the richest households residing in the communities that redistribute the least by income taxes. A more equal distribution of income would result, but it would be caused largely by an outflow of the rich and a consequent fall in the level of per capita income in the jurisdiction under consideration (Oates, 1972). At the central level, this problem does not occur, because mobility is reduced, the larger a jurisdiction.

There are not many theoretical arguments against this line of reasoning. Most of them rely on imperfect mobility of individuals. But the more mobile people become, the better this mechanism should work, the less possible should state and local redistribution by a progressive income tax be. However, according to Buchanan (1975), all individuals have an incentive at the constitutional stage to agree to income redistribution because they are fundamentally uncertain about their future positions in income, health, and employment. At the post-constitutional level the rich might agree to income redistribution by the government because they, first, are interested in a public insurance scheme against fundamental privately uninsurable risks for themselves and their children, and, second, against exploitation by the majority of the poor residents and increasing crime rates. Particularly the second reason is important for the rich: they pay a premium for obtaining social peace. Janeba and Raff (1997) argue that individuals will agree to decentralized income redistribution because it is a credible commitment of the poor majority to the rich minority that the latter is not exploited at the postconstitutional stage. The rich allow for the agreed extent of redistribution because a constitutional competence for income redistribution at the federal level leads to a larger public sector with more income redistribution. From this perspective, decentralized redistribution might be possible and even more efficient since redistribution is closer to citizens' preferences.

Guidelines for a Reasonable Tax Assignment

This reasoning leads to some guidelines for tax assignment of different taxes to the state and local levels. The bottom line is that state and local jurisdictions should finance the services they provide by charges or quasicharges to the beneficiaries (Musgrave, 1983) or they should use those tax bases which have low interregional mobility (Wildasin, 1980). Land and natural resource taxes and to a lesser degree also real estate taxes are thus especially suited for purpose of state or local taxation. A property tax, for example, is a good candidate for being assigned to the local level at least as far as residential property instead of business property is taxed. A property tax is basically a tax on land and construction built on it and it can thus be considered to be a tax on a relatively immobile tax base.

Consumption taxes are inappropriate at the local level because the interjurisdictional mobility leads to cross-border shopping or tax exporting, but they become suitable for the state level if the region is large enough. However, the technological development, in particular Internet shopping, reduces the possibility for state governments to levy a tax upon consumption. If individuals, firms, and financial capital are fully mobile a comprehensive income or profits tax should not be assigned to the state and local levels because it is necessary for such a taxation to be optimal so that a jurisdiction should be able to tax all income of its residents in the other jurisdictions. This is certainly not a realistic case. Beyond this, taxation of wage income tends to reduce this tax base less than the taxation of capital income and it could thus be more easily taxed at the state or even local level unless tax authorities cooperate and exchange information about taxpayers' incomes. And even profits taxes can be levied at the state or local levels if investment is specific to the locality such that a firm cannot easily relocate. The latter is particularly the case for not-in-my-back-yard (NIMBY) firms like waste incinerators or nuclear power plants. These firms have difficulties finding a location and assigning local jurisdictions the power to tax their profits gives these communities an incentive to provide such firms with a location.

Following from this reasoning is another rule for tax assignment: Progressive taxation, designed to secure redistributional objectives, should be assigned to the central level. A progressive income tax should thus be a federal tax. Moreover, as the case of Alaska illustrates, a large natural resource tax base permits local jurisdictions to provide public services at a low tax price leading to an attraction of individuals and firms at an inefficient scale. On these grounds, a case for central taxation of natural resources may be made whereby the central natural resource tax may apply to an excess base only, while leaving the average base for subfederal use.

All in all, these basic considerations led Musgrave (1983) to present the following tax assignment to government levels: the local level should rely on property taxes and payroll taxes, the state level on a resident income tax, a consumption tax, and a natural resource tax, and the federal level on an integrated progressive income tax and a natural resource tax for large natural resource bases. All government levels should additionally rely on fees and user charges as far as public services are attributable to particular users or certain groups of users. This possibility is probably reduced from the local to the federal level. According to Musgrave, corporate source income ideally would be included in the integrated base of the federal income tax. These propositions become substantially more complicated, if interjurisdictional tax agreements providing for tax credits or deductibility are considered. Federal tax deductibility of state and local taxes, for example, may mainly serve taxpayers with higher income and is thus regressive.

Tax Assignment in Some OECD Countries

A first look at real tax systems in OECD countries quickly reveals that tax assignment is not necessarily Musgravian (Pola et al., 1996; Pommerehne and Ress, 1996). The US seem to come closest to those rules of tax assignment. The US states are prohibited by the US constitution from levying taxes on imports or exports but otherwise they may adopt any type of tax that does not egregiously discriminate against citizens of other states. US states have a lot of freedom in defining their tax bases, in setting tax rates, and in offering tax abatements to induce businesses to locate within their state. Finally, the states grant tax authority to differing degrees to local jurisdictions, too. While there are only minimal restrictions on state power to tax in the US, the federal level nevertheless relies mainly on the individual income and corporate income taxes, the states on sales taxes and natural resource taxes, and the most important source of revenue at the US local level is the property tax. Although state tax revenue from individual income and corporate income taxes is nonnegligible, it is only about one fifth of the respective federal revenue. Such a favorable judgment about the US tax system is, however, only justified at a first glance. According to Feldstein and Metcalf (1987), for example, the federal tax deductibility of personal income tax liabilities led many states to rely more heavily on deductible personal income taxes than on other type of revenue.

Apart from the US, for example, in Canada and Australia and particularly in Europe, the assignment of taxes to the state and local levels differs considerably from the Musgravian model and varies widely across countries. In Canada, personal and corporate income taxes are the most important revenue sources of the federal level. The most visible of taxes in Canada, the personal income tax, is, however, also the most important revenue source of the provinces. On the one hand, there are fairly significant income tax rate differences between provinces. On the other hand, arrangements between the Canadian federal and provincial governments on sharing the bases for the personal and corporate income taxes have a long history going back at least to the Second World War. There is also a large variation in sales tax rates from Newfoundland to Alberta as the second important revenue source of the provinces. Canadian local jurisdictions, like those in the US, mainly rely on property taxes.

In Australia, the federal government exclusively levies taxes on personal and company income, sales taxes, and (nearly exclusively) excise taxes. The bulk of a state's tax revenues are derived from payroll taxes, stamp duties, and associated taxes on financial transactions, and so called franchise (i.e., business license) fees on alcohol, tobacco, and petroleum products that are not so much differing from excise taxes as well as lots of smaller revenue sources. Local governments, on the other hand, derive their revenues almost entirely from property taxes. Australian states also obtain revenue from the central government on the basis of a revenue sharing arrangement.

The Australian tax system is not as different from the German and Austrian system of state and local taxation. In particular Germany has established an extensive revenue sharing arrangement of personal and corporate income and value added taxes as the most important taxes accounting for about three quarters of total tax revenue. These are really shared taxes in the sense that German state governments decide on them together with the federal government. The federal government has at its own discretion only the gasoline tax and a surcharge on income taxes while the states have no power to tax. Local government relies on property taxes and local business taxes.

The less federalist European countries usually have taxing authority at the local level. In the UK, however, there is not only no regional or state level of government, but local authorities basically have not much power to tax, only some power in determining tax rates of the local property tax on real property. The formerly existing local business tax in the UK was abolished by the 'Thatcher' administration which cut the contributions of local taxes to local revenue sharply after the unsuccessful attempt to introduce the poll tax in the 1980s. There have been some changes in the UK at the end of the 1990s to establish regional authorities in Scotland, Wales, and Northern Ireland providing at least the Scottish government with some power to tax. The UK thus appears to be following the trend in Europe to a more decentralized government structure with a power to tax for subfederal governments as well.

Since the 1970s, systems of regional government have been given an enhanced role in taxation in France, Italy, and Spain. In France, there are three levels of subnational government: the local level with about 36 000 cities and communities, about 100 départements and 22 regions. Since 1986, the importance of these subnational governments has increased. All three levels have the possibility to levy a property tax on real property (land and construction built on it), a local business tax ('*taxe*

professionelle') and something similar to a residents' income tax ('*taxe d'habitation*'). All three taxes are subject to central regulation, but nevertheless there is a nonnegligible variation in tax rates.

In Belgium, there is a movement toward fiscal federalism since it is agreed that the sales and property taxes will become exclusive taxes for the regional and local levels. Apart from smaller taxes, the three Belgian regions additionally can levy a supplementary rate on the personal income tax which may be positive, a surcharge, or negative, a tax reduction. Although there is some regulation of the local and regional power to tax income, the variation of rates is considerable. In contrast to that, local governments in the Netherlands can only levy property taxes.

An interesting pattern of local taxation can be observed in the Nordic countries. Basically, Denmark, Finland, Norway, and Sweden are unitary countries with only two government levels. According to the studies in Rattsø (1998), local governments have considerable power to tax in Denmark and Finland, less so in Sweden, and only to a limited extent in Norway. In Denmark and Finland, local governments can rely on taxes on property and personal income (nearly) without restrictions for setting tax rates accompanied by a considerable variation. In Sweden, local taxation consists mainly of personal income taxes with less variation in rates due to the central regulation of local governments. In Norway, most tax revenue originates from income and wealth taxes shared with the central government giving local authorities taxing power only within a narrow band. Since 1979, all local governments apply the maximum rate. The property tax is not available to all local governments, but is restricted to urban areas and certain facilities (notably power plants). In addition, local property tax rates are also limited to a narrow band.

In Switzerland, states (cantons) have the basic power to tax personal and corporate income as well as capital. The local jurisdictions can levy a surcharge on cantonal direct taxes and raise their own property taxes. The central government relies mainly on indirect (proportional) taxes, the value added tax, and specific consumption taxes like the mineral oil tax. It can also rely on a source tax on income from interest. There is, finally, a small but highly progressive federal income tax, which, together with revenue from the source tax on interest income, amounted to 34% of total federal tax revenue in 1995, while the cantons and municipalities rely on income and property taxes to about 50% of their total revenue and 95% of their tax revenue. In addition, personal and corporate income taxes in Switzerland vary considerably between the cantons. Taking the value of the (weighted) average for Switzerland as 100, the index of the tax burden of personal income and property taxes varies from 55.6 in the canton of Zug to 135.8 in Fribourg in 1995, while the index of corporate income and capital taxes varied from 57.6 again in Zug to 145.6 in Nidwalden.

As the overview of these selected OECD countries in Table 1 indicates, there is no clear picture of which kinds of taxes are assigned to the state and local levels. Local jurisdictions regularly levy property taxes as well as fees and user charges. Some power to tax personal income, at least labor income, is attributed to state and local governments in the US, Canada, France, Belgium, the Nordic countries, and Switzerland. In some cases, for example, Germany, France, and Switzerland, there exist even local business taxes in the discretion of subfederal jurisdictions. But on the whole, the assignment of taxes does not follow along Musgravian lines and there is a considerable variation in federal regulation of subfederal taxing power.

Table 1 State and local tax assignment in selected OECD countries

Country	*States*	*Local authorities*
USA	Income tax Sales tax Natural resource tax	Property tax Sales tax
Canada	Personal and corporate income tax Sales tax	Property tax
Australia	Payroll taxes Stamp duties Franchise fees	Property taxes
Germany	No taxing power	Property tax Business tax
UK	No state level	Property tax
France	Property tax Business tax Residents' income tax	Property tax Business tax Residents' income tax
Belgium	Personal income tax surcharge Property tax	Personal income tax surcharge Property tax
Switzerland	Personal and corporate income tax Inheritance tax Wealth tax	Personal and corporate income tax surcharge Property tax
Norway	No state level	Personal income tax Property tax
Denmark	No state level	Personal income tax Property tax
Musgravian 'optimal' assignment	Residents' income tax Consumption tax Natural resource tax	Property taxes Payroll taxes

The Political Economy of State and Local Taxation

Given these differences in tax assignment to state and local authorities, the question arises which institutional and political factors explain the tax structure in different countries. An early contribution as to the choice of taxation in a political economy framework originates from Schneider and Pommerehne (1980). They develop a model of Australian tax and expenditure structure along with a model of selfish government behavior assuming that the re-election constraint is not always binding. In choosing among tax instruments, a government in power is only urged to follow the preferences of a majority of voters if an election approaches and the re-election constraint is binding due to low government popularity. Between elections, the government can follow its ideological preferences. This modeling of government behavior proved to be clearly superior compared to the traditional models of government behavior in ex-post predictions of important variables (government spending, GDP, etc.). In contrast, Hettich and Winer (1999) assume that the governing party or coalition always maximizes the expected votes which by themselves depend on the

consequences of policies for the economic welfare of voters. In making fiscal choices, political parties thus take into account economic, political, and administrative factors that determine the nature of electoral support and opposition associated with alternative ways of taxing. The authors present empirical evidence on the reliance on income taxation and income tax structure of the US states in 1985 and 1986 supporting their hypotheses.

Inman (1989) finds that redistributive concerns strongly influenced the decision of the 41 largest US cities between 1961 and 1986 to rely on property taxes, sales taxes, or user charges. The higher the demand of local voters for redistributive services, the less city governments rely on fees and user charges as well as sales taxes and the more on property taxes. Hunter and Nelson (1989) present evidence for the impact of interest groups on the local revenue structure. The higher the share of wealthy homeowners in local jurisdictions in Louisiana, the lower the share of property tax revenue and the higher the share of charges.

Besides vote-maximizing behavior of politicians and interest-group influence for redistribution in their favor, tax competition is a third factor determining the structure of taxation. However, the expected impact is ambiguous. On the one hand, tax competition for mobile capital and workers may lead to a convergence of tax rates to a low level. On the other hand, tax competition may lead to increases in tax rates if a state's geographic neighbors follow suit when it raises its tax rates. Such behavior might occur due to yardstick competition between states and local jurisdictions (Besley and Case, 1995). Due to their incomplete political information, voters evaluate the fiscal performance of representatives in their jurisdiction by comparing it to the fiscal performance of neighboring jurisdictions. If a state's neighbor increases tax rates, this state has an increased possibility to raise its taxes as well. Arkansas State Senator Doug Brandon once said: "We do everything everyone else does."

Finally, the structure of state and local taxation, like federal taxation, is influenced by the degree of fiscal illusion among taxpayers. The more complex a tax system, the more incorrect voters' perceptions about their true tax burden and thus their tax resistance, the more easily representatives can increase taxes in order to finance new spending. Pommerehne and Schneider (1978) present evidence for Swiss local jurisdictions that the extent of fiscal illusion attributed to the complexity of the tax system is less pronounced, the more voters can determine taxes and spending directly in fiscal referenda. Moreover, Matsusaka (1995) finds that US states which allow for the use of initiatives in their constitutions rely more on fees and user charges than on broad based taxes increasing the visibility of contributions to the public sector and the equivalence between public services and tax prices.

Concluding Remarks

Although a comprehensive assessment is still too early, state and local taxation, appears to be the outcome of political decisions shaped by constitutional provisions and influenced by economic and administrative restrictions. Voters, representatives, bureaucrats, and interest groups deciding upon taxation under the decision-making provisions of their constitutions and taking into account the mobility of tax bases and the accompanied tax competition and tax exporting determine to what extent state and local jurisdictions rely on property, income and sales taxes, or on user charges. The differences in these political, institutional, and economic factors explain why there are so many differences among countries in their reliance on tax sources at the state and local levels.

Bibliography

Besley, T., Case, A., 1995. Incumbent behavior: vote-seeking, tax-setting, and yardstick competition. American Economic Review 85, 25–45.

Buchanan, J.M., 1975. The Limits of Liberty. University of Chicago Press, Chicago.

Buchanan, J.M., Goetz, C.J., 1972. Efficiency limits of fiscal mobility: an assessment of the Tiebout model. Journal of Public Economics 1, 25–43.

Feldstein, M.S., Metcalf, G.E., 1987. The effect of federal tax deductibility on state and local taxes and spending. Journal of Political Economy 95, 710–736.

Gordon, R.H., 1983. An optimal taxation approach to fiscal federalism. Quarterly Journal of Economics 98, 567–586.

Hettich, W., Winer, L., 1999. Democratic Choice and Taxation: A Theoretical and Empirical Analysis. Cambridge University Press, Cambridge, UK.

Hunter, W.J., Nelson, M.A., 1989. Interest group demand for taxation. Public Choice 62, 41–61.

Inman, R.P., 1989. The local decision to tax: evidence from large US cities. Regional Science and Urban Economics 19, 455–491.

Janeba, E., Raff, H., 1997. Should the power to redistribute income be (de-) centralized? an example. International Tax and Public Finance 4, 453–461.

Matsusaka, J.G., 1995. Fiscal effects of the voter initiative: evidence from the last 30 years. Journal of Political Economy 103, 587–623.

McLure, C.E., 1967. The interstate exporting of state and local taxes: estimates for 1962. National Tax Journal 20, 49–77.

Metcalf, G.E., 1993. Tax exporting, federal deductibility, and state tax structure. Journal of Policy Analysis and Management 12, 109–126.

Mieszkowski, P., 1983. Energy policy, taxation of natural resources, and fiscal federalism. In: McLure, C.E. (Ed.), Tax Assignment in Federal Countries. Centre for Research on Federal Financial Relations, Canberra, Australia, pp. 129–145.

Musgrave, R.A., 1983. Who should tax, where and what? In: McLure, C.E. (Ed.), Tax Assignment in Federal Countries. Centre for Research on Federal Financial Relations, Canberra, Australia, pp. 2–19.

Oates, W.E., 1972. Fiscal Federalism. Harcourt/Brace/Jovanovich, New York.

Olson, M., 1969. The principle of 'fiscal equivalence': the division of responsibilities among different levels of government. American Economic Review, Papers and Proceedings 59, 479–487.

Pola, G., France, G., Levaggi, R. (Eds.), 1996. Developments in Local Government Finance: Theory and Policy. Edward Elgar, Cheltenham, UK.

Pommerehne, W.W., Ress, G. (Eds.), 1996. Finanzverfassung im Spannungsfeld zwischen Zentralstaat und Gliedstaaten. Nomos, Baden-Baden, Germany.

Pommerehne, W.W., Schneider, F., 1978. Fiscal illusion, political institutions, and local public spending. Kyklos 31, 381–408.

Rattsø, J. (Ed.), 1998. Fiscal Federalism and State-Local Finance: The Scandinavian Perspective. Edward Elgar, Cheltenham, UK.

Schneider, F., Pommerehne, W.W., 1980. Politico-economic interactions in Australia: some empirical evidence. Economic Record 56, 113–131.

Stigler, G.J., 1957. The tenable range of functions of local government. In: US Congress, Joint Economic Committee (Ed.), Federal Expenditure Policy for Economic Growth and Stability. US Government Printing Office, Washington, DC.

Tiebout, C.M., 1956. A pure theory of local expenditures. Journal of Political Economy 64, 416–424.

Wildasin, D.E., 1980. Locational efficiency in a federal system. Regional Science and Urban Economics 10, 453–471.

State and Society: History of the Concept

Gianfranco Poggi, European University Institute, Fiesole, Italy

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14961–14964, © 2001, Elsevier Ltd.

Abstract

The conceptual couple 'state/(civil) society' is a product of the protracted effort through which, starting with the Enlightment, a few generations of European intellectuals sought to characterize the advance of modernization. The juxtaposition of 'state' to '(civil) society' emphasizes the growing differentiation between, on the one hand, the structures and events directly connected with public authority, with the constitution and management of the polity, and on the other the structures and events which individuals generate as they pursue their private interests, express their emotions, acquire and exercise their capacity for judgment. Some institutional outcomes of this process can be still be recognized, according to many authors, in the shape and dynamics of contemporary societies.

An Unsettled Distinction

Speaking of a 'juxtaposition' between two concepts suggests that they lie on the same plane and are unproblematically complementary to one another. But this is not the only relation in which those concepts have stood to one another. In an earlier relation, the two concepts, instead, were closely overlapped. In a philosophical and juridical vocabulary that lasts between Aristotle and Kant, the expression 'civil society' designates a condition in which humans have left the state of nature by bringing about, and subjecting themselves to, political authority; one wrote of *societas civilis SIVE res publica*. Here, what made a society civil was the very fact of its being policed and governed.

Hegel (1953) breaks with this fusion of the two concepts in the *Philosophy of Right* of 1821. The civil society is understood as the set of individuals engaged in the competitive pursuit of their private interests (chiefly economic). The state is understood as an institutional site where identified collective interests – such as the increase in the state's security, or the maintenance of the public order – are pursued through expressly political activities.

But this juxtaposition is loaded with tensions, and Hegel presents conceptual entities – the family, the corporation, or the 'police' (which here means: the administration) – intended to mediate those tensions. Further, the juxtaposition may be perceived as a contrast between state and society, leading thinkers to take sides with the one or the other. Hegel himself opts for the state, but already Lorenz von Stein, who places the paired concepts state/society at the center of his main work, resolutely opts for society.

In a sense, so does Karl Marx, who like Hegel views civil society as the locus of the formation and confrontation of private interests of an economic nature. However, he endows it with a powerful, self-sustaining dynamic, grounded on means and relations of production, with respect to which the state itself loses the creative, autonomous power to drive historical development Hegel attributed to it.

Even in the course of the twentieth century, some authors prioritize one of two concepts and subsume under it the other. Oliver Williamson's (1975) theory of markets and hierarchies, for instance, posits economic exchange as the primordial social relation, and derives from 'market failures' the emergence of hierarchical, command-based relations, and thus of organizations. Implicitly, also in this view society, conceived in the first instance as the locus of market relations, precedes and in a sense justifies the state (if one sees in it a particularly significant organization). This priority of society, naturally, imposes tight limitations on what the state can and should do (see Karlson, 1993).

Such arguments imply that state/society is not a proper dichotomy, for one concept can be made to encompass and posit the other instead of accepting it as its peer. But there are also arguments to the effect not that one concept is sufficient, rather that more than two are needed.

Thus, one may object to the Hegelian/Marxian reduction of the civil society to the political economy or the market, and suggest for it an alternative referent. A model of a social realm focused on intellectual pursuits, and constituted and regulated without political mediation and power relations, can be found in the eighteenth century *Imperium Litterarum*, the international cosmopolis of *savants* and literati. Or one may seek to comprise within civil society the phenomena of culture and religion, or the institutions concerned with play, emotions, aesthetic expression, and other inter-individual relations not reducible to exchange. The market itself may either be made to adjust itself to the presence, under the conceptual umbrella of 'civil society,' of these new and rather different referents, or may be relocated to a third realm all of its own, as if to acknowledge its peculiarly powerful and demanding dynamic.

For instance, according to Cohen and Arato (1992), in advanced societies there are three fundamental institutional complexes – the state, the market, and civil society. The latter is conceived primarily as the site not of competition between private interests, rather of the formation of collective identities, of solidarity, of the formation and transformation of institutions by means of social movements. In this perspective, *tertium datur*. *Politics and Markets* (Lindblom, 1977) do not exhaust the social space.

Historical Roots of the Distinction

One reason why, at the purely conceptual level, the relations between the two terms are so unsettled, may be that 'state' and '(civil) society' should not be conceived as universal categories.

Whether we conceive them as juxtaposed or counterposed to one another, both expressions refer to historically unique ways of establishing and regulating both political and economic processes (and sociocultural ones, if a tripartition is accepted as more plausible).

The state/society distinction presupposes, and profoundly modifies, arrangements which had emerged in various European countries in the late Middle Ages, and which are often characterized by the German expression *Ständestaat* (literally, 'state of the estates'). But the expression is potentially misleading, for it may lead one to think of its referent as a 'state' in the modern sense of the term.

To avoid this misunderstanding the Austrian historian Otto Brunner (1968) has proposed replacing that expression with that of *ständische Gesellschaft*. By this is meant a context where constituted estates, through arrangements partly traditional partly agreed, take an active part in the government of a territory. However they do so chiefly in the exercise of their own privileges, not on a mandate from the central government; and in those privileges political aspects are tightly fused with both economic ones and with status ones, relating to honor and rank. Their activities of government are differentiated, but not as articulations of an overriding political function, rather on the basis of each estate's claims and responsibilities.

Furthermore, typically the component units of each estate are not individuals, but collective entities with a pronounced hierarchical structure. The dominant unit is the 'house,' understood as the seat and patrimony of an aristocratic dynasty. "The lord of the house (*Hausherr*), at the same time that he performed the function, which today we consider private in nature, of head of the family, also exercised an authentic political power (jurisdictional, administrative, representative) over the members of the house subjected to him," which tended to include the bulk of the local rural population (Schiera, 1983). Furthermore, each 'house' tended to be also a fortified place, an assemblage of autonomous military resources, which the members of the family (again, including the rural dependants) would employ at the behest and under the command of the lord. In that capacity the house normally took part in the ruler's military undertakings; but equally could also protect the house's rights, arms in hand, from encroachments by other houses, including the ruler's own.

Finally, the typical house constituted an *oikos*, which is/was the seat of a complex plurality of activities intended to make it economically self-sufficient, as much as possible without recourse to exchange. The monitoring and organization of those activities (including, critically, the periodic delivery to the lord's household of unpaid produce and labor by the rural dependants) engaged the lord's faculties of jurisdiction and coercion.

The Crisis of the Staendische Gesellschaft

The 'decline and fall' of these arrangements was caused by a critical aspect of state formation, the gathering at the center of a country of the faculties and facilities of organized violence, culminating, according to Max Weber's well known formula, in a monopoly of legitimate violence. Exactly because the aristocratic houses originally possessed an autonomous military capacity, that process, when it does not dissolve them outright, turns them into subordinate components of an unchallengeable public order, centered on the sovereign, their councils, standing army, and fiscal, administrative and judicial apparatus. The political autonomy of the houses, and thus of the estates they collectively constitute, is lost. The sovereign's superiority, while it does not suppress the houses' internal hierarchical arrangements (including, of course, 'patriarchy') relativizes them, flattens them. Seen from the summit of sovereign power even the most exalted members of the realm begin to appear relatively similar to one another, relatively powerless.

At the same time that centralized political activities become the prerogative of a sovereign power that performs them in a more unitary and rational manner, economic activities become decentralized. An increasing number of subjects decide what goods and services to produce, in what ways, by interacting according to their own interests and capacities. The provision for the needs of the individuals depends increasingly on their capacity to enter into traffics with one another on the market.

In the countryside, commercialization reduces the part played in the totality of the economic process both by the aristocratic patrimonies and by the traditional rural units constituted by villages, which had shared the aristocrats' preference for self-sufficiency and their hostility to the market. It also diminishes the significance of city-based productive units regulated by guilds, which typically contrasted the market's tendency toward innovation and the free combination and recombination of resources.

To sum up: while state-building dissolves the *staendische Gesellschaft* by taking from it tasks, resources, and legitimacy pertaining to the political sphere, the advance of an economy based on traffics pushes into the market the units let loose by that dissolution. Against an expressly political organization with a strong vertical gradient, concerned with all relations grounded on *imperium*, there stands an open-ended set of increasingly numerous social units which, in the pursuit each of its own interest, establish with one another horizontal, or apparently horizontal relations.

The Political Prerequisites of the Civil Society

This can only take place, however, insofar as the process of state building reflects two peculiar principles: first, the political center takes on all expressly political concerns and resources but only those; second, it acknowledges the existence of private subjects who, while excluded from the political sphere, act autonomously in others. In both ways, political power expressly limits itself, through complex and delicate institutional arrangements: for instance, the progressive denial of political and juridical significance to the religious interests of individuals, or the acknowledgment and guarantee of private property and of contract.

Heinrich Popitz (1986) highlights the historical rarity of such arrangements: "Only rarely has it been possible even just to pose the question of how to constrain institutionalized violence, in a systematic and effective manner. This has only

taken place with the Greek *polis*, republican Rome, a few city-states, and the constitutional state of the modern era. The answers have been surprisingly similar: the principle of the primacy of the law and of the equality of all before the law (*isonomia*); the idea of setting boundaries to all legislation (basic rights), norms of competence (separation of powers, federalism), procedural norms (decisions to be taken by organs, their publicity, appeal to higher organs), norms concerning the distribution of political responsibilities (turn-taking, elections), public norms (freedom of opinion and of assembly)."

Within this small set of experiences of limited political power not all counterpose two realms we might call 'state' and '(civil) society.' For instance, in the Greek polis individuals exist chiefly as soldiers and citizens, and the private dimension of their life is little valued and expressed. Only vis-à-vis the modern constitutional state does (civil) society emerge, through the processes indicated above.

In many other civilizations the tendency of political power to aggrandize itself, to possess itself of all resources (beginning with land and labor power) suppresses all autonomy of the society, and may even reduce all individuals to subjects or slaves. In the modern West, by acknowledging and protecting private property in productive resources, the political center allows at least some subjects to stand against it as autonomous units. This, according to North and Thomas (1973), greatly favors the advance of the division of labor and the enlargement of the traffic economy. But it has also distinctive political effects, best perceived *a contrario* in the 'decline and fall' of empires which fail to acknowledge and protect property rights (Eisenstadt, 1969).

The significance of private property has two implications. First, it justifies to an extent the Hegelian/Marxian understanding of civil society primarily as the site of traffics. Second, it entails that the civil society should not be envisaged as a 'flat' space, harboring no power relations; for private property, especially in the means of production, engenders a power difference between those who do and those who do not possess it. The state/society difference does not mean that power is institutionalized only in the former, but that the latter is structured (among other things) by power relations of its own, foremost among which are those established by property.

State and Society in the Twentieth Century

The complex interactions between the power forms that lie on each side of the divide shape, to a large extent, the dynamics of modern societies. In the twentieth century the prevailing trend could be called 'statism,' for it saw a considerable enlargement of the scope of state action. In totalitarian systems, this trend suppressed all institutional limitations upon that action, and threatened the existence itself of society as a separate realm. In other systems, it pushed back considerably the boundaries of that realm, but this in response also to demands originating from it. For instance, economically underprivileged strata sought to maximize the import of citizenship and to reduce that of their market position; many economically powerful groups sought the state's assistance in securing their profits and stabilizing or correcting the market itself, which for various reasons had lost its self-equilibrating capacity. A further component of the trend consisted in the strategies of groups (sections of the 'political class,' public employees, members of the so-called 'new professions') whose livelihood and whose opportunities for a greater sway on policy depended on an expansive definition of the state's role.

Whatever its protagonists, in non totalitarian countries that definition was much assisted by the existence of a 'public sphere' and by the institutionalization of adversary politics, which gave voice to broader social strata, and found its most recurrent theme in the respective scope of the state and the market. Political parties played the central role in this dynamic, sometimes opposing but more often supporting the 'statist' trend. This expressed itself, among other things, in the formation of bodies and agencies that sat astride the traditional divide between the public and the private realm – as with the Italian *para-stato* or the British *quango*.

In the last decades of the twentieth century the trend was challenged by social forces expressing themselves politically through conservative and/or right-wing parties. These sought to at least slow down 'statism,' and proclaimed the superior rationality of the market, which in their discourse, but to an extent also in practice, reasserted itself as the chief, if not the only institutional site of the civil society. This late revisitation of the economy-centered, Hegelian/Marxian conception of the civil society acknowledges that in contemporary society also human potentialities and social processes – of a cultural, communicative, symbolic, apparent noneconomic nature – which other formulations treat as a *tertium* with respect to both state and market, are in fact targeted and controlled to a large extent by great concentrations of economic power. This is achieved through new material and social technologies, and subjected to the logic of accumulation, which is increasingly focused on informational and financial resources. Mostly those technologies are developed and operate transnationally, and thus increasingly escape the reach of the political competences and facilities vested in the states. This lends a new edge to the attempt to reduce, after decades of 'statism,' the significance of public and political action, and to enhance that of the market.

Also the end of the Cold War operated in that sense, for it reduced the necessity for strong political/military tutelage of the capitalist order. It also loosened the constraints of the Keynesian compact which the dominant strata had previously made with the lower orders and implemented largely through state action.

Both dominant trends – the explicit or implicit reduction of civil society to the market, and the slogan prevalent in the advanced countries at the end of the twentieth century: market, *si*, state, *no* – are potentially worrisome. The market itself remains the site of the formation and interaction between large economic forces, endowed with great power; and these are increasingly operating in a transnational space. In contrast the state – still wedded to the territory – can no longer perform even its most traditional functions, jurisdiction and police, let alone those functions of guidance and intervention in the economic realm it had increasingly claimed during most of the twentieth century.

These recent phenomena compromise the institutional tenability of the couple state/society nearly as much as it had been threatened, earlier, by the totalitarian state or by central

economic planning. A seriously dis-empowered state cannot establish and monitor an authentic civil society complementary to itself. In this sense, the Aristotelian and medieval vision according to which only a politically organized society can truly call itself 'civil,' deserves to be revisited.

Bibliography

Brunner, O., 1968. Neue Wege der Verfassungs- und Sozialgeschichte. Vandenhoeck & Ruprecht, Gottingen, Germany.

Cohen, J.L., Arato, A., 1992. Civil Society and Political Theory. MIT Press, Cambridge, MA.

Eisenstadt, S.N., 1969. The Political Systems of Empires. Free Press, New York.

Hegel, G.W.F., 1953. Philosophy of Right. Clarendon Press, Oxford, UK.

Karlson, N., 1993. The State of State: An Inquiry Concerning the Role of Invisible Hands in Politics and Civil Society. Almkvist & Wiksell, Stockholm, Sweden.

Lindblom, C.E., 1977. Politics and Markets: The World's Political Economic Systems. Basic Books, New York.

Marx, K., 1904. A Contribution to the Critique of Political Economy. Kerr, Chicago, IL.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, London.

Popitz, H., 1986. Phänomene der Macht. Mohr (Siebeck), Tubingen, Germany.

Schiera, P., 1983. Società per ceti. In: Bobbio, N., Matteucci, N., Pasquino, G. (Eds.), Dizionario di Politica. UTET, Turin, Italy.

Williamson, O.E., 1975. Markets and Hierarchies. Analysis and Antitrust Implications: A Study in the Economics of Internal Organization. Free Press, New York.

State: Anthropological Aspects

John Borneman, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Modern statehood is a mode of political organization of 'society' in which the state is the primary agent charged with building its representation. The nation has become the preferred form for this representation, law, the principal means used to anticipate the conduct of its subjects (citizens) and develop normative frameworks (policies) to direct change, and democracy, the dominant form of self-representation. The scope of the state now goes beyond traditional activities such as war and the production of laws that regulate the exercise of power over territoriality and commerce to extend to the generation of forms of subjectivity in people.

Modern statehood is a mode of political organization of 'society' in which the state is the primary agent charged with building the representation of society aware of itself and in harmony with itself. Nation-states have become the preferred mode by which all of the earth's polities organize people and territory into formally comparable and sovereign governing units. The scope of the state now goes beyond traditional activities such as war and the production of laws that regulate the exercise of power over territoriality and commerce; it also generates new forms of subjectivity in people. Anthropologists initially observed this expansion of the state by studying the elimination, transformation, or preservation of indigenous peoples, tribes, peasants, clans, and kinship systems through the creation of encompassing, though also exclusionary, national forms of belonging. One major state activity that anthropologists have identified has been the deployment of ethnographic modes of defining and categorizing human populations and of systematizing knowledge about them, especially in colonial states. Hence, the science of anthropology has followed closely and defined its objects in terms of state projects, alternately supportive and oppositional: itemizing culture traits and social organizations of discrete peoples, tracing efforts at internal and external colonization, mapping the settling and resettling of territory, and tracking diasporic populations and global trade routes.

The 'state' refers to a form of polity that links people and territory in 'societies' and thereby demarcates itself from other states. Although archaeologists date states back to ancient Egypt and Mesopotamia (about 2500 BC), this article deals only with the anthropological aspects of the modern state. It examines the implications of linking people to territory on subject formation; the administration of populations and their cultures; and on the practices of religiosity, kinship, and commerce, as well as on anthropology itself.

The Concept of Modern Statehood

Modern statehood is a mode of political organization of society in which the state is the primary agent charged with building the representation of society aware of itself and in harmony with itself. Today, the nation is the preferred form this representation assumes. It is always an idealization and always in tension with the actual forms of group organization and their heterogeneous, agonistic compositions. Modern law is the principal means by which the state attempts to create this harmony, an ideal state in which there is a coincidence between behavior and rules and where the state can anticipate the conduct of its subjects (citizens) and develop normative frameworks (policies) to direct change (Nader, 1990).

States are only one of many rule-making institutions with administrative capacities. International and supranational bodies also increasingly make law. And while states largely retain for themselves the power of legitimate enforcement, corporations, religions, nongovernmental organizations (NGOs), and other non-state actors are equally organized and rely on other sources of power to circumvent or control state law. In most parts of the contemporary world, however, states remain the institution with the most effective ideological and institutional apparatus, also recognized by other states, to claim to represent the entire people within a particular territorial jurisdiction.

Statehood became the central principle asserted in international law in 1648, with the signing of the Treaty of Westphalia ending the 30 years of war in North–Central Europe. Driven by industrial capitalism as a superior mode of production and accumulation, the state expanded its functions beyond the conduct of war, the administration of justice, and the regulation of commerce, to massive projects in the colonization of populations in the Americas, Asia, Africa, and the mid-East (Dirks, 1993; Stoler, 1995). Western European states, in particular, formulated projects of external and internal colonization, and they began to usurp religious authority's control over the definitions of transcendence, the bodies of their subjects, and the institutionalization of their life courses.

Initially employed as a political program in the French Revolution and subsequently written into the United Nations Charter, the nation form grew out of transformed empires and tribes during the process of European state formation (Tilly, 1990). It also dissolved a feudal model of society by means of cultivation of behaviors based in part on class emulation and conflict (Elias, 1982). At the same time, in a process of reciprocal copying, states in the Americas also emulated the nation form. This transformation of diverse peoples into unified nations appropriated local discourses of consanguinity and affinity and, in turn, was propped onto

and often reconstructed local kinship. Unification of nations was rarely accomplished without intermittent purges or other kinds of homogenizing processes. From the late fifteenth century to the early seventeenth century, all Western European states engaged, with often-disastrous economic consequences, in variant forms of ethnoreligious 'cleansing,' leading to population resettlement and the formation of victim groups.

In the twentieth century, national consolidations and the creation of new types of victim groups reached new heights, especially with the advent of fascist and racist ideologies originating in Europe and also employed elsewhere (Lefort, 1986). The peace treaties concluding World War I introduced into Eastern and Southern Europe the two conditions of West European nation-states – homogeneity of population and rootedness in the soil. In Eastern Europe, the implementation of these principles was largely elided during the Cold War, with the establishment of socialist states based on universal brotherhood and allied with the Soviet Union – a bloc that dissolved following the so-called velvet revolutions of 1989. Western European states emerged weakened from World War II and were unable to resist independence movements in their colonized states, primarily in Africa and Asia. These postcolonial states in turn nearly universally adopted the nation form to integrate diverse groups within territorial boundaries already drawn, arbitrarily, by European states. Nation-states have since become the mode by which all of the earth's polities are organized into formally comparable and sovereign governing units. Appropriation of the nation form effectively broadened the scope of the state to include the planned change of culture, variously defined, and the generation of forms of subjectivity in populations (Borneman, 1992).

Subject Formation through Linking Territory to People

The 1933 Montevideo Convention defined the criteria for statehood as having (1) a permanent population, (2) a defined territory, (3) government, and (4) the capacity to enter into relations with other states (Article 1, League of Nations Treaty Series No. 881) – with only the first two criteria consistently upheld. Between 1945 and 2011, the number of states recognized by international law as sovereign members of the United Nations increased from 51 to 193, with two nonmember states (the Holy See and Switzerland) and a few countries with sovereignty disputes. Defining statehood this way has numerous effects on subject formation. First, it produces pressures to unify ethnically mixed and territorially dispersed populations by centralizing legal and education systems and standardizing the life course. States uniformize by means of both bureaucratic technologies (the use of, e.g., interviews, surveys, and maps) and psychological appeals to the 'shared' symbolic universes they create (through the standardization of grammars and linguistic use, and the shaping of experience in ritual events and everyday routines). Second, it creates populations with mixed loyalties, settlers, or migrants who live in a place among peoples separate from the place and peoples with whom they primarily identify. Third, it generates categories of automatic exclusion (e.g., resident stateless peoples, resident internal 'minorities,' nonresident external others), which are then used frequently to justify war against external or internal enemies. Fourth, the dependence of modern political representation on territorial attachment reinforces identifications with territory and the desire for statehood. These identifications exacerbate territorial conflicts and often lead to wars between potential antagonists who tend to receive international recognition only when their disagreements are articulated in terms of border violations.

Despite the ubiquity of the nation-state, different historical experiences of populations have produced a variety of ways of classifying states, some corresponding to intent of the ruler, others to the relation of ruler to ruled, and still others to their cultural form. The twentieth-century conflict between 'socialist states' and 'liberal democracies' uniformized neither the state form nor the experience of citizenship, as local cultures had enduring influences on actual political forms and relations. Other factors, such as relations with colonial powers or unfriendly neighbors, independence or separatist movements, or dependent economic relations also strongly influenced national cohesion and state functioning, leading to increasing awareness that what some analysts label 'failed states.' In accounting for this variety of state forms and the historical conditions of their production, anthropologists have focused on the organization of public authority and forms of domestic control. The historian Foucault (1977, 1978) and the political scientist Anderson (1983) have inspired much recent scholarship.

The powers unleashed by science since the Enlightenment have generated what Foucault calls 'rituals of truth,' including disciplinary technologies that refigure space and the arts of governance charged with producing society and 'culture' as objects of knowledge and identification. Moreover, the deployment of new procedures of power in the nineteenth century transformed state authority from a sovereign's ability to 'take life or let live' (to order or require someone's death) to an investment in administering life by means of disciplining the body and regulating populations. States, Foucault argues, institutionalize particular cultural forms through projects of reform, education, development, incarceration, war, welfare, and family planning, along with the use of scientific techniques such as surveys, census, questionnaires, actuarial tables, economic forecasting, spatial mapping, case studies, and bureaucratic interviews. Anderson focuses on the historical conditions in which the 'imagined community' of the nation comes into being and achieves global currency, particularly in its relation to states following the collapse of empires. One of these conditions, print capitalism, enables advances in techniques of memory and collective narration through the proliferation of historicist myths and standardized accounts of experiences of national belonging across space and irrespective of local temporalities.

Because modern governing necessitates delineating a population and a territory, rulers invariably deploy ethnographic assumptions about how to define a people. As well, through laws and policies, states produce and circulate information about belonging (through families, schools, and the media); positively and negatively sanction prescriptive and proscriptive behaviors (about sexuality, gender, age, race, ethnicity, and work); and build libraries, monuments, and museums to display particular interpretations of the nation. In sum, modern

states are distinctive in that they produce both laws and people; they not only administer and regulate already existing cultural patterns but also produce new domains of objects and forms of subjectivity.

The Discipline of Anthropology and the State

Early anthropological work on the state, especially among archaeologists, was inspired above all by Friedrich Engels's study of the origin of the state and its relation to private property, Lewis Henry Morgan's evolutionary scheme, and Sir Henry Maine's theory of the progressive movement from status to contract in the development of state law (Service, 1975). Fieldwork-based anthropologists, inspired by Bronislaw Malinowski and Franz Boas, did not initially theorize the state or the nation, although their research was most frequently among indigenous peoples, tribes, peasants, and clans who were either in conflictual relations with states or involved in a process elimination, accommodation, transformation, or preservation in response to state challenges to their territory and resources.

The discipline of anthropology has closely followed and defined its objects in dialogue with state projects, alternately supportive and oppositional: itemizing culture traits and social organizations of 'discrete' peoples; devising evolutionary schemes; tracing efforts at internal and external colonization; mapping the settling and resettling of territory and the transformation of property; and tracking diasporic populations, cultural diffusion, and global trade routes (Moore, 1986). One major state activity that anthropologists have identified has been the deployment of ethnographic modes of defining and categorizing human populations and of systematizing knowledge about them, especially in colonial states (Cohen, 1996; Ferguson, 1990; Pemberton, 1994; Scott, 1998).

By and large, however, anthropological questions have revolved around the production of the cultural or social organization or forms of exchange among people considered marginal or obstacles to state projects (Sahlins, 1981). Throughout the first part of the twentieth century, a preference for fieldwork in villages or small-scale communities not fully integrated into states or supranational exchanges also made states less of a focus to the subjects of anthropological research. Further, long-term observational fieldwork among ordinary people often requires a kind of intimate access that precludes relations with rulers or those in the inner circles of power. Anthropologists in this period, therefore, assumed an essential sociocultural opposition of the people to the state (Clastres, 1987). This claim seemed credible, if not ethically imperative, when, as in the early twentieth century, the objects of study were so-called primitives and indigenous peoples, victimized and threatened with extinction, frequently through state-supported genocide, or, alternatively, when the state was primarily a puppet of a colonial power. Nonetheless, the opposition is misleading, as groups victimized by the state nonetheless must, and increasingly do, appeal to states for protections and remedies. Also, the state often provides the most powerful modality for representation of group interests.

Anthropology has in the meantime become a global discipline of alterity, no longer restricted to questions of origin or to particular kinds of peoples or places. Also, the relation of indigenous people to states has changed considerably, with some states now even guaranteeing the 'preservation of indigenous cultures' and prioritizing 'native' rights over those of more recent settlers. Moreover, the United Nations along with some NGOs and a global tourist industry positively sanction states that protect and support what are called 'first peoples' – a movement that has reversed the trend toward extinction and strengthened the importance of claiming cultural authenticity or temporal priority of settlement.

In any case, the notion of 'kin-based' or 'acephalic' polities as a pre-evolutionary stage to a 'state-based' polity, or of the replacement of status by contract, has become a moot for ethnographic research as the entire planet is organized territorially by states mutually regulated through legal contracts and unevenly integrated by global market exchanges (Moore, 1978). Anthropologists, instead, now investigate the simultaneity of the organization of kinship and the state, of various stages of developmental narratives, and of status and contract. The interaction of people with states may be an unavoidable condition of modern life, but it is by no means unambiguous in outcomes or in the trajectory of change.

This interaction took three meta-narrative forms in the second half of the twentieth century: people living within the bipolar Cold War structure of socialist or communist states (planned economies, single-party rule, so-called Second World of totalitarian states) vs liberal states (free market economies, multiparty democracies, so-called First World) and a contested zone of people living in postcolonial states (mixed economies with diverse political forms liberated from empires, so-called Third World). Until the late 1970s, anthropologists had difficulty obtaining access to people within the totalitarian states, and in other parts of the world they devoted considerable research to reconstructing the precolonial state.

In the mid-1970s, a global trend toward democratization of authoritarian states began in Southern Europe; was advanced in Latin America and parts of Asia in the 1980s; and spread to sub-Saharan Africa, Eastern Europe, and the Soviet Union in the late 1980s and 1990s, making the relation of these peoples to transforming states a subject of anthropological research (Coronil, 1997; Verdery, 1996). In 2011, the same process of challenge and destabilization of authoritarian regimes began among Arab-speaking populations in the Middle East. In effect, the space of the Second World has disappeared and the spaces of the Third and First Worlds are mixed, existing within each other, though in radically uneven proportions.

Beginning in the 1980s, the liberal democracies engaged in market reforms – dubbed 'neoliberalism' – that deregulated the finance sectors as well as many kinds of trade relations and corporate activity, and opened internal economic and cultural relations to global markets (Greenhouse, 2010). The ideology that markets and the private sector are more efficient than states and the public sector received a large bonus by the collapse of the socialist states of East–Central Europe and the disintegration of the Soviet Union. The end of the Cold War divide into East and West blocs accelerated these global exchanges across territories and on a larger scale than

previously possible, suggesting not only a need for states to regulate them but also demonstrating the limits of national law and the inadequacies of existing regulatory agencies to do so. In response to these newly extended and often global markets, the national subject, known as 'citizen,' was increasingly reduced to the identity of a consumer, which has changed the imagination of the individual's relation to the social (Dunn, 2004).

Of note at the beginning of the twenty-first century is the distinctive tensions created by the global integration of financial markets, the loss of state sovereignty over national economies, the proliferation of ethnic affirmation within states, and the denationalization (privatization and outsourcing) of much military conflict. State accommodation to globalization often compromises national sovereignty, although this varies greatly depending on the state's size, wealth, location in the international system, and its domestic legitimacy. States are under increasing global pressures to refrain from protecting local cultural forms, which, even if unable to maintain some autonomy in the face of globalization, are nonetheless not uniformized by global markets, a fact apparent both in the local variations in 'society's' relation to the state, specifically to democratic form, and to the continued patterned diversity of cultures, expressed in everyday routines and rituals. At the same time, however, one of the effects of these tensions is a dramatic rise in inequality within groups and the pauperization of large parts of populations in all parts of the world.

The vast majority of states now call themselves democracies and, therefore, operate in a specific mode of ruling and employ specific devices of power and strategies of self-representation. As a mode of ruling, this entails the introduction of regularized elections; the creation of ruling majorities and minorities; the institutionalization of an opposition; and the strengthening of principles of accountability that circumscribe the exercise of power, usually embodied in principles of 'the rule of law' (Borneman, 1997). As a device of power, however, some of these same democratic modes of ruling, such as majority rule, lend themselves to degeneration into forms of despotism. A state is often as dependent on the externalization of negative others as on facilitating an internal dynamics of citizen formation. Practices of such externalization range from state-tolerated social discrimination to legal exclusion of refugees to war against neighbors. As a strategy of self-representation, each democracy requires not only internal assent but international recognition too, which it obtains only to the extent it is able to claim governance by 'rule of the people' and conformity to international norms and conventions. When states violate such norms and conventions, which they frequently do, they risk censure, sanctions, or other forms of international pressure. Such pressure often strengthens the hand of local groups against the state.

Anthropologists engaged in research requiring listening to others are often asked to advocate in the international arena for the people studied and thus engage with all three registers of democracy. Their research often takes positions on local political struggles. As experts on culture, anthropologists are frequently asked to use legal disputes knowledge obtained in fieldwork that is critical of state modes of rule or strategies self-representation (Good, 2007). Given the increasingly global nature of 'news' and accelerated information flows across state borders, local struggles about representation quickly enter international or supranational arenas, subjecting the state to pressure from authorities external to it and explicitly drawing attention to the ways in which it exercises sovereignty. Such global flows also challenge state sovereignty by bringing to the world's attention local and subregional actors and events that might be in opposition to dominant groups within a state.

The study of conflicts in democratizing states includes research on social and legal citizenship and on the relation of state policy to demands for representation and accountability. Along these lines, four basic contemporary conflicts about representation and states are observable: involving excluded people who want more social, economic, cultural, or legal inclusion into states in which they reside; diasporic people who want representation in both the place in which they reside and the place from which they are displaced; numerical minorities (usually represented as a culture or a 'nation') who do not have their own state but want one; and citizens or social movements that demand more and various kinds of accountability from their rulers.

Legitimacy and the Modern State

Although the conception and functions of the modern state keep changing, a state's legitimacy is always based both on some notion of the 'consent' of its own citizens to rule and on the recognition of other states. This was not the case when rule was justified without recourse to 'the people,' in other words, rule based on coercion alone, on principles of heredity or wealth, the ability to wage war and collect taxes, or the administration of justice. To obtain this consent and recognition, the state must usually go beyond Max Weber's famous dictum that its power rests on a monopoly on the legitimate use of violence.

In many parts of the world, the state no longer even pretends, in fact, to hold this monopoly on violence. In the last 30 years, most states have tended to outsource and privatize many security functions. Where the state has privatized or even dissolved its former functions, the most vulnerable citizens tend to suffer the most from the lack of any protections. Indeed, today weak or 'failed' states are just as likely to lead to violent conflict as were strong imperial states in the last several centuries. In many places, the state's own abuse of this supposed monopoly – through genocide, strategic terrorism, the suppression of dissent, cynical uses of war – has traumatic effects on its citizens, leading to repeated and delayed suffering of past events. Increasingly, then, the anthropology of the state focuses on privatization, security, outsourcing, violence, vulnerability, and trauma (Masco, 2006).

Irrespective of whether they initiate or are victims of violence, states instrumentalize traumatic events, not only by building monuments to fallen heroes or soldiers but also through the construction of nationalist historiographies and support for a heritage industry that especially honors the idea of national sacrifice and organizes the process of mourning for losses after large-scale violence (Borneman, 2011). So long as

the state continues to draw legitimacy from principles of territorial sovereignty and peoplehood, it will be motivated to generate, exacerbate, and institutionalize differences with neighbors or neighboring states. Anthropologists are most focused on the effects of these conflicts and wars on the organization of populations and communities (Tambiah, 1992).

Also, the collapse of the Cold War's bipolar state system has resulted in the relatively unrestricted proliferation of arms, much of it driven by state-supported arms industries, and in a market for new and affordable technologies of death outside the control of states. Where the sources and agents of violence outside state control are increasing – a wide-ranging phenomena found, for example, in Colombia due to drug trafficking, in Sri Lanka and much of Africa due to civil war, and in the United States or Mexico due to poverty and class conflict – people employ private police forces and militias. States are also located ambiguously to many occasions of collective violence, such as genocide, at times instigating, at other times complicitous with local groups, and at others offering remedies (Ghassem-Fachandi, 2012).

Alternatively, many states now lend their militaries for 'peace-keeping' efforts in other countries organized by the United Nations, and still others, such as states within the European Union, are relinquishing military sovereignty to continental or transcontinental military alliances. There is also a growing body of legal doctrine – and an International Criminal Court established in 1999 in the Hague – that attempts to hold states accountable to their own norms and to international covenants and agreements on issues such as human rights, environmental and maritime law, and commercial law. The state's sovereignty and legitimacy, then, is not granted as natural but conferred by its citizens and by the recognition of other states and new international regimes.

Anthropologists now supplement the long history of linking the state to realization of the Idea of the people (Hegel), representation of the social (Durkheim), technical and bureaucratic rationality (Weber), or exploitation and class dominance (Marx), with an acknowledgment that most states still resort to magical practices and cultivate religiosity in order to legitimate rule (Taussig, 1997). State policies and edicts administered through 'law' often work through processes similar to magic: belief in the power of the utterance of the speaker, the performativity of speech, the linguistic rearticulation of group suffering and instrumentalization of memory, nationalistic narrativization of macro-political events, and ad hoc rationalization of failures and successes. Even in the domain of criminal law in what are considered the most rational of states – Germany or Sweden, for example – most crimes are never solved and most criminals go free. Yet, criminal law there still 'works' – that is, contributes to the legitimation of the state – largely through belief in the principles of justice that judicial inquiries, prosecutions, and trials efficaciously represent.

Likewise with respect to religion, states most successful in establishing a secular authority over local or world religions are those that recognize the significance of religiosity and incorporate religious traditions and symbolic forms into the secular rituals of political parties, elections, bureaucratic routines, and life course *rites de passage* (Gauchet, 1997; Herzfeld, 1992). Such practices are forms of group worship and they crystallize in patterns of national belonging when citizen and state narratives are brought into harmony. The struggle between the great world religions and the state to occupy and shape the symbolic forms of everyday life has not quieted, however. Secular narratives organizing life and death have proven incapable of replacing the power of myth and religious authority in accounting for the inexplicable or creating experiences of transcendence. Nonetheless, there has been a dispersion of the power over the control of religious expression, with religions also constantly subject to internal reform movements and no longer holding monopolies over local populations, and the state is one of many benefactors from this process.

In the last quarter of the twentieth century in most places of what is known as the First World, entertainment – especially involving sport and music – has emerged as the field dominating both the management of experiences of transcendence and of people's affective attachments and identifications. State rituals – especially electoral campaigns, legislative debates, memorialization and mourning ceremonies, and political speeches – have, therefore, increasingly incorporated models and techniques of authority drawn from entertainment. This leads to a fusion of actors ('personalities') in the two fields and an incorporation of magic and infusion of 'charisma' in the rationalized arena of state activity.

With the end of the Cold War, First World states increasingly structured their relations to new and Third World states by promoting NGOs. Acting in a negative symbiosis with the state, such NGOs also incorporate magic into their missionary projects of rational reform and economic and political development. In much of the world, they have become strong competitors to organized religion and to the state in providing care (welfare), in development projects, and promoting both local and global accountability in the exercise of power. Often, however, they serve primarily as instruments for the transfer of wealth or power between supranational elites.

Notwithstanding liberal mistrust of the modern state, reaffirmed by the horrors of totalitarian experiments, the nation-state form still provides the possibility for local control of people's own destinies in a manner impossible for prenational communities. Stateless people are indeed the most defenseless actors in the contemporary world. Hence, despite the wars, violent revolutions, and massive repressions of the nation-state, as the sociologist Lepsius (1988) argues, this form of polity has offered particular advantages to West European societies, including the institutionalization of peaceful conflict-solving through the rule of law, guarantees of individual freedom, the organization of interests through parliamentary democracies, and the integration of national economic development in the world economy. In each of these measures, authoritarian and totalitarian states elsewhere have been at a permanent disadvantage, incapable of recognizing the subjects they seek to control for whom they are and unable to organize power, except through coercion. Such states destroy sociality and build false representations of the people, thereby creating a basis for the illusion of society against state.

See also: Citizenship: Political; Governmentality; Kinship in Anthropology; Law: Anthropological Aspects; Nationalism: General; Political Anthropology; Postcoloniality; Socialist Societies: Anthropological Aspects; State Formation, Theory of; State, History of; State, Sociology of The; States and Civilizations, Archaeology of; War: Anthropological Aspects, Historical Development of.

Bibliography

Anderson, B.R.O'.G., 1983. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, London.

Borneman, J., 1992. Belonging in the Two Berlins: Kin, State, Nation. Cambridge University Press, Cambridge, UK.

Borneman, J., 1997. Settling Accounts: Violence, Justice, and Accountability in Post-socialist States. Princeton University Press, Princeton, NJ.

Borneman, J., 2011. Political Crime and the Memory of Loss. Indiana University Press, Bloomington.

Clastres, P., 1987. Society against the State: Essays in Political Anthropology. Zone Books, New York.

Cohen, B., 1996. Colonialism and Its Forms of Knowledge. Princeton University Press, Princeton, NJ.

Coronil, F., 1997. The Magical State: Nature, Money, and Modernity in Venezuela. University of Chicago Press, Chicago.

Dirks, N.B., 1993. The Hollow Crown: Ethnohistory of an Indian Kingdom. University of Michigan Press, Ann Arbor, MI.

Dunn, E., 2004. Privatizing Poland: Baby Food, Big Business, and the Remaking of Labor. Cornell University Press, Ithaca, NY.

Elias, N., 1982. State Formation and Civilization. Basil Blackwell Ltd, London.

Ferguson, J., 1990. The Anti-politics Machine: 'Development', Depoliticization, and Bureaucratic Power in Lesotho. Cambridge University Press, Cambridge, UK.

Foucault, M., 1977. Discipline and Punish: the Birth of the Prison (A. Sheridan, Trans.). Vintage, New York.

Foucault, M., 1978. The History of Sexuality (R. Hurley, Trans.). Penguin, London.

Gauchet, M., 1997. The Disenchantment of the World: a Political History of Religion (O. Burge, Trans.). Princeton University Press, Princeton, NJ.

Geertz, C., 1980. Negara: the Theatre State in Nineteenth-Century Bali. Princeton University Press, Princeton, NJ.

Ghassem-Fachandi, P., 2012. Pogrom in Gujarat. Princeton University Press, Princeton, NJ.

Good, A., 2007. Anthropology and Expertise in the Asylum Courts. Routledge-Cavendish, London.

Greenhouse, C. (Ed.), 2010. Ethnographies of Neoliberalism. University of Pennsylvania Press, Philadelphia.

Herzfeld, M., 1992. The Social Production of Indifference: the Symbolic Roots of Western Bureaucracy. Berg, New York.

Lefort, C., 1986. The Political Forms of Modern Society: Bureaucracy, Democracy, Totalitarianism. MIT Press, Cambridge, MA.

Lepsius, M.R., 1988. Interessen, Ideen und Institutionen. Westdeutscher Verlag, Opladen.

Masco, J., 2006. Nuclear Borderlands: the Manhattan Project in Post-cold War New Mexico. Princeton University Press, Princeton.

Moore, S.F., 1978. Law as Process: an Anthropological Approach. Oxford University Press, Oxford.

Moore, S.F., 1986. Social Facts and Fabrications: 'Customary Law' on Kilimanjaro, 1880–1980. Cambridge University Press, Cambridge, UK.

Nader, L., 1990. Harmony Ideology: Justice and Control in a Zapotec Mountain Village. Stanford University Press, Stanford, CA.

Pemberton, J., 1994. On the Subject of Java. Cornell University Press, Ithaca, NY.

Sahlins, M.D., 1981. Historical Metaphors and Mythical Realities: Structure in the Early History of the Sandwich Islands Kingdom. University of Michigan Press, Ann Arbor, MI.

Scott, J., 1998. Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven, Ct.

Service, E.R., 1975. Origins of the State and Civilization: the Process of Cultural Evolution. Norton, New York.

Stoler, A.L., 1995. Race and the Education of Desire: Foucault's History of Sexuality and the Colonial Order of Things. Duke University Press, Durham, NC.

Tambiah, S.J., 1992. Buddhism Betrayed? Religion, Politics, and Violence in Sri Lanka. University of Chicago Press, Chicago.

Taussig, M., 1997. The Magic of the State. Routledge, New York.

Tilly, C., 1990. Coercion, Capital and European States AD 990–1990. Blackwell, Oxford, MA.

Verdery, K., 1996. What Was Socialism and What Comes Next? Princeton University Press, Princeton, NJ.

State Formation, Theory of

Isebill V Gruhn, University of California, Santa Cruz, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14970–14972, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Most social scientists would agree that the state consists of a set of ongoing institutions for social control and authoritative decision-making and implementation. Beyond this basic description, the concept of the state becomes multidimensional. The modern state emerged in Western Europe in the fifteenth–sixteenth centuries and became more firmly grounded in the seventeenth century. Its development is intimately linked to the political, economic, intellectual, and religious developments of the time. Theorists differ about the dynamic of the formation of the state and its consequences for domestic and international affairs over given historical periods. Because state formation and reformation is continuous, it is useful to examine how and why state formation took shape among the 'old' states, as well as how the 'new' postcolonial states address state formation in the light of contemporary exigencies. Today state formation and reformation vies with the forces for regional integration (Europe) and globalization, brought on by global commerce and the revolutions in communication and technology. State formation continues to be a work in progress.

A state consists of a set of institutions whose functions are social control and authoritative decision-making and implementation processes. In addition, modern states seek autonomy in their relations with other states. Beyond these very general criteria, the concept of the state quickly becomes multidimensional with respect to its historical development, its relevant legal considerations, and its theoretical debates. The institutions of a state are subject to conflict and disagreement over their structure and the role they play. Institutions change over time, and the changes both shape and reflect differing degrees of autonomy and strengths. In the twenty-first century there is considerable interest not only in the historical trajectory of state formation and the more proximate formation of new states – dozens of new states have been created in the second half of the twentieth century – but also in the concept of state capacity. Political scientists and political economists who measure and assess the strength and relative autonomy (sovereignty) of states in national and international arenas concern themselves with the proper role of states in the domestic and international political economy. States are in constant competition with other institutions such as intergovernmental institutions and for-profit and nonprofit institutions. Political scientists study and observe the role and function of each of these various national and international institutions. Not surprisingly, the proper role, the function, as well as the capacity of today's states has become a lively and contentious issue.

The Problematic of the State

Many prominent classical theorists studied state formation and the multifaceted issues surrounding the role and function of states. Theorists such as Machiavelli (*The Prince* 1988), Rousseau (*The Social Contract* 1987), Marx and Engels (*The Communist Manifesto* 1998), and Lenin (*State and Revolution* 1993) considered the formation and role of the state to be a theoretically interesting issue. More recent students of the formation of the modern state have focused their attention on such other topics as the absolutist states of Western Europe and states and revolution (Skocpol, 1979). Contemporary writers address issues such as the state and industrialization, nationalism, colonialism, and the formation of the state, in addition to many other topics concerned with communism and the postcommunist state. Other ongoing themes among students of state formation are the state and the economy, power and legitimacy, and the relationship of states to other states. Twenty-first century analysis of the forces of globalization and their impact on the state are also frequently discussed in the context of the historical evolution of state formation.

History of the State

The modern state is a Western European creation which emerged in the fifteenth and sixteenth centuries and became more firmly grounded in the seventeenth century. The modern state came into existence in the same time frame and in geographical locations that were the wellsprings of modern capitalism, modern science, and philosophy. Many scholars link the evolution of the modern state and Christianity in its Protestant form. It is generally agreed that the emergence of a new political form – the modern state – was intimately linked to the economic, intellectual, and religious developments of the time.

The first modern states to arise were Italy, the UK, France, and Spain. The principal technical tools of modern state building were the development of permanent bureaucracies and a standing army, but it was the availability of money that enabled rulers to acquire them. Once a ruler was able to control the professional bureaucracy and the army, the nobility of medieval times lost its political role and power. The principal producers of cash were the commercial and industrial middle class and the peasantry. By the late eighteenth and early nineteenth centuries the bourgeoisie was able to curtail and even

abolish the privileged status of the aristocracy and the authority of the monarch. The bourgeoisie acquired increasing control over the state. Their victory, signaled by the French Revolution, transformed the dynastic, state into the modern state. However, to define the modern state as 'bourgeois,' meaning 'civic' signifies the civil and legal equality of all subjects regardless of status by birth or occupation. The abolition of the old feudal privileges did not mean that the propertied and educated classes ceased to have privileges in the modern state. Modern theorists of state formation like Charles Tilly (Tilly, 1975) and many Marxist theorists like Miliband (1969), Poulantzes (1969), and Offe (1975) view the state as 'organized coercion' orchestrated by an elite. Other social scientists view the state as the arena of legitimate authority embodied in the rules of politics and in governmental leadership and policies.

Some thinkers, especially Marxists, tend to treat the state as an analytic construct by which to explain the modes of production, class relations, and societal struggles in general. Others, including Theda Skocpol (1979), argue that states are 'actual organizations' that control territories and peoples. Skocpol echoes the legal definition of a state much more closely than does Tilly for example, by pointing out that while the state is anchored in class-divided socio-economic structures it also needs to be seen as part of the international system.

As a practical matter, in order for a state to engage in international relations, other states need to recognize it as a fellow state. Modern sociologists call attention to the fact that a state is a human community that successfully claims the monopoly of legitimate use of physical force within a given territory. Whether a state is considered legitimate and on what basis its legitimacy is acquired and honored in the human community within the territorial boundaries of the state and in the large international community of fellow states, is to say the least, very historically and theoretically complex.

The Modern State

The modern state and the modern state system developed simultaneously. It required coercive capacity for the state to sustain itself both domestically and internationally. The state system of the nineteenth and twentieth centuries led to two world wars and the Cold War. These wars were significant in shaping contemporary states and the international system of states. However, contemporary concern and interest in the state is not confined to questions of the legitimacy of states or even their coercive capacity. At issue in the twenty-first century is the role of states as actors, competing with other actors, to determine what the ongoing proper role and function of a state ought to be with respect to its political, social, cultural, and economic obligations to its citizens.

State formation has never been simply a 'founding.' Quite the contrary, state formation is a process of formation and reformation based on the changing nature of societies within the state and the larger international system of states. The institutions and processes of a state and of states collectively are not static. To say that a state requires the capacity to carry out the proper functions of a state is a truism, but the capacities required at the turn of the twentieth century vary greatly from state capacities required at the turn of the twenty-first century. Furthermore, states compete with other national and international institutions. For example, states vie with interstate organizations, such as the European Union, the United Nations (UN), the World Trade Organization (WTO), and with nonprofit organizations such as Greenpeace and Oxfam, and for-profit corporations such as Microsoft and CNN. Financial capacity, knowledge capacity, monetary capacity, and even legal capacities of non-state actors all challenge state capacity. This is not to suggest that states have become defunct but merely that what it means to be a state today and what capacities a state is expected to possess by its citizens and in the larger community of states continues to go through a process of formation and reformation.

The Future of the State

As the twenty-first century begins there are more than 200 so-called sovereign states in the world. It is useful to think of contemporary states in their subgroups of differing state types: first, the 'old states,' such as the UK, France, Spain, the US, China, and Japan, which have developed reasonably stable nation states over the centuries. Some of these states, like the US for example, are multinational and multiethnic. Others, such as Japan, are more homogeneous. What these states have in common is a citizenry that views the national government as legitimate, and these states on the whole, share stable political economies. Second, there are the 'new' states. Many of these states, for example, most of the sub-Saharan African states and many states in Asia and the Middle East, emerged out of colonialism during the twentieth century. These new states tend to meet the legal criteria for statehood, yet many lack legitimacy as states in the eyes of their citizens. Most often such states are also multiethnic and have yet to achieve a sense of nationhood – a shared belief on the part of citizens that they belong to the nation-state. Some of these states, for example, Somalia or the Congo, are, at the beginning of the twenty-first century, in a state of civil conflict. States which are not able to carry out the basic function of a state, are frequently referred to as 'failed' states. Such states have the legal trappings of a state but the nation-state formation has broken down and the state's capacity to govern its territory has failed. In the category of failed or failing states are all those for which the historical process of state formation has not yet succeeded.

One may observe that the historical timeframe during which state formation takes place will determine both the internal and external factors which have a bearing on the process, as well as the outcome, of state formation. In the post-Cold War era for example, state formation and reformation – in the case of former communist states – is taking place at a time when the dominant paradigm is a democratic form of governance within the context of free market capitalism. Accordingly, state formation in the new states and postcommunist states, takes place in an international setting which exerts a distinctive global leverage on the type of state formation for which political and economic support and acknowledgement is rendered.

However, today state formation and the reformation of all states, not just new or failing states, takes place in the context of globalization. Globalization brings with it active participation

in the national and international life of states by nonstate actors and a new degree of interstate and inter-actor interdependence. Today global and regional institutions, along with states, participate in rule making, rule enforcing, knowledge creation, and in fields such as commerce, banking technologies, medicine, genetic engineering, and in defining the standards for human rights. The globalized world of the twenty-first century is reshaping the idea and the reality of the sovereign state. Yet states still are extremely important in the lives of most people and hence, state formation continues to be an important ongoing process, a process which is not merely a multifaceted historical process but a multidimensional work in progress.

Editor's Note

Since Gruhn's article was first published, the study of comparative state formation has exploded in a resurgence of the 'bringing the state back in' movement in the 1980s. A useful recent attempt to take stock of this growing literature is Vu, 2010. "State failure" is a key topic in this literature; see Eriksen, 2011. Another useful current source is Chase-Dunn and Inoue, 2012.

See also: State, History of; State, Sociology of the; State: Anthropological Aspects.

Bibliography

Chase-Dunn, Christopher, Inoue, Hiroko, 2012. Accerlating democratic global state formation. Cooperation and Conflict 47 (2), 157–175.

Eriksen, Stein Sundstøl, January 2011. 'State failure' in theory and practice: the idea of the state and the contradictions of state formation. Review of International Studies 37 (01), 229–247.

Habermas, J., 1976. Legitimation Crisis. Heineman, London.

Lenin, V.I., 1993. State and Revolution. Penguin, USA, New York.

Machiavelli, N., 1988. The Prince. Cambridge University Press, Cambridge, UK.

Marx, K., 1983. The eighteenth Brumaire. In: Held, O. (Ed.), States and Societies. New York University Press, New York.

Marx, K., Engels, F., 1998. The Communist Manifesto. Signet Classics, New York.

Miliband, R., 1969. The State in Capitalist Society. Weidenfeld and Nicolson, London.

Nozick, R., 1974. Anarchy State and Utopia. Basic Books, New York.

Offe, C., 1975. The theory of capitalist state and the problem of policy formation. In: Lindberg, L.N., Alford, R., Crouch, C., Offe, C. (Eds.), Stress and Contradiction in Modern Capitalism. Lexington Books, Lexington, MA.

Poulantzas, N., 1969. Classes in Contemporary Capitalism. Weidenfeld and Nicolson, London.

Rousseau, J., 1987. The Social Contract. Penguin USA, New York.

Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, Cambridge, UK.

Strange, 1996. The Retreat of the State. Cambridge University Press, Cambridge, UK.

Tilly, C. (Ed.), 1975. The Formation of National States in Western Europe. Princeton University Press, Princeton, UK.

Vu, Tuong, 2010. Studying the state through state formation. World Politics 62 (1), 148–175.

State, Geography of

David Storey, University of Worcester, Worcester, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by L. Staeheli, volume 24, pp. 16435–16439, © 2001, Elsevier Ltd.

Abstract

The state is the world's dominant form of political organization and it reflects a territorial configuration that we tend to take for granted. The focus here is on its geographical dimensions and the article provides an overview of the origins of the state, and focuses on its three key and interconnected geographic dimensions – territory, boundaries, and sovereignty. Drawing on the work of key political geographers it indicates the complex intertwining of ideas of territory and sovereignty and highlights the evolving, contingent, and contested nature of state power and geographical control. While the state might be viewed as a territorial container, it is clear that some states exercise more effective and spatially extensive power than others. Theories of the state suggest it is much more than a mere spatial container, but rather a space in which dominant interests are pursued.

The state is currently the world's dominant form of political organization. Political maps of the world display this territorial configuration in a way that leads us to view it as 'natural' and this impression of solidity engenders a very state-centered view of the world. However, the state system is a political and geographic construct that displays considerable dynamism. In recent years the number of states has risen dramatically, consequent on state collapse (many associated with the fall of communism) and secessionist nationalism in various parts of the world. The changing spatial configuration of states, combined with various processes that suggest both a shifting role for the state and a possible diminution of its power and significance, furthers debates surrounding states, their origins, key features, role, and functions which this article endeavors to draw together.

Origins and Features of States

The focus here is on the modern state system, usually seen as deriving from the Treaty of Westphalia in 1648 which is generally held to have given rise to a territorial apportionment of land between European rulers, though the manner in which this evolved was far from uniform. States emerged through the construction of territories which could be measured and mapped, and could consequently be controlled, apportioned, and bordered (Elden, 2007). In considering how states have come into existence, geographers and others have been instrumental in developing theories of state formation and evolution. The German geographer Friedrich Ratzel likened the state to a living organism arguing that states would expand territorially outward in order to increase their size and power, swallowing smaller surrounding states in a struggle for space (Muir, 1997). In this way a form of social Darwinism is used to explain political processes in territorial terms. Population growth led to pressures which required the acquisition of more territory or extra living space – what Ratzel termed *lebensraum*. Another German geographer, Karl Haushofer, utilized and developed Ratzel's ideas of *lebensraum* prior to Hitler's use of the concept as a justification for German territorial expansion. These 'imperial' political geographers' ideas were bound up with the influence of both physical features, such as mountains and rivers, and human geographical considerations on politics and their role in affecting the strategies of states (Kearns, 2009). Other geographers explored the state in functionalist terms. Richard Hartshorne (1950) highlighted the competing centrifugal and centripetal forces to which states are seen to be subject. The first might include such things as regional inequalities, ethnic diversity, or language differences that might act as an incentive toward secession. Centripetal forces, on the other hand, serve to bind the state together and might include elements such as a common language or religious homogeneity.

While ideas of state evolution and internal coherence have some merit, they imply that states develop as part of some inevitable process. Although internal differences may have a significant bearing on the stability of a state, relative power balances between states need to be taken into account when attempting to understand the evolution of a world state system. Flint and Taylor (2007) utilize Wallerstein's world systems theory through which states are seen within the context of a broader global capitalist system. While it may be tempting to view the state as a set of institutions freely entered into, this is usually very far from the case and states tend to emerge out of violence and involve the imposition of control (Lefebvre, 1991). The creation, maintenance, and vitality of states have much to do with the balance of power between countries, rather than being simply reducible to internal characteristics. States emerge within a state system, not independently of each other.

The state is not a single unambiguous entity but a constellation of various elements and component parts. While precisely what it is may be hard to pin down, the modern state is dependent on three essential, closely entwined, and intrinsically geographical features: territory, boundaries, and sovereignty.

Territory

While the state is a political organization, it must have a territory, a portion of geographic space (land, air, water) over which it exercises power. Territory is a clear necessity for a state to exist and also provides the state with a seat of power and

a functional space in which to operate. States may also deploy internal territorializations in the form of regional units with varying degrees of devolution of power. At an elementary level we can say that states are territorial containers (though increasingly leaky ones) which have jurisdiction (or at least claim it) over politics, economy, society, and culture within their acknowledged geographic spaces (Taylor, 1994). However we need to be mindful of what Agnew (1994) has termed the 'territorial trap' whereby we effectively naturalize the state as a sovereign territorial entity.

It is easy to take territory for granted but, as Stuart Elden (2013) suggests, territory is more than simply land (or geographic space); our ideas of territory are dependent on ways of measuring and mapping it. Without developments in both cartography and geometrical means of areal calculation we would not think of territory in the ways we do. These advances brought with them legal and administrative developments and states emerged alongside bureaucratic systems that codified territorial control. Information gathering on the nature of the land, its features, and resources, as well as on people through censuses and related surveys, became essential components in state building. These led to a way of conceiving geographic space through which it could be owned, controlled, and administered. Mapping, which has generally been bound up with military, political, and economic imperatives, serves to make territory visible, accessible, and governable; as Elden (2013) argues, mapping is not just a tool of power, but a fundamental requirement of it. In addition the state serves to inculcate a sense of territorial or national identity with cartographic techniques appropriated for practical, symbolic, and propaganda purposes through which the state is rendered legible (Escolar, 2003; Scott, 1998). Territory itself results from the processes of state evolution and the mechanics of mapping and calculating space. In a sense then territory only exists because the state does. Territory is much more than a feature of the state in that it is both a facilitator and a consequence of state action (Brenner and Elden, 2009).

Boundaries

If states have control over designated territory, it follows that there must be recognized boundaries separating their own territory from that belonging to neighboring states. Clearly, these boundaries may be contested and may change or disappear over time. The rise of the bounded state as a political unit necessitated a concern with the drawing and redrawing of political borders and the formalization of territorial arrangements. Borders are places where the distinction between 'our' territory and that of others is made clear with the paraphernalia of the state very evident through flags flying at crossing points, passport checks, and police controls. Traditionally, political geographers have distinguished between 'natural' and 'artificial' boundaries. The former category encompasses naturally occurring phenomena, such as rivers and mountain ranges. The Rio Grande forms part of the border between the United States and Mexico, while the Mekong forms part of the boundary between Laos and Thailand. Lines of latitude such as the 38th parallel separating North and South Korea or the 49th parallel dividing the USA and Canada are examples of what are depicted as artificial borders. However, this binary distinction is somewhat misleading given that borders, whether utilizing naturally occurring features or not, are intrinsically artificial. States are human creations and, as a consequence of political decisions, some rivers become borders (or cease to be borders) while others do not. These sorts of boundary classifications are also predicated on an assumption that borders must exist, the only debate being over where the lines should be drawn. It is assumed that such things as ethnic or national identities are fixed and unproblematic and that territorial divisions separating them are 'natural,' desirable, and attainable (Fall, 2010).

Rather than a focus on boundary types, an exploration of the impact of borders and the nature of border disputes may be more worthwhile. Borders are constantly subject to change resulting from disputes over territory. These disagreements may relate to the precise location of the border or they may center on the very existence of a border in the first place. There are numerous examples of shifting borders across Europe including Germany's borders with most of its neighbors. Many borders are subject to ongoing dispute such as those between Venezuela and Guyana, and Surinam's borders with its various neighbors. In the Balkans the border between Macedonia and Kosovo has not been formally agreed and until recently was not an international border at all as both entities were part of Yugoslavia. The relative significance of certain borders changes over time. The Italy–Yugoslavia border was once of huge political import separating a European Union (EU) member state from communist Yugoslavia. Now Italy's border with Slovenia (one of Yugoslavia's successor states) is comparatively insignificant with relatively free movement between two EU members.

While the precise location of a border may give rise to conflict, in other instances its very existence is in dispute. The political conflict in Ireland revolves around the existence of the border between the Republic and the North. For Irish republicans, who wish to see a unified Ireland, the border is at the heart of the conflict, viewed as an imposed and artificial boundary whose removal is essential.

Borders are not just lines dividing territory; they are discursive constructs with important social as well as political ramifications (Paasi, 1999). Borders often cut through areas which are culturally homogenous but the creation and existence of formalized borders may lead to significant differences on opposite sides (Rumley and Minghi, 1991). In many respects, the imposition of a physical border can result in the emergence of a partitionist mentality contributing to the essentializing of national identity and accentuating linguistic and cultural divisions. Cold War boundaries, such as that between Austria and Hungary, had profound impacts on those living in close proximity to them (Meinhof, 2002). Political decisions taken in remote political centers can have dramatic effects on people who find their lives disrupted by shifting borders. The demise of the Soviet Union has had serious repercussions for certain groups such as the Lezghi in Dagestan (part of the Russian Federation) and Azerbaijan who, like others, now find themselves divided by an international boundary, leading to difficulties in maintaining a coherent identity (Birch, 2007).

In a myriad of ways, borders become elements within people's everyday lives and shape their day-to-day being. Their significance stretches from the global importance attaching to

the Berlin Wall in the later part of the twentieth century, through to the local but very real separation of people in cities such as Nogales and other towns and cities along the US–Mexico border. In addition, it is worth pointing out that for many the existence of state boundaries is, in some senses at least, largely meaningless. Many nomadic groups such as Bedouin people living in Iraq, Jordan, and neighboring states tend to ignore (or at least try to) what to them are meaningless lines on a map. Similarly, gypsies in many European countries may regard state borders as of no great significance in terms of identity, having no strong allegiance to the state in which they live. However, such groups may find themselves better treated in some states than in others, while borders often act as impediments on their ability to move freely between states. For those living in regions such as Kashmir or the Ferghana Valley in Central Asia, their lives may be heavily circumscribed by these imposed divisions of political space (Megoran, 2006).

Sovereignty

In addition to territory and the demarcation of bounded space, there is one vitally important additional feature which states must possess: sovereignty. This is a much debated concept but at its heart it refers to the authority of a state to rule over its territory and the people within its borders, without external interference. Sovereignty implies the existence of geographic space over which control is exerted and the emergence of ideas of sovereignty is intimately bound up with ways of conceiving territory (Elden, 2013). The global political map suggests a world of neatly demarcated sovereign spaces but, in reality, sovereignty is contingent, contested, and evolving. In an interstate system, international recognition becomes crucial and sovereignty is dependent on its acceptance by external parties. While a number of pseudostates exist (such as South Ossetia and Transnistria), they lack any widespread international recognition. The attainment of sovereignty depends not just on the nationalist aspiration for a separate state but also on the complexities of international relations (Harris, 2009).

Sovereignty disputes often derive from disagreements over where borders lie. Eritrea attained independence in 1993 following a protracted 30-year struggle to secede from Ethiopia. Both countries continue to claim sovereignty over stretches of disputed territory. The issues associated with enclaves and exclaves (areas belonging to a state but spatially separate from it, being embedded within the territory of another state) also bring with them questions of sovereignty. Although Nakhichivan is a landlocked entity lying between Armenia and Iran, it is neither independent nor a part of either of those states; instead it is an exclave of Azerbaijan. While the region of Nagorno-Karabakh lies entirely embedded within the territory of Azerbaijan (which technically retains sovereignty over it), with a predominantly Armenian population, it has been the center of a long-running dispute, declaring itself a republic in 1991 (though it lacks international recognition).

The US-led intervention in Afghanistan, Iraq, and elsewhere in recent years highlights the complexity of sovereignty in the contemporary era where global powers appear to exert significant and explicit power over territory well beyond their own state borders. Linked to this are more microlevel spaces where 'normal' rules of sovereignty appear to be suspended. The ambiguous territorial status of the US prison camp at Guantánamo Bay (in Cuba) has proven an effective excuse in allowing the indefinite detention without trial of its inmates who find themselves subject to interrogation by the US military authorities (Reid-Henry, 2007). This and other related practices render the relationships between territory and sovereignty ever more complex and blurred (Elden, 2009; Gregory, 2007).

Geographer John Agnew (2005) has attempted to move the sovereignty debate further by positing the idea of sovereignty regimes, referring to the capacity of states to exercise sovereignty. The effective sovereignty of many states is reduced through various processes and global flows, alongside the widening range of nonstate actors who might be seen to exert power across networks rather than over rigidly bordered territory. Part of the issue here is that sovereignty has never been as 'pure' or absolute as some might aver. Agnew identifies classic, imperialist, integrative, and globalist sovereignty regimes. The first might be seen as the 'traditional' view of the state operating within a bounded territorial space; the second alludes to situations where states are relatively weak and beholden to outside forces; the third refers to situations where there might be said to be a pooling of sovereignty though trans-state arrangements. (The EU is probably the most obvious example where absolute sovereignty is wielded neither by the overarching body nor by its constituent members, but rather reflects a form of pooled sovereignty operating hierarchically.) The globalist regime refers to ideas of trans-state sovereignty in an interconnected world but in which some states have considerably more influence than others and where their global reach extends well beyond their own boundaries (as exemplified by the example of the USA mentioned above). Sovereignty, for some, is contingent on adhering to particular ground rules laid out by the more powerful. External interventions by Western powers in a range of regions (including parts of the Middle East and Africa) serve to demonstrate that the territorial sovereignty of some weaker states may be seriously constrained. Some states display a willingness and ability to exert power well beyond their formal territorial boundaries (in pursuit of national strategic interests) while others (sometimes labeled 'failed' or 'rogue' states) become arenas for external intervention. The complex linkages between state sovereignty and territorial control lead to a consideration of effective sovereignty wielded by both states and a range of quasi-state actors and deployed in a variety of territorial contexts (Elden, 2009).

An additional element here is the processes associated with globalization (or linked together under that broad umbrella term), which are seen by many as signaling diminishing sovereignty for the territorial state. Borders are commonly depicted as increasingly porous in a world of transnational flows of people, goods, finance, and ideas. The emergence of regional economic alliances, epitomized by the evolution of blocs such as the EU, is heralded by some as the successor to the state system where boundaries are assumed to have diminishing significance. The financial crisis in recent years points to a considerably reduced ability of the state to maintain economic sovereignty. However, predictions of the demise of the state and the advent of a borderless world seem somewhat exaggerated. While a neoliberal economic orthodoxy (and its attendant crises) centered on trans-global economic flows seems to point toward the irrelevancy of borders, it

simultaneously utilizes and reconfigures them in attempts to increase the returns to capital (Harvey, 2010). In addition, there is the apparent paradox that while capital flows relatively freely across international frontiers, migrants (particularly those from poorer regions) continue to be faced with the intimidating bureaucratic and security paraphernalia of the state at border crossings. Ultimately sovereignty is relative, contingent, and never complete. It is always an ideal aimed toward, rather than a fully achieved condition.

Theorizing the State

In considering the role of the state it might be argued that it performs a number of key functions. It endeavors to regulate the 'national' economy, it provides public goods such as health and education, it maintains legal and other frameworks designed to guide and regulate its citizens' behavior, and it performs a policing and military function in defending its territory and its people against external aggression and internal threats. Viewed more broadly it can be suggested that the state ensures (or claims to ensure) representation, security, and welfare for its citizens. In many countries the period since the late 1970s has witnessed policies of denationalization whereby many state services have been privatized. This is part of a broader neoliberal economic agenda whereby the state's role is seen as a regulator of services (at best) rather than a provider. This has led to the 'hollowing out' of the state as private sector agents now carry out functions previously seen as the preserve of state agencies. In essence what has commonly been conceived of as the 'welfare state' has been eroded with a scalar shift to a heavier emphasis on voluntary and community responses to welfare provision. While 'rolling back the state' can be said to have had a definite effect in reducing the amount of direct contact people have with 'agents' of the state its influence remains pervasive. Even where service provision is by private companies these are usually subject to some form of regulation and the state continues to permeate our lives in a myriad of ways, evident in both our public and private dealings. Despite its seemingly reduced role the state continues to manage a highly regulated framework through which civil society is territorialized (Mann, 1984).

However, there is a need to consider the broader role of the state. Historically states have evolved from largely centralized entities with little popular participation through the development of electoral systems and gradual extensions of the franchise, thereby extending formal political engagement downward from an elite to include ever larger swathes of the populace. While it is tempting to see this as a form of natural evolution it is important to remember that most extensions of the democratic system have resulted from intense and often violent social and political struggle resulting in achievements such as voting rights for women, for example. At a simplistic level there is a tendency to think of the relationship between the state and its citizens as a freely entered into set of arrangements. While this is clearly not the case in authoritarian states, even in so-called liberal democracies there is a need to delve beneath the surface appearance of the state in order to more fully understand its role and functions.

From a pluralist perspective the state within liberal democracies is viewed as a neutral arbiter, above, and separate from, any vested interest. It is seen as an institution shaped according to the citizens' will, reflecting the basic democratic underpinnings of liberal democracy. Power emanates (via the electoral process) from the people who are deemed to be active participants in the processes of societal regulation. From this, it follows that governments merely act as the agents of the electorate. This classic liberal view of the state sees it as the guarantor of the rights of the individual. Pluralist states are seen as the polar opposites of totalitarian ones. Instead of one imposed view, there is a reconciliation of a variety of diverse views. Although at any one time a particular political party may hold power, and although they may represent a particular vested interest, the nature of the pluralist state is such as to ensure a system of checks and balances that safeguards the citizenry against abuses of power. In this way, an independent judiciary, media, and other elements are seen as integral components of the state apparatus. No one faction can have total control and states are seen as neutral and responsive to the various pressures within society. The state delivers the policies and services needed by society and is thus seen largely as a technical apparatus administering its territory for the benefit of its citizens rather than an overtly political entity. States may exhibit what appear to be healthy signs of democracy when judged on a procedural basis reflected in such things as elections and judicial processes. This leads to an assumption that the state is highly democratic and has not been 'captured' by any vested interest.

An alternative view is that, rather than the state being neutral, it acts instead in the interests of specific sections of society. It can be argued that specific elites tend to emerge whose perspectives hold more sway than others. Marxist theories of the state suggest a division along class lines with a basic distinction between a powerful class and a relatively powerless one and the state is seen as a mechanism which acts in defense of that structure. Far from being a neutral mediating party between various vested interests, the state acts in the interests of the prevailing capitalist system. Viewed this way, the state is not an independent structure standing apart from the rest of society, but is deeply embedded within socioeconomic relations (Held, 1989). It defends and protects the class structure by preserving existing power imbalances and it manages conflict so as to ensure the reproduction of the existing social and territorial order.

Despite outward appearances, states are not necessarily the open democracies they purport to be. The underrepresentation of certain strata in society within the arenas of power is one aspect of this (Miliband, 1969). More broadly, the state is constrained, being integrally embedded within a global capitalist system (Poulantzas, 1969). The collapse of communism in Eastern Europe witnessed the reconfiguring of political and state systems and the hurried implantation and global extension of a neoliberal economic regime. The current global financial crisis provides ample evidence of the nature of the state as a tool of capitalist regulation. The banking collapse of 2008 (consequent on a regime of limited state regulation) has been followed by government bailouts in various countries in order to prevent the total collapse of the banking system so that the architecture of the global financial system has remained

largely in place with the state–finance nexus as the 'central nervous system' of capitalist accumulation (Harvey, 2010). Rather than a disinterested and disembodied regulator, the state can be conceived of as an agent that reproduces society, culture, and polity. The administrative and territorial apparatus of the state and its seemingly banal bureaucratic procedures tend to mask its more overt political role, thereby serving to reproduce it (Sharma and Gupta, 2006).

When judged in more substantive terms the outcomes of the enactment of democracy may fall short of what the procedures suggest in terms of inequality and various forms of social and spatial exclusion (Bell and Staeheli, 2001). Despite this the state remains remarkably resilient. Mann (1984) argues that states wield despotic power (reliant on coercion) and infrastructural power (based on forms of negotiation). The first of these refers to institutions such as the police, army, and judiciary, which serve to keep people in line in a very obvious manner. The existence and power of this coercive apparatus is readily apparent in totalitarian states, yet it is no less important in liberal democracies, as evidenced periodically in the policing of protest marches and so on. However, of considerably more importance, especially in liberal democracies, is infrastructural power. In Gramscian terms there is a widely accepted set of 'common sense' propositions that are rarely questioned (1971). This acceptance of a particular way of viewing things is enabled through a variety of filters. Most obvious among these are education and the mass media. It is argued that these act as 'gatekeepers' of ideas, serving to permit discussion of some topics while largely ignoring others. In helping to shape public opinion, they can have a hugely conservative impact, protecting the state against potential opposition. The dominant ideas in society come to be seen as 'common sense,' constituting what Foucault (1980) referred to as 'regimes of truth,' and alternative ideas are portrayed as unreasonable, unrealistic, or unworkable. For example, the idea that cuts to welfare expenditure are necessary in terms of economic crises becomes a virtually unquestioned mantra. Other courses of action that might be pursued, such as cuts on military expenditure, are dismissed. While newspaper headlines target welfare fraud, tax avoidance by large multinational companies is considered almost 'natural' and inevitable.

From a Gramscian perspective, subtle ideological control is bolstered through systems of social welfare, provision of unemployment assistance, pensions, and other forms of support. Various spatial policies ostensibly designed to reduce regional inequalities might be seen as attempts to maintain the state's territorial hegemony. Notwithstanding periodic bouts of austerity and spending cuts, the state generally manages to deflect much opposition through the performance of its welfare functions, serving to 'buy off' the otherwise disaffected, and maintaining the political order. Another mechanism is the invoking of threats. Terrorist attacks can be mobilized as a justification for both foreign and domestic policies: of overseas military adventurism on the one hand, and increasing domestic surveillance and curtailment of civil liberties on the other.

Within this discussion, a note of caution is necessary. It is important to bear in mind that states are not unified, self-contained all-powerful units, but rather consist of a series of diverse and fragmented institutions, operating at and across a range of spatial scales, whose impacts on society are hugely variable (Smith, 2009). While the state obviously wields considerable power, spaces continue to be created in which opposition can be voiced. Increasingly diverse interpretations of citizenship point to the plethora of ways in which state hegemony continues to be challenged at local, national, and transnational scales (Yarwood, 2014). The utilization of public spaces for political protest of various guises is amply demonstrated in such diverse places as Tahrir Square in Cairo and Independence Square in Kiev.

Summary

This article has introduced the concept of the state as a political and geographic entity, it has outlined various theories of the origins of states, highlighted its key features, and cast light on the role and functions of the state in capitalist societies. The state is the world's most obvious territorial unit but it is a human creation resulting from the interplay of political processes. There is a need to be wary of not falling into what Agnew (1994) refers to as the 'territorial trap,' whereby states are assumed to be immutable entities. States are historically contingent, they are dynamic, and they reflect processes through which territory (and those living in that territory) is controlled. States do not just happen; they are created and reproduced, and occasionally subject to challenge. Through the appurtenances of the state we come to think in a territorial way. This state-centered approach reifies current political-territorial structures and portrays the state as a permanent and natural feature, they are historically contingent. States are more than mere spatial containers; they are dynamic entities, constantly reconfigured and reterritorialized through the complex interplay of internal and external forces.

See also: Geopolitics; Political Geography; Territoriality: Geographical; Urban Neoliberalism; War and Terrorism, Geography of.

Bibliography

Agnew, J.A., 1994. The territorial trap: the geographical assumptions of international relations theory. Review of International Political Economy 1 (1), 53–80.

Agnew, J.A., 2005. Sovereignty regimes: territoriality and state authority in contemporary world politics. Annals of the Association of American Geographers 95 (2), 437–461.

Bell, J., Staeheli, L., 2001. Discourses of diffusion and democratization. Political Geography 20 (2), 175–195.

Birch, J., 2007. The Lezghi. In: Brower, B.A., Johnston, B.R. (Eds.), Disappearing Peoples? Indigenous Groups and Ethnic Minorities in South and Central Asia. Left Coast Press, Walnut Creek, pp. 207–220.

Brenner, N., Elden, S., 2009. Henri Lefebvre on state, space, territory. International Political Sociology 3 (4), 353–377.

Elden, S., 2007. Governmentality, calculation, territory. Environment and Planning D: Society and Space 25 (3), 562–580.

Elden, S., 2009. Terror and Territory. The Spatial Extent of Sovereignty. University of Minnesota Press, Minneapolis.

Elden, S., 2013. The Birth of Territory. University of Chicago Press, Chicago.

Escolar, M., 2003. Exploration, cartography and the modernization of state power. In: Brenner, N., Jessop, B., Jones, M., MacLeod, G. (Eds.), State/Space. A Reader. Blackwell, Oxford, pp. 29–52.

Fall, J.J., 2010. Artificial states? On the enduring geographical myth of natural borders. Political Geography 29 (3), 140–156.
Flint, C., Taylor, P.J., 2007. Political Geography. World-Economy, Nation-State and Locality, fifth ed. Prentice Hall, Harlow.
Foucault, M., 1980. In: Gordon, C. (Ed.), Power/Knowledge. Selected Interviews and Other Writings. Harvester Press, Brighton.
Gramsci, A., 1971. Selections from the Prison Notebooks. International Publishers, New York.
Gregory, D., 2007. Vanishing points. law, violence and exception in the global war prison. In: Gregory, D., Pred, A. (Eds.), Violent Geographies: Fear, Terror, and Political Violence. Routledge, London, pp. 205–236.
Harris, E., 2009. Nationalism. Theories and Cases. Edinburgh University Press, Edinburgh.
Hartshorne, R., 1950. The functional approach in political geography. Annals of the Association of American Geographers 40 (1), 95–130.
Harvey, D., 2010. The Enigma of Capital and the Crises of Capitalism. Profile Books, London.
Held, D., 1989. Political Theory and the Modern State. Polity Press, Cambridge.
Kearns, G., 2009. Geopolitics and Empire. The Legacy of Halford Mackinder. Oxford University Press, Oxford.
Lefebvre, H., 1991. The Production of Space. Basil Blackwell, Oxford.
Mann, M., 1984. The autonomous power of the state. European Journal of Sociology 25 (2), 185–213.
Megoran, N., 2006. For ethnography in political geography: experiencing and re-imagining Ferghana Valley boundary closures. Political Geography 26 (10), 622–640.
Meinhof, U.H. (Ed.), 2002. Living (with) Borders. Identity Discourses on East–West Borders in Europe. Ashgate, Aldershot.
Miliband, R., 1969. The State in Capitalist Society. Quartet, London.
Muir, R., 1997. Political Geography. A New Introduction. Macmillan, Basingstoke.
Paasi, A., 1999. Boundaries as social practice and discourse: the Finnish-Russian border. Regional Studies 33 (7), 669–680.
Poulantzas, N., 1969. The problem of the capitalist state. New Left Review 58, 119–133.
Reid-Henry, S.M., 2007. Exceptional sovereignty? Guantanamo Bay and the re-colonial present. Antipode 39 (4), 627–648.
Rumley, D., Minghi, J. (Eds.), 1991. The Geography of Border Landscapes. Routledge, London.
Scott, J.C., 1998. Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven.
Sharma, A., Gupta, A., 2006. Introduction: rethinking theories of the state in an age of globalization. In: Sharma, A., Gupta, A. (Eds.), The Anthropology of the State. A Reader. Blackwell, Malden, pp. 1–41.
Smith, M.J., 2009. Power and the State. Palgrave Macmillan, Basingstoke.
Taylor, P.J., 1994. The state as container: territoriality in the modern world-system. Progress in Human Geography 18 (2), 151–162.
Yarwood, R., 2014. Citizenship. Routledge, Abingdon.

State, History of

Wolfgang Reinhard, Professor Emeritus of Modern History, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

First, the fully developed power state of the nineteenth and twentieth centuries is defined against more loosely structured premodern kingdoms and empires. Next, a complex three-layer theory of state building is presented, and the corresponding growth of state power in European monarchies described. After an assessment of the role of political alternatives such as urban republics, the final development of the national, democratic, total welfare state between the French Revolution and the late twentieth century is analyzed. Finally, the multifaceted decline of state power since the 1970s is discussed together with its theoretical and practical consequences.

Introduction: Problems of Definition

At present, there are about 200 communities in the world that are considered sovereign nation-states, according to standards developed during European history. Officially, alternative political models no longer exist. Therefore, it seems reasonable to adopt the narrow definition of modern statehood created by German professors of law (Jellinek, 1959). According to them, a state is constituted by a homogeneous population inhabiting a contiguous territory under one single government that is characterized by complete independence from any outside authority (sovereignty) and holds the monopoly of jurisdiction and the legitimate use of violence internally. History has shown that this kind of state need not be a nation, and even less a constitutional democracy based on human rights. But today, most states claim to be nations under the rule of law and a formal constitution. Nevertheless, these are recent achievements and rather exceptional in world history.

History abounds with different kinds of political organization that may also be called 'states.' This is not quite correct from the historian's point of view, because in most European languages the term 'state' was originally used to designate social status or real estate. But between the seventeenth and nineteenth centuries, it became the name of a particular type of European political organization, achieving perfection during that same period. Before then, prestate communities can be roughly classified into (1) tribes without rulers, (2) chiefdoms, (3) city-states, and (4) empires (van Creveld, 1999). Historically as well as typologically, humankind was first organized in face-to-face communities, either acephalous societies without rulers based upon kin groups, or chiefdoms based upon force, wealth, and consequent social stratification. In addition, many chiefs were distinguished by religious qualities or supported by a privileged group of priests. All this is even truer of empires, which through religious ideology, a standing army, a kind of bureaucracy, and a tax system solved the problem of permanent control of large territories. This was a necessary precondition of future state building, but not the sufficient one, which was to be the achievement of some city-states of antiquity. Whereas everywhere else the distinction between ruler and owner of a polity did not exist or was at least blurred, in Greece and Rome for the first time citizens appointed certain persons to govern on behalf of the community and not for their own purposes, that is, not as rulers but as magistrates. These were the roots of the state as a transpersonal, abstract institution with an abstract legal personality (van Creveld, 1999). Therefore, the history of the state is first of all the history of the transformation of the loosely structured 'chiefdoms' of European aristocrats and 'empires' of European kings into modern states from the high Middle Ages until the nineteenth century.

Theory

European state formation was a multidimensional process, but most theories of state building are still one dimensional. Therefore, a multifactorial three-level theory of state building that combines (1) the microlevel of individuals and groups, (2) the mesolevel of the political system, and (3) the macrolevel of society is a more promising suggestion (Reinhard, 1992).

State building starts on the microlevel with the self-seeking lust for power of individuals, often with the competitive advantage of holding a kingship. Before the existence of the state as an abstract institution, the necessary transpersonal continuity was provided by the dynasty. Dynastic state building consisted in the elimination or at least control of the rival holders of autonomous power dating from the prestate phase of history – the nobility, the church, and urban and rural communities – in order to establish a monopoly of power. To succeed, dynasties needed the assistance of power elites that in their own interest made the growth of state power their own cause. In the long run, lawyers of bourgeois origin proved better qualified for that role than members of the church or the nobility because, in contrast to the latter, lawyers owed all their status and power to the service of monarchs.

Far-reaching change on the mesolevel of the political system resulted from successful exploitation of war, religion, and patriotism for the purpose of expanding dynastic power. The preexisting rivalry of European monarchs inevitably grew with their power, because it became essential to outstrip their neighbors, to grow at their expense, and to protect themselves against the same objectives in turn. Therefore, they needed continuously growing armies and ever-increasing quantities of money to pay them. In its decisive phase of growth, the modern state was a war state, which expanded its taxation, administrative, and coercive apparatus mainly to wage war. This

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62098-6

resulted in a circular process, the coercion–extraction cycle (Finer, 1997), and finally in the internal and external monopoly of violence. Ultimately, only states wage war. Private wars, such as vendettas or feuds, or noble or popular revolts, were no longer legitimate under the powerful war and police state. 'Necessity' in the service of the common good served as the key argument to legitimate this growth of state power. But when the competing 'confessional' churches lost most of their autonomy to the state after the Protestant Reformation – the price to be paid for political protection – religion became an instrument of emotional identification of the subjects with their country. 'Catholic' and 'Bavarian,' 'Polish,' or 'Spanish' became almost synonymous, as did 'Protestant' and 'English,' 'Prussian,' or 'Swedish' on the other hand.

Essential contributions originated from the societal and cultural environment on the macrolevel. First of all, the geo-historical plurality of Europe was the incentive for the growth of state power through the coercion–extraction cycle. The result was a stable pluralism of internally strictly unitary states, an exceptional case worldwide. Universal empires never had a chance in Europe; the Holy Roman Empire of the Germans was at most a first among equals. But internal unity was not realized before the late eighteenth, the nineteenth, or in some cases even the twentieth century. For a long time, most monarchies were composed of several parts of unequal status such as Castile and Aragon or Polonia and Lithuania. And everywhere monarchs had to cope with a powerful system of autonomous local noble rule, on the one hand, with a countrywide network of partly autonomous urban and rural communities, on the other, once again European particularities. In addition, before the Reformation the church considered itself an independent community, in some respect even a state before the state. This singular European dualism of spiritual and temporal, combined with the equally singular political pluralism, turned out to be a precondition of political freedom, although neither church nor state, and neither noble landlords nor urban oligarchies, were in favor of any liberty except their own. Finally, the strong position of the church derived from its role as keeper of Latin culture. Roman law, to some extent transformed into the canon law of the church, directly and indirectly proved fundamental not only to monarchic state building but also to individual liberty and property.

The Age of Monarchy (Middle Ages to the Eighteenth Century)

European kings of the Middle Ages and the early modern period claimed to be emperors of their kingdoms, which is to say that, like the pope in the church, they held unrestricted power (*plenitudo potestatis*). For the first time in history, the competence of a king was defined by the French constitution of 1791. Before that date, royal power was limited only in the sense that there were some things a king could not do. The English king could not create statutory law or raise taxes without consent of the realm; the French king could not change the succession to the throne or give away parts of the crown domain. For the rest, every monarchy was 'absolute,' that is to say, the king's prerogative included everything necessary to fulfill his fundamental duties as fountain of justice and keeper of peace. In that sense, even Poland's 'republic of nobles' and England's parliamentary system were called 'absolute' by contemporaries. In Britain, where no written constitution defines powers, in theory the monarch today is still entitled to declare war and to conclude peace without consulting Parliament. In reality, however, since the later Middle Ages this was possible only if the necessary funds were raised with the help of representatives of the subjects. The monarchies of the European Ancien Régime were limited not so much by legal restraint as by small resources, traditional rules, and social conventions. Under these circumstances, it was decisive for the process of dynastic state formation that the crowns of Castile, England, and France became hereditary, whereas the Empire and to some extent Denmark, Sweden, Poland, and Hungary remained elective monarchies. Compared with the universal competence of the modern state, including the right to extend this competence at discretion, the prerogatives of early modern monarchs were extremely modest. State activity expanded when, during the early modern period, the so-called police ordinances (*Policeyordnungen*) started to cover new fields such as health and public welfare, moral order and communication, economic behavior, and labor relations. This was no longer mere jurisdiction, reactive government by rescript as in the Middle Ages, but an expanding administration actively interfering with larger and larger parts of its subjects' lives. But it was still not 'absolutism,' an anachronistic term of the nineteenth century that should be dropped (Henshall, 1992), because the kind of monarchy without any legal restriction that this term designates never existed except in Denmark from 1665 to 1834 and in Sweden for a few decades after 1693.

Similar to most monarchies of the world, European ones had a sacred character, practiced elaborate symbolic rituals, and were legitimized 'by the grace of God.' The coronation ceremony together with the coronation oath of the king and the corresponding oath of allegiance of the people were central in this respect. But there were varieties. In France, political rituals and ideology constituted a true 'royal religion' (*religion royale*), and court life with its palaces, ceremonies, and festivities became an all-embracing total work of art. Some of its elements recur in England and elsewhere, especially the ritual of the royal touch, when French and English monarchs touched hundreds of people suffering from scrofula, the tubercular 'royal disease' (*morbus regius*), to cure them. It is worth noting that kings practiced this ritual with particular intensity when their monarchy needed additional emotional support: Henry IV and Louis XIII of France after the religious civil wars, Elizabeth I of England because of her succession problems, and Charles II after the restoration. Other monarchies were more prosaic, including the Holy Roman Empire. And in spite of his politico-religious sanctuary, the Escorial, Philip II of Spain never wore a crown; there was not even a coronation ceremony in Castile. This was a way of demonstrating independence from the church, but even in France political sacrality implied royal control of the church and not the opposite.

Because of the church's monopoly of culture, Europe's early political discourse was theological. The state was considered a consequence of the fall of man. With the reception of Aristotle, natural man became a political being, a concept that was used to create an autonomous discourse of legitimation for the monarchy, erroneously, because 'political being' referred to the

face-to-face community of the city and not to the state. But from Dante Alighieri down to John Locke, political theory remained essentially a theory of monarchical state building, at most harmonizing limited monarchy with aristocracy and democracy into a mixed constitution, supposedly like the ancient Roman republic. Thomas Hobbes was the first to speak of the state instead of the monarchy, but this was not common practice before the eighteenth century, with Montesquieu and Jean-Jacques Rousseau.

This change coincided with the development of central institutions that under the impact of enlightened rationality became more effective. Administrators were no longer the king's personal servants but were transformed into professional officials of an abstract state, into modern civil servants (*Beamte*) as described by Max Weber. The early medieval prince was assisted in his governmental functions by members of his court and council (*curia regis*), and in his written legal work by a chancellor, often an ecclesiastic. Most central institutions of European monarchies were created by segmentation of that curia. The court was reduced to a mere household, and the chancery became a separate body. Professional supreme courts and treasuries were established. To produce political consensus and financial aid, an enlarged council of important subjects was created to become an assembly of estates or a parliament. What was left over of the ancient council was transformed from a multifunctional decision-making instrument such as the German aulic council (the Hofrat) into a differentiated system of governing bodies with a small group of confidants of the ruler at the core (the Privy Council, Geheimer Rat, Staatsrat, etc.). From the sixteenth to the eighteenth centuries, the conciliar principle of collegiality was common practice, to accumulate competence, to balance interests, to prevent disloyalty by mutual control, and to provide positions for as many members of the power elite as possible. Only a few very dedicated rulers such as Philip II of Spain, Louis XIV of France, and Frederic II of Prussia were still able to handle this complicated system of central government in person. Instead, many princes relied on a personal confidant as prime minister, such as Wolsey, Richelieu, or Olivares, or used the new office of a secretary of state for coordination. In England, parliamentary management was a central task of royal confidants. The development of modern cabinet government was the consequence.

In France, Castile, Rome, and elsewhere, many public offices were officially for sale, just another way to tap private accumulation in favor of public finance. The growth of state power was first of all a credit problem, because tax income had to be anticipated by credit to make the large sums of money that were necessary to wage war available on short term. Tax farming was another widespread technique used for that purpose. Direct taxes were based on the land, because movables and income were still not under administrative control. As usual, indirect taxes on mass consumption proved more comfortable, whereas the rich were involved as creditors. But credit was still the personal credit of a prince, not of the state as an institution. Therefore, repayment of loans was expected and interest rates were as high as risks. Under these circumstances, bankruptcy was the regular consequence of power policy until the Dutch and English invention of low-interest, long-term bonds with parliamentary guarantee, a decisive contribution to the further growth of the modern state.

Premodern Alternatives and Their Political Heritage

Traditionally, Europe respected man (and to some extent woman) as proprietor of his (her) person and goods. Therefore, in principle, taxation was possible only with the consent of the taxed or at least their lords and guardians. For this purpose, most princes created structurally similar assemblies of important subjects or representatives of corporations, such as cities, called assemblies of estates, beginning 1188 in León. By the seventeenth century, however, most princes were strong enough to raise taxes without consent and to ignore their estates. This decline of estates has been considered the very essence of 'absolutism' by former research. In contrast, we know today that to some extent estates remained alive in most countries (with the exception of Denmark) and princes still had to bargain with them. But only in England in the seventeenth century did conflicts with the Crown ended with victories of Parliament and the definite establishment of financial control by that body. This system, imitated in most European countries after the French Revolution, finally led to parliamentary government and democracy.

In contrast to other parts of the world, in European political culture the relationship between ruler and subjects was considered a contractual one, based on legal mutuality. In case of a supposed breach of contract by the ruler, subjects felt entitled to armed resistance, because their legitimate lord had degenerated into an illegitimate tyrant. Therefore, early modern discourse and practice of the right of resistance through a series of European revolutions laid the foundations of modern liberty and democracy.

In most countries, assemblies of estates were dominated by the nobility; as holders of local power, they were considered the true lords of the land. Not before the eighteenth century did the administrative institutions of the state start to penetrate this autonomous local power structure, and not before the nineteenth century was noble rule to a large extent replaced by public administration. But beside or below noble rule existed innumerable, more or less self-governing urban and rural communities, from independent republics like Venice, quasi-republics like German imperial towns, privileged English boroughs, down to villages of Polish serfs who were still entitled to regulate certain common affairs on their own. This is the 'Europe of a thousand republics,' the Europe of 'communalism' (Blickle, 2000), a world where members of face-to-face communities decided on their immediate concerns. These are the roots of European republicanism and of the tradition of European self-administration.

When assemblies of estates flourished, ecclesiastical power competing with princes was already declining. But the Roman church had been Europe's first state, with the pope as the first absolute monarch and the priest as the first nonhereditary professional public servant. The church practiced territoriality (*Flächenstaat*) at a time when empires were still mainly based upon personal dependency (*Personenverbandsstaat*). And the church established the rules of political religion, with the consequence that to the present day states remain transcendent entities in need of believers. In this sense, the state is the last church because European political culture is a culture of believers.

In contrast, professional law courts such as the French 'parlements' or the English 'common law and equity courts'

were not originally autonomous alternatives to state-building monarchies like aristocratic rule, communal self-government, or the church but creations of the expanding state power. Nevertheless, because of the loose structure of early modern polities giving room for the development of specific group identity and interests, these judges under certain conditions could transform themselves into opponents of their king's policy and initiate resistance or revolution, as happened in England and France in the seventeenth and eighteenth centuries. But in this way the principles of rule of law, of an independent judiciary, and even of judicial review of legislation were introduced into political culture.

In the end, however, the fully developed modern state replaced noble rule by a king's officials, reduced local self-administration to a commission of the central government, and established complete control of the churches and the legal apparatus. The principle of justice existing beyond, above, and before the state was largely replaced by the arbitrary use of the state's monopoly of legislation. But nevertheless the heritage of political alternatives remained alive: liberty and democracy, rule of law and independent judges, and the claim to self-government and to have a state one could believe in.

The Age of Equality (Nineteenth and Twentieth Centuries)

The modern European state, so far developed by state-building monarchies, reached maturity with the French Revolution and the political modernization of Europe that followed. Modernity means unity: unity of territory, people, government, sovereignty, and violence. The revolutionary process took the last step to modernity and replaced the hierarchy of orders and the plurality of privilege that had characterized the political society of the Ancien Régime with a society where stratification became a private distinction because politically all subjects were now equal before the state and its law. From now on, not families or corporations were the basic components of the state, but individuals. But mobilizing individuals emotionally in favor of the state into some kind of participation – not necessarily a democratic one – resulted in an ultimate increase of state power with incredible dimensions: the total state.

With the quasi-natural order of estates and corporations gone, the modern state of equal individuals needed a new organization. With the exception of Britain and Sweden (until 1975), this was provided from now on by a constitution. But in spite of the American and French models of 1787 and 1791, nineteenth-century European constitutions established rule of law but neither democracy nor fundamental rights. Their first purpose was equal integration, not equal participation. Most countries remained monarchies, but because monarchy was deprived of sacrality since the Enlightenment period, it lost unquestionable legitimacy. Step by step, monarchies were transformed into democratic republics or quasi-republics. But in most countries during the nineteenth century, democracy was still far away. This was not only informally as usual, but even legally: in their right to vote, all subjects were equal, but some were more equal than others (Orwell, 1945), mostly because of their wealth. In 1830, the following percentage of the total population was entitled to vote: in Austria, Denmark, Prussia, and Spain, 0%; France, 0.5%; Belgium, 1.1%; and Britain, 2.1%. In 1910, the numbers were as follows: in Denmark, 17%; Britain, 18%; Austria, Belgium, Germany, and Spain, 21–23%; and France, 29% (Reinhard, 1999).

Percentages in the 20s indicate universal suffrage of adult males. Down to the twentieth century, state building and politics remained almost exclusively a business for men. There were some outstanding queens and empresses in European history, such as Isabella of Castile, Elizabeth I of England, Christina of Sweden, Maria Theresa of Austria, Catherine II of Russia, and others. But some of these great women had their difficulties with male political culture. And there were no female ministers or officials. Even the enlightened philosopher Immanuel Kant considered women political nonentities because they lacked economic independence. After Wyoming (1869) and New Zealand (1893), female suffrage was introduced in Finland (1906); Austria, Germany, and Poland (1918); the United States (1920); Britain (1928); Spain (1931); France (1944); and Switzerland (1971). But in the late twentieth century, female heads of state or government, ministers, parliamentarians, and even female civil servants were still a minority.

Originally, human beings identify themselves mainly with their face-to-face community. 'Patria,' the land of the fathers, and 'Natio,' the community of birth, even during the early modern period quite often designated rather small territorial entities. However, through state building, people learned to apply them to their entire country. Finally, after the French Revolution, self-defined linguistic or cultural communities, peoples in search of a state, considered themselves nationalities, that is, nations to be. In contrast to ideology, nations are not natural but are invented entities. This does not, however, imply complete artificiality and arbitrariness. There is always some kind of basis in reality, such as common descent, language, religion, or territory. Most important is a common history and the consequent option for its continuity. Nationalism as a system of symbols constitutes a community as a nation, integrates its members, and defines the difference from other communities. For radical nationalists, the nation is the highest value and the ultimate source of meaning; nationalism becomes a religion and as such a way to tap the material and emotional resources of subjects, especially in war. Therefore, nation building became the final phase of European state building, whereas new states in Europe and worldwide tried to constitute themselves as nation-states right from their foundation.

Expanding political participation and emotional mobilization of subjects correspond to a further extension of government and of state power. A common denominator is a high degree of social discipline in the subject. Church and State had been working on this since the late Middle Age, but achieved this goal only in the nineteenth and twentieth centuries with the assistance of armies, schools, and factories (Weber, 1964). The citizens of a fully developed modern state enjoyed the legal protection of their person and property and some political participation. On the other hand, they were subject to three fundamental duties: compulsory school attendance, compulsory military service, and general liability to taxation. Based upon the fiction of the sovereignty of the people, the state apparatus was able to extend its competence at its own

discretion and to transform itself into a totalitarian system, as happened under the impact of nationalist, racist, and communist ideologies in the twentieth century. Some of them proclaimed the end of the state, but that proved impossible. In contrast, the totalitarian state turned out to be the final stage of development of the modern state. And, at a closer look, even the modern welfare state proves to be a soft variety of totalitarianism. Beginning with limited social security programs before World War I (1883–1889, the creation of an insurance system for workers in Germany), the welfare state exploded after World War II. Between 1950 and 1983, the Western European average of social security expenses rose from 9% of the Gross National Product to 25%. But total care for citizens means also total regulation of their lives. Even the neoliberal reaction is based upon large-scale government action in the service of a general economic and social policy; the goals, not methods, differ from those of former socialist welfare policy.

Six hundred years of European expansion made the modern European state the world's standard model of political organization. The colonial regimes of Spanish and British America established in the sixteenth and seventeenth centuries adopted premodern forms of European statehood and political culture. Therefore, most Latin American states remain authoritarian but weak states even today, whereas traditional self-government in a comparatively egalitarian society made the United States a successful nation, although it lacked certain traits of the modern state such as belief in the state as a value in itself, in bureaucratic centralism, in a strong professional army, and in an established church. During 1839–1935 Britain's white settler colonies became sovereign states, mostly combining the British and the US models, whereas in the nineteenth and early twentieth centuries East Asian and Muslim empires tried to transform themselves into modern states to cope with the weakness of their traditional regimes when confronted with the West, in the case of Japan with tremendous success. Finally, during several waves of decolonization after World War II, the lot of quite artificially created colonies became equally artificial nation-states. For the first time, a global community of international law and moderately effective international organizations such as the UN became reality. Quantitatively and qualitatively, the modern state reached its culminating point during the twentieth century.

The Disintegration of States and Possible New Forms of Political Life

In the last third of the twentieth century, the culmination of the modern state turned into decline. To begin with, most 'new' states worldwide suffer from serious shortcomings, such as lack of democracy and efficiency, general corruption, and political disintegration. The export of European state models was a success in only three specific constellations: (1) where it was accompanied with the export of European civil society and political culture, as in the cases of the US and the British dominions; (2) where it met a polity and society compatible with its own structure, as in the case of Japan; and (3) where colonial rule lasted long enough to create a Western-style civil society, as in the case of India. For the rest, the new states of the former colonial and the former communist worlds tend to lose their sovereignty internally to the para-statehood of old or new chiefdoms or ethnic communities, and externally to international organizations like the UN or the IMF. They become mere 'quasi-states' (Jackson, 1990). In addition, certain Muslim countries revise their option for secularity, which is an essential trait of the modern European state.

At the same time, European states were also, to some extent, disintegrating into 'tribes,' were losing sovereignty to 'chiefs,' pressure groups, and international organizations; and were no longer able to provide complete social security. As a consequence of totalitarianism and mass participation, the modern state turned self-destructive. The excesses of nationalism, racism, and communism destroyed the necessary faith in the state. Social discipline was questioned by the common desire of self-realization. New ethnic, regional, and social movements started to compete with the state for the loyalty of its subjects, some on the basis of democratic self-determination, and some with the use of violence. Withdrawal of loyalty favored an underground economy expanding at the expense of the tax and welfare systems. The necessity to win elections by bargaining with interest groups and by providing additional welfare benefits for the electorate ended in a debt trap with almost the whole budget allocated in advance. In some cases, an expanding public debt, once the basis of the growth of state power, led to a debt load exceeding national income. State bankruptcy could be avoided only by international intervention. In addition, economic globalization allows important taxpayers to move to less restrictive countries with lower wages and taxes. On the other hand, for economic and security reasons most states lose more and more of their sovereignty to international organization. Despite a common currency (the euro) established by 17 of its members between 1999 and 2002, the European Union has not yet become a federal state and suffers from deficient democracy and social policy. Nevertheless, common economic interest has released remarkable binding forces and still makes the union attractive for new members.

On the other hand, the right to declare and to wage war, once the very essence of sovereignty, has become obsolete. Only the United States is able to start wars at its convenience, and it does it to police certain parts of the globe in their present role of world hegemon. Instead of interstate wars and peace treaties, we have hundreds of 'low intensity conflicts' (van Creveld, 1991) that tend to respect neither noncombatant status nor national borders. International Muslim terrorism may in reality still be marginal, but after the attack on the World Trade Center on September 11, 2001 it is considered as a menace to state power.

Nevertheless, the national state is still there. But it has definitely lost the monopoly of power and policy. In power processes, it tends to become one competitor among others. Unity, the constituent principle of modernity and the monocratic polity, is being replaced by a postmodern plurality of multiple political units coordinated by the necessity to cooperate and not by state power. In this situation, it may no longer be adequate to speak of 'the state' in the essentialist sense of a stable unit but rather to liquefy terminology by employing instead the dynamic concepts of 'stateliness' (*Staatlichkeit*), which may differ in degree, and of 'the state as process' (Schuppert, 2010). Certainly, the new pluralism of political

constructs might end in chaos. But it might also prove to be a new way for humankind to seek the political pursuit of happiness.

See also: Absolutism, History of; Aristocracy, Nobility, and Gentry, European History of; Balance of Power, History of; Bureaucratization and Bureaucracy, History of; Citizenship, Historical Development of; Civil Society, Concept and History of; Communism, History of; Conservatism: Historical Aspects; Democracy, History of; Enlightenment; Ethnic Cleansing, History of; Ethnicity and Ethnic Groups: Historical Aspects; Feudalism; French Revolution, The; Frontiers in History; Global History: Universal and World; Human Rights, History of; Interest Groups, History of; Laicization, History of; Liberalism, Anthropology of; Military History; Modernization and Modernity in History; National Liberation Movements; National Socialism and Fascism; Parliaments, History of; Peace-Making in History; Political History: History of Politics; Political Parties, History of; Racism, History of; Reformation and Confessionalization; Revolution; Social Movements: A Social Psychological Perspective; Socialism: Historical Aspects; Violence, History of; Warfare in History; Welfare State, History of.

Bibliography

Anderson, M.S., 1993. The Rise of Modern Diplomacy, 1450–1919. Longman, London.

Bayart, J.-F., 1989. L'état en Afrique. La politique du ventre. Fayard, Paris (1996 The State in Africa. The politics of the belly. Longman, London).

Berman, H.J., 1983. Law and Revolution. The Formation of the Western Legal Tradition. Harvard University Press, Cambridge, MA (1991 Recht und Revolution. Die Bildung der westlichen Rechtstradition. Suhrkamp, Frankfurt).

Blickle, P. (Ed.), 1997. Resistance, Representation, and Community. Clarendon Press, Oxford (1998 Résistance, représentation et communauté. Presses universitaires de France, Paris).

Blickle, P. (Ed.), 2000, Kommunalismus. Geschichte einer gesellschaftlichen Organisationsform, 2 vols. Oldenbourg, München.

Bonney, R. (Ed.), 1995. Economic Systems and State Finance. Clarendon Press, Oxford (1996 Systèmes économiques et finances publiques. Presses universitaires de France, Paris).

Coleman, J. (Ed.), 1996. The Individual in Political Theory and Practice. Clarendon Press, Oxford (1996 L'individu dans la théorie politique et dans la pratique. Presses universitaires de France, Paris).

Contamine, P. (Ed.), 1998. Guerre et concurrence entre les états européens du XIVe au XVIIIe siècle. Presses universitaires de France, Paris (2000 War and competition between states. Clarendon Press, Oxford).

Creveld, M v., 1991. The Transformation of War. Free Press, New York (1998 Die Zukunft des Krieges. Gerling, München).

Creveld, M v., 1999. The Rise and Decline of the State. University Press, Cambridge (1999 Aufstieg und Untergang des Staates. Gerling, München).

Edelman, M., 1972. Politics as Symbolic Action. University of Illinois Press, Urbana (1976 Politik als Ritual. Die symbolische Funktion staatlicher Institutionen und politischen Handelns. Campus, Frankfurt).

Edelman, M., 1974. The Symbolic Uses of Politics. Markham, Chicago (1976 Politik als Ritual. Die symbolische Funktion staatlicher Institutionen und politischen Handelns. Campus, Frankfurt).

Ellenius, A. (Ed.), 1998. Iconography, Propaganda, and Legitimation. Clarendon Press, Oxford.

Finer, S., 1997. History of government from the earliest times, 3 vols. University Press, Oxford.

Hamilton, K., Langhorne, R., 1995. The Practice of Diplomacy: Its Evolution, Theory and Administration. Routledge, London.

Henshall, N., 1992. The Myth of Absolutism: Change and Continuity in Early Modern European Monarchy. Longman, London.

Hobsbawm, E.J., 1990. Nations and Nationalism since 1780: Programme, Myth, Reality. University Press, Cambridge (1991 Nationen und Nationalismus. Mythos und Realität seit 1780. Campus, Frankfurt; 1997 Nations et nationalisme depuis 1780: programme, mythe, réalité. Gallimard, Paris).

Hoffman, J., 1995. Beyond the State. Polity Press, Oxford.

Jackson, R.H., 1990. Quasi-States. Sovereignty, International Relations, and the Third World. University Press, Cambridge.

Jellinek, G., 1959. Allgemeine Staatslehre. 6th impr, third ed. Wissenschaftliche Buchgesellschaft, Darmstadt (1st edn. 1900).

Jouvenel, B d., 1945. Du pouvoir. Histoire naturelle de sa croissance. Bourquin, Genève (1972 Über die Staatsgewalt. Naturgeschichte ihres Wachstums. Rombach, Freiburg; 1993 On power. The natural history of its growth. Liberty Fund, Turpin Distribution, Letchworth).

Koenigsberger, H.G. (Ed.), 1988. Republiken und Republikanismus im Europa der Frühen Neuzeit. Oldenbourg, München.

Mann, M., 1986–1993. The Sources of Social Power, 2 vols. University Press, Cambridge.

Orwell, G., 1945. Animal Farm. Secker & Warburg, London.

Padoa-Schioppa, A., 1997. Legislation and Justice. Clarendon Press, Oxford.

Reinhard, W., 1992. Das Wachstum der Staatsgewalt. Historische Reflexionen. Der Staat 31, 59–75.

Reinhard, W., 1996. Power Elites and State Building. Clarendon Press, Oxford (1996 Les élites du pouvoir et la construction de l'État en Europe. Presses universitaires de France, Paris; 1997 Las élites del poder y la construcción del Estado. Fondo de Cultura económica, Madrid).

Reinhard, W., 1999. Geschichte der Staatsgewalt. Eine vergleichende Verfassungsgeschichte Europas von den Anfängen bis zur Gegenwart. C. H. Beck, München.

Reinhard, W., 2007. Geschichte des modernen Staates. Von den Anfängen bis zur Gegenwart. C. H. Beck, München (2010 Storia dello stato modern. Il Mulino, Bologna).

Ritter, G.A., 1991. Der Sozialstaat. Entstehung und Entwicklung im internationalen Vergleich, second ed. Oldenbourg, München.

Schuppert, G.F., 2010. Staat als Prozess. Eine staatstheoretische Skizze in sieben Aufzügen. Campus, Frankfurt.

The Origins of the Modern State in Europe, 13th–18th Centuries, 7 vols, 1995–1999. Clarendon Press, Oxford (1996 sq. Les origines de l'état moderne en Europe, XIIIe-XVIIIe siècle. 7 vols., Presses universitaires de France, Paris; 1998 sq. Los Orígenes del Estado moderno en Europa, Siglos XIII a XVIII. 7 vols., Fondo de Cultura Económica, Madrid); see also: Blickle 1997, Bonney, Coleman, Contamine, Ellenius, Padoa-Schioppa, Reinhard 1996.

Tsatsos, D.T. (Ed.), 2010. Die Unionsgrundordnung. Handbuch zur europäischen Verfassung. Berliner Wissenschaftsverlag, Berlin.

Weber, M., 1964. Wirtschaft und Gesellschaft. Studienausgabe. Kiepenheuer und Witsch, Köln, Berlin (1st edn. 1921, 1978 Economy and Society. University of California Press, Berkeley).

State, Sociology of the

Hilton L Root, George Mason University, Arlington, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The formation of states changes the rules governing the behavior of the groups it comprises in ways not easily foreseen. The *emergence* of unanticipated behaviors is a characteristic that states share with other complex systems. These are systems in which there are levels of self-organization, and where agents at one level are the building blocks of the next. Such environments are increasingly referred to as *complex adaptive systems*. A change at one level alters the options for change at another, making it difficult to predict the behavior of individuals in constantly shifting environments and organizational formations. At each level, the agents adapt, they acquire new identities, and adjust their behaviors to new sets of rules and to the expected reactions of others. At each level, the constituents are likely to exhibit new behaviors, and how they adapt and act together to create the behavior of the whole is not easily reducible to the sum of the parts.

States as Complex Systems

States consist of government institutions, and rules, but if devising those institutions, rules, were the whole story, we would not be asking why long and costly efforts at building state institutions – those that enable citizens to perform complex collective tasks – often fail in developing regions. Traditional approaches to state formation emphasize the parts, institution, groups, or culture, but it is the relationship between the parts that gives rise to the collective behavior of the whole.

There are several key reasons why building state capacity involves more than drawing up a plan and then implementing it. A state is capable of performing complex tasks because of how its members interact with each other, and not because of how each member behaves separately. The greater part of the story concerns the transitions of behavior, expectations, and culture that occur when new institutions are introduced. They can form and be effective without being planned. State building refers to process through which a social system acquires new structures and behaviors that its individual components did not possess.

A state's behavior is not reducible even to the most exhaustive catalog of institutions or the rules of the game which includes according to Douglass North "formal rules, written laws, formal social conventions, informal rules of behavior, and shared beliefs about the world, laws a well as the means of enforcement" (North, 1990, pp. 3–4). The rules of interaction matter as much as the specifications of the component parts, and successful adaptations are contingent on other factors in the same environment. The way the component social networks and institutions are interconnected determines their behaviors over time. The behavior of one or more components is modified by the behavior of yet another. New behaviors and characteristics can emerge unintentionally, and even in direct contrast to the goals of those who contribute the resources and labor. Moreover, identical institutions set in a different social network will produce different outcomes; what succeeds in one environment may fail in another.

Even when very specific tasks are involved, such as organizing the central bank or managing the monetary system, no one-way or best practice to accomplish this task is applicable to all states. Managing health services or delivering education is even more dependent on local circumstances. Variations in social network structure will produce qualitatively different outcomes despite the existence of institutions that share a common design.

The formation of states changes the rules governing the behavior of the groups it comprises in ways not easily foreseen. *Emergence* and *interdependence* are two concepts that are critical to understanding the relationships between the parts and the environment that produce the collective behavior of the state. Emergence and interdependence are characteristics that states share with other complex systems. These are systems in which there are levels of self-organization, and where populations or institutions at one level are the building blocks of the next. Such environments are increasingly referred to as *complex adaptive systems*. A change at one level alters the options for change at another, making it difficult to predict the behavior of individuals in constantly shifting environments and organizational formations. At each level the agents adapt, they acquire new identities, and adjust their behaviors to new sets of rules and to the expected reactions of others. At each level, the constituents are likely to exhibit new behaviors, and how they adapt and act together to create the behavior of the whole is not easily reducible to the sum of the parts.

Emergence refers to any behavior that occurs between agents that contributes to a more complex behavior of the whole. A state is a nested hierarchy of interactive components. Lower or inner levels of organization combine into higher levels, i.e., groups of individuals form families, groups of families form neighborhoods, groups of neighborhoods form cities, and groups of cities form metropolitan regions, which connect to provinces. The social networks or institutions at one level are the building blocks of the next, but at each level of aggregation, the behavior of the whole shifts. None of these levels exhibit the identical behaviors characteristic of their individual components. Emergence is a collective property that arises from the relationship between the parts.

States do not exist in isolation from the larger systems in which they form a part, nor do they result solely from the accumulation of events in a local population. They are parts of the international system and acquire structure by being embedded into a larger systemic whole, and they interact with that broader system, even as their constituent parts interact with each other.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32149-3

Interdependence describes a state's relationship with its environment, both its natural and its human environment. Because of interdependence, cause and effect are not obviously related. For example, democratization, along with the concepts typically associated with modernization – individualism, tolerance, freedom of association, and free media – and modernity – urbanization, industrialization, and mass education – all occur in the context of a state (Gellner, 1983). However, process and structure are interdependent and are not readily separated. States are catalysts at the same time that they are outcomes of the processes they affect. Socioeconomic megaforces produce states at least as much as states have an influence on those forces. Direct institutional interventions usually have paradoxical outcomes because of socioeconomic interdependencies.

One consequence of the combined effects of emergence and interdependence is that relations among states and economic/political developmental processes within states show sensitivity to initial conditions and are subject to informational cascades. Another consequence is that some prominent metaphors for large-scale historical change are implausible and misleading such as modernization theory, uniqueness of the west, and the end of history. In terms of policy operations, both emergence and interdependence limit conventional approaches to centralizing authority, building institutional standards, or expanding state capacity.

The Great Discontinuity

What differentiates a state from a nonstate social collectives kinship-based community is the role of kin selection and reciprocal altruism. In nonstate communities, trust and cooperation occurs primarily among members of a society who carry the same genes or who interact repeatedly in the same environment. Kinship-based communities grant "selective advantage … to the altruistic interaction of individuals sharing part of the same genotype." (The definition of kin selection is from Mayr (2001).) The transfer of altruistic norms from families and kin groups to nonkin members is the essence of state making (Fukuyama, 2011).

The social capital that forms in kin-based or homogeneous local societies can thwart the formation of civic culture that state building requires. In its more general use, the term refers to political authority modeled on that of a father within a family. Patrimonialism refers to political recruitment on the basis of kin selection and reciprocal altruism among individuals sharing a common bloodline (Fukuyama, 2011). In this sense, patrimonialism can be considered the opposite of bureaucratic rule. Max Weber refers to patrimonial rule as it is projected beyond the family to an entire society or even nation. Both ancient and contemporary histories offer many examples of bureaucratic offices that are filled by a ruler's kin. A contemporary example of the power that patrimonial identification can have over an entire culture is the domination of Indian politics since independence in 1947 by the Gandhi family and that political incumbency at all levels of political representation depends on family connections.

Patrimonial rule is frequently associated with an additional obstacle to the effectiveness of the state: patronage networks, sometimes referred to as patron–client networks. They are a form of reciprocity based on dependence and inequality in which power brokers create a following by exploiting control over scarce resources, not only material wealth but also any resources that confer status or prestige. The patron gains power by delivering benefits selectively, according to the value of the loyalty that can be extracted from the recipient. These networks are most durable when the debts of the weaker partner can never be paid off or when the clients have limited channels of communication with the outside and depend on the patron for contact with other power brokers in the wider society. The extent of patronage networks was traditionally limited by the need to maintain face-to-face relationships, but modern political parties allow patronage webs to be woven across entire nations. The patron can be the traditional elite who are able to revive their moral authority by representing or protecting their dependent local clientele in national or subnational arenas.

Patronage networks incur significant transaction costs that divert resources from alternative uses. It limits networks of exchange to personal sources of information. Lacking a formal institutional structure relationship-based reciprocity results in fewer transactions and less stable outcomes than state-based norms. Modernization theory tended to assume that patron–clientelism was a residual, or vestigial, characteristic of the old order, but instead of being phased out by economic development, 'neo-patrimonial' regimes have arisen in which formal institutional structures must compete with informal networks. (Authority is personal and familial, modeled upon the mechanics of a household, which contrasts with the concept of a rational bureaucracy governed by formal rules where authority is delegated according to competence. Administrative hierarchies ensure that the private and official roles are clearly separated.) Many political parties in recently formed states function as patronage systems. Contemporary India offers numerous examples, beginning with the Indian Congress Party whose model of creating vote banks via patronage distribution has been copied. Many local political parties later emerged that cater to distinct clienteles by offering public services selectively. In this regard, the history of state building in India parallels that of Brazil, Mexico, and Turkey, as well as many other newly emerging democracies.

No society completely eliminates patrimonial tendencies; the psychological motivations of patrimonial reciprocity are very strong. Nevertheless, state order built upon patronage networks can easily regress once the leader is ousted. In Southeast Asia, once patrimonial leaders Ferdinand Marcos and Suharto were rumored to be ill, the asset values held by their patrons collapsed. Institutional regression is another more long-lasting consequence of patrimonialism to state formation. Once Marcos and Suharto were toppled or after Arab dictators Hosni Mubarak, Muammar Gadhafi, or Ali Abdullah Salah were ousted during the Arab spring, deinstitutionalization occurred. (Patrimonial democracies are less likely to regress when power changes, making them more stable than patrimonial autocracies.) The forfeiture of institutional assets each time power changes hands causes political routines and organization to regress as well. As a result, the gap between the quality of governance in state-based and nonstate societies grows larger.

Time Consistency and State Formation

When patrimonial systems of social authority circumvent loyalty to formal institutions, the result is likely to be 'time-inconsistent' preference functions for the entire society. (Time-consistent choices are those that are consistent with longer time horizons.) In such environments, people make poor choices wisely. They sacrifice the long term for immediate payoffs because the value of assets depends on the value of impermanent interpersonal relationships. Europe and China transitioned from patrimonial domination through different institutional channels. Imperial China had much higher levels of public provision than its neighbors because a rigorous exam system weakened kinship as a basis of administrative authority. The states of Western Europe that sharply separated lawmaking from the exercise of sovereignty afforded a far wider range of public goods than did the societies of Eastern Europe that remained under patrimonial domination.

All societies face a time-inconsistency dilemma when the leadership cannot precommit to rules over discretion. Two ways of making time-consistent behavior a social value are bureaucratic rationalism, or perhaps more broadly to establish public order on a rule-of-law foundation. Both of these strategies entail a compromise of discretion for the dominant coalition; both will eventually circumvent the patrimonial system and clash with hierarchies of birth and lineage.

Leaders who commit to the creation of a meritocratic bureaucracy or rule-bound system with an independent judiciary do so at the cost of sacrificing personal power. Yet without such an infrastructure in place, a state's behavior will not be time consistent. The players will know this. If resource holders have no expectation that a government will carry out its plans for the future, they will make only short-term plans. Such societies will have to forgo long-term payoffs in favor of actions that have only immediate payoffs. A reform program that can be undone by decree is only as good as the paper on which it is written. The time inconsistency of government actions does not disappear when a rule-based framework is established. But the risks to property of discretionary behavior are made more tractable, and this shifts the population's investment horizons toward the longer term.

Credible Commitments and Democratic Transition

The leaders of young democracies face a time-inconsistency dilemma if their promises and commitments are not institutionally credible (Kydland and Prescott, 1977). When the new leadership lacks the mechanisms of an effective bureaucracy or system of rules to make its promise credible, they must resort to patronage if only to stabilize their authority. If the leader calculates that the short-term gain of political survival is more important than the long-term gain of regime survival, then instead of providing the public goods needed to secure democracy in the long term, such as policy-driven political parties or state bureaucracies, a leader may allocate surplus revenues to prevent defection of those groups whose loyalty and resources are needed to prevent a crisis in the present. Voters and politicians in young democracies lack informational mechanisms to ensure that promises to policy performance are met. Voters may not even be able to observe if a policy promise is directed at them, and hence they cannot be expected to respond (Keefer and Vlaicu, 2007). The budgetary trade-off they make in favor of creating more private goods risks long-term stability; they may maximize their chances of staying in power at the risk of a volatile end.

Credible Commitment and Growth

The ability of a state to foster or hinder growth has much to do with its ability to project the assurance of credible rule enforcement. Citizens will be willing to make investments in the projects of nonkin in a rule-of-law system when the rights of all economic actors are heeded. The scope and scale of voluntary exchange will increase and an accelerated rate of capital accumulation will enable states to grow stronger by borrowing or taxing the holders of capital. But when economic advantages are ceded on the basis of prior social ties, or on political access, networks of exchange will be limited by dependence upon personal sources of information, the best plans of the best organizers will fail to materialize. Social responsibility and trust will be limited to the extent of kinship networks.

Strong but limited government has proven to be the most certain path to bind the hands of rulers to time-consistent policy options. The simple reason is the inadequacy of planning for directing the creation of a complex system. The complex tasks that economic development requires cannot be anticipated or performed with certainty by central design. The engineering of a highly complex system requires an evolutionary rather than a centrally planned approach. States are not built, they are grown.

Bibliography

Fukuyama, F., 2011. The Origins of Political Order: From Prehuman Times to the French Revolution. Farrar, Straus and Giroux, New York.

Gellner, E., 1983. Nations and Nationalism. Cornell University Press, Ithaca.

Keefer, P., Vlaicu, R., 2007. Democracy, credibility, and clientelism. Journal of Law, Economics, and Organization 24 (2), 371–406.

Kydland, F., Prescott, E.C., 1977. Rules rather than discretion: the inconsistency of optimal plans. Journal of Political Economy 85 (3), 473–492.

Mayr, E., 2001. What Evolution Is. Basic Books, New York.

North, D.C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, New York.

Root, H.R., 2013. Dynamics Among Nations: The Evolution of Legitimacy and Development in Modern States. MIT Press, Cambridge, Mass.

States and Civilizations, Archaeology of

EM Brumfiel, Albion College, Albion, MI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 14983–14988, © 2001, Elsevier Ltd.

Abstract

This article begins by defining 'civilization' and 'states' and tracing the intellectual history of these two concepts in Western thought. It then introduces the term 'agrarian state' as a less biased, more informative designation for these ancient societies. The key characteristics of agrarian states are described. For the economy, key characteristics include: intensive agriculture, taxes or tribute, craft specialization, long-distance trade, and (sometimes) markets. For social life, the key characterisitcs include social division by class, gender, and ethnicity; coercive institutions including law courts, police, and armies; and (sometimes) cities. For cultural life the key characteristics include: an elite ideology expressed in religious ritual, monumental architecture, and writing, and ideologies of resistance expressed in traditionalism, ethnicity, and revitalization movements. The article concludes with a discussion of theories for the rise of agrarian states (functionalist and conflict theories) and for the fall of agrarian states (external and internal causes).

Civilizations

'Civilization' has had two meanings in reference to ancient societies both of which require critical review. First, civilization has designated complex prehistoric cultures with productive economies, class stratification, and striking intellectual and artistic achievements. Civilization in this sense has contrasted with 'primitive' or 'tribal' societies i.e., societies with modest subsistence economies, egalitarian social relations, and less extravagant art and science. Civilizations have been equated with progress and cultural superiority, a view promoted by European elites of the early modern era. Using ideas of civilization and progress, these elites assumed the role of civilizing agents to legitimate their domination of the working class at home and colonial peoples abroad. Modern anthropologists reject narratives of civilization and progress, preferring less ethnocentric accounts of non-Western cultures. Anthropologists now use more neutral terms to refer to the complex cultures of the ancient world, calling them complex societies or archaic states.

Civilization has also referred to the enduring cultural traditions of the specific regions of the world. Distinctive cultural principles characterized large geographical areas of the world for long periods of time, outliving the expansion and collapse of individual polities. For example, the idea of cyclical time was a fundamental principle among the Olmec, Maya, and Aztec people of ancient Mesoamerica, and it is an idea that persists today among indigenous peoples in Mexico and Guatemala. While contemporary anthropologists recognize the persistence of such orienting concepts, they reject the possible inference that these long-lived ideas imply cultural stasis. In the past as in the present, old concepts were reinterpreted in the face of new situations. Nor do these enduring principles imply uncontested social harmony. People of different social standing might share a set of cultural principles, but still arrive at different conclusions regarding appropriate behavior.

States

State has been a more useful term for analyzing complex societies. States are powerful, specialized institutions for public decision-making and control. State government was the most important element of complex ancient societies. The state created a surplus economy that supported many of the other institutions characteristic of complex societies: infra-structural investments, administrative bureaucracies, full-time craft specialization, long distance trade, market exchange, law courts, armies, and priesthoods. Every archaic state had among its key institutions: (1) some form of surplus extraction e.g., corvée, tribute, or taxation; (2) a bureaucracy that supervised the upward flow of tribute payments and the downward flow of tribute demands, and (3) coercive institutions, including armies, law courts, and police, to ensure tribute payment.

Archaeologists once characterized archaic states as having relatively simple systems of class stratification: nobles and commoners, or tribute-receivers and tribute-payers. But more recently, archaeologists have emphasized the diversity of groups within the state, their complex interactions, and the state's chronic instability. Archaic states were constantly threatened by both internal and external forces. High-ranking nobles led factions that competed for positions of rule. Provincial leaders stood ready to declare their independence from state control. Lineages and ethnic groups resisted the state's appropriation of their collective assets. Households tried to evade the state's demands on their goods and labor. The history of states and civilizations is the history of strategy and counter-strategy deployed by oppositional groups, leading to the development of complex societies and their dissolution.

Key Characteristics of Archaic States – Economics

Economic activity in archaic states was guided by the principles of efficiency, security, and control. All economic actors took into account the efficiency (e.g., the cost:return ratio) of economic alternatives. But peasants and household-based craft specialists often sacrificed efficiency for security; that is, they avoided economic alternatives that placed their annual subsistence at risk. States, on the other hand, often sacrificed efficiency for control; that is, they organized some forms of production so as to monopolize output rather than to maximize returns. Peasants, too, were sometimes guided by their

desire for control; they were inclined to produce goods whose attributes (e.g., perishability) made state appropriation less likely.

Agricultural and Pastoral Systems

Archaic states were characterized by highly productive agricultural systems. The vast majority of these states were dependent upon agriculture augmented by some form of water-control, terracing, and/or fertilizer. Mexico is famous for its chinampas (raised fields), Peru for its terrace systems, Mesopotamia for its irrigation canals, and China for its rice paddies. Such systems enhanced both agricultural productivity and reliability.

At one time, archaeologists believed that states came into existence in order to construct and manage intensified agricultural systems as a response to the problem of population pressure (this is the 'hydraulic hypothesis' of state formation). But periods of state formation do not coincide with periods of population pressure, and the study of contemporary irrigation systems in tribal societies suggests that such systems can be built and managed without centralized forms of government.

States did have good reason to construct agricultural facilities. First, such systems provided income to the state. Agricultural facilities constructed with corvée labor represented prime agricultural land that the state could claim as its own without alienating existing property owners. Second, agricultural facilities attracted tax-paying settlers. Peasants were drawn to newly available, high quality land, and they were willing to pay for access to it. Third, agricultural facilities expanded food production within the immediate hinterlands of political capitals, making it possible to feed the urban populace. The desire for more land, greater crop security, and larger quantities of marketable food also led peasant households to invest in agricultural facilities on their own. State- sponsored agricultural investment can be distinguished from peasant investment through careful study: uniformity in the timing and techniques used in facility construction are the best indicators of state sponsorship.

The long-term environmental consequences of agricultural development was a potentially destabilizing factor in all ancient states. In Mesopotamia, irrigation led to the salinization of the soil, first causing agriculturalists to switch from wheat to more salt-tolerant barley and then to abandon salinized areas all together.

Tribute, Taxes, and Private Estates

All archaic states needed wealth to fund their centralized power. Taxes could be paid in labor, goods, or currency. A tribute in labor was least likely to provoke peasant resistance, since labor could be 'reciprocated' with food and drink, presented to workers by the state in a party-like atmosphere. This form of taxation may characterize the earliest states (i.e., states that still had to contend with kinship groups and kinship norms) or states that legitimated their rule with ideologies of collective wellbeing. A tribute in goods may have its origins in the spoils of war, tribute representing spoils paid on a sustained-yield basis. Agricultural staples and woven cloth were the most common forms of tribute goods, but metals, precious stones, and other exotic valuables were also collected. Tribute foodstuffs were used to feed state personnel. Tribute luxury goods enabled rulers to reward loyal servants and to control the formation of social identities.

Because of their bulk, tribute goods were often stored in provincial warehouses, potentially funding revolts by provincial leaders. Substituting a monetary tax for tribute goods reduced transport costs and provided more centralized control over wealth. But monetary taxation occurred only in conjunction with market systems, where money could be converted into needed goods and services.

Tribute demands often transformed rural landscapes and households. Greek farm families under Roman rule abandoned their rural farmsteads and moved to towns where economic diversification could supplement their incomes and provide the necessary cash for taxes. In the Valley of Oaxaca, the state's demand for maize led to the cultivation of more distant fields which, in turn, altered routines of food preparation. More dry foods (i.e., tortillas) were made so that meals could be carried and eaten away from home.

Peasant households sometimes tried to escape tribute payments. In Mexico under the Aztecs, maguey cultivation expanded, perhaps because its perishable sap was not easily appropriated by the state. In Mesopotamia, peasant families wishing to avoid taxation sometimes fled to the mountains, becoming nomadic pastoralists.

Urbanism

Archaic states provided the setting for the earliest cities i.e., large, socially-heterogeneous, permanent settlements. These cities were religious and political centers; their economic functions were much less important than is the case for modern cities. Cities housed the administrative machinery of the state, usually in a multi-roomed palace complex. Highranking nobles, bureaucrats, and military personnel resided in the capital so that their activities could be directed and monitored by the ruler. Most state capitals also contained shrines, temples, and the residences of high-ranking priests. The concentration of political elites in the capital provided a clientele for enterprising craftspeople, merchants, and other service providers, who swelled the urban populace. Early cities did not supply many of the tools or meager furnishings used by rural peasants; these were usually produced in hinterland towns and villages, by parttime craft specialists.

In several cases, the growth of ancient cities was accompanied by the emptying out of the countryside. Both Uruk (Mesopotamia) and Teotihuacan (Mexico) grew explosively at the expense of rural settlements. In these cities, peasant farmers and farm laborers made up as much as 75% of the urban population. Both push and pull factors may have moved farmers from the country to the city. The security, sanctity, and commercial opportunities of the city provided powerful draws; at the same time, rural people may have been forcibly removed to the cities in order to facilitate supervision by the state and to loosen their claims on the land.

Cities developed only minimally in early Egypt. In Egypt, the concentration of population along the banks of the Nile, and the ease of water-bourne communication up and down the

river, may have made it unnecessary to concentrate elites and rituals in urban centers.

Markets

Markets often appear in conjunction with cities. Anthropologists are divided on the issue of whether markets developed from the 'top down' or from the 'bottom up.' 'Top down' theorists postulate that markets developed from the demands of urban elites for food and status goods. 'Bottom up' theorists view markets as developing from the efforts of peasants to enhance the efficiency and/or economic security of their households by participating in specialized production and exchange.

Regardless of their origins, markets attracted the attention of political leaders. Market taxes were an important source of revenue to rulers, and markets provided an efficient means of provisioning cities. States provided both direct and indirect stimuli to market growth. Rulers promoted markets by offering merchants protection against thieves and bandits and maintaining market order. Indirectly, states stimulated market participation by peasant households. Tribute demands forced peasants to become more efficient producers, engaging in specialized production and market exchange. In addition, the social hierarchies that accompanied state formation sometimes prompted peasants to consume specialist-produced goods to enhance their social standing.

In the long run, markets weakened the state's control over the economy. Merchant guilds, formed during periods of state collapse, resisted state control of the economy. The wide array of goods offered for market sale made it difficult to enforce the state's sumptuary rules limiting certain forms of food and dress to state-designated groups. Buying and selling conferred economic power upon individuals who were not state functionaries.

Full-Time Craft Specialists

Many forms of craft specialization antedate the origins of states. In tribal societies, it is not unusual for households or entire villages to specialize in ceramic production, stone tool production, or even metallurgy. This craft activity is carried out on a part-time basis in conjunction with basic subsistence activities. Craft products are distributed in reciprocal exchanges to strengthen social ties and/or to acquire goods from other specialists. Rural, part-time specialization of this sort was also present in archaic states, often enhanced by the opportunities for exchange provided by urban markets.

In situations of incipient social inequality, some craft specialists produced very elaborate goods that required heavy expenditures of labor. These 'hypertrophic' goods (such as embroidered textiles, jade earspools, stone sculptures, heavy bronze vessels, and gold jewelry) were commissioned by aspiring leaders to legitimate their claims to social worth. The craftsmen filling these commissions were part-time specialists either living in their own households or as members of the leader's high-ranking family.

The earliest full-time craft specialists in archaic states were the dependents of state rulers. Rulers supported craft specialists to maintain exclusive control over their output. Controlling sumptuary goods enabled rulers to consolidate their power by rewarding loyal servants and controlling social identities. Thus, for example, the Inka ruler supported a certain number of yanakuna and aqllakuna, full-time specialists in metallurgy and elaborate textile production. Gold and silver jewelry and embroidered textiles were used by the ruler to denote statuses within the state hierarchy. Maya rulers sponsored elite kinsmen who specialized in the production of elaborate painted ceramic vessels. Used in royal funerary rituals, these vessels confirmed the noble status of families. Egypt's pharaohs supported entire towns of craftsmen dedicated to the preparation of royal tombs. Royal patronage accounts for the high levels of skill and artistry that characterize the finest works of archaic states.

Full-time craft specialization also emerged in more commercial contexts when a high demand for craft goods (usually for export trade) combined with developed transportation networks and a steady supply of marketed food. These conditions were met most fully in the Mediterranean world where the water-bourne transport of food staples from one broad region to another underwrote really large-scale production and exchange. In late precolonial India and China, reliable agricultural systems combined with riverine and oxcart transport yielded similarly high levels of commercial production and exchange.

Key Characteristics of Archaic States – Social Organization

Social heterogeneity is the defining characteristic of complex societies. In addition to the occupational diversity described above, the intersection of class, gender, and ethnicity created an array of social statuses and life histories in the ancient world.

Class Stratification

In the earliest states, the division between classes was based on political rather than economic status. A key division within society was between those who paid tribute and those who received it. Tribute was paid to the ruler who distributed it to servants of the state (usually close relatives). Key personnel were endowed with their own streams of tribute flow. Thus, the tribute-collecting class in archaic states comprised the ruler and his close relatives, high-ranking military personnel, government administrators, priests, and surviving elements of the pre-state nobility. For the governing class, wealth, power, and prestige coincided and were mutually reinforcing. Conversely, tribute-payers suffered a lower standard of living, a lack of formal power, and low social status.

Archaeological and historical research reveals significant variability within the commoner class. For example, Aztec households varied by size, composition, wealth, and ethnic affiliation. Rural households chose different combinations of agricultural production and craft activity. This variability complicates the traditional view that commoners were a homogeneous class. It suggests that commoners were divided by demographic, economic, and social factors.

Slaves constituted a class below commoners. Slaves were captured in warfare, purchased through the slave trade, or sentenced to servitude as punishment for crimes or failed debts.

Among Aztecs and Mayas, slaves were used as household servants, but they were never very numerous. In Mesopotamia, large numbers of female war captives worked at textile production, and this may have set a precedent for the use of slave labor in commodity production. The use of slave labor in commercial enterprises was extremely important in highly commercialized economies of ancient Greece and Rome.

Private property sometimes developed out of warfare-based economic systems. In China and Aztec Mexico, lands that were seized in warfare and redistributed to loyal soldiers constituted an early form of private property. In Mesopotamia, private property developed long after states arose, through the gradual sale of lands belonging to kinship units to secular or religious elites.

Gender

In *The Origin of the Family, Private Property, and the State* (1972 [1884]), Frederick Engels argued that the development of private property brought about the state (to maintain wealth differences) and the patriarchal family (to insure the legitimacy of inheriting offspring). In the process, women's status declined from a position of approximate equality with men to an inferior position. This was the 'world historic defeat of the female sex.' Modern historic and archaeological scholarship suggests that institutions such as private property and the patriarchal family are not universal characteristics of ancient states, nor is the subordination of women. Nevertheless, gender relations and ideologies were frequently renegotiated during the process of state formation, with results that were detrimental to women.

The status of women in states was undermined by several factors. First, states were commonly created and maintained through warfare, and war was usually men's work. Through warfare, men gained access to spoils, a form of wealth that was not household-dependent. In addition, rulers often glorified male warriors and rewarded them with administrative positions in order to maintain their loyalty. Thus, warfare provided avenues of wealth, prestige and power for men that were not available to women. Furthermore, rulers dismantled kinship groups, which were nodes of resistance to state rule. These kinship groups often provided the primary rationale and defense of women's rights in ancient societies, and when they were destroyed, women were left without social recourse. It is not surprising, therefore, that many archaic states favored men. Ancient Rome and ancient China both awarded great authority to male household heads. The Classic Maya and the Aztecs conferred public prestige upon male warriors, and men dominated these state governments.

However, the subordination of women to men was not universal. States that legitimated their rule with ideologies of collective wellbeing did not emphasize warfare and did not stress gender difference. These states might have included Teotihuacan (Mexico), the Harappan states of the Indus Valley, and Egypt under the pharaohs. In the Andes, a long tradition of parallel descent favored gender equality. Even in states where patriarchy was the norm, high-ranking women were wealthier and more powerful than lower class men (i.e., class trumped gender). In peasant families, women's contribution to the household economy often translated into approximate gender equality. In addition, women found forums for resistance outside of the state. Female oracles provided social critique in Oaxaca Mexico, ancient Greece, and the Andes.

Ethnicity

Ethnicity is an identity based upon a presumption of shared history and common cultural inheritance. Ethnic identity is shaped by both ethnic affiliation and ethnic attribution. Ethnic affiliation refers to individuals' own sense of group membership and the characteristics of the group as defined by its members. Ethnic attribution concerns the characteristics of the group as defined by outsiders.

States acted opportunistically and inconsistently in dealing with ethnicity. Sometimes they suppressed ethnic affiliation to weaken the resistance of subject groups to the state. At other times they encouraged ethnic affiliation to accentuate division within the commoner class. States also used derogatory ethnic stereotypes to legitimize their exploitation of subject peoples. For example, Aztecs depicted their Otomí subjects as lazy, improvident, untrained blockheads and gaudy dressers. Stereotyping and ethnic prejudice generally heightened ethnic consciousness and perpetuated ethnic heterogeneity within ancient states.

Key Characteristics of Archaic States – Religion and World Views

State Religion

All archaic states had an official religion with standardized temples and full-time priests. Rulers' patronage of religion may have sprung from either piety or cynical efforts to manipulate the subordinate class. Either way, religion provided a powerful political resource that was actively exploited by ancient rulers.

Many archaeologists argue that state religion was used to forge a unified polity out of distinctive subject peoples and to legitimize the ruler's authority. For example, Chinese rulers tried to force cultural uniformity upon all the groups that they conquered, and Confucian philosophers declared that just rulers held the 'Mandate of Heaven.' The Incas' effort to impose cultural uniformity on their subjects is archaeologically visible in provincial capitals where large plazas were constructed using Inca architectural forms. Local people gathered in these plazas to attend state-sponsored rituals.

In contrast, Aztec state religion was not geared to produce a uniform world view within the empire, nor to sanctify Aztec rulers in the eyes of their subjects. Rather, Aztec religion was intended to win the loyalty of a relatively small target group, the young men who formed the core of the Aztec army, by imbuing their military exploits with cosmological significance. In the Aztec hinterlands, distinctive forms of ceramic figurines suggest that commoners did not accept the state's ideology; on the contrary, commoners successfully maintained their own world view. Distinctive commoner ideologies may have been present in all archaic states, challenging the official point of view.

Monumental Architecture

The earliest monuments were megalithic tombs and earthworks built by groups that were still organized using kinship

principles. These monuments were probably constructed during communal rituals, when aspiring leaders sought social prominence by sponsoring group festivities. Under state rule, monumental construction increased in scale, creating some of the most striking icons of the ancient world: Egypt's pyramids, Mesopotamia's ziggurats, the colossus of Zeus at Rhodes. It is estimated that the great Pyramid of Khafu (Cheops) required 13 440 000 person-days of labor.

Some monuments may have been constructed by a willing populace, motivated by civic pride and religious devotion. This would have been most likely in states that legitimated their rule with ideologies of collective wellbeing. At Teotihuacan, for example, the Temple of the Sun and the Temple of the Moon may have been constructed in a surge of popular enthusiasm. In states relying upon coercive rule, monument construction expressed submission to the state. Aztec rulers invited subject communities to participate in the construction of the Great Temple in Tenochtitlan, and refusals to participate were considered acts of rebellion.

Through their size, monumental structures symbolized the ruler's exclusive link to supernatural forces. Monumental structures also expressed overwhelming power of the state and the fact of commoners' subordination. Large buildings and massive sculptures imposed themselves on social space and human psyches. In the words of Antonio Gilman (1996), they showed 'who was boss.'

Writing

Writing developed in archaic states for many different reasons. The earliest texts, from Mesopotamia at 3300 BCE, recorded the economic transactions of large temple estates. In ancient Egypt and Mesoamerica, writing was first used to record the exploits of rulers. The earliest Chinese inscriptions were questions carved on turtle carapaces and the shoulder blades of cattle for the purposes of divination. In all cases, writing was an elite activity. Writing can be regarded the product of elite efforts to establish and to preserve a single, authoritative record of events and to make its voice stronger than the spoken words of its subjects.

The Rise and Fall of Archaic States

In the 1970s, archaeologists believed that states arose to manage environmental problems. Since states are powerful administrative institutions, state officials can function as highly effective problem solvers. They can construct large-scale irrigation networks, insure populations against crop failure, administer markets, organize warfare, and sponsor long-distance trade. Archaeological evidence confirms that archaic states did oversee some of these activities. However, the evidence for these activities usually dates to a few centuries after state formation; thus, the problem-solving capacity of states does not appear to account for their origins.

More recently, archaeologists have begun to account for state origins by examining how individuals acquire social power. In one model, dynamic, ambitious leaders assemble coalitions of followers by sponsoring feasts and distributing gifts. Each leader with his followers then competes against other leaders and followers to establish dominance, first on the local, then on the regional, level. In these systems, personal prestige is emphasized, and dominance may be maintained through naked force. In a second model, leaders attract a broader following, by appealing to ideologies of collective wellbeing. State formation is more 'volunteeristic,' resembling a religious movement, and there is broad-based participation in collective rituals and public monument-building. Leaders are more anonymous, and when force is used, it is used on behalf of society as a whole. It has been suggested that the flamboyant Maya and the Shang rulers might exemplify the outcome of the first (individualizing) strategy, whereas the Teotihuacan (Mexico) and Harappan states might be products of the second (corporate) strategy.

Archaic states were complex, often fragile, social structures. Social dominance required balancing the needs and power of diverse social segments. Sudden shocks and long-term problems threatened the delicate balance. Over-population, excessive tax burdens, climatic deterioration or invading barbarians, all threatened to topple archaic states at one time or another. Egypt's Old Kingdom collapsed when provincial officials became increasingly independent of the pharaoh's rule; the Middle Kingdom fell because bureaucrats at the center absorbed too much of the state's wealth. Competition among the independent Classic Maya kingdoms led the Maya to overexploit their resources and, finally, to abandon their homes. The surprise is not that archaic states eventually collapsed, but that ruling elites could devise institutional structures that enabled states to endure, sometimes for centuries at a time.

Bibliography

Alcock, S.E., D'Altroy, T.N., Morrison, K.D., Sinopoli, C.M. (Eds.), 2001. Empires: Perspectives from Archaeology and History. Cambridge University Press, Cambridge, UK.

Blanton, R.E., Feinman, G.M., Kowalewski, S.A., Peregrine, P.N., 1996. A dual-processual theory for the evolution of Mesoamerican civilization. Current Anthropology 37, 1–14.

Childe, V.G., 1950. The urban revolution. Town Planning Review 21, 3–17.

Classen, H.J.M., Skalník, P. (Eds.), 1978. The Early State. Mouton, The Hague.

DeMarrais, E., Castillo, L.J., Earle, T., 1996. Ideology, materialization, and power strategies. Current Anthropology 37, 15–31.

Engels, F., 1972[1884]. Origin of the Family, Private Property, and the State. In: Leacock, E. (Ed.). International Publishers, New York.

Feinman, G.M., Marcus, J. (Eds.), 1998. Archaic States. School of American Research, Santa Fe, NM.

Gilman, A., 1996. Comments on 'Ideology, materialization, and power strategies.'. Current Anthropology 37, 56–57.

Gledhill, J., Bender, B., Larsen, M.T. (Eds.), 1988. State and Society: The Emergence and Development of Social Hierarchy and Political Centralization. Unwin Hyman, London.

Patterson, T.C., 1997. Inventing Western Civilization. Monthly Review Press, New York.

Richards, J., Van Buren, M. (Eds.), 2000. Order, Legitimacy and Wealth in Ancient States. Cambridge University Press, Cambridge, UK.

Sanders, W.T., Webster, D., 1988. The Mesoamerican urban tradition. American Anthropologist 90, 521–546.

Silverblatt, I., 1988. Women in States. Annual Review of Anthropology 17, 427–460.

Stein, G.J., 1998. Heterogeneity, power, and political economy: some current research issues in the archaeology of Old World complex societies. Journal of Archaeological Research 6, 1–44.

Yoffee, N., Cowgill, G.L. (Eds.), 1988. The Collapse of Ancient States and Civilizations. University of Arizona Press, Tucson, AZ.

Statistical Analysis Packages

Hans-Jürgen Andreß, Universität zu Köln, Köln, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines the main features of present-day statistical software by reviewing its development from the (late) 1950s to the present time. It puts a focus on the management and analysis of social science data. The article concludes with a discussion of present-day software alternatives for social scientists.

Introduction

Since World War II, social sciences have become increasingly empirical. This does not mean that before World War II social sciences were only doing theories without any links to real world problems or that there were no empirical investigations of real world phenomena. But it is certainly true that empirical studies analyzing quantitative data were clearly outnumbered; and the few that existed were using a methodology that was much less statistically sophisticated than the quantitative analyses that are published nowadays in top social sciences journals. The main reason for this development is obvious: It is the availability of user-friendly software that allows *social scientists* to read, manipulate, and analyze complex data with a few keystrokes or some simple commands, even if they are no specialists in statistics or information science. And it is also true "that the most important changes in *statistics* in the last 50 years are driven by technology. More specifically, by the development and universal availability of fast computers and of devices to collect and store ever-increasing amounts of data" (De Leeuw, 2011; emphasis added).

Quantitative data can be analyzed with all kinds of tools: with paper and pencil, pocket calculators, office software, and specialized statistical software. This article focuses on *statistical software* and not on software that can also be used for statistics (*software for statistics*; e.g., a spreadsheet program). It does so by putting a special emphasis on social science data. While a statistical program simply needs a rectangular matrix of numbers to do its calculations, social science data have a least two characteristics that need special attention. First, the numbers are often not the result of quantitative measurement, they rather indicate ('code') certain characteristics of the unit of analysis (e.g., being a male or female respondent). Second, sometimes, for certain units, no number is available, either because the corresponding variable does not apply to that unit (e.g., the type of job for a respondent that is unemployed) or because the corresponding information is not available (e.g., when an employed respondent refused to provide information about his job). As a consequence, statistical software specifically designed for social science data must provide possibilities to complement the data matrix with additional information about the codes (labels) and missing data. Moreover, statistical computations have to control for the missing information. Finally, but this applies to all kinds of quantitative data, data sets often require a great deal of manipulation before they are ready for analysis. Therefore, any statistical software has to include some functionality for data manipulation, sometimes nearly as complex as a data bank software that allows to index, select, append, and merge all kinds of data sets. One should also note that nowadays statistical software is not confined to the analysis of numbers. The data may also include text and thus statistical software also provides (often limited) possibilities for text analysis.

Due to the rapidly changing software market, it is not possible (and not useful) to give a complete overview of all available statistical software. Instead, this article outlines the main features of present-day statistical software by reviewing its development from the (late) 1950s to the present time (Section A Brief History of Statistical Analysis Packages). The article concludes with a discussion of our present-day alternatives (Section What Choices Do We Have Today?).

A Brief History of Statistical Analysis Packages

The following historical sketch describes the development of statistical analysis packages in three steps that roughly summarize the innovations of the 1960s/1970s, the 1980s, and the 1990s. To understand the historical legacy of the first packages (and the innovations of their followers), we start with a résumé of the tools that were available to social scientists before the advent of statistical analysis packages.

The following historical sketch is closely linked to the development of information technology. From the very beginning, statistical applications were determined by the amount of computation necessary to draw conclusions from the data. In the early times of modern statistics computing time was clearly a restriction. Salsburg (2002) estimated that Sir Ronald Fisher's computations for one table in his 'Studies in Crop Variation. I' took about 185 h on his mechanical desktop calculator. Nowadays, with first the electronic (mainframe) computer and then the personal (micro) computer, computing time is increasingly less of an obstacle. Moreover, the development of programing tools and interfaces made this hardware more and more user-friendly. In addition, the refinement of graphical interfaces has revolutionized the display and exploration of data.

The Time 'Before' Statistical Analysis Packages

Before the advent of statistical software on electronic computers social scientists had to live with mechanical computers and tabulating machines. Larger data sets that could not be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03023-3

handled manually had to be transferred on cards using specialized coding systems that represented the information in digital form by the presence or absence of holes in predefined positions. These punch cards are also called Hollerith or IBM cards, because Herman Hollerith invented both the storage medium and the processing technology (tabulating machines) to analyze the stored information. He founded the Tabulating Machine Company, which was one of four companies that merged to form Computing Tabulating Recording Company, later renamed IBM. The Eleventh United States census in 1890 was the first census to use this technology. From then on, until the 1950s, this technology was used in all industrialized countries to process business, government, and scientific data. In the 1950s, magnetic tapes and new electronic input devices were introduced to store mass data. However, not till the 1960s, the punch card technology was gradually replaced by magnetic tapes as more capable electronic computers became available.

The early times of (quantitative) social research were dominated by punch cards and tabulating machines. Since the punch card provided only 80 columns to store information, all kinds of coding systems were invented to store as much as possible information on one card. If one card did not suffice for the data of one unit of analysis, one had to use several cards for each unit, but then also to make sure that the whole card deck with all the information for the entire sample remained in correct order. Biographies and obituaries of older social scientists are full of anecdotes about how they fought with tabulating machines, multiple punch systems, and disordered card decks that fell down on the floor. It was not before the advent of the transistor based (so-called second generation) electronic computer in the 1960s (e.g., the IBM 7090), that some larger pieces of software developed that allowed nonspecialists to manage and analyze social science data.

One was the DATATEXT package (Armor and Couch, 1972) developed at the end of the 1960s in Harvard, the other was a package of statistical routines for biomedical research (BMD), which was developed at the end of the 1950s at the University of California in Los Angeles (Dixon and University of California, 1967). (For better readability, in this article all package names are written in capital letters.) Both were still carrying the footprints of the punch card times. DATATEXT provided, besides statistics, all kinds of routines to cope with 'messy' card decks and multiple punch systems. BMD was simply a collection of programs; each one for a specific statistical method (e.g., analysis of variance), but all using the same command language and input–output system that expected the data to be prepared in advance in a rectangular data matrix. Program commands had to be written on punch cards using certain columns for command names and other columns for the command parameters. While DATATEXT used command keywords, BMD was controlled entirely by numbers.

The exclusive mode of processing and analyzing data was the so-called batch mode. Program commands were written on punched cards and, together with a card deck of data, were input into the computer. Depending on the power of the computer and the number of other active batch jobs on the computer, it took some time (sometimes even days) until the user would see an output of the program results. Very often, the output only showed an error message, because the user misspelled a command or did not follow the fixed input scheme for the commands. Then the whole process had to be repeated again, often several times until all errors were eliminated. At these times, the statistical analyses were simple (compared to nowadays level of sophistication), but the process was very time consuming and not all social scientists (especially outside the US) had access to electronic computers. Not surprisingly, only a small and selective group of social scientists used this technology.

The Time of the Mainframes

The explosion in the use of computers began with the third generation of electronic computers using integrated circuits or microchips. The IBM System 360, first shipped in 1965, is an example of such a third generation computer. These were still large (mainframe) computers, both in terms of size and prize. Hence, only larger organizations could afford them. In most cases, they were part of a computing center that offered computer time as well as input and output devices to a larger group of users. Input devices could still be card readers, but increasingly became computer terminals equipped with a keyboard and a screen. In principle, this allowed some kind of interaction between the user and the computer or its software. Several users could work concurrently with the computer, as operating systems included routines to share computing resources between users (so-called time-sharing systems).

This new generation of electronic computers was very consequential for all statistical analysis packages that existed so far. Designed for second generation computers and their operating systems, they were simply incompatible and could not be used anymore. As a collection of single statistical programs, BMD easily made the transition and, equipped with a new name: BMDP, used keywords (and not numbers) for the commands (University of California and Dixon, 1975). However, BMDP's target group were health scientists. Social scientists missed an alternative to DATATEXT, whose transition to the new generation of electronic computers came too late.

At that point, a new software, the Statistical Package for the Social Sciences (SPSS), had already started its success (Nie et al., 1970). It was developed around 1968 by social scientists at the University of Chicago and its fast diffusion in the market showed that its designers had built a product that satisfied the needs of its target group, the social scientists. One of the reasons for its success story was the program manual, which – besides describing the commands – provided applied introductions into all statistical techniques; and these introductions were easy to read even for those without a comprehensive statistical education (however, at the cost of not documenting the algorithms used for the statistical techniques). Wellman (1998) considered it one of sociology's most influential books because it allowed ordinary researchers 'to do their own statistical analyses'. Another reason for the success of SPSS was its availability on different computers and operating systems. In the 1970s, SPSS dominated as *the* tool for social science data analysis, while BMDP remained a specialist tool for the statisticians.

Although written for a new generation of computers, both packages still had features that reminded of the punch card era. A prominent example is that SPSS expected parameters for its

commands to start exactly in the 16th column of the command line. Although, as mentioned, user–machine interaction was possible, BMDP and SPSS expected a batch of commands (a 'program'), which then would be executed without intervention by the user. The way designers of statistical analysis packages thought about their users at that time was still not much different from the former 'batch mode' times. When intervention was not possible, and when program output was mostly provided by line printers, possibly in a distant computing center with quite some turnaround time, users chose the 'safe' way of producing results. They rather asked for everything ("…ALL BY ALL, STATISTICS ALL"), instead of asking for the results that were important. Moreover, most of the data management functions of DATATEXT were lost. BMDP as well as SPSS expected a rectangular matrix and the user had to use the supplied data generation commands for data management and cleaning.

All in all, statistical analysis packages introduced with the third generation of electronic computers were characterized, on the one hand, by providing easy access to the statistical tools of data analysis, but also, on the other hand, by the absence of any tools for data management. Moreover, they did not use the potential for user intervention and processed the commands in batch mode. Finally, a more technical point: The typical storage medium of that time was the magnetic tape, which could only be read and written sequentially. Random access memory was scarce and mostly processed sequentially (as if it would be a magnetic tape). Certainly, sequential storage media are a severe restriction for complex data bank systems. Nevertheless, the data management gap was filled in the second half of the 1970s by two software products: SIR/DBMS and SAS.

The mnemonic SIR stands for a data bank management system (Robinson et al., 1980) that was quite popular in the social sciences because it provided, on the one side, simple commands to manage a relational data bank and, on the other hand, the ability to output the contents of this data bank to any kind of statistical analysis package. For example, the German Socio-Economic Panel Study, a renowned panel study of the German resident population was managed for quite a while with SIR/DBMS. However, as more and more statistical analysis packages included the most important data management procedures into their functionality, SIR/DBMS more or less became a niche product.

SAS was developed almost as the same time as SPSS by computational statisticians at North Carolina State University (SAS Institute, 2013). While at the beginning a statistical software for large agricultural data sets, it developed into a comprehensive statistical package with often more functionality than its competitors and with a strong emphasis on handling large and complex data. For a long time, it was available only on IBM mainframes, which clearly restricted its diffusion. Later on, when it was available on all kinds of machines it became quite popular and a serious competitor for SPSS. In 1976, the inventors of SAS formed the SAS Institute Inc. and increasingly marketed SAS as a high-end software product mostly targeted at researchers in private business. The accompanying price policy made SAS difficult to afford for researchers in publicly funded academic organizations and hence restricted its popularity in this part of the market for statistical software.

All in all, the most important statistical software packages during the era of the mainframes were (in alphabetical order) BMDP, SAS, and SPSS. According to De Leeuw, "the three competitors differed mainly in the type of clients they were targeting. And of course health scientists, social scientists, and business clients all needed the standard repertoire of statistical techniques, but in addition some more specialized methods important in their field. Thus the packages diverged somewhat, although their basic components were very much the same" (De Leeuw, 2011). This does not mean that no other statistical analysis packages were available. On the contrary, important centers of statistics or survey research and individual researchers developed similar systems: for example, to name just a few, MINITAB from Penn State, OSIRIS from the Institute for Social Research at the University of Michigan, P-STAT from Princeton, GENSTAT from Rothamsted Experimental Station, or ALMO developed by Kurt Holm.

Some of these 'other' packages also introduced important innovations to the design of statistical software. GLIM, for example, was notable for encouraging an interactive, iterative approach to statistical modeling, besides providing a unified framework for fitting a wide range of generalized linear models (the GLIM syntax could be summarized on two (!) sheets of paper). It also allowed users to define their own macros and hence, provided an open programing environment compared to the 'closed shops' of SPSS and its competitors. SYSTAT, to name another example, was a statistical analysis package that put a special emphasis on statistical graphics.

The Advent of the Microcomputer

Indeed, by the end of the mainframe era, the terrain was prepared for social scientists to do statistical analyses 'themselves'. The software was there, its use was increasingly included in university teaching, and even the computers came a bit closer to the users with so-called mini computers such as Digital Equipment's Vax or Hewlett–Packard's HP3000 that provided computer resources to smaller workgroups. Yet, it was IBM's personal (micro) computer that really 'democratized' the use of statistical analysis packages. Indeed, with a PC on one's desk, everybody could manage and analyze large amounts of data. Besides bringing (rapidly increasing) storage and computing resources to the individual researcher, the main innovation of this fourth generation of computers was its user interface. From the very beginning, the user had total control on what was going on in his computer. But quite soon, operating systems for microcomputers provided graphical user interfaces (GUIs) with windows, icons, menus, and pointers that made user–machine interaction much easier.

All major statistical analysis packages made the transition to this new world of personal computing. Commands could be written and executed in a command window or chosen from a menu, and the results were shown in a separate output window that could be copied and pasted into other software (e.g., in word processors for report writing). From now on, statistical analysis packages could be used interactively. Commands and their results could be easily tested and revised, if the results did not match the research interests of the user. Commands need not to be remembered or taken from an accompanying printed manual, because the packages included

help menus, and manuals were stored electronically with the program. Nevertheless, for replicating the analysis and for computer-intensive tasks, the final commands had to be stored in a syntax file that could be executed again in a batch-mode-like fashion; a small reminiscence of the 'good old times', but useful and necessary. What also remained from the mainframe time was the massive output. Statistical analysis packages transferred from the mainframe era were still showing all numbers that ever could be of interest. Obviously, the interjection of GLIM was not heard to show only those numbers that the user needs for making the next decision.

During the microcomputer era we also observed a second generation of statistical software that made use of the new technologies and did not have the baggage of the mainframe era. Some of them made heavy use of the new graphical capabilities and emphasized visualization and exploratory data analysis. JMP, a computer program developed in the 1980s by the JMP business unit of SAS Institute, was such an example. It was created to take advantage of the GUI introduced by Apple's Macintosh, but nowadays is available also for other operating systems. The most prominent feature of JMP is its ability to really look at multivariate data and to screen the data from different perspectives.

STATA is another example of second generation statistical software. It was created in 1985 by the UCLA graduate William Gould together with his colleagues at Computing Resource Center, a private company, which in 1993 moved to Texas and now is named StataCorp LP (Cox, 2005). STATA is also a general-purpose statistical analysis package and at first sight, looks very similar to its relatives from the first generation (BMDP, SAS, and SPSS). It also had a command line interface, and did not get a GUI until 2003. However, from the start, it emphasized computation speed, extensibility, and user-contributed code. Computation speed is of utmost importance, if one wants to support the interactive use of the program. To this end, STATA is reading and writing all the data into the (fast) computer memory, while its relatives read and write the data (often sequentially) from slower memory devices. Within STATA one can also write macros and programs and in doing so, extend the program functionality to one's own needs. Users can publish their programs on the Internet or on Statacorp's servers. STATA itself makes heavy use of the Internet, both to update the program, but also to download data and user code. More than its relatives (perhaps with the exception of SAS) it tries to tie in its users. One could even speak of the 'STATA community'. It organizes user conferences, has a book series and a peer-reviewed journal, provides an e-mail Listserver for user discussions, and the very informative program documentation, both in terms of applications and technicalities, is available with the program and online. Nowadays, STATA is very popular in economics, sociology, and political science, with a strong focus in academia. It is especially attractive for its computing speed, extendibility, and prize.

From Statistical Program Packages to Statistical Analysis Languages

In the early 1990s, another type rather than a new generation of statistical software came into play. These were not preprogrammed analysis packages, but rather programing languages with a strong focus on statistical analysis. All of them have the S language as common ancestor, which – starting from 1975 – grew out of Bell laboratories; hence, a rather old development and not a new generation (Chambers, 2008: pp. 475–478). According to De Leeuw, a statistician, "the statistical techniques […] were considerably more up-to-date than techniques typically found in SPSS or SAS" (De Leeuw, 2011: p. 5). But it remained a tool for the specialists, although it was freely available for some time to academic institutions. Also its commercial version S-PLUS hardly diffused into the wider circle of applied researchers.

The followers of S in the 1990s were LISP-STAT and R. Both languages are open source products and available on the personal computer, which explains – at least in the case of R – their popularity in and outside the community of statisticians. LISP-STAT was developed by Tierney (2009) and provided statistical tools embedded in a Lisp interpreter (Lisp is a programing language that is often used in artificial intelligence research). Using a (general) programing language as the environment provides lots of opportunities for extending LISP-STAT. According to the Web site, it puts "an emphasis on providing a framework for exploring the use of dynamic graphical methods". But LISP-STAT's development ceased by the end of the twentieth century.

R was written as an alternative implementation of the S language (Ihaka and Gentleman, 1996). Due to its extendibility by user-written programs and easy availability as an open source product, its dispersion in the market has increased dramatically in recent years. De Leeuw says: "R is many things to many people: a rapid prototyping environment for statistical techniques, a vehicle for computational statistics, an environment for routine statistical analysis, and a basis for teaching statistics at all levels. Or, going back to the origins of S, a convenient interpreter to wrap existing compiled code" (De Leeuw, 2011: p. 6). But the extendibility is also a problem. For example, it is quite possible that several user programs exist, which address the same problem. Which one to choose? Sometimes the documentation for this user-written code is insufficient and who tests whether the programing is reliable and efficient? If social scientists want to have the same basic functionality as their old statistical analysis packages, i.e., labeling variables and values, coding missing data, etc., they have to download several user-written programs to their base R installation. Which the most useful ones are, can certainly be found somewhere in the Internet. But how does the applied social researcher find this information, if it is not in textbooks on R, which by the way are mostly written by statisticians and for their audience.

What Choices Do We Have Today?

Without the innovations of information technology, the analysis of social science data would not have come so far. With the help of the computer, we collect and store data in virtually unlimited amounts; and it provides us with the computation power to do all kinds of sophisticated statistical analyses. For example, maximum likelihood as a method of estimation has been known for centuries, but it needed the electronic computer to apply it to real-world data. Moreover, classical parametric

procedures have been supplemented with computer-intensive, nonparametric methods such as resampling to compute standard errors and confidence intervals. Bayesian statistics have become popular in recent years due to the availability of Markov chain Monte Carlo methods, which allow sampling from probability distributions, which is critical in Bayesian Statistics. Finally, the availability of computing power and GUIs has made the dream of really looking at data come true.

Besides the development of the hardware, the Internet of computers worldwide has revolutionized the acquisition, storage, and access to social science data. Whereas the classical statistical problem was to make the most out of limited data, today the challenge is to process large data samples. Only with the power of present-day computers, social scientists have all the means to analyze these vast amounts of information.

With respect to software, our historical account has shown the alternatives that are available to social scientists today. Basically, there are three: (1) general-purpose programs like SPSS, SAS, or STATA (BMDP has become a niche product), (2) statistical analysis languages, most prominently R, or (3) other software that can also be used for statistics. As Brian Ripley, Professor of Applied Statistics at the University of Oxford, once said at a conference of the Royal Statistical Society: "Let's not kid ourselves: the most widely used piece of software for statistics is Excel" (Ripley, 2002). Naturally, statistical add-ons exist also for this third alternative (e.g., ANALYZE-IT, NUMXL, or SIGMAXL). Besides these main alternatives, there are lots of other programs often designed for specific purposes; e.g., MPLUS for structural equation modeling, HLM for multilevel regression, and many others. Wikipedia is an excellent information source for all kinds of software and provides an up-to-date list and comparison of statistical packages. The Journal of Statistical Software, according to its Web site, publishes "articles, book reviews, code snippets, and software reviews on the subject of statistical software and algorithms." Software reviews can also be found in major statistical journals.

This bulk of software and alternatives has pros and cons. One of the advantages is that users can compare. This does not only mean looking for the best alternative, it also implies that users can test the reliability of the available software. In most cases, social scientists will not have the expertise to check the accuracy of the underlying algorithms, but if alternative software does not reproduce the results of their favorite software, then possibly something is wrong and they should consult an expert. For example, regular tests of statistical software show that numerical accuracy cannot be taken for granted (see, e.g., Keeling and Pavur, 2007). One of the disadvantages is, of course, that the uninformed user is easily lost. This is especially true for open source products, as the example of R shows (see the discussion above). This does not mean that proprietary products are without problems, but with such a contested market for statistical software every commercial provider of statistical software has a high interest in providing a reliable product (naturally, the incentive is lower for providers of software that is not designed for statistics in the first place).

Our historical review has shown the rise and fall of various software products. As a final point, this raises the question of what makes a product successful in an obviously very dynamic market with significant competitors, some of them even providing their products for free. One precondition is, of course, being at the forefront of data management and data analysis. A second condition is user friendliness, which does not necessarily mean simplicity. Users accept that data analysis may be a complicated matter and therefore, will accept some complexity, but the easier this complexity is to understand through the handling of the program, the more user-friendly the software is. Third, having a clear and solvent target group seems to be a precondition for commercial success. SAS, focusing primarily on private business, is a typical example of this strategy (SPSS, now owned by IBM, seems to go in the same direction of focusing on users in private business). Academia, on the other hand, is an unreliable and unfaithful clientele. First, they know everything better and try to do it themselves. Second, academia, compared to private business, is rather poor. Hence, if a software provider is not able to tie in its users or has something to sell that others do not supply, it is quite probable that the software will not survive.

See also: Data Bases and Statistical Systems: Applied Social Research; Empirical Social Research, History of; Quantification in the History of the Social Sciences.

Bibliography

Armor, D.J., Couch, A.S., 1972. Data-Text Primer: An Introduction to Computerized Social Data Analysis. Free Press, New York.

Chambers, J.M., 2008. Software for Data Analysis: Programming with R. Springer-Verlag, New York.

Cox, N.J., 2005. A brief history of stata on its 20th anniversary. The Stata Journal 5 (1), 2–18.

De Leeuw, J., 2011. Statistical software – overview. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, Berlin/Heidelberg, pp. 1470–1473.

Dixon, W.J., University of California, 1967. BMD: Biomedical Computer Programs. University of California, Los Angeles. Health Sciences Computing Facility.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

Keeling, K.B., Pavur, R.J., 2007. A comparative study of the reliability of nine statistical software packages. Computational Statistics & Data Analysis 51 (8), 3811–3831.

Nie, N.H., Bent, D.H., Hull, C.H., 1970. SPSS: Statistical Package for the Social Sciences. McGraw-Hill, New York.

Ripley, B.D., 2002. Statistical Methods Need Software: A View of Statistical Computing. Presentation at the conference of the Royal Statistical Society, Plymouth, 3–6 September 2002. Retrieved 7.10.13 from: http://www.stats.ox.ac.uk/~ripley/RSS2002.pdf.

Robinson, B.N., Anderson, G.D., Cohen, E., Gazdzik, W.F., 1980. SIR Scientific Information Retrieval. User's Manual, second ed. SIR INC, Evanston, IL.

Salsburg, D., 2002. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Henry Holt and Company.

SAS Institute, 2013. SAS Company History. Retrieved 4.10.13 from: http://www.sas.com/company/about/history.html.

Tierney, L., 2009. LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics. Wiley, Hoboken, NJ.

University of California, Dixon, W.J., 1975. BMDP: Biomedical Computer Programs. University of California Press, Los Angeles.

Wellman, B., 1998. Doing it ourselves. In: Clawson, D. (Ed.), Required Reading: Sociology's Most Influential Books. University of Massachusetts Press, MA, pp. 71–78.

Statistical Analysis: Multilevel Methods

William M Mason, University of California, Santa Monica, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 22, pp. 14988–14994, © 2001, Elsevier Ltd., with an updated Bibliography supplied by the Editor.

Abstract

In hierarchical data structures, observational units at one level are nested within units at other levels. For longitudinal data on individuals, within-person observations are nested within individuals. Within-person (micro) observations are generally more like each other than observations sampled randomly across individuals (contexts). In generalized regression models, several approaches can accommodate this lack of independence: (a) random effects, (b) fixed effects, (c) marginal models in the non-Gaussian case, and (d) regression coefficient covariance matrix adjustment. Each approach has specific advantages and drawbacks. Fixed effects are often used in situations in which a key assumption of the random effects approach is thought to be implausible. In those instances, the use of marginal models or regression coefficient covariance matrix adjustment would also merit reconsideration. Where more than one approach can justifiably be applied – and there are many such cases – estimation results can be method-dependent, leading to different conclusions about the effects of specific covariates. In addition, particularly for the random effects approach, alternative methods of estimation can yield somewhat different results, as can different algorithmic implementations of the same method across software packages. The impact of small within-context sample sizes in unbalanced designs merits further exploration. Where possible, researchers should consider whether their conclusions are method- or algorithm-dependent.

Multilevel, or hierarchical, data are blocked or clustered. Children within families; individuals within organizational units; longitudinal data on individuals; repeated measures on individuals; organizational units within larger organizations; multistage sample surveys; and time series of cross-sections exemplify multilevel or hierarchical data structures. When data are clustered, each observation counts for less than it would if the data points were independent. If clustering is ignored, statistical inference can be faulty. With clustering, regression methods that assume independent observations provide inefficient and possibly biased coefficient estimates, and associated confidence intervals that are too narrow. In the context of regression modeling, any method that attempts to take account of clustering is multilevel.

Several generalized regression methods are commonly used in the analysis of hierarchical data. Hierarchical or random coefficient models are multilevel. So too are fixed effects models, marginal models, classical single-level estimation followed by regression coefficient covariance matrix adjustment, and hybrid approaches. The choice of an approach for a particular research problem must depend on the goals of the research and assessment of the design and measurement deficiencies presented by a particular body of data, and is often consequential. Multilevel modeling approaches emphasize different aspects of the data, vary in the nature and stringency of their assumptions, and support different kinds of descriptions and interpretations. Poor choice of method can lead to the application of an incorrectly specified model, and thereby to good estimates of the wrong quantities.

Random Coefficient Models

The random coefficient model offers a compelling formulation that is consistent with the social scientific goal of understanding how units at one level affect, and are affected by, units at other levels. Consider a design with two levels, and suppose that there is a single level-1 regressor (X), a single level-2 regressor (G) that is by definition *constant* over all level-1 observations within a context, and that the dependent variable (Y) is Gaussian distributed conditional on X and G. The random coefficient model can be expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij} \qquad [1]$$

with

$$\begin{aligned} \beta_{0j} &= \eta_{00} + \eta_{01}G_j + \alpha_{0j} \\ \beta_{1j} &= \eta_{10} + \eta_{11}G_j + \alpha_{1j} \end{aligned} \qquad [2]$$

where $j = 1,2,\ldots,J$ denotes contexts (level-2 units), and $i = 1,2,\ldots,n_j$ denotes individuals within contexts (level-1 units). The level-2 equations (eqn [2]) assert that the level-1 intercepts and slopes vary over contexts as linear functions of G. There is often justification for supposing that the within-context coefficients depend on contextual characteristics. It is appealing that the dependence includes random error components (α_{0j} and α_{1j}).

Assumptions about the error structure of this two-level model are that, for each context, the level-1 errors follow the Gaussian distribution, are centered on zero, have constant variance, and are independent. The error variances may vary over contexts. The level-2 errors are also Gaussian distributed, centered on zero, independent within equations, and have constant variances. The level-2 errors may be correlated across equations. It is further assumed that the level-1 errors are orthogonal to the level-2 errors, and that all error terms are orthogonal to X and G. The orthogonality assumptions are needed to secure consistent estimates of the ηs.

Many generalizations of this model are possible – coefficients can be constrained to be nonrandom; the

dependent variable may be discrete, counted, truncated, ordinal, or time to event; there may be more regressors at any level as well as more than two levels; the nesting of levels may be cross-classified; different assumptions about the functional forms of the distributions of the errors can be made; and the assumption of independence of errors within level-2 (or higher) units can be relaxed. Nevertheless, the assumptions of orthogonality between regressors and errors, and of cross-level orthogonality between error components, continue to remain important. Until that is no longer so, the credibility of an application of a random coefficient model rests on the plausibility of these assumptions, which should be thought of as inextricably linked to the systematic components of the model, and not as the fine print of a contract that is never read, much less enforced.

However compelling it may be on substantive grounds to specify the random coefficient multilevel model as eqns [1] and [2], substitution of eqn [2] into eqn [1] results in the equivalent mixed model representation,

$$Y_{ij} = \eta_{00} + \eta_{01}G_j + \eta_{10}X_{ij} + \eta_{11}X_{ij}G_j + \left(\alpha_{0j} + \alpha_{1j}X_{ij} + \varepsilon_{ij}\right) \quad [3]$$

which demonstrates that treating intercepts and slopes as dependent on G is the same as supposing that Y depends on X, G, and their cross-level (product) interaction.

Random coefficient models with Gaussian errors are commonly estimated by maximum likelihood or restricted maximum likelihood. For binary response the prevalent methods are marginal and penalized quasi-likelihood, although Bayesian methods based on Markov Chain Monte Carlo estimation are also beginning to be used. A consensus on preferred methods of estimation of multilevel binary response and other nonlinear (non-Gaussian) models has yet to emerge. Estimation of multilevel models is iterative and computationally intensive, more so for non-Gaussian cases. Especially for these cases, researchers are well advised to check numerical estimates across algorithms, methods, and software. Other problems, such as substantial design imbalance in combination with context size, also need to be evaluated for their potential impact on numerical estimates (*see* Estimation: Point and Interval; Guilt; Markov Chain Monte Carlo Methods).

Fixed Effects Models

The fixed effects model does not require the assumption of orthogonality between X and α_0 inherent to the random coefficient model. It can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij}G_j + \delta_j + u_{ij} \quad [4]$$

where the δ_j are between-context contrasts, and the u_{tj} are assumed to be classical Gaussian errors. Relative to the random coefficient model, the additive term in G has been dropped, and the intercept and slope are no longer allowed to vary randomly over contexts, although each context is allowed its own intercept and the cross-level interaction is retained. The δ_j absorb the additive component of *all possible* G-variables, not just the finite number that can be included in eqn [3]. Additive components for any level-1 regressors that do not vary within contexts are also absorbed by the δ_j.

In the fixed effects approach, 'fixing' the between-context contrasts means that the contexts are themselves assumed to be fixed, in the sense that statistical inference is conditional on the particular selection of contexts present in the data. Because the contrasts can be defined as the coefficients of a set of indicator variables that assign individuals to contexts, there is no need to assume that context membership is orthogonal to X. This is a critical difference between the fixed effect and random coefficient models. If the assumption of orthogonality between X and α_0 is implausible in the random coefficient model, and it can be when the list of G-variables is incomplete, then the fixed effects model provides an alternative that solves the problem. In the Gaussian case, the fixed effects model is a conventional regression model. When J is large – and often when it is not – there may be little interest in describing estimated values of the δ_j, in which case estimation of the other covariate coefficients can be performed after within-context centering of the variables. For a binary response variable, conditional logistic regression can be used to estimate fixed effect models (Chamberlain, 1980). In this case, the context contrasts are not estimated, although additive context differences are controlled. Software for fixed effects estimation is widely available.

If it is desired to obtain estimates of the additive component of the contextual variables, then the fixed effects approach is not the method of choice. Also, in the binary response case, conditional logistic regression does not necessarily use all of the data, and in this sense may be inefficient (*see* Linear Hypothesis; Multivariate Analysis: Discrete Variables (Loglinear Models)).

Coefficient Variance Estimates Adjusted for Clustering

The adjusted coefficient variance estimator in common use for multilevel analysis is a robust estimator. In actual applications, care must be taken to ensure that the robust estimator selected includes an extension of the algorithm for the general robust estimator that accommodates clustering. With the robust estimator, it is possible to specify a model whose systematic component is identical to that of eqn [3], and whose error component need not be independently and identically distributed, Gaussian distributed, or orthogonal to the regressors. The robust equivalent of eqn [3] can be expressed as:

$$Y_{ij} = b_0 + b_1 G_j + b_2 X_{ij} + b_3 X_{ij}G_j + e_{ij} \quad [5]$$

Least squares computation can be used to produce coefficients whose estimated variances and covariances are subsequently adjusted for context membership. The coefficients will be unbiased estimates of the corresponding terms in eqn [3], if the orthogonality assumption is satisfied. The standard errors will be adjusted for clustering and other sources of heteroscedasticity regardless of whether this is so. When eqn [3] is true, confidence intervals for its coefficients obtained by prevailing methods will tend to be smaller than confidence

intervals based on the robust clustering adjustment associated with eqn [5].

The robust approach is also an alternative to the fixed effects approach. The latter achieves orthogonality between the regressors and the error component by including context membership in the systematic component of the model, but is unable to include main effects of contextual variables even though it controls for all possible contextual variables. In the robust approach coefficients of contextual regressors are estimated, but all possible contextual variables are not controlled. Coefficient standard errors are adjusted for clustering, but the coefficient estimates are not efficient.

The robust, variance adjustment approach to multilevel analysis has been extended to models of discrete and counted outcomes, and to survival problems, and can be used where a fixed effects approach is not available. Nevertheless, where it can be justified, explicit incorporation of clustering into model specification is preferable to *post hoc* adjustments of estimators that ignore clustering (*see* Sample Surveys: Model-Based Approaches).

Marginal Models

When Y is a discrete or count response variable, it is possible to specify a multilevel model in which the within-context correlatedness of observations is incorporated as a side condition on the model, rather than as a consequence of a random intercept. For Y binary and ω_{ij}, the logit of 'success' for the ijth observation, the multilevel dependence of Y on regressors, can be written as

$$\omega_{ij} = \beta_0 + \beta_1 G_j + \beta_2 X_{ij} + \beta_3 X_{ij} G_j \qquad [6]$$

with an additional specification indicating the pattern of correlation between observations within a context (corr(Y_{ij}, $Y_{i'j}$), $i \neq i'$). Specifying that all pairs of observations within a context are equally correlated corresponds to the assumption of correlatedness induced by a random coefficient model, but virtually any pattern can be specified.

Models that characterize the marginal distribution of Y as a function of regressors are said to be marginal models. The model of eqn [6] and its side condition on correlated observations is a member of this class. The generalized linear model (McCullagh and Nelder, 1989), a unifying theoretical framework for the linear model and discrete and count response variables, has been extended to allow for the analysis of multilevel data (Liang and Zeger, 1986; Zeger and Liang, 1986). This extension, known as the generalized estimating equations approach (GEE), provides a marginal model (population average) alternative to random coefficient models when Y is discrete or a count. Equation [6] can be estimated using the GEE method.

Gaussian random effects models sustain a marginal model interpretation of regression coefficients. Random effects models for discrete and count response variables do not. Marginal models are appropriate when the analytic goal is to infer how Y depends on covariates for a population of level-1 units averaged over contexts. Random effects models are appropriate when within-context estimates are desired. For the logistic response case, the estimated coefficients of marginal models tend to be smaller in absolute value than the corresponding coefficients in random effects models (Diggle et al., 1994: 141).

Marginal models for Y discrete or a count can be estimated for a broad range of structures of potential within-context correlation; random coefficient models for discrete and count response variables currently cannot. Computation is fast (iterative generalized least squares), and estimation is relatively straightforward compared to binary response random coefficient models, although issues surrounding optimal characterization of within-context correlation and estimation of coefficient standard errors continue to attract attention.

Omitted Variables

Omitted variables in random coefficient models are of particular concern when they are endogenous, that is, correlated with Y and the included regressors, because they induce coefficient bias. The fixed effect approach provides one solution to this problem when the omitted variables are contextual. There are others: measurement of omitted variables, random assignment to contexts, and instrumental variables estimation.

Measuring the omitted variables is not an option in secondary analysis or in the midst of primary analysis – by then the data have been collected and returning to the field is generally infeasible. This should not, however, preclude future attempts to extend the range of measurement.

Random assignment to contexts presumes that an experimental design is feasible in the study of a particular phenomenon. Often it is not. Even when randomization is possible, it is not necessarily well suited to answering certain kinds of questions (*see* Experimental Design: Randomization and Social Experiments; Quasi-Experimental Designs; Experimental Design: Large-Scale Social Experimentation).

Instrumental variables estimation is often used in an effort to solve an endogeneity problem. When an instrumental variables approach is used, aspects of the clustering problem may be sacrificed to solve what is perceived as a much more important one – selection into contexts. In eqn [3] replace X by T, suppose that the underlying data structure is quasi-experimental, and let $T_{ij} = 1$ if an individual is treated (0 otherwise). If Y is a health outcome and contexts are hospitals, and if assignment to treatment is not random, then s potentially contains unmeasured variables correlated with T and also with Y. In quasi- and nonexperimental data, whether a patient receives a particular kind of treatment can be nonrandom with some causes known, others not. An instrumental variables approach to this problem attempts to find one or more covariates W correlated with T and Y, where Y and W controlling T are uncorrelated, and W and ε are uncorrelated. It is often difficult to find plausible instruments (W), although contextual variables are sometimes used. Thus, instrumental variables estimation cannot be employed with the automaticity of any of the other approaches to clustered data.

Examples

An example of a Gaussian application, and another of a binary response application, illustrate the kinds of differences in

findings that can occur using alternative multilevel approaches. The data are from a 1% clustered sample of the 1990 Census of China, and are specific to Hainan Province. Clustering is at the enumeration district level, and within clusters enumeration is 100%.

Gaussian Example

In the Gaussian example, there are 17 662 men in the data, distributed over 332 clusters, the smallest having 114 men and the largest having 3735 men. The occupational socioeconomic status (SEI) of men of ages 20–65 is a function of age and years of schooling at the individual level, type of place of residence (rural vs urban), and the square of a contextualized measure of education. Type of place of residence is a characterization of place, and constant within contexts. Substantive theory and exploratory data analysis lead to this random effects model:

$$\begin{aligned} SEI &= \beta_{0j} + \beta_{1j}\text{Age} + \beta_{2j}\text{Ed} + \varepsilon \\ \beta_{0j} &= \eta_{00} + \eta_{01}\text{Rural} + \alpha_{0j} \\ \beta_{1j} &= \eta_{10} + \eta_{11}\text{CED}^2 + \alpha_{1j} \\ \beta_{2j} &= \eta_{20} + \eta_{21}\text{CED}^2 + \alpha_{2j} \end{aligned} \quad [7]$$

Table 1 reports coefficient estimates and ratios of coefficients to standard errors ('t-ratios') for several approaches. Regressions (1)–(3) are the same ordinary least squares (OLS) regression, differentiated to keep track of the type of standard error. When clustering is ignored, the t-ratios are far larger (regressions (1) and (2)). Unstructured adjustment of standard errors makes little difference. Adjusting OLS standard errors for clustering radically reduces the t-ratios (regression (3)). The fixed effects coefficients (regression (4)) are roughly the same as the OLS coefficients, with larger t-ratios than obtained with the clustering adjustment. Only one term is 'significant' when the clustering adjustment is used, whereas three of the four estimable coefficients are highly 'significant' in the fixed effects model. The coefficients of the random effects model (regression (5)) are comparable with those for the fixed effects model (where present), but the t-ratios are smaller. For those terms present in both the fixed effects and random coefficient models, coefficients are either significant in both models, or not significant in both. In computations not displayed here, the random coefficient specification was computed twice within two different programs, once using maximum likelihood and restricted maximum likelihood. There was an agreement across programs and methods. Although the basic conclusion is the same for the fixed effects and random coefficient models and differs somewhat from that provided by the clustering adjustment, the variation in coefficient magnitudes and t-ratios across the models is cautionary. Even in the Gaussian case, it can be misleading to estimate a model using only one approach.

Binary Response Example

In the binary response example, data from the 1990 Hainan sample are again used. Infants and toddlers born in the 18 months prior to the starting date of the Census enumeration are the level-1 units of observation, and enumeration districts are level-2 observations. The available data do not record child mortality, which has been inferred from maternal information elicited in the Census questionnaire. The response variable is whether an imputed death to a child born after 1988 occurred prior to 1 July 1990. There are 43 imputed deaths out of 2572 mother–child pairs. A majority of enumeration districts have no imputed deaths. For this reason, only the within-context intercept is treated as random. For ω, the logit of a death within the first 18 months of life, the specification treats the

Table 1 Alternative multilevel estimates, occupational socioeconomic status of men

	Regressions		
Covariates	*Ordinary least squares Unadjusted/robust/cluster (1)/(2)/(3)*	*Fixed effect (4)*	*Random coefficient (5)*
Age	0.09	−0.01	0.02
	(7.78)/(7.92)/(1.32)	(−0.91)	(0.50)
Education	0.34	0.36	0.43
	(7.29)/(6.00)/(1.00)	(6.96)	(2.36)
Rural	−4.92		−0.72
	(−18.47)/(−13.50)/(−1.01)		(−0.35)
Age·CED^2	0.31	0.66	0.55
	(13.16)/(13.03)/(1.72)	(17.83)	(5.75)
Ed·CED^2	2.67	2.75	2.38
	(25.51)/(21.78)/(3.06)	(20.81)	(4.72)
Intercept	15.50	11.48	12.89
	(33.08)/(32.15)/(4.63)	(29.42)	(6.71)
R^2	0.45	0.57	

$N = 17\,662$. Quantities in parentheses are ratios of coefficients to standard errors. The coefficients of regressions (1)–(3), estimated by ordinary least squares, are identical. The standard errors of (1) are unadjusted; those of (2) are adjusted, but without taking explicit account of clustering; those of (3) are adjusted for cluster membership. The context contrasts in the fixed effects regression (4) are not shown. The level-2 error variances in the random coefficient regression (5), not shown, are significant.
Source: 1% sample, men of 20–65 years of age, Hainan Province, 1990 Census of China.

Table 2 Binary response example, imputed infant mortality

	Regression				
				Random coefficients	
Variables	*Classical cluster adj.* *(1)*	*GEE* *(2)*	*Fixed effects* *(3)*	*PQL* *(4)*	*MCMC* *(5)*
Maternal Ed	−0.25(−2.62)	−0.284(−2.93)	−0.27(−2.77)	−0.29(−3.19)	−0.32(−3.56)
Male	−0.96(−1.26)	−1.18(−1.95)	−1.25(−2.13)	−1.20(−2.07)	−1.43(−2.55)
Ed.male	0.24(1.73)	0.29(2.48)	0.30(2.63)	0.29(2.60)	0.32(2.84)
Rural	2.07(2.17)	1.99(1.48)		1.81(2.03)	1.38(1.38)
Intercept	−4.74(−4.71)	−4.74(−3.50)		−4.75(−5.07)	−4.74(−4.24)
Level-2 var.				0.85(1.78)	2.05(1.53)

$N = 2572$. Quantities in parentheses are ratios of coefficients to standard errors.
GEE, generalized estimating equations; MCMC, Markov Chain Monte Carlo; PQL, penalized quasi-likelihood.
Source: 1% sample, infants to 18 months of age paired with mothers, Hainan Province, 1990 Census of China.

probability of death as a function of maternal years of schooling, the sex of the child, and the interaction between maternal education and sex of child. The intercept is treated as a function of type of place of residence (the same for all residents of an enumeration district) and a random error:

$$\begin{aligned} \omega_{ij} &= \beta_{0j} + \beta_1 \mathrm{Ed} + \beta_2 \mathrm{Male} + \beta_3 \mathrm{Ed} \cdot \mathrm{Male} \\ \beta_{0j} &= \eta_{00} + \eta_{01} \mathrm{Rural} + \alpha_{0j} \end{aligned} \quad [8]$$

Table 2 presents a selection of regressions based on alternative multilevel approaches and algorithms. Regression (1) is the result of classical maximum likelihood estimation followed by an adjustment for clustering. Using this approach, neither the sex of the child nor the sex-education interaction is 'significant.' Using the GEE approach, the coefficient for sex is at the cusp of conventional significance and the sex-education interaction is significant (regression (2)). The fixed effects results, using conditional logistic regression, indicate significance for all included terms (regression (3)). Penalized quasi-likelihood estimates are provided by regression (4). These are consistent with the fixed effects results for those terms present across the fixed and random effects specifications. Regression (5) presents Bayesian results based on MCMC. The type of place of residence coefficient is smallest and not significant in the Bayesian regression. Application of a clustering adjustment to the GEE results (not shown) also substantially affects conclusions about statistical significance.

It is not unusual to encounter sparse data. This binary response example is perhaps even more cautionary than the Gaussian example. Researchers are not yet routinely able to determine whether convergence is to a local or global maximum, whether sufficient iterations have been calculated using MCMC, whether all of the assumptions required for a particular method are valid in a given instance, or whether software is performing as intended.

Conclusions

Analysis of multilevel data structures using methods that ignore the clustering of observations in the data is unsatisfactory because it can lead to confidence intervals that are too narrow, and to coefficient bias. Random coefficient models – and their specializations and generalizations – are attractive tools for the analysis of multilevel data. When there is selection into contexts, the random coefficient approach is subject to coefficient bias, as it is more generally when there are omitted regressors at any level. The fixed effects approach provides one kind of solution to the omitted variables problem. Instrumental variable estimation offers another. A posterior adjustment to accommodate clustering, applied to the estimated variances of coefficients, is robust but inefficient when the random effects model is correct. Although this adjustment does not require orthogonality between errors and covariates, robust standard errors for biased coefficients are far from universally desirable. For discrete and count responses, marginal models estimated using GEE provide an alternative to random effects models. Marginal models are attractive when there is little interest in within-context estimates, or when the pattern of within-context correlation is not simple. The choice of a multilevel modeling approach is ideally governed by research goals, the study design, and the nature of the data. There is also value in cross-checking results across methods and software.

Further Reading

For an introduction to random coefficient models (terminology often used interchangeably with random effects models and hierarchical models), see Snijders and Bosker (1999). The level of exposition in Bryk and Raudenbush (1992) ranges from introductory to advanced, as is also true for Goldstein (1995). The exposition in Longford (1993) is technically demanding. Littell et al. (1996) survey mixed models in general, and include material on hierarchical models. Gilks et al. (1998) and Gelman et al. (2013) are useful for Bayesian random coefficient models. Draper (2002) provides a detailed introduction to random coefficient models from the Bayesian perspective.

Fixed effects modeling is treated in econometrics texts (e.g., Greene, 2000), and is often discussed as one approach to the analysis of panel data (Baltagi, 1995). For the binary response case, see Pendergast et al. (1996), who provide a review with broad scope. Robust adjustments of standard errors for

clustering (e.g., Rogers, 1993) are an extension of the methodology of Huber (1967) and White (1980, 1982). The GEE approach is well exposited in Diggle et al. (1994), who provide detailed comparisons with random coefficient and other models. In this same vein, Allison's (1999) introduction should not be missed. Instrumental variables estimation is a standard subject in econometrics texts (e.g., Greene, 2000).

For an illuminating use of random coefficient models in the study of neighborhoods, see Sampson et al. (1997). Geronimus and Korenman (1992) use a fixed effects approach in the study of teen pregnancy. Hamilton and Hamilton (1997) use a fixed effects approach in the analysis of surgery outcomes. For the same kind of problem, but using a different data structure, McClellan et al. (1994) employ an instrumental variables approach. Applications of marginal models are found largely in the epidemiological and biomedical literature. Multilevel analysis has long been mainstream in economics, and will soon become so in the rest of the social sciences.

Software

Dedicated random coefficient packages include HLM5 (Raudenbush et al., 2000) and MLwiN (Goldstein et al., 1998). HLM5 additionally offers GEE estimation; MLwiN includes Bayesian MCMC estimation. Most major statistical packages include facilities for all of the approaches discussed here, as integral components or contributed macros. Software for multilevel analysis is in continuous development. The multilevel models mail list is an essential portal for current information (multilevel@jiscmail.ac.uk), as is the Web site hosted by the Multilevel Models Project (http://www.ioe.ac.uk/multilevel/).

Bibliography

Allison, P.D., 1999. Logistic Regression Using the SAS System. SAS Institute, Cary, NC.

Andrew Royle, J., Dorazio, Robert M., 2008. Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities. Academic Press, Amsterdam.

Agresti, Alan, 2007. An Introduction to Categorical Data Analysis, second ed. John Wiley and Sons.

Baltagi, B.H., 1995. Economic Analysis of Panel Data. Wiley, New York.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models. Sage, Newbury Park, CA.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Review of Economic Studies 47, 225–238.

Diggle, P.J., Liang, K.-Y., Zeger, S.L., 1994. Analysis of Longitudinal Data. Oxford University Press, Oxford, UK.

Draper, D., 2002. Bayesian Hierarchical Modeling. Springer-Verlag, New York.

Gelman, Andrew, Hill, Jennifer, 2007. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Gelman, Andrew, Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, Aki, Rubin, Donald B., 2013. Bayesian Data Analysis, third ed. CRC Press, Boca Raton, FL.

Geronimus, A.T., Korenman, S., 1992. The socioeconomic consequences of teen childbearing reconsidered. Quarterly Journal of Economics 107, 1187–1214.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1998. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Goldstein, H., 1995. Multilevel Statistical Models, second ed. Halstead, New York.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., Healy, M., 1998. A User's Guide to MLwiN. Multilevel Models Project, Institute of Education, University of London, London.

Greene, W.H., 2000. Econometric Analysis, fourth ed. Prentice-Hall, Upper Saddle River, NJ.

Hamilton, B.H., Hamilton, V.H., 1997. Estimating surgical volume – outcome relationships applying survival models: accounting for frailty and hospital fixed effects. Health Economics 6, 383–395.

Huber, P.J., 1967. The behavior of maximum likelihood estimators under nonstandard conditions. In: LeCam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, CA, pp. 221–233.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., 1996. SAS System for Mixed Models. SAS Institute, Cary, NC.

Longford, N.T., 1993. Random Coefficient Models. Oxford University Press, New York.

McClellan, M., McNeil, B.J., Newhouse, J.P., 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Journal of the American Medical Association 272, 859–866.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, New York.

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M., Fisher, M.R., 1996. A survey of methods for analyzing clustered binary response data. International Statistical Review 64, 89–118.

Raudenbush, S., Bryk, A., Cheong, Y.F., Congdon, R., 2000. HLM 5: Hierarchical Linear and Nonlinear Modeling. Scientific Software International, Lincolnwood, IL.

Rogers, W.H., 1993. Regression standard errors in clustered samples. Stata Technical Bulletin 13, 19–23.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Snijders, T.A.B., Bosker, R., 1999. Multilevel Analysis. Sage Publications, London.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.

Zeger, S.L., Liang, K.-Y., 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121–130.

Statistical Analysis, Special Problems of: Outliers

Toby Lewis, University of East Anglia, Norwich, UK

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15002–15006, © 2001, Elsevier Ltd.

Abstract

Outliers, or outlying observations, are values in data which appear aberrant or unrepresentative. They occur commonly and have to be dealt with. Unless an outlier is explainable, e.g., as a mis-recording, action must be based on the discrepancy between it and the model for the data. If a test shows the outlier to be inconsistent with the model, it is characterized as discordant. A discordant outlier may be identified as of intrinsic interest, rejected as a nuisance item, or incorporated in the main data via a revised model. Multiple outliers may be tested either one at a time, testing continuing so long as successive values are found discordant, or en bloc. The phenomena of masking and swamping are described. A main alternative to detecting and testing outliers is accommodation. Inference procedures are used which down-weight extreme values automatically, without regard to whether or not they are discordant. In complex situations such as regression on several variables, designed experiments, multivariate data, and multilevel data, a data point may be inconsistent with the pattern of the main data, and thus outlying, but not perceivable by inspection. Such an outlier is detectable by a statistical examination of residuals.

Anyone analyzing data may expect sometimes to encounter one or more observed values which appear to be inconsistent or out of line with the rest of the data. Such a value is called an outlying observation or outlier. Judged subjectively, it looks discrepant or surprising. There are two main ways of dealing with outliers, discordancy testing (Sections Outliers in Univariate Samples, Multiple Outliers: Block and Consecutive Tests) and accommodation (Section Accommodation of Outliers). In simpler types of data set, the presence of an outlier is evident by inspection. In more complex data situations such as designed experiments, an outlying observation may disrupt the expected pattern of results, though its presence is not obvious by inspection and a statistical procedure is needed to detect it (Section Outliers in More Complex Data Situations).

Two Simple Examples

A Disputed Paternity Case

In 1921 Mrs Kathleen Gaskill was sued for divorce in the British courts by her husband, a serving soldier, because she had given birth to a child 331 days after he had left the country to serve abroad. He claimed that the child could not be his, since 331 days was not credible as a duration of pregnancy. She denied adultery and claimed, with medical support, that the length of pregnancy, though unusual, was genuine. Viewed as a duration of pregnancy, 331 days is an outlier. In the legal case (which the court decided in favor of the wife), the only evidence of the alleged adultery was the abnormal length of this duration. This exemplifies the situation where the judgment of an outlier depends purely on its position in relation to a main data set. Table 1 gives the durations of pregnancy for 1669 women who gave birth in a maternity hospital in the north of England in 1954 (summarized from Newell, 1964). Clearly 331 days, exceeded certainly by 1 case and possibly by 2 out of 1669 in this data set, cannot be regarded as an incredible duration.

Predicting College Grades

Figure 1 shows a scatter plot of points (x, y) for a small sample of 22 college students; x is the student's score in an aptitude test, used as a predictor of college performance, and y is the student's freshman year grade-point average. There are two

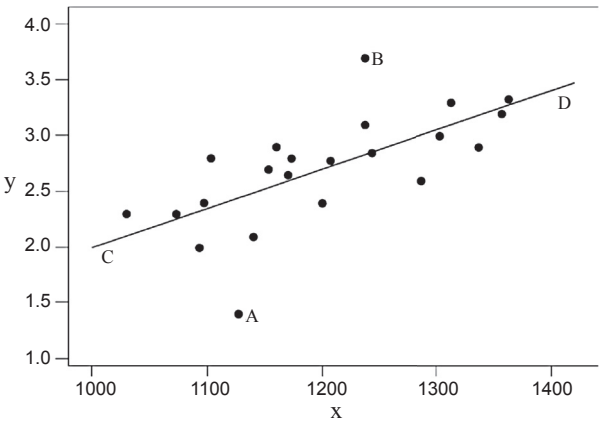

Figure 1 Scatter plot of grade-point average y vs predictive test score x for 22 college students.

Table 1 Durations of pregnancy for 1669 women

Duration (weeks)	*11–20*	*21–25*	*26–30*	*31–35*	*36–38*	*39–41*	*42*	*43*	*44*	*45*	*46*	*47*	*56*
Numbers	4	10	23	113	333	962	144	53	12	10	3	1	1

outlying points A(x_A, y_A) and B(x_B, y_B); the desired predictive relationship of y from x could be modeled satisfactorily by a linear regression omitting these two points. Should they be included in the relationship or not? This is discussed in Section Outliers in More Complex Data Situations. *See* Linear Hypothesis: Regression (Basics).

Outliers in Univariate Samples

The nature of outliers and the range of issues in dealing with them are first discussed in the simple situation of a univariate sample of n observations $x_1, x_2, \ldots, x_n$ of a variable X. Denote the observations arranged in ascending order by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. If, say, the upper extreme observation $x_{(n)}$ appears (to an analyst or investigator examining the data) to be, not only higher than the rest of the sample but surprisingly high, it is declared an outlier. This is a subjective judgment. Is the outlier credible as a value of X? This depends on the form the investigator is assuming for the distribution of X, e.g., whether it is, say, normal or on the other hand a skew distribution such as exponential in which a high value $x_{(n)}$ is less surprising. Call this assumed model for the distribution of X the basic model and denote it by F; $x_{(n)}$ is an outlier relative to model F (e.g., normal), but may not be an outlier relative to a different model (e.g., exponential).

Table 2, with dotplot Figure 2, gives the scores by 30 children in a reading accuracy test. There is an upper outlier $x_{(30)} = 62$ and possibly a second one $x_{(29)} = 53$. The variable X is here the reading accuracy test score obtained by a child in a population of children of the relevant age group and educational characteristics. A reasonable working assumption is that its distribution is normal; this is supported by a plot of ordered values against normal order scores, which is roughly linear.

Causes of Outliers

No statistical issue arises if there is a tangible explanation for the presence of an outlier, e.g., if in the above example the score was misrecorded, or component subscores were wrongly totaled, or if the child tested was known to be unrepresentative of the population under study. The outlying value would simply be removed from the data set, or corrected where appropriate, e.g., in the case of wrong totaling.

Otherwise, a statistical judgment is required for interpreting the outlier and deciding how to deal with it, based on the apparent discrepancy between it and the pattern the investigator has in mind for the data – the basic model F. This judgment requires a test of the null hypothesis or working hypothesis H that all n observations belong to F. On this hypothesis, the upper outlier $x_{(n)}$ is a value from the distribution of a variable $X_{(n)}$, the highest observation in a random sample of size n from F. If H is rejected by a statistical test at

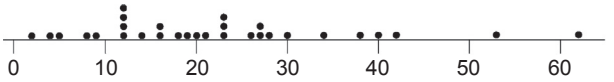

Figure 2 Dotplot of the 30 test scores in Table 2.

some level (e.g., 5 percent), $x_{(n)}$ is judged to be not credible (at the level of the test) as the highest value in a sample of size n from F. It is then said to be a discordant outlier (relative to the basic model F); the test of H is a discordancy test. (*See* Hypothesis Testing in Statistics; Significance, Tests of. See also Section Other Meanings of the Word 'Outlier'.).

In a Bayesian approach, there is no direct analogue of a test of discordancy. Its place is taken by procedures designed to obtain a posterior probability assessment of the 'degree of contamination' of the data. This means the degree to which the data contain observations ('contaminants') generated not by F but by some other model G; a basic model is used which contains additional parameters relating to a possible G (*see* Bayesian Statistics).

Treatment of Outliers: Retention, Identification, Rejection, Incorporation

If the result of the test is to accept H and find $x_{(n)}$ non-discordant, all the data can be taken to come from F, including the outlier. Although it has been thought to look somewhat unusual, it is consistent with having arisen purely by chance under the basic model and can be retained in the data set.

When on the other hand the outlier is adjudged discordant, there are three possibilities.

If the focus of interest is the outlier itself (as in the disputed paternity case), it is identified as a positive item of information, inviting interpretation or further examination; it might reveal, for instance, a novel teaching method causing an exceptional test score.

The other two possibilities, rejection and incorporation, arise when the investigator's concern is with the main data set, and the aim of the research is to study characteristics of the underlying population, e.g., to estimate the mean of a distribution of examination scores. The outlier is only of concern insofar as it affects these inferences about the underlying population. It is not of interest in itself; for example, the students giving rise to points A and B in Figure 1 are effectively unknown!

Either the model F is firmly adopted without challenge, and estimates or tests of the population parameters are to be based on F; or there is the option of changing the model.

If the model is not under question and inferences are to be based on it, the outlier, being adjudged inconsistent with F, is a nuisance item and should be rejected. Historically, outlying values have been studied since the mid eighteenth century,

Table 2 Reading accuracy test scores (in ascending order) for 30 children

2	4	5	8	9	12	12	12	12	14	16	16	18	19	20	21	23	23
23	26	27	27	28	30	34	38	40	42	53	62						

often in the context of astronomical observations. Until the 1870s rejection was often the only procedure envisaged for dealing with outliers, as typified by such publications as Peirce (1852) 'Criterion for the rejection of doubtful observations' and Stone (1868) 'On the rejection of discordant observations.'

The fact that an outlier is omitted for purposes of analyzing a data set does not necessarily mean that it is erroneous or undesirable. Consider, for example, Figure 3. The ordinates are annual figures for the population of the UK, in millions, published in each year's *Annual Abstract of Statistics*. Those for 1985 through 1990 are estimates based on the 1981 census, while the outlier at 1991, the true value from the 1991 census, is the only genuine data point!

The alternative possibility, when the outlier is adjudged discordant relative to F, is that all the observations, including the outlier, belong to the same distribution, but this distribution is not satisfactorily represented by the initial basic model F. The analysis then needs to start afresh with a more appropriate choice of model F* (e.g., a skew distribution instead of a symmetrical F), in relation to which $x_{(n)}$ is not discordant. It was an outlier relative to F, but is not an outlier relative to F*, which satisfactorily models the entire data set. It has been incorporated in the data set in a non-discordant fashion, and the outlier problem disappears.

Other Meanings of the Word 'Outlier'

In the entry, an observation which appears inconsistent with the main data set is called an outlier; this is a subjective judgment. If a statistical test provides evidence that it cannot reasonably be regarded as consistent with the main data set, it is a discordant outlier. It is important to bear in mind, however, that some writers use the word 'outlier' for a discordant outlier, and refer to an observation which looks anomalous but may or may not be discordant as a 'suspicious' or 'suspect' value.

Note also the different meaning given to the word 'outlier' in graphical displays in exploratory data analysis (EDA), such as boxplots. With Q_1 and Q_3 denoting (essentially) the lower and upper quartiles in the sample, observations greater than $Q_3 + k(Q_3 - Q_1)$ or less than $Q_1 - k(Q_3 - Q_1)$ are flagged as outliers. These values are sometimes outliers and sometimes not. With the typical value of 1.5 for k, a normal sample of size 100 has more than 50 percent chance of containing one or more of these 'outliers'! (*See* Exploratory Data Analysis: Univariate Methods).

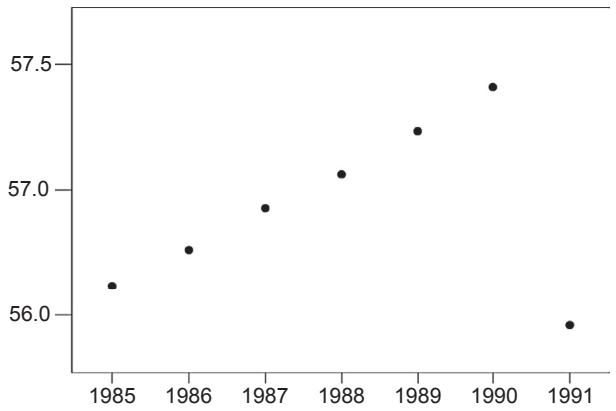

Figure 3 Some annual figures for the UK population.

Multiple Outliers: Block and Consecutive Tests

Any particular discordancy test (taking univariate samples as context) is specified by three items: the particular outlier configuration (e.g., a lower outlier; an upper and lower outlier pair), the distribution of the data on the working hypothesis (e.g., normal with unknown mean and variance), and the test statistic. Taking Table 2 for illustration, there is a discordancy test for $x_{(30)} = 62$, in a normal sample, based on its difference from the sample mean $\bar{x} = 22.53$ in relation to the sample standard deviation $s = 14.13$, with test statistic $T = (x_{(n)} - \bar{x})/s = (62 - 22.53)/14.13 = 2.79$.

For $n = 30$ the tabulated 5 percent point of T on hypothesis H is 2.74; 62 is declared discordant (just) at the 5 percent level. The observation $x_{(29)} = 53$, which is also high, can now be tested for discordancy in relation to the other 28 values; this is a consecutive test procedure for 2 outliers. The value of T is (53–21.17) / 12.22 = 2.61, P≃8 percent, the evidence of discordancy is weak.

Alternatively, 62 and 53 could be considered together as an outlying pair in relation to the other 28 observations. This can be tested, in a block test procedure for 2 outliers, by the statistic $(x_{(n-1)} + x_{(n)} - 2\bar{x})/s$, which here $= (53 + 62 - 2 \times 22.53)$ / 14.13 = 4.95, to be compared with the tabulated 1 percent point 4.92. So, on this block test, there is quite strong evidence that (62, 53) are a discordant pair.

Masking and Swamping

If $x_{(29)}$ were nearer to $x_{(28)} = 42$, the T-value for $x_{(30)}$ would be higher than 2.79, and the evidence for the discordancy of $x_{(30)}$ would be stronger. This is called the masking effect. A discordancy test of the most extreme observation is rendered insensitive by the nearness of the next most extreme observation; the second observation has masked the first.

Conversely, a block test could be insensitive if the second observation $x_{(n-1)}$ is close to $x_{(n-2)}$ and not outlying. The pair $x_{(n)}$, $x_{(n-1)}$ might not reach discordancy level when tested jointly, even though $x_{(n)}$ on its own is discordant in relation to the other $n - 1$ observations. This is called the swamping effect of $x_{(n-1)}$. If $x_{(n-1)}$ is unduly high, there is the risk of masking; if it is unduly low, there is the risk of swamping.

Accommodation of Outliers

When the aim is to make inferences about a population of values of a variable X by analyzing a sample of observations of X, there is an alternative to detecting outliers and testing them for discordancy. This is to use procedures (testing or estimation) which are relatively unaffected by possible outliers, by giving extreme values less weight than the rest of the observations, whether they are discordant or not. Such a procedure is called robust against the presence of outliers (*see* Robustness).

There is a great range of procedures available in the literature. The approach is well illustrated by the following simple methods for robust estimation of the mean and variance of a univariate sample $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

Trimming

The mean is calculated after one or more extreme values have been omitted from the sample. Omission of the r (lowest) and the s (highest) values gives the (r, s)-fold trimmed mean

$$\left(x_{(r+1)} + x_{(r+2)} + \cdots + x_{(n-s)}\right)/(n - r - s).$$

Winsorizing

Instead of omitting the r (lowest) values $x_{(1)},\ldots, x_{(r)}$, they are each replaced by the nearest value which is retained unchanged, i.e., $x_{(r+1)}$; similarly for the s (highest) values. The mean of the resulting modified sample, still of size n, is the (r, s)-fold Winsorized mean

$$\left((r+1)x_{(r+1)} + x_{(r+2)} + \cdots + x_{(n-s-1)} + (s+1)x_{(n-s)}\right)/n.$$

The choice of r and s depends on various factors such as the form of the distribution of X (e.g., symmetrical or asymmetrical) and the level of robustness required. The greatest possible amount of trimming or Winsorizing, with $r = s$, gives the sample median.

Robust estimates of variance can be based in various ways on the deviations $x - m$, where m is a trimmed or Winsorized mean, instead of the usual $x - \bar{x}$.

For an extensive account of accommodation issues and procedures, see Barnett and Lewis (1994).

Outliers in More Complex Data Situations

In a regression of a variable y on a single regressor variable x, an outlying point is apparent by inspection, as instanced by **Figure 1**. For a regression on 2 or more regressor variables x_1, $x_2,\ldots$, this is no longer the case. The data may nonetheless contain an outlying point (or more than one), i.e., a point (x, y) which departs to a surprising degree from the overall pattern of the data. Its observed response value y deviates 'surprisingly' from the fitted value from the regression equation; in other words, it has an outlying residual. Detection of an outlying point requires essentially a statistical examination of residuals.

To illustrate this, consider the regression data in **Figure 1**, with two outlying points A and B. Denote the fitted regression line CD by $y = a + bx (a = -1.528, b = 0.003528)$, with residual mean square $s^2 = 0.148$. The residual for the outlying point A is $y_A - (a + bx_A) = 1.4 - 2.447 = -1.047$, and its estimated standard deviation is

$$s\sqrt{\left\{1 - \frac{1}{n} - \frac{(x_A - \bar{x})^2}{\sum (x - \bar{x})^2}\right\}} = 0.371.$$

A discordancy test statistic for the point A is its studentized residual (the residual divided by its estimated standard deviation), viz. $-1.047 / 0.371 = -2.82$; this is just significant at the 5 percent level. On a similar test the outlying point B is not discordant at all. The most outlying point in the regression (in this case, A) is the one with the largest studentized residual, ignoring sign. *See* Linear Hypothesis: Regression (Basics).

An equivalent criterion to the studentized residual for testing discordancy is the proportional reduction in the residual sum of squares when the outlier is omitted, or, equivalently again, the increase in the maximized log-likelihood. The test then reflects directly how much the fit of the regression model to the data is improved on omission of the outlier. This approach – the deletion method – obviously extends to the detection and testing of multiple outliers, with the choice between block and consecutive deletion (*see* Likelihood in Statistics). It is also the appropriate approach to outliers in multivariate data; *see* Multivariate Analysis: Discrete Variables (Overview).

Similar considerations apply to the detection and testing of outliers in designed experiments and outliers in contingency tables, with use of residual-based or deletion procedures. *See* Experimental Design: Overview. Accommodation procedures are also available in the literature. Outliers also occur in time series (*see* Time Series: General). For details of methods for dealing with outliers in the various data situations mentioned above, see Barnett and Lewis, 1994.

Some work has been done on outliers in sample surveys, but the methodology for outliers in this vital field urgently needs developing. *See* Sample Surveys: Model-based Approaches.

In multilevel, or hierarchical, data the concept of an outlier becomes more complex, and much work needs to be done. *See* Hierarchical Models: Random and Fixed Effects; Statistical Analysis: Multilevel Methods; Lewis and Langford (2001). See also Langford and Lewis (1998), from which the following passage is quoted:

In a simple educational example, we may have data on examination results in a two-level structure with students nested within schools, and either students or schools may be considered as being outliers at their respective levels in the model. Suppose ... that ... a particular school is found under test to be a discordant outlier; we ... need to ascertain whether it is discordant owing to a systematic difference affecting all the students measured within that school or because one or two students are responsible for the discrepancy. At student level, an individual may be outlying with respect to the overall relationships found across all schools or be unusual only in the context of his or her particular school. ... [Again,] masking see Section Masking and Swamping can apply across the levels of a model. The outlying nature of a school may be masked because of another similarly outlying school or by the presence of a small number of students within the school whose influence brings the overall relationships within that school closer to the average for all schools.

Conclusion

'The concern over outliers is old and undoubtedly dates back to the first attempt to base conclusions on a set of statistical data .... The literature on outliers is vast ... ' (Beckman and Cook, 1983). As this article has shown, outliers occur commonly in data and the problem of how to deal with them

in any particular case is an unavoidable one. A massive armory of principle, model, and method has built up over the years. For a review of the relevant literature, see the extensive coverage of outlier methodology by Barnett and Lewis (1994), whose book, which includes some 970 references, still remains the only comprehensive treatment of the subject. A valuable survey of many outlier problems is also given in Beckman and Cook (1983).

Bibliography

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, third edn. Wiley, Chichester, UK.

Beckman, R.J., Cook, R.D., 1983. Outliers (with discussion). Technometrics 25, 119–163.

Chatterjee, S., Hadi, A.S., 1988. Sensitivity Analysis in Linear Regression. Wiley, New York.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, London.

von Eye, A., Schuster, C., 1998. Regression Analysis for Social Sciences. Academic Press, San Diego, CA.

Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall, London.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.

Langford, I.H., Lewis, T., 1998. Outliers in multilevel data (with discussion). Journal of the Royal Statistical Society A 161, 121–160.

Lewis, T., Langford, I.H., 2001. Outliers, robustness and the detection of discrepant data. In: Leyland, A.H., Goldstein, H. (Eds.), Multilevel Modelling of Health Statistics. Wiley, Chichester, UK.

McKee, J.M., 1994. Multiple Regression and Causal Analysis. F E Peacock, Itasca, IL.

Newell, D.J., 1964. Statistical aspects of the demand for maternity beds. Journal of the Royal Statistical Society A 127, 1–33.

Peirce, B., 1852. Criterion for the rejection of doubtful observations. Astronomical Journal 2, 161–163.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points (with discussion). Journal of the American Statistical Association 85, 633–651.

Stone, E.J., 1868. On the rejection of discordant observations. Monthly Notices of the Royal Astronomical Society 28, 165–168.

Statistical Clustering

John A Hartigan, Yale University, New Haven, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 22, pp. 15014–15019, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Statistical methods assist in classification in four ways: in devising probability models for data and classes so that probable classifications for a given set of data can be identified; in developing tests of validity of particular classes produced by a classification scheme; in comparing different classification schemes for effectiveness; and in expanding the search for optimal classifications by probabilistic search strategies. Standard hierarchical and partitioning algorithms are evaluated from a statistical point of view, and some improvements of these algorithms using density estimation methods and mixture models are proposed. Methods are given for identifying multimodality in high dimensional data.

Every formal inquiry begins with a definition of the objects of the inquiry and of the terms to be used in the inquiry, a *classification* of objects and operations. Since about 1950, a variety of simple explicit classification methods, called clustering methods, have been proposed for different standard types of data. Most of these methods have no formal probabilistic underpinnings. And indeed, classification is not a subset of statistics; data collection, probability models, and data analysis all require informal subjective prior classifications. For example, in data analysis, some kind of binning or selection operation is performed prior to a formal statistical analysis.

However, statistical methods can assist in classification in four ways.

1. In devising probability models for data and classes so that probable classifications for a given set of data can be identified;
2. In developing tests of validity of particular classes produced by a classification scheme;
3. In comparing different classification schemes for effectiveness; and
4. In expanding the search for optimal classifications by probabilistic search strategies.

These statistical methods are referred to as *cluster analysis* or *statistical classification*. The second term is preferred because a cluster, from its origin, tends to connote a crowd of points close together in space, but a class is a set of objects assembled together for any kind of reason, and so the more general term *classification* is to be preferred. Part of the reason for the narrower term *cluster analysis* is that the term *classification* was once used in statistics to refer to *discriminant analysis* (*see* Multivariate Analysis: Classification and Discrimination). In discriminant analysis, a classification rule is developed for placing a multivariate observation into one of a number of established classes, using the joint distribution of the observations and classes. These problems are regression problems: the variable to be predicted is the class of the observations. For example, logistic regression is a method for discriminant analysis with two classes. More recently, classification in statistics has come to refer to the more general problem of constructing families of classes, a usage which has always been accepted outside of statistics.

Algorithms

The first book on modern clustering techniques is by Sokal and Sneath (1963). The most comprehensive recent review of a huge diverse literature is Arabie et al. (1996) which includes in particular an excellent survey of algorithms by Arabie and Hubert and of probabilistic methods by Bock.

A classification consists of a set of classes; formal clustering techniques have been mainly devoted to either *partitions* or *hierarchical clustering* (see, however, the nonhierarchic structures in Jardine and Sibson, 1968; and the block clusters in Hartigan, 1975).

Partitions

A partition is a family of *classes*, sets of objects with the property that each object lies in exactly one class of the partition. *Ideal-object* partitions are constructed as follows. Begin with a set of given objects, a class of ideal objects, and a numerical dissimilarity between any given object and any ideal object. A fitted ideal object is associated with each class having minimum sum of dissimilarities to all given objects in the class.

The algorithm aims to determine k ideal objects, and a partition of the objects into k classes, so that the dissimilarity between an object and the fitted ideal object of its class, summed over all objects, is minimized. The usefulness of such a clustering is that the original set of objects is reduced with small loss of information to the smaller set of fitted ideal objects. Except in special circumstances, it is impracticable to search over all possible partitions to minimize the sum of dissimilarities; rather some kind of local optimization is used instead.

For example, begin with some starting partition, perhaps chosen randomly. First fit the ideal objects for a given partition. Second reallocate each object to the class of the fitted ideal object with which it has smallest dissimilarity. Repeat these two steps until no further reallocation takes place. The *k-means*

algorithm, MacQueen (1967), is the best known algorithm of this type; each object is represented by a point in p-dimensional space, the dissimilarity between any two points is squared Euclidean distance, and the ideal object for each class is the centroid of the class. Asymptotic behavior of the k-means algorithm is described in Pollard (1981).

Hierarchical Clustering

Hierarchical clustering constructs *trees* of clusters of objects, in which any two clusters are disjoint, or one includes the other. The cluster of all objects is the root of the tree.

Agglomerative algorithms, Lance and Williams (1967), require a definition of dissimilarity between clusters; the most common ones are *maximum or complete linkage*, in which the dissimilarity between two clusters is the maximum of all pairs of dissimilarities between pairs of points in the different clusters; *minimum or single linkage or nearest neighbor*, Florek et al. (1951), in which the dissimilarity between two clusters is the minimum over all those pairs of dissimilarities; and *average linkage* in which the dissimilarity between the two clusters is the average, or suitably weighted average, over all those pairs of dissimilarities.

Agglomerative algorithms begin with an initial set of singleton clusters consisting of all the objects; proceed by *agglomerating* the pair of clusters of minimum dissimilarity to obtain a new cluster, removing the two clusters combined from further consideration; and repeat this agglomeration step until a single cluster containing all the observations is obtained. The set of clusters obtained along the way forms a hierarchical clustering.

Object Trees

A different kind of tree structure for classification is a graph on the objects in which pairs of objects are linked in such a way that there is a unique sequence of links connecting any pair of objects. The *minimum spanning tree*, Kruskal (1956), is such a graph, constructed so that the total length of links in the tree is minimized. There may be more than one such optimal graph. The minimum spanning tree is related to single linkage clusters: begin by removing the largest link, and note the two sets of objects that are connected by the remaining links, continue removing links in order of size, noting at each stage the two new sets of objects connected by the remaining links. The family of connected sets obtained throughout the removal process constitutes the single linkage clusters.

In a Wagner tree, the given objects are end nodes of the tree, and some set of *ideal* objects are the possible internal nodes of the tree. A Wagner tree for the given objects is a tree of minimum total length in which the given objects are end nodes and the internal objects are arbitrary. A Wagner tree is a minimum spanning tree on the set of all objects and the optimally placed internal nodes. This tree has application in evolutionary biology as the *most parsimonious* tree reaching the given objects (Fitch and Margoliash, 1967).

A third kind of tree structure arises from the CART algorithm of Breiman et al. (1984). This algorithm applies to regression problems in which a given target variable is predicted from a recursively constructed partition of the observation space; the prediction is constant within each member of the partition. If the predictors are numerical variables, the first choice in the recursion looks over all variables V and all cut-points a for each variable, generating the binary variables identifying observations where $V < a$. The binary predictor most correlated with the target variable divides the data set into two sets, and the same choice of division continues within each of the two sets, generating a *regression tree*. It is a difficult problem to know when to stop dividing. These regression trees constitute a classification of the data adapted to a particular purpose: the accurate prediction of a particular variable.

Usually, only the final partition is of interest in evaluating the regressive prediction, although the regression tree obtained along the way is used to decide which variables are most important in prediction, for example, the variables appearing in the higher levels of the tree are usually the most important. However, if there are several highly correlated variables, only one might appear in the regression tree, although each would be equally useful in prediction.

Probabilistic Classification

The aim is to select an appropriate classification C from some family Γ based on given data D

Mixture Models

It is assumed that there are a number of objects, denoted $x_1, \ldots, x_n$, which are to be classified. A probability model specifies the probability of an observation x when x is known to come from a class with parameter value θ, say $p(x|\theta)$. The parameter value θ corresponds to an ideal point in the partition algorithm (section 'Hierarchical Clustering'). The *mixture* model assumes that there is a fixed number of classes k, that an object x_i comes from class C_j with mixing probability p_j, so that the probability of observation x is $p(x) = \Sigma_j p_j p(x|\theta_j)$, and collectively the observations have probability $\Pi_i p(x_i)$. The parameters p_j and θ_j usually are estimated by maximum likelihood in a non-Bayesian approach. No single partition is produced by the algorithm, but rather the conditional probability of each partition at the maximum likelihood values of the parameters. How do these statistical partitionings compare with the dissimilarity based partitions of the section Hierarchical Clustering? Suppose the partition itself is a parameter to be estimated in the statistical model; estimate the optimal partition $C = \{C_1, \ldots, C_k\}$, and the parameter p_j, θ_j to maximize the probability of the data given the partition $\Pi_j \Pi_{x_i \in C_j} p_j p(x_i|\theta_j)$. Thus, the optimal partition in the statistical model is the optimal partition in a dissimilarity model with dissimilarities between the object x_i and the ideal object θ_j

$$d(x_i, \theta_j) = \log(p_j) + \log p(x_i|\theta_j) \qquad [1]$$

It is, thus, possible to translate back and forth between statistical models and dissimilarity models by taking dissimilarities to be log probabilities.

However, it is wrong to estimate the optimal partition by maximizing the probability of the data given the partition and

parameters over all choices of partitions and parameters; as the amount of data increases, the limiting optimal partition does not correspond to the *correct* partition obtained when the parameters p_j, θ_j are known. The correct partition assigns x_i to the class that maximizes $p_j p(x_i|\theta_j)$. In contrast, the estimate of p_j, θ_j that maximizes the data probability averaged over all possible partitions, is consistent for the true values under mild regularity conditions, and the partition determined from these maximum likelihood estimates is consistent for the correct partition.

In a Bayesian approach, the probability model is a joint probability distribution over all possible data D and classifications C. The answer to the classification problem is the conditional probability distribution over all possible classifications C given the present data D. Add to the previous assumptions a prior distribution on the unknown parameters p_j, θ_j (Bock, 1972). Again, consistent estimates of the correct classification are not obtained by finding the classification C and parameters p, θ of maximum posterior probability. Use the marginal posterior distribution of p, θ averaging over all classifications, to give consistent maximum posterior density estimates $\widehat{p}$, $\widehat{\theta}$ say. Then find the classification C of optimal posterior probability C given the data and given $p = \widehat{p}$, $\theta = \widehat{\theta}$. This classification will converge to the optimal classification for the mixture distribution with known p and θ.

If in addition a prior distribution is specified for the number of clusters, a complete final posterior distribution is available for all possible partitions into any number of components.

Hierarchical Models

Many of the hierarchical methods assume a dissimilarity function specifying a numerical dissimilarity $d(i, j)$ for each pair of objects i, j. If d satisfies the *ultrametric inequality* for every triple i, j, k

$$d(i,j) \leq \max[d(i,k), d(j,k)] \qquad [2]$$

then the various hierarchical algorithms discussed in the section 'Hierarchical Clustering' all produce the same clusters. The dissimilarity between a given object and another object in a given cluster C is less than the dissimilarity between that object and another object not in C. Thus hierarchical clustering may be viewed as approximating the given dissimilarity matrix by an ultrametric; methods have been proposed that search for a good approximation based on a measure of closeness between different dissimilarities.

With slightly more generality, *evolutionary distances* between a set of objects conform to a tree in which the objects are end nodes and the internal nodes are arbitrary. A nonnegative link distance is associated with each link of the tree, and the evolutionary distance between two objects is the sum of link distances on the path connecting them. Evolutionary distances are characterized by the following property: any four objects i, j, k, l may be partitioned into two pairs of objects in three ways; for some pairing (i, j), (k, l) say, the *evolutionary equality* is satisfied

$$d(i,k) + d(j,k) = d(i,l) + d(j,k) \qquad [3]$$

The pairing (i, j), (k, l) means that the tree path from i to j has no links in common with the tree path from k to l. We can construct an evolutionary tree on a set of objects by approximating a given dissimilarity by an evolutionary distance.

One case where evolutionary distances arise probabilistically is in the factor analysis model in which the objects to be classified form a set of v variables on which a number of independent v-dimensional observations have been made (Edwards and Cavalli-Sforza, 1964). The variables are taken to be end nodes in an evolutionary tree whose links and internal objects remain to be determined. Associated with each link of the tree is an independent normal variable with mean zero and variance depending on the link; this variance is the link distance. Each observation for the v variables is generated by assigning the random link values; then the variable $V_i - V_j$ is a sum of independent normal variables along the path of directed links connecting i and j. The random normal associated with each link is of opposite sign when the link is traversed in the opposite direction. The quantities $d(i, j) = E(V_i - V_j)^2$ form a set of evolutionary distances, which are estimated by $\widehat{d}(i,j) = \text{sample have } (V_i - V_j)^2$.

For each observation $\{V_1 - V_i, i = 2, \ldots p\}$ is multivariate normal with a covariance matrix determined by the link variances, and the log likelihood may be written, for a certain weight function w

$$\sum_{i,j} \left(\widehat{d}(i,j) - d(i,j)\right) w_{i,j}(\sigma) \qquad [4]$$

This expression is maximized to estimate the unknown σ values. In this case the sample distances form a sufficient statistic for the parameters, and no information is lost in using distances rather than the complete data set. Similar discrete Markovian models due to Felsenstein (1973) reconstruct evolutionary trees from DNA sequence data.

A nonparametric model for hierarchical clustering assumes that the objects $x_1, \ldots, x_n$ to be clustered are sampled randomly from a space with values $x \in S$ and density f with respect to some underlying measure, usually counting measure or Lebesgue measure. It is necessary to have a concept of connectedness in the space S. The family of connected sets of points having density $f(x) > c$, computed for each nonnegative c, together form a hierarchical clustering called *high density* clusters (Bock, 1974; Hartigan, 1975). The statistical approach estimates this hierarchical clustering on the density f from the given sample $x_1, \ldots, x_n$ by first estimating the density f by $\widehat{f}$ say, then forming the estimated clusters as the high density clusters in $\widehat{f}$. There are numerous parametric and nonparametric estimates of density available.

How do the standard hierarchical clustering methods do in estimating these high density clusters? Badly. The single linkage clusters may be reexpressed as the high density clusters obtained from the nearest-neighbor density estimator; the density estimate at point x is a decreasing function of $d(x, x_i)$, where x_i is the closest sample point to x. This estimator is inconsistent as an estimate of the density, and the high density clusters are correspondingly poorly estimated. However, it may be shown that the single linkage method is able to consistently identify distinct modes in the density. On the other hand, complete linkage corresponds to no kind of density estimate at all, and the clusters obtained from complete linkage are not consistent for anything at all, varying randomly as the amount of data

increases. In evaluating and comparing clustering methods, a statistical approach determines how well the various clustering methods recover some known clustering when data is generated according to some model based on the known clustering (Milligan, 1980). The most frequently used model is the multivariate normal mixture model.

Validity of Clusters

A typical agglomerative clustering produces $n - 1$ clusters for n objects, and different choices of dissimilarities and cluster combination rules will produce many different clusterings. It is necessary to decide 'which dissimilarity, which agglomeration method, and how many clusters'.

In partition problems, if a parametric mixture model is used, standard statistical procedures described in the section 'Mixture Models' are available to help decide how many components are necessary. These models require specific assumptions about the distribution within components that can be critical to the results. The nonparametric *high density* cluster model associates clusters with modes in the density. The question about number of clusters translates into a question about number of modes.

The *threshold or kernel width* test (Silverman, 1981) for the number of modes is based on a kernel density estimator, for example, given a measure of dissimilarity, and a threshold d, use the density estimate at each point x consisting of the sum of $1 - d(x, x_i)/d$ over observed objects x_i within dissimilarity d of x. The number of modes in the population is estimated by the number of modes in the density estimate. To decide whether or not two modes are indicated, the quantity d is increased until at $d = \widehat{d}_1$, the estimated density $\widehat{f}_1$ first has a single mode. If the value $\widehat{d}_1$ required to render the estimate unimodal is large enough, then more than one mode is necessary. To decide when $\widehat{d}_1$ is large, resample from the unimodal distribution $\widehat{f}_1$, obtaining a distribution of resampled $\widehat{d}_1$ values for those samples, the values of d in each of those resamples required to make the density estimate unimodal. The tail probability associated with the given $\widehat{d}_1$ is the proportion of resampled $\widehat{d}_1$ that exceed the given $\widehat{d}_1$. A similar procedure works for deciding whether or not a number of modes larger than 2 is justified. An alternative test that applies in one dimension for two modes versus one uses the DIP statistic (Hartigan and Hartigan, 1985), the maximum distance between the empirical distribution function and the unimodal distribution function that is closest to the empirical distribution function. The null distribution is based on sampling from the uniform.

Probabilistic Search

Consider searching for a classification C in a class Γ that will maximize $l(C)$, the log probability of C given the data. Γ is usually too large for a complete search, and a local optimization method is used. Define a neighborhood of classifications $N(C)$ for each classification C. Beginning with some initial classification C_0, steepest ascent selects C_i to be the neighbor of C_{i-1} that maximizes $l(C)$, and the algorithm terminates at C_m when no neighbor has greater probability. Of course, the local optimum depends on the initial classification C_0, and the local optimum will not in general be the global optimum.

The search may be widened probabilistically by *randomized steepest ascent or simulated annealing* (Klein and Dubes, 1989), as follows: rather than choosing C_i to maximize $l(C)$ in the neighborhood of C_{i-1}, choose C_i with probability proportional to $\exp(1(C)/\lambda)$ from the neighborhood of C_{i-1}. In running the algorithm, we keep track of the highest probability classification obtained to date, and may return to that classification to restart the search. The degree of randomness λ may be varied during the course of the search. To carry out randomized ascent at randomness λ add to the log probability of each classification the independent random quantity $-\lambda \log E$, where E is exponential; then do perform deterministic steepest ascent on the randomized criterion.

To be specific, consider the probabilification of the *k-means* algorithm, that constructs say k clusters of n objects in v dimensions to minimize the total within cluster sum of squares S. Use a probability model in which within cluster deviations are independent normal with variance σ^2, the cluster means are uniformly distributed, and σ^2 has density $\sigma^{-(2vk+4)}$ (these prior assumptions match maximum likelihood), the log probability of a particular classification into two clusters given the data is

$$-\left[\frac{1}{2}(n+1)vk+1\right]\log S \qquad [5]$$

Steepest ascent k-means switches between clusters the observation that most decreases S; randomized steepest ascent switches the observation that most increases the randomized criterion

$$-\left[\frac{1}{2}(n+1)vk+1\right]\log S - \lambda \log E \qquad [6]$$

Finally, to choose the number of clusters, allowing k to vary by permitting an observation to 'switch' to a new singleton cluster. Penalize against many clusters by the BIC penalty; the number of parameters (cluster means and the variance) specifying k clusters in v dimensions is $vk + 1$; so the final randomized criterion is

$$-\frac{1}{2}nvk \log S - \lambda \log E - \frac{1}{2}(vk+1)\log n \qquad [7]$$

This formula adjusts only for the increase in log probability due to selecting optimal parameter values, and not for the increase in the log probability caused by selecting the highest probability partitions. A number of alternative choices of information criterion are proposed by Bozdogan (1988).

Conclusion

Although probabilistic and statistical methods have much to offer classification, in providing formal methods of constructing classes, in validating those classes, in evaluating algorithms, and in spiffing up algorithms, it is important to realize that classifications have something to contribute to statistics and

probability as well. The early extensive development of taxonomies of animals and plants in Biology proceeded quite satisfactorily without any formal statistics or probability through 1950; the big methodological change occurred in the late-nineteenth century following Darwin, who demanded that biological classification should follow genealogy.

Classification is a primitive elementary basic form of representing knowledge that is prior to statistics and probability. Many of the standard algorithms were developed without allowing for uncertainty. Many of them, such as complete linkage, say, do not seem supportable by any statistical reasoning. Others will benefit by statistical evaluations and by probabilistic enhancements, as suggested in Hartigan (2000) and Gordon (2000).

Classification is both fundamental and subjective. If an observer classifies a set of objects to be in a class, for whatever reason, that class is valid in itself. Thus, there is no correct natural classification that our formal techniques can aim for, or that our statistical techniques can test for. Since classification is fundamental, it should be used to express formal knowledge, and probabilistic inferences and predictions should be based on these classificatory expressions of knowledge, Hartigan (2000). More formal and refined classifications may be assisted and facilitated by probability models reflecting and summarizing the prior fundamental classifications.

See also: Algorithm; Algorithmic Complexity; Hierarchical Models: Random and Fixed Effects; Mixture Models in Statistics.

Bibliography

Arabie, P., Hubert, L.J., De Soete, G. (Eds.), 1996. Clustering and Classification. World Scientific, Singapore.

Bock, H.H., 1972. Statistiche Modelle und Bayes'sche Verfahren zur Bestimmung einer unbekannten Klassifikation normal-veteilter zufälliger Vektoren. Metrika 18, 120–132.

Bock, H.H., 1974. Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung and Strukturierung von Daten (Clusteranalyse). Vandenhoek & Ruprecht, Göttingen, Germany.

Bozdogan, H., 1988. A new model selection criterion. In: Bock, H.H. (Ed.), Classification and Related Methods of Data Analysis. North Holland, Amsterdam, pp. 599–608.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman and Hall, New York.

Comaniciu, D., Ramesh, V., Del Bue, A., 2002. Multivariate saddle point detection for statistical clustering. In: Computer Vision—ECCV 2002. Springer, New York, pp. 561–576.

Datta, Susmita, Datta, Somnath, 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19 (4), 459–466.

Edwards, A.W.F., Cavalli-Sforza, L.L., 1964. Reconstruction of evolutionary trees. In: Heywood, V.N., McNeill, J. (Eds.), Phenetic and Phylogenetic Classification. The Systematics Association, London, pp. 67–76.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology 22, 240–249.

Fitch, W.M., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279–284.

Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H., Zubrzycki, S., 1951. Sur la liaison et la division des points d'un ensemble fini. Colloquium Mathematicum 2, 282–285.

Gordon, A.D., 2000. Classification. Chapman and Hall, New York.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.

Hartigan, J.A., 2000. Classifier probabilities. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M. (Eds.), Data Analysis, Classification, and Related Methods. Springer-Verlag, Berlin, pp. 3–15.

Hartigan, J.A., Hartigan, P.M., 1985. The dip test of unimodality. Annals of Statistics 13, 70–84.

Jardine, N., Sibson, R., 1968. The construction of hierarchic and non-hierarchic classifications. Computer Journal 11, 77–84.

Klein, R.W., Dubes, R.C., 1989. Experiments in projection and clustering by simulated annealing. Pattern Recognition 22, 213–220.

Kruskal, J.B., 1956. On the shortest spanning subtree of a graph and the travelling salesman problem. Proceedings of the American Mathematical Society 7, 48–50.

Lance, G.N., Williams, W.T., 1967. A general theory of classificatory sorting strategies. I: hierarchical systems. Computer Journal 9, 373–380.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: LeCam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 281–297.

Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.

Pearson, K., 1984. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London, Series A 185, 71–110.

Pollard, D., 1981. Strong consistency of *k*-means clustering. Annals of Statistics 9, 135–140.

Silverman, B.W., 1981. Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society B 43, 97–99.

Sokal, R.R., Sneath, P.H.A., 1963. Principles of Numerical Taxonomy. Freeman, San Francisco.

Swingley, D., 2005. Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology 50 (1), 86–132.

Statistical Distribution Models: Goodness of Fit Tests

John CW Rayner, University of Wollongong, Wollongong, NSW, Australia; and University of Newcastle, Callaghan, NSW, Australia
Oliver Thas, Ghent University, Gent, Belgium; and University of Wollongong, Wollongong, NSW, Australia
Donald J Best, University of Newcastle, Callaghan, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of a one-sample test of fit is to give an objective measure of how well a probability model agrees with observed data. Here we discuss the test of Karl Pearson and derivatives of it, tests based on the empirical distribution function and the construction of the Neyman–Barton smooth tests. In the final section, we then address some modern developments in smooth testing: diagnostics, Cholesky components, data-driven tests and model selection. Other tests of fit, such as correlation tests and Laplace transform tests, are not considered here.

Introduction

The purpose of a one-sample test of fit is to give an objective assessment of how well a probability model agrees with observed data. Otherwise, as Pearson (1900: p. 171) said: "the comparison of observation and theory in general amounts to a remark – based on no quantitative criterion." Reviews of Pearson's paper may be found in Plackett (1983) and Rayner et al. (2009: Sect. 2.2). The multiple sample goodness of fit problem assesses whether data from a number of populations are consistent with having come from the same population. If the answer is affirmative, then the one-sample question may be appropriate. We will not consider the multiple sample problem here.

In many situations it will be valuable to find out if data for a particular event are consistent with an established pattern or model. Specifically, (1) can extreme events, such as flood heights, insurance losses, and equity risks be modeled by the extreme value distribution, (2) are right skewed data, that might arise, for example, in life testing and reliability contexts, consistent with the Rayleigh distribution, and (3) are data, usually binomial in similar scenarios but too overdispersed to be so in the current context, consistent with the beta-binomial? This sometimes occurs with taste test and market research data.

Rayner et al. (2009) identify three benefits of having applied a goodness of fit test:

- a compact description of the data;
- illumination of the data generation mechanisms; and
- validation or not of distributional assumptions necessary for the application of powerful parametric procedures.

The best of modern methods add another benefit: selection of a more valid model than that originally hypothesized.

Here we discuss tests for multinomial models, tests based on the empirical distribution function (EDF), and the construction of the Neyman–Barton smooth tests. In the final section we then address some modern developments in smooth testing: diagnostics, Cholesky components and model selection. Other tests of fit have been suggested, but we shall not consider these here, except to briefly mention two. Correlation tests are sometimes suggested in conjunction with quantile–quantile or probability plots of the data. The Shapiro–Wilk test of normality, for example, may be considered a correlation test. See D'Agostino and Stephens (1986). Tests based on the empirical Laplace transform are a general class of tests and often have competitive power. See, for example, Henze and Meintanis (2002, 2012).

Tests for Multinomial Models

For Pearson's test, sometimes called the Pearson chi-squared test, observations O_1, O_2, …, O_m are assumed to come from m nonoverlapping classes that are expected to contain E_1, E_2, …, E_m observations. The test statistic is

$$X_P^2 = \sum_{i=1}^{m} (O_i - E_i)^2 / E_i$$

Pearson (1900) assumed that parameters of the probability model giving rise to the cell expectations were known, and showed that the asymptotic distribution of X_P^2 is the χ^2_{m-1} distribution.

There is a considerable literature on how, when there is choice, the cells may best be constructed. The recommendation that for the χ^2 approximation to be adequate the cell expectation be at least five, goes back to Sir Ronald Fisher writing in the 1920s. This advice is now seen as extremely conservative.

Most interesting applications involve estimating parameters. Pearson incorrectly asserted that estimating parameters makes no difference to the asymptotic null distribution of the test statistic. If the cell expectations are calculated from parameters estimated efficiently from the grouped data, then Sir Ronald Fisher, again writing in the 1920s, showed that if q parameters are estimated from the grouped data using maximum likelihood the asymptotic distribution is χ^2_{m-q-1}. However, if estimation is via maximum likelihood using the ungrouped data Chernoff and Lehmann (1954) showed that Pearson's statistic no longer had an asymptotic chi-squared distribution. See Rayner et al. (2009) for a fuller discussion of cell construction and multinomial testing when parameters need to be estimated.

There are many possible generalizations of Pearson's test. Neyman's version is based on the test statistic

$$X_N^2 = \sum_{i=1}^{m} (O_i - E_i)^2 / O_i$$

One interesting possibility is the Cressie–Read family, based on the test statistic

$$I_\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{m} O_i \left\{ \left(\frac{O_i}{E_i} \right)^\lambda - 1 \right\}$$

Taking $\lambda = 1$ leads to X_P^2, $\lambda = -1$ leads to a modified likelihood ratio statistic, $\lambda = 0$ is the usual likelihood ratio statistic, $\lambda = -0.5$ is the Freeman–Tukey statistic, and $\lambda = -2$ leads to X_N^2. Other choices can lead to tests with excellent properties; see Read and Cressie (1988).

EDF Tests

Much data are essentially continuous, and grouping it to permit the use of X_P^2 and other statistics based on the multinomial loses information and often power. A different approach that does not group the data is to compare the EDF with the distribution function specified by the hypothesized probability distribution. The Russian mathematician Kolmogorov (1933) proposed the first such test. EDF tests of fit, especially for continuous distributions, are reviewed in D'Agostino and Stephens (1986). Tests of fit for discrete distributions using the EDF approach are considered in Choulakian et al. (1994). See also Spinelli and Stephens (1997).

If we have n data points, the EDF $F_n(x)$ is given by

$$F_n(x) = (\text{number of observations } \leq x)/n$$

Tests of fit for a hypothesized distribution function, $F(x)$, can be made by looking at the difference between $F_n(x)$ and $F(x)$ for each data point. Kolmogorov (1933) suggested the statistic

$$\text{KS} = \max_x |F_n(x) - F(x)|$$

while a generally more powerful test is based on the Anderson and Darling (1954) statistic. Its calculation is illustrated in Example 1 below.

It can be shown that the EDF test statistics are linear combinations of components with decreasing weights. The components are similar to those defined in the following section on Neyman Smooth Tests. See, for example, Stephens (1974). If the null hypothesis specifies a discrete distribution where the number of classes is fixed and not too large (say five or less), we suggest calculating X_P^2 to assess goodness of fit. Otherwise, the Anderson–Darling test is a good choice, as it weights the important early components heavily and gives a powerful omnibus assessment.

Example 1: Testing Normality

Consider the following set of examination marks from 20 students:

53, 15, 70, 73, 79, 48, 91, 20, 24, 91, 87, 15, 3, 78, 78, 62, 16, 15, 20, 32.

Are these data normally distributed?

First we apply the Pearson–Fisher test with class boundaries –12.5, 12.5, 37.5, 62.5, 87.5, 112.5 and midpoints 0, 25, 50, 75, 100. The cell counts are 1, 8, 3, 6, 2. We find $X_{PF}^2 = 5.22$ with p-value 0.07 using the χ_2^2 distribution. The normality assumption would seem to be marginal.

We now apply the Anderson–Darling test. We calculate $\bar{x} = 48.5$ and $s = 30.8$ and then find $y_i = (x_i - \bar{x})/s$ for $i = 1, \ldots, 20$, where the x_i are the 20 exam marks. Next use a computer routine or tables of the standard normal distribution to find

$$z_i = F(y_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_i} \exp(-x^2/2)\mathrm{d}x$$

and order the z_i as $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Then the Anderson–Darling statistic, AD say, is

$$\text{AD} = -n - n^{-1} \sum_{i=1}^{n} (2i-1) \left\{ \log z_{(i)} + \log z_{(n+1-i)} \right\}$$

which, for the above data, takes the value 0.931. From tables of critical values, such as D'Agostino and Stephens (1986: Table 4.9), or using MINITAB's Normplot routine, we find a p-value of 0.014. This implies the normality hypothesis is doubtful and the marks are not well summarized by $\bar{x} \pm s$. We revisit this example later.

Example 2: Testing the Poisson Assumption

Consider the number of people in a shopping mall contributing to a charity collector during 5-min intervals:

Number of contributors	0	1	2	3	4	5	6	7	8
Number of intervals	9	27	26	17	10	8	4	3	1

The sample size is $n = 105$. Let the observed and expected counts assuming a Poisson distribution be o_j and $e_j = np_j$, respectively, for $j = 0, 1, \ldots, 8$. Pooling cells from 7 onwards to get all cell expectations greater than one, we find $X_{PF}^2 = 8.62$ with p-value 0.20 using the χ_7^2 distribution. At all reasonable levels the Poisson assumption is acceptable.

Now let $z_j = \sum_{i=0}^{j} (o_i - e_i)$, in which case

$$\text{AD} = \frac{1}{n} \sum_{j=0}^{7} \frac{\hat{z}_j^2 \hat{p}_j}{\hat{h}_j \left(1 - \hat{h}_j\right)}$$

where the '^' indicates we have estimated the Poisson parameter by the sample mean. We find that AD = 1.178, and, using Spinelli and Stephens (1997: Table 1), the p-value is between 0.025 and 0.050. Thus the distribution is not well fitted by a Poisson model. It seems plausible that contributors come in clusters rather than at random, as a Poisson model would imply. The Pearson–Fisher test is far less critical of the Poisson assumption than the Anderson–Darling test.

Neyman Smooth Tests

One problem with Pearson's chi-squared test is that it is an omnibus test, with some power against general alternative

distributions to the hypothesized distribution, but with relatively poor power against important directional alternatives. Alternatives that are often of great practical importance are differences between the data and the mean, variance and, to a lesser extent, the skewness and kurtosis of the specified distribution. A test that does have good power at detecting these moment alternatives is Neyman's (1937) 'smooth' test. It has the useful property that it can readily be partitioned into *components* that give a powerful and detailed scrutiny of the data. See Rayner et al. (2009: Section 4.1) for a review of Neyman (1937).

The Neyman smooth tests may be applied to both categorized and uncategorized distributions and distributions that may or may not involve nuisance parameters. See Rayner and Best (1990) and Rayner et al. (2011) for overviews. Most modern authors use orthonormal functions, as did Neyman (1937), and permit nuisance parameters, contrary to Neyman (1937). The smooth tests may be used to test for very general distributions, provided orthonormal functions on the specified distribution may be constructed. Recurrence formulae for orthonormal polynomials are given for univariate distributions in Rayner et al. (2008) and for bivariate distributions in Rayner et al. (2013).

The test is based on the score statistic, and hence is weakly optimal, and the test statistic has a convenient asymptotic chi-squared distribution. In our preferred formulation components are readily available, and are asymptotically independent, each asymptotically having the standard normal distribution. The components are often the basis for well-known tests of fit in their own right, but more importantly, may permit a detailed, convenient and informative scrutiny of the data. The order of the test (see below) is at the user's discretion, and it may be chosen to give an omnibus or directional test, or something between.

Suppose we wish to test for a probability density function $f_X(x;\beta)$, where $\beta = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ is a $q \times 1$ vector of parameters (such as μ and σ in the normal case). This null probability density function is nested in the order k probability density function

$$g_k^{\mathrm{N}}(x;\theta,\beta) = C(\theta,\beta)exp\left\{\sum_{i=1}^{k}\theta_i\pi_i(x;\beta)\right\}f_X(x;\beta)$$

(N is for Neyman) where $\theta = (\theta_1, \ldots, \theta_k)^T$ is a vector of k real parameters, $C(\theta, \beta)$ is a normalizing constant that is assumed to exist, and $\{\pi_i(x;\beta)\}$ is a set of functions orthonormal on $f_X(x;\beta)$ with $\pi_0(x;\beta) = 1$ for all x. Thus for $f_X(x;\beta)$ continuous,

$$\int_{-\infty}^{\infty}\pi_i(x;\beta)\pi_j(x;\beta)f_X(x;\beta)\mathrm{d}x = 1 \text{ for } i = j \text{ and zero otherwise.}$$

Sometimes $C(\theta, \beta)$ may not exist but the statistic we are about to define will. Using a random sample $X_1, \ldots, X_n$ we test for $f_X(x;\beta)$ parametrically, by testing H_0: $\theta = 0$ against K: $\theta \neq 0$, with β a vector of nuisance parameters. For small k the alternatives vary 'smoothly' from the null, in the sense of Neyman (1937). As an example of this, in testing for the standard normal distribution, the order one alternative is a normal distribution with a mean shift, the order two alternative is a normal distribution with both a mean shift and a variance shift, and the third order alternative is more involved, with mean, variance, and skewness differences from the hypothesized normal distribution.

Neyman (1937) developed this approach for the case of no nuisance parameters. In this situation the probability integral transformation converts the problem of testing for a completely specified distribution to testing for the uniform (0, 1) distribution. Thomas and Pierce (1979) and Kopecky and Pierce (1979) considered smooth tests for general distributions involving nuisance parameters. Their tests are based on a family of alternatives specified using powers of the cumulative distribution function rather than orthogonal polynomials as Neyman (1937) used, and as a consequence they require tables of coefficients for their implementation. Our preferred formulation permits nuisance parameters and uses orthonormal functions, thereby avoiding the need for tables of coefficients. For a more detailed account of the history of the smooth tests, see Rayner et al. (2009: Section 1.2).

When no nuisance parameters are present, the smooth test statistic based on the model $g_k^{\mathrm{N}}(x;\theta,\beta)$ is

$$S_k = V_1^2 + \ldots + V_k^2$$

in which the components V_r are given by

$$V_r = \sum_{j=1}^{n}\pi_i(X_j;\beta)\Big/\sqrt{n}$$

When testing for a distribution from an exponential family of distributions, such as the binomial, Poisson, geometric, or normal, the likelihood equations for $\widehat{\beta}$, the maximum likelihood estimator of β, are $\widehat{V}_r = V_r(\widehat{\beta}) = 0$ for $r = 1, \ldots, q$. The model $g_k^{\mathrm{N}}(x;\theta,\beta)$ results in a singular asymptotic covariance matrix for $(\widehat{V}_r)$. This can be resolved by removing $\theta_1, \ldots, \theta_q$ from the model, that is, by modifying $g_k^{\mathrm{N}}(x;\theta,\beta)$ to $C(\theta,\beta)exp\{\sum_{i=q+1}^{q+k}\theta_i\pi_i(x;\beta)\}f_X(x;\beta)$. What is happening here is that $\theta_1, \ldots, \theta_q$ are playing the same role as the elements of β, so the presence of all these parameters in the model leads to a redundancy. The test statistic is

$$\widehat{S}_k = \widehat{V}_{q+1}^2 + \ldots + \widehat{V}_{q+k}^2$$

in which $\widehat{V}_r = \sum_{j=1}^{n}\pi_i(X_j;\widehat{\beta})/\sqrt{n}$.

When testing for a distribution outside of an exponential family the score test statistic is not a simple sum of squares: it is a quadratic form. This will be considered in the next section.

It may occur that in $g_k^{\mathrm{N}}(x;\theta,\beta)$ a normalizing constant $C(\theta, \beta)$ cannot be found. An alternative smooth model is due to Barton (1953, 1955, 1956):

$$g_k^{B}(x;\theta,\beta) = \left\{1 + \sum_{i=1}^{k}\theta_i h_i(x;\beta)\right\}f(x;\beta)$$

(B is for Barton). These densities do not require the existence of the normalizing constant, but may be negative. This negativity may be handled using adjustments, such as those proposed in Gajek (1986) and in Glad et al. (2003), but this is only a problem for nonparametric density estimation and not for the smooth tests discussed here. For the Barton model the relevant previous results for the Neyman alternatives remain true.

Example 2 Revisited

For the Poisson, the appropriate orthonormal polynomials are known as Poisson–Charlier polynomials. We have

$$\pi_r(y) = \sqrt{(\lambda^r/r!)}\sum_{j=0}^{r}(-1)^{r-j}\left({}^rC_j\right)(j!)\left(\lambda^{-j}\right)\left({}^yC_j\right)$$

where $y = 0, 1, 2, \ldots$, from which the orthonormal polynomials to second order are

$$\pi_0(y) = 1 \text{ for all y}, \quad \pi_1(y) = (y/\lambda - 1)\sqrt{\lambda}, \quad \pi_2(y) = \left\{y(y-1)/\lambda^2 - 2y/\lambda + 1\right\}\lambda\sqrt{2}$$

Due to the estimation of λ, $\widehat{V}_1 = 0$. For the data here $\widehat{V}_2 = 2.12$, suggesting a model with greater dispersion than the Poisson. Note that $\widehat{V}_2$ is a standardized version of Fisher's Poisson index of dispersion.

For count data the objective is to test for a multinomial distribution with specified cell probabilities $p_1, \ldots, p_m$. When there are no nuisance parameters this is done by imbedding these probabilities in the smooth alternative cell probabilities

$$\pi_j C(\theta)\exp\left\{\sum_{i=1}^{k}\theta_i h_{ij}\right\}p_j, \quad j = 1, \ldots, m$$

and testing H_0: $\theta = 0$ against K: $\theta \neq 0$. In this formulation the θ_i, $i = 1, \ldots, k$ are real parameters, $\theta = (\theta_1, \ldots, \theta_k)^T$, $C(\theta)$ is a normalizing constant so that $\pi_1 + \ldots + \pi_m = 1$, and the h_{rs} satisfy, for $r, s = 1, \ldots, m$,

$$\sum_{j=1}^{m} h_{rj}h_{sj}p_j = \delta_{rs}, \text{ with } h_{mj} = 1, \quad j = 1, \ldots, m$$

Given cell counts $N_1, \ldots, N_m$, the score test statistic when $k = m - 1$ is Pearson's $X_P^2 = \sum_{j=1}^{m}(N_j - np_j)^2/(np_j)$; more familiarly $X_P^2 = \sum_{i=1}^{m}(O_i - E_i)^2/E_i$.

A generalization of this formulation to allow for nuisance parameters leads to the Pearson–Fisher test when the cell probabilities are estimated by maximum likelihood from the multinomial, and to the Rao and Robson (1974) test when the cell probabilities are estimated by maximum likelihood from the ungrouped data. These are thus identified as smooth tests.

It should be emphasized that using different orthonormal functions in the smooth model results in different tests that detect different alternatives. Which orthonormal functions should be chosen depends on which alternatives one hopes to most powerfully detect. If the hypothesized distribution were uniform, we could choose the Legrendre polynomials, as did Neyman (1937). If it were desirable to test for periodic alternatives we could consider using the series $\{\sin(i\pi x)\sqrt{2}\}$, or the Walsh functions as did Hamdan (1964). The aim is to find an orthonormal series that represents the alternatives of interest in as few terms as possible. Greater power results from doing this. A helpful R package is available at http://www.biomath.ugent.be/~othas/smooth2/Home.html.

Each smooth test for a particular distribution requires an independent study. The small sample distributions of the test statistic and its components should be investigated. Generally the approach to the asymptotic chi-squared distribution is so slow that we recommend finding p-values using the parametric bootstrap. Moreover the powers of the smooth tests should be compared with appropriate competitor tests. In general the most powerful smooth test of fit test is likely to be the sum of the squares of the first few nonzero components; $\widehat{S}_{q+k}$ with $k = 2$, 3, or 4. We recommend augmenting this by the use of the components $\widehat{V}_{q+1}^2, \ldots, \widehat{V}_{q+k}^2$ in an exploratory data analysis fashion. However, in particular cases there are variations to this advice.

Sometimes the data are only available in grouped form. See Best and Rayner (2007) for testing for the grouped exponential and Best et al. (2008) for testing for the grouped normal.

Example 1 Revisited

For the exam mark data in Section 2.1, calculate $y_i = (x_i - \bar{x})/s$. The orthonormal polynomials appropriate for testing normality are often called the Hermite–Chebyshev polynomials. We have

$$\pi_0(y) = 1 \text{ for all } y; \pi_1(y) = y;$$

$$\pi_2(y) = \left(y^2 - 1\right)\Big/\sqrt{2};$$

$$\pi_3(y) = \left(y^3 - 3y\right)\Big/\sqrt{6};$$

$$\pi_4(y) = \left(y^4 - 6y^2 + 3\right)\Big/\sqrt{24};$$

$$\pi_5(y) = \left(y^5 - 10y^3 + 15y\right)\Big/\sqrt{120};$$

and

$$\pi_6(y) = \left(y^6 - 15y^4 + 45y^2 - 15\right)\Big/\sqrt{720}.$$

Given these we calculate, for the above data, $\widehat{V}_1 = \widehat{V}_2 = 0$, due to the estimation of μ and σ, and $\widehat{V}_3 = -0.0361$, $\widehat{V}_4 = -1.4656$, $\widehat{V}_5 = -0.0462$, $\widehat{V}_6 = 1.8907$. Note that $\widehat{V}_3$ and $\widehat{V}_4$ are standardized versions of the usual skewness and kurtosis coefficients. As $\widehat{V}_3, \ldots, \widehat{V}_6$ have asymptotic standard normal distributions, we see that $\widehat{V}_6$ is near significance at the 5% level and that $\widehat{V}_4$ and $\widehat{V}_6$ are much larger than the odd-order moment statistics $\widehat{V}_3$ and $\widehat{V}_5$. These observations suggest an alternative probability model that is symmetric and shorter tailed than the normal. Thus a uniform distribution is a possibility so that we could summarize the data as 'evenly' spread over the range (0, 100) rather than as 48.5 ± 30.7. If the odd-order $\widehat{V}_r$ had dominated the even -order values then a skewed alternative model is suggested, while if both odd and even $\widehat{V}_r$ are large, $|\widehat{V}_r| > 1.65$ say, then a symmetric alternative with longer tails then the normal, such as the Laplace, is indicated.

Modern Developments in Smooth Testing

Are Components Diagnostic?

When testing for a distribution from an exponential family the smooth test statistic is the sum of the squares of specified

components. In particular a single component can be used to test for the distribution. There is a considerable literature in recent years on whether a test based on the rth component may be interpreted as diagnosing the failure or not of the data to be consistent with the rth moment of the specified distribution. Simulation studies demonstrate that for virtually any null distribution component tests are not diagnostic. It is now known that if V_r is significantly large, then any or all of the nonnull moments up to the $2r$th could be the cause.

This issue would all be resolved if, in the score test statistics, the asymptotic covariance matrix was replaced by one that estimated the component variances and covariances consistently, under both the null and alternative hypotheses. Henze and Klar (1996), Henze (1997) and Klar (2000) worked in this vein. Unfortunately the simulation studies in Klar (2000) and Thas et al. (2009) show that convergence of these 'rescaled' components to their asymptotic limits is extremely slow, with samples as large as 10 000 required to achieve satisfactory results. Thus rescaling does not create diagnostic components and inference, for example, to obtain p-values, should involve resampling methods rather than the asymptotic distributions of the statistics involved. Nevertheless, although in the finite samples that occur in practice the rescaled components may be somewhat 'tainted' by higher order moments, it is reasonable to say they are more diagnostic than the raw, unscaled components.

In the power studies we have sighted, powers based on the rescaled components are often less than those based on the unscaled components, and where power gains are achieved it is often at the expense of power loss for alternatives elsewhere in the parameter space.

Our personal preference is to not use rescaled components. When testing for distributions within exponential families, for formal testing we prefer to use $\widehat{V}_{q+1}^2 + \ldots + \widehat{V}_{q+k}^2$ with $k = 2, 3,$ or 4, depending on the hypothesized distribution, and use the individual components in a data analytic fashion, as in Example 1 revisited. The first significant component beyond the qth at best *suggests* that the corresponding moment is the cause of the model failure. Outside of exponential families the situation is a little different.

Generalized Smooth Tests and Cholesky Components

The score test statistic for a smooth model is a quadratic form in the components. If the null hypothesis specifies a distribution from an exponential family of distributions then the score test statistic $\widehat{S}_k$ has the appealing form of being a sum of squares of components that under the null hypothesis are asymptotically independent and asymptotically standard normal. Outside of exponential families this convenient form cannot be expected. This includes when testing for the zero-inflated Poisson, extreme value, negative binomial, generalized Pareto (Rayner et al., 2009: Chapter 11) and gamma (De Boeck et al., 2011). The difficulty for these distributions is that the components are not even asymptotically uncorrelated, so the significance of one component may be associated with the significance of others.

Outside of exponential families an alternative approach is to use the generalized score test. Information about generalized score tests is given, for example, in Rippon and Rayner (2010) and Thas (2010). The generalized smooth tests use the generalized score tests, with the $q \times 1$ nuisance parameter β being estimated by solving $V_r = 0$, for $r = 1, \ldots, q$. The solution, $\tilde{\beta}_0$, is a method of moments estimator. These estimators are not usually fully efficient, and their use often means estimating efficiency is sacrificed to gain interpretable components.

The generalized score test statistic for the Neyman and Barton models is of the form $\tilde{V}^T \tilde{\Sigma}^{-1} \tilde{V}$, where $\tilde{\Sigma}$ is a consistent estimator of the asymptotic covariance matrix of the score $\tilde{V} = (\tilde{V}_{q+1}, \ldots, +\tilde{V}_{q+k})^T$. A *Cholesky decomposition* of $\tilde{\Sigma}^{-1}$ gives $\tilde{\Sigma}^{-1} = MM^T$, where M is upper triangular. Putting $\tilde{V}^* = M^T\tilde{V} = (\tilde{V}^*_{q+1}, \ldots + \tilde{V}^*_{q+k})^T$ gives $\tilde{V}^T\tilde{\Sigma}^{-1}\tilde{V} = \tilde{V}^{*T}\tilde{V}^*$. Thus the generalized score test statistic is of the form $(\tilde{V}^*_{q+1})^2 + \ldots + (\tilde{V}^*_{q+k})^2$. Since for most distributions of interest a multivariate central limit theorem applies and $\tilde{V}^*$ has asymptotic covariance matrix the identity, the elements $\tilde{V}^*_r$ of $\tilde{V}^*$ are asymptotically independent and asymptotically standard normal.

Since M^T is lower triangular, $\tilde{V}^*_r$ is the sum of the first r elements of $\tilde{V}$. It follows that the previous discussion about whether or not the components $\widehat{V}_r$ in the smooth test are diagnostic applies equally to the components $\tilde{V}^*_r$ in the generalized smooth test.

The outcome of this discussion is that when testing for *any* distribution, the generalized score test and Cholesky decomposition together yield components that are equally as convenient as those resulting from the score test when testing for distributions from exponential families. Moreover, because of their construction using the Cholesky decomposition, the data and the model agree in moments up to the qth. If the rth ($r > q$) is first significant Cholesky component, this suggests the data and the model do not agree in moments up to the rth, although the significance may be due to moments up to the $2r$th.

The only caveat on this approach is that the orthonormal polynomial of order r requires the existence of the first $2r$ moments of the null distribution. Thus, for example, using this approach it is not possible to directly test for the Cauchy, which has no moments of any order. One option would be to use the probability integral transformation to essentially test for the continuous uniform distribution on (0, 1). However, the resulting components are then difficult to interpret in terms of the original null hypothesis.

Data-Driven Testing

One of the difficulties associated with the definitions of $g_k^N(x;\theta,\beta)$ and $g_k^B(x;\theta,\beta)$ is the determination of the order k. In finite samples misspecifying the order k can seriously affect the power of the test. If k is chosen to be large, the test will give good protection against a wide range of alternatives. However, the power to detect any particular alternative in the parameter space will decrease as k increases. Conversely, if k is chosen to be small, the test will be quite directional, giving good protection against a small set of alternatives, and none against all others.

The idea behind data-driven smooth tests is to let the data make the decision about the order. See Ledwina (1994), Kallenberg and Ledwina (1995, 1997a,b) and Inglot et al. (1997). The order is chosen to optimize a specified criterion. If L_k is the likelihood of a random sample of size n from

a distribution that is an alternative of order k, it would be natural to maximize L_k by choice of k in some set, say $\{1, 2, \ldots, d\}$, where d is specified before sighting the data. Here d is the maximum order one is prepared to accept. However this procedure would simply choose $k = d$. A penalty term is needed to discourage complexity. The Bayesian information criterion (BIC)

$$\mathrm{BIC}_k = -2 \log L_k + k \log n$$

is appropriate, but the computationally easier modified BIC, $\widehat{S}_k + k \log n$ is often preferred. The optimal order, say K, is taken to be the smallest order that maximizes BIC_k. The test statistic is then chosen to be the score or generalized score statistic involving the first K components. This reduces to the sum of squares of the first K components or Cholesky components, as previously defined. An alternative option is to use Aikaike's information criterion (AIC)

$$\mathrm{AIC}_K = -2\log L_k + 2k$$

or its modified form $\widehat{S}_k + 2k$. BIC penalizes complex models more heavily than AIC. The point is that many different model selection rules are possible. In different circumstances different rules will be appropriate.

Since the order is no longer a predetermined constant but a random variable, the test statistic $\widehat{S}_K$ is no longer asymptotically χ^2 distributed, even when testing for distributions from exponential families. Although asymptotic distribution theory has been developed for some of the data-driven tests, critical values and p-values are best determined using resampling methods.

These data-driven smooth tests have competitive power. Inasmuch as they give protection against misspecifying the order, it would be unreasonable to expect them to have greater power than all smooth tests of specified order. If there is any vagueness in the class of alternatives one hopes to best detect, data-driven smooth tests are an excellent choice.

A more flexible method and one of wider applicability, but one which does not result in omnibus consistent tests, is given in Claeskens and Hjort (2004). They fix the maximal order, and they also consider subset selection. This means that for a given maximal order, say d, the model selection criterion can select any subset of indexes from $\{1, \ldots, d\}$ and the data-driven test statistic is then built from the corresponding components V_j with j in the selected index set.

Model Selection

Suppose now that the order k of a smooth alternative has been determined, and a smooth or generalized smooth test of this order applied. If this test statistic or indeed, any of its components is significant, what can be said about the true model? The moment interpretation of the components may not provide helpful insight. The insignificant θ_r in the order k alternative could be replaced by zero, significant θ_r by their method of moments estimators $\tilde{\theta}_r$, and β by its method of moments estimator under the null hypothesis, $\tilde{\beta}_0$. This approach seems intuitively reasonable, but are there better options?

In Rayner et al. (2009: Chapter 10) two approaches are described: model selection through hypothesis testing and model selection using model selection criteria. Here we focus on the first of these. First suppose $S_h = \{1, 2, \ldots, d\}$ is an index set called the *horizon*. We may use either Neyman or Barton models, but in iterative work it is far more convenient to work with the Barton smooth models

$$g_s^B(x; \theta_s, \beta) = \left\{1 + \sum_{i \in S} \theta_i h_i(x; \beta)\right\} f(x; \beta)$$

in which $S \subset S_h$, $\theta_S = \{\theta_i : i \in S\}$ and $\{h_i(x; \beta)\}$ is a set of functions orthonormal on $f(x; \beta)$. The nonpositivity of $g_S^B(x; \theta_S, \beta)$ is not an issue for score tests. Clearly $g_{S_h}^B(x; \theta_{S_h}, \beta)$ is the most complex model we are prepared to accept. A horizon of four clearly limits the maximal order more than one of 44 does.

We describe model selection through hypothesis testing, which is similar to the familiar forward selection and backward elimination techniques used in regression analysis. In forward selection at the uth step, the model is $g_S^B(x; \theta_S, \beta)$ and we consider whether or not to add a single θ_i term to the model, where $\theta_i \in \theta_{S_u}$, for every possible θ_i not already in the model. A slightly modified score test is derived to test each of these hypotheses. If any are significant at a predetermined level, then the θ_i corresponding to the most significant test is added to the model. Backward elimination is similar, with the least significant θ_i being eliminated from successively reduced models until only significant terms remain in the model. The score test statistics change at each iteration.

It has previously been argued that when a data-driven smooth test rejects the null hypothesis it may be informative to investigate the selected components, even though the diagnostic property may not be guaranteed. Another way forward, however, is to plot the density estimate that corresponds to the selected model. The Barton model is particularly convenient for this purpose. Suppose that $\widehat{S}$ contains the indexes of the selected components, and the appropriate nonparametric density estimate is given by

$$g_{\widehat{S}}^B(x; \tilde{\theta}, \tilde{\beta}_0) = \left\{1 + \sum_{i \in \widehat{S}} \tilde{\theta}_i h_i(x; \tilde{\beta}_0)\right\} f(x; \tilde{\beta}_0)$$

This density estimate is referred to as the *improved density estimate*, but it can only be considered as a genuine density estimate after correcting for the nonpositivity, using the methods proposed in Gajek (1986) and Glad et al. (2003). In Rayner et al. (2009: Chapter 10) we suggest using the plot of the improved density as the basis for the formulation of the conclusions. Since the improved density is merely a graphical representation of the information contained in the data-driven smooth test statistic, the conclusions will be consistent.

Examples and Conclusion

Example 3: Angus Data

Angus (1982) gives 20 operational lifetimes in hours:

6278, 3113, 5236, 11584, 12628, 7725, 8604, 14266, 6125, 9350, 3212, 9003, 3523, 12 888, 9460, 13 431, 17 809, 2812, 11 825, 2398.

In testing the null hypothesis that the data follow the exponential distribution, the Chernoff–Lehmann test with

four classes, the Anderson–Darling, Cramer-von Mises, and Shapiro–Wilk tests give bootstrap p-values 0.01, 0.01, 0.01 and 0.00 respectively. It would appear that the data are not consistent with exponentiality, as Angus (1982) also found. See Figure 1.

We also find $\widehat{S}_4 = \widehat{V}_2^2 + \ldots + \widehat{V}_5^2 = 8.8$ with χ_4^2 p-value is 0.07 and bootstrap p-value based on 1000 bootstrap runs of 0.02. In addition $\widehat{V}_2 = -1.7$, $\widehat{V}_3 = -1.9$, $\widehat{V}_4 = -1.4$, $\widehat{V}_5 = -0.8$. The corresponding asymptotic two-sided p-values are 0.10, 0.06, 0.16, and 0.44, with bootstrap p-values based on 1000 bootstrap runs 0.00, 0.00, 0.04, and 0.22, respectively. Here and elsewhere use of the asymptotic null distribution of the components and the test statistic itself can be misleading. In general it is more desirable to use bootstrap p-values.

The first two components contribute most to $\widehat{S}_4$, suggesting at least dispersion and skewness departures from what would be expected under the exponential model. However, the cause of the significance of the order two components could be due to the third and fourth moments of the true distribution differing from those of the exponential; both corresponding components have p-values less than 5%. Similarly the significance of the order three component could be due to moments of the true distribution up to the sixth, and at this point the order six component has not been assessed.

Now suppose a horizon of $\{2, \ldots, 6\}$ is taken; order one is omitted because the rate is a nuisance parameter and fills the same role as θ_1. In forward selection, at the first step the test statistics corresponding to θ_2 to θ_5 are all significant at the 5% level, but that corresponding to θ_3 is most significant, so that term is included in the model, and this new model assessed. This next assessment finds the term corresponding to θ_4 should be included in the model. A new assessment shows that no further terms are significant, and assessment stops.

In backward elimination the initial model includes θ_2 to θ_6 and successively eliminates first θ_6, then θ_5, then θ_2, and then finds that no further terms warrant removal. Although it need not be the case in general, both forward selection and backward elimination agree here.

Figure 1 shows the hypothesized and the improved density estimates. From this it may be concluded that the true density is probably not a monotonically decreasing density such as the hypothesized exponential density. The improved density suggests that there is a mode close to five and that the distribution is skewed with a tail toward the larger lifetimes.

Statistical model assessment is at the heart of good statistical practice, and is the genesis of modern statistics. In general we suggest using the Anderson–Darling test as an omnibus test. For a more detailed scrutiny of the data, we recommend testing using the smooth or generalized smooth tests with asymptotically uncorrelated components. These tests generally have competitive power and the asymptotically uncorrelated components facilitate exploratory data analysis. In exploratory data analysis, a significant rth component suggests nonzero θ_r and hence the data and the null model differ in moments up to the rth. However with arbitrary alternatives moments up to the $2r$th may be the cause of the model failure. Model selection techniques may be used to construct smooth models consistent with the data. It is more informative and accurate to base exploratory data analysis on these models and their graphical representations than on the

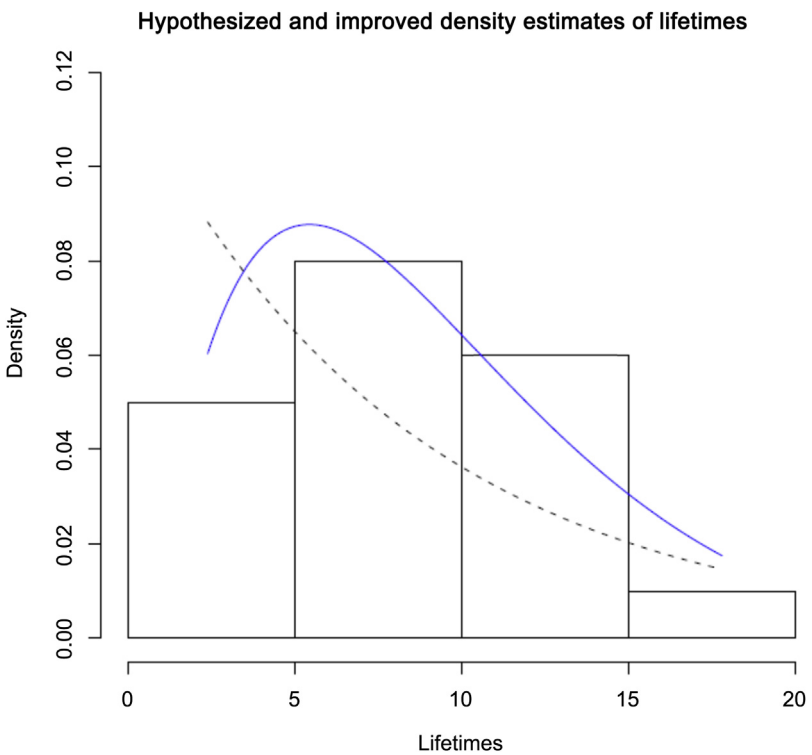

Figure 1 Histogram, density estimates of the hypothesized exponential distribution (dashed line) and the improved density (full line) for the Angus lifetimes data. Lifetimes are given in thousands of hours. The counts in the four classes are 5, 8, 6, and 1.

components themselves. However, these techniques may use many criteria, so there are many possible density plots. Each selected model has implications about which moments may be the cause of model failure, but it is important to remember that this is exploratory data analysis, and perhaps it is the envelope of possible densities that gives insights rather than any particular choice.

See also: Distributions, Statistical: Special and Discrete; Estimation: Point and Interval; Hypothesis Testing in Statistics; Model Selection and Model Averaging.

Bibliography

Anderson, T.W., Darling, D.A., 1954. A test of goodness of fit. Annals Mathematical Statistics 23, 193–212.

Angus, J.E., 1982. Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. Journal of Statistical Planning & Inference 6, 241–251.

Barton, D.E., 1953. On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. Skandinavisk Aktuarietidskr 36, 24–63.

Barton, D.E., 1955. A form of Neyman's Ψ^2 test of goodness of fit applicable to grouped and discrete data. Skandinavisk Aktuarietidskr 38, 1–16.

Barton, D.E., 1956. Neyman's Ψ^2 test of goodness of fit when the null hypothesis is composite. Skandinavisk Aktuarietidskr 39, 216–245.

Best, D.J., Rayner, J.C.W., 2007. Chi-squared components as tests of fit for the grouped exponential distribution. Computational Statistics & Data Analysis 51, 3946–3954.

Best, D.J., Rayner, J.C.W., Thas, O., 2008. X^2 and its components as tests of normality for grouped data. Journal of Applied Statistics 35 (5), 481–492.

Chernoff, H., Lehmann, E.L., 1954. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Annals Mathematical Statistics 25, 579–586.

Choulakian, V., Lockhart, R.A., Stephens, M.A., 1994. Cramer-Von Mises statistics for discrete distributions. Canadian Journal of Statistics 22, 125–137.

Claeskens, G., Hjort, N., 2004. Goodness of fit via non-parametric likelihood ratios. Scandinavian Journal of Statistics 31, 487–513.

D'Agostino, R.B., Stephens, M.A., 1986. Goodness of Fit Techniques. Marcel Dekker, New York.

De Boeck, B., Thas, O., Rayner, J.C.W., Best, D.J., 2011. Smooth tests for the gamma distribution. Journal of Statistical Computation & Simulation 81, 843–855.

Gajek, G., 1986. On improving density estimators which are not bona fide functions. Annals of Statistics 14, 1612–1618.

Glad, I., Hjort, N., Ushakov, N., 2003. Correction of density estimators that are not densities. Scandinavian Journal of Statististics 30, 415–427.

Hamdan, M.A., 1964. A smooth test of goodness of fit based on the Walsh functions. Australian Journal of Statistics 6, 130–136.

Henze, N., 1997. Do components of smooth tests of fit have diagnostic properties? Metrika 45, 121–130.

Henze, N., Klar, B., 1996. Properly rescaled components of smooth tests of fit are diagnostic. Australian & New Zealand Journal of Statistics 38, 61–74.

Henze, N., Meintanis, S., 2002. Tests for exponentiality based on the empirical Laplace transform. Statistics 36, 147–161.

Henze, N., Meintanis, S., 2012. Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. Communications in Statistics-Theory and Methods 41, 1543–1556.

Inglot, T., Kallenberg, W., Ledwina, T., 1997. Data driven smooth tests for composite hypotheses. Annals of Statistics 25, 1222–1250.

Kallenberg, W., Ledwina, T., 1995. Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. Annals of Statistics 23, 1594–1608.

Kallenberg, W., Ledwina, T., 1997a. Data driven smooth tests for composite hypotheses: comparison of powers. Journal of Statistical Computation and Simulation 59, 101–121.

Kallenberg, W., Ledwina, T., 1997b. Data driven smooth tests when the hypothesis is composite. Journal of the American Statistical Association 92, 1094–1104.

Klar, B., 2000. Diagnostic smooth tests of fit. Metrika 52, 237–252.

Kolmogorov, A.N., 1933. Sulla determinazione empirica di una legge di sitribuzione. Giornale dell' Instituto Italiano desli Attuari 4, 83–91.

Kopecky, K.J., Pierce, D.A., 1979. Efficiency of smooth goodness-of-fit tests. Journal of the American Statistical Association 74, 393–397.

LaRiccia, V.N., 1991. Smooth goodness of fit tests: a quantile function approach. Journal of the American Statistical Association 86, 427–431.

Ledwina, T., 1994. Data driven version of Neyman's smooth test of fit. Journal of the American Statistical Association 89, 1000–1005.

Neyman, J., 1937. 'Smooth' test for goodness of fit. Skandinavisk Aktuarietidskr 20, 150–199.

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. Philosophy Magazine, 5th Series 50, 157–175.

Plackett, R.L., 1983. Pearson Karl and the chi-squared test. International Statistical Review 51, 59–72.

Rao, K.C., Robson, D.S., 1974. A chi-square statistic for goodness-of-fit tests with the exponential family. Communication in Statistics 3, 1139–1153.

Rayner, J.C.W., Best, D.J., 1990. Smooth tests of goodness of fit: an overview. International Statistical Review 58 (1), 9–17.

Rayner, J.C.W., Thas, O., De Boeck, B., 2008. A generalised Emerson recurrence relation. Australian and NZ Journal of Statistics 50 (3), 235–240.

Rayner, J.C.W., Thas, O., Best, D.J., 2009. Smooth Tests of Goodness of Fit: Using R, second ed. Wiley, Singapore.

Rayner, J.C.W., Thas, O., Best, D.J., September/October 2011. Smooth tests of goodness of fit. Wiley Interdisciplinary Reviews: Computational Statistics 3 (5), 97–406. http://dx.doi.org/10.1002/wics.171.

Rayner, J.C.W., Thas, O., Pipelers, Peter, Beh, E.J., 2013. Calculating bivariate orthonormal polynomials by recurrence. Australian & NZ Journal of Statistics 55 (1), 15–24.

Read, T.R.C., Cressie, N.A.C., 1988. Statistics for Discrete Multivariate Data. Springer-Verlag, New York.

Rippon, P., Rayner, J.C.W., 2010. Generalised score and Wald tests. Advances in Decision Sciences Article ID 292013, 8. http://dx.doi.org/10.1155/2010/292013.

Spinelli, J.J., Stephens, M.A., 1997. Cramer-Von Mises tests of fit for the Poisson distribution. Canadian Journal of Statistics 25, 257–268.

Stephens, M.A., 1974. Components of goodness of fit statistics. Annals of the Institute of Henri Poincare 10, 37–54.

Thas, O., 2010. Comparing Distributions. Springer, New York.

Thas, O., Rayner, J.C.W., Best, D.J., De Boeck, B., 2009. Informative statistical analyses using smooth goodness of fit tests. Journal of Statistical Theory and Practice 3, 705–733.

Thomas, D.R., Pierce, D.A., 1979. Neyman's smooth goodness-of-fit test when the hypothesis is composite. Journal of the American Statistical Association 74, 441–445.

Statistical Identification and Estimability

Gordon M Kaufman, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15025–15031, © 2001, Elsevier Ltd.

Abstract

The problem of statistical identification appears in different guises in many diverse fields. The most interesting form of the problem occurs when a sample of infinite size, assumed to be generated by one member of a class of statistical models indexed by a parameter, does not lead to certain knowledge of one or more elements of that parameter. A different type of identification problem arises when sample size is less than the number of elements of the parameter indexing the class of statistical models assumed to have generated the sample. For example, in the case of ordinary multiple regression, when sample size is less than the number of independent (explanatory) variables, it is not possible to estimate regression coefficients uniquely, even in the absence of collinearity among vectors of independent variable observations. This article focuses on the former type of identification problem and takes a broad view of what is often considered a narrow problem particular to one's given field of interest. Following a brief history of statistical identification, the concept of 'identifiability' is examined and explained from both Bayesian and non-Bayesian viewpoints.

A statistical identification problem arises when a sample of infinite size, assumed to be generated by an element of a class of statistical models indexed by a parameter, does not lead to certain knowledge of one or more components of that parameter. Roughly, a model is identified if enough observed data could allow one to distinguish elements of the parameter set from one another.

Statistical identification problems are ubiquitous, occurring in fields as diverse as reliability, optimal control theory, experimental design, and epidemiology. A rich array of statistical identification problems appear in the social sciences: latent structure analysis, factor analysis, and analysis of covariance structures are based on multivariate statistical models that possess, by construction, this problem. Nonresponse and data censoring in the conduct of sample surveys often lead to an identification problem and the problem is a central feature of the analysis of econometric structural equation models and of errors-in-variables regression models. Experimental designs in which treatments are confounded lead to an identification problem. A nonparametric identification problem occurs in the theory of competing risks.

Understanding if a statistical model is identified, nonidentified, or unidentified is a necessary prelude to the construction of a sensible method for making inferences about model parameters from observed data.

Background

The problem of statistical identifiability appears in economics as early as Pigou (1910). Thurstone discusses identification in his applications of factor analysis to intelligence (1935, 1947) and, in his latent structure analysis of opinion research and attitudes, Lazarsfeld deals with identification issues (1950).

Principal contributions to a rigorous mathematical treatment of the problem began in the 1950s with the work of Haavelmo (1944), Hurwicz (1950), Koopmans et al. (1950), and Koopmans and Reiersøl (1950) and spawned a huge literature on the subject. Fisher (1966) provides a thorough rigorous treatment of identification for simultaneous equation systems. He laid to rest the issue of whether or not identifiability is a discrete or continuous property of parameter estimation by demonstrating that estimators, which are consistent given a correct model specification, converge in probability to true parameter values as specification error approaches zero. Thus, identifiability may be regarded as a continuous property of a set of parameters, and small specification errors lead to small inconsistencies in parameter estimation. Rothenberg (1971) and Bowden (1973) extend identification theory to embrace nonlinear models, distinguishing local from global identifiability. Hannan (1969, 1971) generalizes the theory of identification for linear structural equation systems with serially uncorrelated residual errors in another direction: he establishes conditions for identifiability for the important case of autoregressive moving average error terms (*see* Time Series: General).

Probabilistic models employed in linear system optimal control theory have much in common with econometric structural (simultaneous) equation systems in which observations are measured with random error. In the psychometric literature, the later class of models goes under the acronym LISREL (Linear Structural Relationships). Jöreskog and Sörbom (1984) have developed methods for specifying LISREL model structure, algorithms for calculating various types of parameter estimators, and a framework for interpretation of results. As with econometric structural equation systems, applications of LISREL span a wide domain of social science applications, marketing, psychology, sociology, and organizational behavior among them.

A Bayesian treatment of identification began with Dawid (1979) and Kadane's (1974) studies of identification in Bayesian theory and Drèze's (1972), (1974) analysis of econometric simultaneous equation systems from a Bayesian perspective. Identification through a Bayesian lens is currently

a vigorous area of research yielding new insights into how to construct meaningful prior distributions for parameters.

Example 1

A sample survey is conducted to determine the attitudes of individuals in a population toward a particular subject. The sample frame consists of N individuals, each of whose response is classified as either 'Favorable' or 'Unfavorable.' This sample yields $N_{R\cdot}$ respondents and $N_{S\cdot} = N - N_{R\cdot}$ nonrespondents. Of the $N_{R\cdot}$ respondents, N_{FR} are classified as 'Favorable' and $N_{UR} = N - N_{FR}$ are classified as 'Unfavorable.' In the absence of information bearing directly on the fraction of nonrespondents who would be classified as 'Favorable,' statistics available for inference are N_{FR}, N_{UR}, $N_{S\cdot}$ subject to $N_{FS} + N_{US} = N_{S\cdot}$ and to $N_{FR} + N_{UR} + N_{S\cdot} = N$. Suppose that the process generating data about both respondents and nonrespondents is modeled as a double dichotomy with probabilities p_{FR}, p_{UR}, p_{FS}, and $p_{US} > 0$ and $p_{FR} + p_{UR} + p_{FS} + p_{US} = 1$ as shown in Table 1. Table 2 shows the corresponding tables for the model of observed data.

The probability distribution for observations N_{FR}, N_{UR}, and $N_{S\cdot}$ in Table 1 is proportional to

$$(p_{FS} + p_{US})^{N_{S\cdot}} p_{FR}^{N_{FR}} p_{UR}^{N_{UR}}$$

Observations N_{FR}, N_{UR}, and $N_{S\cdot}$ are sufficient for joint inference about $p_{S\cdot} = p_{FS} + p_{US}$, p_{FS}, and p_{UR}, but even an infinite sample will not lead to knowledge of p_{FS} and p_{US} with certainty: p_{FS} and p_{US} are not identifiable in the presence of nonresponse as an infinite sample yields knowledge only that p_{FS} and p_{US} satisfy $p_{FS} + p_{US} = p_{S\cdot\cdot}$.

Definitions

A statistical 'model' is a family $\{P_\theta, \theta \in \Theta\}$ of probability distributions indexed by a parameter $\theta \in \Theta$. A statistical 'structure' is P_θ for a particular $\theta = \theta_0 \in \Theta$. Parameters θ and $\theta' \in \Theta$ are 'observationally equivalent' if and only if $P_{\theta'}(A) = P_{\theta'}(A)$ for all A in a measurable space $\mathscr{L}$.

Table 1

	Parameters			Observations		
Favorable (F)	p_{FR}	p_{FS}	$p_{F\cdot}$	N_{FR}	N_{FS}	$N_{F\cdot}$
Unfavorable (U)	p_{UR}	p_{US}	$p_{U\cdot}$	N_{UR}	N_{US}	$N_{U\cdot}$
	$p_{R\cdot}$	$p_{S\cdot}$	1.0	$N_{R\cdot}$	$N_{S\cdot}$	N

Table 2

	Parameters		Observations		
Favorable (F)	p_{FR}		N_{FR}		
Unfavorable (U)	p_{UR}		N_{UR}		
	$p_{R\cdot}$	$p_{S\cdot}$	$N_{R\cdot}$	$N_{S\cdot}$	N

Definition: Θ is 'identifiable' if and only if $\theta = \theta'$ for all observationally equivalent pairs $\theta, \theta' \in \Theta$.

In practice, it is useful to couch these definitions in terms of distribution functions: let X be an observable random variable with range $\mathscr{L}_x$ and distribution function $F_\theta(x) \equiv P(X \le x|\theta)$ belonging to a model $\{F_\theta, \theta \in \Theta\}$. If there exists at least one pair (θ, θ') for which $\theta, \theta' \in \Theta$, and $\theta \ne \theta'$ such that $F_\theta(x) = F_{\theta'}(x)$ for all $x \in \mathscr{L}_x$, then we say that Θ is 'nonidentifiable' by X. If Θ is not nonidentifiable by X it is identifiable. These definitions of identifiability can be recast to embrace nonparametric models as well (Basu, 1981; Berman, 1963).

Example 2

In general, unobservables in an 'econometric simultaneous equation model' are restricted to residuals (errors in model equations). A 'factor analysis model,' on the other hand, possesses both unobservable latent and unobservable explanatory variables. Each is a special case of a more general 'structural equation model' in which a 'latent variable model' displaying interactions among latent variables and a 'measurement model' specifying how observable variables are related to latent variables are conjoined. Properties of the covariance matrix of observable variables are at center stage and structural equation test procedures are designed to measure the quality of fit between an observed (sample) covariance matrix Σ and a model based covariance matrix $\Sigma(\theta)$ indexed by a vector θ of model parameters.

Even when the structure of relations between observable variables and latent variables and the structure of relations among latent variables are both linear, the model covariance matrix $\Sigma(\theta)$ may be a complicated, nonlinear function of elements of the parameter vector θ. In this setting, for an element of the parameter vector θ to be identified it must be an explicit, known function of one or more elements of the sample covariance matrix Σ under the hypothesis that $\Sigma = \Sigma(\theta)$. (See Bollen (1989) for a comprehensive treatment and also Hoyle (1995) for applications.) If PDS $\Sigma(\theta)$ is $(k \times k)$, it has at most $\frac{1}{2}k(k+1)$ functionally independent components, so for θ to be identifiable, it can possess at most $\frac{1}{2}k(k+1)$ functionally independent components. This is a generic, easy-to-apply, necessary but not sufficient condition for identifiability of θ.

In the notation of Bollen (1989), we have a latent variable model and a measurement model.

Latent Variable Model

$$\eta = \mathbf{B}^*\eta + \Gamma\xi + \zeta$$

The $(m \times m)$ coefficient matrix $\mathbf{B}^*$ induces interactions among latent endogenous variables $(m \times 1)\eta$, the $(m \times r)$ coefficient matrix Γ establishes the effect of exogenous variables $(r \times 1)\xi$ on η, and $(m \times 1)\zeta$ represents a random residual error term assumed to be uncorrelated with ξ and to possess mean zero and finite (PDS) variance matrix Ψ. In addition $\mathbf{B} \equiv \mathbf{I} - \mathbf{B}^*$ is assumed to be nonsingular.

Measurement Model

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$
$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

The elements of the $(p \times m)$ coefficient matrix $\mathbf{\Lambda}_y$ determine the magnitude of the effect of levels and changes in levels of latent variables $\boldsymbol{\eta}$ on observable variables $\mathbf{y}$; $(q \times r)\mathbf{\Lambda}_x$ relates exogenous variables $\boldsymbol{\xi}$ to observable variables in $\mathbf{x}$ in a similar fashion. Residual errors $(p \times 1)\boldsymbol{\varepsilon}$ and $(q \times 1)\boldsymbol{\delta}$ are random variables, generally assumed to be uncorrelated with one another, with $\boldsymbol{\eta}$, and with $\boldsymbol{\xi}$ and to possess common mean zero, but possibly different (PDS) covariance matrices $\mathbf{\Theta}_{\varepsilon}$ and $\mathbf{\Theta}_{\delta}$, respectively.

To be specific, focus on the composition of the variance matrix $\mathrm{Var}(\mathbf{y})$ of $\mathbf{y}$ as a function of latent variable model parameters $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$ and measurement model parameters $\mathbf{\Lambda}_y$ and $\mathbf{\Theta}_{\varepsilon}$:

$$\mathrm{Var}(\mathbf{y}) = \mathbf{\Lambda}_y \mathbf{B}^{-1}\left[\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^t + \mathbf{\Psi}\right]\left(\mathbf{B}^{-1}\right)^t \mathbf{\Lambda}_y^t + \mathbf{\Theta}_{\varepsilon}$$

The $(p \times p)$ PDS matrix $\mathrm{Var}(\mathbf{y})$ possesses $\frac{1}{2}p(p+1)$ functionally independent components. The variance matrix $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{\Theta}_{\varepsilon}$ alone also possesses $\frac{1}{2}p(p+1)$ functionally independent components and, in the absence of restrictions, other than just $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are PDS, $\mathbf{\Lambda}_y$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$ possess $\frac{1}{2}m(m+1) + \frac{1}{2}r(r+1) + m(m+p+r)$ additional functionally independent components. As the above equation suggests, it can be a formidable task to verify by direct algebraic analysis whether or not a particular model parameter is, or is not identifiable.

Clearly, parameter matrices $\mathbf{\Lambda}_y$, $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Phi}$, and $\mathbf{\Theta}_{\varepsilon}$ cannot be expressed as unique functions of elements of $\mathrm{Var}(\mathbf{y})$ unless restrictions are placed upon them. When one or more components of $\mathbf{B}$ cannot be recovered uniquely as a function of elements of $\mathrm{Var}(\mathbf{y})$, such components are called 'underidentified' or 'nonidentified.' A central task then is to select constraints on latent variable model parameters and measurement model parameters that renders desired parameters 'identifiable.'

If a priori restrictions on parameters of latent variable and measurement model parameters are specified in a way that leads to two or more distinct representations of some components of $\mathbf{B}$ as a function of components of $\mathrm{Var}(\mathbf{y})$, those components of $\mathbf{B}$ are called 'overidentified.' It can happen, of course, that some elements of $\mathbf{B}$ are identified, some are underidentified, and others are overidentified. Bartels (1985) provides an easily accessible treatment of these concepts.

A common form of a priori constraint is to 'specify' that some parameter values are equal to zero. For example, if the modeler assumes that $\mathbf{y}$ and $\mathbf{x}$ are measured without error and that $\mathbf{\Lambda}_y = (m \times m)\mathbf{I}$, $\mathbf{\Lambda}_x = (r \times r)\mathbf{I}$, $\mathrm{Var}(\boldsymbol{\eta}) = \mathrm{Var}(\boldsymbol{\xi}) = \mathbf{\Theta}_{\varepsilon} = \mathbf{\Theta}_{\delta} = 0$ and the variance of $\mathbf{y}$ becomes

$$\mathrm{Var}(\mathbf{y}) = \mathbf{B}^{-1}\left[\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^t + \mathbf{\Psi}\right]\left(\mathbf{B}^{-1}\right)^t.$$

If, in turn, as is often the case in econometric applications, exogenous variables $\mathbf{x}$ are assumed to be predetermined and known with certainty, $\mathbf{\Phi} = \mathbf{0}$. With all these restrictions in force $\boldsymbol{\eta} = \mathbf{y}$, $\boldsymbol{\xi} = \mathbf{x}$, and the latent variable model becomes $\mathbf{By} = \mathbf{\Gamma z} + \boldsymbol{\delta}$. Then, observable $\mathbf{y} = \mathbf{\Pi z} + \mathbf{u}$, with $\mathbf{\Pi} \equiv -\mathbf{B}^{-1}\mathbf{\Gamma}$, $\mathbf{u} = \mathbf{B}^{-1}\boldsymbol{\delta}$, and $\mathrm{Var}(\mathbf{y}) = \mathbf{B}^{-1}\mathbf{\Psi}(\mathbf{B}^{-1})^t \equiv \mathbf{\Omega}$, is called the 'reduced form' of this model.

Provided that predetermined variables $\mathbf{z}$ are uncorrelated with variables $\mathbf{u}$ (weaker assumptions are possible) and that the asymptotic variance matrix of the $\mathbf{z}$'s exists, both $\mathbf{\Pi}$ and $\mathbf{\Omega}$ can be consistently estimated by ordinary least squares. However, the distribution of $\mathbf{y}$, given $\mathbf{z}$, now depends exclusively on parameters $\mathbf{\Pi}$ and $\mathbf{\Omega}$. Premultiplication of both sides of $\mathbf{By} = \mathbf{\Gamma z} + \boldsymbol{\delta}$ by any nonsingular $(m \times m)$ matrix leads to the same reduced form, so all such structures are observationally equivalent and without further restrictions the system is underidentified.

Fisher (1966) provides an elegant and thorough exposition of conditions for identifiability of linear simultaneous equation system model parameters. Among the most important are 'order' and 'rank' conditions established by Koopmans et al. (1950). Suppose interest focuses on a single equation and that exclusion restrictions alone on $\mathbf{y}$ and $\mathbf{z}$ are in force. A (necessary but not sufficient) 'order condition' for parameters of this equation to be identifiable is that the number of variables excluded be greater than or equal to $m - 1$ – one less than the number of components in (endogenous) $\mathbf{y}$. Alternatively, the total number of excluded predetermined variables must be greater than or equal to the number of variables not excluded from this equation, less one (Fisher, 1966: 40–41).

The following argument due to Fisher is a transparent explanation of the origin of the 'rank condition' for linear simultaneous equation systems, first proved by Koopmans et al. (1950). Exclusion of the lth element of the first row $\mathbf{a}_1$ of $m \times (m+r)\mathbf{A} \equiv [\mathbf{B}, \mathbf{\Gamma}]$ is representable as $\mathbf{a}_1\mathbf{c}^{(l)} = 0$, $\mathbf{c}^{(l)}$ an $(m+r) \times 1$ column vector with a one at element l and zeros elsewhere. Thus, k such exclusions of distinct variables may be written in terms of $(m+r) \times k$ $\mathbf{C}$ as $\mathbf{a}_1\mathbf{C} = \mathbf{0}$. Clearly if $\mathbf{a}_1$ is some linear combination of other rows of $\mathbf{A}$, it is not identifiable. Thus, a sufficient condition for $\mathbf{a}_1$ to be identified is that $\mathbf{xAC} = \mathbf{0}$ holds only for $(1 \times m)\mathbf{x}$ of the form $= (\alpha, 0, \ldots, 0)$, α a scalar. In the absence of other a priori restrictions on elements of $\mathbf{a}_1$, this condition is also necessary. When it obtains, the rank of $\mathbf{AC}$ must be $m - 1$. The order condition follows from the observation that the rank of $\mathbf{AC}$ must be less than or equal to that of $\mathbf{C}$, so the number of independent restrictions imposed by $\mathbf{C}$ must be at least $m - 1$.

Estimation

When a linear statistical model is just identified, both consistent and unbiased estimators of its parameters exist. A more general notion of estimability for linear statistical models, first proposed by Bose (1944, 1947), says that a parameter $\boldsymbol{\theta}$ of such a model is (linearly) estimable if there exists a linear function of observations with expectation equal to $\boldsymbol{\theta}$ for each $\boldsymbol{\theta}$ in a prespecified set $\mathbf{\Theta}$. This special case of identifiability, apparently developed independently of the more general theory of identification outlined in this review, has generated a large literature of its own (*see* Factor Analysis and Latent Variable Models in Personality Psychology).

Deistler and Seifert (1978) address the interplay between identification and estimability for a quite general (possibly nonlinear) version of an econometric structural equation model without measurement error. They answer the question "When, for such systems, does identifiability imply existence of consistent estimators?"

Jöreskog (1973) and Wiley (1973) were among the first to specify practical procedures for the estimation of parameters by analysis of the covariance structure of observations generated by a general structural equation model of the form described in Example 2. For a detailed description of parameter estimation via maximum likelihood, instrumental variables, and least squares see, for example, Jöreskog and Sörbom (1984).

Simultaneous Equation Estimates (Exact and Approximate), Distribution of is a crisp summary of methods designed to recover consistent estimates of structural equation parameters, such as Full Information Maximum Likelihood (FIML), Limited Information Maximum Likelihood (LIML), R-class estimators, and two- and three-stage least squares (2 SLS, 3 SLS). A more leisurely presentation of econometric simultaneous equation system identification issues, couched in terms of a two-equation supply and demand model, and the nature of estimation problems they pose, appears in Simultaneous Equation Estimation: Overview.

Bayesian Identification

Non-Bayesians typically assign exact numerical values to a subset of parameters of a statistical model in order to achieve identification of one or more other parameters. Fisher (1961) suggested broadening this special type of a priori specification by replacing exact numerical values with probabilistic constraints. Drèze (1962) took up the challenge in his treatment of Bayesian identification for simultaneous equation estimation models. As before, let $\mathbf{By} + \mathbf{\Gamma z} = \mathbf{u}$ be a 'structural equation system' indexed by matrix parameters consisting of nonsingular $\mathbf{B}$, $\mathbf{\Gamma}$, and PDS $\mathrm{Var}(\mathbf{u}) = \sum$ with reduced form $\mathbf{y} = \mathbf{\Pi z} + \mathbf{v}$, $\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{\Gamma}$ and disturbance variance $\mathrm{Var}(\mathbf{v}) = \mathbf{B}^{-1}\sum(\mathbf{B}^{-1})^t \equiv \mathbf{\Omega}$. Because the likelihood function for observations $\mathbf{y}$ can be written as a function of reduced form parameters $\mathbf{\Pi}$ and $\mathbf{\Omega}$ alone, full knowledge of the exact numerical values of elements of $\mathbf{\Pi}$ and $\mathbf{\Omega}$ does not permit recovery of unique values for elements of $\mathbf{B}$, $\mathbf{\Gamma}$, and $\sum$. The map from $\mathbf{\Pi}$ and $\mathbf{\Omega}$ to $\mathbf{B}$, $\mathbf{\Gamma}$, and $\sum$ is one to many. In effect, all nonsingular matrices $\mathbf{B}$ are observationally equivalent and $\mathbf{B}$ is only identifiable in the special case when $f(\mathbf{B}|\mathbf{\Pi}, \mathbf{\Omega})$ is concentrated on a single point in the space of nonsingular matrices $\mathbf{B}$. Nevertheless, one can assign a prior density $g(\mathbf{B}, \mathbf{\Gamma}, \sum)$ to structural parameters, observe data $(\mathbf{Y}, \mathbf{Z})$ and compute a posterior density $f(\mathbf{B}, \mathbf{\Gamma}, \sum|\mathbf{Y}, \mathbf{Z})$ for $\mathbf{B}$, $\mathbf{\Gamma}$, and $\sum$ whether or not the model is identified in a classical sense. Because identifiability is a property of the likelihood function alone, if a proper prior distribution is assigned to a complete set of parameters ($\mathbf{B}$, $\mathbf{\Pi}$, and $\mathbf{\Omega}$ here) the posterior for these parameters is proper whether or not the likelihood function is identified. Such priors must be chosen judiciously: too tight a prior will anchor the posterior on the prior until sample size becomes very large; a very loose proper prior may fail to reflect relevant prior knowledge. To streamline computation of posteriors, Raiffa and Schlaifer (1961) introduced the concept of 'natural conjugate' priors. The functional form of a natural conjugate prior is derived by interpreting parameter as sufficient statistic and vice versa in the likelihood function. A consequence is that both prior and posterior possess the same functional form, a form dictated by the shape of the likelihood function. For simultaneous equation systems, multivariate-, matric- and poly-t posterior distributions for parameters follow and figure prominently in work by Drèze (1976), Drèze and Morales (1976), Drèze and Richard (1983), and Mouchard (1976) (*see* Bayesian Statistics).

To minimize the impact of a priori evidence or, in some cases, to exploit invariance with respect to mapping from one representation of a parameter to another, a diffuse (vague, noninformative) prior can be the prior of choice. However, in some settings, this choice comes at a cost as a diffuse prior may lead to an improper posterior. Lindley and El Sayeed (1968) foreshadow much recent work on the character of diffuse priors. They were perhaps the first to recognize several important features of prior to posterior analysis. First, a proper prior always yields a proper posterior. Second, "…in the functional relationship problem [improper priors] can cause considerable difficulties…" Third, even when the likelihood function is not identifiable, observed data can lead to a posterior that is informative, i.e., different from the prior. Fourth, "uncritical use of improper distributions can lead to inconsistent estimates."

A simple example often used to introduce the idea of nonidentifiability highlights issues raised by diffuse priors.

Example 3

Suppose observations to be independent and identically normal with mean $\boldsymbol{\theta} = \mu_1 + \mu_2$ and variance 1. Then μ_1 and μ_2 are nonidentifiable. Nevertheless, assignment of a proper prior to μ_1 and μ_2 leads to a proper posterior. As the number of observations approaches infinity, the posterior for $\boldsymbol{\theta}$ converges in distribution to a true value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ and the (proper) posterior for μ_1 and μ_2 concentrates on the line $\boldsymbol{\theta}_0 = \mu_1 + \mu_2$.

If, however, a diffuse prior $d\mu_1$, $-\infty < \mu_1 < \infty$, is assigned to μ_1 and proper prior $g(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$, the posterior for μ_1 remains diffuse irrespective of sample size. Replace the proper prior $g(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ with a proper prior $h(\mu_2)$ for μ_2. Then, provided the prior for μ_1 is diffuse, the posterior for μ_2 remains the same as the prior $h(\mu_2)$, no matter how many observations are made.

These features of Example 3 suggest that uncritical choice of a diffuse prior can lead to a posterior with unsatisfactory or even unacceptable properties, in particular when the parameter space is of large dimension. For example, Kleibergen and Van Dijk (1997) point out that for simultaneous equation systems "… the order condition reflects overall identification while the rank condition reflects local (non) identification." They show that local nonidentification coupled with a diffuse prior leads to posterior pathology when one attempts Limited Information analysis of a single equation and suggest an

alternative approach. (See Gelfand and Sahu (1999) for other examples.)

Dawid (1979) and Kadane (1974) use conditional independence as a device for a definition of nonidentifiability which embraces both Bayesian and non-Bayesian perspectives. Suppose that an observable X possesses a distribution $F_{\boldsymbol{\theta}}$ determined by a parameter $\boldsymbol{\theta}$ which we partition into two pieces $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. If the distribution of $\boldsymbol{X}$ is fully determined by $\boldsymbol{\theta}_1$ alone, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is nonidentifiable.

When a Bayesian assigns a prior distribution $g(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$, Bayes theorem tells her that the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ for a parameter $\boldsymbol{\theta}$ given data $\mathbf{y}$ is proportional to the product of the likelihood function $\mathbf{L}(\boldsymbol{\theta}|\mathbf{y})$ for $\boldsymbol{\theta}$ given $\mathbf{y}$ and the prior $g(\boldsymbol{\theta})$ assigned to $\boldsymbol{\theta}$. The (posterior) distribution for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ and data $\mathbf{y}$ given prior $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1)$ is

$$f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\ \mathbf{y}) \propto \mathbf{L}(\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2|\mathbf{y})g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1)$$

If $\boldsymbol{\theta}_2$ is absent from the likelihood function,

$$f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\ \mathbf{y}) \propto g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1|\mathbf{y}) \propto g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$$

so that while data will modify her prior judgements about $\boldsymbol{\theta}_1$, no amount of observable data $\mathbf{y}$ will change her a priori probability assessment $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$. Data $\mathbf{y}$ is, however, unconditionally informative about $\boldsymbol{\theta}_2$; as data accrues and we learn more about $\boldsymbol{\theta}_1$, prior judgements about $\boldsymbol{\theta}_2$ are modified in accord with

$$f(\boldsymbol{\theta}_2|\mathbf{y}) \propto \int g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)h(\boldsymbol{\theta}_1|\mathbf{y})\mathrm{d}\boldsymbol{\theta}_1$$

provided that both $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ and $h(\boldsymbol{\theta}_1|\mathbf{y})$ are proper distributions.

Return to Example 1. As Bayesians, (noting that $p_{\mathrm{FR}} + p_{\mathrm{UR}} + p_{\mathrm{FS}} + p_{\mathrm{US}} = 1$), we assign a prior distribution $g(p_{\mathrm{FR}}, p_{\mathrm{UR}}, p_{\mathrm{FS}})$ to $p_{\mathrm{FR}}, p_{\mathrm{UR}}, p_{\mathrm{FS}}$. If observed data $\mathbf{y} = (N_{\mathrm{S}\cdot}, N_{\mathrm{FR}}, N_{\mathrm{UR}})$ is generated in accord with Table 2, the likelihood function for $p_{\mathrm{FR}}, p_{\mathrm{UR}}, p_{\mathrm{FS}}$,

$$[1 - p_{\mathrm{FR}} - p_{\mathrm{UR}}]^{N_{\mathrm{S}\cdot}} p_{\mathrm{FR}}^{N_{\mathrm{FR}}} p_{\mathrm{UR}}^{N_{\mathrm{UR}}}$$

does not depend on $p_{\mathrm{FS}\cdot}$ Consequently,

$$f(p_{\mathrm{FS}}|p_{\mathrm{FR}}, p_{\mathrm{UR}}, \mathbf{y}) \propto g(p_{\mathrm{FS}}|p_{\mathrm{FR}}, p_{\mathrm{UR}})$$

This matches our intuition: given $p_{\mathrm{FR}} + p_{\mathrm{UR}} = \alpha$ we then know with certainty that $p_{\mathrm{FS}} + p_{\mathrm{US}} = 1-\alpha$, $p_{\mathrm{FS}}, p_{\mathrm{US}} \geq 0$. Data of type $(N_{\mathrm{S}\cdot}, N_{\mathrm{FR}}, N_{\mathrm{UR}})$ refines our knowledge about the value of α, but even an infinite sample of this type will not lead to knowledge certain of $p_{\mathrm{FS}\cdot}$ The conditional distribution of p_{FS} given $p_{\mathrm{FR}} + p_{\mathrm{UR}} = \alpha$ prior to observing $\mathbf{y}$ is the same as the conditional distribution of p_{FS} given $p_{\mathrm{FR}} + p_{\mathrm{UR}} = \alpha$ posterior to observing $\mathbf{y}$.

While the theory of identification in a classical statistical context is now well understood, the story of identification from a Bayesian viewpoint is still unfolding. Markov Chain Monte Carlo (MCMC) methods breathe new life into Bayesian inference, allowing efficient computation of normalizing constants and parameter estimates for complex models. An explosion of Bayesian application of MCMC methods surfaces new identification issues. In their study of Bayesian identifiability for 'Generalized Linear Models' (GLMs), Gelfand and Sahu (1999) present conditions under which assignment of an improper prior distribution to some components of a parameter set leads to a proper posterior distribution. More particularly, if a posterior distribution for parameters is improper, under what conditions do there exist a lower-dimensional parameter set that possesses a proper posterior distribution? Ghosh et al. (2000) prove that when data $\mathbf{y}$ is uninformative about $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ so that $\boldsymbol{\theta}_2$ is not identifiable, the posterior density $f(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1|\mathbf{y})$ is proper if and only if both $h(\boldsymbol{\theta}_1|\mathbf{y})$ and $g(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ are proper; a fact that Gelfand and Sahu use to establish that for GLMs, at least, even when the posterior distribution for all parameters is improper, there are embedded lower dimensional models that possess a proper posterior distribution, although, in general, that distribution is not uniquely determined. (See Gelfand and Sahu (1999) for conditions that guarantee uniqueness.) They warn, however, that model fitting using MCMC methods is "…somewhat of an art form, requiring suitable trickery and turning to obtain results in which we can have confidence."

Bibliography

Bartels, R., 1985. Identification in econometrics. The American Statistician 39 (2), 102–104.

Basu, A.P., 1981. In: Taillie, C., Patil, Baldessari (Eds.), Statistical Distributions in Scientific Work, vol. 5. D Reidel, Dordrecht, The Netherlands, pp. 335–348.

Berman, S.M., 1963. Annals of Mathematical Statistics 34, 1104–1106.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Bose, R.C., 1944. In: Proceedings of the 31st Indian Science Congress, Delhi, vol. 3, pp. 5–6.

Bose, R.C., 1947. In: Proceedings of the 34th Indian Science Congress, Delhi, Part II, pp. 1–25.

Bowden, R., 1973. The theory of parametric identification. Econometrica 41 (6), 1069–1074.

Dawid, A.P., 1979. Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society Series B 41, 1–31.

Deistler, M., Seifert, H.-G., 1978. Identifiability and consistent estimability in econometric models. Econometrica 46 (4), 969–980.

Drèze, J.H., 1962. The Bayesian Approach to Simultaneous Equation Estimation. ONR Research Memorandum No. 67. Northwestern University.

Drèze, J.H., 1972. Econometrics and decision theory. Econometrica 40, 1–17.

Drèze, J.H., 1974. Bayesian Theory of Identification in Simultaneous Equations Models. Studies in Bayesian Econometrics and Statistics. North-Holland, Amsterdam (presented to Third NBER-NSF Seminar on Bayesian Inference Econometrics, Harvard University, October 1971). pp. 159–174.

Drèze, J.H., 1976. Bayesian limited information analysis of the simultaneous equations model. Econometrica 44, 1045–1075.

Drèze, J.H., Morales, J.A., 1976. Bayesian full information analysis of simultaneous equations. Journal of the American Statistical Association 71 (376), 919–923.

Drèze, J.H., Richard, J.F., 1983. Bayesian analysis of simultaneous equation systems. Handbook of Econometrics 1.

Fisher, F.M., 1961. On the cost of approximate specification in simultaneous equation estimation. Econometrica 29 (2).

Fisher, F.M., 1966. The Identification Problem in Econometrics. McGraw-Hill, New York.

Gelfand, A., Sahu, S.K., 1999. Identifiability, improper priors, and Gibbs sampling for generalized linear models. Journal of the American Statistical Association 94 (445), 247–253.

Ghosh, M., Chen, M.H., Ghosh, M., Agresti, A., 2000. Non-informative priors for one parameter item response models. Journal of Statistical Planning and Inference 88, 99–115.

Hannan, E.J., 1969. The identification of vector mixed auto-regressive moving average systems. Biometrika 57, 223–225.

Hannan, E.J., 1971. The identification problem for multiple equation systems with moving average errors. Econometrica 39, 751–765.

Haavelmo, T., 1944. The probability approach in econometrics. Econometrica 12 (Suppl.).

Hoyle, R.H., 1995. Structural Equation Modeling: Concepts, Issues and Applications. Sage, Thousand Oaks, CA.

Hurwicz, L., 1950. Generalization of the Concept of Identification. Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph 10. Wiley, New York.
Johnson, W.O., Gastwirth, J.L., 1991. Bayesian inference for medical screening tests: approximations useful for the analysis of AIDS. Journal of the Royal Statistical Society Series B 53, 427–439.
Jöreskog, K.G., 1973. In: Krishnaiah, P.R. (Ed.), Analysis of Covariance Structures. Multivariate Analysis III. Academic Press, New York.
Jöreskog, K.G., Sörbom, D., 1984. LISREL VI Analysis of Linear Structural Relations by Maximum Likelihood, Instrumental Variables, and Least Square Methods. User's Guide. Department of Statistics, University of Uppsala, Uppsala, Sweden.
Kadane, J.B., 1974. The Role of Identification in Bayesian Theory. Studies in Bayesian Econometrics and Statistics. North-Holland, Amsterdam (presented to Winter Meeting of Econometric Society, Toronto, December 1972). pp. 175–191.
Kleibergen, F., Van Dijk, H.K., 1997. Bayesian Simultaneous Equations Analysis Using Reduced Rank Structures. Erasmus University. Technical Report 9714/A.
Koopmans, T.C., Reiersøl, O., 1950. The identification of structural characteristics. Annals of Mathematical Statistics 21, 165–181.
Koopmans, T.C., Rubin, H., Leipnik, R.B., 1950. Measuring the Equation Systems of Dynamic Economics. Statistical Inference in Dynamic Economic Models. Cowles Commission Monograph 10. Wiley, New York.
Lazarsfeld, P.F., 1950. The Logical and Mathematical Foundation of Latent Structure Analysis. The Interpretation of Some Latent Structures. Measurement and Prediction 4: Studies in Psychology of World War II.
Leamer, E.E., 1978. Specification Searches. Wiley, New York.
Mouchard, J.-F., 1976. Posterior and Predictive Densities for Simultaneous Equation Models. Springer-Verlag, New York.
Neath, A.A., Samaniego, F.J., August 1997. On the efficacy of Bayesian inference for nonidentifiable models. The American Statistician 51 (3), 225–232.
Pigou, C., 1910. A method of determining the numerical values of elasticities of demand. Economic Journal 20, 636–640.
Poirier, D.J., 1996. Revising Beliefs in Non-identified Models. Technical Report, University of Toronto, Canada.
Raiffa, H., Schlaifer, R.O., 1961. Applied Statistical Decision Theory. Harvard Business School, Boston.
Rothenberg, T.J., May 1971. Identification in parametric models. Econometrica 39, 577–591.
Thurstone, L.L., 1935. The Vectors of Mind. University of Chicago Press, Chicago.
Thurstone, L.L., 1947. Multiple-Factor Analysis. University of Chicago Press, Chicago.
Van der Genugten, B.B., 1997. Statistica Neerlandica 31, 69–89.
Wald, A., 1950. Statistical Inference in Dynamic Economic Models. Note on the identification of economic relations. Cowles Commission Monograph 10. Wiley, New York.
Wiley, D.E., 1973. The identification problem for structural equation models with unmeasured variables. In: Goldberger, A.S., Duncan, O.D. (Eds.), Structural Equation Models in the Social Sciences. Seminar Press, New York.

Statistical Pattern Recognition

Richard M Golden, University of Texas at Dallas, Richardson, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Statistical pattern recognition is concerned with the problem of designing machines that can detect and classify complex patterns in data. Statistical pattern recognition problems frequently arise in all fields of science and engineering. In particular, statistical pattern recognition methods are widely used by social and behavioral scientists in data analysis and in the course of modeling complex social, behavioral, and neural systems. This article reviews the major components of statistical pattern recognition systems and discusses methods for classification and learning. In addition, the importance of evaluating generalization performance using cross-validation measures and other out-of-sample measures is emphasized.

Statistical pattern recognition is concerned with the problem of using statistical techniques to design machines that can detect and classify patterns of measurement data in useful ways. Such techniques are successfully applied in areas such as object recognition, speech recognition, fingerprint recognition, data analysis, document classification, image analysis, information retrieval, bioinformatics, signal processing, cognitive modeling, machine learning, and artificial intelligence. Examples of statistical pattern recognition machines include many popular connectionist and computational neuroscience models (Golden, 1996). After identifying key 'features' of information in the measurement data, a statistical pattern recognition machine uses its expectations about the 'likelihood' of those features and the 'losses' of making situation-specific decisions to select a classification action that minimizes an appropriate expected risk function. In many cases, the statistical pattern recognition machine must 'learn' through experience the relative likelihood of occurrence of patterns of features in its environment. Although typically viewed as a branch of artificial intelligence or engineering science, statistical pattern recognition methods have been extensively and successfully applied to many behavioral modeling and data analysis problems.

Generic Statistical Pattern Recognition Machine

As shown in Figure 1, a generic statistical pattern recognition machine may be described in terms of four major components: (1) feature extraction, (2) knowledge representation of both loss and probabilistic uncertainty, (3) a Bayes decision rule for classification, and (4) an optional mechanism for learning.

Feature Selection and Extraction

The pattern of measurement data to be classified is represented as a measurement vector (i.e., a point in 'measurement space'). The choice of the measurement space is primarily determined by problem-imposed considerations such as the types of available transducers and the types of patterns that must be processed. The particular pattern classification methodology embodied in the pattern recognition machine usually has little or no influence over the choice of the measurement space. Each measurement vector is then mapped into a feature vector (i.e., a 'point' in feature space) whose dimensions are primarily chosen to facilitate pattern classification by the pattern recognition machine. This (possibly nonlinear) transformation is a crucial component of the researcher's design of the statistical pattern recognition machine (Duda et al., 2001; Hastie et al., 2009). The design of the measurement space to feature space transformation may be informed by prior knowledge (Duda et al., 2001; Hastie et al., 2009; Nilsson, 1965), optimally chosen using appropriate statistical tests (see Golden, 2003; for an example of such a test), or learned using support vector machine or multilayer artificial neural network learning methods such as backpropagation (Duda et al., 2001; Golden, 1996; Hastie et al., 2009; Nilsson, 1965). The process that maps a measurement vector $\mathbf{s}$ into a d-dimensional feature vector $\mathbf{x}$ is called the feature selection and extraction stage of the pattern recognition process.

For example, Figure 1 shows a digital image of a dog represented by a matrix of image pixels. The matrix has 227 rows and 304 columns of image pixels. Each image pixel, in turn, is represented by a positive integer ranging in value from 1 to 256 indicating the image intensity of that pixel. Thus, the measurement vector in this example is a real vector consisting of $227 \times 304 = 69\,008$ elements.

The feature space in this example is generated by the feature selection and extraction stage that transforms a 69 008-dimensional measurement vector into a 7-dimensional feature vector by an appropriate feature map. Useful feature maps for this type of problem include spatial Fourier transforms, spatial derivative transformations, support vector machines, kernel methods, or multilayer feedforward artificial neural network preprocessing transformations. Such approaches as well as others are discussed by Duda and Hart (1973), Duda et al. (2001), and Hastie et al. (2009).

Probabilistic Knowledge Representation

Knowledge is implicitly represented in a statistical pattern recognition machine in several ways. The particular process for mapping points in measurement space into feature space discussed in the previous section is one important knowledge source. However, the expected likelihood of observing a particular feature vector and the associated losses for making different situation-specific decisions are also critical knowledge sources.

Specifically, let the symbol ω_j denote the jth pattern class in a set of J possible pattern classes (i.e., $j = 1, \ldots, J$). Thus, ω_1 refers to the first pattern class (e.g., a set of photographs of

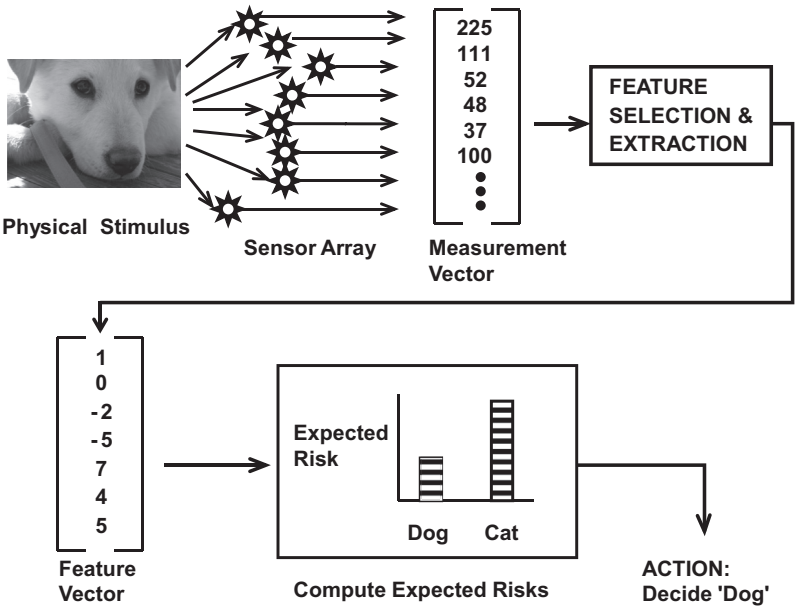

Figure 1 Flow chart of a generic statistical pattern recognition machine. In this example, the sensor array (represented as a measurement vector) detects a pattern of light intensities produced from light reflecting off the head of a dog. A feature selection and extraction stage transforms the measurement vector into a feature vector. Using the feature vector and its probabilistic knowledge representation and loss function representation of the consequences of its decision actions, the pattern recognition machine selects the minimum expected risk decision (i.e., 'Decide Dog').

'dogs') that might have generated the observed feature vector $\mathbf{x}$, while the symbol ω_2 refers to a second competing pattern class (e.g., a set of photographs of 'cats') that might have generated $\mathbf{x}$. The likelihood of observing a feature vector $\mathbf{x}$ corresponding to a particular pattern class is expressed in terms of two types of probability densities. The a priori class probability density $p(\omega_j)$ indicates the likelihood that a feature vector is generated from pattern class ω_j when information about the feature vector $\mathbf{x}$ is not available. The class conditional probability density $p(\mathbf{x}|\omega_j)$ indicates the conditional likelihood of observing the feature vector $\mathbf{x}$ given knowledge that the feature vector $\mathbf{x}$ was generated from pattern class ω_j.

Knowledge regarding the expected consequences of taking a particular classification action α_k given that the feature vector came from pattern class ω_j is also required. Assume there are K possible actions so that $k = 1, \ldots, K$. Typically, the classification action takes the form of a decision indicating which of a finite number of pattern classes generated the feature vector. Formally, it is assumed that a loss $\lambda(\alpha_k|\omega_j)$ is incurred when classification action α_k is taken given that the feature vector was in fact generated from pattern class ω_j.

Referring to the example in Figure 1 assumes two pattern classes $\omega_1 =$ 'Dog' and $\omega_2 =$ 'Cat' and three possible actions: $\alpha_1 =$ 'Decide Dog,' $\alpha_2 =$ 'Decide Cat,' and $\alpha_3 =$ 'Decide Uninterpretable.'

Finally, although assumptions of a finite number of classes and actions are commonly made in the literature, one might extend the traditional statistical pattern recognition framework to include situations involving infinite numbers of pattern classes or infinite numbers of possible classification actions. In such situations, an 'action' is a real vector whose values are estimated (e.g., the amount of pressure to be applied to an accelerator or brake pedal of a car).

Bayes Decision Rule for Classification

The optimal decision rule for selecting a classification action α^* generated by pattern class ω_j is to select α^* such that $\lambda(\alpha^*|\omega_j) \geq \lambda(\alpha_k|\omega_j)$ for $k = 1, \ldots, K$. In practice, however, the pattern class ω_j that actually generated the feature vector is unobservable. The statistical pattern recognition machine does, however, have knowledge of (1) the observed feature vector $\mathbf{x}$, (2) a representation of environmental uncertainty (i.e., the probability density $p(\mathbf{x}|\omega_j)$ and probability mass function $p(\omega_j)$), and (3) a representation of the consequences of the machine's actions (i.e., the loss function $\lambda(\alpha_k|\omega_j)$).

One possible rational inductive decision-making strategy is called the Bayes decision rule for classification; it forms the basis for classification decisions in most statistical pattern recognition algorithms. Bayes decision rules can be derived from classical axioms of rational decision making (see Golden, 1996; for a review). The Bayes decision rule for classification dictates that the pattern recognition machine should select the classification action α^* that minimizes the expected loss or conditional risk given feature vector $\mathbf{x}$ is observed. Formally, the conditional Bayes risk function for classification $R(\alpha_k|\mathbf{x})$ is defined by the formula

$$R(\alpha_k|\mathbf{x}) = \sum_{j=1}^{J} \lambda(\alpha_k|\omega_j) p(\omega_j|\mathbf{x}) \qquad [1]$$

The important minimum probability of error case of eqn [1] arises when $\lambda(\alpha_k|\omega_j) = 1$ when $j \neq k$ and $\lambda(\alpha_k|\omega_j) = 0$ when $j = k$. In this '0–1 loss function case', eqn [1] becomes

$$R(\alpha_k|\mathbf{x}) = \sum_{j \neq k, j=1}^{J} p(\omega_j|\mathbf{x}) = 1 - p(\omega_k|\mathbf{x}) \qquad [2]$$

The action α^* that minimizes the probability of error case eqn [2] of the conditional risk function eqn [1] is usually called a maximum a posteriori (MAP) estimate since α_k maximizes the a posteriori density $p(\omega_k|\mathbf{x})$ (i.e., the likelihood of pattern class ω_k after observing the feature vector $\mathbf{x}$).

Thus, a minimum probability of error decision for the example in Figure 1 is implemented by penalizing the machine with a loss of one when the statistical pattern recognition machine does not decide 'cat' when a photograph of a cat is presented and does not decide 'dog' when a photograph of a dog is presented. All other decisions by the machine are penalized with a loss value of zero.

Now note that the risk $R(\alpha_k|\mathbf{x})$ in eqn [1] can be expressed in terms of the a priori class likelihoods and class conditional likelihoods using the formula $p(\omega_j|\mathbf{x}) = p(\mathbf{x}|\omega_j)p(\omega_j)/p(\mathbf{x})$. Substituting this expression into eqn [1], it follows that the risk, $R(\alpha_k|\mathbf{x})$, for selecting action α given feature vector $\mathbf{x}$ is observed can be expressed as

$$R(\alpha_k|\mathbf{x}) = \left(\frac{1}{p(\mathbf{x})}\right)\sum_{j=1}^{J}\lambda(\alpha_k|\omega_j)p(\mathbf{x}|\omega_j)p(\omega_j) \qquad [3]$$

In practice, direct computation of eqn [3] may be computationally intractable, and so for computational reasons, some monotonically decreasing function of $R(\alpha_k|\mathbf{x})$ is usually defined as a discriminant function $\phi_k(\mathbf{x})$ (e.g., Duda et al., 2001; Hastie et al., 2001). For example, a useful discriminant function for eqn [3] that does not require computing $p(\mathbf{x})$ is given by the formula $\phi_k(\mathbf{x}) = \sum_{j=1}^{J}\lambda(\alpha_k|\omega_j)p(\mathbf{x}|\omega_j)p(\omega_j)$.

Such discriminant functions define 'decision boundaries' in feature space. For example, suppose that there are only two actions α_1 and α_2 corresponding, respectively, to selecting pattern classes ω_1 and ω_2. To implement a decision rule that minimizes the expected risk, the feature vector $\mathbf{x}$ is classified as a member of pattern class ω_1 if $\phi_1(\mathbf{x}) < \phi_2(\mathbf{x})$ and the feature vector $\mathbf{x}$ is classified as a member of pattern class ω_2 if $\phi_1(\mathbf{x}) > \phi_2(\mathbf{x})$. The 'decision boundary' between the pattern classes ω_1 and ω_2 in feature space is defined by the set of points $\{\mathbf{x} : \phi_1(\mathbf{x}) = \phi_2(\mathbf{x})\}$.

Learning

Knowledge Representations

The generic strategy for classification outlined in the previous selection requires knowledge of the (1) a priori class probabilities $p(\omega_1), \ldots, p(\omega_J)$, (2) class conditional likelihoods/densities $p(\mathbf{x}|\omega_1), \ldots, p(\mathbf{x}|\omega_J)$, and (3) loss function $\lambda(\alpha_k|\omega_j)$ associated with selecting classification action α_k given the correct classification of the feature vector is class ω_j. Typically, it is assumed that the a priori class probabilities $p(\omega_1), \ldots, p(\omega_J)$ either are known or can be reliably estimated from previous data. It is also typically assumed that the incurred loss function $\lambda(\alpha_k|\omega_j)$ is always explicitly known.

Thus, the objective of the learning problem is to identify the class conditional likelihoods. A typical strategy for approaching this problem is to define a class conditional probability model for pattern class ω_j that consists of a collection of alternative choices of class conditional likelihoods for pattern class ω_j where each element, $p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)$, of the class conditional probability model is a probability density uniquely identified by a particular value of the q-dimensional parameter vector $\boldsymbol{\beta} \in \mathscr{R}^q$. Note that these modeling assumptions imply the training vectors generated from a particular pattern class ω_k for a particular parameter value $\boldsymbol{\beta}^*$ are a realization of a sequence of independent and identically distributed (i.i.d.) random vectors. The a priori preference for a particular element $p(\mathbf{x}|\boldsymbol{\beta}_j, \omega_j)$ in the class conditional probability model is given by the a priori parameter density $p(\boldsymbol{\beta})$. Given $p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)$, $p(\boldsymbol{\beta})$, and the training data $\mathscr{D}_n$ the objective of the learning problem is to identify the class conditional density $p(\mathbf{x}|\omega_j)$ through the examination of the elements of the class conditional probability model $\{p(\mathbf{x}|\boldsymbol{\beta}, \omega_j) : \boldsymbol{\beta} \in \mathscr{R}^q\}$.

Environmental Modeling Assumptions

We assume that $(\mathbf{x}^i, \omega^i)$, $i = 1, \ldots, n$ is a realization of a sequence of i.i.d. random vectors $(\mathbf{X}^i, \boldsymbol{\omega}^i)$, $i = 1, \ldots, n$ where the notation $(\mathbf{x}^i, \omega^i)$ indicates that the ith pattern vector originated from pattern class $\omega^i \in \{\omega_1, \ldots, \omega_J\}$. We assume $\mathbf{x}^i$ takes on a countable number of possible values in a set $\Omega \subseteq \mathscr{R}^d$. However, we do not necessarily assume that $(\mathbf{X}^i, \boldsymbol{\omega}^i)$ is completely observable. Rather we assume that the training data $\mathscr{D}_n$ is defined as a collection of n data observable vectors $\mathbf{y}^1, \ldots, \mathbf{y}^n$ that are realizations of a completely observable stochastic process of (conditionally) i.i.d. random vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ with common density $p(\mathbf{y}^i|\mathbf{x}^i, \boldsymbol{\omega}^i)$. The marginal density $p(\mathbf{y}^i|\boldsymbol{\beta}) \equiv \sum_{j=1}^{J}\sum_{\mathbf{x}\in\Omega} p(\mathbf{y}^i|\mathbf{x}, \omega_j)p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)p(\omega_j)$ is used to model the observable component of the data in terms of the class conditional densities and a priori class probabilities. The assumption that $(\mathbf{X}^i, \boldsymbol{\omega}^i)$ and $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are i.i.d. discrete random vectors can be considerably relaxed and is made only for expository reasons.

Supervised, Unsupervised, and Reinforcement Learning Paradigms

For the case of supervised learning, the observable training data has the property that $p(\mathbf{Y}^i = [\mathbf{x}^i, \omega^i]|\mathbf{x}^i, \omega^i) = 1$ if the pattern $\mathbf{x}^i$ and the pattern class ω^i from which $\mathbf{x}^i$ originated are both completely observable. Thus,

$$p(\mathbf{y}^i|\boldsymbol{\beta}) \equiv \sum_{j=1}^{J}\sum_{\mathbf{x}\in\Omega} p(\mathbf{y}^i|\mathbf{x}, \omega_j)p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)p(\omega_j) = p(\mathbf{x}^i, \omega^i|\boldsymbol{\beta})$$

For the case of unsupervised learning, the observable training data has the property that $p(\mathbf{Y}^i = \mathbf{x}^i|\mathbf{x}^i, \omega^i) = 1$ corresponding to the case where every pattern $\mathbf{x}^i$ is completely observable yet the pattern class ω^i from which $\mathbf{x}^i$ originated is not observable. Thus,

$$\begin{aligned} p(\mathbf{y}^i|\boldsymbol{\beta}) &\equiv \sum_{j=1}^{J}\sum_{\mathbf{x}\in\Omega} p(\mathbf{y}^i|\mathbf{x}, \omega_j)p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)p(\omega_j) \\ &= \sum_{j=1}^{J} p(\mathbf{x}^i|\boldsymbol{\beta}, \omega_j)p(\omega_j) = p(\mathbf{x}^i|\boldsymbol{\beta}) \end{aligned}$$

Note that the unsupervised learning case is formally equivalent to a 'clustering analysis' problem where the training data originates from different clusters $\omega_1, \ldots, \omega_J$ yet the identity of the cluster for a given pattern vector is not observable.

For the case of reinforcement learning, one receives only a 'hint' regarding the identity of the pattern $\mathbf{x}^i$ and pattern class ω^i through observing $\mathbf{y}^i$. That is, the identity of the pattern $\mathbf{x}^i$

and pattern class ω^i are only partially observable. For example, $\mathbf{y}^i$ might be a scalar-valued variable that reveals only a subtle clue regarding the identity of pattern $\mathbf{x}^i$ and the pattern class ω^i from which $\mathbf{x}^i$ originated.

Estimative Approach: A Bayes Risk Function for Learning

A Bayes risk function estimative approach (Aitchison, 1975; Patrick, 1972) may be defined as finding $\hat{\boldsymbol{\beta}}$ that minimizes the conditional Bayes risk function for learning: $R(\boldsymbol{\beta}|\mathscr{D}_n) \equiv \int c(\boldsymbol{\beta}, \ddot{\boldsymbol{\beta}}) p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n) d\ddot{\boldsymbol{\beta}}$ where $c(\hat{\boldsymbol{\beta}}, \ddot{\boldsymbol{\beta}})$ denotes the loss incurred in taking the action of 'plugging in' the constant $\hat{\boldsymbol{\beta}}$ into the class conditional model formula $p(\mathbf{x}|\hat{\boldsymbol{\beta}}, \omega_j)$ to obtain the class conditional density $p(\mathbf{x}|\omega_j) \equiv p(\mathbf{x}|\hat{\boldsymbol{\beta}}, \omega_j)$ when the data was, in fact, generated from the class conditional density $p(\mathbf{x}|\ddot{\boldsymbol{\beta}}, \omega_j)$. This methodology is analogous to minimizing expected risk during pattern classification as described in Section Generic statistical pattern recognition machine. The a posteriori density $p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n)$ is calculated using the formula $p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n) = p(\mathscr{D}_n|\ddot{\boldsymbol{\beta}}) p(\ddot{\boldsymbol{\beta}}) / p(\mathscr{D}_n)$ where $p(\mathscr{D}_n|\ddot{\boldsymbol{\beta}})$ is called the likelihood function because it specifies the likelihood of the observed data given the parameter vector $\ddot{\boldsymbol{\beta}}$.

For expository reasons, assume that $p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n)$ is a probability mass function so that $\ddot{\boldsymbol{\beta}}$ takes on only a finite number of values. We now consider the important '0–1 loss function special case' where $c(\hat{\boldsymbol{\beta}}, \ddot{\boldsymbol{\beta}}) = 1$ if $\hat{\boldsymbol{\beta}} \neq \ddot{\boldsymbol{\beta}}$ and $c(\hat{\boldsymbol{\beta}}, \ddot{\boldsymbol{\beta}}) = 0$ only if $\hat{\boldsymbol{\beta}}_j = \ddot{\boldsymbol{\beta}}_j$ (i.e., it is assumed that there is a uniquely correct choice for the parameter values). Furthermore, to simplify the exposition, we also temporarily assume that $p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n)$ is a probability mass function so that $\ddot{\boldsymbol{\beta}}$ takes on only a finite number of values (e.g., $\ddot{\boldsymbol{\beta}}$ is 'quantized'). Then we have $R(\boldsymbol{\beta}|\mathscr{D}_n) \equiv \sum_{\ddot{\boldsymbol{\beta}} \in \Omega} c(\boldsymbol{\beta}, \ddot{\boldsymbol{\beta}}) p(\ddot{\boldsymbol{\beta}}|\mathscr{D}_n) = 1 - p(\boldsymbol{\beta}|\mathscr{D}_n)$. Thus, the 0–1 loss function special case corresponds to a minimum probability decision rule of selecting a 'most probable' value of the quantized $\ddot{\boldsymbol{\beta}}$ given the observed data $\mathscr{D}_n$. Because this learning strategy involves maximizing the a posteriori density $p(\boldsymbol{\beta}|\mathscr{D}_n)$, it is referred to as a MAP estimation learning strategy. A global maximizer $\tilde{\boldsymbol{\beta}}$ of the a posteriori density $p(\boldsymbol{\beta}|\mathscr{D}_n)$ is called a MAP estimate. Since $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d., the negative normalized log-likelihood $l_n(\boldsymbol{\beta}) \equiv -\frac{1}{n} \log p(\mathscr{D}_n|\boldsymbol{\beta})$ where $-\frac{1}{n} \log p(\mathscr{D}_n|\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}^i|\boldsymbol{\beta})$. A global minimizer $\hat{\boldsymbol{\beta}}_n$ of the negative normalized likelihood is called a maximum likelihood (ML) estimate.

The negative normalized log-likelihood $l_n(\boldsymbol{\beta})$ measures the likelihood of the observed data for a given parameter value in contrast to $p(\boldsymbol{\beta}|\mathscr{D}_n)$, which is the likelihood of a parameter value given the observed data. As $n \to \infty$, $l_n(\boldsymbol{\beta})$ converges to the cross-entropy $l(\boldsymbol{\beta}) \equiv -\sum_{\mathbf{y}} p(\mathbf{y}) \log p(\mathbf{y}|\boldsymbol{\beta})$ between the observable data distribution and the researcher's model of the observable data distribution.

Using the relation $p(\mathscr{D}_n|\boldsymbol{\beta}) = \exp[-n l_n(\boldsymbol{\beta})]$, an explicit formula for $p(\boldsymbol{\beta}|\mathscr{D}_n)$ is given by

$$p(\boldsymbol{\beta}|\mathscr{D}_n) = \exp[-n(l_n(\boldsymbol{\beta}) - \frac{1}{n} \log p(\boldsymbol{\beta}) + \frac{1}{n} \log p(\mathscr{D}_n))] \quad [4]$$

where inspection of eqn [4] shows that for large sample size n, a global minimizer of $l_n(\boldsymbol{\beta})$ (i.e., an ML estimate) will be an approximate global maximizer of $p(\boldsymbol{\beta}|\mathscr{D}_n)$ (i.e., a MAP estimate) because $\frac{1}{n} \log p(\mathscr{D}_n)$ is a constant on the parameter space and $\frac{1}{n} \log p(\boldsymbol{\beta}) \to 0$ (under typically assumed conditions).

Predictive Approach: Bayesian Learning

The key idea of the estimative approach in Section Estimative approach: a Bayes risk function for learning was that the class conditional density $p(\mathbf{x}|\omega_j)$ is obtained by 'plugging in' a particular parameter value $\hat{\boldsymbol{\beta}}$ (that minimizes a Bayes risk function) into the class conditional model formula $p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)$ to obtain $p(\mathbf{x}|\omega_j) = p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)$. An alternative Bayesian approach referred to as the 'predictive approach' by Aitchison (1975) is to integrate $p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)$ using the formula

$$p(\mathbf{x}|\omega_j) = \int p(\mathbf{x}|\boldsymbol{\beta}, \omega_j) p(\boldsymbol{\beta}|\mathscr{D}_n) d\boldsymbol{\beta} \quad [5]$$

Aitchison (1975) provides a theoretical analysis demonstrating that for a wide range of situations the predictive approach is superior to the estimative approach based upon a cross-entropy goodness-of-fit criterion. However, the practical difficulty of applying the predictive approach is that the integral in eqn [5] is typically computationally intractable. In this section, we will discuss analytic and simulation methods for evaluating the integral in eqn [5].

Analytic Methods for Applying the Predictive Approach

In this section, we informally discuss an analytic formula for approximating the integral in eqn [5] valid for sufficiently large sample sizes (Tierney and Kadane, 1986). The integral in eqn [5] may be rewritten as

$$\begin{aligned} p(\mathbf{x}|\omega_j) &= \int p(\mathbf{x}|\boldsymbol{\beta}, \omega_j) \left(\frac{p(\mathscr{D}_n|\boldsymbol{\beta}) p(\boldsymbol{\beta})}{p(\mathscr{D}_n)} \right) d\boldsymbol{\beta} \\ &= \frac{\int p(\mathbf{x}|\boldsymbol{\beta}, \omega_j) p(\mathscr{D}_n|\boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}}{\int p(\mathscr{D}_n|\boldsymbol{\beta}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}} \end{aligned}$$

Now let $\psi_j(\boldsymbol{\beta}) = -n^{-1} \log(p(\mathbf{x}|\boldsymbol{\beta}, \omega_j) p(\mathscr{D}_n|\boldsymbol{\beta}) p(\boldsymbol{\beta}))$ and $\overline{\psi}(\boldsymbol{\beta}) = -n^{-1} \log(p(\mathscr{D}_n|\boldsymbol{\beta}) p(\boldsymbol{\beta}))$. Let $\hat{\boldsymbol{\beta}}_j$ and $\overline{\boldsymbol{\beta}}$ be the respective global minimizers of ψ_j and $\overline{\psi}$. Let the notation $\mathscr{I}^j(\psi_j) = \int \exp(-n\psi_j(\boldsymbol{\beta})) d\boldsymbol{\beta}$ and the notation $\overline{\mathscr{I}}(\psi_j) = \int \exp(-n\overline{\psi}(\boldsymbol{\beta})) d\boldsymbol{\beta}$ so that $p(\mathbf{x}|\omega_j) = \mathscr{I}^j(\psi_j) / \overline{\mathscr{I}}(\overline{\psi})$.

An informal discussion of a method for approximating quantities such as $\mathscr{I}^j(\psi_j)$ and $\overline{\mathscr{I}}(\overline{\psi})$ is now provided. Let $\psi(\boldsymbol{\beta})$ be a smooth unimodal function whose gradient evaluated at the global minimizer $\hat{\boldsymbol{\beta}}$ vanishes. Expand $\psi(\boldsymbol{\beta})$ in a second-order Taylor series expansion about $\hat{\boldsymbol{\beta}}$ to obtain $\psi(\boldsymbol{\beta}) \approx \psi(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} \hat{-} \boldsymbol{\beta})^T [\nabla^2 \psi(\hat{\boldsymbol{\beta}})](\boldsymbol{\beta} \hat{-} \boldsymbol{\beta})$ where $\nabla^2 \psi(\hat{\boldsymbol{\beta}})$ is positive definite because $\hat{\boldsymbol{\beta}}$ is a global minimizer. The integral $\mathscr{I}(\psi) = \int \exp(-n\psi(\boldsymbol{\beta})) d\boldsymbol{\beta}$ can be approximated by $\tilde{\mathscr{I}}(\psi) = \int \exp(-n[\psi(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} \hat{-} \boldsymbol{\beta})^T [\nabla^2 \psi(\hat{\boldsymbol{\beta}})](\boldsymbol{\beta} \hat{-} \boldsymbol{\beta})]) d\boldsymbol{\beta}$ when the sample size n is sufficiently large (Tierney and Kadane, 1986). Note that since the integral of a multivariate Gaussian density with mean $\hat{\boldsymbol{\beta}}$ and covariance matrix $[n\nabla^2 \psi(\hat{\boldsymbol{\beta}})]^{-1}$ equals one we have

$$(2\pi)^{-q/2} (DET(n\nabla^2 \psi(\hat{\boldsymbol{\beta}})))^{1/2} \int \exp\left(-\frac{1}{2}(\boldsymbol{\beta} \hat{-} \boldsymbol{\beta})^T \times [n\nabla^2 \psi(\hat{\boldsymbol{\beta}})](\boldsymbol{\beta} \hat{-} \boldsymbol{\beta}) \right) d\boldsymbol{\beta} = 1 \quad [6]$$

Using the identity in eqn [6],

$$\hat{\mathscr{I}}(\psi) = \exp[-n\psi(\hat{\boldsymbol{\beta}})](2\pi)^{q/2}(DET(n\nabla^2\psi(\hat{\boldsymbol{\beta}})))^{-1/2} \quad [7]$$

Thus, $p(\mathbf{x}|\omega_j) \approx \hat{\mathscr{I}}(\psi_j)/\hat{\mathscr{I}}(\overline{\psi})$.

Predictive Approach with Nonparametric Bootstrap Simulation

A nonparametric bootstrap simulation method also called 'bagging' (Hastie et al., 2009) provides a simple simulation method for implementing the predictive approach.

- Step 1. Estimate a density $\hat{p}(\mathbf{y}|\mathscr{D}_n)$ for generating the observed data $\mathscr{D}_n$ by assigning a probability mass of $1/n$ to each of the original n observed data samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
- Step 2. Let $b = 1$.
- Step 3. Sample n times with replacement from density $\hat{p}(\mathbf{y})$ to generate the bth bootstrap sample: $\mathbf{y}_1^b, \ldots, \mathbf{y}_n^b$.
- Step 4. Use $\mathbf{y}_1^b, \ldots, \mathbf{y}_n^b$ to compute an MLE or MAP estimate $\hat{\boldsymbol{\beta}}^b$.
- Step 5. Compute $\hat{p}^b(\mathbf{x}|\omega_j) = \frac{1}{b}\sum_{m=1}^{b} p(\mathbf{x}|\hat{\boldsymbol{\beta}}^m, \omega_j)$.
- Step 6. Let $b = b + 1$ and then go to step 3.

Predictive Approach with Parametric Bootstrap Simulation

- Step 1. Let $b = 1$.
- Step 2. Sample with replacement b times from $p(\boldsymbol{\beta}|\mathscr{D}_n)$ to obtain $\hat{\boldsymbol{\beta}}^1, \ldots, \hat{\boldsymbol{\beta}}^b$.
- Step 3. Compute $\hat{p}^b(\mathbf{x}|\omega_j) = \frac{1}{b}\sum_{m=1}^{b} p(\mathbf{x}|\hat{\boldsymbol{\beta}}^m, \omega_j)$.
- Step 4. Let $b = b + 1$ and then go to step 2.

As $b \to \infty$, $\hat{p}^b(\mathbf{x}|\omega_j) \to p(\mathbf{x}|\omega_j) = \int p(\mathbf{x}|\boldsymbol{\beta}, \omega_j)p(\boldsymbol{\beta}|\mathscr{D}_n)d\boldsymbol{\beta}$ under general conditions. However, implementing step 2 requires sampling from the a posteriori parameter density $p(\boldsymbol{\beta}|\mathscr{D}_n)$. This sampling process is typically computationally intractable so Monte Carlo Markov Chain methods such as Gibbs sampling or the Metropolis algorithms are often used to implement computationally tractable sampling algorithms (Hastie et al., 2009; Jackman, 2000).

Nonparametric and Parametric Learning Algorithms

Learning algorithms that make minimal assumptions regarding how the data was generated are referred to as nonparametric learning algorithms. For such algorithms, the parametric form of the class conditional density is often quite flexible. Nonparametric Parzen-window type learning algorithms e.g., (Duda et al., 2001; Hastie et al., 2009) essentially count the observed occurrence of different values of the feature vector $\mathbf{x}$. Learning algorithms that incorporate considerable prior knowledge about the data generating process are referred to as parametric learning algorithms. The classical Bayes classifier, which assumes that the class conditional densities are Gaussian, is a good example of a parametric learning algorithm (Duda et al., 2001; Patrick, 1972).

Nonparametric learning algorithms have the advantage of being relatively 'unbiased' in their analysis of the data but may require relatively large amounts of the right type of training data for effective learning. Moreover, despite the fact that nonparametric learning algorithms make minimal assumptions about the data, these minimal assumptions may still be incorrect. Parametric learning algorithms are 'biased' algorithms, but if the right type of prior knowledge is 'built-in' to the learning algorithm then a parametric learning algorithm can exhibit superior learning from even poor quality training data. Statistical pattern recognition algorithms that are typically viewed as 'nonparametric' can be transformed into parametric learning algorithms if a sufficient number of constraints upon the class conditional densities are imposed. For example, multilayer backpropagation neural network and perceptron-type learning algorithms (Duda et al., 2001; Golden, 1996; Hastie et al., 2009) can learn to approximate almost any desired form of the class conditional density function (given sufficiently many free parameters). Thus, such learning algorithms with large numbers of free parameters are naturally viewed as nonparametric statistical pattern recognition algorithms. On the other hand, the use of prior knowledge to construct the parametric form of the class conditional probability models results in parametric variations of such classical nonparametric learning algorithms. In other words, a multilayer backpropagation neural network learning algorithm is also naturally viewed as a parametric learning machine when the number of free parameters in such algorithms is constrained to be small (Golden, 1996).

Evaluation of Generalization Performance

In order to evaluate the performance of a statistical pattern recognition machine, it is not sufficient to report classification performance using the pattern vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ that constitute the original data sample $\mathscr{D}_n$. The performance of a poorly designed statistical pattern recognition machine may be quite good on the original data sample $\mathscr{D}_n$ yet exhibit abysmal performance on new data that is not included in the original data sample $\mathscr{D}_n$.

One approach to this problem is to collect an additional data sample $\mathscr{D}'_n \equiv [\mathbf{y}'_1, \ldots, \mathbf{y}'_n]$ (called the 'test data') with the same number of pattern vectors as the original data sample $\mathscr{D}_n \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_n]$. The original data sample $\mathscr{D}_n$, which is used to estimate the parameters of the model (and possibly specify the model architecture), is called the 'training data'. Classification results can then be reported using both the training data $\mathscr{D}_n$ and the test data $\mathscr{D}'_n$. Figure 2 illustrates the performance of a pattern recognition machine that does not evidence poor generalization performance. Notice that classification performance on the training data is consistently better than classification performance on the test data, yet this difference is relatively small in magnitude.

Figure 2 illustrates the performance of a poorly designed statistical pattern recognition machine applied to the same training and test data sets. Although the classification performance of the 'poor' machine on the training data appears superior to performance results of the 'good' machine on the training data, the performance of the 'poor' machine on the test data is very poor. These types of results are strong indicators that the poor statistical pattern recognition machine will exhibit poor generalization performance. A typical interpretation of the pattern of results shown in Figure 2 is that the poor pattern recognition machine has failed to learn crucial statistical regularities that are common to many different data samples but instead has learned 'statistical flukes' peculiar to the specific training data sample.

In practice, there are two major difficulties with using a training data and test data set approach. First, it is possible that crucial statistical regularities that are necessary for optimal

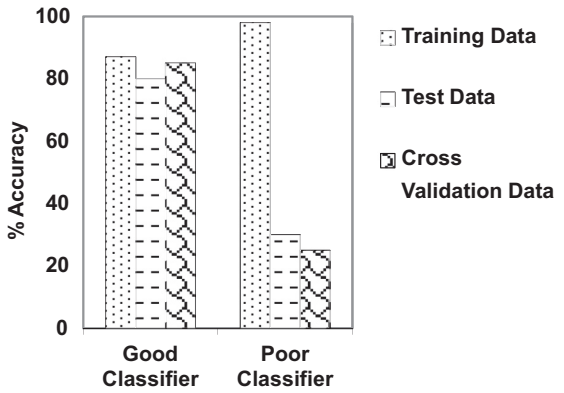

Figure 2 Evaluation of generalization performance in a statistical pattern recognition machine. In this example, the performance of the 'good' classifier evidences good classification performance on the training data, good classification performance on the test data, and good classification performance on a 'leave-one-out' cross-validation estimator using the training data. On the other hand, the performance of the 'poor' classifier is expected to exhibit poor generalization performance because it shows good classification performance on the training data yet poor classification performance on the test data and poor classification on the 'leave-one-out' cross-validation estimator on the training data.

learning are not present in the training data but are, in fact, present in the test data. Thus, learning performance might be improved by training the system with all the collected data. However, this would imply that there would not be a test data sample to support evaluation. Second, it is possible that all the crucial statistical regularities necessary for optimal learning are, in fact, in the original training data sample but the test data sample is not representative of a typical 'novel' data sample. Poor classification results on the test data sample in such a situation would thus be misleading. To address these issues, a number of approaches have been developed.

A widely used computationally intensive approach to addressing these problems is K-fold cross-validation (Hastie et al., 2009; pp. 214–216; Duda et al., 2001). The K-fold cross-validation method involves dividing the original data set $\mathscr{D}_n$ into K equal groups of observable data records. Next, $K-1$ of the groups are used as the training data and the remaining group of data is used as the test data. The classification performance statistics on the training and test data are then recorded. Next, a different choice of $K-1$ groups are defined as the training data, leaving the remaining group as the new test data group and again classification performance statistics on the training and test data recorded. Since there are K groups, this process can be repeated K times so that every one of the original K groups of data has the opportunity to be the test data group. The average and the standard deviation of the K 'training performance' and the K 'test performance' classification statistics are computed to, respectively, obtain estimators for test and training classification performance scores with standard errors. In the special case where the number of groups K is equal to the number of data samples n, then K-fold cross-validation is referred to as 'leave-one-out' cross-validation or 'jackknifing.' Figure 2 illustrates the ideas that even though the training data is used to compute the cross-validation estimator, the cross-validation estimator behaves in a manner that is similar to an estimator based upon 'test data.'

Another important approach for assessing generalization performance is called nonparametric bootstrapping (Hastie et al., 2009; pp. 214–221). An advantage of this approach over the K-fold cross-validation method is that the sample size of the training data and the sample size of the test data are both equal to the sample size n of the original data set. Training data performance is assessed simply by training the statistical pattern recognition machine using the original data sample $\mathscr{D}_n$ and recording the classification performance statistics. The mth nonparametric bootstrap data sample is generated by sampling n times with replacement from the original data sample $\mathscr{D}_n \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ to obtain $\mathscr{D}_n^m = [\mathbf{y}_1^m, \ldots, \mathbf{y}_n^m]$, $m = 1, \ldots, b$. By construction then, a particular nonparametric bootstrap sample will typically omit some of the elements of the original data sample and have replications of other elements. The classification performance statistics are then computed from the mth nonparametric bootstrap sample. Nonparametric bootstrap classification performance estimators are obtained by computing the average and standard deviation of the b classification performance statistics from b nonparametric bootstrap samples. As $b \to \infty$, the average and standard deviation of the b classification performance statistics may be used to, respectively, estimate the nonparametric bootstrap classification performance estimator and its standard error. One difficulty with this form of the nonparametric bootstrap estimator is that parts of the training data may overlap with the test data. Hastie et al. (2009) review an approach for adjusting the estimator to compensate for this bias.

Applications

The number of applications of statistical pattern recognition is so diverse that any review cannot possibly do justice to indicating the range of applications (but see Jain et al., 2000; for a recent review). However, to provide some insights into the relevance of this area, a few arbitrary chosen examples of work in this area are provided.

Artificial Intelligence

Statistical pattern recognition methods have been extensively applied in the field of artificial intelligence. Successful applications of these methods in the field of computer vision include extraction of low-level visual information from visual images, edge detection, extracting shape information from shading information, object segmentation, and object labeling (e.g., Duda and Hart, 1973; Duda et al., 2001). Statistical pattern recognition methods such as hidden Markov models play an important role in speech recognition algorithms and natural language understanding (e.g., Duda et al., 2001). In the fields of robot control and neurocomputing, optimal control theory may be interpreted within a statistical pattern recognition framework where the control signal corresponds to the pattern recognition machine's 'action.'

Behavioral Modeling Applications

Econometricians and psychometricians make extensive use of statistical pattern recognition techniques for investigating the

behavior of individuals and groups of individuals. In addition, researchers have developed statistical pattern models of both human perceptual and human cognitive processes (Chater et al., 2006; Jacobs and Kruschke, 2011). Such probabilistic pattern recognition modeling applications include auditory–visual spatial integration, sensorimotor learning, associative learning, concept learning, multinomial processing tree cognitive process stage models, item response theory, multidimensional general recognition theory, psychometric cognitive diagnostic assessment, and language processing and language acquisition.

Data Analysis Applications

Classical statistical pattern recognition techniques such as factor analysis and principal component analysis, latent class modeling, latent and structural equation modeling, as well as other more general types of missing data analysis problems have been widely used by social scientists since their inception (Jackman, 2000; Raftery, 2001).

Statistical Pattern Recognition: Summary and Future Prospects

Most statistical pattern recognition systems consist of four major components. The first component is a feature selection and extraction stage where critical informational features about the data are identified for classification purposes. The second component is a probabilistic knowledge representation for representing the expected likelihood of particular features and the associated losses for making situation-specific decisions. The third component is a decision rule for making classification decisions that minimize an appropriate expected loss function. In many cases, a fourth component is required to learn the probabilistic knowledge representation from training data.

The computational goal of the learning process can be 'estimative' where parameters are estimated, or it can be 'predictive' where Bayesian methods are used to integrate out parameter effects. It is important to assess the performance of a statistical pattern recognition machine by using some combination of test data or cross-validation data. Important advantages of formulating behavioral and social models within a statistical pattern recognition framework include the availability of both existing and state-of-the-art methods for (1) making robust inferences in the presence of model misspecification (Golden, 2003; White, 1994) as well as model misspecification and missing data (Kashner et al., 2010), (2) comparing competing models (Golden, 2003), and (3) detecting the presence of model misspecification (Golden et al., 2013; White, 1994).

Statistical pattern recognition techniques are widely used in the fields of artificial intelligence and machine learning. In addition, such methods have been developed and widely used by social and behavioral scientists since their inception as well. Nevertheless, the field of statistical pattern recognition has so much to offer that methods of statistical pattern recognition remain underutilized in many fields including the social and behavioral sciences. For example, statistical pattern recognition methods such as linear regression or structural equation modeling are often applied without appropriate evaluations of generalization performance. As another example, missing data handling methods that are robust to the effects of model misspecification are rarely used. Continued efforts to more effectively utilize such methods not only will provide improved models in all areas of science and engineering (including the social and behavioral sciences) but also will additionally improve the quality of inferences for existing and future data.

See also: Artificial Intelligence: Connectionist and Symbolic Approaches; Artificial Intelligence: Genetic Programming; Artificial and Natural Computation; Auditory Scene Analysis: Computational Models; Cognitive Modeling: Connectionist Approaches; Computational Approaches to Model Evaluation; Concept Learning and Representation: Models; Constraint-Satisfaction Models; Face Recognition Models: Computational Approaches; Fourier Analysis; Information Theory; Knowledge Representation; Mathematical Psychology; Model Testing and Selection, Theory of; Motor Control Models: Learning and Performance; Optimal Control Theory; Psychometrics; Signal Detection Theory; Structural Equation Modeling.

Bibliography

Aitchison, J., 1975. Goodness of prediction fit. Biometrika 62 (3), 547–554. http://dx.doi.org/10.1093/biomet/62.3.547.

Chater, N., Tenenbaum, J.B., Yuille, A., 2006. Probabilistic models of cognition: conceptual foundations. Trends in Cognitive Science 10 (7), 287–291.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley, New York.

Golden, R.M., 1996. Mathematical Methods for Neural Network Analysis and Design. MIT Press, Cambridge, MA.

Golden, R.M., 2003. Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. Psychometrika 68 (2), 229–249. http://dx.doi.org/10.1007/bf02294799.

Golden, R.M., Henley, S.S., White, H., Kashner, T.M., 2013. New directions in information matrix testing: eigenspectrum tests. In: Chen, X., Swanson, N.R. (Eds.), Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis. Springer, New York, pp. 145–177.

Golden, R.M., Henley, S.S., White, H., Kashner, T.M., 2011. New Directions in Information Matrix Testing: Eigenspectrum Tests.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second edition. Springer.

Jackman, S., 2000. Estimation and inference via Bayesian simulation: an introduction to Markov Chain Monte Carlo. Journal of Political Science 44 (2), 375–404.

Jacobs, R.A., Kruschke, J.K., 2011. Bayesian learning theory applied to human cognition. Wiley Interdisciplinary Reviews: Cognitive Science 2 (1), 8–21. http://dx.doi.org/10.1002/wcs.80.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 4–36.

Kashner, T.M., Henley, S.S., Golden, R.M., Byrne, J.M., Keitz, S.A, Cannon, G.W., Chang, B.K., Holland, G.J., Aron, D.C., Muchmore, E.A., Wicker, A., White, H., 2010. Studying the Effects of ACGME duty hours limits on resident satisfaction: results from VA Learner's Survey. Academic Medicine 85, 1130–1139.

Nilsson, N.J., 1965. Learning Machines. McGraw-Hill, New York.

Patrick, E.A., 1972. Fundamentals of Pattern Recognition. Prentice-Hall, New Jersey.

Raftery, A.E., 2001. Statistics in sociology, 1950–2000: a selective review. Sociological Methodology 31 (1), 1–45.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81 (393), 82–86. http://dx.doi.org/10.2307/2287970.

White, H., 1994. Estimation, Inference, and Specification Analysis. Cambridge University Press, Cambridge.

Statistical Sufficiency

Nancy Reid, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the language of statistical theory, a statistic is a function of a set of data, and a sufficient statistic contains as much information about the statistical model as the original set of data. Statistical sufficiency has served as a powerful concept in the theory of inference in helping to clarify the role of models and data and in providing a framework for data reduction.

Introduction

Statistical sufficiency is a concept in the theory of statistical inference that is meant to capture an intuitive notion of summarizing a large and possibly complex set of data by relatively few summary numbers that carry the relevant information in the larger data set. For example, a public opinion poll might sample one or two thousand people, but will typically report simply the number of respondents and the percentage of respondents in each of a small number of categories. A study to investigate the effectiveness of a new treatment for weight loss may well report simply the average weight loss of a group of patients receiving the new treatment. This study would usually report as well some measure of the range of weight loss observed; variability in this setting would be a crucial piece of information for evaluating the effectiveness of the new treatment.

In order to make precise the vague notion of an informative data summary, it is necessary to have a means of defining what we mean by relevant information. This in turn is highly dependent on the use we wish to make of the data. The formal mechanism for doing this is to use a *statistical model* to represent an idealized version of the problem we are studying. Statistical sufficiency and information are defined relative to this model. A statistical model is a family of probability distributions, the central problem of statistical inference being to identify which member of the family generated the data currently of interest. The basic concepts of statistical models and sufficiency were set out in Fisher (1922, 1925) and continue to play a major role in the theory of inference.

For example, we might in an idealized version of the weight loss problem assume that the potential weight loss can be described by a model in the family of *normal distributions*. This specifies a mathematical formula for the probability that the weight loss of an individual subject falls in any possible interval, and the mathematical formula depends on two unknown *parameters*, the mean and standard deviation. If we are willing to further assume that the same statistical model applies to all the potential recipients of the weight loss treatment, then the data collected in a careful study will enable us to make an inference on the two parameters that depends only on two data summaries: the average weight loss in the data and the standard deviation of the weight loss in the data.

The normal distribution is also referred to as the 'bell curve,' but there are in fact a family of bell curves, each differing from the other by a simple change of mean, which fixes the location of the bell curve, and standard deviation, which fixes the scale of the bell curve. Standardized tests for measuring IQ are designed so that the scores follow a bell curve model with parameter values for the mean and standard deviation of 100 and 15, respectively. The parameters for the potential weight loss under our hypothetical new treatment are unknown, and the role of the data collected on the treatment is to shed some light on their likely values. Should the weight loss bell curve turn out to be centered at or near zero, one would be wise not to invest too heavily in the new company's stock just yet. As long as we are willing to accept the bell curve model for the weight loss data, the information about that model is contained in two numbers: the average and the standard deviation of the weight loss recorded in our data set.

An idealized model that lies behind most media reporting of polls is a *Bernoulli* or *binomial* model, in which we assume that potential respondents will choose one of two categories with a fixed but unknown probability, that respondents choose categories independently of each other, and with the same, constant probability. In such an idealized version of the problem, we have complete information about this unknown probability from the number of responses and the proportion of responses in each of the two categories.

These two examples are by construction highly idealized, and it is easy to think quickly of a number of complexities that have been overlooked. An educated reader of a report on the weight loss study would want to know how the subjects were selected to receive the new treatment, whether the subjects were in any way representative of a population of interest to that reader, whether any comparison has been made with alternative, possibly conventional, treatments, the length of time over which the study was conducted, and so on. Public opinion polls typically offer more than two categories of response, even to a 'yes/no' question, including something equivalent to 'don't know,' or 'won't say,' and also often break down the results by gender, geographic area, or some other classification. Randomized selection of respondents by computerized telephone polls is much different than a reader response survey often conducted by magazines and media outlets. These and a wide variety of other issues are studied in the fields of experimental design and sample survey design, by statisticians and researchers in a variety of subject matter areas, from economics to geophysics. *See* Observational Studies: Overview; Sample Surveys: The Field and Methods; Experimental Design: Overview.

Statistical sufficiency is concerned with a much narrower issue, that of formalizing the notion that for a given family of models, a complex set of data can be summarized by a smaller

 http://dx.doi.org/10.1016/B978-0-08-097086-8.42178-1

collection of numbers, and lead to the same statistical inference. The following sections describe statistical sufficiency with reference to families of models and related concepts in the theory of statistics.

Statistical Models and Sufficiency

Basic Notation

We describe a statistical model by a *random variable* Y and a family of *probability distributions* for Y, $\mathcal{F}$. A random variable is a variable that takes values on a *sample space*, and a probability distribution describes the probability of observing a particular value or range of values in the sample space. The sample space is in a great many applications either *discrete*, taking values on a countable set such as (some subset of) the nonnegative integers, or *continuous* taking values on (some subset of) the real line $\mathbb{R}$ or p-dimensional Euclidean space $\mathbb{R}^p$. A typical data set will be a number of realizations of Y, and the goal is to use these observations to deduce which member of $\mathcal{F}$ generated these realizations. The members of $\mathcal{F}$ are often conveniently indexed by the probability mass function or probability density function $f(y)$, where

$$f(y) = \text{Prob}(Y = y) \qquad [1]$$

if the sample space is discrete, and

$$f(y) = \lim_{h \to 0} h^{-1}\{\text{Prob}(Y \le y + h) - \text{Prob}(Y \le y)\} \qquad [2]$$

if the sample space is continuous. If Y is a vector then Prob $(Y \le y)$ is defined component-wise as Prob $(Y_1 \le y_1, \ldots, Y_p \le y_p)$.

Very often the mathematical form of the probability distributions in $\mathcal{F}$ is determined up to a finite number of unknown constants called parameters, so we write $\mathcal{F}_\theta = \{f(\bullet;\theta); \theta \in \Theta\}$, where Θ is a set of possible values for θ, such as the real numbers $\mathbb{R}$, or real-valued vectors of fixed length p, $\mathbb{R}^p$, or [0, 1], and so on. The problem of using data to make an inference about the probability distribution is now the problem of making inference about the parameter θ, and is referred to as parametric inference. Sufficiency also plays a role in nonparametric inference as well, and we illustrate this at the end of this section.

We assume in what follows that the random variable $Y = (Y_1, \ldots, Y_n)$ is a vector of length n with observed value $y = (y_1, \ldots, y_n)$. A *statistic* is a function $T = t(Y)$ whose distribution can be computed from the distribution of Y. Examples of statistics include the average, $T_1 = n^{-1}\Sigma Y_i$, the sample variance $T_2 = (n-1)^{-1}\Sigma(Y_i - T_1)^2$, the standard deviation $T_3 = T_2^{1/2}$, the range $T_4 = \max Y_i - \min Y_i$, and so on. The probability distribution of any statistic is obtained from the probability distribution for Y by a mathematical calculation.

Sufficiency and Minimal Sufficiency

A statistic $T = t(Y)$ is sufficient for θ in the family of models $\mathcal{F} = \{f(y;\theta); \theta \in \Theta\}$ if and only if its conditional distribution

$$f_{Y|T}(y|t) \qquad [3]$$

does not depend on θ.

Example 1. Suppose $Y_1, \ldots Y_n$ are independent and identically distributed with probability mass function

$$f(y_i;\theta) = \theta^{y_i}(1-\theta)^{1-y_i}, \quad y_i = 0, 1; \quad \theta \in [0, 1] \qquad [4]$$

which can be used to model the number of successes in n independent Bernoulli trials, when the probability of success is θ. The joint density for $Y = (Y_1, \ldots, Y_n)$ is

$$f(y;\theta) = \prod \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} \qquad [5]$$

and the marginal density for $T = \Sigma Y_i$ is

$$f(t;\theta) = \sum_{y:\Sigma y_i = t} \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i} = \theta^t(1-\theta)^{n-t}c(t,n) \qquad [6]$$

where $c(t,n) = n!/t!\,(n-t)!$ is the number of vectors y of length n with t ones and $n-t$ zeroes. The conditional density of Y given T is the ratio of [5] and [6] and is thus free of θ.

The intuitive notion that the definition is meant to encapsulate is that once the sufficient statistic T is observed no further information about the θ is available from the original data Y.

The definition of sufficiency in terms of the conditional distribution is not very convenient to work with, especially in the case of continuous sample spaces where some care is needed in the definition of a conditional distribution. The following result is more helpful:

Factorization theorem: A statistic $T = t(Y)$ is sufficient for θ in the family of models $\mathcal{F} = \{f(y;\theta); \theta \in \Theta\}$ if and only if there exist functions $g(t;\theta)$ and $h(y)$ such that for all $\theta \in \Theta$

$$f(y;\theta) = g(t;\theta)h(y) \qquad [7]$$

Example 2. Let $Y_1, \ldots Y_n$ be independent, identically distributed from the normal distribution with mean μ and standard deviation σ:

$$f(y_i;\mu,\sigma) = \frac{1}{\sqrt{(2\pi)}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \qquad [8]$$

The joint density of $Y = (Y_1, \ldots Y_n)$ is

$$f(y;\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right\},$$
$$= \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\bar{y})^2 + \frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\}\left(\frac{1}{\sqrt{2\pi}}\right)^n \qquad [9]$$

where $\bar{y} = n^{-1}\Sigma y_i$ is the sample average, $\theta = (\mu, \sigma)$, and $\Theta = \Re \times \Re^+$. From [9] we see that $T = \{\bar{Y}, \sum(Y_i - \bar{Y})^2\}$ is sufficient for θ, with $g(t;\theta)$ identified with all but the final factor in [9].

Strictly speaking, there are a number of other sufficient statistics in both the examples above. In Example 2, $(\sum_1^4 Y_i, \sum_5^n Y_i, \sum_1^7 (Y_i - \bar{Y})^2, \sum_8^n (Y_i - \bar{Y})^2)$ is also sufficient, as is, trivially, the original vector Y itself. Clearly $(\bar{Y}, \sum_1^n (Y_i - \bar{Y})^2)$ is in some sense 'smaller' than these other sufficient statistics, and to be preferred for that reason. A *minimal sufficient statistic* is defined to be a function of every other sufficient statistic. One-to-one functions of the minimal sufficient statistic are also minimal sufficient and are not distinguished, so in Example 2, $(\sum_1^n Y_i, \sum_1^n Y_i^2)$ is also minimal sufficient. Any statistic defines a partition of the sample space in which all sample points leading to the same value of the statistic fall in the

same partition, and the minimal sufficient statistic defines the coarsest possible partition.

Since the central problem of parametric statistical inference is to reason from the data back to the parameter θ, extensive use is made of the *likelihood function*, which is (proportional to) the joint density of the data, regarded as a function of θ:

$$L(\theta; y) = c(y) f(y; \theta) \qquad [10]$$

See Likelihood in Statistics. The factorization theorem suggests, and it can be proved, that the minimal sufficient statistic in the family of models $\mathcal{F}$ is the likelihood statistic $L(\cdot; Y)$. This result is easily derived from the factorization theorem if Θ is a finite set, but extending it to cases of more practical interest involves some mathematical intricacies (Barndorff-Nielsen, 1978: Chapter 4; Fraser and Naderi, 2007). *See* Likelihood Function and Maximum Likelihood Estimation.

Example 3. To illustrate the concept of sufficiency in a nonparametric setting, suppose that $Y_1, \ldots, Y_n$ are independent, identically distributed on $\mathbb{R}$ with probability density function f; i.e., $\mathcal{F}$ consists of all probability distributions which have densities with respect to Lebesgue measure. Denote by $Y_{(\cdot)}$ the vector of ordered values $Y_{(\cdot)} = (Y_{(1)} \le Y_{(2)} \le \ldots \le Y_{(n)})$. Then $Y_{(\cdot)}$ is sufficient for $\mathcal{F}$, as

$$f_{Y|Y(\cdot)}(y|y_{(\cdot)}) = \begin{cases} \dfrac{1}{n!} & \text{if y is a permutation of } y_{(\cdot)} \\ 0 & \text{otherwise} \end{cases} \qquad [11]$$

is the same for all members of $\mathcal{F}$.

Example 4. Let $Y_1, \ldots, Y_n$ be independent, identically distributed according to the uniform distribution on $(\theta - 1, \theta + 1)$, where $\theta \in \mathbb{R}$. The order statistic $Y_{(\cdot)}$ is sufficient, as this is a special case of Example 3 above, but is not minimal sufficient. The minimal sufficient statistic is determined from the likelihood function:

$$L(\theta; y) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left(\frac{1}{2}\right) & \text{if } y_{(n)} - 1 \le \theta \le y_{(1)} + 1 \\ 0 & \text{otherwise} \end{cases} \qquad [12]$$

from which we see that the minimal sufficient statistic is $T = (Y_{(1)}, Y_{(n)})$, as this is a one to one function of the likelihood statistic.

Exponential Families

A random vector $Y \in \mathbb{R}^p$ is said to be distributed according to an exponential family of dimension k if the probability mass or density function can be expressed in the form

$$f(y; \theta) = \exp\left\{\varphi(\theta)^T t(y) - c(\theta) - d(y)\right\} \qquad [13]$$

for some functions $\varphi(\theta) = \{\varphi_1(\theta), \ldots, \varphi_k(\theta)\}$, $t(y) = \{t_1(y), \ldots, t_k(y)\}$, $c(\theta)$, and $d(y)$.

Many common distributions can be expressed in exponential family form, including Example 1 with $\varphi(\theta) = \log\{\theta/(1-\theta)\}$, $t(y) = \sum y_i$ and Example 2 with

$$\varphi(\theta) = \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma^2}\right), \quad t(y) = \left(\sum y_i, \sum y_i^2\right) \qquad [14]$$

Other examples of exponential families include the Poisson, geometric, negative binomial, multinomial, exponential, gamma, and inverse Gaussian distributions. Generalized linear models are regression models built on exponential families that have found wide practical application. *See* Analysis of Variance; Generalized Linear Models.

From [13] we can see that exponential families are closed under independent, identically distributed sampling, and that the minimal sufficient statistic for θ is the k-dimensional vector $T = \sum t(Y_i)$. Exponential families are said to permit a *sufficiency reduction* under sampling; we can replace the original n observations by a vector of length k. Possibly for this reason exponential families are widely used as models in applied work.

A partial converse of this is also true, that if a finite dimensional minimal sufficient statistic exists in independent sampling from the same density, and the range of values over which the density is positive does not depend on θ, then that density must be of exponential family form.

Versions of the material in this appear in most mathematical statistics textbooks. Suggested references, roughly in order of increasing difficulty, are Azzalini (1996), Casella and Berger (1990), Lehmann and Casella (1998), Cox and Hinkley (1974), Pace and Salvan (1997), and Barndorff-Nielsen and Cox (1994). Generalized linear models are treated in Dobson (2001) and McCullagh and Nelder (1989).

Some Related Concepts

Ancillary Statistics

A companion concept for the reduction of data to a number of smaller quantities is that of *ancillarity*. A statistic $A = a(Y)$ is said to be ancillary for θ in the family of models $\mathcal{F}$ if the marginal distribution of A

$$f_A(a) = \int_{\{y; a(y) = a\}} f(y; \theta) \mathrm{d}y \qquad [15]$$

is free of θ. If every other ancillary statistic is a function of a given ancillary statistic, the given statistic is a maximal ancillary statistic.

Example 4 cont'd: In independent sampling from the $U(\theta, \theta + 1)$ distribution, $T = (Y_{(1)}, Y_{(n)})$ is a minimal sufficient statistic and $A = Y_{(n)} - Y_{(1)}$ is a maximal ancillary statistic.

If an ancillary statistic exists all the information about θ is in the conditional distribution of Y, given the ancillary statistic. If a minimal sufficient statistic exists, all the information about θ is in the marginal distribution of T. These two data reductions, to A or to T, are complementary, and can sometimes be combined.

Example 5: Let $Y_1, \ldots, Y_n$ be independent, identically distributed from the one parameter *location* family $f(y_i; \theta) = f_0(y_i - \theta)$, $\theta \in \mathbb{R}$, where $f_0(\cdot)$ is known. Then the vector of residuals $A = (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ is ancillary. An extension of this to the location scale family is possible: let $f(y_i; \theta) = \sigma^{-1} f_0\{\sigma^{-1}(y_i - \mu)\}$, then

$$A = \left\{(Y_1 - \overline{Y})/S, \ldots, (Y_n - \overline{Y})/S\right\} \qquad [16]$$

where $S^2 = \sum (Y_i - \overline{Y})^2$ is a maximal ancillary. The normal distribution of Example 4 is a location-scale model, with minimal sufficient statistic $T = (\overline{Y}, S^2)$, which in this case is independent of A.

A result due to Basu (1964) is that a sufficient statistic that is boundedly complete is independent of any ancillary statistic. Bounded completeness of a sufficient statistic is a technical condition that is occasionally useful in obtaining optimality results, particularly with respect to minimum variance unbiased estimation, a topic that is somewhat out of fashion.

The importance for the theory of statistics in concepts of dimension reduction such as ancillarity and sufficiency is that inference for a k-dimensional parameter is encapsulated by a probability distribution on a k-dimensional space. This immediately provides tests of significance, confidence intervals, and point estimates for the unknown parameter. From this point of view, it now appears that ancillarity is possibly a more central notion to the theory of statistics, although sufficiency is widely regarded as more natural and intuitive.

It should be noted that inference from a Bayesian point of view automatically provides a k-dimensional distribution for inference for a k-dimensional parameter, which is the posterior distribution given the observed data y. This is achieved at the expense of constructing a probability model for the unknown parameter θ. *See* Bayesian Statistics.

Asymptotic Sufficiency

The normal distribution is widely used for modeling data from a variety of sources, but is also very important in the theory of inference in providing approximate inferences, when exact inferences may be unavailable. While the likelihood function is equivalent to the minimal sufficient statistic, a special role in approximate inference is played by some functions computed from the likelihood function.

The maximum likelihood estimator of θ in the model $f(y;\theta)$ is defined by.

$$\sup_{\theta} L(\theta;y) = L\left(\widehat{\theta};y\right) \qquad [17]$$

and the observed and expected information functions for θ are

$$j(\theta) = -\partial^2\left[\log\{L(\theta;y)\}\right]/\partial\theta\partial\theta', \quad i(\theta) = \mathrm{E}\{j(\theta)\} \qquad [18]$$

Under regularity conditions on the family of models for $Y=(Y_1,\ldots,Y_n)$, we have the following asymptotic result for a scalar parameter θ as $n \to \infty$:

$$\left(\widehat{\theta}-\theta_0\right)i^{1/2}(\theta_0) \xrightarrow{d} N(0,1) \qquad [19]$$

which means that for sufficiently large n, the distribution of the maximum likelihood estimator is arbitrarily close to a normal distribution with mean zero and variance $i^{-1}(\theta_0)$, under sampling from $f(y;\theta_0)$. A similar result is available for vector parameters. Usually in practice, the asymptotic variance of $\widehat{\theta}$ will be estimated by $j^{-1}(\widehat{\theta})$.

The maximum likelihood estimator has by construction the same dimension as the unknown parameter θ, so [19] provides a means of constructing an inference for θ using the normal approximation. This suggests that a notion of *asymptotic sufficiency* could be formalized, the maximum likelihood estimator having this property. This formalization relies on showing that the likelihood function is asymptotically of normal form, with minimal sufficient statistic $\widehat{\theta}$; see Fraser (1976: Chapter 8). The argument is very similar to that showing that the posterior distribution for θ is asymptotically normal. More generally, in models with smooth likelihood functions, a Taylor series expansion of the likelihood function around $\widehat{\theta}$ suggests a refinement of asymptotic sufficiency: the set of statistics $\widehat{\theta}$ and second and higher order derivatives of the likelihood function evaluated at $\widehat{\theta}$ form a set of approximately sufficient statistics in a certain sense (Barndorff-Nielsen and Cox, 1994: Chapter 7).

In general models, the maximum likelihood estimator is not the minimal sufficient statistic, although it is always a (many-to-one) function of the minimal sufficient statistic. Much of the recent development of the parametric theory of inference has been concerned with combining sufficiency or approximate sufficiency with ancillarity or an appropriate notion of approximate ancillarity, to effect a dimension reduction with minimal loss of information. In general models if the maximum likelihood estimator is used in place of the minimal sufficient statistic, information is lost. The idea of recovering this information, via conditioning on an ancillary statistic, was put forth in Fisher (1934). With some exceptions this idea was not fully exploited in the statistical literature until about 1980, when it was discovered that an asymptotic version of Fisher's argument holds in very general situations. The past 20 years have seen continued development and refinement of Fisher's original idea. A brief overview is given in Reid (2000), and the books Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997) give detailed accounts.

Inference in the Presence of Nuisance Parameters

Many statistical models have vector valued parameters θ, but often some components or scalar valued functions of θ are of particular interest. For example, in evaluating the hypothetical new weight loss treatment under a normal model, the mean parameter μ is in the first instance of more interest than the standard deviation σ. In studying results from a poll, we may wish to introduce parameters describing various aspects of the population, such as gender, age group, income bracket, and so on, but the main purpose of the poll is to construct an inference about the proportion likely to vote for our candidate, for example.

The idealized modeling of these types of situations is to describe θ as consisting of two components $\theta=(\psi,\ \lambda)$, of *parameters of interest*, ψ, and *nuisance parameters*, λ. An extension of this would be to define $\psi=g(\theta)$ as determined by a set of constraints on θ, but only the component version will be considered here.

There are in the literature several different definitions of sufficiency for a parameter of interest ψ in the presence of a nuisance parameter λ. To illustrate some of the issues we consider again the normal distribution with $\theta=(\mu,\ \sigma)$, $T = (\overline{Y}, S^2)$. We have

$$f_T(t;\mu,\sigma) = f_{\overline{Y}}(\overline{y};\mu,\sigma)f_{S^2|Y}\left(s^2;\sigma^2\right), = f_{\overline{Y}}(\overline{y};\mu,\sigma)f_{S^2}\left(s^2;\sigma^2\right) \qquad [20]$$

where $f_{\overline{Y}}(\cdot)$ is the density of a normal distribution with mean μ and standard deviation σ/n, and $f_{S^2}(\cdot)$ is proportional to that of a chi-squared distribution on $n-1$ degrees of freedom. As the conditional distribution of S^2, given $\overline{Y}$ is free of μ, we could define $\overline{Y}$ as sufficient for μ, by analogy to [3]. However, it is not sufficient for μ in the sense of providing everything we need for

inference about μ, since its distribution depends also on σ. In this latter sense S^2 is sufficient for σ.

There have been a number of attempts to formalize the notions of partial sufficiency and partial ancillarity, but none has been particularly successful. However, the notion of dimension reduction through exact or approximate sufficiency and ancillarity has played a very important role in the theory of parametric inference.

Sufficiency in the presence of nuisance parameters is considered in Barndorff-Nielsen (1978: Chapter 4), Jorgensen (1997), and Bhapkar (1991). A discussion of partitions like [20] and the role of conditional and marginal inference is given in Reid (1995). *See* Likelihoods that Eliminate Nuisance Parameters.

Complex Models

In some applications, the model is too complex to enable a likelihood function to be constructed; it may entail a large number of integrals over latent random variables, for example. Several approaches to inference in such settings involve computing some version of approximate likelihoods, either by simulation, or by ignoring some higher-order dependencies. Two examples of these approaches are the method of indirect inference (Gouriéroux and Montfort, 1996) and approximate Bayesian computation (ABC) (Marin et al., 2011); the latter is focused on Bayesian inference, but may be useful for approximate likelihood inference as well. Since the likelihood map is sufficient, this might suggest that there is no role for the concept of sufficiency or approximate sufficiency in such settings, but Robert et al. (2011) show that the ABC method will not be consistent for model selection unless it is built on sufficient statistics for the model of interest. Indirect inference also uses sufficient statistics, although for an auxiliary model, as a step in simulating the likelihood of the model of interest. Connections between indirect inference and ABC methods are mentioned in Fearnhead and Prangle (2011) and a subject of ongoing research.

Conclusion

A sufficient statistic is defined relative to a statistical model, usually a parametric model, and provides all the information in the data about that model or the parameters of that model. In many cases, the sufficient statistic provides a substantial reduction of complexity or dimension from the original data, and as a result the concept of sufficiency makes the task of statistical inference simpler. The related concept of ancillarity has a similar function. Recent developments in likelihood inference have emphasized the construction of approximately sufficient and approximately ancillary statistics in cases where further dimension reduction is needed after invoking sufficiency and ancillarity. Excellent textbook references for the definitions and examples include Casella and Berger (1990) and Azzalini (1996). More advanced accounts of recent work are given in Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997).

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Approximations; Experimental Design: Overview; Likelihood in Statistics; Nonparametric Regression; Observational Studies: Overview; Order Statistics.

Bibliography

Azzalini, A., 1996. Statistical Inference. Chapman and Hall, London.

Barndorff-Nielsen, O.E., 1978. Information and Exponential Families in Statistical Theory. Wiley and Sons, New York.

Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall, London.

Basu, D., 1964. Recovery of ancillary information. Sankhya A 26, 3–16.

Bhapkar, V.P., 1991. Sufficiency, ancillarity and information in estimating functions. In: Godambe, V.P. (Ed.), Estimating Functions. Oxford University Press, Oxford.

Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth/Brooks Cole, Pacific Grove CA.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.

Dobson, A., 2001. An Introduction to Generalized Linear Models, second ed. Chapman and Hall, London.

Fearnhead, P., Prangle, D., 2011. Constructing summary statistics for approximate Bayesian computation. Journal of the Royal Statistical Society Series B 74, 1–28.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions Royal Society London, Series A 222, 309–368 (Reprinted in: Kotz, S., Johnson, N.L. (Eds.), 1992. Breakthroughs in Statistics. Springer-Verlag, New York.).

Fisher, R.A., 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700–725.

Fisher, R.A., 1934. Two new properties of mathematical likelihood. Proceedings of the Royal Society Series A 144, 285–307.

Fraser, D.A.S., 1976. Probability and Statistics. Wadsworth, Belmont CA.

Fraser, D.A.S., Naderi, A., 2007. Minimal sufficient statistics emerge from the observed likelihood function. International Journal of Statistical Science 6, 55–61.

Gouriéroux, C., Montfort, A., 1996. Simulation-Based Econometric Methods. Oxford University Press, Oxford.

Jorgensen, B., 1997. The rules of conditional inference: is there a universal definition of non-formation. Journal of the Italian Statistical Society 3, 355–384.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation, second ed. Springer-Verlag, New York.

Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J., 2011. Approximate Bayesian computational methods. Statistics and Computing 22, 1167–1180.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

Pace, L., Salvan, A., 1997. Principles of Statistical Inference from a Neo-Fisherian Perspective. World Scientific, Singapore.

Reid, N., 1995. The roles of conditioning in inference (with discussion). Statistical Science 10, 138–157.

Reid, N., 2000. Likelihood. Journal of the American Statistical Association 95, 1335–1340.

Robert, C.P., Cornuet, J.-M., Marin, J.-M., Pillai, N.S., 2011. Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences 108, 15112–15117.

Statistics, History of

Alain Desrosières, INSEE, Paris, France

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15080–15085, © 2001, Elsevier Ltd.

Abstract

The history of statistics can be told in several ways: as a quantitative description of the social world, or as a branch of applied mathematics making it possible to build models in any area featuring large numbers. There are many historical connections between these narratives. This history began in the eighteenth century, with German '*Statistik*' and English *political arithmetic*. In the nineteenth century Quetelet established a relation between the probabilistic theory of the 'average man' and the gathering of official statistics by governments, with a large use of the Gaussian distribution and of the relative stability of macrosocial phenomena. In the 1890s, Francis Galton and Karl Pearson created a mathematical theory of statistical distribution, and introduced the ideas of variance, regression, correlation, and chi-square test. That theory was then developed by Ronald Fisher, Jerzy Neyman, and Egon Pearson, with the inferential statistics theory, giving a strong probabilistic foundation to estimation, tests, maximum likelihood, and sampling theory. Simultaneously, official statistical bureaux, compiling administrative statistics, and organising censuses and sampling surveys, were also developed. These institutions, often perceived only as suppliers of data assumed to reflect reality, are also places where this reality is instituted through co-constructed operations of social representations, public action, and statistical measurement.

The word 'statistics' today has several different meanings. For the public, and even for many people specializing in social studies, it designates numbers and measurements relating to the social world: population, gross national product, and unemployment, for instance. For academics in 'Statistics Departments,' however, it designates a branch of applied mathematics making it possible to build models in any area featuring large numbers, not necessarily dealing with society. History alone can explain this dual meaning. Statistics appeared at the beginning of the nineteenth century as meaning a 'quantified description of human–community characteristics.' It brought together two previous traditions: that of German '*Statistik*' and that of English *political arithmetic* (Lazarsfeld, 1977).

When it originated in Germany in the seventeenth and eighteenth centuries, the *Statistik* of Hermann Conring (1606–81) and Gottfried Achenwall (1719–79) was a means for classifying the knowledge needed by kings. This 'science of the state' included history, law, political science, economics, and geography, that is, a major part of what later became the subjects of 'social studies,' but presented from the point of view of their utility for the state. These various forms of knowledge did not necessarily entail quantitative measurements: they were basically associated with the description of specific territories in all their aspects. This territorial connotation of the word 'statistics' would subsist for a long time in the nineteenth century.

Independently of the German *Statistik*, the English tradition of political arithmetic had developed methods for analyzing numbers and calculations, on the basis of parochial records of baptisms, marriages, and deaths. Such methods, originally developed by John Graunt (1620–74) in his work on 'bills of mortality,' were then systematized by William Petty (1627–87). They were used, among others, to assess the population of a kingdom and to draw up the first forms of life insurance. They constitute the origin of modern demography.

Nineteenth-century 'statistics' were therefore a fairly informal combination of these two traditions: taxonomy and numbering. At the beginning of the twentieth century, it further became a mathematical method for the analysis of facts (social or not) involving large numbers and for inference on the basis of such collections of facts. This branch of mathematics is generally associated with the *probability theory*, developed in the seventeenth and eighteenth centuries, which, in the nineteenth century, influenced many branches of both the social and natural sciences (Gigerenzer et al., 1989). The word 'statistics' is still used in both ways today and both of these uses are still related, of course, insofar as quantitative social studies use, in varied proportions, inference tools provided by mathematical statistics. The probability calculus, on its part, grounds the credibility of statistical measurements resulting from surveys, together with random sampling and confidence intervals.

The diversity of meanings of the word 'statistics,' maintained until today, has been heightened by the massive development, beginning in the 1830s, of *bureaus of statistics*, administrative institutions distinct from universities, in charge of collecting, processing, and transmitting quantified information on the population, the economy, employment, living conditions, etc. Starting in the 1940s, these bureaus of statistics became important 'data' suppliers for empirical social studies, then in full growth. Their history is therefore an integral part of these sciences, especially in the second half of the twentieth century, during which the mathematical methods developed by university statisticians were increasingly used by the so-called 'official' statisticians. Consequently the 'history of statistics' is a crossroads history connecting very different fields, which are covered in other articles of this encyclopedia: general problems of the quantification' of social studies (Theodore Porter), mathematical 'sampling' methods (Stephen E. Fienberg and J. M. Tanur), ' survey research' (Martin Bulmer), ' demography' (Simon Szreter), 'econometrics,' etc. The history of these different fields was the object of much research in the 1980s

and 1990s, some examples of which are indicated in the Bibliography. Its main interest is to underscore the increasingly closer connections between the so-called 'internal' dimensions (history of the technical tools), the 'external' ones (history of the institutions), and those related to the construction of social studies 'objects,' born from the interaction between the three foci constituted by university research, administrations in charge of 'social problems,' and bureaus of statistics. This 'co-construction' of objects makes it possible to join historiographies that not long ago were distinct. Three key moments of this history will be mentioned here: Adolphe Quetelet and the average man (1830s), Karl Pearson and correlation (1890s), and the establishment of large systems for the production and processing of statistics.

Quetelet, the Average Man, and Moral Statistics

The cognitive history of statistics can be presented as that of the tension and sliding between two foci: the *measurement of uncertainty* (Stigler, 1986), resulting from the work of eighteenth-century astronomers and physicists, and the *reduction of diversity*, which will be taken up by social studies. Statistics is a way of taming chance (Hacking, 1990) in two different ways: chance and uncertainty related to protocols of observation, chance and dispersion related to the diversity and the indetermination of the world itself. The Belgian astronomer and statistician Adolphe Quetelet (1796–1874) is the essential character in the transition between the world of 'uncertain measurement' of the probability proponents (Carl Friedrich Gauss, Pierre Simone de Laplace) and that of the 'regularities' resulting from diversity, thanks to his having transferred, around 1830, the concept of average from the natural sciences to the human sciences, through the construction of a new being, the average man.

As early as the eighteenth century, specificities appeared from observations in large numbers: drawing balls out of urns, gambling, successive measurements of the position of a star, sex ratios (male and female births), or the mortality resulting from preventive smallpox inoculation, for instance. The radical innovation of this century was to connect these very different phenomena thanks to the common perspective provided by the 'law of large numbers' formulated by Jacques Bernoulli in 1713. If draws from a constant urn containing white and black balls are repeated a large number of times, the observed share of white balls 'converges' toward that actually contained by the urn. Considered by some as a mathematical theorem and by others as an experimental result, this 'law' was at the crossroads of the two currents in epistemological science: one 'hypothetical deductive,' the other 'empirical inductive.'

Beginning in the 1830s, the statistical description of 'observations in large numbers' became a regular activity of the state. Previously reserved for princes, this information henceforth available to 'enlightened men' was related to the population, to births, marriages, and deaths, to suicides and crimes, to epidemics, to foreign trade, schools, jails, and hospitals. It was generally a by-product of administration activity, not the result of special surveys. Only the population census, the showcase product of nineteenth-century statistics, was the object of regular surveys. These statistics were published in volumes with heterogeneous contents, but their very existence suggests that the characteristics of society were henceforth a matter of scientific law and no longer of judicial law, that is, of observed regularity and not of the normative decisions of political power. Quetelet was the man who orchestrated this new way of thinking the social world. In the 1830s and 1840s, he set up administrative and social networks for the production of statistics and established – until the beginning of the twentieth century – how statistics were to be interpreted.

This interpretation is the result of the combination of two ideas developed from the law of large numbers: the generality of 'normal distribution' (or, in Quetelet's vocabulary: the 'law of possibilities') and the regularity of certain yearly statistics. As early as 1738, Abraham de Moivre, seeking to determine the convergence conditions for the law of large numbers, had formulated the mathematical expression of the future 'Gaussian law' as the limit of a binomial distribution. Then Laplace (1749–1827) had shown that this law constituted a good representation of the distribution of measurement errors in astronomy, hence the name that Quetelet and his contemporaries also used to designate it: the law of errors (the expression 'normal law,' under which it is known today, would not be introduced until the late nineteenth century by Karl Pearson).

Quetelet's daring intellectual masterstroke was to bring together two forms: on the one hand the law of errors in observation, and on the other, the law of distribution of certain body measurements of individuals in a population, such as the height of conscripts in a regiment. The likeness of the 'Gaussian' looks of these two forms of distribution justified the invention of a new being with a promise of notable posterity in social studies: the average man. Thus, Quetelet restricted the calculation and the legitimate use of averages to cases where the distribution of the observations had a Gaussian shape, analogous to that of the distribution of the astronomical observations of a star. Reasoning on that basis, just as previous to this distribution there was a real star (the 'cause' of the Gaussian-shaped distribution), previous to the equally Gaussian distribution of the height of conscripts there was a being of a reality comparable to the existence of the star. Quetelet's average man is also the 'constant cause,' previous to the observed controlled variability. He is a sort of model, of which specific individuals are imperfect copies.

The second part of this cognitive construction, which is so important in the ulterior uses of statistics in social studies, is the attention drawn by the 'remarkable regularity' of series of statistics, such as those of marriages, suicides, or crimes. Just as series of draws from an urn reveal a regularity in the observed frequency of white balls, the regularity in the rates of suicide or crime can be interpreted as resulting from series of draws from a population, some of the members of which are affected with a 'propensity' to suicide or crime. The average man is therefore endowed not only with physical attributes but also 'moral' ones, such as these propensities. Here again, just as the average heights of conscripts are stable, whereas individual heights are dispersed, crime or suicide rates are just as stable, whereas these acts are eminently individual and unpredictable. This form of statistics, then called 'moral,' signaled the beginning of sociology, a science of society radically distinct from a science of the

individual, such as psychology (Porter, 1986). Quetelet's reasoning would ground the one developed by Durkheim in *Suicide: A Study in Sociology* (1897).

This way of using statistical regularity to back the idea of the existence of a society ruled by specific 'laws,' distinct from those governing individual behavior, dominated nineteenth-century, and, in part, twentieth-century social studies. Around 1900, however, another approach appeared, this one centered on two ideas: the distribution (no longer just the average) of observations, and the correlation between two or several variables, observed in individuals (no longer just in groups, such as territories).

Distribution, Correlation, and Causality

This shift of interest from the average individual to the distributions and hierarchies among individuals, was connected to the rise, in late-century Victorian England, of a eugenicist and hereditarian current of thought inspired from Darwin (MacKenzie, 1981). Its two advocates were Francis Galton (1822–1911), a cousin of Darwin, and Karl Pearson (1857–1936). In their attempt to measure biological heredity, which was central to their political construction, they created a radically new statistics tool that made it possible to conceive partial causality. Such causality had been absent from all previous forms of thought, for which A either is or is not the cause of B, but cannot be so somewhat or incompletely. Yet Galton's research on heredity led to such a formulation: the parents' height 'explains' the children's, but does not entirely 'determine' it. The taller fathers are, the taller are their sons *on average*, but, for a father's given height, the sons' height dispersion is great. This formalization of heredity led to the two related ideas of regression and correlation, later to be extensively used in social studies as symptoms of causality.

Pearson, however, greatly influenced by the antirealist philosophy of the physicist Ernst Mach, challenged the idea of 'causality,' which according to him was 'metaphysical,' and stuck to the one of 'correlation,' which he described with the help of 'contingency tables' (Pearson, 1911: Chap. 5). For him, scientific laws are only summaries, brief descriptions in mental stenography, abridged formulas, a condensation of perception routines for future use and forecasting. Such formulas are the limits of observations that never perfectly respect the strict functional laws. The correlation coefficient makes it possible to measure the strength of the connection, between zero (independence) and one (strict dependence). Thus, in this conception of science associated by Pearson with the budding field of mathematical statistics, the reality of things can only be invoked for pragmatic ends and provided that the 'perception routines' are maintained. Similarly, 'causality' can only be insofar as it is a proven correlation, therefore predictable with a fairly high probability. Pearson's pointed formulations would constitute, in the early twentieth century, one of the foci of the epistemology of statistics applied to social studies. Others, in contrast, would seek to give new meaning to the concepts of reality and causality by defining them differently. These discussions were strongly related to the aims of the statistical work, strained between scientific knowledge and decisions.

Current mathematical statistics proceed from the works of Karl Pearson and his successors: his son Egon Pearson (1895–1980), the Polish mathematician Jerzy Neyman (1894–1981), the statistician pioneering in agricultural experimentation Ronald Fisher (1890–1962), and finally the engineer and beer brewer William Gosset, alias Student (1876–1937). These developments were the result of an increasingly thorough integration of so-called 'inferential' statistics into probabilistic models. The interpretation of these constructions is always stretched between two perspectives: the one of science, which aims to prove or test hypotheses, with truth as its goal, and the one of action, which aims to make the best decision, with efficiency as its goal. This tension explains a number of controversies that opposed the founders of inferential statistics in the 1930s. In effect, the essential innovations were often directly generated within the framework of research as applied to economic issues, for instance in the cases of Gosset and Fisher.

Gosset was employed in a brewery. He developed product quality-control techniques based on a small number of samples. He needed to appraise the variances and laws of distribution of parameters calculated on observations not complying with the supposed ' law of large numbers.' Fisher, who worked in an agricultural research center, could only carry out a limited number of controlled tests. He mitigated this limitation by artificially creating a randomness, itself controlled, for variables other than those for which he was trying to measure the effect. This 'randomization' technique thus introduced probabilistic chance into the very heart of the experimental process. Unlike Karl Pearson, Gosset and Fisher used distinct notations to designate, on the one hand, the theoretical parameter of a probability distribution (a mean, a variance, a correlation) and on the other, the estimate of this parameter, calculated on the basis of observations so insufficient in number that it was possible to disregard the gap between these two values, theoretical and estimated.

This new system of notation marked a decisive turning point: it enabled an inferential statistics based on probabilistic models. This form of statistics was developed in two directions. The estimation of parameters, which took into account a set of recorded data, presupposed that the model was true. The information produced by the model was combined with the data, but nothing indicated whether the model and the data were in agreement. In contrast, the hypothesis tests allowed this agreement to be tested and if necessary to modify the model: this was the inventive part of inferential statistics. In wondering whether a set of events could plausibly have occurred if a model were true, one compared these events – explicitly or otherwise – to those that would have occurred if the model were true, and made a judgment about the gap between these two sets of events.

This judgment could itself be made according to two different perspectives, which were the object of vivid controversy between Fisher on the one hand, and Neyman and Egon Pearson on the other. Fisher's test was placed in a perspective of truth and science: a theoretical hypothesis was judged plausible or was rejected, after consideration of the observed data. Neyman and Pearson's test, in contrast, was aimed at decision making and action. One evaluated the respective costs of accepting a false hypothesis and of rejecting a true one,

described as errors of Type I and II. These two different aims – truth and economy – although supported by close probabilistic formalism, led to practically incommensurable argumentative worlds, as was shown by the dialogue of the deaf between Fisher on one side, and Neyman and Pearson on the other (Gigerenzer et al., 1989: pp. 90–109).

Official Statistics and the Construction of the State

At the same time as mathematical statistics were developing, so-called 'official' statistics were also being developed in 'bureaus of statistics,' for a long time on an independent course. These latter did not use the new mathematical tools until the 1930s in the United States and the 1950s in Europe, in particular when the random sample-survey method was used to study employment or household budgets. Yet in the 1840s, Quetelet had already actively pushed for such bureaus to be set up in different countries, and for their 'scientification' with the tools of the time. In 1853, he had begun organizing meetings of the International Congress of Statistics,' which led to the establishment in 1885 of the 'International Statistical Institute' (which still exists and includes mathematicians and official statisticians). One could write the history of these bureaus as an aspect of the more general history of the construction of the state, insofar as they developed and legitimized a common language specifically combining the authority of science and that of the state (Anderson 1988; Desrosiéres 1998; Patriarca 1996).

More precisely, every period of the history of a state could be characterized by the list of 'socially judged social' questions that were consequently put on the agenda of official statistics. So were co-constructed three interdependent foci: representation, action, and statistics – a way of describing and interpreting social phenomena (to which social studies would increasingly contribute), a method for determining state intervention and public action, and finally, a list of statistical 'variables' and procedures aimed at measuring them.

Thus for example in England in the second third of the nineteenth century, poverty, mortality, and epidemic morbidity were followed closely in terms of a detailed geographical distribution (counties) by the General Register Office (GRO), set up by William Farr in 1837. England's economic liberalism and the Poor Law Amendment Act of 1834 (which led to the creation of workhouses) were consistent with this form of statistics. In the 1880s and 1890s, Galton and Pearson's hereditarian eugenics would compete with this 'environmentalism,' which explained poverty in terms of the social and territorial contexts. This new 'social philosophy' was reflected in news forms of political action, and of statistics. Thus the 'social classification' in five differentiated groups used by British statisticians throughout the twentieth century is marked by the political and cognitive configuration of the beginning of the century (Szreter 1996).

In all the important countries (including Great Britain) of the 1890s and 1900s, however, the work of the bureaus of statistics was guided by labor-related issues: employment, wages, workers' budgets, subsistence costs, etc. The modern idea of unemployment emerged, but its definition and its measurement were not standardized yet. This interest in labor statistics was linked to the fairly general development of a specific 'labor law' and the first so-called 'social welfare' legislation, such as Bismarck's in Germany, or that developed in the Nordic countries in the 1890s. It is significant that the application of the sample survey method (then called 'representative' survey) was first tested in Norway in 1895, precisely in view of preparing a new law enacting general pension funds and invalidity insurance: this suggests the consistency of the political, technical, and cognitive dimensions of this co-construction.

These forms of consistency are found in the statistics systems that were extensively developed, at a different scale, after 1945. At that time, public policies were governed by a number of issues: the regulation of the macroeconomic balance as seen through the Keynesian model, the reduction of social inequalities and the struggle against unemployment thanks to social-welfare systems, the democratization of school, etc. Some people then spoke of 'revolution' in government statistics (Duncan and Shelton 1978), and underscored its four components, which have largely shaped the present statistics systems. National accounting, a vast construction integrating a large number of statistics from different sources, was the instrument on which the macroeconomic models resulting from the Keynesian analysis were based. Sample surveys made it possible to study a much broader range of issues and to accumulate quantitative descriptions of the social world, which were unthinkable at a time when the observation techniques were limited to censuses and monographs. Statistical coordination, an apparently strictly administrative affair, was indispensable to make consistent the observations resulting from different fields. Finally, beginning in 1960, the generalization of computer data processing radically transformed the activity of bureaux of statistics.

So, 'official statistics,' placed at the junction between social studies, mathematics, and information on public policies, has become an important research component in the social studies. Given, however, that from the institutional standpoint it is generally placed outside, it is often hardly perceived by those who seek to draw up a panorama of these sciences. In fact, the way bureaus of statistics operate and are integrated into administrative and scientific contexts varies a lot from one country to another, so a history and a sociology of social studies cannot omit examining these institutions, which are often perceived as mere suppliers of data assumed to 'reflect reality,' when they are actually places where this 'reality' is instituted through co-constructed operations of social representation, public action, and statistical measurement.

See also: Estimation: Point and Interval; Galton, Sir Francis (1822–1911); Neyman, Jerzy (1894–1981); Pearson, Karl (1857–1936); Probability and Chance: Philosophical Aspects; Quantification in History; Quantification in the History of the Social Sciences; Quetelet, Adolphe (1796–1874).

Bibliography

Anderson, M.J., 1988. The American Census. A Social History. Yale University Press, New Haven, CT.

Desrosiéres, A., 1998. The Politics of Large Numbers. A History of Statistical Reasoning. Harvard University Press, Cambridge, MA.
Duncan, J.W., Shelton, W.C., 1978. Revolution in United States Government Statistics, 1926–1976. US Department of Commerce, Washington, DC.
Gigerenzer, G., et al., 1989. The Empire of Chance. How Probability Changed Science and Everyday Life. Cambridge University Press, Cambridge, UK.
Hacking, I., 1990. The Taming of Chance. Cambridge University Press, Cambridge, UK.
Klein, J.L., 1997. Statistics Visions in Time. A History of Time Series Analysis, 1662–1938. Cambridge University Press, Cambridge, UK.
Lazarsfeld, P., 1977. Notes in the history of quantification in sociology: Trends, sources and problems. In: Kendall, M., Plackett, R.L. (Eds.), Studies in the History of Statistics and Probability, vol. 2. Griffin, London, pp. 213–269.
MacKenzie, D., 1981. Statistics in Britain, 1865–1930. The Social Construction of Scientific Knowledge. Edinburgh University Press, Edinburgh, UK.
Patriarca, S., 1996. Numbers and Nationhood: Writing Statistics in Nineteenth-century Italy. Cambridge University Press, Cambridge, UK.
Pearson, K., 1911. The Grammar of Science. Third ed rev. and enl. A. and C. Black, London.
Porter, T., 1986. The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton, NJ.
Stigler, S.M., 1986. The History of Statistics: The Measurement of Uncertainty Before 1900. Belknap Press of Harvard University Press, Cambridge, MA.
Szreter, S., 1996. Fertility, Class and Gender in Britain, 1860–1940. Cambridge University Press, Cambridge, UK.

Status and Role, Social Psychology of

Lynn G Chin, Washington & Lee University, Lexington, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G.M. Platt, volume 22, pp. 15090–15095, © 2001, Elsevier Ltd.

Abstract

Status and roles are fundamental concepts in sociology, but have been conceptualized in a myriad of different ways. For example, do status and roles refer to a stable set of positions and expectations or do they refer to more fluid processes of interaction? The way these concepts have been defined over the years has impacted the way their relationship has been characterized. The social psychological study of the relationship between status and roles has traditionally fallen under two main camps: structural-functionalism and symbolic interaction. This article discusses differences in how the definitions of status and role have changed over time under both the functionalist and symbolic interactionist camps. Although these two lines of thought have developed under different theoretical trajectories, they actually create a complementary understanding of the relationship between status and roles. In sum, this article makes the case for the need to develop both lines of thought more fully in conjunction with one another.

Social psychological research in the relationship between status and roles has had a rich and complex history. This relationship has been complicated by the fact that the definitional content of what 'roles' and 'status' are has varied across different areas of research. Social psychological research on 'roles and status' has traditionally fallen under two major paradigmatic camps: one whose roots have structural-functionalist underpinnings and the other whose roots are embedded in symbolic interaction. In the quest to characterize the relationship between status and roles, structural-functional research and symbolic interactionist research have differentially emphasized the extent to which they conceive of roles and status as being stable sets of attributes, expectations, and behaviors versus being flexible dynamic processes.

On the one hand, the structural-functionalist camp has over the years viewed the development of intragroup status hierarchies as a process, which in turn impacts and is impacted by the extent to which actors take on stable role expectations. These expected behaviors act as an extension of individuals' position within a group's status hierarchy. On the other hand, the symbolic interactionist camp has tended to view the development of role behaviors as a shared social process, which impacts and is impacted by the extent to which interactants hold or aspire to hold certain stable status positions. As a result, these two lines of research conceptualize the interplay between status and roles differently and focus on different social outcomes. The structural-functionalist camp has tended to be more concerned with the development of patterned intragroup interactions, such as the process of role differentiation and the creation of ingroup status hierarchies, while the symbolic interactionist camp has tended to be more concerned with the development of inter- and intrapersonal interactions, such as the processes of role creation and the maintenance of one's self-identity. Given that the two major paradigms have differing conceptualizations of the definitions of 'roles' and 'status,' this article will discuss the interplay of status and roles along the way their definitions have traditionally developed within both the structural-functionalist and symbolic interactionist paradigms.

Structural-Functionalism

'Status' as a Social Position and 'Roles' as Associated Behavioral Expectations

In early structural-functionalist works, 'status' was not always considered a dynamic process, but instead referred more to stable societal positions, which were associated with certain 'roles,' or normative behaviors. In his book, *The Study of Man*, anthropologist Ralph Linton has famously been credited for being the first to define 'status' as a societal position imbued with certain normative privileges and responsibilities, where status is a "… a position in particular pattern" (1936: p. 113), which "… is simply a collection of rights and duties" (1936: p. 113). For Linton, roles referred to the behaviors associated with enacting or embodying one's status or position in society: "A role represents the dynamic aspect of a status…. When he puts the rights and duties which constitute the status into effect, he is performing a role" (1936: p. 114).

Linton divided statuses into two types. For Linton, status positions could either be ascribed or achieved, a distinction that is still utilized to this day. Ascribed statuses are social positions that are accorded to an individual without reference to the individual possessing any special personal talents or skills, such as those assigned to individuals based on their age or gender, while achieved statuses are those social positions that individuals can attain competitively through skill and effort. However, according to Linton (1936), most statuses in society are ascribed. Even those statuses that are supposedly achieved are in actuality frequently circumscribed because the opportunities to attain many achieved statuses are likely to be limited by individuals' ascribed statuses. For example, even if they are equally skilled at academics, not all individuals have the equal opportunity to become a college student. Other factors, such as family income, can limit many students.

Linton's idea of status and roles as inseparable building blocks around which society is built is representative of the early structural-functionalist view that status and roles are fundamentally important to the successful workings of society. Since functionalists conceptualize society as being built upon

patterned reciprocal interdependencies among people, this highlights the importance of statuses as positions and roles as behaviors that create and maintain the social interdependencies around which society functions in order to survive. As Linton stated: "[S]ocial systems serve as guide to the shaping of individuals and their arrangement in certain relations to each other, this combination of shaping and arrangement making it possible for the mutually adjusted individuals to function together as a society," (1936: p. 107) for "Status and role serve to reduce the ideal patterns for social life to individual terms. They become models for organizing the attitudes and behavior of the individuals so that these will be congruous with those of the other individuals participating in the expression of the pattern" (1936: p. 114). In other words, carrying out expected roles or behaviors associated with certain social positions is what helps maintain the social order of interdependent relationships that undergird society.

Talcott Parsons (1951), one of the originators of modern functionalist theory, also took roles and status as fundamental building blocks of social life in describing the social system. For Parsons, 'the status–role' is fundamental for individual participation in social life because positional status always places individuals in relation with one another. It was this relational aspect of status as positions that determined interaction between individuals through associated role behavior: "…[P]articipation has two principle aspects. On the one hand there is the positional aspect-that of where the actor in question is 'located' in the social system relative to other actors. This is what we will call his status, which is his place in the relationship system considered as a structure, that is a patterned system of parts. On the other hand there is the processual aspect, that of what the actor does in his relations with others seen in the context of its functional significance for the social system. It is this which we shall call his role" (Parsons, 1951: p. 25). Along this focus on the relational aspects of different status positions, Robert K. Merton extended the original conception of having one role associated with each status position, to the idea of a role set, where "each social status involves not a single associated role, but an array of roles" (1957: p. 110). A role set is a series of different behaviors associated with a single status position. Individual actors in a specific status position will play different roles depending on the status positions of the others they are interacting with. By invoking the idea of a 'role set,' Merton focused researchers' attention onto the plethora of role relationships associated with holding a certain status position. After all, the expected behavior of a role occupant differs depending on the social position of the 'other' they are interacting with. For example, the status of a police officer would invoke different role expectations for behavior depending on if the officer was interacting with a potential criminal, a victim, the press, his or her colleagues, or his or her superior.

Under Parsons and Merton's direction, functional role theory focused on understanding (1) how stable social relations could invoke role behavior from status inhabitants and (2) how social structure can shape role relations in ways that impact individuals' ability to carry out different roles. Since all individuals hold several different social positions associated with different roles, they are subject to tensions that arise from inconsistent and incompatible role expectations. All individuals face differing degrees of role congruency (Sarbin and Allen, 1968), or compatibility, among their different roles or in their different expectations for a single role. In turn, these inconsistencies subject individuals to varying degrees of role conflict and role strain (Goode, 1960). Role conflict is produced when an individual holds different statuses that produce incompatible role expectations, such as the tensions that might arise when an individual is faced with the dilemma of whether to work overtime as he or she tries to fulfill both the role of devoted parent and loyal employee. Role strain is created when an individual experiences difficulty in carrying out their role obligations, such as when the expectations of the different roles associated with a particular status position conflict. For example, parents often experience role strain when buying presents for their children. On the one hand, the parent role entails the expectation of making sound financial decisions that fit under the family's budget, but on the other hand, the parent role also entails trying to provide their children happiness by giving them things they may want.

'Status' as Social Prestige and 'Role Differentiation'

Another path by which early structural-functionalists considered the relationship between 'status' and 'roles' was in examination of the processes by which individuals in task-oriented groups attain social positions high in status or social prestige. Under this line of research, the term 'status' took on extra meaning besides referring to status as a social position. Although Linton is often credited for propounding a definition of status as a stable social position, he actually mentioned the duality of status both as a social position, and as the social honor or prestige derived across the aggregate of one's different social positions: "*the status* of any individual means the sum total of all the statuses which he occupies. It represents his position with relation to the total society" (italics in original, p. 113). Similarly, in this line of work, the definition of status is more similar to Max Weber's 'status,' which refers to the degree of honor, social prestige, or respect accorded to an individual by others based on one's social position (1964[1947]).

Delving into status as positions that are due deference or respect led to work focused on the process of role differentiation between members of a group. In particular, Robert Freed Bales and colleagues in their group laboratory experiments at Harvard noticed the tendency of small task-oriented groups to parse group members into differentiated role positions over time even when the groups were essentially homogenous and without a preexisting status order to begin with. What Bales and colleagues noticed was the development of consistent patterns in the extent to which group members came to a consensus around who in the group (1) provided quality ideas toward solving the group task, (2) guided the group forward on the group task, and (3) was well liked or popular. They also noted that members' ratings of group members correlated with group members' actual behaviors in interaction. The interaction process analysis method developed by Bales (1950) allowed researchers to categorize group members' behaviors along 12 different interactive behaviors (shows solidarity, shows antagonism, shows tension release, shows tension, shows agreement, shows disagreement, gives suggestions, asks for suggestions, gives opinion, asks for opinion, gives information, and asks for information). This allowed

researchers to rank group members along different criteria: (1) who talked the most, (2) who received the most interaction, (3) who provided the best ideas on the group task, (4) who was most effective at guiding the group forward, and (5) who was the most liked. What the laboratory study patterns suggested was the emergence of different functional leadership roles into two complementary types of leaders: an instrumental task leader and a socioemotional leader.

On the one hand, the instrumental task leader is a position where other group members highly rate the occupant as the most competent or skilled group member on the group task, but who is not well liked by others. On the other hand, the socioemotional leader is a position where the occupant is highly liked by others, but who is not the most highly rated on competence at the group activity. What Bales and colleagues found was that evaluations of these different group leaders mapped onto the actual role behavior of group members (Bales, 1958; Bales and Slater, 1955; Slater, 1955). Task leaders were more likely to not only contribute problem-solving ideas and receive the most agreement, but also exhibit more antagonism. Socioemotional leaders were more likely to initiate positive reactions to others, receive more solidarity and tension release. As a result, the roles of these different leaders became differentiated, such that task leaders pushed the group forward with regard to focusing attention on the instrumental task at hand, while socioemotional leaders kept the group together by focusing attention on solving intragroup tensions.

According to work in this field, the emergence of role differentiation, particularly with respect to leaders, helped maintain a group's survival – this was because they conceived of task contributions in a zero-sum way such that the more one individual engages in task acts, the more he or she limits others' opportunities for doing so. This line of research was interested in how role differentiation enhanced the positive functioning of the group as different specialists catered their role behaviors to solve different needs of the group. For example, having a task leader helped the group fulfill the need to coordinate task activity toward successful achievement, while having a socioemotional leader fulfilled the need to promote positive integrative feelings and relations within the group that might deteriorate while engaging in conflict around task-oriented activities. Research on the differentiation between leadership roles fell off by the 1970s, as researchers realized these processes were not universal and that the situations under which role differentiation occurred were limited. For example, Burke (1967, 1968, 1971) found that role differentiation into separate roles for task and socioemotional leaders was less likely to occur when group members accept the task activity as legitimate, such that group members are more incentivized to engage in the task and are asked to achieve consensus. Under these conditions, the instrumental task leader becomes more likely to engage in socioemotional activity and is more liked, which lessens the need and the opportunity for a socioemotional leader to arise.

'Status' as a Social Process and 'Roles' as Stable Expectations

Bales' findings spawned another very productive line of contemporary structural-functionalist research regarding the process of status differentiation. Interested in explaining the process by which stable hierarchies could emerge even in nondifferentiated groups, such that some group members come to hold more influence over group decision making and respect than others, Joseph Berger and colleagues created expectation states theory (Berger et al., 1972). The development of this new theoretical research program led the structural-functionalist line of research to move from conceptualizing status as set positions per se, and conceive status to be more of a process where some individuals achieve more influence over others in small group interaction. Under expectation states theory, members of task-oriented groups become differentiated as the group develops consensual expectations regarding which group members are more skilled or competent at the collective task. Those whose contributions are expected to be more beneficial for helping the group achieve their goal are accorded more influence in the group, such that others in the group are more likely to defer decisions to group members who are expected to be skilled at the group task. More recent work has also found that influence or status is provided to those who are perceived to be more group oriented and willing to sacrifice personal gain for the group (Hardy and Van Vugt, 2006; Ridgeway, 1982; Willer, 2009). Not only are these 'high status' group members more likely to be given more opportunity to state their opinions and have their opinions evaluated highly and validated by other group members, but also they take on the behaviors of a leader, such that they are more likely to interrupt, disagree, or emit behaviors to guide the group along. The opposite is true for lower status group members, who become less likely to state their opinions and more likely to concede to the suggestion of others. Thus, under expectation states theory, roles become relegated to being stable set of complementary expectations of the skills and behaviors attributed to those accorded different positions along an intragroup status hierarchy. Under expectation states theory, expectations, or roles produce status or intragroup influence, which in turn gives rise and legitimacy to differentiated status positions.

This view of status as a dynamic process with roles as static differing expectations attached to different status positions continues to hold even among several theoretical offshoots of expectations states theory like status characteristics theory (Berger et al., 1972) and status construction theory (Ridgeway, 2001, 2006). Status characteristics theory explains how salient status characteristics, such as gender, race, or class, become stable bases of hierarchical status inequality in groups. Status construction theory focuses on explaining the development and spread of status beliefs in society. Recent work on the social psychology of status has focused on functional importance of status hierarchies in creating differential role expectations between high and low status group members. It has been found that people respond positively when they form working relationships with others who complement them in terms of dominance, such that 'dominance complementarity' has also led to higher levels of group cohesion (Tiedens and Fragale, 2003; Tiedens et al., 2007). Indeed, Ridgeway and Johnson (1990) found that frustration and negative emotional behaviors are more likely to be engaged in when groups are unable to come to agreement on a status order. They state: "Status struggles emerge out of task-based disagreements when the

disagreeing members do not mutually agree on a resolution after an initial statement of arguments. This may … happen when two members disagree about their relative performance expectations or have nearly equal performance expectations" (p. 1203).

In recent years, work on social status and its impact on role expectations has proliferated greatly. However, an underlining theme across much recent research highlights the idea that as people are differentiated into varying status positions, the characteristics, behaviors, and even emotions attributed to these status-differentiated individuals differ in predictable ways. For example, Tiedens et al. (2000) found that high- and low-status members of groups were expected to have different emotional reactions to positive and negative group outcomes. High-status actors were expected to feel angrier when group outcomes were negative and more pride when group outcomes were positive, while lower status actors were expected to feel more sad and guilty when the group failed and more appreciation and gratitude when the group was successful. Work on the stereotype content model (Fiske et al., 1999; Fiske et al., 2002; Glick and Fiske, 2001) has found that stereotypes associated with high status groups often assume members of these groups are highly competent. Even more interesting has been their finding that many stereotypes hold an inverse relationship between perceptions of competence and perceptions of warmth of different social groups, such that many groups that hold high societal social status, but whose positions threaten status quo, are considered competent, but not warm (e.g., business women or rich people). Indeed, work on real-world status differences has found that those who hold higher social positions, such as those with more resources, are less likely to act in prosocial manner (Piff et al., 2010), more likely to violate norms of ethical behavior to achieve desired personal ends (Piff et al., 2012), and less likely to attribute social and personal outcomes with contextual explanations (Kraus et al., 2009).

To summarize, early structural-functional role theorists were interested in understanding how occupants of different social statuses learned to internalize, conform, and juggle the existing agreed-upon role expectations that their status positions subject them to. Over time, structural-functional research in roles and status moved toward a more dynamic picture of the process of both role and status differentiation. However, this move toward a more dynamic look at the differentiation processes between those of differing positions and status comes at the expense of understanding how undertaking role behavior itself can also be a dynamic process. Structural-functional explanations of role behavior tend to presuppose static normatively shared consensuses around the boundaries and expectations of a role and they also tend to assume that individuals engage in differential role behavior due to individual internalization of normative role expectations. The areas in which both early and contemporary functional role theorists are not strong in explaining include understanding role-based interactions, in particular: (1) how variety and change in the enactment of role behaviors can occur and (2) understanding how individuals resolve conflicting motivations when enacting or rejecting role expectations.

Symbolic Interactionism

'Roles' as a Process and 'Status' as Stable Extensions of Position

The symbolic interactionist approach to understanding the relationship of roles and status complements the structural-functionalist approach. While structural-functionalists have concentrated on explaining (1) how individual behaviors fall into line by following stable role expectations based around structural social inequalities and (2) how conflicting role expectations can be adjudicated by changes in the larger social structure, symbolic interactionists have reacted against this oversocialized view of individuals who unquestioningly take on and carry out expected role behaviors. Instead, the symbolic interactionist approach to the relationship between role and status allows individual identity and motivations to play a large part in carrying out role behaviors. Under symbolic interaction, the connection between individual identity and social interaction is of utmost importance in the creation of shared social meanings behind different social roles. This is because social identities help individuals frame and interpret the social situation and their place in the scenes they find themselves in. Furthermore, roles link individuals to society because each individual holds a role repertoire, which contains all the different role sets of the individual, and as such individuals' identities are a reflection of the larger society under which the individual lives and functions.

George Herbert Mead's conception of the role forms the basis around which symbolic interactionist research regarding the relationship between roles and status are founded (1934). For Mead, 'taking the role of the other' is essential to social interaction. To take on the role of the other, one must be able to view the world (including oneself and one's behaviors) through the standpoint of others. Taking on the role of the other is important because it is only through being able to view how others might see one and one's actions that an individual is able to adjust his/her actions to fulfill his/her roles. As Mead stated: "It is through taking this role of the other that he is able to come back on himself and so direct his own process of communication…. The immediate effect of such role-taking lies in the control which the individual is able to exercise over his own response" (1934: p. 254). Further, according to Mead, it is only through taking the role of the 'generalized other,' such that one can see oneself through the eyes of society or all others who may hold related positions, that one will come to know oneself.

This conception of identities as a lens to the world works when one assumes that social identities are attached to relatively stable status positions. When one assumes that the underlying identity attached to a status position is stable, it allows for individual agency in the interpretation and performance of the role expectations attached to those stable status positions. Each individual has the capacity to make the role identity their own and will use that role behavior to try to draw others into their own definition and interpretation of the interaction. As such, the symbolic interactionist approach to roles and status moves the focus of research away from trying to understand how social structural positions impact individual behavior, and instead concentrates attention on how individuals may affect and change the behavior of others in interaction

by the way they play the role attached to their situationally salient role identity. The explanation for why individuals perform role behavior differs between structural-functionalists and symbolic interactionists. For structural-functionalists, because individuals occupy certain social positions, they conform or are expected to conform to the role behaviors associated with their social positions. For symbolic interactionists, individuals carry out the actions they do because they are trying to frame or uphold a certain identity in the eyes of others. As a result, individuals will use their behaviors to try to shape the way others interpret and define the situation. As such, role behavior becomes the means by which individuals negotiate the creation of shared meanings with others in interaction.

The enactment of roles as a performance was propounded very early on in Erving Goffman's (1959) work on impression management. A founder of the dramaturgical school of symbolic interactionism, Goffman discusses social interaction and the development of the self as a performative experience, where the setting and roles provide individuals scripts by which they can engage with one another. Individuals attempt to achieve their aims through the way they perform the role identities that they take on. In his work, Goffman articulated the importance of role performance while inhabiting a particular social position or identity: "A status, a position, a social place is not a material thing, to be possessed and then displayed; it is a pattern of appropriate conduct, coherent, embellished, and well articulated. Performed with ease or clumsiness, awareness or not, guilt or good faith, it is none the less something that must be enacted and portrayed, something that must be realized" (1959: p. 75). Under impression management, Goffman suggested that people tailor their behavior to try to produce the view of themselves that they want others to see. In other words, individuals use role performance to shape others' impressions of themselves for "Regardless of the particular objective which the individual has in mind and of his motive for having this objective, it will be in his interest to control the conduct of others, especially their responsive treatment of him. This control is achieved by influencing the definition of the situation which others come to formulate, and he can influence this definition by expressing himself in such a way as to give them the kind of impression that will lead them to act voluntarily in accordance with his own plans" (Goffman, 1959: pp. 3–4). As such, Goffman's work highlights how individuals can instrumentally use role behavior to dynamically affect how others view their social position in interaction.

Ralph Turner further shifted symbolic interactionists' attention away from 'role taking' to one of 'role making' (1975 [1962]). By proposing researchers focus on how individuals create shared meanings for roles, Turner emphasized the importance of individual agency in the creation of role behavior. According to Turner, the way that people perform roles differs and that much of role behavior is not predetermined by shared normative expectations before interaction occurs, but instead arises in each interaction. Under symbolic interaction, consensus, and the negotiation of shared meanings is the goal of interaction, rather than a given. Turner states that: "The idea of role-taking shifts emphasis away from the simple process of enacting a prescribed role to devising a performance on the basis of an imputed other-role. The actor is not the occupant of a position for which there is a neat set of rules … but a person who must act in the perspective supplied in part by his relationship to others whose actions reflect roles that he must identify," (Turner, 1975[1962]: p. 111). Variation in role-behavior stems not only from individual agency, but also because the content of roles can be ambiguous themselves: "Roles 'exist' in varying degrees of concreteness and constancy, while the individual confidently frames his behavior as if they had unequivocal existence and clarity. The result is that in attempting from time to time to make aspects of the roles explicit he is creating and modifying roles as well as merely bringing them to light; the process is not only role-taking but role-making" (Turner, 1975[1962]: p. 110).

For Turner, the actual enactment of a role is shaped not by the specific actions an individual undertakes in carrying out the role, but more in the 'mind-set' an individual associates with the role. Turner argued that, "The individual acts as if he were expressing some role through his behavior and may assign a higher degree of reality to the assumed role than to his specific actions. The role becomes the point of reference for placing interpretations on specific actions, for anticipating that one line of action will follow upon another, and for making evaluations of individual action" (1975[1962]: p. 111). Depending on which role out of one's role repertoire is being engaged in, individuals create and adjust their definitions of a situation. Turner's work highlights an inherent tension between one's self-identity and the enactment of one's role identities. On the one hand, an individual's conception of his/her self is an entity that is beyond all the different roles they have in their role repertoire and this aspect of the self colors how the individual engages in any one of his/her roles. On the other hand, a person's self is also linked to the roles that the individual often engages in, such that one's self-concept is tied to "… those roles that are still played when not called for and color the way in which other roles are played" (Turner, 1978: p. 2).

Turner also created the idea of a role–person merger to describe the occurrence when a role becomes so salient to the individual that the role becomes integrated into their personal behavior. A role–person merger occurs when people become too attached to a role leaving them unable to dissociate from or compartmentalize that role even when it may be to their benefit to take on an alternative role. This can create role conflict for individuals, as "Many of the discrepancies between role prescription and role behavior … can be explained by the individual's inability to shed roles that are grounded in other settings and other stages of the life cycle" (Turner, 1978: p. 1).

However, as much as people come to embrace their roles, they also leave them. The concept of role distance was created by Goffman (1975[1961]) to account for the process by which individuals disassociate themselves from certain roles. Sometimes, role distance comes from people transitioning out of roles, such as when a grade-schooler refuses to hold hands with parents because it is a sign of being 'babyish.' At other times, role distance comes from a disinclination to fully embrace a role, for example, when teens are unwilling to be seen as 'dependents' of their parents. It is interesting to note that

many examples of this type of role distance can be found when an individual is struggling with accepting the (usually low) status of the position they are holding, such as when a working parent loses their job and becomes a 'stay-at-home' parent. As Goffman himself notes: "Situated roles that place an individual in an occupational setting he feels is beneath him are bound to give rise to much role distance…. Some of the most appealing data on role distance come from situations where a subordinate must take orders or suggestions and must go along with the situation as defined by superordinates" (1975[1961]: p. 128).

If people's self-identities are bound to a set of role identities (Burke, 2004; Burke and Reitzes, 1981), this begs the question of how do different roles come to gain or lose importance, such that it is possible for an individual to have a role–person merger or engage in the process of role distance? Identity salience refers to the probability that a certain role identity will be enacted (Stryker, 1968). According to Sheldon Stryker (1968), all the identities in an individual's role repertoire are rank ordered in a salience hierarchy. The higher a role identity is in the salience hierarchy, the more likely it is that the individual will enact and utilize that role across different social situations. According to Stryker's identity theory (1968), a role identity will be more salient the higher his/her commitment to that identity. Commitment to a role heightens the more an individual believes that occupying a certain role is essential in maintaining important social relationships. Therefore, commitment to a role for an individual is essentially based on two main factors – the number of relationships that depend on the individual fulfilling a certain role and the extent to which the individual values each of those relationships.

Symbolic interactionist work brings insight into how roles connect the individual self to larger society. First, roles are central to conceptions of self-concepts because ideas of the self are composed of different role identities (McCall and Simmons, 1966; Stryker, 1968). Second, each individual's 'self' as a ranked hierarchy of role identities is actually a reflection of an individual's place in society, and a mirror image of the network of social relationships that comprise society itself (Stryker and Macke, 1978). Recent work on the enactment of roles, in particular research in identity control theory on self-verification, focuses more on how internal processes impact how individuals choose to engage and carry out their roles (Burke, 1991, 2006; Burke and Cast, 1997). This type of work focuses on understanding how individuals may change their behaviors when others' reflected meanings do not match the identity standard they want to achieve. In other words, individuals' behaviors are important because they produce meanings that support and maintain individuals' self-meanings in interaction. In this line of research, leaders do not engage in leadership behaviors just because they are rising or fulfilling expectations of others for how leaders should behave. Rather, their actions stem from a motivation to control perceptions of others to make others' reactions conform to what the actors themselves believe is fitting for their role. In other words, leaders engage in behavior to try to elicit others' actions to match their identity standard of what reactions are due to a leader in a given situation. When individuals perceive that their actions are being interpreted as fulfilling the identity they want to achieve, they are not likely to change their behaviors. However, when they feel that their identity is being perceived differently than the ideal they would like to attain, this dissonance is grounds for behavioral change. This type of path can shed light on what the behaviors individuals may engage in when they are not accorded the proper amount of esteem or respect they believe they are entitled to. Work in this vein has also found that the more an individual has personal, interpersonal, and structural resources, the higher their ability to self-verify their own personal and role identities (Cast, 2003; Stets and Cast, 2007). This creates the potential for a reinforcing feedback pattern that higher status individuals are more able to take control of situational meanings, since the more an individual is able to self-verify the more resources he/she gains from interaction that can be used toward future identity verification.

Summary

Social psychological work on the relationship between roles and status not only has had a long and rich history, but also continues to thrive. Although work on the relationship between roles and status is often characterized as falling into two oppositional camps, it is clear that the functionalist and the symbolic interactionist approaches also complement one another in their understandings of how social status affects social roles and vice versa. Contemporary structural-functional approach has moved away from considering social status as static and currently focuses on the process of status differentiation and its impact on subsequent role expectations (Rashotte and Webster, 2005; Ridgeway, 2006). There has also been large growth in structural-functional research examining the impact of status hierarchies, status position, and status strivings on people's expectations and enactment of role behaviors (Anderson and Kilduff, 2009; Anderson et al., 2012a,b; Bendersky and Hays, 2012). However, the structural functionalist perspective is still weak at explaining the wide variation in how certain behaviors come to gain shared meaning or describing the processes by which individuals choose to engage in the roles that they do. On the other hand, the symbolic interactionist approach focuses heavily on enactment of role expectations as a process. Both traditional and contemporary symbolic interactionist research has emphasized individual agency in engaging in role behavior. Current work has delved even deeper into the factors that impact the process by which individuals engage in different role behaviors in the process of identity verification and has closely examined how resources, including those that stem from different social positions can impact individuals' ability to successfully achieve their desired identities by manipulating perceptions created in interaction (Cast, 2003; Stets and Cast, 2007; Burke, 2006). However, unlike structural-functional work on the emergence and outcomes that result from the creation of status hierarchies, the symbolic interactionist approach is not strong in explaining how such inequalities between resources and status positions may arise in the first place. In all, research in both traditions complement each other highly. As work continues to progress along both paradigms, it seems quite promising that these

paradigms' differing definitions and emphasis on roles and status will develop an even fuller picture of complex relationship among all the different social psychological processes that connect roles to status and back again.

See also: Conformity: Sociological Aspects; Differentiation: Social; Dramaturgical Analysis; Elites: Sociological Aspects; Interactionism, Symbolic; Social Identity in Sociology; Social Influence, Sociology of; Status and Role: Structural Aspects.

Bibliography

Anderson, C., Kilduff, G.J., 2009. The pursuit of status in social groups. Current Directions in Psychological Science 18, 295–298.

Anderson, C., Kraus, M.W., Galinsky, A.D., Keltner, D., 2012a. The local-ladder effect: social status and subjective well-being. Psychological Science 23, 764–771.

Anderson, C., Willer, R., Kilduff, G.J., Brown, C.E., 2012b. The origins of deference: when do people prefer lower status? Journal of Personality and Social Psychology 102, 1077–1088.

Bales, R.F., 1950. Interaction Process Analysis: A Method for the Study of Small Groups. Addison-Wesley, Reading, MA.

Bales, R.F., 1958. Task roles and social roles in problem-solving groups. In: Maccoby, E., Newcomb, T., Hartley, E. (Eds.), Readings in Social Psychology. Holt, Rinehart & Winston, New York, pp. 437–447.

Bales, R.F., Slater, P.E., 1955. Role differentiation in small-decision making groups. In: Parsons, T., Bales, R.F., Olds, J., Zelditch, M., Slater, P.E. (Contributors), Family, Socialization and Interaction Process. The Free Press, Glencoe, IL, pp. 259–306.

Bendersky, C., Hays, N.A., 2012. Status conflict in groups. Organization Science 23, 323–340.

Berger, J., Cohen, B.P., Zelditch, M., 1972. Status characteristics and expectation states. American Sociological Review 37, 241–255.

Biddle, B.J., 1986. Recent development in role theory. Annual Review of Sociology 12, 67–92.

Burke, P.J., 1967. Development of task and socio-emotional role differentiation. Sociometry 30, 379–392.

Burke, P.J., 1968. Role differentiation and the legitimation of task activity. Sociometry 31, 404–411.

Burke, P.J., 1971. Task and socio-emotional leadership role performance. Sociometry 34, 22–40.

Burke, P.J., 1991. Identity processes and social stress. American Sociological Review 56, 836–849.

Burke, P.J., 2004. Identities and social structure: the 2003 Cooley-Mead award address. Social Psychology Quarterly 67, 5–15.

Burke, P.J., 2006. Identity change. Social Psychology Quarterly 69, 81–96.

Burke, P.J., Cast, A.D., 1997. Stability and change in the gender identities of newly married couples. Social Psychology Quarterly 60, 277–290.

Burke, P.J., Reitzes, D.C., 1981. The link between identity and role performance. Social Psychology Quarterly 44, 83–92.

Cast, A.D., 2003. Power and the ability to define the situation. Social Psychology Quarterly 66, 185–201.

Fiske, S.T., Xu, J., Cuddy, A.C., Glick, P., 1999. (Dis)respecting versus (dis)liking: status and interdependence predict ambivalent stereotypes of competence and warmth. Journal of Social Issues 55, 473–491.

Fiske, S.T., Cuddy, A.J., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Glick, P., Fiske, S.T., 2001. Ambivalent stereotypes as legitimizing ideologies: differentiating paternalistic and envious prejudice. In: Jost, J., Major, B. (Eds.), The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice and Intergroup Relations. Cambridge University Press, New York, pp. 278–306.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Bantam Doubleday Dell Publishing Group, Inc., New York.

Goffman, E., 1975[1961]. Role-distance. In: Brissett, D., Edgeley, C. (Eds.), Life as Theater: A Dramaturgical Sourcebook. Aldine Publishing Company, Chicago, pp. 123–132.

Goode, W.J., 1960. A theory of role strain. American Sociological Review 4, 483–496.

Hardy, C., Van Vugt, M., 2006. Nice guys finish first: the competitive altruism hypothesis. Personality and Social Psychology Bulletin 32, 1402–1413.

Kraus, M.W., Piff, P.K., Keltner, D., 2009. Socioeconomic status, the sense of control, and social explanation. Journal of Personality and Social Psychology 97, 992–1004.

Linton, R., 1936. The Study of Man. D. Appleton-Century Company, Inc., New York.

McCall, G.J., Simmons, J.L., 1966. Identities and Interactions. The Free Press, New York.

Mead, G.H., 1934. Mind, Self and Society: From the Standpoint of a Social Behaviorist. University of Chicago, Chicago.

Merton, R.K., 1957. The role-set: problems in sociological theory. British Journal of Sociology 8, 106–120.

Parsons, T., 1951. The Social System. The Free Press, Glencoe, Illinois.

Piff, P.K., Kraus, M.W., Cote, S., Cheng, B.H., Keltner, D., 2010. Having less, giving more: the influence of social class on prosocial behavior. Personality and Social Psychology 99, 771–784.

Piff, P.K., Stancato, D.M., Cote, S., Mendoza-Denton, R., Keltner, D., 2012. Higher social class predicts increased unethical behavior. Proceedings of the National Academy of Sciences 109, 4086–4091.

Rashotte, L.S., Webster Jr., M., 2005. Gender status beliefs. Social Science Research 34, 618–633.

Ridgeway, C.L., 1982. Status in groups: the importance of motivation. American Sociological Review 47, 76–88.

Ridgeway, C.L., 2001. The emergence of status beliefs: from structural inequality to legitimizing ideology. In: Jost, J., Major, B. (Eds.), The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice and Intergroup Relations. Cambridge University Press, New York, pp. 257–277.

Ridgeway, C.L., 2006. Status construction theory. In: Burke, P.J. (Ed.), Contemporary Social Psychological Theories, pp. 301–323.

Ridgeway, C.L., Johnson, C., 1990. What is the relationship between socioemotional behavior and status in task groups? American Journal of Sociology 95, 1189–1212.

Sarbin, T.R., Allen, V.L., 1968. Role theory. In: Lindzey, G., Aronson, E. (Eds.), The Handbook of Social Psychology, second ed. Addison-Wesley, Reading, MA, pp. 488–507.

Slater, P.E., 1955. Role differentiation in small groups. American Sociological Review 20, 300–310.

Stets, J.E., Cast, A.D., 2007. Resources and identity verification from an identity theory perspective. Sociological Perspectives 50, 517–543.

Stryker, S., 1968. Identity salience and role performance: the relevance of symbolic interaction theory for family research. Journal of Marriage and the Family 30, 558–564.

Stryker, S., Macke, A.S., 1978. Status inconsistency and role conflict. Annual Review of Sociology 4, 57–90.

Tiedens, L.Z., Fragale, A.R., 2003. Power moves: complementarity in dominant and submissive nonverbal behavior. Journal of Personality and Social Psychology 84, 558–568.

Tiedens, L.Z., Ellsworth, P.C., Mesquita, B., 2000. Sentimental stereotypes: emotional expectations for high- and low-status group members. Personality and Social Psychology Bulletin 26, 560–575.

Tiedens, L.Z., Unzeueta, M.M., Young, M.J., 2007. An unconscious desire for hierarchy? The motivated perception of dominance complementarity in task partners. Journal of Personality and Social Psychology 93, 402–414.

Turner, R.H., 1975[1962]. Role-taking: process vs. conformity. In: Brissett, D., Edgeley, C. (Eds.), Life as Theater: A Dramaturgical Sourcebook. Aldine Publishing Company, Chicago, pp. 109–122.

Turner, R.H., 1978. The role and the person. American Journal of Sociology 84, 1–23.

Weber, M., 1964[1947]. The Theory of Social and Economic Organization (A.M. Henderson, T. Parsons, Trans.). The Free Press, New York.

Willer, R., 2009. Groups reward individual sacrifice: the status solution to the collective action problem. American Sociological Review 74, 23–43.

Status and Role: Structural Aspects

John Scott, University of Copenhagen, Copenhagen, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the development of the concepts of status (position) and role within anthropology and sociology, showing how a specific form of role theory was elaborated within structural functionalism. Critical perspectives and elaborations are reviewed, and it is shown that the concepts have a continuing relevance within a variety of theoretical perspectives.

This article examines the development of concepts of status (social position) and role in social anthropology and sociology, tracing their development into mainstream concepts and into the development of role theory. It also explores the criticisms leveled against these structural ideas from symbolic interactionist and from ethnomethodological points of view. It is suggested that these criticisms have, in fact, broadened the structural focus and provide a number of important complementary ideas.

The closely connected concepts of 'status' and 'role' have a long and complex history in a number of different traditions of social thought, but their contemporary meanings were developed largely during the first half of the twentieth century. Despite it having been Shakespeare who claimed that 'all the world's a stage,' it was the social psychologists of the early twentieth century who systematically employed the concept of role and so paved the way for later 'dramaturgical' approaches to social life. This work also prepared the way for those who developed the idea of status as the structural position corresponding to a role. The term status, however, has been used in two quite different ways, only one of which is directly relevant to this discussion. One usage has referred to hierarchical differences of prestige and social honor in systems of social stratification (see Social Stratification); the other has referred, more generally, to any structural position or location in a social system. It is the second sense that will be discussed in this article.

Status as Structural Position

The hierarchical concept of status was popularized by Max Weber, who contrasted stratification by status with stratification by class (Scott, 1996). In the 1930s, however, cultural anthropologists used the word status when referring more generally to positions or locations in a social structure, regardless of whether these involved any ranking by prestige (Linton, 1936). This usage was adopted rapidly by structural-functionalist sociologists. Although this led initially to some confusion about whether or not particular 'statuses' were hierarchically ranked, the new usage soon became established (Parsons, 1942b). Some investigators resolved this confusion by using the words 'position' or 'location' in preference to status (Warner, 1952: 46), and this unambiguous terminology is adopted in this article and is recommended for wider use.

The idea of status as structural position originated in the conceptualization of social structure as comprising culturally defined institutions. This had its origins in the sociology of Durkheim, but it was the development of a clear and systematic understanding of culture by American anthropologists (Kroeber, 1917; White, 1949; see Kluckhohn, 1954) that gave the idea its contemporary form. The culture of a society was seen as defining a set of standardized normative patterns that people are expected to follow in their social actions (Parsons, 1951). These social institutions regulate individual and group behavior by defining the nature of their social relations with each other. Institutions define "proper, legitimate or expected modes of action or of social relationship" (Parsons, 1940: 53), and so guide and channel behavior. This view of social structure as "a patterned system of social relationships of actors" (Parsons, 1945: 230) rapidly became a central concept in sociological analysis.

The institutionalized expectations that people share define their social positions in relation to other positions and specify their expectations in relation to the anticipated actions of the individuals who occupy these other positions (Parsons, 1942b: 143). The knowledge that people have about social positions provides them with conceptual maps that they can use to organize their social actions. This knowledge, of course, is contained in individual minds. Such *shared* knowledge, however, involves a reciprocity of expectations among actors, who sanction each other's behavior through their reactions to any deviance from the expectations. People do not simply 'know' the social positions that make up their society, they experience social pressure to conform to the expectations they involve. This solidifies expectations as institutionalized collective representations (Davis, 1948: 87), and social positions, then, are to be seen as Durkheimian social facts that constrain the behavior of individuals.

The behavioral expectations associated with a position define the role that their occupants are to play. Roles are definitions of what people are expected to do in different situations. Roles include the specific rights and obligations entailed in a social position (Dahrendorf, 1958: 36). They tell us what to do and what to expect others to do (Parsons, 1940: 54). They provide us with a definition of the situation that defines the goals, the means, and the attitudes that are recognized culturally as legitimate for those who occupy particular social positions and sets the limits within which they may legitimately act (Williams, 1960).

Some writers have used the concept of role to describe the *actual* behavior associated with a position, but most have recognized that a normative focus requires a concern for

expected behavior (Levy, 1952: 158–160). The relationship between expected behavior and actual behavior is complex and involves multiple determinants, so there is rarely a one-to-one correspondence between norms and behavior. The fully institutionalized social relation, in which there is a complete consensus and perfect conformity allows us to analyze one set of determinants among the many affecting other people's behavior (Parsons, 1951: 39).

These ideas were taken up during the 1950s as the basis of a 'role theory' that set out a comprehensive paradigm for role analysis. The most influential formulation of this was that of Gross et al. (1958), who used it to explore the behavior of school superintendents. They did this using Merton's (1957) argument that each social position is associated with an array of role-specific forms of behavior that together make up a 'role set.' A medical student, for example, must act as a student not only in relation to other students, but also in relation to teachers, doctors, nurses, social workers, patients, medical technicians, and so on. In each of these relations, the student is likely to encounter different expectations about her or his behavior (Merton, 1957: 112). Gross et al. (1958) showed that the school superintendent must negotiate the conflicting expectations held by teachers, parents, children, governors, and ancillary staff. It was recognized that this definition of role expectations is likely to produce varying degrees of strain or conflict in role expectations. Occupants of different positions may legitimately expect behavior that conflicts with that expected by others, and it may be difficult for an actor to meet all expectations equally or simultaneously. They must balance one set of demands against another and resolve the resulting tensions. This idea of 'role conflict' has been central to many applications of role theory.

Socialization and Conformity

Although this work concentrated on the structural aspects of position and role, it was complemented by those who worked on a social psychology of roles (Newcomb, 1950). These writers explored the learning processes through which roles are acquired and the social pressures to conformity that exist in face-to-face situations. Socialization and conformity were central principles for role theorists, and their overemphasis has been subject to persistent criticism.

Role theorists see the norms that define social positions as being learned through a process of socialization. The expectations of others are *internalized* in the personality as it is formed. They become objects of moral commitment, and individuals are likely to be strongly attached to them. They tend to regard conformity as a duty (Parsons, 1940).

This is associated with the argument that there is consensus over social norms. Members of a society share their moral commitments and so violations are met with guilt or shame in the violator and indignation or condemnation in those whom the deviance affects. Deviance from these shared norms is likely to result in 'corrective forces' or sanctions that may vary from mild disapproval to overt coercion. For many, this led to the conclusion that deviance is both infrequent and automatically corrected. Societies were seen as self-regulating systems in which actors had little choice but to conform.

The degree of consensus in a society rarely reaches this level, however, and individuals are never perfectly socialized (Wrong, 1961). These points were raised especially forcefully against Parsons, but even he recognized that the 'completely institutionalized' social relationship was a rarity. He held that, although broad descriptive contrasts could be drawn between societies in terms of their common values, "when dynamic problems of directions and processes of change are at issue, it is essential to give specific attention to the elements of malintegration, tension and strain in the social structure" (Parsons, 1942a: 117). Nevertheless, the mainstream of role theory overemphasized both consensus and internalization. It was conflict theorists such as Dahrendorf (1958) who showed that role expectations have to be seen in relation to the distribution of power and that this can give them a solidity that does not depend on a consensus of opinion.

What passes for consensus actually may be the dominance of the ideas of a powerful group (Rex, 1961; Shils, 1961). A minority may be able to use its resources to institutionalize its own norms for the population as a whole. This may be especially effective where the majority of the population forms a relatively unorganized mass without strong counterinstitutions of its own.

Mapping Social Positions

The structural-functional approach to positions and roles identified social positions as unintended cultural patterns that form an institutional structure. This was seen as corresponding to actual social relations among occupants of the positions and hence pointed to a conception of relational structure alongside institutional structure (López and Scott, 2000). Thus, social positions could be mapped as structures of relations independent of the cultural meanings that constitute them. The earliest attempt to set out a formal model of relationally defined positions was that of Nadel (1957), who pioneered a mathematical approach that has been taken up in social network analysis by advocates of block modeling (White et al., 1976; Boorman and White, 1976). In a block model, the social relations of individuals are represented in a matrix of connections that discloses 'structurally equivalent' individuals and defines these as social positions. Although some such positions may correspond to culturally defined norms, those such as, for example, social classes may be purely *de facto* social positions. White (2008) has attempted to bring together cultural and relational points of view (see Emirbayer and Goodwin, 1994).

Other approaches to social network analysis have identified relational positions of centrality and peripherality, seeing central intermediaries as, for example, brokers in power relations (Burt, 2005). Combining these approaches with the identification of cliques and clusters in networks of relations has made it possible to identify social positions as associated with varying degrees of social capital (Lin, 2001; Burt et al., 2001). Although some of these approaches depart from the tenets of role theory, they have provided an important and powerful extension of the idea of social position.

Role-Taking and Role-Making

Structural-functional and conflict theories of position and role, supported by social psychological theories of socialization and conformity, lay at the heart of mainstream sociology from the 1940s to the 1960s. During the 1960s, however, radical and forceful objections were raised against these theories . The immediate origin of the sociological concept of role had been Mead's (1927) account of taking the role of the other, and critics of the mainstream turned to Mead for support. They questioned the cultural determinism and oversocialized image of the actor who saw people as the mere puppets of structural forces, which had been espoused in many of the simpler textbook versions of role theory that had popularized its ideas. Blumer (1962: 189) reiterated the original insight that a structure of positions is simply a 'framework inside of which action takes place.' It is a conceptual guide for action, not an external, substantial entity or an independent coercive power.

Turner (1962) took this argument to the heart of role theory. He held that the emphasis on conformity gives only a partial view of role behavior. It stresses imitation and ignores innovation, failing to recognize that role-takers are also 'role-makers': that is, individuals do not simply take over roles as templates for conformist behavior but must, of necessity, improvise. The role knowledge that individuals learn during their primary and secondary socialization cannot give them precise, programmed instructions for behavior in all of the many unique and unpredictable circumstances in which they are likely to find themselves, especially when faced by role conflict. Individuals tentatively interpret and reinterpret each other's actions in the situations that they encounter, recreating their roles from the raw materials provided to them during socialization (Turner, 1962: 23). Shared knowledge of social positions sets the general conditions for action, but it does not fully determine it (Blumer, 1962: 190). The normative functionalist approach to role-taking has to be broadened into a more comprehensive account.

The fullest development of this view was that of Erving Goffman (1959, 1963), who explored the ways in which individuals actively and creatively construct the images that they present in their role behavior. Goffman systematically employed the theatrical metaphor of role-playing, holding that individuals are never merely actors following a script, but are authors as well. Using such ideas as props, scenery, front stage, and back stage, Goffman showed how public performances depend on the more private situations to which individuals can withdraw. He introduced a number of novel concepts, most notably 'role distance' to describe the psychological distancing that actors may establish from the role expectations that they find uncomfortable or uncongenial.

A Sense of Social Structure

The criticisms made from symbolic interactionism added a whole new dimension to the analysis of position and role. Although some critics of mainstream views saw this as a complete alternative to the orthodoxy, others saw it as complementing the structural account. These differences persist and have been compounded by a more radical line of argument from phenomenology and ethnomethodology. Aaron Cicourel has produced the most systematic statement of this critique.

Cicourel asked the fundamental question: How is role-taking possible? His answer was that the taking and making of roles rests on a set of cognitive processes through which actors give meaning to the world and so sustain a 'sense of social structure' (Cicourel, 1972: 11). People do carry role information in their heads, but they also must be able to recognize when one particular position or role is relevant, and they must be able to infer what expectations others have of their behavior. They cannot make sense of their social world simply by drawing on the role and positional knowledge that they have learned during their socialization. Before they can apply norms in particular situations, they must arrive at an understanding of what kind of situation it is. This means that "members of a society must acquire the competence to assign meaning to their environment so that surface rules and their articulation with particular cases can be made" (Cicourel, 1968: 52). This ability to infer and to impute meaning to situations is a practical skill that is an essential condition for any social life at all. Cicourel saw this skill as an interactional competence, making explicit parallels with Chomsky's concept of linguistic competence. Although he did not adopt Chomsky's own rationalist theory of the mind, Cicourel did take over his stress on the generative capacities that are provided by human competences. It is their practical, meaning-making skills that allow people to use their knowledge of social norms to generate appropriate role behavior. The structural aspects of positions and roles, therefore, were seen by Cicourel as resting on the possession of a complex set of cognitive procedures (also termed inductive, interpretive, or inference procedures) that operate in the same way as the deep structure grammatical rules of a language.

Cicourel illustrated these cognitive procedures by drawing on Alfred Schutz's discussion of the assumptions that people must make for social interaction to be possible. Schutz held that individuals must assume a reciprocity of perspectives between themselves and their potential partners, they must fill in the gaps in their knowledge through the *et cetera* principle, and they must assume that things occur as 'normal form.' These and similar cognitive procedures constitute the mental module that makes it possible for actors to generate appropriate but innovative responses in changing circumstances, despite the fact that they have only fragmentary and uncertain evidence available to them. They allow people to assign meaning and relevance to the objects in their environment and to construct definitions of the situation that allow them to infer which of the norms stored in their memories are relevant. People build a sense of social structure that allows them to orient themselves appropriately in the various situations that they encounter. Once the meaning of a situation has been decided, norms can be invoked on the assumption that there is a consensus among those with whom they interact and that these are, indeed, the appropriate norms. Normative order and role behavior, therefore, are negotiated and constructed on the basis of the underlying sense of social structure that interactional competence makes possible.

Gender and Sex Roles

Liberal feminist work on sex roles was one of the earliest applications of the Parsonian view of role, but it is in wider feminist approaches that novel ideas also have developed. Betty

Friedan (1962) took up Margaret Mead's (1950) earlier work to set out a view of socialization into culturally defined gender roles corresponding to biological sex differences (see also Oakley, 1972). The ways in which boys learn the norms of masculinity and girls learn those of femininity involve unconscious learning through interactions with their parents and others, through their play with specific types of toys, and in their games with other children. Nancy Chodorow (1978) applied a psychological perspective to this process of gender socialization. This differentiation of roles is reinforced in adult; life through a differential participation in employment and public life and the overall structure of gender positions has been seen as constituting a structure of patriarchy (Walby, 1997). These arguments on the nature of gender positions and roles have been extended into wider issues of sexuality and sexual orientation through discussions of norms of masculinity and the formation of gay, lesbian, bisexual, and transgendered identities (Connell, 1997; Weeks, 2000).

Judith Butler (1990) took a poststructuralist approach to these issues and provided a novel extension of Goffman's ideas through her work on performativity. Gender differences, she has argued, result from performances that represent idealized gender norms. Masculinity and femininity, therefore, are the results of repeated actions that create the appearance of fixed biological differences. Drawing on Foucault's (1961, 1975) work on practices of surveillance, discipline, and self-control, she saw these performances as discursive constructions that regulate subjectivity and identity.

Conclusion

Status – or position – and role remain central concepts in accounts of social structure (see Biddle, 1986). They do not tell us everything that we need to know about social structure (Lockwood, 1956, 1964), but they are essential elements in any structural account. The use of these normative concepts, however, must incorporate an awareness of the creative and innovative role-making activities that are integral to all role behavior. Role behavior is deeply implicated in the dispersed practices of surveillance and control through which actions are organized. Improvisation and negotiation are essential features of the construction of social action; however, these processes are possible only because individuals possess certain cognitive skills that allow them to infer the meaning of social situations and to build a sense of the social structures in which they must act. Structural positions and roles are human accomplishments, but they have real consequences for the people who occupy them.

See also: Dramaturgical Analysis; Institutions; Norms; Phenomenology in Sociology; Social Network Analysis; Social Stratification; Status and Role, Social Psychology of.

Bibliography

Biddle, B.J., 1986. Recent developments in role theory. Annual Review of Sociology 12.
Blumer, H., 1962. Society as symbolic interaction. In: Rose, A. (Ed.), Human Behavior and Social Processes. Routledge and Kegan Paul, London.
Boorman, S.A., White, H.C., 1976. Social structure from multiple networks: II. American Journal of Sociology 81.
Burt, R.S., 2005. Brokerage and Closure: An Introduction to Social Capital. Oxford University Press, New York.
Burt, R.S., Cook, K.S., Lin, N. (Eds.), 2001. Social Capital: Theory and Research. Transaction, New Brunswicvk, NJ.
Butler, J., 1990. Gender Trouble. Feminism and the Subversion of Identity. Routledge, London.
Chodorow, N., 1978. The Reproduction of Mothering: Psychoanalysis and the Sociology of Gender. University of California Press, Berkeley.
Cicourel, A.V., 1968. The acquisition of social structure: towards a developmental sociology of language and meaning. In: Cicourel, A.V. (Ed.), Cognitive Sociology. Penguin, Harmondsworth, UK, 1973.
Cicourel, A.V., 1972. Interpretive procedures and normative rules in the negotiation of status and role. In: Cicourel, A.V. (Ed.), Cognitive Sociology. Penguin, Harmondsworth, UK, 1973.
Connell, R., 1997. Gender Transformations. Routledge, London.
Dahrendorf, R., 1958. Homo sociologicus: on the history, significance, and limits of the category of social role. In: Dahrendorf, R. (Ed.), Essays in the Theory of Society. Routledge and Kegan Paul, London, 1968.
Davis, K., 1948. Human Society. Macmillan, New York.
Emirbayer, M., Goodwin, J., 1994. Network analysis, culture, and the problem of agency. American Journal of Sociology 99, 1411–1454.
Foucault, M., 1961. Madness and Civilization. Vintage Books, New York, 1973.
Foucault, M., 1975. Discipline and Punish. Allen Lane, London, 1977.
Friedan, B., 1962. The Feminine Mystique. Dell, New York.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Penguin, Harmondsworth.
Goffman, E., 1963. Relations in Public Places. Free Press, New York.
Gross, N., Mason, W.S., Mc Eachern, A.W., 1958. Explorations in Role Analysis: Studies of the School Superintendency Role. John Wiley and Sons, New York.
Kluckhohn, C., 1954. Culture and behavior. In: Lindzey, G. (Ed.), Handbook of Social Psychology, vol. 2. Addison-Wesley, Cambridge, MA.
Kroeber, A., 1917. The superorganic. American Anthropologist 19, 162–213.
Levy, M.J., 1952. The Structure of Society. Princeton University Press, Princeton, NJ.
Lin, N., 2001. Social Capital: a Theory of Social Structure and Action. Cambridge University Press, New York.
Linton, R., 1936. The Study of Man. D. Appleton Century, New York.
Lockwood, D., 1956. Some remarks on the social system. British Journal of Sociology 7.
Lockwood, D., 1964. Social integration and system integration. In: Zollschan, G., Hirsch, W. (Eds.), System, Change and Conflict. Houghton-Mifflin, New York, 1976.
López, J., Scott, J., 2000. Social Structure. Open University Press, Buckingham.
Mead, G.H., 1927/1934. Mind, Self and Society from the Standpoint of Social Behaviorism. University of Chicago Press, Chicago.
Mead, M., 1950. Male and Female. Gollancz, London.
Merton, R.K., 1957. The role set: problems in sociological theory. British Journal of Sociology 8.
Nadel, S.F., 1957. The Theory of Social Structure. Free Press, Glencoe.
Newcomb, T.M., 1950. Social Psychology. Dryden Press, New York.
Oakley, A., 1972. Sex, Gender and Society. Temple Smith, London.
Parsons, T., 1940. The motivation of economic activity. In: Parsons, T. (Ed.), Essays in Sociological Theory, second ed. Free Press, New York. 1954.
Parsons, T., 1942a. Democracy and social structure in pre-Nazi Germany. In: Parsons, T. (Ed.), Essays in Sociological Theory, second ed. Free Press, New York. 1954.
Parsons, T., 1942b. Propaganda and social control. In: Parsons, T. (Ed.), Essays in Sociological Theory, second ed. Free Press, New York. 1954.
Parsons, T., 1945. The present position and prospects of systematic theory in sociology. In: Parsons, T. (Ed.), 1954 Essays in Sociological Theory, second ed. Free Press, New York.
Parsons, T., 1951. The Social System. Free Press, New York.
Rex, J.A., 1961. Key Problems of Sociological Theory. Routledge and Kegan Paul, London.
Scott, J., 1996. Stratification and Power: Structures of Class, Status and Command. Polity Press, Cambridge, UK.
Shils, E.A., 1961. Centre and periphery. In: Shils, E.A. (Ed.), Centre and Periphery: Essays in Macrosociology. University of Chicago Press, Chicago, 1975.
Turner, R., 1962. Role-taking: process versus conformity. In: Rose, A. (Ed.), Human Behavior and Social Processes. Routledge and Kegan Paul, London.
Walby, S., 1997. Gender Transformations. Routledge, London.

Warner, W.L., 1952. The Structure of American Life. University of Chicago Press, Chicago.
Weeks, J., 2000. Making Sexual History. Polity, Cambridge.
White, H.C., 2008. Identity and Control: How Social Formations Emerge, second ed. Princeton University Press, Princeton NJ.
White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks: I. American Journal of Sociology 81.
White, L.A., 1949. The Science of Culture: A Study of Man and Civilization. Grove Press, New York.
Williams, R., 1960. American Society: A Sociological Investigation, second ed. Alfred A Knopf, New York.
Wrong, D., 1961. American Sociological Review 26, 183–193.

Stem Cell Research and Cloning

Catherine Waldby, The University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the 1990s, human embryonic stem cell research and mammalian cloning emerged as two of the most controversial areas of biomedical research, provoking complex social responses and contestations over the status of the embryo, the reproductive rights and obligations of women, the entitlements of patient groups and aging populations, and the commercialization of reproductive biology. These developments placed unprecedented levels of pressure on governments to formulate regulations that are able to reconcile competing ethical concerns while also facilitating scientific innovation and the production of therapies to treat the medical conditions associated with disability and aging. This article will consider the ways that various social science disciplines, particularly science and technology studies and medical sociology, have responded to this situation.

Introduction

In the late 1990s, human embryonic stem cell (hESC) research and mammalian cloning emerged as two of the most controversial and promising areas of biomedical research, provoking both intense scientific interest and complex social responses. Both techniques emerged from advances made in human and animal reproductive sciences and *in vitro* fertilization (IVF) throughout the twentieth century (Clarke, 1998) but particularly since the 1980s (Edwards, 2001). A particular category of stem cell is hESCs, a term used to describe undifferentiated cells that can renew themselves and also give rise to one or more specialized cell types with specific functions in the body. Stem cells are crucial to embryonic and fetal development, but they are also found in adult tissues, particularly in the bone marrow, and they are crucial agents in healing and the replenishment of some tissues. Different categories of stem cell have different qualities. The most celebrated type of stem cell is 'pluripotent', meaning that it has the capacity to develop into almost all of the body's tissue types. Other cell types may be classified as totipotent (able to produce the entire organism) and multipotent (able to differentiate into a more limited range of cell types) (Petersen and Munsie, 2012). The most viable source of pluripotent stem cell lines is human embryos. Other stem cell sources do not appear to be as flexible or active as tissue derived from embryos.

Non-human ESCs were first isolated during the 1980s, but it was not until 1998 that the first hESCs were cultured. In that year, a research team led by James Thomson at the University of Wisconsin derived hESCs from embryos donated by couples in fertility treatment. The inner cell mass was isolated from 5-day-old embryos, and the team developed five cell lines from the tissues (Thomson et al., 1998). Thomson and colleagues demonstrated that the cells could be cultured in an undifferentiated state, frozen and thawed, and then induced to differentiate into the three main groups of embryonic tissue – the ectoderm, endoderm, and mesoderm – which in turn form into the main adult cell types.

The clinical implications of this development are far reaching. If the research could be successfully developed into therapies, it might be used to treat currently intransigent degenerative conditions – spinal cord injuries, Parkinson's disease, Alzheimer's disease – by introducing regenerating tissue into appropriate sites. It might serve as an alternative to solid organ donation, either by repairing a damaged organ in the body or creating a new organ without the need for a donor. In the process it would treat two significant, twenty-first century population health problems. First, it would provide a new biomedical method to treat conditions associated with the aging of the population in the developed world, and the associated increase in health and social care budgets. It would contribute to what is now known as the 'compression of morbidity', the contraction of disabling illness to the shortest possible period prior to death, by improving treatments for chronic degenerative conditions. Second, it would address the ever-increasing gap between tissue donation levels and patient waiting lists. While rates of organ donation have increased in many developed countries over the last 10 years, the demand for organs has far outstripped supply, as more and more conditions are considered treatable through transplant.

The clinical possibilities presented by hESC research were made even more promising by their intersection with advances in mammalian cloning. After nearly a decade of research (Franklin, 2007), Ian Wilmut and other scientists at the Roslin Institute in Scotland announced the birth of Dolly the sheep in 1997, the first mammal to be conceived by somatic cell nuclear transfer (SCNT) (Wilmut et al., 1997). SCNT involves the creating an embryo not by the usual process of *in vivo* conception, fusion of egg and sperm, but through the insertion of the nucleus of a cell from an adult's tissues into an unfertilized egg. The egg has in turn had its own nucleus removed to make way for the introduced nucleus. This creates an embryo with the genome of the adult from whom the nucleus was taken. Prior to the cloning of Dolly, biologists assumed that the nuclei of adult cells had lost their ability to proliferate into other cell types. That is, once dedicated to the production of a particular cell type – muscle, blood, skin – the nuclei could not regain their embryonic pluripotency. As Wilmut and colleagues put it, "the fact that a lamb was derived from an adult cell confirms that differentiation of that cell did not involve the irreversible modification of genetic material required for development to term" (Wilmut et al., 1997: 810).

In addition to enabling whole new animals to be cloned, these techniques are thought to have potential therapeutic applications. It may be possible to use some aspects of the SCNT technique to develop embryonic stem cell lines with the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85060-6

genetic material of an adult donor, and so it could be used to create tissue compatible with that of the patient in need of a transplant. This would avoid the problems of tissue rejection and the need for immunosuppressive drugs, as is the case with solid organ donation treatment today. However, to date, no human SCNT lines have been derived, and it is unclear if the technique can be successfully adapted for human biology. More recently, induced pluripotent stem cells have somewhat displaced SCNT approaches, as they appear to offer similar possibilities for the creation of patient-specific tissues using adult somatic cells rather than embryonic cells (Yu et al., 2007).

These developments provoked an energetic social response, as various constituencies, jurisdictions, and religious, civil, and scientific groups mobilized to try and shape the social ordering of the new field. Since the late 1990s, the areas of stem cell science and cloning have attracted extended debate, both in the public domain and in the social science and humanities literature and disciplines, regarding the status of the embryo, the reproductive rights and obligations of women, the entitlements of patient groups and aging populations, the place of State interests, and the commercialization of reproductive biology. In what follows I will give a broad overview of these debates, and the ways different social scientific disciplines have responded to the challenge of analyzing science–society relationships in this area.

Bioethics, Biopolitics, and the Uses of Embryonic Life

The advent of hESC technology and mammalian cloning reanimated a history of public debate and regulatory activity that had its origins in the IVF debates of the early 1980s. At that time, the advent of a new suite of reproductive technologies, including new techniques for handling and researching human embryos as well as distributing gametes, produced a flurry of inquiries and regulatory activity around the world. A suite of complex legislation ensued in many countries after a protracted period of public debate; for example, the Human Fertilisation and Embryology Act (1990) in the UK, and the Embryo Protection Act (1990) in Germany, which set out regulations for embryo research and handling, clinic licensing systems, and the management of gametes. The exception was the USA, where the federal government was silent on social regulation, in part because of a *laissez-faire* attitude toward privately funded research, and in part because of political anxieties over the bitter abortion politics that fiercely polarized American society after the 1973 *Roe v. Wade* Supreme Court decision. Hence the USA did not settle on a coherent regulatory system for embryo research (Jasanoff, 2005).

After the announcement of hESC technology and mammalian cloning in 1997–98, the trajectory of public debate was shaped considerably by this earlier history, although these new techniques also introduced novel quandaries and provocations. They reanimated concerns about the civil and religious status of embryos (Holland et al., 2001), along similar lines to earlier abortion and IVF debates, about the relationship between human origins and human dignity (Habermas, 2003), about the uses and abuses of cloning, with its many dystopian associations (Franklin, 2007), and about the general instrumentalization of reproduction, particularly as it related to the rights of women (Dodds, 2004).

These concerns, while grave, were more inflammatory in some contexts than others. Hence in the USA, with an absence of regulation and a history of polarization around the status of the unborn, the political power of the anti-abortion, anti-stem cell research religious groups was rapidly evidenced by a cascade of bills introduced into the legislature to ban human reproductive cloning and to prevent federal funding for stem cell research. At the same time, supporters of hESC research, including patient groups, research scientists, and the biotechnology industry, mobilized to argue for the importance of medical research, the therapeutic promise of stem cell research, and the status of the United States as a leader in science (Gottweis et al., 2009). While the Clinton administration initially supported a liberal approach, and the National Bioethics Advisory Commission advocated a federal regulatory panel to facilitate research (Jasanoff, 2005), the G.W. Bush, Republican Party administration, after being elected in 2000 gave far more ground to the religious conservatives, and instituted a policy that restricted federal funding to research only on hESC lines already derived at the time of the policy announcement, August 2001. This compromise was sufficiently workable to both appear to be protecting future embryonic life and to allow privately funded research to continue unhindered, while publicly funded research could be carried out in more restricted ways.

In Germany, the advent of Dolly and hESC research also provoked powerful social responses, for reasons somewhat different from the USA. Here the reception of stem cell research was conditioned by the history of eugenics and Nazi medical experimentation, and by a unique set of statutes designed to prevent the recurrence of such a history. The *Grundgesetz*, the German Constitution, makes the protection of the human right to dignity and life the highest duty of the State, and this protection is extended to human embryos through the 1990 *Embryo Protection Act*. This Act falls within the criminal law, and prohibits, among other things, embryo research, germ line manipulation, reproductive and therapeutic human cloning, and hybridization. It prescribes that all eggs fertilized through an IVF procedure must be transferred to the woman from whom the eggs had been taken and the number of embryos transferred is limited to three. No fertilized egg may be stored or destroyed. Its effect is to prevent the generation of surplus embryos through IVF, and to ensure that all embryos created are treated as potential life.

The existence of the *Embryo Protection Act* meant that there was no pressure initially for a government response to the stem cell announcements of the late 1990s. However, the adequacy of this arrangement was thrown into question by an attempt by German medical researchers, supported by the German Research Association, to import embryonic stem cells from Israel. They argued that

> German patients and German scientists should not be excluded from scientific progress. The embryo's right to life should be balanced against the scientists' freedom to conduct research, which is also, argued the association, a high-ranking constitutional guarantee. As long as there was a ban on embryo research in Germany, the association continued, the cell lines should be imported.
>
> (Braun, 2005: 47)

This request unleashed a complex, protracted debate within the Bundestag, regarding Germany's special mission to uphold human dignity, the ambit of the *Embryo Protection Act* and the extent to which it should be applied to non-German embryos, and the appropriate forms of compromise between scientific progress and public morality (Jasanoff, 2005). Eventually, the Stem Cell Law (*Stammzellgesetz*) was introduced in 2002, which permits the importation of existing hESC lines if they meet specific requirements, a decision justified with reference to the fact that cell lines were not embryos, and that the destruction of the embryo had already taken place and was not reversible (Sperling, 2004; Jasanoff, 2005).

Space constraints prevent the canvassing of other national debates. However, numerous policy-oriented studies have been carried out into national, regional, and transnational ethical debates about hESC research, including in Israel (Prainsack, 2006), India (Bharadwaj and Glasner, 2009), the UK (Mulkay, 1997), Australia (Harvey, 2008), South Korea (Gottweis and Kim, 2009), the European Union (Gottweis et al., 2009), Singapore (Lysaght and Capps, 2012), China (Salter, 2009a; Sleeboom-Faulkner and Hwang, 2012), and Jasanoff's (2005) comparative study of Germany, the UK, and the USA.

The State and Regulation

It is evident from the discussion above that provincial, national, and regional administrations have been heavily involved in modulating the public response to hESC research and cloning. While this was also true for the earlier public debates around Assisted Reproductive Technologies, the post-1990s state involvement in stem cell debates and regulation has taken place in a significantly different *economic context*. Two factors in particular have shaped this context. First is the aging of the population in the developed nations, and the search for policy options that can address the growing proportion made up of the retired as compared with the working. Aging populations place large burdens on welfare, care, and pension provisions, at a time when governments in developed nations are working through the fiscal contraction precipitated by the global financial crisis of 2007–08. In this context, medical innovation is highly desirable, as it offers the possibility that currently intractable medical conditions associated with aging and the reduction of employment capacity might be newly treatable, rendering the aging body less dependent and able to remain longer in the workforce (Neilson, 2012). This prospect is particularly appealing in the case of stem cell research, precisely because of its promise to *regenerate* adult tissues, and reverse some aspects of the aging process itself (Cooper, 2006). As Sperling (2004) argues in his study of the German debates around the importation of hESC lines described above, those in favor of importation articulate the need to regenerate the German population as a whole through medical progress.

The second factor is the increasing centrality of the knowledge sector in the national economies of both the developed and the emerging nations. The commercial exploitation of basic science underpins the OECD nations' transition from industrial mass-manufacture economies to knowledge economies, dependent on extensive cooperation among university, industrial and government sectors to pursue promising research directions and strategic advantages. Since the late 1990s, the life sciences have assumed an ever-more prominent place in this model of economic development, a strategy thematized by the OECD notion of the *Bioeconomy* (OECD, 2006). While North America and Western Europe have been the major research economies historically, the advent of embryonic and other kinds of stem cell research has also coincided with the changing geopolitics of life science research, particularly with East Asian and South Asian embrace of innovation-driven development (Salter, 2009b). India and China, as well as South Korea, Taiwan, and Singapore, have invested heavily in biomedical research capacity since the late 1990s, and each of these countries now has a vigorous stem cell research sector (Sleeboom-Faulkner and Patra, 2011; Lysaght and Capps, 2012). Hence stem cell research has become an element in international economic competitiveness.

For these reasons, since the early 2000s many governments have sought to support stem cell research through the development of sector-specific innovation funding and infrastructure building (Harvey, 2012). In the UK, for example, these initiatives have included the creation of the UK Stem Cell Bank and Stem Cell Network, and the allocation of ring-fenced funding through the various Research Councils. In 2007 the Singapore government allocated $8 billion (US) to 2010 for life sciences, including stem cell science, and has made massive investments in research infrastructure since the early 2000s with the creation of the Biopolis life sciences research park (Waldby, 2009). In the USA, in the face of Federal funding restrictions, individual states have undertaken the task of supporting and coordinating innovation, notably with California Proposition 71, which established a $3 billion program in stem cell science by raising state bonds to fund the California Institute for Regenerative Medicine (Adelson and Weinberg, 2010). It is notable that, in some cases at least, research funding has been extended to the social sciences, in recognition of the social complexity of stem cell research. This has been particularly true of the UK, through the ESRC Social Science Stem Cell Initiative, and through project work carried out under the auspices of the ESRC Genomics Network (2002–13).

At the same time, as mentioned above, some states have initiated deliberative processes in order to adjudicate various social views and formulate workable regulations. One notable response has been the creation, with varying degrees of success, of peak bioethics councils or cognate bodies to formulate robust regulations. So, for example, in the USA, President Clinton's relatively liberal National Bioethics Advisory Commission was displaced in 2002 by the Bush Administration's President's Council on Bioethics, with a more conservative membership to appease the right to life constituencies in the Republican Party and the polity at large. Its first action was to advocate a ban on human 'cloning-to-produce children' and a moratorium on human 'cloning-for-research'. In Germany in 2001, Chancellor Schroder formed the National *Ethikrat* to advise him on hESC research in a very hostile social climate. The *Ethikrat* however never developed social acceptance, because it was perceived as too aligned with ruling party interests (Gottweis et al., 2009). In Singapore, a Bioethics Advisory Committee was formed in 2001 to formulate regulations for hESC research, and it has successfully established

itself as a legitimate forum for ethical reasoning and social deliberation on a number of other areas of biomedical research (Reubi, 2010).

The importance of these peak forums is demonstrated in the breach by the Hwang scandal in South Korea. The Hwang case attracted an extensive body of analysis and commentary from science and technology studies scholars, in particular Hong (2008), Kim (2008), Leem and Park (2008), and Gottweis and Kim (2009), as they try to come to grips with different aspects of the case. In terms of regulatory governance, the Hwang case demonstrates the legitimacy problems created both for science and public administration when medical research is poorly regulated. Professor Hwang Woo Suk and his team published two fraudulent papers in *Science* (in 2004 and 2005) falsely claiming to have created the world's first human SCNT cell lines. The fraud was eventually exposed, with dramatic consequences not only for Professor Hwang but also for the government. In subsequent analysis, it became evident that the research fraud was partly the result of granting exceptional status to Hwang's laboratory, and the removal of all obstacles from his research, including oversight. Hence while the Ministry of Science formed a national Bioethics Advisory Committee in 2001, in response to stem cell controversy, its recommendations went largely ignored by government. It was later chaired by Hwang's lawyer Yang Sam-sung. Hwang's SCNT research was later granted an exemption from statutory regulation by the 2005 *Bioethics and Biosafety Act*. The distribution of grants for stem cell research was not based on peer review, but was rather a government decision, and the South Korean government was intent on promoting Hwang as a national hero, so his laboratory received extensive support without scientific oversight. As a consequence, the exposure of the fraudulent research nearly brought down the government. Both the Minister for Science and the head of the BAC resigned.

Feminist Critiques

The Hwang scandal also galvanized and brought wider attention to the feminist critiques of stem cell research. This was the case because Hwang and his team, in their attempt to create human SCNT cell lines, had procured 2221 eggs from 121 women, most from brokers that normally sold Korean oöcytes to Japanese fertility tourists. The rest were donated, including from two research assistants from Hwang's own laboratory team (Leem and Park, 2008). When the research fraud scandal broke, the world's media were also alerted to an ethical issue that had gone relatively unremarked in the early 2000s debates about the status of embryonic life. That is, the dependence of SCNT and many other kinds of stem cell research (embryonic, hematopoietic, and fetal) on women's reproductive biology and the often-onerous nature of the medical processes that precede donation. This dependence has motivated a range of feminist scholarship since the early 2000s. This includes in particular a body of ethnographic and interview-based research that has engaged with the presumptions of regulation to test out the effects of particular policy settings in the clinic itself. So, for example, in most countries that permit and regulate hESC research, embryonic tissue is obtained through the solicitation of what are termed 'spare' embryos – produced as part of *in vitro* reproductive procedures but designated as being surplus to reproductive requirements (Gottweis et al., 2009). However, clinical studies and interviews with both IVF patients and clinicians have demonstrated that the designation of an embryo as 'spare' is not particularly straightforward, either for the donors or medical staff, and that the consent procedures around this process often fall short in recognizing this complexity (Scott et al., 2012). Studies have probed the ways women and couples understand their embryos, including what it means to either bank or discard them (Nisker et al., 2010), and the differences between donating for other couples' reproductive needs, for reproductive medical research and stem cell research (Waldby and Carroll, 2012). It is notable that, despite the fact that almost all of the initial controversy around embryonic stem cell science revolved around the special status of the embryo as an entity with a potential biographical life, a finding of most studies is that women prefer to donate to research where the embryo will be disaggregated, rather than to feel continuing responsibility for what they regard as a genetic child and sibling of their own children; one that has been effectively 'adopted' by unknown parents. A body of feminist research has also been generated around other kinds of stem cell research, notably egg donation for SCNT (Dickenson, 2005; Leem and Park, 2008; Baylis, 2009; Haimes et al., 2012), and the donation of terminated fetuses (Kent, 2012; Pfeffer, 2008) for the derivation of various kinds of stem cells. This latter body of research is concerned to demonstrate the ethical and political complexities of these forms of tissue procurement, despite the absence of embryos.

Translational Medicine, Experimentation, and Medical Tourism

The changing nature and geography of stem cell research since the early 2000s has produced a body of recent social scientific research addressed to a different set of issues. This section will address two very significant developments: attempts to develop stem cell therapies, as opposed to simply carrying out laboratory research, and the growth of stem cell tourism.

The term 'translational medicine' came into wide use in the 2000s to describe the attempts to develop actual clinical therapies from laboratory innovations in stem cell research, genomics, proteomics, and other areas of life science experimentation. This model of clinical development has formed the basis for considerable public funding in some jurisdictions (Harvey, 2012). For this reason, it has also attracted critical social science attention, particularly from scholars working in the sociology of expectations, concerned with how particular ways of imagining the future, including the future of innovation, shape understandings of the present. Along these lines, Wainwright and colleagues interviewed stem cell scientists to identify the nonlinear complexities involved in the purportedly linear process involved in translating research from bench to bedside (Wainwright et al., 2006). Meanwhile, Martin and colleagues approached the issue from a historical perspective, investigating the experimental relations between bench and bedside in the history of bone marrow transplantation using hematopoietic stem cells to treat blood disorders (Martin et al., 2008). They demonstrate that the first experimental bone

marrow transplants (BMTs) during the 1960s and 1970s followed a model of clinical innovation, initiated by clinical hematologists to treat patients without other treatment options. These early attempts had catastrophic results, before laboratory research identified the immunological causes of graft vs host disease and developed methods of precise tissue matching. These laboratory findings provoked a second round of much more successful BMT, recently redefined as hematopoietic stem cell therapy, which is now the standard treatment for many blood disorders. It is notable that, in studies that investigate actual clinical trial and hospital practice around stem cell therapies, this model of hospital innovation and clinical tinkering to find workable treatments for very ill patients can also be identified (Raval et al., 2008), and it is thought that the hospital-based approach is likely to dominate stem cell translational medicine for the foreseeable future (Webster et al., 2011). This domination is in part due to the simpler regulatory approvals process involved in batch-by-batch hospital treatments, while formal clinical trials using hESC lines in particular have been caught up with complex compliance and safety issues.

The demand for new therapies among patient groups, and the social expectations built up around stem cell science have also seen the development of a vigorous stem cell 'tourism' sector, as clinics offer generally unproven stem cell-based treatments to patients willing to travel in order to evade clinical regulation constraints in their own countries. This development is consonant with a more general growth of transnational medical tourism, in which patients travel to gain access to treatments that are unavailable or unaffordable in their own locations. It is estimated that around 700 clinics worldwide currently offer stem cell treatments that have not been validated through clinical trial nor formal evidence-based regulatory assessment (Einsiedel and Adamson, 2012). The potential for harm to patients and the use of direct-to-consumer advertising, which relies on anecdotal accounts of therapeutic success rather than validated scientific evidence (Petersen and Seear, 2011), have alarmed regulators and scientific organizations like the International Society for Stem Cell Research, who have produced patient handbooks and voluntary guidelines for clinics in an attempt to introduce a degree of self-regulation (Harvey, 2012). However, the small amount of fieldwork carried out among stem cell tourists suggests that many make highly informed decisions to access unproven treatments, with complex knowledge of the risks and benefits that may be involved (Song, 2010; Petersen and Munsie, 2012).

See also: Biocitizenship; Bioethics in the Post-genomic Era; Bioethics: Genetics and Genomics; Biomedical Sciences, History and Sociology of; Biotechnology; Genomics, Ethical Issues In; Patenting in the Biological Sciences; Pharmaceutical Industry: Political Economies of Drugs and Knowledge; Risk, Sociology and Politics of; Science and Law; Science and The State.

Bibliography

Adelson, J.W., Weinberg, J.K., 2010. The California stem cell initiative: persuasion, politics, and public science. American Journal of Public Health 100 (3), 446–451.

Baylis, F., 2009. For love or money? The saga of Korean women who provided eggs for embryonic stem cell research. Theoretical Medicine and Bioethics 30 (5), 385–396.

Bharadwaj, A., Glasner, P.E., 2009. Local Cells, Global Science: The Rise of Embryonic Stem Cell Research in India. Routledge, London, New York.

Braun, K., 2005. Not just for experts: the public debate about reprogenetics. Hastings Center Report 35 (3), 42–49.

Clarke, A., 1998. Disciplining Reproduction: Modernity, American Life Sciences, and the Problems of Sex. University of California Press, Berkeley, CA.

Cooper, M., 2006. Resuscitations: stem cells and the crisis of old age. Body & Society 12 (1), 1–23.

Dickenson, D., 2005. The lady vanishes: what's missing from the stem cell debate. International Journal of Ethics 3, 43–54.

Dodds, S., 2004. Women, commodification and embryonic stem-cell research. In: Humber, J.M., Almeder, R.F. (Eds.), Biomedical Ethics Reviews: Stem Cell Research. Humana Press, Totawa, NJ, pp. 151–174.

Edwards, R.G., 2001. IVF and the history of stem cells. Nature 413, 349–351.

Einsiedel, E., Adamson, H., 2012. Stem cell tourism and future stem cell tourists: policy and ethical implications. Developing World Bioethics 12 (1), 35–44.

Embryo Protection Act, 1990. Gesetz zum Schutz von Embryonen (Embryonenschutzgesetz – ESchG). Promulgated 13 December, Federal Law Gazette I, p. 2746.

Franklin, S., 2007. Dolly Mixtures: The Remaking of Genealogy. Duke University Press, Durham, NC.

Gottweis, H., Kim, B., 2009. Bionationalism, stem cells, BSE, and Web 2.0 in South Korea: toward the reconfiguration of biopolitics. New Genetics and Society 28 (3), 223–239.

Gottweis, H., Salter, B., Waldby, C., 2009. The Global Politics of Human Embryonic Stem Cell Science: Regenerative Medicine in Transition. Palgrave, Basingstoke, UK.

Habermas, J., 2003. The Future of Human Nature. Polity, Cambridge.

Haimes, E., Taylor, K., Turkmendag, I., 2012. Eggs, ethics and exploitation? Investigating women's experiences of an 'egg sharing' scheme. Sociology of Health and Illness 34 (8), 1199–1214.

Harvey, O., 2008. Regulating stem cell research and human cloning in an Australian context: the Lockhart review. New Genetics and Society 27 (1), 33–42.

Harvey, O., 2012. Commercialising the Stem Cell Sciences. Woodhead Publishing, Cambridge.

Holland, S., Lebacqz, K., Zoloth, L. (Eds.), 2001. The Human Embryonic Stem Cell Debate: Science, Ethics and Public Policy. MIT Press, Cambridge, MA.

Hong, S., 2008. The Hwang scandal that "shook the world of science". East Asian Science. Technology and Society 2 (1), 1–7.

Jasanoff, S., 2005. Designs on Nature: Science and Democracy in Europe and the United States. Princeton University Press, Princeton, NJ.

Kent, J., 2012. Regenerating Bodies: Tissue and Cell Therapies in the Twenty-first Century. Routledge, Abingdon and New York.

Kim, L., 2008. Explaining the Hwang scandal: national scientific culture and its global relevance. Science as Culture 17 (4), 397–415.

Leem, S., Park, J., 2008. Rethinking women and their bodies in the age of biotechnology: feminist commentaries on the Hwang affair. East Asian Science, Technology and Society: An International Journal 2 (1), 9–26.

Lysaght, T., Capps, B.J., 2012. Public discourses of stem cell science in Singapore (1997–2010). New Genetics and Society 31 (4), 342–358.

Martin, P., Brown, N., Kraft, A., 2008. From bedside to bench? communities of promise, translational research and the making of blood stem cells. Science as Culture 17 (1), 1–13.

Mulkay, M.J., 1997. The Embryo Research Debate: Science and the Politics of Reproduction. Cambridge University Press, New York.

Neilson, B., 2012. Ageing, experience, biopolitics: life's unfolding. Body & Society 18 (3–4), 44–71.

Nisker, J., Baylis, F., Karpin, I., McLeod, C., Mykitiuk, R. (Eds.), 2010. The 'Healthy' Embryo: Social, Biomedical, Legal and Philosophical Perspectives. Cambridge University Press, Cambridge, NY.

OECD, 2006. The Bioeconomy to 2030: Designing a Policy Agenda. OECD International Futures Programme, Paris.

Petersen, A., Munsie, M., 17 October 2012. Marketing Hope: The Dynamics of Expectation in Relation to Stem Cell Treatments. Annual Meeting of the Society for Social Studies of Science, Copenhagen Business School. Denmark, Fredericksburg.

Petersen, A., Seear, K., 2011. Technologies of hope: techniques of the online advertising of stem cell treatments. New Genetics and Society 30 (4), 329–346.

Pfeffer, N., 2008. What British women say matters to them about donating an aborted fetus to stem cell research: a focus group study. Social Science & Medicine 66, 2544–2554.

Prainsack, B., 2006. 'Negotiating life': the regulation of human cloning and embryonic stem cell research in Israel. Social Studies of Science 34, 1–33.
Raval, A., Kamp, T., Hogle, L., 2008. Cellular therapies for heart disease: unveiling the ethical and public policy challenges. Journal of Molecular and Cellular Cardiology 45, 593–601.
Reubi, D., 2010. The will to modernize: a genealogy of biomedical research ethics in Singapore. International Political Sociology 4 (2), 142–158.
Salter, B., 2009a. China, globalisation and health biotechnology innovation: venture capital and the adaptive state. East Asian science. Technology and Society 3, 401–420.
Salter, B., 2009b. State strategies and the geopolitics of the global knowledge economy: China, India and the case of regenerative medicine. Geopolitics 14 (1), 47–48.
Scott, R., Williams, C., Ehrich, K., Farsides, B., 2012. Donation of 'spare' fresh or frozen embryos to research: who decides that an embryo is 'spare' and how can we enhance the quality and protect the validity of consent? Medical Law Review 20 (3), 255–303.
Sleeboom-Faulkner, M., Hwang, S., 2012. Governance of stem cell research: public participation and decision-making in China, Japan, South Korea and Taiwan. Social Studies of Science 42 (5), 684–708.
Sleeboom-Faulkner, M., Patra, P.K., 2011. Experimental stem cell therapy: bio-hierarchies and bionetworking in Japan and India. Social Studies of Science 41 (5), 645–666.
Song, P., 2010. Biotech pilgrims and the transnational quest for stem cell cures. Medical Anthropology 29 (4), 384–402.
Sperling, S., 2004. From crisis to potentiality; managing potential selves: stem cells, immigrants, and German identity. Science and Public Policy 31 (2), 139–149.
Thomson, J., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, A.A., Swiergiel, J.J., Marshall, V.S., Jones, J.M., 1998. Embryonic stem cell lines derived from human blastocysts. Science 282, 1145–1147.
Wainwright, S.P., Williams, C., Michael, M., Farsides, B., Cribb, A., 2006. From bench to bedside? Biomedical scientists expectations of stem cell science as a future therapy for diabetes. Social Science & Medicine 63 (8), 2052–2064.
Waldby, C., 2009. Singapore biopolis: bare life in the city state. East Asian science. Technology and Society 3 (2/3), 367–383.
Waldby, C., Carroll, K., 2012. Egg donation for stem cell research: ideas of surplus and deficit in Australian IVF patients' and reproductive donors' accounts. Sociology of Health & Illness 34 (4), 513–528.
Webster, A., Haddad, C., Waldby, C., 2011. Experimental heterogeneity and standardisation: stem cell products and the clinical trial process. Biosocieties 6 (4), 401–419.
Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., Campbell, K.H., 1997. Viable offspring derived from fetal and adult mammalian cells. Nature 385 (6619), 810–813.
Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., et al., 2007. Induced pluripotent stem cell lines derived from human somatic cells. Science 318 (5858), 1917–1920.

Relevant Websites

www.genomicsnetwork.ac.uk/ – ESRC Genomics Network. 2002–13 (accessed 17.02.14.).
www.york.ac.uk/res/sci/introduction.htm – ESRC Social Science Stem Cell Initiative. 2005 (accessed 17.02.14.).

Stereotypes in Social Psychology

Olivier Klein and Philippe Bernard, Université Libre de Bruxelles, Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M.R. Banaji, volume 22, pp. 15100–15104, © 2001, Elsevier Ltd.

Abstract

Social stereotypes fulfill a variety of functions and influence social judgment via both automatic and controlled cognitive processes. The formation of shared stereotypes reflects a group's attempt to appraise the social position of relevant out-groups in line with its preexisting norms and beliefs. Competence and warmth organize the content of stereotypes. The position of an out-group on these two dimensions is a function of its perceived status and level of cooperation toward the in-group. When they are shared within majority groups, stereotypes can exert a variety of undesirable influences on the members of minority group.

What Are Stereotypes?

A stereotype is an individual's set of beliefs about the characteristics or attributes of a group (Judd and Park, 1993). Such attributes can typically consist in general psychological traits. A stereotype can also include behaviors considered typical of the members of this social category. Other possible aspects of the stereotype include information regarding the typical norms of this social category or its social status. A stereotype needs not to be disparaging: It can include desirable traits. Stereotypes, being beliefs, should be clearly distinguished from prejudice, which refers to a negative evaluation of a target group. Naturally, of course, stereotypes and prejudice often go hand in hand. Negative stereotypes can foster or justify prejudice.

According to some theorists (e.g., Oakes et al., 1994), a trait can be considered as stereotypical of a given group only if it is perceived as more likely to be present in this group than in other relevant comparison groups. Thus, stereotypes are eminently context dependent. Often, the reference group will be one's own group, in such a way that stereotypes of out-groups are mirror images of one's own (Oakes et al., 1994).

Another definitional question concerns sharedness. Most stereotypes are known by many members of a social group. Hence, they can inform a group's behavior and policies in a much more powerful way than the unique beliefs of a single individual could. This makes it all the more difficult for the members of the target group to escape their influence. The shared character of stereotypes has led Devine (1989) to distinguish cultural stereotypes from personal beliefs. The former refers to stereotypes known to be held in the community but that one does not necessary endorse. The latter concerns beliefs that are subjectively endorsed.

Stereotypes are inherently associated with social categorization, the cognitive process by which individuals are sorted into various social categories. Categorization is a flexible process that depends both on characteristics of the perceived stimuli (in this case individuals) and of the perceiver. People can categorize others in two groups because members of these two groups blatantly differ in their appearance or behavior. In this case categorization is driven by observed regularities as a function of the relevant categorization dimension or *comparative fit* (Oakes et al., 1994). Consider now a situation in which a teacher expects girls to be stronger in literature and boys better in maths (a common stereotype). He may use the gender category to organize his perceptions of the students *even if performance differences between the two genders do not correspond to these expectations*. Here, what determines categorization is 'normative fit' (Oakes et al., 1994). Thus stereotype emerge as much from categorization (comparative fit) as they determine it (normative fit).

It is common to consider stereotypes as an inventory of unrelated traits. Yet, stereotypes often function as 'naive theories' the content of which is dependent on the nature of the associative links between the traits. For example, two individuals may endorse the view that members of a group are 'poor' or 'lazy'. However, they may differ in the theories explaining this association. One may view 'laziness' as causing 'unemployment' whereas another may consider this trait as stemming from discouragement at not being able to find work due to pervasive discrimination. Thus stereotypes not only describe members of a social group but explain *why* they are *how* they are.

Stereotypes often invoke the existence of deep underlying properties that may account for the homogeneity of the group through a form of perception known as *psychological essentialism* (Yzerbyt et al., 1997). Essentialism can be biological when it involves considering that the group's traits stem from shared biological (e.g., genetic) properties. It can also be cultural, when a culture is viewed as determining deep psychological orientations of its members.

The stereotype, as a cognitive structure, should be distinguished from the process of *stereotyping*. Stereotypes involve applying knowledge included in this cognitive structure to specific individuals.

Mental Representation of Stereotypes

Stereotypes have been initially conceptualized as lists of traits describing the relevant category. This representation is then used to appraise new information regarding members of the target group (Katz and Braly, 1933). Critics of this view (e.g., Brewer et al., 1981) have argued that it did not consider variations in the *typicality* of group members. The stereotype is

not applied in an 'all-or-none' manner as the classical view would argue. Rather, the probability that a person belongs to a given category depends on the degree to which it matches the 'typical' member of the category, the 'prototype'.

Both the 'classical' and the 'prototypical' view consider that stereotypes are generalizations based on a large number of observations. A contrasting view (Smith and Zarate, 1992) argues that people hold information about group members in the form of 'exemplars' i.e., individual members of whom they have formed an impression rather than in the form of general abstract knowledge. When judging a member of a social group, or this group as a whole, they rely on the relevant members of this social category that are cognitively accessible rather than on general traits ascribed to the entire group. In this view, stereotypes do not exist as abstract knowledge structures but are just judgments based on whatever exemplars have come to mind.

This approach allows to take into account an important aspect of people's representations of social groups: Even if people strongly adhere to a stereotype, they still acknowledge that group members may vary in their conformity to this stereotype, or even sometimes strongly deviate from it.

In an attempt to reconcile these diverging views, Sherman (1996) has suggested that we need not necessarily choose between an 'abstract' and a 'concrete' model of stereotypes. Indeed, when people form stereotypes of unfamiliar groups, they initially rely on the few exemplars they know. Increasingly, as they become more familiar with the group, they develop more abstract, schematic, representations.

Assessment of Stereotypes

Historically, stereotypes have been assessed through self-report methods, that is, by asking people to explicitly report their beliefs about social groups. For example, the 'checklist approach' pioneered by Katz and Braly (1933) involved asking participants to select five traits descriptive of the target group in a list of 84. Another self-report approach involves asking respondents to estimate the extent to which a group can be characterized by a given trait on rating scales. Thanks to statistical techniques, it is possible to inspect the correlations between the ratings on the individual traits and, in turn, to extract 'dimensions' assumed to organize the stereotype (cf *infra*: the content of stereotypes).

Stereotypes can be conceptualized as representations of the distribution of a set of traits in a group. If this is the case, assessment of stereotypes should provide information on several aspects of these distributions. One such aspect is called its central tendency: What is the 'middle' of the distribution? For example, what level of warmth is expected of the average Italian? Central tendency has received the greatest attention in stereotyping research. Another more neglected aspect of a distribution however relates to dispersion. To what extent are members of the target group perceived to be similar on the given trait, to 'cluster' around the central tendency? People can perceive a trait as very prevalent and yet consider that there is a wide variation in the extent to which members of the target group display this trait. Judd and Park (1993) have developed rating scales assessing the perceived dispersion of a trait in a group.

In spite of their widespread use, rating scales suffer from an important limitation. In general, they do not allow to explore the richness of the explanatory structures underlying stereotypes. Often only the use of more open ended, qualitative methods, does more justice to this richness.

Social norms proscribe the expression and the endorsement of stereotypes about certain groups. If this is the case, self-report measures may be inappropriate as people may hide their true beliefs toward the target group. To circumvent this problem, new measures of stereotypes have been developed by relying on computerized techniques. Typically, these techniques assess *associations* between the social category and a given set of traits. A popular method, the 'Implicit Association Test' (Greenwald et al., 1998), assesses people's associations between the target category and a given trait or set of traits based on reaction times. Unfortunately, such methods remain sensitive to social desirability concerns (although less so than self-report methods). As we should expect, they are not always consistent with explicit ratings and the relation between implicit and explicit ratings is on average quite low although it can increase under specific conditions.

Functions of Stereotypes

Tajfel (1981) distinguishes the *individual* and *social functions* of stereotypes. At an individual level, stereotype first serve a function of cognitive economy: Indeed, stored knowledge about an individual's group may be readily applied to interpret his or her behavior, or even to evaluate this person, without having to engage in in-depth impression formation. This function is particularly meaningful when cognitive resources are taxed or when facing a large number of individuals.

Stereotypes may also help guide behavior. By relying on stored information regarding the members of a social group, one can more easily behave in appropriate ways toward members of this group.

At a collective level, stereotypes are used by groups to explain societal events. This function is particularly salient in conspiracy theories depicting members of a social group as the organizers of a plot leading to an unfortunate event.

One step further, stereotypes can be used to *justify* the existing social structure and/or the discriminatory behaviors displayed by members of the dominant groups. By attributing 'essential' traits to members of a stigmatized group, they may conceal the role of power dynamics and oppression, portraying the group's position as the outcome of its inner characteristics (cf Jost et al., 2004). For example, gender inequality may be attributed to women's natural inclinations toward child rearing and family life.

Finally, according to Tajfel, stereotype may be used to differentiate the in-group from relevant out-groups. Thus, by portraying the in-group in a favorable light, compared to the out-group, stereotypes may promote or maintain a 'positive social identity'. This may explain why self-stereotypes (i.e., stereotypes of the in-group) are generally more positive than heterostereotypes (i.e., stereotypes of the out-group).

In view of these functions, stereotypes appear like interpretational devices that help individuals and group appraise and understand others' behaviors. Stereotyping does not

necessarily imply imposing an a priori view of someone on his or her behavior. Much more subtly, it may help fill the gaps in our understanding of others when we lack all the information necessary to make a judgment.

Since Tajfel's seminal work, other functions of stereotypes have been uncovered.

At the individual level, stereotypes may help individuals cope with the fear of death. Indeed, according to terror management theory, people hold 'cultural worldviews', i.e., a set of beliefs about the organization of the world and about important values shared within a culture (Schimel et al., 1999). According to this theory, the primary function of these worldviews, of which cultural stereotypes are part, is to protect people against the fear of their own mortality.

A collective function that has attracted increasing attention is collective mobilization. Stereotypes are very efficient rhetorical devices, which can be used to mobilize other group members into supporting the speaker's political projects (Klein et al., 2007). As stereotypes are collectively understandings of the important groups composing one's social environment, they provide a readily accessible 'image' of the social structure on which group leaders can capitalize to mobilize support for their political projects.

Stereotype Content

The content of stereotypes is generally organized around two broad dimensions (Fiske et al., 2002): One pertains to the group's skills and capacities (the competence dimension); another to his or her intentions toward the self or one's in-group (the warmth dimension). As a consequence, it is possible to classify groups in four quadrants depending on their position on these two dimensions. Such an analysis reveals that most out-groups are perceived ambivalently as high on one dimension and low on the other. Thus, envied groups (e.g., Jews and Asians in the US) are high on competence, and low on warmth. By contrast, other groups, victims of paternalistic prejudice (e.g., housewives and disabled people in the US) are low on competence and high on warmth (cf **Table 1**).

Table 1 Four types of out-groups, combinations of status and competition, and corresponding forms of prejudice as a function of perceived warmth and competence

	Competence	
Warmth	*Low*	*High*
High	Paternalistic prejudice Low status, not competitive Pity, sympathy (e.g., elderly people, disabled people, housewives)	Admiration High status, not competitive Pride, admiration (e.g., in-group, close allies)
Low	Contemptuous prejudice Low status, competitive Contempt, disgust, anger, resentment (e.g., welfare recipients, poor people)	Envious prejudice High status, competitive Envy, jealousy (e.g., Asians, Jews, rich people, feminists)

Reproduced from Fiske, Susan T., Cuddy, Amy J.C., Glick, Peter, Xu, Jun, 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Importantly, the position of a group on these two dimensions can be predicted by sociostructural variables. Perceived warmth is a function of perceived cooperation (vs competition) between the in-group and the out-group whereas the competence dimension is a function of perceived status.

This is consistent with the suggestions that stereotypes not only serve to describe the out-group but to explain, and even justify, this group's relative position within the social structure vis-à-vis the in-group.

Recent research on these two dimensions of social judgment suggests that they may be negatively related. Thus when a group is perceived as high on one dimension, it may tend to be perceived as low on the other. A similar pattern holds when making intergroup comparisons: when compared to a more competent group, a group is also perceived as warmer.

A specific form of group perception that has attracted attention recently concerns the perceived *humanity* of groups (Haslam and Loughnan, 2014). Historically, foreign, or unknown, groups have often be portrayed in animal terms (e.g., Rwandan Tutsis as 'cockroaches' by the Hutu propaganda during the genocide). An indicator of humanization, as opposed to animalization, is the attribution of emotions. Some emotions, called 'secondary emotions' are typically human (e.g., nostalgia, guilt,…) whereas others (e.g., fear, joy), called 'primary emotions', are common to both humans and animals. People tend to routinely 'infrahumanize' out-groups by attributing them fewer secondary emotions than their in-group. Such a form of group perception seems particularly likely when the out-group is potentially threatening to the in-group.

Stereotype Accuracy

Whereas stereotypes have often been characterized as 'inaccurate' or 'exaggerated', current definitions do not consider inaccuracy as a property of stereotypes.

Researchers trying to estimate the accuracy of stereotypes have found that, on average, they tended to be relatively accurate at predicting the central tendency of the relevant traits in the target group (Jussim, 2012). If stereotypes are sometimes good at predicting group level tendencies, they are generally of little use in predicting an individual's characteristics. This is because social category membership explains very little variance on most psychological traits.

Nevertheless, comparing stereotypes with 'reality' poses methodological and conceptual difficulties.

First, given that stereotypes help explain relations between social groups, it should come as no surprise that they may also change as a function of the evolution of these relations (e.g., Poppe, 2001). An individual's stereotype may also vary depending on the situation in which the target group is evaluated, and of the group to which it is compared (Oakes et al., 1994). Thus, the stereotype to be gauged in terms of accuracy is often an elusive target.

Second, the norm against which stereotypes are compared, 'reality', can change very quickly as well. For example, in a study by Guimond et al. (2006), differences between the

self-concepts of men and women were a function of the context in which participants were asked to describe themselves. Gender differences, with men being more 'independent' and less 'interdependent' than women, were most likely to emerge when group members compared themselves to members of the other gender (i.e., because it activated their gender identity).

Third, when assessing stereotype accuracy, it is important to distinguish the accuracy of the description from the accuracy of the explanation. Establishing a correlation between a social category and a stereotypical trait does not demonstrate that the social category plays a direct role in the existence of this covariation (i.e., correlation does not imply causation). Thus, finding that a stereotype is 'accurate' does not tell us anything as to why there is a correlation between the category and the trait.

Fourth, a reason why stereotypes can be accurate is that they can produce the reality they describe even when they are initially inaccurate (i.e., 'self-fulfilling prophecy'). Thus, when large segments of a powerful group hold a stereotype of another, less powerful group, they may influence members of the latter into adopting social roles and positions that call for the qualities that are part of the stereotype (Klein and Snyder, 2003).

Development and Diffusion of Stereotypes

Children acquire stereotypes as early as four. It is tempting to view children's minds as a 'blank slate' on which adults' views are progressively imprinted. Reality is more nuanced. According to developmental intergroup theory (Bigler and Liben, 2007), stereotypes develop in three stages.

First, children identify person dimensions (e.g., sex, gender, occupation) as relevant. This may occur because a given dimension is highly visible or discriminable. It may also be because it is used by peers and adults to characterize others.

Second, children use these dimensions to characterize others.

Finally, they develop stereotypes. The content of these stereotypes depends on four different factors:

- Essentialism or the tendency to attribute essences (cf supra) to social groups.
- In-group bias refers to a motivation to describe one's group in a favorable light compared to other groups even in the absence of evidence. Thus, children may literally invent stereotypes to depict their group in a positive light.
- Stereotypes can be shaped by adults' or peers' explicit remarks about the characteristics of social groups.
- Finally, children develop stereotypes based on the observation of correlations between social categories and attributes. For example, children may notice that most teachers are female or that most politicians are male.

Thus, the development of social stereotypes is not simple brainwashing. Neither should it be understood as a pure inductive process based on the solitary observation of the social world. Children elaborate stereotypes to the extent that these offer them as satisfactory understanding of the world. They depend on the frame of references provided by their group's norms and beliefs. Peers and parents provide them cognitive tools that may help them appraise the humans that surround them. For this reason, children's, like adults', stereotypes are malleable and highly dependent on the social context in which the target group is judged (Oakes et al., 1994).

Generally, stereotypes are not maintained if they are not shared and communicated within the group, a process known as consensualization. As people communicate about others, they tend to preferentially select information that conforms to existing stereotypes, thereby contributing to their diffusion and maintenance (cf Kashima et al., 2007). By being communicated within the group, via interpersonal communication or the media, stereotypes are socially validated and become shared understandings that are incorporated in the group's view of the organization of the social world, i.e., its ideology.

Stages Involved in Stereotyping

How do stereotypes affect the impression human form of other persons? In their influential *continuum model*, Fiske et al. (1999) suggest that, when encountering another person (the 'target'), one may engage in two types of processing. At one extreme, one can attend to the person's appearance and behavior and derive an individualized impression of this person. At the other extreme, one can categorize a person in a social group and perceive this person in terms of the stereotype associated with this social category regardless of his or her observable attributes. The latter type of processing is less demanding than the former in terms of cognitive resources. As a consequence, it will often tend to be preferred especially when attentional resources are scarce, when the perceiver is under time pressure or when there are limited opportunities for more effortful deliberate processing.

The continuum model (Fiske et al., 1999) considers impression formation as involving sequential stages: People initially try to categorize the person in terms of the activated stereotype, process information relevant to this stereotype only. If they do not succeed, they recategorize the person in terms of a new stereotype and consider information relevant to this new category. They attend to individuating information that does not fit into any accessible stereotype only when all attempts at categorization fails. In contrast to such a *serial model*, Kunda and Thagard (1996) have suggested that people often process category based and individuating information *in parallel*. In this view, all the information about a target is activated concurrently and may influence the impression simultaneously.

To impact on judgment of others and behavior, stereotypes first need to be made accessible for further treatment, a process known as *activation*. An important question concerns whether such an activation is automatic or controlled: After having categorized a person in a given group, can observers refrain from activating the associated stereotype? A process is considered automatic, as opposed to controlled, if it is unintentional, efficient (i.e., requires few cognitive resources), occurs outside awareness and is uncontrollable once triggered. Stereotype activation satisfies many of these criteria. When people are confronted with cues associated with a person's social category, stereotype consistent information becomes more easily accessible. This occurs even when these cues are presented subliminally or cannot be consciously processed (Devine, 1989).

This means that stereotype activation often occurs without control, awareness of intention.

However, stereotype activation remains susceptible to controlled processes (Blair, 2002). The activation of stereotypes can be influenced by the perceiver's goals. In certain cases, the regulation of stereotypes may become controllable. Also, activation seems to require some cognitive resources. Hence, it is best characterized as 'conditionally' automatic.

In addition, members of a cultural group do not necessarily activate the same stereotype when confronted with a social category. People low in prejudice, or those chronically motivated by egalitarian norms, may activate more positive stereotypes than others. They may also be more capable of regulating the impact of stereotypes on their judgment of others. This form of control may become so habitual in 'egalitarians' that it occurs swiftly and without awareness (Moskowitz, 2010). Such findings blur the boundary between 'automatic' and 'controlled' processes.

Once stereotypes are activated they may influence several stages of impression formation: Thus, they may lead observers to focus on dimensions of the target's personality that are associated with the stereotype and fail to notice aspects of the target's personality that are not represented in the stereotype. They may also lead observers to interpret ambiguous behavior in line with their stereotypical expectations. The activation of stereotype seems to trigger associations that thoughtlessly influence the interpretation of ambiguous behavior (cf Devine, 1989).

Finally, under certain conditions, the activation of stereotypes by a 'prime' may induce people to unknowingly behave in line with the traits that are part of the stereotype. For example, in a classic study, people walked more slowly after the concept of old age was activated (Bargh et al., 1996). Due to a common representational basis, the activation of the trait may lead to the production of the behavior. Whether such effects are truly automatic has been a matter of controversy (cf Doyen et al., 2012).

Stereotype activation should be distinguished from stereotype application: People can inhibit the application of a previously activated stereotype, for example because this stereotype does not conform to social norms. However, such inhibition attempts could backfire (Macrae et al., 1994) by making it more accessible. As a consequence, inhibition may subsequently increase the chances that it will color judgment of others once the perceiver does no longer try to suppress it, a pattern known as a 'rebound' effect.

Stereotype application may also be encouraged in response to specific motivations. For example, people have been found to apply stereotype when their social identity, self-esteem, or core aspects of their cultural worldview were threatened.

Stereotype Change

Changing an individual's beliefs regarding a target group is rarely a simple matter. Intergroup contact (Allport, 1954/1979) is a much-studied strategy for modifying stereotypes and decreasing prejudice. This intervention is presumed on the assumption that negative stereotypes may be due to insufficient knowledge of the out-group. Under certain conditions, developing environments allowing intergroup contact has indeed been shown to decrease prejudice and stereotyping (Pettigrew and Tropp, 2008). However, many hurdles have to be cleared in order for exposure to counterstereotypical group members to undermine stereotypes (Klein and Snyder, 2003). First, counterstereotypical behaviors need to be noticed and interpreted as such. If this is the case, people may still characterize such behavior as driven by situational factors rather than by the persons' dispositions. Even when observers attribute such behavior to the target's dispositions, they may still 'fence off the target' and characterize her as an atypical exemplar of her social category, a process known as 'subtyping'. Thus, for stereotype change to take place, the target needs to be perceived as typical of the group in spite of her stereotype inconsistent behavior. Finally, although a single deviant group member may sometimes affect perceived group variability ('they are not all alike'), observers may need to be exposed to many instances of stereotype disconfirming behaviors, from different members of the target group, to expect an alteration in this group's perceived 'central tendancy' on the given trait.

Thus the journey to stereotype change through the cognitive pathway is a rocky one. If contact may lead to a change in stereotypes and prejudice, it may be primarily by influencing the observer's affective ties with the out-group (e.g., by making friends in the out-group) rather than via cognitive processes exclusively (Pettigrew and Tropp, 2008).

Since stereotypes are inherently associated with the social position of the in-group and out-group, the best hope for change is the evolution of these relations. It is only insofar as stereotypes successfully explain the organization of the social structure in which groups are placed that they can persist.

Influence of Stereotypes on the Target Group

Members of groups targeted by social stereotypes are affected by these in many ways. Some of these consequences are relatively direct, such as discrimination. Indeed, stereotypes can prompt or justify discrimination (Allport, 1954/1979).

However, they are more subtle consequences of negative stereotypes. An important such consequence has to do with the difficulty of identifying whether other people's negative behaviors toward the self are due to the stereotype or, worse, to prejudice than to other factors. Attributing such behavior to prejudice may protect self-esteem (Crocker et al., 1991). For most people, performance feedback may also be used to correctly assess one's capacities. When one belongs to a stigmatized group, suspecting that this feedback is biased by stereotypes may promote a sense of uncertainty about one's capacities.

When stereotypes depict the abilities of members of a group in a negative way, stigmatized group members may be tempted to psychologically withdraw ('disengage') from this domain. If they are nevertheless confronted with the stereotype, they may also adopt compensatory strategies to help maintain their self-esteem in case of failure, such as devoting less effort to the task in an attempt to 'explain away' a possible failure.

People who keep identifying with a domain in which their group is viewed negatively may succumb to another hazard known as 'stereotype threat'. In a pioneering study, Steele and Aronson (1995) have found that the activation of the African-

American stereotype lead members of the stereotyped group to perform more poorly in a stereotype relevant task (an intelligence test) than if they had not been reminded of the stereotype. Similar findings have been reported for other target groups and stereotypes (Schmader et al., 2008). This phenomenon depends on the situational salience of a stereotype and does not specifically target chronically stigmatized groups.

Stereotype threat effects on cognitive performance occur in domains that demand a high level of mental concentration. The activation of the stereotype seems to mobilize much needed cognitive resources. And indeed, impaired working memory capacity has been shown to explain womens' tendency to underperform in maths when the stereotype is salient (Schmader et al., 2008).

To combat the stigma associated with negative stereotypes, members of the stigmatized group may engage in two types of strategies (Klein and Snyder, 2003).

One type of strategy is individual and involves dissociating the self from the negative stereotypes to achieve positive relations with members of the nonstigmatized group. Thus, the stigmatized compensatory behaviors, such as behaving more assertively or devoting more effort to establish smooth interpersonal relationships. Reacting to the stereotype is of course much easier when it is blatantly endorsed or expressed by members of the nonstigmatized group. More often than not, however, the stereotype is conveyed subtly or implicitly and it can be difficult for members of the target group to notice it and therefore to react effectively.

Rather than dissociating themselves from their stigmatized group, people may adopt a 'collective strategy' and choose to confront the image held by the dominant group about their in-group, or 'metastereotypes'. Faced with negative metastereotypes, group members may collectively strive at changing it in an effort to upgrade their image (Klein et al., 2007). However, in other circumstances, stigmatized group members actively conform to negative metastereotypes in order to distance themselves from the dominant group (Kamans et al., 2009).

Whether people resort to the 'individual' or 'collective' strategies seem to depend once again on the nature of the relations between the groups and especially on several factors: first, whether the boundaries between the in-group and the out-group are permeable (i.e., whether it is possible to 'pass' from one group to the other). Second, whether the status differences between their stigmatized in-group and the nonstigmatized group is perceived as legitimate (which may be especially the case, when the stigma, such as obesity for example, is perceived as controllable) and stable. According to social identity theory (Tajfel and Turner, 1986), the use of collective strategies aimed at upgrading a stigmatized social identity is most likely when group boundaries are perceived as impermeable, and status differences are perceived as either unstable and/or illegitimate. These predictions have generally been supported although little research has considered how these applied to stereotypes per se (Klein and Snyder, 2003).

See also: Essentialism; Implicit Social Cognition; Intergroup Relations; Self-Categorization Theory; Social Categorization; Social Cognition; Social Identity in Social Psychology; Social Psychology; Stigma, Social Psychology of.

Bibliography

Allport, Gordon W., 1954/1979. The Nature of Prejudice: 25th Anniversary Edition, Unabridged. Basic Books.

Bargh, John A., Chen, Mark, Burrows, Lara, 1996. Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology 71, 230–244.

Bigler, Rebecca S., Liben, Lynn S., 2007. Developmental intergroup theory explaining and reducing children's social stereotyping and prejudice. Current Directions in Psychological Science 16, 162–166.

Blair, I.V., 2002. The malleability of automatic stereotypes and prejudice. Personality and Social Psychology Review 6 (3), 242–261.

Brewer, Marilynn B., Dull, Valérie, Lui, Layton, 1981. Perceptions of the elderly: stereotypes as prototypes. Journal of Personality and Social Psychology 41, 656–670.

Crocker, Jennifer, Voelkl, Kristin, Testa, Maria, Major, Brenda, 1991. Social stigma: the affective consequences of attributional ambiguity. Journal of Personality and Social Psychology 60, 218–228.

Devine, Patricia G., 1989. Stereotypes and prejudice: their automatic and controlled components. Journal of Personality and Social Psychology 56, 5–18.

Doyen, Stéphane, Klein, Olivier, Pichon, Cora-Lise, Cleeremans, Axel, 2012. Behavioral priming: it's all in the mind, but whose mind? PLoS ONE 7, e29081.

Fiske, Susan T., Cuddy, Amy J.C., Glick, Peter, Xu, Jun, 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Fiske, Susan T., Lin, Monica, Neuberg, Steven L., 1999. The continuum model: ten years later. In: Chaiken, Shelly, Trope, Yaacov (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, pp. 231–254.

Greenwald, Anthony G., McGhee, Debbie E., Schwartz, Jordan L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology 74, 1464–1480.

Guimond, Serge, Chatard, Armand, Martinot, Delphine, Crisp, Richard J., Redersdorff, Sandrine, 2006. Social comparison, self-stereotyping, and gender differences in self-construals. Journal of Personality and Social Psychology 90, 221–242.

Haslam, Nick, Loughnan, Steve, 2014. Dehumanization and infrahumanization. Annual Review of Psychology 65, 399–423.

Jost, John T., Banaji, Mahzarin R., Nosek, Brian A., 2004. A decade of system justification theory: accumulated evidence of conscious and unconscious bolstering of the status quo. Political Psychology 25, 881–919.

Judd, Charles M., Park, Bernadette, 1993. Definition and assessment of accuracy in social stereotypes. Psychological Review 100, 109–128.

Jussim, Lee, 2012. Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-fulfilling Prophecy. Oxford University Press, New York.

Kamans, Elanor, Gordijn, Ernestine H., Oldenhuis, Hilbrand, Otten, Sabine, 2009. What i think you see is what you get: influence of prejudice on assimilation to negative meta-stereotypes among Dutch Moroccan teenagers. European Journal of Social Psychology 39, 842–851.

Kashima, Yoshihisa, Klein, Olivier, Clark, Anna E., 2007. Grounding: sharing information in social interaction. In: Fiedler, Klaus (Ed.), Social Communication. Psychology Press, New York, pp. 27–77.

Katz, Daniel, Braly, Kenneth, 1933. Racial stereotypes of one hundred college students. The Journal of Abnormal and Social Psychology 28, 280–290.

Klein, Olivier, Snyder, Mark, 2003. Stereotypes and behavioral confirmation: from interpersonal to intergroup perspectives. Advances in Experimental Social Psychology 35, 153–234.

Klein, Olivier, Spears, Russell, Reicher, Stephen, 2007. Social identity performance: extending the strategic side of SIDE. Personality and Social Psychology Review 11, 28–45.

Kunda, Ziva, Thagard, Paul, 1996. Forming impressions from stereotypes, traits, and behaviors: a parallel-constraint-satisfaction theory. Psychological Review 103, 284–308.

Macrae, Neil C., Bodenhausen, Galen V., Milne, Alan B., Jetten, Jolanda, 1994. Out of mind but back in sight: stereotypes on the rebound. Journal of Personality and Social Psychology 67, 808–817.

Moskowitz, Gordon B., 2010. On the control over stereotype activation and stereotype inhibition. Social and Personality Psychology Compass 4, 140–158.

Oakes, Penelope J., Haslam, Alexander S., Turner, John C., 1994. Stereotyping and Social Reality. Wiley-Blackwell, Oxford, UK.

Pettigrew, Thomas F., Tropp, Linda R., 2008. How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. European Journal of Social Psychology 38, 922–934.

Poppe, Edwin, 2001. Effects of changes in GNP and perceived group characteristics on national and ethnic stereotypes in Central and Eastern Europe. Journal of Applied Social Psychology 31, 1689–1708.

Schimel, Jeff, Simon, Linda, Greenberg, Jeff, Pyszczynski, Tom, Solomon, Sheldon, Waxmonsky, Jeanette, Arndt, Jamie, 1999. Stereotypes and terror management: evidence that mortality salience enhances stereotypic thinking and preferences. Journal of Personality and Social Psychology 77, 905–926.

Schmader, Toni, Johns, Michael, 2003. Converging evidence that stereotype threat reduces working memory capacity. Journal of Personality and Social Psychology 85, 440–452.

Schmader, Toni, Johns, Michael, Forbes, Chad, 2008. An integrated process model of stereotype threat effects on performance. Psychological Review 115 (2), 336.

Sherman, Jeffrey W., 1996. Development and mental representation of stereotypes. Journal of Personality and Social Psychology 70, 1126–1141.

Smith, Eliot R., Zárate, Michael A., 1992. Exemplar-based model of social judgment. Psychological Review 99, 3–21.

Steele, Claude M., Aronson, Joshua, 1995. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology 69, 797–811.

Tajfel, Henri, 1981. Human Groups and Social Categories. Cambridge University Press, Cambridge.

Tajfel, Henri, Turner, J.C., 1986. The social identity theory of inter-group behavior. In: Worchel, S., Austin, L.W. (Eds.), Psychology of Intergroup Relations. Nelson-Hall, Chicago, pp. 7–24.

Yzerbyt, Vincent Y., Rocher, Steve, Schadron, Georges, 1997. Stereotypes as explanations: a subjective essentialistic view of group perception. In: Spears, Russel Ed, Oakes, Penelope J., Ellemers, Naomi Ed, Haslam, Alexander S. (Eds.), The Social Psychology of Stereotyping and Group Life. Blackwell Publishers, Oxford, pp. 20–50.

Stevens, Stanley Smith (1906–73)

Robert Teghtsoonian, Smith College, Northampton, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

S. Smith Stevens (1906–73) was a leading figure in the field of experimental psychology during the middle part of the twentieth century. He is best known for his work in psychophysics and especially for the development of the psychophysical power law, an empirical description of the relation between the intensity of stimuli such as sound and light and the perceptual magnitude experienced by an observer. He was also a pioneer in the study of hearing and did important work on the theory of measurement. Aside from his many research reports, his most influential publications include the classic *Hearing: Its Psychology and Physiology* (1938), coauthored with Hallowell Davis, the *Handbook of Experimental Psychology* (1951), and *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects* (1975), edited by his wife Geraldine Stevens.

Beginnings

To understand fully the significance of any scientist's contribution, it is important to know the context in which it occurred. It is particularly useful to be acquainted both with the scientists who paved the way and with those who followed. Historians of experimental psychology have traced the beginnings of the field to the German scientist Wilhelm Wundt (1832–1920), who established the first laboratory for psychological research in 1879. Although other scientists had made important contributions to the emerging field, it was Wundt who first identified an entire range of questions in psychology amenable to laboratory investigation, and it was to his laboratory in Leipzig that students from many countries came for training in experimental psychology. Stevens, whose distinguished career culminated in his position as Professor of Psychophysics at Harvard University, stands in a direct line of scholarly descent from Wundt as a member of the fourth generation. He was trained by E.G. Boring (1886–1968) at Harvard, who in turn had studied at Cornell with E.B. Titchener (1867–1927), an Englishman who had received his degree in Leipzig from Wundt in 1892. And whereas Wundt may be considered the first to devote himself to Experimental Psychology, it may well be that Stevens was the last. His work marked and contributed to a transition from one era to another in the evolution of scientific psychology; an understanding of that transition makes an interesting story.

Stevens was born on 4 November 1906 in Ogden, Utah, home to the Mormon Church. Following the customs of that institution, he departed at the age of 17 to spend 3 years as a missionary in Belgium and later in France, despite a complete ignorance of French. It seems likely that this experience contributed to the development of persistence and determination, traits that were to serve him well in his career as a scientist. On his return he entered the University of Utah and, after 2 years, transferred to Stanford, where he graduated in 1931 having taken such a wide variety of courses that the identification of a major was problematic. He was accepted by the Harvard Medical School but demurred on the grounds that the required payment of a $50 fee and the need to spend the summer taking a course in organic chemistry seemed equally unattractive. But he did decide to study at Harvard, and duly enrolled in the School of Education as the easiest (and most economical) path to Harvard's resources. The transforming experiences there were E.G. Boring's course on perception and service as Boring's unpaid research assistant. By the end of his second year, Stevens had earned a PhD in the Department of Philosophy (from which the Psychology Department emerged a year later). A year studying physiology under Hallowell Davis and another as a research fellow in physics led to appointment in 1936 as an instructor in the Psychology Department, where he remained until his death on 18 January 1973. By 1946 he achieved the rank of full professor, and in 1962 the University acceded to his request that he be named Professor of Psychophysics in honor of the field founded a century earlier by G.T. Fechner, whose theories, ironically, Stevens worked hard to refute.

Stevens's Psychophysical Power Law

Scholars familiar with his work quickly could agree in making a list of Stevens's major achievements, but could and do disagree about their relative importance and their lasting value. For some, the single most important accomplishment was his discovery of the psychophysical power law that sometimes does and always should carry his name (Stevens, 1957). He not only revived a view broached by some in the nineteenth century, but also created a body of evidence that firmly established what is sometimes called the psychophysical power law and sometimes Stevens's Law. It concerns the relation between the strength of some form of energy, such as the sound pressure level of a tone, and the magnitude of the corresponding sensory experience, loudness in this example. It is easily discovered that sensation strength is nonlinearly related to stimulus intensity (two candles do not make a room seem twice as bright as one), but it is harder to say what that relation is. For over a century before Stevens approached the problem, the prevailing view, Fechner's Law, asserted that as stimulus strength grew geometrically (by constant ratios), sensation strength increased arithmetically. This logarithmic relation had little supporting evidence but remained in many texts for lack of a good alternative. Stevens came to a different conclusion on the basis of data obtained by asking observers

who were presented with stimuli varying in intensity to make numerical judgments of their subjective experience. In the case of loudness he found that judged loudness grew as a power function of sound pressure (with an exponent of about 0.6), but not as the logarithmic relation predicted by Fechner (Stevens, 1961). A similar relation was found to hold between judged brightness and luminance, and eventually, such a relation was found for dozens of other continua. In each case, as the amount or intensity of a stimulus grew by ratio steps, so too did the observer's estimate of his subjective experience of it. Stevens's Law seemed applicable to all intensive perceptual continua.

Equally interesting was the discovery that for each perceptual continuum, there was a distinctive value of the exponent relating numerical judgment to stimulus intensity. In hearing, for example, a huge range of sound pressures was compressed into a much smaller range of loudness, whereas for electric shock applied to the fingertips, a quite small range of currents between detectability and pain was expanded into a larger range of perceived intensities.

Although Stevens originally argued that the method of magnitude estimation provided a 'direct measure' of sensation and that the power law described how energy in the environment was transformed through the relevant sensory apparatus into sensation, critics soon showed this assumption to be untenable. What remained was an empirical principle relating judgment to stimulus intensity.

Stevens went on to show that numerical magnitude estimation was a special case of a more general paradigm, cross-modal matching. If the observer is asked to match loudness, for example, by manipulating the luminance of a light source until its brightness matches the level of the target loudness, the power relation is again the result. And, neatly enough, the exponent of that function is exactly predictable by the exponents identified through magnitude estimation of brightness and loudness alone (Stevens, 1969). It is this ability of observers to match perceived magnitudes across many different perceptual attributes, and the discovery that in all cases the result is described by a power function that are the two pillars on which Stevens's Law rests. Despite the many controversies over its meaning, this simple empirical principle stands securely on a mountain of evidence and must be accommodated by any theory of psychophysics. And indeed, it has been incorporated into a more general model that, starting with the premise that a power law describes the relation between sensory magnitudes and stimulus intensity, adds the idea that the various sensory systems vary in dynamic range and in resolving power, but all map into a single scale of perceived magnitude with its own characteristic dynamic range and its own resolving power (Teghtsoonian, 2012). It remains to be seen whether such a theoretical construct has some observable neural foundation.

Scales of Measurement

Others attach greater importance to Stevens's ideas about scales of measurement. In a seminal paper (1946) he created a taxonomy of measurement, which identified four classes of scales – nominal, ordinal, interval, and ratio – distinguishing among them on the basis of the operations that leave the form of the scale invariant. Nominal scales assigned numbers that serve only to distinguish one event or class of events from another (the classic example is the set of numbers used to identify each member of a team); any one-to-one substitution is permissible because player identity is preserved. Ordinal scales of measurement permit a rank ordering of measured events (relative hardness of materials, for example) and will remain invariant under monotonic increasing transformations. Interval scales entail a constant unit of measurement and so permit the calculation of differences between any two values (as for the common scales of temperature) and remain invariant under linear transformations. Ratio scales (the bread and butter of physics) also feature constant units of measurement, but in addition allow the expression of two values as a ratio, since a true zero exists; weight is a good example. Such a scale is invariant under a simple multiplicative transformation.

It is harder to assess the impact of this conceptualization of measurement than that of the work on cross-modal matching judgments. The latter was instrumental in creating new bodies of knowledge, especially in the field of sensory psychology, and continues to play an influential role in psychophysical theory and research. But Stevens's four-step hierarchy of measurement scales encountered an assortment of reactions that leaves its present status in some doubt.

Contemporary theorists find the nominal scale of little interest because it imposes no ordering on the measured entities. And several authorities have noted the omission of a log interval scale from what Stevens had proposed as an exhaustive listing. The relevance of this classification to the selection of statistical procedures was hotly debated (e.g., see Binder, 1984). Stevens had argued that the type of measurement placed constraints on which statistics can be used, but statisticians were quick to cite seeming exceptions, and the usefulness of Stevens's measurement taxonomy as a guide to the selection of statistical techniques became uncertain.

But, for psychologists, his ideas about measurement were, and in some ways continue to be, of great importance. At a time when Stevens was trying to provide support for his newly-created scale of loudness, he felt it important to show that it met the same criteria of legitimacy as did the familiar scales of the more developed sciences. It was a time when the status of psychology as a science was a topic of lively debate, and Stevens was much involved in efforts to identify the critical features of science and the ways in which psychology shared them. During the 1940s and 1950s, questions about the nature of science and psychology's place in it were discussed at length, and Stevens's analysis of measurement in general and the ways in which psychologists could use scales of measurement was very influential. The idea that a measurement scale – or, for that matter, any lawful relation – could be expressed in terms of the operations that left it invariant played an important role in the thinking of his students, and no doubt for scientists in other disciplines as well. The work that Stevens began has now evolved into a mathematical specialty in its own right, yielding results that confirm some of what Stevens intuited, and disconfirming other aspects of his work. (It must be remembered that the ideas he was advancing lent themselves to

mathematical formalization, though he himself did not provide it.) That his ideas about measurement have continued to be developed and refined by other scientists today may be the ultimate accolade for his pioneering work in this field (e.g., see Luce and Narens, 1999).

Hearing

If Stevens had never established the psychophysical power law and had left the study of measurement untouched, he would still be known and remembered as a pioneer in the field of psychoacoustics, the scientific study of the perception of sound (e.g., see Scharf, 1999). His collaboration with Hallowell Davis in the late 1930s resulted in *Hearing*, a book that for many years was the fundamental source on the subject and has been ranked by some scholars as second only to the work of Helmholtz in its originality and breadth. In addition to his measurements of loudness, he did important work on auditory localization, and on the relation of pitch to intensity. At the behest of the US Army Air Corps he established the Psycho-Acoustic Laboratory at Harvard in 1940 to study the effect of noise on pilot psychomotor efficiency. As its director he was able to attract a group of students and colleagues of extraordinary ability; among them were W.R. Garner, K.D. Kryter, J.C.R. Licklider, and G.A. Miller, all of whom made important contributions to psychoacoustics and beyond. It was an interesting aspect of his personality that, despite a rather gruff manner, he functioned as an intellectual magnet, attracting to his laboratory some of the most talented researchers in the world. Not the least of these was G. von Békésy, who spent almost 20 years in Stevens's basement lab and won a Nobel Prize during that period for his early work (in Budapest and Stockholm) on the ear.

Throughout his career there were several periods of intense activity as a laboratory scientist. But in the period before World War II, he was also active in the discussion of philosophical questions about the nature of psychology. He was impressed by the work of the physicist P.W. Bridgman on the central role of operationism in science, and argued for its use in defining the concepts of psychology. And for a period he was associated closely with a group of logical and methodological positivists whose ideas about the nature of science seemed especially hospitable to a behaviorist psychology (Stevens, 1939). Perhaps, one manifestation of this outlook was a reluctance to extend his thinking very far beyond his data. In a period when many experimental psychologists were developing ever more complex theoretical systems, Stevens's own forays into the realm of theory were few and limited in scope. He seemed happiest examining his data (as well as those of others) to see what useful information could be extracted. Much of his time in the laboratory was spent plotting and replotting data in the pursuit of clarity and simplicity, and his research reports were models of both qualities.

Impact on Experimental Psychology

Two additional characteristics of his work deserve mention. While the majority of his contemporaries were engaged in a search for relevant variables (e.g., does 'meaningfulness' affect ease of remembering?), Stevens's goal was the discovery of functional relations, mathematical descriptions of the association of one factor with another, as epitomized by his psychophysical power law. A corollary of this belief was an attitude toward sampling theory and statistical testing that ranged from indifference to disdain. The object for him was not to see whether 80 dB tones received an average loudness judgment greater than did 60 dB tones (at the 5% level of significance), the equivalent of what most of his contemporaries were doing. What was important to Stevens was whether a lawful relation could be seen when data were plotted in a thoughtful way. Large numbers of experimental subjects were not needed to reveal the kind of robust relations he sought; a few colleagues or students recruited in the hallways were enough for most of his purposes.

A landmark publication for research psychologists in the 1950s was the *Handbook of Experimental Psychology* edited by Stevens (1951). It was, as he described it in the preface, "a technical survey that would systematize, digest, and appraise the mid-century state of experimental psychology." It was not the first of its kind, but it was the most recent, and thus defined for teachers and students alike the set of shared interests that constituted experimental psychology. It was a time still when one might be expert in one aspect of the field and fairly knowledgeable about many others. Since then the amount of knowledge in psychology has grown at an explosive rate, and now the achievement of competence in one area increasingly restricts knowledge of any other. At the beginning of the twenty-first century, the tent of 'experimental psychology' has few remaining occupants; there are instead cognitive scientists, brain researchers, vision scientists, and a host of others in specialized niches. Just as Stevens's mentor E.G. Boring may have been among the last to teach an all-inclusive account of psychology in a single course, so Stevens may have been among the last to view the phrase 'experimental psychology' as a useful rubric.

Conclusion

Stevens can be seen as a transitional figure in the evolution of psychological science, yet much of what he did continues to influence the way it is practiced today. His lasting contributions to the fields of psychoacoustics and psychophysics have been discussed here. But, in addition, he created a roadmap for much of modern sensory neurophysiology and for those who theorize in that area. Many of the phenomena for which today's brain researchers seek explanations were established by Stevens or by others using the methods he developed. Were he alive today, it seems likely that he would embrace the new technology and would be developing ingenious techniques for exploring the neural aspects of sensory events. Although an account of his influence remains an unfinished story, at the least he will be remembered as one who brought new and powerful ideas to the study of psychophysics, leaving it a field reinvigorated and evolving as a subject of great intrinsic interest and many useful applications (e.g., see Baird, 1997; Norwich, 1993; Stevens, 1966).

See also: Experimentation in Psychology, History of; Galton, Sir Francis (1822–1911); Helmholtz, Hermann Ludwig Ferdinand von (1821–94); Lashley, Karl Spencer (1890–1958); Psychology, History of (Twentieth Century); Sensation and Perception: Direct Scaling; Speech Processing, The Cortical Organization of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Baird, J.C., 1997. Sensation and Judgment: Complementarity Theory of Psychophysics. Erlbaum, Mahwah, NJ.

Binder, A., 1984. Restrictions on statistics imposed by method of measurement: some reality, much mythology. Journal of Criminal Justice 12, 467–481.

Luce, R.D., Narens, L., 1999. S.S. Stevens, measurement, and psychophysics. In: Killeen, P.R., Uttal, W.R. (Eds.), Fechner Day 99: The End of Twentieth Century Psychophysics. The International Society for Psychophysics, Tempe, AZ.

Miller, G.A., 1974. Stanley Smith Stevens: 1906–1973. American Journal of Psychology 87, 279–288.

Norwich, K.H., 1993. Information, Sensation, and Perception. Academic Press, San Diego, CA.

Scharf, B., 1999. S.S. Stevens's contributions to an understanding of hearing. In: Killeen, P.R., Uttal, W.R. (Eds.), Fechner Day 99: The End of Twentieth Century Psychophysics. The International Society for Psychophysics, Tempe, AZ.

Stevens, S.S., 1939. Psychology and the science of science. Psychological Bulletin 36, 221–263.

Stevens, S.S., 1946. On the theory of scales of measurement. Science 103, 677–680.

Stevens, S.S. (Ed.), 1951. Handbook of Experimental Psychology. Wiley, New York.

Stevens, S.S., 1957. On the psychophysical law. Psychological Review 64, 153–181.

Stevens, S.S., 1961. To honor Fechner and repeal his law. Science 133, 80–86.

Stevens, S.S., 1966. A metric for the social consensus. Science 151, 530–541.

Stevens, S.S., 1969. On predicting exponents for cross-modality matches. Perception & Psychophysics 6, 251–256.

Stevens, S.S., 1975. Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects. Wiley, New York.

Stevens, S.S., Davis, H., 1938. Hearing: Its Psychology and Physiology. Wiley, New York.

Teghtsoonian, R., 2012. The standard model for perceived magnitude: a framework for (almost) everything known about it. American Journal of Psychology 125, 165–174.

Stigler, George Joseph (1911–91)

Milton Friedman, Hoover Institution, Stanford University, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15108–15112, © 2001, Elsevier Ltd.

Abstract

George Stigler, a key member of the 'Chicago School' of economic thought, made major contributions to the history of thought, price theory, economics of information, organization of industry, and the theory of economic regulation. He taught at Iowa State University, the University of Minnesota, Brown University, Columbia University, and the University of Chicago, where he also served as editor of the *Journal of Political Economy*. He was President of the American Economic Association in 1964, and of the History of Economics Society in 1977. He received the Alfred Nobel Memorial Prize in Economic Science in 1982.

Personal History

George Stigler, one of the great economists of all time, with a gift for writing matched among modern economists only by John Maynard Keynes, was born 17 January 1911 in Renton, Washington, a suburb of Seattle, the only child of Joseph and Elisabeth Stigler. His parents had separately migrated to the United States at the end of the nineteenth century, his father from Bavaria, his mother from what was then Austria-Hungary. Stigler went to public schools in Seattle, and then to the University of Washington, receiving a BA in 1931. Although he says that he had 'no thought of an academic career' when he graduated from college, he applied for and was awarded a fellowship at Northwestern University for graduate study in the business school, receiving an MBA in 1932 (Stigler, 1988, p.15). More important, at Northwestern he developed an interest in economics as an object of study in its own right and decided on an academic career. After one further year of graduate study at the University of Washington, he moved to the University of Chicago where he found the intense intellectual atmosphere he sought. Chicago became his intellectual home for the rest of his life, as a student from 1933 to 1936, a faculty member from 1958 to his death in 1991, and a leading member of, and contributor to, the 'Chicago School' throughout. He received his PhD in 1938, with a thesis, now a classic, on the history of neoclassical theories of production and distribution (Stigler, 1941).

At Chicago, Stigler was particularly influenced by Frank H. Knight, under whom he wrote his dissertation, Jacob Viner, who taught economic theory and international economics, John U. Nef, economic historian, and their younger colleague, Henry Simons, who became a close personal friend and whose *A Positive Program for Laissez Faire* greatly influenced many of Stigler's contemporaries.

'At least as important to me,' Stigler writes (1988, pp. 23–5), 'as the faculty were the remarkable students I met at Chicago,' and goes on to list W. Allen Wallis, Milton Friedman, Kenneth Boulding, and Robert Shone from Britain, Sune Carlson from Sweden, Paul Samuelson, and Albert G. Hart – all of whom subsequently had distinguished careers.

In 1936, Stigler was appointed an assistant professor at Iowa State College (now University), and shortly thereafter was married to Margaret Mack. The Stiglers had three sons: Stephen, a professor of statistics at the University of Chicago, David, a lawyer in private practice, and Joseph, a businessman. Mrs. Stigler died in 1970, tragically without any advance warning. George never remarried.

In 1938, Stigler accepted an appointment at the University of Minnesota, going on leave in 1942 to work first at the National Bureau of Economic Research and then at the Statistical Research Group of Columbia University, where he joined a group directed by W. Allen Wallis that was engaged in war research on behalf of the armed services. When the war ended in 1945, George returned to the University of Minnesota, but remained only one year, leaving in 1946 to accept a professorship at Brown University, where he also remained only one year before moving to Columbia University in 1947 and then, in 1958, to the University of Chicago where he remained the rest of his life. At Chicago, he became an editor of the *Journal of Political Economy*, established the Industrial Organization Workshop, which achieved recognition as the key testing ground for contributions to the field of industrial organization, and in 1977 founded the Center for the Study of Economy and the State, serving as its director until his death in Chicago on December 1, 1991.

Stigler served as President of the American Economic Association in 1964, and of the History of Economics Society in 1977. He was elected to the National Academy of Sciences in 1975. He received the Alfred Nobel Memorial Prize in Economic Science in 1982 and the National Medal of Science in 1987.

Stigler's governmental activities included: serving as a member of the Attorney General's National Committee to Study the Antitrust Laws, 1954–5; Chairman, Federal Price Statistics Review Committee, 1960–1; member, Blue Ribbon Panel of the Department of Defense, 1969–70; Vice Chairman, Securities Investor Protection Corporation, 1970–3; Co-Chairman, Blue Ribbon Telecommunications Task Force, Illinois Commerce Commission. 1990–1.

Scientific Contributions

Stigler made important contributions to a wide range of economic areas: (a) history of thought, (b) price theory, (c) economics of information, (d) organization of industry, and (e) theory of economic regulation. In addition, (f) he wrote many influential essays and reviews, and was a long-time editor.

Intellectual History

Intellectual history was Stigler's first field of specialization and remained a continuing interest throughout his life. His doctoral dissertation, published as *Production and Distribution Theories*, was followed by a steady flow of perceptive, thoughtful, and beautifully written articles and books interpreting the contributions of his predecessors. Almost single-handedly, he kept the field alive, despite widespread lack of interest within the profession. More recently, he succeeded in bringing about a revival of interest in the field, which is currently flourishing. In the words of one of his students, 'The versatility of his work in [the history of economic thought] has been matched by its depth and insight – and unmatched by any other historian of economic thought' (Sowell 1987, p. 498; the rest of this paragraph draws from Sowell's excellent summary of 'Stigler as an Historian of Economic Thought'). In addition to analyzing the theories of particular economists, he examined particular concepts, such as perfect competition and utility, traced trends in the economics profession, considered the influence of economic theories on events and vice-versa, and the meaning of originality. Many of his important articles in this area are collected in a 1965 volume, *Essays in the History of Economics*.

The history of thought provided a rich seedbed for Stigler's wide-ranging scientific work. Stigler regarded economic theory, in the words of Alfred Marshall, as 'an engine for the discovery of concrete truth,' not as a subject of interest in its own right, a branch of mathematics. Few economists have so consistently and successfully combined economic theory with empirical analysis, or ranged so widely.

Price Theory

Stigler's first important publication after his doctoral thesis was a textbook, *The Theory of Competitive Price*, published in 1942, followed by four revised versions under the title *The Theory of Price* (1946, 1952, 1966, 1987). His textbook is unique among intermediate textbooks in price theory in its systematic linking of highly abstract theory to observable phenomena. It is unique also in its concise yet rigorous exposition.

One of his earliest papers (1939) broadened the concept of a short-run cost curve to accommodate the observation that firms faced with fluctuating demand could build flexibility into their fixed equipment at a cost of not achieving the lowest possible cost for a particular output.

Another early paper (1945) presented a formal analysis of determining the minimum cost of an adequate diet, supplemented by an empirical calculation of the minimum cost – one of the earliest applications of linear programming. 'The procedure,' he explained, 'is experimental because there does not appear to be any direct method of finding the minimum of a linear function subject to linear conditions' (p. 310). Two years later George Dantzig provided such a direct method, the simplex method, now widely used in many economic and industrial applications (1987).

Like these two examples, Stigler's many later contributions to price theory were a by-product of seeking to understand puzzling real world phenomena, and are best discussed in that context.

Economics of Information

'The Economics of Information' is the title of a seminal article, published in 1961. In his intellectual autobiography, Stigler termed it, 'My most important contribution to economic theory' (1988, pp. 79–80). The article begins, 'one should hardly have to tell academicians that information is a valuable resource: knowledge is power. And yet it occupies a slum dwelling in the town of economics. Mostly it is ignored' (Stigler, 1961, p. 213).

This brief 12-page article is a splendid illustration of several of Stigler's unique virtues: creativity, which he defined as consisting 'of looking at familiar things or ideas in a new way'; the capacity to extract new insights about those seemingly familiar things; and the ability to state his main points in a provocative and eminently readable way.

Stigler illustrates the importance of subjecting information to economic analysis with two familiar observations—that there is generally no single 'market price' but a variety of prices that 'various sellers (or buyers) quote at any time' and that 'advertising ... is an enormously powerful instrument for the elimination of ignorance ... and is normally "paid" for by the seller.' In each case, he presents a sophisticated analysis that explains what determines the observed phenomena, an analysis that has empirical implications capable of being contradicted by observation. He concludes that 'The identification of sellers and the discovery of their prices are only one sample of the vast role of the search for information in economic life' (quotes from Stigler, 1988, pp. 213, 220, 221, 224).

As occurred throughout his career in every field Stigler touched, this article was less important for the particular familiar observations that it analyzed than for giving birth to an essentially new area of study for economists. There followed an outpouring of further studies by him but even more by his students and others that have radically altered the accepted body of economic theory.

This aspect of his influence was stressed in the citation for his Nobel Prize, which noted that Stigler's 'studies of the forces which give rise to regulatory legislation have opened up a completely new area of economic research ... Stigler is also recognized as the founder of "economics of information" and "economics of regulation," and one of the pioneers of research in the intersection of economics and law.'

The Organization of Industry

The Organization of Industry is the title of a book published in 1968. Its 'main content,' Stigler says in the preface, 'is a reprinting of 17 articles I have written over the past two decades [including the "Economics of Information"] in the area of industrial organization. Although the main topics in industrial organization are touched upon, the touch is often light. The ratio of hypotheses to reasonably persuasive confirmation is distressingly high in all of economic literature, and it must be my chief and meager defense that I am not the worst sinner in the congregation.' Far from being the worst sinner, Stigler's main contribution to the field, both in this book and later writing, was the use of empirical evidence to test hypotheses designed to explain features of industrial organization. Article after article combines subtle theoretical analysis

with substantial nuggets of empirical evidence, presented so casually as to conceal the care with which the data were compiled and the effort that was expended to determine what data were both relevant and accessible. These articles record the shift in Stigler's views on antitrust – from initial support of an activist antitrust policy to skepticism about even a minimalist policy.

One of the most influential of the articles in the book was on 'The Economies of Scale' in which he introduced 'the survivor principle' as a way to use market data to determine the efficient size of an enterprise in a particular industry. The approach has since become widely accepted.

In 1962, he published an article, in collaboration with Claire Friedland, his long-time associate, entitled 'What Can Regulators Regulate? The Case of Electricity,' which concluded that regulation of electric utilities had produced no significant effect on rates charged. Schmalensee (1987, p. 499) notes that 'This essay stimulated a great deal of important empirical work on public regulation in the USA. It also posed a basic problem: if regulation does not generally achieve its stated objectives, why have so many agencies been established and kept in existence?'

Economic Regulation

'The Theory of Economic Regulation' (1971b) attempts to answer that question. It is one of a series of contributions by Stigler to what has since come to be called 'public choice' economics: the shift from viewing the political market as not susceptible to economic analysis, as one in which disinterested politicians and bureaucrats pursue the 'public interest,' to viewing it as one in which the participants are seeking, as in the economic market, to pursue their own interests, and hence subject to analysis with the usual tools of economics. The seminal work that deserves much of the credit for launching 'public choice,' *The Calculus of Consent*, by James Buchanan and Gordon Tullock (1962), appeared in the same year as the Stigler–Friedland article.

The 'central thesis of the article,' Stigler wrote, 'is that, as a rule, regulation is acquired by the industry and is designed and operated primarily for its benefit.' He notes that two 'alternative views of the regulation of industry are widely held. The first is that regulation is instituted primarily for the protection and benefit of the public at large or some large subdivision of the public. The second view is essentially that the political process defies rational explanation.' He then gives example after example to support his own thesis, which is, by now, the orthodox view in the profession, concluding: 'The idealistic view of public regulation is deeply imbedded in professional economic thought. The fundamental vice of such a [view] is that it misdirects attention' – to preaching to the regulators rather than changing their incentives (quotes from Stigler, 1971b, p. 3).

'Smith's Travels on the Ship of State' (1971a), published in the same year as 'The Theory of Economic Regulation,' raises the same question on a broader scale. Smith gives self-interest pride of place in analyzing the economic market, but does not give it the same role in analyzing the political market. Smith's failure to do so constitutes Stigler's main – indeed, nearly only – criticism of the *Wealth of Nations*, that 'stupendous palace erected upon the granite of self interest' (Rosenberg, 1993, p. 835). The same theme pervades many of Stigler's later publications.

Essays, Editorship, and Reviews

In addition to his prolific academic publications, Stigler published a great many essays for a broader audience, many given as talks to a wide variety of audiences. The most important are collected in three volumes, *The Intellectuals and the Marketplace* (1963), *The Citizen and the State* (1975), and *The Economist as Preacher* (1982). 'There he (the intelligent layman) will find a potpourri of wit and seriousness blended with a high writing style' (Demsetz, 1982, p. 656). 'For 19 years Stigler was a very successful editor of the *Journal of Political Economy*. Under his leadership this journal solidified its high reputation among economists' (Becker, 1993, p. 765).

Stigler was also a prolific reviewer of books. His complete bibliography lists 73 reviews in 24 publications ranging from strictly professional, such as the *Journal of Political Economy* (22) and the *American Economic Review* (10) to the popular, such as the *Wall Street Journal* (5), and the *New York Times* (3), and dating from 1939 to 1989.

Stigler's last book, his intellectual autobiography, published in 1988, *Memoirs of an Unregulated Economist*, is a delight to read. As I described it at the time: 'Stigler's memoirs are a gem: in style, in wit, and above all, in substance, they reflect accurately his own engaging personality and his extraordinarily diverse contributions to our science.'

Concluding Comments

An evaluation of Stigler's scientific contribution would be incomplete without recognizing his role as a teacher and colleague. As teacher, he inspired his students by his own high standards, and instilled a respect for economics as a serious subject concerned with real problems. As one of his students, Thomas Sowell, wrote (1993, p. 788): 'What Stigler really taught, whether the course was industrial organization or the history of economic thought, was intellectual integrity, analytical rigor, respect for evidence – and skepticism toward the fashions and enthusiasms that come and go.'

Stigler supervised many doctoral dissertations at both Columbia and the University of Chicago. According to Claire Friedland, Stigler's associate for many years, he served on more than 40 thesis committees at Chicago, and perhaps 40 more at Columbia, and chaired a considerable fraction of those committees. His students come close to dominating the field of industrial organization.

Stigler was an extremely valuable colleague. He provided much of the energy and drive to the interaction among members of the Chicago economics department, business school, and law school that came to be known as 'The Chicago School.' His Workshop in Industrial Organization was an outgrowth of a law school seminar started by Aaron Director that Stigler cooperated in running when he came to Chicago. His relations were especially close with Aaron Director, Gary Becker, Richard Posner, Harold Demsetz, and myself, enhancing significantly the scientific productivity of all of us.

See also: Economics, History of; Employment and Labor, Regulation of; Information, Economics of; Law and Economics; Market Structure and Performance; Political Economy, History of; Regulation, Economic Theory of; Utilitarian Social Thought, History of.

Bibliography

Becker, G.S., 1993. George Joseph Stigler, January 17, 1911–December 1, 1991. Journal of Political Economy 101, 761–767.

Buchanan, J.M., Tullock, G., 1962. The Calculus of Consent. University of Michigan Press, Ann Arbor, MI.

Dantzig, G.B., 1987. Linear programming. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 3. Stockton Press, New York, pp. 203–205.

Demsetz, H., 1982. The 1982 Nobel Prize in economics. Science 218, 655–657.

Rosenberg, N., 1993. George Stigler: Adam Smith's best friend. Journal of Political Economy 101, 833–848.

Schmalensee, R., 1987. Stigler's contributions to microeconomics and industrial organization. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 4. Stockton Press, New York, pp. 499–500.

Sowell, T., 1987. Stigler as an historian of economic thought. In: Eatwell, J., Milgate, M., Newman, P. (Eds.), The New Palgrave: A Dictionary of Economics, vol. 4. Stockton Press, New York, pp. 498–499.

Sowell, T., 1993. A student's eye view of George Stigler. Journal of Political Economy 101, 784–792.

Stigler, G.J., 1939. Production and distribution in the short run. Journal of Political Economy 47, 305–327.

Stigler, G.J., 1941. Production and Distribution Theories, 1870 to 1895. Macmillan, New York.

Stigler, G.J., 1942. The Theory of Competitive Price. Macmillan, New York.

Stigler, G.J., 1945. The cost of subsistence. Journal of Farming Economy 27, 303–314.

Stigler, G.J., 1946. The Theory of Price. rev. edns. 1952, 1966, 1987. Macmillan, New York.

Stigler, G.J., 1961. The economics of information. Journal of Political Economy 69, 213–225.

Stigler, G.J., 1963. The Intellectual and the Market Place, and Other Essays. Free Press of Glencoe, New York.

Stigler, G.J., 1965. Essays in the History of Economics. University of Chicago Press, Chicago, IL.

Stigler, G.J., 1968. The Organization of Industry. Irwin, Homewood, IL.

Stigler, G.J., 1971a. Smith's travels on the ship of state. History of Political Economy 3, 265–277.

Stigler, G.J., 1971b. The theory of economic regulation. Bell Journal of Economics and Management Strategy. Sci 2, 3–21.

Stigler, G.J., 1975. The Citizen and the State: Essays on Regulation. University of Chicago Press, Chicago, IL.

Stigler, G.J., 1982. The Economist as Preacher, and Other Essays. University of Chicago Press, Chicago, IL.

Stigler, G.J., 1988. Memoirs of an Unregulated Economist. Basic Books, New York.

Stigler, G.J., Friedland, C., 1962. What can regulators regulate? The case of electricity. Journal of Law and Economics 5, 1–16.

Stigma, Social Psychology of

Lauri Hyers and Eleanor Brown, West Chester University, West Chester, PA, USA
Jacqueline Sullivan, TW Ponessa & Associates, Exton, PA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J.K. Swim, L.L. Hyers, volume 22, pp. 15112–15115, © 2001, Elsevier Ltd.

Abstract

A stigma is an attribute of an individual or a group that compromises full acceptance from others. Those who possess a stigma are stigmatized, referring to the process of derogation and devaluing that may lead to immediate and long-term social, psychological, economic, and health-related injustices that divide human society. Over the last century, research on people with stigma has evolved from controversial origins into a model social science research program, with contributors from diverse social backgrounds and academic disciplines, working on a range of issues related to stigma.

A stigma is an attribute of an individual or a group that compromises full acceptance from others. Those who possess a stigma are stigmatized, referring to the process of derogation and devaluing that may lead to immediate and long-term social, psychological, economic, and health-related injustices that divide human society. Stigmatizing attributes are often esoteric, context dependent, and changing across human eras and cultures, so that the very trait that is highly stigmatizing in one society or era may be deemed irrelevant or even valued in another. The term stigma itself finds unsettling origins in the ancient practice of marking the skin of those who are socially rejected, although in present usage a stigma may be a figurative rather than a literal marking. While stigmatization remains one of the regrettably enduring features of human society, it is matched by an equally long history of stigmatized individuals, groups, and their allies mobilizing in interpersonal and collective resistance against externally imposed disadvantages.

History of Stigma Research

Over the last century, research on people with stigma has evolved from controversial origins into a model social science research program, with contributors from diverse academic disciplines and social backgrounds, working on a range of issues.

Early 'Race Theory'

In the early part of the 1900s, scientists did not have an integrated theory of the psychology of stigma, per se. Instead, an assortment of individual small-scale disaggregated studies were published (in many academic disciplines), usually 'exposing' an unseemly or tragic nature to the lives of a specific marginalized social group (e.g., people with disabilities, African-Americans, or people living in urban poverty), with the aim of documenting deviation, promoting social concern, and discussing intervention. At this time of dramatic intentional social reorganization, science was the newly recognized authority on how to manage and direct social interventions. These new experts on the human social order employed ethically and scientifically dubious research methods. Portrayed as the downfall of contemporary social progress, stigmatized groups were cataloged and classified to include people referred to variously as epileptics, the feeble-minded, the insane, the morally deficient, criminals, drunkards, the poor, the blind, the deaf-mute, the migraneous, neurotics, hysterics, the paralytic, the sexually depraved, unmarried mothers, the syphilitic, the tubercular, wanderers, and runaways. Some of the earliest writers in the scientific community weighing in on stigma would be the Social Darwinists who minimally tolerated the existence of lower social castes of 'defectives' and 'degenerates' as inevitable by-products of human evolution. Fueled with eugenic derision, some published bogus claims about the innate moral inferiority of nonwhite Western peoples around the globe while others turned their attentions locally, deriding low-income 'degenerate families' in their own communities (Guthrie, 2004). Condescension and victim blaming began to grow in the medical community. For example, Baker and Smith (1939) concluded in the *Journal of the American Medical Association* that facial 'disfigurement' was often inextricably linked with defective personality. With a small sample of a mere 12 plastic surgery patients "taken for detailed study," the authors lamented that many who "experienced very creditable surgical results, have remained unsatisfied and have shown that they are sick and maladjusted persons" (p. 301). Similar derision can be found in the field of education, as in a study of 'mental defectives' written by Leta Hollingworth (1920) warning teachers that 'subnormal children' were social menaces:

> At the extremely low end of our curve we have the undesirable and unfortunate deviates, the dull, the subnormal. They constitute a great burden, economically and socially. …By intermarriage with the more intelligent, subnormals tend to degrade all the social body. By marrying or cohabitation among themselves they produce a numerous offspring and increase the financial burden on public and private charity. Psychologists working in penal institutions and courts have shown conclusively that crime and delinquency of all kinds are closed related to mental subnormality. Alcoholism is closely bound up with mental deficiency… Disease easily originates…and is spread broadcast by them. The problem of the unmarried mother is to be understood more clearly in light of mental inferiority. (p. 18)

Sharing the sentiments of her contemporaries, Hollingworth also explains that Black Americans have lower intelligence quotients because any ancestors "caught and led into slavery

were on the average more stupid" and that the relationship between "feeble-mindedness" and poverty is such "that the former is the cause for the latter" (p. 15). Many early scientists endorsed these 'race theories,' proposing that stigmatized groups are separate and inherently defective races of humans. Influential race theorists such as Herbert Spencer, Edward Thorndike, and Francis Galton advocated for social policies to manage inferior races: lower educational standards (for inferior minds), moving people back to farms (to prevent urban slums), no minimum wage and no unions (so that work was only merit based), and no social charity of any kind (to inhibit poor parents from having many children). Initial scientific efforts to address marginalized stigmatized groups encouraged rigid segregation, institutionalization, sterilization, imprisonment, hospitalization, restrictive marriage, deportation, and extermination.

The Social Gospel

Subsequently, a new agenda was championed by minister and activist Washington Gladden, dubbed the 'social gospel,' promoting Protestant Christian sympathies for the 'downtrodden' and offered an alternative to the race theories of the time (Hilgard, 1986). These advocates, including such notable social scientists as Goodwin Watson, Kurt Lewin, and Gynnar Myrdal, favored sociological explanations over notions of willful degeneracy or innate deviance. Thus, the scientific value of essentialist work on stigmatized groups was increasingly contested in the years approaching World War II (WWII), including criticism by the newly formed activist psychological organization Society for the Psychological Study of Social Issues in the late 1930s. Growing in influence were the ideas of anthropologist Franz Boas who introduced 'culture' as an antidote to eugenicists' naturalistic arguments of essential inferiority. Merton (1938) argued that being poor does not induce crime, but instead crime is induced by being poor in a culture that values material wealth while preventing everyone from having equal access to it. The most impactful research during this era involved in the famous 'doll studies' conducted to prove that cultural and psychological damage had been done to the self-concept of African-Americans (Clark and Clark, 1939). Although based on what later scholars found to be exaggerated, misrepresented, and inconsistent findings (Cross, 1991; Jackson, 2001), concerns about African-Americans' shame and self-hatred led the United States to end segregation in the 1954 *Brown v. Board of Education* Supreme Court decision. Despite their virtuous sympathies, these social scientists still favored an elitist scientific 'objectivity' that spoke for stigmatized groups, "fixing" social problems on behalf of stigmatized communities, instead of empowering stigmatized communities to define and address their social problems from within. The social gospel was noticeably patronizing and denigrating in tone by contemporary standards, leaving a legacy of etiological attributions to 'inferior culture' that later social scientists would need to redress. In time, stigmatized group members would redirect the discourse on stigma, rejecting both the race theorist's charge of innate inferiority and the social gospel's charge of deficient culture.

Once culture supplanted inheritance as an explanation for degeneracy, removing essentialist justifications for discrimination against stigmatized groups, the general public needed an explanation/justification for why the selfsame public was so unjust and segregationist (Jackson, 2001), and hence came in the study of prejudice. Before the concept was 'operationally defined,' prejudice was not typically viewed as a social problem or even a scientific construct for that matter; instead, negative intergroup attitudes were generally seen as natural and inevitable responses to group differences. This monumental change in the research agenda on stigmatized groups provided researchers and the general public a logical explanation for why there was so much antipathy, and the social scientific attention turned away from the 'miserable' lives of stigmatized individuals and onto the perpetrators of prejudice who made the stigmatized so 'miserable.'

Allport's Theory of the Nature of Prejudice

As the shock of WWII was settling in and societal and academic focus turned toward explaining and obliterating anti-Semitism, racism, and segregation, Gordon Allport's (1954) *The Nature of Prejudice* offered a unifying social psychological explanation of these processes. Although focused more narrowly on racism and ethnocentrism at the expense of attention to many other forms of prejudice, Allport made four significant contributions to the understanding of stigmatization with his comprehensively unmatched theory. First, he provided an integrated sociological and psychological framework for understanding stigmatization. Second, he did not restrict his analysis to only one stigmatized group, as in past research, but instead synthesized research on general stigmatization processes, regardless of which social groups were the target. Third, he integrated research on both dominant groups and stigmatized groups – focusing on the perpetrations of the former and the responses of the latter – a rarity in a body of literature that to this day tends to focus exclusively on only one or other. Gladden's 'social gospel' and Boaz's concepts of 'culture' were quite evident in Allport's chapter entitled "Discussion of Traits Due to Victimization" about the character of people who are subject to stigmatization: "One's reputation, whether false or true, cannot be hammered, hammered, hammered, into one's head without doing something to one's character" (p. 142). Allport's fourth contribution was his multidimensional model of stigmatizing factors that recognized stigma to be a layered phenomenon. He offered an elaborate sociological model that presages current research on nearly every level of analysis in the social structure of prejudice and stigmatization: history, sociocultural influences, situational factors, personality differences, phenomenological (social cognitions), and the traits of the target. Contemporaries of Allport in sociology followed a similar view of stereotyping as resulting from cognitive biases, bad social policy, and miseducation, rather than from instinct or direct animosity. Allport's model could accommodate a wide range of intergroup theories including scapegoat theory, social dominance theory, authoritarianism, realistic conflict theory, self-fulfilling prophecy, and social identity theory (Tajfel and Turner, 1979). Allport's (1954) model suggests that these may be 'normal' processes, but at the same time malleable from a social engineering perspective. In a concluding section of the book, he discusses the success of social change programs and asks prophetically in a chapter title "Ought there to be a law?" (p. 461). The

response of impatient social scientists, as noted by Jackson (2001) was to build "an argument that it was not necessary to wait and do nothing while society 'evolved' more egalitarian" (p. 9). Social scientific attention to unfair treatment of and attitudes toward stigmatized people has continued well into the twenty-first century (Falk, 2001).

Goffman's Stigma and the Spoiled Interaction

Just after Allport redirected the social science of stigma onto prejudice and the prejudicial, Erving Goffman brought the focus right back to stigma and stigmatized people in his 1963 book *Stigma: Notes on the Management of Spoiled Identity*. This was the first social psychological synthesis of stigma research in the twentieth century, discussing an array of stigmatized groups. Using a symbolic interactionist framework, Goffman characterized stigma as a situation in which a devalued attribute of an individual jeopardizes acceptance from others who anticipate a usual, expected, or 'standard' person but instead are faced with an unusual, unexpected, or discounted one. Goffman emphasized the interactional nature of interpersonal encounters between stigmatized and nonstigmatized individuals, including an analysis of ways in which stigmatized people cope with their situation. Goffman (1963) described three general types of stigma: bodily 'abominations' (e.g., physical irregularities), blemishes of individual character (e.g., criminal behavior), or tribal stigma (e.g., ethnicity). A person with a stigma may be viewed as abnormal due to their possession of a 'shortcoming' and thus be subject to either deliberate or unintentional marginalization and oppression including prejudice, stereotyping, and discrimination. Further emphasizing the relational nature of stigma, Goffman distinguished two types of stigma, the discredited (who are already known by others to have a stigmatizing trait) and the discreditable (who have the trait, but others may not already know).

There has been no other treatment of the subject of stigma that parallels Goffman's approach to categorizing and classifying the nature and types of stigma, excepting Jones and colleagues' (1984) *Social Stigma: The Psychology of Marked Relationships* published 20 years later. Jones and colleagues described six dimensions of stigma that expand upon Goffman's discredited/discreditable dichotomy, adding concealability of the stigmatizing attribute, course (duration and permanence) of the stigmatizing attribute, disruptiveness of the stigmatizing attribute in social interactions, aesthetic qualities of the stigmatizing trait, origin or cause of the stigmatizing trait, and perilousness of the stigmatizing trait to others. Together, these works ensured that research on the experiences of stigmatized people would continue; unfortunately, traces of the perspectives of eugenicists, race theorists, and the social gospel remained. Several of Goffman's source materials were products of these earlier agendas that preceded him and much of his terminology would be deemed overtly degrading by modern standards (e.g., contrasting 'stigmatized' 'cripples,' 'illiterates,' and 'very Orthodox Jews' with 'normals'). Goffman criticized stigmatized individuals who "flaunt their refusal to accept their place" (p. 145) and described coping in patronizing tones as in "the tortured performance… when an individual confined to a wheelchair manages to take the dance floor with a girl in some kind of mimicry of dancing" (p. 10). He described 'defiant' stigmatized individuals as those who have the audacity to think of themselves as 'normal.' He also roused old themes of depraved culture in his description of the culture of deviant elements in society:

> Prostitutes, drug addicts, delinquents, criminals, jazz musicians, bohemians, gypsies, carnival workers, hobos, winos, show people, full time gamblers, beach dwellers, homosexuals, and the urban unrepentant poor… failing to use available opportunity for advancement in the various approved runways of society…show open disrespect for their betters; they lack piety; they represent failures in the motivational schemes of society. (p. 171)

Some of these early themes of deviance continue to haunt the public's views. For instance, people are more supportive of stigmatizing people whose status is seen as uncontrollable/ascribed rather than controllable/achieved, and stigmatized group members make difficult choices about concealing, passing, or revealing the origins of their stigma. Goffman implied that the responsibility to ease the problems of stigma lies with stigmatized individuals not with 'normals' because "normals really mean no harm; they don't know better" (p. 116). Reflected in contemporary psychological work that favors individualistic sociocognitive coping over group solidarity and activism, Goffman saw 'good adjustment' as in the hands of the stigmatized:

> It requires that the stigmatized individual cheerfully and unselfconsciously accept himself…voluntarily withholds himself from those situations in which normals would find it difficult…it means that normals can remain relatively uncontaminated by intimate contact with the stigmatized, relatively unthreatened in their identity beliefs. (p. 121)

Goffman laments, that although this means taking on all this responsibility for society's weakness, this strategy promises "the best return he can get on his money" (p. 122).

Goffman's version of the social gospel guaranteed that in the second half of the century researchers would continue to wrestle with the idea that the possession of stigma is universally burdensome. His fatalistic view of stigma reinforced self-hate themes even as integration and upward social mobility meant that many stigmatized individuals were gaining in status. Research turned to characterizing stigmatized individuals in intergroup settings as being hassled, vigilant, distracted, and tokenized. Indeed, even Goffman recognized that context is key, although his obsession with the demoralizing treatment of stigma assumes a backdrop of a 'mainstream' 'normal' cultural context. Goffman cautions that all people have experience negotiating stigma from both sides, because we all possess attributes that in some contexts can be stigmatizing. The individual, contextual, and cultural relevance of any stigmatizing attribute fluctuates (Campbell and Deacon, 2006). There are even contexts in which culturally valued social groups can experience stigmatization due to 'positive deviance' and the 'model minority stereotype.'

Influence of the Civil Rights Movement

The cultural context began to change dramatically throughout the civil rights movement, and researchers had to disentangle

the erroneous confounding of stigmatized identity with intellectual inferiority, mental illness, and low self-esteem (Crocker et al., 1998; Cross, 1991). The conclusion that self-hatred is inevitable and only artificially correctable has been questioned by findings that members of stigmatized groups typically have high self-esteem. The work on stigma moved into a new phase in view of new cultural norms, new rights, higher social standing, celebration of group pride, and a rising number of minority scholars who changed not only how members of stigmatized groups were perceived in society, but also how they were portrayed in academia. The old idea that marginalized groups are immoral and their damaged culture pathological was replaced with a reclaiming of culture of as "valuable and worth preserving" (Jackson, 2001: p. 3).

During this time, complex and affirmative theories of self-concept emerged acknowledging the within-group variability in the group identities of members of stigmatized social groups, including identities that are politicized and empowering. Group identity has been shown to be positively associated with well-being, motivation, resilience, performance, self-protection, and willingness to stand up to social injustice. Research has revealed a diverse and multifaceted concept of group identity that has universal and group specific characteristics, is flexible to changes throughout the life span and across situations, and may be both empowering as well as challenging to achieve and sustain, especially for groups who are highly stigmatized or geographically isolated.

Contemporary Levels of Analysis

Social psychologist Thomas Pettigrew has noted how psychologists "rarely combine levels of analysis in order to put individual and situational phenomena in their macro-contexts" (Pettigrew, 2010: p. 615). Allport's (1954) exhaustive theory was rare in its attention to the individual and intergroup levels, as well as sociocultural and historical perspectives. Research on stigma combining all its multileveled complexity has proved to be a significant challenge for researchers. As one's focus is crisply positioned at one level, there is another level that is made fuzzy.

Stigma at the Individual Level

At the turn of the twenty-first century, a focus on stigma from the *Target's Perspective* contributed to the understanding of stigma in two major ways (Swim and Stangor, 1998). First, the *Target's Perspective* prioritized the insider's perceptions on their daily lives with qualitative reporting of their experiences by and for a diversifying group of academics. Concepts of 'everyday prejudice' and 'microaggressions' emerged to characterize the mundane experiences with prejudice in daily living. This approach represented a marked change from Allport, Goffman, and their contemporaries who had (with some exceptions) been writing secondhand about stigma to a nonstigmatized audience. Thus, the *Target's Perspective* made it possible for stigmatized individuals to represent their daily experiences in their own words in the scientific literature. This writing improved the language and assumptions made in past social scientific work and revealed some unexpected results, such as the reserve stigmatized group members exercise when perceiving and responding to disadvantage. Second, the *Target's Perspective* benefited from the opportune timing of its emergence post-Allport and post-Goffman, which afforded researchers the ability to move from the contextualized experiences of a single stigmatized group to a larger general theory of prejudice and stigma, and vice-versa. The localized study of certain stigmatized groups' experiences from the insiders' voice became valuable in its own right, and at the same time essential to developing the larger theory of stigma and prejudice in society. There is now a more empowering view of stigmatized individuals, not as victims constrained by a predetermined set of inferior life outcomes, but as active copers adjusting strategies to meet individual and situational needs. Individual differences in how stigmatized people cope have been given new attention. *Target's Perspective* research that explores 'confrontation' over 'coping' has revealed that targets have a well-developed repertoire to draw from in acts of interpersonal resistance (Hyers, 2010). Shih (2004) has argued that the most resilient in the battle against stigma are those who adopt 'empowerment' strategies rather than framing their efforts merely as 'coping' with the stigma they face.

When focusing on stigma from the *Target's Perspective*, two major limitations become apparent. First, it could be misinterpreted as implying that the responsibility for correcting oppressive social stratification lies solely in hands of stigmatized people without also addressing sociostructural inequalities (Hyers, 2010). Second, it may falsely represent stigmatized groups as orthogonal and one-dimensional (as if people in the category 'lesbian' do not have an ethnicity, people in the category 'African-American' do not have a gender, etc.). Theorists coming out of gender and ethnic studies traditions offer a range of metaphors for the complex ways in which stigmatization and identity manifest at the individual level. Crenshaw's (1991) influential intersectionality theory emphasizes that all people are multidimensional, group memberships are not orthogonal once embodied, stigmatizing forces intersect in unique ways, and when people possess multiple group memberships (e.g., lesbian women of color with disabilities), these dimensions cannot be easily separated for analysis. This problematizes discussions of singular prejudice or measures of one-group identity, since experiences and identity are shaped by a combination of factors such as race/ethnicity, social class, gender, income/socioeconomic status, and disability status. Research comparing the experiences of two stigmatized groups (e.g., lesbians being compared to people with disabilities) implies that the groups are mutually exclusive, neglecting the 'double' and 'triple jeopardy' of multiple stigmas. Stigmatizing attributes are not simply additive; for example, there is not just racism and sexism, but also gendered racism, and gendered racism may be further impacted by socioeconomic status.

Stigma at the Interpersonal/Intergroup Level

Allport's (1954) attention to the intergroup situation along with Kurt Lewin's work on group dynamics encouraged research on intergroup contact that has grown quite prominent in the psychological research on stigma, especially

with regard to identifying and enhancing the intergroup processes that lead to internal cognitive and emotional change. According to social identity theory, individual identity is shaped by group memberships; therefore, individuals engage in strategies that elevate the status of their groups (Tajfel and Turner, 1979). Expanding upon Allport's (1954) distinction between intropunitive actions aimed at adapting to others and extropunitive actions aimed at blaming outside forces, social identity theory proposes three strategies: social 'creativity' (valuing one's in-group), individual mobility (striving to leave one's stigmatized status), and social competition (action against injustices). This work has been flexible enough to be relevant to work on stigma and stigmatization, perpetrators and targets, and prejudice formation and reduction. Intergroup research has been criticized for its use of artificial laboratory settings, minimal groups design, and experimentally created and randomly assigned 'stigmas.' A second criticism is in regard to the negative light that has been cast on intergroup encounters, especially with regard to the stigmatized participants, since research shows that both stigmatized and nonstigmatized individuals exhibit intergroup anxiety and avoidance coping. A third criticism is in regard to the promotion of the strategies of the common in-group identity model and 'colorblindness,' which contradict other work at the individual level on the importance of contextualized identity (e.g., intersectionality theory). Intergroup research that focuses on group-honoring intergroup processes, such as appreciating difference and self-reflection, may fit more nicely with perspectives on stigma embodied and *Target's Perspective* research on the 'educational moment' in dynamic exchanges when stigmatized individuals confront others. Regretfully, similar to the *Target's Perspective* research, interpersonal/intergroup research has not given enough attention to the dynamic exchange of actions and interactions of both players in the immediate intergroup context.

Stigma at the Macrosocial Level

Research and intervention on stigma directed at the societal level complements the previous two levels of analysis. As Steele et al. (2002) note, stigma does not reside in the person "but in the unfortunate circumstance of possessing an attribute that, in a given social context, leads to devaluation" (p. 506). In other words, a stigmatizing attribute is possessed by an individual, but defined by larger social and cultural constructions of power. The field of psychology tends to direct the focus away from vast macrosocial forces including economic, political, and social unrest; institutional discrimination; global geopolitical power dynamics; and historical inequities. Research on interdependence and power shows that power impacts not only how prejudice is expressed, but also how one perceives and responds to it. Coping with one stigmatizing attribute (e.g., disease, poverty, and criminal status) is affected by the possession of other hierarchically organized identities (e.g., skin color, nationality, and sexual orientation) and the impact of existing cultural institutions (e.g., hospitals, schools, churches, and marriage rites). Stigma has been most promisingly addressed in educational institutions and workplace settings.

New Directions

Addressing the Subtlety of Stigma

The response of the public to stigma has changed since the civil rights movement, and so, too, has the approach of social science. *De facto* stigmatization is illegal yet *de jure* stigmatization remains, making it harder for the public to understand, victims to identify, the legal system to prosecute, and for researchers to study. Nier and Gaertner (2012) dedicated a special volume of the *Journal of Social Issues* to "The Challenges of Detecting Contemporary Forms of Discrimination." The shift toward the subtle is reflected in modern, symbolic, aversive, ambivalent, and benevolent prejudice theories and in work on dominant group privilege, entitlements, and conferred dominance. Stereotype threat research similarly suggests that situational threat can be quite subtle. Research methods and measures have grown more discrete using subtle measures of implicit associations, response latency, heart rate, eye blinks, functional magnetic resonance imaging, and delayed memory deficits. Even *Target's Perspective* research, with its focus on the personal narrative and voice of stigmatized individuals, recognizes that chronic, subtle, everyday prejudice can be intentionally or unintentionally forgotten or dismissed, ergo the use of daily diaries, experience sampling, checklists, palm pilots, and photo diaries to document the subtle but accumulating stigmatization.

Stigmatized Groups in Focus in the Early Twenty-First Century

Link and Phelan (2001) remind us that the concern about stigma is first and foremost a concern over unfair limits placed upon human chances for health, economic security, housing, and education, as well as psychological well-being. As has been the case throughout the history of academic work on stigma, contemporary social problems shape the discourse and research agenda. Several stigmatized groups facing pressing injustices have been accorded more intensive focus of late.

Stigma of Poverty

Early Race Theorists had described poverty as a moral failing resulting from the intergenerationally transmitted deviant 'culture of poverty.' For decades, psychologists interested in prejudice still left concerns over the stigma of poverty to sociologists and economists, because social class was not seen as a psychological identity, per se. With income inequality increasing in the United States and globally, the American Psychological Association finally established a Task Force on Socioeconomic Status in 2007. International commerce has intricately tied the socioeconomic statuses of people in cultures around the globe. In the increasingly connected international labor force, the stigma of relative poverty is as relevant as the stigma of absolute poverty. Ironically, race theories from the past continue to resurface, some actually suggesting that stigma and shame are functional because they motivate the otherwise "indolent masses" to strive (Murray, 2012).

Stigma of Aging

Another rising stigmatized group that has been calling for more attention is the aging population. Entrenched institutionalized ageism and derogatory cultural stereotypes of aging are occurring at a time when cosmetic and medical advances to 'turn back

the clock' are booming, and populations are living longer. With a disproportionate aging population in many parts of the world where aging is viewed negatively, the stigma of aging is coming to the fore.

Stigma of Disability

Goffman's writing on disability was some of the most offensive, especially when he described disability as a situational 'failing' of the stigmatized person for not being the 'normal' person others expected. In dramatic contrast, a major social shift propelled by disability activism emphasizes that in fact, the person making normality assumptions has 'failed.' This represents a new standard in which people in all contexts must anticipate and accommodate all variations of human competencies. Standards of the Americans with Disabilities Act compliance, ecological design, the social (and the radical antisocial) models of disability, and even demand for 'safe spaces' have changed the playing field dramatically since Goffman. After decades of efforts by the mental patients' rights movement, the World Psychiatry Association established guidelines for reducing the stigma of mental illness in 2010. This effort is further strengthened by new attention to war veterans and first responders who may face disability, trauma, and social isolation, but feel inhibited from asking for help.

Future Frontiers

The last century has brought substantial progress to the research and realities of stigma in our culture; however, there is more work ahead (Barreto and Ellemers, 2010). From a global perspective, the progress of Western Psychology has not crossed boundaries as much as it should. Stigma remains critically divisive to the social order within and across many cultures, in the relationships between 'fourth world' indigenous and 'first world' colonizing nations, and in the struggles of immigrants and refugees around the globe. Furthermore, in the virtual world of cyberspace, there is a new unregulated domain wherein stigmatization has been transferred from the real world and sometimes back again in the form of 'cyber bullying.' Interestingly, a number of 'antibullying' and 'allies' programs have developed into widespread and unified responses to extreme incidents of cyber bulling. Stigma research also has the potential to make major contributions to environmental protection. There are obvious parallels between humans' stratification of each other and their stratification of nonhuman animals and the environment – differentiating, collecting, classifying, labeling as primitive, exoticizing, and thereby legitimizing exploitation (Swim et al., 2009). As the understanding of and sensitivity toward stigmatization continues to grow, one can hope to see increasingly more successful collaborations between activists and applied science scientists in the years to come.

See also: Dehumanization; Desegregation; Essentialism; Identity and Identification, Social Psychology of; Implicit Social Cognition; Intergroup Relations; Meritocracy and Tokenism; Sexism; Social Dominance Orientation; Social Identity in Social Psychology.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley, Reading, MA.
Baker, W., Smith, L., 1939. Facial disfigurement and personality. Journal of the American Medical Association 112, 301–304.
Barreto, M., Ellemers, N., 2010. Current issues in the study of social stigma: some controversies and unresolved issues. Journal of Social Issues 66, 431–445.
Campbell, C., Deacon, H., 2006. Unravelling the contexts of stigma: from internalisation to resistance to change (Special issues on stigma). Journal of Community and Applied Social Psychology 16, 411–417.
Clark, K.P., Clark, M.P., 1939. Segregation as a factor in the racial identification of Negro pre-school children: a preliminary report. The Journal of Experimental Education 8, 161–163.
Crenshaw, K.W., 1991. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Review 43, 1241–1299.
Crocker, J., Major, B., Steele, C., 1998. Social stigma. In: Gilbert, D., Fiske, S.T., Lindzey, G. (Eds.), Handbook of Social Psychology, fourth ed., vol. 2. McGraw-Hill, Boston, pp. 504–553.
Cross, W.E., 1991. Shades of Black: Diversity in African-American Identity. Temple University Press, Philadelphia.
Falk, G., 2001. Stigma: How We Treat Outsiders. Prometheus Books, New York.
Goffman, E., 1963. Stigma: Notes on the Management of Spoiled Identity. Simon and Schuster, New York.
Guthrie, R.V., 2004. Even the Rat Was White: A Historical View of Psychology. Pearson Education, Upper Saddle River, NJ.
Hilgard, E., 1986. From the social gospel to the psychology of social issues. Journal of Social Issues 42, 107–110.
Hollingworth, L., 1920. The Psychology of Subnormal Children. Macmillan Company, Norwood, MA.
Hyers, L.L., 2010. Choosing alternatives to silence in face-to-face encounters with everyday heterosexism. Journal of Homosexuality 57, 539–565.
Jackson, J.P., 2001. Social Scientists for Social Justice: Making the Case against Segregation. New York University Press, New York.
Jones, E.E., Farina, A., Hastorf, A.H., Markus, H., Miller, D., Scott, R.A., 1984. Social Stigma: the Psychology of Marked Relationships. Freeman, New York.
Link, B.G., Phelan, J.C., 2001. Conceptualizing stigma. Annual Review of Sociology 27, 363–385.
Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.
Murray, C., 2012. Coming Apart: the State of White America, 1960–2010. Random House, New York.
Nier, J., Gaertner, S.L., 2012. The challenge of detecting contemporary forms of discrimination. Journal of Social Issues 68, 207–220.
Pettigrew, T., 2010. Looking to the future. In: Dovidio, J.F., Hewston, M., Glick, P., Esses, V.M. (Eds.), The Sage Handbook of Prejudice, Stereotyping, and Discrimination. Sage, Thousand Oaks, CA.
Shih, M., 2004. Positive stigma: examining resilience and empowerment in overcoming stigma. The Annals of the American Academy of Political and Social Science 591 (1), 175–185.
Steele, C.M., Spencer, S.J., Aronson, J., 2002. Contending with group image: the psychology of stereotype and social identity threat. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 34. Academic, San Diego, CA, pp. 379–440.
Swim, J., Clayton, S., Doherty, T., Gifford, R., Howard, G., Reser, J., Weber, E., 2009. Psychology and global climate change: addressing a multi-faceted phenomenon and set of challenges. A report by the American Psychological Association's task force on the interface between psychology and global climate change. Retrieved 15.04.10.
Swim, J.K., Stangor, C., 1998. Prejudice: the Target's Perspective. Academic Press, New York.
Tajfel, H., Turner, J.C., 1979. An intergrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–47.

Stochastic Models

David J Bartholomew, London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Human behavior is, to some degree, unpredictable. The description and prediction of social systems therefore requires models of such systems to be stochastic. Most social systems can be described at any time by the numbers of people in various categories. Change over time is described by probabilities and these enable one to trace the trajectory of the system over time using the theory of stochastic processes. This article describes some of the basic models that have been widely used for such systems and suggests that the way forward depends mainly on constructing realistic simulation models of real-life systems.

Terminology and Background

The term stochastic model is used in a broad sense and a narrow sense in the social sciences and beyond. In the broad sense, it refers to any model involving probability and thus includes, for example, the standard regression model and random sampling from finite populations. In the narrow sense, it refers to models for social systems and it is in that sense that it is used here.

The characteristic feature of such a model is that it involves the unpredictable behavior of a collection of people. It attempts to describe the behavior of such a system, usually over time. Its starting point is with individuals who behave unpredictably, to some extent, and then it seeks to deduce the aggregate properties of the system. It is an empirical fact that the behavior of individuals varies. In making purchases in a supermarket, for example, individuals will choose different brands, apparently haphazardly and in voting in an election they will back different candidates. In a purely deterministic world, every such action and choice would be determined, and in principle at least, the future would then be entirely predictable. The reality is quite different.

Individual unpredictability has no particular philosophical implications. It does not imply, for example, that decisions are made as if by the toss of a coin. It could be, of course, that each of us is a random automaton with some neural randomizer that determines our every thought and action but that is not necessary. As a matter of fact, choice does mimic chance and patterns of choice can often be well described by probability distributions. In technical language, the element of chance in our models is not ontological but epistemic.

The rapid growth in the power of computers in the last few decades has radically changed the context in which stochastic modeling takes place. In order to delineate this change, it is useful to see the field as consisting of two parts. First, there is the construction of stochastic models and second, their analysis. It is on the analysis that computing has had the greatest effect. Models that formerly would be regarded as involving too much heavy computation have now come within the realm of the possible. This might be seen as a source of encouragement to construct ever more complicated models. But on the construction side, the effect has been much more subtle. The limiting factor there is no longer the amount of computing power available but the capacity of the analyst's mind to encompass the results of the exercise. The mere proliferation of models in number and complexity achieves little unless it provides insight into the fundamental character of the processes being modeled. The prime purpose of a stochastic model is to obtain a firmer intellectual grip on the reality underlying the process studied. The strategy in this article is, therefore, to focus on simpler models which, although may not always be fully realistic, give genuine insight into the reality that the model is designed to capture.

Uncertainty in human behavior is not confined to the social sciences as commonly understood, of course, but it occurs also in many contiguous areas. Population genetics and epidemiology are two examples and some of the models mentioned had their origins in these fields. Nevertheless, there is a unity in models used in the social sciences, which justifies their being treated as a coherent whole.

In this article, the author traces a progression from the simple to the complex. Beginning with an example simple enough to easily comprehend and yet being capable of displaying some of the subtlety of more complex models, the author goes on to develop the ideas in larger and more realistic examples.

A Simple Example (The Poisson Model)

The simplest of all probability models stems from the idea behind coin tossing. If a coin is tossed independently and repeatedly, one observes a succession of heads and tails in roughly equal numbers but in no particular order. It is a straightforward extension of this model to suppose that the chances of heads and tails are not equal and to note in particular, that there is nothing to prevent the probability of heads being very small. There are, in fact, many situations in which human choice making shows these characteristics. For example, the chance that any individual meets a long-lost friend on any particular day may be very small indeed, but in the course of a year they may meet several. The actual number is uncertain and that number will have a probability distribution. It is a remarkable fact that the probability distribution of this number often closely follows what is

known as a Poisson distribution. The probabilities do not even have to be equal but providing they are all very small and the number of occasions is large, the probability of observing exactly n occurrences will be

$$P(n) = m^n/_{n!} \exp(-m) \qquad [1]$$

where m is the mean number of occurrences and n can take any nonnegative integer value. This is a natural place to start since it makes minimal assumptions about behavior. In practice, however, it often fails to represent the pattern of variation adequately. For this reason, it is sometimes referred to as a baseline model as it forms the starting point for a deeper investigation. One useful property of the Poisson distribution is that its mean and variance are equal. A quick check on the adequacy of the fit of any distribution is therefore to compare the mean and variance. It is often found that the observed distribution is 'overdispersed' meaning that there is more variation than one would expect on the Poisson hypothesis. An important function of modeling is to try to explain this departure by generalizing the model.

There are many possible generalizations. The most natural, perhaps, is to relax the assumption that all individuals experience the same risk. The parameter m may be thought of as a measure of risk because it is the mean number of accidents per unit time. One could suppose that there were two risk groups present – a high-risk group and a low-risk group. The observed distribution would then be a mixture of two Poisson distributions one with parameter m_1, say, and the other with m_2. Alternatively, the sample might have come from some larger population in which risk varied continuously. A common assumption is that m varies continuously according to a gamma distribution whose density function may be written

$$h(m) = \frac{c^q}{\Gamma(q)} m^{q-1} e^{-cm}. \qquad [2]$$

The resulting distribution of counts is then

$$P(n) = \binom{n+q-1}{q-1} \left(\frac{1}{c+1}\right)^n \left(\frac{c}{c+1}\right)^q \qquad [3]$$

This is known as the negative binomial distribution because its successive terms can be generated by the expansion of a binomial expression with negative index. It is easy to show that its variance is greater than its mean, as anticipated.

There are many other possible generalizations. It may be that individuals do not differ, but instead, that successive accidents are not independent. For example, it could be that having one accident makes one more liable to have another. Neyman proposed what are called 'contagious' distributions, which embody this feature. If accidents tended to cluster, there would be relatively more periods with more accidents than expected and others with fewer, since the mean is fixed. Qualitatively, this has a very similar effect to mixing.

These examples illustrate how successively more complicated models may help to identify new features and, more importantly, perhaps, demonstrate the limits of modeling. An important aspect of this is the indistinguishability of quite different models for the same phenomena.

Time has played no explicit role in the models so far. Accidents do occur in time, of course, but the mere passage of time was supposed to play no part. Most of the models to be discussed in this article develop over time, and time itself is an important factor in determining what happens.

One can introduce time into the simple coin tossing model by supposing that tosses take place at unit intervals of time. One then looks at the intervals between successive 'heads.' For a head to occur at the nth toss after the last one, there must be $n - 1$ tails in succession followed by a head. The probability of a head thus has probability $(1-p)^{n-1}p$. If n is the interval length, the probability distribution of n is therefore

$$P(n) = (1-p)^{n-1}p, \ldots n = 1, 2, \ldots \qquad [4]$$

This is the geometric distribution. The random variable n now represents the time before a head first occurs. One speaks of this as a discrete time model.

Suppose one imagines the time interval between tosses becomes very small, of length δt. On the new scale of measurement, the time interval between heads will be $t = n\delta t$. If one were to leave p unchanged, the interval between heads would become very short. To compensate for this, one needs to allow the probability to become small at a rate that exactly balances the increased rate of tossing. Thus, suppose $p = \lambda\delta t$, then

$$\Pr\{\text{head occurs in } (t, t+\delta t)\} = .\lambda\delta t \exp(t/\delta t - 1)\text{In}(1 - \lambda\delta t) \qquad [5]$$

As $\delta t \to 0$ this approaches

$$\lambda e^{-t}\delta t. \qquad [6]$$

The probability $p = \lambda\delta t$ may now be called an infinitesimal probability, λ the rate of occurrence, and $f(t) = \lambda e^{-\lambda t}$ the probability density function of the time interval.

This link between the discrete time version of a model and its continuous time counterpart arises frequently in stochastic modeling and one shall meet it again in connection with Markov models below.

As with the Poisson distribution, the very simple assumptions on which the waiting time distribution depends are seldom met in practice but provide a good baseline model from which more realistic models may be derived.

Accounting Equations

A very large number of social systems have the following structure. People are distributed over a number of categories. Over time, individuals move from one category to another, or leave the system altogether. If these moves are, in some degree, unpredictable, one needs a stochastic model to describe the development of the system. The starting point for almost all models is a set of accounting equations connecting the stocks and the flows. The numbers in the categories at any time are the stocks and the number in category i at time T is denoted by $n_j(T)$. Similarly, the number moving from i to j in the interval from $T-1$ to T may be denoted by $n_{ji}(T-1,T)$ and these are the flows. If $i = j$, this is the

number that does not move. All internal movements are then accounted for by the equations

$$n_j(T+1) = \sum_{j=0}^{k} n_{ji}(T, T+1) \qquad [7]$$

where k is the number of categories and i runs from 1 to k. Movements into the system from outside may be covered by letting the subscript 0 denote the outside world. Losses in the interval are not in the system at time T and therefore do not appear in eqn [7]. Any stochastic model of the system will have to specify the process of change.

In a Markov model, it is assumed that the probability that an individual moves from i to j in any interval depends only on the origin state and that that all such moves are independent. At first sight, the requirement for the probability to depend only on the current state seems very restrictive but, as seen later, this difficulty can be considerably ameliorated by an appropriate definition of the states.

Markov Models

The Discrete Time Closed Model

Three kinds of Markov models are described but they all have one thing in common, which justifies the terminology. This is the set of transition probabilities between states, which may be set out as follows.

$$\begin{pmatrix} p_{11} & p_{12} \cdots \cdots \cdots & p_{1k} \\ p_{21} & p_{22} \cdots \cdots \cdots & p_{2k} \\ p_{k1} & p_{32} \cdots \cdots \cdots & p_{3k} \end{pmatrix} \qquad [8]$$

where p_{ij} is the probability of moving from state i to state j at any stage. In practice, some of the elements may be zero as when the states form a hierarchy of some kind in which downward movement is impossible. If the system is closed, so that no one enters or leaves, the row sums will all be unity. Such a system is 'closed' and the existing theory of Markov chains can be applied with little modification.

Totally closed systems are quite rare in practice but they may be a reasonable approximation especially when any replacements are equivalent to those that have been lost (see also the third type of Markov model discussed below where the correspondence is exact).

The development of a Markov chain over time can be traced by means of the expected numbers in the categories. Thus, let $n_i(T+1)$ denote the number of individuals in state i at time $T+1$. These numbers may be collected into a row vector $n(T+1)$. To calculate the expectation of $n(1)$ given $n(0)$, say, one must enumerate all the paths from each state at time 0 to each state at time 1 and compute their expectations. For example, state j can be reached at time 1 in one step from state i at time 0 and the probability that this happens, conditional on being in state 0 initially, is p_{ji} and hence the expected number in state i is $n_i(0)p_{ji}$ or, in matrix notation,

$$n(1) = n(0)P$$

where $\boldsymbol{P}$ is the transition matrix defined above. Repeated application of this argument for successive points in time gives

$$n(T+1) = n(0)P^{T+1}.$$

It follows at once that the elements of P^{T+1} give the probabilities of going from state i to state j in $T+1$ steps. These are known as the $(T+1)$ step probabilities and are sometimes written

$$p_{il}^{(T+1)}.$$

There are two state vectors that are of particular interest. These are the 'limiting' vector, if it exists, and the 'stationary' vector. If one follows the system for extended periods of time, the state vector might settle down to a limiting value or it might vary indefinitely. The former situation applies under certain conditions, which amount, roughly speaking, to saying that it must be possible to reach any state from any other state. In that case, a limiting state vector exists. It may be found empirically by repeated powering of the transition matrix, or by other means, but it turns out that P^{∞} has identical rows, which may easily be deduced from eqn [9] below.

The second state vector of particular interest is the stationary vector. A vector is stationary if it is unchanged by the transition process. That is, it must satisfy

$$n = n\,P. \qquad [9]$$

This stationary vector turns out to be identical to the limiting vector.

In practice, the Markov model is often used to predict the numbers of individuals who will be in the categories at different times in the future. Thus, if initially there are N people distributed across the categories according to $\boldsymbol{n}(0)$, the expected numbers in the categories at time T will be given by $n(T)$. Similarly, numbers in the categories in the limiting, or stationary, state will be n. The stationary state is of particular interest in many applications because it represents the stable structure that can be maintained indefinitely. Any other state distribution cannot be maintained under the operation of the given transition matrix.

The foregoing theory is concerned only with expected values. It is easily possible to find recurrence relations from which the variances, covariances, and higher moments of the category sizes may be determined. These will be found, for example, in Bartholomew (1982: pp. 20ff).

Open Markov Systems with Fixed Input

Many systems, such as, for example, educational systems, are fed by a constant flow of new entrants. This total number may vary over time but, in order to concentrate on the principle of the matter, it is assumed to be a fixed number, R, say. Suppose that these are allocated to categories with probabilities $\{r_{\downarrow}i\}$. The expected flow into category i, which is denoted by $n_{0i}\,(T-1, T)$ in eqn [7] will then be Rr_i. Taking expectations on both sides of eqn [7] and assuming that internal movements are governed by the transition matrix $\boldsymbol{P}$, one then finds

$$n(T+1) = n(T)P + R\boldsymbol{r}. \qquad [10]$$

If there were no losses, the size of the system would grow without limit, but otherwise, the structure will approach a limiting value satisfying

$$n = nP + Rr_i \qquad [11]$$

or

$$n = Rr(I - P)^{-1} \qquad [12]$$

where the existence of the inverse is guaranteed by the fact that some, at least, of the row sums of $\boldsymbol{P}$ must be strictly less than one.

Open Markov Systems with Fixed Size

There are many fields where there are constraints on the size of the whole system or, even, on the size of individual categories. Manpower systems are a good example. In such cases, the model needs to be modified and this is easily done when it is the overall size that is fixed. All that one needs to do is to ensure that the total input matches the total loss. For this purpose, a symbol is needed to denote the probability of leaving state i, which is obviously one minus the total probability of moving elsewhere in the system. It is convenient and natural to denote this by p_{i0}. The expected number of leavers between T and $T+1$ is then $n(T)p_0'$ where p_0' is the column vector with elements $\{p_{i0}\}$. This must be precisely equal to the number of new entrants; hence, one may substitute in eqn [11] to give

$$n(T+1) = n(T)P + n(T)p_0'r = n(T)\left[P + p_0'r\right]. \qquad [13]$$

It is easy to verify that the matrix in square brackets is a stochastic matrix (i.e., its rows all sum to 1). Hence, this model may be treated exactly as the closed model discussed above with the matrix $P + p_0'r$ replacing $\boldsymbol{P}$ of eqn [9].

Practical Issues with Markov Models

The Choice of Categories in a Markov Model

The most restrictive assumption of Markov models is that each transition probability depends only on the current state and not on any preceding states. In practice, this is often not realistic. For example, when studying a manpower system, the chance of an upward move (promotion) may depend on both the category (grade) and the length of time spent in that grade. If one ignores that dependence, predictions are likely to be in error. This problem can be circumvented by subdividing the grades and forming subgrades according to length of service. Thus, one would have a new grade consisting of those originally in grade 1 and having length of service up to 1 year; there would be a new second category consisting of those originally in grade 1 with between 1 and 2 years of service and so on. One could then model the new system more accurately because the effect of length of service would now have been taken account of in forming the grades. Before making the final predictions, the new categories can be recombined. In principle, the system could be made ever more Markovian by incorporating any other variables on which propensity to be promoted depended. However, there is a practical limit to the extent to which this device can be used. In practice, the transition probabilities will usually be unknown and so have to be estimated from flow data. The finer the subdivision, the less precise will the estimates of the transition probabilities become and, consequently, the errors of any predictions will be increased. A balance has to be struck between making the model more realistic and reducing the uncertainties of prediction.

Discrete Time and Continuous Time Systems

The choice between discrete and continuous time models is partly a matter of practical realism and partly of mathematical convenience. The development of the theory viewed in a historical perspective shows that realism and convenience have often combined to favor one or the other. In the case of Markov models, for example, changes of state often do, in fact, occur at fixed intervals of time, as when students in an educational system change state once every academic year. It is thus natural to formulate models of such systems in discrete time. Even in manpower systems, where change may take place at any time, it may be convenient to account for movements on a weekly or monthly basis. For such reasons, most of the relevant theory has been developed in discrete time. Nevertheless, it is worth observing that the theory described is closely paralleled by models in continuous time. Thus, in the case of the waiting time distribution discussed in the section A Simple Example (The Poisson Model), the time interval of the discrete model was allowed to become very small. The transition probabilities then became infinitesimal and the coefficients of δt became transition rates. Fuller details are given in, for example, Bartholomew (1982, Chapters 4 and 5).

When one moves on to other models, much of the theory required has been obtained, for various reasons, from first principles, in continuous time and this fact has largely determined the direction that research has taken. For example, this is true in the case of renewal models and most of the models for the diffusion of news and innovations. In the latter case, the pioneering work was in the field of epidemics where it is more natural and convenient to work in continuous time. As this is a review article, the author has followed established practice and the reader will note a shift from predominantly discrete time models to continuous models at this point.

Deterministic and Stochastic Models

The derivation of some of the models considered here, and certainly their use, has often proceeded as if all change was deterministic. A model devised on such principles may be described as a deterministic model. The distinction may be made clear by taking a particular case. In the Markov model, transition probabilities were introduced to describe the movement of individuals between states. Thus, p_{ij} was the probability of moving from i to j. The actual number moving would be a random variable with a binomial distribution and parameters n and p_{ij} and mean np_{ij}. If, by contrast, the number moving were supposed to be the exact number np_{ij}, one would have the deterministic equivalent of the Markov model. The expectations of the numbers in the stochastic version thus correspond with actual numbers in the deterministic version. If all one is interested in is the expectations, it might seem simpler to work deterministically from the start using averages instead of actual numbers. There is no distinction in the case of Markov models because the equations are linear. The stochastic model gives us much more, of course, enabling us, for example, to compute variances and covariances of the state numbers. But in

much practical work, in manpower planning, for example, predictions have not gone beyond the expectations and the deterministic interpretation suffices.

When one moves on to diffusion models, the linearity is lost and it is no longer true that the expectations derived from the stochastic model coincide with predictions of its deterministic counterpart. At best, the deterministic model serves as an approximation to the stochastic model when the numbers are large.

Renewal Models and Vacancy Chains

Basic Ideas

The author has considered a Markov model with a fixed total size. In some fields, the sizes of all categories are fixed. In such cases, the basic accounting relations expressed in eqn [7] still hold but the stock numbers on the left-hand side are now given numbers. This places severe constraints on the flows. Under these circumstances, one must begin by specifying what drives the system or, what initiates movement. Normally, such a system will be open with outgoing flows, which create the need to maintain the stock levels. This movement will then generate consequential flows within the system. Two approaches to modeling such systems have been explored. One draws on renewal theory whose origins are in population mathematics and the theory of industrial replacement. The author describes this below. The second was proposed by Harrison White (1970) in work on what he called 'chains of opportunity.' Modeling vacancy chains inverts the usual viewpoint by modeling the flows of vacancies rather than the flow of people. If a person leaves the system, the assumption that each category is of fixed size means that a vacancy is created. When that vacancy is filled from elsewhere in the system, one may think of the vacancy as moving to the place from where that replacement came. That movement will create a new vacancy (which can be thought of as moving elsewhere in the system) from where that replacement came. In general, every movement of a person is associated with a movement of a vacancy in the opposite direction. If one could assume that the movement of a vacancy was governed by a Markov chain, one could model the movements of vacancies in the system using Markov chain theory. This possibility has been explored for manpower systems in Bartholomew et al. (1991). More recently, the idea has been used in the study of housing markets (see Magnusson Turner, 2010), for example, where the creation of new housing stock creates vacancies into which people move from existing housing.

A Simple Renewal Model

Renewal models are another kind of stochastic model that has been used in the study of social processes and here the author describes a simple example. They have almost always been developed in continuous time, although a discrete time version of the most elementary example was actually included, as a special case, in the Markov model for a system of fixed size. Here, one deals with a similar case in continuous time because it has important, and perhaps surprising, practical implications.

Suppose that there is a system consisting of N persons and that that number remains fixed. The individuals who populate the system are assumed to behave independently and the uncertainty arises from not knowing when they will leave. This uncertainty is expressed by a hazard function defined as follows.

$$\Pr\{\text{loss occurs in } (t, t+\delta t)\} = \lambda(t)\delta t + o(\delta t). \qquad [14]$$

An increasing hazard function means that the individual is more likely to leave as time passes; a decreasing function means the opposite and more subtle changes can be expressed by more complicated functions.

The number of people who are expected to leave, which of course is exactly equal to the number joining, is given by the 'renewal function.' This function is denoted by $r(T)$ and is derived by a straightforward probability argument that shows that it satisfies the following integral equation

$$Nr(T) = Nf(T) + N\int_0^T r(T-t)f(t)\mathrm{d}t \qquad [15]$$

where $f(t)$ is the probability density function of lifetime in the system and $Nr(T)\delta t$ is the expected number of replacements required in the interval $(T, T+\delta T)$.

In principle at least, this equation can be solved mathematically by the use of Laplace transforms or numerically, using a discrete approximation. However, there is one case where the solution is immediate. This is when the life distribution is exponential – a distribution already met in eqn [6]. In that case, the hazard function is constant and so it turns out is the renewal function. Furthermore, the renewal function is also constant and equal to μ^{-1}, where μ is the average length of stay. This result accords with intuition because it says that if the average stay, for example, were 4 years, then the rate of leaving and replacement would be 25%. If this were a good approximation when the life distribution was not exponential, there would be little more to be said. But unfortunately, this is not the case for the life distributions that occur, for example, in labor wastage or house occupancy. In such cases, the hazard function is not constant but often declines quite steeply over most of its range. This means that the longer an individual has been in the system, the less likely they are to leave. If this happens, the short-term loss rate may be very far from the equilibrium value and then a solution to eqn [15] is necessary.

Nonlinear (Diffusion) Models

In this section, one turns to models that may be described as nonlinear. The Markov models discussed above were linear in the sense that the number moving between category i and category j depended only on the number in category i and this made the model linear. There are many social processes where transition depends on the numbers in more than one category in a nonlinear fashion. Although the fields of application are diverse, they share a common mathematical structure. The earliest application goes back to the 1950s when Anatol Rapport was studying what he called the random net, which had to do with the connectedness of neural arcs and the possibility of random connections

allowing transmission through the net. Next came the epidemic theory, where an infectious disease might spread through a population as a result of random contacts between members of the population. This was easily extended and modified to cover the spread of news or rumors through populations of randomly mixing populations. Less obvious, perhaps, is the example of social groups competing for membership and this is the example the author shall use here to illustrate the typical problems that arise. Cross-fertilization has been hampered by the varied disciplinary allegiances of researchers in the field and the journals that have published their work, for example, the spread of information as described in Rabajante (2012).

In the simplest case, there are two groups, political parties say, competing for members in a population of fixed size. At any time, the population may be divided into three groups: those who are members of group 1, those who are members of group 2, and those who are unaffiliated. The composition of the population changes as a result of the following events. A member of group 1 meets an uncommitted person and recruits them to membership of that group, a member of group 2 similarly recruits an uncommitted member, or members of one group meet members of the other group and convert them to membership of their own group. Members of either group may also withdraw from membership and so become unaffiliated. It is assumed that the population is homogeneously mixing, which means that any pair is just as likely to meet as any other pair. On meeting, either member of the pair may change state according to rules that must be specified. This may be formalized as follows.

The state of the whole system may be described by the number of members of the two groups. In any small interval of time of length δT, any one of the following transitions may take place:

(1) $(n_1, n_2) \rightarrow (n_1 + 1, n_2)$
(2) $(n_1, n_2) \rightarrow (n_1, n_2 + 1)$
(3) $(n_1, n_2) \rightarrow (n_1 - 1, n_2)$
(4) $(n_1, n_2) \rightarrow (n_1, n_2 - 1)$
(5) $(n_1, n_2) \rightarrow (n_1 - 1, n_2 + 1)$
(6) $(n_1, n_2) \rightarrow (n_1 + 1, n_2 - 1)$

The probabilities of these various transitions will depend on two things. First, on the parameters chosen to reflect the effectiveness of the effort made to affect recruitment, conversion, or in the case of (3) and (4), the propensity for membership to lapse. Second, on the chance that a contact of the designated kind will take place. Thus, for example, the infinitesimal probability of transition (1) may be written,

$$\beta_1 n_1(T)\{N - n_2(T) - n_2(T)\}\ \delta T \qquad [16]$$

where $\beta_1 \delta T$ is the infinitesimal probability that a member of group 1 will recruit an uncommitted person when contact is made. The number of such pairs is given by the remaining part of the expression in eqn [16]. It is clear from eqn [16] why the model is called nonlinear; this probability is a quadratic function of the numbers in the states.

It is not possible to obtain exact expressions for the probabilities of the system being found in any state (n_1, n_2) at time T, but various partial or approximate methods are available. One is to exploit the power of computers to simulate the process. This permits experimentation with different parameter values and the relaxation of the highly restrictive assumption of homogeneous mixing. The direction in which recent research on these lines is taking can be seen in Gong and Xiao (2007). A second approach is to ignore the time element of the model and study the sequence of states through which the system passes with a view to studying the terminal behavior. This is facilitated by the fact that the process defined can be shown to be an absorbing Markov chain defined on the set of states of the form (n_1, n_2). A third approach is to use what is called a deterministic approximation, which replaces the infinitesimal transition probabilities by proportions. In this case, the rates of change of the state numbers can be expressed as linear functions of the state numbers. By this method, one can also investigate possible equilibria that occur when the derivatives are zero.

Bibliography

Bartholomew, D.J., 1982. Stochastic Models for Social Processes, third ed. John Wiley and Sons, Chichester.

Bartholomew, D.J., Forbes, A.F., McClean, S., 1991. Statistical Techniques for Manpower Planning, second ed. John Wiley and Sons, Chichester.

Gong, Xiaoguang, Xiao, Renbin, 2007. Research on multi-agent simulation of epidemic news spread characteristics. Journal of Artificial Societies and Social Simulation 10 (3), 1.

Magnusson Turner, L., 2010. Residential Mobility: Vacancy Chains. The International Encyclopedia of Housing and Home. Elsevier Ltd, Oxford, UK.

Rabajante, J.F., 2012. Investigating the propagation and death of information in human subpopulation networks. International Journal of Applied Mathematical Research 1, 433–451.

Stochastic Optimization

Lauren A Hannah Columbia University, New York, NY, USA

Published by Elsevier Ltd.

Abstract

In statistical science, stochastic optimization refers to a collection of methods for minimizing or maximizing an objective function when randomness exists. Over the last few decades these methods have become important tools in both academia and industry. Specific applications include: running simulations to refine the placement of acoustic sensors on a beam, deciding when to release water from a reservoir for hydroelectric power generation, and optimizing the parameters of a statistical model for a given dataset. Randomness usually enters these problems in two ways: through the cost function or the constraint set. Although stochastic optimization refers to any optimization method that employs randomness, this article only addresses settings where the objective function or constraints are random.

Introduction

Stochastic optimization refers to a collection of methods for minimizing or maximizing an objective function when randomness is present. Over the last few decades these methods have become essential tools for science, engineering, business, computer science, and statistics. Specific applications are varied, but include: running simulations to refine the placement of acoustic sensors on a beam, deciding when to release water from a reservoir for hydroelectric power generation, and optimizing the parameters of a statistical model for a given data set. Randomness usually enters the problem in two ways: through the cost function or the constraint set. Although stochastic optimization refers to any optimization method that employs randomness within some communities, we only consider settings where the objective function or constraints are random.

Like deterministic optimization, there is no single-solution method that works well for all problems. Structural assumptions, such as limits on the size of the decision and outcome spaces, or convexity, are needed to make problems tractable. Because solution methods are tied to problem structure, we divide this article according to problem type and describe the associated solution methods.

The most prominent division is between solution methods for problems with a single-time period (single-stage problems) and those with multiple time periods (multistage problems). Single-stage problems try to find a single, optimal decision, such as the best set of parameters for a statistical model given data. Multistage problems try to find an optimal sequence of decisions, such as scheduling water releases from hydroelectric plants over a 2-year period. Single-stage problems are usually solved with modified deterministic optimization methods. However, the dependence of future decisions on random outcomes makes direct modification of deterministic methods difficult in multistage problems. Multistage methods are more reliant on statistical approximation and strong assumptions about problem structure, such as finite decision and outcome spaces, or a compact Markovian representation of the decision process. The second section describes solution methods for single-stage stochastic optimization problems and the third section gives methods for sequential problems. Conclusions and areas for advancement are given in the last section.

Single-Stage Stochastic Optimization

Single-stage stochastic optimization is the study of optimization problems with a random objective function or constraints where a decision is implemented with no subsequent recourse. One example would be parameter selection for a statistical model: observations are drawn from an unknown distribution, giving a random loss for each observation. We wish to select model parameters to minimize the expected loss using data.

To generalize the problem, we begin by introducing some formal concepts and notation. Let $\mathcal{X}$ be the domain of all feasible decisions and x a specific decision. We would like to search over $\mathcal{X}$ to find a decision that minimizes a cost function, F. Let ξ denote random information that is available only after the decision is made. Unless otherwise noted, we will limit our discussion to random cost functions, denoted $F(x, \xi)$ (random constraint sets can be written as random cost functions). As we cannot directly optimize $F(x, \xi)$, we instead minimize the expected value, $\mathbb{E}[F(x,\xi)]$. The general single-stage stochastic optimization problem becomes.

$$\zeta^* = \min_{x \in \mathcal{X}} \{f(x) = \mathbb{E}[F(x,\xi)]\} \qquad [1]$$

Define the set of optima as $\mathcal{S}^* = \{x \in \mathcal{X} : f(x) = \zeta^*\}$. For all single-stage problems, we assume that the decision space $\mathcal{X}$ is convex and the objective function $F(x, \xi)$ is convex in x for any realization ξ. Problems that do not meet these assumptions are usually solved through more specialized stochastic optimization methods like the stochastic ruler (Alrefaei and Andradottir, 2001), nested partitions (Shi and Olafsson, 2000), branch and bound (Norkin et al., 1998), or tabu search (Battiti and Tecchiolli, 1996; Glover and Laguna, 1999). In this section, we begin by outlining a set of problem considerations and then review four broad solution methods problem [1]: sample average approximation (SAA), stochastic approximation, response surfaces, and metamodels.

Problem Considerations

Stochastic optimization has been studied in a broad set of communities that each developed methods to solve problems that were important to their own discipline. As such, there is no consensus about how data are generated or processed.

Each community has a set of tacit assumptions, such as observational data with exogenous noise in statistical inference problems, or generative data with endogenous noise in engineering design problems. The types of assumptions are rarely stated, but we find it helpful to discuss them and how they interact with solution methods.

Noise Generation: Exogenous versus Endogenous

The random outcome ξ is exogenous if the decision x does not affect the distribution of ξ. Likewise, noise is endogenous if the choice of x may affect the distribution of ξ. An example of exogenous noise would be returns for a portfolio of stocks: portfolio choice will not affect stock market returns. Endogenous noise can arise in problems like acoustic sensor placement on a steel beam (Sadegh and Spall, 1998): placement affects the recorded distortions.

Data Generation: Constructive versus Black Box

An observation of $F(x, \xi)$ is constructive if, for a given ξ, $F(x, \xi)$ can be evaluated for any x. This is common when F is a parametric model, like a log likelihood of parameters x given data realization ξ. Conversely, $F(x, \xi)$ is generated by a black box if $F(x, \xi)$ can only be generated for a specific x. Constructive data generation is only possible if the noise is exogenous.

Data Treatment: Batch versus Online

Solution methods have batch data evaluation if all of the observations of $(x_i, F(x_i, \xi))_{i=1}^n$ are used at once to generate a decision. An example would be maximum likelihood estimation for a parametric statistical model. Solution methods have online data evaluation if the decision is updated in a sequential manner using one observation at a time.

Data Type: Observational versus Generative

Data are observational if $(x_i, F(x_i, \xi))_{i=1}^n$ are generated a priori and the decision maker cannot generate any more. An example would be using n observations to fit a parametric statistical model. Data are generative if the decision maker can obtain samples for a new decision value, $(x_i, F(x_i, \xi_i))$, as needed. This is common in experimental settings where new observations can be generated by changing experimental parameters and rerunning the experiment. A poorly studied but common middle ground is partially observational data, where a decision maker can obtain new samples by controlling a subset of the parameters. An example would be temperature data in a building that is controlled by a thermostat set point: the set point can be changed by the decision maker, but other variables that influence the optimal decision, like occupancy level, cannot be changed.

Sample Average Approximation

SAA (Healy and Schruben, 1991; Robinson, 1996; Shapiro and Wardi, 1996; Shapiro et al., 2002) is a two-part method that uses sampling and deterministic optimization to solve eqn [1]. The first step in SAA is sampling. While directly computing the expected cost function, $\mathbb{E}[F(x, \xi)]$, is not possible for most problems, it can be approximated through Monte Carlo sampling in some situations. Let $(\xi_i)_{i=1}^n$ be a set of independent, identically distributed realizations of ξ, and let $F(x, \xi_i)$ be the cost function realization for ξ_i. The expected cost function is approximated by the average of the realizations:

$$\mathbb{E}[F(x, \xi)] \approx \frac{1}{n} \sum_{i=1}^{n} F(x, \xi_i) \qquad [2]$$

The second step in SAA is search. The right hand side of eqn [2] is deterministic, so deterministic optimization methods can be used to solve the approximate problem:

$$\zeta_n^* = \min_{x \in \mathcal{X}} \left\{ f_n(x) \frac{1}{n} \sum_{i=1}^{n} F(x, \xi_i) \right\} \qquad [3]$$

The set of approximate optima is $\mathcal{S}_n^* = \{x \in \mathcal{X} : f_n(x) = \zeta_n^*\}$.

Deterministic search is the main benefit of SAA. Many commercial software packages, including Matlab and R, offer implementation of basic deterministic optimization methods, while more specialized packages like CPLEX and Gurobi provide a wider array of deterministic methods. To guarantee convergence of the search method to a global optimum, it is assumed that $\mathcal{X}$ is convex and that $F(x, \xi_i)$ is convex in x for every realization of ξ. These assumptions can be lifted in exchange for a guarantee of only local optimality.

One of the main limitations is that SAA is only available for problems with exogenous noise and constructive data generation. It cannot accommodate any problems with black box generation. The data are processed in a batch manner. These rather strong assumptions allow the use of either observational or generative data.

Because of its inherent simplicity, SAA has been independently proposed in a variety of fields under a variety of names. A form of SAA was derived in the early statistics community as maximum likelihood estimation (Fisher, 1922), although it was not generalized to other types of stochastic optimization problems. SAA-type methods were generalized in the operations research community under a variety of names: sample path optimization (Robinson, 1996; Shapiro and Wardi, 1996; Gürkan et al., 1999), stochastic countermethods (Rubinstein and Shapiro, 1993), retrospective optimization (Healy and Schruben, 1991), scenario optimization (Birge and Louveaux, 1997), and SAA (Shapiro et al., 2002; Kleywegt et al., 2002). Applications for SAA include: chance constrained optimization (Pagnoncelli et al., 2009), discrete stochastic optimization (Kleywegt et al., 2002), stochastic routing (Verweij et al., 2003), queuing models (Atlason et al., 2004), asset allocation (Blomvall and Shapiro, 2006), and solving (Partially Observable) Markov Decision Processes ((PO)MDPs) (Ng and Jordan, 2000; Perkins, 2002).

Theoretical Properties of SAA Estimates

Like all stochastic optimization methods, SAA relies upon a collection of random variables to produce a statistical estimate. Most theoretical results follow directly from the Law of Large Numbers and Central Limit Theorem due to the construction of SAA. Under mild regularity conditions, for any fixed x the limiting distribution of the SAA estimate is Gaussian with mean $f(x)$ and variance $\sigma^2(x) = \mathrm{Var}(F(x, \xi))/n < \infty$:

$$n^{\frac{1}{2}}\left[f_n(x) - f(x)\right] \xrightarrow{d} N\left(0, \sigma^2(x)\right)$$

The symbol $\xrightarrow{d}$ denotes 'convergence in distribution.' Under more stringent conditions, including Lipschitz continuity of $F(\cdot, \xi)$, convexity of $F(x, \xi)$ and $\mathcal{X}$, and $f(x)$ having a unique optimum $S^* = \{x^*\}$, a similar result holds for the optimal values:

$$n^{\frac{1}{2}}\left[\zeta_n^* - \zeta^*\right] \xrightarrow{d} N(0, \sigma^2(x^*)) \qquad [4]$$

The limiting Gaussian distribution in eqn [4] can be used to determine the number of samples needed to generate an $\varepsilon-$optimal solution with at least probability $1 - \alpha$. An $\varepsilon-$optimal solution is defined as any solution $\tilde{x} \in \mathcal{X}$ such that $f(\tilde{x}) \leq \zeta^* + \varepsilon$ for a specified $\varepsilon > 0$. However, there is a tradeoff between the accuracy of f_n and the time needed to search f_n. In the general case, with efficient search algorithms the additional computational complexity is roughly on the order of n (Shapiro, 2013). See Shapiro (1991), Rubinstein and Melamed (1998), and Shapiro et al. (2009) for a more complete discussion of SAA consistency and convergence rates.

Stochastic Approximation

Stochastic approximation is an iterative method that uses noisy observations to find the root of a function, g, i.e., the set of x where $g(x) = 0$ (Robbins and Monro, 1951). When that function is the gradient of the expected cost function, $g(x) = \nabla_x f(x)$, stochastic approximation finds local optima of f for an unconstrained decision set using noisy subgradient observations, $g(x_i, \xi_i)$. In a generalization of stochastic gradient descent, a decision x_i is updated recursively by moving a distance a_i in the direction opposite the gradient:

$$x_{i+1} = x_i - a_i g(x_i, \xi_i) \qquad [5]$$

A constraint set, $\mathcal{X}$, can be included through a projection operator $\Pi_{\mathcal{X}}$ that maps $x_i - a_i g(x_i, \xi_i)$ into the closest point of $\mathcal{X}$. The main considerations for eqn [5] are selecting a step size sequence $(a_i)_{i=1}^{\infty}$ and obtaining a gradient $g(x_i, \xi_i)$ through direct observation or estimation.

Stochastic approximation has a less restrictive set of assumptions than SAA: data generation may be either constructive or black box, and noise generation may be either endogenous or exogenous. If the data have exogenous noise and constructive generation, then observational data may be used. If those conditions are not met, generative data are required. Finally, stochastic approximation uses online data processing. This has made stochastic approximation one of the most prominent methods for large-scale statistical inference; see Bottou (2010), Hoffman et al. (2010), and Agarwal et al. (2012) for examples. For n observations, SAA requires n evaluations of the derivative of a log-likelihood function per iteration of the deterministic solver, which can become unreasonably expensive. In extremely large-scale inference, it may even be problematic to store n observations in RAM. Stochastic approximation requires only n evaluations of the derivative of the log-likelihood function and allows data streaming from an external storage device.

Gradient Approximation

Gradients can be computed through either direct observation or statistical estimation. Direct computations are usually possible when the loss function has a closed form, as in case of parametric statistical inference. In contrast, the gradient must be estimated when the data are generated by a black box. Finite difference (Kiefer and Wolfowitz, 1952) and related methods (Spall, 1992) approximate the gradient by using the current decision, x_t, as a starting point and adding a perturbation, $e_j c_i$. Here e_j denotes a vector of all 0's with a 1 in the jth location and $(c_i)_{i=1}^{\infty}$ is a sequence of perturbation sizes that decreases to 0. The jth entry of the gradient is approximated by:

$$[g(x_i, \xi_i)]_j = \frac{F(x_i + c_i e_j, \xi_{i,1}) - F(x_i - c_i e_j, \xi_{i,2})}{2c_i} \qquad [6]$$

Estimates from eqn [6] can be made more stable by using minibatches. Full finite difference computations may become infeasible when $F(x_i, \xi_i)$ is computationally expensive or p is large. Alternatives include one-sided differences, requiring $p + 1$ evaluations, and simultaneous perturbation analysis (Spall, 1992), which generates an estimate by adding and subtracting a random perturbation vector to get an estimate with two observations.

Step Size Selection

The step size sequences $(a_i)_{i=1}^{\infty}$ and $(c_i)_{i=1}^{\infty}$ affect whether the stochastic approximation estimator converges to a locally optimal solution and the rate at which it converges. In general, convergence for stochastic approximation with a directly computed gradient is guaranteed if

$$\sum_{i=1}^{\infty} a_i = \infty, \quad \sum_{i=1}^{\infty} a_i^2 < \infty, \quad a_i > 0 \qquad [7]$$

Generally, $(a_i)_{i=1}^{\infty}$ is chosen such that $a_i = Ci^{-\alpha}$ for $\alpha \in (½, 1)$ and a positive constant C. There is a tradeoff between the rate of convergence and robustness of the estimator: $\alpha = 1$ generally promises the best asymptotic rate in many circumstances, but is sensitive to misspecification of C or nonstrong convexity of f. If second order information is available, a_i can be replaced by an approximation of the inverse of the Hessian of f (Bottou, 2010).

When gradients are estimated, the perturbation sequence $(c_i)_{i=1}^{\infty}$ needs to be chosen in conjunction with $(a_i)_{i=1}^{\infty}$. Traditional constraints require that in addition to eqn [7], the sequences satisfy (Kiefer and Wolfowitz, 1952):

$$\sum_{i=1}^{\infty} \frac{a_i^2}{c_i^2} < \infty, \quad \sum_{i=1}^{\infty} a_i c_i < \infty, \quad c_i > 0$$

However, slightly looser constraints do exist (Broadie et al., 2011). Parameters are commonly set as polynomial sequences with $a_i = Ci^{-1}$ and $c_i = Di^{-1/4}$ for positive constants C and D.

Theoretical Properties of Stochastic Approximation

Convergence theory for stochastic approximation generalizes convergence theory for gradient descent algorithms. Rates depend on the shape of f and whether the gradients are exact or estimated. Common shape assumptions are smoothness – whether f is Lipschitz and has Lipschitz first order derivatives – and strong convexity:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq m\|x - y\|_2^2$$

for all $x, y \in \mathcal{X}$, where $m > 0$ is a curvature constant. In all instances, we use n iterations and assume a few mild regularity conditions. If the gradients are not estimated, f is smooth and strongly convex, and $a_i = Ci^{-1}$, then the convergence rate is $\mathcal{O}(n^{-1})$:

$$\mathbb{E}[f(x_n) - f(x^*)] \leq Kn^{-1}$$

for a constant $K > 0$ if $mC > 2$. However, if $mC < 2$, then an arbitrarily poor rate $\mathcal{O}(n^{-mC/2})$ is obtained (Nemirovski et al., 2009; Bach and Moulines, 2011). If the gradient is estimated, rates of $\mathcal{O}(n^{-1/2})$ can be obtained using $a_i = Ci^{-1}$ and $c_i = Di^{-1/4}$ (Broadie et al., 2011). When f is nonstrongly convex or nonsmooth, the best general convergence rate is $\mathcal{O}(n^{-1/2})$ (Bach and Moulines, 2011).

Stochastic approximation can be made more robust to parameter misspecification through a combination of step sizes larger than $\mathcal{O}(n^{-1})$ and iterate averaging (Polyak and Juditsky, 1992; Ruppert, 1988; Nemirovski et al., 2009). The most common averaging scheme is Polayk-Ruppert averaging:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, \ldots, x_n$ are generated through the usual stochastic approximation recursion. For a strongly convex, smooth objective function, a step size $a_i = Ci^{-\alpha}$ with $\alpha \in [½, 1]$ will yield rates of $\mathcal{O}(n^{-1})$ using iterate averaging (Bach and Moulines, 2011). For convex, nonsmooth or nonstrongly convex objective functions, iterate averaging will give a rate of $\mathcal{O}(n^{-1/2})$ for $\alpha = 1/2$ (Nemirovski et al., 2009). In general, using $\alpha = 1/2$ and iterate averaging is a good choice because it is adaptive to strong convexity while being robust to nonsmoothness and nonstrong convexity.

Response Surfaces

Response surface methodology (RSM) is a collection of methods that fit a surface to a set of decision–response pairs and then search the surface to find a new decision. Traditionally, RSM refers to online methods for decision making: a collection of decision–response pairs are sampled around a current decision, a 'response surface' is fitted locally to the observations, and then decision is updated by searching the local surface. Some authors have placed batch methods under the umbrella of RSM (Downing et al., 1985; Lim, 2010), but we categorize these as global metamodels. RSM originated as methodology for optimizing conditions for chemical investigation in the 1950s (Box and Wilson, 1951), and has since been widely applied in experimental design, process optimization, and simulation optimization. Recent surveys and books include Barton and Meckesheimer (2006), del Castillo (2007), Myers et al. (2009), and Barton (2013).

RSM proceeds in a manner similar to stochastic approximation. A starting point, x_0, is selected. Then configurations around this point are sampled and used to fit a first or second order polynomial (Box and Draper, 1987). Expected values in a local area around the current decision x_i are estimated by:

$$\begin{aligned} \widehat{y}(x_i) &= \beta_0 + \beta_{1:p}^T x_i && \text{(linear)} \\ \widehat{y}(x_i) &= \beta_0 + \beta_{1:p}^T x_i + x_i^T \mathbf{B} x_i. && \text{(quadratic)} \end{aligned}$$

Typically, a linear model is used in earlier iterations while a quadratic model is used in later ones. Sampling design near x_i is often based on experimental design literature for parametric models (Montgomery, 2012) and model fit is evaluated with standard parametric statistical tools (Kleijnen and Sargent, 2000). The decision is updated by moving from the current decision in the direction of steepest descent. For a locally linear model, this is

$$x_{i+1} = x_i - a_i \beta_{1:p}$$

The step size a_i is often set as the quasi-Newton step size with the local surface as the objective function (Kleijnen, 2008) or through a line search (Barton and Meckesheimer, 2006). Once the estimated slopes get close to zero, a second order model is used to check optimality conditions through hypothesis testing.

Response surfaces have the same data generating requirements as stochastic approximation: data generation may be either constructive or black box, and noise generation may be either endogenous or exogenous. If the data have exogenous noise and constructive generation, then observational data may be used; otherwise, generative data are required.

Global Metamodel Optimization

Metamodels relate a set of inputs to an expected output through regression. Response surfaces are one form of metamodel, as are the Monte Carlo objective function approximations formed by SAA. In contrast to local response surfaces, global metamodels define a relationship that is valid across all feasible input values. A variety of regression methods have been used to fit global metamodels, including radial basis functions (Shin et al., 2002), neural networks (Anjum et al., 1997), Gaussian processes/stochastic kriging (Ankenman et al., 2010), and shape constrained nonparametric regression (Lim, 2010; Hannah et al., 2013). Metamodels have been used for optimization for both batch and online problems.

In the batch setting, global metamodel optimization is done through a two-step process much like SAA. Before any modeling is done, n decision–response pairs, $(x_i, y_i)_{i=1}^n$ where $y_i = F(x_i, \xi_i)$, are collected. Then a metamodel, $f_n : \mathcal{X} \to \mathbb{R}$, is fit to $(x_i, y_i)_{i=1}^n$ through regression. Finally, the decision is generated by solving

$$\zeta_n^* = \min_{x \in \mathcal{X}} f_n(x)$$

Batch global metamodel optimization is primarily useful for inferring a decision from observational data generated by a black box with either endogenous or exogenous noise. Note that most regression methods, like neural networks, basis function regression, and Gaussian processes, are not shape-constrained: f_n is not guaranteed to be convex and must either be optimized locally or through global optimization methods. Recently, there has been some work on using shape-constrained regression to maintain metamodel convexity (Lim, 2010; Hannah et al., 2013).

Global metamodel optimization has recently gained popularity for online optimization of functions that are costly to evaluate, such as hyperparameter tuning for Bayesian

models (Scott et al., 2011; Snoek et al., 2012). The metamodels used are generative models, like Gaussian processes, which give a distribution over possible expected value functions. Design locations are then selected recursively to minimize an error metric, such as the expected minimal value given the new design location. After each design location is selected, the metamodel is updated. Selected design locations generally try to balance gaining knowledge about a region with high uncertainty against tuning decisions in an area that is known to perform well. Online global metamodels require generative data.

Along with potential search difficulties, global metamodel optimization is limited by the amount of data for the regression step. Nonparametric metamodels require that n grows exponentially with the dimension of $\mathcal{X}$. This issue can be bypassed by using models with strong assumptions, like parametric relationships.

Multistage Stochastic Optimization

Multistage stochastic optimization problems aim to find a sequence of decisions, $(x_t)_{t=0}^{T}$, that minimize an expected cost function. The subscript t denotes the time at which decision x_t is made. Usually decisions and random outcomes at time t affect the value of future decisions. An example would be making a move in a chess game. With a move, the player may capture one of his or her opponent's pieces, change his or her board position, and alter his or her possible future moves. He or she needs to account for these issues to select the move that maximizes his or her probability of winning. Mathematically, we can describe multistage stochastic optimization problems as an iterated expectation:

$$\begin{aligned}\zeta^* = \min_{x_0\in\mathcal{X}_0} \mathbb{E}\Big[&\inf_{x_1\in\mathcal{X}_1(x_0,\xi_1)} F_1(x_1,\xi_1) \\ &+ \mathbb{E}\Big[\cdots + \mathbb{E}\Big[\inf_{x_T\in\mathcal{X}_T(x_{0:T-1},\xi_{1:T})} \gamma^{T-1}F_T(x_T,\xi_T)\Big]\Big]\Big]\end{aligned} \quad [8]$$

Here T is the number of time periods; $x_{0:t}$ is the collection of all decisions between 0 and t; ξ_t is a random outcome observable at time t; $\mathcal{X}_t(x_{0:t-1},\xi_{1:t})$ is a decision set that depends on all decisions and random outcomes between times 0 and t; $F_t(x_t, \xi_t)$ is a cost function for time period t that depends on the decision and random outcome for period t; and γ is the discount rate. The time horizon T may be either finite or infinite.

Unlike the methods outlined in Section Single Stage Stochastic Optimization, there are no multistage solution methods that work well for all problems within a broad class, like convex problems or Markov decision processes. The decision sequence space is affected by the curse of dimensionality: the size of the space grows exponentially with T, the number of possible outcomes for ξ_t, and the size of the decision space each time period, $\mathcal{X}_t$. Most successful methods are tailored to problem subclasses with exploitable structure, including bandit problems and Markov decision processes (MDPs). In this section, we break describe solution methods according to problem classes: bandit problems, MDPs, and convex multistage problems.

Bandit Problems

Bandit problems are sequential decision problems where each time period a decision maker has a choice of playing one of K slot machines, or 'one-armed bandits.' Each machine has a payoff that can take values zero or one according to a Bernoulli distribution with an unknown, arm-dependent probability of success p_k. The goal of the decision maker is to maximize his or her expected revenue or to minimize regret after T time periods. The decision $x_{t-1} \in \{1,\ldots, K\}$ is the selection of the arm to be played at time t, the random outcome ξ_t is a Bernoulli random variable with parameter $p_{x_{t-1}}$. The cost functions in eqn [8] for maximizing expected rewards and minimizing regret are as follows:

$$\begin{aligned} F_t(x_{t-1},\xi_t) &= -\xi_t && \text{(expected reward)} \\ F_t(x_{t-1},\xi_t) &= \mathbb{E}[\xi_t] - \max_{k=1,\ldots,K} p_k && \text{(regret)}\end{aligned}$$

Difficulty comes from the tradeoff between exploiting machines that have been observed to have high payouts and exploring new machines that have an uncertain payout probability. Solution methods can be divided by the size of K. When K is small each arm can be tried many times. When K is large the decision maker needs to borrow information from similar arms to estimate p_k, usually through regression. All bandit problems are online and assume generative data from a black box with endogenous noise.

Small K methods rely on computing a summary statistic for each arm, such as a mean value, upper confidence bounds (UCB) for p_k (Lai and Robbins, 1985; Audibert et al., 2009), knowledge gradient (Frazier et al., 2008), or Gittins index (Gittins, 1979). Once these have been computed, the arm with the highest value according to the statistic is played in the next round. These policies often result in asymptotic optimality.

When K is large, the arm space is often treated as a continuous space, $\mathcal{X}$. An arm mean function, $g(x)$, is assumed to have structure, often linearity, Lipschitz continuity, or a local Lipschitz continuity. The ellipsoid method creates an ellipse that has high probability of containing the optimal solution, which works well with linear mean functions (Dani et al., 2008). UCBs can be generated for problems with either Lipschitz or locally Lipschitz mean functions by using hierarchical models (Bubeck et al., 2011) or Gaussian processes (Srinivas et al., 2010).

Bandit problems can be extended to include 'side information,' which is an observable random variable that gives information about p_k. For instance, if p_k is an average click-through rate for an online advertisement, side information may be the advertiser, the user click history, zip code, etc. Bandit problems with side information are called 'contextual bandits.' Solution methods approximate the value of the arms as a function of the side information using regression methods like linear combinations of basis functions (Li et al., 2010), discretization of the state space (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013), random histograms (Yang and Zhu, 2002), nearest neighbors (Yang and Zhu, 2002) or adaptive partitioning (Slivkins, 2011). While all bandit problems are broadly applicable to many e-commerce problems like advertising and page layout, contextual bandits are particularly useful models for online advertising when there is information about the user and/or the advertisement.

Markov Decision Processes

MDPs simplify eqn [8] by assuming that the information required to make a decision at time t can be contained within a state, s_t. Naturally, this formulation includes all sequential decision processes the state is defined as the entirety of the decision and outcome history. However, MDPs are usually used to model problems where the state can take a small number of values that recur over time periods. For example, a state may be the position of a robot within a room where the goal is to navigate to a specific location while avoiding obstacles. Previous decisions, $x_{0:t}$, and random outcomes, $\xi_{1:t}$, are also summarized by the state. Due to the Markov nature of the problem, the cost function F_t is modified by removing the time index and including a dependence on the state to produce $F(x,s)$. The probability of transitioning from state s to state s' with decision x is $P_x(s,s')$. Traditionally, the time horizon is infinite. This allows optimality conditions to be expressed in terms of a value function $V(s)$, defined as the value of being in state s given that all subsequent decisions are optimal. The conditions are given by the Bellman optimality equation (Bellman, 1961):

$$\begin{aligned} V(s) &= \min_{x\in\mathcal{X}(s)} F(x,s) + \gamma \sum_{s'} P_x(s,s')V(s') \\ &= \min_{x\in\mathcal{X}(s)} F(x,s) + \gamma \mathbb{E}[V(s')|s,x] \end{aligned} \quad [9]$$

As eqn. [9] is recursive, it is convenient to use value functions that are generated by a policy π, which maps each state to a decision:

$$V^{\pi}(s) = F(x,s) + \gamma \mathbb{E}[V^{\pi}(s')|s,x] \quad [10]$$

where x is chosen according to a policy π.

Solution methods for MDPs usually involve two steps: estimating value functions for a given policy π, and using then the estimated value functions to find a better policy π',

$$\pi'(s) = \arg\min_{x\in\mathcal{X}(s)} F(x,s) + \gamma \mathbb{E}[V^{\pi}(s')|s,x] \quad [11]$$

Two of the most common solution methods are policy iteration and value iteration. Policy iteration selects a policy through eqn [11], and then updates value function estimates in eqn [10] using that policy. Value iteration combines the policy search and approximation steps by repeatedly solving:

$$\overline{V}_{i+1}(s) = \min_{x\in\mathcal{X}(s)} F(x,s) + \gamma \mathbb{E}[\overline{V}_i(s')|s,x] \quad [12]$$

where $\overline{V}_i$ is the estimated function after i iterations. See Sutton and Barto (1998) for a summary of these methods. Other solution techniques include rollout methods, which generate sample paths a few time periods into the future (Ng and Jordan, 2000), Q-learning, which fits a regression model to state–decision pairs (Watkins and Dayan, 1992), and dynamic programming, which is a computational method to solve finite horizon, discrete state, discrete decision problems (Sutton and Barto, 1998). MDPs usually assume generative, black box data with endogenous noise. Most solution methods are online, although some methods like fitted-Q iteration (Murphy, 2003; Antos et al., 2007) accommodate observational data in a batch setting.

MDP solution methods have many open research problems, including value function approximation, value attribution, and search over the decision space. Value function approximation is a form of regression, and it suffers the same problems as regression with regard to data sparsity in high dimensional state spaces. Proposed approximation methods have included lookup tables for small state spaces, linear combinations of a set of basis functions (Lagoudakis and Parr, 2003), nonparametric functions (Engel et al., 2003), or other problem specific representations that include hierarchical structure or convexity (Powell, 2011). Additionally, attributing value to actions when MDP has a single payout after a long string of decisions, such as winning a table tennis game, is notoriously difficult. Finally, all solution methods require a search over the decision space, as in eqns [11] and [12]. This is trivial when the decision space is finite and small, but can become difficult or impossible when the space is continuous and multidimensional.

Convex Multistage Problems

As in single-stage optimization, convexity of the cost function and decision space allows decision makers to leverage deterministic convex optimization methods. Convexity occurs in many operations research problems like resource allocation, inventory management, and portfolio management. We define convex multistage problems with respect to eqn [8]:

$F_t(x_t, \xi_t)$ is convex in x_t for each ξ_t and $\mathcal{X}_t(x_{0:t-1}, \xi_{1:t})$ is convex for all $x_{0:t-1}$ and $\xi_{1:t}$. In almost all situations the time horizon T is finite and relatively small. Solution methods include stochastic programming and approximate dynamic programming (ADP).

Stochastic programming solves eqn [8] by approximating it with a deterministic mathematical program, similar to SAA. One important but often unstated assumption is exogenous noise. It is described with a scenario tree. When n scenarios, $\omega_1, \ldots, \omega_n$, are expressed as a tree, each scenario is a path from the current state to a leaf of the tree, $\{\xi_1(\omega_i), \ldots, \xi_T(\omega_i)\}$. Scenario trees are often constructed through moment matching (Høyland and Wallace, 2001) or Monte Carlo methods (Blomvall and Shapiro, 2006), and each scenario can have nonuniform probability weights, $\{p_1(\omega_i), \ldots, p_T(\omega_i)\}_{i=1}^n$. Decisions are also indexed by scenario, $x = \{x_0(\omega_i), \ldots, x_T(\omega_i)\}_{i=1}^n$. The objective function is a linear combination of the costs for each of n scenarios:

$$f(x_{0:T}(\omega_1), \ldots, x_{0:T}(\omega_n)) = \sum_{i=1}^{n} \sum_{t=1}^{T} p_t(\omega_i) F_t(x_t(\omega_i), \xi_t(\omega_i)) \quad [13]$$

As with SAA, this is a batch method using observational data and constructive observed costs. The tree structure is maintained through nonanticipativity constraints: if ω_i and ω_j share a node at time t, then $x_t(\omega_i) = x_t(\omega_j)$. Each decision also needs to satisfy the original constraints for a given scenario, $x_t(\omega_i) \in \mathcal{X}_t(x_{0:t-1}(\omega_i), \xi_{1:t}(\omega_i))$. A solution is found by minimizing eqn [13] subject to the original and nonanticipativity constraints, often using dual decomposition methods (Ruszczyński, 1997).

Stochastic programming methods have been successfully implemented for short and medium term scheduling problems, including reservoir management (Tejada-Guibert et al., 1995), unit commitment for electricity generation (Takriti

et al., 2000), and asset management (Carino et al., 1994). Most computational difficulties come from large-scenario trees, which occur when there are long time horizons, finer time scales, or outcome spaces that are not well described by a sparse tree.

ADP uses ideas from MDPs to solve eqn [8], but it often has a finite time horizon and imposes shape constraints to allow search over continuous decision spaces. As in the MDP setting, the relevant historical decision and random outcome information is stored in a state variable, $(S_t)_{t=0}^{T}$, and the methods have the same data assumptions. ADP estimates a set of value functions to satisfy:

$$V_t(S_t) = \arg\min_{x_t \in \mathcal{X}_t(S_t)} \{F_t(x_t, \xi_t) + \gamma \mathbb{E}[V(S_{t+1})|S_t, x_t]\} \qquad [14]$$

As long as a convex approximation of $V_t(S_t)$ with respect to x_{t-1} is maintained, the convexity assumption allows efficient search over the decision space (Powell, 2011). The value functions are often fit through methods similar to value iteration or policy iteration. While ADP can scale to much larger problems than stochastic programming (Anderson et al., 2011), approximating value functions is a nontrivial statistical challenge: state variables can be high dimensional, simulation outputs can be limited, and shape constrained inference is difficult in multiple dimensions (Hannah and Dunson, 2013). Approximation methods have included basis functions linear projections (Tsitsiklis and Van Roy, 2001; de Farias and Van Roy, 2003; Keller et al., 2006), Benders cuts (Higle and Sen, 1991; Zakeri et al., 2000), and nonparametric regression (George et al., 2008; Deisenroth et al., 2009; Taylor and Parr, 2009). Consistency of ADP has not been shown in a general setting, although it has been demonstrated for certain cases (Nascimento and Powell, 2009).

Future Challenges and Closing Remarks

Due to the swift progress of data storage and analysis capabilities, stochastic optimization has been an area of extremely active research. Summaries of progress within specific fields are given by Fu (2013), Hutchison and Spall (2013), and Powell (2014). However, many important areas lacking solutions include the following.

Efficient Methods for Black Box, Observational Data

Observational data are currently handled through a global regression over the decision space, and problems with partially observational data receive little study. Methods that exploit convexity could drastically reduce data requirements and provide more reliable search.

Efficient Methods for Large Multistage Problems

While much attention has been given to value function approximation, discovering new problem subclasses with exploitable structure may prove more useful. Possible areas include the interplay between large decisions, such as 3-year operating goals for a hydroelectric plant, and small decisions, such as hourly water releases.

Solution Evaluation Methods

Asymptotic theory provides estimates on the quality of single-stage solutions, but in the multistage case asymptotic bounds either do not exist or they only provide pessimistic bounds. Some work in the operations research literature uses information relaxations to provide dual bounds (Haugh and Kogan, 2004; Brown et al., 2010), but these are computationally infeasible for general problems. Fast, reliable bounds would be useful for practitioners to assess decision quality before implementation.

Optimization in the presence of randomness is a fundamental problem in many fields. We fully expect that current methods for stochastic optimization will be refined over the next few decades while computational advances and new problems will lead to new opportunities for stochastic optimization research.

See also: Distributions, Statistical: Approximations; Probability Theory: Formal.

Bibliography

Agarwal, A., Negahban, S., Wainwright, M.J., 2012. Fast global convergence of gradient methods for high-dimensional statistical recovery. The Annals of Statistics 40 (5), 2452–2482.

Alrefaei, M.H., Andradottir, S., 2001. A modification of the stochastic ruler method for discrete stochastic optimization. European Journal of Operational Research 133 (1), 160–182.

Anderson, R.N., Boulanger, A., Powell, W.B., Scott, W., 2011. Adaptive stochastic control for the smart grid. Proceedings of the IEEE 99 (6), 1098–1115.

Anjum, M.F., Tasadduq, I., Al-Sultan, K., 1997. Response surface methodology: a neural network approach. European Journal of Operational Research 101 (1), 65–73.

Ankenman, B., Nelson, B.L., Staum, J., 2010. Stochastic kriging for simulation metamodeling. Operations Research 58 (2), 371–382.

Antos, A., Munos, R., Szepesvári, C., 2007. Fitted Q-iteration in continuous action-space MDPs. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20, Vancouver, BC, pp. 9–16.

Atlason, J., Epelman, M.A., Henderson, S.G., 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127 (1–4), 333–358.

Audibert, J.-Y., Munos, R., Szepesvári, C., 2009. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science 410 (19), 1876–1902.

Bach, F., Moulines, E., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems 24, Granada, Spain, pp. 451–459.

Barton, R.R., 2013. Response surface methodology. In: Gass, S.I., Fu, M. (Eds.), Encyclopedia of Operations Research and Management Science, third ed. Springer, New York.

Barton, R.R., Meckesheimer, M., 2006. Metamodel-based simulation optimization. In: Henderson, S.G., Nelson, B.L. (Eds.), Simulation, Handbooks in Operations Research and Management Science, vol. 13. Elsevier, pp. 535–574.

Battiti, R., Tecchiolli, G., 1996. The continuous reactive tabu search: blending combinatorial optimization and stochastic search for global optimization. Annals of Operations Research 63 (2), 151–188.

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour, vol. 4. Princeton University Press, Princeton, NJ.

Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming, second ed. Springer Verlag, New York.

Blomvall, J., Shapiro, A., 2006. Solving multistage asset investment problems by the sample average approximation method. Mathematical Programming 108 (2), 571–595.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (Eds.), Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT 2010). Springer, Paris, France, pp. 177–187. http://leon.bottou.org/papers/bottou-2010.

Box, G.E.P., Draper, N.R., 1987. Empirical Model-Building and Response Surfaces. John Wiley & Sons, Hoboken, NJ.

Box, G.E.P., Wilson, K.B., 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B 13 (1), 1–45.

Broadie, M., Cicek, D., Zeevi, A., 2011. General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm. Operations Research 59 (5), 1211–1224.

Brown, D., Smith, J., Sun, P., 2010. Information relaxations and duality in stochastic dynamic programs. Operations Research 58 (4), 785–801.

Bubeck, S., Munos, R., Stoltz, G., Szepesvari, C., 2011. X-armed bandits. The Journal of Machine Learning Research 12, 1655–1695.

Carino, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T., 1994. The Russell-Yasuda Kasai model: an asset/liability model for a Japanese insurance company using multistage stochastic programming. Interfaces 24 (1), 29–49.

del Castillo, E., 2007. Process Optimization: A Statistical Approach. Springer, New York.

Dani, V., Hayes, T.P., Kakade, S.M., 2008. Stochastic linear optimization under bandit feedback. In: Servedio, R., Zhang, T. (Eds.), Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, pp. 355–366.

Deisenroth, M.P., Rasmussen, C.E., Peters, J., 2009. Gaussian process dynamic programming. Neurocomputing 72 (7), 1508–1524.

Downing, D.J., Gardner, R.H., Hoffman, F.O., 1985. An examination of response-surface methodologies for uncertainty analysis in assessment models. Technometrics 27 (2), 151–163.

Engel, Y., Mannor, S., Meir, R., 2003. Bayes meets Bellman: the Gaussian process approach to temporal difference learning. In: Fawcett, T., Mishra, N. (Eds.), Proceedings of the 20th International Conference on Machine Learning, Washington, DC, pp. 154–161.

de Farias, D.P., Van Roy, B., 2003. The linear programming approach to approximate dynamic programming. Operations Research 51 (6), 850–865.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222, 309–368.

Frazier, P.I., Powell, W.B., Dayanik, S., 2008. A knowledge-gradient policy for sequential information collection. SIAM Journal on Control and Optimization 47 (5), 2410–2439.

Fu, M.C., 2013. Simulation optimization. In: Gass, S.I., Fu, M. (Eds.), Encyclopedia of Operations Research and Management Science, third ed. Springer, New York.

Gass, S.I., Fu, M. (Eds.), 2013. Encyclopedia of Operations Research and Management Science. Springer, New York.

George, A., Powell, W.B., Kualkarni, S.R., 2008. Value function approximation using multiple aggregation for multiattribute resource management. Journal of Machine Learning Research 9 (10), 2079–2111.

Gittins, J.C., 1979. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society: Series B 41 (2), 148–177.

Glover, F., Laguna, M., 1999. Tabu Search. Springer, New York.

Gürkan, G., Yonca Özge, A., Robinson, S.M., 1999. Sample-path solution of stochastic variational inequalities. Mathematical Programming 84 (2), 313–333.

Hannah, L.A., Dunson, D.B., 2013. Multivariate convex regression with adaptive partitioning. Journal of Machine Learning Research 14, 3207–3240.

Hannah, L.A., Powell, W.B., Dunson, D.B., 2013. Semi-convex regression for metamodeling-based optimization. SIAM Journal on Optimization to appear.

Haugh, M.B., Kogan, L., 2004. Pricing American options: a duality approach. Operations Research 52 (2), 258–270.

Healy, K., Schruben, L.W., 1991. Retrospective simulation response optimization. In: Proceedings of the 1991 Winter Simulation Conference, pp. 901–906.

Higle, J., Sen, S., 1991. Stochastic decomposition: an algorithm for two-stage linear programs with recourse. Mathematics of Operations Research 16 (3), 650–669.

Hoffman, M., Bach, F.R., Blei, D.M., 2010. Online learning for latent Dirichlet allocation. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23, Vancouver, BC, pp. 856–864.

Høyland, K., Wallace, S., 2001. Generating scenario trees for multistage decision problems. Management Science 47 (2), 295–307.

Hutchison, D.W., Spall, J.C., 2013. Stochastic approximation. In: Gass, S.I., Fu, M. (Eds.), Encyclopedia of Operations Research and Management Science, third ed. Springer, New York.

Keller, P.W., Mannor, S., Precup, D., 2006. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In: Cohen, W., Moore, A. (Eds.), Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, pp. 449–456.

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23 (3), 462–466.

Kleijnen, J.P.C., 2008. Response surface methodology for constrained simulation optimization: an overview. Simulation Modelling Practice and Theory 16 (1), 50–64.

Kleijnen, J.P.C., Sargent, R.G., 2000. A methodology for fitting and validating metamodels in simulation. European Journal of Operational Research 120 (1), 14–29.

Kleywegt, A.J., Shapiro, A., Homem-de Mello, T., 2002. The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization 12 (2), 479–502.

Lagoudakis, M.G., Parr, R., 2003. Least-squares policy iteration. The Journal of Machine Learning Research 4, 1107–1149.

Lai, T.L., Robbins, H., 1985. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics 6 (1), 4–22.

Li, L., Chu, W., Langford, J., Schapire, R., 2010. A contextual-bandit approach to personalized news article recommendation. In: Rappa, M., Jones, P., Freire, J., Chakrabarti, S. (Eds.), Proceedings of the 19th International Conference on the World Wide Web, Raleigh, NC, pp. 661–670.

Lim, E., 2010. Response surface computation via simulation in the presence of convexity constraints. In: Johansson, B., Jain, S., Montoya-Torres, J., Hugan, J., Yücesan, E. (Eds.), Proceedings of the 2010 Winter Simulation Conference, Baltimore, MD, pp. 1246–1254.

Montgomery, D.C., 2012. Design and Analysis of Experiments, eighth ed. Wiley, Hoboken, NJ.

Murphy, S.A., 2003. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B 65 (2), 331–355.

Myers, R.H., Montgomery, D.C., Anderson-Cook, C.M., 2009. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley Sons, Hoboken, NJ.

Nascimento, J., Powell, W., 2009. An optimal approximate dynamic programming algorithm for the lagged asset acquisition problem. Mathematics of Operations Research 34 (1), 210–237.

Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A., 2009. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization 19 (4), 1574–1609.

Ng, A.Y., Jordan, M., 2000. PEGASUS: a policy search method for large MDPs and POMDPs. In: Laskey, K.B., Boutilier, C., Goldszmidt, M. (Eds.), Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, Stanford, CA, pp. 406–415.

Norkin, V.I., Pflug, G.C., Ruszczynski, A., 1998. A branch and bound method for stochastic global optimization. Mathematical Programming 83 (1–3), 425–450.

Pagnoncelli, B.K., Ahmed, S., Shapiro, A., 2009. Sample average approximation method for chance constrained programming: theory and applications. Journal of Optimization Theory and Applications 142 (2), 399–416.

Perchet, V., Rigollet, P., 2013. The multi-armed bandit problem with covariates. The Annals of Statistics 41 (2), 693–721.

Perkins, T.J., 2002. Reinforcement learning for POMDPs based on action values and stochastic optimization. In: Dechter, R., Kearns, M., Sutton, R. (Eds.), Proceedings of the 18th National Conference on Artificial Intelligence, Edmonton, Alberta, pp. 199–204.

Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30 (4), 838–855.

Powell, W.B., 2011. Approximate Dynamic Programming: Solving the Curses of Dimensionality, second ed. John Wiley Sons, Hoboken, NJ.

Powell, W.B., 2014. Energy and Uncertainty: Models and Algorithms for Complex Energy Systems. AI Magazine p. to appear.

Rigollet, P., Zeevi, A., 2010. Nonparametric bandits with covariates. In: Kalai, A.T., Mohri, M. (Eds.), Proceedings of the 23rd Annual Conference on Learning Theory, Haifa, Israel, pp. 54–66.

Robbins, H., Monro, S., 1951. A stochastic approximation method. The Annals of Mathematical Statistics 22 (3), 400–407.

Robinson, S.M., 1996. Analysis of sample-path optimization. Mathematics of Operations Research 21 (3), 513–528.

Rubinstein, R.Y., Melamed, B., 1998. Modern Simulation and Modeling. Wiley, Hoboken, NJ.

Rubinstein, R.Y., Shapiro, A., 1993. Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. John Wiley Sons Inc., Hoboken, NJ.

Ruppert, D., 1988. Efficient estimations from a slowly convergent Robbins-Monro process, Technical report, Cornell University Operations Research and Industrial Engineering.

Ruszczyński, A., 1997. Decomposition methods in stochastic programming. Mathematical Programming 79 (1), 333–353.

Sadegh, P., Spall, J.C., 1998. Optimal sensor configuration for complex systems. In: Chow, J.H. (Ed.), Proceedings of the 1998 American Control Conference, Philadelphia, PA, pp. 376–380.

Scott, W., Frazier, P., Powell, W., 2011. The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. SIAM Journal on Optimization 21 (3), 996–1026.

Shapiro, A., 1991. Asymptotic analysis of stochastic programs. Annals of Operations Research 30 (1), 169–186.

Shapiro, A., 2013. Sample average approximation. In: Gass, S.I., Fu, M. (Eds.), Encyclopedia of Operations Research and Management Science, third ed. Springer, New York.

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2009. Lectures on Stochastic Programming: Modeling and Theory. Society for Industrial Mathematics, Philadelphia, PA.

Shapiro, A., Homem-de Mello, T., Kim, J., 2002. Conditioning of convex piecewise linear stochastic programs. Mathematical Programming 94 (1), 1–19.

Shapiro, A., Wardi, Y., 1996. Convergence analysis of stochastic algorithms. Mathematics of Operations Research 21 (3), 615–628.

Shi, L., Olafsson, S., 2000. Nested partitions method for stochastic optimization. Methodology in Computing and Applied Probability 2 (3), 271–291.

Shin, M., Sargent, R.G., Goel, A.L., 2002. Gaussian radial basis functions for simulation metamodeling. In: Snowdon, J.L., Charnes, J.M. (Eds.), Proceedings of the 2002 Winter Simulation Conference, San Diego, CA, pp. 483–488.

Slivkins, A., 2011. Contextual bandits with similarity information. In: Kakade, S., von Luxborg, U. (Eds.), Proceedings of the 24th Annual Conference on Learning Theory, Budapest, Hungary, pp. 679–701.

Snoek, J., Larochelle, H., Adams, R., 2012. Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, pp. 2960–2968.

Spall, J.C., 1992. Approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control 37 (3), 332–341.

Srinivas, N., Krause, A., Kakade, S.M., Seeger, M., 2010. Gaussian process optimization in the bandit setting: no regret and experimental design. In: Fürnkranz, J., Joachims, T. (Eds.), Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, pp. 1015–1022.

Sutton, R.S., Barto, A.G., 1998. Introduction to Reinforcement Learning. MIT Press, Cambridge.

Takriti, S., Krasenbrink, B., Wu, L.S.-Y., 2000. Incorporating fuel constraints and electricity spot prices into the stochastic unit commitment problem. Operations Research 48 (2), 268–280.

Taylor, G., Parr, R., 2009. Kernelized value function approximation for reinforcement learning. In: Bottou, L., Littman, M. (Eds.), Proceedings of the 26th International Conference on Machine Learning, Montreal, Quebec, pp. 1017–1024.

Tejada-Guibert, J.A., Johnson, S.A., Stedinger, J.R., 1995. The value of hydrologic information in stochastic dynamic programming models of a multireservoir system. Water Resources Research 31 (10), 2571–2579.

Tsitsiklis, J.N., Van Roy, B., 2001. Regression methods for pricing complex American-style options. IEEE Transactions on Neural Networks 12 (4), 694–703.

Verweij, B., Ahmed, S., Kleywegt, A.J., Nemhauser, G., Shapiro, A., 2003. The sample average approximation method applied to stochastic routing problems: a computational study. Computational Optimization and Applications 24 (2–3), 289–333.

Watkins, C.J., Dayan, P., 1992. Q-learning. Machine Learning 8 (3–4), 279–292.

Yang, Y., Zhu, D., 2002. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. Annals of Statistics 30 (1), 100–121.

Zakeri, G., Philpott, A.B., Ryan, D.M., 2000. Inexact cuts in Benders decomposition. SIAM Journal on Optimization 10 (3), 643–657.

Stock Market Predictability

Burton G Malkiel, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This survey reviews the findings of the empirical work concerning stock market predictability. It reviews the studies suggesting that future returns can be predicted on the basis of such fundamental variables as dividend yields, price earnings, price-to-book-value ratios, measures of systematic risk, and market capitalization (size). It also examines evidence that future prices exhibit short-term momentum and longer run mean reversion. The survey concludes by arguing that whatever predictable patterns do exist are unlikely to provide investors with a method to obtain extraordinary returns.

One of the most enduring cornerstones of modern financial theory has been that stock markets are reasonably efficient in reflecting new information and that price movements closely fit a random walk model where future prices cannot be predicted from an examination of past prices. Toward the end of the twentieth century, however, financial economists and statisticians began to question received doctrine, and many came to believe that stock prices were at least partially predictable. For example, stocks that perform the best over the preceding 3–12 months continue to outperform over the subsequent 5–12 months. Moreover, although far more controversial, many have argued that these predictable patterns enable investors to earn excess risk-adjusted rates of returns and provide evidence that the efficient market theory should be rejected. This survey reviews the findings of the empirical work concerning stock market predictability. It reviews the work done on stock prices that appears to reject the random walk hypothesis and looks at the body of studies suggesting that future returns can be predicted on the basis of such fundamental variables as dividend yields, price earnings (P/E), price-to-book-value (P/BV) ratios, measures of systematic risk, and market capitalization (size). The survey also examines the relationship between predictability and market efficiency.

The Efficient Market Theory and Its Critics

The basic idea behind the efficient market theory is that securities markets are extremely efficient in digesting information about individual stocks or about the stock market in general. When information arises about a stock (or the market as a whole), the news spreads very quickly and is immediately incorporated into the prices of securities. Thus, neither technical analysis (an analysis of past price patterns to determine the future) nor fundamental analysis (an analysis of a company's earnings, dividends, future prospects, etc., to determine a stock's proper value) will help investors to achieve returns greater than would be obtained by buying and holding one of the broad stock market indices. The efficient market theory is associated with the idea of a 'random walk,' which is a term loosely used in the finance literature to characterize a price series where all subsequent price changes represent random departures from previous prices. The logic of the random walk idea is that if the flow of information is unimpeded and information is immediately reflected in stock prices, then tomorrow's price change will reflect only tomorrow's news and will be independent of the price changes today. But news is by definition unpredictable and, thus, resulting price changes must be unpredictable and random. As a result, prices fully reflect all known information, and even uninformed investors buying a diversified portfolio at the tableau of prices given by the market will obtain a rate of return as good as that achieved by market professionals. Discussions of the random walk hypothesis can be found in Samuelson (1965) and Fama (1970).

Academic work published mainly after the 1980s has tended to cast doubts on many of the tenets of the efficient market theory. This work has suggested that the random walk model is not supported and that future stock prices and returns can be predicted, at least in part, by an examination of past stock prices and certain 'fundamental' valuation methods such as the ratio of stock prices to earnings and book values. Moreover, some of this work has been interpreted as implying that markets are inefficient in the sense that arbitrage opportunities exist that enable investors to earn excess risk-adjusted returns.

The Major Predictable Patterns That Have Been Discovered

This article will now focus on the findings of empirical work about stock market predictability done from 1980 to 2011. The general thrust of much of the published academic work has been along the following two lines. First, it appears that there are many predictable patterns in the stock market. The stock market is *not* random. Indeed, a book published by Lo and MacKinlay (1999) has the arresting title, *A Non-Random Walk Down Wall Street*. Second, although this is far more controversial, many have argued that these predictable patterns enable an investor to earn big risk-adjusted rates of return and are proof that the efficient market theory is wrong.

There Is Some Evidence of Short-Run Momentum in the Market

When stock returns for individual stocks and market averages are calculated over various periods during the preceding year, there is positive serial correlation. For example, Jegadeesh and Titman (1993, 2001) find that momentum strategies based on 3-, 6-, and 12-month returns have been persistently profitable

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71024-5

during the 1980s and the 1990s. A positive rate of return in one period is more likely than not to be followed by a positive return in the next period. The market is not a perfect random walk. Further evidence provided by Lo and MacKinlay uses a simple specification test based on variance estimators. Their test exploits the fact that the variance of the increments of a random walk is linear in the sampling interval. For example, if stock prices are generated by a random walk, then the variance of monthly sampled log-price relatives must be four times as large as the variance of a weekly sample. Using data over a 23-year time frame from 1962 to 1985, Lo and MacKinlay reject the random walk hypothesis. The rejection cannot be explained completely by infrequent trading (where news affects smaller company stocks with a lag because they trade less frequently) or by time-varying volatilities.

This work suggests that quantitative managers who use momentum to help guide purchases have some empirical support on their side. Whether this is a true inefficiency or not, however, is open to more question. Certainly, any investor who pays transactions costs is not able to outperform a buy-and-hold strategy because the kind of momentum that has been found is small relative to the transaction costs involved. In empirical tests, momentum strategies appear to work better with small stocks (where the strategy is harder to implement because of high transactions costs) than with larger stocks.

The source of momentum profits and the interpretation of the evidence are controversial. For example, Hong and Stein (1999) present a behavioral model based on the hypotheses that delayed overreaction to information pushes the price of winners (losers) above (below) their fundamental value. In subsequent periods, the stocks display momentum as they revert to their fundamental values. Others, such as Conrad and Kaul (1998), suggest that returns from momentum strategies are simply compensation for risk. Carhart (1997), in his study of mutual fund returns, uses momentum as a risk variable (along with capital asset pricing model (CAPM) beta, value, and size) to risk adjust returns. Positive momentum is considered to be strong relative performance over the preceding 12 months (not including the most recent month to allow for any short-term return reversals).

One aspect of the empirical work on momentum has been referred to as 'postearnings announcement drift.' The pattern goes back to Ball and Brown (1968), Rendelman et al. (1982), and Bernard and Thomas (1989). Stocks whose earnings exceed forecasts (whether from analysts or some autoregressive earnings model) do react positively, but the reaction is not immediate. The price keeps rising over subsequent days, weeks, and even months. This underreaction of market prices to earnings surprises appears to be one of the more robust short-horizon predictable patterns and has been explained in terms of 'limits to arbitrage' and market frictions. However, the effect is particularly strong for smaller, less liquid stocks, where transaction costs and market impact costs are likely to eliminate apparent arbitrage profits, as suggested by Chung and Hrazdil (2011).

Since 2010, there have been a number of exchange traded funds (ETFs) formed on the basis of momentum. Depending upon the benchmarks used, some of these momentum strategies executed with real money have outperformed standard buy-and-hold index funds, but most others have underperformed. (See, for example, the international momentum fund, ticker AIMOX, and compare its performance with the standard international index fund for institutional investors, ticker VINX, during 2011 and 2012. Also compare ASMOX (a small-cap momentum fund) with IWM (a Russell 2000 index fund).) These funds have also failed to perform consistently over time. It is therefore far from clear how dependable any momentum pattern that can be discovered will be. The market is not statistically a random walk, but it is nearly so.

But There Is Mean Reversion

Returns may be positively correlated for short periods, such as months or quarters, but when returns are measured over longer periods, such as 3 or 4 years, there tends to be negative correlation. Just as in the Bible 7 lean years followed 7 fat years, so in the stock market extraordinary returns are somewhat more likely to be followed by lackluster returns. For example, Fama and French (1988) demonstrated that long holding period returns are negatively serially correlated. The serial correlation is significant and implies that 25–40% of the variation in long holding period returns can be predicted on the basis of past returns. Similar results were found by Poterba and Summers (1988), who conclude that there is substantial mean reversion in stock market returns at longer horizons. Some studies have attributed this forecastability to the tendency of stock market prices to 'overreact.' DeBondt and Thaler (1985), for example, argue that investors are subject to waves of optimism and pessimism that cause prices to deviate systematically from their fundamental values and later to exhibit patterns of reversal. These findings give some support to investment techniques that rest on a 'contrarian' strategy, that is, buying the stocks, or groups of stocks, that have been out of favor for long periods of time, and avoiding those stocks that have had large run-ups over the last several years.

While there is considerable support for this view, it should be pointed out that such mean reversion is quite a bit weaker in some decades than it is in other periods. Indeed, the strongest empirical results come from periods including the Great Depression. The effect was relatively weak during the 1990s and the first decade of the 2000s. Moreover, such return reversals for the market as a whole may be quite consistent with the efficient functioning of the market and could result, in part, from the volatility of interest rates. There is a tendency when interest rates go up for stocks to go down, and as interest rates go down for stocks to go up. If, in fact, interest rates fluctuate over time, one will tend to get return reversals, or mean reversion, and this is quite consistent with the efficient functioning of markets where stock returns will need to go up or down to be competitive with bonds.

Moreover, it may not be possible to profit from the tendency for individual stocks to exhibit patterns of return reversals. For example, Fluck et al. (1997) simulated a strategy of buying stocks that had particularly poor returns over the preceding 3–5 years. They found very strong statistical evidence of return reversals, but it was truly reversion to the mean, *not* an opportunity to make extraordinary returns. They found that stocks with very low returns over the preceding 3–5 years had higher returns in the next period, but the returns in the next period were similar for both groups. While they found strong evidence of mean reversion, they could not confirm that

a contrarian approach would yield higher than average returns. There was a statistically strong pattern of return reversal, but not one that implied an inefficiency in the market that would enable investors to make excess returns.

There Do Seem To Be Some Seasonal Patterns in the Stock Market

A number of researchers have found that January has been a very unusual month for stock market returns. Returns from an equally weighted stock index have tended to be unusually high during the first 2 weeks of the year. The return premium has been particularly evident for small stocks, i.e., those with relatively small total capitalizations, as has been demonstrated by Keim (1983). Haugen and Lakonishok (1988) detail the high January returns in a book entitled *The Incredible January Effect*.

There also appear to be a number of seasonal effects. For example, French (1980) documents significant Monday returns. There appear to be significant differences in average daily returns in countries other than the United States, as summarized by Hawawini and Keim (1995). There also appear to be some patterns in returns around the turn of the month, as shown by Lakonishok and Smidt (1988), as well as around holidays, as shown by Ariel (1990).

The general problem with these anomalies, however, is that they are *not* dependable from period to period. The January effect has not been useful to investors through much of the 1990s and the early 2000s. Moreover, these nonrandom effects (even if they were dependable) are small relative to the transaction costs involved in trying to exploit them. They do not appear to offer investors any arbitrage opportunities that would enable them to make excess returns.

The Size Effect

Probably one of the stronger effects that investigators have found is the so-called size effect: the tendency over long periods of time for small stocks to generate larger returns than those of large stocks. Since 1926, small-company stocks in the United States have produced a rate of return about 1.5 percentage points larger than the returns from large-company stocks.

Fama and French (1992) looked at data from 1963 to 1990 and divided all stocks into deciles according to their size as measured by total capitalization. Decile 1 contained the smallest 10% of all stocks, while decile 10 contained the largest stocks. The results plotted in **Figure 1** show a clear tendency for the deciles made up of portfolios of smaller stocks to generate higher average monthly returns than deciles made up of larger stocks.

If the 'beta' measure of systemic risk from the CAPM is accepted as the correct risk measurement statistic, the size effect can be interpreted as indicating market inefficiency. This is so because portfolios consisting of smaller stocks have excess risk-adjusted returns. But any conclusion that an inefficiency exists involves accepting the joint hypothesis that a size effect is present *and* that the CAPM is correct. Fama and French point out, however, that the average relationship between 'beta' and return is flat – not upward sloping as the CAPM predicts. Moreover, within beta deciles, 10 portfolios constructed by size display the same kind of positive relationship shown in **Figure 1**. On the other hand, within size deciles, the

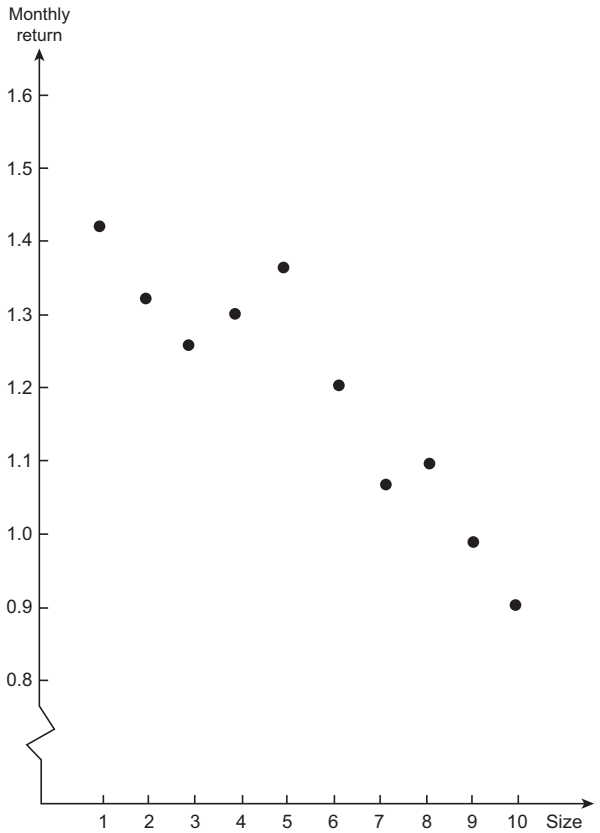

Figure 1 Average monthly returns for portfolios, formed on the basis of size; 1963–90. Source: Fama, E.F., French, K.R., 1997. The cross-section of expected stock returns. Journal of Finance 47(2), 427–465.

relationship between beta and return continues to be flat. Fama and French conclude that size may well be a far better proxy for risk than beta, and therefore that their findings should not be interpreted as indicating that markets are inefficient.

The dependability of the size effect is open to question. The size effect has not been consistent from one decade to the next. During the first decade of the twenty-first century, there was a pronounced size effect in United States and other world stock markets, where small-cap stocks produced substantially higher returns than large-cap stocks. But during the 1990s the opposite was true. Higher returns were earned on large-cap stocks. Moreover, some portion of the size effect may result from survivorship bias in currently available computer sources of past returns. Today's lists of companies include only small firms that have survived, not the ones that later went bankrupt. Thus, a researcher who examined the 10-year performance of today's small companies would be measuring the performance of those companies that survived – not of the ones that failed.

There Is Evidence That Stocks with Low P/E Multiples Generate Higher Returns Than Those with High P/E Multiples

As **Figure 2** shows, evidence first published by Nicholson (1960) and later confirmed by Ball (1978) and Basu (1983) indicates that low-P/E stocks provide higher rates of return than high-P/E stocks. This finding is consistent with the views of behavioralists (see, for example, Kahneman and Riepe,

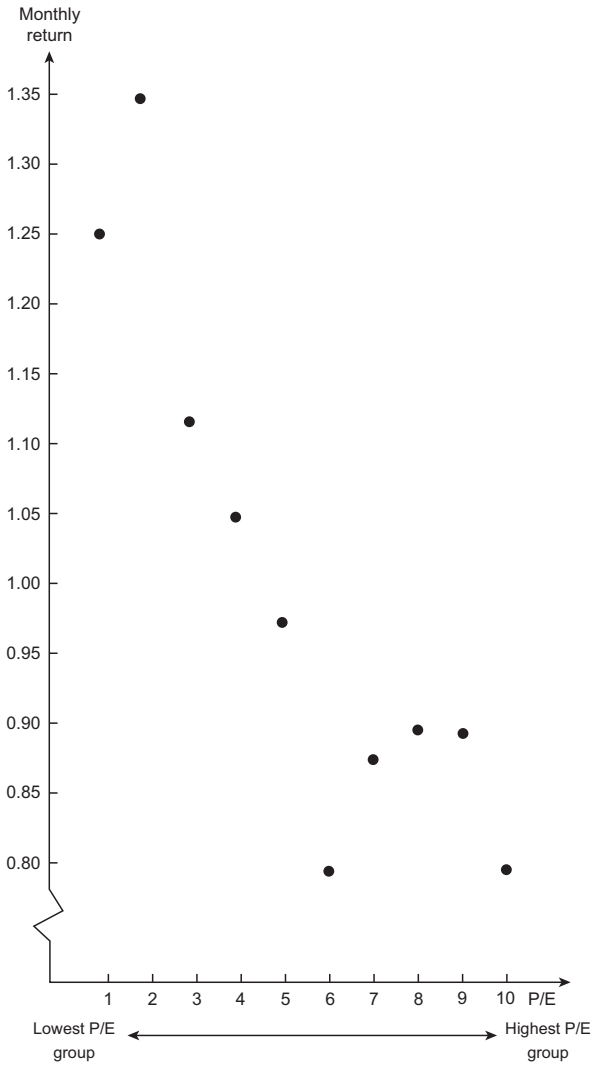

Figure 2 Average monthly returns for portfolios, formed on the basis of price earnings (P/E) ratio; 1962–90. Source: Hawawini, G., Keim, D.B., 1995. On the predictability of common stock returns: worldwide evidence. In: Jarrow, R., Maksimovic, V., Ziemba, W. (Eds.), Handbooks in Operations Research & MS, vol. 9. Elsevier Science, Amsterdam, New York, pp. 497–544.

1998) that investors tend to be overconfident of their ability to project high earnings growth and thus overpay for 'growth' stocks. The finding is also consistent with the views of Benjamin Graham and David Dodd (1934), first expounded in their classic book on security analysis and later championed by the legendary US investor, Warren Buffett. Similar results have been shown for price/cash-flow multiples, where cash flow is defined as earnings plus depreciation and amortization, (see Hawawini and Keim, 1995).

Stocks with Low P/BV Ratios Generate Larger Returns than Stocks with High P/BV Ratios

The ratio of stock P/BV (the value of a firm's assets minus its liabilities divided by the number of shares outstanding) has also been found to be a useful predictor of future security returns. Low P/BV (along with low P/E) is considered to be another hallmark of the so-called 'value' in equity securities, and is also consistent with the view of behavioralists that investors tend to overpay for 'growth' stocks that subsequently fail to live up to expectations. Fama and French (1992) concluded that size and P/BV together provide considerable explanatory power for future returns and once they are accounted for, little additional influence can be attributed to P/E multiples. Fama and French (1997) also conclude that the P/BV effect is important in many world stock markets other than that of the United States.

Lakonishok et al. (1994) argue that such results raise questions about the efficiency of the market. But these findings do not necessarily imply inefficiency. P/BV may be a more accurate proxy for the systematic risk priced in the stock market than traditional estimates of the CAPM beta because of measurement errors in the traditional estimates. Indeed, this is the explanation Fama and French give for this predictable pattern.

While there has been a pronounced value effect over the entire period from the early 1960s to the 2011s, the effect is not dependable in all periods. The so-called 'value' stocks substantially underperformed 'growth' stocks (those with high P/E and high P/BV ratios) during the 1990s. Value stocks then produced outstanding relative returns from early 2000 to the end of 2006. For the next 5 years, through 2011, however, growth stocks enjoyed higher returns. In general, value stocks produce the best relative returns when valuation metrics are widely dispersed, as they were in January 2000 near the top of the 'dot-com bubble.' Finally, it should be noted that considerable merger and restructuring activity has made book value increasingly hard to interpret. Thus, it is far from clear that the pattern will continue.

The High Risk—Low Return Findings

As indicated above, the relationship between the beta and subsequent return is not upward sloping as the CAPM predicts. But then CAPM may turn out to be a particularly useful model that could generate exceptional returns. Papers by Frazzini and Pedersen (2011) and Baker et al. (2011) have simulated a strategy of initiating long positions in low-beta stocks and short positions in high-beta stocks to produce a beta-neutral portfolio. The simulated strategy has produced extraordinary results over a more than 40-year period from January 1968 to December 2008. A dollar invested in the lowest quintile beta portfolio grew to $60.46 ($10.25 in real terms). A dollar invested in the highest quintile beta portfolio grew to only $3.77 and fell to 69 cents in real terms. The same kind of result is obtained when using price volatility and idiosyncratic risk (see Ang et al., 2006) as well as in international markets. Whether such a pattern continues to hold if constrained investors no longer bid up the prices of high-beta stocks is open to question, as is the feasibility of the strategy in the presence of transactions costs and short-sale constraints. It should be noted, however, that both high-beta and low-beta ETFs began to be marketed in mid-2011. For at least the next year, the high-beta ETF underperformed its benchmark, while the low-beta fund exceeded its benchmark performance (see, for example, the performance of the Power Shares S&P 500 high-beta portfolio (SPHB) during 2011 and 2012).

Predicting Market Returns from Initial Dividend Yields

Another apparently predictable relationship concerns the returns realized from stocks and the initial dividend yields at which they were purchased. The data since 1926 for the United States are presented in Figure 3. The figure was produced by measuring the dividend yield of the broad stock market (in this case, the Standard and Poor's 500 Stock Index) each quarter between 1926 and 2001, and then calculating the market's subsequent 10-year total return through 2011. The observations were then divided into deciles depending upon the level of the initial dividend yield. In general, the figure shows that investors have earned a higher rate of return from the stock market when they purchased stocks with an initial dividend yield that was relatively high (the first three deciles), and a below-average rate of return when they purchased them at initial dividend yields that were low (the last three deciles).

Formal statistical tests of the ability of dividend yields (dividend-price ratios) to forecast future returns have been conducted by Fama and French (1988) and Campbell and Shiller (1988). Depending on the forecast horizon involved, as much as 40% of the variance of future returns can be predicted on the basis of initial dividend yields.

These findings are not necessarily inconsistent with efficiency. Dividend yields of stocks tend to be high when interest rates are high, and they tend to be low when interest rates are low. Consequently, the ability of initial yields to predict returns may simply reflect the adjustment of the stock market to general economic conditions. Moreover, the dividend behavior of US corporations may have changed over time (see Fama and French 1999). Companies in the twenty-first century may be more likely to institute a share repurchase program rather than increase their dividends. Thus, dividend yield may not be as meaningful as in the past. Moreover, it is worth pointing out that dividend yields were unusually low at the start of 1995. An investor who avoided US stocks then would have missed out on one of the steepest 5-year increases in stock prices in history.

Finally, this phenomenon does not work consistently with individual stocks. Investors who simply purchase a portfolio of individual stocks with the highest dividend yields in the market will not earn a particularly high rate of return.

One popular implementation of such a 'high-dividend' strategy is the 'Dogs of the Dow Strategy,' which involves buying the 10 stocks in the Dow Jones Industrial Average with the highest dividend yields. For some periods, this strategy handily outpaced the overall average, and so several 'Dogs of the Dow' funds were brought to market during the mid-1990s and aggressively sold to individual investors. Such funds generally underperformed the market averages during the next 5-year period.

Predicting Market Returns from Initial P/E Multiples

The same kind of predictability for the market as a whole has been shown for P/E ratios. The data are shown in Figure 4. The figure presents a decile analysis similar to that described for dividend yields above. Investors have tended to earn larger long-horizon returns when purchasing stocks at relatively low P/E multiples. Campbell and Shiller (1988) report R^2s of over 40% and conclude that equity returns have been predictable in the past to a considerable extent. Whether such historical relations are necessarily applicable to the future is another matter. Such an analysis conducted by Campbell and Shiller in December 1996 predicted a long-horizon (nominal) return for the US stock market close to zero.

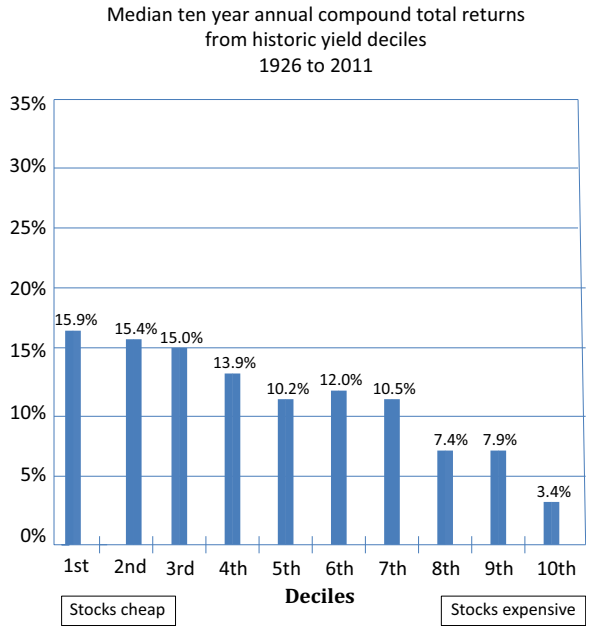

Figure 3 The future 10-year rates of return when stocks are purchased at alternative initial dividend yields (divident/price (D/P)). Source: The Leuthold Group.

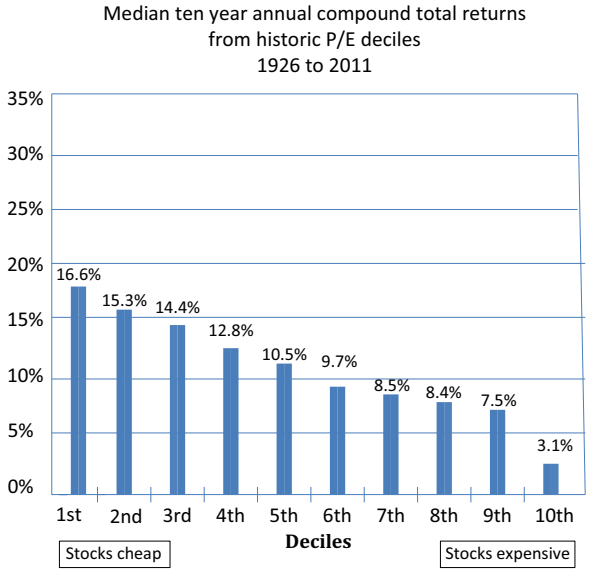

Figure 4 The future 10-year rates of return when stocks are purchased at alternative initial P/E multiples. Source: The Leuthold Group.

The findings presented to the Board of Governors of the Federal Reserve System led Alan Greenspan to question whether stock prices were exhibiting irrational exuberance, but the stock market continued to rally strongly through the first quarter of 2000.

The Adaptive Market Hypothesis

Financial economists have increasingly entertained the notion of temporary instabilities in the relationship between underlying economic relationships, as suggested by Stock and Watson (2003). This may be equally true of the question of stock market predictability. One useful development that has been championed by Lo (2004, 2008) is the 'adaptive market hypothesis' (AMH), where time-varying predictability may exist (see also Farmer and Lo, 1999; Farmer, 2002). For Lo's evolutionary AMH, the lack of stock market predictability is not an all-or-nothing condition, but rather one that can vary over time, depending on the characteristics of existing limits to arbitrage, market imperfections, and psychological biases. These factors can give rise to periods when stock returns are predictable and departures from the efficient market hypothesis (EMH) appear to exist. In this sense, EMH and behavioral finance can comfortably coexist. But investors do learn from past history, and it is not reasonable to expect that any predictable patterns will persist, except those that are based on differences in risk. While it is possible that $100 bills may be found on the ground from time to time, they should be picked up right away. They are unlikely to be sitting there for long and, as Grossman and Stiglitz (1980) have emphasized, they will be picked up only if their value exceeds the cost. To be sure, there appear to be many examples where asset prices reached such 'bubble' levels that, at least in retrospect, it seems obvious that corrections were inevitable and that markets were inefficient. The housing bubble that peaked in 2002 is one example, as is the Internet stock market bubble that peaked in 2000. And no one can reasonably argue that the prevailing prices in these examples were 'correct.' But the EMH does not require that prices be correct. Indeed, prices are always wrong. What EMH posits is that no one knows for sure whether the prices are too high or too low and that no obvious arbitrage possibilities exist to earn abnormal returns.

As discussed above, many analysts claimed during the 1990s that the stock market had reached bubble levels at prices substantially lower than were actually realized and many years before the actual peak. Similarly, the collapse of the housing bubble that led to the most severe world economic downturn since the Great Depression was not predicted by the monetary authorities. In the author's view, the correct conclusion is that we should be modest in our appraisal of how predictable asset prices really are. EMH remains an extremely useful hypothesis that is strongly supported by experience.

Concluding Comments

In summary, there are many statistically significant predictable patterns in the stock market. It is important to emphasize, however, that these patterns are not dependable in each and every period, and that some of the patterns based on fundamental valuation measures of individual stocks may simply reflect a better proxy for measuring risk. Moreover, many of these patterns could self-destruct in the future, as many of them have already done. Indeed, this is the logical reason why one should be cautious not to overemphasize these anomalies and predictable patterns.

Suppose, for example, that one of the anomalies is really true. Suppose that there is a truly dependable January effect, and that the stock market – especially stocks of small companies – will generate extraordinary returns during the first 5 days of January. What will investors do? They will buy late in December, on the last day of December, and sell on January 5. But then investors find that the market rallied on the last day of December and so they will begin to buy on the next-to-last day. Moreover, because there is so much 'profit taking' on January 5, investors will have to sell on January 4 to take advantage of this effect. Thus, to beat the gun, investors will have to be buying earlier and earlier in December, and selling earlier and earlier in January so that eventually the pattern will self-destruct. Any truly repetitive pattern that can be discovered in the stock market and can be arbitraged away will self-destruct. Indeed, the January effect became undependable after it received considerable publicity.

Similarly, suppose there is a general tendency for stock prices to underreact to certain news events, leading to abnormal returns to investors who exploit the lack of full immediate adjustment. 'Quantitative' investment managers will then develop strategies in an attempt to exploit the pattern. Indeed, the more potentially profitable a discoverable pattern is, the less likely it is to survive.

Moreover, many of the predictable patterns that have been discovered may simply be the result of data mining. The ease of experimenting with financial data banks of almost every conceivable dimension makes it quite likely that investigators will find some seemingly significant but wholly spurious correlation between financial variables or among financial and nonfinancial data sets. An amusing one that has been uncannily accurate in the past for US markets is the Super Bowl Indicator, which predicts the sign of market returns for the year by the winner of January's football Super Bowl. Given enough time and massaging of data series, it is possible to tease almost any pattern out of every data set. Moreover, the published literature is likely to be biased in favor of reporting such results. Significant effects are likely to be published in professional journals, while negative results, or boring confirmations of previous findings, are relegated to the file drawer or discarded. Data-mining problems are unique to nonexperimental sciences, such as economics, which rely on statistical analysis for their insights and cannot test hypotheses by running repeated controlled experiments.

An exchange at an academic symposium (Roll-Shiller, 1992) between Robert Shiller (1981), an economist who is sympathetic to the argument that stock prices are partially predictable and skeptical about market efficiency, and Richard Roll, an academic financial economist who was also a businessman managing billions of dollars of investment funds, is quite revealing. After Shiller stressed the importance

of inefficiencies in the pricing of stocks, Roll responded as follows:

> I have personally tried to invest money, my client's money and my own, in every single anomaly and predictive device that academics have dreamed up … I have attempted to exploit the so-called year-end anomalies and a whole variety of strategies supposedly documented by academic research. And I have yet to make a nickel on any of these supposed market inefficiencies…a true market inefficiency ought to be an exploitable opportunity. If there's nothing investors can exploit in a systematic way, time in and time out, then it's very hard to say that information is not being properly incorporated into stock prices (Roll and Shiller, 1992).

Finally, there is a remarkably large body of evidence suggesting that professional investment managers are not able to outperform index funds that simply buy and hold the broad stock market portfolio (see, for example, Malkiel, 2012). Three-quarters of professionally managed funds are regularly outperformed by a broad index fund with equivalent risk, and those that do appear to produce excess returns in one period are not likely to do so in the next. The record of professionals does not suggest that sufficient predictability exists in the stock market to produce exploitable arbitrage opportunities.

Pricing irregularities and predictable patterns in stock returns do appear over time and even persist for periods. Undoubtedly, with the passage of time and with the increasing sophistication of databases and empirical techniques, further apparent departures from efficiency, and further patterns in the development of stock returns, will be documented. But the end result will not be an abandonment of the belief of many in the profession that the stock market is remarkably efficient in its utilization of information and that whatever patterns do exist are unlikely to provide investors with a method to obtain extraordinary returns.

To be sure, there appear to be many examples where asset prices reached such 'bubble' levels that, at least in retrospect, it seems obvious that corrections were inevitable and that markets were inefficient. The housing bubble that peaked in 2002 is one example as is the Internet stock market bubble that peaked in 2000. And no one can reasonably argue that the prevailing prices in these examples were 'correct.' But the EMH does not require that prices be correct. Indeed, prices are always wrong. What EMH posits is that no one knows for sure whether the prices are too high or too low and that no obvious arbitrage possibilities exist to earn abnormal returns.

As discussed above, many analysts claimed during the 1990s that the stock market had reached bubble levels at prices substantially lower than were actually realized and many years before the actual peak. Similarly, the collapse of the housing bubble that led to the most severe world economic downturn since the Great Depression was not predicted by the monetary authorities. In the author's view, the correct conclusion is that we should be modest in our appraisal of how predictable asset prices really are. EMH remains an extremely useful hypothesis that is strongly supported by experience.

See also: Behavioral Economics; Economic History.

Bibliography

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. Journal of Finance 61 (1), 259–299.

Ariel, R.A., 1990. High stock returns before holidays: existence and evidence on possible causes. Journal of Finance 45, 1611–1626.

Asness, C., Frazzini, A., Pedersen, L.H., 2012. Leverage aversion and risk parity. Financial Analysts Journal 68, 47–59.

Baker, M., Bradley, B., Wurgler, J., 2011. Benchmarks as limits to arbitrage: understanding the low-volatility anomaly. Financial Analysts Journal 67, 1–15.

Ball, R., 1978. Anomalies in relationships between securities' yields and yield-surrogates. Journal of Financial Economics 6, 103–126.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6 (2), 159–178.

Basu, S., 1983. The relationship between earnings yield, market value and the returns for NYSE common stocks: further evidence. Journal of Financial Economics 12, 129–156.

Bernard, V.L., Thomas, J.K., 1989. Post-earnings announcement drift: delayed price response or risk premium? Journal of Accounting Research 27, 1–36.

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton, NJ.

Campbell, J.Y., Shiller, R.J., 1988. Stock prices, earnings, and expected dividends. Journal of Finance 43, 661–676.

Carhart, M.M., 1997. On persistence in mutual fund performance. The Journal of Finance 52 (1), 57–92.

Chung, D.Y., Hrazdil, K., 2011. Market efficiency and the post-earnings announcement drift. Contemporary Accounting Research 28 (3), 926–956.

Conrad, J., Kaul, G., 1998. Time variation expected returns. Journal of Business 61 (4), 409–425.

DeBondt, W.F.M., Thaler, R., 1985. Does the stock market overreact? Journal of Finance 40, 793–805.

DeBondt, W.F.M., Thaler, R., 1995. Financial decision-making in markets and firms: a behavioral perspective. In: Jarrow, R., Maksimovic, V., Ziemba, W. (Eds.), Handbook in Operations Research & MS, vol. 9. Elsevier Science.

Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. Journal of Finance 25, 383–417.

Fama, E.F., French, K.R., 1988. Permanent and temporary components of stock prices. Journal of Political Economy 96, 246–273.

Fama, E.F., French, K.R., 1997. Value vs. growth: the international evidence. Journal of Finance 53, 1975–1999.

Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. Journal of Finance 47 (2), 427–465.

Fama, E.F., French, K.R., 2001. Disappearing dividends: changing firm characteristics or increased reluctance to pay? Journal of Financial Economics 60, 3–43.

Farmer, J.D., 2002. Market force, ecology and evolution. Industrial and Corporate Change 11, 895–953.

Farmer, J.D., Lo, A., 1999. Frontiers of finance: evolution and financial markets. Proceedings of the National Academy of Sciences 96, 9991–9992.

Fluck, Z., Malkiel, B.G., Quandt, R.E., 1997. The predictability of stock returns: a cross-sectional simulation. Review of Economics and Statistics 79, 176–183.

Frazzini, A., Pedersen, L., October 9, 2011. Betting against beta. In: Swiss Finance Institute Research Paper No. 12–17.

French, K.R., 1980. Stock returns and the weekend effect. Journal of Financial Economics 8, 55–69.

Graham, B., Dodd, D.L., 1934. Security Analysis: Principles and Techniques. McGraw-Hill, New York.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of the informationally efficient markets. The American Economic Review 70 (3), 393–408.

Haugen, R.A., Lakonishok, J., 1988. The Incredible January Effect. Dow Jones-Irwin, Homewood, AL.

Hawawini, G., Keim, D.B., 1995. On the predictability of common stock returns: worldwide evidence. In: Jarrow, R., Maksimovic, V., Ziemba, W. (Eds.), Handbooks in Operations Research & MS, vol. 9. Elsevier Science, Amsterdam, New York, pp. 497–544.

Hong, H., Stein, J.C., 1999. A unified theory of underreaction, momentum trading and overreaction in asset markets. The Journal of Finance 54, 2143–2184.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. Journal of Finance 48 (1), 65–91.
Jegadeesh, N., Titman, S., 2001. Profitability of momentum strategies: an evaluation of alternative explanations. The Journal of Finance 56 (2), 699–726.
Kahneman, D., Riepe, M.W., 1998. Aspects of investor psychology. Journal of Portfolio Management 24, 52–65.
Keim, D.B., 1983. Size-related anomalies and stock return seasonality: further empirical; evidence. Journal of Financial Economics 12, 13–32.
Lakonishok, J., Schleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation, and risk. Journal of Finance 49, 1541–1578.
Lakonishok, J., Smidt, S., 1988. Are seasonal anomalies real? A ninety-year perspective. Review of Financial Studies 1, 403–425.
Lo, A.W., 2004. The adaptive markets hypothesis: market efficiency from an evolutionary perspective. Journal of Portfolio Management 30, 15–29.
Lo, A.W., 2008. Where do alphas come from?: A new measure of the value of active investment management. Journal of Investment Management 6, 1–29.
Lo, A.W., MacKinlay, A.C., 1999. A Non-Random Walk Down Wall Street. Princeton University Press, Princeton, NJ.
Malkiel, B.G., 1995. Returns from investing in equity mutual funds 1971 to 1991. Journal of Finance 50, 549–572.
Malkiel, B.G., 2012. A Random Walk Down Wall Street, tenth ed. Norton, New York.
Nicholson, S.F., July/August 1960. Price-earnings ratios. Financial Analysts Journal, 43–50.
Poterba, J.M., Summers, L.H., 1988. Mean reversion in stock returns: evidence and implications. Journal of Financial Economics 22, 27–59.
Rendleman Jr., R.J., Jones, C.P., Latane, H.A., 1982. Empirical anomalies based on unexpected earnings and the importance of risk adjustments. Journal of Financial Economics 10, 269–287.
Roll, R., Shiller, R.J., 1992. Comments: symposium on volatility in US and Japanese stock markets. Journal of Applied Corporate Finance 1, 25–29.
Samuelson, P., 1965. Proof that properly anticipated prices fluctuate randomly. Industrial Management Review 6, 41–49.
Shiller, R.J., 1996. Price-Earnings Ratios as Forecasters of Returns: The Stock Market Outlook in 1996. Unpublished manuscript. Yale University.
Shiller, R.J., 1981. Do stock prices move too much to be justified by subsequent changes in dividends? American Economic Review 71, 421–436.
Stock, J.H., Watson, M.W., 2003. Has the bussiness cycle changed and why? NBER Macroeconomics Anual 2002 (17), 159–230.

Stockholders' Ownership and Control

Michael Useem, University of Pennsylvania, Philadelphia, PA, USA
Gerald F Davis, The University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Useem, G.F. Davis, volume 22, pp. 15133–15137, © 2001, Elsevier Ltd.

Abstract

Corporate governance research traditionally focused on how American corporations overcame potential pathologies created by the separation of ownership and control. Theorists described an array of mechanisms that oriented managers toward shareholder value, from incentive compensation and vigilant directors to a 'market for corporate control.' Research revealed, however, that top management retained wide discretion. Subsequent comparative and historical studies concluded that American-style corporate governance is only one of many possible formats. On the other hand, globalizing equity markets encourage common models for optimizing shareholder value and performance transparency, and at least partial convergence in the relations among managers, directors, and owners.

Theory and research on the relations among top managers, company directors, stockholders, and external contenders for corporate control experienced a remarkable flowering during the 1990s and 2000s. Early work addressed the central puzzle raised by the widespread separation of ownership and control among large American corporations, namely, why would any sensible person – much less thousands of them – invest their savings in businesses run by unaccountable professional managers? As Berle and Means (1932) framed the problem, those who ran such corporations would pursue 'prestige, power, or the gratification of professional zeal' in lieu of maximizing profits. Shareholders weakened by their fractionation could do little to stop them. Yet generations of individuals and financial institutions continued to invest in these firms. Why?

Resolving the Separation of Stock Ownership from Company Control

Answering this question led to the creation of a new theory of the firm that portrayed the public corporation as a 'nexus of contracts.' In this model, the managers of the corporation were disciplined to optimize shareholder value by an array of mechanisms, from incentive compensation and vigilant directors to an unforgiving 'market for corporate control.' Taken together, these devices worked to vouchsafe shareholder interests even when ownership was widely dispersed. Research in this tradition flourished as takeovers of underperforming firms in the United States became common and restive institutional investors made their influence known. Studies focused on assessing the effectiveness of such contrivances as governing boards dominated by outsiders, executive pay tied to stock performance, and susceptibility to hostile acquisition (Bebchuk and Weisbach, 2010; Gillan, 2006; Kang and Zardkoohi, 2005; Larcker and Tayan, 2011).

Three developments, however, expanded research on ownership and control beyond these mechanisms and beyond American business. The first was for investigators to view the governance structure of the firm – the set of devices that evolve within the firm to monitor managerial decision making – as an ensemble. Rather than regarding any particular aspect of the firm's structure as essential, researchers came to study them as complements or substitutes. Remuneration highly leveraged around share price, for instance, could act as a substitute for a vigilant board, and together they could really tighten the leash.

The second development was a growth of comparative and historical research that highlighted the idiosyncrasy of the American obsession with ownership and control. American-style corporate governance, aimed at 'solving' the problems created by the separation of ownership and control, came to be seen as only one of several feasible systems. Indeed, among some advanced economies with well-developed business enterprise, the problem needed no solving at all since the separation of ownership from control remained much the exception (Conger, 2009).

The third development was the articulation of a reflexive stance on the theory of ownership and control. While agency theory can be viewed as an empirical theory of the corporation, it also came to be considered a prescriptive theory, an explanation of not only what *is* but also a vision of what *could* or *should* be. Its influence on public policy debates in recent years is evident in the widespread rise of the rhetoric of shareholder value in the business press and executive suite (Davis, 2009; Lounsbury and Hirsch, 2010; Useem, 1996).

Shareholder Influence on Top Management and the Corporation

When ownership is widely dispersed, shareholders possess few direct devices to ensure their control – or at least their influence – on the company's direction and performance. This is the essence of the 'agency problem' identified by Berle and Means (1932), and it remains potentially potent among major US firms. Ownership is still relatively dispersed by global standards, although a small number of mutual funds and exchange traded funds have gathered relatively large ownership positions in hundreds of US corporations over the past 20 years (Davis, 2008), and by 2011 BlackRock (the world's largest asset manager) has become the largest

shareholder of one in five US corporations, including Citigroup, JP Morgan Chase, ExxonMobil, Chevron, AT&T, and Verizon.

Scattered shareholders in effect relinquish their control to directors whom they elect to serve as their agents in selecting and supervising top management. Dispersed holders can influence the firm only through trading shares and voting directors. They have no real voice in the selection of executives, or even who serves on the board. By one argument, shareholder inattention is in fact the only rational course for small shareholders since the cost of learning about the governance of the firm and mobilizing change is sure to be far overshadowed by the marginal and nonselective benefits of any resulting performance improvement. This problem of collective action could result in management unaccountability in some capital markets, as was evident in the financial crisis of 2008–09, but investors nonetheless rest more easy in the United States because of a well-developed set of mechanisms that have arisen to ensure share price maximization without overt investor control (Easterbrook and Fischel, 1991).

One forceful instrument for riveting attention on share price is the 'market for corporate control.' By hypothesis, a poorly run corporation suffers a depressed market valuation, and this opens an opportunity for outsiders with better management skills to acquire the firm at an implicit discount, oust top management, and rehabilitate the firm, thus restoring value to its long-suffering stockholders. Unwanted takeovers and their threats provide a potent mechanism to shock top managers who underserve shareholder interests; they know that executives generally find themselves on the street after an unfriendly takeover. The market for corporate control is thus one of the more visible hands of Darwinian discipline, weeding out poorly run firms and protecting shareholders from bad management (Manne, 1965). During the 1980s, more than one-quarter of the Fortune 500 largest industrials in the United States received takeover bids, although the risk of unwanted takeover declined dramatically after 1990.

Hostile takeovers are a blunt instrument for corporate control, and as a result, some US institutional investors moved toward more fine-grained methods of influence. Large pension funds, for example, often field shareholder proxy proposals for better governance. Evidence suggests that such efforts push targeted firms to improve their governance and thereby further open them further to investor influence (Karpoff, Unpublished). On the other hand, conflicts of interest often limit investor activism; for instance, mutual funds are more likely to vote with management on shareholder proposals to the extent that they manage corporate-sponsored pension plans, as many of the largest mutual funds do (Davis and Kim, 2007).

If hostile takeovers are often too blunt an instrument and shareholder activism too diffuse, stock analysts provide an intermediate form of shareholder influence on company behavior. Analysts investigate companies in order to render judgments on their prospects for future earnings and thus to estimate their appropriate valuation. In principle, the rewards go to the most accurate analysts, giving them incentives to act as corporate watchdogs (Useem, 1996). In practice, however, analysts often simply follow the lead of their peers rather than forming independent judgments (Rao et al., 2001).

Between Ownership and Control: Top Management

A handful of company executives make the defining decisions, whether to launch a new product, enter a promising region, or resist a tender offer. It is they who most directly control the fate of the firm. It is they who stand between corporate owners and company results (Cannella et al., 2008).

Firms often define their 'top' as no more than the seven or eight most senior officers. To much of the world beyond the company walls, however, top management is personified almost solely by the chief executive. Academic researchers had long been drawn to the same pinnacle of the pyramid, partly on the conceptual premise that the chief executive is the manager who really matters, and partly on the pragmatic ground that little is publicly available on anybody except the CEO.

We have also benefited from a long accumulation of work on chief executive family histories, educational pedigrees, and political identities, and we know as a result that their origins are diverse, credentials splendid, and instincts conservative. More contemporarily, we have learned much about their tangled relations with directors and investors as well. To know the CEO's personal rise and board ties is to anticipate much of the firm's strategic intent and performance promise. They are also a good predictor of the CEO's own fortunes. An elite MBA degree accelerates movement to the top, a wealthy background attracts outside directorships, and a hand-picked board enhances pay and perquisites (Stern and Westphal, 2010; Useem and Karabel, 1986).

The conceptual and pragmatic underpinnings for shining the light solely on the CEO as the fulcrum for ownership and control, however, have eroded in recent years, as companies redefined their operations and researchers reconceived their methods. At many companies, CEOs have fashioned teams among their top officers that meet frequently and resolve jointly, their vertical control giving way to lateral leadership. As a result, researchers expanded their field of view from the chief executive to the entire upper echelon. They applied fresh strategies, ranging from personal interviews to direct observation and Internet surveys, to acquire data on top managers other than the chief executive who now matter far more to company performance and stock analysts. A host of studies subsequently confirmed what many executives already appreciated. A better predictor of a company's performance than the chief executive's capability is the quality of the team that now runs the show (Cannella et al., 2008).

Yet even this widening of research attention beyond the chief executive has not broadened enough. Many companies have expanded the concept of top management to include one, two, or even 300 senior managers whose decisions have decisive bearing on firm performance. The concept of top management, then, has expanded in the minds of company managers and, to a lesser extent, academic researchers, from just the boss to the boss's team and now to the boss's court. The board of directors of many companies has also become a partner with management in leading the firm (Charan et al., forthcoming).

If shareholders are increasingly insistent that company directors require top managers to deliver steadily rising returns on their investments, their action is premised on a critical assumption that these top management teams do make the

difference. While the premise may seem intuitively obvious to those inhabiting this world, it is far from self-evident to those who study it. Observers diverge from the assumption of influential executives in two opposing directions, one viewing top managers as all powerful, others seeing them as all but powerless.

Vilfredo Pareto and kindred 'elite' theorists early articulated the all-powerful view, but C. Wright Mills captured it best for the United States in his classic work, *The Power Elite* (1956). Company executives, in his acid account, had joined with government officials and military commanders in an unholy alliance to exercise a commanding control over companies and the country. To the extent that top management is indeed all powerful, shareholders possess a silver bullet for righting whatever has gone wrong in a company: install new management.

For the community of management scholars, Pfeffer (1981) well articulated the opposite, all-*powerless* view with the contention that market and organizational constraints so tied top management's hands that it was much the captive, not a maker, of its own history. Production technologies and competitive frays in this view are far more determining of company results than the faceless executives who sit in their suites. It counts little who serves in top management, and, by extension, on the board. If top management is as powerless as the imagery suggests, investors seeking new management in response to poor performance are surely wasting their time.

Company executives do complain of a seeming helplessness at times, but almost anybody in working, consulting, or research contact with top managers reports just the opposite. Direct witnesses typically describe instead a commanding presence of top management, and they characterize its capabilities as making the difference between a company's success or failure. Investors themselves think otherwise as well. When companies announce executive succession, money managers and stock analysts are quick to place a price on the head of the newly arrived, and, depending on the personality, billions can be added to or subtracted from the company's capitalization. Ample evidence confirms that the chief executive does make a significant difference in company performance (Thomas, 1988).

The importance that investors and directors attribute to top management for growing their fortunes can also be seen in the ultimate punishment for failure to do so: dismissal. Whether the United States, Japan, or Germany, the likelihood of a CEO's exit in the wake of a stock free fall is increased by as much as half, and investors are often the engine behind the turnover. Study of Japanese companies stunned by a sharp reversal of fortune, for example, reveals that those whose top 10 shareholders control a major fraction of the firm's stock are more often than others to dismiss the president and replace directors (Kaplan, 1997).

Similarly, company boards are observed to reward successful executives far more generously than in the past. For each additional $1000 added to a company's value in 1980, directors on average provided their chief executive an extra $2.51. By 1990, the difference in their payments had risen to $3.64, and by 1994 to $5.29. Put differently, in 1980 the boards of companies ranked at the 90th percentile in performance gave their CEOs $1.4 million more than did boards whose firms ranked at the 10th percentile (in 1994 dollars). But by 1990, the boards had expanded that gap to $5.3 million and by 1994 to $9.2 million (Hall and Liebman, 1998).

Between Stockholders and Top Management: The Governing Board

If companies make semiconductors and concrete much the same way the world around, they organize their directors in almost as many ways as there are national economies. American boards tend to include only one or two insiders, while Japanese boards traditionally included only a handful of outsiders. German and Dutch governance is built around a two-tier governance structure, employees holding half of the upper-tier seats; British and Swiss governance is designed around a single-tier, management-dominated structure. Norway requires that two-fifths of board seats are held by women; Spain and the Netherlands have mandated the same. Some systems give formal voice to labor, others none: German law requires that labor representatives serve on the board, while French law places labor observers in the boardroom. American law mandates less, but US boards are now required to designate a lead director and to create an audit committee composed of independent nonexecutive directors (Charkham, 1994).

Yet such varieties in national practices may well diminish for two reasons. First, as equity markets internationalize, with companies seeking capital from all corners of the globe, investors predictably prefer relatively consistent director models that they believe will optimize shareholder value and performance transparency regardless of country, and investing has become far more crossborder in recent years. In 1970, foreign investors owned just 5% of Japanese shares, but by 2009 they held 26%; in the United Kingdom, foreign holdings rose from 7% of the share market in 1963 to 42% in 2008; in Korea, foreign ownership climbed from 5% in 1992 to 32% in 2009 (Federation of European Securities Exchanges, 2008). Moreover, thousands of firms have chosen to list their shares on foreign stock markets. By 2000, there were more foreign firms listed on Nasdaq and the New York Stock Exchange (NYSE) than there were German firms listed on the Deutsche Boerse.

International investors' mental models of what's right may not always be right, but that is beside the point. Consider the preference expressed by some institutional investors in the United States for a board chair who is not a sitting executive, believing that separation improves monitoring, a practice that is more common in the United Kingdom and elsewhere. Whatever the separatist penchant and however well it may have served some companies, research reveals that such a division creates no clear advantage for company performance (Baliga et al., 1996).

The diversity in directorship practices is likely to diminish, however, for a second, more factual reason. Certain practices do engender better performance, regardless of the setting, and as companies increasingly compete worldwide for customers and investors, they are likely to adopt what indeed does prove best. A case in point here is the number of directors on the board.

Research on teams' success suggests that bigger is not better when boards exceed 15 or 20 members. When the number of directors climbs far above middle range, the engagement of each diminishes, and so too does their capacity to work in unison. Studies of American and European firms reveal that smaller boards displayed stronger incentives for their chief executive, greater likelihood of dismissing an underperforming CEO, and larger market share and superior financial performance (Conyon and Peck, 1998).

It is not surprising, then, to see companies migrating worldwide toward smaller boards in search of improved performance. Directors in principle protect owner interests, direct company strategy, and select top management. In practice, they had concentrated more on executive selection and less on owner interests. But the rising power of investors has made for greater director focus on creating value and less coziness with top management. American and British directors, following the 'Anglo-Saxon' model, are already more focused on value than most. But directors in other economies can be expected to gravitate slowly toward the mantra of shareholder supremacy as well.

Shareholders, Directors, and Executives at the Table

With the rise of professional investors and their subjugation of national boundaries, those who occupy the executive suite and those who put them there are drawing far more research and policy attention, and justifiably so. In a bygone era when markets were more steady and predictable, when airline and telephone executives confidently knew what to expect next year, the identity of top management mattered less. When shareholders were far smaller and decidedly quieter, when airline and telephone directors comfortably enjoyed inconsequential board meetings, the composition of the governing body mattered less also. In recent years, however, all of this has changed in the United States and the United Kingdom with other economies close behind. As shareholders have sought to reclaim authority over what they owned, they have brought top management and company directors back in. Company strategy and financial results can no longer be understood without understanding the capabilities and organization of those most responsible for delivering them.

The Sarbanes–Oxley Act of 2002 in the United States and the introduction of stronger listing requirements by the NYSE in 2003 sought to enhance director capacity to exercise vigilance on behalf of shareholders. The Sarbanes–Oxley legislation requires that the board of publicly traded companies constitutes an audit committee of the board composed entirely of independent nonexecutive directors, and NYSE requires the same for the compensation and governance committee. The intention of their strengthening of the board committees in this way has been to improve their fiduciary oversight, and the behavioral results confirm the expected outcome. The revised stock-exchange listing requirements, for instance, have been found to significantly reduce executive compensation of listed companies – especially at firms that had least been subject to the pressures from institutional investors – implying that the new regulations served to improve the hands of directors in monitoring management (Chhaochharia and Grinstein, 2009). The Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010 in the United States pointed in the same direction, giving shareholders greater say on executive compensation and requiring more independence for board compensation committees and their advisors.

The China Securities Regulatory Commission and Chinese regulations moved in the same direction since the Company Law of 1994 and Securities Law of 1998, as had the Securities and Exchange Board of India with its Clause 49 of 2004. Their thrust has been to strengthen the monitoring role of directors, though some legislative and regulatory provisions are found to have little of the intended impact (Brown and Caylor, 2009). Yet the effort to strengthen the directors' monitoring role could also have the unintended consequence of strengthening the directors' partnering role. By pressing directors to be stronger and more independent of management for the purpose of monitoring, government legislation and state regulations could at the same time be creating a platform for directors to seek a more active partnership with executives as well (Charan et al., forthcoming).

The market is no longer viewed as so impersonal, the company no longer so isolated. Ownership and financial decisions are embedded in a complex network of working relations among money managers, stock analysts, company directors, senior executives, and state regulators. Given the right combination of features, that lattice can yield high investment returns and robust national growth; the wrong amalgam, however, can lead instead to self-dealing, frozen form, and economic stagnation. The road ahead thus depends on how well researchers understand what control mechanisms and leadership styles work well both within national settings and across cultural divides, and on the extent that top managers, company directors, and stockholders learn and apply what is best.

See also: Business Law; Capitalism: Global; Capitalism; Corporate Finance: Financial Control; Corporate Governance; Corporate Law: Publicly Traded Corporations; International Law and Treaties; International Trade: Commercial Policy and Trade Negotiations; Law and Economics: Empirical Dimensions; Law and Economics; Monetary Policy Since the 2007–2009 Financial Crisis; Regulation, Economic Theory of; Stock Market Predictability; Venture Capital.

Bibliography

Baliga, B.R., Moyer, R.C., Rao, R.S., 1996. CEO duality and firm performance: what's the fuss? Strategic Management Journal 17, 41–53.

Bebchuk, L.A., Weisbach, M.S., 2010. The state of research in corporate governance. Review of Financial Statistics 23 (3), 939–961.

Berle Jr., A.A., Means, G.C., 1932. The Modern Corporation and Private Property. Commerce Clearing House, New York.

Brown, L.D., Caylor, M.L., 2009. Corporate governance and firm operating performance. Review of Quantitative Finance & Accounting 32 (2), 129–144.

Cannella Jr., B.A., Finkelstein, S., Hambrick, D.C., 2008. Strategic Leadership: Theory and Research on Executives, Top Management Teams, and Boards. Oxford University Press, New York.

Charan, R., Carey, D., Useem, M., forthcoming. Boards that Lead, Harvard Business Publishing, Cambridge, MA.

Charkham, J., 1994. Keeping Good Company: A Study of Corporate Governance in Five Countries. Oxford University Press, New York.

Chhaochharia, V., Grinstein, Y., 2009. CEO compensation and board structure. Journal of Finance 64 (1), 231–261.

Conger, J. (Ed.), 2009. Leading Corporate Boardrooms: The New Realities, the New Rules. Jossey-Bass, San Francisco.

Conyon, M.J., Peck, S.I., 1998. Board size and corporate performance: evidence from European companies. European Journal of Finance 4, 291–304.

Davis, G.F., 2008. A new finance capitalism? Mutual funds and ownership re-concentration in the United States. European Management Review 5 (1), 11–21.

Davis, G.F., 2009. Managed by the Markets: How Finance Re-Shaped America. Oxford University Press, New York.

Davis, G.F., Kim, E.H., 2007. Business ties and proxy voting by mutual funds. Journal of Financial Economics 85, 552–570.

Easterbrook, F.H., Fischel, D.R., 1991. The Economic Structure of Corporate Law. Harvard University Press, Cambridge, MA.

Federation of European Securities Exchanges, 2008. Share Ownership Structure in Europe. Federation of European Securities Exchanges, Brussels.

Gillan, S.L., 2006. Recent developments in corporate governance: an overview. Journal of Corporate Finance 12, 381–402.

Hall, B.J., Liebman, J.B., 1998. Are CEOs really paid like bureaucrats? Quarterly Journal of Economics 113, 653–691.

Kang, E., Zardkoohi, A., 2005. Board leadership structure and firm performance. Corporate Governance: An International Review 13 (6), 785–799.

Kaplan, S.N., 1997. Corporate governance and corporate performance: a comparison of Germany, Japan, and the U.S. Journal of Applied Corporate Finance 9, 86–93.

Karpoff, J.M., Unpublished. The Impact of Shareholder Activism on Target Companies: A Survey of Empirical Findings, University of Washington.

Larcker, D., Tayan, B., 2011. Corporate Governance Matters: A Close Look at Organizational Choices and Their Consequences. Pearson, Upper Saddle River, NJ.

Lounsbury, M., Hirsch, P.M. (Eds.), 2010. Markets on Trial: The Economic Sociology of the U.S. Financial Crisis. Emerald Group Publishing, Bingley, UK.

Manne, H.G., 1965. Mergers and the market for corporate control. Journal of Political Economy 73, 110–120.

Pfeffer, J., 1981. Management as symbolic action: the creation and maintenance of organizational paradigms. In: Cummings, L.L., Staw, B.M. (Eds.), Research in Organizational Behavior, vol. 3. JAI, Greenwich, CT.

Rao, H., Greve, H.R., Davis, G.F., 2001. Fool's gold: social proof in the initiation and abandonment of coverage by Wall Street Analysts. Administrative Science Quarterly 46, 502–526.

Stern, I., Westphal, J.D., 2010. Stealthy footsteps to the boardroom: how the backgrounds of corporate leaders predict the sophistication and subtlety of their interpersonal influence tactics, and the consequences for board appointments. Administrative Science Quarterly 55, 278–319.

Thomas, A.B., 1988. Does leadership make a difference to organizational performance? Administrative Science Quarterly 33, 388–400.

Useem, M., 1996. Investor Capitalism: How Money Managers Are Changing the Face of Corporate America. Basic Books, New York.

Useem, M., 2012. The ascent of shareholder monitoring and strategic partnering: the dual functions of the corporate board. In: Clarke, T., Branson, D. (Eds.), Sage Handbook on Corporate Governance. Sage Publications, Los Angeles.

Useem, M., Karabel, J., 1986. Pathways to top corporate management. American Sociological Review 51, 184–200.

Strain Theories and Crime

Sung Joon Jang, Baylor University, Waco, TX, USA
Robert Agnew, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Strain theories state that certain strains or stressors lead to negative emotions, which create pressure for corrective action. Crime is one possible response, especially when people lack the ability to cope in a legal manner; the costs of criminal coping are low, and there is some disposition for criminal coping. Classic strain theories focus on one type of strain, the inability to achieve conventional success goals. Agnew's general strain theory (GST) focuses on a much broader range of strains and factors that influence the likelihood of criminal coping. Empirical support for GST has rejuvenated criminological research on strain.

Classic Strain Theories

Robert K. Merton (1938) developed the first major strain theory in criminology. He argues that *all* people in the United States are encouraged to strive for the cultural goal of monetary success. Lower-class individuals, however, are often prevented from achieving this goal through legitimate means. Their parents may fail to provide them with the skills and attitudes necessary for school success, they may live in communities with inferior schools, and their parents often lack the money to send them to college or set them up in business. As a consequence, lower-class individuals more often experience strain, with this strain being a function of the disjunction between their goals and the legitimate means for achieving them.

Merton states that there are several ways to cope with this strain, some of which involve crime. Individuals may attempt to achieve monetary success through illegitimate channels, such as theft, drug selling, and prostitution. They may strike out at others in their frustration. They may engage in drug use to alleviate their frustration. And they may reject the goal of monetary success and focus on the achievement of other goals, certain of which involve crime. Merton noted that most strained individuals live with their strain rather than cope through crime, and he went on to describe factors that influence the likelihood of criminal coping – such as the extent to which individuals are socialized to condemn crime.

Merton's theory led to the development of other strain theories, including that of his student, Albert Cohen. Cohen (1955) drew on Merton's theory to explain the origins of the lower-class, adolescent, male gangs that were common in the 1950s. Cohen states that lower-class individuals desire middle-class status, which includes both money and the respect of others. Lower-class people, however, have trouble achieving such status through legitimate channels. In particular, lower-class juveniles are often unprepared for school and are sometimes humiliated by their middle-class teachers and the middle-class students with whom they compete. But unlike monetary success, middle-class status is not a goal that can be readily achieved through illegitimate channels (i.e., one cannot steal middle-class status). As a result, certain lower-class adolescents cope with their strain in a different manner.

These adolescents develop an alternative status system in which they can successfully compete. Their hostility toward the middle class, among other things, leads them to value those things that the middle class rejects. The middle class values private property and polite behavior, for example, while these lower-class juveniles value theft and aggression. This alternative status system takes the form of juvenile gangs, with the gangs rejecting the cultural goal of middle-class status and substituting new, criminally oriented goals in its place. Researchers have questioned Cohen's argument that lower-class gangs fully reject conventional values and embrace crime (Agnew, 2000). But his argument that individuals sometimes adapt to strain by forming deviant subcultures appears to have some merit (Agnew, 2000).

Cloward and Ohlin (1960) developed yet another major version of strain theory. Like Cohen, they wanted to explain the formation and nature of lower-class juvenile gangs. And they too argue that the inability to achieve conventional success goals through legitimate channels contributes to the creation of gangs. But what most distinguishes their theory is the argument that any explanation of crime must consider not only the individual's opportunities to achieve conventional success goals but also the individual's illegitimate opportunities. That is, it must consider opportunities to engage in violence, theft, drug use and sales, and other crimes. Specifically, are individuals in contact with others who provide the skills, beliefs, and resources that facilitate these types of crime? This idea has found contemporary expression in routine opportunities and rational choice theories, which devote much attention to variation in the opportunities to engage in crime.

Despite their differences, the 'classic' strain theories of Merton, Cohen, and Cloward and Ohlin share in common the idea that strain involves the inability to achieve conventional success goals through legitimate means. These theories were perhaps the dominant theories of crime in the 1950s and 1960s. They also had a major impact on public policy, being an inspiration for the War on Poverty (Agnew, 2000). The War on Poverty was designed to make it easier for people to achieve material success through legal channels. It involved programs such as Project Headstart, a preschool enrichment program, and Job Corps, a jobs training program. But by the 1970s and 1980s, classic strain theory was being heavily criticized

and came close to being abandoned (e.g., Agnew, 2000; Agnew et al., 1996; Kornhauser, 1978).

Classic strain theory predicts that crime is concentrated in the lower classes, but many self-report studies in the 1960s and beyond found that the relationship between class and crime was weak. Relatedly, research revealed the existence of much white-collar and corporate crime. Furthermore, crime rapidly increased during the 1960s, despite increasing economic prosperity and a sustained effort to reduce poverty. Finally, several studies attempted to test classic strain theory by examining the relationship between aspirations or ideal goals and expectations or expected levels of goal achievement. These studies focused on educational and occupational goals, and argued that if classic strain theory is correct, crime should be highest among those with high aspirations but low expectations (i.e., those who desire a lot but expect little). Instead, crime was highest among those with low aspirations and expectations (e.g., Agnew, 2000; Agnew et al., 1996; Kornhauser, 1978).

Classic strain theory, however, has experienced a revival in recent years. It has been argued that the theory can explain higher class crime, since success goals are relative rather than absolute. That is, people look to others in their reference group when deciding what goals to pursue and evaluating their levels of goal achievement. Those in higher social classes often experience strain because they compare themselves to more privileged others (Passas and Agnew, 1997). This argument also helps explain the increase in crime during the 1960s. While many people improved their economic situation, others did not. These others may have experienced intense strain as they witnessed the economic progress of their friends and neighbors.

Finally, those studies examining the disjunction between aspirations and expectations were criticized for not focusing on economic goals (Agnew, 2000; Agnew et al., 1996; Passas and Agnew, 1997). Also, the inability to achieve aspirations or ideal goals was said to be a poor measure of strain. Aspirations have something of the utopian in them, so failure to achieve them does not generate much strain. More recent studies have found that dissatisfaction with one's monetary status and the experience of economic problems, such as trouble paying bills, are related to crime; these findings are supportive of classic strain theory (e.g., Agnew et al., 1996, 2008). Contemporary research on stress in psychology and sociology, however, suggested that the focus of classic strain theory on the inability to achieve conventional success goals was too narrow.

General Strain Theory

Building on his earlier works (Agnew, 1985, 1989), Agnew (1992) proposed a general strain theory (GST) of crime and delinquency, which focuses on a much broader range of strains. But like the classic strain theorists, he argued that these strains pressure individuals into engaging in crime through the negative emotions they generate. He also took an inclusive approach to other theories of crime, especially social control and social learning theories (Akers, 1985; Hirschi, 1969), by using their key concepts to help explain why some people are more likely than others to cope with strains through crime. Agnew's (1992) introduction of GST rejuvenated criminological research on strain, which had been out of favor for about two decades.

Foundation for a GST

Agnew's (1992) GST was built on his critical evaluations of Merton's (1938) classic strain theory and its revisions of Cohen (1955) and Cloward and Ohlin (1960). Presenting GST as another revision at the social psychological level, Agnew (1995a) argued that previous theories failed to realize the full potential or unique contribution that strain theory could make to the explanation of crime. To this end, Agnew suggested that strain theory should broadly define strain, place more explicit emphasis on the mediating role played by negative emotions, and identify a broader range of factors that influence the likelihood of criminal coping.

Strain, Negative Effect, and Conditioning Factors

Social control theory (Hirschi, 1969) explains crime in terms of the absence of positive relationships with conventional others, such as parents and teachers. Social learning theory (Akers, 1985) explains crime in terms of positive relationships with deviant others or at least those who reinforce deviance. GST, by contrast, posits that crime is a consequence of "negative relationships with others", or strain (Agnew, 1992: p. 48). Strain includes not only negative interpersonal relationships but also situations, conditions, and events that are objectively expected or subjectively perceived to be negative (e.g., parental abuse, chronic unemployment, homelessness, discrimination, and the immediate need for much money). Specifically, Agnew proposes three types of strain: strain as the actual or anticipated failure to achieve positively valued goals, which includes the classic conception of strain (i.e., the inability to achieve monetary goals); strain as the actual or anticipated removal of positively valued stimuli (e.g., the violent death of a close friend); and strain as the actual or anticipated presentation of negative stimuli (e.g., child abuse).

Next, GST proposes that strain generates negative effective states, such as anger and frustration, which create pressure for corrective action, including crime. This focus on negative effect helps distinguish GST from other theories. The effects of many types of strain on crime, such as child abuse and unemployment, can be explained using social learning and control theories (Agnew, 1995b). For example, child abuse provides the juvenile with a violent model, while unemployment weakens one's stake in conformity. But GST is the only theory to argue that these variables impact crime at least partly through the negative emotions they create (see Agnew, 2001, 2006 for other links between strain and crime).

GST focuses on the negative emotion of anger, which is said to energize individuals for action, reduce concern for the consequences of one's behavior, and create a desire for revenge. Agnew (1995b: p. 384) suggests that researchers should examine both "long-term or chronic anger and the frequency of angry episodes," that is, trait anger and state anger. State anger is a transitory emotional condition due to annoying or frustrating situations, whereas trait anger refers to an individual's disposition to perceive a range of situations as annoying or frustrating and respond to such situations with marked elevations in state anger. Thus, individuals with high levels of

trait anger are more likely to experience sustained elevation in state anger in reaction to strain than their peers with low levels of the trait. Because of this trait–state causality, while GST focuses on state anger, trait anger is also relevant to the theory. In addition, state anger can further be conceptualized and measured as 'situational anger,' produced by a particular strain that functions as a 'situational event' (Agnew, 1992: p. 60).

Although, for Agnew (1992: p. 59), anger is the "most critical emotional reaction for the purpose of the general strain theory", other negative emotions, such as depression and anxiety, are important as well. In fact, Agnew (1992) proposes that different types of negative emotions may lead to different crimes. For example, the inner-directed emotion of depression may be more conducive to internalized criminal acts, such as drug use, while the outer-directed emotion of anger may be more conducive to externalizing acts such as aggression.

Finally, while classic strain theories tend to be limited in explaining why not all strained individuals turn to crime for strain adaptation, Agnew (1992) addresses this issue by identifying a range of factors that affect the likelihood of criminal versus noncriminal coping. In particular, these factors affect a person's ability to engage in legal and criminal coping, costs and benefits in regard to criminal and legal coping, and disposition for criminal coping. They include both individual characteristics (e.g., self-efficacy) and environmental factors (e.g., social support).

Empirical Tests of GST

Early tests of GST generally provided empirical support for the theory. For example, Agnew and White (1992) found strain to have positive contemporaneous relationships with adolescent delinquency and drug use, while their lagged counterparts were significant only for delinquency, controlling for the previous measure of adolescent deviance. This first empirical test of GST was replicated by Paternoster and Mazerolle (1994), who found a positive strain–delinquency association to be not only direct but also indirect via control and social learning variables. Other researchers also reported positive (both direct and indirect) relationships between strain and crime based on longitudinal as well as cross-sectional data (Agnew, 1998; Hoffmann and Cerbone, 1999). While these findings were consistent with GST, the early studies were limited because of their omission of negative effect, and thus they could not attribute the positive association fully to GST (Agnew, 1995b).

On the other hand, Agnew (1985) found trait anger to explain most of the effects of strain on delinquency, whether confrontational (interpersonal aggression) or escapist (being late for or skipping class or school). Similarly, Brezina (1996) found strain to have significant effects on trait measures of resentment, anxiety, and depression as well as anger, while he did not examine the effects of the negative emotions on delinquency. In a later study, however, he found anger–delinquency as well as strain–anger associations to be positive (Brezina, 1998).

Other researchers examined state or situational as well as trait emotions because GST focuses on the former rather than the latter in explaining the criminogenic influence of strain, while recognizing the theoretical relevance of emotional traits to GST (Capowich et al., 2001; Mazerolle et al., 2000, 2003; Piquero and Sealock, 2000; Jang and Johnson, 2003; Aseltine et al., 2000; Broidy, 2001). Their findings together tend to show that state and, to a greater extent, situational emotions are more likely than trait emotions to mediate the influence of strain on criminal and deviant coping. In addition, Jang and Johnson (2003) tested the hypothesis about differential relationships between inner- and outer-directed negative emotions and deviance, and they reported supportive evidence: that is, anger was found to be more strongly associated with interpersonal aggression than drug use, while depression and anxiety were more strongly associated with drug use than aggression. In sum, prior research tends to show that negative emotions, however measured, are positively associated with both strain and crime, as GST posits.

Finally, studies tested the hypothesis that several factors condition the effects of strains and negative emotions on crime (Agnew and White, 1992; Paternoster and Mazerolle, 1994; Capowich et al., 2001; Aseltine et al., 2000). These studies examined conditioning variables such as self-esteem, self-efficacy, conventional bonding, and delinquent peer association. Despite some evidence that conditioning factors partly explain why strain does not always lead to crime, support for the hypothesized conditioning factors has been limited. This may be due to the difficulty of detecting interaction effects in survey research or the fact that researchers usually consider one conditioning factor at a time, with other factors controlled (Agnew, 2006). It may also be that individuals have to score high on several conditioning factors before criminal coping is likely (see Mazerolle and Maahs, 2000). Furthermore, researchers should examine other, understudied factors, like biopsychological traits (Agnew et al., 2002; Eitle and Taylor, 2011), religion (Jang and Johnson, 2003), and social support (Jang and Lyons, 2006).

Extensions of GST

After GST was introduced, it was extended to explain gender differences in crime (Broidy and Agnew, 1997), stability and change in crime over the life course (Agnew, 1997), community differences in crime (Agnew, 1999), and other patterns of offending (Agnew, 2006).

Gender and GST

Broidy and Agnew (1997) extended GST to explain gender differences in crime because the original GST does not offer an immediate explanation of the fact that, while males and females are not necessarily different in the amount of strain, females tend to report *higher* levels of negative emotions but *lower* levels of crime than males. To address the issue, they advanced three propositions. First, they propose that females are at least as strained as males but commit less crime partly because they are less likely than men to experience criminogenic types of strain. Second, they suggest, whereas females tend to report higher levels of anger than males, their anger is likely to be accompanied by less criminogenic emotions (e.g., fear, anxiety, guilt, and shame) and thus is mitigated by the concurrent nonangry emotions; and, as a result, females are less likely than males to turn to crime for strain adaptation. Third, they propose that females are less likely to turn to crime and deviance in response to strain partly because they are more likely to have prosocial

conditioning factors, such as coping skills and conventional social support.

Prior research tends to provide evidence of gender differences in types of strain (Agnew and Brezina, 1997; Hay, 2003; Mazerolle, 1998; Kaufman, 2009). For example, Jang (2007) found African-American women to be more likely to report strains related to physical health, interpersonal relations, and gender roles in the family, and less likely to mention racial and job strain than their male counterparts. Previous research also reports gender differences in the relationships between types of strains and emotional responses to strain (Broidy, 2001): for example, females' strains were more likely to be associated with inner-directed (e.g., depression and anxiety) than outer-directed emotions (e.g., anger), and vice versa was true for males' strains (Jang, 2007; Jang and Johnson, 2005). Finally, while the third proposition has been tested less often than the other two, Jang and Johnson (2005) found the distress-buffering effects of religiosity as a conditioning factor to be larger for women than men, which partly explains gender differences in deviant coping.

Life-Course Criminology and GST

Pointing out that life-course criminology was being dominated by control and social learning as well as biological and psychological theories, Agnew (1997, 2006) suggested that GST has a unique role in explaining stability and change in crime over the life course. First, GST posits that the individual traits of negative emotionality and low constraint explain the stability of crime. These traits increase exposure to criminogenic strains, contribute to negative emotional reactions, and foster criminal coping. For example, people high in negative emotionality (irritability) are more likely to elicit negative treatment from others, interpret such treatment as aversive (which increases subjective strain), blame others for adversity (which increases anger), and have limited conventional problem-solving skills (which increases the likelihood of crime being used as a coping strategy). These processes not only increase the probability of criminal coping but also help maintain and enhance the trait of negative emotionality, which in turn further stabilizes the pattern of criminal behavior over the life course.

Second, Agnew suggests that patterns of offending over the life course should at least roughly track levels of strain, negative effect, and conditioning factors. For example, according to GST, the 'age–crime curve' of delinquency, which increases, reaches a peak, and then decreases during adolescence and young adulthood, is likely to be in part a function of changes in the extent of one's negative relations with others (strain) and the tendency to interpret such relations as aversive, thereby experiencing anger and frustration, and cope with adversity through delinquency.

Although life-course research on GST is limited, findings tend to support the theory (Agnew, 1997, 2006). Hoffmann and Cerbone (1999) found a positive association between the number of stressful life events that adolescents experienced over time and the growth rate of delinquent involvement between early and middle adolescence, and the strain effect on the escalation of delinquency was found to not vary by potential conditioning factors, including self-efficacy and self-esteem. While not intended to test GST, Broidy et al.'s (2003) study revealed a pattern relevant to GST: chronic physical aggression during the elementary school years – a proxy of the aggressiveness trait – was positively associated with both violent and nonviolent forms of delinquency during adolescence. More recently, Eitle (2010) found changes in strain to be positively associated with changes in criminal activity among young adult males, which implies the effect of decreased strain on criminal desistance (see also Gunnison and Mazerolle, 2007); whereas Eitle et al. (2010) reported the criminogenic effect of strain over the life course, between early adolescence and young adulthood.

Macrolevel GST

Drawing on his original microlevel theory, Agnew (1999) proposed macrolevel GST (MGST) to explain community differences in crime, which have been studied based mostly on social disorganization and, to a lesser extent, subcultural deviance as well as relative deprivation theory. His proposed MGST model includes not only the key GST variables – strain, negative effect, and conditioning factors – at an aggregate level but also the exogenous variable of 'community characteristics,' like economic deprivation, inequality, overcrowding, high population mobility, and a high percentage of nonwhite members. In addition to increasing community-level strain, these characteristics tend to contribute to the selection and retention of strained individuals because they, especially those who are economically strained, are likely to move into deprived communities where they can afford to live and their migration would meet little resistance. Thus, the community characteristics in the MGST model help control for self-selection bias as well as specify the source of community strain.

Community strain includes all three types of strain in GST: the failure to achieve positively valued goals, the loss of positive stimuli, and the presentation of negative stimuli (e.g., goal blockage, family disruption, incivilities, relative deprivation, and frequent interactions with angry or frustrated people). MGST posits that community strain generates community-level (or aggregate) negative effect, which in turn increases community crime rates, while conditioning factors moderate the positive associations between community strain and negative effect and between community negative effect and crime rate. It also posits that community crime tends to increase community negative effect, proposing reciprocal causal relationships between the two.

Research on MGST has been limited partly because data to test the theory are relatively unavailable. Specifically, unlike the measures of community characteristics, community strain, as well as community crime that can be drawn from the US Census and FBI Uniform Crime Reports data, those of community negative effect and conditioning factors are not easy to construct because it requires aggregating individual-level data on those variables from a survey conducted in multiple community-level units, each of which includes a large enough number of respondents for the aggregation. Despite this operational challenge, several studies have been conducted to test MGST.

Based on national survey data, Brezina et al. (2001) found a positive association between school-level anger among students and fighting or arguing with other students, as MGST posits. Next, although failing to include community negative effect, Warner and Fowler (2003) reported not only positive associations between the Census-based variables of deprived neighborhood characteristics and aggregated measures of

experienced and vicarious strain (violent behaviors witnessed or heard in the neighborhood) but also an aggregated measure of social support and social capital conditioning the effects of the strain on violence. Finally, although not intended to test MGST, Stockdale et al. (2007) found community strain to be positively associated with alcohol, drug, and mental health disorders, and the positive association was conditioned by community-level social support.

Looking Back and Looking Forward

Based on the first 10–15 years of its empirical tests, Agnew (2001, 2006) critically assessed GST for further conceptual specifications and theoretical elaboration. First, distinguishing between objective and subjective strains, he proposed strains to be the most criminogenic when they (1) are seen as unjust, (2) are high in magnitude, (3) are associated with low social control, and (4) create some pressure or incentive for criminal coping. Examples include child abuse and neglect, abusive peer or marital relations, criminal victimization, and racial and gender discrimination. Agnew (2002) further divided strains into experienced, vicarious, and anticipated strains, arguing that while experienced strains should be most strongly related to crime, vicarious and anticipated strains may contribute to crime in some circumstances (e.g., Jang and Rhodes, 2012).

Second, Agnew calls for more research examining understudied types of strains (e.g., work in the secondary labor market, and discrimination), the 'causal' effect of strain based on panel data with a short interval, and indirect effects of strains on crime via social control, social learning, and personality variables. Third, he also points out a relative lack of research not only on emotional states (as opposed to traits) but also on other emotions than anger. Fourth, in response to the limited evidence for conditioning factors, Agnew suggests that studies based on other methods than survey (e.g., experimental, vignette, and observational studies) should be conducted given the difficulty of detecting interaction effects with survey data. In addition, he suggests that researchers should examine conditioning factors in combination rather than in isolation from one another.

Finally, Agnew stresses that GST is more than a microlevel theory explaining individual differences in crime, and thus needs to be applied to account for group (other than gender) or community differences in crime. For example, GST can shed light on racial and ethnic differences in crime (e.g., Kaufman et al., 2008). Furthermore, research opportunities for cross-national studies to extend GST to account for societal or cultural differences in crime are another area where more research is needed (see Agnew, 2006). These opportunities and the topics discussed here for future research confirm that GST is still "a work in progress" (Agnew, 2006: p. 202).

See also: Crime and Gender; Crime, Sociology of; Delinquency, Sociology of; Life Course Criminology.

Bibliography

Agnew, R., 1985. A revised strain theory of delinquency. Social Forces 64 (1), 151–167.

Agnew, R., 1989. A longitudinal test of the revised strain theory. Journal of Quantitative Criminology 5 (4), 373–387.

Agnew, R., 1992. Foundation for a general strain theory of crime and delinquency. Criminology 30 (1), 47–88.

Agnew, R., 1995a. The contribution of social–psychological strain theory to the explanation of crime and delinquency. In: Adler, F., Laufer, W.S. (Eds.), The Legacy of Anomie Theory. Transaction Publishers, New Brunswick, NJ, pp. 113–137.

Agnew, R., 1995b. Testing the leading crime theories: an alternative strategy focusing on motivational processes. Journal of Research in Crime and Delinquency 32 (4), 363–398.

Agnew, R., 1997. Stability and change in crime over the life course: a strain theory explanation. In: Thornberry, T.P. (Ed.), Developmental Theories of Crime and Delinquency. Transaction Publishers, New Brunswick, NJ, pp. 101–131.

Agnew, R., 1998. The approval of suicide: a social–psychological model. Suicide and Life-Threatening Behavior 28 (2), 205–225.

Agnew, R., 1999. A general strain theory of community differences in crime rates. Journal of Research in Crime and Delinquency 36 (2), 123–155.

Agnew, R., 2000. Sources of criminality: strain and subcultural theories. In: Sheley, J.F. (Ed.), Criminology: A Contemporary Handbook. Wadsworth, Belmont, CA, pp. 349–371.

Agnew, R., 2001. Building on the foundation of general strain theory: specifying the types of strain most likely to lead to crime and delinquency. Journal of Research in Crime and Delinquency 38 (4), 319–361.

Agnew, R., 2002. Experienced, vicarious, and anticipated strain: an exploratory study on physical victimization and delinquency. Justice Quarterly 19 (4), 603–632.

Agnew, R., 2006. Pressured into Crime: An Overview of General Strain Theory. Roxbury Publishing Company, Los Angeles, CA.

Agnew, R., Brezina, T., 1997. Relational problems with peers, gender, and delinquency. Youth & Society 29 (1), 84–111.

Agnew, R., Brezina, T., Wright, J.P., Cullen, F.T., 2002. Strain, personality traits, and delinquency: extending general strain theory. Criminology 40 (1), 43–71.

Agnew, R., Cullen, F.T., Burton Jr., V.S., Evans, T.D., Dunaway, R.G., 1996. A new test of classic strain theory. Justice Quarterly 13 (4), 681–704.

Agnew, R., Matthews, S.K., Bucher, J., Welcher, A.N., Keyes, C., 2008. Socioeconomic status, economic problems, and delinquency. Youth & Society 40 (2), 159–181.

Agnew, R., White, H.R., 1992. An empirical test of general strain theory. Criminology 30 (4), 475–499.

Akers, R.L., 1985. Deviant Behavior: A Social Learning Approach, third ed. Wadsworth, Belmont, CA.

Aseltine Jr., R.H., Gore, S., Gordon, J., 2000. Life stress, anger and anxiety, and delinquency: an empirical test of general strain theory. Journal of Health and Social Behavior 41 (3), 256–275.

Brezina, T., 1996. Adapting to strain: an examination of delinquent coping responses. Criminology 34 (1), 39–60.

Brezina, T., 1998. Adolescent maltreatment and delinquency: the question of intervening processes. Journal of Research in Crime and Delinquency 35 (1), 71–99.

Brezina, T., Piquero, A.R., Mazerolle, P., 2001. Student anger and aggressive behavior in school: an initial test of Agnew's macro-level strain theory. Journal of Research in Crime and Delinquency 38 (4), 362–386.

Broidy, L.M., 2001. A test of general strain theory. Criminology 39 (1), 9–36.

Broidy, L.M., Agnew, R., 1997. Gender and crime: a general strain theory perspective. Journal of Research in Crime and Delinquency 34 (3), 275–306.

Broidy, L.M., Nagin, D.S., Tremblay, R.E., Bates, J.E., Brame, B., Dodge, K.A., Fergusson, D., Horwood, J.L., Loeber, R., Laird, R., Lynam, D.R., Moffitt, T.E., Pettit, G.S., Vitaro, F., 2003. Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: a six-site, cross-national study. Developmental Psychology 39 (2), 222–245.

Capowich, G.E., Mazerolle, P., Piquero, A., 2001. General strain theory, situational anger, and social networks: an assessment of conditioning influences. Journal of Criminal Justice 29 (5), 445–461.

Cloward, R.A., Ohlin, L.E., 1960. Delinquency and Opportunity. Free Press, New York.

Cohen, A.K., 1955. Delinquent Boys. Free Press, New York.

Eitle, D., 2010. General strain theory, persistence, and desistance among young adult males. Journal of Criminal Justice 38 (6), 1113–1121.

Eitle, D., Taylor, J., 2011. General strain theory, BIS/BAS levels, and gambling behavior. Deviant Behavior 32 (1), 1–37.

Eitle, D., Taylor, J., Pih, K.K., 2010. Extending the life-course interdependence model: life transitions and the enduring consequences of early self-derogation for young adult crime. Youth & Society 41 (4), 519–545.

Gunnison, E., Mazerolle, P., 2007. Desistance from serious and not so serious crime: a comparison of psychosocial risk factors. Criminal Justice Studies 20 (3), 231–253.

Hay, C., 2003. Family strain, gender, and delinquency. Sociological Perspectives 46 (1), 107–135.
Hirschi, T., 1969. Causes of Delinquency. University of California Press, Berkeley, CA.
Hoffmann, J.P., Cerbone, F.G., 1999. Stressful life events and delinquency escalation in early adolescence. Criminology 37 (2), 343–373.
Jang, S.J., 2007. Gender differences in strain, negative emotions, and coping behaviors: a general strain theory approach. Justice Quarterly 24 (3), 523–553.
Jang, S.J., Johnson, B.R., 2003. Strain, negative emotions, and deviant coping among African Americans: a test of general strain theory. Journal of Quantitative Criminology 19 (1), 79–105.
Jang, S.J., Johnson, B.R., 2005. Gender, religiosity, and reactions to strain among African Americans. The Sociological Quarterly 46 (2), 323–357.
Jang, S.J., Lyons, J.A., 2006. Strain, social support, and retreatism among African Americans. Journal of Black Studies 37 (2), 251–274.
Jang, S.J., Rhodes, J.R., 2012. General strain and non-strain theories: a study of crime in emerging adulthood. Journal of Criminal Justice 40 (3), 176–186.
Kaufman, J.M., 2009. Gendered responses to serious strain: the argument for a general strain theory of deviance. Justice Quarterly 26 (3), 410–444.
Kaufman, J.M., Rebellon, C.J., Thaxton, S., Agnew, R., 2008. A general strain theory of racial differences in criminal offending. Australian & New Zealand Journal of Criminology 41 (3), 421–437.
Kornhauser, R.R., 1978. Social Sources of Delinquency: An Appraisal of Analytic Models. University of Chicago Press, Chicago, IL.
Mazerolle, P., 1998. Gender, general strain, and delinquency: an empirical examination. Justice Quarterly 15 (1), 65–91.
Mazerolle, P., Burton, V.S., Cullen, F.T., Evans, T.D., Payne, G.L., 2000. Strain, anger, and delinquent adaptations specifying general strain theory. Journal of Criminal Justice 28 (2), 89–101.
Mazerolle, P., Maahs, J., 2000. General strain and delinquency: an alternative examination of conditioning influences. Justice Quarterly 17 (4), 753–778.
Mazerolle, P., Piquero, A.R., Capowich, G.E., 2003. Examining the links between strain, situational and dispositional anger, and crime: further specifying and testing general strain theory. Youth & Society 35 (2), 157.
Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3 (5), 672–682.
Passas, N., Agnew, R. (Eds.), 1997. The Future of Anomie Theory. Northeastern University Press, Boston, MA.
Paternoster, R., Mazerolle, P., 1994. General strain theory and delinquency – a replication and extension. Journal of Research in Crime and Delinquency 31 (3), 235–263.
Piquero, N.L., Sealock, M.D., 2000. Generalizing general strain theory: an examination of an offending population. Justice Quarterly 17 (3), 449–484.
Stockdale, S.E., Tang, L., Zhang, L., Belin, T.R., Wells, K.B., 2007. The effects of health sector market factors and vulnerable group membership on access to alcohol, drug, and mental health care. Health Services Research 42 (3), 1020–1041.
Warner, B.D., Fowler, S.K., 2003. Strain and violence: testing a general strain theory model of community violence. Journal of Criminal Justice 31 (6), 511–521.

Strategic Adaptation

Torben J Andersen, Center for Global Strategic Responsiveness, Copenhagen Business School, Frederiksberg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article provides an overview of theoretical contributions that have influenced the discourse around strategic adaptation including contingency perspectives, strategic fit reasoning, decision structure, information processing, corporate entrepreneurship, and strategy process. The related concepts of strategic renewal, dynamic managerial capabilities, dynamic capabilities, and strategic response capabilities are discussed and contextualized against strategic responsiveness. The insights derived from this article are used to outline the contours of a dynamic process of strategic adaptation. This model incorporates elements of central strategizing, autonomous entrepreneurial behavior, interactive information processing, and open communication systems that enhance the organization's ability to observe exogenous changes and respond effectively to them.

Introduction

The topic of strategic adaptation is as old as the strategy field and can most likely be traced in the earlier management literature. The idea that organizations must be able to develop strategies that suit the circumstances seems so obvious that we hardly think about it. With that also follows that the ability to adapt the strategy as the surrounding business environment changes is an essential managerial concern. The common view is that managers try to develop a proper strategy for the environmental context and design an organizational structure to match it. When the environmental conditions change then the strategy and structure will also have to be adapted accordingly. As argued by Ansoff and Brandenburg (1971: B-706): "Rapid changes in customer needs and product-process technologies require organizations that are flexible and responsive." Therefore, it is hardly surprising that strategic management in its essence relates to the process of adapting the strategy to dynamic changes in the surrounding environment.

Substantial energy has been devoted to uncover how the business environment is changing. It has been characterized as, e.g., dynamic, turbulent, and hypercompetitive with frequent technological changes, shorter product cycles, and ongoing innovation that continue to challenge existing competitive advantages (Thomas and D'Aveni, 2009). The environment is becoming more unpredictable and calls for strategic response capabilities (Bettis and Hitt, 1995), adaptive capabilities (Volberda, 1996), and dynamic capabilities (Teece et al., 1997; Teece, 2007) to facilitate adaptation to ongoing changes.

So much has been written about this theme over the years that it is literally impossible to cover it all comprehensively in a single article. It might also be rather futile to give an encyclopedic account of the entire literature since much of the discourse revolves around the same basic ideas but often with new labels and under new headings. The area subscribes substantially to differentiating concepts and terminologies that can lead into something of a semantic quagmire. Nonetheless, there are traces of development in central elements of strategic adaptation that seems to evolve around the firm's ability to respond to change. This article will try to uncover these subtle, possibly subjective, observations more or less in chronological order and use them to outline an evolutionary path that may be underway.

The article will first discuss the contingency perspective that implicitly promotes an underlying idea about fit. Then corporate entrepreneurship is presented as driver of strategic renewal and adaptation. Finally, the more recent concepts of dynamic managerial capabilities, dynamic capabilities, and strategic responsiveness are used to theorize about the dynamic process of strategic adaptation.

Environmental Contingency and Fit Perspectives

The effectiveness of managerial decisions and actions depends on the situation and the circumstances around them including the conditions of the firm's internal and external environment. This is an underlying premise of contingency theory. However, the early focus on planning processes in strategy-paid lip service to this premise at least to some extent, which led to a move away from strategy process toward an emphasis on strategy content and the environmental conditions that affect strategic choices (Hofer, 1975). This drive toward a focus on content considered organizational characteristics and environmental circumstances that could be "of major importance in determining the content of specific business strategies" (Hofer, 1975: 792–793). These included, for example, economic, demographic, social, and political/legal conditions all of which today are recognized as essential elements of an environmental analysis.

In considering the effect of environmental conditions was also an implicit concern for the ability to deal with external changes since the environment by experience is not stable in the longer run. In fact, it can be considered of the uttermost importance and Chakravarthy (1982: 35) went as far as arguing that "the essence of management is coping with change." He conceived of three types of change that would require three states of adaptation: an unstable, a stable, and a neutral state. In the unstable state the firm buffers itself from environmental events. In a stable state the firm operates under benign conditions and can react to changes. In a neutral state the firm anticipates the changes, absorbs them as they arise, and adapts

the organization to accommodate the new conditions (Chakravarthy, 1982).

This analysis assumes a better match between the adaptive state and choices between mechanistic and organic structures. An organic structure with little formality and much decentralization and information-sharing works better in a neutral state where the organization has a high degree of adaptability. It might not be as simple though because the strategic choices made between organic and mechanistic structures can be achieved in many different ways (equifinality) and the environment can be more or less forgiving of the underlying decisions. That is, the strategic choice and the environmental pressure constitute two distinct dimensions that can frame different adaptation typologies (Hrebiniak and Joyce, 1985). Hence, conscious structural choices can achieve adaptation either by design or by chance, or adaptation can be constrained if the firm is not simply selected out by competitive forces.

The adaptation contingencies and the implied performance effects are often taken for granted as automatic outcomes. As Drazin and Van de Ven (1985: 514) state so eloquently "it appears that our concepts of fit are drawn from a general and often implicit pool of domain assumptions and methodological conventions." The core argument is that organizational structure and processes should fit the environmental context for the firm to survive. But fit is not a clear-cut concept and can be seen from at least three increasingly complex views: (1) congruence between structure and context, (2) match between pair wise structure–context factors, and (3) consistency between multiple contingencies and multiple structural elements. This alludes to the fact that dealing with turbulent environments requires more complex responsive organizational structures to achieve high performance (Drazin and Van de Ven, 1985).

Structural change is not just an automatic response but rather evolves from a structural adjustment process of regaining fit as the environment changes (Donaldson, 1987). Contingency determinism assumes that changes in external conditions lead to changes in the structural variables directly whereas structural adjustment alludes to an underlying adjustment process. Shifts in environmental contingencies will create a disequilibrium that leads to declining effectiveness. This in turn will create internal pressures for change that eventually drive actions of structural adaptation aimed at regaining fit.

Strategic Fit

The concept of fit between the environment, strategy, and the supportive organizational structure is referred to as strategic fit where performance is seen as partially determined by a match between firm strategy, structure, and the environment. In the words of Jennings and Seaman (1994: 461): "organizations having an optimal strategy-structure match will give higher performance." It is formalized in strategic fit theory (Fiegenbaum et al., 1996) arguing that higher performance is achieved when the strategy content and the organizational structure are aligned with the external conditions. Hence, dynamic environments require adaptive firms associated with an organic structure and they perform better if they pursue a prospector strategy at the same time (Jennings and Seaman, 1994). In contrast, nonadaptive firms in more stable environments perform better when they have a mechanistic structure and follow a defender strategy.

The strategic fit concept can be extended to consider aspects of corporate culture where proper aims and values are seen as equally important. Scholz (1987: 78) refers to strategic fit as the situations "in which all the internal and external elements relevant for a company are in line with each other and with the corporate strategy." He argues that this entails *intrastrategy* fit where the strategy elements are consistent, *intrasystem* fit with congruence between structural elements, *strategy-system* fit for consistency between strategy and structural elements, and *internal–external* fit where the structural system is tailored to fit the environment. However, corporate culture forms the human behaviors and intensions in strategic actions and, therefore, should also suit the chosen strategy. Scholz (1987: 80) refers to corporate culture as "the implicit, invisible, intrinsic, and informal consciousness of the organization which guides the behavior of individuals." He further argues that it can be helpful to aggregate various dimensions of corporate culture to form generic cultural types and use these typologies to assess whether the culture is a good match to the strategy (Scholz, 1987).

Organizational Perspectives

The vast literature on new organizational forms points toward a need for responsive decision structures. A centralized structure confines decisions to top management whereas decentralization authorizes managers at lower levels to take responsive initiatives. The structural solution to deal with increasing information flows and pressures from organizational dependencies seems to be a move of decision power down the organization closer to where the relevant information is located (Child and McGrath, 2001; Volberda, 1996). Hence, a broad literature advocates decentralization under pervasive uncertainty.

It is generally recognized that adaptive strategy-making also needs integrative organizational features. Effective organizations engage in central integrative processes with dispersed decisions embedded in more rigid organizational structures (Hill et al., 2000). The real challenge, therefore, seems to be to create the right combination of strategic control features and experimentation within the organization.

Adaptive Organizations

In empirical studies of strategy development patterns, Miller and Friesen (1980) observed that extreme periods of stable momentum and dramatic revolution seem to interchange reflecting a certain resistance to reverse strategic direction. They further noted that continuous changes are more common than fundamental reversals and that the interdependencies between structural variables lead to ongoing adjustments when the smaller adjustments are carried out (Miller and Friesen, 1980).

Organizations seem to go through major epochs of rather stable strategy–structure constellations broken off by shorter periods of quantum change and strategic reorientation (Mintzberg and Waters, 1982). This phenomenon often referred to as punctuated equilibrium consists of relatively long periods of equilibrium intercepted by shorter periods of revolutionary

adaptations. The punctuation of stable periods can be caused by major environmental changes, such as, the emergence of new disruptive technologies (Romanelli and Tushman, 1994). Hence, periods of revolutionary adaptation seems to occur when significant changes happen in the environment that disrupts the prior conditions for a structural equilibrium and calls for new structure forms that can respond to those changes.

This may speak for advantages of continuous changes that can modify the strategy–structure relationship on an interdependent basis as the organization gradually adapts to accommodate ongoing changes in the environment. To this effect Mintzberg and Waters (1985) consider a more responsive low-level emergent approach to strategy making in conjunction with a high-level intended planning approach. They argue that emergent strategy means that management is "open, flexible and responsive, in other words, willing to learn" and that "strategy formation walks on two feet, one deliberate, the other emergent" (Mintzberg and Waters, 1985: 271). This dual approach may enable top management to direct strategic actions toward realization of intended outcomes while at the same time leaving room for organizational members to respond to the environmental changes as they unfold.

A population ecology perspective conceive of internal and external selection mechanisms as low-level business initiatives prove their commercial viability and eventually may grow into corporate business saviors when market changes may require a new strategic focus. Burgelman (1991: 240) argue that these "strategic initiatives come about, in part, as the result of individual strategists seeking expression of their special skills and career advancement through the pursuit of different types of strategic initiatives." A number of internal mechanisms, including decision rules, hurdle rates, behavioral norms, etc., determine the advancement of initiatives that can develop into viable businesses. The low-level strategic initiatives can be venues to shape technologies and products that address new customers while developing essential capabilities that can support future adaptive moves. Hence, the process is a way to obtain strategic renewal through internal experimentation and selection.

Various studies illustrate the significance of decentralized initiatives developed within the organization as the source of opportunities that can have potential strategic implications (Burgelman, 1988). The so-called Bower–Burgelman process model (Burgelman, 1996) depicts strategy making as a function of business initiatives pursued by low-level managers that, if they are successful, can be promoted for consideration as part of the official corporate strategy. Top management influences strategy by formulating policies and organizing the structure, processes, and systems that support the progression of business initiatives. Hence, responsive strategy making relies on effective decision processes and communication channels between hierarchical levels to determine which initiatives may successfully adjust the corporate strategy.

Corporate Entrepreneurship

Entrepreneurial initiatives try to exploit opportunities and create capabilities that respond to environmental developments where "corporate entrepreneurial initiatives are depicted as casually adjacent to new and more favorable strategic positions" (Garrett and Covin, 2007). Entrepreneurial opportunities can be pursued as intended and emergent initiatives where strategic goals are targeted while leaving room for new relevant business leads. The ability to seize new markets and capitalize on technology development to gain adapted strategic positions requires leadership support, resource availability, and decision-making autonomy.

If the firm maintains a reasonable fit between the organization and the environment it will perform well and generate slack resources (Chakravarthy, 1982), which in turn provides discretionary means for ongoing entrepreneurial initiatives. That is, an entrepreneurial model of strategic adaptation can lead in "a virtuous circle to the generation of rents that make the pursuit of future entrepreneurial initiatives possible" (Garrett and Covin, 2007: 11). Hence, strategic adaptation can be seen as the outcome of entrepreneurial initiatives that continuously pursue new opportunities where new resources are deployed "according to imperatives created by changing environmental conditions" (Garrett and Covin, 2007: 20–21; Figure 1).

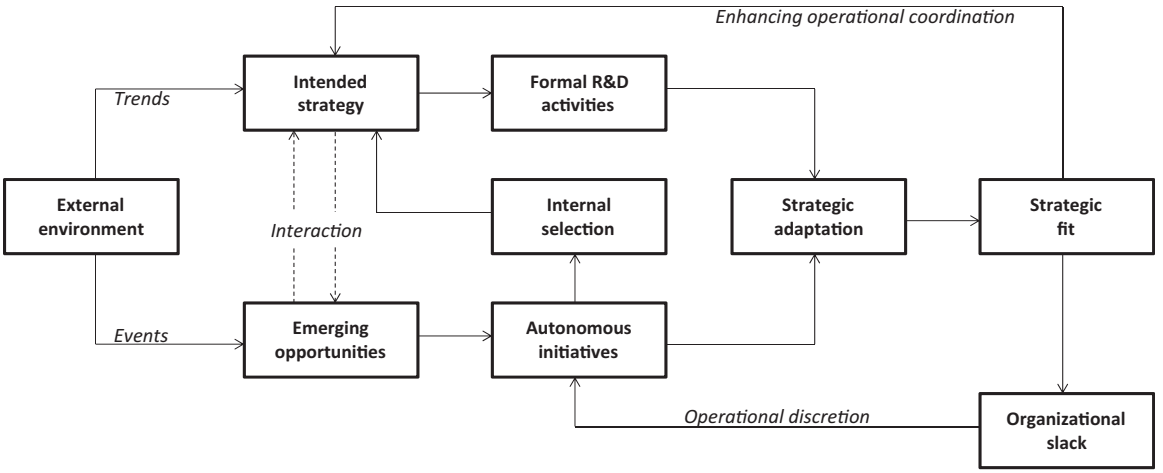

Figure 1 An entrepreneurial emergent model for strategic adaptation. Note: Inspired by Garrett, R.P., Covin, J.G., 2007. A model of corporate entrepreneurship as a strategic adaptation mechanism. Advances in Entrepreneurship, Firm Emergence and Growth 10, 9–31.

The gradual molding of entrepreneurial initiatives resonates with the concept of effectuation as entrepreneurial processes that "take a set of means as given and focus on selecting between possible effects that can be created with that set of means" (Sarasvathy, 2001: 245). According to this logic entrepreneurial initiatives use available resources and competencies to discover emergent strategic opportunities at acceptable risks and thereby generate low-cost exploratory search processes for adaptive strategic ventures. In line with the prior organization theoretical arguments, the effectuation approach is also hypothesized to "build strong participatory cultures, rather than hierarchical, procedures-based ones" (Sarasvathy, 2001: 261).

Strategic Renewal and Dynamic Capabilities

Managers differ in their ability to deal with change and, therefore, approach adaptive decisions differently, which affects the strategic outcomes. These dynamic managerial capabilities are defined as "the capabilities with which managers build, integrate, and reconfigure organizational resources and competences" (Adner and Helfat, 2003: 1020). This is not far from the definition of dynamic capabilities as "the capacity of an organization to purposefully create, extend, or modify its resource base" (Helfat et al., 2007: 4). However, the former is focused on the top management echelon whereas the latter looks more into the firm-level mechanisms that may involve a broader set of organizational constituents. Asset orchestration and the ability to coordinate resource decisions are considered central as the "fit between resource investments and their means of deployment is important for firm performance" (Sirmon and Hitt, 2009: 1376). Hence, the reconfiguration of resources is an essential element of strategic renewal and dynamic (managerial) capabilities.

Strategic renewal includes "the process, content, and outcome of refreshment or replacement of attributes of an organization that has the potential to substantially affect its long-term prospects" (Agerwal and Helfat, 2009: 282). There are fundamentally two types of transformation: discontinuous and incremental renewal. Incremental strategic renewal is conceived as a more proactive approach that "enable firms to cope with changes in the external environment as they take shape, and thereby reduce the need for a much larger and more difficult transformation later on" (Agerwal and Helfat, 2009: 283). Both incremental continuous renewal and discontinuous transformations can eventually contribute to significant organizational changes as "a series of small incremental changes can accumulate into a much larger change when viewed over a longer time span" (Agerwal and Helfat, 2009: 284).

The new competitive landscape characterized by increasing uncertainty and a decreasing ability to forecast developments within ambiguous industry borders calls for strategic response capabilities (e.g., Bettis and Hitt, 1995). Effective dynamic responses must incorporate the ability to sense the environmental changes, find proper responses to the changes, and reorganize available resources to execute the responses. This calls for dispersion of decision power to local decision nodes as "a flatter hierarchy reduces both transmission time and the distortion as information moves up and down the hierarchy, and, thus, improves performance" (Bettis and Hitt, 1995: 16).

Along the same lines, Teece et al. (1997) argue that successful firms demonstrate timely innovation where internal and external competences are coordinated and redeployed in effective responses. The underlying theme is that firm competencies must be renewed to remain congruent with a changing environment and both internal and external resources and competencies must be reconfigured to match the changing conditions. On this basis they define dynamic capabilities "as the firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments" (Teece et al., 1997: 516).

A more formalized conceptualization of the dynamic capabilities approach comprises three elemental functions. First, the ability to *sense* ongoing changes and identify new opportunities and threats in the evolving business environment. Second, the ability to *seize* opportunities in new products, processes, and services and develop them into viable business propositions. Third, the ability to *reconfigure* and recombine resources and competences into modified organizational structures that respond to market and technology changes (Teece, 2007).

Responsive Strategy Conduct

The strategy conduct approach relates to managerial decisions but differs from the contingency perspective. It does not try to explain management behavior but attempts to show how good responsive practices can lead to better strategic adaptation and favorable outcomes. The underlying idea is to develop response capabilities that allow the firm to retain strategic fit by assessing external changes accurately, formulating appropriate responses, and reconfiguring internal resources in effective responses. "We will refer to this bundle of capabilities to assess the environment, identify firm resources, and mobilize them in effective responsive actions (achieving strategic fit over time) as strategic responsiveness" (Andersen et al., 2007: 411).

The ability to adapt to environmental change may affect organizational performance as good responsive management practices can lead to higher returns and lower variation of returns at the same time. A simple model of strategic responsiveness related to the organization's ability to assess environmental conditions, mobilize available resources, and implement responsive actions that adapt the organization toward a better strategic fit shows how this can come about.

The model assumes that a firm's resource endowment can achieve an optimal performance K, if it applies these resources effectively to obtain perfect strategic fit. If a firm is unable to assess the environment and respond effectively to environmental changes there will be a strategic mismatch and performance will be below K. If poor responsiveness leads to wide discrepancies between the firm's position compared to the environment, performance will be particularly low. The performance function of a given firm in period t is then expressed as:

$$R_t = K - b|c_t - d_t|^a$$

Here c_t is interpreted as a key environmental parameter whereas d_t is a central structural dimension in the firm corresponding to its chosen position with respect to parameter c at time t. Managers in the firm observe environmental changes from period to period, identify solutions based on available resources, and deploy them in responses that try to move the firm closer to the environment. The coefficients (a, b) determine how severe the effect of a misfit will be. If the firm is unable to match the environment, performance suffers in proportion to the size of the misfit.

To achieve optimal performance a firm must be able to match d_t to c_t at all times, but in reality firms cannot always achieve a perfect fit, due to uncertainty, imprecise foresight, judgmental errors, biased information, etc. Hence, one must expect that firms have different response capabilities with some firms more likely to reach fit than others. The performance function together with the heterogeneity assumption leads responsive firms to have higher profitability and lower variance in profitability, and vice versa for nonresponsive firms, which is expressed in an inverse risk-return relationship (Andersen et al., 2007). That is, responsive firms that can adapt effectively to obtain better strategic fit will gain higher performance outcomes as well as they will achieve them with lower variation in outcomes over time.

Responsive Strategy Making

How do we synthesize the preceding insights derived from different theoretical angles toward describing a responsive strategy-making process? There seems to be general agreement about a need to decentralize decisions to induce entrepreneurial behavior and responsive initiatives at lower hierarchical levels. There seems to be good arguments for incremental responses to achieve continuous adaptation of strategy and structural parameters as the environmental context evolves. It further projects more complex hybrid structural forms that combine elements of intended planning and strategic emergence. This suggests that a combination of analytics-based strategizing and autonomy to pursue entrepreneurial initiatives together with interactive exchanges of information in open communication systems can improve environmental sensing and the ability to identify new opportunities for strategic responses.

Strategic emergence relates to decentralized, or autonomous, initiatives where strategy is shaped by experimentation with new opportunities pursued in actions taken at lower hierarchical levels (Andersen, 2004; Burgelman, 1988). The dispersion of decision power allows low-level managers to take actions in response to environmental changes they identify and thereby experiment with alternative business solutions. To avoid too divergent experimentation a more complete model incorporates intended planning modes with interactive management information systems levered by intraorganizational communication (**Figure 2**).

While it is often assumed that top management has more updated information about organizational performance, it can also be argued that frontline employees that learn from daily transactions have a good immediate sense of the effectiveness of firm capabilities. Their hands-on experience from the ongoing dealings with major stakeholders gives them updated information to assess current competencies and their effect on performance (Hallin et al., 2013). This unique information can be collected and channeled in concise form as updated inputs to the strategic decision processes, which can accommodate faster and more accurate responses to incremental environmental changes.

The combination of decentralized experimentation and central strategic reasoning constitutes the central building blocks of a responsive organization. In this organization current observations and experiential insights gained by operating managers and frontline employees are linked to the forward-looking planning that is considered and articulated around the top management team. This identifies a responsive organization that is able to observe and react to frequent and often abrupt environmental changes and adapt organizational activities to a gradually changing reality (Andersen and Fredens, 2011). This perspective highlights the significance of dynamic interaction between the operational managers involved in the execution of business activities and the top managers involved in forward-looking long-term strategy discussions. Hence, the firm's strategic considerations are updated and revised partially based on experiential insights from autonomous initiatives and, therefore, provide more

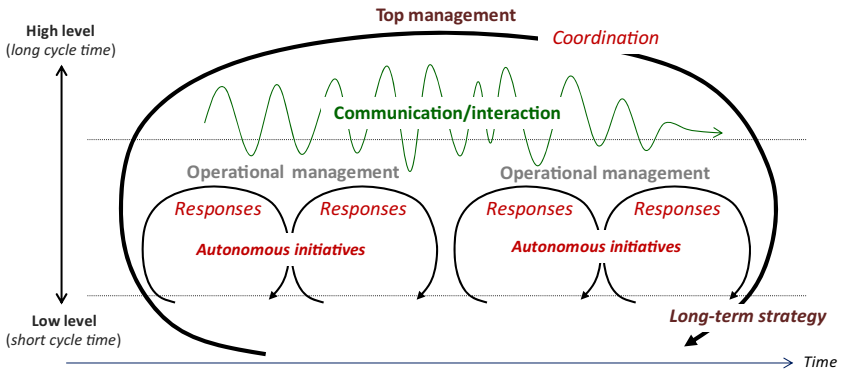

Figure 2 An interactive long-term responsive strategic adaptation model. Note: Inspired by Andersen, T.J., 2013. Short Introduction to Strategic Management. Cambridge University Press, Cambridge, UK.

updated and accurate environmental sensing for better and more responsive strategic decisions.

The decentralized structure that allows local experimentation may uncover new business opportunities (Burgelman, 1996; Burgelman and Grove, 2007) that can be fed into the periodic strategy deliberations in the long-term strategic planning process. Hence, strategic management and information processing can support updated forward-looking evaluations of opportunities uncovered by insights gained from autonomous initiatives. "That is, an integrative structure circumscribed by dispersed actions can be construed as a cohesive dynamic system where the elemental parts constitute the organization's response capabilities" (Andersen and Fredens, 2011).

Discussion and Conclusion

The preceding overview of article linked to the strategic adaptation topic has uncovered a considerable legacy of contributions forming different aspects of the concept although not necessarily providing a clear depiction of the adaptive strategy processes. The ability to create strategic fit and retain it as the environment changes is a central theoretical argument for strategic adaptation and the importance ascribed to the concept is its necessity for survival and the fact that it simply pays off to adapt. This is because the highest returns are awarded to those firms that can fulfill the changing customer needs, technological capabilities, operational competencies, and so forth at all times. The underlying processes of adapting the internal strategy-structure system to fit the changing external conditions are carried out through relatively complex adjustment mechanisms. Strategic fit is not achieved as an automatic reaction.

The strategic adaptation process is typically a piecemeal encounter where incremental adjustments in local entities lead to coevolutionary adjustments in other related parts of the organization. Continuous changes are more common than revolutionary changes but may nonetheless lead to substantial movements in strategy over time as many small steps add up. Adapting things incrementally is a way to avoid dramatic need-based and forced change scenarios. The continuous incremental changes can come about in the form of autonomous low-level initiatives pursuing entrepreneurial opportunities that subsequently may lead the way toward new strategic positions. Adhering to a long-term strategy process at the same time gives direction toward intended goals and coordinates strategic actions.

A formal model of strategic responsiveness and adaptation can show how firms will gain higher returns with lower variability in returns when they adapt particularly well to environmental changes and consistently obtain superior strategic fit. This inverse risk-return effect is particularly pronounced in dynamic industries where change is frequent and substantial so it is more difficult to assess and adapt to the changing conditions. This shows a need to further elaborate on the core strategic response capabilities as important elements of effective strategic management processes. Here the ability to sense environmental changes and reconfigure competencies around effective responses to the changing environment is essential.

A suggested model for strategic adaptation combines intended long-term strategy with emergent strategy making in the form of autonomous initiatives experimenting with strategic opportunities for changing conditions. The structure incorporates updated information processing and open communication to facilitate interaction and mutual adjustments between the intended and emergent processes. This integrative structure comprised by forward-looking strategic reasoning and entrepreneurial experimentation constitutes a dynamic responsive system.

See also: Business Models; Business Platforms; Cyclical Industrial Dynamics; Entrepreneurship and Organizations; Organization: Overview; Organizational Path Dependence; Strategic Management; Strategizing.

Bibliography

Adner, R., Helfat, C., 2003. Corporate effects and dynamic managerial capabilities. Strategic Management Journal 24, 1011–1025.

Agerwal, R., Helfat, C.E., 2009. Strategic renewal of organizations. Organization Science 20 (2), 281–293.

Andersen, T.J., 2013. Short Introduction to Strategic Management. Cambridge University Press, Cambridge, UK.

Andersen, T.J., 2004. Integrating decentralized strategy making and strategic planning processes in dynamic environments. Journal of Management Studies 41, 1271–1299.

Andersen, T.J., Fredens, K., 2011. Strategy as Complementary Interaction between Central and Peripheral Processes. Strategic Management Society Annual International Conference. Miami, FL.

Andersen, T.J., Denrell, J., Bettis, R.A., 2007. Strategic responsiveness and Bowman's risk-return paradox. Strategic Management Journal 28, 407–429.

Ansoff, H.I., Brandenburg, R.G., 1971. A language for organization design. Management Science 17 (12), 705–716.

Bettis, R.A., Hitt, M.A., 1995. The new competitive landscape. Strategic Management Journal 16 (Special Issue), 7–19.

Burgelman, R.A., 1988. Strategy making as a social learning process: the case of internal corporate venturing. Interfaces 18, 74–85.

Burgelman, R.A., 1991. Intraorganizational ecology of strategy making and organizational adaptation: theory and field research. Organization Science 2 (3), 239–262.

Burgelman, R.A., 1996. A process model of strategic business exit: implications for an evolutionary perspective on strategy. Strategic Management Journal 17, 193–214.

Burgelman, R.A., Grove, A.S., 2007. Let chaos reign, then rein in chaos – repeatedly: managing strategic dynamics for corporate longevity. Strategic Management Journal 28 (10), 965–979.

Chakravarthy, B.S., 1982. Adaptation: a promising metaphor for strategic management. Academy of Management Review 7 (1), 33–44.

Child, J., McGrath, R.G., 2001. Organizations unfettered: organizational form in an information-intensive economy. Academy of Management Journal 44, 1135–1148.

Donaldson, L., 1987. Strategy and structural adjustment to regain fit and performance: in defence of contingency theory. Journal of Management Studies 24 (1), 1–24.

Drazin, R., Van de Ven, A.H., 1985. Alternative forms of fit in contingency theory. Administrative Science Quarterly 30 (4), 514–539.

Fiegenbaum, A., Hart, S., Schendel, D., 1996. Strategic reference point theory. Strategic Management Journal 17, 216–236.

Garrett, R.P., Covin, J.G., 2007. A model of corporate entrepreneurship as a strategic adaptation mechanism. Advances in Entrepreneurship. Firm Emergence and Growth 10, 9–31.

Hallin, C.A., Andersen, T.J., Tvetterås,, S., 2013. Who Are the Better Predictors: Frontline Employees or Executive Managers? Strategic Management Society Annual International Conference. Atlanta, GA.

Helfat, C.E., Finkelstein, S., Mitchell, W., Peteraf, M., Singh, H., Teece, D., Winter, S.G., 2007. Dynamic Capabilities: Understanding Strategic Change in Organizations. John Wiley & Sons.

Hill, S., Martin, R., Harris, M., 2000. Decentralization, integration and the post-bureaucratic organization: the case of R&D. Journal of Management Studies 37, 563–585.

Hofer, C.W., 1975. Towards a contingency theory of business strategy. Academy of Management Journal 18 (4), 784–810.

Hrebiniak, L.G., Joyce, W.F., 1985. Organizational adaptation: strategic choice and environmental determinism. Administrative Science Quarterly 30 (3), 336–349.

Jennings, D.F., Seaman, S.L., 1994. High and low levels of organizational adaptation: an empirical analysis of strategy, structure, and performance. Strategic Management Journal 15, 459–475.

Miller, D., Friesen, P.H., 1980. Momentum and revolution in organizational adaptation. Academy of Management Journal 23 (4), 591–614.

Mintzberg, H., Waters, J., 1982. Tracking strategy in an entrepreneurial firm. Academy of Management Journal 25 (3), 465–499.

Mintzberg, H., Waters, J., 1985. Of strategies, deliberate and emergent. Strategic Management Journal 6, 257–272.

Romanelli, E., Tushman, M., 1994. Organizational transformation as punctuated equilibrium: an empirical test. Academy of Management Journal 37 (5), 1141–1166.

Sarasvathy, S.D., 2001. Causation and effectuation: towards a theoretical shift from economic inevitability to entrepreneurial contingency. Academy of Management Review 26 (2), 243–263.

Scholz, C., 1987. Corporate culture and strategy: the problem of strategic fit. Long Range Planning 20 (4), 78–87.

Sirmon, D.G., Hitt, M.A., 2009. Contingencies within dynamic managerial capabilities: interdependent effects of resource investment and deployment of firm performance. Strategic Management Journal 30, 1375–1394.

Teece, D.J., Pisano, G., Shuen, A., 1997. Dynamic capabilities and strategic management. Strategic Management Journal 18, 509–533.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28, 1319–1350.

Thomas, L.G., D'Aveni, R., 2009. The changing nature of competition in the US manufacturing sector, 1950–2002. Strategic Organization 7 (4), 387–431.

Volberda, H.W., 1996. Toward the flexible form: how to remain vital in hypercompetitive environments. Organization Science 7, 359–374.

Strategic Management

Robert A Burgelman, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Strategic management has become a well-established, though still somewhat fragmented, academic discipline. Two of its theoretical foundations – the strategic positioning view and the resource-based view – together with an integrative view on strategy formulation and implementation are combined into one conceptual framework. Similarly, strategizing and economizing are combined into a conceptual framework that integrates external-oriented perceived value (focused on strategizing) and internal-oriented delivered cost (focused on economizing) to provide further insight into the performance challenges facing strategic management. Finally, a framework is constructed that integrates four models of the strategy-making process proposed in different academic literature relevant for strategic management.

Genesis and Development

The academic field of strategic management grew out of the earlier pedagogical field of business policy, which was conceptualized as an integrative capstone course to be taught at the end of the business school curriculum to teach students how to integrate the various functional areas, but was not considered a field of research in its own right or a broader scientific discipline (Schendel and Hofer, 1979). In the context of becoming viewed as a new academic field, strategic management was initially defined as "…a process that deals with the entrepreneurial work of the organization, with organizational renewal and growth, and more particularly, with developing and utilizing strategy which is to guide the organization's operations" (Schendel and Hofer, 1979: 11). The creation, in 1982, of the Strategic Management Society provided scholars of the new academic field with an institutional home, and the yearly Strategic Management Society Conference and the establishment of the *Strategic Management Journal* (*SMJ*) offered them an outlet for their research.

By the late 1980s, the strategic management field had gained enough momentum to motivate organizing a second conference in Napa, California, in 1991 to examine progress made since the 1977 conference in Pittsburgh, Pennsylvania, and to provide further directions for the development of the field. Synthesizing the insights from the conference papers and discussions into book form, the organizers (R. Rumelt, D. Schendel, and D. Teece) identified four key issues, translated into research questions, that could be helpful in defining the field of strategic management, its boundaries as they existed at the time, its concerns, and how the field could make a contribution to the practice of management: (1) How do firms behave? (2) Why are firms different? (3) What is the function of, or value added by, the headquarters unit in a multibusiness firm? and (4) What determines the success or failure of the firm in international competition? (Rumelt et al., 1994: 2). By the time they were able to publish their book 3 years after the conference, the organizers realized they had failed to pay enough attention to the nature of what they call the 'policy process' – the strategy-making process – and to several other important questions (1994, Afterword). In other words, it had become clear that there simply was no way to guide scholars in advance to what would be the important contributions to the strategic management field.

By the mid-2000s, *SMJ* had become a well-respected A-level academic journal and the Strategic Management Society Conference every year attracted some 1000 academics, consultants, and business leaders (the ABCs) from all over the world. Yet the burgeoning new academic field was still viewed by many as fragmented and lacking of a coherent identity. As the result of empirically identifying the common substantive core of the field and the implicit consensus among its diverse scholars about what constitutes the field's coherence and identity, Nag, Hambrick, and Chen were able to arrive at a new definition of the field: "The field of strategic management deals with the major intended and emergent initiatives taken by general managers on behalf of owners, involving utilization of resources, to enhance the performance of firms in their external environments" (2007: 944). The researchers realized that this definition of the field was in many ways too restrictive, but suggested that "…the greatest benefit of having this definition is that it allows strategic management scholars to frame the debate about what they want the field to become, or how they want to change it" (2007: 951).

Some Theoretical Foundations

Soon after the initial formation of the strategic management field, Porter (1980) appeared on the scene to provide the inchoate new field with a solid foundation rooted in the economics of industrial organization. Porter's seminal discussion of the importance of 'strategic position' and 'generic strategy' for competitive advantage stimulated the development of the 'positional school' in the strategic management literature (for an updated view, see Stonehouse and Snowdon, 2007).

Shortly after Porter's breakthrough contribution, Wernerfelt's (1984) articulation of the so-called resource-based view offered an alternative to the 'position-based view' of competitive advantage. In this view, resources are broadly defined to encompass technical, commercial, as well as knowledge, administrative and cultural factors. Some of the

foci of the resource-based view are on the competitive advantages that may derive from resource heterogeneity among firms and on the sustainability of resource-based advantages (for a critical review, see Kraaijenbrink et al., 2010). In an important sense, the resource-based view also relates to Bower's finding that strategy making in reality is constituted, to a large extent, by the resource allocation process (Bower and Gilbert, 2005).

Since 1990, 'core competence' (Prahalad and Hamel, 1990) has gained great currency among practitioners. So much so that Porter (1996) felt the need to respond with a strong reaffirmation of the importance of strategic positioning for competitive advantage. More recently, the concept of 'dynamic capabilities' (e.g., Teece, 2007) has focused strategic management researchers' attention on how an organization's resources and competencies may be combined into 'capabilities' and how these may change over time and drive the organization into new strategic directions.

It is worth noting that the positional school and the resource-based view are part of a larger set of 10 identifiable 'schools' in strategic management (Mintzberg et al., 1998), which testifies to the field's scholarly fertility but also its increasing complexity that Nag et al. (2007) tried to reduce.

An Integrative Conceptual Framework

Longitudinal field research, which examined the role of the internal ecology of strategy making in helping shape Intel Corporation's evolution (Burgelman, 2002a), provided a response to important research in organizational ecology that fundamentally questioned the role of strategic management in organizational survival and adaptation. The response was based on the view that organizations, such as Intel, are ecologies in their own right within which strategic initiatives compete for its resources, and that the role of strategic management is to orchestrate and guide this internal ecological process in light of the dynamics of the higher level, external ecological processes.

This longitudinal field research produced an integrative conceptual framework for strategic management – the 'strategy diamond' – of the dynamic forces that drive a company's evolution. This framework brings together ideas from the strategic management literature that emphasize the importance of product-market position (the industrial organization perspective) with ideas that emphasize the importance of distinctive (or core) competencies (the resource-based view), and proposes that both product-market positioning and distinctive competencies are important for the attainment of competitive advantage. In addition, the framework proposes that the reality of strategy (as opposed to the rhetoric of strategy) lies in its execution through strategic actions.

In doing so, the strategy diamond framework integrates five dynamic forces: (1) the organization's 'official (or stated) corporate strategy,' which defines the scope of the business(es) it wants to be a winner in and its intended competitive advantage relative to the other players in the industry; (2) the 'basis of competitive advantage in the industry' associated with the organization's chosen product-market position in the industry; (3) the organization's 'distinctive competencies,' which encompass the technical, commercial, and administrative and managerial competencies in which it excels to create customer value relative to competitors; (4) the organization's 'strategic actions,' which refer to the 'consequential' actions; that is, actions that commit the organization in a strategic direction and are not easily reversed (see also Ghemawat, 1991), and through which it actually uses its product-market position and distinctive competence to achieve competitive advantage; and (5) its 'internal selection environment,' which comprises the internal contextual elements that help maintain alignment of the other four forces (organization structure, planning and control systems, measurement and reward systems, recruitment and executive development, and leadership style, values, and norms). This framework serves to analyze strategic dynamics at the organization–environment interface level of analysis. Figure 1 shows the strategy diamond framework.

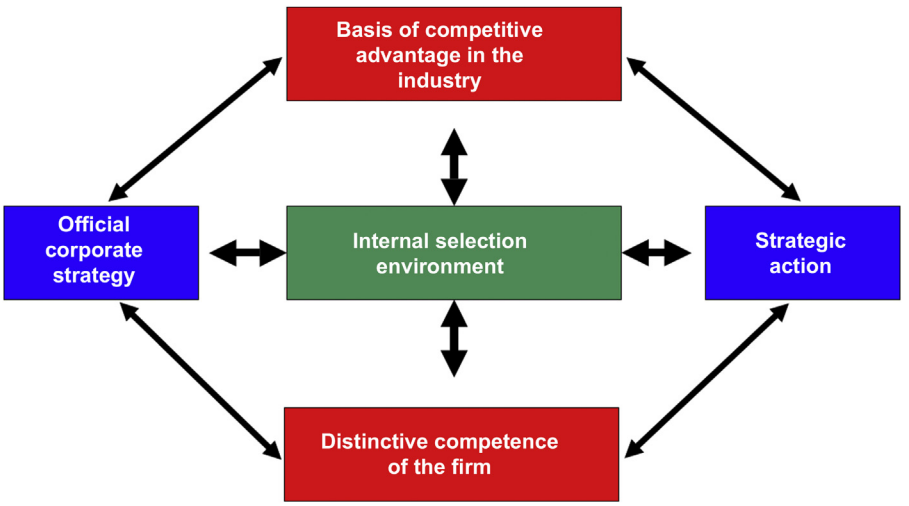

Figure 1 The strategy diamond: dynamic forces driving company evolution. Source: Adapted from Burgelman, R.A., March 1994. Fading memories: A process theory of strategic business exit in dynamic environments, Administrative Science Quarterly and Burgelman, R.A., 2002. Strategy is Destiny: How Strategy-Making Shapes a Company's Future, Free Press, New York.

The scalability of the strategy diamond framework, which can be scaled up and down levels of analysis (individual, functional, single business, multibusiness, or corporate), facilitates strategic integration across levels in the organization. For instance, it can be used to study the strategic situation facing the multibusiness firm as well as the strategic situation of any functional area within the firm.

The strategy diamond draws attention of strategic management scholars and practitioners to the importance of considering strategic position together with distinctive competence in order to understand and secure an organization's capacity to survive and thrive. This turns out to be surprisingly difficult. Two examples of the computer industry serve to illustrate this point. First, Novell's preeminent strategic position in network operating system market segment in the early 1980s (more than 70% market share) may have led the company to allow its distinctive competence in network operating systems to wither while pursuing nonrelated opportunities in the PC application software market segment, leading to delays in the launch of its next-generation core product and giving Microsoft an opportunity to attack Novell's core market with Windows NT. Second, Apple Computer has always had extraordinary strong distinctive competencies in software development, with many experts agreeing that the Mac operating system was superior to Microsoft's Windows operating system. Yet, because Apple's top management did not appreciate the importance of achieving a strategic position with a large installed base to gain 'increasing returns to adoption' (attracting independent software developers increases the value of the operating system, which increases the installed base and attracts even more software developers, and so on) Apple Computer was relegated to a small niche player in the PC industry.

Strategic Dynamics

Think of the connections between the basis of competitive advantage in the industry, distinctive competence, official strategy, and strategic action as 'rubber bands.' During periods of strategic alignment, the rubber bands are evenly stretched because the dynamic forces are in harmony, and the strategy diamond looks like Figure 1. Over time, however, the alignment will come under pressure as the dynamic forces become disharmonious, causing the rubber bands to stretch unevenly. The stretching of the rubber bands causes 'strategic dissonance' between the company's top and senior executives, which signals that the company likely faces an impending 'strategic inflection point' (Burgelman and Grove, 1996). Having established a baseline of alignment with the help of the strategy diamond may make it easier to understand the sources of the strategic dissonance. This, in turn, may make it easier to face up to the changes that top management must make to re-establish alignment over time in the face of changing environmental dynamics.

Role of the Internal Selection Environment

Most organizations face a tough external selection environment constituted by evolving market forces, nonmarket forces (technological change, government regulation), and general economic conditions (booms and busts). Most startup companies succumb to the pressures of the external selection environment and disappear, and even large companies or business units often face nonlinear strategic dynamics that can quickly change their relative success (Burgelman and Grove, 2007). Organizations that survive external selection gain the opportunity to substitute to some extent internal selection for external selection, which allows them to continue to outrun and outsmart the external selection environment. At the same time, however, as organizations grow and become more complex with multiple functional, product and geographic groups, and multiple levels of management, it becomes more difficult to accurately and in a timely fashion assess the pressures of a highly dynamic external selection environment.

Hence, a key set of analytical questions concerns how well the company's internal selection environment works to dynamically align strategic action with stated strategy and distinctive competence with the basis of competition in the face of changes in the external selection environment. First, to what extent do the criteria guiding the company's resource allocation – the types of projects funded and the funding levels – reflect external competitive reality? If they do not, they will probably reflect purely internal standards of performance, internal politics, or simple inertia. (General electric's determination that a business has to be #1 or #2 in its industry to remain in the corporate portfolio and Intel's rule of maximizing margin-per-wafer-start to allocate scarce manufacturing capacity to different semiconductor products are examples of criteria that reflect competitive reality.) Second, to what extent does the strategic planning process offer the opportunity to debate new opportunities? These debates typically involve those who make money today (existing opportunities) and those who promise to make money tomorrow (new opportunities), and it is rational to expect that today will usually win against tomorrow. Third, a central concern is therefore the capacity of the senior team for 'strategic recognition'; that is, how strong is the senior team's ability to see the implications of internal and external changes often *after* they are already happening but *before* others see their strategic implications? And fourth, how strong is the senior team's capacity to go from strategic recognition to strategic action; that is, their capacity for the sort of strategic leadership that forcefully drives change?

Finding valid and reliable answers to these questions may help top management diagnose the capacity of the organization's internal selection environment to maintain the key alignments over time in the face of environmental dynamics. In the case of divergences between strategy and action, for instance, this may help top management get clear about the extent to which these divergences are due to a poorly designed internal selection environment in the face of a relatively stable external selection environment, or to the fact that the internal selection environment is actually correctly picking up signals of external changes – caused either by the organization's own strategic actions to shape the external environment to its advantage or by the strategic actions of others in the external environment to which it must respond – that are putting pressure on the alignment between strategy and action.

Implications

Linking firm-level strategy and strategic action in one framework draws attention to their dynamic reciprocal relationship. For instance, as further discussed in section Strategy-Making Process, the received strategy at any given time will 'induce' (intended) action; but entrepreneurial, 'autonomous' (emergent) action may, in turn, influence the evolving strategy (Burgelman, 1983; Mintzberg, 1978). Explicitly drawing attention to both corporate strategy and strategic action also highlights the reality that the effectiveness of capabilities ultimately depends on human actors: Capabilities in and of themselves only constitute a potential. Also, the strategy diamond may help top management appreciate better that in the end strategy is only as good as strategic action, and that while strategy without capabilities is powerless, capabilities without strategy are aimless.

The strategy diamond framework also draws attention to two inertial hazards associated with potential tradeoffs between positional advantages and advantages based on distinctive competence (e.g., Barnett, 1997; Burgelman, 2002b). Such tradeoffs may create two types of strategic traps. Companies that rely heavily on positional advantages shield themselves from competitive pressures but face a potential 'position trap': The security of their positional advantage may relax their diligence in continuing to hone and develop their distinctive competencies. As a result, their existing competencies may lose some of their efficiency or strength, which may make them potentially vulnerable to new, fitter competitors attacking their strategic position. On the other hand, companies that rely heavily on distinctive competence to compete vigorously with similar others may be able to sharply hone these competencies and become best in class. However, such efforts potentially create a 'competence trap': The relentless efforts to hone existing distinctive competence may make the company vulnerable to new competitors with different distinctive competencies as competitive dynamics in the industry change (Levitt and March, 1988), or they may simply fail to appreciate the competitive importance of achieving a dominant strategic position.

Integrating the dynamic links between strategic positioning and the deployment of distinctive competence, the strategy diamond framework may potentially serve as a research tool for examining the microfoundations of dynamic capabilities, which as Teece points out can for analytical purposes "…be disaggregated into the capacity (1) to sense and shape opportunities and threats, (2) to seize opportunities, and (3) to maintain competitiveness through enhancing, combining, protecting, and, when necessary reconfiguring the business enterprise's intangible and tangible assets" (2007: 1319).

Examining dynamic capabilities in the framework of the strategy diamond may help prepare top management to better face the transient nature of all sources of competitive advantage in highly dynamic, 'hypercompetitive' environments (D'Aveni, 1994).

The strategy diamond framework also highlights the critical role of the internal selection environment in helping top management cope with such hypercompetitive external selection environments. Introducing the concept of an internal selection environment into the strategic management literature made it possible to link the field to evolutionary and ecological organization theories. It directs scholarly attention to how the internal selection environment may help a company to dynamically align strategic action with strategy and distinctive competencies with the basis of competition (given its product-market position) in the face of changes in the external selection environment and/or in response to its own strategic moves. It also provides a potential lens for studying the interplays between action and cognition in strategy-making, and may thereby become a potentially useful concept in the further development of what is now the emerging subfield of 'behavioral strategy' (e.g., Levinthal, 2011).

Finally, while it takes hard work to clearly define, and get the organization to agree upon, each of the framework's concepts in light of a specific situation, it may also help significantly speed-up the strategic recognition process because it may help top management achieve faster 'knowledge churns'; that is, the framework could serve as a theory about how to make an organization successful, a means to initiate experimental strategic actions that test that theory, and a means of interpreting the results and quickly incorporating the learning of these experimental strategic actions into the strategy going forward (through the functioning of the internal selection environment). Interestingly, the process involved in generating fast knowledge churns through the use of the strategy diamond seems to support the three capacities associated with dynamic capabilities as discussed in Teece (2007). Hence, the most important advantage of using the strategy diamond may be to help avoid the loss of precious time that moving in the wrong strategic direction inevitably entails, and which puts at least some dent in the potential competitive advantage that moving in the right strategic direction earlier offers. This also suggests that the strategy diamond framework may have the potential to serve as a research tool for helping guide the field toward a dynamic theory of strategy.

Strategizing, Economizing, and Performance

Enhancing the performance of firms in their external environment is a key factor in the definition of the strategic management field, but the link between strategic management and organizational performance, however, remains somewhat murky (Nag et al., 2007). Williamson (1994) observes that business strategy encompasses both strategizing (related to market power) and economizing (related to efficiency) and proposes that while performance can be a function of both strategizing and economizing, economizing – efficiency analysis, including governance costs as well as production costs, and comparative economic organization – is the most important factor in determining performance because all firms need economizing to survive and only relatively few firms have enough market power to do strategizing.

While Williamson emphasizes the difference between strategizing and economizing, it may be more fruitful to view these two aspects of business strategy as inherently complementary. This can be readily seen if strategizing is cast in terms of creating 'perceived value' (PV) on the part of target customers and economizing is cast in terms of 'delivered costs'

(DCs) associated with creating and delivering PV. This is shown in Figure 2 (different versions of this framework have been proposed in the literature; e.g., Porter, 1996; Saloner et al., 2001).

Figure 2 shows that at any moment in time, there exists a possibilities frontier linking DC to PV as a function of the best use of available technology (T_1) and modes of organizing (O_1). This possibilities frontier, which rules out the attainment of 'bliss' (DC-unconstrained achievement of PV), can be discovered through benchmarking to find best in class competitors at the high-end, mid-end, and low-end of the market. Most organizations will find themselves inside the frontier, and it is an important task of strategic management to try to move their organization closer to the frontier through programs such as total quality control and/or the re-engineering of business processes (see Figure 2). This requires careful analysis of the link between cost drivers and PV, and the readiness of top management to move costs from lower value-creation activities to higher value-creation activities on an ongoing basis.

Combining strategizing and economizing in a PV/DC framework also offers the opportunity to represent different types of innovation. Incremental innovation moves the possibilities frontier commensurately and makes it possible for most industry incumbent companies to adapt, sometimes taking the lead and sometimes having to catch up. Radical innovation, however, shifts the possibilities frontier in a discontinuous fashion. This involves new technological solutions (T_2 in Figure 2), for instance, the use of the Internet, new ways of organizing (O_2 in Figure 2), for instance, various forms of outsourcing, and the introduction of new business models, for instance, shifting revenue generation from selling products to providing subscription services. Such changes are often introduced by new entrants to an industry and are usually very difficult for incumbent companies to adapt to because of various forms of strategic inertia.

Strategy-Making Process

After the 1991 Napa conference, it had become clear that the strategy-making process (previously known as the policy process), also deserved systematic attention of strategic management researchers (Rumelt et al., 1994). As noted earlier, the interplay of strategy and action in the diamond model sheds light on the existence of intended/induced and emergent/autonomous strategy processes, which can be related to current work in strategic management that build on March's (1991, 2006) influential discussion of exploitation and exploration in organizational learning. Top management sets the corporate strategy and induces strategic actions by lower level leaders that are aligned with it in order to exploit opportunities in the familiar environment. The autonomous strategy process, in contrast, explores new opportunities that are outside the scope of the existing corporate strategy, relate to new environmental segments, and are often based, at least in part, on new distinctive competencies. An important top management responsibility and challenge is to balance resource allocation between the intended/induced strategy process (so as to maintain 'fit') and the emergent/autonomous strategy process (so as to maintain 'evolvability') over time; in particular, the scaling up and vectoring of resources related to autonomous initiatives that demonstrate viability through the process of 'strategic context determination' (Burgelman and Grove, 2007).

Taking into account the existence of intended/induced and emergent/autonomous strategy processes, the overall strategy-making process of a complex organization can be further characterized in terms of two key dimensions: (1) the degree of concentration (versus distribution) of strategic decision-making power, and (2) the degree to which strategy execution involves all relevant parties simultaneously (or sequentially). The combination of these two dimensions

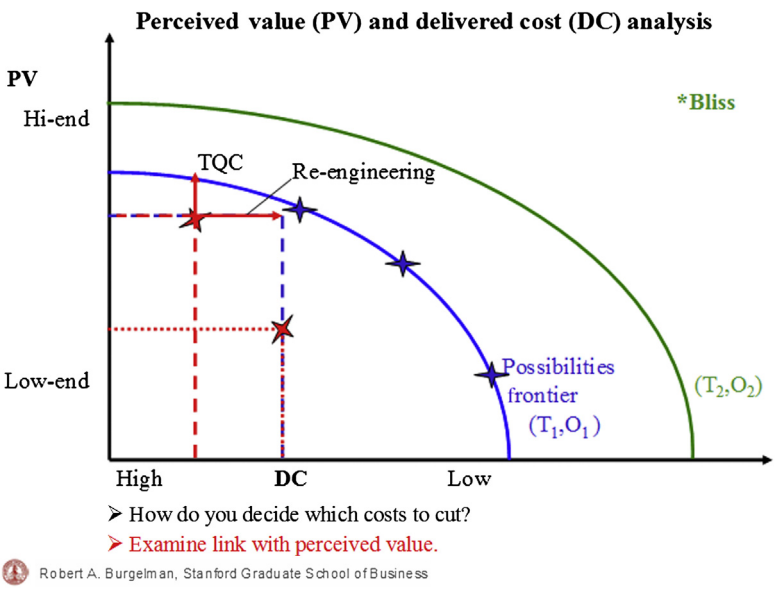

Figure 2 Strategizing and economizing: perceived value and delivered cost. *Source: Robert A. Burgelman, Lecture Material, Stanford Business School.

Table 1 Models of strategy making

		Strategic decision-making power	
		Concentrated	*Distributed*
Mode of execution	Simultaneous	Rational actor model[a]	Internal ecology model[c]
	Sequential	Bureaucratic model[b]	Garbage can model[d]

[a]Also called 'Model I' in Allison, G., Zelikow, P., 1999. Essence of Decision: The Cuban Missile Crisis, second ed. Addison Wesley Longman, New York.
[b]Also called 'Model I' in Allison, G., Zelikow, P., 1999. Essence of Decision: The Cuban Missile Crisis, second ed. Addison Wesley Longman, New York.
[c]Burgelman, R.A., 1991. Internal ecology of strategy-making and organizational adaptation: theory and field research. Organization Science. 239–262. The internal ecology model is perhaps closest to Allison and Zelikow''s Model III: 'governmental politics.'
[d]March, J.G., Olsen, J.P., 1976. Ambiguity and Choice in Organizations. Universitetsforlaget, Bergen.
Reproduced from Burgelman, R.A., 2002. Strategy is Destiny: How Strategy-Making Shapes a Company's Future. Free Press, New York, pp. 4–6.

makes it possible to integrate into one conceptual framework four organizational decision-making models previously developed in the literature (Burgelman, 2002a). Table 1 shows these four strategy-making processes.

Rational actor model. A comprehensively rational top management (individual leader or leadership team) formulates the overall strategy and is able to get all the interdependent actors in the organization to simultaneously engage in the actions necessary to implement it. In this model, there is strong alignment between strategy and action. It is often viewed as the ideal type. However, it may be most effective to respond to environmental dynamics that can be reasonably well anticipated and influenced. It may also be the best model for coping with a 'clear and present danger' or for exploiting an extraordinary opportunity. While a comprehensively rational top management is in principle able to effectively balance induced and autonomous strategy processes for some period of time, they are likely to eventually start favoring the intended/induced strategy process.

Bureaucratic model. In this model, the overall strategy is still formulated by a comprehensively rational top management, but implementation is less immediate because various parts of the organization are independent of each other and translate the strategy in terms of the logic of their own operations before taking action to implement it. This model has advantages in slow-moving environments because each part of the system has time to optimize its strategic actions in light of the overall strategy. In rapidly changing environments, however, it will lead to sluggish execution of the overall strategy. While autonomous strategic initiatives will undoubtedly spring up in different parts of the system, scaling them up will be difficult. By default, the intended/induced strategy process will become dominant.

Internal ecology model. This model views organizational-level strategy as the result of successful strategic initiatives of interdependent actors (individuals or groups), who are in a position to commit the organization and who continuously try to do so. In this model, strategy making is a highly dynamic process that capitalizes on anticipated and unanticipated variations in the internal and external environments. It views the strategy making process as constituting an opportunity structure for strategic leaders in the organization, but one in which individual opportunity seeking is constrained, to some extent, by the imperative of organizational survival. This model may be most effective in highly uncertain, opportunity-rich environments. The emergent/autonomous strategy process is likely to be dominant here. Coherence of system-level strategic action depends on the characteristics of the internal selection environment.

Garbage can model. In this model, strategy making results from various independent actors taking action as a function of the sequence in which problems, solutions, and decision opportunities arise. The effectiveness of system-level strategic action depends on the sequence in which problems, solutions, and decision opportunities arise. There is neither an explicit or implicit overall strategy, nor a clear ecological survival force, serving as reference point for determining whether an initiative is induced or autonomous. Hence, by default the autonomous strategy process dominates. Arriving at a coherent overall strategy for the organization is to a large extent governed by chance. The normative implication of this model is to 'just hang in there and keep trying.'

Conclusion

Strategic management has come a long way since 1979. As Nag et al. (2007) observe, in addition to the great diversity among strategic management scholars proper, a wide variety of scholars in adjacent fields (economics, sociology, psychology, marketing, etc.) are doing significant work that has a bearing on strategic management. The launch of two new journals – *Strategic Entrepreneurship Journal* and *Global Strategy Journal* – in recent years, as well as the continued growth of the Strategic Management Society, attest to the further expansion of the field.

While this expansion is important for the continued intellectual fertility of the field and will likely lead to unanticipated, radically novel insights over time, it also poses a potential problem in terms of an ever-expanding topical universe and concomitant fragmentation of the field into specialized communities (or even camps) that have increasingly less in common. Hence, it is perhaps useful to engage in some efforts along the way to continue to integrate various important insights generated within this diverse field. That is the main purpose of the framework of dynamic forces driving company evolution (the strategy diamond), the PV/DC framework, and the framework of models of strategy making.

See also: Business Models; Closed and Open Systems: Organizational; Ecology: Organizations; Entrepreneurship and Organizations; Strategizing.

Bibliography

Barnett, W.P., 1997. The dynamics of competitive intensity. Administrative Science Quarterly 42, 128–180.

Bower, J.L., Gilbert, C.G., 2005. From Resource Allocation to Strategy. Oxford University Press, Oxford.

Burgelman, R.A., 1983. A model of the interaction of strategic behavior, corporate context, and the concept of strategy. Academy of Management Review 8, 61–70.

Burgelman, R.A., 2002a. Strategy is Destiny: How Strategy-Making Shapes a Company's Future. Free Press, New York.

Burgelman, R.A., 2002b. Strategy as vector and the inertia of coevolutionary lock-in. Administrative Science Quarterly 47, 325–357.

Burgelman, R.A., Grove, A.S., 1996. Strategic dissonance. California Management Review 38 (2), 3–23.

Burgelman, R.A., Grove, A.S., 2007. Let chaos reign, then rein in chaos – repeatedly: managing strategic dynamics for corporate longevity. Strategic Management Journal 28, 965–979.

D'Aveni, R.A., 1994. Hypercompetition: Managing the Dynamics of Strategic Maneuvering. Free Press, New York.

Ghemawat, P., 1991. Commitment. Free Press, New York.

Kraaijenbrink, J., Spender, J.-C., Groen, A.J., 2010. The resource-based view: a review and assessment of its critiques. Journal of Management 36, 349–372.

Levinthal, D.A., 2011. A behavioral approach to strategy: what's the alternative? Strategic Management Journal 32, 1517–1523.

Levitt, B., March, J.G., 1988. Organizational learning. In: Scott, W.R. (Ed.), Annual Review of Sociology 14, 319–340.

March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science 1, 71–87.

March, J.G., 2006. Rationality, foolishness, and adaptive intelligence. Strategic Management Journal 27, 201–206.

Mintzberg, H., 1978. Patterns in strategy formation. Management Science 24, 934–948.

Mintzberg, H., Ahlmstrand, B., Lampel, J., 1998. Strategy Safari. Free Press, New York.

Nag, R., Hambrick, D.C., Chen, M.-J., 2007. What is strategic management really? Inductive derivation of a consensus definition of the field. Strategic Management Journal 28, 935–955.

Porter, M.E., 1980. Competitive Strategy. Free Press, New York.

Porter, M.E., November–December 1996. What is strategy? Harvard Business Review.

Prahalad, C.K., Hamel, G., 1990. The core competence of the corporation. Harvard Business Review.

Rumelt, R.P., Schendel, D.E., Teece, D.J., 1994. Fundamental Issues in Strategy. Harvard Business School Press, Boston.

Saloner, G., Shepard, A., Podolny, J., 2001. Strategic Management. John Wiley, New York.

Schendel, D.E., Hofer, C.W., 1979. Strategic Management: A New View of Business Policy and Planning. Little, Brown and Company, Boston.

Stonehouse, G., Snowdon, B., 2007. Competitive advantage revisited: Michael Porter on strategy and competitiveness. Journal of Management Inquiry 16, 256–273.

Teece, D.J., 2007. Explicating dynamic capabilities: the nature and micro-foundations of (sustainable) enterprise performance. Strategic Management Journal 28, 1319.

Wernerfelt, B., 1984. A resource-based view of the firm. Strategic Management Journal 5, 171–180.

Williamson, O.E., 1994. Strategizing, economizing, and economic organization. In: Rumelt, R.P., Schendel, D.E., Teece, D.J. (Eds.), Fundamental Issues in Strategy. Harvard Business School Press, Boston.

Strategic Planning for Public and Nonprofit Organizations

John M Bryson, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Strategic planning is a deliberative, disciplined approach to producing fundamental decisions and actions that shape and guide what an organization (or other entity) is, what it does, and why it does it. Strategic planning is not a substitute for strategic thinking, acting, and learning, or for leadership. Instead, strategic planning is a leadership tool meant to enhance strategic thinking, acting, and learning. Strategic planning can produce several benefits, but there is no guarantee that it will. There are a variety of approaches to strategic planning; some approaches emphasize process, while others emphasize content. Strategic management is a broader concept that links planning and implementation.

Strategic planning may be defined as 'a deliberative, disciplined approach to producing fundamental decisions and actions that shape and guide what an organization (or other entity) is, what it does, and why it does it' (Bryson, 2011: pp. 7–8). This definition implies that strategic planning is not a single thing, but is instead a set of concepts, procedures, and tools that must be carefully tailored to situations if desirable outcomes are to be achieved. Poister and Streib (1999), in a review of the field of strategic planning and management, asserted that strategic planning should:

1. Be concerned with identifying and responding to the most fundamental issues facing an organization.
2. Address the subjective question of purpose and the often competing values that influence mission and strategies.
3. Emphasize the importance of external trends and forces, as they are likely to affect the organization and its mission.
4. Attempt to be politically realistic by taking into account the concepts and preferences of internal, and especially external, stakeholders.
5. Rely heavily on the active involvement of senior-level managers, and often elected officials, assisted by staff support where needed.
6. Require the candid confrontation of critical issues by key participants in order to build commitment to plans.
7. Be action-oriented and stress the importance of developing plans for implementing strategies.
8. Focus on implementing decisions now in order to position the organization favorably for the future.

In process terms, a generic strategic planning process typically might consist of the following steps, although not necessarily in this order (Bryson, 2011):

1. Initiate and agree on the need for a strategic planning process, who should be involved and how, what key decision points will be, and a work program.
2. Clarify the various legal and other requirements that must be addressed.
3. Clarify organizational mission, vision, and values.
4. Assess the external and internal organizational environments to identify strengths, weaknesses, opportunities, and threats.
5. Identify the strategic issues, which are the fundamental policy questions or critical challenges affecting an organization's mandates, mission and values, product or service level and mix, clients, users or payers, costs, financing, organization, and/or management (Bryson, 2011: p. 55), facing the organization.
6. Formulate strategies to manage the issues.
7. Review and adopt the strategic plan or plans.
8. Perhaps establish a more refined organizational vision.
9. Develop an effective implementation process.
10. Reassess strategies and the strategic planning process.

In addition to key decision makers, there typically will be a strategic planning team, and possibly various task forces to address different strategic issues. Varied analysis, synthesis, group process, negotiation, and decision tools will be used in the different steps depending on the specific situation.

It is important to highlight what strategic planning is *not*. Strategic planning is not a substitute for strategic thinking (thinking in context about how to pursue purposes or achieve goals (Bryson, 2011: pp. 14–15)), acting (acting in context in light of future consequences to achieve purposes and/or to facilitate learning (Bryson, 2011, p. 15)), and learning (any change in a system (which could be an individual) that, by adapting it better to its environment, produces a more or less permanent change in its capacity to pursue its purposes (Bryson, 2011: p. 15). It may help people do that, but used unwisely it may hinder strategic thinking, acting, and learning. Strategic planning is not a substitute for leadership. At least some key actors must be committed to the process, or it is bound to fail. Strategic planning is also not a substitute for an organizational or community strategy. Strategies have numerous sources, both planned and unplanned. Strategic planning is likely to result in a statement of organizational or community intentions, but what is realized in practice will be some combination of what is intended with what emerges along the way. Finally, strategic planning is not synonymous with what is called 'comprehensive planning' for communities in the United States, or what has been called 'structure planning' or 'strategic spatial planning' in Europe. There may be little difference if the agency or cross-organizational group doing the comprehensive or structure planning has strong ties to government and other important decision makers. However, in practice there may be significant differences.

First, the plans are often prepared to meet legal requirements and must be formulated according to a legally prescribed process with legally prescribed content. As legal instruments, these plans have an important influence. However, the plans'

typical rigidity can conflict with the political process with which public officials must cope. Therefore, strategic plans can provide a bridge from legally required and relatively rigid policy statements to actual decisions and operations.

Second, comprehensive or structure plans are usually limited to a shorter list of topics than a government's full agenda of roles and responsibilities. For that reason, they are of less use to key decision makers than strategic planning, which can embrace all of a government's actual and potential roles before deciding why, how, where, and when to act.

Third, strategic planning on behalf of a community is typically more action oriented, more broadly participatory, more emphatic about the need to understand the community's strengths and weaknesses as well as the opportunities and threats it faces, and more attentive to intercommunity competitive behavior. Strategic planning is thus typically more comprehensive than comprehensive planning or structure planning, while at the same time producing a more limited action focus.

To summarize, strategic planning for communities comprises a set of concepts, procedures, and tools that are broader than those of more traditional urban planning. Community strategic planning is typically viewed as an important precursor of comprehensive or spatial strategic planning, since it tends to focus on the community holistically as an integrated social, economic, political, and environmental system. Strategic planning thus can help frame and articulate the broad strategies that show how land use, transportation, housing, the environment, education, economic development, health, social services, and urban branding and marketing efforts, among others, might be linked; what key points of leverage and priorities among specific strategies might be; how progress might be measured and communicated; and what lessons might be learned from ongoing implementation. United Cities and Local Governments (UCLG) (http://www.uclg.org), the largest association of local governments in the world (whose member governments cover half the world's population), calls such precursor plans City Development Strategies (CDSs). The specific form and content of CDSs vary across the globe. For example, in Latin America, many local governments have used them to address issues of inequality; in Europe, CDSs often address issues of urban competitiveness; in Asia, some cities have developed CDSs to address issues of economic development, climate change, and the need for social dialog; in North America, CDSs often address issues of financial and infrastructure crises and poverty; and in Africa and Eurasia, CDSs are being used to facilitate moves toward more decentralized planning (United Cities and Local Governments, 2010).

Strategic planning also is typically distinguished from strategic management. Strategic management, also known as performance management systems or results-based management systems are ongoing organizational designs or arrangements for strategically managing the implementation of agreed-upon strategies, assessing the performance of those strategies, and formulating new or revised strategies (Bryson, 2011: p. 323), is a far more encompassing process and is concerned with managing an organization in a strategic manner on an ongoing basis. Strategic management links strategic planning and implementation by adding ongoing attention to budgeting, performance measurement, evaluation, and management, and by adding feedback relationships among those elements (Poister et al., 2010).

Brief History of the Development and Use of Strategic Planning

Strategic planning for public nonmilitary and nonprofit purposes has its roots primarily in the business world. This history has been amply documented by others (Eden and Ackermann, 1998; Mintzberg et al., 2009). However, since around 1980, public and nonprofit use of strategic planning has skyrocketed. This experience and a growing body of literature have indicated that strategic planning approaches, when adapted to public and nonprofit organizational or community purposes, can help these entities deal effectively with their often dramatically changing environments.

In the United States – where strategic planning has proceeded the furthest – surveys indicate that strategic planning is widely used in state agencies and by local governments and jurisdictions. In addition, strategic planning is mandated for all federal departments as a result of the Government Performance and Results Act of 1993 (GPRA) and the Government Performance and Results Modernization Act of 2010. Other nations are also making increased use of strategic planning concepts, procedures, and tools for public and nonprofit organizations and communities (Bovaird and Löffler, 2009; Hughes, 2012).

Benefits of Strategic Planning

When done well, strategic planning offers a number of benefits. Advocates usually point to six main potential benefits. The first is *promotion of strategic thinking, acting, and learning*. The second is *improved decision making*, while the third is *enhanced organizational effectiveness, responsiveness, and resilience*. The fourth is *enhanced organizational legitimacy*. The fifth is *increased effectiveness of broader societal systems* to which public and nonprofit organizations contribute. Finally, strategic planning can *directly benefit the organization's people* by helping them better perform their roles, meet their responsibilities, and enhance teamwork and expertise. There is no guarantee, however, that these benefits will be achieved. For one thing, strategic planning is simply a set of concepts, procedures, and tools that must be applied wisely to specific situations. Also, even when they are applied wisely, there is no guarantee of success.

The empirical evidence on whether strategic planning actually helps is unclear. Reports published by the US Government Accountability Office (2004, 2005) indicate that while federal agency strategic planning has improved, there are still problems in establishing outcome-oriented goals, dealing with cross-agency issues, ensuring adequate consultation with external stakeholders, linking long-term and short-term goals, and identifying the financial, human, and other resources needed to implement strategies. Some also argue that the federal government's mandated top-down, fairly rigid approach limits agencies' abilities to tailor strategic planning efforts to their particular situations. However, there are numerous case study examples of successful strategic planning in public and nonprofit

organizations. In addition, surveys of public managers in the United States and United Kingdom have found that strategic planning efforts generally have helped improve organizational capacities and performance. An important feature of newer research is its emphasis on strategic planning as a *way of knowing* (Bryson et al., 2009; Balducci et al., 2011). A way of knowing means, in the first instance, a practice or customary way of doing something that allows its users jointly to develop, show, or possess knowledge or understanding. In this sense, what clearly matters is the extent to which strategic planning helps its users think, act, and learn strategically.

Approaches to Strategic Planning

Strategic planning may be divided into *process* and *content* approaches. We start by discussing broadly encompassing process approaches, then consider more narrowly conceived process approaches, and finally discuss two content approaches.

The Harvard Policy Model

The Harvard policy model was developed as part of the business policy courses that have been taught at the Harvard Business School since the 1920s (Bower et al., 1991). The approach provides the principal (although often implicit) inspiration behind the most widely cited recent models of public and nonprofit sector strategic planning. The main purpose of the Harvard model is to help an organization develop the best 'fit' between itself and its environment; that is, to develop the best strategy for the organization. One discerns the best strategy by analyzing the internal strengths and weaknesses of the company and the values of senior management, and by identifying the external threats and opportunities in the environment and the social obligations of the firm. Then one designs the appropriate organizational structure, processes, relationships, and behaviors necessary to implement the strategy, and focuses on providing the leadership necessary to implement the strategy.

Effective use of the model presumes that senior management can agree on the organization's situation and the appropriate strategic response, and has enough authority to enforce its decisions. A final important assumption of the model, common to all approaches to strategic planning, is that if the appropriate strategy is identified and implemented, the organization will be more effective. For the model to be useful for public and nonprofit purposes, it typically has to be supplemented with other approaches, such as the portfolio and strategic issues management approaches (also discussed in this section). A portfolio approach is needed because a principal strategic concern at the organizational level is oversight of a portfolio of agencies, departments, or programs. Strategic issues management is needed because much public and nonprofit work is typically quite political, and articulating and addressing issues are at the heart of political decision making.

The systematic assessment of strengths, weaknesses, opportunities, and threats – known as a SWOT analysis – is a primary strength of the Harvard model. This element of the model appears to be applicable in the public sector to organizations, functions, and communities. Another strength is its emphasis on the need to link strategy formulation and implementation in effective ways. The main weaknesses of the Harvard model are that (1) it does not draw attention to strategic issues; (2) it does not consider the full range of stakeholders; and (3) it does not offer specific advice on how to develop strategies, except to note that effective strategies will build on strengths, take advantage of opportunities, and overcome or minimize weaknesses and threats.

Strategic Planning Systems

Strategic planning is often viewed as a system whereby leaders and managers go about making, implementing, and controlling important decisions across functions and levels in the organization. When viewed this way, strategic planning is synonymous with strategic management. Strategic planning systems vary along several dimensions: the comprehensiveness of decision areas included, the formal rationality of the decision process, and the tightness of control exercised over implementation of the decisions, as well as how the strategy process itself will be tailored to the organization and managed. The strength of these systems is their attempt to coordinate the various elements of an organization's strategy across levels and functions. Their weakness is that excessive comprehensiveness, prescription, and control can drive out attention to mission, strategy, and innovation, and can exceed the ability of participants to comprehend the system and the information it produces (Mintzberg et al., 2009). In other words, the systems are prone to driving out the very strategic thinking, acting, and learning they are meant to produce.

Strategic planning systems are potentially applicable to public organizations (and, to a lesser extent, communities), for, regardless of the nature of the particular organization, it makes sense to coordinate at least some decision making across levels and functions and to concentrate on whether the organization is implementing its strategies and accomplishing its mission. However, it is important to remember that a strategic planning system characterized by substantial comprehensiveness, formal rationality in decision making, and tight control will work only in an organization that has a clear mission, clear goals and objectives, relatively simple tasks to perform, centralized authority, clear performance indicators, and information about actual performance available at reasonable cost. While some public and nonprofit organizations – such as hospitals and police and fire departments – operate under such conditions, most do not. As a result, most public and nonprofit strategic planning systems typically focus on a few issues, rely on a decision process in which politics plays a major role, and control something other than program outcomes (e.g., budget expenditures).

Stakeholder Management Approaches

Freeman (1984) states that strategy can be understood as an organization's mode of relating to or building bridges with its stakeholders. For Freeman, a stakeholder may be defined as any individual, group, or organization that is affected by, or that can affect, the future of the organization. He argues, as do others who emphasize the importance of attending to stakeholders, that a strategy will be effective only if it satisfies

the needs of multiple groups. Because it integrates economic, political, and social concerns, the stakeholder model is one of the most applicable approaches to the public sector. Many interest groups have stakes in public organizations, functions, and communities. If the model is to be used successfully, however, there must be the possibility that key decision makers can achieve reasonable agreement about who the key stakeholders are and what the response to their claims should be.

The strengths of the stakeholder model are its recognition of the many claims – both complementary and competing – placed on organizations by insiders and outsiders and its awareness of the need to satisfy at least the key stakeholders if the organization is to survive. Because of its attention to stakeholders, the approach may be particularly useful in working collaboratively with other organizations, since each party to a collaboration must be viewed as a stakeholder (Huxham and Vangen, 2005). The weaknesses of the model are the absence of criteria with which to judge competing claims and the need for more advice on developing strategies to deal with divergent stakeholder interests.

A number of other encompassing process approaches to strategic planning have been developed, primarily in the United Kingdom in the field of operations research. These include Ackermann and Eden's (2011) strategy mapping approach and Friend and Hickling's (2005) strategic choice approach. They are now beginning to be used more widely around the world.

Strategic Issues Management Approaches

These approaches are process components, pieces of a larger strategic planning process. The concept of strategic issues first emerged when practitioners of corporate strategic planning realized that a step was missing between the SWOT analysis of the Harvard model and the development of strategies. That step was the identification of strategic issues, or crucial challenges facing the organization. Many organizations now include a strategic issue identification step as part of full-blown strategy revision exercises, and also as part of less comprehensive annual strategic reviews between major strategy revisions. In addition, many organizations have developed strategic issue management processes actually separated from annual strategic planning processes. Many important issues emerge too quickly to be handled as part of an annual process.

Strategic issue management is clearly applicable to public organizations, since the agendas of these organizations consist of issues that should be managed strategically (Ackermann and Eden, 2011). The strength of the approach is its ability to recognize and analyze key issues quickly. The approach also applies to functions or communities, as long as some group, organization, or coalition is able to engage in the process and manage the issue. The main weakness is that, in general, the approach offers no specific advice on exactly how to frame the issues other than to precede their identification with a situational analysis of some sort. Nutt and Backoff (1995) have gone the furthest in remedying this defect.

The final four process approaches to be discussed are *process strategies*. Two are primarily organizational: logical incrementalism and strategic planning as a framework for innovation. Two are primarily multiorganizational: contract approaches and collaboration. Process strategies are approaches to implementing a strategy that typically has already been developed in very broad outline and is subject to revision based on experience with its implementation. Other important process approaches not discussed in this article, due to space limitations, include total quality management (Cohen and Eimicke, 1998), the mutual gains approach (Susskind and Field, 2010), and the management of culture (Schein, 2010).

Logical Incrementalism

In incremental approaches, strategy is a loosely linked group of decisions that are handled incrementally. Decisions are handled individually below the organizational level because such decentralization is politically expedient – organizational leaders should reserve their political clout for crucial decisions. Decentralization also is necessary since often only those closest to decisions have enough information to make good ones. The incremental approach is identified principally with Quinn (1980), although the influence of Lindblom (1959) is apparent. Quinn developed the concept of *logical incrementalism* – or incrementalism in the service of overall organizational purposes – and, as a result, transformed incrementalism into a strategic approach. Logical incrementalism is a process approach that, in effect, fuses strategy formulation and implementation. The strengths of the approach are its ability to handle complexity and change, its emphasis on minor as well as major decisions, its attention to informal as well as formal processes, and its political realism. A related strength is that incremental changes in degree can add up over time into changes in kind. The major weakness of the approach is that it does not guarantee that the various loosely linked decisions will add up to fulfillment of organizational purposes.

Logical incrementalism would appear to be very applicable to public and nonprofit organizations, as long as it is possible to establish some overarching set of strategic objectives to be served by the approach. When applied at the community level, there is a close relationship between logical incrementalism and collaboration. Indeed, collaborative purposes and arrangements typically emerge in an incremental fashion as organizations individually and collectively explore their self-interests and possible collaborative advantages, establish collaborative relationships, and manage changes incrementally within a collaborative framework (Huxham and Vangen, 2005; Innes and Booher, 2010).

Strategic Planning as a Framework for Innovation

As noted in the 'Strategic Planning Systems' section, strategic planning systems can drive out attention to mission, strategy, and innovation. The systems, in other words, can become ends in themselves and drive out creativity, innovation, and new ventures without which the organization might become irrelevant or die. Many organizations, therefore, have found it necessary to emphasize innovative strategies as a counterbalance to the excessive control orientation of many strategic planning systems (Osborne and Brown, 2012). The framework-for-innovation approach relies on many elements of the approaches discussed in this article, such as SWOT analyses and portfolio methods. This approach differs from earlier ones in four emphases: (1) innovation as a strategy; (2) specific management practices to

support innovation; (3) the development of a mission or 'vision of success,' which provides the decentralized and entrepreneurial parts of the organization with a common set of superordinate goals toward which to work; and (4) the nurturing of an entrepreneurial organizational culture.

The main strength of the approach is that it allows for innovation and entrepreneurship while maintaining central control of direction. The weaknesses of the approach are that typically – and perhaps necessarily – a great many, often costly, mistakes are made as part of the innovation process and that there is a certain loss of accountability in very decentralized systems. Those weaknesses reduce its applicability to the public sector, in particular, in which mistakes are less acceptable and the pressures to be accountable for details (as opposed to results) are often greater. Many nonprofit organizations will also have trouble pursuing this approach because a shortage of important resources will magnify the risks of failure.

Contract Approach

The contract approach is another popular system of institutionalizing strategic planning and management, especially in simple to moderately complex shared-power environments. The contract model is employed for much of the planning and delivery of many publicly financed social services in the United States via either public or nonprofit service providers (Milward and Provan, 2003). The system is also used to institutionalize strategic planning and management in school districts with site-based management.

In this system, there is a *center* that establishes strategic objectives for the jurisdiction or organization as a whole, negotiates contracts with individual units of management, monitors performance, and ensures the integrity of the system. In the language of economics and *principal–agent models*, the center is the *principal* and the individual units of management are the *agents*. The contract between the center and a unit outlines the unit's expected performance, defines its resources, lists other support the unit can expect from the center, and describes a review and renegotiation sequence. Within the framework and legal strictures of the contract, general managers of individual units and their staffs are free to do whatever they think is necessary or desirable to ensure adequate performance. At its best, this approach allows both the center and the individual units to focus on what is important for them – both are empowered to do their jobs better. In such a system, there would be a strategic plan for the center and one for each of the units. Key system concerns would include the content and approach embodied in the center's plan, the center's difficulties in acquiring adequate information, the proper alignment of incentives for the principal and the agents, the difficulties that the center may have in exercising control in the face of a large number of contractors, and ways to ensure adequate investments by the units if they cannot be sure of a long-term contract.

Collaboration Approach

Collaboration represents the final process approach to be discussed. Like contracting, collaboration is increasingly being used to govern and manage in shared-power environments. In fact, the contract system represents a classic form of collaboration, but there are many different (and often more complicated) approaches to collaboration that are more suitable than competitive contracting for situations involving moderate to high levels of complexity and ambiguity (Agranoff, 2007). Human service systems often embody contracting and additional collaboration approaches – contracts for what can be specified and governed with reasonable ease, and supplemental collaboration for those situations involving higher levels of complexity and ambiguity, and therefore requiring greater reliance on trust, shared norms, professionalism, and learning-by-doing for effective governance and management (Huxham and Vangen, 2005). In the latter case, collaboration overlaps with strategic issue management and stakeholder management approaches.

Collaboration is particularly useful when addressing problems for which no organization is fully in charge. Situations of this sort occur when, for example, there is a marked degree of separation between the source and use of funds, services are jointly produced (that is, service recipients are at least partly responsible for effective production, as in mental health services), or the key governance and management task is arranging networks rather than managing hierarchies (Milward and Provan, 2003).

Collaboration involves varying degrees of sharing power and resources (such as information, money, clients, and authority) between units to achieve common ends that could not be achieved separately. There are three different archetypal approaches to collaborative network management (Provan and Kenis, 2007): *lead organization*, *shared governance*, and *partnership administrative organization*.

In the *lead organization* approach, a single partner organization coordinates the major collaboration activities and key decisions. The lead organization has more power than the other partners, who typically are moderate in number. Milward and Provan (2003), in their longitudinal study of mental health service delivery networks, have found that network effectiveness is greatest when there is a strong central integrating unit; clear and consistent lines of authority and accountability embodied in contracts; aligned incentives that give everyone a stake in the success of the network; system stability; and munificent resources. These factors allow constructive norms, social capital, and network learning capabilities to develop, and needed incremental investments and changes to be made.

A key system concern with the lead organization approach is how to achieve the right balance between network stability and adaptability. Because of the importance of stability to performance, a system probably should be changed infrequently and, if possible, incrementally. Another key concern will be the continual need to make sure that incentives are aligned properly so that participants have an incentive to maintain the network and high performance levels.

The *shared governance* approach is likely when no partner has significantly greater power and resources than the others and no external governance organization is formed or mandated. The viability of the approach depends on each organization's involvement and commitment, as the partners are responsible for managing internal and external relations (Provan and Kenis, 2007). Viability also depends on reasonable goal consensus, since exit is always an option for member organizations. If the number of organizations participating in shared governance becomes too large, and if trust levels decline, goal consensus

becomes a bit shaky, and the collaborators have limited collaboration abilities, the partners may create a separate administrative entity – a *partnership administrative organization* (PAO) – to govern the collaboration and its activities and decisions. The PAO is a separate organization whose purpose is to manage the collaboration. A strength of the PAO approach is that presumably the capacity exists to manage the network. A weakness is that the PAO may become disconnected from the concerns of the organizations in the network.

In sum, 'process' approaches do not prescribe answers, although good answers are presumed to emerge from appropriate application. In contrast, 'content' approaches do yield answers. In fact, the models are antithetical to process when process concerns get in the way of developing the 'right' answers. In the 'Portfolio Models' and 'Competitive Analysis' subsections, we discuss two content approaches: portfolio models and competitive analysis. Other important primarily content approaches not covered in this article, due to space limitations, include reengineering and systems analysis.

Portfolio Models

The idea of strategic planning as managing a portfolio of 'businesses' is based on an analogy with investment practice in which an investor manages a portfolio of stocks to balance risk and return. Although the applications of portfolio theory to the public sector may be less obvious than those of the other approaches described in this section, they are nonetheless just as powerful (Bryson, 2011). Many public and nonprofit organizations consist of 'multiple businesses' that are only marginally related. Often, resources from various sources are committed to these unrelated businesses. That means that public and nonprofit managers must make complex portfolio decisions, although usually without the help of portfolio models that frame those decisions strategically.

The strength of portfolio approaches is that they provide a method of measuring entities of some sort (e.g., programs, proposals, or problems) against dimensions that are deemed to be of strategic importance (usually the attractiveness of the option versus the capability to implement it). Weaknesses include the difficulty of knowing what the appropriate strategic dimensions are, the difficulties of classifying entities against dimensions, and the lack of clarity about how to use the tool as part of a larger strategic planning process.

Portfolio approaches can be used in conjunction with an overall strategic planning process to provide useful information on an organization, function, or community in relation to its environment. However, unlike the process models, portfolio approaches provide an 'answer'; that is, once the dimensions for comparison and the entities to be compared are specified, the portfolio models prescribe how the organization or community should relate to its environment. Such models will work only if a dominant coalition is convinced that the answers they produce are correct and the coalition is willing to act based on the answers.

Competitive Analysis

Another important content approach that assists strategy selection has been developed by Michael Porter (1998a) and his associates and adapted to public purposes by Vining (2011). Porter (1998a) hypothesizes that five key competitive forces shape an industry: the relative power of customers, relative power of suppliers, threat of substitute products, threat of new entrants, and amount of rivalry activity among the players in the industry. The assumption is that the stronger the forces, the greater the profitability potential for businesses and industries. For many public and nonprofit organizations, there are equivalents to the forces that affect private industry. An effective organization in the public and nonprofit sector, therefore, must understand the forces at work in its 'industry' in order to compete effectively, and must offer value to its customers that exceeds the cost of producing it. On another level, planning for a specific public function (health care, transportation, or recreation) can benefit from competitive analysis if the function can be considered an industry. Vining (2011) has adapted the five-forces model specifically for a public agency. He hypothesizes that the relevant five competitive forces are the relative power of agency sponsors and customers (a modification), the relative power of suppliers, the threat of substitute products and new entrants (combining two of Porter's forces), political influence (a new force), and the intensity of rivalry among agencies. Vining posits that the stronger the forces, the less the agency's autonomy and the poorer its performance, because it will have little room or incentive to make improvements. Similarly, the weaker the forces, the greater the public agency's autonomy and, up to a point, the better the agency's performance. Too much autonomy is presumed to be bad for performance.

In addition, Porter (1998b) points out that for the foreseeable future, self-reinforcing agglomerations of firms and networks are crucial aspects of successful international economic competition. In effect, not just firms and nations but also metropolitan regions (Shanghai, the Silicon Valley, New York, London, and Paris) are key economic actors. Regions interested in competing on the world stage should therefore try to develop the infrastructure necessary for virtuous (rather than vicious) cycles of economic growth to unfold. In other words, wise investments in education, transportation and transit systems, water and sewer systems, parks and recreation, housing, and so on can help firms reduce their costs – particularly the costs of acquiring an educated labor force – and thus improve the firms' ability to compete internationally.

The strength of competitive analysis as proposed by Porter is that public and nonprofit organizations can use competitive analysis to discover ways to help the private firms in their regions. When applied directly to public and nonprofit organizations, however, competitive analysis has two weaknesses: it is often difficult to know what the 'industry' is and what forces affect it, and the keys to organizational success in the public and nonprofit world are often success in accommodating political forces and using collaboration instead of competition. Competitive analysis for public and nonprofit organizations, therefore, must be suitably adapted and coupled with a consideration of social and political forces and the possibilities for collaboration (Vining, 2011; Huxham and Vangen, 2005).

Conclusions

It should be noted that careful studies of corporate-style strategic planning in the public and nonprofit sectors are few in number, but the evidence on balance indicates that strategic planning in general appears to be helpful (Poister et al., 2010; Andrews et al., 2012). Beyond that, it is possible to reach some tentative conclusions. First, it should be clear that strategic planning is not a single concept, procedure, or tool. In fact, it embraces a range of approaches that vary in their applicability to the public and nonprofit sectors and in the conditions that govern their successful use. Second, while any generic strategic planning process may be a useful guide to strategic thinking, acting, and learning, it will have to be applied with care in a given situation, as is true of any planning process. Because every planning process should be tailored to fit specific situations, every process in practice will be a hybrid. Third, public and nonprofit strategic planning is well on its way to becoming part of the standard repertoire of public and nonprofit leaders, managers, and planners. Because of the often dramatic changes these people and their organizations confront, we can hypothesize that the most effective leaders, managers, and planners at the beginning of the twenty-first century are, and will be increasingly in the future, the ones who are best at 'strategic' planning. Finally, research must explore a number of theoretical and empirical issues in order to advance the knowledge and practice of public sector strategic planning. In particular, strategic planning approaches that are responsive to different situations must be developed and tested. These approaches should specify key situational factors governing their use; provide specific advice on how to formulate and implement strategies in different situations; be explicitly political; indicate how to deal with plural, ambiguous, or conflicting goals or objectives; link content and process; indicate how collaboration as well as competition are to be handled; and specify roles for those involved in the process. Other topics in need of attention include the nature of strategic leadership; ways to promote and institutionalize strategic planning across organizational levels, functions that bridge organizational boundaries, and intra and interorganizational networks; the ways in which information technologies can help or hinder the process; and how best to engage the public in strategic planning. Progress has been made on all of these fronts, but work is clearly necessary if we are to understand better when and how to use strategic planning.

See also: Governance; Innovation; Leadership in Organizations, Sociology of; Learning: Organizational; Organizational Control; Organizational Decision Making; Organizations: Negotiated Orders; Public Administration: Organizational Aspects; Public Sector Organizations; Social Innovation; Strategizing.

Bibliography

Ackermann, F., Eden, C., 2011. Making Strategy: Mapping Out Strategic Success. Sage Publications, Thousand Oaks, CA.

Agranoff, R., 2007. Managing Within Networks: Adding Value to Public Organizations. Georgetown University Press, Washington, DC.

Andrews, Rhys, Boyne, George, Law, Jennifer, Walker, Richard, 2012. Strategic Management and Public Service Performance. Palgrave Macmillan, London.

Balducci, A., Fedeli, V., Pasqui, 2011. Strategic Planning for Contemporary Urban Regions. Ashgate, Farnham, UK.

Bovaird, T., Löffler, E., 2009. Public Management and Governance, second ed. Routledge, London.

Bower, J., Bartlett, C., Christensen, C., Pearson, J., 1991. Business Policy: Text and Cases, seventh ed. Irwin, Homewood, IL.

Bryson, J.M., 2011. Strategic Planning for Public and Nonprofit Organizations, fourth ed. Jossey-Bass, San Francisco, CA.

Bryson, J.M., Crosby, B.C., Bryson, J.K., 2009. Understanding strategic planning and the formulation and implementation of strategic plans as a way of knowing: the contributions of actor-network theory. International Public Management Journal 12 (2), 172–207.

Cohen, S., Eimicke, W., 1998. Tools for Innovators: Creative Strategies for Strengthening Public Sector Organizations. Jossey-Bass, San Francisco, CA:

Eden, C., Ackermann, F., 1998. Making Strategy: The Journey of Strategic Management. Sage Publications, Thousand Oaks, CA.

Freeman, R.E., 1984. Strategic Management: A Stakeholder Approach. Pitman Publishing, Boston, MD.

Friend, J., Hickling, A., 2005. Planning Under Pressure: The Strategic Choice Approach. Butterworth-Heinemann, Oxford, UK.

Hughes, Owen, 2012. Public Management and Administration, fourth ed. Palgrave Macmillan, New York.

Huxham, C., Vangen, S., 2005. Managing to Collaborate: The Theory and Practice of Collaborative Advantage. Routledge, New York.

Innes, J., Booher, D., 2010. Planning with Complexity. Routledge, New York.

Lindblom, C.E., 1959. The science of muddling through. Public Administration Review 19, 79–88.

Milward, H.B., Provan, K.G., 2003. Managing the hollow state. Public Management Review 5 (1), 1–18.

Mintzberg, H., Ahlstrand, B., Lampel, J., 2009. Strategy Safari: A Guided Tour Through the Wilds of Strategic Management, second ed. Financial Times – Prentice Hall, London.

Nutt, P.C., Backoff, R.W., 1995. Strategy for public and third sector organizations. Journal of Public Administration Research and Theory 5 (2), 189–211.

Osborne, Stephen P., Brown, Louise, 2012. Managing Public-Sector Innovation. Routledge, London.

Poister, T.H., Pitt, D., Edwards, L., 2010. Strategic management research in the pubic sector: a review, synthesis, and future directions. American Review of Public Administration 40 (5), 522–545.

Poister, T.H., Streib, G., 1999. Strategic management in the public sector: concepts, models, and processes. Public Management and Productivity Review 22 (3), 308–325.

Porter, M., 1998a. Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, New York.

Porter, M., 1998b. The Competitive Advantage of Nations: With New Introduction. Free Press, New York.

Provan, K.G., Kenis, P., 2007. Modes of network governance: structure, management, and effectiveness. Journal of Public Administration Research and Theory 18, 229–252.

Quinn, J.B., 1980. Strategies for Change: Logical Incrementalism. R D Irwin., Homewood, IL.

Schein, E., 2010. Organizational Culture and Leadership, fourth ed. Jossey-Bass, San Francisco, CA.

Susskind, L., Field, P., 2010. Dealing with an Angry Public: The Mutual Gains Approach to Resolving Disputes. Free Press, New York.

United Cities and Local Governments, 2010. Policy Paper on Urban Strategic Planning: Local Leaders Preparing for the Cities of the Future. United Cities and Local Governments, Barcelona, Spain.

US Government Accountability Office (GAO), 2004. Results Oriented Government: GPRA Has Established a Solid Foundation for Achieving Greater Results. US Government Printing Office, Washington, DC. GAO-04-38.

US Government Accountability Office (GAO), 2005. Results Oriented Government: Improvements to DHS's Planning Process Would Enhance Usefulness and Accountability. US Government Printing Office, Washington, DC. GAO-05-1011T.

Vining, Aidan R., 2011. Public agency external analysis using a modified "five forces" framework. International Public Management Journal 14 (1), 63–105.

Strategizing

J-C Spender, Kozminski University, Warsaw, Poland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.M. Bryson, volume 22, pp. 15145–15151, © 2001, Elsevier Ltd.

Abstract

The business strategy field covers business practice and academic research but the two are not strongly related. A history of the topic is presented and the two views are contrasted. One is the academic discipline's view that strategy is the rigorous pursuit of optimal performance. The other, a practitioner's view, sees strategizing as addressing the 'What do we do now?' question, always present for strategists to inhabit uncertain situations. We follow Knight's intuition that profit only arises as strategists engage real world uncertainties through their firm's practice.

The Strategy Field's History

As strategy textbook authors are quick to point out, strategizing is an ancient topic. The beginnings of todays' theorizing came with Chandler's analysis of the relationship between 'strategy' as the firm's chosen market engagements and 'structure' as its chosen administrative means (Chandler, 1962). He sought an efficient 'fit' between environment and firm, between strategy and structure, a view so taken for granted that it is difficult to find any other definition. The external–internal fit (EIF) paradigm dominates the strategy textbooks – as in the external assessment of threats and opportunities and the internal assessment of strengths and weaknesses (SWOT). EIF was jolted by Porter's 'competitive strategy' analysis (Porter, 1980). His thinking derived from industrial organization economics and paid little attention to efficiency. It focused on protecting the firm's rent streams (above average returns) from those economic actors able to disturb or appropriate them. Though interfirm rivalry is considered, Porter's is not an industry analysis, as many assume, for many of the players in his familiar 5-box model are actually outside the focal firm's industry (Spender and Kraaijenbrink, 2011). The underlying presence of market power ('force' in Porter's terminology) showed that firms are not the passive adapters implied in the EIF analysis but can push back aggressively against those eroding their rent streams or threatening to – with repricing, R&D, new products, negotiating, and signaling. Porter summarized his competitive approach by contrasting (1) differentiation, the rent-protection paradigm, and (2) cost leadership, restating the EIF paradigm – though he noted later that his 'generic strategies' analysis was an 'afterthought.' The shift from efficiency to rents was also in tune with changing politics, as the Reagan and Thatcher administrations rolled antitrust regulations back to give monopolists freer rein.

By the end of the 1980s, strategy researchers were looking for ways to explain how firms might acquire and protect their rent streams in dynamic competitive situations (Rumelt et al., 1991). One conjecture was that a firm's rents would derive only from its ownership of those valuable, rare, and inimitable (VRIN) resources that could be integrated into the firm's operations – the resource-based view (RBV). Missing, though, was any explanation of how firms might (1) acquire VRIN resources without their rent potential being stripped in the factor markets or (2) transform these resources into added economic value. Answering either question would escape the tautology that valuable resources beget value. Wildly popular among strategy academics, in spite of being little supported by empirical research or having clear managerial implications, the RBV provoked a second conjecture, sometimes dubbed the knowledge-based view (KBV) (Kraaijenbrink et al., 2010). The intuition was that the firm's 'learning' can produce new rent-yielding VRIN resources internally, where they will be shielded from market forces and be inimitable. Theorists borrowed earlier notions of firms as 'bundles of resources, competences, and capabilities.' There was attention to the dynamics of learning with notions like 'dynamic capabilities' (DCs), resources that would enable firms to generate new rents when changing circumstances eroded their existing rent streams. But missing, as with the RBV, was an appropriate 'theory of the firm' as comprising resources, competences, or capabilities – dynamic or otherwise.

The strategy discourse was again reshaped or 'colonized' by new connections with post-WWII microeconomics, especially with responses to Coase's questions about the 'nature of the firm': "Why do firms exist, why are their boundaries where they are, why are their internal arrangements as they are, and why is their performance so varied?" (Coase, 1937). Researchers began to incorporate transaction cost analysis, principal-agent theory, information asymmetry, moral hazards, contracts, and allocation of property rights. There was empirical research into vertical and horizontal integration, interfirm alliances, mergers and acquisitions, initial public offerings (IPOs), networks, and alternative modes of financing and governance. Countering the microeconomic impulse was renewed attention to the EIF paradigm and strategizing as planning for the long term. This focused on the intraorganizational processes of strategic plan formulation (strategy process, strategy-as-practice, case analysis), their behavioral and psychological aspects (top management teams, HR strategy, incentive and pay schemes, organizational communication, knowledge management, and behavioral strategy), and on improved environmental analysis (business intelligence, data mining, scenario planning, and risk analysis). In the background a single metaphor – the strategist viewing the firm's situation objectively, gathering information, and processing into a rigorous actionable conclusion. It dominates strategy research and helps

explain why journals are packed with quantitative data that probes for causal relationships between outcomes, as dependent variables, and situational information as independent variables. There are plenty of reasons to doubt this metaphor's practical usefulness, for Simon's notion of 'bounded rationality' looms. No business people – or academics – are 'perfectly rational.' But still our discipline's textbooks and articles are written as if conclusive data are available, meaningful, and computable. Theory is the thing in question, the data to be computed is not – and thus practice does not appear.

But the essence of strategizing is the practical difficulty of choosing action in the face of a real situation's uncertainties. Strategizing deals with the challenges of acting 'mindfully' in one-off uncertain situations, unique or, more specifically, situations open to being made unique by the strategist's choices. Strategizing is how human beings pursuing their goals identify what is specific to their situation, the features that are material to action that cannot be generalized, and so neither covered by theory, nor ignored. There is a connection to Polanyi's notion of tacit knowing; strategy points directly to the tacit dimensions of human action rather than to its explicit rational dimensions. The strategist's judgment particularizes the firm's situation. Given the pervasiveness of uncertainty, it is difficult to understand why our discipline has devoted so much effort to analyzing situations that deny it and make the analysis irrelevant to practitioners.

Engaging the Unique

Strategic practitioners exercise their judgment under uncertainty in ways that are subjective and personal. In practice, the strategist's world is uncertain in three ways – their own ignorance, or the situation's indeterminacy, or the incommensurability of differing aspects of what the strategist's knows (Spender, 2007). The strategist must construct a view within which to think. This takes five steps: constraint discovery, constraint identification, judgment, creating a language for persuading others' judgment, and implementing practice – which includes managing the learning from the implementation experience. First, step 1, the strategist must discover the degree or manner in which the situation (knowledge-absence chosen) is underdetermined and open to her/his judgments yet constrained in ways beyond her/his power to change. This identifies the strategist's 'option space.' The positivist way of thinking is that independent variables must fully determine the strategist's decision. But while real-world choice is always constrained by some 'facts' because of uncertainty it is never constrained to the point of being fully determined. As actors in an uncertain world we are always strategists and always have choices, even under the most constraining circumstances. It follows that the facts to be taken into account must be selected, and they are never self-evident and beyond doubt. What must be attended to? Simon argued that managerial attention was the organizational resource in shortest supply, so the essence of strategic choice lies in knowing what to pay attention to, on the grounds that much is relevant but not everything.

The private sector business strategist has little choice over government regulations affecting the firm's goods and services, patents, and other forms of valuable knowledge to which the firm has no access, labor contracts made on an industry-wide basis, constraints to the supply of the relevant factors of production, many costs, such as transportation costs and taxes, and so on. This leads to our realizing that each firm or, rather, its business model, is unique. One question is "How many constraints need to be taken into account?" for the list threatens to be endless, almost anything might matter. The strength of strategic consulting tools is their confidence that the strategist's problem can be boiled down to a few essentials: strengths, weaknesses, opportunities, and threats (SWOT) or cash cows, question marks, and stars Boston Consulting Group (BCG) or, in the case of the balanced scorecard – matters financial, marketing, internal, and growth. Empirical research by the present author suggested most business models address around a dozen strategic issues (Spender, 1989).

With an inventory of constraints in place, the strategic opportunity space is not finally framed until the practical nature of each constraint has been explored. This is step 2, exploring whether the specific constraint is a matter of ignorance, of indeterminacy, or of incommensurability. In practice the constraints to a firm's strategic maneuvering are seldom known as brute facts, but are always open to interpretation and manipulation. Constraints of ignorance can be shifted advantageously by research. Perhaps R&D has an impact, or 'data mining,' or 'industrial espionage.' The firm's activity is here directed toward what is in principle knowable. But sometimes other actors, such as competitors or regulators, put a different type of constraint in place, not knowable until enacted. But they can also be moved to strategic advantage by exercising power, perhaps negotiating directly or indirectly with the others, or lobbying, signaling, or bribery. Finally the constraint in question might be incommensurability "How do these facts or relationships fit together?" The strategist's capacity here might be labeled an intellectual or imaginative power. Overall the strategist always has a degree of power to act on or refashion the strategic opportunity space, to address it proactively rather than remain passive.

Step 3 is the practical act of strategic judgment, and this cannot be theorized, of course, no more than it is possible to theorize why Picasso painted in one way rather than another – even after the constraints of the materials, subject, and timing are appreciated. Nonetheless some useful things can be said about acts of judgment. Von Clausewitz wrote of personal preparation for military strategizing, deep study of military history and biography, study of facts such as the terrain and the military resources available, but also of the need to clear the mind so that strategic insights can arrive; perhaps sleep on a major decision or take a walk to clear the head, and so on (Sumida, 2008). Von Clausewitz was especially interested in strategizing as choosing the place, time, and manner of engagement. Timing may be everything.

With step 3, the business model comes together and the language created becomes the means to justify the choice of constraints and how they are interpreted. It is an innovation bringing something new into the world as a result of the practice it precipitates. Any action in an uncertain world is an innovation, of course, but the most profound and advantage-yielding innovation a strategist can generate is a new business model, a new way of adding economic value in the world. Its novelty may be modest, a new way of dealing with the same set

of constraints (incremental), or more innovative, provoking an entirely new way of dealing with them (radical). But the most significant innovations arise from changing or repositioning constraints, especially those impacting others' business models (game changing).

With the new business model in place, step 4 follows and leads toward strategic practice. The ultimate test of strategizing is 'Does it work?' The relationship between analysis and practice hinges on the possibility of an analysis that tests the business model's viability ahead of implementing it, irreversibly committing resources. So long as bounded rationality is taken seriously there can be no possibility of taking everything into account, no preparation can secure the actors from surprise. The doing is always different from the thinking about it. There comes that crucial point when the actor must 'jump in with both feet' and expose her/himself to the unconsidered. In business situations the rawness of this act can be shared because most activity is collaborative. Others can help. So the essence of step 4 is strategic conversation with those others whose judgment might bear usefully on the situation, increasing the strategist's confidence without, of course, ever entirely eliminating the risk (Spender and Strong, 2014). The strategist always remains responsible.

Engaging Others

With steps 1, 2, 3 completed and step 4 in hand, the strategizing process focuses on the creation of the firm. The bridge between entrepreneurial/strategic ideas and the real world is crossed by practice. Instead of viewing the firm as a taken for granted command and control machine to be directed this way or that, the firm must be considered problematic, something to be constructed by the strategizing process. It is a dynamic and tentative bundle of collaborative practice – never tidy, with elements of error and surprise intermingling. Value is only created by practice, instantiation. Planning and thinking are fine, but the firm's purpose is the added value that arises through its practices and their specificities. The result is the firm and as close as one can get to its bundle of activities, a situationally specific language that embraces, shapes, and justifies the practices of those involved. As a result the business is a discursive life space, its actors inhabit as they apply their skills and generate value. But, who gets told what is in this language?

The difference between analysis and synthesis has a huge impact on communications within the firm and the network of those who can respond to the strategist's choosing. There are the external agents indicated in stakeholder theory: moneymen, customer action groups, and focus groups. There are the R&D people. There are the employees, especially valuable sources of judgment when adding value calls for skills in the transformation of inputs into outputs or dealing with customers' wishes. There are the lead users, open innovators, and market makers. All may have important judgments to contribute to the enterprise. The wise strategist resists the temptation that comes with responsibility, to cut herself or himself off and hide out in isolation. Judgment is necessary to manage the delicate balance between being present as a leader, remaining in charge and responsible, or as a listener, open to others' opinions, being aware of what is going on in and around the situation being strategized. The judgments to be made about who and what to listen to are complementary to the process of promulgating the strategic choices arrived at (Spender and Strong, 2014). Who should be heard and who told? Organizational communications, especially about strategy, matter greatly and interact with the patterns of power and authority that are essential to the firm's smooth functioning.

The difference between analysis and synthesis illuminates communication. Given the strategist's primary function is to lead the firm toward engaging uncertainties in ways that result in added value, strategic communication has little to do with broadcasting conclusions or giving directions. The uncertainties chosen must be grasped and resolved if the firm is to hold together, function, and prosper. When the leader absorbs all the uncertainty those acting in the situation are aware of, the result may not be entirely beneficial. The leader's principal contribution is to shape workplace activity – which includes the firm's network of external arrangements and distribution and purchasing channels – so that the remaining uncertainties can be engaged and resolved by those doing the work. No matter how thorough the leader's strategizing, there will be uncertainties remaining that can only be resolved as the work is being done. But the worker's judgment and synthesizing capabilities must be applied in ways that reflect the strategist's choice of objective. On the one hand the experience of working is enriched when there are uncertainties to be resolved and judgments to be made in addition to completing the task. For many this marks the difference between empowering work and mindless labor. But if too much of the uncertainty remains unresolved or too little support is provided, the work situation becomes overly stressful, discouraging, and maybe undoable. On the other hand if too little uncertainty remains this may dull the subordinates' work. They find little that engages their creative capabilities, whose exercise binds them into the activity.

By making uncertainty central, strategizing goes way beyond developing a plan of action to address a chosen objective. It becomes a matter of building a human apparatus with which the strategist grasps and engages selected knowledge-absences that are only resolved through creative practice, and so yield an added value. It is the vital core of the firm as the most important apparatus invented to advance the human condition. Highly specific language is its key as a managed domain of interpersonal practices. The business model – as the firm's specific language – builds the firm by communicating with those whose judgment and activities bring it to life (Spender, 2014). Direction and instruction can be contrasted against encouraging proactive creative engagement. Direction and instruction, especially when both necessary and sufficient to completing the task, positions the worker as the strategist's passive instrument, channeling the strategist's thinking – vs the different relationship implied when the strategist passes specific uncertainties to others to resolve, seeking their judgment in their resolution, to deal with uncertainties the strategist was not able to resolve. The one presumes a submissive authority relation, the other presumes respectful creative collaboration; and nontrivial strategizing demands the second because the strategist cannot arrive at a full and final resolution of all the uncertainties arising from the knowledge-absences chosen. How is a collaboration of diverse judgment achieved?

Much of the literature on organizational communication presumes the firm is like an electrical network with instruction signals passed without loss through the administrative apparatus to individual work roles that are thereby fully controlled – while accountability and performance signals flow back to the strategic nerve center. The firm is envisaged as the rigorous solution to a communication system design problem. Strategy has little relationship to this metaphor. The distinction made earlier between strategically trivial decision making in circumstances in which full data are available and the nontrivial situations in which uncertainties are engaged, is paralleled by the distinction between instruction and the human communication that results in creative engagement. If the strategic task is nontrivial, it follows the firm is more than a design problem. Rather, strategizing works by capturing the vitality of the firm's participants, and thereby the firm gains the capacity to absorb and resolve uncertainty, and make profit. The firm is a group of creative people unified by a firm-specific language. They collaborate, applying their judgment to engage chosen uncertainties far beyond those that can be handled by the individuals alone – a movie production, perhaps, or an expedition to the Moon, or a modern hospital. The key to managing the distribution of uncertainties is not to pass them to others to deal with as they wish. To the contrary, the others' judgment has to be harnessed and channeled toward objectives the strategist has chosen.

It is useful to contrast mechanical or electrical communication – such as that which takes place between computers – with the broad range of human communications considered within the study of rhetoric. Machine communication depends on the presence or absence of a signal, of data. The meaning of the data is not at issue. Because of our bounded rationality and doubts about facts, human communication always concerns both data and meaning (Spender, 2007). Meanings are what we attach to data and managing them is as much of a challenge as is getting data. Human communication is generally less about data than about meaning. Paying attention to the distinction between data and meaning separates the strategic task into two complementary spheres. On the one hand there are steps 1, 2, and 3 – discovery, applying judgment and getting to grips with an uncertain situation, and selecting and creating the business model's facts and meanings. On the other, step 4, the process of drawing other peoples' judging capacity into the firm by managing the meanings they attach to the data considered in their work.

Rhetoric shows how meanings are managed, for they cannot be communicated as data. Ultimately meanings are what we create and add in order to make sense of the data we receive. As Aristotle told us, human communication has three parts or 'modes' that are incommensurate – *logos*, *ethos*, and *pathos*, the different types of communication that reflect the different ways in which human beings know, understand, and debate meaning. *Logos* concerns communication that appeals to an audience's rationality and objectivity and is the mode of communicating that works well with data, when meanings are not problematic and can be taken for granted – time of day, number of points in a basketball game. *Ethos* concerns an appeal to the social relationship between audience and rhetor (leader/strategist) and draws people toward letting their meanings change. When an admired person speaks, we listen; we may be persuaded to think in a new way because we respect the speaker. *Pathos* appeals to the audience's emotions. Emotion is the driver to human action; persuasion alone is insufficient. Getting people to agree is not the same as getting them to act as you wish. Rhetoric tells us that effective communication toward chosen actions synthesizes all three modes into a situated discourse.

Strategizing arcs from the realm of thinking into purposive practice but cannot ever be separated from its action dimensions. There can be no abstract or rigorous theory of strategizing. The language that the leader/strategist needs to harness the judgment of others toward the firm's chosen goals in step 4, and thereby transform the chosen strategy into the added value sought, stands on the empirical constraints explored in steps 1, 2, and 3. The language developed persuades the other actors to support the strategist by pointing out (1) what to pay attention to, and (2) how to interpret and act on what they discover. There is an additional wrinkle here in that human communication is far from being bounded by data and meaning – taken together to comprise 'information.' Strategizing reaches beyond this, and the various forms of nonverbal signaling and into the firm's tacit knowledge resources. The strategist seeks to harness the others' skills, often tacit, as well as their explicit information resources, for their support is crucial to implementing the strategy successfully. Rhetoric's focus on practice reaches beyond merely shaping how people think and delves into engaging others to act resourcefully, imaginatively, and purposively in the specific context of their action.

At the same time, since these supportive and skillful others are being drawn into contributing imaginatively by being immersed in the value adding practices – the antithesis of participating dumbly – so they become aware of the nature of the uncertainties being engaged and of subtlety of their relationship with 'the leadership.' Whose judgment rules? Given uncertainty, the balance is constantly open to question. But what questions? Rhetoric can help clarify here, for one branch of rhetoric covers 'question theory.' Most know of rhetoric as the process of persuasion, a speaker engages the audience and moves them to act. Those in the audience are thereby transformed into people they were not previously – they get 'with the program.' They are immersed in a new way of living and, because that is also under uncertainty, a new set of existential questions arises for each of them. So everyone involved and committed to the firm looks to the strategist/s to reassure them that the firm (1) has a viable business model that adds value, (2) has a means of estimating its viability ahead of implementing it, (3) appreciates the principal threats to this viability, (4) has a sense of how to change the firm's business model as needed, and (5) has a sense of how long it will remain viable.

This places the business model between the strategist's identity, intentions, and context in two different ways. First, it encapsulates the firm's identity and intentions as sources of the judgments articulated into its strategy. These shape how those involved view the firm and its context. But the way the actors view their experience of these is likewise shaped for all learning is mediated and conditioned by what is known already. The business model is both a bridge out from thinking to practice, and a bridge back from practice to better thinking. Thus, step 5 of the strategizing process does not stop with measuring the firm's performance against its strategic intentions. It embraces managing what those involved have learned from the practices

that resulted from the previous steps. Step 5 completes the cycle of strategizing and finalizes the business model as a dynamic pattern of managed responses in an ongoing and incompletely predictable context.

Conclusion

Business strategy is a field of research and of managerial practice. The last few decades of academic research have produced a big literature but business practice has not been much informed; though there are important exceptions in highly specialized areas that bring high-speed and large-scale computing and data analytics to business decision making. Beyond these it seems clear that researchers and practitioners have moved in separate directions. Academic research has flourished nonetheless; to the point the field now comprises a large variety of mutually inconsistent ways of looking at its central problematic – the generation of theory that explains firm performance and competitive advantage. Attempts to rein in this proliferation have failed and show no prospect of succeeding.

Instead of hoping for a long-sought overarching theory of business strategy that would underpin rigorous research and show businesses how best to meet their goals and gain competitive advantage, there is a different way of proceeding. It follows from Knight's distinction between circumstances that are (1) knowable vs (2) situations of irrevocable uncertainty (Knight, 1921). Instead of presuming, as do most theorists today, that an economy is fully knowable, Knight adopted a humbler position, that man's understanding of reality is always limited – and economies that are real call for strategizing. Knight's approach was not new, obviously; but he went further and presumed the possibility of economic profit and added value lay in imaginative (entrepreneurial) engagement with real-world uncertainties. Firms are an apparatus for successfully and collaboratively engaging these. Objectivity, as much as possible, is our discipline's dominant metaphor, the realm of rational man theorizing – but permits no consideration of Knightian uncertainty or bounded rationality. Consequently a rigorous analysis must proceed on the assumption that the 'answer,' the optimum action, lies in the data. This makes strategizing 'trivial' in the sense that it does not manifest the identity, intention, imagination, and values of the strategist as a creative human agent. The analysis is dehumanized and abstracted from the world of socioeconomic action where real value is created. More fundamentally, it does not take the firm as an apparatus for creating economic value seriously; yet the whole point of business practice is to generate new value and seize some part of it and thereby advance the human condition.

Adopting Knight's notions of engaging uncertainty, strategizing becomes profoundly subjective, the process of applying personal judgment to generate and manage practices that manifest the will of the actors. The difference between the metaphors is stark. The 'rational man' based objectivist metaphor considers only what is known or can be discovered using the methods of natural science. The second takes off from the strategist's subjective engagement with selected knowledge-absences that seem to offer opportunities for profit. The first metaphor's process is well understood. Facts are gathered and computed. The conclusion is articulated into instructions that are issued to subordinates whose function is to implement them as closely as they are able. Their progress is monitored; problems are handed to specialists to 'reengineer' the process and its controls. The metaphor is mechanical, logical, and rigorous. The second metaphor's process is utterly different. Aside from the exercise of judgment about what to pay attention to, how to interpret what is known in the light of the firm's goals, and how to manage the resulting activity, the relationship between those making the choices and those transforming them into value are very different for the subordinates' judgment is drawn into the process rather than being dismissed.

Overall, strategizing is the process of regarding situations, gathering data, imagining possibilities, choosing goals and means, and communicating conclusions. Because everyone's knowledge and judgment is limited, strategizing calls for energetic and imaginative commitment and collaboration, getting all hands to the pumps, making many types of contribution, for a firm's situation is always full of uncertainties but also of potential, to be seen as a glass half full, awaiting being filled with new value, rather than a glass half empty, a result of being not fully known and predictable.

See also: Behavioral Theories of Organization; Business Models; Entrepreneurship and Economic Growth; Innovation; Narrative Approaches to Organizations; Organization: Overview; Organizational Decision Making; Organizational Emergence and Firm Formation; Strategic Adaptation; Strategic Management.

Bibliography

Chandler, Alfred D., 1962. Strategy and Structure: Chapters in the History to the American Industrial Enterprise. MIT Press, Cambridge, MA.

Coase, Ronald H., 1937. The nature of the firm. Economica N.S. 4 (16), 386–405.

For a comprehensive overview of mainstream academic strategy:

Grant, Robert M., 2010. Contemporary Strategic Analysis, seventh ed. John Wiley & Sons, Chichester.

Knight, Frank Hyneman, 1921. Risk, Uncertainty and Profit. Houghton Mifflin Company, Boston, MA.

Kraaijenbrink, Jeroen, Spender, J.-C., Groen, Aard, 2010. The resource-based view: a review and assessment of its critiques. Journal of Management 36 (1), 349–372.

Porter, Michael E., 1980. Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, New York.

Rumelt, Richard P., Schendel, Dan, Teece, David J., 1991. Strategic management and economics. Strategic Management Journal 12, 5–29 (winter special issue).

Spender, J.-C., 1989. Industry Recipes: The Nature and Sources of Managerial Judgement. Blackwell, Oxford.

Spender, J.-C., 2007. Data, meaning and practice: how the knowledge-based view can clarify technology's relationship with organizations. International Journal of Technology Management 38 (1/2), 178–196.

Spender, J.-C., 2014. Business Strategy: Managing Uncertainty, Opportunity, and Enterprise. Oxford University Press, Oxford.

Spender, J.-C., Kraaijenbrink, Jeroen, 2011. Why competitive strategy succeeds – and with whom. In: Huggins, Robert, Izushi, Hiro (Eds.), Competition, Competitive Advantage, and Clusters: The Ideas of Michael Porter. Oxford University Press, Oxford, pp. 33–55.

Spender, J.-C., Strong, Bruce, 2014. Strategic Conversations: Creating and Directing the Entrepreneurial Workforce. Cambridge University Press, Cambridge.

Sumida, Jon Tetsuro, 2008. Decoding Clausewitz: A New Approach to on War. University Press of Kansas, Lawrence, KS.

Strauss, Anselm (1916–96)

Isabelle Baszanger, CNRS-CERMES, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Together with an impressively imaginative and productive cohort of University of Chicago-trained sociologists, Anselm Strauss is emblematic of the long-lasting fecundity of American symbolic interactionism and its major impacts on sociology, too often neglected today. Strauss' personal contributions are as extensive as they are diverse, covering an amazing range of areas: sociological theory, social psychology, urban sociology, sociology of occupations and professions, sociology of work, qualitative research methodology, sociology of organizations, and medical sociology. He also helped launch new subdisciplines including medical sociology and sociology of the body and of emotion. The concept of action is the Ariadne's thread, which weaves the work of Anselm Strauss together.

Together with an impressively imaginative and productive cohort of University of Chicago-trained sociologists (Freidson, Goffman, Becker, Davis, Roth), Anselm Strauss is emblematic of the long-lasting fecundity of American symbolic interactionism and its major impacts on sociology, too often neglected today. Strauss' personal contributions are as extensive as they are diverse, covering an amazing range of areas: sociological theory, social psychology, urban sociology, sociology of occupations and professions, sociology of work, qualitative research methodology, sociology of organizations, and medical sociology. He also helped launch new subdisciplines including medical sociology and sociology of the body and of emotion.

The concept of action is the Ariadne's thread, which weaves the work of Anselm Strauss together. There is action around personal and group identity (*Mirrors and Masks*, 1959; *Status Passage*, 1971), action around death (*Awareness of Dying*, 1965; *Time for Dying*, 1968; Anguish 1970, all with Barney Glaser), the action of various professional and nonprofessional groups negotiating order and change in organizations (*Psychiatric Ideologies and Institutions*, 1964, with Leonard Schatzman, Rue Bucher, Danuta Erlich, and Melvin Sabshin) and, more generally, the action of negotiation and other processual forms in the doing of social life (*Negotiations: Varieties, Contexts, Processes, and Social Order*, 1979), action to control the impact of new technologies and chronic illnesses on medical work (*The Social Organization of Medical Work*, 1985, with Shizuko Fagerhaugh, Barbara Suczek, and Carolyn Wiener), and action around the management of chronic illness in everyday life (*Chronic Illness and the Quality of Life*, 1975, with Barney Glaser; *Unending Work and Care: Managing Chronic Illness at Home*, 1988, with Juliet Corbin). This concept of action, accompanied by its structural environment (contexts and conditions, and social worlds and arenas), marks the specificity of his interactionist approach culminating in his last theoretical book, *Continual Permutation of Action* (1993).

Intellectual Genesis

Strauss was born in 1916 in New York and brought up in an 'ordinary' Jewish American family. His father was a high school teacher and his mother a housewife. He wanted to study medicine and took his bachelor's degree in biology at the University of Virginia (1939). However, along the way he discovered psychology and sociology. Eventually, he opted for sociology and took his master's and doctoral degrees at the University of Chicago (PhD 1945). At Chicago, Strauss was influenced especially by Herbert Blumer and Everett Hughes and the sustained presence of Mead, Dewey, Thomas, and Park.

From 1944 to 1952, Strauss taught sociology, first (until 1946) at Lawrence College in Wisconsin, then at Indiana University (until 1952). There, with his colleague Ralph Lindesmith, he worked on a social psychology textbook (*Social Psychology* 1949, now in its eighth edition). In 1952, Strauss returned to the University of Chicago as an assistant professor. Embarking on a new program of research dealing with occupations and professions, in collaboration with Hughes, he started with a study of art students and young artists. In 1953, at the request of Nelson Foote, he wrote a paper on social psychology for an edited book aimed at graduate students, never published. This paper, distributed in a mimeographed version under the title, "An Essay on Identification," became the first part of *Mirrors and Masks: The Search for Identity*. However, completing the book took 4 years as he wrestled to introduce a more structural dimension. As then developed in sociology, social psychology left very little room for interaction between the individual and organizations. Strauss then laid claim to a new sociological approach to identity through synthesizing symbolic interactional and social organizational perspectives into a workable, suggestive social psychology. He used the concept of identity not to define, but as a conceptual lever to reveal heterogeneous relationships between individual identities and group identities, and to establish explicit linkages between structure and interaction. This theme of identity through collective commitment persists throughout Strauss' work and informs other concepts, such as social worlds, negotiated order, and trajectory.

The book appeared in 1959. It is a major milestone in Strauss' intellectual trajectory. The theoretical basis for a number of themes in his subsequent work is sketched. However, the process of construction through abduction, taking back and forth between empirical materials and conceptualizing, is still absent from *Mirrors and Masks*. The first

elements of grounded theory were laid down in 1957 when he did his short, first 'real' fieldwork as part of the famous research program initiated by Hughes and carried out primarily by Howard Becker, Blanche Geer, and Strauss. Using comparisons among the different hospital services to which the young resident doctors were allocated in succession as a methodological tool, Strauss highlighted a period of reciprocal assessment among interns, residents, and the hospital's permanent medical and nursing staff. In each iteration, the shift to a new service gave birth to new structural contexts, requiring afresh the work of defining the social order in ongoing processes of negotiated assessment. Strauss integrated the idea of active individuals responding creatively to the situations they encounter with the fact that these responses, these individual and group actions, can only be analyzed as embedded in sets of prior conditions, which themselves influence the present situation. The present consequences of an action become, in turn, the prior conditions for actions yet to come. Capturing the flows of such trajectories and their consequences was his goal.

This study was Strauss' first step in an area of research – medical sociology – that he contributed develop to for the rest of his career. Strauss left the University of Chicago in 1958 to head a research project on psychiatric institutions, and in 1960 took a new post at the University of California, San Francisco (UCSF), where he established both the Department of Social and Behavioral Sciences and the Doctoral program in Sociology in the School of Nursing.

A Privileged Field of Research: Medicine and Illness

Psychiatric Institutions

In 1957 Strauss led an interdisciplinary research team in an ethnographic study of two psychiatric hospital departments deliberately selected for their major differences. This work – *The Hospital and Its Negotiated Order* (1963) and *Psychiatric Ideologies and Institutions* (1964) – offers a new framework for analyzing complex organizations. Conceptualized as sites of permanent negotiation, continuous processes of generating and acting upon tacit agreements allow organizations to function. The negotiations include unofficial 'working' arrangements, official decisions among the various professional and other groups concerning the overall strategy of organization, and ways of literally dividing up the work.

Parallel to this vision of organization, professions like medicine are analyzed as nonhomogeneous worlds constantly changing and composed of segments, each of which harbors different identities and ideologies concerning the specificity of their practices and their central activities (*Professions in Process*, 1961). The general organization of a profession is shaped by conflicts and competition among segments in movement and change. Numerous compromises are established by the representatives of the different ideologies of the professional segments involved both with other professional segments and with clients/patients as well. There are also microlevel negotiations regarding the philosophies specific to each person, linked to their individual biographies. Professions are thereby contextualized, and Strauss has integrated the mesolevel into his work. All of Strauss' later work assumes such meaningful heterogeneities.

Working on Death

Anticipating one of the major issues in American society, in collaboration with Barney Glaser, Strauss participated in opening up a new field of research – the management of death in hospitals. The question of interactions among dying patients, hospital personnel, and family members was at the core of their work. In a ground-breaking series of books and papers, they introduced major concepts now part of the contemporary stock-in-trade of sociology. First, a particular 'awareness context' is seen as patterning the mode of interaction around the dying person, who is also defined as an actor in the situation. Different possible combinations of 'who knows what' exist, characterizing different types of awareness contexts. These are organized around various dimensions such as the number of people involved, acknowledgment of the awareness (pretense or not), the degree of awareness, the object of the awareness (here the dying person), or the awareness that the others have of the object (*Awareness of Dying*, 1965). Each context of awareness characterizes different modes of interaction. Strauss and Glaser defined four major contexts: closed awareness, suspected awareness, mutual pretense awareness, and open awareness. For each, the social regulations and other structural conditions entering into the interaction were detailed. But Glaser and Strauss went further in their sociological treatment of interaction. At a theoretical level, they were preoccupied with changes that may occur during the trajectory of interaction, not only with interactional stability. So they also focused on "the tendency for interaction to move out of regulated social bounds and into new interactional modes." Their approach differed from that of Goffman, who examined interactions in terms of the social relations that subtly govern them and which, from his point of view, are therefore a direct incarnation of social order. In contrast, for Glaser and Strauss it is the indeterminate and problematic nature of interactions that holds attention.

The second set of concepts is linked to temporality. Dying takes time. This observation, less banal than it might appear, opens the way to a theoretical construct, which, associated with the concept of trajectory, continued to inform Strauss' work. Conceiving of death as a process allowed the authors to see it as a series of transitional statuses in a passage from life to death. By questioning the facts and comparing them with other situations of individual status passage, the authors defined the trajectories of death as a status passage and retraced its major dimensions. This immediately forces the researcher to place the discussion within the framework of a social structure. The study of trajectories of death in hospital organizations allowed Glaser and Strauss to advance the idea of a structure in movement (here the hospital), a structural process. Strauss later drew on this notion of flexible structuring to develop a conceptual network organized around one central point: the heterogeneous and complicated work necessary for the management of chronic illness on a daily basis.

Chronic Illness, Management Work, and Trajectories of Illness

Starting as a battle cry ("Chronic Illness", 1973) to draw attention to problems raised for the health system by the

growing prevalence of chronic illnesses and the technologies developed to deal with them, Strauss and his collaborators opened up a major new field for medical sociology and anthropology. Across multiple writings (including four books) around the concept of trajectory, they developed a dense theoretical framework for future research on medical work, patient work, and the experience of illness. Elaborated in two stages, this framework has at its core the notion of work defined as action.

Trajectories of Illness and Medical Work

The concept of trajectory, this time associated with chronic illness, refers "not only to the physiological unfolding of a patient's disease but to the total organization of work done over that course, plus the impact on those involved with that work and its organization. For different illnesses, the trajectory will involve different medical and nursing actions, different kinds of skills and other resources, a different parcelling out of tasks among the workers (including perhaps, kin and the patient), and involving quite different relationships – instrumental and expressive both – among the workers" (Strauss et al., 1985). This definition points to a transformation of the concept itself, moving from the substantive field of illness to the more general one of work. Work itself, a form of action, would become the center of the web of concepts organized around the concept of trajectory. From a sociological point of view, in Strauss' usual fashion, this concept is eventually understood to transcend the substantive setting and become a generic process and part of his version of 'formal theory.'

Thus, one of the first merits of the concept of trajectory is to remove the study of chronic illnesses from biological, medical, and psychological frames of thought and instead equip the researcher with a reflective stance based on analyzing the work involved – clearly and distinctively social phenomena. The initial framework of this formulation is a research monograph analyzing the combined impacts of chronic illness and new medical technologies on medical work in the hospital, *The Social Organization of Medical Work* (Strauss et al., 1985). The scene is complex: the illnesses are extremely varied (new technologies prolong the trajectories and even create new ones), and the work is carried out in different places in the hospital. This diversity can be conceptualized only by tackling the content of the work per se and its concrete execution. This is the juncture at which work and trajectory meet. Meticulous ethnographic analysis of the different types of work composing medical work is built up around the concept of trajectory to allow an analytic understanding of the organization and execution of the therapeutic work.

The trajectory of the illness can be seen as composed of series of successive, overlapping tasks, which, taken together, make up 'arcs of work.' The interaction of efforts to control the illness and the contingencies (predictable or not) shapes each trajectory. As an heir to the Chicago School, particularly Hughes, Strauss has a different way of looking at work. His point of departure, in fact, is neither the existence of a formal division of labor nor the problem of allocation of tasks. Quite the reverse. His central argument concerns the primacy of the work: The work processes themselves shape the division of work, as well as the experiences and the careers of the individuals doing this work, and influence the very structure of the organizations within which it takes place. Different types of medical work have been identified and analyzed: working with machines, working on clinical safety, the work of comfort, working with feelings, working on errors, working with the body, the work of information, the work of negotiation, and the work of articulation (Strauss et al. 1981, 1982b, 1985; Strauss, 1988, 1993; Corbin and Strauss, 1988). However, the types of work and their combinations vary across the overall trajectory of the illness and its specific arcs of work. Hence, the goal is to try to understand the links between tasks and workers as accountability chains.

Because the same task can be carried out by different actors, with different qualifications, at different moments of the trajectory, one of the essential analytic effects here is to establish a distinction between acts and actors. By basing the analysis not only on occupations and professions or organizational structures, but also on the work itself (its conception, organization, the concrete actions of doing it), concepts such as 'mistakes at work,' 'dirty work' (both Hughes' ideas), group work, and work with customers or clients are seen in a different light. Moreover, through this approach to work, Strauss brought to the fore the concept of invisible work, tasks that are requisite but usually ignored and usually carried out by lower level workers – nurses, psych techs, orderlies, janitors. Seeing all the work and workers gave him the means to make visible in the early 1960s the neglected work of technicians in psychiatric hospitals, an exploration that continued in the work on chronic illness.

This focus on the work itself has another major implication: it allows conceptualization of the patient as an actor in medical work and in the analysis of the different components of patients' work. In fact, patients are involved virtually throughout the work process, but their work is often difficult to identify if they are not systematically conceived as actors. The issue is not of so much knowing the patient's place in the division of work, as it is of understanding how the different types of work that she or he accomplishes fit into the overall work of the group and the relationships among their respective trajectories. In sum, for Strauss, work is synonymous with action. This concept of work is further developed in his final research projects.

The Work of Management: Biography and Trajectory

Triggered in part by his own massive heart attack, Strauss realized that the lion's share of the trajectory of a chronic illness takes place not at the hospital but at home, and the sick person is not the only person affected. Every facet of a person's life and these of his or her family and friends are implicated to differing degrees. The duration of the illness, the person's age, sex, and income bracket interact with the work of managing the new conditions of life brought into existence by the illness. In his last empirical project, in collaboration with Juliet Corbin, Strauss analyzed the dimensions of the work of management carried out every day by couples in which one of the spouses is afflicted with a chronic illness or disability (*Unending Work and Care: Managing Chronic Illness at Home*, 1988). Here he deployed a complex theoretical web linking trajectory and

biography with work conjugated as action starting from the experiences, which people afflicted with a chronic illness go through – especially the sense that their bodies have betrayed them. Leaving behind problems of body image or psychological and psychiatric concepts of stress and coping, he focused instead on biographical work processes – an idea rooted in fieldwork. An illness is always placed in a biographical context: what happened before, what life was like in the past, what hopes and dreams had been shattered or altered. For Strauss and Corbin, a biography is made up of three elements: biographical time, concepts of the self, and the body. For each aspect of himself or herself, a person must accomplish various groups of tasks in relation to biographical management. All these tasks fit into biographical time, since they are part of this individual's past, present, and future.

The accomplishment or execution of these tasks linked to the self requires that the body functions in appropriate ways. Through the expression biographical body conception, Strauss and Corbin sought to conceptualize the interdependence of these three elements: conceptions of the self (identity) are generated directly or indirectly through the body and evolve in line with the unfolding of biographical time.

The structural conditions arising from the illness – the individual biography and everyday life on the one hand, and on the other, more general political, social, and economic conditions – together influence the effectiveness, speed, and even the mood in which the work is done. The structural backdrop against which trajectory management takes place is important because of its implications for this work. The immediate interaction is combined with close and distant structural conditions to create a work context. The context then throws light on the core of the action by showing how interaction becomes the means whereby the work itself is accomplished under different – and specified – conditions. In this sense, structure and interaction are intimately linked.

The Web of Negotiation and Social Worlds

Drawing on his work on professionals and hospitals, Strauss progressively developed theories about social space and social relationships. It is important to stress that these theories simultaneously nourish the work presented so far and are also generated by it, as this is distinctive to Strauss' way of thinking. He continued this vein of work in his book *Negotiations: Varieties, Contexts, Processes and Social Order* in 1979. Here he proposed a "paradigm of negotiations," a grid for analyzing negotiation processes, and applied it to a wide range of case studies. In the development of this paradigm, we find the same theoretical leverage point as in all of Strauss' research. He starts, as in his analysis of work, from the negotiation itself and its secondary processes (e.g., bargaining, obtaining gratification, paying debts, and negotiating agreements), interactions, types of actors, their strategies and tactics, and proceeds to examine the consequences. He then builds a context of action on two levels: on the one hand, the structural context within which the negotiations take place and, on the other hand, the context of negotiation that refers specifically to structural properties acting directly as conditions on the flow of negotiation. Strauss uses this context of negotiation as he used contexts of awareness: "Just as there are many types of awareness context pertaining to interaction with respect to dying persons, so also there are many specific types of negotiation context pertaining to interaction among negotiating parties." These different types can be reconstituted in terms of the combinations of the general properties of any negotiation context.

The importance given to negotiation in social life is not surprising in a sociologist who always sought to empirically address one of the fundamental tensions of sociology: that between constraint and freedom, and to do so with attention to range of variation. Strauss' analysis of society as composed of social worlds and arenas is closely associated with this preoccupation. Moreover, this demonstrates his desire to uphold the central position of Thomas and Park on social change – that it is produced by the shock and encounters of groups, by the reciprocal effects and reactions of groups of all kinds encountering each other and their attempts to deal with such encounters (see also Blumer, 1958). Strauss formed a loose map of the territory to provide a perspective in terms of social worlds: "We are confronting a universe marked by tremendous fluidity; it won't and can't stand still. It is a universe where fragmentation, splintering, and disappearance are the mirror images of appearance, emergence and coalescence."

The strength of this approach comes not from defining in advance either the actors in social worlds or their conventions but, on the contrary, from the attempt to empirically identify and represent them. Thus the approach opens up questions of boundaries of the social worlds and the 'authenticity' of what is produced (whether it is a work of art or a medical treatment). This perspective is deeply attuned with analysis of a 'project' or a trajectory, as described above. In two successive articles, Strauss (1982b, 1984) frames various processes of segmentation of social worlds, their interaction through processes of intersection, and processes of legitimation or recognition as developed in this kind of 'social matrix.'

The development of ideas initiated decades earlier culminates in Strauss' capstone book, *Continual Permutation of Action* (1993) in which he offered his own interactionist theory of action. Reviewing the trajectory of his life work through the lens of action, he made clear that action for him is synonymous with acting and he redefined this perspective of negotiated order using a verb rather than a name "processual ordering," stating that "this concept of processual ordering of reactivity is meant to embrace every main topic discussed in this book and of course many more." In what can be described as a theoretically grounded legacy, he gently and firmly gives us important means to participate as sociologists and human beings in this ever-changing world with our own sociological awareness.

Grounded Theory and Thinking through Fieldwork

Strauss is perhaps internationally known best for his work in qualitative research methods. In the postwar era, quantitative procedures were the most commonly used methodological tools in the 'social world' of American sociology, though it was possible to see the beginning of a renewed interest in fieldwork. Glaser and Strauss (1967) gave this movement significant

momentum with their now famous book *The Discovery of Grounded Theory*, which sought to legitimize qualitative sociological research. The book was and remains a benchmark for qualitative sociology. Over and above the deeper investigation of methodological issues, in *The Discovery of Grounded Theory* one finds a concept of sociology as a discipline centered on and through prolonged fieldwork. Qualitative analysis can generate the only kind of theoretical statements they saw as legitimate: 'grounded theoretical statements' derived from painstaking field studies. Theoretical generalization remains possible, but only when based in empirically rooted statements.

The singularity of this approach resides in the deliberately reflexive style of the analysis and the rhythm at which the research proceeds, along with the reasoned (if highly unusual) overlapping of its different stages. From its first formulation, two aspects of grounded theory make it unique: (1) analysis must begin as soon as there are data and (2) variations in the data must be looked for through 'theoretical sampling,' that is sampling observation sites or persons to find new data on emergent ideas and concepts. The goal is to force the researcher in an organized way to generate ideas, thinking through fieldwork. In *Qualitative Analysis for Social Scientists* (1978), Strauss presented a loose 'coding paradigm' around concrete examples taken from his own research and that of others in 'working groups' or analysis sessions that he led weekly. This book is itself a living ethnographic testimony to how Strauss taught and worked. Procedures for analysis have been developed in a series of basics books (1990, 1998, 2008).

Strauss' Influence on Social Sciences and Beyond

With more than thirty books, many translated into other languages, and over hundred articles and chapters, the impact of Anselm Strauss is impressive and lasting. Moreover, Strauss crossed boundaries in at least two ways: across academic disciplines (sociology; social psychology; anthropology; linguistics; nursing; education; social work; communication; cultural studies; etc.), and specialty areas (work, professions and occupation; heath and illness; urban sociology; organization studies; etc.). He deliberately wrote for multiple audiences.

Not only was Strauss one of the founders of the specialty of medical sociology, but he was ahead of his time in two major areas, contemporary studies on death and dying (in sociology, in nursing, medical, social work, physician professional spheres, and bioethics) that still draw on Glaser and Strauss' conceptual framework 50 years later. Second, Strauss and his coauthors work on the sociological study of chronic illness laid the ground work for all future studies on 'the experience of illness,' influencing research directions in sociology and anthropology, and beyond. Strauss pioneered a radical understanding of patients and kin as active people, not passive. His conceptualization of the work of patients and their families was a profound breakthrough that not only changed sociological research but also influenced the political economy of health. Moreover, it enlarged political debate about physician–patient relationships and legitimated patient empowerment in the vein of the Chicago tradition. This exemplified the professional and personal commitments of Strauss toward improved and more democratic participation of lay and sick persons in the health system.

Strauss' perspectives on identity, social worlds and arenas, and negotiated ordering have penetrated many specialties including social studies of science and technology through some of his students' works (Star, 1989; Bowker and Star, 1999; Clarke, 1998; Clarke and Star, 2007; Fujimura, 1996.) Grounded theory is today the most cited qualitative approach globally (Clarke and Charmaz, 2014). While sometimes used superficially, it is also a living tradition, and groups of researchers continue to embrace its reflexive position and work at deepening and also expanding it (Charmaz, 2014). Adele Clarke (2005), a former student of Strauss, created an extension, which integrates aspects of postmodernism, supplementing basic grounded theory with a situation-centered approach.

The depth and diversity of Strauss' intellectual contributions will continue to have long-term impacts on sociology and beyond, highly visible and less visible.

See also: Death and Dying, Sociology of; Dewey, John (1859–1952); Disability: Sociological Aspects; Grounded Theory: Methodology and Theory Construction; Interactionism, Symbolic; Medical Sociology; Qualitative Methods, History of; Social Psychology.

Bibliography

Blumer, H., 1958. Race prejudice as a sense of group position. Pacific Sociological Review 1, 3–8.

Bowker, G., Star, S.L., 1999. Sorting Things Out: Classification and Its Consequences. MIT Press, Cambridge, MA.

Clarke, A.E., 2005. Situational Analysis. Grounded Theory after the Post Modern Turn. Sage publication, London.

Clarke, A.E., 1998. Disciplining Reproduction: Modernity, American Life Sciences and the 'Problem of Sex. University of California Press, Berkeley.

Clarke, A.E. Everyone Was Negotiating about Something: Forty Years of Straussian Negotiated Order Research. (To be submitted to Studies of Symbolic interaction), in preparation.

Clarke, A.,E., Charmaz, K. (Eds.), 2014. Grounded Theory and Situational Analysis. Sage Benchmarks in Social Research Series, 4 vols. Sage, London.

Clarke, A.E., Star, S.L., 2007. Social worlds/arenas as a theory-methods package. In: Hackett, E., Amsterdamska, O., Lynch, M., Wacjman, J. (Eds.), Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge, MA, pp. 113–137.

Clarke, A.E., Star, S.L., 2008. The social worlds framework: a theory-methods package. In: Hackett, E., Amsterdamska, O., Lynch, M., Wacjman, J. (Eds.), Handbook of Science and Technology Studies. MIT Press, Cambridge, MA, pp. 113–137.

Charmaz, K., 2006/2014. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis, first & second eds. Sage, London.

Corbin, J., Strauss, A.L., 1988. Unending Work and Care: Managing Chronic Illness at Home. Jossey-Bass, San Francisco.

Fujimura, J., 1996. Crafting Science: A Socio-history of the Quest for the Genetics of Cancer. Harvard University Press, Cambridge.

Glaser, B.G., Strauss, A.L., 1965. Awareness of Dying. Aldine, Chicago.

Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory. Aldine, Chicago.

Glaser, B.G., Strauss, A.L., 1968. Time for Dying. Aldine, Chicago.

Maines, D.R. (Ed.), 1991. Social Organization and Social Process: Essays in Honor of Anselm Strauss. Aldine, New York.

Star, S.L., 1989. Regions of the Mind: Brain Research and the Quest for Scientific Certainty. Stanford University Press, Stanford.

Strauss, A.L., 1959/1997. Mirrors and Masks: The Search for Identity. Free Press, Glencoe, IL (Reprinted with a new introduction by Transaction Pubs., New Brunswick, NJ.).
Strauss, A.L., 1978a. A social worlds perspective. Studies in Symbolic Interaction 1, 119–128.
Strauss, A.L., 1978b. Negotiations: Varieties, Contexts, Processes and Social Order. Jossey Bass, San Francisco.
Strauss, A.L., Fagerhaugh, S., Suczek, B., Wiener, C., 1981. Patients' work in the technologized hospital. Nursing Outlook 26, 404–412.
Strauss, A.L., 1982a. Interorganizational negotiation. Urban Life 11 (3), 350–367.
Strauss, A.L., 1982b. Social worlds and legitimation processes. In: Denzin, N. (Ed.), Studies in Symbolic Interaction, vol. 4. JAI Press, Greenwich, CT, pp. 171–190.
Strauss, A.L., Fagerhaugh, S., Suczek, B., Wiener, C., 1982. The work of hospitalized patients. Social Science and Medicine 16 (9), 977–986.
Strauss, A.L., 1984. Social worlds and their segmentation processes. In: Denzin, N. (Ed.), Studies in Symbolic Interaction, vol. 5. JAI Press, Greenwich, CT, pp. 123–139.
Strauss, A.L., 1987. Qualitative Analaysis for Social Scientists. Cambridge University Press, New York.
Strauss, A.L., 1988. The articulation of project work: an organizational perspective. Sociological Quarterly 29, 163–178.
Strauss, A.L., 1991. Creating Sociological Awareness: Collective Images and Symbolic Representation. Transaction Pubs., New Brunswick, NJ.
Strauss, A.L., 1993. Continual Permutations of Action. Aldine, New York.
Strauss, A.L., Fagerhaugh, S., Suczek, B., Wiener, C., 1985. The Social Organization of Medical Work. University of Chicago Press, Chicago.
Strauss, A.L., Corbin, J., 1990/1998. The Basics of Qualitative Analysis: Grounded Theory Procedures and Techniques. Sage, Thousand Oaks, CA.

Relevant Website

http://sbs.ucsf.edu/medsoc/anselmstrauss.

Street Children: Cultural Concerns

Elizabeth A Swart, University of Southern California, Los Angeles, CA, USA
James D Wright, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Panter-Brick, volume 22, pp. 15154–15157, © 2001, Elsevier Ltd.

Abstract

This article reviews contemporary literature on the so-called street children. There has been a paradigm shift since the late 1970s in the way social scientists understand and seek solutions to the global problem of street children that abandons misleading categorizations of street children for highly contextualized studies that emphasize the heterogeneity, agency, and gender of the population. This article seeks to understand culturally constructed intersections and the associated material conditions within which children negotiate their own identities. To this end, participatory methods and intervention models help situate the discourse in historical and ethnographic context.

Introduction

Since the late 1970s, there has been a paradigm shift in the way social scientists conceptualize 'street children.' Although governments and the media continue to refer to street children as an undifferentiated group or to categorize them in misleading ways, e.g., by differentiating between children 'in the street' and 'of the street,' contemporary research emphasizes the heterogeneity, agency, and gender of street children and this has given rise to unique research questions and participatory strategies for aid and intervention.

Definitions

Defining 'street children' has itself been problematic. The term 'street children' is still used by many as though there are two (but only two) categories of children – 'street children' who are without homes and families and 'normal children' who are not – thereby creating a false dichotomy that belies the heterogeneity of the population of street children, ignores issues of social construction and cultural context, and overlooks the many lines of differentiation that exist among street children.

In 1979, the United Nations 'Year of the Child,' the term 'street children' was popularized to designate the growing population of young people who spent some or all of their time in urban public spaces. Although the designation was meant to neutralize negative terms then in common usage – street urchins, vagrants, street Arabs, ragpickers, and the like (Panter-Brick and Smith, 2000) – it did not solve the definitional problem. In fact, it was only the beginning of an ongoing debate about how to define, delimit, and desensitize the discussion about this controversial population, the conception of which is closely related to cultural understandings of urbanization, homelessness, and structural issues related to poverty.

In 1983, the Inter-NGO Program for Street Children and Street Youth came up with a new slant on the definition, dubbing street children as "those for whom the street (in the widest sense of the word, that is, unoccupied buildings, wasteland, etc…) more than their family has become their real home, a situation in which there is no protection, supervision, or direction from responsible adults" (Ennew, 1994: p. 15). The definition led to calls for clarification of terms like 'protection,' 'home,' and 'family' (Hecht, 1998) because such terms have a great range of meaning, both historically and culturally. Critics pointed out, for example, that the term 'home' may refer to a situation of affiliation with other people or to transience, instead of access to a physical abode, as the notion of 'home' is usually understood in Western scholarship (Desjarlais, 1996).

Responses attempted to recognize various *categories* of street children and to clarify the heterogeneous nature of the population. These, too, proved confusing. The United Nations International Children's Emergency Fund's (UNICEF's) notion of categorizing children as 'of the street' or 'on the street' presented a seeming solution. According to this definition, children *of the street* made the streets their home in a consistent and constant manner. Children *on the street* (or *in the street* in an alternative but equivalent formulation) were those who had families to return to but often worked on the streets or chose to be there for periods of time before returning home. Other attempts were made to categorize street children by their 'degree of family involvement' or their 'psychological characteristics' (Panter-Brick and Smith, 2000). But such attempts at categorization raised semantic and definitional issues because they were neither analytically precise nor culturally grounded in dealing with the fluid and ephemeral nature of street child populations (Panter-Brick and Smith, 2000). By the beginning of the twenty-first century, the discourse had turned from *categorization* of street children themselves to the *contextualization* 'of their position in particular cultural situations' and how the concept of street children is structurally related to regional issues of homelessness and poverty (Panter-Brick and Smith, 2000: p. i). For example, research began to recognize that "abandonment in its most recent manifestations may include a wider range of childhood experiences than explored" previously and that the discussion is embedded in a historical context "when changes in family form, attitudes to sexual identity and reproductive technology" united to put both concepts of child and home, and even parent, into question (Panter-Brick and Smith, 2000: p. xv). As society began to consider children more subjects of human rights than objects of concern (Ennew, 1994), the

implications for research, policy, and intervention also shifted enormously, reflecting less a welfare-based understanding of street children than a growing concern that research correctly portray these children's views of themselves and their relationship to society. Although the term 'street children' is still widely used by governments, nongovernmental organizations (NGOs), and helping agencies to refer to youth who occupy urban public spaces, many categories such as 'of the street' and 'on the street' have lost credibility. Social scientists have moved the debate beyond overly facile noncontextualized categorization and into issues of diversity and intersectionality. The latter refers to systematic study of the interactions of multiple systems of oppression, e.g., race, class, gender, etc.

Twenty-first century research on street children has become increasingly focused on (1) the concept of street children as a socially constructed one, (2) the cultural understanding of poverty and homelessness, (3) the need for longitudinal studies that describe lifetime career outcomes of street children's lives, (4) the development of participatory research methods, and (5) the development of effective research strategies that foster children's human rights (Panter-Brick, 2002). With these changes came a new interest in the intersections of age, gender, class, ethnicity, sexuality, and ability as these variables relate to street children, their coping strategies, and life-course paths (Beazley, 2002; Evans, 2006).

History

Abandoned and homeless children have doubtlessly been present since the beginning of civilization, appearing sporadically and tangentially in Western histories since Roman times. In fact, Western historians note that street children made up the largest population of boy prostitutes in Augustan Rome (Boswell, 1988). Similarly, in CE 422 – anthropologists use CE (Current Era) – a church council expressed concern that street children were treated with ruthlessness and violence (Boswell, 1988). Street children were common in Russia, both under the Tsarist rule and after the Russian Revolution.

In both Britain and the United States, the industrial revolution affected families in a variety of ways, contributing to rural–urban migration, child labor, family disruption, caregiver death, and the phenomenon of working, abandoned, or homeless children. In 1848, Lord Ashley remarked that there were at least 30 000 "naked, filthy, roaming, lawless, and deserted children" in metropolitan London (Hutson and Liddiard, 1994). Near the turn of the twentieth century, American photographer and social historian Jacob Riis captured images of huddled children sleeping on streets adjacent to tenement houses in New York in his shocking *How the Other Half Lives* (Riis, 1890). World War I further increased the number of street children, especially in Russia where, by 1922, there were an estimated 7 million homeless children (Ball, 1996).

After World War II, the presence of homeless children on the streets of Western Europe prompted United Nations Educational, Scientific and Cultural Organization in 1951 to utilize the term 'street children' to define this growing social problem. In the 1970s, the United Nations Year of the Child (1979) brought attention to the growing street child population and popularized it as a social cause, thereby rendering the term 'street children' as a call to action and intervention. By 1986, UNICEF was raising alarms about the growing population of homeless youth in the *developing* world, citing skyrocketing numbers in India, Brazil, and Thailand. Today, street children, particularly those in the developing world, are often residents of huge slums, adjacent to large metropolitan areas. In these 'megaslums,' street youth are especially vulnerable to prostitution and survival sex, addiction (to glue or toluene, changa'a, and petrol), sex trafficking, and human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) infection (Wittig et al., 1997; Davis, 2006; Swart, 2009).

Throughout the twentieth century, the social construction and understanding of 'childhood' have played into our representation of street children. The Western discourse on childhood posits that the street is not an appropriate place for a child and the child who is on the street is 'out of place,' socially deviant, or in dire need of intervention (Boyden, 1990; Hecht, 1998). Ennew has pointed out, however, that such a dichotomous understanding of the population – as either victim or villain – is unhelpful because it "tends to ignore social networks and coping strategies developed by the children" (Ennew, 1994; Panter-Brick and Smith, 2000). Social scientists today are engaging in child-centered research perspectives that are likely to yield more accurate data about populations of street children, as well as more meaningful interventions.

Statistics

In the late 1980s, UNICEF estimated there were about 100 million 'street children' around the world. In 2005, it admitted what a moment's reflection makes obvious, namely, that "the exact number of street children is impossible to quantify." UNICEF, as of 2012, will say only that there are "tens of millions" of street children around the world and that the numbers are increasing (UNICEF, 2012). The 100 million estimate is still commonly cited by researchers and policy makers because attempts to more accurately estimate the population of street children are frustrated by the fluidity of the population and differing agency and governmental agendas. When NGOs issue divergent estimates, it is sometimes because, according to Panter-Brick, they were not talking about the same children. That is, the definition of street children is embedded in the social construction, maintenance, and management of homelessness as a social problem (Ennew, 1994; Panter-Brick and Smith, 2000) and with how it is defined and quantified in various contexts.

Gaps in the Research

Judith Ennew was perhaps the first to insist that most national and regional estimates of street children "had no basis in fact" (Ennew, 1994: p. 32) because they were "rarely referenced to counting methods" (Thomas de Benitez, 2011). Although Ennew asserted in 1994 and in 2000 that "the arithmetic" used to arrive at estimated street child numbers "is as symbolic as the children themselves" (Ennew, 2000: p. 170), the 100 million number was still commonly referenced in articles and academic journals and remained in common usage for more than two

decades. Experts today are more cautious in their 'guesstimates,' using the tens-of-millions estimate as a go-to figure that is neither overblown nor understated. Nevertheless, more precise numbers – and the methodologies that produce them – are emerging.

Improving Localized Counting

Throughout the 1990s, collaborations between UNICEF and governments at the national or city level yielded rapid situation assessments that were, unfortunately, not tied to uniform methodologies or definitions. The lack of definitional and methodological uniformity led to an outpouring of 'head counts' that, when considered together, were either misleading or nonsensical. In 1989, for example, UNICEF estimated that 100 million children were on the streets of the world's major cities, while two other estimates claimed (for Brazil alone) 20 million (Connelly, 1990) and 50 million (De Moura, 2005), implying that nearly half the world's street children are Brazilian, which one researcher remarked seemed 'rather unlikely' (Marrengula, 2010).

Since the beginning of the twenty-first century, though, efforts to count street children have "become more careful in their research methods and more circumspect" in their estimates (Thomas de Benitez, 2011: p. 5). Efforts today are yielding improved localized counting, made possible in part by a new definitional understanding of the 'street connectedness' of children (Thomas de Benitez, 2011: p. 5). Defining children as 'street connected' permits tossing the dichotomous 'on the street' or 'of the street' thinking into the rubbish bin, along with the temptation to *categorize* children in misleading and unhelpful ways. Instead, using a definition that emphasizes 'street connectedness' – that is, defining street children as those who depend on the street as a *central reference point* in their lives – adds both a broader cultural base to the definitional debate, as well as a sharper focus on individualized context. The concept of 'street connectedness' sheds light on the 'multiple deprivations' of street children within a global context, while it also situates street children 'within the wider policy agendas and interventions' of particular cities and countries (Thomas de Benitez, 2011).

Definitional improvements are accompanied by innovative and extremely localized data collection, using participatory research to help determine local numbers. For example, Ensing and Strehl's (2010) study of street-working and street-living children in Peru used local knowledge and local investigators to determine numbers of street children in Cusco and Lima. Their methodology consisted of a 35-question survey administered by six enquirers to 343 street children, who were recruited through painstaking and personalized knowledge of the children themselves and their locale:

> The enquirers responsible for approaching and interviewing the children in Cusco were all Peruvians living in Cusco. They were male and female, students and graduates, between 22 and 34 years old. Most of them had experience in working with street children or working children. The enquirers worked in couples and were scheduled to work in one specific area only. This way, they could become familiar with the neighborhood and the children on the streets. (Ensing and Strehl, 2010: p. 3)

Similarly, Marrengula (2010) used localized, participatory research to determine numbers of street children in Maputo, Mozambique. After allying themselves with a local NGO known to and trusted by street children, Marrengula's team constructed a methodology that they hoped would provide the most accurate head count possible for the city. It consisted of visiting street children's 'zones' at night, something local helping professionals, being used to providing services to children during the daytime, had rarely attempted. Marrengula describes precise, localized methods for determining numbers:

> The research team…came to a consensus that it was very worthwhile to work during the night, as most of the street children were working during the day, and we would not have any information about their real location during the day and this idea would give us a close to real picture of the social landscape of the street children in terms of their numbers and group members…. (Marrengula, 2010: p. 139)

The head count was accomplished "first by making several rounds of the city area, then making several visits to different places where children had been living and asking those children about the location of other places" street children lived (Marrengula, 2010: p. 143). Researchers walked long distances around the city during late evening hours, working from 7 to 10 p.m. 3 days a week for 5 weeks to gather information about children living in each *zone*. The data collection team asked authorization to visit the children at night. Once that was given, they began discussions with children about their lives, feelings, and problems. With the help of children, other zones were determined where the same methodology was put into place:

> We had to visit each zone two or three times in order to make sure the information gathered on previous visits matched, because it happens that those children moved constantly or that they provided false information about the group. During these visits, we attempted to gather information that could give us detail in the following aspects: number of children in the zone; the name used to identify the zone; the names of the street children; and, if possible, their ages, their interests, their difficulties and challenges. (Marrengula, 2010: p. 143)

During the course of the study, Marrengula's team identified "five communities of street children with a total of 42 zones where 445 street children were living, of which 46 were girls and 5 were babies under 2 years old" (Marrengula, 2010: p. 145). This number and its demographic diversity had heretofore *not* been known (or had been underestimated) by local NGOs, even those with a close relationship to the city's children.

As of this writing (in 2013), research aspires to definitional *and* methodological uniformity and transparency. Some progress is being made. Nevertheless, major gaps still exist in research methodologies and data collection. The Consortium for Street Children (Thomas de Benitez, 2011) has pointed out several of the most egregious gaps in research about street children. For example, academic research about street children is 'fragmented' (Thomas de Benitez, 2011: p. ix), being

conducted in a variety of disciplines and departments without consultation or collaboration. Thus, much of the data remains in "small islands of research within separate disciplines" (Thomas de Benitez, 2011: p. ix), which are not informed by one another and are often used at cross purposes. Similarly, academic and *development* studies are also conducted without reference to one another, meaning that academic knowledge "rarely informs development practice and vice versa" (Thomas de Benitez, 2011: p. ix). Finally, experts contend that "no motor exists for the building of a body of research on street children" and that a concerted effort is needed to coordinate interdisciplinary research and to make data available in a proactive way to relevant audiences.

Agency

The issue of children's agency has become an increasingly important concept since the beginning of the twenty-first century, with participatory research being used to improve numbers' estimates and to enhance our contextual understanding of street children's lives. Panter-Brick has pointed out that in the late 1990s, fundamental assumptions about children and appropriate methods of research began to be reevaluated in the context of international laws about children's rights (2002). A growing awareness developed in the social sciences that children were not helpless victims but capable agents, able to make informed choices about their lives and future aspirations. Researchers attempting to view street children's lives from the children's own perspectives have recognized children's agency, competence, and participation in their own life choices. A partnership methodology has emerged, in which social scientists design projects that work 'with' children not 'for' them and view that children are "not objects of concern but people … who need respect, not pity" (Ennew, 1994: p. 35). Such participatory research has given rise to reflexive methodologies that are perhaps more effective than the commonly used questionnaire (Panter-Brick, 2002), such as participant observation, group workshops and discussions, in-depth interviews, and innovative techniques, such as encouraging children to draw, act out, and 'map' their own life experiences (Young and Barrett, 2001; Beazley, 2002; Evans, 2006). Although these methods have drawbacks, many social scientists have found that they yield data that more accurately reflect the experience of street children as seen through the children's own eyes (Beazley, 2002; Evans, 2006). These data are, therefore, more potentially helpful in understanding the fluidity of the population and devising culturally specific interventions.

Gender

The dramatic rise in the number of street girls, particularly since the mid-1990s, has resulted in a more *gendered* discourse about street children. Several important issues arise in the discourse on street girls: they are doubly marginalized in terms of social attitude and the dominant discourse on gender and they are more vulnerable to discrimination and to sexual exploitation on the street; but, nevertheless, many street girls subvert dominant paradigmatic gender constructions to create their own gendered spaces on the street. For example, street girls may select and maintain boundaries around their own social spaces where they gather with one another, excluding male street youth. Beazley has observed that, in their own gendered gathering spaces, street girls in Indonesia are "more vocal, gregarious, and confident" (Beazley, 2002: p. 1675). Such gendered spaces insulate and protect street girls from discrimination at the hands of both the dominant culture and male street youth.

Research on street girls has revealed that girls' double-marginalization is defined in terms of social attitudes toward deviance. Girls suffer discrimination not only for being *children* on the street but also for being *females* in a place not traditionally defined as acceptable for them. Lucchini has described the predicament of street girls in Argentina by saying

> The street world is before all masculine. The presence of girls in the street upsets the values and representations of adults. Uneasiness takes place and the explanation of this presence goes through a pathologizing of the girl and her family. The marginal girls and the delinquent girl are soon associated to sexuality. (Lucchini, 1994: p. 8)

In Indonesia, for example, street *children* are referred to as 'tikyan' (a generalized category referring to homeless youth) but street *girls* are called 'rendan,' or 'other' by street boys, who use the designation to stigmatize female street youth (Beazley, 2002).

Furthermore, street girls suffer from higher vulnerability to sexual abuse on the street, both from society in general and from street boys, who find in street girls a group they can dominate and control. Research in Indonesia, for example, reveals that:

> …street girls particularly encounter abusive discrimination on the streets of Indonesia because they are seen to be committing a 'heretical geography' by violating ideas of femininity in state discourse and by 'invading' the street that is fundamentally a male space….Although street boys are themselves powerless children, suppressed by the control of an oppressive regime, street girls are one group to whom them feel they are superior and over whom they have power. (Beazley, 2002: p. 1679)

Girls also suffer from lower income-generating potential on the street, where the division of labor is highly gendered. While street boys engage in odd jobs, scavenge for salable items, and wash car windows (in many urban cities, such as Nairobi and Kampala), street girls are confined to labor that is defined as appropriate for females, e.g., laundry, house-cleaning, food selling, and the selling of sex. Such jobs are less lucrative and, in the case of prostitution and 'survival sex,' more dangerous, since the girl risks pregnancy, as well as sexually transmitted diseases including HIV/AIDS infection (Swart, 2009, 2012).

Nevertheless, many of these girls actively subvert gender norms (Beazley, 2002; Evans, 2006) to construct their own gendered identities and their own gendered sense of place (Beazley, 2002; Evans, 2006). For example, in Indonesia,

"street girls negotiate social and personal spaces that are different from those of boys …. These socio-spatial patterns are explained as 'geographies of resistance,' which are a response to the pervasive patriarchal discourse within Indonesian society, a society which believes that girls, in particular, should not be on the streets" (Beazley, 2002: p. 1665). For example, girls create their own gathering place in a city park, where they gather with other girls, sleep, keep possessions in a small building by the park gate, and generally 'hang out.' By contrast, street boys gather in the city center, near a public toilet. Girls also create their own style and behavioral patterns, including adopting short hair cuts, wearing men's shirts, and adopting behaviors, such as smoking and body piercing, which run counter to accepted gender norms. Street girls also create their own patterns of negotiating social space, which involve not only separate 'hangouts' from those that boys use but also different *patterns* of movement (in smaller groups than boys) and *times* of within-city movement (night instead of day) (Beazley, 2002: p. 1677). Similarly, in Tanzania, Evans has discussed how young people "reconfigure gender norms and the concept of 'family' on the street" (Evans, 2006: p. 109) and how girls "actively subvert gender norms by seeking independent lifestyles in towns and cities" (Evans, 2006: p. 125). While street boys sleep in gutters, bushes, or verandas, for example, girls often team up with other girls or boyfriends to stay in guesthouses or other temporary shelters that afford them protection from the street. Within such settings, Evans' study of girls' photographic accounts of their lives show them playing a variety of gender roles in temporary domestic settings – washing dishes, playing with children, dressing in Western-style clothes, looking 'tough,' and putting on evening makeup. Girls present a variety of *selves* in pseudo-domestic street environments, sometimes even parodying dominant paradigms, such as motherhood, homemaker, or prostitute by overemphasizing the stereotypic characteristics of that role, such as wearing excessive lipstick to parody the prostitute image they deplore (Evans, 2006). For such girls, the street presents an opportunity to "construct their own gendered places, where identities are contested and negotiated, and where stigmatizing labels and restrictive norms can be resisted and reconfigured" (Evans, 2006: p. 125).

Responses and Solutions

The Human Rights Discourse

A strong and pervasive discourse on children's human rights began when the United Nations Convention of the Rights of the Child (1989) articulated a set of acknowledged rights for children worldwide and shifted the discourse from the needs of street children to their *rights* (Panter-Brick, 2002: p. 155). Thus, the dialogue shifted from one characterized by humanitarian pity (children's needs) to one seeking to empower children with their rights as citizens (Ennew, 1994). As the paradigm shifted from understanding children as victims to perceiving them as rights-holding citizens, social scientists began to view children as agents of change with power over their own lives. This new interest in children's agency, participation, and point of view gave rise to new research questions and strategies.

Participatory Methodologies and Partnership Ethics

One response to the human rights discourse has been to find alternatives to unhelpful and inaccurate categorizations of street children by focusing on the *diversity* of experience within the street children population. To that end, research about street children has become vividly descriptive and more broadly focused on intersectionality. Current research is increasingly focused on street children as a highly differentiated group and on the ways that "their differences affect their experiences" (Beazley, 2002; Ennew and Swart-Kruger, 2003; Ruvero and Bourdillon, 2003; Evans, 2006). The understanding of this intersectionality of age, gender, ethnicity, class, and ability enables social scientists to articulate meaningful differences within a heretofore (seemingly) monolithic population.

There has also been a marked increase in participatory research methods and child-centered research. For example, one study on the negotiation of social identity among street children in Tanzania used "participatory methods, such as focus groups, drawings and photographic exercises" (Evans, 2006) to engage young people in the research. Others have used such child-centered methods to give children more control over the representation of their lives (Young and Barrett, 2001).

Research on street children has also articulated a partnership ethic in relation to the population being studied. Instead of doing research 'on' street children, social scientists conduct research 'with' street children, using such methods not only to engage children in the research but also to enable them to participate in describing their own lives (Evans, 2006; Beazley, 2002).

Interventions

Social scientists in the twenty-first century seek responses and solutions that are contextual and tailored to individual subpopulations of children. Recognizing both the heterogeneity of street children around the world, as well as the necessity of adapting interventions to unique populations, social scientists are creating unique intervention models for specific populations, such as refugee children (Hinton, 2000), slum children (Montgomery, 2000; Swart, 2012), children displaced by war (Charnley, 2000), and many more. For more examples of contextualized intervention models and strategies, see Panter-Brick and Smith (2000) and Subbarao and Coury (2004).

See also: Child Protection: International Issues; Children, Rights of: Cultural Concerns; Disposable Youth in America in the Age of Neoliberalism; Homelessness in the United States; Human Rights, Anthropology of; Participatory Action Research in Social Research; Youth Culture, Anthropology of.

Bibliography

Ball, A.M., 1996. And Now My Soul is Hardened: Abandoned Children in Soviet Russia 1918–1930. University of California Press, Berkeley.

Boyden, J., 1990. Childhood and the policy makers: a comparative perspective on the globalization of childhood. In: James, A., Prout, A. (Eds.), Constructing and Reconstructing Childhood: Contemporary Issues in the Sociological Study of Childhood. Falmer Press, London.

Beazley, H., 2002. 'Vagrants wearing make-up': negotiating spaces on the streets of Yogyakarta, Indonesia. Urban Studies 39 (9), 1665–1683.

Boswell, J., 1988. The Kindness of Strangers: Abandonment of Children in Western Europe from Late Antiquity to the Renaissance. Harpers, New York.

Charnley, H., 2000. Children separated from their families in the Mozambique war. In: Panter-Brick, C., Smith, M.T. (Eds.), Abandoned Children. Cambridge University Press, Cambridge.

Connelly, M., 1990. Adrift in the city: a comparative study of street children in Bogota, Colombia, and Guatemala. In: Boxill, N.A. (Ed.), Homeless Children: The Watchers and Waiters. Haworth Press, New York.

Crenshaw, K., July 1991. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Review 43, 1241–1299.

Davis, M., 2006. Planet of Slums. Verso Books, New York.

De Moura, S.L., 2005. The prevention of street life among young people in Sao Paulo, Brazil. International Social Work 48 (2), 193–200.

Desjarlais, R., 1996. Some causes and cultures of homelessness. American Anthropologist 98 (2), 420–425.

Ennew, J., 1994. Street and Working Children: A Guide to Planning. Save the Children, London.

Ennew, J., 2000. Why the Convention is not about street children. In: Fottrell, D. (Ed.), Revisiting Children's Rights: 10 Years of the UN Convention on the Rights of the Child. Kluwer Law International, The Hague.

Ennew, J., Swart-Kruger, J., 2003. Homes, places and spaces in the construction of street children and street youth. Children, Youth and Environments 13 (1). http://cye.colorado.edu.

Ensing, A., Strehl, T., 2010. Street-Working and Street-Living Children in Peru: Quantitative Report Cusco. IREWOC.

Evans, R., 2006. Negotiating social identities: the influence of gender, age and ethnicity on young people's 'street careers' in Tanzania. Children's Geographies 4 (1), 109–128.

Hecht, T., 1998. At Home in the Street: The Street Children of Northeastern Brazil. Cambridge University Press, Cambridge.

Hinton, R., 2000. Seen but not heard: refugee children and models for intervention. In: Panter-Brick, C., Smith, M.T. (Eds.), Abandoned Children. Cambridge University Press, Cambridge.

Hutson, S., Liddiard, M., 1994. Youth Homelessness: The Construction of a Social Issue. MacMillan, London.

Lucchini, R., 1994. The Street Girl: Prostitution, Family, and Drugs. University of Fribourg Press, Fribourg, Switzerland.

Marrengula, M., 2010. Addressing Socio-Cultural Animation as Community Based Social Work with Street Children in Maputo, Mozambique (Ph.D. dissertation). Univesity of Tampere, Department of Social Work Research, Finland.

Montgomery, H., 2000. Abandonment and child prostitution in a Thai slum community. In: Panter-Brick, C., Smith, M.T. (Eds.), Abandoned Children. Cambridge University Press, Cambridge.

Panter-Brick, C., Smith, M.T., 2000. Abandoned Children. Cambridge University Press, Cambridge.

Panter-Brick, C., 2002. Street children, human rights, and public health: a critique and future directions. Annual Review of Anthropology 31, 147–171.

Panter-Brick, C., 2004. Homelessness, poverty and risks to health: beyond at risk categorizations of street children. Children's Geographies 2 (1), 83–94.

Riis, J., 1890. How the Other Half Lives. Charles Scribner's Sons, New York.

Ruvero, R., Bourdillon, M., 2003. Girls: the less visible street children of Zimbabwe. Children, Youth and Environments 13 (1), 150–166. http://cye.colorado.edu.

Subbarao, K., Coury, D., 2004. Reaching Out to Africa's Orphans: A Framework for Public Action. World Bank, Washington, DC.

Swart, E.A., 2009. Words are a place to stand: journals by young women in Kibera, Kenya. Affilia: Journal of Women and Social Work 24 (1), 19–30.

Swart, E.A., 2012. Gender-based violence in a Kenyan slum: creating local, woman-centered interventions. Journal of Social Service Research 38 (4), 427–438.

Thomas de Benitez, S., 2011. State of the World's Street Children: Research. Consortium for Street Children, London.

UNICEF, 2005. State of the World's Children. United Nations, New York.

UNICEF, 2012. State of the World's Children: Children in an Urban World. United Nations, New York.

Wittig, M.C., Wright, J.D., Kaminsky, D.C., 1997. Substance use among street children in Honduras. Substance Use and Misuse 32 (7–8), 805–827.

Young, L., Barrett, H., 2001. Adapting visual methods: action research with Kampala street children. Area 33 (2), 141–152.

Stress and Cardiac Response

Jaime Vila, José L Mata, and Pedro Guerra, Facultad de Psicología, University of Granada, Granada, Spain

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by J. Vila, volume 22, pp. 15160–15163, © 2001, Elsevier Ltd.

Abstract

The notion that organisms react physiologically to danger and that this reactivity represents a long-term health risk, is essential to the concept of stress. This article summarizes the literature on cardiac defense. This article begins with a brief introduction to the physiology of cardiac reactivity. Then it focuses on the history of the two major research traditions in the study of cardiac defense: the cognitive and the motivational. It then presents a new approach to the study of cardiac defense based on the analysis of the complex pattern of heart-rate changes elicited by intense auditory stimulation.

The origin of the concept of stress in health-related disciplines is closely linked to the idea that organisms react physiologically to danger or threat. This reactivity has a protective function since it provides the logistic and instrumental basis for adaptive behaviors such as fight or flight. However, if maintained for long periods, this protective or defensive reactivity may become a health risk, compromising the normal functioning of the organs involved. For many scientists, the excessive physiological reactivity associated with stress is the main mechanism linking stress to illness.

Cardiac Response

The heart is a vital organ that reacts rapidly to physical and psychological demands. This reactivity is controlled by the central nervous system through neural and humoral pathways.

Neural Pathways

The neural pathway to the heart involves the two branches of the autonomic nervous system: the sympathetic and the parasympathetic. When the sympathetic nerves are activated, the heart rate accelerates and the heart contraction becomes stronger, resulting in a greater blood volume discharge and greater blood pressure in the vessels. When the parasympathetic nerves are activated, the heart rate decelerates, reducing the blood volume discharge and the blood pressure. The two branches can act reciprocally (one activated and the other inhibited), working in the same direction (increasing or decreasing the heart activity), or nonreciprocally (both activated or inhibited simultaneously), competing against each other for the final cardiac outcome.

Humoral Pathways

The humoral pathway involves the neural activation of the endocrine system, resulting in the secretion by the adrenal glands of various hormones into the bloodstream. One of these hormones, adrenaline, when it reaches the heart, has the same effect as the sympathetic nerves: it increases the heart rate and the stroke volume. Other adrenal hormones, such as cortisol, facilitate the transfer of lipids and glucose into the bloodstream and also have an important indirect effect on the heart and on the arteries that supply blood to the heart muscle: the coronaries. The humoral effects on the heart are slower than the neural effect but they last longer, contributing to maintaining cardiac reactivity during prolonged periods.

Cardiac–Respiratory Interactions

Both the cardiovascular and the respiratory systems are responsible for maintaining the continuous exchange of gases, oxygen, and carbon dioxide between the atmosphere and the tissues. The cardiovascular system reacts to changes in the concentration of these gases in the blood. Therefore, changes in the respiratory pattern, which induce changes in gas concentration, produce changes in the blood flow and cardiac activity in order to restore the appropriate gas concentration. Two examples of cardiac reactivity due to changes in respiration are the hyperventilation and respiratory sinus arrhythmia phenomena.

Hyperventilation

Hyperventilation is caused by rapid respiratory patterns provoking reduced carbon dioxide concentration in the blood. The cardiovascular response is a reduced blood flow to the brain and heart arteries, and an increase in cardiac activity due to parasympathetic inhibition.

Respiratory Sinus Arrhythmia

In contrast, respiratory sinus arrhythmia is caused by slow and relaxed respiratory patterns. The consequence is a rhythmic deceleration of heart rate, due to parasympathetic activation, coinciding with each expiratory cycle.

Stress and Cardiac Defense

Cardiac defense can be defined as the cardiac response to a perceived danger or threat. The concept of defense has a long tradition in physiology and psychology and it is in this context that the concept of stress was first introduced.

Historical Antecedents

At the beginning of the twentieth century, Ivan Pavlov and other Russian reflexologists used the term defense reflex to refer to protective physiological responses elicited by noxious stimulation, such as hand withdrawal from an electric shock, eyeblink to a puff of air, or vomit to bad food. A few years later, Walter Cannon used the term defense to refer to the 'fight or flight' response, a sympathetically mediated cardiovascular response to emergency situations aimed at providing energy supply to the body to facilitate adaptive behaviors such as attack (fight) or escape (flight). By the middle of the century and following Cannon's ideas, Hans Selye had introduced the concept of stress and used the term 'alarm' response to refer to the first stage of the physiological response to stressful situations.

Classical Approaches to Cardiac Defense

There have been two major approaches to the study of cardiac defense: (1) the cognitive or sensory processing approach built on Pavlov's distinction between orienting and defense reflexes and (2) the motivational or emotional approach built on Cannon's ideas on the fight or flight response.

The Cognitive Approach

The cognitive approach assumes that cardiac changes to environmental stimuli reflect attentional and perceptual mechanisms aimed at facilitating or inhibiting the processing of the stimuli (Sokolov, 1963; Graham, 1979). The orienting response (a deceleration of the heart rate to moderate or novel stimulation) facilitates attention and perception of the stimuli, whereas the defense response (an acceleration of the heart rate to intense or aversive stimulation) reduces attention and perception as a form of protection against the aversive stimuli. To study cardiac defense, this approach has developed tasks that present discrete sensory stimuli, such as noises, lights, pictures, or words, manipulating their physical features – intensity, duration, and rise time – or their affective content – pleasantness versus unpleasantness. According to this approach, cardiac defense is characterized by a heart-rate accelerative response, peaking between 3 and 6 s after the onset of stimulus, which habituates slowly with stimulus repetition.

The Motivational Approach

The motivational approach assumes that cardiac changes in response to environmental demands reflect metabolic mechanisms aimed at providing the necessary energy to the body to support behavioral adjustments (Obrist, 1981). If the appropriate behavior is to be quiet and relax, then the cardiac response will be a heart-rate deceleration. If the appropriate behavior is to be active, either physically or psychologically, then the cardiac response will be a heart-rate acceleration. To study cardiac defense, tasks have been developed that involve emotionally or cognitively challenging situations maintained throughout several minutes, such as keeping the hand in cold water, pressing a key as fast as possible to avoid shocks, or doing complex arithmetic tasks. The term 'mental stress' was introduced in this context to refer to tasks that require 'mental effort' instead of 'physical effort.' Under such conditions, cardiac defense is characterized by heart-rate acceleration maintained during the task duration, accompanied by reduced cardiac variability linked to hyperventilation and loss of respiratory sinus arrhythmia.

A New Approach to Cardiac Defense

The cognitive and motivational approaches to cardiac defense have been difficult to reconcile in the past. However, new data have been accumulated on defense responses, both in animals and humans, and on the neurophysiological brain circuits controlling such responses, which not only facilitate integration of the classical approaches but also have potential implications for the concept of stress and its relation to cardiovascular illness.

Natural Defense

In natural setting, such as the imminence of a predator, the defense reaction is not a unique response that could be conceptualized as cognitive versus motivational, or vice versa. Rather, it follows a dynamic sequential process with initial phases in which attentional factors predominate, aimed at the detection and analysis of the potential danger, and if the process continues, later phases in which behavioral factors predominate, aimed at preparation for active defense such as flight or fight. This sequential process (defense cascade) can be divided into three stages: pre-encounter, postencounter, and circa strike. Depending on the spatial and temporal proximity of the predator, different components of the defense reaction may take place successively.

Protective Reflexes

Motor startle and cardiac defense are two protective reflexes that have been extensively studied in recent years in relation to both cognitive and emotional modulatory factors (for an extensive review, see Bradley and Lang, 2007; Vila et al., 2007). Startle is a whole-body motor flexor response, which involves a quick closing of the eyes accompanied by stiffening of the head, dorsal neck, body wall, and limbs, as if to protect from a predator. It is elicited by abrupt stimulation such as a brief, intense noise or light. The response pattern habituates rapidly on repeated stimulus presentations, although some components of the reflex, such as the eyeblink, tend to persist and habituate more slowly. The startle reflex is modulated by emotional and attentional factors. The startle magnitude is augmented when the organism is in an aversive emotional state and inhibited when the organism is in a positive emotional state. Similarly, attentional factors may inhibit the startle reflex if another stimulus is presented in short proximity (prepulse inhibition) or increase it if the organism is in the attentive phase of the defense cascade.

On the other hand, studies on the cardiac components of the defense response to unexpected noises or shocks have questioned the classical models on stimulus–response characteristics, physiological mediation, and psychological significance. In the next sections, we summarize the results of a large

series of studies on cardiac defense using as basic paradigm the presentation of an unexpected intense noise under different stimulus and task conditions.

Characteristics of Cardiac Defense

Descriptive Pattern of the Response

Figure 1 represents the average second-by-second heart-rate response to an unexpected intense noise at 105 dB intensity, 500 ms duration, and instantaneous rise time. The response is represented along the 80-s poststimulus onset in terms of differential scores with respect to a baseline of 15 s prior to the stimulus. The response pattern shows two accelerative and two decelerative components in alternating order: acceleration–deceleration–acceleration–deceleration. The first acceleration/deceleration is faster, reaching its accelerative peak around second 3, whereas the second acceleration/deceleration is slower, reaching its accelerative peak around second 35. Both accelerations have larger amplitudes than the two decelerations. This response pattern can also be represented in a simplified form by selecting 10 points corresponding to the medians of 10 progressively longer intervals: two intervals of 3 s, two of 5 s, three of 7 s, and three of 13 s. This simplified representation, as shown in Figure 2, facilitates statistical analysis without altering the topographic characteristic of the response.

Stimulus Characteristics

The classical model on cardiac defense assumes that any stimulus modality sufficiently intense or aversive should be capable of eliciting the defense response. However, studies comparing four stimulus parameters – sensory modality, intensity, duration, and rise time – have found that (1) not all stimulus modalities elicit response (the auditory and electrocutaneous modalities do elicit the response pattern, but the visual modality does not); (2) stimulus intensity is an important factor increasing the amplitude of the two accelerative components (for example, a white noise at 105 dB), but the response pattern can also be elicited with auditory stimuli of moderate intensity (for example, a white noise at 79 dB); (3) stimulus duration does affect the amplitude of the first and second accelerative components: the first acceleration increases linearly as a function of duration (50–500 ms), whereas the second acceleration/deceleration is elicited only with long stimulus durations (500 ms or longer); and (4) rise time has no effect on the response pattern. The auditory modality has been the most frequently used modality to investigate the two protective reflexes mentioned above (cardiac defense and eyeblink startle). However, there are two stimulus parameters that differentiate cardiac defense and eyeblink startle with regard to their capacity to elicit the response (Ramírez et al., 2005): stimulus duration and rise time. Stimulus duration is critical to elicit cardiac defense (longer than 500 ms), whereas rise time is critical to elicit eyeblink startle (shorter than 24 ms). Thus, on manipulating these two parameters it is possible to evoke cardiac defense and eyeblink startle (intense noises of long duration and short rise time), only cardiac defense (intense noises of long duration and long rise time), or only eyeblink startle (intense noises of short duration and short rise time).

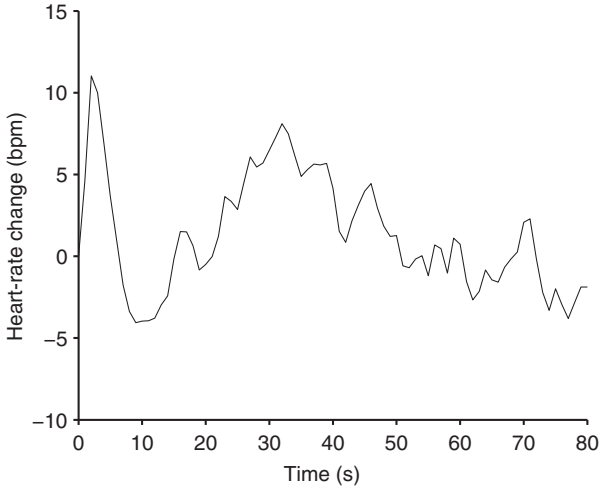

Figure 1 Cardiac defense: second-by-second heart-rate response to an unexpected intense noise. The horizontal axis represents 80 s after noise onset. The vertical axis represents the cardiac changes in beats per minute (bpm) with respect to a baseline of 15 s prior to the onset of noise.

Habituation and Dishabituation

Repetition of the acoustic stimulus results in rapid habituation of the response pattern, the second acceleration/deceleration almost disappearing after the first stimulus presentation. Although habituation is a fundamental issue in the scientific literature on cardiac reflexes (Graham, 1979), the rapid habituation of this response has been inadequately investigated. According to Graham's classical model, cardiac defense should be resistant to habituation. However, the rapid habituation of the response, especially the second acceleration/deceleration, is a consistent finding in studies where the intense noise is repeated several times with intertrial intervals (ITI) between 100 and 120 s. The first acceleration/deceleration also shows habituation, but the habituation trend is less pronounced (Ramírez et al., 2005). Two recent studies have examined the habituation and dishabituation of cardiac defense within a single session manipulating (1) the time interval between stimulus repetition (Mata et al., 2009) and (2) the acoustic quality and meaning of the stimulus (Vila et al., 2007).

In the first study, an intense white noise was presented three times by maintaining the time interval constant between the first and third presentation (30 min) and manipulating the moment of the second stimulus presentation, closer to the first one or closer to the third one. This was done in such a way as to increase and decrease symmetrically, in six-group conditions, the ITI between the three stimulus presentations. For the ITI between the first and second presentation, the time intervals were 2.5, 7.5, 12.5, 17.5, 22.5, and 27.5 min. The corresponding ITI between the second and third presentation for the same six conditions were 27.5, 22.5, 17.5, 12.5, 7.5, and 2.5 min Figure 2 presents the results for the two most extreme conditions: condition 1 (2.5 min between trial 1 and 2; and 27.5 min between trial 2 and 3) and condition 6 (27.5 min between trial

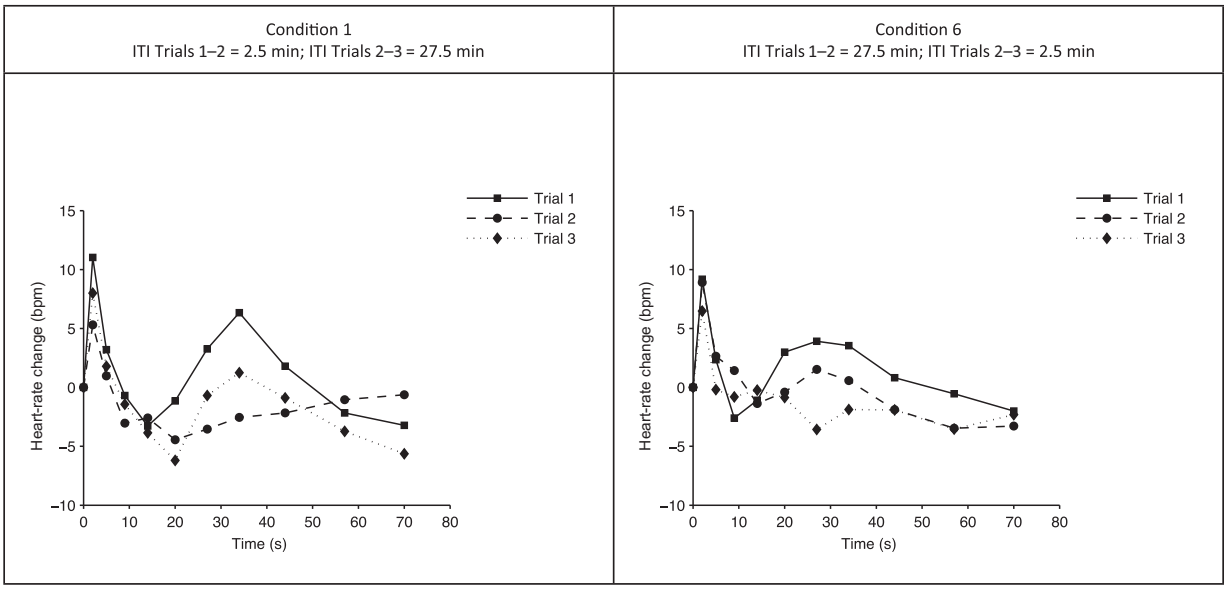

Figure 2 Habituation and dishabituation of cardiac defense: simplified heart-rate response to the three noise presentations in conditions 1 (left) and 6 (right) as a function of intertrial interval (ITI).

1 and 2; and 2.5 min between trial 2 and 3). It can be seen in Figure 2 that the expected pattern of the response, with its two accelerative/decelerative components, is observed for both conditions in the first stimulus presentation (trial 1). The expected habituation of the response – faster for the second acceleration/deceleration – is also observed in the second stimulus presentation (trial 2) when the ITI was the shortest (2.5 min). However, reduced habituation in trial 2 and increased dishabituation in trial 3 (recovery) is also observed when the ITI was the longest (27.5 min).

In the second study, a white noise and a human scream, matched in intensity and duration (95 dB and 2 s duration), were presented three times in the following sequence: 10-min rest period followed by three presentations of either the white noise or the human scream with an ITI of 120 s. Immediately after the last presentation, the same sequence was repeated with the other stimulus. As expected, the response showed rapid habituation within each sequence. However, the response pattern with its two accelerative/decelerative components was fully dishabituated (recovered) when the human scream was presented in the first sequence and the white noise in the second sequence (see Figure 3 left panel). In the opposite order, the dishabituation effect was not evident (Figure 3 right panel).

Sympathetic versus Parasympathetic Mediation

Studies examining indices of sympathetic (i.e., pre-ejection period, stroke volume, and pulse transit time) and parasympathetic (i.e., respiratory sinus arrhythmia and baroreflex sensitivity) cardiac control during the evocation of cardiac defense have found both sympathetic and parasympathetic influences (Fernández and Vila, 1989; Reyes del Paso et al., 1993, 1994). The first acceleration/deceleration is mediated by parasympathetic influences: withdrawal during the acceleration and activation during the deceleration. The second acceleration/deceleration is mediated by both sympathetic and parasympathetic influences working reciprocally: sympathetic activation/parasympathetic inhibition during the acceleration and sympathetic inhibition/parasympathetic activation during the deceleration. A coactivation of sympathetic and parasympathetic influences, with parasympathetic dominance, is also observed during the first deceleration.

Attentional versus Motivational Significance

The classical cognitive model of cardiac defense assumes an attentional interpretation of the heart-rate changes. The organism protects itself against the aversive stimulus by reducing sensory processing ('closing the sensory input'). From this cognitive perspective, defense is the opposite of orienting ('opening the sensory input'). Accordingly, cardiac defense should correlate negatively with indices of sensory intake (external attention) and positively with indices of sensory rejection (internal attention). Several studies have addressed this issue by superimposing a secondary attentional task on the evocation of cardiac defense (Pérez et al., 2000; Ramírez et al., 2010). The task could be of external attention (for example, tracking an external light on and off every half second, or performing a visual search task) or internal attention (for example, tracking one's own heartbeats or performing a memory search task). The task always started immediately after the onset of the defense stimulus. Results of these studies are very consistent confirming the opposite prediction: a positive relationship of cardiac defense with attentional processes of sensory intake (external attention).

The classical motivational model of cardiac defense assumes an energetic interpretation of the heart-rate changes. The organism protects itself by preparing for active defense

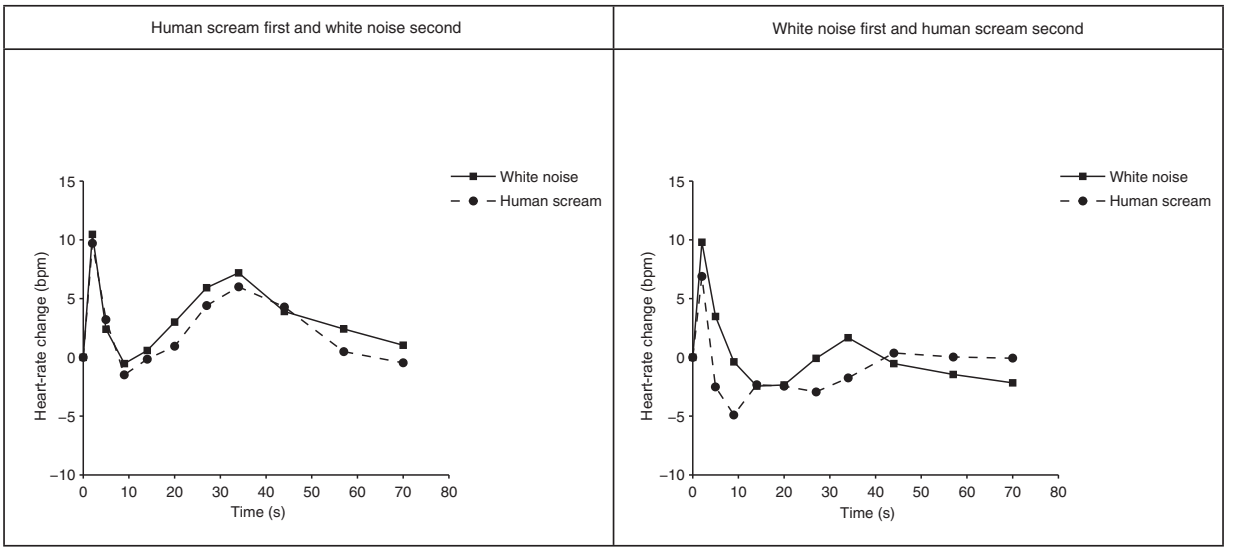

Figure 3 Dishabituation of cardiac defense: simplified heart-rate response to the human scream vs white noise when the human scream was presented in the first sequence and the white noise in the second sequence (left panel) and when the white noise was presented in the first sequence and the human scream in the second sequence (right panel).

such as fight or flight. Accordingly, the accelerative components of cardiac defense should be potentiated when participants are under negative emotional states triggered by danger or threat signals. Several studies have addressed this issue by presenting aversive pictures prior to the defense stimulus using a procedure similar to Lang's startle probe paradigm (Lang, 1995). Results of these studies are very consistent showing a marked increase in the accelerative components of the response when the emotional pictures are aversive (phobic animals or mutilated bodies). Moreover, the first deceleration disappears completely, the response being a single large acceleration finishing with the late deceleration. When the pictures were pleasant (erotic pictures) or neutral (household objects), the typical response pattern is maintained although attenuated (Sánchez et al., 2002, 2009; see Figure 4).

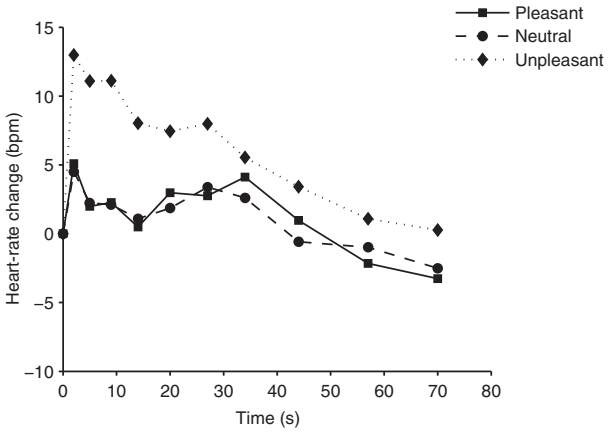

Figure 4 Emotional modulation of cardiac defense: simplified heart-rate response to the noise when an unpleasant, neutral, or pleasant picture preceded the noise.

Individual Differences

The data on cardiac defense reported in the previous sections are based on group averages. When the responses are analyzed on individual bases, research has consistently demonstrated important differences in the response pattern. The differences affect mainly the second acceleration/deceleration, and even absent in some individuals. Thus, a simple classification divides individuals into two categories: those showing the second acceleration/deceleration (called 'accelerators') and those not showing the second acceleration/deceleration (called 'decelerators'). However, cluster analysis has revealed four major response patterns: two patterns showing the second acceleration and two patterns not showing the second acceleration. Of the two patterns showing the second acceleration, one displays the pattern described in Figure 1, with the initial deceleration overpassing the baseline level; the other pattern does not display the initial deceleration (or the deceleration does not overpass the baseline level), the overall pattern being a single prolonged acceleration until the final deceleration or return to baseline level. Of the two patterns not showing the second acceleration, one displays a prolonged deceleration after the initial acceleration, which is maintained during the 80 s; the other pattern displays a return toward baseline after the initial acceleration. These individual differences have been found to be associated with various biological and psychological conditions, such as menstrual cycle (Vila and Beech, 1978), gender (Vila et al., 1992), personality traits (Cloete, 1979), excessive worry (Delgado et al., 2009), fear learning (López et al., 2009), and pathological anxiety (Sánchez-Navarro et al., 2012; Schalinski et al., 2013). In particular, the second acceleration has been found to be augmented in phobic women during the premenstrual phase, in men as compared to women, in people with high scores in worry and emotional instability, in people showing affective fear learning, and in patients diagnosed of some anxiety disorders. Moreover,

people with excessive worry and patients with generalized anxiety disorders or post-traumatic stress disorder tend to show the accelerative pattern without the initial deceleration, whereas people with subclinical social anxiety and patients with specific phobias tend to show the accelerative pattern with the initial deceleration, as described in Figure 1.

Dynamics of Cardiac Defense and the Brains' Defense System

The description of cardiac defense as a response pattern with both accelerative and decelerative components, parasympathetic and sympathetic influences, and attentional and motivational significance highlights the dynamic character of the defense reaction. As mentioned above, in natural settings, such as the imminence of a predator, the defense reaction is not a single response that could be conceptualized as attentional or motivational. Depending on the type and severity of the danger, its spatial proximity, the availability of a scape route, and either the success or failure in the initial responses to cope with it, different components of the defense reaction may take place successively. The complex pattern of heart-rate changes that characterizes cardiac defense, together with its complex physiological and psychological significance, can be better understood within this naturalistic perspective.

The response pattern, with its two accelerative/decelerative components, seems to reflect the succession of two defensive phases: an attentional protective phase linked to the first acceleration/deceleration, aimed at increasing attention to the potential danger, and a motivational protective phase linked to the second acceleration/deceleration, aimed at preparation for active defense. This sequential process, however, can be altered if threatening signals precede the eliciting stimulus or if the individual is already in an aversive emotional state. The data reported above on modulation of cardiac defense by aversive pictures and individual differences support this view. The change in the response pattern when phobic pictures precede the eliciting stimulus (transformed into a single large acceleration) suggests that the attentional phase has been preactivated by the aversive pictures and that the motivational phase has been advanced in time to better match successful adaptation. A similar interpretation may apply to individuals who show a single large acceleration in response to an unexpected intense noise. Such response pattern has been observed in people with high scores in excessive worry (Delgado et al., 2009) and in patients with post-traumatic stress disorder (Schalinski et al., 2013). A preexisting state of anticipatory anxiety or worry may act in the same way as the preceding aversive pictures (i.e., advancing the motivational phase).

The biological foundation of the cardiac defense response is in the brain's defensive system. What is known about this system comes from animal and human research on defensive reactions and fear conditioning (LeDoux, 1996; Lang et al., 2000). Neuroscientists have traced the chain of neural activation in the brain, starting from the sensory input, proceeding through the subcortical and cortical connecting structures, and finishing in the autonomic and motor effectors. A critical structure in this circuit is the amygdala, a subcortical center located within the temporal lobes that receives inputs from the sensory thalamus and various cortical structures, and projects on subcortical and brainstem areas that directly control specific defensive and fear reactions such as freezing (central gray), motor startle (nucleus reticularis pontis caudalis), and autonomic responses (lateral hypothalamus). This hierarchical organization of the brain's fear circuit, with common modulatory structures and different local pathways for different defense reactions, provides a considerable variety in the pattern of defensive output. Depending on the situational context and on the weight of the local pathways within the circuit, the aversively motivated responses can be freezing or escape, aggressive behavior, hormonal and autonomic responses, or the augmentation of motor startle.

Conclusion: Implications for Stress and Cardiovascular Disease

The new approach to cardiac defense has implications for the stress concept and stress-related illnesses. The classical models did not emphasize the dynamic character of the defense response or the simultaneous involvement of different physiological and psychological processes. The new model better matches the prevalent model of stress in health-related disciplines: the cognitive transactional model. The continuous dynamic interaction between the situation and the person, leading to activation or deactivation of the stress response, is the main feature of the transactional model. In addition, the new approach to defense can help to understand the specific nature of stress, differentiating it from other related concepts such as fear, anxiety, or arousal. Based on the new defense model, stress can be defined as the sustained activation of the brain's defense motivational system. A state of maintained activation of this system implies a variety of specific defense responses being continuously or intermittently elicited, including neural, endocrine, and immunological responses that correlate, in the long term, with poor cardiovascular health: maintained increases in heart rate, stroke volume, and blood pressure; and maintained decreases in respiratory sinus arrhythmia and cardiac variability. These sustained responses are instrumental for the development of cardiovascular diseases such as essential hypertension and coronary heart disease.

Acknowledgments

This work was supported by grants from the Junta de Andalucía (SEJ-02964) and the Spanish Ministry of Economy and Innovation (PSI2008-04372, PSI2011-28530).

See also: Cardiovascular Reactivity and Recovery; Natural Disasters: Health-Related Aspects; Startle Reflex and Health; Stress, Coping and Health.

Bibliography

Bradley, M.M., Lang, P.J., 2007. Emotion and motivation. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), Handbook of Psychophysiology, third ed. Cambridge University Press, New York, pp. 581–607.

Cloete, N., 1979. Autonomic responsivity of subjects with body boundary differences during white noise stimulation. Acta Psychologica 43, 177–183.

Delgado, L.C., Guerra, P., Perakakis, P., Mata, J.L., Pérez, M.N., Vila, J., 2009. Psychophysiological correlates of chronic worry: cued versus non-cued fear reaction. International Journal of Psychophysiology 74, 280–287.

Fernández, M.C., Vila, J., 1989. Sympathetic-parasympathetic mediation of the cardiac defense response in humans. Biological Psychology 28, 123–133.

Graham, F.K., 1979. Distinguishing among orienting, defense and startle reflexes. In: Kimmel, H.D., van Olst, E.H., Orlebke, F.J. (Eds.), The Orienting Reflex in Humans. Lawrence Associates Ltd, Hillsdale, NJ, pp. 137–167.

Lang, P.J., 1995. The emotion probe: studies of motivation and attention. American Psychologist 50, 372–385.

Lang, P.J., Davis, M., Öhman, A., 2000. Fear and anxiety: animal models and human cognitive psychophysiology. Journal of Affective Disorders 61, 137–159.

LeDoux, J.E., 1996. The Emotional Brain. Pergamon, New York.

López, R., Poy, R., Pastor, M.C., Segarra, P., Moltó, J., 2009. Cardiac defense response as a predictor of fear learning. International Journal of Psychophysiology 74, 229–235.

Mata, J.L., Rodríguez, S., Ruiz-Padial, E., Turpin, G., Vila, J., 2009. Habituation and sensitization among protective reflexes: cardiac defense and eye-blink startle. Biological Psychology 81, 192–199.

Obrist, P.A., 1981. Cardiovascular Psychophysiology: A Perspective. Plenum Press, New York.

Pérez, N., Fernández, M.C., Vila, J., Turpin, G., 2000. Attentional and emotional modulation of cardiac defense. Psychophysiology 37, 275–282.

Ramírez, I., Sánchez, M.B., Fernández, M.C., Lipp, O.V., Vila, J., 2005. Differentiation between protective reflexes: cardiac defense and startle. Psychophysiology 42, 732–739.

Ramírez, I., Guerra, P., Muñoz, M.A., Perakakis, P., Anllo-Vento, L., Vila, J., 2010. The dynamics of cardiac defense: from attention to action. Psychophysiology 47, 879–887.

Reyes del Paso, G., Godoy, J., Vila, J., 1993. Respiratory sinus arrhythmia as an index of parasympathetic cardiac control during the cardiac defense response. Biological Psychology 35, 17–35.

Reyes del Paso, G., Vila, J., García, A., 1994. Physiological significance of the defense response to intense auditory stimulation: a pharmacological blockade study. International Journal of Psychophysiology 17, 181–187.

Sánchez, M.B., Ruiz-Padial, E., Pérez, M.N., Fernández, M.C., Cobos, P., Vila, J., 2002. Modulación emocional de los reflejos defensivos mediante visualización de imágenes afectivas. Psicothema 14, 702–707.

Sánchez, M.B., Guerra, P., Muñoz, M.A., Mata, J.L., Bradley, M.M., Lang, P.J., Vila, J., 2009. Communalities and differences in fear potentiation between cardiac defense and eyeblink startle. Psychophysiology 46, 1–4.

Sánchez-Navarro, J.P., Martínez-Selva, J.M., Torrrente, G., Pineda, S., Murcia-Liarte, J.B., Carrillo-Verdejo, E., 2012. Preattentive processing of feared stimuli in blood-injection-injury fearful subjects. International Journal of Psychophysiology 84, 95–101.

Schalinski, I., Elbert, T., Schauer, M., 2013. Cardiac defense in response to imminent threat in women with multiple trauma and severe PTSD. Psychophysiology 50, 691–700.

Sokolov, E.N., 1963. Perception and the Conditioned Reflex. Pergamon Press, Elmsford, New York.

Vila, J., Guerra, P., Muñoz, M.A., Vico, C., Viedma, M., Delgado, L.C., Perakakis, P., Kley, E., Mata, J.L., Rodriguez, S., 2007. Cardiac defense: from attention to action. International Journal of Psychophysiology 66, 169–182.

Vila, J., Beech, H.R., 1978. Vulnerability and defensive reactions in relation to the human menstrual cycle. British Journal of Clinical Psychology 17, 93–100.

Vila, J., Fernández, M.C., Godoy, J., 1992. The cardiac defense response in humans: effects of stimulus modality and gender differences. Journal of Psychophysiology 6, 140–154.

Stress at Work

Johannes Siegrist, University of Duesseldorf, Duesseldorf, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Stress in modern work life is subject to profound changes. This article describes these changes and identifies the role of stress at work, with emphasis on health-adverse effects. Following a definition of core terms the function of theoretical models of stressful work is explained, and major models are introduced. Research findings are summarized documenting adverse health effects. Practical implications are discussed at two levels, national labor policies and companies and their stakeholders. In addition to reducing the gap between science and policy further progress of scientific knowledge is needed, and some promising future directions are proposed.

The Changing Nature of Work and the Role of Stress

Far-reaching changes of work and employment occurred during the past few decades in economically advanced societies. Together with rapid expansion of new information technologies and automation the proportion of people working in service occupations and professions grew, and many employment arrangements became more flexible, including an increase in self-employment and informal work contracts. The current labor force of economically advanced societies is composed of a substantial proportion of women, often working in part-time arrangements, and of an increase in older workers. Although physically strenuous work and exposure to chemical, biological, and physical hazards continue to pose serious challenges to workers' health and safety, the threats of a health-adverse psychosocial work environment are now becoming more visible, and they contribute to a growing burden of work-related diseases. This latter development has been critically intensified by the recent expansion of economic globalization. Economic globalization is characterized by free market principles spreading over the world in conjunction with technological innovations, by a stimulation of large flows of transnational capital, trade, and labor force. This process is often perceived as enhancing the gross national product and as increasing the national employment rates, particularly so in rapidly developing countries, whereas its adverse consequences on the health and well-being of workers are less strongly emphasized. However, given a growing competition and pressure toward an increase in return on investment a general intensification of work has been reported as an outcome of this process (Schnall et al., 2009). In many instances, this intensification of work goes along with an increase in job instability and insecurity, due to threats of job loss or redundancy, downsizing, outsourcing, and the privatization of public services. A different consequence of economic globalization concerns the segmentation of the labor market and a related increase in income inequality. On the one hand, there is a well-trained, skilled, and flexible workforce with fair promotion prospects and adequate earnings. On the other hand, large parts of the workforce suffer from precarious employment, low wages, hazardous jobs, and low safety at work. This segmentation occurs in the context of weakened welfare state policies and reduced power of trade unions, thus augmenting the burden of stressful work among the most vulnerable and deprived groups of national workforces.

Why is work so important for human well-being, and how does work contribute to the burden of stress and its adverse effects on health? In all advanced societies work and occupation in adult life are accorded primacy for the following reasons. First, having a job is a principal prerequisite for continuous income and, thus, for independence from traditional support systems (family, community welfare, etc.). Increasingly, the level of income determines a wide range of life chances. Second, training for a job and achievement of occupational status are the most important goals of socialization. It is through education, job training, and status acquisition that personal growth and development are realized, a core social identity outside the family is acquired, and goal-directed activity in human life is shaped. Third, occupation defines an important criterion of social stratification. The amount of esteem in interpersonal life largely depends on the type of job and level of occupational achievement. Furthermore, the type and quality of occupation, and especially the degree of self-direction at work, strongly influence personal attitudes and behavioral patterns in areas that are not directly related to work, such as leisure or family life (Kohn and Schooler, 1983). Finally, occupational settings produce the most pervasive continuous demands during one's lifetime, and they absorb the largest amount of active time in adult life, thus providing a source of recurrent negative or positive emotions. It is for these reasons that stress research in organizations, where paid work takes place, is of particular relevance.

It is important to recognize that traditional occupational hazards, such as exposure to toxic substances, heat, cold, or noise, are no longer the dominant challenges of health at work in advanced societies. Although they continue to pose serious challenges, mainly to manual workers and other less qualified occupational groups, a majority of working people are now more often exposed to mental and emotional demands and threats at work. There has been a recognition that the importance of work goes beyond traditional occupational diseases and, indeed, it is likely that work makes a greater contribution to diseases not thought of as 'occupational' in conventional terms.

For instance, the organization of work plays an important role in triggering adverse health risks. This holds particularly true for shift work and long work hours. Shift work is frequent in the production sector and in some service occupations and professions. Long-standing exposure to shift work in combination with night shift increases the risk of cardiovascular disease, and a similar health risk is attributable to recurrent overtime work (e.g., working more than 50 h per week) (Härmä, 2006). In several jobs it is increasingly difficult to disentangle work obligations from private life, in particular with the expansion of electronic communication and surveillance devices. Other stressful organizational features concern overload at work, such as high demands and work pressure, frequent interruptions, or a high level of responsibility. Interpersonal conflicts at work may accelerate the burden of stressful work, in terms of lack of support and respect and unfair treatment by supervisors or coworkers. Moreover, many jobs are characterized by monotony and a low degree of skill utilization and autonomy. If combined with reduced job stability, forced part-time or downward mobility, and low wages and salaries, these challenges are likely to increase ill health at work (Schnall et al., 2009).

There is now growing awareness among all parties of the labor market that psychosocial stress at work produces considerable costs, most importantly a high level of absenteeism, reduced productivity, elevated risks of disability pensions, and direct medical expenses. However, there is a clear need to further develop scientific evidence on the links between stressful psychosocial work environments and the burden of work-related diseases. Based on improved knowledge, interventions at different levels are required to improve healthy work at large. The following paragraphs deal with these challenges in more detail.

Concepts of Work-Related Stress and Selected Findings

Research on psychosocial work-related stress differs from traditional biomedical occupational health research by the fact that stressors cannot be identified by direct physical or chemical measurements. Rather, theoretical models are needed to analyze the particular nature of the psychosocial work environment. A theoretical model is best understood as a heuristic device that selectively reduces complex reality to meaningful components. Components are meaningful to the extent that they provide the material from which the researcher can deduce explanations and, thus, produce new knowledge. Ideally, a theoretical model of psychosocial stress at work with relevance to health should encompass a wide variety of different occupations and should not be restricted to a specific time or space.

Before describing some prominent theoretical models of occupational stress, the basic terminology needs to be clarified. Much critique has been raised against the ambiguity of the term 'stress.' To avoid this, the following terms were suggested: 'stressor' is defined as an environmental demand or threat that taxes or exceeds a person's ability to meet the challenge; 'strain' is defined as the person's response to such a situation in psychological and physiological terms. Psychological responses include specific cognitions and negative emotions (e.g., anger, frustration, anxiety, and helplessness), whereas physiological responses concern the activation of the autonomic nervous system and related neurohormonal and immune reactions. Stressors, in particular novel or dangerous ones, are appraised and evaluated by the person, and as long as there is some perception of agency on the part of the exposed person efforts are mobilized to reverse the threat or to meet the demands. Such efforts are termed 'coping,' and they occur at the behavioral (even interpersonal), cognitive, affective, and motivational level. Clearly, when judging strain, the quality and intensity of a stressor as well as the duration of exposure have to be taken into account, as well as individual differences in coping and in vulnerability to strain reactions. Recent research indicates that only part of human strain reactions are subject to conscious information processing, whereas a large amount bypasses awareness. The term 'stressful experience' is introduced to delineate that part of affective processing that reaches consciousness. Stressful experiences at work are often attributed to adverse working conditions by exposed people themselves. While they usually refer to some commonsense notions of stress, it is crucial to note that these attributions differ from the explanatory constructs of stressful experience at work that have been identified by science. Next, several such explanatory constructs are described, combined with information about their empirical contribution and their practical implications. These constructs are selected from a larger number of approaches (Cartwright and Cooper, 2009) on the basis of their theoretical significance, their standardized measurement method, and the amount of empirical evidence supporting their role in explaining stress-related disorders.

Person–Environment Fit

The first significant contribution to modern occupational stress research dates back to studies conducted more than 40 years ago at the Institute for Social Research at the University of Michigan. They were guided by a theoretical concept termed 'person–environment fit' (Caplan et al., 1980). Stressful experience at work, in this model, is conceptualized in two ways: first, as an experience where the work environment does not provide adequate supplies to meet the person's needs; second, as an experience where the abilities of the person fall short of demands that are prerequisite to receiving supplies. In both conditions, stressful experience results from a misfit between needs or abilities of the working person and demands or opportunities of the work environment. Further job opportunities may fail to fulfill the person's needs as a consequence of unmet demands. As an example, a worker with limited skills is excluded from any promotion prospects, although, to meet his level of living, he badly needs better pay.

This approach represents a type of theorizing that is termed process theory. As such, it is applicable to a wide range of work situations but does not specify the particular content dimensions on which a person and environment should be examined (Edwards et al., 1998). This lack of specificity and related difficulties of measuring the model by a uniform, psychometrically validated method may explain why empirical evidence of health-adverse consequences has been relatively sparse up to now, compared to the other constructs. Nevertheless, this

pioneering approach was important in raising awareness of the combined role of situational and personal features in triggering stressful experience at work.

Demand, Control, and Resources at Work

The most extensively tested theoretical model of a health-adverse psychosocial work environment has been labeled 'demand–control' or 'job strain' model (Karasek and Theorell, 1990). Rooted in evidence from psychobiological stress research on pathological outcomes of lack of control in challenging situations (Henry and Stephens, 1977), this construct posits that strain results from job task profiles that combine high quantitative demands with lack or low degree of control. In this model, control at work is defined in two ways, as the degree of decision latitude of the working person while performing the task and as the level of skill discretion, i.e., the opportunity of using one's competencies and of learning from one's job. High-demand–low-control jobs are more prevalent among employees in lower socioeconomic positions. Therefore, they might contribute to the explanation of the social gradient of some common stress-related chronic disorders, such as coronary heart disease (CHD) and depression, leaving people in lower socioeconomic positions at higher risk (Marmot, 2004). In fact, several studies indicate that part of the social gradient of stress-related disorders is mediated by conditions measured by this construct, whereas other studies suggest that the effect of stressful work is particularly strong among workers with low social status (Siegrist et al., 2012).

Although 'job strain' can be described as a 'black box' model that does not include the person's appraisal and coping efforts at work, it nevertheless offers the advantage of being measured by job characteristics that can be objectively assessed, either by observer ratings or by classification of job titles (Karasek and Theorell, 1990). However, in the frame of epidemiological investigations, a standardized, psychometrically validated self-completed questionnaire is usually applied for economic reasons.

Based on this important, innovative model, a substantial body of research evidence was produced over the past 20 years, and several far-reaching policy implications were drawn (see below). The model was also subject to conceptual criticism, and some relevant extensions were developed. Importantly, the availability of social support at work was included in the model as a protective resource, mitigating to some extent the health-adverse effects of stressful work. Accordingly, it was found that workers confined to high-demand–low-control jobs who additionally are socially isolated on their job suffer the highest risks of stress-related disorders (Schnall et al., 2009). The notion of a fit or misfit between extrinsic demands and the working person's resources of coping with these demands has been elaborated in more recent extensions of the original model (Bakker and Demerouti, 2007), but the empirical evidence of these extensions is still limited.

Reciprocity and Fairness at Work

While the notion of personal control or autonomy in dealing with adverse circumstances in one's work and life represents a powerful psychobiological and sociological approach toward studying human health and disease (Marmot, 2004), there are additional psychosocial dimensions of work and employment that matter for health. One such dimension concerns the balance between 'give' and 'take' at work and the fairness of exchange relationships inherent in the employment contract. Scientific studies of exchange inequity (Adams, 1965) and organizational injustice (Greenberg, 2010) were developed in social and organizational psychology several decades ago, but the theoretical link of these notions with stress research, their standardized measurement, and the test of their role in explaining health-adverse effects in epidemiological and experimental studies were more recent achievements. In this context, 'effort–reward imbalance' evolved as a widely studied theoretical model of stressful work complementing the 'job strain' model (Siegrist, 1996). This model is concerned with unjust exchange at work, i.e., the lack of reciprocity between efforts spent and rewards received in return. Rewards at work include money, esteem, and career opportunities (promotion and job security). High effort in combination with low reward matters for health and well-being, as it violates a basic 'grammar' of social exchange rooted in the notions of reciprocity and fairness. High-cost–low-gain conditions at work occur frequently in modern economies. For instance, workers who have no alternative choice in the labor market, employees working in highly competitive jobs, and people with excessive work-related motivations and attitudes ('overcommitment') are particularly vulnerable to the recurrent experience of failed reciprocity at work.

The notions of personal control and social reward seem particularly relevant in explaining health-adverse effects of stressful work, partly because they address the important psychological needs of self-efficacy and positive self-esteem. Therefore, it is not surprising that the demand–control model and the effort–reward imbalance model were shown to predict stress-related disorders independently and, if experienced simultaneously, to produce additive effects on poor health (Schnall et al., 2009; Siegrist et al., 2012).

To some extent, this holds true for a related, more recently tested concept, organizational justice (Greenberg, 2010; Elovainio et al., 2002). According to this approach, adverse effects on health are triggered by the experience of perceived unfairness of distribution of valued resources, perceived unfairness of decision-making processes in organizations, and perceived unfairness of being treated by superiors and colleagues. Despite its conceptual importance, this latter approach, until now, has been less frequently tested than the two former models.

Empirical Evidence

There are three study designs that allow a test of causal associations of work stress (as measured by theoretical models as the ones described) with health. The first one is the prospective cohort study that assesses work stress (as exposure) at study entry among an otherwise healthy working population, and that follows this population over some time, identifying the incidence of a disease under study (e.g., depression or CHD), and testing the strength of association of the exposure with the risk of incident disease. The second design concerns the assessment of psychobiological or behavioral pathways that are assumed to link the exposure with the disease outcome, either by classical stress experiments or by 'naturalistic' studies of

working people's physiological activity in everyday life (e.g., ambulatory blood pressure assessment; for review Steptoe, 2006). Third, intervention studies are used to demonstrate that a reduction in exposure is followed by a subsequent reduction in disease risk.

These designs were applied with regard to established work stress models. For instance, job strain (or at least one or several components of the model) was associated with an elevated relative risk of incident CHD, ranging from about 1.3 to about 2.0. Similar results were observed for effort–reward imbalance and organizational injustice, although based on a smaller number of studies (for review Schnall et al., 2009; Siegrist et al., 2012). With respect to depression the current evidence is comparable, documenting a twofold elevated risk of incident affective disorder among those reporting high job strain or high effort–reward imbalance (Nieuwenhuijsen et al., 2010). For other stress-related disorders with relevance to occupational life, such as musculoskeletal disorders (Rugulies and Krause, 2008) or alcohol dependence (Head et al., 2004), prospective confirmation is still rare.

In addition to epidemiological information, a substantial body of research supports the notion of psychobiological and behavioral pathways linking work stress with disease. Recent studies have explored basic regulatory processes within the body, such as inflammation or cell immunity (Chandola et al., 2010), to validate the stress-physiological bases of these models. Less evidence so far is available from intervention studies, given the procedural difficulties of conducting such studies. However, a few promising results underline the importance of intervention research as further means of justifying practical consequences derived from scientific findings (Schnall et al., 2009).

In summary, despite some inconsistencies and methodological challenges, a robust body of scientific evidence is now available demonstrating associations of an adverse psychosocial work environment, as assessed by established theoretical models, with a range of stress-related disorders.

Practical Implications and Future Directions

Unhealthy work produces substantial economic costs and reduces the health, well-being, and work performance of a large proportion of the workforce. It is therefore disappointing to observe that investments into good quality of work and employment are still relatively modest, at least in an international perspective. In addition to scientific evidence there are now several models of good practice available that demonstrate how successful solutions can be implemented at different levels. One level concerns national labor and social policies that contribute to a reduction in the burden of stressful work, e.g., by protecting workers from job instability and redundancy; by controlling overtime work, shift work, and exposure to occupational hazards; and by providing comprehensive occupational health and safety services that also meet prevention and rehabilitation needs of occupational risk groups. To this end, it is important to implement reliable and valid monitoring and surveillance systems to assess the burden of work-related stress and to guide preventive activities. Several pioneering examples are available from European countries, such as Denmark or England (Siegrist et al., 2012).

A second level addresses single companies, organizations, and stakeholders. Here, knowledge derived from theoretical models of stressful work can be used to promote employees' health. As stated above, person–environment fit theory concludes that work demands need to be tailored according to the worker's abilities. This involves both specific measures of organizational development and individual skill training. In the demand–control model, special emphasis is put on autonomy, skill discretion, and personal growth. Measures of work reorganization include job enlargement, job enrichment, skill training, enhanced participation, and teamwork. Practical implications of the effort–reward imbalance model concern the development of compensatory wage systems, the provision of models of gain sharing, and the strengthening of nonmonetary gratifications, including leadership behavior. Fair promotion prospects and procedures, and the development of a culture of trust, transparency, and sustainability, are additional measures derived from this model as well as from the organizational justice model. Importantly, these practical applications are not restricted to primary prevention but are equally valid for secondary prevention, aiming at an improvement of return to work among chronically ill, disabled, or unemployed people (Schnall et al., 2009; Siegrist et al., 2012).

To summarize, the quality of work is an important determinant of health and well-being. Recent scientific progress has identified promising entry points of interventions at the macro-, meso-, and microstructural level of policy measures. Thus, a major effort should be directed toward reducing the gap between available scientific knowledge and the implementation of policies that contribute to improved health at work. An additional effort concerns the achievement of further scientific progress.

Future Challenges for Research

There are several promising directions of future occupational stress research. First, although we witness a rapid growth of empirical evidence in the frame of epidemiological and experimental investigations, there is still a substantial lack of theory-based intervention studies demonstrating beneficial long-term effects of structural and/or behavioral changes at work. Many studies address the individual or interpersonal level rather than measures of organizational change. Positive findings are often derived from relatively privileged groups of employees, less so from low-skilled groups who suffer most from adverse employment and working conditions. In this area of research, far-reaching methodological challenges prevail as it proves to be difficult to conduct randomized controlled trials, the gold standard of evidence-based medicine. Eventually, interventions are compromised by unanticipated business changes, and control groups may be exposed to some extrinsic developments with beneficial effects. Thus, the advancement of appropriate methodological approaches seems to be a topic of high priority.

A second direction of research addresses the link or interface between work and nonwork settings, in particular work and family ('work–life balance'). In advanced societies, work and family have been identified as the two most central domains, especially so in midlife. Therefore, stress-enhancing and stress-buffering experiences in these two domains significantly contribute to personal health and well-being. This interface is

conceptualized in terms of spillover, compensation, or cumulation. Spillover refers to the transfer of experience and behavior from one domain to the other. For instance, work overload or shift work schedules may have negative effects on family life. Alternatively, experience of competence and success at work can have beneficial effects on marital life and relationships between parents and children (Kohn and Schooler, 1983). Compensation represents efforts to reduce stressful experience in one domain by improving satisfaction in another domain. For instance, leisure and family activities are enhanced to the extent that decreasing involvement in a stressful work environment is feasible. Compensation often occurs in terms of resource management in a stressful situation. In conceptual terms, there is some overlap between the buffering model of social support and compensation processes. Cumulation refers to the fact that the critical components of stressful psychosocial experience, such as lack of control or effort–reward imbalance, generalize across core life domains. For instance, disease vulnerability may be the highest among people with insecure, badly paid, effortful jobs who also live in a family arrangement that offers little emotional and material satisfaction. It is desirable to extend the theoretical models discussed, both conceptually and operationally, to life domains other than work.

Third, given the challenges of a globalized economy, the advances of work stress research that were mainly achieved in countries with modern economies need to be transferred to rapidly developing countries. In this bidirectional process of exchange, theoretical and methodological extensions and adaptations may enrich the field of research, given the powerful diversity of cultural traditions and socioeconomic developments. However, it is equally important to generate robust knowledge about those unhealthy work and employment conditions that produce similar adverse effects in different countries and branches across continents. It is only at the level of international organizations and concerted political actions that such global trends can be tackled. In conclusion, in view of the importance of work for health, it is well justified to advance the field of occupational stress both at the science and policy level.

See also: Occupational Stress Management; Organizational Climate in the Work Setting Stress and Cardiac Response; Stress, Coping and Health.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic, New York, pp. 267–299.

Bakker, A.B., Demerouti, E., 2007. The job demands-resources model: state of the art. Journal of Managerial Psychology 22, 309–328.

Caplan, R.D., Cobb, S., French, J.R.P., Harrison, R.V., Pinneau, S.R., 1980. Job Demand and Worker Health: Main Effects and Occupational Differences. Michigan: Institute for Social Research, Ann Arbor.

Cartwright, S., Cooper, C.L. (Eds.), 2009. The Oxford Handbook of Organizational Well-being. Oxford University Press, Oxford.

Chandola, T., Heraclides, A., Kumari, M., 2010. Psychophysiological biomarkers of workplace stressors. Neuroscience and Biobehavioral Reviews 35, 51–57.

Edwards, J.R., Caplan, R.D., van Harrison, R., 1998. Person–environment fit. In: Cooper, C. (Ed.), Theories of Organizational Stress. Oxford University Press, Oxford, UK, pp. 28–67.

Elovainio, M., Kivimäki, M., Vahtera, J., 2002. Organizational justice: evidence of a new psychosocial predictor of health. American Journal of Public Health 92, 105–108.

Greenberg, J., 2010. Organizational injustice as an occupational health risk. The Academy of Management Annals 4, 205–243.

Härmä, M., 2006. Workhours in relation to work stress, recovery and health. Scandinavian Journal of Work, Environment and Health 32, 502–514.

Head, J., Stansfeld, S.A., Siegrist, J., 2004. The psychosocial work environment and alcohol dependence: a prospective study. Occupational and Environmental Medicine 61, 219–224.

Henry, J.P., Stephens, P.M., 1977. Stress, Health, and the Social Environment. Springer, Berlin.

Karasek, R.A., Theorell, T., 1990. Healthy Work. Stress, Productivity, and the Reconstruction of Working Life. Basic Books, New York.

Kohn, M., Schooler, C., 1983. Work and Personality: An Inquiry into the Impact of Social Stratification. Ablex, Norwood, NJ.

Marmot, M., 2004. Status Syndrome. Bloomsbury, London.

Nieuwenhuijsen, K., Bruinvels, D., Frings-Dresen, M., 2010. Psychosocial work environment and stress-related disorders, a systematic review. Occupational Medicine 60, 277–286.

Rugulies, R., Krause, N., 2008. Effort-reward imbalance and incidence of low back and neck injuries in San Francisco transit operators. Occupational and Environmental Medicine 65, 525–533.

Schnall, P., Dobson, M., Rosskam, E. (Eds.), 2009. Unhealthy Work: Causes, Consequences, Cures. Baywood, New York.

Siegrist, J., 1996. Adverse health effects of high-effort/low-reward conditions. Journal of Occupational Health Psychology 1, 27–41.

Review of social determinants of health and the health divide in the WHO European region. In: Siegrist, J., Rosskam, E., Leka, S. (Eds.), Employment and Working Conditions Including Occupation, Unemployment and Migrant Workers. World Health Organisation, Copenhagen (Unpublished report).

Steptoe, A., 2006. Psychobiological processes linking socio-economic position with health. In: Siegrist, J., Marmot, M. (Eds.), Social Inequalities in Health. Oxford University Press, Oxford, pp. 101–126.

Stress, Coping and Health

Kate Faasse and Keith J Petrie, Faculty of Medical and Health Sciences, The University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by H.W. Krohne, volume 22, pp. 15163–15170, © 2001, Elsevier Ltd.

Abstract

Stress has been consistently associated with negative health outcomes, including increased rates of heart disease, slower wound healing, and compromised immune function. Interventions designed to improve peoples' ability to cope with stress can improve health outcomes. Such interventions include relaxation training, emotional expression, benefit finding, and cognitive-behavioral stress management and mindfulness-based stress reduction programs. Coping interventions have the potential to improve health outcomes for patients undergoing a stressful illness experience.

Introduction

There is a large body of literature surrounding the impact of stress on health, and more recently on how coping effectively with stress might influence this relationship. Most people are able to identify a stressful event or experience when it is happening to them. However, attempting to provide a definition of 'stress' that accurately encompasses the vast range of experiences that people might report as stressful elucidates stress as a broad and subjective concept.

Within the social and behavioral sciences, much of the literature on stress stems from the work of Lazarus and Folkman (1984). Their model views stress as occurring when an individual is faced with a situation that places demands on them, but for which the individual does not have the necessary resources to manage these demands. Importantly, this situation must be appraised as personally relevant to the individual, making stress a transactional process between the person and their environment. Segerstrom and O'Connor (2012) highlight that 'stress' can originate from a variety of sources, including from the environment, personal appraisal, and response to a situation.

Stress and coping are difficult to disentangle, and it is hard to imagine the need to attempt to cope without being first faced with a stressor of some sort. Indeed, coping is often viewed as an attempt to deal with a demanding situation in a way that removes the threat or at least minimizes its impact on the individual (Carver, 2006). These attempts may involve the use of cognitive and behavioral strategies in order to manage the stressful situation (Folkman and Moskowitz, 2004).

Lazarus and Folkman (1984) proposed different broad types of coping. Problem-focused coping is aimed at the source of the stress, and involves attempts to remove or avoid the threat itself. For example, a person who feels that they are at high risk of developing cancer may engage in problem-focused coping by quitting smoking, changing their diet, or taking part in an exercise program. Emotion-focused coping is aimed at managing the emotions associated with stressful experiences. In the example above, the same person may engage in emotion-focused coping by adjusting their personal perception of risk, or reframing the threat; 'we all have to die of something.' Importantly, while coping is typically thought of as a positive process, this is not necessarily the case. Likewise, it cannot be assumed that coping strategies that are adopted will necessarily be appropriate or effective in helping an individual avoid or reduce the impact of a stressor (Carver, 2006).

The focus of this article is on the relationship between stress, coping, and physical health. Health in this context is seen as the presence or absence of disease states, physiological markers of underlying disease processes, rates of healing from physical injury, and the functioning of the immune system.

Stress and Health

There is evidence spanning many fields of scientific enquiry that links stress and health. While the mechanisms may not be fully understood at this stage, it is clear that experiencing an event or situation that is perceived as stressful can have negative consequences for physical health. While the majority of evidence points to the negative impact of stress on health outcomes, it is worth noting that in some cases stress can have a beneficial impact; for example, in encouraging people to more rapidly seek assessment of worrying physical symptoms (Meechan et al., 2003).

Disease Outcomes

Perhaps one of the largest literature linking stress to disease outcomes involves cardiovascular disease. The experience of a relatively brief stressor has been associated with increased risk of experiencing a cardiac event. For example, the death of a loved one substantially increases the risk of a cardiac event in the month following the loss (Kaprio et al., 1987). Other acute stressors, including natural disasters, missile attacks, or even watching a tense sports contest have also been associated with increased mortality from heart attacks (Leor et al., 1996; Meisel et al., 1991; Wilbert-Lampen et al., 2008).

Chronic or ongoing stress has also been reliably associated with increased cardiac morbidity. Greater exposure to chronic work stress has been linked to higher blood pressure (Schnall et al., 1998) and a greater risk of coronary heart disease (Steptoe and Kivimaki, 2013). Actual and remembered hostile marital interactions can increase blood pressure (Carels et al., 1998; Ewart et al., 1991), and women with coronary heart disease who experience chronic marital stress are almost three

times more likely to experience recurrent cardiac events (Orth-Gomér et al., 2000). The chronic stress of social isolation also confers an increased risk of coronary heart disease (Steptoe and Kivimaki, 2013). A recent review of the role of stress in cardiovascular disease highlights the impact that stress can have across decades; stressors experienced during childhood, including abuse and socioeconomic adversity, are associated with higher rates of later cardiovascular disease in adulthood (Steptoe and Kivimaki, 2013).

There is also evidence for the role of stress in other disease processes. Reported physical illness increased 18% in the 3 years following the terror attacks of 9/11 in the United States (Holman and Silver, 2011). In a group of men undergoing prostate-specific antigen level testing, higher levels of distress at initial testing was associated with a 23% increased risk of receiving a prostate cancer diagnosis (Turner et al., 2009). Psychological distress has also been associated with more rapid HIV disease progression and associated $CD4^{+}$ T cell decline (Yoichi and Vedhara, 2009).

Wound Healing

Another body of work focuses on the relationship between the experience of stress and the rate with which wounds heal over time. A recent review and meta-analysis demonstrated a robust relationship between stress and wound healing, with higher levels of stress being associated with slower healing rates or other physiological markers of impaired healing (Walburn et al., 2009). These findings were seen in naturally occurring and experimental wounds, acute and chronic wounds, using different definitions and assessments of stress.

Much early work in this area has focused on standardized wounds created for the purposes of research. When comparing wound healing rates in women who were caregivers for a family member with Alzheimer's disease to matched controls, Kiecolt-Glaser and colleagues found that the experimentally inflicted wounds took a significantly longer time to heal in caregivers compared to the control participants (Kiecolt-Glaser et al., 1995). Similar findings were seen when comparing students across examination and vacation time periods; punch biopsy wounds took approximately 40% longer to heal during the stressful examination period (Marucha et al., 1998).

The impact of stress on wound healing is also seen in wounds created through surgical intervention. In patients undergoing hernia surgery, those with greater reported preoperative stress levels demonstrated an impaired immune response following surgery, as well as experiencing a more painful and slower recovery (Broadbent et al., 2003).

Immune Function

Wound healing appears to be mediated by a more adaptive immune response following tissue injury. Immune activity and the impact of stress can also be assessed using other measures of immune function, including susceptibility to viral infection and immune response to vaccination. When exposing healthy participants to the common cold virus, those with higher reported stress levels were more susceptible to infection than their less stressed counterparts (Cohen et al., 1991). Similarly, when comparing elderly participants who were caring for a spouse with dementia to control participants, caregivers produced fewer antibodies in response to a pneumococcal bacterial vaccination both 3 and 6 months after immunization (Glaser et al., 2000). However, when looking at a younger sample of spousal caregivers, similar declines in immune function have not been seen (Vedhara et al., 2002).

The effects of stress on infection appear to be seen predominantly in response to chronic stressors, which are associated with an increased risk of infection with a cold virus; however, more acute stress lasting less than 1 month was not (Cohen et al., 1998). Indeed, there is some evidence that acute stress can result in an enhanced immune response, as evidenced by higher antibody titers following vaccination (Edwards et al., 2006). A recent meta-analysis supports this and indicates that brief or acute stress may result in increased activity of natural immunity including natural killer cells and large granular lymphocytes, although acute stress does not appear to be beneficial in all instances or across all immune parameters (Segerstrom and Miller, 2004).

Coping and Health

A large body of work has sought to understand what coping strategies are associated with better health outcomes. Additionally, recent research has focused on developing coping interventions with the aim of improving patient outcomes. Peoples' coping reactions to a stressful event can vary markedly. Carver and colleagues investigated the effects of different coping strategies in a group of women undergoing treatment for early stage breast cancer (Carver et al., 1993). Lower levels of distress both before and after surgery were seen in women who reported using the coping strategies of acceptance, positive reframing, and humor. Denial and behavioral disengagement were associated with greater distress. All the women in the study were experiencing a similar stressor, yet their coping strategies and resultant distress levels varied widely.

A review of the coping literature notes that coping is associated with regulating emotions; while our focus tends to be on negative emotions and distress, positive emotions also occur alongside negative emotions, which can lead to health benefits (Folkman and Moskowitz, 2004). Importantly, the authors also note that coping skills can be taught. It is this process, and the utility of coping interventions in improving health outcomes, that will be the focus of this section. A recent review by Broadbent and Koschwanez (2012) highlights recent advances in wound healing and psychological interventions in clinical populations.

Relaxation Training

Relaxation training aims to reduce stress by lowering levels of physiological arousal, and it is included in many stress management interventions (Taylor, 2012). The effectiveness of relaxation training has been investigated as a possible intervention for improving markers of wound healing following surgery (Broadbent et al., 2012). Patients undergoing

laparoscopic cholecystectomy were randomly assigned to receive a guided imagery and relaxation intervention or standard care. Participants who received the intervention reported reduced perceived stress and increased hydroxyproline deposition, a marker of collagen deposition and healing, in the surgical wound. However, the mechanism of action remains unclear, as changes in perceived stress were not associated with rates of wound healing.

Emotional Expression

Another coping strategy that has been examined as an intervention in the context of health is that of emotional expression. It is thought that inhibiting negative emotions takes physiological effort, thus greater inhibition must be facilitated by increased physiological activity (Pennebaker, 1997). The physiological effort necessary for such inhibition may result in fewer resources being available to support other important bodily processes. Being asked to suppress thoughts, whether they are related to trauma or more mundane events, can result in a reduction in circulating immune cells, including T lymphocytes (Petrie et al., 1998). The ability to express or disclose negative emotions may reduce the need for such inhibition, reducing physiological demand (Stanton, 2010). Written emotional disclosure has the ability to increase reported physical health as well as physiological functioning in healthy participants (Smyth, 1998).

Emotional expression does not appear to be beneficial for all people or across all stressors (Stanton and Low, 2012). Expressing emotions appears to be of greatest utility for helping people adjust in situations in which they have low levels of control, where emotion-focused coping strategies tend to be of greatest benefit. Additionally, emotional expression seems to be of most benefit to people who already have a tendency toward emotional expression before encountering a stressor, and who experience intense emotions. Finally, the benefit of disclosure depends on the reactions of those to whom people choose to express their emotions. Written disclosure of a secret while imagining an accepting response from another person results in fewer reported illnesses following the experiment compared to a group imaging a nonaccepting response (Rodriguez and Kelly, 2006).

Stanton and colleagues investigated the coping strategies of women with stage I or II breast cancer, and found that women who expressed their emotions about cancer had better psychological and physical outcomes following medical treatment, including less distress, fewer cancer-related medical appointments, and enhanced physical health (Stanton et al., 2000).

Early experimental work in this area found that writing about traumatic events once a day for 4 days resulted in higher blood pressure and more negative mood immediately after writing. However, the students who were randomly assigned to write about a trauma visited the student health center fewer times during the 6 months following the writing task than those who wrote about trivial events (Pennebaker and Beall, 1986). Similarly, medical students who were randomly assigned to write about traumatic events had higher antibody levels following a Hepatitis B vaccination compared to control participants (Petrie et al., 1995). Wound healing is also enhanced in older adults engaging in expressive writing compared to control participants (Koschwanez et al., 2013).

Expressive writing also appears to confer some benefits to patient populations, and such interventions appear to have a high degree of acceptability. In HIV patients, writing about emotional topics was associated with increased CD4+ lymphocyte counts following the intervention compared to participants writing about control topics (Petrie et al., 2004). Breast and prostate cancer patients have been shown to experience a reduction in physical symptoms and less medical care utilization following writing interventions; however, the mechanism is unclear, as levels of emotional distress do not appear to be affected by the expressive writing tasks (Corter and Petrie, 2011).

Mindfulness-Based Stress Reduction

Mindfulness-based stress reduction (MBSR) involves using meditation training to help people develop skills to enable them to regulate their emotional reaction to stress (Brown and Ryan, 2003). MBSR does appear to be an effective intervention in improving coping and health across a wide range of conditions, including various cancers, coronary artery disease, and fibromyalgia (Grossman et al., 2004). In women who were recently diagnosed with early stage breast cancer, an MBSR intervention resulted in increased natural killer cell activity and cytokine production, as well as increased coping effectiveness compared to the control group (Witek-Janusek et al., 2008). A recent trial found that patients with irritable bowel syndrome who received MBSR training experienced significant improvement in symptom severity compared to a wait-list control group (Zernicke et al., 2013). However, MBSR training was not found to be effective in reducing ambulatory blood pressure in patients with hypertension when compared to wait-list controls (Blom et al., 2014).

Cognitive-Behavioral Stress Management

Cognitive-behavioral stress management (CBSM) interventions are composed of training in techniques including relaxation, cognitive restructuring, social skills, and recognizing environmental stressors (Lichstein, 1988). Following a CBSM intervention, elderly spouses caring for dementia patients demonstrated an improved immune response to an influenza vaccination when compared to both nonintervention caregivers and noncaregiver control participants (Vedhara et al., 2003). Improvements in immune function, including increased cytokine production and reduced cortisol, have also been found in postsurgical breast cancer patients who have undergone a CBSM intervention when compared to control patients (Antoni et al., 2008).

In a recent study, Antoni et al. (2012) found that a CBSM program for women with breast cancer reversed leukocyte gene expression, downregulating proinflammatory and metastasis-related genes. This suggests that stress management interventions may impact directly on immune activity by influencing proinflammatory processes at the level of gene expression. There also is some evidence to suggest that the effectiveness of CBSM may be mediated by increased benefit finding (Cruess et al., 2000).

Benefit Finding

Benefit finding is a commonly utilized emotion-focused coping strategy; many patients identify positive effects that follow a serious illness. In a group of heart attack and breast cancer patients, almost two-thirds reported some positive changes following their illness; myocardial infarction patients most often reported a healthy lifestyle change, while a closer relationship with others was the most common theme reported by breast cancer patients (Petrie et al., 1999). An experimental written emotional disclosure study with breast cancer patients found that participants who were encouraged to write about positive aspects of their breast cancer experience required fewer cancer-related medical appointments in the 3 months following the intervention compared to control participants (Stanton et al., 2002).

Conclusions

Stress is a transactional process involving an individual and their environment. Situations that place demands on a person that they do not see themselves as having the resources to meet are typically experienced as stressful. However, there is much variation regarding what situations will generate such a response between individuals and even within the same person on different occasions.

Research into the impact of stress on health involves both specific experiences that are likely to be stressful, as well as more general personal perceptions of stress. Acute stressors discussed in this article include the death of a loved one, environmental disasters, and examination stress. Chronic stressors that have been investigated include marital discord, being a caregiver for an ill spouse, and workplace stress. Importantly, the experience of a serious illness is also often perceived as a highly stressful event. Across a broad range of stressors, it is clear that stress can have negative consequences for physical health. Such consequences include increased rates of cardiovascular disease, slower healing of both experimental and surgical wounds, and dysregulation of immune function.

While stress can be detrimental to health, it is worth noting that, under some circumstances, the experience of stress can also prove to be beneficial. The stress process involves the experience of both negative and positive emotions, and these positive emotions can result in health benefits. There is also some evidence that certain acute stressors can enhance the functioning of some aspects of the immune system. Finally, the experience of distress following the discovery of physical symptoms is associated with more rapid health care seeking, enabling prompt access to treatment should it be needed.

While the experience of stress tends to be associated with negative health outcomes, coping provides a potential pathway by which to mitigate this damage. Coping refers to attempts to remove or reduce the threat that a stressful situation presents. Such attempts may involve cognitive or behavioral strategies, directed at the problem itself, or at managing negative emotions associated with the stressful experience.

However, the coping strategies that people engage in may not necessarily be helpful or successful in reducing stress or eliminating a threat. This highlights the importance of interventions that are aimed at teaching adaptive coping strategies that can improve health outcomes. Research has investigated a number of coping interventions, including relaxation training, emotional expression, benefit finding, and MBSR and CBSM programs. All of these have been associated with improved health outcomes.

It is likely that these interventions will have the most benefit when they are appropriately matched to both the individual and the situation. Participants with a natural tendency for emotional expression are likely to benefit more from such interventions, and emotion-focused techniques appear to be more useful in situations where participants have low levels of control over the stressful situation.

The mechanisms by which such interventions work require more investigation. Some studies report better health outcomes in the intervention group, but do not demonstrate a reduction in perceived stress, although it is possible that physiological changes may precede psychological adaptation. Regardless, the strong evidence linking high levels of stress to worse health outcomes and the promising results from a number of coping studies, suggest that with appropriate interventions health outcomes for many patients could be improved.

See also: Behavioral Medicine; Cardiac Disease, Coping with; Caregiver Burden; Chronic Illness, Psychosocial Coping with; Coping with Job Stress; Coronary Heart Disease: Psychosocial Aspects; Environmental Stress and Health; Health: Definitions; Mindfulness; Psychoneuroimmunology; Social Support and Recovery from Disease and Medical Procedures; Socioeconomic Status and Health.

Bibliography

Antoni, M.H., Lechner, S., Diaz, A., Vargas, S., Holley, H., Phillips, K., Blomberg, B., 2008. Cognitive behavioral stress management effects on psychosocial and physiological adaptation in women undergoing treatment for breast cancer. Brain, Behavior, and Immunity 23, 580–591.

Antoni, M.H., Lutgenforf, S.K., Blomberg, B., Carver, C.S., Lechner, S., Diaz, A., Cole, S.W., 2012. Cognitive-behavioral stress management reverses anxiety-related leukocyte transcriptional dynamics. Biological Psychiatry 15, 366–372.

Blom, K., Baker, B., How, M., Dai, M., Irvine, J., Abbey, S., Tobe, S.W., 2014. Hypertension analysis of stress reduction using mindfulness meditation and yoga: results from the harmony randomized controlled trial. American Journal of Hypertension 27, 122–129.

Broadbent, E., Kahokehr, A., Booth, R.J., Thomas, J., Windsor, J.A., Buchanan, C.M., Hill, A.G., 2012. A brief relaxation intervention reduces stress and improves surgical wound healing response: a randomised trial. Brain, Behavior, and Immunity 26, 212–217.

Broadbent, E., Koschwanez, H.E., 2012. The psychology of wound healing. Current Opinion in Psychiatry 25, 135–140.

Broadbent, E., Petrie, K.J., Alley, P.G., Booth, R.J., 2003. Psychological stress impairs early wound repair following surgery. Psychosomatic Medicine 65, 865–869.

Brown, K.W., Ryan, R.M., 2003. The benefits of being present: mindfulness and its role in psychological well-being. Journal of Personality and Social Psychology 84, 822–848.

Carels, R.A., Sherwood, A., Blumenthal, J.A., 1998. Psychosocial influences on blood pressure during daily life. International Journal of Psychophysiology 28, 117–129.

Carver, C.S., 2006. Stress, coping, and health. In: Friedman, H.S., Cohen Silver, R. (Eds.), Foundations of Health Psychology. Oxford University Press, Cary, NC, USA, pp. 117–144.

Carver, C.S., Pozo, C., Harris, S.D., Noriega, V., Scheier, M.F., Robinson, D.S., Clark, K.C., 1993. How coping mediates the effect of optimism on distress: a study of women with early stage breast cancer. Journal of Personality and Social Psychology 65, 375–390.

Cohen, S., Frank, E., Doyle, W.J., Skoner, D.P., Rabin, B.S., Gwaltney Jr., J.M., 1998. Types of stressors that increase susceptibility to the common cold in healthy adults. Health Psychology 17, 214–223.

Cohen, S., Tyrrell, D.A., Smith, A.P., 1991. Psychological stress and susceptibility to the common cold. New England Journal of Medicine 325, 606–612.

Corter, A., Petrie, K.J., 2011. Expressive writing in patients diagnosed with cancer. In: Nyklicek, I., Vingerhoets, A., Zeelenberg, M. (Eds.), Emotion Regulation and Well-being. Springer Science and Business Media, New York, pp. 297–306.

Cruess, D.G., Antoni, M.H., McGregor, B.A., Kilbourn, K.M., Boyers, A.E., Alferi, S.M., Kumar, M., 2000. Cognitive-behavioral stress management reduces serum cortisol by enhancing benefit finding among women being treated for early stage breast cancer. Psychosomatic Medicine 62, 304–308.

Edwards, K.M., Burns, V.E., Reynolds, T., Carroll, D., Drayson, M., Ring, C., 2006. Acute stress exposure prior to influenza vaccination enhances antibody response in women. Brain, Behavior, and Immunity 20, 159–168.

Ewart, C.K., Taylor, C.B., Kraemer, H.C., Agras, W.S., 1991. High blood pressure and marital discord: not being nasty matters more than being nice. Health Psychology 10, 155–163.

Folkman, S., Moskowitz, J.T., 2004. Coping: pitfalls and promise. Annual Review of Psychology 55, 745–774.

Glaser, R., Sheridan, J., Malarkey, W.B., MacCallum, R.C., Kiecolt-Glaser, J.K., 2000. Chronic stress modulates the immune response to a pneumococcal pneumonia vaccine. Psychosomatic Medicine 62, 804–807.

Grossman, P., Niemann, L., Schmidt, S., Walach, H., 2004. Mindfulness-based stress reduction and health benefits: a meta-analysis. Journal of Psychosomatic Research 57, 35–43.

Holman, E.A., Silver, R.C., 2011. Health status and health care utilization following collective trauma: a 3-year national study of the 9/11 terrorist attacks in the United States. Social Science and Medicine 73, 483–490.

Kaprio, J., Koskenvuo, M., Rita, H., 1987. Mortality after bereavement: a prospective study of 95 647 persons. American Journal of Public Health 77, 283–287.

Kiecolt-Glaser, J.K., Marucha, P.T., Malarkey, W.B., Mercado, A.M., Glaser, R., 1995. Slowing of wound healing by psychological stress. Lancet 346, 1194–1196.

Koschwanez, H.E., Kerse, N., Darragh, M., Jarrett, P., Booth, R.J., Broadbent, E., 2013. Expressive writing and wound healing in older adults: a randomized controlled trial. Psychosomatic Medicine 75, 581–590.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer Publishing, New York.

Leor, J., Poole, W.K., Kloner, R.A., 1996. Sudden cardiac death triggered by an earthquake. New England Journal of Medicine 334, 413–419.

Lichstein, K., 1988. Clinical Relaxation Strategies. Wiley, New York.

Marucha, P.T., Kiecolt-Glaser, J.K., Favagehi, M., 1998. Mucosal wound healing is impaired by examination stress. Psychosomatic Medicine 60, 362–365.

Meechan, G., Collins, J., Petrie, K.J., 2003. The relationship of symptoms and psychological factors to delay in seeking medical care for breast symptoms. Preventive Medicine 36, 374–378.

Meisel, S.R., Kutz, I., Dayan, K.I., Pauzner, H., Chetboun, I., Arbel, Y., David, D., 1991. Effects of Iraqi missile war on incidence of acute myocardial infarction and sudden death in Israeli civilians. Lancet 338, 660–661.

Orth-Gomér, K., Wamala, S.P., Horsten, M., Schenck-Gustafsson, K., Schneiderman, N., Mittleman, M.A., 2000. Marital stress worsens prognosis in women with coronary heart disease: the Stockholm female coronary risk study. Journal of the American Medical Association 284, 3008–3014.

Pennebaker, J.W., 1997. Writing about emotional experiences as a therapeutic process. Psychological Science 8, 162–166.

Pennebaker, J.W., Beall, S.K., 1986. Confronting a traumatic event: toward an understanding of inhibition and disease. Journal of Abnormal Psychology 95, 274–281.

Petrie, K.J., Booth, R.J., Pennebaker, J.W., 1998. The immunological effects of thought suppression. Journal of Personality and Social Psychology 75, 1264–1272.

Petrie, K.J., Booth, R.J., Pennebaker, J.W., Davison, K.P., Thomas, M.G., 1995. Disclosure of trauma and immune response to a hepatitis B vaccination program. Journal of Consulting and Clinical Psychology 63, 787–792.

Petrie, K.J., Buick, D.L., Weinman, J., Booth, R.J., 1999. Positive effects of illness reported by myocardial infarction and breast cancer patients. Journal of Psychosomatic Research 47, 537–543.

Petrie, K.J., Fontanilla, I., Thomas, M.G., Booth, R.J., Pennebaker, J.W., 2004. Effect of written emotional expression on immune function in patients with human immunodeficiency virus infection: a randomized trial. Psychosomatic Medicine 66, 272–275.

Rodriguez, R.R., Kelly, A.E., 2006. Health effects of disclosing secrets to imagined accepting versus nonaccepting confidants. Journal of Social and Clinical Psychology 25, 1023–1047.

Schnall, P.L., Schwartz, J.E., Landsbergis, P.A., Warren, K., Pickering, T.G., 1998. A longitudinal study of job strain and ambulatory blood pressure: results from a three-year follow-up. Psychosomatic Medicine 60, 697–706.

Segerstrom, S.C., Miller, G.E., 2004. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. Psychological Bulletin 130, 601–630.

Segerstrom, S.C., O'Connor, D.B., 2012. Stress, health and illness: four challenges for the future. Psychology & Health 27, 128–140.

Smyth, J.M., 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. Journal of Consulting and Clinical Psychology 66, 174–184.

Stanton, A.L., 2010. Regulating emotions during stressful experiences: the adaptive utility of coping through emotional approach. In: Folkman, S. (Ed.), Oxford Handbook of Stress, Health and Coping. Oxford University Press, New York.

Stanton, A.L., Danoff-Burg, S., Cameron, C.L., Bishop, M., Collins, C.A., Kirk, S.B., Twillman, R., 2000. Emotionally expressive coping predicts psychological and physical adjustment to breast cancer. Journal of Consulting and Clinical Psychology 68, 875–882.

Stanton, A.L., Danoff-Burg, S., Sworowski, L.A., Collins, C.A., Branstetter, A.D., Rodriguez-Hanley, A., Austenfeld, J.L., 2002. Randomized, controlled trial of written emotional expression and benefit finding in breast cancer patients. Journal of Clinical Oncology 20, 4160–4168.

Stanton, A.L., Low, C.A., 2012. Expressing emotions in stressful contexts: benefits, moderators, and mechanisms. Current Directions in Psychological Science 21, 124–128.

Steptoe, A., Kivimaki, M., 2013. Stress and cardiovascular disease: an update on current knowledge. Annual Review of Public Health 34, 337–354.

Taylor, S.E., 2012. Health Psychology, eighth ed. McGraw-Hill, New York.

Turner, E.L., Lane, J.A., Metcalfe, C., Down, L., Donovan, J.L., Hamdy, R., Vedhara, K., 2009. Psychological distress and prostate specific antigen levels in men with and without prostate cancer. Brain, Behavior, and Immunity 23, 1073–1078.

Vedhara, K., Bennett, P.D., Clark, S., Lightman, S.L., Shaw, S., Perks, P., Shanks, N.M., 2003. Enhancement of antibody responses to influenza vaccination in the elderly following a cognitive-behavioural stress management intervention. Psychotherapy and Psychosomatics 72, 245–252.

Vedhara, K., McDermott, M.P., Evans, T.G., Treanor, J.J., Plummer, S., Tallon, D., Schiffitto, G., 2002. Chronic stress in nonelderly caregivers: psychological, endocrine and immune implications. Journal of Psychosomatic Research 53, 1153–1161.

Walburn, J., Vedhara, K., Hankins, M., Rixon, L., Weinman, J., 2009. Psychological stress and wound healing in humans: a systematic review and meta-analysis. Journal of Psychosomatic Research 67, 253–271.

Wilbert-Lampen, U., Leistner, D., Greven, S., Pohl, T., Sper, S., Volker, C., Guthlin, D., Plasse, A., Knez, A., Kuchenhoff, H., Steinbeck, G., 2008. Cardiovascular events during World Cup soccer. New England Journal of Medicine 358, 475–483.

Witek-Janusek, L., Albuquerque, K., Rambo Chroniak, K., Chroniak, C., Durazo-Arvizu, R., Mathews, H.L., 2008. Effect of mindfulness based stress reduction on immune function, quality of life and coping in women newly diagnosed with early stage breast cancer. Brain, Behavior, and Immunity 22, 969–981.

Yoichi, C., Vedhara, K., 2009. Adverse psychosocial factors predict poorer prognosis in HIV disease: a meta-analytic review of prospective investigations. Brain, Behavior, and Immunity 23, 434–445.

Zernicke, K.A., Campbell, T.S., Blustein, P.K., Fung, T.S., Johnson, J.A., Bacon, S.L., Carlson, L.E., 2013. Mindfulness-based stress reduction for the treatment of irritable bowel syndrome symptoms: a randomized wait-list controlled trial. International Journal of Behavioral Medicine 20, 385–396.

Stress: Effects on Neonatal and Child Development

Vivette Glover, Imperial College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

If a mother is stressed, anxious, or depressed while pregnant, her child is at greater risk of a range of problems, including increased anxiety, attention deficit hyperactivity disorder, and conduct disorder, as well as lower birthweight. The mechanisms underlying this fetal programming are starting to be understood and probably involve altered placental function. There may be an evolutionary value in the changes observed, which helped the child to survive in a stressful environment in early human communities, but are maladaptive now. The clinical implications are that better emotional care of pregnant women should help to improve the outcomes for their children.

Introduction

If we want to understand the child development, or the origins of psychopathology, we need to start before birth. It has been known for many decades that how we turn out depends on an interaction between our genes and our environment. What we now realize is that an important part of our environment is in the womb. This is because of fetal programming, the fact that the milieu in the womb, during different sensitive periods for specific outcomes, can alter the development of the fetus, with a permanent effect on the child and later adult. David Barker has conducted epidemiological studies, which have now been confirmed by many independent groups across the world, that show that a healthy lower birthweight baby is at greater risk of dying from cardiovascular disease in later life, than a larger newborn baby (Barker, 2007). If organs such as the kidney or the pancreas develop less well in the womb they may cope less well with an adverse environment later.

There is now considerable evidence that fetal programming is also important for neurodevelopment and that the mother's emotional state during pregnancy can affect the development of her baby's brain (Van den Bergh et al., 2005; Talge et al., 2007; Charil et al., 2010). This has been known for many decades from animal studies (Weinstock, 2001). It has been shown that in rats, guinea pigs, and monkeys, if the pregnant animal is stressed there are long-term effects on the offspring. These include increased anxiety, reduced attention, and many structural and functional changes in the brain. In animal studies, the effects can be proved to be prenatal rather than postnatal because the offspring can be crossed fostered to a mother not stressed during pregnancy, or in the case of monkeys, all the offspring can be reared together in a nursery.

Longitudinal Human Studies – Types of Outcomes Affected

The risk for a wide range of different outcomes has been shown to be raised in humans in association with prenatal stress (see Table 1). These include both neurodevelopmental, such as emotional and behavioral problems, and physical problems, such as asthma. It is important to note that in all these studies, the findings show only an increased risk. Most children are not affected. But an increased risk, for example, a doubling of emotional/behavioral problems from 5% to 10%, if the mother is in the top 15% of anxiety or depression as shown in a normal population (O'Connor et al., 2002), is of real clinical significance.

Different studies have examined children at times from birth until adulthood. Many have shown that prenatal stress is associated with somewhat lower birthweight and reduced gestational age. Another finding at birth is an altered sex ratio, with fewer males to females being born than in an unstressed population. Studies have found an increased proportion of children who are mixed handed, rather than right handed, after prenatal stress, and also altered fingerprint patterns. These physical alterations are of interest because they are features that are known to develop *in utero*. Being mixed handed is not a problem in itself, but it is known that people with a range of neurodevelopmental problems such as autism and schizophrenia are also more likely to be mixed handed. A recent study has shown that prenatal stress is associated with reduced telomere length. Telomeres are the end pieces of DNA on each chromosome, which protect them from deterioration. This is an intriguing finding, as well as of concern, as reduced telomere length is associated with a reduced life span.

Table 1 Neurodevelopmental and physical outcomes shown to be associated with prenatal stress

Neurodevelopmental outcomes	*Physical outcomes*
Altered neonatal behavior	Reduced birthweight
More difficult infant temperament	Reduced gestational age
Altered sleep pattern in infants	More mixed handedness
	Altered sex ratio; fewer boys born
Increased anxiety and depression in children	Altered fingerprint pattern
Increased behavioral problems: ADHD, conduct disorder in children	Decreased telomere length
Impaired cognitive development, especially language in children	Increased asthma
Increased autism in children?	Reduced immune function
Increased schizophrenia in adults	Increased obesity
	Increased diabetes

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23003-1

Some investigators have looked at the newborns of mothers who report stress during pregnancy and found a poorer performance on the Neonatal Behavioral Assessment Scale relative to newborns of mothers who do not report stress during pregnancy, showing that adverse behavioral outcomes are also observable from the very beginning. Studies of infants and toddlers have shown more difficult temperament, sleep problems, lower cognitive performance, and increased fearfulness associated with higher maternal stress during pregnancy. Other studies have examined the association between prenatal stress and neurodevelopmental outcomes in children aged 3–16 years. Many independent groups have shown that prenatal stress increases child emotional problems, especially anxiety and depression, and symptoms of attention deficit hyperactivity disorder (ADHD) and conduct disorder. Other studies have shown a reduction in cognitive performance associated with prenatal stress.

Some research has found an association between prenatal stress and increased risk of autism, although a large population study has failed to confirm this. Two studies have found an increased risk of schizophrenia in adults born to mothers who experienced stress during pregnancy. Both showed effects with severe stress, the death of a relative or exposure to the invasion of the Netherlands in 1940, and both showed that the sensitive period of exposure was during the first trimester. One group, using magnetic resonance imaging, has shown associations between prenatal stress and specific regional reductions in gray matter density in the brain. Such altered gray matter may be associated with neurodevelopmental and psychiatric disorders as well as cognitive and intellectual impairment.

There is little consistency in the literature as to the most sensitive time in gestation for the influence of prenatal stress. It is likely that there are different times of sensitivity dependent on the outcome studied and the stage of development of the relevant brain structures. The two studies of schizophrenia found that the most sensitive period was in the first trimester. This is when neuronal cells are migrating to their eventual site in the brain, a process previously suggested to be disrupted in schizophrenia. In contrast, two studies of conduct disorder, or antisocial behavior, found the greatest associations with stress in late pregnancy.

Types of Prenatal Stress Associated with Altered Child Outcome

Stress is a generic term, which includes a wide range of different types of exposure, and many have been shown to be associated with altered outcome for the child. They include maternal anxiety and depression, pregnancy-specific anxiety, daily hassles, bereavement, life events, a bad relationship with the partner, and exposure to acute disasters such as a Canadian ice storm, 9/11, Chernobyl, a hurricane in Louisiana, or war. It is clear that it is not just a diagnosable mental illness or very extreme or 'toxic stress' that can alter outcome. Exposures that can have an effect vary from the very severe, such as the death of an older child to quite mild stresses, such as daily hassles.

It has been suggested that mild to moderate stress may actually improve some outcomes. Mild prenatal stress has been shown in some studies to accelerate motor development and cognitive ability. This is an interesting idea and deserves further investigation. Other studies have found a linear dose response between prenatal maternal anxiety and emotional/behavioral outcomes for the child. It is possible that prenatal stress has different patterns and direction of effect for different outcomes. For example, mild to moderate stress may accelerate physical maturation and cognitive function while also increasing symptoms of anxiety.

Is It Causal?

Many human studies, as discussed earlier, have shown that there is an association between maternal stress during pregnancy and an altered outcome for the child. The evidence for this is very strong and has been shown in many independent prospective studies from around the world. What is harder to establish is that the association is causal. If a mother is stressed while she is pregnant she may well be stressed postnatally and this could affect her parenting. There can be other associated confounding factors such as smoking or alcohol consumption, which may affect her child, for example, in behavior and birthweight. There also could be genetic continuity. The mother may have certain genes which make her more likely to become anxious or depressed and she may pass these genes on to her child, which in turn makes them more prone to emotional or behavioral problems.

Several studies have tried to address these points but the first evidence to consider is that from animals. With animal studies it is much easier to establish that prenatal stress has a direct effect on the outcome for the offspring. Newborn rat pups of prenatally stressed mothers can be cross fostered to nonstressed mothers on the first day after birth, with control pups of unstressed mothers cross fostered also. This can establish that any differences in outcome are caused by stress in the prenatal period. Many such studies have shown that there are definite programming effects of prenatal stress on behavior, cognitive development, and brain structure of the offspring. The nature of the effects can be affected by the timing of the exposure in gestation, the type of the stress, the strain of the animal, the age at which the offspring was tested, and the sex of the offspring. Some altered outcomes are not observed in the youngest offspring, but become apparent as they mature. The effects can also depend on the sex of the offspring. In general, although not always, prenatal stress increases anxiety and depressive behavior to a greater extent in female offspring and impairs learning and cognition more in the males (Weinstock, 2007).

In the animal studies, there has been much research on the underlying mechanisms. The effects of prenatal stress on the offspring can often be mimicked by giving the stress hormone cortisol (corticosterone in rodents), or a synthetic glucocorticoid such as dexamethasone, to the pregnant animal. Some of the effects of prenatal stress can also be moderated by the quality of the postnatal maternal care. All is not over at birth.

In humans there are several types of evidence which also suggest that prenatal stress is causing fetal programming, but it is harder to be definitive. There is good evidence for an association between the mother's emotional state and the behavior or heart rate of her fetus. Experiments in which a pregnant mother is asked to carry out a stressful computer task, while the

fetal heart rate is monitored, showed that the fetal heart rate went up during the task, but only in mothers who rated themselves as anxious. Thus even before birth the fetus can be affected by the maternal emotional state, although we do not know what the mechanism is for this are (it is too quick to be caused by the stress hormone cortisol). There is evidence for continuity between fetal behavior and neurological maturation in the first weeks after birth. The fact that maternal stress during pregnancy is associated with altered outcomes at birth including reduced birthweight and reduced scores on a neonatal assessment is some evidence for prenatal, independent of postnatal, effects.

Another approach to establishing that the associations between the maternal emotional state and long-term outcome for the child are, at least in part, causal, is by controlling for confounding factors such as prenatal smoking and alcohol consumption, and for postnatal maternal mood. Several studies have done this and found a strong signal remaining for prenatal anxiety or depression, thus controlling for impaired parenting due to postnatal depression or anxiety (O'Connor et al., 2002). If the observed associations are primarily due to an anxious mother passing on predisposing genes to her child, it would not be expected to be specifically associated with prenatal as opposed to postnatal mood. There clearly are additional effects of both postnatal maternal mood and parenting. For example, the association between prenatal anxiety and child fearfulness was found to be greater in those children with an insecure attachment to their mother. However, many studies have shown that there is good evidence for a prenatal causal component also.

An interesting study compared the association between prenatal stress and child outcome, in children born after *in vitro* fertilization, in those who were genetically related to the mother with those who were not (Rice et al., 2010). They showed that there was an association between maternal stress in pregnancy and child symptoms of ADHD and conduct disorder, and that the association with conduct disorder was apparent in the unrelated mothers. This gives strong support to the idea that the association between prenatal stress and child conduct disorder is independent of genetic factors. However, the fact that the increase in symptoms of ADHD was apparent only in those with related mothers does not conclusively rule out a prenatal environmental component. There may be a gene–environment interaction. Prenatal stress may only have the effect of increasing symptoms of ADHD in the genetically vulnerable mother and child pairs. More research is needed to disentangle the role of genetic factors for all outcomes.

One further indication that the effects of prenatal stress are not just due to genetic continuity is the group of studies that have shown children of mothers exposed to acute disasters, such as the Canadian ice storm (Laplante et al., 2008). With these 'natural experiments,' the level of stress was objectively assessed, and the exposure was of a specific duration. This reduces the confounding effects of pre-existing emotional problems and genetic continuity, and also postnatal emotional and parenting effects.

Although there is some room for skepticism, the evidence is mounting that in humans, as in animals, prenatal stress has a direct causal effect on fetal development including neurodevelopment. However, the early postnatal environment is equally important and can either exacerbate or ameliorate the prenatal effects.

Evolutionary Significance

The changes that we observe in the child whose mother was exposed to stress during pregnancy may be explained in evolutionary terms. Maternal prenatal stress and anxiety may program the offspring to respond more effectively in a future dangerous or stressful environment (Glover, 2011). The programming effects do not seem to be specifically associated with a particular clinical diagnosis in the mother, as discussed earlier, but with a range of different conditions which make her feel stressed. Our emotions and behaviors have been shaped by natural selection. Attacks by predators, exclusion from the group, and opportunities for mating were sufficiently important and frequent to have shaped special patterns of behavior. Some of the altered outcomes observed after prenatal stress, may well, in their milder forms, have also been adaptive for our ancestors. But a mismatch between what was adaptive in an earlier environment and the world in which we now live, can lead to pathology. Mammals have developed the capacity to respond in an appropriate way *in utero*, presumably to improve fitness, and be better prepared for the environment in which they would find themselves after birth (Gluckman et al., 2005).

It is of interest to speculate how the various changes observed in children exposed to prenatal stress may have been adaptive in a stressful environment for early human communities. One of the most common findings in children who have been exposed to prenatal stress is an increase in anxiety. Normal anxiety is an emotion that helps organisms defend against a wide variety of threats. This evolutionary perspective on anxiety explains why we are more likely to develop phobias to archaic dangers like snakes or spiders, and not to modern dangers, such as cars. It is easy to see how extra vigilance may be more necessary and adaptive in a more stressful environment. A mild threat causes a general increase in anxiety that helps to locate the source and type of danger and to plan possible ways to deal with it.

Another common outcome associated with prenatal stress is ADHD. Distracted attention may be adaptive in a dangerous environment, which the pregnant mother has experienced as stressful. Sensitivity to small signals, such as a new sound, may alert one to the approaching presence of a predator or a hostile human. The hyperactive/impulsive dimension of ADHD, novelty seeking, risk taking and seeking maternal attention as well as increased sexual activity, all may have been adaptive too. Novelty seeking and increased exploration could propel migration, which could be important in a stressful environment. Exploration does not always have to be well organized. Random search strategies can be beneficial too.

An increased risk of conduct disorder is also associated with prenatal stress. It is easy to understand how more aggressive behavior may be adaptive in a stressful environment in which there is a shortage of resources or more danger from outsiders. Our early human ancestors could have been at risk from carnivorous predators and also from other humans, such as

neighboring tribes. An ability to react quickly, and to be aggressive, as well as to mount stress responses, may have been helpful to escape or to fight back. Other aspects of conduct disorder may be adaptive in such circumstances too. For example, breaking the rules, or behaving in a way different from the group, may aid the discovery of new environments or new strategies. In times of stress or difficulty, those with these dispositions may be the ones who try out new things or literally explore new areas.

We do not know to what extent the cognitive impairment detected after prenatal stress is a side effect of readily distracted attention or other behavioral changes. However, there may be advantages in having cognitive variation in a population, similar to the possible benefits of impulsivity. Individuals who learn, understand, and interpret things in different ways may also have had adaptive advantage under varying conditions in an earlier society.

Underlying Biological Mechanisms

Little is understood about the mechanisms that may underlie fetal programming by prenatal stress in humans. One early suggestion was a decrease in blood flow to the fetus. However, it is not clear if the decrease observed would be clinically significant, and others have failed to replicate the original finding. Another possible mediating factor is increased exposure of the fetus to cortisol, as has been shown in animal studies. Glucocorticoids (cortisol in humans and primates, corticosterone in rodents) are known to have a range of effects on the developing fetus. While they are essential for fetal development and tissue maturation, overexposure of the fetus to glucocorticoids can have effects that predispose to ill health in later life, such as a reduction in birthweight and epigenetic changes in the brain (Harris and Seckl, 2011).

Fetal overexposure to glucocorticoids could occur through an increase in maternal cortisol associated with anxiety and stress, which then crosses the placenta into the fetal environment. In animal models, this has been shown to be one mechanism. If pregnant rhesus monkeys are hormonally stimulated to have increased cortisol production, there are altered offspring neurodevelopmental outcomes, similar to those seen in response to prenatal stress. The effects of prenatal stress in rats have been shown to be prevented by removing the adrenal gland, and, at least for some outcomes, reinstated by corticosterone administration. However, the human hypothalamic–pituitary–adrenal (HPA) axis which makes cortisol functions differently in pregnancy from in most animals. The maternal HPA axis becomes gradually less responsive to stress as pregnancy progresses. Only a weak, if any, association between maternal mood and her cortisol level exists, especially later in pregnancy. It thus seems unlikely that an increase in maternal cortisol is the major mediating mechanism between prenatal maternal stress, anxiety or depression in later pregnancy, and altered fetal development.

It is possible that the effects of prenatal stress on fetal programming are mediated by cortisol without increases in maternal levels. An alternative explanation is that stress or anxiety causes increased transfer of maternal cortisol across the placenta to the fetus. The placenta clearly plays a crucial role in moderating fetal exposure to maternal factors and presumably in preparing the fetus for the environment in which it is going to find itself (O'Donnell et al., 2009). Thus another mechanism by which the fetus could become overexposed to glucocorticoids is through changes in placental function, especially the enzyme 11β-hydroxysteroid dehydrogenase type II (11β-HSD2), the barrier enzyme which converts cortisol to the inactive cortisone. If there is less of this barrier enzyme then the fetus will be exposed to more maternal cortisol, independently of any change in the maternal cortisol level. There is some evidence in both rat models and in humans that prenatal stress can indeed reduce placental 11β-HSD2.

Maternal cortisol is correlated with fetal cortisol levels, and so higher maternal levels will result in higher fetal levels. It is thus possible that both maternal cortisol and maternal anxiety/depression are independent determinants of fetal exposure, with the maternal anxiety inducing increased placental permeability to the raised cortisol. There is some evidence that this is the case. The question remains of what changes in maternal chemistry, associated with maternal mood in pregnancy, have this influence on placental function. One possibility is the group of chemicals called the cytokines, which are involved in the immune system and in inflammation. Psychosocial stress during pregnancy has been shown to be associated with raised levels of inflammatory cytokines such as interleukin-6. Although, how, if at all, these affect placental function remains to be determined.

The level of cortisol in the amniotic fluid, surrounding the baby in the womb, has been shown to be inversely correlated with infant cognitive development at about 18 months. The higher the cortisol in the amniotic fluid, the lower the infant performance on a baby IQ test. However, the study also measured the attachment between the mother and baby and found that this inverse relationship was strong in insecurely attached infants but absent in the securely attached. This suggests that the quality of attachment may be able to buffer against, at least this effect, of raised *in utero* cortisol. The quality of the postnatal mothering is as important as what happens in the womb.

Epigenetic changes, which involve reversible changes to the structure of DNA, such as the addition of a methyl group, control the amount of mRNA and protein produced. Such epigenetic changes can be induced by the environment and may underlie many of the processes of fetal programming. Prenatal stress has been shown to cause epigenetic changes in the rodent brain, in the DNA codes for the receptor which binds the stress hormone corticosterone. In humans, stress during pregnancy caused by violence from the partner has been shown to cause epigenetic changes in the DNA for this same receptor in the blood of their adolescent children. This suggests that the HPA stress responses in these children may be augmented, and this in turn may underlie an increased vulnerability to anxiety or depression.

However, several other biological systems are likely to be involved as well as that producing cortisol. Serotonin (also known as 5-hydroxytryptamine) is another possible mediator of prenatal stress-induced programming effects on offspring neurocognitive and behavioral development. During gestation serotonin acts as a trophic factor regulating cell division,

differentiation, and synaptogenesis. Animal studies have shown that increased serotonin exposure during gestation is associated with alterations in many neuronal processes and subsequent changes in offspring behavior. Recent work has identified an endogenous serotonin biosynthetic pathway in the human placenta, suggesting a possible role for alterations in placental serotonin in human fetal programming.

Implications for Intervention

The main implication arising from the evidence of the potentially harmful effects of stress during pregnancy on the child is to improve the emotional care of pregnant women. While the physical care of pregnant women has improved dramatically over the last century, their emotional care has not. It is still true that most anxiety, depression, and emotional and physical abuse experienced by pregnant women is undetected by health professionals and little help seems to be available, even in the developed world. In the developing world, the situation is even worse. It is not sufficiently known that symptoms of anxiety and depression are at least as common during pregnancy as postnatally and domestic abuse can increase. Midwives and obstetricians need to be trained to detect all this and refer women to further appropriate help when needed. Up to 15% of pregnant women and their children would benefit from relevant extra support. For most depressed or anxious women, some sort of counseling or talking therapy, such as cognitive behavioral therapy, will be most suitable. For the more severely depressed, antidepressants may be the best option.

The magnitude of the increased risk to the child is clinically significant, although most children of stressed mothers are not affected, and those that are, are affected in different ways. In the large population Avon Longitudinal Study of Parents and Children study, it was found that if the mother was in the top 15% for anxiety, her child was at double the risk for emotional/behavioral problems later (O'Connor et al., 2002). The risk was raised from about 5 to 10%. Similar results have been found more recently for prenatal depression. The effects of prenatal stress on cognitive development also appear to be clinically significant. In a study correlating prenatal life events with child cognitive development at 17 months, Bergman et al. (2007) found that prenatal stress accounted for 17% of the variance in cognitive ability, after allowing for a range of confounders, including postnatal maternal mood. Prenatal partner relationship strain accounted for most of the life event stress-related variance in cognitive ability. If the woman said she had suffered from three or more prenatal partner-related life events (e.g., 'your partner was emotionally cruel to you') her infant had a mean Mental Developmental Index of 89, compared with 98 for the rest of the group (the general population norm being 100).

As maternal stress during pregnancy increases the risk for an adverse outcome for the child, interventions to reduce such stress should improve the outcome. There has not yet been any study of an intervention specifically designed to reduce maternal stress in pregnancy with a follow-up study of outcome for the child. The only intervention which starts during pregnancy and which has included long-term follow-up studies of the child is the Nurse Family Partnership (Olds et al., 2004). This is targeted at deprived especially teenage and poor single mothers and includes multiple home visits by specially trained nurses during pregnancy and for the first 2 years. It was not designed to target stress but help the mothers with diet, health care, their own education, reduction in smoking, and with parenting. However, these special nurses do provide much support for these mothers from early pregnancy onward. The results of this intervention have been impressive, with many long-term improvements in self-care, parenting behavior, and child outcome. A notable finding was a reduction of criminal behavior, especially in girls. Relative to the comparison group, by 19 years of age, girls in the pregnancy and infancy nurse-visited group were less likely to have been arrested and convicted. There was no comparable benefit for boys. As prenatal stress causes an increased risk for conduct disorder, ADHD and cognitive problems, and all these are major risk factor for later criminal behavior, one would expect a reduction in prenatal stress to reduce crime as well as improving other outcomes. It may be that, for boys, a 'top up' intervention in the teenage years is needed.

Although there have been no long-term follow-up studies on the child of interventions that start in pregnancy, apart from the Nurse Family Partnership, there have been several studies of interventions that reduce aspects of stress during pregnancy. Interpersonal therapy has been shown to be effective at reducing depression in pregnant women. Listening to music has been found to decrease maternal plasma cortisol and self-reported state anxiety score in women awaiting amniocentesis. Maternal relaxation has also been shown to improve indices of fetal neurobehavior, such as heart rate variability. One study compared active (directed by a therapist) and passive (sitting in a chair) relaxation. Both active and passive relaxation significantly reduced self-rated anxiety and maternal heart rate, but the effect was significantly greater with the active relaxation. In contrast, the passive relaxation significantly reduced noradrenaline levels whereas active did not. Both methods significantly reduced cortisol. However, overall there was a striking lack of correlation between the psychometric and biological indices. In order to reduce specific biological effects of anxiety or depression during pregnancy, different methods may be needed from those that are most effective at reducing subjective psychological symptoms.

New studies on the ways of reducing stress during pregnancy, together with long-term follow-up studies of the child, are needed.

See also: Anxiety and Anxiety Disorders; Anxiety and Fear Conditioning, Neural Basis of; Autism Spectrum Disorders; Depression; Stress in Childhood: Effects on Development; Stress: Effects on Neonatal and Child Development.

Bibliography

Barker, D.J., 2007. The origins of the developmental origins theory. Journal of Internal Medicine 261, 412–417.

Bergman, K., Sarkar, P., O'connor, T.G., Modi, N., Glover, V., 2007. Maternal stress during pregnancy predicts cognitive ability and fearfulness in infancy. Journal of the American Academy of Child & Adolescent Psychiatry 46, 1454–1463.

Charil, A., Laplante, D.P., Vaillancourt, C., King, S., 2010. Prenatal stress and brain development. Brain Research Reviews 65, 56–79.

Glover, V., 2011. Annual research review: prenatal stress and the origins of psychopathology: an evolutionary perspective. Journal of child psychology and psychiatry, and allied disciplines 52, 356–367.

Gluckman, P.D., Hanson, M.A., Spencer, H.G., 2005. Predictive adaptive responses and human evolution. Trends in Ecology & Evolution 20, 527–533.

Harris, A., Seckl, J., 2011. Glucocorticoids, prenatal stress and the programming of disease. Hormones and Behavior 59, 279–289.

Laplante, D.P., Brunet, A., Schmitz, N., Ciampi, A., King, S., 2008. Project Ice Storm: prenatal maternal stress affects cognitive and linguistic functioning in 5 1/2-year-old children. Journal of the American Academy of Child & Adolescent Psychiatry 47, 1063–1072.

O'connor, T.G., Heron, J., Golding, J., Beveridge, M., Glover, V., 2002. Maternal antenatal anxiety and children's behavioural/emotional problems at 4 years. Report from the Avon Longitudinal Study of Parents and Children. British Journal of Psychiatry 180, 502–508.

O'donnell, K., O'connor, T.G., Glover, V., 2009. Prenatal stress and neurodevelopment of the child: focus on the HPA axis and role of the placenta. Developmental Neuroscience 31, 285–292.

Olds, D.L., Kitzman, H., Cole, R., Robinson, J., Sidora, K., Luckey, D.W., Henderson Jr., C.R., Hanks, C., Bondy, J., Holmberg, J., 2004. Effects of nurse home-visiting on maternal life course and child development: age 6 follow-up results of a randomized trial. Pediatrics 114, 1550–1559.

Rice, F., Harold, G.T., Boivin, J., Van Den Bree, M., Hay, D.F., Thapar, A., 2010. The links between prenatal stress and offspring development and psychopathology: disentangling environmental and inherited influences. Psychological Medicine 40, 335–345.

Talge, N.M., Neal, C., Glover, V., 2007. Antenatal maternal stress and long-term effects on child neurodevelopment: how and why? Journal of Child Psychology and Psychiatry 48, 245–261.

Van Den Bergh, B.R., Mulder, E.J., Mennes, M., Glover, V., 2005. Antenatal maternal anxiety and stress and the neurobehavioural development of the fetus and child: links and possible mechanisms. A review. Neuroscience & Biobehavioral Reviews 29, 237–258.

Weinstock, M., 2001. Alterations induced by gestational stress in brain morphology and behaviour of the offspring. Progress in Neurobiology 65, 427–451.

Weinstock, M., 2007. Gender differences in the effects of prenatal stress on brain development and behaviour. Neurochemical Research 32, 1730–1740.

Relevant Website

beginbeforebirth.org.

Stress in Adolescence: Effects on Development

Nada Krapić, Jasna Hudek-Knežević, and Igor Kardum, University of Rijeka, Rijeka, Croatia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Adolescents frequently have to cope with various stressors that could be potential threats to their healthy development and well-being. Most important mental health consequences of stress in adolescence, such as depression, anxiety, suicide, substance use, and antisocial behavior are reviewed. Physiological mechanisms through which stress exerts its effects on health, as well as some physical health outcomes, such as somatic symptoms, immune changes, and illnesses (cancer, type 1 diabetes, and dermatological conditions) are described. Furthermore, the effects of cognitive appraisal, coping, and social support as mediating and moderating factors between stress and various health outcomes also are explained.

Adolescence is an important period of human development characterized by neurobiological, hormonal, psychological, and social changes. During this developmental period, adolescents have to cope with various common stressors, including physical and sexual changes related to puberty, school life demands, problems with initiating and maintaining friendships and romantic relationships, career choice and beginnings of their working life, and a gradual independence from their families. Additionally, some teenagers have to cope with more unusual stressors that could be related to family (e.g., mental or physical illness, death of a loved one, drug or alcohol abuse, parental separation or divorce), social problems (e.g., poverty, violence), as well as their individual problems (e.g., teenage pregnancy, serious illness, school failure, and various kinds of abuse). Therefore, adolescence often has been thought of as a time of storm and stress, a period of rebellion characterized by conflicts with parents. Contrary to the traditional considerations, however, more recent research findings support the view that storm and stress is not a universal and exclusively biologically based phenomenon (Susman and Rogol, 2004). Contemporary understanding suggests that adolescence is a period when specific types of stressors are more likely to occur than in other developmental periods and that these developmental changes also include adaptation to cultural expectations of becoming an adult. Nonetheless, adolescents frequently have to cope with various stressors that could be potential threats to their healthy development and well-being. For example, cumulative and chronic stressors may lead to increased emotional distress (e.g., depression, anxiety), antisocial behaviors (e.g., aggression, delinquency), risky health behaviors (e.g., substance use, risky driving), and physical health outcomes (e.g., physical symptoms, common colds). Traumatic stressors, such as physical, sexual, and relational victimization, may result in posttraumatic stress symptoms. Furthermore, many studies have found that stressful experiences exert short-term as well as long-term effects on adolescents' healthy development.

Mental Health Consequences of Stress in Adolescence

Depression

Research on recent birth cohorts showed that adolescence is one of the developmental periods during which depression most frequently occurs (Hammen, 2009). Although estimates of childhood depressive episodes for children under the age of 12 years vary from 1 to 3%, several studies found adolescence lifetime prevalence of major depression to be from 9.3 to 24.0% (Merikangas and Knight, 2009), with an additional 11% of adolescents reporting minor depression (Hammen and Rudolph, 2003). Longitudinal studies have found that depression in adolescence might have long-term negative consequences regarding some other mental disorders (e.g., anxiety disorders, substance use) and various negative outcomes (e.g., academic failure, teenage pregnancy, conflicts with family and peers, and suicide). Additionally, it has been found that among those with early onset of depression, very high percentages experience depressive episodes as adults and that early onset of depression is related to a wide range of psychiatric, physical, and functional health outcomes (Abela and Hankin, 2008).

Although the majority of stressful events do not lead to depressive episodes, most depressive episodes are preceded by stressful events. Exposure to stressful events, such as parental death, serious illness, and sexual and physical abuse, as well as less traumatic events and hassles are frequently found to be risk factors for depressive symptoms and disorders in both clinical and community samples of children and adolescents. For example, a review of 60 studies exploring the associations between stressors and psychological problems, including depressive symptoms in children and adolescence showed that stressful experiences predict increases in mental health problems over time (Grant et al., 2004). It has been found that stress and depression are bidirectionally influenced, that is, those with depression histories have greater chance to experience additional distress in the future. A better understanding of the reciprocal influence of stress and depression may

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23031-6

have important implications for the course, treatment, and prevention of depression.

Strong evidence exists for gender differences in depressive disorders in adolescence. Recent meta-analysis reveals a higher prevalence of depression among teenage adolescent girls compared with boys (Costello et al., 2006). One longitudinal birth-cohort study found gender difference in major depressive episodes, with girls having higher rates of depression beginning from age 13 years, with this difference becoming more pronounced by age 15 years (Hankin et al., 1998). Therefore, it seems that the gender difference in depression emerges between ages 12 and 15 years, and by early adulthood women are twice as likely as men to experience depressive episodes. Several biological and psychosocial factors that can explain this difference have been identified. Besides biological risk factors (e.g., dysregulation of the hypothalamic-pituitary-adrenocortical (HPA) axis, genetic factors, and dysregulation of neurotransmitter systems regulating mood), psychosocial risk factors (e.g., negative cognitive style, interpersonal orientation, and negative life stress) also proved to be important (Hilt and Nolen-Hoeksema, 2009). Namely, many studies demonstrated that girls are exposed to greater levels of stressors, as well as that they report more interpersonal stressors compared with boys who report more academic stressors (Hammen, 2009). Girls also more often than boys are the victims of physical and sexual abuse, with both types of abuse being risk factors for the development of depression. Additionally, girls appear to generate more dependent stressors that contribute to their depression, and they are more reactive to stress than boys. Research results consistently show that under high levels of stress, girls have greater tendency to express depressive symptoms than boys.

Anxiety Disorders

Anxiety and depression share important features, but are nevertheless distinguishable phenomena. Although most fears during childhood and adolescence are considered appropriate for the developmental periods and the situations in which they occur, and although anxiety usually is thought of as adaptive because it helps in anticipating threat or harm, its clinical expression usually refers to developmentally inappropriate fears or to those that are developmentally appropriate but that produce excessive distress or dysfunction (Moore et al., 2010). Anxiety disorders usually begin relatively early in life, often manifesting before late adolescence.

Epidemiological data show that the anxiety disorders often are evident in children and adolescents. Overall, lifetime prevalence estimates in children and adolescents for any anxiety disorder range from approximately 10 to 20% (Costello et al., 2004). Specific types of anxiety disorders, such as panic disorder (PD), specific phobias, social phobia, obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), acute stress disorder, separation anxiety disorder (SAD), and generalized anxiety disorder (GAD) occur with different frequency during the period of childhood and adolescence. Epidemiological studies show that GAD is the most common, followed by SAD, social phobia, and the specific phobias. Much less frequent are PTSD, OCD, and PD, with less than 1% prevalence in population samples. Considering gender differences, research shows that girls are affected more, and the overall gender ratio for anxiety disorders throughout childhood and adolescence is ~2:1 (Costello et al., 2004). Untreated anxiety disorders tend to become chronic, with only a minority of adults (12–30%) exhibiting spontaneous remission. Adolescents show somewhat higher rates of recovery, and the majority of early onset anxiety disorders remit by early adulthood.

Early anxiety disorders, however, especially those that develop in late adolescence, represent major risk factors for the development of anxiety disorders in adulthood. Several disorders, such as PD, social phobia, GAD, and OCD, frequently manifest waxing and waning symptoms in which the exacerbation of symptoms may be related to stressful experience (DSM-IVTR; APA, 2000).

Ample evidence indicates that the most prominent factors that contribute to the development and maintenance of anxiety symptoms and disorders are biological predispositions (e.g., family-genetic, neurobiological, and hormonal), but environmental factors should not be underestimated. Namely, although more consistent findings were obtained for depressive disorders, many studies showed that children and adolescents who are exposed to chronic stressors and unpredictable circumstances may be at greater risk for developing anxiety disorders. Research suggest that GAD may result after family relocations, school changes, or other stressful life events (Vasey and Ollendick, 2000), although stressful events related to health and family discord are predictive of increases in anxiety sensitivity (McLaughlin and Hatzenbuehler, 2009). Parenting practices also may lead to anxiety disorders. For example, overprotective parents usually reinforce avoidant and inhibited behavior in their children encouraging them to perceive many situations as new or negative, which in turn lead to their impaired ability to cope and raises their anxiety. Furthermore, a moderate degree of specificity exists in the associations between types of life events and types of anxiety disorders. For example, it has been found that parental separation was related with GAD, while parental death was related with increased risk for phobia (Kendler et al., 1992).

Prevailing theoretical models of the causes of internalizing disorders (depression and anxiety) are variants of diathesis-stress models. The fundamental assumption of a diathesis-stress model is that both the diathesis (also interchangeably referred to as vulnerability) and various environmental conditions (e.g., stressors) are necessary for the development of a disorder and that neither one alone is sufficient. Therefore, it is assumed that internalizing symptoms results from the interaction between different vulnerabilities (e.g., neurobiological, genetic, cognitive and environmental) and stressful events.

Suicide

According to the World Health Organization (2001), suicide is one of the three leading causes of death among people from 15 to 34 years of age in all countries. The results of a meta-analysis of 128 studies from different countries showed that 10% of youths reported a suicide attempt, and 20–30% reported suicidal ideations at some point in their life (Evans et al., 2005). The higher rate of death by suicide is

observed in adolescent boys, while after puberty adolescent girl show higher rates of suicidal ideation and suicide attempts.

The strongest risk factors for attempted and completed suicide in youth are previous suicide attempts, suicidal ideation, and the presence of psychiatric disorders, especially depressive disorder, alcohol and drug abuse, conduct disorder, and aggressive-impulsive behavior (Gould et al., 2003).

Furthermore, environmental stressors also have been recognized as having a role in suicide attempts and suicidal ideation. For example, several studies have found that life events, such as interpersonal loss and legal or disciplinary problems have been linked to completed suicide (Jacobson and Gould, 2009). Childhood physical abuse (e.g., corporal punishment), sexual abuse, and neglect also have been related to suicidal behavior. The younger the person was at the time of the first abuse, the higher the number of suicide attempts reported (Berk et al., 2010).

Bullying also is linked to suicidal behavior among adolescents, especially girls, as indicated by a large epidemiological study showing that bullying perpetrators and their victims, as well as youth who are both victims and perpetrators, have a higher risk for suicidal ideation and suicide attempts (Brunstein-Klomek et al., 2007). Parental conflict is found to be related more frequently to suicide in younger adolescents, whereas romantic relationship difficulties are associated more strongly with suicide in older adolescents (Brent et al., 1999).

Neither psychiatric disorders nor stress alone, although important, are sufficient to trigger suicidal behavior. It is likely that the combination of underlying psychiatric disorders paired with an acute stressor typically leads to suicide.

Substance Use

Adolescence is a period of life characterized by the experimentation with risky behaviors and is a critical developmental period for the onset of substance use and, sometimes, the beginning of the trajectory toward substance dependency. Evidence shows that approximately 50% of adolescents have tried an illicit drug at least once before the end of their high school education, more than 75% have tried alcohol, and approximately 60% of high school seniors report some experience with cigarette smoking, with 19% of them being daily smokers (Johnston et al., 2001, 2002). Substance use and addictive disorders are topics of great importance because of their pervasive negative consequences that may affect the course of adolescent development as well as outcomes in adulthood. For example, accidents, suicides, and homicides account for more than 80% of adolescence deaths, of which at least 50% include drugs and alcohol (Soderstrom and Dearing-Stuck, 1993). Studies also suggest that long-term consequence that is most reliably related to substance use in adolescence is the continued use of that substance. It has been found that adolescence substance use is associated with psychological distress and mental health problems in adulthood.

Research shows many risk factors for drug abuse, each having a different impact depending on the phase of development. These factors can be related to individual characteristics (e.g., genetic susceptibility to addiction, high sensation seeking and impulsivity, alienation, academic failure, and low commitment to school), as well as environmental stressful events (e.g., family problems, problems with peers and community environment).

Studies consistently have found that adolescents who experience high levels of life stress are more likely to use alcohol or drugs and to increase the quantity and the frequency of substance use over time (Chassin et al., 2004). The most frequent explanation of the role of stress in substance use is that adolescents who experience high level of stress and negative affect use alcohol or drugs as a way to decrease them. This is in accord with research findings showing that emotional distress, negative affect, and clinically diagnosed depression are associated with adolescent substance use, although some results show that negative affect might be a result, rather than a cause, of adolescent substance use.

Furthermore, although substance use often is considered as a motivation to reduce negative affect, it also is seen as a way to increase positive affect (Chassin et al., 2004). This distinction in the motive for substance use is important because it has been suggested that adolescents who use substances to maintain or enhance positive affect may use them moderately in quantity and frequency, whereas those who use substances to decrease negative affect may use them more excessively.

Antisocial Behavior

Among several types of mutually intercorrelated antisocial behaviors the most important are conduct disorder, aggression, and delinquency. They are related logically and empirically and, therefore, risk factors that apply to one of them also likely apply to the other two (Farrington, 2004). The prevalence estimates for antisocial behaviors in adolescence are usually very high. For example, the prevalence of conduct disorders in adolescent boys varies from 6 to 16%, and in adolescent girls it ranges from 2 to 9% (Mandel, 1997).

The etiology of antisocial behaviors includes multiple risk factors that do not occur in isolation. They can be grouped into individual, family, and extrafamilial factors (Connor, 2002). Genetic factors and individual differences in temperament are the most important individual factors. Among family and extrafamilial risk factors, chronic and acute stressful life events are predominant. Several family factors found to be strongly associated with antisocial behaviors are severe marital discord, overcrowding or large family size, paternal criminality, maternal psychiatric disorder and the family history of admission of a child into the care of local child protective services. Extrafamilial risk factors for antisocial behaviors include relationship with peers, social deprivation, neighborhood violence, and media violence (Connor, 2002). Except for heritable genetic factors, the majority of risk factors are nonspecific. Namely, they increase the risk for wide variety of general adolescent psychopathology, including different types of antisocial behaviors. Research also suggest that cumulative effects of multiple risk factors exert a much stronger influence on antisocial behavior in children and adolescents than any specific type of risk factors.

Physiological and Physical Health-Related Consequences of Adolescent Stress

When an event has been appraised as stressful, our body triggers a sequence of physiological reactions. Physiological stress response and its control are regulated by the two axes of the neuroendocrine system: HPA and sympathetic-adrenal-medullary (SAM) pathways. The activation of the SAM axis is related strongly to acute emotions, such as fear and anger, and describes the first part of the stress response. During stress, the hypothalamus helps the body to prepare for 'fight or flight' by triggering sympathetic impulses to various organs, stimulating the release of epinephrine, norepinephrine, and other catecholamines in the blood circulation. Catecholamines can cause changes in blood pressure, heart rate, sweating, and pupil dilatation, which result in a feeling of arousal. They also have an effect on a variety of body tissues and can lead to changes in immune function. In addition to the sympathetic activation, stress also triggers changes in the HPA system. At the same time that sympathetic activity increases, the hypothalamus releases corticotropin-releasing hormone, which stimulates the anterior pituitary gland to secrete adrenocorticotropic hormone (ACTH), which increases the adrenal cortex's secretion of cortisol. Cortisol supplies cells with amino acids and extra energy sources and diverts glucose from skeletal muscles to brain tissue. Stress also stimulates the release of glucagon from the pancreas, growth hormone from the anterior pituitary, antidiuretic hormone from the posterior pituitary gland, and renin from the kidneys. Through feedback to the hypothalamus, epinephrine can increase secretions of ACTH to higher levels and thus can maintain the increased activation of HPA axis (Tsigos and Chrousos, 2002). The neurohormonal and immune changes that develop during stress are physiological pathways through which stress can lead to mental and physical health outcomes.

Many factors during adolescence simultaneously contribute to the heightened sensitivity to stressful events, especially neurobiological processes, such as the continued maturation of stress-responsive brain regions, the changes in hormonal reactivity, and the quantity and quality of stressful events. For example, research shows that gender differences in the functioning of stress-related physiological pathways begin during adolescence and persist into adulthood. Available evidence suggests that HPA response in girls differs from those in boys, generally, and is more under the influence of genetic factors and previous environmental exposures. Studies suggest that variants in specific genes, parental depression, and stressful events affect cortisol responses in girls more than those in boys. A greater reactivity to internal and external influences under stress than may be adaptive in many situations also might be related to risks (e.g., increased probability of depression) (Oldehinkel and Bouma, 2011).

Most studies dealing with the impact of stress on adolescent health and psychological well-being have found that adolescents suffer from a large number of body complaints that might be considered as psychosomatic symptoms (e.g., headaches, stomachaches, indigestion, circulatory problems, and nausea). Among those most frequently reported are abdominal pain, chest pain, and headaches (Seiffge-Krenke, 1998). There are indications that daily hassles are better predictors of somatic symptoms than major life events. Among stressful situations that are most important for elevated somatic symptoms in this developmental period are school-related stressors, such as interpersonal problems with peers and teachers, schoolwork pressure, and fear of school failure. Additionally, the transition from primary to secondary school, as well as conflicts with parents, are stressful experiences that in many adolescents increase the frequency and intensity of somatic health symptoms.

Stress also leads to a significant decrease in the function of cellular immunity such as lower NK-cells activity, lower lymphocyte proliferation, and production of gamma interferon (IFN-γ), as well as the increases in herpes virus antibody production. These changes are related to a greater susceptibility to infectious diseases, such as herpes reactivations, infections of the upper respiratory tract, and HIV infections. Exam stress, a stressor typical for adolescents, frequently leads to these immune changes and related infectious diseases (Kiecolt-Glaser et al., 2002).

Research also shows that stressful life events have an impact on chronic illnesses in adolescents. Especially high incidence of stressful life events has been found in children and adolescents with cancer, who most frequently reported the experience of marital separation, death of a close family member, change in residence, and change in health of family members before the onset of cancer (Seiffge-Krenke, 1998). Although these findings could not be considered as evidence of causal relations between stressful events and diseases, and although physiological mechanisms of these associations are not fully understood, it seems that not a single event, but rather the accumulation of several major stressors, could have negative effects on health outcomes.

Although the strongest associations between life events and the onset of chronic illness have been found in cancer patients, the negative impact of life events also was found in patients with other chronic diseases. Studies on diabetic patients also show greater frequency of stressful life events before the onset of diabetes, and it has been found that stress could contribute to the onset of type 1 diabetes through its effects on the immune function (Cox and Gonder-Frederick, 1992). Furthermore, higher levels of diabetes-specific stress as well as general stress are associated with less adherent behavior and subsequently more problems with metabolic control, at least in patients with type 1 diabetes (Farrell et al., 2004).

Stress can be an important factor in the onset, clinical course, and asthma-related mortality. For example, stressful life events such as family dysfunction and patterns of behavior that increase psychological conflict as well as stressful school exams can accelerate the onset and recurrence of asthma after the remission. Furthermore, studies have found that children and adolescents from socially isolated families have more frequent symptoms of asthma, more days of restricted activities because of the symptoms, worse illness management, and more frequent visits made to physicians (Wright et al., 1998). It seems that high levels of ongoing chronic stress plus an acute negative life event could be an especially important risk combination for an asthma attack in children and adolescents. Research highlights immune mechanisms through which psychosocial variables, such as stress, may plausibly affect asthma outcomes. For example, stress has

been linked to inflamed bronchial pathways implicated in asthma (Chen et al., 2006).

Psychosocial stress also has been related to the onset or worsening of various dermatological disorders. It may influence dermatological symptoms in primary stress-reactive skin disorders in which neuroendocrine and psychoneuroimmunologic factors have a crucial role (e.g., psoriasis, atopic dermatitis or eczema, urticaria, acne, and alopecia areata), in disorders characterized by physiological response (e.g., excessive perspiration and blushing), as well as in skin disorders resulting from psychiatric condition such as OCD, PTSD, or major depressive disorder (e.g., dermatitis artefacta, neurotic excoriations, or trichotillomania) (Gupta, 2010).

Stressful events frequently experienced by adolescents, such as exams, have been found to worsen some dermatological diseases (e.g., acne). Additionally, an increase in distress is related with the increase in acne severity. Acne has a peak incidence during adolescence, a period typical for intensive preoccupation with body image and appearance. Therefore, besides the contribution of psychosocial stress to the onset of dermatological disease, the disease itself may result in significant stress for the patient. For example, in vulnerable adolescents, even mild acne could result in severe depression.

Several neurophysiological mechanisms have been implicated in stress-mediated skin disorders. Because stressful emotional experiences disturb the regulation of the HPA and the SAM systems, they may result in the alternations of glucocorticoid levels. For example, studies have found a relation between stress and an imbalance in the epidermal permeability barrier function, which is mediated by the increase in endogenous glucocorticoids. Additionally, higher anxiety in acne patients have been associated with high blood catecholamine levels, which decreased with the treatment of the acne, suggesting that the anxiety in acne patients is associated with the SAM system activation (Gupta, 2010).

Effects of Cognitive Appraisal, Coping, and Social Support on the Experience of Stress

The same stressful event in different individuals may lead to the distress of various intensities. What is extremely stressful in one individual may not be at all stressful in another. Among the most important sources of variation in the effects of stress during adolescence are personality traits, cognitive appraisal, coping with stress, and social support. According to the transactional model of stress and coping, the appraisal of a situation determines the degree of stress experienced by the individual (Lazarus and Folkman, 1984). The model describes two main types of cognitive evaluations of stressful situations, primary and secondary appraisals. Primary appraisal includes the individual's assessment of the stressfulness of the situation. Situations appraised as stressful can involve harm, loss, threat, or challenge. Secondary appraisal includes the individual's assessment of controllability of the stressful episode and the choice of coping strategies. These two types of appraisals are parallel processes that may be automatic and usually occur without conscious awareness. Coping is most often defined as constantly changing cognitive and behavioral efforts to manage specific external or internal demands that are appraised as taxing or exceeding the resources of the person (Lazarus and Folkman, 1984). According to this theory, coping efforts are broadly classified into two types: problem-focused and emotion-focused coping. Problem-focused coping can be directed outward to alter some aspect of the environment or inward to alter some aspect of the self. They include cognitions and behaviors directed toward analyzing and solving a problem and may as well include dividing the problem, seeking information, and considering alternative courses of actions as well as direct action. The efforts directed toward self are various forms of reappraisals, such as changing the meaning of the situation or event, reducing ego involvement, or recognizing the existence of personal resources or strengths. Emotion-focused coping strategies are directed toward decreasing emotional distress. They include various ways of coping usually described as expression and suppression of emotions as well as seeking emotional support from others. Besides problem- and emotion-focused coping, avoidance coping also is included in some taxonomies of coping with stress. It refers to cognitive or behavioral avoidance or denial of the problem.

During puberty and adolescence, the differentiation of coping resources, skills, and strategies oriented toward problem solving are under the influence of cognitive development and social experience. Adolescence is a period in human life characterized by a relatively quick development and the use of a broader range of coping strategies, as well as the tendency to abandon ineffective coping. For example, between 13 and 15 years of age, a significant decrease of avoidance coping has been noted. Also, it has been observed that older adolescents use wishful thinking and distancing less than before. Adolescence is also a critical period for the development of some ineffective coping strategies, especially social withdrawal and alcohol and substance use (Aldwin, 2007). Regarding gender differences, the most consistent result found in the literature is that girls use emotion-focused coping more often than boys. To the lesser extent, this is also true for problem-focused coping. This usually is explained by their faster maturation and the experience of higher levels of stress during early adolescence when compared with boys. Also, it has been found that boys use drugs, alcohol, and repressive coping more often than girls.

Among major antecedents of coping with stress are individual differences in temperament and other personality traits as well as in self-concept. These factors influence the availability of coping strategies in stressful situations as well as the appraisal of the stressfulness of events. For example, the majority of research shows that the personality trait of extraversion is related to active and more efficacious coping strategies, such as problem-focused coping, whereas neuroticism is related to passive and less efficacious coping strategies, such as avoidance. Neuroticism is related to higher exposure and reactivity to stressful events and also is considered to be a disposition for stressful experience even in the absence of stressors. Furthermore, adolescents who are high in self-esteem usually adopt problem-focused coping strategies, whereas adolescents with low self-esteem often adopt avoidant and emotion-focused coping strategies.

Coping strategies are not defined as generally effective or ineffective, but rather their effectiveness may depend on the type of situation, specific combinations between different coping strategies, and timing in their use during the course of a stressful transaction. For example, avoidance strategies are more effective in reducing emotional distress in the short term, and problem-focused strategies are more effective over the long term. When the stress is perceived as controllable, problem-focused coping strategies are associated with fewer psychological symptoms, whereas in uncontrollable stressful situations, emotion-focused coping is related to fewer symptoms.

Considering various health consequences of coping with stress, numerous studies show that some coping strategies more often are related to positive, whereas others to negative, health outcomes. For example, longitudinal research on adolescents reveals that the use of avoidance and emotion-focused coping have been linked with more intensive depressive and anxiety symptoms, whereas the use of problem-focused coping has been associated negatively with depression and hopelessness. Furthermore, problem-focused coping has been found to correlate negatively with health problems and health risk behaviors, whereas avoidance coping correlated positively with these domains (Wilson et al., 2005). Generally, it could be concluded that when adolescents lack adequate coping skills, there is a risk from various negative consequences, including psychological distress and suicidal attempts, smoking and substance use, and high-risk sexual behaviors (Rew, 2005).

Social support is considered to be one of the most important psychosocial factors that influence the relationship between stress and health outcomes. It has been defined broadly as interpersonal social resources that involve either the presence or the implication of stable human relationships. Social support is regarded as structural (e.g., number of social contacts, size, and density of social networks) or functional (e.g., informative, emotional, instrumental). Evidence suggests that social support generally contributes to better adjustment of adolescents and that their low satisfaction with social support is associated with anxiety, depression, psychosomatic symptoms, and interpersonal sensitivity.

The protective roles of social support on health outcome are explained by the two models. The main-effect model states that social support has independent effects on mental and physical health outcomes regardless of the level of stress experienced, while the stress-buffering model posits that social support has indirect effects on health outcomes by acting as a buffer against stress (Cohen and Wills, 1985).

Among adolescents, the social support of family and peers has great significance. Parental support has been found to be a predictor of better adjustment and less distress in adolescence. The presence of familial social support has been associated with higher levels of resilience. On the other hand, different studies show that two characteristics of a risky family social environment – conflict and aggression and a cold, unsupportive, or neglectful home – have adverse developmental effects on adolescents. It has been shown that parental support, involvement, and monitoring reduce the probability that adolescents will engage in a variety of health-compromising behaviors. Furthermore, longitudinal studies have found that lack of support and deficient nurturing during childhood leads to higher rates of illness and physical complaints several years later (Repetti et al., 2002).

Without the support of a peer group, adolescents experience loneliness and a sense of lack of belonging and acceptance, which, in turn, can contribute to frustration and feelings of rejection and emergence of antisocial behavior. Along with positive effects of the peer support, adolescence is also a period during which relationships with peers can undermine healthy development. For example, drug use and risky sexual behavior may be influenced greatly by peers. Regarding drug use, adolescents who become involved with substance use usually have friends who do the same.

Social support most effectively acts as a buffer against stress when it matches the needs elicited by a stressful event. Even social relationships that are close and supportive are not uniformly positive and may have negative consequences. For example, being close to and having highly supportive antisocial parents might result in various types of problematic behaviors of adolescents, such as substance abuse or conduct disorder (Rook, 2003). Also, it has been found that diabetic adolescents who turn to their peers for coping strategies may have worse metabolic control than those who turn to their parents (Aldwin, 2007).

Cross-Cultural Differences in the Effects of Stress on the Development of Adolescents

As mentioned, developmental changes during adolescence, among other factors, also include adaptation to cultural expectations of becoming an adult. The greatest part of our knowledge about the effects of stress on development in adolescence was drawn from the research on the samples from Western societies. Theoretical assumptions as well as empirical evidence suggest the existence of cultural differences in the effects of stress on the adolescent development. Cultures differ in the norms, beliefs, and values that provide prescriptions for behavior and therefore have a different impact on the stress process. Aldwin (2007) stated that culture can affect the stress and coping process in four ways. It shapes the types of stressors that a person is likely to experience, affects the appraisal of the stressfulness of a given situation, affects the choice of coping strategies, and provides different institutional mechanisms by which a person can cope with stress. Furthermore, because of the intercultural exchange that has increased markedly over the past several decades throughout the world, the problem of acculturative stress has become highly pronounced. Also, it has been emphasized that stress increases with rapid social change, most often in the form of westernization and a change in values (e.g., individualism, competition, and consumerism) (Dasen, 2000).

Considering ethnic differences in the stress consequences, Dornbusch et al. (1991) have found significant relationships between stressful life events and symptoms of depression among Asian and African American, Latino, and white adolescents, but the relationship was the weakest for African Americans. Similarly, Gerard and Buehler (2004) have found that the relationship between risk factors and depression in adolescence was moderated by racial–ethnic group membership and that this relationship was stronger

for whites compared with both blacks and Latinos. Also, rates of completed as well as attempted suicide vary according to ethnicity. For example, the elevated rates of suicide among Native American youth are supposed to be associated with low social integration that results in limited access to various resources. On the other hand, lower rates of suicide among African American youth have been associated with their higher levels of religiosity compared with white youth. Highest levels of suicidal ideation and attempts are found in Hispanic females and can be attributed to a combination of cultural factors, including familism, authoritarian parenting style, and conflicts with caregivers over the issue of individuality (Jacobson and Gould, 2009). Comparing adolescent coping behaviors from seven different European nations, Gelhaar et al. (2007) found the highest cultural differences for coping with job-related problems, while similarities were found for coping with self- and future-related problems. Cross-cultural differences in quality and availability of social support also should be taken into account. Namely, despite social change, stress of adolescents is reduced more efficiently in those societies that manage to maintain a strong cultural identity and a part of their value system, such as family solidarity.

Conclusion

Although studies on the effects of stress on development of adolescence are numerous, more longitudinal and integrative research is needed to broaden our understanding of how complex relations between stress and mediating and moderating processes (personality, cognitive appraisal, coping, social support) exert their effects on (un)healthy development of adolescents. Also, the problem of the lack of comprehensive research exploring the factors associated with ethnicity that influence the relationship between stress and adolescent health is becoming more and more pronounced because the current intercultural exchange has considerably increased worldwide.

It is well known that stressful experience can lead to long-term changes in the body by destabilizing different physiological pathways, which often result in various mental and physical health outcomes. Therefore, knowledge of the deleterious effects of stress on adolescent health outcomes may be of help in planning and organizing prevention and treatment programs and thus improve adolescents' quality of life.

See also: Adolescence, Sociology of; Adolescent Health and Health Behaviors; Adolescent Sexual Risk; Alcohol Use among Young People; Alcohol Use and Abuse; Antisocial Behavior; Anxiety and Anxiety Disorders; Coping across the Lifespan; Depression, Pessimism, and Health; Depression; Suicide, Sociology of; Suicide.

Bibliography

Abela, J.R.Z., Hankin, B.L., 2008. Depression in children and adolescents: causes, treatment, and prevention. In: Abela, J.R.Z., Hankin, B.L. (Eds.), Handbook of Depression in Children and Adolescents, first ed. The Guilford Press, New York, pp. 3–5.

Aldwin, C., 2007. Stress, Coping, and Development: An Integrative Perspective, second ed. The Guilford Press, New York.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed. American Psychiatric Association, Washington, DC.

Berk, M., Suddath, R., Devich-Navarro, M., 2010. Adolescent suicide. In: Fink, G. (Ed.), Stress Consequences: Mental, Neuropsychological and Socioeconomic, first ed. Elsevier, Oxford, pp. 316–317.

Brent, D.A., Baugher, M., Bridge, J., Chen, T., Chiappetta, L., 1999. Age- and sex-related risk factors for adolescent suicide. Journal of the American Academy of Child and Adolescent Psychiatry 38, 1497–1505.

Brunstein-Klomek, A., Marrocco, F., Kleinman, M., Schonfeld, I., Gould, M., 2007. Bullying, depression, and suicidal ideation in adolescents. Journal of the American Academy of Child and Adolescent Psychiatry 46, 40–49.

Chassin, L., Hussong, A., Barrera Jr., M., Molina, B.S.G., Trim, R., Ritter, J., 2004. Adolescent substance use. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. John Wiley and Sons, New Jersey, pp. 665–696.

Chen, E., Hanson, M.D., Patterson, L.Q., Griffin, M.J., Walker, H.A., Miller, G.E., 2006. Socioeconomic status and inflammatory processes in childhood asthma: the role of psychological stress. Journal of Allergy and Clinical Immunology 117, 1014–1020.

Cohen, S., Wills, T.A., 1985. Stress, social support, and the buffering hypothesis. Psychologica Bulletin 98, 310–357.

Connor, D.F., 2002. Aggression and Antisocial Behavior in Children and Adolescents. The Guilford Press, New York.

Costello, E.J., Egger, H.L., Angold, A., 2004. Developmental epidemiology of anxiety disorders. In: Ollendick, T.H., March, J.S. (Eds.), Phobic and Anxiety Disorders in Children and Adolescents: A Clinician's Guide to Effective Psychosocial and Pharmacological Interventions, first ed. Oxford University Press, New York, pp. 61–91.

Costello, E.J., Erkanli, A., Angold, A., 2006. Is there an epidemic of child or adolescent depression? Journal of Child Psychology and Psychiatry and Applied Disciplines 47, 1263–1271.

Cox, D.J., Gonder-Frederick, L., 1992. Major developments in behavioral diabetes research. Journal of Consulting and Clinical Psychology 60, 628–638.

Dasen, P.R., 2000. Rapid social change and the turmoil of adolescence: a cross-cultural perspective. International Journal of Group Tensions 29, 17–49.

Dornbusch, S.M., Mont-Reynaud, R., Ritter, P.L., Chen, Z., Steinberg, L., 1991. Stressful events and their correlates among adolescents of diverse backgrounds. In: Colton, M.E., Gore, S. (Eds.), Adolescent Stress: Causes and Consequences, first ed. Aldine de Gruyter, Hawthorne, NY, pp. 111–130.

Evans, E., Hawton, K., Rodham, K., 2005. Suicidal phenomena and abuse in adolescents: a review of epidemiological studies. Child Abuse and Neglect 29, 45–58.

Farrell, S.P., Hains, A.A., Davies, W.H., Smith, P., Parton, E., 2004. The impact of cognitive distortions, stress, and adherence on metabolic control in youths with type 1 diabetes. Journal of Adolescent Health 34, 461–467.

Farrington, D.P., 2004. Conduct disorders, aggression, and delinquency. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. John Wiley and Sons, NJ, pp. 627–664.

Gelhaar, T., Seiffge-Krenke, I., Borge, A., Cicognani, E., Cunha, M., Loncaric, D., Macek, P., Steinhausen, H.-C., Winkler Metzke, C., 2007. Adolescent coping with everyday stressors: a seven-nation study of youth from central, eastern, southern, and northern Europe. European Journal of Developmental Psychology 4 (2), 129–156.

Gerard, J.M., Buehler, C., 2004. Cumulative environmental risk and youth maladjustment: the role of youth attributes. Child Development 75 (6), 1832–1849.

Gould, M., Greenberg, T., Velting, D.M., Shaffer, D., 2003. Youth suicide risk and preventive interventions: a review of the past 10 years. Journal of the American Academy of Child and Adolescent Psychiatry 42, 386–405.

Grant, K.E., Compas, B.E., Thurm, A.E., McMahon, S.D., Gipson, P.Y., 2004. Stressors and child and adolescent psychopathology: measurement issues and prospective effects. Journal of Clinical Child and Adolescent Psychology 33, 412–425.

Gupta, M.A., 2010. Dermatological conditions. In: Fink, G. (Ed.), Stress Consequences: Mental, Neuropsychological and Socioeconomic, first ed. Elsevier, Oxford, pp. 420–424.

Hammen, C., 2009. Stress exposure and stress generation in adolescent depression. In: Nolen-Hoeksema, S., Hilt, L.M. (Eds.), Handbook of Adolescent Depression, first ed. Routledge, New York, pp. 305–333.

Hammen, C., Rudolph, K.D., 2003. Childhood mood disorders. In: Mash, E.J., Barkley, R.A. (Eds.), Child Psychopathology, second ed. Guilford Press, New York, pp. 233–278.

Hankin, B.L., Abramson, L.Y., Moffitt, T.E., McGee, R., Silva, P., Angell, K.E., 1998. Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. Journal of Abnormal Psychology 107, 128–140.

Hilt, L.M., Nolen-Hoeksema, S., 2009. The emergence of gender differences in depression in adolescence. In: Nolen-Hoeksema, S., Hilt, L.M. (Eds.), Handbook of Depression in Adolescents, first ed. Routledge, New York, pp. 111–135.

Jacobson, C.M., Gould, M., 2009. Suicide and nonsuicidal self-injurious behaviors among youth: risk and protective factors. In: Nolen-Hoeksema, S., Hilt, L.M. (Eds.), Handbook of Depression in Adolescents, first ed. Routledge, New York, pp. 207–235.

Johnston, L.D., O'Malley, P.M., Bachman, J.G., 2001. Monitoring the Future National Survey Results on Drug Use, 1975–2000. In: Secondary School Students (NIH Publication No. 02-5106), vol. I. United States Department of Health and Human Services, National Institute on Drug Abuse, Bethesda, MD.

Johnston, L.D., O'Malley, P.M., Bachman, J.G., 2002. Monitoring the Future National Survey Results on Drug Use, 1975–2001. In: College Students and Adults Ages 19–40, vol. II. United States Department of Health and Human Services, National Institute on Drug Abuse, Bethesda, MD (NIH Publication No. 02–5107).

Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C., Eaves, L.J., 1992. Childhood parental loss and adult psychopathology in women: a twin study perspective. Archives of General Psychiatry 49 (2), 109–116.

Kiecolt-Glaser, J.K., McGuire, L., Robles, T.F., Glaser, R., 2002. Psychoneuroimmunology: psychological influences on immune function and health. Journal of Consulting and Clinical Psychology 70, 537–547.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Mandel, H.P., 1997. Conduct Disorder and Underachievement: Risk Factors, Assessment, Treatment, and Prevention. Wiley, New York.

McLaughlin, K.A., Hatzenbuehler, M.L., 2009. Stressful life events, anxiety sensitivity, and internalizing symptoms in adolescents. Journal of Abnormal Psychology 118 (3), 659–669.

Merikangas, K.R., Knight, E., 2009. The epidemiology of depression in adolescents. In: Nolen-Hoeksema, S., Hilt, L.M. (Eds.), Handbook of Depression in Adolescents, first ed. Routledge, New York, pp. 53–74.

Moore, P.S., March, J.S., Albano, A.M., Thienemann, M., 2010. Anxiety disorders in children and adolescents. In: Stein, D.J., Hollander, E., Rothbaum, B.O. (Eds.), Textbook of Anxiety Disorders, second ed. American Psychiatric Press, Washington, DC, pp. 629–651.

Oldehinkel, A.J., Bouma, E.M.C., 2011. Sensitivity to the depressogenic effect of stress and HPA-axis reactivity in adolescence: a review of gender differences. Neuroscience and Biobehavioral Reviews 35 (8), 1757–1770.

Repetti, R.L., Taylor, S.E., Seeman, T.E., 2002. Risky families: family social environments and the mental and physical health of offspring. Psychological Bulletin 128 (2), 330–366.

Rew, L., 2005. Adolescent Health: A Multidisciplinary Approach to Theory, Research, and Intervention. Sage, Thousand Oaks, CA.

Rook, K.S., 2003. Exposure and reactivity to negative social exchanges: a preliminary investigation using daily diary data. Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 58, 100–111.

Seiffge-Krenke, I., 1998. Adolescents' Health: A Developmental Perspective. Lawrence Erlbaum, Mahwah, NJ.

Soderstrom, C.A., Dearing-Stuck, B.A., 1993. Substance misuse and trauma: clinical issues and injury prevention. Adolescent Medicine 4, 423–438.

Susman, E., Rogol, A., 2004. Puberty and psychological development. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, second ed. John Wiley and Sons, NJ, pp. 15–44.

Tsigos, C., Chrousos, G.P., 2002. Hypothalamic-pituitary-adrenal axis, neuroendocrine factors and stress. Journal of Psychosomatic Research 53 (4), 865–871.

Vasey, M.W., Ollendick, T.H., 2000. Anxiety. In: Sameroff, A.J., Lewis, M., Miller, S.M. (Eds.), Handbook of Developmental Psychopathology, second ed. Plenum, New York, pp. 511–529.

Wilson, G.S., Pritchard, M.E., Revalee, B., 2005. Individual differences in adolescent health symptoms: the effects of gender and coping. Journal of Adolescence 28 (3), 369–379.

World Health Organization, 2001. The World Health Report 2001. WHO, Geneva.

Wright, R.J., Rodriguez, M., Cohen, S., 1998. Review of psychosocial stress and asthma: an integrated biopsychosocial approach. Thorax 53 (12), 1066–1074.

Stress in Childhood: Effects on Development

Steven G Little, Walden University, Minneapolis, MN, USA
Angeleque Akin-Little, Akin-Little and Little Behavioral Psychology Consultants, Malone, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Stressful events can occur at any point in child and adolescent development. Events that can cause stress include parental divorce, sexual or physical abuse, severe accidents, cancer or other life-threatening illness, natural or man-made disasters, war, terrorism, or the sudden death of a parent, sibling, or peer. However, not every child exposed to a traumatic event develops stress-related symptoms. Many children demonstrate resiliency, the ability to thrive and excel even when exposed to severe stressors. This article discusses the types of stressors that may impact upon children throughout the world, details the effect of stress on development, and discusses other factors that may either exacerbate or minimize the effect of stress on development.

Unfortunately, stressful events are not an unusual occurrence in child and adolescent development. Parents divorce, grandparents die, children and others in their lives get injured. These may not be considered extreme trauma, however. According to the *Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition, Text Revision* (*DSM-IV-TR*; American Psychiatric Association, 2000), an extreme traumatic stressor is one that involves

> direct personal experience of an event that involves actual or threatened death or serious injury, or other threat to one's physical integrity; or witnessing an event that involves death, injury, or threat to the physical integrity of another person; or learning about unexpected or violent death, serious harm, or threat of death or injury experienced by a family member or other close associate. (p. 463)

Examples include sexual or physical abuse, severe accidents, cancer or other life-threatening illness, natural or man-made disasters, war, terrorism, or the sudden death of a parent, sibling, or peer (Cohen et al., 2006a). However, it should also be noted that not every child exposed to a traumatic event will develop stress-related symptoms. Many children demonstrate resiliency, the ability to thrive and excel even when exposed to severe stressors (Leckman and Mayes, 2007). It is the intent of this article to discuss the types of stressors that may impact upon children throughout the world, detail their effects on development, and discuss other factors that may either exacerbate or minimize their effects on development.

Incidence of Stress-Inducing Incidents in Children

Children and adolescents may encounter a number of different experiences which may induce a stress response. These can include direct experience of abuse (e.g., sexual, physical), grief, domestic and community violence, wars, terrorism, natural disasters, or a combination of these (Little et al., 2009). In a retrospective study of over 17 000 adults, Felitti et al. (1998) reported that more than one half of their sample had experienced at least one adverse event in childhood with approximately one quarter reporting having experienced two or more. Clearly, exposure to possible stress-inducing events is not uncommon in childhood. Increasing numbers of children and adolescents have also been victims of natural disasters. For example, in 2011 there were 63 earthquakes worldwide of 6.0 magnitude or greater (United States Geological Survey, 2012). This includes a magnitude 9.0 earthquake off the coast of Honshu, Japan which resulted in at least 15 550 people killed, 5344 missing, 5314 injured, 130 927 displaced, and at least 332 395 buildings, 2126 roads, 56 bridges, and 26 railways destroyed or damaged by the earthquake and resultant tsunami (United States Geological Survey, 2012). The past decade has also seen earthquake-induced tsunamis that have caused devastation in countries bordering the Indian Ocean and, more recently, in Chile and Samoa and hurricane-related devastation in many islands of the Caribbean as well as the central Gulf Coast of the United States (i.e., New Orleans), Florida, and the US northeast (i.e., New Jersey and New York). In addition, there have been major floods on virtually every continent and destruction caused by cyclones and other natural disaster throughout the world. While there are similarities across natural disasters, each is unique in some way, which, in turn, may influence the type of reaction of the victims (Evans and Ohlner-Stinnett, 2006). One thing is clear, however, natural disasters are a common source of stress in childhood, resulting from a loss of property, displacement, and disruption of the normal routine (e.g., attending school).

Child sexual abuse is another source of trauma and stress in children. A 1998 meta-analysis found that in the United States the incidence of sexual abuse ranged from 3 to 37% for boys and 8–71% for girls with an average of 17% for boys and 28% for girls (Rind et al., 1998). In addition, the US Department of Education estimates 9.6% of students are targets of educator sexual misconduct sometime during their school career (Shakeshaft, 2004). In the United Kingdom, it is estimated that 11% of boys and 21% of girls under 16 have experienced sexual abuse (Cawson et al., 2000) and in New Zealand it is estimated that 14% of boys and 26% of girls have experienced unwanted sexual experiences (Fleming et al., 2007). Data from the rest of world are not as clear but one study from India indicated 53% of respondents reported having faced one or more forms of sexual abuse (Ministry of Women and Child Development, Government of India, 2007). As high as these numbers are, many believe these numbers are underestimates (Freyd et al., 2005).

Data from the UN Secretary-General's Study on Violence Against Children (2006) indicate that almost 53 000 children died worldwide in 2002 as a result of homicide and up to 80–98% of children worldwide suffer physical punishment in their homes, with a third or more experiencing severe physical punishment resulting from the use of implements. In addition, 2002 data indicate that 150 million girls and 73 million boys under 18 experienced forced sexual intercourse or other forms of sexual violence during 2002 and that between 100 and 140 million girls and women in the world have undergone some form of female genital mutilation/cutting. Furthermore, in 2004, 218 million children were involved in child labor, of whom 126 million were in hazardous work, and estimates suggest that 1.8 million children were forced into prostitution and pornography. School violence (e.g., school shootings, bullying) is also not uncommon in western or Asian cultures (Chen and Astor, 2009).

There are also wars and terrorist activity causing havoc in the lives of children and families in much of the world. According to data compiled by UNICEF (2007), over the last decade 20 million children have had to flee their home because of war, with one million children orphaned or separated from family. War also contributed to the deaths of two million children with six million children being permanently disabled or seriously injured. As many as 300 000 children have been forced to be soldiers in more than 30 conflicts worldwide.

An even more common source of stress for children is grief. Studies have found that approximately 4% of American children experience the death of a parent before the age of 18 (Social Security Administration, 2000), with the number of parental deaths even higher in other countries. The death of a loved one can be one of the most severe stress-inducing events a child may encounter (Ayyahh-Abdo, 2001). Each year 7000–12 000 American children and adolescents are affected by suicide of a parent (Wilcox et al., 2010). In addition, with the death rate of children in the United States from birth to 17 being 62.4 per 100 000 (CDC, 2007), having a close friend or sibling die is not uncommon. Notably, these data do not include the number of children who experience the death of a grandparent or someone else close to them. Unfortunately, the data are clear that experiencing the death of someone close, and the resulting stress, is not uncommon in childhood.

Another source of stress in childhood is bullying. Bullying can take the form of overt or covert behaviors intended to physically harm or ostracize a particular student (Little et al., 2011a). While schools are reporting a lower rate of student bullying in the recent past, in 2005, 28% of students reported having been bullied during the last 6 months (Dinkes et al., 2007). Additionally, the last few years have seen an increase in the occurrence of cyber-bullying. There are reports of 20–40% of middle- and high-school students being bullied at least one time online (Stover, 2006).

The data are unequivocal that exposure to stress-inducing traumatic events is not uncommon in childhood and adolescence. Furthermore, this exposure appears to be increasing particularly in the areas of natural disaster, wars, and terrorism. In addition, some children experience stress in the absence of clear environmental determinants. It is the intent of the remainder of this article to discuss the psychobiology of stress, the effects of stress on development, the concept of resiliency, and approaches that adults can take to minimize adverse effects of stress in children.

Psychobiology of Stress

In the literature, the term 'stress' has been used inconsistently as either a stimulus (stressor), a response to a stimulus, or the physiological response to that stimulus (Kemeny, 2003). Kemeny also reports there are a number of factors that can influence the outcome of exposure to a stressor. These include whether the individual has perceived control over the stressor, the ambiguity of the stressor, and the level of demand placed on the individual as well as its duration and novelty. In addition, psychological stress in children might be caused by negative family circumstances such as parenting stress, experiences of negative life events, parental unemployment, or low socioeconomic status, as children can sense stress that is being experienced by their parents (Essex et al., 2002).

Stress and the resultant anxiety are natural human reactions that help an individual respond to danger or threat. When the body reacts to a stressful event, the individual experiences physical sensations like rapid heartbeat, changes in respiration, or sweaty or shaky hands and feet. These physical reactions, often called the fight–flight response, are caused by a rush of adrenaline and other stress hormones that prepare the body to make a quick retreat or 'flight' from danger. This physiological response happens instantly, but it usually takes slightly longer for the cortex to process the situation, evaluate the threat, and determine an appropriate response. When the cortex sends the all-clear signal, the fight–flight response is deactivated and the nervous system starts to calm down. Chronic or severe threat situations such as those caused by the environmental situations discussed earlier may perpetuate the physiological and cognitive response, however.

From a psychobiological perspective, the focus is on the functioning of the hypothalamic–pituitary–adrenal (HPA) axis stress–response system (Cicchetti, 2002; Sturge-Apple et al., 2012; Susman, 2006). The HPA axis serves as a primary means for organizing the body's response to environmental threat and stress. The outcome of HPA activation is the glucocorticoid (GC) hormone cortisol. This increase in cortisol levels increases cognitive processing of significant events, as well as organizing energy and physiological resources toward responding to the perceived stressor (Gunnar and Quevedo, 2007). Exposure to stressors and the resultant physiological stress response have been associated with a variety of negative health consequences including pain, sleep deficiencies, and adverse effects on the cardiovascular, endocrine, and immune systems (Maunder and Hunter, 2001; Uchino et al., 1996). For example, stress resulting from psychosocial strain in the family, has been found to be involved in the induction, or progression, of diabetes-related autoimmunity in the child during the first year of life (Sepa et al., 2005). In addition, De Bellis (2001) proposed that maltreatment in childhood and its resultant stress may be more detrimental than trauma experienced in adulthood because stress in childhood may cause delays in behavioral, cognitive, and emotional

regulation. The following section will elaborate on these developmental delays.

Stress and Development

Stress has been identified as a factor affecting both behavioral and physiological development (Bremner and Vermetten, 2001). In addition, Bremner and Vermetten report that repeated exposure to stressors early in life "has effects on stress-responsive neurobiological systems that persist throughout life" (p. 474), including dysregulation of the HPA axis. They further hypothesize that stress-induced changes in neurobiology underlie the development of psychopathology in those who do develop psychiatric symptoms.

Cohen et al. (2006a) categorize trauma symptoms into affective, behavioral, cognitive, and physical domains. Fear, depression, anger, and frequent mood changes are common affective symptoms with behavioral symptoms centering on the avoidance reminders of the traumatic experiences. Cognitive symptoms usually involve distorted cognitions the child may have about themselves, others, the event, or the world (e.g., 'the event is my fault'). There is not always a rational explanation for why bad things happen so it is common for irrational beliefs to occur. This coping mechanism appears to help the child develop a sense of control or predictability over their environment. Unfortunately, the most common irrational belief in children is blaming themselves (Cohen et al., 2006a). Physical symptoms are related to the chronic stress and can include elevated pulse rates and blood pressure, increased physical/muscle tension, and hypervigilance.

Biologically, early stress has been found to have an effect on the hippocampus, the area of the brain important in learning and memory. Neuropsychological studies in adults have shown that individuals with posttraumatic stress disorder (PTSD) have impaired attention, working memory, and learning (Vasterling et al., 2002). Studies also suggest heightened processing of trauma-related information, but weaker processing of neutral information (Karl et al., 2006). While the totality of studies with children has resulted in equivocal findings with regard to hippocampal volume, Tupler and De Bellis (2006) found that the total volume of the hippocampus was positively correlated with age and the severity of psychological symptoms of trauma. Although studies on children and adolescents have produced equivocal results with regard to hippocampal volume, biochemical studies have consistently identified an increase in urine cortisol and salivary cortisol levels in children with PTSD or abused children with PTSD symptoms (Carrion et al., 2002; De Bellis et al., 1999). Adult patients with PTSD often report a wide range of cognitive problems in memory, concentration, attention, planning, and judgment (Moradi et al., 1999). However, little such research exists with children. This led Moradi and colleagues to examine cognitive factors in children and adolescents with PTSD. Samuelson et al. (2010) found that children and adolescents diagnosed with PTSD showed poorer overall memory performance than controls, slower and less effective learning, heightened sensitivity to interference, and impaired effect of rehearsal on memory acquisition on a word list learning task.

Resiliency and Posttraumatic Growth

Not every child exposed to a traumatic stressor will develop trauma symptoms. Many children demonstrate resiliency and the ability to thrive and excel even when exposed to severe stressors (Leckman and Mayes, 2007). Posttraumatic growth (PTG) has been defined as positive psychological change experienced as a result of the struggle with highly challenging life circumstances (Tedeschi and Calhoun, 2004). It should also be noted that PTG is not simply a return to baseline after experiencing trauma, it is an experience of improvement that for some persons is significant (Tedeschi and Calhoun, 2004).

Calhoun and Tedeschi (2006) identify five domains of PTG: personal strength, new possibilities, relating to others, appreciation of life, and spiritual change. Personal strength refers to the realization people have that they may be stronger than they imagined. In other words, he/she may think, "I survived a major life challenge and I am therefore quite strong." New possibilities are defined as developing new interests, sometimes related to the trauma they experienced. For example, a child who was physically or sexually abused may become a psychologist or social worker in order to help others who have similar experiences. Relating to others refers to a greater connection to other people and increased compassion for others who may be suffering (Calhoun and Tedeschi, 2006). Especially when death or injury is part of the traumatic experience, individuals may develop a new appreciation of life and a reexamination of priorities. Finally, some individuals report deeper spiritual and existential meaning in their lives although this is more likely to happen with older adolescents and adults (Calhoun and Tedeschi, 2006).

Hefferon et al. (2009) investigated PTG and illness-related stress. They argue for the need to extend PTG research from the current five main areas mentioned earlier to include positive benefits unique to illness. They suggest a sixth area of growth: "a new awareness of the body by way of a diagnosis of life threatening physical illness" (p. 373). They report that illness-related survivors take greater responsibility for their health, engage in healthier behaviors, and reduce participation in risky behaviors. One difference they found between the respondents recovering from illness and those found in other PTG studies was that many of the respondents recovering from illness believed their circumstances to be "a gift which added value to their lives" (p. 374) while in the majority of other PTG studies, such as with bereavement, the participants would change the circumstances if it were possible.

Cryder et al. (2006) studied children who experienced Hurricane Floyd in 1999. Results indicated that there was a significant correlation between indicators of PTG and the children's competency beliefs. Although both social support and ruminative thinking correlated with positive competency beliefs, they did not correlate directly with PTG or with each other. Unlike previous research, the results of this study found that trauma severity was not associated with PTG or the other measured constructs.

Wong et al. (2009) examined the long-term effects of parental cancer in children through retrospective interviews with adults. Results indicated that 44% of participants spontaneously reported PTG when asked to describe how

their parent's illness affected their childhood. Three of Calhoun and Tedeski (2006) five areas of PTG were reported by the participants: increased appreciation for life, stronger personal relationships, and the possibility of taking new paths in life. Additionally, Wong et al. found that PTG was reported by participants regardless of whether their parent survived their illness, indicating that the threat of a parent dying may be sufficient to bring about PTG in children.

Salter and Stallard (2004) examined PTG among a population of children (aged 7–18 years) who had been involved in a road traffic accident. Results indicated that 42% of the children in the study reported some evidence of PTG following the road traffic accident although 37% were also identified as experiencing PTSD. The majority of reports of PTG came from the older children in the sample. Salter and Stallard found very limited support for positive growth in the areas of perception of self and interpersonal relationships. However, the reports from the participants did indicate an increased appreciation for life.

Alisic et al. (2008) examined the relationship between posttraumatic stress reactions, PTG, and quality of life in a large sample ($n = 1770$) of primary school children. Results indicated that posttraumatic stress reactions and PTG were positively correlated, supporting previous research which suggests that the two constructs can coexist (e.g., Salter and Stallard, 2004; Tang, 2007). There was a significant negative correlation between posttraumatic stress reactions and quality of life, and a significant but small positive correlation between PTG and quality of life.

Tang (2007) examined the prevalence and predictors of PTG among Thai adult survivors of the 2004 Southeast Asian earthquake-tsunami. Tang emphasized that worldviews in Asian cultures often significantly differ from those in western cultures and asserted that depending on one's culture, a traumatic event may indeed confirm these worldviews rather than violate them. The participants in the study were adult Thai survivors with mild physical injuries. The majority of the sample population witnessed others being seriously injured or killed, and reported friends or family members being injured, missing, or killed. Results indicated that 88% of the participants experienced varying levels of PTG and that PTG and PTSD were not mutually exclusive. Factors such as disaster-related cognitive emotional states (intrusive, hyperaroused, and avoidant states), perceived emotional familial support, and being married predicted both inter- and intrapersonal PTG. While Tang did not specifically examine children's response, clearly, cultural considerations are important.

Working with Children Experiencing Stress

While it is true that many children exposed to trauma are resilient (i.e., quickly return to prior functioning, Malchiodi et al., 2008) and others may actually improve in functioning via a process called PTG (Leckman and Mayes, 2007), many children who are exposed to traumatic experiences manifest psychological difficulties such as those described above. The development of PTSD, while differing in frequency across different trauma populations (Roberts et al., 2009), is not unusual in children and may become chronic if not addressed.

Cognitive behavioral interventions are the most common approaches used for individuals experiencing behavioral and emotional difficulties as the result of experiencing stress. A large body of research has begun to emerge on the efficacy of treatments for those exposed to severe trauma, mostly associated with abuse or exposure to an accident or disaster (Little et al., 2011b). One intervention which has been empirically supported for use with children who have experienced trauma is Trauma-Focused Cognitive Behavior Therapy (TF-CBT). While most of the research (i.e., controlled, randomized trials) has supported the efficacy of TF-CBT with sexually abused children (Cohen et al., 2004, 2005), this intervention has also proved efficacious for other types of trauma (Cohen, 2005; Cohen et al., 2006a,b). There is also a wealth of evidence to support the efficacy of TF-CBT over treatments such as nondirective play therapy and supportive therapies for children who were the victims of sexual abuse (Cohen et al., 2004). More recently, Roberts et al. (2009) conducted a meta-analysis of 25 randomized controlled trials of interventions focused on treating symptoms of PTSD within 3 months of a traumatic event occurring. Their results indicated that for individuals who displayed traumatic stress symptoms, the only significantly effective intervention was TF-CBT. The TF-CBT was particularly effective for individuals with clinical diagnoses relating to the trauma they had experienced.

TF-CBT consists of a collection of core skills that are individually based and involve activities such as relaxation training, reviewing the traumatic event via cognitive behavior therapeutic techniques and ending, usually, with some sort of permanent product (e.g., a journal). It is believed that for treatment to be effective it must be consistent with the families' religious, community, and cultural values as the therapist works closely with the family and not just with the child in isolation. A key factor in TF-CBT is restoring trust and functioning in the child and to accomplish this the therapist must model trust, empathy, and acceptance. The ultimate goal of TF-CBT is an optimally functioning individual and family and to achieve this, the therapist attempts to establish a sense of self-efficacy in the client's affect, behavior, and cognitions (Little et al., 2009).

Within the United States, TF-CBT has been demonstrated to be equally effective with both White and African American children and youths (Weiner et al., 2009). In addition, Walker et al. (2010) demonstrate the effectiveness of a model of TF-CBT that incorporates the client's preexisting religious and spiritual functioning. Internationally, TF-CBT has been used in Sri Lanka, Indonesia, and Thailand following the 2004 tsunami and in Pakistan for children affected by the 2005 earthquakes (Cohen and Mannarino, 2008). In addition, TF-CBT has been modified for use with HIV-affected sexually abused children in Zambia (Cohen and Mannarino, 2008). Other countries where TF-CBT has been successfully adapted include Palestine/Israel, the Netherlands, Germany, Norway, and Russia (National Child Traumatic Stress Network, 2008a).

Another empirically validated intervention for children experiencing trauma-related stress is Cognitive Behavioral Intervention for Trauma in Schools (CBITS), a skills-based

group intervention for children aged 10–15 (Jaycox, 2004). The focus of this intervention is on PTSD symptomatology, depression, and anxiety (National Child Traumatic Stress Network, 2008b). CBITS has been used with children who have witnessed, or been the victims of violent acts (Stein et al., 2003), been exposed to a natural disaster (Cohen et al., 2009; Jaycox et al., 2010; Walker, 2008), a house fire, or been physically injured/abused (National Child Traumatic Stress Network, 2008b). Since it was first used in the 2000–2001 school year, CBITS has been implemented widely across the United States with different cultural groups including Latinos (Kataoka et al., 2003), Native Americans (Morsette et al., 2009), and African Americans (Jaycox et al., 2010). Internationally, it has been adapted for use in Australia, China, Guyana, and Japan (SAMHSA, 2012).

CBITS involves six main components: education, relaxation training, cognitive therapy, real-life exposure, stress or trauma exposure, and social problem-solving (Jaycox, 2004). CBITS is conducted in 10 group sessions and one to three individual sessions and was designed for use specifically in schools. It also includes teacher and parent components. CBITS is activity focused with each session involving a review of activities assigned in the previous session and new activities assigned. Intervention techniques such as relaxation are also taught and practiced (Jaycox, 2004).

Conclusion

Stress, even stress that results from exposure to traumatic events is not an unusual occurrence in childhood. Unfortunately, children experience stress; be it from bullying, family crises, abuse, or a natural disaster. In fact, in a retrospective study of adults, Felitti et al. (1998) found that approximately 50% of their sample had experienced at least one stress-producing adverse event in childhood with 25% reporting having experienced two or more. Chronic stress can result in cognitive (e.g., self-blame), affective (e.g., fear, depression, anger), behavioral (e.g., avoidance, withdrawal), and physical symptoms (e.g., autonomic nervous system arousal). Deficits in learning and memory have also been reported. Biologically, this stress response is primarily due to activation of the HPA axis stress–response system and release of the GC hormone cortisol.

While many children exhibit maladaptive symptoms as the result of stress, not all children exposed to stressors go on to develop maladaptive behaviors. The fact is, many are resilient and return to baseline levels of functioning and some even experience something known as PTG and improve their psychological functioning, both without medical or psychological assistance (Roberts et al., 2009). For those children who do need assistance in dealing with stress, cognitive behavioral interventions such as TF-CBT (Cohen et al., 2006) or CBITS (Jaycox, 2004) have been empirically validated and are available.

See also: Bereavement and Complicated Grief across the Lifespan; Child Neglect; Child Protection; Child Soldiers as Victims and Actors: Lessons from Ethiopia; Children and Families in Social Work; Concepts and Methods of Cognitive Therapies; Coping across the Lifespan; Disasters in Social, Cultural and Political Context; Evidence-Based Psychotherapy; Mass Trauma: Psychopathological Effects across the Life Span; Military and Disaster Psychiatry; Post-Traumatic Stress Disorder; Psychological Treatments, Empirically Supported; Social Work and Disasters; Stress in Adolescence: Effects on Development; Stress in Infancy and Early Childhood: Effects on Development; Stress: Effects on Neonatal and Child Development.

Bibliography

Alisic, E., van der Schoot, T.A.W., van Ginkel, J.R., Kleber, R.J., 2008. Looking beyond posttraumatic stress disorder in children: posttraumatic stress reactions, posttraumatic growth, and quality of life in a general population sample. Journal of Clinical Psychiatry 69, 1455–1461. http://dx.doi.org/10.4088/JCP.v69n0913.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, fourth ed., Text Revision. American Psychiatric Association, Washington, DC.

Ayyahh-Abdo, H., 2001. Childhood bereavement: what school psychologists need to know. School Psychology International 22, 417–433. http://dx.doi.org/10.1177/0143034301224003.

Bremner, J.D., Vermetten, E., 2001. Stress and development: behavioral and biological consequences. Development and Psychopathology 13, 473–489. http://dx.doi.org/10.1017/S0954579401003042.

Calhoun, L.G., Tedeschi, R.G., 2006. The foundations of posttraumatic growth: an expanded framework. In: Calhoun, L.G., Tedeschi, R.G. (Eds.), Handbook of Posttraumatic Growth: Research and Practice. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–24.

Carrion, V.G., Weems, C.F., Ray, R.D., Glaser, B., Hessl, D., Reiss, A.L., 2002. Diurnal salivary cortisol in pediatric posttraumatic stress disorder. Biological Psychiatry 51, 575–582. http://dx.doi.org/10.1016/S0006-3223(01)01310-5.

Cawson, P., Wattam, C., Brooker, S., Kelly, G., 2000. Child Maltreatment in the United Kingdom: A Study of the Prevalence of Child Abuse and Neglect. NSPCC, London.

CDC, 2007. Mortality by Underlying Cause among Children: US/state, 1990-2007. http://205.207.175.93/HDI/TableViewer/tableView.aspx?ReportId=530.

Chen, J.K., Astor, R.A., 2009. The perpetration of school violence in Taiwan: an analysis of gender, grade level and school type. School Psychology International 30, 568–584. http://dx.doi.org/10.1177/0143034309107076.

Cicchetti, D., 2002. How a child builds a brain: insights from normality and psychopathology. In: Hartup, W., Weinberg, R. (Eds.), Minnesota Symposia on Child Psychology: Child Psychology in Retrospect and Prospect, vol. 32. Erlbaum, Mahwah, NJ, pp. 23–71.

Cohen, J.A., 2005. Treating traumatized children: current status and future directions. In: Cardeña, E., Croyle, K. (Eds.), Acute Reactions to Trauma and Psychotherapy: A Multidisciplinary and International Perspective. Haworth Press, New York.

Cohen, J.A., Deblinger, E., Mannarino, A.P., Steer, R.A., 2004. A multi-site, randomized controlled trial for sexually abused children with PTSD symptoms. Journal of the American Academy of Child and Adolescent Psychiatry 43, 393–402. http://dx.doi.org/10.1097/01.chi.0000240839.56114.bb.

Cohen, J.A., Mannarino, A.P., 2008. Trauma-focused cognitive behavioral therapy for children and parents. Child and Adolescent Mental Health 13, 158–162. http://dx.doi.org/10.1111/j.1475-3588.2008.00502.x.

Cohen, J.A., Mannarino, A.P., Deblinger, E., 2006a. Treating Trauma and Traumatic Grief in Children and Adolescents. The Guilford Press, New York.

Cohen, J.A., Mannarino, A.P., Knudsen, K., 2005. Treating sexually abused children: one year follow-up of a randomized controlled trial. Child Abuse & Neglect 29, 135–145. http://dx.doi.org/10.1016/j.chiabu.2004.12.005.

Cohen, J.A., Mannarino, A.P., Staron, V.R., 2006b. A pilot study of modified cognitive-behavioral therapy for childhood traumatic grief (CBT-CTG). Journal of the

American Academy of Child and Adolescent Psychiatry 45, 1465–1473. http://dx.doi.org/10.1097/01.chi.0000237705.43260.2c.

Cohen, J.A., Jaycox, L.H., Walker, D.W., Mannarino, A.P., Langley, A.K., DuClos, J.L., 2009. Treating traumatized children after Hurricane Katrina: Project Fleur-de Lis. Clinical Child and Family Psychology Review 12, 55–64.

Cryder, C.H., Kilmer, R.P., Tedeschi, R.G., Calhoun, L.G., 2006. An exploratory study of posttraumatic growth in children following a natural disaster. Journal of Orthopsychiatry 76, 65–69. http://dx.doi.org/10.1037/0002-9432.76.1.65.

De Bellis, M.D., 2001. Developmental traumatology: the psychobiological development of maltreated children and its implications for research, treatment, and policy. Development and Psychopathology 13, 539–564.

De Bellis, M.D., Baum, A.S., Birmaher, B., Keshavan, M.S., Eccard, C.H., Boring, A.M., Jenkins, F.J., Ryan, N.D., 1999. A.E. Bennett research award. Developmental traumatology. Part I: Biological stress systems. Biological Psychiatry 45, 1259–1270.

Dinkes, R., Cataldi, E.F., Lin-Kelly, W., 2007. Indicators of School Crime and Safety: 2007 (NCES 2008-021/NCJ 219553). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, and Bureau of Justice. Statistics, Office of Justice Programs, U.S. Department of Justice, Washington, DC.

Essex, M.J., Klein, M.H., Cho, E., Kalin, N.H., 2002. Maternal stress beginning in infancy may sensitize children to later stress exposure: effects on cortisol and behavior. Biological Psychiatry 52, 776–784. http://dx.doi.org/10.1016/S0006-3223(02)01553-6.

Evans, L., Ohlner-Stinnett, J., 2006. Children and natural disasters: a primer for school psychologists. School Psychology International 27, 33–55. http://dx.doi.org/10.1177/0143034306062814.

Felitti, V.J., Anda, R.F., Nordenberg, D., Williamson, D.F., Spitz, A.M., Edwards, V., Koss, M.P., Marks, J.S., 1998. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) study. American Journal of Preventive Medicine 14, 245–258.

Fleming, T., Watson, P., Robinson, E., Ameratunga, S., Dixon, R., Clark, T., et al., 2007. Violence and New Zealand Young People: Findings of Youth 2000 – A National Secondary School Youth Health and Wellbeing Survey. The University of Auckland, Auckland.

Freyd, J.J., Putnam, F.W., Lyon, T.D., Becker-Blease, K.A., Cheit, R.E., Siegel, N.B., Pezdek, K., 2005. The science of child sexual abuse. Science 308, 501. http://dx.doi.org/10.1126/science.1108066.

Gunnar, M., Quevedo, K., 2007. The neurobiology of stress and development. Annual Review of Psychology 58, 145–173. http://dx.doi.org/10.1146/annurev.psych.58.110405.085605.

Hefferon, K., Grealy, M., Mutrie, N., 2009. Post-traumatic growth and life threatening physical illness: a systematic review of the qualitative literature. British Journal of Health Psychology 14, 343–378. http://dx.doi.org/10.1038/135910708X332936.

Jaycox, L.H., 2004. Cognitive Behavioral Intervention for Trauma in Schools. Sopris West, Longmont, CO.

Jaycox, L.H., Cohen, J.A., Mannarino, A.P., Walker, D.W., Langley, A.K., Gegenheimer, K.L., Scott, M., Schonlau, M., 2010. Children's mental health care following Hurricane Katrina: a field trial of trauma-focused psychotherapies. Journal of Traumatic Stress 23, 223–231. http://dx.doi.org/10.1002/jts.20518.

Karl, A., Malta, L.S., Maercker, A., 2006. Meta-analytic review of event-related potential studies in post-traumatic stress disorder. Biological Psychology 71, 123–147. http://dx.doi.org/10.1016/j.biopsycho.2005.03.004.

Kataoka, S.H., Stein, B.D., Jaycox, L.H., Wong, M., Escudero, P., Tu, W., Zaragoza, C., Fink, A., 2003. A school-based mental health program for traumatized Latino immigrant children. Journal of the American Academy of Child & Adolescent Psychiatry 42, 311–318. http://dx.doi.org/10.1097/00004583-200303000-00011.

Kemeny, M.E., 2003. The psychobiology of stress. Current Directions in Psychological Science 12, 124–129. http://dx.doi.org/10.1111/1467-8721.01246.

Leckman, J.F., Mayes, L.C., 2007. Nurturing resilient children. Journal of Child Psychology and Psychiatry 48, 221–223. http://dx.doi.org/10.1590/S1516-44462007000100003.

Little, S.G., Akin-Little, A., Gutierrez, G., 2009. Children and traumatic events: therapeutic techniques for psychologists working in the schools. Psychology in the Schools 46, 199–205. http://dx.doi.org/10.1002/pits.20364.

Little, S.G., Akin-Little, A., Medley, N., 2011a. Interventions to address school crises and violence. In: Kehle, T., Bray, M. (Eds.), The Oxford Handbook of School Psychology. Oxford University Press, New York, pp. 483–503.

Little, S.G., Akin-Little, A., Somerville, M., 2011b. Response to trauma in children: an examination of effective intervention and post-traumatic growth. School Psychology International 32, 448–463. http://dx.doi.org/10.1177/0143034311402916.

Malchiodi, C.A., Steele, W., Kuban, C., 2008. Resilience and posttraumatic growth in traumatized children. In: Malchiodi, C.A. (Ed.), Creative Interventions with Traumatized Children. The Guilford Press, New York, pp. 285–301.

Maunder, R.G., Hunter, J.J., 2001. Attachment and psychosomatic medicine: developmental contributions to stress and disease. Psychosomatic Medicine 63, 556–567.

Ministry of Women and Child Development, Government of India, 2007. Study on Child Abuse in India 2007. New Delhi, Ministry of Women and Child Development, Government of India. Retrieved from wcd.nic.in/childabuse.pdf.

Moradi, A., Neshat-Doost, H., Taghavi, R., Yule, W., Dalgleish, T., 1999. Everyday memory performance deficits in children and adolescents with PTSD: performance on the Rivermead Behavioural Memory Test. Journal of Child Psychology and Psychiatry 40, 357–362. http://dx.doi.org/10.1111/1469-7610.00453.

Morsette, A., Swaney, G., Stolle, D., Schuldberg, D., van den Pol, R., Young, M., 2009. Cognitive behavioral intervention for trauma in schools (CBITS): school-based treatment on a rural American Indian reserve. Journal of Behavior Therapy and Experimental Psychiatry 40, 169–178. http://dx.doi.org/10.1016/j.jbtep.2008.07.006.

National Child Traumatic Stress Network, 2008a. Trauma-focused Cognitive Behavior Therapy: Culture Specific Information. http://nctsnet.org/nccts/nav.do?pid=ctr_top_trmnt_prom.

National Child Traumatic Stress Network, 2008b. Cognitive Behavioral Intervention for Trauma in Schools (CBITS): General Information. Retrieved from http://nctsnet.org/nccts/nav.do?pid=ctr_top_trmnt_prom.

Rind, B., Tromovitch, P., Bauserman, R., 1998. A meta-analytic examination of assumed properties of child sexual abuse using college samples. Psychological Bulletin 124, 22–53. http://dx.doi.org/10.1037//0033-2909.124.1.22.

Roberts, N.P., Kitchiner, N.J., Kenardy, J., Bisson, J.I., 2009. Systematic review and meta-analysis of multiple-session interventions following traumatic events. The American Journal of Psychiatry 166, 293–301. http://dx.doi.org/10.1176/appi.ajp.2008.08040590.

Salter, E., Stallard, P., 2004. Posttraumatic growth in child survivors of a road traffic accident. Journal of Traumatic Stress 17, 335–340. doi: 10.1023/B:JOTS.0000038482.53911.01.

SAMHSA, 2012. National Registry of Evidence-based Programs and Practices: Cognitive Behavioral Intervention for Trauma in Schools. http://nrepp.samhsa.gov/ViewIntervention.aspx?id=153.

Samuelson, K.W., Krueger, C.E., Burnett, C., Wilson, C.K., 2010. Neuropsychological functioning in children with posttraumatic stress disorder. Child Neuropsychology 16, 119–133. http://dx.doi.org/10.1080/09297040903190782.

Sepa, A., Wahlberg, J., Vaarala, O., Frodi, A., Ludvigsson, J., 2005. Psychological stress may induce diabetes-related autoimmunity in infancy. Diabetes Care 28, 290–295. http://dx.doi.org/10.2337/diacare.28.2.290.

Shakeshaft, C., 2004. Educator Sexual Misconduct: A Synthesis of Existing Literature. US Department of Education, Washington, DC.

Social Security Administration, 2000. Intermediate Assumptions of the 2000 Trustees Report. Office of the Chief Actuary of the Social Security Administration, Washington, DC.

Stein, B.D., Jaycox, L.H., Kataoka, S.H., Wong, M., Tu, W., Elliot, M.N., Fink, A., 2003. A mental health intervention for schoolchildren exposed to violence: A randomized controlled trial. The Journal of the American Medical Association 290, 603–611.

Stover, D., 2006. Treating cyberbullying as a school violence issue. Education Digest 72, 40–42.

Sturge-Apple, M.L., Davies, P.T., Cicchetti, D., Manning, L.G., 2012. Interparental violence, maternal emotional unavailability and children's cortisol functioning in family contexts. Developmental Psychology 48, 237–249. http://dx.doi.org/10.1037/a0025419.

Susman, E.J., 2006. Psychobiology of persistent antisocial behavior: stress, early vulnerabilities, and the attenuation hypothesis. Neuroscience and Biobehavioral Reviews 30, 376–389. http://dx.doi.org/10.1016/j.neubiorev.2005.08.002.

Tang, C.S., 2007. Posttraumatic growth of southeast Asian survivors with physical injuries: six months after the 2004 southeast Asian earthquake-tsunami. Retrieved from The Australasian Journal of Disaster 2007-1 http://www.massey.ac.nz/~trauma/issues/2007-1/tang.htm.

Tedeschi, R.G., Calhoun, L.G., 2004. Posttraumatic Growth: Conceptual Foundation and Empirical Evidence. Lawrence Erlbaum Associates, Philadelphia, PA.

Tupler, L.A., De Bellis, M.D., 2006. Segmented hippocampal volume in children and adolescents with posttraumatic stress disorder. Biological Psychiatry 59, 523–529. http://dx.doi.org/10.1016/j.biopsych.2005.08.007.

Uchino, B.N., Cacioppo, J.T., Kielcolt-Glaser, J.K., 1996. The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. Psychological Bulletin 119, 488–531. http://dx.doi.org/10.1037//0033-2909.119.3.488.

UNICEF, 2007. Child Protection from Violence, Exploitation and Abuse: Children in Conflict and Emergencies. http://www.unicef.org/protection/index_armedconflict.html.

United Nations Secretary General, 2006. Violence against Children: United Nations Secretary-general's Study. http://www.unicef.org/violencestudy/.

United States Geological Survey, 2012. Significant Earthquakes of the World: 2011. http://earthquake.usgs.gov/earthquakes/eqarchives/significant/sig_2011.php.

Vasterling, J.J., Duke, L.M., Brailey, K., Constans, J.I., Allain Jr., A.N., Sutker, P.B., 2002. Attention, learning, and memory performances and intellectual resources in Vietnam veterans: PTSD and no disorder comparisons. Neuropsychology 16, 5–14.

Walker, D.F., Reese, J.B., Hughes, J.P., Troskie, M.J., 2010. Addressing religious and spiritual issues in trauma-focused cognitive behavior therapy for children and adolescents. Professional Psychology: Research and Practice 41, 174–180. http://dx.doi.org/10.1037/a0017782.

Walker, D.W., 2008. A school-based mental health service model for youth exposed to disasters: Project Fleur-de-lis. The Prevention Researcher 15 (3), 11–13.

Weiner, D.A., Schneider, A., Lyons, J.S., 2009. Evidence-based treatments for trauma among culturally diverse foster care youth: treatment retention and outcomes. Children and Youth Services Review 31, 1199–1205. http://dx.doi.org/10.1016/j.childyouth.2009.08.013.

Wilcox, H.C., Kuramoto, S.J., Lichtenstein, P., Langström, N., Brent, D.A., Runeson, B., 2010. Psychiatric morbidity, violent crime, and suicide among children and adolescents exposed to parental death. Journal of the American Academy of Child & Adolescent Psychiatry 49, 514–523. http://dx.doi.org/10.1016/j.jaac.2010.01.020.

Wong, M.L., Cavanaugh, C.E., Macleamy, J.B., Sojourner-Nelson, A., Koopman, C., 2009. Posttraumatic growth and adverse long-term effects of parental cancer in children. Family, Systems. & Health 27, 53–63.

Stress in Infancy and Early Childhood: Effects on Development

Jenalee R Doom and Megan R Gunnar, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.R. Gunnar, pp. 12336–12339, © 2001, Elsevier Ltd.

Abstract

Developmental changes during infancy and early childhood guide the maturation of the hypothalamic–pituitary–adrenal (HPA) axis, and experiences during this time may program stress-responsive systems, influencing mental and physical health into adulthood. A growing body of research focuses on social, cognitive, environmental, and genetic moderators of the stress system, as well as how the HPA axis mediates physiological adaptations to stress in childhood. The effects of early stressors, including poverty, maltreatment, and neglect, are being examined to inform interventions that target children with these life histories.

The ability to mobilize behavior and physiological reactions to cope with stressors that threaten well-being is critical to survival, yet stress reactions can impair healthy development if too frequent or too long-lasting. Experiences early in life may program stress-responsive systems, affect how individuals develop, and impact lifelong health. The activity of the hypothalamic–pituitary–adrenal (HPA) system plays an important role in transducing the effects of experience on physical and psychological health (Danese and McEwen, 2012).

In this article, we will review the anatomy and physiology of the HPA system, followed by ways to measure activity of the system in infants and young children. Normative changes in HPA regulation and reactivity from infancy to early childhood will be discussed, including recent research on moderators of the HPA axis such as social relationships and child temperament. Recent studies have implicated the HPA system as a mediator of changes in brain structure and function, immune system activity, and the epigenome. In addition, there is a growing literature on the effects of severe early life stressors such as poverty, maltreatment, and neglect on the HPA axis and early child development. The extensive impact of the stress system on development will be discussed looking ahead to future research in this area.

Psychobiology of the HPA System

The majority of research on the HPA axis has been conducted using animals, which is a challenge for investigators who must translate this knowledge to studies of humans. A review of both animal and human studies of HPA axis functioning will be presented to gain a better understanding of stress system reactivity and regulation in infancy and early childhood.

The glucocorticoid (GC) cortisol is considered to be the end product of the HPA axis, produced by the adrenal cortex after the hypothalamus of the brain initiates a signaling cascade. When stimulated, the paraventricular nucleus of the hypothalamus produces corticotropin-releasing hormone (CRH) and arginine vasopressin (AVP), neuropeptides that signal the pituitary gland of the brain to produce adrenocorticotropic hormone (ACTH). ACTH is then released into the bloodstream and travels to the cortex of the adrenal gland, where it interacts with receptors that stimulate cortisol production. Finally, cortisol is secreted into the bloodstream in order to exert effects throughout the body and to regulate HPA activity through negative feedback loops primarily in the hypothalamus, hippocampus, and medial prefrontal cortex. Additionally, CRH and ACTH participate in negative feedback loops that regulate their own production and the synthesis of peptides at other levels in the cascade.

In response to psychological stressors, the amygdala simultaneously activates the HPA axis and the sympathetic adrenomedullary (SAM) system. The SAM system is a component of the sympathetic arm of the autonomic nervous system that controls the release of epinephrine from the adrenal medulla. The release of epinephrine into the bloodstream promotes rapid physiological and behavioral responses to stressors, unlike the relatively slow-acting HPA system. Although the SAM and HPA systems act on a different timeline, they do not function in isolation. Rather, they interact in a cascade of neural events that facilitate fight-or-flight responses followed by recovery to baseline.

Following the HPA cascade, cortisol interacts with cell surface receptors to signal changes in gene transcription within the cell, which may explain the relatively long time (from minutes to hours) it takes for cortisol's actions to be measurable. The effect of cortisol varies depending on the type of receptor that is being stimulated. In the brain, two types of receptors, called mineralocorticoid receptors (MR) and glucocorticoid receptors (GR), are the targets of GCs. MRs have higher affinity for GCs than do GRs, and the balance of MR/GR occupation is imperative for HPA axis functioning. At basal levels, MRs are 80–90% occupied, while GRs are usually only occupied at the peak of the circadian rhythm or when the system is reacting to a stressor. Thus, MRs are typically responsible for basal functioning while GRs become involved to mediate functioning under conditions of stress. GRs and MRs in the hippocampus are also involved in negative feedback of the HPA axis, and the location of these receptors may explain impairment in hippocampal activities following chronic exposure to GCs. Specifically, the occupation of GRs inhibits glucose utilization in the brain, contrary to MRs, which facilitate glucose metabolism. As a result, the occupation of GRs

may have deleterious effects if stress is prolonged, and alterations in the ratio of GR/MR may increase these risks (Hostinar and Gunnar, 2012).

At basal levels, cortisol exhibits a diurnal rhythm that consists of a sharp increase at wakening with a peak 30 min later called the cortisol awakening response (CAR). After the CAR, cortisol levels decrease throughout the day, with the lowest levels 30 min after sleep onset. In addition to fluctuations due to the diurnal rhythm, cortisol shows significant increases in response to stressors that activate the HPA axis. Typically, cortisol rises sharply at the onset of a stressor, peaks 20 min after onset, and slowly decreases to baseline. However, individual variation in stress responsiveness exists in that not all individuals will activate the HPA axis in response to the same stressor, and there is considerable variability in the level of the response mounted by the system as well as the ability to return to baseline. Significant developmental changes in infancy and early childhood are discussed in detail below in relation to cortisol reactivity and the diurnal rhythm (*see* Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of).

Cortisol is not solely produced under conditions of stress, but rather it is necessary for the normal functioning of several physiological systems, including the nervous, gastrointestinal, reproductive, cardiovascular, and immune systems. McEwen (2008) explains that both hyperactivity and hypoactivity of the HPA axis have been linked to pathological functioning of the brain and other physiological systems, with alterations in basal regulation and reactivity to stressors implicated as possible mechanisms of disorder (reviewed in Hostinar and Gunnar, 2012). It is hypothesized that the relationship between stress and optimal functioning can be represented as an inverted U-curve, with too much or too little stimulation associated with lower functioning than medium-level stimulation. Animal models of chronic prenatal and postnatal stress show increases in basal glucocorticoid levels, alterations of GR/MR ratios, and increased vulnerability to stressors later in life. Cirulli et al. (2003) report that in rodent models, prolonged maternal separation as a pup is related to larger startle responses, greater freezing and anxiety behaviors, and two to three times the ACTH and GC response to stress as adults compared to those who experienced only brief separations as pups (reviewed in Hostinar and Gunnar, 2012). Increased vulnerability to stress and altered stress system activity are also seen in adults and children who have experienced childhood maltreatment, deprivation, or chronic stress.

Measurement Issues in Infants and Young Children

In children, researchers commonly measure cortisol levels using free-flowing saliva (see for review, Gunnar and Adam, 2012). Although useful, it is unfortunate that sampling higher levels of the HPA axis can only be done invasively (e.g., spinal tap for CRH or plasma for ACTH). Drugs that are used in adults to assess stress system physiology – to control for individual differences in stressor interpretation – are administered invasively and are thus not used if not medically indicated in children. Nonetheless, researchers have used saliva to sample both diurnal cortisol and cortisol reactivity to an acute stressor, although not without measurement difficulties.

Due to the diurnal rhythm of cortisol, saliva samples must be collected at the same time each day, as levels will vary greatly from morning to evening. Children often have difficulty producing enough saliva or using the passive drool method for saliva collection, which are challenges for researchers who need cortisol levels to remain unaffected by collection methods. Investigators must carefully consider the child's health, recent foods or drinks ingested, and the effects of a novel lab environment on cortisol production. The use of medication in children may alter cortisol levels or even the child's subjective experience of stress. Sufficient screening must be done to ensure all potential confounding variables are recorded. Finally, stressor paradigms intended to elicit cortisol responses may or may not be successful in children depending on age, the availability of coping resources, and the presence of threats to the social self.

Hair cortisol analysis is a relatively new method of sampling that aims to measure long-term cortisol exposure as opposed to real-time fluctuations in cortisol levels captured by saliva. Each 1 cm length of hair represents approximately 1 month of development, and thus cumulative cortisol measurements can be obtained for specific time periods. Further research on hair cortisol analysis must be conducted to verify its routine use in samples experiencing chronic stress.

Developmental Changes in the HPA Axis

Newborns have the capacity to mount physiologically significant cortisol and ACTH responses to environmental stressors. Specifically, stressors that elicit pain, such as heel lances, promote sensitization to the response while handling stressors, such as physical examinations, show habituation over time. Although newborns show robust cortisol responses to even mild stressors, across the first year of life it becomes increasingly difficult to elicit cortisol responses to the same types of stressors. This period of hyporesponsiveness mimics the stress hyporesponsive period that researchers have studied in rat pups between 4 and 14 days of age, and it is uncertain whether infants enter the functional equivalent of this period during the first year of life. By 1 year of age, infants typically show no increases in cortisol to stressors such as separations or stranger approach, even while exhibiting significant behavioral distress (Hostinar and Gunnar, 2012). Research suggests that physiological changes including improved negative feedback regulation of the HPA axis and decreased adrenocortical sensitivity to ACTH, may at least partially explain hyporesponsiveness of the system during this period. Relationships with caregivers also act as an important moderator of the HPA response. Children who experience supportive adult care are better able to buffer cortisol responses, while children who have been separated from their caregiver exhibit elevations in cortisol to the same stressors. Over time, children develop the ability to regulate HPA responses in the absence of caregivers. Although able to mount a significant cortisol response, newborns lack the typical diurnal cortisol rhythm seen in adults and instead show two peaks that are 12 h apart and do not depend on the time of day. By 3 months of age, infants usually develop a more adult-like rhythm that includes high morning and

low evening cortisol levels. Watamura et al. (2004), reviewed in Hostinar and Gunnar (2012), report that napping in toddlers seems to facilitate decreases in cortisol across the day, possibly involving development of the circadian HPA axis rhythm through age 3.

Levels of corticosteroid-binding globulin (CBG), a factor that binds circulating cortisol and renders it biologically inactive, may explain differences in cortisol levels across early childhood. In adults, about 80% of cortisol is bound to CBG, but CBG levels are low in newborns, and adult levels are not reached until approximately 6 months of age. Thus, although total plasma cortisol is low in newborns, levels of free cortisol in infants under 6 months often surpass older infants and children because of increasing cortisol levels and low levels of CBG. Young infants also show marked increases in free cortisol to very minor stressors such as undressing. Remarkably, the HPA axis is relatively mature by 6 months of age (Hostinar and Gunnar, 2012).

Relationships as Regulators of the HPA System

Until approximately 6 months of age, the infant HPA axis reacts robustly to a variety of stressors, but between 6 and 12 months of age, cortisol responses are mediated by caregiver interactions. Children's attachment to caregivers is an important moderator of physiological reactivity. In the presence of a caregiver, toddlers with a secure attachment to the caregiver have a buffered cortisol response to stressors, while those with insecure attachments show elevations in cortisol (Hostinar and Gunnar, 2012). Ahnert and colleagues (2004) report that attachment to caregivers does not predict the magnitude of the cortisol response during daily separations such as day care, but securely attached children are more likely to show increased cortisol responses that match their level of behavioral distress (reviewed in Hostinar and Gunnar, 2012). Further, Luijk et al. (2010) found that children who have disorganized attachment to caregivers – often observed for children whose caregivers are depressed or abusive – exhibit elevations in cortisol and a more flattened diurnal cortisol pattern than children with non-disorganized attachment (reviewed in Hostinar and Gunnar, 2012). A history of responsive care also predicts whether cortisol will be buffered in response to stress, as responsive caregiving is thought to give children an outlet to express emotions without triggering the neuroendocrine response (Hostinar and Gunnar, 2012).

Social stressors within the family such as child conflict and interparental violence are associated with altered HPA axis reactivity and regulation in children. Infants at 7 months who experienced more home violence and used less caregiver-oriented regulation strategies were more likely to show high cortisol reactivity (Towe-Goodman et al., 2012). However, there is evidence that by toddlerhood, children who experience more violence and maternal unavailability actually show lower cortisol reactivity to a social stressor (Sturge-Apple et al., 2012). Consistent with hyporeactivity observed in toddlers, 4–5-year olds who experienced greater child conflict in the home had lower morning cortisol and flatter diurnal cortisol rhythms than children who experienced less conflict (Slatcher and Robles, 2012).

A more normative but less potent social stressor for children is the transition to child care, which involves peer interaction and separation from primary caregivers. A meta-analysis by Vermeer and van IJzendoorn (2006) found that while most infants and toddlers show expected decreases in cortisol across the day at home, many children at day care actually show increases from morning to afternoon (reviewed in Phillips et al., 2011). Ouellet-Morin and colleagues (2010) report that by 3 years of age, children are more likely to show decreases at both home and day care, especially if they have more day care experience (reviewed in Phillips et al., 2011). Sensitive and responsive caregiving from both the primary caregiver and day care staff helps children to adapt effectively. Separation from the caregiver is less likely to activate the HPA axis if high-quality substitute care is provided. Attachment security to the teacher predicts falling cortisol levels across the day, and a secure attachment to the caregiver interacts with day care quality so that those with better-quality care have buffered cortisol levels (Badanes et al., 2012). In addition, peer relations are a source of significant stress in day care, and child social competence is a skill that reduces stressful interactions with peers and lowers cortisol production as a result (Hostinar and Gunnar, 2012).

Temperament

Behavioral inhibition in young children has been consistently linked to increased physiological reactivity, including greater activity of the HPA axis. Behavioral inhibition parallels the activation of conditioned and unconditioned fear circuits in the brain, which influence behavioral responses as well as autonomic and HPA axis reactivity. However, HPA axis activity may not necessarily be a reflection of fear-related responses but rather the child's ability to cope with a stressful situation. For example, children who are behaviorally inhibited may avoid social situations that would trigger an increase in cortisol, but in situations that they must confront, cortisol reactivity may indicate whether children are coping successfully. The presence of a caregiver to whom the child is securely attached may help the child to cope and thus explain why some studies do not report higher cortisol in behaviorally inhibited children.

Children who are behaviorally inhibited or have negative emotional temperaments may be more susceptible to high HPA reactivity under conditions of low-quality care than children with more moderate temperaments (Phillips et al., 2011). Important gender differences may exist in the components of temperament associated with cortisol reactivity. Gunnar et al. (2010) report that in 3–4-year-old girls, those who exhibited more anxious and vigilant behavior were more likely to show a rise in cortisol across the day at day care. For boys, however, this rise was related to angry, aggressive behavior (reviewed in Phillips et al., 2011).

In a peer context, children from generally low-risk backgrounds who are rejected by others show higher cortisol levels. Gunnar et al. (2003) report that preschoolers who are overly controlled and less exuberant are more likely to have high cortisol levels, but in a social situation, it is the less controlled (more impulsive) and more exuberant children who are more likely to be rejected by peers and show higher cortisol levels as

a result (reviewed in Hostinar and Gunnar, 2012). Although cortisol levels for average or exuberant preschoolers lessen over time, highly inhibited children do not show this decrease on average. A recent study found that inhibited children who were more dominant and popular than other inhibited children showed increases in cortisol over time. On the other hand, exuberant children who were not socially integrated showed consistently high cortisol levels across the year (Tarullo et al., 2011). Gunnar et al. (2011) report that behavioral inhibition may also moderate the effect of cortisol reactivity on later internalizing symptoms, with preschoolers who are highly inhibited and reactive at the greatest risk for internalizing symptoms (reviewed in Phillips et al., 2011).

Effortful control is a higher order dimension of temperament that includes the ability to focus attention and inhibit a prepotent response. Effortful control has been identified as an important moderator of the relationship between negative emotionality and behavior in children. Children who score highly on measures of effortful control are better able to suppress both positive and negative emotions. In general, preschoolers who exhibit more effortful control show better regulation and less activation of the HPA axis as well as a more adult-like diurnal cortisol rhythm.

Moderators of Stress System Reactivity and Gene Expression

Individual differences in child cognitive and emotional factors explain significant variation in HPA axis activity. Specifically, executive function, the top-down ability to manage lower order cognitive processes, has been shown to moderate cortisol levels. Blair and colleagues (2011) found that infants and toddlers with higher cortisol levels tend to have lower levels of executive function in early childhood, and cortisol partially mediates the positive effect of parenting on executive function (reviewed in Blair, 2010). In addition, Blair et al. (2005) reported that the ability to mount a moderate cortisol increase to challenge and appropriate return to baseline is positively associated with executive function and teacher-reported self-regulation (reviewed in Blair, 2010). The child's ability to regulate emotion is another important moderator of HPA activity. As children gain emotion regulation skills, they also become better at regulating physiological reactivity to psychosocial stressors. Good emotion regulation skills, including the ability to utilize appropriate coping strategies, have been shown to buffer cortisol responses to social stress in preschoolers.

Unfortunately, child cognitive and emotional factors may not be enough to protect children from stress system dysregulation in response to unfavorable home environments. Essex et al. (2002), as reviewed in Hostinar and Gunnar (2012), report that maternal depression, a condition that significantly disrupts family life and maternal care, is related to elevated cortisol levels and heightened HPA axis reactivity in early childhood. Maternal comorbid anxiety and depression has also been linked to increased infant cortisol reactivity. More studies are needed to understand how specific types of parental psychopathology, as well as the timing and treatment of disorders, affect child stress reactivity and regulation.

Gene–environment interactions explain at least part of the variance in stress regulation and responsiveness in children. A study of 19-month-old twins aimed to understand the unique contributions of genes and the environment to differences in stress reactivity under differing levels of adversity (Ouellet-Morin et al., 2008). In addition to unique environmental influences, under conditions of low familial adversity, shared genetic influences most strongly affect cortisol responses, while under conditions of high familial adversity, shared environmental influences showed greater associations with cortisol responsiveness. As a result, experiences in early childhood may have a powerful programming effect on the stress system in highly stressful conditions.

Allelic variation in genes that influence the HPA axis and related systems (e.g., the sympathetic nervous system) operate through various mechanisms to influence development. These interactions between the gene and the environment may operate differently in conditions of stress versus basal functioning, and multiple genes likely interact with each other simultaneously to influence protein synthesis and processes that affect downstream behavior. The ability of genes to influence behavior may differ depending on the timing during development. For instance, the activity of a gene during a sensitive period could have greater implications for development than during another period. Alterations to the genome itself may occur in response to a particular environment, including methylation of specific functional genes and shortening of DNA telomeres. Oberlander and colleagues (2008) report that maternal depressed/anxious mood in the third trimester predicted specific methylation of the *NR3C1* gene, which was then associated with an increased cortisol response in infants at 3 months (reviewed in Miller et al., 2011). Drury et al. (2011) noted telomere shortening in 4-year olds who had spent most of their lives in institutions, showing the impact of stress even in early childhood (reviewed in Danese and McEwen, 2012). Studies that integrate measurement of the environment, changes to the genome, and subsequent physiological alterations will be essential to understanding the mechanisms by which the environment and the genome interact to affect individual development.

HPA System Impacts the Brain and Immune System

Interactions between the environment and the HPA axis produce further alterations in the brain, the immune system, and the epigenome, among other physiological systems. Research in animals has begun to elucidate the mechanisms behind these changes, so most of what is known about stress, the HPA axis, and other physiological systems has been extrapolated from animal research. However, studies in infants and young children are emerging that allow us to understand how these mechanisms might be operating early in human development.

The HPA axis provides a way in which stress can shape brain development. Cortisol passes through the blood–brain barrier, and brain regions such as the hippocampus and amygdala are dense with glucocorticoid receptors. Although moderate levels

and brief increases in cortisol appear to enhance brain function, frequent or chronically high levels of cortisol may produce cell death in addition to other functional or structural changes in the brain. In maltreated (highly stressed) children, the most consistent neurological finding is impairment in the prefrontal cortex (PFC). The PFC tends to be smaller in volume, and children are more likely to experience impulsiveness, inattention, and deficits in executive function (Danese and McEwen, 2012). Many – but not all – studies report increased amygdala volume in maltreated children, but there is little evidence to date of changes in hippocampal volume. The animal literature suggests that HPA axis hyperactivity is a plausible mechanism by which neurological deficits arise.

Stress and elevated cortisol levels in early childhood have been linked to immune system impairments including decreased antibody levels and increased inflammation. Watamura et al. (2010) found that children whose cortisol increased across the day at child care tended to have lower antibody levels at home, and children with higher home cortisol were more likely to experience parent-reported illness (reviewed in Vermeer et al., 2012). In addition, children in both the home and at child care showed lower antibody levels when caregiver sensitivity was low (Vermeer et al., 2012). Unlike the decreases observed in antibody levels, stress in early childhood has been linked to increased levels of inflammatory proteins. This elevation could be due to excessive stimulation of the sympathetic nervous system and impairments in HPA axis signaling, but it is possible that stress is not increasing inflammation but rather the effect is the reverse; inflammation is increasing glucocorticoid levels (Danese and McEwen, 2012). Thus longitudinal work is needed to understand the direction of effects.

Early Adversity: Poverty, Neglect, and Maltreatment

Early adversity significantly impacts stress system regulation, gene expression, and a myriad of other developmental processes from prenatal to adult life. Significant adversity consistently produces stress system disruptions in infants and young children (Danese and McEwen, 2012). Chronic stress has been linked to alterations in HPA axis functioning, but the patterns of disruption vary based on the type, timing, and subjective meaning of the stressor. Such variation is to be expected. When the HPA axis is activated frequently for prolonged periods, it undergoes remodeling that results in normal to low levels of cortisol production in reaction to brain signals of threat. At the same time, chronic threat increases the activity of brain circuits that process threat, including increasing the production of CRH from sites such as the central nucleus of the amygdala. The result is hypocortisol production in the face of heightened sympathetic and central brain stress signals (as reviewed in Hostinar and Gunnar, 2012). For the researcher measuring cortisol, however, the result is sometimes finding elevated and sometimes finding suppressed cortisol activity in response to adverse life conditions in children. For example, Chen et al. (2010) report an association between low socioeconomic status (SES) and greater increases in diurnal cortisol from birth to age 2, and the effect was partially mediated by increased perception of threat and family chaos in low SES children (reviewed in Miller et al., 2011). By preschool, children in poverty have been shown to exhibit blunted morning cortisol levels and a flattened diurnal slope (Zalewski et al., 2012). This pattern could be a continuation of the increased cortisol output shown from birth for low SES children, putting children at risk for the deleterious effects associated with chronic activation of stress systems.

Children in foster care show a similar pattern of stress system dysregulation to children living in poverty, with lower morning cortisol levels and a flatter slope across the day. Dozier and colleagues (2006), as reviewed in Hostinar and Gunnar (2012), report that children between the ages of one and five living in foster care are more likely to show atypical diurnal cortisol patterns (high or low) than children not living in foster care. Morning cortisol in foster children may signal types of maltreatment the child has experienced. Bruce et al. (2009) found low morning cortisol to be related to higher incidence of severe physical neglect in preschool-age foster children, and high morning cortisol was linked to severe emotional maltreatment (reviewed in Hostinar and Gunnar, 2012).

Although there are few studies of cortisol regulation and reactivity in infants and young children who have experienced maltreatment, there is evidence that poor early care affects neuroendocrine levels early in life (Hostinar and Gunnar, 2012). Specifically, maltreated preschoolers may exhibit lower cortisol reactivity to stress than children who are not maltreated, and on days of high conflict they may produce even lower cortisol levels. This hypoactivity could be due to down-regulation following periods of high GC levels after repeated exposure to maltreatment. Bugental et al. (2003) report that less severe forms of maltreatment, such as maternal emotional withdrawal, have been related to elevated baseline cortisol levels, and increased cortisol reactivity has been noted in infants who receive frequent corporal punishment (reviewed in Sturge-Apple et al., 2012). Increases in cortisol may result when infants have limited use of their attachment figure to help them buffer stressors such as maltreatment. Excess glucocorticoid levels in maltreated children may be particularly harmful if they affect the developing brain, especially neural circuits involved in interpreting threats and regulating the stress response (Danese and McEwen, 2012). It is likely that HPA hyperactivity is adaptive during the short-term response to stress, but these effects may no longer be adaptive when the response is prolonged or once the child is removed from the context of maltreatment.

Early social deprivation has been shown to significantly alter HPA axis function. Toddlers living in institutions with poor quality care typically show blunted morning cortisol levels and little to no decrease in cortisol across the day. Children may also have impaired HPA axis regulation in social situations even after adoption into a supportive home. Post-institutionalized children who had been with their adoptive families for at least 3 years showed prolonged cortisol elevations during an interaction with their mother, but not a stranger. Further, children who experienced more severe neglect and deprivation are more likely to exhibit dysregulation of the HPA axis years following rescue (as reviewed in Hostinar

and Gunnar, 2012). Current research is being done to examine recovery of the HPA axis activity following the transition from early deprivation to more optimal care in post-institutionalized children.

Lasting Effects of Early Adversity

Recent evidence strongly suggests that psychosocial stress and altered HPA activity during the first few years of life have the potential to impact physiological systems, cognition, and behavior across the lifespan. In a sample of adults who were separated from their parents as children during World War II, those who experienced separation had higher baseline cortisol and ACTH levels as well as greater cortisol reactivity than nonseparated adults (Pesonen et al., 2010). The effects were greatest for children separated during early childhood compared to infancy or school-age separations. Trauma during early childhood has also been linked to alterations in adult physiological systems that are sometimes inconsistent with those reported in childhood. For instance, although the amygdala seems to be enlarged in maltreated children, adults who experienced maltreatment as children have normal to decreased amygdala volumes (Danese and McEwen, 2012). In addition, adults maltreated as children show decreased hippocampal volume, although there were no differences between maltreatment groups in hippocampal size in childhood. Abnormalities in brain regions following childhood stress could be the result of excess glucocorticoid or inflammatory protein exposure over time. These findings underscore the importance of studying effects of early childhood stress and altered HPA activity developmentally because mechanisms that produce certain outcomes in childhood may result in different outcomes later in life.

Future of Research

A proliferation of research using human models of early stress has occurred during the past decade, but future psychobiological studies are needed to further elucidate the mechanisms of normative and atypical development in children. Both animal and human researches will be necessary to understand the interactions between the HPA axis, immune system, genome, and the brain. In addition, further studies of social, cognitive, and environmental moderators of stress system reactivity will help researchers understand how early experiences affect later mental and physical health. Recent interventions have successfully targeted the stress system as a mediator of recovery from early adversity, and future applied work will be necessary to understand how these mechanisms operate and how to apply scientific knowledge to help children in need.

See also: Stress in Childhood: Effects on Development; Stress: Effects on Neonatal and Child Development; Temperament and Human Development.

Bibliography

Badanes, L.S., Dmitrieva, J., Watamura, S.E., 2012. Understanding cortisol reactivity across the day at child care: the potential buffering role of secure attachments to caregivers. Early Childhood Research Quarterly 27 (1), 156–165.

Blair, C., 2010. Going down to the crossroads: neuroendocrinology, developmental psychobiology, and prospects for research at the intersection of neuroscience and education. Mind, Brain, and Education 4 (4), 182–187.

Danese, A., McEwen, B.S., 2012. Adverse childhood experiences, allostasis, allostatic load, and age-related disease. Physiology & Behavior 106, 29–39.

Gunnar, M.R., Adam, E.K., 2012. What can neuroendocrine measures reveal about emotion? In: Dennis, T.A., Buss, K.A., Hasting, P.D. (Eds.), Physiological Measures of Emotion from a Developmental Perspective: State of the Science. Society for Research in Child Development Monograph 77(2), 109–119.

Hostinar, C.E., Gunnar, M.R., 2012. The developmental psychobiology of stress and emotion in childhood. In: Lerner, R.M., Esterbrooks, M.A., Mistry, J. (Eds.), Handbook of Psychology, vol. 6: Developmental Psychology. Wiley & Sons, Hoboken, NJ.

Miller, G.E., Chen, E., Parker, K.J., 2011. Psychological stress in childhood and susceptibility to the chronic diseases of aging: moving toward a model of behavioral and biological mechanisms. Psychological Bulletin 137 (6), 959–997.

Ouellet-Morin, I., Boivin, M., Dionne, G., Lupien, S.J., Arsenault, L., Barr, R.G., et al., 2008. Variations in heritability of cortisol reactivity to stress as a function of early familial adversity among 19-month-old twins. Archives of General Psychiatry 65, 211–218.

Pesonen, A.K., Räikkönen, K., Feldt, K., Heinonen, K., Osmond, C., Phillips, D.I., et al., 2010. Childhood separation experience predicts HPA axis hormonal responses in late adulthood: a natural experiment of World War II. Psychoneuroendocrinology 35, 758–767.

Phillips, D.A., Fox, N.A., Gunnar, M.R., 2011. Same Place, different experiences: Bringing individual differences to research in child care. Child Development Perspectives 5 (1), 44–49.

Slatcher, R.B., Robles, T.F., 2012. Preschoolers' everyday conflict at home and diurnal cortisol patterns. Health Psychology (advance online publication).

Sturge-Apple, M.L., Davies, P.T., Cicchetti, D., Manning, L.G., 2012. Interparental violence, maternal emotional unavailability and Children's cortisol functioning in family contexts. Dev Psychol 48 (1), 237–249.

Tarullo, A.R., Mliner, S., Gunnar, M.R., 2011. Inhibition and exuberance in preschool classrooms: associations with peer social experiences and changes in cortisol across the preschool year. Developmental Psychology 47 (5), 1374–1388.

Towe-Goodman, N.R., Stifter, C.A., Mills-Koonce, W.R., Granger, D.A., The Family Life Project Key Investigators, 2012. Interparental aggression and infant patterns of adrenocortical and behavioral stress responses. Developmental Psychobiology (advance online publication).

Vermeer, H.J., van IJzendoorn, M.H., Groeneveld, M.G., Granger, D.A., 2012. Down-regulation of the immune system in low-quality child care: the case of secretory immunoglobulin A (SIgA) in toddlers. Physiology & Behavior 105 (2), 161–167.

Zalewski, M., Lengua, L.J., Kiff, C.J., Fisher, P.A., 2012. Understanding the relation of low income to HPA-axis functioning in preschool children: cumulative family risk and parenting as pathways to disruptions in cortisol. Child Psychiatry and Human Development (advance online publication).

Stress: Psychological Perspectives

Scott M Monroe and Lori F Cummins, University of Notre Dame, Notre Dame, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.M. Monroe, volume 22, pp. 15198–15201, © 2001, Elsevier Ltd.

Abstract

Psychological stress is a popular term denoting processes believed to contribute to the onset and maintenance of a variety of mental and physical conditions. Despite widespread interest in psychological stress and its consequences for health and well-being, debate remains about how to best define the term. Three perspectives for defining and studying psychological stress are reviewed with respect to their history, development, and current status. These three perspectives on psychological stress differ in terms of the relative emphasis each places on the environment, the organism, and the interaction between the organism and environment over time. Conceptual and methodological considerations and implications of the three perspectives are reviewed. Recent elaborations of the concept of stress that attempt to expand its biological consequences over time are also discussed. Finally, promising topics for future inquiry on psychological stress are addressed.

Psychological stress, like the general term 'stress,' is a commonly used expression that defies a universally agreed-upon meaning. These definitional challenges are not due to a lack of interest in the concept of psychological stress. Indeed, the opposite is the case. Stress is a concept that fascinates scientists, clinicians, and laypersons alike, for its apparent ability to explain a wide-range of conditions and adverse outcomes. From understanding immediate emotional states through understanding a variety of mental and physical disorders, stress is a commonly invoked cause. It is the resulting complexity of this broad explanatory enterprise that both attracts and vexes scientific investigators.

To understand current thinking about psychological stress, historical considerations are addressed first in the following section. Subsequently, the three different perspectives that have evolved to define stress are described, along with their own conceptual and methodological traditions for operationalizing psychological stress and conducting research on its effects. Following this, key conceptual and methodological issues regarding psychological stress are presented. We conclude with a discussion of contemporary trends and future directions regarding psychological stress.

Historical Considerations and Concerns

Curiosity about the implications of emotions for health is age-old, dating back to the origins of holistic approaches to medicine as exemplified by the writings of Hippocrates (400 BC). Interest in the influence of psychological factors on emotional functioning has consequently represented a related theme of interest over the centuries (*see* Environmental Stress and Health). The basic premise is that stress gives rise to distressing emotions, which in turn are associated with perturbations in biological systems, which in turn contribute to disease onset or exacerbation. Yet the connections between various related terms and their intended ideas are quite rich, and more complex than the current vocabulary suggests. Indeed, 'stress' is derived as a shortened form of 'distress'; disease, in turn, originally referred to 'dis-ease,' or distress (Rees, 1976). It is in this lexicographic context that the varied connections between stress, distress, and disease can be appreciated, only some of which may be useful for scientific purposes.

As rich in explanatory power as psychological stress might be for understanding the causes of distress and disease, the linguistic overlap and interplay of these terms also raises problems. The intuitive appeal of these ideas, terms, and relations can foster false theories and findings. There is a tendency to 'explain away' disorders of unknown etiology based on prevailing beliefs about the unwanted or adverse social or psychological circumstances (Monroe, 2008). For example, before the physical causes of specific infectious diseases were discovered, the origins of these illnesses were commonly ascribed to psychological stress and distressing emotions (Sontag, 1978). Prior to the discovery of the *Treponema pallidum* spirochete, the common and deadly mental condition of general paresis (or tertiary syphilis) in the nineteenth century commonly was attributed to the 'demands of civilization on physical and mental powers,' including 'overwork,' 'overstudy,' 'competition,' and the 'feverish pursuit of wealth and social position' (Kellogg, 1897). In contemporary parlance, psychological stress frequently provides a satisfying pseudoexplanation for poorly understood mental and medical conditions. Whether it is in the scientific journals, the popular media, or the physician's office, 'stress' is often resorted to for explaining mystifying maladies and miseries. The intuitive appeal of explanations based on psychological stress and related ideas require firm thinking and sound science to differentiate valid associations from plausible and convenient fictions.

Research Traditions and Major Definitional Perspectives

Two research traditions can be discerned which, together, have given rise to three definitional perspectives on psychological stress. The first research tradition derives from clinical observations and field studies of humans exposed to threatening and adverse circumstances. The second tradition arises from animal

laboratory studies, where experimental control over stressful conditions and response alternatives of the organism has been employed. Emerging from and uniting these two traditions is the general interest in the psychological concomitants and sequelae associated with adverse environments and stimuli, particularly with regard to their implications for subsequent health and well-being. How this general interest in environment, responses, and disorder has been more precisely conceptualized and defined, though, has resulted in three differing definitional perspectives on defining stress. The differences between these perspectives, as explained next, revolve around the relative emphasis placed on the environment, organism, or environment–organism interactions over time as the focal feature for defining and operationalizing psychological stress.

The Stimulus or Environmental Perspective

The stimulus or environment perspective adopts the position that psychological stress derives from the characteristics of the environment the person or organism confronts. The underlying assumption is that objective external circumstances can be conceptualized and defined as psychologically demanding, threatening, benign, and so forth. Prior research on both humans and animals has advanced understanding of this perspective. In human field or human laboratory research, the stimulus conditions were recorded in naturalistic contexts or manipulated in ethically responsible ways. Animal laboratory research, on the other hand, allowed for experimental control over the environmental conditions, thereby providing opportunities for greater latitude in exposure to diverse challenging and threatening stimuli.

Other factors, too, were influential in promoting the environmental perspective. Attention to the environment for creating stress and illness became especially evident toward the mid-twentieth century; the experiences of World War II reshaped the ways in which a breakdown in human functioning was viewed. Prior to this time, it was thought that individual vulnerability (i.e., biological) largely accounted for who succumbed to stress. Normal people were believed to maintain health and survive the ravages of stress; the biologically predisposed or tainted were those who broke down. With the awareness that one-third of American casualties in World War II sustained no physical injury (Weiner, 1992), the idea that severe stress could precipitate psychological and physical breakdown in previously normal individuals was incorporated into medical and psychiatric thinking (Dohrenwend, 1998). This brought to the foreground the importance of extreme and challenging environments for the defining psychological stress and for investigating its health consequences in normal as well as vulnerable populations.

Finally, a surge of popular interest and research activity in psychological stress from the environmental perspective came from another avenue of inquiry: life events questionnaires. Thomas Holmes and Richard Rahe promoted the idea that life stress could be defined and assessed in an objective manner through surveying people's recent experiences. These researchers developed a list of life events observed to occur prior to the onset of disease, including items such as marriage, trouble with the boss, change in sleeping habits, death in the family, and vacation. They named the list of 43 life events as the Schedule of Recent Experiences (SRE). The events on the SRE subsequently were collated and weighted using psychophysics procedures, producing the Social Readjustment Rating Scale (SRRS; Holmes and Rahe, 1967). Literally thousands of studies have been conducted with the SRE, SRRS, and the proliferation of derivative life event checklists they spawned (Monroe, 2008).

The Response Perspective

From the response perspective, psychological stress is viewed as the complex of *reactions* of the organism to the adaptive demands imposed by environment change and challenge. This perspective can be traced to the writings of Darwin and the theory of natural selection. It is in these ideas that attention began to shift toward the *implications* for the organism of the dangers and threat associated with environmental demands, bringing to light an awareness of the importance of the ongoing *adaptations* of the organism to the everaltering external circumstances. The response approach to defining stress takes on special importance from this evolutionary vantage point, as it recognizes that not only biological and behavioral repertoires are needed for enhanced fitness and adaptive capacities. Most importantly from a psychological perspective on stress, evolution also fostered skills in *anticipation* of dangers and adversity (Weiner, 1992). In other words, the ability to not just respond to environmental exigencies, but to anticipate and to prepare for such potential exigencies, provided a more complete and effective adaptive system. Being able to foresee, prepare for, and prevent potential problems in the future drew upon mental, emotional, and behavioral repertoires that, over time, enhanced survival and reproduction.

The twentieth century brought more sustained and intense interest in psychological stress from the response perspective. Claude Bernard's attention to the body's regulatory functions and homeostasis, along with the physiological regulation of emotions by Walter Cannon, furthered scientific interests in how the body maintained its regulatory functions in the face of changing external circumstances. Most importantly, Hans Selye introduced the term 'stress' into formal scientific arenas. By defining stress as "the nonspecific response to the body of any demand made upon it" (Selye, 1976: p. 74), Selye grounded the concept explicitly in the response of the organism to environmental challenge. (He reserved the term 'stressor' to denote the stimuli or environmental conditions eliciting stress responses.) Widely regarded as the father of modern stress theory and research, Selye and his writings on the stress response essentially launched the field and contributed greatly to the appeal of the stress concept to both researchers and laypersons alike.

The Environment–Organism Transaction Perspective

The stimulus and response perspectives on psychological stress each possess advantages, yet they also have important limitations. For example, the objective characteristics of the environment do not always predict the organism's responses with precision; individual differences in response to identical or similar stimuli are common. Moreover, when considering the

response perspective, there are many levels of response (e.g., psychological, behavioral, biological) and numerous potential response indicators within these levels (e.g., specific emotions, behaviors, psychophysiologic, or hormonal reactions) to draw upon, many of which display different patterns of relationships with distinct environmental challenges (Weiner, 1992).

To address these limitations and enlarge the scope of inquiry, psychological stress has been proposed to represent a particular *relationship over time* between the organism and environment. This third perspective holds that psychological stress is the product of the individual's ongoing transactions with ambient circumstances, particularly circumstances that are perceived as taxing or exceeding adaptive resources. Several possibilities are entertained within this perspective, ranging from an emphasis on stressful *experiences* (Weiner, 1992) through emphases upon the ongoing process of event occurrence, appraisal, coping, reappraisal, and eventual resolution (Lazarus and Folkman, 1984). For instance, a person is faced with the threat of job loss. This individual appraises the degree of threat in relation to his or her personal circumstances (e.g., economic conditions, ability to find alternative employment, possible embarrassment). He or she undertakes measures to cope with the possible termination of employment (e.g., works harder, begins looking for alternative work). As a result of these efforts, the environmental conditions are changed, hopefully for the better (but not necessarily) and the resulting responses altered as well. The core feature of this perspective is an *active process* between a dynamic environment and an adapting organism over time, whereby stress is the accruing and collective net result of the process.

Conceptual Challenges and Problems

After the 1950s, stress became a common term, appearing across diverse research literature in sociology, psychology, medicine, physiology, and biology. A broad array of loosely related constructs was united under this common rubric. At least two issues arise from this explosive growth in research on stress: (1) individual differences; and (2) generality versus specificity of psychological stress.

Individual Differences

Despite the promising leads of early research on psychological stress, it became apparent that variability in people's responses to stressful circumstances was common. Not all individuals succumb to even the most extreme of situations, while some people break down in the face of apparently minimal demands. Theory and research were needed to account for such variation across individuals, thereby introducing factors hypothesized to moderate stress effects.

A multitude of factors has been proposed as potential stress moderators. Attention to this issue has focused on individual difference variables such as appraisal of threat and coping (Lazarus, 2000; Monroe, 2008), and a variety of other considerations believed to moderate the effects of stress (e.g., social support; Brown and Harris, 1989; Holt-Lunstad et al., 2010). The net effect has been to broaden the conceptual arena and research operations considerably. Multifactorial models have been introduced taking into account a range of factors that could influence stressful features of the environment and responses over time (*see* Coping across the Lifespan).

Generality versus Specificity of Psychological Stress

A second issue related to the rapid growth of stress research pertains to the omnibus nature of stress. On the one hand, the stress concept has united a variety of cognate constructs and emphasized the general nature of the stress response with regard to a wide variety of adaptive demands (e.g., Selye, 1976). On the other hand, the unique implications of specific environmental challenges and the distinctive responses they may elicit could be obscured within such a general viewpoint (e.g., Weiner, 1992). Research from both the animal laboratory and human field studies suggests that responses to stressors can be relatively specific with respect to the particular the demands imposed, and recognition of such specificity may be key for further understanding many of the adaptive and maladaptive implications (Dohrenwend, 1998; Weiner, 1992; *see* Stress, Neural Basis of)

Overall, the challenge for research on psychological stress has been to remain true to its roots in extracting the 'essence' of psychological phenomena related to adversity, while at the same time not becoming so overly inclusive as to become essentially meaningless or to obscure attention to other more specific associations. Such an agenda raises important methodological issues, as discussed next.

Methodological Considerations

Measuring psychological stress in a reliable and valid manner has been a continuing challenge for stress researchers. For example, animal laboratory research has proven useful for more precise specification of both stimulus parameters and response characteristics associated with psychological stress. Yet the methods adopted for such research in the past often constrained the array of possible responses the animal could make, creating unnatural environmental conditions. For instance, early studies on stress commonly subjected animals to extreme stressors that were not representative of the kinds of challenges the animals would typically face (e.g., electric shocks). Furthermore, the animals were not able to respond in ways that they typically would in the wild (e.g., fight or flee). These artificial paradigms in turn could yield misleading information and views about how adaptation proceeds under more ecologically valid conditions (Weiner, 1992; *see* Measurement of Stress).

In terms of human field research, there has been considerable and ongoing debate about measurement procedures for assessing stressful life events. Research on self-report life event checklists, such as the SRE, has revealed serious psychometric limitations (Monroe, 2008). A major concern with self-report life event checklists is the problem of intracategory variability (Dohrenwend, 2006). This refers to the inconsistencies in how respondents interpret and respond any particular life event item on such checklists. For example, one person might interpret the event of 'Serious illness of a close family member' being fulfilled if his or her child had a case of the flu, another if

a distant relative had a heart attack, and finally another if a first-degree family member had a stroke. In this example, the myriad ways in which a person might interpret 'serious' and 'close' leads to an unacceptably large range of response alternatives. Many if not most listed items can produce such undesired and uncontrolled variation. The use of naïve research designs with self-report checklists, too, frequently has led to spurious findings. For example, a depressed person actually may cause some life events to occur, such as 'trouble at work,' and 'difficulties with spouse.' This means that instead of the life event stress causing depression, the reverse is true: The depression causes the life event. Furthermore, many of the life events on the SRE and similar checklists are direct indicators of psychological disorder or illness (i.e., 'major change in eating habits, major change in sleeping habits'). This means that if the study relies only upon the self-report checklist completed by the study participant, the resulting measure of stress will be confounded with disorder and thereby of little scientific value. Although more refined and comprehensive self-report measures have been developed in an attempt to address such concerns, fundamental problems with this particular methodology remain (Dohrenwend, 2006; Monroe, 2008).

In response to these measurement concerns about self-report life event checklists, other stress researchers have developed semi-structured interviews. These interview-based measures provide detailed inquiry into the lives of the participants, and adopt explicit guidelines, decision rules, and operational criteria for defining and rating life events (Brown and Harris, 1989; Dohrenwend, 2006). In general, utilizing these procedures enhances the reliability of life event assessments and provides better prediction of outcomes. Although these approaches are more time and labor intensive, they represent among the best available tools for measuring psychological stressors. Importantly, too, these approaches provide flexibility in measuring life stress, allowing for ratings of specific dimensions of stress (e.g., loss, threat, danger, humiliation) and for taking into account the influences of moderators of stress impact. Thus, in theory, such approaches can help to bridge the gap between a strictly stimulus perspective, and to incorporate useful features of the response and organism–environment interaction perspectives (Monroe, 2008).

Recent Advances and Future Directions

Over recent times it has become increasingly evident that life stress and other psychosocial factors influence susceptibility to a variety of mental and medical conditions (Cohen et al., 2007). This emerging consensus has directed the field toward identification of the behavioral and biological mechanisms via which stress influences disease susceptibility. Broadly speaking, these collective research activities have produced a wealth of research findings substantiating how stress impacts neuroendocrine, immunologic, and genetic dynamics in ways that, in theory, can lead to illness. These avenues of investigation hold considerable promise for future research addressing the central question of how environmental adversity 'gets under the skin' (Miller et al., 2009). We point to three emerging areas of interest that will likely be of importance over the coming years.

One conceptual development that has been part of these recent advances is a shift in thinking about the nature of stress and its effects on major biological regulatory systems of the body. The concept of *allostasis* has been advanced to take into account how the organism achieves stable biological functioning over time despite ongoing adaptive challenges and change. Allostasis enlarges theoretical attention to take into account a broader range of whole-organism regulatory mechanisms that are recruited in adaptation to constantly changing environmental circumstances. The cumulative burden on the body's major regulatory systems of repeated or poorly controlled allostatic activation over time is referred to as *allostatic load*. In other words, allostatic load represents the cumulative 'wear and tear' on the major body systems resulting from either too much stress or from poor regulatory management of stress, and is thought to be a key part of the process that results in vulnerability to disorder and disease. Allostasis and allostatic load have figured prominently in recent research neuroendocrine and immune mechanisms involving a wide range of conditions, but especially coronary heart disease, obesity, diabetes, and the metabolic syndrome (McEwen and Wingfield, 2003; Monroe, 2008).

A second significant trend over the past decade has been the integration of psychological stress and genetic information in relation to health and well-being. With the completion of the human genome project in 2003, the aspiration to study gene-by-environment interactions in relation to health and disease became readily feasible and fashionable. Psychological stress, as a key conceptual bridge linking variation in environmental adversity with gene function, has been a core construct in much of this research. A landmark article by Caspi et al. (2003) reported that individuals with one or two short alleles of the serotonin transporter gene promoter polymorphism (5-HTTLPR) particularly were susceptible to depression if they have been exposed to adversity early in life or to recent stressful life events. The Caspi et al. (2003) findings have helped to spawn a new era of research on life stress and genetic predisposition (i.e., gene-by-environment interactions) in relation to a wide variety of psychological and physical conditions. One very important lesson learned from this rapidly developing literature is that care must be taken to appropriately define and measure life stress if the hypothetical interactions with genetic vulnerability are to be detected reliably (Monroe and Reid, 2008; Uher and McGuffin, 2010).

Finally, the conceptual developments involving allostasis and the attention to gene-by-environment interactions point to the potential importance of greater specificity in thinking about, and assessment of, psychological stress in future research. For instance, early childhood adversity is well recognized as an important risk factor for a wide array of medical and mental problems later in life. How early adversity impacts development, particularly with regard to allostatic load and ongoing psychological stress, may be a key topic for understanding how psychological stress can influence disease susceptibility (Miller et al., 2009). With regard to gene–environment research, it is likely that different forms of psychological stress may be important for particular genetic vulnerabilities to specific diseases. Thus, for example, chronic forms of stress may be important for particular types of genetic predisposition and disease, whereas acute major life events

may be especially important for activating specific genetic predisposition to other kinds of conditions. Much as attention to proper definition and assessment of psychological stress has been a lesson learned from the recent gene–environment interaction literature, care in defining and assessing particular forms of psychological stress will be important in the coming years as more sophisticated genetic and biological multifactorial models are investigated.

See also: Burnout, Psychology of; Coping across the Lifespan; Coping with Job Stress; Coronary Heart Disease: Psychosocial Aspects; Emotion, Neural Basis of; Environmental Stress and Health; Occupational Stress Management; Post-Traumatic Stress Disorder; Self-Regulated Learning; Social Integration, Social Networks, and Health; Stress at Work; Stress in Infancy and Early Childhood: Effects on Development; Stress, Coping and Health.

Bibliography

Brown, G.W., Harris, T.O., 1989. Life Events and Illness. Guilford Press, London.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., et al., 2003. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386–389.

Cohen, S., Janicki-Deverts, D., Miller, G.E., 2007. Psychological stress and disease. Journal of the American Medical Association 298, 1685–1687.

Dohrenwend, B.P., 1998. Overview of evidence for the importance of adverse environmental conditions in causing psychiatric disorders. In: Dohrenwend, B.P. (Ed.), Adversity, Stress, and Psychopathology. Oxford University Press, New York, pp. 523–538.

Dohrenwend, B.P., 2006. Inventorying stressful life events as risk factors for psychopathology: toward resolution of the problem of intracategory variability. Psychological Bulletin 132, 477–495.

Holmes, T.H., Rahe, R.H., 1967. The social readjustment rating scale. Journal of Psychosomatic Research 11, 213–218.

Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. PLoS Medicine 7 (7), e1000316. http://dx.doi.org/10.1371/journal. pmed.1000316.

Kellogg, T.H., 1897. A Textbook of Mental Diseases. Wood, New York.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Lazarus, R.S., 2000. Toward better research on stress and coping. American Psychologist 55, 665–673.

McEwen, B.S., Wingfield, J.C., 2003. The concept of allostasis in biology and biomedicine. Hormones & Behavior 43 (1), 2–15.

Miller, G.E., Chen, E., Cole, S.W., 2009. Health psychology: developing biologically plausible models linking the social world and physical health. Annual Review of Psychology 60, 501–524.

Monroe, S.M., 2008. Modern approaches to conceptualizing and measuring human life stress. Annual Review of Clinical Psychology 4, 33–52.

Monroe, S.M., Reid, M.W., 2008. Gene-environment interactions in depression: genetic polymorphisms and life stress polyprocedures. Psychological Science 19 (10), 947–956. http://dx.doi.org/10.1111/j.1467-9280.2008.02181.x.

Rees, W.L., 1976. Stress, distress, disease. British Journal of Psychiatry 128, 3–18.

Selye, H., 1976. The Stress of Life, second ed. McGraw-Hill, New York.

Sontag, S., 1978. Illness as Metaphor. Farrar, Straus and Giroux, New York.

Uher, R., McGuffin, P., 2010. The moderation by the serotonin transporter gene of environmental adversity in the etiology of depression. Molecular Psychiatry 15, 18–22.

Weiner, H.W., 1992. Perturbing the Organism: The Biology of Stressful Experience. The University of Chicago Press, Chicago.

Strikes

John Kennan, University of Wisconsin-Madison, Madison, WI, USA; NBER, Cambridge, MA, USA; and IZA, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Franzosi, volume 22, pp. 15204–15208, © 2001, Elsevier Ltd.

Abstract

Strikes involve an apparent waste of resources that might otherwise have been divided between the employer and the workers. Although it has been argued that strikes might be caused by political conflicts within unions, or as a means to demonstrate the union's toughness with an eye to future negotiations, economic models of strikes generally view the apparent waste as a necessary cost of credibly transmitting private information regarding the actual size of the surplus available to be divided. For example, an employer who knows that sales prospects are poor may find it difficult to convince the union that this is the truth, as opposed to self-serving cheap talk. A strike then serves to verify the truth of the employer's claim, since if sales prospects are in fact good, the employer will be anxious to settle, even on terms favorable to the union. Empirically, strikes are rare, and in recent years strike activity has diminished considerably. The level of strike activity has been shown to be sensitive to fluctuations in the general level of economic activity, with fewer strikes occurring during recessionary periods. Collective bargaining legislation also has been shown to have some influence on the level of strike activity, and this is especially true of legislation constraining the use of replacement workers, depending on whether the replacements are temporary or permanent.

In perfectly competitive markets, wages are fully determined by supply and demand, and it is futile for workers to attempt to raise their wages by threatening to strike. But labor markets rarely conform to the competitive ideal, and it is generally the case that the value of labor to the employer is greater than the workers' opportunity cost, so that there is a surplus to be divided between the worker and the employer. A strike is a suspension of production while both the sides argue about how to divide this surplus.

Strikes are generally viewed by economists as a wasteful activity. After the strike is eventually settled, it is obvious that agreement could have been reached without any suspension of productive activity, on terms that would have been more favorable for both sides. Indeed, if one had a theory capable of predicting how long the strike would last, and what the settlement terms would be, the theory would be self-contradictory, in the sense that the employer and the union would both do better by reaching an agreement without a strike, rather than follow the predictions of the theory. This has been called the Hicks Paradox (Kennan, 1986), since it is implicit in Hicks's analysis of the relationships between anticipated strike durations and settlement terms guiding the employer's decisions of how much to concede to the union, and the union's decisions about how long to resist. Hicks gave an insightful discussion of the conditions under which some strikes will nevertheless occur:

> Weapons grow rusty if unused, and a Union which never strikes may lose the ability to organize a formidable strike, so that its threats become less effective. The most able Trade Union leadership will embark on strikes occasionally, not so much to secure greater gains on that occasion (which are not very likely to result) but in order to keep their weapon burnished for future use, and to keep employers thoroughly conscious of the Union's power.
>
> Under a system of collective bargaining, some strikes are more or less inevitable for this reason; but nevertheless the majority of actual strikes are doubtless the result of faulty negotiation …. Any means which enables either side to appreciate better the position of the other will make settlement easier; adequate knowledge will always make a settlement possible. The danger lies in ignorance by one side of the other's dispositions, and in hasty breaking off of negotiations.
>
> …
>
> Yet conciliation, however intelligently operated, cannot prevent strikes altogether. There will still be some strikes necessary to keep the Union "in training", and further and more important, there remains the possibility of a difference of opinion between the Union leaders and their rank and file. The leaders may be convinced that they have got the best that could be got by any method, but they may fail to convince their supporters.
>
> Hicks, 1932, The Macmillan Company

This points to three possible lines of thinking along which an economic theory of strikes might be developed. First, it might be rational for the union to endure a costly strike this time in order to establish a credible strike threat that can be used to extract concessions from the employer in future negotiations. Second, the union is a political entity rather than a single agent, and there may be differences of opinion among the union members regarding acceptable terms of settlement. Third, either side might have private information that cannot be credibly transmitted to the other side without going through a costly work stoppage; in effect, a claim by one side that the pie is actually smaller than the other side thinks can be proved only by demonstrating a willingness to forgo some known fraction of the pie.

The first of these lines has not really been pursued in the literature, while the second was developed by Ross (1950), and used by Ashenfelter and Johnson (1969) to provide an informal basis for an empirical analysis of strike activity. But the main developments in the more recent literature have been based on the idea that private information is the fundamental cause of strikes. Indeed, Hicks's remark that "adequate knowledge will always make a settlement possible" applies to all three of the theoretical approaches just mentioned.

The main idea underlying private information theories of strikes can be illustrated using a simple example. Suppose

a union negotiates a contract with an employer who has private information about the value of the workers' labor. From the union's point of view this value, V is a random variable with a distribution F that is assumed to be common knowledge; on the other hand the employer knows the realized value of this random variable, which always exceeds the union workers' opportunity cost, w_0. The employer's informational advantage may be offset by giving the union the power to make offers that the employer must accept or reject. If the union demands wage w, the probability of acceptance is $1 - F(w)$, so the union's optimal demand, w^* satisfies the condition

$$w^* = \arg\max_w w(1 - F(w)) + w_0 F(w)$$

Suppose that the support of F lies in the interval (a, b), with $w_0 \leq a$, so that there is a positive surplus to be shared between the employer and the union for any realization of the random variable V. It is easy to see that the union should make a demand such that some employer type is indifferent between acceptance and rejection. In the simplest case, V takes just the two values b and a, with probabilities p and $1 - p$. If the union demands the high value, the expected payoff is $pb + (1 - p)\, w_0$, while if the union demands the low value, it will be accepted for sure. Thus, it is optimal to make the more aggressive demand if

$$pb + (1 - p)w_0 \geq a$$

If this condition holds, there is a strike whenever the realized value of V is low. This strike condition can be rewritten as

$$p \geq \frac{a - w_0}{b - w_0}$$

Thus, the union's decision on whether to use the strike threat is influenced on the one hand by the likelihood p that the employer's valuation is high, and on the other hand by the difference between high and the low valuations (both measured relative to the workers' opportunity cost w_0). A large spread between a and b means that the firm's private information is relatively important, so that a strike threat is worthwhile even if the probability of acceptance is low. The right side of the above inequality gives a threshold for p below which the strike threat is not worthwhile. In this case there is a 'pooling' outcome, in the sense that the wage settlement is the same for both employer types. When p rises above the threshold it is optimal to threaten a strike, but further increases in p reduce the probability that a strike will actually occur. In this case there is a 'screening' outcome: the high employer type agrees to a high wage, while the low type rejects the union's demand. This simple model is analyzed in more detail in Kennan (1986).

When the firm's valuation has a continuous distribution, the union's optimal policy again might involve either pooling or screening. If there is an interior solution, it must satisfy the first-order condition

$$(w^* - w_0)h(w^*) = 1$$

where h is the hazard function, $h(w) = \frac{f(w)}{1-F(w)}$. For example, if the distribution of V is exponential, so that $1 - F(w) = \exp(\theta(a - w))$, then $h(w) = \theta$ for $w \geq a$. In this constant–hazard case there is only one possible interior solution, and the optimal demand is given by

$$w^* = \max\left\{w_0 + \frac{1}{\theta}, a\right\}$$

If θ is large, the distribution of V is concentrated near a, so there is not much to be gained by making an aggressive demand, and the pooling solution is optimal. Otherwise the union demands $w_0 + \frac{1}{\theta}$, and the probability of a strike is $1 - \exp(\theta(a - w_0) - 1)$.

Another natural example is the case where V has a uniform distribution on (a, b). Then the first-order condition is

$$\frac{w^* - w_0}{b - w^*} = 1$$

Thus, the optimal demand is given by

$$w^* = \max\left\{\frac{1}{2}w_0 + \frac{1}{2}b, a\right\}$$

In this case the probability that the employer accepts the union's demand is

$$q = \frac{1}{2}\frac{b - w_0}{b - a}$$

Bulow and Roberts (1989) show that the analysis of bilateral exchange with private information is closely related to standard models of monopoly pricing. Consider a monopolist facing a linear demand curve

$$p = b - (b - a)Q$$

with constant marginal cost c, and a capacity constraint $Q \leq 1$. The monopolist equates marginal revenue and marginal cost, unless marginal revenue is still above marginal cost at the capacity constraint, in which case the optimal output is $Q = 1$.

Equating marginal revenue and marginal cost means choosing Q so that $c = b - 2(b - a)Q$, with $p = \frac{1}{2}c + \frac{1}{2}b$, while the price at the capacity constraint is $p = a$. Thus the optimal price is

$$p^* = \max\left\{\frac{1}{2}c + \frac{1}{2}b, a\right\}$$

which matches the solution of the union's problem, with $c = w_0$. Indeed, the union acts as a monopoly seller of labor, facing a tradeoff between a low wage and a high quantity, or the reverse, where the quantity is the probability that the employer will accept the union's wage demand.

A much more extensive analysis of bargaining models with private information, allowing for private information on both sides, is presented in Kennan and Wilson (1993), with applications to legal conflicts as well as labor negotiations. A general treatment of games with private information, with many bargaining examples, can be found in Myerson (1991). Kennan and Wilson (1993) discuss three varieties of private information models. The simple examples above are screening models, which become more interesting when the union makes a sequence of offers (with a delay period between offers), rather than a single 'first and final' demand. When the informed party can make offers (or counteroffers) additional complications arise, because the offer that is made conveys a signal of the private information underlying the offer. In contrast to these screening and signaling models, attrition models treat the surplus to be divided as an indivisible object, with each side trying to convince the other that it can tolerate a longer strike; eventually the weaker side concedes and the other side takes the

whole surplus (although a large fraction of the surplus is generally lost in a protracted conflict).

In a series of papers, beginning with Cramton (1992) and Cramton and Tracy (1992), and summarized in Cramton and Tracy (2003), Cramton and Tracy developed a private information model of strikes and 'holdouts,' which occur when the old contract expires without agreement on a new one, but work continues without a strike until a new agreement is reached (and the new contract is backdated to the expiration of the old contract). Cramton and Tracy (1992) show that holdouts are empirically important, and that they can be interpreted as the outcome of a signaling game which is like a strike except that the stakes are smaller. The crucial assumption is that, although the opportunity cost of a strike is privately known to the firm, it is common knowledge that the opportunity cost of a holdout is proportional to the opportunity cost of a strike. In effect, the parties have a way of using a small-scale conflict to reveal information which could otherwise lead to a strike.

The 'joint cost' model of strikes, due to Reder and Neumann (1980) and Kennan (1980), simply postulates that whatever the negotiation process might be, it is reasonable to suppose that there will be less strike activity with strikes being more costly, where the relevant measure of cost is the joint loss suffered by both sides taken together. This simple model has been used quite extensively to guide empirical work, although in practice it is often difficult to isolate differences in joint costs in a sample of contract negotiations, and in particular it is difficult to isolate variables that affect the cost of a strike without affecting the size of the surplus.

Empirical Analysis of Strike Activity

Strikes are rare. In 2013, there were only 15 strikes in the United States involving 1000 workers or more. The total number of workers involved in these strikes was about 54 500, and the number of work-days lost was about 289 900. Most of these strikes involved local government workers. There were only 6 strikes in the private sector, with a total of 13 200 workers involved, and a loss of 39 500 work days. There are about 11 million workers employed in establishments with 1000 workers or more, so the chance of being involved in a strike (in the US private sector in 2013) was about one in a 1000.

Figure 1 shows the number of work-days lost due to large strikes in the United States and Canada over the period 1946–2013 (a large strike in Canada means a strike involving 500 workers, so the levels of the two series are not comparable). Strike activity has declined quite dramatically, especially in the United States, although one large strike can cause a big jump in the data (as in the case of the 6-month strike by 135 000 actors involved in making TV commercials in the United States in 2000). Although it is difficult to obtain comparable data for many countries, the decline in strike activity over the last 40 years or so seems to have been quite widespread.

Fluctuations in Strike Activity

There is an extensive empirical literature analyzing the relationship between strike activity and business cycle fluctuations. The main conclusion from early work in this area is that strikes are more frequent when general economic conditions are good. Although this conclusion is supported by a considerable body of evidence, it is of limited interest because it lumps together strikes of all sorts, including many minor disputes that occur during the term of ongoing labor contracts. More recent work has attempted to determine whether economic conditions affect the incidence of contract strikes. This work largely relies on North American data, because unions in the United States and Canada generally negotiate contracts covering clearly defined periods of a few years, so that one can count the

Figure 1 Strike activity.

number of negotiations that might lead to a strike, and use this to measure strike incidence.

The empirical results on strike incidence are well summarized by Card (1990); some more recent findings are reviewed by Cramton and Tracy (2003). Surprisingly, the evidence indicates that strike incidence and duration move in opposite directions over the business cycle. Strike incidence is generally found to be procyclical, although the relationship between strikes and general economic conditions is not strong enough to dominate other sources of variation, so that a long time series is needed to establish the result. Although less work has been done on cyclical movements in strike duration, there is solid evidence that duration moves counter-cyclically. Some attempts have also been made to distinguish between the effects of cyclical fluctuations in product markets and in labor markets. There is no clear pattern in these results, and there is no clear theoretical reason for making the distinction in the first place.

There is also an old empirical literature, starting with Yoder (1938), showing pronounced seasonal patterns in strike activity. The evidence on seasonality is discussed by Kennan (1986), who pointed out that although the seasonal fluctuations are large there has been no serious attempt to explain them, a comment which apparently still applies in 2014.

Effects of Collective Bargaining Legislation

Strike activity is clearly affected to a large extent by laws governing the tactics available to workers and employers as they negotiate over the division of the employment surplus. This is a large subject in itself, which is not dealt with here; Cramton and Tracy (2003) give an overview of some of the main issues with respect to North American legislation, and Campolieti et al. (2014) analyze the empirical relationships between variations in collective bargaining laws and strike incidence and duration across Canadian provinces.

One topic worth considering briefly is the use of replacement workers. As Schnell and Gramm (1994) point out, it is important to distinguish between temporary replacements, used to maintain production during the course of a strike, and replacement workers who permanently displace the striking workers from their jobs. Gunderson and Melino (1990) show that prohibiting replacement workers (in Quebec) increases strike duration, and also increases strike incidence (Table 3). They also argue that replacement workers reduce the joint cost of a strike, so prohibiting replacements should reduce strike incidence according to the joint cost theory. But since the Quebec prohibition (which came into effect in 1977) covered both temporary and permanent replacements, there is no way to disentangle the separate effects of these variables. A reasonable interpretation is that temporary replacements reduce the cost of a strike, while permanent replacements reduce the size of the surplus to be divided between the employer and the union.

The analysis of the effects of replacement workers in Kennan and Wilson (1989) is concerned with strike duration, rather than incidence. The availability of replacement workers truncates the distribution of employer valuations from above, so if replacement workers are banned, longer strike durations are needed to reveal the firm's valuation (because strikes continue beyond the point where it has been established that the valuation is above the truncation point).

The same logic implies that prohibiting replacement workers increases the incidence of strikes. In the case of the exponential distribution discussed above, if there is an interior solution without replacement workers, then the wage demand is

$$w^* = w_0 + \frac{1}{\theta}$$

and the strike incidence is

$$1 - \exp\left[-\theta(w^* - a)\right]$$

If the truncation point $\widehat{w}$ is greater than w^*, there is no effect on strike incidence: the union makes the same demand regardless of whether replacements are allowed, and there is a strike if $V < w^*$.

If $\widehat{w}$ is less than w^*, then the optimal demand when replacement workers are permitted is $\widehat{w}$, and there is a strike only if the firm's valuation is below this point. But this means that the replacement workers are irrelevant – the employer would not hire them in any case (although the union does not know this). When replacements are not allowed, there are also strikes when $V \in (\widehat{w}, w^*)$. Thus allowing replacement workers increases the incidence of strikes.

See also: Labor Unions; Minimum Wages.

Bibliography

Ashenfelter, O., Johnson, G.E., 1969. Bargaining theory, trade unions, and industrial strike activity. The American Economic Review 59, 35–49.

Bulow, J., Roberts, J., 1989. The simple economics of optimal auctions. The Journal of Political Economy 97, 1060–1090.

Campolieti, M., Hebdon, R., Dachis, B., 2014. The impact of collective bargaining legislation on strike activity and wage settlements. Industrial Relations: A Journal of Economy and Society 53 (3), 394–429.

Card, D., May 1990. Strikes and bargaining: a survey of the recent empirical literature. American Economic Review 80, 410–415.

Cramton, P., Tracy, J.S., 2003. Unions, Bargaining and Strikes. Edward Elgar Publishing, Cheltenham, UK, pp. 86–117 (Chapter 4).

Cramton, P.C., 1992. Strategic delay in bargaining with two-sided uncertainty. Review of Economic Studies 59 (1), 205–225.

Cramton, P.C., Tracy, J.S., March 1992. Strikes and holdouts in wage bargaining: theory and data. American Economic Review 82, 100–121.

Gunderson, M., Melino, A., 1990. The effects of public policy on strike duration. Journal of Labor Economics 8 (3), 295–316.

Hicks, J.R., 1932. The Theory of Wages. MacMillan, London. Reprinted in 1957, by Peter Smith, Gloucester, MA.

Kennan, J., 1980. Pareto optimality and the economics of strike duration. Journal of Labor Research 1 (1), 77–94.

Kennan, J., 1986. The economics of strikes. In: Ashenfelter, Orley, Layard, Richard (Eds.), Handbook of Labor Economics, vol. II. Elsevier Science Publishers, Amsterdam.

Kennan, J., Wilson, R., December 1989. Strategic bargaining models and interpretation of strike data. Journal of Applied Econometrics 4 (Suppl.), S87–S130.

Kennan, J., Wilson, R., 1993. Bargaining with private information. Journal of Economic Literature 45–104.

Myerson, R.B., 1991. Game Theory. Harvard University Press, Cambridge MA.

Reder, M.W., Neumann, G.R., 1980. Conflict and contract: the case of strikes. The Journal of Political Economy 867–886.

Ross, A.M., 1950. Trade Union Wage Policy. University of California Press.

Schnell, J.F., Gramm, C.L., 1994. The empirical relations between employers' striker replacement strategies and strike duration. Industrial and Labor Relations Review 47, 189–206.

Yoder, D., 1938. Seasonality in strikes. Journal of The American Statistical Association 23, 687–693.

Strong Program, Sociology of

David Bloor, University of Edinburgh, Edinburgh, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Strong Program in the sociology of knowledge is a set of methodological requirements about the proper way in which social scientists and historians should analyze and explain knowledge, including scientific knowledge. Such investigations should be causal, impartial, symmetrical, and reflexive. The article explains and illustrates these requirements. The basis of the argument in their favor is that these requirements are expressive of a truly scientific orientation in which knowledge is treated as a natural phenomenon. The article ends with a brief introduction to some case studies from the history of science that show the Strong Program in action.

The 'Strong Program' is a set of methodological requirements about the proper way to conduct sociological enquiry into the nature of knowledge, including scientific knowledge. The program enjoins the investigator to adopt four requirements: (1) causality, (2) impartiality, (3) symmetry, and (4) reflexivity. The requirement of 'causality' means that the aim of the sociologist is not merely descriptive but also explanatory. What brings about and sustains a body of knowledge? 'Impartiality' means that, for the purposes of the investigation, the investigator disregards his or her sense of whether the beliefs under scrutiny are true or false, or rational or irrational. Equal curiosity should be expended on both sides of the evaluative divide. This does not mean that an investigator should have no opinions. The requirement is that these opinions should not, for example, lead them to study one side of a scientific dispute while leaving the other side unexamined because it seems right or obvious. 'Symmetry' means that this equally distributed curiosity should issue in the same general kinds of sociological explanation regardless of how the knowledge is evaluated. All beliefs confront the same problems of credibility and depend on the same contingencies. True beliefs have no more intrinsic credibility than false ones. Finally, 'reflexivity' means that sociologists of knowledge must not claim access to any transcendent standpoint. Put simply, no sociological theory of knowledge is acceptable unless it applies to itself. These requirements all stem from the desire that the approach of the sociologist of knowledge should be scientific in character. It should embody the same naturalistic spirit that now routinely informs work in such fields as physics, engineering, and biology. Despite its scientific motivation, the Strong Program is widely perceived by critics as being antiscientific. To offset such misunderstandings, each of the four requirements calls for amplification and defense.

Causality

The sociological causes at work in the creation of a body of knowledge shared by the members of a group range from the traditions, customs, conventions, and institutions of the group to the current problems and specific interests of its members. A sociologist of science will attend to the role of different disciplinary and professional traditions, the presence of dominant or competing research schools, and the interactions between scientists and other significant institutions such as the government, the church, and the military as well as the institutions by which knowledge is transmitted, such as universities and professional bodies.

The commitment to causality does not mean that sociologists believe that only sociological causes are at work. It is self-evident that all belief systems, whether scientific or nonscientific, are the achievements of biological entities (namely, human animals) that interact with one another as social animals and who also (and simultaneously) interact collectively with their material environment. People have their eyes open, and their sensory inputs influence belief just as do their susceptibilities to one another as social animals. Unless human agents had effective sense organs, as well as a range of innate propensities, they could neither create nor transmit a shared culture. Knowledge cannot be 'purely' social and no responsible sociological analysis has ever carried such an implication.

It follows that sociologists can only contribute part of a complete explanation of cognition. Sociologists must always work alongside other specialists, such as psychologists. Conversely, no adequate account of human knowledge can ever proceed without due acknowledgment of the role of the social. The 'strength' of the program lies in the necessity, not the sufficiency, of the social dimension.

The False Charge of Idealism

The idea that material reality plays no role in human knowledge is called 'idealism,' that is, the metaphysical doctrine that there is no material world that can constitute an independent object of knowledge. Idealists fuse the subject and object of knowledge together and treat both as spiritual, or mental, in nature. The accusations of idealism that have been directed at the Strong Program are incorrect. They have arisen through the misrepresentations of the critics. For example, Bruno Latour (1993) says the Strong Program is one of a family of theories based on the idea that knowledge has two ingredients, 'nature' and 'society,' which trade off against one another in a zero-sum game – the more of the one ingredient meaning the less of the other. Viewing the Strong Program from within this framework, Latour

 http://dx.doi.org/10.1016/B978-0-08-097086-8.85033-3

presents it as an extreme stance where one of the ingredients, nature, actually makes zero contribution. The 'strength' of the program is thus wrongly said to reside in the claim that knowledge (and indeed reality) is purely social.

This false reading is widespread. Thus, some critics implicitly accept a zero-sum picture but insist that nature plays some role rather than no role. It is assumed to be an empirical question to decide just how much influence is exerted by nature in comparison to that exerted by society. A variant of this argument underlies the claim that the Strong Program is incompatible with acknowledging that the natural sciences have a good record of predictive success. The accusation rests on the premise that if scientists were only attending to what is going on in society, and were not attending to what is happening in nature, their ability to make predictions would be unintelligible. The inference is correct, but the premise is false. The assumption behind the Strong Program is that scientists understand nature collectively, through society, not in spite of it. An analogy may be of help in seeing where the critics have gone wrong. Society no more interferes with knowledge than the eye interferes with vision. The eye, like society, operates as the relevant organ of cognition and as a necessary constituent of it. No one asks 'how much' of our visual experience is contributed by the eye, and 'how much' by the object seen. We might ask 'what' is contributed, but not 'how much.' The latter question makes no sense, and neither does the allegedly 'empirical' question of 'how much' of knowledge depends on nature and 'how much' on society.

The Role of the Social

How do social processes enable scientists to have access to the world? The answer given many years ago by Thomas Kuhn (1962) illustrates the kind of answer available to the sociologist. Kuhn argued that cognition within a group of scientists is coordinated by reference to a shared exemplar or paradigm, that is, a concrete achievement, recognized by the group and used as a resource on which to model their creative activity. By virtue of their shared orientation toward that achievement, and their knowledge of its standing in the eyes of others, members of a group endow it with a normative status. The paradigm indicates the proper way to proceed. It will be evident that something is constituted as a paradigm by being accepted as a paradigm – in the way that money is constituted as money by being regarded as money. In these self-referential and self-reinforcing ways, society gets woven into the fabric of knowledge and into the struggle to make sense of intransigent, real-world processes.

Impartiality and Symmetry

The reference to Kuhn also points to the correct way to understand impartiality and symmetry. Kuhn did not say that the idea of a paradigm merely helps the historian to make sense of old-fashioned theories. He applied the idea to all cases of shared knowledge whether past or present, or whether accepted or rejected by the historian. He was searching for invariant, social principles of scientific cognition. Thus, it is always true that a shared interpretation of the world has to be sustained by some mechanism of social coordination and that mechanism must work in the face of continuing difficulty, anomaly, and threat of fragmentation. Every shared approach to nature must overcome the problem that our experience can be given more than one interpretation. Every cultural institution requires that the behavior of its participants be aligned rather than allowed to lapse into individualistic anarchy. To understand the ways in which these ubiquitous problems are confronted and solved calls for a perspective that picks out the general mechanisms by which social and cultural order are maintained. This is the point of the methodological demand for impartiality and symmetry.

Reflexivity

In the eyes of its critics, the requirement that the Strong Program must be applicable to itself immediately raises the specter of self-refutation. If the claims of the sociologist themselves transcend sociological determination, then the sociologist is admitting that the sociology of knowledge is false. If the claims of the sociologist are not exceptions, but are also socially determined, then why should anyone believe what the sociologist says? The claims will be mere ideology or the product of social bias. The basis of this objection goes back to the theme of causation. Reflexivity looks self-refuting, but only if it is tacitly assumed that the sociologist's vision of knowledge sees knowledge as wholly the product of custom, tradition, prejudice, or ideology, that is, treats it as a collective fantasy. Once it is accepted that society is a necessary ingredient of all knowledge, including science, then the role of society can be studied in a matter of fact manner. It can no longer be assumed that sociological causation operates as a fatal contaminant.

Ideological Limitations on Causality

In adopting the requirements of causality, impartiality, symmetry, and reflexivity, the sociologist or historian of science faces a major obstacle. These requirements seem to run counter to deep-rooted aspects of common sense. In general, human curiosity has a markedly partial and asymmetric structure. Social actors typically confine the request for causal explanation to the breakdowns in their routine expectations or evaluations. When things go smoothly, they do not ask why, because they typically feel that nothing more needs to be said. This common-sense asymmetry makes sense in the context of daily life, but sociologists are seeking general results and are working within a broader context. Their professional curiosity is symmetrical, not asymmetrical like common sense.

For these reasons, the Strong Program is in direct opposition to the conclusions of philosophers such as Imre Lakatos (1971) who have sought to recast common-sense asymmetry into a set of general methodological guidelines for the writing of the history and sociology of science. These writers require the analyst to identify what is deemed to be 'rational' in a body of scientific work and treat this as a baseline. They encourage the historian and social scientist to think that the normal workings of causality do not apply to what is rational, or that qualitatively different kinds of causes are in operation when things

proceed rationally compared with when there is a deviation from rationality. Although this is not their terminology, they explain rational behavior teleologically, by appeal to 'final causes,' and deviations from rationality in terms of 'efficient causes.' Remarkably, some philosophers have even taken this argument to the extreme of claiming that while illusory perception calls for causal explanation, veridical perception does not. No psychologist or physiologist would approach perception in this way – just as no medical scientist would say there are causes for disease but no causes for health, and no engineer would say there are causes for why bridges collapse but none for why they stay up. But if these specialists can overcome the assumptions and asymmetries of everyday life and give their professional curiosity a locally symmetrical character, then so can the sociologists.

The universality of the causal perspective, and the need to avoid placing arbitrary restrictions on the scope of causality, should be self-evident from a scientific standpoint. How can philosophers oppose this when, at the same time, they see themselves as defending science against its detractors? Anthropologists have long known the answer to this question. The selective use of causality is a technique widely used to make the basic norms and beliefs of a culture appear distinct from nature and hence take on a form that renders them deserving of special respect.

Durkheim (1915) drew attention to the important distinction between the sacred and the profane and traced the way in which groups protected their norms and institutions by treating them as sacred. To be sacred was to be untouchable, and that means untouchable by the mundane categories of cause and effect. It appears that these techniques are not confined to members of the traditional societies whose culture interested Durkheim. Such strategies are alive and well in the writings of philosophers of science. Nevertheless, the strategy of 'defending' science by treating it as sacred is a strange one. Today's science has no place in it for the sacred so, in thus defending it, the philosophers are betraying it.

Crucial Experiments

One of the most profound insights into the working of science was furnished by the physicist Pierre Duhem in his classic *Aim and Structure of Physical Theory* (1914). Duhem argued that there is no such thing as a crucial experiment and, in developing this claim, he uncovered important general truths about the nature of meaning and reference.

Consider two rival theories T1 and T2, both of which are capable of explaining a body of experimental results. How can scientists decide which is correct? The standard view was that it was possible to perform a 'crucial' experiment. An experiment must be found in which T1 and T2 both make predictions regarding the outcome but where the predictions diverge, so T1 predicts E while T2 predicts not-E. Such an experiment might permit the elimination of one of the theories and the (cautious) retention of the other.

Duhem argued that this analysis does not do justice to the real predicament, or the real resources, of the experimenter. In reality, argued Duhem, no theory can be tested in isolation from a mass of background assumptions (A). Some of these assumptions can be made explicit, but many will be tacit or unknown. Once the role of the background is brought into the picture, then it is clear that what is under test is not T1 or T2, but a logical conjunction of theory and background assumption, that is, (A&T1) or (A&T2). A false prediction, say by T1, thus carries the implication, not that T1 is false, but that either T1 is false or A is false. This makes every refutation equivocal. It means that T1 might be reconciled with the experimental outcome by altering the background assumption A, for example, by blaming the outcome on impurities in the chemicals, or faults in the wiring, or inadequate experimental shielding, or hitherto unappreciated friction effects, etc.

It might look as if Duhem is merely showing how the discipline of experiment can be evaded, but the point goes much deeper. Duhem had put his finger on the basic cognitive mechanics by which a Kuhnian paradigm is elaborated in the face of an endless string of anomalies. Every anomaly could be counted as a refutation, but it is not. It is seen as a puzzle and past achievements are seen as resources for solving the puzzle. Duhem was not describing bad science, but normal science. Duhem had his own answer to how, and to what extent, this room for maneuver should be exploited. He said that, when adjusting the background assumptions, a scientist must use 'good sense.' As far as it goes this is correct. Not any attempt to rescue a theory from refutation will carry credibility. It may demand contradicting other results, and this may be felt to impose too great a cost. The rescue may not be worth it. A sociologist of science will see at once that the argument has now moved decisively into the realm of the social. When a group makes a decision about such costs, they are making decisions about the conventions and institutions they want to sustain. The voice of 'good sense' is the voice of the group.

Meaning Finitism

Duhem's argument can be generalized. It applies to every single act of concept application, whether inside or outside of science, and takes the sociologist to the heart of the phenomena of meaning, concept learning, and concept use. The 'negotiability' of scientific prediction translates into the 'negotiability' of all meaning. Unfortunately, the generality of Duhem's insight has escaped the philosophical critics of the sociology of knowledge. To see what is at stake, consider how simple words are learned and applied.

All learning must go back to a process known as 'ostensive definition.' The word 'cat' is learned by competent users pointing out a number of examples in the hope that the learner will grasp the meaning. The learner must then move from the finite number of examples in the learning set to a new application to a putative cat that has not been encountered before. The attempts will be monitored and, if necessary, corrected by the teachers. Sometimes concepts are learned by verbal definitions, but the words in the definition must be familiar and, ultimately, these will have been defined ostensively.

It is a mistake to think that ostensive definition can 'fix' meanings. It cannot, but this is not because ostensive definition is defective but because the notion of 'fixed' meanings is wrong. A concept can no more be associated with a fixed meaning than a theory can be associated with a fixed prediction. Past

instances of a concept do not determine future applications because a whole network of background assumptions and contingencies is always in play. This important principle is called 'meaning finitism.'

Meaning finitism rules out a familiar line of criticism directed at the sociology of knowledge. This criticism depends on the myth of fixed meanings and is exemplified in the work of the philosopher John Searle (1995). Searle acknowledges that a word such as 'cat' is learned on the basis of a finite number of examples, but he assumes that the language learner has the capacity to extract 'fixed' meanings from the examples. Thus,

> What counts as a correct application of the term 'cat' … is up to us to decide and is to that extent arbitrary. But once we have fixed the meaning of such terms in our vocabulary by arbitrary definitions, it is no longer a matter of any kind of relativism or arbitrariness whether representation-independent features of the world satisfy those definitions …. (Searle, 1995: p. 166)

Expanding on this theme, Searle says that how one uses the word 'cat' is up to him or her, but whether there is an object that "satisfies that use, is a plain matter of (absolute, intrinsic, mind-independent) fact" (p. 166).

But an object 'satisfies' a concept in the same way that an experimental result 'satisfies' the prediction from a theory. Had Searle taken note of Duhem's expert testimony, he would have known that 'satisfaction' is always a negotiable relationship, not something that can ever be treated as 'fixed.' Rather than shutting down sociological curiosity, as Searle thinks, the conventions of concept application are as sociologically problematic as Duhem's 'good sense.'

Relativism

One of the most frequent objections to the Strong Program is that it is 'relativist.' The supporters of the Strong Program are happy to embrace the label provided the definition of the word 'relativism' is clear and appropriate. To meet these requirements, the central idea of relativism must be disentangled from other ideas with which it is often confused. Thus, it can be accepted that sociologists of knowledge are relativists insofar as they see knowledge as a natural phenomenon, located in space and time and subject to causal laws. Knowledge will then have a contingent character that precludes it ever achieving a status that could justifiably be described as 'absolute.' Absolute cognition is a phantom and would place knowledge outside the realm of nature. Thus, if 'relativism' is simply the denial of 'absolutism,' and the rejection of absolutism is a necessary and sufficient condition for relativism, then the Strong Program is relativist, and rightly so.

Unfortunately, many self-proclaimed antirelativists reject this simple definition of relativism. They claim that they are not absolutists, but insist that this does not commit them to relativism. They claim to be neither relativists nor absolutists. Thus, the rejection of absolutism is taken as a necessary, but not a sufficient, condition for relativism. The assumption is that the definition of relativism must involve a further necessary condition. Typically, and as already indicated, this extra condition involves imputing to the relativist some form of philosophical idealism or some form of irrationalism e.g., the belief that 'anything goes'). Critics thus use an eclectic definition, but in so doing they conflate questions that should be kept separate. They run together different intellectual traditions and fail to draw obvious distinctions. The dichotomy between absolutism and relativism is not the same as the dichotomy between idealism and materialism or between rationalism and irrationalism.

Converging Evidence

It was a slogan of the Enlightenment that a rational man proportions his belief to the evidence. What do sociologists have to say about the role of evidence in science? Critics maintain that sociological causes may explain beliefs that are not warranted by evidence, but insist that the attempt to explain warranted beliefs is misguided. Knowledge of the evidence itself is sufficient to explain warranted belief.

Consider the accumulation of evidence over time and its capacity to modify belief. Suppose two rival theories T1 and T2 are adopted, respectively, by subgroups G1 and G2. Members of G1 give a high a priori probability to T1 and a low a priori probability to T2, while members of G2 do the opposite. Let it be granted to the sociologist that these a priori probabilities can be explained by observing that the theories help further the interests of the respective groups.

Now, suppose that unbiased experiments are carried out to test the consequences of the two theories. Evidence accumulates that yields a new a posteriori probability for the two theories, making, say, T1 more probable and T2 less probable in the eyes of both G1 and G2.

Will it still be true that the beliefs of the members of G1 and G2 are socially explicable or should it be said that the burden of explanation has, in some way, begun to shift from the social to the rational? There are some mathematical results, called convergence theorems, which seem to accomplish this transformation. They show how, as evidence accumulates, the a posteriori probability will eventually converge on 0 (false) or 1 (true). Evidence eventually 'washes out' the initial, socially caused, distribution of belief and, it seems, leads to new beliefs based solely on rational rather than social considerations.

Such an interpretation, however, cannot be right. Convergence theorems only work under highly restricted conditions. They require the randomized and logically independent nature of the incoming information. The theorems also apply to a limiting case, that is, to the convergence of an infinite sequence of events. These conditions are not met in reality. All empirical sequences are finite, and experiments are not performed randomly or independently. The beliefs that are the end product of experimentation are therefore no less the proper objects of sociological enquiry than the a priori probabilities at the beginning.

Inductive Strategies

The lesson to be learned is that evidence-based belief should not be set in opposition to belief caused by social processes. To drive

this point home, it will be appropriate to examine another celebrated argument, namely, Rudolf Carnap's (1952) theory of inductive confirmation. Carnap saw evidential support as a logical relation that, under suitable circumstances, was capable of rigorous definition and mathematical formulation. Here, if anywhere, would be a demonstration of the irrelevance of the social to the rational conduct of science. In fact, Carnap's analysis generated the opposite result. It turned into a demonstration of the ineradicably social character of inductive confirmation.

Carnap began by confining his analysis to a simplified model of a world with a specified number of objects and properties and a formally defined 'language' for the description of possible states of affairs. The aim was to make it possible to specify the precise logical content of any proposition, whether a hypothesis h, or a piece of evidence e. It would then be possible to measure inductive support by defining it in terms of an overlap (or partial entailment) between the states of affairs specified by the propositions in the language. These restrictions are hardly designed to be friendly to the sociological perspective, but this is what makes the exercise an interesting one. It makes it a severe test of the sociological approach.

Carnap found that there was no unique confirmation function, that is, no unique formula $c(h,e)$ that would accord a numerical value to the degree of evidential support of e for h. It transpired that there were an infinite number of possible functions that could perform this role. To accommodate this fact, Carnap introduced a parameter λ (lambda) into the confirmation function. This parameter could assume any value between zero and infinity. Each value of λ, said Carnap, represented a possible inductive strategy, that is, a greater or lesser willingness to make risky generalizations. Thus, all inductive inferences depended on an inductive strategy, and inductive strategies must be supplied by the subject who is seeking to learn about the world.

Carnap expressed the point psychologically and presented the choice of inductive strategy as a matter of individual preference. If confirmation theory is intended as a model of science, however, this individualistic formulation needs to be amended. Inductive strategies are matters of consensus not subjective preference. They are embodied in the conventions and institutions of groups of scientists. In short, λ is a social parameter and inductive inference is a socially structured form of cognition. Sociological enquiry is needed to identify the causal conditions determining the inductive strategy of any given group.

Can There Be a Sociological Analysis of Mathematics?

Can there be a Strong Program analysis of mathematics – that is, not just a sociology of mathematicians, their recruitment, organizations, status hierarchies, quarrels, and professional lapses, but one that deals with the knowledge itself? One way to appreciate the scope of this question, and the possibilities for answering it, is to imagine mathematics organized into a deductive system. In such a system, mathematical results take the form of theorems, which follow rigorously, according to clearly stated rules of reasoning, from clearly stated axioms. In reality, mathematics is much more disjointed and ad hoc but the idealized case is a useful starting point because (once again) it makes the minimum of concessions to the sociologist.

In an axiom system, there are two points of entry for sociological enquiry. First, where do the axioms come from? How are they chosen and what gives them meaning, credibility, and interest? Second, where do the rules of inference come from? What gives these rules their power to compel assent? It is always possible to give a metaphysical answer to such questions and thereby create the impression that there is no need to exercise empirical curiosity on such matters. It might be said that the axioms are 'self-evident' or that one possesses 'intuition' into the validity of basic deductive arguments. This will not satisfy those with a scientific commitment to causal enquiry. To meet this requirement, the answers will have to take a different form, namely, an appeal to psychosocial causation, that is, the form codified in the Strong Program.

The status of the basic concepts and modes of inference used in mathematics has long been a matter of professional interest for mathematicians themselves. There has been an important strand in the history of mathematics that has attempted to illuminate such problems. The fate of this foundational enterprise therefore needs to be understood before assessing the prospects for a sociological analysis of mathematics.

The development of the differential and integral calculus in the eighteenth century greatly increased the power of mathematics but its procedures lacked rigor. To put the calculus on a firm foundation, its fundamental concepts were eventually defined in terms of more basic arithmetical operations. These successes led to further questions. If the calculus is justified by arithmetic, what is the justification for arithmetic? Guiseppe Peano cast arithmetic into an axiom system and, at the turn of the twentieth century, Gottlob Frege and Bertrand Russell set out to justify these axioms in terms of yet more basic logical concepts. Despite enormous effort, they failed to achieve their aims. Mathematicians use techniques that could not be defined in terms of the concepts of logic. Worse was to follow. Logic itself turns out to be susceptible to a variety of different formulations that are inconsistent with one another. And then there is the inevitable question of the justification of the deductive steps within these systems. Can deduction be justified? The answer is no – no more than induction can be justified (see Haack, 1976).

The collapse of the Frege-Russell program, and a number of similar foundational attempts, provided the motivation for Wittgenstein's *Remarks on the Foundations of Mathematics* (1956). Wittgenstein adopted a radically new approach. He acknowledged the failure of the justificatory project and represented mathematics in a radically different manner – as a sociological phenomenon. This is the meaning of his famous, but frequently misunderstood, analysis of rule following. Wittgenstein used the example of following a simple rule for generating a sequence of numbers such as 2, 4, 6, 8, etc. What is the character of the logical compulsion to say 8 after 6, and so on? In what does the necessity of the rule consist? All the candidate explanations that have been traditionally invoked turn out to be question begging. They take the form of rules for following rules and thus fail to engage with the problem at a fundamental level. By a process of elimination, Wittgenstein

excluded all explanations apart from the causal influence exerted on one another by the members of the community of rule followers as they seek to coordinate themselves and pursue their collective interests. Wittgenstein said that reasons come to an end, but he did not mean that two things are at work, reason taking the rule follower so far, and then social processes take over. His account was not dualistic. His blunt conclusion was that following a rule is to be understood as a custom, convention, or institution. If Wittgenstein's argument is right, then the sociology of mathematics is not merely possible – it is a necessary stance if the true nature of mathematical activity is to be brought to light.

Conclusion

The Strong Program was explicitly formulated by an interdisciplinary group in Edinburgh in the early 1970s. (For the development of this approach, see Barnes et al., 1996.) The aim was to codify the assumptions behind an emerging body of case studies of scientific practice – the better to defend them against outside criticism of the kind offered by philosophers. The lifeblood of the sociology of scientific knowledge, however, lies with empirical work. To name but a small sample of work that can be read as supporting the Strong Program, see Mackenzie (1981) on the history of statistics and its connections with eugenics, Shapin and Schaffer (1985) on the disputes surrounding Boyle's experiments with the air pump, Harwood (1993) on styles of research in German genetics before 1933, Kusch (1999) on the different theories of mind generated in different psychological laboratories in the early years of the twentieth century, and Bloor (2011) on the disputes in aerodynamics in the early decades of aviation. The Strong Program continues to serve its function because the importance of the rich body of empirical work on the nature of science is still obscured by misguided criticisms directed at its methodological foundations.

See also: Cultural Studies of Science; Experiment in Science and Technology Studies; Kuhn, Thomas S. (1922–96); Physical Sciences: History and Sociology; Popper, Karl Raimund (1902–94); Reflexivity in Science and Technology Studies; Relativism: Philosophical Aspects; Science and Technology Studies, Ethnomethodology of; Scientific Controversies; Scientific Knowledge, Sociology of; Truth and Credibility in Science.

Bibliography

Barnes, B., et al., 1996. Scientific Knowledge: A Sociological Analysis. University of Chicago Press, Chicago.
Bloor, D., 2011. The Enigma of the Aerofoil: Rival Theories in Aerodynamics, 1909–1930. University of Chicago Press, Chicago.
Carnap, R., 1952. The Continuum of Inductive Methods. University of Chicago Press, Chicago.
Duhem, P., 1914, 1962. The Aim and Structure of Physical Theory (P.P. Wiener, Trans.). Atheneum, New York.
Durkheim, E., 1915, 1961. The Elementary Forms of the Religious Life (J.W. Swain, Trans.). Collier Books, New York.
Haack, S., 1976. The justification of deduction. Mind 85, 112--119.
Harwood, J., 1993. Styles of Scientific Thought: The German Genetics Community 1900–1933. University of Chicago Press, Chicago.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Kusch, M., 1999. Psychological Knowledge: A Social History and Philosophy. Routledge, London.
Lakatos, I., 1971. History of science and its rational reconstructions. In: Buck, R.C., Cohen, R.S. (Eds.), Boston Studies in the Philosophy of Science, vol. 8. North Holland Publishing Company, Amsterdam, pp. 91–136.
Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.
Mackenzie, D.A., 1981. Statistics in Britain 1865–1930. Edinburgh University Press, Edinburgh.
Searle, J.R., 1995. The Construction of Social Reality. Allen Lane, London.
Shapin, S., Schaffer, S., 1985. Leviathan and the Air Pump: Hobbes, Boyle and the Experimental Life. Princeton University Press, Princeton, NJ.
Wittgenstein, L., 1956. Remarks on the Foundations of Mathematics (G.E.M. Anscombe, Trans.). Blackwell, Oxford.

Structural Adjustment in Africa: Implications for Social Work

Lengwe-Katembula J Mwansa, University of Botswana, Gaborone, Botswana

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Africa, rich in diversity and natural resources, continues to reel from abject poverty and underdevelopment, especially in sub-Saharan countries. Its genesis in economic crises of the 1980s is disputed. This article contends that recognizing the complexities of these crises is essential to understanding Africa's current predicament and ways forward. I argue that the model of structural adjustment that the International Monetary Fund and World Bank imposed destabilized economic development and led to this policy's abandonment. In promoting social development, social workers can utilize models of empowerment through community practice to facilitate social and economic innovation in sub-Saharan Africa.

Introduction

The African continent, the second largest in the world at 30.2 million square kilometers (www.google.co.za) is home to more than 1 billion people, and the second most populated continent (UNDESA, 2006). According to UNESCO, Africa has 2000 languages and dialects, and almost as many clans (Mokhtar, 1990), mineral wealth, natural beauty, and a rich history. However, for many Westerners, Africa conjures up images of the Dark Continent and underdevelopment (Bates, 2012). Some, including scholars of international repute, forget that Africa is not a country but a continent of rich diversity, with 55 countries having their own sovereignty, distinct traditions and beliefs despite multifarious cultural nuance, bound together in a kaleidoscope of confounding perplexity.

Many Westerners do not realize how Africa has been locked into poverty, seeing its people as mendicants. Their reasons for the stagnation of the continent are similar to those expressed by advocates of Western economic models like Vishnu Padayachee (2010: p. 47) and 'ignored Africa's diversity in their rush to peddle simple nostrums of dubious merit'. Other than the economy, liberal economists have not viewed diversity and conflict as fundamental factors in the matrix for African development, utilizing instead Keynesian neoliberal economic models that do not incorporate original thinking or fit African realities. This includes African economists who claim allegiance to neoliberal economic ideological models espoused by the International Monetary Fund (IMF) and World Bank and the efficacy and necessity they recommend.

The Structural Adjustment Programmes (SAPs) were hailed as the best strategy for the economies of Africa, particularly Ghana, Uganda, and Zimbabwe, which had been celebrated as shining models of development (Devarajan et al., 2001: p. 54; Onimode, 1989a). However, the SAPs did not spare their economies because the ensuing economic crisis wreaked havoc and untold misery on their populations, particularly women, older people, rural people, and children. Life expectancy plummeted to the lowest levels from 49.0 to 45.1 and 61.7 to 47.4 for Uganda and Zimbabwe, respectively from 1987 to 1998 (CIA, 2012a,b), while maternal and infant mortality rates worsened. The majority of people in Ghana and Uganda continue to experience abject poverty with over 50% of Uganda's population living on $1 a day (United Nations Development Programme, 2012; *Millennium Development Goals Report*, 2012b).

Despite receiving substantial amounts of foreign aid, sub-Saharan Africa still suffers from wretched poverty, massive illiteracy, disease, squalor, and a poor general standard of living. This article recognizes the relevance of economic problems that bedevil the continent, but contends that remaining open-minded and recognizing the complex structural problems that have equally adversely affected human development in Africa is essential in avoiding flawed solutions to its problems. Not attributing the continent's woes to one factor is also significant in this.

Origins of the Economic Crisis

The 1980s and beyond are regarded as 'lost decades' for development in many sub-Saharan African countries. Some observers (Onimode, 1989b) located the genesis of deteriorating conditions of life at the time led many observers wrongly to locate the genesis of the economic crisis in sub-Saharan Africa to the Organisation of Petroleum Exporting Countries oil shock of the 1970s. Others locate the origin of the economic crisis in the period shortly after political independence (Hope, 1997). Hope blames the failure of a "policy framework, which put considerable faith in the role of government to plan and allocate resources for economic development" (p. 11). Easily forgotten or not sufficiently appreciated by many observers of Structural Adjustment Policies are the antecedents of development that fundamentally affected the internal logic of the structure of African economies given their diverse nature. The determining antecedents cited are: corruption, the Cold War, negative sequel of colonialism, slave trade, 'the rape of Africa' through blatant exploitation of natural resources, unequal terms of trade, and unending conflict, often instigated by Westerners pursuing their own interests (Rodney, 1972: p. 106; Kim, 2010).

The Cold War was a period of intense posturing with a great deal of uneasiness among countries such as Britain, the USA, and Russia. The aid-recipient nations had to demonstrate compliant behavior especially voting at global level, which had to be seen to support the fight against communism (Foot et al., 2003). Thus, creditor nations

through their lending institutions controlled the behavior of African nations – a carrot and stick approach. This patronage, had little or no benefits for the recipients and, in the process, the opportunity for development was lost.

Many observers (Heldring and Robison, 2012) question the role of colonialism in disrupting the fabric of Africa and creating conditions that foster enmity between ethnic groups leading to ongoing confliction in the continent. Peace in Africa has largely been undermined by conflicts, which seem endemic in many countries including the Democratic Republic of the Congo (DRC), Somalia, Uganda, Sudan, Nigeria, and Central African Republic (Gettleman, 2010). However, like other ancient civilizations, tribes indigenous to Africa experienced conflict among themselves and intertribal periods of war and peace. Historically, they used war to gain wealth and notoriety and established peace treaties with neighbors that would benefit them.

European colonizers, attempting to maintain control of their subjects, used intertribal conflict to divide and rule. Consequently, previously established interethnic relations were jeopardized and effectively destroyed, as the colonial powers pitted tribes against one another (Bates, 2012). Rather than overtly establishing discord among their subjects, a favorite ploy was to incite jealousy and envy by showing favor to some tribes but not to others, and arbitrarily installing chiefs who were manipulated to oppress and suppress other ethnic groupings. This sowed seeds of bitterness and anger in the foundations of many future nations, set the stage for intra and international conflict. Besides separating indigenous ethnic groups, colonial powers totally discouraged interclan interactions, creating prejudices by fostering tribalistic attitudes based on incorrect perceptions of others. A poignant example of this strategy was the apartheid regime in South Africa, which ended in 1994. Although the move to democracy was relatively peaceful, the country is now beset by internecine strife, which threatens economic and social development. Despite escaping the SAPs, the majority of black people in South Africa continue to live in squalor and suffer the indignities of poverty like those in many sub-Saharan countries.

Freedom movements arose in response to colonialism. National groups were established and nations came together across the continent to form the Pan-African Movement. The impact was massive, creating a domino effect of newly formed states. As colonial powers pulled out, a new era was born, with the Organization of African Unity being formed in 1963 by 32 member states. This became known as the African Union in 2002.

The fairy tale of an independent Africa was short lived. An unfortunate consequence of this historical process was that the prejudices that had been created and exploited under colonialism did not disappear overnight, but continued as 'bones of contention' within and among newly formed nations. This was seen during the genocide that took place in Rwanda in 1994 (Semujanga, 2003), which led to the massacre of more than 800 000 Tutsi men, women, and children, and the displacement of hundreds of thousands more. At the root of the genocide was the belief that Tutsi people were flaunting and abusing the historical favor they had gained from Belgian colonial masters, having carried the attitude of superiority into the postcolonial era to the disadvantage of the Hutu people (Des Forges, 1999; Adedeji, 1999; Semujanga, 2003).

Immediately after independence, young nations on the continent thrived. Wealth and investments flowed in liberally and spent without accountability including in the public sector. During such economic prosperity, albeit based on primary commodities, there was money to spend. Lack of scrutiny of government spending began a trend of resource abuse that opened the door to corruption (Wrong, 2009). Millions of aid money went into the pockets of a few in leaders such as heads of state who plundered state coffers with impunity. Their insatiable greed was described as pathological (Wrong, 2009).

As corruption grew politics became dirtier, the chasm between rich and poor people widened, and economies spiraled downward at breathtaking speed. The drop in commodity prices, the failure to produce adequate food, and concomitant inability to service accumulating debt and cope with growing inflation added to the rapid free fall of many sub-Saharan economies. Only certain groups of people were benefiting economically and many people had been better off at independence than afterward. Ideals such as national unity and a common identity were dismissed, and a mad scramble to gain resources, through whatever means possible, gained momentum. Notably, corrupt rulers of the new political dispensation took away power from traditional authorities and structures to make people powerless (Gibbon et al., 1992). Instability spread across the continent. Conflicts began as warlords and politicians attempted to bend the masses to their will. Again, it seemed that, at the root of the political instability that ended in conflict were corruption, tribalism, and xenophobia (Wrong, 2009). Those in power were intent on taking advantage of the remnants of colonialism to spur people on to war. A pattern was established: instability led to conflict, which caused further deterioration in stability, and led to more conflict as people fought to survive. DRC, Somalia, Mali, Sudan, and Uganda are examples of ongoing hostility and ethnic tensions causing people to flee for their lives, while warlords amass wealth and exploit the region (Gettleman, 2010; Adedeji, 1999).

Structural Adjustment Programmes

This context of widespread prejudice and misrepresentation on a continental scale calls for debate on African development that lacks a monolithic approach, which focuses entirely on economics. Unfortunately, various scholars (Campbell and Loxley, 1989; Gibbon, 1995; Devarajan et al., 2001) have expressed a neocolonialist mind-set that prescribes the one-size-fits-all solutions inherent in the SAPs. SAPs are "a set of free market economic policy reforms imposed on developing countries by the Bretton Woods institutions: the IMF and the World Bank in the 1980s and 1990s … as a condition for receipt of loans" imposed on almost all sub-Sahara African countries (Gibbon, 1995; Savas, 1987). This led social welfare programs to plummet from massive budgetary cuts in social funding resulting from the

economic financial crisis that affected almost all sub-Saharan Africa.

The financial crisis on the continent worsened from the 1970s to the 1990s, causing the social welfare of millions of people to decline to the lowest level ever. According to the Human Development Report (2012), many countries in sub-Saharan Africa ranked low in the Human Development Index, combined with poor measures in human capabilities regarding health, education, and income (p. 16). The inability of many families to acquire food due to lack of agricultural input and disposable income meant millions of people in the region experienced food insecurity at household and personal level, and weakened health status from poor nutrition. In sub-Saharan Africa, 31% of the population was undernourished in 1990–1992 (UN, Millennium Development Goals Report, 2012b). In 1990, over 29% of children under age five were either moderately or severely underweight (UN, Millennium Development Goals Report, 2012b). Hungry children suffer from communicable diseases that are ordinarily preventable and treatable and succumb easily to death. The result is high infant mortality rates. Children in sub-Saharan Africa tend to join school late and drop out early. Malnourished expectant mothers in rural areas far from health facilities face the prospect of death while giving birth. Agricultural production is hampered by ill-health from malnutrition and diseases such as HIV and AIDS, malaria, and tuberculosis, which impact negatively on food production. All these factors feed into a vicious circle of poverty, which has scaled down in other parts of the world, but spiraled out of control in sub-Saharan Africa (Khumalo, 2009).

At the same time, the economic crisis escalated as a result of a burgeoning external debt burden, the brain drain, capital flight, poor balance of payments, rising unemployment, infrastructure deterioration, growing budget deficits, inability to service internal and external debts, social inequalities, rampant inflation, negative interest rates (Ajayi and Khan, 2000) and corruption. Unequal terms of trade contributed to poor growth and caused the continent to become marginalized on the world trade stage and entrenched its position vis-a-vis lack of investment, negative growth, and poor economic performance (Adedeji, 1976, 2002). Dearth of investment in sub-Saharan Africa affected its capacity to export and caused the region to rely primarily on imported goods and services, turning these countries into net importers. The option for salvation, claimed the IMF and World Bank, was for sub-Saharan African nations to encourage a liberalized economic approach that would bring about growth and prosperity in both the short and long terms. The major thrust of the approach was based on growth of the economy, which, it was assumed, would eventually trickle down to deprived groups, including rural poor people. However, before that could happen, the state had to find money to launch a program of economic recovery. Without other ways of paying for it, many sub-Saharan countries were forced to seek financial assistance from the World Bank and IMF as the only source of funding (Gibbon, 1995; Devarajan et al., 2001). Funding came with conditionalities, including strict adherence to the adoption of the market economic model with private sector driven and free-market policies. Without other feasible options, the countries of sub-Sahara Africa were coerced into contracts for loans (Campbell and Loxley, 1989). This led many economies in the region to lose the ability to reorganize and relaunch their development agendas. The scarcity of capital and lack of adequate domestic savings ability led to financial chaos and economic collapse owing to hyperinflation, high unemployment rates, and lack of net investment for growth and development. The balance of payments could not be serviced and accumulated into huge debt, which over time, grew into an interminable debt burden (Ajayi and Khan, 2000). The concomitant offshoots of the crisis included severe shortage of goods and services and general hardships especially diswelfare and falling standards of living. The burden was heavier for rural areas than urban ones.

The SAPs, as a model of economic recovery based on free market economic policy reforms, consisted of a series of measures that had to be accepted and strictly adhered to by debt-burdened African countries to obtain loans. SAPs were applied to every country in Africa, regardless of its historical, cultural, and material circumstances or particular country context. These SAPs consisted mainly of the following components:

1. Elimination of subsidies
2. Reduction of public expenditure on social services
3. Devaluation of national currencies coupled with foreign-exchange auctions
4. Reliance on free market forces
5. Privatization of public enterprises
6. Liberalization of trade and inputs
7. Downsizing of the public labor force

(Adedeji, 2002; Chikulo, 1997; Gibbon, 1995, 1993).

The situation is akin to the contemporary phase of economic crisis being experienced in Southern Europe with consequences of extreme reduction in social spending and uneven distribution of incomes (Dominelli, 1990, 2010; Martin, 2012; Papadopoulos, 2012). There is total agreement about the economic crisis that engulfed the economies of sub-Saharan Africa and need to carry out reforms to resolve the crisis. However, disagreement arose over the particular model of economic recovery that had to be implemented. While some observers contend that SAPs were a necessity, others argue the contrary, and categorically believe that another credible alternative could have been implemented to address the continent's economic problem. SAPs were based on nondemocratic and unwarranted harsh methods without regards to the human aspect of development, which led to underdevelopment as is currently occurring in Southern Europe. For example, the UN Economic Commission for Africa provided an all-inclusive and sound alternative to SAPs in 1989 (Adedeji, 2002). This essentially offered an approach to development that considered the plight of humans. Another example was the African Alternative Framework, which called for 'Adjustment with transformation'. The emphasis in these and similar models was on reducing the continent's reliance on external trade and financing, fostering food self-sufficiency, and promoting popular participation in economic planning and decision making.

These all go against the imperialist expectations of the Bretton Woods institutions, their ideology and philosophy as Ioakimidis argued regarding Southern Europe (Ioakimidis, 2011).

SAPs' focus was supposed to correct the economy but it ended up finding new sources of capital accumulation for multinational corporations who initially used the state apparatus in their countries of origin to secure the conditions that suited them. Thus, insufficient attention was paid to the social consequences of the proposed economic growth model with the introduction of free market enterprise to 'eliminate trade and investment regulations', improve foreign-exchange income, increase exports, and reduce domestic borrowing and expenditure on subsidies and social welfare. Despite the harsh conditions imposed by the IMF and the World Bank, countries had no choice but to take the loans and strings attached. Many donor countries tied their aid to the adoption of the SAPs (Devarajan et al., 2001; Jordan and Drakeford, 2012). The World Bank and IMF were adamant that the SAPs were necessary to bring about economic development and recovery with the usual 'trickle down' theory, which continued to have appeal (Jordan and Drakeford, 2012). In retrospect, the SAPs constituted a project doomed to failure even before it was implemented as argued by many observers (Adedeji, 1999: pp. 521–522, 2002; Dominelli, 1999; Campbell and Loxley, 1989: pp. 1–3).

The countries of Africa were trapped for long periods in a never-ending debt repayment process through which the Bretton Woods institutions ensured rescheduling that enabled the arrangement to continue. The debt servicing design became the order of the day with a constant rise in rates of interest, which Onimode (1987) termed a 'Ponzi scheme' or fraudulent scam. Payments from debtor nations boosted domestic profits and made possible employment creation for the lenders while indebtedness, unemployment, and poverty in the debtor nations grew by leaps and bounds (Onimode, 1987: vol. 2).

Despite the immense, growing burden caused by the SAPs, the IMF, World Bank, and creditor nations remained oblivious to the harsh economic hardships of poorer countries. This led to global outcry by organizations, individuals, activists, political leaders, the church, and Jubilee 2000 (http://www.guardian.co.uk/global-development/2011/nov/15/vulture-funds-jersey-decisions; Carrasco et al., 2007) that felt that the debt had caused untold misery among children, women, and older people. Women and children, with their vulnerable socioeconomic positions, bore a disproportionate share of the burden created by SAPs (Mupedziswa and Gumbo, 1998). The world's conscience through leading figures, the Jubilee 2000 Movement and Jeffrey Sachs (www.sgiquartely.org/feature/2000Jly.5html), and Greene and Khan (1990) spoke out about the negative aspects of SAPs and made a determined effort to spearhead the cancellation of debt to give the ailing economies of sub-Saharan Africa an opportunity to recover. However, extreme poverty affected millions of people in the region as economies continued to experience deep recession. The absolute number of people living in extreme poverty reached 395 million in 2005 but dropped to 386 million by 2008 (a 5% reduction), while other regions demonstrated significant reductions in poverty. Nevertheless, the majority of people in sub-Sahara Africa continue to live below the absolute poverty rate of $1.25 a day (Millennium Development Goals Report, 2012b). For those in rural areas, or places of conflict, the debt burden was disproportionate. They faced crumbling infrastructure, lack of food security, relentless poverty, unemployment, and the devastation of diseases such as HIV and AIDS, tuberculosis, and malaria. They were confronted daily by declining life expectancy, unacceptably high rates of maternal and child mortality, and famine; while a few lived in unimaginable luxury, funded through ill-gotten gains, usually in urban areas or abroad. While the donor countries forgave the highly indebted countries by dropping the demand to pay any outstanding debt, nobody knows the extent of the negative impact of the SAPs and concomitant injury caused to the people of Africa. The African situation may sound outdated but should alarm Southern Europe. The crisis will be long and hard with consequences of rising unemployment, budgetary austerity, slowing of domestic economy, growth deficit, and a deterioration in living standards for majority of citizens (Das, 2008; Martin, 2012).

Many groups including nongovernmental and humanitarian and grassroots organizations and churches saw through the actions of donor countries and lending institutions like the IMF and World Bank: the debtor nations were totally incapable of paying back the outstanding loans (Jeffrey Sachs, 2000). The debtor nations had more than paid back the original principal on the loans and the whole exercise of continued paying was a mockery by any standard. The situation led to worsening poverty outcomes and perpetual unequal relations thus entrenching inequalities and ever-growing gap between the rich and poor nations and individuals within sub-Saharan countries themselves.

SAP's schemes led to the emergence of a new form of inequality between debtor and donor nations. Previously the parties had comprised the colonizers and colonized – a correlation, which ended at the time of independence when the countries of sub-Saharan Africa attained sovereign statehood. The new relationship, however, consisted of debtor and creditor countries. This relationship ushered in a new form of colonialism, which essentially ensured continuation of the domination of social, political, cultural, and economic relations by former colonial masters. Although Nkrumah (1965) warned of such a relationship as neocolonialism and imperialism in which former colonial powers continued to maintain control or hegemony over former colonized countries and thus perpetuate colonialism in a new form, the issue has gone beyond this in the last few decades. Like in Southern Europe, the rise of large multinational corporations that have more economic clout than nation states and the rise of China as a global economic player has made Western governments play into the pockets of financial elites who are unaccountable to none but themselves (Dominelli, 2009). This situation has led multinational corporations to act with impunity in their fraudulent activities throughout the globe to maintain their influence. Essentially, apart from Botswana, Mauritius, and South Africa the countries of Africa have become locked into new forms of independent/dependent relationships with Western governments and financial institutions. The unequal relationships imposed on

these countries have undermined their sovereignty and virtually crafted them into marionettes of the West (Campbell and Loxley, 1989; Onimode, 1987). The end of the Cold War has not substantially altered these political and economic dynamics because developed countries continue to use methods of control over African countries. The entry of China on to the global scene has not changed these power relations, but worsened the situation for African countries (Heller, 2006; Ghosh, 2006).

Implications for Social Work Practice

The first implication for social workers wishing to depart from neocolonial relationships is that social work must be redefined to fit the realities of Africa. The need to reconfigure African societies' structures and economic functioning has grown exponentially over the past quarter of a century. The African continent demands a corresponding realignment in the social sciences and social work practice. This is particularly evident in social work, addressing, as it does, social functioning at individual, familial, and societal levels. International conceptualizations of social work focus on these three levels with, ostensibly, equal emphasis. However, the reality is that people in the developed world are tending to become more and more isolated and individualistic and social work practice is increasingly responsive to this phenomenon. Thus, although the consensus is that person in environment is the mantra for assessment and intervention, reality is not always in line with this contention (Dominelli, 1999, 2002; Mullaly, 1997).

Social work practice should identify its inherent purpose in relation to the needs of vulnerable, marginalized, and minority groups in society and work toward social justice for all. Also important is for a definition of social work in Africa to incorporate a common point of departure. The community would serve well in this regard as it situates the practice within an environmental perspective demonstrating the fundamental and traditional importance of kinship systems and other social structures of value in daily life (Dominelli, 1999, 2002; Mullaly, 1997). This also recognizes that social issues in Africa are essentially community issues that cannot be effectively dealt with by focusing on individuals without acknowledging that society, family, and individuals are essential elements of a coordinated approach to social work practice. Such issues have to be incorporated in a universal definition of social work and its practice if the needs of all peoples are to be addressed adequately.

In Africa, like elsewhere, care has to be taken to ensure that social work takes cognizance of culture and related environmental factors as populations are integrated into these aspects of life. Although individualistic trends are becoming more apparent, the situation on the ground is bound up with societal and familial expectations. Manipulating the relationship between development and custom calls for special skills in social workers and these have to be incorporated into Western definitions and theoretical frameworks for practice to avoid neocolonial relationships.

Secondly, in searching for the appropriate implications of social work in Africa in current global and regional economic trends and demands, social workers worldwide must continue to base their practice on the well-established principles and values of human rights and social justice while focusing on marginalized and disadvantaged peoples. Thus curricula changes to incorporate interventions that promote empowerment and strengths and human rights–based approaches (Mullaly, 1997) for sustainable development are required in programs. Empowering interventions are suitable and efficacious for sustainable development because they advocate equity and social justice in all development processes. The attraction of a human rights–based approach is that it ensures that human rights' values and principles distinctively guide the development and capacity of all stakeholders (Hardcastle and Powers, 2004; Midgley and Conley, 2010) and stimulates a consciousness and awareness of human rights in the process of working with various groups of people to resolve their problems. The human rights–based approach is also essential in realizing the Millennium Development Goals, the Convention on the Elimination of all Forms of Discrimination Against Women (CEDAW), and other global initiatives. Human rights relate well to globalization and neoliberalism in addressing the recent global financial crisis, which has demonstrated the interdependence among peoples of the world and carries implications for social work practice (Dominelli, 2013).

Professionals in Africa must constantly remember that the continent shares this rapidly changing and globalizing environment. Hence, continuously developing the profession's standards, principles, methods, and skills base in relation to the specific challenges experienced on a daily basis is important. In the African context, challenges of practice include highly disparate levels of socioeconomic development, widespread and complex conflicts and wars, large-scale poverty, and a ravaging HIV and AIDS epidemic. To develop responses that are locally relevant and appropriate, it is important to draw on existing local knowledge, skills, and value systems. These are strengths to be incorporated into a profession seeking to make a meaningful contribution to social and economic development in Africa.

The focus of practice in Africa, as elsewhere, must be to develop strategies for capacity building among people to free them from control and oppression. Control and oppression have pervasive effects on the lives of people and undermine the value base of social work regarding personal dignity and worth, egalitarianism, social equality, and democratic ideals (Mullaly, 1997). This is why community practice approaches that emphasize economic development, physical environment, and those models that critique models of capitalist models of socioeconomic development must be promoted as engines of practice through models incorporating social action and planning (Zastrow, 1993; Shragge, 1997; Sherraden and Ninacs, 1998; Hardcastle and Powers, 2004; Ambrosino et al., 2005; Zapf, 2010).

Conclusion

Economic crises have happened in many parts of the world but, as Carol Lancaster has observed, "the economic crisis in Africa has involved problems much deeper than those posed by

short-term financial obligation" (1983: pp. 149–150; De Menil, 2012). Lancaster made these remarks in relation to the debt crisis then taking place in Latin America. Africa's economic crisis is indeed multifaceted, complex, and structural, which cannot be explained unilaterally, whatever the attractions of that view. The observers (Nissanke, 2003) of the African phenomenon should exercise patience and caution in theorizing its nature and origins, attempting instead, to examine a plethora of circumstances some of which are historical and intricate. I do not disagree that there has been a crisis and intervention was necessary to resolve it. However, the question has always been: could there have been an alternative to the SAPs imposed by the IMF and World Bank? The liberal perspective enabled SAPs to impose free market enterprise models that ushered in new forms of relationships between the indebted and lender nations that perpetuated old power relationships embedded in colonizer and colonized binaries and created new relationships that produced a new order or form of colonialism popularly referred to as neocolonialism. The SAPs created more hardships among people as unemployment, poverty (with its offshoots), higher infant and maternal mortality rates, and unprecedented levels of illiteracy. Social work, as professional practice based on principles of humanitarianism and social justice, must adopt strategies and approaches that liberate marginalized and disenfranchised populations. This can be achieved through a redefinition of its purpose to reflect the unique circumstances of the African continent.

See also: Global Agenda for Social Work and Social Development; International Association of Schools of Social Work; International Council on Social Welfare; Marxist Approaches to Social Work; Neoliberalism; Social Work Theory.

Bibliography

Adedeji, A., 1976. Africa, the Third World and the Search for a New Economic Order. Government Printer, Georgetown.

Adedeji, A. (Ed.), 1999. Comprehending and Mastering African Conflict: The Search for Sustainable Peace and Good Governance. Zed Books, London.

Adedeji, A., 2002. Structural adjustment policies in Africa. International Social Science Journal 51 (162), 521–528.

Ajayi, S., Khan, M.S. (Eds.), 2000. External Debt and Capital Flight in Sub-Saharan Africa: International Monetary Fund. IMF Institute, Washington, DC.

Ambrosino, R., Heffernan, J., Shuttlesworth, G., Ambrosino, R., 2005. Social Work and Social Welfare an Introduction, fifth ed. Brooks/Cole – Thomas Learning, Belmont.

Bates, R., 2012. History of Africa Through Western Eyes. Think Africa Press, Guardian African Network, London.

Bos, H.C., Sanders, M., Secchi, C., 1974. Private Foreign Investment in Developing Countries: A Quantitative Study on the Evaluation of the Macro-Economic Effects, International Studies in Economics and Econometrics. Dordrecht-Holland, Boston.

Campbell, K.B., Loxley, J. (Eds.), 1989. Structural Adjustment in Africa. Macmillan Press Ltd., Southampton.

Carrasco, E., McClellan, C., Ro, J., 2007. Foreign Debt Forgiveness and Repudiation. University of Iowa, Center for International Finance and Development E-book. http://en.wikipedia.org/wiki/Jubilee_2000 (accessed on 31.01.14.).

Chandrasekhar, C.P., Ghosh, J., 2008. China's African Hinterland. http://www.macroscan.com/fet/mar08/China_Hinterland.pdf.

Chikulo, B., 1997. Structural adjustment policies and the new conditionalities: towards development in Africa. In: Hope, K.R. (Ed.), Structural Adjustment, Reconstruction and Development in Africa. Ashgate Publishing Company, Vermont, pp. 47–73.

CIA World Factbook, 2012a. Life Expectancy at Birth of Uganda. Central Intelligence Agency, Washington, DC.

CIA World Factbook, 2012b. Life Expectancy at Birth of Zimbabwe. Central Intelligence Agency, Washington, DC.

Das, D.K., 2008. Contemporary phase of globalisation: does it have a serious downside? Global Economic Review 38 (4), 507–526.

De Menil, G. (Ed.), Economic Policy 72. Economic Policy Series. http://eu.wiley.com/WileyCDA/Section/id-397558.html.

Dent, M., 1999. Jubilee 2000 The Genesis of the Campaign. Available from: http://www.guardian.co.uk/global-development/2011/nov/15/vulture-funds-jersey-decisions (accessed on 10.10.13.).

Des Forges, A., 1999. Leave No One to Tell the Story: Genocide in Rwanda. Human Rights Watch, New York. International Federation of Human Rights, Paris.

Devarajan, S., Dollar, D.R., Holmgren, T., 2001. Aid and Reform in Africa. Lessons from Ten Case Studies. The International Bank for Reconstruction and Development, The World Bank, Washington, DC.

Dominelli, L., 1990. Women and Community Action. University of Chicago Press, Chicago.

Dominelli, L., 1999. Neo-liberalism, social exclusion and welfare clients in a global economy. International Journal of Social Welfare 8 (1), 14–22.

Dominelli, L., 2002. Anti-oppressive Social Work Theory and Practice. Pelgrave MacMillan, Basingstoke.

Dominelli, L., 2009. Social Work in a Globalising World. Polity Press, Cambridge.

Dominelli, L., 2010. Social work in a Globalising World. Polity Press, Cambridge.

Dominelli, L., 2013. Green social work: from environmental crises to environmental justice. International Journal of Social Welfare 22 (3), 328–329.

Foot, R., MacFarlane, S.N., Mastanduno, M. (Eds.), 2003. US Hegemony and International Organization. The United States and Multilateral Institutions. Oxford University Press, Oxford.

Gettleman, J., March/April 2010. African Forever Wars. Why the Continent's Conflict Never End. Foreign Policy. www.foreignpolicy.com/node/103221.

Gibbon, P. (Ed.), 1993. Social Change and Economic Reform in Africa. Nordiska Afrikainstitutet, Uppsala.

Gibbon, P. (Ed.), 1995. Structural Adjustment and the Working Poor in Zimbabwe. Nordiska Afrikainstitutet, Uppsala.

Gibbon, P., Bangura, Y., Ofstad, A. (Eds.), 1992. Authoritarianism Democracy and Adjustment. The Politics of Economic Reform in Africa, Seminar Proceedings No. 26. The Scandinavian Institute of African Studies, Uppsala.

Greene, J.E., Khan, S.M., 1990. African Debt Crisis, AERE Special Paper 3. Initiative Publishers, Nairobi.

Hardcastle, D.A., Powers, P.R., 2004. Community Practice, Theories and Skills for Social Workers. Oxford University Press, Oxford.

Heldring, L., Robison, J.A., 2012. Colonialism and Development in Africa. Harvard University Working Paper. NBER Working Paper No. 18566. Issued in November 2012. The National Bureau of Economic Research. NBER Program(s): DAE http://www.nber.org/papers/w18566.

Heller, T.C., 2006. African Transition and Resource Curse: An Althernative Perspective. 26 Economic Affairs 24. http://www.law.stanford.edu/publications/african-transitions-and-the-resource-curse-an-alternative-perspective.

Hope, K., 1997. Structural Adjustment, Reconstruction and Development in Africa. Ashgate, Aldershot.

Jordan, B., Drakeford, M., 2012. Social work and social policy under austerity. In: Adams, R., Dominelli, L., Payne, M. (Eds.), Reshaping Social Work Series. Palgrave Macmillan, London, p. 256.

Khumalo, T.E., 2009. An Assessment of the Department of Social and Community Development Poverty Alleviation Programmes. University of Botswana, Gaborone.

Kim, S., 2010. The role of cultural diversity in conflict resolution in Africa. Peace and Conflict Monitor. From http//www.monitor.upeace.org/printer.cfm?id article=747 (accessed on 24.10.13.).

Lancaster, L., 1983. Who Gets to Run the World. Africa Economic Crisis in Foreign Policy. Washington Newsweek International, Washington, pp. 149–166.

Loakimidis, L., 2011. Social Work in Extremis. Policy Press, New York.

Martin, C.J., 2012. The consequences of the global financial crisis: the rhetoric of reform and regulation. In: Grant, W., Wilson, G.K. (Eds.), Social Solidarity in Scandinavia after the Fall of Finance Capital. Oxford University Press, Oxford, pp. 187–205.

Midgley, J., Conley, A., 2010. Social Work and Social Development: Theories and Skills for Developmental Social Work. University Press, Oxford.

Mokhtar, G., 1990. UNESCO General History of Africa. In: Abridged Edition: Ancient Africa, vol. 2. University of California Press, Berkeley, CA.

Mullaly, B., 1997. Structural Social Work. Ideology, Theory, and Practice, second ed. Oxford Press, Oxford.

Mupedziswa, R., Gumbo, P., 1998. Structural Adjustment and Women Informal Sector Traders in Harare, Zimbabwe. Nordiska Afrikainstitutet Research Report No. 106. Nordiska Afrikaninstitutet, Uppsala.

Nissanke, M., 2003. The Neo-liberal Doctrine and the African Crisis (Draft). Rethinking Development Economics. UNRISD, Palais de Nations, Switzerland. http://www.unrisd.org/unrisd/website/document.nsf/(httpPublications)/B00DCEBF29494C59C1256BC900482F49?OpenDocument (accessed 31.01.14.).

Nkrumah, K., 1965. Neo-Colonialism: The Last Stage of Imperialism. Thomas Nelson & Sons, London.

Onimode, B., 1987. The African crisis and Nigeria. African Crisis. Institute for African Alternatives, London.

Onimode, B. (Ed.), 1989a. The IMF, the World Bank and the African Debt, vol. 1. Zed Books Ltd, London.

Onimode, B. (Ed.), 1989b. The IMF, The World Bank and the African Debt The Social and Political Impact, vol. 2. Zed Books Ltd, London.

Papadopoulos, T., Roumpakis, A., 2012. The Greek welfare state in the age of austerity: anti-social policy and the politico-economic crisis. In: Kilkey, M., Ramia, G., Farnsworth, K. (Eds.), Social Policy Review 24: Analysis and Debate in Social Policy, vol. 24. The Policy Press, Bristol.

Padayachee, V. (Ed.), 2010. The Political Economy of Africa. Routledge, New York. http://books.google.co.za/books?id=PU9xkf2e0soC&pg=PA61&lpg=PA61&dq=Padayachee,V.++(ed).,+2010+The++Political+economy+of+Africa,+Routledge&source=bl&ots=2uSqXP5v2j&sig=66xF8oruKyNCqMrmsSEfLRHG_wg&hl=en&sa=X&ei=-.

Rodney, W., 1972. How Europe Underdeveloped Africa. London and Tanzanian Publishing House, Bogle-L'Ouverture Publications, Dar-es-Salaam.

Sachs, J., 2000. Debt Cancellation: A Moral and Economic Perspective. Interview with Professor Jeffrey Sachs www.sgiquartely.org/feature/2000Jly.5html (accessed 20.06.13.).

Sachs, J., 2005. The End of Poverty. How We Can Make It Happen in Our Lifetime. Penguin Books, New York.

Savas, E.S., 1987. Privatisation: The Key to Better Government. Chatham House Publishers, Chatham, NJ.

Semujanga, J., 2003. Origins of the Rwandan Genocide. Humanity Books, Amherst, New York.

Sherraden, M.S., Ninacs, W.A., 1998. Community Development and Social Work. Routledge, London.

Shragge, E., 1997. Community Economic Empowerment. Black Rose Books, Montreal.

United Nations Development Programme, 2012. Africa Human Development Report 2012. Towards a Food Secure Future. 1 UN Plaza, The United Nations Development Programme, Regional Bureau for Africa (RBA), New York.

United Nations Development Programme, 2012a. African Human Development Report 2012. 1 UN Plaza New York, New York.

United Nations Development Programme, 2012b. The Millennium Development Goals Report 2012. We Can End Poverty 2015 Millennium Development Goals. 1 UN Plaza New York, New York.

Wrong, M., 2009. The Story of a Kenya Whistleblower. It's Our Turn to Eat. Fourth Estate, London.

Zapf, M.K., 2010. Social Work and the Environment: Understanding People and Place. Critical Social Work 11 (3). http://www1.uwindsor.ca/criticalsocialwork/social-work-and-the-environment-understanding-people-and-place.

Zastrow, C., 1993. Introduction to Social Work and Social Welfare, fifth ed. Brooks/Cole Publishing Company, Pacific Grove, CA.

Structural Balance Theory: Recent Developments

Peter Abell, London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Structural balance analyzes the statics and dynamics of signed graphs (digraphs), which experience stress induced by cycles with an uneven number of negative links. The theory postulates two alternative equilibria (balanced states), either k-sets of (predominately) positively related nodes separated by negative links or a single set of positively linked nodes.

Introduction

Structural balance theory concerns the formation, dynamics, and ultimate stable (i.e., balanced) disposition of positive (P) and negative (N) expressive pairwise binary relations/links on a finite set (n) of actors, which may be individuals or collective entities. Friendship and enmity provide typical examples of expressive relations; though the theory covers any type of relationship carrying positive and negative emotional orientations, it has in addition been extended to trust and mistrust relationships. Although the theory has been applied predominately to small groups of individuals, relations between nations have also been explored.

The structures generated by the relations on n nodes can be depicted as either signed graphs, G(n;P,N), where the relations are symmetric (reciprocal) or digraphs when the relations are directed. Directed structures are usually generated by studies of individual choices and symmetric structures when the focus is upon sustained reciprocal relationships. The pairwise relations/links/edges are conveniently labeled as either $+1$ or -1. Correspondingly, they can also be depicted as the associated incidence matrices (M). The theory can be extended by allowing the links to be weighted; either normed in the interval -1 through $+1$ or drawn from a continuous distribution on the real line. The entries in M are then adjusted accordingly. Reflexive (self) relations are sometimes introduced usually for mathematical purposes. Occasionally, 'ambivalent' relations have been suggested whereby both a positive and a negative link can exist between two nodes (symmetric links implying mutual ambivalence and directed links implying unilateral ambivalence).

Origins and Development of Balance Theory

The theory of structural balance developed from Heider's (1946) exploration of an individual (p) with negative- and positive-directed orientations to both another individual (o) and simultaneously to objects/concepts (x), with (o) also having a signed orientation to (x). It is further assumed that (p) is aware of the orientation of (o) toward (x). The links to (x) were denominated as 'unit formations.' Eight types of possible, so-called 'pox' triples, centered upon (p) are then possible. It was assumed that only those triples with an even number of negative links are stable (balanced), the others induce stress, which will in due course be resolved by the change of sign of a link that was left undetermined.

Cartwright and Harrary (1956) extended this basic picture to incorporate interpersonal relations across many actors, leading to graph-theoretic depictions of signed structures. In the process the orientation to objects was lost, though more recently objects have been reintroduced within the framework of bipartite graphs (Mrvar and Doreian, 2009). Graph-theoretic depiction introduced the distinction between symmetric (reciprocal) and directed links. Early empirical studies of the signed structures tended to consider small groups with complete structures where all the pairs of nodes/actors are directly linked to either P or N (symmetric or directed). More recently, incomplete structures have been incorporated in attempts to apply the theory to larger structures. Incompletion allows for unlinked (null) pairs, which may be either unacquainted or entertain a 'neutral' affective orientation. In the former case the sign of the link is fixed at null (up to the formation of a new link) whereas in the latter case the link can subsequently acquire a value (P or N). These distinctions have not been systematically incorporated into the theory, though the relative density of P, N, and null links plays a role in recent developments.

Balance Theory

The standard assumptions (axioms) of balance theory are often colloquially expressed in terms of friendship and enmity, which give a direct pointer to the intuition behind the theory. So, stable (i.e., balanced) situations run as follows:

1. A friend of a friend is a friend,
2. An enemy of a friend is an enemy,
3. A friend of an enemy is an enemy, and
4. An enemy of an enemy is a friend.

In the context of incomplete structures, however, a slight reformulation better expresses stable situations:

5. A friend of a friend is not an enemy (i.e., is a friend or null),
6. An enemy of a friend is not a friend (i.e., is an enemy or null),
7. A friend of an enemy is not a friend (i.e., is an enemy or null), and
8. An enemy of an enemy is not an enemy (i.e., is a friend or null).

It is postulated that those involved in a particular imbalanced 3-(semi)cycle (with either directed or symmetric links) experience stress and consequently seek to remedy the situation by either deleting or changing the sign of one of the offending links (thus, by either eradicating imbalance or inducing balance).

In the case of weighted links the state of balance of a cycle (or directed triple) is computed by the product of the weightings. However, Montgomery (2009) underscores the fact that imbalanced triples only induce stress if individuals are aware of the nature of the links between others (or from others to concept (x) in the case of 'pox' triples). Thus, it is only with complete awareness (which is often an unrealistic supposition in large structures) that imbalance implies instability.

It is usually assumed that structures do not evolve according to the precepts of balance but rather by the random creation of links, thus initially acquiring imbalanced cycles, which then subsequently adjust toward balance with the passage of time. Since links in the graph G(n;P,N) are involved in up to $(n - 2)$ 3-cycles in a complete structure, it becomes a matter of detailed calculation as to what the impact of changing the sign of a particular link would be. The dynamics are consequently often left undetermined. In this respect the tendency toward balance/stability is treated not unlike many equilibrium theories, which remain silent about the inherent complexities of the off-equilibrium dynamic path.

Nevertheless, the global balance of a structure can be followed by a number of possible coefficients each of which is applicable to both directed and symmetric links, though the counting procedures have to be adjusted accordingly. A line index gives the number of links, which need to flip sign or, equivalently, be deleted in order to attain balance. An alternative is to compute the proportion of balanced 3-cycles ($\beta(3)$) at any particular juncture. The node balance ($\beta_i(3)$) and link balance ($\beta_{ij}(3)$) can likewise be defined in terms of the proportion of balanced cycles subtended by respectively node, i and link, ij. These coefficients can easily be extended to compute the proportional difference between the balanced cycles and their random expectation, given the proportion of P links. Alternatively, the stress is expressed as $(1 - \beta(\cdot))$. A logically equivalent conception (deriving from physicists who have taken an interest in balance) is to interpret stress as potential energy (PE) and the move to balance minimizes the PE. Then, in a complete structure, $(PE) = -Tr(M^3)/{}^nC_3$, acquiring a value of -1 at balance.

The various methods of tracking balance/stress can be extended to incorporate weighted P and N links, which are usually allowed to vary between -1 and $+1$, though recently dynamic models (below), which allow for indefinite continuous scaling, have been introduced.

In a complete structure with symmetric links counting the 3-cyles is equivalent to counting cycles of all lengths. A complete balanced structure will contain no cycles of any length with an uneven number of negative links. But this is not so with incomplete structures, which may contain longer imbalanced cycles and no imbalanced 3-cycles. It is possible then to compute $\beta(n)$ rather than $\beta(3)$, which embraces cycles of all lengths. Although incorporating longer cycles into balance theory is often necessary to maintain its mathematical precision, whether long cycles induce stress remains controversial.

The Structural Theorems

A complete balanced structure G(n;P,N) will, according to the First Structural Theorem (Cartwright and Harary, 1956), comprise either a single set (group) of P-linked nodes or two such sets (groups) connected by N links (bipolarization). It is this theorem which has given balance theory much of its prominence by providing a potential theory of subgroup formation in signed structures.

It leaves open, however, the conditions under which balance will surrender either a single or two groups. In addition, though balance theory has largely been treated as deterministic, the issue arises as to how tolerant near-balance is to the presence of a small numbers of imbalanced 3-cycles. A theorem due to Easley and Kleinberg (2010) states: In a complete signed graph with at least $(1 - e)$ balanced 3-cycles where $0 \leq e \leq 1/8$ and $d = (e)^{1/3}$, then either at least $(1 - d)$ of the nodes are linked by at least $(1 - d)$ of positively related pairs (approximate unity) or the structure consists of two groups where at least $(1 - d)$ pairs are positively related in both groups, and at least $(1 - d)$ pairs are negatively related between the groups. So when $e = 0$, we recover the first structural theorem but note that balance falls away strongly as e increases.

Incomplete but connected structures will probably contain null links in a unified group or both within and between the polarized groups. In this respect the densities of P and N links within and between groups become the focus of attention.

Davis (1967) proposed that all negative 3-cycles (triples) may well be relatively stable. The Second Structural Theorem, incorporating this assumption states that: a k-balanced structure will comprise of either a single or a number, k, of positively related node sets (groups), again, in the latter case each negatively linked.

We can, thus, speak of k-balance, where $k = 1$ or 2 when all negative 3-cycles do contribute to imbalance, and potentially $k > 2$ when they do not. The terms 'clusters' and 'plus sets' referring to k-sets are to be found in the literature.

The instability of 3-cycles with one negative link has also been questioned on the grounds that they can provide a functional bridge between conflicting nodes. If both all negative and one negative 3-cycles were to be deemed stable (stress free) then balance theory would fail. These considerations prompt the notion that thresholds (see below) may operate in respect of the instability of different 3-cycle types.

Statics: Studying Balanced and Near-Balanced States

Since (k)-balance theories predict the evolution of a variety of alternative possible balanced (stable) states, recording the comparative equilibrium statics becomes of central importance. Visual inspection of small structures gives way to formal procedures as n increases and near k-balanced states, particularly in incomplete structures, also enter the picture.

Network block-modeling techniques, which involve permutations of the rows and columns of the appropriate incidence matrix, M, have been extended to incorporate signed links for both one- and two-mode networks (Brusco et al., 2011). In the latter respect, Heider's unit relations are reintroduced into the theory (Doreian and Mrvar, 2009). Balanced

states are located by finding a permutation on M, which locates blocks of positively (and null relations in incomplete structures) related nodes down the leading diagonal. The algorithms minimize the number of misplaced links (i.e., negative links (weighted or binary) internal to the k-sets and positive ones between them). Other mechanisms like differential popularity/unpopularity (thus, degree power laws) can interfere with the likelihood of finding clean k-sets. Relaxed Structural Balance (Doreian and Mrvar, 2009) allows for an approximation to an equilibrium at k-sets within the framework of block modeling by disturbing the neat pattern in any permutations on M and also introducing the possibility of mutually negative nodes (negative blocks). The two structural theorems are then special cases of the more general theory. Although methods of extracting pictures of near-balanced structures are now well developed, balance dynamics that introduce probability in accounting for the variety of outcomes have not been explored. Furthermore, since the time line of stress-reducing dynamics is little understood, whether or not a particular empirical structure has come to rest is often by no means clear.

Dynamics: Studying Adjustment toward Balance

Since more than one stable state is predicted by k-balance theory it cannot be regarded as complete without a specification of how the alternatives arise. It is only recently that the complexities of possible dynamics have been addressed and the literature is, so far, confined to $k = 2$ theory.

If complete structures were to evolve in accordance with the precepts of $k = 2$ balance then the proportion of N links would be maximized and P links minimized (both at 0.5) at $k = 2$ equal-sized groups. The proportion of negative links would fall to zero as a single ($k = 1$) group forms. The dynamics would be determined by the proportion of positive and negative links created at the early period of the evolution and balance would determine the transitive closure of 3-cycles. However, only a few studies conceive of balance in this respect (which would require actors to anticipate stress and, consequently, not to establish imbalanced cycles) but rather assume that stress develops over time in already established imbalanced cycles. Thus, balance is usually construed as an adjustment mechanism in structures that have evolved under the impact of other mechanisms. If P and N links are assigned randomly the probability that a balanced network would form decreases exponentially with n. Current lines of research, however, seek to understand the evolution of signed networks under the impact of multiple mechanisms (e.g., small world and scale free) including balance.

Doreian and Krackhardt (2001), studying transitive closure with directed links, found that adjustments eventuating in PPP, PNN, PNP, and PNP cycles were all consistent with balance, the first two increasing and the latter two decreasing over time, whereas the remaining types were not; NNN increased consistently with $k > 2$ balance.

The dynamics of both binary and weighted models have been studied combining a mixture of simulation and analytical results in both discrete and continuous time. The starting point is usually a structure with randomly distributed varying initial proportions of positive and negative links (sometimes null links also and thus varying density); n is usually fixed, though some analytical results as $n \rightarrow$ infinity have been established.

Queries arise as to whether and the extent to which balance adjustment depends upon the structure size (n), the initial proportion of P, N, and null links (thus density), the relative propensity of P and N links to change sign, and above all, the precise nature of the adjustment mechanism. In the last respect it is important to contrast 'local' mechanisms, which make only limited demands upon an actor's cognitive capacities with 'global' mechanisms, which assume a capacity to assay all the cycles in which an actor is involved.

Antal et al. (2005, 2006) showed that complete structures with binary, symmetric links, adjusting in discrete time, can sometimes fail to reach either of the balanced states, becoming 'jammed' while still imbalanced. Nevertheless, simulations of a local adjustment rule which allows that a randomly selected imbalanced 3-cycle flips the sign of a constituent link, without any regard to the impact on other cycles can procure $k = 1$ balance when the probability that an N link flips to P in one negative 3-cycles is greater than 0.5. If the 'local' adjustment rule is replaced by a 'global' one which selects a link at random and changes its sign only if this decreases the number of imbalanced cycles in the structure, then a sharp transition from $k = 2$ to $k = 1$ balance occurs when the initial proportion of P links is about 0.6. The transition is invariant to a large range in n.

Kulakowski et al. (2008) allowed for weighted links (normed between $-R$ and $+R$), continuous rather than discrete time and global adjustment, enabling a simple differential equation treatment of balancing dynamics. They showed (by simulation) that the time to balance falls with both the size ($n^{-1/2}$) of the network and R (up to $n = 500$). Further, by increasing the initial proportion of P links, a sharp transition from convergence at polarization to convergence at unity is encountered. Balance is always attained and persists.

The continuous time model with global adjustment was further developed by Marvel et al. (2011) providing closed-form demonstration that, over a wide range of initial conditions, balance is necessarily achieved in a finite time; polarization or unity occurs depending on the initial proportion of P links, a result supporting the earlier simulations. However, the dynamics conceived as $dM/dt = M^2$ in complete structures generate a divergence to infinite values for the link weights, though balance is achieved prior to this divergence. Nevertheless, the values required at balance are unrealistic and whether this is pertinent to empirical investigation remains an open question. The authors opine that other mechanisms may be signaled, when large weightings occur, which also drive toward balance.

Abell and Ludwig (2009) combine the growth and balance adjustment of network rather than allowing balance to operate upon an exogenously given structure. They also allowed nodes to tolerate imbalance up to a varying threshold and studied, by simulation, a model where varying proportions of symmetric P and N links are introduced to a node set and any induced imbalance is subsequently solved by cascading the deletion of offending links. The distribution of the sizes of the cascades was reported. Simulations in the space generated by allowing both the threshold and proportion of P links introduced to vary, demonstrate sharp transitions in the degree of balance

achieved and completion at a steady state. One of the regions in the space exhibits power-law distribution in cascade sizes perhaps characteristic of self-organized criticality.

Deng and Abell (2010) studied local ('myopic') balance adjustment as a Markov process with varying transition probabilities between the four types of 3-cycles. The density of links (d), initial proportion of P links (α), network size (n), and transition probabilities were also allowed to vary. The convergence to balance depends critically upon the transition probabilities. Local adjustment does not generate balance but a steady state where $\beta(3) = 0.5$, unless the transition probabilities reflect a strong bias toward P links, when $k = 1$ is favored (cf Antal et al., 2005). This result is immune to variations in a range of network sizes ($n > 30$), densities ($d > 0.3$), and initial proportion of P links ($\alpha < 0.9$). However, below these various parameter values the balance outcome exhibits marked fluctuations, which suggests that the outcomes of small-group studies, which support balance, must involve a global adjustment. Different combinations of 3-cycles were allowed to operate as attractors to explore $k > 2$ balance theory.

Our current understanding of balance dynamics is far from comprehensive. Controlling for other possible conflating mechanisms and possible thresholds, do signed structures achieve balance and if so at which value of k? Analytical results demonstrate that global adjustment will achieve balance in complete structures ($k = 1$ when the initial proportion of P links $\alpha > 0.5$, $k = 2$ when $\alpha < 0.5$) but for incomplete and more realistic local adjustment the picture is more ambiguous. This ambiguity is, however, in accord with accumulated empirical evidence.

Recent Empirical Support

Although recently claims have been made for the predictive capacity of balance theory using very large data sets (e.g., Szell et al., 2010), empirical studies have over the years been dogged by many problems. Paramount among these have been the paucity of longitudinal studies, the difficulties encountered in the identification of negative links, and the interaction of balance with other mechanisms which all appear to have different adjustment time scales. Furthermore, the social context in which the adjustment occurs can have a profound impact (perhaps by virtue of exogenously varying thresholds). At the individual level a small number of data sets (often not designed to address balance issues) have been mined and repeatedly reworked. Recently (Doreian and Mrvar, 2009), two-mode analysis has been applied to voting patterns of the Supreme Court (USA). Balance has proved effective in the analysis of polarization in international relations (Antal et al., 2005). Guha et al. (2004) extended $k = 2$ balance theory to trust- and distrust-directed relations inviting a theoretical connection between balance and game theory. Nevertheless, the parallel between positive and negative affect and trust and distrust has to be handled with considerable care.

The current conclusion which may be drawn from the empirical literature is that balance is often a detectable adjustment mechanism operative in otherwise generated affective structures at varying scales. There is little evidence, though that complete polarization into two or k plus-sets is a universal phenomenon suggesting that local adjustment is often operative. However, in structures with either a high initial proportion of P links, or a disproportionate propensity for P links to form, (which may in practice correlate) approximate adjustment to one set is probably more assured (also supported by both simulation and analytical studies in both continuous and discrete time). In the future, balance dynamics will be studied under the auspices of local adjustment but accompanied by other mechanisms. As it appears that the distinction between local and global adjustment appears so crucial, one may anticipate studies of dynamics with adjustment lying between these extremes.

See also: Bayesian Graphical Models and Networks; Clustering and Cohesion in Networks: Concepts and Measures; Network Centrality, Measures of; Social Networks, Algebra of; Social Networks; Statistical Clustering; Structural Balance Theory: Recent Developments.

Bibliography

Abell, P., Ludwig, M., 2009. Structural balance: a dynamic perspective. Journal of Mathematical Sociology 3, 129–155.
Antal, T., Krapivsky, P.L., Redner, S., 2005. Dynamics of social balance on networks. Pyhsical Review E 72, 036121.
Antal, T., Krapivsky, P.L., Redner, S., 2006. Social balance on networks: the dynamics of friendship and enmity. Physica D 224, 130.
Brusco, M.P., Doreian, P., Mrvar, A., Steinley, D., 2011. Two algorithms for relaxed partitioning: linking theory, models and data to understand social network phenomena. Sociological Methods and Research 40, 57–87.
Cartwright, D., Harary, F., 1956. A generalization of Heider's theory. Psychological Review 63, 277–292.
Davis, J.A., 1967. Clustering and structural balance in graphs. Human Relations 20, 181–187.
Deng, H., Abell, P., 2010. A study of local sign change adjustment in balancing structures. Journal of Mathematical Sociology 34, 253–282.
Doreian, P., Krackhardt, D., 2001. Pre-transitive balance mechanisms for signed networks. Journal of Mathematical Sociology 25, 43–67.
Doreian, P., Mrvar, A., 2009. Partitioning signed social networks. Social Networks 31, 1–11.
Easley, D., Kleinberg, J., 2010. Networks Crowds and Markets. Cambridge University Press, Cambrigdge.
Guha, R.V., Kumar, R., Raghaven, P., Tomkins, A., 2004. Propagation of trust and distrust. In: Proceedings 13th International World Wide Web Conference, pp. 403–412.
Heider, F., 1946. Attitudes and cognitive organization. Journal of Psychology 21, 107–112.
Kulakowski, K., Gawronski, P., Gronek, P., 2008. The Heider balance – a continuous approach. International Journal of Modern Physics C 16, 707–716.
Marvel, S.A., Kleinberg, J., Kleinberg, R.D., Strogatz, S.H., 2011. Continuous-time model of structural balance. Proceedings of the National Academy of Sciences 108, 1771–1776.
Montgomery, J.D., 2009. Balance theory with incomplete awareness. Jornal of Mathematical Sociology 33, 69–96.
Mrvar, A., Doreian, P., 2009. Partitioning signed two-mode networks. Journal of Mathematical Sociology 33, 196–221.
Szell, M., Lambiotte, R., Thurner, S., 2010. Multirelational organization of large-scale networks in the online world. Proceedings of the National Academy of Sciences 107, 1004–1008.

Structural Contingency Theory

Lex Donaldson, School of Management, UNSW-Business School, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Structural contingency theory holds that the effect on organizational performance of organizational structure depends upon how far the structure fits the contingencies, such as uncertainty, strategy, and size. Organizations facing low uncertainty are fitted by specialized and centralized hierarchical structures, whereas organizations facing high uncertainty are fitted by lower specialization and decentralization (i.e., decisions being taken at lower levels of the hierarchy). Undiversified strategy is fitted by a functional structure, whereas diversified strategy is fitted by a multidivisional structure. Larger size is fitted by more specialized and decentralized structure. Various changes over time in focus are identified, such as from differentiation to interdependence.

Definitions

Structural contingency theory is a theory about how an organizational structure needs to fit factors, such as its strategy and size, in order to produce high performance. These factors are referred to as contingencies, because the performance resulting from an organizational structure is contingent upon how far it fits those factors.

Intellectual Background

Organizational structure refers to features such as the degree of specialization of roles in an organization, centralization of decision-making, and so on. Early attempts to prescribe effective organizational structures produced Principles of Management in what came to be referred to as Classical Management. This tended to prescribe structures that had specialized roles, detailed role definitions, standard procedures, and centralized, top-down structures. Subsequently this view was challenged by Human Relations Theory which prescribed loose role definitions and bottom-up structures in which lower level employees exercised relative autonomy or had influence on decisions. Structural contingency theory sought to resolve this tension by saying that each theory was right in its own place. Thus the optimal structure depended upon fitting the contingency factors.

For instance, where organizations faced environments low on uncertainty, they required top-down type of structures that were specialized and centralized (i.e., Classical Management); whereas, where organizations faced environments high on uncertainty, they required bottom-up type of structures that were low on specialization and were decentralized (i.e., Human Relations). This widely received branch of structural contingency theory was created by Burns and Stalker (1961).

Other important early advocates of structural contingency theory were Child, Galbraith, Lawrence, Lorsch, Perrow, Thompson, and Woodward. In particular, Lawrence and Lorsch (1967) described their approach as being a contingency theory. There had already been a contingency theory of leadership, but theirs was the first application of this term to organization structure. This became known as structural contingency theory. Structural contingency theory became the cornerstone of teaching in business schools about which structures are most effective in courses such as Organizational Design or Strategic Management, and is prominent in textbooks on such subjects to the present day, though not always called contingency theory. For a collection of milestone journal articles on research into structural contingency theory see Donaldson (1995).

Current Knowledge

It is convenient to consider there to be three major contingencies of organizational structure: uncertainty, size, and strategy.

The Contingency of Uncertainty

Uncertainty is the inability to predict what course of action will lead to the desired results because of unpredictable variation or lack of knowledge. The environment is a major source of uncertainty, but uncertainty can also arise from the task inside the organization. Much uncertainty comes from seeking to innovate in product or service, or from seeking to provide these in new ways (e.g., online). Burns and Stalker (1961) argued that when uncertainty is low, operations become routine and can be captured by rules that are specified from the top-down, while organizational roles are specialized and closely defined; this type of structure they call mechanistic. In contrast, where uncertainty is high, operations need to rely on the initiative of employees who view their role more broadly; this type of structure they call organic. Burns and Stalker view these two organizational forms as lying at either extreme of a continuum, according to the degree of uncertainty involved.

Lawrence and Lorsch (1967) argued that in predictable environments, departments would be similar, because all functional departments have high certainty, so they can all rely on rules and the like; hence they all tend toward being mechanistic. Therefore, there is little internal differentiation between departments. The differentiation within the organization being low, the organization requires only low levels of integration, such as meetings between members of the departments. In contrast, in unpredictable environments, departments would

differ, with production being certain and mechanistic and research being uncertain and organic, so that the differentiation within the organization would be high, requiring high levels of integration, such as by project teams.

The Contingency of Strategy

Structural contingency theory considers strategy in terms of diversification and vertical integration. A corporate strategy of low diversification is fitted by a functional structure in which the major organizational subunits are functions and their heads report directly to the head of the corporation (CEO). This fosters specialist expertise in each function, promoting efficiency. The functions are cost centers and the only profit center is the whole firm. The CEO can be involved in coordination of the functions, so that the structure is relatively centralized. However, higher diversification is fitted by a multidivisional structure in which the major organizational subunits are businesses, each of whose heads report directly to the head of the corporation (CEO). This fosters coordination across functions for serving particular products or customers. The CEO lacks detailed knowledge of all the divisions' operations and so the divisions are granted some autonomy and the structure is relatively decentralized. The more diversified is an organization, the more decentralized is its structure and so the smaller in size and functions the corporate head office needs to be. Similarly, with more diversification, rewards for subunit heads can shift from being based on corporate performance to being based on subunit performance.

Divisions need to be defined by whichever variable forms the highest degree of diversification: product, customer, or geography. Where more than one variable is quite high on diversification then a matrix structure becomes appropriate.

If corporate strategy favors more rather than less vertical integration, then the stages in adding value to raw material (e.g., mining, refining, and manufacturing) may each become divisions, but subject to central planning of production and transportation across the divisions. Thus, divisional autonomy needs to be restricted and the head office needs to be relatively large (Lorsch and Allen, 1973).

The Contingency of Size

Organizational size refers to the number of people in an organization. As size increases, the number of people reporting directly to the CEO becomes suboptimal, and so an intermediary management level is added. This process needs to be repeated as organizational size grows. Thus, larger size is fitted by more levels in the hierarchy. The greater complexity from increased size and hierarchy requires the top managers to delegate more decision-making. Similarly as size increases, work has to be divided between more people and so role specialization needs to increase, as does the number of departments. Larger size also makes administrative decisions routine, such as hiring, so that rules can be developed that allow faster and more effective decision-making. The structuring of work that comes from these increases in specialization and rules compensate for the increased decentralization, providing compensatory means of control by top management over lower level employees.

Changes in Focus or Emphasis over Time

From Single to Multiple Contingency Fits

Over time, structural contingency theory research has moved from considering one single contingency at a time to considering multiple, simultaneous contingencies. Early structural contingency theory research concentrated on making the argument that, rather than there being a 'one best way,' optimal structures were contingent upon fitting to some factor, e.g., technology (Woodward, 1965) or change in the organization's environment (Burns and Stalker, 1961). They tended to make the case for a particular contingency variable, such as technology or uncertainty. Some subsequent research has been concerned with empirically establishing that structure really is subject to such contingencies in its effect on performance or similar outcomes. This reflects anxiety that not all studies confirm the original contingency relationships, so that empirical validity of individual structural contingencies is still an ongoing concern in the research literature.

Child (1975), however, argued that both size and uncertainty were simultaneously contingencies of structure. And other research has sought to draw together multiple contingences as bearing simultaneously upon various aspects of structure. However, even after several decades, much of the research can be condensed into models in which about seven or so structural variables are each on average fitted to about two contingencies. Thus, fears that structural contingency theory must degenerate into bewildering complexity may be unwarranted.

In one particularly comprehensive synthesis, there are many contingencies which are considered to bear on many aspects of structure (Burton et al., 2002). This might seem to lead to the conclusion that managers have to attend to many simultaneous misfits of the multiple structural variables to the multiple contingency variables, rendering structural change complex. However, Burton et al. (2002) found empirically that the largest misfit accounted for much of the performance loss. Thus, while there can be many combinations of contingencies and structures that produce performance loss, managers only need to identify which is the main misfit in their organization to obtain most of the benefit to be had from changing the structure.

The Contingency of Technology?

While technology is the first structural contingency and pioneered the structural contingency approach its study has not fared well over the years. The pioneering researcher, Woodward (1965), presented sparse evidence of a link between fit and performance. While replication by others was claimed (Zwerman, 1970), this was disputed (Child and Mansfield, 1972; Donaldson, 1976). There have been attempts to maintain the idea of a technology contingency by changing the nature of the contingency or by arguing it exists only for parts of the organization (Reimann, 1980). But these have not been conclusive. Technology should be thought of as part of the research effort that pioneered structural contingency theory, but not as part of the established contemporary core structural contingency theory. This is important because the neo-institutional theory of organizations school gives as a reason for rejecting contingency theory that technology failed to be a useful contingency in their studies of educational systems

(Meyer and Scott, 1983). But if technology is not a core contingency of structural contingency theory its failure in that setting does not justify a rejection of structural contingency theory.

From Curvilinear to Linear Fits

The simplest fit of a structural variable to the contingency variable is that as the contingency level increases so too does the fitting level of the structural variable. Most research takes this simple, linear approach (e.g., Keller, 1994) and this may make the best theoretical sense for many fits of structural variables to their contingencies. However, the fitting structural level could have a curvilinear relationship to the contingency variable. Indeed the early depiction of a structural contingency relationship by Woodward (1965) shows that the level of the structural variable that has the highest performance first increases and then decreases as the technology contingency variable increases. The theoretical interpretation is that work initially involves skill and requires close supervision, but then mass production assembly lines entail less skill and supervision, while, finally, continuous production (e.g., oil refinery) has few highly skilled workers responsible for expensive, automated production systems. However, it can be questioned whether the apparent curvilinearity is anything more than sampling error in an otherwise linear relationship. Therefore, the use of linear, rather than curvilinear, fits of structure to contingencies may be justifiable.

From Differentiation to Interdependence

Initially, Lawrence and Lorsch (1967) proposed that organizations facing predictable environments only needed small differences between their functional departments and so only needed weak integration to fit (as explained above). In contrast, organizations facing unpredictable environments had larger differences between their functional departments and so needed strong integration to fit. Thus, differentiation must be matched by integration for high performance to result. This became known as the 'differentiation and integration (DnI),' approach. It held in their original study of fairly undiversified business firms, but in a later study of diversified corporations, there was no relation between the match of differentiation–integration and performance. This led Lawrence and Lorsch to reinterpret their findings in light of the theory of interdependence of Thompson (1967).

Thompson distinguished degrees of interdependence between parts of an organization: pooled where there was no direct connection, sequential where one part did its job before another worked on it, and reciprocal where work passes back and forth between two parts. He argued that pooled interdependence was fitted by rules, sequential interdependence was fitted by planning, and reciprocal interdependence was fitted by mutual adjustment (for this to hold these three types of interdependence have to be successively higher in their uncertainty).

Lawrence and Lorsch saw that firms of low rate of product change have only sequential interdependence between their functional departments (e.g., Sales and Production). This can be handled effectively by centralized planning. In contrast, firms in environments of high rate of product change have work that flows back and forth unpredictably between Sales, Production, and R&D. Thus, there is reciprocal interdependence between their functional departments. To be handled effectively, this requires individuals and project teams that liaise between the functional departments. Therefore, a medium level of integration fits a medium level of interdependence (i.e., sequential), while a high level of integration fits a high level of interdependence (i.e., reciprocal). Hence, integration primarily needs to fit interdependence.

In a diversified corporation, its divisions have no direct task connections, that is to say, there is only pooled interdependency, the lowest level of interdependence between the divisions. Thus, they can be adequately coordinated by a low-level integration, such as by having rules about the financial reporting of the divisions to the corporate head office. It does not matter whether the divisions are lowly differentiated from each other, or highly differentiated, because there is no direct task connection between them and so little integration needed. Hence, their integration is not causally related to their differentiation – DnI does not hold.

Configurationism or Cartesianism?

Some early research postulated that each level of a contingency variable (e.g., size) is fitted by a different level of a structural variable (e.g., specialization), so that fit is a continuous line (Child, 1975). This has been called the Cartesian view. In contrast, some researchers hold that there are only few fits, which they call configurations or gestalts. These are held to be widely separated, so that an organization rarely moves from one fit to another, other than by a quantum jump, and may fall back to the first fit. Thus, such configurationism implies that organizations take only few values on the contingency and structural variables (e.g., only simple structures that are of small size and low formalization, or machine bureaucracies that are large and high on formalization), whereas empirical research shows that they take many values, supporting Cartesianism. Configurationism also implies that organizations seldom change their values on the contingency and structural variables, whereas empirical research shows that they frequently do so. Furthermore, configurationism implies that when organizations do change it is through large changes in the contingency and structural variables, whereas empirical research shows that they typically make only incremental changes. Cartesianism holds that, because fits form a continuous line, an organization can move from one fit to an adjacent fit, and in that way make frequent changes and change considerably over time through an incremental process. This is a more plausible and empirically backed view than Configurationism (Donaldson, 2001). However, some configurationists (Greenwood and Hinings, 1993) are referring to the cognitive configurations in the minds of people and these can be few and discrete, so the idea of cognitive configurations is sound.

Structures Fitting Contingencies or Other Structures?

The idea that the fit of structure to a contingency affects performance was extended to the idea that structural variables needed to fit *other structural variables* to attain high performance.

This opens the door to dispensing with there being any fit to contingency factors, which means managers need not be restricted in their choices of structure to those that fit some contingency factor.

Equifinality

Equifinality here means many ways of achieving the same end. Applied in the mainstream contingency literature, it means that there is not just one structure that fits the circumstances of an organization, so that multiple structures or combinations of structural variables fit the organization, allowing its managers a wide realm of choice.

Contingency Imperative or Strategic Choice?

The picture of organizational change in the pioneering work is that changes in contingencies cause changes in structure, so the organization moves from one fit to another. This seems an almost automatic process in which managerial choices of structure are determined by a contingency imperative. Child (1972) challenged this view and asserted strategic choice. He argues that various factors intervene to loosen the connection between contingency change and structural change, so that managers exercise choice. Performance lost from misfit may be compensated for by performance from other sources, so that performance maintains an acceptable level and change is avoided. Misfit may be regained through adjusting contingency to structure, avoiding having to adopt a new structure. Managerial perceptions, beliefs, and interests may mediate in the change process. Also, if fit is really just to other structural variables, this also avoids having to bow to contingency dictates. And if structural combinations are equifinal then this also avoids having to adopt a particular structure because of contingencies.

This strategic choice (Child, 1972) revision of structural contingency theory has proved popular. However, Donaldson (2001) has offered a rebuttal of many of the arguments for strategic choice. In its place he has offered structural adaptation to regain fit (Donaldson, 2001) which holds that contingency change causes structural change but after a delay until performance drops below the acceptable level, thereby retaining one of Child's mediating processes.

Current and Emerging Directions in Theory and Research

In the early 1980s, there were suggestions that the topic of organizational design was matured out. And there has been the claim that research on structural contingency theory has stopped from the early 1980s. A critique of structural contingency theory was published around that time that was quite influential (Schoonhoven, 1981), though not clearly amounting to a refutation of structural contingency theory. However, subsequent research into structural contingency theory has been conducted and continues to appear in the published literature (e.g., Hill et al., 1992; Jennings and Seaman, 1994; Keller, 1994; for a fuller review see Donaldson, 2001). This is in part because organization design remains a topic of practical importance in business schools and structural contingency theory continues to be the core of the way it is studied (Jensen et al., 2010; Burton and Obel, 2011). As lately as 2010, scholars have called for renewed efforts into organizational design research involving structural contingency theory (in synthesis with other contemporary organizational theories such as institutional theory) (Greenwood and Miller, 2010). There are also moves to apply structural contingency theory out to other domains, such as operations management (Sousa and Voss, 2008), and to topics such as outsourced work (Hsiao-Hui et al., 2011) and supply chains (Wong et al., 2002).

Isoperformance and Heteroperformance

The core of structural contingency theory is that fit of structure to contingency leads to higher performance than does misfit between them. Early work provided some empirical evidence that firms in misfit had lower performance than firms in fit, with misfit occurring through structure being either below (underfit) or above (overfit) the fitting level (Woodward, 1965). Subsequent work held that the farther away from fit, that is, the greater the misfit, the lower performance would be as a consequence. Moreover, each fit produces the same level of high performance as any other fit, a view known as isoperformance (iso meaning equal).

For contingencies that are internal to the organization, however, the problem arises as to why its managers should increase its level of the contingency variable such as size, which is clearly costly, if there is no performance gain. An attempt to resolve this problem has led to the creation of heteroperformance theory ('hetero' because performance varies) (Donaldson, 2001). This states that the fit to a higher level of the contingency variable (e.g., size) produces higher performance than the fit to the next lower level of the contingency variable. For example, an increase in size means there are more employees in a firm to produce output, so that, if it adopts the level of (say) specialization that fits that size, that fit will produce more performance than will the fit of specialization to the lower level of size. Specialization fitting size produces performance, but also increasing size increases performance. Hence, there is an incentive for managers to move their organization from it being in fit to a lower level of size to it being in fit to a higher level of size. By moving incrementally from one fit to the next, the organization can grow over time from being small to large. This may be interpreted as an explanation as to why there is such a range in size of organizations. The same argument applies for other internal contingencies.

Isoperformance holds for external contingencies, however. Where, for example, the environment of an organization becomes more uncertain, that raises the level of that contingency variable without the managers of that organization doing anything to increase the uncertainty of their organization's environment. The managers require no incentive from some increased performance due to higher levels of the contingency variable. Thus, even where fits all produce the same performance level (i.e., isoperformance) the external contingency can take various values across organizations.

These considerations lead to a modification in structural contingency theory to say that heteroperformance holds for internal contingencies (e.g., size) whereas isoperformance still

holds for external contingencies (e.g., environmental uncertainty). Thus, the core fit-performance model of structural contingency theory has become elaborated. This remains at this time a theoretical development and testing its empirical validity is a task for future research.

Asymmetric Effects on Performance of Overfit and Underfit

A further elaboration of structural contingency theory is the suggestion that the effects of misfit on performance are asymmetric. Traditionally, in structural contingency theory, the effect of misfits on performance are presumed to be symmetric, in that overfit is seen as producing the same performance loss as underfit (Woodward, 1965). However, it has been argued that overfit produces less performance loss than underfit. An organization in overfit has too high a structural level, and therefore has an unnecessarily complex and costly structure. This means the organization attains its objectives, but with excessive costs, which adversely affects its performance, but not seriously. In contrast, an organization in underfit has too low a structural level to fit the contingency and therefore cannot attain its objectives, which is more serious for performance. This theoretical idea needs more empirical research in the future to ascertain its validity and the degree of difference between overfit and underfit in their effects on performance.

Matrix Structures

Matrix structures were seen as a beneficial innovation in the 1970s but became a less popular idea. However, twenty-first century research has found that large corporations are adding matrices to deal with the increasing complexity they face. Scholars had said they were needed for organizations with dual focus (e.g., function and products). Wolf and Egelhoff (2002) defined the fits of matrix structures to various contingencies, based on the information-processing needs posed by the contingencies. Scholars had argued that in multinational corporations, matrix structures would be the fit where product and area diversification were both high, but empirical research found little support. More recently, it has been shown that matrix structures of several types all fit the combination of high local responsiveness and high global integration. Adding foreign product diversification back in, creates a model of three contingencies whose combinations are each fitted by one of nine different structures (Qiu and Donaldson, 2010). This model requires further empirical testing.

Systems Fits?

Structural contingency theory holds that there are bivariate fits of a structural variable to a contingency variable that affects performance. In contrast, it has been asserted that multiple fits between numerous structural and contingency variables lead to the existence of systems fits. These are said to have holistic or emergent types of properties such that the effect on performance of a systems fit is different from that would obtain if the bivariate fits were just summed together. While this idea has gained some currency, however, it is noteworthy that in the model proposing systems fits the overall effect on performance is really just the sum of the effects of the bivariate fits (see Donaldson, 2001).

Selection?

On occasion, the word selection has been used to describe the fact that structures are brought into fit. However, given the influence of organizational (or population) ecology theory in organizational theory, selection has connotations of natural selection through death and birth processes, whereas the idea of 'selection' here is that managers choose structures. The insightful point that accompanies this language is that managers may be constrained in the range of structures they can choose because of constraints on them, e.g., middle managers' limited authority in the face of head office instructions. The term managerial choice (Donaldson, 2001) might be preferable to selection in conveying this idea.

New Contingency Variables?

Above we discussed three main contingencies: uncertainty, size, and strategy. However, structural contingency theory is amenable to the identification of more contingency factors. Knowledge has more recently become of interest to organizational theories and structural variables about organizational knowledge have been shown to have fits with contingencies (Birkinshaw et al., 2002). The development of new contingency-structural fits is an open agenda for structural contingency theory in the future.

Links to Databases and Software

Computer software is available that gives the structure needed for an organization to fit its contingencies. The analyst enters the levels of the contingencies of an organization and the software gives the optimal structure. It is called 'OrgCon.' It draws on a comprehensive set of contingencies and corresponding structural factors (Burton and Obel, 2004).

See also: Automation: Organizational Studies; Bureaucracy and Bureaucratization; Design and Form: Organizational; Diversification, Relatedness, and the New Logic of Co-Creation; Ecology: Organizations; Learning: Organizational; Management: General; Multinational Corporations; Organizational Decision Making; Organizations: Authority and Power; Strategic Adaptation; Strategic Management; Technology and Organization.

Bibliography

Birkinshaw, Julian, Nobel, Robert, Ridderstrale, Jonas, 2002. Knowledge as a contingency variable: do the characteristics of knowledge predict organization structure? Organization Science 13 (3), 274–289.

Burns, Tom, Stalker, G.M., 1961. The Management of Innovation. Tavistock, London.

Burton, R., Lauridsen, J., Obel, B., 2002. Return on assets loss from situational and contingency misfits. Management Science 48, 1461–1485.

Burton, R., Obel, B., 2004. Strategic Organizational Diagnosis and Design: The Dynamics of Fit, third ed. Kluwer, Boston.

Burton, Richard M., Obel, Børge, 2011. Computational modeling for what-is, what-might-be, and what-should-be studies – and triangulation. Organization Science 22 (5), 1195–1202.

Child, John, 1972. Organizational structure, environment and performance: the role of strategic choice. Sociology 6, 1–22.

Child, John, 1975. Managerial and organizational factors associated with company performance, part 2: a contingency analysis. Journal of Management Studies 12, 12–27.

Child, John, Mansfield, Roger, 1972. Technology, Size and Organization Structure. Sociology 6, 369–393.

Donaldson, Lex, 1995. Contingency theory. In: Pugh, D.S. (series Ed.), History of Management Thought Series, vol. 9. Dartmouth Publishing, Aldershot, Hants, UK.

Donaldson, Lex, 1976. Woodward, Technology, organisational structure and performance - a critique of the universal generalisation. Journal of Management Studies 13, 255–273.

Donaldson, Lex, 2001. The Contingency Theory of Organizations. Sage, Thousand Oaks, California.

Greenwood, Royston, Hinings, C.R., 1993. Understanding strategic change: the contribution of archetypes. Academy of Management Journal 36 (5), 1052–1082.

Greenwood, Royston, Miller, Danny, 2010. Tackling design anew: getting back to the heart of organizational theory. Academy of Management Perspectives 24, 78–88.

Hill, Charles W.L., Hitt, Michael A., Hoskisson, Robert E., 1992. Cooperative versus competitive structures in related and unrelated diversified firms. Organization Science 3, 501–521.

Jennings, Daniel F., Seaman, Samuel L., 1994. High and low levels of organizational adaptation: an empirical analysis of strategy, structure and performance. Strategic Management Journal 15, 459–475.

Jensen, Wickstrøm Kent, Håkonsson, Dorthe Døjbak, Burton, Richard M., Obel, Børge, 2010. The effect of virtuality on the functioning of centralized versus decentralized structures – an information processing perspective. Computational and Mathematical Organization Theory 16, 144–170.

Keller, Robert T., 1994. Technology-information processing fit and the performance of R&D project groups: a test of contingency theory. Academy of Management Journal 37, 167–179.

Lawrence, Paul R., Lorsch, Jay W., 1967. Organization and Environment: Managing Differentiation and Integration. Harvard University, Graduate School of Business Administration, Division of Research, Boston.

Lee, Hsiao-Hui, Pinker, Edieal J., Shumsky, Robert A., 2011. Outsourcing a two-level service process. Simon School Working Paper No. FR 11–23, Simon Graduate School of Business. University of Rochester.

Lorsch, Jay W., Allen, Stephen A., 1973. Managing Diversity and Inter-Dependence: An Organizational Study of Multidivisional Firms. Division of Research, Graduate School of Administration. Harvard University, Boston, MA.

Meyer, John W., Scott, W. Richard, Rowan, B., Deal, T.E., 1983. Organizational Environments: Ritual and Rationality. Sage, Beverly Hills, CA.

Qiu, J.X.J., Donaldson, L., 2010. the cubic contingency model: towards a more comprehensive international strategy-structure model. Journal of General Management 36 (1), 81–100.

Reimann, Bernard C., 1980. Organization structure and technology in manufacturing: system versus work flow level perspectives. Academy of Management Journal 23, 61–77.

Schoonhoven, C.B., 1981. Problems with contingency theory: testing assumptions hidden within the language of contingency theory. Administrative Science Quarterly 26, 349–377.

Sousa, R., Voss, C.A., 2008. Contingency research in operations management practices. Journal of Operations Management 26 (6), 697–713.

Thompson, James D., 1967. Organizations in Action. McGraw-Hill, New York.

Wolf, J., Egelhoff, W.G., 2002. Research notes and commentaries: a re-examination and extension of international strategy-structure theory. Strategic Management Journal 23 (2), 181–189.

Wong, C.Y., McFarlane, D., Zaharudin, A., Ahmad, Agarwal, V., 2002. The intelligent product driven supply chain. Systems, Man and Cybernetics, IEEE International Conference vol. 4, pp. 1–6.

Woodward, Joan, 1965. Industrial Organization: Theory and Practice. Oxford University Press, Oxford, UK.

Zwerman, William L., 1970. New Perspectives on Organization Theory: An Empirical Reconsideration of the Marxian and Classical Analyses. Greenwood, Westport, CT.

Structural Equation Modeling

Shawn Bauldry, University of Alabama at Birmingham, Birmingham, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Kaplan, volume 22, pp. 15215–15222, © 2001, Elsevier Ltd.

Abstract

Structural equation modeling refers to a broad statistical framework that involves the specification of a substantive or theoretical model, potentially involving latent or unobserved variables, and the testing of the model with respect to observed data. This article presents the general model, discusses model identification, provides an overview of estimators for structural equation models, outlines testing both the overall fit of a hypothesized model and specific parameters, and notes a couple of notable developments with structural equation models.

Structural equation modeling refers to a broad statistical framework that involves the specification of a substantive or theoretical model and the testing of the model with respect to observed data. Structural equation models, with roots in psychometrics, econometrics, sociological methods, and earlier work in genetics, allow for the presence of unobserved, or latent, variables and the specification of systems of equations (Bollen, 1989). The inclusion of latent variables allows researchers to explicitly address measurement error in observed variables, which is an important feature of the structural equation modeling framework. This article provides an overview of structural equation models including a consideration of both the traditional general model as well as an overview of important recent developments in structural equation modeling.

The article is organized as follows. The first section presents the traditional general structural equation model. The second section discusses model identification, an issue that is typically more prominent in structural equation models than in many other statistical models. The third section provides a brief overview of commonly used estimators for structural equation models, including estimators for categorical data. One particularly useful feature of structural equation models is the ability to test or assess the overall fit of the models. The fourth section presents a number of different approaches to testing structural equation models. Finally, the fifth section provides a brief note of three important extensions or developments of structural equation modeling.

The General Model

The general structural equation model as outlined by Jöreskog (1973) consists of two parts: (1) the latent variable model links latent variables to each other via systems of simultaneous equations and (2) the measurement model links latent variables to observed variables via a restricted (confirmatory) factor model. Using a slightly modified version of Jöreskog's (1977) LISREL notation, the latent variable model is

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha}_\eta + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \qquad [1]$$

where i indexes observations, $\boldsymbol{\eta}$ is a vector of endogenous latent variables, $\boldsymbol{\xi}$ is a vector of exogenous latent variables, and $\boldsymbol{\zeta}$ is a vector of disturbance terms for each of the endogenous latent variables. The parameters in the model include a vector of intercepts ($\boldsymbol{\alpha}_\eta$), a matrix of regression coefficients that gives the effects of the latent endogenous variables on each other ($\mathbf{B}$), and the effects of the latent exogenous variables on the latent endogenous variables ($\boldsymbol{\Gamma}$). The matrix $\boldsymbol{\Phi}$ contains the variances and covariances among the exogenous latent variables and the matrix $\boldsymbol{\Psi}$ contains the variances and covariances among the disturbances for the latent endogenous variable.

The measurement model links observed variables to endogenous and exogenous latent variables. The equations for the measurement model are

$$\mathbf{y}_i = \boldsymbol{\alpha}_\mathrm{y} + \boldsymbol{\Lambda}_\mathrm{y}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \qquad [2]$$

and

$$\mathbf{x}_i = \boldsymbol{\alpha}_\mathrm{x} + \boldsymbol{\Lambda}_\mathrm{x}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i \qquad [3]$$

where $\mathbf{y}$ is a vector of observed indicators of the ηs (latent endogenous variables), $\mathbf{x}$ is a vector of observed indicators of the ξs (latent exogenous variables), and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are, respectively, vectors of disturbances for the observed indicators of the latent variables. The matrix of coefficients, often referred to as factor loadings, $\boldsymbol{\Lambda}_\mathrm{y}$ gives the effects of the ηs on the ys and similarly $\boldsymbol{\Lambda}_\mathrm{x}$ gives the effects of the ξs on the xs. The measurement models also include vectors of intercepts $\boldsymbol{\alpha}_\mathrm{y}$ and $\boldsymbol{\alpha}_\mathrm{x}$. Finally, the matrix $\boldsymbol{\Theta}_\varepsilon$ contains the variances and covariances among the $\boldsymbol{\varepsilon}$ disturbances and the matrix $\boldsymbol{\Theta}_\delta$ contains the variances and covariances among the $\boldsymbol{\delta}$ disturbances.

There are a number of assumptions that typically underlie the latent variable model and measurement model. First, it is assumed that expected values for all the disturbances equal 0, i.e., $E[\boldsymbol{\zeta}_i] = E[\boldsymbol{\varepsilon}_i] = E[\boldsymbol{\delta}_i] = 0$. Second, it is assumed that the exogenous latent variables are uncorrelated with the disturbances in the latent variable model and that the endogenous and exogenous latent variables are uncorrelated with the disturbances in the measurement model, i.e., $\mathrm{Cov}[\boldsymbol{\xi}_i, \boldsymbol{\zeta}_i] = \mathrm{Cov}[\boldsymbol{\eta}_i, \boldsymbol{\varepsilon}_i] = \mathrm{Cov}[\boldsymbol{\xi}_i, \boldsymbol{\delta}_i] = 0$. Finally, it is generally assumed that the disturbances across eqns [1]–[3] are uncorrelated with each other, but this assumption is less essential than the previous two.

Structural equation modeling is a methodology designed primarily to test substantive theories. As such, a theory should be sufficiently developed to suggest that certain latent variables do not affect other latent variables, certain observed variables do not load on certain latent variables, and certain disturbances and measurement errors do not covary. This implies that some

of the elements of $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Lambda}_y$, $\mathbf{\Lambda}_x$, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}_\varepsilon$, and $\mathbf{\Theta}_\delta$ are fixed to zero by hypothesis. The remaining parameters are free to be estimated. The pattern of fixed and free parameters implies a specific structure for the moment matrices of the observed variables. In this way, structural equation modeling can be seen as a special case of a more general covariance structure model defined as $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are, respectively, the population mean vector and covariance matrix and $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\mathbf{\Sigma}(\boldsymbol{\theta})$, referred to as the 'implied moment matrices,' are matrix valued functions of the parameter vector $\boldsymbol{\theta}$ containing all the parameters of the hypothesized model.

Identification

A prerequisite for the estimation of the parameters of the model is to establish whether the model is identified, i.e., whether it is possible to find unique values for each of the parameters in the model contained in $\boldsymbol{\theta}$. The elements in $\boldsymbol{\theta}$ are identified if they can be expressed uniquely in terms of the elements of the population moment matrices $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. If all elements in $\boldsymbol{\theta}$ are identified then the model is identified. If there are just enough elements in the population moment matrices to identify the parameters, then the model is said to be 'just identified.' If there are more than enough elements, then the model is said to be 'overidentified.' The distinction is important in that global tests of model fit (discussed below) are available only for overidentified models. Identification concerns the existence of unique values for the parameters and thus should be established prior to estimation (although, as noted below, one of the popular means of establishing identification occurs in the process of estimation).

There are three approaches to establishing the identification of a structural equation model in the general model framework. The first approach is to determine whether the model conforms with one of the known rules for identified models (see Bollen (1989) for a discussion of a number of well-known rules of identification). The second approach, which is perhaps the most frequently used in practice, is to rely on empirical identification tests that are typically incorporated into the software used to estimate structural equation models (SEMs). Empirical tests of identification, however, assess local rather than global identification (i.e., whether unique values for the parameters exist in a local neighborhood rather than across the complete parameter space) and thus can be misleading in some cases. The third approach is to prove through algebraic means that all the parameters can be uniquely expressed in terms of the population moment matrices $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. This approach can be quite challenging with large models as the algebra becomes complex, although computer algebra systems can assist in the process (Bollen and Bauldry, 2010). Nonetheless, it is the only foolproof method of establishing identification for models that do not conform to one of the known rules of identification.

Estimation

Parameter estimation involves obtaining estimates of the free parameters of the model, subject to the constraints imposed by the fixed parameters of the model. The goal is to apply a model to the sample moment matrices $\mathbf{z}$ and $\mathbf{S}$, which are estimates of the population moment matrices $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ that are assumed to follow a hypothesized structural equation model. Estimates are obtained such that the discrepancy between the estimated moment matrices $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}(\hat{\boldsymbol{\theta}})$ and the sample moment matrices $\mathbf{z}$ and $\mathbf{S}$ are as small as possible. There are a number of estimators that have been developed for structural equation models.

Maximum Likelihood Estimators

The most frequently used estimator for structural equation models is the maximum likelihood (ML) estimator. The ML estimator has a number of desirable properties (Bollen, 1989). ML estimators are asymptotically unbiased, consistent, and asymptotically efficient. In addition, ML estimators are asymptotically normally distributed, which ensures that the ratio of the parameters to their standard errors approximates a Z distribution in large samples. ML estimators also provide an overall model fit test (for overidentified models). Finally, as described below, ML estimators can be extended relatively easily to address missing data subject to certain assumptions.

The ML estimator was originally derived under the assumption of multivariate normality. Browne (1987), however, demonstrated that ML estimators maintain their desirable asymptotic properties even with nonnormal data as long as there is no excess multivariate kurtosis. In addition, numerous simulation studies find that the ML estimator is largely robust to departures from normality in finite samples across a range of research contexts. Finally, corrections have been developed for chi-square test statistics and standard errors of parameter estimates that adjust for nonnormality when using ML estimators (Satorra and Bentler, 1994).

Model-Implied Instrumental Variable Estimators

ML estimators are sometimes referred to as full-information estimators in that they involve the simultaneous estimation of all of the parameters of a given model using information from the whole system of equations. By contrast, limited information estimators operate on single equations or subsets of equations. An important class of limited information estimators is instrumental variable estimators. In a series of papers, Bollen and colleagues developed what are now known as model-implied instrumental variable estimators for structural equation models (Bollen, 1996; Bollen and Maydeu-Oliveres, 2007; Bollen et al., forthcoming). These estimators include families of two-stage least squares (2SLS) and generalized method of moments (GMM) estimators. The instruments used in these estimators arise from the specification of the model, hence the term 'model-implied' instruments.

As limited information estimators, the model-implied instrumental variable estimators are not as asymptotically efficient as full-information estimators, but they have a number of advantages over the more commonly used ML estimators for structural equation models. First, the fact that they operate on single equations or subsets of equations helps isolate potential bias due to misspecification in other parts of the model. Second, these estimators are

asymptotically distribution free and thus do not rely on a multivariate normality assumption. Finally, simulation studies suggest that the efficiency trade-off as compared with a full-information estimator is not that large. To date, however, model-implied instrumental variable estimators have not enjoyed widespread use in empirical research using structural equation models.

Estimators for Categorical Observed Variables

Although ML estimators are robust to moderate departures from normality, it is still desirable to have estimators that explicitly account for categorical observed variables. There are broadly two classes of estimators for categorical observed variables in structural equation models. The first class is based on positing that each of the categorical observed variables has an underlying latent variable that is continuous and normally distributed (Olsson, 1979; Muthén, 1984). The correlations among these underlying latent variables, sometimes referred to as polychoric correlations, can be used in combination with a weight matrix in diagonally weighted least squares estimator (Jöreskog and Sörbom, 1993; Muthén, 1984). The second class of estimators involves extending the measurement model given in eqns [2] and [3] to be a generalized linear model in which different link functions allow for different types of variables (Skrondal and Rabe-Hesketh, 2004). ML estimation with this approach, in contrast to the model assuming multivariate normality, cannot be based on sufficient statistics and thus relies on more computationally demanding approaches that typically involve 'integrating out' the latent variables.

Complex Sampling Designs and Missing Data

In the absence of explicit information to the contrary, estimation methods such as ML assume that data are generated according to simple random sampling. Perhaps more often than not, however, structural equation models are applied to data that have been obtained through something other than simple random sampling. For example, when data are obtained through multistage sampling, the usual assumption of independent observations is violated resulting in predictable biases. Estimators for structural equation models that account for complex sampling designs have been developed and are available in most of the widely used software packages for estimating structural equation models (Muthen and Satorra, 1995).

Traditionally, statistical procedures such as structural equation modeling assume that each unit of analysis has complete data. However, for many reasons, units may be missing values on one or more of the variables under investigation. Numerous ad hoc approaches exist for handling missing data, including listwise and pairwise deletion. However, these approaches assume that the missing data are unrelated to the variables that have missing data and any other variable in the model (i.e., missing completely at random in Little and Rubin's (2002) terminology), an often untenable assumption.

There are two contemporary approaches to addressing missing data that can be used with structural equation models and rely on the less stringent assumption that the missing data is unrelated to the variable missing data conditional on the variables in the model (i.e., missing at random in Little and Rubin's (2002) terminology). The first procedure, the one most commonly used for structural equation modeling, is to use a casewise ML (or direct ML) estimator. This estimator is an extension of the standard ML estimator that allows one to maintain all the cases in the estimation of parameters (Arbuckle, 1996). The second procedure is to use multiple imputation, which involves constructing M complete data sets using an imputation model to fill in the missing data along with some random error to reflect the imputation process, estimating the structural equation model on each complete data set, and then combining the estimates using Rubin's rules (Little and Rubin, 2002).

Testing

Once the parameters of the model have been estimated, the model can be tested for global fit with the data. In addition, hypotheses concerning individual parameters can be evaluated.

Testing Global Model Fit

A feature of ML estimators is that with overidentified models one can explicitly test the hypothesis that the model fits data. The null hypothesis states that the specified model holds in the population, i.e., H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. The corresponding alternative hypothesis is that the population covariance matrix is some other symmetric positive definite matrix. With an ML estimator, the fitting function is given by F_{ML}. The function $(N-1)$ F_{ML} has an asymptotic chi-square distribution with $\frac{1}{2}(p+q)(p+q+1)-t$ degrees of freedom where p is the number of observed indicators of $\boldsymbol{\eta}$ (ys), q is the number observed indicators of $\boldsymbol{\xi}$ (xs), and t is the number of free parameters in the model. A significant chi-square statistic indicates that $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and thus H_0 is rejected and at least one component of the specified model is incorrect.

Similar model fit statistics are available with model-implied instrumental variable estimators. In the context of structural equation models, the well-known overidentification tests associated with instrumental variable estimators, such as Sargan's chi-squared statistic for 2SLS estimators and Hansen's J statistic for GMM estimators (Sargan, 1958; Hansen, 1982), serve as tests of model misspecification.

Testing Hypotheses regarding Individual Parameters

In addition to a global test of whether the model holds exactly in the population, one can also test hypotheses regarding the individual fixed and free parameters in the model. There are a few alternative ways to evaluate the fixed and freed elements of the model vis-à-vis overall fit.

One important method rests on the difference between the chi-square statistics comparing a given model against a less restrictive model. Recall that the initial hypothesis, say $H_{0\mathrm{R}}$, is tested against the alternative hypothesis H_{a} that $\boldsymbol{\Sigma}$ is some other symmetric positive definite matrix. Consider a second hypothesis, say $H_{0\mathrm{U}}$ that differs from $H_{0\mathrm{R}}$ in that a single parameter that was restricted to 0 in model R is now relaxed in model U. Note that the alternative hypothesis is the same

in both cases. Therefore, the change in the chi-square value between the two models can be used to test the improvement in fit due to the relaxation of the restriction. The distribution of the difference between the two chi-square tests is itself distributed as chi-square with degrees of freedom equal to the difference in degrees of freedom between the model under H_{0R} and the less restrictive model under H_{0U}. In the case of a single restriction described here, the chi-square difference test is evaluated with one degree of freedom.

Alternative Methods of Assessing Model Fit

Since the early 1980s, attention has focused on the development of alternative indices that provide relatively different perspectives on the fit of structural equation models. The development of these indices has been motivated, in part, by the known sensitivity of the chi-square statistic to large sample sizes. Other classes of indices have been motivated by a need to rethink the notion of testing exact fit in the population – an idea that is deemed by some to be unrealistic in most practical situations.

In this section, the discussion of alternative fit indices is divided into three categories: (1) measures based on comparative fit to a baseline model, (2) measures based on population errors of approximation, and (3) the Bayesian information criterion (BIC). In addition to the chi-square test statistic, each of these alternative fit indices have strengths and weaknesses, so the current recommendation for empirical research using structural equation models is to report multiple fit statistics and assess the pattern of information across the statistics regarding model fit.

Measures Based on Comparative Fit to a Baseline Model

The basic idea behind comparative fit indices is that the fit of the model is compared to the fit of some baseline model that usually specifies complete independence among the observed variables. The fit of the baseline model will usually be fairly large and thus the issue is whether one's model of interest is an improvement relative to the baseline model. The index is typically scaled to lie between 0 and 1, with 1 representing perfect fit relative to this baseline model. The sheer number of comparative fit indices precludes a detailed discussion of each one. A subset of indices that serve to illustrate the basic ideas is considered here.

The most straightforward, although rarely used, example of a comparative fit index (CFI) is the normed fit index (NFI) proposed by Bentler and Bonett (1980). This index can be written as

$$\mathrm{NFI} = \frac{T_b - T_m}{T_b}, \qquad [4]$$

where T_b is the chi-square test statistic for the model of complete independence (the so-called baseline model) and T_m is the chi-square test statistic for the target model of interest. This index, which ranges from 0 to 1, provides a measure of the extent to which the target model is an improvement over the baseline model.

The NFI assumes a true null hypothesis and therefore a central chi-square distribution of the test statistic. However, an argument could be made that the null hypothesis is never exactly true and that the distribution of the test statistic can be better approximated by a noncentral chi-squared with noncentrality parameter λ. An estimate of the noncentrality parameter can be obtained as the difference between the statistic and its associated degrees of freedom. Thus, for models that are not extremely misspecified, an index developed by McDonald and Marsh (1990) and referred to as the relative noncentrality index (RNI) can be defined as

$$\mathrm{RNI} = \frac{(T_b - \mathrm{df}_b) - (T_m - \mathrm{df}_m)}{T_b - \mathrm{df}_b}, \qquad [5]$$

where df_b and df_m are the degrees of freedom, respectively, for the baseline and target model. The RNI can lie outside the 0–1 range. Bentler (1990) adjusted the RNI so that it would lie in the range of 0–1. This adjusted version is referred to as the CFI and is often preferred in practice, although it can obscure overfit as indicated by values of the RNI that are greater than 1.

Measures Based on Errors of Approximation

It was noted earlier that the chi-square test assesses an exact null hypothesis that the model fits perfectly in the population. In reality, it is fairly unlikely that the model will fit the data perfectly. In addition, as the sample size grows larger, the power of the chi-square test increases and thus the ability to detect potentially trivial model misspecifications arises. An alternative approach is to assess whether the model fits approximately well in the population.

Based on the work of Steiger and Lind (1980), it is possible to assess the approximate fit of a model in the population. The method of Steiger and Lind (1980) utilizes the root mean square error of approximation (RMSEA). The RMSEA is given by

$$\mathrm{RMSEA} = \sqrt{\frac{T_m - \mathrm{df}_m}{\mathrm{df}_m(N-1)}}. \qquad (6)$$

In utilizing the RMSEA for assessing approximate fit, a formal hypothesis testing framework is employed. Typically an RMSEA value ≤ 0.05 is considered a 'close fit.' This value serves to define a formal null hypothesis of close fit. In addition, a 90% confidence interval around the RMSEA can be formed, enabling an assessment of the precision of the estimate (Steiger and Lind, 1980). Practical guidelines suggest that values of the RMSEA between 0.05 and 0.08 are indicative of fair fit, while values between 0.08 and 0.10 are indicative of mediocre fit. Simulation studies, however, indicate that these guidelines do not perform well in small samples.

Bayesian Information Criterion

The BIC is a well-known general approach to model selection that favors more parsimonious models over more complex models (i.e., it adds a penalty based on the number of parameters being estimated in the model) (Schwarz, 1978; Raftery, 1995). One form for calculating the BIC is given by

$$\mathrm{BIC} = T_m - \mathrm{df}_m \ln(N), \qquad (7)$$

where T_m is the chi-square statistic for the hypothesized model. In this form, a BIC greater than 0 favors the saturated model (i.e., the model that allows all observed variables to be intercorrelated with no assumed model structure), while a BIC less

than 0 favors the hypothesized model. Furthermore, the BIC can be used to assess two competing models. Following Jeffrey-Raftery's (1995) guidelines, if the difference in BICs between the two models is 0–2, this constitutes 'weak' evidence in favor of the model with the smaller BIC; a difference in BICs between 2 and 6 constitutes 'positive' evidence; a difference in BICs between 6 and 10 constitutes 'strong' evidence; and a difference in BICs greater than 10 constitutes 'very strong' evidence in favor of the model with smaller BIC.

Developments and Extensions

Structural equation modeling continues to be a dynamic area of research with numerous areas of development and extensions to other types of analytic strategies. Space constraints make it impossible to highlight all the modern developments in structural equation modeling. This section focuses on two increasingly popular classes of structural equation models, growth curve models and latent class models, and the connection between graphical models and SEMs.

Growth Curve Models

Social science researchers increasingly have access to and rely on longitudinal data for analyzing processes that unfold over time and interindividual differences in these processes. Traditionally, the specification of growth models falls within the class of random effects models or multilevel linear models in which multiple observations over time are nested within individuals (Raudenbush and Bryk, 2001; Wooldridge, 2002). Growth curve models have also been incorporated into the structural equation modeling framework by treating the growth parameters as latent variables (Bollen and Curran, 2006). The advantages of the structural equation modeling framework for growth curve modeling lies in its flexibility in specifying and testing models of substantive interest.

Latent Class Models

The defining characteristic of latent class models is the inclusion of categorical latent variables as opposed to the continuous latent variables assumed in the traditional structural equation modeling framework. Latent class models are typically used when a researcher suspects that distinct subgroups or categories of individuals exist in a population. These models have an affinity with the person-oriented approach advocated by Bergman and Magnusson (1997). Recent interest in latent class models has focused on their extension to longitudinal data and the modeling of different growth trajectories (Collins and Lanza, 2010).

Graphical Models and Structural Equation Models

Judea Pearl's work using directed acyclic graphs to represent causal models has been widely influential in statistics and is gaining traction in the social sciences (Pearl, 2000, 2010). There is a natural affinity between Pearl's graphical models and SEMs given their shared focus on causal modeling and figure-oriented notation, although there remain some important differences. For instance, Pearl's framework is non-parametric, whereas SEMs traditionally rest on a linearity assumption (although this can be relaxed to some extent (Bollen and Pearl, 2013)). On the other hand, the SEM framework pays more attention to latent variables and measurement error than has so far been the case with graphical models.

See also: Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis; Factor Analysis and Latent Variable Models in Personality Psychology; Growth Curve Analysis; Indicator: Methodology; Latent Structure and Causal Variables; Linear Hypothesis: Regression (Basics); Statistical Analysis: Multilevel Methods.

Bibliography

Arbuckle, J.L., 1996. Full information estimation in the presence of incomplete data. In: Marcoulides, G.A., Schumacker, R.E. (Eds.), Advanced Structural Equation Modeling: Issues and Techniques. Lawrence Erlbaum Associates, Inc, Mahwah, NJ.

Bentler, P.M., 1990. Comparative fit indices in structural models. Psychological Bulletin 107, 238–246.

Bentler, P.M., Bonett, D.G., 1980. Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin 88, 588–606.

Bergman, L.R., Magnusson, D., 1997. A person-oriented approach in research on developmental psychopathology. Development and Psychopathology 9, 291–319.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Bollen, K.A., 1996. An alternative 2SLS estimator for latent variable models. Psychometrika 61, 109–121.

Bollen, K.A., Bauldry, S., 2010. Model identification and computer algebra. Sociological Methods & Research 39, 127–156.

Bollen, K.A., Curran, P.J., 2006. Latent Curve Models: A Structural Equation Perspective. John Wiley & Sons, Inc, Hoboken, NJ.

Bollen, K.A., Kolenikov, S., Bauldry, S., Model-implied instrumental variable – generalized method of moments (MIIV-GMM) estimators for latent variable models. Psychometrika, forthcoming.

Bollen, K.A., Maydeu-Oliveres, A., 2007. A polychoric instrumental variable (PIV) estimator for structural equation models with categorical variables. Psychometrika 72, 309–326.

Bollen, K.A., Pearl, J., 2013. Eight myths about causality and structural equation models. In: Morgan, S.L. (Ed.), Handbook of Causal Analysis for Social Research. Springer, New York.

Browne, M.W., 1987. Robustness of statistical inference in factor analysis and related models. Biometrika 74, 375–384.

Collins, L.M., Lanza, S.T., 2010. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. John Wiley & Sons, Hoboken, NJ.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Jöreskog, K., 1977. Structural equation models in the social sciences: specification, estimation, and testing. In: Krishnaiah, P.R. (Ed.), Applications of Statistics. North-Holland, Amsterdam.

Jöreskog, K., Sörbom, D., 1993. LISREL 8: Structural Equation Modeling. Scientific Software International, Lincolnwood, IL.

Jöreskog, K.G., 1973. A general model for estimating a linear structural equation system. In: Goldberger, A.S., Duncan, O.D. (Eds.), Structural Equation Models in the Social Sciences. Seminar Press, New York.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley, Hoboken, NJ.

Mcdonald, R.P., Marsh, H.W., 1990. Choosing a multivariate model: noncentrality and goodness-of-fit. Psychological Bulletin 107, 247–255.

Muthén, B., 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 49, 115–132.

Muthen, B.O., Satorra, A., 1995. Complex sample data in structural equation modeling. Sociological Methodology 25, 267–316.
Olsson, U., 1979. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 44, 443–460.
Pearl, J., 2000. The logic of counterfactuals in causal inference. Journal of the American Statistical Association 95, 428–435.
Pearl, J., 2010. The foundations of causal inference. Sociological Methodology 40, 75–149.
Raftery, A., 1995. Bayesian model selection in social research. Sociological Methodology 25, 111–163.
Raudenbush, S.W., Bryk, A.S., 2001. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications, Inc, Thousand Oaks, CA.
Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26, 393–415.
Satorra, A., Bentler, P.M., 1994. Corrections to test statistics and standard errors in covariance structure analysis. In: Eye, A.V., Clogg, C.C. (Eds.), Latent Variable Analysis: Applications for Developmental Research. Sage, Thousand Oaks, CA.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.
Steiger, J.H., Lind, J.C., 1980. Statistically-based tests for the number of common factors. Annual Meeting of the Psychometric Society (Iowa City, IA).
Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge.

Structural Equivalence: Meaning and Measures

Stephen P Borgatti and Travis J Grosser, LINKS Center for Social Network Analysis, Gatton College of Business and Economics, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In social network analysis, two nodes are considered structurally equivalent if they have the same neighborhoods – that is, they are connected to the same others. Initially introduced as a convenience for creating reduced models of networks, it was soon seen as a way to formalize the concept of relational role or position. To the extent that characteristics of nodes are shaped by their social environments, we expect structurally equivalent nodes to develop similar characteristics. Structural equivalence has been used to explain similarities in beliefs and attitudes, the adoption of innovation, the evolution of interfirm networks, political affiliation, the structure of trade among nations, and the effects of technology change.

In social network analysis, two nodes are structurally equivalent to the extent that they have ties to the same third-parties. An example is shown in Figure 1, where nodes B and D are structurally equivalent. Both nodes have an outgoing tie to E, and both have incoming ties from A and C. Two key intuitions underlie the concept. One is the notion of a graph-theoretic niche – i.e., the idea of two nodes sharing the same social environment. The other is the notion of indistinguishableness, meaning that structurally equivalent nodes are structurally identical and indeed would be indistinguishable if not for their labels. Both of these ideas lead to a generic expectation that structurally equivalent nodes will tend to be similar in other, nonstructural, ways as well. For example, to the extent that a node's function, outcomes, behavior, attitudes, etc., are the result of its environment, two nodes with similar environments should develop other similarities as well. Likewise, to the extent that there are consequences to occupying certain structural positions in a network, structurally equivalent nodes should be more likely to have similar outcomes than nonequivalent nodes.

Formal Definition

The most common definition of structural equivalence is based on the same-neighborhood or social niche idea. Given an undirected graph G(V,E), a pair of vertices u and v are structurally equivalent if $N(u) = N(v)$, where $N(x)$ is the set of vertices adjacent to x. In other words, u and v are connected to the same vertices. For directed graphs, we require that $N^i(u) = N^i(v)$ and $N^o(u) = N^o(v)$, meaning that u and v have incoming arcs from the same vertices and outgoing arcs to the same vertices.

Unfortunately, this definition does not quite satisfy the indistinguishableness criterion. For example, when G is undirected and nonreflexive (i.e., no ties from a vertex to itself) if u and v are tied, it is not possible for u and v to be structurally equivalent, even though it is obvious that they are indistinguishable (see vertices A and C in Figure 1). The reason is that u's tie to v requires v to have a tie to v as well, in order to satisfy the equality of neighborhoods condition, but this is impossible if G is nonreflexive. We could patch the definition by requiring only that $N(u) - \{v\} = N(v) - \{u\}$, but this would fail if G was directed with an arc from u to v, as it would find u and v equivalent when clearly they are not (see Figure 2). A variety of other possibilities of this type fail in one circumstance or another.

A better definition is provided by Everett and Borgatti (1994), which they refer to as weak structural equivalence. In their definition, $u \equiv v$ (that is, are equivalent) if there exists an automorphism π of G of the form $\pi = (u,v)$, meaning that u and v are mapped to each other and all other vertices are mapped to themselves. With this definition, A and C in Figure 2 would not be equivalent, but would be if there were a reciprocating arc from C back to A. The Everett and Borgatti definition actually yields a lattice of equivalence relations in which the maximal element is the equivalence in which *all* vertices that can be mapped to each other are placed in the same class.

The lattice of structural equivalences is a sublattice of the lattice of regular equivalences (Everett and Borgatti, 1994). An equivalence is regular if $u \equiv v$ implies that $C(N^i(u)) = C(N^i(v))$ and $C(N^o(u)) = C(N^o(v))$, where $C(w)$ is the equivalence class of w and $C(N(w))$ is the set of equivalence classes observed in the (in- or out-) neighborhood of w. The frequency of occurrence of each class is ignored.

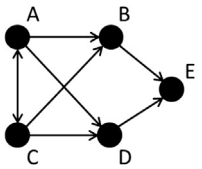

Figure 1 Directed graph.

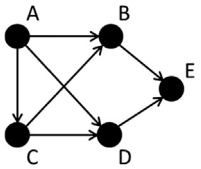

Figure 2 Vertices A and C are not structurally equivalent.

Another way to think about structural equivalences is in terms of blockmodels. Blockmodeling refers to the practice of reordering and partitioning the rows and columns of an adjacency matrix so that the rows and columns of all vertices in an equivalence class are placed next to each other and are visually separated from the vertices in other equivalence classes (see Figure 3). This creates matrix blocks consisting of all dyads (u,v) whose members are involved in the same two equivalence classes.

A distinctive property of structural equivalence is that all matrix blocks consist of either all 1s or all 0s. As a result, there is no loss of information in representing the network by a reduced matrix – the blockmodel – whose rows and columns correspond to the classes of the structural equivalence (see Figure 4).

Blockmodels provide a concise way to describe both empirical and theoretical social structures. For example, the structure shown in Figure 5 is a clique structure, in which ties occur only within class, and Figure 6 is a status hierarchy, in which vertices prefer not to have ties with those below them.

The blockmodel is the adjacency matrix of the image graph of a graph homomorphism induced by the structural equivalence. That is, given a graph G and a structural equivalence S, we construct an image graph H whose vertices are the equivalence classes of S. An edge occurs in H if and only if any member of one class has a tie in G to any member of the other class. Of course, in the case of structural equivalence, if any member of one class has a tie to a member of another class, then all members of the first class have ties to all members of the second one. Hence, there is no loss of information in viewing the image graph. An example of an image graph is shown in Figure 7.

A property of the maximal structural equivalence is that the image graph contains no structurally equivalent vertices. Therefore, we cannot apply structural equivalence iteratively to obtain increasingly abstract representations of a network. Relatedly, it can be shown that, for an adjacency matrix A and a structural equivalence represented by a matrix E where $e_{ij} = 1$ if the vertices corresponding to i and j are equivalent, the Boolean product AE is equal to EA.

	A	B	C	D	E	F	G	H	I	J
A		1	1	1	0	0	0	0	0	0
B	1		1	1	0	0	0	0	0	0
C	1	1		1	0	0	0	0	0	0
D	1	1	1		0	0	0	0	0	0
E	0	0	0	0		0	1	1	1	1
F	0	0	0	0	0		1	1	1	1
G	1	1	1	1	1	1		0	0	0
H	1	1	1	1	1	1	0		0	0
I	1	1	1	1	1	1	0	0		0
J	1	1	1	1	1	1	0	0	0	

Figure 3 Adjacency matrix blocked by structural equivalence.

	{A,B,C,D}	{E,F}	{G,H,I,J}
{A,B,C,D}	1	0	0
{E,F}	0	0	1
{G,H,I,J}	1	1	0

Figure 4 Reduced blockmodel of matrix in Figure 3.

	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

Figure 5 Clique structure.

	1	2	3
1	1	0	0
2	1	1	0
3	1	1	1

Figure 6 Status structure.

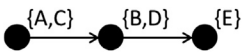

Figure 7 Reduced model of the graph in Figure 1.

Methods

The earliest popular method for finding structurally equivalent nodes in empirical graphs was CONCOR (Breiger et al., 1975). Given the adjacency matrix for an undirected graph, what CONCOR does is correlate the rows of the matrix (ignoring the values along the main diagonal) to produce an initial correlation matrix C(0), then correlate the rows of the correlation matrix, producing C(1), and repeat this process k times until the cells of C(k) are all +1s or −1s. At this point, the vertices can be partitioned into two sets, with all +1s for cells within set and all −1s in cells between sets. The algorithm appears to converge in all cases where correlations are calculable (i.e., each row has positive variance), although this property has not been proved mathematically.

To obtain solutions with more than two equivalence classes, it suffices to apply the algorithm to the subgraph(s) defined by the members of one (or both) initial sets. This process can be applied iteratively until a partition with the desired number of classes is obtained. The final partition can then be used to block the adjacency matrix and construct a reduced image matrix or blockmodel.

The problem of deciding which sets to further bifurcate and which to leave alone is typically handled by using a measure of clustering adequacy. In particular, Panning (1982) recommends an eta coefficient, which is equivalent to the multiple R obtained by regressing the original adjacency matrix on a series

of dummy matrices $X^{(m)}$ where $X^{(m)}(i,j) = 1$ if the (i,j) cell is a member of the mth matrix block. This effectively constitutes an analysis of variance that measures the extent to which the variance in block densities exceeds the mean within-block variances.

An alternative to CONCOR is to submit the initial correlation matrix C(0) to any clustering algorithm. Burt (1976) argued for using the Euclidean distance between rows instead of Pearson correlation, and then applying Johnson's (1967) hierarchical clustering. Euclidean distance allows differences in means and standard deviations of rows to affect the estimate of structural equivalence, whereas correlation ignores such differences (Cronbach and Gleser, 1953). If rows are standardized to have mean 0 and standard deviation 1, the square of the Euclidean distance between two rows equals $2n(1 - r)$, where n is the number of entries in each row and r is the correlation between them.

A completely different approach is to use combinatorial optimization algorithms such as simulated annealing or tabu search (Glover, 1989) to find partitions that minimize errors (an error being a 0 in a 1-block and 1 in a 0-block). This is the approach taken by Batagelj et al. (1992).

Once a partition of nodes into (approximately) structurally equivalent nodes is obtained, it is typically used as a categorical factor in predicting node-level outcomes. For example, the dependent variable could be attitude toward a political candidate measured on a 5-point scale where +2 indicates complete support and −2 indicates complete opposition. A significant relationship between the structural equivalence partition and attitude would suggest that the attitude is, at least in part, determined by position in the network.

An alternative approach is to use the initial correlation (or distance) matrix directly rather than first finding discrete clusters of equivalent nodes and using cluster membership as a variable. In this alternative approach, the correlation between two nodes gives the degree of equivalence between them, and this is used to predict the degree of similarity in their behavior, attitudes, or other outcomes. Besides taking advantage of the continuous degrees of equivalence, this approach has the benefit of being less 'massaged' than the cluster membership variable. Clustering algorithms are numerous and can give widely different answers, leaving open the question of whether the relationship discovered between structural equivalence and an outcome variable would still be found if a different clustering algorithm had been used.

For directed graphs, we correlate not only the rows of the adjacency matrix but the columns as well. This is done by appending the $n - 1$ column entries to the $n - 1$ row entries to obtain vectors of length $2(n - 1)$ for each vertex. It is these vectors that are then correlated with each other. The same approach is used to handle multiple relations. For instance, if both friendship and advice ties were measured for a set of nodes, we would build vectors consisting of the $n - 1$ row entries in the friendship adjacency matrix, the $n - 1$ column entries in the same matrix, the $n - 1$ row entries in the advice matrix, and the $n - 1$ column entries in this matrix, creating a vector of length $4(n - 1)$. When the edges of any of the graphs are weighted, it is advisable to standardize the values in all matrices so that each of them has mean 0 and standard deviation 1. Otherwise, the social relations with the largest edge weights would dominate the measure of equivalence.

History and Significance of the Concept

The term structural equivalence was introduced and formally defined by Lorrain and White (1971) in an article about role structures. Although the title of the piece includes the term structural equivalence, the concept played a minor role in the article. Their interest was in examining compositions of social relations. For example, given the relations 'parent of' (P) and 'spouse of' (S), we can compose the PS relation, which is the 'parent of the spouse of' relation, otherwise known as the in-laws. Similarly, $P^{-1}P$ is the 'child of the parent of' relation, which we know as the sibling relation. A more complex compound relation is $P^{-1}P^{-1}PP$, which is cousin. Some of these compositions are identical to others, a condition that can be expressed as an equation. For example, if it were true that friends of friends are friends, we would find that FF = F. Lorrain and White noted that they could also declare equalities like this by fiat, either as an expression of a cultural concept (e.g., brother of spouse = spouse of sister = brother-in-law), or because empirically friends of friends were, often enough, friends. Making these kinds of identifications could radically simplify the system of compositions generated by a set of starter relations (such as P and S), and simplification or 'reduction' was the goal of their analysis.

Lorrain and White realized that if relations could be found or declared equivalent, so could nodes, and this would serve the same purpose of creating a simplified abstract model of a complex observed set of concrete relationships among a set of nodes. For example, if, in **Figure 1**, we are able to equate A and C, and also B and D, we can describe the network in more abstract terms, as shown in **Figure 7**, as a system in which there are terminal nodes (E), intermediaries (B and D), and sources (A and C).

Structural equivalence was initially seen as a way to formalize a relational view of social roles. Theorists like Nadel (1957) viewed societies as a system of positions or roles such that individuals occupying these positions had characteristic relations with other persons in their capacity as incumbents of their own positions. In support of this view, Walker et al. (1997) showed that organizations tend to remain structurally equivalent over time, even as they formed and dissolved ties with third-parties.

Over time, however, there was dissatisfaction with structural equivalence as an embodiment of the social role concept (Sailer, 1978; Borgatti and Everett, 1992). It was noticed that while structural equivalence correctly identifies a mother and father as both playing the parent role to their children (and grandparents to their children's children, and so on), it does not identify two mothers of different children as playing the same role because, by definition, structural equivalence demands having the same type of tie with the *same* others. Eventually, the notion of regular equivalence, first defined in print by White and Reitz (1983), came to supplant structural equivalence as a structural definition of social role.

However, structural equivalence remains an important concept as an explanatory factor in the study of social

homogeneity (Erickson, 1988). As noted above, a number of mechanisms can be proposed for why structurally equivalent nodes might be similar in behavior, attitudes, knowledge, or, in the case of collective actors such as firms, internal structures. One mechanism is third-party contagion since structurally equivalent nodes have third-parties in common. Another mechanism is based on the indistinguishableness property of structural equivalence. For any structural process (i.e., one whose course depends entirely on the structure of the network and not the characteristics of the nodes), we expect structurally equivalent nodes to have similar outcomes. For example, if the process involves random walks through the network, the number of times something lands on each node should have the same expected value for structurally equivalent nodes.

Another mechanism is suggested by social comparison theory (Festinger, 1954), which argues that individuals are especially influenced by those they see as similar. Hence, assuming structurally equivalent nodes are aware of their equivalence and are aware of each other's attitudes, they should tend to imitate each other. This perspective was used by Burt (1987) to explain why physicians were more likely to adopt a new drug which their structural equivalents adopted.

Since then, there has been considerable debate in the sociological literature about whether structural equivalence or dyadic cohesion (e.g., being adjacent) is the better predictor of similarity on outcomes like behavior and attitude (Mizruchi, 1993). At a theoretical level, it makes sense to distinguish between the similar opportunities and constraints afforded by a common structural position and the diffusion of something flowing from node to adjacent node. Unfortunately, structural equivalence was not the best way to examine this distinction empirically because structural equivalence embodies both poles of the debate (Borgatti and Everett, 1992).

Empirical Applications

Measures of structural equivalence have been used in a number of empirical studies in the social sciences, particularly in sociology and organizational science. The list of phenomena that have been studied by taking a structural equivalence approach includes the evolution of interfirm networks (Madhavan et al., 1998; Stuart, 1998; Walker et al., 1997), the diffusion of innovations (Burt, 1987), the effects of technological change (Barley, 1990), patterns of political affiliation (Padgett and Ansell, 1993), and attitude and belief similarity among individuals (Burkhardt, 1994; Erickson, 1988; Galaskiewicz and Burt, 1991). Below, we review four highly cited empirical studies that have used measures of structural equivalence as a major component of their methodology. This review is not meant to be an exhaustive review of empirical studies of structural equivalence, but rather to provide a summary overview of how structural equivalence has been applied empirically. We begin with studies at a macrolevel of analysis and then progress to studies at a more microlevel.

Structural equivalence has been widely used in studies of interorganizational relationships in the management field. One example of such a study is that of Walker et al. (1997), who used structural equivalence as a measure of social capital in their study of network formation in the biotechnology industry. Walker et al. gathered data on the partnership ties of biotech start-up firms between the years 1984 and 1988. They then ran CONCOR analyses, coupled with the CALCOPT algorithm, in order to group firms into structurally equivalent blocks. These researchers sought to determine whether the start-up firms would have a tendency to form new ties within their structural blocks in order to capitalize on the social capital benefits of network closure (Coleman, 1990), or form ties outside of their groupings in order to capitalize on network brokerage opportunities (Burt, 1992). This study found that, over time, the biotech start-ups tended to form new ties with partners with whom they were structurally equivalent, thus providing evidence that the firms in this industry tend to reproduce network structure. Walker et al. also examined structural equivalence among the start-ups at the dyadic level and found that if two firms were structurally equivalent at the beginning of the time period of their study, then those two firms were likely to remain structurally equivalent over time. Conversely, start-ups that were not structurally equivalent where highly likely to remain so, thus providing evidence for the relative rigidity of this particular network.

A second example of a study conducted at the firm level of analysis is that of Robins and Wiersema (1995), who conducted a study in which they sought to examine the business portfolios of a sample of large US manufacturing firms having multiple businesses. The aim of the study was to determine whether firms that maintained portfolios of relatively similar businesses enjoyed higher financial performance than those firms that maintained portfolios of differentiated businesses. Robins and Wiersema used structural equivalence analyses to model the similarity of businesses in each firm's portfolio. In order to do this, they first calculated the structural equivalence of a set of defined industry categories by using data on technology flows (e.g., patent transfers). In other words, one industry category was considered to be structurally equivalent to another industry category to the extent that they had technology flow ties to the same partners. From these industry-level structural equivalence data, they then created an index for each multibusiness firm in their study that captured the extent of the relatedness of the businesses in its portfolio. A firm was considered to have a highly related portfolio to the extent that each of its businesses were in industry categories that were structurally equivalent with one another. Robins and Wiersema found that the firms with highly related portfolios financially outperformed firms with lower levels of relatedness.

Moving to a lower level of analysis, Padgett and Ansell (1993) analyzed the intermarriage and economic relationships (e.g., trading ties, joint ownerships or partnerships, real estate ties) among families of 15th-century Florence in order to gain insight into the dynamics that led up to the political dominance of the Medicean political party over the reigning oligarch regime in Renaissance Florence. Padgett and Ansell constructed the social networks that existed among 92 elite Florentine families from 1427 to 1434 based on detailed historical accounts of this time. They then computed the structural equivalence of these families and partitioned the network into two structurally equivalent blocks: one affiliated with the Medici family and one affiliated with the oligarch regime. Padgett and Ansell find that their blockmodels are

highly predictive of the political affiliations of each of the families. Padgett and Ansell argue that the social structure uncovered in their analysis is an important factor in understanding how the Medici family was positioned to gain long-lasting political dominance in 1434.

At the individual level of analysis, Burt (1987) examined the effect played by structural equivalence in the diffusion of innovations by reanalyzing the data from Coleman et al.'s (1966) classic study, *Medical Innovation*. In the original study, Coleman et al. (1966) examined the diffusion of the newly released drug tetracycline among a sample of 125 physicians in the midwestern United States in the mid-1950s. Through their analysis of the social networks of these physicians, they concluded that the adoption of the new drug among the physicians was due to informal professional discussions whereby adopter physicians influenced nonadopters through direct contact (i.e., cohesion). Burt (1987) sought to reanalyze these data in order to more thoroughly examine the social structural conditions by which the apparent social contagion effects occurred by incorporating structural equivalence into the analysis. He compared the adequacy of models in which cohesion and structural equivalence were used to predict the adoption behavior of the physicians. Burt concludes that the social contagion effect seen in Coleman et al.'s data is due to structural equivalence, with cohesion having virtually no effect. These results suggest that the physicians in the Coleman et al. study were influenced not by those with whom they had direct interactions, but rather by the adoption behaviors of those with whom they were structurally equivalent. Thus, the results of this reanalysis suggest that one is more likely to feel social pressure to adopt innovations from those who are structurally equivalence as opposed to those to whom one directly interacts with. This study, and others like it (e.g., Burt and Dorian, 1982), provided counterevidence to the long-held belief that the effects of social contagion are explained by cohesion as opposed to structural equivalence. It was one of the first empirical articles to do so, and it gave rise to the debate about whether cohesion or structural equivalence was the primary mechanism through which social contagion occurs (for more on this debate, see Borgatti and Foster, 2003). This study, along with others (e.g., Galaskiewicz and Burt, 1991), provided support for the argument that structural equivalence is of primary importance in social contagion processes.

See also: Bayesian Graphical Models and Networks; Clustering and Cohesion in Networks: Concepts and Measures; Network Centrality, Measures of; Social Networks, Algebra of; Social Networks; Statistical Clustering.

Bibliography

Barley, S.R., 1990. The alignment of technology and structure through roles and networks. Administrative Science Quarterly 35, 61–103.

Batagelj, V., Ferligoj, A., Doreian, P., 1992. Direct and indirect methods for structural equivalence. Social Networks 14, 63–90.

Borgatti, S.P., Everett, M.G., 1992. Notions of position in social network analysis. Sociological Methodology 22, 1–35.

Borgatti, S.P., Foster, P.C., 2003. The network paradigm in organizational research: a review and typology. Journal of Management 29, 991–1013.

Breiger, R., Boorman, S., Arabie, P., 1975. An algorithm for clustering relational data with applications to social network analysis and comparison with multi scaling. Journal of Mathematical Psychology 12, 328–383.

Burkhardt, M.E., 1994. Social interaction effects following a technological change: a longitudinal investigation. Academy of Management Journal 37, 869–898.

Burt, R.S., 1976. Positions in networks. Social Forces 55, 93–122.

Burt, R.S., 1987. Social contagion and innovation: cohesion versus structural equivalence. American Journal of Sociology 92, 1287–1335.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R.S., Dorian, P., 1982. Testing a structural model of perception: conformity and deviance with respect to journal norms in elite sociological methodology. Quality and Quantity 16, 109–150.

Coleman, J.S., 1990. Social capital in the creation of human capital. American Journal of Sociology 94, 95–120.

Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation. Bobbs-Merrill, New York.

Cronbach, L.J., Gleser, G., 1953. Assessing similarity between profiles. The Psycho Bulletin 50, 456–473.

Erickson, B., 1988. The relational basis of attitudes. In: Wellman, B., Berkowitz, S.D. (Eds.), Social Structures: A Social Network Approach. Cambridge University Press, Cambridge, pp. 99–121.

Everett, M.G., Borgatti, S.P., 1994. Regular equivalence: general theory. Journal of Mathematical Sociology 19 (1), 29–52.

Festinger, L., 1954. A theory of social comparison processes. Human Relations 7 (2), 117–140.

Galaskiewicz, J., Burt, R.S., 1991. Interorganizational contagion in corporate philanthropy. Administrative Science Quarterly 36, 88–105.

Glover, F., 1989. Tabu search – Part I. ORSA Journal on Computing 1, 190–206.

Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 2, 241–254.

Lorrain, F., White, H.C., 1971. Structural equi of individuals in social networks. Journal of Math Sociology 1, 49–80.

Madhavan, R., Koka, B.R., Prescott, J.E., 1998. Networks in transition: how industry events (re)shape interfirm relationships. Strategic Management Journal 19, 439–459.

Mizruchi, M., 1993. Cohesion, equivalence and similarity of behavior: a theoretical and empirical assessment. Social Networks 15, 275–307.

Nadel, S.F., 1957. The Theory of Social Structure. Cohen & West, London.

Padgett, J.F., Ansell, C.K., 1993. Robust action and the rise of the Medici: 1400-1434. American Journal of Sociology 98, 1259–1319.

Panning, W., 1982. Fitting blockmodels to data. Social Networks 4, 81–101.

Robins, J., Wiersema, M.F., 1995. A resource-based approach to the multi-business firm: empirical analysis of portfolio interrelationships and corporate financial performance. Strategic Management Journal 16, 277–299.

Sailer, L.D., 1978. Structural equivalence: meaning and definition, computation and application. Social Networks 1, 73–90.

Stuart, T.E., 1998. Network positions and propensities to collaborate: an investigation of strategic alliance formation in a high-technology industry. Administrative Science Quarterly 43, 668–698.

Walker, G., Kogut, B., Shan, W., 1997. Social capital, structural holes, and the formation of an industry network. Organization Science 8, 109–125.

White, D.R., Reitz, K.P., 1983. Graph and semigroup homomorphisms on networks and relations. Social Networks 5, 193–234.

Structuralism

Mauro W Barbosa de Almeida, Departamento de Antropologia, Universidade Estadual de Campinas, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Structuralism is a mode of knowledge of nature and human life that is interested in relationships rather than individual objects or, alternatively, where objects are defined by the set of relationships of which they are part and not by the qualities possessed by them taken in isolation. Claude Lévi-Strauss was the spokesperson for structuralism in Anthropology, incorporating the work of many authors along the twentieth century. Three meanings of 'structuralism' will be distinguished, corresponding to different timescales: structuralism as a French intellectual movement of the 1960s, structuralism as a wider epistemological attitude and Lévi-Strauss' structuralism which is a link between the two.

In a wider sense, structuralism is a mode of knowledge of nature and human life that is interested in relationships rather than individual objects or, alternatively, where objects are defined by the set of relationships of which they are part and not by the qualities possessed by them taken in isolation. Claude Lévi-Strauss was the spokesperson for a structuralist anthropology, but his structuralism incorporated the methods and ideas of predecessors and contemporaries such as Lewis Morgan, for whom the object of the investigation were "systems of relationships" and the "structure of [such] systems" (Trautmann, 2008), A.R. Radcliffe-Brown, who focused on the structure of social relationships instead of individual customs, and E.E. Evans-Pritchard, who dealt with structural time rather than lived time, and the structural forms of clans instead of individual clans. Lévi-Strauss recognized these influences, as well as his debt to Marcel Mauss, the Leiden school and Marcel Granet for the structural analysis of kinship (Hénaff, 1998), and to Georges Dumézil regarding the analysis of myths. In the 1940s, the concept of structure was widespread in social anthropology, which is clear in the works by L. Murdock and C. Lévi-Strauss published in 1949 with the word 'structure' in their titles (Lévi-Strauss, 1967(1949)). However, in contrast with other anthropologists and sociologists who shared structural approaches to human phenomena, Lévi-Strauss explicitly attempted to formulate a structural method and demonstrate its performance through structural analyses conducted in different domains. We will distinguish here three meanings of 'structuralism,' corresponding to different temporal scales: structuralism as a French intellectual movement of the 1960s, structuralism as a wider epistemological attitude who took shape in early nineteenth century, and Lévi-Strauss' structuralism which is a link between the two.

French Structuralism

In the first sense, structuralism refers to an intellectual movement that reached its height in the 1960s in France, with no attachment to any organized group or specific institution. Instead, structuralism was a label for theses associated with R. Barthes, L. Althusser, M. Foucault, and J. Lacan in books following the publication of two seminal books by Claude Lévi-Strauss. These were *The Savage Mind* (1962) and *The Raw and the Cooked* (1964): both were chiefly responsible for placing the movement at the forefront of intellectual debate. This disparate group of *maîtres à penser* can be seen as representative of the French version of the linguistic turn in the Humanities. Faithful to its inspiration in F. De Saussure (rather than in C.S. Peirce or L. Wittgenstein, as in the noncontinental versions), French structuralism focused on 'semiology' as a science that would study 'the life of signs at the heart of social life' in areas such as literature (R. Barthes), philosophy (M. Foucault and L. Althusser), psychoanalysis (J. Lacan), and anthropology (C. Lévi-Strauss). French structuralism was also, in part, a reaction to Jean-Paul Sartre's 'philosophy of the subject,' but it also represented a stance against the empiricism and antimentalism of Anglo-Saxon philosophy. Methodologically, French structuralism emphasized the *system* of signs (synchrony) instead of their history (diachrony), and the *social* character of such systems (*langue*) instead of individual performance (*parole*) (Saussure, 1967(1916), pp. 30–33; Lévi-Strauss, 1950, p. xix), although this characterization has been questioned recently, through the publication of Saussure's original writings (Maniglier, 2006). French structuralism also insisted on the connection between meaning and form or semantics and syntax, again according to Saussure's doctrine of the inseparability of *signifiant* and *signifié* thus distinguishing itself from *formalism* (Lévi-Strauss, 1973, p. 139), and often described structure as systems of binary semantic contrasts, inspired in phonology. Lévi-Strauss distinguished his version 'structuralism' from the 'formalism' of Vladimir Propp's transformational model of folk narratives, which he rejected, as well as F Lounsbury's structural approach, on the grounds of allegedly dissociating 'form' from 'content,' that is to say, dealing in the first case with folktales as sequences of actions and objects without the 'bundles of semantic oppositions' underlying them, and, in the second case, analyzing 'kinship terminology' without 'marriage rules' they are supposed to express (Lévi-Strauss, 1973, p. 139; 1967, p. xxx). N. Chomsky's program of generative linguistics is similarly given short thrift (Lévi-Strauss and Eribon, 1988, p. 160). Philosophically, French structuralists rejected the concept of an autonomous subject endowed with the attributes of freedom to act and to seek knowledge (as in Sartre, 1946), affirming the priority of *structures* over *subjects* (see Caws, 2009 for a reassessment of Sartre as a structuralist). This "destitution of the subject" (Balibar, 2005)

is illustrated by Louis Althusser's formula, "Ideology interpellates individuals as subjects," which condenses the idea that individuals are *made into subjects* by 'State apparatuses' such as family, school, and church. The 'destitution of the subject' as a structuralist trope appears variously as the 'death of the author,' 'structural causality,' 'man is a recent invention,' and in Lévi-Strauss' dictum that "myths think about themselves among themselves … independent of any subject" (Lévi-Strauss, 1964, pp. 19–20). French structuralism proposed to substitute 'unconscious structures' for the 'subject' although the concept of structures that are innate to the human mind (or '*sprit humain*'), or 'immanent' as opposed to self-conscious (Lévi-Strauss, 1950, p. xxxi), does not convey precisely what they consist of or where they are to be located.

Although the rejection of the subject and the refusal of history have been pointed to as diagnostic traits of French structuralism, there are historically oriented versions of structuralism, sometimes allowing for the creative action of subjects. Hellenist Jean-Pierre Vernant, profusely praised by Lévi-Strauss, is a major example (Vernant, 1986). Jean Piaget's 'genetic structuralism' characterizes cognitive structures as relatively stable stages in the development of cognitive abilities which are generated through the interaction between the organism and the environment (Piaget, 1968); Pierre Bourdieu rehabilitated the role of the acting subject, retaining the notion of a structure as schemes for action ("principles of the generation and structuring of practices and representations") embodied in a subject's *habitus*; M Sahlins argued in favor of a view of structure phrased in the Boasian sense of "the symbolic relations of cultural order," as both constraining human practice and being transformed by it (Sahlins, 1985: vii). T. Turner's 'recursively hierarchical' systems of social production and reproduction were proposed as a synthesis of Marxist and epigenetic views of structure, influenced by Piaget's and by French marxism (Turner, 1973). Clearly, there is more to structuralism than is contained in the so-called French structuralism.

Structuralism in Wider Sense

In the wider sense indicated above, structuralism refers to the attitude toward knowledge that has roots within a philosophical and scientific epistemological revolution that took place in the late eighteenth century, represented by G. Leibniz and A. Dürer (Petitot, 2009). In the transition from the eighteenth century to the nineteenth century, W. Goethe formulated the program of a science of 'metamorphoses' or 'transformations,' the goal of which would be nothing short of depicting the underlying formal unity in wide domains of living beings and human works, a seminal idea which was inspired by the Kantian notion of a 'transcendental schema,' which purported to explain how abstract concepts could subsume a virtually infinite variety of particular empirical perceptions (*phenomena*). Goethe's program was invoked by biologist D. Thompson (1992, p. 1032) who argued that similar life forms, such as fish species, could be thought of as related to each other by continuous transformations (Thompson, 1992, pp. 5–7, 1094), although distinct phyla, such as fishes and worms, cannot. Dürer, Goethe, and D'Arcy Thompson, as well as Kant, were thus understandably acknowledged by C. Lévi-Strauss as forerunners of structuralism as a science of 'transformations.' In the 1960s, the Field Medal mathematician R. Thom proposed that his theory of 'morphogenesis' was the appropriate tool for depicting discontinuous changes of form in biological and cultural systems, an idea which J. Petitot applied to Lévi-Strauss's analysis of myths (Thom, 1975(1972); Scubla, 1998; Petitot, 2009). Along another line of thought, Goethe's goal to find a key with which "new plants might be devised *ad itum*" would be realized in Chomsky's concept of a generative grammar (1966)."

Mathematics, which had long been defined as the science of numbers and geometric shapes, came to be seen by the end of the nineteenth century as the study of conceptual-practical structures, the paradigm of which are transformation groups. According to this changed view, groups of transformations embody the notion of what is common in a family of objects, namely, the structural properties which are conserved by the transformations which constitute the group (Weyl, 2012, pp. 170–171). The mathematical concept of a group of transformations was extended to domains such as crystallography, physics, biology, and psychology (Piaget, 1968). It was however with a grain of surprise that it was realized that 'elementary structures of kinship' could be represented as permutation groups, as did André Weil in his invited appendix to Lévi-Strauss treatise (1967(1949)), which was the seed for an extensive body of literature (e.g., White, 1963). Lévi-Strauss came to see this use of modern mathematics to apprehend rules governing marriage rules and kinship terminologies usage as evidence of genuine mathematical ability among illiterate people. It should be mentioned that André Weil's receptivity to Lévi-Strauss's ideas was no accident, as he himself was a forefront member of the group of French mathematicians that, since the 1930s, devoted itself to the task of reconstructing mathematics as a theory of structures under the collective *nom de plume* of Nicolas Bourbaki (1962). According to this research program, groups are the simplest algebraic structure, expressing symmetries and inversions; lattices and trees are examples of structures of order, present in classifications and genealogies, while topological structures involve the notion of continuity and conservation of form.

The Structuralism of Lévi-Strauss

Lévi-Strauss's scientific and literary output ranges from 1945 to 1991. Over the course of nearly half a century, Lévi-Strauss produced three highly influential works, each of which dealt with a distinct domain of ethnographical phenomena: kinship (1945, 1949), classifications (1962a,b), and mythologies (1964, 1966, 1969, 1971; 1979, 1985, 1991). His study of kinship contains a sociological theory based on the concept of exchange, while his theory of mythical throught hinged on the notion of transformation. Coming between these works, *The Savage Mind* was a study of classification, being a philosophical and methodological interlude that launched Lévi-Straussian structuralism into a wider intellectual arena not only as a social theory but also as a philosophy of history and even as a epistemic paradigm. It could be said that the 'mother structures' in the sense the previous section run through the theory

of marriage of *The Elementary Structures of Kinship* (permutation groups), the systems of classification of *The Savage Mind* (taxonomies and conceptual lattices), and the chains of mythical transformations studied in *Mythologiques* (mythical transformations, topological forms such as 'Klein bottles' and Möbius strips). However, there is more to it than that, for according to Lévi-Strauss' insistent reminders, in the course of history events inevitably lead to the loss of structure – a melancholy view of history, which he expressed by borrowing from physics the metaphor of entropy, the assertion of ineluctable and irreversible change from ordered states toward disordered states of any system (Lévi-Strauss, 1967(1949). This view is reasserted in his latest and posthumous writings (Lévi-Strauss, 2013). Lévi-Strauss' analogical borrowing of scientific concepts, however, should be nuanced by the recent appreciation of his importance as a literary artist, influenced by Breton, Mallarmé, and Wagner (Mehlman, 2006) and as a original philosopher, rather than as a scientist strictu sensu (Balibar, 2005; Debaenne, 2008; Viveiros de Castro, 2008; Wyseman, 2009).

Kinship and Group Structures

An early example of Lévi-Strauss's structural approach, algebraic in spirit although phrased with the vocabulary of phonological linguistic, is his article on the 'atom of kinship,' originally published in 1945 (Lévi-Strauss, 1958). The 'atom of kinship' refers to a set of four interconnected relationships, namely husband/wife, brother/sister, father/son, and nephew/maternal uncle, each one associated with an 'attitude,' which Lévi-Strauss characterizes with a '+' or '−' mark. The structural approach consist here in seeking a law behind the patterns in the sets of 'attitudes' found in particular societies, rather than explaining each attitude by social, historical, or psychological causes. This is the proposed law: the husband/wife relationship and the sister/brother relationship have opposite signs, as do the father/son and the avuncular relationships. There are four models which satisfy this requirement: (+/−, +/−), (+/−, −/+), (−/+, +/−), and (−/+, −/+). This model provides an example of group structure known as the Klein group. For let *S* transform (+/−, +/−) into (+/−, −/+) and let *T* transform (+/−, +/−) into (−/+, +/−); *ST* will then be the transformation which takes (+/−, +/−) into (−/+, −/+) and will have the same effect as *TS*. We call the transformation that leaves (+/−, +/−) unchanged the 'identity' *I*. The set of four transformations *I*, *S*, *T*, and *ST* is closed (any composition of transformations is again a transformation) and for each transformation there is an inverse transformation. Lévi-Strauss' own language, the 'family of models' for attitudes within the kinship atom is represented as a group of transformations (Lévi-Strauss, 1958) which preserve as invariant feature the contrast between affinity and consanguinity (the husband/wife and the brother/sister relations have opposite signs, as the maternal uncle/sister's son and the father/son relations do). The structure of the atom of kinship illustrates all the features that Lévi-Strauss listed as being required by 'structural analysis,' and which may be summarized as the specification of the transformations which generate a 'family of models' having in common the same 'pairs of oppositions' (Lévi-Strauss, 1958). It also illustrates the importance of the distinction between empirical data about social relationships and models built up on the basis of those relationships from which structural laws are obtained. Underlying the 'atom of kinship' model there is a social theory according to which kinship is not fundamentally a matter of consanguinity ties of siblingship and filiation, but necessarily includes alliance, either as affinity between in-laws in the same generation, as between brothers-in-law, or across generations, as between nephew and maternal uncle. The contrasted attitudes displayed by the 'atom of kinship' are explained as the necessary opposition between consanguinity and affinity in human social life, which ultimately express the contrast between nature and culture. This thesis is unfolded in Lévi-Strauss's *The Elementary Structures of Kinship*. In Lévi-Strauss's analysis, rules of marriage that prescribe marriage among cross-cousins but forbids it among parallel cousins, as well as kinship terminologies which equate cross-cousins with spouses, are viewed as cultural expressions of an unconscious principle that requires the obligation to give sisters away in order to receive wives from others. In other words, rules and vocabularies are the conscious expressions of unconscious structures of reciprocity. According to Lévi-Strauss, 'elementary structures of kinship' exist in societies where people belong to descent groups, and where the choice of wife is governed by obligatory rules that determine the descent group in which a man can seek a wife. On a more specialized level, Lévi-Strauss classified these 'elementary structures' as two major types. In the first major type, or restricted exchange, women are exchanged reciprocally between paired descent groups, a global, or statistical, social process which the result, at the level of norms regulating individual action, of a rule of marriage according to which a man marries a woman classified as his mother's brother's and as his father's sister's (a bilateral cross-cousin). In the second major type, 'generalized exchange,' women exchange is unilateral, with two possibilities: matrilateral cross-cousin marriage (a man takes for a wife a woman classified as his mother mother's brother's daughter), or patrilateral cross-cousin marriage (a man takes for a wife his father's sister's daughter). Matrilateral cross-cousin marriage leads to cycles of exchange which require several generations to close, in contrast with the short cycles resulting from the patrilateral rule. Such consequences can be be taken as the global results of the strict working of the rules; they may also be seen as statistical effects of the existence of rules that affect the micro-level of individual behavior (Lévi-Strauss, 1967(1949)). Characteristically, such consequences are deduced under the assumption that rules are applied strictly in accordance to a model.

As in the case of the explanation of attitudes between uncles and nephews which motivated the theory of the 'atom of kinship', the explanation of the opposition between cousins and sisters leads theory of social life in which society is depicted as a fabric of exchanges of things and persons, either immediate or along longer circuits. This theory can be seen as an application to the domain of kinship of Marcel Mauss' theory of gift exchange as a basic institution of social integration, which was already present in the work of the Leyden school and Marcel Granet (Héran, 2009, p. 395), and may be compared to economic models of social reproduction through exchange of goods (Rosman and Rubel, 2009).

The interpretation of marriage as exchanges among descent groups leads to the mathematical analysis of elementary

structures a dance of chairs in which lineages or clans interchange the roles of donors and receivers of wives, a clue taken by mathematicians since A. Weil as a sociological version of permutation groups. It should be noted that Lévi-Strauss does not assumes that the real social world must obey the rules in a strict sense; as he made it clear in later writings, he only assumes that social rules, although not followed to the letter, inflect statistical phenomena towards a global picture described by the idealized structure. But the theory of elementary structures is not only formalism, since it also contains the idea that human kinship is based on the insoluble tension between the incestuous desire for autonomy and the cultural obligation of exchange. This tension is a universal human trait for Lévi-Strauss, expressed in the desire to reject the rule of culture and return to an imagined state of nature (Lévi-Strauss, 1967(1949)). This view of culture, where the influence of Freud is clearly seen, is an example of the pessimism that runs through all of Lévi-Strauss's work like a melancholy meta-anthropology: humanity does not have an internal trend toward betterment, but is a result of a nonoriented process throughout which certain structures crystallize in the shorter or longer term, only to dissolve inevitably under the influence of internal contradictions and demographic and environmental change (Lévi-Strauss, 1955). Lévi-Strauss's kinship theory was maybe the part of his opus that fared worst after it formulated in the 1940s, in a empirical sense. His factual basis was contested early on by E. Leach (1961), while Lévi-Strauss himself noted in 1967 that 'elementary structures' as social systems for the regulation of marriages, implying unilineal descent according to his definition, were relatively rare in the world where cognatic systems are more frequent (Lévi-Strauss, 1967). The notion of cross-cousin marriage as the basic form of alliance was also subject to criticism for not accounting for cases as 'Arab marriage,' a prescribed union between paternal parallel cousins (Barry, 2000; Coelho de Souza, 2009). Recent research movements focused on modeling real social relationships through graph theory can be seen as a mode of structural analysis by means of rigorous analysis and ethnography (Hamberger et al., 2011).

Savage Thought: Structures of Order

The broader impact of Lévi-Strauss's work was not a consequence of his 1949 study of kinship, although it was at once reviewed by Simone de Beauvoir (who praised it, pace later feminist critiques) and received a detailed critique by Jean-Paul Sartre, as well as awakening the sympathies of philosopher Merleau-Ponty. Lévi-Strauss's ideas broke through national and academic borders through *La Pensée Sauvage*, later translated as *The Savage Mind* – a multifaceted book containing general ideas about science, art, and the science of nonliterate peoples, a theory of totemism and classification systems, and a polemical, nonteleological view of history. Whereas in the decades between the 1940s and 1960s Lévi-Strauss introduced the structural method as an extension of the scientific use of models in the humanities, *The Savage Mind* emphasizes the idea that 'savage' or untamed thought is another form of scientific thought, which amounts to say, is capable of applying logical reasoning to sense perception. He pictured this idea through the famous analogy of the *bricoleur*, a tinker who builds structured object out of fragments found haphazardly and with no previous master plan. Such objects, be they myths or kinship terminologies, body paintings or musical pieces, technical artifacts or domesticated plants, are according to this view both historical objects and mental constructions, and therefore the same time contingent and subject to constraints. Lévi-Strauss extols the inventions and techniques of the 'Neolithic revolution' as lasting results of the workings of such '*science du concret*' in the past, at the same time reminding us that this creative mind is currently active in traditional knowledge and art. *The Savage Mind* gave a detailed account of how this science of illiterate people was active in conceiving complex classification systems encompassing plants, animals, people, and cosmic phenomena. Classification systems are by no means mere sets of 'binary oppositions,' as in the 'atom of kinship' model, or permutation groups as in the 'elementary structures' form. They involve *order structures*, described by mathematicians as 'conceptual lattices' or as 'formal ontologies,' which attempt to show relations of inclusion and implication between categories (Lévi-Strauss, 1962, p. 201; Parrochia and Neuville, 2013, p. 30). The investigation of 'savage thought' as wide-reaching modes of classifying the human and natural worlds was pursued by P. Descola (2005).

It has been said that *The Savage Mind* contains a cybernetic theory of social stability essential for structuralism: according to this interpretation, the classificatory 'structures' there described are regulatory devices to ensure social stability (Bloch, 1996, p. 531). This view is apparently anchored in Lévi-Strauss's contrast between 'cold' and 'hot' societies, the first seeking to 'annihilate time' and retain structures, the latter valuing change and accumulation; however, as a theory of history, this view is contradicted by Lévi-Strauss insistence on the idea that every society and every symbolic system is subject to 'entropy,' to inevitable degradation of the structure over time (Lévi-Strauss, 1955). One example is myths, which not only transform into each other, but may lose their structure and eventually die at the end of long transformational trajectories and in contact with other narrative systems.

Myths: Transformations and Continuity

Lévi-Strauss devoted his longest study to the structural analysis of the myths of the Americas, both in the *Mythologiques* (published in four volumes from 1964 to 1971) (Lévi-Strauss, 2008; comprising three independent little volumes originally published from 1979 to 1991). These analyses give rise again to a new way of employing structural analysis. Instead of finite models having the structure of groups, one is now confronted with myths connected by an open set of transformations. Thus, close to a 1000 myths that extend from the Chaco region of South America to Canada are viewed as part of chains of transformations, both historical-geographic and logical-metaphysical. Furthermore, in this work, Lévi-Strauss takes to the extreme his tendency to attribute the same dignity to the construction of myths as he does to scientific and artistic works: mythical thinking is the work of a collective author, persistently tackling themes of essential interest for the Amerindian mind, such as the relationship between the cosmic, terrestrial, and subterranean planes of life, as well as the opposition between animals and humans, cultural and natural objects, the

continuous and the discrete, the finite and the infinite. One can found an almost didactic example of Lévi-Strauss's use of the 'transformation' concept in the opening chapters of *The Jealous Potter*, one of the "small mithologiques," where Lévi-Strauss provides a detailed commentary on a set of myths from the South American Jivaro Amerindians that suggests clues to "the place of jealousy in myths on the origins of pottery," connecting "an art of civilization, a moral sentiment and a bird" (Lévi-Strauss, 1985, p. 32; Gow, 2009). This assertion is argued by Lévi-Strauss as follows. He firsts shows that myths recorded for the Jivaro area assert a connection between the art of pottery, jealousy and a bird, the Night-Hawk. These 'empirically' recorded connections (in myth narratives) suggest that there should be a bird linking pottery and conjugality. Or, as Lévi-Strauss reconstructs the mythical reasoning at play: as a *jealous* bird (Night-Hawk) is related to the *pottery-making* ability of Woman, the Woman's *jealousy* should be related to a *pottery-making bird*. The problem, says Lévi-Strauss, is that there is no such bird in the Achuar área – namely, a bird with the required in the set of myths available in the Achuar area. This situation 'puts certain fundamental principles into play in the structural analysis of myth.' And this is the fundamental principle: given a logically incomplete set of myths, we must seek a missing piece in *another* sets of myths, perhaps in nearby geographical settings, perhaps in another semantic context, in which the mythical reasoning comes to a closure. Let us illustrate this mechanism: Lévi-Strauss finds a conclusion of the logical sequence initiated in the Peruvian rain forest in myths recorded in the Chaco plains, where there is indeed a pottery-making bird, the *Oven-bird*. While the Night-Hawk is a jealous and lonely bird that cries on moonlit nights remembering his lost cosmic partner, the oven-bird is a happily married bird that lives in a house he built from clay together with his wife. In Lévi-Strauss's terms, the *Oven-bird* is a 'transformation' of the *Night-Hawk* that preserves the logical structure underlying the explanation of the origins of pottery, while connecting two different ecological setting, that of the Peruvian pre-andean rain forest and that of the Chaco plains. In sum, the analysis of myths must combine 'empirical deductions,' that is, those which are based directly on the mythical material, and 'transcendental deductions' that make it possible to "close a cycle of transformations" (Lévi-Strauss, 1985, p. 80).

Myths transform themselves in history (Gow, 2001) and are geographically connected by a process of dissemination that is, however, obliged to follow logical procedures (Carneiro da Cunha, 1973). It has been argued that the mythical transformations invoked by Lévi-Strauss are closer to the grammar of dreams and to surrealistic aesthetics than to formal logic (Viveiros de Castro, 2008). Instead of the turalism of binary pairs that form closed and finite groups, we find here, according to a post-modern interpretation of Lévi-Strauss' views pioneered by Viveiros de Castro, a situation where "the subject is always a particular state of a system of transformations whose boundaries are radically contingent," and in which the concept of structure "does not favor any desire for closure, completion or compactness" (Viveiros de Castro, 2008). These characteristics are particularly supported by *Story of Lynx*, the last work that Lévi-Strauss devoted to the analysis of myths (Lévi-Strauss, 1991). In it, the author explains that 'binary pairs' are just a particular example of 'transformations' that include inversions, equivalencies, symmetries, and contradictions (pp. 249ff.); that transformations occur along multiple semantic axes that include time and space, the body and stars, sexuality and cuisine; finally, that 'transformation' can take place 'by degrees,' that 'unstable dualism' is the rule, that mythological systems are in a state of 'perpetual imbalance,' and that mythological transformations are therefore open (pp. 306–316). Furthermore, mythological transformations do not appear as the product of a mythical mind floating in space, but as the work of myth builders whose work must respect the historical boundaries imposed by the settlement of the Americas. Mechanical and symmetrical group structures can occur locally, but in the overall analysis, they give way to irreversible historical processes such as the colonization of the America.

The Legacy of Structuralism

Structuralism was the target of criticisms that demanded either a return to history to the detriment of 'synchrony,' a return to 'thick ethnography,' preferably to the pursuit of doces,' or a rehabilitation of the active and revolutionary subject from his alleged position of an 'effect' of structures (Turner, 1973, 2009). The influence of structuralism remained strong among specialists in lowland South America, where Lévi-Strauss had his original ethnographic experience. Vigorous research programs on the social structure of indigenous societies of Central Brazil were conducted in dialogue with his theses although in a critical vein (Maybury-Lewis, 1979). Carneiro da Cunha pointed to the role of mythical structures dealing with messianic movements in Amazonia (1973) and of the connection between structure and the *persona*; S. Hugh-Jones (1979b) and C. Hugh-Jones (1979a) have shown how a close ethnography of ritual and body practices could be framed in structuralism terms, while P. Gow demonstrated the possibility of combining a historical-biographical approach with the structural analysis of myths (Gow, 2001, 2009). E. Viveiros de Castro is a foremost example of Lévi-Strauss' influence on contemporary South American anthropological thinking, combined with a Deleuzian 'post-structural' twist (Viveiros de Castro, 1992), while P. Descola has developed a wide-ranging theory of sociocosmologies rooted in the structuralist heritage (Descola, 2005).

In a more general sense, what is left of structuralism in the twenty-first century? Perhaps the notion that the world of natural and human phenomena can either be scrutinized in depth toward the complexity of particular objects – whether a painting or a novel, an equation or an animal species, a year in history and a village in the country, a concept or a musical piece – or analyzed in pursuit of generality and abstraction, placing the individual thing so speak within a set of variations on a same theme. In this sense, both 'French structuralism' with its mentalist tradition and literary style and 'British structuralism' with its adherence to empirical sobriety are as many varieties of the encompassing nineteenth-century epistemological strive for generalization and model building which made its marks in sciences, humanities, and arts along the twentieth century, and was since then not so much refuted as assimilated to the point of becoming a set of truisms in domains such as social network

analysis, having been thus incorporated into common sense, structuralism may be as well be forgotten.

Translated by H. Sabrina Gledhill.

See also: Anthropology and History; Archaeology, Theory in; Cognitive Anthropology; Culture: Contemporary Views; Exchange in Anthropology; Kinship in Anthropology; Primitive Society; Psychological Anthropology; Time, Anthropology of.

Bibliography

Balibar, E., January 2005. Le structuralisme, une destitution du sujet? Revue de Métaphysique et de Morale (special issue).

Barbosa de Almeida, M.W., 1990. Symmetry and entropy. Mathematical metaphors in the work of Lévi-Strauss. Current Anthropology 31 (4), 367–385.

Barry, L.S. (Ed.), 2000. Question de parenté. L'Homme (special issue), pp. 154–155.

Bloch, M., 1996. Structuralism. In: Barnard, A., Spencer, J. (Eds.), Encyclopedia of Social and Cultural Anthropology. Routledge, London.

Bourbaki, N., 1962. L'architecture des mathématiques. In: Le Lionnais, F. (Ed.), Les Grands Courants de la Pensée Mathématique. Albert Blanchard, Paris, pp. 35–47.

Carneiro da Cunha M., 1973. Loqique du mythe e de l'action. Le mouvement messianique Canela de 1963. L'Homme 13 (4), 5–37.

Caws, P., 2009. Sartrean Structuralism? In: Wyseman 2009, pp. 293–317.

Coelho de Souza, M., 2009. The Future of the structural theory of kinship. In: Wiseman, B. (Ed.), The Cambridge Companion to Lévi-Strauss. Cambridge University Press, Cambridge, pp. 80–99.

Debaenne, V., 2008. Préface. In: Lévi-Strauss, C. (Ed.), Oeuvres. Gallimard, Paris, pp. ix–xlii.

Descola, P., 2005. Par-delà nature et culture. Éditions Gallimard, Paris.

Gow, P., 2001. An Amazonian Myth and Its History. Oxford University Press, Oxford.

Gow, P., 2009. Of the Story of Lynx: Lévi-Strauss and Alterity. In: Wyseman 2009, pp. 219–236.

Hamberger, K., Houseman, M., White, D.R., 2011. Kinship network analysis. In: Scott, J., Carrington, P.J. (Eds.), The SAGE Handbook of Social Network Analysis. SAGE Publications, London.

Hénaff, M., 1998. Lévi-Strauss and the Making of Structural Anthropology. The University of Minnesota Press, Minnesota.

Héran, F., 2009. Figures de la Parenté. Presses Universitaires de France, Paris.

Hugh-Jones, C., 1979a. From the Milk River. Cambridge University Press, Cambridge.

Hugh-Jones, S., 1979b. The Palm and the Pleiades. Initiation and Cosmology in Northwest Amazonia. Cambridge University Press, Cambridge.

Izard, M. (Ed.), 2004. Claude Lévi-Strauss. Éditions de L'Herne, Paris.

Kronenfeld, D., Decker, H.W., 1979. Structuralism. Annual Review of Anthropology 8, 503–541.

Leach, E.R., 1961. Rethinking Anthropology. The Athlone Press, Oxford.

Lévi-Strauss, C., 2008. Ouvres. Gallimard, Paris.

Lévi-Strauss, C., 1950. Introduction. In: Mauss, M. (Ed.), Sociologie et Anthropologie. Presses Universitaires de France, Paris.

Lévi-Strauss, C., 1955. Tristes Tropiques. Plon, Paris (Tristes Tropiques (J. Weightman, D. Weightman, Trans.). Penguin Books, 1976.).

Lévi-Strauss, C., 1958. Anthropologie Structurale. Plon, Paris (Structural Anthropology (C. Jacobson, B.G. Schoepf, Trans.). Basic Books, New York, 1963.).

Lévi-Strauss, C., 1962. La Pensée Sauvage. Plon, Paris (The Savage Mind, 1966).

Lévi-Strauss, C., 1964. Mythologiques I. Le Cru et le Cuit. Plon, Paris (The Raw and the Cooked, 1969).

Lévi-Strauss, C., 1967(1949). Les Structure Élementaires de la Parenté, second ed. Plon, Paris (first ed. 1949).

Lévi-Strauss, C., 1973. Anthropologie Structurale II. Plon, Paris (Structural Anthropology 2).

Lévi-Strauss, C., 1985. La Potière Jalouse. Plon, Paris.

Lévi-Strauss, C., 1991. L'Histoire de Lynx. Plon, Paris.

Lévi-Strauss, C., 2013. Nous Sommes Tous Cannibales. Seuil, Paris.

Lévi-Strauss, C., Eribon, M., 1988. De près et de loin. Odile Jacob, Paris.

Lévi-Strauss, C., 1971. Mythologiques IV. L'Homme Nu. Plon, Paris.

Maniglier, P., 2006. La vie énigmatique des signes, Saussure et la naissance du structuralisme. Non & Non. Léo Scheer, Paris.

Maybury-Lewis, D., 1979. Dialectical Societies: The Gê and Bororo of Central Brazil. Harvard University Press, Harvard.

Mehlman, J., 2006. Structuralism, Poetry, Music: Lévi-Strauss between Mallarmé and Wagner. In: Wyseman 2009, pp. 257–274.

Parrochia, D., Neuville, P., 2013. Towards a General Theory of Classifications. Birkhäuser, Heidelberg.

Petitot, J., 2009. Morphology and Structural Aesthetics: From Goethe to Lévi-Strauss. In: Wyseman 2009, pp. 275–295.

Piaget, J., 1968. Le Structuralisme. Presses Universitaires de France, Paris.

Rosman, A., Rubel, P.G., 2009. Structure and Exchange. In: Wyseman 2009, pp. 59–79.

Sahlins, M., 1981. Historical Metaphors. Structure in the Early History of the Sandwich Islands Kingdom. The University of Michigan Press, Ann Arbor.

Sahlins, M., 1985. Islands of History. The Chicago University Press, Chicago.

Sartre, J., 1946. L'existentialisme est un humanisme. Éditions Nagel, Paris. (Existentialism Is a Humanism (C. Macomber, Trans.). Yale University Press, New Haven, 2007).

Saussure, F., 1967(1916). In: Mauro, T. (Ed.), Cours de Linguistique Génerale. Payot, Paris.

Scubla, L., 1998. Lire Lévi-Strauss. Le Déploiement D'une Intuition. Éditions Odile Jacob, Paris.

Thom, R., 1975(1972). Structural Stability and Morphogenesis. An Outline of a General Theory of Models. W.A. Benjamin Reading.

Thompson, D.W., 1992. On Growth and Form (the complete revised edition). New York, Dover.

Trautmann, T.R., 2008. Lewis Henry Morgan and the Invention of Kinship, second ed. University of Nebraska Press, Lincoln and London.

Turner, T.S., 1973. Review of Piaget's genetic epistemology and le structuralisme. American Anthropologist 75 (2), 351–373.

Turner, T.S., 2009. The crisis of late structuralism. Perspectivism and animism: rethinking culture, nature, spirit, and bodylines. Tipití: Journal of the Society for Anthropology of Lowland South America 7 (1), 3–39.

Vernant, J.P., 1986. Le Tyran boiteux: d'Oedipe à Périandre. In: Vernant, J.P., Vidal-Nacquet, P. (Eds.), Oedipe et ses Mythes. Éditions La Découverte, Paris.

Viveiros de Castro, E., 1992. From the Enemy's Point of View. Humanity and Divinity in an Amazonian Society. The University of Chicago Press, Chicago.

Viveiros de Castro, E., 2008. Claude Lévi-Strauss, fundador do pós-estruturalismo. Lecture at the colloquium Lévi-Strauss: un siglo de reflexión. Museo Nacional de Antropología, Mexico.

Weyl, H., 2012. Levels of Infinity. Selected Writings on Mathematics and Philosophy. New York, Dover.

White, H.C., 1963. An Anatomy of Kinship. Prentice Hall, New Jersey.

Wyseman, B. (Ed.), 2009. The Cambridge Companion to Lévi-Strauss. Cambridge University Press, Cambridge.

Subaltern History

Partha Chatterjee, Columbia University, New York, NY, USA; and Centre for Studies in Social Sciences, Calcutta, India

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Subaltern Studies is a series of writings on modern South Asian history in which Antonio Gramsci's idea of the subaltern classes was used to study the autonomous struggles of peasants in colonial India. Even when peasants participated in movements led by nationalist elites, they retained their own objectives, methods, and ideologies. Since there were few written sources produced by peasant activists themselves, and since elite sources usually depicted the subaltern as passive and subservient, subalternist historians argued that the forms of autonomous subaltern consciousness were best revealed in the archives of peasant revolt. The idea of autonomous subaltern consciousness was criticized, following which the project reoriented itself to consider several other facets of subaltern life in which elite and subaltern domains were intertwined. This included topics such as religion, caste, gender, nationalism, modernity, etc. An important feature here is the argument that Western modernity, encountering resistance from subaltern practices in South Asia, were taking the shape of different and hybrid modernities.

Gramsci in India

In the form in which it is currently known, subaltern historiography is derived from the writings of a group of historians of modern South Asia whose work appeared between 1982 and 2005 in a series of 12 volumes entitled *Subaltern Studies* (Amin and Chakrabarty, 1996; Arnold and Hardiman, 1994; Bhadra et al., 1999; Chatterjee and Jeganathan, 2000; Chatterjee and Pandey, 1992; Guha, 1982–1989; Mayaram et al., 2005). The term 'subaltern' was borrowed by these historians from Antonio Gramsci, the Italian Marxist, whose writings in prison in the period 1929–35 had sketched a methodological outline for a 'history of the subaltern classes' (Gramsci, 1971). In these writings, Gramsci uses the word 'subaltern' (*subalterno* in the Italian) in at least two senses. In one, he used it as a code for the industrial proletariat. But against the thrust of orthodox Marxist thinking, he emphasized that in its rise to power, the bourgeoisie did not simply impose a domination through the coercive apparatus of the state but transformed the cultural and ideological institutions of civil society to construct a hegemony over society as a whole, even eliciting in the process the acquiescence of the subaltern classes. In Gramsci's analysis of capitalist society, the central place is occupied by questions such as the relation of state and civil society, the connections between the nation; the people; the bourgeoisie; and other ruling classes, the role of intellectuals in creating the social hegemony of the bourgeoisie, strategies for building a counterhegemonic alliance, etc.

In the second sense, Gramsci talked of the subaltern classes in precapitalist social formations. Here he was referring to the more general relationship of domination and subordination in class-divided societies. But specifically in the context of southern Italy, he wrote about the subordination of the peasantry. Gramsci was very critical of the negative and dismissive attitude of European Marxists toward the culture, beliefs, practices, and political potential of the peasantry. Positioning himself against this attitude, he wrote of the distinct characteristics of the religious beliefs and practices, the language and cultural products, the everyday lives and struggles of peasants, and of the need for revolutionary intellectuals to study and understand them. But he also highlighted the limits of peasant consciousness, which in comparison with the comprehensiveness, originality, and active historical dynamism of the ruling classes, was fragmented, passive, and dependent. Even at moments of resistance, peasant consciousness remained enveloped by the dominant ideologies of the ruling classes. These discussions by Gramsci were turned to productive use by South Asian historians writing in the 1980s.

Against Elitist Historiography

Ranajit Guha, who edited the first six volumes of *Subaltern Studies*, wrote in the introductory statement of the project: "The historiography of Indian nationalism has for a long time been dominated by elitism – colonialist elitism and bourgeois-nationalist elitism" (Guha, 1982). The objective of subaltern historiography was to oppose the two elitisms. The field of modern South Asian history was dominated in the 1970s by a debate between a group of historians principally located in Cambridge, UK, and another based mainly in Delhi, India. The former argued that Indian nationalism was a bid for power by a handful of Indian elites who used the traditional bonds of caste and communal ties to mobilize the masses against British rule. The latter spoke of how the material conditions of colonial exploitation created the ground for an alliance of the different classes in Indian society and how a nationalist leadership inspired and organized the masses to join the struggle for national freedom. Guha argued that both these views were elitist – the former representing a colonial elitism and the latter a nationalist elitism. Both assumed that nationalism was wholly a product of elite action. Neither history had any place for the independent political actions of the subaltern classes.

In setting their agenda against the two elitisms, the historians of *Subaltern Studies* focused on two main issues. One was the difference between the political objectives and methods of colonial and nationalist elites on the one hand and those of the subaltern classes on the other. The second was the autonomy of subaltern consciousness. Pursuing the first question, the historians of *Subaltern Studies* showed that the claim of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.62129-3

colonialist historians that the Indian masses had been, so to speak, duped into joining the anticolonial movement by Indian elites using primordial ties of kinship and clientelism was false. They also showed that it was untrue to say, as nationalist historians did, that the political consciousness of the subaltern classes was only awakened by the ideals and inspiration provided by nationalist leaders. It was indeed a fact that the subaltern classes had often entered the arena of nationalist politics. But it was also a fact that in many instances they had refused to join, despite the efforts of nationalist leaders, or had withdrawn after they had joined. In every case, the goals, strategies, and methods of subaltern politics were different from those of the elites. In other words, even within the domain of nationalist politics, the nationalism of the elites was different from that of the subaltern classes.

The second question followed from the first. If subaltern politics was indeed different from that of elite politics, what was the source of its autonomy? What were the principles of that politics? The answer that was suggested was that subaltern politics was shaped by the distinct structure of subaltern consciousness. That consciousness had evolved out of the experience of subordination – out of the struggle, despite the daily routine of servitude, exploitation, and deprivation, to preserve the collective identity of subaltern groups. Where was one to look for the evidence of this autonomous consciousness? It could not be found in the bulk of the archival material that historians conventionally use, because that material had been prepared and preserved by and for the dominant groups. For the most part, these documents only show the subaltern as subservient. It is only at moments of rebellion that the subaltern appears as the bearer of an independent personality. When the subaltern rebels, the master realizes that the servant too has a consciousness – interests and objectives, methods and organization. If one had to look for evidence of an autonomous subaltern consciousness in the historical archives, then it would be found in the documents of revolt and counterinsurgency.

The first phase of the work of *Subaltern Studies* was dominated by the theme of peasant revolt. Ranajit Guha's *Elementary Aspects of Peasant Insurgency in Colonial India* (1983) is the key text in this area. But most other scholars associated with the project wrote on the history of peasant revolts from different regions and periods of South Asian history. They were able to discover a few sources where the subaltern could be heard telling his or her own story. But it was always clear that there would be few such sources. What became far more productive were new strategies of reading the conventional documents on peasant revolts. The historians of subaltern politics produced several examples of ways in which reports of peasant rebellion prepared by official functionaries could be read from the opposite standpoint of the rebel peasant and thus used to shed light on the consciousness of the rebel. They also showed that when elite historians, even those with progressive views and sympathetic to the cause of the rebels, sought to ignore or rationally explain away what appeared as mythical, illusory, millenarian, or utopian in rebel actions, they were actually missing the most powerful and significant elements of subaltern consciousness. The consequence, often unintended, of this historiographical practice was to somehow fit the unruly facts of subaltern politics into the rationalist grid of elite consciousness and make them understandable in terms of the latter. The autonomous history of the subaltern classes, or to put it differently, the distinctive traces of subaltern action in history, were completely lost in this historiography.

The Imbrication of Elite and Subaltern Politics

The analysis of peasant insurgency in colonial India and of subaltern participation in nationalist politics by the historians of *Subaltern Studies* amounted to a strong critique of bourgeois-nationalist politics and the postcolonial state. Writing about peasant revolts in British India, Ranajit Guha (1983) and Gautam Bhadra (1994) showed how this powerful strand of anticolonial politics, launched independently of bourgeois-nationalist leaders, had been denied its place in established historiography. Gyanendra Pandey (1984, 1990), David Hardiman (1984), Sumit Sarkar (1984), and Shahid Amin (1995) wrote about the two domains of elite and subaltern politics as they came together in the nationalist movement led by the Congress. Dipesh Chakrabarty (1989) wrote about a similar split between elite and subaltern politics in the world of the urban working class. Partha Chatterjee (1986, 1993) traced the development of nationalist thought in India in terms of the separation of elite and subaltern politics and the attempts by the former to appropriate the latter. The post-colonial nation-state had, it was argued, included the subaltern classes within the imagined space of the nation but had distanced them from the actual political space of the state. Although the political emphasis was not the same in each writer, there was nevertheless a strong flavor in *Subaltern Studies* of the Maoist politics that had hit many parts of India in the 1970s. Many critics thought they could detect in it a romantic nostalgia for a peasant armed struggle that never quite took place. Others alleged that by denying the unifying force of the nationalist movement and stressing the autonomous role of the subaltern classes, the historians of *Subaltern Studies* were legitimizing a divisive and possibly subversive politics (see Ludden, 2001).

Another connection that was often made in the early phase of *Subaltern Studies* was with the 'history from below' approach popularized by British and French historians. Clearly, the work of British Marxist historians and the *Annales* school was eagerly mined by subalternist historians for methodological clues for doing popular history. But there was a crucial difference because of which subaltern history could never be a 'history from below.' The forgotten histories of the people, pulled out from under the edifice of modern capitalist civilization in the West, had undoubtedly made the story of Western modernity more detailed and complete. But there did not seem to be any narrative of 'history from below' that could persuasively challenge the existence, stability, or indeed the historical legitimacy of capitalist modernity itself. Not surprisingly, 'history from below' was invariably written as tragedy. But in countries such as India, 'history from below' could not be confined within any such given narrative limits. The subalternist historians refused to subscribe to the historicist orthodoxy that what had happened in the West, or in other parts of the world, had to be repeated in India. They rejected the framework of modernization as the necessary emplotment of history in the formerly colonial countries. They were skeptical about the established

orthodoxies of both liberal nationalist and Marxist historiographies. As a result, they resisted in their writings the tendency to construct the story of modernity in India as the actualization of modernity as imagined by the great theorists of the Western world. This resistance, apparent even in the early phase of *Subaltern Studies*, was to be expressed later in arguments about 'other modernities.'

Reorientations

One line of argument of the early *Subaltern Studies* project that ran into serious problems concerned the existence of an autonomous subaltern consciousness. Much of the study of the insurgent peasant was a search for a characteristic structure of peasant consciousness, shaped by its experience of subordination but struggling ceaselessly to retain its autonomy. One problematic question that arose was about the historicity of this structure. If subaltern consciousness was formed within specific historical relations of domination and subordination, then could that consciousness change? If so, why could it not be said that the Indian peasantry was transformed, modernized, turned into citizens by its experience of nationalist politics? Or was there another narrative of a changing subaltern consciousness? It was a classic structuralist impasse to which there was no easy historical answer.

A related problem was about the notion of the subaltern as an active historical agent. Research into subaltern history had shown that the subaltern was both outside and inside the domains of colonial governance and nationalist politics. To the extent that it was outside, it had retained its autonomy. But it had also entered those domains, participated in their processes and institutions and thereby transformed itself. Every bit of historical evidence was pointing to the fact that the subaltern was 'a deviation from the ideal.' Why then the search for a 'pure structure' of subaltern consciousness? Moreover, argued Gayatri Spivak (1987, 1988) in two influential articles, subaltern history had successfully shown that the 'man' or 'citizen' who was the sovereign subject of bourgeois history writing was in truth only the elite. Why was it necessary now to clothe the subaltern in the costume of the sovereign subject and put him on stage as the maker of history? Subaltern historiography had in fact challenged the very idea that there had to be a sovereign subject of history possessing an integral consciousness. Why bring back the same idea into subaltern history? It was only a myth that the subaltern could directly speak through the writings of the historian. In fact, the historian was only representing the subaltern on the pages of history. The subaltern, announced Spivak, cannot speak.

The new turn in *Subaltern Studies* began more or less from the fifth and sixth volumes published in 1987–89. It was now acknowledged with much greater seriousness than before that subaltern histories were fragmentary, disconnected, incomplete, split within itself, and constituted by elements drawn from the experiences of both dominant and subordinate classes. Alongside the evidence of autonomy displayed by subalterns at moments of rebellion, the forms of subaltern consciousness undergoing the everyday experience of subordination now became the subject of inquiry. Once these questions entered the agenda, subaltern history could no longer be restricted to the study of peasant revolts. Now the question was not 'What is the true form of the subaltern?' The question had become 'How is the subaltern represented?' 'Represent' here meant both 'present again' and 'stand in place of.' Both the subjects and the methods of research underwent a change.

One direction in which the new research proceeded was the critical analysis of texts. Once the question of the 'representation of the subaltern' came to the fore, the entire field of the spread of the modern knowledges in colonial India was opened up for subaltern history. Much studied aspects such as the expansion of colonial governance, English education, movements of religious and social reform, the rise of nationalism – all of these were opened to new lines of questioning by the historians of *Subaltern Studies*. The other direction of research concentrated on the modern state and public institutions through which modern ideas of rationalism and science and the modern regime of power were disseminated in colonial and postcolonial India. In other words, institutions such as schools and universities, newspapers and publishing houses, hospitals, doctors, medical systems, censuses, the industrial labor process, scientific institutions – all of them became subjects of subaltern history writing.

Other Modernities

A major argument that has been developed in the more recent phase of writings in and around *Subaltern Studies* is that of alternative or hybrid modernities. The focus here is on the dissemination of the ideas, practices, and institutions of Western modernity under colonial conditions. The framework of modernization theory invariably produces the history of modernity in non-Western countries as a narrative of lag or catching up. As Chakrabarty (2000) has put it, these societies seem to have been consigned forever to 'the waiting room of history.' The universality of Western modernity erases the fact that like all histories, it too is a product of its local conditions. What happens when the products of Western modernity are domesticated in other places? Do they take on new and different shapes – shapes that do not belong to the original? If they do, are we to treat the changes as corruptions? As deviations from an ideal? Or are they valid as examples of a different modernity?

To argue the latter is both to provincialize Europe and to assert the identity of other cultures even as they participate in the presumed universality of modernity. Chakrabarty (2000), Arnold (1993), Prakash (1999), and Spivak (1999), for example, have explored various aspects of this process of the 'translation' of modern knowledges, technologies, and institutions. They have tried to show that the encounter between Western forms of modernity and non-Western cultures was not a simple imposition of the one on the other, nor did it lead to corrupt or failed forms of modernity. Rather, it produced different forms of modernity whose marks of difference still remain subject to unresolved contestations of power. Kaviraj (2005) observed that the process of translation was better understood as involving different historical sequences of modernity in which its constitutive elements were serially produced in different countries, leading to different end products.

Here, the work of South Asian subalternist historians often overlapped with and contributed to what has become known as postcolonial studies. The historical study of modern discourses and institutions of power in colonial India fed into a growing literature on the production of hybrid cultural forms in many different regions of the formerly colonial world. More significantly, postcolonial studies extended the argument about hybrid cultural forms to the understanding of contemporary cultures in the Western metropolitan countries themselves, instanced most immediately in the diasporic cultures of immigrants but also, less obviously, in the role of the colonial experience in the formation of Western modernity even in its purportedly 'original' form.

Rethinking the Political

What then is the current understanding of subaltern politics? The early emphasis on peasant rebellion and consciousness had widened, even in the first phase of the project, to include the resistance of other dominated and marginalized groups in colonial society. Once the idea of a paradigmatic structure of subaltern consciousness became less persuasive, the subject of resistance began to be approached from a variety of angles and without a fixed design either of the reproduction of a traditional structure or of the necessary transition to new structures. Recent subaltern historiography has vigorously participated in three sets of South Asian debates – over religious minorities, caste, and gender. These debates have opened up the way to rethinking the political formation of the nation as well as the political process of postcolonial democracy.

In South Asia, the broader political debate over religious minorities has been carried out between two opposed groups – Hindu and Muslim chauvinists on the one hand and the secularists on the other. What the researches of subalternist historians have shown is that the debate between secularism and communalism is in no way a struggle between modernity and backwardness. Both the rival political positions are firmly planted in the soil of modern government and politics. Second, the two groups are pursuing two different strategies of consolidating the regime of the modern nation-state. Both strategies are elitist. But they involve two different modes of representation and appropriation of the subaltern. Third, faced with these rival elitist strategies, subaltern groups are devising, in their own ways, independent strategies of coping with communal as well as secularist politics. Recent experiences of ethnic violence and authoritarian politics in Sri Lanka and Pakistan have raised even more fundamental questions about the adequacy of the nation-form as a field where subaltern politics might be negotiated (Chatterjee and Jeganathan, 2000).

The second question on which there has been significant recent discussion is caste. There has been a transformation in the politics of caste in India following the agitations in 1990–91 over the Mandal Commission report. By looking at the politics of caste in its discursive aspect, it has become clear that the supposedly religious basis of caste divisions has now completely disappeared from public debate. The conflicts now are almost exclusively centered on the relative positions of different caste groups in relation to the state. Second, the debate over whether or not to recognize caste as a criterion of affirmative action by the state once again reflects two different elitist strategies of representing and appropriating the subaltern. And subaltern groups too, in their efforts to establish social justice and self-respect, are devising various strategies of both resisting the state as well as of using the opportunities offered by its electoral and developmental functions. Strategies of alliance of lower castes with middle or upper castes or with other oppressed groups such as tribals and religious minorities have produced significant electoral successes. But with the creation of new political elites out of subaltern groups, the questions of 'who represents?' and 'to what end?' are being asked with a new urgency (Mayaram et al., 2005).

The third debate concerns the social position of women. In one sense, all women living in patriarchal societies are subalterns. Yet it is not true that women are not identified by class, race, caste, and community. Hence, just as it is valid to analyze the subordination of women in a society ruled by men, so also is it necessary to identify how the social construction of gender is made more complex by the intervention of class, caste, and communal identities. Recent discussions on the question have focused on the Indian social reform movements of the nineteenth century, especially the various legal reforms to protect the rights of women, in the context of the colonial state and nationalist politics. Subaltern feminist writings have raised questions about the adequacy of a modernizing agenda of legal reform from the top without facing up to the challenge of reforming the actual structures of patriarchal power within the local communities that continue to flourish outside the reach of the law.

The recent debates have raised new questions about conceptualizing old ideas such as nation, citizenship, and democracy. The forms of subaltern participation in governmental processes and electoral politics in postcolonial democracies are now being conceptualized as modalities of democratic practice unknown in Western democracies (Chatterjee, 2004, 2011). By virtue of these implications, the recent subaltern history writings from South Asia have been productively used in writings on the history of modernity in other parts of the formerly colonized world, such as, for instance, on nationalism and gender in the Middle East or on the politics of peasant and indigenous groups in Latin America. Having traveled from Italy to India, the idea of subaltern history has now produced a generally available methodological and stylistic approach to modern historiography that could be used anywhere.

See also: Caste; Colonialism, Anthropology of; Colonization and Colonialism, History of; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historical Thought and Historiography: South Asia; Minorities; Peasants and Rural Societies in History (Agricultural History); Peasants in Anthropology.

Bibliography

Amin, S., 1995. Event, Metaphor, Memory: Chauri Chaura 1922–1992. Oxford University Press, Delhi, India.

Amin, S., Chakrabarty, D. (Eds.), 1996. Subaltern Studies IX. Oxford University Press, Delhi, India.
Arnold, D., 1993. Colonizing the Body: State Medicine and Epidemic Disease in Nineteenth-Century India. University of California Press, Berkeley, CA.
Arnold, D., Hardiman, D. (Eds.), 1994. Subaltern Studies VIII. Oxford University Press, Delhi, India.
Bhadra, G., 1994. Iman O Nishan: banglar Krishak Chaitanyer Ek Adhyaya [Faith and the Flag: An Aspect of Peasant Consciousness in Bengal]. Subarnarekha, Calcutta, India.
Bhadra, G., Prakash, G., Tharu, S. (Eds.), 1999. Subaltern Studies X. Oxford University Press, Delhi, India.
Chakrabarty, D., 1989. Rethinking Working-class History: Bengal 1890–1940. Princeton University Press, Princeton, NJ.
Chakrabarty, D., 2000. Provincializing Europe: Postcolonial Thought and Historical Difference. Princeton University Press, Princeton, NJ.
Chatterjee, P., 1986. Nationalist Thought and the Colonial World: A Derivative Discourse? Zed Books, London.
Chatterjee, P., 1993. The Nation and Its Fragments: Colonial and Postcolonial Histories. Princeton University Press, Princeton, NJ.
Chatterjee, P., 2004. The Politics of the Governed. Columbia University Press, New York.
Chatterjee, P., 2011. Lineages of Political Society. Permanent Black, Delhi, India.
Chatterjee, P., Jeganathan, P. (Eds.), 2000. Subaltern Studies XI. Permanent Black, Delhi, India.
Chatterjee, P., Pandey, G. (Eds.), 1992. Subaltern Studies VII. Oxford University Press, Delhi, India.
Gramsci, A., 1971. Selections from the Prison Notebooks (Q. Hoare, G. Nowell Smith, Trans.). International Publishers, New York.
Guha, R. (Ed.), 1982–1989. Subaltern Studies I–VI. Oxford University Press, Delhi, India.
Guha, R., 1982. On the historiography of Indian nationalism. In: Guha, R. (Ed.), Subaltern Studies I. Oxford University Press, Delhi, India, pp. 1–8.
Guha, R., 1983. Elementary Aspects of Peasant Insurgency in Colonial India. Oxford University Press, Delhi, India.
Hardiman, D., 1984. Peasant Nationalists of Gujarat. Oxford University Press, Delhi, India.
Kaviraj, S., 2005. An outline of a revisionist theory of modernity. Archives européennes de sociologie 46 (3), 497–526.
Ludden, D. (Ed.), 2001. Reading Subaltern Studies. Permanent Black, Delhi, India.
Mayaram, S., Pandian, M.S.S., Skaria, A. (Eds.), 2005. Subaltern Studies XII. Permanent Black, Delhi, India.
Pandey, G., 1984. The Ascendancy of the Congress in Uttar Pradesh, 1926–1934. Oxford University Press, Delhi, India.
Pandey, G., 1990. The Construction of Communalism in Colonial North India. Oxford University Press, Delhi, India.
Prakash, G., 1999. Another Reason. Princeton University Press, Princeton, NJ.
Sarkar, S., 1984. Popular Movements and Middle-Class Leadership in Late Colonial India. K.P. Bagchi, Calcutta, India.
Spivak, G.C., 1987. Subaltern studies: deconstructing historiography. In: Guha, R. (Ed.), Subaltern Studies IV. Oxford University Press, Delhi, India, pp. 338–363.
Spivak, G.C., 1988. Can the subaltern speak? In: Nelson, C., Grossberg, L. (Eds.), Marxism and the Interpretation of Culture. University of Illinois Press, Urban, IL, pp. 271–313.
Spivak, G.C., 1999. A Critique of Postcolonial Reason. Harvard University Press, Cambridge, MA.

Subculture, Sociology of

Rebecca G Adams and Amy M Ernstes, The University of North Carolina at Greensboro, Greensboro, NC, USA
Kelly M Lucey, North Carolina State University, Raleigh, NC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by F. Dubet, volume 22, pp. 15245–15247, © 2001, Elsevier Ltd.

Abstract

In this article, 'subculture' is defined simply as a culture embedded within a broader culture. Despite theoretical debates concerning its usage within the two main schools of thought, the Chicago School and the Birmingham Centre for Contemporary Cultural Studies, the concept remains important. Discussion of the underlying unity of conceptualizations and their variation invites reflection on the development of research on subculture and also foreshadows its future. Consideration of the emergence of similar terms, some proposed as replacements to subculture, and of the benefits of the application of such terminology, suggests pathways for future research. Specific issues include the changing relevance and understanding of subculture with advancing technology, as well as the insights to be gleaned from the related literature on community. This article concludes with a discussion of methodological approaches used to study subculture, with considerations for the future of its conceptualization and study in an increasingly digital age.

Introduction

In the most general sense, a 'subculture' is a culture embedded within a broader culture. The terminology used here is purposefully vague. This definition does not, for example, stipulate the relationship between a subculture and the broader culture in which it is embedded; it does not specify whether they are distinctly different, share some similarities, coexist in harmony, or are in conflict with each other. Furthermore, this definition does not describe the relationships among members of the subculture, whether they interact only online, sometimes see each other face-to-face, know each other, or are strangers. Thus, though extremely basic, this simple definition of a subculture as a culture within another culture seems the most reasonable place to start this article, especially given the degree to which the concept has been debated historically. This initial simplicity does not deny the complexities, controversies, and challenges bound to the concept of 'subculture,' but to circumvent these issues, focuses on 'culture' as its defining feature.

Honing in on culture provides an entrée into this discussion of subculture. Despite slight variations in terminology specified and prioritized, most social science textbooks provide a definition of 'culture' as resting upon a foundation of certain elements. Elements commonly included as part of this list are beliefs, values, norms, customs, rituals, ideas, knowledge, physical artifacts, language, symbols, and gestures. Often, these elements are separated into those that constitute material culture (physical and material objects) and those that constitute nonmaterial culture (values, norms, etc.). These disparate elements are united by a web of meanings shared by the members of a culture, or more pertinent here, by the members of a subculture. Thus, to elaborate on the basic definition of subculture provided here, we can now add that a subculture finds its distinction from another subculture or from the broader culture on the basis of these same disparate elements and the meanings attributed to them differing to some degree.

At this point the simplicity of this definition begins to dissolve. The difficulty in unpacking this definition any further arises from the problem of operationalizing the notion of difference. The question becomes how different and how self-consciously different a social group must be in order to be considered distinct. Establishing such parameters has become more problematic as our world has become increasingly globalized and as modern technological advancements in information, communication, and transportation have served to erode temporal and geographic boundaries that once confined the members of social groups to shared times and spaces (Adams, 1998). Descriptions that define a subculture as a distinct culture from a dominant culture have grown less relevant in a postmodern age in which templates of cultural hierarchies have at the least become blurred, if not identifiable as entirely historically inaccurate and useless. Furthermore, the notion of a dominant culture has long been critiqued for implying 'normalization' through which subcultures have often been narrowly cast as deviant, whether subtly or overtly.

Yet, despite the challenges of defining subculture and questions regarding the ongoing relevance of the concept, studies of subculture have clearly added to the understanding of society. Studying subcultures has allowed both a detailed and nuanced perspective on cultural groups not necessarily well known or understood by the 'mainstream' (as the larger, dominant culture is often called) and at the same time has provided a broad and inclusive approach in terms of recognizing instead of ignoring these distinct groups. The study of subcultures thus provides a more realistic picture of a society by recognizing complex layers of divergent groupings within it. For, as is the case given the nature of humans and humanity, as long as societies have existed, so too have a variety of ways of building culture within them. This necessarily social process of providing meaning to that existence will inevitably vary in form and expression across different societies and social groupings.

Furthermore, although much of the research on subcultures focuses on social groups within the English-speaking world,

culturally distinct groups exist within all societies. In other words, subcultures have been present throughout time (i.e., history) and across space (i.e., the globe); the elements of subcultures have been transmitted from one generation to the next and diffused across geographic boundaries. Subcultures serve as the basis for individual identity and group solidarity, which in turn leads to both social integration and intergroup dynamics. Not only are subcultures intrinsic parts of societies and thus important to study, but doing so has also allowed for the recognition, understanding, and sometimes even acceptance or appreciation of these social groups, which has had important consequences, especially for those subcultures that were considered deviant and consequently were marginalized (e.g., see Decade in Brief (2012) for a discussion of the Decade of Roma Inclusion 2005–15, "an unprecedented political commitment by European governments to eliminate discrimination against Roma and close unacceptable gaps between Roma and the rest of society").

Theoretical Issues: The Definition of Subculture

Despite the initial simplistic definition, and the undeniable value of studying subcultures, the trajectory of its use in the social sciences has not been so straightforward. Although two main schools of thought have dominated conceptualizations of subculture – the Chicago School and the Birmingham Centre for Contemporary Cultural Studies (CCCS) – so too have the critiques of these conceptualizations.

Yet, despite the debates that have come to hallmark the term, questioning its validity and even arguing against its relevance, beneath this divergence is an arguably unified understanding of the term. In other words, while broadened and narrowed, expanded, and simplified, and redirected and refocused, the various conceptualizations of subculture have nonetheless generally unfolded within a shared framework. Although early anthropologists used the term to mean universal tendencies that occur in all societies (e.g., Linton, 1936), that definition was later eclipsed by broader ones resting on such facets as those discussed earlier in this entry; an example is "normative systems of groups smaller than a society, to give emphasis to the ways these groups differ in such things as language, values, religion, diet, and style of life from the larger society of which they are a part" (Yinger, 1960: p. 626). Chicago School sociologists popularized this type of definition in research to explain deviant behavior, arguing that behavior perceived to be deviant from a mainstream perspective could be in keeping with the norms of a smaller group within society (Becker, 1963; Matza and Sykes, 1961). As early as 1960, Yinger noted that a third definition of subculture as determined by "norms that arise specifically from a frustrating situation or from conflict between a group and a larger society" had also been used in the literature; rather than recommending the element of conflict be incorporated into the definition of subculture, he suggested these subcultures be designated as 'contracultures' to distinguish them from subcultures that existed more harmoniously with the mainstream (Yinger, 1960: p. 629).

This distinction between subcultures and contracultures continued to be recognized in much of the research that was conducted in the 1960s and 1970s (e.g., Dorn, 1969; Kando, 1975; Rich, 1972; Sebald, 1975); the two notions were not reconflated until the CCCS began its research on post–World War II working-class youth subculture. Although building on Chicago School studies of deviant subcultures, the publication of *Resistance through Rituals* (Hall and Jefferson, 1976) and *Subculture: The Meaning of Style* (Hebdige, 1979) shifted the discussion from an examination of the patterns across a variety of deviant subcultures to a societal perspective on the ongoing class struggle in Great Britain and the stylistic responses of youth.

Although some scholars have continued to apply the more general definition of subculture in their work on youth subcultures (e.g., Hunt, 2008, 2010; Latysheva, 2011), and especially in examinations of subcultures not associated exclusively with youth (Lieske, 2004; Spencer, 1997), after the introduction of the notion of youth subcultures as 'cultures of resistance,' this more specified definition has dominated the subculture literature and much of the critique of the notion of subculture since then is a response to this more specific usage (e.g., Crosset and Beal, 1997; Miles, 2000; Straw, 1999). Muggleton and Weinzierl (2003), for example, question the validity of this approach in a postmodern world, and champion "a more pragmatic approach compared to the romanticism of the CCCS" (p. 4). Short of a direct critique, others have suggested moving beyond this narrow scope, such as Fincham (2007), who champions the aspect of 'enjoyment,' as one often overlooked as an attraction to membership in a subculture. Interestingly enough, Simon Frith (2004), a major contributor to this literature on British youth subculture, recently observed that what is being called 'postsubcultural theory' (Bennett and Kahn-Harris, 2004; Muggleton and Weinzierl, 2003) does not focus on youth at all. So, despite all the controversy, it appears that the usage of the concept of 'subculture' may have come full circle and is now being used in the same two ways Yinger noted were common in 1960, in a more general sense but sometimes also in what is now identified as a Birmingham School approach incorporating a notion of resistance or conflict.

Albeit the approach to subculture inspired by the Birmingham School seems to have run its course, now much more is known about youth subculture, not just in Britain, but globally (e.g., Baron, 1989; Jasper, 2004; Johnston and Snow, 1998; Maira, 1999; Mitchell, 2003; Platek, 1990). Furthermore, the debates over the validity and utility of the concept of subculture have led to the application and sometimes development of related concepts that have shed light on other aspects of social groups that could have been studied using a subcultural theoretical lens. So, although as Muggleton and Weinzierl (2003: p. 5) observe, "a new, dominant paradigm has yet to develop," a plethora of concepts have been proposed to replace or supplement the old one. Although most of the critiques of the concept of subculture have been in response to attempts to apply it to study a specific type of social group that has characteristics not revealed by its application, the tendency of most authors seems to be to try to propose a definition of subculture or an alternative concept that will be consistently useful to apply across social groups. Here we propose a different approach, which is to embrace the use of a variety of concepts, including subculture, to study the component social groupings of a larger one and thereby to increase what is known about them.

To elaborate on this proposal, it is useful to understand subculture and its related concepts as what Vaihinger (1925) called 'neglective fictions.' Although fictions are devoid of reality, they are not devoid of utility. According to Vaihinger, 'fictions' have been used to advance scientific understanding of phenomena, much like hypothesis testing has. We are suggesting here that concepts such as 'subculture,' 'contraculture,' and others to be discussed in this article be applied to social groupings, much as Adam Smith (1776/1976) wrote about human behavior 'as if' it were all egoistic. He deliberately substituted a fraction of reality, what Vaihinger called a 'neglective fiction,' for the total range of possible explanations and facts. Without this substitution, Smith would have been unable to construct an ordered interpretation of the political economy. Similarly, without using neglective fictions such as subculture, contraculture, culture of resistance, lifestyle, tribe, scene, subworld, or subsociety as lenses, the understanding of the role of these groupings in shaping identity and society would not be as great. We are suggesting that researchers could use and have used various concepts related to subculture either independently or in sequence as lenses to study subgroups of a larger culture or society as Simmel (1986/1907) would apply various forms (i.e., various neglective fictions) to the same content (Adams, 2010; Kaern, 1983). Through this process, which is already ongoing though the researchers using this approach might not associate it with Vaihinger's scientific fictions, not only will we learn more about the social groupings under scrutiny, we will also reach greater understanding of the utility and patterns of relevance of the theoretical concepts themselves.

In keeping with this perspective on subculture and related concepts as neglective fictions, many of the related concepts just mentioned have been developed in response to attempts to understand specific social groupings by first using subculture as a lens and finding it lacking for some theoretical or substantive reason. For example, Crosset and Beal (1997) propose the use of the concept 'subworld' in addition to the concept 'subculture.' Building on Unruh's (1983) definition of social worlds "as large and highly permeable, amorphous and spatially transcendent forms of social organization made up of people sharing common interests and sharing common channels of communication," they note that "[t]he concept of 'social world' pushes the researchers' gaze in a slightly different direction than does 'subculture'" (p. 81). Unlike analyzing sports as subculture, analyzing it as a social world allows the discovery of subworlds within social worlds, and at least unlike analyzing sports from a CCCS perspective, it "frees the researcher from looking for signs of cultural resistance in every sporting practice." Similarly, although the concept of 'scene' is not a new one (Irwin, 1977) and originally described expressive, voluntary, public gatherings that take place in urban locations, Straw (1999: p. 373) and others have recently introduced the term scene in part to replace older conceptions of music communities as having relatively permanent membership focused on specific musical idioms. Straw's reconceptualization of 'scene' casts it as a space in which a range of musical activities coexist and cross-fertilization and change take place. Miles (2000) provided a third example of a concept introduced in order to understand a specific phenomenon when he argued that the application of the term 'subculture' to heavy drug users makes sense, but suggested 'lifestyle' as a more useful term when researching people whose drug use is intermittent.

Rather than suggesting 'subculture' be replaced with other concepts because of the inadequacy of subculture to highlight the important features of a specific social grouping, sometimes theorists have recommended it be replaced by other concepts in response to widespread social changes such as globalization and multiculturalism. Like Miles (2000), Chaney (2004) also suggests the use of the term 'lifestyle' might be more useful in the contemporary context but for very different reasons. He argued that the idea of a multicultural world has rendered the distinction between dominant and subcultural norms moot and therefore the notion of a dominant culture as a context for subcultures archaic. Furthermore, he (Chaney, 2004) argued that the widespread availability of a greater variety of cultural styles and identities suggests the notion of subculture with that of lifestyles as the basis for the negotiation of identity, community, affiliation, and difference. Implicitly he argues that subcultures no longer exist.

More recently, the definition of subculture is being scrutinized for its usefulness in the context of improved communications and transportation technology (Bennett, 2004). The Chicago School studies of subcultures such as taxi dancers (Cressey, 1932) and gangs (Thrasher, 1927) were situated in local contexts and the need for close proximity and face-to-face interaction were assumed to provide a foundation for their development, persistence, and stability. Although it is the advent of the Internet that has led subculture scholars to challenge these assumptions, other earlier improvements in communications and transportation technology such as the postal system, telegraph, telephone, steam ships, locomotives, and automobile began changing the way subcultures developed and persisted much earlier (Adams, 1998). Nonetheless, the Birmingham School's examination of punks in the United Kingdom in the 1970s continued the Chicago tradition of focusing on subcultures in the local context, and much of the work accomplished by scholars elsewhere but working within these traditions often did the same.

In this regard, the subculture literature has lagged behind the community literature in which the discussion of non-territorial communities had already begun in the 1960s (Webber, 1963; Effrat, 1974) and has continued since (Adams, 1998; Wellman, 1999). As a result, scholars who have studied migrating groups or intermittently territorial groups such as the New Age Travellers (Frediani, 1998), some music fans (Gardner, 2004), or National Association for Stock Car Auto Racing spectators (Wright, 2002) have sometimes implicitly if not explicitly used 'community' as a conceptual lens rather than 'subculture.' For example, in her study of Deadheads, the fans of the Unites States rock group the Grateful Dead, Adams originally intended to use 'subculture' as a conceptual lens (Adams, 1990), but subsequently shifted to using the concept of 'community' for this purpose in part because it allowed a focus on the development and continuity of identity and solidarity across time and space (Adams et al., 2014). Although the members of such communities do not share a permanent territory, like members of local subcultures and traditional communities, they are bound by affect, customs, and common interests (Nels, A. and Council of Social Agencies of Chicago,

1923; Tönnies, 1963/1887). Also like local subcultures and traditional communities, these intermittently territorial communities sometimes serve the functions of production, distribution, and consumption; socialization; social control; social participation; and mutual support (Humphreys, 1970; Warren, 1978).

In contrast to Chaney's (2004) position that the notion of subculture is archaic in a multicultural world, it could be argued that with the advent of the Internet and the related development of virtual media of various forms, it is even easier for subcultures to transcend space and time and for true global subcultures to exist. As one celebrates the 25th anniversary of the World Wide Web, it seems particularly appropriate to reflect on its impact on culture. As early as 1994, Rheingold discussed the 'virtual community' and connectivity within it. Shortly thereafter, Healy (1997) described it as a 'loose collection' or 'ecosystem' of subcultures (p. 65). Subsequently, scholars in both the community (Smith and Kolluck, 1999) and subculture traditions (Bennett, 1999) began considering how the Internet challenged the relevance of concepts developed in an earlier time period. The literature on migrating or intermittently territorial communities is somewhat useful here, because like the literature on online subcultures, it does not assume permanent shared territory and once the Internet developed these communities also manifested themselves online. Although as Bennett (2004) astutely observed, subcultural studies often overlook the continuities between online and offline interactions, not all virtual communities manifest themselves locally. The development of the Internet has thus exacerbated concerns of postmodernists such as unstable identities (Lash and Urry, 1987, 1994) and the lack of emotional and moral depth of communities based solely on common interests (Bauman, 1992).

Taken together, these concerns beg the current question of whether this technology detracts from the development of subcultures that provide the context for stable identities and feelings of solidarity with a subculture of like-minded people or leads to shifting identities and weak commitments to multiple subcultures. This question is particularly salient in light of the emergence of 'one device' portable technology and the ability of people to remain 'plugged in' constantly and to manage activity in multiple subcultures or communities at one time. On the one hand, simultaneous ties to multiple communities can be hypothesized to weaken the individual connection to a subculture, thereby making it more likely that individual identities will shift over time and weaken subcultural solidarity. On the other hand, even before the advent to of the Internet, Fine and Kleinman (1979: p. 12) noted that multiple group memberships promote cultural diffusion and the spread of subcultures, and more recently Thornton (1996) has argued that media, including the Internet, are integral to the formation of subcultures, playing a significant role in both their origin as well as prolonging their life cycle. McArthur's study of the virtual subculture of 'geeks' supports Thornton's observation as it would have been very difficult for his participants to meet in person as they were from many nations including The Philippines, England, Argentina, and Czech Republic (McArthur, 2009). Future scholarship will hopefully reveal the waxing or waning of the significance of subcultures in a global society.

Methodology

Subculture researchers have relied mainly on ethnographic methods, as qualitative approaches afford insight into valuable details difficult to capture with a broader lens. Due to previous conceptions of subculture as local and therefore convenient (both in the traditions of the Chicago and Birmingham Schools) and the tendency of researchers to study deviant subcultures that are difficult to enter, much of the research on subcultures has specifically involved participant observation. Nonetheless, as many researchers, including Fine and Kleinman (1979), acknowledged years ago, the narrowness of such methods embodies their own limitations. For example, a case study of one gang could not claim representation of all gangs in a given population or yield any definite findings regarding the degree of interrelatedness between them.

Some subculture researchers have thus instead conducted surveys of samples of general populations (e.g., King, 2007; Latysheva, 2011), but this method can be ineffective, especially if the interviewer is not identified with the subculture or when the interviewer's values differ from the subculture's. Returning to the aforementioned limitations of a finer lens, "while case studies provide valuable information about the culture, it is unclear to what extent they provide evidence of the culture as a population segment" (Fine and Kleinman, 1979: p. 4). Yet, the broader study of general populations also yields difficulty, particularly in identifying subcultural groups (King, 2007). Given that in order to demonstrate a subculture's existence, a subpopulation's relative agreement on a set of attitudes, behaviors, or values must be identified, the researcher must necessarily show that communication occurs within that subpopulation and that its members identify as a group. For this reason, even in the challenging case of studying migrating or intermittently territorial subcultures, researchers examining them have tended to supplement any survey or documentary research with their observations (e.g., Hunt, 2008, 2010; Riley, 1988).

The globalization of culture and the introduction of the Internet and related technologies have not only challenged theoretical conceptions of subculture, but have also led to changes in how they are studied. As exemplified by McArthur (2009) in his study of 'geeks,' the Internet allows new formats for data to be gathered on subcultures, at the same time that it allows for new forms of expressing subcultural identity, as well as influencing those identities themselves. The future of subculture analysis may benefit from the accessibility of information and even the resurfacing of past information and data previously obscured by data security policies (although individuals participating in subcultures through different mediums in the past may never have realized that information would one day be made available electronically and via search engines). The criticism that studies are focused on subcultures at one point in time can now be addressed by expanding the search backward and more broadly to include an ever-increasing amount of information available online. The Internet also makes it possible to study distant subcultures remotely. Finally, videos and multimedia files produced by other researchers are shared in the advent of a very open attitude toward information sharing.

Although the advent of the Internet facilitates research on subcultures, it also complicates it. Despite the growth of the Internet, it is not equally accessible to all populations. The 'digital divide' represents the absence or limitation of access to the Internet for certain people or groups based on affordability, knowledge, or motivation. Furthermore, even in the hypothetical scenario that all potentially desired participants would have similar access to and know-how regarding these technological means, self-selection would remain an issue.

Nonetheless, the increased availability of information will certainly aid in research on subcultures that exist in a variety of modes and will address some of the previously articulated, but nonetheless still valid, challenges of qualitative research identified by Fine and Kleinman (1979): the Internet makes it possible to discover linkages among groups, track the transmission and forms of information and exchanges of communication, and identify kinds of information that is transmitted. Furthermore, given the progression of technology to allow for more remote and anonymous participation in a subculture, many studies themselves may in turn become focused purely on online cultures (King, 2007). While ethnography remains an important method for studying subculture, the existence of the Internet may lead to the realization of a new form of interdisciplinary research, combining ethnography with methods such as textual analysis (Pollex, 1997), an approach some have called 'postmodern ethnography' (Tyler, 1986). Such a new method would have to be focused on mining data and organizing it in such a way that large numbers of disparate pieces can be integrated and interpreted in meaningful ways. The Internet spreads information quickly and retains it forever, so future subcultural analysts will need to develop methods for sampling this vast array of information and deciding how valid it is. As the information age continues to unfold, the challenges and benefits the Internet poses to theorists and researchers will no doubt lead to substantial changes in how subculture is understood and studied.

See also: Community Sociology; Community Studies: Anthropological; Community and Society: History of the Concepts; Counterculture: 1960s and Beyond; Counterculture: The Classical View; Cultural Sociology, History of; Gay, Lesbian, and Bisexual Youth; Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Popular Culture; Popular Sexual Cultures in Western Europe (1930s to the Present); Roma People; Youth Culture, Anthropology of; Youth Culture, Sociology of; Youth Movements; Youth, Geography of; Youth, Music, and Peace Building.

Bibliography

Adams, R.G., 1998. The demise of territorial determinism: online friendships. In: Adams, R.G., Allan, G. (Eds.), Placing Friendship in Context. Cambridge University Press, Cambridge, pp. 153–182.

Adams, R.G., 2010. "I can't do anything but lie": studying Deadheads while wearing Simmelian lenses. In: Tuedio, J., Spector, S. (Eds.), The Grateful Dead in Concert. McFarland, Jefferson, NC and London, pp. 310–322.

Adams, R.G., 1990. (Associate Producer and Narrator). Deadheads: An American subculture (28 minute video). Produced by UNCG Office of Continuing Education, directed by Emily Edwards, distributed by Films for the Humanities & Sciences, and broadcast on PBS stations nationally. Re-released by PBS in 1995. Available at: www.youtube.com/watch?v=QJ0JvjhQTcY.

Adams, R.G., Ernstes, A.M., Lucey, K.M., 2014. After Jerry's death: achieving continuity in Deadhead identity and community. In: Duffet, M. (Ed.), Popular Music Fandom: Identities, Roles and Practices. Routledge, New York, pp. 186–206.

Baron, S., Summer 1989. The Canadian West Coast punk subculture: a field study. Canadian Journal of Sociology 14 (3), 289–316.

Bauman, Z., 1992. Intimations of Postmodernity. Routledge, London.

Becker, H.S., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press Glencoe, Oxford, England.

Bennett, A., Kahn-Harris, K., 2004. After Subculture: Critical Studies in Contemporary Youth Culture. Palgrave MacMillan, New York.

Bennett, A., 1999. Subcultures or neo-tribes? Rethinking the relationship between youth, style and musical taste. Sociology, 33–59.

Bennett, A., 2004. After subculture Chapter 11: Virtual subculture? Youth, identity, and the Internet. In: Bennett, A., Kahn-Harris, K. (Eds.), After Subculture: Critical Studies in Contemporary Youth Culture. Palgrave MacMillan, New York, pp. 162–172.

Chaney, D., 2004. Chapter 2: Fragmented culture and subcultures. In: Bennett, A., Kahn-Harris, K. (Eds.), After Subculture: Critical Studies in Contemporary Youth Culture. Palgrave MacMillan, New York, pp. 36–48.

Cressey, P., 1932. The Taxi-Dance Hall: A Sociological Study in Commercialized Recreation and City Life. University Chicago Press, Chicago.

Crosset, T., Beal, B., 1997. The use of "Subculture" and "Subworld" in ethnographic works on sport: a discussion of definitional distinctions. Sociology of Sport Journal 14, 73–85.

Dorn, D., 1969. Self-concept, alienation, and anxiety in a contraculture and subculture: a research report. Journal of Criminal Law and Criminology 59 (4), 531–535.

Effrat, M.P. (Ed.), 1974. The Community Approaches and Applications. Free Press, New York.

Fincham, B., 2007. 'Generally speaking people are in it for the cycling and the beer': bicycle couriers, subculture and enjoyment. The Sociological Review 55 (2), 189–202.

Fine, G.A., Kleinman, S., 1979. Rethinking subculture: an interactionist analysis. American Journal of Sociology 85 (1), 1–20.

Frediani, M., 1998. La Baraque. An alternative experience. In: Dearling, A. (Ed.), No Boundaries: New Travellers on the Road (Outside of England). The Enabler Publications, Lyme Regis, pp. 133–139.

Frith, S., 2004. Afterword. In: Bennett, A., Kahn-Harris, K. (Eds.), After Subculture: Critical Studies in Contemporary Youth Culture. Palgrave MacMillan, New York, pp. 173–178.

Gardner, R.O., 2004. The portable community. Symbolic Interaction 27 (2), 155–178.

Hall, S., Jefferson, T., 1976. Resistance through Rituals: Youth Subcultures in Post-war Britain. Routledge.

Healy, D., 1997. Cyberspace and place: the internet as middle landscape on the electronic frontier. In: Porter, D. (Ed.), Internet Culture. Routledge, London, pp. 55–68.

Hebdige, D., 1979. Subculture: The Meaning of Style. Routledge.

Humphreys, L., 1970. Tearoom Trade: Impersonal Sex in Public Places. Aldine Publishing Company, Chicago.

Hunt, P.M., 2008. From festies to tourrats: examining the relationship between Jamband subculture involvement and role meanings. Social Psychology Quarterly 71 (14), 356–378.

Hunt, Pamela M., 2010. Are you kynd? conformity and deviance within the Jamband subculture. Deviant Behavior 31, 521–551.

Irwin, J., 1977. Scenes. Sage, Beverly Hills.

Jasper, A., 2004. I am not a goth! The unspoken Morale of authenticity within the Dutch Gothic subculture. Etnofoor 17 (1/2), 90–115.

Johnston, H., Snow, D., 1998. Subcultures and the emergence of the Estonian nationalist opposition 1945–1990. Sociological Perspectives 41 (3), 473–498.

Kaern, M., 1983. Understanding Georg Simmel. Sociological Focus 16 (3), 169–179.

Kando, T., 1975. Popular culture and its sociology: two controversies. The Journal of Popular Culture 9 (2), 438–455.

King, A.B., 2007. Finding online subcultures in shared meanings. Social Science Computer Review 26, 137–151.

Lash, S., Urry, J., 1987. The End of Organized Capitalism. Wisconsin.

Lash, S., Urry, J., 1994. Economies of Signs and Space. Sage.

Latysheva, T.V., 2011. The essential nature and types of the youth subculture phenomenon. Russian Education and Society. M.E. Sharpe, Inc. 53, 73–88.

Lieske, J., 2004. The changing regional subcultures of the United States: a new cultural measure for understanding political behavior. Prepared for Delivery at the 2004 Annual Meeting of the Midwest Political Science Association, Chicago Palmer House, 15–18 April.

Linton, Ralph, 1936. The Study of Man. Appelton-Century, New York.
Maira, S., 1999. Identity dub: the paradoxes of an Indian American youth subculture (New York mix). Cultural Anthropology 14 (1), 29–60.
Matza, D., Sykes, G., 1961. Juvenile delinquency and subterranean values. American Sociological Review 26 (5), 712–719.
McArthur, J.A., 2009. Digital subculture: a geek meaning of style. Journal of Communication Inquiry 33 (1), 58–70.
Miles, S., 2000. Youth Lifestyles in a Changing World. Open University Press, London.
Mitchell, T., 2003. Australian hip hop as a subculture. Youth Studies Australia 22 (2), 40–47.
Muggleton, D., Weinzierl, R., 2003. The Post-subcultures Reader. Berg Publishers, Oxford.
Nels, A., Chicago Council of Social Agencies, 1923. The Hobo: The Sociology of the Homeless Man. University of Chicago Press.
Platek, M., 1990. Prison subculture in Poland. International Journal of the Sociology of Law 18 (4), 459–472.
Pollex, M., 1997. Review: Ken Gelder and Sarah Thornton, the subcultures reader. Canadian Journal of Sociology 24 (2), 303–306.
Rich, A., 1972. Contracultural language codes: the language of survival. Paper Presented at the Western Speech Communication Association Convention, Honolulu.
Rheingold, H., 1994. The Virtual Community. Harper Perennial, New York.
Riley, P., 1988. Road culture of international long term budget travelers. Annals of Tourism Research 15 (3), 312–328.
Sebald, H., 1975. Subculture: problems of definition and measurement. International Review of Modern Sociology 5 (1), 82–89.
Simmel, G., 1986/1907. Schopenhauer and Nietzsche (H. Loiskandl, D. Weinstein, and M. Weinstein, Trans.). The University of Massachusetts Press.
Smith, A., 1976/1776. In: Campbell, R.H., Skinner, A.S., Todd, W.B. (Eds.), The Wealth of Nations, vol. 1 and 2. Clarendon Press, Oxford.
Smith, M.A., Kolluck, P., 1999. Communities in Cyberspace. Routledge, London and New York.
Spencer, Nancy E., 1997. Once upon a subculture: professional women's tennis and the meaning of style, 1970–1974. Journal of Sport and Social Issues 21, 363.
Straw, W., 1999. Systems of articulation, logics of change: communities and scenes in popular music. Cultural Studies 5 (3), 368–388.
Thornton, S., 1996. Club Cultures: Music, Media and Subcultural Capital. Wesleyan University Press, Hanover, NH.
Thrasher, Frederick, 1927. The Gang. University of Chicago Press.
Tönnies, F., 1963/1887. In: Loomis, C.P. (Ed. and Trans.), Community and Society. Harper & Row, New York.
Tyler, S., 1986. Postmodern ethnography from document of the occult to occult document. In: Clifford, J., Marcus, G. (Eds.), Writing Culture: The Poetics and Politics of Ethnography, pp. 122–140.
Unruh, D.R., 1983. Invisible Lives: Social Worlds of the Aged. Sage, Newbury Park.
Vaihinger, H., 1925. The Philosophy of the 'as if'. Harcourt, New York.
Warren, R.L., 1978. The Community in America. Rand McNally, Chicago.
Webber, M., 1963. Order and diversity: community without propinquity. In: Wingo Jr., L. (Ed.), Cities and Space. Johns Hopkins, pp. 23–54.
Wellman, B. (Ed.), 1999. Networks in the Global Village: Life in Contemporary Communities. Westview Press, Boulder, CO.
Wright, Jim, 2002. Fixin to Git: One Fan's Love Affair with NASCAR's Winston Cup. Duke University Press, Durham, NC and London.
Yinger, J.M., 1960. Contraculture and subculture. American Sociological Review 25 (5), 625–635.

Relevant Websites

https://www.youtube.com/ – Cantelo, Erin (series creator), 2013. Subcultures – a documentary series spotlighting niche communities you might want to be part of. From "Roller Derby Girls" to "Guerilla Gardeners".
http://www.romadecade.org/ – Decade in Brief, 2012. Decade of Roma Inclusion Secretariat Foundation.
http://designtaxi.com – A Guide To Recognizing 21st Century Subcultural Tribes.
http://www.theguardian.com/ – A Journey through London Subculture: 1980s to Now, 2013.

Subjective Wellbeing and Culture

Sim Jing Han and Sharon Koh, Singapore Management University, Singapore, Singapore
Christie Napa Scollon, School of Social Sciences, Singapore Management University, Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Subjective well-being (SWB) is what most people refer to as happiness. It is people's subjective sense that life is going well. Many factors influence SWB, and culture is of course an important one. The present article describes the cultural differences in SWB, beginning with the desirability of happiness around the world. Next, the authors discuss the major cultural differences in SWB, for example, which regions and cultures are happiest. Next, the article reviews the literature that has sought to explain the cause of these differences. Finally, the article concludes with a view toward future research in culture and SWB.

What does it mean to be happy? What is a good life? Most of us have thought about these questions at some point in our lives. Researchers studying subjective well-being (SWB), however, seek to empirically study happiness through the use of scientific methods like surveys and laboratory experiments to investigate, for example, who is happy and what are the societal conditions that contribute to happiness.

Within this field, the impact of culture on SWB has been of interest to researchers. At the most basic level, one of the key questions that researchers try to understand is whether happiness is a universal value that is important to people around the world. Next, if some cultures are happier than others, we would want to know why this occurs. After all, if some cultures are doing things that promote happiness, perhaps less happy nations could learn a thing or two from the happy ones. In addition, researchers constantly face the challenge of different cultural conceptions of happiness and how to make meaningful comparisons in light of these differences. In this article, we review the current literature in the study of SWB among cultures and discuss the factors that contribute to these differences.

Definitions

What Is Subjective Well-Being?

SWB is what most people recognize as happiness. Among researchers, SWB broadly refers to people's subjective evaluations of their lives. It is people's subjective sense that life is going well. The specific components are typically measured as the frequency of pleasant over unpleasant emotions as well as a cognitive evaluation of one's life as a whole (i.e., life satisfaction) (Diener, 1984). Although these components are highly correlated – that is, someone with high SWB typically experiences lots of pleasant emotions, few unpleasant ones, and has high life satisfaction – the components have been demonstrated to be empirically separate as well. Importantly, because of the subjective nature of SWB, most researchers measure SWB simply by asking people how they feel about their lives. This straightforward approach has been shown to have strong validity and is conceptually meaningful. Of course, SWB researchers do not argue that objective indicators of a life are not important, rather that SWB is about the subjective evaluation of one's life and therefore people are the best sources of information about their own happiness.

What Is Culture?

When thinking about culture, we typically consider various aspects of societies like religion, food, and language. Psychologists, however, understand culture to be something deeper and define culture as an evolved collection of practices and beliefs that people share and transmit between generations in order to allow them to coordinate their actions and satisfy their needs (Chiu et al., 2010). Simply put, culture is a shared understanding of how we go about living together in groups. Researchers studying culture commonly examine how people understand themselves, their values, as well as their environment.

Broadly speaking, the ideas of individualism and collectivism provide us with a useful way to think about different cultures. In individualist cultures, the uniqueness of individuals and their choices and achievements are the focus of a person's identity (Hofstede, 1980; Triandis, 1989). In collectivist cultures, group memberships shape a person's identity and group goals are more important than personal desires (Triandis, 1989). For example, individualists are more likely to finish the sentence 'I am …' with personal traits (e.g., 'I am a boy' or 'I am very intelligent'). Collectivists, on the other hand, tend to finish the same sentence with social roles or group memberships (e.g., 'I am a son' or 'I am a student') (Diener and Diener, 2008). There are variations in the classification of cultures within the individualistic–collectivistic spectrum. As a general guideline, most Western societies like the United States and Australia are characterized as individualistic while Eastern Asian societies like Japan and South Korea are collectivistic cultures.

Differences in SWB around the World

The Desirability of Happiness across Cultures

Happiness appears to be a universal value for people around the world. Diener (2000) asked college students from 42 different countries about the importance of happiness, love, going to heaven, and being rich. Happiness emerged as the top

value in every country studied, even edging out values such as eternal salvation and wealth. Happiness scored well above the midpoint in every single country, typically around a 7 or an 8 on a 10-point scale where 10 equaled 'extremely important.' The lowest rating given to happiness was 5.03 in Bulgaria. Although there were differences among countries in the importance assigned to the various values, in no society sampled did happiness emerge as less important than wealth.

The Experience of Happiness across Cultures

Though most people desire to be happy, actual scores of happiness vary widely across cultures. There are now several comprehensive world-wide surveys that provide an annual rank order of nations in terms of the happiness of their citizens. For example, Gallup recently conducted a survey of happiness that involved over 100 000 respondents in over 100 countries, with representative samples in each country. In addition, the United Nations also releases an annual publication, the World Happiness Report that incorporates data from over 150 countries. Similar to the study by Diener (2000), the World Happiness Report uses a variant of the satisfaction with life scale, asking participants to rate their overall satisfaction with life on a 10-point scale where 10 equaled 'completely satisfied.' Based on data collected from 2010 to 2012, this report ranked Denmark, Norway, and Switzerland as the most happy in the world, while Central African Republic, Benin, and Togo ranked lowest (Helliwell et al., 2013).

While it is silly to debate whether the number one spot belongs to Denmark or to Luxembourg, and whether Singapore is 30th or 33rd, there are clear and consistent regional differences in reports of SWB. In general, countries in North America and Australia/New Zealand are quite happy, as are countries in Western Europe and Latin America. The African nations are almost always overrepresented in the bottom of the rankings, while the Middle Eastern states are also not very happy. Asia is not miserable, but is also not nearly as happy as the Western societies. While the unhappiness of the African states makes sense in light of their economic conditions, the happiness of Asian countries is surprisingly low, given their gross domestic product (GDP), and this continues to puzzle scientists interested in culture and SWB (Diener et al., 1995). Interestingly, Rice and Steele (2004) examined Americans who varied in ancestry and found that people's SWB levels were more strongly related to that of their ancestors than that of their current neighbors. In other words, a Chinese American is more like the Chinese in China in reports of SWB than an Irish American!

Why Do Cultures Differ in SWB?

Gross Domestic Product

Part of the differences in SWB across countries can be explained by the differences in their GDP. Across multiple studies, the wealth of nations is one of the strongest and most reliable predictors of national levels of life satisfaction (Diener and Suh, 1999). In other words, countries with the highest SWB levels also tend to be wealthy (e.g., Switzerland and Denmark), while countries with the lowest SWB levels tend to be the underdeveloped nations (e.g., Bulgaria and Tanzania).

However, while GDP is a predictor of SWB across countries, Myers (2000) and Diener and Oishi (2000) have also found that mean happiness levels have not been increasing dramatically over time, although income has. Known as the 'Easterlin Paradox,' this is likely due to the rising expectations associated with the rise in income, where people now have higher economic desires that negate the gains in their happiness levels.

Although countries with higher GDP experience higher SWB, the association between GDP and SWB is not perfectly linear and there are some countries that report unexpectedly higher or lower life satisfaction levels. For example, the high wealth of the four Asian tigers, Hong Kong, Singapore, South Korea, and Taiwan, should predict relatively high levels of SWB in those countries. Yet in many of the happiness surveys conducted worldwide, citizens from these four countries typically do not report experiencing high levels of SWB. In fact, the recent Gallup World survey reported that Singaporeans were the least likely in the world to report experiencing emotions of any kind on a daily basis (Clifton, 2012). Meanwhile, the low wealth of the Latin American countries should predict lower SWB than is observed, but countries like Chile and Brazil consistently report relatively high levels of SWB (Helliwell et al., 2013), suggesting that GDP alone is insufficient in providing us with a complete explanation of the cultural differences in SWB among countries.

Individualism-Collectivism

Across multiple studies, researchers have found that individualism correlates with higher SWB (Diener et al., 1995), indicating that individualistic cultures experience happiness more frequently and more intensely than collectivistic cultures (Eid and Diener, 2001). Countries high in individualism, like the United States and Canada, tend to report higher SWB levels compared to collectivistic cultures like China and Japan. At the individual level, individualists, compared to collectivists, also tend to report higher happiness levels, and this correlation holds even if the individualist is in a collectivistic society like Asia. For example, an American living in Japan is more likely to report higher happiness levels compared to a Japanese person living in the same country (Diener, 2000). It may be that collectivist values are at times incompatible with personal happiness, for example, when an individual chooses to sacrifice his/her own personal happiness in order to maintain group harmony.

Although the relationship between individualism and happiness is fairly strong, there are however exceptions to this trend where individualism correlated with higher SWB. Across multiple studies, researchers have found that Asia and Latin America are similar in their levels of collectivism. Thus, it follows that they should report similar happiness levels, and yet Latin Americans have consistently outscored Asians in well-being surveys, including the recent World Happiness Report 2013, puzzling researchers.

Norms Regarding Happiness and Other Positive Emotions

Diener et al. (2000) examined norms toward happiness by asking participants in 40 different societies to rate the ideal level of life satisfaction an ideal person would experience.

Interestingly, Diener found a good deal of variation in responses to this question around the world. To the surprise of Americans, not everybody said that the highest levels of happiness were the most ideal. Although all respondents agreed that the ideal level of happiness was above neutral, some respondents were much closer to neutral while others were much closer to the top of the scale. Ideal levels of SWB also correlated with actual levels of SWB and corresponded to a nation's general positivity bias as well. Overall, Asian countries displayed little positivity bias and lower ideal SWB scores, while Latin American countries displayed the most positivity bias and the highest ideal SWB scores. North American and European countries were in between.

Why is there less of an inclination toward positivity in Asian cultures? And why do Asians report lower levels of ideal SWB than say Latin Americans and North Americans? The answer may lie in the value that negative emotions have in Asian societies. Feelings such as guilt, although unquestionably unpleasant, can serve to enhance group harmony. On the other hand, although pride may feel good, pride about the individual's goal accomplishments may set the individual apart from group members. Eid and Diener (2000) found that respondents in North America and Australia overwhelmingly preferred positive emotions and devalued negative ones, whereas in Asia, respondents were more balanced in their views, noting that even unpleasant feelings have value.

Different Definitions of Happiness

Societies like the United States recognize happiness as a high-arousal positive state, where happiness is an upbeat and energetic emotion, featuring intense emotions like excitement and exhilaration while cultures like Japan tend to define happiness as a low-arousal positive state, which involves being calm, relaxed, and in a peaceful state of mind, with everything in harmony (Tsai et al., 2006). Consistent with self-reports of these emotional states, Tsai et al. (2007) examined best-selling children storybooks in the United States and Taiwan and found that American storybooks had more excited expressions, wider and open-mouth smiles, and more arousing activities. On the other hand, happiness in Taiwanese storybooks was often depicted with closed-mouth smiles and closed eyes.

Different Strategies in Pursuing Happiness

Cultures could also differ in happiness due to the strategies they use to pursue happiness. Individualistic cultures adopt a promotion-focused orientation toward achieving a goal, focusing more strongly on information that is relevant to attaining their desires and they also place more emphasis on positive than on negative information regarding themselves (Lee et al., 2000). Thus, individualists will actively seek out situations that could maximize their well-being. For example, they may seek out happiness-inducing activities like dinner with friends or exciting activities like bungee jumping, allowing them to experience more happiness.

A collectivistic culture, however, adopts a prevention-focused orientation toward goal achievement, and more attention is placed on information that prevents them from violating social norms and also emphasizes negative information about themselves (Lee et al., 2000). Thus collectivists might actively avoid situations that could negatively impact their happiness levels as they might have the mindset that "I will achieve happiness if nothing bad happen to me throughout the day." Therefore, they would try to avoid conflict with people in an attempt to protect themselves from unhappiness.

Cultural Differences in Materialism

Materialism, which is the tendency to believe that consumer goods and services provide the greatest source of satisfaction in life (Belk, 1985), could also explain some of the cultural differences in happiness. Research on materialism and happiness has consistently shown that materialistic people are less happy. To the extent that cultures differ in their focus on materialism, we might expect there to be differences in SWB.

Wirtz and Scollon (2012) and others have shown that, indeed, materialism differs among cultures. In their study, they found that Asians and Americans both strongly valued happiness and meaning in life when judging life quality. However, Americans perceived life desirability to be less dependent on material success than on the presence of happiness and meaning in life. Asians, on the other hand, were more likely to view a life characterized by high material success as more favorable, suggesting that materialism may be more congruent in Asian societies. Wirtz and Scollon (2012) further demonstrated that priming a collectivist perspective enhanced the importance of wealth in judgments of a good life.

Measurement

Are there cultural differences in response styles that could account for the cultural differences in SWB? One argument that has been made is that Asians tend to avoid providing extreme values in a survey, preferring to use the middle of scales more. However, Diener et al. (1995) tested this hypothesis and found little support for it, as they found that compared to Americans, Asians reported a greater range of moods, indicating that Asians were not adverse to reporting extreme moods.

Still, there is still good evidence that how SWB is measured does contribute to cultural differences in SWB in several systematic ways (Scollon et al., 2011). First, cultural differences are often more pronounced with positive emotions. For example, Scollon et al. (2004) compared momentary reports of positive and negative effects provided by European Americans, Asian Americans, Hispanic Americans, Japanese, and Indian students and found greater cross-cultural differences in positive compared to negative emotions. Why are positive emotions more susceptible to cultural influences than negative emotions? Researchers believe that positive emotions may be subjected to greater social learning influences than negative emotions that have stronger genetic and biological constraints. Tellegen et al. (1988), for example, found that there were stronger genetic effects for negative emotions between twins. For positive emotions, common environmental experiences such as parenting and household environment had a stronger impact. This suggests that the learning of positive emotions might require greater social interactions and reinforcement. Considering that positive emotions are valued differently

amongst different cultures, it is likely that this reinforcement of positive emotions would also differ across cultures.

Second, larger cultural differences in SWB have also been found as the time frame of measurement is increased. For example, when considering how satisfied you were with your life in general over the past year, compared to your life satisfaction yesterday or in the past hour, researchers tend to find greater cultural differences when participants consider their satisfaction over the past year. Furthermore, Oishi (2002) found that there were differences in how cultural groups remember their experiences because while Asian Americans and European Americans did not differ in their daily self-report of emotions, European Americans, compared to Asian Americans, recalled a higher degree of well-being.

Cultural differences are larger for such retrospective reports because such reports are not exact duplicates or summations of momentary experience. Instead, these memories are reconstructed (Robinson and Clore, 2002). When evaluating their experiences, people tend to first consider their event-specific information, piecing together every detail that they can remember about the event, before concluding with an overall judgment of their experience. For example, when asked to recount whether lunch was an enjoyable experience, you may seek to systematically recollect the various details of lunch, the food you ate, and the company you were with. Once you have established all the facts, you then cognitively appraise all these details to conclude whether lunch was enjoyable or not.

However, as the time frame increases, it becomes increasingly hard for people to integrate all aspects of the experience together. Imagine trying to recount how pleasant your overall lunch experience was over the past year. It is almost impossible for us to remember all the lunches we had over the past month, much less a year. People then begin to use their nonevent-specific information, which includes their cultural values and beliefs, to fill in these memory gaps. Using such nonevent-specific information could thus result in differences to their retrospective evaluations as cultures hold different views about the desirability of particular emotions such as happiness (Scollon et al., 2011).

Future Research and Conclusions

While research in SWB and culture has explained some of the discrepancies between countries, many questions still remain and new findings warrant further investigation. Although there are some impressive worldwide data sets, most of the studies in the field compare two countries on a limited set of SWB variables at a single point in time. More comprehensive and longitudinal studies, and more importantly, interdisciplinary approaches are sorely needed. Although psychologists and economists have made important discoveries in SWB thus far, more progress will be made at the intersection of these fields and others such as policy studies, genetics, sociology, and environmental studies. As policy makers grow increasingly aware and interested in national indicators of SWB, the data will become available in time to examine the impact of social policies on SWB. Cultural explanations for cultural differences will need to be further unpacked to examine the source and development of cultural beliefs; thus, the field could benefit from collaborations with anthropologists and ethnographers as well as those who study consumer behavior. The possibilities are limitless.

One example of the exciting developments in interdisciplinary approaches to understanding cultural differences in SWB for this area combines behavioral genetics, epidemiology, and well-being research. In the first study of its kind, De Neve (2011) found support for a direct link from genes to happiness. Specifically, people with a certain variant (long alleles) of the serotonin transporter gene (5HTTLPR) tend to report higher satisfaction. Meanwhile, people with the short alleles of 5HTTLPR tend to report lower life satisfaction. Interestingly, the distribution of the polymorphisms of 5HTTLPR appears not to be the same across different ethnic groups and different regions of the world, which means that cross-cultural comparisons on SWB have been confounded with differences in genetic variation. Compared to North American and European countries, Asian countries have more of the short alleles represented in their populations, which could possibly account for the lower levels of SWB in those parts of the world as suggested by Proto and Oswald (2013) and Scollon et al. (2011). Clearly, however, more research is needed.

In summary, people around the world want to be happy. Yet, scores of happiness vary widely across cultures. Past research on culture and SWB has revealed that cultural differences in SWB can be explained in part by societal wealth, individualism, materialism, cultural norms regarding positive emotions, and possibly genetic differences. Future research on culture and SWB remains an exciting area with opportunities for interdisciplinary explorations.

See also: Collectivism and Individualism: Cultural and Psychological Concerns; Cross-Cultural Psychology; Cultural Psychology; Culture and Emotion; Culture and the Self: Implications for Psychological Theory; Subjective Wellbeing, Psychology of.

Bibliography

Belk, R.W., 1985. Materialism: trait aspects of living in the material world. Journal of Consumer Research 12, 265–280.

Chiao, J., Blizinsky, K., 2010. Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene. Proceedings of the Royal Society B-Biological Sciences 277 (1681), 529–537.

Chiu, C.-Y., Leung, K.-Y., Hong, Y.-Y., 2010. Cultural processes: an overview. In: Leung, A.K.-Y., Chiu, C.-Y., Hong, Y.-Y. (Eds.), Cultural Processes: A Social Psychological Perspective. Cambridge University Press, New York.

Clifton, J., 2012. Singapore Ranks as Least Emotional Country in the World. Retrieved from http://www.gallup.com/poll/158882/singapore-ranks-least-emotional-country-world.aspx.

De Neve, J., 2011. Functional polymorphism (5-HTTLPR) in the serotonin transporter gene is associated with subjective well-being: evidence from a US nationally representative sample. Journal of Human Genetics 56 (6), 456–459. http://dx.doi.org/10.1038/jhg.2011.39.

Diener, E., 1984. Subjective well-being. Psychological Bulletin 95, 542–575.

Diener, E., 2000. Subjective well-being: the science of happiness and a proposal for a national index. American Psychologist 55 (1), 34–43. http://dx.doi.org/10.1037/0003-066X.55.1.34.

Diener, E., Diener, M., 1995. Cross-cultural correlates of life satisfaction and self-esteem. Journal of Personality and Social Psychology 68 (4), 653–663. http://dx.doi.org/10.1037/0022-3514.68.4.653.

Diener, E., Diener, R.B., 2008. Happiness: Unlocking the Mysteries of Psychological Wealth. Blackwell Publishing Ltd.

Diener, E., Suh, E.M., 1999. National differences in subjective well-being. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well-being: the foundations of hedonic psychology. Russell Sage Foundation, New York, pp. 434–452.

Diener, E., Diener, M., Diener, C., 1995. Factors predicting the subjective well-being of nations. Journal of Personality and Social Psychology 69 (5), 851–864. http://dx.doi.org/10.1037/0022-3514.69.5.851.

Diener, E., Oishi, S., 2000. Money and happiness: Income and subjective well-being across nations. In: Diener, E., Sub, E. (Eds.), Culture and Subjective Well-Being across Nations. MIT Press, Cambridge, MA, pp. 185–218.

Diener, E., Scollon, C.K., Oishi, S., Dzokoto, V., Suh, E., 2000. Positivity and the construction of life satisfaction judgments: global happiness is not the sum of its parts. Journal of Happiness Studies 1 (2), 159–176.

Eid, M., Diener, E., 2001. Norms for experiencing emotions in different cultures: inter- and intranational differences. Journal of Personality and Social Psychology 81 (5), 869–885. http://dx.doi.org/10.1037/0022-3514.81.5.869.

Helliwell, J., Layard, R., Sach, J., United Nations, 2013. World Happiness Report. Retrieved from: http://unsdsn.org/resources/publications/world-happiness-report-2013/.

Hofstede, G., 1980. Culture's consequences. Sage, Beverly Hills, CA.

Lee, A.Y., Aaker, J.L., Gardner, W.L., 2000. The pleasures and pains of distinct self-construals: the role of interdependence in regulatory focus. Journal of Personality and Social Psychology 78, 1122–1134.

Myers, D.G., 2000. The funds, friends, and faith of happy people. American Psychologist 55, 56–67.

Oishi, S., 2002. The experiencing and remembering of well-being: A cross-cultural analysis. Personality And Social Psychology Bulletin 28 (10), 1398–1406. http://dx.doi.org/10.1177/014616702236871.

Rice, T., Steele, B., 2004. Subjective well-being and culture across time and space. Journal of Cross-Cultural Psychology 35 (6), 633–647.

Scollon, C.N., Diener, E., Oishi, S., Biswas-Diener, R., 2004. Emotions across cultures and methods. Journal of Cross-Cultural Psychology 35, 304–326.

Robinson, M.D., Clore, G.L., 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychological Bulletin 128 (6), 934–960. http://dx.doi.org/10.1037/0033-2909.128.6.934.

Scollon, C., Koh, S., Au, E.M., 2011. Cultural differences in the subjective experience of emotion: when and why they occur. Social & Personality Psychology Compass 5 (11), 853–864. http://dx.doi.org/10.1111/j.1751-9004.2011.00391.x.

Tellegen, A., Lykken, D.T., Bouchard Jr., T.J., Wilcox, K.J., Segal, N.L., Rich, S., 1988. Personality similarity in twins reared apart and together. Journal of Personality and Social Psychology 54, 1031–1039.

Triandis, H.C., 1989. The self and social behavior in differing cultural contexts. Psychological Review 96, 506–505.

Tsai, J., Louie, J., Chen, E., Uchida, Y., 2007. Learning what feelings to desire: socialization of ideal affect through children's storybooks. Personality and Social Psychology Bulletin 33 (1), 17–30.

Tsai, J.L., Knutson, B., Fung, H.H., 2006. Cultural variation in affect valuation. Journal of Personality and Social Psychology 90 (2), 288–307. http://dx.doi.org/10.1037/0022-3514.90.2.288.

Uchida, Y., Kitayama, S., 2009. Happiness and unhappiness in east and west: themes and variations. Emotion 9, 441–456.

Wirtz, D., Scollon, C., 2012. Culture, visual perspective, and the effect of material success on perceived life quality. Journal of Cross-Cultural Psychology 43 (3), 367–372.

Subjective Wellbeing, Psychology of

Christina N Armenta, Peter M Ruberton, and Sonja Lyubomirsky, University of California Riverside, Riverside, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Diener, volume 24, pp. 16451–16454, © 2001, Elsevier Ltd.

Abstract

Although people have been intrigued by happiness for centuries, subjective well-being became a thriving topic of scientific inquiry only relatively recently, with the creation of valid and reliable well-being measures. Growing evidence suggests that happiness is associated with success in multiple domains, such as relationships, work, and health. However, objective life circumstances do not impact well-being as much as people believe, in part because of hedonic adaptation – one of the biggest obstacles in the pursuit of happiness. Fortunately, researchers have found that positive activities, like expressing gratitude or doing acts of kindness, can have a significant impact on well-being.

Historical Background

Although people have contemplated what makes life meaningful and worthwhile for millennia (Keyes, 2006), happiness did not become a subject of scientific inquiry until relatively recently. Prior to the 1970s, researchers focused on topics such as obedience, conformity, and aggression in an attempt to understand the devastation of World War II. As the world began to recover from the effects of wartime, psychological scientists became interested in monitoring and improving social change. This interest led to a newfound appreciation of the individual, as many believed that the key to understanding social change could be found in the study of the individual's attitudes, expectations, and values. The outcome was an increased focus on the importance of an individual's sense of meaning and views on life. Measures of quality of life were developed to monitor social change and influence social policy, and subjective well-being began to emerge as a field of scientific study.

Definition and Measurement

Today, researchers generally agree that subjective well-being is a broad construct that refers to evaluations of the quality of one's life and includes both an affective and a cognitive component (Diener et al., 1999). The affective component refers to the frequency of experienced emotions, such that individuals with high levels of subjective well-being report high levels of positive affect (i.e., many pleasant emotions) and low levels of negative affect (i.e., few unpleasant emotions). The cognitive component comprises overall life satisfaction, as well as a person's evaluations of multiple life domains, such as work, health, and relationships. Importantly, because respondents determine whether their own lives are worthwhile and meaningful, the concept of well-being is inherently subjective and democratic. Hence, researchers commonly use the term subjective well-being (SWB), as well as happiness, and these terms are used synonymously throughout this article.

Since the field's beginnings, researchers have relied almost exclusively on self-report measures of happiness. As such, multiple self-report scales of well-being are now widely used. Investigators initially developed simple measures prompting participants to make general evaluations about their lives (Krueger and Schkade, 2008). Specifically, participants are asked a single general question about their overall happiness or satisfaction with life. Such global measures are still most popular, having been incorporated into large-scale surveys, such as the Gallup World Poll and World Values Survey. Indeed, global scales have been invaluable for gaining insight on well-being around the world and have allowed for comparisons in happiness and satisfaction in large representative samples across cultures and across decades (Helliwell et al., 2013).

Although single-item measures of well-being are simple and easy to administer, it is difficult to capture how happy a person is with a single question. To address this limitation, global, multiitem measures of well-being were created and validated. Today, a number of validated self-report measures of happiness, affect, and life satisfaction are in common usage, including the Satisfaction with Life Scale (Diener et al., 1985), Positive and Negative Affect Scale (Watson et al., 1988), and Subjective Happiness Scale (Lyubomirsky and Lepper, 1999).

Other measures of well-being are intended to capture its affective component. For example, with the Experience Sampling Method, participants are prompted to report their current emotions and thoughts at random intervals throughout the day (see Csikszentmihalyi and Larson, 1987). This method, however, can be difficult to administer because it requires a large time commitment from participants. The Day Reconstruction Method (DRM) is another popular approach to measuring well-being. Participants are asked to consider the preceding day, break it up by time into different episodes, describe each episode, and report their well-being during that part of the day (Kahneman et al., 2004). Given that people in a happy mood are relatively more likely to recall positive information (Schwarz and Clore, 2003), with their current mood coloring their interpretations of the past, memory biases may be minimized with the DRM because participants are pressed to recall what they were feeling during a specific time (e.g., while commuting to work) after describing the episodes of the previous day.

Self-report measures of well-being have their limitations. Fortunately, researchers have found that self-report ratings of

happiness are reliable over time (Diener, 1994), correlate with theoretically relevant constructs (Larsen et al., 1985), and converge with non-self-report measures, such as reports of family and friends (Sandvik et al., 1993). With the creation of valid and reliable measures of well-being, research on happiness has grown rapidly and developed into a thriving new field. The ability to assess well-being has opened the door for researchers to investigate numerous previously elusive questions, such as what are the determinants and correlates of happiness, what are its benefits or costs, and, perhaps most important, whether and how happiness can be increased.

Benefits of Subjective Well-Being

Happiness involves much more than simply an emotion that feels good. The experience of positive affect allows happy people to build social, intellectual, physical, and psychological resources (Fredrickson, 2001). It may also signal to the individual that everything is going well, thus allowing him or her to be more creative, productive, approach oriented, and social. Accordingly, happiness leads to a variety of beneficial outcomes via an increase in behaviors that offer individuals the opportunity to achieve success in multiple domains (Lyubomirsky et al., 2005a). For example, happy individuals have a relatively greater tendency to seek out relationships, exhibit helping behavior, be productive at work, and cope successfully with life changes. As such, happiness has been found to benefit one's interpersonal relationships, work life, and health.

Social Relationships

Happy individuals tend to have successful social relationships. For example, happier people are relatively more likely to seek out both formal and informal social activities (Okun et al., 1984), and, thus, engage in frequent social interactions. This high level of social engagement may allow them to build stronger, deeper bonds with others. It is therefore not surprising that happy people tend to report having more friends and social support than their less happy peers (Baldassare et al., 1984; Mishra, 1992). Happier individuals are also relatively more likely to get married and to report higher marital satisfaction (Harker and Keltner, 2001; Headey and Veenhoven, 1989; Marks and Fleming, 1999). In turn, having a solid social network may make people relatively better equipped to handle negative life events. Belonging to a social group promotes better mental and physical health, granting people the resources to manage life changes and cope with illnesses (Jetten et al., 2012).

Work

A large portion of an individual's time and energy is spent at work or pursuing work-related activities. Indeed, work is a vital source of income and self-esteem for many people. Importantly, happier individuals are relatively more likely to have successful work lives and tend to be more satisfied with their jobs. For example, happy people are more likely to obtain a job, have high job satisfaction, and feel financially independent at an young age than their less happy counterparts (Roberts et al., 2003). Furthermore, supervisors tend to rate happy workers relatively higher on productivity, work quality, creativity, and dependability (Cropanzano and Wright, 1999). Happiness has also been found to predict increases in income and lower rates of unemployment over time (Graham et al., 2004; Marks and Fleming, 1999).

Physical Health

Subjective well-being is associated with a number of positive health outcomes. For example, happy people report having better health, missing less work due to illness, and making fewer visits to the hospital than less happy people (Graham et al., 2004). One possible reason for superior physical health may be the association between happiness and strong immune functioning. For example, one study found that healthy individuals with a positive emotional style were relatively less likely to get a cold after exposure to a rhinovirus (Cohen et al., 2003). Notably, happiness is even related to longevity (Chida and Steptoe, 2008). Happy people are relatively less likely to experience be diagnosed with cardiovascular disease (Davidson et al., 2010) or suffer fatal accidents (Kirkcaldy and Furnham, 2000), and they are more likely to survive illnesses like HIV (Ickovics et al., 2006).

The Role of Genetics in Well-Being

Researchers have found that well-being is moderately stable (Headey and Wearing, 1989; Lykken and Tellegen, 1996), such that, for example, happier individuals tend to remain consistently happy over time. One of the reasons for this stability is that people show shifts in well-being after positive and negative events, but, as described in greater detail below, gradually return to their previous happiness levels over time. A growing literature has accumulated evidence that this baseline level of happiness is in part influenced by genetics (Bartels and Boomsma, 2009; Lykken and Tellegen, 1996; Rietveld et al., 2013). The genetic component of well-being is investigated by comparing the correlations between the reported SWB of monozygotic (identical) twins to that of dizygotic (fraternal) twins. Identical twins have been found to report more similar levels of happiness than fraternal twins, even if they were raised in different households (Bartels and Boomsma, 2009; Lykken and Tellegen, 1996). For example, one study found that up to 80% of the stable component of happiness is heritable, with nonshared environmental factors accounting for the remaining variance (Lykken and Tellegen, 1996). Although studies have found varying heritability coefficients (ranging from 0.25 to 0.55), the overwhelming conclusion of this research is that happiness has a large genetic component.

Related work suggests that personality may underlie the genetic influence on happiness. Subjective well-being is highly positively correlated with extraversion and negatively correlated with neuroticism (Costa and McCrae, 1980; DeNeve, 1999; Furnham and Brewin, 1990; Hayes and Joseph, 2003), with mixed evidence on its relationship with other personality factors, such as agreeableness and conscientiousness. Because, much like SWB, personality has a genetic component,

a relationship between the genetic influence on personality and the genetics of happiness may exist. One study found evidence to suggest that the genetic influence on extraversion, neuroticism, and conscientiousness entirely explains the heritability of SWB (Weiss et al., 2008).

Correlates of Subjective Well-Being

The nonheritable correlates of subjective well-being can be broadly divided into two categories: objective environmental circumstances (such as income or marital status) and intentional cognitions and behaviors (such as prosocial behavior; Lyubomirsky, 2001). Laypeople are often surprised to learn that objective circumstances or demographic variables are weaker predictors of well-being than volitional behaviors. For example, men and women in nations across the world report being happy and satisfied with their lives with equal frequency (Inglehart, 1990), although the two sexes tend to be unhappy in different ways (Myers, 2000). Furthermore, the well-being gains or losses from major life events, such as winning the lottery or sustaining a debilitating injury, tend to be short-lived (Clark et al., 2008). Nevertheless, despite the finding that life circumstances account for only a small portion of individual differences in happiness (Diener et al., 1999), research on the objective correlates of happiness has revealed a number of important links between subjective well-being and such factors as wealth, spirituality, and social relationships.

As it is with most life circumstances, the relationship between wealth and well-being is nuanced and often unexpectedly small. Contrary to popular perceptions that more money is the key to greater happiness, more money appears to be more important for avoiding unhappiness: At income levels above those needed to fulfill basic sustenance needs, increasing income is not strongly associated with more positive feelings (Diener and Biswas-Diener, 2002; Myers, 2000). At the global level, income is associated with well-being far more strongly between nations than within them: People in wealthier nations tend to be happier, on average, than people in less developed countries, but the wealthiest people in a nation are not necessarily happier than their poorer compatriots (see Diener and Biswas-Diener, 2002). Furthermore, wealth is more strongly associated with certain kinds of well-being than with others: Both personal and national income are highly correlated with global life evaluations, but only modestly so with positive feelings (Kahneman and Deaton, 2010). Challenging these findings, however, recent research has found that increases in household income, but not national income (i.e., gross domestic product), are associated with sustained increases in subjective well-being (Diener et al., 2013).

Despite the weak correlation between general life circumstances and well-being, two objective psychosocial variables have consistently been found to be important to happiness. First, as mentioned earlier, people with strong, close social relationships tend to have high levels of happiness and psychological health (e.g., resilience to stress) (Myers, 2000). For example, married individuals in 16 nations reported higher levels of well-being than nonmarried individuals (Stack and Eshleman, 1998), although the happiness boost from getting married does not often persist past the first few years of marriage (Lucas et al., 2003; Lucas and Clark, 2006). Second, when controlling for other life circumstances, religious people tend to be slightly happier than nonreligious people, a relationship that may be due to greater social support and feelings of purpose from religious activity (Diener et al., 2011; Ellison and Levin, 1998).

Hedonic Adaptation as a Barrier to Well-Being

The relatively small impact of life circumstances on well-being may be due to the phenomenon of hedonic adaptation. People experience boosts in well-being following positive life events and declines after negative ones, but, interestingly, these shifts in well-being do not last. A growing literature has shown that individuals become accustomed to changes in their lives via hedonic adaptation (Frederick and Loewenstein, 1999). How rapid and how complete this adaptation is differs for negative versus positive experiences. To begin, longitudinal panel studies suggest that individuals have a remarkable capacity to adapt to traumatic life events, such as disability, widowhood, and unemployment (Clark and Georgellis, 2012; Lucas, 2005, 2007), although the amount of adaptation depends on the individual and the type of event. Some studies have found that individuals only partially adapt to events such as divorce (Lucas, 2005) and widowhood (Lucas et al., 2003), whereas others have found evidence for complete adaptation following these adverse negative events (Clark et al., 2008).

By contrast, the evidence is fairly consistent that people, on average, adapt completely to major positive life changes such as getting married, acquiring a new job, and even winning the lottery (Boswell et al., 2005; Brickman et al., 1978; Clark and Georgellis, 2012). This tendency to adapt to positive events explains in part why changing one's circumstances (e.g., receiving a raise) contributes little to a person's well-being. Although adaptation to many circumstances may be evolutionarily adaptive, becoming accustomed to positive changes in one's life is an obstacle in the pursuit of happiness (Lyubomirsky, 2011).

Increasing Subjective Well-Being

The fact that happiness is heritable and impacted only minimally by life circumstances does not mean that it cannot be intentionally and effortfully increased (Lykken, 1999; Lyubomirsky et al., 2005b). By observing the behaviors that happy people naturally engage in, researchers have found that happy individuals are not passive actors in their lives. Relative to their less happy peers, they seek out rewarding situations (Luhmann et al., 2013), interpret ambiguous events positively, and perceive the silver lining and reappraise negative situations (Lyubomirsky, 2001). This research suggests that happiness is not necessarily static, and that the activities individuals engage in are likely to influence their well-being. If naturally happy people perform particular behaviors, then deliberately choosing to perform those behaviors may increase well-being in less happy people. As such,

observations of happy people have steered researchers to specific ways to increase happiness and informed the designs of future interventions.

Interventions

Happiness is highly valued in cultures around the world (Diener, 2000) and is a subject of great interest to the general public. Fortunately, in recent years, growing work has revealed multiple ways that ordinary people can boost happiness in their daily lives. Specifically, researchers have designed positive activity interventions, which prompt participants to engage in brief, simple, self-administered activities, as an approach to increase well-being and uncover the mechanisms underlying their success (e.g., motivation, effort, and dosage). Many of these activities, such as doing acts of kindness, predicting a positive future, and expressing gratitude, are modeled after the behaviors that naturally happy people do. Notably, a meta-analysis of 51 randomized controlled experimental interventions has found support for the efficacy of positive activities both to increase well-being and to decrease depressive symptoms (Sin and Lyubomirsky, 2009). These efficacious intentional activities include writing letters of gratitude (Boehm et al., 2011; Lyubomirsky et al., 2011; Seligman et al., 2005), counting one's blessings (Emmons and McCullough, 2003; Froh et al., 2008), performing acts of kindness (Dunn et al., 2008; Layous et al., 2012; Sheldon et al., 2012), and practicing optimism (Boehm et al., 2011; Lyubomirsky et al., 2011).

Expressing Gratitude

Gratitude is the felt sense of thankfulness and appreciation toward an individual who has done something above and beyond what is expected or required (Emmons, 2008). Researchers typically induce gratitude in participants by prompting them to count their blessings (i.e., list things for which they are grateful) or write gratitude letters to important people in their lives. One random controlled intervention found that participants who counted their blessings felt better about their lives as a whole, were more optimistic about the coming week, complained less, and exercised more than did controls (Emmons and McCullough, 2003). Expressing gratitude was also found to reduce negative affect and increase positive affect. Another study randomly assigned participants over the course of 6 weeks either to write letters of gratitude to friends or family members who had done something for which they are grateful or to write about neutral weekly experiences (Boehm et al., 2011). Participants who expressed gratitude experienced relatively greater increases in satisfaction with life. The expression of gratitude may allow people to savor the positive experiences and situations in their lives and thus extract the maximum amount of satisfaction from them. Furthermore, gratitude may lead an individual to strengthen social bonds and friendships with others (Algoe et al., 2010) and to be inspired to pay back their benefactors by being a better person.

Performing Acts of Kindness

Naturally happy people report performing more altruistic acts, spending more time helping others, and helping their colleagues at work (Lyubomirsky et al., 2005a). Kind acts are often described as behaviors that are not expected or required and involve some sacrifice in effort, time, and energy (Sheldon et al., 2012). A study by Dunn et al. (2008) found that participants randomly assigned to perform a kindness for another person (i.e., spend money on him or her) reported greater happiness than those assigned to do a kind act (i.e., spend money) for themselves. One possible account of this finding is that kindness tends to increase one's evaluation of oneself as a good person (Williamson and Clark, 1989). For example, after performing acts of kindness for another person, an individual may begin to view himself or herself as a compassionate, helpful person and thus increase in feelings of confidence and efficacy. Furthermore, kind acts may foster social rewards. For example, in a 4-week intervention, 9- to 11-year-old children were randomly assigned to perform three acts of kindness or to visit three places (the comparison group; Layous et al., 2012). Participants also listed classmates with whom they would like to spend time. Children who performed three acts of kindness were significantly more likely to be socially accepted by their peers. This finding suggests that people who perform acts of kindness may not only become happier but better liked by others.

Visualizing Best Possible Selves

Happier individuals are more optimistic and view their lives in more positive ways (DeNeve, 1999; Lyubomirsky, 2001). Furthermore, research has found that individuals prompted to practice optimism show increases in well-being. In one study, participants were instructed to write about their best possible selves – an activity that compels optimism about the future – and complete measures of SWB every day for 4 days (King, 2001). Cultivating optimism led to an increase in happiness that persisted at least until the 3-week follow-up. In another study, participants wrote narratives about their best possible future self with regards to family, friends, hobbies, career, romantic partner, and health – a different domain each week (Boehm et al., 2011). Those prompted to express optimism experienced relatively higher satisfaction in life across time. This activity may lead to greater happiness because it allows people to learn about themselves, restructure their priorities, and gain insight into their motives and desires.

Current and Future Directions

Over the past several decades, researchers have developed and tested a variety of positive activities that happiness seekers can practice. However, much of the research conducted thus far has relied on self-report measures of well-being. This approach can be problematic because self-report measures are subject to social desirability and memory biases (Diener et al., 2009). To this end, investigators are beginning to use more objective and unobtrusive measures of well-being. For example, Chancellor et al. (in press) conducted a study using technology that assessed participants' behavioral rhythm throughout the day over 6 weeks. In this study, behavioral data were collected using sociometric badges that monitor how an individual's body oscillates in three-dimensional space. These badges are sensitive enough to differentiate between a range of movements, including sitting, walking, and talking. The researchers found that participants who wrote down "three things that went well at

work" each week came to the office with more energy and engaged in less office chitchat than those who wrote down three tasks that they accomplished (Chancellor et al., in press).

The rapid growth in technology in recent years has revealed a largely untapped potential for new methods of measuring well-being. For example, the popularity of Facebook™ and Twitter™ gives researchers the opportunity to collect massive data sets from large diverse samples by coding updates for such variables as positive language, emoticons, and humor (see Young et al., 2009; Hughes et al., 2012). New technological advances have also led to innovative devices, such as the Electronic Activated Recorder, that allow researchers to obtain snippets of conversations throughout the day. These recordings can be coded for humor, laughter, or positive words to achieve a clearer and more precise portrait of naturalistic fluctuations in well-being – and its antecedents, correlates, and consequences – over the course of the day (Mehl et al., 2010).

Although subjective well-being is highly valued around the world, a great deal of cross-cultural variation has been observed (Diener, 2000). Some research suggests that carrying out specific positive activities may not be as accepted or endorsed in non-Western cultures. For example, one study found evidence to suggest that Asian-Americans benefit less from practicing gratitude and optimism (Boehm et al., 2011). In another set of studies, U.S. participants, but not S. Korean ones, reported increased well-being after writing gratitude letters (Layous et al., 2013). One possibility is that expressions of gratitude may elicit feelings of guilt or indebtedness in members of Eastern nations. Cultures vary on numerous dimensions, such as individualism/collectivism, power distance, hierarchy, and strength of the economy. Future research could inform positive interventions by considering how such cultural dimensions predict the extent to which individuals might benefit – or even be harmed – by positive activity interventions.

The scientific study of happiness has come a long way since Aristotle proclaimed that it is the meaning and the purpose of life, the whole aim and end of human existence. In the past several decades, researchers have developed valid and reliable measures to assess well-being, explored its correlates, determinants, and consequences, and revealed the ways and mechanisms by which it can be increased. Policy makers around the world have become concerned with the well-being of their citizens, and companies have become invested in implementing programs to boost the morale and well-being of their employees. As interest in happiness continues to spread, research on well-being will begin to inform social policies in countries around the world. Therefore, although research on well-being has a short past, it has a lengthy future.

See also: Happiness and Work; Social Psychological Theory, History of; Subjective Wellbeing and Culture.

Bibliography

Algoe, S.B., Gable, S.L., Maisel, N.C., 2010. It's the little things: everyday gratitude as a booster shot for romantic relationships. Personal Relationships 17, 217–233.

Baldassare, M., Rosenfield, S., Rook, K., 1984. The types of social relations predicting elderly well-being. Research on Aging 6, 549–559.

Bartels, M., Boomsma, D.I., 2009. Born to be happy? The etiology of subjective well-being. Behavior Genetics 39, 605–615.

Boehm, J.K., Lyubomirsky, S., Sheldon, K.M., 2011. A longitudinal experimental study comparing the effectiveness of happiness-enhancing strategies in Anglo Americans and Asian Americans. Cognition & Emotion 25, 1263–1272.

Boswell, W.R., Boudreau, J.W., Tichy, J., 2005. The relationship between employee job change and job satisfaction: the honeymoon-hangover effect. Journal of Applied Psychology 90, 882–892.

Brickman, P., Coates, D., Janoff-Bulman, R., 1978. Lottery winners and accident victims: is happiness relative? Journal of Personality and Social Psychology 36, 917–927.

Chancellor, J., Layous, K., Lyubomirsky, S. Recalling positive events at work makes employees feel happier, move more, but interact less: a 6-week randomized controlled intervention at a Japanese workplace. Journal of Happiness Studies, in press.

Chida, Y., Steptoe, A., 2008. Positive psychological well-being and mortality: a quantitative review of prospective observational studies. Psychosomatic Medicine 70 (7), 741–756.

Clark, A.E., Diener, E., Georgellis, Y., Lucas, R.E., 2008. Lags and leads in life satisfaction: a test of the baseline hypothesis. The Economic Journal 118, F222–F243.

Clark, A.E., Georgellis, Y., 2012. Back to baseline in Britain: adaptation in the British household panel survey. Economica 80, 496–512.

Cohen, S., Doyle, W.J., Turner, R.B., Alper, C.M., Skoner, D.P., 2003. Emotional style and susceptibility to the common cold. Psychosomatic Medicine 65, 652–657.

Costa, P.T., McCrae, R.R., 1980. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. Journal of Personality and Social Psychology 38, 668–678.

Cropanzano, R., Wright, T.A., 1999. A 5-year study of change in the relationship between well-being and job performance. Consulting Psychology Journal: Practice and Research 51, 252–265.

Csikszentmihalyi, M., Larson, R., 1987. Validity and reliability of the experience-sampling method. The Journal of Nervous and Mental Disease 175, 526–536.

Davidson, K.W., Mostofsky, E., Whang, W., 2010. Don't worry, be happy: positive affect and reduced 10-year incident coronary heart disease: the Canadian Nova Scotia Health Survey. European Heart Journal, ehp603.

Diener, E., 1994. Assessing subjective well-being: progress and opportunities. Social Indicators Research 31, 103–157.

Diener, E., 2000. Subjective well-being: the science of happiness and a proposal for a national index. American Psychologist 55, 34–43.

Diener, E., Biswas-Diener, R., 2002. Will money increase subjective well-being? Social Indicators Research 57, 119–169.

Diener, E.D., Emmons, R.A., Larsen, R.J., Griffin, S., 1985. The satisfaction with life scale. Journal of personality assessment 49, 71–75.

Diener, E., Oishi, S., Lucas, R.E., 2009. Subjective well-being: the science of happiness and life satisfaction. In: Snyder, C.R., Lopez, S.J. (Eds.), Oxford Handbook of Positive Psychology. Oxford University Press, Oxford, pp. 187–194.

Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L., 1999. Subjective well-being: three decades of progress. Psychological Bulletin 125, 276–302.

Diener, E., Tay, L., Myers, D.G., 2011. The religion paradox: if religion makes people happy, why are so many dropping out? Journal of Personality and Social Psychology 101, 1278–1290.

Diener, E., Tay, L., Oishi, S., 2013. Rising income and the subjective well-being of nations. Journal of Personality and Social Psychology 104, 267–276.

DeNeve, K.M., 1999. Happy as an extraverted clam? The role of personality for subjective well-being. Current Directions in Psychological Science 8, 141–144.

Dunn, E.W., Aknin, L.B., Norton, M.I., 2008. Spending money on others promotes happiness. Science 319 (5870), 1687–1688.

Ellison, C.G., Levin, J.S., 1998. The religion-health connection: evidence, theory, and future directions. Health Education & Behavior 25, 700–720.

Emmons, R.A., 2008. Gratitude, subjective well-being, and the brain. In: Eid, M., Larsen, R.J. (Eds.), The Science of Subjective Well-being. Guilford Press, New York, pp. 469–489.

Emmons, R.A., McCullough, M.E., 2003. Counting blessings versus burdens: an experimental investigation of gratitude and subjective well-being in daily life. Journal of Personality and Social Psychology 84, 377–389.

Fredrickson, B.L., 2001. The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. American psychologist 56, 218–226.

Frederick, S., Loewenstein, G., 1999. Hedonic adaptation. In: Kahneman, D., Diener, E., Schwarz, N. (Eds.), Well-being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York, pp. 302–329.

Froh, J.J., Sefick, W.J., Emmons, R.A., 2008. Counting blessings in early adolescents: an experimental study of gratitude and subjective well-being. Journal of School Psychology 46, 213–233.

Furnham, A., Brewin, C.R., 1990. Personality and happiness. Personality and Individual Differences 11, 1093–1096.

Graham, C., Eggers, A., Sukhtankar, S., 2004. Does happiness pay? An exploration based on panel data from Russia. Journal of Economic Behavior & Organization 55, 319–342.

Harker, L., Keltner, D., 2001. Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood. Journal of Personality and Social Psychology 80, 112–124.

Hayes, N., Joseph, S., 2003. Big 5 correlates of three measures of subjective well-being. Personality and Individual Differences 34, 723–727.

Headey, B., Veenhoven, R., 1989. Does happiness induce a rosy outlook? In: Veenhoven, R. (Ed.), How Harmful Is Happiness? Consequences of Enjoying Life or Not. Universitaire Pers Rotterdam, Rotterdam, pp. 106–127.

Headey, B., Wearing, A., 1989. Personality, life events, and subjective well-being: toward a dynamic equilibrium model. Journal of Personality and Social psychology 57, 731.

Helliwell, J.F., Layard, R., Sachs, J.D. (Eds.), 2013. World Happiness Report 2013. http://unsdsn.org/wp-content/uploads/2014/02/WorldHappinessReport2013_online.pdf.

Hughes, D.J., Rowe, M., Batey, M., Lee, A., 2012. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Computers in Human Behavior 28 (2), 561–569.

Ickovics, J.R., Milan, S., Boland, R., Schoenbaum, E., Schuman, P., Vlahov, D., HIV Epidemiology Research Study (HERS) Group, 2006. Psychological resources protect health: 5-year survival and immune function among HIV-infected women from four US cities. AIDS 20, 1851–1860.

Inglehart, R., 1990. Cultural Shift in Advanced Industrial Society. Princeton University Press, Princeton.

Jetten, J., Haslam, C., Haslam, A.S., Alexander, S.H. (Eds.), 2012. The Social Cure: Identity, Health and Well-being. Psychology Press, Hove, England.

Kahneman, D., Deaton, A., 2010. High income improves evaluation of life but not emotional well-being. Proceedings of the National Academy of Sciences of the United States of America 107, 16489–16493.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: the day reconstruction method. Science 306, 1776–1780.

Keyes, C.L., 2006. Subjective well-being in mental health and human development research worldwide: an introduction. Social Indicators Research 77 (1), 1–10.

King, L.A., 2001. The health benefits of writing about life goals. Personality and Social Psychology Bulletin 27, 798–807.

Kirkcaldy, B., Furnham, A., 2000. Positive affectivity, psychological well-being, accident-and traffic-deaths and suicide: an international comparison. Studia Psychologica 42, 97–104.

Krueger, A.B., Schkade, D.A., 2008. The reliability of subjective well-being measures. Journal of Public Economics 92, 1833–1845.

Larsen, R.J., Diener, E.D., Emmons, R.A., 1985. An evaluation of subjective well-being measures. Social Indicators Research 17, 1–17.

Layous, K., Lee, H., Choi, I., Lyubomirsky, S., 2013. Culture matters when designing a successful happiness-increasing activity: a comparison of the United States and South Korea. Journal of Cross-cultural Psychology 44, 1294–1303.

Layous, K., Nelson, S.K., Oberle, E., Schonert-Reichl, K.A., Lyubomirsky, S., 2012. Kindness counts: prompting prosocial behavior in preadolescents boosts peer acceptance and well-being. PLOS One 7, e51380.

Lucas, R.E., 2005. Time does not heal all wounds: a longitudinal study of reaction and adaptation to divorce. Psychological Science 16, 945–950.

Lucas, R.E., 2007. Long-term disability is associated with lasting changes in subjective well-being: evidence from two nationally representative longitudinal studies. Journal of Personality and Social Psychology 92, 717–730.

Lucas, R.E., Clark, A.E., 2006. Do people really adapt to marriage? Journal of Happiness Studies 7, 405–426.

Lucas, R.E., Clark, A.E., Georgellis, Y., Diener, E., 2003. Reexamining adaptation and the set point model of happiness: reactions to changes in marital status. Journal of Personality and Social Psychology 84, 527–539.

Luhmann, M., Lucas, R.E., Eid, M., Diener, E., 2013. The prospective effect of life satisfaction on life events. Social Psychological and Personality Science 4, 39–45.

Lykken, D., 1999. Happiness: What Studies on Twins Show Us about Nature, Nurture, and the Happiness Set-Point. Golden Books, New York.

Lykken, D., Tellegen, A., 1996. Happiness is a stochastic phenomenon. Psychological Science 7, 186–189.

Lyubomirsky, S., 2001. Why are some people happier than others? The role of cognitive and motivational processes in well-being. American Psychologist 56, 239–249.

Lyubomirsky, S., 2011. Hedonic adaptation to positive and negative experiences. In: Folkman (Ed.), Oxford Handbook of Stress, Health, and Coping. Oxford University Press, New York, pp. 200–224.

Lyubomirsky, S., Dickerhoof, R., Boehm, J.K., Sheldon, K.M., 2011. Becoming happier takes both a will and a proper way: an experimental longitudinal intervention to boost well-being. Emotion 11, 391–402.

Lyubomirsky, S., King, L., Diener, E., 2005a. The benefits of frequent positive affect: does happiness lead to success? Psychological Bulletin 131, 803–855.

Lyubomirsky, S., Sheldon, K.M., Schkade, D., 2005b. Pursuing happiness: the architecture of sustainable change. Review of General Psychology 9, 111–131.

Lyubomirsky, S., Lepper, H.S., 1999. A measure of subjective happiness: preliminary reliability and construct validation. Social indicators research 46, 137–155.

Marks, G.N., Fleming, N., 1999. Influences and consequences of well-being among Australian young people: 1980–1995. Social Indicators Research 46, 301–323.

Mehl, M.R., Vazire, S., Holleran, S.E., Clark, C.S., 2010. Eavesdropping on happiness: well-being is related to having less small talk and more substantive conversations. Psychological Science 21, 539–541.

Mishra, S., 1992. Leisure activities and life satisfaction in old age: a case study of retired government employees living in urban areas. Activities, Adaptation & Aging 16 (4), 7–26.

Myers, D.G., 2000. The funds, friends, and faith of happy people. American Psychologist 55, 56–67.

Okun, M.A., Stock, W.A., Haring, M.J., Witter, R.A., 1984. The social activity/subjective well-being relation: a quantitative synthesis. Research on Aging 6, 45–65.

Rietveld, C.A., Cesarini, D., Benjamin, D.J., Koellinger, P.D., De Neve, J.E., Tiemeier, H., Johannesson, M., Magnusson, P.K.E., Pedersen, N.L., Krueger, R.F., Bartels, M., 2013. Molecular genetics and subjective well-being. Proceedings of the National Academy of Sciences of the United States of America 110, 9692–9697.

Roberts, B.W., Caspi, A., Moffitt, T.E., 2003. Work experiences and personality development in young adulthood. Journal of Personality and Social Psychology 84, 582–593.

Sandvik, E., Diener, E., Seidlitz, L., 1993. Subjective well-being: the convergence and stability of self-report and non-self-report measures. Journal of Personality 61, 317–342.

Schwarz, N., Clore, G.L., 2003. Mood as information: 20 years later. Psychological Inquiry 14, 296–303.

Seligman, M.E., Steen, T.A., Park, N., Peterson, C., 2005. Positive psychology progress: empirical validation of interventions. American Psychologist 60, 410–421.

Sheldon, K.M., Boehm, J.K., Lyubomirsky, S., 2012. Variety is the spice of happiness: the hedonic adaptation prevention (HAP) model. In: Boniwell, I., David, S. (Eds.), Oxford Handbook of Happiness. Oxford University Press, Oxford, pp. 901–914.

Sin, N.L., Lyubomirsky, S., 2009. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. Journal of Clinical Psychology 65, 467–487.

Stack, S., Eshleman, J.R., 1998. Marital status and happiness: a 17-nation study. Journal of Marriage and Family 60, 527–536.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of personality and social psychology 54, 1063–1070.

Weiss, A., Bates, T.C., Luciano, M., 2008. Happiness is a personal(ity) thing: the genetics of personality and well-being in a representative sample. Psychological Science 19, 205–210.

Williamson, G.M., Clark, M.S., 1989. Providing help and desired relationship type as determinants of changes in moods and self-evaluations. Journal of Personality and Social Psychology 56, 722–734.

Young, S., Dutta, D., Dommety, G., 2009. Extrapolating psychological insights from Facebook profiles: a study of religion and relationship status. CyberPsychology & Behavior 12 (3), 347–350.

Sub-Saharan African Legal Traditions

Tinenenji Banda, Cornell Law School, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sub-Saharan African legal systems are pluralistic in nature and incorporate an amalgam of heterogeneous influences. These states face the ongoing task of reconciling and maintaining these diverse legal traditions in a constitutionally compliant and structurally cohesive way. This article traces the inception of legal pluralism in sub-Saharan Africa and explores the various tensions and complexities created by the complex interplay of ostensibly opposed systems of law and the attendant quest to balance completing values in the context of a unique mélange of legal cultures and tradition.

The concept of law in sub-Saharan Africa is multidimensional, and the simultaneous presence of disparate legal systems operating within the same composite space is a defining feature of its legal tradition. Indeed, all African legal systems are pluralistic in nature and incorporate varied heterogeneous influences that reflect an amalgam of African customary law, religious law, received law, legislation, and more recently, the interaction of international norms with local legal systems.

Legal pluralism with its various contradictions and convulsions has created a variety of challenges as well as opportunities for postcolonial African states that have, in the main, matured into constitutional democracies. These states face the ongoing and nebulous task of reconciling and maintaining these diverse legal traditions in a structurally cohesive way, while allowing these legal systems to operate in a way that does not fall short of constitutional imperatives and human rights obligations.

This article seeks to explore the pluralistic nature of the sub-Saharan legal tradition and the competing interests generated. The aim is not to give a comprehensive discourse on the operation of legal pluralism in sub-Saharan Africa – as that would require a more extensive substantive project, neither is it to assert that these pluralistic experiences are uniform. The goal is rather to explore the various tensions and complexities created by the complex interplay of ostensibly opposed systems of law and the attendant quest to balance competing values in the context of a unique mélange of legal cultures and tradition.

The article proceeds as follows: First, African customary law, the founding legal tradition of sub-Saharan Africa, is introduced. Here, the defining features of customary law are described. Next, the genesis of legal pluralism as legal transplants imported from the colonizing powers is explained. A discussion of the landscape of the postdemocratization legal traditions of sub-Saharan Africa follows. Finally, the interesting challenges that legal pluralism generates in the modern post-democratic sub-Saharan African state are examined. Here, the focus is on the intersection of customary succession law with modern constitutional imperatives and human rights norms using South Africa as a case study.

African Customary Law: The Founding Legal Tradition

Prior to the advent of colonialism, the sole legal tradition of sub-Saharan Africa was what is described today as 'African customary law.' African customary law still governs many sub-Saharan African populations today and remains a widespread and dominant legal tradition.

The singular source of customary law is social practice as exercised by indigenous populations and as passed down from generation to generation through oral tradition (Bennett, 2008). For indigenous populations, custom is law.

While it may be tempting to envisage precolonial African customary law as a singular concept, it was by no means such; rather its content differed among different tribal groupings and it was an eclectic mix of varying ethnic norms and practices.

Distinct Features of African Customary Law

Arguably the most notable difference between African customary law and received law (whether common or civil) is that while the subject of the latter is the individual, the main concern of customary law is the preservation of the cohesiveness of the community and kinship rights. As noted elsewhere, "in traditional African Societies, the individual citizen is primarily seen as a link of a community-chain … the full respect of social order and the necessity to achieve group harmony are crucial values which must be maintained" (Frémont, 2009: 150). The distinct interest of African customary law is therefore the community and not the individual. Individual rights are subservient to broader communitarian objectives, and this greater concern with building consensus within the community is the golden thread that runs through the normative framework of African customary law.

This communitarian approach is probably best understood by examining the dispute resolution systems of indigenous African legal systems. Unlike the adversarial approach of modern legal systems where litigants vie before an ostensibly impartial arbiter who settles the question, dispute resolution in African customary law is conciliatory and involves the community, not just the disputing parties (Owasanoye, 2001). This method seeks first and foremost to preserve social harmony. With this approach, adversaries are encouraged to devise their own solutions with the assistance of the community, and in particular, 'the elders,' who are a prominent feature of traditional African dispute resolution and viewed as a bastion of wisdom, experience, and good judgment (Owasanoye, 2001).

Another key feature of African customary law is that in its most authentic form, it is unwritten. By definition, African customary law is the "unwritten African traditional law which

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86140-1

consists of a variety of different types of principles, norms and rules …"(Assefa, 2012). As noted above, the elders are the repositories of the various norms and regulations of customary law and these norms are passed on orally.

It is thus not entirely possible to ascertain the precise content of African customary law at any given time in history. While this may seem like a recipe for anarchy when viewed in light of the Western understanding of law as a predominantly written amalgamation of rules and regulations, it is precisely this aspect of customary law that is its greatest strength, making it dynamic, amenable to change, and capable of adapting in the face of changing realities, without the long, expensive, and sometimes politically charged process of changing official law (whether through precedent or enactments).

While several African regimes have attempted to codify customary law, the net result of these efforts is often that what is actually practiced on the ground by the indigenous populations years after codification, bears little resemblance to the customary law as it exists in codified, ossified form (*Bhe v Magistrate Khayelitsha, 2005*).

This process of capturing and fixing customary law was a prominent feature of apartheid South Africa. Of course codification had a practical appeal; it aided the ascertainment and proof of customary law and as such made it easier to administer. However, the injustice done to customary law through codification was well noted by the Constitutional Court of South Africa in the landmark decision of *Bhe v Magistrate Khayelitsha* (2005) when it held that:

> because of the dynamic nature of society, official customary law as it exists in the text books and in the Act is generally a poor reflection, if not a distortion of the true customary law. True customary law will be that which recognizes and acknowledges the changes which continually take place.

The *Bhe* court went on to endorse the observations of African customary law expert Tom Bennett in holding that "a critical issue … about customary law will therefore be the question [of] whether a particular rule is a mythical stereotype, which has become ossified in the official code, or whether it continues to enjoy social currency" (Bennett, 1997).

Another distinct feature of African customary law is the manner in which it is administered. Indigenous customary law is administered by traditional leaders known as chiefs, kraal heads, and headmen who ascend to these positions through lineage. As alluded to earlier, traditional leaders in precolonial Africa were seen as custodians of the customs, values, and aspirations of the people and presided over disputes and social institutions like marriage. While traditional leaders continue to preside over customary law today, these leaders are now, in almost all cases, regulated by statute and as such the debate on the proper role of the traditional leader in a postcolonial modern African state is an ongoing one (Rambe and Madwere, 2011).

The Genesis of Legal Pluralism: Legal Transplants

Virtually all nations of sub-Saharan Africa are former colonies or protectorates. When the former colonial powers first descended upon the various African territories, the customary legal system described above was already in operation within these native communities. These laws governed far-reaching aspects of traditional life including property, succession, crime, and family life.

However, the advent of the colonialists and their attendant interests required a reordering of indigenous African institutions. Interestingly, the colonial powers did not do away with customary law but allowed it to continue operating, provided it did not impinge on their commercial or political interests (Bennett, 1982). Their reasons for doing so were of course strategic. It made both economic and political sense to preserve these institutions, as dismantling them forthwith would have placed an enormous strain on the colonial powers, which were already stretched both financially as well as in terms of human capital (Allott, 1970). Moreover had the customary law systems been summarily done away with, language and cultural barriers may have rendered the majority of African communities ungovernable.

Initially, the question of how the two legal systems (African customary law and European law) would apply was resolved fairly simply: European law would apply to Europeans and their business partners, while indigenous law would apply to native Africans (Seidman, 1969). However as time progressed, this dichotomy proved to be too simplistic. Conflicts of laws soon arose as Europeans and Africans began to have increased economic dealings with one another, and as Africans urbanized and began to adopt a Western lifestyle. The colonial administrators soon realized that a more complex mechanism to deal with this burgeoning pluralism was required (Bennett, 1985).

The manner in which choice of law rules evolved in the colonies reflected the different governing styles and legal traditions of the colonial powers. The French, the Belgians, and the Portuguese, who were of the civil law tradition, took an assimilative approach in the colonies they controlled, and crafted choice of law rules that reflected this assimilationist policy (Robert, 1959). However, the British, who were common law adherents, were not predominantly concerned with the assimilation of the local populations, but rather pursued a policy of 'indirect rule'; they adopted choice of law rules to promote this governance style (Gifford et al., 1971).

The Civil Law Colonies

In the African territories governed by colonies following the civil law tradition (France, Portugal, and Belgium), the choice of law rules were determined by looking at the legal status of the litigant (Robert, 1959). In the French colonies, the distinction between persons of *statut coutumier* status (customary status) and those of *status civil francais* (civil status) grounded the choice of law framework (Gifford et al., 1971; Salacuse, 1969).

Salacuse (1969) and Robert (1959) explain how the French colonies dealt with native Africans and customary law. By decree, it was established that all legal matters arising between African litigants must be governed by customary law, while legal matters arising between litigants with civil status (initially only Europeans) would be governed by civil law.

Thus the status of the litigants was the only consideration. It was left to the courts to extrapolate on this basic choice of

law rule in individual situations that required a more complex approach. The status of a litigant was determined by citizenship, which was in turn determined by race. If a person was of European origin, they were considered a citizen and therefore civil law applied. If the person was of African descent, they were not a citizen and customary law applied.

Prior to 1946, it was impossible for an African to be considered a citizen unless they underwent a naturalization process sealed by an administrative or judicial act. In order to be naturalized, an African had to demonstrate that they were of good moral character, had attained a basic level of education or served in the military, had renounced customary law, and was not part of a polygamous marriage. When this was proven to the satisfaction of the governing authorities, an African could be granted citizenship and civil law would govern his or her affairs.

In 1946, the method by which indigenous Africans in the French-administered colonies could be granted civil status was changed. Through two enactments, all French subjects in overseas territories were granted citizenship subject to the proviso that those who did not renounce their personal status retained that status and did not attain French civil status. And then in 1955, the law was again changed to allow any African over the age of 21 to renounce his or her customary status by means of a declaration of a court applying French law. However, the proviso that such person could not be a party to a polygamous marriage was retained. It also had to be shown that the subject had been notified of the irrevocable consequences of the revocation. The French therefore pursued an all or nothing approach. If a subject attained civil status, their revocation of customary law was complete and final.

Similarly in the Belgian-controlled territory of the Congo, the 1908 Colonial charter gave express recognition to customary law, but also provided that Africans could renounce customary law and become subject to civil law by way of a process called "immatriculation" (Robert, 1959).

In the Portuguese-controlled territories, while African customary law was recognized, it was seen only as a transitory measure pending the full assimilation of Africans into the Portuguese way of life. An African could attain civil status through exercising one of two options. The first was a partial acceptance of civil law, in which civil law would administer limited aspects of one's estate (property, inheritance, contracting, and family) and the second option was a full acceptance of civil status in its totality and a total and irrevocable renunciation of customary law (Robert, 1959). In order to exercise the second option, one had to be 18 years or older, fluent in Portuguese, employed, and not an evader of military service (Robert, 1959). Application was made through a court-administering Portuguese law. Later, in 1953, a decree was passed to provide that metropolitan Portuguese law would be applied to every inhabitant of a Portuguese territory whether inland or overseas (Robert, 1959).

The Common Law Colonies

In the Anglophone colonies, however, the race or the personal status of the litigant was not the principal thing used to determine which law would apply to a particular individual, rather the subject matter of the case was the most important consideration. In addition, the choice of law system adopted differed in varying British colonies. In West Africa, for instance, Section 19 of the Gold Coast Supreme Court Ordinance No. 4 of 1876 provided that

> such laws and customs shall be deemed applicable in causes and matters where the parties thereto are natives of the Colony, and particularly, but without derogating from their application in other cases, in causes and matters relating to marriage and to the tenure and transfer of real and personal property, and to inheritance and testamentary dispositions, and also in causes and matters between natives and Europeans where it may appear to the Court that substantial injustice would be done to either party by a strict adherence to the rules of English law.

However in East Africa, Sections 3 and 4 of the 1897 Native Court Regulations held thus:

> The native courts … shall, as far as practicable … be guided by, and have regard to any native laws and customs not opposed to natural morality and humanity. The native courts … shall be guided, subject to the provisions of these regulations, by the native laws or customs existing in their respective jurisdictions.

As can clearly be seen from the excerpts above, in contrast to what obtained in the civil law countries, the Anglophone choice of law system was not concerned primarily with race, but rather with the subject matter of the transaction. Moreover, in certain situations where an African was following what could be deemed a Western lifestyle, English law could be used to govern their affairs without the detailed renunciation process required in the civil law colonies (Bennett, 1982). Partial application of the common law was also acceptable; it was entirely possible for one to have English law apply to his commercial transactions, for instance, but to choose customary law for his marriage and other personal affairs.

In the Anglophone territories, where a conflict between customary law and common law arose, the common law trumped customary law (Ndulo, 2011). However, where there was no conflict, customary law were not tampered with unless a particularly customary law rule was found to be 'repugnant.' The repugnant aspects of customary law were dealt with through the introduction of repugnancy clauses. (Ndulo, 2011). These clauses were enacted in virtually all the Anglophone colonies.

In its basic form, the repugnancy clauses prohibited the enforcement of a customary law rule that was 'repugnant to natural justice, equity, or good conscience.' Explaining what he understood to be the necessity of the repugnancy clause, Lord Atkin in the case of *Eshugbaye Eleko v. Government of Nigeria (1930)* stated that "[T]he court cannot itself transform a barbarous custom into a milder one. If it stands in its barbarous character it must be rejected as repugnant to natural justice, equity and good conscience."

Ultimately, the British settlers gave themselves authority to "establish such laws, institutions and ordinances, and to constitute such courts and offices as may be necessary for the peace, order and good government (Section 1 of the British Settlements Act 1943)" in the territories under their control.

But, as noted above, this far-reaching authority was not exercised in a way that did away with African customary law. While certain aspects of customary law were done away with through the use of the repugnancy clause, most aspects were preserved.

Hybrid Colonies

Hybrid colonies are those sub-Saharan nations that were colonized by more than one power, South Africa and Namibia being cases in point. Namibia was colonized by the Germans from 1884 to 1915 and then by South Africa from 1915 to 1989 (Ruppel and Schlichting, 2011).

South Africa's colonization history is less straightforward. South Africa was first occupied by the Portuguese in 1488, but the Portuguese settlers did not establish or claim it a colony (Thompson, 1990). In 1652, the Dutch East India Company arrived in the Cape of South Africa and established a settlement. The British entered the scene in the 1800s and captured the Cape from the Dutch (the 'Boers'), and in 1803 established sovereignty in the area. Soon the Dutch grew weary of British sovereignty and embarked upon a 'Great Trek' – a mass movement into the interior of South Africa in search of independence from the British (Thompson, 1990). As this movement occurred, the Dutch clashed with African tribes and experienced several crushing defeats at the hands of Zulu warriors (Thompson, 1990). However, they persevered in their quest for independence and established Boer republics in the Transvaal and the Orange Free State (Farwell, 1976). Tensions between the British settlers in the Cape and the newly formed Dutch colonies lead to two installments of an Anglo-Boer war that occurred between 1877 and 1902 (Farwell, 1976). A truce was reached in 1902 and in 1909, the Union of South Africa, a conglomeration of British-controlled territory and Boer republics, was born (Farwell, 1976).

South African law is thus an amalgamation of Roman Dutch law, English law, and African customary law. Prior to the new constitutional dispensation of 1994, African customary law had second-class status in South Africa, but this has since changed and the South African Constitution places customary law on an equal footing with the common law. Namibia as a former South African colony is similarly pluralistic.

The Postdemocratization Era

While the dismantling of colonialism across the African continent in many cases led to a 'romanticized notion of restoring [true] customary law,' it was indisputable that the received law had had an indelible and irreversible effect on the former colonies (Quarmyne, 2011). Even after attaining independence, the former colonies retained the received law and did not revert back to the precolonial era of universal application of African customary law. A legal transplant had occurred. And, even though in most independent states African customary law continued to enjoy recognition and in some cases was elevated to equal status with the received law, legal pluralism had become the hallmark of African legal systems.

The former civil law colonies for the most part did away with the distinction between civil law and customary law and instead created a unified code of law that is an amalgamation of civil law and customary law and is applicable to all citizens (Bennett, 1982). The common law countries including Namibia and South Africa have maintained pluralistic legal systems in which common law and customary law exist side by side. What one finds in most of these sub-Saharan African countries then is an amalgamation of laws from disparate sources and these sources are typically listed in the Constitution.

In Ghana, for instance, a former British colony, the Constitution lists the country's sources of law as follows:

"The laws of Ghana shall comprise (a) the constitution; (b) enactments made by or under the authority of the parliament established by this constitution; (c) any Orders, Rules, and Regulations made by any person or authority under a power conferred by this Constitution; (d) the existing law and (e) the common law." The Ghanaian Constitution goes on to state that the common law includes customary law (Chapter 004 of the Constitution of the Republic of Ghana).

In Zambia, the Constitution of the republic avers that the Laws of Zambia consist of:

1. this Constitution;
2. laws made by or under the authority of Parliament;
3. any orders, rules, regulations, and other statutory instruments made by any person or authority under a power conferred by this Constitution or any other law;
4. the British laws and statutes, which apply or extend to Zambia as prescribed by an Act of Parliament;
5. Zambian customary law, which is consistent with this Constitution;
6. the common law of England, which is consistent with this Constitution;
7. the rules of law generally known as the doctrines of equity; and
8. the law as determined by the superior courts (Section 14 of the Constitution of the Republic of Zambia).

The sources of law in other countries are not dissimilar and it is clear that sub-Saharan African legal systems are heavily pluralistic.

The Survival of African Customary Law

Legal pluralism can and does pose several interesting challenges in the modern postdemocratic sub-Saharan African state. There pluralistic systems often interact and transpose upon one another in ways that are anything but neat and circumspect. There are, for instance, inherent tensions between African customary law and the human rights norms that many African states have voluntarily subscribed to. Embedded within African customary laws are various practices that are discriminatory in nature in their application to women, for instance. Some of the more hotly contested issues include bride price, polygamy, and primogeniture.

For constitutionally democratic African states, the challenge is and continues to be how to preserve customary law (which in most cases is constitutionally protected) and the rights of indigenous populations to be governed by it, while simultaneously ensuring that its practice is in consonance with constitutional and human rights imperatives. As noted by

Bennet in reference to the South African context "[B]y implicitly recognizing customary law … the Constitution has brought about a head on confrontation between two opposed cultures, admittedly a confrontation that has long been gathering force" (Bennett, 1994).

So, for instance, while most sub-Saharan African Constitutions contain an equality clause that prohibits discrimination, the protection of African customary law (which is inherently discriminatory in several respects) is also guaranteed by the Constitution. This creates a conflict that the text of the Constitution often does not provide a mechanism of resolving. So, for instance, the Zambian Constitution in article 23 (2) states that no person shall be treated in a discriminatory manner, but this is qualified by s 23 (6) (d) that permits discrimination "for the application in the case of members of a particular race or tribe, of customary law with respect to any matter to the exclusion of any law with respect to that matter which is applicable in the case of other persons."

The area of succession and inheritance norms provides an apt example of the clash between customary law and broader normative considerations of equality under the law and will now be considered in some detail.

The rule of male primogeniture features in almost all customary law succession systems. The basic tenet of the rule is that only male heirs are entitled to inherit property to the exclusion of females. In recent times, this has come under judicial scrutiny and African courts have been forced to grapple with the question of whether such a rule can stand scrutiny under the equality clauses (and in some cases bill of rights) that are a prominent feature of almost all postcolonial African Constitutions. The contentiousness of the issue lies in the fact that customary law is constitutionally protected and is a recognized and enforced system of law in all post-democratized constitutional African states.

In some African nations, the repugnancy clause has ironically been retained as a mechanism to deal with the aforementioned clash. As noted above, the British colonialists used a repugnancy test to determine which aspects of customary law should be done away with. The repugnancy clauses were and still are controversial, because they ignite old debates about Western universalism and cultural relativism. What is interesting is that the language of repugnancy has for the most part been retained by African courts and African customary law is still in many instances subject to this test even postdemocratization.

For example, in *Mojekwu v Mojekwu* (1997) the Nigerian High Court examined the Nnewi custom, which dictated that if a man died and was survived by a female child or children and not a male child, the deceased's brother or brother's male descendants were the rightful heirs. In the case in question, the deceased had a daughter who was excluded from inheriting from her father's estate in favor of her late-father's brother's son. Using the language of repugnancy, the court held:

> Any form of societal inhibition on grounds of sex, apart from being unconstitutional, is antithesis to a society built on the tenets of democracy which we have freely chosen as a people. On my part, I have no difficulty in holding that the 'Oli-Ekpe' custom of Nnewi is repugnant to natural justice, equity and good conscience.

Repugnancy aside, what is clear from an observation of the jurisprudence of many African courts is that certain aspects of customary law are being invalidated in favor of constitutional imperatives. The courts have frequently used the flexible nature of customary law to justify these decisions. In Botswana, for instance, the Supreme Court in invalidating the rule of primogeniture held concerning customary law.

> …Customary law is not static; it develops and is modernized with the times, harsh and inhumane aspects of custom being discarded as time goes on. More liberal and flexible aspects consistent with the changing ethos being retained and probably being continuously modified on a case by case basis or at the insistence of the traditional leadership to keep pace with the times.…For after all what is customary law but a set of rules developed by society to address issues around certain values which protect the community's social fabric and cohesion … there is no rational and justifiable basis for sticking to the narrow norms of days gone by when such norms go against current value systems.
>
> *Ramantele v Mmusi and Others*, 2012

However, the question of how customary law is to be developed by the courts is one that is a subject of contention. It has been argued that courts have a mandate to develop customary law and not simply obliterate it, and that it is not enough for Courts to pay lip service to customary law but that they have the responsibility to ensure its survival (Mbatha, 2002).

The Case of South Africa

In South Africa, the interesting question of whether the Bill of Rights is directly or indirectly applicable to customary law has arisen with some frequency. The distinction is not a superfluous one, because if the Bill of Rights is directly applicable, the result is that any offending rule must be struck down (Himonga and Bosch, 2000). As Justice Sachs noted in *Du Plessis and others v De klerk and another* (1996) "The direct enforceability of [the bill] could require this Court, if asked to do so, to indulge in a wholesale striking down of customary law because of [its] violation of the equality clause …" (*Du Plessis and others*, 1996). Further, customary law practitioner Alastair Kerr argues in the South African context that "the Constitutional Court would not have certified the Constitution had it thought that in doing so it was bringing about … a destructive confrontation between the Bill of rights … and indigenous law …"(Kerr, 1997).

On the other hand, indirect applicability of the Bill of Rights mandates the court to develop the contravening rule in line with the spirit, purport, and objects of the Bill of Rights in accordance with S 39 (2) of the South African Constitution, which states "when interpreting legislation and when developing the common law or customary law, every court, tribunal or forum must promote the spirit, purport and objects of the Bill of Rights." This approach ensures that an offending customary law rule is not just obliterated but developed.

The constitutionality of the customary principle of male primogeniture as well as the direct or indirect applicability of the Bill of Rights was tested in the South African landmark case of *Bhe v Magistrate Khelitsha* (2005). In that case, the Constitutional Court of South Africa declined to develop the

customary male principle of primogeniture because, according to the Court, such development would create problems of ascertainment. That is, the Court felt that in order to develop the offending customary law rule, the Court would first have to ascertain the breadth of the rule as it was practiced on the ground, and that this would present problems of proof. In light of this, the Court applied the Bill of Rights directly and thus invalidated the rule.

This approach was criticized by customary law scholars who felt the Court reneged on their duty to develop the customary law and instead took the easier route of rendering it invalid (Mbatha, 2002). It was argued that in declining to develop the rule, the Court missed a crucial opportunity to narrow the gap between the official customary rule and the day to day practices on the ground because, as has been observed by Likhapa Mbatha, "… communities that lead a customary life have no problem with allowing a woman to inherit property" (Mbatha, 2002).

Conclusion

Sub-Saharan legal systems are fundamentally pluralistic. This pluralism is a legacy of their colonial past and is reflected in unified civil codes that reflect both customary and received law, as well as in the sources of law typically enumerated in a nation's Constitution. The coexistence of different systems of law has not been without its difficulties. As the place of customary law within constitutional democracies evolves, African courts will continue to grapple with the question of how to preserve what is a distinct African heritage, while complying with constitutional and human rights imperatives.

See also: Constitutionalism, Comparative; Law: Change and Evolution; Legal Pluralism; Legal Systems, Classification of; Middle Eastern and North African Legal Traditions; Postcolonial Law.

Bibliography

Allott, A.N., 1970. New Essays in African Law. Butterworths, London.

Assefa, A.G., 2012. Customary laws in Ethiopia: a need for better recognition. The Danish Institute of Human Rights. Working Paper.

Bennett, T.W., 1982. Conflict of laws: the application of customary law and the common law in Zimbabwe. The International and Comparative Law Quarterly 30 (1), 59–103.

Bennett, T.W., 1985. The Application of Customary Law in Southern Africa: The Conflict of Personal Laws. Juta, Cape Town.

Bennett, T.W., 1994. The Equality clause and customary law. South African Journal of Human Rights 10, 122–130.

Bennett, T.W., 1997. Human Rights and African Customary Law under the South African Constitution. Juta, Cape Town.

Bennett, T.W., 2008. Comparative law and African customary law. The Oxford Handbook of Comparative Law 642–671.

Bhe v Magistrate Khayelitsha and Others 2005 (1) BCLR (CC).

Du Plessis and others v De Klerk and another 1996 (3) SA 850 (CC), 1996 (5) BCLR 658 (CC).

Eshugbaye Eleko v. Government of Nigeria Appeal No 42 of 1930.

Farwell, B., 1976. The Great Anglo-Boer War. Harper & Row, New York.

Fremont, J., June 01, 2009. Legal pluralism, customary law and human rights in Francophone African countries. Victoria University of Wellington Law Review 40 (1), 149–165.

Gifford, P., Louis, W.R., Yale University, 1971. France and Britain in Africa: Imperial Rivalry and Colonial Rule. Yale University Press, New Haven.

Himonga, C., Bosch, C., 2000. The application of African customary law under the constitution of South Africa: problems solved or just beginning? South African Journal of Human Rights 117 (2), 306–341.

Kerr, A.J., 1997. The bill of rights in the new constitution and customary law. South African Journal of Human Rights 114, 346–352.

Mbatha, L., 2002. Reforming the customary law of succession. South African Journal of Human Rights 18, 259–286.

Mojekwu v Mojekwu [1997] 7 NWLR 283.

Ndulo, M., 2011. African customary law, customs and women's rights. Indiana Journal of Global Legal Studies 18 (1), 87–120.

Owasanoye, B., 2001. Dispute Resolution Mechanisms and Constitutional Rights in Sub-saharan Africa. UNITAR Resource Center. Chapter 2, Document No.14, 15–25.

Quarmyne, M., 2011. Witchcraft: a human rights conflict between customary/traditional laws and the legal protection of women in contemporary Sub-Saharan Africa. William & Mary Journal of Women and the Law 17, 475–507.

Rambe, P., Madwere, M., 2011. African traditional customs potentialities and dilemmas: conflict over thanksgiving to chiefs in Zimbabwean rural villages. International Journal of Politics and Good Governance 3 (3.2), Quarter 2.

Robert, A., 1959. A comprehensive study of legislation and customary law courts in the French, Belgian and Portuguese territories in Africa. Journal of African Administration 124–131.

Ruppel, O.C., Ruppel-Schlichting, K., 2011. Legal and judicial pluralism in Namibia and beyond: a modern approach to African legal architecture? Journal of Legal Pluralism and Official Law 64, 33–63.

Ramantele v Mmusi and Others (CACGB-104-12) [2013] BWCA (3 September 2012).

Salacuse, J.W., 1969. An Introduction to Law in French-speaking Africa. Va: Michie Co, Charlottesville.

Seidman, R., 1969. The reception of English law in colonial Africa revisited. East Africa Law Review 47, 53–54.

Thompson, L.M., 1990. A History of South Africa. Yale University Press, New Haven.

Suburbanization and Suburbanism

Richard Harris, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A global phenomenon, suburbanization entails low-density settlement at the urban fringe. It is caused especially by urbanization, improved transportation, job decentralization, public policy, and a common preference for suburban living. Most suburbs are developed by private agents guided by the state; in the global South many are produced informally by squatters or 'pirate' developers. They take many forms: residential, industrial, mixed; exurban/peri-urban, outer, inner; polycentric, dispersed; planned, unplanned; elite, middle class, low income; and ethnic. Arguably, they reflect and influence suburbanism, a private, family-focused way of life. Surbanization is necessary, and reduces housing costs, but entails major public and private expenditures, reduces farmland, and disrupts ecosystems.

Suburbanization is how urban areas expand and change. It happens when urban populations increase, incomes rise, transportation technology improves, and jobs decentralize. Channeling capital and shaped by private-sector agents, it is usually constrained by municipal regulations and sometimes promoted by governments (Ekers et al., 2012). It takes a variety of low-density physical forms, including sprawl, and involves most types of people. Because it is especially favored by families with children, it is arguably associated with a distinctive way of life, known as suburbanism. It has consequences for population health, the economy, the environment, and municipal governance.

Suburbanization is global but its evolution in the global North and South has been interpreted separately, using different terminology (Angel, 2012; Champion and Hugo, 2004; Harris, 2010; Phelps et al., 2010). The literature on suburbanization has focused on North America and, to a lesser extent, the United Kingdom and Australasia (Clapson, 2003; Palen, 1995; Teaford, 2008). This follows the geographical distribution and interests of urban researchers and publications. Those same researchers have spoken of urban fringe development in the global South in other terms, emphasizing squatter, or other informal-sector settlements. This difference is not surprising. By almost any definition, suburbs first developed in the West. In the nineteenth century, England became the first urban nation – where more than half of the national population lived in urban areas – and the first to see large-scale suburban development, notably around London. In the twentieth century, rising prosperity in the United States created extensive suburban development, which took especially decentralized forms (Schneider and Woodcock, 2008). Because of America's geopolitical influence, and the reach of its media, images of its suburbs are ubiquitous. They have influenced consumers and developers elsewhere, embodying what many still think of as the quintessential suburban form.

Definitions

Definitions are correct only in so far as they are useful. Suburbanization entails the growth of suburbs but, because there is no generally accepted definition of this term, the meaning of both must be explored. The issue is complicated by the fact that, although 'suburb' or equivalent terms is used in many parts of the world, its connotations vary. In the United States, but not elsewhere, a true suburb is politically distinct; this is significant because under the US Constitution, no higher level of government can abolish or amalgamate municipalities without their permission (Nicolaides and Wiese, 2006; Teaford, 2008). In Spanish and Italian-speaking countries, 'suburbio' does not carry strong connotations of prestige, and in France 'les banlieues' are stigmatized. China has no precise equivalent to suburb, and India lacks a generic term. In some sense suburbs exist everywhere, but they are defined and understood variously. Differences of definition and understanding can even exist within a city, perhaps especially in the global South. Western-educated politicians and planners speak of suburbs, but this may have no resonance for residents who speak other languages and have a different understanding of their city's geography.

In the English-speaking world, 'suburb' can mean different things (Clapson, 2003; Forsyth, 2012; Harris, 2010; Palen, 1995; Phelps et al., 2010). A narrow definition, meaningful because influential as a model in Anglo-America and beyond, is that of the low-density, socially exclusive, and purely residential area located at or near the urban fringe. But, in the past as in the present, extensive urban fringe areas have failed to conform to this stereotype. In particular, many house low- and moderate-income households and, although they contain residential subdivisions, they often also include commercial and industrial land uses. And the Anglo-American stereotype is not generally employed elsewhere. In general, two aspects are emphasized: moderate residential density and urban fringe location. These are related but distinct, which raises the unresolved question as to whether a suburb must satisfy both criteria, or whether one takes priority.

Reflecting the fact that land at or near the edge of the built-up area is intermediate in price between urban and rural lands, so too are residential densities. In many cities, the typical urban dwelling is a townhouse (row or terraced) or apartment building, while a suburban home is likely to be detached or semidetached. Collectively, the latter constitute a 'middle landscape' that, depending on one's point of view, marries the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74044-X

best or the worst elements of city and country (Clapson, 2003; Palen, 1995: 95–99). Commonly, residential densities decline from city center to urban fringe, although this may be complicated in at least two ways. First, where incomes are low and/or land prices are high (e.g., Hong Kong, Singapore, Istanbul) even fringe development may occur as high-rise towers. At the scale of the building, densities are very high, but where towers are widely spaced, area-wide densities are moderate. In regions where land market forces had little influence, including communist-era Russia, China, and central Europe, closely spaced high-rise buildings were erected, so that urban fringe development took an unusually dense character (Hirt, 2012). A second and increasingly common complication is the growth of major, competing centers of employment and retailing in fringe locations. Often known as Edge Cities, these become nodes of higher land costs and building densities. Polynuclear and dispersed growth can create very complex patterns of land use, but in general densities still decline toward the fringe.

Regardless of density, districts may be viewed as suburban simply because they are located near the edge of the built-up area. 'Near' has no precise meaning. Not all urban areas have sharp physical boundaries: sprawl may involve the creation of isolated subdivisions scattered across an 'exurban,' semirural landscape. Then, too, as urban areas expand, districts once located at the fringe will be engulfed by later development; it is difficult and indeed arbitrary to determine at what point they cease to be suburban. Commonly, a distinction is now made between outer-ring and inner-ring ('mature' or 'older') suburbs (Hanlon, 2010). In growing urban areas, outer-ring suburbs will be newly built and settled. Their novelty may be socially significant, but is not usually regarded as a defining characteristic.

Inevitably, suburbanization occurs as the population of an urban area increases. Growth must usually be accommodated mainly through fringe development. There is limited vacant land near the urban center, while redevelopment at higher densities is expensive, in part because it typically involves the purchase and assembly of several land parcels from sometimes-reluctant owners. Most challenging of all, it arouses NIMBY (Not In My Back Yard) opposition, which even authoritarian governments may wish not to arouse. Suburbanization may also be said to occur surreptitiously, as densities everywhere decline. For many decades over the past century, through the demolition of slums, a reduction in average household size, and the relocation of many families to the suburbs, residential densities declined in the central areas and inner suburbs of almost all major cities (Angel, 2012). Densities in some city neighborhoods today are similar to those of outer suburbs a century ago. Now, as outer suburbs boom, it may seem that urban areas are becoming more suburban in character. In that sense, it has been suggested that the United States (or, more probably, Australia) became the first suburban nation. Extrapolating, a world that is now predominantly urban will increasingly become suburban.

Depending on how suburbs are defined, this trend is neither inevitable nor in the long run likely. As urban areas expand, outer suburbs become inner, while inner suburbs become city neighborhoods. In the process, residential areas cease to be, and to be thought of, as suburban. (Even in Australia, where residents refer to city neighborhoods as 'suburbs' everyone recognizes that such areas are physically and socially different from new subdivisions.) It is not just the relative location of inner suburban areas that has changed. Some, especially elite neighborhoods with restrictions that prevent redevelopment, may preserve their character over many decades. But most undergo physical and social changes, typically involving the incursion of new land uses and redevelopment at higher building densities. Much of the residential density decline that has occurred in inner areas over the past century has been due to the conversion of residential to other forms of land use. Now, even in many North American cities, it is reversing: the popularity of condominium tenure since the 1970s has made it possible for people to live at high densities and still own their own homes – a common aspiration. High-rise Singapore has the highest level of home ownership of any major city in the world. In this unobtrusive manner, over a period of decades residential areas cease to be suburban in density as well as location. Whether a city is becoming more or less suburban depends on what balance is struck between unobtrusive 'citification' and the more obvious development of outer suburbs. Currently, in most cities, the balance favors suburbanization. But the larger urbanization process is necessarily finite, while current fertility trends suggest that world population may peak within 50 years. As the population of cities levels off, and especially if environmental considerations favor higher density forms of settlement, the balance will shift, and perhaps dramatically.

Causes

Although many accounts overlook the fact, over the past two centuries the main cause of suburbanization has been the growth of urban populations. In most regions, such growth was promoted by an overall increase in population and also by urbanization, that is, an increase in the proportion of people living in urban areas. Until the mid-twentieth century, because cities were unhealthy, urbanization was driven by rural–urban migration. It still is, most dramatically in China. But recently, in most world regions, improvements in public health have made cities healthier than rural areas. Urbanization, and therefore suburbanization, has depended in part on natural increase.

Suburbanization has happened, and continues to occur, even in the absence of urban population growth, as population densities decline (Angel, 2012). Five related influences are commonly invoked to explain this trend; two others have also been important though often overlooked. Their relative importance has varied across place and time. The most important has been improved transportation technology. At first, railways, horsecars, and then electric streetcars (trams) enabled people to live at increasing distances from centrally located workplaces, clustering along transit lines and around stations. Affluent nineteenth-century railway suburbs, and middle-income streetcar suburbs, resulted. Then, the personalized technology of bicycles, cars, and mopeds/scooters enabled a further, and also more dispersed, pattern of decentralization. Cars began to shape suburbs in North America in the 1920s, Europe after 1945, Latin America in the 1960s, India

in the 1970s, and then China in the 1980s. Automobiles have played such a large role in shaping modern suburbs that many observers view suburbs as places where people rely on them for commuting, shopping, and other errands.

All of these technologies, even streetcars but especially cars, are costly. Their widespread use has depended on rising incomes, initially among the middle class and then for workers. Growing affluence is a second influence on the growth of suburbs, and suburbanization is one of the most visible signs of urban prosperity. The most affluent nations are also the most suburban.

A third major cause of suburbanization has been the decentralization of employment, which has made it possible and preferable for many employees to live in adjacent suburban areas. Although the fact is not always acknowledged, many industries were moving to the suburbs of North American and European cities from the nineteenth century, and by the early postwar era these were soon followed by office employment and retailing (Harris and Lewis, 2001). This trend, too, was facilitated by improvements in transportation, and notably the growing use of truck transportation from the 1920s. At first, workers settled within walking distance of suburban factories in industrial suburbs. This pattern is still common in the global South. Automobiles and mopeds have given employees wider options.

Transportation technology allows, but does not compel, workers to live in the suburbs. If they have chosen to do so, that is because of a widespread preference for suburban life. Where workers also have jobs in the suburbs, this involves no tradeoff. Historically, however, and even today in many cities, workers are willing to spend time and money commuting downtown in order to enjoy the greater quiet, spaciousness, and privacy of suburban living. These features depend on larger homes and gardens, readier access to parks and open space, and removal from centers of commercial activity. Historically, suburbs were also perceived to be healthier, with cleaner air and less industrial pollution. Because of relatively cheap land, suburban home ownership is affordable to more households and that has been a major additional motive. Owner-occupancy is widely preferred because it offers greater security of tenure than tenancy, and more opportunities to adapt domestic space to changing family needs. Transportation technology and decentralizing jobs have made it possible for millions of households to realize their suburban dream.

Recognizing the strength of this popular aspiration, governments have often subsidized it. These usually implicit subsidies, a fifth significant cause of suburbanization, vary greatly in strength by country and take many forms. In the United States, federal support for highway (motorway) construction, coupled with income tax deductions for mortgage interest and property taxes, created unusual subsidies for low-density development. There, and in many other countries, implicit subsidies for all types of road construction, coupled with property taxes that do not reflect the full cost of servicing low-density development, have produced the same effect (Blais, 2010). So too has the general failure of governments to levy effective taxes (betterment charges) on the dramatic price increases that occur when land is converted from rural to urban use (Evans, 2004). This failure continues to provide major incentives for land conversion and development.

A sixth and often overlooked influence on suburbanization has been a delay in the creation of effective methods of public health and for the disposal of human, animal, and industrial wastes. The consequences are usually worst in the most densely settled districts close to city centers, and this has shaped people's belief that suburbs are healthier. In the nineteenth century, the problems of public health were those of ignorance, notably about how diseases spread. Today, the challenge usually faced by municipalities is how to pay for the necessary infrastructure.

All discussions of suburbanization in the global North assume that those who move to the suburbs have chosen to do so, recognizing that in some cases (Paris, New York City, London) their choices are constrained by the unaffordability of centrally located real estate. In fact, some households have had almost no choice at all; involuntary relocation has sometimes been a seventh significant force for suburbanization. In Europe, housing agencies have allocated low-income households to peripheral estates. Often families moved willingly but sometimes not, because peripheral estates were far from jobs and stores, often poorly served by transit, and soon stigmatized. A comparable situation exists elsewhere, for example Delhi, where hundreds of thousands of families, displaced through slum clearance, have been moved to isolated resettlement colonies (Dupont, 2005). The undesirability of these places is underlined by the fact that many residents abandon modest but decent houses and return to cramped squatter quarters in the city.

The Development of Suburbs

Suburban development entails the conversion of rural or unoccupied land to urban use through the legal subdivision of land into smaller parcels, and the laying out of streets. Municipal regulations commonly require the provision of hard infrastructure, notably piped water and sewers, and sometimes services such as schools.

The development and governance of suburbs involves a variable balance of public and private agencies (Ekers et al., 2012). Development is sometimes initiated by a municipality, or public agency. This happened in communist countries, and municipalities in China are still entrepreneurial (Hsing, 2010; Wu and Shen, in press; Zhang, 2010). More modestly, social-democratic governments develop large-scale public housing (Clapson, 2003). In Britain, state agencies built New Towns and suburban estates; in France, modernist projects (*grands ensembles*) defined many banlieues; the City of Stockholm linked comprehensive suburban development to the construction of fixed-rail transit. More typically, municipalities play a permissive and regulatory role. Land use master plans specify which areas may be developed, and for what purposes; building regulations control designs and methods of construction; housing bylaws mandate maintenance; charges levied on developers cover the costs of providing infrastructure, including roads; and property taxes support public services and the maintenance of infrastructure (Evans, 2004). By taxing, or failing to tax, fuel and capital gains, national governments shape the suburbs. Governments at all levels affect the suburbanization process.

Suburban development is usually undertaken by private companies and coordinated through the markets for land, housing, retail, office, and industrial space (Rybczynski, 2008). Today, developers often install infrastructure, which they pass to the municipality for maintenance. An increasingly common alternative is the common-interest development (CID), sometimes gated, where common facilities are owned jointly by property owners who pay fees to cover upkeep (McKenzie, 1994; Nicolaides and Wiese, 2006). Facilities may include play areas, miniparks, pools, community halls, and even roads. Once developed, responsibility for maintenance lies with a homeowner association. Developers may require buyers to sign covenants that govern who may live there and how property is maintained. Many buyers believe such regulations help maintain standards and property values.

Developers usually delegate construction to builders. Because of the large capital investments involved, they rely on credit. To minimize risk, lenders favor conventional schemes and, for ease of later sale, so do home buyers. In any city, region, or nation, most suburban developments – whether of homes, apartments, malls, or offices – conform to a limited range of types. As some lenders and developers now operate across borders, and because some suburban models (such as malls and gated communities) have become popular, standardization is becoming international (Glasze et al., 2006; Rimmer and Dick, 2009).

In North America, Europe, Australasia, and Japan, suburban development occurs within the formal sector: it conforms to prevailing laws and regulations. In many cities elsewhere, informal development is common. Sometimes, squatters settle and build on land they do not own. This is commonest on publicly owned land because governments are wary of evicting citizen-voters. As public lands are growing scarce, squatting is becoming rarer. Instead, a growing amount of suburban development is undertaken by 'pirate' entrepreneurs who acquire and sell land legally, but where residential development is not authorized (Doebele, 1977; Dupont, 2005; Klaufus, 2012). Commonly, these developers fail to install required infrastructure; dwellings, commercially or owner-built, often violate building regulations. Although informal settlements are illegal, they may be tolerated or implicitly encouraged by municipalities because they relieve the shortage of housing for low- and moderate-income families. Around cities in the global South, the formal and informal land markets are interdependent, and together produce suburbanization (Benjamin, 2004; Rimmer and Dick, 2009; Wu and Shen, in press).

Types of Suburbs

Today, any large metropolitan area may include a wide variety of suburbs. How much more this is true of the suburbs of a nation or, even more obviously, the world (Harris, 2010).

Physically, the main dimensions of difference are those of age, land use, pattern of development, and degree of planning. Urban areas expand into the countryside. Occasionally, as with Cairo's desert, fringe territory is unoccupied; sometimes, as with areas of wet-rice cultivation in Southeast Asia, it is densely settled; normally it contains a mixture of farms, villages, and towns. Expanding cities engulf such areas through the dispersal of urban families, seeking space and a rural ambience, or by attracting rural migrants who seek urban jobs but cannot afford urban rents. Dispersal is more common in the global North, where it creates exurbia (Taylor, 2011). Rural–urban migration is the norm in China and the global South, where it helps create what Western researchers refer to as a rural or peri-urban fringe (Simon, 2008; Hsing, 2010). Geographer Terry McGee has labeled the very dense urban fringes of Southeast Asia *desakotas* (villagecities). By whatever name, such areas are the first, presuburban phase of urban expansion. They contain rapidly changing and strangely juxtaposed elements, including truck farms, billboards, villages, waste dumps, individual houses, gravel pits and quarries, small subdivisions, and industry, in no apparent order.

Suburban development emerges with consolidation, or the redevelopment, of fringe development into a continuously built-up landscape. Consolidation involves infill, perhaps in the uncoordinated form of individual houses and small subdivisions. Redevelopment requires planning and often government approval and produces a more uniform landscape. The result is an outer suburb, at first raw, with limited vegetation, new residents, and few services. Such suburbs mature: lawns, bushes, and trees grow; residents adapt their dwellings, introducing variety into standard subdivisions; they make friends of neighbors; stores open and services develop. In time, dwellings may deteriorate or become obsolete; lower income households move in; homes may be subdivided and maintenance declines; and stores change. Regardless, once-outer suburbs become inner, and redevelopment pressure grows. The outcome is unpredictable and, in the United States as elsewhere, inner suburbs are varied (Hanlon, 2010). But in terms of density as well as location, age distinguishes suburbs and adjacent fringe areas.

Although individual suburbs may appear residential in character, all have some nonresidential land, commonly parks, schools, churches, retail activity, and recreational facilities. Retailing may be extensive, taking the form of auto-centered malls or big-box centers. Most suburbs also contain office clusters, perhaps lining highways or major arterial roads, and many have light industry. A surprising number of suburban areas, especially those close to airports or major highways, are dominated by light industry, truck and road crew depots, warehouses, and related activities, and so contain more jobs than resident workers. Suburbs, then, may be classified as 'residential' or 'employing,' or by the main type of employment, whether industry, office, or diversified (Harris and Lewis, 2001).

Suburbs were once oriented toward and subsidiary to cities. This is still the case around small- and medium-sized isolated urban centers. Where such centers form regional clusters, and in large metropolitan areas, it is no longer true. Metropolitan development is now multicentered, as suburban Edge Cities form nodes of office, retail, and government employment, or decentered, as jobs scatter, with only minor clusters at highway interchanges (Phelps et al., 2010). Commonly, in varying degrees, these forms of suburbanization coexist.

Large-scale clustering may reflect metro-wide planning, but Edge Cities can emerge spontaneously. At the scale of the individual suburb or subdivision, the role and degree of

planning is more obvious (Relph, 1987). Suburbs planned by a single developer, or tightly regulated by a municipality, have integrated street plans, nonrandom patterns of land use, and some homogeneity in dwelling types and architectural styles. Postwar suburbs tend to boast curvilinear streets, with loops and cul-de-sacs, and large-scale segregation of land uses. More recent suburbs, including Planned Unit Developments and those inspired by New Urbanism, may combine residential and commercial land use on a connected street network. If organized as CIDs, they may feature fences, walls, and gates. Popularized in the United States, and drawing on local traditions, versions have become common in many cities in emerging economies and the global South (Rimmer and Dick, 2009; Fleischer, 2011; Glasze et al., 2006; Grant, 2009; Hirt, 2012).

At the other extreme, unplanned suburbs may feature incomplete street grids, juxtapositions of incompatible land uses, and modest dwellings built of different materials in varied styles. Irregularity is especially true of those that first developed in exurban or peri-urban areas, where regulation is limited or weakly enforced. These display varied orientations and setbacks from the street. In the case of squatter settlements, lot boundaries may be unclear or absent. Unplanned suburbs were common in North America before the 1930s, and persist in outer fringe areas, around smaller urban centers or, as *colonias*, close to the Mexican border. They are common in Japan, in part because of patterns of land ownership (Sorensen, 2001), in parts of southern and central Europe, and throughout the global South, where many began as informal settlements.

Physical contrasts are often associated with social differences. The stereotypical suburb was once exclusive; it is now middle-class (Nicolaides and Wiese, 2006). Both are typically planned. But working-class suburbs, many unplanned, have always been common, historically around European cities such as Paris, Madrid, Rome, and Athens, and now in informal settlements in China, India, sub-Saharan Africa, and Latin America (Benjamin, 2004; Klaufus, 2012; Wu and Shen, in press). The Anglo-American stereotype is that suburbs are for the native-born, as immigrants settle first in city neighborhoods. This was never wholly true: many North American industrial suburbs were first occupied by Irish, southern European, and eastern European immigrants. Today, many of Europe's lower income suburbs house immigrants, especially from North Africa or Turkey; in the global South, rural migrants bring cultures and sometimes languages that locals view as foreign; in North America, the decline of some inner suburbs has enabled many immigrants to settle there; and waves of prosperous immigrants, notably Chinese, have created ethnic enclaves in outer suburbs, such as Toronto's Markham. In cities such as Accra, Ghana, remittances and expatriates are shaping many outer suburbs (Grant, 2009). Various versions of the ethnoburb are now common.

Suburbanism

Faced with such diversity, it may seem impossible to identify a shared 'suburbanism' or way of life. The term was coined in the 1950s by sociologist Sylvia Fava, in counterpoint to the idea that urbanism, or city living, entailed social anonymity and freedom. But the idea that 'suburbanites' have a distinctive way of life is older, dating from late-nineteenth century London, when 'suburbia' referred to both a type of place and a type of person. Ever since, writers have tried to disentangle the interrelation of built environment and social behavior.

To understand suburbanism we must distinguish the popular mythology from the social aspiration and the lived reality. Propagated by academics and the media, the mythology has varied (Clapson, 2003; Nicolaides and Wiese, 2006: 291–320; Palen, 1995). It has commonly referred to social conformity, and with privatized forms of family living at the expense of higher cultural aspirations. In the United States in the 1950s, it apparently meant enforced socializing with neighbors; now, supposedly, the opposite. Such mythologies were often based on the impressions of well-educated and snobbish outsiders. Ethnographic studies provide valuable corrections to stereotypes (Fleischer, 2011; Grant, 2009; Zhang, 2010). Early research in the United States, for example, showed that elements of the supposedly suburban lifestyle were in fact specific to a social class (Gans, 1968).

Mythologies have often been modified or contradicted by studies of people's aspirations in moving to the suburbs, and of their level of satisfaction and pattern of living after their move. People who move voluntarily to the suburbs do so for a common, if variable, mix of reasons (Michelson, 1977). They want more space within, around, and adjacent to their home; they value privacy; and they believe suburbs are healthier and safer for raising children. Their preference is confirmed where suburban schools are perceived to be better. Prestige is sometimes a consideration. There is a widespread preference for home ownership, although sometimes, as in Singapore and Beijing, this takes the form of condominiums (Zhang, 2010). Socially diverse but demographically unrepresentative, suburbanites have moved for particular reasons. If there is a distinctively suburban way of life, it must reflect these facts.

Evidence indicates that suburban residents are satisfied with their environment. They devote above-average time and money to their homes, and in childrearing. Young families have little disposable income, especially those with mortgages. They entertain in and around the home, and spend little on exotic holidays or cultural events. To save money, and/or as a hobby, they undertake DIY (Do-It-Yourself) house repairs and modifications. In low-density, transit-poor suburbs, especially where both parents work, they need cars. Committed to privatized consumption, they resist taxes and many are fiscally conservative. This is especially true of marginal, self-made suburbanites who often deplore state benefits. The conservative political orientation of suburbanites is not inevitable, but some patterns of expenditure and behavior are strongly associated with suburban living.

These patterns of behavior were mostly anticipated and chosen even if, in the case of auto-dependency and long commutes, not always welcomed (Michelson, 1977). However, evidence suggests that the physical environment negatively guides behavior more than residents expected. This is especially true of the time spent driving, to work, to shop, to school, visiting friends, and running errands, commitments that peak when children reach their early teens, years after parents originally planned their move to the suburbs. Research also indicates that suburban environments discourage active transportation (walking, cycling), promote obesity, and have harmful side effects on health. Although

these influences have been best documented with reference to suburbs in the global North, they are valid in varying degrees elsewhere. The suburban way of life is mostly chosen, but partly imposed by the physical environment.

Such impositions are experienced as confining or intolerable by specific groups. Feminists have pointed that, even when they help make the decision to move, women in one-car families find auto-dependent suburbs confining. This is less an issue in two-car households, although because they still bear primary responsibility for shuttling children around, many women experience an extra travel burden. In general, suburbs are often felt by teenagers to be oppressive, whether because of their social conformity or because they are challenging to navigate without a car and drivers' license. The elderly – those who age in place or those who move in with their children – together with the disabled, may also feel constrained and dependent. Although suburbs entail a certain pattern of living for most residents, they can help prevent the remainder from participating in a full social experience.

Consequences

Beyond its effects on social behavior and health, suburbanization has wider consequences. These vary with residential density, being greatest for sprawl. In regions experiencing rapid urban growth, suburbanization is inevitable. Three effects are especially important: economic, environmental, and governmental.

The lower density environments typical of suburbs are more costly to service with roads, tracks, pipes, and wires, as well as with police and fire protection. They also require large expenditures on personal transportation. Their production requires major capital investments; their maintenance and use demands high property taxes and also municipal and personal debts. This is the cost of obtaining the environment many people prefer. However, because of market distortions introduced by government policy, discussed earlier, the density of suburban development is lower, and its costs greater, than is socially optimal. Directed elsewhere, some investment and expenditure currently channeled into suburbs might raise productivity and the standard of living. Against this, evidence suggests that even in the global North, land and housing costs less in lower density metropolitan places. In many cities in the global South, housing conditions would improve if urban areas were less dense (Angel, 2012).

Another type of economic significance arises from linkages between the urban built environment and flows of capital (Harvey, 1978). Built environments are immobile and enduring, tying up capital for decades. In metropolitan areas that stagnate, their devaluation creates problems for families, companies, investors, banks, and municipalities. In rapidly developing areas, problems can also arise. Booms invite speculative investment. When bubbles burst, vacancies and devaluation may bankrupt individuals and lenders alike, especially during periods of overaccumulation when lending institutions have excess funds to invest. Periodic crises are endemic to capitalism, and the scale of fallout has grown with the globalization of capital markets. The financial crisis of 2008, triggered by overinvestment in suburban real estate in Ireland, southern Europe, and above all the United States, underlined the economic ramifications of suburbanization.

The environmental impacts of suburbanization are also clear. They include the loss of productive farmland, as well as threats to biodiversity and ecosystems, notably wetlands and watersheds. These effects are most apparent in rapidly changing peri-urban zones, but they affect all urban residents (Simon, 2008). Some but not all can be mitigated through planning. Their significance varies greatly, depending on local biota and physiography. Many researchers have focused on the effects of low-density development on energy use, carbon emissions, ozone levels, and climate change. By comparison with smaller homes in denser city neighborhoods that are well served by transit, suburban living leaves a larger carbon footprint. But in many regions, suburbanization is caused primarily by rural–urban migration. Depending on the lifestyles of rural residents, their relocation to suburban and urban fringe areas may reduce carbon emissions. The net effects cannot be taken for granted.

Third, apart from raising infrastructure costs, suburbanization challenges suburban governance. As urban settlement expands beyond municipal limits, outer suburban areas require infrastructure and land use planning that existing local governments cannot provide. Where effective suburban governments do exist, they may adopt land use and other regulatory practices that raise housing costs and exclude suburban living from many people. And under all circumstances, the proliferation of suburban jurisdictions makes effective metro-wide planning of land use and transportation more difficult. In principle, planning can be improved through municipal amalgamation or the formation of specialized metro-wide agencies with wide powers. These initiatives are commonly unpopular among suburbanites. Upper levels of government, which have the power to implement them, hesitate to do so.

Suburbanization is inevitable, and often desirable. The practical challenge is to guide it in ways that minimize the negative social, economic, and environmental effects. This requires the cooperation of municipal and upper levels of government, whether produced by rare political vision, grassroots pressure, or a major crisis.

Acknowledgments

Research on which this article is based was supported by the Social Sciences and Humanities Research Council of Canada through funding from the Major Collaborative Research Initiative 'Global Suburbanisms' (2010–17).

See also: Cities: Internal Structure; Gated Communities; Real Estate Development; Urban Informality: The Production and Regulation of Space; Urban Internal Spatial Structure; Urban Planning (Western), History of; Urban Sociology; Urban Sprawl; Urbanization.

Bibliography

Angel, S., 2012. Planet of Cities. Lincoln Institute of Land Policy, Cambridge, MA.
Benjamin, S., 2004. Urban land transformation for pro-poor economies. Geoforum 35, 177–187.

Blais, P., 2010. Perverse Cities. Hidden Subsidies, Wonky Policies, and Urban Sprawl. UBC Press, Vancouver.
Champion, A.G., Hugo, G. (Eds.), 2004. New Forms of Urbanization. Beyond the Rural–Urban Dichotomy. Ashgate, Aldershot, Hants.
Clapson, M., 2003. Suburban Century. Social Change and Urban Growth in England and the USA. Berg, Oxford.
Doebele, W.A., 1977. The private market and low income urbanisation. The 'pirate' subdivisions of Bogotá. American Journal of Comparative Law 25, 531–564.
Dupont, V., 2005. Residential practices, creation and use of urban space. unauthorised colonies in Delhi. In: Hust, E., Mann, M. (Eds.), Urbanization and Governance in India. Manohar, Delhi, pp. 311–341.
Ekers, M., Hamel, P., Keil, R., 2012. Governing suburbia. Modalities and mechanisms of suburban governance. Regional Studies 46, 405–422.
Evans, A., 2004. Economics and Land Use Planning. Blackwell, Oxford.
Fleischer, F., 2011. Suburban Beijing. Housing and Consumption in Contemporary China. University of Minnesota Press, Minneapolis.
Forsyth, A., 2012. Defining suburbs. Journal of Planning Literature 27, 270–281.
Gans, H., 1968. Urbanism and suburbanism as ways of life. In: Gans, H. (Ed.), Essays on Urban Problems and Solutions. Basic Books, New York.
Glasze, G., Webster, C., Frantz, K. (Eds.), 2006. Private Cities. Global and Local Perspectives. Routledge, London.
Grant, R., 2009. Globalising City. The Urban and Economic Transformation of Accra, Ghana. New York University Press, Syracuse, NY.
Hanlon, B., 2010. Once the American Dream. Inner-ring Suburbs of the United States. Temple University Press, Philadelphia.
Harris, R., Lewis, R., 2001. The geography of North American cities and suburbs, 1900–1950. A new synthesis. Journal of Urban History 27, 262–292. Reprinted in Nicolaides and Wiese.
Harris, R., 2010. Meaningful types in a world of suburbs. In: Clapson, M., Hutchison, R. (Eds.), Suburbanization in Global Society, Research in Urban Sociology, vol. 10. Emerald, Bingley, UK, pp. 15–47.
Harvey, D., 1978. The urban process under capitalism. International Journal of Urban and Regional Research 2, 101–131.
Hirt, S., 2012. Iron Curtains. Gates, Suburbs and Privatisation of Space in the Post-socialist City. Wiley, Malden, MA.
Hsing, Y.-T., 2010. The Great Urban Transformation. Politics and Property in China. Oxford University Press, New York.
Klaufus, C., 2012. Urban Residence. Housing and Social Transformations in Globalising Ecuador. Berghahn, New York.
McKenzie, E., 1994. Privatopia. Homeowner Associations and the Rise of Residential Private Government. Yale University Press, New Haven, CT.
Michelson, W., 1977. Environmental Choice, Human Behavior and Residential Satisfaction. Oxford University Press, Oxford.
Nicolaides, B., Wiese, A. (Eds.), 2006. The Suburb Reader. Routledge, New York.
Palen, J.J., 1995. The Suburbs. McGraw-Hill, New York.
Phelps, N.A., Wood, A.M., Valler, D.C., 2010. A post-suburban world? An outline of a research agenda. Environment and Planning A 42, 366–383.
Relph, E., 1987. The Modern Urban Landscape. Johns Hopkins University Press, Baltimore, MD.
Rimmer, P.J., Dick, H., 2009. The City in Southeast Asia. Patterns, Processes and Policy. National University of Singapore Press, Singapore.
Rybczynski, W., 2008. Last Harvest. How a Cornfield Became New Daleville. Scribner, New York.
Schneider, A., Woodcock, C.E., 2008. Compact, dispersed, fragmented, extensive? A comparison of urban growth in twenty-five global cities using remotely-sensed data, patterns metrics and census information. Urban Studies 45, 659–691.
Simon, D., 2008. Urban environments. Issues on the peri-urban fringe. Annual Review of Environment and Resources 33, 167–185.
Sorensen, A., 2001. Building suburbs in Japan. Continuous unplanned change on the urban fringe. Town Planning Review 72, 247–273.
Taylor, L., 2011. No boundaries. Exurbia and the study of contemporary urban dispersion. GeoJournal 76, 303–322.
Teaford, J., 2008. The American Suburb. The Basics. Routledge, New York.
Wu, F., Shen, J. Suburban development and governance in China. In: Hamel, P., Keil, R. (Eds.), Suburban Governance. A Global View. University of Toronto Press, Toronto (in press).
Zhang, L., 2010. In Search of Paradise. Middle Class Living in a Chinese Metropolis. Cornell University Press, Ithaca, NY.

Relevant Websites

http://www.theatlanticcities.com/ – Atlantic Cities. A site that includes summaries of, and links to, current research and commentary on metro areas worldwide, with a U.S. emphasis.

http://www.brookings.edu/research/topics/u-s-metro-areas – Brookings. Authoritative and up-to-date analysis of U.S. metropolitan areas.

http://www.lincolninst.edu/aboutlincoln/planning_urban.asp – Lincoln Institute for Land Policy. Research on urban/suburban land and planning issues, especially in the global South.

http://www.planetizen.com/ – Planetizen. Includes summaries of, and links to, urban research world wide, with an emphasis on U.S. and planning issue.

Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model

Frieder R Lang, Institute of Psychogerontology, University of Erlangen-Nuremberg, Nürnberg, Germany
Margund K Rohr, Institute of Psychogerontology, University of Erlangen-Nuremberg, Nürnberg, Germany; and University of Leipzig, Leipzig, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The processes and outcomes of successful aging are viewed in a life span developmental perspective that emphasizes adaptive person–environment transactions across adulthood and the maximization of gains while minimizing losses. This article is organized around three questions: (1) What is successful aging?, (2) How is successful aging different from normal or pathological aging?, and (3) What are processes that lead to successful aging? – The model of selection, optimization, and compensation points to a standard of optimality in successful life management in old age.

Introduction

The notion of successful aging is a long runner in the field of behavioral and cognitive aging. In more than 60 years, theoretical and empirical research has generated a wealth of findings on adaptive or positive processes of aging. In this vein, there are three pertinent issues: (1) What is the meaning of successful aging, (2) What are criteria that differentiate successful from normal or pathological aging, and (3) What are the processes that bring forth desirable outcomes across adulthood.

Since the early beginnings of empirical research on successful aging, the term was used to describe the ways through which people "become happier, more successful in their pursuit of the goals of life, or better adjusted" (Havighurst and Orr, 1955, p. 477). Already then, 60 years ago, it was widely accepted that successful aging refers to a process rather than merely reflecting an outcome in old age. Today there is a strong agreement that successful aging involves adaptive processes in response to the changing gain–loss dynamics across adulthood that involve a maximization of gains while minimizing potential losses (Baltes and Baltes, 1990; Baltes and Carstensen, 1996; Marsiske et al., 1995).

In addition, successful aging may not be limited to positive subjective experience in old age but also pertains to objective conditions such as health, competence, and functioning (Palmore, 1978; Rowe and Kahn, 1987). However, it is still debated how different criteria of success may be identified in the context of multifaceted challenges in the aging process. For example, in an early study on successful aging Havighurst and Albrecht (1953) observed that good health did not guarantee high levels of social activity or positive psychological adjustment. In the biomedical sciences though, successful aging has been more strongly related to good physical and mental health (Rowe and Kahn, 1987), and also to a high level of functioning of the older individual. In more recent theoretical approaches, there is agreement that successful aging requires both subjective and objective sources of evaluation that rely on age-adjusted norms of judgment (cf Baltes and Carstensen, 1996; Lang et al., 2011; Freund et al., 2013). In line with this, we discuss processes and outcomes of successful aging in a life span perspective that emphasizes adaptive person–environment transactions across adulthood and the maximization of gains while minimizing losses in the course of becoming old. In the following, we organize this article around the following three central questions:

1. What is successful aging?
2. How is successful aging different from normal or pathological aging?
3. What are the processes that lead to successful aging?

What is Successful Aging?

Since the year 1961, about 350 articles and books abstracted in PsycInfo (American Psychological Association) and more than 400 articles in Medline included the term successful aging in the publication title, not counting articles on 'positive aging' or on 'aging well.' In the literature, there exist many definitions of successful aging, and even more heterogeneity with regard to the measurement of successful aging. According to modern dictionaries of everyday language (cf Merriam-Webster Dictionary, Oxford Advanced Learner's Dictionary), *success* is generally defined as an attainment of desired or ideal states, events, or outcomes, that is, wealth, health, or high social status. In the life span sciences success of aging goes beyond such limited views to also reflect the individual's capacity of adaptation and mastery of challenges. In specific, we argue that successful aging pertains to establishing a best possible (i.e., *optimal*) match between an individual's capacities or resources and the individual's contextual demands and opportunities (Baltes and Carstensen, 1996; Baltes and Lang, 1997; Lang et al., 2011; Marsiske et al., 1995). For instance, a just widowed older woman may seek to refresh old social bonds, intensify existing close ties, or seek for new social contact in order to master her personal loss (cf Zettel and Rook, 2004). In general, adaptation to aging-related losses may depend both on personal resources (e.g., sensory functioning, being extra- or introverted) as well as on contextual resources (e.g., access to supporting devices, availability of supporting infrastructure such as access to Internet or the existence of meeting places or self-help groups in the neighborhood).

Most scholars agree that subjective and objective criteria should be integrated in some way when evaluating the aging process (Bowling, 2006, 2007; Depp and Jeste, 2006; Pruchno et al., 2010). However, there is disagreement regarding the relevance and role of subjective evaluations in the process of aging (Brandtstädter and Greve, 1994; Freund et al., 2013; Lang et al., 2011). Surprisingly enough, in their early classic study, Havighurst and Albrecht (1953) defined successful aging as subjective evaluation of the personal aging experience. However, such subjective evaluations of the self may be interpreted in different ways.

Some authors argued that subjective self-appraisals serve to protect the ability to exert control, and may not be used as criteria of successful aging 'in their own right' (Schulz and Heckhausen, 1996, p. 705). Other authors claimed that subjective criteria reflect a positive adaptation of the self that need to be respected even when appearing to be in contradiction to more objective conditions of individuals' aging situation (Brandtstädter and Greve, 1994; Schulz and Hanusa, 1980). In the first perspective, the subjective evaluation of one's aging experience points to potential processes of adaptation. In the latter perspective, the subjective evaluation of an individual's aging situation is seen as an outcome of an adaptation process. Obviously, both perspectives are complementary rather than contradictory. Subjective evaluations of the aging self-reflect both functional processes as well as adaptive outcomes in the process of aging. Successful aging may follow from a correspondence between subjective standards and objective constraints (e.g., happiness in low income contexts; Chung and Park, 2008). Also, successful aging may occur even after stroke when an older person has adjusted well psychologically and physically in a life with handicaps and limited autonomy (Donnellan et al., 2012). This means that successful aging may not be limited to the pursuit of happiness or life goals, but rather should also reflect an *optimality* norm. Optimality, in this vein, is defined as a standard of *best practice* given the individual's available resources and contextual challenges.

Many theories on successful aging emphasize contextual demands and constraints associated with adaptation in later adulthood. According to such theoretical approaches, successful aging is reflected in the ecology of the older individual and the proactive coconstruction of person–environment fit (Lang et al., 2011; Kahana and Kahana, 1996; Schulz and Hanusa, 1980). This means that definitions of what is positive, adequate, or functional in old age should consider the respective contextual challenges and resources of the aging person. For example, the ability to drive a car may be more essential in a rural area as compared to a metropolitan area where persons may rely on a more developed public transportation system or a better infrastructure. Conditions of successful aging may differ between cultures, but the contextual demands may at times differ strongly within one culture. In this vein, images of successful aging, and the ascribed responsibility for aging well may vary across cultures and cohorts (e.g., Levy and Langer, 1994; Hess, 2006; Kotter-Grühn and Hess, 2012).

On a more general level, the ecology of aging may often entail contextual resources that serve to activate, enhance, and maintain adaptation in the process of successful aging. Illustrations for such contextual regulation of adaptation in old age pertain to social, medical, and technological advancements in modern societies, ranging from improved health care services to new prosthetic technological devices (Lang et al., 2011; Lindenberger et al., 2008). For instance, in most aging societies new societal structures evolve that provide novel options for social participation, mobility, and education for older adults. New opportunity structures generate proactivity in the selection of goal engagement or the refinement and exercising of competence, in addition to prosthetic or supportive environments. One way of how the culture and language scaffolds successful aging may be illustrated with the proverb study of Freund and Baltes (2002). Proverbs reflect culturally proliferated thumb rules on successful life management. Some of the proverbs were identified as reflecting adaptive life management strategies (e.g., 'a rolling stone gathers no moss'; 'the early bird catches the worm'), while other proverbs were more neutral in this sense (e.g., 'out of sight, out of mind'). Consistent with the expectation, people of all ages preferred proverbs that were identified as capturing adaptive life management strategies. Thus, the study underscores that on the cultural level individuals appear to be provided with functional knowledge on successful aging as embedded in social norms.

Another example illustrates individual efforts in creating the best possible positive match with existing resources, desires and needs, and adapting to certain contextual constraints. In her classical observation studies in the context of nursing homes, Margret Baltes (1996) reported a dependency-support script: According to these classic findings, nursing home staff reinforced dependent behavior while generally ignoring independent behavior of residents. Consequently, nursing home residents were found to behave more depending on staff, and less competent, which increased the likelihood of obtaining attention and contact of staff. After a retraining of nursing home staff to better acknowledge independent behaviors, this pattern was reversed (Baltes et al., 1994). Such findings also serve to illustrate that even late in life older adults have much adaptive competence for mastery of environmental demands. Again, such illustration from the nursing home context implies that successful aging may also be defined with regard to a standard of optimality.

How Does *Successful* Aging Differ from Normal and Pathological Aging?

A critical issue in research on successful aging pertains to the evaluation of patterns of stability or change in old age. There is much agreement that growth, stability, and decline may co-occur at the same time across different behavioral and psychological functions. For example, a person may experience a personal loss, while having stable self-esteem, and feeling even an increasing sense of control. The evaluation of successful aging may depend on functional (e.g., tests of everyday competence) or on statistical norms (e.g., divergence from average). In this vein, a critical issue pertains to discontinuity in age-related changes (e.g., Carlesimo et al., 1998). Discontinuity typically involves sudden changes, for example, related to a fall, a stroke, a loss, a failure, or an external event. In general, discontinuity in the

process of aging is more likely associated with undesirable or decline aspects of aging, whereas continuity or stability in the course of aging is more likely associated with positive aging (Gerstorf et al., 2006). The differentiation of successful to normal and pathological aging is strongly related to three fundamental aspects of the aging phenomenon, that is, (a) multidimensionality and multidirectionality, (b) the dynamics of gains and losses, and (c) plasticity (Baltes, 1987).

First, aging as a multidirectional and *multidimensional process* implies that success, stability, and pathology may alternate in time, and may even coexist within one individual across different domains of functioning (e.g., Morack et al., 2013). For instance, there is ample evidence that people can maintain high levels of life satisfaction and view themselves as successful agers even when facing chronic illness and age-related health impairments (Montross et al., 2006; Strawbridge et al., 2002). Multidimensionality reflects intraindividual differentiation across and within domains of functioning, while decline in the process of aging was observed to be associated with progressing dedifferentiation (Lindenberger and Baltes, 1997). It may be concluded that the observation of multidimensionality across adulthood is a critical premise for positive aging.

Second, life span psychology has put forth the understanding that there is *no gain without a loss*, and that there is no loss without a gain (Baltes, 1987). For example, advancement and progression often involve moving on and *away from* previous states (i.e., leaving behind something). Moreover, many investments in life (e.g., exercise, effort, acquiring new skills) imply that some of the invested resources (e.g., time, physiological functions) are depleted. And also, any loss may provide an opportunity for experiencing further personal growth (Baltes, 1987; Ryff, 1989). Thus, successful aging was defined as a balancing of gains and losses (Baltes and Carstensen, 1996). In this vein, studies demonstrate that the acknowledgment of meaning and personal growth is linked to positive outcomes (e.g., Levy et al., 2002 for the impact of positive aging perceptions; Zacher and Frese, 2011 for the work context). However, this notion involves that success may involve experiences of loss or failure. Here again, successful aging pertains to the standard of optimality, that is, a definition of what is the best possible adaptation in a given situation. For example, Ebner et al. (2006) were able to demonstrate that a switch in goal strivings may be functional across adulthood. In their study, young adults endorsed more growth-oriented goals (e.g., becoming healthy), whereas older adults rather preferred goals with a maintenance focus (e.g., staying healthy). Such age differences in goal focus were associated with well-being, and thus reflect an adaptation to the individual's available resources.

Third, aging entails much heterogeneity and malleability until very late in life. *Behavioral plasticity* refers to the flexibility of the self, as well as to cognitive and functional reserve capacities. This means that older individuals may be able to improve their abilities or even acquire a new skill even in the last years of life. In general, individuals are always capable to adapt to new environmental challenges. However, there also exist limits of such plasticity. While it is a robust finding that cognitive performance can be improved through training, it was also observed that there are age-associated limits in the best possible training performance (Baltes et al., 1995; Zinke et al., 2012). Furthermore, individual resources as well as contextual surroundings also enhance or hinder one's ability to act flexibly and proactively to age-related demands (Baltes and Lang, 1997; Lang et al., 2002). Plasticity (in its limits) may be seen as an important concomitant of successful aging, while the absence of plasticity points to less positive trajectories of aging. Consequently, behavioral plasticity constitutes, an important process characteristic of successful aging.

Together, multidimensionality, gain–loss dynamics, and behavioral plasticity point to the individual's potentials to co-organize and coconstruct positive aging processes within the boundaries of biological and cultural contexts. This conception of optimal aging builds on the idea that individuals are capable of intentional self-development and of proactive adaptation to changing life span contexts, which involves the question of what are the processes that contribute to successful aging.

What are Processes of Successful Aging? Selection, Optimization, and Compensation

Grounding on a wealth of theoretical and empirical findings regarding the potential and course of aging, Paul Baltes and Margret Baltes (1980, 1990) proposed the model of selection, optimization, and compensation (SOC) as a theoretical framework of successful aging. According to this model there are three interrelated developmental principles of behavioral adaptation and mastery in response to biocultural challenges in the aging process: *selection* as leading principle of behavioral choice, *optimization* as principle of enhancement and refinement, and *compensation* as principle of counterbalancing any kind of loss. In a first publication on this model, Baltes and Baltes (1980) introduced the model of SOC as follows: "*The adaptive task […] for the aging individual is to select and concentrate on those domains which are of high priority and involve a convergence of environmental demands, individual motivations and skills, and biological capacity*" (p. 59). In their work, Margret and Paul Baltes together with their colleagues have provided ample evidence that the processes of SOC contribute to proactivity and improved person–environment fit in later adulthood.

For each of the three adaptive processes the model assumes contextual conditions that may activate the respective behavioral mechanism. **Table 1** gives an overview on the contextual conditions, definitions, and involved behavioral processes involved in each of the three principles of SOC. In the following we shortly describe each of the three principles, and give empirical illustrations for how such successful aging principles may contribute to positive aging outcomes.

Selection encompasses strategies that aim at the reduction of and at focusing on personal meaningful goals or behavioral domains. At the same time, selection implies turning away one's efforts or behavioral resources from alternatives. This may include an adaption of goals, or an investment in new goals. Selection occurs in response to three premises of the life course: (1) opportunity structures provide options for differing individual pathways, (2) remaining time in life is limited, and (3) behavioral resources are exhaustible (e.g., health, mobility, cognitive performance). Two modes of the mental organization

Table 1 Definition, preconditions, and behavioral processes of selection, optimization, and compensation

Principle	*Contextual precondition*	*Definition*	*Behavioral process*	*Illustrations*
Selection	● Limitation of resources ● Finitude of life time ● Opportunity structures	Canalization of pathways and trajectories Emergence of opportunity structures	Mental organization of choices (tasks or goals); in response to loss (*loss-based S.*) or to positive events or outcomes (*elective S.*)	Giving direction to efforts or investments, commitment, abandonment, renewal of goals.
Optimization	● Amendability of resources ● Efficiency of resources ● Plasticity	Improving the function of a system or organism	Enhancing, refining, or strengthening a behavioral function (i.e., resources required for cognitive or physical activity)	Exercise, timing of activity, increased effort, cognitive motor control, solicit advice, generate information or new opportunities
Compensation	● Error-proneness (susceptibility to loss) ● Available, not used resources or opportunities	Replacing or restoring lost or failed functions Balancing failure or loss	Establishing new or restored behavioral functions in response to loss, limits, or failure (i.e., activating or using idle resources)	*Note: in response to anticipated future loss* (= *loss-based selection*), *In response to actual failure/loss*: Disengagement, motivation protection

Adapted from Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, NY, USA; Lang, F.R., Rohr, M.K., Williger, B., 2011. Modeling success in life-span psychology: the principles of selection, optimization, and compensation. In: Fingerman, K.L., Berg, C.A., Antonucci, T., Smith, J. (Eds.), Handbook of Lifespan Development. Springer, New York; and Marsiske, M., Lang, F.R., Baltes, P.B., Baltes, M.M., 1995. Selective optimization with compensation: life-span perspectives on successful human development. In: Dixon, R.A., Bäckman, L. (Eds.), Compensation for Psychological Defects and Declines: Managing Losses and Promoting Gains. Lawrence Erlbaum Associates.

of choices in life may be distinguished, that is, elective selection as a deliberate response to having options, and loss-based selection as a response to the limitation of resources.

Optimization captures efforts to enhance means and resources in order to reach one's desired goals. This process is highly dependent on the type and the amount of resources that an individual is able to invest. Clearly, optimization may also adhere to standards or norms of *optimality*, defined as a *best practice* in a given situation or context. Optimization pertains to those behavioral and contextual resources that are efficient, malleable, and amendable. As long as there is behavioral plasticity, individuals may refine, enhance, or strengthen positive aging outcomes. A debated issue in this vein is whether behavioral optimization requires goal pursuit or may also occur automatically (e.g., effortful activity without purpose; Lang et al., 2011).

Compensation occurs in the face of losses and constraints. It refers to the search and use of new means or resources that allow substitution of loss and therewith contribute to the stability or maintenance of performance. In line with the person–environment fit as criterion of successful aging two sources of compensatory efforts are possible, namely the increase of internal and external demands (cf Baltes and Baltes, 1990; Lang et al., 2011).

For instance, a person who notices memory problems may prefer to spend time in familiar contexts (selection), continue to invest in an active lifestyle (optimization), and use aids such as shopping lists (compensation). In the following we give a few notable empirical illustrations with regard to the social behavior, the pursuit of goals, and the use of technology in everyday life.

Empirical illustrations of SOC can be found in research on the formation of social relationships across adulthood (Lang, 2001). It is a robust finding that social networks tend to become smaller in old age. However, a closer look has revealed that such reduction of social networks was not just due to uncontrollable reasons such as increased immobility, health impairments, or the morbidity and mortality of social partners, but also reflects older adults' active attempts of matching resources, goals, and needs with contextual demands and challenges. More concretely, the reduction in network size is often restricted to peripheral social partners and acquaintances, but at the same time an increasing importance of close contacts and an intensification of family relationships is observable (Antonucci and Akiyama, 1987; Lang, 2001; Lang and Carstensen, 2002). Furthermore, individuals strive for positive social contacts, try to adapt social networks to their needs and resources, and proactively attempt to balance potential downsides and challenges of interpersonal relationships (Rohr and Lang, 2009). As mentioned earlier, the importance of long-lasting and close ties is increasing across life span and individual seek to maximize positive and emotional meaningful interactions (Birditt et al., 2005; Fredrickson and Carstensen, 1990; Lang, 2001). The enhancement of emotional needs in the regulation of one's social world is known to be associated with positive aging outcomes such as improved well-being (Lang and Schütze, 2002). Moreover, empirical research points to the benefits of social context for health and autonomy, physical activity, and for nutritional behavior (cf Rohr and Lang, 2009). In a series of studies, Levenson and colleagues demonstrated that older adults seem to be better in balancing relationship conflicts by using less escalating strategies, staying calmer, and expressing more affection to the partner (e.g., Carstensen et al., 1995; Levenson et al., 1993).

A second set of examples stems from research on the development of personal goals across the life span. In an impressive series of studies on the pursuit of goals across adulthood, Freund and her colleagues (e.g., Freund, 2006; Freund and Baltes, 2002; Freund et al., 2013; Riediger and

Freund, 2006) addressed the action-theoretical foundations of successful aging with regard to SOC. For example, Riediger and Freund (2006) explored whether similarity and ego-centrality of goals and the number of goals contributed to goal pursuit. Participants reported personal goals, which were rated for centrality (central vs marginal) and similarity (similar, belonging to one are of life vs dissimilar, spanning several life domains). Greater goal focus was positively associated with persistence and intensity of goal pursuit 3 months later. In another study (Freund, 2006), participants completed a sensorimotor working task in an optimization condition (i.e., improving performance, "try to get as good as possible"), or in a compensation condition (i.e., responding to loss, "try to get back to the level of performance in the first session"). Young adults were more persistent with the task in optimization instruction, while older adults spent more time on the task in the compensation condition.

Third, the model of SOC also contributes to improved understanding of how the use and adoption of technical solutions may improve successful aging outcomes. Based on the model of successful aging, Lindenberger et al. (2008) extracted three principles that enhance the usability of technical devices. First, technical solution should set free more resources than are needed for their use. Second, they should allow person-specific use, i.e., be adaptable to one's biography, needs, and competencies – and to individual changes in these aspects over time. Third, technological solutions should be assessed regarding their short-term and long-term benefits and risks. The aim should be that risks counteract benefits while the difficulty is to be aware of the principles of multidimensionality and plasticity. Melenhorst et al. (2006) underscore the importance of gain–loss dynamics for use of technical innovations in old age. In focus-groups, older adults aged 65–80 years reported to rely on anticipated gains in order to decide about using or not using a newly developed communication media device. In contrast, potential costs of use or nonuse seem less important in the selection and adoption process.

In general, research on successful aging puts forward the benefits of proactivity (e.g., Kahana et al., 2002; Klusmann et al., 2012). For instance, using longitudinal data from the Aging in Manitoba Study Menec (2003) showed that active engagement was linked to outcomes of successful aging such as happiness, life satisfaction, and physical functioning. However, each of these strategies goes along with chances and risks and functionality clearly depends on the person–environment fit. For instance, concentrating on a goal and dropping others (e.g., 'I want to stay mobile'), investing in a specific training (e.g., going to the gym), or choosing a certain way to reach one's goals (e.g., use of a wheeled walker) may also lead to potential misfits and failure (e.g., disease-related impairments such as hemiplegia, missing physical resources, being confronted with stairs in the house). Moreover, individuals who can draw on multiple resources such as sensorimotor and cognitive capabilities or social contacts did not only adapt better to age-related challenges, but were also able to delay functional losses by using strategies more efficiently (Baltes and Lang, 1997).

Conclusions and Outlook

Successful aging results from a fit between personal capacities and environmental demands. Individuals are proactive in selecting and shaping their environments in accordance with personal goals, needs, and capacities in order to obtain positive aging outcomes related to well-being and health. In line with this we proposed an *optimality* norm as a process-oriented criterion of successful aging.

The model of SOC includes three fundamental principles contributed to adaptive mastery and competence in the aging process. All together the model of SOC contributes importantly to improved understanding of how individuals may achieve successful aging outcomes even when being confronted with functional or personal loss experience. The principles of SOC were shown to be associated with improved functioning and strong well-being.

However, predictions of the model of SOC still lack further evidence from short- and long-term longitudinal data that empirically demonstrate the dynamic changes in the interplay of the three principles, and their impact on adaptation over time (cf Boker, 2013). Moreover, more research on cultural, societal, and historical influences on the processes of SOC is needed to better understand the contextual challenges in the aging process.

In all, the model of SOC serves to improve understanding of successful aging in a process perspective. SOC are ways through which individuals achieve improved and optimal person–environment fit in late, and thus contribute to positive aging outcomes in cognitive, behavioral, and emotional domains of functioning in later life.

See also: Affect-Regulation Motivation; Motivation and Actions, Psychology of; Personal Projects; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Self-Regulation in Adulthood.

Bibliography

Antonucci, T.C., Akiyama, H., 1987. Social networks in adult life and a preliminary examination of the convoy model. Journal of Gerontology 42, 519–527.

Baltes, M.M., 1996. The Many Faces of Dependency in Old Age. Cambridge University Press, New York, NY, USA.

Baltes, M.M., Carstensen, L.L., 1996. The process of successful aging. Aging & Society 16, 397–422.

Baltes, M.M., Kühl, K.-P., Gutzmann, H., Sowarka, D., 1995. Potential of cognitive plasticity as a diagnostic instrument: a cross-validation and extension. Psychology and Aging 10, 167–172.

Baltes, M.M., Neumann, E.-M., Zank, S., 1994. Maintenance and rehabilitation of independence in old age: an intervention program for staff. Psychology and Aging 9, 179–188.

Baltes, M.M., Lang, F.R., 1997. Everyday functioning and successful aging: the impact of resources. Psychology and Aging 12, 433–443.

Baltes, P.B., 1987. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23, 611–626.

Baltes, P.B., Baltes, M.M., 1980. Plasticity and variability in psychological aging: methodological and theoretical issues. In: Gurski, G.E. (Ed.), Determining the Effects of Aging on the Central Nervous System. Schering, Berlin.

Baltes, P.B., Baltes, M.M., 1990. Psychological perspectives on successful aging: the model of selective optimization with compensation. In: Baltes, P.B., Baltes, M.M. (Eds.), Successful Aging: Perspectives from the Behavioral Sciences. Cambridge University Press, New York, NY, USA.

Birditt, K.S., Fingerman, K.L., Almeida, D.M., 2005. Age differences in exposure and reactions to interpersonal tensions: a daily diary study. Psychology and Aging 20, 330–340.

Boker, S.M., 2013. Selection, optimization, compensation, and equilibrium dynamics. GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 26, 61–73.

Bowling, A., 2006. Lay perceptions of successful aging: findings from a national survey of middle aged and older adults in Britain. European Journal of Aging 3, 123–136.

Bowling, A., 2007. Aspirations for older age in the 21st century: what is successful aging? The International Journal of Aging & Human Development 64, 263–297.

Brandtstädter, J., Greve, W., 1994. The aging self: stabilizing and protective processes. Developmental Review 14, 52–80.

Carlesimo, G.A., Mauri, M., Graceffa, A.M., Fadda, L., Loasses, A., Lorusso, S., Calagirone, C., 1998. Memory performances in young, elderly, and very old healthy individuals versus patients with Alzheimer's disease: evidence for discontinuity between normal and pathological aging. Journal of Clinical and Experimental Neuropsychology 20, 14–29.

Carstensen, L.L., Gottman, J.M., Levenson, R.W., 1995. Emotional behavior in long-term marriage. Psychology and Aging 10, 140–149.

Chung, S., Park, S.-J., 2008. Successful aging among low-income older people in South Korea. Aging & Society 28, 1061–1074.

Depp, C.A., Jeste, D.V., 2006. Definitions and predictors of successful aging: a comprehensive review of larger quantitative studies. American Journal of Geriatric Psychiatry 14, 6–20.

Ebner, N.C., Freund, A.M., Baltes, P.B., 2006. Developmental changes in personal goal orientation from young to late adulthood: from striving for gains to maintenance and prevention of losses. Psychology and Aging 21, 664–678.

Donnellan, C., Hevey, D., Hickey, A., O'neill, D., 2012. Adaptation to stroke using a model of successful aging. Aging, Neuropsychology, and Cognition 19, 530–547.

Fredrickson, B.L., Carstensen, L.L., 1990. Choosing social partners: how old age and anticipated endings make people more selective. Psychology and Aging 5, 335–347.

Freund, A.M., 2006. Age-differential motivational consequences of optimization versus compensation focus in younger and older adults. Psychology and Aging 21, 240–252.

Freund, A.M., Baltes, P.B., 2002. The adaptiveness of selection, optimization, and compensation as strategies of life management: evidence from a preference study on proverbs. Journals of Gerontology: Psychological Sciences 57, P426–P434.

Freund, A.M., Nikitin, J., Riediger, M., 2013. Successful aging. In: Lerner, R.M., Easterbrooks, M.A., Mistry, J., Weiner, I.B. (Eds.), Handbook of Psychology, Developmental Psychology, second ed., vol. 6. John Wiley & Sons Inc., Hoboken, NJ, USA.

Gerstorf, D., Smith, J., Baltes, P.B., 2006. A systemic-wholistic approach to differential aging: longitudinal findings from the Berlin Aging Study. Psychology and Aging 21, 645–663.

Havighurst, R.J., Albrecht, R., 1953. Older People. Longmans, Green, New York.

Havighurst, R.J., Orr, B.E., 1955. Aging and personality development. Review of Educational Research 25, 477–486.

Heckhausen, J., Schulz, R., 1995. A life-span theory of control. Psychological Review 102, 284–304.

Hess, T.M., 2006. Attitudes toward aging and their effects on behavior. In: Birren, J.E., Schaire, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Elsevier, Amsterdam, The Netherlands.

Kahana, E., Lawrence, R.H., Kahana, B., Kercher, K., Wisniewski, A., Stoller, E., Tobin, J., Stange, K., 2002. Long-term impact of preventive proactivity on quality of life of the old-old. Psychosomatic Medicine 64, 382–394.

Kahana, E., Kahana, B., 1996. Conceptual and empirical advances in understanding aging well through proactive adaptation. In: Bengtson, V.L. (Ed.), Adulthood and Aging. Research on Continuities and Discontinuities. Springer, New York, pp. 18–40.

Klusmann, V., Evers, A., Schwarzer, R., Heuser, I., 2012. Views on aging and emotional benefits of physical activity: effects of an exercise intervention in older women. Psychology of Sport and Exercise 13, 236–242.

Kotter-Grühn, D., Hess, T.M., 2012. The impact of age stereotypes on self-perceptions of aging across the adult lifespan. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 67B, 563–571.

Lang, F.R., 2001. Regulation of social relationships in later adulthood. Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 56B, P321–p326.

Lang, F.R., Carstensen, L.L., 2002. Time counts: future time perspective, goals, and social relationships. Psychology and Aging 17, 125–139.

Lang, F.R., Rieckmann, N., Baltes, M.M., 2002. Adapting to aging losses: do resources facilitate strategies of selection, compensation, and optimization in everyday functioning? The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 57, P501–P509.

Lang, F.R., Rohr, M.K., Williger, B., 2011. Modeling success in life-span psychology: the principles of selection, optimization, and compensation. In: Fingerman, K.L., Berg, C.A., Antonucci, T., Smith, J. (Eds.), Handbook of Lifespan Development. Springer, New York.

Lang, F.R., Schütze, Y., 2002. Adult children's supportive behaviors and older parents' subjective well-being: a developmental perspective on intergenerational relationships. Journal of Social Issues 58, 661–680.

Levenson, R.W., Carstensen, L.L., Gottman, J.M., 1993. Long-term marriage: age, gender, and satisfaction. Psychology and Aging 8, 301–313.

Levy, B., Langer, E., 1994. Aging free from negative stereotypes: successful memory in China among the American deaf. Journal of Personality and Social Psychology 66, 989–997.

Levy, B.R., Slade, M.D., Kasl, S.V., 2002. Longitudinal benefit of positive self-perceptions of aging on functional health. Journals of Gerontology: Series B: Psychological Sciences and Social Sciences P409–p417.

Lindenberger, U., Baltes, P.B., 1997. Intellectual functioning in old and very old age: cross-sectional results from the Berlin Aging Study. Psychology and Aging 12, 410–432.

Lindenberger, U., Lövdén, M., Schellenbach, M., Li, S.-C., Krüger, A., 2008. Psychological principles of successful aging technologies: a mini-review. Gerontology 54, 59–68.

Marsiske, M., Lang, F.R., Baltes, P.B., Baltes, M.M., 1995. Selective optimization with compensation: life-span perspectives on successful human development. In: Dixon, R.A., Bäckman, L. (Eds.), Compensation for Psychological Defects and Declines: Managing Losses and Promoting Gains. Lawrence Erlbaum Associates.

Melenhorst, A.-S., Rogers, W.A., Bouwhuis, D.G., 2006. Older adults' motivated choice for technological innovation: evidence for benefit-driven selectivity. Psychology and Aging 21, 190–195.

Menec, V.H., 2003. The relation between everyday activities and successful aging: a 6-year longitudinal study. The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 58B, S74–S82.

Montross, L.P., Depp, C., Daly, J., Reichstadt, J., Golshan, S., Moore, D., Sitzer, D., Jeste, D.V., 2006. Correlates of self-rated successful aging among community-dwelling older adults. The American Journal of Geriatric Psychiatry 14, 43–51.

Morack, J., Ram, N., Fauth, E.B., Gerstorf, D., 2013. Multidomain trajectories of psychological functioning in old age: a longitudinal perspective on (uneven) successful aging. Developmental Psychology 49, 2309–2324.

Palmore, E., 1978. Predictors of successful aging. The Gerontologist 18, 110–111.

Pruchno, R.A., Wilson-Genderson, M., Rose, M., Cartwright, F., 2010. Successful aging: early influences and contemporary characteristics. The Gerontologist 50, 821–833.

Riediger, M., Freund, A.M., 2006. Focusing and restricting: two aspects of motivational selectivity in adulthood. Psychology and Aging 21, 173–185.

Rohr, M.K., Lang, F.R., 2009. Aging well together – a mini-review. Gerontology 55, 333–343.

Rowe, J.W., Kahn, R.L., 1987. Human aging: usual and successful. Science 237, 143–149.

Ryff, C., 1989. In the eye of the beholder: views of psychological well-being among middle-aged and older adults. Psychology and Aging 4, 195–210.

Schulz, R., Hanusa, B.H., 1980. Experimental social gerontology: a social psychological perspective. Journal of Social Issues 36, 30–46.

Schulz, R., Heckhausen, J., 1996. A life span model of successful aging. American Psychologist 51, 702–714.

Strawbridge, W.J., Wallhagen, M.I., Cohen, R.D., 2002. Successful aging and well-being: self-rated compared with Rowe and Kahn. The Gerontologist 42, 727–733.

Zacher, H., Frese, M., 2011. Maintaining a focus on opportunities at work: the interplay between age, job complexity, and the use of selection, optimization, and compensation strategies. Journal of Organizational Behavior 32, 291–318.

Zettel, L.A., Rook, K.S., 2004. Substitution and compensation in the social networks of older widowed women. Psychology and Aging 19, 433–443.

Zinke, K., Zeintl, M., Eschen, A., Herzog, C., Kliegel, M., 2012. Potentials and limits of plasticity induced by working memory training in old-old age. Gerontology 58, 79–87.

Suicide

Thomas Bronisch, Max-Planck-Institute of Psychiatry, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term suicide is used in the international scientific literature to avoid judgment as do terms like self-murder and voluntary death. Suicide belongs to the 10 most frequent causes of death among adults in industrialized European and North American countries. There is a considerable increase of suicides in societies with a high degree of secularization, industrialization, and individualization starting in the middle of the nineteenth century. The most important motive for suicidal behavior is certainly the impending loss of a human relationship, even if the wish to die outweighs the wish to survive in the suicidal act, i.e., in the seriousness of the suicide attempt. Biological studies could demonstrate a dysfunction of the serotonergic system in suicide victims and suicide attempters, which in psychopathological terms is correlated with a loss of impulse control, aggressiveness, affective instability, and violent suicide attempt. Empirical data about successful prevention of suicidal behavior is still lacking and successful therapeutic interventions only exist for subgroups of suicide attempters such as depressives with a treatment with lithium and mood stabilizers as well as cognitive behavioral therapy and psychodynamic psychotherapy for patients with borderline personality disorder.

Definition

The term suicide was mentioned for the first time in medieval times in 1177 (van Hooff, 1990) and comes from the Latin words *sui cidium* (self-killing) and *sui caedere* (to kill himself or herself). Now, this term is used in the international scientific literature. It avoids judgment, as do terms like 'self-murder' and 'voluntary death.'

In the realm of the term suicide, there are other terms that have to be defined. Suicide ideas include thinking about death either of one's own or someone else's or of death in general, the wish to die, and suicide ideas in a more narrow sense. It is not easy to define the term suicide attempt. The definition most commonly accepted is the following: "An act with nonfatal outcome, in which an individual deliberately indicates a non-habitual behavior that, without the intervention of others, will cause self-harm, or deliberately ingests a substance in excess of the prescribed or generally recognized therapeutic dosage, and which is aimed at realizing changes which the subject desired via the actual or expected physical consequences" (Platt et al., 1992). This definition includes an active intention for self-harm but not necessarily self-killing. Suicide and suicide attempt are also considered as suicidal behavior.

Importance of Suicide

Suicide belongs to the 10 most frequent causes of death among adults in industrialized European and North American countries. In these countries, suicide is considered the second most frequent cause of death among adolescents and young adults (15–35 years). Suicide in young male adults has increased constantly during the past 20 years. **Table 1** shows the frequencies of causes of death worldwide (WHO, 1999) disregarding the most frequent causes of death such as cardiovascular diseases, stroke, cancer, infectious diseases, and malnutrition. These figures have to be taken *cum grano salis* because of a dearth of inaccurate statistics on death in developing countries. As shown in **Table 1**, only fatal car accidents had a similar frequency of death as compared with suicide rates. War activities, crime, or even acquired immune deficiency syndrome at the beginning of the 1990s had a considerably lower death rate.

Epidemiology

The most recently available suicide rates from all over the world are depicted in **Figure 1** (see Schmidtke et al., 1999). This shows that a high variability of suicide rates exists across cultures and countries. Looking at the rank order of the frequency of suicide rates, a relatively stable pattern can be described – at least for European countries. There exists a North–South downgradient and an East–West downgradient of suicide rates: eastern and northern countries have higher suicide rates than western and southern countries. Hereby, sociological factors seem to play an important role in terms of secularization and industrialization with the collapse of the family group as a consequence as well as climate factors and genetic underpinnings, which cannot be disentangled sufficiently.

There exist distinct sex differences with a preponderance of males in suicide and a preponderance of females in suicide attempt. This distribution is stable across all countries and cultures. Only China reports a higher prevalence of suicide among women in selected areas.

The highest rate of suicide is found in people of older age and that of suicide attempt is found in younger people.

Table 1 WHO statistics: causes of death worldwide (1999)

856 000	*Death because of car accidents*
818 000	Death because of suicide
322 000	Death because of war
292 000	Death because of crime
291 000	Death because of acquired immune deficiency syndrome

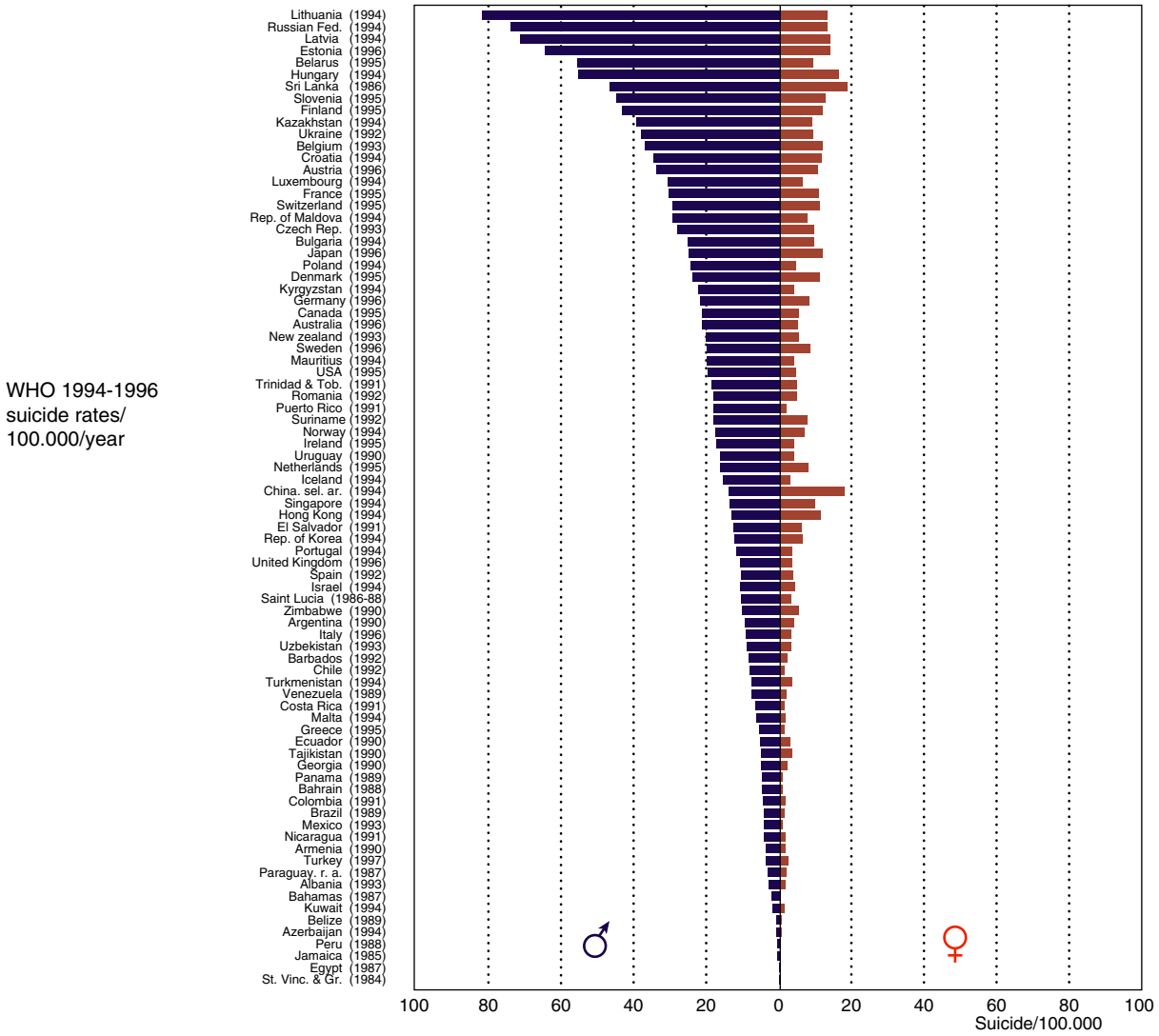

Figure 1 Rank order of suicide rates (0–99+ years) of various countries. Reproduced from Schmidtke, A., Weinacker, B., Apter, A., Batt, A., Berman A., Bille-Brahe, U., Botsis, A., De Leo, D., Doneux, A., Goldney, R., Grad, O., Haring, C., Hawton, K., Hjelmeland, H., Kelleher, M., Kerkhof, A., Leenars, A., Lönnqvist, J., Michel, K., Ostamo, A., Salander-Renberg, E., Sayil, I., Takahashi, C., van Heeringen, C., Värnik, A., Wasserman, D., 1999. Suicide in the world: update. Archives Suicide Research 5, 81–89.

However, suicide occurring in children before prepuberty is very rare, but, without doubt, there is clinical evidence that children from 4 years of age have suicide ideas and can make a suicide attempt or kill themselves. Furthermore, it has to be taken into consideration that until prepuberty (8–10 years) children normally do not experience death as something irreversible (Pfeffer, 1986). Cultural differences exist in terms of using different suicide methods. The so-called weak suicide methods (ingestion of drugs, wrist cutting) and hard suicide methods (shooting, hanging, jumping, drowning, etc.) will be distinguished. Weak suicide methods prevail in suicide attempts, whereas hard methods prevail in successful suicide. In general, females more often prefer weak methods than males. The most frequent method for suicide in females and males in North America is shooting; in Europe, hanging; and in Asia, intoxicating with pesticides and rodenticides especially in rural areas.

Suicide rates in several European countries such as Germany, England and Wales, France, Hungary, and the Nordic countries show an astonishing stability over decades and even a century. The same holds true for German counties such as Saxony, Brandenburg, and Thuringia, where suicide rates have been stable over a century. Figure 2 depicts suicide rates over one century in Germany (Deutsches Reich and Germany (West and East), Felber, 2010). Such stability can even be traced (in other countries) back to the nineteenth century such as in Prussia, France, and the Nordic countries. On the other hand, there have been dramatic changes in suicide rates such as during the nineteenth century, where secularization, industrialization, and individualization have taken place in Western

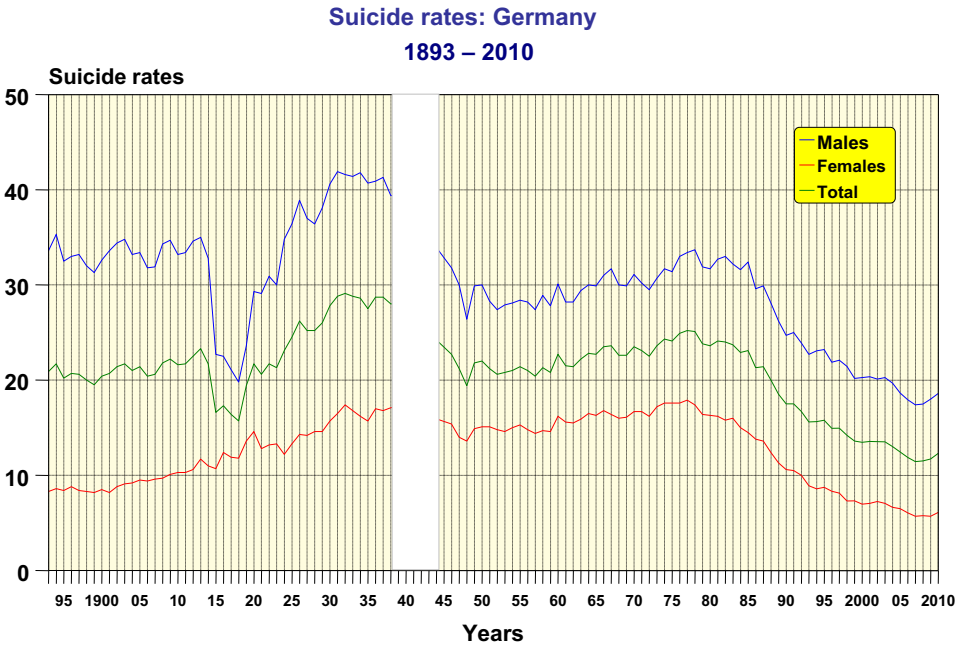

Figure 2 Suicide rates per 100 000 inhabitants in the German Reich from 1900 to 1940 and in Germany (West and East) from 1945 to 2010. Reproduced from Felber, 2010.

and Northern European societies. During World Wars I and II, further dramatic changes occurred affecting not only males and females but also countries that have not been involved in these wars, such as Switzerland. Finally, the last changes have occurred after the fall of the Iron Curtain in Eastern Europe between 1985 and 1993 with a dramatic decrease in suicide rates by up to 40% and an increase after 1993 because of the economic, social, and political worsening in these countries (Mäkinen, 2000). However, the suicide rates are decreasing continuously again during the past 10 years because of amelioration of the economic, social, and political situation (Bronisch and Hegerl, 2010; *see* Mental Disorders, Epidemiology of).

Differing and Common Aspects between Suicide and Suicide Attempts

There are considerable differences between suicide attempters and suicide victims (Stengel, 1964). The latter group is rather small in comparison with the former group. The sociodemographic characteristics are quite different: suicide victims are older and predominantly male, whereas suicide attempters are younger and predominantly female. The tendency of a transition from the suicide attempters group to the suicide victims group is rather small with a 10% chance. The life of the victim is terminated by suicide, whereas a nonfatal suicide attempt becomes a very important event in the life of the person, often with a change in the further life trajectory.

Nevertheless, there are also some common aspects between these two groups: the probability of suicide is higher, the more often suicide attempts happen. Despite the fact that more males are committing suicide and more females make a suicide attempt, there is a considerable overlap between both groups. The same holds true for age distribution: a considerable proportion of young males and females commit suicide, while older males and females make suicide attempts. Most of the risk factors are equally valid for both suicide victims and suicide attempters (see below).

Risk Factors for Suicide and Suicide Attempts

Risk factors for suicide and suicide attempts are as follows: increased age; being separated, divorced, widowed, or single; unemployment; as well as gender and economical status (suicide occurs more often in males, suicide attempts occur more often in females and the lowest social class). The most important risk factors for suicide and suicide attempts, however, are a previous suicide attempt and a psychiatric disorder (Kreitman, 1977).

Gender

As already mentioned there is a higher suicide rate for males than for females. This seems to be surprising since the prevalence of depression is twice as high in females than in males (see below psychiatric diagnoses). This so-called gender paradox might be primarily based upon a preponderance of hard suicide methods in males as opposed to females, whereas in Asian countries such as China, the deadly intoxications with pesticides and rodenticides explain the preponderance of females (Möller-Leimkühler, 2003). Finally, a sociobiological explanation could be that women as guarantees for the transfer of genes from one generation to the other cannot afford suicide. This theory is backed by epidemiological data whereby females show the lowest suicide rates in all cultures and

societies depending on the number of children (Bronisch, 2007). Psychiatric hospitals bear a great risk of losing a patient to suicide. It has been estimated that 15% of depressed patients, 15% of patients with drug or alcohol addiction, and 10% of patients with schizophrenia, who have been treated in a psychiatric hospital, finally die by committing suicide. Patients with anxiety disorders or personality disorders also have a considerable risk of committing suicide. From an epidemiological point of view, subjects with a depression or an addiction have a 15- to 60-fold higher risk of committing suicide when compared with the general population (Bronisch and Hegerl, 2010).

There are differences between urban and rural areas with respect to suicide rates. The highest suicide rates are seen in urban areas, whereas the lowest rates are seen in rural areas. However, there are exceptions: the states of the former USSR as well as Asian countries such as China, where there is a preponderance of suicide rates in rural areas.

Climate also seems to play an important role in the distribution of suicide rates. The highest suicide rates are found during late spring and summer time. This distribution can be found across the Northern as well as the Southern Hemispheres. Countries in southern parts of Europe show a lower suicide rate than countries in the northern parts of Europe.

Religious affiliation also plays an important role in the frequency of suicides. In countries with a predominance of Catholics, suicide rates are lower, whereas in countries with a predominance of Protestants, suicide rates are higher. However, if the degree of urbanization and the distribution of social classes are taken into consideration, the differences between Catholic and Protestant countries are fading (Kreitman, 1977). The same holds for the Islamic countries where a rather low suicide rate can be observed (Okasha and Okasha, 2010).

Origins of Suicidal Behavior

Suicide as a Basic Human Existential Problem

Ever since the written history of mankind, suicide has been considered a basic human existential phenomenon that has been reflected in many philosophical and religious contributions as well as in all kinds of arts. The attitude toward suicide did not only depend on the *Zeitgeist* of the respective historical period but also on the individual stance of different philosophical schools such as of Stoa and Epikur in antiquity and different religions such as Christianity and Buddhism. The sociological theories are a consequence of the Enlightenment of the seventeenth and eighteenth centuries. During the first half of the nineteenth century, medical, especially psychiatric, explanations emerged and during the second half of the nineteenth century, psychological theories developed upon the influence of psychological thinking and emerging psychoanalysis. During the second half of the twentieth century, biological theories were finally stimulated by the natural sciences focusing on the brain.

Suicide requires self-reflecting properties of an individual, i.e., an individual who can differentiate between an observing and experiencing ego. Suicide does not exist in animals. Even nonhuman primates do not show suicidal behavior, neither in natural settings nor in preservations. However, suicide has been found in Neolithic societies.

Sociological Aspects

Durkheim's Theory

Sociological theories are influenced by the theory of Durkheim (1897[1951]) based upon the evaluation of suicide statistics of Western European countries (France and Prussia) during the second half of the nineteenth century. Durkheim proposed four different types of suicide depending on the degree of successful adaptation of the individual to different types of societies. A prerequisite to adaptation of the individual to society is that individualization is neither too weak nor too strong. If the individualization is too strong, the individual will be isolated from society and the danger of an egoistic suicide increases. If the individualization is too weak, the danger of an altruistic suicide increases.

On the other hand, a prerequisite to the equilibrium between the individual and society is that social norms are neither too strong nor too unspecific. Norms that are too strong facilitate fatalistic suicide such as an expiatory death. Norms that are too unspecific favor anomic suicide. With the help of the existing suicide statistics, Durkheim tried to verify three of the four types (he did not discuss the fatalistic suicide). He tried to prove that Protestant countries have a higher suicide rate than Catholic countries, since Protestantism allows the individual a higher degree of freedom than does Catholicism: it is called egoistic suicide. He explained the increase in suicide rates during political crises (Dreyfuss affair in France) in terms of an egoistic suicide. During times of flourishing economic growth, implying a derestriction of social norms, suicide rates should lead to an anomic suicide, whereas during times of war, stronger norms lead to a stronger coherence of society, resulting in a decrease in suicide rates.

General Sociological Trends

During the nineteenth century, only parts of Durkheim's theories could be confirmed such as the differences between Catholic and Protestant countries in Western Europe and the decrease of suicide rates during both World Wars. Generally speaking, Durkheim's data and statistical data from the twentieth century could demonstrate an increase in suicide in Europe during the nineteenth and twentieth centuries depending on the development of industrialization, secularization, and individualization of the respective countries, i.e., the more pronounced these developments were, the more frequently did suicide and suicide attempts occur (Retterstøl, 1992). Another quite remarkable trend of the past decades has been the increase in suicide rate of 15- to 35-year-old adolescents and young adults. In most Western societies, this development has been correlated with an increasing rate of depression and addiction potential. An increasing urbanization, loss of social structures and familial bonds, as well as changes in sex roles may play an important role in this development (Bronisch and Hegerl, 2010).

Psychological Aspects

Psychological models include psychodynamic, attachment, and cognitive behavioral theories that are only partially substantiated by empirical research. The main theories are

concerning the loss of a symbiotic relationship, and a severe personal insult resulting in severe loss of self-esteem. Finally, classical conditioning, operant conditioning, and model learning are, in terms of the behavioral point of view, also of great importance.

Aggression Directed against the Ego

S. Freud (1917) conceptualized one of the first psychological theories about the development of suicidal behavior in *Mourning and Melancholia*. To him, psychodynamics of suicidal behavior and depression were indistinguishable, i.e., suicidal behavior and depression were seen as a turn of aggression against the ego and an ambivalent attitude toward other people. Aggression directed against other people will be followed by loss of a desperately needed person. Turn of the aggression against the ego is expressed by feelings of guilt, self-devaluation, and finally self-killing.

Distorted Cognitive Schemes

Beck (1967) also emphasized a strong connection between suicidal behavior and depression. Depressive persons see themselves, the world, and the future (the so-called cognitive triad) negatively because of their distorted cognitive schemes. Catastrophization, minimizing, maximizing, and over-generalization belong to typically distorted cognitive schemes. As a consequence of such a cognitive triad, helplessness, hopelessness, and powerlessness result. In this case, the only solution for an individual may be to end his or her miserable life by committing suicide.

The entrapment in a vicious cycle of a negative evaluation of own coping strategies is been observed as an essential aspect of suicidal behavior in cognitive behavior theory (Williams, 2001). The result is a lack of development of distinctive problem-solving strategies. The basis for that is a deficient autobiographical memory for all kinds of life events, not only the positive but also the negative ones. A sufficient autobiographical memory is the basis of efficient judgment of future events. Suicidal persons are vague in regard to their future and therefore cannot develop effective problem-solving strategies. They do not know any solution, and experience the situation as hopeless and senseless so that suicide remains the only solution.

Narcissistic Crisis

Henseler (1974/1984) formulated the first narcissistic theory of suicidal behavior. Individuals with narcissistic personality traits are people who can be hurt easily by significant others through criticism or lack of special attention. They are lacking a stable self-esteem and their self-perception is distorted: they oscillate between overestimation and underestimation of their own abilities. This is not restricted to one's own character but also extends onto a corresponding distorted view of other people. Insults through criticism or lack of special attention lead to aggressive outbursts that can be directed either toward others or their own character in terms of a suicide or suicide attempt.

Appeal to Human Bonding

Attachment of the newborn to the mother is a characteristic common to all mammals. In all mammals, the newborn is very dependent and the reproductivity rate at least in the nonhuman and human primates is rather small. Therefore, the strong attachment is a life-preserving measure of the evolution to protect the dependent newborn until it is grown up. The emotional bonding in primates is very intense; separation of the mother will be experienced as very painful (Bowlby, 1977a,b).

As has been known from many empirical studies, most suicides and suicide attempts are consequences of imminent separation, separation, or divorce of males and females. The intention to die is ambivalent in most suicide attempts, i.e., besides a true tendency to end life in nearly all suicide attempts there are tendencies to survive. This can be seen easily in the choice of suicide methods, which often are not so dangerous, i.e., ingestion of small amounts of pills, or in the arrangement of suicide attempt that leaves a rescue open, i.e., by attempting suicide in the presence of a significant other.

The painfully expected or experienced separation from a significant other and the appeal of the significant others not to be left behind both seem to play an important role in the meaning of a suicide attempt (Bronisch, 2007).

Classical and Operant Conditioning

Basic principles of behavioral theory and therapy can be described using the Stimulus Organism Reaction Kontingency Consequence Scheme (SORKC scheme) of Kanfer (Schmidtke, 1988; see **Figure 3**). The SORKC scheme tries to describe each behavior in a broader context of conditions, which can be placed linearly in a chain or circle. The constituents of the SORKC scheme for suicidal behavior are the following:

1. Life events and social situations as Stimuli variables (S), which may trigger suicidal behavior.
2. Behavioral repertoire based upon organic conditions of the individual such as psychiatric disorders and the personality (e.g., cognitive styles and stances) as Organism variable (O), which may predispose to suicidal behavior.
3. Suicidal behavior as a reaction (R) to the S and O variables.
4. Kontingency (K), i.e., the reinforcement conditions, which support or suppress suicidal behavior, and behavior according to a regular or irregular order of the consequences.
5. Consequences (C) in the sense of negative reinforcement (e.g., death) or positive reinforcement (attention of the environment) of suicidal behavior.

The process hereby is rather circular, i.e., consequences can again influence further suicidal behavior. This might be possible by changing the individual (O variable) and the surroundings (S variable).

Model Learning

Imitation of suicidal behavior is a new area in suicide research. The observation that suicidal behavior is running in certain families, cultures, and countries could be explained by genetic transmission, psychological conflicts, or model learning. In regard to model learning, three hypotheses have been proposed: the imitation hypothesis, the contagion hypothesis, and the suggestion hypothesis. All three hypotheses have in common that suicidal behavior of a model will be imitated, whereby predisposition of the individual (the so-called

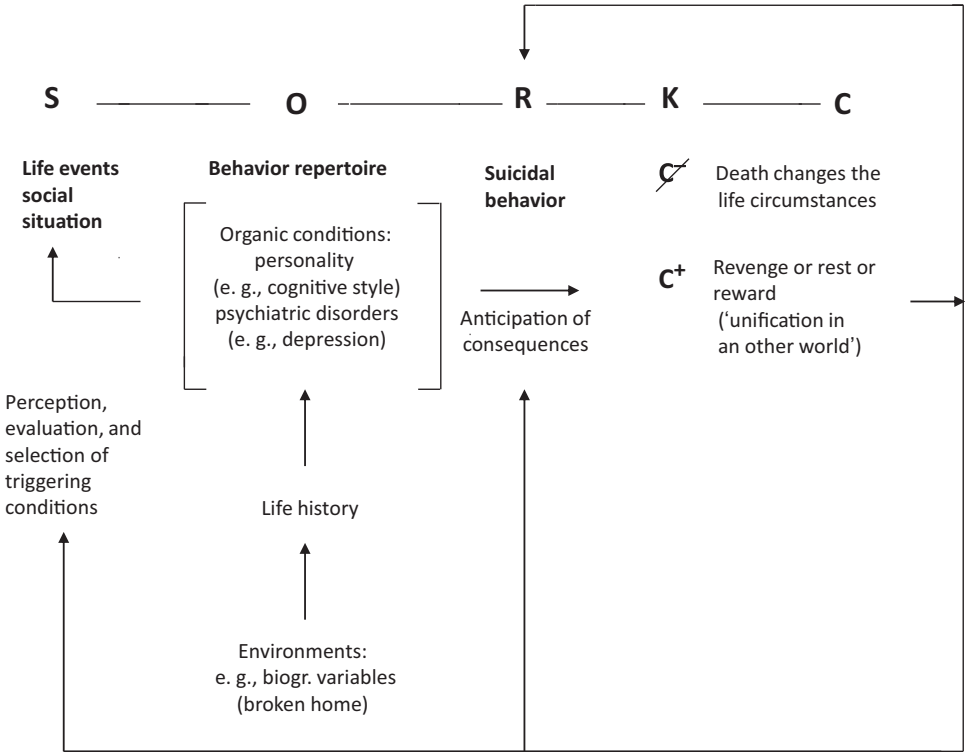

Figure 3 The cognitive behavioral scheme for the development of suicidal behavior. Reproduced from Schmidtke, A., 1988. Verhaltenstheoretisches Erklärungsmodell Suizidalen Verhaltens. S. Roderer, Regensburg, Germany.

'presuicidal personality'), lack of social support in coping with life crises, and enduring social burden are the contributing factors. Furthermore, a high degree of social suggestibility of predisposed individuals plays an important role.

During the 1980 and the 1990s the influence of different media, such as books, newspapers, journals, television, and movies, on increasing suicide rates can be demonstrated (Phillips and Lesyna, 1995). Suicide victims have resembled the person portrayed by the respective media: there have been similarities between this person and the suicide victim in terms of age, gender, motives for the suicide, and the methods used for suicide. The increase of suicide rate was in temporal relationship to the reported suicide cases and was not compensated by a decrease in later time, i.e., these suicides were committed earlier than planned. Also very impressive are studies that demonstrate a decrease in suicide rates not only by changing the information toward being more informative and being less sensational but also by reducing information about suicide (Phillips and Lesyna, 1995). One example is the studies of railway suicides reported in newspapers, which could be reduced by cooperation with local newspapers to reduce and confine to nonsensational reports on this matter. Mass media including Internet play an ever more important role in model learning of suicidal behavior. However, they could be also used for interventions such as restricting information about suicide methods as well as promoting preventive measures. These preventive measures could provide information, offer help, as well as control suicide-glorifying information in web sites, etc. (Westerlund et al., 2009).

In a broader sense, traditions in different cultures and countries concerning either permissive or restrictive attitudes toward suicide can also be regarded as a form of model learning. These attitudes may be transferred over generations such as in Hungary where an extremely high suicide rate can be observed throughout an entire century.

Biological Aspects

During the nineteenth century, Esquirol (1838) was the first to observe that suicidal behavior ran in families and, therefore, postulated that suicidal behavior was a kind of mental illness. Family, twin, and adoption studies gave hints about suicidal behavior having a genetic component.

There is a familial burden due to suicide and suicide attempts (of suicide victims or suicide attempters as compared with subjects without suicide or suicide attempts). Separation of nature (heritage) and nurture (rearing) is not possible within family studies, but it is possible within twin and adoption studies. In twin studies, identical (identical genes) and fraternal twins (half of the genes are nonidentical as in sisters and brothers) are compared with regard to suicide or suicide attempts. The respective studies demonstrate a significantly higher frequency of suicide and suicide attempts in identical as compared with fraternal twins (Roy et al., 1991). In adoption studies, identical twins have been compared with each other, whereby one twin has been raised with his biological parents, while the other twin has been adopted right away after birth. Significantly fewer twins raised by their biological parents

developed suicidal behavior or committed suicide than twins who had been adopted following birth. Furthermore, within biological families, a significantly higher rate of suicides or suicide attempts was shown than within adoptive families of the respective identical twin. All these studies also demonstrate that suicidal behavior has to be seen apart from a specific psychiatric diagnosis, especially apart from a depressive illness. On the other hand, the distinctive feature that might be inherited may not be suicidal behavior as such but rather personality traits such as lack of impulse control or aggressiveness (Wender et al., 1986).

Modern biological research on suicidal behavior started with the investigation by the working group of Marie Asberg (1976) of suicidal and nonsuicidal depressive patients. The following results of Asberg and coworkers could be replicated in several studies (Mann and Currier, 2007):

1. Patients after a suicide attempt showed lower values of 5-hydroxyenodolacetic acid (5-HIAA), the degradation product of serotonin, in the cerebrospinal fluid (CSF) than the psychiatric controls.
2. Patients with 'hard' suicide attempt methods showed lower 5-HIAA values in the CSF than patients with soft methods.
3. Patients with a higher lethal risk of a suicide attempt also showed lower 5-HIAA values in the CSF than those with lower lethality of a suicide attempt.

These results are further corroborated by postmortem studies, neuroendocrine studies, and investigations in peripheral parameters such as platelets, whereby the phenotype is not suicidal behavior per se but impulsive aggressive behavior that could be directed inwardly or outwardly (Mann, 1998).

The second biological system that contributes to the biological underpinnings of suicidal behavior is the stress hormone system. Based upon the preclinical findings, the stress hormone system with the hypothalamic–hypophysis-adrenal axis enhances the deficit of the serotonergic system and leads therefore to impulsive aggressive behavior (Bronisch and Hegerl, 2010).

During the past 10–20 years, molecular genetic research prevailed with rather inconsistent results. In pooled data from multicenter studies as well as meta-analyses polymorphisms in the serotonergic genes seem to predispose primarily to the endophenotypes of suicidal behavior such as impulsivity, aggressiveness, irritability, and hopelessness (Bondy et al., 2006). Further research is directed to molecular genetic aspects of suicidal behavior beyond the serotonergic system (Rujescu et al., 2007).

Prevention

Prevention can be divided into primary, secondary, and tertiary prevention. With regard to primary prevention, the main focus implies a reduction of tendencies within the society that support suicidal tendencies. A verdict of suicidal behavior by means of social ostracizing or punishment has not been successful in reducing suicide rates in medieval and modern times (Minois, 1995). However, primary prevention also concerns the prevention of suicidal behavior by ameliorating life conditions in general and socioeconomic conditions in particular as well as effective treatment of psychiatric disorders. It also refers to a support of tolerance for individual lifestyles. Furthermore, the restricted approach to frequently used methods, such as guns for shooting, can be helpful. Primary prevention also implies the amelioration of power of resistance against suicidal tendencies in the individual. Education of lay people and professionals in detecting and treating suicidal persons could be the most effective strategies. Secondary and tertiary prevention provided by fellow citizens (e.g., Samaritans) or medical or psychological professional personnel means prevention of relapses in suicidal behavior or rehabilitation of patients with suicidal behavior. Empirical studies proving the effect of the above-mentioned preventive strategies are still lacking. Physician education in depression and treatment and restricting access to lethal methods have reduced suicide rates so far (Mann et al., 2005). There is now an emerging trend to develop national suicide prevention programs in several Western and Northern European countries.

Therapy

Therapy has to be divided into crisis intervention, with actual psychiatric-psychotherapeutic intervention; pharmacotherapy; and psychotherapy. Crisis intervention primarily means prevention of an imminent danger of suicide. Crisis intervention should provide care and protection to the individual, i.e., protection of life, since suicide attempts often require medical care such as admission to an internal intensive care unit for detoxification or a surgical intervention. Crisis intervention also includes immediate evaluation of still existing suicidal tendencies of the patient in order to make decisions for further preventive measures: admission to a closed psychiatric ward for relapse prevention or crisis intervention on an in- or outpatient basis. Crisis intervention also includes acute psychotherapy and pharmacotherapy of the patient. The most important determinant of the psychotherapy is hereby the development of a supportive relationship between the suicidal person and the therapist. The greater the time lapse between the suicide attempt and the help offered, the more often are suicidal people inclined to disregard this help. A supportive relationship means that the therapist is accepting the patient's 'cry for help' and is making a commitment trying to understand the patient; however, the therapist is not willing to accept the patient's decision to end his or her life as a form of reasonable problem solving. The therapist clarifies the triggering situation that has led to the suicide attempt. Together with the patient, the therapist tries to develop other strategies to resolve seemingly insoluble conflicts in the patient's life. Also, an important determinant of a successful intervention is the involvement of significant others of the patient as well as the social environment in general. Acute pharmacological intervention means primarily sedation of the patient in order to keep him or her from further suicidal behavior.

Psychotherapy means treatment of psychopathological states such as depression, dysfunctional behaviors like impulsiveness, or pathological personality traits such as low self-

esteem. It includes the classical types of psychotherapy, i.e., psychodynamic approaches and cognitive behavioral approaches. So far, empirical studies could only demonstrate efficacy for some groups of patients, such as patients with borderline personality disorders by cognitive behavioral and psychodynamic therapies (Bronisch, 2011).

Pharmacotherapy of suicidal behavior is primarily focused on the underlying psychiatric illness of the patient such as schizophrenia (antipsychotics), depressive disorders (antidepressants), anxiety disorders (anxiolytics), and addiction (drugs against withdrawal symptoms). However, during the 1990s drugs specifically targeting the typical characteristics of suicide attempters such as impulsiveness and aggressiveness seem to have been successful in reducing suicide and suicide attempts: lithium and mood stabilizers. Antidepressants reduce suicidality in association with reduction of depressive symptoms, and it is assumed that suicidal behavior is also reduced as a consequence (Möller, 2010).

Concluding Remarks

Suicide and suicide attempt are a behavior confined to human beings. They are rooted fundamentally in the human existence and have as a prerequisite self-reflecting properties of the individual, i.e., conscious acting with the consequence or the attempt to extinguish one's own existence.

The attempts to describe suicide and suicide attempt based on a single cause in only one dimension has been tried by Esquirol (1838) in biology, Durkheim (1897[1951]) in sociology, or Freud (1917) in psychology and seems to be inappropriate. Suicide and suicide attempt are multidetermined acts (Roy, 2000). However, the weighting of different aspects of suicide and suicide attempt seems to be rather difficult to explain, so that in the following, only a synopsis of the empirical findings will be given.

Suicide is lacking in animals and even in nonhuman primates. However, it has already been found in Neolithic societies as well as later on in the old Greek and Latin societies. Precise empirical data about suicide in different countries and cultures has been lacking up to the nineteenth century so that a description of the development of suicide from the antiquity via the medieval to modern times is impossible. However, there is a considerable increase of suicides in societies with a high degree of secularization, industrialization, and individualization starting in the middle of the nineteenth century. For individual countries, there is often a stability of suicide rates on a high or low level depending at least in Europe on the degree of industrialization, religious affiliation, and climate: northern countries with a high degree of industrialization, Protestant religious affiliation, and rather few sunny days have considerably higher suicide rates than southern countries with a low degree of industrialization, Catholic religious affiliation, and many sunny days. The differences could best be explained by human bonding as a protective factor, whereas climate could influence biological rhythms, hence triggering suicidal behavior.

In regard to sex differences and suicidal behavior, very stable differences exist between males and females with a male preponderance in suicide and female preponderance in suicide attempts. The reasons for these sex differences, which seem to be stable across countries, cultures, and times, have not been well investigated and need further empirical research. There is no doubt that psychological motives play not only an important role as triggers for suicidal behavior but also are important determinants for the development of suicidal behavior. The most important motive for suicidal behavior is certainly the impending loss of a human relationship, even if the wish to die outweighs the wish to survive in the suicidal act, i.e., in the seriousness of the suicide attempt. Besides that, narcissistic personality traits predispose to suicidal behavior. From the cognitive behavioral point of view, classical conditioning, operant conditioning, and model learning are basic learning principles, which also can be transferred to behaviors such as suicide and suicide attempt. This way, model learning is not only important for perpetuating suicidal behavior in families and peer groups but also in societies with traditional high (Hungary) suicide rates (the so-called imitation hypothesis).

In certain families, a genetic burden of suicidal behavior could be confirmed by twin and adoption studies. This genetic burden is often connected to psychiatric illnesses such as depressive disorder, addiction, anxiety disorder, schizophrenia, and personality disorder. Biological studies (postmortem brain and peripheral biochemical studies) could demonstrate a dysfunction of the serotonergic system in suicide victims and suicide attempters, which in psychopathological terms is correlated to a loss of impulse control, aggressiveness, affective instability, and violent suicide attempt. These results are corroborated by molecular genetic studies.

As empirical studies could confirm, suicidal behavior is not a 'normal' behavior, but is rooted in psychological and biological abnormalities of the respondent with the corollary of prevention and therapy. However, empirical data about successful prevention of suicidal behavior is still scarce. Only physician education in depression and treatment and restricting access to lethal methods have reduced suicide rates so far. Successful therapeutic interventions only exist for subgroups of suicide attempters such as depressives with a successful treatment with lithium and mood stabilizers as well as cognitive behavioral and psychodynamic therapy for patients with borderline personality disorder. Life as a precious good, desperation and suffering of the respondents as well as suffering of the relatives should bring the focus on suicide more into the scientific scope of view as other life-threatening illnesses already have been for a long time.

See also: Bowlby, John (1907–90); Depression; Durkheim, Emile (1858–1917); Freud, Sigmund (1856–1939); Suicide, Sociology of.

Bibliography

Åsberg, M., Träskman, L., Thoren, P., 1976. 5-HIAA in the cerebrospinal fluid: a biochemical suicide predictor? Archives General Psychiatry 33, 1193–1197.

Beck, A.T., 1967. Depression: Clinical, Experimental and Theoretical Aspects. Harper and Row, New York, Evanston, IL, London.

Bondy, B., Büttner, A., Zill, P., 2006. Genetics of suicide. Molecular Psychiatry 11, 336–351.

Bowlby, J., 1977a. The making and breaking of affectional bonds I. British Journal of Psychiatry 130, 201–210.
Bowlby, J., 1977b. The making and breaking of affectional bonds II. British Journal of Psychiatry 130, 421–431.
Bronisch, T., 2007. Der Suizid. Ursachen—warnsignale—prävention. Beck, Munich, Germany.
Bronisch, T., 2011. Suizidalität. In: Dulz, B., Herpertz, S., Kernberg, O.F., Sachsse, U. (Eds.), Handbuch der Borderline-Störungen. Schattauer, Stuttgart, New York, pp. 406–411.
Bronisch, T., Hegerl, U., 2010. Kap. 80: Suizidalität. In: Möller, H.-J., Laux, G., Kapfhammer, H.-P. (Eds.), Lehrbuch Psychiatrie, Psychosomatik, Psychotherapie, 4. neu bearbeitete Auflage. Springer, Berlin, pp. 1469–1501.
Durkheim, E., 1897[1951]. In: Spaulding, J.A., Simpson, G. (Eds.), Le Suicide: Etude de Sociologie. Free Press trans, Glencoe, IL.
Esquirol, J.E.D., 1838. Des maladies mentales: Considerées sous les rapports médicale, hygienique et medicolegale; Accompagneé des 27 Planches gravées. Bruxelles. TI & II, Paris.
Freud, S., 1917. Trauer und Melancholie. Mourning and Melanchiolia. In: Johnes, E. (Ed.), Collected Papers, vol. 4. Hogarth Press, London, pp. 137–156.
Henseler, H., 1974/1984. Narzißtische Krisen. Zur Psychodynamik des Selbstmords. Westdeutscher Verlag, Obladen, Germany. Reinbek bei Hamburg: Rowohlt.
van Hooff, A.J.L., 1990. From Autothanasia to Suicide: Self-killing in Classical Antiquity. Routledge, London.
Kreitman, N., 1977. Parasuicide. Wiley, New York.
Mäkinen, I.C.H., 2000. Eastern European transition and suicide mortality. Social Science Medicine 51, 1405–1420.
Mann, J.J., 1998. The neurobiology of suicide. Nature Medicine 4, 25–30.
Mann, J.J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., Haas, A., Hegerl, U., Lönnqvist, J., Malone, K., Marusic, A., Mehlum, L., Patton, G., Phillips, M., Rutz, W., Rihmer, Z., Schmidtke, A., Shaffer, D., Silverman, M., Takahashi, Y., Varnik, A., Wasserman, D., Yip, P., Hendin, H., 2005. Suicide prevention strategies: a systematic review. JAMA 294, 2064–2074.
Mann, J.J., Currier, D., 2007. A review of prospective studies of biological predictors of suicidal behavior in mood disorders. Archives Suicide Research 11, 3–16.
Minois, G., 1995. Histoire du suicide. Fayard, Paris.
Möller, H.-J., 2010. Pharmacological and other biological treatments of suicidal individuals. In: Wasserman, D., Wasserman, C. (Eds.), Oxford Textbook of Suicidology and Suicide Prevention. A Global Perspective. Oxford University Press, Oxford, pp. 395–405.
Möller-Leimkühler, A.M., 2003. The gender gap in suicide and premature death or: why are men so vulnerable? European Archives Psychiatry Clinical Neuroscience 253, 1–8.
Okasha, A., Okasha, T., 2010. Suicide and Islam. In: Wasserman, D., Wasserman, C. (Eds.), Oxford Textbook of Suicidology and Suicide Prevention. A Global Perspective. Oxford University Press, Oxford, pp. 49–55.
Pfeffer, C.R., 1986. The Suicidal Child. Guilford Press, New York, London.
Phillips, D.P., Lesyna, K., 1995. Suicide and the media research and policy implications. In: Diekstra, R.F.W., Gulbinat, W., Kienhorst, I., De Leo, D. (Eds.), Preventive Strategies on Suicide. Brill, Leiden, Germany, pp. 231–261.
Platt, S., Bille-Brahe, U., Kerkhof, A., Schmidtke, A., Bjerke, T., Crepet, P., De Leo, D., Haring, C., Lonnqvist, J., Michel, K., Philippe, A., Pommereau, X., Querejeta, I., Salander-Renberg, E., Temesvary, B., Wasserman, D., Sampaio Faria, J., 1992. Parasuicide in Europe: the WHO/EURO multicentre study on parasuicide I. Introduction and preliminary analysis for 1989. Acta Psychiatrica Scandinavica 85, 97–104.
Retterstøl, N., 1992. Suicide in the Nordic countries. Psychopathology 25, 254–265.
Roy, A., 2000. Suicide. In: Sadock, B.J., Sadock, V.A. (Eds.), Kaplan & Sadock's Comprehensive Textbook of Psychiatry II. Lippincott Williams & Wilkins, Philadelphia, Baltimore, New York, London, Buenos Aires, Hong Kong, Sydney, Tokyo, pp. 2031–2040.
Roy, A., Segal, N.L., Centerwall, B.S., Robinette, C.D., 1991. Suicide in twins. Archives General Psychiatry 48, 29–32.
Rujescu, D., Thalmeier, A., Moeller, H.-J., Bronisch, T., Giegling, I., 2007. Molecular genetic findings in suicidal behaviour. What is beyond the serotonergic system? Archives Suicide Research 11, 17–40.
Schmidtke, A., 1988. Verhaltenstheoretisches Erklärungsmodell Suizidalen Verhaltens. S. Roderer, Regensburg, Germany.
Schmidtke, A., Weinacker, B., Apter, A., Batt, A., Berman, A., Bille-Brahe, U., Botsis, A., De Leo, D., Doneux, A., Goldney, R., Grad, O., Haring, C., Hawton, K., Hjelmeland, H., Kelleher, M., Kerkhof, A., Leenars, A., Lönnqvist, J., Michel, K., Ostamo, A., Salander-Renberg, E., Sayil, I., Takahashi, C., van Heeringen, C., Värnik, A., Wasserman, D., 1999. Suicide in the world: update. Archives Suicide Research 5, 81–89.
Stengel, E., 1964. Suicide and Attempted Suicide. Penguin, UK.
Wender, P.H., Kety, S.S., Rosenthal, D., Schulsinger, F., Ortmann, J., Lunde, I., 1986. Psychiatric disorders in the biological and adoptive families of adopted individuals with affective disorders. Archives General Psychiatry 43, 923–929.
Westerlund, M., Schaller, S., Schmidtke, A., 2009. The role of mass-media in suicide prevention. In: Wasserman, D., Wasserman, C. (Eds.), Oxford Textbook of Suicidology and Suicide Prevention. A Global Perspective. Oxford University Press, Oxford, pp. 515–524.
Williams, M., 2001. Suicide and Attempted Suicide. Penguin Books, London.
World Health Organization, 1999. Consultation on Strategies for Reducing Suicidal Behaviors in European Region: Summary Report. World Health Organization, Geneva.

Suicide, Sociology of

Julie A Phillips, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by F. Gratton, volume 22, pp. 15273–15278, © 2001, Elsevier Ltd.

Abstract

Suicide, the act of deliberately killing oneself, is shaped by a confluence of biological, psychological, social, economic, and cultural factors. When attempting to explain suicide, psychologists emphasize individual-level factors that are partly determined by biology, such as a history of mental illness or substance abuse. In contrast, sociologists recognize the important social aspects of suicide, focusing on variation in suicide at the aggregate level. Numerous cross-sectional and longitudinal studies provide support for the Durkheimian perspective that social integration and regulation, proxied by religious and demographic composition, family structure, and economic conditions, are associated with suicide rates. Contagion and access to means also play a role. A multidisciplinary approach to understanding suicide is required, integrating the micro and macro, social and psychological, and quantitative and qualitative approaches.

Definition and Measurement of Suicide

Suicide derives from the Latin word *sui* (of oneself) and *cide* (killing) and is defined by the World Health Organization as the act of deliberately killing oneself or intentional self-harm resulting in death. Historical accounting of suicide is fragmented but reveals substantial variation in our attitudes and beliefs about the act. Some in ancient Greece believed fervently in the right of an individual to choose the means and timing of death, and Eskimo and Norse cultures encouraged altruistic suicide among the elderly and sick. Yet Aristotle viewed suicide as an act of cowardice and a form of treason, the Romans prohibited suicide, and major religions imposed strong sanctions and prohibitions against the act (Jamison, 1999). Today, virtually all Western societies have decriminalized suicide, most recently Ireland in 1993, but attempted suicide remains a punishable offense in some countries and is an issue of debate among the US military (Dishneau, AP April 30, 2013). Although public understanding of the phenomenon has improved, considerable misinformation and stigma still surround the act.

Researchers often distinguish between completed suicide in which a death has occurred and other forms of suicidal behavior, such as attempts, suicidal ideation, and other types of self-harm such as excessive alcohol or drug use, promiscuity, or reckless driving. Completed suicides are widely believed to be underreported with estimates of the underreporting ranging from 10 to 50%. Underreporting is common largely because it can be difficult to ascertain intent since the majority of suicides do not leave notes or have witness reports. Measurement is also complicated by the fact that different areas may have varying legal and administrative procedures to report suicide. Moreover, relatives may apply pressure on coroners to avoid the stigma of suicide, although studies suggest the influence of this is marginal (Timmermans, 2005). Most researchers have concluded that suicide underreporting is randomly distributed and that official statistics can reasonably be used (Pescosolido and Mendelsohn, 1986). Still, caution should be used when making comparisons over geographic area and time due to potential differences in reporting practices.

Schools of Thought on the Etiology of Suicide

Historically, suicide was believed to be caused by supernatural forces or by an insufficient embrace of religion; by the 1500–1600s, the notion of insanity as a cause of suicide emerged. It is clear from current research that the etiology of suicide is complex, shaped by a confluence of biological, psychological, social, economic, and cultural factors that operate at both the individual and societal levels. Thus, the topic of suicide lies at the intersection of multiple disciplines, among them sociology, psychology, psychiatry, biology, public health, genetics, neuroscience, and philosophy.

Psychological and Biological Approaches to Understanding Suicide

Not surprisingly, psychological theories of suicide emphasize psychological pain as the root cause of suicide. The early psychiatrists, Sigmund Freud and Karl Menninger, asserted that three emotions – hate, depression, and guilt – characterize suicide and that these feelings give rise to an excessively critical superego and a wish to die. In the contemporary period, Edwin Shneidman's prominent 10-point definition of suicide emphasizes the intense 'psychache' experienced by suicide victims that leads them to view suicide as the only solution to their problems. Shneidman, the founder of the American Association of Suicidology, suggests that all suicides share 10 common characteristics, among them the need for a solution, unbearable psychological pain, hopelessness and helplessness, and a lifelong pattern of self-destruction (Maris et al., 2000).

As a result of this general theoretical framework, research coming out of the psychiatry/psychology fields focuses on the role of individual-level risk factors that are partly determined by biology, such as a history of mental illness or substance abuse. Using psychological autopsies in which data are collected through interviews and other sources about persons who have completed suicide, some studies indicate that close to 90% of individuals completing suicide suffered from some severe depressive disorder at the time of death. Twin and family studies show a strong genetic component to depression and

suicide, with current research estimating the genetic contribution to suicidal behavior somewhere between 30 and 50% (Joiner et al., 2005). Neurobiological work suggests that there are important serotonergic differences in depressed and suicidal individuals compared with others (see work by John Mann and colleagues), and some studies have linked the development and increased use of new antidepressants (selective serotonin reuptake inhibitors) to declines in suicide rates during the 1990s and early 2000s, particularly among the elderly (e.g., Gibbons et al., 2005). However, despite the clear significance of mental disorders in contributing to suicidal behavior, it remains the case that the vast majority of individuals who suffer from some mental illness do *not* commit suicide, suggesting the importance of other factors.

Individual-level analyses also reveal that a loss or lack of connection to others is an important underlying risk factor for suicide. Hopelessness has been identified as a significant correlate of suicide, along with other psychosocial factors such as burdensomeness, impulsivity and aggression, and ineffectiveness. These psychosocial factors may explain the strong links found at the individual level between (changes in) marital or employment status, physical or mental health problems, and substance abuse and the risk for suicide (Joiner et al., 2005; Maris et al., 2000). Psychologist Thomas Joiner interweaves these components in his influential interpersonal theory of suicide to identify conditions under which suicides are more likely to occur. According to Joiner's theory, suicides are more common when both the desire for suicide, spurred by feelings of social alienation and burdensomeness, and the ability to suicide, a fearlessness regarding death whereby one is capable of suicide due to prior exposure to life-threatening accidents or self-injurious behavior, coexist (Joiner, 2007).

Yet the psychological approach cannot effectively account for group-level differences in suicide risk, be they across geographic areas or different demographic groups. Nor can the individual-level approach explain other social aspects of suicide, such as contagion and the fact that suicide *rates*, a collective measure, may operate on a different explanatory level from individual suicides due to the effect of social interaction.

Sociological Approaches to Understanding Suicide

To address some of these limitations to psychiatric and psychological perspectives on suicide, sociological approaches tend to study suicide at the aggregate level to understand variation across groups. Sociologists' work on suicide is heavily influenced by Emile Durkheim, a founding father of sociology who argued that, although suicide is seemingly an individualistic act, it is also very much a function of a society's social structure – what he termed a 'social fact.' Thus, societal level characteristics, such as marriage rates, religious composition, and economic conditions, may produce variation in suicide rates, and a vast number of studies have tested Durkheim's ideas using data from a variety of countries and time periods. More recently, sociologists have introduced the notion of contagion in producing suicide as well as the role of networks. All these sociological approaches imply that social interaction is critical in affecting the well-being of societies and their members.

Durkheim's Original Thesis

In his book entitled *Suicide*, Durkheim examined patterns of suicide in nineteenth-century Europe, finding that some societies displayed higher suicide rates than others and that a country's suicide rate tended to correlate with other social characteristics. In explaining these patterns, he identified four types of suicide, which relate to levels of social integration and regulation within a society, as illustrated by **Figure 1**. Low levels of social integration within a society lead to instability and lack of cohesion, producing excessive individualism and high rates of *egoistic suicide*. Conversely, extreme levels of social integration can overwhelm individual needs, leading to *altruistic suicide*. A lack of social regulation produces anomie, an absence of norms and an inability of society to meet the population's needs and expectations, and corresponds to high rates of *anomic suicide*. Too much social regulation, on the other hand, in which individuals may feel oppressed, may produce *fatalistic suicide*, although Durkheim asserted that this is a rare form of suicide (Durkheim, 1951). Using this framework, Durkheim was able to explain why Protestant societies exhibited higher suicide rates than their Catholic counterparts and higher proportions of married persons are associated with lower suicide rate. He used the theory to predict that suicide rates would rise during periods of economic depression and boom as a result of rising anomie.

Critiques and Extensions of Durkheim

Durkheim's treatise has been criticized on a few grounds, most notably that he may have committed the ecological fallacy by drawing conclusions about individual behavior based on group-level statistics. Subsequent studies have found that differential recording practices across religious groups rather than differing levels of social integration and regulation may have accounted for some of the observed patterns (van Poppel and Day, 1996; but see van Tubergen et al., 2005). Others have

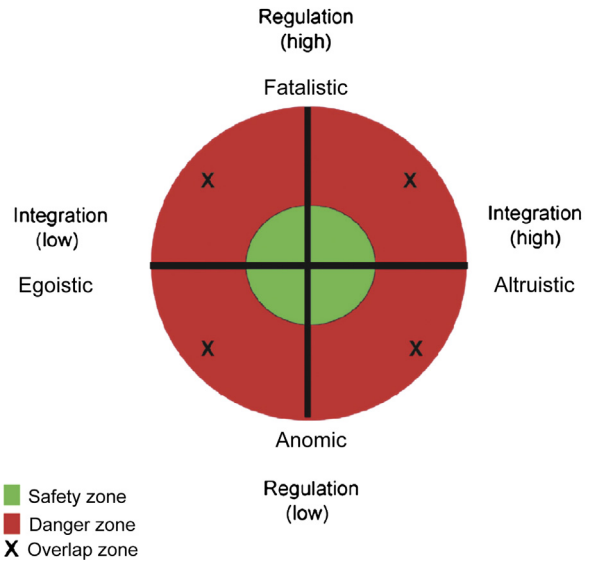

Figure 1 Durkheim's typology of suicide. Source: Marson, S.M., Powell, R.M., 2011. Suicide among elders: a Durkheimian proposal. International Journal of Ageing and Later Life 6 (1), 59–79.

noted that some typologies of suicide, namely, fatalistic and altruistic suicide, are not fully developed and that the distinction between anomie and egoism may be false (Johnson, 1965).

Jack Douglas (1967) in particular was highly critical of the Durkheimian approach in his book *The Social Meanings of Suicide*. His critique rests largely on the argument that suicide does not have a tidy definition – rather, it is a multidimensional and complex phenomenon. As a result, official statistics like death records or death certificates cannot reliably indicate the extent of suicide, as evidenced by the fact that coroners and medical examiners tend to underreport suicide and apply varying criteria to make the determination. Without a well-measured dependent variable, Douglas argues that quantitative approaches to suicide and their attempts to link the phenomenon to measures of social isolation or regulation should be abandoned. Instead, researchers should consider the subjective meaning of suicide by applying an ethnomethodological approach that focuses on the first-person accounts of actual suicidal individuals, namely, those who have made a prior serious attempt at suicide. Despite this criticism, it remains the case that the vast amount of sociological work on suicide is quantitative in nature.

Half a century after the publication of Durkheim's *Suicide*, Henry and Short (1954) revisited Durkheimian themes by examining how anomie and frustration can produce different types of aggressive responses. Like Durkheim, Henry and Short recognized the importance of external restraints (such as religion and particularly the economy) on suicide rates, but they introduced the notion of internal restraints (psychological components such as a strong conscience or propensity toward guilt) and the ways in which they may interact with varying levels of external restraints to produce different forms of aggression. Those experiencing weak external restraints (e.g., high social status and/or social isolation) and strong internal restraints (a well-developed superego) are more likely to take responsibility for frustration and internalize aggression (suicide) while those with strong external restraints and weak internal restraints are more likely to blame others for frustration and externalize their aggression (homicide). Although the theory is among the first to document the ways in which business cycles may affect social status and in turn levels of frustration and aggression, it has been criticized on empirical grounds since suicide rates tend to be higher among those of lower socioeconomic status; it cannot explain within-group variation in suicide/homicide rates and focuses primarily on level rather than change in status.

Gibbs and Martin (1964) expanded Durkheim's conceptualization of social integration by introducing the notion of status integration to explain societal-level variation in suicide rates. Individual behavior is largely governed by one's social identification or status, which carries with it a set of demands and expectations. An individual may occupy multiple social roles or statuses in the realm of family, work, community, and elsewhere. Status integration is determined by the degree to which the demands associated with one status are at odds with those dictated by another status; when roles are incompatible, social relationships are disrupted. Gibbs and Martin predicted that when a large proportion of the population faces a conflict in roles or lack of status integration, suicide rates rise. The authors pay close attention to the operationalization and advocate empirical testing of these concepts. Studies find support for the role of both occupational and marital integration in affecting levels of suicide (e.g., Gibbs, 2000).

Empirical Tests of the Durkheimian Perspective

Since Durkheim's original formulation, a large number of empirical studies have been conducted to determine the extent to which his theoretical framework is supported. Perhaps the earliest application to the United States following Durkheim emanated from the Chicago School of Sociology. Faris (1955) showed that socially disorganized neighborhoods in Chicago during the 1930s exhibited higher levels of social problems, including suicide and mental illness. These types of studies have proliferated in recent decades with the advent of large data sets and increased computer capabilities. The studies adopt both cross-sectional approaches, examining variation in suicide rates across countries or smaller geographic areas such as states or counties, and longitudinal designs, exploring factors that can explain variation in suicide rates over time within a given locale. Taken as a whole, the studies provide substantial support from many of the ideas put forward by Durkheim.

The Role of Religion

The role of religion in affecting suicide rates was central in Durkheim's argument. He presented evidence that Catholic societies exhibit lower suicide rates than their Protestant counterparts and attributed the difference to the degree of social integration and regulation present within the two Churches. While the Catholic Church is more communal and steeped in rituals with strong prohibitions against suicide, the Protestant Church, although prohibiting suicide, is more individualistic in orientation and places the responsibility for salvation on individual members. This tenet of Durkheim's *Suicide* – the idea that Catholics are more socially integrated and regulated and present lower suicide rates as a result – was once viewed as so infallible as to be coined sociology's 'one law' (Pope and Danigelis, 1981).

Recent research adds greater nuance to this finding, showing important differentiation by denomination. At the aggregate level, studies reveal that a greater proportion of more conservative Protestant faiths (such as Evangelical and Fundamental Christians) tend to protect against suicide while a greater prevalence of more liberal Protestant faiths (such as Mainline or Institutional Protestant churches) seem to raise suicide rates. The proportion of those adhering to Judaism tends to fall in between, having a small but protective effect (Pescosolido and Georgianna, 1989). A recent analysis of suicide rates across US states and over time between 1976 and 2000 showed that states with rising numbers of Catholics and greater declines in Episcopalians exhibited larger declines in suicide rates over the period (Phillips, 2013).

Contemporary work also offers insight into the mechanisms by which religious composition may affect suicide rates. The degree of religious heterogeneity appears to matter – an analysis of almost 300 metropolitan areas (MSAs) in the United States in 1979–81 revealed that MSAs with more homogeneous religious composition exhibit lower suicide rates, particularly in the Northeast and the South (Ellison et al., 1997).

Pescosolido (1990) attributes these differences to the strength of social networks – the more coadherents there are in an individual's direct environment, the more strongly involved and the greater support that individual will receive from the religious community, which in turn will lower the risk for suicide rate. van Tubergen et al. (2005) tested competing hypotheses about how and why religious communities (i.e., those with a high level of religious involvement) may be associated with lower suicide rates. Is the protective effect due to the fact that more religious communities and their social networks enhance the social and emotional support that their members experience, as Pescosolido (1990, 1994) proposes? Or do more religious communities impose stronger prohibitions against suicide than secular communities? Their study employs a rare multilevel analysis of suicide rates in the Netherlands to show that religious communities have a more general protective effect against suicide. As the proportion of religious individuals within a municipality increases, the risk of suicide declines within that municipality, regardless of an individual's own denomination and for non-Church members as well. These findings are consistent with the idea that a shared religious norm prohibiting suicide extends to all within the community and lowers the suicide risk, rather than the idea that social support from homogeneous others protects people from suicide.

The Role of Family Structure

Marital status and family are known to be among the strongest protective factors against suicide at both the individual and aggregate levels. Just as Durkheim observed in nineteenth-century Europe and explained using the idea of social integration, widowed and divorced individuals, men in particular, exhibit among the highest suicide rates while married individuals (over age 20 years) with children present the lowest rates. Of course, family relations are protective against suicide only when they are positive and strong. Evidence shows, not surprisingly, that family tension and conflict can elevate the risk of suicide at the individual level (Gibbs, 2000; Maris et al., 2000). For example, the risk of suicidal attempts among adolescents was reduced in families with higher levels of emotional and material support (forms of social integration) and with firm boundaries and greater parental supervision (Thorlindsson and Bjarnason, 1998).

At the aggregate level, research typically confirms that the higher the rate of divorce, the higher the rate of suicide, on average. This is particularly true for cross-sectional studies that examine variation across nations, states, and/or cities, but longitudinal studies that explore how changes in divorce over time are associated with changes in suicide rates often show a link as well. For example, Phillips (2013) found that US states with higher percentages of divorced persons exhibited higher firearm and nonfirearm suicide rates between 1976 and 2000, but that changes over the time period in the percentage divorced were positively associated only with nonfirearm suicide rates. Lester (1995) demonstrated that the divorce rate of a place is associated not only with the suicide rate of divorced people but also with those of the married, widowed, and single populations, suggesting that the divorce rate is a general indicator of the vitality of marriage as an institution and perhaps societal well-being.

The Role of the Economy

Durkheim hypothesized that suicide rates would rise during periods of economic depression *and* economic boom because both extremes can be viewed as indicators of anomie. With rapidly changing economic prospects in either direction, a society may be unable to adapt quickly enough to sufficiently meet its members' needs and expectations, weakening social regulation. Furthermore, the rise in unemployment during periods of economic recession has implications for social integration. Social integration may be undermined directly as unemployed individuals have more limited access to social support systems and indirectly because unemployment can produce relationship problems and/or financial difficulties that create stress, as well as harmful coping mechanisms, such as increased alcohol consumption. Studies show, for instance, an increase in excessive alcohol consumption during periods of economic downturn (Popovici and French, 2013). To test these assertions, research has adopted multiple measures of the economy, including unemployment rates, gross domestic product, real income, and sales, to determine its association with suicide rates.

Although cross-sectional analyses typically find limited support for a connection between economic conditions and suicide rates, longitudinal studies generally do find an association, supporting the more intuitive notion suggested by the Henry and Short model that suicide rates rise only during periods of economic recession (Stack, 2000a,b). One of the most comprehensive studies by Luo et al. (2011) examined business cycles and suicide rates in the United States between 1928 and 2007, showing that the total suicide rate tends to rise during periods of economic recession and fall during expansions. The study found some variation in the relationship by age group, with this association observed for those between the ages of 25 and 64 years, but no such relationship for those aged 15–24 years and over age 65 years. We might anticipate this variation by age, as those between the ages of 25 and 64 years are likely to be family breadwinners and supporting dependents and thus face greater hardship during periods of economic recession while the young and elderly are less likely to be in the labor force and often have the support of family and in the case of the elderly, social security. Although this analysis does not include the period of the 2009–10 Great Recession in the United States, US suicide rates have risen for those between the ages of 35 and 64 years during this period. A leading explanation for the recent spike among this particular age group relates to economic factors.

Studies using individual-level data confirm a powerful link between unemployment status and suicide risk (Maris et al., 2000). This relationship is particularly strong for men, who, historically at least, have been the family breadwinner with closer ties to work and careers than women and thus may experience more stress and feelings of failure when faced with unemployment. However, the pathways linking employment status to suicide at the individual level are less clear. While unemployment may increase suicide risk by creating financial strain, feelings of worthlessness, and relationship difficulties, it is also possible that those who possess a risk for suicide (e.g., through depression or substance abuse) may be more likely to become unemployed.

The Role of Social Stratification

Durkheim was relatively silent about the effect of social and economic stratification on suicide. He hypothesized that the poor would exhibit relatively low levels of suicide – essentially because he viewed poverty as a form of social regulation that protects against anomie during periods of recession. Henry and Short (1954) argued that the poor possessed strong familial and community relations but a disadvantaged social status, leading them to blame external factors for their difficulties. As a result, the poor were predicted to exhibit high homicide rates but low suicide rates. In reality, these predictions are not borne out by empirical data. The poor exhibit higher death rates from both homicide *and* suicide. Research on the effects of income inequality on suicide is limited and results are mixed, but scholars have explored race, gender, and life course disparities in suicide.

Race

Suicide rates are highest for Native Americans and whites and relatively low for Hispanics and blacks. Whites have rates three to four times those of blacks, a pattern that has remained remarkably stable over time. Scholars have put forward a number of explanations for the relatively low rates of suicide among blacks and Hispanics despite their economic disadvantage relative to whites. Some stress factors of social integration, such as the strong extended family networks and religious involvement within certain minority communities. Others note that suicide is highly stigmatized among the black community and viewed as 'a white thing' (Gibbs, 1997).

At the aggregate level, Kubrin et al. (2006) found that disadvantage (measured by factors such as male joblessness, poverty, educational attainment, and household structure) is associated positively with suicide for both young black and white men, but that industrial employment levels and changes therein as a result of deindustrialization are significant predictors of black male suicide rates only across US cities. Among Hispanics, higher levels of income inequality (relative to whites) are positively associated with suicide rates for the native-born but for foreign-born Hispanics, higher levels of black–Hispanic economic inequality are tied to lower suicide rates (Wadsworth and Kubrin, 2007).

Gender

With regard to gender, males are three to four times more likely to die by suicide than are females. Women, however, attempt suicide at higher rates but tend to use less lethal means such as poisoning, which are less likely to result in death. In terms of a sociological understanding of this differential, Durkheim, writing at the turn of the twentieth century, explained women's lower risk of death by suicide by suggesting that women are 'simpler social beings' with 'a less developed mental life.' As a result, women possess weaker ties to the collective and their behavior is not so heavily influenced by the social structure. Durkheim's arguments are now dismissed, and scholars today note that women may be protected from suicide because they tend to possess stronger and more frequent ties with family, friends, and community than do men and exhibit lower rates of alcohol abuse and higher rates of help seeking, among other explanations.

As women's status has changed over time, with increases in female labor force participation and divorce rates, one may expect to see changes in the sex suicide differential. Using data on 18 nations over a 40-year period, Pampel (1998) found that male and female suicide rates start to converge as changes in labor force participation and marital stability emerge. However, as women become increasingly involved in the labor force and institutions adjust, the female advantage in suicide reappears as their rates fall relative to men's, a pattern that occurs more quickly in collectivist societies. These findings are consistent with the conjectures of Gibbs and Martin, who argued that as status configurations (e.g., a married woman with children in the labor force) become more common, those in them may experience less role conflict and stress and lower suicide risk as a result.

Birth Cohorts

Demographers have focused attention on variation in suicide risk across birth cohorts or social generations. Cohort members share not only common contemporaneous but also historical sociocultural contexts that can influence behavior. Unlike prior work on suicide that tends to be divided along biological and social lines, the concept of a birth cohort implicitly connects biology and social structure by recognizing that the implications of a particular social or historical event can be distinct depending on the stage of life course at which that event occurs, with events during adolescence generally believed to be most formative in shaping attitudes and beliefs.

Most well known is Richard Easterlin's theory of relative cohort size in explaining variation in cohort behavior, which suggests that members of relatively large birth cohorts experience a number of social and economic disadvantages that persist throughout the life course (Easterlin, 1987). In childhood and adolescence, social integration and regulation are reduced due to factors such as large classroom size and reduced parental attention due to bigger average family size; in adulthood, members of large birth cohorts face greater competition in college and in the labor market. Age-period-cohort analyses of suicide rates find support for this theory, with cohorts of larger relative size and with a greater proportion of nonmarital births exhibiting higher suicide rates, net of age and period effects (Stockard and O'Brien, 2002a,b). Others have noted the possibility of a cohort-based argument for the unusual and marked increase in suicide rates between 1999 and 2010 in the United States among middle-aged baby boomers (Phillips et al., 2010).

The Theory of Contagion and Suicide Clusters

A contemporary of Emile Durkheim, Gabriel Tarde (1903 [1895]) introduced an alternative perspective on suicide in a book entitled *Laws of Imitation*, in which he proposed that individuals will imitate those behaviors and attitudes to which they are exposed in their social environment. The closer the proximity to those behaviors and the more frequent the exposure, the greater the likelihood of imitation. Although Durkheim largely dismissed the notion of imitation, historical and contemporary analyses of suicide rates over time or space reveal that suicides can occur more frequently or cluster in close temporal or geographic proximity. More contemporary versions of imitation theory or suicide contagion are known as the

'Werther effect,' named for the rise in suicides in 1774 following the publication of Goethe's novel about copycat suicides, whereby suicide rates may increase in the wake of well-publicized individual suicide cases. Simple awareness of a suicide act raises the possibility in the public's mind, akin to Joiner's notion of fearlessness. These ideas have been important in affecting guidelines for media reporting of suicide, which specify that details on method and location should not be included in public reports.

The sociologist David Phillips has been most influential in the recent period in promoting these ideas. Phillips (1974) demonstrated that coverage of celebrated suicides (such as Marilyn Monroe) on the front page of the *New York Times* newspaper was associated with a statistically significant increase in the national suicide rate 7–10 days after the publicity. The increase in suicide was greater the more extensive the front-page coverage and in geographic areas where the story ran. In another compelling study, Phillips (1982) tracked the association between the airing on television soap operas of fictional suicides and the actual suicide rate over a 10-year period, finding that the actual suicide rate spiked in the days following exposure to a soap opera suicide. Although the studies are suggestive of contagion, the ecological nature of the studies with minimal statistical controls limits our conclusions. However, research that includes more extensive controls for factors like unemployment shows more convincingly that contagion producing suicide clusters is more likely among teens and when the suicide involves a celebrity.

Suicide Prevention

While sociologists and psychologists have focused primarily on understanding the causes of suicide, public health practitioners are more likely to consider effective means of suicide prevention and postvention efforts. Historically, suicide prevention efforts rely on psychological approaches to understanding suicide and target those who possess well-known individual-level risk factors for suicide, such as a prior attempt or history of depression. Wray et al. (2011) lament the absence of sociological theories and explanations in suicide prevention research and call for a more active role moving forward, particularly in light of the expansion of community-based approaches to suicide prevention.

Perhaps the one exception to this individualistic approach to suicide prevention is the focus among public health scholars on the ways in which reducing access to lethal means of suicide can be a powerful prevention tool. This is most famously illustrated by the British Coal Gas story (Kreitman, 1976). During the 1950s, the British government began to move away from coal gas, which contained large quantities of carbon monoxide, and toward natural gas for domestic purposes. In the 10 years following this conversion, the national suicide rate declined markedly for all age–sex subgroups, sometimes by as much as one-third. The leading means of killing oneself – poisoning by gas inhalation – had been removed and suicides by other means increased only slightly, leading to an overall decline in suicide rates, particularly among females.

In the United States, Matthew Miller and colleagues at the Harvard School of Public Health have been most vocal recently in promoting the important role that means reduction must play in any comprehensive suicide prevention program (http://www.hsph.harvard.edu/means-matter). Research shows that a substantial number of suicide attempts are impulsive, that access to a lethal means, such as a firearm, is an important risk factor for suicide, and that individuals do not necessarily substitute an alternative method if access to a firearm is removed. For example, ecological studies show that for both sexes and all age groups in the United States, states with higher rates of household firearm ownership have higher suicide rates by firearm and overall, yet rates of nonfirearm suicide are not associated with gun ownership rates, indicating that displacement (from one method to another) does not occur (Miller et al., 2007). In sum, research shows importantly that any mechanism to slow down an action, whether requiring trigger locks on guns, placing barriers on bridges, or packaging pills individually, can be an effective way to prevent suicide.

Conclusion

A recent report by the Institute for Health Metrics and Evaluation on the Global Burden of Disease is a grim reminder of the importance of the topic. Worldwide, suicide was responsible in 2010 for more deaths than war, homicide, and natural disasters combined, and it was identified as the leading cause of death for those aged 15 to 49 years. Thousands of articles have been published on the topic of suicide, appearing in a wide variety of journals across multiple disciplines and underscoring both the complexity of and the continued interest in the mystery of suicide. Sociology has played a critical role in the understanding of the phenomenon, reminding us that even the most individualistic and intimate behaviors are products of the society we inhabit. Yet, as Wray et al. (2011) note in the *Annual Review of Sociology*, scholars are increasingly recognizing that a multidisciplinary approach to understanding suicide is required. Achieving that goal means that moving forward, sociologists must make concerted efforts to integrate the micro and the macro, the social and the psychological, and to employ both quantitative and qualitative methods in their quest to understand and ultimately prevent suicide (Wray et al., 2011).

See also: Adult Mortality in Industrialized Societies; Anomie; Gene–Environment Interplay; Inequality, Social; Life Course: Sociological Aspects; Medical Sociology; Mortality, Epidemiological, and Health Transitions; Social Relationships in Adulthood.

Bibliography

Douglas, J.D., 1967. The Social Meanings of Suicide. Princeton University Press, Princeton, NJ.

Durkheim, É., 1951. In: Spaulding, J.A., Simpson, G. (Eds. and Trans.), Suicide: A Study in Sociology. Free Press, Glencoe, IL.

Easterlin, R.A., 1987. Birth and Fortune: The Impact of Numbers on Personal Welfare. University Chicago Press, Chicago.

Ellison, C.G., Burr, J.A., McCall, P.L., 1997. Religious homogeneity and metropolitan suicide rates. Social Forces 76 (1), 273–299.

Faris, R.E.L., 1955. Social Disorganization, second ed. The Ronald Press Company, New York, NY.
Gibbons, R.D., Hur, K., Bhaumik, D.K., Mann, J.J., 2005. The relationship between antidepressant medication use and rate of suicide. Archives of General Psychiatry 62 (2), 165–172.
Gibbs, J.P., 2000. Status integration and suicide: occupational, marital, or both? Social Forces 78, 949–970.
Gibbs, J.P., Martin, W.T., 1964. Status Integration and Suicide: A Sociological Study. University of Oregon Press, Eugene.
Gibbs, J.T., 1997. African American suicide: a cultural paradox. Suicide and Life-Threatening Behavior 27 (1), 68–79.
Henry, A.F., Short, J., 1954. Suicide and Homicide. Free Press, Glencoe, IL.
Jamison, K.R., 1999. Night Falls Fast: Understanding Suicide. Vintage Books, New York.
Johnson, B.D., 1965. Durkheim's one cause of suicide. American Sociological Review 30, 875–886.
Joiner Jr., T.E., 2007. Why People Die from Suicide. Harvard University Press, Cambridge, MA.
Joiner Jr., T.E., Brown, J.S., Wingate, L.R., 2005. The psychology and neurobiology of suicidal behavior. Annual Review of Psychology 56, 287–314.
Kreitman, N., 1976. The coal gas story. United Kingdom suicide rates, 1960–71. British Journal of Preventive and Social Medicine 30 (2), 86–93.
Kubrin, C.E., Wadsworth, T.P., DiPietro, S., 2006. Deindustrialization, disadvantage and suicide among young black males. Social Forces 84, 1559–1579.
Lester, D., 1995. Is divorce an indicator of general or specific social malaise? Journal of Divorce and Remarriage 23, 203–205.
Luo, F., Curtis, S.F., Quispe-Agnoli, M., Ouyang, L., Crosby, A., 2011. Impact of business cycles on US suicide rates, 1928–2007. American Journal of Public Health 101 (6), 1139–1146. Online, e1–e8.
Maris, R.W., Berman, A.L., Silverman, M.M., 2000. Comprehensive Textbook of Suicidology. Guilford Press, New York.
Miller, M., Hemenway, D., 2008. Guns and suicide in the United States. New England Journal of Medicine 359, 989–991.
Miller, M., Lippmann, S.J., Azrael, D., Hemenway, D., 2007. Household firearm ownership and rates of suicide across the 50 United States. The Journal of Trauma, Injury, Infection, and Critical Care 62 (4), 1029–1035.
Pampel, F.C., 1998. National context, social change, and sex differences in suicide rates. American Socioogical Review 63, 744–758.
Pescosolido, B.A., 1990. The social context of religious integration and suicide: pursuing the network explanation. Sociological Quarterly 31, 337–357.
Pescosolido, B.A., 1994. Bringing Durkheim into the 21st century: a network approach to unresolved issues in the sociology of suicide. In: Lester, D. (Ed.), Le Suicide 100 Years Later. Charles, Philadelphia, pp. 264–295.
Pescosolido, B.A., Georgianna, S., 1989. Durkheim, suicide, and religion: toward a network theory of suicide. American Sociological Review 54, 33–48.
Pescosolido, B.A., Mendelsohn, R., 1986. Social causation or social construction of suicide? An investigation into the social organization of official rates. American Sociological Review 51, 80–100.
Phillips, D.P., 1974. The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. American Sociological Review 39, 340–354.
Phillips, D.P., 1982. The impact of fictional television stories on US adult fatalities: new evidence on the effect of the mass media on violence. American Journal of Sociology 87, 1340–1359.
Phillips, J.A., 2013. Factors associated with temporal and spatial patterns in suicide rates across US states, 1976–2000. Demography 50 (2), 591–614.
Phillips, J.A., Robin, A.V., Nugent, C., Idler, E., 2010. Understanding recent changes in suicide rates among the middle-aged: period or cohort effects? Public Health Reports 125 (5), 680–688.
Pope, W., Danigelis, N., 1981. Society's 'one law'. Social Forces 60, 495–516.
Popovici, I., French, M.T., 2013. Does unemployment lead to greater alcohol consumption? Industrial Relations 52 (2), 444–466.
Stack, S., 2000a. Suicide: a 15-year review of the sociological literature part I: cultural and economic factors. Suicide and Life-Threatening Behavior 30 (2), 145–162.
Stack, S., 2000b. Suicide: a 15-year review of the sociological literature part II: modernization and social integration perspectives. Suicide and Life-Threatening Behavior 30 (2), 163–176.
Stockard, J., O'Brien, R.M., 2002a. Cohort effects on suicide rates: international variations. American Sociolocial Review 67, 854–872.
Stockard, J., O'Brien, R.M., 2002b. Cohort variations and changes in age-specific suicide rates over time: explaining variations in youth suicide. Social Forces 81, 605–642.
Tarde, G.D., 1903[1895]. The Laws of Imitation. H. Holt, New York.
Thorlindsson, T., Bjarnason, T., 1998. Modeling Durkheim on the micro level: a study of youth suicidality. American Sociological Review 63, 94–110.
Timmermans, S., 2005. Suicide determination and the professional authority of medical examiners. American Sociological Review 70, 311–333.
van Poppel, F., Day, L.H., 1996. A test of Durkheim's theory of suicide—without committing the "ecological fallacy." American Sociological Review 61, 500–507.
van Tubergen, F., te Grotenhuis, M., Ultee, W., 2005. Denomination, religious context, and suicide: neo-Durkheimian multilevel explanations tested with individual and contextual data. American Journal of Sociology 111, 797–823.
Wadsworth, T., Kubrin, C.E., 2007. Hispanic suicide in US metropolitan areas: examining the effects of immigration, assimilation, affluence, and disadvantage. American Journal of Sociology 112, 1848–1885.
Wray, M., Colen, C., Pescosolido, B., 2011. The sociology of suicide. Annual Review of Sociology 37, 505–528.

Relevant Websites

http://www.apa.org/topics/suicide/index.aspx – American Psychological Association.
http://www.suicidology.org/home – American Association of Suicidology.
https://www.afsp.org/ – American Foundation for Suicide Prevention.
http://www.cdc.gov/violenceprevention/ – Centers for Disease Control.
http://durkheim.uchicago.edu/ – The Durkheim Pages.
http://www.hsph.harvard.edu/means-matter/ – Harvard School of Public Health, Means Matter.
http://www.nimh.nih.gov/ – National Institute of Mental Health.
http://www.who.int/mental_health/en/ – World Health Organization.

Sumner, William (1840–1910)

Robert Bannister, Swarthmore, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The son of British immigrants, William Graham Sumner led a varied career as clergyman, social theorist, and sociologist. After studying at Yale, Göttingen, and Oxford, he was an Episcopalian minister from 1867 to 1872. Resigning his ministry, he was then professor of political and social science at Yale from 1872 to 1909. In numerous books and essays he championed free trade, sound money, and minimalist government. An anti-imperialist, he opposed the Spanish–American War (1898). His book *Folkways* (1906) coined the terms 'folkways' and 'mores'. Despite this important contribution to early American sociology, Sumner remains better known as America's leading social Darwinist.

Overview

Born in Paterson, New Jersey, William Graham Sumner (30 October 1840–12 April 1910), political economist and sociologist, was the son of Thomas and Sarah Graham Sumner, British immigrants who left their native Lancashire at a time when technological advances were rendering obsolete the skills of an entire generation of artisans.

From a background 'as humble as any', as he later described it, Sumner became one of Yale's most popular and influential professors, the author of a dozen books and some 150 articles in America's leading journals of opinion. A staunch anti-imperialist, he opposed the Spanish–American war of 1898. His *Folkways* (1906) added the terms 'folkways', 'mores', and 'in-group' to the sociological lexicon. In 1908–09, Sumner served as a president of the recently created American Sociological Association (ASA). His defense of individualism, free market capitalism, and limited government later earned him a reputation as America's leading social Darwinist, although this characterization has been contested.

Early Life

A less than happy childhood shaped Sumner's convictions and personality. A 'plain mechanic', as Sumner described his father, Thomas escaped the hard times of Lancashire only to fall victim to the American depression of 1837. Having trained himself to build and repair the sort of machinery that drove his own father into poverty, Thomas left Paterson after 5 years, traveling throughout Pennsylvania, Ohio, New York, and finally Connecticut in an unsuccessful search of employment.

Thomas's lack of formal education and his economic failures often set son against father. But his influence on William is difficult to overstate. William's causes echoed the aspirations and fears of Jacksonian era artisans – among them, antimonopoly, free trade, and a belief in the power of education. Although often characterized as an uncritical defender of industrial capitalism, William did not share his generation's faith in progress. When asked years later where he acquired his critical, scientific outlook, he credited his father. As he sorted out his deceased father's nonexistent estate, his estimate of Thomas grew, making him a model for the 'forgotten man', the tax paying, hard working, artisan hero of Sumner's What Social Classes Owe to Each Other (1884).

The death of Sumner's mother when he was eight, and his father's immediate remarriage, placed him in the care of a penny-pinching stepmother who was as stingy with affection as with worldly goods. Sumner disliked her intensely, at one point planning half-seriously how he and his younger brother Joe might kill her. He tempered his opinion when she encouraged his application to Yale, even trimming the family budget to provide him funds. Although she died before this largesse could be appreciated, the damage was done. Throughout his life Sumner built walls to contain emotion. On the surface he was all 'root, hog, or die', the sour-faced theology student who neighborhood boys taunted with shouts of 'Hey mister, why don't you smile'. Feelings for Sumner were personal, to be channeled, for example, into the adventure stories wrote in his early school years, his performances in college theatricals, love letters to his wife-to-be, and a surprising fondness of children – affection once desired but not received (Bannister, 1979, 1987, 1992; Bellomy, 1980; Curtis, 1981).

Education

Education was Sumner's salvation. Although some neighbors found 'little Willie Sumner' to be a conceited, overbearing prig, teachers in Hartford's North Grammar school remembered the perfect teacher's pet. Forced by finances to leave school for 2 years to work in a dry goods store, he was rescued by a Hartford clergyman named Elias Beadle, who guided him to Hartford High School's collegiate program. From there he entered Yale College, financed initially by family funds suddenly available thanks to Thomas's brief success in real estate, and later supplemented by aid from Skull and Bones classmates. This support eventually included purchase of a Civil War substitute, overseas studies, and a sabbatical when Sumner suffered an emotional collapse in the early 1890s (Bellomy, 1980).

As a Yale undergraduate (1859–63) Sumner followed the narrowly prescribed course laid out in the Yale Report of 1828: 2 years of the classics, drilled mechanically by young tutors, a third year that added some physics, chemistry, and

astronomy, and a fourth year where students got to hear lectures on history, politics, and law by Theodore Dwight Woolsey and Noah Porter, Yale's current and future presidents. Sumner's views on political economy, however, were acquired not at Yale, but earlier in a workman's institute library in Hartford. There he discovered Harriet Martineau's *Illustrations of Political Economy* (1832), an introduction for laymen to the theories of Malthus, Ricardo, and other classical economists through fictionalized accounts of mankind's efforts to wrest a living from nature. When Sumner later returned to these authors in the original, his favorite was Malthus, the bedrock of his thought in the later 1870s and 1880s.

While the Civil War raged at home, Sumner spent 2 years studying theology at the Universities of Geneva, Göttingen, and Oxford. Although each taught different lessons, a common thread was the impact of science, not just the much touted 'warfare of science and religion' but the nature of science itself. Despite some overlap, Sumner was exposed to three conflicting views. At Yale, and again at Oxford, science was a strictly inductive Baconian realism rooted in the Common Sense Scottish philosophy, which had dominated discourse in American colleges since early in the century. In teaching by example, Martineau effectively translated theoretical classical economics into the daily struggles of real men and women, stories to which laypersons could easily relate.

In Germany, Sumner dabbled briefly with the rationalistic Higher Criticism, an attempt to discover the meaning of biblical texts through literary and historical analysis. Oxford returned him to the common sense of Richard Hooker's *Of The Laws of Ecclesiastical Polity* (1894–97) and finally to Anglicanism.

At Oxford Sumner also had passing contact with a tradition of positivism going back to August Comte and now restated in Thomas Henry Buckle's *History of Civilization in England* (1865–66). So viewed, a science of history provided predictions based on a perceived regularity in human affairs, a probabilistic conception of natural law that later underlay the statistical sociology that gained popularity in the decades after Sumner's death. Although Sumner grappled with this view in his final years, he rejected it then just as he had ignored it in his Oxford year (Bannister, 1987).

Preacher to Scholar in Politics

In 1868, Sumner returned to Yale as a tutor in classics, combining it with service as Deacon in New Haven's Trinity Episcopal Church. The following spring (1869) he gave up tutoring to serve as assistant to the director of Cavalry Church, New York, and the editor of the newly launched publication, the Living Church. From 1870 to 1872 he was rector of the Church of the Redeemer in Morristown, New Jersey.

On 17 April 1871 Sumner married Jennie Elliot the daughter of a New York merchant. Reared to privilege, she found the role of professor's wife exhausting, falling victim to recurrent illnesses that made her a burden rather than partner. Sumner was nonetheless a loving husband and good father to their two sons, Eliot and Graham (Curtis, 1981). For Sumner, home life provided a welcome refuge from his professional duties, reinforcing the walls he built between feeling and facts. Sumner's presidential address to the ASA. in 1908 was a defense of the traditional family.

Although the first few years of their marriage were relatively happy, Sumner later recalled the 1870s as the worst time of his life. Anglicans were then divided between High and Low Church factions, the first concerned with tradition and ritual, the second tinged with an unacceptable evangelicalism ("Calvinist in doctrine, Methodistical in behavior", he wrote his wife-to-be) (Curtis, 1981). Sumner allied himself with the moderate Broad Church faction, serving as editor of their journal, the Living Church – living because it addressed current issues, church because it was an institution rooted in tradition.

Religious Views

Preaching to his Morristown parishioners, Sumner tackled issues that would concern him throughout his career. The enemy was rationalism which led to a solipsistic individualism and a sterile materialism, each the subject of sermons in the early 1870s. In surprisingly rare references to evolutionary theory and the threat science posed to religion, he also sought a middle ground: Huxley and Darwin were correct in 'their original works' (their method being the careful observation of facts) but not in the speculations others drew from their work. Likewise Sumner outlined a middle way between 'Tradition and Progress', the title of another of his weekly sermons.

Sumner never openly repudiated religion, perhaps mindful of his father's warning that Christianity should not be junked until humanity had something better to put in its place (Bellomy, 1980). After leaving the clergy he continued to attend weekly services, shocking a student who observed his tough-minded professor taking up collection at New Haven's Trinity Episcopal Church. By the late 1870s any religious fervor was gone. Sumner put his religious beliefs in a drawer, he later quipped, only to find 'nothing there at all' the next time he looked (Curtis, 1981).

Politics

Sumner made his transition from preacher to social theorist the hard way, through immersion in practical politics, locally and nationally. In 1872, he accepted a chair of political and social science at Yale, a position he held until his retirement in 1909. From this platform he launched a brief career as scholar-in politics, a term, being popularized by the Nation magazine. Locally, he joined the Young Yale movement to modernize the curriculum and increase alumni members of the Yale Corporation. He again sought a middle ground between radical demands to 'keep up with the times' and conservative defense of what he termed 'mere traditionalism and stagnation' (Bannister, 1992).

In 1873–77, Sumner served as a New Haven alderman. Joining the New Haven Chamber of Commerce, he undertook a crusade for free trade and sound currency, the subjects of *A History of American Currency* (1874) and *Lectures on the History of Protection* (1877). He served on the Finance Committee of the American Social Science Association, a forerunner of the ASA, only to resign in protest over its congenital reformism. In 1877, he joined an electoral commission to investigate fraud in the Hayes–Tilden

presidential election. The following year he testified before a Congressional Committee investigation of labor unrest.

The corruption in local and national politics left Sumner profoundly disillusioned. He concluded that he could do more harm than almost any man in politics because he did not know the rules and did not want to learn them. He declined to vote in the 1880 election and eventually supported the crusade for civil service reform. Joining the Mugwumps, he voted for Grover Cleveland in 1884 and 1888, but later found little to choose between the protectionist McKinley and the silverite Bryan. By the late 1870s his focus shifted from narrow partisan issues to broader social concerns (Bannister, 1992; Curtis, 1981).

Social Theorist

Several controversies in the late 1870s and early 1880s laid the basis for Sumner's reputation as America's leading Spencerian, and eventually as a social Darwinist who buttressed his arguments for a dog-eat-dog competition, private property, and laissez-faire with bogus analogies drawn from Darwinian biology. The first confrontation was with Yale President Noah Porter over Sumner's use of Spencer's Study of Sociology (1874) in a senior class in social science in 1878. Their battle, which soon went public, was one of several early attempts to defend academic freedom at a time when academics were establishing the legitimacy of still fragile disciplines. As with the theological battles of his Göttingen-Oxford years, the issue was not so much religion vs science but the definition of science. As Porter put it in his *The American College and the American Public* (1870) the issue was not whether the colleges would teach theology but whether it would be 'theology according to Comte and Spencer, or according to Bacon and Christ' (Heyl and Heyl, 1976).

Sumner at the time was indulging a new-found fondness of Spencer, whose *Social Statics* (1851) he had earlier dismissed as excessively metaphysical. During the 1870s, Spencerisms crept into his prose in references to 'first principles', the effects of 'action and reaction' and even 'progress … from the simple toward the complex', although not yet the tell-tale Darwinian phrases 'natural selection' and 'survival of the fittest', the latter a phrase Spencer himself used sparingly in his sociological writings after coining it in the early 1860s. The Porter crisis finally resolved itself when Sumner quietly dropped the Spencer text, and offered sociology only to graduate students while giving top billing to anthropology, a discipline longer on facts than on speculative analogies.

That might have been the end of it had Sumner not indulged some unmistakable Darwinian flourishes in his escalating battles against socialism and other radical reforms. In an article 'Socialism' in Scribner's (1878) he based his argument almost entirely on the works of Thomas Malthus, soon to be the target of Single Taxer Henry George in *Progress and Poverty* (1879). Sumner then charged that George 'wasted … effort' because 'the Malthusian law is swallowed up in a great biological law' – an apparent allusion to evolutionary theory. In a speech in the Free Trade Club that same year, and in several published and unpublished pieces over the next half decade, Sumner went a step further. "The law of the survival of the fittest was not made by man," he wrote in probably his best-known statements. "We can only, by interfering with it, produce the survival of the unfittest" (Bannister, 1979).

Seemingly unaware that he had stepped on a third rail of Gilded Age discourse, Sumner soon faced a firestorm. *The New York Times*, in an editorial titled 'The Selfish Sciences' condemned his 'lame analogy'. A member of the Nineteenth Century Club called it 'the baldest of bald truisms'. Even the Nation, which shared many of Sumner's views, commented that his way of expressing himself was 'occasionally unfortunate'. Sumner never recanted, although he tacitly admitted that he had crossed the line. His grasp of sociology was weak, he confessed. All he could 'affirm with certainty' was that social laws were 'in their entire character like the laws of physics'. Regarding struggle and survival, he sharpened a distinction between the 'competition of life' and 'the struggle for existence', the first being a social contest among individuals, subject to social rules and conventions, the second the collective struggle of humans against nature. The first could be altered, the second could not. Most social reformers confused the two, ascribing to the competition of life hardships due to the collective struggle with nature. Following a public exchange in 1884 on the issue of survival and fitness in the journal of the Free Religious Association, he dropped the fittest-unfittest epigram, never to repeat it again (Bannister, 1979).

The Forgotten Man

A new tone in Sumner's thinking appeared with the publication of *What Social Classes Owe to Each Other* (1884), next to *Folkways* his best-known book. Some readers then and later recoiled at his seemingly callous reduction of all humanitarian reforms to a school book logic wherein 'A and B decide what C shall do for D' – 'C' in this calculus being the forgotten man. What critics failed to note was the absence of any references to struggle, survival, or fitness. Instead he added a chapter 'Wherefore we should love one Another', a feel-good sermon designed to counter his 'unfortunate' way of expressing himself during the previous half decade. Also missing in the reviews was any note of the fact that Sumner had lost his appetite for cosmic laws and abstract rights.

Returning to the crude induction of Harriet Martineau's *Illustrations of Political Economy*, he now defined nature as the physical rocks and soil from which men wrested their living. Rights were nothing more than the freedom to pursue this activity without interference. Sumner's conception of science had always been a seesaw affair. While Baconian empiricism (induction) seemed the safest route, its dangers surfaced in the reform efforts of the social worker faction of the American Social Science Association as well as in the work of some representatives of the German Historical School of economics, the so-called Socialists of the Chair, both of whom abandoned theory for catalogs of social ills needing attention.

Sumner's forgotten man resurfaced almost four decades later when in one of his radio Fireside Chats (April 1932) Franklin Roosevelt appropriated the term, not to describe the hard working, tax-paying artisan Sumner had pictured, but the down-and-out victim of economic depression. Although a contributor to *The New York Times* (18 September 1932) attempted to set the record straight, the phrase soon entered popular culture in ways Sumner would not have recognized,

for example, in singer Joan Blondell's rendition of 'Remember my Forgotten Man' in Gold Diggers of 1933, as well as in several movies and a TV show. Year later he reappeared in Amity Shales' *The Forgotten Man* (2007), a critical account of the New Deal that drew acclaim in Tea Party and libertarian circles.

Farewell to Reform

In the late 1880s, Sumner's attention shifted from Henry George to Karl Marx and Edward Bellamy, author of the Utopian romance *Looking Backward* (1888), and also from free trade and civil service reform to plutocracy. Rejecting the metaphysical caste of reformers' appeals to liberty, rights, and equality, Sumner continued to provide crudely empirical translations of traditional liberal values. Liberty meant only the ability to gather the fruits of labor unimpeded by restraints. Liberty was inseparable from property, which in turn was the basis of liberty. Rights were measured by stuff you could see and touch.

This distinction was subtle but important. Historically the language of rights implied restraints on the powers of the state and escape from the past. In Sumner's translation, these terms validated the status quo. He adopted a similar tactic to cut through the haze of Marxian (and Hegelian) rhetoric. The proletariat was simply the 'House of Want', the state nothing more than 'all of us'.

'The Absurd Effort to Make the World Over' (1894) reflected this shift in method and mood. In place of logic or natural law, he answered social reformers with an 'appeal to the facts', historical facts in particular. "[All] the allegations of general mischief, social corruption, wrong, and evil in our society must be referred back to those who make them for particulars and specifications," he wrote. "That is why it is the greatest folly of which a man can be capable, to sit down with a slate and pencil to plan out a new social world." But had not Sumner himself devoted two decades to restoring America to the Republican virtues he inherited from his father? To this extent, as historian Donald Bellomy has noted, 'The Absurd Effort' was Sumner's personal farewell to reform.

Sumner's attack on plutocracy was more than a passing phase. As with many Mugwumps, he initially used the term to describe the vulgar nouveau riche, but by the later 1880s narrowed the definition to those who made money through politics, and in turn polluted politics with their money "[Plutocracy] is the thing which seems to me to be really new and really threatening," he wrote in one of three articles that appeared in the Independent in 1888–89. "There have been states in which there have been large plutocratic elements, but none in which wealth seemed to have such absorbing and controlling power as it threatens us" (Bannister, 1992).

Anti-Imperialist to Sociologist

During the 1890s, the power of money in politics was a centerpiece in his battle against imperialism and militarism. "The great foe of democracy now and in the near future is plutocracy…," he declared in 'The Conquest of the United States by Spain' a speech delivered on 16 January 1899 before a packed house in New Haven's old College Street Hall. "In that war militarism, expansion, and imperialism will all favor plutocracy." Spain's 'conquest' was in the realm of values, not on the battlefield. In pursuing overseas expansion the United States was adopting the values that had brought Spain to lowly estate. "My patriotism is of the kind which is outraged by the notion that the United States never was a great nation until in a petty three months' campaign it knocked to pieces a poor, decrepit, bankrupt old state like Spain," he concluded, in a line that drew applause from a previously skeptical audience.

Sumner's battle against militarism and imperialism had important consequences for his developing sociology. The disjuncture between America's professed intentions (spreading liberty) and the war's outcome (acquisition of the Philippines) dramatized the difference between purposes and consequences, the title of an unpublished essay written sometime after the turn of the century. "Motives and purposes are in the heart and brain of man," he wrote. "Consequences are in the world of fact".

During these years, Sumner discovered the work of Ludwig Gumplowicz and Gustav Ratzenhofer, representatives of the 'struggle school' of Austrian sociology. From them he learned the importance of warfare in the early stages of society although he preferred Gumplowicz's realpolitik to Ratzenhofer's ethical idealism. He also embraced the work of Austrian cultural historian Julius Lippert who stressed a historical view of traditions and the importance of group effort in the struggle for existence against nature, a process Sumner soon termed 'antagonistic cooperation'.

The growing gap between Sumner's beliefs and the imperialist bluster of modern America caught him between a system of values he could not accept but despaired of changing. With the Spanish–American war hero Theodore Roosevelt in the White House, America was entering its dangerous 'glory days'. The alternative of the silver-tongued William Jennings Bryan seemed even worse. He would vote for T.R., in 1908, Sumner wrote, but only "in order to ward off Bryan and Hades". His list of pet peeves grew by the day, finally including 'advertisers who exaggerate', journalistic excess, 'electioneering devices' and 'oratorical and dithyrambic extravagances in politics'. Although these things were not 'properly part of the mores they were symptoms of them', he added, using a term he coined in 1899 when he first began work on his science of society (Bannister, 1992).

Folkways

In *Folkways* (1906) Sumner poured his disillusion and growing pessimism into one of the most important books in early American sociology. In all societies people are driven by four basic instincts of hunger, love, vanity, and fear, he argued. Through trial and error, one method of satisfying these demands becomes customary for a significant part of a society. These folkways, as he called them, gain a moral sanction through comparison and reflection. The mores are folkways grown moral and reflective. Good and bad have no meaning outside of the mores. Mores that survive command the support of the most powerful groups. Thus 'nothing but might has ever made right', he added, using a phrase that drew fire from critics then and later. To complicate things Sumner acknowledged

that the mores could sometimes be wrong and 'positively harmful', a curious judgment if good and bad existed only within the mores.

His reply to this criticism was that an educated elite (the classes) had introduced the scientific outlook into the mores in the modern period. Using history as its laboratory, science could ascertain which mores were conducive to social survival. The mores that is, contained a self-correcting element. The distinction between an anterior and a posterior view explained why he was not advocating a might-is-right policy as often charged. Before the fact, the might that could be employed in no way determined the right of an action. A scientific view after the fact (posterior) proved the superiority of some values over others. Rather than being an enemy of tradition, as historically had been the case since the Enlightenment, science properly defined could be its best defense.

This resolution was not without its problems as Sumner at times appeared to advocate rule by a scientific elite of experts not that different from the 'sociocracy' that sociologist Lester Ward and many others were proposing. But Sumner's narrowly Baconian definition of the mores as facts – the other half of his reading of Darwin – translated into the view that the mores have a coercive character that transcends mere utility – a view similar to that of the French sociologist, Émile Durkheim. These facts did not rule out prudent change, but explained why it was folly to sit down with pen and paper and design a new world.

Sociological Legacy

Since Sumner died before his magnum opus *The Science of Society* (1927) was completed, the full scope and final direction of his sociology remains unknown. At Yale both sociology and graduate studies generally were poorly funded. Sumner had only six graduate students in sociology, four of them predominantly in social work. At least two of these students translated Sumner's teaching into proposals for government action he would have rejected: the socialist Henry Pratt Fairchild in calls for immigration restriction, eugenics, and birth control, and Frederick E. Lumley in *Means of Social Control* (1925). The later Cross-Cultural Survey in Yale's Institute of Human Relations owed a debt to Folkways but was more directly the work of anthropologist G.P. Murdock.

Reviews of Folkways were mixed. Some praised Sumner for putting sociology on a more scientific (objective) path. But even this praise was mingled with snide asides. Chicago sociologist George Vincent wrote that the ethnological data "seem at times to overweigh the book by their sheer bulk and multiplicity". Charles Ellwood, a Chicago PhD now professor at the University of Missouri, observed that the mores were not facts in the same sense as chairs and tables. "Whoever saw a tradition", he asked? Robert Park of Chicago, an advocate of a more objective sociology, said that Sumner's method 'if it can be called a method', consisted mostly of collection and description. Some refused to consider Sumner a sociologist at all. Vincent called him a folk psychologist. Albion Small, chair of Chicago's department, confessed that he never thought of Sumner as a sociologist until the Yale professor was elected president of the ASA (Bannister, 1987).

In the short run, Sumner's greatest influence was among several graduate students elsewhere. At Chicago, Luther Bernard drew directly on Folkways for his thesis 'An Objective Standard of Social Control', although like Sumner's own students he translated his sociology into a radical politics. At Columbia, F. Stuart Chapin, later the source of the widely used 'living room scales', wrote a thesis on Education and the Mores (1911). William F. Ogburn, later author of *Social Change* (1922) and the concept of cultural lag, used Sumner's work directly in a thesis on child labor legislation (1912). Franklin Giddings, chair at Columbia and one of the pioneers of American sociology, later called Sumner "perhaps the most consistently sociological if not the greatest of sociologists" (Bannister, 1987).

Despite their common dedication to objective sociology, the work of Ogburn, Chapin, and their many students differed from Folkways in being statistical rather than descriptive. In his thesis on child labor legislation, Ogburn charted existing laws statistically, defining 'best' as those closest to the norm, not by examining social mores regarding children and education. Like Sumner this new breed of statistical sociologist urged a more scientific approach to social problems. But, they generally supported Franklin Roosevelt's New Deal rather than the policies Sumner championed half a century earlier.

Sumner left no school of sociology, only a handful of disciples, mostly at Yale, and a book his biographer has described as an isolated masterpiece with no forebears or successors. However, the many references to mores and folkways in the scholarly literature of the past century reveal that his work has had wide-ranging if indirect influence outside as well as within sociology.

Social Darwinism

Sumner's contribution to sociology was soon eclipsed in the public mind by his reputation as America's premiere social Darwinist. Although this label as applied to Sumner has its roots in his rhetorical skirmishes in the early 1880s, it was Spencer who first earned the label 'le Darwinisme sociale' from European socialists, especially in response to his *Man versus the State* (1884). The term in its English translation 'social Darwinism' did not appear in the United States until the turn of the century, typically applied to a diverse group of thinkers who had little in common beyond their alleged misuse of Darwinism – eugenists, Nietzcheans, and German militarists heading the list (Bannister, 1979).

Until the 1940s the term social Darwinism rarely appeared in either scholarly literature or the popular press. This changed dramatically with the appearance of Richard Hofstadter's *Social Darwinism in American Thought* (1944). A brilliant tour deforce synthesizing more than five decades of debate over social applications of Darwinism, Hofstadter's book became one of the most influential works in American history. Although reviewers quibbled over his definitions, most praised his work. In their enthusiasm, however, no one noted that William Graham Sumner was included in the ranks of social Darwinists for the first time.

An exception was Sumner's student-turned-acolyte Albert G. Keller. "I find a scattering of social scientists as contrasted to social Darwinists on the Hofstadter roster," Keller wrote with Sumner in mind. "Some persons hopelessly allergic to 'isms' and 'ismics' have been ushered into the wrong pew". Ironically, Keller himself had contributed to Hofstadter's image of Sumner. During the years immediately preceding the publication of Folkways, he lectured Sumner on lessons to be learned from the Origin of Species, a work Keller himself first encountered in debates over neo-Darwinism in the 1890s. The first lesson concerned the importance of fact gathering, the second the role of struggle, selection, and transmission in the shaping of human cultures, a theme Keller later developed in his *Social Evolution* (1915) (Bannister, 1987; Smith, 1979; Smith and Hinkle, 1979).

Keller's celebration of facts was Baconian in the extreme. Fact gathering provided mental discipline formerly served by Yale's now defunct classical curriculum. Facts build character. With Darwin there was 'no sensitiveness about personal glory'. Facts were manly. "Lord knows there are enough 'It-Books', not to mention the 'She-Books'," Keller later wrote to sociologist Edward A. Ross. On this scale Sumner outdid even Spencer. "Spencer was something of an old maid and perhaps a little feline now and then," Keller reminisced. "But Sumner, with all his hard hitting … ought to appeal to anyone who, in these times, prays: 'O God, give us a Man!'" (Bannister, 1979).

Concerning facts, Sumner needed little persuasion. In more than 50 files boxes he collected thousands of note cards on traditions ranging from sports to incest. In Folkways he treated readers to a massive array of facts. But speculation based on analogies drawn from biology remained off-limits. For this every reason, he would not accept Keller's evolutionary view of the mores or his social selectionism. He underlined this point in a talk to Yale's Anthropology Club in 1905, although in deference to his mentor did not include these remarks in Folkways as originally intended.

Folkways were Darwinian only in the sense that it consisted of what Keller once termed a 'shovelfuls of facts', that is, hardly Darwinian in any recognizable sense. The word 'fittest' appeared only once with no reference to Darwin or biology. A chapter titled 'Societal Selection', described the triumph of some mores over others not in terms of struggle and survival of individuals or groups, but a subtler process he termed 'syncretism'. He continued to insist that the 'competition of life' and 'struggle for existence' were distinct albeit parallel phenomena.

Despite Keller's failure to convert Sumner to an evolutionist view of the mores, he did his best in later years to convey the impression that Sumner would eventually have capitulated. In a review of Folkways Keller stated that Sumner's book "constitutes a distinct and characteristic amplification of the evolutionary theory", adding that Sumner would agree.

In editing *The Challenge of Facts* (1914), the first of four volumes of Sumner's essays, Keller included several pieces emphasizing Sumner's debt to Darwin and Spencer. The first was a 'Sketch' of Sumner that appeared in the Spencerian journal Popular Science in 1889 based on Sumner's answers to a series of questions seemingly designed by the editors to show his familiarity with the work of Spencer and Darwin. The two others were previously unpublished manuscripts 'The Challenge of Facts' (originally 'Socialism') and 'The Predicament of Sociological Study'. As it happened, each included the epigram that the alternative to the 'survival of the fittest' was 'the survival of the unfittest', two of a handful of such comments in the early 1880s. Keller dated the first as sometime in the 1880s, and the second as after 1900s, leaving the impression that Sumner used such phraseology throughout his career. In Sumner's posthumous four-volume Study of Society, coauthor Keller gave the mores a Darwinian selectionist spin, echoing his own Societal Evolution (1915) (Bannister, 1987; Smith, 1979; Smith and Hinkle, 1979).

Keller's efforts paid off. "The fidelity of the work [Science of Society] to Sumner's major conceptions," Hofstadter wrote "is such that I have not hesitated to use it as a source". Writing in the journal Social Forces a few years later, sociologist Donald W. Calhoun agreed. "In evaluating Sumner's total contribution to the science of society, we shall follow these other commentators in treating the Science of Society as a source fundamentally parallel to Folkways and the essays".

Calhoun went on to praise the cosmic sweep of Sumner's investigations, his thoroughly sociological view of human society as embodied in the terms folkways, mores, and in-group, his divorce of value judgments from sociological inquiry, and his contribution to 'the behavioristic treatment of sociological data'. These points conceded, he then launched a blistering critique of Sumner's social views, capped by an attack on his social Darwinism (Calhoun, 1945).

A revisionist critique of the Hofstadter thesis first appeared in the early 1950s, initially in an analysis of the 'self-made man' ideology (Wyllie, 1959) and eventually as related to Sumner himself (Bannister, 1973; Bannister, 1979; Bellomy, 1984). Moderate revisionists argued that charges of social Darwinism misrepresented Darwin's influence on social thought in the Gilded age, Sumner in particular. Applied to society, Darwin's image of nature red in tooth and claw pictured a social order few if any Americans wished to defend in those terms. Scattered attempts to provide such a defense, met the vigorous opposition that greeted Sumner's rhetorical flourishes in the early 1880s. Portrayals of society in Darwinian terms were not intended to defend competition or laissez-faire but typically prefaced proposals that included the full range of progressive era reforms, including socialism, eugenics, immigration restriction, and Jim Crow racial restrictions.

A more extreme revision of the Hofstadter thesis argued that social Darwinists exist only in the minds of those who invented them, a socially constructed label with special appeal in a nation where almost half the population professes not to believe in Darwinism (Bannister, 1979; Bellomy, 1984).

While casting a shadow on his sociological legacy, a one-dimensional picture of Sumner as social Darwinist has also complicated efforts to place him in an American conservative tradition. In *The Conservative Mind* (1953), Russell Kirk, one of the leaders of the New Conservatism of the 1950s, pointedly excluded Sumner. More than a century after Sumner's death self-identified conservatives remain uneasy including him in the conservative canon. While opponents of the Tea Party have pictured the movement as heir to Sumner's social Darwinism (meaning a brutal, unfettered capitalism), many Tea Party supporters themselves sidestep Darwinism and instead embrace Sumner's views on the

importance of family or, alternately, see themselves as twenty-first century reincarnations of Sumner's forgotten men.

For some libertarians, Sumner's alleged use of Darwinism is itself the problem. If we can successfully distinguish evolution from liberty, one libertarian wrote, "we can save the Social Darwinists from themselves".

Historical revisionism has had limited success countering Sumner's reputation as dog-eat-dog social Darwinist. Some historians concede that Sumner and other conservative social Darwinists cited Darwin less than once believed. Others argue that Sumner is better termed a social Spencerian, a kinder, gentler social Darwinist. However, a Google/Lexis Nexis search of 'William Graham Sumner' and 'social Darwinism' tells another story, as hundreds of hits reveal that, in the public mind, Sumner is best known as America's leading social Darwinist (Hawkins).

See also: Capitalism; Comte, Auguste (1798–1857); Conservatism: Historical Aspects; Darwin, Charles Robert (1809–82); Durkheim, Emile (1858–1917); Evolution, Natural and Social: Philosophical Aspects; Liberalism: Historical Aspects; Libertarianism; Malthus, Thomas Robert (1766–1834); Marx, Karl (1818–83); Spencer, Herbert (1820–1903); The American Conservative Movement.

Bibliography

Bannister, Robert C., 1973. William Graham Sumner's 'Social Darwinism'. History of Political Economy 5, 89–109.
Bannister, Robert C., 1979. Social Darwinism: Science and Myth. Temple University Press, Philadelphia, PA.
Bannister, Robert C., 1987. Sociology and Scientism: The American Quest for Objectivity 1880–1940. University of North Carolina Press, Chapel Hill, NC.
Bannister, Robert C. (Ed.), 1992. On Liberty, Society and Politics, The Essential Essays of William Graham Sumner. Liberty Fund, Indianapolis, IN.
Bellomy, Donald C., 1980. The Moulding of an Iconoclast: William Graham Sumner, 1840–1885. Harvard University, unpublished doctoral dissertation.
Bellomy, Donald C., 1984. Social darwinism revisited. Perspectives in American History 1, 1–129.
Burton William L., 1959–1960. The Conservatism of William Graham Sumner. Modern Age 4, 45–51.
Byrne, William F., 2010. William Graham Sumner and the problem of liberal democracy. Review of Politics 72, 571–597.
Calhoun, Donald W., 1945. William Graham Sumner. Social Forces 24, 15–32.
Curtis, Bruce, 1969. William Graham Sumner on the concentration of wealth. Journal of American History 55, 823–832.
Curtis, Bruce, 1977. Victorians abed: William Graham Sumner on the family, women, and sex. American Studies 18, 101–122.
Curtis, Bruce, 1981. William Graham Sumner. Twayne, Boston, MA.
Hawkins, Mike, 1997. Social Darwinism in European and American Thought. Cambridge University Press, Cambridge.
Heyl, John, Heyl, Barbara, 1976. The Sumner-Porter controversy at Yale. Sociological Inquiry 46, 41–49.
Keller, Albert G., 1933. Reminiscences Mainly Personal of William Graham Sumner. Yale University Press, New Haven, CT.
Lemert, Edwin M., 1942. The folkways and social control. American Sociological Review 7, 394–399.
Marshall, Jonathan, 1979. William Graham Sumner: critic of progressive liberalism. Journal of Libertarian Studies 3, 261–277.
Notestein, Robert B., 1955. The moral rigorism of william Graham Sumner. Journal of the History of Ideas 16, 389–400.
Notestein, Robert B., 1959. William Graham Sumner. American Journal of Economics and Sociology 18, 397–413.
Park, Robert E., 1933. Sumner's conception of society. Chinese Social and Political Science Review 17, 230–241.
Pickens, Donald K., 1968. William Graham Sumner: moralist as social scientist. Social Science 43, 202–209.
Pickens, Donald K., 1987. Scottish common sense philosophy and folkways. Journal of Thought 22, 39–44.
Richard, Hofstadter, 1944. Social Darwinism in American Thought. University of Pennsylvania Press, Philadelphia, PA.
Robert, Garson, Richard, Maidment, 1981. Social darwinism and the liberal tradition. South Atlantic Quarterly 80, 61–76.
Shone, Steve J., 2004. Cultural relativism and the savage, the alleged inconsistency of William Graham Sumner. American Journal of Economics and Sociology 63, 697–715.
Sijuwade, Philip, 1999. William Graham Sumner and the Scottish moral philosophers of the eighteenth century. International Review of Modern Sociology 29, 49–57.
Smith, Norman E., 1979. William Graham Sumner as an anti-social Darwinist. Pacific Sociological Review 22, 332–347.
Smith, Norman E., Hinkle, Roscoe C., 1979. Sumner versus Keller and the social evolutionism of early American sociology. Sociological Inquiry 49, 41–48.
Starr, Harris E., 1925. William Graham Sumner. Holt, New York.
Trask, H. Scott, 2004. William Graham Sumner: against democracy, plutocracy, and imperialism. Journal of Libertarian Studies 18, 1–27.
Wyllie, Irvin G., 1959. Social darwinism and the businessman. Proceedings of the American Philosophical Society 103 (5), 629–635.

Relevant Website

http://oll.libertyfund.org/ – LibFund. The full URL for Sumner is: http://oll.libertyfund.org/?option=com_staticxt&staticfile=show.php?person%3D236&Itemid=28.

Sun Exposure and Skin Cancer Prevention

Natalie Schüz, School of Medicine, University of Tasmania, Hobart TAS, Australia
Michael Eid, Educational Science and Psychology, Freie Universität Berlin, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Eid, volume 22, pp. 15278–15281, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of the risks of sun exposure and reviews the psychological literature on sun-tanning behavior, sun protection behavior, and skin cancer prevention. The different forms of skin cancer, their incidence, and causation are described briefly. The psychological predictors of sun exposure and sun protection behavior are reviewed. Finally, the effectiveness of skin cancer prevention programs is examined.

Sun exposure can be defined as the exposure of human beings to natural and artificial ultraviolet radiation (UVR). UVR emitted by the sun and devices such as tanning beds is a known human carcinogen (El Ghissassi et al., 2009). Accordingly, exposure to UVR is considered to be the main causal factor of all types of skin cancer (Armstrong, 2004). Appropriate sun protection behavior is an effective and essential means to prevent premature skin aging, as well as skin cancer at any age.

This article provides an overview of the incidence and causation of skin cancer, of the psychological determinants of sun exposure and sun protection behavior, and of different types of skin cancer prevention programs.

Skin Cancer: Incidence and Causation

There are two major classes of skin cancer: melanoma and nonmelanoma skin cancers (NMSC; subtypes: basal cell carcinoma (BCC), squamous cell carcinoma (SCC)). Three-fourths of all skin cancer-related deaths result from melanoma, indicating that melanoma is a more severe form of skin cancer than NMSC, which is often self-detected and successfully treated in doctor's surgeries.

Melanoma and NMSC Incidence

While the worldwide incidence rate of melanoma is lower than that of other types of cancer such as breast-, prostate-, and lung cancer, melanoma is the most common type of cancer diagnosed in Australian adolescents and young adults, accounting for more than 25% of all cancers in the age group between 15 and 29 years (Stiller, 2007). The age-standardized incidence rate for melanoma in Australia/New Zealand is 54.0 cases per 100 000, 19.7 cases in Northern America, and 11.7 cases in Europe, while incidence rates in Africa and Asia are much lower (0.9 and 0.3 cases per 100 000 respectively) (Ferlay et al., 2010).

Even though NMSC is a less severe type of skin cancer than melanoma, it is usually the most common form of cancer diagnosed in light-skinned populations, with incidence rates in some countries as high or higher as those for all other cancers combined. For example, in the annual report of the American Cancer Society (American Cancer Society, 2013), it is estimated that 3.5 million new cases of NMSC were diagnosed in the United States. This compares to a total number of 1.66 million annually diagnosed cancers across all body sites. Thus, NMSC constitutes a major burden in terms of both quality of life and treatment costs. As most national cancer registries do not routinely collect data on NMSC, it is difficult to obtain standardized estimates of worldwide incidence rates, especially for dark-skinned populations. A systematic review suggests a yearly increase of 5.5% in BCC across Europe (Lomas et al., 2012). The same study reported the highest worldwide incidence rate of BCC for Australia (ranging between 884 and 1813 per 100 000), while rates in Europe and Northern America ranged between 33.6 and 147.0 per 100 000. Rates of SCC are lower than those of BCC, ranging between 8.9 and 290 per 100 000 in Europe and Northern America, and between 387 and 1035 per 100 000 in Australia (Lomas et al., 2012). For both, BCC and SCC, incidence rates differ according to latitude, with more diagnosed cases of NMSC in light-skinned populations living closer to the equator.

Skin Cancer Risk Factors

There are several constitutional, environmental, and behavioral risk factors for skin cancer. Paler skin that burns easily, a personal or family history of skin cancer, and certain medical conditions that suppress the immune system are the most common constitutional factors. Additionally, atypical, numerous moles (i.e., more than 50) are risk factors for melanoma and there is an increased risk for NMSC developing on the sites of old scars or burns. UVR is the most important risk factor for any type of skin cancer (Armstrong, 2004). Thus, a history of excessive sun exposure, sunburns, and use of tanning booths increases the risk for skin cancer. Skin cancer sites and types vary according to exposure patterns and latitudes. Melanoma on less frequently exposed body sites is often associated with recreational sun exposure and sunburns across all latitudes, while melanoma on more exposed sites is associated with total lifetime and occupational sun exposure, especially at lower latitudes (Chang et al., 2009). NMSC is most

frequently found on commonly exposed body sites, such as ears, face, neck, and forearms. Studies suggest that childhood and recreational sun exposure is more related to BCC, whereas cumulative and occupational exposure mainly determines the risk for SCC (Armstrong, 2004). Sun exposure up until the age of 18 years seems to be causing the most critical damage with regard to cancer and premature skin aging; however, damage is still occurring in people aged 50 years and over (Kimlin and Guo, 2012). Overall, UVR was estimated to have caused the loss of 1.5 million disability-adjusted life years and 60 000 premature deaths in the year 2000 (Lucas et al., 2008).

There are three different types of radiation: UVA, UVB, and UVC radiation. UVC radiation does not play a role for the development of skin cancer, since it is blocked by the ozone layer. In the past, UVA radiation was thought to be less carcinogenic than UVB, since it induces less direct DNA damage. Recent studies, however, show that UVA is more carcinogenic than previously thought and that the basal layer of the skin is particularly vulnerable to UVA radiation (Agar et al., 2004). Therefore, many countries have introduced policies that require UVA protection labeling on sunscreens next to the sun protection factor (SPF), which only refers to UVB protection. UVR levels vary depending on season, time of day, latitude, cloud cover, altitude, and ground surface. The UV Index (UVI) is often reported in conjunction with weather forecast to alert the general public of the amount of harmful UVR on a particular day. The UVI is presented on a scale from 0 to 11+ and protection is required above the threshold value of 3 (Allinson et al., 2012).

Studies suggest that sun exposure is causing 99% of all NMSC and 95% of all melanoma (Armstrong, 2004). As skin cancer is a highly preventable disease, it is important to explore the factors influencing sun exposure and adequate sun protection measures.

Predictors of Sun Exposure

Sun exposure not only accompanies many occupational and recreational outdoor activities, it is often sought intentionally. One of the most important predictors of sun exposure is the desire for a suntan (Schüz and Eid, 2013). Even though the preference for tanned skin seems to be decreasing in countries such as Australia, where skin cancer awareness campaigns are well established, a significant proportion of light-skinned people are still actively seeking a tan because of a perceived increase in attractiveness (Stanton et al., 2004). Voluntary sun exposure is related to thinking more about one's physical appearance, to a higher fear of negative evaluations by others, and to a higher body self-consciousness and body self-esteem (Leary and Jones, 1993). Even though intention is usually a good predictor of behavior, appearance motivation seems to be a stronger predictor of sun exposure than the intention to avoid overexposure to the sun (Schüz and Eid, 2013). Next to the perception of improved appearance, relaxation also seems to play a role as a driver of UVR exposure. More than 76% of tanning bed users state that relaxation is a reason for tanning (Schneider et al., 2013). Moreover, when frequent tanners are given the choice between UV and non-UV tanning beds under blinded conditions, they are not only able to correctly distinguish between the conditions, but also actively seek UV exposure because of its relaxing effects (Feldman et al., 2004). In addition to appearance motivation and relaxation, sun tanning is related to a positive attitude toward risk-taking in general and to having friends who sunbathe and use tanning beds. Age and gender are important demographic predictors of sun exposure: Adolescents tend to spend more time in the sun than adults, and women report more deliberate tanning behavior than men (Stanton et al., 2004).

Predictors of Sun Protection Behavior

Deliberate sun exposure and the use of sun protection are two largely independent and only weakly correlated behaviors. Moreover, the predictors of sun protection behavior differ from the predictors of sun exposure. In contrast to sun exposure, appearance-oriented variables are less important for predicting sun protection.

The four most important sun protection behaviors are (1) avoiding direct sunlight, particularly during the peak hours of daylight (10 a.m.–4 p.m.), (2) seeking shade, (3) wearing a wide-brimmed hat, (4) wearing long-sleeved, dark clothing, and (5) using and reapplying sunscreen with an SPF of at least 15. These protective measures are endorsed whenever the UVI reaches a level of three or above. Additional recommendations include wearing sunglasses to avoid eye damage caused by UVR and tinting of car windows to block UVR. There have been some controversies regarding the effectiveness of sunscreen use for the prevention of skin cancer (Weinstock, 1999). However, studies show that adequate sunscreen use reduces the risk for both melanoma and SCC (Green et al., 2011). It has to be noted, however, that most people do not adhere to the usage guidelines for sunscreen and apply far less than what is recommended by manufacturers (around 35 ml per application), or do not reapply sunscreen regularly. Furthermore, sunscreen is often seen as a justification to overextend deliberate sun exposure (Stanton et al., 2004).

Even when awareness for the need of sun protection is high, many people do not adequately adhere to the recommendations, even in high-risk countries such as Australia (Stanton et al., 2004). There are several demographic predictors of sun protection: Women tend to report better sun protection than men, particularly with regard to sunscreen use. Light-skinned individuals protect themselves more than dark-skinned individuals, and adults report better protection levels than adolescents (Bränström et al., 2010).

Self-efficacy beliefs and perceived benefits and barriers of different sun protection behaviors are important sociocognitive predictors of sun protection, while many studies show that risk perception, operationalized as perceived vulnerability and/or severity of skin cancer, plays less of a role (Craciun et al., 2012b; Bränström et al., 2010). Sun protection self-efficacy beliefs pertain to a person's confidence that he or she is able to actually perform sun protection even in the face of difficulties (e.g., forgot to bring along sunscreen). There are both direct effects of self-efficacy on protection behavior and indirect effects mediated via sun protection intentions (Craciun et al.,

2012b). Commonly reported barriers for sunscreen use include the belief that sunscreen is greasy and sticky and the fact that sunscreen has to be applied repeatedly. Barriers against wearing hats are beliefs that hats are inconvenient to wear and cause problems with hairstyle. The major barriers against wearing shirts are the discomfort from heat and the feeling of being overdressed (Paul et al., 2008). Therefore, a goal of skin cancer prevention programs must be the removal of these specific barriers against sun protection behavior. Because these barriers partly depend on temporary fashion influences and differ between target groups, current and target group specific barriers of sun protection behaviors have to be explored before a prevention campaign is planned.

Social factors also have an important influence on sun protection. Use of sun protection is correlated with peers' sun protection behavior, parental influence, and parental sun protection behavior. Furthermore, people who know a person who has been diagnosed with skin cancer use sun protection more often (Leary and Jones, 1993).

To summarize, sun exposure and sun protection are two largely independent behaviors, influenced by a distinct set of determinants. While appearance motivation is an important predictor of sun exposure, self-efficacy beliefs and perceived benefits and barriers are more important drivers of sun protection. Moreover, behavior and beliefs differ according to sociodemographic group; thus, potential interventions may benefit from a targeted approach.

Skin Cancer Prevention

Interventions targeting sun exposure and sun protection behavior can take place on the population-, community- or individual level. Moreover, such interventions can focus on different aspects such as changes in policy (e.g., prohibiting the sale of sunscreen with an SPF under 30), education and communication (e.g., providing individual advice or making use of media campaigns), introducing technology (e.g., use of UV dosimeters to measure UVR), or providing people with resources (e.g., distributing free sunscreen). Due to the costs involved in appropriately evaluating different types of interventions, there is usually better evidence available for individual-oriented approaches than for population-based or community programs.

Examples of Population- and Community-Based Programs

The Australian SunSmart program is an example of a successful and well-evaluated mass-media campaign (for publications, see Relevant Websites). First introduced in 1988, more than 95% of the population is now familiar with its slogan Slip! (on a clothing) Slop! (on sunscreen) Slap! (on a hat). It is currently estimated that for every dollar spent on the program, $2.30 are saved in terms of cost for the health care system. The campaign aims to prevent 10 000 premature deaths, reduce melanoma incidence by 80 000 and NMSC incidence by 111 000 cases over the next 20 years (Shih et al., 2009). In conjunction with the SunSmart mass-media campaigns, sun protection in Australia is also enforced by policy: a study showed that 95% of childhood services reported to have a written sun protection policy which required all children to wear a hat and to use sunscreen (Ettridge et al., 2011). Moreover, having such a written sun protection policy is a requirement for accreditation of childcare centers. With regard to providing resources, a majority of Australian outdoor workers report that their employer provides them with sunscreen, protective clothing, and hats (Safe Work Australia, 2010).

The Pool Cool program is a successful and well-evaluated community-based program disseminated throughout the United States. With a theoretical basis in social cognitive theory, a main focus of the program is to provide people with sun protection skills, while taking into account environmental and structural changes, such as the availability of shaded areas and sun protection policies. The program was shown to increase sun protection among children and their parents, as well as to support the introduction of sun protection policies and environmental changes (for publications, see Relevant Websites).

More recently, smart phone applications that target sun protection behavior and skin cancer detection behavior have become available (e.g., the 'Solar Cell' app or the 'Know your skin' app). These provide real-time advice, e.g., via text messages, UVI forecasts, or educational and motivational messages, sometimes even tailored to the specific user. One prominent example is the 'Know your own skin' app that lets users take pictures of their moles and track potential changes over time by sending regular reminders to perform a skin self-check. Future evaluations of these types of intervention programs will show whether they are useful tools to prevent skin cancer.

Individual-Oriented Interventions

In contrast to population- and community-based programs, individual-oriented approaches to skin cancer prevention allow for better targeting of intervention contents. For example, as Schüz and Eid (2013) have shown, health-related time perspective (i.e., the extent to which future health is discounted over short-term outcomes such as getting a tan) is an important predictor of sun exposure intentions in adolescents, which can be successfully targeted in interventions. However, time perspective is less likely to play a role in adults, who are usually more aware of the future consequences of their behaviors as teenagers. Likewise, individuals who do not intend to use sun protection benefit more from an intervention involving resource communication, whereas motivated individuals profit more from an intervention involving planning when, where, and how to use sunscreen and how to face barriers toward sunscreen use (Craciun et al., 2012a).

In order to target appearance motivation, many interventions make use of UV photography. These are pictures taken under UV light that highlight damage to skin cells caused by UVR. Damaged areas appear as dark spots and are indicative of premature skin aging and higher risk for skin cancer. Some studies found UV photos to be effective in increasing sun

protection intentions and decreasing the use of tanning booth; however, the overall evidence for UV photos is inconclusive (Hollands et al., 2010). Moreover, messages that emphasize the negative effects of the sun on one's appearance can even lead to increased sun exposure intentions among individuals with high appearance motivation (Jones and Leary, 1994). This boomerang effect can be interpreted as a defensive reaction; messages highlighting the negative effects of tanning on appearance are challenging for individuals with a high appearance motivation, as they suggest that sun exposure has the opposite effect of what they want to achieve: it reduces their attractiveness rather than increasing it. However, a recent study has shown that an intervention combining UV photos with self-affirmation messages that enable individuals to process risk information less defensively reduces sun exposure among individuals with a high appearance motivation (Schüz et al., 2013).

Even though a range of evaluation studies on individual-oriented interventions targeting sun exposure and sun protection behavior is available, to date no comprehensive systematic review or *meta*-analysis exists that summarizes effective behavior change techniques in this area. However, a more general *meta*-analysis on health behavior change confirms the notion that tailored messages outperform more general health and risk messages, and that interventions addressing theory-based determinants of health behavior change are more effective than non-theory-based interventions (Noar et al., 2007).

Future Directions

Systematic reviews and *meta*-analyses summarizing the effects of interventions targeting sun exposure and sun protection behavior are needed in order to determine the effectiveness of different intervention components, thus providing much needed evidence for future intervention programs. A relatively recent field of research involves interventions using smart phone technology to target sun protection behavior. While this is a promising approach as smart phone applications are able to provide real-time advice in a tailored manner whenever the user asks for it, future evaluation studies are needed to prove the effectiveness of these applications. Moreover, future research will profit from the use of multimethod assessment strategies of behavior that does not only assess self-reported behavior but also uses observational and physical measures to validate self-reports (e.g., spectrophotometers for measuring the melanin skin content). As memory biases play a role when assessing sun protection and sun exposure via self-reports, ecological momentary assessment studies that assess behavior as it occurs in real time are a promising avenue for research on skin cancer prevention.

See also: Adolescent Health and Health Behaviors; Health Behavior, Psychosocial Theories of; Health Behaviors, Assessment of; Health Education and Health Promotion; Health Interventions, Community-based; Health Risk Perception; Self-Efficacy and Health.

Bibliography

Agar, N.S., Halliday, G.M., Barnetson, R.S., et al., 2004. The basal layer in human squamous tumors harbors more UVA than UVB fingerprint mutations: a role for UVA in human skin carcinogenesis. Proceedings of the National Academy of Sciences of the United States of America 101, 4954–4959.

Allinson, S., Asmuss, M., Baldermann, C., et al., 2012. Validity and use of the UV index: report from the UVI working group, Schloss Hohenkammer, Germany, 5–7 December 2011. Health Physics 103, 301–306.

American Cancer Society, 2013. Cancer Facts and Figures 2013. American Cancer Society, Atlanta.

Armstrong, B.K., 2004. How sun exposure causes skin cancer: an epidemiological perspective. In: Hill, D., Elwood, J.M., English, D.R. (Eds.), Prevention of Skin Cancer. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 89–116.

Bränström, R., Kasparian, N.A., Chang, Y-m, et al., 2010. Predictors of sun protection behaviors and severe sunburn in an international online study. Cancer Epidemiology Biomarkers & Prevention 19, 2199–2210.

Chang, Y.-M., Barrett, J.H., Bishop, D.T., et al., 2009. Sun exposure and melanoma risk at different latitudes: a pooled analysis of 5700 cases and 7216 controls. International Journal of Epidemiology 38, 814–830.

Craciun, C., Schüz, N., Lippke, S., et al., 2012a. Enhancing planning strategies for sunscreen use at different stages of change. Health Education Research 27, 857–867.

Craciun, C., Schüz, N., Lippke, S., et al., 2012b. A mediator model of sunscreen use: a longitudinal analysis of social–cognitive predictors and mediators. International Journal of Behavioral Medicine 1, 65–72.

El Ghissassi, F., Baan, R., Straif, K., et al., 2009. A review of human carcinogens – part D: radiation. Lancet Oncology 10, 751–752.

Ettridge, K.A., Bowden, J.A., Rayner, J.M., et al., 2011. The relationship between sun protection policy and associated practices in a national sample of early childhood services in Australia. Health Education Research 26, 53–62.

Feldman, S.R., Liguori, A., Kucenic, M., et al., 2004. Ultraviolet exposure is a reinforcing stimulus in frequent indoor tanners. Journal of the American Academy of Dermatology 51, 45–51.

Ferlay, J., Shin, H.R., Bray, F., et al., 2010. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 (accessed 24.05.13.).

Green, A.C., Williams, G.M., Logan, V., et al., 2011. Reduced melanoma after regular sunscreen use: randomized trial follow-up. Journal of Clinical Oncology 29, 257–263.

Hollands, G.J., Hankins, M., Marteau, T.M., 2010. Visual feedback of individuals' medical imaging results for changing health behaviour. Cochrane Database of Systematic Reviews, CD007434.

Jones, J.L., Leary, M.R., 1994. Effects of appearance-based admonitions against sun exposure on tanning intentions in young adults. Health Psychology 13, 86–90.

Kimlin, M.G., Guo, Y., 2012. Assessing the impacts of lifetime sun exposure on skin damage and skin aging using a non-invasive method. Science of the Total Environment 425, 35–41.

Leary, M.R., Jones, J.L., 1993. The social psychology of tanning and sunscreen use: self-presentational motives as a predictor of health risk. Journal of Applied Social Psychology 23, 1390–1406.

Lomas, A., Leonardi-Bee, J., Bath-Hextall, F., 2012. A systematic review of worldwide incidence of nonmelanoma skin cancer. British Journal of Dermatology 166, 1069–1080.

Lucas, R.M., McMichael, A.J., Armstrong, B.K., et al., 2008. Estimating the global disease burden due to ultraviolet radiation exposure. International Journal of Epidemiology 37, 654–667.

Noar, S.M., Benac, C.N., Harris, M.S., 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin 133, 673–693.

Paul, C., Tzelepis, F., Parfitt, N., et al., 2008. How to improve adolescents' sun protection behavior? Age and gender issues. American Journal of Health Behavior 32, 387–398.

Safe Work Australia, 2010. National Hazard Exposure Worker Surveillance–Exposure to Direct Sunlight and the Provision of Sun Exposure Controls in Australian Workplaces. Commonwealth of Australia, Canberra.

Schneider, S., Diehl, K., Bock, C., et al., 2013. Sunbed use, user characteristics, and motivations for tanning results from the German population-based sun-study 2012. JAMA Dermatology 149, 43–49.

Schüz, N., Eid, M., 2013. Beyond the usual suspects: target group- and behavior-specific factors add to a theory-based sun protection intervention for teenagers. Journal of Behavioral Medicine 36, 508–519.
Schüz, N., Schüz, B., Eid, M., 2013. When risk communication backfires: randomized controlled trial on self-affirmation and reactance to personalized risk feedback in high-risk individuals. Health Psychology 32, 561–570.
Shih, S.T.F., Carter, R., Sinclair, C., et al., 2009. Economic evaluation of skin cancer prevention in Australia. Preventive Medicine 49, 449–453.
Stanton, W.R., Janda, M., Baade, P.D., et al., 2004. Primary prevention of skin cancer: a review of sun protection in Australia and internationally. Health Promotion International 19, 369–378.
Stiller, C.A., 2007. International patterns of cancer incidence in adolescents. Cancer Treatment Reviews 33, 631–645.
Weinstock, M.A., 1999. Do sunscreens increase or decrease melanoma risk: an epidemiologic evaluation. Journal of Investigative Dermatology Symposium Proceedings 4, 97–100.

Relevant Websites

http://www.cancer.gov/publications/patient-education/wyntk-skin-cancer – National Cancer Institute, What You Need To Know About™ Melanoma and Other Skin Cancers.
http://www.knowyourownskin.com.au/checking-your-skin – Know your own skin.
http://www.med.upenn.edu/poolcool – University of Pennsylvania, Pool Cool.
http://www.sunsmart.com.au – SunSmart.
http://www.who.int/uv/en/ – World Health Organization, Ultraviolet radiation.

Supervision, Abusive

Christine MY Kermond, Michigan State University, East Lansing, MI, USA
John M Schaubroeck, The Eli Broad Graduate School of Management, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes research on abusive supervision in organizations. Abusive supervision refers to "subordinates' perceptions of the extent to which supervisors engage in the sustained display of hostile verbal and nonverbal behaviors". We review consequences of abusive supervision as well as the explanatory mechanisms of its relationship with these outcomes. We also discuss supervisor and subordinate factors contributing to abusive supervision perceptions. Finally, we review current methodological issues and describe promising future research directions.

Over the past two decades, the topic of abusive supervision has generated substantial interest in the popular press and among organizational scholars. Based on survey data reported by Schat et al. (2006), nearly 50% of employees in the United States have experienced abusive supervision in their working careers. One in seven employees reports that his or her current supervisor is abusive. Tepper et al. (2006) estimated that abusive supervision costs organizations $23.8 billion per annum. Since Tepper's (2000) seminal article that introduced the construct of abusive supervision, there have been over 100 journal articles published on the topic (Kermond et al., 2013a). Tepper defined abusive supervision as "subordinates' perceptions of the extent to which supervisors engage in the sustained display of hostile verbal and nonverbal behaviors, excluding physical contact" (p. 178). Examples of abusive supervisor behaviors include belittling and ridiculing subordinates, giving subordinates the silent treatment, making negative comments about subordinates to others, invading subordinates' privacy, and taking personal credit for the work of subordinates.

The purpose of this article is to provide a summary of the current state of knowledge about abusive supervision and to discuss emerging directions in theory and research. We organize our review in six sections: (1) Conceptual facets of abusive supervision, (2) Consequences of abusive supervision, (3) Mechanisms linking abusive supervision to outcomes, (4) Predictors of abusive supervision, (5) Methodological issues in abusive supervision research, and (6) Promising new directions in abusive supervision research.

Conceptual Facets of Abusive Supervision

Abusive supervision is usually examined in the context of a conventional supervisor–subordinate dyad, wherein the supervisory relationship is formal and position power is asymmetrical. It is conceptualized as a persistent and ongoing hostile behavioral pattern, meaning a one-time hostile outburst would not be considered abusive supervision. In addition, abusive supervision is an employee's subjective assessment of his or her supervisor's hostile behaviors toward him or her personally. Thus, the same behavior can potentially be perceived as abusive by one subordinate but not the other. Likewise the same behavior, for example, yelling, can be perceived to be abusive in one context (e.g., an office environment) but not abusive in another (e.g., a football coach directing players on the field).

Consequences of Abusive Supervision

Most abusive supervision research has focused its connections to individual victim's behavioral responses, attitudinal reactions, and personal well-being.

Work Behaviors and Attitudes

Aggression and Deviance

Victims of abusive supervision are often seen to engage in aggressive or deviant behaviors directed toward the supervisor, coworkers, and/or the organization (Mitchell and Ambrose, 2007). Meta-analyses have reported that on average, abusive supervision exhibits moderate relationships with deviant and counterproductive behaviors ($\rho = 0.39$, $k = 59$, $N = 19\ 700$; Kermond et al., 2013a; see also Schyns and Schilling, 2013). The relationship between perceptions of supervisory abuse and organization-directed retaliation was found to be stronger when the victim perceived that the organization is responsible for the abuse (Bowling and Michel, 2011). Tepper et al. (2008) observed that aggressive responses to abusive supervision were more likely to arise when the organizational context itself was perceived to be hostile, such as when coworkers condone deviance and engage in deviant acts themselves. Aggressive responses to abusive supervision also appear to be more prevalent when there is a hostile climate within a work group (Mawritz et al., 2012), or when there is a climate of uncertainty concerning the organization's leadership (Thau et al., 2009). Victims who intend to leave the organization were theorized to be less concerned with the repercussions of deviance, and thus are more likely to engage in retaliatory deviance against their supervisor (Tepper et al., 2009).

Performance and Prosocial Behaviors

Subordinates who report higher levels of abusive supervision tend to receive lower supervisory ratings of job performance

($\rho = -0.23$, $k = 14$, $N = 3185$; Kermond et al., 2013b; see also Schyns and Schilling, 2013). Abusive supervision victims are also less likely to engage in extra-role organizational citizenship behaviors in support of the organization (e.g., attend non-mandatory company functions, put in extra effort into work; $\rho = -0.23$, $k = 13$, $N = 2953$) or toward coworkers (e.g., adjust their work schedule to accommodate coworkers; $\rho = -0.20$, $k = 10$, $N = 2525$; Kermond et al., 2013b). Liu et al. (2012) reported that abusive supervision exhibited a stronger negative relationship with subordinates' creativity levels when subordinates attributed supervisory abuse to the supervisors' injurious motives rather than to performance-promotion motives.

Attitudes and Justice Perceptions

Abusive supervision is negatively related to subordinates' work and organization-related attitudes. For example, employees experiencing higher levels of supervisory abuse report lower job satisfaction ($\rho = -0.38$, $k = 24$, $N = 8953$), lower organizational commitment ($\rho = -0.21$, $k = 15$, $N = 4354$), and higher turnover intention ($\rho = 0.32$, $k = 21$, $N = 7982$; Kermond et al., 2013b; see also Schyns and Schilling, 2013). Abusive supervision is also negatively associated with perceptions of the supervisor's interpersonal justice practices (i.e., fairness of interpersonal treatment; $\rho = -0.61$, $k = 10$, $N = 2290$) and procedural fairness (i.e., fairness of procedures used to make allocation decisions; $\rho = -0.39$, $k = 6$, $N = 1658$; Kermond et al., 2013b; see also Schyns and Schilling, 2013).

Victim Personal and Family Well-Being

Psychological and Physical Health

Much research has examined the influence of abusive supervision on employee psychological and physical well-being. Evidence suggests that experiencing abusive supervision is psychologically draining for subordinates and impairs their capacity to self-regulate (Lian et al., 2012a). Abused employees often suffer from emotional exhaustion (Aryee et al., 2008). Victims of supervisory abuse also report higher levels of insomnia (Rafferty et al., 2010) and a greater tendency to self-medicate with alcohol (Bamberger and Bacharach, 2006). Meta-analyses report moderate correlations between abusive supervision and psychological distress ($\rho = 0.36$, $k = 40$, $N = 13\,330$; Kermond et al., 2013a; see also Schyns and Schilling, 2013), whereas relationships between abusive supervision and physical health symptoms are lower in magnitude ($\rho = 0.16$, $k = 10$, $N = 4662$; Kermond et al., 2013a). The strength of the relationship between abusive supervision and psychological distress appears to be more pronounced among subordinates with fewer job mobility options (Tepper, 2000). Abused employees also tend to be more depressed and anxious (Tepper, 2000).

Family

Individuals who report higher levels of abusive supervision tend to experience higher levels of conflict between their work and family demands (Tepper, 2000). They are also more inclined to take negative work emotions out on family members or domestic partners by exhibiting family-directed aggression (Hoobler and Brass, 2006). Similarly, Carlson et al. (2011) reported crossover effects of abusive supervision from work to family; higher levels of abuse were associated with greater relationship tension, lower family satisfaction, and more work-to-family conflict. A separate study found that subordinates' psychological distress explained the relationship between abusive supervision and subordinates' aggression toward their spouse or partner (Restubog et al., 2011).

Mechanisms Linking Abusive Supervision to Outcomes

A good deal of scholarly effort has been devoted to identifying explanatory mechanisms that link abusive supervision to adverse employee outcomes. Most theories of abusive supervision can be broadly distinguished between those that focus on relational mechanisms, such as exchange theories and other theories that concern perceptions of fairness, and theories that emphasize self-regulatory mechanisms.

Relational Theories

Relational theories as frameworks that describe how employees respond to abusive supervision based on inferences they draw from the abuse about their relationship with the supervisor and/or the organization. One such framework is fairness theory, in which judgments of fairness are driven by individuals' assignments of responsibility (Folger and Cropanzano, 2001). Uncertainty management theory highlights how being mistreated is especially important for employees who are uncertain about themselves or their relationship with the organization (Lind and Van den Bos, 2002). Social exchange theory describes the costs and benefits to employees of maintaining different forms of reciprocity-based relationships with their supervisor and the organization (Blau, 1964). Each of these theories has been invoked in studies highlighting fairness of some kind, such as supervisor interpersonal or procedural fairness, or in studying the influence of abusive supervision on the victim's level of trust in the supervisor. For example, an exchange perspective may highlight how abusive supervision elicits negative reciprocity, such that subordinates respond to negative supervisor treatment with negative behaviors (e.g., Peng et al., forthcoming). An uncertainty management or a fairness theory perspective may emphasize how being abused activates unfairness cognitions that confirm or disconfirm an existing schema about the organization, and this in turn influences the employee's motivation (e.g., Mayer et al., 2012; Thau et al., 2009). Below we highlight studies of abusive supervision that examined mediators that were consistent with one or more of these relational theories of abusive supervision.

Justice Perceptions

Tepper (2000) reported that overall justice perceptions partially mediated the relationship between abusive supervision and emotional exhaustion and anxiety. Aryee et al. (2007) demonstrated that abusive supervision elicits perceptions of interactional justice (i.e., the combination of interpersonal and informational fairness), which in turn influences subordinate organizational commitment and organizational citizenship behaviors. In a study reported by Burton and Hoobler (2011),

interactional justice mediated the relationship between abusive supervision and subordinate aggression. Zellars et al. (2002) reported that subordinates' procedural justice perceptions mediated the relationship between abusive supervision and subordinate organizational citizenship behaviors.

Leader–Member Exchange

Not surprisingly, abusive supervision appears to be associated with subordinate–supervisor relationships of poorer quality, in the organizational behavior literature leader-follower relationship quality is commonly operationalized by subordinate perceptions of the leader–member exchange relationship. Leader–member exchange quality refers to one's level of positive emotional connection to the supervisor, including affect-based trust and loyalty directed toward him or her (Scandura and Graen, 1984). When subordinates are in a poor exchange relationship, they are seen to withhold valuable exchange resources from the supervisor. Xu et al. (2012) reported that leader–member exchange quality fully mediated the relationship between abusive supervision and subordinate performance and organizational citizenship behaviors. Martinko et al. (2011) suggested that the relationship between abusive supervision and leader–member exchange may be less clear-cut, arguing that leader–member exchange might even be antecedent to abusive supervision rather than the reverse.

Attributing Blame to the Organization

Abused subordinates can indirectly reciprocate supervisor abuse by engaging in organization-directed deviance. Abusive supervisors may not only be viewed as representing the organization, many subordinates perceive them as embodying the organization (Eisenberger et al., 2010). Such employees may hold the organization responsible for the abuse and direct retributive actions against the organization. Shoss et al. (2013) examined the mediating role of perceived organizational support, defined as subordinate's perception that the organization values them and cares about their well-being. Their findings suggested that employees attributed their mistreatment partly to a lack of organizational support, and thus victims tended to engage in organization-directed deviance and withhold supportive extrarole contributions to the organization. Similarly, Tepper et al. (2008) reported results suggesting that abused subordinates also blamed the organization for the abuse, resulting in lower organizational commitment.

Emotions and Self-Regulation

Victims of abusive supervision may experience distinctly high levels of anger, fear, and/or anxiety (e.g., Rafferty et al., 2010; Tepper et al., 2009; Wu and Hu, 2009). Yet, because of the asymmetrical balance of power with the supervisor, abused subordinates may be deterred from directly confronting the supervisor or engaging in retributive actions against him or her. This can lead to pent-up aggression, which is often redirected toward more vulnerable targets. A displaced-aggression perspective (Dollard et al., 1939) is suggested by certain findings. For example, as noted above (Section Family), studies find that abusive supervision victims tend to experience a variety of troubling cross-over effects in their home life. These effects have been attributed in part to displaced aggression (Carlson et al., 2011; Hoobler and Brass, 2006).

The self-regulation impairment perspective argues that experiencing abusive supervision diminishes the victims' capacity for self-regulation (e.g., behaving normatively and rationally), because they are preoccupied with processing and interpreting the abusive events (Thau et al., 2007). Studies have reported that when abused subordinates had limited in self-regulatory capacities, as reflected in poorer impulse control (Lian et al., forthcoming-a; Thau and Mitchell, 2010) or uncertainty about their job competence (Mayer et al., 2012), they were more likely to respond to abuse by acting out in a counterproductive manner. Conversely, us of effective coping strategies appear to weaken the relationship between perceptions of abuse and subordinate deviance (Nandkeolyar et al., 2014).

Predictors of Abusive Supervision

Researchers have examined several antecedents of abusive supervision. The studied antecedents are primarily characteristics of supervisors and subordinates.

Supervisor Characteristics

Supervisor Traits

A small number of studies have examined relationships between supervisor traits such as authoritarian leadership style and supervisor Machiavellianism on abusive supervision. Subordinates who viewed their leader as adopting an authoritarian leadership style tended to also perceive higher levels of abusive supervision (Aryee et al., 2007). Authoritarian leadership style is defined as a "leader's behavior that asserts absolute authority and control over subordinates and demands unquestionable obedience from subordinates" (Cheng et al., 2004: p. 91). Supervisor Machiavellianism, which is the extent to which an individual manipulates and exploits others to maximize his or her self-interests (Christie et al., 1970), was positively related to supervisor authoritarian leadership style, which in turn predicted abusive supervision (Kiazad et al., 2010). Furthermore, subordinates with lower levels of organization-based self-esteem tend to perceive higher levels of abusive supervision when their supervisors exhibited an authoritarian leadership style (Kiazad et al., 2010).

Supervisors' Experience of Abusive Supervision

Just as abused subordinates may often displace their aggression onto individuals who are more vulnerable than the perpetrator of the abuse, supervisors may also displace the abuse they experience from their manager onto their own subordinates. Supervisors who perceive higher levels of procedural injustice (Tepper et al., 2006), interactional injustice (Hoobler and Hu, 2013; Rafferty et al., 2010), and psychological contract violations (Hoobler and Brass, 2006; Rafferty et al., 2010) are seen as more abusive by their subordinates. Abusive supervision can also 'trickle down' (or 'cascade' down) a chain of command. Abusive behaviors have been found to exhibit indirect effects

on abusive behaviors across two hierarchical levels, such that higher level manager abuse predicted lower level manager abuse (Liu et al., 2012; Mawritz et al., 2012).

In further support of the displaced-aggression hypothesis, Tepper et al. (2006) reported that supervisor depression mediated the relationship between procedural injustice perceptions and abusive behaviors. In addition, supervisors' conflicts with other supervisors at the same level were positively related to their subordinates' perceptions of abusive supervision (Harris et al., 2011). In addition, when supervisors' workplace stress levels were high, their direct reports reported significantly more abusive supervision (Burton et al., 2012). Finally, Kiewitz et al. (2012) drew upon social learning theory to explain how supervisor childhood experiences affect their abusive tendencies. They found that supervisors' experiences of family undermining (such as parent criticism, insults, unpleasant or angry treatment) during childhood were linked to their subordinates' reports of abusive supervision. This connection was weaker among supervisors with higher levels of impulse control.

Subordinate Characteristics

Subordinate Precipitating Factors

According to the victim precipitation model (Elias, 1986), targets of abuse often 'precipitate' abuse, meaning they indirectly or directly provoke the abuse. For example, abusive supervision victims may tend to be those individuals who are seen by the supervisor as the most vulnerable persons on whom they can release their own pent-up aggression. Victims may also provoke abuse because of their own negative dispositions, such as by behaving in a hostile manner toward others. A subordinate trait that is examined most frequently in this context is negative affectivity, which is characterized by frequent negative mood states such as feeling distressed, jittery, angry, depressed, or upset (Watson et al., 1988). Negative affectivity of subordinates has been linked to their reports of abusive supervision (e.g., Hoobler and Hu, 2013; Tepper et al., 2006), although it is uncertain whether this reflects the tendency of high-negative-affectivity individuals to provoke abuse or rather a general tendency of high-negative-affectivity individuals to interpret ambiguous stimuli in negative ways. Low job-related competence or low job performance quality may also make a subordinate more likely to become a victim of abusive supervision (Tepper et al., 2006).

Subordinate Attributions

Martinko et al. (2011) reported that employees with a hostile attribution style (i.e., attribute negative events to external and stable causes) perceived more abusive supervision than others, and that this relationship was mediated by perceptions of leader–member exchange quality. Subordinates' levels of psychological entitlement may also affect their attributions concerning their supervisor's abusive behaviors. Because psychologically entitled subordinates feel that they deserve high levels of praise irrespective of their performance or ability, compared to less psychologically entitled subordinates, they are more likely to perceive supervisor behaviors as being abusive (Harvey et al., 2014).

Methodological Issues in Abusive Supervision Research

Similar Constructs

In recent years, there has been a proliferation of constructs that conceptually overlap with abusive supervision. These constructs include petty tyranny (Ashforth, 1994), supervisor bullying (Fox and Stallworth, 2005), aversive leadership (e.g., Bligh et al., 2006; Pearce and Sims, 2002), supervisor social undermining (e.g., Duffy et al., 2002), supervisor verbal aggression (e.g., Chang and Lyons, 2012; Grandey et al., 2007), and supervisor incivility (e.g., Cortina et al., 2001; Leiter et al., 2011). See **Table 1** for the definitions of these constructs.

These constructs all refer to hostile and/or aggressive verbal or nonverbal behaviors directed by a supervisor toward a specific subordinate (Hershcovis, 2011). Scholars have raised concerns about construct proliferation in this domain, while also cautioning against ignoring key conceptual differences between the constructs, such as the intensity, intentionality, and perseverance of the abuse (e.g., Hershcovis, 2011; Tepper

Table 1 Definitions of constructs similar to abusive supervision

Supervisor abuse variable	*Definition*
Abusive supervision	"Abusive supervision refers to subordinates' perceptions of the extent to which supervisors engage in the sustained display of hostile verbal and nonverbal behaviors, excluding physical contact" (Tepper, 2000: p. 178).
Petty tyranny	A petty tyrant is defined as "an individual who lords his or her power over others … to overcontrol other and to treat them in an arbitrary, uncaring and punitive manner" (Ashforth, 1994: pp. 755–756).
Supervisor bullying	"A situation where one or several individuals persistently over a period of time perceive themselves to be on the receiving end of negative actions from one or several persons, in a situation where the target of bullying has difficulty in defending him or herself against these actions" (Einarsen and Skogstad, 1996: p. 191).
Aversive leadership	"Leadership that primarily relies on coercive power and is rooted in punishment" (Pearce and Sims, 2002: p. 173).
Supervisor social undermining	"Behavior intended to hinder, over time, the ability to establish and maintain positive interpersonal relationships, work-related success, and favorable reputation" (Duffy et al., 2002: p. 332).
Supervisor verbal aggression	"Verbal forms of aggression involve efforts to inflict harm on others through words rather than deeds" (Baron and Neuman, 1996: p. 163).
Supervisor incivility	"Workplace incivility is low-intensity deviant behavior with ambiguous intent to harm the target, in violation of workplace norms for mutual respect" (Andersson and Pearson, 1999: p. 457).

and Henle, 2011). Conversely, the metaconstruct of destructive leadership refers to generally undesirable characteristics or behaviors of leaders that are not necessarily targeted toward a specific subordinate (Einarsen et al., 2007; Padilla et al., 2007). Constructs included under the rubric of destructive leadership are distinct from those just mentioned because they refer to subordinates' perceptions of leader behaviors directed toward others in general, rather than toward a specific subordinate (e.g., despotic leadership; De Hoogh and Den Hartog, 2008), or to behaviors or traits that do not constitute abuse, per se, but which are seen as undesirable for other reasons (e.g., authoritarian leadership style, narcissistic supervision, negative leadership, psychopathic leadership; Schyns and Schilling, 2013).

Self-Reported Data and Cross-Sectional Research Designs

Abusive supervision researchers have mainly relied on variables measured at one time point or measured outcome variables at a separate point in time (e.g., staggering data collection over two to three periods). Many studies have relied exclusively on subordinates' self-reports about both their leaders' behaviors and the outcome variables. Because abusive supervision is conceptualized as a dyadic process, the potential benefits of aggregating perceptions across multiple subordinates of the same leaders in reducing common method variance are not realized. Thus, cross-sectional studies and studies that take little advantage of time separation or which do not focus on moderated relationships are prone to spurious inflation of studied relationships owing to common method variance. Among the few exceptions, Thau and Mitchell (2010: sample 3) employed a repeated measures design that measured abusive supervision and deviant behaviors over 10 time points. The authors found that the within-person relationship between abusive supervision and deviant behaviors was mediated by self-regulation impairment. Another example was provided by Lian et al. (forthcoming-b). They obtained repeated measures of abusive supervision and subordinate deviant behavior, measuring both variables at two time points. In one study in which the measures were separated by a 20-month lag, their cross-lagged analyses suggested that subordinate deviant behavior leads to abusive supervision rather than the reverse. A second study with a 6-month lag between repeated measures suggested a reciprocal relationship between subordinate deviant behavior and abusive supervision. They argued that the differences between the two studies in time lag explained the diverging results.

Inadequate Model Specification

Although there are a few exceptions (e.g., Aryee et al., 2007; Lian et al., 2012b; Peng et al., forthcoming), abusive supervision studies have tended to focus on one mediating variable at a time. Examining a single mediating pathway may not adequately reflect the various theoretical mechanisms responsible for the relationship between abusive supervision and employee outcomes. This is of particular concern in studies in which abusive supervision and the putative mediator are not separated in time and common method variance is a concern, as this can exaggerate the estimated sizes of indirect effects.

Promising New Directions in Abusive Supervision Research

In addition to promising trends in the direction of longitudinal designs and investigating multiple pathways of mediation as noted above, there are other positive trends in the way abusive supervision is studied and in the depth and scope of the research questions addressed. Researchers have moved toward examining the social context of abusive supervision, including investigations of the role of coworkers' abusive supervision levels in how individuals respond to their own experiences of abuse, testing multilevel processes related to abusive supervision, examining abusive supervision in non-Western cultural contexts and investigating cultural differences as moderator variables, and giving long-overdue empirical attention to means of preventing abusive supervision.

Third-Party Observers of Abusive Supervision

The deontic justice perspective suggests that observers of injustices inflicted on others experience outrage and often engage in retributive actions to punish the source of the injustice and aid the victims (e.g., Lotz et al., 2011; Turillo et al., 2002). A third-party observer may learn about the abuse of another individual through word of mouth from the abused employee or from other observers. If the observer deems the target of abuse as undeserving of the ill treatment, he or she may attempt to restore justice by lashing out to the supervisor, or report the abuse to a higher authority in the organization (Skarlicki and Kulik, 2005). Umphress et al. (2013) found that third-party observers displayed anger toward abusive supervisors and they reported an intention to take corrective or punitive response, especially when they perceived that the abuse was intended to harm the target employee.

Multilevel Research

Tepper (2007) called for more multilevel research on abusive supervision, noting that "abusive supervision is a multilevel phenomenon and the field needs to move beyond individual-level research" (p. 281). One approach to multilevel research on abusive supervision examines the interaction between subordinates' own reports of experiencing abusive supervision and the average group experience (i.e., average levels of abusive supervision as reported by work unit peers (i.e., 'work unit,' 'coworker,' or 'peer' abusive supervision)). Hannah et al. (2013) observed an interaction between subordinates' own reports of abuse and the average level of work unit member abuse in predicting subordinates' ethical behaviors and intentions among soldiers deployed in combat. Subordinate's own experience of abusive supervision was more strongly related to outcomes when coworker unit abusive supervision was lower. This study drew from a self-regulation perspective of abusive supervision, emphasizing the strength of moral standards and moral courage as mediators of these interactive effects. Drawing from a social exchange perspective, Peng et al. (forthcoming) examined the interaction of peer and own abusive supervision on employees' leader–member exchange perceptions and

affect-based trust in peers. These exchange variables partially mediated the interactive effects of own and peer abusive supervision on task performance and helping behaviors directed toward peers. Only employees who were low in both own and peer abusive supervision exhibited high levels of helping behavior and task performance, a pattern consistent with Hannah et al. (2013).

In sum, a higher level of work unit abuse is associated with weaker relationships between individual abusive supervision and subordinate outcomes, reflecting a pattern of substitution moderation. Thus, the social context of abusive supervision affects how employees respond to their own abusive treatment by the supervisor. Employees are especially impacted by such abusive behavior when they are singled out for abuse. It is also noteworthy that the findings of the studies reviewed above concerning own and coworkers' experiences of abusive supervision are not consistent with the deontic justice literature, in which coworkers' experiences of injustice are met with aversive psychological reactions and aggressive responses, much as the individuals react to experiencing the injustice themselves. Peng et al. (forthcoming) noted that conflicting results across the two approaches may be due to the fact that research on deontic reactions to injustice has been conducted primarily in laboratory settings. Unlike a workplace, laboratory participants would be less likely to fear repercussions for confronting or reporting an abusive supervisor or expressing moral outrage for the abuse.

Cross-Cultural Research

Abusive supervision research using samples from outside of the United States has been growing steadily, spanning many different regions of the world (e.g., Australia, Belgium, Canada, China, India, Philippines, Singapore, Slovenia, South Korea, New Zealand, Nigeria, Norway, and Taiwan). Hu et al. (2011) compared participant responses from Taiwan and the United States and found that employees generally experienced abusive supervision as a single latent construct. However, Taiwanese employees reported that supervisor belittling was more essential to their conceptualization of 'abusive supervision' than workers from United States. Belittling behaviors may reflect an underlying construct of 'saving face,' which Confucianism-based societies (e.g., Taiwan) value as an important part of interpersonal interactions (House et al., 2004).

Power distance orientation has been the most commonly studied cultural construct in the abusive supervision literature. It is defined as the extent to which individuals accept an unequal distribution of power in organizations and institutions. As one might expect, subordinates who score high on power distance orientation exhibit weaker relationships between abusive supervision and subordinates' psychological health and job satisfaction (Lin et al., 2013). Subordinates high in power distance orientation viewed abusive supervision as more fair and were less likely to react with interpersonal deviance than subordinates low in power distance orientation (Lian et al., 2012a). Another cultural predisposition variable of interest to abusive supervision scholars is Chinese traditionality (i.e., the degree to which an individual is fatalistic and values respect for authorities and elders). Liu et al. (2010) observed that the relationship between abusive supervision and supervisor-directed deviance was stronger for individuals who did not report high traditionality.

Preventing and Managing Abusive Supervision

Preventing or managing abusive supervision ought to be on the forefront of research given its deleterious impact on employees and organizations. Yet there have been few published empirical assessments of abusive supervision intervention programs. A study conducted by Leiter et al. (2011) is an exception. They evaluated the efficacy of a general civility intervention in a sample of health care workers. The program consisted of trained facilitators meeting with employees on a weekly basis for 6 months. Training content included learning to identify incivility behaviors (e.g., rude and condescending behavior), effectively respond to incivility, and practicing new ways to interacting positively. Compared to a contrast group that did not formally participate in the sessions, treatment group participants reported lower experienced incivility from supervisor and coworkers, lower report of self-enacted incivility behaviors, and higher civility within the work unit overall (e.g., spirit of teamwork in work group, disputes are resolved fairly).

A number of legislative developments may encourage organizations to devote greater attention to preventing abusive supervision. In many countries, employees may seek redress in civil courts for abusive supervision as well as for coworker abuse. As of 6 February 2014, 26 state legislatures in the United States had introduced the Healthy Workplace Bill, which includes specific penalties for abuse of employees. Similar legal protections have been become law in other countries, including Australia, Belgium, Canada, Denmark, Finland, France, Great Britain, Iceland, Ireland, and Norway.

Conclusion

Given the detrimental impact of abusive supervision on employee outcomes, researchers have examined a wide range of antecedents and mechanisms through which abusive supervision arises and affects victims' behaviors, attitudes, and well-being. The literature has become very vibrant in more recent years, broadening in the depth and scope of research questions examined, improving in methodological sophistication, and engaging a fairly large network of scholars worldwide. While the future of abusive supervision research looks promising, we hope that abusive supervision scholars and those investigating similar constructs (e.g., supervisor incivility) can begin to work together in addressing their common questions. We also encourage researchers to become more active in producing evidence-based approaches for preventing supervisory abuse and for enabling workers to more effectively confront abuse without suffering adverse personal consequences.

See also: Job Satisfaction; Leader–Member Exchange Theory; Managerial and Organizational Cognition; Negotiation and Conflict, The Psychology of; Organizational Climate in the Work Setting; Organizational Commitment; Organizational Justice; Power, Politics, and Influence in Organizations; Workplace Deviance; Workplace Stress.

Bibliography

Andersson, L.M., Pearson, C.M., 1999. Tit for tat? the spiraling effect of incivility in the workplace. Academy of Management Review 24, 452–471.

Aryee, S., Chen, Z.X., Sun, L.Y., Debrah, Y.A., 2007. Antecedents and outcomes of abusive supervision: test of a trickle-down model. Journal of Applied Psychology 92, 191–201.

Aryee, S., Sun, L.Y., Chen, Z.X.G., Debrah, Y.A., 2008. Abusive supervision and contextual performance: the mediating role of emotional exhaustion and the moderating role of work unit structure. Management and Organization Review 4, 393–411.

Ashforth, B.E., 1994. Petty tyranny in organizations. Human Relations 47, 755–778.

Bamberger, P.A., Bacharach, S.B., 2006. Abusive supervision and subordinate problem drinking: taking resistance, stress and subordinate personality into account. Human Relations 59, 723–752.

Baron, R.A., Neuman, J.H., 1996. Workplace violence and workplace aggression: evidence on their relative frequency and potential causes. Aggressive Behavior 22, 161–173.

Blau, P.M., 1964. Exchange and Power in Social Life. John Willey & Sons, New Jersey, New York.

Bligh, M.C., Pearce, C.L., Kohles, J.C., 2006. The importance of self–and shared leadership in team based knowledge work: a meso-level model of leadership dynamics. Journal of Managerial Psychology 21, 296–318.

Bowling, N.A., Michel, J.S., 2011. Why do you treat me badly? the role of attributions regarding the cause of abuse in subordinates' responses to abusive supervision. Work & Stress 25, 309–320.

Burton, J.P., Hoobler, J.M., 2011. Aggressive reactions to abusive supervision: the role of interactional justice and narcissism. Scandinavian Journal of Psychology 52, 389–398.

Burton, J.P., Hoobler, J.M., Scheuer, M.L., 2012. Supervisor workplace stress and abusive supervision: the buffering effect of exercise. Journal of Business and Psychology 27, 271–279.

Carlson, D.S., Ferguson, M., Perrewe, P.L., Whitten, D., 2011. The fallout from abusive supervision: an examination of subordinates and their partners. Personnel Psychology 64, 937–961.

Chang, C.H., Lyons, B.J., 2012. Not all aggressions are created equal: a multi-foci approach to workplace aggression. Journal of Occupational Health Psychology 17, 79–92.

Cheng, B.S., Chou, L.F., Wu, T.Y., Huang, M.P., Farh, J.L., 2004. Paternalistic leadership and subordinate responses: establishing a leadership model in Chinese organizations. Asian Journal of Social Psychology 71, 89–117.

Christie, R., Geis, F.L., Berger, D., 1970. Studies in Machiavellianism. Academic Press, New York.

Cortina, L.M., Magley, V.J., Williams, J.H., Langhout, R.D., 2001. Incivility in the workplace: incidence and impact. Journal of Occupational Health Psychology 6, 64–80.

De Hoogh, A.H., Den Hartog, D.N., 2008. Ethical and despotic leadership, relationships with leader's social responsibility, top management team effectiveness and subordinates' optimism: a multi-method study. The Leadership Quarterly 19, 297–311.

Dollard, J., Miller, N.E., Doob, L.W., Mowrer, O.H., Sears, R.R., 1939. Frustration and Aggression. Yale University Press, New Haven.

Duffy, M.K., Ganster, D.C., Pagon, M., 2002. Social undermining in the workplace. The Academy of Management Journal 45, 331–351.

Einarsen, S., Aasland, M.S., Skogstad, A., 2007. Destructive leadership behaviour: a definition and conceptual model. The Leadership Quarterly 18, 207–216.

Eisenberger, R., Karagonlar, G., Stinglhamber, F., Neves, P., Becker, T.E., Gonzalez-Morales, M.G., Steiger-Mueller, M., 2010. Leader-member exchange and affective organizational commitment: the contribution of supervisor's organizational embodiment. Journal of Applied Psychology 95, 1085–1103.

Einarsen, S., Skogstad, A., 1996. Bullying at work: epidemiological findings in public and private organizations. European Journal of Work and Organizational Psychology 5, 185–201.

Elias, R., 1986. The Politics of Victimization: Victims, Victimology, and Human Rights. Oxford University Press, New York.

Folger, R., Cropanzano, R., 2001. Fairness theory: justice as accountability. In: Greenberg, J., Cropanzano, R. (Eds.), Advances in Organizational Justice. Stanford Business Books, Stanford, pp. 1–55.

Fox, S., Stallworth, L.E., 2005. Racial/ethnic bullying: exploring links between bullying and racism in the U.S. workplace. Journal of Vocational Behavior 66, 338–456.

Grandey, A., Kern, J., Frone, M., 2007. Verbal abuse from outsiders versus insiders: comparing frequency, impact on emotional exhaustion, and the role of emotional labor. Journal of Occupational Health Psychology 12, 63–79.

Hannah, S.T., Schaubroeck, J.M., Peng, A.C., Lord, R.G., Trevino, L., Kozlowski, S., Avolio, B., Dimotakis, N., Doty, J., 2013. Joint influences of individual and work unit abusive supervision on ethical intentions and behaviors: a moderated mediation model. Journal of Applied Psychology 98, 579–592.

Harris, K.J., Harvey, P., Kacmar, K.M., 2011. Abusive supervisory reactions to coworker relationship conflict. The Leadership Quarterly 22, 1010–1023.

Harvey, P., Harris, K.J., Gillis, W.E., Martinko, M.J., 2014. Abusive supervision and the entitled employee. The Leadership Quarterly 25, 204–217.

Hershcovis, M.S., 2011. Incivility, social undermining, bullying… oh my!: A call to reconcile constructs within workplace aggression research. Journal of Organizational Behavior 32, 499–519.

Hoobler, J.M., Brass, D.J., 2006. Abusive supervision and family undermining as displaced aggression. Journal of Applied Psychology 91, 1125–1133.

Hoobler, J.M., Hu, J., 2013. A model of injustice, abusive supervision, and negative affect. The Leadership Quarterly 24, 256–269.

House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W., Gupta, V., 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage, Thousand Oaks.

Hu, C., Wu, T.Y., Wang, Y.H., 2011. Measurement equivalence/invariance of the abusive supervision measure across workers from Taiwan and the United States. The Journal of Psychology 145, 111–131.

Kermond, C.M.Y., Malonson, B., Schaubroeck, J.M., 2013a. Meta-analytic investigation of abusive supervision in the workplace. Paper presented at the Annual Congress of European Association of Work and Organizational Psychology. Munster, Germany.

Kermond, C.M.Y., Schaubroeck, J.M., Malonson, B., 2013b. Supervisory abuse and follower outcomes: a meta-analysis. Unpublished manuscript.

Kiazad, K., Restubog, S.L.D., Zagenczyk, T.J., Kiewitz, C., Tang, R.L., 2010. In pursuit of power: the role of authoritarian leadership in the relationship between supervisors' Machiavellianism and subordinates' perceptions of abusive supervisory behavior. Journal of Research in Personality 44, 512–519.

Kiewitz, C., Restubog, S.L.D., Zagenczyk, T.J., Scott, K.D., Garcia, P.R.J.M., Tang, R.L., 2012. Sins of the parents: self-control as a buffer between supervisors' previous experience of family undermining and subordinates' perceptions of abusive supervision. The Leadership Quarterly 23, 869–882.

Leiter, M.P., Laschinger, H.K.S., Day, A., Oore, D.G., 2011. The impact of civility interventions on employee social behavior, distress, and attitudes. Journal of Applied Psychology 96, 1258–1274.

Lian, H., Ferris, D.L., Brown, D.J., 2012a. Does power distance exacerbate or mitigate the effects of abusive supervision? it depends on the outcome. Journal of Applied Psychology 97, 107–123.

Lian, H., Ferris, D.L., Brown, D.J., 2012b. Does taking the good with the bad make things worse? How abusive supervision and leader–member exchange interact to impact need satisfaction and organizational deviance. Organizational Behavior and Human Decision Processes 117, 41–52.

Lian, H., Brown, D., Ferris, D.L., Liang, L., Keeping, L., Morrison, R. Abusive supervision and retaliation: a self-control framework. Academy of Management Journal, forthcoming-a.

Lian, H., Ferris, D.L., Morrison, R., Brown, D.J. Blame it on the supervisor or the subordinate? Reciprocal relations between abusive supervision and organizational deviance. Journal of Applied Psychology, forthcoming-b.

Lin, W., Wang, L., Chen, S., 2013. Abusive supervision and employee well-being: the moderating effect of power distance orientation. Applied Psychology: An International Review 62, 308–329.

Lind, E.A., Van den Bos, K., 2002. When fairness works: toward a general theory of uncertainty management. Research in Organizational Behavior 24, 181–223.

Liu, J., Kwan, H.K., Wu, L.Z., Wu, W.K., 2010. Abusive supervision and subordinate supervisor-directed deviance: the moderating role of traditional values and the mediating role of revenge cognitions. Journal of Occupational and Organizational Psychology 83, 835–856.

Liu, D., Liao, H., Loi, R., 2012. The dark side of leadership: a three-level investigation of the cascading effect of abusive supervision on employee creativity. Academy of Management Journal 55, 1187–1212.

Lotz, S., Okimoto, T.G., Schlösser, T., Fetchenhauer, D., 2011. Punitive versus compensatory reactions to injustice: emotional antecedents to third-party interventions. Journal of Experimental Social Psychology 47, 477–480.

Martinko, M.J., Harvey, P., Sikora, D., Douglas, S.C., 2011. Perceptions of abusive supervision: the role of subordinates' attribution styles. The Leadership Quarterly 22, 751–764.

Mawritz, M.B., Mayer, D.M., Hoobler, J.M., Wayne, S.J., Marinova, S.V., 2012. A trickle-down model of abusive supervision. Personnel Psychology 65, 325–357.

Mayer, D.M., Thau, S., Workman, K.M., Van Dijke, M., De Cremer, D., 2012. Leader mistreatment, employee hostility, and deviant behaviors: integrating self-uncertainty and thwarted needs perspectives on deviance. Organizational Behavior and Human Decision Processes 117, 24–40.

Mitchell, M.S., Ambrose, M.L., 2007. Abusive supervision and workplace deviance and the moderating effects of negative reciprocity beliefs. Journal of Applied Psychology 92, 1159–1168.

Nandkeolyar, A.K., Shaffer, J.A., Li, A., Ekkirala, S., Bagger, J., 2014. Surviving an abusive supervisor: the joint roles of conscientiousness and coping strategies. Journal of Applied Psychology 99, 138–150.

Padilla, A., Hogan, R., Kaiser, R.B., 2007. The toxic triangle: destructive leaders, susceptible followers, and conducive environments. The Leadership Quarterly 18, 176–194.

Pearce, C.L., Sims, H.P., 2002. Vertical versus shared leadership as predictors of the effectiveness of change management teams: an examination of aversive, directive, transactional, transformational, and empowering leadership behaviors. Group Dynamics: Theory, Research and Practice 6, 172–197.

Peng, A.C., Schaubroeck, J.M., Li, Y. Social exchange implications of own and coworkers' experiences of supervisory abuse. Academy of Management Journal, forthcoming.

Rafferty, A.E., Restubog, S.L.D., Jimmieson, N.L., 2010. Losing sleep: examining the cascading effects of supervisors' experience of injustice on subordinates' psychological health. Work & Stress 24, 36–55.

Restubog, S.L.D., Scott, K.L., Zagenczyk, T.J., 2011. When distress hits home: the role of contextual factors and psychological distress in predicting employees' responses to abusive supervision. Journal of Applied Psychology 96, 713–729.

Scandura, T.A., Graen, G.B., 1984. Moderating effects of initial leader member exchange status on the effects of a leadership intervention. Journal of Applied Psychology 69, 428–436.

Schat, A.C.H., Frone, M.R., Kelloway, E.K., 2006. Prevalence of workplace aggression in the U.S. workforce: findings from a national study. In: Kelloway, E.K., Barling, J., Hurrell, J.J. (Eds.), Handbook of Workplace Violence. Sage, Thousand Oaks, pp. 47–89.

Schyns, B., Schilling, J., 2013. How bad are the effects of bad leaders? A meta-analysis of destructive leadership and its outcomes. The Leadership Quarterly 24, 138–158.

Shoss, M.K., Eisenberger, R., Restubog, S.L.D., Zagenczyk, T.J., 2013. Blaming the organization for abusive supervision: the roles of perceived organizational support and supervisor's organizational embodiment. Journal of Applied Psychology 98, 158–168.

Skarlicki, D., Kulik, C., 2005. Third party reactions to employee mistreatment: a justice perspective. In: Staw, B., Kramer, R. (Eds.), Research in Organizational Behavior. JAI Press, Greenwich, pp. 183–230.

Tepper, B.J., 2000. Consequences of abusive supervision. Academy of Management Journal 43, 178–190.

Tepper, B.J., 2007. Abusive supervision in work organizations: review, synthesis, and research agenda. Journal of Management 33, 261–289.

Tepper, B.J., Carr, J.C., Breaux, D.M., Geider, S., Hu, C.Y., Hua, W., 2009. Abusive supervision, intentions to quit, and employees' workplace deviance: a power/dependence analysis. Organizational Behavior and Human Decision Processes 109, 156–167.

Tepper, B.J., Duffy, M.K., Henle, C.A., Lambert, L.S., 2006. Procedural injustice, victim precipitation, and abusive supervision. Personnel Psychology 59, 101–123.

Tepper, B.J., Henle, C.A., 2011. A case for recognizing distinctions among constructs that capture interpersonal mistreatment in work organizations. Journal of Organizational Behavior 32, 487–498.

Tepper, B.J., Lambert, L.S., Henle, C.A., Giacalone, R.A., Duffy, M.K., 2008. Abusive supervision and subordinates' organization deviance. Journal of Applied Psychology 93, 721–732.

Thau, S., Aquino, K., Poortvliet, P.M., 2007. Self-defeating behaviors in organizations: the relationship between thwarted belonging and interpersonal work behaviors. Journal of Applied Psychology 92, 840–847.

Thau, S., Bennett, R.J., Mitchell, M.S., Marrs, M.B., 2009. How management style moderates the relationship between abusive supervision and workplace deviance: an uncertainty management theory perspective. Organizational Behavior and Human Decision Processes 108, 79–92.

Thau, S., Mitchell, M.S., 2010. Self-gain or self-regulation impairment? Tests of competing explanations of the supervisor abuse and employee deviance relationship through perceptions of distributive justice. Journal of Applied Psychology 96, 1009–1031.

Turillo, C.J., Folger, R., Lavelle, J.J., Umphress, E.E., Gee, J.O., 2002. Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. Organizational Behavior and Human Decision Processes 89, 839–865.

Umphress, E.E., Simmons, A.L., Folger, R., Ren, R., Bobocel, R., 2013. Observer reactions to interpersonal injustice: the roles of perpetrator intent and victim perception. Journal of Organizational Behavior 34, 327–349.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology 54, 1063–1070.

Wu, T.Y., Hu, C.Y., 2009. Abusive supervision and employee emotional exhaustion dispositional antecedents and boundaries. Group & Organization Management 34, 143–169.

Xu, E., Huang, X., Lam, C.K., Miao, Q., 2012. Abusive supervision and work behaviors: the mediating role of LMX. Journal of Organizational Behavior 33, 531–543.

Zellars, K.L., Tepper, B.J., Duffy, M.K., 2002. Abusive supervision and subordinates' organizational citizenship behavior. Journal of Applied Psychology 87, 1068–1076.

Supply Chain Management

Lauren Xiaoyuan Lu, Jayashankar M Swaminathan, Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous article by J.M. Swaminathan, volume 22, pp. 15281–15285, © 2001, Elsevier Ltd.

Abstract

Supply chain management is a critical aspect of conducting any business. In this article, we provide an overview of the advancements in supply chain management. In the initial section, we present alternative definitions and key issues related to supply chain management followed by a discussion of complexities associated with managing supply chains. Subsequently, we discuss major inefficiencies of poor supply chain management. Finally, a brief summary of research activity to date and a discussion of future challenges related to supply chain management are presented.

Introduction

Supply chain management is one of the most essential aspects of conducting business. Many people outside of the direct community (in research and industry) do not realize this because an ordinary consumer often experiences only its effects. Recall the times when the item that you wanted was not available in your favorite garments or grocery store, recall how many times you got a great 'deal' at the end of the season, recall the sudden increases in gas prices due to shortages, recall the times when your e-commerce site promised availability but later could not send the required product or sent you the wrong product, or recall the times when your customized product (like a personal computer or kitchen cabinet) was delayed to a great extent. All the above and several other experiences that consumers have on a routine basis are direct consequences of supply chain practices followed by firms. As opposed to business-to-consumer transactions, supply chain practices have immediate impact on business-to-business transactions.

In the late 2000s, due to glitches of its extensively outsourced supply chain for Dreamliner 787, Boeing experienced substantial delay in launching the new aircraft and incurred more than $2 billion in charges to support and expedite component supplies. Less than 2 years after the first delivery of Dreamliner 787 in 2011, Boeing was ordered to shut down production of the aircraft due to quality issues with batteries. In 2007, Mattel had to recall tens of millions of toys made in China, which becomes poster child for concerns about quality of offshore goods. While some firms suffered from consequences of bad supply chain management, firms such as Amazon, Wal-Mart, and Zara have consistently outperformed competition due to great supply chain capabilities.

Definition

Since Keith Oliver, a consultant at Booz Allen Hamilton, coined the term in 1982, supply chain management has evolved from originally being understood as only logistics to a complex multifunctional corporate undertaking that ranges from procurement and demand forecasting to distribution and after-sales service. Supply chain management is such a vast topic that as a result people often give it a different definition based on their own personal experience. To some, supply chain management is all about managing the supplier base, determining what to outsource and to whom, and managing relationships with the various suppliers. To some others it is efficient ways of transferring goods from one place to another taking into account the distribution and transportation costs. To another set of people it is all about how the different firms in the distribution channel or value chain are integrated in terms of information systems and inventory management practices. To yet another group it is effective management of fixed and variable assets required for running the business. In a sense all these definitions are like the blind men defining the elephant based on its different organs. A comprehensive definition of supply chain management can be given as follows.

A supply chain is the set of entities that are involved in the design of new products and services, procuring raw materials, transforming them into semifinished and finished products, and delivering them to the end customers (Swaminathan, 2001). Supply chain management is efficient management of the end-to-end process starting from the design of the product or service to the time when it has been sold, consumed, and finally disposed of by the consumer. This complete process includes product design, procurement, planning and forecasting, production, distribution, fulfillment, and after-sales support (see **Figure 1**).

Supply chain management issues can be classified into two broad categories – configuration and coordination. Configuration-level issues relate to the high-level design and basic infrastructure of the supply chain and coordination-level issues relate to the tactical decisions and day-to-day operations of the supply chain.

Configuration-Level Issues

Supply Base Decisions

How many and what kinds of suppliers to have? Which parts to outsource and which to keep in-house? How to standardize and streamline procurement practices? Should one use vertical marketplaces for auctions or should one invest in developing highly integrated supply partnerships? How long or short contracts with suppliers should be?

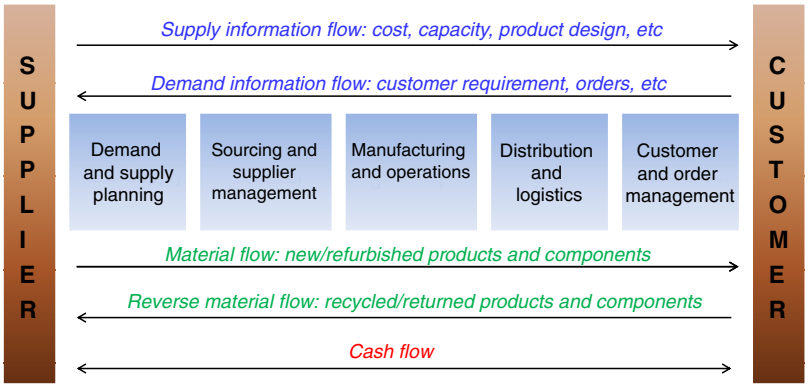

Figure 1 Supply chain process.

Plant Location Decisions

Where and how many manufacturing, distribution, or retail outlets to have in a global production distribution network? How much capacity should be installed at each of these sites? What part of the supply chain should be kept onshore, near-shore, or offshore? What kind of distribution channel should a firm utilize – traditional brick and mortar, direct to consumer via Internet or phone, or a combination?

Product Portfolio Decisions

What kinds of products and services are going to be supported through the supply chain? How much variety to provide to customers? What degree of commonality to have across the product portfolio?

Information Support Decisions

Should enterprise resource planning software be standardized across functional units of a firm? Should the supply chain work on standard protocols such as extended markup language (XML) or on proprietary standards?

Coordination-Level Issues

Material Flow Decisions

How much inventory of different types of products should be stored? Should inventory be carried in finished form or semi-finished form? How often should inventory be replenished? Should a firm make all of its inventory decisions or is it better to have the vendor manage the inventory? Should suppliers be required to deliver goods just in time?

Information Flow Decisions

In what form is information shared between different entities in the supply chain – paper, voice via telephone, electronic data interchange (EDI), XML? To what degree does collaborative forecasting take place in the supply chain? What kind of visibility is provided to other entities in the supply chain during execution? How much collaboration takes place during new product or service development among the supply chain partners?

Cash Flow Decisions

When do suppliers get paid for their deliveries? What kinds of cost reduction efforts are taken across the supply chain (or expected of suppliers)? In a global firm, in which currency will a supplier be paid?

Capacity Decisions

How to optimally utilize the existing capacity in terms of manpower and machines? How to schedule on a manufacturing line to complete jobs on time? How much buffer capacity to have for abnormal situations with excess demand?

As is evident, configuration and coordination issues are interdependent. Configuration issues can be viewed as strategic long-term decisions whereas coordination issues are medium- to short-term decisions. Generally, firms develop a strategy for the configuration-level decisions and then constrain the coordination decisions based on those.

Complexities Associated with Supply Chains

As evident from discussions in the earlier section, supply chain management spans several functional and geographical areas. This introduces complexities both in terms of design and execution of supply chains. Some of the pertinent factors that complicate supply chain management decisions are as follows.

Multiple Agents

Supply chain issues need to be decided by different entities sometimes having different interests. For example, a retailer may want that the distributor provide very high availability for the products but at the same time not charge anything additional to the retailer. The distributor may sometimes agree to that but in turn may want information about actual customer sales which the retailer may not want to share. Even when decisions have to be made within the same firm there could be incentive issues. For example, the marketing or sales department, typically a revenue center, presents the future demand forecast to the manufacturing department that is a cost center. Clearly, there is incentive for the former to overforecast and the latter to underproduce (as compared to the forecast). This creates several difficulties while deciding on the amount of inventory to be stocked. Another related issue is encountered where the marketing department may push for huge amount of

variety in the product/service offerings; the manufacturing department may not want to embrace that because additional complexities are created during execution.

Uncertainty

Accurately matching supply and demand is the ultimate goal of effective supply chain management but that is complicated by uncertainty at various levels of the process. There is uncertainty in product and technology development, in predicting customer demand, in day-to-day operations and manufacturing, and supply. Typically, uncertainty creates more inefficiency in the system. For example, if the final demand for a sweater at a store cannot be predicted accurately then the firm either stocks too little (in which case it suffers from stock-outs) or produces too much (in which case it has to salvage the inventory through a huge sale at the end of the season). Similarly, the uncertainty in supply may necessitate additional buffer inventory.

Information Asymmetry

Since supply chain processes extend across multiple functional units within a firm and often across different firms there is a high degree of asymmetry in terms of information. This is caused primarily by two main reasons – one relates to lack of adoption of information technology and the other relates to reluctance to share information with other supply chain partners. The lack of information causes several problems during actual fulfillment. For example, when a consumer goes to an e-commerce site and buys an item off the electronic catalog, the consumer expects to receive the product on time. The consumer is not aware that the inventory status on the product may be updated only once a week and that the information on the site may be outdated. As a result, the consumer is disappointed when the product does not arrive on time.

Lead Time

Each and every task in the supply chain process needs time to be completed and the resources (labor, machines, or computers) have limited processing capacity. As a result, not all tasks can be completed after the actual demand is known and some of the tasks need to be done up front (which may or may not get utilized based on the actual demand realized). Further, the limited capacity associated with the resources creates variability in the actual realized lead-time, which in turn necessitates greater resource requirements at the next stage in the supply chain. The above complexities lead to several types of inefficiencies in the supply chain that are often perceived as the 'bad effects' of inefficient supply chain management. Some of the major inefficiencies can be classified into the following categories.

Competition

When making supply chain decisions, firms have to take market competition into consideration. For example, forming exclusive trade relationship with key suppliers helps prevent proprietary technology to be leaked to competitors who may source from the same suppliers. As another example, firms' outsourcing decisions can be shaped by what competitors do (see Feng and Lu, 2012). In the computer and electronics industries, many companies rely on contract manufacturers to assemble final products. Due to their location in low-cost regions and scale economies, these contract manufacturers generally have a lower cost structure than original equipment manufacturers. For a long time, Dell, operated its own assembly plants in North America to support its build-to-order model. Under increasing market pressure to lower costs, Dell had to change its supply chain strategy by outsourcing assembly work to contract manufacturers, a strategy that has long been adopted by its key competitors.

Inefficiencies of Supply Chain Management

Poor Utilization of Inventory Assets

One common effect of poor supply chain management is having excess inventory at various stages in the supply chain, at the same time having shortages at other parts of the supply chain. Since inventory forms a substantial part of working assets of a firm, poor management could lead to huge inefficiencies. Lee and Billington (1992) provide an excellent overview of pitfalls and opportunities associated with inventory management in supply chains.

Distortion of Information

Another effect relates to lack of visibility of demand and supply information across the supply chain which causes the bullwhip effect. This effect describes how a small blip in customer demands may get amplified down the supply chain because the different entities in the supply chain generate and revise their individual forecasts and do not collaborate and share actual demand information. Lee et al. (1997) describe the causes and controls for this effect.

Stock-Outs

Poor supply chain management also results in late deliveries and large stock-outs. Fundamentally, these effects are caused due to an inability of the firm to predict the requirement for raw material and equipment capacity together with the uncertainty associated with obtaining deliveries of products on time from its suppliers. Fisher et al. (1994) describe how accurate forecasts in the apparel industry could potentially reduce this inefficiency.

Customization Challenges

As the degree of customization has increased in the marketplace, one of the immediate effects of poor supply chain management relates to late deliveries of customized products. Firms are developing several strategies in order to provide variety while keeping costs under control. These include delaying differentiation of the product and introducing more commonality and modularity in product lines (see Swaminathan and Tayur, 1998).

New Developments in Supply Chain Management

Global Supply Chains

In the increasingly globalized world economy, even a small manufacturing firm may face the challenge of managing an oversea supply relationship. Globalization is not a new phenomenon, but the globalization trend developed in the first decade of the twenty-first century created a more flattened world than history has seen. However, in the recent years the cost differential between the emerging economies and the developed ones is gradually disappearing. This forces many firms to reconsider their supply chain configurations, particularly their cost-driven offshoring strategy. In the last few years, some manufacturing firms, large and small, have brought offshored production back onshore. Meanwhile, the growing market size of the economies outside the developed countries has made expansion of one's global supply chain footprint a more attractive option. Striking the right balance between onshore and offshore sourcing will continue to be a key task in firms' supply chain strategy.

Sustainable Supply Chains

In the last decade, building sustainable supply chains has gathered tremendous attention from environmentalists, NGOs, and businesses. This society wide sustainability initiative has influenced many firms' supply chain strategies. For instance, Wal-Mart in 2005 launched three overarching sustainability goals: (1) use 100% renewable energy; (2) produce zero waste; (3) sell products that sustain our resources and environment (Denend and Plambeck, 2007). To achieve these goals, the company examined various aspects of its supply chains to identify areas that offered the most potential for sustainability. And it used various incentives to motivate its suppliers to contribute to its sustainability goals.

Humanitarian Supply Chains

Between 1974 and 2003, there were 6637 natural disasters worldwide with more than 5.1 billion people affected and a reported damage of $1.38 trillion USD (Ergun et al., 2009). Although the occurrence of these events was hard to forecast, the social and financial impact could have been reduced with proper planning. In particular, humanitarian supply chains, i.e., supply chains for disaster planning and response, can be well designed and maintained for the purpose of quick response and relief effort after disasters strike. Humanitarian supply chains differ from regular supply chains in terms of their supply and demand patterns. Refer to Ergun et al. (2009) for a detailed review on the subject. Another important type of humanitarian supply chains relates to delivery of essentials (food supplements, bed nets, and vaccines) to people in need during nonemergency times (see Swaminathan, 2009; Komrska et al., 2013).

New Technologies Impacting Supply Chains and Big Data

In the last two decades, digital technologies have had a lasting impact on how supply chain activities are conducted. According to a recent study by Intermec, a leading supply chain solutions provider, the top 10 technologies that have the most impact on supply chain operations include: (1) comprehensive connectivity – from 802.11 wireless LAN technologies, cellular networks, and Bluetooth; (2) voice and GPS communication integrated into rugged computers; (3) speech recognition; (4) digital imaging; (5) portable printing; (6) 2D and other bar coding advances; (7) radio-frequency identification; (8) real-time locating system; (9) remote management; (10) wireless and device security.

The proliferation of the aforementioned digital and communication technologies in business and consumer uses has created tremendous amount of data on business transactions, logistical activities, customer characteristics, etc. According to IDC, approximately 750 exabytes (EB) of data were created online in 2009, and it is forecasted to exceed 35 zettabytes (ZB) by 2021. Big data, according to Gartner, is defined as "high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." The real impact of big data on improving supply chain efficiency is yet to be seen, and companies are only starting to tap into the potential offered by the vast amount of data being collected in their information systems. In the face of explosive growth of data, the information industry is going through a revolution, and sophisticated tools are yet to be developed to analyze supply chain-related data for performance improvement.

Supply Chain Research: Past, Present, and Future

The science related to supply chain management traces its history back to the early 1950s when several researchers were interested in understanding the optimal policies related to inventory management. One of the first pieces of work in this stream relates to the models developed by Clark and Scarf (1958) for managing inventories at multiple echelons. Several hundreds of researchers have studied related inventory problems under stochastic and deterministic environments since the 1950s.

This research is captured concisely in the research handbook edited by Graves et al. (1993). There is a large amount of literature in the area of transportation and distribution as well as plant location models in the context of supply chain management. Traditional researchers focused on developing optimal policies and rules for specific supply chain issues assuming a centralized control of the supply chain. In the 1990s researchers have started to study problems which take a decentralized multiagent approach to analyzing supply chain problems, integrate information availability across the supply chain with logistics decisions, develop new models for supply contracts, and demand forecasting and integrate product design with supply chain management. A collection of prominent pieces of research in this area is contained in Tayur et al. (1998) and de Kok and Graves (2003). After several decades of studying supply chains with mathematical models, researchers started to use data and empirical methods to validate supply chain theories and to systematically characterize supply chain practices. For example, using industry-level U.S. data, Cachon et al. (2007)

document the strength of the bullwhip effect, i.e., the phenomenon that demand variability increases moving from the downstream of a supply chain to the upstream. Using firm-level data, Cachon and Olivares (2010) examine the factors that contribute to the difference in finished goods inventory between several leading auto manufacturers. The empirical stream of research will continue to grow as researchers become more creative in finding relevant data on supply chains.

In addition to academic research, several firms in the 1990s developed successfully and employed large analytical and simulation models for supply chain optimization and execution. Arntzen et al. (1995) describe one such system developed for Digital Equipment, and in 2004, Motorola was awarded the Franz Edelman Award by the Institute for Operations Research and Management Sciences (INFORMS) for using operations research methods to launch a comprehensive online negotiation system to support the company's sourcing process which led to $600 million in savings.

In the twenty-first century, firms face severe challenges in terms of global competition and customer requirement for greater variety, shorter and reliable delivery times, and lower prices. The advent of e-commerce has created immense opportunities but at the same time has made firms more vulnerable to logistics pitfalls. Today customers do not just buy products but they buy delivered products. As a result fulfillment is as important as making the sale. As opposed to traditional channels where inventory could be stored to hide other inefficiencies in terms of lead-time and poor forecasts, in the fast-paced electronic business environment such arrangements are not as useful. As a result, firms are beginning to pay more attention toward supply chain management. Both business-to-consumer and business-to-business e-commerce environments have introduced several issues related to supply chain management which are likely to be studied by researchers in the near future.

The prevalence of the Internet has led to the development of vertical marketplaces that promise to reduce the inefficiencies in the buying process in several industries. On one hand, these marketplaces are likely to reduce the cost of goods for the manufacturer due to more competition leading to better prices. This line of thought indicates that in the future supply chains may be more agile and supplier relationships may be short-term oriented. On the other hand, several firms realize that greater benefits can be attained if some of these marketplaces can in fact be used for process integration and collaboration across the supply chain. In such an environment, firms need to develop greater trust so that they would be willing to share information with their supply chain partners. Researchers today are trying to identify under what conditions one or the other scenario may play out and what kinds of new models and analysis need to be developed. A related effect of the Internet is the expansion of global supply chains. Today it is much easier for any supplier located in a remote part of the world to bid on contracts from large firms in developed nations with whom they may not have done business in the past. Issues related to coordination of global supply chain management are likely to be an integral part of supply chain management research in the future. Another important research topic is sustainable supply chain management. Traditionally, researchers have only concerned themselves with efficient movement of goods from supplier to the customer. Now a greater number of researchers are studying problems related to disposal of used products, refurbishing old products, making packaging more environmentally friendly, and basing supplier selection on environmental criteria in addition to traditional criteria related to cost, quality, and reliability. Another new stream of research is the study of supply chains in the service industry. As opposed to traditional manufacturing-oriented supply chains, service supply chains are more complicated due to the inability to store inventory. Uncertainty is handled in those cases using additional buffer capacity. Finally, researchers are beginning to look at behavioral issues such as trust, bounded rationality, mental accounting, etc. that may arise in supply chain management.

See also: Commodity Chains; Location Theory; Market Structure and Performance; Marketing Strategies.

Bibliography

Arntzen, B.C., Brown, G.G., Harrison, T.P., Trafton, L.L., 1995. Global supply chain management at Digital Equipment Corporation. Interfaces 25 (1), 69–93.

Cachon, G., Olivares, M., 2010. Drivers of finished goods inventory in the U.S. automobile industry. Management Science 56 (1), 202–216.

Cachon, G., Randall, T., Schmidt, G., 2007. In search of the bullwhip effect. Manufacturing & Service Operations Management 9 (4), 457–479.

Clark, A.J., Scarf, H., 1958. Optimal policies for a multi-echelon inventory problem. Management Science 6, 475–490.

Denend, L., Plambeck, E., 2007. Wal-Mart' Sustainability Strategy (A). Stanford Graduate School of Business Case OIT-71.

Ergun, O., Karakus, G., Keskinocak, P., Swann, J., Villarreal, M., 2009. Humanitarian supply chain management - an overview. In: Barnhart, C., Clausen, U., Lauther, U., Mohring, R.H. (Eds.), Models and Algorithms for Optimization in Logistics. Germany.

Feng, Q., Lu, L.X., 2012. The strategic perils of low cost outsourcing. Management Science 58 (6), 1196–1210.

Fisher, M.L., Hammond, J.H., Obermeyer, W.R., Raman, A., 1994. Making supply meet in an uncertain world. Harvard Business Review 72 (3), 83–89.

Graves, S.C., Rinnooy Kan, A.H.G., Zipkin, P.H., 1993. Logistics of Production and Inventory. In: Handbook in OR and MS, vol. IV. Amsterdam.

de Kok, A.G., Graves, S.C., 2003. Handbooks in Operations Research and Management Science: Supply Chain Management. Elsevier, Amsterdam.

Komrska, J., Kopczak, L., Swaminathan, Jayashankar M., 2013. When supply chains save lives. Supply Chain Management Review (January–February).

Lee, H.L., Billington, C., 1992. Managing supply chain inventory: pitfalls and opportunities. Sloan Management Review 33 (3), 65–73.

Lee, H.L., Padmanabhan, V., Whang, J., 1997. Information distortion in a supply chain: the bullwhip effect. Management Science 43 (4), 546–548.

Swaminathan, J.M., 2001. Supply Chain Management, International Encyclopedia of the Social and Behavioral Sciences. Elsevier Sciences, Oxford, England.

Swaminathan, J.M., 2009. UNICEF case: getting food to disaster victims. Financial Times, 2010.

Swaminathan, J.M., Tayur, S., 1998. Managing broader product lines through delayed differentiation using vanilla boxes. Management Science 44 (12), S161–S172.

Tayur, S., Ganeshan, R., Magazine, M.J., 1998. Quantitative Models for Supply Chain Management. Kluwer, MA.

Suprasegmentals

Carlos Gussenhoven, Radboud University Nijmegen, Nijmegen, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The prosodic structure of languages is divided into word prosody and sentence prosody. Beyond an organization of vowels and consonants into syllables, languages may have segmental length, tone, and stress, in any combination. The prosodic structure of words can be inferred from their pronunciation by abstracting away from the phonological phrasing structure (the prosodic hierarchy) and the tonal structure of the intonational melody. Boundary tones delimit larger phonological phrase types and some languages have phrase-internal tones (pitch accents). Word prosody distinctions encode lexical and grammatical morphemes, prosodic phrasing imperfectly reflects syntactic structure, and melodic contrasts signal discourse and focus meanings.

Introduction

The prosody of language covers all aspects of speech that are not related directly to the articulation of the vowels and consonants in linguistic expressions, a negative definition that is echoed by the traditional term 'suprasegmentals.' The topic can be approached in positive terms either from the perspective of the phonetic signal or from that of the linguistic structure. The first perspective leads to the analysis of the acoustic signal into four dimensions of variation. These are the spectrum, duration, intensity, and fundamental frequency. It was pointed out by Ilse Lehiste (1970) that the last three dimensions can be observed as properties of the first. We can say how long and how intense the speech signal for some vowel is and what its fundamental frequency is, but we cannot say how intense duration is or how long the fundamental frequency, she argued. In this sense, suprasegmental features are overlaid features, while spectral features are inherent. The second perspective presupposes an understanding of the prosodic structures of languages. It seems fair to say that this understanding has not yet achieved the level of analytic accuracy that has been reached for segmental structures, but considerable progress has been made in the past decades.

Two general points are made before we move on to linguistic structures. One is that the neat distinction between spectral variation and the three suprasegmental dimensions of variation must blur again, because phonological elements are famously promiscuous with respect to the four dimensions of phonetic variation. Thus, linguistic categories that typically are encoded in one of the suprasegmental dimensions of variation frequently also make use of spectral variation, including variation in voice quality. This applies to stress and tone as well as to intonation. The second point is that a great deal of the communicative impact of intonated speech is to be attributed to paralinguistics, the systematic variation that speakers of all languages use to signal universal meanings not encoded in the linguistic structure (Ladd, 2008). Variations in overall pitch range, for example, may signal such general meanings as surprise, authoritativeness, and emphasis (Ohala, 1994). Generally, paralinguistic meanings can be seen as metaphorical interpretations of anatomical and physiological effects on vocal fold vibration (Gussenhoven, 2004, Chapter 4). Ohala's Frequency Code connects the effect of larynx size on frequency of vocal fold vibration with the inferred size of the animal or human producing it, causing high pitch to signal 'small' meanings like submissiveness, vulnerability, and uncertainty and low pitch to signal their opposites. The disentanglement of these paralinguistic meanings – from the meanings encoded in intonational morphemes – will remain an intriguing area of research. A related point is that intonation differs from other components of linguistic structure in that deviations in the timing or pitch range from a canonical pronunciation will trigger paralinguistic interpretations of speaker intentions, while comparable spectral deviations may be interpreted just as indications of articulatory casualness or of temporary physical constraints on speech production.

The approach adopted in the remainder of this article takes the phonological representation as the starting point, selecting those elements in that representation whose realization typically or in large measure involves nonspectral variation, much as has been done in Fox (2000). Since few languages will exemplify all the prosodic phenomena that have been reported, this approach inevitably looks at the languages of the world as a single body of data and will identify aspects of prosodic structure cross-linguistically. One way to divide the topic is to separate word prosody from sentence prosody.

Word Prosody

Broadly, the phonological structure of words can be more than a syllabified string of consonants and vowels in three ways. One is the existence of long segments by the side of short ones, a second is the presence of metrical structure, and a third is the presence of lexical tone. Languages do not require any of these properties. French, for instance, has been claimed to be a language without any word prosodic structure, which is supported by the finding that French listeners perform poorly in short-term memory tasks involving the position of the stress (Peperkamp and Dupoux, 2002).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52024-8

Quantity

When duration differences are encoded in the phonology, they are referred to as quantity differences or length differences. Vowels as well as consonants may contrast for quantity. Finnish has quantity contrasts for both consonants and vowels, as has Japanese. Hawaiian contrasts long and short vowels only, while Italian contrasts long and short consonants only. Spanish and French have neither long vowels nor geminate consonants. Quantity is described with the help of moras, which make up the syllable rime, the part of the syllable minus the initial (onset) consonant(s), but including the closing (coda) consonants. A Finnish syllable rime containing a short vowel, like the first syllable of *eri* 'different,' has one mora, while a syllable with a long vowel, like the first syllable of *vaara* 'danger,' has two. A long vowel is prosodically identical to a short vowel plus a coda consonant, making *vaara* and *arki* 'workday' prosodically equivalent. Three moras occur in the first syllable of *arkki* 'sheet of paper' or *aarre* 'treasure,' where the first half of the geminate consonants [kk] and [rr] is a coda in the first syllable and the second half an onset in the second. Syllables with four moras, like [aark] in the proper name *Jotaarkka*, are extremely rare (Suomi et al., 2008). Japanese syllables have one, two, and very rarely three moras, as illustrated by, respectively, [o] 'tail,' [on] 'speech sound,' and [aan] in (*Nyuu*) *Jiraando* '(New) Zealand.' Trimoraic rimes in which all three moras are filled by the same segment are extremely rare and have been reported for Estonian and related languages and for Dinka. The most frequent phonological context of geminates is intervocalic. A language with word-initial, word-medial, and word-final geminates has been reported in Kraehenmann (2001), together with a discussion of their moraic representation.

Word Prosodic Metrical Structure

A language has metrical word structure if the syllables in a word differ in prominence or weight from other syllables in the same word. In such cases, the more prominent syllable is stressed and the less prominent ones are unstressed. Stress systems show a number of tendencies. One of these is alternation, the tendency for stressed and unstressed syllables to alternate. Another is culminativity, the restriction to a single most prominent syllable. When there are two or more stressed syllable, the most prominent one is referred to as the primary stress (aka main stress, word stress) and the other or others as secondary stress. A third tendency is for the main stress to occur at or near a word edge, the demarcative tendency. These three aspects are illustrated by Pintupi (Australia), where words always have the primary stress on the first syllable, while in careful pronunciation, secondary stresses occur on odd syllables after the first, as in the 12-syllable [ˈjumaˌɻiŋkunˌtamaˌɻacaˌnampaˌluɻa] 'avoid (the camp) of all those mothers-in-law' (stresses inferred, Hansen and Hansen, 1969). To capture this alternation, stress systems are described in terms of feet, disyllabic strong-weak (trochee) or weak-strong (iamb) templates, which are assigned sequentially through the word. Pintupi is analyzed as having trochees assigned from left to right, which accounts for the fact that any odd numbered word-final syllable is not stressed (i.e., the last two syllables of words with an odd number of syllables are unstressed). Culminativity and demarcativeness are described by a higher level prominence assignment to either the left or right. Further complexity may be due to interdependencies between stress and the segmental composition of the syllable. In quantity-sensitive languages, long vowels, or long vowels and closed syllables, do not tolerate being in a weak position, and thus will interrupt any regular count through the word. Hawaiian is a quantity-sensitive trochaic language that assigns feet right-to-left, so that [ko'hola] 'reef' has penultimate stress, but [koho'laa] 'whale' has final stress (Elbert and Pukui, 1979). Another type puts no limit on the distance of the heavy syllable from the word edge (unbounded feet). Classical Arabic has initial stress, as in ['balaħatun] 'date,' but if there is a later heavy syllable, the rightmost one has the stress, as in [manaa'diilu] 'kerchiefs.'

At the right edge, the most frequent location of main stress is penultimate (e.g., Mohawk, Polish, Quechua, Swahili), but final and antepenultimate locations are common, often in the same language. On the left, the most frequent position is initial (e.g., Bengali, Czech, Finnish, Icelandic), but stress on the second syllable is common. Extrametricality is a descriptive device used to explain regular or exceptional stress locations further away from the edge. Exceptional antepenultimate stress in languages with regular penultimate stress can be explained by making the last syllable or the last consonant 'invisible' to foot assignment. The opposite ploy is called catalexis, the postulation of an extra syllable at the word end, which is presumed to be parsed as a weak syllable in a foot. Languages may have lexicalized stress within an otherwise-regular stress system, in which case stress will be lexically contrastive, as in English '*insight* vs *in'cite*. Stress location may be sensitive to morphological structure, as in Spanish '*ablo* vs *a'blo* or English '*overhaul* (noun) vs *over'haul* (verb). It generally is believed that no language *only* has lexicalized stress, such that stress locations do not yield to any generalization. A recent survey is van der Hulst et al. (2010).

From a distributional point of view, the notion stress is reasonably uncontroversial. Hyman (2006) characterized a language as having 'stress' if it occurs on a syllable, as opposed to a mora; if it is culminative, it being the most prominent syllable in the word; and if it is obligatory, meaning that there should be no words without stress, discounting unstressed words that cannot appear in an utterance without stressed words, like the English articles. From the perspective of its phonological and phonetic substance, stress is more elusive. On the assumption that the origin of stress is a rhythmic alternation of hyper- and hypoarticulation, it is not surprising that stressed syllables are typically longer and pronounced with greater articulatory care than unstressed syllables, resulting in more canonical articulations with less articulatory reduction and a sustained amplitude across the spectrum (e.g., Sluijter et al., 1997), which enhances intelligibility (Lu and Cooke, 2009).

In many languages, hyperarticulation tendencies are reflected in the phonology. English is a good example of a language in which the composition of a stressed syllable is phonologically different from an unstressed one. Unstressed syllables usually have a three-way vocalic distinction, with details depending on the dialect (Bolinger, 1986: p. 347 ff), as

illustrated by the vowels in the second syllables of *cranial*, *usual*, and *mineral*. Any other vowel would attract the stress, as in *denial*, *bestowal*, *betrayal*, and so on. Second, the stressed syllable in English acts as the attractor of the intonational pitch accent. Less than 40% of the syllables in English prose texts are stressed, and in a competent reading, only about 65% of these are pitch accented. Stressed syllables of other languages may have similar phonological features distinguishing them from unstressed syllables. In Egyptian Arabic, stressed syllables are always pitch accented, for instance (Hellmuth, 2007). In Finnish, which has a quantity distinction on both consonants and vowels and primary stress on the first syllable, all vowels that appear in stressed syllables also can appear in unstressed syllables. (According to Suomi et al. (2008: p. 77), some of the more complex syllable structures occur only in initial position.) If languages have no segmental distinction between stressed and unstressed syllables, the stressed syllable may be variably pitch accented (Beckman and Edwards, 1994), or else the difference is phonetic. Spanish distinguishes unaccented stressed and unstressed syllables with the help of a phonetic duration difference (Ortega-Llebaria and Prieto, 2010).

A fairly widely honored, but deeply confusing, tradition in English phonology is to collapse unaccented stressed and pitch-accented stressed syllables into a single paradigm of syllables with different degrees of stress. The presence of a pitch accent on a word's stressed syllable does not alter any of its segmental features, but it adds pitch features in and around the accented syllable that make up the intonation contour of the utterance (Bolinger, 1986). The determination of what is a stressed syllable belongs virtually exclusively in the lexicon; a minor role for syntax is found in the obligatory stressed pronunciation of function words before deletion sites, as illustrated in *Sue can* (kən) *come, but I don't know if Gary can* (kæn) (Kaisse, 1985; Selkirk, 1986, 1995). The reasons for the presence of pitch accents on words are more varied. They are morphological, phonological, or syntactic or they have to do with information structure (cf Gussenhoven, 2011). Compound formation may be marked by lack of a pitch accent in the second constituent, as in *COURSE requirements*, while phonological phrasing is responsible for the difference in accentuation in *The gave [the CHInese DISHES]φ* ('They gave the dishes that were of Chinese origin') *They gave [the CHINESE]φ [DISHes]φ* ('They gave dishes to the Chinese') (Hayes, 1989). Syntax is involved in the generalization that predicates may lack a pitch accent when adjacent to an argument. The predicate–object combination in *I used chopsticks* typically is pronounced with a pitch accent on *CHOPsticks* only, while the noun phrase *USED CHOPsticks* will have two pitch accents (Schmerling, 1976; Selkirk, 1986; Birch and Clifton, 1995). A condition on the deletion of the accent on the verb is that the proposition should be eventive (Gussenhoven, 1983; cf the distinction between stage-level and individual-level predicates in Kratzer, 1995). For instance, the verb in *Too many cooks spoil the broth* has an obligatory pitch accent (in addition to the pitch accents on *Too*, *cooks*, and *broth*), because the sentence is a 'contingency' sentence ('If there are …' etc.). The absence of that pitch accent would invite the interpretation that the number of cooks that spoil the broth is in fact large. Information structure is involved in the difference between *Is that an ENTrance requirement or a COURSE requirement?* In the literature that collapses the presence of a pitch accent with word-stressed syllables, the compound *COURSE requirements* would be said to have secondary stress on the penultimate syllable of *requirements* and primary stress on *course*, while in the case of the contrastive focus on *course* in *COURSE requirement*, the word *requirement* would be said to be deaccented. The prosodic structures of these two expressions are the same, however.

Lexical Tone

A frequently applied morphosyntactic criterion for lexical tone, and hence of the notion *tone language*, is whether tone(s) enter into the specification of words or morphemes. If they do, the language is a *tone language*. Examples are found in Asia (Sino-Tibetan languages, Japanese), Africa (Bantu languages, Khoi-San languages, Nilo-Saharan languages), and many languages in the Americas (Yip, 2002). Mandarin (Sino-Tibetan) has five words that consist of the syllable [ma]. Four of these have a lexical tone: [ma˥] 'mother,' [ma˧˥] 'hemp,' [ma˨˩˦] 'horse,' [ma˥˩] 'scold,' where the Chao tone letters give an iconic indication of the pitch contour to the left of a vertical bar, which indicates the pitch span (Chao, 1930). One [ma] is toneless, a question particle ('neutral tone'). Notation of tones in tone languages spoken in Africa and, to a lesser extent, in the Americas use accents over vowels (high level [á], low level [à], mid level [ā] falling [â], rising [ǎ], etc.), whereby long vowels are given as [aa], as in Mamvu [máàkà] 'type of seasoning' vs [màákà] 'cat.' Frequently, tone is involved in the morphology, as in Makonde [nìndì tálèékà] 'I cooked' vs [nìndà tàlèékà] 'I will cook' and Mamvu [áfù] 'man,' [àfú] 'men' (examples from Kaji, 1999). Estimates of the proportion of tone languages run from 50% (Maddieson, 1978) to 70% (Yip, 2002).

A subclass of tone languages frequently is singled out as being pitch-accent languages. Hyman (2009) argued that languages that have been identified as such differ from other tone languages in the distribution of the tone contrast over the word. Tone contrasts typically are not located on every syllable or mora of the word. Sino-Tibetan languages may have very 'dense' distributions, with only a minority of words, like the interrogative particle [ma] of Mandarin, having no tone of their own ('neutral tone'). Japanese is an example of a language with a 'sparse' distribution of tone. Only about half of the Japanese words, the accented words, have a lexical pitch accent (HL), the other half being lexically toneless, or unaccented. Since the pitch accent can be lexically specified once per word for any syllable, the number of prosodic patterns for words of n syllables is $n - 1$. A triplet of disyllables is [ha˥si] 'chopsticks,' [hasi˥] 'bridge,' and [hasi] 'end,' where the right hook indicates a pitch accent on the preceding syllable.

The study of tonal structures in African languages in the second half of the last century led to the development of autosegmental representations (Goldsmith, 1976), whereby the phonological specification of words consists of at least two parallel linear tiers, in which the elements (the 'segments') are synchronized by means of associations. In

the case of the synchronization of tones on the one hand and vowels and consonants on the other, the associations are not direct, but rather they are mediated by the prosodic structure. Some language use the syllable as the tone-bearing unit (TBU), others the mora, whereby the tone-bearing status of the mora is usually conditional on the presence of a vowel or sonorant consonant in it, thus excluding obstruent consonants from being tone bearers. Although the Japanese pitch accent is assigned to a syllable, the tones associate with moras. That is, even though the number of moras in [hai] is the same as that in [hasi], there could be only two words with the monosyllabic structure, [hai] and [hai˥], both of which happen to exist, 'lungs' and 'yes,' respectively. The associations of the lexical pitch accents with the previous five words are shown in (1) for the disyllables and (2) for the monosyllables. Structurally, there is no (2b).

```
(1) a.   h a s i        b.   h a s i          c. h a s i
         |   |                  |
         H   L                 H L
         'chopsticks'           'bridge'       'edge'

(2) a.   h a i          b.   -                 c.  h a i
         | |
         H L
         'yes'                                 'lungs'
```

Autosegmental representations explain, as in (1a) and (2a), how the same tone string, here HL, can have different syllabic pitch effects. The first syllable in (1a) is high, the second is low, and the syllable in (2a) has a fall. For this reason, linguists are careful to distinguish phonological tones from their phonetic realization.

Tone languages spoken in Africa have varying degrees of density. Somali is comparable to Japanese in that it has maximally one lexical H-tone (or 'pitch accent') per word, but it is assigned to the mora, which means that a monosyllable like [ei] can have H either on the first mora ('dog') or on the second ('dogs'). Unaccented moras have L-tone. The location and presence of the H pitch accent in Somali largely signal morphological and syntactic, as opposed to lexical information. In addition to marking number in some nouns by its location, the H pitch accent is deleted in WH-phrases, for instance, as in [ei geé] 'which dog'? (Hyman, 1981). Other languages have a number of word melodies, like H, HL, LHL, that are specified along with the vowels and consonants for each word. The way tone strings distribute themselves over the TBUs of the word formed a significant motivation behind the development of autosegmental phonology, whose essential claim is that more than one segment on one tier may correspond to a single unit on the other. This is illustrated for the Mende words in (5) (from Leben, 1978). All three words happen to have the LH word melody, but because of the different numbers of syllables, their syllabic tone patterns come out as 'rise' for (3a), 'low-high' for (3b), and 'low-high-high' for (3c). This is achieved by assuming that the first tone of the melody associates with the first syllable and subsequent tones with subsequent syllables, if there are any, that leftover tones are associated with the final syllable and leftover syllables are associated with the last tone. (Not all the words of the language follow precisely this pattern of association.)

```
(3) a. mba          b. navo          c. ndavula
       |\               | |                | |/
       L H              L H              L H
       'rice'           'money'          'sling'
```

Tone grammars can be quite complex. There is a general tendency for tones to be pronounced later than the syllables with which they are lexically associated, which leads to a wide variety of rules for spreading and moving tones rightward. Also, in the way they regulate the associations of tones, grammars may effectively ignore one of the tones, leaving empty TBUs to be filled up with a 'default tone' (Hyman, 2001). Both features are illustrated by Nochixtlán Mixtec, where a final H-tone in morphemes associates with the rightmost mora of the next word. At the same time, all moras must be provided with tones. In (4a), the tone of [kasi^{H}] 'eat' associates with the last syllable of [lana^{MH}]. The preceding H of the verb prefix [nā^{H}] will fill up the first syllable of [lana^{MH}], while the initial three empty syllables are filled with default M (not shown). The final MH of [lana] are deleted. The resulting pronunciation is (4b) (McKendry, 2013).

```
(4) a. nã          kasi lana        b. nā̃ kāsī láná
                                       'May the child eat'
       H           H   MH
       SUBJUNCTIVE eat child
```

Many two-tone systems in Bantu are underlyingly H vs Ø, as opposed to H vs L, whereby Ø is filled in with L on the surface. The motivation for such privative analyses is provided by the presence of rules that refer to H and the absence of any rules that refer to L (Hyman, 2001). For instance, no morpheme could consist of, contain or assign L, while for H at least one of these operations exists. Whereas H is typically subject to distributional restrictions, for instance, in never occurring in nonperipheral syllables in the underlying word or in being banned from final syllables in the phrase on the surface, the distribution of L can be expressed only in terms of where H does *not* appear. Languages may have the same tone as a default tone and an underlying tone. In Nochixtlán Mixtex, an L-tone between H and M is raised to M, causing HLM to be HMM on the surface. If the third position is empty, however, a preceding HL sequence remains intact, even though an M will appear as a default tone on the surface, giving HLM (McKendry, 2013). If there is a default tone, it is most likely to be L in a two-tone system and M in a three-tone system, but these patterns are not exclusive.

Sentence Prosody

The prosody of the sentence is determined by its prosodic phrasing and its tonal structure. The prosodic phrasing consists of a hierarchically structured set of phonological constituents (Selkirk, 2011; Nespor and Vogel, 1986; Hayes, 1995). The tonal structure has different sources. Boundary tones come with one or more of the phonological constituents, intonational pitch accents, and any lexical tones that are introduced by the words.

The Prosodic Hierarchy

In (4), an example of a prosodic representation is given of an English sentence, together with frequently used symbols. The boundary tones are subscripted for the phonological constituent they come with, in this case the IP.

H-tone, pronounced after the pitch accent on *cooks*. Boundaries of this type often are felt to correspond with a comma in writing. The utterance, finally, roughly corresponds with the intuitive notion of spoken sentence. There are three phonological phrases (*Too many cooks*, *spoil*, and *the broth*), two intonational phrases (*Too many cooks* and *spoil the broth*), and one utterance (*Too many cooks spoil the broth*).

The higher ranking phonological constituents are less predictable from other aspects of the linguistic structure than the lower ranking ones. The most important factor for the phonological phrase and beyond is the syntactic structure, but great length of constituents may lead to the insertion of constituent breaks. For instance, while *Swans like to swim in the river* may well be one IP, a replacement of *Swans* by *Hippopotamuses* is likely to cause a separate IP to arise for the subject. Conversely, a short length may lead either to restructuring, the merging of constituents that would be expected to be separate

(5)

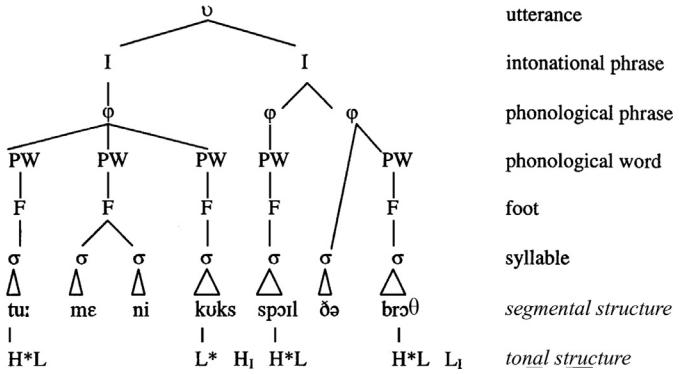

The entire structure above the syllables is assumed to determine how these syllables are rhythmically organized, just as the structure above the foot does this for feet, and so on. The constituent that dominates the foot is the phonological word. In (4), each foot happens to be dominated by a phonological word. A diagnostic for the phonological word in Germanic languages is syllabification: within this constituent, consonants syllabify as onsets with vowels to their right. The phonological word may be smaller or larger than the morphological word. In Germanic languages, each lexical constituent is a separate syllabification domain (e.g., English *cat's eyes* syllabify as [kæts.aɪz], not as *[kæt.saɪz], cf the single phonological word *capsize*). Clitics are morphemes that are included in a following or preceding phonological word (e.g., English *What is he* [wɒt.si] *saying?*). The phonological phrase may determine the distribution of accents, favoring first and last accents. Although, *Too many*! will have two accents, *Too many* COOKS is likely to lose the pitch accent on *many*.

The term 'stress shift' is applied to cases in which such rhythmic deaccentuation leads to the sole presence of a pitch accent on a foot with secondary stress, as in JAP*a*NESE, but JAP*a*nese FUR*n*iture (Gussenhoven, 2011). The intonational phrase typically comes with *boundary tones* at its edges. The intonational phrase *Too many cooks* in (4) ends with an

on the basis of the syntax, or cliticization, the inclusion of a short syntactic constituent into a lower ranked constituent than it would have formed by itself if syntax were the only determinant. An example of restructuring occurs in verb–adverb combinations in American English (Nespor and Vogel, 1986, p. 178). If the adverb is a single phonological word, it is likely to merge with the verb into a single phonological phrase. Thus, in *Rabbits reproduce quickly*, the words *reproduce* and *quickly* will combine in a single phonological phrase, while in *Rabbits reproduce quickly and quietly* the verb will retain a phonological phrase to itself. In some varieties of English, the difference is shown by the pronunciation of *reproduce*. In the first example, a pitch accent will occur on *re-*, but in the second, there will (also) be one on *-duce*. In verb–object combinations, a similar situation exists. Thus, *spoil the broth* may form a single phonological phrase, as shown by varieties that allow stress shift in verb–object combinations where the object is a single phonological word. The pronunciation of the verb in MAIN*tain* OR*der* will then contrast with that in OR*der to main*TAIN. In *We need to maintain order and discipline*, however, the pronunciation *MAIN*tain* is ungrammatical in those same varieties. In English, cliticization applies to pronouns and auxiliaries, for instance, as illustrated by the pronunciation of *What is he* in the previous paragraph. There is no consensus on how many

prosodic constituents there are across the languages of the world and terminology will vary. Japanese has an *accentual phrase*, which seems to lie halfway between PW and the phonological phrase, also referred to as a *minor phrase*, while a *clitic group* has been claimed for Italian and English, for instance.

Preboundary lengthening refers to the phenomenon that the last syllable of a higher phonological constituent is longer than a medial syllable, an effect that is larger as the constituent rank is higher. At the level of the phonological word in English, it is responsible for the difference in pronunciation between *sin tax* and *syntax*, where [sɪn] in the first structure is longer than in the second. And at the higher end, the insertion of an intonational phrase boundary after *go* in *Let's go and have dinner* will considerably lengthen *go* (e.g., Beckman and Edwards, 1990; Wightman et al., 1992; Gussenhoven and Rietveld, 1992; Byrd et al., 2006).

Initial strengthening is the more careful pronunciation of initial segments of a phonological constituent. For instance, the utterance-initial [t] in *Tie it up*! is likely to have a wider area of contact, a longer period of aspiration (a longer, positive Voice Onset Time) than the word initial one in *Stronger ties*, while the same will be true for the latter [t] relative to the foot-initial one in *Valentine* (Cho and Keating, 2001; Keating et al., 2003; Cho et al., 2007).

Finally, the constituents in the prosodic hierarchy often define the context of segmental restrictions. For instance, syllable final plosives and affricates are voiceless in Turkish; syllable final obstruents are voiceless in Dutch; within Italian prosodic words, intervocalic [z] occurs to the exclusion of [s]; and in Kimatuumbi, all vowels are shortened except for any long vowel that ends up in the last word in the phonological phrase (Nespor and Vogel, 1986).

Tone Structure

The tones in a sentence have three sources. First, there are the tones that come with the morphemes in the structure (i.e., the lexical tones). Second, phonological constituents like the intonational phrase and the phonological phrase may begin or end with boundary tones, and third, intonational pitch accents may occur on certain words. Boundary tones will signal the prosodic phrasing of utterances, and since this largely reflects the morphosyntactic structure, will play a crucial role in parsing speech (Cutler et al., 1997; Carlson, 2009; Pannekamp et al., 2005). In addition, there may be a choice of boundary tones, in which case, they signal discourse meanings like 'statement,' 'finality,' 'question,' and so on. Japanese sentences, for instance, have tones from the first two sources (Pierrehumbert and Beckman, 1988). In addition to the lexical pitch accents, the accentual phrase invariably begins with an LH melody, notated LαHα, and the utterance is ended by either an L, signaling declarative meaning, or H, for interrogative meaning. Most European languages derive their tones from the second and third sources, in each case providing a large number of choices, giving a wide variety of discourse meanings (e.g., Pierrehumbert, 1980; Grice, et al. 2005; Frota, 2000; Ladd, 2008; Gussenhoven, 2005; Chen et al., 2007). Swedish and Norwegian have tones from all three sources (Kristoffersen, 2000; Riad, 2013), as do Northern Bizkaian Basque (Hualde et al., 1994), Belgrade Serbian (Smiljanić, 2002), and the dialects in Germany, Belgium, and the Netherlands known as Central Franconian (*Mittelfränkisch*) and Limburgish (e.g., Gussenhoven and Peters, 2004) (cf van der Hulst, 1999). There is considerable variation in the proportions of intonational and lexical tones. For Standard Chinese, the existence of an interrogative utterance boundary tone has been claimed (Duanmu, 2007), where all other tones are lexical. At the other extreme, there is the existence of a pitch-accent-marked focus governing words like [o] 'even' in Bengali, which otherwise only has intonational tones (Lahiri and Fitzpatrick-Cole, 1999). When all the tones have been sequenced into a string, there may be rule-governed assimilations or deletions of tones, after which the surface representation is converted to tonal targets (Pierrehumbert, 1980). Interpolations may be linear or may have somewhat-convex or convex shapes (Niebuhr, 2007; Barnes et al., 2013).

In (5a), the tones for the Japanese word for *New Zealand* are illustrated. The word has a lexical pitch accent (LH) on the penultimate syllable; the declarative tone is notated Lʊ. Together with the initial boundary tones of the accentual phrase, these tones will associate with the available moras in the word, and the target of each tone is indicated by a bullet. (Many speakers avoid rising pitch on a single syllable, as in *nyuu* in (5), replacing it with high pitch.) In (5b), the Portuguese equivalent is shown, which will have a prenuclear H* pitch accent on the stressed syllable [no] and a nuclear H + L* pitch accent in the stressed syllable [lã], where the leading H-tone is pronounced on the preceding syllable, followed by a boundary L-tone of the intonational phrase (Lɪ). In (5c), the Venlo Limburgish word is shown, which illustrates a case of tone assimilation. The language has a lexical H-tone (Accent 2), which assimilates to L if it occurs on the last mora of the intonational phrase and is immediately followed by Lɪ. Tone bearers are the sonorant moras in accented syllables ([zee] in (5c)) and IP-final syllables ([lɑnt] in (5c)), which require a tonal association. This is the reason why lexical L spreads to the available second mora in [zee], which has been left unused by the H*, which associates to the first mora of [zee]. The phonetic reflex of the combination of lexical L and the final boundary Lɪ is a leveling off or slight reversal of the falling pitch movement, as shown in (5c), which contrasts with the straight fall used for Accent 1.

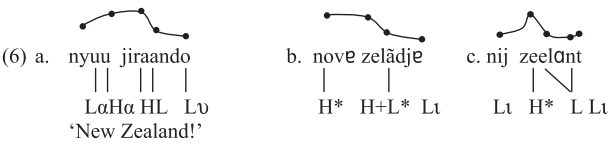

Thus, just as lexical tones are subject to a tone grammar, so too are intonational tones. The distinction between lexical and postlexical phonology (*see* Phonology) is relevant in that no lexical tone rules can affect intonational tones, but as example (6c) shows, postlexical rules can affect both lexical and intonational tones. As a further example, Japanese has rules that govern the priority of tonal associations in cases in which there are more tones than moras, such as when a [LαHα HL Lʊ] tone string (see (6a)) needs to be pronounced on a monomoraic accented word like [ka˥] 'passable school

grade' or a bimoraic one like [hai˥] 'yes' (Gussenhoven, 2004, Chapter 10).

The Expression of Focus

Languages frequently exploit prosodic structure for the expression of *focus*, the marking of information status in sentences (REF Information structure/focus; *see* Information Structure in Linguistics). Frequently, this occurs through the compression of the pitch range over words after the focus constituent. In European languages, this is achieved by the deletion of pitch accents after the focus, with conditions for deletion varying from language to language (Vallduví, 1992; Swerts et al., 2002; Ladd, 2008). In other languages, the words after the focus constituent have reduced pitch range, but the tonal structure remains intact, meaning that lexical distinctions in Chinese are retrievable postfocally. It has been suggested that postfocal compression is an areal feature of Europe and North Asia (Xu et al., 2012). Different meanings or categories of information structure are discussed in Lee et al. (2007). For many languages, an understanding of the prosodic expression of focus structure is impossible without considering the syntactic constraints on focus, because these may implicate the distribution or identity of boundary tones or lexical tones (cf Aboh et al., 2007; Fiedler and Schwarz, 2010). Still, in many languages, information structure may show no prosodic effect of prosody at all (e.g., Zerbian, 2007).

See also: Linguistics: Overview; Phonology; Speech Perception; Speech Production, Psychology of; Speech Recognition and Production by Machines.

Bibliography

Aboh, E.O., Hartmann, K., Hartmann, M., 2007. Focus Strategies in African Languages: The Interaction between Focus and Grammar in Niger-Congo and Afro-Asiatic. Walter de Gruyter, Berlin.

Barnes, T.J., Veilleux, N., Brugos, A., Shattuck-Hufnagel, S., 2013. Tonal center of gravity: a global approach to tonal implementation in a level-based intonational phonology. Laboratory Phonology 3, 337–383.

Beckman, M.E., Edwards, J., 1994. Articulatory evidence for differentiating stress categories. In: Keating, P.A. (Ed.), Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III. Cambridge University Press, pp. 7–33.

Birch, S., Clifton, C.J., 1995. Focus, accent, and argument structure: effects on language comprehension. Language and Speech 38, 365–391.

Bolinger, D., 1986. Intonation and Its Parts: Melody in Spoken Language. Stanford University Press, Stanford, CA.

Byrd, D., Krivokapic, J., Lee, S., 2006. How far, how long: on the temporal scope of prosodic boundary effects. Journal of the Acoustical Society of America 120, 1589–1599.

Carlson, K., 2009. How prosody influences sentence comprehension. Language and Linguistic Compass 3, 1188–1200.

Chao, Y.-R., 1930. ə sistim əv "toun-letəz". Le Maître Phonétique 45, 24–27.

Chen, A., den Os, E., de Ruiter, J.P., 2007. Pitch accent type matters for online processing of intonation status: evidence form natural and synthetic speech. The Linguistics Review 24, 317–344.

Cho, T., Keating, P.A., 2001. Articulatory and acoustic studies of domain-initial strengthening in Korean. Journal of Phonetics 29, 155–190.

Cho, T., McQueen, J.M., Cox, E.A., 2007. Prosodically driven phonetic detail in speech processing: the case of domain-initial strengthening in English. Journal of Phonetics 35, 210–243.

Cutler, A., Dahan, D., van Donselaar, W., 1997. Prosody in the comprehension of spoken language: a literature review. Language and Speech 40, 121–201.

Duanmu, S., 2007. The Phonology of Standard Chinese. Oxford University Press, Oxford.

Elbert, S.H., Pukui, M.K., 1979. Hawaiian Grammar. University Press of Hawaii, Honolulu.

Fiedler, I., Schwarz, A., 2010. The Expression of Information Structure: A Documentation of Its Diversity across Africa. Benjamins, Amsterdam.

Fox, A., 2000. Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals. Oxford University Press, Oxford.

Frota, S., 2000. Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation. Garland Press, New York.

Goldsmith, J., 1976. Autosegmental Phonology. MIT, Doctoral dissertation.

Grice, M., Baumann, S., Benzmüller, R., 2005. German intonation in autosegmental-metrical phonology. In: Jun, S.-A. (Ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, pp. 55–83.

Gussenhoven, C., 1983. Focus, mode and the nucleus. Journal of Linguistics 19, 377–417.

Gussenhoven, C., 2004. The Phonology of Tone and Intonation. Cambridge University Press, Cambridge.

Gussenhoven, C., 2005. The Phonology of Tone and Intonation. Cambridge University Press, Cambridge.

Gussenhoven, C., 2011. Sentential prominence in English. In: van Oostendorp, M., Ewen, C.J., Hume, E., Rice, K. (Eds.), The Blackwell Companion to Phonology. Wiley- Blackwell, Malden, MA and Oxford, pp. 2780–2806.

Gussenhoven, C., Peters, J., 2004. A tonal analysis of Cologne Schärfung. Phonology 25, 251–285.

Gussenhoven, C., Rietveld, A.C.M., 1992. Intonation contours, prosodic structure and preboundary lengthening. Journal of Phonetics 20, 293–303.

Hansen, K.C., Hansen, L.E., 1969. Pintupi phonology. Oceanic Linguistics 8, 153–170.

Hayes, B., 1989. The prosodic hierarchy in meter. In: Kiparsky, P., Youmas, G. (Eds.), Rhythm and Meter. Academic Press, Orlando, FL, pp. 201–260.

Hayes, B., 1995. Metrical Stress Theory. University of Chicago Press, Chicago.

Hellmuth, S., 2007. The relationship between prosodic structure and pitch accent distribution: evidence from Egyptian Arabic. The Linguistic Review 24, 291–316.

Hualde, J.M., Elordieta, G., Elordieta, A., 1994. The Basque Dialect of Lekeitio. Universidad del País Vasco, Bilbao.

van der Hulst, H. (Ed.), 1999. Word Prosodic Systems in the Languages of Europe. Mouton de Gruyter, Berlin.

van der Hulst, H., Goedemans, R., van Zanten, E. (Eds.), 2010. A Survey of Accentual Systems in the Languages of the World. Mouton de Gruyter, Berlin.

Hyman, L.M., 1981. Tonal accent in Somali. Studies in African Linguistics 12, 169–203.

Hyman, L.M., 2001. Privative tone in Bantu. In: Kaji, S. (Ed.), Proceedings of the Symposium on Cross-Linguistic Studies of Tonal Phenomena: Tonogenesis, Japanese Accentology, and Other Topics. Tokyo University of Foreign Studies, Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA), pp. 237–257.

Hyman, L.M., 2006. Word-prosodic typology. Phonology 23, 225–257.

Hyman, L.M., 2009. How (not) to do phonological typology: the case of pitch-accent. Language Sciences 31, 213–238.

Kaisse, E., 1985. Connected Speech: The Interaction of Syntax and Phonology. Academic Press, Orlando, FL.

Kaji, S. (Ed.), 1999. Proceedings of the Symposium on Cross-Linguistic Studies of Tonal Phenomena: Tonogenesis, Typology, and Related Topics. Tokyo University of Foreign Studies: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).

Keating, P., Cho, T., Fougeron, C., Hsu, C., 2003. Domain-initial articulatory strengthening in four languages. In: Local, J., Ogden, R., Temple, R. (Eds.), Papers in Laboratory Phonology, vol. 6. Cambridge University Press, Cambridge, pp. 143–161.

Kraehenmann, A., 2001. Swiss German stops: geminates all over the word. Phonology 18, 109–145.

Kratzer, A., 1995. Stage-level/individual-level predicates. In: Carlson, G.N., Pelletier, F.J. (Eds.), The Generic Book. University of Chicago Press, Chicago, pp. 125–175.

Kristoffersen, G., 2000. The Phonology of Norwegian. Oxford University Press, Oxford.

Ladd, D.R., 2008. Intonational Phonology, second ed. Cambridge University Press, Cambridge, UK.

Lahiri, A., Fitzpatrick-Cole, J., 1999. Emphatic clitics in Bengali. In: René Kager, R., Zonneveld, W. (Eds.), Phrasal Phonology. Foris, Dordrecht, pp. 119–144.

Leben, W.R., 1978. The representation of tone. In: Fromkin, V. (Ed.), Tone: A Linguistic Survey. Academic Press, pp. 177–219.

Lee, C., Gordon, M., Büring, M., 2007. Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation. Springer, Dordrecht, The Netherlands.
Lehiste, I., 1970. Suprasegmentals. MIT Press, Cambridge, MA.
Lu, Y., Cooke, M., 2009. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Communication 51, 1253–1262.
Maddieson, I., 1978. Universals of tone. In: Greenberg, J. (Ed.), Universals of Human Language, Phonology, vol. 2. Stanford University Press, Stanford, CA, pp. 335–365.
McKendry, I., 2013. Tonal Association, Prominence and Prosodic Structure in South-eastern Nochixtlán Mixtec. Ph.D. dissertation, University of Edinburgh.
Nespor, M., Vogel, I., 1986. Prosodic Phonology. Foris, Dordrecht, The Netherlands.
Niebuhr, O., 2007. The signalling of German rising-falling intonation categories – the interplay; of synchronization, shape, and height. Phonetica 64, 174–193.
Ohala, J.J., 1994. The frequency code underlies the sound-symbolic use of voice pitch. In: Hinton, L., Nichols, J., Ohala, J.J. (Eds.), Sound Symbolism. Cambridge University Press, Cambridge, UK, pp. 325–347.
Ortega-Llebaria, M., Prieto, P., 2010. Acoustic correlates of stress in Central Catalan and Castilian Spanish. Language and Speech 54, 73–97.
Pannekamp, A., Toepel, U., Alter, K., Hahne, A., Friederici, A.D., 2005. Prosody-driven sentence processing: an event-related brain potential study. Journal of Cognitive Neuroscience 17, 407–421.
Peperkamp, S., Dupoux, E., 2002. A typological study of stress 'deafness'. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology, vol. 7. Mouton de Gruyter, Berlin, pp. 203–240.
Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English Intonation. PhD dissertation. MIT.
Pierrehumbert, J., Beckman, M.E., 1988. Japanese Tone Structure. MIT Press, Cambridge, MA.
Riad, T., 2013. The Phonology of Swedish. Oxford University Press, Oxford.
Rietveld, T., Gussenhoven, C., 1990. Preboundary lengthening and intonational and prosodic constituents. Journal of the Acoustical Society of America 88 (S1), S130.
Schmerling, S.F., 1976. Aspects of English Sentence Stress. University of Texas Press, Austin.
Selkirk, E.O., 1986. Phonology and Syntax: The Relation between Sound and Structure. MIT Press, Cambridge, MA.
Selkirk, E., 1995. Sentence prosody: intonation, stress and phrasing. In: Goldsmith, J. (Ed.), The Handbook of Phonological Theory. Blackwell, Cambridge, MA, pp. 550–569. ISBN: 9780631180623.
Selkirk, E., 2011. The syntax-phonology interface. In: Goldsmith, J., Riggle, J., Yu, A. (Eds.), The Handbook of Phonological Theory, second ed. Blackwell, Oxford, pp. 535–549.
Sluijter, A.M.C., van Heuven, V.J., Pacilly, J.J.A., 1997. Spectral balance as a cue in the perception of linguistic stress. Journal of the Acoustical Society of America 101, 503–513.
Smiljanić, R., 2002. Lexical, Pragmatic and Positional Effects on Prosody in Two Dialects of Croatian and Serbian: an Acoustic Study (Ph.D. dissertation), University of Illinois at Urbana-Champaign.
Suomi, K., Toivanen, J., Ylitalo, R., 2008. Finnish Sound Structure: Phonetics, Phonology, Phonotactics and Prosody. University of Oulu.
Swerts, M., Krahmer, E., Avesani, C., 2002. Prosodic marking of information status in Dutch and Italian: a comparative analysis. Journal of Phonetics 30, 629–654.
Vallduví, E., 1992. The Informational Component. Garland, New York.
Wightman, C., Shattuck-Hufnagel, S., Ostendorf, S., Price, P., 1992. Segmental durations in the vicinity of prosodic boundaries. Journal of the Acoustical Society of America 91, 1707–1717.
Xu, Y., Chen, S.-W., Wang, B., 2012. Prosodic focus with and without post-focus compression (PFC): a typological divide within the same language family? The Linguistic Review 29, 131–147.
Yip, M., 2002. Tone. Cambridge University Press, Cambridge.
Zerbian, S., 2007. Investigating prosodic focus marking in Northern Sotho. In: Aboh, E.O., Hartmann, K., Hartmann, M. (Eds.), Focus Strategies in African Languages: The Interaction between Focus and Grammar in Niger-Congo and Afro-Asiatic. Walter de Gruyter, Berlin, pp. 55–79.

Supreme Courts

Charles R Epp, University of Kansas, Lawrence, KS, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes multiple features of supreme courts in common-law and civil-law countries: their characteristics, decision-making processes and dynamics, and relations to other governmental institutions and important organized interests and parties in society. The survey is limited to courts of final appeal that take cases from lower courts within national judicial systems (thus, constitutional courts that take cases only on referral from other governmental branches are not included in this definition). Particular, but not exclusive, attention is devoted to the growing policy-making role of many supreme courts in recent decades, to significant limitations on judicial power, and to mechanisms used by governments in many countries to compel their supreme courts to defer to governmental policies.

Supreme courts are courts of final appeal within national judicial systems. Civil-law countries have, effectively, three (or more) separate court systems covering different areas of the law; there are separate institutions at the head of each, of which only the supreme civil court is called a 'court.' Federal systems commonly have, along with a national supreme court, separate supreme courts for their states or provinces. The development of supranational courts (e.g., the European Court of Justice) further complicates the matter, as these have become the courts of final appeal on some matters. Among these many institutions that arguably function in one way or another as 'supreme courts,' this article focuses on national, not supranational or subnational, courts, and on courts whose cases arrive principally by way of appeal from lower courts.

Over the past 30 years, scholarly interest in these courts has grown significantly as globalization, democratization, human rights movements, and a weakening of centralized governmental power have encouraged many supreme courts, in a wide array of countries, to intervene increasingly frequently and actively in the policy process (Alter, 2001; Domingo and Sieder, 2001; Epp, 1998; Helmke and Rosenbluth, 2009; Henkin and Rosenthal, 1990; Keller and Stone Sweet, 2008; Tate and Vallinder, 1995). For this reason, many observers have contended that strong and independent supreme courts play a key role in fostering the conditions for constitutional democracy, the rule of law, the protection of individual and minority rights, and economic development. These are heavy burdens to bear, and the capacity of supreme courts to achieve a wide range of policy goals is often exaggerated. Although supreme courts are not without significant influence, their power is constituted and thus also limited by their countries' particular political, social, and economic structures. Nonetheless, the conditions favoring active policy intervention by supreme courts show no signs of abating, and scholarly attention to them is likely to increase.

Institutional Structure

The Main Types

Supreme courts take a wide variety of forms and are related to other political institutions in a variety of ways, but they nonetheless fall within two main types, reflecting the broad division between the common-law and civil-law traditions (see Legal Systems, Classification of; Common Law; Civil Law). In common-law countries, a single national supreme court has general jurisdiction over a wide range of cases in virtually all areas of the law arising in lower courts. These courts are recognized as *de facto* law-making bodies (through their development of case law), and traditionally have been valued for their roles in checking executive discretion and lending coherence and consistency to their country's case law. They exercise these roles almost exclusively in cases arising out of concrete disputes between two or more parties; the losing party in a lower court may appeal the decision to a superior court, and a small proportion of these appeals ultimately reaches the country's supreme court, where review, in theory, is confined to disputes over law (and not facts) (see Appeals: Legal). This model is found today in countries throughout the former British empire, particularly in the United States, Canada, Australia, India, Ghana, Kenya, Nigeria, as well as South Africa.

In some common-law countries, particularly the United States and countries following its model, the supreme court possesses the power of judicial review to strike down legislation as unconstitutional, but typically only incidentally to the resolution of disputes between two or more parties (a system known as 'concrete review') (see Judicial Review). In other common-law countries – those with parliamentary sovereignty based on the British model – the supreme court has no such power but nonetheless exercises relatively broad authority to check administrative discretion. Following the enormous symbolic influence of the US Supreme Court's attack on racial segregation and its creation of new individual rights in the 1950s and 1960s, some supreme courts in common-law countries, particularly Canada and India, shifted markedly from the British to the American model by increasingly exercising the power of judicial review, particularly in the area of individual rights (Epp, 1998; Knopff and Morton, 1992). Under the growing influence of European law, even the Appellate Committee of the House of Lords, Britain's supreme court, has been legislatively reformed as a new Supreme Court, and its decisions appear to be edging toward a frank exercise of judicial review, a trend that is likely to be reinforced as Britain's new bill of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.86115-2

rights gains authority (Blom-Cooper et al., 2009; Dickson, 2007).

Civil-law systems, by contrast, have been influenced by the French Revolution's efforts to limit the power of courts and to foreclose the development of an independent, powerful, and active judiciary headed by a single supreme court. Consequently, the supreme judicial power in many civil-law countries is divided into three (or more) distinct institutions, each with a specialized jurisdiction (Jacob et al., 1996; Merryman, 1985). Thus, in France, the ordinary judiciary, which is headed up by the Court of Cassation, traditionally has been strictly forbidden from exercising judicial review or in any other way challenging legislative or administrative acts. The value of some mechanism for checking administrative acts was quickly recognized, however, resulting in the creation of a separate administrative court, the Contentious Section of the Council of State (hereafter, the Council of State), which now functions as the supreme court for the administrative court system. With the growing interest in constitutionalism after World War II, many civil-law countries added a separate constitutional court with the power of judicial review over legislation, reflecting the belief that the ordinary supreme court was insufficiently statesmanlike, too conservative, or too tied to a prior regime to give meaningful effect to constitutional provisions (Cappelletti, 1989, p. 145; Stone Sweet, 2000) (see Constitutional Courts). In countries in which the con-stitutional court decides challenges to existing laws, that court and the ordinary supreme court have at times developed sharply conflicting interpretations of the validity of particular laws, which has led to significant tensions between these courts (see, e.g., Krug, 1997). The civil-law model – with multiple courts of last resort – has been adopted throughout much of Europe and in many parts of Africa, the Middle East, and Latin America.

The strength of the pressures favoring active supreme court participation in the policy process is evident in the history of the French system. The French Court of Cassation originated as a legislative tribunal with power only to quash incorrect judicial interpretations of law, but over time it adopted a forthright appellate function and eventually was formally converted into the ordinary court system's supreme court; it has since engaged in significant judicial policy making (Merryman, 1985) (for parallels in Britain, see Stevens, 1978). Similarly, although the Council of State is not a part of the regular judiciary, it has evolved to a position of independence within the administrative branch, adjudicates the legal claims of individuals against the state bureaucracy, exercises judicial review over administrative regulations (which constitute a substantial proportion of French law), and has served as a model for supreme courts elsewhere (Brown and Bell, 1993; Merryman, 1985). Although formally there is no doctrine of *stare decisis* (the rule of precedent) in civil-law systems, in practice the decisions of the Court of Cassation and the Council of State gain wide adherence in lower courts and constitute a body of 'jurisprudence' that is similar in many ways to case law (Merryman, 1996).

The institutional structures of supreme courts in Latin America mix elements of the US and French models (see, e.g., Garcia, 1998; Navia and Rios-Figueroa, 2005). Thus, most Latin American countries have a single supreme court with the power of judicial review and with general jurisdiction over appeals from a variety of ordinary courts, but many deny to the inferior courts the power of judicial review. Even the supreme court's power of judicial review in most of these countries is limited to suspending the application of a law in a particular case, not invalidating the law itself. Some countries in the region follow the French model by denying the supreme court jurisdiction over the administrative courts; in others, the supreme courts hear only constitutional issues that are referred by the legislature. Judicial reforms of the 1980s and 1990s moved some Latin American countries closer still to the US model. Thus, in a 1994 reform the Mexican Supreme Court gained the power of judicial review, its justices gained a 15-year tenure, and some of the Court's decisions now have the force of precedent (Rios-Figueroa and Taylor, 2006).

Agenda Setting, Workload, and Internal Structure

The workload of many supreme courts has increased tremendously in recent years, providing significant opportunities for creative judicial policy making. To better manage rising caseloads, some countries have created intermediate appellate courts and have granted their supreme court significant discretion over which cases to decide. Such reforms acknowledge and enhance the policy-making role of the supreme court, and courts that have this discretion typically use it to expand their policy-making activities by building a coherent agenda that commonly has focused on constitutional issues (Pacelle, 1991). Granting the supreme court discretion over its judicial agenda is more typical of common-law than civil-law systems, as appellate review traditionally has been viewed in the former as an extraordinary resort, whereas in the latter it is viewed as a regular aspect of the judicial process (Damaska, 1986; Shapiro, 1981). Countries that have not granted discretion over case selection to their supreme courts – like India and France – are virtually forced by rising caseloads to expand the number of judges on the court; this is often accompanied by a division of the court into several panels (or 'chambers' or 'benches') specializing in different areas of the law.

Supreme courts differ, too, in how decisions are issued. The decisions of common-law supreme courts typically report significant disagreements among the deciding justices. Some, like the supreme courts of Canada and the United States, issue a single majority opinion along with dissenting opinions; others report separate (*seriatim*) opinions written by each of the justices. Civil-law courts, by contrast, commonly present a public face of consensus that conceals politically motivated disputes among the judges. Generally important decisions are published, but this varies too.

These various internal structures provide senior supreme court justices with different opportunities and constraints for strategic influence. Thus, in courts that are divided into separate panels, the chief justice may influence outcomes in particular cases through the exercise of authority to assign justices and cases to particular panels. In courts that meet as a whole, as in the United States, the chief justice or the senior justice in the majority similarly may influence case outcomes by exercising authority to assign the responsibility for writing the majority opinion (see, e.g., Epstein and Knight, 1998, pp. 125–135).

Judges may also manipulate the timing and publication of decisions for strategic effect. Thus, in some cases Israel's High Court of Justice has influenced the Israeli government's anti-terrorism policies by communicating behind-the-scenes threats to rule against the government unless policy is changed; some progovernment decisions in response to changed policy thus mask a more complex behind-the-scenes interaction (Hofnung and Margel, 2010).

Supreme Courts in the Political Process

Political Influences on Supreme Courts and Their Legitimacy

Supreme courts, while forthrightly legal institutions, are influenced by both legal norms and the political process, but the ways in which they are political, and the nature and paths of political influence, vary considerably. Although countries have adopted a wide array of constitutional and statutory provisions aimed at striking a balance between judicial independence from undue political manipulation and judicial responsiveness to widespread popular values, the effects of these measures are strongly shaped by several key factors related to the broader structure and distribution of political and organizational powers in each society.

First, the more centralized governmental power is, particularly in the executive, the greater the possible degree of direct manipulation of the supreme court by officials in these other institutions is. At one extreme, military coups or the exercise of emergency powers by governing regimes are commonly associated with sharp limitations on the authority of a country's supreme court and direct manipulation of the court by the regime (Halliday et al., 2012; Helmke, 2005; Tate, 1993). At the other extreme, in countries in which governing power is divided by a federal structure, an American-style separation of powers, or both, the paths of influence on the supreme court are more attenuated, affording the supreme court greater room for maneuver (Helmke and Rosenbluth, 2009). Between these ends of the continuum lie many intermediate instances. Thus, the supreme courts of Latin America vary considerably in degree of control by the executive and level of judicial intervention in key policy disputes (Helmke, 2005; Hilbink, 2007; Couso et al., 2010). Nonetheless, even supreme courts in pluralistic systems – for example, the United States and Canada – are influenced by the national governing coalition and may be active in many policy areas yet rarely challenge directly the key policies favored by that coalition (Whittington, 2005).

Second, political influences on supreme courts are conditioned by the degree of pluralistic conflict over selection of justices and the degree of diversity in the pool from which justices are drawn (Malleson and Russell, 2006). At one extreme, in a pattern more typical of common-law than civil-law countries, are courts whose appointment processes are subject to open political conflict between competing factions and whose justices are drawn from a relatively diverse population that includes politicians, government officials, and law professors. Thus, in the United States, the constitutional requirement for Senate confirmation of nominations made by the president has afforded the opportunity for open, pluralistic conflict over appointments to the Supreme Court (Yalof, 1999). At the other extreme are systems that insulate appointment processes from popular pressure and that draw judges from an insular elite, a pattern typical of civil-law countries. In some civil-law countries, the judiciary is largely self-regulating; judges are selected by competitive examination after receiving specialized legal training, and they advance through the judicial hierarchy on the basis of seniority and the approval of superior judges. In other civil-law countries, the executive dominates judicial recruitment and promotion to the supreme courts. Often, the two factors are found together, resulting in highly insular supreme courts.

For instance, in Chile the Supreme Court annually evaluates and ranks all of the country's judges, and appointments to the Court are made by the executive from a short list of nominees selected by the Court itself from among the lower judiciary; the system discourages the development of diversity of opinion and creativity within the judicial ranks, and, as a consequence, the Court has remained very quiescent over most of its history (Hilbink, 2007). Although the differences between civil-law and common-law patterns of appointment to supreme courts thus loom large, the dichotomy does not account for all variations. In the common-law world, the process of appointment to the US Supreme Court is notoriously open and conflictual, while for the British higher judiciary it is highly insular. In France, judges on the Court of Cassation are drawn exclusively from inferior courts; by contrast, members of the Council of State are drawn not only from inferior administrative tribunals but also from the administrative bureaucracy and academia (Brown and Bell, 1993, pp. 78–80). In general, the insularity of the recruitment process in continental Europe appears to be gradually eroding, with potentially significant implications for the judicial role.

Third, the degree of popular legitimacy of supreme courts and the nature and diversity of external coalitions of support for them place constraints on direct manipulation of these courts by other governmental officials. Some supreme courts have gained significant popular legitimacy, while others remain relatively unsupported. Support from the organized bar and other groups in civil society has often played a critical role in building these coalitions of support (Epp, 1998; Halliday et al., 2012). Even in authoritarian regimes, supreme courts have sometimes gained considerable support and leverage from broad external coalitions of popular support (see, e.g., Ginsburg and Moustafa, 2008; Lopez-Ayllon and Fix-Fierro, 2003). Judicial legitimacy, ironically, may be cultivated by courts themselves through active intervention in the policy process that attracts external coalitions of support (Gibson et al., 1998). Where coalitions of support are diverse and politically strong, they have played key roles in protecting supreme courts against measures aimed at curbing their power. This is clear, for instance, in comparing the limited success of attacks on the US Supreme Court prior to 1937 (Ross, 1994) with the Gandhi government's relatively greater influence over the Supreme Court of India in the 1970s (Baxi, 1980; Epp, 1998). In general, judges in common-law countries tend to be accorded relatively high status and respect, while the civil-law tradition has cultivated a deeply entrenched distrust for the judicial role and for judges

(Merryman, 1996). But beginning in the 1990s, many Latin American countries experienced growth in support for active judiciaries and enacted reforms to enhance judicial independence and power (Couso et al., 2010).

Although the various factors just summarized are conceptually distinct, they are mutually reinforcing in practice. Governing systems in which political power is decentralized have tended to accord greater symbolic status to their judges, particularly to their supreme court justices; develop pluralistic and even conflictual appointment processes that select justices from a diverse population; and cultivate broad coalitions of support around their supreme courts. Moreover, their supreme courts have tended to be relatively actively engaged in the policy process. By contrast, governing systems in which political power is centralized have tended to accord their judges, even their supreme court justices, the lowly status of legal bureaucrats and to select these justices through an executive-dominated process from highly insular pools of legal professionals. As a consequence, these courts have developed highly formalistic decision processes that yield politically acquiescent decisions and rarely cultivate support in the broader population. The latter extreme is exemplified by the insular judicial bureaucracies of some Latin American countries that have few sources of popular support – a pattern that is surely compounded by the effects of economic inequality on the degree of organizational pluralism in civil society (see, e.g., Hilbink, 2007).

The Growth in Policy Intervention by Supreme Courts, and the Limits of Their Power

Supreme courts, particularly in governing systems in which political power is not highly centralized, have increasingly intervened in the policy process in recent decades, particularly with regard to individual rights, environmental protection, and social welfare. The literature on these developments is large and growing (see, e.g., Couso et al., 2010; Dickson, 2007; Epp, 1998; Ginsburg and Moustafa, 2008; Hirschl, 2004; Tate and Vallinder, 1995; Widner, 2001).

Several partially conflicting strains have developed in these studies. One is over the degree to which high courts should be viewed as frankly political institutions. On this topic, some scholars have argued that supreme court judges are straightforwardly political. The 'attitudinal model' posits that justices' decisions directly reflect their political attitudes; this theory, although originally based on studies of the US Supreme Court (Segal and Spaeth, 1993), has been applied elsewhere as well (Morton et al., 1995). A closely related theory posits that justices' decisions are strategic, meaning that justices try to enact their political preferences but must modify, check, or restrain these desires in light of the expected responses of other actors or institutions (see, e.g., Epstein and Knight, 1998; Maltzman et al., 2000). This, too, has proved fruitful in studies of judicial policy making worldwide (see, e.g., Helmke, 2005; Hirschl, 2004). Other scholars, by contrast, favoring a theory sometimes labeled 'new institutionalism,' posit that justices' decisions are influenced by institutions, understood as organizational norms, rules, and structures, both among formal governmental structures and in the structure of civil society (see, e.g., Clayton and Gillman, 1999). Thus, legal assumptions establish the foundations and boundaries of judicial decision making and thereby constrain judicial decisions (Bartels, 2009; Dickson, 2007; Gillman, 1993); and the institutional relationship of the supreme court, the executive, and the legislature shapes the range of discretion available to the court (Navia and Rios-Figueroa, 2005). This institutionalist perspective appears in many studies of supreme court decisions worldwide (see, e.g., Couso et al., 2010; Epp, 1998; Ginsburg and Moustafa, 2008; Perry, 1991).

The extent to which supreme courts are capable of effecting substantial changes in governmental policy, and whether the policy effects are mainly egalitarian or antiegalitarian, remains a matter of dispute. It is clear that supreme courts have no enforcement authority of their own and thus depend for implementation of their decisions on the actions of public officials, private individuals, and organizations. On this basis, Gerald Rosenberg (1991) has argued that policy reform decisions of the US Supreme Court have led to actual reforms only in very limited circumstances, particularly when they are aided by sympathetic officials or a sympathetic population, or are subject to implementation by market forces. Michael McCann (1994), however, has shown that judicial decisions provide strategic resources that may be used by social movements, organized groups, and public officials to substantial effect (albeit often in ways unanticipated by the judiciary) (see Law, Mobilization of). These competing perspectives share the core observation that supreme courts depend on external actors to carry out their decisions, and this support varies. In the absence of organized support, court decisions often have little force, but in its presence they may produce considerable policy change (see, e.g., Epp, 2009; Gauri and Brinks, 2008). Taken as a whole, the increasing policy intervention of supreme courts, and of courts in general, has contributed to the 'judicialization' or 'legalization' of policy and admin-istrative processes, but whether this has contributed to democratization or has benefited the least well off varies considerably (Brinks, 2008; Couso et al., 2010; Epp, 2009; Tate and Vallinder, 1995).

The State of Social Scientific Research on Supreme Courts

Research on supreme courts has advanced remarkably in the past 20 years. Scholars have made careful studies of a wide range of courts that previously had been ignored and have developed sophisticated quantitative databases of decision-making patterns in some of these courts. Additionally, research agendas are increasingly based on social scientific theories and thus increasingly attempt to explain rather than merely describe. Yet important gaps remain. Most of the existing social scientific research on supreme courts has focused on a handful of highly active courts, while many other courts, particularly the supreme courts of cassation in the civil-law world, have been largely ignored. Most of the research has focused on the conditions shaping supreme court decisions; little has been done on the policy consequences of these decisions, even though it is widely recognized that decisions are often implemented only in part, if at all. Finally, although research on

supreme courts in many parts of the world has expanded, only a few truly comparative studies have been done.

Nonetheless, our understanding of supreme courts has grown. The body of social scientific research suggests both that supreme courts may contribute to the broader goals that many observers in recent years have pinned on them – the protection of individual and minority rights and the cultivation of constitutional law more generally – and that the extent to which they support these goals is conditioned by a range of factors, particularly the degree to which political power in the governing system is centralized and the degree to which these courts enjoy popular legitimacy and coalitions of support among civil society organizations.

See also: Civil Law; Constitutional Courts; Courts and Adjudication; Judicial Review.

Bibliography

Alter, K.J., 2001. Establishing the Supremacy of European Law: The Making of an International Rule of Law in Europe. Oxford University Press, Oxford, UK.

Bartels, B.L., 2009. The constraining capacity of legal doctrine on the U.S. Supreme Court. American Political Science Review 103, 474–495.

Baxi, U., 1980. The Indian Supreme Court and Politics. Eastern Book Company, Lucknow, India.

Blom-Cooper, L., Dickson, B., Drewry, G., 2009. The Judicial House of Lords, 1876–2009. Oxford University Press, Oxford, UK.

Brinks, D., 2008. The Judicial Response to Police Killings in Latin America: Inequality and the Rule of Law. Cambridge University Press, New York.

Brown, L.N., Bell, J.S., 1993. French Administrative Law, fourth ed. Clarendon, Oxford, UK.

Cappelletti, M., 1989. The Judicial Process in Comparative Perspective. Clarendon, Oxford, UK.

Clayton, C.W., Gillman, H., 1999. Supreme Court Decision-Making: New Institutionalist Approaches. University of Chicago Press, Chicago.

Couso, J.A., Huneeus, A., Sieder, R. (Eds.), 2010. Cultures of Legality: Judicialization and Political Activism in Latin America. Cambridge University Press, New York.

Damaska, M.R., 1986. The Faces of Justice and State Authority: A Comparative Approach to the Legal Process. Yale University Press, New Haven, CT.

Dickson, B. (Ed.), 2007. Judicial Activism in Common Law Supreme Courts. Oxford University Press, Oxford.

Domingo, P., Sieder, R (Eds.), 2001. Rule of Law in Latin America: The International Promotion of Judicial Reform. Institute of Latin American Studies, London.

Epstein, L., Knight, J., 1998. The Choices Justices Make. CQ Press, Washington, DC.

Epp, C.R., 1998. The Rights Revolution: Lawyers, Activists, and Supreme Courts in Comparative Perspective. University of Chicago Press, Chicago, IL.

Epp, C.R., 2009. Making Rights Real: Lawyers, Activists and the Creation of the Legalistic State. University of Chicago Press, Chicago, IL.

Garcia, F.S., 1998. The nature of judicial reform in Latin America and some strategic considerations. American University International Law Review 13, 1267–1325.

Gauri, V., Brinks, D. (Eds.), 2008. Courting Social Justice: Judicial Enforcement of Social and Economic Rights in the Developing World. Cambridge University Press, New York.

Gibson, J.L., Caldeira, G.A., Baird, V.A., 1998. On the legitimacy of national high courts. American Political Science Review 92, 343–358.

Gillman, H., 1993. The Constitution Besieged: The Rise and Demise of Lochner Era Police Powers Jurisprudence. Duke University Press, Durham, NC.

Ginsburg, T., Moustafa, T. (Eds.), 2008. Rule by Law: The Politics of Courts in Authoritarian Regimes. Cambridge University Press, New York.

Halliday, T.C., Karpik, L., Feeley, M. (Eds.), 2012. The Fates of Political Liberalism in the British Post-colony. Cambridge University Press, New York.

Helmke, G., 2005. Courts under Constraints: Judges, Generals and Presidents in Argentina. Cambridge University Press, New York.

Helmke, G., Rosenbluth, F., 2009. Regimes and the rule of law: judicial independence in comparative perspective. Annual Review of Political Science 12, 345–366.

Henkin, L., Rosenthal, A., 1990. Constitutionalism and Rights: The Influence of the United States. Constitution Abroad. Columbia University Press, New York.

Hilbink, L., 2007. Democracy and Dictatorship: Lessons from Chile. Cambridge University Press, New York.

Hirschl, R., 2004. Towards Juristocracy: The Origins and Consequences of the New Constitutionalism. Harvard University Press, Cambridge, MA.

Hofnung, M., Margel, K.W., 2010. Judicial setbacks, material gains: terror litigation at the Israeli High Court of Justice. Journal of Empirical Legal Studies 7, 664–692.

Jacob, H., Blankenburg, E., Kritzer, H.M., Provine, D.M., Sanders, J., 1996. Courts, Law and Politics in Comparative Perspective. Yale University Press, New Haven, CT.

Keller, H., Stone Sweet, A. (Eds.), 2008. A Europe of Rights: The Impact of the ECHR on National Legal Systems. Oxford University Press, New York.

Knopff, R., Morton, F.L., 1992. Charter Politics. Nelson Canada, Scarborough, Canada.

Krug, P., 1997. Departure from centralized model: the Russian Supreme Court and constitutional control of legislation. Virginia Journal of International Law 37, 725–787.

Lopez-Ayllon, S., Fix-Fierro, H.F., 2003. "Faraway, so close!" The rule of law and legal change in Mexico 1970–2000. In: Friedman, Lawrence, Perez Perdomo, Rogelio (Eds.), Legal Culture in the Age of Globalization: Latin America and Latin Europe. Stanford University Press, Stanford, CA.

Malleson, K., Russell, P.H. (Eds.), 2006. Appointing Judges in an Age of Judicial Power: Critical Perspectives from around the World. University of Toronto Press, Toronto, Canada.

Maltzman, F., Spriggs, J.F., Wahlbeck, P.J., 2000. Crafting Law on the Supreme Court: The Collegial Game. Cambridge University Press, New York.

McCann, M., 1994. Rights at Work: Pay Equity Reform and the Politics of Legal Mobilization. University of Chicago Press, Chicago, IL.

Merryman, J.H., 1985. The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America, second ed. Stanford University Press, Stanford, CA.

Merryman, J.H., 1996. The French deviation. American Journal of Comparative Law 44, 109–119.

Morton, F.L., Russell, P.H., Riddell, T., 1995. The Canadian charter of rights and freedoms: a descriptive analysis of the first decade, 1982–1992. National Journal of Constitutional Law 5, 1–60.

Navia, P., Rios-Figueroa, J., 2005. The constitutional adjudication mosaic of Latin America. Comparative Political Studies 38, 189–216.

Pacelle Jr., R.L., 1991. The Transformation of the Supreme Court's Agenda: From the New Deal to the Reagan Administration. Westview Press, Boulder, CO.

Perry, H.W., 1991. Deciding to Decide: Agenda Setting in the United States Supreme Court. Harvard University Press, Cambridge, MA.

Rios-Figueroa, J., Taylor, M., 2006. Institutional determinants of judicialization of policy in Brazil and Mexico. Journal of Latin American Studies 38, 739–766.

Rosenberg, G.N., 1991. The Hollow Hope: Can Courts Bring about Social Change? University of Chicago Press, Chicago, IL.

Ross, W.G., 1994. A Muted Fury: Populists, Progressives, and Labor Unions Confront the Courts, 1890–1937. Princeton University Press, Princeton, NJ.

Segal, J.A., Spaeth, H.J., 1993. The Supreme Court and the Attitudinal Model. Cambridge University Press, New York.

Shapiro, M.M., 1981. Courts: A Comparative and Political Analysis. University of Chicago Press, Chicago, IL.

Stevens, R.B., 1978. Law and Politics: The House of Lords as a Judicial Body, 1800–1976. University of North Carolina Press, Chapel Hill, NC.

Stone Sweet, A., 2000. Governing with Judges: Constitutional Politics in Europe. Oxford University Press, New York.

Tate, C.N., 1993. Courts and crisis regimes: a theory sketch with Asian case examples. Political Research Quarterly 46, 311–338.

Tate, C.N., Vallinder, T. (Eds.), 1995. The Global Expansion of Judicial Power. New York University Press, New York.

Whittington, K.E., 2005. "Interpose your friendly hand": political supports for the exercise of judicial review by the United States Supreme Court. American Political Science Review 99, 583–596.

Widner, J.A., 2001. Building the Rule of Law. W.W. Norton, New York.

Yalof, D.A., 1999. Pursuit of Justices: Presidential Politics and the Selection of Supreme Court Nominees. University of Chicago Press, Chicago, IL.

Surveillance and Privacy, Geography of

Till F Paasche, Soran University, Soran City, Kurdish Region Iraq
Francisco R Klauser, Institut de Géographie, Université de Neuchâtel, Neuchâtel, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Surveillance and privacy are mutually exclusive: if one increases, the other decreases. After defining these terms and their relation to each other, this article introduces key concepts in the field of surveillance studies, followed by a discussion of classical forms of surveillance and privacy invasion, namely, forms of visual surveillance. The second half of the article departs from the surveillance of individuals to new technological trends that represent new challenges to privacy concerns and the social sorting of populations through software algorithms.

Surveillance

Surveillance is a complex issue that has sparked a number of definitions. Because we have to start somewhere, we defer to two intellectual authorities in the field, Lyon (2007) and Marx (2012), who define surveillance in general terms outside the risk and security discourse in which it is often applied. Lyon (2007, p. 14) defines surveillance as follows:

> (...) the focused, systematic and routine attention to personal details for purposes of influence, management, protection or direction. Surveillance directs its attentions in the end to individuals (even though aggregated data, such as those available in the public domain, may be used to build up a background picture). It is focused. By systematic, I mean that this attention to personal details is not random, occasional or spontaneous; it is deliberate and depends on certain protocols and techniques.

Marx applies a similar individually focused angle to his definition:

> At the most general level surveillance of human (which is often, but need not, be synonymous with human surveillance) can be defined as regard or attendance to a person or to factors presumed to be associated with a person.
>
> Marx, 2012, p. xxv

While outlining various more detailed subfields of surveillance, Marx maintains a relatively strong focus on one individual or group attending to another individual or group. Thus,

> some common classificatory notion can be applied. In the case of surveillance social structures, for example, we can identify the *surveillance agent* (weather as watcher/observer/seeker/inspector/auditor/tester), while the person about whom information is sought or reported is the *surveillance subject*.
>
> Marx, 2012, p. xxv, emphasis in original

Inherently linked to these definitions is privacy, which we will discuss in the context of surveillance. The two terms can almost be understood as antipodes. The more (or less) surveillance there is, the less (or more) privacy there is. Thus, in an argument about omnipresent surveillance, we commonly speak about the death of privacy.

Privacy

Privacy is commonly seen as a good that needs to be protected against increasing surveillance activities. From this position, privacy is often defined in terms of access to information and control over this access. If surveillance is access to individual or personal information, privacy is the control over this access to information (Rössler, 2001). If we can decide who has access to what information about ourselves, we have full privacy. However, Rössler (2001), a key figure in privacy studies, argues that privacy is not only a desirable good but also a precondition for liberal democracies in which the autonomy to decide over one's own life according to one's abilities is of fundamental importance. To guarantee autonomy, a degree of privacy is needed to protect the individual. If, for example, an employer knew all the private details of an employee's life, such as mistakes in the past, diseases that do not impact the job, and relatives with criminal records, these details would likely influence employment and thus decrease autonomy. Nonetheless, as Rössler acknowledges, privacy in its purest form, i.e., complete control over all personal information, is impossible to achieve given participation in contemporary societies. The state, employers, banks, etc. will always possess some personal information about an individual that is beyond his/her control. Discussing surveillance and privacy is thus a question of balance; how much surveillance is needed to have, For example a functioning government able to provide security, and when is the invasion of the state into the privacy of its citizens unnecessary? Because many governments and governance bodies have a tendency to collect an abundance of data on their populations and thus tip this balance in one direction (Foucault, 2007), different parts of civil society, as well as many of the studies on surveillance and privacy, engage critically with the topic of increasing surveillance and its negative implications on privacy. However, as mentioned, this article approaches privacy in relation to surveillance. Therefore, we only refer to one of what Rössler (2001) calls the three dimensions of privacy. The first dimension, decisional privacy, is concerned, for example, with the freedom to decide what

clothes I put on in the morning. The dimension discussed in this article, informational privacy, is concerned with issues related to access to personal information. The last dimension, local privacy, is concerned with the ability to retreat into personal space, such as one's own home. Elements of this last dimension play a role when discussing visual surveillance in the form of drones.

Surveillance and Privacy Studies

In recent years, a rapidly developing international body of research has sought to explore the various causes, implications, and problems of the contemporary proliferation and intensification of surveillance. The emerging cross-disciplinary field of 'surveillance studies' (Lyon, 2007) has raised a number of critical issues with regard to the intensification and generalization of data gathering, transfer, and analysis in the information age. The resulting critical debates about surveillance fit largely into two complementary and often interrelated categories. The first category – predominant in sociology and urban studies – has sought to challenge the use of surveillance as a powerful tool for social sorting (Lyon, 2003) and is related to concerns of social exclusion and spatial justice (Coleman, 2004). The second category, traditionally predominant in regulatory approaches, has focused mainly on surveillance-related privacy and data protection issues.

In this article, we will touch on both categories. Following an introduction of the concepts and theories that explain surveillance, we introduce surveillance in its classical, human-focused form, drawing on the example of visual surveillance. We then shift our attention to the monitoring of the nonhuman, the foremost of which is urban aspects. As surveillance studies have shown, ubiquitous computing embodied by the various smart campaigns, can have a significant impact on the private even though the focus is not directly on the individual. We outline the social consequences of the various surveillance approaches in our final section.

Competing Concepts

There are different approaches to theorizing surveillance in its multiple facets, applications, and effects. Space does not permit an exhaustive review of the existing approaches, but some of the most important conceptual perspectives on surveillance shall be outlined here to add a more theoretical dimension to our discussion of surveillance and privacy that follows (for a more exhaustive discussion, see, for example, Lyon, 2006).

Surveillance scholars have found a range of conceptual tools in Michel Foucault's theorization of power and governmentality for dealing with the dynamics inherent in contemporary forms of control and regulation. Most notably, Foucault's analysis of disciplinary power, exemplified in his discussion of the Panopticon in *Discipline and Punish* (Foucault, 1977), has inspired a number of empirically and conceptually informed discussions on the normalizing prowess of novel surveillance technologies on individuals and social groups. For Foucault, the Panopticon exemplifies a particular economy of power that characterizes European modernity shaping, working through, and developing from a range of milieus such as hospitals, schools, army barracks, and prisons. The exercise of power in all of these settings implies a specific way of managing multiplicities through techniques of individualized surveillance and normalization.

However, while Foucault's Panopticon paradigm has been used prominently to problematize surveillance in its inherent power problems and implications, additional approaches have been mobilized to create a more conceptually and empirically accurate picture of the actual effects and limits of contemporary surveillance. For example, drawing on Deleuze's analysis of the contemporary shift from a Foucauldian 'disciplinary society' to a postmodern 'society of control' (Deleuze, 1992), various scholars have emphasized the changing modalities and functions of surveillance, from rigid and permanent monitoring and enclosure to more flexible and adaptable forms of regulation and control. Similar insights have been gained from Foucault's recently translated *Security, Territory, Population* lectures, given in 1977/1978 at the College de France (Foucault, 2007). Here, Foucault distinguishes between 'disciplinary' and 'security apparatuses' to capture the differing regulatory dynamics and mechanisms of power in the governing of men and things. A third source of inspiration in counterbalancing the Panopticon paradigm has been found in Michel de Certeau's work, who stresses the microtactics and strategies deployed by individuals and social groups to resist surveillance (De Certeau, 1984). On these grounds, the effects and implications of surveillance on social life (including privacy) have been found to be much more contingent and complex than initially expected.

While Foucault, Deleuze, and de Certeau allowed for the theorization of the effects and implications of surveillance, other theoretical sources have been mobilized to conceptualize the actual internal functioning, logic, and organization of specific surveillance systems. On a rather general level, approaches inspired by Marx and especially Giddens (1985) have portrayed (bureaucratic) surveillance as indeed a defining and constitutive feature of modern capitalist life. More specifically, rooted in an actor-network theory-based line of thinking, as developed by Bruno Latour and Michel Callon (Latour, 2005), various scholars have studied the complex actor networks underpinning and shaping the making of particular surveillance systems. Placing centrally the processes and relationships through which surveillance is conditioned and coproduced, this research has developed an understanding of surveillance in its complex sociotechnical associations and compositions that bring together, and are the result of, a wide range of actors, objects, instruments, intentions, and domains of expertise. This focus also resonates with the longstanding conceptual and empirical interest of Deuleuzian literature in the heterogeneous 'assemblages' of human and nonhuman entities that make up specific milieus or, more specifically speaking, surveillance systems (Deleuze and Guattari, 1987; Haggerty and Ericson, 2000). In terms of privacy, these studies have shown that the implications of surveillance depend on a variety of system-inherent factors, including not only its technical specificities but also resulting from the modalities, dynamic actor

networks, and interacting domains of expertise shaping the system 'in action.'

Visual Surveillance

CCTV

In this section, we discuss one of the most classic themes in surveillance and privacy studies, the CCTV (closed circuit television) camera, and some of its contemporary modifications, such as facial recognition software and drones. The first major boom in the use of CCTV cameras for surveillance and security purposes started in the late 1970s and continued into the 1980s. Particularly in the beginning, CCTV cameras in public space were hailed as a revolution in crime prevention, a claim that failed to manifest itself clearly in empirical data. Although the installation of surveillance cameras often results in a short-term reduction of crime rates, longer follow-up periods show that the cameras' long-term effects in deterring crime have to be interpreted more critically (Welsh and Farrington, 2002; Gill and Spriggs, 2005). Furthermore, a growing range of CCTV evaluations suggest that the efficiency of CCTV strongly depends on which type of crime is being analyzed. For example, CCTV is generally found to be effective in reducing vehicular crimes in car parks, whereas empirical research suggests that CCTV has little or no effect on vandalism and acts of aggression on public transportation systems and in city-center settings. The functioning and impacts of CCTV systems also depend on a variety of system-inherent factors, from collaborations between operators to technical dimensions of the cameras. Nonetheless, the number of cameras in various spaces continues to grow. In the United Kingdom, a country where the CCTV camera enjoys particular popularity, there are an estimated 1.85 m cameras ('The Guardian' http://www.guardian.co.uk/uk/2011/mar/02/cctv-cameras-watching-surveillance). Today, in many countries, high-resolution CCTV cameras can be bought in home improvement shops, and where they can be installed is only limited by national legislation.

Intelligent CCTV Systems

Throughout the 1980s and 1990s, CCTV systems were established and simply continued to exist more or less effectively; however, recent coded technologies have generated a new wave of excitement about CCTV cameras. The availability of large affordable digital storage facilities and fiber-optic cables has made cameras more effective for investigatory proposes and made intelligent networked CCTV systems possible. With continued research on increasingly sophisticated algorithms that can recognize faces, license plates, and even behavior patterns, cameras are becoming exponentially more effective (Graham, 2005). Where there used to be a human agent sitting in front of a wall of screens trying to detect suspicious individuals, code now automates this process. For example, by comparing faces to a database of terror suspects or registered hooligans, the algorithm can sound an alarm once a suspect is sighted or near a strategic facility. Although code is not yet capable of such sophisticated procedures, vast amounts of resources are being invested into the advancement of already existing recognition technologies.

Drones

Another technology that led camera-based surveillance into a new era is the unmanned aerial vehicle (UAV), commonly referred to as drones. UAVs further exemplify the close link between military and civilian surveillance technologies. During the War on Terror and various campaigns of counterinsurgency warfare, drones became increasingly popular and have undergone rapid development in recent years. UAVs are now integral to the warfare practices of technologically advanced armies. What started as surveillance tools soon became armed machines, as the latest versions of the Reaper drone show. Using drones in conflicts, intelligence can be collected and (extraterritorial) attacks carried out without putting pilots in danger (in fact, drone pilots can command their vehicles for a mission over Pakistan from their 'cockpits' in the United States). In their original function as surveillance tools that film or take images from high altitudes, UAVs thus complement satellites but are not bound to a single orbit. Like other developments originating in the military, UAVs soon caught the attention of law enforcement and other civilian agencies. As Graham argues,

> The crossover between the military and the civilian applications of advanced technology – between the surveillance and control of everyday life in Western cities and the prosecution of aggressive colonial and resource wars – is at the heart of a much broader set of trends that characterize the new military urbanism. (…) Fundamental to the new military urbanism is the paradigmatic shift that renders cities' communal and private spaces, as well as their infrastructure – along with their civilian populations – a source if targets and threats.
>
> Graham, 2010, p. xiii

With wars and conflicts being fought increasingly in urban areas, the line between military and civilian surveillance and other technologies is becoming blurred as technologies increasingly serve double functions. Today, UAVs come in all shapes and sizes, ranging from high-tech military drones with wingspans of well over 20 m to small surveillance UAVs that can land on the palm of a hand. Smaller UAVs are becoming increasingly popular with police forces all over the world for obtaining intelligence on armed suspects, groups of protesters, and sports fans or simply for monitoring deserted urban landscapes more efficiently. As mentioned above, UAVs constitute a quickly developing technological field whose development cannot be foreseen. Given their relative low costs and multifaceted functions for surveillance purposes, it is unlikely that drones will disappear from the CCTV/surveillance landscape soon. Combined with high-definition cameras and facial recognition software, drones are the latest high-tech advancement in a surveillance trend that started with the classic CCTV camera in the 1980s. Whereas the usual CCTV camera is stationary, although it may rotate, drones are highly mobile and able to survey many spaces previously unseen by CCTVs. In this context, local privacy, or the freedoms in one's own home, is under threat by the image of a drone hovering in front of a bedroom window (Rössler, 2001). As the technological possibilities increase exponentially, substantial further invasions of

privacy continue to be protected only by a critical civil society and the law, at least in Western democracies.

Computerization of Surveillance

Having outlined the more obvious forms of direct surveillance, we now focus on a more recent surveillance trend beyond the visual that focuses on the monitoring of the nonhuman. Today, we have reached an area of ubiquitous computing, where code penetrates most aspects of life and enables the management of urban environments, including both human and nonhuman objects. The logic of this approach of managing the urban environment is not new, although the first cases were human-oriented. The basic ideas of these current trends date back to the late nineteenth and early twentieth centuries and the first use of punch cards that enabled the structuring of relatively complex data on populations according to different characteristics. A similar logic was utilized by the German police in the late 1970s when *Rasterfahndung* (dragnet policing) was introduced in an attempt to identify members of the *Rote-Armee Fraktion* (Red Army Faction) and other left-wing terror groups. Using a state-of-the-art computer, various available sources of information, such as student loans, insurance, and registration data, were analyzed to obtain a manageable number of locations and suspects assumed to be typical for terrorists. In principle, the process included analyzing large data sets according to set parameters to generate a small number of individualized data.

However, only in recent years have computers and technologies such as GPS tracking and mobile phones become advanced and widespread enough to integrate algorithms into the functioning of everyday lives. Such advancements include RFID (radio-frequency identification) devices, which, by now, are ubiquitous in various trades and have become an unavoidable standard technology. RFID chips store information that is being transmitted via radio frequencies. If the chip comes close enough to a receiver from a few centimeters to hundreds of meters, the information can be read. RFID chips, some of which are not much larger than a grain of rice, can be found in access, food court and student cards, pets and livestock, shipping containers and other cargo, among other things. With a starting cost of just a few pence, the only reason RFID chips are not used even more extensively is data protection concerns, which are why chips have not been put in clothes (the washing machine would automatically know what temperature to use) or directly implemented in humans (this would make it impossible to forget ID cards and easier for police to identify individuals).

This increased computerization of life and dependenc on code led Dodge and Kitchin (2011) to formulate their code/space theory. Some years ago, code supported the functioning of space, as it eased daily operations. If the code broke away or malfunctioned, all operations could, although more slowly and less efficiently, continue. Although this coded space was supported by computing, it did not depend on it. In many instances, this has changed. Today, various spaces depend on code in a way that if the code breaks away, the whole space stops functioning. One example of such a code/space is the modern airport. If the various algorithms that enable check-in, baggage sorting, the functioning of the tower, or the functioning of the planes themselves stop working, the space is no longer an airport, but a chaotic agglomeration of hectic and panicking people. Dodge and Kitchin (2011) provide other examples of how we interact with more banal forms of code/space every day, including diaries in our smartphones, access to our e-mails, the engine of our car managed through the inboard computer, the electronic till of our favorite coffee place, and possibly even the entrance to our office building controlled by a swipe card or RFID chip. All of those examples are spaces that stop functioning in their designed purpose if the code breaks away.

Privacy and Contemporary Smart Surveillance

With the popularized smart city ambitions that are currently symptomatic for the management of urban environments, the idea of code/space is developed to a further extreme. In the context of the hyped smart city technologies, the linkages between increasing computerization and surveillances and privacy become clear. As explained with the example of punch cards and the *Rasterfahndung*, large data sets, are being combined and analyzed. In a process called data mining, linkages within data are made, and new information is extracted. Urban processes are managed on the basis of this new information, possibly resulting in the automation of parts of cities and buildings.

The smart city is currently one of the most prominent urban development technology and is being discussed around the world. Spearheaded by global IT companies, the smart city aims to optimize different aspects of urban space through algorithms. Targeted areas of cities include governance structures, water and electricity grids, traffic and transport, and the emergency services and police. Although part of different sectors, their automations all follow a similar logic, the analysis and interconnection of large quantities of data, including existing data being made available as well as the generation of new data either by digitizing analogous processes, such as electricity meters and sensors on highways or in public transport, or connecting different existing databases. Data are cumulated in vast databases, and once the data are centralized, algorithms can make the city, smart, by rendering the data usable for different optimization proposes, a process called data mining. Such optimization processes include more effective governance function as decisions can be based on the analysis of more data and thus be made faster, optimizing public and individual transport, for example in terms of travel time, or easing access to medical data.

Nonetheless, the data sets these mining algorithms analyze often comprise sensitive and private information. One example is the use of a smart grid, one component of the smart city. Smart grids are one technology experts see as the future of energy production and consumption; they are based on the analysis of the energy patterns of individual households through algorithms and, subsequently, the automation of some electronic appliances through code. The data collected in this process provide insights on individual habits and movement patterns (for example when one takes a shower, leaves the house, watches TV, uses electronic appliances).

The primary function of smart grids is the optimization of energy consumption and increases in alternative energy sources. Although these are honorable ambitions, to get there,

unrepresented amounts of private data are collected with increasing standardization of these technologies. Additionally, these data can potentially serve a secondary purpose; detailed personal information is of considerable value to the advertisement industry, which can create localized and personalized advertisements. This information could be of value to companies interested in the activities of their employees and police who would gain *Rasterfahndungs* capabilities that were unimaginable 30 years ago. Differing from the classical definition of surveillance of Lyon and Marx or the example of the CCTV camera that singles out suspicious individuals, the current trend of surveillance is to collect a variety of data on different aspects and to extract individualized information out of this data pool. Because they are not open or direct, these surveillance functions become blurred to the public eye and are ultimately trivialized by both the surveying agencies and the surveyed population.

Thus, the smart city exemplifies an increasing belief that cities cannot be managed without substantial invention by code and that the complexity of urban lives requires algorithms to regulate them. Indeed, there are arguments that underline the advantage of codes regulating and managing the urban space. However, these developments are rarely discussed with regard to surveillance, and the substantial invasion of privacy is of limited concern. In many contemporary examples, code/space issues of privacy and surveillance are deferred to the legal and institutionalized theme of data protection and detailed accounts on technical solutions against, for example hackers that want to invade privacy and/or abuse private data. What data are really needed in the first place is of lesser concern. Furthermore, with smart city ambitions, it is often automatically assumed that the state is immune to any form data abuse and thus can be trusted with vast amounts of very detailed personalized data, an assumption that was often contested in the past. In the concluding section, some of the implications of this trend are sketched out.

Social/Software Sorting

In this concluding section, we outline the consequences of surveillance, especially the impacts new coded forms of surveillance have on society and privacy. The increasing metamorphosis of space supported by code into code/space creates a situation where it is increasingly difficult, and in some parts impossible, to have control over personal information, who has access to this information, and how this information is being used. This difficultly is evident in our digital traces on the World Wide Web or mobile communication networks, the RFID chips in the cards in our wallets, and the sophisticated CCTV cameras that are watching us.

It is even more difficult for an individual to control his/her personal information with the spatial manifestations of smart cities and related visions based on data mining. By combining different information, code generates new data and information. Because we usually do not know the source information the data mining uses or the workings and capabilities of the code, we no longer have the chance to understand what knowledge about us exists. Linking this back to Rössler's privacy discourse, our autonomy is threatened by this development. In an environment of data mining and networked information, we cannot be certain what our opposite in a meeting might see on his/her screen about us; thus, we lose the ability to control our lives, and the chance of discrimination on the basis of private information grows concurrently. Privacy is increasingly no longer a right.

In addition to its implication on privacy, the computerization of space also enables an increasingly sophisticated sorting of populations along set parameters, as three examples outlined by Graham (2005), Lyon (2003), and Jones (2000) will show.

- Using code to control and regulate traffic flows: Utilizing software tools, road pricing becomes flexible and adaptable, different from the static vignette system some countries apply. Steering congestion through road pricing, the usage of streets during peak hours can become too costly for some at certain times. Although this is a specific example, it relates to the wider issue of managing space and populations, governing at a distance through code (Jones, 2000).
- Using codes to structure the city and manage police patrols: Using software such as geographic information systems (GIS), cities are becoming increasingly digitized. Attributing data to parts of the city, individual streets and so forth, space is being branded, identities are being created, or property prices are being influenced (Graham, 2005). A popular example is the online GIS-based crime mapping tool of the British police (http://www.police.uk/).
- Direct forms of coded exclusion and sorting: Code enables different form of exclusion. Although some of the following could theoretically be achieved, for example using ordinary keys, code eases processes and renders them useable, especially on a large scale. The most banal forms of exclusions include PIN codes, relatively cheap swipe cards, and using biometrics, such as fingerprints or iris scanners, to allow access to buildings. Here different levels of access can be allocated to individuals. The same can be applied to the virtual worlds of inter- and intranets. With the increasing importance of codes for access, the rejection of access privileges can be used in a disciplining way, punishing behavior.

This list could continue. However, the aim is not to produce a complete picture, but to indicate the issues that will arise with an increasing codification of live and space and rapidly advancing computing possibilities. However, the above is not a phenomenon that has emerged because of new possibilities in surveillance, but is rather a reflection of the long-existing desire of various agencies to survey and to manage the surveyed.

> If surveillance as social sorting is growing, this is not merely because some new devices have become available. Rather, the devices are sought because of the increasing number of perceived and actual risks and the desire more completely to manage populations – whether those populations are citizens, employees, or consumers.
>
> Lyon, 2007, p. 20

New code-based technologies support this desire for surveillance and data collection. What gives the development

a novel spin is that the omnipresence of a code, very different from CCTV cameras, begins to blur the line between surveillance and the invasion of privacy. This is especially true when the focus is not safety and security, but the improvement of urban space and the quality of life.

See also: Policing; Privacy: Theoretical and Legal Issues; Smart Cities; Surveillance Studies; Technology and Social Control.

Bibliography

Coleman, R., 2004. Reclaiming the streets: closed circuit television, neoliberalism and the mystification of social divisions in Liverpool, UK. Surveillance and Society 2, 145–160.

De Certeau, M., 1984. Practice of Everyday Life. University of California Press, Berkeley.

Deleuze, G., October 1992. Postscript on the societies of control, vol. 59. MIT Press, Cambridge, MA, 3–7.

Deleuze, G., Guattari, F., 1987. A Thousand Plateaus. University of Minnesota Press, Minneapolis.

Dodge, M., Kitchin, R., 2011. Code/Space. Software and Everyday Life. MIT Press, Cambridge, MA.

Foucault, M., 2007. Security, Territory, Population. Palgrave Macmillan, London.

Foucault, M., 1977. Discipline and Punish. Pantheon Books, New York.

Giddens, A., 1985. The Nation-State and Violence. Polity Press, Cambridge.

Gill, M., Spriggs, A., 2005. Assessing the Impact of CCTV, Home Office Research Study 292. Home Office, London.

Graham, S., 2005. Software-sorted geographies. Progress in Human Geography 29, 562–580.

Graham, S., 2010. Cities under Siege. The New Military Urbanism. Verso, London.

Haggerty, K., Ericson, R., 2000. The surveillance assemblage. British Journal of Sociology 51, 605–621.

Jones, R., 2000. Digital rule: punishment, control and technology. Punishment & Society 2, 5–22.

Latour, B., 2005. Reassembling the Social. University Press, Oxford.

Lyon, D. (Ed.), 2006. Theorizing Surveillance: The Panopticon and Beyond. Willan, Cullompton.

Lyon, D., 2007. Surveillance Studies. An Overview. Polity, Cambridge.

Lyon, D., 2003. Surveillance as social sorting: computer codes and mobile bodies. In: Lyon, D. (Ed.), Surveillance as Social Sorting. Privacy, Risk, and Digital Discrimination. Routledge, London, pp. 13–30.

Marx, G., 2012. "Your papers please": personal and professional encounters with surveillance. In: Lyon, D., Ball, K., Haggerty, K. (Eds.), International Handbook of Surveillance Studies. Routledge, London, pp. xx–xxxi.

Rössler, B., 2001. Der Wert des Privaten. Suhrkamp, Frankfurt a.M.

Welsh, B.C., Farrington, D.P., 2002. Crime Prevention Effects of Closed Circuit Television: A Systematic Review, Home Office Research Study 252. Home Office, London.

Surveillance Studies

Gary T Marx, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article suggests some basic terms for surveillance analysis. The analysis requires a map and a common language to explain and evaluate its fundamental properties, contexts, and behaviors. Surveillance is neither good nor bad but context and comportment make it so. Topics considered in this article include: a broad definition of surveillance, its strategic and nonstrategic forms, and the traditional and new surveillance. A family of related terms – privacy, publicity, confidentiality, and secrecy – is also considered. The discussion next focuses on characteristics of the social structures that organize behavior, the characteristics of the *means* used, and some value conflicts and social processes seen with the emergent, interactive character of much surveillance behavior.

Queen Elizabeth (1533–1603) introduced modern ideas about the rights of the person including protection against "windows into men's hearts and secret thoughts." In this view, she draws on an English proverb: "the eyes are the windows into the soul" that in turn reflects the biblical claim (Matthew 6: 22–23) that probative looking into another's eyes reveals who they are (questions of deception and validity apart). Queen Elizabeth realized that the dignity of the person requires limits on looking, particularly when coercion and inequality are present, as with state power. Yet as a ruler concerned with the welfare of her subjects she needed information about them, as well as about rule breakers and those who would overthrow her government. Her challenge – juggling the protection of individuals' hearts and secret thoughts and the protection of state security – is one that faces democratic leaders everywhere.

This article suggests some basic terms for surveillance analysis. A map and a common language are required to explain and evaluate its fundamental properties, contexts, and behaviors. The empirical richness of watching and being watched (whether involving the eye or other senses and various kinds of data) and the uses of surveillance results need to be disentangled and parsed into basic categories and dimensions. After offering a brief comment on surveillance studies and a broad definition of surveillance, attention is given to strategic and nonstrategic forms, and traditional and the new forms of surveillance. A family of terms related to surveillance – privacy, publicity, confidentiality, and secrecy – is considered. The discussion next focuses on characteristics of the *social structures* that organize behavior and the characteristics of the *means* used and some *social processes* seen with the emergent, interactive character of much surveillance behavior. I next turn to value conflicts that make surveillance often so controversial. The discussion that follows draws from Marx (2015, and articles at www.garymarx.net) and is informed by the sources in **Table 1**.

What Is Surveillance?

Today's *surveillance society*, as is involves the state, the private sector, and interpersonal relations, brings forth the same paradox faced by Queen Elizabeth noted in the preceding paragraph – whether the National Security Agency's worldwide gathering of metadata on telecommunications, or merchants, employers, and parents watching customers, workers, and children, respectively. In a world where surveillance is seen as both a response to threats and a threat, before asking "Is surveillance good or bad?" we need to ask, "What concepts are needed to capture its basic structures and processes?" Surveillance as such is neither good nor bad, but context and comportment do make it so. The same can be said for the related concept

Table 1 Surveillance essays: names of the new society and its key aspects[a]

The panopticon (Bentham, 1995)
Disciplinary society, the gaze and bio-power (Foucault, 1977, 1998)
Surveillance society, the new surveillance and maximum security society (Marx, 1985, 2015)
Net widening (Cohen, 1985)
Dossier society (Laudon, 1986)
Dataveillance (Clarke, 1988)
Super-panopticon (Poster, 1990)
Society of control (Deleuze, 1992)
L'anamorphose de l'etat-nation (Palidda, 1992)
Panoptic sort (Gandy, 1993)
Minimum security society (Blomberg, 1987)
Synopticon (Mathiesen, 1997)
Securitization (Buzan et al., 1998)
Telematic society (Bogard, 1996)
Techno-policing (Nogala, 1995)
Transparent society (Brin, 1998)
The maximum surveillance society (Norris and Armstrong, 1999)
Liquid modernity (Bauman, 2000)
Information empire (Hardt and Negri, 2001)
Surveillant assemblage (Haggerty and Ericson, 2006)
Post-panopticon (Boyne, 2000)
Glass cage (Gabriel, 2005)
Ban-opticon (Bigo, 2006)
High policing (Brodeur and Leman-Langlois, 2006)
Ubiquitous computing (Greenfield, 2006)
Ubereveillance (Michael et al., 2008)
Ambient intelligence (Wright et al., 2008)
Safe society (Lyon, 2007)
Thick and thin surveillance (Torpey, 2007)

[a]A representative, although hardly exhaustive, list.

of privacy. Context refers to the type of institution and organization in question and to the goals, rules, and expectations they are associated with. *Comportment* refers to the kind of behavior expected (whether based on law or less formal cultural expectations) of, and actually shown by, those in various surveillance roles.

While sharing some elements, differences in surveillance contexts involving coercion (government), care (parents and children), contracts (work and consumption) and free-floating accessible personal data (the personal and private within the public) need consideration. Surveillance is a generic process characteristic of living systems with information borders and not something restricted to governments, spying, or secrecy. Surveillance and privacy are not necessarily in opposition and the latter can be a means of insuring the former as with access controls to information. While media attention to the problems associated with inappropriate surveillance (particularly by government) is present, there are also problems associated with the failure to use surveillance when it is appropriate. The emerging interdisciplinary field of surveillance studies analyzes these issues.

Surveillance Studies

The watchful and potentially wrathful (although also sometimes loving and protective) eye of the Biblical God of the Old Testament offers an early example of surveillance. More modern authors include Hobbes, Rousseau, Bentham, Marx, Nietzsche, Weber, and Taylor. Foucault (1977) (although writing about earlier centuries) is the grandfather of contemporary surveillance studies.

The field of surveillance studies came to increased public and academic attention after the events of 9/11 (Monaghan, 2006; Ball et al., 2012). But the topic in its modern form has been of interest to scholars at least since the 1950s. This is related to greater awareness of the human rights abuses of colonialism, fascism, and communism and anti-democratic behavior within democratic societies. The literary work of Huxley, Orwell, and Kafka and the appearance of computers and other new technologies with their profound implications for social behavior, organization, and society are also factors in the field's development.

In the form of *the surveillance essay* current writers from many disciplines and perspectives (e.g., political economy, social control, law and society, and criminology) draw on and extend the earlier theorists to describe the appearance of a new kind of society with new forms of social ordering (**Table 1**).

As ideal types, the terms in **Table 1** such as the 'panopticon,' 'disciplinary society,' or 'maximum security society' combine many strands that need to be separated if we are to move beyond sweeping claims made about surveillance. The concepts discussed in this article seek to bring greater precision and to add some leaves to the trees. One way to do that is to develop a middle range approach that fills out a general concept such as the *maximum security society* (as well as most of the other broad surveillance society naming concepts in **Table 1**) by identifying subsocieties that compose the surveillance society.

The maximum society concept draws parallels to the *total institution* and the maximum security prison and suggests that forms of control traditionally associated with the prison are diffusing into the broader society. But as an abstract notion, it does little to analyze variation in surveillance practices and changes over time. To do that the threads of the tapestry must be unwound in a series of subsocieties. Among components of the contemporary surveillance society are: a *hard engineered society*; a *soft and seductive engineered society*; a *dossier society*; an *actuarial society*; a *transparent society*; a *self-monitored society*; a *suspicious society*; a *networked society* of ambient and ubiquitous sensors in constant communication; a *safe and secure society* with attenuated tolerance for risk and with a strong emphasis on prevention; a '*who are you society?*' of protean identities, both asserted by and imposed upon individuals; and a '*where are you, where have you been, and who else is there?' society* of documented mobility, activity, and location. The broad approach of the surveillance essays is important in calling attention to contemporary changes, but such work generally does not take us beyond broad statements, does not adequately define surveillance, nor identify components that would systematically permit differentiating the new from the old forms, or making comparisons between various surveillance uses, contexts, and societies. In offering neither inclusive and nuanced definitions nor adequately elaborating dimensions, they fail to call attention to what is universal in human societies or to offer ways to analyze what is different.

Defining Surveillance

The English noun *surveillance* comes from the French verb *surveillir*. It is related to the Latin term *vigilare* with its hint that something vaguely sinister or threatening lurks beyond the watchtower and town walls. Still, the threat might be successfully warded off by the vigilant. This ancient meaning is reflected in the association many persons still make of surveillance with the activities of police and national security agencies. Yet in contemporary society the term has a far wider meaning.

What is surveillance? The dictionary, thesaurus, and popular usage suggest a set of related activities: look, observe, watch, supervise, control, gaze, stare, view, scrutinize, examine, checkout, scan, screen, inspect, survey, glean, scope, monitor, track, follow, spy, eavesdrop, test, or guard. While some of these are more inclusive than others and can be logically linked (e.g., moving from look to monitor), and while we might tease out subtle and distinctive meanings for each involving a particular sense, activity, or function, they all reflect what the philosopher Ludwig Wittgenstein calls a family of meanings within the broader concept.

At the most general level surveillance of humans (which is often, but need not be synonymous with human surveillance) can be defined as regard or attendance to others (whether a person, a group, or an aggregate as with a national census) or to factors presumed to be associated with these. A central feature is gathering some form of data connectable to individuals (whether as uniquely identified or as a member of a category).

A verb such as 'observe' is not included in the definition because the nature of the means (or the senses involved) suggests subtypes and issues for analysis and ought not to be foreclosed by a definition (e.g., how do visual, auditory, text, and other forms of surveillance compare with respect to factors such as intrusiveness or validity?). If such a verb is needed, to 'scrutinize,' 'regard,' or 'attend to' is preferable to observe, with its tilt toward the visual.

Many contemporary theorists offer a narrower definition tied to the goal of control (e.g., Dandeker, 1990; Lyon, 2001; Manning, 2008; Monahan, 2010). Taking a cue from Foucault's earlier writings, control as domination is emphasized (whether explicitly or implicitly) rather than as a more positive direction or neutral discipline. Yet, as Lianos (2003) observes, the modern role of surveillance as control must be placed in perspective alongside its fundamental importance in enhancing institutional efficiency and services.

Surveillance – particularly as it involves the state and organizations, but also in role relationships as in the family – commonly involves power differences and on balance favors the more powerful. Understanding this is furthered with comparisons to settings where control and domination are not central as with other goals such as surveillance for protection, entertainment, or contractual relations; where surveillance is reciprocal; and where it does not only, or necessarily, flow downward or serves to disadvantage the subject.

Authority and power relations are closely related to the ability to collect and use data. The conditions for accessing and using information are elements of a democratic society (Haggerty and Samatas, 2010). The greater the equality in subject–agent settings, the more likely it is that surveillance will be bilateral. Given the nature of social interaction and a resource-rich society with civil liberties, there is appreciable data collection from below as well as from above and also across settings. Reciprocal surveillance can also be seen in many hierarchal settings.Mann et al. (2003) refer to watchful vigilance from below as *sousveillance*.

The definition of surveillance as hierarchical watching over or social control is inadequate. The broader definition offered here is based on the generic activity of surveilling (the taking in of data). It does not build in the goal of control, nor specify directionality. In considering current forms we need to appreciate bidirectionality and horizontal as well as vertical directions. Control needs to be viewed as only one of many possible goals and/or outcomes of surveillance. When this is acknowledged, we are in a position to analyze variation and note factors that may cut across kinds of surveillance.

In his analysis of "The Look" Sartre (1993) captures a distinction between *nonstrategic* and *strategic surveillance*. He describes a situation in which an observer is listening from behind a closed door while peeking through a keyhole when "all of a sudden I hear footsteps in the hall." He becomes aware that he himself will now be observed. In both cases he is involved in acts of surveillance, but these are very different forms. In the latter case he simply responds and draws a conclusion from a state of awareness. In the former he has taken the initiative, actively and purposively using his senses.

Nonstrategic surveillance refers to the routine, autopilot, semiconscious, often even instinctual awareness in which our sense receptors are at the ready, constantly receiving inputs from whatever is in perceptual range. Smelling smoke or hearing a noise that might or might not be a car's backfire are examples. In contrast, strategic surveillance involves a conscious strategy to gather information. This may be in a cooperative or adversarial setting – contrast parents watching a toddler with corporations intercepting each other's telecommunications.

Within the strategic form, which to varying degrees ferrets out what is not freely offered, we can identify two mechanisms intended to create (or prohibit) conditions of visibility and legibility: material *tools* that enhance (or block) the senses and *rules* about the surveillance itself. While these are independent of each other, they show common linkages, as with rules requiring reporting when there are no available tools for discovery or rules about the conditions of use for tools that are available. A stellar example is the 'Lantern Laws' that prohibited slaves from being out at night unless they carried a lantern (Browne, 2015). Here, the emphasis is on requiring the subject to make him or herself visible given the limitations brought by darkness. But note also efforts to alter environments to make them more visible as with the creation of 'defensible space': via taking down shrubs or using glass walls (Newman, 1972) or less visible ala the architecture of bathrooms.

Within the strategic form we can distinguish traditional from the new surveillance. Examples of the new surveillance include computer matching and profiling, big data sets, video cameras, DNA analysis, GPS, electronic work monitoring, drug testing, and the monitoring made possible by social media and cell phones. The new surveillance tends to be more intensive, is extensive, extends the senses, is based on aggregates and big data, has lower visibility, involves involuntary (often categorical) compliance of which the subject may be unaware, tends to decrease cost, and reach remote locations. While the historical trend here is clear, it is more difficult to generalize about other characteristics such as whether or not surveillance has become more deceptive or more difficult to defeat than previously. Many forms are more omnipresent and often presumed to be omnipotent.

The new surveillance may be defined as scrutiny of individuals, groups, and contexts through the use of technical means to extract or create information. In this definition the use of 'technical means' to extract and create the information implies the ability to go beyond what is naturally offered to the senses and minds unsupported by technology, or what is voluntarily reported. Many of the examples extend the senses and cognitive abilities by using material artifacts, software, and automated processes, but the technical means for rooting out can also involve sophisticated forms of manipulation, seduction, coercion, deception, infiltrators, informers, and special observational skills.

Including 'extract and create' in the definition calls attention to the new surveillance's interest in overcoming the strategic or logistical borders that inhibit access to personal information. These inhibitors may involve willful hiding and deception on the part of subjects or limits of the natural world, senses, and cognitive powers. Create also suggests that data reflect the output of a measurement tool. The tool itself reflects a decision about what to focus on and the results are an artifact of the way they were constructed. Of course, constructions vary in their usefulness, validity, and reliability. Our perceptions of the empirical world are conditioned by where and how we look and these may vary in their fidelity to that world.

The use of 'contexts' along with 'individuals' recognizes that much modern surveillance attends to settings, or patterns of relationships and groups, beyond focusing on a given, previously identified individual. Meaning may reside in cross-classifying discrete sources of data (as with computer matching and profiling) that, when considered separately, are not revealing. Systems as well as persons are of interest. The collection of group data or the aggregation of individual into group data offers parameters against which inferences about individuals are drawn for purposes of classification, prediction, and response. Depending on the parameters, this may bring rationality and efficiency, but there is an inferential leap in going from *group* characteristics based on *past events* to *future* predictions about a given *individual*.

This definition of the new surveillance excludes the routine, nontechnological surveillance that is a part of everyday life, such as looking before crossing the street or seeking the source of a sudden noise or an unusual scent. It also excludes the routine attentiveness to, and interaction with, others that is fundamental to being a social being (as with mannerly behavior such as opening the door for another or offering a seat to an elderly person). An observer on a nude beach or police interrogating a cooperative suspect would also be excluded, because in these cases the information is volunteered and the unaided senses are sufficient.

Related but Distinct: Surveillance and Privacy, Privacy, and Publicity

How do surveillance and privacy relate? Surveillance is often wrongly seen to be the opposite of privacy. Kelvin (1973) emphasized this role of privacy as a nullification mechanism for surveillance. But at the most basic level, surveillance is simply a way of discovering and noting data that may be converted to information. Thus, depending on the context and role played, individuals or groups may be required, find it optional, or be prohibited from engaging in these activities, whether as subjects or agents of surveillance and communication. This obviously can involve invasions of privacy, as it was with the employee in a lab testing for AIDS who sold information on positive results to a mortuary.

Yet surveillance can also be the means of protecting privacy. Consider biometric identification and audit trails required to use some databases, or defensive measures such as a home security video camera. Privacy for whom, surveillance of whom, by whom, and for what reasons need to be specified.

Privacy like *surveillance* is a multidimensional concept whose contours are often ill-defined, contested, negotiated, and fluid, dependent on the context and culture. Among the major forms are *informational* (Westin, 1967), *aesthetic* (Rule et al., 1980), *decisional* (Decew, 1997), and *proprietary* (Allen, 2007) privacy. Informational privacy (Westin, 1967) is the most significant and contested contemporary form and involves the rules and conditions around personal information.

Breaches of decisional or proprietary privacy involve application or use of private information, rather than information discovery. While distinct, informational privacy shares with the other forms the key factor of control over access to the person or at least the person's data and the forms may be temporally connected. Thus, if individuals can control their personal information – whether not having to reveal their purchase of birth control pills (when this was illegal) or keeping paparazzi from taking pictures – they need not worry about that information being used.

Informational privacy encompasses *physical privacy*. The latter can refer to insulation resulting from natural conditions such as walls, darkness, distance, skin, clothes, and facial expression. These can block or limit outputs and inputs. Bodily privacy is one form of this, and its borders can be crossed by implanting something such as a chip or birth control device or removing something, such as tissue, fluid, or a bullet. Within informational privacy we find the conditions of anonymity and pseudo-anonymity, often referred to as being necessary for another type of privacy involving seclusion and being left alone. Personal information borders are obviously more difficult to cross if an individual cannot be reached via name or location.

Informational privacy can be considered as it ties to institutional setting (e.g., financial, educational, health, welfare, employment, criminal justice, national security, voting, and census); places and times; the kind of data involved, such as about religion or health; participant roles; and aspects of technology and media, such as audio or visual, wire or wireless, print, phone, computer, radio, or TV. Considerations of setting, location, time, data type, and means are central to legislation and regulation and rich in anomalies and cross-cultural differences.

A concept related to privacy is publicity. The two can be linked within the same framework. The common elements are rules about the protection and revelation of information. In some countries such as Canada the same officials are responsible for privacy and for freedom of information. In the first case there are rules giving individuals the right to control their personal information and in the second rules requiring that information not be restricted – that is, that it be made public. While sharing elements, for policy purposes there are major differences between the privacy of individuals and the secrecy of organizations. The standards for the latter should not automatically be applied to the former.

As nouns privacy and publicity can be seen as polar ends of a continuum involving rules about withholding and disclosing, and seeking or not seeking, information. Thus, depending on the context and role played, individuals or groups may be required, find it optional, or be prohibited from engaging in these activities, whether as subjects or agents of surveillance and communication.

When the rules specify that a surveillance agent is not to ask certain questions of (or about) a person and the subject has discretion about what to reveal, we can speak of *privacy norms*. When the rules specify that the subject must reveal the information or the agent must seek it, we can speak of *publicity norms* (or, better perhaps, disclosure norms). With publicity norms there is no right to personal privacy that tells the agent not to seek information, nor does that give the subject discretion regarding revelation. Rather, the reverse – the subject has an obligation to reveal and/or the agent to discover (Marx, 2011).

The moral expectations surrounding information as a normative phenomenon (whether for protection as with

privacy or revelation as with publicity and whether based on law, policy, or custom) can be differentiated from the empirical status of the information as known or unknown. To understand this distinction, we need the related terms, private and public – adjectives that can tell us about the status of information. Is information known or unknown, does it have an objective quality, can it be relatively easily measured? For example, in face-to-face encounters one generally knows the gender and face of a stranger, whether this is in the street, an office, or a home. The information is 'public' as in readily accessible, and this may be supported by anti-mask laws and requirements to wear symbolic items of clothing, tattoos, or badges. Absent such rules, the stranger's political or religious beliefs are more likely to be invisible and unknown.

Normative expectations of privacy and publicity do not always correspond to how the adjectives public and private are applied to empirical facts. Thus, the cell phone conversations of politicians and celebrities that have privacy protections may become public. Information subjected to publicity requirements such as government and corporate reports and disclosure statements may be withheld, destroyed, or falsified. Information not entitled to privacy protections, such as child or spouse abuse, may be unknown because of the inaccessibility of the home to broader visibility. The distinction calls for empirical analysis of the variation in the fit between the rules about information and what actually happens to it.

Privacy and publicity can be thought of in literal and metaphorical spatial terms involving invisibility–visibility and inaccessibility–accessibility. The privacy offered by a closed door or a wall and an encrypted e-mail message share information restriction, even as they differ in many other ways. Internet forums are not geographically localized, but in their accessibility can be usefully thought of as public places, not unlike the traditional public square where exchanges with others are possible or where others are visible as with open architecture. Erving Goffman (1971) in writing of 'relations in public' and 'public life' attends to the elements and possibilities within the immediacy of physical copresence. This is the strand of 'publicness' as visibility. It suggests the 'public' as known to at least one other person rather than to any rules about the status of information (that it must be revealed or concealed or to a legally defined place such as private golf course) as noted above. Thus, we can paradoxically speak of 'public order in private places' (Goffman, 1971: XIV).We can also speak about expectations of the private even within the public (Nissenbaum, 1998; Marx, 2001). In the latter case, since the information is available (such as with someone's appearance or a conversation overheard in a restaurant). But limits remain as with expectations about not staring or listening too closely to what others nearby are saying.

Structures and Processes

Regardless of whether we are dealing with traditional or the new surveillance, some common classificatory notions can be applied. In the case of surveillance social structures, for example, we can identify the *surveillance agent* (whether as watcher/observer/seeker/inspector/auditor/tester), while the person about whom information is sought or reported is the *surveillance subject*. The agent role can be further separated into the *sponsor*, *data collector*, and *initial* or *secondary user*.

Many contemporary concerns over surveillance involve the practices of large organizations relative to employees, clients, or the public. *Organizational surveillance* is distinct from the *nonorganizational surveillance* carried about by individuals. At the organizational level, formal surveillance involves a constituency. Organizations have varying degrees of internal and external surveillance (Rule, 1973). Erving Goffman (1961) has identified many kinds of employee or inmate monitoring, such as within 'total institutions.'

Within an organization *internal constituency surveillance* (scrutiny of insiders as with work monitoring) contrasts with *external constituency surveillance* (attending to outsiders such as customers, patients, and travelers). *External nonconstituency surveillance* involves organizations monitoring their broader environment in watching other organizations and social trends. The rapidly growing, understudied field of business intelligence fits here.

In *nonorganizational surveillance* the watching is apart from a formal organizational role. It may involve *role relationship surveillance* as with family members (parents and children, the suspicious spouse) or friends looking out for and at each other (e.g., monitoring location through a cell phone). Or it can involve *non-role relationship surveillance* – as with the free-floating activities of the voyeur whose watching is unconnected to a legitimate role.

Agent-initiated surveillance illustrated by compliance checks such as an inspection of a truck or a boat is distinct from *subject-initiated surveillance* – as with volunteering to participate in consumer reward programs. The agent and subject of surveillance can merge with *self-surveillance* – where individuals watch themselves (home alcohol or reproductive tests).

While distinct, subject and agent can be intertwined, as with parallel or *co-surveillance* (e.g., remote health monitoring in which both the monitored person and a health agency simultaneously receive signals about the subject). In the latter case co-surveillance is *nonreciprocal* with personal data from the watched going to the watcher. This tends to be the case with employers, merchants, doctors, teachers, and parents and reflects power and resources differences. In contrast, *reciprocal surveillance* is by definition bidirectional, as with social networking sites. But reciprocal, need not mean equal. Surveillance that is reciprocal may be *asymmetrical* or *symmetrical*. In a democratic society citizens and government engage in reciprocal but distinct, and shifting forms and degrees of mutual surveillance. These questions draw attention to who is entitled to and/or able to play the agent role and who is the subject? New tools may bring increased democratization (or a better term – equalization) as with readily available cell phone cameras and Internet access or the tools may be restricted as with access to satellites, private databases, and sophisticated data mining.

A measure of democracy is the extent of restrictions on and mandatory requirements for information flows across actors and sectors. For example, what is the ratio over time of what governments and large organizations are expected to (or may) reveal about themselves (e.g., freedom of information, truth in advertising laws and policies, and conflict of interest statements) and what citizens are expected to reveal about

themselves to governments and large organizations. Contrast the extremes of a *totalitarian government* that has no obligation to reveal anything to citizens (who must reveal all to government), with the unrealistic case of a *fully open government* that must reveal all to citizens (who in turn reveal only what they choose to government).

The concepts above emphasize structure (i.e., patterned forms of behavior and organization). Structure implies that the topic is static and fixed at one point in time, yet surveillance also needs to be viewed as a fluid, ongoing process involving interaction, strategic calculations, and negotiations over time. Among major processes are the *softening of surveillance* (in which minimal invasiveness, manipulation, persuasion, and low visibility replace traditional coercion); efforts to create the *myth of surveillance* (generating fears supportive of the need for increased surveillance and/or claims that the solution is more effective than it is); the *monetarization* of personal data so it can be sold to marketers, governments, and individuals for purposes of sorting and the *commodification* of surveillance in which it (or protection from it) become products to be purchased; *surveillance slack* – the relative gap between potentials and practices; and various techniques of neutralization. The latter are strategic moves by which subjects of surveillance seek to subvert the collection of personal information such as direct refusal, discovery, avoidance, switching, distorting, counter-surveillance, cooperation, blocking, and *masking*. Equivalent counter-neutralization moves by agents are also present (Marx, 2003).

Another way to think about process is to consider the links between the distinct activities covered by the umbrella term surveillance. The most common meaning refers to acts of data collection, but these occur within a broader system of connected activities. Seven kinds of activity conceived as *surveillance strips* can be noted: *tool selection*, *subject selection*, *collection*, *processing/analysis*, *interpretation*, *uses/action*, and *data fate*. Considered together these strips constitute the *surveillance occasion* and offer a way to bind a given application.

Viewing surveillance in stages also permits us to see one link between *surveillance* and *communication* – the two related topics that are rarely connected. However, at the use stage suggested above surveillance results can be used for targeted marketing, solicitation, and propaganda (Gandy, 1993). Both can involve the crossing of personal borders and respect for, or the disavowal of, a person's dignity – in one case to take from and in another to impose upon, the person in contexts of roles, rules, and expectations. They also share some issues of interpretation and meaning regarding the data surveillance collects and the content of a message.

The 'career' of a particular surveillance tool may also be tracked as it emerges and then may diffuse – whether in a jagged or in a more linear direction. In the case of the latter this may be through *surveillance creep* or *gallop*, often displacing other means along the way and bringing new goals and users. There may also be *surveillance constriction* as a new, unregulated tactic becomes subject to limitations and even prohibitions. Thus, the rampant surveillance that computerization and related forms of electronic monitoring and communication made possible has been somewhat curtailed by laws and policies. An early example was the Electronic Privacy Protection Act of 1986 or with the new concerns that DNA brought the 2008 Genetic Information Nondiscrimination Act. Robert Smith (2013) reports on the vast expansion of state and federal laws regarding surveillance and privacy.

Understanding Surveillance and Value Conflicts

There are many ways of approaching surveillance from the social sciences, and the perspective and methods chosen need to reflect specific questions and local situations (e.g., How were suspects identified before photography was invented?, What are the most important causes of the emergence of a surveillance society?, and Do young people differ in their assessment of the privacy implications of social media from older people?) However, regardless of the kind of question, our understanding will benefit from awareness of factors such as the level of analysis and conflicts in values discussed next.

Understanding surveillance will be furthered to the extent that we:

1. Identify questions and levels of analysis. Central here for both explanation and policy purposes is awareness of social and cultural factors present within different contexts and institutional settings. In addition, how we account for and judge surveillance should depend on the role played and the characteristics of the tool, goals, the kind of data involved, and the values at stake.
2. Identify variation and then look for correlations and explanatory causes/drivers within a framework of soft determinism, but do not confuse correlation with causality.
3. Realize that the same causal factor(s) can have different outcomes and the equivalent outcomes can have different causes.
4. Appreciate the advantages of a loose systems approach with some open-ended borders and room for (but limits upon) exogenous inputs.
5. Attend to beginnings (or at least prior circumstances) – everything was preceded by something before and new ways of meeting human needs must be compared with old ways.
6. Note that while some things change, others remain the same.
7. View surveillance as a process, as well as a structure or a tool. Central to that is studying interactions involving agents, subjects, third parties, and audiences over varying time periods.
8. Ask about the appropriateness of both means and ends. Desirable ends do not justify doubtful means, and good means can be misused. Good goals and purity of motives are not sufficient justification. Consider the acceptability of means and ends independently, as well as in their relationship to each other. Recognize that a given tool can serve a variety of goals and that a given goal can be met by a variety of tools.
9. Differentiate facts from values. No matter how sound the method or clear the findings, a leap to values, ethics, and political choices always remains in what we come to see as facts and in the ends for which surveillance is used.
10. With respect to values (and the goals and specific applications to which they are related) be aware of how their

abstract nature and conflicts between them mean that surveillance will often be subject to legitimate disagreements among well-meaning persons. (The issue of value conflicts is vital and I discuss it in the following paragraphs.)

Value conflicts are everywhere, like the weather. Thus, we seek privacy and often in the form of anonymity, but we also know that secrecy can hide dastardly deeds and that visibility can bring accountability. On the other hand, too much visibility may inhibit experimentation, creativity, and risk taking. And while we value disclosure and 'permanent records' in the name of fairness and just deserts, we also believe in redemption. New beginnings after individuals have been sanctioned, or after they have otherwise overcome limitations or disadvantages, are fundamental to the American reverie.

In our democratic, media-saturated, impression-management societies, many of us want to both see and be seen (e.g., social media) even as we also want to look the other way and be left alone. We may want to know but also be shielded from knowing.

We value freedom of expression and a free press but do not wish to see individuals defamed, harassed, or unduly humiliated (whether by the actions of others or their own). Also as ideals, we desire honesty in communication and also civility and diplomacy. In interpersonal relations (in contrast to the abrasive talk shows) we may work hard to avoid embarrassing others by not seeking certain information or by holding back on what we know. We value the right to know, but also the right to control personal information. The absence of surveillance may bring freedom from censorship, but also open the door to the worst demagogues, liars, and self-deluded snoops. Yet undue surveillance chills nonconforming communication and is the companion of repression.

Individuals expect organizations to treat them in a fair, accurate, and efficient manner, and to judge them as unique, not as undifferentiated members of a general category, while at the same time, they hesitate to reveal personal information and desire to have their privacy and confidentiality protected. Meeting the first set of goals necessarily requires giving up personal data, and up to some point, the more one gives up, the more accurate and distinctly reflective it will be of the unique person. Yet, the more data one reveals, the greater the risk of manipulation, misuse, and privacy violation. At the same time, knowing more can bring new questions and less certainty to surveillance agents. Depending on their role and social location, individuals and groups differ in the relative importance they give to privacy as compared to accuracy.

The individual's expectation to be assayed in his or her full uniqueness may conflict with an organization's preference for responding to persons as part of broad common aggregates – something seen as more rational, effective, and even efficient. The idea of due process and fairness to be determined in each individual case can radically conflict with an organization's utilitarian goals and bottom line. In the criminal justice context, for example, civil liberties sometimes conflict with the goal of effective enforcement. The case for *categorical surveillance* (without cause) versus *particularized surveillance* (only with cause) and for prevention versus after-the-violation responses can be well argued either way.

Culture sends contradictory messages. On the one hand, individuals are expected to submit to surveillance as members of a community that supports the common good and fairness (e.g., the required census or social security number that apply to all) or that allows one to participate in certain behaviors such as traveling, buying on credit, or obtaining an entitlement. Yet fairness apart, when such surveillance goes beyond minimal verification and is done in a coercive manner, it may conflict with the expectation that before personal information borders are crossed, there needs to be some grounds for suspicion. If agents have to wait to do surveillance until they have cause in situations where there is evidence of preparatory actions or where violations are of low visibility or hidden outright, many violators get a free ride. This limitation protects the innocent against unnecessary searches. Yet, it can also mean failing to prevent terrible events – for example, in the case of 9/11, where well-intentioned policies from another era as well as many informal factors blocked the FBI and CIA from exchanging information about the perpetrators.

If your tools work and if you search them all, you will likely get the guilty, not to mention the innocent and this is likely inefficient. Profiling as a surveillance tool permeates society far beyond ethnicity, religion, or national origin. In contemporary society, with its emphasis on prevention, the push is toward broader and deeper searching for the absent cause. The dilemma can be identified but not solved because observers differ in judging the trade-offs between equality, fairness, grounds for suspicion, invasiveness, prevention and effectiveness and the likelihood and seriousness of risks, and the potential for identifying and remedying failures.

The existence of practices with a good potential for abuse traditionally leads to demands for regulation. Thus, we see bureaucratic and legalistic responses. These may lessen problems, but ironically, can also lead to expanded use of potentially troubling means. In contrast, without a formal mandate legitimating and acknowledging the tactic, agents may hesitate to use it because of uncertainty about where to draw the lines.

The above discussion involves conflicts between abstract values. But more concrete conflicts may also appear in applying the tools. The intrinsic properties of a device may work against the agent's desire for secrecy. While much contemporary surveillance is defined by its ability to root out the unseen and unknown, it also paradoxically may reveal itself through electrical, chemical, and other forms of data. That which silently gathers the emanations of others, if not exactly a mirror image, nonetheless emanates itself, offering discovery possibilities and means of neutralization to technically competent adversaries. The watchers may also be watched by the means they apply to others. Also, if an agency publicizes a surveillance system that has as one goal, making citizens feel more secure, it may in addition have the opposite effect because it sends the message that dangers are so great as to warrant the need for such a system. Or this same publicity may alert committed malefactors to the presence of surveillance, triggering evasive, blocking, or displacement means – a kind of unfair (at least to the law-abiding public) warning. Thus, advertising the means versus keeping them secret highlights the potential conflict of goals between deterrence and apprehension so apparent with undercover work. The existence of practices with a good potential for abuse traditionally leads to demands

for regulation. A bureaucratic and legalistic response may lessen problems, but ironically, it can also lead to expanded use of potentially troubling means. In contrast, without a formal mandate legitimating and acknowledging the tactic, agents may hesitate to use it because of uncertainty about where to draw the lines.

A final mandate: with respect to the significant very complex issues new and old forms of surveillance bring to a democratic society, neither a pessimist nor an optimist be. Unless, of course, your view is based on empirical evidence, and a logical and well developed argument that defines its terms and identifies its background assumptions regarding both values and the empirical workings of the world.

Surveillance practices need to be understood within specific settings in light of history, culture, social structure and the give and take of interaction, and require the appreciation (if not necessarily the welcoming) of the ironies, unintended consequences, and value conflicts that limit the best laid plans. Mushrooms do well in the dark, but so does injustice. Sunlight may bring needed accountability through visibility, but it can also blind and burn.

There is a path, however twisting, changing, and bramble and illusion filled between Tennyson's early nineteenth century optimism, "For I dipt into the future, far as the eye could see, saw the world, and all the wonders that would be (Ricks, 1989) and Einstein's 20th century worry that technological progress can become like an axe in the hand of a criminal" (Folsing, 1998). Yet, hope must trump dread as we keep the faith with respect to both the importance of having a dream and the ameliorative potential of critical analysis. Empirical and scientific knowledge about human and social conditions can result in the improvement of those conditions, or maybe stop them from getting worse.

The study of surveillance is a reminder that while they (whether the state, commercial interests or new, expanding public–private hybrid forms) are watching us, we as citizens must watch them. And as baseball player Yogi Berra said, "you can see a lot by looking." Making surveillance (as with any technology) more visible and understandable hardly guarantees a just and accountable society, but it is surely a necessary condition for one.

See also: Community Supervision and Diversion in the United States; Goffman, Erving (1922–1982); Government Statistics; Governmentality; Internet and Privacy; Power in Society; Privacy: Theoretical and Legal Issues; Surveillance and Privacy, Geography of.

Bibliography

Allen, A., 2007. Privacy Law and Society. West Group.
Ball, K., Haggerty, K., Lyon, D., 2012. Handbook of Surveillance Studies. Routledge, New York.
Bauman, Z., 2000. Liquid Modernity. Polity, Cambridge.
Bentham, J., 1995. Introduction. In: Bozovic, Miran (Ed.), The Panopticon Writings. Verso, London.
Bigo, D., 2006. Globalized (in)Security: the field and the Ban-opticon. In: Sakai, N., Solomon, J. (Eds.), Translation, Biopolitics, Colonial Differences. Hong Kong University Press,, Hong Kong, pp. 109–156.
Blomberg, T., 1987. Criminal justice reform and social control: are we becoming security socitey. In: Lowman, J., Menzies, R.J., Plays, T.S. (Ed.), Transcarceration: Essays in the Sociology of Social Control. Gower Publishing Co., Surrey, England, pp. 218–226.
Bogard, W., 1996. The Simulation of Surveillance: Hyper Control in Telematic Societies. Cambridge University Press, Cambridge.
Boyne, R., 2000. Post-Panopticon. Economy and Society 29 (2), 285–307.
Brodeur, J.P., Leman-Langlois, S., 2006. Surveillance fiction or high policing. In: Haggerty, K., Ericson, R. (Eds.), The New Politics of Surveillance and Visibility. University of Toronto Press, Toronto.
Brin, D., 1998. The Transparent Society: Will Technology Force Us to Choose between Privacy and Freedom? Perseus Book Group, Cambridge, MA.
Browne, S., 2015. Dark Matters: On the Surveillance of Blackness. Duke University Press, Los Angeles, CA.
Buzan, et al., 1998. Security: A New Framework for Analysis. Lynne Reiner, Boulder, CO.
Clarke, R., 1988. Information technology and dataveillance. Communications of the ACM 31 (5), 498–512.
Cohen, S., 1985. Visions of Social Control. Polity Press, Cambridge.
Dandeker, C., 1990. Surveillance, Power, and Modernity: Bureaucracy and Discipline from 1700 to the Present Day. Polity Press, Cambridge.
Decew, J.W., 1997. In: Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Cornell University Press, Ithaca, NY.
Deleuze, G., 1992. Postscript on the societies of control. October 59, 3–7.
David, W., et al., 2010. Safeguards in a World of Ambient Intelligence. Springer, New York.
Folsing, A., 1998. Albert Einstein: A Biography. Penguin Books, New York.
Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Vintage, New York.
Foucault, M., 1998. The History of Sexuality, vol. 1. Penguin, London.
Gabriel, Y., 2005. Glass cages and glass palaces: images of organizations in image-conscious times. Organization 12 (1), 9–27.
Gandy, O., 1993. The Panoptic Sort. Westview Press, Boulder, CO.
Greenfield, A., 2006. Everyware: The Dawning Age of Ubiquitous Computing. New Riders, Berkeley.
Goffman, E., 1961. Asylums. Anchor Books, Garden City, NJ.
Goffman, E., 1971. Relations in Public. Micro Studies of the Public Order. Basic Books, New York.
Haggerty, K., Ericson, R., 2006. The surveillant assemblage. British Journal of Sociology 51 (4), 605–622.
Haggerty, K., Samatas, M., 2010. Surveillance and Democracy. Routledge, London.
Hardt, M., Negri, A., 2001. Empire. Harvard University Press, Cambridge, MA.
Kelvin, P., 1973. A social-psychological examination of privacy. British Journal of Clinical Psychology 12 (3), 248–261.
Laudon, K.C., 1986. Data quality and due process in large record systems: criminal record systems. Communications of the ACM 29 (1), 4–11.
Lianos, M., 2003. Social control after foucault. Surveillance & Society 1 (3), 412–430.
Lyon, D., 2001. Surveillance and Society: Monitoring Everyday Life, Open University Press, Buckingham and Philadelphia.
Lyon, D., 2007. Surveillance Studies: An Overview. Polity Press, Cambridge, England.
Lyon, D., Ball, K., Haggerty, K. (Eds.), 2012. International Handbook of Surveillance Studies. Routledge, London.
Mann, S., Nolan, J., Wellman, B., 2003. Sousveillance: inventing and using wearable computing devices for data collection in surveillance environments. Surveillance & Society 1 (3), 331–355.
Manning, P., 2008. A view of surveillance. In: Leman-Langlois, S. (Ed.), Technocrime: Technology, Crime, and Social Control. Willan Publishing, Devon, UK, pp. 209–242.
Marx, G.T., June 1985. The Surveillance Socitey. The Futurist XIX(3).
Marx, G.T., 2001. Murky conceptual waters: the public and the private. Ethics and Information Technology 3 (3), 157–169.
Marx, G.T., 2003. A tack in the shoe: neutralizing and resisting the new surveillance. Journal of Social Issues 59 (2).
Marx, G.T., 2011. Turtles, firewalls, scarlet letters and vacuum cleaners: rules about personal information. In: Aspray, W., Doty, P. (Eds.), Making Privacy. Scarecrow Press.
Marx, G.T., 2015 Windows into the Soul: Surveillance and Society in an Age of High Technology. University of Chicago Press, Chicago.
Mathiesen, T., 1997. The viewer society: Michel Foucault's 'panopticon' revisited. Theoretical Criminology 1 (2), 215–234.
Michael, M., Fusco, S., Michael, K., 2008. A research note on ethics in the emerging age of uberveillance. Computer Communications 31 (6), 1192–1199.
Monaghan, P., 17 March 2006. Watching the watchers. Chronicle of Higher Education 52 (28), 1 and 18–25.
Monahan, T., 2010. Surveillance in the Time of Insecurity. Rutgers University Press, New Brunswick.

Newman, O., 1972. Defensible Space: Crime Prevention through Urban Design. Macmillan Books, New York.
Nissenbaum, H., 1998. Protecting privacy in an information age: the problem of privacy in public. Law and Philosophy 17, 559–596.
Nogala, D., 1995. The future role of technology in policing. In: Brodeur, Jean-Paul (Ed.), Comparisons in Policing: An International Perspective. Aldershot, Avebury, UK, pp. 191–210.
Norris, C., Armstrong, G., 1999. The Maximum Surveillance Society: The Rise of CCTV. Oxford University Press, New York.
Palidda, S., 1992. L'anamorphose de l'État-Nation: le cas italien. Cahiers internationaux de sociologie 269–298.
Poster, M., 1990. The Mode of Information: Postculturalism and Social Context. University of Chicago Press, Chicago.
Ricks, C., 1989. Tennyson: a Selected Edition. University of California Press, Berkeley, CA.
Rule, J., 1973. Private Lives, Public Surveillance. Allen-Lane, London.
Rule, J., MCadam, D., Stearns, L., Uglow, D., 1980. The Politics of Privacy. New American Library, New York.
Sartre, Jean-Paul, 1993. Being and Nothingness. Washington Square Press, New York.
Smith, R.E., 2013. Compilation of State and Federal Privacy Laws. The Privacy Journal, Providence, RI.
Torpey, J., 2007. Through thick and thin: surveillance after 9/11. Contemporary Sociology 36 (2), 116–119.
Westin, A., 1967. Privacy and Freedom. Columbia University Press, New York.
Wright et al., 2008. Safeguard in a World of Ambient Intelligence. Springer, Dordrecht.

Survey and Excavation (Field Methods) in Archaeology

Martha Sharp Joukowsky, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15306–15311, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

This article presents a general introduction to the subject of archaeological field methods, and therefore does not deal with many distinct areas or challenges to the discipline. Archaeology is a dynamic, inherently interdisciplinary field of study in that new research techniques continue to revolutionize its modes of thought and methods, and researchers are asking different questions of the data today from those they asked only a few years ago. On both scientific and theoretical fronts, therefore, the field is developing at a fast pace. In the past few years, archaeology has become a global discipline with many researchers applying the same systems such as geographic information systems along with aerial reconnaissance and remote sensing. We begin with 'archaeology' as a term, and cover a brief history of the discipline, its importance, types of archaeological sites, fields of archaeological study, and dating systems used in the discipline. The focus then changes to survey, excavation and ethical considerations, excavation and the modern developments in the discipline.

The Term 'Archaeology'

The term 'archaeology' comes from the Greek word, *arkhaiologia*, meaning the 'discussion or discourse about ancient things' (*see* Archaeology, Theory in). By definition, archaeology is the science by which the remains of the ancients can be methodically and systematically studied to obtain a picture as complete as possible of ancient culture and society, and thereby to reconstruct past ways of life (Gamble, 2008). An antiquarian doctor of Lyon, Jacques Spoon, first used the term 'archaeology' in the seventeenth century. Archaeology is a discipline that involves the scientific method of study, observation, recording, and experiment (Moshenska, 2013). Archaeologists study in detail the complete life of cultures and their patterns of change in an attempt to delineate the causes and effects of the cultural process. The fascination of archaeology lies in the solving of long-standing problems and in the discovery of the meaning of things. The excitement of archaeology lies in the new problems which recent findings produce.

The History of Archaeology

The earliest 'archaeologist' was the sixth-century BCE King Nabonidus of Babylon who excavated a temple down to its foundations, which had been laid thousands of years before (Crowell, 2007). Even the Roman emperor Augustus was a collector of 'giant' bones and ancient weaponry. With time, Europeans began to collect burial urns, chipped flint stone, and ground stone tools which some believed were amulets. In 1719, the Austrian Prince d'Elboef led the first excavation to Herculaneum, which had been destroyed along with Pompeii in AD 79 when Mount Vesuvius erupted. Thereafter, when pieces of classical sculpture began to turn up, artists began to study their shapes and decoration and collectors began to collect, proudly display, and discuss their forms. With time, more people began to describe what they recognized as ancient monuments and excavations were undertaken to retrieve more and more objects. In the mid-eighteenth and early nineteenth centuries, digging for ancient objects became the pursuit of gentlemen of the leisured class and professionals: they became known as antiquarians and formed clubs to discuss the merits of ancient artifacts.

Archaeology as a pursuit using scientific systems of recovery began in the nineteenth century, along with the growing realization that humankind had an ancient past that could be investigated in a methodical way. The names of most important archaeologists from the early days include Sir William Matthew Flinders Petrie in Egypt, Robert Koldewey in Babylon, the duplicitous Heinrich Schliemann in the Aegean at Troy, Mycenae and Tiryns, and General Augustus Pitt-Rivers in Britain. Later pioneers were Sir R.E. Mortimer Wheeler in Britain and India, George Andrew Reisner and Sir Leonard Woolley in the Near East, Max Uhle and Alfred Vincent Kidder in America, and François Bordes and André Parrot in France.

In the present day, archaeology has grown into a multidisciplinary, methodical, problem-oriented search for the past (Insoll, 2007). It now involves above-ground research as well as excavation and a host of specialists is required, depending on the site. Background research as well as excavation has become meticulous. The study of pottery, for example, involves not only typologies but sophisticated scientific testing by neutron activation analysis to understand the chemical components and to identify the sources and cultural patterns as well as to assess trade relationships.

Archaeological questions extend far beyond the excavation and interpretation of any single cultural group, geographic region, or chronological period of time. In theory, all archaeological study is aimed toward an explanation that will provide a conceptual basis for the reconstruction of the past, and the where, how, and why of the changes that took place.

The Importance of Archaeology

Archaeology is unique in that it encompasses the whole world over the past 2.5 million years in order to reconstruct the past. It includes the research of all kinds of sites – underwater wrecks or port installations, caves, jungles, deserts, and mountain-top

sites; wherever past activity has taken place. Archaeology wants to discover where and when humankind developed and began to practice art, agriculture, and technology and became urbanized. It is our route to our understanding of ourselves, for in order to know where we are going we have to reconstruct and understand the past and where we have come from.

Archaeology continues to give us a broad and long-term view about the past and the events that conditioned us. There is an implosion of new answers to research on how, where, and when humankind began; our development of technology and the arts; the expression of our ideas in writing; and how we harnessed the environment brought about by the domestication of plants and animals, and developed agricultural systems (Bahn and Renfrew, 2005). Archaeology helps us to understand the processes of becoming urbanized, how we developed complex societies, and to understand our basic needs.

Archaeological Sites

Archaeology covers a wide diversity in time, space, subject matter, and approach. An archaeological site can be any place that has been a scene of past human activity, including a briefly settled campsite, a shipwreck, a monumental city, or the evidence of any number of different combinations of human remains. Any place where human beings have established themselves, even momentarily, is considered a site.

As each archaeological site is a unique time capsule, each has its own distinct character and problems. Types of sites are often labeled by their prehistoric or historic context, their function, their general topographic nature, or according to whether they are composed of one or many levels of occupation. Sites represent a body of data relevant to their setting and their cultural patterning and must be interpreted both in relation to this local setting and as a link between cultures. Sites may be small or large. Small sites are far more numerous throughout the world and may range from hunting or fishing stations to small village communities. There are prehistoric plains or riverside terraces such as Olduvai Gorge in Tanzania, Africa (Blumenschine, 1969–present); there are caves such as Chou Kou Tien near Peking, China where the Peking Man was found; or those containing Upper Paleolithic art, like Lascaux in France (Wallis, 2013).

Large sites are generally cities, temples, large tells – artificially built-up mounds in the Middle East with one city built on top of the other (known as *höyük* in Turkish or *tepe* in Persian). The ancient mound of Jericho, Israel is particularly important, revealing evidence dating back to 8000 BCE (Lalleman, 2008). A more recent tell of the Mississippi Basin was built by the Mound Builders of North America in about 1300 CE (*see* North America, Archaeology of). There are monumental cities such as classical Rome, Mesoamerican Tikal in Guatemala (*see* Mesoamerica, Archaeology of), or Angkor Vat in Cambodia (*see* Southeast Asia, Archaeology of). There are sacred areas reserved for ritualistic functions such as Stonehenge in England or the statues of Easter Island. There are cemeteries and tombs such as those of Mycenae in Greece; the Pyramids in Egypt; or the mounds of Ohio, Illinois, or Mississippi. Famous sites also include underwater shipwrecks and harbors discovered worldwide off the shores of the Mediterranean, or near Oslo, Norway, where early Viking ships have been recovered.

The selection of a site for excavation therefore depends upon what the excavator (principal investigator) wants to find, and what problems will be answered by excavation. What is the nature of the evidence? It usually involves a problem to be solved – How has the site been built up over time? What is its environment and how has this changed? Using a research design, a review process is undertaken which includes the preliminary analysis of the site, the evaluation of the probable results, the examination of the costs, and the evaluation of organizational abilities and needs to carry out the tasks involved.

Archaeological Studies

Many types of archaeology have evolved over the years. There are Old World (Eastern Hemisphere) and New World (Western Hemisphere) prehistoric specialties, based on the stone tools humankind produced. The Paleolithic, Mesolithic, Neolithic, Bronze Age, and Iron Age are specialties (Toth and Schick, 2009). These are broad artificial classifications with subdivisions. As far as the discipline in which it has received attention, Old World archaeology is a separate subject in some schools, or is associated with departments of classics or near-Eastern studies, history, or art history. In the New World, prehistoric specialists may concentrate on the Paleo-Indian, Archaic, and Woodland periods (Kelley, 2003). Then both Old and New World archaeology move into historical archaeology. In the Old World there are experts who are devoted to Classical archaeology (the worlds of Greece and Rome) or Biblical archaeology (the areas encompassed by the Holy Land) (Cline, 2009).

Similarly, many kinds of specialties have developed in archaeology, for example, industrial archaeology; the study of manufacturing and salvage archaeology; the rescue work done before a modern construction takes place such as land development, or urban renewal, or when a pipeline, road, or dam is built. There are also technical specialties of people who concentrate on specific techniques such as ceramics, lithics, paleobotany, palynologists (specialists in pollen), and dendrochronologists (specialists of tree ring growth) among many others, including computer specialists who lend support to all these areas.

Since the 1960s in America, archaeology as a subdiscipline also has been associated with anthropology (Hicks, 2013). As anthropology is the study of humanity, anthropologists generally observe the behavior of present-day people in order to explain and reconstruct their behavior patterns. In so doing, they have to make assumptions about past behavioral patterns as set against the environment of the present.

Anthropology is divided into several subfields among which are social and/or cultural anthropology; physical or biological anthropology; linguistic anthropology, which includes the interpretation of oral history; demographic anthropology; ethnic anthropology; cognitive and symbolic anthropology; enthno-archaeology; and archaeoastronomy (the study of ancient astronomical events) (Harrison-Buck, 2014).

Dating Systems in Archaeology

The term 'culture' is used when assuming that a group of people exhibits shared behavior. Archaeologists want to be able to establish accurate space and time frameworks or dates for the cultures they are researching. In short, archaeologists want to establish the site context, culture, and chronology. Laboratory methods provide what are known as absolute dating methods, and relative dates are ascribed to archaeological strata based on the site deposition or stratigraphy and the discovery of cultural artifacts similar to those found at other sites (*see* Chronology, Stratigraphy, and Dating Methods in Archaeology).

Absolute dating is a date given in specific years, in terms of chronometrical dating or the calendar. A date given as BP is a date before 1950. BC means 'before Christ' (although nowadays BCE is often used, meaning 'before the Common Era'). CE (Common Era) is used instead of the old term AD, meaning *Anno Domini*, or after the birth of Christ. Absolute dates can be established by laboratory analyses like radiocarbon dates (C-14) (Boaretto and Poduska, 2013), potassium–argon, thermoluminescence, archaeomagnetism, and palaeomagnetism, among others. These are geochronologic methods used for determining the absolute date of a deposit.

Relative dates are based on the stratigraphic deposition – the principles of superposition, association, and correlation (Gamble, 2008). Superposition assumes that the uppermost stratum (soil level) was deposited last and the lowest stratum was deposited first; thus the reasoning assumes that lower strata are generally older than the layers above them. Association assumes that the artifacts found in the same closed and sealed archaeological context are contemporary and are therefore presumed to be culturally and temporally associated. Correlation assumes that, once cultural associations have been determined, artifacts can be cross-culturally correlated with artifacts from other site cultures and they can be expected to have been manufactured at the same time, and therefore to possess the same relative date.

Archaeologists use seriation as a method of relative dating (Porčić, 2013). Seriation involves the ordering of specific artifact types from an earlier to a later time sequence by the use of a typology. In creating a type series, the archaeologist assumes that artifacts have formal attributes that can be measured like raw material, length, width, thickness, color, weight, and so on, and that these nonrandom attribute factors were used as a standard by the maker. Therefore, a typology is comprised of types which are a cluster of attributes in a class of objects and these can be given relative dates that can be used to date the stratification.

Once associations have been defined cultural levels can be inferred. A cultural level represents the choices of a particular social group, such as the classical Greeks. The data collected in such a cultural level depict a known and recognizable set of ideas that are then represented as the cultural products of that society.

Archaeological Survey

Survey includes exhaustive historical research, map reading, geomorphologic and geographic analysis, aerial photography, geospatial survey like Light Detection and Ranging (LiDAR), and ground-level surveys including geophysical surveying techniques (Johnson and Millett, 2013). Historical research includes all previous modern and historic research of an area to be surveyed – excavation reports, artifact studies, and the conclusions drawn by previous researchers. In-depth bibliographic research provides essential background information.

All archaeologists are concerned with surveying as a means of planning and laying out the area to be studied or excavated. Surveying is generally the first step in all archaeological projects, and is often the last step before the completion of the excavation. Using maps, aerial photos, airborne LiDAR, satellite imagery, and whatever background information is available, ground-level surveys can be undertaken on foot to understand the physical features of the site as well as the artifacts – in short, all the material cultural remains of the site or the culture (Johnson and Millett, 2013). The site is then located according to Global Positioning Systems on a map or within a geographic information system (GIS), and its location is established and officially reported.

As a branch of engineering, surveying is the science that accurately determines the shape, area (size), and positions of a site's surface by the measurement of certain points. It includes the measurements of horizontal distances and elevations, and determines the direction and relationships of angles by the use of optical instruments. There are several principal classes of survey: *geodetic* surveys, *photogrammetric* (picture-measured) aerial surveys, and *plane* surveys. Geodetic surveys are used for the study of large areas like countries or continents, and take into account the curvature of the earth's surface. Satellite photographs have become important to the archaeological analyses of large landforms. Photogrammetric surveys are the science of taking detailed vertical or oblique photographs from an aircraft for the reliable measurement of the earth's features. Plane surveying is used for small areas of archaeological focus in which the curvature of the earth's surface is not measured. In the past few years, archaeologists have begun to conduct digital surveys using plane-mounted LiDAR units that generate an accurate three-dimensional model of the ground surface, archaeological features, bodies of water, and vegetation (Chase et al., 2012; *see* Geospatial Applications in Archaeology). Preliminary survey also consists of a systematic walking over the site recording standing structures and features as well as artifacts and material scatters. Depending upon what type of survey is to be conducted, artifact concentrations may be collected, processed, and studied to indicate the type of site and its artifact range. The survey should establish the site size, what kind of site it may be, its time span, and, if possible, its depth of deposit.

Depending on the site, three principal types of survey are usually conducted before excavation: (1) the land survey for the proper planning and laying out of the site which includes the site grid; (2) the topographic survey which is the measurement of elevations representing the surface of the area; and (3) the ongoing excavation survey, which is made during excavation to control the location of archaeological structures, features, and objects and providing these with elevations (Dupras et al., 2011). Anomalies and physical features are then thoroughly mapped and described as well as factors like vegetation and soils.

Decisions then may be taken as to what additional surveying equipment will be used to look at the subsurface features. Geophysical surveying techniques employ devices that either transmit an electromagnetic wave into sediments and record

the round-trip travel time of these waves or passively receive ambient electromagnetic waves (Oswin, 2009). Such devices include ground-penetrating radar, electrical resistivity meters, and magnetometers. These devices and others can identify subsurface anomalies that can either be mapped using only geophysical devices or tested through excavation. Only when the comprehensive survey has been completed can the researcher decide if excavation is justified, and if it is, where it is to begin and what course it is to take.

Excavation and Ethical Considerations

Even before a site is excavated, preservation, and consolidation measures must be seriously considered and implemented. How will the site be left after the excavation is completed? As excavation is destruction, one of the most important considerations is the site's stratigraphy, its depth of deposit, and its characteristics. We need to conserve and protect the excavated sites and monuments from all periods and in all parts of the world. Furthermore, we need to protect sites against the threats from human intervention, mass tourism, and building development, as well as from the elements of natural deterioration and neglect. Looting is an age-old problem. There are those who clandestinely excavate to sell artifacts at an enormous profit. Although artifact-rich countries have the legislation to prevent the illegal excavation and export of their country's patrimony, they lack consistent regulatory measures to police looters, dealers, and the exporters of their antiquities. If we are to preserve our common human heritage, the protection of cultural patrimony has to become everyone's priority.

There have been changes in the ethical standards observed by archaeologists today. Ethical standards in archaeology may be defined as the obligations of a professional or an amateur to the excavation; to the country, state, or area in which it is located; to the public at large; and to fellow archaeologists (Scarre and Scarre, 2006). Today's archaeologists strive to work and live according to a high standard of behavior, applying their knowledge to enhance the reputation and achievements of archaeology as a whole and for the benefit of human knowledge. Many archaeological associations have drawn up ethical codes that have been accepted by international archaeological societies: the codes of the Archaeological Institute of America, the Register of Professional Archaeologists, and the Society for American Archaeology.

Who will be responsible for the care and upkeep of the site over time? And what should be done to care for site artifact collections? Where are they going to be stored, who is going to study them, and on what time schedule? What is to be their final disposal – a museum, reburial, and, if the latter, where will they be buried? Today the archaeologist has the responsibility of addressing such issues and these should be best addressed before excavation takes place (*see* Archaeology, Politics of).

Excavation

Excavation is the deliberate recovery of buried objects in relation to the layer or stratum and other associated objects in its original deposit. The principle of excavation is that artifacts, nonartifactual materials, and architectural features are not randomly distributed, but are deposited by people. Their position in the earth reflects the cultural, economic, and social behavior of humankind that can be studied by means of excavation. This is the basis for the functional reconstruction of the site. It is only through excavation that the archaeologist can collect and extract a wide range of physical information or obtain a sense of the people who lived, worked, and perhaps worshipped at a site. The archaeologist does this by describing the evidence the people left behind – the architectural features and objects. It is only through excavation that we can align these physical features in time and space so that the functional reconstruction of the processes involved can be determined. In this way we can better understand and interpret the causes, changes, and processes that the ancients used to bring about change.

Excavation is the method of removing objects and uncovering stationary features that have been concealed by later deposits (Sharer and Ashmore, 2003). It involves the removal of layers or strata in reverse order to the way in which they were laid down, gradually revealing each successive stage in the history of the people who used the site. Because excavation is the means by which information is unearthed, it must be conducted methodically so that whatever is found can be seen and studied within its own context. The main objectives of excavation are to uncover the position of objects as well as floral and faunal data, and to plot their location on a horizontal plan (Sharer and Ashmore, 2003). It also involves the meticulous recording and sequencing of successive layers on the vertical section by position and depth measurements. Finally, the goal is to relate such data to the total environment of the site and to assess its relationship to other sites in the area.

Modern Developments in Archaeology

Major developments in archaeology took place in the late 1990s with the advent of computer technology: remote sensing using images from airplanes, satellites, and drones; web pages; virtual reality; computer simulations; digital recording devices like cameras and tripod-mounted LiDAR; digital record keeping systems like relational databases; geospatial analytic softwares like GIS (Hritz, 2014); and the greater use of the hard sciences including DNA analysis (Grosman et al., 2014).

In the 1960s, American archaeologists began to challenge the traditional archaeological academy and argued that the discipline had become simplistic. As a result of this critique the 'new archaeology' was developed, emphasizing the importance of the processes within the society being studied. This was known as 'processual archaeology,' and it emphasized the value of objective, logical explanation, and the descriptive processes that can produce testable assumptions. They believed that culture should be studied as sets of systems and subsystems as they relate to the environment, the economy, and patterns of subsistence. They also emphasized that the interrelationships between different sectors of the society should be explainable so that specific developments could be established throughout time (Harrison-Buck, 2014).

They failed to stimulate the traditionalist majority who practiced archaeology as they had before and who believed that there were no universal laws governing human behavior.

The merit to the then-new approach to archaeological data opened the way to a vast array of new theories including Marxist, neo-Marxist positivist, structuralist, poststructuralist, postprocessualist, deconstructionist, and probabilist views along with other theoretical constructs (McGee and Warms, 2011). The postprocessual school of archaeologists adheres to aninterpretative archaeology, which insists on emphasizing history, philosophy, and, whenever possible, literary studies. Interpretative archaeologists place importance on the fundamental uniqueness and diversity of culture, believing there is no one governing law that can apply either to archaeological research or to the interpretation of the past. These theories hold some truth, provide stimulating hypotheses for our explanation of the past, and are used by many who thoughtfully interpret the probabilities of the evidence. Now archaeologists are wrestling with the ideas of multivariate explanations – and many continue to build idealized models for their interpretations.

Conclusion

By and large, archaeologists are dedicated researchers and teachers. As for future directions in theory and research, evaluation can be said to drive archaeological standards. By clearly specifying the outcomes in measured ways, archaeologists will change the way they excavate and publish. In the training of students, field directors hope they will train people who possess the qualities and attributes desired in a competent archaeologist. As an archaeological project is expected to foster a shared sense of a shared mission between the field researcher and the principal investigator, both will strive to reach a common goal. Archaeology is best taught through active learning processes, and principal investigators help students by gauging their progress and identifying and overcoming barriers to the excavation as a whole. The developmental process of archaeology is most effective when the goals are known so that the whole team can work together toward the shared goals, collaboratively devising strategies to overcome any obstacles. Our understanding of the past is in a constant state of change due to new discoveries and advances in our knowledge. It will never be finished.

Today archaeology has developed into a complex discipline with a vast literature and a range of specialties that are well beyond the competence of a single researcher. More and more data are being produced, museums are overwhelmed with objects, and, even with the advent with the computer, many artifacts have remained unanalyzed or unregistered. There is a crisis in the discipline due to the fact so much data have been produced, and as excavation involves the destruction of the site, archaeologists must be aware that the discipline may self-destruct unless it puts its house in order. It faces considerable challenges and risks ahead and it must develop an international dialogue among its leadership that is committed to a strategy which will ensure the saving of the past for the future.

See also: Archaeology, Politics of; Archaeology, Theory in; Ceramics in Archaeology; Chronology, Stratigraphy, and Dating Methods in Archaeology; Environmental Archaeology; Geoarchaeology; Geospatial Applications in Archaeology.

Bibliography

Bahn, P.G., Renfrew, C., 2005. In: Renfrew, Colin, Bahn, Paul (Eds.), Archaeology: The Key Concepts. Routledge, London; New York.

Blumenschine, R.J., Stanistreet, I.G., Masao, F.T., 2012. Olduvai Gorge and the Olduvai landscape paleoanthropology project. Journal of Human Evolution, 63 (2, Sp. Iss. Sl), 247–250.

Boaretto, E., Poduska, K., 2013. Materials science challenges in radiocarbon dating: the case of archaeological plasters. JOM: The Journal of The Minerals, Metals & Materials Society (TMS) 65 (4), 481–488.

Chase, A., Chase, D., Fisher, C.T., Leisz, S.J., Weishampel, J.F., 2012. Geospatial revolution and remote sensing LiDAR in mesoamerican archaeology. Proceedings of the National Academy of Sciences 109, 12916–12921.

Cline, E.H., 2009. Biblical archaeology: a very short introduction. In: Cline, Eric H. (Ed.). Oxford University Press, Oxford; New York.

Crowell, B.L., 2007. Nabonidus, as-Sila', and the beginning of the end of Edom. Bulletin of The American Schools of Oriental Research 348, 75–88.

Deetz, J., 1996. In Small Things Forgotten: An Archaeology of Early American Life. Anchor Books/Doubleday, New York.

Dupras, T., Schultz, J., Wheeler, S., Williams, L., 2011. Methods of geophysical survey. In: Forensic Recover of Human Remains: Archaeological Approaches, second ed. Taylor & Francis Group, LLC.

Greene, K., 1995. Archaeology: An Introduction: The History, Principles and Methods of Modern Archaeology. University of Pennsylvania Press, Philadelphia, PA.

Grosman, L., Karasik, A., Harush, O., Smilansky, U., 2014. Archaeology in three dimensions: computer based methods in archaeological research. Journal of Eastern Mediterranean Archaeology and Heritage Studies [serial online] 1 (2), 48–64.

Gamble, Clive, 2008. Archaeology: The Basics, second ed. Routledge, London; New York.

Harrison-Buck, E., 2014. Anthropological archaeology in 2013: the search for truth(s). American Anthropologist 116 (2), 338–351.

Hicks, D., 2013. Four-field anthropology Charter Myths and time warps from St. Louis to Oxford. Current Anthropology 54 (6), 753–763.

Hodder, I., 1999. The Archaeological Process: An Introduction. Blackwell, Oxford, UK.

Hodder, I., 1991. Reading the Past: Current Approaches to Interpretation in Archaeology. Cambridge University Press, Cambridge, UK.

Hritz, C., 2014. Contributions of GIS and satellite-based remote sensing to landscape archaeology in the Middle East. Journal of Archaeological Research 22 (3), 229–276.

Insoll, T., 2007. Archaeology: the conceptural challenge. In: Insoll, Timothy (Ed.). Duckworth, London.

Johnson, P., Millett, M., 2013. Archaeological survey and the city. In: Johnson, Paul, Millett, Martin (Eds.). Oxbow Books, Oxford, UK. David Brown Co., Oakville, CT.

Kelley, D., 2003. The rise of prehistory. Journal of World History 17 (1).

Lalleman, P.J., 2008. The old testament and archaeology with a personal top ten of discoveries. Evangel 26 (3), 83–88.

McGee, Jon R., Warms, Richard, 2011. Anthropological Theory: An Introductory History. McGraw-Hill Humanities/Social Sciences/Languages, New York, New York.

Moshenska, Gabriel, 2013. Why the history of archaeology matters: a response to Mees. Post-Medieval Archaeology 47 (1), 247–251.

Oswin, J., 2009. A Field Guide to Geophysics in Archaeology. Praxis Publishing, Chichester.

Pettitt, Paul, 2009. The rise of modern humans. In: Scarre, Chris (Ed.), The Human Past: World Prehistory and the Development of Human Societies, second ed. Thames and Hudson, London, pp. 149–151.

Porčić, Marko, 2013. The goodness of fit and statistical significance of seriation solutions. Journal of Archaeological Science 40 (12), 4552–4559.

Renfrew, C., Bahn, P.G., 2000. Archaeology: Theories, Methods and Practice, third ed. Thames & Hudson, New York.

Scarre, G., Scarre, C., 2006. The ethics of archaeology: philosophical perspectives on archaeological practice. In: Scarre, Chris, Scarre, Geoffrey (Eds.). Cambridge University Press, Cambridge, UK; New York.

Sharer, R., Ashmore, W., 2003. Excavation. In: Archaeology: Discovering Our Past, third ed. McGraw-Hill Higher Education, New York, New York.

Schiffer, M.B., 1987. Formation Processes of the Archaeological Record. University of Utah, New Mexico Press, Albuquerque, NM.

Trigger, B., 1989. A History of Archaeological Thought. Cambridge University Press, Cambridge, UK.

Toth, Nicholas, Schick, Kathy, 2009. African origins. In: Scarre, Chris (Ed.), The Human Past: World Prehistory and the Development of Human Societies, second ed. Thames and Hudson, London, pp. 67–68.

Wallis, R.J., 2013. Animism and the interpretation of rock art. Time & Mind: The Journal of Archaeology, Consciousness & Culture 6 (1), 21.

Willey, G.R., Sabloff, J.A., 1995. A History of American Archaeology. W.H. Freeman, San Francisco, CA.

Survey Research Centers and Companies

Helmut K Anheier and Nikolas Scherer, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by P.P. Mohler, volume 22, pp. 15311–15314, © 2001, Elsevier Ltd.

Abstract

Social scientists seek answers by way of research methods. One such method is survey research, which involves the collection and analysis of data obtained from individuals, usually through questionnaires. A specific 'know-how' and infrastructure is needed to conduct surveys and obtain reliable and valid data. Survey research centers and companies provide such expertise and the services. The following entry presents some of the most important survey research centers and companies in the US and Europe. By referring briefly to their focus and history, the article reviews their standing in the field and its contribution to social sciences research methods generally.

Survey research draws on different disciplines. Insights from economics, sociology, psychology, statistics, geography, and other related subjects have all made their contributions to a highly interdisciplinary and cooperative field. The early roots of survey research might go back to the late nineteenth century; its takeoff, however, was not earlier than in the 1940s often supported by (radio) audience research and the American military during World War II. At the same time, the impact of some survey centers and companies for the development of survey research and their subsequent standing in the field cannot be overestimated. This article reviews eight centers and companies in three different countries. It distinguishes academic and public policy-driven survey research from survey research conducted by private companies. Although their difference is not clear-cut in practice, the application of academic methodological standards into the applied survey research process and the training and education of survey professional is what tends to distinguish survey centers from survey companies and justify such a distinction. Among the masses of survey centers and companies we will introduce four survey centers and companies as current 'blueprints' of contemporary survey research. These are, for (academic/national) survey centers the National Opinion Research Center (NORC, Chicago), the Institute for Social Research (ISR, Ann Arbor), the National Centre for Social Research (NATCEN, London), the Leibniz-Institute for the Social Sciences (German Social Science Infrastructure Institutes (GESIS), Mannheim) and for private survey companies Gallup (Washington D.C.), Westat (Rockville), Ipsos MORI (London), and Taylor Nelson Sofres (TNS) Emnid (Bielefeld). By presenting briefly their history and their contribution to survey research, this article will review their standing in the field and the future development of survey research. Since the emphasis is on survey research, this article will next give a short overview of what distinguishes surveys from other methods of data collection.

Survey Research – Definition and Brief Overview

Surveys are understood here as "a method to gather information from a sample of individuals" (Scheuer, 2004), using interviews as the source of information and systematic sampling as the base for selecting information sources (cf De Heer et al., 1999). They are instruments to identify distributions of societal characteristics. Questions or items in a survey are understood as indicators pointing to a latent construct (e.g., opinions, attitudes) or dimensions. Survey research is not interested in information about individuals. Instead, it is interested in information on larger groups, even societies. These target groups are termed populations. The information gathered by questioning individuals is used to make inferences to the population being studied.

At the first sight conducting surveys seems simple: it is basically about asking questions, recording answers and inferring characteristics, and dimensions from a small sample of individuals to a larger population. However, survey research is fairly complex. It involves specific methods and techniques of sampling and measurement. It has to deal with problems such as nonresponsiveness, social desirability bias, or construct validity. The way a question is worded, for instance, has subsequent effects on how a question is answered. To deal with these issues survey research has from its very beginning embarked on range of disciplines such as cognitive psychology and statistics. Apart from these methodological challenges survey research also involves organizational and administrative challenges. While a survey conducted in a closed classroom should not pose a greater problem for a survey researcher it looks entirely different when a survey is conducted in a national, regional, or even global space. This requires not only different methods how to reach the individuals that come with own problems but also enough interviewers to collect the data and a number of supervisors to oversee the whole process. This complexity requires adequate technical, organizational, and administrative capacities. Large commercial survey companies such as Gallup, Ipsos MORI, and RAND with various site offices and thousands of employees in numerous countries provide these capacities and are able to conduct large-scale cross-national surveys for the public and private sector. They partly compete with other large noncommercial (academic/national) survey centers of global research such as NORC, ISR, and GESIS that match these capacities.

Distinguishing Survey Centers from Companies

The survey industry can be thought of as compromising at least two distinct actors: survey centers and commercial firms.

What distinguishes these are basically whom they do research for and what type of research they do. While commercial survey companies largely have other private firms as customer, and concentrate on market research as well as research for media corporations, noncommercial survey centers serve mainly academic communities, governments, or public interests more in general. As such the latter engage more in methodological and long-term policy and social research. By definition commercial survey companies are more likely to sell the information gathered, while survey centers distribute the information mostly free to a wider public audience.

Noncommercial survey centers can be thought of as compromising two subtypes that are closely related: academic and national survey centers. In contrast to commercial survey companies, academic survey centers do not only promote rigorous methodological standards and perspective but they engage also in methodological research. Their research goes beyond the 'we use the method because it works to find out XYZ' as it also aims to find out 'why and how this method works to find out XYZ.' Academic survey centers, such as NORC, ISR in the US, GESIS in Germany, or RECSM in Spain are, thus, also known centers of methodological excellence. They often host long-term surveys and collaborations on national, regional, and international issues.

National survey centers focus mainly on domestic issues of public interest and collect information on public attitudes, opinions, demographic characteristics such as age, gender, urban/rural, and other characteristics of national residents. The NORC and the ISR in the US, the National Centre for Social Research (NATCEN) in the UK, the Norwegian Social Science Data Services, GESIS are prime examples and carry out major national surveys. Their annual studies, the General Social Survey (GSS, conducted by NORC), the German General Social Survey (conducted by ZUMA/GESIS), or the British Election Studies (conducted by NATCEN) feed into many social science research projects, and inform public policy making.

ZUMA and NORC are also founding members of the International Survey Programme (ISSP). The ISSP consists of cross-national collaborations that bring together numerous survey centers and projects, thereby adding a cross-national perspective to the individual, national studies (www.issp.org). The ISR, in turn, took the lead in conducting the World Value Survey (WVS). The WVS brings together a global network of social scientists and provides a cumulative study of almost 100 countries on changing values and their impact on social and political life. It is the largest ongoing, cross-national investigation of human beliefs and values (www.worldvaluessurvey.org).

Generally speaking, what differentiates survey institutions are primarily the target audience (private firms vs public agencies), the survey purpose, the publication of methodological research, the temporal aspect of the surveys (short-term vs long-term), and the free distribution of major nation-wide data sets. However, it might be noteworthy that these distinctions are not clear-cut in practice. In search for funding national and academic survey centers have also been charging fees for their services or set up own service centers within an academic department instead of relying only public research grants and contracts (Guterbock, 2010).

There are numerous survey centers and thousands of commercial survey companies across the globe. No reliable figure of their actual number exists. Since no other country has as many survey centers and companies as the US (Converse, 2009; Guterbock, 2010), this article will concentrate in particular on American centers and companies but makes reference to Europe.

Survey Centers

National Opinion Research Center, Chicago

NORC is one of the largest national survey centers in the US and is based at the University of Chicago (www.norc.org). It was established in 1941 as a nonprofit organization through a grant from the University of Denver and the Field Foundation. It moved to the Chicago campus in 1947. The center conducts research for public and private entities throughout the world. Its key areas of research focus on public issues such as health, education, economics, crime, justice, energy, security, and the environment. NORC is primarily known for its large national surveys and its innovations in survey methodology. Since 1972, NORC conducts the GSS. The GSS is funded by the National Science Foundation (www.nsf.gov) and is a general data-collection program that monitors the structure and development of American society over time (http://www3.norc.org/GSS+Website). It is one of the most frequent cited sources of information and, as such, has been for the benefit of the sociological profession at large. The GSS is released every year. Many of the core questions have been unchanged since 1972 to allow for time-trend studies. Interested parties have free access to historical and up-to-date survey material, which they can analyze on their own terms. Other strongholds of NORC are its methodological research on measurement errors.

Institute for Social Research, Ann Arbor

The ISR belongs to the University of Michigan and is the world's largest academic survey centers in the world (http://home.isr.umich.edu). It was established in 1949 and conducts research for the public and private sector. From its very beginning the ISR understood itself as a laboratory of and for the social sciences and advocated interdisciplinary research. Today it compromises five centers: The Center for Political Studies, the Inter-university Consortium for Political and Social Research, the Population Studies Center, the Research Center for Group Dynamics, and the Survey Research Center. The academics come from a broad range of fields including sociology, psychology, anthropology, economics, sociology, history, political science, and statistics. The ISR is known for its broad expertise on broad range of subject matter. Key areas of research focus include drug use, poverty, health, group processes and social cognition, aging, human development, political participation, voting behavior, institutions and political processes, and so on. The ISR has conducted groundbreaking methodological research and paved the way of modern survey research in areas such as survey sampling theories and methods and survey cognition processes. Among its major achievements are the development of the Likert Scale, one of the most widely used measure of attitudes and options academic, government,

and private sector research today and foundation of the field of organizational behavior. The ISR conducted several surveys which had major impacts on politics and society: In 1954, for example, it fielded the double-blind experimental trials for the Salk polio vaccine on a national sample of the US schoolchildren, revealing that vaccine was both safe and effective. Another impact area is data storage, education, and training. The Inter-university Consortium for Political and Social Research, which was founded in 1962, is one of the world's largest archives for quantitative social science data. It provides interested research access to thousands of data files. Moreover, it offers different education and training programs for graduates, postgraduates, and researcher in all phases of survey research. As such, the ISR contributes to secure and expand the existing knowledge base in survey methodology and methods.

National Centre for Social Research, London

The NATCEN is Britain's largest nonprofit survey research center. It counts about 250 employees and was founded in 1969 as charity governed by a Trust Deed (http://www.natcen.ac.uk). Its goal is to disseminate social, economic, and political research to the public for debate and educational purposes. NATCEN provides services to both the public and private sector agencies. Its customers include the UK government departments, think tanks, charities, businesses, universities, and other research organizations. The scope of NATCEN's research is broad and focus on issues such as health, demographics, crime, education, political attitudes, transport, income, and housing. Within the tradition of British social research NATCEN conducts quantitative and qualitative studies. Researchers at NATCEN have produced a range of highly applicable survey studies. Among their major accomplishments are the British Social Attitudes Survey (BSA), Britain's general social survey. The BSA has been carried out since 1983 and involves about 3000 in-depth interviews about what is like to live in Britain. BSA informs the development of British public policy and is well used by opinion leaders and social commentators as an important barometer of public attitudes. NATCEN forms the base for a number of mid-range and long-range programs and projects such as the Centre for Research into Elections and Social Trends. Funded by the Economic and Social Research Council the research center is a joint venture of the Department of Sociology in Oxford, on the one hand, and NATCEN on the other. The Scottish Centre for Social Research, in turn, is an integral part of NATCEN and conducts applicable social and political research reflecting the different social and political environment in Scotland. Apart from these projects NATCEN has recently launched its NATCEN learning platform, which provides training and education for interested researcher and the policy community interested in survey methods and methodology. It thus contributes to secure and boost the quality of applied survey research in the UK.

GESIS: Leibniz-Institute for the Social Sciences, Mannheim

With more than 250 employees GESIS is the largest social science research institute in Germany (www.gesis.org). GESIS was established in 1986 as the publicly funded German Social Science Infrastructure Services and consisted of a close collaboration between the three independent institutes: The Social Science Information Centre (IZ) in Bonn, Central Archive for Empirical Social Research in Cologne (ZA), and Centre for Survey Research and Methodology (ZUMA) in Mannheim. In 2007, they merged into one institute and rebrand itself as GESIS – Leibniz-Institute for the Social Sciences. GESIS is fully financed through public funds as provided by the federal government and the German Länder although foundations and other sponsors are increasingly involved in single project funding. Being primarily publicly funded GESIS is obliged to have a strict supportive function for academic research. Accordingly, the institute provides expertise in all production processes of social sciences research. Key areas of research concentrate on basic research in survey methodology, social research on attitudes, behavior and social structure, applied informatics and information sciences. Moreover, GESIS provides research-related services such as data storage, consulting, education and training in survey design, methods and archiving. All services are free or nominally charged. GESIS, as such, provides a comprehensive and unique infrastructure for academics and social sciences as a whole. Apart from the particular institutional design GESIS and its different organizational units have contributed in a range of large-scale surveys and cross-national projects. Among these are the ISSP, the European Social Survey, the German General Social Survey (ALLBUS) and the Programme for the International Assessment of Adult Competencies. The GESIS's Central Archive for Empirical Social Research (ZA), founded in 1960, is the German national archive for quantitative survey data and played a major role in establishing the international archive network. The Social Science Information Centre (IZ) keeps track of German social sciences publications, ongoing research projects and serves as a platform for information exchange between Eastern and Western Europe. GESIS is without doubt Germany's 'engine room' for social and political survey research.

Survey Companies

Gallup, Washington D.C.

Gallup is among the largest private market and opinion research companies of world (www.gallup.com). It is employee-owned and engages primarily in customized public opinion polling, market research, and consulting. Its clients are major businesses and organizations from around the world. Gallup employs around 2000 people and has offices in almost 40 countries. It was set up in 1935 by George Gallup in Washington D.C. Gallup was one of the first companies that engaged professionally and at large in survey research. It became suddenly famous in 1936 as it rightly predicted the victory of Franklin D. Roosevelt in the US presidential election, thereby contradicting the popular polls of that time that predicted a landslide victory for Alfred Landon. The company relied on a relatively small (5000 respondents) but carefully elected sample from the US population using face-to-face interviews instead of usual indiscriminate mail surveys that reached more better-off households and got it

right. Gallup was the first commercial survey organization using nation-wide probability sampling. Gallup's correct prediction had substantial influence on the practice of sampling. Gallup's methodology were copied not only in the US but in the emerging polling in Europe and Australia (Fink, 2003: 108). Gallup was also one of the pioneers engaging in global polling. In the early years the Gallup International Research Institute held annual meetings at which various affiliates reported and analyzed what the world's people thought about the multitude of sensitive and controversial issues that affected their lives such communism, capitalism, religion, etc. Since 2005, it conducts the Gallup World Poll bringing together survey data from more than 160 countries measuring attitudes, opinions, and trends. The insights have been used by different academic and public institutions such as the UN, the World Bank, or the OECD. Although Gallup regularly conducts political, social, and economic opinion polls and data analyses for large media corporations such as CNN or USA Today, market research has always been a second cornerstone of the corporation. Gallup moved more into business consulting in the 1990s as the company was purchased by Selection Research International in 1988 following the death of founder George Gallup. Today Gallup engaged also in employee engagement, customer engagement, talent management, and well-being using polls. It also issues its own journal, the *Gallup Business Journal* and has published several books. Some of the content and data produced by Gallup is free on its Web site or accessible via its customers, full data set is, however, only accessible via a paid subscription. Moreover, Gallup contributes to the Survey Research and Methodology program at the University of Nebraska-Lincoln through grants, fellow- and studentships.

Gallup's main contribution was to establish nation-wide probability sampling. Gallup experimented with different survey methods and helped establish surveys as a systematic method of data collection.

Westat, Rockville

Along with Gallup and other big corporations such as RTI International, RAND, Survey Sampling International, Nielsen, Opinion Research Corporation, and Abt Associates, Westat is one of the largest private survey companies of the US (www.westat.com). Westat was founded in 1963 by a professor and two former students. Its establishment and rise has its historical context in the massive demand for statistical information by the US government and businesses in the 1960s (Groves, 2011: 865). Today it employs around 2000 people at its headquarter in Rockville while it maintains several research offices in the US and abroad, in particular in Asia. Westat is employee-owned and conducts research for the public and the private domain. Among its customers are various departments and agencies of the federal US government but also foundations, businesses, states, and local governments. Key research areas concentrate on issues such as on health and medicine, education and literacy, social policy, workforce, recruitment and retention of military personnel. In addition to its survey research, Westat engages also in custom and evaluation research and communications development. Moreover it developed WestVar, a statistical software program. Westat has a broad expertise for all phases of survey research – sample design; instrument design; execution of mail, telephone, web, and personal interview data collection; data processing; analysis and report writing. Westat is well known for collecting the annual National Assessment of Educational Progress, the National Health and Nutrition Examination Survey, and the National Medical Expenditure Penal Survey which all inform the US public policy.

Ipsos MORI, London

Ipsos MORI is a leading private research survey company in the UK and was founded by a merger of Ipsos UK and MORI in 2005 (www.ipsos-mori.com). It is part of the Paris-listed Ipsos Group and, thus, part of one of the largest private survey research organizations in the world. Its offices distributed in more than 80 countries. Ipsos MORI conducts research for the public and the private sector. It is specialized in market research, media research, marketing and advertising research, reputation research, loyalty research, social and political research. Ipsos MORI investigates industry structures and trends, tries to understand what motivates and influence costumers and engages in research to test market products. It does so on a national and cross-national basis. Ipsos MORI's Social Research Institute is often consulted by the UK government and other major and smaller public and private authorities to inform public debate and decision making. It examines public attitudes toward key public services and provides information to assist social policy makers.

Researcher at Ipsos MORI uses a wide range of quantitative and qualitative methods. Ipsos MORI's predecessor RLS were leading way in using Computer-Assisted Telephone Interviews on a large-scale basis and were among first to use Computer-Assisted Personal Interviews. Moreover, RLS were among the pioneers in the development of regional surveys as The European Businessman Readership Survey, The Asian Businessman Readership Survey, and the British Businessman. Regional surveys involve an added complexity in terms of research design, database modeling, overseeing and coordinating the survey. MORI, instead, was known for its research on public attitudes to key public services that came with trends toward privatization and outsourcing of public services and its research on social policies and the trend of evidence-based public policy making. Ipsos MORI has been at the forefront of using new digital tools such as Internet surveys as well as mobile research and biometric tools.

TNS Emnid, Bielefeld

TNS Emnid is the oldest and one of leading private survey companies in Germany (www.tns-emnid.com). It was established in 1945 and serves since 1997 as the national branch of the larger London-based TNS Group, which is, in turn, owned by the London stock exchange-listed WPP Group since 2008. The WPP Group is by 2014 the largest market research holding of the world (www.wpp.com). TNS Emnid counts about 50 employees and conducts political, social, and media research. It does so primarily but not exclusively on the national level. To its customer belong major public entities and private organizations, including businesses, foundations, and other

survey research centers. TNS Emnid investigates political opinions and trends such as the weekly Sonntagsfrage ('Which party would you vote for if general elections would be held next Sunday?') and social issues such as crime, religion, cross-cultural attitudes and engages in advertisement research. TNS Emnid pioneered political opinion polling in postwar (West) Germany which became popular in the 1960s and has been frequently contracted by political actors and Germany's largest media corporations. Since 1955, TNS Emnid is member of the Gallup International Association and contributed to different of cross-national surveys such as the 'State of the Mankind' – study. TNS Emnid mostly known for its ad hoc polls following (inter-)national political events and uses primarily Computer-Assisted Telephone Interviewing and Computer-Assisted Personal Interviewing based on random dialing. More recently they have turned to mobile research tools and Internet surveys.

Conclusion: Surveys, Survey Centers and Companies, and the Social Sciences

Survey research centers are institutions that provide databases, methodological research, education and training for clients of the academic community, and are increasingly international in scope. Some academic survey centers such as ISR and GESIS have now combined social analysis, data archiving, storage, and information sciences to secure a functioning infrastructure of the social sciences at large.

More recently survey companies have shown a tendency to merge. Some of the originally national survey companies form now or part of large global survey groups as it is the case with Ipsos MORI and TNS Emnid mentioned above. Such global mergers cannot be observed in the academic realm, however. Instead, the smaller academic survey centers have created and relied on global networks and collaborations. These collaborations demand different capacities, require substantial financial means and might involve a future reconfiguration of the roles between survey centers and between survey centers and companies.

A challenge that might affect future survey research and their infrastructural carriers are the threats and opportunities that arise with the Internet. The Internet has enabled the collection in massive amounts of data (Big Data). Private survey companies have been at the forefront in using new modes of knowledge production such as Web-based tools or mobile research tools. Overall, it remains to be seen how survey research methods and techniques will change as Big Data approaches evolve.

See also: Applied Social Research, History of; Experimental Methods in Survey Research in Demography; Health Surveys; International Social Survey Programme; Large-Scale Social Surveys, Analysis of; Sample Survey Methodology, History of; Surveys and Polling: Ethical Aspects; Surveys: Question Wording and Response Categories.

Bibliography

Converse, J., 2009. Survey Research in the United States: Roots and Emergence, 1890–1960. Transaction Publishers, Piscataway, NJ.

De Heer, W., De Leeuw, E., van der Zouwen, J., 1999. Methodological issues in survey research: a historical review. Bulleting de Methodologie Sociologique 64, 25–48.

Fink, A., 2003. The Survey Handbook, second ed. SAGE Publications, Thousand Oaks, CA.

Groves, R., 2011. Three eras of survey research. Public Opinion Quarterly 75 (5), 861–871.

Guterbock, M., 2010. The future of academic research organization. Survey Research 41 (1).

Mohler, P., 2001. Survey Research: National Centers. In: Smelser, N., Baltes, P. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Oxford, pp. 15311–15314.

Scheuer, F., 2004. What Is a Survey?. http://www.amstat.org/sections/srms/pamphlet.pdf (01.08.2014).

Relevant Websites

http://www.degruyter.com – Journal of Official Statistics.
http://poq.oxfordjournals.org – Public Opinion Quartely.
http://surveypractice.org – Survey Practice.

Associations

www.aapor.org – American Association for Public Opinion Research (AAPOR).
www.amstat.org – American Statistical Association (ASA).
www.asc.org.uk – Association for Survey Computing (ASC).
www.esomar.org – European Society for Opinion and Market Research (ESOMAR).
http://www.europeansurveyresearch.org – European Survey Organization (ESRA).
https://www.mrs.org.uk – Market Research Society (MRS).
http://wapor.org – World Association for Public Opinion Research (WAPOR).

Surveys and Polling: Ethical Aspects

Constance F Citro, Committee on National Statistics, National Research Council/National Academy of Sciences, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Singer, volume 22, pp. 15314–15318, © 2001, Elsevier Ltd.

Abstract

Polls and surveys gather data from respondents. Two broad aspects of ethical practice for them are: (1) the extent to which research designs, instruments, data collection procedures, and analysis techniques permit valid and reliable conclusions from respondents' answers; and (2) the researcher's obligations toward respondents, sponsors, and the larger public. The American Association for Public Opinion Research (AAPOR) Code of Professional Ethics and Practices defines best practices for survey research (e.g., scientific sampling) and unacceptable practices (e.g., selling under the guise of research). The World Association for Public Opinion Research (WAPOR) has similar ethical standards. Obligations to sponsors include documentation of methods used, and, increasingly for U.S. federal agency sponsors, sharing of research data. Obligations to the public include disclosing key information about the research (e.g., sponsor, size of sample). Obligations to respondents (embodied for research sponsored by U.S. federal agencies in Federal Policy for the Protection of Human Research Subjects, or Common Rule) include providing the opportunity for informed consent to participate and protecting the confidentiality of responses.

Polls and surveys are used to investigate opinions and attitudes and to obtain factual information from respondents. Two broad aspects of ethical practice with respect to polls and surveys can be distinguished. The first pertains to the extent to which research designs, instruments, data collection procedures, and analysis techniques permit valid and reliable conclusions from respondents' answers. The second pertains to the researcher's obligations toward others, not only toward respondents, but also toward sponsors and the larger public.

Polls and surveys are essentially indistinguishable, although the terms have somewhat different connotations. The word 'poll' is typically used for a survey of public opinion, such as whether the public thinks a president is doing a good job. This usage accords with another dictionary definition of 'poll,' which is 'the process of voting in an election.' The word 'survey' is typically used for a more comprehensive examination of one or more topics, such as a survey of household income and program participation or a survey of business research and development. This usage accords with a dictionary definition of 'survey' as 'to examine or look at comprehensively.' Because methodological issues of validity and reliability and issues of obligations of researchers toward others are the same, the present discussion of ethical practices is intended to apply to both polls and surveys.

While the discussion is framed primarily in terms of the code of ethics of one organization in a particular country – the Code of Professional Ethics and Practices (hereafter 'Code of Ethics') of the American Association for Public Opinion Research (AAPOR), a professional organization established in 1947, which includes academic, commercial, and government survey researchers among its members – the principles presented here apply more generally. Indeed, the World Association for Public Opinion Research (WAPOR), also established in 1947 with members in over 60 countries, has similar standards of ethics and good practices for opinion polls. (The WAPOR Code of Ethics was last revised in December 2011.) Because legal requirements for protecting respondent confidentiality and adhering to other ethical imperatives, as well as the ability to do so in practice, vary from one country to another, such variations are not discussed here. Also excluded from the discussion are the fundamentally different ethical issues raised by nonstatistical uses, such as law enforcement, program administration, or regulation, of data that were originally gathered from polls or surveys for statistical purposes (see National Research Council, 2013b; Seltzer and Anderson, 2008).

Standards for the Conduct of Research

AAPOR's Code of Ethics obligates members to adhere to the Code and includes provisions for reporting and reviewing violations, with appropriate punishment, including public censure. The Code, last revised in May 2010, is necessarily couched in general terms, simply asserting, for example, that members will "exercise due care in developing research designs and instruments, and in collecting, processing, and analyzing data, taking all reasonable steps to assure the reliability and validity of results." In May 1996, AAPOR adopted two more specific statements, one defining a set of performance standards that embody 'best practices' in survey research; the other, in contrast, defining a set of practices that are universally condemned. Among 'best practices' are scientific methods of sampling, which ensure that each member of the population has a measurable chance of selection; follow-up of those selected to achieve an adequate response rate; careful development and pretesting of questionnaires, with attention to known effects on response tendencies of question wording, question order, and interview mode (mail, telephone, Internet, in-person); and adequate training and supervision of interviewers. Among 'unacceptable' practices are fund raising, selling, or canvassing under the guise of research; representing the results of a poll of self-selected or volunteer respondents (for example, readers of a magazine who happen to return a questionnaire, callers to an 800- or 900-number poll, or haphazard respondents to a web-based survey) as if they

resulted from legitimate research; and 'push polls,' which feed false or misleading information to voters under the pretense of taking a poll. For example, a push-poll might ask if the respondent would object to learning that candidate X has been accused of child abuse, has defaulted on a loan, etc. Also among 'unacceptable' practices in the AAPOR Code are revealing the identity of individual respondents without their permission and requiring or soliciting monetary payments from members of the public as part of a research process.

Standards for Dealing with Sponsors, the Public, and Respondents

Obligations to Sponsors

Ethical behavior toward sponsors or clients has several elements, some of which apply to specific categories of sponsors. Applicable to all sponsors is the duty of researchers to undertake only those survey research assignments that can reasonably be accomplished, given the limitations of survey research techniques or limited resources of the researcher or sponsor.

Obligations to U.S. Federal Agency Sponsors

More and more federal agency sponsors are requiring timely sharing of research results and the underlying data, including survey data. For example, the data-sharing policy of the National Institutes of Health (NIH) requires all investigator-initiated grant applications of $500 000 or more in direct costs to propose a data-sharing plan that allows for 'timely release and sharing,' defined as 'no later than the acceptance for publication of the main findings from the final data set.' Final peer-reviewed journal manuscripts reporting on NIH-funded research are required to be accessible to the public via the digital archive PubMedCentral no later than 12 months after publication. Similarly, the data-sharing policy of the National Science Foundation (NSF) stipulates that "investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants." In February 2013, the White House Office of Science and Technology Policy issued a memorandum on 'Increasing Access to the Results of Federally Funded Scientific Research,' which requires all federal agencies that annually fund over $100 million in scientific research to develop a plan for public access to digital data resulting from federally funded scientific research and a plan for public access to scientific publications, using a 12-month postpublication embargo period as a guideline. These policies all stem from the premise that federally funded data collections are of value to and should be 'made available to and useful for the public, industry, and the scientific community.' They all also recognize the imperative to maintain the confidentiality of individual responses as part of a data-sharing or public access plan (National Research Council, 2013b).

Obligations to Commercial or Private Sponsors

For commercial or other private sponsors, it is usually the case that the researcher has a duty to hold information about the research and the sponsor confidential, except as the sponsor expressly authorizes or as other provisions of the AAPOR Code require. Chief among the latter is a requirement that researchers who become aware of serious distortions of their findings in public must disclose whatever is necessary to correct these distortions to the audience to which the distorted findings were presented. In practice, this provision may obligate researchers to issue a corrective statement to a legislative body, a regulatory agency, the media, or another appropriate group as soon as they become aware of the release of distorted results. Moreover, increasingly, researchers who seek to publish findings from a commercially funded project in peer-reviewed journals should be prepared to secure the sponsor's agreement to allow the underlying data (suitably protected against disclosure of individual responses) to be shared with other researchers to permit verification and replication of results.

Obligations to the Public

The AAPOR Code of Ethics addresses the profession's obligations to the larger public by mandating the disclosure of information about a poll or survey whose findings are publicly released. (The Code obligates such information to be made available to a private sponsor at the time the results are delivered.) The elements that must, at a minimum, be disclosed are: (1) who sponsored the survey, and who conducted it; (2) the exact wording of questions asked, including text of any preceding instruction or explanation to the interviewer or respondent that might be expected to affect the response; (3) a definition of the population under study and a description of the sampling frame; (4) a description of the sample selection procedure, making clear whether respondents were selected by the researcher or were self-selected and whether probability or nonprobability methods were used; (5) the size of the sample and, if applicable, completion rates and information on eligibility criteria and screening procedures; (6) a discussion of the precision of the findings, including estimates of sampling error and a description of any weighting or estimating procedure used; (7) which results, if any, are based on parts of the sample rather than the entire sample; and (8) methods, locations, and dates of data collection. Together, these items of information communicate an idea of the error likely to be associated with estimates from a given survey, or, conversely, the degree of confidence warranted by the results. The Code requires much of this information to be included in or to accompany any report of research results, or, in some instances, to be available within 30 days of a request. In short, while the Code encourages researchers to follow best professional practices, it requires them to disclose the practices they used, relying on the 'marketplace of ideas' to drive out bad practices over time.

In the international arena, there are countries that impose barriers to release of public opinion polls. WAPOR has conducted five studies on the freedom to publish opinion polls – the first study included 49 countries and was conducted in 1984; the fifth and latest study included 85 countries and was conducted in 2012. The results of the 2012 study indicate that many countries impose embargoes on preelection and exit polls, presumably so as not to influence voting. More troubling, 12 countries indicated that there were legal or self-imposed bans on public opinion polls on particular topics, such as the

political leadership, religion or ethnicity, foreign policy, private matters (e.g., psychological problems), and values. Finally, about one-half of the 60 countries in which polls were frequently conducted indicated that access to poll results by researchers or the public was difficult or not possible.

Obligations to Respondents

Few, if any, surveys put respondents at physical or psychological risk in ways that psychological experiments or, particularly, medical research may do. Ordinarily, the biggest risk to survey respondents comes from the possibility that their identity will be disclosed, either inadvertently, because inadequate precautions were taken to maintain the confidentiality of the data, or deliberately, as a result of special efforts by an 'intruder' to obtain the identity of one or more respondents. Thus, survey organizations have a special responsibility both to protect the confidentiality of the data they collect and to inform respondents of the risks of disclosure that remain. Adding to the challenges of protecting confidentiality is the growing propensity of data collection programs to include not only traditional survey responses, but also such data as biospecimens (for example, blood or tissue samples) collected from respondents, geospatial information obtained from Internet sources and linked to respondents, and administrative records (for example, social security earnings records) linked to respondents. Yet another ethical issue pertaining to respondents arises from declining survey response rates and corresponding attempts to compensate by special persuasion strategies, including the offer of refusal conversion payments. Preceding the discussion of these issues is a brief summary of federal requirements for the protection of human research participants, including survey respondents. Although in the United States these requirements legally apply only to federally funded research that involves human subjects, many survey and polling organizations use institutional review boards (IRBs) to review new and ongoing survey projects to accord with federal requirements. In the international arena, the extent of regulation of research with human subjects varies from country to country. For example, member states of the European Union are subject to the provisions of the Clinical Trials Directive adopted in 2001, but this directive does not cover surveys or other noninterventional research.

U.S. Federal Agency Protections for Human Research Participants

The U.S. government first instituted policies designed to protect human research participants in the 1960s. These policies gained regulatory force in 1974 and evolved into the current "Federal Policy for the Protection of Human Subjects," Title 45 of the *Code of Federal Regulations*, Part 46. The Common Rule, as 45 *CFR* 46 is known, has been adopted by over 15 agencies and departments; it was largely put in place in 1991 with minor modifications in 2005 and 2009. All organizations that conduct research funded by a participating agency must register and operate an IRB to ensure that covered research complies with the Common Rule regulations.

A seminal document shaping the Common Rule was the 1978 report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, known as the *Belmont Report* (see Citro, 2008; National Research Council, 2003). This report identified three major ethical principles for the conduct of research on humans:

- *Respect for persons*, or the obligation to treat people as autonomous agents by respecting their decisions on whether to participate in research. From this principle flows a requirement for researchers to obtain voluntary informed consent from participants, with special protective procedures when dealing with people who are immature, incapacitated, or whose autonomy is constrained.
- *Beneficence*, or the obligation to protect participants from harm to the extent possible and to maximize benefits that are likely from the research, not only for them, but also for society. From this principle flows a requirement for researchers and IRBs to assess risks of harm and probability of benefits in a systematic manner, including risks of disclosure of confidential information.
- *Justice*, or the obligation to show fairness in the selection of research participants with regard to who bears the burden of research and who benefits from the results.

Provisions of the Common Rule that are relevant to surveys and polls include:

- 'Research' is defined to include not only production but also pilot testing.
- Surveys or interviews are exempt from IRB review if they record responses so that individuals cannot be identified, directly or through identifiers linked to them, and if any reidentification would not reasonably be expected to place respondents at risk of criminal or civil liability or be damaging to their financial standing, employability, or reputation.
- Surveys or interviews are exempt from IRB review if their respondents are elected or appointed public officials or candidates for public office.
- Surveys that are not exempt may qualify for expedited IRB review if they present no more than 'minimal risk' to respondents as defined in the regulations.
- Surveys, like other research with human participants, must obtain the informed consent of individuals to participate, via written consent documents unless the IRB determines that the existence of a signed written consent could lead to a breach of confidentiality or the research is minimal risk and involves procedures for which written consent is not normally required outside the research context.
- Surveys conducted with prisoners or children are subject to additional rules for protection of these vulnerable population groups.

A major overhaul of the Common Rule was proposed in an Advance Notice of Proposed Rule-Making, issued in July 2011 (see National Research Council, 2014). The goals of the proposed modifications are "to enhance the effectiveness of the research oversight system by improving the protections for human subjects while also reducing burdens, delays, and ambiguity for investigators and research subjects." As of this writing, no action has been taken to finalize the proposed changes.

Other federal laws may affect surveys that plan to link administrative records to survey responses. They include the Family Educational Rights and Privacy Act of 1974 (FERPA, 34

CFR 99), which governs the use of school-based records, and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the accompanying Privacy rule issued by the U.S. Department of Health and Human Services (DHHS) in 2002 (45 *CFR* 160, 164), which govern use of patient medical records.

Informed Consent in Surveys and Polls

Informed consent means that a person's decision to participate is made by that person (or his or her legally authorized representative), without pressure to hurry decisions, coercion, or undue influence from the researcher, and with relevant, understandable information about the research (National Research Council, 2003). One important purpose of informed consent is to protect respondents from harm. Another purpose is to give them meaningful control over information about themselves, even if the question of harm does not arise. For example, informing potential respondents of the voluntary nature of their participation is essential to ethical survey practice, even if this leads some potential respondents to refuse to participate. (All private sector and all government surveys of households and individuals in the United States are voluntary, with the exception of the decennial census and the American Community Survey conducted by the U.S. Census Bureau under the authority of Title 13 of the U.S. Code and the U.S. Constitution.) Similarly, obtaining respondents' permission to record an interview should be standard practice, even though no harm is likely to come to them as a result.

At present, obtaining informed consent is only legally mandated for surveys funded by federal agencies under the Common Rule and subject to IRB approval. For the many other surveys conducted for nonfederal sponsors, the AAPOR Code of Ethics mandates that researchers "provide all persons selected for inclusion with a description of the research study sufficient to permit them to make an informed and free decision about their participation, … make no false or misleading claims as to a study's sponsorship or purpose, and … provide truthful answers to direct questions about the research. If disclosure could substantially bias responses or endanger interviewers, it is sufficient to indicate that some information cannot be revealed or will not be revealed until the study is concluded." (The American Sociological Association Code of Ethics, adopted in 1999, also includes a requirement to obtain informed consent.) The AAPOR Code further mandates that researchers "avoid practices or methods that may harm, endanger, humiliate, or seriously mislead survey respondents or prospective respondents" and that researchers "respect respondents' desires, when expressed, not to answer specific survey questions or provide other information to the researcher." Similarly, the American Statistical Association (ASA) Ethical Guidelines for Statistical Practice, adopted in 1999, cautions researchers to avoid excessive imposition on the time and privacy of respondents and also to use no more (or less) numbers of research subjects to satisfy the study requirements for precision of results.

Historically, a point of contention of the survey research community with IRBs has been the preferred status in the Common Rule accorded to written consent. Research has demonstrated that a small but significant group (around 10%) of potential respondents is willing to participate in a survey but not to sign a consent form (Singer, 2003). Consequently, IRB insistence on written consent in minimal-risk surveys does not appear to protect respondents and can reduce the generalizability of the findings. The National Research Council (2003) supports the waiver of written consent by IRBs as a matter of routine for mail and Internet surveys of the general population and most telephone surveys, primarily because the interaction between the interviewer and respondent is absent or at arm's length, and the respondent can easily choose not to respond. For in-person interviews, the National Research Council finds that tacit consent (that is, willingness of the respondent to sit for the interview) is usually adequate. For a more-than-minimal-risk survey, explicit consent is appropriate, although it is possible that verbal consent is preferable to written consent if the existence of a paper record could intentionally (by subpoena) or accidentally become public and cause harm to the respondent. Finally, additional separate consent procedures are likely to be required when a survey is to collect biospecimens that will be stored for future research and when survey responses are to be linked to administrative records or other independent data sources.

Protection of Respondents' Confidentiality

As noted above, probably the most important source of risk and potential harm to survey respondents comes from identification of their responses – that is, from deliberate or inadvertent breaches of confidentiality. The more sensitive the information requested – for example, HIV status or whether the respondent has engaged in some illegal activity – the greater the harm that a breach of confidentiality might entail. Not unexpectedly, respondents require strong assurances of confidentiality when they are asked to provide such data about themselves (Singer et al., 1995). There is also evidence of rising public concerns about confidentiality and privacy, which is problematic for maintaining survey response rates given the finding that people with stronger concerns in this area are less likely to participate in surveys (Singer et al., 2003).

A variety of threats to the confidentiality of survey data exist (National Research Council, 2005). Most common is simple carelessness – not removing identifiers from questionnaires or electronic data files, leaving cabinets unlocked, or not encrypting files containing identifiers. Although there is no evidence of respondents having come to harm as a result of such negligence, it is important for data collection agencies to be continually vigilant to raise the consciousness of their employees with respect to these issues, provide guidelines for appropriate behavior, and ensure that these are observed.

Less common, but potentially more serious, threats to confidentiality are legal demands for identified data, either in the form of a subpoena or as a result of a Freedom of Information Act (FOIA) request. Surveys that collect data about illegal behaviors such as drug use, for example, are potentially subject to subpoena by law enforcement agencies. To protect against this, researchers studying mental health, alcohol and drug use, and other sensitive topics, whether federally funded or not, may apply for Certificates of Confidentiality from DHHS. Such certificates, which remain in effect for the duration of the study, protect researchers from being compelled in most circumstances to disclose names or other identifying characteristics of survey respondents in federal, state, or local proceedings (42 *CFR* 2a.7, 'Effect of Confidentiality

Certificate'). The Certificate protects the identity of respondents; it does not protect the data, nor will it protect against 'statistical disclosure' discussed below.

When surveys are done expressly for litigation, demands for data from one party or the other are not uncommon, and protection against such demands may not be available. Several such cases are reviewed by Presser (1994). In one, the case was settled before the data were to be produced; in two others, the courts ordered the researchers to turn over raw data, including respondent identifiers, to the defendants. Presser (1994) discusses several actions that might be taken by professional survey organizations to enhance protection of confidentiality in such cases – for example, mandating the destruction of identifiers in surveys designed for adversarial proceedings.

A final threat to data confidentiality that has become much more serious since the advent of the Internet is what is known as 'statistical disclosure.' There is increasing awareness that even without explicit identifiers, such as name, address, or social security number, reidentification of respondents may be possible given the existence of high-speed computers, external data files readily available on the Internet to which the survey data could be matched, and sophisticated matching software. Longitudinal studies may be especially vulnerable to such reidentification because of the breadth and depth of data typically available on each respondent. Both cross-sectional and longitudinal studies are also more vulnerable to the extent that the data collection includes links to administrative records or other data sources. Finally, not only data provided in aggregate form, but also and especially microdata records for individual respondents are vulnerable to reidentification.

There are several ways to reduce the risk of statistical disclosure. For microdata, one method is to restrict access, for example, by requiring researchers to work in research data centers or enclaves, such as those maintained by the U.S. Census Bureau at a number of universities, or by requiring researchers and their institutions to agree to the terms of a license for access to microdata at their own site or through electronic communication to sites that house the data in a secure environment. For both aggregate and microdata, another way to reduce risk is to restrict the data that are released, for example, by top-coding age and income and releasing detailed data only for large geographic areas. Yet a third way to safeguard aggregate or microdata is to create synthetic data derived from respondents' actual information by using multiple imputation and other sophisticated statistical methods. There are tradeoffs for the researcher in each of these forms of access (see Duncan in National Research Council, 2003; Federal Committee on Statistical Methodology, 2005; for special issues for surveys that collect biospecimens, see National Research Council, 2010).

In view of the diverse threats to confidentiality reviewed above, there is growing consensus among researchers that absolute assurances of confidentiality should seldom, if ever, be given to respondents. The ASA Ethical Guidelines, for example, caution statisticians to "anticipate secondary and indirect uses of the data when obtaining approvals from research subjects." At a minimum, the Guidelines urge statisticians to provide for later independent replication of an analysis by an outside party. They also warn researchers not to imply protection of confidentiality from legal processes of discovery unless they have explicit authorization to do so, for example, by a Certificate of Confidentiality. Research by Singer et al. (1995) suggests that for surveys on nonsensitive topics, neither the response rate nor data quality is likely to suffer as a result of such qualified assurances of confidentiality.

Efforts at Persuasion

A third ethical issue concerning respondents is the dilemma created by the need for high response rates from populations that spend less time at home, have less discretionary time, and are less inclined to cooperate in surveys (National Research Council, 2013a). The problem of falling response rates leads survey organizations to greater use of incentives to promote response, which research shows generally improves cooperation (Singer and Ye, 2013), and to more follow-up attempts for respondents who do not respond to the initial contact attempt, often using different methods, such as telephone follow-ups to a mail survey. The problem also leads survey organizations to attempt more conversions of people who actually refuse to participate and to make greater use of refusal conversion payments. Respondents see both practices as undesirable (Groves et al., 1999). They also regard the paying of refusal conversion payments as unfair, but awareness of such payments does not appear to reduce either the willingness to respond to a future survey or actual participation in a new survey (Singer et al., 1995). Surveys are increasing turning to so-called 'responsive design,' whereby models and paradata about incoming responses are used to determine which nonrespondents it is most cost-effective to pursue.

Polling on the Net

Polls and surveys conducted on the Internet have shown explosive growth in the last several years, and such surveys not only bring entirely new ethical problems to the fore, but also new dimensions to old problems. For example, maintenance of privacy and confidentiality in such surveys requires technological innovations (for example, 'secure' Websites, encryption of responses) beyond those required by older modes of data collection. Other ethical issues arise for Internet surveys that have never pertained to older modes of conducting surveys – for example, the possibility that a respondent might submit multiple answers, thus deliberately biasing the results – and these, too, call for new technical solutions requiring sophisticated knowledge of computer hardware and software. These problems, and proposed solutions, are still too new to permit more than brief mention of them here.

Concluding Observations

A number of developments in recent years have increased ethical dilemmas for survey researchers. There is growing demand for information about individuals on the part of both business and government. At the same time, there is growing concern about personal privacy on the part of the public. Technological developments have made safeguarding privacy and confidentiality increasingly difficult, and for this as well as other reasons, there is increasing reluctance on the part of the public to participate in polls and surveys, including

government censuses. How these dilemmas are resolved in the next decade will have profound implications for the future of survey research.

See also: Common Rule: US Regulatory Protections for Human Research Participants; Ethical Practices: Institutional Oversight and Enforcement; Fieldwork: Ethical Aspects; Informed Consent of Research Subjects; Internet Research: Ethical Concerns; Privacy of Individuals in Social Research: Confidentiality; Sample Survey Methodology, History of; Surveys: Question Wording and Response Categories.

Bibliography

Citro, C.F., 2008. Legal and human subjects considerations in surveys. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research, second ed. Emerald Group Publishing Limited, Bingley, UK.

Federal Committee on Statistical Methodology, 2005. Working Paper 22: Report on Statistical Disclosure Limitation Methods. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC (originally published 1994).

Groves, R.M., Singer, E., Corning, A.D., Bowers, A., 1999. A laboratory approach to measuring the joint effects of interview length, incentives, differential incentives, and refusal conversion on survey participation. Journal of Official Statistics 15, 251–268.

National Research Council, 2003. Citro, C.F., Ilgen, D.R., Marrett, C.B. (Eds.), Protecting Participants and Facilitating Social and Behavioral Sciences Research Committee on National Statistics and Board on Behavioral, Cognitive, and Sensory Sciences. The National Academies Press, Washington, DC.

National Research Council, 2005. Expanding Access to Research Data: Reconciling Risks and Opportunities. Committee on National Statistics. The National Academies Press, Washington, DC.

National Research Council, 2010. Committee on National Statistics and Committee on Population. Hauser, R.M., Weinstein, M., Pool, R., Cohen, B. (Eds.), Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biodata. Committee on National Statistics and Committee on Population. The National Academies Press, Washington, DC.

National Research Council, 2013a. Nonresponse in Social Science Surveys: A Research Agenda. Committee on National Statistics. The National Academies Press, Washington, DC.

National Research Council, 2013b. Committee on National Statistics. Citro, C.F., Straf, M.L. (Eds.), Principles and Practices for a Federal Statistical Agency. The National Academies Press, Washington, DC.

National Research Council, 2014. Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences. Board on Behavioral, Cognitive, and Sensory Sciences. Committee on National Statistics, and Committee on Population. The National Academies Press, Washington, DC.

Presser, S., 1994. Informed consent and confidentiality in survey research. Public Opinion Quarterly 58, 446–459.

Seltzer, W., Anderson, M., 2008. Using population data systems to target vulnerable population subgroups and individuals: issues and incidents. In: Asher, J., Banks, D., Scheuren, F.J. (Eds.), Statistical Methods for Human Rights. Springer, New York.

Singer, E., 2003. Exploring the meaning of consent: participation in research and beliefs about risks and benefits. Journal of Official Statistics 19 (3), 273–285.

Singer, E., Von Thurn, D.R., Miller, R.E., 1995. Confidentiality assurances and response: a quantitative review of the experimental literature. Public Opinion Quarterly 59, 266–277.

Singer, E., Groves, R.M., Corning, A.D., 1999. Differential incentives: beliefs about practices, perceptions of equity, and effects on survey participation. Public Opinion Quarterly 63, 251–260.

Singer, E., Van Hoewyk, J., Neugebauer, R.J., 2003. Attitudes and behavior: the impact of privacy and confidentiality concerns on survey participation: the case of the 2000 census. Public Opinion Quarterly 67, 368–384.

Singer, E., Ye, C., 2013. The use and effects of incentives in surveys. In: Massey, D.S., Tourangeau, R. (Eds.), The Nonresponse Challenge to Surveys and Statistics, Annals of the American Academy of Political and Social Science, vol. 645. Sage Publications, Thousand Oaks, CA.

Relevant Websites

http://www.aapor.org/AAPOR_Code_of_Ethics/4249.htm#.UoUSifnrws0 – AAPOR Code of Professional Ethics and Practices.

http://www.asanet.org/images/asa/docs/pdf/Ethics%20Code.pdf – American Sociological Association Code of Ethics.

http://www.amstat.org/about/ethicalguidelines.cfm – American Statistical Association Ethical Guidelines for Statistical Practice.

http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html – ANPRM (Protection of Human Subjects).

http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html – Belmont Report (Protection of Human Subjects).

http://www.hhs.gov/ohrp/humansubjects/commonrule/ – Common Rule (Protection of Human Subjects).

http://www.ed.gov/policy/gen/guid/fpco/ferpa/index.html – FERPA (Privacy Protection for Student Records).

http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/ – HIPAA Privacy Rule (Privacy Protection for Health Records).

http://grants.nih.gov/grants/policy/coc/ – NIH Grants Policy (Certificates of Confidentiality).

http://grants.nih.gov/grants/policy/nihgps_2013/nihgps_ch8.htm#_Toc271264947 – NIH Grants Policy (Data Sharing).

http://www.nsf.gov/bfa/dias/policy/dmp.jsp – NSF Grants Policy (Data Sharing).

http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf – OSTP Memorandum on Data Access.

http://wapor.org/wapor-code-of-ethics/ – WAPOR Code of Ethics.

http://wapor.org/freedom/ – WAPOR Report on the Freedom to Publish Opinion Polls.

Surveys: Extreme Response

Daniel M Bolt, University of Wisconsin-Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Extreme response style (ERS) is a frequently observed response style in which respondents disproportionately select the extreme end points of a rating scale (e.g., 1 = strongly disagree and 7 = strongly agree on a 7-point Likert rating scale). ERS has been demonstrated to contribute to bias in the measurement of latent constructs and in estimates of their association with other variables. It frequently correlates with other person and survey-related variables, and has also been shown to vary in its prevalence across cultures. Different statistical methods for both measuring ERS and controlling for the biasing effects of ERS have been presented in the research literature.

Rating scales in survey measures are often used in unique ways by different respondents. Response styles (alternatively 'response sets') refer to stylistic tendencies in how respondents assign ratings to items unrelated to item content (Cronbach, 1946). Response styles are often reflected by disproportionate tendencies toward using particular rating scale categories. A common form of response style that has been the focus of much methodological research is 'extreme response style' (ERS). ERS is defined by the tendency to select the extreme end points of a rating scale (such as 1 = 'strongly disagree' and 7 = 'strongly agree' on a 7-point Likert scale). Although various response style types have been documented, ERS has likely received the largest amount of attention in the research literature. This is probably due both to the high frequency with which it is observed as well as its demonstrated tendency to correlate with both person- and survey-related variables. The observation of ERS in a survey often coincides with the observation of an 'anti-ERS' response style, namely, a disproportionate tendency to avoid the extreme end points.

Different explanations have been given for the occurrence of ERS (Baumgartner and Steenkamp, 2001), some of which may be more relevant than others depending on the context. To the extent that response styles are often viewed as relatively stable characteristics of persons, the explanations have often focused on person-related variables, although characteristics of the survey, or an interaction between person and survey features likely also play a role (see below). One explanation is that ERS behavior is the manifestation of a personality characteristic or trait of the respondent, such as an intolerance of ambiguity (Naemi et al., 2009), for example. A related explanation focuses on different response norms, as might be associated with different cultural backgrounds, for example, whereby some respondents are encouraged to respond in modesty (i.e., anti-ERS) and others to respond in sincerity (i.e., ERS) (Hui and Triandis, 1989). Such explanations would appear to apply best to self-report surveys. Yet another common perspective connects ERS to a respondents' attempt to simplify the process of responding to multicategory items. Distinguishing between multiple levels of agreement to a survey item, such as on a 7-point rating scale, for example, can be cognitively demanding, especially when applied repeatedly across a large number of items. From this perspective, respondents may only choose among rating scale end points as a way of reducing the complexity of the response process to a more simple agree/disagree response. Such an explanation appears consistent both with the relationships seen between ERS and more simplistic thinking/faster response times (e.g., Naemi et al., 2009), and the tendency to observe higher levels of response style behavior on surveys carrying more substantial cognitive load (Van Vaerenbergh and Thomas, 2012).

Consequences of ERS

ERS can be viewed as a form of systematic measurement error that contributes to bias in survey scores. Depending on the use of the survey instrument, the bias introduced can have significant consequences both in the interpretation of individual scores and in evaluating how the intended-to-be-measured substantive trait relates to other variables. At the individual level, ERS can systematically bias survey scores upward or downward depending on the substantive trait level of the respondent and the difficulty levels of the items (Baumgartner and Steenkamp, 2001; Bolt and Johnson, 2009). As a simple example, for a respondent whose expected response to an item (based on substantive trait level only) would be a 6 on a 1–7 scale, the presence of ERS will likely lead the respondent to select a 7 as opposed to a 6 (upward bias), while for a respondent whose expected score to an item (based on substantive trait level only) is a 2, ERS will likely lead to selecting a 1 (downward bias). ERS can also bias estimates of the distributional characteristics of the substantive trait across groups, such as in cross-cultural comparisons of scale scores. For example, Mõttus et al. (2012) evaluate ERS as a possible cause of unexpected country-level differences in self-reported levels of conscientiousness. Along these lines, ERS can also bias estimated associations between the survey and other external variables, a consequence that is also related to the tendency for ERS behavior to correlate with other person variables (see below). For example, Moors (2012) illustrates the deleterious effects of ERS on the estimated correlations among various leadership style types. In addition, the presence of ERS can interfere with the study of other psychometric properties of tests and test items, such as the study of differential item functioning/measurement invariance across groups (Morren et al., 2011; Bolt and Johnson, 2009). For these reasons, it

can be important to evaluate the potential effects of ERS on survey scores, and where possible, correct for such effects.

Index-Based Measurement of ERS

Various approaches have been taken to measure ERS. The different approaches often represent subtle differences in how ERS can be formally defined. A common strategy is to construct an index directly based on the scores a respondent assigns to items. Examples of such indices include the proportion of times a respondent selects the extreme categories, or alternatively, the standard deviation of the ratings assigned by the respondent across items. A limitation of index methods is their likely sensitivity to a respondent's substantive trait level. For example, a respondent with an extremely high level on the substantive trait (e.g., an extreme attitude) may be selecting primarily extreme categories due to the intended-to-be-measured substantive trait and not due to ERS, which by definition is response behavior unrelated to item content.

As a result of this limitation, a more attractive alternative is to add additional self-report items solely for the purpose of measuring response style (Greenleaf, 1992). An ERS index can then be computed as above using only the added items. Ideally, the added items will tap constructs poorly related to the trait intended to be measured, so as to ensure that the ERS index serves as a purer indicator of response style. Such an approach assumes ERS to be a relatively stable characteristic of a respondent across constructs. A challenge in implementing the method involves identification of such added items, as in many disciplines it can be difficult to find constructs that are nearly uncorrelated. A promising approach that overcomes the limitation of the Greenleaf strategy is the use of anchoring vignettes. Anchoring vignettes are brief hypothetical scenarios that are rated by respondents using the same rating scale as is used for the survey instrument. The primary difference from the Greenleaf approach is that the vignettes do not relate to a self-report construct, and instead reflect scenarios designed to instill a common subjective response across all respondents. The administration of vignettes aligned with different levels of the rating scale can thus lead to insight in how respondents of varying trait levels are using the rating scale differently. While appealing, the method does assume that respondents use the rating scale in the same way for self-report ratings as when responding to the anchoring vignettes (King et al., 2004).

Psychometric Modeling of ERS

Another way of overcoming the limitations of index-based ERS measurement is to apply an item response model. There have been a number of different item response models advanced for the measurement of ERS. In describing the different models, it can be helpful to distinguish those that attempt to address response styles more generally from those that handle ERS as a qualitatively distinct response style type. An example of the former is Rossi et al.'s (2001) scale heterogeneity model. Under the scale heterogeneity model, individual differences in the use of ratings scale categories are captured by respondent-specific location (τ) and scale (σ) parameters, corresponding roughly to the mean and variance of ratings assigned across items. Such an approach can be viewed as addressing a wide variety of potential response style types, in this case all that can be understood by a linear translation of the rating scale. Under such an approach, ERS respondents emerge as respondents having large positive σ parameters. Such models can also be generalized to include respondent-specific category threshold parameters, in which case ERS respondents would tend to have thresholds for the extreme categories pulled much closer to the center of a subjective continuum. An approach of this kind is presented by Johnson (2003).

A different class of modeling approaches emphasizes ERS as a particular type of response style. These models associate with certain respondents a greater propensity toward selecting the extreme categories, specifically, as opposed to a general variability in item scores, and thus provide a more direct representation of ERS as traditionally defined. Such tendencies can be captured through latent class models, for example (Moors, 2003; Van Rosmolen et al., 2010), whereby different classes are defined by different response style types. Under such approaches, an ERS class commonly emerges in an exploratory analysis. Under the Moors (2003) approach, specifically, the latent classes emerge in the context of a latent class factor model, whereby one latent class factor corresponds to the intended substantive trait, and others to response style tendencies. Bolt and Johnson (2009) consider a continuous generalization of the Moors (2003) latent class factor approach. As in the Moors (2003) model, the item response is influenced simultaneously by the underlying substantive trait as well as an ERS trait that are related to the probability of selecting a particular score category using a multinomial logistic link function. For purposes of illustration, the author considers the Bolt and Johnson (2009) approach in more detail. Assuming $U_j = k$ to denote the event that a respondent selects score category k on item j, the probability of that event is given by

$$P(U_j = k|\theta_1, \theta_{\mathrm{ERS}}) = \frac{\exp\left(a_{jk1}\theta_1 + a_{jk,\mathrm{ERS}}\theta_{\mathrm{ERS}} + c_{jk}\right)}{\sum_h \exp\left(a_{jh1}\theta_1 + a_{jh,\mathrm{ERS}}\theta_{\mathrm{ERS}} + c_{jh}\right)}$$

where θ_1 and θ_{ERS} denote the continuous substantive and extreme response style latent traits, respectively (often assumed bivariate normal with a mean of 0 on each trait). In this model, the category slope parameters (for a seven-category item) are specified as a_{jh1}=(0, 1, 2, 3, 4, 5, 6) and $a_{jh,\mathrm{ERS}}$=(1, −0.4, −0.4, −0.4, −0.4, −0.4, 1), while the c_{jh} category intercept parameters are estimated but subject to a normalization constraint $\sum_h c_{jh} = 0$. Specification of the category slope parameters permits identification of the traits as conforming to the substantive and ERS traits, respectively.

The Bolt and Johnson (2009) model can be viewed as a multidimensional generalization of the nominal response model (Bock, 1972). As such, the model can be understood in a 'divide-by-total' framework where the numerator defines a respondent's propensity toward score category k, while the denominator captures the sum of propensities across all score categories. As a result, respondents with higher levels of

θ_1 are more prone to select the higher score categories (assuming items have been coded such that higher scores reflect a higher level of the substantive trait) and respondents with higher levels of θ_{ERS} are more prone to select the end-point categories.

Table 1 displays examples of item response patterns and estimated levels of ERS when the data are fit using the Bolt and Johnson (2009) and Rossi et al. (2001) models. The example item response patterns are from an analysis reported in Bolt and Johnson (2009) to a 68-item survey instrument measuring tobacco smoking dependence where each item is scored using a 7-point rating scale. Response patterns representing varying levels of ERS are shown, with Respondents 1 and 2 showing evidence of anti-ERS, Respondents 3 and 4 neither anti-ERS nor ERS, and Respondents 5 and 6 ERS. As seems apparent from these examples, both the scale heterogeneity (e.g., Rossi et al., 2001) and response style type-based (Bolt and Johnson, 2009) models of ERS generally provide comparable descriptions of respondents in term of their ERS tendencies. Importantly, both models are able to detect ERS tendencies for respondents across varying levels of the substantive trait.

Latent class representations of response styles can also be incorporated into models that characterize the substantive trait as continuous. Mixture item response models (e.g., Mixture Rasch models) provide a common statistical methodology in which such an approach can be applied (see e.g., Eid and Zickar, 2007). Under such models, class differences are commonly represented in the form of different thresholds associated with rating scale categories across items. ERS is seen through a class that has thresholds associated with the extreme categories that imply a greater tendency to select those response categories. Figure 1 displays latent class profiles for a mixture Rasch solution applied to data from a 10-item scale measuring the sixth graders enjoyment of science. The survey was administered as part of the Programme for International Student Assessment (PISA) 2003 survey (Bolt and Newton, 2011). All items were coded using four score categories ranging from 1 = strongly agree to 4 = strongly disagree. In the current analysis, a mixture partial model was applied that assumed a constant shift in the category threshold parameters associated with rating scale categories across items. Statistical criteria identified a three-class solution as superior. (Based on model coding, higher category estimates correspond to an increased likelihood of selecting the category.) The three classes that emerged reflect unique forms of response style, including an ERS class (class 3), an anti-ERS class (class 2), and a class with a disproportionate tendency to select category 4 (class 1), potentially a disacquiescent response style class.

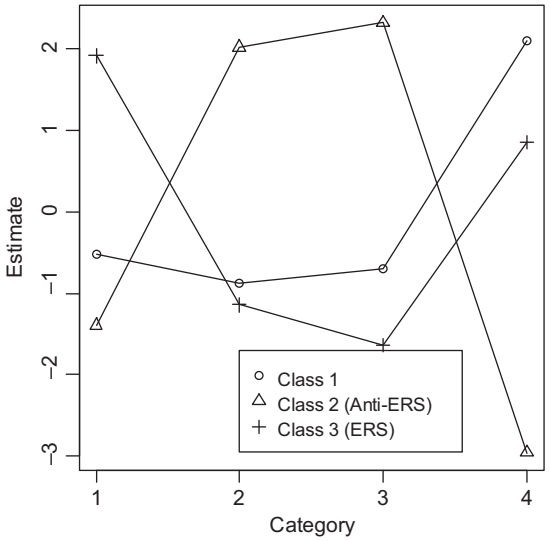

Figure 1 Category thresholds associated with three class mixture partial credit solution, 2003 PISA survey measuring enjoyment of science (sixth grade).

Table 1 Example item score patterns from smoking dependence scale illustrating different levels of ERS behavior and measurement of ERS under different statistical models

Respondent	*Response pattern*	$\hat{\theta}_1$	$\hat{\theta}_{ERS}$	$\hat{\tau}$	$\hat{\sigma}$
1	5655656664554653555554 6 6665532656564655666646 5 5566673455565234634664	1.17	−1.35	1.04	1.07
2	4654555554554435433 4 4555534455554544565335 3 5455445334544455554454	0.34	−1.34	0.38	0.77
3	3132434474524757223223 3 3221427217242413257122 5 2732273212113761452117	−0.51	0.00	−0.62	1.96
4	2113242463711513143535 1 6416651417332641125163 3 5675272634532154655432	−0.43	0.00	−0.46	1.90
5	7777777777771711777777 7 1771717171717771777177 7 7177777771777777777111	0.82	1.92	1.67	2.51
6	1111711711171711171711 1 7111711117717771171117 1 1117177111711171711111	−0.31	2.01	−1.06	2.93

One of the appealing aspects of item response models for ERS is not only their capability to measure ERS, but also their potential in controlling for the biasing effects of ERS on estimates of the substantive trait. This possibility is described in a later section.

Evaluating the Biasing Effects of ERS

Another advantage of a model-based approach to ERS is the ability to examine more specifically how ERS contributes to bias in survey scores. The author considers this issue again using the model of Bolt and Johnson (2009) as an example. For any given levels of θ_1, θ_{ERS}, one can initially compute an expected total score on the survey as

$$\text{Expected Score } (\theta_1, \theta_{\text{ERS}}) = \sum_{j=1}^{N} \sum_{h=1}^{n} k \times P(U_j = k | \theta_1, \theta_{\text{ERS}})$$

where $P(U_j = k|\theta_1,\theta_{\text{ERS}})$ is as defined in the previous section, and N and n denote the total number of items and rating scale categories, respectively. If one assumes the mean of θ_{ERS} (=0) to define an appropriate reference point, one can also use the equation to determine an expected score at this level of ERS, namely, Expected Score $(\theta_1,\theta_{\text{ERS}} = 0)$. Then bias can be estimated as

$$\text{Bias}(\theta_1, \theta_{\text{ERS}}) = \text{Expected Score } (\theta_1, \theta_{\text{ERS}}) - \text{Expected Score}(\theta_1, 0)$$

It is important to note that since there exists no 'correct' level of ERS, $\theta_{\text{ERS}} = 0$ is to some extent an arbitrary reference point. However, as the location represents the mean of θ_{ERS}, and due to the likely presence of both ERS and anti-ERS response styles, it provides a natural reference point. In addition, because survey scores will be strictly increasing in relation to θ_1 for any specified level of θ_{ERS}, any alternative θ_{ERS} reference location should lead to a similar result in terms of relative bias when comparing respondent scores.

An examination of bias curves allows for an evaluation of the practical influence of ERS on a survey. Figure 2 presents curves that illustrate the bias in total score for different levels of θ_{ERS} again in reference to the tobacco smoking dependence survey analyzed in Bolt and Johnson (2009). In this case, the y-axis illustrates average bias at the item score level. The curves make apparent how the biasing effects of ERS are dependent on θ_1. It is also important to note that these curves are dependent on the category intercept estimates of the items (not shown here), which will naturally vary across survey instruments. For the survey that is the focus of the current analysis, an appealing feature of the bias curves is that they consistently go to 0 at a θ_1 level of approximately 0.2, a location very near the mean of θ_1. As a larger number of respondents are located near the mean, it is apparent that for many respondents, ERS has very little biasing effect. Such a result can be attributed to the fact that for respondents near this trait location, there are nearly as many items with expected scores below 4 as there are items with expected scores above 4, leading to a cancellation of ERS bias effects across items. For other survey instruments, this naturally may not be the case, making plots such as those in

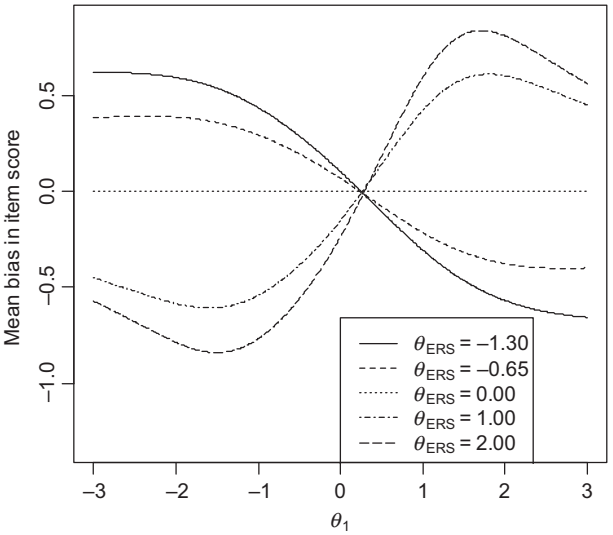

Figure 2 Bias due to ERS as calculated using Bolt and Johnson (2009) model, 68-item smoking dependence scale.

Figure 2 useful to inspect in determining the consequences of ERS.

Factors Related to ERS

A substantial amount of research has examined variables connected to display of ERS. ERS is frequently linked to person characteristics. For example, prior research has shown that respondents of higher education levels and socioeconomic status are often less prone to display ERS (Meisenberg and Williams, 2008) while respondents of more advanced age are more likely to display ERS (Kieruj and Moors, 2013). Other personality characteristics, such as trait anxiety (Lewis and Taylor, 1955), extraversion, and conscientiousness (Austin et al., 2006) have also been positively correlated with ERS, although findings across studies in regard to these factors are not always consistent (Van Vaerenbergh and Thomas, 2012). The effects of such respondent characteristics frequently align with explanations for ERS as given earlier, especially when viewing ERS as a stable person characteristic.

ERS has perhaps been most extensively studied in regard to its variability across ethnicity and culture. For example, some evidence suggests that respondents from Asian countries are less prone to display ERS, while those of Mediterranean or Latin American countries display on average higher levels of ERS. Similar effects are also seen across ethnic groups within a country (e.g., Marin et al., 1992). As noted earlier, one explanation for such findings may involve the application of different response norms across cultures (Hui and Triandis, 1989). Relatedly, Johnson et al. (2005) point to cross-cultural variability in the dimensions of masculinity and power distance as factors. To the extent that country-level comparisons of ERS are not always consistent, it is also possible that the content of surveys elicit stronger reactions in one culture compared to another, in which case cross-cultural differences that emerge as ERS might not be viewed as a response style

difference at all. Still other explanations for cross-cultural differences relate to issues associated with the translation of anchor points on rating scales or wording of items (Arce-Ferrer, 2006). Given the increased use of cross-cultural assessments and appeal in studying cultural differences in the social sciences, attempting to understand and address ERS in this context is an important area of future research.

While the response style literature has tended to emphasize response styles as stable person characteristics, there is also evidence that response style tendencies can become more prominent in some contexts than others. For example, to the extent that a respondent has a higher level of involvement with a topic, ERS can be increased (Gibbons et al., 1999). Other moderating factors may relate to characteristics of the administration of items and/or rating scales. One potentially relevant factor is the number of rating scale points, with some evidence suggesting scales with more scale points, as well as scales that include a middle point, can be less prone to lead to ERS (Weijters et al., 2010). In addition, administering items using a two-stage approach, whereby an initial filter item (e.g., agree/disagree) is followed by a subsequent item reflecting intensity (e.g., extent of agreement), has at times been shown to increase ERS, but in other contexts to decrease it (see e.g., Albaum et al., 2007). Other studies have examined the format of administration, such as Web-based versus paper/pencil-based responding, where Web-based administration is often seen to lead to greater ERS (e.g., Heerwegh, 2009). Although (as with person characteristics) there is a fair amount of inconsistency in the literature regarding the effects of characteristics of the survey instrument on ERS, most research appears supportive of some general principles in minimizing response style effects. As a general rule, the more that can be done to reduce the ambiguity of the responding process, the more likely the effects of response style, including ERS, are reduced. Reduced ambiguity can be achieved in various ways, including more explicit rating scale anchor labels ('every day' as opposed to 'frequently,' for example), more labels attached to the rating scale, and careful wording of items so as to avoid ambiguity. In addition, attempts to reduce the cognitive demand of responding (e.g., fewer items) or else instill greater effort on the part of respondents can also be effective.

Methods for Correcting ERS Bias

Under conditions where ERS is detected and found to contribute to bias, various statistical techniques can also be used to correct the bias. When using an index-type method to quantify ERS, a common approach applies a simple linear correction to the scale score. Such a correction can be applied, for example, through a linear regression of scale scores on the ERS index and using the residuals of the regression in place of the original scale scores in subsequent analyses. One limitation of this strategy is that it assumes a constant biasing effect of ERS on the scale score, a highly questionable assumption (Baumgartner and Steenkamp, 2001). As seen in **Figure 1**, the biasing effects of ERS can clearly vary as a function of the substantive trait level of the respondent. A second limitation is that the approach fails to account for the possibility that ERS may correlate with the intended-to-be-measured substantive trait. As a result, the method can lead to an 'overcorrection' for ERS bias by removing that association.

As an alternative, many of the statistical models for ERS described earlier also implicitly perform a bias correction by including the substantive trait in the model. For example, in the Bolt and Johnson (2009) model, estimates of the θ_1 trait can be viewed as bias-corrected, given the simultaneous presence of the θ_{ERS} trait in the model underlying the item scores. A challenge with these types of strategies, however, is the frequent difficulty in disentangling the potential effects of ERS from the effects of the substantive trait on the item responses. Often, large standard errors are found for estimates of θ_1 in such statistical models, for example, in relation to a response pattern where only the highest rating scale categories are endorsed for all items.

Such limitations point to the value of anchoring vignettes, as described earlier. Some studies have illustrated potential ways in which responses to vignettes can be used to perform bias corrections. One approach uses vignette responses to define respondent-specific threshold locations for score categories that can then be applied in reference to the self-report responses (King et al., 2004). In effect, the method attempts to reproduce the relative spacing and location of rating scale categories corresponding to the respondent's own internal representation of the rating score categories against a subjective continuum. A related approach makes use of vignette responses within a hierarchical ordinal probit (hopit) model, where bias corrections are performed in the context of relating the self-report items or survey scores to other variables (Grol-Prokopczyk et al., 2011). Each of these methods is more consistent with a scale heterogeneity approach to modeling response styles, and thus would be implicitly accounting for any form of response style heterogeneity as opposed to ERS specifically.

Finally, it is naturally also possible to address ERS bias by simply collapsing score categories (e.g., converting 7-point Likert responses into agree/neutral/disagree) so as to remove information from the item scores regarding extremity. While such an approach would remove effects related to ERS, it comes at the expense of reducing potential meaningful variance among respondents as well as items (Baumgartner and Steenkamp, 2001).

Future Directions

There remain many directions for future work in relation to ERS. One important issue relates to how best to design survey instruments so as to limit the effects of ERS. In particular, survey design issues that consider both (1) how to minimize the occurrence of ERS (e.g., through item writing and/or features of rating scales) and (2) how to make it reliably assessed at the respondent level when present (e.g., through including items of widely varying difficulty levels) remain important issues. Further applications of anchoring vignette strategies and related techniques will be important in evaluating the potential value of this methodology and the plausibility of its assumptions. Other techniques that can be adopted so as to generate more consistency in rating scale usage across respondents would also be worth exploring.

A second issue relates to further study of statistical models and their relative effectiveness in addressing problems introduced by ERS. In particular, attempts to compare methods such as the scale heterogeneity model that account for all forms of rating scale heterogeneity against models that address only response style types such as ERS will be useful.

Finally, one large challenge in synthesizing research on ERS relates not only to the diversity of statistical methods applied in measuring and controlling for its effects, but also to the widely varying conditions associated with the different survey administration contexts in which it is studied. It seems clear from past research that there are likely many different reasons respondents engage in ERS behavior, many of which will be more prominent in some settings than in others. It is likely that progress in this area of research will need to deal more explicitly with this diversity, and that final recommendations for how to address ERS may well need to be closely linked to the context of measurement.

See also: Psychometrics; Reliability: Measurement; Surveys: Question Wording and Response Categories.

Bibliography

Albaum, G., Roster, C., Yu, J.H., Rogers, R.D., 2007. Simple rating scale formats: exploring extreme response. International Journal of Market Research 49, 633–650.

Arce-Ferrer, A.J., 2006. An investigation into the factors influencing extreme response style. Educational and Psychological Measurement 66, 374–392.

Austin, E.J., Deary, I.J., Egan, V., 2006. Individual differences in response scale use: mixed Rasch modeling of responses to NEO-FFI items. Personality and Individual Differences 40, 1235–1245.

Baumgartner, H., Steenkamp, J.B.E.M., 2001. Response styles in marketing research: a cross-national investigation. Journal of Marketing Research 38, 143–156.

Bock, R.D., 1972. Estimating item parameters and latent ability when items are scored in two or more nominal categories. Psychometrika 37, 29–51.

Bolt, D.M., Johnson, T.R., 2009. Addressing score bias and DIF due to individual differences in response style. Applied Psychological Measurement 33, 335–352.

Bolt, D.M., Newton, J.R., 2011. Multiscale measurement of extreme response style. Educational and Psychological Measurement 71, 814–833.

Cronbach, L.J., 1946. Response sets and test validity. Educational and Psychological Measurement 6, 475–494.

Eid, M., Zickar, M.J., 2007. Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In: von Davier, M., Carstensen, C.H. (Eds.), Multivariate and Mixture Distribution Rasch Models: Extensions and Applications. Springer, New York, pp. 255–270.

Gibbons, J.L., Zelner, J.A., Reduk, D.J., 1999. Effects of language and meaningfulness on the use of extreme response styles by Spanish-English bilinguals. Cross-Cultural Research 33, 369–381.

Greenleaf, E.A., 1992. Measuring extreme response style. Public Opinion Quarterly 56, 328–350.

Grol-Prokopczyk, H., Freese, J., Hauser, R.M., 2011. Using anchoring vignettes to assess group differences in general self-related health. Journal of Health and Social Behavior 52, 246–261.

Heerwegh, D., 2009. Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. International Journal of Public Opinion Research 21, 111–121.

Hui, C.H., Triandis, H.C., 1989. Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology 20, 296–309.

Johnson, T.P., 2003. On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. Psychometrika 68, 563–583.

Johnson, T., Kulesa, P., Llc, I., Cho, Y.I., Shavitt, S., 2005. The relation between culture and response styles: evidence from 19 countries. Journal of Cross-cultural Psychology 36, 264–277.

Kieruj, N.D., Moors, G.B.D., 2013. Response style behavior: question format dependent or personal style? Quality and Quantity 47 (1), 193–211.

King, G., Murray, J.L., Murray, J.A., Salomon, J.A., Tandon, A., 2004. Enhancing the validity and cross-cultural comparability of survey research. American Political Science Review 98, 191–207.

Lewis, N.A., Taylor, J.A., 1955. Anxiety and extreme preferences. Educational and Psychological Measurement 15, 111–116.

Marin, G., Gamba, R.J., Marin, B.V., 1992. Extreme response style and acquiescence among Hispanics. Journal of Cross-Cultural Psychology 23, 498–509.

Meisenberg, G., Williams, A., 2008. Are acquiescent and extreme response styles related to low intelligence and education? Personality and Individual Differences 44, 1539–1550.

Moors, G., 2003. Diagnosing response style behavior by means of a latent-class factor approach: socio-demographic correlations of gender role attitudes and perceptions of ethnic discrimination reexamined. Quality and Quantity 37, 277–302.

Moors, G., 2012. The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. European Journal of Work and Organizational Psychology 21, 271–298.

Morren, M., Gelissen, J., Vermunt, J., 2011. Dealing with extreme response style in cross-cultural research: a restricted latent class factor analysis approach. Sociological Methodology 41, 13–47.

Mõttus, R., Allik, J., Realo, A., Pullman, H., Rossier, J., Zecca, J., Tseung, C.N., 2012. Comparability of self-reported conscientiousness across 21 countries. European Journal of Personality 26, 307–317.

Naemi, B.D., Beal, D., Payne, S.C., 2009. Personality predictors of extreme response style. Journal of Personality 77, 261–282.

Rossi, P.E., Gilula, Z., Allenby, G.M., 2001. Overcoming scale usage heterogeneity: a Bayesian hierarchical approach. Journal of the American Statistical Association 96, 20–31.

van Rosmalen, J., van Herk, H., Groenen, P.J.F., 2010. Identifying response styles: a latent-class bilinear multinomial logit model. Journal of Marketing Research 47, 157–172.

Van Vaerenbergh, Y., Thomas, T.D., 2012. Response styles in survey research: a literature review of antecedents, consequences, and remedies. International Journal of Public Opinion Research. Online publication June 19, 2012.

Weijters, B., Cabooter, E., Schillewaert, N., 2010. The effect of rating scale format on response styles: the number of response categories and response category labels. International Journal of Research in Marketing 27, 236–247.

Surveys: Question Wording and Response Categories

Nora C Schaeffer and Jennifer Dykema, University of Wisconsin Survey Center, University of Wisconsin-Madison, Madison, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A large literature provides the basis for recommendations about how to improve the reliability and validity of standardized measurement, whether in survey interviews or in other contexts. Building on an analysis of question characteristics and the decisions required when writing questions, this article summarizes implications of relevant research. We present illustrations of revisions that apply guidelines for improving question.

Introduction

Writing questions for standardized survey measurement requires attention to how survey constructs are conceptualized and operationalized as well as to the technical issues that arise in composing the items themselves. Questionnaire designers are increasingly focusing on the characteristics of questions to better understand how they affect measurement and nonresponse error (e.g., Alwin, 2007; Willis, 2005). Based on an extensive review of the characteristics of survey questions (Schaeffer and Dykema, 2011), the authors proposed that question writers and researchers consider the following classes of characteristics: question topic, question type, response dimension, conceptualization and operationalization of the target object, question structure, question form, response categories, question wording, and implementation. Although only a subset of characteristics are discussed here, they all have implications for decisions that question writers have to make, decisions that may affect the response task faced by respondents and therefore the reliability and validity of the resulting data. These technical issues, on which this article focuses, involve a series of related decisions. Our orientation leads us to consider how characteristics of questions are related to each other. For example, a 'middle category' means something different for a bipolar versus a unipolar rating scale. Details of the literature on which this discussion and recommendations are based can be found in recent reviews including Krosnick and Presser (2010), Schaeffer and Dykema (2011), and Schaeffer and Presser (2003).

To be practical, questions must ask for information that respondents are willing and able to provide (see Fowler and Cosenza, 2008). To meet the requirements of measurement, all respondents must understand the question in the way intended by the question writer. Ideally, we would evaluate survey questions based on whether they obtain reliable (low random error and low interviewer variance) and valid (measure the intended concept without systematic error) data. Because studies of reliability and validity are expensive, decisions are often based on surrogate assessments that are plausibly related to these criteria, including cognitive interviews and response patterns such as straight-lining. For example, question characteristics that appear to facilitate cognitive processing by improving comprehension or retrieval are presumed to improve measurement. Additionally, in interviewer-administered surveys, interviewers must be able to read questions accurately (Schaeffer et al., 2010). For example, question characteristics that increase the frequency with which interviewers follow-up are presumed to reduce data quality (Mangione et al., 1992; Schaeffer and Dykema, 2004).

Parts of a Survey Question

We use a fictitious survey question to introduce some basic vocabulary:

> I am going to read you a list of household tasks. (HAND CARD.) For each task, I'd like you to use the categories on this card and tell me if you find the task not at all boring, a little boring, somewhat boring, very boring, or extremely boring.
>
> The first task is vacuuming. How boring you find vacuuming - not at all boring, a little boring, somewhat boring, very boring, or extremely boring?
>
> And how boring do you find mopping floors – not at all boring…

The *preamble* to this question includes an announcement of the *class of events or objects* to be asked about ('household tasks'), *task instructions* (an announcement about how the question is structured and request for the respondent to rate the events on the response dimension 'how boring'), an announcement of the *response dimension* ('how boring'), and a listing of the exact *response categories* ('not at all boring, a little boring, somewhat boring, very boring, extremely boring'). The *structure of the task* is then reinforced by using a *transitional phrase*, "The first task is…" to introduce the first *target object*, that is, the object to be rated ('vacuuming'), which is then followed by the *target question* – the question that requires the respondent's answer ("How boring do you find vacuuming…") – and a repetition of the *response categories*. The part of the target question that is (sometimes optionally) repeated for a set of items (in this case, "How boring do you find…") is commonly called the *question stem*, but we use the label *repeated part* to distinguish it from the *changing part*; in this case, the changing part is a set of objects, that is, specific household tasks, but it may also be a set of statements that the respondent is asked to evaluate. Alwin (2007: pp. 167–172)

describes a set of questions that share common response categories, as these do, as a *battery* and questions that are not a battery but share a common topic as a *series*.

Overview of Question Characteristics

Question Topic, Question Type, and Response Dimension

Question topic can be analyzed at different levels, although surveys are so varied that there is yet no comprehensive analysis that covers both broad and specific aspects of topic (but see Alwin, 2007; Saris and Gallhofer, 2007; Turner and Martin, 1984). We use *question type* to refer to groups of survey questions that require similar decisions or have a similar structure, for example, because they share possible response dimensions. We focus on two important types of questions: questions about events and behaviors and questions that ask for evaluations or judgments using rating scales. (We do not consider other common types of survey questions such as social characteristics, open nominal choice questions (Schuman and Presser, 1981), or rankings (Alwin and Krosnick, 1985).) **Figures 1** and **2**, respectively, summarize decisions that are made for each type of question. Questions about events and behaviors (henceforth 'events') ask about target objects that have occurred in real time because of the actions of self (behaviors) or others (events). The events may be internal (such as a headache) or external (such as being laid off or going to the doctor). Questions about evaluations or judgments (henceforth 'evaluations') require that respondents evaluate a target object (such as 'my job') using an evaluative response dimension (such as 'how satisfied'). We use 'evaluation' to describe this type of question in order to describe the task we ask the respondent to perform. Evaluations may be absolute or comparative, and they are often used as measures of attitudes.

A *response dimension* is the dimension that the question asks the respondent to think about in constructing an answer. As **Figure 1** suggests, questions about events or behaviors can ask about whichever of the response dimensions of occurrence, frequency, duration, rate of occurrence, or date or timing are relevant. For example, for a question about doctor visits, researchers might ask about whether or not the respondent went to the doctor in the past year (occurrence), how many times the respondent saw a doctor (frequency), how long most visits lasted (duration), at what interval they occurred (rate), and when the last visit occurred (date).

Figure 2 summarizes decisions that are made for evaluations that use rating scales. For these questions, the label for the response dimension is chosen based on the research question, and it is usually conceptualized as an ordered or intensity rating task. For rating scales, several technical decisions need to be made about the response dimension: whether to structure the response dimension as unipolar (from low to high) or bipolar (from high negative to high positive), how to orient the response dimension (e.g., beginning with the positive or negative end), how many response categories to include, whether and how to label the response categories, and how to present the response categories (e.g., in self-administered modes, categories can be laid out horizontally or vertically).

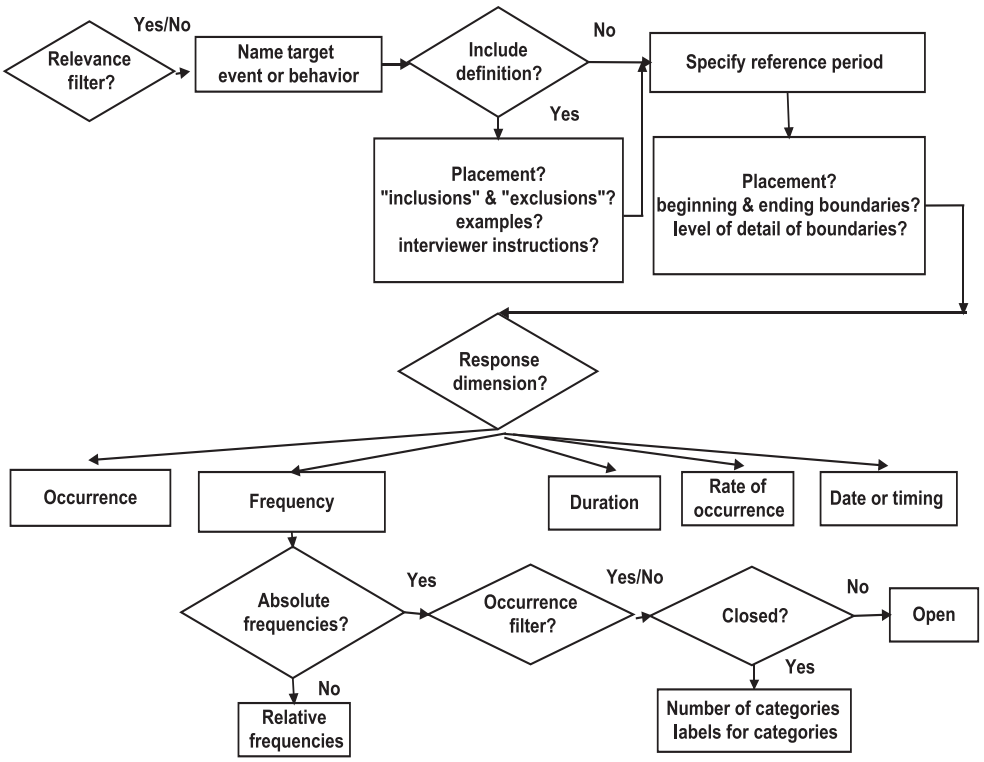

Figure 1 Simplified schematic of decisions for questions about events and behaviors Reprinted from Schaeffer, N.C., Dykema, J., 2011. Questions for surveys: current trends and future directions. Public Opinion Quarterly 75 (5), 909–961.

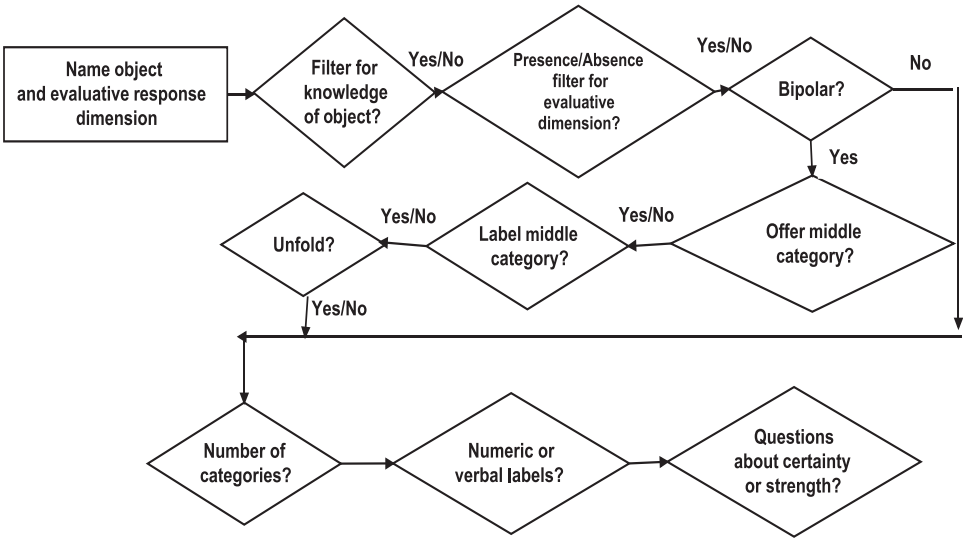

Figure 2 Simplified schematic of decisions for questions that use rating scales Reprinted from Schaeffer, N.C., Dykema, J., 2011. Questions for surveys: current trends and future directions. Public Opinion Quarterly 75 (5), 909–961.

Different types of questions (e.g., events or evaluations) are associated with different topics and different response dimensions. As a result, they are associated with different question structures and question forms (discussed below) and resulting differences in response processes and response errors. Various common response dimensions (e.g., occurrence, frequency, agree/disagree, intensity of a feeling) might be considered a further specification of the type of question. For example, recent work suggests that a type of question in which an object (such as 'vacuuming') is evaluated directly using a response scale built around an appropriate response dimension ('how boring') can be expected to have higher reliability and validity than a similar indirect question in which respondents are asked to rate how much they agree or disagree with a statement about the object (e.g., "How strongly do you agree or disagree that vacuuming is extremely boring?") (Saris et al., 2010; also Dykema et al., 2012; Fowler and Cosenza, 2008; Krosnick and Presser, 2010).

Conceptualization and Operationalization of the Target Object

Most survey questions are written about target objects – events, objects being rated, policy positions, and so on. The *conceptualization and operationalization of the target object* involve important decisions about how to label the target object and the response dimension. For example, in writing a question about the frequency of visits to the doctor in the past year, the question could refer to 'doctor,' 'medical doctor,' or 'health care provider.'

The operationalization of the concept and the resulting wording of the question can be further complicated by conditions, usually using conjunctions, prepositional phrases, subordinate clauses, or modifiers (e.g., "Last week, did you do ANY work for either pay or profit?"); by instructions to include or exclude (Redline, 2013); by examples, such as pictures in a web survey (Couper, 2008); or by other explicit definitions.

The operationalization of the target object in a series of questions may include a type of decomposition. One approach is to decompose a class of objects (e.g., contacts with your doctor), into subclasses (e.g., face-to-face, telephone, email, webchat, electronic chart system). Asking about each subclass separately implements a definition of what constitutes a 'contact with your doctor' in the structure of the questions themselves as an alternative to reading a complex definition. The occurrence of each subclass of contact can be asked about separately using a 'check-all-that-apply' question, a yes/no 'checklist,' or stand-alone questions; or a frequency can be obtained for each type of contact, that is, each subclass. In selecting among the approaches to measure occurrence, the researcher would rely on the small body of research comparing checklist to check-all response formats (see discussion in Schaeffer and Dykema, 2011). If frequencies are obtained for the subclasses, research would predict that the sum of the reports for the subclasses would be larger than the number that would be obtained by simply asking for a total (Belli et al., 2000).

Question Structure

As illustrated by the discussion about decomposition above, decisions about *question structure* concern the relationship between questions (e.g., Dykema and Schaeffer, 2000). Researchers must often decide between using a single question or two closely related questions, and in these cases, 'structure' and 'form' are extremely closely related. Examples include whether or not to use an occurrence filter followed by a frequency question versus a frequency question alone (Schaeffer and Dykema, 1994) or how to group multiple occurrences of a filter and follow-up question (Kreuter et al., 2011).

A decision for some topics is whether to precede a rating scale with a 'no opinion filter' (or another relevant 'don't know' option). The results from such filtered questions can be

compared to those obtained with unfiltered questions, which may or may not offer a 'no opinion' or 'don't know' category or record the answer only if volunteered (Krosnick and Presser, 2010). There is no evidence that using such a filter or offering such a category increases reliability, and using them complicates analysis. Similarly, although there is not much research about this question structure, for a unipolar evaluative response dimension, using a presence/absence filter (e.g., "Are you concerned about food safety?") and an intensity follow-up question appears less desirable than asking the rating question without the filter (Sterngold et al., 1994; review in Schaeffer and Dykema, 2011). In contrast, implementing a bipolar response dimension by asking about valence followed by an intensity question (with three categories) as in 'unfolding' (Malhotra et al., 2009; review in Krosnick and Presser, 2010) appears to increase reliability.

Question Form

Decisions about *question form* involve selecting a format to obtain the desired information. In some cases, there are alternative forms for obtaining the same or similar information, although the resulting data may have different properties. For example, to determine which of a set of objects the respondent might choose (e.g., as the most important problem facing the country), one could present candidate choices and ask the respondent to choose one, use an open question that provided categories for the interviewer to 'field code' the answer into, or simply record the answer to the open question verbatim (e.g., Schuman and Presser, 1981). The following list illustrates some common question forms that are used in standardized interviews:

1. Closed – selection from a set of ordered categories
   a. Ratings
      – (Direct) rating scales
      – (Indirect) statement + categories (e.g., agree/disagree)
   b. Frequencies
      – Grouped frequency categories
      – Rate-based frequency categories
      – Relative frequency categories
2. Closed – choose category or object
   a. Classification questions (e.g., "Are you currently married…?")
   b. Choose one or nominal choice (e.g., "What is the most important problem facing the country these days?")
   c. Choose all that apply
3. Rankings
4. Yes/No
5. Discrete value
   a. Numerical values of various sorts (e.g., dates, age, amounts, times)
   b. Questions that require calculations (e.g., number of eggs eaten in a typical week, percentages)
6. Open
   a. (Record answer verbatim and) interviewer field codes answer in a set of provided categories
   b. Record answer verbatim for coding later

Response Categories

As Figures 1 and 2 suggest, the decisions about response categories for questions about events are different from those for rating scales. As a result, there are dependencies among question characteristics, and they could interact in their effects on reliability and validity (e.g., Alwin, 2007). Consequently, there is some uncertainty in generalizing from studies that do not take the relationship among these question characteristics into account. Analyses of nonexperimental data (e.g., Alwin, 2007) are further limited both by researchers' analysis of the relationship among item characteristics and by the distribution of item characteristics in the set of available items (e.g., the net impact of polarity in rating scales can be investigated only if there is a sufficient number of unipolar and bipolar items in relevant combinations of other characteristics). Thus, for rating scales, results of these studies are potentially affected by whether the response dimension is unipolar or bipolar; for frequency questions, grouped frequency categories (e.g., '1–10 times') and relative frequency categories (i.e., labeled with adverbial phrases such as 'very often') are difficult to compare. The mode of administration and implementation of the items can also affect how categories are interpreted and used by respondents, and so these should, ideally, be controlled as well. Understanding the impact of the way the item is implemented is increasingly important in mixed-mode survey designs, for which researchers want measurement to be comparable across different implementations.

Studies about the number or labeling of frequency categories are fewer than those about rating scales. Some studies compare questions that ask for discrete values (which obtain absolute frequencies) and grouped frequency questions (sometimes called open and closed questions) or different sets of grouped frequency categories (review in Schaeffer and Presser, 2003). For questions that use relative frequency categories, the issues are probably similar to those for unipolar rating scales, although we do not know of any direct comparisons between grouped frequencies and relative frequencies. Taken together, relevant studies have led to a common recommendation to use discrete value rather than grouped frequency or relative frequency categories for frequency questions (e.g., Schaeffer and Presser, 2003), unless there is a specific reason to do otherwise. Asking the respondent to report a number eliminates decisions about the number of and labels for response categories.

For rating scales, there is considerable research about the effect of the number and labeling of response categories. There is substantial evidence from a wide range of experimental and nonexperimental studies across modes that verbal category labels increase reliability, presumably because they are interpreted more consistently than unlabeled numbers (review in Krosnick and Presser, 2010). The choice of labels can be informed by studies that scale adverbial phrases and intensifiers (see Cliff, 1959; Dobson and Mothersill, 1979 for examples). Cliff's (1959) results suggest that the following six categories would be approximately equidistant for a unipolar rating scale: *not at all, slightly, somewhat, quite, very, extremely* (Schaeffer and Presser, 2003). With only five categories, it is more difficult to select labels that have roughly equidistant

scale values. The labels in the fictitious item we gave earlier were chosen for five categories using Cliff (1959). Using Dobson and Mothersill (1979), one might suggest *not at all, slightly, somewhat (or moderately), very, extremely*. All these sets of labels end in 'extremely,' because these studies strongly suggest that 'very' is simply not intense enough to be the highest category for most purposes (see Schaeffer and Presser, 2003).

The choice of the number of categories for a rating scale requires balancing the researcher's desire to make distinctions, the respondent's ability to discriminate, and the capacity of language to supply sensible labels. It is difficult to summarize the evidence about the number of categories for rating scales, for the reasons given in the introduction to this section. Five to seven categories is a common recommendation (Krosnick and Presser, 2010; Schaeffer and Presser, 2003). Often the choice depends on mode: five or fewer is more common in telephone surveys, and there may not be room for seven categories on a paper self-administered instrument. In a mixed-mode study, researchers are likely to consider the most constrained mode. The polarity of the scale may also be a consideration: five categories offer little discrimination on either side of the midpoint, but unfolding makes it possible to increase the number of categories in the final scale by asking valence and then intensity (e.g., with three categories) on the side the respondent chooses (Malhotra et al., 2009; review in Krosnick and Presser, 2010). For a bipolar scale that is not unfolded, there is experimental evidence (using five-category agree–disagree items) that including a labeled midpoint reduced random error without affecting validity (review in Krosnick and Presser, 2010).

Question Wording

There are additional matters of *question wording*. In addition to definitions and task instructions, the preamble or introduction for a question can reinforce reference periods (e.g., "By 'last week' I mean the 7 days beginning Monday March 1st and ending yesterday"), motivate (e.g., "This information is important, so please take your time"), attempt to desensitize (e.g., "In talking to people about elections, we often find a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time…"), or improve recall (e.g., Cannell et al., 1981).

Finally, a substantial group of item characteristics concerns language, particularly usage that affects comprehension (e.g., because the vocabulary or grammar is complex) or retrieval (e.g., because the question provides cues or time for retrieval) (Tourangeau et al., 2000). These include question length; reading difficulty of the vocabulary; number of words that are vague, ambiguous, or technical; and grammatical complexity (e.g., number of clauses, number of conjunctions) (Graesser et al., 2006).

Guidelines for Writing Survey Questions

Technical guidelines for writing questions draw on research examining implementation (e.g., ease of processing and respondents' preferences) and the consequences for measurement (e.g., validity, reliability) of various question characteristics. The literature cited earlier offers guidelines about various aspects of question writing. We offer a few examples here. The revised questions in these illustrations are not intended to model all 'best practices,' but only to show a revision of the problem under discussion.

For example, with regard to decisions that have to be made about vocabulary and word choice, question writers are advised to use native vocabulary in place of analytic vocabulary in order to improve both comprehension and retrieval. The analytic vocabulary used by the analyst includes construct labels, technical terms, and jargon. Native vocabulary, the language used by the respondent, can be uncovered in developmental and cognitive interviews (Schaeffer and Dykema, 2004). Aids such as QUAID (Question Understanding AID) (Graesser et al., 2006) can be helpful in identifying technical and unfamiliar words. The examples below illustrate how technical vocabulary can be simplified:

Analytic vocabulary	*Revision*	*Source*
Occupants of this household	People who live here	Dillman et al. (2009)
Your responses	Your answers	Dillman et al. (2009)

Other technical guidelines are about the relationship between question structure and the response categories. For example, question writers are advised to match the format provided by the response categories to those requested by or projected by the question. Often, survey questions are written so that the question requests an answer in a format different from that provided. In Problematic Question 1 below, the response categories require the interviewer to enter the month and year. The question is less specific, asking only 'when' the respondent last saw his child, a question that could elicit a variety of responses only some of which will be a month and year. As already noted, if the interviewer has to probe for a codable response, interviewer variability and random error increase. The revised question clearly projects the categories in which the response is desired (see Fowler and Cosenza, 2008).

Problematic question 1

When did you last see CHILD?
MO _____ YR _____

Revision

In what month and year did you last see CHILD?
MO _____ YR _____

Similarly, the response dimension in the question should match the response categories in the question. In Problematic Question 2, the response dimension implied by the question, 'occurrence' (i.e., 'did you feel'), does not match the response dimension reflected in the response categories, 'frequency expressed as a rate' (i.e., 'never, about once a month, about once a week, or more often than once a week'). The response

dimension in the question and the response categories match in the revision.

Problematic question 2

During the last 12 months, did you feel like you just didn't know where the next day's food was going to come from? Never, about once a month, about once a week, or more often than once a week?

Revision

During the last 12 months, how often did you feel like you just didn't know where the next day's food was going to come from? Never, about once a month, about once a week, or more often than once a week?

In Problematic Question 3, the response dimension in the question (i.e., 'how often') does not match the response dimension in the response categories, frequency expressed as a number of times.

Problematic question 3

How often did you see CHILD last month?
ENTER NUMBER OF TIMES _____

Revision

How many times did you see CHILD last month?
ENTER NUMBER OF TIMES _____

In Problematic Question 4, the response dimension in the question, occurrence, does not match the response dimension in the response categories (i.e., 'often true, sometimes true, never true'). In Revision (1), the question is rewritten to ask about occurrence; in Revision (2), the question is rewritten to ask for relative frequencies; and in Revision (3), the question is rewritten to ask for absolute frequencies.

Problematic question 4

"I worried about whether my food would run out before I got money to buy more." Was that often true, sometimes true, or never true for you in the last 12 months?

Revision 1

In the last 12 months, did you worry about whether your food would run out before you got money to buy more?
YES
NO

Revision 2

In the last 12 months, how often did you worry about whether your food would run out before you got money to buy more? Often, sometimes, or never?

Revision 3

In the last 12 months, on how many days did you worry about whether your food would run out before you got money to buy more?
____ NUMBER OF DAYS

Conclusion

In this article, we have provided an overview of several classes of question characteristics and described some of their effects on measurement. The usefulness of research about survey questions depends, ultimately, on the underlying analysis of the characteristics of questions, which characteristics are compared, and how the dependencies among question characteristics are taken into account in the study design. For example, what characteristics of questions should be held constant when the effect of another characteristic is being estimated? Making progress in understanding how the characteristics of questions affect measurement requires that we also develop our understanding of what those characteristics are and how they covary or nest within each other. Additional topics, such as how the structure of the entire instrument or mode in which the instrument is administered affect data quality, also need exploration with stronger research designs.

See also: Ecology and Health: Large-Scale Social Surveys, Analysis of; Sample Survey Methodology, History of; Sample Surveys: Cognitive Aspects of Survey Design; Surveys: Extreme Response.

Bibliography

Alwin, D.F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. John Wiley & Sons, Inc., Hoboken, New Jersey.
Alwin, D.F., Krosnick, J.A., 1985. The measurement of values in surveys: a comparison of ratings and rankings. Public Opinion Quarterly 49 (4), 535–552.
Belli, R.F., Schwarz, N., Singer, E., Talarico, J., 2000. Decomposition can harm the accuracy of behavioural frequency reports. Applied Cognitive Psychology 14, 295–308.
Cannell, C.F., Miller, P.V., Oksenberg, L., 1981. Research on interviewing techniques. In: Leinhardt, S. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco, pp. 389–437.
Cliff, N., 1959. Adverbs as multipliers. Psychological Review 66 (1), 27–44.
Couper, M.P., 2008. Designing Effective Web Surveys. Cambridge University Press, Cambridge.
Dillman, D.A., Smyth, J.D., Christian, L.M., 2009. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, third ed. John Wiley & Sons, New York.
Dobson, K.S., Mothersill, K.J., 1979. Equidistant categorical labels for construction of Likert-type scales. Perceptual and Motor Skills 49 (2), 575–580.
Dykema, J., Schaeffer, N.C., 2000. Events, instruments, and reporting errors. American Sociological Review 65 (4), 619–629.
Dykema, J., Schaeffer, N.C., Garbarski, D., 2012. Effects of Agree-Disagree Versus Construct-Specific Items on Reliability, Validity, and Interviewer-Respondent Interaction. Paper presented at the annual meeting of the American Association for Public Opinion Research, May, Orlando, FL.
Fowler Jr., F.J., Cosenza, C., 2008. Writing effective questions. In: de Leeuw, E.D., Hox, J.J., Dillman, D.A. (Eds.), International Handbook of Survey Methodology. Lawrence Erlbaum Associates, New York, pp. 136–160.
Graesser, A.C., Cai, Z., Louwerse, M.M., Daniel, F., 2006. Question understanding AID (QUAID): a web facility that tests question comprehensibility. Public Opinion Quarterly 70 (1), 3–22.
Kreuter, F., McCulloch, S., Presser, S., Tourangeau, R., 2011. The effects of asking filter questions in interleafed versus grouped format. Sociological Methods & Research 40 (1), 88–104.
Krosnick, J.A., Presser, S., 2010. Question and questionnaire design. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research, second ed. Emerald Group Publishing Limited, Bingley, UK, pp. 263–313.
Malhotra, N., Krosnick, J.A., Thomas, R.K., 2009. Optimal design of branching questions to measure bipolar constructs. Public Opinion Quarterly 73 (2), 304–324.

Mangione, T.W., Fowler, J., Floyd, J., Louis, T.A., 1992. Question characteristics and interviewer effects. Journal of Official Statistics 8 (3), 293–307.

Redline, C., 2013. Clarifying categorical concepts in a web survey. Public Opinion Quarterly 77 (S1), 89–105.

Saris, W.E., Gallhofer, I.N., 2007. Design, Evaluation, and Analysis of Questionnaires for Survey Research. Wiley, New York.

Saris, W.E., Revilla, M., Krosnick, J.A., Shaeffer, E.M., 2010. Comparing questions with agree/disagree response options to questions with item-specific response options. Survey Research Methods 4 (1), 61–79.

Schaeffer, N.C., Dykema, J., 1994. Open and Closed Questions about the Frequency of Events: Replication, Extension, Paradox, Center for Demography and Ecology, University of Wisconsin-Madison, Working Paper 94-18.

Schaeffer, N.C., Dykema, J., 2004. A multiple-method approach to improving the clarity of closely related concepts: distinguishing legal and physical custody of children. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Singer, E. (Eds.), Methods for Testing and Evaluating Survey Questionnaires. Springer-Verlag, New York, pp. 475–502.

Schaeffer, N.C., Dykema, J., 2011. Questions for surveys: current trends and future directions. Public Opinion Quarterly 75 (5), 909–961.

Schaeffer, N.C., Dykema, J., Maynard, D.W., 2010. Interviewers and interviewing. In: Marsden, P.V., Wright, J.D. (Eds.), Handbook of Survey Research, second ed. Emerald Group Publishing Limited, Bingley, UK, pp. 437–470.

Schaeffer, N.C., Presser, S., 2003. The science of asking questions. Annual Review of Sociology 29 (1), 65–88.

Schuman, H., Presser, S., 1981. Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context. Academic Press, Orlando, FL.

Sterngold, A., Warland, R.H., Herrmann, R.O., 1994. Do surveys overstate public concern? Public Opinion Quarterly 58 (2), 255–263.

Tourangeau, R., Rips, L.J., Rasinski, K., 2000. The Psychology of Survey Response. Cambridge University Press, Cambridge, England.

Turner, C.F., Martin, E., 1984. Surveying Subjective Phenomena. Russell Sage Foundation, New York.

Willis, G.B., 2005. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Sage, Thousand Oaks, CA.

Survival Analysis: Introduction

Niels Keiding and Per K Andersen, University of Copenhagen, Copenhagen K, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Survival analysis is the study of the distribution of time-to-event data, that is, the times from an initiating event to some terminal event. A distinguishing feature of survival data is the occurrence of right censoring, where the terminal event for some individuals is not observed; instead, it is only known that this event is at least later than a given point in time.

This article presents a brief historical review and focuses on central statistical approaches: non- and semiparametric estimation (Kaplan–Meier, Cox regression) and hypothesis testing (log-rank). We also add brief comments on parametric models and list some important reference works.

Introduction

Survival analysis is the study of the distribution of time-to-event data, that is, the times from an initiating event (birth, start of treatment, employment in a given job) to some terminal event (death, relapse, disability pension). A distinguishing feature of survival data is the inevitable presence of incomplete observations, particularly when the terminal event for some individuals is not observed; instead, it is only known that this event is at least later than a given point in time: *right censoring*.

History

Survival analysis is one of the oldest statistical disciplines with roots in demography and actuarial science in the seventeenth century; see Westergaard (1901: Chapter 2; 1932) for general accounts of vital statistics and Hald (1990) for specific accounts of the work before 1750. More specifically, competing risks had been studied since at least 1760 by Bernoulli (see Seal, 1977 and Dietz and Heesterbeek, 2002), indirect and direct standardization since at least 1777 (Keiding, 1987; Keiding and Clayton, to appear), and the simultaneous handling of age, period, and cohort since the 1870s (Lexis, 1875; cf Keiding, 2011).

In the midtwentieth century the above techniques were well established, usually formalized in discrete time with consequent necessary approximations when describing the inherently continuous underlying time. The penetrating study by Kaplan and Meier (1958), the fascinating genesis of which was chronicled by Breslow (1991), eliminated in principle the need for these approximations in the situations where all survival and censoring times are known precisely. Following Böhmer (1912), Kaplan and Meier shrank the observation interval to include at most one observation per interval. Though overlooked by many later authors, Kaplan and Meier also formalized the age-old handling of delayed entry – again, see also Böhmer (1912) – through the necessary adjustment for the risk set, the set of individuals alive and under observation at a particular value of the relevant time variable.

Other remarkable contributions from the midtwentieth century include the pioneering study by Harris et al. (1950) on what is now termed interval-censored survival data, where each event is only known to have occurred in a time interval, usually different for each individual, and the combination by Ederer et al. (1961) of the old calculation of expected number of deaths with survival distribution estimation into the relative survival rate.

Nonparametric Estimation and Testing

The simplest situation encountered in survival analysis is the nonparametric estimation of a survival distribution function based on a right-censored sample of observation times $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$. Here, each $\widetilde{X}_i$ is either a survival time X_i, in which case the failure/censoring indicator D_i takes the value 1, or it is a right-censoring time, say U_i, and then $D_i = 0$. The true survival times X_i, $i = 1, \ldots, n$, are assumed to be independent and identically distributed with common survival distribution function $S(t) = P(T_i > t)$, whereas as few assumptions as possible are usually made about the right-censoring times U_i except for the assumption of *independent censoring*, see Andersen et al. (1993: Section III.2.2) or Kalbfleisch and Prentice (2002: p. 195). The concept of independent censoring has the interpretation that the fact that an individual, i, is alive *and uncensored* at time t, say, should not provide more information on the survival time for that individual than $X_i > t$, i.e., the right-censoring mechanism should not remove individuals from the study who are at a particularly high or a particularly low risk of dying. Under these assumptions, $S(t)$ is estimated by the *Kaplan–Meier estimator* (Kaplan and Meier, 1958). This is given by

$$\widehat{S}(t) = \prod_{\widetilde{X}_i \le t} \left[1 - D_i / Y(\widetilde{X}_i)\right] \qquad [1]$$

where $Y(t) = \sum I(\widetilde{X}_i \ge t)$ is the number of individuals *at risk* just before time t. The Kaplan–Meier estimator is a nonparametric maximum likelihood estimator (Kaplan and Meier, 1958; Johansen, 1978; Andersen et al., 1993: Section IV.3.1) and its asymptotic properties as $n \to \infty$ are very well studied since Breslow and Crowley (1974), see Andersen et al. (1993: Chapter IV).

Thus, in large samples, $\widehat{S}(t)$ is approximately normally distributed with mean $S(t)$ and a variance which may be estimated by Greenwood's (1926) formula:

$$\widehat{\sigma}^2(t) = \left[\widehat{S}(t)\right]^2 \sum_{\widetilde{X}_i \le t} \frac{D_i}{Y(\widetilde{X}_i)\left[Y(\widetilde{X}_i) - 1\right]} \qquad [2]$$

From this result, point-wise confidence intervals for $S(t)$ are easily constructed and, since one can also show weak convergence of the entire Kaplan–Meier curve $\{\sqrt{n}[\widehat{S}(t) - S(t)]; 0 \le t \le \tau\}, \tau \le \infty$ to a mean zero Gaussian process, simultaneous confidence bands for $S(t)$ on $[0, \tau]$ can also be set up, see, e.g., Andersen et al. (1993, Section IV.3.3) for details.

As an alternative to describing the distribution of survival times via the survival function $S(t)$, the *hazard function* $\alpha(t) = -S'(t)/S(t)$, or the *cumulative hazard function* $A(t) = -\log S(t)$ may be applied. Thus, the latter may be estimated by the *Nelson–Aalen estimator*,

$$\widehat{A}(t) = \sum_{\widetilde{X}_i \le t} D_i / Y(\widetilde{X}_i) \qquad [3]$$

(Nelson, 1969; Aalen, 1978; see also Andersen et al., 1993: Section IV.1). The relation between the estimators $\widehat{S}(t)$ and $\widehat{A}(t)$ is given by the *product integral* (Gill and Johansen, 1990) from which it follows that their large sample properties are equivalent. Though the Kaplan–Meier estimator has the advantage that a survival *probability* is easier to interpret than a cumulative hazard function, the Nelson–Aalen estimator is easier to generalize to multistate situations beyond the survival data context (*see* Event History Analysis in Continuous Time). To give a nonparametric estimate of the hazard function $\alpha(t)$ itself requires some smoothing techniques to be applied.

Right censoring is not the only kind of data-incompleteness to be dealt with in survival analysis; in particular, individuals may not all be followed from time 0 but may be from a later entry time V_i conditionally on having survived until V_i, so-called *left truncation* or *delayed entry*. Dealing with left truncation only requires a redefinition of the *risk set* from the set $\{i : \widetilde{X}_i \ge t\}$ of individuals still alive and uncensored at time t to the set $\{i : V_i < t \le \widetilde{X}_i\}$ of individuals with entry time $V_i < t$ and who are still alive and uncensored. With $Y(t)$ still denoting the size of the risk set at time t, eqns [1], [2], and [3] are applicable though one should be aware of the fact that estimates of $S(t)$ and $A(t)$ may be ill-determined for small values of t due to the left-truncation.

When the survival time distributions in a number, k, of homogeneous groups have been estimated nonparametrically it is often of interest to test the hypothesis H_0 of identical hazards in all groups. Thus, based on censored survival data $[(\widetilde{X}_{hi}, D_{hi}), i = 1, \ldots, n_h]$ for group h, $h = 1 \ldots k$, the Nelson–Aalen estimates $\widehat{A}_h(t)$ may be computed, and based on the combined sample of size $n = \sum_h n_h$ with data $[(\widetilde{X}_i, D_i), i = 1, \ldots, n]$ an estimate of the common cumulative hazard function $A(t)$ under H_0 may be obtained by a Nelson–Aalen estimator $\widehat{A}(t)$. As a general statistic for testing H_0, one may then use a k-vector of sums of weighted differences between the increments of $\widehat{A}_h(t)$ and $\widehat{A}(t)$ (Aalen, 1978; Andersen et al., 1982):

$$Z_h = \sum_{i=1}^{n} K_h(\widetilde{X}_i)\left[\Delta\widehat{A}_h(\widetilde{X}_i) - \Delta\widehat{A}(\widetilde{X}_i)\right] \qquad [4]$$

Here, $\Delta\widehat{A}_h(t) = 0$ if t is not among the observed survival times in the hth sample and $K_h(t)$ is 0 whenever $Y_h(t) = 0$, in fact all weight functions used in practice have the form $K_h(t) = Y_h(t) K(t)$. With this structure for the weight function, the covariance between Z_h and Z_j given by eqn [4] is estimated by

$$\sigma_{hj} = \sum_{i=1}^{n} K^2(\widetilde{X}_i) \frac{Y_h(\widetilde{X}_i)}{Y(\widetilde{X}_i)} \left[\delta_{hj} - \frac{Y_j(\widetilde{X}_i)}{Y(\widetilde{X}_i)}\right] D_i \qquad [5]$$

and, letting Z be the k-vector $(Z_1, \ldots, Z_k)'$ and $\sum$ the k by k matrix $(\sigma_{hj}, h, j = 1, \ldots, k)$ the test statistic $X^2 = Z' \sum^- Z$ is asymptotically chi-squared distributed under H_0 with $k-1$ degrees of freedom if all n_h tend to be infinity at the same rate, cf Andersen et al. (1993: Section V.2). Here, $\sum^-$ is a generalized inverse for $\sum$.

Special choices for $K(t)$ correspond to test statistics with different properties for particular alternatives to H_0. An important such test statistic is the *log-rank* test (Peto and Peto, 1972) obtained for $K(t) = I[Y(t) > 0]$. For this test, which has particularly good power for *proportional hazards* alternatives, Z_h given by eqn [4] reduces to $Z_h = O_h - E_h$ with O_h the total number of observed failures in group h and $E_h = \sum D_i Y_h(\widetilde{X}_i)/Y(\widetilde{X}_i)$ an 'expected' number of failures in group h. For the two-sample case ($k = 2$), one may of course use the square root of X^2 as an asymptotically normal test statistic for the null hypothesis. For the case where the k groups are *ordered*, and where a *score* x_h (with $x_1 \le \ldots \le x_k$) is attached to group h, a *test for trend* is given by $T^2 = (x'Z)^2/x' \sum x$ with $x = (x_1, \ldots, x_k)'$ and it is asymptotically chi-squared with 1 df, cf Andersen et al. (1993: Section V.3.3).

The above linear rank tests have low power against certain important classes of alternatives such as 'crossing hazards.' Just as for uncensored data, this has motivated the development of test statistics of the Kolmogorov–Smirnov and Cramèr–von Mises types, based on maximal deviation or integrated squared deviation between estimated hazards, cumulative hazards, or survival functions, cf Andersen et al. (1993: Section V.4.1).

Parametric Inference

The nonparametric methods outlined in the previous section have become the standard approach to the analysis of simple homogeneous survival data without covariate information. However, parametric survival time distributions are sometimes used for inference, and we shall here give a brief review. Assume again that the true survival times $X_1, \ldots, X_n$ are independent and identically distributed with survival distribution function $S(t; \theta)$ and hazard function $\alpha(t; \theta)$ but that only a right-censored sample $(\widetilde{X}_i, D_i), i = 1, \ldots, n$, is observed. Under independent censoring the likelihood function for the parameter θ is

$$L(\theta) = \prod_{i=1}^{n} \left[\alpha(\widetilde{X}_i; \theta)\right]^{D_i} S(\widetilde{X}_i; \theta) \qquad [6]$$

The function of eqn [6] may be analyzed using standard large sample theory, see e.g., Andersen et al. (1993: Chapter VI). Thus, standard tests, i.e., Wald-, score-, and likelihood ratio

tests are used as inferential tools. Two frequently used parametric survival models are the *Weibull* distribution with hazard function $\alpha\rho(\alpha t)^{\rho-1}$, and the piecewise exponential distribution with $\alpha(t, \theta) = \alpha_j$ for $t \in I_j$ with $I_j = [t_{j-1}, t_j)$, $0 = t_0 < t_j < \ldots < t_m = \infty$. Both of these distributions contain the very simplest model, the exponential distribution with a constant hazard function, as null cases.

Comparison with Expected Survival

As a special case of the nonparametric tests discussed above, a *one-sample* situation may be studied. This may be relevant if one wants to compare the observed survival in the sample with the *expected survival* based on a standard life table. Thus, assume that a hazard function $\alpha^*(t)$ is given and that the hypothesis H_0: $\alpha = \alpha^*$ is to be tested. One test statistic for H_0 is the one-sample *log-rank test* $(O - E^*)/(E^*)^{1/2}$ where E^*, the 'expected' number of deaths is given by $E^* = \sum[A^*(\widetilde{X}_i) - A^*(V_i)]$ (with A^* the cumulative hazard corresponding to α^*). In this case $\widehat{\theta} = O/E^*$, the *standardized mortality ratio*, is the maximum likelihood estimate for the parameter θ in the model $\alpha(t) = \theta\alpha^*(t)$ (Kilpatrick, 1962; Breslow, 1975; Keiding, 1987; Nielsen, 1997; Keiding and Clayton, to appear). Thus, the standardized mortality ratio arises from a *multiplicative* model involving the known population hazard $\alpha^*(t)$. Another classical tool for comparing with expected survival, the so-called *expected survival function*, arises from an *additive or excess* hazard model (Andersen and Væth, 1989).

The Cox Regression Model

In many applications of survival analysis the interest focuses on how *covariates* may affect the outcome; in clinical trials adjustment of treatment effects for effects of other explanatory variables may be crucial if the randomized groups are unbalanced with respect to important prognostic factors, and in epidemiological cohort studies reliable effects of exposure may be obtained only if some adjustment is made for confounding variables. In these situations, a *regression model* is useful and the most important model for survival data is the Cox (1972) *proportional hazards regression model*. (See Prentice (1991) and Reid (1994) for descriptions of the genesis of the Cox model.) In its simplest form it states the hazard function for an individual, i, with covariates $Z_i = (Z_{i1}, \ldots, Z_{ip})'$ to be

$$\alpha_i(t; Z_i) = \alpha_0(t)\exp(\beta' Z_i) \qquad [7]$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression coefficients and $\alpha_0(t)$, the *baseline hazard*, is the hazard function for individuals with all covariates equal to 0. Thus, the baseline hazard describes the common shape of the survival time distributions for all individuals while the *relative risk* function $\exp(\beta' Z_i)$ gives the level of each individual's hazard. The interpretation of the parameter, β_j for a dichotomous $Z_{ij} \in \{0,1\}$ is that $\exp(\beta_j)$ is the hazard ratio for individuals with $Z_{ij} = 1$ compared to those with $Z_{ij} = 0$, all other covariates being the same for the two individuals. Similar interpretations hold for parameters corresponding to covariates taking more than two values.

The model is *semiparametric* in the sense that the relative risk part is modeled parametrically while the baseline hazard is left unspecified. The regression coefficients β are estimated by maximizing the Cox (1975) *partial likelihood*

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' Z_i)}{\sum_{j \in R_i} \exp(\beta' Z_j)} \right]^{D_i} \qquad [8]$$

where $R_i = \{j : \widetilde{X}_j \geq \widetilde{X}_i\}$, the *risk set* at time $\widetilde{X}_i$, is the set of individuals still alive and uncensored at that time. Furthermore, the cumulative baseline hazard $A_0(t)$ is estimated by the *Breslow estimator*

$$\widehat{A}_0(t) = \sum_{\widetilde{X}_i \leq t} D_i / \sum_{j \in R_i} \exp\left(\widehat{\beta}' Z_j\right) \qquad [9]$$

which is the Nelson–Aalen estimator one would use if β were known and equal to the maximum partial likelihood estimate $\widehat{\beta}$. The estimators based on eqns [8] and [9] also have a nonparametric maximum likelihood interpretation. In large samples, $\widehat{\beta}$ is approximately normally distributed with the proper mean and with a covariance, which is estimated by the information matrix based on eqn [8]. This means that approximate confidence intervals for the relative risk parameters of interest can be calculated and that the usual large sample test statistics based on eqn [8] are available. Also, the asymptotic distribution of the Breslow estimator is normal; however, this estimate is most often used as a tool for estimating *survival probabilities* for individuals with given covariates, Z_0. Such an estimate may be obtained by the product integral $\widehat{S}(t; Z_0)$ of $\exp(\widehat{\beta}' Z_0)\widehat{A}_0(t)$. The *joint* asymptotic distribution of $\widehat{\beta}$ and the Breslow estimator then yields an approximate normal distribution for $\widehat{S}(t; Z_0)$ in large samples.

As is the case for many other statistical models for discrete data and survival data, the Cox regression model has no separate variance parameter but estimates the 'noise' from the mean value parameters. However, since Lin and Wei (1989), robust 'sandwich' variance estimates have been available for the Cox model, and these are now included in many software packages. See Keiding (2014) for some general remarks in this connection.

A number of useful extensions of this simple Cox model are available. Thus, in some cases the covariates are *time-dependent*, e.g., a covariate might indicate whether or not a given event had occurred by time t, or a time-dependent covariate might consist of repeated recordings of some measurement likely to affect the prognosis. In such cases the regression coefficients β are estimated replacing $\exp(\beta' Z_j)$ in eqn [8] by $\exp[\beta' Z_j(\widetilde{X}_i)]$.

Also a simple extension of the Breslow estimator eqn [9] applies in this case. However, the survival function can, in general, no longer be estimated in a simple way because of the extra randomness arising from the covariates, which is not modeled in the Cox model. This has the consequence that the estimates are more difficult to interpret when the model contains time-dependent covariates. To estimate the survival function in such cases a joint model for the hazard and the time-dependent covariate may be used. See our article on Event History Analysis in Continuous Time for some further remarks on time-dependent covariates and time-dependent confounding.

Another extension of eqn [7] is the *stratified* Cox model where individuals are grouped into a number, k of strata each of which has a separate baseline hazard. This model has important applications for checking the assumptions of eqn [7]. The model assumption of proportional hazards may also be *tested* in a number of ways, the simplest possibility being to add interaction terms of the form Z_{ij} $f(t)$ between Z_{ij} and time where $f(t)$ is some specified function. Also various forms of *residuals* as for normal linear models may be used for model checking in eqn [7]. A careful review of current methods for goodness-of-fit procedures of the Cox model was given by Martinussen and Scheike (2006).

There are many general reviews of the Cox model; among them we mention Sasieni (2005) and Oakes (2013).

Other Regression Models for Survival Data

Though the semiparametric Cox model is the regression model for survival data, which is applied most frequently, other regression models, e.g., *parametric* regression models also play important roles in practice. Examples include models with a multiplicative structure, i.e., models like eqn [7] but with a parametric specification, $\alpha_0(t) = \alpha_0(t;\ \theta)$, of the baseline hazard, and *accelerated failure time models*.

A multiplicative model with important epidemiological applications is the *Poisson regression* model with a piecewise constant baseline hazard (Breslow and Day, 1987; Clayton and Hills, 1993). In large data sets with categorical covariates this model has the advantage that a sufficiency reduction to the number of failures and the amount of person-time at risk in each *cell* defined by the covariates and the division of time into intervals is possible. This is in contrast to the Cox regression model eqn [7] where each individual data record is needed to compute eqn [8]. The substantial computing time required to maximize eqn [8] in large samples has also led to modifications of this estimation procedure. Thus, in *nested case–control studies* the risk set R_i in the Cox partial likelihood is replaced by a random sample $\widetilde{R}_i$ of R_i (see Borgan et al., 1995).

In the accelerated failure time model the focus is not on the hazard function but on the survival time itself much like in classical linear models. Thus, this model is given by $\log X_i = \alpha + \beta' Z_i + \varepsilon_i$ where the error terms are assumed to be independent and identically distributed with expectation 0. Examples include normally distributed (ε_i, $i = 1 \ldots n$), and error terms with a logistic or an extreme value distribution, the latter giving rise to a regression model with Weibull distributed life times.

Finally, we shall mention some nonparametric hazard models. In Aalen's additive model, $\alpha_i(t) = \beta_0(t) + \beta(t)' Z_i(t)$ (Aalen, 1980; Aalen et al., 2008), the regression functions $\beta_0(t), \ldots, \beta_p(t)$ are left completely unspecified and estimated nonparametrically much like the Nelson–Aalen estimator discussed above. This model provides an attractive alternative to the other regression models discussed in this section. There also exist more general and flexible models containing both this model and the Cox regression model as special cases (Martinussen and Scheike, 2006).

Other Kinds of Incomplete Observation

Salient features of survival data are right censoring and left truncation, and we have referred to them throughout this survey. In some cases only events (such as AIDS cases) that occur *before* a certain time are included (right-truncation).

More generally, event times may only be known to lie in an interval. If these intervals are the same for all individuals one may use the grouped time approach inherent in the Poisson regression model with piecewise constant intensities, or the discrete time approach, see e.g., Fahrmeir and Tutz (2001). If the intervals are not synchronous, more general techniques for interval-censored data (Sun, 2006) become relevant.

Concluding Remarks, Main References

Survival analysis is a well-established discipline in statistical theory, in biostatistics, and with significant presence both in econometrics and social statistics. Important general texts are the documentation of the actuarial and demographical traditions by Elandt-Johnson and Johnson (1980), the comprehensive texts by Kalbfleisch and Prentice (2002) and Lawless (2002) and the concise text by Cox and Oakes (1984). Andersen et al. (1993) covered the counting process approach and gave many details on the methods outlined in the present article; for follow-up texts see Martinussen and Scheike (2006) and Aalen et al. (2008). *The Encyclopedia of Biostatistics* (Second ed., Armitage and Colton, 2005) had many articles on survival analysis, collected in the separate volume edited by Andersen and Keiding (2006). The Bayesian approach was covered by Ibrahim et al. (2001). The recent *Handbook of Survival Analysis* (Klein et al., 2014) provides a fine overview, with some emphasis on biostatistical applications.

Among texts specifically intended for social and behavioral science users we mention the classic by Allison (1984), of which a second edition should be forthcoming in 2014, and texts by Blossfeld and Rohwer (1995), Courgeau and Lelièvre (1992) and Mills (2011).

See also: Event History Analysis in Continuous Time; Event History Analysis: Applications; Life Course and Event History Analysis; Life Table; Longitudinal Data: Event–History Analysis in Discrete Time; Multistate Transition Models in Demography; Stochastic Models.

Bibliography

Aalen, O.O., 1978. Nonparametric inference for a family of counting processes. Annals of Statistics 6, 701–726.

Aalen, O.O., 1980. A model for non-parametric regression analysis of counting processes. In: Klonecki, W., Kozek, A., Rosinski, J. (Eds.), Springer Lecture Notes in Statistics, Mathematical Statistics and Probability, vol. 2. Springer, New York, pp. 1–25.

Aalen, O.O., Borgan, Ø., Gjessing, H.K., 2008. Survival and Event History Analysis. Springer, New York.

Allison, P.D., 1984. Event History Analysis: Regression for Longitudinal Event Data. Sage, Beverly Hills/London.

Andersen, P.K., Borgan, Ø., Gill, R., Keiding, N., 1982. Linear non-parametric tests for comparison of counting processes, with applications to censored survival data (with discussion). International Statistical Review 50, 219–258. Amendment vol. 52, 225(1984).

Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Andersen, P.K., Keiding, N. (Eds.), 2006. Survival and Event History Analysis. Wiley, Chichester.

Andersen, P.K., Væth, M., 1989. Simple parametric and nonparametric models for excess and relative mortality. Biometrics 45, 523–535.

Armitage, P., Colton, T. (Eds.), 2005. Encyclopedia of Biostatistics, second ed. Wiley, Chichester.

Bernoulli, D., 1766. Essai dune nouvelle analyse de la mortalité, causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. In: Histoire avec les Mémoires, académie Royal des Sciences Paris 1760, pp. 1–45.

Blossfeld, H.-P., Rohwer, G., 1995. Techniques of Event History Modeling. Lawrence Erlbaum, Mahwah, NJ.

Böhmer, P.E., 1912. Theorie der unabhängigen Wahrscheinlichkeiten. Rapports, Mémoires et Procés – verbaux du 7e Congrés International d'Actuaires, Amsterdam 2, 327–343.

Breslow, N.E., 1991. Introduction to Kaplan and Meier (1958). Nonparametric estimation from incomplete observations. In: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics, vol. II. Springer, New York, pp. 311–318.

Borgan, Ø., Goldstein, L., Langholz, B., 1995. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. Annals of Statistics 23, 1749–1778.

Breslow, N.E., 1975. Analysis of survival data under the proportional hazards model. International Statistical Review 43, 45–58.

Breslow, N.E., Crowley, J.J., 1974. A large sample study of the life table and product limit estimates under random censorship. Annals of Statistics 2, 437–453.

Breslow, N.E., Day, N.E., 1987. Statistical Methods in Cancer Research. Vol. II The Design and Analysis of Cohort Studies. IARC Publications 82, International Agency for Research on Cancer, Lyon, France.

Clayton, D.G., Hills, M., 1993. Statistical Models in Epidemiology. University Press, Oxford.

Courgeau, D., Lelièvre, É., 1992. Event History Analysis in Demography. Clarendon, Oxford.

Cox, D.R., 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society Series B 34, 187–220.

Cox, D.R., 1975. Partial likelihood. Biometrika 62, 269–276.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman and Hall, London.

Dietz, K., Heesterbeek, J., 2002. Daniel Bernoullis epidemiological model revisited. Mathematical Biosciences 180, 1–21.

Ederer, F., Axtell, L.M., Cutler, S.J., 1961. The relative survival rate: a statistical methodology. National Cancer Institute Monographs 6, 101–121.

Elandt-Johnson, R.C., Johnson, N.L., 1980. Survival Models and Data Analysis. Wiley, New York.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modeling Based on Generalized Linear Models, second ed. Springer, New York.

Gill, R.D., Johansen, S., 1990. A survey of product-integration with a view towards application in survival analysis. Annals of Statistics 18, 1501–1555.

Greenwood, M., 1926. The natural duration of cancer. In: Reports on Public Health Medical Subjects, vol. 33. His Majesty's Stationery Office, London, pp. 1–26.

Hald, A., 1990. A History of Probability and Statistics and Their Applications before 1750. Wiley, New York.

Harris, T.E., Meier, P., Tukey, J.W., 1950. The timing of the distribution of events between observations. Human Biology 22, 249–270.

Ibrahim, J.G., Chen, M.-H., Sinha, D., 2001. Bayesian Survival Analysis. Springer, New York.

Johansen, S., 1978. Product limit estimator as maximum likelihood estimator. Scandinavian Journal of Statistics 5, 195–199.

Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data, second ed. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481, 562–563.

Keiding, N., 1987. The method of expected number of deaths 1786–1886–1986. International Statistical Review 55, 1–20.

Keiding, N., 2011. Age-period-cohort analysis in the 1870s: diagrams, stereograms, and the basic differential equation. Canadian Journal of Statistics 39, 405–420.

Keiding, N., 2014. Event history analysis. Annual Review of Statistics and Its Application 1, 333–360.

Keiding, N., Clayton, D. Standardization and control for confounding in observational studies: a historical perspective. Statistical Science, to appear.

Kilpatrick, S.J., 1962. Occupational mortality indexes. Population Studies 16, 175–187.

Klein, J.P., van Houwelingen, H.C., Ibrahim, J.G., Scheike, T.H. (Eds.), 2014. Handbook of Survival Analysis. Chapman and Hall, Boca Raton, FL.

Lawless, J.F., 2002. Statistical Models and Methods for Lifetime Data, second ed. Wiley, New York.

Lexis, W., 1875. Einleitung in die Theorie der Bevölkerungsstatistik. Trübner, Strassburg.

Lin, D.Y., Wei, L.J., 1989. The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association 64, 1074–1078.

Martinussen, T., Scheike, T.H., 2006. Dynamic Regression Models for Survival Data. Springer, New York.

Mills, M., 2011. Introducing Survival and Event History Analysis. Sage, Los Angeles.

Nelson, W., 1969. Hazard plotting for incomplete failure time data. Journal of Quality Technology 1, 27–52.

Nielsen, B., 1997. Expected survival in the Cox model. Scandinavian Journal of Statistics 24, 275–287.

Oakes, D., 2013. An introduction to survival models: in honor of Ross Prentice. Lifetime Data Analysis 19, 442–462.

Peto, R., Peto, J., 1972. Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society: Series A 135, 185–206.

Prentice, R.L., 1991. Introduction to Cox (1972) regression models and life-tables. In: Kotz, S., Johnson, N.L. (Eds.), Breakthroughs in Statistics, vol. II. Springer, New York, pp. 519–526.

Reid, N., 1994. A conversation with Sir David Cox. Statistical Science 9, 439–455.

Sasieni, P., 2005. Cox regression model. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, second ed., vol. 2. Wiley, Chichester, pp. 1280–1294.

Seal, H.L., 1977. Studies in the history of probability and statistics, XXXV. Multiple decrements or competing risks. Biometrika 64, 429–439.

Sun, J., 2006. The Statistical Analysis of Interval-censored Failure Time Data. Springer, New York.

Westergaard, H., 1901. Die Lehre von der Mortalität und Morbilität, vol. 2. Aufl. Jena: Fischer.

Westergaard, H., 1932. Contributions to the History of Statistics. King, London.

Survival Analysis: Overview

Ross L Prentice, University of Waterloo, Waterloo, ON, Canada
John D Kalbfleisch, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15318–15325, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Survival analysis, sometimes referred to as failure-time analysis, refers to the set of statistical methods used to analyze time-to-event data. Time-to-event or failure-time data, and associated covariate data, may be collected under a variety of sampling schemes, and very commonly involves right censoring. The distribution of a failure-time variate is usefully characterized in terms of its conditional failure rate, or hazard, function. Hazard-function modeling integrates nicely with the aforementioned sampling schemes, leading to convenient techniques for statistical testing and estimation. For univariate failure-time data those techniques include Kaplan–Meier estimators of the survivor function, censored data rank tests to compare the survival distributions of two or more groups, and relative risk (Cox) regression procedures for associating the hazard rate with a vector of study subject characteristics. Alternatively, linear models for the logarithm of failure time, for example, may be used for the regression analysis of failure-time data. Multivariate failure times may arise in the form of multiple events of various types on individual study subjects, or in the form of correlated failure times on distinct study subjects. The methods for the analysis of these types of data are still being actively researched. There are many additional specialized methods in survival analysis, including covariate measurement-error procedures, cohort-sampling strategies, and testing and estimation under sequential monitoring, to name just a few.

Failure-Time Data and Representation

The term survival analysis has come to embody the set of statistical techniques used to analyze and interpret time-to-event, or failure-time data. The events under study may be disease occurrences or recurrences in a biomedical setting, machine breakdowns in an industrial context, or the occurrence of major life events in a longitudinal-cohort setting. It is important in applications that the event of interest be defined clearly and unambiguously. A failure-time variable T is defined for an individual as the time from a meaningful origin or starting point, such as date of birth, or date of randomization into a clinical trial, to the occurrence of an event of interest. For convenience, we refer to the event under study as a failure.

A cohort study involves selecting a representative set of individuals from some target population and following those individuals forward in time in order to record the occurrence and timing of the failures. Often, some study subjects will still be without failure at the time of data analysis, in which case their failure times T will be known only to exceed their corresponding censoring times. In this case, the censoring time is defined as the time elapsed from the origin to the latest follow-up time, and the failure time is said to be right censored. Right censorship may also arise because subjects drop out the study, or are lost to follow-up. A key assumption underlying most survival analysis methods is that the probability of censoring an individual at a given follow-up time does not depend on subsequent failure times. This independence assumption implies that individuals who are under study at time t are representative of the cohort and precludes, for example, the censoring of individuals at time t because, among individuals under study at that time, they appear to be at unusually high (or low) risk of failure. (This assumption can be relaxed to independence between censoring and failure times conditional on the preceding histories of covariates.) The reasonableness of the independent censoring assumption, which usually cannot be tested directly, should be considered carefully in specific applications.

The distribution of a failure time variate $T > 0$ is readily represented by its survivor function $F(t) = P(T > t)$. If T is absolutely continuous, its distribution can also be represented by the probability density function $f(t) = -dF(t)/dt$ or by the hazard, or instantaneous failure rate, function defined by.

$$\lambda(t) = \lim_{\Delta t \downarrow 0} P(t \le T < t + \Delta t | T \ge t)/\Delta t = f(t)/F(t^-). \quad [1]$$

Note that

$$F(t) = \exp\{-\int_0^t \lambda(u)\mathrm{d}u\}, \quad \text{and}$$

$$f(t) = \lambda(t)\exp\{-\int_0^t \lambda(u)\mathrm{d}u\}$$

$$\text{for } 0 < t < \infty$$

In a similar way, the distribution of a discrete failure time variate T, taking values on $x_1 < x_2 < \cdots$, can be represented by its survivor function, its probability function$f(x_i) = P(T = x_i), i = 1,2\ldots$, or its hazard function $\lambda(x_i) = P(T = x_i | T \ge x_i) = f(x_i)|F(x_i^-)$. Note that

$$F(t) = P(T > t) = \prod_{i|x_i \le t} [1 - \lambda(x_i)] \quad [2]$$

as arises through thinking about the failure mechanism unfolding sequentially in time. To survive past time t where $x_j \le t < x_{j+1}$, an individual must survive in sequence each of the points $x_1,\ldots, x_j$ with conditional probabilities $1 - \lambda(x_i), i = 1,\ldots, j$.

It is often natural and insightful to focus on the hazard function. For example, human mortality rates are known to be elevated in the first few years of life, to be low and relatively flat

for the next three decades or so, and then to increase rapidly in the later decades of life. Similarly, comparison of hazard rates as a function of time from randomization for a new treatment compared with control can provide valuable insights into the nature of treatment effects. Hazard function modeling also integrates nicely with a number of key strategies for selecting study subjects from a target population. For example, if one is studying disease occurrence as it depends upon the age of study subjects, hazard function models readily accommodate varying ages at entry into the study cohort since the conditioning event in (1) specifies that the individual is surviving and event-free at age t. Again one needs to assume that selected subjects have hazard rates that are representative (conditional on covariates) of those at risk for the target population.

Estimation and Comparison of Survival Curves

Suppose that a cohort of individuals has been identified at a time origin $t = 0$ and followed forward in time. Suppose that failures are observed at times $t_1 < t_2 < \cdots < t_k$. Just prior to the failures observed at time t_j, there will be some number, r_j, of cohort members that are under active follow-up. That is, the failure and censoring times for these r_j members occur at time t_j or later. Of these r_j individuals, some number, d_j, are observed to fail at t_j. The empirical failure rate or estimated hazard at t_j is then d_j/r_j, and the empirical survival rate is $(1 - d_j/r_j)$. The empirical failure rate at all times other than $t_1,\ldots, t_k$ in the follow-up period is zero. The Kaplan and Meier (1958) estimator of the survivor function F at follow-up time t is.

$$\widehat{F}(t) = \prod_{\{t_j \le t\}} (1 - d_j/r_j) \qquad [3]$$

the product of the empirical survivor rates across the observed failure times less than or equal to t. Note that $\widehat{F}$ estimates the underlying survivor function F at those times where there are cohort members at risk. For (3) to be valid, the cohort members at risk at any follow-up time t must have the same failure rate operating as does the target population at that follow-up time. Thus, the aforementioned independent censoring assumptions are needed.

The hazard rate estimators $d_j/r_j, j = 1,\ldots, k$ can be shown to be uncorrelated, giving rise to the Greenwood variance estimator.

$$\mathrm{var}\{\widehat{F}(t)\} = \widehat{F}(t)^2 \sum_{\{t_j \le t\}} d_j/\{r_j(r_j - d_j)\} \qquad [4]$$

for $\widehat{F}(t)$. In-order to show asymptotic results for the estimator $\widehat{F}(t)$, for example that $\widehat{F}(t)$ is a consistent estimator of $F(t)$, it is necessary to restrict the estimation to a follow-up interval $[0, \tau)$ within which the size of the risk set (i.e., number of individuals under active follow-up) becomes large as the sample size (n) increases. Under these conditions, $n^{1/2}\{\widehat{F}(t) - F(t)\};\ 0 \le t < \tau$ can be shown to converge weakly to a mean-zero Gaussian process, permitting the calculation of approximate confidence intervals and bands for F on $[0, \tau)$. Applying the asymptotic results to the estimation of $log\{-log F(t)\}$, which has no range restrictions, is suggested as a way to improve the coverage rates of asymptotic confidence intervals and bands (Kalbfleisch and Prentice, 1980: pp. 14, 15).

Consider a test of equality of m survivor functions $F_1,\ldots, F_m$ on the basis of failure-time data from each of m populations. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered failure times from the pooled data, and again let d_j failures occur out of the r_j individuals at risk just prior to time t_j, $j = 1,\ldots, k$. Also, for each sample $i = 1,\ldots, m$, let d_{ij} denote the number of failures at t_j, and r_{ij} denote the number at risk at t_j^-. Under the null hypothesis $F_1 = F_2 = \cdots = F_m$ the conditional distribution of $d_{1j},\ldots, d_{mj}$ given $(d_j, r_{1j},\ldots, r_{mj})$ is hypergeometric from which the conditional mean and variance of d_{ij} are respectively $u_{ij} = r_{ij}(d_j/r_j)$ and $(V_j)_{ii} = r_{ij}(r_i - r_{ij})d_j(r_j - d_j)r_j^{-2}(r_j - 1)^{-1}$ The conditional covariance of d_{ij} and $d_{\ell j}$ is.

$$(V_j)_{i\ell} = r_{ij}r_{\ell j}d_j(r_j - d_j)r_j^{-2}(r_j - 1)^{-1}. \qquad [5]$$

Hence the statistic $v_j' = (d_{1j} - u_{1j}, \ldots, d_{mj} - u_{mj})$ has conditional mean zero and variance V_j under the null hypothesis. The summation of these differences across failure times.

$$v = \sum_1^k v_j \qquad [6]$$

is known as the logrank statistic (Mantel, 1966; Peto and Peto, 1972). The contributions to this statistic across the various failure times are uncorrelated, and $v'V^{-1}v$ generally has an asymptotic chi-square distribution with $m - 1$ degrees of freedom where $V = \sum_1^k V_j$ and a 'prime' denotes vector transpose.

This construction also shows that a broad class of 'censored data rank tests' of the form $\sum_1^k w_j v_j$ also provide valid tests of the equality of the m survival curves. In these tests, the 'weight' w_j at time t_j can be chosen to depend on the above conditioning event at t_j, or on data collected before t_j, more generally. The null hypothesis variance of these test statistics is readily estimated by $\sum_1^k w_j^2 V_j$, again leading to an asymptotic chi-square $(m - 1)$ test statistic. The special case $w_j = \widehat{F}(t_j^-)$ generalizes the Wilcoxon rank sum test to censored data in a natural fashion (e.g., Peto and Peto, 1972; Prentice, 1978) and provides an alternative to the logrank test that assigns greater weight to earlier failure times. More adaptive tests can also be entertained (e.g., Harrington and Fleming, 1982).

Failure-Time Regression: the Cox Model

Model Development

The same type of conditional probability argument generalizes to produce a flexible and powerful regression method for the analysis of failure-time data. This method was proposed in the fundamental paper by Cox (1972) and is referred to as Cox regression or as relative risk regression.

Suppose that each subject has an associated vector of covariates $z' = (z_1, z_2,\ldots)$. For example, the elements of z may include primary exposure variables and confounding factor measurements in an epidemiologic cohort study, or treatment indicators and patient characteristics in a randomized clinical trial. Let $\lambda(t; z)$ denote the hazard rate at time t for an absolutely continuous failure time T on an individual with covariate vector z. Now one requires the censoring patterns, and the cohort entry patterns, to be such that the hazard rate $\lambda(t; z)$ applies to the cohort members at risk at follow-up

time t and covariate value z for each t and z. The Cox regression model specifies a parametric form for the hazard rate ratio$\lambda(t;\, z)/\lambda(t;\, z_0)$, where z_0 is a reference value (e.g., $z_0 = 0$). Because this ratio is nonnegative, it is common to use an exponential parametric model of the form $exp\{x(t)\beta\}$. The vector $x(t) = \{x_1(t),\ldots,\, x_p(t)\}$ includes derived covariates that are functions of the elements of z (e.g., main effects and interactions) and possibly product (interaction) terms between z and t. The vector $\beta' = (\beta_1,\ldots,\, \beta_p)$ is a p-vector of regression parameters to be estimated. Denoting $\lambda_0(t) = \lambda(t;\, z_0)$ allows the model to be written in the standard form.

$$\lambda(t;z) = \lambda_0(t)\exp\{x(t)\beta\}. \qquad [7]$$

As in previous section let $t_1 < t_2 < \cdots < t_k$ denote the ordered distinct failure times observed in follow-up of the cohort, and let d_j denote the number of failures at t_j. Also let $R(t_j)$ denote the set of labels of individuals at risk for failure at t_j^-. We suppose first that there are no ties in the data so that $d_j = 1$. The conditional probability that the ith individual in the risk set with covariate z_i fails at t_j, given that $d_j = 1$ and the failure, censoring (and covariate) information prior to t_j, is simply.

$$\frac{\lambda(t_j;z_i)}{\sum_{\ell\in R(t_j)}\lambda(t_j;z_\ell)} = \frac{\exp\{x_i(t_j)\beta\}}{\sum_{\ell\in R(t_j)}\exp\{x_l(t_j)\beta\}}. \qquad [8]$$

The product of such terms across the distinct failure times.

$$L(\beta) = \prod_{j=1}^{k}\frac{\exp\{x_j(t_j)\beta\}}{\sum_{\ell\in R(t_j)}\exp\{x_\ell(t_j)\beta\}} \qquad [9]$$

is known as the partial likelihood function for the relative risk parameter β.

If one denotes $L(\beta) = \prod_{j=1}^{k} L_j(\beta)$, then the fact that the scores $\partial \log L_j(\beta)/\partial\beta,\ j = 1,\ldots,\, k$ have mean zero and are uncorrelated (Cox, 1975) suggests that, under regularity conditions, the partial likelihood (9) has the same asymptotic properties as does the usual likelihood function. Let $\widehat{\beta}$ denote the maximum partial likelihood estimate that solves $\partial \log L(\beta)/\partial\beta = 0$. Then, $n^{1/2}(\widehat{\beta} - \beta)$ will generally converge, as the sample size n becomes large, to a p-variate norma1 distribution having mean zero and covariance matrix consistently estimated by $-n^{-1}\partial^2 \log L(\widehat{\beta})/\partial\widehat{\beta}\,\partial\widehat{\beta}'$. This result provides the usual basis for tests and confidence intervals on the elements of β.

Even though we are working with continuous models, there will often be some tied failure times. If the ties are numerous, it may be better to model the data using a discrete failure-time model. Often, however, a simple modification to (9) is appropriate.When the d_j are small compared with the r_j, the approximate partial likelihood (Peto and Peto, 1972; Breslow, 1974)

$$L(\beta) = \prod_{j=1}^{k}\frac{\exp\{s_j(t_j)\beta\}}{\left[\sum_{\ell\in R(t_j)}\exp\{x_l(t_j)\beta\}\right]^{d_j}} \qquad [10]$$

is quite adequate for inference, where $s_j(t_j) = \sum_{i\in D_j} x_j(t_j)$ and D_j represents the set of labels of individuals failing at t_j. Efron (1977) suggested an alternate approximation that is only slightly more complicated and is somewhat better. Both the Breslow and Efron approximations are available in software packages.

Special Cases and Generalizations

Some special cases of the modeled regression vector $x(t)$ illustrate the flexibility of the relative risk model (7). If one restricts $x(t) = x$, to be independent of t, then the hazard ratios at any pair of z-values are constant. This gives the important proportional hazards special case of (7). For example, if the elements of z include indicator variables for $m - 1$ of m samples and $x(t) = (z_1,\ldots,\, z_{m-1})$ then each e^{β_j}, $j = 1,\ldots,\, m - 1$ denotes the hazard ratio for the jth sample compared with the last (mth) sample. In this circumstance the score test for $\beta = 0$ arising from (9) reduces to the logrank test of the preceding section.

The constant hazard ratio assumption, in this example, can be relaxed by defining $x(t) = \{g(t)z_1,\ldots,\, g(t)z_{m-1}\}$ for some specified function g of follow-up time. For example $g(t) = log\ t$ would allow hazard ratios of the form t^{β_j} for the jth sample compared with the mth. This gives a model in which hazard ratio for the jth versus the mth population increase ($\beta_j > 0$) or decrease ($\beta_j < 0$) monotonely from unity as t increases. The corresponding score test would be a weighted logrank test for the global null hypothesis. More generally one could define $x(t) = \{z_1,\ldots,\, z_{m-1},\, g(t)z_1,\ldots,\, g(t)z_{m-1}\}$. This model is hierarchical with main effects for the sample indicators and interactions with $g(t)$. A test that the interaction coefficients are zero provides an approach to checking for proportionality of the hazard functions. Departure from proportionality would imply the need to retain some product terms in the model, or to relax the proportionality assumption in some other way.

As a second special case consider a single quantitative predictor variable z. Under a proportional hazards assumption one can include various functions of z in the modeled (time independent) regression vector $x(t) = x$ for model building and checking, just as in ordinary linear regression. The proportionality assumption can be relaxed in various ways including setting $x(t) = \{xI_1(t),\ldots,\, xI_q(t)\}$ where $I_1(t),\ldots,\, I_q(t)$ are indicator variables for a partition of the study follow up period. The resulting model allows the relative risk parameters corresponding to the elements of $x = x(z)$ to be separately estimated in each follow-up interval. This provides a means for testing the proportionality assumption, and relaxing it as necessary to provide an adequate representation of the data.

The above description of the Cox model (7) assumed that a fixed regression vector z was measured on each study subject. It is important to note that (7) can be relaxed to include measurements of stochastic covariates that evolve over time. This does not complicate the development of (9) or (10). Specifically, if $z(t)$ denotes a study subject's covariates at follow-up time t, and $Z(t) = \{z(u);\, u < t\}$ denotes the corresponding covariate history up to time t, then the hazard rate $\lambda(t;\, Z(t)\}$ can be modeled according to the right side of (7) with $x(t)$, comprising functions of $Z(t)$ and product terms between such functions and time.

Careful thought must be given to relative risk parameter interpretation in the presence of stochastic covariates. For example, the interpretation of treatment effect parameters in a randomized controlled trial may be affected substantially by

the inclusion of postrandomization covariates when those covariates are affected by the treatment. However, careful use of such covariates can serve to identify mechanisms by which the treatment works.

Survivor Function Estimation

The cumulative 'baseline' hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)\mathrm{d}u$ (7) can be simply estimated (Breslow, 1974) by.

$$\widehat{\Lambda}_0\left(t;\widehat{\beta}\right) = \sum_{t_j \le t}\left[d_j \Big/ \sum_{\ell \in R(t_j)} \exp\left\{x_\ell(t_j)\widehat{\beta}\right\}\right]. \qquad [11]$$

If $x(t) = x$ is time-independent, the survivor function $F(t; z) = p\{T > t|z\} = exp\{-\Lambda_0(t)e^{x\beta}\}$ is estimated by.

$$\widehat{F}(t;\, z) = \exp\left\{-\widehat{\Lambda}_0\left(t;\, \widehat{\beta}\right)e^{x\widehat{\beta}}\right\} \qquad [12]$$

A slightly more complicated estimator $\widehat{F}(t; z)$ is available if x is time-dependent.

A practically important generalization of (7) allows the baseline hazard to depend on strata $s = s(z)$, or more generally $s = s(z, t)$ so that

$$\lambda(t; z) = \lambda_{0s}(t)\exp\{x(t)\beta\} \qquad [13]$$

where s takes integer values 1,2,…, m indexing the strata. This provides another avenue for relaxing a proportional hazards assumption. The stratified partial likelihood is simply a product of terms (9) over the various strata.

Asymptotic Distribution Theory

The score function $U(\beta) = \partial \log L(\beta)/\partial\beta$ from (9) involves a sum of terms that are uncorrelated but not independent. The theory of counting processes and martingales provides a framework in which this uncorrelated structure can be described, and a formal development of asymptotic distribution theory. Specifically, the counting process N for a study subject is defined by $N(t) = 0$ prior to the time of an observed failure, and $N(t) = 1$ at or beyond the subject's observed failure time. The counting process can be decomposed as

$$N(t) = \Lambda(t) + M(t) \qquad [14]$$

where, under independent censorship,

$$\Lambda(t) = Y(t)\int_0^t \lambda\{u; Z(u)\}\mathrm{d}u \qquad [15]$$

and $M(t)$ is a locally square integrable martingale. Note that in the 'compensator' Λ the 'at risk' process Y is defined by $Y(t) = 1$ up to the time of failure or censoring for the subject and zero thereafter. The decomposition (14) allows the logarithm of (9), and its derivatives, and the cumulative hazard estimator (11), to be expressed as stochastic integrals with respect to martingales, to which powerful convergence theorems apply. Anderson and Gill (1982) provide a thorough development of asymptotic distribution theory for the Cox model using these counting process techniques, and provide weak conditions for the joint asymptotic convergence of $n^{1/2}(\widehat{\beta} - \beta)$ and to $\widehat{\Lambda}(\cdot, \widehat{\beta})$ to Gaussian processes. The books by Fleming and Harrington (1991) and Andersen et al. (1993) use the counting process formulation for the theory of each of the statistics mentioned above, and for various other tests and estimators.

Other Failure-Time Regression Models

Parametric Models

Parametric failure-time regression models may be considered as an alternative to the semiparametric Cox model (7). For example, requiring $\lambda_0(t)$ in (7) to be a constant or a power function of time gives exponential and Weibull regression models respectively. The maximum partial likelihood estimate $\widehat{\beta}$ of the preceding section is known to be fully efficient within the semiparametric model with $\lambda_0(t)$ in (7) completely unspecified. While some efficiency gain in β estimation is possible if an assumed parametric model is true, such improvement is typically small and may be more than offset by a lack of robustness.

Of course there are many parametric regression models that are not contained in the Cox model (7). One such class specifies a linear model for the logarithm of time to failure.

$$\log T = \alpha + x\beta + \sigma u \qquad [16]$$

where $x = x(z)$ is comprised of functions of a fixed regression vector z. The parameters α and σ are location and scale parameters, β is a vector of regression parameters and u is an error variate. Because regression effects are additive on $\log T$, or multiplicative on T, this class of models is often referred to as the accelerated failure-time model. If the error u has an extreme value distribution with density $f(u)exp(u - e^u)$, then the resulting model for T is the Weibull regression model previously mentioned. More generally, any specified density for u induces a hazard-function model $\lambda(t; z)$ for T to which standard likelihood estimation procedures apply, and asymptotic distribution theory can be developed based again on counting process theory.

A Competing Semiparametric Model

The model (16) has also been studied as a semiparametric class in which the distribution of the error u is unspecified. Prentice (1978) considered the likelihood for the censored data rank vector based on the 'residuals' $v(\beta_0) = logT - x\beta_0$ to generate a linear rank statistic.

$$\sum_{j=1}^{k}\left(x_j - \bar{x}_j\right)w_j v_j(\beta_0) \qquad [17]$$

for testing $\beta = \beta_0$. In (17), $v_j(\beta_0)$ is the jth largest of the uncensored $v(\beta_0)$ values, x_j is the corresponding modeled covariate value, $\bar{x}_j$ is the average of the x-values for the individuals having 'residual' equal or greater than $v_j(\beta_0)$ and w_j is a weight, usually taken to be identically equal to one (logrank scores). An estimator of $\widehat{\beta}$ of β in (17) can be defined as a value of β_0 for which the length of the vector (17) is minimized. The calculation of $\widehat{\beta}$ is complicated by the fact that (17) is a step function that requires a reranking of the residuals at each trial value of β_0. However, simulated annealing methods have been developed (Lin and Geyer, 1992) for this calculation, and resampling methods have been proposed for the estimation of confidence intervals for the elements of β (Parzen et al., 1994). See Ying (1993) for an account of the asymptotic theory for $\widehat{\beta}$.

The model (16) implies a hazard function of the form.

$$\lambda(t;z) = \lambda_0(te^{-x\beta})e^{-x\beta}. \quad [18]$$

This form suggests a time-dependent covariate version of the accelerated failure-time model (e.g., Cox and Oakes, 1984; Robins and Tsiatis, 1992) given by

$$\lambda\{t, Z(t)\} = \lambda_0\{\int_0^t e^{-x(u)\beta}du\}e^{-x(t)\beta} \quad [19]$$

which retains the notion of the preceding covariate history implying a rescaling or acceleration, by a factor $e^{-x(t)\beta}$, of the rate at which a study subject traverses the failure-time axis.

Multiple Failure Times and Types

Multiple Failure Types and Competing Risks

Suppose that individuals are followed for the occurrence of a univariate failure time, but that the failure may be one of several types. The type i specific hazard function at follow-up time t, for an individual having preceding covariate $Z(t)$ can, for example, be specified as

$$\lambda_i\{t; Z(t)\} = \lambda_{0i}(t)\exp\{x(t)\beta_i\} \quad [20]$$

where λ_{0i} is a baseline hazard for the ith failure type and β_i is a corresponding relative risk parameter. Estimation of $\beta_1, \beta_2, \ldots$, under independent censorship is readily based on a partial likelihood function that is simply a product of terms (9) over distinct failure types (e.g., Kalbfleisch and Prentice, 1980: Chapter 7). In fact, inference on a particular regression vector β_i and corresponding cumulative hazard function Λ_{0i} is based on (9) upon censoring the failures of types other than (i) at their observed failure times. Accelerated failure time regression methods similarly generalize for the estimation of type-specific regression parameters.

Very often questions are posed that relate to the relationship between the various types of failure. It should be stressed that data on failure time and failure type alone are insufficient for studying the interrelation among the failure mechanisms for the different types, rather one generally requires the ability to observe the times of multiple failures on specific study subjects.

Recurrent Events

Suppose that there is a single type of failure, but that individual study subjects may experience multiple failures over time. Examples include multiple disease recurrences among cancer patients, multiple job losses in an employment history study, or multiple breakdowns of a manufactured item. The counting process N at follow-up time t is defined as the number of failures observed to occur on the interval $(0, t]$ for the specific individual. The asymptotic theory of Andersen and Gill (1982) embraced such counting processes in which the failure rate at time t was of Cox model form.

$$\lambda\{t; H(t)\} = \lambda_0(t)\exp\{x(t)\beta\}. \quad [21]$$

In this expression, the conditioning event $H(t)$ includes the individual's counting process history, in addition to the corresponding covariate history prior to follow-up time t. Hence the modeled covariate $x(t)$ must specify the dependence of the failure rate at time t both on the covariates and on the effects of the individual's preceding failure history. Various generalizations of (21) may be considered to acknowledge these conditioning histories in a flexible fashion. For example, λ_0 can be stratified on the number of preceding failures on the study subject; or the argument in λ_0 may be replaced by the gap time from the preceding failure (Prentice et al., 1981).

A useful alternative to (21) involves modeling a marginalized failure rate in which one 'forgets' some aspects of $H(t)$. For example, one could replace $H(t)$ in (21) by the covariate history $Z(t)$ and the number $s = s(t)$ of failures for the individual prior to time t. A Cox-type model $\lambda_{0s}(t)exp\{x(t)\beta_s\}$ for the resulting failure rate function would allow one to focus on covariate effects without having to model the timing of the s preceding failures. While the parameters β_s can again be estimated using the score equations from a stratified version of (9) a corresponding sandwich variance formula is needed to acknowledge the variability introduced by the nonincreasing conditioning event (Wei et al., 1989; Pepe and Cai, 1993; Lawless and Nadeau, 1995).

Correlated Failure Times

In some settings, such as group randomized intervention trials, the failure times on individuals within groups may be correlated, a feature that needs to be properly accommodated in data analysis. In other contexts, such as family studies of disease occurrence in genetic epidemiology, study of the nature of the dependency between failure times may be a key research goal.

In the former situation, hazard-function models of the types described above can be specified for the marginal distributions of each of the correlated failure times, and marginal regression parameters may be estimated using the techniques of the preceding sections. A sandwich-type variance estimator is needed to accommodate the dependencies between the estimating function contributions for individuals in the same group (Wei et al., 1989; Prentice and Hsu, 1997).

If the nature of the dependence between failure times is of substantive interest it is natural to consider nonparametric estimation of the bivariate survivor function $F(t_1, t_2) = p(T_1 > t_1, T_2 > t_2)$. Strongly consistent and asymptotically Gaussian nonparametric estimators are available (Dabrowska, 1988; Prentice and Cai, 1992), though these estimators are nonparametric efficient only if T_1 and T_2 are independent (Gill et al., 1995). Nonparametric efficient estimators away from this independence special case can be developed through a grouped data nonparametric maximum likelihood approach (Van der Laan, 1996), though this estimation problem remains an active area of investigation.

Frailty Models

There is also substantial literature on the use of latent variables for the modeling of multivariate failure times. In these models, failure times are assumed to be independent given the value of a random effect, or frailty, which is usually assumed to affect the hazard function in a multiplicative (time-independent) fashion.

One of the simplest such models arises if a frailty variate has a gamma distribution. This generates the Clayton (1978) bivariate model with joint survivor function.

$$F(t_1, t_2) = \left\{F_1(t_1)^{-\theta} + F_2(t_2)^{-\theta} - 1\right\}^{-1/\theta} \qquad [22]$$

where θ controls the extent of dependence between T_1 and T_2 with maximal positive dependence as $\theta \to \infty$, maximal negative dependence as $\theta \to -1$, and independence at $\theta = 0$ (e.g., Oakes, 1989).

Nielsen et al. (1992) generalize this model to include regression variables and propose an expectation-maximization algorithm with profiling on θ for model fitting. Andersen et al. (1993) and Hougaard (2000) provide fuller accounts of this approach to multi variate failure-time analysis.

Additional Survival Analysis Topics

Space permits only brief mention of a few additional survival analysis topics: Considerable effort has been devoted to model building and testing aspects of the use of failure-time regression models. The monograph by Therneau and Grambsch (2000) summarizes much of this work in the context of the Cox model (7).

The cost of assembling covariate histories has led to a number of proposals for sampling within cohorts, whereby the denominator terms in (9) are replaced by suitable estimates, based on a subset of the subjects at risk. These include nested case–control sampling in which the denominator at t_j is replaced by a sum that involves the individual failing at t_j and controls randomly selected from $R(t_j)$, independently at each distinct t_j, and case-cohort sampling in which the denominator summation at t_j consists of the failing individual along with the at risk members of a fixed random subcohort (see Samuelson (1997) for a recent contribution).

Considerable effort has also been directed to allowing for measurement error in regression variables, particularly in the context of the Cox model (7) with time-independent covariates. Huang and Wang (2000) provide a method for consistently estimating β in (7) under a classical measurement error model, and provide a summary of the literature on this important topic.

Testing and estimation on regression parameters has also been considered in the context of clinical trials under sequential monitoring. The recent book by Jennison and Turnbull (2000) summarizes this work.

Additional perspective on the survival analysis literature can be obtained from the overviews of Anderson and Keiding (1999) and of Oakes (2000). Goldstein and Harrell (1999) provide a description of the capabilities of various statistical software packages for carrying out survival analyses.

See also: Survival Analysis: Introduction.

Bibliography

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. Annals of Statistics 10, 1100–1120.

Andersen, P.K., Keiding, N., 1999. Survival analysis, overview. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, vol. 6. Wiley, New York, pp. 4452–4461.

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Breslow, N.E., 1974. Covariance analysis of censored survival data. Biometrics 30, 89–99.

Clayton, D., 1978. A model for association in bivariate life tables and its application in studies of familial tendency in chronic disease incidence. Biometrika 65, 141–151.

Cleves, M., 2008. An Introduction to Survival Analysis Using Stata. Stata Press, College Station, TX.

Cox, D.R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187–220.

Cox, D.R., 1975. Partial likelihood. Biometrika 62, 269–276.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman & Hall, London.

Dabrowska, D., 1988. Kaplan-Meier estimate on the plane. Annals of Statistics 16, 1475–1489.

Efron, B., 1977. Efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 72, 557–565.

Fleming, T.R., Harrington, D.D., 1991. Counting Processes and Survival Analysis. Wiley, New York.

Gill, R.D., Van der Laan, M., Wellner, J.A., 1995. Inefficient estimators of the bivariate survival function for three models. Annals of the Institute Henri Poincaré, Probability and Statistics 31, 547–597.

Goldstein, R., Harrell, F., 1999. Survival analysis, software. In: Armitage, P., Colton, T. (Eds.), Encyclopedia of Biostatistics, vol. 6. Wiley, New York, pp. 4461–4466.

Harrell, F.E., 2001. Regression Modeling Strategies: with applications to linear models, logistic regression, and Survival Analysis. Springer, New York.

Harrington, D.P., Fleming, T.R., 1982. A class of rank test procedures for censored survival data. Biometrika 69, 133–143.

Hougaard, P., 2000. Analysis of Multivariate Survival Data. Springer, New York.

Huang, Y., Wang, C.Y., 2000. Cox regression with accurate covariates unascertainable: a nonparametric correction approach. Journal of the American Statistical Association 95, 1209–1219.

Jennison, C., Turnbull, B.W., 2000. Group Sequential Methods with Application to Clinical Trials. Chapman & Hall, New York.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data. Wiley, New York.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457–481.

Lawless, J.F., Nadeau, C., 1995. Some simple and robust methods for the analysis of recurrent events. Technometrics 37, 158–168.

Lin, D.Y., Geyer, C.J., 1992. Computational methods for semi-parametric linear regression with censored data. Journal of Computational and Graphical Statistics 1, 77–90.

Lin, D.Y., Wei, J.L., Ying, Z., 1998. Accelerated failure time models for counting processes. Biometrika 85, 608–618.

Mantel, N., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer and Chemotherapy Reports 50, 163–170.

Nielsen, G.G., Gill, R.D., Andersen, P.K., Sorensen, T.I.A., 1992. A counting process approach to maximum likelihood estimation in frailty models. Scand Journal of Statistics 19, 25–43.

Oakes, D., 1989. Bivariate survival models induced by frailties. Journal of the American Statistical Association 84, 487–493.

Oakes, D., 2000. Survival analysis. Journal of the American Statistical Association 95, 282–285.

Parzen, M.I., Wei, L.J., Ying, Z., 1994. A resampling method based on pivotal estimating functions. Biometrika 68, 373–379.

Pencina, M.J., D'Agostino, Ralph B., 2004. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine 23 (13), 2109–2123.

Pepe, M.S., Cai, J., 1993. Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates. Journal of the American Statistical Association 88, 811–820.

Peto, R., Peto, J., 1972. Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society Series A 135, 185–206.

Prentice, R.L., 1978. Linear rank tests with right censored data. Biometrika 65, 167–179.

Prentice, R.L., Cai, J., 1992. Covariance and survivor function estimation using censored multivariate failure time data. Biometrika 79, 495–512.

Prentice, R.L., Hsu, L., 1997. Regression on hazard ratios and cross ratios in multivariate failure time analysis. Biometrika 84, 349–363.

Prentice, R.L., Williams, B.J., Peterson, A.V., 1981. On the regression analysis of multivariate failure time data. Biometrika 68, 373–379.

Robins, J.M., Tsiatis, A.A., 1992. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. Biometrika 79, 311–319.

Samuelsen, S.O., 1997. A pseudolikelihood approach to analysis of nested case-control studies. Biometrika 84, 379–384.

Therneau, T., Grambsch, P., 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York.

Van der Laan, M.J., 1996. Efficient estimation in the bivariate censoring model and repairing NPMLE. Annals of Statistics 24, 596–627.

Wei, I.J., Lin, D.Y., Weissfeld, L., 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of American Statistical Association 84, 1065–1073.

Ying, Z., 1993. A large sample study of rank estimation for censored regression data. Annals of Statistics 21, 76–99.

Survival Analysis: Whether, and If So, When?

Suzanne E Graham, Division of Educational Studies, University of New Hampshire, Durham, NH, USA
John B Willett and Judith D Singer, Harvard University Graduate School of Education, Harvard University, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, we introduce methods for addressing questions about the occurrence and timing of critical events in a person's life. Such questions are typically framed by asking whether and, if so, when the event will occur? Examples range from the occurrence and timing of macro-events such as high school graduation and dropout, marriage, abortion and divorce, onset of depression and relapse into drug and alcohol abuse, and so forth.

We discuss key issues in the design of event-history data collection and the ubiquitous problem of censoring that renders standard quantitative data analyses inhospitable for such data. We distinguish two approaches – discrete-time and continuous-time survival analysis – and focus on the former because of its conceptual clarity, flexibility, and ease of application. Throughout the article, we illustrate each step in this process using a running example of the occurrence and timing of first divorce in Wisconsin, in the latter half of the twentieth century. We end by revealing compelling extensions of the basic approach to the case of time-varying predictors, repeated spells, competing risks, and continuous-time methods. We suggest follow-up readings.

Scholars in the social and behavioral sciences often investigate the occurrence and timing of critical events in a person's life. Developmentalists pose questions about whether a child has achieved some moral, affective, or cognitive milestone, and by what age. Psychiatrists seek to understand the occurrence and timing of suicide, or relapse into drug abuse. Educators map out the patterning of school dropout and graduation, or entry into and exit from college. Whatever the nature of the underlying event, their questions all deal with the occurrence and timing of critical events. They inquire whether a critical event (such as relapse into drug abuse or first experience of sexual activity) has occurred, and if it occurred, then when. Beyond this, we also usually want to know how answers to such questions differ by features of participants' backgrounds, life experiences, treatment, and the like. For instance, when investigating the event of relapse into drug abuse, psychiatrists may want to confirm that the patterning of relapse differed for addicts who had been assigned to an innovative treatment versus those who were not. These are all questions about systematic interindividual differences in the patterning of critical events in a person's life.

In this article, we illustrate how survival (or event-history) analysis can be used to address these questions. Even though we refer to these methods as 'survival' analysis, they are ideal for investigating the occurrence and timing of any event that can be defined clearly and observed cleanly, both those that are of 'negative' consequence (e.g., death, divorce) and those that are 'positive' (e.g., birth of a first child, achievement of life goals). In presenting the methods, we use longitudinal data from the Wisconsin Longitudinal Study (WLS). This dataset contains data on a random and representative sample of 10 317 men and women in Wisconsin, followed from high school graduation in 1957 through 2004, with follow-ups in 1957, 1964, 1975, and 1992. We focus on the subsample of 7860 participants who entered into a first marriage during this period, some of whom divorced and some who did not. For this group, we also know their educational attainment on entry into the marriage (recorded as values of the dichotomous predictor, *DEGREE*, and coded 1 if the participant possessed a bachelor's degree at the start of the marriage, 0 otherwise).

Three Design Decisions That Underpin Survival Analysis

To study event occurrence, you must first define the event whose occurrence you wish to investigate. Events occur when someone makes a transition from one 'state' to another. In a survival-analytic application, these bounding states must be mutually exclusive and exhaustive. In our example, for instance, the two states in question are 'in a first marriage' versus 'divorced from a first marriage.' We are not saying that divorce is the only way a marriage can end – clearly, the death of either partner would also suffice. Nor are we saying that survival methods apply only in situations where such an either/or choice exists. The basic survival-analytic approach can be extended to situations in which folk can occupy three or more states, or in which a transition is repeated over and over. A single on/off event is a good place to start. Thus, here, we ask whether those who entered a first marriage in Wisconsin in the latter half of the twentieth century experienced a first divorce and, if so, when?

After defining the event of interest, you must also specify an occasion to be your 'origin' or 'beginning of time.' At this moment, everyone in the population must occupy the first of the two possible states, so that no one has yet experienced the event but is eligible to do so. In the language of survival analysis, you start the clock when everyone is 'at risk' of experiencing the event. In our example, participants are 'at risk' of experiencing a first divorce once they enter a first marriage, and so our 'beginning of time' is the date of the first marriage.

Next, you must decide on the scale you will use to measure time. Sometimes, time can be recorded very precisely. For example, we might know the exact number of days that a first

marriage lasted if we knew the explicit dates of marriage and divorce. Unfortunately, our WLS dataset provides only the years in which these milestones occurred, and so we measure the duration of the marriage very coarsely. In survival analysis, we distinguish between lifetimes recorded in narrow precise units and those recorded in coarser intervals by referring to the former as continuous-time data versus the latter discrete-time. Despite the obvious advantages of measuring everything as precisely as possible, scholars often find themselves with discrete-time data, as we do in our data example here. Some events may occur only at discrete points in time. Prospective students, for example, can often only graduate from college at preset times during the year. Or, financial or data-collection constraints may force researchers to use discrete intervals to record the passage of time to the nearest month or year.

The distinction between the continuous- and discrete-time approaches to survival analysis is not cosmetic. Every aspect of the process – the specification of statistical models, their parameter definition, methods of estimation, the testing, and interpretation – all depend on this choice. In this article, we focus on methods for event-history data collected in discrete time. In our conclusion, we comment on the extension to continuous-time methods.

The Ubiquitous Problem of Censoring

Whenever you investigate event occurrence, you face a difficulty that is not encountered by scholars working with other kinds of outcomes. What should you do with participants who do not experience the target event while you are observing them? In our WLS dataset, for instance, most of the participants whose first marriages were tracked did not divorce at all by 2004, the last year of data collection. Thus, for them, we have no information on the very 'time-to-divorce' that we wish to investigate and summarize. These are the censored cases (Cox and Oakes, 1984; Kalbfleisch and Prentice, 1980).

You cannot ignore censored respondents in analyses of 'time-to-divorce' because the very fact that their first marriages exceeded the length of data collection tells you much about the time to divorce, the event at the center of the research. Thus, although you do not know whether the event of first divorce ever occurred for the censored cases, perhaps after data collection was ended, you must incorporate this information legitimately into your summary statistics, for the latter to be credible (unbiased). For instance, to estimate the average length of a first marriage, in Wisconsin, we would need to include both uncensored and censored cases in your estimation; otherwise, we would be ignoring many of the folk with the longest marriages, longer than the period of data collection. To estimate such a statistic, it seems as though we must impute some sensible value of 'time-to-first-divorce' for the censored cases simply to retain them in our estimation, even though they never experienced the event while we watched. Otherwise, they would wind up excluded. But, what imputed value of 'time-to-divorce' would be most appropriate, and does assigning an arbitrary imputed value make sense, given that many of them may never divorce at all?

This is the conundrum at the heart of investigating event occurrence. No matter when data collection begins and how long it lasts, some members of the sample will be censored and have unknown event times. How can we remain true to the research question and have our findings reflect the representative sample with which we began the research? It is the resolution of this problem that is the centerpiece of survival analysis. The solution follows.

First, however, think about the mechanisms by which censoring occurs. In any study, the amount of censoring depends on two things. It depends on the rate at which events occur and on the length of data collection. If the event of interest is common and your data collection is long, then most participants in your sample may experience the event while you watch. There will be very few censored cases. On the other hand, if the event of interest is rare or your data collection curtailed, then censoring will be widespread. In our WLS dataset, even though data collection lasted decades, the target event – divorce from a first marriage – was not common. Among the 7860 respondents in our analytic subsample of 'first marrieds,' a huge number – 6826 – did not experience divorce while watched and were therefore censored.

Censoring is also exacerbated by other features of the event-history process under investigation. For instance, in addition to being censored if they remained in their first marriage at the end of data collection in 2004, participants in Wisconsin were censored for two other reasons. First, they could not still be married if either they or their spouses died during data collection. Therefore, in our data, partners who died or experienced the death of a spouse were also censored, not by the end of data collection but at the year of their spouse's death. Second, in large longitudinal studies like our WLS survey, respondents disappear via attrition. In our case, almost one-third of respondents were missing data in the last wave of data collection (2004). But, many of them had responded legitimately in the previous wave, and so their event histories are censored in 1992.

The most critical distinction is whether censoring was noninformative or informative. A noninformative censoring mechanism operates independently of the risk of event occurrence and is exogenous to it. Thus, if censoring is determined in advance by design or by the ending of grant funding, say, then it is noninformative. On the other hand, it would not be the case if participation in the study was ended because of endogenous actions taken by study participants (i.e., if they chose when they stopped participating) or because of unanticipated consequences of the study's outcomes that then impacted the research design retroactively (as when the evaluation of a drug treatment is cut short because of the declining health of some participants). If censoring is noninformative, we can assume that participants have been censored at random, and consequently the sample of uncensored participants who continue into the following time period remains representative of the population from the sample that was drawn in the previous period.

If censoring pops up in ways connected to the event occurrence itself, then censoring is informative and the unbiasedness of your analysis is threatened. For instance, let us suppose you conduct a study of time-to-relapse following treatment for cocaine abuse, contacting former addicts monthly to inquire if they are still abstinent. During your 2-year follow-up, you find that some of the former addicts are censored because

they stopped responding to your outreach. If this attrition was unrelated to their former habit, the censoring would remain uninformative. However, if they had relapsed to drug abuse and as a consequence stopped responding to you, through embarrassment or by dropping off the social map, then censoring would be informative, and connected to the occurrence of the event. In this case, you could no longer assume that the subsample of folk who remained uncensored after each wave of data collection continues to be representative of the participants in the preceding time interval, and your estimation of time-to-event will be challenged. No survival-analytic strategy can produce unbiased estimates of event occurrence and timing if censoring is informative.

Exploring Event Occurrence in Discrete-Time Data

The presence of censoring in event-history data makes standard statistical tools such as the estimation of means, the fitting of linear regression models, and so on inappropriate. This is because the presence of censoring indicates that important outcome information is absent for a legitimate portion of the sample. While you know that a censored individual did not experience the target event by the time of censoring, you do not know when s/he will experience it ultimately, if at all. Consequently, you do not have a value for what appears to be the conceptual outcome featured in the research question; you do not know the 'when,' the requisite time-to-event, for the censored person. The resolution that underpins survival methods is that while continuing to focus on the central notion of time-to-event, you must abandon statistics such as the sample mean and linked strategies like linear-regression analysis because they become meaningless in the face of censoring.

Instead, we focus on innovative approaches requiring the definition of new quantities known as the *hazard probability*, the *survival probability*, and the *median lifetime*, each of which can be estimated unbiasedly in the presence of censoring (Allison, 1984). These three new parameters are easily comprehended via an example of life-table analysis, a traditional technique for summarizing event-history data. It is useful not only for introducing the concepts of survival analysis but also for exploring discrete-time event-history data before embarking on model fitting and statistical inference to address your research questions formally.

In Table 1, we present a life table summarizing the occurrence of first divorce, among 'first marrieds' from Wisconsin. Having defined the start of the first marriage as the 'beginning of time,' we have constructed the table so that, following this inauguration, there is one row for each of the discrete-time 4-year intervals observed. Incidentally, we carved up the passage of time into these coarser 4-year discrete-time periods to reduce the complexity of subsequent statistical models, and then later confirmed that our conclusions were robust to the choice. In establishing the coarsening, we wanted to make sure that at least one person experienced the event of interest within each bin in order to avoid later difficulties in estimation. Finally, because no one in our first-married sample who remained married for more than 44 years actually got divorced before the end of data collection, we truncated the marriage span at 44 years (after 11 discrete 4-year time periods). We label the discrete-time intervals in the first two columns. There are 11, labeled intervals from '1' through '11.' We have used a standard notation (Singer and Willett, 2003) of an opening bracket '[' and a closing parenthesis ')' to indicate that each interval includes the start time and runs up to, but does not include, the closing time. Thus, interval [1, 4) includes the first day of the first year of marriage and runs up through the last day of the fourth year of marriage, and so on.

In the next three columns, we list frequencies that summarize event distribution in the sample, including the number of participants who entered each time interval still married (Column #3), the number who experienced a first divorce during the time interval (Column #4), and the number who were censored by the end of the time period (Column #5). All 7860 participants are in their first marriage at the beginning of time, and therefore 'at risk' of experiencing the event of interest, a first divorce. Then, during the first discrete-time interval (first 4 years of marriage), 150 participants divorced

Table 1 Life table displaying the sample distribution of divorce among the first married in Wisconsin, as a function of time (partitioned into discrete 4-year time periods)

		Number of participants…			*Proportion of participants…*	
Time period	*Covered duration (years)*	*In a first marriage at the beginning of each time period (risk set)*	*Divorced during the time period*	*Censored at the end of the time period*	*In a first marriage at the beginning of the time period, and divorced during it (hazard probability)*	*Still married at the end of the time period (survival probability)*
1	(1, 5)	7860	150	6	0.019	0.981
2	(5, 9)	7704	253	26	0.033	0.949
3	(9, 13)	7425	258	74	0.035	0.916
4	(13, 17)	7093	190	204	0.027	0.891
5	(17, 21)	6699	43	217	0.006	0.886
6	(21, 25)	6439	17	108	0.003	0.883
7	(25, 29)	6314	12	209	0.002	0.882
8	(29, 33)	6093	20	382	0.003	0.879
9	(33, 37)	5691	36	639	0.006	0.873
10	(37, 41)	5016	43	820	0.009	0.866
11	(41, 45)	4153	12	2061	0.003	0.863

and 6 were censored, leaving 7704 (=7860 − [150 + 6]) to enter the next interval. Of these, in their turn, 253 divorced and 26 were censored, leaving 7425 individuals to enter the third discrete-time interval and so on.

In survival analysis, we refer to those first marrieds who entered each successive discrete-time period as the risk set for that period, to indicate that they were eligible for or 'at risk' of experiencing a first divorce in that period. At the beginning of each discrete-time interval, the risk set contains everyone from the prior period's risk set minus those who divorced or were censored in the prior period. The definition of the risk set is inherently conditional. Any participant, in the risk set in any particular period, can only experience first divorce (or be censored) in that time period if s/he has not experienced it, nor been censored, in a prior period. Once s/he has experienced the event, or been censored, s/he disappears from the risk set for all future periods. This conditionality is a centerpiece of subsequent data summaries because it permits folk to contribute to those summaries up to their last moment of eligibility, regardless of termination by divorce or censoring.

We can define the 'risk' of event occurrence in any discrete-time period by defining the hazard probability, a summary that remains credible in the face of censoring (Allison, 1984). It is the conditional probability that someone will experience the event of interest during a particular discrete-time period, given that s/he did not already experience it, in an earlier period. As usual, hazard probability is defined in the population as a parameter and then estimated in the sample. Thus, in the population, the hazard probability is

$$h(t_{ij}) = \Pr[T_i = j | T_i \geq j] \qquad [1]$$

where T is a discrete random variable whose values identify the time period j in which an individual i can experience the event. The logic of the definition is that it stipulates hazard as the probability that the event occurs in the jth discrete time period, given that it must occur in this, or some future, period – in other words, that it had not occurred earlier. A list of discrete-time hazard probabilities, for all j, in the population is referred to as the *discrete-time hazard function*.

We can obtain maximum-likelihood estimates of the discrete-time hazard probabilities straightforwardly, by computing the sample proportions of participants who were at risk at the beginning of each time period and who experienced the event by the end of the time period (Allison, 1984). These sample hazard probabilities describe the pattern of first divorce in Wisconsin (Column #6). For example, during time period #1, out of 7860 high school graduates who entered a first marriage, 150 divorced before their fourth anniversary, at the end of the first period. Thus, the estimated hazard probability is 0.019 (or 150/7860) in this period. At the end of the period, six respondents had also been censored, leaving 7704 (or 7860 − [150 + 6]) participants, now married for at least 4 years, to enter the risk set in the second discrete-time period. Of these, 0.033 (or 253/7704) then divorced before their eighth anniversary, providing the estimated hazard probability in the second period, and so on.

We plot these estimated discrete-time hazard probabilities against time to provide the estimated hazard function, in the top panel of **Figure 1**. In the figure, we have adopted some conventions. First, we have plotted the estimated risks vertically

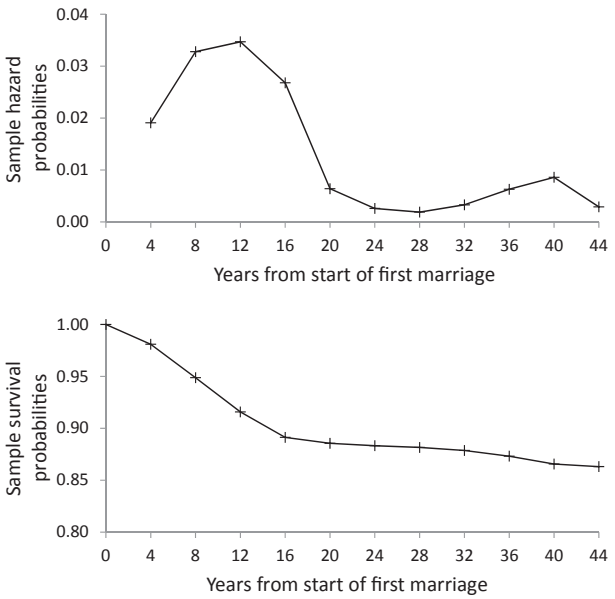

Figure 1 Sample hazard and survivor functions displaying the risk of first divorce as a function of the duration of a first marriage (scaled in discrete 4-year periods) (from **Table 3**).

above the number of full years of marriage that had occurred by the end of each period. Thus, even though the first period is labeled as [1, 5) in **Table 1**, the estimated hazard probability for the first period is plotted at '4 years' to indicate that 4 years of marriage had passed by the end of this period. Second, we joined the plotted points by line segments to suggest an envelope of risk (Miller, 1980).

Inspection of the estimated hazard function helps identify especially risky time periods, when respondents will most likely experience a first divorce. For instance, the estimated hazard function rises over the first three discrete time periods, peaking at 0.035 after 12 years. In other words, among participants who remain in first marriages for 12 years, 3.5% divorce in the third time period. Then, the estimated hazard function drops dramatically, and risk of divorce between 20 and 32 years of marriage is almost zero. Finally, there is a second lower peak at around 40 years, suggesting a marginally greater risk of divorce later in marriage.

You can also cumulate this period-by-period risk information and redisplay it in a second profile, the survivor function, to describe the distribution of event occurrence over time in an alternative way (Allison, 1984). The population survival probability, $S(t_{ij})$, is the probability that an individual i will survive beyond a particular time period j in the population. With no censoring, survival probability at any time can be estimated directly by computing the proportion of individuals who have not experienced the target event by the end of the requisite discrete-time period. However, when censoring is present, we use an indirect approach to estimate the survival probability in each time period:

$$\widehat{S}(t_j) = \widehat{S}(t_{j-1})\left[1 - \widehat{h}(t_j)\right] \qquad [2]$$

Here, the estimated survival probability in any time period (j) equals the estimated survival probability in the previous

time period ($j-1$) multiplied by the complement of the estimated hazard probability for the current time period (j). The logic is straightforward. For instance, the survival probability at the start of the first period is 1.000 by definition, because all marriages must be intact at the beginning of time. By the end of that time period (year 4 of marriage), 0.981, or 98.1%, of the respondents had not yet divorced. We know this because, during this period, the estimated hazard probability is 0.019 and so this proportion of the entering risk set did divorce and its complement ($1-0.019$) did not. Since the period began with everyone at risk (survival probability of 1.000), we know that 0.981 out of 1.000, or 0.981, must have survived the period undivorced and continued into the second period. The algorithm can then be executed sequentially, over successive periods. Thus, we enter the second period with 0.981, or 98.1% of first marrieds surviving. During the second period, 0.033 of the entering risk set experienced the event of interest, and so ($1-0.033$), or 0.967 of them, did not. Thus, at the end of the second period, a proportion of 0.967 of the original 0.981 (0.949) remained undivorced, and so on.

Graphing the estimated survivor function over time summarizes trends in event occurrence visually (bottom panel of Figure 1). While its slope is always negative, the shallowness of the slope between 0 and 4 years indicates that most marriages 'survive' during this period. Between years 4 and 16, the function declines more steeply, reflecting the larger risk of first divorce here. By the end of year 16 of a first marriage, 89% of our participants have not yet divorced. The remainder of the estimated survivor function is then relatively flat, with another slight dip between years 36 and 40, where risk of a first divorce again elevated slightly.

It can be useful to provide a single number summary in response to the question, 'How long does a first marriage typically last in Wisconsin?' If there were no censoring, event times would be known for everyone and a sample average time to divorce could be computed. But, in the face of censoring, we need an estimate of central tendency that respects the long-time presence of undivorced participants in the sample. The median lifetime is suitable for this purpose and is the time by which the survival probability falls to exactly one-half. It can be estimated by constructing a line across the estimated survivor function (Miller, 1980), at a probability of 0.5, and then projecting downward onto the time axis. By the median lifetime, half of all first married have divorced and half have not. Unfortunately, here, marriages last so long that the 50–50 point is never reached during data collection, and we cannot estimate a median lifetime. However, we can estimate other quantiles instead, such as a '90th Percentile Lifetime,' which, by inspection in Figure 1, is around 12.5 years.

Specifying and Fitting Statistical Models for the Discrete-Time Hazard Probability

Life-table analysis provides useful summaries of the event history, even when censoring has occurred. More intriguingly, these same estimates of risk can be replicated by formatting the event-history data in a special way – in a person-period dataset – and conducting standard logistic-regression analyses in that dataset. We illustrate this strategy below. It not only replicates our life-table results but also provides a regression-like framework within which event occurrence can be investigated formally, with predictors of risk incorporated into the statistical model and supported by the tools of statistical inference. This latter strategy is known as discrete-time hazard modeling (Allison, 1984).

Formatting the Event-History Data in a Person-Period Dataset

To engage the new approach, you must first format your event-history data in a person-period format, in which each participant contributes one record (row) for each time period that they are at risk of the event of interest. In Table 2, we present an example of event-history data, for three folk in our sample. The first two participants divorced before the end of data collection and are therefore not censored; each has a known divorce time. The first (#1025) divorced during the second 4-year

Table 2 Fragment of a person-period dataset containing event-history data on the duration of a first marriage (in discrete units of 4 years) for three participants who began a first marriage in Wisconsin between 1957 and 2004 and whose marriages were terminated either by divorce or censoring, with information on whether they entered the marriage with a bachelor's degree

ID	*Period*	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*Divorce*	*DEGREE*
1025	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1025	2	0	1	0	0	0	0	0	0	0	0	0	1	0
1050	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1050	2	0	1	0	0	0	0	0	0	0	0	0	0	0
1050	3	0	0	1	0	0	0	0	0	0	0	0	0	0
1050	4	0	0	0	1	0	0	0	0	0	0	0	1	0
1078	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1078	2	0	1	0	0	0	0	0	0	0	0	0	0	1
1078	3	0	0	1	0	0	0	0	0	0	0	0	0	1
1078	4	0	0	0	1	0	0	0	0	0	0	0	0	1
1078	5	0	0	0	0	1	0	0	0	0	0	0	0	1
1078	6	0	0	0	0	0	1	0	0	0	0	0	0	1
1078	7	0	0	0	0	0	0	1	0	0	0	0	0	1

discrete-time period, between 4 and 8 years after marriage, and the second (#1050) divorced between 12 and 16 years after marriage, in their fourth discrete-time period. On the other hand, the third participant (#1078), who married in 1977, had not gotten divorced by the end of data collection in 2004 and so was censored at the end of their seventh time period. To record this information, the person-period dataset includes a variable, *PERIOD*, to label the time period j to which each record (row) refers, ranging from 1 to 7, in this case. In the dataset, we have also described the particular time period to which each record refers by including a vector of dichotomous time-indicator variables, P_1 through P_{11}. If the row corresponds to the first period of the record for a particular participant, then the value of indicator P_1 is 1, and P_2 through P_{11} are zero. In the second row for any person, indicator P_2 has been set to 1 (and all others to zero), and so on. Later, we use these dichotomous indicators in our discrete-time hazard models to provide a general specification for time.

Our person-period dataset also contains an explicit dichotomous event indicator, *DIVORCE*, to describe whether first divorce occurred, during each discrete-time period (=1; 0 otherwise). Because of the conditionality implicit in our definition of risk, for each person who ultimately divorces (such as #1050), the value of the event indicator must be 0 in every record except the last one, in which the divorce occurred, and there it is set to 1. Whereas censored folk (like #1078) never experience a first divorce in any time period, for them, the value of *DIVORCE* remains 0 throughout all their records. Finally, the values of other important variables, potential predictors of risk, can also be recorded in the person-period dataset. In our case, we have included hypothesized dichotomous predictor *DEGREE*, whose values record whether the participant entered their first marriage with (=1) or without (=0) a college degree. Notice that, by virtue of its definition, the values of this predictor are time invariant and remain identical across all the rows contributed by each particular individual. However, it is also easy to imagine how the values of a time-varying predictor might also be recorded straightforwardly in the person-period dataset, with values that differed across the rows within person.

Fitting an Initial Discrete-Time Hazard Model in the Person-Period Dataset

If you imagine the complete person-period dataset of which Table 2 is a fragment, you will recognize that questions about whether, and if so when, first divorce occurs in Wisconsin are simple questions about the hypothesized relationship between the dichotomous outcome *DIVORCE* and time. Consequently, you can address this question by conducting a logistic-regression analysis of the relationship between *DIVORCE* and one of the specifications for time – either the linear effect of *PERIOD* or the general specification of time using indicators P_1 through P_{11} – in the person-period dataset. You would fit the following hypothesized discrete-time hazard model:

$$h(t_{ij}) = \frac{1}{1 + e^{-(\alpha_1 P_{1ij} + \alpha_2 P_{2ij} + \alpha_3 P_{3ij} \ldots + \alpha_{11} P_{11ij})}} \qquad [3]$$

Usefully, you can manipulate this model algebraically into a mathematically equivalent 'logit-hazard' specification, becoming

$$\begin{aligned} \text{logit } h(t_{ij}) &= \log_e \left\{ \frac{h(t_{ij})}{1 - h(t_{ij})} \right\} \\ &= \alpha_1 P_{1ij} + \alpha_2 P_{2ij} + \alpha_3 P_{3ij} \ldots + \alpha_{11} P_{11ij} \end{aligned} \qquad [4]$$

where 'logit' defines the natural logarithm of a probability (h) divided by its complement ($1 - h$) or as the log-odds of divorce occurrence (Singer and Willett, 2003).

Notice that we have not included a formal intercept in either specification. Such a 'no-intercept' specification is always available, in any statistical model, providing that other facets of the model are present to play the role. Here, instead of including a single overarching intercept (and omitting one of the time dummies to create a reference category), we have included several α parameters (α_1–α_{11}) to serve as multiple intercept terms, each capturing the risk of a first divorce during the corresponding discrete-time period. Parameter α_1 describes the logit-risk, or log-odds, of divorce in the first time period, α_2 in the second time period, α_3 the third, and so on. This model is our initial 'time-only' discrete-time hazard model.

Be assured of the veracity of our claim that standard logistic-regression methods can be used to fit this initial discrete-time hazard model, in the person-period dataset, and provide unbiased estimates of risk, along with suitable estimates of precision, fit, and inference. This claim is valid even though the model contains no explicit residual (as does a linear-regression model, say) and even though the data in the person-period dataset are multilevel in structure, with multiple rows being contributed by each person. The omission of an explicit 'residual' term is standard in regular logistic-regression models (Hosmer and Lemeshow, 2000) because any required uncertainty in the outcome is introduced, not by including an unobserved residual (with assumed distributional properties) but by the formulation of the outcome as a probability. In addition, interconnections among the unobserved components of the outcome, across the multiple rows of data within a person, are embedded implicitly in the model by our definition of the outcome as a conditional probability. We provide a formal proof of this claim in our companion technical paper (Singer and Willett, 1993). Consequently, you can use any standard logistic-regression routine from any statistical package to fit the models, provided your data are recorded in the person-period format. In this article, we used PROC LOGISTIC in SAS.

Recovering Estimated Hazard Probability from a Fitted Initial Discrete-Time Hazard Model

Having fitted the initial discrete-time hazard model in eqn [3] or eqn [4], you can recover period-by-period maximum-likelihood estimates of the hazard probabilities as fitted values. In the fourth column of Table 3, we list the estimated α intercept parameters. These are period-by-period estimates of the initial 'logit-hazards.' Although they are all negative (because the corresponding fitted hazard probabilities are less than 0.5), their increasing values for time periods 1–3 demonstrate an increase

Table 3 Parameter estimates and fitted hazard probabilities from an initial fitted discrete-time hazard model in which the risk of first divorce is presented as a function of a general specification of time, measured in discrete periods of 4 years ($n = 7860$)

Discrete-time period	*Predictor*	*Parameter*	*Parameter estimate* $(\widehat{\alpha}_j)$	*Fitted hazard probability* $\widehat{h}(t_j) = \frac{1}{1+e^{-\widehat{\alpha}_j}}$
1	P1	α_1	−3.940	0.019
2	P2	α_2	−3.383	0.033
3	P3	α_3	−3.324	0.035
4	P4	α_4	−3.593	0.027
5	P5	α_5	−5.042	0.006
6	P6	α_6	−5.934	0.003
7	P7	α_7	−6.264	0.002
8	P8	α_8	−5.716	0.003
9	P9	α_9	−5.057	0.006
10	P10	α_{10}	−4.751	0.009
11	P11	α_{11}	−5.844	0.003

in the risk of divorce, followed by a decrease through time period 7, a slight uptick through time period 10, and an eventual downtick in time period 11. We have then converted the parameter estimates from their 'log-odds' form into fitted hazard probabilities. Substituting respective values for dichotomous time indicators, P_1 through P_{11}, into the logistic-model specification in eqn [3], the fitted hazard probability in discrete-time period j is

$$\widehat{h}(t_j) = \frac{1}{1 + e^{-(\widehat{\alpha}_j)}} \qquad [5]$$

Substituting estimated values of the α parameters into eqn [5], we obtain the fitted hazard probabilities in Column #5. They are identical to the life-table estimates in Table 1.

Adding Substantive Predictors to the Initial Discrete-Time Hazard Model

The fact that you can obtain equivalent estimates of risk using logistic-regression analysis in the person-period dataset has an important ramification. It means that you now have a regression-like framework for addressing questions about 'Whether, and if so, when?' In particular, this discrete-time hazard modeling approach can be extended to include the effects of additional substantive predictors, to test hypotheses about their impact on event occurrence. For example, we can investigate whether the risk profile of divorce differs between those who were more or less educated, at the point of marriage, by adding dichotomous predictor *DEGREE* to distinguish between those who had received a bachelor's degree on entry into marriage versus those who had not, as follows:

$$\text{logit } h(t_{ij}) = (\alpha_1 P_{1ij} + \alpha_2 P_{2ij} + \alpha_3 P_{3ij} \ldots + \alpha_{11} P_{11ij}) + \beta_1 DEGREE_i \qquad [6]$$

Here, parameter β_1 represents the hypothesized influence of educational attainment on the logit-hazard function, quantifying the overarching difference in logit-hazard between participants who entered their marriages with, and without, bachelor's degrees.

Table 4 Parameter estimates and approximate *p*-values for three fitted discrete-time hazard models in which the risk of a first divorce depends on the duration of the first marriage (in discrete periods of 4 years) and whether the participant possessed a bachelor's degree on entry into the marriage ($n = 7860$)

	Fitted discrete-time hazard		
Predictor	*A*	*B*	*C*
Time period			
P1	−3.940***	−3.876***	−3.896***
P2	−3.383***	−3.319***	−3.330***
P3	−3.324***	−3.260***	−3.268***
P4	−3.593***	−3.529***	−3.455***
P5	−5.042***	−4.976***	−4.893***
P6	−5.934***	−5.867***	−5.819***
P7	−6.264***	−6.199***	−6.272***
P8	−5.716***	−5.650***	−6.143***
P9	−5.057***	−4.992***	−5.214***
P10	−4.751***	−4.687***	−4.609***
P11	−5.844***	−5.791***	−5.844***
Bachelor's degree at first marriage			
DEGREE		−0.516***	
DEGREE × PERIOD			
DEGREE × PERIOD^2			
DEGREE × PERIOD^3			
DEGREE × PERIOD^4			
DEGREE × P1			−0.335
DEGREE × P2			−0.413*
DEGREE × P3			−0.438*
DEGREE × P4			−1.734***
DEGREE × P5			−2.069*
DEGREE × P6			−1.121
DEGREE × P7			0.049**
DEGREE × P8			1.479**
DEGREE × P9			0.744*
DEGREE × P10			−2.023*
DEGREE × P11			–
Goodness-of-fit			
−2LL Statistic	9996.13	9968.80	9922.56

∼p < .10; * p < .05; ** p < .01; *** p < .001.

We report parameter estimates and fit statistics for this new model in the third column of Table 4, as Model B. For comparison, we also include parallel statistics from our initial discrete-time hazard model fit, as Model A. Parameter β_1 has an estimated value of −0.516 in Model B. Because the estimate is negative, we know that the logit-hazard function for folk who were more educated on entry into marriage is lower in elevation than for those less educated. We illustrate this in the top panel of Figure 2, plotting the fitted logit-hazard profiles by group. The estimated logit-hazard profile for prototypical participants who had college degrees at marriage sits 0.516 points, on the logit-hazard scale, below that of folk without degrees, at every time point.

You can antilog a logistic-regression parameter estimate to provide an estimated odds ratio (Hosmer and Lemeshow, 2000). Here, antilogging −0.516 we obtain 0.597, which contrasts the odds of divorce in the college-degree-at-marriage group (numerator of the ratio) with the same odds in the no-degree group (denominator of the ratio). Generally, because it is easier to interpret an odds ratio that is greater

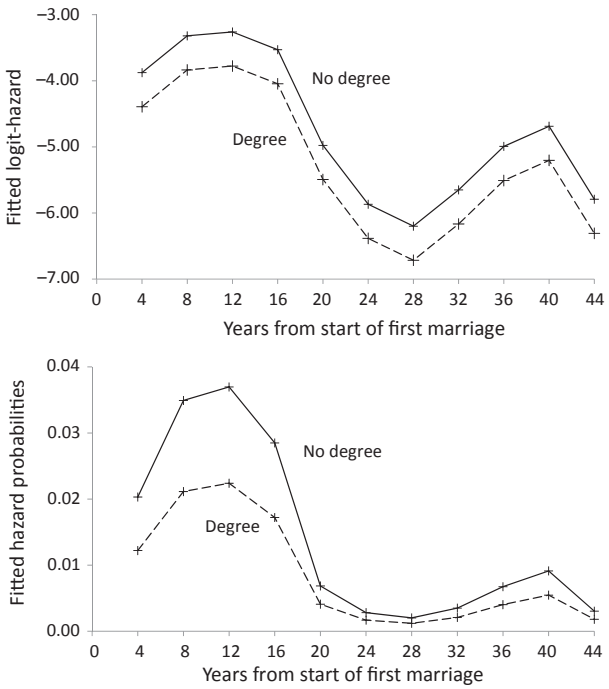

Figure 2 Fitted logit-hazard and hazard functions displaying the risk of a first divorce, in Wisconsin, as a function of first-marriage duration (in discrete periods of 4 years) for prototypical participants who had attained a bachelor's degree prior to entering into marriage (dashed profile) and did not (solid profile) (from fitted Model B, containing only the main effect of *DEGREE*, in **Table 4**).

than 1, we can flip this over to give an estimate of 1.68 (=1/0.597) and reinterpret. The estimated odds of a first divorce for a prototypical participant with no college degree at marriage are 1.67 times the odds for one with a college degree – risks of divorce are higher among the less educated, on entry into marriage. Typically, in any fitted logistic-regression model, each parameter estimate is accompanied by a single-parameter test, based on the z-statistic obtained by dividing the estimate by its standard error (or the equivalent Wald (χ^2) statistic, which is its square). In **Table 4**, we have not listed standard errors and z-statistics, to save space. But, in Model B, the standard error of $\widehat{\beta}_1$ was 0.106, and the z-statistic 4.89 ($p < .001$). We do list the latter approximate p-value as a superscript on the estimate, in the table, concluding – with great confidence – that there was a relationship between the main effect of *DEGREE* and risk of divorce, among first marrieds in Wisconsin.

We have added one more panel to **Figure 2**. In it, we have antilogged and detransformed to plot fitted hazard functions for prototypical participants who did, and did not, possess college degrees at marriage. Regardless of group, the profile of risk has a higher 'early' peak around 8–12 years of marriage and a lower 'later' peak around 36–40 years of marriage. However, the groups differ in their risk. Specifically, the estimated (hazard) probability of divorce during the first peak-risk period is around 4% (0.037) for folk with no degree at marriage, and about 2% (0.022) for those with degrees, given no divorce in a prior period.

One difference between the two panels is a consequence of the antilogging and detransforming that converted the fitted logit-hazard values in the top panel into the fitted hazard probabilities of the middle panel. This transformation underpins a ubiquitous property of these models. The inclusion of the main effect of *DEGREE* in the discrete-time hazard model seems to have simply shifted the fitted lower risk function ($DEGREE = 1$) vertically, and en bloc, to become the upper function ($DEGREE = 0$) in logit-hazard space. However, the same transition shows up as a magnification of the risk profiles in the space of hazard probabilities (bottom plot). In the latter, the vertical separation of the profiles is no longer identical in each time period. These visual differences are a mathematical property of the specification of the logistic model. (Actually, the magnification evident in the bottom panel is approximate, and true at low levels of risk, for hazard probabilities of a few percent. In reality, it is the corresponding odds profiles that are magnified with respect to one another, or proportional, because they are exponentiated versions of the corresponding vertically shifted logit-hazard functions.)

Adding Interactions between a Substantive Predictor and Time to a Discrete-Time Hazard Model

The parallel estimated logit-hazard functions in the upper panel of **Figure 2** imply that the effect of *DEGREE* on logit-risk is identical in every time period. This is often referred to as the proportional-odds assumption (Allison, 1984). This homogeneity of effect was compelled by our decision to include predictor *DEGREE* in eqn [6] only as a main effect. But, it is possible that the predictor's effect could differ from time period to period, particularly over a large portion of the human lifetime. Fortunately, we can relax this potentially unrealistic constraint by including statistical interactions between the predictor and time into the discrete-time hazard model, as follows:

$$\begin{aligned}\text{logit } h(t_{ij}) = &(\alpha_1 P_{1ij} + \alpha_2 P_{2ij} + \ldots + \alpha_{11} P_{11ij}) \\ &+ (\beta_1 DEGREE_i \times P_{1ij} + \beta_2 DEGREE_i \times P_{2ij} \\ &+ \ldots \beta_{11} DEGREE_i \times P_{11ij})\end{aligned} \quad [7]$$

Here, we have elected to drop the main effect of *DEGREE* and have, instead, included its two-way interactions with all of the time period indicators (rather than retaining the main effect of *DEGREE* and dropping one of the interaction terms, to avoid multicollinearity). From an interpretive perspective, we prefer this latter specification. In eqn [7], the original 'multiple intercept' α parameters now represent the population logit-hazard function of a 'baseline' group of persons who entered marriage without a college degree ($DEGREE = 0$). The new β parameters capture the unique interval-by-interval deviations in logit-hazard between this baseline group and those with a college degree at marriage ($DEGREE = 1$). Because there are 11 such interaction terms, spanning the observed period of marriage, the impact of *DEGREE* on the logit-risk of divorce can then be unique to each time interval. The parameter estimates and fit statistic for this new model are presented as Model C, in **Table 4**. Ironically, we could not actually estimate the last

interaction term, between *DEGREE* and time in the final period (#11), because during this last time period, no one who divorced had a college degree.

To test whether the proportional-odds assumption has been violated, we can use general linear hypothesis (GLH) testing to compare fitted models B and C. This will allow us to check whether the addition of the *DEGREE* by time interaction terms have made a statistically significant contribution to goodness-of-fit, as summarized by the −2LL statistic. In this particular comparison, we refer to Model C, with the greater number of parameters, as the 'full' model, and to Model B as the 'reduced' model (because we could constrain some of the parameters in Model C to zero, and 'reduce' it to become Model B). While the magnitude of the −2LL statistic itself is difficult to interpret directly, it is an index of 'badness-of-fit' – the worse the fit of a model, the higher the value of the −2LL statistic. Here, the −2LL statistic increases in magnitude from 9922.56 in the full model (C) to 9968.8 in the reduced model (B), a difference of 46.24, documenting the worsened fit of the simpler model on omission of the *DEGREE* by time interactions. However, is 46.24 a 'big' difference? In large samples, under maximum-likelihood theory, differences in the −2LL statistic between nested models have a χ^2 distribution with degrees of freedom equal to the number of additional parameters added to the full model to represent the additional hypothesized effect. In this case, we have added nine parameters (the difference between the 21 parameters in Model C and the 12 in Model B). The critical value of a χ^2 distribution with 9 degrees of freedom, at a Type I error of 0.05, is 16.92, which is less than the actual difference in the −2LL statistic between Models B and C, 46.24. Therefore, we judge the latter to be 'big,' and we reject the null and conclude that the interaction terms are needed, as a group. We prefer Model C to Model B, the proportionality assumption is violated, and the impact of *DEGREE* on logit-hazard is not identical in each time interval.

Consequently, in the top panel of Figure 3, based on Model C, we present revised fitted logit-hazard functions describing the risks of first divorce among prototypical participants who possessed a college degree at first marriage (dashed line) versus those who did not (solid line). In contrast with the fitted functions obtained from Model B, in Figure 2, there are now differences from time interval to interval in the impact of predictor *DEGREE*. In the first two discrete-time periods, for instance, after years 4 and 8, differences in logit-hazard between the two groups are small. Between 8 and 24 years, logit-hazard values describing the risks of divorce for those with degrees at marriage then drop more rapidly, representing a declining lower risk of a first divorce for these folk. During period 7 (28 years into marriage), the pattern of risk reverses and folk who entered the marriage with a college degree now have a higher logit-hazard of first divorce. By year 40, the pattern has flipped back, and logit-risk for the college-educated folk has become lower again. Overall, although they enjoy lower risks of divorce in most time periods, it seems as though the peaks in risk now arrive somewhat earlier for participants who entered their marriages with degrees, compared to those that did not.

To better understand differences in risk of divorce over time for folk who entered their marriages with, and without, college degrees, we have plotted fitted hazard and survivor functions

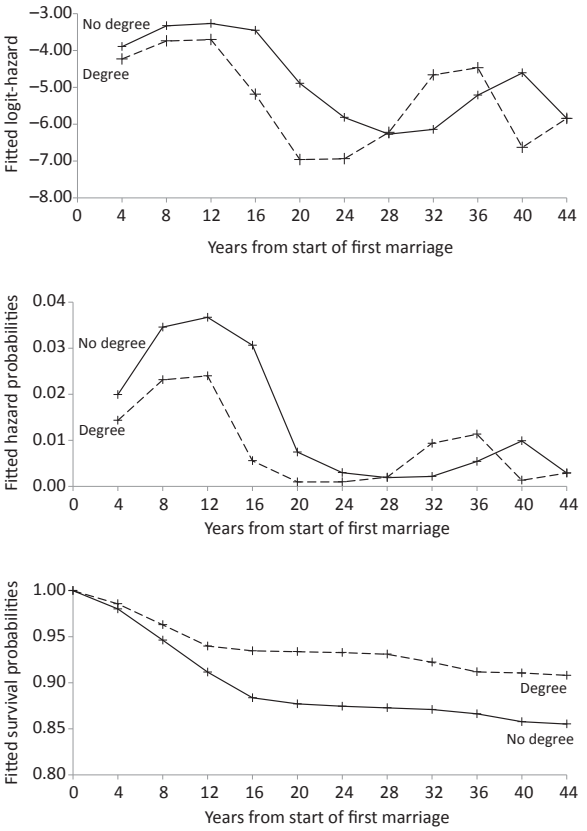

Figure 3 Fitted logit-hazard, hazard, and survivor functions displaying the risk of a first divorce, in Wisconsin, as a function of first-marriage duration (in discrete periods of 4 years) for prototypical participants who had attained a bachelor's degree prior to entering into marriage (dashed profile) and did not (solid profile) (from fitted Model C, containing the two-way interaction between *DEGREE* and a general specification of time, in Table 4).

computed from Model C, in the middle and bottom panels of Figure 3. The estimated (hazard) probability of a first divorce for prototypical holders of college degrees at marriage is largest in the second time period. Almost 3% of participants who were married at least 4 years were divorced by the end of 8 years of marriage. At this point, the hazard probability for the no-degree group is only slightly larger (0.034). However, after the second time period, the risk of a first divorce plummets for the group with degrees, while continuing to rise for the no-degree group. These latter participants attain their maximum risk in the third time period (12 years), where almost 4% of them divorce. In addition, a second, and lower, peak in the risk of first divorce is still observed for both groups, but at different later times: at period 9 (36 years) for the degree holders and at period 10 (40 years) for those with no college degree. From the fitted survivor functions, we estimate that 85% of the participants who entered their marriages without degrees have not experienced a first divorce by the end of time period 11 (44 years), compared with 91% of those who entered with degrees. These latter estimates of survival probability are very similar to those provided by the earlier fitted main-effects model. Therefore, the time-varying impact of

DEGREE does not affect the long-term probability of 'surviving' a first divorce, through 44 years of marriage; it just distributes the annual risk differentially, depending on educational attainment at marriage.

In general, with these discrete-time methods, you can continue the addition of further predictors and covariates to test hypotheses about their impact on the risk of event occurrence, in the same way as you would with regular linear- or logistic-regression analysis. You simply include the new variables in the person-period dataset and add them as main effects, or test their interaction with time, as we have done above with predictor *DEGREE*. You can also test for interactions among new substantive predictors themselves, in the usual way, by forming and adding the relevant cross-products. With discrete-time hazard modeling, all of the standard skills that you already possess in the credible modeling of relationships between outcome and predictors carry over into the investigation of risk, making it something that is readily available to scholars throughout the social and behavioral sciences.

Conclusion: Extending Discrete-Time Survival Analysis

We have used our limited space here to introduce the methods of discrete-time survival analysis. However, these methods can be extended easily in important ways, and we invite you to follow up on these extensions by consulting our companion volume (Singer and Willett, 2003) and those of our colleagues (Allison, 2010; Hosmer and Lemeshow, 2002). Below, we touch briefly on the most important extensions.

First, the type of event does not matter, as long as you can define and detect it explicitly. Thus, you can use survival methods to model the risks of marriage and divorce, as easily as milestones in the educational or professional career. Other potential questions could involve the onset of reckless behavior, such as alcohol and drug abuse during adolescence, or relapse to those addictions following treatment. Or, if you can identify explicit stages in cognitive, affective, or moral development with clarity, then their occurrence and timing also offer fertile ground for survival analysis.

Second, discrete-time hazard models can easily incorporate time-varying predictors, whose values differ over time. You have actually seen this implicitly in Model C, because the values of the cross-product of *DEGREE* and time that served to represent our two-way interaction are time varying in value. However, the utility of these methods for modeling the impact of time-varying predictors is broader than this. For instance, had we possessed time-varying information on the educational attainment of each participant, during marriage, rather than only knowing whether they had a college degree at the start of their marriage, we could have treated educational attainment as a time-varying predictor, updating its values in each row of the person-period dataset as the participant acquired more education, and including the now time-varying predictor in another discrete-time hazard model.

Third, as in most statistical modeling, discrete-time hazard models can be extended to include nonlinear effects of predictors, both temporal and substantive. In additional work with our WLS example, not reported here, for instance, we showed that the general specification for time could be replaced parsimoniously by a smooth quartic polynomial specification of the time predictor, without a dramatic reduction in goodness-of-fit. Of course, such nonlinear specifications could also be used with substantive predictors such as AGE at first marriage, if we were to add it to the model as a predictor of divorce. (Incidentally, we do find age at marriage to affect the risk of divorce – those who married younger were at much greater risk of divorce, especially in the early years of their marriage.) Generally, when needed any kind of transformation of the predictors can be incorporated into survival models. By the judicious use of both continuous and dichotomous versions of time, for example, you can specify spline and segmented profiles for the risk of event occurrence. The application is very broad.

Fourth, although we focused here on the risks of the single event of first divorce, you are not limited to investigating the occurrence and timing of events that occur only once over the lifetime. Survival methods extend readily to the modeling of repeated events, as when a first marriage and divorce are followed by a second, a third, and so on. The occurrence and timing of such 'multiple spells' can be modeled simultaneously in a single discrete-time hazard model expanded to incorporate dichotomous predictors to distinguish the spells. We describe this process in a technical paper (Willett and Singer, 1995).

Fifth, you can extend the basic discrete-time hazard modeling strategy to a 'competing-risks' condition, in which individuals can occupy three or more states, all of which compete for attention. Thus, you might examine the termination of marriage while treating the death of the respondent, the death of their spouse, or divorce as three viable events that 'compete' to end marriage. The extension is straightforward. Under a competing-risks assumption, a participant whose 'life' is terminated by one of the several competing events is simply censored on the others. This requires some minor modifications in the person-period dataset, but analyses remain the same.

Finally, the methods of discrete-time survival analysis can be extended to the case of continuous time. We ourselves have found discrete-time methods to be the most useful, because the event-history data that we encounter the most often are collected in discrete time and the principles underpinning the strategy are simple and conceptually appealing. However, a parallel methodological development, referred to as 'Cox-regression analysis,' is available for handling event-histories measured continuously (Cox, 1972). We also describe these latter methods in our companion volume (Singer and Willett, 2003). Reassuringly, as the width of discrete-time periods tends to zero, the findings of discrete- and continuous-time survival analysis converge.

See also: Policy Analysis; Prevention Research; Quantification in the History of the Social Sciences; Survival Analysis: Introduction.

Bibliography

Allison, P.D., 1984. Event-History Analysis: Regression for Longitudinal Data. Sage, Beverly Hills, CA.

Allison, P.D., 2010. Survival Analysis Using SAS: A Practical Guide, second ed. SAS Institute Inc, Cary, NC.
Cox, D.R., 1972. Regression models and life-tables. Journal of the American Statistical Association, Series B 34, 187–220.
Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman and Hall, London.
Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. John Wiley & Sons, Inc., New York.
Hosmer, D.W., Lemeshow, S., 2002. Applied Survival Analysis: Regression Modeling of Time to Event Data. John Wiley & Sons, Inc, New York.
Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure-Time Data. Wiley, New York.
Miller, R.G., 1980. Survival Analysis. Wiley, New York.
Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford, New York.
Singer, J.D., Willett, J.B., 1993. It's about time: using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics 18, 155–195.
Willett, J.B., Singer, J.D., 1995. It's deja-vu all over again: using multiple-spell discrete-time survival analysis. Journal of Educational Statistics 20, 41–67.

Sustainability: A Dialog of Values

John Blewitt, Aston Business School, Aston University, Birmingham, UK

Published by Elsevier Ltd.

Abstract

Sustainable development and sustainability are related concepts whose origins can be traced to the nineteenth century but have gained increasing prominence from the middle years of the twentieth century. Increasingly, scientists, business leaders, academics, politicians, and other leaders have recognized that the dominant model of industrialization, economic growth, and development was exceeding the natural biological limits of what the planet could practically sustain. This has led to a series of ongoing international conferences attempting to envisage theories, concepts, and practices that could reconcile sustainability with a development "that meets the needs of the present without compromising the ability of future generations to meet their own needs."

Some History

Until the industrialization of Europe in mid-eighteenth century, wood was the primary material used for fuel, construction, smelting, and shipbuilding. World trade and the navies of the great maritime powers relied on a seemingly inexhaustible supply of timber. But these people were wrong. Although timber is a renewable resource European nations were harvesting more trees than were being planted and nurtured to maturity. Governments in Britain, France, and particularly Germany slowly recognized that consuming so much timber was unsustainable. In *Sustainability: A Cultural History*, Ulrich Grober (2012: 88) discusses how many foresters and enlightened government ministers, including Johan Wolfgang von Goethe of Weimar, believed that 'the true capacities of the forests' should become the basis for their use and exploitation. The modern science of ecology, the concept of sustainability, and the practice of sustainable development were born. Many years later in the mid-twentieth century, Rachel Carson's (Carson, 1962) *Silent Spring*, first published in 1962, analyzed the devastating ecological impact chemical pesticides and huge agricultural businesses were having on the American countryside. This seminal book probably has come symbolizing the beginning of what became known as Earth Politics and inspired for establishment of the modern environmental movement.

In Europe and America the 1960s' and 1970s' economic growth and rampant consumerism led to a growing recognition that Western lifestyles and the ideology of development were compromising the ecological balance, stability, and security of the planet. This concern was amplified when a single image of our lonely and luminous planet Earth, taken from the Apollo Eight spacecraft in 1968, was published and disseminated across the world. The photograph, titled Earthrise, was interpreted by many people as revealing the fragile beauty of our blue planet. It had a truly powerful emotional impact. In 1972, a further photograph from the Apollo Project, Blue Marble, soon became the most published image in history thereby gaining an iconic status among many new sustainability advocates and the wider environmental movement. World famous pressure groups were formed, like Friends of the Earth and Greenpeace, and soon hit the headlines. Ecologically minded writers, artists, poets, and scientists, following in Rachel Carson's footsteps, came to prominence. Theodore Roszak's *Where the Wasteland Ends* (1972), and E.F. Schumacher's game changing *Small is Beautiful* (1973) were international best sellers and from the dismal science of economics Kenneth E. Boulding's *The Economics of the Coming Spaceship Earth*, created a metaphor which captured the public imagination. In this work, he clearly stated there were no unlimited reservoirs of anything and that humanity needed to find its place in the planet's cyclical ecological system which, although capable of continuous reproduction, continually needed inputs of energy to maintain itself. Thus, in 1970, Earth Day was the first major environmental event to have any real social, public, and cultural impact was held in the US. Two years later, in 1972, a global think tank known as The Club of Rome published what would soon become its landmark study, *Limits to Growth*, and the first serious international conference on global environmental issues occurred at the United Nations Conference on the Human Environment in Stockholm, Sweden.

Limits to Growth

The Club of Rome's report (Meadows et al., 1972) combined an optimism regarding the human potential to innovate and capacity to transcend environmental and demographic limitations with a series of detailed models and scenarios warning that if existing trends continued, negative economic and ecological consequences would undoubtedly follow. The Club of Rome's model for global development was constructed to investigate five major trends – accelerating industrialization, rapid population growth, widespread malnutrition, depletion of nonrenewable resources, and a deteriorating environment (*see also* Environment and Development; Limits to Growth). The authors looked to the future and posed some difficult questions: What do we want our world to be like? Can we continually keep expanding production and consumption? The answer was a clear: No. But achieving a self-imposed limitation to economic growth would require considerable effort, a massive change of mindset, and political practice. It would necessarily involve learning to do many things in new ways and even to stop doing others. It would tax the ingenuity, the

flexibility, will power, moral sense, and self-discipline of the human race. It would create disagreement and conflict and challenge many vested economic and financial interests. Bringing a deliberate, controlled end to growth would clearly be a tremendous challenge, not easily met and extremely difficult to countenance by governments, businesses, and ordinary citizens from across the world. There were many doubts, arguments, and questions too. Would the final result be worth it? What would humanity gain by such a transition and what would it lose? In the 30 years preceding the report publication, global economic growth continued and the world's unsustainable practices became increasingly obvious. Thirty years later, three of the original authors produced an update (Meadows et al., 2005). They reviewed many of the debates and criticisms their earlier work had generated, analyzed new evidence, rectified the flaws in, and sharpened their models and as a consequence amended their position somewhat. Nonetheless, they firmly reaffirmed and demonstrated that the theory of necessary limits to growth remained valid, vital, and significant. Indeed, more so in the twenty-first century than ever before.

Toward Sustainable Development

At the same time as the Club of Rome was doing its work, the General Assembly of the IUCN (World Conservation Union), a body established a little after World War II, met in New Delhi. Together with the newly formed World Wildlife Fund (later remained Worldwide Fund for Nature), the IUCN focused on developing new strategic thinking for animal and habitat conservation, ecological restoration, and human well-being. The relatively unfamiliar and under utilized concept 'quality of life,' as oppose to the more material and comprehensible concept 'standard of living,' became the centerpiece of IUCN thinking and policy development. The IUCN's aim was to intelligently link cultural diversity with ecological or what has become known today as biodiversity. In 1980, the IUCN published its World Conservation Strategy and so launched into the global public sphere the rather hopeful concept of sustainable development. For the IUCN, humanity's relationship with the biosphere would continue to deteriorate until a new ecologically sensitive international economic order and a new environmental ethic was developed. Indeed, economic development would continue but it needs to be socially equitable and environmentally sustainable. In the same year, the Brandt Commission published its *North-South: A Programme for Survival* which planted the responsibility for human survival firmly in the political arena at a time when leaders still seemed more concerned with the vagaries, conflicts, and posturing of the Cold War than addressing the equally, and perhaps even more, pressing issues of global poverty, disease, social inequality, injustice, self-determination, human rights, and the constant depletion of the world's natural resources. The Commission did not wish to redefine development, but duly noted:

> One must avoid the persistent confusion of growth with development, and we strongly emphasize that the prime objective of development is to lead to self fulfillment and creative partnership in the use of a nation's productive forces and its full human potential.
> Brandt, 1980: 23

In other words, the Brandt report suggested that development strategy should not be based upon unlimited and infinite economic growth or GDP as the sole measure of progress (*see* Economic Growth). The world as a whole should not use as a model for future generations the ecologically destructive industrial development paths pursued that had created and defined the West. The standard of life is not the same as the quality of life. More stuff does not necessarily mean a better happier life although a certain degree of material wealth and comfort is essential for human and social well-being. Development should focus on nurturing human well-being and ecological sensitivity rather than the relentless accumulation of material products. Every region of the world has its own ecological and cultural heritage and should as a result be enabled to chart its own distinct and distinctive path toward sustainability. Thus, wealth and development took on a qualitative as well as a quantitative aspect. Material and spiritual poverty both need addressing in policy and practice at local and global levels.

In 1983, work started on a major study by the World Commission on Environment and Development that would firmly establish sustainable development as the most significant concept and practice of our time. In 1987, the results were published as *Our Common Future* (the Brundtland Report). More than half of the Commission were representatives from developing countries, ensuring that the global environmental concerns of the developed Western nations would not overwhelm the desire to eradicate problems of human need and poverty. Unlike Brandt, Brundtland did offer the now globally famous definition of sustainable development (WCED, 1987: 43):

> Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.

This definition is still commonly used despite it attracting serious criticisms for perhaps, a term which some came to replace the Cold War notion of the West, inferring that economic growth, industrial modernization, and market imperatives should be key drivers and goals for all nations. Whereas the industrialized North seemed to be, and in many ways still is, concerned with environmental impacts the issues confronting the majority of the South included poverty, health, income, agricultural sustainability, food security, educational opportunity and achievement, shelter, sanitation, desertification, and armed conflict. For the South, economic development including growth was seen as essential and the more ecological concerns of the North were often perceived as being luxuries only the affluent nations could afford and countenance. For others, particularly within the academic world, sustainable development seemed to be either hopelessly vague or a basic contradiction in terms. Whatever the case, sustainable development entered the global policy lexicon and has remained there ever since.

The 1992 Rio Earth Summit

Five years later, in 1992, the UN Conference on Environment and Development, the follow-up to Stockholm, was held in Rio de Janeiro. This meeting, known as the Earth Summit, produced a number of very important agreements, including the Rio Declaration on Environment and Development, the Framework Convention on Climate Change, the Convention on Biological Diversity (CBD), a nonbinding Statement on Forest Principles, and the hugely cumbersome but nonetheless foundational agreement known as Agenda 21. The CBD and the negotiations before and after the 1997 Kyoto Protocol on climate mitigation are two very important examples of multilateral environmental agreements. Maintaining biological diversity is key to maintaining the planet's overall health and those ecosystem services that replenish natural resources and offering all creatures the dynamic equilibrium upon which life depends. Thus, if plant and animal species disappear, as they are doing at an unprecedented rate, then monocultures will emerge which are highly susceptible to disease, the effects of climate change, and other damaging ecological transformations. Industrialized systems of agricultural production have forced governments and the large corporations to recognize that these impacts must be mitigated and managed – biological diversity must be conserved, resources must be used more sustainably, and the material and cultural benefits derived from the planet's genetic resources shared more equitably. Business as usual is no longer an option. Following Rio, many national strategies have been based on these broad international agreements and statements of principle and intent. However, indigenous peoples and local communities have not always found their ideas, values, and needs accepted when the actual implementation processes are enacted but there has more recently been an acceptance that the developed world has a great deal to learn from the traditional ecological knowledge of peoples who retain a close affinity to the natural environment. However, global trade and commercial imperatives have led to rather weak attachments to sustainable development. Probably most depressing have been the limited, tortuous, and hesitant agreements and accommodations around the 1997 Kyoto Protocol – so far the only international, legally binding agreement on climate change.

The realities of climate change, global poverty, economic inequality, and water shortage also highlight the significance of gender in sustainable development. Although much NGO attention has focused inevitably on the appalling hardships many women experience, gender issues cannot be separated from wider social, cultural, or environmental concerns which has sometimes been the case. The Women's Environment and Development Organization (WEDO) has campaigned vigorously to combat the intergovernmental blindness to the gender implications of environmental policy and actions. Global climate change negotiations, including the Kyoto Protocol and the reports of the United Nations' Intergovernmental Panel on Climate Change, concentrate almost exclusively on reducing greenhouse gas emissions, largely ignoring the wider social and gender impacts. A United Nations Environment Programme (UNEP) (UNEP, 2006) survey, "Gender mainstreaming among environment ministries," discovered that just two countries involved in climate change activities had incorporated a gender perspective. However, as well as arguing that women often suffer disproportionately from unsustainable development, UNEP frequently promotes women as important agents for community empowerment, social leadership, and positive change. As the World Conservation Union has shown (IUCN, 2007), communities often cope more effectively during natural disasters when women play a leadership role in early warning systems and post-disaster reconstruction than when they do not. The IUCN also notes that women's local knowledge and skills offer tangible benefits such as Inuit, the women of Northern Canada, having a deep understanding of weather conditions because of their traditional responsibility for evaluating hunting conditions. When a drought occurs in the small islands of Micronesia, local women who have a sound knowledge of island hydrology find potable water by digging new wells. WEDO (2007: 3) adds that women tend to share information related to community well-being, choose less polluting energy sources, and adapt more easily to environmental changes when their families 'survival is at stake'.

The 40 chapters of Agenda 21 offered an action plan for sustainable development, integrating environmental with social and economic concerns, and articulating a participatory, community-based approach to a variety of issues, including population control, transparency, partnership working, equity and justice, and placing market principles within a regulatory framework. Local Agenda 21 (LA21), its local realization, was not legally binding, although by the end of 2000, many countries including the UK, had detailed policies and elaborate frameworks for sustainable development at local and regional levels. Municipal governments in many countries have taken a strong lead in global sustainability and today many of the world's 'global cities' have made considerable progress toward realizing sustainability goals. With an increasingly urban global population the sustainability of towns and cities can only increase in importance (*see* Urbanization; Development and Urbanization). In those, particularly Scandinavian countries where local government has a considerable degree of autonomy to raise income locally and regulate environmental matters, LA21 has been highly successful inspiring many communities to take action. However, throughout the world, even though local government priorities and powers may differ, global structures of economic, financial, and political power which include support for the neoliberal free trade system have compromised some attempts to fashion sustainable development from the bottom up. The local cannot be disassociated or disconnected from the global, conceptually or practically. Nonetheless, the LA21 process continued with, from 2002, Local Agenda 21 turning into Local Action 21. In 2004, the Aalborg Commitments was published, showing many local authorities within the European Union to be firmly embracing the need for urban sustainability and good governance. In addition, many grassroots movements such as those associated with Transition Towns have more recently made considerable progress with local municipalities becoming partners in rather than leaders of local actions such as local food growing, community enterprise and the creation of energy or carbon descent plans in our era of Peak Oil (*see* Mineral Resources, Limits to: The Case of Peak Oil).

Rio + 10, and + 20

Rio was, despite all the compromises and shortfalls, a significant achievement which over the years has gained in stature and authority. Paradoxically this enhanced status has grown because of the reluctance of the US to accept sustainable development policies, its frequent refusal to recognize the importance of the precautionary principle as a guide to environmental law, and the failure to be a party to global agreements on cutting greenhouse gas emissions. America's continuing support for neoliberal economic globalization has also had damaging consequences not least to the global economy following the financial crash of 2007–08 but also to the global environment more generally. However, the early years of President Obama's second administration has seen a shift in tone, policy, and action regarding the environment including particularly the increasingly evident impacts of climate change. Also, and again somewhat paradoxically, the fact that the Rio Declaration was seriously criticized by many radical green groups made its achievements all the more valuable and politically symbolic.

In 2002, the Johannesburg (Earth) Summit reviewed a decade of progress. The tensions apparent in 1992 remained obvious to all although the final declaration noted that global disparities in wealth and environmental degradation now risked becoming entrenched. Unless the world decides to act in a manner that fundamentally changes the lives of the poor, many citizens could lose confidence in democratic systems of government perceiving their representatives as little more than 'sounding brass or tinkling cymbals.' Indeed, general public disillusion and disaffection with established political institutions, ostensibly democratic procedures and processes, have characterized a great deal of the politics and protest occurring in the first decade or so of the twenty-first century at every spatial scale – the global, national, and local. At Johannesburg, little was said about financing international development, though in the same year, at an International Conference on Finance for Development in Monterrey, northeast Mexico, a consensus was reached on financing sustainable development, fostering health and education, providing shelter, eradicating poverty, and sustaining economic growth. The role of trade and overseas development aid, the importance of debt reduction and good governance in the developing world, and the mobilization of national economic resources and external investment were directly addressed although the perennial problems remained unresolved. Economic crises, the conference report stated, underscored the importance of the need for effective social safety nets.

The twenty-first century has witnessed progress, with regard to sustainable development, stall. In December 2009, a major climate conference was convened in Copenhagen to decide the successor to Kyoto but no legally binding treaty emerged from the tortuous negotiations that were frequently deadlocked (*see* Climate Change). China, India, and the US were each in turn blamed for the conference failure by politicians from other nations, international NGOs, and media commentators. However, delegates did agree that global warming should not exceed 2 °C but set no actual targets for cutting emissions. For some critics the real culprits were not the negotiators at Copenhagen but a general lack of political will and leadership in most countries. However, over the next few years talks continued at Cancun, Durban, Bangkok, Bonn, and in November–December 2012 in Doha, Qatar. A number of documents were produced at Doha, collectively known as The Doha Climate Gateway, which extended the Kyoto protocol to 2020 but limited the scope of global carbon emissions to 15% because Japan, Russia, Canada, and the US did not participate. Additionally, because China, India, and Brazil were still classified as developing nations at Kyoto, they have not been subject to agreed emission reduction targets. Climate campaigners and others reviewing these negotiations have not surprisingly expressed their exasperation and frustration. In 2011, Kevin Anderson and Alice Bows published a paper for the Royal Society in the UK warning that there was 'no chance' of keeping global warming below 2 °C and in any case recent studies relating to the impacts associated with such a rise have been revised upwards from 'dangerous' to 'extremely dangerous.' Consequently, they write (Anderson and Bows, 2011: 41), "with tentative signs of global emissions returning to their earlier levels of growth, [the year] 2010 represents a political tipping point." Energy expert Dieter Helm (2013) argues that massive investment needs and rigorous economic instruments are needed to combat climate change and businesses, governments, and citizens need to change their behavior radically. We are in a hole, he writes, but we can stop digging.

Although progress since the first Earth Summit in 1992 was carefully evaluated, commemorated, and celebrated at the Rio+20 Summit no new agreements or targets were issued in 2012 but there were plenty of 'reaffirmations' and 'recognitions' in the conference's final statement, The Future We Want (United Nations: 2012: para19; p.4). Indeed, the clear admission expressed in the clause, "we emphasize the need to make progress in implementing previous commitments," indicates both why and how so many delegates felt so deflated with the conference outcomes. Much of the debate was polarized around the meaning of the 'green economy' which for many seemed to coalesce around the desire for green energy technologies rather than defining the need for a new economic paradigm that favored social equity and quality of life above economic growth. UNEP, for example, clearly advocated and advocates a series of policy prescriptions characterized by the key principles of ecological modernization, the low carbon economy, and eco-efficiency. At a moment when the global economy was experiencing considerable stress as a result of the serial failures of finance capitalism, the Rio+20 vision for the future was hesitant, modest, and accommodative. Having said this, if progress was made at Rio in 2012 it was in realizing that no single assessment matrix for sustainable development had been previously devised and accepted. Consequently, an immediate task for the future became the fashioning of a set of sustainable development goals (SDGs) which in effect would supersede the Millennium Development Goals formulated at the turn of the century which largely focused on human health, poverty eradication, and general well-being. These new SDGs would be action orientated, concise, easy to communicate, limited in number, aspirational.

Sustainability as a 'Dialogue of Values'

There have been many academic critiques of the various discourses on sustainable development (Blewitt, 2014). For many observers the sustainable development has an equivocal relationship to the dominant economic growth paradigm and has also been informed by the residue of colonial thought processes which have disempowered many rural populations and the urban poor in the developing world. Although some critics acknowledge that the idea of sustainable development includes notions of plurality and dialogue, the commitment to and attraction of biotechnology, Western science, the commoditization of biodiversity via corporate patenting and intellectual property rights, has marginalized the many who still depend on their land for their livelihood. Indeed, a great deal of the discussion around green business still focuses on technological solutions and eco-efficiency – doing more with less – although there are signs that some very large business corporations are taking sustainability and social responsibility extremely seriously. Unilever is often cited as an example of a company in the process of changing its business model in a major way.

However, for conservation and sustainability academic Bill Adams (2001: 381) the fact remains that in the developing world, the global South, there is "no magic formula for sustainable development," no easy reformist solution to poverty and that, contrary to dominant practice, development "ought to be what human communities do to themselves" rather than what is done to them by states, bankers, experts, agencies, centralizing planners, and others. Instead, a 'green development' is required for which there can be no hard and fast blueprints because there is an overwhelming need to be open-ended, open minded, participatory, and overtly democratic. Green development is about who has the power and how it is managed. It is about empowerment and self-determination. For the eco-economist Ignacy Sachs (1999), there is a real need to emphasize the importance of social sustainability as a counterweight to the ideological and political dominance of the economic discourse on growth. Social sustainability introduces the need to address the recurring incidents of war and serious violence, state oppression of citizens, and a growing sense that there is a lack of meaning in today's world. Thus, Ignacy Sachs argues that development is akin to liberation and transformation if understood as a self-organizing and intentional process which actually frees people from poverty and exploitation. However Sachs, like the World Bank, recognizes that trade-offs will occur but some are totally, ethically, unacceptable. He writes (1999: 33),

> Concern for the environment should not become a diversion from the paramount imperatives of social justice and full democracy, the two basic values of whole development.

The Indian development economist and philosopher Amartya Sen (1999) has argued that realizing human capabilities in a sustainable society means equity, democracy, human and civil rights, and a continuing enhancement of people's ability to do what they have good reason to value. It also entails envisioning alternatives, acting and thinking differently, and most importantly having the capacity, freedom, and opportunity to do so. It will involve protecting biodiversity because human social and economic relationships are closely interwoven in a coevolutionary relationship with the biosphere. It means conceiving and practicing development holistically and systemically, not simply economically or socially, politically or anthropocentrically. What renders sustainability and sustainable development especially valuable is their resilient compatibility which affirm the ideas of freedom, of fulfillment, of being and securing what is truly valuable – the freedom to achieve, to effect solidarity with others, to develop capabilities and alternatives, and to live justly and meaningfully (Sen, 1999). Thus, social and environmental justice is essential to the satisfaction of human needs, of securing inter- and intra-generational equal opportunity, partnership, and cooperation. For Sen, this is the meaning of development within sustainable development.

For environmental scientist Richard B. Norgaard (1994), Western science, the environment, and material resources are connected within mutually interactive coevolving systems where one does not control any of the others. In nineteenth-century Europe, the application of scientific knowledge facilitated the use of coal and hydrocarbons which in turn directed and intensified scientific activity, agricultural, technological, and industrial development and the emergence of a new social, moral, and political order. Urbanization, class division, multinational business, global trade, and bureaucratic management systems helped concentrate economic and political power and the strategy of imposing this Western practice and ideology of development on non-Western others. Consequently, "correcting the unsustainability of development is not simply a matter of choosing different technologies for intervening in the environment. The mechanisms of perceiving, choosing, and using technologies are embedded in social structures which are themselves products of modern technologies" (Norgaard, 1994: 29). This coevolutionary approach to historical explanation offers tremendous insights but does not help us predict the future, for in this theory there can be no simple causal relationships. Prediction can therefore becomes either dangerous or ridiculous or both.

Given this, the concepts of and practices associated with sustainable development and sustainability therefore encompass multidimensional social, ecological, and economic goals and perspectives whose breadth and unscientific nature has led to many attempts over the decades to formulate precise sustainability measures, tools, indicators, and frameworks as well as the vaguer aspirational goals and intended inspirational aims articulated in various declarations, treaties, and political speeches. However, as sustainable development does not endorse 'calculative thinking' as a sole or perhaps even the most way of understanding what is the future we actually want and can achieve. Perhaps then, it is the sociologist Blake Ratner's notion of sustainability as a 'dialogue of values' which points to a way of effectively salvaging, understanding, and engaging with the various theories, values, perspectives, and assumed and actual contradictions of sustainable development. He identifies three basic tendencies in sustainable development practice, namely the technical, ethical, and dialogic. He writes (2004: 62),

> The sustainability concept is meaningful, therefore, not because it provides an encompassing solution to different notions of what is good, but for the way it brings such differences into a common field of dispute, dialogue, and potential agreement as the basis of collective action.

One of the main obstacles to developing a common conceptual framework or dialogue of values incorporating social, economic, and ecological concerns is the lack of genuine consensus among experts in each discipline concerning how ecological, economic, and social systems relate to one another. The resulting 'trifocal' vision makes understanding the world, international and national policy development, and effective action a major challenge. This is sometimes exacerbated by occasional tendencies for one group of experts to see their approach as being more fundamental than any other. When this happens, dialogue can become extremely difficult. Interdisciplinary and transdisciplinary approaches need to be discussed and enacted within and outside the academy. Political and practical strategies, policies and actions facilitating the emergence of conditions allowing for possible reconciliation between perspectives are essential. A recognition that systems thinking and concepts drawn from the study of nature and ecology can be applied, adapted, and developed in new contexts, such as education, economics, and politics, is an important way forward (*see also* Human Ecology; Social Ecology). Only if this is possible, can there be an effective understanding of, and engagement with, the overlapping and interconnected systems making up the biosphere, the economy, and human society. The intellectual basis for this engagement lies in the processes of dematerialization and resocialization of economic activity and by placing greater significance to well-being rather than materialistic criteria for assessing progress and development. To effect this democratically, there needs to be an authentic social dialogue on sustainability values and sustainable development policies. Without such dialogue political decision-making is likely to be blinkered, misinformed, or uninformed.

Ultimately, as Canadian academic John Robinson (2004) argues, sustainability is a political act, and what must inform this act is learning, understanding, and agreement. Meanings and practices will change as the world changes, as our skills, knowledge, and capabilities develop, and as communication and dialogue continues over time. From the neighborhood to global levels, different interests come together and sometimes collide but it is only through discussion, critical reflection, and dialogue that agreement and action on sustainability can come about. A step toward this is the remarkable document known as The Earth Charter which could only have been reached by people listening, talking, and learning from one another … and being willing to do so. Talk can become or lead to action for different countries have different levels of development, different values, cultures, and traditions. They are endowed with differing amounts of natural resources and so have, certainly according to Brundtland, differentiating responsibilities in promoting and realizing SDGs. This very heterogeneity has and can continue to allow a coming together, and an identification of common ground on which to build further action toward sustainability.

Conclusion

One of the main reasons why 'sustainable development' – the process – and 'sustainability' – the end in view or goal – have generated so much discussion is because they tend to reflect the political and philosophical value base of those involved in the conversation. John Robinson (2004: 370) prefers the term sustainability to sustainable development because, as he writes, it "focuses attention where it should be placed, on the ability of humans to continue to live within environmental constraints." For those who want an unambiguous scientific, technical, discipline-specific, and/or operationable definition, this causes problems unless a great deal is ignored or discounted. What is needed, then, and what the creative ambiguity surrounding 'sustainability' can offer, is the possibility of integration, synthesis, and synergy – of a social learning process that bridges the divisions between the social and ecological, the scientific and spiritual, the economic and the political. Thus, in practice, technical fixes are necessary but not sufficient if ecological, economic, and social imperatives are to be reconciled. For John Robinson, this cannot be done scientifically, only politically – in dialogue and in partnership, making sustainability "the emergent property of a conversation about what kind of world we collectively want to live in now and in the future." Finally, it is probably pragmatic, at the very least, to acknowledge Robinson's important insight (2004: 382) that within the field of sustainability multiple conflicting views do clearly exist and it may not always be possible to find ways of reconciling them.

> In other words, no single approach will, or indeed should be, seen as the correct one. (...). [W]e are inescapably involved in a world in which there exist multiple conflicting values, moral positions and belief systems that speak to the issue of sustainability. While it is crucial to identify points of empirical disagreement and to resolve those with better research and analysis, the ultimate questions are not susceptible to empirical confirmation or disconfirmation. What is needed, therefore, is a process by which these views can be expressed and evaluated, ultimately as a political act for any given community or jurisdiction.

In this way, 'sustainable development' and 'sustainability' may productively function as a heuristic, in other words a learning process by which people are enabled to find things out for themselves and to fully appreciate the contested nature of knowledge, the environment and sustainability, and the impact human actions have upon the Earth. Human life and the environment are not two separate entities: we are on and of the world. We must think and act wisely. We share the planet with other creatures and act accordingly in a just and equitable manner.

See also: Eco-cities and Sustainable Urbanism; Environmental Education and Education for Sustainable Development; Planning Issues and Sustainable Development; Sustainability and Sustainability Science; Sustainable Development and Social Work; Sustainable Development: An Economic Perspective; Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption.

Bibliography

Adams, W.M., 2001. Green Development: Environment and Sustainability in the Third World. Routledge, London.

Anderson, K., Bows, A., 2011. Beyond 'dangerous' climate change: emission scenarios for a new world. Philosophical Transactions of the Royal Society A 369, 20–44.

Blewitt, J., 2014. Understanding Sustainable Development, second ed. Routledge-Earthscan, London.

Brandt, W., 1980. North–South: A Programme for Survival. The Brandt Commission. Pan Books, London.

Carson, R., 2000, Original 1962. Silent Spring. Penguin Books, London.

Grober, U., 2012. Sustainability: A Cultural History. Green Books, Totnes UK.

Helm, D., 2013. The Carbon Crunch. Yale University Press, New Haven.

IUCN, 2007. Gender Aspects of Climate Change. IUCN. Gland.

Meadows, D.H., Meadows, D.L., Randers, J., Behrens III, W.W., 1972. The Limits to Growth: A Report to the Club of Rome. Universe Books, New York.

Meadows, D., Randers, J., 2005. Limits to Growth: The Thirty Year Update. Behrens III, W.W. Earthscan, London.

Norgaard, R.B., 1994. Development Betrayed: The End of Progress and a Coevolutionary Revisioning of the Future. Routledge, London.

Ratner, B.D., 2004. Sustainability as a dialogue of values: challenges to the sociology of development. Sociological Inquiry 74 (1), 59–69.

Robinson, J., 2004. Squaring the circle? Some thoughts on the idea of sustainable development. Ecological Economics 48, 369–384.

Roszak, T., 1972. Where the Wasteland Ends: Politics and Transcendence in Post-industrial Society. Doubleday, New York.

Sachs, I., 1999. Social sustainability and whole development: exploring the dimensions of sustainable development. In: Becker, E., Jahn, T. (Eds.), Sustainability and the Social Sciences. Zed Books, London.

Schumacher, E.F., 1973. Small is Beautiful. Abacus, London.

Sen, A., 1999. Development as Freedom. Oxford University Press, Oxford, UK.

United Nations UN, 2012. The Future We Want: Rio+20 Outcomes Document. United Nations, Rio de Janeiro. Available at: http://www.uncsd2012.org/thefuturewewant.html.

UNEP, 2006. Gender Mainstreaming Among Environmental Ministries. UNEP, Nairobi.

WCED (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, Oxford.

Sustainability and Sustainability Science

Robert W Kates, Trenton, ME, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sustainability, beyond its dictionary meaning, combines economic and human development with environmental maintenance, preservation, or improvement as sustainable development. It is characterized by various indicators, scenarios, goals, and targets. As a new subfield of science, it has an active solution-oriented research agenda, a theoretical perspective in coupled socio-environment systems, a decade of major publication, dedicated centers and institutes in many parts of the world, and a growing number of degree programs at every level.

Sustainability is everywhere, as in its dictionary use as the ability to be maintained at a certain rate or level, but also linked to economic and human development, applied to environmental maintenance, preservation, or improvement, or increasingly as a new subfield of science. How widespread its use has become is illustrated by Bettencourt's and Kaur's (2011) extended analysis. They assembled a database of scientific publications in English written between 1974 and 2010 that contained in their title, abstract, or key words 'sustainability' and/or 'sustainable development,' but eliminated items that simply used sustainable as in general usage. The data set contained about 20 000 papers, authored by about 37 000 authors found in 174 countries and in 2200 cities. The range of publications is quite varied as illustrated in the word cloud (Figure 1). Sustainable development, environmental sustainability, and sustainability science constitute the substance of this article.

Sustainable Development

The reconciliation of society's human developmental goals with the planet's environmental limits over the long term is the foundation of the concept known as sustainable development. The origins of sustainable development can be traced back through the 1980 World Conservation Strategy and the 1972 Stockholm Conference on the Human Environment to the early days of the international conservation movement (O'Riordan, 1988; Adams, 1990). But the contemporary linking of environment and development is now a quarter of a century old, stemming from *Our Common Future*, the 1987 report of the World Commission on Environment and Development (also known as the Brundtland Commission). The Commission defined sustainable development as the ability of humanity "to ensure that it meets the needs of the present without compromising the ability of future generations to meet their own needs" (WCED, 1987).

An extraordinarily diverse set of groups and institutions have taken this widely accepted statement of sustainable development and projected upon it their own hopes and aspirations. While sharing a common concern for the fate of the earth, proponents of sustainable development differ in what they think is to be developed, what is to be sustained, how each is related to the other, and for how long a time.

The most common view of what is to be developed is the economy that provides employment, needed income, desired consumption, and the means for investment in physical and human capital as well as environmental maintenance and restoration. An alternative emphasis is on people-centered human development which focuses on the quantity of life as seen in the survival of children and their increased life expectancy and the quality of life through

Figure 1 Word cloud for bigrams titles of sustainability papers. Bigrams (consecutive two-word combinations, obtained after frequent stop words are removed) often give a better sense of subjects addressed than single words. The frequency of *sustainable development* has been reduced by 35%, so that other terms would be more visible. Taken from Figure S3 in Bettencourt, L.M.A., Kaur, J., 2011. The evolution and structure of sustainability science. Proceedings of the National Academy of Sciences 108, 19540–19545 with permission of PNAS.

education and equal opportunity. Finally, some discussions of what is to be developed emphasize society itself, addressing the well-being and security of national states, regions, and institutions.

Environmental Sustainability

The most common view as to what is to be sustained are planetary life support systems, particularly those supporting human life. These systems include natural resources as well as environments that provide food, fiber, water, regulate atmosphere, climate, disease, pests, water, and culture as aesthetics, spirit, and recreation (Daily, 1997). The Millennium Ecosystem Assessment (MEA, 2005) found that these ecosystems have been changed more in the past 50 years than in any comparable period of human history, primarily to provide increased amounts of fiber, food, fuel, timber, and water. These changes have led to substantial net gains in economic development and human well-being, but not for all people and locations. The environmental systems that provide these services have been degraded or used unsustainably in many places and with possibilities of even more abrupt, rapid, or irreversible change that may well diminish the benefits to future generations (Turner et al., 1990).

Contrasting this anthropocentric view of nature as human life support is one of sustaining nature itself in all its biodiversity and assemblages of life forms. These ought to be sustained, not only for their utilitarian service to humans, but because of the moral obligations of humanity arising either as 'stewardship' – still acknowledging the primacy of humans – or as a form of 'natural rights' in which earth and its other living things have equal claims for existence and sustenance. Also for some, cultural diversity is seen as a counterpoint to biological diversity, with various communities that include distinctive cultures, particular groups of people, and specific places that need to be sustained.

Linkages between Development and Environmental Sustainability

The relationship between what is to be developed and what is to be sustained also differ. Some conceptual statements provide equal emphasis to both; others while paying homage to sustainable development, seem to be saying 'sustain only' or 'develop mostly' or 'sustain or develop subject to some minimal constraint on the other.' Similarly the implied time horizons for sustainable development differ. There is universal acceptance that such time horizons are intergenerational, but these then range between a single generation of 25 years, several generations, or as an unstated, but implicit, forever. Each of these time periods present very different prospects and challenges for sustainable development. Over the space of a single generation, almost any development appears sustainable, over an infinite forever none do.

In 2002, these relationships were expanded when the World Summit in Sustainable Development addressed growing concerns over the defining of development as economic development, while ignoring human development, and issues of equity and social justice. In its Johannesburg Declaration (2002) it asserted "a collective responsibility to advance and strengthen the interdependent and mutually reinforcing pillars of sustainable development – economic development, social development and environmental protection – at local, national, regional and global levels." These three pillars of sustainable development, i.e., economic, social, and environmental, are widely used, but there has been no agreement on the specific content of the three pillars, especially the social pillar.

Web searches can find three major variants of social development. There is a generic non-economic social designation described as 'social,' 'social development,' and 'social progress.' An alternative uses terms such as 'human development,' 'human well-being,' or just 'people' to address human development as opposed to economic development, while such terms as 'social justice,' 'equity,' and 'poverty alleviation' characterize concerns over justice and equity. Still the commitment to the three pillars persists and in the most recent of the world summits, Rio +20, while generally found disappointing (Halle, 2012), restated in its very first paragraph its declaration that we, "renew our commitment to sustainable development, and to ensure the promotion of economically, socially and environmentally sustainable future for our planet and for present and future generations."

Indicators and Scenarios

To add detail to sustainability, major efforts were made early to choose indicators of sustainable development in order to indicate directions of action or to measure progress of a sustainability transition (UN-CSD, 1996; Moldan et al., 1997; Meadows, 1998). These continue and if global, national, and local efforts are included, there are hundreds of efforts to identify appropriate indicators and to measure them. Typical of these efforts is the set of 40 economic, environmental, and social indicators for the US drawn from existing data sets by an intergovernmental committee for each of the three pillars of sustainable development (**Table 1**).

A review of global, regional, country, or local indicator efforts (Parris and Kates, 2003a) found that large numbers of indicators were considered and selected, with 40 or more indicators as in **Table 1** were quite common. But despite the intergenerational perspective of the standard definition of sustainability, the emphasis of most indicator studies is on the present or very short term. The large number of indicators seems to emerge from a participatory process where the ostensible need for indicators actually enables a broad discussion by very diverse sets of stakeholders as to what sustainable development should be and a negotiation over their differing views and aspirations. Thus to overcome the essential ambiguity of 'sustainable development' and to be inclusive of stakeholder views, the number of indicators becomes very large.

In contrast to indicators such as these that can inform society to what extent progress is being made in a transition toward sustainability, alternative pathways for a transition were explored through long-range development scenarios that reflect the uncertainty about how the global system might unfold. While most are rigorous, using the insights of science and modeling, these scenarios are told in the language of words as well as

Table 1 Sustainable development in the USA: an experimental set of indicators

Economic	*Environmental*	*Social*
Capital assets	Surface-water quality	Population of USA
Labor productivity	Ratio of renewable water supply to withdrawals	Life expectancy at birth
Domestic product	Fisheries utilization	Births to single mothers
Income distribution	Acres of major terrestrial ecosystems	Children living in families with only one parent present
Consumption expenditures per capita	Invasive alien species	Educational achievement rates
Unemployment	Conversion of cropland to other uses	Educational attainment by level
Inflation	Soil-erosion rates	Teacher training level and application of qualifications
Investment in R&D as a percentage of GDP	Timber growth to removals balance	People in census tracts with 40% or greater poverty
Federal debt to GDP ratio	Outdoor recreational activities	Crime rate
Energy consumption per capita and per $ of GDP	Contaminants in biota	Participation in the arts and recreation
Materials consumption per capita and per $ of GDP	Quantity of spent nuclear fuel	Contributing time and money to charities
Homeownership rates	Identification and management of superfund sites	
Percentage of households in problem housing	Metropolitan air quality nonattainment	
	Status of stratospheric ozone	
	Greenhouse gas emissions	
	Greenhouse climate response index	

numbers because assumptions about culture, values, lifestyles, and social institutions require qualitative description. Sets of scenarios were created that reflect current trends and reform proposals, various forms of social and environmental breakdown, and more fundamental transitions or transformations (Hammond, 1998; Raskin, 2002; Carpenter et al., 2006).

Goals and Targets

A transition to sustainability is also marked by specific and concrete goals that become targets for action that both define sustainable development and illustrates what it seeks to achieve (Parris and Kates, 2003b). For example, consider three sets of goals that use different time horizons. They are: (1) the short-term (2015) goals of the Millennial Declaration of the United Nations; (2) the two-generation goals (2050) of the Sustainability Transition of the NAS-BSD; and (3) the three generation (beyond 2075) goals of the Great Transition of the Global Scenario Group.

Millennial Declaration of the United Nations

In September 2000, the General Assembly of the United Nations adopted some 60 millennium goals that addressed peace, development, environment, human rights, the vulnerable, hungry, and poor, Africa, and the United Nations. Many of these contained specific targets to be reached by 2015 and for eight of the major goals (with 21 targets) progress is monitored by international agencies (United Nations, 2013) Three major goals address hunger and poverty, education, and gender equality. Three goals address health related to children, mothers, and specific diseases. Another is concerned with environmental sustainability, but only as drinking water and sanitation availability, and one on slum improvement. The last goal is to attain a global partnership for development including aid, debt, trade, and technology, with special attention to the least-developed, landlocked, and small island states.

As of 2012, 3 years before the 2015 achievement date, a number of the targets have been or will soon be achieved (United Nations, 2013). The first goal, the target of reducing extreme poverty by half has been reached 5 years ahead of the 2015 deadline, although the target of halving hunger will not be reached. Universal primary education will not be reached but major progress has been made especially in sub-Saharan Africa. For the third goal of gender equality, primary school enrollment of girls has already equaled that of boys. Among the health goals, there has been accelerated progress in reducing child but not maternal mortality, tuberculosis will be halved by 2015, and progress will be made on malaria and HIV-AIDS. For environmental sustainability, the target of halving the proportion of people who lack dependable access to improved sources of drinking water was met by 2010 and conditions for more than 200 million people living in slums have been ameliorated – double the 2020 target. Plans are underway for a new set of goals and targets beyond 2015.

Sustainability Transition of the NAS-BSD

From 1995 to 1999, the National Academy of Sciences-National Research Council's Board on Sustainable Development (1999) focused on a two-generation time horizon to address the minimal needs of a global population of 9 billion in which the world provides the energy, materials, and information to feed, nurture, house, educate, and employ the many more people of 2050, and reducing their hunger and poverty – while preserving the basic life support systems of the planet. They called this a sustainability transition.

Specific targets for these goals for human well-being and environmental preservation had been defined over the previous decades through extensive processes of international political debate and intergovernmental action. In the area of

human needs, internationally agreed-on targets existed for providing food and nutrition, nurturing children, finding shelter, and providing an education, although not for finding employment. Thus, there was an implicit hierarchy of needs that favored children and people in disasters, and favored feeding and nurturing first, followed by education, housing, and employment.

Compared with targets for meeting human needs, quantitative targets for preserving life-support systems were fewer, more modest, and more contested. Global targets existed for ozone-depleting substances and greenhouse gases, and regional targets existed for some air pollutants. Absolute prohibitions (zero targets) existed for ocean dumping of radioactive wastes and some toxics, for the taking and/or sale of a few large mammals (whales, elephants, and seals), migratory birds when breeding or endangered, and certain regional fishing stocks. Water, land resources, and ecosystems such as arid lands and forests had, at best, qualitative targets to achieve sustainable management or restoration. International standards existed for many toxic materials, organic pollutants, and heavy metals that threatened human health, but not for ecosystem health.

In choosing such targets, the Board anticipated the Millennium Declaration goals, with its own analysis of goals and targets. Less sanguine than the UN, the Board determined it would take a generation (25 years) to reach the 2015 goals of the Declaration and another generation to achieve the Declaration's goals to reduce hunger and poverty. The Board thought that this transition could be achieved without extraordinary or still unavailable technologies or radical transformations of human societies. Now 15 years later, based on the progress and lack of progress (reported above) of the Millennium Development Goals, development seems on track for the 2050 sustainability transition, but environmental goals, especially those dealing with the oceans and atmosphere, are lagging far behind.

Great Transition of the GSG

In exploring the sustainability transition, the Board on Sustainable Development (BRC-BSD, 1999) undertook with the assistance of the Global Scenario Group, an analysis of the proposed Sustainability Transition, using for illustration, specific goals on hunger and the greenhouse gas emissions. It found that a sustainability transition is clearly possible, without requiring fundamental changes in technology or society, but it is 'just' possible. To move from business as usual without changing the economic system and people's lifestyles or values is daunting.

That analysis was an application of four widely used scenarios of future sustainability. The four, developed by the Global Scenario Group (Raskin, 2002), included a business as usual 'Market Forces' scenario, a 'Fortress World' scenario of defended haves and remaining have-nots, a possible but daunting 'Policy Reform' scenario, and a 'Great Transition' scenario that sought to achieve for all humankind "a rich quality of life, strong human ties and a resonant connection to nature." Such a transition envisaged major system change in environment, economy, and society and in the lifestyles and values held by people and institutions.

With almost two decades of hindsight, further research, and continuing experience, how likely is a Great Transition? In terms of needed values, a review of eight multinational surveys (Kates et al., 2006) of attitudes and behavior related to the three Great Transition values of quality of life, human solidarity, and ecological sensibility was undertaken. For the values of solidarity and ecological sensibility, there is widespread attitudinal support, although these strong beliefs are not necessarily reflected in the actual behavior of individuals, countries, or societies. Thus for these two goals of a Great Transition, it is bridging the gap between attitudes and behavior gap that is needed, not profound value change.

But the somewhat ambiguous quality of life values, including the rejection of material consumption beyond what is needed for a 'good life,' are not evident today, nor can large system changes be seen in major trends. There is widespread support for economic growth. A third of humanity continues to struggle with extreme scarcity, while another third is just beginning to aspire to an orgy of consumption. Indeed the greatest challenge of the Great Transition might be to create the growth and material conditions, while preserving the life support systems, that could eventually allow for the emergence of values of fulfillment and not wealth as the definition of well-being.

Sustainability Science

Sustainability science, as described by the journal, *Proceedings of the National Academy of Sciences* (PNAS) is: "…an emerging field of research dealing with the interactions between natural and social systems, and with how those interactions affect the challenge of sustainability: meeting the needs of present and future generations while substantially reducing poverty and conserving the planet's life support systems." The origins of sustainability science lie both in the grassroots efforts of scientists around the world and in the elite institutional efforts beginning with the National Research Council NRC-US Board of Sustainable Development.

A measure of the grassroots effort, as mentioned above, is in the 20 000 English-language scientific publications, with 37 000 authors who wrote articles between 1974 and 2010 that contained in their title, abstract, or key words 'sustainability' and/or 'sustainable development.' (Bettencourt and Kaur, 2011). Beginning in the 1990s, the number of authors of articles grows rapidly, doubling about every 8 years, and evidence a distinctive pattern of author locations and disciplines (**Figure 2**).

In more traditional sciences, authorship is concentrated in Japan, the United States, and Western Europe, but sustainability science authorship is much more widely distributed and includes significant numbers of authors from developing countries such as Kenya and Nigeria and almost all the emerging economies (Brazil, Russia, India, China, and South Africa). The authors' institutions include corporate laboratories, government and nongovernmental organizations, as well as universities large and small. Many of the papers originate in political capitals, such as Beijing, Canberra, London, Tokyo, and Washington rather than in academic centers.

Sustainability Science is also extraordinarily multidisciplinary, spanning the natural, social, and technological sciences. Using the disciplinary classification of the journals,

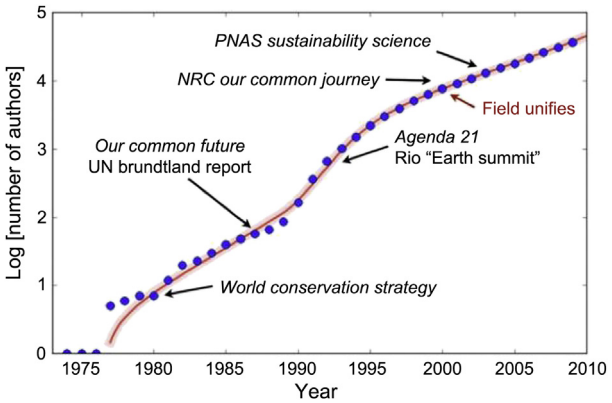

Figure 2 The graph illustrates the number of authors of English-language scientific publications, who wrote articles between 1974 and 2010 that contained in their title, abstract, or key words 'sustainability' and/or 'sustainable development.' Five key institutional and publication initiatives and field unifications are also shown by their approximate dates. Taken from Bettencourt, L.M.A., Kaur, J., 2011. The evolution and structure of sustainability science. Proceedings of the National Academy of Sciences 108, 19540–19545 with permission of PNAS.

a third of the papers were published in social science journals, a quarter in biological journals, and a fifth in engineering journals. In the judgment of Bettencourt and Kaur, sustainability science unifies around the year 2000, with most scholars and places connected with network links of co-authorship.

The elite institutional development of sustainability science is partly shown in Figure 2 as key institutional initiatives. Sustainable development emerged in the early 1980s from scientific perspectives on the interdependence of society and environment (ICSU, 1980). It entered the high political agenda through the publication by the Brundtland-led World Commission on Environment and Development (1983–87) report: *Our Common Future* and the subsequent United Nations Conference on Environment and Development held in Rio de Janeiro in 1992.

While a great political success, many scientists found it difficult both to conceptualize and measure sustainable development and make it a subject of scientific inquiry. To assist, the US National Academy of Sciences created a Board on Sustainable Development that began a 5-year effort to better connect science and technology to sustainable development including creating a sustainability science (NRC-BSD, 1999). Subsequent meetings of the World Academies of Sciences, the Initiative for Science and Technology for Sustainability (ISTS), the International Council for Science (ICSU), and the Third World Academy of Sciences in Japan, Sweden, Nigeria, Thailand, Germany, Canada, Chile, and Mexico embraced the notion of sustainability science (Clark et al., 2002), identified its core questions and methodologies, and launched sustainability science around the world (Kates et al., 2001). After the field unified around 2000, the PNAS, a major journal, started a section on sustainability science as one of its many disciplinary sections. By the end of 2012, 378 articles had been published in this unique section on sustainability science.

Both the grassroots and elite accounts find that sustainability science is a different kind of science that is primarily use-inspired as are agricultural and health sciences, with significant fundamental and applied knowledge components and commitment to moving such knowledge into societal action. From the grassroots effort it was clear that sustainability science was already in the air and in print prior to Board of Sustainable Development, but the elite international and regional scientific institutions saw that it needed theoretical and methodological framing beyond the topics of individual research. One such framing was the focus on socioenvironment interactions (also nature-society, socioecological, human environment interactions) and the search for core questions or research themes beyond the local or sectoral problems considered in most of the sustainability science literature. Major attempts to put together core questions (Kates et al., 2001) and thematic research (Levin and Clark, 2010) for sustainability science were made in 2000 and 2011. Combining these core questions and thematic research, there are seven major questions for research.

1. What shapes the long-term trends and transitions that provide the major directions for this century?
2. What determines the adaptability, vulnerability, resilience of human–environment systems?
3. How can theory and models be formulated that better account for the variation in human–environment interactions?
4. What are the principal tradeoffs between human well-being and the natural environment?
5. Can scientifically meaningful 'limits' be defined that would provide effective warning for human–environment systems?
6. How can society most effectively guide or manage human environment systems toward a sustainability transition?
7. How can the sustainability of alternative pathways of environment and development be evaluated?

Most recently, several proposed research frameworks (e.g., Dasgupta and Duraiappah, 2012; Arrow et al., 2012) have sought to make the Brundtland definition researchable and measurable by defining sustainable development as the well-being of people over the long run. Thus, society's development is sustained over a period of time if during that period intergenerational well-being does not decline. In Figure 3, human well-being (W) is an aggregate of various constituents such as material needs, health, education, opportunity, community, and security. These are a function of the goods and services that provide these constituents when consumed by people and in turn these are produced from a productive base composed of the assets: natural capital (Cn), human capital (Ch), manufactured capital (Cm), institutions and social capital (I), and knowledge (K). This productive base can then be considered to be the ultimate determinants of well-being, a relationship succinctly expressed as follows: W = +(Cn, Ch. Cm, I, K). This formulation suggests that assets of the productive base should have a worth or value to society. It has been shown that under an interesting range of conditions, this social worth of the productive base, aggregated as indices of 'inclusive wealth,' tracks the intergenerational well-being that can flow from it and thus serve as a conceptually grounded metric of sustainable development (e.g., UNU-IHDP and UNEP (2012); World Bank (2010)).

Sustainability Science is now at most two decades old, whether dated by the accelerated publication of the 1990s, the proposal by the National Academy of Sciences-Board on Sustainable Development in 1999, or as explicated in *Science* in

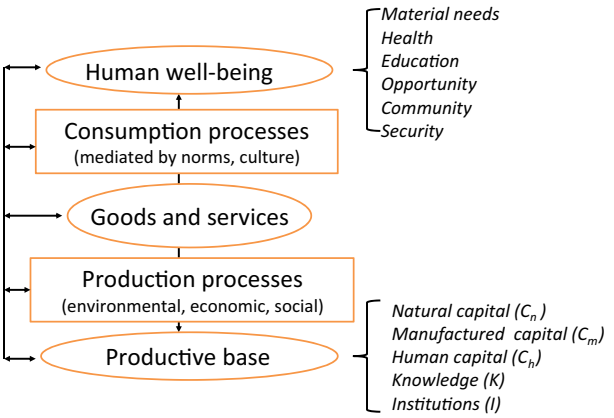

Figure 3 A research framework for sustainable development defined as long-run human well-being, i an aggregate of various constituents such as material needs, health, education, opportunity. community and security. These are a function of the goods and services that provide these constituents when consumed by people and in turn these are produced from a productive base composed of the assets: natural capital (Cn), human capital (Ch), manufactured capital (Cm), institutions and social capital (I), and knowledge (K) (drawn from Clark et al., forthcoming, with permission of the authors).

2001. Today the field has achieved an active solution-oriented research agenda, a theoretical perspective in coupled socio-environment systems, a large published library of new work (Kates, 2010), dedicated centers and institutes in many parts of the world, and a growing number of degree programs at every level. While sustainability science has become rooted in the scientific enterprise, it is being pursued somewhat differently in the US, the EU, Japan, and S. Africa. Its research questions are now better defined, but of the core questions and themes major, progress lags in half of them. Numerous degree programs have been created in colleges and universities, but common sets of needed competencies for students have not yet emerged. There are many examples of joint undertakings with practitioners from local communities, industry, government, and civil society, but fewer examples of solutions offered or problems solved. Thus the growing and exciting large body of research may be only slowly yielding solutions for the great environment and development challenges of this century.

See also: Eco-cities and Sustainable Urbanism; Environmental Education and Education for Sustainable Development; Sustainability: A Dialog of Values; Sustainable Agriculture; Sustainable Development and Social Work; Sustainable Development: An Economic Perspective; Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption; Water Resources and Sustainable Water Management.

Bibliography

Adams, W.M., 1990. Green Development: Environment and Sustainability in the Third World. Routledge, London.

Arrow, K.J., Dasgupta, P., Goulder, L.H., Mumford, K.J., Oleson, K., 2012. Sustainability and the measurement of wealth. Environment and Development Economics 17, 317–353.

Bettencourt, L.M.A., Kaur, J., 2011. The evolution and structure of sustainability science. Proceedings of the National Academy of Sciences 108, 19540–19545.

Carpenter, S.R., Bennett, E.M., Peterson, G.D., 2006. Scenarios for ecosystem services: an overview. Ecology and Society 11 (1), 29.

Clark, W.C., et al., May 20–23, 2002. Science and Technology for Sustainable Development: Consensus Report of the Mexico City Synthesis Workshop, 2002. Initiative on Science and Technology for Sustainability, Cambridge, MA.

Dasgupta, P.S., Duraiappah, A.K., 2012. Well-being and wealth. In: United Nations University International Human Dimensions Programme on Global Environmenatal Change. Cambridge University Press, Cambridge, UK, pp. 13–26. Inclusive Wealth Report 2012: Measuring Progress toward Sustainable Development.

Daily, G.C. (Ed.), 1997. Nature's Services. Societal Dependence on Natural Ecosystems. Island Press, Washington, DC.

Halle, M., 2012. Life after Rio. The International Institute for Sustainable Development. http://www.iisd.org/pdf/2012/com_life_after_rio.pdf (accessed 21.06.13.).

Hammond, A., 1998. Which World? Scenarios for the 21st Century. Island Press, Washington, DC.

International Union for the Conservation of Nature and Natural Resources (IUCN), 1980. World Conservation Strategy: Living Resource Conservation for Sustainable Development. IUCN, Gland, Switzerland.

The Johannesburg Declaration on Sustainable Development, 2002. http://www.un-documents.net/jburgdec.htm (accessed 21.06.13.).

Kates, R.W., et al., 2001. Sustainability science. Science 292, 641–642.

Kates, R., Leiserowitz, A.A., Parris, T.M., 2006. Great transition values: present attitudes, future changes. In: Kriegman, Orion, Raskin, Paul (Eds.), Frontiers of a Great Transition: GTI paper series, 9. Tellus Institute, Boston, MA.

Kates, R.W. (Ed.), 2010. Readings in Sustainability Science and Technology. CID Working Paper 213 (Center for International Development. Harvard University, Cambridge, MA.

Levin, S.A., Clark, W.C. (Eds.), 2010. Report from Toward a Science of Sustainability Conference. Center for Biocomplexity, Princeton University, Princeton, NJ. Sustainability Science Program, Harvard University, Cambridge, MA.

Meadows, D., 1998. Indicators and Information Systems for Sustainable Development. A Report to the Balaton Group. The Sustainability Institute, Hartland Four Corners, VT.

Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-Being. Island Press, Washington.

Moldan, B., Billharz, S., Matravers, R. (Eds.), 1997. Sustainability Indicators: A Report on the Project on Indicators of Sustainable Development. Wiley, New York and Chichester, UK.

National Research Council, Board on Sustainable Development, 1999. Our Common Journey: A Transition toward Sustainability. National Academy Press, Washington, DC.

O'Riordan, T., 1988. The politics of sustainability. In: Turner, R.K. (Ed.), Sustainable Environmental Management: Principles and Practice Boulder. Westview Press, Boulder, CO.

Parris, T.M., Kates, R.W., 2003a. Characterizing and measuring sustainable development. Annual Review of Environment and Resources 28, 559–586.

Parris, T.M., Kates, R.W., 2003b. Characterizing a sustainability transition: goals, targets, trends, and driving forces. Proceedings of the National Academy of Sciences 100 (14), 8068–8073.

Raskin, P., 2002. Great Transition: The Promise and Lure of the Times Ahead. Stockholm Environment Institute, Boston.

Turner II, B., Clark, W., Kates, R., Richards, J., Mathews, J., Meyer, W., 1990. The Earth as Transformed by Human Action: Global and Regional Changes in the Biosphere over the Past 300 Years. Cambridge University Press, Cambridge, UK.

United Nations, Commission on Sustainable Development (UN-CSD), 1996. Indicators of Sustainable Development: Framework and Methodologies. United Nations, New York. http://www.un.org/esa/sustdev/csd/csd9_indi_bp3.pdf (accessed 21.06.13.).

United Nations, 2013. Millennium Development Goals Report, 2012. http://www.un.org/millenniumgoals//reports.shtml (accessed 21.06.13.).

UNU-IHDP and UNEP (United Nations University, International Human Dimensions Programme on Global Environmental Change and United Nations Environment Programme), 2012. Inclusive Wealth Report 2012: Measuring Progress toward Sustainability. Cambridge University Press., Cambridge, UK.

World Bank, 2010. The Changing Wealth of Nations: Measuring Sustainable Development in the New Millennium. World Bank, Washington, DC.

World Commission on Environment and Development, 1987. Our Common Future. Oxford University Press, Oxford, UK.

Sustainable Agriculture

Alison Heslin, Emory University, Atlanta, GA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Barkin, volume 6, pp. 3597–3599, © 2001, Elsevier Ltd.

Abstract

Revolutions in agricultural technology and policy throughout the twentieth century have dramatically changed farming practices. While these changes created necessary, short-term increases in food production, they also created environmental and social consequences that have called this agricultural system into question. Beginning in the 1960s, the environmental movement drew attention to the impacts of industrialized, chemical-intensive agricultural practices on the health of individuals and the planet. In addition, the consequences of agribusinesses on farmers and consumers throughout the world have sparked reactions against the injustices of a corporate, capitalist food system. These negative consequences of modern agriculture have shifted the discourse of agriculture and food security away from one of strictly growing more food to one of growing food in a way that is environmentally, socially, and economically sustainable.

Introduction

Throughout the process of industrialization and development, agriculture has been profoundly altered, disconnected from nature, and fit into the capitalist system of production (McMichael, 2000). Changes in food production resulted from technological, demographic, political, and economic changes, which increased yields, supported growing urban populations, affected domestic and international trade and trade laws, and grew lucrative, powerful agricultural businesses. Toward the end of the twentieth century, however, this system of agriculture was called into question as people perceived the negative environmental consequences, as well as the shortfalls of the system in meeting current and future demands.

By 2050, the projected increase in world population combined with increased meat consumption will require the world to produce twice as much grain as at the start of the twenty-first century (Tilman et al., 2002). This increased demand requires not only a sheer increase in production, but also a departure from the current system of agriculture, which has proved socially and environmentally unsustainable. The current methods of agricultural land use are "trading short-term increases in food production for long-term losses in ecosystem services, including many that are important to agriculture" (Foley et al., 2005, pp. 570–571).

The call for a more sustainable agricultural system came as a reaction to the numerous negative externalities associated with modern, industrialized agriculture. While 'sustainable agriculture' and 'sustainability' more generally are commonly used, they are quite recent in origin and variable in definition. Entering mainstream discourse after the 1987 UN World Commission on Environment and Development report, sustainable development, in general, was defined as "development that meets the needs of today without compromising ability of future generations to meet their own needs" (WCED, 1987). Sustainability has since been used to describe many different facets of development, including agriculture.

Although numerous definitions of sustainable agriculture exist, most agree on three basic, overlapping components: ecological, economic, and social sustainability (Jordan and Constance, 2008; Kaur, 2013; Pilgeram, 2013; Thompson, 2010). Environmentally, sustainable agriculture requires that the process of food production does not deplete natural resources faster or to a greater extent than they can replenish themselves. Second, to be sustainable within a capitalist economy, the food system must be profitable in the long-term for those involved in growing and consuming. Last, the facet of sustainability which receives comparatively little attention stresses the quality of life of farm workers and rural communities as well as the health and safety of consumers. These three dimensions of sustainability must each be addressed to meet the growing demand without compromising future generations' food access.

Agriculture in the Twentieth Century

During the twentieth century, agriculture has undergone changes that have dramatically altered the world's environment, landscape, and diet. Beginning with the shift from animal to machine plowing, and then from natural to chemical fertilizers and pesticides, and most recently with the advent of genetically modified seeds, revolutions in agricultural practices have created a food system disjointed from nature (Jordan and Constance, 2008; Kenney et al., 1991). While these shifts in agriculture accomplished a very necessary increase in yield to meet the demands of an increasing world population, they did so at the expense of the long-term capacity of our land, water, air, and traditional farming knowledge. In place of the formerly intimate connection between food and land, the ecosystem became something to alter and control to produce food as an "input in urban diets and industrial processing plants" (McMichael, 2000, p. 21).

Environmental

With 40% of the world's land area devoted to cropland and pastures, the modern system of intensive agriculture is one of the largest contributors to greenhouse gas emissions and is responsible for extensively altering the world's ecosystems (Foley et al., 2005). Modern farming often uses the practice of monocropping, alters the landscape, decreases biodiversity,

and relies on chemical fertilizers and pesticides that damage soil fertility, disrupt local ecosystems, and pollute runoff and ground water (Kaur, 2013). Animal grazing, which utilizes 25% of the world's land area, contributes to desertification, woody encroachment, and deforestation (Asner et al., 2004) as well as producing greenhouse gases. In addition to such land-related impacts of agriculture, crop irrigation accounts for 70% of fresh water extracted from underground (Thompson, 2010). These processes can in turn negatively affect future agriculture and population health by "diminishing the capacity of global ecosystems to sustain food production, maintain fresh water and forest resources, regulate climate and air quality, and mediate infectious diseases" (Foley et al., 2005, p. 572).

Due to the sheer scale of agriculture, its impact on the environment is crucial to consider. In addition, the demand for food is only increasing as the world population grows and diets shift to more resource-intensive foods. This increased need requires that the productivity of the land actually increases over time, and not merely remains constant. The focus on strictly growing more food in any way possible to combat hunger has lessened the productive capacity of our land and depleted or adulterated many water sources. To ensure food security for the growing population, the mind-set is shifting to prioritize for the way such increased production is accomplished and the long-term consequences of these practices (Lang and Barling, 2012).

Economic

In addition to the environmentally unsustainable aspects of modern agriculture, the technology-intensive system has created immense profits for large agriculture companies at the expense of farmers and consumers around the world. The initial infrastructure and technology for manufacturing chemical agricultural inputs in the United States came from chemical companies that had been manufacturing explosives during World War II. Following the War, these factories shifted to producing chemical agricultural inputs. A new form of agriculture and associated products were then exported throughout the world. Throughout the Cold War, the export of such technology, and the model of high-yield capitalist agriculture, to the developing world was encouraged and supported politically as an attempt to prevent more countries from falling into communism's orbit (Pilgeram, 2013).

As the twentieth century progressed, this industry bred high profit and powerful multinational corporations. These corporations, often termed "agribusinesses," are characterized by "large, industrial-type operations along the commercial food chain" (Walsh, 1975, p. 531). Agribusinesses have gone through numerous mergers, resulting in single companies such as Monsanto, which controls seed and chemical production and owns patents in the corn, cotton, and wheat sectors in multiple countries around the world. Between 1995 and 1998 alone, Monsanto spent over $8 billion buying seed companies (Shiva, 2000). The increase in size and scope of these corporations has increased corporate influence over international trade and intellectual property law, environmental policy, agricultural regulations, the chemical market, the media, and so on (Tokar, 2003). Such influence has disempowered farmers who are now dependent upon multinational corporations for expensive seeds and chemical inputs rather than on traditions of seed saving and natural forms of fertilizers and pesticides with significantly lower input costs. In addition, those purchasing food are subject to the potential risks associated with consuming highly processed products from chemical-intensive agricultural sources.

So, while economically advantageous for the companies themselves, the health of the environment, the quality of life of farmers, and the safety of consumers are all compromised on behalf of such profits. These profits are reaped strictly through the process by which the "world price for agricultural commodities strikingly divorced from cost" (McMichael, 2005, p. 271). The costs of producing and trading agricultural commodities are absorbed throughout the process by states, laborers, consumers, and unaccounted environmental degradation.

Social

As introduced above, the increase in necessary inputs has placed farmers in an economically difficult and disempowered situation; given the size and concentration of input corporations, farmers completely lack control over the price of inputs and, often, the selling price of crops. Farmers have become more vulnerable to price fluctuations as the world agricultural market has become increasingly integrated and the social structure of many countries has changed. Due to increased trade, many farmers shifted from subsistence farming or farming crops that could not compete with imports to producing crops they could sell on the international market. Shifting from diverse, subsistence farming to cash crop production encourages efficiency and specialization, which is often achieved through increasing farm size (i.e., economies of scale). This increase in farm size throughout the world consolidated land previously farmed by multiple small farmers, transforming subsistence and small, for-market farm owners into agricultural wage laborers. Unlike subsistence farmers and agricultural small holders, who frequently engage in a degree of subsistence farming, wage laborers must purchase all of their food, leaving them vulnerable to price fluctuations. In addition, many rural dwellers throughout the world, through either the push of farm consolidation or the pull of industrial development in cities, moved to urban areas and became entirely dependent on wages to acquire food. The increase of trade, combined with a decrease in farmers, has created a growing population of consumers dependent on the international food market for their food security, often with negative consequences on their food security (Walton and Seddon, 1994).

In addition to these products and technologies that augment the natural biological growing cycle, more recent innovations in agricultural input attempt to fit agricultural production into the capitalist, highly industrial production mold. From a capitalist standpoint, "the problem with agriculture, then is how to make nature step to the tune of the capitalist clock, that is how to revolutionize the biological processes themselves, not just how to use machinery or fertilizer to augment natural processes" (MacLennon and Walker, 1980, p. 28). These attempts to fit farming into an industrial

model, when successful, have contributed to a loss of traditional knowledge about sustainable farming practices that have existed for the history of agriculture (Barndt, 1997).

Origins

This move toward unsustainable development practices, especially in the realm of food production and the environment, gained popular attention in the 1960s. Rachel Carson's book *Silent Spring*, published in 1962, is often cited as the beginning of the modern environmental movement in the United States. This book raised awareness regarding the negative side-effects of chemical pesticides on the environment, particularly the impact of DDT on bird populations. In the years since *Silent Spring*, environmental concern has grown into a global movement, embedded in domestic and international policies, governmental organizations, and NGOs. The term 'sustainable development' and subsequent sustainability agendas followed in 1987, when the UN explicitly addressed the issue of sustainability in its report 'Our Common Future.' Agriculture became a major component of these environmental and sustainability movements, as critics argued that despite the gains in output, "when all its hidden costs are considered, modern industrial agriculture is not the bargain it appears to be" (Beus and Dunlap, 1990, p. 591). With the growing awareness of the environmental consequences of agriculture, the perspective on global food security also shifted. Initially, the approach to fighting world hunger was to grow more food by whichever means possible, but by the first decade of the twenty-first century, the influence of the sustainability discourse and new challenges to food security broadened and reshaped critiques of the food and agricultural systems (Lang and Barling, 2012). The current criticisms and modifications of the modern agricultural system stress sustainability across the different dimensions of agriculture and food access including environmental, economic, and social considerations.

As the sustainable agriculture paradigm diverged from that of conventional agriculture, some key differences emerged in critiques and prescriptions. According to Beus and Dunlap (1990), these key differences between conventional and alternative agriculture fall, respectively, into six major categories: centralization/decentralization, dependence/independence, competition/community, domination/harmony with nature, specialization/diversity, and exploitation/restraint. Within each of these categories, alternative or sustainable agricultural practice stresses smaller-scale, resource-conserving practices that value local communities.

What is Sustainable Agriculture?

Various parties engaged in sustainable agriculture define it in different ways, ranging from slight modifications needed within the existing structure, to drastic overhaul of power relations that currently govern food production. At its broadest, following the UN definition, sustainable agriculture is a method for growing food that does not compromise the future of food production. Most authors extend this definition, similarly to Beus and Dunlap, by defining sustainable agriculture in opposition to conventional agricultural practices. Carried further, sustainable agriculture is sometimes viewed as a complete rejection of locating agriculture with the corporate, capitalist system. Despite the range of possible definitions of sustainable agriculture, most facets of the process fall into one of three major categories: environmental/ecological, economic, and social sustainability.

Ecological

Those components that come to mind when one thinks of sustainability are likely related to the health of ecosystems and the environment. Impact studies and enacted sustainable programs and practices address different aspects of environmental consequences depending on the geographic location and the type of food production, from organic fertilizers to nonpolluting fish farming, the scope of ecological changes is enormous. Generally, the mind-set stresses organic practices over blind faith in technology to mimic traditional farming methods and natural ecosystem processes as much as possible. To accomplish this, sustainable agriculture proponents have called for reduced dependence on external, nonrenewable energy sources, recommending a shift to renewable and local resources for energy and farm inputs. This theme of decreasing the scale of dependence as much as possible occurs often in sustainability discourse, beginning at the level of the farm. Creating closed loop farming requires practices such as crop rotation, composting, combined livestock and crop production, diversified plant types and species, water collection systems, and sun or wind-power harvesting. For those inputs which a farm is not able to self-produce, local production is prioritized, favoring small-scale community-based production to dependence on centralized corporations. This stress on local production extends to sale and consumption. In contrast to the existing system, which "produces food for long storage and transportation times using vertically integrated processes of production and distribution, increasingly large corporate farms, and heavy use of pesticide and other chemicals" (Gasteyer et al., 2008, p. 48), local consumption requires less transport time, and therefore a decreased need for preservation in the form of chemicals or climate control.

Economic

The economic aspects of sustainable agriculture vary in purpose and form more than those devoted to protecting the environment, as practices that are economically advantageous for corporations may not benefit consumers, those beneficial for farmers may hurt corporate profits, and so on. Within discourse on sustainable agriculture, however, certain themes emerge regarding the economic direction in which agriculture should move. In general, sustainable agriculture entails a move toward smaller-scale, less capital- and technology-intensive production. In this production, farmers and other food-system actors are called to consider the external costs and long-term economic outcomes of present-day actions rather than strictly immediate monetary costs and short-term profits. In addition to farm specific economic concerns, much sustainability discourse encourages a move away from materialism more generally to valuing the

nonmonetary dimensions of nature. Whereas some economic analyses of sustainability attempt to monitor long-term, externalized costs to the environment by calculating the future profit losses or monetizing the negative impacts on the environment, others instead encourage valuing nature and the environment for its own sake, rather than fitting it into the man-made capitalist framework of costs and profits.

Social

The final category of agricultural sustainability addresses issues of social and economic equity in the current food system. This dimension of sustainability questions whether food can be deemed sustainable if, while not endangering the environment or future farm solvency, still creates or depends upon inequality and exploitation for production. Thus, models of sustainable agriculture feature decentralized control and value local communities. Decentralization of agricultural production means more farms and farmers, as well as local and regional processing and sales (Beus and Dunlap, 1990). This decentralized structure reduces the capacity of the food system to hide the costs "in the poverty of migrant and third world agricultural workers" (Cone and Myhre, 2000, p. 188). Through buying food directly from farmers and abstaining from purchasing food grown or processed by distant, transnational corporations, consumers become more informed regarding the food production process, often purchasing food from farms or farmers' markets directly. This direct interaction between farmer and consumer removes some potential for hidden short-term costs and exploitation of farmers, by allowing the consumer to interact with the farmer and have access to the farm where food is grown.

A criticism sometimes waged against the sustainable agriculture movement, however, is its primary market of privileged consumers (Pilgeram, 2011). If the concept of sustainable agriculture is used to represent an ideal or a crucial aspect of development, then one cannot solely focus on the effects on the environment and on wealthy, first-world consumers. Such critics argue that to be properly sustainable, the food system must ensure access to nutritious, affordable food for the world's population, and not only for a privileged few.

Sustainable Agriculture in Practice

One form of sustainable agriculture that has gained momentum in the United States is Community Supported Agriculture (CSA). CSA exemplifies the ways environmental, economic, and social aspects of sustainable agriculture can be carried out. Ecologically, CSAs are organic, ecologically diverse farms often located near urban consumers. To provide a range of food products to members of the CSA, farmers must grow many types of food on the same farm, allowing for easier crop rotation practices than on specialized farms. Economically and socially, this system seeks to restructure food production by operating through collective ownership and planned production. In CSAs, members purchase 'shares' at the beginning of a season, which provides initial, necessary capital for farmers to buy production inputs for the growing season. In addition to this preliminary capital influx, farmers also have a guaranteed income for the season regardless of market fluctuations or crop failure. Producers and consumers share risks associated with farming and receive weekly shares of food according to what and how much was harvested that week. Members of CSAs believe agriculture is "unable to survive in an economy dominated by the market, in which prices radically fluctuate, the cost of land escalates, and resources—natural and human—are exploited" (Cone and Myhre, 2000, p. 188). Unlike other forms of supporting local agriculture, such as farmers' markets, this form of local, sustainable agriculture movements seeks not only to support environmentally sustainable forms of agricultural production, but also to radically restructure the way people view and participate in the economic structure of food production and consumption.

Future of Food

While sustainable agriculture has made substantial gains over the past few decades in terms of awareness, research, policy, and producer and consumer behaviors, much more is needed in all components before we have a just and sustainable global food system. The health of ecosystems, society, and individuals is highly influenced by our food production process. Ensuring safe, long-term food access for the world's growing population is a defining challenge of our time, and whether we can effectively address it will dictate future development of our society.

See also: Agrarian Political Economy; Agriculture, Economics of; Alternative Food Movements; Climate Change and Global Poverty; Development and Urbanization; Environment and Development; Environmental Movements; Environmental Sociology; Food Deserts in US Cities; Food Security and 'Green Revolution'; Food, Sociology of; Land Change: The Merger of Land Cover and Land use Dynamics; Near Middle East/North Africa Studies: Society and History; Slow Food as a Social Movement; Sustainability and Sustainability Science; Sustainable Development: An Economic Perspective; Sustainable Development.

Bibliography

Asner, G.P., et al., 2004. Grazing systems, ecosystem responses, and global change. Annual Review of Environment and Resources 29 (1), 261–299.

Barndt, D., 1997. Bio/cultural diversity and equity in post-NAFTA Mexico. In: Drydyk, J., Penz, P. (Eds.), Global Justice, Global Democracy. Fernwood Publishing, Winnipeg/Halifax, pp. 55–59.

Beus, C.E., Dunlap, R.E., 1990. Conventional versus alternative agriculture: the paradigmatic roots of the debate. Rural Sociology 55 (4), 590–616.

Carson, R., 1962. Silent Spring, first ed.

Cone, C., Myhre, A., 2000. Community-supported agriculture: a sustainable alternative to industrial agriculture? Human Organization 59 (2), 187–197.

Foley, J.A., et al., 2005. Global consequences of land use. Science 309 (5734), 570–574.

Gasteyer, S., et al., 2008. Produce sections, town squares, and farm stands; comparing local food systems in community context. Southern Rural Sociology: SRS 23 (1), 47–71.

Jordan, J., Constance, D., 2008. Sustainable agriculture and the social sciences: getting beyond best management practices and into food systems. Southern Rural Sociology 23 (1), 1–22.
Kaur, G., 2013. Sustainable development in agriculture and green farming in India. OIDA International Journal of Sustainable Development 6 (12), 61–64.
Kenney, M., et al., 1991. Agriculture in U.S. Fordism: the integration of the productive consumer. In: Friedland, W., et al. (Eds.), Towards a New Political Economy of Agriculture. Westview Press, Boulder, CO, pp. 173–188.
Lang, T., Barling, D., 2012. Food security and food sustainability: reformulating the debate. Geographical Journal 178 (4), 313–326.
MacLennon, C., Walker, R., 1980. Crisis and change in U.S. agriculture: an overview. In: Agribusiness in the Americas. Monthly Review Press, New York, NY, pp. 20–40.
McMichael, Philip, 2005. Global development and the corporate food regime. Research in Rural Sociology and Development 11, 265–299.
McMichael, Phillip, 2000. The power of food. Agriculture and Human Values 17 (1), 21–33.
Pilgeram, R., 2011. "The only thing that isn't sustainable … is the farmer": social sustainability and the politics of class among Pacific Northwest farmers engaged in sustainable farming. Rural Sociology 76 (3), 375–393.
Pilgeram, R., 2013. The political and economic consequences of defining sustainable agriculture in the US. Sociology Compass 7 (2), 123–134.
Shiva, V., 2000. Stolen Harvest: The Hijacking of the Global Food Supply. Zed Books.
Thompson, P.B., 2010. The Agrarian Vision: Sustainability and Environmental Ethics. University Press of Kentucky.
Tilman, D., et al., 2002. Agricultural sustainability and intensive production practices. Nature 418 (6898), 671–677.
Tokar, B., 2003. Monsanto: a profile of corporate arrogance. In: Goldsmith, E., Mander, J. (Eds.), The Case against the Global Economy & for a Turn towards Localization. Earthscan Publications, London.
Walsh, J., 1975. U.S. agribusiness and agricultural trends. Science 188 (4188), 531–534.
Walton, J., Seddon, D., 1994. Free Markets and Food Riots: The Politics of Global Adjustment. Blackwell, Cambridge, Massachusetts.
WCED, 1987. Our Common Future, World Commission on Environment and Development. WCED.

Sustainable Development

Karl-Werner Brand, Technical University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following the UNCED conference in Rio 1992, the concept of sustainable development has developed into a new model of societal development institutionalized in a large number of international programs and agreements. While the concept itself is widely accepted as a normative guiding principle, fundamental differences appear at the implementation level. The debate on sustainable development thus can be analyzed as a controversially structured, politically contested field of discourse. This article reconstructs both the hegemonic version of sustainable development as 'sustainable growth' and the various counterdiscourses, which – in the North as in the South – refer back to either a stronger ecocentric understanding of nature or put the model of global industrial growth and development as such into question.

Even though the concept of sustainability has been long anchored in the context of forestry, the much more comprehensive idea of sustainable development first entered the international policy debate via the Brundtland Commission report 'Our Common Future' (WCED, 1987) and the subsequent United Nations Conference on Environment and Development (UNCED) in Rio de Janeiro in 1992. The concept of sustainable development has been significantly shaped by the institutional context of the United Nations and refers to previous UN debates, reports, commissions, conferences, and declarations. Within this context, the focus is firstly on *global* problems, and secondly on the goals and action strategies which take the welfare of *humanity as a whole* into consideration. Facing threatened ecological conditions of human development, this aim had already been served some years before by the creation of the United Nations Environment Program at the UN Conference on Human Environment in Stockholm 1972 and the elaboration of the 'World Conservation Strategy' by the International Union for Conservation of Nature. In the early 1980s, two UN commissions had also addressed the growing poverty gap between North and South, dealing with the lack of social and economic development in most developing countries (Brandt, 1980).

Both lines of discussion initially remained largely unconnected, however, because environmental issues constituted a high priority for the industrialized countries while for the southern countries, the development issue stood in the foreground. In addition, environmental issues were negotiated at that time in terms of 'Limits to Growth' (Meadows et al., 1972), which created fundamental conflicts between environmental and development interests. Against this background, the World Commission on Environment and Development (WCED), headed by Gro Harlem Brundtland, was expected to merge the controversial lines of discussion. With the concept 'sustainable development,' the commission appeared to have found the magic formula needed to integrate environmental and social issues into a productive vision of development.

The Hegemonic Sustainability Discourse

What is the core of this concept? The fundamental idea on which the WCED report and the documents of the subsequent UNCED Conference in Rio is based (Rio Declaration, Agenda 21) is the vision of a global development of mankind that allows the needs of all people, present as well as future, to be met. This anthropocentric vision refers to the normative principles of both intra- and intergenerational fairness. It aims to integrate environmental, social, and economic dimensions of development and to translate them into verifiable objectives and measures with the aid of a broad participation of society. It calls for "harmony among human beings and between humanity and nature" (WCED, 1987: 65) and it is supported by an optimistic belief in the feasibility of this vision.

Although the WCED report discusses key problems, their interconnections, and possible solutions (population and human resources, food security, species and ecosystems, energy and development, sustainable industrial development, the urban challenge), it is not a scientific but rather a political document. It attempts to bridge the existing controversies between the commission members, their divergent interests, and interpretation of problems. It therefore remains vague in many respects. But it is precisely this vagueness that allows for a broad consensus.

The question is *how*, by which kind of framing, the WCED report and the Rio documents succeed to bridge the contradiction between global environmental and developmental interests. This 'reconciliation' is accomplished in two argumentative steps (Dingler, 2003: 222f.). The first is the transformation of 'sustainability' from a concept focused on resource conservation to a concept focused on the satisfaction of basic human needs, including the needs of future generations. The second transformation step is to no longer view economic growth as a 'limit' but as a prerequisite for sustainable development. Because the negative environmental consequences of industrialization and economic growth are understood as a result "of the present state of technology and social organization on environmental resources" and not as a result of growth per se (WCED, 1987: 8), and, at the same time, the solution to poverty is seen in economic development, it is technology and social organizations that have to be improved. The solution to global environmental and developmental problems, therefore, would not be in the absence of growth, but in a new era of reinforced, ecologically sound '*sustainable growth.*' As a means to promote this goal (1) an

increase in efficiency, (2) dematerialization, and (3) a rational, global management of resources are recommended (Dingler, 2003: 227).

The Rio Conference takes up this conceptual reframing of sustainability. In the 'Agenda 21,' a comprehensive action plan for the twenty-first century, the emphasis is on making the production and consumption patterns of the North more sustainable. In addition to increasing efficiency and the demand for new, environmentally sound technologies, the importance of science for rational resource management is particularly emphasized, thus significantly enhancing the scientific nature of the discourse on sustainable development. In addition, Agenda 21 is concentrating on environmental economic instruments of trade liberalization and the opening of markets, which promotes the economic reformulation of the concept of sustainability into categories of economic, environmental, and human 'capital.' Through this, the discourse on development policy with its focus on poverty and basic needs, which still dominated the Brundtland report, was tacitly transformed into a *discourse of ecological modernization of economic growth*. Sustainable development in the South then means the transfer of 'green' growth patterns from the industrialized countries to developing countries.

This hegemonic conceptual framing of sustainability is accompanied, however, by another, procedural element: the emphasis on the *participatory* nature of searching for and trying out new paths of sustainable development. In particular, the action plan put forth by Agenda 21 has provided socially and ecologically engaged groups and organizations all over the world a new legitimation for participation in planning processes and the collaborative search for sustainable solutions to local problems.

Meanwhile, this hegemonic variant of the sustainability discourse has been included in many policy areas of Western countries. Local Agenda 21 processes have initiated a multitude of sustainability projects. New approaches to sustainable production and consumption were tested by businesses and committed consumers, and have evolved in some cases, e.g., in the field of organic food consumption, to absolute trendsetters. At the international level, following the Rio conference, a veritable marathon of negotiations and conferences emerged to bring about mandatory international regimes into the areas of climate, biodiversity, forests, fisheries, water conservation, toxic waste transports, etc. (cf Winter, 2006).

Counterdiscourses on Sustainable Development

The hegemonic understanding of sustainable development, however, encountered considerable criticism in southern as well as in northern countries. This criticism can be assigned to two lines of discourses (cf Banerjee, 2003; Dingler, 2003; Dryzek, 1997; Escobar, 1995; Milton, 1996; Sachs, 1999): a first counterdiscourse that follows a different understanding of nature (1) and a second counterdiscourse that follows a different understanding of development and growth (2).

All three positions, the hegemonic discourse and its criticisms, as well as the various subthreads of the debate, can be graphically located in a field of discourses on sustainable development that can be spanned along two axes (see **Figure 1**)

- along a horizontal axis of divergent images of nature and technology with the opposing poles 'techno-centrism' (technical-industrial model of progress) versus 'eco-centrism' (embedding society in the web of nature), and
- along a vertical axis of divergent models of economy and societal organization with the opposing poles 'world-market model of industrial growth' versus 'models of local self-sufficiency and self-determination.'

Nature-Related Counterdiscourses

As part of the hegemonic discourse of sustainability, the preservation of ecosystems and natural resources is of instrumental character; it serves the sustainable satisfaction of human needs. According to this view, the development of human needs is not subject to any absolute natural barriers, but only to (rectifiable) social and technical limitations. This typical modern discourse on industrial progress is put into question from two positions: (1) from radical ecologists who reject the anthropocentric premise of the sustainability discourse and postulate that nature has its own, intrinsic values that must be respected; (2) from a neo-Malthusian understanding of 'limits to growth' which postulates that the limited carrying capacity of the Earth may be exceeded only at the cost of collapse.

Radical Ecology

'Radical ecologists' (Eckersley, 1992; Merchant, 1992) conceive human society not as the center, but only as one element of a larger living community whose natural elements, animals, plants, ecosystems, landscapes, the planet as a whole, have the same rights for development as humans (see Dryzek, 1997: 155 ff.). Throughout the past century this concept of man and nature was one of the basic motives of conservation and environmental movements. It formed its romantic and spiritual pole whose other pole was represented by a rational, utilitarian concept of sustainable resource management (cf Hays, 1959; Nash, 1967). In the US context, ecocentric positions are closely linked to 'deep ecology,' 'ecofeminism,' and 'bioregionalist' models of social life.

Deep ecology (cf Devall and Session, 1985; Naess, 1973) disengages itself from environmental policy approaches that would only like to reform industrial society in some technical or economic aspects. Deep ecology is concerned with more fundamental changes. "Its two basic principles are self-realization and biocentric equality. Self-realization means identification with a larger organic 'self,' beyond the individual person. […] The idea is to cultivate a deep consciousness of organic unity, of the holistic nature of the ecological webs in which every individual is enmeshed. […]. Biocentric equality means that no species, including the human species, is regarded as more valuable or in any sense higher than other species" (Dryzek, 1997: 156).

Ecofeminism (cf Diamond and Orenstein, 1990; Plumwood, 1993) "covers a broad set of ideas, analyses, and action forms that have evolved since the 1970s in the context of women's activism in the environmental and peace movements in the North and South. These approaches are interconnected by looking for common causes of natural exploitation and discrimination of women and the goal of linking

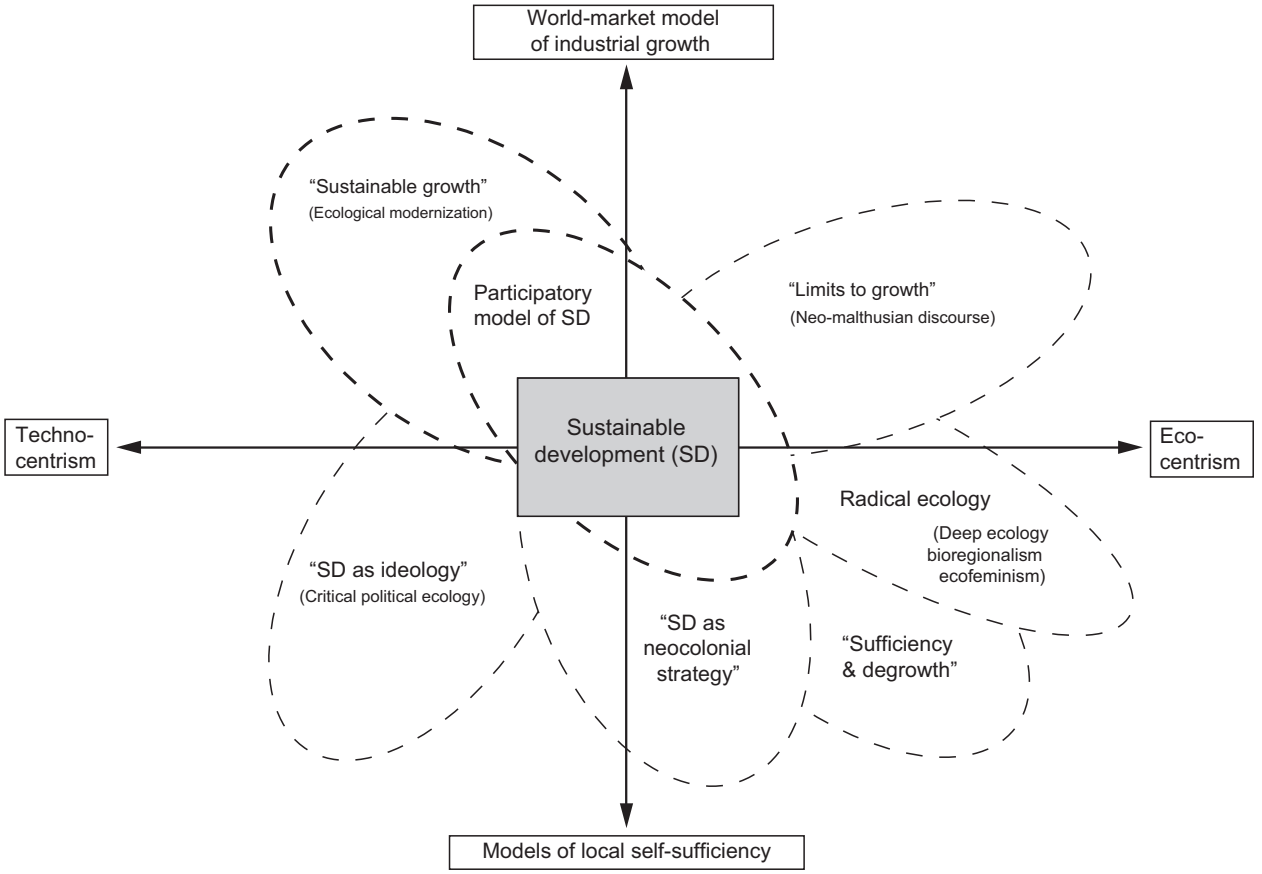

Figure 1 The field of discourses on sustainable development.

environmental policy to the empowerment of women" (Vinz, 2005: 7). Vinz distinguishes between two schools of thought within ecofeminism, 'cultural' and 'social' ecofeminism. While the latter examines the social recoding of nature and gender with the development of modern science, giving way to a new exploitation of nature and oppression of women, the former plays a more central role in radical ecology. This cultural variant suggests a "special closeness of females to nature due to bodily (menstruation, childbirth) and spiritual reasons"; female-specific values (feelings, empathy, corporeality) – in contrast to an instrumental, male approach to nature – are seen as the "guarantor for a respectful relationship with nature and for a life in peace" (Vinz, 2005: 7).

Bioregionalists (cf Sale, 1985) share, together with cultural ecofeminism, the practical, anti-industrial utopia of agricultural self-sufficiency and reruralization. They also propagate small, decentralized settlements, adapted to their natural environments, which should enable people to live in 'ecological balance' with the surrounding nature.

In densely industrialized and urbanized countries these are scarcely feasible options for action, however. Ecocentric positions, therefore, are frequently expressed as direct action protest in diverse fields of conflict like transportation, agriculture, or energy policy. Not only the foundation of new radical ecological organizations like Greenpeace (1971), the Sea Shepherd Conservation Society (1979), or Earth First! (1979), but also the direct actions of animal rights activists or opponents of green genetic engineering are all fed by the unconditional moral impetus of the 'defense' of the maltreated nature, whether animals, landscapes, or people.

Limits to Growth

Biologists, environmental scientists, and environmental activists usually share the idea that human society is only one subsystem of nature. Human use of natural resources is part of the overall dynamics of planetary ecosystems. This is analyzed in categories of material and energy flows as well as in categories of the Earth's human carrying capacity. According to the 'limits to growth' approach, both the resource availability and the sink function of planetary ecosystems then have clear natural boundaries that limit economic growth (Daly, 1996). Whereas this discourse is compatible with various reformist top-down and bottom-up strategies, the prominent neo-Malthusian discourse of 'Survivalism,' advocated by Hardin, Ehrlich, and Meadows et al., relies on radical, hierarchical control strategies that are to limit both population and economic growth (Dryzek, 1997: 23ff.). The need for such a strict control strategy is justified by the argument that humankind has already overshot the Earth's carrying capacity by far. This argument is impressively underlined by exponential growth curves, global computer models, and dire predictions about the consequences of the impending ecological collapse.

Criticism of the Conventional Concepts of Development and Growth

This counterdiscourse includes a broad range of critical positions. The most prominent of them are (1) the anti-imperialist discourse of small farmers and indigenous people in the South, (2) a complementary Northern critique of consumerism, and (3) the Marxist critique of social and ecological exploitation by the dynamics of capitalist accumulation.

Sustainable Development as a Neocolonial Strategy

From this perspective, the hegemonic strategy of sustainable development is marginalizing subsistence-based ways of life of small farmers and indigenous peoples in the South. The central criticism is that the industrial model of 'sustainable growth' follows a neocolonial pattern of 'development' that accelerates the destruction of smallholder agriculture close to nature only under new, ecological auspices (Escobar, 1995; Mies and Shiva, 1993; Sachs, 1999). Linked to this is the criticism of the hegemonic model of Western rationality that devalues local knowledge and uses it only instrumentally as a patentable resource for pharmaceutical purposes. The winners of this model of sustainable development seem to be multinational companies, national governments, and the urban consumer class who all profit from 'green' innovative technologies and from 'sustainable growth.' The losers are the approximately 2 billion people who depend on land, forests, and water as their traditional means of livelihood.

The conventional notion of progress and development, therefore, should be fundamentally questioned. Local rural protest movements are thus seen as supporters of emancipatory visions, fighting not only for a decentralized control over natural resources and habitats, but also for the recognition of cultural diversity. The myth of the 'primitive ecological wisdom' of indigenous and traditional peoples who lived across many generations in 'harmony with nature' and as responsible guardians of environmental resources links this local community discourse with the discourse of radical ecologists (Milton, 1996).

Sufficiency and Degrowth Discourse

In the industrialized countries of the North, a complementary criticism of the hegemonic sustainability discourse is found. It is based on the normative principle of ecological justice as well as on a new concept of 'good life.' Given the social inequalities and the ecologically disastrous consequences of global economic growth, the proponents of this position plead for both the dismantling of global industrial dependencies and the restraint of consumption in the North. New models of prosperity have to be developed which follow a 'cosmopolitan localism' and are guided by principles like "living well instead of possessing much" (Sachs, 1999; Sachs et al., 1998). Among economists, this vision has been discussed for some years now under the label 'degrowth' (Jackson, 2009; Latouche, 2009). The central argument is first that the possibility of dematerializing economic growth and decoupling it from energy consumption (not only in relative but also in absolute terms) is an illusion, and second that an increase in material goods above a certain level by no means increases the subjective well-being, but only leads to additional burdens and dependencies. The aim is, therefore, to strengthen regional economies and to achieve a new balance between self-sufficiency and market economy. Instead of consumerism, life has to be refocused on the idea of 'modest wealth' (Latouche, 2009).

Green Marxism/Critical Political Ecology

This position generally assumes that local and global environmental problems are an inherent part of a capitalist economy that is pushed forward by a limitless need for capital valorization and accumulation. This implies both social exploitation and the exploitation of nature. Thus, capitalism is not only responsible for repeated social and economic crises, but it also increasingly undermines the natural conditions of human life. Nevertheless, the possibilities of regulating the ecological crisis are considered controversial in the eco-Marxist debate. For some, the exacerbation of global environmental problems are regarded as a new frontier of capitalism (Altvater, 2005; Clark and York, 2005; O'Connor, 1988), while others see in the Rio model of 'sustainable growth,' at least a possibility for regulating the ecological crisis (Brand and Wissen, 2012; Peet et al., 2011). This, however, does not solve the inherent exploitative character of capitalism. Rather, the sustainability concept is seen as an ideological, depoliticizing concept that interprets global environmental problems as a common and 'external' threat to all mankind, while it conceals the existing power relationships and inequalities in the reconstruction of a 'green capitalism' (Swyngedouw, 1997).

The program of 'sustainable growth' thus has been strongly questioned from different positions. The resulting conflicts, however, were significantly mitigated by the participatory character of the sustainability process inspired by Rio. On the one hand, this increased the importance of Non Governmental Organizations in the various fields of global environmental governance. On the other, this also led, on the local and national level, to the flourishing of new forms of citizen participation, dialogical forms of politics, and cooperative networks to promote sustainable development. This vivid field of activities, in nationally varying forms, also included representatives of critical positions (for the German case see Brand, 2011).

Perspectives of Sustainable Development

The first wave of sustainability initiatives that were carried by the 'spirit of Rio' waned at the beginning of the new century. New topics began to dominate the international agenda. Although the dramatic warnings about the catastrophic consequences of climate change remained a key issue in the media, '9/11,' the 'war on terrorism,' and the deep financial and economic crisis that swept the Western world since 2008 have changed political priorities. In this context the concept of 'green economy,' a new political focus on 'low-carbon technologies,' and renewable energies, seemed to be an adequate answer both to global economic and environmental problems. It provides a new magic formula "that embodies the promise of a new development paradigm, whose application has the potential to ensure the preservation of the earth's ecosystem along new economic growth pathways while contributing at the same time to poverty reduction" (UN DESA, 2011: v). Considering

the dependence of profound changes in the human relationship to nature to equally fundamental changes in the structural organization of societies, it is unlikely, however, that an economic conversion to green technologies per se will achieve the desired general solution to global environmental and poverty problems (Brand, 2012: 30f.). As ecological transformation and issues of global justice are inseparably linked with each other a wider, integrative concept of sustainable development will remain on the agenda in the foreseeable future.

See also: Degrowth; Eco-cities and Sustainable Urbanism; Environmental Sociology; Limits to Growth; Planning Issues and Sustainable Development; Sustainability and Sustainability Science; Sustainability: A Dialog of Values; Sustainable Development and Social Work; Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption; Water Resources and Sustainable Water Management.

Bibliography

Altvater, E., 2005. The End of Capitalism as We Know It. A Radical Critique of Capitalism (In German). Westfälisches Dampfboot, Münster.

Banerjee, S.B., 2003. Who sustains whose development? Sustainable development and the reinvention of nature. Organization Studies 24 (1), 143–180.

Brand, K.-W., 2011. Sociological perspectives on sustainability communication. In: Godemann, J., Michelsen, G. (Eds.), Sustainability Communication. Interdisciplinary Perspectives and Theoretical Foundations. Springer, Dordrecht, pp. 55–68.

Brand, U., 2012. Green economy – the next oxymoron. No lessons learned from failures of implementing sustainable development. GAIA 21 (1), 28–32.

Brand, U., Wissen, M., 2012. Crisis and continuity of capitalist society-nature relationships: the imperial mode of living and the limits to environmental governance. Review of International Political Economy, iFirst, 1–25.

Brandt, W. (Ed.), 1980. North-South: A Program for Survival. The Report of the Independent Commission on International Development Issues under the Chairmanship of Willy Brandt. MIT Press, Cambridge, MA.

Clark, B., York, R., 2005. Carbon metabolism: global capitalism, climate change, and the biospheric rift. Theory and Society 34 (4), 391–428.

Daly, H., 1996. Beyond Growth: The Economics of Sustainable Development. Beacon Press, Boston.

Devall, B., Session, G., 1985. Deep Ecology: Living as if Nature Mattered. Gibbs Smith, Layton, UT.

Diamond, I., Orenstein, G.F. (Eds.), 1990. Reweaving the World: The Emergence of Ecofeminism. Sierra Club Books, San Francisco.

Dingler, J., 2003. Postmodernity and Sustainability. A Discourse-theoretical Analysis of the Social Construction of Sustainable Development (In German). Oekom, München.

Dryzek, J., 1997. The Politics of the Earth. Environmental Discourses. Oxford University Press, Oxford.

Eckersley, R., 1992. Environmentalism and Political Theory: Towards an Ecocentric Approach. University College London Press, London.

Escobar, A., 1995. Encountering Development: The Making and Unmaking of the Third World, 1945–1992. Princeton University Press, Princeton, NJ.

Hajer, M., 1995. The Politics of Environmental Discourse. Ecological Modernization and the Policy Process. Oxford University Press, Oxford.

Hays, S.P., 1959. Conservation and the Gospel of Efficiency: The Progressive Conservation Movement, 1890–1920. Harvard University Press, Harvard.

Jackson, T., 2009. Prosperity without Growth: Economics for a Finite Planet. Earthscan, London.

Latouche, S., 2009. Farewell to Growth. Polity Press, Cambridge.

Meadows, D., Meadows, D., Randers, J., Behrens III, W., 1972. The Limits to Growth. Universe Books, New York.

Merchant, C., 1992. Radical Ecology. Routledge, London.

Mies, M., Shiva, V., 1993. Ecofeminism. Spinifex Press, Melbourne.

Milton, K., 1996. Environmentalism and Cultural Theory. Exploring the Role of Anthropology in Environmental Discourse. Routledge, London.

Naess, A., 1973. The shallow and the deep. Long-range ecology movement: a summary. Inquiry 16, 95–100.

Nash, R., 1967. Wilderness and the American Mind. Yale University Press, New Haven.

O'Connor, J., 1988. Capitalism, nature, society: a theoretical introduction. Capitalism, Nature, Socialism 1 (3), 15–50.

Peet, R., Robbins, P., Watts, M., 2011. Global nature. In: Peet, R., Watts, M. (Eds.), Global Political Ecology. Routledge, London, pp. 1–47.

Plumwood, V., 1993. Feminism and the Mastery of Nature. Routledge, London.

Sachs, W., 1999. Planet Dialectics. Explorations in Environment and Development. Zed Books, London.

Sachs, W., Loske, R., Linz, M., 1998. Greening the North. A Post-Industrial Blueprint for Ecology and Equity. Zed Books, London.

Sale, K., 1985. Dwellers in the Land: The Bioregional Vision. Sierra Club Books, San Francisco.

Swyngedouw, E., 1997. Impossible/undesirable sustainability and the post-colonial condition. In: Krueger, J.R., Gibbs, D. (Eds.), The Sustainable Development Paradox. Guilford, New York, pp. 13–40.

UN DESA (United Nations Department of Economic and Social Affairs, 2011. The Great Green Technological Transformation. World Economic and Social Survey. UN DESA, New York.

Vinz, D., 2005. Sustainability and Gender: Environmental Policy from the Perspective of Gender Studies (In German). Gender-Politik. Online, January 2005. http://www.gender-politik-online.de (accessed 20.03.14.).

WCED (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, Oxford.

Winter, G. (Ed.), 2006. Multilevel Governance of Global Environmental Change. Cambridge University Press, Cambridge.

Sustainable Development and Social Work

Susan Buckingham, Brunel University, Uxbridge, UK
Victoria J Kina, University of Dundee, Dundee, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the emergence of sustainable development, considers the problems attached to it, and argues that a critical and holistic understanding of sustainable development in which the environmental, economic, cultural, and social are deeply interwoven may be one potential pathway for social work to contribute to the human and nonhuman world.

Introduction

The aim of this article is to define sustainable development in a way that is useful to social workers in their teaching, study, and practice. Using the conventional definition of sustainable development as its starting point, the authors argue that this is both too limited, and in its universality too unspecific, to be meaningful. They draw attention to the indivisibility of the social, environmental, and economic, as well as the cultural, in the ways in which we live, the ways in which our natural and social world has become unsustainable, and therefore the ways in which we need to address this unsustainability. Each of these characteristics (the first three of which are sometimes called 'pillars' or 'legs of a stool') imbues the others, so that one cannot understand the economic without understanding how it is socially, culturally, and naturally re/produced. In effect, each gives us a perspective, or a starting point, but only inasmuch as it enables us to understand the whole.

Having established the common definition, the authors review the evidence that suggests that its promise has not materialized and that even where local, small-scale sustainability appears to have been achieved, this is compromised by the outsourcing of pollution and resource exploitation needed to achieve sustainability in a single place. The authors argue, then, by introducing notions of environmental justice and injustice, that nothing short of global sustainable development is fair, both internationally and intranationally. The authors consider how the sustainable development agenda is being replaced in ways that they consider to be both positive (through considerations of well-being) and more problematic (for example, through resilience). They consider a number of examples that point to both the opportunities and challenges for sustainable development, foremost among which is climate change.

Following this more general discussion, the authors identify how social workers need to engage with sustainable development, from being explicit as to how environmental problems negatively affect the people and groups who social workers work with, and how an holistic approach to social, economic, and environmental problems is likely to have a positive impact on social work service users. Ultimately, the authors argue, a failure to holistically attend to the problems that create unsustainable development will affect social workers, social service providers, users, and carers, which in turn makes the case for social workers to be aware of, and active in, working toward a more sustainable environment.

Definition

The term sustainable development has come into common usage since *Our Common Future* was published in 1987. This was prepared by the World Commission on Environment and Development (WCED), a United Nations–sponsored enquiry into how the world community could negotiate an environmentally, socially, and economically sustainable future with costs and benefits shared between rich and poor countries. While this definition coined by the WCED (also known as the Brundtland Commission after its chairperson, Gro Harlem Brundtland) has become well rehearsed, it is also much misused, and often misunderstood:

> Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. It contains within it two key concepts:
>
> - the concept of 'needs,' in particular the essential needs of the world's poor, to which overriding priority should be given; and
> - the idea of limitations imposed by the state of technology and social organization on the environment's ability to meet present and future needs. (WCED, 1987: p. 43)

As Erik Swyngedouw has commented, sustainable development is supported by everyone, from Greenpeace to Bill Gates and the labor unions (2007). The WCED's definition was a compromise necessitated by a group of policy makers from countries across the rich and poor worlds. While much of the rich West demanded selective restrictions on growth to protect rapidly degrading environments, more powerful poorer countries such as Brazil, Malaysia, and Indonesia challenged these demands on the grounds that their capacity to develop their economies would be compromised. Sustainable development, then, was a compromise, and a compromised concept, which promoted the need for economic growth, while acknowledging that this should not come at the expense of environmental and social degradation. Others argue that sustainable development provides a discursive justification for 'ecological modernization,' or development inflected with a pale green tinge of bike paths and biofuels, but that does not fundamentally challenge development trajectories or balances of power (Krueger and Gibbs, 2007; Pepper, 1996).

The Performance

Almost 30 years after this definition was published, the world is far less sustainable than it was in 1987. Sustainable development comprised one of the United Nations Millennium Goals, set out in 2000, with clear targets to be achieved by 2015. The United Nations acknowledges that there is scant chance for Goal 7, to ensure environmental sustainability, to be met, with the only achievement appearing to be successful being to halve the proportion of people without sustainable access to safe drinking water.

The world has also seen significant changes since 1987, which have affected its ability to achieve sustainable development: the fall of the Berlin Wall following the collapse of the centrally planned economies of the USSR and central and eastern Europe; the first Gulf War fought over oil resources in the Middle East; the invasions of Afghanistan and Iraq (arguably as much motivated by the energy security fears of the United States and the United Kingdom, as by other security issues); the rise of the new economies of Brazil, Russia, India, and China whose energy and resource needs pose a significant threat to environmental sustainability, and which change the balance of power worldwide; the concomitant precarity of Western economies, which are still attached to traditional growth models; and the advance of global warming evidenced by temperature rises, retreating glaciers and ice caps, and increasingly turbulent weather. Add to this, the apparent inability of rich countries to reduce their resource consumption, while also failing to contribute as much as they initially agreed to funding poor countries' attempts to better manage their environments, and the widening gulf between the richest and the poorest, and it is clear that sustainable development has been, at the global scale, a chimera. The attachment of international organizations and individual states to an economic accounting system measured by gross domestic product (GDP) developed in the wake of the Second World War to stimulate production, is outdated and ill-equipped to address contemporary problems, such as climate change (Waring, 1988).

Scale

In order to achieve sustainable development globally, there also needs to be action at national and subnational levels and it is here that social work could take a central role. The articulation of this interrelationship by the 'Think Global, Act Local' entreaty has inspired many local communities and governments to tackle unsustainable practices, where these are under their control, yet social work – particularly in the Global North – has been slow to actively engage with these possibilities. The individualized nature of social work practice has shifted focus away from the creative possibilities offered through global/local initiatives even when change could potentially be within their grasp. While some localities have been imaginative and energetic in pursuing this, as evidenced by various alliances (for example, International Council for Local Environmental Initiatives, Local Governments for Sustainability), including some that have made significant achievements (such as through the Transition Towns movement), social work has broadly failed to engage with these social and community development movements.

One reason for this is the continued emphasis on economic growth driving government policies worldwide, which severely compromises the ability of both local governments and social work departments to follow policies and practices that prioritize social and environmental sustainability. There is now a groundswell of opinion from a number of prominent economists, among others, that historic economic growth models are themselves unsustainable, and increasingly unable to deliver economic sustainability, let alone able to foster environmental and social sustainability. For example, Nicholas Stern, former Chief Economist at the World Bank, has argued in a major report for the UK Treasury for the precautionary principle to guide legislation on climate change (Stern, 2006), while Joseph Stiglitz, another former Chief Economist at the World Bank, and who won the Nobel Memorial Prize in Economic Sciences in 2001, reported in 2009 for a Commission set up by the then President of France, Nicolas Sarkozy, that economies could not sustain continual growth at the expense of the environmental and social damage they were doing internationally as well as within countries (Stiglitz et al., 2009). Also in 2009, Tim Jackson published a report for the United Kingdom's Sustainable Development Commission, which was highly critical of economic growth, and advocated an alternative model of 'Prosperity without Growth' in order to achieve environmental and social, as well as economic, sustainability. Given that most rich countries have, despite a reliance on traditional economic growth models, experienced economic recession since these reports have been published emphasizes that these models do not have the capacity to deliver sustainable development.

Within social work in the Global North, neoliberal approaches to national development have promoted an increasingly individualized approach to social work practice. Casework has come to dominate service provision, resulting in neglect of opportunities for community development and action. This highlights both the need for, and the challenge of, adopting a global perspective on sustainable development. The individualized notion of practice results in the impact of environmental issues tending to be addressed superficially through the experiences of one family, individual, or service user group rather than through the explicit recognition of the interconnectedness of the modern world. To overcome this, a global perspective on sustainable development is required for a number of reasons. First, without this, it would be possible for individual countries (or towns, cities, or regions) to claim sustainability internally, while contracting out their unsustainable practices (from exporting polluting industries to countries with cheaper labor costs or lower health and safety regulations, to moving vulnerable populations to poorer areas with cheaper housing costs). This only accentuates the divide between rich, 'clean,' 'sustainable' areas and poor, 'dirty,' 'unsustainable' ones. On this basis, sustainability (the goal of sustainable development) can only be achieved globally, and until then it is likely that sustainable development in one area is contingent on unsustainable development somewhere else, directly contradicting the social justice principles on which social work is based. Second, given the highly internationalized nature of industry (from oil and gas to finance), some form of

global governance is necessary to prevent multinational companies from exploiting differences in sustainable development legislation to continue unsustainable practices, rather than to change the ways in which they produce goods and do business. This highlights the need for social work both to develop an international voice, as advocated by the International Federation of Social Workers (IFSW) (2013), and to engage politically with arenas traditionally deemed outside the social work 'remit' (see how this is being developed in article Global Agenda for Social Work and Social Development).

Well-Being

There is increasing emphasis from academics from various disciplines across the social sciences, think tanks, and nongovernmental organizations, on alternative concepts to 'growth,' which are focusing on 'well-being' (Jackson, 2009; NEF, 2009). The decoupling of economic wealth from well-being has been identified as the 'Easterlin paradox' (Easterlin, 1974) after work undertaken by the American economist who identified that as a society gains in economic wealth this is not necessarily reflected by a rise in happiness, relative to other countries (NEF, 2009). This is likely to be aggravated in societies that privilege economic growth above any other kind of development, such as the United States and the United Kingdom, but will not characterize countries such as Bhutan (ranking 141 of 187 on the 'World Development Index' measured by the United Nations Development Programme Human Development Indicators), which consistently comes top of world rankings of happiness. In 1971, Bhutan decided to measure its wealth not by GDP, but by gross national happiness, which measures "the quality of a country in a more holistic way [than GDP] and believes that the beneficial development of human society takes place when material and spiritual development occurs side by side to complement and reinforce each other" (Ura et al., 2012: pp. 6–7). Bhutan's achievements, which prioritize environmental conservation and sustainability, include a doubling of life expectancy over 20 years, enrolling almost 100% of its children in primary school, and an overhaul of its infrastructure. By doing so, it offers the international community an alternative set of values and aspirations, and a mechanism to measure these. The irony is, however, that their ability to achieve these is compromised by global environmental degradation, which affects water resources and pollution across the Himalayan region, over which Bhutan has no control (Kelly, 2012).

Research undertaken by NEF (2009) identifies the aspects that people interviewed across Europe attribute to their own happiness and well-being. In this, well-being is categorized into personal and social attributes, such as a satisfying life, vitality, resilience and self-esteem, positive functioning, supportive relationships, trust and belonging, and well-being at work. Not only did this identify countries whose populations experience higher states of well-being (Scandinavian countries, led by Denmark, and Switzerland did better on both measures; Central and Eastern European countries did worst), but also highlighted intranational differences. For example, across Europe, certain circumstances indicate stronger or weaker well-being. Being a woman, belonging to an ethnic minority group, living with children, being ill or disabled, and being unemployed (but not retired, which had a positive correlation with well-being) were all negatively associated with well-being. While this may not be particularly surprising to social workers from across the globe whose daily work is to engage with the impacts of these subjective inequities, it highlights the role that social workers can play in these newly emerging debates.

These attributes suggest that the reach of the original definition of sustainable development is neither wide nor specific enough to capture all dimensions of people's lives. Moreover, a further dimension – a fourth 'leg' of the sustainability 'stool' – has been proposed by the United Nations Educational, Scientific and Cultural Organization's World Commission on Culture and Development so that cultural policies become a key part of sustainable development (Leach, 1998).

Climate Change

Just how ineffective sustainable development has been to date is evident from the global approach toward climate change. While identified as long ago as 1896, climate change first reached international prominence outside specialist circles in 1992 when the United Nations Conference on Environment and Development held in Rio de Janeiro drafted a climate change protocol whose purpose would be to hold or return global CO_2 emissions to 1990 levels. This would be achieved primarily by Annex 1 (rich) countries, which, once having signed up to the protocol, would develop their own climate change reduction strategies. The Kyoto Protocol, so named for the city of its launching, was initiated in 1997, and eventually came into force when the required number of signatories had been achieved in 2005. Russia and Australia signed at this point, although the world's largest CO_2 emitter, the United States, never endorsed the protocol. This was a mere 7 years before it was due to expire in 2012, and the questions of how to replace it, and which countries should be included in its renegotiation, have vexed annual climate change conferences since (notably Copenhagen in 2009). For further details on Kyoto and its successors, see the United Nations Framework Convention on Climate Change Web site. Of particular note, and germane to the discussion of sustainable development here, is the gradual and almost imperceptible change of the discourse around climate change, from 'mitigation' (preventing the conditions that create climate change) to 'adaptation' (adapting to a climate changed world). This has invoked a shift from creating 'sustainable development,' which might suggest some changes to the ways in which resources are transformed, used, consumed, and disposed of, to 'resilience,' which implies changes in the ability of people, communities, and/or countries to cope with the effects of climate change, be these increased flooding, changes in food supply, or patterns of disease. However, resilience (and capability, which enables people or communities to become resilient) is limited by lack of basic resources, as Amartya Sen has noted in his body of work over many years (see, for example, Sen, 1999). People who are poorly sheltered, hungry, or malnourished, who are illiterate

and/or poorly educated, are least able to withstand environmental problems. Moreover, these conditions disproportionately expose them to environmental problems such as pollution, which in turn affects their capacity for resilience. This not only affects people in poor countries but also is a recognized challenge for social work across both the Global North and the Global South.

Environmental In/Justice

The WCED definition of sustainable development extends the issue of intergenerational equity to a concern that "must logically be extended to equity within each generation" (Chapter 2, point 3). It continues:

> A world in which poverty and inequity are endemic will always be prone to ecological and other crises. Sustainable development requires meeting the basic needs of all and extending to all the opportunity to satisfy their aspirations for a better life.
>
> Living standards that go beyond the basic minimum are sustainable only if consumption standards everywhere have regard for long-term sustainability. (Chapter 2, points 4 and 5)

While there is an assumption that this intragenerational equity needs to be achieved between rich and poor countries, it is also important to note that current intragenerational inequalities persist within the richest countries (see, for example, Wilkinson and Pickett, 2009).

Even in times of expanding GDP, wealth does not 'trickle down' from the wealthiest to the poorest within the richest nations. Despite continuous rises in GDP in the United Kingdom and the United States, both countries have increasing levels of wealth inequality (Reuters, 2012). According to Reuters, using the Gini coefficient (where 0 equals perfect parity and 1 represents all wealth going to a single person), inequality in income distribution in the United States has risen in each of the past two decades. Between 1989 and 2011, the index rose 6% nationally, to its highest level since the US census started measuring it in the 1960s. Using census data to compare average income, the top 5% of income earners on average earned 24 times as much as the bottom 20% in 1989. By 2011, that ratio was 28 to 1 (Reuters, 2012). According to Danny Dorling, in 2012, the richest 1% of people in the United Kingdom took home 15% of all income, compared to 6% in 1979, higher than at any time since 1940 (Dorling, 2012). Wilkinson and Pickett (2009) have identified how greater income inequality correlates with higher rates of crime and mental illness, with countries such as the United States and Portugal, and the United States, the United Kingdom, and Australia dominating these respectively.

The concept of the 'Fourth World' was introduced by George Manuel, the founding president of the World Council of Indian Peoples, in the mid-1970s, initially to describe people effectively disenfranchised within their own countries, such as the native North Americans or the Australian aboriginal people (Manuel and Posluns, 1974). The concept has now expanded to include stateless people, and people in both rich and poor countries who are significantly impoverished. Manuel Castells (1998: p. 164) defined the Fourth World as the "multiple black holes of social exclusion" regardless of the wealth of the countries in which these black holes exist, from the slums of Mumbai, rural areas of Latin America, or US inner city ghettos.

This represents a huge challenge for sustainable development in general and social work in particular, and is a social indication of collective failure to engage effectively with the processes necessary to achieve sustainable development. People inhabiting this Fourth World are not 'resilient' to environmental problems which, when they occur, are more likely to impact them more harshly than others. Consider, for example, the victims of aggravated natural weather events such as Hurricane Katrina, which affected the poor and black communities of New Orleans much more than those with savings, insurance, and private mobility (Cutter et al., 2006). The environmental in/justice literature also suggests that the most marginalized of populations are more likely to be exposed to environmental problems than the less marginalized, whether it is because they can only afford housing in the poorest, most polluted neighborhoods, or lack the education, expertise, and connections to fight off polluting activities (see, for example, Bullard, 1999; Kurtz, 2007; Pulido, 1996; Taylor, 2000). Over and above the effect that poverty has on the sustainability of people's lives, race and ethnicity have been found to be disproportionately related to environmental problems (Agyeman, 2013). Gender is less well-recognized in the environmental justice literature, but nonetheless it is clear that being female also disproportionately exposes women, on average, to more environmental problems than men (Buckingham and Kulcur, 2009).

Meanwhile, excesses of consumption are becoming lethal; obesity, a disease of the rich West, is diminishing life expectancy. High consumption of meat products not only compromises health, but also affects the amount of land available to grow food for direct human consumption. Hunger/malnutrition and the inability to feed one's own family are the most fundamental attacks on human autonomy and dignity. And yet, even in the United Kingdom, the world's sixth richest economy (as reported by the Centre for Economics and Business Research in December 2012), a new informal measure of deprivation is the rise of the food bank, with an estimated 200 000 people using food banks in 2012, up from 26 000 in 2008/2009 (Trussell Trust, 2012).

Sustainable Transitions

Among the dispiriting reading about the lack of progress on achieving sustainable development from and between the local to the global scales, there are inspiring projects that address sustainability jointly from economic, environmental, social, and cultural perspectives.

The 'Coriander Club' women's community gardening project in East London, supported by the Women's Environmental Network, illustrates well how women from marginalized Bangladeshi households in East London have been supported with help in transforming common areas within and between council flat buildings. The Bangladeshi community is one of the poorest and most marginalized in the United

Kingdom (Bernard and Turner, 2011). Many are refugees from the 1972 war of independence fought by Bangladesh, and originated in a rural area in which Muslim women experienced living in purdah, did not earn independent incomes, and were largely illiterate. In the United Kingdom, Bangladeshi people have some of the poorest employment prospects, and women continue to be isolated, and more likely to suffer physical and mental ill health than women from other ethnic groups. Meanwhile, their families tend to be larger. Growing their own fruit and vegetables not only enabled these women's families to eat more healthily, but also empowered the women, reducing their isolation and improved a derelict urban neighborhood (Rycroft, 2000).

Local food gardening projects are small but significant ways of achieving greater economic, environmental, social, and cultural sustainability in cities. Such projects can create greater neighborhood cohesion, and offer opportunities for people isolated from the communities in which they live to develop social bonds, which in turn can lead to better physical and mental well-being. As well as refreshing or providing new skills, enabling families to eat more healthily on restricted incomes, and to grow their own food, community gardening reduces the amount of food that is transported unfeasibly long distances. Urban gardening programs have flourished in many cities across North America, in Detroit (see the history of the Detroit Angels who helped recover derelict space at Detroit Urban Sustainability Web site, the link for which is provided below), and notably New York City, although their hold on land here is tenuous, given the pressures on land for housing and other commercial development (Englander, 2001).

Sustainable Development and Social Work

Social work is slowly waking up to the role of the environment in practice. Despite a number of authors campaigning for the recognition of the natural world within social work since the early 1990s (see Gray and Coates, 2012 for an excellent summary), it is only more recently that the debates have begun to really take hold. The IFSW now calls on social workers to "recognise the importance of the natural and built environment to the social environment" and "to develop environmental responsibility and care for the environment in social work practice" (IFSW, 2012). However, research has indicated that social workers in the United States, for example, are no more environmentally aware than members of the general population (Shaw, 2013). Conversely, in Cuba, young social workers are known as 'doctors of the soul,' who recognize the imbrication of the social, economic, and environmental. For instance, in 2006, they visited homes replacing incandescent light bulbs with free energy-saving compact fluorescent bulbs as part of Cuba's 'energy revolution,' and have also been involved in a program to replace household appliances such as fridges and televisions with new energy-efficient models (see the Direct Action Web site listed below).

To date, discussion of social work's potential relationship with the environment has been from two standpoints: first, that social work as a profession needs to think about the natural world as it impacts on humans (environmental justice); second, that the environment is in fact inseparable from the social world and therefore the understanding of the human, social, and natural are inherently intertwined (ecological justice). As summarized by Miller et al. (2012: p. 271), "In environmental justice theories, humans remain the metaphorical sun around which questions of natural resources orbit. In ecological justice theories, nature or Earth expands in scope to become the universe, with recognition that humans are one part of that universe." For social work, the differentiating factor between these two standpoints is an ethical one – as effectively summarized by Gray and Coates (2012): Should one think of sustainable development as being sustainable human development or should social workers also take nonhuman interests into account?

Yet, there is also a moral dimension to the debate that connects with discussions of power, socioeconomic systems, and political structures. It asks us to reflect on not just the interconnectedness of the well-being of the planet and the well-being of humankind (Dominelli, 2012: p. 199) but to connect this socioecological perspective with the critical tradition in social work (Peeters, 2012). In other words, to understand how the exploitation of the Earth's resources is unequally controlled as a direct result of current neoliberal structures and processes. To address this, this dimension of the discussion highlights that social work needs to draw on its critical tradition through working with social movements to build capacity at the local level to resist change as well as advocate for policy change at the governmental level (Peeters, 2012). From this perspective, the indivisibility of the Earth's and people's well-being (Dominelli, 2012: p. 199) further strengthens the argument for the need to bring an end to the 'market-based free-for-all' that not only has prevented poverty eradication but has resulted in preventing equal access to the world's resources (Dominelli, 2012: p. 197).

To date, however, it is sustainable human development that has clearly dominated both social work theory and practice. This can be illustrated by one of the central theoretical approaches to social work practice, the ecological approach, of which the person-in-environment model (Bronfenbrenner, 1979) has come to dominate. This model focuses almost entirely on the social environment which, as Dominelli (2012: p. 3) notes, tends to "ignore power relations based on geo-political social structures that have a deleterious impact upon the quality of life of poor and marginalized populations and the Earth's flora and fauna." This indicates a weakness within the current rights-based approach to social work practice. This can clearly be seen in contemporary notions of social justice as these fail to recognize that social justice can only ever be achieved if environmental justice is also ensured. Therefore, there is an immediate need to expand current understanding of rights to include environmental rights whereby the disproportionate impact of environmental degradation on marginalized or poor communities is explicitly acknowledged. As noted above, there is also an underlying ethical argument, in which social work should also start to challenge itself to move beyond the current human-centered approach to practice through expanding the social work universe so that both human and nonhuman become incorporated into practice models. What the authors have tried to argue throughout this article is that the holistic understanding of sustainable

development in which the environmental, economic, social, and cultural are deeply interwoven highlighted here may be one potential pathway for social work to begin to contribute to this expanded – and sustainable – universe.

See also: Disasters in Social, Cultural and Political Context; Ecofeminism; Ecological Social Work: Shifting Paradigms in Environmental Practice; Global Agenda for Social Work and Social Development; Green Social Work; International Association of Schools of Social Work; Risk Assessment in Social Work; Social Work and Disasters; United Nations and its Agencies.

Bibliography

Agyeman, J., 2013. Introducing Just Sustainabilities. Policy, Planning and Practice. Zed Books, London.

Bernard, H., Turner, C., 2011. Poverty and Ethnicity. A Review of Evidence. Joseph Rowntree Foundation, York.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Buckingham, S., Kulcur, R., 2009. Gendered geographies of environmental injustice. Antipode 41 (4), 659–683.

Bullard, R.D., 1999. Dismantling environmental racism in the USA. Local Environment 4 (1), 5–20.

Castells, M., 1998. End of Millennium. Blackwell Publishers, Oxford.

Coates, J., Gray, M., 2012. The environment and social work: an overview and introduction. International Journal of Social Welfare 21 (3), 230–238.

Cutter, S.L., Emrich, C.T., Mitchell, J.T., Boruff, B.J., Gall, M., Schmidtlein, M.C., Burton, C.G., Melton, G., 2006. The long road home: race, class, and recovery from Hurricane Katrina. Environment: Science and Policy for Sustainable Development 48 (2), 8–20.

Dominelli, L., 2012. Green Social Work: From Environmental Crises to Environmental Justice. Polity Press, Cambridge.

Dorling, D., 2012. Fairness and the Changing Fortunes of People in Britain 1970–2012. Beveridge Memorial Lecture to the Royal Statistical Society, London.

Easterlin, R.A., 1974. Does economic growth improve the human lot? In: David, Paul A., Reder, Melvin W. (Eds.), Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz. Academic Press, Inc., New York.

Englander, D., 2001. New York's Community Gardens. A Resource at Risk. The Trust for Public Land, New York.

Gray, M., Coates, J., 2012. Environmental ethics for social work: social work's responsibility to the non-human world. International Journal of Social Welfare 21 (3), 239–247.

International Federation of Social Workers, 2013. United Nations Calls for a More Sustainable, Just and Prosperous World for All. http://ifsw.org/news/united-nations-calls-for-a-more-sustainable-just-and-prosperous-world-for-all/ (accessed 12.09.13.).

Jackson, T., 2009. Prosperity without Growth. Economics for a Finite Planet. Earthscan, London.

Kelly, A., Saturday December 1, 2012. Gross national happiness in Bhutan: the big idea from a tiny state that could change the world. The Observer, 14.30 GMT.

Krueger, R., Gibbs, D., 2007. The Sustainable Development Paradox. Urban Political Economy in the United States and Europe. Guilford, New York.

Kurtz, H.E., 2007. Gender and environmental justice in Louisiana: blurring the boundaries of public and private spheres. Gender, Place and Culture 14 (4), 409–426.

Leach, M., 1998. Culture and sustainability. In: Emmerji, L., Streeton, P. (Eds.), World Culture Report. United Nations Educational, Scientific and Cultural Organisation, Paris, pp. 93–104.

Manuel, G., Posluns, M., 1974. The Fourth World: An Indian Reality. Collier-Macmillan Canada, Toronto.

Miller, S., Hayward, A., Shaw, T., 2012. Environmental ethics for social work: a principles approach. International Journal of Social Welfare 21 (3), 270–277.

NEF, 2009. National Accounts of Well-being: Bringing Real Wealth onto the Balance Sheet. New Economics Foundation, London.

Peeters, J., 2012. The place of social work in sustainable development: towards ecosocial practice. International Journal of Social Welfare 21 (3), 287–298.

Pepper, D., 1996. Modern Environmentalism. An Introduction. Routledge, London.

Pulido, L., 1996. Environmentalism and Economic Justice. Two Chicano Struggles in the Southwest. University of Arizona Press, Tucson.

Rycroft, V., 2000. Taste of a Better Future: A Participatory Evaluation. Women's Environmental Network, London.

Sen, A., 1999. Development as Freedom. Oxford University Press, Oxford.

Shaw, T., 2013. Is social work a green profession? An examination of environmental beliefs. Journal of Social Work 13 (1), 3–29.

Stern, N., 2006. Stern Review: The Economics of Climate Change. HM Treasury, London.

Stiglitz, J., Sen, A., Fitoussi, J.-P., 2009. Report of the Commission on the Measurement of Economic Performance and Social Progress. CMEPSP, Paris.

Swyngedouw, E., 2007. Impossible "sustainability" and the postpolitical condition. In: Krueger, R., Gibbs, D. (Eds.), The Sustainable Development Paradox. Urban Political Economy in the United States and Europe. Guilford Press, New York.

Taylor, D., 2000. The rise of the environmental justice paradigm: injustice framing and the social construction of environmental discourses. American Behavioural Scientist 43 (4), 508–580.

Ura, K., Alkire, S., Zangmo, T., Wangdi, K., 2012. A Short Guide to Gross National Happiness Index. The Centre for Bhutan Studies, Thimphu, Bhutan.

Waring, M., 1988. Counting for Nothing. What Men Value and What Women Are Worth. Allen and Unwin, New Zealand.

Wilkinson, R., Pickett, K., 2009. The Spirit Level: Why More Equal Societies Almost Always Do Better. Allen Lane, London.

World Commission on Environment and Development, 1987. Our Common Future. Oxford University Press, Oxford.

Relevant Websites

www.cebr.com – Centre for Economics and Business Research, 2012 (accessed 15.08.13.).

http://www.detroiturbansustainabilitytraining.org/ – Detroit Urban Sustainability Training (accessed 13.08.13.).

http://directaction.org.au/issue9/cuban_doctors_of_the_soul – Direct Action for Socialism in the 21st Century.

http://www.theguardian.com/sustainable-business/bhutan-advise-united-nations-happiness?INTCMP=SRCH – The Guardian, 2012.

http://www.iclei.org – ICLEI Local Governments for Sustainability.

IFSW, 2012. Globalisation and the Environment: Policy Statement. International Federation of Social Workers. Available at: http://ifsw.org/policies/globalisation-and-the-environment/ (accessed 27.01.13.).

Reuters, 2012. The Unequal State of America: How We Calculated Inequality. Tuesday December 18, 2012, 7.12 a.m. EST. Available at: http://uk.reuters.com/article/2012/12/18/us-equality-washington-methods-idUKBRE8BH0JG20121218 (accessed 11.1.13.).

http://www.trusselltrust.org/foodbank-projects – The Trussell Trust.

http://www.un.org/millenniumgoals/environ.shtml – United Nations, 2013 (accessed 08.01.13.).

http://unfccc.int/kyoto_protocol/status_of_ratification/items/2613.phpU – United Nations Framework Convention on Climate Change (accessed 12.09.13.).

http://www.wen.org.uk/your-wen/local-food/ – Women's Environmental Network.

Sustainable Development: An Economic Perspective

Edward B Barbier and Joanne C Burgess, University of Wyoming, Laramie, WY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sustainable development is concerned with meeting the needs of the present without compromising the ability of future generations to meet their own needs. Recent advances in three areas of the sustainable development literature are ecosystem services and resilience; environmental Kuznets curves and environment–growth relationships; and technological innovation, endogenous growth, and resource dependency. Progress toward sustainable development requires measuring the social, ecological, and economic impact of declining natural resources; improving our estimates of the costs of depleting natural capital; establishing appropriate incentives, institutions, and investments for sustainable management of natural capital; and encouraging interdisciplinary collaboration.

In the late 1980s, the World Commission on Environment and Development, otherwise known as the Brundtland Commission, defined sustainable development as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs" (WCED, 1987). Since then, sustainable development has received significant attention from the global community at the local, national, and international level amid concerns about climate change, biodiversity loss, tropical deforestation, ecological degradation, and other environmental depletion and degradation problems.

This article provides a broad definition of sustainable development and discusses its various interpretations. Recent advances in three areas of the sustainable development literature are ecosystem services and resilience; environmental Kuznets curve (EKC); and technological innovation, endogenous growth, and resource dependency. For each area, the basic concept is explained and the implications for sustainable development are discussed. Finally, four key policy options for achieving sustainable development are identified: measuring the ecological, social, and economic impacts of declining natural resources; improving our estimates of the economic costs of depleting natural capital; establishing appropriate incentives, institutions, and investments for sustainable management of natural capital; and encouraging interdisciplinary collaboration between the relevant social and behavioral sciences undertaking research on sustainable development.

Concept of Sustainable Development

Definition of Sustainable Development

The broad concept of sustainable development encompasses considerations of equity across and within generations, taking a longer-term perspective, and accounting for the value of the environment in decision-making (Pearce et al., 1989). This interpretation is easily translated into normative terms: an increase in well-being today should not have as its consequences a reduction in well-being tomorrow. That is, future generations should be entitled to at least the same level of social and economic opportunities – and thus, at least the same level of economic welfare – as currently available to present generations. Consequently, economic development today must ensure that future generations are left no worse off than present generations. Or, as some have succinctly put it, per capita welfare should not be declining over time (Arrow et al., 2012; Pezzey, 1989).

Stated in this way, the objective of 'sustainability' is different from that of the standard economic goal of 'efficiency' (Bishop, 1993). That is, there are potentially an infinite number of development paths for an economy, only some of which are sustainable. Efficiency, therefore, does not guarantee sustainability, as some efficient paths are not sustainable. At the same time, there is no reason why an economy could not be both efficient and sustainable.

Consequently, sustainable development continues to receive increasing international recognition, and it has become a key guiding principle for the global society (Barbier and Markandya, 2012; National Research Council, 1999; UNU-IHDP-UNEP, 2012).

Total Capital Stock

The above interpretations of sustainable development also imply that it is the *total* stock of capital employed by the economic system, including natural capital, which determines the full range of social and economic opportunities, and thus well-being, available to both present and future generations. Society must decide how best to use its total capital stock today to increase current economic activities and welfare and how much it needs to save or even accumulate for tomorrow, and ultimately, for the well-being of future generations.

However, it is not simply the aggregate stock of capital in the economy that may matter but also its composition, in particular whether present generations are using up one form of capital to meet the needs of today. For example, much of the interest in sustainable development has risen out of concern that current economic development may be leading to rapid accumulation of reproducible and human capital, or *human-made capital*, but at the expense of excessive depletion and degradation of *natural capital*. The major concern has been that by depleting the world's stock of natural wealth irreversibly, the development path chosen today will have detrimental implications for the well-being of future generations. In other words, according to this view, current economic development is essentially unsustainable.

Some economists argue that the more critical issue of debate is not whether natural capital is being irreversibly depleted

but what are the costs of these losses and can society today compensate future generations for the current loss of natural capital. For example, as Pearce et al. (1989: p. 3) state, "future generations should be compensated for reductions in the endowments of resources brought about by the actions of present generations." A key question is in what form should this compensation take place? On this issue, opinions diverge. This difference of view is often referred to as *weak sustainability* versus *strong sustainability* perspectives.

Weak and Strong Sustainability

Two distinct interpretations of sustainable development are referred to commonly in the literature as 'weak' and 'strong' sustainability (Barbier and Markandya, 2012; Barbier et al., 1994; Neumayer, 2010; Turner, 1993). According to the *weak sustainability* view, there is no inherent difference between natural and other forms of capital, and hence the same compensation rules ought to apply to both. As long as the natural capital that is being depleted is replaced with even more valuable, reproducible, and human capital, then the value of the aggregate stock – comprising human, reproducible, and the remaining natural capital – is increasing over time. Maintaining and enhancing the total stock of all capital alone is sufficient to attain sustainable development.

In contrast, proponents of the *strong sustainability* view argue that reproducible or human capital cannot substitute for all the environmental resources comprising the natural capital stock or all of the ecological services performed by nature. Consequently, the strong sustainability viewpoint questions whether human, reproducible, and natural capital comprise a single homogeneous total capital stock. Instead, proponents of strong sustainability maintain that some forms of natural capital are *essential* to human welfare, particularly key ecosystem goods and services, unique environments and natural habitats, and even irreplaceable natural resource attributes such as biodiversity. Uncertainty over the true value to human welfare of these important assets, in particular the value that future generations may place on them if they become increasingly scarce, further limits our ability to determine whether we can adequately compensate future generations for irreversible losses in such essential natural capital today. Thus, the strong sustainability view suggests that environmental resources and ecological goods and services that are essential for human welfare and cannot be easily *substituted* by human and reproducible capital should be protected and not depleted. The only satisfactory compensation rule for protecting the welfare of future generations is to keep essential natural capital intact. That is, maintaining or increasing the value of the total capital stock over time in turn requires keeping the nonsubstitutable and essential components of natural capital constant over time.

Aspects of Sustainable Development

Ecosystem Services and Resilience

The multiple benefits arising from natural ecosystems have become increasingly critical to many conservation and development decisions worldwide. It is increasingly common to refer to these multiple benefits as *ecosystem services*. Or, as the Millennium Ecosystem Assessment (2005) defines them, "ecosystem services are the benefits people obtain from ecosystems."

The growing literature on ecological services also implies that ecosystems are assets that produce a flow of beneficial goods and services over time (Barbier, 2011; Daily et al., 2000; Dasgupts, 2008). For example, as Daily et al. (2000: p. 395) state, "the world's ecosystems are capital assets. If properly managed, they yield a flow of vital services, including the production of goods (such as seafood and timber), life-support processes (such as pollination and water purification), and life-fulfilling conditions (such as beauty and serenity)." In principle, then, ecosystems should be treated as in other form of wealth in the economy, and by accounting for their contribution to current and future economic well-being, we would have a better measure of economic progress.

If ecosystems are also considered capital assets – or *ecological capital* for short – then efforts to attain sustainable development should account for the contributions of ecosystems as well. However, as Dasgupta (2008: p. 3) reminds us, "ecosystems are capital assets" that "differ from reproducible capital in three ways: (1) depreciation of natural capital is frequently irreversible (or at best the systems take a long time to recover), (2) except in a very limited sense, it isn't possible to replace a depleted or degraded ecosystem by a new one, and (3) ecosystems can collapse abruptly, without much prior warning."

There is mounting evidence that under pressure from external disturbances, ecosystems may switch abruptly to a contrasting alternative stable state, or *ecological regime* (Folke et al., 2004; Holling, 1986; Levin and Lubchenco, 2008; Scheffer et al., 2001). Although a variety of disturbances may trigger the regime shift, what determines whether an ecosystem will change to an alternative state depends on its *resilience* or *robustness* – the ecosystem's ability to absorb large shocks or sustained disturbances and still maintain internal integrity and functioning.

The loss of ecosystem functioning and resilience will lead to a decline in the availability of some economically important biological resources and ecological services. For example, the loss of fish stocks as the result of pollution impacting on marine systems, the inability of rangeland pastoral systems to recover from drought, and the decline of the watershed protection function of degraded forests. At first, these ecological and welfare impacts may be localized, but if they occur on a sufficiently large scale, they may generate wider national, regional, or even global effects such as climate change (*see* Climate Change: Economics). Although the scale of such impacts and their ultimate effects on human welfare are often uncertain and difficult to determine, the irreversible loss of key biological resources and ecological 'services' through the disruption and breakdown of ecosystems could constitute a growing problem of *ecological scarcity* (Barbier, 1989, 2011). The rising 'scarcity' of ecological resources and services will mean that they will increase in value relative to human-made goods and services produced in economic systems. Moreover, persistent and rising ecological scarcity may be an indicator of more widespread and frequent disruptions to the functioning and resilience of ecosystems and increases the likelihood of ecological collapse and catastrophe on ever-widening scales (Folke et al., 2004; Levin and Lubchenco, 2008; Scheffer et al., 2001).

Implications of Ecological Functioning and Resilience for Sustainable Development

Given the rate at which ecological disruption is occurring globally as a result of the cumulative impacts of pollution, habitat modification, and conversion and overexploitation of some biological resources, the ecological functioning and resilience upon which many important ecological resources and services depend may be undermined. Thus, not only is there a danger of depleting essential ecological components of natural capital but the widespread disruptions to the functioning and resilience of ecosystems may mean that these essential components are being lost irreversibly. Therefore, the ecological economics literature generally endorses the strong sustainability view.

Ensuring sustainable development first will require identifying those essential ecological resources and services that are most at risk from current patterns of economic activity and development. Second, it is necessary to take appropriate policy measures to protect the irreplaceable components of natural capital. Finally, it is important to maintain the functioning and resilience of the key ecosystems on which they depend. This implies a fundamental rethinking in the way in which countries currently manage their environments, as well as the relative importance of policies for environmental protection in the process of economic development. The additional costs that are required to adjust current patterns of economic activity and development could be significant, especially in the short and medium term, and may constrain such a comprehensive shift in policy.

EKC and the Environment–Growth Debate

There has been a long-standing debate about the relationship between economic growth and the environment. An early view was expressed in the publication of the Club of Rome's *Limits to Growth* in 1972 that there exists a *trade-off* between economic growth (measured by rising real per capita incomes) and the environment, whereby an improvement in one implies a reduction in the other (Meadows et al., 1972; Daly, 1987). The modern sustainable development debate has focused more on the potential *complementarity* between growth and the environment, where growth is considered compatible with improved environmental quality (see Pearce et al., 1989; Barbier and Markandya, 2012).

An extension to this literature looks at the analysis of EKC. This is the hypothesis that there exists an 'inverted U'-shaped relationship between a variety of indicators of environmental pollution or resource depletion and the level of per capita income (Kuznets, 1955; Grossman and Kreuger, 1995; Shafik, 1994; Stern et al., 1996). The implication of this hypothesis is that environmental degradation should be observed to increase initially, but to eventually decline, as per capita income increases. To some extent, this combines the two previous views on the relationship between economic growth and the environment. That is, at low incomes, there exists a trade-off between economic growth and environmental quality, while at higher incomes, economic growth is complementary to improved environmental quality.

These studies suggest that EKC relationships are more likely to hold for only a few pollutants with short-term and local impacts such as sulfur dioxide (SO_2) and to a lesser extent solid particulate matter. In addition, even when an EKC relationship is estimated, the 'turning point' on the curve, where environmental degradation starts to decline with per capita income, is very high relative to the current per capita GDP levels of most countries. This is a particular problem for less developed countries with low per capita income levels (*see* Environment and Development). However, several recent studies have also demonstrated that EKCs are highly susceptible to structural economic shifts and technological changes, which are in turn influenced by policy initiatives such as the Montreal Protocol and the Second Sulfur Protocol.

Implications of EKCs for Sustainable Development

Some commentators have argued that the empirical evidence of EKCs supports the general proposition that the solution to combating environmental damage is economic growth itself (Beckerman, 1992). For example, the EKC literature offers limited evidence that for certain environmental problems, particularly air pollutants with localized or short-term effects, there is an eventual reduction in emissions associated with higher per capita income levels, which may be attributable to the 'abatement effect,' which arises as countries become richer (Panayotou, 1997). Also, both the willingness and the ability of political jurisdictions to engage in and enforce improved environmental regulations, to increase public spending on environmental research and development, or even to engage in multilateral agreements to reduce emissions may also increase with per capita income levels.

Other commentators have been more cautious, noting that conclusive evidence of an EKC relationship applies only to a few pollutants, which makes it difficult to support more general growth–environment linkages. It has also been noted that even for those pollutants displaying EKC characteristics, aggregate global emissions are projected to rise over time, demonstrating that the existence of an EKC relationship for such pollutants does not imply necessarily that, at the global level, any associated environmental damage is likely to disappear with economic growth. Therefore, the EKC literature should not be taken to imply that economic growth on its own will foster environmental improvement automatically. As Panayotou (1997) concluded, "when all effects are considered, the relationship between growth and the environment turns out to be much more complex with wide scope for active policy intervention to bring about more desirable, and in the presence of market failures, more efficient economic and environmental outcomes." Or, as Arrow et al. (1995) have succinctly put it, "economic growth is not a panacea for environmental quality; indeed it is not even the main issue."

Endogenous Growth, Technological Innovation, and Resource Dependency

Other literature with important implications for sustainable development looks at *endogenous growth*, *technical innovation*, and *resource dependency*. A key feature of endogenous growth models is that technological innovation, that is, the development of new technological ideas or designs, is *endogenously*

determined by private and public sector choices within the economic system (hence the term 'endogenous' growth) rather than outside (i.e., 'exogenous') the system (Romer, 1990). This implies that the level of technology in an 'endogenous growth' economy can be advanced perpetually through public and private investments such as research and development expenditures and increases in human capital skills. Since *technical innovation* adds to physical capital already used in production, the economy can potentially 'sustain' growth rates indefinitely. Furthermore, as the economy develops, its growth rate should eventually diminish until it reaches its long-run growth rate (Barro and Sala-I-Martin, 2004).

This debate over the role of innovation in economic growth has led to empirical studies across countries and regions in the factors underlying long-term economic growth. Empirical evidence seems to support the notion that rich countries have higher rates of savings or more 'effective' institutions and implies that persistently bad government policies and ineffective institutions are inhibiting the long-run growth prospects of low-income countries. In addition, poor countries may fail to achieve higher rates of growth because they fail to generate or use new technological ideas to reap greater economic opportunities.

Institutional and policy failures in poor economies are important determinants of their inability to innovate sufficiently to achieve higher long-term growth rates. Another important factor may be the structural economic dependence of these economies on their natural resource endowments. Explanations as to why *resource dependence* may be a factor in influencing economic growth point to a number of possible fundamental linkages between environment, innovation, and long-term growth relevant to poor economies. Previous studies of an economy's dependence on natural resources and its long-run growth (Dasgupta and Heal, 1979; Stiglitz, 1974) have been extended by Barbier (1999) to include an endogenous-growth economy in order to determine whether an economy's dependence on an exhaustible natural resource is necessarily a binding constraint on long-term growth. The results of the analysis show that sufficient allocation of human capital to innovation will ensure that in the long-run resource exhaustion can be postponed indefinitely, and the possibility exists for sustained long-term per capita consumption. However, when negative feedback effects between the rate of resource utilization and innovation exist, the conditions under which it is possible to sustain per capita consumption over the long run are much more stringent and again assume that the economy is able to build up sufficient human capital and to innovate.

Implications of Endogenous Growth Theory, Technical Innovation, and Resource Dependency for Sustainable Development

Endogenous growth theory offers some support for the weak sustainability view. That is, provided that obstacles such as persistent policy distortions, political instability, and institutional failures can be overcome, any economy – even a low-income and resource-dependent developing country – should be able to foster endogenous innovation to substitute human and physical capital for a declining natural capital base in order to sustain economic opportunities and welfare indefinitely. The key appears to be developing effective policies and institutions, including the necessary public and private investments to enhance research and development and human capital skills.

Unfortunately, in most low-income countries current economic policies and investments in agriculture, forestry, and other resource-based sectors have led to rapid changes – frequently with adverse economic consequences – in resource stocks and patterns of use. Demographic trends have often worsened the relationship between population and resource-carrying capacity in many regions, and chronic environmental degradation may itself be contributing to social conflict and political instability, thus undermining the social and economic conditions necessary to foster long-term innovation and development (Homer-Dixon, 1995). In addition, the type of 'resource-saving' innovations envisioned in most endogenous growth models is likely to be technologies to abate pollution and other waste products and to conserve the use of raw material and energy inputs. However, many ecological resources and services, such as biodiversity, amenity services, and ecological functions, are less amenable to substitution by the conventional resource-saving innovations developed in the economic process. Thus, innovations are less likely to lead to opportunities for the substitution of the more unique, and possibly essential, range of ecological resources and services, including the general life-support functioning and resilience of ecosystems.

Policy Issues and Options for Sustainable Development

This section discusses the policy issues and options for sustainable development. There are some important areas of common ground for policy makers that make good progress toward sustainable development, regardless of the viewpoint on strong and weak sustainability. For example, as there exists considerable scope for substituting human and physical capital for natural capital through conservation and substitution of resource inputs, reducing pollution, and improving environmental protection, then designing policies for sustainable development must begin with these basic objectives in mind. To do this requires progress in four areas.

First, it is important to improve existing efforts to measure the economic, social, and ecological consequences of natural capital depletion and degradation. That is, in order to consider how well an economy is substituting physical and human capital for natural capital, it is necessary to measure how much the value of the natural capital stock has been 'depreciated' through current efforts to use environmental resources and services to promote greater levels of consumption and investment in the economy. Determining the 'sustainable' level of income and consumption, that is, accounting for the increases in the total available goods and services in the economy *net* of the depreciation in the value of the natural capital stock, is an important economic approach to measuring sustainability. Equally, determining the ecological 'limits' to exploitation of the natural capital stock, including ecological thresholds; carrying capacities; and key ecological goods, services, and

natural environments that might be considered 'essential,' is important ecological contribution to measuring sustainability.

Second, many of the resources and services of our natural capital endowment are either 'unpriced,' meaning that unlike other goods and services in our economy there are no markets for them, or 'underpriced,' meaning that where markets do exist for some environmental goods and services, the prices in these markets do not necessarily reflect their true contribution to human welfare. As long as environmental goods and services continue to be 'unpriced' and 'underpriced,' then their contribution to human welfare relative to other goods and services in the economy is effectively 'undervalued.' Thus, in order to promote more effective policies for sustainable development, it is necessary to adequately capture the economic values for environmental goods and services and to develop suitable tools for assessing these values.

Third, determining an appropriate 'mix' of human, reproducible, and natural capital to ensure sustainable development is ultimately about designing appropriate incentives, institutions, and investments for efficient and sustainable management of the natural environment. In order to promote sustainable development, it is important to determine the causes of environmental degradation, particularly where the failures of institutions, markets, and government policy are the main contributing factors, and to demonstrate how correcting these failures can lead to improved incentives and investments for efficient and sustainable management of natural capital (*see* Environmental Policy: Protection and Regulation).

Finally, in order to make genuine progress toward sustainable development, it is necessary to foster collaboration across the social and natural sciences, such as economics and ecology, to attain the common goal of sustainable development.

See also: Sustainable Agriculture.

Bibliography

Arrow, K.J., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C.S., Jansson, B.O., Levin, S., Mäler, K.G., Perrings, C.A., Pimentel, D., 1995. Economic growth, carrying capacity, and the environment. Science 268, 520–521.

Arrow, K.J., Dasgupta, P., Goulder, L.H., Mumford, K.J., Oleson, K., 2012. Sustainability and the measurement of wealth. Environment and Development Economics 17, 317–353.

Barbier, E.B., 1989. Economics, Natural Resource Scarcity and Development: Conventional and Alternative Views. Earthscan Publications, London.

Barbier, E.B., 1999. Endogenous growth and natural resource scarcity. Environmental and Resource Economics 14 (1), 51–74.

Barbier, E.B., 2011. Capitalizing on Nature: Ecosystems as Natural Assets. Cambridge University Press, Cambridge.

Barbier, E.B., Burgess, J.C., Folke, C., 1994. Paradise Lost? The Ecological Economics of Biodiversity. Earthscan, Publications, London.

Barbier, E.B., Markandya, A., 2012. A New Blueprint for a Green Economy. Routledge/Taylor & Francis, London.

Barro, R.J., Sala-I-Martin, X., 2004. Economic Growth, second ed. McGraw Hill, New York.

Beckerman, W., 1992. Economic growth and the environment: whose growth? World Development 20, 481–496.

Bishop, R.C., 1993. Economic efficiency, sustainability and biodiversity. Ambio 22 (2–3), 69–73.

Daily, G.C., Ehrlich, P.R., 1992. Population, sustainability and earth's carrying capacity. BioScience 42, 761–771.

Daly, H.E., 1987. The economic growth debate: what some economists have learned but many have not. Journal of Environmental Economics and Management 14 (4), 323–336.

Daily, G.C., Söderqvist, T., Aniyar, S., Arrow, K.J., Dasgupta, P.S., Ehrlich, P.R., Folke, C., Jansson, A., Jansson, B.-O., Kautsky, N., Levin, S., Lubchenco, J., Mäler, K.-G., Simpson, D., Starrett, D., Tilman, D., Walker, B., 2000. The value of nature and the nature of value. Science 289, 395–396.

Dasgupta, P.S., 2008. Nature in economics. Environmental and Resource Economics 39, 1–7.

Dasgupta, P.S., Heal, G.E., 1979. The Economics of Exhaustible Resources. Cambridge University Press, Cambridge, UK.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., Holling, C.S., 2004. Regime shifts, resilience, and biodiversity in ecosystem management. Annual Review of Ecology, Evolution, and Systematics 35, 557–581.

Grossman, G.M., Kreuger, A.B., 1995. Economic growth and the environment. Quarterly Journal of Economics 110 (2), 353–377.

Holling, C.S., 1986. Resilience of ecosystem: local surprise and global change. In: Clark, E.C., Munn, R.E. (Eds.), Sustainable Development of the Biosphere. Cambridge University Press, Cambridge, UK.

Homer-Dixon, T., 1995. The ingenuity gap: Can poor countries adapt to resource scarcity. Population and Development Review 21, 587–588.

Kuznets, S., 1955. Economic growth and income inequality. American Economic Review 49, 1–28.

Levin, S.A., Lubchenco, J., 2008. Resilience, robustness, and marine ecosystem-based management. BioScience 58, 27–32.

Meadows, D.H., Meadows, D.L., Randers, J., Behrens, W., 1972. The Limits to Growth. Universe Books, New York.

Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC.

National Research Council Board on Sustainable Development, Policy Division, 1999. Our Common Journey: A Transition towards Sustainability. National Academy Press, Washington, DC.

Neumayer, E., 2010. Weak versus Strong Sustainability: Exploring the Limits of Two Opposing Paradigms, third ed. Edward Elgar, Cheltenham, UK.

Panayotou, T., 1997. Demystifying the environmental Kuznets curve: turning a black box into a policy tool. Special Issue on Environmental Kuznets Curves, Environment and Development Economics 2 (4), 465–484.

Pearce, D.W., Markandya, A., Barbier, E.B., 1989. Blueprint for a Green Economy. Earthscan Publications, London.

Pezzey, J., 1989. Economic Analysis of Sustainable Growth and Sustainable Development Environment Department Working Paper No. 15. World Bank, Washington, DC.

Romer, P., 1990. Endogenous technological change. Journal of Political Economy 98, S71–S102.

Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B., 2001. Catastrophic shifts in ecosystems. Nature 413, 591–596.

Shafik, N., 1994. Economic development and environmental quality: an econometric analysis. Oxford Economic Papers 46 (Suppl. S), 757–773.

Stern, D.I., Common, M.S., Barbier, E.B., 1996. Economic growth and environmental degradation: the environmental Kuznets curve and sustainable development. World Development 24 (7), 1151–1160.

Stiglitz, J.E., 1974. Growth with exhaustible natural resources: efficient and optimal growth paths. Review of Economic Studies. Symposium on the Economics of Exhaustible Resources, 123–138.

Turner, R.K., 1993. Sustainability principles and practice. In: Turner, R.K. (Ed.), Sustainable Environmental Management: Principles and Practice, second ed. Belhaven Press, London, pp. 3–36.

United Nations University (UNU)-International Human Dimensions Programme (IHDP) on Global Environmental Change and United Nations Environment Programme (UNEP), 2012. Inclusive Wealth Report 2012. Measuring Progress toward Sustainability. Cambridge University Press, Cambridge.

WCED (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, Oxford.

Sustainable Lifestyles

Stewart Barr, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sustainable lifestyles is a widely recognized but often contested description of the ways in which citizens/consumers conduct conscious or habitual practices that have a benefit to the local or global environment. Often utilized as a policy instrument, the term has been applied widely in countries like the United Kingdom through top-down behavioral change programs. The understanding of sustainable lifestyles as a tool for governing the environment, therefore, is critical to an appreciation of how states seek to deploy environmental policy through engaging what we will call citizen/consumers. The outworking of these policy goals is manifested in a range of approaches to encouraging sustainable living, such as individual contributions to energy saving, water conservation, environmentally friendly travel practices, and sustainable waste management. Indeed, the range of practices we can describe as 'sustainable' is also reflected in the variety of intellectual approaches within academic research, focusing on the psychological, sociological, and pragmatic components of behavioral change. This article will explore these approaches in turn, arguing that the challenge for social science is to provide a more ambitious vision of what a sustainable lifestyle might look like, which draws more on the social than the individual as a frame of reference.

Introduction

Writing an article about such a common term as 'sustainable lifestyles' might seem one of the simpler tasks for an author in an encyclopedia on social and behavioral sciences, yet this broad term is one that comprises a range of definitional, political, and intellectual challenges. Even if the notion of sustainability is put to one side for a moment, the term itself is one that comes with a great deal of political baggage because of the ways in which nation-states, particularly in Western capitalist economies, have attempted to draw more and more upon behavioral change as a mechanism for delivering policy goals formerly ascribed to the central state. There are important questions that arise about why individuals have become the focus of policy making on the environment and why, given the challenge posed by anthropogenic climate change, it should be individuals who bear the responsibility for tackling carbon emissions. Indeed, these often poorly addressed questions are frequently accompanied by a simplistic view of behavioral change, and thus sustainable lifestyles, through modes of policy making and communication that fail to draw on the considerable intellectual insights gained from the social and behavioral sciences. Psychology has added much to our understanding of so-called proenvironmental behavior since the late 1960s and has recently been joined by sociology, bringing insights into the contextual and situated nature of sustainable lifestyles. Indeed, one consequence of the increased policy attention being focused on behavior change is the emergence of a further strand of research (largely an extension of the psychological approach), which places emphasis on pragmatic change programs in community settings.

These intellectual contributions mean that an academic understanding of sustainable lifestyles is complex and that the translation of academic research into policy making can often be problematic, given the epistemological assumptions underlying much of the research within different disciplines. Such research is often focused on contemporary framings of environmental issues, like anthropogenic climate change, and perhaps lacks the ambitiousness of research that asks a much needed set of questions about proactive social transformation in the light of environmental change, rather than mitigative and reactionary approaches. Indeed, the intellectual trajectory of research on the notion of sustainable lifestyles has largely been a Western pursuit, dominated by a discourse of maintaining current standards of living and economic growth, and disenfranchising other (and perhaps innovative) ways of living sustainably that have emerged in developing world contexts. This article therefore explores these key themes and challenges, focusing initially on the political framing of sustainable lifestyles, before examining the three major intellectual approaches in detail and concluding with a discussion on the future of research in the field.

Governing Sustainable Lifestyles

A brief look back in history (and quite recent history at that) reveals that our current approach to understanding and promoting environmental action is quite different from that deployed in the early years of the environmental crisis. Roberts (2011) notes how the first decade of significant environmental political debate (the 1970s) was dominated by international negotiations and a broad acceptance that global environmental problems were to be dealt with by state action through regulation and legislation. In this era of the 'big state,' governments often owned many national assets like industrial plants and large quantities of land, and were able to exercise considerable influence, frequently underlined by a strong belief that the state was best placed to deal with global problems like environmental pollution. We might rightly ask the question, therefore, of how it is the case that the predominant policy mechanism for tackling global issues like anthropogenic climate change in the twenty-first century is not state intervention, large-scale regulation or a command and control economy, but rather a reliance on citizens, as consumers, to effect the changes required.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.91052-3

The answer may be found in the underlying political and economic conditions that began to radically change from the late 1970s in Western nations. As Giddens (1991) has noted, the period from the late 1970s to the 1990s was marked by a major transformation in Western states, epitomized by the emergence of neoliberalism as a driving force in right of center politics, particularly in the United States and United Kingdom. The basic philosophy of neoliberalism can be described as

> …scepticism over the capacities of political authorities to govern everything for the best; vigilance over the attempts of political authorities to seek to govern. Its language is familiar and needs little rehearsal. Markets are to replace planning as regulators of economic activity. Those aspects of government that welfare construed as political responsibilities are, as far as possible, to be transformed into commodified forms and regulated according to market principles.
>
> Rose and Miller, 1992: 98

The impact of this type of political philosophy was a wholesale transfer of responsibilities for managing the environment (as well as other assets) to the market place. The state's role would be to provide light touch regulation and at a distance, but in essence, market mechanisms had to drive environmental management and in so doing, the priorities of environmental protection and economic growth were brought into ever sharper focus as neoliberal states considered the most effective ways of maintaining economic competitiveness and tackling environmental problems.

It was the notion of sustainability that enabled politicians to manage this increasingly difficult balancing act. Published in 1987, the United Nations Environment Program's *Our Common Future* (WCED, 1987) institutionalized the term sustainable development, defining it as

> Development that meets the needs of current generations without compromising the ability of future generations to meet their own needs.
>
> WCED, 1987: 8

The era of sustainability was formalized through the United Nations World Summit in Agenda *21* (UNCED, 1992), a prospectus and policy framework for promoting sustainable development. And so was born the current political consensus that now dominates Western discourse on the environment. In this discourse, it is possible to have both economic growth and protect the environment, through carefully balanced relationships between the state, markets, private enterprises, and, most crucially, 'citizen/consumers.'

The notion of the citizen/consumer emerged in the early 2000s as a way of recognizing the ways in which the state simultaneously encourages prescribed forms of citizenship and does so through a framework that places choice at the heart of decision-making. Clarke et al.'s (2007) authoritative study of the citizen/consumer concept emphasizes many of the inherent contradictions that lie at the heart of encouraging citizens to be environmentally responsible, while at the same time promoting such responsibility through a choice model that drives economic growth and ever greater consumption.

This apparent contradiction in the development of sustainable lifestyles policy, as it has been framed in countries like the Untied States and the UK, has been the subject of critical writing from social scientists (Barr et al., 2011), who have argued that this particular framing of environmental responsibility and its operationalization through the consumer growth-orientated model, is a form of libertarian paternalism (Jones et al., 2011). In this way, States promote existing forms of capitalist hegemony through free-market principles and consumer 'choice,' but they also exercise power at a distance through paternalistic notions of 'right' ways of living, for example, by guiding consumers in the ways that they should consume products, use energy, save water, and recycle waste. Indeed, in recent years, the evocation of libertarian paternalism has been formalized through the publication of the highly influential book *Nudge* (Thaler and Sunstein, 2008). This book, published by two academics at the Massachusetts Institute of Technology, highlights the ways in which consumer choice can be maintained while guiding consumers to 'better' choices through alterations in what they term the choice architecture. Accordingly, a nudge can be described as

> … any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting the fruit at eye level counts as a nudge. Banning junk food does not.
>
> Thaler and Sunstein, 2008: 8

In this model, behavior change and the promotion of sustainable lifestyles is about managing choice within a consumption-based framework. As such, it has much in common with a closely aligned approach is social marketing. Defined by French et al. (2009) as the use of conventional marketing principles to promote a social good, social marketing aims to use positive messages, targeted at specific consumer segments, to promote incremental behavioral change and as Doug McKenzie-Mohr (2000) has noted, it offers practitioners a feasible means of achieving better results in programs designed to promote sustainable forms of behavior, like recycling or energy saving.

In these ways, the political notion of sustainable lifestyles is one that is largely locked down into a particular framing, based on a set of state-led prescriptions about citizenship responsibilities on the one hand, and facilitated by architectures of choice on the other. Pragmatically, sustainable lifestyles are defined by a series of consumer-based decisions. For example, the UK's Department for the Environment, Food and Rural Affairs (DEFRA) published a *Framework for Pro-environmental Behaviors* (2008) in which it advocated the following 12 types of action:

- install insulation products,
- better energy management and usage,
- install domestic microgeneration through renewables,
- increase recycling and segregation,
- waste less (food),
- more responsible water usage,
- buy/use more energy-efficient (low-carbon) vehicles,

- use car less – seek alternatives for short trips (<3 miles),
- reduce nonessential flying (short haul),
- buy energy-efficient products,
- eat food locally in season, and
- adopt a diet with lower environmental impacts.

In essence, these actions mirror what many current attempts to encourage sustainable lifestyles have become – a set of discrete and often quite inflexible prescriptions for behavioral change. As we will see at the end of this article, this approach is being challenged by some scholars, who are arguing for a greater emphasis to be placed on more ambitious and creative ways of understanding sustainable lifestyles, which challenge the existing economic hegemony.

The Psychology of Sustainable Lifestyles

If neoliberalism has framed the notion of sustainable lifestyles as a viable approach for tackling global environmental concerns, then it is the psychological approach to behavior that has characterized the intellectual tradition, which has itself become deeply implicated in the ways in which policy makers understand behavior and seek to change it (Jones et al., 2011). The tradition of psychological approaches to understanding environmental behavior (which is used interchangeably here with sustainable lifestyles) is one that started in the late 1960s with concerns over energy use, later developing to incorporate concerns over recycling and water conservation. There is also a tradition in transport psychology that has explored how we understand travel habits and behaviors (Heath and Gifford, 2002).

Psychological studies of sustainable lifestyles have tended to follow the epistemological tradition of mainstream psychology, which is largely theory-led and utilizes quantitative social science methods to develop understandings of behavior. In this tradition, it is familiar to see generic models of behavior applied to specific issues, like the prediction of participation in environmental programs. For example, models like Fishbein and Ajzen's (1975) Theory of Reasoned Action have been used to explore the role that attitudes, norms, and perceived behavioral control exert on an individual's intention to act and their reported behavior. In these models, measures of particular constructs are developed, tested, and applied to ensure item validity and are operationalized through questionnaire surveys, being analyzed using multivariate statistical techniques like regression analysis, path analysis, and factor analysis (Fishbein and Ajzen, 2009). A useful example of this type of research is Anable's (2005) survey of tourists and their environmental attitudes, and the extent to which they were willing to reduce car driving as a way of reducing their carbon emissions. Using the Theory of Planned Behavior, Anable constructed appropriate measurement items for a questionnaire of National Trust Property visitors in the UK and utilized a factor analysis to explore the ways in which the constructs used in this model were reflected in her own data. In so doing, she was able to identify six segments of the population according to their environmental attitudes toward car driving and to use the Theory of Planned Behavior to suggest the reasons for variation in each segment.

A second strand of research in this area draws on a range of empirical insights from social and environmental psychology to develop a series of frameworks for understanding proenvironmental behavior. Since the early 1970s, psychologists have attempted to correlate a large number of variables to proenvironmental behavior and these have often been focused around the sociodemographic composition of populations. As documented in seminal works, Hines et al. (1987), Schultz et al. (1995), and Bamberg and Möser (2007) have identified, for example, that those with greater levels of environmental consciousness and reported proenvironmental behavior tend to be younger, female, have more formal education, be on higher incomes, and come from a politically liberal constituency. Indeed, meta-analyses of this type of research reveal the tendency for environmentalism to be a largely middle-class activity (although this is based on the explicit link between stated environmental attitudes and corresponding behaviors, and says little about actual environmental impact). As Barr and Gilg (2006) have outlined, a broader reading of the psychological literature on proenvironmental behaviors indicates that the correlation of demographic factors to behaviors relates to a set of situational factors that influence proenvironmental behaviors, such as environmental and behavior-specific knowledge, past experience of engaging in sustainable behaviors, and access to the resources and facilities that promotes or prevents participation. Within the household context, Barr and Gilg (2006) highlight, for example, the role of access to curbside recycling facilities as a major situational factor influencing participation in recycling.

Two other broad categories of influences on proenvironmental behavior can be identified from this psychological tradition. As Barr and Gilg (2006) argue, both social and environmental values and psychological variables are of importance. In the first instance, the seminal work of Stern et al. (1995) has demonstrated the role of social values in relation to proenvironmental behavior, not least the tendency for those with altruistic values and with an openness to change to be more likely to adopt proenvironmental practices (as opposed to those with egoistic and conservative values). Indeed, in describing specific values toward the environment, Dunlap et al.'s (2000) New Ecological Paradigm highlights the ways in which those individuals with ecocentric and biospheric values (e.g., human equality with nature, acceptance of environmental limits) tend to have more proenvironmental attitudes and behaviors.

Alongside the role of values, psychological factors are those variables, which are comprised of individual psychological attributes that have been found to influence an intention to act in a proenvironmental manner. These are many and varied and include the role of social and subjective norms (the influence of others and the acceptance of this by individuals), self-efficacy (the belief in one's ability to act), response efficacy (the belief in one's actions to have the desired effect), perceived threat of the problem at hand, and the perception that the behavior is convenient and easy to undertake. As Schultz et al. (1995) note, the 'list' of factors is seemingly every growing and one challenge for this broad strand of psychological inquiry is to explore how there can be conceptual integrity with such a large number of variables.

Accordingly, psychological approaches to understanding sustainable lifestyles have tended to focus on theory-led or theory-informed empirical studies of quantitatively measured constructs of behavior, attitudes, and other variables. The popularity of this approach has been significant in influencing policy and it is to the second broad strand of research approaches that has sought to apply this approach in a policy setting that this article now turns.

Policy Approaches for Promoting Sustainable Lifestyles

As Jones et al. (2011) point out, in countries like the UK, the psychological approach has been favored above others in its adoption as a pragmatic tool for creating and implementing policy. Often known as behavioral economics because of the dominance of economics as a policy-determining discipline in government, the psychological approach has offered a reassumingly 'scientific' basis for exploring and understanding behavior, underlain as it is by logical positivist assumptions, quantitative methods, and statistical rigor. Indeed, psychological approaches have been translated from theoretical ideas into apparently workable policy tools through the twin approaches outlined earlier in this article (Nudge and social marketing) and have been enthusiastically adopted by government departments and agencies seeking to apply psychological theories for policy making. Although many agencies have been keen to adopt the principles of choice architecture and social marketing (see the UK's National Social Marketing Centre as an example; NSMC, 2013), one of the best examples of how social marketing has been incorporated into government policy making in a country like the UK is DEFRA's (2008) *Framework for Pro-environmental Behaviors*. Published at a time when behavioral change was beginning to make a major impact in the UK central government policy making, this document highlights the influence and role of social marketing as a technique for guiding policy. Within the document, key behavioral goals are outlined (listed earlier in this article) and seven core segments of the UK's population are identified according to their (quantitatively measured) environmental attitudes and reported behaviors. As the document outlines, policy for promoting sustainable lifestyles rests on determining the right marketing mix for promoting specific changes in individual behavior through engagement with particular segments of the population. Critically, these behavioral changes are framed within the context of making 'better' choices, a clear example of Jones et al.'s (2011) argument concerning the role of libertarian paternalism in behavioral change policies.

Accordingly, the translation of psychological insights from quantitative research into policy making and program implementation has been achieved through a high level engagement of politicians and civil servants with the notion of individualized, quantifiable, and theory-led understandings of behavior, to the extent that the UK government currently has a Behavioral Insights Team (2013) at the very center of government, adjacent to the Prime Minister's Office.

Recognizing this particular framing of sustainable lifestyles is important because it illustrates the links between behavioral science and policy making that have come to dominate Western approaches to understanding and promoting sustainable lifestyles. Yet these particular understandings do have their critics and it is to a different version of sustainable lifestyles that this article now turns, initially by outlining the third intellectual approach to behavioral change and then to the broader issue of how we might envision sustainable lifestyles for the future.

Social Practices and Sustainable Lifestyles

In framing sustainable lifestyles as a set of discrete and quantifiable behaviors acted at the individual level, there is a danger that the broader context of such academic and policy initiatives is missed. Sociologist Elizabeth Shove (2010) has strongly argued for social science to adopt a different approach to that adopted by psychologists and the majority of the policy making community by exploring the following three key issues: the role of 'environment' in understanding sustainable lifestyles; the importance of the individual as the scale of reference; and finally the mode of research and analysis as a way of uncovering social practices.

In the first instance, Shove (2010) argues that in focusing the attention of analysis on specific environmental concerns (for example, focusing on changing energy behaviors as a way of tackling carbon emissions) researchers fail to recognize the complexity of 'social practices,' which

> … are conceived as being routine-driven, everyday activities situated in time and space and shared by groups of people as part of their everyday life.
>
> Verbeek and Mommaas, 2008: 634

Accordingly, Shove argues that to understand the environmental implications of behaviors, researchers need to explore the underlying practices that relate to how people live their everyday lives. Using the example of energy behaviors cited above, what is of interest is not a specific behavior that might have an impact on energy use and carbon emissions, but rather the social practices that lead to ways of living that result in high carbon emissions. Such an analysis requires a deep understanding of how social practices have formed over time and what broader social contexts come to frame such practices.

This leads to a second concern, which is a need to move beyond the individual as the frame of reference, to look at how individual behaviors are reflections of wider social practices, which are connected through shared understandings of what constitutes normal or everyday practice. A good example of this might be the fundamental changes we have witnessed in water use in developed nations. Practice researchers would not necessarily start with an isolated analysis of individual water use behaviors nor indeed the environmental attitudes of individuals, but would rather seek to appreciate how it is the case that in recent years the amount of personal water use has increased. The questions that such researchers would ask relate to how water use practices have changed over time, how common norms of water use have altered across populations, and what drivers can be identified for such changes, such as

shifting notions of what it means to be clean and to experience comfort from water-based practices. Inevitably, this takes the scale of analysis away from the individual and toward the social and cultural, emphasizing the role of changing architectures and contexts that surround individuals.

A third and final argument advocated by social practice researchers relates to the ways in which scholars explore these issues. Whereas psychological and policy-orientated approaches place emphasis on quantification and statistical representativeness, social practice researchers focus on the role of context, situation, and do so through the use of qualitative methods. Although there is variation in the types of qualitative methods adopted, popular techniques include participant observation, in-depth interviews, focus groups, and other forms of ethnography. Critically, to understand social practices, many researchers have chosen not to begin with a theoretical framework, but instead to use a grounded theory approach, which starts by accumulating significant amounts of data and information from participants and builds a situated theory from these data.

In this way, a social practices approach acts as a counter (and often critic) to the psychological and dominant policy-making approaches. Advocates often contest the logical positivist assumptions of the psychological perspective, arguing that this encourages policy makers to prefer rationalistic, theory-driven, and linear models of behavior that do not reflect social reality (Owens, 2000).

Social Transformation and Sustainable Lifestyles

Thus far this article has explored the current state of social scientific understandings of so-called sustainable lifestyles and has emphasized both the embedding of this term in contemporary forms of political economy and the ways in which policy makers have adopted particular definitions and methods for understanding and promoting sustainable lifestyles. The sociological influence over debates on such research and the advancement of thinking on social practices has, to a certain extent, provided a counter to this intellectual dominance of psychology. However, to really understand sustainable lifestyles there is a need to explore what this notion might look like if we extracted it from contemporary policy debates on incremental behavioral change, dominated by the 'small change' agenda of recycling habits, energy conservation, and minor changes in travel practices. Rather, can we be more ambitious in our understanding and promotion of living more sustainably?

Crompton and Thøgersen (2009) have highlighted that if all we ever advocate is small change, the outcome will be a proportionally small change for the environment. They argue for a more radical rethink of how our economic system and associated social practices relate to the environment and what changes in society will be needed to radically reduce carbon emissions associated with consumption and, ultimately, to promote ways of living that enable humans to live with less, not more. This is an argument, which has been pursued by a number of social scientists, most passionately by Shove (2010) and Dobson (2010). Although they focus on different components, their approach is to argue that the high carbon lifestyles that most people in the Western world lead are the direct result of a largely unquestioned and seemingly unchangeable set of economic assumptions, grounded in the argument that economic growth is the only way to develop and prosper. It is not until these basic assumptions are unpacked and changed that shifts in social practices will be achievable that will offer the prospect of tackling global climate change. In this sense, the argument becomes one not about the specific of sustainable lifestyles, but rather one about economic assumptions, contemporary living arrangements, and the very basis of what most people in the West consider it means to live a 'good life.'

In challenging researchers and policy makers to reconsider sustainable lifestyles in this radicalized context, advocates of social transformation argue that such changes are not instrumental but necessary for encouraging greater social equity and for dealing with an emergent, albeit often dismissed, second global emergency, that of resource scarcity. Epitomizing what it might come to mean to live a 'sustainable lifestyle,' authors like Kunstler (2005) argue that with the onset of oil price rises, a peak in economically viable oil production and shortages of other energy sources, individuals and communities will have to reengage with lost practices of sustainable living that reconnect us with nature and different forms of place-based community.

Although more radical changes in sustainable lifestyles as advocated by these authors are certainly off the policy agenda at present, they merit attention because they have been fundamental in framing local and growing community responses to environmental change and resource scarcity in many nations. As Hopkins (2008) points out, a transition to a low-carbon, low-oil-dependent society will require major social and economic change and this has been reflected in the growth of the Transition Movement (Transition Network, 2013), which advocates place-based, local strategies for living sustainably and for thinking about what a quality of life without abundant cheap energy might look like. This is likely to involve major changes in all parts of our lives, from our personal mobility (working and socializing close to home) to our shopping practices (buying local, seasonal produce) and to changes in how we view 'work' (with manual occupations like food growing and mechanical trades becoming more important once again). In sum, this important and emerging strand of sustainable lifestyles research argues that our approach to behavioral change will need to be much more radical to deal with the challenges of a climate-changed and resource scarce world.

Conclusions

Sustainable lifestyles is a deceptive term; it appears to be something we might all be able to describe, but in fact it is one that is politically implicated, intellectually complex, and definitionally problematic. Indeed, it is a term that is likely to mean something quite different in a decade's time. From an academic perspective, the intellectual schism between psychological and sociological approaches is problematic and is likely to endure for as long as one approach is favored over another from a policy perspective. Indeed, because of the dominance in policy making of quantifiable, 'scientific' approaches, there is a sense that the notion of sustainable lifestyles has effectively become politicized and narrowly defined as a relatively

unambitious set of behaviors that reflect the political priorities of the day.

Rather, this article has tried to demonstrate that we need to think about sustainable lifestyles as a problematic term, poorly defined and politically loaded. If we can accept this, then we must also move on intellectually to more fertile ground. We must consider what a sustainable lifestyle might look like in times of scarce resources and a world changed by anthropogenic global warming. This requires a much more ambitious social science, one that challenges economic hegemony in the name of promoting social and environmental justice and a role for citizens that is democratic and driven by the principles of participation and localized decision-making. Placed in this context, the achievement of sustainable lifestyles is a long way off.

See also: Energy and Society; Environmental Cognition, Perception, and Attitudes; Environmental Sociology; Sustainable Practices.

Bibliography

Anable, J., 2005. 'Complacent car addicts' or 'aspiring environmentalists'? Identifying travel behaviour segments using attitude theory. Transport Policy 12, 65–78.

Bamberg, S., Möser, G., 2007. Twenty years after Hines, Hungerford, and Tomera: a new meta-analysis of psycho-social determinants of pro-environmental behaviour. Journal of Environmental Psychology 27, 14–25.

Barr, S.W., Gilg, A.W., 2006. Sustainable lifestyles: framing environmental action in and around the home. Geoforum 37, 906–920.

Barr, S.W., Gilg, A.W., Shaw, G., 2011. Citizens, consumers and sustainability: (re) framing environmental practice in an age of climate change. Global Environmental Change 21, 1224–1233.

Cabinet Office, 2013. Behavioral Insights Team. https://www.gov.uk/government/organisations/behavioural-insights-team (accessed 25.11.13.).

Clarke, J., Newman, J., Smith, N., Vidler, E., Westmarland, L., 2007. Creating Citizen—Consumers; Changing Publics and Changing Public Services. Sage, London.

Crompton, T., Thøgersen, J., 2009. Simple and Painless? The Limitations of Spillover in Environmental Campaigning. WWF UK, London.

Department of the Environment Food and Rural Affairs (DEFRA), 2008. Framework for Environmental Behaviours. DEFRA, London.

Dobson, A., 2010. Environmental Citizenship and Pro-Environmental Behaviour: Rapid Research and Evidence Review. Sustainable Development research network, London.

Dunlap, R.E., Van Liere, K.D., Mertig, A.G., Jones, R.E., 2000. Measuring endorsement of the new ecological paradigm: a revised NEP scale. Journal of Social Issues 56, 425–442.

Fishbein, M., Ajzen, I., 1975. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Addison-Wesley, Reading, MA.

Fishbein, M., Ajzen, I., 2009. Predicting and Changing Behavior: The Reasoned Action Approach. Psychology Press, Oxford.

French, J., Blair-Stevens, C., McVey, D., Merritt, R., 2009. Social Marketing and Public Health: Theory and Practice. Oxford University Press, Oxford.

Giddens, A., 1991. Modernity and Self-identity. Polity Press, Cambridge.

Heath, Y., Gifford, R., 2002. Extending the theory of planned behavior: predicting the use of public transportation. Journal of Applied Social Psychology 32, 2154–2189.

Hines, J.M., Hungerford, H.R., Tomera, A.N., 1987. Analysis and synthesis of research on responsible environmental behavior: a meta analysis. Journal of Environmental Education 18, 1–8.

Hopkins, R., 2008. The Transition Handbook: From Oil Dependency to Local Resilience. Green Books, Dartington.

Jones, R., Pykett, J., Whitehead, M., 2011. Governing temptation: changing behaviour in an age of libertarian paternalism. Progress in Human Geography 35, 483–501.

Kunstler, J.H., 2005. The Long Emergency: Surviving the Convergent Catastrophes of the 21st Century. Simon and Schuster, London.

McKenzie-Mohr, D., 2000. New ways to promote proenvironmental behavior: promoting sustainable behavior: an introduction to community-based social marketing. Journal of Social Issues 56, 543–554.

National Social Marketing Centre, 2013. National Social Marketing Centre. http://www.nsmcentre.org.uk/ (accessed 25.11.13.).

Owens, S., 2000. Engaging the public: information and deliberation in environmental policy. Environment and Planning A 32, 1141–1148.

Roberts, J., 2011. Environmental Policy. Routledge, London.

Rose, N., Miller, P., 1992. Political power and the state: problematics of government. British Journal of Sociology 43, 173–205.

Schultz, P.W.S., Oskamp, S., Mainieri, T., 1995. Who recycles and when? A review of personal and situational factors. Journal of Environmental Psychology 15, 105–121.

Shove, E., 2003. Comfort Cleanliness and Convenience: The Social Organization of Normality. Berg, Oxford.

Shove, E., 2010. Beyond the ABC: climate change policy and theories of social change. Environment and Planning A 42, 1273–1285.

Stern, P.C., Dietz, T., Guagnano, G.A., 1995. The new ecological paradigm in social-psychological context. Environment and Behavior 27, 723–743.

Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving Decisions about Health, Wealth and Happiness. Yale University Press, New Haven, CT.

Transition Network, 2013. Transition Network Website. http://www.transitionnetwork.org (accessed 25.11.13.).

United Nations Conference on Environment and Development (UNCED), 1992. Agenda 21-Action Plan for the Next Century. Endorsed at UNCED. UNCED, Rio de Janeiro.

Verbeek, D., Mommaas, H., 2008. Transitions to sustainable tourism mobility: the social practices approach. Journal of Sustainable Tourism 16, 629–644.

World Commission on Environment and Development (WCED), 1987. Our Common Future: The World Commission on Environment and Development. Oxford University Press, Oxford.

Sustainable Practices

Michael Jonas and Beate Littig, IHS, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article gives an overview of the 'practical turn' in the social sciences and its reception within sustainable development research and debates more generally, and studies of (un)sustainable consumption in particular. It offers a critical review of opportunities and limits of practice research on sustainability issues and outlines potential further developments in this field.

Introduction

Practice theories are at the cutting edge of sustainable consumption research and related policies. They claim to offer an alternative route to conceptualizing the dynamics of social change by addressing both the continuity and innovativeness of human behavior. Much practice-theoretical research focuses on daily routines, that is, the everyday doings of people that are carried out with the aid of objects, such as tools, instruments, or the surrounding infrastructure. Since practices involve both physical and verbal activities, they need to be investigated using a mix of research methods, including interviews, observations, document analysis, and statistical data. Practice theories draw attention to core issues that frequently remain unaddressed by behavioral approaches, such as the role of large-scale infrastructure or material objects as participants in everyday life. When it comes to behavioral change toward a more sustainable way of living, practice theories pose the critical question whether it is actually possible to steer or govern social practices and, if so, how this can be done.

This article seeks to outline the main characteristics of practice theories before linking them to sustainability issues. The main focus will be on current sustainable consumption research. The third section discusses methodological issues surrounding empirical research on sustainable practices, while the concluding part outlines the main aspects of research required to foster sustainability in various fields of society.

The 'Practice Turn' and Its Consequences

The 'practice turn' was proclaimed some years ago and has since triggered considerable debate in the social sciences on how to conceptualize human behavior and the social (Schatzki et al., 2001). Despite the heterogeneity within the field of practice theory, the unifying assumption is that societal phenomena (e.g., individual behavior, collective activities, or sustainability processes) can be explained neither by methodological individualism (e.g., rational choice theory, sociological action theories) nor by structuralist or normativist approaches (Reckwitz, 2002). Instead, practices are treated as the core units of investigation and analysis. Despite the multitude of potential definitions, the view of a practice as a stable, yet adaptable nexus of doings and sayings that exists independently of the individuals who engage in it has largely established itself in the debate. More far-reaching conceptualizations view practices either as agents between individual activities and social structural moments (Giddens, 1979) or substitutes for social structures.

From an analytical perspective, it makes sense to view a practice as an organized entity on the one hand, and as the performance of such an entity on the other (Schatzki, 2002). While the former concentrates on the different aspects that constitute the organization of a practice, the latter looks at the activities carried out by the actors involved. Organizational aspects include understandings, rules, and teleoaffective structures or *leitmotifs* that constitute a practice. Understandings refer to those physical and mental skills that individuals require to engage in a practice. Rules include instructions for actions that are used either consciously or subconsciously by the actors involved. They can be implicit or explicit and range from virtually uncodifiable (e.g., rules of thumb, hints) to strongly codified (e.g., regulations, laws). A teleoaffective structure incorporates all target-oriented and affective orientations that determine which activities are suitable and appropriate enactments for a practice and which are not. The sociomaterial context in which a practice is carried out is sometimes viewed as a further organizational aspect (Shove et al., 2012). In contrast, the understanding of a practice as the performance of an activity assumes that the practices of the actors must be enacted through appropriate doings and sayings to have an effect in their societal context. Accordingly, the focus is directed toward various discursive and nondiscursive, physical, and linguistic activities which human (and in some approaches also nonhuman (Mol, 2002)) actors carry out when they engage in a practice. At the same time, there is a strong focus on the material embeddedness of these activities, for example, how a sequence of motions is carried out within a particular sociomaterial space, what artifacts and behavioral scripts are needed to do so as well as the level of involvement of other beings (Schatzki, 2010).

The sociomaterial context of practices is often (not always) understood as networks of human and nonhuman beings, organisms, and artifacts, which together with the performance of corresponding practices constitute the social site in which an individual is located. Here, individuals are viewed as carriers of social practices, a perspective that contrasts with both the consciousness philosophy foundations of (neo-)Marxist practice theories and more recent Foucaultdian subjectivity-oriented approaches. Consequentially, individual subjectivity or the adoption of a particular lifestyle is not attributed to individual performance. Instead, the focus is on how practices

and their sociomaterial contexts facilitate, encourage, shape, or prevent the formation of subjectivity or the adoption of a particular lifestyle.

There is great diversity in the field of practice theory and research. Different social science disciplines have adopted different practice-theoretical approaches. Moreover, a large number of empirical studies exists on different practices like driving (Thrift, 2004), dancing tango (Littig, 2013), enacting urban imaginaries (Jonas, 2014a), or eating (Shove and Spurling, 2013). These studies often highlight the need to combine practice-theoretical concepts with elements drawn from other areas of social scientific inquiry (such as the sociology of consumption, time-space geography, or gender studies) to produce fertile insights. On the other hand, these existing studies also reveal that individuals are involved in the performance of many different practices, which may or may not complement, overlap, or exclude each other.

As a result, the focus of both theoretical debates and empirical research is now moving toward bundles of practices, the extent to which the practices in these bundles are linked to each other, how they emerge, reproduce, and change as well as their connections with other bundles of practices and associated sociomaterial orders. As far as the types of links between them are concerned, experts differentiate between randomly ordered ensembles, networks, closely linked webs and stable complexes (Jonas, 2014b), or bundles and complexes (Shove et al. 2012). These in turn are positioned either on one level, thereby suggesting a 'flat ontology' (Marston et al., 2005), or on several societal levels (micro–meso–macro) that characterize more traditional multilevel approaches. There are also conceptual alternatives with regard to how practices are created, reproduced, and transformed. These draw on evolutionary theory, innovation theory, or cooperation and conflict sociology and address issues of overlap, competition, or relationships of dominance between practices. Practice theory approaches thus treat societal phenomena in a new way by focusing on the 'space' between individuals' activities and wider social structures, which is often ignored in more individualist or structuralist approaches.

Sustainability and Praxeological Research

The term 'sustainability' is one of those normative concepts that defies clear-cut definition. Prevailing approaches describe sustainability using a so-called three pillar model which assumes that economic development, environmental protection, and social justice and equity can be reconciled and should be given equal priority. In practice, however, economic and environmental interests are frequently placed over and above the social pillar (Littig and Grießler, 2005). The field of interdisciplinary sustainability research is highly diverse (Fahy and Rau, 2013); however, most sustainability policies have all taken advantage of the relative fuzzyness of the term sustainability and primarily propagated and promoted technological advancements ('green' technologies). Moreover, measures to change individual attitudes, behaviors, and habits have been given priority over other forms of intervention. A strong focus on responsible consumerism and corresponding choices demonstrates this. These approaches to (un)sustainable consumption build on authoritative behavior theory, adopting the simple formula that (frequently moralizing) education and instruction will change people's attitudes, which in turn transform their behavior and choices (e.g., OECD, 2013). This 'ABC logic' has recently been criticized, particularly from a practice theory perspective (Shove, 2010). Others suggest that dominant approaches to sustainability research and policy that endorse ABC thinking and that adopt phrases such as 'green economy' or 'win–win situation for the market, the state and the environment' are acquiescent to prevailing neoliberalization processes (Banerjee, 2003), neglect aspects of social inequality and justice and promote a feminization of environmental protection by shifting responsibility for sustainability efforts to private households (Littig, 2001). In this context, practice theory approaches have been recognized as a credible alternative.

As a consequence, mainstream sustainability policy is neither able to grasp the societal relationship with nature (Becker and Jahn, 1998) nor can it offer effective sustainability solutions. In practice, many sustainability strategies continue to pursue ecological modernization, above all through a greening of the economy. From a global perspective, such modernization projects are frequently exclusive and selective (Brand, 2010). Given the rapid growth in the world population, the increasing consumption of resources from the emergence of a global middle class (Brand and Wissen, 2012), the consequences of climate change (IPCC, 2007), and the unprecedented surge in both urbanization and slumification (Davis, 2006), sustainability cannot be separated from questions of social justice. In fact, as far as the industrialized nations are concerned, they require a radical change in production and consumption processes that goes far beyond the promotion of green technology advancement and individual changes in behavior and that bring about a transition between different sociometabolic regimes (*see* Social Ecology).

Empirical research based on practice-theoretical approaches has been conducted for diverse areas of consumption and different aspects of society. These studies not only illustrate the limitations of stand-alone technological solutions to sustainability problems, they also suggest that individual choices or changes in behavior are only of marginal relevance. Even though they cover a broad spectrum that extends from the exploration and transformation of nonsustainable work practices (Hargreaves, 2011), the adaption of sustainable agricultural practices and relevant research in the fields of mobility, energy use in private households, or leisure through to nutrition, most of the studies conducted to date can be assigned to the field of sustainable consumption (e.g., Shove and Spurling, 2013). The prevailing concept in sustainable consumption (*see* Sustainable Production and Consumption) is based (above all in the global north) on the notion that private consumer behavior, i.e., the purchase and consumption of market goods and services, has a significant influence on global economic, social, and environmental development. A goal of sustainable consumption is thus the development and distribution of (technologically) efficient and sufficient collective consumption patterns, requiring fundamental changes in both individuals' behavior and prevailing structures or 'systems of provision' to ultimately reduce the use of resources and energy.

From a praxeological perspective, and given the criticism directed at this prevailing understanding of sustainable consumption (Dolan, 2002), the term contains a broad range of practices in which individual actors participate to perceive, acquire, use, exploit, and dispose of not only market-brokered but also nonmarket-brokered entities without damaging themselves, their environment, or the need to satisfy future generations (Jonas, forthcoming). It becomes clear that, strictly speaking, there is no such thing as a sustainable consumption practice per se; instead, consumption is seen as a key component of many very different bundles of practices that trigger distinction effects in the social space. This notion assumes that sustainability must always be measured in relation to other aspects (e.g., power), and any assessment of the effects of the corresponding practices must give due consideration to related contexts. The activities of the individuals involved have more in common with the image of a practitioner than that of a consumer. Here, connections to lifestyle and milieu concepts prove fruitful because they emphasize the involvement of collectives of actors in practice bundles. However, corresponding studies carried out, for example, in Europe (Barr et al., 2011) suggest there is no such thing as a consistently sustainable lifestyle (i.e., sustainable way of living). Instead, nonsustainable aspects (e.g., climate-damaging flights) coexist alongside sustainable aspects (e.g., eco-vacations in faraway destinations) in the lifestyles studied and can be seen as the result of lifestyle-segment-specific practice bundles (in this case, mobility and vacation) (*see* Sustainable Lifestyles). Collectively accepted, environmentally aware rhetorics of action and the individual balancing of environmentally friendly and harmful behavior help to reduce cognitive dissonances and legitimize individual wrongdoings (Littig, 2001). If such practice bundles from different lifestyle segments are closely and complexly linked to each other, the chance of developing a consistently sustainable lifestyle becomes very slim.

Other studies offer more detailed insights into the sustainability, performance, and organization of very different practices, particularly in the above-mentioned fields of mobility (Middleton, 2011), energy use (Spinney et al., 2012), and other areas of everyday life (Gregson et al., 2009). These studies show that the performance of everyday practices, like driving or walking, eating, showering, or using a laptop, is primarily shaped by habits, routines, and day-to-day rules. Rational deliberations and 'freedom of choice' that are so popular in mainstream sustainability policies are the exception rather than the rule. They also underline the key role of different perceptory senses in carrying out practices. Moreover, affects and emotions lead people to engage in some practices and exclude others. In addition to corresponding practice-specific competences and tacit and explicit rules, teleoaffective *leitmotifs* also reproduce dominant practices, e.g., images of motorized personal transport that have established themselves in many societies in relation to driving. At the same time, they suppress the development and spread of alternative, so-called low-CO_2 practices like walking and cycling (or using public transport). More importantly, dominant practices such as car-based mobility are facilitated by corresponding sociomaterial infrastructures, the planning and implementation of which took only insufficient (if any) account of alternative, more sustainable practices such as walking and cycling, and which can often only be changed with considerable effort and at great cost.

Interestingly, everyday habits that are core elements of (un) sustainable practices prove to be far more flexible than is often assumed. Furthermore, they serve as core anchor points that supply the necessary rhythms and predictabilities of everyday life. In this connection, practices and the customs they involve are always tied to sociospatial settings. In the United States, but also in many European countries, the above-mentioned practices of walking or cycling, which are far more energy efficient than driving a car, have been "marginalized by autocentric planning, engineering, politics, and legal regimes that manifest as laws and infrastructure incompatible with active modes" (Nixon, 2012: 1664). Accordingly, initiatives to promote environmentally friendly mobility will remain of little success until such regimes are adapted to the needs of walking or cycling practices. At the same time, these studies point out that there is no one single practice of walking. Instead, walking practices vary from strolling and sauntering through to purposeful hikes. While they share a whole number of commonalities, they also have their differences. Moreover, we also have to determine the relevance of gender or milieu-specific, age-related, or other nonagent-related aspects (e.g., the weather) for such practices and ascertain whether the performing of such practices in the available or planned sociomaterial spaces are linked with positive sensory perception opportunities. Hence, infrastructural changes to facilitate more sustainable practices also need to be taken into consideration when trying to change behavior (Dangschat and Segert, 2011).

The relevance of habits for the course of societal processes currently receives renewed attention not only in practice theory based sustainability debates but also in related areas of sustainability research and practice (Warde and Southerton, 2012; Strengers and Maller, forthcoming). From a praxeological perspective, it makes sense to differentiate between habits and practices. All habits can be designated as practices, but practices are often more than just a set of habits. Many of the empirical studies mentioned above demonstrate the flexibility of practices for a context in which habits are grasped not as primarily rigid patterns of behavior but as partially flexible and creative processes that facilitate adaptation to different situations. Habits and routines are thus not opposites to reflexive action but form the necessary conditions for it. This discussion can provide important impetus for the formulation and development of promising sustainability policies because it shifts the emphasis away from trying to suppress unsustainable habits (e.g., traveling extensively by car and/or plane, maintaining an energy-intensive lifestyle) by raising awareness and increasing reflexivity. Instead, attention is drawn toward how established habits can be replaced by sustainable alternatives and for which changes need to be made in the sociomaterial context to ensure that more sustainable (bundles of) practices appeal more to their potential agents than currently established ones.

Methodology and Methods

Practice theory is not only a specific strand of social theory. As demonstrated above, many examples exist of empirical

research into sustainable practices. Interestingly, a practice theory perspective also requires a specific methodological approach which takes a 'practical turn' toward research objects, i.e., the phenomena studied using scientific practices. Praxeological analyses situate their research objects in fields of embodied, materially mediated activities and processes which are organized along collectively shared forms of practical knowledge and competences. The praxeological perspective is a particular lens on social phenomena which emphasizes the collective organization and mutual interdependence of social life. In their investigation of practices, practice theory based approaches often seek to examine the very research practices that underpin these investigations. In other words, they include 'doing science/research' aspects, i.e., the research practices in relation to the research object. This immediately raises interesting methodological questions.

Firstly, there is considerable disagreement among advocates of practice theory regarding the methods that are suitable for studying social practices. Participant observation was long held to be the primary research method. Meanwhile, other methods such as discourse analysis, qualitative interviews, and extended ethnographic research increasingly form part of the methodological repertoire. John Law, Annemarie Mol, and Marianne de Laet all use the term 'praxiography' to refer to a special research approach which stresses the need to allow the actual field to define the methods used. With analogies to Actor-Network-Theory, they view practice as 'enactment' in the sense of an interaction of human and nonhuman participants. In their empirical studies to date, they even select objects as the starting points of their research, such as the 'Zimbabwe Bush Pump' (de Lait and Mol, 2000) or the 'Cumbrian Sheep' (Law and Mol, 2008). Despite the dominance of qualitative modes of inquiry, some argue that a practice theory perspective does by no means preclude quantitative research. Nonetheless, there are currently only a few examples which demonstrate the extent to which quantitative methods can be used beyond the numeric-descriptive recording of practices (Cheng et al., 2007).

Outlook

Practice theories can be seen as an alternative approach to activity or behavioral theory approaches. Despite their breadth and heterogeneity, practice theory approaches are unified by their fundamental criticism of individualistic notions of sociality on the one hand, and structural determinism on the other. This said, sophisticated practice theory debates in the social sciences are often contrasted with a comparatively limited use of their concepts in empirical research. These empirical studies are also scattered across relatively distinct research fields that rarely interact (e.g., interpretive policy analysis, practices in and of organization, sustainability, etc.). To assess the merits of practice theory for empirical research, greater interlinkages between these disparate research fields would be expedient. Recent multidisciplinary international conferences (The Institute for Advanced Studies in Vienna hosted an international conference on practice theories and research in June 2013. Publications are planned and will be forthcoming in 2014. See: http://www.ihs.ac.at/conferences/practices/(accessed 25.09.13.).) and workshops are first steps toward achieving this (Warde and Southerton, 2012). Moreover, research into (un)sustainable practices would also profit from comparative perspective.

One key criticism directed toward social practice theorists is their apparent inability or unwillingness to suggest strategies and policies that seek to intervene in the trajectories of practices to address 'real-world' policy problems. Few published studies exist that seek to apply social practice theories in the territory of changing practices, i.e., to identify opportunities to disrupt, reorient, or redirect social practices in more sustainable directions (e.g., Shove and Spurling, 2013; Davies et al., 2014). There is clearly still a lack of research into particular questions like how to design effective political interventions against a backdrop of practice theory considerations. Last but not least, the question remains whether and to what extent the (political) steering of practices is possible given the complex linkages between different practices. Thus, more empirical praxeological experiments are needed in the future.

See also: Degrowth; Environmental Policy: Protection and Regulation; Sustainable Lifestyles; Sustainable Production and Consumption; Time Use and Resource Consumption.

Bibliography

Banerjee, S.B., 2003. Who sustains whose development? Sustainable development and the reinvention of nature. Organization Studies 24 (1), 143–180.
Barr, S., Shaw, G., Coles, T., 2011. Sustainable lifestyles: sites, practices, and policy. Environment and Planning A 43, 3011–3029.
Becker, E., Jahn, T., 1998. Growth or development? In: Keil, R., Bell, D.J., Penz, P., Fawett, L. (Eds.), Political Ecology: Global and Local. Routledge, London, pp. 68–83.
Brand, U., 2010. Sustainable development and ecological modernization – the limits to a hegemonic policy knowledge. Innovation: The European Journal of Social Science 23 (2), 135–152.
Brand, U., Wissen, M., 2012. Global environmental politics and the imperial mode of living: articulations of state-capital relations in the multiple crisis. Globalizations 9 (4), 547–560.
Cheng, S.-L., Olsen, W., Southerton, D., Warde, A., 2007. The changing practice of eating: evidence from UK time diaries, 1975 and 2000. British Journal of Sociology 58 (1), 39–61.
Dangschat, J.S., Segert, A., 2011. Nachhaltige Alltagsmobilität – soziale Ungleichheiten und Milieus. Österreichische Zeitschrift für Soziologie 36 (2), 55–73.
Davies, A., Fahy, F., Rau, H. (Eds.), 2014. Challenging Consumption: Pathways to a More Sustainable Future. Routledge, London.
Davis, M., 2006. Planet of Slums. Verso, London.
Dolan, P., 2002. The sustainability of sustainable consumption. Journal of Macromarketing 22 (2), 170–181.
Fahy, F., Rau, H. (Eds.), 2013. Methods of Sustainability Research in the Social Sciences. Sage, London.
Giddens, A., 1979. Central Problems in Social Theory – Action, Structure and Contradiction in Social Analysis. MacMillan, London.
Gregson, N., Metcalfe, A., Crewe, L., 2009. Practices of object maintenance and repair. Journal of Consumer Culture 9 (2), 248–272.
Hargreaves, T., 2011. Practice-ing behaviour change: applying social practice theory to pro-environmental behaviour change. Journal of Consumer Culture 1 (1), 79–99.
IPCC, 2007. Climate Change 2007. Synthesis Report.
Jonas, M., 28 October 2014a. About the enactment of a new economic imaginary: the dortmund case. International Journal of Urban and Regional Research, 38(6), 2123–2140.
Jonas, M., 2014b. Zur Inszenierung eines Wirtschaftsclusters – eine praxeologische Analyse. Springer VS, Wiesbaden.
Jonas, M. Nachhaltiger Konsum – eine praxeologische Kritik. In: Schäfer, H. (Ed.), Praxistheorie. Ein Forschungsprogramm. Transcript, Bielefeld, forthcoming.
de Lait, M., Mol, A., 2000. The Zimbabwe bush pump: mechanics of a fluid technology. Social Studies of Science 30 (2), 225–263.

Law, J., Mol, A., 2008. The actor-enacted: Cumbrian sheep in 2001. In: Knappett, C., Malafouris, L. (Eds.), Material Agency. Springer, New York, pp. 57–77.

Littig, B., 2001. Feminist Perspectives on Environment and Society. Pearson Education, London.

Littig, B., 2013. On high heels. A Praxeography of doing argentine tango. European Journal of Women's Studies 20 (4), 455–467.

Littig, B., Grießler, E., 2005. Social sustainability. A catchword between political pragmatism and social theory. In: Spangenberg, J., Giljum, S. (Eds.), Governance for Sustainable Development. International Journal of Environment and Sustainable Development 8 (2), 65–79 (special issue).

Marston, S., Jones III, J.P., Woodward, K., 2005. Human geography without scale. Transactions of the Institute of British Geographers 32, 416–432.

Middleton, J., 2011. "I'm on autopilot, I just follow the route": exploring the habits, routines, and decision-making practices of everyday urban mobilities. Environment and Planning A 43, 2857–2877.

Mol, A., 2002. The Body Multiple: Ontology in Medical Practice. Duke University Press, Durham, NC.

Nixon, D.V., 2012. A sense of momentum: mobility practices and dis/embodied landscapes of energy use. Environment and Planning A 44, 1661–1678.

OECD, 2013. Greening Household Behaviour: Overview from the 2011 Survey. http://www.oecd-ilibrary.org/environment/greening-household-behaviour_9789264181373-en (accessed 30.09.13.).

Reckwitz, A., 2002. Toward a theory of social practices: a development in culturalist theorizing. European Journal of Social Theory 5 (2), 243–263.

Schatzki, T., Knorr-Cetina, K., Von Savigny, E. (Eds.), 2001. The Practice Turn: Contemporary Theory. Routledge, London.

Schatzki, T., 2002. The Site of the Social. A Philosophical Account of the Constitution of Social Life and Change. Pennsylvania State University Press, State College, PA.

Schatzki, T., 2010. Materiality and social life. Nature and Culture 5 (2), 123–149.

Shove, E., 2010. Beyond the ABC: climate change policy and theories of social change. Environment and Planning A 42, 1273–1285.

Shove, E., Pantzar, M., Watson, M., 2012. The Dynamics of Social Practice: Everyday Life and How It Changes. Sage, London.

Shove, E., Spurling, N. (Eds.), 2013. Sustainable Practices – Social Theory and Climate Change. Routledge, London.

Spinney, J., Green, N., Buringham, K., Cooper, G., Uzzell, D., 2012. Are we sitting comfortably? Domestic imaginaries, laptop practices, and energy use. Environment and Planning A 44, 2629–2645.

Strengers, Y., Maller, C. (Eds.), 2014. Beyond Behaviour Change – Intervening in Social Practices for Sustainability. Routledge, London.

Thrift, N., 2004. Driving in the city. Theory, Culture & Society 21 (4/5), 41–59.

Warde, A., Southerton, D. (Eds.), 2012. The Habits of Consumption. COLLeGIUM. Studies across Disciplines in the Humanistics and Social Sciences, 12, 87–99.

Sustainable Production and Consumption

Daniel Welch, The Sustainable Consumption Institute, The University of Manchester, Manchester, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Sustainable production and consumption (SPC) is both a policy area and a multi- and interdisciplinary research agenda that seeks to address the transformation of society toward environmental and social sustainability. The SPC field encompasses diverse research topics and theoretical approaches, often characterized as falling into stronger and weaker versions, defined by commitment to more or less systemic socioeconomic change. Fault lines within the field reflect fundamental controversies within the social and behavioral sciences more broadly, such as the structure–agency antinomy. Major theoretical developments within SPC include novel conceptual frameworks for production–consumption systems and reconceptualizations of consumption.

Introduction

Sustainable production and consumption (SPC) is best understood as both a policy area and a multi- and interdisciplinary research agenda rather than a fully coherent social and behavioral scientific field. The field is also referred to as 'Sustainable Consumption and Production' (SCP), usually with a concomitant research emphasis, or simply 'sustainable consumption' (often, however, without signaling an exclusion of interest in production). The generic area is referred to as SPC throughout this article without any intended emphasis on a particular research agenda or viewpoint. Research in SPC is conducted by scholars in, primarily, economics, geography, sociology, social anthropology, and social psychology, as well as, inter alia, development studies, marketing, management studies, planning, political science, and public policy. SPC research is also to be found in the interdisciplinary areas of consumer behavior, innovation studies, and organizational studies. Research in the field often seeks to transcend narrowly disciplinary perspectives and this is reflected in inter- and multidisciplinary research programs and publications (e.g., Cohen et al., 2013; Shove and Spurling, 2013; Southerton and Ulph, 2014; Reisch and Thøgersen, 2014).

Definitions of SPC are highly contested, particularly around the manner in which SPC addresses: systemic change on a global level, the demarcation between production and consumption, and between consumption and other social activities. The scope of SPC research is therefore more useful to introduce the field. Useful distinctions can be made between research on the material aspects of SPC (commonly addressed as 'life cycle assessment' (LCA), 'material flow,' or 'environmentally extended input–output analysis' – *see* Industrial Ecology and Life Cycle Assessment), economic aspects (studied at both micro and macro levels), and psychological, social, and cultural aspects.

Technically and instrumentally oriented SPC research commonly addresses improving the environmental performance of production processes and supply chains; product, service, and system innovation for sustainability; and mechanisms for, and barriers to, making consumption more sustainable (including consumer behavior, lifestyles, and symbolic aspects of consumption). Research further addresses corollary policy areas concerning technological innovation and the management of sociotechnical transitions toward sustainability and consumer-oriented policy, as well as related areas of corporate social responsibility, such as stakeholder engagement. Conceptually oriented work in SPC theorizes relations between production and consumption, and systemic change of sociotechnical and socioeconomic systems (e.g., Tukker et al., 2008; Lebel et al., 2008).

A broad range of additional topics are found under the rubric of sustainable consumption, including critiques of consumerism and critiques of SPC policy. Conceptual work in the area theorizes consumption and its relation to wider social activity, as well as the roles and capacities of consumers, social movements, and other collective actors in promoting sustainability. Such theoretical work is commonly in dialogue with empirical studies, and vice versa. Empirical studies address, on the one hand, environmental and consumption-oriented politics, including ethical and political consumption, and the cultural dimensions of sustainable consumption, and on the other, more broadly, the dynamics of everyday practices, the evolution of sociotechnical systems, and the relation of both to transitions toward sustainability (e.g., Cohen and Murphy, 2001; Cohen et al., 2013; Jackson, 2006; Reisch and Thøgersen, 2014; Southerton et al., 2004).

Sustainable consumption is often explored through the broad domains of consumption most critical for environmental impact. There is widespread agreement that transport (or mobility), food, water, housing, and energy-using products account for 70–80% of consumption-related environmental impacts in industrialized nations when measured over the life cycle (Tukker et al., 2010: p. 17). Life cycle here refers to the production of goods and services, their consumption, and their disposal, drawing our attention to the systemic relations between production and consumption. SPC scholars thus conceptualize domains as production–consumption systems and analyze the complexities of changing such systems.

SPC has sought to develop a comprehensive view of the environmental impacts of the economy as a system and to analyze the impacts, dynamics, and potential for change of component production–consumption subsystems. This represents a shift in environmental policy and analysis from the supply-side perspective: the sole focus on producer

responsibility for environmental impacts and the modification of productive processes alone as the mechanism for achieving Sustainable Development (*see* Sustainable Development: An Economic Perspective). Consumption and demand-side management has come to be viewed as a legitimate area for environmental policy making and for social scientific research into the environmental impacts of the economy. Indeed, much of the normative controversy surrounding SPC concerns where on the continuum between production and consumption responsibility for environmental impact lies.

Approaches to SPC, and sustainable consumption especially, are often considered as falling into strong and weak programs, where the weak program reflects incremental reforms to the status quo, and the strong program radical systemic change to the economy and society (e.g., Fuchs and Lorek, 2005; Hobson, 2013). Public and corporate policy approaches, dominated by a focus on technological innovation on the production side and understandings drawn from neoclassical and behavioral economics and social psychology on the consumption side, commonly reflect the weaker program, while the stronger program is usually advocated by critical social scientists and social movements. Proponents of the strong form of SPC reflect quite divergent theoretical and political commitments and often (not always) draw on strong normative critiques of consumption. Important schools of thought include Eco-Marxism (see the journal *Capitalism Nature Socialism*) and Degrowth (*see* Degrowth), although it should be noted discourses of sustainability per se are inherently normative, given their orientation to socioeconomic change.

A common fault line between the weak and strong versions is whether sustainability can be achieved within the conventional economic growth paradigm, or whether 'degrowth,' or absolute demand reduction, is required. Those supporting the former view contend that a decoupling of economic growth from increasing environmental impact is technically feasible, while those supporting the latter view point out that while instances of relative decoupling can be observed, there is no evidence of absolute decoupling on a global level (see Harris, 2013). Associated with this division is an emphasis on the ecoefficiency of production and consumption on the one hand, and on the other, on radical changes in the patterns and levels of consumption. The weak program is criticized as approaching consumption as simply an addition to the older productivist agenda of cleaner production, regarding consumption primarily in terms of problems and opportunities in the demand for more environmentally friendly products. The position toward 'consumer sovereignty' – or the framing of individuals as autonomous consumers who express their preferences through market relations and purchasing decisions – is thus another commonly invoked fault line between the weak and strong programs. The weak program typically endorses a view of sustainable consumption as the aggregate effects of individual consumers' environmental preferences, while the strong program typically endorses a more expansive notion of 'environmental citizenship' and normative criticism of consumer society (Hobson, 2002). While critics of the weak program accuse it of collusion with the neoliberal economic order culpable for the ecological crisis, critics of the strong program accuse it of political infeasibility, or of hijacking sustainability to promote existing political commitments.

The following section will examine the origins and development of SPC as a policy area; SPC as an area of research is intimately linked with the development of international policy initiatives. The final section will examine the variety of SPC research and some key developments within the field.

The Development of SPC Policy and Research

The origins of the field lie in the growth of environmental social science (*see* Environmental Sociology) in the 1970s, following the 1972 publication of the Club of Rome report 'The Limits to Growth' and the United Nations (UN) Stockholm Conference on the Human Environment. However, SPC did not develop as a recognizable field of research until the 1990s when its emergence coincided with the strengthening of international SPC policy. Proponents of the SPC policy agenda have consistently recognized the importance of, and funded, social science research. As well as instrumental insights and analysis, social scientists have offered an ongoing critique of how the epistemological and ideological foundations of policy frames the SPC agenda and associated public discourse (e.g., Cohen, 2001; Hobson, 2002, 2013; Fuchs and Lorek, 2005).

The 'Brundtland Report' of the World Commission on Environment and Development (WCED, 1987) made an important shift in environmental policy discourse, recognizing that sustainable development (*see* Sustainable Development: An Economic Perspective)

> requires the promotion of values that encourage consumption standards that are within the bounds of the ecologically possible and to which all can reasonably aspire…Major changes in policies will be needed to cope with the industrial world's current high levels of consumption.
>
> WCED, 1987: p. 57

The WCED laid the foundations for the 1992 UN Conference on Environment and Development, or 'Rio Earth Summit,' which was the watershed for international sustainability policy. Agenda 21, the summit's voluntary action plan for the implementation of sustainable development called for, in addition to established areas of environmental policy such as pollution control and conservation, "a better understanding of the role of consumption and how to bring about more sustainable consumption patterns" (UNCSD, 1992: p. 4.23), as well as provisions to address global economic and political inequities. The previous focus of global environmental policy on developing world population growth was displaced by 'sustainable livelihoods' for the global South and 'sustainable production and consumption' for the North (UNCSD, 1992).

Another important shift came in the early 1990s with the European Commission's promotion of the 'ecological modernization' of production (see Mol et al., 2009). Ecological modernization proposed a new relationship between production, policy, and environment. The former approach of

'command and control' environmental legislation shifted to one in which economic competitiveness and growth on the one hand and environmental protection on the other were seen as mutually reinforcing rather than antagonistic. Demanding environmental standards were seen as opportunities for, and stimulators of, innovation. The minimization of waste, the increase in resource efficiency, and the maximization of output were seen to go hand in hand.

Over the course of the 1990s the conventional attribution of environmental responsibility for sustainability to producers was increasingly supplemented by a focus on the role of consumers. Consumption became an organizing theme for environmental policy making, especially in Europe, where sustainable consumption was seen as the 'downstream' complement to the ecological modernization of production. This was initially conceived primarily in terms of consumers' influence through demand for more environmentally friendly products. Early definitions of sustainable consumption, such as the 1995 formulation of the UN Commission on Sustainable Development focused on the use of goods and services that improved quality of life while minimizing environmental impacts, over their life cycle, within boundaries that did not compromise the needs of future generations (Luskin, 2007). The concept of product life cycle – from production through consumption to 'end-of-life' management – was important in foregrounding production–consumption systems as integrated wholes.

The focus on sustainable consumption was expanded in the late 1990s through initiatives such as the United Nations Environment Programme (UNEP)'s Sustainable Consumption program and the Organisation for Economic Co-operation and Development (OECD)'s Sustainable Consumption and Production program. The latter supplemented the conventional focus on sustainable products and eco-labeling by seeking to understand "both the driving forces behind specific consumption patterns, and the implications of using different policy instruments to try to change them" as well as "innovative approaches to modifying consumer behavior" (Cohen, 2001: p. 31).

A step change in activity occurred following the 2002 World Summit on Sustainable Development in Johannesburg, which initiated a 10-Year Framework of Programmes on SCP (the 'Marrakech Process'), including activities around 'life cycle thinking,' public policy tools and approaches, advertising and communications, sustainable products and services, and education and training (see Clark, 2007). Numerous national and international policy-related SPC research initiatives were launched in the aftermath of the 2002 Johannesburg summit. The 'Rio + 20' UN Summit in 2012 has adopted a further 10-year SCP framework program that reiterates the tenets of sustainable development laid down in 1992. It calls for the scaling up of SCP and resource efficiency initiatives to contribute to the decoupling of environmental impacts from economic growth while increasing social development (Hobson, 2013). Rio + 20 also saw the launch of a Global Research Forum on Sustainable Production and Consumption, which frames the research agenda in terms of "how to understand the present lock-in in unsustainable consumerist lifestyles and production patterns, and to understand how a possible transition to more sustainable production and consumption patterns and lifestyles could be accomplished" (Vergragt et al., 2014: p. 2).

Sustainable Consumption

The emphasis on consumption must also be seen in the context of the growth of post-Fordist consumer society from the 1970s onward and a concomitant cultural emphasis on consumption activities. While in developed nations CO_2 emissions from the industrial sector fell in absolute terms and there has been success in reducing pollution and waste from production, overall CO_2 emissions, energy use, and material throughput have continued to grow. Moreover, a growing proportion of these environmental impacts can be directly or indirectly related to private households (Michaelis, 2003). The rapid emergence of a global consumer class, particularly in the Asia–Pacific region, has further focused attention on the sustainable consumption agenda.

The area of sustainable consumption attained a certain degree of autonomy from the wider SPC field following 2002, marking a divergence apparent today. The view that consumption should be viewed purely as an economic activity that can be modified through top-down approaches (primarily information provision, such as ecolabeling, and market signals), and the adequacy of using consumers as the agents of upstream change to the productive sphere, began to be challenged by policy makers (Cohen, 2001: p. 34). The UNEP, for example, identified clarifying the meaning of consumption and a more appropriate conceptual schema for production–consumption as strategic areas requiring research (Fuchs and Lorek, 2005: p. 269). The OECD broadened the definition of consumption beyond market choices to use, disposal, and noncommodified goods and services (Southerton et al., 2004: p. 3). This said, mainstream policy initiatives such as the European Union Sustainable Consumption and Production Action Plan launched in 2008 often continue to be framed in terms of improving the environmental performance of products and production technologies and increasing the demand for more sustainable goods.

While the focus on developed world consumption as the central cause of the ecological crisis was welcomed by many scholars, critical social science has been skeptical of the shifting of responsibility onto consumers, of the adequacy of the characterization of human behavior that has tended to go with it, and of the ameliorative mechanisms thus proposed. This framing of the consumption issue has been consistently critiqued as serving neoliberal ideological interests and obscuring the complex relations to production and macroeconomic issues (Fuchs and Lorek, 2005; Hobson, 2002; Southerton et al., 2004). Furthermore, the divergence of sustainable consumption from the earlier sustainable development agenda has tended to see a focus on environmental impacts to the exclusion of social aspects of sustainability, such as issues of equity.

However, a strong case can be made for the focus on consumption through the importance of behavioral or systemic offsets to potential emissions savings through energy efficiency gains, known as rebound effects. Changes in patterns of consumption can offset the environmental gains achieved

through sustainability improvements in productive processes. Improved refrigeration technology, for example, may simply lead to the use of increased refrigerated space at the same cost. Or on a household level financial savings through energy efficiency may simply be recycled into goods and services with their own life-cycle impacts. Efficiencies also spur innovations in products and services that themselves increase demand. Exponential increases in efficiency in computing power, for example, have driven the development of new products and the creation of whole new markets. On the macro level this leads to the argument that increases in energy efficiency actually cause increases in aggregate energy use (Jevons Paradox, see Sorrell, 2009).

Furthermore, increased efficiency may inadvertently naturalize less sustainable needs, conventions, norms, and expectations; such as the need for Internet access to participate in modern life, or conventions of adequate room temperature (see Shove, 2003). Understanding the dynamics of changing social conventions and expectations can lead to deeper understanding of the dynamics of escalating demand and suggest possible modes of intervention.

Moreover, the case for a focus on consumption lies in the need for rapid emissions reduction to ameliorate catastrophic climate change. Short-term demand opportunities for reduction in CO_2 emissions dwarf the short-term opportunities afforded from the supply side. This is principally due to the long lead times in replacing energy infrastructure compared to the life span of consumer goods and vehicles, as well as inefficiencies in electricity supply systems favoring end-use energy reductions over supply-side savings. As the timescale for effective emission reductions has narrowed, it is widely held that timely action is impossible without radical changes to consumption patterns, however radical the changes made in the productive sphere.

Varieties of SPC Research and Key Theoretical Developments

As noted, SPC policy has been dominated by models of human behavior drawn from economics and social psychology. Both share ontological commitments to methodological individualism and a fortiori the assumption that consumption is driven by individuals' choices, which in turn reflect relatively stable preferences, attitudes, or values. Theoretical developments in the field, especially from within sociology (e.g., Southerton et al., 2004) and social anthropology (e.g., Wilk, 2002), have offered a fundamental critique of this position, emphasizing instead social and cultural processes, and the coevolution of social practices on the one hand, and technologies and infrastructures on the other. That said, there have also been important developments within psychology and economics addressing SPC.

Social psychology, and a specific branch of environmental psychology, as well as lesserly sociology, has addressed attitudes toward the environment and 'proenvironmental behavior' (*see* Environmental Cognition, Perception, and Attitudes), the 'spillover' of environmental behavior from one domain to another, and the 'value–action' or 'attitude–behavior' gap, which describes the discrepancy between reported proenvironmental values and obdurately unsustainable behavior (see Kollmus and Agyeman, 2002).

Economic psychology and behavioral economics address sociopsychological aspects of economic phenomena and processes, and have contributed significantly to sustainable consumption behavioral change policies (see Carlsson and Johansson-Stenman, 2012). Emphasizing the role of unconscious and contextual factors in individuals' behavior, this body of work questions the conventional economic notion of rational preferences, and offers insights into a range of irrational consumption behaviors (e.g., the sunk cost fallacy). Economists concerned with sustainable consumption have also provided insights into traditionally sociological areas such as the role of acquired dispositions in consumption patterns and competitive or conspicuous consumption (see Southerton and Ulph, 2014).

Environmental and ecological economics address macrolevel issues within the SPC research agenda. Ecological economics addresses the systemic relationships between economic systems and ecosystems. Environmental economics studies the economics of environmental policies, offering insights into a range of issues such as the valuation of consumption and capital assets from an environmental perspective (*see* Sustainable Development: An Economic Perspective), for example, using 'shadow prices' rather than market prices to make valuations including natural capital (see Southerton and Ulph, 2014). The emerging area of New Economics draws on these heterodox approaches to develop an alternative economic paradigm for SPC. The field addresses topics such as the drivers and effects of unsustainable consumption, the economics of well-being, grassroots socioeconomic innovations, and the macroeconomics of sustainability (such as issues of degrowth and steady state economics) (see Cohen et al., 2013).

Further major theoretical developments within SPC that have generated new questions and insights are found in the development of novel conceptual frameworks for production–consumption systems and in the reconceptualization of consumption. Conceptualizations of consumption are grounded in fundamental assumptions about human nature, connected to deep philosophical issues about the causes of human behavior (Wilk, 2002). The field thus reflects controversies within the social and behavioral sciences more broadly, and brings some of the key conceptual debates, over, for example, human agency and the relation between social and material phenomena, to the fore.

Processes of appropriating resources from the environment have for most of the modern era conventionally been ascribed to the sphere of production. Their ascription to the sphere of consumption in addition is a foundational innovation of SPC. However, consumption must be distinguished conceptually to avoid conflation with production, overall economic activity, other categories of environmental concern such as population or technology, or normative categories such as materialism (Princen, 1999).

Princen (1999) suggests that there are two fundamental approaches to the issue of sustainable consumption. The first is to work within the production–consumption, supply–demand, producer–consumer dichotomy, which does not debar examining product use and nonmarket activity. The second is to

challenge that dichotomy. Developments in the conceptual framework of SPC have navigated between these two fundamental positions. Princen (1999) acknowledges that both approaches have their limitations but argues for an ecologically grounded definition of consumption as 'material provisioning,' transcending the traditional dichotomy while maintaining the specificity of consumption as an analytical category. This is to treat all resource use (including that traditionally ascribed to production) as consumption or the using up of materials and energy in purposive human activity. Conventional approaches to SPC, such as LCA, often adopt versions of this position. Arguably, however, they have done so in a manner that lacks insight into the specific dynamics of use and demand, and consequently the importance of the latter for production–consumption systems over all.

In the first social science collection addressing sustainable consumption, Cohen and Murphy (2001), the editors, argued against Princen's position, suggesting that the physical dimension of material provisioning occludes the sociocultural dimensions of consumption. Material provisioning, they argued, ignores the phenomenology of consumption and its socially constructed aspects, including the social construction of the notion of the environment itself, and its role in policy formulation. Leaving aside whether this criticism is justified, this approach brings to the topic of sustainable consumption questions of meaning, identity, and communication in consumption associated with 'the cultural turn' in the social sciences. Cohen and Murphy (2001) here define consumption as people's acquisition (not necessarily through the market) and use of objects, services, and places, specifically excluding forms of institutional consumption (Cohen and Murphy, 2001: p. 8).

Shove and Warde (2002), however, criticized this cultural theory approach as offering limited analytical purchase to the routine, nondeliberative and 'inconspicuous' aspects of environmentally significant consumption. The reaction against the 'cultural turn' sought to move beyond the emphasis on consumer choice and changing lifestyles to a focus on social practices, the socially structured character of consumption and its mediation through infrastructures of provision, or the sociotechnical organization of consumption (see Southerton et al., 2004).

Warde reconceptualized consumption as "not itself a practice but…rather, a moment in almost every practice" (2005: p. 137), accommodating both the semiotic and utilitarian functions of consumption. Sustainability here chiefly relates to the way in which resources are appropriated in the pursuit of social practices. The practice perspective explores phenomena normally concealed to cultural analysis, including the temporal ordering of practices, the coevolution of social and technical systems, and the dynamics of sociotechnical change. Theories of practice, furthermore, have a particular affinity to sustainable consumption: offering explanatory resources for the use of goods and services in the accomplishment of everyday routine tasks (e.g., showering or commuting) and the role of the dynamics of social practices in escalating demand (see Welch and Warde, 2015). While in principle the productive sphere is equally amenable to practice theoretical accounts, thus far within SPC research practice theory has generally been applied to microsociological accounts of end use consumption.

A further key innovation is the multilevel perspective (MLP) on sociotechnical transitions, which has offered SPC a framework with which to analyze innovation and change in production–consumption systems. MLP offers a heuristic framework that distinguishes between levels of niche (where technical and social innovations form), sociotechnical regime (or incumbent production–consumption systems), and landscape (or broad cultural and historical context) (see, e.g., Elzen et al., 2004). MLP has also focused attention on the importance of intermediary levels between production and end-use consumption. While practice theoretical accounts have tended to neglect the productive sphere, MLP accounts have arguably lacked sophistication in their approach to consumption. The emerging pragmatic synthesis between practice theory and MLP therefore represents an important contemporary development within SPC (see Cohen et al., 2013; McMeekin and Southerton, 2012).

Considerations of the environmental impact of production have traditionally privileged the material perspective, while considerations of consumption have traditionally been formulated in normative terms. The main contributions of SPC research have therefore been to consider the environmental impacts of production also as social phenomena and to address the sustainability of consumption patterns as an object of social scientific enquiry. The promise of SPC research is to offer insights into the complex, systemic linkages between social, natural, and technical phenomena; how these systems change; and what possibilities for guiding that change we might be afforded.

See also: Environmental Sociology; Industrial Ecology; Sustainable Development.

Bibliography

Carlsson, F., Johansson-Stenman, O., 2012. Behavioral economics and environmental policy. Annual Review of Resource Economics 4, 75–99.

Clark, G., 2007. Evolution of the global sustainable consumption and production policy and the United Nations Environment Programme's (UNEP) supporting activities. Journal of Cleaner Production 15, 492–498.

Cohen, M.J., 2001. The emergent environmental policy discourse on sustainable consumption. In: Cohen, M.J., Murphy, J. (Eds.), Exploring Sustainable Consumption: Environmental Policy and the Social Sciences. Pergamon, London.

Cohen, M.J., Murphy, J. (Eds.), 2001. Exploring Sustainable Consumption: Environmental Policy and the Social Sciences. Pergamon, London.

Cohen, M.J., Szejnwald, H., Vergragt, P.J. (Eds.), 2013. Innovations in Sustainable Consumption: New Economics, Socio-Technical Transitions, and Social Practices. Edward Elgar, Northampton, MA.

Elzen, B., Geels, F.W., Green, K. (Eds.), 2004. System Innovation and the Transition to Sustainability: Theory, Evidence and Policy. Edward Elgar, Cheltenham.

Fuchs, D., Lorek, S., 2005. Sustainable consumption governance: a history of promises and failures. Journal of Consumer Policy 28 (3), 261–288.

Harris, J.M., 2013. The macroeconomics of development without throughput growth. In: Cohen, M.J., Szejnwald, H., Vergragt, P.J. (Eds.), Innovations in Sustainable Consumption: New Economics, Socio-Technical Transitions, and Social Practices. Edward Elgar, Northampton, MA.

Hobson, K., 2002. Competing discourses of sustainable consumption: does the 'rationalization of lifestyles' make sense? Environmental Politics 11 (2), 95–120.

Hobson, K., 2013. 'Weak' or 'strong' sustainable consumption? Efficiency, degrowth, and the 10 Year Framework of Programmes. Environment and Planning C: Government and Policy 31, 1082–1098.

Jackson, T. (Ed.), 2006. The Earthscan Reader in Sustainable Consumption. Earthscan, London.

Kollmus, A., Agyeman, J., 2002. Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? Environmental Education Research 8 (3), 239–260.

Lebel, L., Lorek, S., Daniel, R. (Eds.), 2008. Sustainable Production Consumption Systems: Knowledge, Engagement and Practice. Springer, Dordrecht.
Luskin, J., 2007. Introduction to the special issue on, sustainable production and consumption: making the connection. Journal of Cleaner Production 15, 489–491.
McMeekin, A., Southerton, D., 2012. Sustainability transitions and final consumption: practices and socio-technical systems. Technology Analysis and Strategic Management 24 (4), 327–330.
Michaelis, L., 2003. Sustainable consumption and greenhouse gas mitigation. Climate Policy 3 (Suppl. 1), S135–S146.
Mol, A.J.P., Sonnenfeld, D.A., Spaargaren, G. (Eds.), 2009. The Ecological Modernization Reader: Environmental Reform in Theory and Practice. Routledge, London.
Princen, T., 1999. Consumption and environment: some conceptual issues. Ecological Economics 31, 347–363.
Reisch, L., Thøgersen, J. (Eds.), 2014. Handbook of Research on Sustainable Consumption. Edward Elgar Publishing, Cheltenham, UK.
Shove, E., 2003. Comfort, Cleanliness and Convenience: The Social Organization of Normality. Berg, Oxford.
Shove, E., Spurling, N. (Eds.), 2013. Sustainable Practices: Social Theory and Climate Change. Routledge, London.
Shove, E., Warde, A., 2002. Inconspicuous consumption. In: Dunlap, R.E., et al. (Eds.), Sociological Theory and the Environment. Rowman & Littlefield, London.
Sorrell, S., 2009. Jevons' Paradox revisited: the evidence for backfire from improved energy efficiency. Energy Policy 37, 1456–1469.
Southerton, D., et al. (Eds.), 2004. Sustainable Consumption: The Implications of Changing Infrastructures of Provision. Edward Elgar, Cheltenham, UK.
Southerton, D., Ulph, A. (Eds.), 2014. Sustainable Consumption: Multi-Dimensional Perspectives in Honour of Professor Sir Partha Dasgupta. Oxford University Press, Oxford.
Tukker, A., Charter, M., Vezzoli, C., Sto, E., Andersen, M.M. (Eds.), 2008. System Innovation for Sustainability: Perspectives on Radical Changes to Sustainable Consumption and Production. Greenleaf, Sheffield, UK.
Tukker, A., Cohen, M.J., Hubacek, K., Mont, O., 2010. The impacts of household consumption and options for change. Journal of Industrial Ecology 14 (1), 13–30.
UNCSD (United Nations Commission for Sustainable Development), 1992. Agenda 21 (viewed 10.01.14.). http://sustainabledevelopment.un.org/content/documents/Agenda21.pdf.
Vergragt, P., Akenji, L., Dewick, P. (Eds.), 2014. Sustainable Production, Consumption, and Livelihoods: Global and Regional Research Perspectives. Journal of Cleaner Production, vol. 63, pp. 1–12.
WCED (World Commission on Environment and Development), 1987. Our Common Future. Oxford University Press, Oxford.
Welch, D., Warde, A., 2015. Theories of practice and sustainable consumption. In: Reisch, L., Thøgersen, J. (Eds.), Handbook of Research on Sustainable Consumption. Edward Elgar Publishing, Cheltenham, UK.
Wilk, R., 2002. Consumption, human needs, and global environmental change. Global Environmental Change 12, 5–13.

Relevant Websites

http://esa.un.org/marrakechprocess/ – The Marrakech Process.
http://www.eupopp.net/default.htm – Policies to Promote Sustainable Consumption Patterns.
http://grfscp.wordpress.com – Global Research Forum.
http://scorai.org/ – Sustainable Consumption Research and Action Initiative.
http://www.score-network.org – Sustainable Consumption Research Exchange.
http://www.scp-centre.org/ – Collaborating Centre on Sustainable Consumption and Production.
http://scp.eionet.europa.eu/about – European Topic Centre on Sustainable Consumption and Production.
http://sustainabledevelopment.un.org – United Nations, Sustainable development.
http://www.uneptie.org/scp/ – United Nations Environment Programme, Sustainable Consumption & Production Branch.
http://www.weforum.org/issues/sustainable-consumption – World Economic Forum, Sustainable Consumption.

Sustainable Transportation

David L Greene, Oak Ridge, National Laboratory, Oak Ridge, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 22, pp. 15335–15339, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Sustainable transportation applies the concept of sustainable development to the movements of people and goods. Sustainable development requires that in meeting their own needs, generations must not compromise the ability of future generations to meet their needs. How to apply this to a single sector of society is unresolved, yet sustainable transport is clearly a necessary condition for societal sustainability. Motorization, concentrated in industrialized economies during most of the twentieth century, is rapidly globalizing. Transportation depends on petroleum for energy, yet half of total resources of conventional petroleum will have been consumed by 2025. Transition to alternative energy sources will be essential. Transportation is a major source of air pollution and noise. Since 1960, emissions per kilometer have been reduced but in many cities, the rate of traffic growth exceeds that of improving emissions controls. Solid waste from parts, fluids, scrapped vehicles, and infrastructure may require full recycling. Many consider the toll of deaths and injuries in accidents a sustainability issue. Transportation systems affect habitats directly, yet indirect effects on settlement patterns may pose the most difficult challenge. Equity requires increasing mobility in developing economies and for the less advantaged in general. How to measure progress toward sustainable transport is a key, unresolved issue.

The concept of sustainable transportation applies the idea of societal sustainability to a particular area of human activity: the movement of people and goods. Sustainability, or sustainable development, has been adopted by the United Nations agencies as an overarching goal of economic and social development (UNCED, 1992). Although there are several alternative definitions of sustainability, the most widely cited definition and the one generally recognized as initiating the global consideration of the subject, is that of the World Commission on Environment and Development (also known as the Brundtland Commission).

> Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs.
>
> (WCED, 1987, p. 43).

This and most other definitions constitute a normative concept of intergenerational responsibility or fairness. It is also significant that the context of the Brundtland definition was the issue of economic development and resulting conflicts with the environment. The concept of sustainability balances the right of less advantaged nations to economic development, with the imperative to preserve the resources, especially the environmental resources, that future generations will need for their own well-being. Thus, a concept of equity, both intergenerational and international, has been inherent in the idea of sustainability since its inception.

Sustainability pertains to the responsibility of an entire generation of society to future generations; whether it can meaningfully be applied to a single area of human activity such as transportation has been a subject of debate. That is, sustainability must be satisfied by the integral activities of a society and so, in this sense, it is not possible to judge whether one sector of society is sustainable on its own. On the other hand, it would not be possible for society to achieve sustainability, at least not in the long run, if even one sector were not sustainable. Thus, sustainable transportation may be considered a necessary but not sufficient condition for the sustainability of society.

Economic development has been central to the issue of sustainability since its inception. Economists have argued that the Bruntland definition implies non-decreasing well-being, and some have suggested that measures such as gross domestic product or real consumption per capita may be appropriate measures of sustainable growth. Others argue that well-being must be defined more broadly to include other areas of human development, such as education, health, environmental quality, and social justice (e.g., Pearce et al., 1994). In this broader view, the human institutions passed from one generation to another are just as important a resource as any other institution.

The question of how to translate the Brundtland imperative into a measurable and verifiable concept is still being worked out. In the economic literature, efforts to resolve this question have revolved around the concepts of weak and strong sustainability. The doctrine of weak sustainability holds that manufactured capital (machinery, buildings, infrastructure, etc.) is perfectly substitutable for natural capital (natural and environmental resources). This view holds not only that manufactured capital can be substituted for resources such as coal or iron ore, but that it can be substituted for environmental functions such as the ability of the air to absorb pollution or even the atmospheric and ocean processes that produce the global climate. Strong sustainability, on the other hand, maintains that there are limits to the substitutability of manufactured and natural capital. In particular, it maintains that there are certain environmental processes and functions that must be protected and preserved because they cannot be replaced by created capital. Ecosystems and biodiversity are often-cited examples of irreplaceable environmental processes.

How economic growth and expanding human welfare could continue indefinitely in the face of intensifying environmental impacts and finite natural resources is a central

question for theorists and researchers. Human activity, and especially economic activity, are inextricably linked to the environment by the ever-expanding quantities of natural resources extracted, processed and consumed (Pearce and Warford, 1993). Because mass and energy must be conserved due to the laws of nature, all the resources and energy used must end up somewhere in the form of waste (unless recycled). A useful way to conceive of the relationship between human activity and the environment is to imagine a coefficient of environmental intensity that relates total environmental damage to the total level of activity, perhaps measured, albeit incompletely by gross domestic product. Similarly, one can imagine a coefficient of natural resource requirements for input to the world economy. Because emissions controls, recycling, and greater energy efficiency can change the environmental coefficient, technological change and behavioral change clearly must play a key role in achieving sustainability. If the environmental coefficient can be reduced as fast or faster than the growth of human activity, and if increasing efficiency and recycling can similarly reduce the requirements for energy and other natural resource inputs, then it should be possible for human well-being to expand in a sustainable way.

The above line of reasoning identifies sustainability not as an equilibrium or final state of being able to continue indefinitely, but as a steady state in which the rate of improvement exceeds the rate of exhaustion of manufactured capital, natural resources, environmental services, and human and institutional capital. In this view, sustainability requires reducing demands on the environment per unit of activity as fast as or faster than the growth in activity. It requires increasing the efficiency of energy and resource use, or expanding the scope of useful resources at a rate at least as fast as the growth of human activity. It may even be interpreted more broadly as a matter of improving human institutions faster than challenges emerge to threaten them. While these concepts can and have been applied to transportation, a comprehensive theory of sustainable transportation that is measurable and verifiable is still yet to emerge.

Sustainable Transportation and Economic Development

Throughout history, the expansion and improvement of transportations systems has been an engine of economic development (Greene and Wegener, 1997). The fast and reliable movement of persons and commodities contributes more than simply a reduction in the costs of production. It facilitates the reorganization of production and consumption to permit specialization, scale economies, inventory reduction, and more. Today, transportation activity continues to increase at roughly the rate of economic growth in both the developed and developing economies.

Motorized transport is the source of essentially all the demands transportation imposes on natural and environmental resources. The motorization of transport, fully accomplished in the developed economies during the twentieth century, is now proceeding rapidly in the developing world. In 1950, 90 percent of the world's 50 million motor vehicles could be found in Europe or the United States. Today, half of the world's roughly 700 million vehicles are to be found outside of those countries. The developing economies own more than one-third of the world's motor vehicles, and vehicle ownership outside of the developed economies is growing at twice the rates seen in the developed economies. The global potential for motorization is enormous. In developed economies, motor vehicle ownership rates average 500 per 1000 persons. In the US, there is more than one motor vehicle per person of driving age. Ownership rates in the developing economies are about one-twentieth as high.

Expansion of modern, motorized transportation in the developing economies appears to be essential to their sustainable development. The World Bank has identified three broad types of sustainability in considering transport investments in developing countries: (1) economic and financial sustainability; (2) environmental sustainability; and, (3) social sustainability (World Bank, 1996). These categories reflect the view that to be sustainable, transportation systems must have stable financing in the context of market economies, must address the many adverse impacts of transportation on health and ecosystems, and must serve the mobility needs of the poor as well as those with the ability to pay.

Environmentally Sustainable Transport

Transportation systems have diverse and widespread impacts on the environment, affecting climate, air and water quality, noise, habitats, and biodiversity. Worldwide, transportation accounts for one-fourth of global carbon dioxide emissions. Scientists have concluded that the accumulation of carbon dioxide and other less common anthropogenic greenhouse gases in the world's atmosphere since the onset of the industrial revolution are now having a perceptible impact on the earth's climate. While it is not yet clear what the impacts on climate will be, they are likely to include increases in storm activity, changes in the patterns of rainfall, a general warming leading to higher sea levels, and other potentially harmful or even catastrophic effects. The quantity of carbon dioxide emission produced by transportation depends on the carbon content of the transportation fuels and the energy efficiency of transport systems. Today, motorized transport is entirely dependent on fossil fuels. The combustion of fossil fuels is the predominant source of greenhouse gas emissions from human activities. A typical passenger car adds 1 to 2 tonnes of carbon to the atmosphere each year. Despite 50–100 percent increases in the fuel economy of automobiles over the past 30 years, there appears to be no prospect of reducing global greenhouse gas emissions from transportation for at least 20 to 30 years. Clearly, continuing to force alterations in the earth's climate systems in unpredictable ways does not satisfy the sustainability doctrine.

Transportation is a major source of local and regional air pollution. In Europe and the United States, transportation vehicles emit roughly 45 percent of nitrogen oxides and 40 percent of volatile organic compounds that mix in the atmosphere to form ozone, a pollutant that exacerbates a variety of respiratory problems and causes damage to vegetation as well. Air pollutants from transport have an immediate impact on human health and on vegetation. Ozone pollution has been

observed to change the distribution of tree species in forests, retard the growth of vegetation, and damage crops. Prolonged exposure to ozone is strongly believed to reduce biodiversity of plant communities and diminish variability in their gene pools (NRC, 1997). Emissions of oxides of nitrogen and sulfur from motor vehicles contribute to acid rain with widespread negative impacts on terrestrial and aquatic ecosystems. Volatile organic compounds also include a number of toxic pollutants, including known and suspected carcinogens, such as benzene, formaldehyde, cyanide, and dioxin.

Transportation vehicles produce significant quantities of fine particulate pollution via several mechanisms. Vehicle tailpipe emissions contain particles produced by incomplete combustion. Particles and aerosols are formed when certain motor vehicle emissions react in the air. Finally, dust from roads is a major source of air-borne particulate matter. Fine particulate matter can be inhaled deeply into the lungs, contributing to a variety of respiratory problems. Epidemiological studies have correlated elevated morbidity and mortality rates with long-term, low-level exposure to fine particulates, as well as to episodes of acute pollution, suggesting that particulate pollution may be the most serious form of motor vehicle pollution from the perspective of human health (NRC, 1997).

Since the 1950s, enormous technological progress has been made in reducing the pollutant emissions from new transportation vehicles. A properly operating motor vehicle manufactured in 2000 produces 1 percent to 10 percent of the pollution (depending on the pollutant) of a car made 35 years earlier. Still, even this rate of progress has been insufficient to achieve air quality standards in many urban areas because the growth of travel has outpaced the rate of technological improvement. Further significant improvements in conventional vehicles and fuels will be made from 2000 to 2010. Yet achieving sustainability will probably require entirely new energy sources and propulsion technologies ultimately.

Solid Waste

Transportation infrastructure and vehicles generate enormous amounts of solid waste. Some of this waste, such as iron and steel, has a long history of recycling. Other components like plastics most often end up in landfills or incinerators. Waste oil and fluids, batteries, and other components often contain toxic materials that pose serious health and environmental hazards if not properly disposed of. The maintenance and reconstruction of transportation facilities also generate substantial quantities of mostly asphalt and concrete waste. Not only increased recycling, but life-cycle management of materials for transportation systems has been recognized as essential to achieving sustainable transportation systems (OECD, 1997).

Noise

It may seem odd to include noise among the issues affecting transportation's sustainability. Noise, after all, is inherently transitory and its impacts on society and the environment would seem to be also. On the other hand, transportation systems operate more or less continuously, generating noise around the clock. Studies carried out in Europe have shown that about 20 percent of the population are exposed to daytime transport noise exceeding acceptable levels (65dB(A)) (European Commission, 1995). Another 170 million are exposed to noise levels that cause serious annoyance as defined by the World Health Organization. Road noise is the dominant source, followed by rail and air transport. Furthermore, data for the 1980–95 period showed no significant improvement in exposure to traffic noise for the population of Europe. The principal impacts of transportation noise are annoyance, interference with work and sleep and, less commonly, hearing loss. Estimates of the costs of noise pollution in Europe range between 0.1 percent and 2 percent of GDP.

Natural Resources

Transportation is a major consumer of two critical, limited resources: land and petroleum. Worldwide, transportation accounts for 56 percent of global petroleum consumption and between 1973 and 1997 was responsible for all of the increase in world oil consumption (IEA, 2000). Recent assessments of the world's supplies of conventional petroleum indicate that sometime between 2005 and 2025, the world will have consumed half of all the conventional petroleum that is known or believed to exist (EIA, 2000 and USGS, 2000). Once this point is passed, it is virtually certain that production of conventional petroleum will begin to decline, and unconventional energy sources will have to be substituted at an increasing rate. The world's transportation system, which is 97 percent dependent on petroleum for energy, will not be sustainable unless it can make a transition to alternative energy sources. While there are more than adequate resources of fossil fuels that can be converted to transport fuels (e.g., coal, oil shale, and tar sands), the greater environmental impacts associated with using these fuels raises serious questions about the sustainability of a motorized world transport system based on fossil fuels.

Perhaps even more serious are the questions raised by transportation systems' expanding direct and indirect uses of land. Not only does transportation infrastructure such as road, airports, and railroad tracks, occupy millions of hectares of land, but the infrastructure and vehicles that use it have significant impacts on surrounding ecosystems. Linear transportation structures serve as deadly barriers to the migration and movements of species, and alter water flow and groundwater. But perhaps more importantly, transportation infrastructure indirectly alters the patterns and density of human settlements, with even more pervasive impacts on ecosystems and biodiversity (NRC, 1997).

Sustainable Communities

Transportation is recognized as a key element of sustainable communities (PCSD, 1997). Human development patterns and the transportation systems that serve them are strongly interdependent. Transportation choices made by present generations will constrain the opportunities of future generations because

built environments can be changed only slowly and at great cost. The high level of mobility provided by automotive transport has engendered dispersed low-density development, sometimes referred to as sprawl. The area of built-up urban land in the United States increased by an estimated 14 million acres between 1982 and 1992, driven in large part by the expansion of highway infrastructure (BTS, 1996, p. 162).

As demand for mobility has continued to grow, while the land available for expanding transportation infrastructure in urban areas has not, chronic congestion has become a major problem in cities around the world. Increasing levels of traffic congestion are not only a source of stress and waste the time of present generations, but can restrict the opportunities for mobility of future generations. If the capacities of transportation systems cannot be expanded to accommodate future growth without incurring unacceptable trade-offs, the options available to future generations will be diminished.

Safety

Safety must also be included in assessing the sustainability of transportation systems (e.g., Black, 2000). While absolutely compelling as a human tragedy, the sense in which transportation safety is a sustainability issue may be less obvious. Globally, there are about 700 000 motor vehicle fatalities annually, and many millions of serious injuries. On the other hand, fatality rates (e.g., per 100 000 vehicle miles) are generally declining. In developed economies such as the United States, recent years have seen declines in total fatalities even while motor vehicle travel and population size continue to grow. In countries undergoing rapid motorization, however, total fatalities are increasing at the same time rates are declining. When both fatalities and fatality rates are increasing, the sustainability of transportation systems is in question. The resource at issue is human life itself, and the needs of future generations would clearly appear to be threatened by increasingly unsafe transportation systems. Where both total fatalities and fatality rates are declining, an argument can be made for sustainability. Without a doubt, safety is the single most important problem facing transportation. Yet, to date there is little agreement on how safety should be interpreted in the context of the sustainability doctrine.

Equity

Equity among generations is clearly integral to the concept of sustainability. For sustainability to exclude equity considerations among individuals within a generation would be inherently inconsistent. The Charter of European Cities and Towns Towards Sustainability states that the goals of sustainable development are '… to achieve social justice, sustainable economies, and environmental sustainability' (Greene and Wegener, 1997). Furthermore, the context of the Brundtland sustainability definition was clearly the need to preserve the global environment while permitting developing countries the economic development that is their right. Clearly social justice and particularly equity have been integral to the concept of sustainability. It has been argued that the concept of intergenerational fairness embodied in sustainability, in fact arises from a societal commitment to equality of opportunity among contemporaries (Howarth, 1997). Exactly how equity concerns are to be incorporated in an operating definition of sustainability has yet to be resolved, as have the implications for transportation. At a minimum, equity considerations include the interests of developing economies to pursue higher levels of mobility, as well as the needs of disadvantaged citizens for access to mobility in all societies.

Measuring Sustainability

If policies are to be formulated and implemented to achieve sustainable transportation, indicators must be found to measure it (OECD, 1994). If the steady-state concept of sustainability is accepted, these indicators should show:

- whether the world's environmental resources are being reduced by transportation's impacts on air or water pollution;
- whether the global climate is being altered by greenhouse gas emissions;
- whether habitats are being destroyed through direct or indirect land-use impacts; and
- whether the generation of solid waste is increasing or being restrained by recycling.

Indicators must show whether the fundamental natural resources required for transportation, such as energy and land, are being expanded as fast or faster than they are being consumed, and whether transportation systems are being developed and managed so as to enhance social equity by improving access to mobility for all. Until the consequences of transportation for sustainability are made visible measuring them, progress toward sustainable transportation will be limited.

Editor's Note

Sustainable transportation continues to be be a concern of urban planners, transportation specialists, and environmentalists. Three recent books provide helpful updates to the themes, findings, and literature reviewed in Dr. Greene's original entry: Black, 2010; Schiller et al, 2010; Worrall, 2005.

See also: Ecotourism; Environmental Economics; Resource Geography; Safety, Economics of; Tourism, Geography of; Transportation Geography; Transportation Planning; Transportation, Supply, and Congestion.

Bibliography

Black, W.R., 2000. Socio-economic barriers to sustainable transport. Journal of Transport Geography 8, 141–147.

Black, William R., 2010. Sustainable Transportation: Problems and Solutions. Guilford, New York.

Bureau of Transportation Statistics (BTS), 1996. Transportation Statistics Annual Report. US Department of Transportation, Washington, DC.

Energy Information Administration (EIA), 2000. Long-term World Oil Supply. US Department of Energy, Washington, DC. Available on the World Wide Web at. http://www.eia.doe.gov.

European Commission, 1995. Towards Fair and Efficient Pricing in Transport. DG-VII, Brussels, Belgium. COM(95)691, Directorate-General for Transport.

Greene, D.L., Wegener, M., 1997. Sustainable transport. Journal of Transport Geography 5 (3), 177–190.

Howarth, R.B., 1997. Sustainability as opportunity. Land Economics 73 (4), 569–579.

International Energy Agency (IEA), 2000. Key Energy Statistics. OECD, Paris. Available on the World Wide Web at. http://www.iea.org.

National Research Council (NRC), Committee for a study on transportation and a sustainable environment, 1997. Toward a Sustainable Future. Special Report 251, Transportation Research Board. National Academy Press, Washington, DC.

Organization for Economic Co-operation and Development (OECD), 1994. Environmental Indicators. OECD, Paris.

OECD, 1997. Recycling strategies for road works. Road Transport Research. Organization for Economic Co-operation and Development, Paris.

President's Council on Sustainable Development (PCSD), 1997. Sustainable Communities Task Force Report. Washington, DC.

Pearce, D.W., Warford, J.J., 1993. World Without End. International Bank for Reconstruction and Development, Washington, DC.

Pearce, D.W., Atkinson, G.D., Dubourg, W.R., 1994. The economics of sustainable development. Annual Review of Energy and the Environment 19, 457–474.

Schiller, Presaton L., Bruun, Eric C., Kenworthy, Jeffrey R., 2010. An Introduction to Sustainable Transportation. Earthscan, London.

United Nations Conference on Environment and Development (UNCED), 1992. Agenda 21. United Nations, New York.

United States Geological Survey (USGS), 2000. US Geological Survey World Petroleum Assessment 2000 – Description and Results. Department of the Interior, Denver, CO.

World Commission on Environment and Development (WCED), 1987. Our Common Future. Oxford University Press, Oxford, UK.

World Bank, 1996. Sustainable Transport. International Bank for Reconstruction and Development, Washington, DC.

Worrall, Harold, 2005. A System at Risk: The Economics of Transportation. iUniverse, Bloomington, IN.

Symbolic Boundaries

Michèle Lamont, Harvard University, Cambridge, MA, USA
Sabrina Pendergrass, University of Virginia, Charlottesville, VA, USA
Mark Pachucki, Harvard Medical School, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M. Lamont, volume 23, pp. 15341–15347, © 2001, Elsevier Ltd.

Abstract

'Symbolic Boundaries' are the lines that include and define some people, groups, and things while excluding others. These distinctions can be expressed through normative interdictions (taboos), cultural attitudes and practices, and patterns of likes and dislikes. They play an important role in the creation of inequality and the exercise of power. The term 'symbolic boundaries' also refers to the internal distinctions of classification systems and to temporal, spatial, and visual cognitive distinctions in particular. This article focuses on boundaries within and between groups. It discusses the history, current research, and future challenges of work on this topic.

Definition and Intellectual Context

'Symbolic Boundaries' are the lines that include and define some people, groups, and things while excluding others (Epstein, 1992: p. 232). These distinctions can be expressed through normative interdictions (taboos), cultural attitudes and practices, and patterns of likes and dislikes. They play an important role in the creation of inequality and the exercise of power. The term 'symbolic boundaries' also refers to the internal distinctions of classification systems and to temporal, spatial, and visual cognitive distinctions in particular (Wagner-Pacifici, 2000; Zerubavel, 1997). This article focuses on boundaries within and between groups. It discusses the history, current research, and future challenges of work on this topic.

The literature on symbolic boundaries has gained importance since the 1960s due to a convergence between research on symbolic systems and indirect forms of power. Writings by Pierre Bourdieu, Mary Douglas, Norbert Elias, Erving Goffman, and Michel Foucault on this topic have influenced several disciplines internationally, particularly anthropology, history, literary studies, and sociology. In North America, a renewed cultural sociology has produced wide-ranging empirical research agendas on symbolic boundaries and inequality. In other fields including community, cognition, deviance, gender, immigration, knowledge and science, nationalism, professions, race and ethnicity, and social movements, boundaries issues have gained analytical prominence.

History

Two of the founding fathers of sociology played central roles in shaping the literature on symbolic boundaries: Emile Durkheim and Max Weber. We review their contributions before turning to the 'neo-classical' writings of Mary Douglas, Norbert Elias, and Thorstein Veblen, which illustrate the lasting influence of Durkheim and Weber on this literature up to the 1960s. While Durkheim emphasized classification systems and their relationship with the moral order, Weber emphasized their impact on the production and reproduction of inequality. (For a more encompassing historical overview, see Schwartz, 1981.)

One widely used example of symbolic boundaries is taken from Durkheim's work, *Les formes élementaires de la vie religieuse* (*The Elementary Forms of Religious Life*, 1965[1911]). Durkheim argues that the religious experience is distinct from other types of experiences because it involves a symbolic distinction between the realms of the sacred and the profane (pp. 234, 250). The meanings of these realms are mutually exclusive and are defined relationally, through interdictions and rituals that isolate and protect the former from the latter (e.g., a Roman-Catholic sinner cannot receive communion until he is purified through confession) (p. 271).

The distinction between the sacred and the profane extends to the whole universe of objects and people in which it takes place. For instance, the status of members of a community is defined by the types of relationship they have with sacred objects (e.g., Roman Catholic women cannot celebrate mass). In this sense, religious systems provide a cosmology, i.e., a general interpretation of how the world is organized and how its elements relate to one another and to the sacred. This cosmology acts as a system of classification and its elements are organized according to a hierarchy (e.g., counterposing the pure with the impure). The belief invested in this 'order of things' structures people's lives to the extent that it limits and facilitates their action.

Durkheim also points to the existence of a moral order (i.e., a common system of perception of reality that regulates, structures, and organizes relations in a community). This system operates less through coercion than through intersubjectivity (p. 238). In fact, Durkheim defines society by its symbolic boundaries: it is the sharing of a common definition of the sacred and the profane, of similar rules of conduct and a common compliance to rituals and interdictions that defines the internal bonds within a community. Hence, he posits that the boundaries of the group coincide with those delimiting the sacred from the profane.

Unlike Durkheim, Max Weber focuses on the role of symbolic boundaries (honor) in the creation of social inequality rather than social solidarity. In *Wirtschaft und Gesellschaft* (*Economy and Society*, 1922/1956), he describes human beings as engaged in a continuous struggle over scarce resources. To curb competition, they discriminate toward

various groups on the basis of their cultural characteristics, such as lifestyle, language, education, race, or religion. In the process, they form status groups whose superiority is defined in relation to other groups. They cultivate a sense of honor, privilege relationships with group members, and define specific qualifications for gaining entry to the group and for interacting with lower status outsiders (e.g., opposing miscegenation). They invoke their higher status and shared rules of life to justify their monopolization of resources. Hence, cultural understandings about status boundaries have a strong impact on people's social position and access to resources.

Thorstein Veblen's *The Theory of the Leisure Class* (1899/1979) parallels Weber's writings. This American economist suggests that habits of thought ('classifying and demarcating') are central to the mechanisms that produce boundaries between status groups, and they are often organized around notions of superiority and inferiority concerning employment, consumption, and leisure. For instance, idleness symbolizes status because it signifies pecuniary status. He writes, "refined tastes, manners, and habits of life are useful evidence of gentility because good breeding requires time, application, and expenses" (p. 49) and are therefore not available to those "whose time and energies are taken up with work." Veblen also developed the concept of 'conspicuous consumption.' He argues that the possession and display of wealth confers honor: as an invidious distinction, it symbolizes ranking within a group. This way of manifesting superiority is more common when predatory aggression or war are less frequent. Veblen's analysis assumes that there is a tendency to change standards of sufficiency as one's pecuniary situation improves; one becomes restless with creating 'ever-widening distance' between herself and the average standard.

Also paralleling Weber's work, is the work of German sociologist Norbert Elias. In *Über den Prozess der Zivilisation* (*The Civilizing Process*, 1982[1939]), Elias analyzes the emergence of a boundary between civilized and barbarian habits using evidence from Western manner manuals written between the late middle ages and the Victorian period. He examines 'natural' bodily functions such as spitting and defecating to demonstrate the growing centrality of shame and embarrassment in instituting norms of behavior in public and private over time. More broadly, he argues that standards of behavior and feelings and personality structures (what he calls 'habitus,' or habits emerging from social experience) vary across hierarchical groups in society and that these variations are key to pacification and the exercise of power. In a later study, *The Established and the Outsiders* (1965/1994, with John L. Scotson), Elias shows how residents of one part of a town, 'the Village' had higher status than residents of another part of the town, 'the Estate,' because the former had the social cohesion and resources to stigmatize the outsiders and impose their own definition of self.

To turn now to the lasting influence of Durkheim's work, in *Purity and Danger* (1966), Mary Douglas examines the order-producing, meaning-making and form-giving functions of classification systems and the role of rituals in creating boundaries grounded in fears and beliefs. In *Natural Symbols* (1970), she describes how the structure of binary symbolic systems reflects group structures. Like Elias, she also probes the moral order and focuses on the system of social control as expressed through the body and through observable artifacts of everyday life (food, dirt, and material possessions). She argues that the basis of order in social life is the presence of symbols that demarcate boundaries.

One of Douglas's main concerns is how communities differentiate themselves from one another and how they are internally differentiated. She distinguishes groups on the basis of their degree of social control and of the rigidity of their grid (i.e., the extent to which their system of classification competes with other systems). In societies with high social control and great cultural rigidity (i.e., what she calls high grid and group), there is a concern to preserve social boundaries; the role structure is clearly defined; and formal behavior is highly valued and well defined in publicly insulated roles. Through 'the purity rule,' formality screens out irrelevant organic processes.

Current Theory and Research

In the contemporary literature on symbolic boundaries, both the neo-Weberian and neo-Durkheimian heritage remain strong. The question of how boundaries intersect with the production of inequality has attracted interest since the publication of Pierre Bourdieu's impressive corpus. In the United States, cultural sociologists have assessed Bourdieu's theoretical claims and used his work to illuminate the cultural aspects of class, gender, and racial inequality. Other developments concern the study of identity through boundary work, and research on moral order, community, symbolic politics, and religion. While some of this research is reviewed in Pachucki et al. (2007) and in Lamont (2012), the discussion below considers new contributions as well as less recent ones.

Culture and Inequality

In the last 30 years, a large neo-Weberian literature emerged about processes of closure. Frank Parkin (1979) drew on Weber to study the distributive struggle for monopolizing or usurping resources within and across classes. He emphasized the right of ownership and credentialism, i.e., the use of educational certificates to monopolize positions in the labor market. Equally inspired by Durkheim, Collins (1998) extended his earlier work on credentialism and interaction rituals to analyze how intellectuals compete to maximize their access to key network positions, cultural capital, and emotional energy, which generates intellectual creativity. These contributions coincide with those of Pierre Bourdieu and his collaborators.

In *Reproduction* (1977[1970]), Pierre Bourdieu and Jean-Claude Passeron proposed that the lower academic performance of working-class children cannot be accounted for by their lower ability but by institutional biases. They suggest that schools evaluate children based on their familiarity with the culture of the dominant class (or cultural capital), thus penalizing lower-class students. Extensive vocabulary, wide-ranging cultural references, and command of high culture are valued by the school system and students from higher social backgrounds are exposed to this class culture at home. Lower-class children remain under the spell of the dominant class culture. They blame themselves for their failure, and

consequently drop out or sort themselves into lower prestige educational tracks.

This work extends Karl Marx and Friedrich Engels' (1848/1960) 'dominant ideology thesis,' which centers on the role of ideology in cementing relations of domination by camouflaging exploitation and differences in class interests. However, Bourdieu and Passeron are more concerned with classification systems than with representations of the social world itself, i.e., with how representations of social relationships, the state, religion, and capitalism contribute to the reproduction of domination. Implicitly building on Gramsci (1971), they focus on the control of subjectivity in everyday life through the shaping of common sense and the naturalization of social relations. They broaden Marx and Engels by suggesting that crucial power relations are structured in the symbolic realm proper, and are mediated by meaning. They *de facto* provide a more encompassing understanding of the exercise of hegemony by pointing to the incorporation of class-differentiated cultural dispositions mediated by both the educational system and family socialization.

In *Distinction* (1984[1979]), Bourdieu applies this analysis to the world of taste and cultural practice at large. He shows how the logic of class struggle extends to the realm of taste and lifestyle, and that symbolic classification is key to the reproduction of class privileges: dominant groups define their own culture as superior. Thereby they exercise 'symbolic violence,' i.e., impose a specific meaning as legitimate while concealing the power relations that are the basis of its force (Bourdieu and Passeron, 1977[1970]: p. 4). They define legitimate and 'dominated' cultures in opposition: the value of cultural preferences and behaviors are defined relationally around binary oppositions (or boundaries) such as higher/lower, pure/impure, distinguished/vulgar, and aesthetic/practical (p. 245). The legitimate culture they thereby define is used by dominant groups to mark cultural distance and proximity, monopolize privileges, and exclude and recruit new occupants to high status positions (p. 31). Through the incorporation of '*habitus*' or cultural dispositions, cultural practices have inescapable and unconscious classificatory effects that shape social positions.

A large American literature applying, extending, assessing, and critiquing the contributions of Bourdieu and his collaborators developed in the wake of their translation in English (for an early review, see Lamont and Lareau, 1988.) For instance, DiMaggio (1987) suggests that boundaries between cultural genres are created by status groups to signal their superior status. DiMaggio and Mohr (1985) found that levels of cultural capital influence higher education attendance and completion as well as marital selection patterns in the United States. Lamont (1992) critiqued Bourdieu (1984) for exaggerating the importance of cultural capital in upper-middle-class culture and for defining salient boundaries a priori, instead of inductively. Based on interviews with professionals and managers, she showed that morality, cultural capital, and material success are defined differently and that their relative importance vary across national contexts and by subgroups. Lamont also showed variations in the extent to which professionals and managers tolerate the lifestyles and tastes of other classes, and argued that cultural laissez-faire is more important in American society than French society. High social and geographic mobility, strong cultural regionalism, ethnic and racial diversity, political decentralization, and relatively weak high culture traditions translate into less highly differentiated class cultures in the United States than France (also see Lamont, 2010).

Other sociologists also argue that cultural boundaries are more fluid and complex than cultural capital theory suggests. Hall (1992) emphasized the existence of heterogeneous markets and of multiple kinds of cultural capital. He proposes a 'cultural structuralism' that addresses the multiplicity of status situations in a critique of an overarching market of cultural capital. Crane (2000) analyzes how social change disrupts the relationship between cultural capital and social class strata during nineteenth- and twentieth-century France and the United States. In separate studies of visual art consumption, Halle (1993) and Banks (2009) show how meanings attached to artwork in the home are influenced by factors afield of class. For Halle's participants, art consumption does not necessarily generate social boundaries, and he finds that the meaning attached to living room art is somewhat autonomous from professional evaluations. Banks explores how middle-class blacks feel they contribute to black cultural advancement through patronage of black artists and cultural institutions. Pachucki (2012) explores how art professionals make symbolic distinctions in their everyday work with art, and shows how these boundaries shape status hierarchies across museums. Lareau (2003) analyzed how middle- and working-class parents socialize their children for middle- and working-class jobs. Researchers have continued to provide a more complex approach to cultural capital through ethnographies of how teenagers learn status and privilege in elite boarding schools (Gaztambide-Fernandez, 2009; Khan, 2011) and an ethnography of hiring practices in elite professional firms (Rivera, 2012).

Research on cultural omnivorousness suggests that cultural breadth is a highly valued resource in the upper and upper-middle classes, countering Bourdieu's postulate that the value of tastes is defined relationally through a binary or oppositional logic between upper- and lower-class cultures. Studies of musical tastes (Bryson, 1996; Peterson and Kern, 1996; Peterson and Rossman, 2008), familiarity with popular culture (Erickson, 1996), arts participation (Alderson et al., 2007; DiMaggio and Mukhtar, 2004), and food consumption (Johnston and Baumann, 2010) are among a growing body of scholarship that highlights the complexities of how individuals distinguish themselves through the diversity of what they consume. Lena's (2012) work on music genres suggests that aesthetic boundaries are only one element in a complex interplay with audiences and industry forces. Her careful analysis shows how types of music share similar trajectories in their emergence and transformation.

Identity and Boundary Work

The literature on identity is another arena where the concept of symbolic boundaries has become more central. Sociologists and psychologists have studied boundary work, a process central to the constitution of the self. Thomas Gieryn originally proposed the concept of boundary work in the early 1980s to designate "the discursive attribution of selected qualities to scientists, scientific methods, and scientific claims for the

purpose of drawing a rhetorical boundary between science and some less authoritative residual non-science"(1999: pp. 4–5). Since then, sociologists have analyzed this process by looking at self-definitions of ordinary people and in a range of contexts beyond science.

Social psychologists working on group categorization have studied the segmentation between 'us' and 'them.' Brewer's (1986) social identity theory suggests that "pressures to evaluate one's own group positively through in-group/out-group comparison lead social groups to differentiate themselves from each other." This process of differentiation aims "to maintain and achieve superiority over an out-group on some dimension" (Tajfel and Turner, 1985: pp. 16–17). While these authors understand the relational process as a universal tendency, sociologists are concerned with analyzing precisely how boundary work is accomplished, i.e., with what kinds of typification systems, or inferences concerning similarities and differences, groups mobilize to define who they are.

The study by Jenkins (1996) on social identity also contributes to the study of boundary work. He describes collective identity as constituted by a dialectic interplay of processes of internal and external definition. On one hand, individuals must be able to differentiate themselves from others by drawing on criteria of community and a sense of shared belonging within their subgroup. On the other hand, this internal identification process must be recognized by outsiders for an objectified collective identity to emerge. These insights continue the early efforts of Du Bois (1903) who emphasized the work of managing the 'color line' or 'veil' that defines race and of Barth (1969) who analyzed the process of ethnic boundary making. They also mirror work on racial group positioning (Blumer, 1958; Bobo and Hutchings, 1996).

During the 1990s, seminal studies of self-definitions among ordinary people emerged. For instance, Newman (1999) analyzes how poor fast-food workers define themselves in opposition to the unemployed poor. Lamont (1992) studies the boundary work of professionals and managers while Lamont (2000) examines how workers in the United States and France define worthy people in opposition to the poor, 'people above,' blacks, and immigrants, drawing moral boundaries toward different groups across the two national contexts. Lichterman (1999) explores how volunteers define their bonds and boundaries of solidarity by examining how they articulate their identity around various groups. These mappings translate into different kinds of group responsibility, in "constraining and enabling what members can say and do together." Binder (1999) analyzes boundaries that proponents of Afrocentrism and multiculturalism build in relation to one another in conflict within the educational system. Becker (1999) studies how religious communities build boundaries between themselves and 'the public.' Finally, Gamson (1992) analyzes how the injustice frames used in social movements are organized around 'us' and 'them' oppositions.

Studies of boundary work and identity continue to flourish in the fields of race and ethnicity and immigration. For instance, Carter (2005) examines the diverse ways black and Latino youth manage ethnic boundaries in school and their peer groups and the consequences for their academic achievement. Warikoo (2011) analyzes how second-generation Indo-Caribbean teenagers negotiate ethnic boundaries in London and New York and the role of hip-hop and other aspects of youth culture. Wimmer and Lewis (2010) use Facebook data to explore how racial homogeneity in Americans' social networks is linked with social stratification. Other scholars have considered the relationship between racial and class boundaries. Young (2006) focuses on the ways poor black men think about status mobility and opportunity. Lacy (2007) analyzed the boundary work of race and class among blacks across different types of middle-class neighborhoods. Pattillo (2007) considers boundaries between middle-class and poor blacks in a mixed income neighborhood and how they are negotiated as neighborhood residents promote black political action. In the area of immigration and ethnicity, Alba (2009) and Wimmer (2013) have focused on group formation processes. Also, Roth (2012) analyzes how migration and transnationalism influence definitions of panethnic identity among Puerto Ricans and Dominicans. Future research on the process of collective identity formation may benefit to focus on the dynamic between self-identification and social categorization.

Moral Order, Community, Symbolic Politics, and Religion

A third strand of work on symbolic boundaries presents more palpable neo-Durkheimian influences. Several studies have centered on moral order and on communities. Wuthnow (1987: p. 69) writes, "Order has somehow to do with boundaries. That is, order consists mainly of being able to make distinctions – of having symbolic demarcations – so that we know the place of things and how they relate to one another." A recent example of this neo-Durkheimian line of work is Alexander's (1992) semiotic analysis of the symbolic codes of civic society. The author describes these codes as "critically important in constituting the very sense of society for those who are within and without it." He also suggests that the democratic code involves clear distinctions between the pure and the impure in defining the appropriate citizen. His analysis locates those distinctions at the levels of people's motives and relationships, and of the institutions that individuals inhabit (with 'honorable' being valued over 'self-interested' or 'truthful' over 'deceitful' in the case of the democratic code).

The last decades have produced studies of status politics that documented how groups sharing a lifestyle made such distinctions, engaged in the maintenance of the moral order, and simultaneously bolstered their own prestige. Particularly notable is Gusfield (1963) who analyzed the nineteenth-century American temperance movement in favor of the prohibition and the Eighteenth Amendment to the constitution. Gusfield understands this movement as a strategy used by small-town Protestants to bolster their social position in relation to urban Catholic immigrants. Along similar lines, Luker (1984) describes the worldviews of antiabortion and pro-choice activists. She shows that they have incompatible beliefs about women's careers, family, sexuality, and reproduction, and that they talk past one another and define themselves in opposition to one another. Beisel (1997) has studied Anthony Comstock's nineteenth-century antipornography movement to protect the morality of children in the context of important social changes that threatened the reproduction of upper class privileges. However, she argued against the distinction between

symbolic and class politics and showed how the two often operate hand in hand, particularly in the drawing of moral boundaries. The literature on social movements includes studies that focus on the process by which categories of people are turned into categories of enemies (Jasper, 1997). Using the case of the 11 September attacks, Wagner-Pacifici (2010) also advanced a 'political semiosis' methodology for understanding the bounding and unbounding of events.

Scholars recently have shown how symbolic boundaries play a key role in issues around religious identification. Bean et al. (2008) offer a comparative-historical investigation of how Canadian evangelicals mobilize around moral issues, finding that in both Canada and the United States, political identification, beliefs, and religious group membership evolve in a path-dependent process. Religious boundaries and everyday interactions intersect for Tavory (2010), who explores how members of an Orthodox Hassidic group in Los Angeles signal Jewish identification. Yukich (2010) focuses on how the Catholic Worker movement, which considers itself inclusive, constructs exclusionary boundaries against other service and charitable organizations. For their part, Edgell and Tranby (2010) investigate how various types of sub-group identity relates with a shared American identification. They find three types of citizen that they term 'cultural preservationists,' 'critics of multiculturalism,' and 'optimistic pluralists.' Also, Edgell and Gerteis (2006) identify atheism as a salient group boundary in the United States.

Challenges and Future Directions

A persistent challenge since the emergence of this field of scholarship is understanding the connection between objective boundaries and symbolic boundaries. As Lamont (1992) argued, symbolic boundaries are a necessary but insufficient condition for the creation of objective boundaries. Early research on objective social boundaries focused on topics such as the relative importance of educational endogamy versus racial endogamy among the college-educated (Kalmijn, 1991); racial hiring and firing (Silver and Zwerling, 1992); residential racial segregation (Massey and Denton, 1993); the relative permeability of class boundaries (Wright and Cho, 1992); and the creation of professional boundaries (Abbott, 1988). Tilly (2004) provides a synthesis of social boundary mechanisms. More recent theoretical consolidation has contributed to our understanding of links between symbolic and social boundaries. Lamont and Molnar (2002) issued a challenge to find similarities of boundary configurations across cases. These efforts have been extended to the study of ethnicity and race (see special issue of *Poetics*, Pachucki et al., 2007), immigrant and national identity (Zolberg and Woon, 1990; Bail, 2008; Rivera, 2008; Wimmer, 2013; Brubaker, 2009), evaluation processes (Lamont, 2012), and responses to stigmatization (see special issue of *Ethnic and Racial Studies*, Lamont and Mizrachi, 2012 and the special section of the *Du Bois Review*, Lamont et al., 2012).

See also: Art and Socialisation; Collective Identity; Cultural Capital and Education; Cultural Mediators and Gatekeepers; Cultural Participation, Trends In; Culture and Networks; Culture, Cognition and Embodiment; Culture, Production of: Prospects for the Twenty-First Century; Globalization and World Culture; Leisure and Cultural Consumption: US Perspective; Nation-State as Symbolic Construct, Networks and Cultural Consumption; Social Inequality in Cultural Consumption Patterns.

Bibliography

Abbott, A., 1988. The System of Professions: An Essay on the Division of Expert Labor. University of Chicago Press, Chicago.
Alba, R., 2009. Blurring the Color Line: The New Chance for a More Integrated America. Harvard University Press, Cambridge.
Alderson, A., Junisbai, A., Heacock, I., 2007. Social status and cultural consumption in the United States. Poetics 35, 191–212.
Alexander, J., 1992. Citizens and enemy as symbolic classification: on the polarizing discourse of civil society. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago.
Bail, C., 2008. The configuration of symbolic boundaries against immigrants in Europe. American Sociological Review 73, 37–59.
Banks, P., 2009. Black cultural advancement: racial identity and participation in the arts among the black middle class. Ethnic and Racial Studies 33, 272–289.
Barth, F., 1969. Introduction. In: Barth, F. (Ed.), Ethnic Groups and Boundaries: The Social Organization of Culture Difference. George Allen and Unwin, London.
Bean, L., Kaufman, J., Gonzalez, M., 2008. Why doesn't Canada have an American-style Christian right? A comparative framework for analyzing the political effects of Evangelical subcultural identity. Canadian Journal of Sociology 33, 899–943.
Becker, P., 1999. Congregations in Conflict: Cultural Models of Local Religious Life. Cambridge University Press, New York.
Beisel, N., 1997. Imperiled Innocents: Anthony Comstock and Family Reproduction in Victorian America. Princeton University Press, Princeton.
Binder, A., 1999. Friend and foe: boundary work and collective identity in the afrocentric and multicultural curriculum movements in American public education. In: Lamont, M. (Ed.), The Cultural Territories of Race: Black and White Boundaries. University of Chicago Press, Chicago and Russell Sage Foundation, New York.
Blumer, H., 1958. Race prejudice as a sense of group position. Pacific Sociological Review 1, 3–7.
Bobo, L., Hutchings, V.L., 1996. Perceptions of racial group competition: extending Blumer's theory of group position to a multiracial social context. American Sociological Review 61, 951–972.
Bourdieu, P., 1979. Distinction, critique sociale du jugement. Editions de Minuit, Paris (1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge).
Bourdieu, P., Passeron, J.-C., 1970. Reproduction. éléments pour une théorie du système d'enseignment. Editions de Minuit, Paris (1977. Reproduction in Education, Society, and Culture. Beverly Hills, Sage).
Brewer, M.B., 1986. The role of ethnocentrism in intergroup conflict. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Nelson-Hall Publishers, Chicago.
Brubaker, R., 2009. Ethnicity, race, and nationalism. Annual Review of Sociology 35, 21–42.
Bryson, B., 1996. 'Anything but heavy metal': symbolic exclusion and musical dislikes. American Sociological Review 61, 884–899.
Carter, P., 2005. Keepin' It Real: School Success beyond Black and White. Oxford University Press, New York.
Collins, R., 1998. The Sociology of Philosophies: A Global Theory of Intellectual Changes. Harvard University Press, Cambridge.
Crane, D., 2000. Fashion and Its Social Agendas: Class, Gender, and Identity in Nineteenth and Twentieth Century Clothing. University of Chicago Press, Chicago.
DiMaggio, P., 1987. Classification in art. American Sociological Review 52, 440–455.
DiMaggio, P., Mohr, J., 1985. Cultural capital, educational attainment, and marital selection. American Journal of Sociology 90, 1231–1261.
DiMaggio, P., Mukhtar, T., 2004. Arts participation as cultural capital in the United States, 1982–2002: signs of decline? Poetics 32, 169–194.
Douglas, M., 1966. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Routledge and Kegan Paul, London.
Douglas, M., 1970. Natural Symbols. Vintage, New York.

Du Bois, W.E.B., 1903. The Souls of Black Folks. Bantam Classic, New York.
Durkheim, E., 1911. Les formes élementaires de la vie religieuse. Alcan, Paris (1965. The Elementary Forms of Religious Life. Free Press, New York).
Durkheim, E., Mauss, M., 1903. De quelques formes primitives de classification: contribution à l'étude des représentations collectives. Année Sociologique 6, 1–72 (1963. Primitive Classification. University of Chicago Press, Chicago).
Edgell, P., Gerteis, J., 2006. Atheists as 'other': moral boundaries and cultural membership in American society. American Sociological Review 71, 211–234.
Edgell, P., Tranby, E., 2010. Shared visions? Diversity and cultural membership in American life. Social Problems 57, 175–204.
Elias, N., 1939. Uber den Prozess der Zivilisation. Soziogenetische und Psychogenetische Untersushungen, vols. 2. Haus Zum Flakin, Basel (1982. The Civilizing Process. Blackwell, New York).
Elias, N., Scotson, J.L., 1965/1994. The Established and the Outsiders: A Sociological Enquiry into Community Problems. Sage Publications, London.
Epstein, C.F., 1992. Tinker-bells and pinups: the construction and reconstruction of gender boundaries at work. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago.
Erickson, B., 1996. Culture, class, and connections. American Journal of Sociology 102, 217–251.
Gamson, W., 1992. Talking Politics. Cambridge University Press, New York.
Gaztambide-Fernandez, R., 2009. The Best of the Best: Becoming Elite at an American Boarding School. Harvard University Press, Cambridge.
Gieryn, T.F., 1999. Cultural Boundaries of Science: Episodes of Contested Credibilities. University of Chicago Press, Chicago.
Gramsci, A., 1971. Selections from the Prisons Notebooks. Lawrence and Wishart, London.
Gusfield, J., 1963. Symbolic Crusade. University of Illinois Press, Urbana.
Hall, J.R., 1992. The capital(s) of cultures: a non-holistic approach to status situations, class, gender, and ethnicity. In: Lamont, M., Fournier, M. (Eds.), Cultivating Differences: Symbolic Boundaries and the Making of Inequality. University of Chicago Press, Chicago.
Halle, D., 1993. Inside Culture: Art and Class in the American Home. University of Chicago Press, Chicago.
Jasper, J., 1997. The Art of Moral Protest: Culture, Biography, and Creativity in Social Movements. University of Chicago Press, Chicago.
Jenkins, R., 1996. Social Identity. Routledge, London.
Johnston, J., Baumann, S., 2010. Foodies: Democracy and Distinction in the Gourmet Foodscape. Routledge, New York.
Kalmijn, M., 1991. Status homogamy in the United States. American Journal of Sociology 97, 496–523.
Khan, S., 2011. Privilege: The Making of an Adolescent Elite at St. Paul's Boarding School. Princeton University Press, Princeton.
Lacy, K., 2007. Blue-Chip Black: Race, Class, and Status in the New Black Middle-class. University of California Press, Berkeley.
Lamont, M., 1992. Money, Morals, and Manners: The Culture of the French and American Upper-Middle Class. University of Chicago Press, Chicago.
Lamont, M., 2000. The Dignity of Working Men: Morality and the Boundaries of Race, Class, and Immigration. Harvard University Press, Cambridge (Russell Sage Foundation, New York).
Lamont, M., 2010. Looking back at Bourdieu. In: Silva, E., Warde, A. (Eds.), Cultural Analysis and Bourdieu's Legacy: Settling Accounts and Developing Alternatives. Routledge, London.
Lamont, M., 2012. How has Bourdieu been good to think with? The case of the United States. Sociological Forum 27, 228–237.
Lamont, M., Lareau, A., 1988. Cultural capital: allusions, gaps, and glissandos in recent theoretical developments. Sociological Theory 6, 153–168.
Lamont, M., Mizrachi, N., 2012. Ordinary people doing extraordinary things: responses to stigmatization in comparative perspective. Ethnic and Racial Studies 35, 365–381.
Lamont, M., Molnar, V., 2002. The study of boundaries in the social sciences. Annual Review of Sociology 28, 167–195.
Lamont, M., Thévenot, L. (Eds.), 2000. Rethinking Comparative Cultural Sociology: Polities and Repertoires of Evaluation in France and the United States. Cambridge University Press, New York (Presses de la Maison des Sciences de l'Homme, Paris).
Lamont, M., Welburn, J., Fleming, C., 2012. Varieties of responses to stigmatization: macro, meso, and micro dimensions. Du Bois Review 9, 43–49.
Lareau, A., 2003. Unequal Childhoods: Class, Race, and Family Life. University of California Press, Berkeley, CA.
Lena, J.C., 2012. Banding Together. Princeton University Press, Princeton.
Lichterman, P., 1999. Civic Culture Meets Welfare Reform: Religious Volunteers Reaching Out. Paper presented at the annual meetings of the American Sociological Association. August, Chicago.
Luker, K., 1984. Abortion and the Politics of Motherhood. University of California Press, Berkeley.
Marx, K., Engels, F., 1848/1960. Die Deutsche Ideologie. Dietz, Berlin (1979. The German Ideology. Lawrence and Wishart, London).
Massey, D., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Cambridge University Press, New York.
Newman, K., 1999. No Shame in My Game: The Working Poor in the Inner City. Knopf and the Russell Sage Foundation, New York.
Pachucki, M., 2012. Classifying quality: cognition, interaction, and status appraisal of art museums. Poetics 40, 67–90.
Pachucki, M., Pendergrass, S., Lamont, M., 2007. Boundary processes: recent theoretical developments and new contributions. Poetics 35, 331–351.
Parkin, F., 1979. Marxism and Class Theory: A Bourgeois Critique. Columbia University Press, New York.
Pattillo, M., 2007. Black on the Block: The Politics of Race and Class in the City. University of Chicago Press, Chicago.
Peterson, R.A., Kern, R., 1996. Changing highbrow taste: from snob to omnivore. American Sociological Review 61, 900–907.
Peterson, R.A., Rossman, G., 2008. Changing arts audiences: capitalizing on omnivorousness. In: Ivey, B., Tepper, S. (Eds.), Engaging Art: The Next Great Transformation of Americans' Cultural Life. Routledge, New York.
Rivera, L., 2008. Managing 'spoiled' national identity: war, tourism, and memory in Croatia. American Sociological Review 73, 613–634.
Rivera, L., 2012. Hiring as cultural matching: the case of elite professional service firms. American Sociological Review 77, 999–1022.
Roth, W., 2012. Race Migrations: Latinos and the Cultural Transformation of Race. Stanford University Press, Stanford.
Schwartz, B., 1981. Vertical Classification: A Study in Structuralism and the Sociology of Knowledge. University of Chicago Press, Chicago.
Silver, H., Zwerling, C., 1992. Race and job dismissals in a federal bureaucracy. American Sociological Review 57, 651–660.
Tajfel, H., Turner, J., 1985. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Nelson-Hall, Chicago.
Tavory, I., 2010. Of yarmulkes and categories: delegating boundaries and the phenomenology of interactional expectation. Theory and Society 39, 49–68.
Tilly, C., 2004. Social boundary mechanisms. Philosophy of the Social Sciences 34, 211–236.
Veblen, T., 1899/1979. The Theory of the Leisure Class. Penguin, New York.
Wagner-Pacifici, R., 2000. Theorizing the Standoff: Contingency in Action. Cambridge University Press, New York.
Wagner-Pacifici, R., 2010. Theorizing the restlessness of events. American Journal of Sociology 115, 1351–1386.
Warikoo, N., 2011. Balancing Acts: Youth Culture in the Global City. University of California Press, Berkeley, CA.
Weber, M., 1922/1956. Wirtschaft und gesellschaft. grundisse der verstehenden soziologie. Hohrn, Tubingen (1978. Economy and Society, vol. 1. University of California Press, Berkeley).
Wimmer, A., 2013. Ethnic Boundary Making: Institutions, Power, Networks. Oxford University Press, New York.
Wimmer, A., Lewis, K., 2010. Beyond and below racial homophily: ERG models of a friendship network documented on facebook. American Journal of Sociology 116, 583–642.
Wright, E., Cho, D., 1992. The relative permeability of class boundaries to cross-class friendships: a comparative study of the United States, Canada, Sweden, and Norway. American Sociological Review 57, 85–102.
Wuthnow, R., 1987. Meaning and Moral Order: Explorations in Cultural Analysis. University of California Press, Berkeley.
Young, A., 2006. The Minds of Marginalized Black Men: Making Sense of Mobility, Opportunity, and Future Life Chances. Princeton University Press, Princeton.
Yukich, G., 2010. Boundary work in inclusive religious groups: constructing identity; at the New York Catholic Worker. Sociology of Religion 71, 172–196.
Zerubavel, E., 1997. Social Mindscapes: An Invitation to Cognitive Sociology. Harvard University Press, Cambridge.
Zolberg, A.R., Woon, L.L., 1990. Why Islam is like Spanish: cultural incorporation in Europe and the United States. Politics and Society 27, 5–38.

Symbolic Interaction: Methodology

Andrea Fontana, University of Nevada Las Vegas, Las Vegas, NV, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15347–15350, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

This article first introduces the origins of symbolic interactionism (SI) and its early methodological developments. It then discusses the more recent influence of postmodernism on the methodological orientations of SI. It ends with a survey of recently developed methods of SI. It argues that recent methods diverge from Blumer's initial guiding ideas drastically, but the diversity among them may be exaggerated.

This article will examine the different types of methodology used by Symbolic Interactionism. The bulk of the contribution will outline Blumer's qualitative approach, based on Mead's theories. Blumer coined the term symbolic interaction and suggested participant observation as the preferred method to study interaction. The article will also discuss briefly the method attempted by the Iowa School of Symbolic Interactionism, which was psychologically based. Finally, it will outline the new postmodern-informed, interactionist method used by a growing number of symbolic interactionists led by Denzin. Mention will be made of radical interactionist methods suggested by Ellis, Clough, and Richardson, among others.

The term 'symbolic interactionism' (SI) was coined in an article by Blumer (1937). SI presented a radical break from the traditional, macrostructural theories of sociology. SI rejected the scientific method and its reliance on predetermined hypotheses. SI's goal was not predicting about society, as structuralist methodologies did, but understanding the nature of the social world. In basing its research in the real world of everyday life, SI relied heavily on the theoretical work of the pragmatists, especially that of Mead (1934).

Blumer and Naturalistic Research

In his book containing reprints of his earlier articles, Blumer (1969) applies a methodology, naturalistic research, to the theoretical work of Mead. In symbolic interactionism, theory and methodology are intertwined very closely. Mead's social pragmatism advocated a study of society that takes place in small group interaction; namely, the relation of self and others. This reciprocal interaction is derived, for Mead, through the use of symbols, primarily language but others as well. Thus, society, for Mead, is a consensual, intersubjective world where the sharedness of meanings among its members allows some stability in an ever-changing flux (cf Adler and Adler, 1980).

Blumer, who was first a student and then a colleague of Mead at the University of Chicago, followed the latter's theoretical ideas and stated that society exists in action, in the reciprocal exchange of symbols, and the meaning these have for the members of society. If society is to be found in the interactional processes of its members we should study it in 'the real world,' i.e., the world created by societal members in their interaction, rather than in artificial settings or hypothetical situations. For Blumer, it follows that rather than constraining the study of social processes with controlled variables we should proceed with minimal intervention by the researcher. Also, in trying to maintain the integrity of the phenomenon under investigation we should not fit the 'real world' to our method, as typical in survey research, but should adapt out methodology to the world and thus use flexible data collection strategies. Finally, the model of the study should not be predetermined and then tested, but should be ever-changing to fit the changes and events in the real world.

For Blumer the naturalistic research has two phases: exploration and inspection. Exploration allows the researcher to gain a clearer understanding of the study, what methods to use, and to conceptualize the area of study in order to be able to talk from facts rather than speculations (Blumer, 1969). Exploration should be flexible and change as necessary to understand the topic at hand and thus should rely on a multiplicity of methods including observation, interviewing, life histories, and personal documents. Exploration is the first step of naturalistic inquiry for Blumer but must be complemented by another, more analytic step, inspection. This latter step aids the researcher in developing an analytic understanding of the elements of inquiry and their relations. This understanding is sensitizing rather than definitive, as the world studied is a process, i.e., ever changing. Furthermore, inspection allows for the testing of theoretical propositions. Thus, naturalistic research uses exploration to define the parameters of the study, then turns to inspection to develop sensitizing concepts about the study, finally developing theoretical concepts about society and its members (Hammersley, 1989).

The Chicago School

Blumer mentioned some works of the Chicago School and particularly *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 1918–20) as exemplars of the naturalistic method. Indeed, many works of the Chicago School are considered the antecedents of SI, as they tended to rely on some facet of the naturalistic method. Both Mead and Blumer were at the University of Chicago in the 1920s and 1930s when the school had great influence on American sociology. Among many other important contributors to the naturalistic studies of the Chicagoans, besides W.I. Thomas, who was one of the

theoretical leaders (but left as early as 1918), Robert Park must be mentioned. He was a former journalist trained in philosophy and took some courses from the pragmatists, John Dewey and William James. From his previous profession Park brought an understanding of studying the real world 'out there,' thus, being the inspirator of many studies of naturalistic nature. It must be noted that the Chicago School has been criticized for saying very little about their methods in their writings (Hammersley, 1989) and for not really being a school of thought at all (Becker, 1999). Some of the most notable works from the Chicago naturalistic tradition are *The Ghetto* (Wirth, 1928), *The Delinquent Gang* (Thrasher, 1927), *The Gold Coast and the Slum* (Zorbaugh, 1929), *The Hobo: The Sociology of the Homeless Man* (Anderson, 1923). For a detailed study of the Chicago School see Faris (1967) and Bulmer (1984).

Other Methods Based on Mead

While Blumer's adaptation of Mead's theories is the methodological mainstay of SI, there are other methodologies based on SI, and these will be mentioned next. Analytic induction was first discussed by Znaniecki (1928). It was later used, with minor variations by Lindesmith (1937, 1968) (he was a graduate student of Blumer), Cressey (1950) (a student of Lindesmith), Becker (1963) (see Hammersley, 1989), and others. Analytic induction, according to Znaniecki, recognizes the fact that objects in the world are open to an infinite number of description and, thus, our account of them must be selective; this selectivity will be based on the interest at hand, which for sociologists is primarily social and cultural systems; commonly used sociological methods relying on preidentification (deductive) or superficial description (inductive) will not work, only analytic induction will accomplish the task. The researcher will select a small number of cases (10–12, usually) and study them in depth, continually defining and redefining the event and formulating and reformulating theoretical propositions until they will fit all cases. Negative cases must also be examined (this was Lindesmith's idea).

Another student of Blumer, Strauss, together with Glaser, developed another SI method, grounded theory (Glaser and Strauss, 1967). Closely related both to Blumer's methodology and to analytic induction, grounded theory placed more emphasis on the generation and development of theory. Relying on the inductive method, grounded theory is akin to Blumer's inspection, only much more elaborate. Rather than relying on a priori population, in analytic theorizing one continues to study new cases until the point of saturation, generating theoretical categories.

Not all SI methods followed the constructionist approaches outlined above. A notable exception came from the Iowa School of Sociology. Kuhn (1964) adopted a much more deterministic approach to Mead's discussion of the self and the nature of the 'me,' the various roles and images we have of ourselves. The methodology he adopted to discover the nature of the self was called the Twenty Statements Test (TST), a series of open-ended questions about the self. Kuhn felt that rather than use the oblique method of observing people one ought to ask them directly about the nature of their inner feelings and they would honestly disclose them to the researcher. The results of TST would be used, by Kuhn, to outline generic laws that would apply to human beings in different situations. Other positivistic oriented symbolic interactionists are Sheldon Stryker, described as a 'structural role theorist,' who influenced numerous students at the University of Indiana and Carl Couch, who was a stalwart of the discipline, with his 'Behavioral Sociology' at the University of Iowa (cf Reynolds, 1993).

In the 1960s and 1970s, a plethora of theoretical approaches, largely based on the naturalistic method, appeared. Some were based on basic Meadian tenets, such as dramaturgy (Goffman, 1959), and labeling (Becker, 1963). Others based their constructionist approach not only on the ideas of Mead but on those of the phenomenologists (Husserl, Schutz, Heidegger, Dilthey) and the existentialists (Merleau-Ponty, Sartre), and ordinary language philosophers (Wittgenstein). They are phenomenological sociology, existential sociology, ethnomethodology, and the sociology of emotions (see Douglas et al. (1980) for a survey of these sociologies and a list of references to them; also, see Adler et al. (1987), Moran (2002), and Grossman (2013)).

Postmodern-Informed Symbolic Interactionism

More recently, with the influence of postmodernism on the social sciences, the methodology of SI has undergone some dramatic changes (Marcus and Fischer, 1986; Dickens and Fontana, 1994). Again, the closeness of theory and methodology in SI affects methodological approaches to SI. Postmodernism (Lyotard, 1984) advocates a disenchantment with metatheories, claiming all we can possibly know about society is much smaller in scope, and it is relative due to its historical, temporal, regional nature, as well as other elements. Thus, society becomes fragmented and we should question and deconstruct all overarching metatheories and reduce the scope and intent of our studies to our immediate surroundings, to the details of everyday life.

These beliefs impact the naturalistic method of SI in a number of ways. First, we can no longer claim to be aiming at universal laws of human behavior as that would be a metatheory. Second, metatheories come under scrutiny and are deconstructed, i.e., their basic assumptions are rendered problematic and questioned. This leads to the questioning of the role of the researcher and author in terms of authority of selectively choosing data to record and report, thus biasing the study. Postmodern-informed interactionism attempts to remedy this authorial bias by the use of polyphony, by this meaning to allow the many voices of the respondents to speak for themselves. Additionally, researchers disclose as much information as possible about themselves – role, interest, possible bias, and more. This is only a partial solution to the authorial influence, which is ultimately, irremediable, as we cannot help but to perceive and report selectively.

Other methodological changes take aim at the relation between the researcher and the researched. Traditionally, interactionists have attempted to remain as neutral and invisible as possible in order not to influence the study. Postmodern-informed interactionists make their presence felt

in an attempt to establish a one-to-one partnership with the respondents. In so doing researchers are no longer aloof observers collecting data but become a part of the interaction and influence the research findings. Now data stems from a negotiated process between researcher and respondent (Gubrium and Holstein, 1997) (notice that this goal contradicts the previous goal of reducing authorial bias).

Another change involves the notion of 'value free' research. Some postmodern-informed interactionists wish to overthrow traditional sociology (Clough, 1998; Denzin, 1997, 1999a, 2001; Fontana and McGinnis, 2003) which they consider patriarchal, instead they envision a new researcher who is a "feminist, communitarian researcher … who builds collaborative, reciprocal, trusting, and friendly relations with those studied" (Denzin, 1997).

Current Trends

Postmodern-informed interactionists not only have made dramatic methodological changes from Blumerian interactionism but are experimenting with radically different reporting styles, which are closely intertwined with the new methodological ethos. Some of these styles include poetry and performances, as well as the writing of short stories about one's personal past (autoethnographies). While some of these reports are highly engaging (Richardson, 1997; Denzin, 1999b), they raise a question about standards – should these works be judged and evaluated by the standards of sociology or poetry and literature? Some interactionists have been highly critical of postmodern-informed interactionism (Seidman, 1991; Shalin, 1993; Best, 1995; Prus, 1996; Perrotta, 2004). As a result the field of interactionism is currently divided in many groups with different methodological approaches.

Adler and Adler (1999), in describing the varied groups of interactionists at the turn of the millenium, use the metaphor of 'the ethnographers' ball' and sit the factions at various tables. We will rely upon the Adlers' description in the following section. The first table sits the 'postmodern-informed' ethnographers, next to them is the 'autoethnography group,' a bit further down sit the 'interpretive discourse analysts' who attempt to integrate interactionism, ethnomethodology, and postmodernism. At another table the Adlers have the 'classic interactionists,' in the wake of the old Chicago School. Next to them sit the 'no mo pomo' (Adler and Adler, 1999), the Blumerian group critical of the new postmodern influence. Near by are the followers of Erving Goffman, the 'dramaturgists,' studying everyday life by using the metaphor of the theater. They are followed by the 'grounded theorists,' next to the 'conversational analysts' who are continuing the work of ethnomethodologists such as the late Harvey Sacks and others. Next, the Adlers have a table for those who practice 'the extended case study,' using ethnography to do applied research. Finally we see the 'visual interactionists' who capture symbolic interaction through the eye of the camera lens. As the Adlers themselves say, 'the list can go on' (Adler and Adler, 1999).

We would add to the Adlers' list three more tables, before we run out of room in the ballroom. One would be the 'phenomenological/existential' table placed near Jack Douglas and his former students at the San Diego campus of the University of California. Another table, right next to the postmodernists, would be the 'feminist interactionists' who are 'united in their criticisms of "Eurocentric masculinist approaches"' (Collins, quoted in Denzin, 1997). Finally, there would be a table of 'theoretical interactionists' who have written extensively in the area without actually engaging in ethnographic work (cf Lyman and Scott, 1989).

The methodology of Symbolic Interactionism has become the methodologies of symbolic interactionism, and can no longer be grouped solely under Blumer's guiding ideas, yet their diversity is exaggerated and exacerbated by rivalry among various subgroups (Fontana, 2005). By and large, while differing in some methodological points, the great majority of interactionists still wish to study the real world and the meaning that its members make of it and how they manage to maintain and achieve social order as a negotiated interactional accomplishment.

See also: Ethnomethodology, General; Explanation: Conceptions in the Social Sciences; Grounded Theory: Methodology and Theory Construction; Interactionism, Symbolic; Participant Observation; Postmodernism: Methodology; Postmodernism: Philosophical Aspects; Qualitative Methods, History of.

Bibliography

Adler, P., Adler, P., 1980. Symbolic interactionism. In: Douglas, J.D., Adler, P.A., Adler, P., Fontana, A., Freeman, C.R., Kotarba, J.A. (Eds.), Introduction to the Sociologies of Everyday Life. Allyn and Bacon, Boston.

Adler, P., Adler, P., 1999. The ethnographers' ball revisited. Journal of Contemporary Ethnography 28 (5), 442–450.

Adler, P., Adler, P., Fontana, A., 1987. Everyday life sociology. Annual Review of Sociology 13, 217–235.

Anderson, N., 1923. The Hobo: The Sociology of the Homeless Man. University of Chicago Press, Chicago.

Becker, H.S., 1963. Outsiders: Studies in the Sociology of Deviance. Free Press of Glencoe, London.

Becker, H., 1999. The Chicago school, so-called. Qualitative Sociology 22 (1), 3–12.

Best, J., 1995. Lost in the ozone again. Studies in Symbolic Interaction 17, 120–130.

Blumer, H., 1937. Social psychology. In: Schmidt, E.P. (Ed.), Man and Society. Prentice-Hall, New York.

Blumer, H., 1969. Symbolic Interactionism: Perspective and Method. Prentice-Hall, Englewood Cliffs, NJ.

Bulmer, M., 1984. The Chicago School of Sociology. University of Chicago Press, Chicago.

Clough, P.T., 1998. The End(s) of Ethnography: From Realism to Social Criticism, second ed. Peter Lang, New York.

Cressey, D., 1950. The criminal violation of financial trust. American Sociological Review 15, 738–743.

Denzin, N.K., 1989. Interpretive Interactionism. Sage, Newbury Park, CA.

Denzin, N.K., 1997. Interpretive Ethnography: Ethnographic Practices for the 21st Century. Sage, Thousand Oaks, CA.

Denzin, N.K., 1999a. Interpretive ethnography for the next century. Journal of Contemporary Ethnography 28 (5), 510–519.

Denzin, N.K., 1999b. Performing Montana. In: Glassner, B., Hertz, R. (Eds.), Qualitative Sociology as Everyday Life. Sage, Thousand Oaks, CA.

Denzin, N.K., 2001. Interpretive Interactionism. Sage, Thousand Oaks.

Dickens, D.R., Fontana, A., 1994. Postmodernism and Social Inquiry. The Guilford Press, New York.

Douglas, J.D., Adler, P.A., Adler, P., Fontana, A., Freeman, C., Kotarba, J.A. (Eds.), 1980. Introduction to the Sociologies of Everyday Life. Allyn and Bacon, Boston.

Faris, R.E.L., 1967. Chicago Sociology 1920–1932. Chandler Pub. Co., San Francisco.
Fontana, A., McGinnis, T.A., 2003. Ethnography since postmodernism. Studies in Symbolic Interaction 26, 215–234.
Fontana, A., 2005. The postmodern turn in interactionism. Studies in Symbolic Interaction 28, 239–254.
Glaser, B.G., Strauss, A.L., 1967. The Discovery of Grounded Theory. Aldine, Chicago.
Goffman, E., 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City, New York.
Grossman, R., 2013. Phenomenology and Existentialism: An Introduction. Routledge, New York.
Gubrium, F., Holstein, J.A., 1997. The New Language of Qualitative Methods. Oxford University Press, New York.
Hammersley, M.H., 1989. The Dilemma of Qualitative Methods: Herbert Blumer and the Chicago Tradition. Routledge, London.
Kuhn, M., 1964. Major trends in symbolic–interaction theory in the past twenty-five years. Sociological Quarterly 5, 61–84.
Lindesmith, A.R., 1937. The Nature of Opiate Addiction. The University of Chicago Libraries, Chicago.
Lindesmith, A., 1968. Addiction and Opiates. Aldine, Chicago.
Lyman, S., Scott, M., 1989. Sociology of the Absurd, second ed. Harcourt Brace, New York.
Lyotard, J.-F., 1984. The Postmodern Condition: A Report on Knowledge (G. Bennington, B. Massumi, Trans.). University of Minnesota Press, Minneapolis, PA.
Marcus, G.E., Fischer, M.M.J., 1986. Anthropology as Cultural Critique: An Experimental Moment in the Human Sciences. University of Chicago Press, Chicago.
Mead, H.G., 1934. Mind, Self and Society. University of Chicago Press, Chicago.
Moran, D., 2002. Introduction to Phenomenology. Routledge, New York.
Perrotta, R., 2004. Mirror, mirror on the wall, who is the fairest of them all? Studies in Symbolic Interaction 27, 297–301.
Prus, R., 1996. Symbolic Interaction and Ethnographic Research. State University of New York Press, New York.
Reynolds, L., 1993. Interactionism: Exposition and Critique. General Hall, Dix Hills, New York.
Richardson, L., 1997. Fields of Play: Constructing an Academic Life. Rutgers University Press, New Brunswick, NJ.
Sanders, C., 1995. Stranger than fiction: insights and pitfalls in postmodern ethnography. Studies in Symbolic Interaction 17, 89–104.
Schmidt, E.P. (Ed.), 1937. Man and Society. Prentice-Hall, New York.
Seidman, I.E., 1991. Interviewing as Qualitative Research. Teachers College Press, New York.
Shalin, D., 1993. Modernity, postmodernism and pragmatist inquiry: an introduction. Symbolic Interaction 16, 303–332.
Thomas, W.I., Znaniecki, F., 1918–20. The Polish Peasant in Europe and America, vols.1–5. University of Chicago Press, Chicago.
Thrasher, F.M., 1927. The Gang: A Study of 1,313 Gangs in Chicago. University of Chicago Press, Chicago.
Wirth, L., 1928. The Ghetto. University of Chicago Press, Chicago.
Znaniecki, F., 1928. Social research in criminology. Sociology and Social Research 12, 307–322.
Zorbaugh, H.W., 1929. The Gold Coast and the Slum. University of Chicago Press, Chicago.

Symbolism in Anthropology

Janet Hoskins, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the importance of symbolism in anthropology, noting its role in the development of a 'school' of symbolic anthropology in the 1960s and 1970s, and the various positive and critical responses to this embrace of the ethnographic study of symbols. It concentrates on the works of Clifford Geertz, Victor Turner, David Schneider, Mary Douglas, and Sherry Ortner, since each of these thinkers has been seminal in developing this approach, although many others could also be included. I also explore the 'genealogical' criticism of the notions of religion and ritual as symbolic systems, particularly as articulated by Talal Asad.

This article explores the importance of symbolism in anthropology, noting its role in the development of a 'school' of symbolic anthropology in the 1960s and 1970s, and the various positive and critical responses to this embrace of the ethnographic study of symbols. It concentrates on the works of Clifford Geertz, Victor Turner, David Schneider, Mary Douglas, and Sherry Ortner, since each of these thinkers has been seminal in developing this approach, although many others could also be included. I will argue that Turner developed an approach to analysis of symbolism in ritual (defined most notably in his 1975 review article on *Symbolic Studies*), while Geertz explored symbols as 'vehicles of culture.' Schneider studied culture itself as a symbolic system, and Mary Douglas looked at symbols of purity and pollution as demarcating problems in social classification. Symbolic anthropology was more a number of shared concerns than a unified school, and some have argued that rather than being in constructive dialogue or conflicting with each other they managed only to 'talk past one another' (Ortner, 1984: p.131). But the tools developed to analyze and argue about symbols have remained useful in anthropological discussions into the twenty-first century.

In the final section of this article, I will explore what could be called the 'genealogical' criticism of the notions of religion and ritual as symbolic systems, particularly as articulated by Talal Asad. I will assess whether or not this criticism is based on a certain misunderstanding of the properties of symbolism as studied by anthropologists. I will also cite recent developments in relating the study of symbolism to theories of practice, especially Pierre Bourdieu's idea of a 'religious field of practice.' The study of symbolism in anthropology has shifted in new directions in the twenty-first century, orienting itself more toward ideas of discipline and control, and less toward the free play of multivocal symbols. But the topic has remained important and has shaped the history of the discipline.

Defining Symbolism

The Oxford English Dictionary defines a symbol as "a thing regarded by general consent as naturally typifying or representing or recalling something by possession of analogous qualities or by association in fact or thought" (Victor Turner, 1975: p.151). Applying this concept to anthropological analysis, the idea of 'general consent' would have to be limited to a particular cultural context, since meanings are in many senses culture bound. Symbolism has been central to definitions of culture ever since Kroeber and Kluckhohn wrote "culture considers of patterns explicit and implicit, of and for behavior acquired and transmitted by symbols, constituting the distinctive achievement of human groups" (Kroeber and Kluckhohn, 1952: p. 357). A quarter century later, Clifford Geertz echoed this definition with his own formulation of the culture concept as "an historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic form by means of which men communicate, perpetuate, and develop their knowledge about and attitudes toward life" (Geertz, 1973: p. 89).

Victor Turner has more than any other single anthropologist to define the forms and functions of symbols, and their relation to language. He argues that it is important to distinguish a symbol from a sign. A symbol can have multiple meanings and it signifies through a more complex process than a sign. The sensorily perceptible vehicle or 'outward form' is the signifier, and the meaning, sense, or designation is the signified. For signs, the signified is relatively straightforward, while for symbols it may be more complex and ambiguous. The linguistic sign signifies in ways that are arbitrary and conventional, while a symbol has a metaphoric or metonymic relationship to what it designates. Signs tend to univocality and are often deliberate constructs for precise communicative purposes, while symbols tend to be iconic, "where a sensory likeness relation is intended or interpreted" (Turner, 1975: p.152).

Turner defines the properties of symbols – multivocality, complexity of association, ambiguity, open-endedness, primacy of feeling, and a propensity to ramify into further semantic systems – as connected with their dynamic quality (Turner, 1975: p.155). Symbols can trigger social action and can give personal action form in the public arena. From Turner's perspective, "we master the world through signs, ourselves by symbols" (Turner, 1975: p.159). Turner developed a view of symbols as operators in the social process, which could produce social transformations,

particularly in the context of ritual. He applied this analysis to understanding Ndembu ritual, such as the temporary shrine to hunter-ancestors called the *chishing'a*. The *chishing'a* is a three or four pronged branched whose ends are sharpened to represent the 'sharpness' of huntsmanship. The bark is peeled to reveal its auspicious inner quality of whiteness associated in Ndembu thought with milk, and hence fertility and nurturance. Braided grass is tied under lowest prong to represent animals and ancestral shades who come up from the red grave to drink the blood offered to them by the hunters. These three aspects "form part of a complex verbal hermeneutic which encompasses the innumerable nonverbal symbols of Ndembu culture in webs of explicit significance" (1975: p.154).

Turner developed and elaborated this analysis of symbols not only to the Ndembu, but also to a wide range of other rituals, which he distinguished as 'performances' (Turner 1967, 1974, 1986, 1987). Theater was "one of the many inheritors of that great multifaceted system of preindustrial ritual that embraces ideas and images of cosmos and chaos, interdigitates clowns and their foolery with gods and their solemnity, and uses all the sensory codes, to produce symphonies in more than music" (1990: p.12). Ritual teases the senses as part of a process similar to what Turner had earlier called a 'social drama': There are four phases – a breach (of social rules), a crisis, redress, and either reintegration or recognition of schism (Turner, 1987: pp. 34–35). The stylized, scripted action of ritual defines a creative tension between "what should be performed" and "what actually happens," providing the potential for both continuity and change, since ritual is a way for members of a society to reflect on their own actions. As Turner has put it, "cultural performances are not simple reflectors or expressions of culture or even of changing culture but may themselves be active agencies of change, representing the eye by which culture sees itself" (Turner, 1986: 24). If, as Geertz argues, anthropological interpretation is defined as 'constructing a reading of what happens,' then ritual is one of the best places to isolate and analyze that 'reading' (Turner, 1975: p.148).

Religious Symbols and Cultural Systems

Turner's explicit acknowledge of Clifford Geertz in this final phrase signals the ways in which his approach is compatible with, but quite distinct from, the more cultural approach to symbolic interpretation, which Geertz has proposed. While Turner sees symbols as "instrumentalities of various forces – physical, moral, economic, political, and so on – operating in isolable, changing fields of social relationships" (Turner, 1975: p.145), Geertz sees them as part of a cultural system. Turner studied symbols as they are is performed in ritual, where he identified forms of dramas, fields, and metaphors in a series of detailed studies moving from the Ndembu out to a wide range of other societies. His famous notion of liminality (an 'in-between status') is developed through an analysis of the ritual process, in which the meaning of particular symbols is constituted through their effectiveness in producing transformations in rites of passage. He then extended the drama metaphor to studies in the 'anthropology of performance,' suggesting that that non-Western societies were cyclically enacting a dramatic script in successive performances of pilgrimages, rites of affliction, and calendrical festivals.

Geertz innovated by arguing that culture is not something private, hidden away inside individuals, but is embodied in public symbols. He defined his new paradigm for the study of religion by asserting that "sacred symbols function to synthesize a people's ethos – the tone, character, and quality of their life, its moral and aesthetic style and mood – and their world view – the picture they have of the way things in actuality are, their most comprehensive ideas of order" (Geertz, 1973: p. 89). The study of symbols was a way to understand the complex ways in which people communicate with each other and transmit their way of life to future generations. Geertz argued that the anthropologist should study 'the native's point of view,' as a sort of text that could read over the shoulders of native actors, and interpreted through an analysis that would make it intelligible to nonnatives. He defined anthropological interpretation, as 'constructing a reading of what happens,' and this cannot be done without having a sense of the whole to which it belongs. To divorce it from what happens "from what, in this time and place, specific people say, what they do, what is done to them, from the whole vast business of the world, is to divorce it from its applications and render it vacant" (Geertz, 1973: p.18).

The study of symbols for Geertz was, first and foremost, the study of religion: "Religious symbols formulate a basic congruence between a particular style of life and a specific (if often implicit) metaphysic, and in so doing sustain each with the borrowed authority of the other" (Geertz, 1973: p. 90). He also argued that symbols were part of a series of cultural patterns that provided both a 'sense of direction' and a 'power of self control,' adding "Man depends upon symbols and symbol systems with a dependence so great as to be decisive for his creatural viability" (Geertz, 1973: p. 99). Without them, we would be reduced to "a formless monster… a chaos of spasmodic impulses and vague emotions."

Religion as a cultural system creates 'an aura of utter actuality,' which all its symbolic activities are "devoted to producing, intensifying, and, so far as possible, rendering inviolable by the discordant revelations of secular experience" (1973: p.112). Ritual performances, like the elaborate Balinese dance drama opposing Rangda and Barong, which he analyzes, make "the model for and the model of aspects of religious belief mere transpositions of one another" (Geertz, 1973: p.118). Religious symbols offer a 'model of' the relations between the self and the world, and also a 'model for' how to act in this world, a set of distinctive 'mental dispositions' that motivate specific forms of action. Victor Turner argues that symbols can be used instrumentally, to trigger social action, while Geertz stresses the fact that they are first of all 'conceptions of a general order of existence' that provide both the map (the model of the world) and the specific instructions for the journey we should take across that territory.

Core Symbols and the Domain of Kinship

The attention that Clifford Geertz gave to religion as a locus for symbolism has been transferred to the domain of kinship in

the work of David Schneider. Schneider sought to study the internal logic of systems of symbols, and in particular the ways in which they might be organized around core symbols. In his methodology a symbolic system is studied by isolating its elements, specifying the internal relationships among those elements, and then characterizing the whole system in some way, stressing its core symbols, ideological principles, and the underlying structures of which is a surface expression. Culture is more radically cut off from social action in this approach, in order to explore the precise relationship between various terms. In an article examining kinship, nationality and religion in American culture, Schneider found that the three terms constituted a 'system' dependent

> on a distinction between the "pure" domain of kinship, defined in terms of the symbol of coitus and differentiated into two major aspects, relationship as natural substance and relationship as code for conduct, and a "conglomerate" domain of kinship, differentiated into "the family" on the one hand and an articulated system of person-defined statuses (genealogical) on the other. (Schneider, 1969: 123)

He found that each cultural domain contains a hierarchy of symbols devolving from a semantically bipolar core symbol: substance or natural order is opposed to, yet united with, code for conduct. The symbol of 'love' links conjugal and cognatic love and relates them to the central core symbol of coitus. The cultural notion of the "person" is "constructed out of symbols and definitions from various cultural domains … in relation to particular contexts and systems" (Schneider, 1969: p.17).

Researchers in this tradition should search for core symbols of 'diffuse, enduring solidarity' as centers of organization in cultural systems and subsystems of symbols. The dual aspects of relationship as natural substance and relationship as code for conduct are defined and structured in almost identical terms in the domains of nationality and religion as well as kinship.

Symbols, Classification, and Taboo

Anthropological studies of symbolism address what Geertz called 'the problem of meaning' in a number of different ways. The British anthropologist Mary Douglas stresses the power of classification to order the world and make it conceptually manageable. There have been a number of other theories of taboo. Relational theories, like that of Levi-Strauss in his work on *Totemism*, project the relations of society upon the relations in nature. Moral theories argue that taboos create dangers that reinforce the moral code. This was the approach taken by Freud in his analysis of Moses and monotheism. Her own approach is a 'housekeeping model,' which views pollution as an aid to 'imposing system' on 'an inherently untidy experience,' equating anything that does not fit into tidy conceptual categories as dirt, since "dirt is matter out of place." As she argued in 2000 (p. 2288): "Taboos stop the cosmos from disintegrating. It is a process of learning by not doing. The strange connections taboos make between actions and consequences are not due to faulty logic but to a formidable intellectual effort to link nature and morality into a single scheme."

In 1966, Douglas published her most celebrated work, *Purity and Danger: an Analysis of Concepts of Pollution and Taboo*. This book is best remembered for its stylish demonstration of the ways in which all schemes of classification produce anomalies: whether the pangolin for the Lele, or the God incarnate of Catholic theology. Some of this classificatory 'matter out of place' – from humble house dust in her Highgate house to the abominations of Leviticus for the Hebrews – was polluting, but other breaches of routine classification had the capacity to renew the world symbolically.

In *Natural Symbols* (1970), Douglas began to answer the question that arose from this: what kinds of society are best able to support the complex symbolism that makes the best society possible? The book took exception to some 1960s cultural trends – notably the liturgical reforms of the Second Vatican Council and student revolts – and presented the first of many attempts to correlate features of institutional organization with patterns of belief and morality. Her conclusion – that differentiated, hierarchical and bounded institutions provided the most conducive environments for complex thinking and symbolism – remained the same, whatever changes her theories underwent.

Mary Douglas affirmed that all human beings have a profound need for a theory of a general order of existence, and religious symbols fulfill that need. In *Purity and Danger* she wrote that it was almost inconceivable for a person in a traditional society to be without religion: "The person without religion would be the person content to do without explanations of certain kinds, or content to behave in society without a single unifying principle violating the social order" (1975, p. 76). The deep dread of disorder that she assumed was common to all human beings was also evoked in Geertz's discussion of why religions come into being: "there are at least three points where chaos – a tumult of events, which lack not just interpretations but *interpretatability* – threatens to break in upon man: at the limits of his analytic capabilities, at the limits of his powers of endurance, and at the limits of his moral insight" (1973: p.100). Religious symbols must work to meet these perceived threats of intellectual, physical, and moral dissolution.

While the problem of meaning affirmed "the inescapability of ignorance, pain, and injustice on the human plane," it also had to simultaneously deny that "these irrationalities are characteristics of the world as whole. And it is in terms of religious symbolism a symbolism relating to man's sphere of existence to a wider sphere within which it is conceived to rest, that both the affirmation and the denial are made" (Geertz, 1973: p.108). Faced with resolving this problem of meaning, Geertz asserts that "He who would know must first believe" (1973: p.110), so a commitment to a religious perspective begins as a state of mind rather than a constituting activity in the world.

The Genealogical Critique of Religion as a Symbolic System

Perhaps the most influential criticism of the paradigm of symbolic anthropology has come from Talal Asad's post-structuralist perspective. He begins by saying "the theoretical

search for an essence of religion invites us to separate it conceptually from the domain of power" (1993: p. 29). Using the Foucauldian notion of the genealogy, he traces a history of notions of religion that emphasizes power, discipline, and coercion more than symbolic expressiveness. Asad argues that Geertz's 'interior' view of religion (focusing on subjective states of belief or faith) and ritual is both Eurocentric and presentist: it has emerged from a modern conception of religion that is not applicable to many other societies, or even to the earlier history of Christianity. But Asad's attack on Geertz' approach as secretly 'Protestant,' since it focuses on the believing individual as the prime locus of religion, does not take account of the fact that one reason to formulate a 'symbolic' definition of religion was to avoid the more conventional definitions (widely used in religious studies), which require a textual tradition, an institutionalized set of religious specialists, and physical structures – 'books, clergy, and temples.'

Geertz was writing as an anthropologist who was determined to grant the term 'religion' to less institutionalized practices like shamanism, ancestor worship, spirit possession séances, and forms of 'popular religion' usually excluded from earlier definitions of religion as incoherent and premodern forms not yet worthy of the term. But placing his emphasis on shared symbols, he does not have to require written doctrine or a formalized set of scriptures, but can include religious activity that is more improvisational, context dependent, dynamic, and emerging.

Both Geertz and Asad have tried to escape the limits of the modern Christian perspective, but they have done so in different ways. The first significant difference lies in ideas of the concept of the symbol. Geertz says that "religion is a system of symbols" and he defines a symbol as any object, act, event, quality, or relation that serves as a vehicle for a conception – the conception is the symbol's 'meaning.' Asad, in contrast, argues that a symbol is "a set of relationships between objects and events uniquely brought together as complexes or as concepts, having at once an intellectual, instrumental, and emotional significance"(1993: p. 31).

This difference of opinion may seem technical and of little consequence, but it has great importance for the kinds of research questions that can be investigated. While Geertz sees a system of symbols in place, which has to be decoded by the researcher, Asad is interested in the question of how (under what conditions) symbols come to be constructed and accepted as natural and authoritative, while others are opposed. Asad argues that the most important object of anthropological inquiry is "the conditions (discursive and nondiscursive) that explain how symbols come to be constructed, and how some of them are established as natural or authoritative as opposed to others" (Asad, 1993: p. 31).

Geertz's argument that religious symbols act as models of and models for reality follows in the tradition of Durkheim and Talcott Parsons, but Asad rejects this implied isomorphism. He asks instead how religious power has created an idea of religious truth through a series of embodied practices that make religion more a form of social activity than a system of beliefs. Looking at the early history of the Christian church, Asad argues that Augustine used discipline to push his disciples to experience Christian truth: "It was not mind that moved spontaneously to religious truth, but power that created the conditions for experiencing that truth. Particular discourses and practices were to be systematically excluded, forbidden, denounced – made as much as possibly unthinkable; others were to be included, allowed, praised, and drawn into the narrative of sacred truth" (Asad, 1993: p.35).

Geertz and Asad also differ on the role of ritual. While Geertz sees ritual as a signifying activity distinct from instrumental behavior, Asad sees it as an "apt performance of what is prescribed, something that depends on intellectual and practical disciplines but does not itself require decoding" (Asad, 1993: p. 62). It is coded in a regulatory way rather than in a semantic way. Instead of opposing symbolic to instrumental behaviors, he looks at the varied disciplinary programs through which a subject with virtuous desires was created. Asad says the West's emphasis on possessive individualism is in contrast with Islam's communitarian emphasis on virtue making as a collective responsibility. Current ideas of religious privacy and individual autonomy are incompatible with Muslim ideas of communitarianism and discipline.

Asad concludes that no universal definition of religion can be established, since all such definitions are the historical product of culturally specific discursive practices. "It is not mere symbols that implant true Christian dispositions, but power," and this power is deployed in an "entire range of available disciplinary activities… through which the possibilities of attaining the truth are marked out" (p. 50). He also accuses Geertz of conflating cognitive and communicative issues.

The modern Christian perspective marginalizes "the authorizing process by which 'religion' is created" (p. 37) and overlooks the fact that one can be religious "without being able to articulate that knowledge" (p. 36). Geertz' interpretive theory is said to deny the material, coercive side of the state, and material aspects of religion, which have shaped embodied practices and the experience of religion.

Geertz vs Asad and Bourdieu on Public Symbols and Private Discipline

The role of belief vs practice is a key point of disagreement. Asad accuses Geertz of tying his notion of religion to a cognitive notion of belief, as "a set of propositions to which believers gave assent," which allows religion to be separated off into a private domain. "Geertz's treatment of religious belief…is a modern, privatized Christian one because and to the extent that it emphasizes the priority of belief as a state of mind rather than as a constituting activity in the world" (Asad, 1993: p. 47). Asad argues that this shows a "Christian bias": "It is preeminently the Christian church that has occupied itself with identifying, cultivating and testing belief as a verbalizable inner condition of true religion" (p. 48).

Geertz distinguishes religious from nonreligious practices by claiming that a religious practice is one which affirms (something) about the fundamental nature of reality: "What any particular religion affirms about the fundamental nature of reality may be obscure, shallow. Or all too often perverse, but it must, if it does not consist of the mere collection of received practices and conventional sentiments we usually refer to as moralism, affirm something" (Geertz, 1973: pp. 98–99,

Asad, 1993: p. 43). Asad argues that this definition tends to exclude the "unevangelized" who come to be seen "typically either as those who have practices but affirm nothing, in which meaning can be attributed to their practices (thus making them vulnerable), or as those who do affirm something (probably 'obscure, shallow, or perverse'), an affirmation that can therefore be dismissed" (Asad 43).

The claim may not hold water. 'Belief' itself is not even mentioned in Geertz' definition, although his idea of a 'model for' and conceptions 'clothed in an aura of factuality' might be taken as leading to a certain metaphysics. His examples of religious symbols ("elaborate initiation rites, dramatic shamanistic exhibitions, cruel human sacrifice rites, obsessive curing ceremonies, large communal feasts") are all focused on the religious practices of 'unevangelized' non-Western groups (Australian aborigines, the Maori, the Eskimo, the Navaho, Polynesian groups). Rather than accepting a modernist strategy that religion occupies the space of noncognitive feeling or experience, Geertz explicitly insists that religions are interested in teaching a world view.

Geertz also argues strenuously that symbols are public. He says culture is "not anybody's knowledge or belief about anything" (1973: p. 12). It is instead embodied in public acts: "Culture is public because meaning is. … Culture consists of social established structures of meaning in terms of which people do such things as sign conspiracies and join them or perceive insults and answer them" (Geertz, 1973: pp.12–13). His famous essay on *Thick Description* is primarily an attack on 'the cognitivist fallacy' that culture consists of mental phenomena, which can be analyzed by formal methods. Here, Asad seems to have fundamentally misunderstood Geertz' position. After all, it is Geertz – not Asad – who argues that "cultural acts, the construction, apprehension and utilization of symbolic forms, are social events like any other" (Geertz, 1973: p. 91).

In a later essay (which could be interpreted as a veiled rejoinder to Asad), Geertz argues that religion is "the symbolic contrivances by means of which individuals imagined themselves as persons, as actors, sufferers, knowers, judges, as, to introduce the exposing phrase, participants in a form of life" (Geertz, 2000: p.15). The crux of Asad's argument is that one cannot assume a symbolic system separate from practices. Geertz' famous position that culture can be studied as an 'ensemble of texts' implied that these cultural elements are disembodied, abstracted from action. Drawing on the hermeneutic philosophy of Paul Ricoeur and a Weberian vocabulary of meaning, pattern and interaction, Geertz proposes that we can study the systems of meaning on their own. To his critics, this seems to detach meaning from practice, leading Bourdieu (1977, p. 96) to criticize the hermeneutic stance as 'objectivism.'

Alternatives to Symbolism: A Religious Field of Practice

What alternative approaches to the study of religion would allow us to study symbols in the context of relations of power? Pierre Bourdieu has developed a concept of the 'religious field of practice' (Bourdieu, 1971). A religious field comes into being when a class of religious specialists emerge and try to centralize, systematize, and control a body of knowledge. In doing so, they assert their religious authority and create a field of power in which others are disqualified as laypeople or dismissed as practicing 'superstition,' 'magic,' or some lower form of popular religion. The 'search for meaning' that Geertz describes as motivating the elaboration of symbolic systems is part of this field of practice, but Asad would say that it would be naive to separate this search from the historical dynamics of relations of power.

Asad follows his critique of the definition of religion with another chapter on the concept of ritual. He insists that the idea that ritual is essentially a symbolic, signifying behavior is entirely modern. In his view, anthropologists and historians are performing a theological task when they establish a master meaning of representations by declaring that the (often conflicting) explanations offered by the performers of rituals are inadequate. It is not the role of the scholar to interpret these traditions, he says, but to trace their histories and the ways in which they have been discursively constructed.

Asad's genealogy of ritual concludes with this observation: "Symbols… call for interpretation, and even as interpretative criteria are extended, so interpretations can be multiplied. Disciplinary practices, on the other hand, cannot be varied so easily, because learning to develop moral capabilities is not the same thing as learning to invent representations" (1993: p. 79). His phrasing seems to imply that interpretive studies of symbolism may multiply throughout a hall of mirrors without leading us on to greater knowledge, while the more grounded study of disciplinary practices and the ways in which they come to establish an authoritative discourse rests on somewhat firmer ground.

The question of religion and ritual cannot, in Asad's eyes, be raised without also considering the contrasting process of secularization. Asad's position is that the boundary between secular and religious has been constantly redrawn in Christianity, and anthropologists need to be aware of the fact that their discipline is defined by its conceptual site in the modern world: "It is not just that anthropology is a modern creation born out of Europe's encounter with non-Europeans. It is that the major ideas it used to grasp its subjects (nonmodern, local, traditional) are often dependent on the contrastive sense of the modern" (Asad, 1993: p.19).

This then brings up the question of how one could develop a critical examination of the modernity of symbolic systems within the discipline of anthropology. Asad himself considers David Schneider's 1984, *Critique of the Study of Kinship* an 'outstanding example' of such a study by an anthropologist (Asad, 1993: p.19). Presumably, this is because it calls into question the naturalized concepts of family, sexuality and marriage within American culture, and shows how their symbolic construction has a specific history. The first reader to try to programmatically describe the project of 'symbolic anthropology' was in fact coedited by David Schneider and J.L. Dolgin (Dolgin and Schneider, 1977). Schneider's approach has been very influential in inspiring a new generation of anthropologists who have fused the insights of symbolic anthropology with an awareness of practices and their relation to systems of power, celebrated in Ortner's 1984 review of emerging 'theories of practice.'

See also: Ancestors, Anthropology of; Belief, Anthropology of; Carnival; Cognitive Anthropology; Color Classification and Symbolism; Interpretation in Anthropology; Liminality; Magic, Anthropology of; Play, Anthropology of; Religion, Anthropology of; Ritual; Sacrifice; Spirit Possession, Anthropology of; Taboo; Totemism; Values, Anthropology of; Witchcraft; Worldview.

Bibliography

Asad, Talal, 1993. Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam. Johns Hopkins University Press, Baltimore, MD.

Bourdieu, Pierre, 1971. Genèse et structure du champs religieux. Revue française de sociologie 12, 295–334.

Dolgin, J.L., Schneider, David M. (Eds.), 1977. Symbolic Anthropology: A Reader in the Study of Symbols and Meanings. Columbia University Press, New York.

Douglas, Mary, 1970. Natural Symbols. Cresset, London.

Douglas, Mary, 2000. Deep thoughts on the forbidden: a review of Valerio Valeri's Forest of taboos. Science 289, 2288.

Geertz, 1973. The Interpretation of Cultures. Basic Books, New York.

Geertz, 2000. Available Light: Anthropological Reflection on Philosophical Topics. Princeton University Press, Princeton, NJ.

Kroeber, Albert L., Kluckhohn, Clyde, 1952. Culture: A Critical Review of Concepts and Definitions. The Peabody Museum, Cambridge, MA.

Ortner, Sherry, (Ed.), January 1984. Theory in anthropology since the sixties. Comparative Studies in Society and History 26 (1), 126–166.

Schneider, David M., 1969. Kinship, nationality and religion in American culture. In: Spencer, R.F. (Ed.), Forms of Symbolic Action. University of Washington Press, Seattle, WA, pp. 116–125.

Turner, Victor, 1967. The Forest of Symbols. Cornell University Press, Ithaca.

Turner, Victor, 1974. Dramas, Fields and Metaphors. Cornell University Press, Ithaca.

Turner, Victor, 1975. Symbol studies. Annual Review of Anthropology 4, 145–161.

Turner, Victor, 1986. Dewey, Dilthey and Drama: An essay in the anthropology of experience. In: Victor, Turner, Edward, M. Bruner (Eds.), The Anthropology of Experience. University of Illinois, Urbana, IL.

Turner, Victor, 1987. The Anthropology of Performance. PAJ Publications, New York.

Symptom Perception, Awareness and Interpretation

Omer Van den Bergh, Katleen Bogaerts, and Ilse Van Diest, University of Leuven, Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by B. Knäuper, volume 23, pp. 15357–15362, © 2001, Elsevier Ltd.

Abstract

Self-reported health complaints result from a complex integration of interoceptive (bottom-up) information emerging from peripheral physiological activity with (top-down) perceptual-cognitive and affective processes. Interoceptive signals have to be sensed, perceived, attended to, appreciated and interpreted, put into language and expressed, mostly in a social context. In this article, we will discuss the role of perceptual and attentional processes, anticipation, symptom schemata and illness beliefs, emotion and personality, and memory processes. Their role may importantly influence the relationship between indicators of peripheral physiological activity and self-reported health complaints.

Interoception and Symptom Perception

Although all self-reported health complaints are the result of processes in the brain, a general assumption is that they accurately reflect dysfunctional processes in the peripheral body. The afferent or sensory processes transmitting information from the body to the brain can be called interoception. However, although interoception and symptom perception are related processes, they are not the same.

Interoception

Interoception is poorly defined and several definitions are used. Some equalize it rather narrowly with visceroception (Dworkin, 2000), whereas others (Craig, 2004) also include proprioception and somatoception (the perception of stimuli interacting with the body surface). Still others conceive of interoception more broadly as a meta-representation of the body's state allowing the brain to distinguish the inner from the outer world (Mosely et al., 2012).

Recent neurobiological research has shown that sensations from within the body, such as hunger, thirst, stomach cramps, fullness, rectal urgency, temperature, itch, muscle ache, dyspnea, pain, etc., are processed in a shared evolutionarily ancient neural pathway that serves to maintain the integrity of the body or, in other words, homeostasis. Several comprehensive reviews (Craig, 2002, 2004) suggest that interoceptive pathways include a sensory branch carrying signals to the lamina I and the medullary nucleus of the solitary tract, which are (in humans) integrated in the parabrachial nucleus to travel via parallel ascending pathways to the ventromedial thalamic nuclei further to the mid/posterior insular cortex, in which a modality-selective representation is produced underlying distinctive sensations.

Via ascending afferent pathways providing a direct thalamocortical connection to the anterior cingulate cortex (ACC), interoceptive sensations also become endowed with affective-motivational value (e.g., unpleasantness). The interoceptive image is further rerepresented in the right anterior insula, which is also involved in subjective awareness of feelings and emotions. This process of rerepresenting progresses from the posterior to the anterior parts implies a successively increasing "integration of homeostatic, environmental, hedonic, motivational, social and cognitive activity to produce a 'global emotional moment,' which represents the sentient self" (Craig, 2009: p. 67). This allows for a bidirectional influence of emotion and interoception (Critchley et al., 2001; Zaki et al., 2012). For example, accurate cardiac perception is associated with more intense emotional experience (Herbert et al., 2010), and state negative affect enhances the activation of the right anterior insula and the dorsal ACC (Mayer et al., 2006). The ACC has further strong interconnections with cortical prefrontal (PFC) regions as well as strong descending projections, particularly, to the periaquaductal gray (PAG) and the brain stem, allowing for top-down facilitatory and inhibitory modulation on interoception from higher cortical regions (Mayer et al., 2006). Top-down modulatory processes may go as far as to influence neural activity at the level of the spinal cord (Eippert et al., 2009).

This short neurobiological account shows that there is room for multiple top-down processes to modulate the experience of sensations from the body and different functional aspects of interoception (Garfinkel and Critchley, 2012) have been distinguished. For example, on both neurobiological and functional grounds, a distinction can be made between interoceptive accuracy, which refers to the ability to accurately perceive one's own internal activity, and interoceptive or body awareness, which refers to the state of being cognizant, mindful of one's own internal state (Khalsa et al., 2008, 2009; Ceunen et al., 2013). Interestingly, heightened body awareness can be both adaptive and maladaptive for perceived health (Mehling et al., 2009), whereas evidence shows that elevated symptom reporting may be related to reduced interoceptive accuracy (see further).

Symptom Perception

Symptom perception can be defined as the process of becoming aware of bodily dysfunction. Although symptom perception can theoretically be distinguished from symptom reporting or self-reported health complaints, these concepts are

often used interchangeably. Perceived internal sensations turn into health complaints when they become endowed with negative affective value and become part of a meaning network associated with potential threat to the integrity of the body. In this respect, bodily symptoms can be conceived of as 'homeostatic emotions' (Craig, 2003): they involve a sensory component serving the experience of intensity, location, and other qualities of internal sensations, and an affective-motivational component providing the drive for behavioral action to promote survival. Accurate and aware perception of bodily dysfunction may, therefore, be considered the behavioral layer in a hierarchical defense system to protect the integrity of the body: the behavioral action system is engaged when automatic, more local, automatic and reflexive regulatory systems within the body fail. For example, the experience of breathlessness interrupts ongoing activities, rearranges processing priorities and is associated with a compelling drive to gasp, opening the window, and/or fleeing from closed places to compensate for inaccurate automatic regulation of blood gases.

The process of becoming aware of bodily dysfunction typically includes putting perceived sensations into words, which are part of, or constitute, meaning networks. The semantics of symptom words are particularly relevant for symptom assessments. Most questionnaires use a variety of symptom words, which are clustered according to physiological systems, such as cardiorespiratory and gastrointestinal symptoms. However, other relevant underlying aspects in symptom reporting tend to be overlooked and more sophisticated analyses of dimensions underlying somatic symptom reporting are needed (see Jasper et al., 2012). Indeed, items used in assessment instruments may differ according to other aspects as well. Some inquire about a mere sensory aspect (e.g., fast/deep breathing), while other items also probe an implicit appraisal of its unpleasantness (i.e., 'breathlessness'). Both aspects can reliably be distinguished (von Leupoldt et al., 2006; Wan et al., 2009), but individuals differ in how they intuitively integrate both components (Petersen et al., 2011). Several other relevant dimensions can be distinguished, such as the extent to which a symptom word refers to a specific location in the body or to a systemic state (compare 'tooth ache' with 'tiredness') and the probability of a symptom to be life threatening (compare 'running nose' with 'blood in stools'), etc. Few studies have looked at the role of psychological processes to select or endorse particular wordings to describe one's own internal state (but see Van Diest et al., 2005).

Just like any other type of information entering conscious awareness, information from within the body is sensitive to modulation by functional psychological processes. The seminal work of Pennebaker (1982) on the psychology of physical symptoms introduced most processes elaborated on in later, more detailed models. These processes include attention, expectation, memory, attributions and illness beliefs, emotions and personality (Rief and Broadbent, 2007). In the following, we will discuss a number of relevant psychological processes and how they influence symptom perception. Although several processes are discussed separately, it should be clear that they are mutually interacting to construct a unified somatic experience for the 'sentient self.'

Psychological Modulation of Symptom Perception

Perception and Attention

Critical dimensions for symptom perception are intensity and location of interoceptive stimulation. A general rule seems to be that psychological modulation of bodily symptom perception is less pronounced when the interoceptive stimulation is intense and unambiguous in time (on/off) and location. In these conditions, a rather direct relationship between the interoceptive stimulation and symptom reports exists (Stegen et al., 1998; Put et al., 2004; Martin et al., 2003). Most often, however, interoceptive information is low- to mildly intense and has no clear spatial and/or temporal boundaries, particularly, when the physiological dysfunction is systemic. In such conditions, attention seems to play an important role. Attention, a process involving the allocation of processing resources to stimuli, typically amplifies psychophysiological and behavioral responses to attended stimuli, whereas distraction from these stimuli reduces such responses (Bushnell et al., 1999). Attentional focus can result from local, bottom-up stimulus characteristics (e.g., novelty) or from top-down processes, such as anticipation.

Two important questions emerge as regards attention and symptom perception: is attention 'directed inward' influencing the number and intensity of perceived bodily symptoms, and is it influencing the accuracy of symptom perception (defined as the within-subject correspondence between physiological and self-reported changes)? As to the first question, direct manipulations of attention to the body and studies assessing body awareness generally show that attention focused on the body enhances the sensitivity to interoceptive stimulation and increases symptom reports. This has been documented by studies on tactile stimulation (Mirams et al., 2012), respiration (Stegen et al., 2001), pain (Villemure and Bushnell, 2009), itch (Van Laarhoven et al., 2010), dyspnea (von Leupoldt et al., 2007), and general symptom reporting (Verkuil et al., 2007). Interestingly, not just symptom reports but also objective cough frequency is enhanced by focusing attention on internal sensations (Van den Bergh et al., 2012). Consistent with such findings, it has been shown that performance on concurrent cognitive tasks is disrupted as a result of attentional resources being consumed by noticing interoceptive stimulation (Stegen et al., 2001; Eccleston and Crombez, 1999). Conversely, distraction from bodily sensations generally reduces the perceived intensity of interoceptive stimulation and symptom reports (Pennebaker and Lightner, 1980; Coen et al., 2008; Accarino et al., 1997). The role of attention for symptom reports has been captured in the 'cue-competition hypothesis' (Pennebaker, 1982), assuming that the amount of interoceptive information reaching awareness is a function of the ratio between the amount of interoceptive and exteroceptive information. Findings that persons living in boring environments tend to report more symptoms than persons living in rich and stimulating environments are in line with this hypothesis.

As to the second question, evidence suggests that an interoceptive attentional focus is not necessarily associated with better interoceptive accuracy, but rather with a tendency to overestimate the intensity of somatic sensations. In a study applying low-intensity tactile stimuli, it was shown that a more

liberal response criterion was used when attention was directed inward, leading to more false positives (Mirams et al., 2012). Some groups, such as persons with high-trait negative affectivity (NA) who tend to focus attention inward during a physiological challenge (Stegen et al., 2001), report more symptoms, which are less closely related to the induced physiological changes (Van den Bergh et al., 2004; Bogaerts et al., 2005). Also persons with somatoform disorders, who are typically preoccupied with sensations in the body, reported more symptoms as shown by their interoceptive accuracy to be lower (Schaefer et al., 2012) compared to healthy controls.

Yet an unresolved question is whether the emotional significance of an interoceptive sensation interacts with its sensory processing. Substantial evidence shows that the amygdala serves as a 'neurological hub' swiftly assessing the emotional value of exteroceptive sensory events and subsequently enhancing perceptual processing in early sensory pathways through connections with sensory areas in more distant cortical regions (Pourtois et al., 2013). Claims are made that similar attentional prioritization and subsequent enhancement of processing occurs for potentially noxious stimuli in the 'peripersonal space' (the body and the space surrounding it) (Legrain et al., 2011), but it is not known whether similar mechanisms operate when processing stimuli originating from within the body.

Anticipation, Symptom Schemata and Illness Beliefs

A wealth of evidence on placebo and nocebo documents the dramatic effects of expectations on symptom perception and reporting. Anticipation can be induced either by classical conditioning, verbal instructions, and all kinds of contextual factors. It should be noted that conditioned placebo and nocebo effects do not necessarily depend on conscious expectation (Benedetti, 2013). Effects can impact the intensity, but also determine the sheer presence/absence of symptoms. Particularly, pain analgesia has been thoroughly investigated and a so-called descending pain modulatory network has been described, involving prefrontal (Dorsolateral PFC), limbic (ACC, amygdala), midbrain (PAG), and even spinal cord areas. Knowledge about the mechanisms involved in placebo and nocebo phenomena related to other sensations than pain is more sketchy, but sufficient to document the pervasive effects of anticipation on neurotransmitter function, including the endogenous opioid system, and on autonomic, endocrine, immune, and motor functions (Finniss et al., 2010; Enck et al., 2013). Such findings show that symptom perception involves a bidirectional process without a marked boundary between central and peripheral processes. Bottom-up processes relay information of peripheral physiological activity to the brain, but cognitive processes (critically requiring the involvement of prefrontal areas, such as the dorsolateral PFC) in turn substantially influence activity in other central neurobiological and in peripheral physiological systems. These specific systems may differ according to the type of placebo or nocebo response (Benedetti, 2013).

Symptom schemata typically result from repeated symptom episodes. It is assumed that the person's history with symptoms episodes is recorded in memory and represented in the form of schemata, which capture the commonalities among different symptom episodes. Because some people had more episodes than others, while repeated symptom episodes can result from the same or from different health problems, symptom schemata can differ in complexity and coherence (Petersen et al., 2011). Symptom schemata act like perceptual categories: When activated or primed, a readiness to perceive an interoceptive stimulus configuration in a particular way is facilitated. The benefit is that less information has to be checked and controlled for a symptom experience to emerge, speeding up the perceptual process, but the ease of perception may come at a cost showing up as an elevated probability toward biased perception. For example, the slightest change in respiratory effort may be noticed by an 'experienced' asthma patient, whereas the same change may not reach awareness of the not-yet diagnosed or novice asthma patient. Conversely, an asthma patient who is concerned about potential attacks may easily misperceive respiratory distress caused by stress-induced hyperventilation as signs of an impending attack. Even simple magnitude judgments of respiratory resistance and related affective and behavioral responses are affected by primed perceptual categories of high or low respiratory effort (Petersen et al., 2014).

Just like the perception of a chair is automatic, categorical, meaningful and constructive, so is symptom perception the result of similar basic and automatic perceptual-cognitive processes. These involve a process of mapping sensory evidence on perceptual hypotheses, representing some kind of implicit reasoning. This was captured by Brown's model when postulating a primary attentional system (Brown, 2004). In line with research on perception, it hypothesizes that sensory stimuli are activating perceptual hypotheses in a feed-forward process and that actual information is gradually mapped onto them, giving rise to awareness of only this perceptual hypothesis that obtains the highest level of evidence (e.g., becomes a percept). The higher the activation state of symptom schemata, the less evidence from peripheral stimulation is needed in order to result in percepts. Ultimately, percepts may emerge without peripheral input at all, such as in placebo or nocebo symptoms, which then can be characterized as 'somatovisceral illusions.' For example, when participants were given an air mixture to breathe for a number of times, consisting of a harmless odor and CO_2-enriched air causing bodily symptoms, elevated symptom reports emerged upon subsequently breathing a mixture of the harmless odor only. These nocebo symptoms were similar to the symptoms originally induced by CO_2-inhalation and apparently resulted from automatically activated symptom schemata biasing the subject's perception of his/her somatic state (Van den Bergh et al., 1997, 1998).

However, once symptoms are felt, people may contemplate upon them, link them to an illness label, infer potential causes, anticipate about their consequences, and worry and ruminate about them. In other words, they develop an illness theory, elaborate on it and behave accordingly. The interpretation of palpitations as a symptom of an acute heart condition will prompt completely different behaviors compared to believing that they result from drinking too much coffee. This secondary process may, however, prime perceptual categories and modify the amount of evidence needed for a somatic percept to emerge, with potentially beneficial or disadvantageous

outcomes. In extreme cases, chronic activation of perceptual hypotheses, possibly as a result of catastrophizing, worrying and ruminating, may contribute to somatovisceral illusions, underlying medically unexplained symptoms (MUS) (see further).

Emotion and Personality

A robust association between symptom reporting and trait NA is consistently found ($r = 0.40$ to 0.50). Trait NA is a broad and stable disposition to appraise situations as more threatening and to experience negative mood states and emotions (Watson and Clark, 1984). It can be described as an overreactive evaluative system combined with poor or deficient emotion regulation when processing emotional stimuli (Yiend, 2010). It appears to have a genetic basis, is associated with distinct brain circuit function and neurotransmitter activity (Hariri, 2009) and is considered a vulnerability factor for emotional disorders, such as anxiety and depression (Lonigan and Vasey, 2009).

The association with symptom reporting reflects the effect of NA upon symptom reporting rather than vice versa (Watson and Pennebaker, 1989). The association appears in non-consulting healthy persons (Van Diest et al., 2005), in primary care patients where MUS cooccur with elevated anxiety and depression levels (Kroenke, 2003), and in patients with functional syndromes showing elevated psychiatric comorbidity (Wessely et al., 1999). In addition, also in known diseases a substantial correlation between symptom reports and NA exists (Janssens et al., 2009), reflecting a tendency to overreport symptoms in high NA persons. Closer inspection suggests that the depressive component of NA is mainly associated with 'overreporting' of past symptoms, whereas the anxiety component is more related to excessive reporting of concurrent symptoms (Suls and Howren, 2012). Also, not all symptoms are associated with NA. Van Diest et al. (2005) investigated the relationship between 73 individual symptoms and NA in a large healthy sample and found a wide variability (r's 0–0.45), which could be explained by two factors, namely severity (whether a symptom was potentially life-threatening) and 'somatic versus psychological' (whether a symptom was vague and possibly associated with anxiety; e.g., compare 'stuffed nose' vs 'loosing contact with reality').

Interestingly, also state negative affect impacts symptom perception. When presenting unpleasant emotional cues during experimentally induced bodily sensations, like pain (de Wied and Verbaten, 2001; Meagher et al., 2001), dyspnea (Von Leupoldt et al., 2006), or esophageal stimulation (Philips et al., 2003), symptom reports are more elevated. Even very short presentations of unpleasant pictures can result in increased symptom reports, particularly, in persons with high-habitual symptom levels, and/or high negative affect (Bogaerts et al., 2005, 2008; 2010; Constantinou et al., 2013) and in patients with functional syndromes (Montoya et al., 2005).

Several hypotheses have been advanced to understand this relationship. One idea is that persons with NA and/or during negative affective states have higher levels of sympathetic activity, which would translate into elevated self-reported symptoms. However, several extensive laboratory and ambulatory studies have not been able to document significant differences in a wide variety of peripheral physiological stress or arousal indicators (Houtveen and van Doornen, 2007). Recent studies are inspired by the hypothesis that inflammatory load and immune-related mechanisms are involved and that the relationship between negative emotional states and elevated symptom reports results from stress-related sensitization of the brain-immune communication, but only scanty evidence exists so far in support of this idea (Lacourt et al., 2013, 2012).

Another idea emphasizes attentional mechanisms leading to lower perceptual thresholds for (normal) physiological arousal responses to enter awareness. A set of studies by Gendolla et al. (2005) suggested that self-focused attention in conjunction with negative mood is critical for elevated symptom reports to occur. The interpretation advanced for these findings is that attentional focus on one's own negative mood would activate symptom schemata biasing reports of one's somatic state. However, self-focused attention seems not critical: simply viewing pictures with negative valence also elicits elevated symptom reports in high NA persons (Bogaerts et al., 2010; Constantinou et al., 2013). Apparently, simply processing negative cues, either in the internal or the external world, triggers symptom schemata and makes people prone to biased symptom perception, particularly, in persons with high NA.

Memory

Symptom assessments during diagnosis and treatment evaluations importantly rely on memory information. "How did you feel?" is an often heard question in the doctor's office, but also screening and assessment instruments measuring symptoms in general or related to a particular disease typically use retrospective frequency estimates over a long, often unspecified time window. Questionnaires typically require the respondent to rate a set of symptoms along intuitive standards such as seldom, frequent, almost daily, etc. (Zijlema et al., 2013). In such conditions, symptom reports are mainly based on memory of somatic episodes. The importance of symptom assessment being memory based is often overlooked despite the fact that memory is not just storing an experience, but is actively reconstructing it. Only when actual symptoms are assessed (how do you feel now?), either during doctor visits or with (often disease-specific) questionnaires, momentary information is collected. Evidence suggests that even then, there is no clear one-to-one relationship between physiological responses and self-reported complaints (Walentynowicz et al., 2013). Increasingly, time- or event-related multiple momentary assessment is used to assess symptoms in daily life (Shiffman et al., 2008), allowing for interesting comparisons between concurrent and retrospective ratings spanning the same symptom episodes and assessment period.

The majority of findings suggest a consistent overestimation of past symptoms as measured with global retrospective symptom ratings (Stone et al., 2004; Giske et al., 2010; Linton and Melin, 1982; Broderick et al., 2008). However, several factors relating to characteristics of the symptom episodes, the duration of the recall period, the context during recall and individual differences seem to moderate this overestimation, often resulting in a complex picture. For example, retrospective overestimation of pain was greater with lower initial pain,

while high initial pain was associated with underestimation (Feine et al., 1998). Also the variability of symptoms is an important variable (Stone et al., 2005): greater momentary variability was associated with greater discrepancy between momentary and retrospective ratings and higher retrospective ratings overall. Overestimation typically also increases with longer recall periods (Broderick et al., 2008), although in a study recording physical symptoms in daily life in students, a gradual increase in overestimation of experienced symptoms with longer time frames was observed only among high, but not low habitual symptom reporters (Houtveen and Oei, 2007). Furthermore, actual state during recall plays a role: lower current pain intensity at the moment of recall was associated with underestimation of recalled pain, whereas higher current pain was associated with overestimation (Smith and Safer, 1993; Lefebvre and Keefe, 2002; Gendreau et al., 2003).

A well-investigated memory bias is the so-called peak-end effect in retrospective evaluations. This refers to the observation that not all constituent elements of an experience are equally important when representing it in memory: the intensity of an experience at both the peak and the end receive relatively more weight as these episodes convey the most relevant information. The effect in symptom studies shows up as lower retrospective discomfort ratings when, for example, pain or dyspnea ended gradually at a lower discomfort level rather than abruptly at the peak, despite an equal level of peak discomfort and a longer duration of total discomfort (Kahneman et al., 1993; Bogaerts et al., 2012). Interestingly, patients with MUS do not show this effect, suggesting substantial differences in the way they represent symptom episodes in memory (Bogaerts et al., 2012).

In general, it can be concluded that research on memory for symptoms needs more attention in view of its importance for diagnostic and therapeutic assessment of self-reported health.

Medically Unexplained Symptoms: A Symptom Perception Pathology?

For a substantial part of the patients consulting medical doctors, no physiological dysfunction can be related to their health complaints. Hence, the latter are often called MUS. The share of MUS in primary care consultations is estimated to range from 20 up to 50%, while prevalence rates in secondary care are even higher (Nimnuan et al., 2001). Typical symptoms are fatigue, weakness, headache, muscle aches, nausea and other gastrointestinal complaints, joint pain, palpitations, chest pain, dyspnea, dizziness, etc. (Barsky and Borus, 1999) In a clinical context, the symptoms often appear as functional syndromes, such as chronic fatigue, fibromyalgia, irritable bowel disease, multiple chemical sensitivity, etc. Despite specific diagnostic criteria for such syndromes, overlap and comorbidity between the different categories is large, leading to a debate between 'lumpers' (who consider the different syndromes as basically identical) and 'splitters' (who emphasize the different specificities of the syndromes) (Wessely et al., 1999; Lacourt et al., 2013; White, 2010).

MUS challenge the traditional disease model, which assumes a direct relationship between a physiological dysfunction and self-reported complaints. The absence of such a relationship is the source of both frustrated medical doctors and patients, it feeds somatization processes in the patient and leads to overuse of health care resources and puts a substantial burden on the health care system (Barsky et al., 2005).

Several strategies are followed in search of an explanation. One strategy is motivated by the assumption that as yet unknown (stress-related?) dysfunction in peripheral physiological systems and/or in their interaction with the brain, such as the immune-to-brain communication, is critical and that symptom reports are a true reflection of such dysfunctions. Several reviews revealed evidence in support of this strategy, but the critical involvement of specific physiological systems is overall not considered convincing enough because of the lack of evidence documenting the specificity, consistency, and/or causal direction of the findings (Rief and Barsky, 2005). Other strategies rely on the idea that the critical mechanisms are centrally mediated distortions in the perception of one's bodily state. While substantial evidence is consistently showing distortions in perceptual-cognitive mechanisms of symptom perception, cause–effect relationships remain often unclear, and the critical hypothesized mechanisms remain untested (Rief and Broadbent, 2007).

Both strategies are neither mutually exclusive nor theoretically incompatible. The important advances in placebo research of the recent decade illustrate how deeply psychological variables, such as learning, anticipation, and social context are intertwined with central and peripheral physiological mechanisms (Benedetti, 2013). Such advances are likely the prelude of a paradigm shift in which the opposition between psychological and physiological processes in the study of MUS is considered elusive and will be replaced by research and theorizing on MUS as a nocebo phenomenon, emerging from an intricately intertwined mind-body system.

Conclusion

Several processes modulate the relationship between peripheral physiologic dysfunction and self-reported symptoms. In extreme cases, there is no link at all. This is insufficiently recognized in a biomedical disease model, which assumes a direct relationship between peripheral dysfunction and symptom reports. A more advanced symptom model (Kroenke and Harris, 2001), in which both peripheral and central/psychological processes are included in both a diagnostic and therapeutic step, is needed to further improve health care.

See also: Chronic Pain: Models and Treatment Approaches; Doctor–Patient Interaction in the West: Psychosocial Aspects; Explanatory Style and Health; Health Risk Perception; Illness Behavior and Care-Seeking; Interoception.

Bibliography

Accarino, A.M., Azpiroz, F., Malagelada, J.R., 1997. Attention and distraction: effects on gut perception. Gastroenterology 113, 415–422.

Barsky, A.J., Borus, J.F., 1999. Functional somatic syndromes. Annals of Internal Medicine 130, 910–921.
Barsky, A.J., Orav, E.J., Bates, D.W., 2005. Somatization increases medical utilization and costs independent of psychiatric and medical comorbidity. Archives of General Psychiatry 62, 903.
Benedetti, F., 2013. Placebo and the new physiology of the doctor-patient relationship. Physiological Reviews 93, 1207–1246.
Bogaerts, K., Notebaert, K., Van Diest, I., Devriese, S., De Peuter, S., Van den Bergh, O., 2005. Accuracy of respiratory symptom perception in different affective contexts. Journal of Psychosomatic Research 58, 537–543.
Bogaerts, K., Millen, A., Li, W., De Peuter, S., Van Diest, I., Vlemincx, E., Fannes, S., Van den Bergh, O., 2008. High symptom reporters are less interoceptively accurate in a symptom-related context. Journal of Psychosomatic Research 65, 417–424.
Bogaerts, K., Van Eylen, L., Wan, L., Bresseleers, J., Van Diest, I., Stans, L., Decramer, M., Van den Bergh, O., 2010. Distorted symptom perception in patients with medically unexplained symptoms. Journal of Abnormal Psychology 119, 226–234.
Bogaerts, K., Wan, L., Van Diest, I., Stans, L., Decramer, M., Van den Bergh, O., 2012. Peak-end memory bias in laboratory-induced dyspnea: a comparison of patients with medically unexplained symptoms and healthy controls. Psychosomatic Medicine 74, 974–981.
Broderick, J.E., Schwartz, J.E., Vikingstad, G., Pribbernow, M., Grossman, S., Stone, A.A., 2008. The accuracy of pain and fatigue items across different reporting periods. Pain 139, 146–157.
Brown, R.J., 2004. Psychological mechanisms of medically unexplained symptoms: an integrative conceptual model. Psychological Bulletin 130, 793–812.
Bushnell, M.C., Duncan, G.H., Hofbauer, R.K., Ha, B., Chen, J.-I., Carrier, B., 1999. Pain perception: is there a role for primary somatosensory cortex? Proceedings of the National Academy of Sciences of the USA 96, 7705–7709.
Ceunen, E., Van Diest, I., Vlaeyen, J., 2013. Accuracy and awareness of perception: related, yet distinct (commentary on Herbert et al., 2012). Biological Psychology 92, 426–427.
Coen, S.J., Aziz, Q., Yágüez, L., Brammer, M., Williams, S.C., Gregory, L.J., 2008. Effects of attention on visceral stimulus intensity encoding in the male human brain. Gastroenterology 135 (6), 2065–2074.
Constantinou, E., Bogaerts, K., Van Diest, I., Van den Bergh, O., 2013. Inducing symptoms in high symptom reporters via emotional pictures: the interactive effects of valence and arousal. Journal of Psychosomatic Research 74, 191–196.
Craig, A.D., 2002. How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3, 655–666.
Craig, A.D., 2003. A new view of pain as a homeostatic emotion. Trends in Neuroscience 26, 303–307.
Craig, A.D., 2004. Human feelings: why are some more aware than others? Trends in Cognitive Sciences 8, 239–241.
Craig, A.D.B., 2010. The sentient self. Brain Structure and Function 214, 563–577.
Craig, A.D., 2009. How do you feel now? The anterior insula and human awareness. Nature Review Neuroscience 10, 59–70.
Critchley, H.D., Mathias, C.J., Dolan, R.J., 2001. Neuroanatomical basis for first- and second-order representations of bodily states. Nature Neuroscience 4, 207–212.
Dworkin, B.R., 2000. Interoception. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), Handbook of Psychophysiology, second ed. Cambridge University Press, Cambridge, pp. 482–506.
Eccleston, C., Crombez, G., 1999. Pain demands attention: a cognitive-affective model of the interruptive function of pain. Psychological Bulletin 125, 356.
Eippert, F., Finsterbusch, J., Bingel, U., Büchel, C., 2009. Direct evidence for spinal cord involvement in placebo analgesia. Science 326, 404.
Enck, P., Bingel, U., Schedlowski, M., Rief, W., 2013. The placebo response in medicine: minimize, maximize or personalize? Nature Reviews Drug Discovery 12, 191–204.
Feine, J.S., Lavigne, G.J., Thuan Dao, T.T., Morin, C., Lund, J.P., 1998. Memories of chronic pain and perceptions of relief. Pain 77, 137–141.
Finniss, D.G., Kaptchuk, T.J., Miller, F., Benedetti, F., 2010. Biological, clinical, and ethical advances of placebo effects. The Lancet 375, 686–695.
Garfinkel, S.N., Critchley, H.D., 2012. Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on: "anterior insular cortex mediates bodily sensibility and social anxiety" by Terasawa et al. (2012). Social Cognitive and Affective Neuroscience 8, 231–234.
Gendolla, G.H., Abele, A.E., Andrei, A., Spurk, D., Richter, M., 2005. Negative mood, self-focused attention, and the experience of physical symptoms: the joint impact hypothesis. Emotion 5, 131.
Gendreau, M., Hufford, M.R., Stone, A.A., 2003. Measuring clinical pain in chronic widespread pain: selected methodological issues. Best Practice & Research Clinical Rheumatology 17, 575–592.
Giske, L., Sandvik, L., Røe, C., 2010. Comparison of daily and weekly retrospectively reported pain intensity in patients with localized and generalized musculoskeletal pain. European Journal of Pain 14, 959–965.
Hariri, A.R., 2009. The neurobiology of individual differences in complex behavioral traits. Annual Review of Neuroscience 32, 225–247.
Herbert, B.M., Pollatos, O., Flor, H., Enck, P., Schandry, R., 2010. Cardiac awareness and autonomic cardiac reactivity during emotional picture viewing and mental stress. Psychophysics 47, 342–354.
Houtveen, J.H., Oei, N.Y.L., 2007. Recall bias in reporting medically unexplained symptoms comes from semantic memory. Journal of Psychosomatic Research 62, 277–282.
Houtveen, J.H., van Doornen, L.J., 2007. Medically unexplained symptoms and between-group differences in 24-h ambulatory recording of stress physiology. Biological Psychology 76, 239–249.
Janssens, T., Verleden, G., De Peuter, S., Van Diest, I., Van den Bergh, O., 2009. Inaccurate perception of asthma symptoms: a cognitive-affective framework and implications for asthma treatment. Clinical Psychology Review 29, 317–327.
Jasper, F., Hiller, W., Rist, F., Bailes, J., Witthöft, M., 2012. Somatic symptoms reporting has a dimensional latent structure: results from taxometric analyses. Journal of Abnormal Psychology 121, 725–738.
Kahneman, D., Fredrickson, B.L., Schreiber, C.A., Redelmeier, D.A., 1993. When more, pain is preferred to less: adding a better end. Psychological Science 4, 401–405.
Khalsa, S.S., Rudrauf, D., Damasio, A.R., Davidson, R.J., Lutz, A., Tranel, D., 2008. Interoceptive awareness in experienced meditators. Psychophysiology 45, 671–677.
Khalsa, S.S., Rudrauf, D., Feinstein, J.S., Tranel, D., 2009. The pathways of interoceptive, awareness. Nature Neuroscience 12, 1494–1496.
Kroenke, K., Harris, L., 2001. Symptoms research: a fertile field. Annals of Internal Medicine 134, 801–802.
Kroenke, K., 2003. Patients presenting with somatic complaints: epidemiology, psychiatric comorbidity and management. International Journal of Methods in Psychiatric Research 12, 34–43.
Lacourt, T.E., Houtveen, J.H., Smeets, H.M., Lipovsky, M.M., van Doornen, L.J.P., 2013. Infection load as a predisposing factor for somatoform disorders: evidence from a dutch general practice registry. Psychosomatic Medicine 75, 759–764.
Lacourt, T., Houtveen, J., van Doornen, L., 2012. "Functional somatic syndromes, one or many?": An answer by cluster analysis. Journal of Psychosomatic Research 74 (1), 6–11.
Lefebvre, J.C., Keefe, F.J., 2002. Memory for pain: the relationship of pain catastrophizing to the recall of daily rheumatoid arthritis pain. Clinical Journal of Pain 18, 56–63.
Legrain, V., Iannetti, G.D., Plaghki, L., Mouraux, A., 2011. The pain matrix reloaded: a salience detection system for the body. Progress in Neurobiology 93, 111–124.
Linton, S.J., Melin, L., 1982. The accuracy of remembering chronic pain. Pain 13, 281–285.
Lonigan, C.J., Vasey, M.W., 2009. Negative affectivity, effortful control, and attention to threat-relevant stimuli. Journal of Abnormal Child Psychology 37, 387–399.
von Leupoldt, A., Ambruzsova, R., Nordmeyer, S., Jeske, N., Dahme, B., 2006. Sensory and affective aspects of dyspnea contribute differentially to the Borg scale's measurement of dyspnea. Respiration 73, 762–768.
von Leupoldt, A., Taube, K., Schubert-Heukeshoven, S., Magnussen, H., Dahme, B., 2007. Distractive auditory stimuli reduce the unpleasantness of dyspnea during exercise in patients with COPD. Chest 132, 1506–1512.
Martin, R., Rothrock, N., Leventhal, H., Leventhal, E., 2003. Common sense models of illness: implications for symptom perception and health-related behaviors. In: Suls, J., Wallston, K. (Eds.), Social Psychological Foundations of Health and Illness. Blackwell, Oxford, pp. 199–225.
Mayer, E.A., Naliboff, B.D., Craig, A.D., 2006. Neuroimaging of the brain-gut axis: from basic understanding to treatment of functional GI disorders. Gastroenterology 131, 1925–1942.
Meagher, M.W., Arnau, R.C., Rhudy, J.L., 2001. Pain and emotion: effects of affective picture modulation. Psychosomatic Medicine 63, 79–90.
Mehling, W.E., Gopisetty, V., Daubenmier, J., Price, C.J., Hecht, F.M., Stewart, A., 2009. Body awareness: construct and self-report measures. PLoS One 4, e5614.
Mirams, L., Poliakoff, E., Brown, R.J., Lloyd, D.M., 2012. Interoceptive and exteroceptive attention have opposite effects on subsequent somatosensory perceptual decision making. Quarterly Journal of Experimental Psychology 65, 926–938.
Montoya, P., Sitges, C., García-Herrera, M., Izquierdo, R., Truyols, M., Blay, N., et al., 2005. Abnormal affective modulation of somatosensory brain processing among patients with fibromyalgia. Psychosomatic Medicine 67, 957–963.
Moseley, G.L., Gallace, A., Spence, C., 2012. Bodily illusions in health and disease: physiological and clinical perspectives and the concept of a cortical 'body matrix'. Neuroscience and Biobehavioral Reviews 36, 34–46.

Nimnuan, C., Hotopf, M., Wessely, S., 2001. Medically unexplained symptoms: an epidemiological study in seven specialities. Journal of Psychosomatic Research 51, 361–367.

Petersen, S., van den Berg, R., Janssens, T., Van den Bergh, O., 2011. Illness and symptom perception: a theoretical approach towards an integrative measurement model. Clinical Psychology Review 31, 428–439.

Petersen, S., Schroijen, M., Mölders, C., Zenker, S., Van den Bergh, O., 2014. Categorical interoception: perceptual organization of sensations from inside. Psychological Science 25 (5), 1059–1066.

Pennebaker, J.W., 1982. The Psychology of Physical Symptoms. Springer, New York.

Pennebaker, J.W., Lightner, J.M., 1980. Competition of internal and external information in an exercise setting. Journal of Personality and Social Psychology 39, 165.

Phillips, M.L., Gregory, L.J., Cullen, S., Cohen, S., Ng, V., Andrew, C., Aziz, Q., 2003. The effect of negative emotional context on neural and behavioural responses to oesophageal stimulation. Brain 126, 669–684.

Pourtois, G., Schettino, A., Vuilleumier, P., 2013. Brain mechanisms for emotional influences on perception and attention: what is magic and what is not. Biological psychology 92 (3), 492–512.

Put, C., Van den Bergh, O., Van Ongeval, E., De Peuter, S., Demedts, M., Verleden, G., 2004. Negative affectivity and the influence of suggestion on asthma symptoms. Journal of Psychosomatic Research 57, 249–255.

Rief, W., Barsky, A.J., 2005. Psychobiological perspectives on somatoform disorders. Psychoneuroendocrinology 30, 996–1002.

Rief, W., Broadbent, E., 2007. Explaining medically unexplained symptoms – models and mechanisms. Clinical Psychology Review 27, 821–841.

Schaefer, M., Egloff, B., Witthöft, M., 2012. Is interoceptive awareness really altered in somatoform disorders? Testing competing theories with two paradigms of heartbeat perception. Journal of Abnormal Psychology 121, 719.

Shiffman, S., Stone, A.A., Hufford, M.R., 2008. Ecological momentary assessment. Annual Review of Clinical Psychology 4, 1–32.

Smith, W.B., Safer, M.A., 1993. Effects of present pain level on recall of chronic pain and medication use. Pain 55, 355–361.

Stegen, K., Neujens, A., Crombez, G., Hermans, D., Van de Woestijne, K.P., Van den Bergh, O., 1998. Negative affect, respiratory reactivity and somatic complaints in a CO_2 inhalation paradigm. Biological Psychology 49, 109–122.

Stegen, K., Van Diest, I., Van de Woestijne, K.P., Van den Bergh, O., 2001. Do persons with negative affect have an attentional bias to bodily sensations? Cognition and Emotion 15, 813–829.

Stone, A.A., Broderick, J.E., Shiffman, S.S., Schwartz, J.E., 2004. Understanding recall of weekly pain from a momentary assessment perspective: absolute agreement, between-and within-person consistency, and judged change in weekly pain. Pain 107, 61–69.

Stone, A.A., Schwartz, J.E., Broderick, J.E., Shiffman, S.S., 2005. Variability of momentary pain predicts recall of weekly pain: a consequence of the peak (or salience) memory heuristic. Personality and Social Psychology Bulletin 31, 1340–1346.

Suls, J., Howren, M.B., 2012. Understanding the physical-symptom experience: the distinctive contributions of anxiety and depression. Current Directors in Psychological Science 21, 129–134.

Van den Bergh, O., Stegen, K., Van de Woestijne, K.P., 1997. Learning to have psychosomatic complaints: conditioning of respiratory behavior and complaints in psychosomatic patients. Psychosomatic Medicine 59, 13–23.

Van den Bergh, O., Stegen, K., Van de Woestijne, K.P., 1998. Memory based symptom reporting in a respiratory learning paradigm. Health Psychology 17, 241–248.

Van den Bergh, O., Winters, W., Devriese, S., Van Diest, I., Vos, G., De Peuter, S., 2004. Accuracy of respiratory symptom perception in persons with high and low negative affectivity. Psychology Health 19, 213–222.

Van den Bergh, O., Van Diest, I., Dupont, L., Davenport, P., 2012. On the psychology of cough. Lung 190, 55–61.

Van Diest, I., De Peuter, S., Eertmans, A., Bogaerts, K., Victoir, A., Van den Bergh, O., 2005. Negative affectivity and enhanced symptom reports: differentiating between symptoms in men and women. Social Science & Medicine 61, 1835–1845.

Verkuil, B., Brosschot, J.F., Thayer, J.F., 2007. A sensitive body or a sensitive mind? Associations among somatic sensitization, cognitive sensitization, health worry, and subjective health complaints. Journal of Psychosomatic Research 63, 673–681.

Van Laarhoven, A.I., Kraaimaat, F.W., Wilder-Smith, O.H., Evers, A.W., 2010. Role of attentional focus on bodily sensations in sensitivity to itch and pain. Acta Dermato-Venereologica 90, 46–51.

Von Leupoldt, A., Mertz, C., Kegat, S., Burmester, S., Dahme, B., 2006. The impact of emotions on the sensory and affective dimension of perceived dyspnea. Psychophysiology 43, 382–386.

Villemure, C., Bushnell, M.C., 2009. Mood influences supraspinal pain processing separately from attention. Journal of Neuroscience 29, 705–715.

Walentynowicz, M., Verlinden, A.-M., Van Diest, I., Van den Bergh, O., 2013. How painful was it? retrospective overestimation of pain in high habitual symptom reporters. In: Poster Presented on the 8th Congress of the European Federation of IASP Chapters (EFIC), eighth ed. Florence, Italy, 9–12 October 2013.

Wan, L., Van Diest, I., De Peuter, S., Bogaerts, K., Van den Bergh, O., 2009. Repeated breathlessness experiences induced by hypercapnia: differential effects of intensity and unpleasantness. Chest 135, 455–461.

Watson, D., Clark, L.A., 1984. Negative affectivity: the disposition to experience aversive emotional states. Psychological Bulletin 96, 465–490.

Watson, D., Pennebaker, J.W., 1989. Health complaints, stress, and distress: exploring the central role of negative affectivity. Psychological Review 96, 234–254.

Wessely, S., Nimnuan, C., Sharpe, M., 1999. Functional somatic syndromes: one or many? Lancet 354, 936–939.

White, P.D., 2010. Chronic fatigue syndrome: is it one discrete syndrome or many? Implications for the "one vs. many" functional somatic syndromes debate. Journal of Psychosomatic Research 68, 455–459.

de Wied, M., Verbaten, M.N., 2001. Affective pictures processing, attention, and pain tolerance. Pain 90, 163–172.

Yiend, J., 2010. The effects of emotion on attention: a review of attentional processing of emotional information. Cognition and Emotion 24, 3–47.

Zaki, J., Davis, J.I., Ochsner, K.N., 2012. Overlapping activity in anterior insula during interoception and emotional experience. Neuroimage 62, 493–499.

Zijlema, W.L., Stolk, R.P., Löwe, B., Rief, W., White, P.D., Rosmalen, J.G., 2013. How to assess common somatic symptoms in large-scale studies: a systematic review of questionnaires. Journal of psychosomatic research 74 (6), 459–468.

Synaptic Mechanisms and Protein Synthesis in Memory

Fabio Benfenati, University of Genova Medical School, Genova, Italy; and Fondazione Istituto Italiano di Tecnologia, Genova, Italy
Flavia Valtorta, San Raffaele Scientific Institute, University Vita-Salute S. Raffaele, Milano, Italy; and International School of Psychotherapy with Imaginative Procedures (SISPI), Milano, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Learning and memory are fundamental higher brain functions that allow individuals to adapt to the environment, build up their own history as unique creatures, and widen the personal cultural background, and, ultimately, the population culture. The molecular and cellular mechanisms that contribute to short- and long-term memory are extremely conserved across evolution from mollusks to man and among various forms of memory and consist short- to long-lived rearrangements in synaptic efficiency and in the structure of neuronal networks.

Introduction

Memory is commonly seen as a positive ability of the individual to improve performance, indispensable for survival and social success. Forgetting, on the other hand, generally has a negative connotation, which is often associated with pathological states and/or aging (Della Sala, 2010). This common view is also reflected in our knowledge of the underlying biological processes. Memory and forgetting are daily processes of our life, which allow us to select, from billions of experiences, the most relevant for our personal history and our culture. The French psychobiologist Henry Laborit (1914–95) said that "Without the memories of what was pleasant or unpleasant, it is not possible to be happy, sad, anguished, one cannot get angry, infatuated. One could say that an organism is a memory that acts."

On the other hand, in the absence of forgetting, memory would be completely useless. This concept is very well exemplified by the short fantasy story '*Funes el memorioso*' by Jorge Luis Borges (1942). Describing Funes, Borges says: "On falling from the horse, he lost consciousness; when he recovered it, the present was almost intolerable it was so rich and bright; the same was true of the most ancient and most trivial memories. …And now, his perception and his memory were infallible. …Without effort, he had learned English, French, Portuguese, Latin. I suspect, nevertheless, that he was not very capable of thought. To think is to forget a difference, to generalize, to abstract. In the overly replete world of Funes there were nothing but details, almost contiguous details. …It occurred to me that each one of my words (each one of my gestures) would live on in his implacable memory; I was benumbed by the fear of multiplying superfluous gestures."

The inability to forget details prevents the process of generalization that is necessary for abstract thought and ultimately for making sense of our experiences.

The Synaptic Bases of Learning and Memory

The major and most distinctive feature of the nervous system is its astonishing ability to adapt to the environment and to improve its performance over time and experience. In 1906 this special/unique property, collectively named 'plasticity,' was precisely defined by Santiago Ramon y Cajal as "the property by virtue of which sustained functional changes occur in particular neuronal systems following the administration of appropriate environmental stimuli or the combination of different stimuli." Since the neural changes evoked by the stimuli can persist for a very long time, virtually for the whole life of the individual, neural plasticity could represent an attractive basis for learning and memory. Conversely, the built-in property of neural plasticity might allow experience to functionally and structurally shape the nervous system. The latter aspect, which is also a fundamental feature of neural development and maturation, was envisaged in the third century BC by the Greek philosopher Epicurus who wrote that "it's because something of the external objects penetrates in ourselves that we can identify shapes and think."

The first assembly of neuronal networks is driven by genetic factors, i.e., by the size of the physiological targets and the expression of chemotactic and/or cell adhesion 'recognition' proteins whose genes are specifically transcribed and translated by the various neuronal populations (Kandel et al., 2012). After this first gene-driven developmental period, neuronal circuits are continuously modified and shaped by experience (epigenetic development): synaptic connections that are scarcely used become weaker and eventually disappear, whereas synapses that are heavily used become stronger and eventually increase in number. Synaptic strength can be finely tuned over a short or even a long timescale by a combination of factors including previous activity of the network, generation of second messengers, functional changes in pre- and postsynaptic proteins, as well as regulation of the expression of genes implicated in growth, survival, and synaptic transmission. This results in changes in the efficiency of synaptic transmission that can last from a fraction of a second to minutes in the case of short-term synaptic plasticity (facilitation or depression) to hours, days, and months in the case of long-term synaptic plasticity (long-term potentiation (LTP), long-term depression (LTD)). These changes profoundly affect the processing carried out between input and output information and, ultimately, filter and shape the flow of information within the neural network.

Interestingly, after the cornerstone discoveries of Camillo Golgi and Santiago Ramon y Cajal, but some 50 years before Donald Hebb (1949) formulated the idea of synaptic plasticity as the basis of psychological functions, Sigmund Freud proposed in his *Entwurf einer Psychologie* (1895) that the physical structure of memories consists of a long-lasting activity-dependent modification of information transfer between neurons (Freud, 1953; Sejnowski, 1999; Centonze et al., 2004). Freud drew attention to the synapse, which he called 'contact barrier,' and to the quantity of information '$Q\eta$' that passes through the synapse during the process of neural excitation, i.e., the equivalent of synaptic strength (**Figure 1**). He identified two types of communication, that of "permeable or φ neurons that behave as if they have no contact barriers" (i.e., neurons which transfer information across the synapse without resistance) and that of "impermeabile or ψ neurons which act in such a way as to permit only a difficult or partial passage of $Q\eta$." Thus, the activity of a network depends on the mosaic of facilitated and nonfacilitated barriers since, as Freud says, "$Q\eta$ in an α neuron will be directed toward a more facilitated barrier… and the higher $Q\eta$ during the course of excitation, the greater the facilitation." Thus, memories can be represented as sequences of activity patterns distributed across a population of neurons, which in turn are associated with a different subsequent pattern of encoding (**Figure 1**). During retrieval, a memory cue may cause neural activity to evolve toward one of these activity patterns (attractor state; Hasselmo and McClelland, 1999).

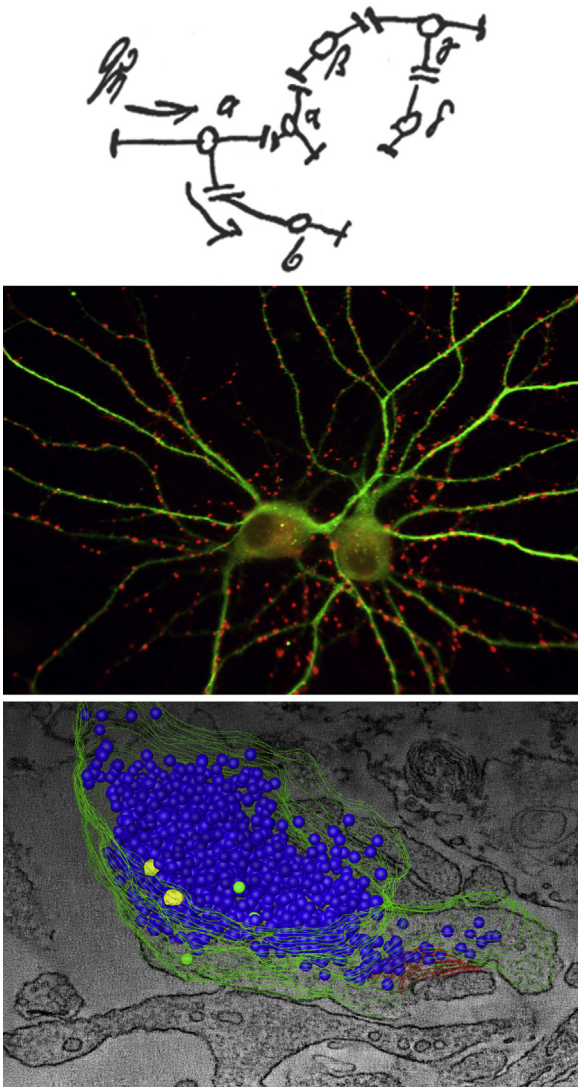

Figure 1 The synapse is at the basis of learning and memory. Upper panel: Sigmund Freud's drawing describing the possibility that changes in the gain of synaptic connections in a neuronal network are at the basis of learning and memory. Middle panel: Two large hippocampal neurons in culture (stained in green) receiving multiple synaptic inputs labeled for the presynaptic protein synaptophysin (red puncta). Lower panel: Three-dimensional reconstruction of an axon terminal from a chemical synapse with its complement of synaptic vesicles storing and releasing neurotransmitter (courtesy of Andrea Raimondi, IIT, Genova).

Memory Consolidation at the Cellular Level

Learning induces cellular and molecular changes that facilitate or impair communication among neurons and are fundamental for memory storage. If learning brings about changes in 'synaptic strength' within neuronal circuits, the persistence of these changes represents the way memories are stored. Short-term memory is believed to involve only functional changes in pre-existing neuronal networks, mediated by a fine-tuning of multiple intracellular signal transductions systems. These short-lived changes can undergo either of two processes: either fade out with time (forgetfulness) or be reinforced and transformed into long-term memory by a process called memory consolidation. Forgetfulness is at least as important as consolidation. Since only a minimal part of what we perceive is useful, the brain needs a mechanism to prevent itself from being burdened by negligible information. To be consolidated, functional changes have to be followed by gene transcription and protein synthesis that produce permanent phenotypic changes in the neuron associated with structural rearrangements in neuronal networks. Thus, memory consolidation is abolished by mRNA and protein synthesis inhibitors. Consolidation is not a high fidelity process: stored memories gradually change and fade with time and only the most relevant and useful aspects are retained over time (Kandel and Pittenger, 1999; Silva and Josselyn, 2002).

Several molecular actors and biochemical processes underlie short-term memory processes. Two processes that appear to be a final common pathway are phosphorylation and dephosphorylation of synaptic proteins (Greengard et al., 1993; Greengard, 2001). Phosphorylation has dramatic effects on protein conformation, interactions, and functions. Thus, the balance between phosphorylation and dephosphorylation is tightly regulated in neurons via the activation of kinase and phosphatase enzymes by specific intracellular signaling molecules called 'second messengers' that include cyclic AMP and calcium. The finely tuned balance between phosphorylation and dephosphorylation has a strong impact on synaptic transmission both at the presynaptic level (neurotransmitter release) and at the postsynaptic level (density and affinity of receptors, activation of signal transduction systems, generation of

retrograde messengers). The activation of the molecules involved in these signaling pathways can last for minutes and thereby represents a sort of short-term 'molecular memory' which can operate as a '*push–pull*' mechanism playing a key role in regulating the delicate balance between storage and disposal of short-term memories (Greengard et al., 1993; Elgersma and Silva, 1999; Abel and Lattal, 2001; Silva and Josselyn, 2002; Genoux et al., 2002).

These purely functional changes cannot survive for long times in the absence of a structural rearrangement of the neurons participating in the modulated synapse. The sustained activation of the same pathways promotes memory consolidation by affecting gene transcription and translation. Sustained stimulation leads to persistent activation of the cyclic AMP-dependent protein kinase and the extracellular signal-regulated kinase pathways. In turn, the cyclic AMP-dependent protein kinase phosphorylates and activates the transcriptional activator CREB1a, whereas the extracellular signal-regulated kinase phosphorylates and inactivates the transcriptional repressor CREB2. The CREB family of transcription regulators is highly conserved across evolution and represents the major switch involved in the transformation of short-term memory into long-term memory. The CREB target genes, whose transcription is regulated during consolidation, include a set of immediate-early genes (such as C/EBP or *zif*268) that affect transcription of downstream genes. This results in changes, either increases or decreases, in the expression of an array of proteins involved in protein synthesis, axon growth, synaptic structure, and function (Kandel and Pittenger, 1999; Alberini, 2005; Bozon et al., 2003). When synaptic strength has to be permanently potentiated (LTP), ribosomal proteins, neurotrophins, Ca^{2+}-binding proteins, and proteins involved in the exo–endocytic cycle of synaptic vesicles and neurotransmitter receptors become upregulated, whereas cell adhesion molecules that usually maintain synaptic stability become downregulated. These specific changes in protein expression favor growth of terminal axon branches and establishments of novel synaptic contacts. Opposite phenomena are believed to occur in the case of LTD of synaptic strength, favoring a decrease in the number of synaptic connections and/or a decreased activity of the existing synapses.

Such properties of synapses were remarkably emphasized by Donald Hebb. He wrote: "when an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949; Sejnowski, 1999). In other words, a '*synaptic learning rule*' exists by which synapses learn from the pattern of afferent stimulation and persistently change synaptic strength accordingly (**Figure 2**). Although LTP and LTD were originally referred to specific synapses (in the hippocampus and the cerebellum, respectively), it turned out that virtually every synapse can finely tune its strength entering a potentiated or depressed state that can last for long

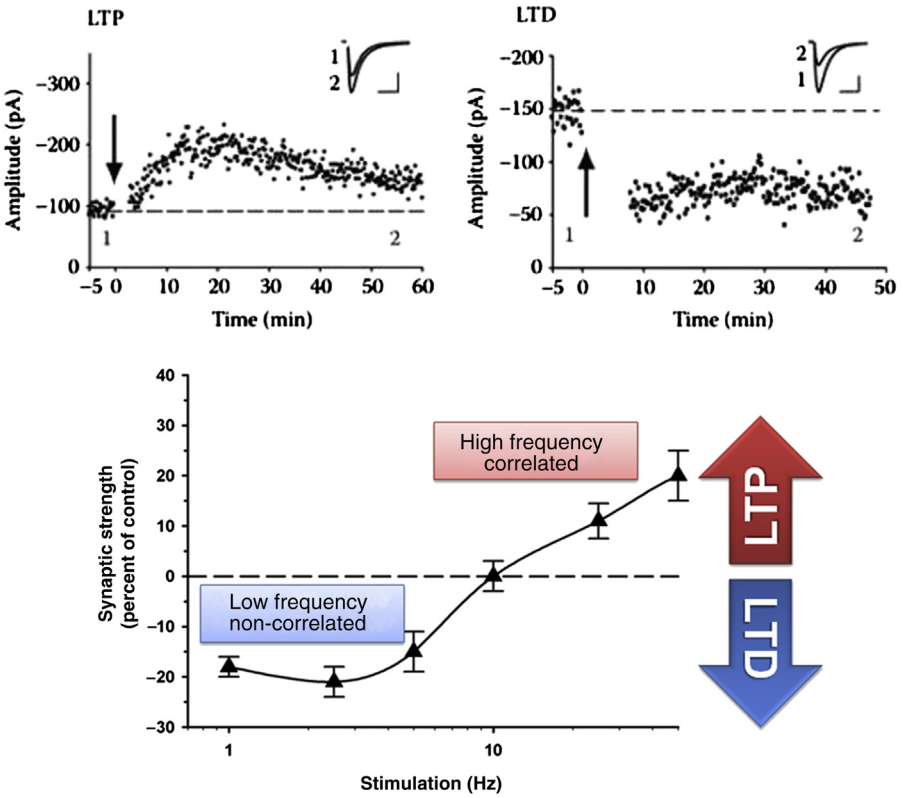

Figure 2 Plasticity paradigms as a '*synaptic learning rule*.' Upper panels: Long-term potentiation (LTP) and long-term depression (LTD) in the dentate gyrus of the hippocampus. Point plot representing the magnitude synaptic in awake rats evoked once every minute. Lower panel: Bidirectional modification of synaptic plasticity as a function of frequency and correlation of the presynaptic activity.

periods of time. As Albert Einstein wrote: "The mind that opens up to a new idea, never comes back to the previous state."

Distinct Forms of Memory Share Similar Cellular Mechanisms

Two major types of memory exist, one for skills and one for knowledge. The first one refers to information storage to perform various reflexive or perceptual tasks and is also referred to as nondeclarative or implicit memory because it is recalled unconsciously. The second form of memory, called declarative or explicit memory because it is recalled by a deliberate and conscious effort, concerns factual knowledge of persons, things, notions, and places. Declarative memory can be further subdivided into episodic or autobiographic memory and semantic memory.

Neuropsychological studies, pioneered by the Canadian neuropsychologist Brenda Milner, have shown that the multiple memory systems involve distinct brain areas and exhibit distinctive features. Thus, explicit memory requires an intact medial temporal lobe (hippocampus), while implicit memory systems are integrated at various levels in the central nervous system including reflex pathways, striatum, cerebellum, amygdala, and neocortex. Moreover, the kinetics of the learning, consolidation, and recall phases of memories are quite different. Implicit memory, e.g., learning to ride a bike, takes time and many attempts to build up, while explicit memory, such as learning a page of history or a telephone number, is more immediate and implies a smaller effort. However, while explicit memory fades relatively rapidly in the absence of recall and refreshing, implicit memory is much more robust and may last for all our life even in the absence of further practice (Blackemore, 1977; Kandel and Pittenger, 1999).

The simplest paradigms of implicit memory are elementary forms of nonassociative and associative behaviors, which are present in primitive animals. These paradigms have been effectively studied in mollusks, particularly the sea snail *Aplysia californica*, which has a very simple central nervous system made up of a few thousand neurons (the human brain in comparison is made up of about 10^{11}–10^{12} neurons). *Aplysia* is able to learn specific behaviors that, upon practice, can be consolidated into long-term memories. The animal progressively learns to respond more weakly to repeated innocuous stimuli (e.g., a light tactile stimulus), a behavior called *habituation*, and to reinforce the response to repeated noxious stimuli (e.g., a painful electrical shock), a behavior known as *sensitization*. In both cases, the synaptic efficiency in the integration center of a sensory-motor reflex is changed by experience, leading to an increased response of the reflex in the case of sensitization or to a reflex inhibition in the case of habituation. Both changes are integrated at the presynaptic level, mediated by changes in calcium influx in response to the action potential. In habituation, calcium influx is decreased into the sensory neuron terminal, resulting in depression of glutamate release. In sensitization, on the other hand, the activity of a facilitating serotonergic interneuron induces phosphorylation of a potassium channel and lengthening of the depolarization, thus increasing the influx of calcium and the release of glutamate (facilitation; Kandel and Pittenger, 1999; Kandel et al., 2012).

Aplysia also exhibits a more complex form of associative learning, typical of higher animals, known as classical conditioning. In this learning paradigm, the animal is given a strong and painful unconditioned stimulus (that if administered alone would produce sensitization) in association with a weak, innocuous, conditioned stimulus (that if administered alone would produce habituation). Following the repeated pairing of these two stimuli over the trials, the animal learns to associate the two stimuli and to react to the isolated conditioned stimulus with an enhanced response (greater than sensitization to the noxious stimulus). Classical conditioning is reflected in the neural circuitry as a greatly enhanced synaptic strength of the input connections between the sensory neuron and the motor neuron. In contrast to nonconditioned learning, this potentiation involves both presynaptic and postsynaptic mechanisms. The coincidence of the two stimuli is revealed by specific coincidence detectors located on both sides of the synapse and including adenylyl cyclase at the presynaptic level, and the ligand- and voltage-operated *N*-methyl-D-aspartate (NMDA)-type glutamate receptor at the postsynaptic level (Kandel and Pittenger, 1999; Kandel et al., 2012). This simple model, highly conserved phylogenetically, tells us that association of events, sensed by synaptic coincidence detectors, generates long-term synaptic changes (such as LTP) that maintain the associative memory with time.

The studies on the mechanisms involved in explicit memory are more complex, as explicit memory involves conscious recall and the integration of multiple sensory inputs. Thus, these studies are not feasible in invertebrates and lower vertebrates but instead require the complexity of the mammalian nervous system. Experimental work in rodents and man has provided strong evidence for the involvement of the hippocampus in many kinds of explicit memory, and particularly in spatial memory. The famous case of Henry Gustav Molaison (the patient H.M.; 1926–2008), who underwent a bilateral hippocampectomy in 1953 to cure an intractable form of epilepsy, first revealed that in the absence of the hippocampus we lose the ability to consolidate new explicit memories. Interestingly, in H.M., short-term memory, old explicit memories, and implicit memory processes were completely normal. The intuition by Brenda Milner on the role of hippocampus in the consolidation of new explicit memories was later confirmed by functional magnetic resonance imaging studies, which have demonstrated an activation of the medial temporal lobe in all tasks in which the subject memorizes a map or mentally rehearses an itinerary. Moreover, studies on the rodent hippocampus have revealed the existence of '*place cells*,' whose firing is primarily controlled by the position of the animal and by distant visual cues that create an internal representation of the animal's location with respect to the surrounding environment (Colgin et al., 2008). In analogy with the rodent 'place cells,' which are very selective and, from a very low baseline, fire strongly to their preferred stimulus, giving an explicit representation of the environment, recording of multiple single neurons in awake and behaving human subjects has shown that sparse neurons in the human medial temporal lobe (and particularly in the hippocampus) respond in a remarkably selective and abstract manner to particular

persons or objects, such as *Jennifer Aniston*, *Luke Skywalker*, or the *Tower of Pisa*. These neurons have been named 'Jennifer Aniston neurons' or, more recently, 'concept cells.' Concept cells can be activated by multiple sensory modalities and encode an abstract representation of the concept triggered by the stimulus. These neurons are probably not required for memory per se, but are crucial for the creation of associations that lead episodic memories and the flow of consciousness (Quiroga, 2012).

The key role of the hippocampus in the formation of new explicit memories and the fact that this brain area exhibits the most known and extensively studied form of synaptic plasticity, namely LTP, has directed research toward the demonstration of a direct link between LTP and memory (Kandel and Pittenger, 1999; Kandel et al., 2012; Moser and Paulsen, 2001; Colgin et al., 2008). A large number of studies have demonstrated that LTP is indeed a valid model of 'memory storage': hippocampal LTP can be induced by animal experience and, conversely, conscious learning is impaired under conditions in which LTP is impaired or abolished. LTP has all the features required to be the cellular mechanism of explicit memory as it is associative in nature, is triggered by the coincidence of events and can be activated by endogenous patterns of electrical activity (e.g., the Θ rhythm). The molecular mechanisms that mediate the generation of hippocampal LTP are surprisingly conserved across evolution and are closely similar to the mechanisms of associative learning identified in invertebrates. Thus, both pre- and postsynaptic mechanisms participate in the early phase of LTP expression, with a coincidence detector represented in most cases by NMDA-type glutamate receptors that trigger activation of multiple kinase pathways, including calcium/calmodulin dependent kinases, and generation of retrograde messengers. Moreover, the late phase of LTP involves activation of transcription factors (*CREB*, *C/EBPβ*, *Arc*, *c-fos*, etc.) and regulation of transcription of the target genes (Miller and Mayford, 1999; Matynia et al., 2002; Alberini, 2005, 2008, 2009; Won and Silva, 2008; Miyashita et al., 2008).

Memory needs time to be stabilized in the hippocampus before the final storage. In fact, LTP induced by an experience is inhibited by a novel experience administered soon after the first one (interference), whereas an LTP established for more than 1 h is immune to this reversal mechanism. These observations suggest that the critical event in determining the retention of information may consist in the stabilization of the potentiated hippocampal synapses in order to resist the LTP reversal upon new information (Miller and Mayford, 1999). Although the hippocampus is fundamental to the acquisition of new memories, it appears to be dispensable after the memory has been fully consolidated. Although patient H.M. was totally unable to lay down new memories, he was still able to remember his past life preceding the bilateral ablation of the hippocampi. This indicates that permanent memories are distributed among different cortical regions according to the various perceptual features ('feature map') and that these various aspects are linked in a 'master map' so that, upon recall, the different components of a memory are bound together to reproduce the memory in its integrity (Seung, 2009). It is currently believed that this memory transfer process occurs largely during sleep, particularly the rapid eye movement phase of sleep (Kandel and Pittenger, 1999; Alberini, 2005). Thus, the hippocampus may represent both the site of the imprinting of the memory and the temporary store for this trace during the progressive formation of neocortical memory representations (Hasselmo and McClelland, 1999).

Memory Recall and Reconsolidation

Upon new learning, a short-lived memory is formed that can be either stabilized over the following several hours or pruned out. If this consolidation process takes place, a long-term memory is formed that is thought to be rather stable over time and stored as permanent modifications in the wiring of the brain in modality-specific areas. Traditionally, consolidation has been considered as an event occurring only once in the biological history of a memory, and recall of a given memory has been often considered a good exercise against forgetting. However, it has only recently become clear that retrieval does not directly reinvigorate memories, rather it makes them return to a labile state susceptible to disruption and interference which needs further consolidation (the so-called reconsolidation process).

This process, originally proposed in the 1960s (Lewis, 1979), has only been recently studied in detail. Reconsolidation appears to be a highly dynamic process that occurs every time memories are reactivated. From a general point of view, consolidation and reconsolidation should be considered as part of the fundamental process of memory stabilization that allows a memory to be preserved, recalled, refreshed over the years. It is commonly found that memories do not remain unchanged over time, but that they undergo transformations in their basic elements and emotional content that have nothing to do with fading. This phenomenon implies that consolidation that is carried out soon after the salient experience cannot be the unique mechanism involved. Rather, reconsolidation provides a dynamic mechanism for updating and modifying memories while they are recalled. What is a bit counterintuitive in this general scheme of memory processes is that a stable memory goes back to a labile state when it is recalled, that is to say that memory recalling is per se an amnesic challenge. However, as memories are not printed as tracks in a compact disk, but are dynamically stored as changes in activity patterns in networks of neurons which in turn depend on modulation of synaptic strength, it is understandable that reactivation of these activity patterns during recall may change the plastic substrate of the memory, so that additional plasticity changes are needed to preserve it. Moreover, the temporal dynamics and the extent by which a memory is deconsolidated upon retrieval strongly depend on the strength of the initially consolidated trace, on the intensity of reactivation, and on the number of reactivation episodes over time. As a general rule, stronger memories are less susceptible to forgetfulness, and the stronger the reactivation, the more labile the memory becomes. It is as if reactivation subtracts part of the memory body and the subtracted part has to be 'rebuilt' by reconsolidation.

Moreover, although every time a memory is reactivated, it regresses to a labile state and needs reconsolidation; each successive reactivation task requires a progressively smaller reconsolidation. The memory therefore becomes rather stable after several cycles, and successive retrieval episodes will not disrupt the trace; the respective reconsolidation episodes will only modify it (Nader, 2003a,b; Dudai and Eisenberg, 2004).

As mentioned earlier, consolidation (i.e., the transformation of short-term memory into long-term memory) requires transcription of specific genes and protein synthesis and involves an array of highly conserved signaling pathways including calcium, cyclic AMP, cyclic AMP-dependent and extracellular signal-regulated kinases, as well as tyrosine kinases that collectively render the memory resistant to cell turnover. Distinct areas are engaged over time in a precise temporal and spatial sequence at both cellular and systems levels. It is well known that hippocampus-related memories are hippocampus-dependent only over a limited period and that at later times they become hippocampus-independent (remote memories), indicating that other brain regions connected to the hippocampus have undergone a sequential memory imprinting (Nader, 2003a; Nader and Hardt, 2009).

When the memory is reactivated, it regresses to a short-term memory labile state (*postreactivation* short-term memory) that is again hippocampus-dependent and sensitive to interference. Thus, the memory needs to be reconsolidated in order to become *postreactivation* long-term memory. However, if the memory is not reactivated, it will remain in a stable state that will slowly fade away over time. Thus, recall appears to disrupt this process of slow decay of long-term memory. While the exact mechanisms of recall are not fully understood, reconsolidation has been thoroughly studied by the use of transcription or translation inhibitors or of genetically altered mice lacking specific proteins involved in synaptic plasticity. The most conservative mechanism that can be envisaged for memory reconsolidation is that this process employs the very same molecular mechanisms used for consolidation of short-term memory and that both cellular and systems reconsolidation processes occur. Under conditions of inhibition of protein synthesis, the functional and structural changes that mediate long-term memory become either dysfunctional or actively removed in 4–24 h after reactivation (Alberini, 2005; Nader and Hardt, 2009).

Reconsolidation has been found to occur in many species from invertebrates to vertebrates (including mammals) and therefore represents a highly conserved fundamental process in memory storage. A typical experiment is the following (**Figure 3**). An animal is subjected to a classical conditioning trial and, after the memory is fully stabilized, it is exposed to the conditioning stimulus to reactivate the memory. If the animal is treated with the protein synthesis inhibitor anisomycin in the reactivation session, it exhibits an intact *postreactivation* short-term memory, but the *postreactivation* long-term memory is markedly impaired. Interestingly, if the animal is not challenged for memory reactivation, protein inhibition is ineffective on the long-term memory acquired in the conditioning session, demonstrating that only the reactivated memory becomes sensitive to disruption, unless synthesis of new proteins is allowed.

The sensitivity to protein inhibition applies only to a narrow time window: if the anisomycin 'amnesic' treatment is

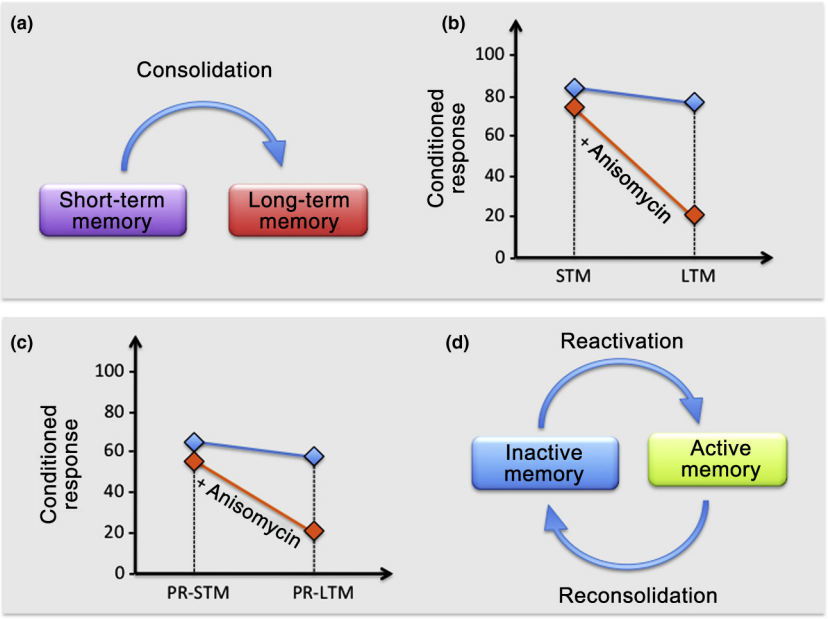

Figure 3 Properties of consolidation and reconsolidation of memories. (a) Model of consolidation of short-term memory (STM) to long-term memory (LTM). (b) Blockade of consolidation of a fear memory by administration of anisomycin after conditioning. The treatment does not affect short-term memory, but strongly impairs long-term memory. (c, d) Model of long-term memory reconsolidation. Long-term memory and post-reactivation long-term memory (PR-LTM) are stable traces in an inactive state. Reactivated memories return in an active, labile state that needs reconsolidation for long-term storage.

administered several hours after reactivation, it is ineffective, indicating that reconsolidation, like consolidation, is a time-dependent mechanism and that the time needed for reconsolidation is generally shorter than that needed for consolidation. This picture was observed in the case of diverse memory paradigms, including contextual or fear conditioning, passive avoidance, object recognition, taste aversion, motor sequence learning, etc. As for consolidation, reconsolidation is not demonstrated only in behavioral tasks, but it also has neurophysiological correlates. It has been demonstrated that if anisomycin is given 2 h after LTP induction, it does not affect LTP maintenance. The most conservative explanation is that both consolidation and reconsolidation share the same mechanisms of synaptic rearrangement and permanent tuning of the strength of synaptic connections. Indeed, this seems to be often the case, although similarities between the two processes are not complete and important differences exist in either the molecular actors involved or in the target brain regions or both (Nader, 2003a; Alberini, 2005; Nader and Hardt, 2009).

Recent data have shown that in humans existing declarative memories can be selectively rewritten by disrupting reconsolidation using a noninvasive retrieval–relearning technique. This reconsolidation-associated amnesia can be achieved 48 h after formation of the original memory, but only if relearning occurs soon after retrieval (Chan and LaPaglia, 2013). These results demonstrate that also declarative memories, whose encoding and retrieval rely on a distributed network, are susceptible to reactivation-induced lability, and that memory impairment can be obtained in a behavioral paradigm, without the use of pharmacological agents.

In conclusion, memory recall destabilizes long-term memory and poses the need for memory reconsolidation. Successive cycles of reconsolidation make the memory more stable, even in the presence of successive retrievals and, at the same time, rearrange and slowly modify the memory trace, so that the most salient and emotionally significant features are preserved or even enhanced (Figure 4). Consolidation and reconsolidation often occur in distinct brain regions or subregions, consistent with the idea that consolidated memories are sorted to diverse brain areas, but the molecular mechanisms involved largely overlap.

Conclusions

Memory is not just an ability to store information, but the essence of our beings, the basis of our individuality and our consciousness. Each individual knows that he/she is unique, not merely because of his/her external appearance, but because of his/her personal history, behavior, and ability to face daily life. Learning and memory are achieved by permanently shaping neuronal circuits and interneuronal connections. These modifications are initially labile, but if they are perceived as useful and salient, then they are consolidated and eventually reconsolidated to become a stable memory. However, most of the percept is quickly discarded, in a process of selection: what we think is relevant is remembered, although it tends to fade and change with time: "Life is not what one lived, but what one remembers and how one remembers it in order to recount it. The heart's memory eliminates the bad and magnifies the good; and thanks to this artifice we manage to endure the burdens of the past" Gabriel Garcia Marquez said (Marquez, 2002). Although many questions remain open, the astonishing progress in the field of molecular and cellular neuroscience is greatly contributing to the understanding of the exact role of

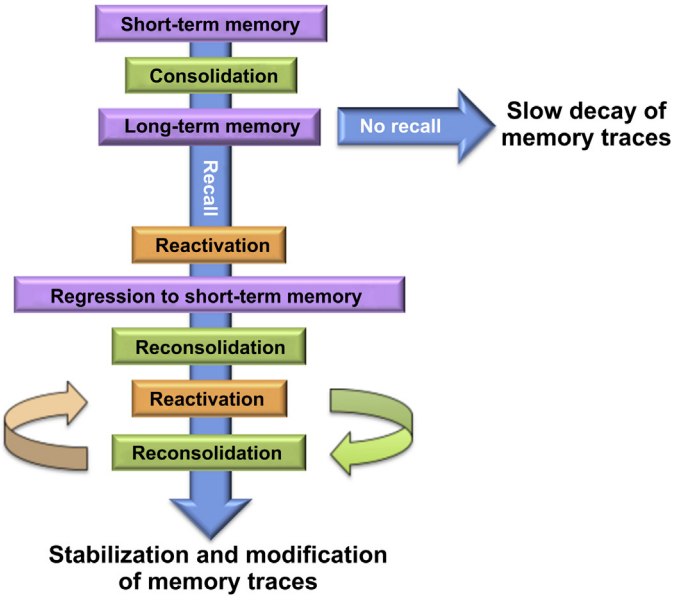

Figure 4 Lifetime and transformations of memory traces. Once short-term memories have been consolidated into long-term memories, they will slowly fade away if not recalled. On the contrary, if recalled, long-term memories are destabilized and need a process of reconsolidation to be preserved. Successive cycles of recall and reconsolidation make the memory more stable, but, at the same time, rearrange and progressively modify the memory trace.

gene products and signaling pathways in distinct processes of memory and forgetting, and it is likely that within a few years this knowledge will be translated into the development of novel therapeutic approaches to memory disorders.

See also: Amnesia: General; Autobiographical Memory; Declarative Memory, Neural Basis of; Episodic Memory; Genetic Approaches to Memory; Hippocampal Amnesia; Hippocampus and Related Structures; Implicit Memory; Learning and Memory, Neural Basis of; Memory in the Bee; Non Declarative (Procedural) Memory; Retrieval; Retroactive Interference; Semantic Memory; Short Term Memories, Theories of; Short-Term Memory: Psychological and Neural Aspects; Sleep and Memory.

Bibliography

Abel, T., Lattal, K.M., 2001. Molecular mechanisms of memory acquisition, consolidation and retrieval. Current Opinion in Neurobiology 11, 180–187.

Alberini, C.M., 2005. Mechanisms of memory stabilization: are consolidation and reconsolidation similar or distinct processes? Trends in Neurosciences 28, 51–56.

Alberini, C.M., 2008. The role of protein synthesis during the labile phases of memory: revisiting the skepticism. Neurobiology of Learning and Memory 89, 234–246.

Alberini, C.M., 2009. Transcription factors in long-term memory and synaptic plasticity. Physiological Reviews 89, 121–145.

Blackemore, C., 1977. Mechanics of the Mind. Cambridge University Press, Cambridge.

Borges, J.L., 1942. Funes el memorioso. In: Ficciones (part 2), Editorial Sur.

Bozon, B., Kelly, A., Josselyn, S.A., Silva, A.J., Davis, S., Laroche, S., 2003. MAPK, CREB and zif268 are all required for the consolidation of recognition memory. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 358, 805–814.

Centonze, D., Siracusano, A., Calabresi, P., Bernardi, G., 2004. The project for a scientific psychology (1895): a Freudian anticipation of LTP-memory connection theory. Brain Research. Brain Research Reviews 46, 310–314.

Chan, J.C.K., LaPaglia, J.A., 2013. Impairing existing declarative memory in humans by disrupting reconsolidation. Proceedings of the National Academy of Sciences of the United States of America 110, 9309–9313.

Colgin, L.L., Moser, E.I., Moser, M.B., 2008. Understanding memory through hippocampal remapping. Trends in Neurosciences 31, 469–477.

Della Sala, S. (Ed.), 2010. Forgetting. Psychology Press, East Sussex.

Dudai, Y., Eisenberg, M., 2004. Rites of passage of the engram: reconsolidation and the lingering consolidation hypothesis. Neuron 44, 93–100.

Elgersma, Y., Silva, A.J., 1999. Molecular mechanisms of synaptic plasticity and memory. Current Opinion in Neurobiology 9, 209–213.

Freud, S., 1953. The Standard Edition of the Complete Psychological Works of Sigmund Freud. Hogarth Press, London.

Genoux, D., Haditsch, U., Knobloch, M., Michalon, A., Storm, D., Mansuy, I.M., 2002. Protein phosphatase 1 is a molecular constraint on learning and memory. Nature 418, 970–975.

Greengard, P., Valtorta, F., Czernik, A.J., Benfenati, F., 1993. Synaptic vesicle phosphoproteins and regulation of synaptic function. Science 259, 780–785.

Greengard, P., 2001. The neurobiology of slow synaptic transmission. Science 294, 1024–1030.

Hasselmo, M.E., McClelland, J.L., 1999. Neural models of memory. Current Opinion in Neurobiology 9, 184–188.

Hebb, D.O., 1949. Organization of Behavior: A Neuropsychological Theory. John Wiley & Sons, New York.

Kandel, E.R., Pittenger, C., 1999. The past, the future and the biology of memory storage. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 354, 2027–2052.

Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J., 2012. Principles of Neural Sciences. McGraw Hill Professional, New York.

Lewis, D.J., 1979. Psychobiology of active and inactive memory. Psychological Bulletin 86, 1054–1083.

Marquez, G.G., 2002. Vivir para contarla. Grupo Editoriale, Bogota.

Matynia, A., Kushner, S.A., Silva, A.J., 2002. Genetic approaches to molecular and cellular cognition: a focus on LTP and learning and memory. Annual Review of Genetics 36, 687–720.

Miller, S., Mayford, M., 1999. Cellular and molecular mechanisms of memory: the LTP connection. Current Opinion in Genetics and Development 9, 333–337.

Miyashita, T., Kubik, S., Lewandowski, G., Guzowski, J.F., 2008. Networks of neurons, networks of genes: an integrated view of memory consolidation. Neurobiology of Learning and Memory 89, 269–284.

Moser, E.I., Paulsen, O., 2001. New excitement in cognitive space: between place cells and spatial memory. Current Opinion in Neurobiology 11, 745–751.

Nader, K., 2003a. Memory traces unbound. Trends in Neurosciences 26, 65–72.

Nader, K., 2003b. Re-recording human memories. Nature 425, 571–572.

Nader, K., Hardt, O., 2009. A single standard for memory: the case for reconsolidation. Nature Reviews. Neuroscience 10, 224–234.

Quiroga, R., 2012. Nature Reviews. Neuroscience 13, 587–597.

Sejnowski, T.J., 1999. The book of Hebb. Neuron 24, 773–776.

Seung, H.S., 2009. Reading the book of memory: sparse sampling versus dense mapping of connectomes. Neuron 62, 17–29.

Silva, A.J., Josselyn, S.A., 2002. The molecules of forgetfulness. Nature 418, 929–930.

Won, J., Silva, A.J., 2008. Molecular and cellular mechanisms of memory allocation in neural networks. Neurobiology of Learning and Memory 89, 285–292.

Syncretism

André Droogers, Vrije Universiteit, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The meaning of the term 'syncretism' has changed drastically over the course of time. Currently, the term refers to the mixing together of elements from different religions. This mixing is sometimes viewed with contempt, especially by the clergy of the religions concerned. Scholarly use of the term is more objective and does not entail this kind of value judgment. In this article we will describe the history of the term. Special attention will be given to current debates on power mechanisms in syncretism. Syncretism will be compared with fundamentalism. Finally, the tendency to move on from defining syncretism to studying the process of syncretization will be discussed.

History of the Term 'Syncretism'

Earliest Uses

The first known use of the term 'syncretism' is found in Plutarch's (46?–125) *Moralia* (Morals), a treatise on moral principles. When suggesting that one should consider one's brother's enemies and friends as one's own enemies and friends, he uses the example of the inhabitants of Crete. Plutarch uses the word syncretism to refer to their habit of forgetting internal conflicts and differences when confronting a common enemy. It may be that the use of the word was inspired not only by the Cretans' example but also by a word play, since the Greek *syngkrasis* means a 'mixing together' (Stewart and Shaw, 1994: p.3).

The second author known to have used the term was the humanist Erasmus of Rotterdam (1496–1536). He quoted Plutarch and, like him, recommended syncretism, although here as a way of seeking a synthesis of seemingly disparate philosophical and theological opinions and as an attempt to understand the other completely, despite differences of view.

After the Reformation, with its increase in divergent theological views, the term acquired a negative connotation for the first time. Opponents of the Danish–German Lutheran theologian George Calixtus (1586–1656), an early defendant of ecumenism and theological reconciliation, used the term pejoratively as a mixing of points of view that, in their opinion, was to be avoided. The debate became known as the 'syncretistic controversy.'

Views from Religious Studies and Missionary Practice

In the nineteenth century, when research on world religions began to flourish and colonialism brought cultures and religions into contact, syncretism was understood more objectively as the mixing of elements from different religions, although purity continued to be viewed implicitly as the norm (Rudolph, 1979: pp.196–197).

Christianity also reached the height of its expansion in the nineteenth century. Missionaries objected to the tendency of new converts to continue practices from their prior religion. Syncretism then became the concept used for this – condemned and combated – tendency. This meant that Christian theological use of the term was no longer restricted to internal differences between theologians but referred to the mixing of religious elements, especially Christian and non-Christian. Unlike its usage in religious studies, syncretism was considered impure, contradictory, and inconsistent, and therefore censurable and illegitimate. The term still retains this meaning, even though ecumenical missionary circles have become more understanding of the need to translate the Christian message into local cultural terms (so-called inculturation; see Schreiter, 1993). It has also been suggested that Christianity itself is the result of a syncretistic process (Pannenberg, 1970: pp.85–88 and note 37). Yet, despite this development, the theological criterion continues to be how intact the core message of Christianity remains. An exception is the Korean theologian Chung Hyun-Kyung, known for her address to the Assembly of the World Council of Churches in Canberra in 1991.

For a long time, scholars in religious studies have worked with the term and defined its meanings (Rudolph, 1979 offers a useful summary), whereas, as will be shown in the next section, the term has a much shorter history in anthropology. Scholars in the field of religious studies, such as historians and phenomenologists of religion, view syncretism objectively and point to its general presence at the founding of religions (even in religions that now oppose syncretism; cf. Maroney, 2006) and in their mutual contact. Scholars distinguish between symmetrical and equal relationships between religions on the one hand, with mutual exchange, and asymmetrical and unequal relationships on the other hand, with unilateral influencing. The basic ambiguity of meanings in syncretistic contexts has been emphasized by Pye (1971: p.93). Typologies of the degree to which elements can be mixed have been proposed, using terms such as 'identification,' 'amalgamation,' and 'symbiosis.' It has been acknowledged that syncretism may involve elements other than religious ones, because religions are linked to cultural contexts. Syncretism then appears as a religious form of a general cultural process of meaning-making. Sometimes, syncretism is presented as the result but more often as the process itself.

Whereas syncretistic clergy is more the exception than the rule, syncretism can be found as a normal part of popular religion, generally condemned by the representatives of official religion. Correspondingly, syncretism has often been depicted as an unconscious, implicit, and spontaneous routine, as typical of what Redfield called the little tradition of the

unreflective many (Rudolph, 1979: p.208). In terms of Saussure's linguistic distinction between *langue* (the system) and *parole* (speech), institutional religion resembles the first, whereas syncretism can be compared to the latter (Leopold and Jensen, 2004: pp.381–382).

There have also been proposals to excise the term from scholarly vocabulary (Baird, 1971; Pakkanen, 1996; Schineller, 1992; Van der Veer, 1994), for being too vague, subjective, and too broad. If syncretism occurs in all religions and if borrowing and blending are universals in any event, the term seems to be insufficiently specific and therefore superfluous. But the term 'syncretism' is useful for showing the relativity of established and rather essentialist categories, such as 'world religion,' 'tribal religion,' or 'big and little tradition.' The focus could move from bounded institutionalized systems of beliefs and practices to the individual believer who is finding his or her way through these systems, not only between religions but also within one religion, and including the wider cultural context. Whereas institutionalized forms of religion can be easily located and studied, the individual syncretistic believer, including his or her idiosyncratic practice of plural meaning-making, is not as easily researched.

Precise definition or even redefinition may further the debate on the use of the concept (see Leopold and Jensen, 2004 for a fundamental discussion of the issues and options at stake). In view of the increase in contact between believers of different religions, some term is needed to refer to the inevitable mixing of religious elements. In addition, the long history of the term has shown that its meaning is open to change.

Views from Cultural Anthropology

Although students of religion from outside cultural anthropology focused on world religions, the first anthropological texts on syncretism referred to Latin American and Caribbean examples, especially the Afro-Christian syncretism that followed from the presence of slaves and their descendants in the region. Of course, syncretism was present in other areas, but where the negative missionary use of the term was the norm, as in Africa, anthropologists avoided using the term, even with respect to obvious cases of syncretism, as occurred in the so-called African independent churches founded by Africans and not by missionaries (Stewart and Shaw, 1994: pp.14–16).

When the American anthropologist Melville Herskovits began using the term in anthropology, following his Brazilian student Arthur Ramos, he took his examples from Afro-American religions (Greenfield and Droogers, 2001: pp.26–28; Stewart and Shaw, 1994: pp.5–6). His interest in the topic was partly ideological, because he understood syncretism as a tool for promoting a unified culture in the United States, to be formed from the many cultures that immigrants, including those who had involuntarily come from Africa, had brought with them. This melting-pot idea was, of course, a way of rehabilitating the Afro-American strand in the US culture over against white racism. Moreover, Herskovits saw syncretism as a halfway stage in the process of acculturation, the intermingling of cultures in contact. He was very much interested in the way people gave meanings to selected old and new elements, mixing cultural standards, and thus initiating change toward a new synthesis. The alternative course is an emphasis on authenticity and a return to tradition and to the past, as in the case of the black consciousness movement. Nourished by the political ideal of integration the concept of acculturation may have suggested a one-sided linear evolution.

Another scholar who used the term 'syncretism' in the Afro-American context was the French ethnologist Roger Bastide (1978). On the basis of fieldwork in Afro-Brazilian religions, he emphasized the systematic way in which analogous elements from different religious sources are mixed. He also pointed to the role of power mechanisms, especially in the contact between slaves and their masters. The structural similarity between African, Catholic, and Amerindian worldviews facilitated syncretism. Thus, African gods could be identified with Catholic saints and Amerindian spirits. In the oppression that the slaves suffered, supernatural help was most welcome. This redressed the balance of power somewhat, such as when masters were the object of slave rituals intended to harm them. The redress of this balance also became clear when Catholic elements were selectively adopted through the application of African criteria, and not vice versa, as the masters wished when they baptized their slaves. In practice, as a strategic device, identification with and differentiation from the masters' religion were both used in a double game. Thus, the apparent adoption of a Catholic ritual attitude in Mass could serve as an alibi for the continuation of African ritual practices.

One might ask whether syncretism is primarily a matter of convictions and ideas, or of practices and rituals. The fact that syncretists are usually not aware of their mixing attitude points to a trial-and-error integration of new religious ideas and practices in their routine, more in an unconscious, natural way than an emphatic, intercultural way. Because rituals often serve a practical purpose, any new element that promises to be effective will be welcomed, thereby making religious practice the preferred scene for syncretism, as was the case in Afro-American syncretism. Interreligious pilgrimage is another case in point. Those religious leaders who manifest themselves against syncretism may combat either ideas or practices, or both, though the practices are by nature more visible than the ideas. On the other hand, the clergy's theological interest may be more focused on beliefs, especially those that obviously were in contrast with their own doctrine – which served as a sieve. In any case, much syncretism goes in fact unobserved, partly for lack of information on the part of the religious leadership, and partly for not drawing attention, as is especially the case with ideas.

Current Issues

Power

The political use of the concept of syncretism has received more attention in the 1990s (Droogers, 1989; Greenfield and Droogers, 2001; Stewart and Shaw, 1994). The negative sense of the term is an indication of a conflict between clergy and laity on the right to produce legitimate religion. Leaders of a religion whose adepts are involved in a syncretistic process will, from a purist point of view, use their power to condemn any effort to mix elements from different religions as heresy. Stewart and Shaw (1994) accordingly proposed the term 'antisyncretism.'

As suggested earlier in the article, the subjective meaning of the term could be a reason for abandoning the term in scholarly use. Yet it has the advantage of revealing the power dimension of the syncretism process. In that sense, syncretism is one of a number of terms used in anthropology despite their pejorative meanings, such as magic or popular religion, all having the connotation of deviation, resistance, and rebellion. Moreover, religious innovators, including founders of world religions, operated in a manner that can be labeled syncretistic, often being critical of the established religion of their days.

An attitude of antisyncretism is not only held by leaders of affected religions but can also be found in representatives of religions that are the result of a process of syncretism, once these religions have transformed themselves from movement into institution. Thus, some leaders of the very syncretistic Afro-Brazilian religions adopt an antisyncretistic attitude, favoring a return to the African roots and the removal of all Catholic influence.

Syncretism and Fundamentalism

In various senses, syncretism and fundamentalism are opposites (Droogers, 2005, 2008; Maroney, 2006). Whereas syncretism uses more than one religious source, fundamentalism defends the uniqueness of the one source, often codified in a sacred text. Syncretism implies that more versions of the experience with the sacred are acceptable and can be combined, whereas fundamentalists defend one exclusive version. Syncretists experiment with religious symbols, as many as seem to be applicable to their situation, whereas fundamentalists jealously guard the one central symbol of their religion. Syncretism is usually criticized by religious specialists who feel that they are losing control and who may react by defending a fundamentalist view. Syncretists interpret the opportunity offered by the growing contact between cultures and religions as a chance to combine religious repertoires for thought and action, whereas fundamentalists react to the moral side effects of this contact by retreating to the traditional repertoire, even though they may make use of modern means of communication to spread their message. Both syncretism and fundamentalism seem to flourish as a consequence of modernization and globalization but for different reasons: one embracing the new situation, the other being critical of it because it is felt to threaten cherished traditions and morals. Syncretists behave as global citizens, whereas fundamentalists divide the world into 'us' and 'them.' Syncretists are not organized and rarely present themselves as such, whereas fundamentalists tend to organize themselves and to occupy power positions. Syncretists embrace the stream of information on religious ideas and practices produced by modern mass media, whereas fundamentalists tend to view this type of information as inappropriate and therefore to be barred. Syncretists easily admit inconsistencies or are not even aware that they do so, while fundamentalists tend to build a system with its own logic and principles of causality.

From Definition to Process

Whereas for decades the main issue in the debate on syncretism was its definition, more recently the central question has become what actually happens in the syncretistic process. Now that the field of study has been outlined, more attention can be given to the way syncretism works. To reach this point, it was necessary to take leave from an essentialist way of thinking that takes institutional and official religion as the normal and main manifestation of religion, considering deviations from that model as exceptional and even dysfunctional. In fact, though indirectly, scholars for decades copied the religious leadership's perspective and framework. This was reflected in the way syncretism was defined, that is, in contrast to bounded institutional religion. When more attention was given to the way believers deal with the meanings of their beliefs and practices, as happened in the study of popular religion and also by introducing power as a crucial framework in meaning-making, syncretism could become an accepted topic on the research agenda, as normal as institutionalized religion. This also led research from a focus on coherent systems of thought to the process of often inconsistent and complex meaning-making.

This approach was inspired by the increasing tendency to look at culture not as a set of customs connected with a particular society, in the same way as world religions were categorized, but as the universal human capacity for meaning-making. Where formerly culture was the repertoire of meanings that a person should know in order to be able to participate in his or her society, now culture became the tool needed to get along in a globalized world, full of perforated cultural boundaries. Religion could accordingly be understood as a particular form of universal meaning-making, addressing transcendental experiences. The concept of syncretism expressed the same trend in the study of religion. Therewith syncretism became a domain of religious globalization and part of a broader cultural transformation. Similar to the condemnation of religious syncretism by representatives of the clergy in a particular religion, cultural globalization comes with a nationalist or ethnic countermovement, leading to a kind of cultural orthodoxy. The methods used for the study of globalized meaning-making are applicable to the study of syncretism as well. This is also reflected in the more recent interdisciplinary approach in syncretism research, using notions from linguistics and cognitive studies (Leopold and Jensen, 2004). The emphasis on syncretism as deviating from institutional religion becomes less important, whereas its role in the globalization process receives more attention. This also stimulates a view on syncretism as being a normal phenomenon, instead of being a deviation. Its role in the rise and sustenance of institutional religion is acknowledged.

Symptomatic for this development is also the extension of the concept of syncretism to culture in general, as was proposed recently (Adogame et al., 2008). The attention given to the power aspect points in the same direction, showing the political dimension of religion. Parallel to the emphasis on the process, the term then becomes 'syncretization' (Adogame et al., 2008: p.4). This expansion of the study field of syncretism would solve the problem of including or excluding the gray zone of the connections between religious syncretism and the wider social and cultural context. Such an elaboration would put syncretism in one category of concepts with creolization, hybridity, and *bricolage* (on the religious aspects of the latter term, see McGuire and Maduro, 2005),

with a focus on the hermeneutics of blending, negotiating, borrowing, and appropriating. In a way, as a concept, syncretism refers primarily to a form of meaning-making, pointing to religious and, by extension, cultural mixing.

At the same time, the history of the concept is itself an example of ongoing scholarly meaning-making, including the mixing of terms, disciplines, methods, and meanings. Both parties, researchers and researched, appear to be engaged in the same cultural game. Consequently, academia can be said to have its syncretists, just as it has its orthodox and fundamentalist scholars (Droogers, 2008).

See also: Creolization: Sociocultural Aspects; Globalization, Anthropology of; Hybridity; Mission: Proselytization and Conversion.

Bibliography

Adogame, A., Echtler, M., Vierke, U. (Eds.), 2008. Unpacking the New: Critical Perspectives on Cultural Syncretization in Africa and Beyond. LIT, Zürich and Berlin.

Baird, R.D., 1971. Category Formation and the History of Religion. Mouton, The Hague.

Bastide, R., 1978. The African Religions of Brazil, Toward a Sociology of the Interpenetration of Civilizations. Johns Hopkins University Press, Baltimore, MD.

Droogers, A., 1989. Syncretism: the problem of definition, the definition of the problem. In: Gort, J.D., Vroom, H.M., Fernhout, R., Wessels, A. (Eds.), Dialogue and Syncretism: An Interdisciplinary Approach. Eerdmans and Rodopi, Grand Rapids, MI and Amsterdam, pp. 7–25.

Droogers, A., 2005. Syncretism and fundamentalism: a comparison. Social Compass 52, 463–471.

Droogers, A., 2008. Syncretists, fundamentalists and scholars compared. In: Adogame, A., Echtler, M., Vierke, U. (Eds.), Unpacking the New: Critical Perspectives on Cultural Syncretization in Africa and Beyond. LIT, Zürich and Berlin, pp. 47–69.

Greenfield, S., Droogers, A. (Eds.), 2001. Reinventing Religions: Syncretism and Transformation in Africa and the Americas. Rowman & Littlefield, Lanham, MD.

Leopold, A.M., Jensen, S.P. (Eds.), 2004. Syncretism in Religion: A Reader. Equinox, London.

Maroney, E., 2006. The SCM Core Text: Religious Syncretism. SCM, London.

McGuire, M.B., Maduro, O. (Eds.), 2005. Beyond Syncretism: The Bricolage Debate. Social Compass, 52, pp. 411–471 (special issue).

Pakkanen, P., 1996. Interpreting Early Hellenistic Religion: A Study Based on the Mystery Cult of Demeter and the Cult of Isis. Layias and Souvatzidakis, Athens.

Pannenberg, W., 1970. Basic Questions in Theology. SCM, London.

Pye, M., 1971. Syncretism and ambiguity. Numen 18, 83–93.

Rudolph, K., 1979. Synkretismus vom theologische Scheltwort zum religionswissenschaftlichen Begriff. In: Humanitas religiosa: Festschrift für Haralds Biezais zu seinem 70. Geburtstag. Almqvist & Wiksell, Stockholm, pp. 193–212.

Schineller, P., 1992. Inculturation and syncretism: what is the real issue? International Bulletin of Missionary Research 16, 50–53.

Schreiter, R.J., 1993. Defining syncretism: an interim report. International Bulletin of Missionary Research 17, 50–53.

Stewart, S., Shaw, R. (Eds.), 1994. Syncretism/Anti-syncretism: The Politics of Religious Synthesis. Routledge, London.

Van der Veer, P., 1994. Syncretism, multiculturalism and the discourse of tolerance. In: Stewart, S., Shaw, R. (Eds.), Syncretism/Anti-syncretism: The Politics of Religious Synthesis. Routledge, London, pp. 196–211.

Syntax (General)

Masayoshi Shibatani, Rice University, Houston, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Fundamental issues in syntax are reviewed including part of speech systems, phrase structures, constituency, grammatical relations, and several syntactic constructions such as possessives, relative clauses, and nominalizations. Chomskyan generative grammar and functional-typological approaches are illustrated and contrasted with reference to several phenomena, including a discussion on a new approach to embedding structures and relative clauses, in particular.

Until the middle of the last century, syntax, along with semantics, was one of the least-developed fields in linguistics. The field, however, witnessed the advent of a new age with the publication of Noam Chomsky's groundbreaking monograph *Syntactic Structures* (1957), which radically altered the outlook on grammar, from individual languages to native speakers' mental knowledge of their languages, and the analytical methods, from the discovery procedures to the positing of deep structures and generative transformational rules, as well as the focus of investigation, from historical comparative linguistics to the theoretical study of the universal properties of human language. Besides these innovations in theoretical linguistics, which have dominated the field over the past 50 years, the 'Chomskyan revolution' aroused interest in syntax-focused descriptions of the world's languages, with the expectations that they would provide empirical foundations for the theoretical inquiries into the nature of human language. In more recent years, however, this latter effort of dovetailing the descriptive-grammar writing and the search for language universals has come to be strongly aligned with the Greenbergian functional-typological, as opposed to the Chomskyan formal, approach to language universals, whose methods are laid out in the influential 1963 paper "Some universals of grammar with particular reference to the order of meaningful elements" by the late Joseph Greenberg. The discussion below scrutinizes some fundamental issues in both theoretical and descriptive studies in syntax in the light of the recent developments in formal and functional-typological approaches to grammar.

Parts of Speech

A grammatical classification of words into lexical categories is prompted by the simple fact that words, the minimal syntactic units, do not behave uniformly. English adjectives, for example, modify a noun, as in *a large dog*, whereas English verbs cannot do so (**a run dog*; cf *a running dog*). Do all languages divide words into similar lexical classes or parts of speech, and if so what are universal parts of speech? These have been perennial questions that have been asked by ancient and modern grammarians alike including Chomskyan generative syntacticians.

Chomsky (1970) proposed feature-based representations of four lexical categories of N, V, A, and P in terms of the specifications of [±N] and [±V] as in (1). The use of features of N and V as basic classificatory features is reminiscent of the ancient division of sentence components into *onoma* (a nominal part) and *rhema* (a verbal part) by Plato (*Cratylus*):

1. N=[+N, –V]
   V=[−N, +V]
   A=[+N, +V]
   P=[−N, –V]

This neat formal representation of parts of speech and its advantages of allowing cross-categorical classifications of words have quickly found their way to syntax textbooks such as Radford (1997: 76), where N, V, A, and P are introduced as 'the four primary lexical categories.' Whether or not this kind of feature notation is adopted, Noun, Verb, Adjective, Preposition, and a few other categories are routinely introduced in textbooks (e.g., Sportiche et al., 2013) that are meant to inform the recent efforts and progress in search of the universal properties of human language.

Similarly, many, if not most, descriptive grammars take it for granted that these lexical categories, or at least so-called open-class categories of noun, verb, adjective, and possibly adverb, exist in all languages and include them without questioning the universal validity of these lexical categories or their realizations in individual languages. While it is unclear what is meant by the characterizations of the four lexical categories given in (1) as the 'major' or 'primary' lexical categories in the works cited above, it is increasingly becoming clear that adjective is not a major or primary lexical category in many languages, and some languages simply do not have prepositions or their positional variants, postpositions (e.g., Dyirbal and other Pama-Nyungan languages in Australian – see Dixon, 2002).

Of A and P (covering both prepositions and postpositions), the problem is more acute for the former since many descriptive grammars assume the existence of an adjective class since languages in general have property-concept words (PCWs) for such semantic categories as size, color, and age, without realizing that parts of speech must be established on morpho-syntactic grounds for each language. While most textbooks offer morphological criteria

such as a tense inflection for verbs, a number inflection for nouns, and comparative and superlative inflections for adjectives, these morphological properties are, for one thing, language specific and are not applicable universally. For another, these morphological features are semantically based, and as such they do not represent the properties of the relevant lexical categories. For example, whether nouns inflect for number or not is determined by the countability of nominal referents, and accordingly there are numerous nouns that do not show the plural inflection in English. Similarly, it is only adjectives denoting gradable property concepts that inflect for the comparative and superlative degrees.

A more reliable method in identifying parts of speech of individual languages is in terms of the morpho-syntactic patterning of lexical items in the three major functional domains of reference, predication, and modification (Croft, 1991). The idea is that those items that display similar morpho-syntactic patterns in the three functional domains form a unified part of speech, while those that pattern differently warrant different lexical classes. In English, nouns and adjectives are similar to a large extent in the ways they behave in both predication and modification functions, and they distinguish themselves from verbs in these functions (see (2a) and (2b)). English nouns differ from adjectives in that while the former can perform a referential function, the latter cannot (2c):

2. a. Predication *This is cotton/expensive/*complain.*
   *They complain/*cotton/*expensive.*
   b. Modification *cotton shirt/expensive shirt/*complain customers.*
   c. Reference *Cotton is a useful commodity.*
   **Expensive is annoying.*

This method of identifying parts of speech also allows cross-linguistic labeling of lexical categories without resorting to the semantic properties of the relevant lexical items. Forms that perform a referential function without undergoing a category-changing derivation are labeled as noun, those that perform a modification function, but not a referential function, without a category-changing derivation form an adjective class, and those that perform a predication function by themselves constitute a verb class. Languages in general have category-changing derivations that permit derived forms to perform functions different from those associated with the relevant lexical categories. For example, English has nominalization processes that convert verbs and adjectives into a nominal category, and these derived nominals behave like basic nouns performing those functions associated with nouns (cf *Smoking* is a bad habit/*Sincerity* is an admirable quality).

One of the limitations of the feature-based analysis of parts of speech like the one shown in (1) is that it fails to capture cross-categorial similarities in a more fine-grained manner. Possible similarities between adjectives and nouns are captured by the shared [+N] specification, and the [+V] specification singles out adjectives and verbs for any cross-categorial similarities they may display. But this system fails to capture the indication seen in (2) above that English adjectives are more similar to nouns than to verbs, a trait that English has inherited from classical Indo-European languages such as Ancient Greek and Sanskrit, and which is more pronounced in Romance and other European languages than in English (see below).

The above point is indicative of the fact that the adjective class is not really a 'major,' 'primary,' or a universal lexical category on a par with the noun and the verb class. While there is a language like Japanese, where PCWs form two adjective classes of verbal adjective (mainly consisting of native PCWs behaving similarly to verbs) and nominal adjective (mainly Chinese loans behaving similarly to nouns), many languages of the world treat PCWs either like nouns, or as a subclass of noun (Quechua, Tarascan (Mexico), Yokuts, Sothern Sierra Miwok, Shaptin, most languages of Europe, North Africa, North Asia, and Australia), or like verbs or as a subclass of verb (many East and Southeast Asian and Pacific languages (e.g., Chinese, Korean, Lahu, and Fijian)) and most of North American languages (e.g., Mohawk) (Dixon 1993; Dixon and Aikhenvald, 2006).

Particularly interesting are languages that divide PCWs into two subclasses of adjectival noun and adjectival verb. The Austronesian language Lamaholot spoken in the eastern tip of Flores Island in eastern Indonesia treats PCWs like *okĩ* 'old,' *wuʔũ* 'new,' *ləmeʔ* 'deep,' *briŋĩ* 'sick,' *blara* 'injured,' *gələ* 'tired,' and so on like nouns in their role in a referential and a modification function, where they can head an NP and can modify another noun without undergoing nominalization. On the other hand, PCWs like *beləʔ* 'big,' *kreʔ* 'small,' *blahaʔ* 'long,' *blolo* 'tall,' *nureʔ* 'young,' *blega* 'wide,' *knipu* 'narrow,' *gə̃* 'sharp,' and so on behave like verbs in that they must be nominalized in order to perform both referential and modification functions (Nagaya, 2011). In Malagasy, another Austronesian language, PCWs like *avaesatora* 'heavy,' *alalaka* 'wide' inflect for present, past, and future tenses just like verbs, while other PCWs like *mandu* 'wet' and *fuhi* 'short' are noun-like in inflecting only for future tense.

Phrase-Structure Rules, Phrasal Categories, and Constituency

In an earlier generative syntax framework, possible syntactic structures of a given language and the constituency (chunking) of the internal phrasal units were defined by phrase-structure (PS) rules of the following form:

3. a. S → NP Aux VP
   b. NP → (D)N′
   c. N′ → N (S′)
   d. VP → V (NP)
   e. S′ → (C) S

These PS rules 'generate' or recognize the following phrase structure in (4) to be well-formed, while disallowing those in (5) from the possible phrase structures of English:

4.

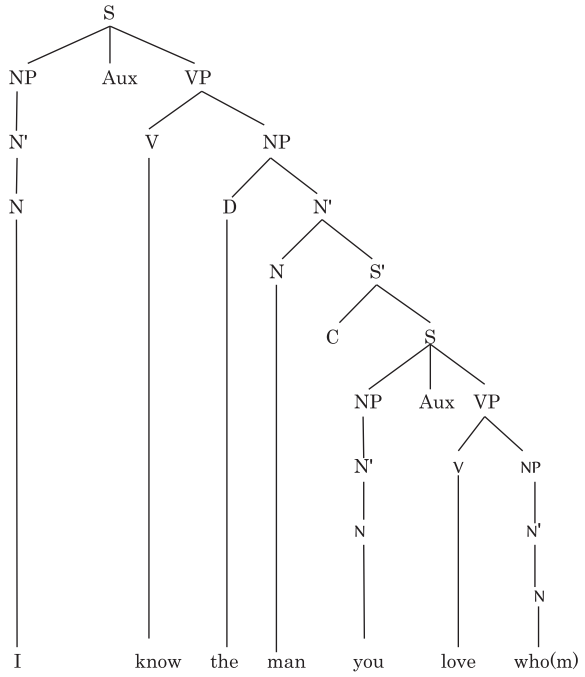

I know the man who(m) you love.

5. a.

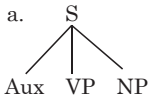

b.

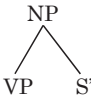

c.

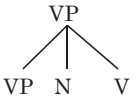

The PS rules in (3) also capture the recursive property of syntactic structures that is instantiated by the relative clause construction in (4), where the recursion of S (or a sentence) together with its internal phrasal categories is seen.

As it turns out, a theory incorporating PS rules of the type shown in (3) is too strong since it allows the positing of any kind of PS rules sanctioning phrase structures that are unlikely in human language. For example, PS rules (3b) and (3c) define an NP structure as one containing an N as its obligatory member. Similarly, PS rule (3d) stipulates V to be an obligatory member of a VP structure. These NP and VP structures are widely seen across languages. On the other hand, nothing prevents us from positing PS rules of the following type, which would sanction unlikely structures like those in (7):

6. a. NP → V S
   b. VP → D NP

7. a.

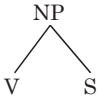

b.

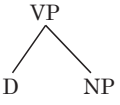

Realization that NPs, VPs, and other phrasal categories of APs (adjectival phrases) and PPs (prepositional/postpositional phrases) all contain a core or 'head' constituent of N, V, A, and P, respectively, prompted the positing of X-bar schemata, as in (8), which would replace the individual PS rules generating these phrasal categories (Chomsky, 1970; Jackendoff, 1977). (This framework is known as 'X-bar' syntax because the original proposal was to use bars above the lexical categories to indicate the phrasal status of the relevant categories. The practice in the field has been to write $\overline{X}$ ($\overline{N}$ for example) as X′ (N′) for a typographic reason.)

8. a. XP = … X′
   b. X′ = … X

   where … is optional phrasal elements; X varies over N, V, A, and P; word order unspecified.

Unlike the traditional PS rules, X-bar schemata are meant to be universal constraints on possible phrase structures of human language and would predict that no language permits phrase structures of the type given in (7), where a phrasal category is not headed by a constituent of a similar category.

The X-bar schemata of the type in (8) handle endocentric phrasal structures, where a larger phrasal category is headed by a subcategory of the same kind; for example, NPs headed by an N′ and the latter by an N. However, they would not define the structures of clauses and sentences as these have been traditionally considered exocentric, where a higher structure is of different constituent type (e.g., S) from any of the immediately lower constituents (e.g., NP, Aux, and VP). Positing PS rules like (3a) and (3e) for clauses and sentences undermines the effort to constrain the theory sufficiently so as to bar arbitrary PS rules that make wrong predictions about possible phrase structures. Besides, permitting both endocentric and exocentric structures goes against Occam's razor. What has been proposed is a radical departure from the traditional conception about the structures of clauses and sentences, whereby a functional category representing inflectional features, I, has been recognized as the head of a clause, and the complementizer category C has been introduced as the head of a sentence. In other words, IP, the maximal projection of I, and CP, the maximal projection of C, have been proposed as the clausal and sentential structures, as in the following representations:

9.

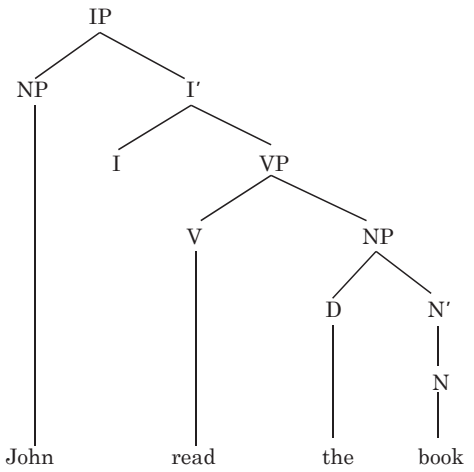

10.

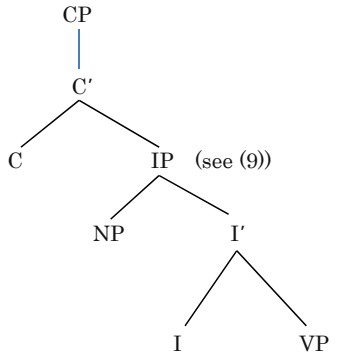

Besides the two functional (as opposed to lexical categories such as N and V) categories of I and C, D, for determiners, have been proposed as another functional category that heads a noun phrase. Now the traditional NPs have been reanalyzed as DPs of the following structures:

11.

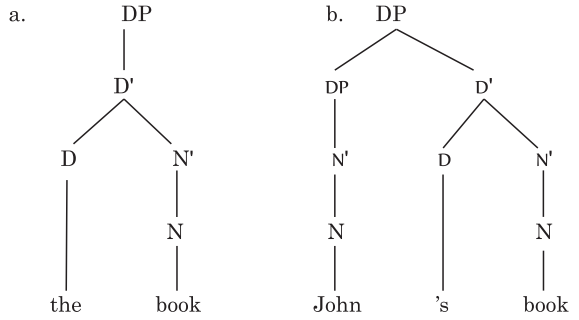

Notice that IP, CP, and DP are all endocentric structures, unifying them with other endocentric structures of VP, AP, and PP. This generalization over various types of phrase structures, however, has been achieved at the great cost of losing the fundamental understanding of the notion of endocentricity. Endocentric structures are defined as those that function similarly to one of their constituents, namely to their heads. Endocentric structures result from structural elaborations of head constituents such that functioning endocentric structures are replaceable by head constituents without affecting the relevant functions. For example, the VP *slept three hours in the office* in *John slept three hours in the office* can be replaced by the head *slept* without affecting the grammaticality of the sentence as *John slept* indicates. Similarly the NP constituent *these very expensive books* can be replaced by the head *books* without affecting the well-formed NP structure. In the case of the proposed IP, CP, and DP, however, they are certainly not elaborations of I, C, and D in a similar way VPs and APs are of their head V and A. Indeed IPs, CPs, and DPs do not function like their heads, and certainly cannot be, respectively, replaced by I, C, and D without affecting the grammaticality of the resulting structures. In other words, IP, CP, and DP are not functionally endocentric at all. This is easy to see if we consider the function of NPs, now reanalyzed as DPs. The central function of NPs is reference, that is, a speech act of pointing out and distinguishing an entity in the universe of discourse from other similar entities. This function of NPs is based on the entity-denotation property of an N head, not that of a determiner modifying the N head.

In functional terms exocentric structures for clauses and sentences make good sense. While VPs and NPs, respectively, have a predication and referential function whether they simply consist of a head V or N, or they have elaborate endocentric structures with modifying elements, clauses, and sentences perform different functions from either VPs or NPs. Clauses have the function of materializing the predication function of a VP by ascribing the relational property denoted by a VP to the referent of a subject NP. Sentences, on the other hand, perform speech acts of various sorts. A declarative sentence, for example, asserts the truth of the predication made by a clause. These properties of clauses and sentences distinguish them sharply from nominalizations, whose function is to denote entity-concepts just like nouns (see Section Structure, Use, and Form).

The purging of exocentric structures from linguistic theory seems premature. For one thing, exocentric structures are seen elsewhere in grammar. Compound words show both hierarchical internal structures and the headedness, as well as endocentric categorical patterns similar to phrasal structures. There are, however, many compounds in various languages that are exocentric in both semantic and structural terms. Japanese deverbal noun *nobori* 'climbing,' for example, is a noun distinguishing itself from the members of the verbal noun class. Verbal nouns such as *benkyoo* 'studying' and *ryokoo* 'travel' allow light-verb constructions supported by the verb *suru* 'do,' as *benkyoo*(=*o*) *suru* (studying (=ACC) do) 'do studying,' but nouns, including deverbals like *nobori* 'climbing,' do not permit light-verb constructions. When *nobori* is compounded with the noun *yama* 'mountain' as in *yama-nobori* 'mountain climbing,' we obtain an exocentric structure as in (12), where the entire word is a verbal noun, which allows the light-verb construction *yama-nobori* (=*o*) *suru* 'do mountain-climbing.'

12.

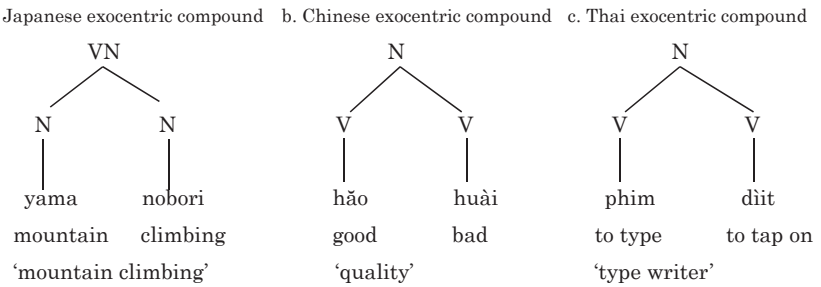

Chinese and Thai have both right-headed and left-headed N–V compounds. In addition they both have exocentric V–V compounds whose overall structures are N. For example, Chinese *hǎo-huài* (good–bad) 'quality,' *dà-xiǎo* (big–small) 'size,' *cháng-duǎn* (long–short) 'length,' *lěng-rè* (cold–hot) 'temperature'; Thai *phim-đìit* (to type-to tap on) 'typewriter,' *kan-chon* (to prevent-to collide) 'buffer,' *kam-phân* (to grasp-to mold) 'fist.'

Observations that words forming a sentence organize themselves into hierarchically structured constituents underlie the notion of phrasal categories such as NPs, VPs, and PPs that PS rules or X-bar schema are meant to define. However, whether or not these phrasal categories are universal has been a contentious issue. Informed of free word-order languages like Japanese and many Australian aboriginal languages, Hale (1983) proposed what is known as the configurational parameter dividing the languages of the world into configurational (e.g., English) and non-configurational languages (e.g., Japanese) with the assumption that while the former type of languages have phrasal categories such as NP and VP, the latter may not. The extremely 'flat' Pamangyngian language of Australia Kalkatungu is reported to allow reordering of sentence elements in any manner even dispersing semantically related words at a great distance as in the following sentence, where *japacara-tu* 'clever-ERG' is placed far away from what it semantically modifies, namely *kuḷa-ji* 'father-ERG' (Blake, 1979):

13. Ṇa-ci kuḷa-ji ḷaji ṯuar-Ø malṯa j apacara-ṯu.
    1SG-DAT father-ERG kill snake-ABS mob-ABS clever-ERG
    'My clever father killed the snakes.'

The pairing of the semantic head and its modifier is made possible thanks to the case markers (glossed as ABS and ERG in the example above) that identify semantically related words. What is of interest here is that these discontinuous 'constituents' call into question some of the syntactic tests identifying constituents such as the ones below that are standardly introduced in introductory syntax textbooks:

14. Some proposed constituency tests
    a. Only a constituent can answer a question, while retaining the meaning of the original sentence.
    b. Only constituents can be substituted with pro-forms like *it*, *that*, *them*, *then*, *there*, *do that*, *that way*, ….
    c. Only constituents (of the same type) can be coordinated using conjunction words like *and*, *or*, and *but*.

Test (14b) tells us that while the underlined part of (15a) is a constituent, the underlined portion of (15b) is not. Similarly (14c) predicts that while the underlined parts of (16a) are constituents, the underlined portions of (16b) are not:

15. a. Jane married *the man whom she loved*, although **he** was poor.
    b. *Jane married *the man* whom she loved, although **he** whom Mary also loved was poor.
16. a. *The trip to London* and *the party in East London* were both enjoyable.
    b. John gave *a book to Mary* and Bill gave *a CD to Janet*. **The book to Mary* and *the CD to Janet* were both expensive.

If these tests were valid, we would expect the discontinuous elements *kuḷa-ji* 'father-ERG' and *japacara-tu* 'clever-ERG' in the Kalkatungu example in (13) to fail them. However, this does not appear to be the case; (13) may be following by the following sentence, where the pronominal form *li-ji* 'he-ERG' substitutes not just *kuḷa-ji* 'father-ERG' but the entire discontinuous elements *kuḷa-ji* 'father-ERG' … *japacara-tu* 'clever-ERG,' since (17) is likely to mean "The clever father cooked the meat":

17. Ḷi-ji ṯuji-ṇa ati.
    he-ERG cook-PAST meat
    'He cooked the meat.'

Moreover, a nominal phrase and a noncontiguous modifier together may be conjoined with another nominal phrase, as in the following example, where the modifier *japacara-ṯu* 'clever-ERG' can be associated with either *na-ci kuḷa-ji* 'my father-ERG' or *marapai-ṯu* 'woman-ERG' – notice that (18) does not mean "my father and the woman, who are clever …" (Van Valin, 2001):

18. Ṇa-ci kuḷa-ji marapai-ṯu-jama ḷaji ṯuar-Ø malṯa
    1SG-DAT father-ERG woman-ERG-and kill snake-ABS mob-ABS
    japacara-ṯu.
    clever-ERG
    'My clever father and the woman killed the snakes.'
    'My father and the clever woman killed the snakes.'

These possibilities in Kalkatungu, where discontinuous 'constitutes' of the above type abound, suggest that some of the constituency tests are really for semantically connected units rather than syntactic or phonological groupings. It so happens that in those languages used to demonstrate the notion of constituency semantically related elements tend to be grouped together phonologically, as in English.

Grammatical Relations

The theoretical status of grammatical relations or functions such as subject, object, and indirect object, and the question of whether or not they – the subject in particular – are universal relational categories have also been contentious perennial issues in syntax. While descriptive grammars in general tend to continue the traditional definition of subject as the NP representing an agent, Chomskyan generative grammar made a radical departure from this tradition in defining grammatical relations subject and object in structural terms, in part due to the fact that the traditional equation of agent = subject does not work in many cases including the subject of a passive sentence. The earlier version of the structural definitions holds subject to be an NP immediately dominated by S and object to be an NP immediately dominated by VP (Chomsky, 1965). Structurally-based definitions like these and those based on the more recent X-bar version of PS rules, where subject is defined as a Specifier of XP, however, assume that all languages having these grammatical relations have configurational structures similar to English (cf the earlier discussion on non-configurational languages).

While the strict Chomskyan generative tradition holds grammatical relations to be derived notions definable in structural terms, other approaches such as Relational Grammar (Perlmutter, 1983) and Lexical Functional Grammar (Bresnan, 2001) regard them to be universal theoretical primitives necessary in capturing universal properties of syntactic rules (e.g., passivization is defined as a relation-changing rule of turning object into subject) and constraints on the distribution of grammatical relations (e.g., a single clause may have at most a single subject, this constraint being responsible for disallowing a passive sentence from having two subjects).

The position that grammatical relations are language universals has posed a serious challenge in identifying subject and other grammatical relations in two groups of languages, namely so-called ergative languages, where what, on morphological grounds, appear to be the subject of a transitive sentence corresponds to a direct object of languages like English, and Western Malayo-Polynesian languages, where what appears to be a subject is not always aligned with an agentive NP even in what seem to be basic structures. The problem of identifying grammatical relations in different types of language even prompted a checklist approach, such as the one proposed by Keenan (1976), which lays out a set of subject properties that would help us identify a subject in any given language.

Much more empirically-minded typologists adopt a phenomenon-driven approach to grammatical relations, which recognizes grammatical relations only insofar as a phenomenon demands them (Dryer, 1979; Haspelmath, 2007). Whether a language possesses the grammatical relation subject, for example, can be determined on the basis of whether or not it displays a phenomenon that requires a grouping of argument types S (the argument of an intransitive sentence) and A (the agentive argument of a transitive sentence) to the exclusion of P (the patientive argument of a transitive sentence), since the S/A grouping is what the notion of subject represents in English and other European languages, in which the term subject has been invented. The following examples show the syntactic grouping of S and A arguments with regard to the distribution of pronominal case forms and the agreement pattern. The P argument is distinguished from these in case form (*her* as opposed to *she*) and in not triggering agreement:

19. a. She runs all the time. (S in 'she' form and agrees with the verb)
    b. She scolds them all the time. (A in 'she' form and agrees with the verb)

The phenomenon-driven approach to grammatical relations reveals that many of the so-called ergative languages involve subject (which groups S and A) and object (for P) relations as well as some other relations, which may be termed 'absolutive' (which groups S and P) and 'ergative' (for A) relations. In the following Warlbiri examples, agreement (or clitic placement) is controlled by the subject/object relations. The first person subjects (S and A) trigger *-ṇ* marking in the verb (see (20a) and (20c), whereas the first person object (P) triggers *-tju* marking (see (20b)). Thus, there is a clear-cut subject/object opposition in this language. Case forms, however, are controlled by a different set of relations. The absolutive relation (a grouping of S and P) demands the *ɲatju* form for the first person pronoun, as in (20a) and ((20b). On the other hand, the A argument holding the ergative relation appears in the ergative-marked *ɲatju-luḷu* form (20c):

20. Warlpiri (Australia; Hale, 1976)
    a. ɲatju ka-ṇa puḷa-mi
       1SG.ABS PRES-1SG shout-NONPAST
       'I am shouting/I shout.'
    b. ɲjuntu-luḷu ka-npa-tju ɲatju njanji
       2SG-ERG PRES-2SG-1SG 1SG.ABS see-NONPAST
       'You see me.'
    c. ɲatju-luḷu ka-ṇa-ɲku njuntu njanji
       1SG-ERG PRES-1SG-2SG 2SG.ABS see-NONPAST
       'I see you.'

Somewhat similar pattern emerges among Western Malayo-Polynesian languages, in which there are phenomena calling for both the subject relation (defined as a union of S and A, as above) and a distinct topic relation marked specially in some of these languages, as the *ang*-marking in Tagalog and the *'o*-marking in the Tsou language of Taiwan. Sasak, spoken in the Lombok Island in eastern Indonesia, presents one of the clearest cases of a language requiring both subject and topic relations. Sasak, like most Austronesian languages in Indonesia, has a two-way focus system, a remnant of the proto-Austronesian four-way focus system (see below), whereby transitive clauses can be coded either in the Actor-focus (AF) construction, where the A argument is aligned with the topic relation, or in the Patient-focus (PF) construction, where the P argument is the topic, as in (21a) and (21b):

21. Sasak (Puyung meno-mené dialect; Shibatani, 2008)
    a. Ali mu = *n* empuk aku. (AF construction)
       Ali PAST = 3 hit I
       'Ali hit me.'
    b. Aku mu = *n* empuk siq Ali. (PF construction)
       I PAST = 3 hit by Ali
       'Ali hit me.'
    c. Aku mu = *k* te-empuk siq Ali. (Passive)
       I PAST = 1 PASS-hit by Ali
       'I was hit by Ali.'
    d. Inaq mu = *n* lalo jok peken (Intransitive)
       mother PAST = 3 go to market
       'Mother went to the market.'

Cliticization operates according to the subject relation since S in (21d), the derived S of passive sentence (21c), and As in (21a.b) call for the cliticization of appropriate form, *-n* for a third person singular subject (21a.b.d) and *-k* for a first person singular subject (21c).

The topic in Sasak is not marked specially but appears sentence initially. Notice that PF constructions and passive constructions are two distinct constructions (cf (21b) and (21c)); PF constructions have the alignment patterns of A = SUB = Non-Topic and P = OBJ = Topic, whereas passives align P with subject as well as with topic, as P = SUB = Topic. There are several phenomena in which topics play a crucial role, including the controlling of a gap in coordinate constructions of the following type. Notice that in (22b) 'Ali' is an object

but it functions as the 'antecedent' of the gap in the second clause, while the subject 'Siti,' which triggers cliticization (see (21b)), does not:

22. Ganti meno-mené Sasak (author's field notes)
    a. Siti empuk Ali, terus Ø pelai=n (Ø =Siti)
       Siti hit Ali then run=3
       'Siti hit Ali, and ran away.'
       (AF: 'Siti'=A=SUB=TOP)
    b. Ali empuk=en siq Siti, terus Ø pelai=n (Ø =Ali, *Siti)
       Ali hit-3 by Siti then run=3
       'Siti hit Ali, and (Ali) ran away.'
       (PF: 'Ali'=P=OBJ=TOP)
    c. Ali te-empuk siq Siti, terus Ø pelai=n (Ø =Ali, *Siti)
       Ali PASS-hit by Siti then run-3
       'Ali was hit by Siti, and ran away.'
       (Pass: 'Ali'=P=SUB=TOP)

Besides these phenomena that specifically call for the subject and the topic relation, there are those that require the alignment of SUB = TOP such as the controlling of the gap in the '*try*-constructions' (Ali nyobaq [Ø ngiduk Siti] 'Ali tried to kiss Siti').

Just as the properties of subjects differ across different languages, topic properties also differ from one language to another. While topics of Western Malayo-Polynesian languages are involved in various syntactic phenomena, resembling subjects in European languages to a varying extent, topics in Asian languages such as Japanese and Korean are less relevant to those phenomena controlled by subjects in these and other languages.

Although the above survey shows that many languages, including controversial ergative and Western Malayo-Polynesian languages, may have a subject, the universality of this relational category must still be empirically established. Toward this end, future descriptive grammars must each demonstrate the need for positing grammatical relations of various sorts by the kind of procedure described above.

Structure, Use, and Form

Both traditional and modern grammarians have not been very successful in distinguishing structures (linguistic units or constructions), their uses (functions), and forms that may vary according to use and function. For example, 'possessor' expressions such as *my/mine*, *your/yours*, *his/his*, and so on have traditionally been analyzed as belonging to separate parts of speech (i.e., as distinct structures) depending on their use. They are recognized as 'possessive pronouns' ('posesivos pronombres' in Spanish) when they function as the head of a noun phrase, as in (23a), and as 'possessive adjectives' ('posesivos adjetivos') when they modify, as in (23b):

23. a. [Mine/yours/his/my younger brother's] are those piled up on the table over there.
    b. [my/your/his/my younger brother's] books.

Besides the problem of identifying forms such as *my younger brother's* either as pronouns or adjectives – they are not words – this kind of analysis confuses structures and their use. While many structures need to undergo category change when they play different functions, some do not. In particular, nouns in many languages can modify another noun without changing their category from noun to adjective. For example, the English noun *cotton* typically heads a noun phrase, as in (24a), where it realizes a referential function of the NP. A noun may, in addition, function as a modifier of another noun as in (24b):

24. a. $[[\text{Cotton}]_N]_{NP}$ is a useful commodity. (NP-use of a noun)
    b. $[\text{a } [[\text{cotton}]_N\ [\text{shirt}]_N]_{N'}]_{NP}$ (Modification-use of a noun)

That nouns do not become adjectives in their modification function is seen from the fact that they continue to be modified by an adjective, not an adverb that adjectives call for, as in the expressions *an Egyptian cotton shirt*, *developmental psychology students*, and so on. Interestingly so-called relative clauses also have the two types of use exemplified above:

25. a. You should marry [who you love].
    b. You should marry the man [who you love].

Capturing the parallelism between the two uses across these constructions requires proper understandings of the 'possessive' forms and so-called relative clauses so that we can describe these patterns in an analogous form as we do for the two uses of nouns seen in (24); that is, where structure X is used as the head of an NP as 'an NP-use of X,' and where X is used as a modifier as 'a modification-use of X.' Some measure of effort in capturing the relationship between the bracketed portions in (25a) and (25b) is seen in the typological literature, where forms like the one in (25a) are considered a type of relative clause, namely 'headless relative clauses' (Keenan, 1985; Kuroda, 1976; Andrews, 2007). A major problem with this widely accepted approach is that it is based on a wrong-headed perspective similar to characterizing the relevant forms in (23a) as 'headless possessive adjectives' and those in (24a) as 'headless modifying nouns.' Terms like 'headless relatives' and the alternative 'free relatives' are oxymora – things cannot be 'relatives' when they have nothing to relate to.

Indeed, the relevant form in (25a) does not modify, while the function of true relative clauses is to modify a head noun as in (25b). Instead of modifying, the form in (25a) and its equivalents in other languages actually denote thing-like concepts (e.g., those people evoked and characterized by the structure [who you love] in (25a)). That is, the form in question has a nominal property of denoting thing-like concepts. A more promising line of analysis for the patterns exemplified by (25a) and (25b) recognizes processes and structures termed 'grammatical nominalizations' (Shibatani, 2009; Yap et al., 2011). Unlike lexical nominalizations, which are nouns registered in the lexicon, grammatical nominalizations are not nouns – they typically have structures larger than words – and are created for the nonce.

Grammatical nominalization is a metonymic process that yields nominalized structures associated with various entity-denotations such as events, facts, propositions, places, reasons, times, and participants associated with the events evoked by the nominalized structures. Languages in general make

a formal distinction between 'event nominalizations' and 'argument/participant nominalizations,' as in the following examples, where these may differ in terms of marking and whether the structure contains a missing argument or not:

26. a. I know [that you love that man]. (Event nominalization)
    b. Marry [who you love __] (Argument nominalization)

The gap in an argument nominalization indicates the grammatical role that the evoked entity is to be understood to play. In (26b) the 'object nominalization' denotes a person that the addressee loves. While many languages (e.g., Chinese, Japanese, and Korean) have no marking other than the position of a gap in indicating the role of the argument denoted by an argument nominalization, many languages have special markers that indicate such roles (as well as some other information). The basic functions of the so-called focus markers in Austronesian languages are to mark the grammatical roles in argument nominalizations. The Actor-focus (AF), Patient-focus (PF), Locative-focus (LF), and Circumstantial-focus (CF) markers, respectively, indicate that the denoted argument is agentive, patientive, locative, and benefactive/instrumental, as in the following Mayrinax Atayal (Taiwan) examples:

27. Mayrinax Atayal argument nominalizations (based on Huang, 1995)
    a. Actor nominalization
    ßaq-un = mu ku? [m<in>aniq cu? goqiluh] ka? haca?
    know-PF = 1SG.GEN NOM.REF AF<PERF>eat ACC banana LINK that
    'I know that one who ate banana.'
    b. Patient nominalization
    ßaq-un = mu ku? [t<in>uting = si hisa]
    know-PF = 1SG.GEN NOM.REF <PF.PERF>beat = 2SG yesterday
    'I know the one you beat yesterday.'
    c. Locative nominalization
    ɣaɣhapuyan ku? [naniq-an cu? ßuŋa? nku? ?ulaqi?]
    kitchen NOM.REF eat-LF ACC.NONREF yam GEN.REF child
    'The kitchen is where the child eats yam.'
    d. Circumstantial (benefactive/instrumental) nominalization
    ini = mu s?wa? = i ku? [si = ghahapuy nku? kanairil]
    NEG = 1SG like = LF NOM.REF CF = cook GEN.REF woman
    'I don't like the one for whom the woman cooks.'

So-called German relative pronouns *der*, *den*, *dem*, *dessen* are in fact nominalizers of the same sort as the Austronesian focus markers, respectively, marking subject, object, indirect object, and genitive nominalizations. English nominalization markers include *who*, *whom*, *whose*, *why*, and *where*, which indicate the grammatical roles, and *which* and *what* contrasting with *who/whom* in marking inanimate denotations.

Now, parallel to the usage pattern of nouns in (24), event nominalizations have two major uses, as below:

28. a. I know [that you love that man]. (NP-use)
    b. I know the fact [that you love that man]. (Modification-use)

In (28a) an event nominalization heads an NP functioning as the object of the verb *know*, while in (28b) the same form modifies the head noun *fact*. This is precisely the way the pattern in (25) should be analyzed. Namely (25a) instantiates an NP-use of an argument nominalization, and (25b) initiates a modification-use of the same argument nominalization. What have been considered relative clauses are nothing but a modification-use of argument nominalizations. See the following Mayrinax Atayal examples, where the Actor and Patient nominalizations in (27) are used as a modifier together with the linker.

29. a. Modification-use of an Actor nominalization
    kanairil ka? [m<in>aniq cu? goqiluh]
    woman LINK AF-<PERF>eat ACC banana
    'woman who ate a banana'
    b. Modification-use of a Patient nominalization
    ?uliaqi? ka? [t<in>uting = si hisa]
    boy LINK <PF.PERF>beat = 2SG yesterday
    'boy whom you beat yesterday.'

Besides the ones based on clausal units as seen above, various structures can be nominalized grammatically. What we observed earlier in (23) exemplify nominalizations of pronouns and noun phrases. *Mine*, *my younger brother's*, and so on evoke denotations closely associated with the referents of the pronoun *I* and the noun phrase *my younger brother* – the most immediately evoked denotations being things that are possessed by these referents. Then, (23a) exemplifies an NP-use of nominal-based nominalizations, and (23b) a modification-use of the same N-based nominalizations. Notice that forms of the N-based nominalizations differ for some (*mine* vs *my*) according to their use/function, while some (*his* vs *his*) remain the same. It is often the case that the NP-use of nominalization calls for a special marker indicating its referential function as in the case of *one*-marking seen in nominalized prepositional phrases (*one[on the table]*) and nominalized adjectival phrases (*one [highly expensive and valuable]*). Notice that *one*-marking does not occur in the modification-use (*the book [on the table]*, *the book [highly expensive and valuable]*.) This is indeed what we observe between the two uses of argument nominalizations, as below (The use of numerals as

markers of the NP-use of nominalizations is also seen in the Indo-Aryan language Sinhala.):

30. a. I know the one [who __ came yesterday]. (NP-use)
    I know the man [who __ came yesterday]. (Modification-use)
31. b. Show me the one [which you bought __ yesterday]. (NP-use)
    Show me the book [which you bought __ yesterday]. (Modification-use)

The marginal NP-use of argument nominalizations without *one*-marking seen in Modern English above (see (26b)) is a remnant of a much more widespread use of these argument nominalizations as an NP-head without *one*-marker in an earlier form of English. The development of *one*-marking for the NP-use of argument nominalizations since the time of Middle English parallels the rise of *no*-marking of the NP-use of grammatical nominalizations in Japanese in the beginning of the sixteenth century. These markers of the NP-use of nominalizations, such as English *one*, Japanese *no*, and Korean *kes*, must be distinguished from nominalizing particles like Mandarin Chinese *de* and Nepali *ko*, which occur in both NP and modification functions. Compare the Korean and Mandarin Chinese examples below:

32. Korean argument nominalizations
    a. Na = nun [apeci = ka cwu-n] = *kes* = ul ilk-ko-iss-ta. (NP-use)
       I = TOP father = NOM give-NMLZR = NM = ACC read–CON–be-IND
       'I am reading what father gave (me)'
    b. [apeci = ka cwu-n] chaek (Modification-use)
       father = NOM give = NMLZR book
       '(the) book which father gave me'
33. Korean N-based nominalizations
    a. [emeni = uy] = *kes* = un ku chaek i-ta. (NP-use)
       mother = NMLZR = NM = TOP that book COP = IND
       'Mother's is that book.'
    b. [emeni = uy] chaek (Modification-use)
       mother = NMLZR book
       'mother's book'
34. Mandarin Chinese argument nominalizations
    a. Nǐ méi yǒu [wǒ xǐhuān = *de*] (NP-use)
       you not have I like = NMLZR
       'You don't have what I like.'
    b. [wǒ xǐhuān = *de*] yīfu (Modification-use)
       I like = NMLZ clothes
       '(the) clothes that I like'
35. Mandarin Chinese N-based nominalizations
    a. Zhèi běn shū shì [wǒ = *de*]. (NP-use)
       this CLF book is I=NMLZR
       'This book is mine.'
    b. [wǒ = *de*] shū (Modification-use)
       I=NMLZR book
       'my book'

The fact that some of these two types of nominalization markers are used for both argument nominalizations and N-based nominalizations supports the analysis of the latter (*mine/yours/my younger brother's*, and so on) as nominalizations (see also Matisoff, 1972).

Reanalyzing so-called relative clauses and so-called complement clauses such as (26a) as nominalizations has profound implications to the past traditional and generative, as well as typological studies of these constructions. Firstly, contrary to the traditional label relative 'clauses,' nominalizations are not clauses even if their internal structures look clausal with verbal syntax (see e.g., (28a)). Once various grammatical units are nominalized, they behave like nouns. In particular, nominalizations denote entity concepts rather than predicate, which is the function of clauses, and neither do they assert like (declarative) sentences. The nominalization *that you love that man* in (28a), for example, does not predicate or assert; instead it denotes a fact. The impression that this expression predicates is due to the presupposition that it is associated with. The same presupposition property obtains with so-called relative clauses.

The nominalization analysis also casts doubt on both traditional and more recent analyses of relative clauses within the Chomskyan generative framework, which assume that there is recursion/embedding of S or IP/CP, that is, either sentences or clauses (see (4)). The nominalization analysis claims that sentences and clauses cannot be embedded unless they are first nominalized. This is a desirable consequence. As discussed in the Section Phrase-Structure Rules, Phrasal Categories, and Constituency, constraining the grammar/theory has been one of the prime concerns of generative syntacticians. The current

theory is still very powerful, indeed too strong, since it allows embedding of S/IP/CP within a phrase smaller in structure than these, namely NP (in the case of N complements and relative clauses) and VP (for V complements). The nominalization analysis entails a much more constrained theory of syntax allowing recursion/embedding of only elements equivalent to or smaller than the hosting structures.

Finally, the nominalization analysis of relative clauses undermines the fundamental assumption underling the notion of NP-accessibility developed in one of the most influential typological studies of relative clauses by Keenan and Comrie (1977), where it is assumed that there are relativization processes that target NPs holding specific grammatical relations such as subject, direct object, and indirect object. This has yielded a typological classification of relative clauses – and the languages possessing them – on the basis of whether relativization is accessible/applicable only to subjects, to subjects and direct objects, to subjects, direct objects, and indirect objects, and so on. Within the nominalization analysis, relative clause formation is simply juxtaposing of a head noun and an argument nominalization, and no notion of grammatical relations is involved in the relativization process per se. Argument nominalizations are characterizable as subject nominalization, object nominalization, and so on, but this is totally independent from the process of relativization.

See also: Case, Morphological: Linguistics; Control, Syntactic; Functional Approaches to Grammar; Generative Grammar; Grammatical Relations; Linguistic Presupposition; Linguistic Typology; Parts of Speech.

Bibliography

Andrews, A., 2007. Relative clauses. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, Complex Constrictions, second ed., vol. 2. Cambridge University Press, Cambridge, pp. 206–236.
Blake, B.J., 1979. A Kalkatungu Grammar. Pacific Linguistics, Canberra.
Bresnan, J., 2001. Lexical Functional Syntax. Blackwell.
Chomsky, N., 1957. Syntactic Structures. Mouton, Hague.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Chomsky, N., 1970. Remarks on nominalization. In: Jacobs, R., Rosenbaum, P. (Eds.), Readings in Transformational Grammar. Ginn, Waltham, MA, pp. 184–221.
Croft, W., 1991. Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information. The University of Chicago Press, Chicago.
Dixon, R.M.W., 2002. Australian Languages: Their Nature and Development. Cambridge University Press, Cambridge.
Dixon, R.M.W., 1993. Adjective. Encyclopedia of Language and Linguistics, first ed. Elsevier.
Dixon, R.M.W., Aikhenvald, A. (Eds.), 2006. Adjective Classes: A Cross-Linguistic Typology. Oxford University Press.
Dryer, M., 1979. Are grammatical relations universal? In: Bybee, J., Haiman, J., Thompson, S.A. (Eds.), Essays on Language Function and Language Type. John Benjamins, Amsterdam, pp. 115–143.
Greenberg, J., 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (Ed.), Universals of Grammar. MIT Press, Cambridge, MA, pp. 73–113.
Hale, K., 1976. On ergative and locative suffixal alternations in Australian languages. In: Dixon, R.M.W. (Ed.), Grammatical Categories in Australian Languages. A.I.A.S, Canberra, pp. 414–417.
Hale, K., 1983. Warlpiri and the grammar of non-configurational languages. Natural Language & Linguistic Theory 1, 5–47.
Haspelmath, M., 2007. Pre-established categories don't exist: consequences for language description and typology. Linguistic Typology 11.1, 119–132.
Huang, M.-C., 1995. A Study of Mayrinax Syntax. Crane, Taipei.
Jackendoff, R., 1977. X-Bar-Syntax: A Study of Phrase Structure. MIT Press, Cambridge, MA.
Keenan, E.L., 1976. Towards a universal definition of 'subject'. In: Li, C.N. (Ed.), Subject and Topic. Academic Press, New York, pp. 303–333.
Keenan, E.L., 1985. Relative clauses. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, Complex Constrictions, vol. 2. Cambridge University Press, Cambridge, pp. 141–170.
Keenan, E.L., Comrie, B., 1977. Noun phrase accessibility and universal grammar. Linguistic Inquiry 8, 63–100.
Kuroda, S.-Y., 1976. Types of Japanese relatives. In: Kuroda, S.-Y. (Ed.), Japanese Syntax and Semantics: Collected Papers. (1992). Kluwer, Dordrecht, pp. 157–169.
Matisoff, J.A., 1972. Lahu nominalization, relativization, and genitivization. In: Kimball, J.P. (Ed.), Syntax and Semantics, vol. 1. Seminar Press, New York and London, pp. 237–257.
Nagaya, N., 2011. The Lamaholot Language of Eastern Indonesia (Doctoral thesis). Rice University.
Perlmutter, D.M. (Ed.), 1983. Studies in Relational Grammar 1. Chicago University Press, Chicago.
Radford, A., 1997. Syntactic Theory and the Structure of English. Cambridge University Press, Cambridge.
Shibatani, M., 2008. Relativization in Sasak and Sumbawa, Eastern Indonesia. Language and Linguistics 9.4, 865–916.
Shibatani, M., 2009. Elements of complex structures, where recursion isn't: the case of relativization. In: Givón, T., Shibatani, M. (Eds.), Syntactic Complexity. John Benjamins, Amsterdam, pp. 163–198.
Sportiche, D., Koopman, H., Stabler, E., 2013. An Introduction to Syntactic Analysis and Theory. Wiley-Blackwell.
Van Valin, R.D., 2001. An Introduction to Syntax. Cambridge University Press, Cambridge.
Yap, F.H., Grunow-Hårsta, K., Wrona, J. (Eds.), 2011. Nominalization in Asian Languages: Diachronic and Typological Perspectives. John Benjamins, Amsterdam.

Syntax–Phonology Interface

Elisabeth Selkirk, University of Massachusetts, Amherst, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15407–15412, © 2001, Elsevier Ltd.

Abstract

The bulk of evidence concerning the syntax–phonology interface shows an influence of syntax on phonology. The domain structure for sentence-level phonological and phonetic phenomena, which forms part of the surface phonological representation (PR) of the sentence, is defined through an interaction of two types of constraints: syntax–prosodic structure interface constraints, which call for certain properties of the surface syntactic representation of the sentence (PF) to be reflected in domain structure in PR, and prosodic structure markedness constraints, which call for the surface prosodic structure to display patterns of unmarked prosodic structure. The effects of prosodic markedness constraints argue against direct access theories, which see phonological phenomena as defined directly on the surface syntax. Distinguishing PF and PR raises the question whether PF is input to the phonological component, with PR the output, as in standard models of generative grammar, or whether there may be mutual influence. Current models of grammar would countenance effects in the other direction, with the possibility of phonological principles constraining the range of acceptable surface syntactic representations, and research is beginning to explore this area.

The bulk of evidence concerning the syntax–phonology interface shows an influence of syntax on phonology. The domain structure for sentence-level phonological and phonetic phenomena, which forms part of the surface phonological representation (PR) of the sentence, is defined through an interaction of two types of constraints: *syntax–prosodic structure interface constraints*, which call for certain properties of the surface syntactic representation of the sentence (PF) to be reflected in domain structure in PR, and *prosodic structure markedness constraints*, which call for the surface prosodic structure to display patterns of unmarked prosodic structure. The effects of prosodic markedness constraints argue against direct access theories, which see phonological phenomena as defined directly on the surface syntax. Distinguishing PF and PR raises the question whether PF is input to the phonological component, with PR the output, as in standard models of generative grammar, or whether there may be mutual influence. Current models of grammar would countenance effects in the other direction, with the possibility of phonological principles constraining the range of acceptable surface syntactic representations, and research is beginning to explore this area.

An individual word, pronounced in isolation, will have a characteristic pronunciation. Yet the pronunciation of a sentence, made up of a sequence of words, is not merely a stringing together of these individual pronunciations. The pronunciation of a sentence, i.e., its phonetic realization, is a function of the surface phonological representation of the sentence. Standard models of the organization of a grammar hold that this phonological representation interfaces with the surface syntactic representation of the sentence, the latter being made up of the sequence of the underlying phonological representations of the words and morphemes that are actually pronounced, and their organization into a syntactic word and phrase structure. This organization of the grammar permits in principle that the syntactic representation of a sentence may influence its phonological representation, and therefore allows an explanation for why sentence phonology is not simply the phonology of the individual words of the sentence strung together. Consider, for example, the sentence pairs *No people will go* vs. *No, people will go* and *People will go happily* vs. *People will go, happily*. In both cases, commas indicate a difference in syntactic structure within the pairs. That syntactic structure difference is reflected in a difference in pronunciation, specifically here in a difference in intonation. Effects like these of syntax on phonology have been studied most commonly, in part because earlier models of generative grammar sanctioned only effects in this direction. It is also possible that effects might go in the opposite direction, with phonological principles constraining the range of acceptable syntactic representations at the interface; some recent research explores this question.

Two major lines of thinking on the influence of syntax on phonology can be distinguished. The *prosodic structure hypothesis* holds that the phonological representation of a sentence is organized into a prosodic constituent structure which is independent of, but related to, the surface syntactic structure of a sentence (Selkirk, 1986, 1995; Nespor and Vogel, 1986; Truckenbrodt, 1999). It claims, moreover, that the syntax of a sentence can impinge directly only on this prosodic structure, namely on the organization into phonological words or phrases, and on the distribution of prosodic heads (or stresses). Other apparent effects of syntax on phonology, e.g., the choice of segmental variants or the placement of certain phrasal tones, are claimed to be indirect, mediated by the prosodic structure organization. And since the phonetic interpretation of a sentence is based on surface phonological representation, it follows that there are no direct effects of syntax on phonetic form. If the prosodic structure hypothesis embodies the correct theory of the domain structure for sentence phonology, then the theory of grammar must distinguish between the surface phonological representation of a sentence (PR), which would include this prosodic structure, and its surface syntactic representation (PF). The *direct access hypothesis* (Kaisse, 1985; Odden, 1995), by contrast, holds that phenomena of sentence phonology and phonetics may be defined directly from the surface syntactic word and phrase structure. It does not require a theory of grammar that distinguishes the surface syntactic representation of a sentence crucially from its surface

phonological representation, at least in its hierarchical structure. Evidence appears to favor the prosodic structure hypothesis.

The Prosodic Structure Hypothesis

A prosodic structure is a well-formed labeled bracketing or tree. The constituents of prosodic structure belong to distinct prosodic categories, arranged in a *prosodic hierarchy*.

(1) The Prosodic Hierarchy:
Utterance (Utt)
Intonational Phrase (IP)
Major Phonological Phrase (MaP)
Minor Phonological Phrase (MiP)
Prosodic Word (PWd)
Foot
Syllable

In the unmarked case, it is claimed, prosodic structure is *strictly layered*, in the sense that a constituent of a higher level in the hierarchy immediately dominates only constituents of the next level down in the hierarchy. In addition, within a prosodic constituent, in the unmarked case, one of the daughter constituents constitutes the *prosodic head*, the locus of prominence or stress. Compelling support for the claim that the domain structure of sentence phonology and phonetics has the formal properties of such a structure comes from evidence of *domain convergence* and *domain layering* within individual languages. Yet a direct access approach could potentially model these effects. Crucial support for the prosodic structure hypothesis comes from the fact that the constraints which define the hierarchical phonological domain structure are heterogeneous in type, including *prosodic structure markedness constraints* which are properly phonological in character, appealing only to properties of surface phonological representation, and *syntax-prosodic structure interface constraints*, which call for features of the syntactic representation to be reflected in phonological representation. Effects of prosodic markedness constraints show that the domain structure of surface phonology is not strictly determined by the syntax, only influenced by it.

Domain Convergence

In English declarative sentences, the right edge of a major phrase (MaP) in prosodic structure is marked by the presence of a low tone. In (1a,b) a L- marks the low target tone found on the final syllable of a MaP.

(2) (a) (No^{L-}) (animals are allowed^{L-}) =
No, animals are allowed.
[ɨænɨmlz]
(b) (No animals are allowed^{L-}) =
No animals are allowed.
[ænɨmlz]

We also see a segmental reflex of that same MaP organization: a glottal stop appears as the onset of the first syllable of *animals*, [ɨænɨmlz], in (2a), where it is phrase-initial, but not in (2b). This glottalization effect is arguably phonetic in character (Dilley and Shattuck-Hufnagel, 1996). In (3a,b) we see an additional MaP effect: a contrast in the pronunciation of the function word *to*.

(3) (a) (They're allowed to graze there^{L-}) (by law^{L-}).
[tə]
(b) (They're allowed to^{L-}) (by law^{L-})
[tuw]

To appears in its stressless, vowel-reduced weak from [tə] in (3a), where it is medial within the MaP, but in its stressed, full vowel strong form [tuw] when it appears at the right edge of MaP in (3b) due to the ellipsis of the following verb phrase. In English, monosyllabic function words like *to* generally appear in strong form only at the right edge of MaP (or when they bear contrastive stress) (Selkirk, 1995). These three different phenomena which converge on major phrase in English together show that both edges of these constituents are simultaneously relevant to defining phonetic and phonological phenomena.

In Bengali sentence phonology (Hayes and Lahiri, 1991), MaP is marked at its right edge by a H- tone. The left edge of MaP is the locus of further phenomena: the word that is leftmost in a MaP is the head of that phrase, and a Low pitch accent, L*, is located on the initial syllable of this stressed word:

(4) (ʃæ^{L*}moli^{H-}) (ra^{L*}m-er bari^{H-}) (ɖ^{h}u^{L*}ketʃ^{h}il^{H-})$^{L°}$
Shamoli Ram's house entered
Shamoli entered Ram's house.

In addition, there are segmental assimilation phenomena which apply optionally both within and between words. These assimilations operate across the span of the MaP, but are blocked if the sequence of segments belongs to different phrases:

(5) (ɔ^{L*}mor^{H-}) (tʃa^{L*}dor^{H-}) (ta^{L*}ra-ke^{H-}) (díe^{L*}tʃ^{h}e^{H-})$^{L°}$
[r] [tʃ] [r] [t]
Amor scarf Tara–to gave
Amor gave the scarf to Tara.
(6) (ɔ^{L*}mor tʃador tara-ke^{H-}) (díe^{L*}tʃ^{h}e^{H-})$^{L°}$
[tʃtʃ] [tt]
Amor gave the scarf to Tara.

The faster pronunciation in (6) is organized into fewer MaP than in the more deliberate pronunciation of the same sentence in (5). This difference in domain organization is indicated both by the patterns of tone distribution and by the assimilation of /r/. This Bengali phrasal domain, then, shows right and left domain-edge phenomena, as well as phenomena defined on the prominent head of the domain and across the span of the domain.

Domain Layering

The edge of a clause in the syntax typically coincides with the edge of what prosodic structure theory calls the intonational phrase (IP), the prosodic constituent immediately superordinate to the MaP in the prosodic hierarchy. In English, a final rising tonal contour on the last syllable of word (notated$^{L\text{-}H°}$), sometimes referred to as the continuation rise, is taken to indicate the presence of the right edge of IP (Beckman and Pierrehumbert, 1986), indicated with curly braces.

(7) {(Since herds of grazing cows^{L-})
[hrdz]
(have been allowed in her meadow^{L-})$^{H\%}$}
[əv] [ɾ]
{(her flowers have all disappeared^{L-})}
[hɾ] [əv]

The right edge of an IP typically is accompanied by a short pause or a significant lengthening of the final syllable. In English (and other languages) the left edge of IP is, moreover, the site of phonetic strengthening effects that are potentially different in kind or in degree from those seen with the major phrase (Keating et al., 1998). For example, an underlying /h/ in English will typically fail to be pronounced if it is in an unstressed syllable (compare stressed *herds* to stressless *her* in the first clause), but at the beginning of an IP /h/ must necessarily be pronounced, even in a stressless syllable (compare stressless *her* in the second clause to stressless *her* in the first). By contrast, there is the possibility of absence of /h/ in the stressless *have* which is MaP-initial in the first clause. This difference in appearance of [h] provides further testimony of the difference in the two levels of domain structure.

In Japanese, a rise (LH) in pitch which is referred to as Initial Lowering is found at the left edge of a minor phrase (MiP) (Poser, 1984; Kubozono, 1993), indicated with angle brackets in (8b).

(8) (a) Yamaai-no yamagoya-no uraniwa-no umagoya-ni kabi-ga hae^{H*}-ta
[[[[mountain village-GEN] hut-GEN] backyard-GEN] barn-LOC] [mold-NOM] [grow-PAST]
The barn in the backyard of a hut in a mountain village grew moldy.
(b) $^{L}<^{H}$ Yamaai-no yamagoya-no$>^{L}$ $<^{H}$ uraniwa-no umagoya-ni$>\uparrow^{L}<^{H}$ kabi-ga hae$^{H*}\downarrow$-ta$>)^{L}$

Sentence (8b) shows an organization into binary MiP. In (9b), by contrast, where all the words have a lexical accent, we see that there are as many instances of MiP, and hence of Initial Lowering, as there are of accents:

(9) (a) Yama^{H*} gata-no yama^{H*} dera-no one:^{H*} san-ga mayone:^{H*} zu-o ho:ba^{H*} tteiru-wa
[[[Yamagata-GEN] mountain temple-GEN] young-lady-NOM] [[mayonnaise-ACC] filling-her-mouth]
The young lady from a mountain temple in Yamagata was filling her mouth with mayonnaise.
(b) ($^{L}<^{H}$yama$^{H*}\downarrow$gata$>^{L}<^{H}$yama$^{H*}\downarrow$dera$>^{L}$ $<^{H}$one:$^{H*}\downarrow$san-ga$>\uparrow^{L}<^{H}$ mayone:$^{H*}\downarrow$zu-o$>^{L}<^{H}$ho:ba$^{H*}\downarrow$tteiruwa$>)^{L}$

The difference in patterns of minor phrasing seen in (8) and (9) are the result of prosodic markedness constraints, to be discussed below.

At the level of MaP, the level above MiP in the prosodic hierarchy, two tonal phenomena, are defined: catathesis (↓), and upward pitch reset (↑) (Poser, 1984; Pierrehumbert and Beckman, 1988). Catathesis is an accent-induced downstepping of the pitch range, while upward pitch reset returns the pitch range to a higher level. Only material within the same MaP undergoes catathesis. Upwards reset, seen in both (13) and (14), appears at the left edge of MaP. These examples, then, illustrate the basic strict layering of unmarked hierarchical prosodic structure: here a MaP is parsed into a sequence of MiP.

Nonsyntactic Determinants of Phonological and Phonetic Domain Structure

A direct access theory could model patterns of edge-sensitive phonetic and phonological phenomena as long as the edges were definable solely in syntactic terms. But demonstrably nonsyntactic factors in phonological domain organization, such as prosodic markedness constraints and speech rate, show that the influence of syntax is indirect.

The interaction of prosodic markedness constraints and interface constraints

In optimality theory (Prince and Smolensky, 1993), phonological markedness constraints call for some ideal phonological target output shape. Of particular importance to our topic, markedness constraints compete with a class of constraints called faithfulness constraints, which call for properties of the surface phonological representation to reflect, or be identical to, properties of a related grammatical representation. Syntax–phonology interface constraints can be understood as a variety of faithfulness constraint, one of the input–output variety, if surface syntactic representation (PF) and surface phonological representation (PR) are in an input–output relation, or one of an output–output variety, if the PF and PR relation are evaluated in parallel, not serially.

In the examples of Japanese minor phrasing cited in (8) and (9), prosodic markedness constraints complement the effects of syntax–phonology interface constraints, producing phrasing where it would not be called for by the syntax. A prosodic constraint calling for a MiP to be binary in its PWd composition, call it Binary MiP, results in the binary phrasing seen in (8). Another prosodic constraint, call it MiP Accent, calls for a MiP to contain at most one accent; it is responsible for the appearance of the nonbinary MiP in (14). In Japanese, MiP Accent takes precedence over Binary MiP, as in (9), a relation that can be expressed in the grammar of Japanese with an optimality-theoretic constraint ranking: MiP Accent ≫ Binary MiP.

The ideal binary shape for MiP, which appears in (8), fails to appear in sentences with somewhat different syntactic structures, showing that certain syntax–phonology interface constraints also outrank the Binary MiP markedness constraint in the grammar of Japanese. The simple three word subject–object verb sentence (10) divides into two major phrases (MaP) at the break between subject and object, with the result that the subject consists of a nonbinary minor phrase (MiP), in violation of Binary MiP:

(10) (a) Inayama-ga yuujin-o yonda
[[Inayama-NOM] [[friend-ACC] call-PST]]
Mr Inayama called his friend.
(b) ($^{L}<^{H}$Inayama-ga$>\uparrow$($^{L}<^{H}$yuujin-o yonda$>)^{L}$

The MaP break here is imposed by a syntax–phonology interface constraint expressible as the alignment constraint (Selkirk and Tateishi, 1988, 1991):

(11) Align-L (XP, MaP)
The left edge of a maximal projection in syntactic representation (PF) corresponds to the left edge of a major phrase in surface phonological representation (PR).

The noun phrase (12), which consists of an accentless Adjective–Adjective–Noun sequence, consists of a single MaP (as shown by lack of internal upward pitch range reset), but nonetheless shows a MiP break between the two adjectives, leaving the first MiP nonbinary.

(12) (a) Amai akai ame-ga
[sweet [red [candies]]]
(b) $({}^{L}<{}^{H}\text{amai}>{}^{L}<{}^{H}\text{akai ame-ga}>)^{L}$
NOT: $*({}^{L}<{}^{H}\text{amai akai ame}>)^{L}$

At play here is an interface constraint sensitive to syntactic branching (Kubozono, 1993):

(13) Align- L (X^{br}, MiP).
The left edge of a branching constituent in syntactic structure corresponds to the left edge of a minor phrase in prosodic structure.

These two syntax–phonology alignment constraints outrank Binary MiP in the grammar of Japanese, causing the violations of Binary MiP seen in (10) and (12):

(14) Align-L (XP, MaP), Align- L (X^{br}, MiP) ≫ Binary MiP.

It is because neither of these interface constraints is applicable in the left-branching structure of the subject noun phrase in (8) that we see the effects of the lower-ranked Binary MiP emerge there. In other words, where the lower-ranked Binary MiP is not forced to be violated by the higher ranked interface constraints, it produces effects that are *complementary* to the effects of the interface constraints.

We also find cases where an expected effect of a syntax–phonology interface constraint *fails to appear* because of pressure from a competing prosodic markedness constraint. This can be modeled by ranking the prosodic constraint higher than the relevant interface constraint. (15) shows the expected MaP break at the left edge of VP in a neutral Focus sentence in Japanese.

(15) A neutral Focus sentence in Japanese

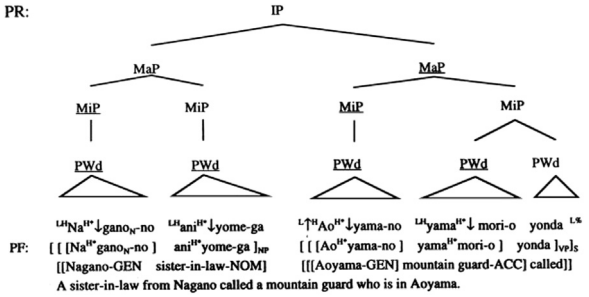

Sentence (16) contains a text which differs only in having a contrastive Focus on the last word of the subject NP, *ani'yome*. The phonological domain structure of this sentence shows an absence of the syntactically expected MaP break at the left edge of the VP, as well as the unexpected presence of a MaP break at the left edge of the Focus constituent (Poser, 1984; Pierrehumbert and Beckman, 1988; Nagahara, 1994).

(16) The same sentence, except with *aniyome* as FOCUS.

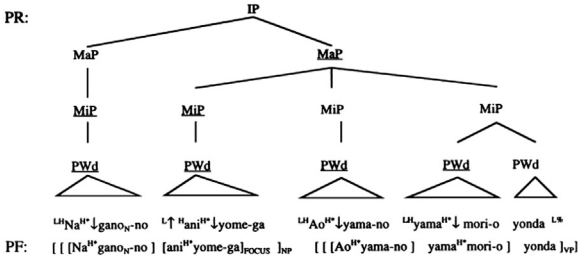

Truckenbrodt (1995) argues that the source of these phrasing effects in Focus sentences lies in a syntax–phonology interface constraint calling for a Focus to be more prosodically prominent than any other element within the Focus domain.

(17) Focus Prominence
The string in PR which corresponds to a string in PF dominated by Focus must contain a prosodic prominence that is greater than any other prominence contained in the Focus domain.

In the case where the domain of the Focus is the entire sentence, as in (16), respect of this constraint means that the Focus element and all the prosodic phrases that dominate it will be the head(s) of the intonational phrase corresponding to that sentence in PR (underlined). A prominent constituent is subject to a prosodic markedness constraint that calls for it to be aligned with an edge of the superordinate constituent (McCarthy and Prince, 1993). The violation in (17) of the interface constraint (11) calling for a MaP edge at the left edge of VP is claimed to be driven by the higher-ranked prosodic constraint Align-R (MaP_{head}, IP).

This constraint calls for the prominent MaP in the IP that dominates the Focus in (17) to align with the right edge of the intonational phrase, and so prevents the existence of any other MaP following the Focus. This analysis is captured in the ranking in (18).

(18) Focus Prominence, Align-R (MaP_{head}, IP) ≫ Align-L (XP, MaP).

So, here, a syntactic effect is overridden by a prosodic effect.

Speech Rate

One additional sort of nonsyntactic factor on phonological domain structure that has been recognized for some time in the literature is speech rate. In Bengali, a faster speech rate gives rise to larger phonological phrases (see Section Domain Convergence). Speech rate plays a role as well in MiP organization in Japanese; in slower speech, single words may easily constitute a MiP (Selkirk and Tateishi, 1988). Since syntactic structure does not vary with speech rate, these

rate-based differences in phrasing give additional evidence for an autonomous, nonsyntactic, prosodic structural representation of phonological domains.

The Syntax–Prosodic Structure Interface: Serialist or Parallelist?

The classic generative model of the relation between syntax and phonology is an input–output model, in which surface syntactic representation, the output of the syntactic component, is input to the phonological component, whose output is the surface phonological representation. Most approaches to modelling the syntax–prosodic structure interface have assumed this serialist, unidirectional, input–output model, with the prediction that syntax may influence phonology, but not vice versa. They have posited constraints, algorithms or rules relating the two representations which have the general form 'If the surface syntactic representation of the sentence has property S, then the surface phonological representation will have property P.'

In the Align/Wrap theory of the interface (Selkirk, 1986, 1995; Truckenbrodt, 1999), the demarcational Align class of constraints calls for the edge of any syntactic constituent of type C_S in the surface syntax to align with the edge of a prosodic constituent of type C_P in the surface phonology, while the cohesional Wrap class of constraints calls for a syntactic constituent of type C_S in the surface syntax to be contained within a prosodic constituent of type C_P in the surface phonology. There have been other, relation-based, proposals concerning the nature of syntax–prosodic structure interface constraints as well, including Nespor and Vogel, 1986; McHugh, 1990; Kim, 1997, all understood to be consistent with the prosodic structure hypothesis, and with the input–output architecture.

An input–output model of the syntax–phonology interface makes the essential claim that constraints on phonological output representation do not interact with constraints on the surface syntactic representation. Two sorts of apparent challenges for the input–output model have recurred in recent literature. One sort concerns clitics whose distribution is governed by both syntactic and prosodic constraints, such as the second position clitics of Serbo-Croatian (Inkelas and Zec, 1990). Yet because the syntax itself affords options in the positioning of second position clitics, it is not necessary to construe the prosodic subcategorization of a clitic as outranking a constraint on syntactic word order. Another sort of challenge concerns the alleged dependence of word order on the positioning of focus-related prosodic prominences in the sentence, as suggested by Vallduvi (1991) for Catalan, or Zubizarreta (1998) for Spanish. However, since the latter sort of phenomenon potentially can receive a purely syntactic treatment in which the Focus properties of syntactic representation are crucial in determining word order (see, e.g., Grimshaw and Samek-Lodovici, 1998), this sort of case is not yet compelling. It remains to be seen whether the syntax–phonology interface is an input–output relation or whether it is instead a two-way street.

See also: Generative Grammar; Phonology.

Bibliography

Beckman, M., Pierrehumbert, J., 1986. Intonational structure in English and Japanese. Phonology 3, 255–309.

Dilley, L., Shattuck-Hufnagel, S., 1996. Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24, 424–444.

Grimshaw, J., Samek-Lodovici, V., 1998. Optimal subjects and subject universals. In: Barbosa, P., Fox, D., Hagstrom, P., McGinnis, M., Pesetsky, D. (Eds.), Is the Best Good Enough? Optimality and Competition in Syntax. MIT Press, Cambridge, MA, pp. 193–220.

Hayes, B., Lahiri, A., 1991. Bengali intonational phonology. Natural Language and Linguistic Theory 9, 47–96.

Inkelas, S., Zec, D., 1990. Prosodically constrained syntax. In: Inkelas, S., Zec, D. (Eds.), The Phonology–Syntax Connection. CSLI, Chicago, IL, pp. 365–378.

Kaisse, E.M., 1985. Connected Speech: The Interaction of Syntax and Phonology. Academic Press, San Diego, CA.

Keating, P., Fougeron, C., Cho, T., Hsu, C.-S., 1998. Domain-initial articulatory strengthening. In: Papers in Laboratory Phonology, 6. Cambridge University Press, Cambridge, UK.

Kim, N.-J., 1997. Tones, Segments, and their Interaction in North Kyungsang Korea. Ohio State University Ph.D. diss. Columbus, OH.

Kubozono, H., 1993. The Organization of Japanese Prosody. Kurosio, Tokyo.

McCarthy, J.J., Prince, A., 1993. Generalized alignment. In: Booij, G., Marle, J. (Eds.), Yearbook of Morphology. Kluwer, Dordrecht, The Netherlands, pp. 79–153.

McHugh, B., 1990. The phrasal cycle in Kirunjo Chaga tonology. In: Inkelas, S., Zec, D. (Eds.), The Phonology–Syntax Connection. University of Chicago Press, Chicago, IL, pp. 217–242.

Nagahara, H., 1994. Phonological Phrasing in Japanese. UCLA Ph.D. diss. Los Angeles.

Nespor, M., Vogel, I., 1986. Prosodic Phonology. Foris, Dordrecht, The Netherlands.

Odden, D., 1995. Phonology at the phrasal level in Bantu. In: Katamba, F. (Ed.), Bantu Phonology and Morphology. LINCOM EUROPA, Munich, Germany, pp. 40–68.

Pierrehumbert, J., Beckman, M., 1988. Japanese Tone Structure. MIT Press, Cambridge, MA.

Poser, W.J., 1984. The Phonetics and Phonology of Tone and Intonation in Japanese. Massachusetts Institute of Technology Ph.D. diss. Cambridge, MA.

Prince, A., Smolensky, P., 1993. Optimality theory: Constraint interaction in generative grammar. Ms. Rutgers University Center for Cognitive Science, New Brunswick, NJ.

Selkirk, E., 1986. On derived domains in sentence phonology. Phonology 3, 371–405.

Selkirk, E., 1995. The prosodic structure of function words. In: Beckman, J., Walsh Dickey, L., Urbanczyk, S. (Eds.), Papers in Optimality Theory. GLSA, Amherst, MA, pp. 439–470.

Selkirk, E., Tateishi, K., 1988. Constraints on minor phrase formation in Japanese. In: Proceedings of the 24th Annual Meeting of the Chicago Linguistics Society. Chicago Linguistic Society, Chicago, IL.

Selkirk, E., Tateishi, K., 1991. Syntax and downstep in Japanese. In: Georgopoulos, C., Ishihara, R. (Eds.), Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda. Kluwer, Dordrecht, The Netherlands, pp. 519–544.

Truckenbrodt, H., 1995. Phonological Phrases: Their Relation to Syntax, Focus and Prominence. Massachusetts Institute of Technology Ph.D. diss. Cambridge, MA.

Truckenbrodt, H., 1999. On the relation between syntactic phrases and phonological phrases. Linguistic Inquiry 30, 219–256.

Vallduvi, E., 1991. The role of plasticity in the association of focus and prominence. In: Proceedings of the Eastern States Conference on Linguistics (ESCOL), 7, pp. 295–306.

Zubizarreta, M.L., 1998. Prosody, Focus and Word Order. MIT Press, Cambridge, MA.

Syntax–Semantics Interface

Erkki Luuk, University of Tartu, Tartu, Estonia; and Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

I introduce the topic by pointing out some common problems and disagreements in defining the notions of syntax and semantics. Avoiding theory-specificity and adhering to no particular linguistic school, the article makes minimal assumptions about the notions to arrive at a description of syntax–semantics interface. Among the topics discussed are logic, truth, meaning and their relevance to syntax and semantics, syntactic and semantic validity, and semantics' relations to core areas of syntax (word order, grammatical relations, and constituency). The main claims are (1) a robust correlation between semantic and syntactic validity and (2) that grammatical relations and constituency are largely determined by semantics, while certain forms of word order are autonomous.

Introduction: Syntax/Semantics

The notions of syntax and semantics are widely used beyond linguistics (e.g., in mathematical logic and computer science, not to mention their colloquial usage) but the present article focuses on their application on natural languages (Semantics). In linguistics, the notions are by no means strictly defined, and estimates as to what is what and where to draw the line between the two vary considerably (Jackendoff, 2002: pp. 268–277). Thus, at this stage, it is useful to think of the domain we are interested in as syntax/semantics, and to keep in mind that different theories partition it into syntax and semantics in different ways. A (or perhaps the) crucial difference between different schools is whether phrasal and sentential semantics is allowed. In generative grammar (originating in Chomsky, 1964 [1957]), at least in its recent emanation (Berwick et al., 2013), only lexical items (basically, words and morphemes) can have semantics (*see* Generative Grammar; Lexicon). By contrast, there are approaches focusing mainly on phrasal and sentential semantics, e.g., frame and formal semantics and semantic syntax (originating in Fillmore, 1976; Montague, 1970; Seuren, 1996, respectively). This picture is further complicated by the fact that syntax and semantics are already very different by their nature (see Sections Syntax; Semantics; and Syntax–Semantics Interface: Preliminaries for details). While everybody regards meaning as central to semantics, views on what is (central to) syntax are much more variegated. Syntax has at least three distinct focal parts – word order, grammatical relations, and constituency – and their roles, contents, and interrelations are widely debated by different theories (see Sections Constituency; Grammatical Relations; and Word Order for details; (cf Word Order; Grammatical Relations)). In some ramified approaches to sentential semantics, it may be possible to derive grammatical relations and hierarchical sentence structure (constituency) from semantics, leaving only the cross-linguistically highly variable word order for syntax. The gulf between what could be termed as syntacto- and semantocentric views on syntax/semantics is also evident in their claims about the universality of syntax or semantics. For generative grammarians, syntax is universal and semantics (except for the general principle of form-meaning relation) highly language-specific, while a semantocentric view is the direct opposite. In addition, the approaches have generally opposing research agendas. As referred above, a semantocentric agenda is largely shaped by questions like how much of constituency, grammatical relations, and word order can be derived from semantics and pragmatics. By contrast, in the history of generative grammar, the role of semantics has gradually diminished and the function of syntax expanded (cf Berwick et al., 2013; Chomsky, 1964[1957]; Jackendoff, 2002).

Due to innumerable fundamental discrepancies between different approaches and schools, as well as for some peculiarities of natural language and its study, it seems impossible to give an objective overview of syntax/semantics, much less of the syntax–semantics interface. However, if we ever plan to get anywhere with this we would better start somewhere.

Partitioning Syntax/Semantics

To arrive at a description of syntax–semantics interface, we must start with a few axioms about syntax and semantics. Given the state of theorizing in this field, one cannot hope for our axioms to be both strong and universal (or uncontroversial); nevertheless, if we want to progress we have to posit some. In reading indiscriminately what has been written about syntax/semantics in the last few hundred years, one may observe the following general principle:

1. Syntax pertains to the ordering of elements, while semantics deals with their meanings.

Usually, the 'elements' are meaningful elements – thus, for example, the ordering of phonemes does not belong to syntax. However, this is by no means uncontested. According to Berwick et al. (2013), syntax is also responsible for the ordering of phonemes. Admittedly, this is something with which no other approach that I know of concurs. Besides, when viewed like this, the scope of syntax would be much wider than that of semantics, which deals only with meaningful elements by our definition. Thus we are tempted to set the additional condition:

2. The elements in (1) are meaningful.

From (1) and (2), one can deduce an important dependency:

3. Syntax depends on semantics but the converse does not hold.

As a test that the converse does not hold, take any known word in natural language, and observe that it has meaning irrespective of its context. This follows from the definition of elementary word: a minimal unit of language understood (but not necessary used) outside context (Luuk, 2010). As another test, consider the sentence:

4. *He teached them*,

which is morphosyntactically (see Section Syntax) invalid but semantically valid (i.e., readily understood). We end this section with an observation which, although not particularly illuminating or important, is still worth mentioning: Both the objects of study and the disciplines studying them are called syntax and semantics, so we should alert ourselves to this potential difference in these words' meanings.

Syntax

The context, i.e., the ordering meant (and studied) by syntax, is fundamentally determined by the condition of temporal representation over serial channel. In spoken languages this manifests itself as a sequence of sounds, in written language as the scanning sequence in reading, in signed language as a sequence of manual signs. There is a subtle point as to whether the order of morphemes (the smallest meaningful units in natural language) belongs to the domain of syntax. On one hand, many morphemes are words, but on the other hand, there is an entire discipline (and level of language) – morphology – generally viewed as distinct from syntax dealing with them (*see* Morphology in Linguistics). In fact, quite a few researchers would be willing to dismiss syntax and morphology as independent disciplines altogether and speak of morphosyntax instead (cf Hengeveld and Mackenzie, 2008). From a logical viewpoint, as a study of orderings of linguistic signs, morphosyntax would make perfect sense. (In this article, linguistic sign and linguistic expression mean the same.) However, unfortunately 'morphosyntax' has been used in another sense already, and for historical (as well as synchronic, typological, and terminological reasons) syntax has, at least for the last 100 years, vastly overshadowed morphology. For a major typological reason, there are languages like Classical Chinese that admittedly lack morphology (in the sense of lacking morpheme compounding, not in the sense of lacking morphemes, which would be impossible for any language). On the other hand, no language could lack syntax because no language lacks words, the main building blocks of syntax (note that word compounding is never an issue, and probably lacking in a vast number of languages). Thus the reason for this superiority of syntax is an important definitional asymmetry: syntax is mainly about word ordering, while morphology is about morpheme compounding and ordering. This last 'and' makes morphology's domain much more restrictive, essentially equating it with the intersection of morpheme compounding and morpheme ordering. The domain of syntax is more elementary, although in the analysis of levels (of linguistic representation, semantics, etc.) morphemes are more elementary than words.

The vast majority of syntactic rules are language-specific (see Section Syntax–Semantics Interface: Preliminaries). However, some cross-linguistically universal synchronic (e.g., (1), (7), and head-dependent relation – see Section Constituency) and diachronic principles remain. In general, the syntax of any language at a particular moment is partly coincidental and partly determined by the phase of grammaticalization cycle that the language is currently in (cf Grammaticalization). By coincidental I mean the result of a complex interplay of diverse historical factors such as migration, occupation, technology, war, language contact, language prestige, true accident, etc. that we will never disentangle (cf Language Contact).

Semantics

When one considers levels of linguistic representation – phoneme, morpheme, word, phrase, sentence, and discourse – semantics pertains uniformly to all of them except phoneme. Put differently, all these five levels have meaning, thus semantics by (1). However, the meaning constructed (or extracted) on one level may in some respect differ from that on the others. Semantics is uniform in the sense that it has only one general rule:

5. Form-meaning correspondence.

Note that this view of semantics does not address the definition of meaning, which must be specified separately (see Section Predicates, Arguments, and Meaning). The universality of rule (5) (essentially, a definition of sign) stretches far beyond language, as (5) describes all natural and artificial communication systems. Semantics is also uniform in the sense that it is cross-linguistically universal (as testified by the possibility of translation – cf Haspelmath, 2007). Succinctly speaking, semantics is a function that, given a form as input, returns its meaning(s) – assuming there are any. If there is more than one meaningful input, the relation will attempt to compose the meanings of its arguments. For example, as the sentence is a combination of meaningful units, its meaning is generally compositional (as usual in language, there are few exceptions, like *it rains*, where *it* has clearly no meaning whatever – although *it* as a pronoun has a meaning). On the other hand, the meaning of a morpheme is usually elementary (of which there is a plethora of examples, e.g., the noun stem *cat* or the noun class (or gender) markings for feminine and masculine). Several other morpheme types (e.g., most cases, adpositions, tense-aspect-mood) mark relations between more than one input, and have thus compositional meanings involving predicates with two to four arguments. Four arguments seem to be (close to) the attested grammatical maximum, probably due to the constraints on short-term memory (Cowan, 2001).

Predicates, Arguments, and Meaning

Mathematically, predicate is a function from a set A to a special two-element set. In the context of language (and mathematical logic), the two elements are usually taken to be true and false (cf Logic and Linguistics). Assuming that the values are true and false, and since function is a subcase of relation, predicates describe the truth or falsity of relations. In the

simplest case, the relation could be, e.g., belonging to a subset B of A. If this is the case, the predicate P that represents the relation is true of an argument x (usually written P(x) or Px) iff x is an element of B and false (¬Px) otherwise. In this case we say that P is the characteristic (or indicator) function of B. Thus, predicates are a convenient way of addressing truth, falsity, sets, relations, functions, and (other) predicates. Given all this, it might occur that they are also a convenient way of addressing meaning. In the traditional model-theoretic sense, the meaning of a sentence is its truth-conditions (Davidson, 1967). In a more refined (but still entirely model-theoretic) sense, the meaning of a sentence is the proposition expressed by the sentence as the set of circumstances in which it is true (Soames, 1992: p. 29). Admittedly, I have a difficult time distinguishing these definitions, but they are purportedly different (Soames, 1992), so it makes sense to mention both. In either case, meaning boils down to a set of circumstances. Now circumstances are relations, and relations can be expressed as predicates – which is all the more appropriate since the circumstances are required to be true (and it is difficult to get around with truth without predicates). The frequent argument that only (declarative) sentences can have meanings is immaterial, because one can take, e.g., *cat* as a one-place predicate (which is true if used in reference to the right type of creature). In a suitable context, even a single word can produce a reference. Also, one should not be too daunted by the standard counterargument to this type of theory of meaning, concerning the truth-conditions of sentences that are always true (such as *it rains or it does not rain or both* and *1 = 1*). It is namely claimed that all such sentences have exactly the same truth-conditions, whereas in fact there are ways to filter out the relevant circumstances (but we will not delve further into this issue here). In sum, we seem to have a mathematically precise way of addressing meaning as a (set of) predicate(s). Note that this does not tell anything about how or whether these predicates are implemented in the brain, so there is no conflict with representation or concept-based theories of meaning (of which there are many – e.g., Gärdenfors, 1998; Jackendoff, 2002; Langacker, 1990; Luuk, 2013a). After all, predicates are somehow implemented as concepts in our consciousness. If so, why could not they be implemented on some lower conceptual level? (See Hurford (2003a,b) for a claim that predicates and arguments can be distinguished on a perceptual level already.)

Syntax–Semantics Interface: Preliminaries

Both semantics and syntax are rule systems but the rules involved are very different. As we saw above (see Section Semantics), there is only one general rule in semantics, holding far beyond natural language. Differently from animal communication systems, there is usually no one-to-one correspondence between linguistic form and meaning (Ferrer-i-Cancho et al., 2005). This (possibly) unique property of natural language facilitates expressivity, creativity, ambiguity, and making mistakes. Syntax, on the other hand, is a relatively complex

6. combinatorial system of dependency, agreement, and linear ordering rules,

the sole purpose of which appears to be mapping meanings to a serial channel (by (1)). Differently from (5), (6) and even the weaker (7) are uniquely human. In addition, the vast majority of syntactic rules are language-specific, and only a tiny fraction of them is known, because there are (or recently were) 6000–7000 languages in the world. The staggering typological diversity of syntactic rules is a testament to the degree of freedom the meaning mappings have. Indeed, it seems as if anything is allowed as far as the rules strike a (sub)optimal balance between minimum effort and maximum precision. This is a complex task, with trade-offs in at least three dimensions: both for the speaker and hearer individually and between them (producing vs processing effort). The sole criteria for such a balance seem to be communicative and propagative success (by the latter is meant the propagation rate of a syntactic structure in the community).

In a sense, "mapping meaning to a serial channel" is an understatement – if a mere mapping would be involved in syntax, a sequence (a concatenation) of signs would do. Granted, perhaps not any sequence but, even if we constrain the sequences, the core of full syntax does more than just concatenate signs – it also compounds and embeds them. The difference between concatenation and compounding is that in the latter the signs are necessarily interpreted together. The difference between compounding and embedding is that only the latter allows for center insertion and (potentially) more than two-level nesting. Note that embedding has a broader definition (a meaningful unit in another meaningful unit) here than in the traditional syntactic literature. For example, a morpheme in a word is an instance of embedding in this broader sense.

As mentioned above, differently from the principle of semantics (5), the principle of syntax:

7. a combination of meaningful elements by (morpho-) syntactic rules,

is uniquely human (Hurford, 2004; Jackendoff, 1999; Ujhelyi, 1998). This claim is not particularly surprising, given that (morpho)syntactic rule as such is uniquely human (if we adhere to definitions (1) and (2) – if phonological sequences are included in syntax, birds may have it). Thus syntax must have emerged relatively recently by the evolutionary scale. In fact, we have also some evidence to base our conjectures on how it evolved. For example, it can be shown that syntax evolved from concatenation to compounding to embedding, whilst retaining all the preceding procedures (Luuk and Luuk, 2014). As virtually all modern analyses concur, protolanguage started with few isolated signs (cf Bickerton, 2007; Jackendoff, 1999; Johansson, 2006). Only after that could certain rules be applied to them. Despite possible claims to the contrary (Bickerton, 1998; Chomsky, 2010), it is implausible that a complex rule system like syntax emerged suddenly full-fledged as a result of some mutation or otherwise. A gradualist scenario is much more likely (Pinker and Bloom, 1990). To be precise, the claims (at least that of Chomsky) are about the abrupt emergence of language faculty, the core of which would be a capacity for syntax. This is something completely different from an abrupt emergence of full syntax

in a linguistic community, which, as said, is implausible (and probably also impossible).

Syntactic and Semantic Validity

It is frequently claimed that syntax and semantics are two entirely different systems, domains or dimensions, which have little (if anything) in common except for the need for a common interface. This may be so – since (and as long as) the notions are not strictly defined, not much seems to hinge on the truth or falsity of this claim anyway. Another frequently advanced claim is that there is a double dissociation between syntactic and semantic validity. If this were the case, we would have rather substantive evidence for syntax and semantics being autonomous systems of linguistic representation. As the claim about dissociation seems quite well-articulated, we might be even able to test it. So let us examine this claim in more detail.

We mentioned above the principle of syntax: a combination of meaningful elements by (morpho)syntactic rules. It has been allegedly shown that this function can be separated from semantics. The commonest example is (Chomsky, 1964[1957]):

8. *Colorless green ideas sleep furiously.*

This is taken to be an example of a sentence that is syntactically valid but meaningless (that is, presumably, semantically void). However, the truth-conditions for this sentence (the circumstances under which it could be true) are

9. If green could be colorless and ideas green, and if sleep could be furious and ideas sleep (then (8) could be true).

If we accept the standard model-theoretic stance that truth-conditions are (or yield) the meaning, the meaning of (8) is somehow given by (9). The circumstance that the relations in (8) happen to be false in our world is irrelevant. Whether or not a truth-condition is satisfied in a (or in any) model is immaterial for its existence. What is required is a structure where it can be interpreted. Observe that the actual world is such a structure since it has all the necessary predicates (coinciding with the words in (8)). This structure allows for the interpretation of the meaning (i.e., truth-conditions) of the sentence (8) that is always false in the structure. After all, if it did not, how would one know that the sentence is false in the actual world?

In sum, and in light of what we said about the asymmetric dependency between syntax and semantics (3), one is led to the view that all syntactically valid expressions are also semantically valid but not vice versa. As a further test, consider the strings

10. *askldfjkldjkl sdfaklj fkkdfj*
11. *the wug drunted geperfully*
12. *you runs*
13. *this is a house*

(10–13) exhibit a gradation of general semantic/syntactic validity (well-formedness, if you will) from zero to maximum. In each of them, a robust correlation between syntactic and semantic validity is evident. In (10), both are (near) zero. Pseudo-expressions like (11) are sometimes erroneously viewed as examples of syntax–semantics dissociation. In fact, it is easy to see that everything syntactically valid about (11) is also meaningful, i.e., valid semantically. Let us try to analyze this string, starting from syntax. First, there is the noun phrase (NP) *The wug*, signaled by the article as well as 'word' order. Second, the 'verb' *drunted* – technically not an English verb but readily interpreted as one owing to the past tense suffix *-ed* and 'word' worder. Finally, the 'adverb' *geperfully* modifying (as adverbs do) the 'verb.' We do not know how exactly it does that but a modification (as such) is evident. Like the 'verb,' the 'adverb' is signaled by a suffix (*-fully*) and 'word' order. Now turn to semantics. It is not difficult to see that all we said about syntax reflects in the string's meaning. In fact, everything in (11) except for the 'stems' *wug*, *drunt*, and *geper* is meaningful. We have a definite (*the*) agent that was engaged in an activity (in a loose sense) in a specific way (although we do not know, which) in the past. Admittedly, this is a lot of meaning (as compared to, e.g., (10)), and the fact that it was extracted mostly from syntactic rules and categories does not make it less valid. It merely shows that not only stems but all morphemes and even specific word orders (such as *the x*) have meaning, and that semantics probably cannot be dissociated from syntax (while the converse is possible by (3)). (12) is defective morphosyntactically (the agreement failure) and semantically (2nd and 3rd person clash), but its syntactic and semantic interpretations are constrained much better than those of (10–11). In fact, only three readings suggest themselves – *you run*, *your runs*, and *(s) he runs* – so as compared to the previous examples the expression seems more ambiguous than invalid. (13) is valid syntactically and semantically. In sum, there is a robust correlation between syntactic and semantic validity in all examples (cf Kako and Wagner, 2001), corroborating the claim or suspicion (whichever you prefer at this point) that syntax is not an autonomous system of linguistic representation.

Constituency

Constituency refers to the whole–part relation in language expressions. Although one may speak of phonological and morphological constituents, the term 'constituent' has a predominately syntactic usage, being mainly reserved for words, phrases, and sentences. In the context of syntax, constituent structure thus refers to words, phrases, or sentences in a hierarchical structure. There are numerous ways to model constituent structure of natural language. This (besides modeling grammatical relations – cf Section Grammatical Relations) is a bread-and-butter work of theoretical linguists, and most grammatical theories also stipulate their specific constituent structures. Likewise, there is no shortage of methods for describing them (e.g., trees, rewrite rules, bracketing, predicate logic, frame semantics, lambda calculus, categorial grammar, model theory, etc. (cf Logic and Linguistics)). Far from being equivalent or complementary, the methods frequently lead to very different views on syntax, semantics, and their interface.

The single most important (i.e., widespread and universal) morphosyntactic rule is head-dependent relation. The essence of a head-dependent relation is that in combining two linguistic units (e.g., in a word or phrase), one of them 'projects' or starts governing the other (although it may be not easy to determine

which one). The one projecting is called the head and the other is dependent. Since every head-dependent relation specifies two whole–part relations and three constituents, head-dependent relation yields constituency for free. Since agreement marks syntactic constituents, head-dependent relation provides also the foundation for agreement (cf Grammatical Agreement). In natural language, head-dependent relation and constituency go far beyond syntax, possibly even to phonology. An interesting question for syntax–semantics interface is what determines which elements become heads at phrase and sentence levels. At least in some cases semantics has probably some role in it. For example, the finite verb or flexible (see Luuk, 2010) is the head of the clause and the clause predicate, suggesting that the head may be determined by the semantic predicate-argument structure. Another example is agreement. Lehmann (1982) has argued that only constituents that relate to the same referent may overtly agree with each other.

Grammatical Relations

Grammatical relations (or syntactic functions, as they are sometimes also called (*see* Grammatical Relations)) hold between clause constituents. Standard constituents in this sense are verb, object, and subject but there are others like modifier, specifier, determiner, etc. (there are many others, most of them specific to particular grammatical theories, of which there are more than 100). Every grammatical relation is stipulated by at least one head-dependent relation (see Section Constituency). A crucial question is how much of grammatical relations are derivable from semantics. As it turns out, a lot. The Chomskian term for this phenomenon is theta criterion: each argument of the (finite) verb is assigned exactly one theta role (semantic role, e.g., agent, patient, goal, etc.) and vice versa. Because of this one-to-one correspondency, it seems that the two – semantic role and syntactic argument – are merely two facets of the same thing, viz. meaning and ('syntactic') form, respectively. But if so, then grammatical relations are fundamentally semantic phenomena, because form-meaning is the general principle of semantics. Furthermore, the supposedly syntactic relations like V, S, O, etc. can be also described with semantic roles in frame semantics or predicates and arguments in a suitable predicate logic. A closely related notion used in generative grammar and elsewhere is subcategorization frame, identifying finite verb's potential positions in phrase structure trees by its class (intransitive, transitive, ditransitive, ergative, etc. (cf Valency and Argument Structure in Syntax)). In generative grammar, subcategorization frame (detached from the verb that appears in it) was believed to be meaningless. However, Kako (1997) experimentally showed this not to be the case (with stimuli resembling (11)). Since subcategorization frames have meanings, it seems that a core part of syntax is semantically driven.

Word Order

All possible (S, V, O) word orders have been attested in the world languages, although some are very rare (the *World Atlas of Language Structures*, http://wals.info/chapter/81; (*see* Word Order)). In languages with (relatively) free word order, semantics and especially pragmatics (e.g., information structure) are responsible for at least some ordering principles. In languages with fixed word order, semantics has probably a minor role in it. The reason for this is that natural language semantics is order-neutral – nowhere except in formal notations is there a rule to the effect that a predicate should follow its argument(s) or vice versa. If a verb has three arguments (like *give*), it probably makes sense (for processing considerations at least) not to put the indirect object first but this is a pragmatic rather than semantic constraint. Granted, certain fixed orders are meaningful as they specify constituents with specific meanings, be they referents (such as *the x* mentioned above) or relations (e.g., subcategorization frames and prepositional phrases – cf Luuk, 2013b). However, there is no semantic reason for choosing this particular order – we could have determiners following nouns and postpositions instead of prepositions. So, of all the syntactic phenomena we have discussed, at least fixed word order seems largely independent from semantics (and other levels of linguistic representation), thus qualifying syntax as an autonomous level after all (cf Section Syntactic and Semantic Validity). Although we have so far concentrated on cases of semantics determining syntax (partly because they are more difficult to see), this qualification allows us, at least in languages with fixed word order, to consider also the converse cases. There are many such examples, e.g.,

14. *j loves m*
15. *m loves j*

Conclusion

As was noted in the beginning (cf Section Introduction: Syntax/Semantics), an objective account of the syntax–semantics interface of natural language seems unattainable. Although I have tried to shun all major schools and to keep my assumptions to a minimum, I do not know how successful or convincing my case has been in this short introduction to what must be one of the most obscure, complex, and technical areas of natural language. This obscurity and complexity is due to several reasons, which for this particular field unhappily coincide: the great number of opposing linguistic schools; the still greater number of specific grammatical theories; the direct relevance of grammar (and hence of grammatical theory) for a description of syntax–semantics interface; the field's uncomfortable position between two established (yet arbitrarily defined) linguistic levels, hence dependence from both; the relatively central position with regard to all levels (and thus a near-maximum dependency from all of them (and hence also of their descriptions, of which there are many)); and lastly (but most importantly) the scarcity of clear, universal definitions and the (maybe even necessary) lack of a theory-neutral research program. One can only hope that perhaps in a not-so-distant future someone will succeed in advancing this area beyond mere wishful thinking and haphazard theorizing. On a more optimistic note, the great number of linguistic schools and grammatical theories in itself is rather positive – the more theories we have, the greater (albeit likely still very low) the probability of least one of them being (sufficiently) correct.

See also: Grammatical Agreement; Grammaticalization; Language Contact; Logic and Linguistics; Morphology in Linguistics; Valency and Argument Structure in Syntax; Word Order.

Bibliography

Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J., 2013. Evolution, brain, and the nature of language. Trends in Cognitive Sciences 17 (2), 89–98.

Bickerton, D., 1998. Catastrophic evolution: the case for a single step from protolanguage to full human language. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language: Social and Cognitive Bases. Cambridge University Press, Cambridge, pp. 341–358.

Bickerton, D., 2007. Language evolution: a brief guide for linguists. Lingua 117 (3), 510–526.

Chomsky, N., 1964[1957]. Syntactic Structures. Mouton, The Hague.

Chomsky, N., 2010. Some simple evo-devo theses: how true might they be for language? In: Larson, R.K., Yamakido, H., Deprez, V. (Eds.), Evolution of Human Language: Biolinguistic Perspectives. Cambridge University Press, Cambridge, pp. 45–61.

Cowan, N., 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behavioral and Brain Sciences 24 (1), 87–114 discussion 114–185.

Davidson, D., 1967. Truth and meaning. Synthese 17, 304–323.

Ferrer-i-Cancho, R., Riordan, O., Bollobas, B., 2005. The consequences of Zipf's law for syntax and symbolic reference. Proceedings of the Royal Society of London B Biological Sciences 272 (1562), 561–565.

Fillmore, C.J., 1976. Frame semantics and the nature of language. In: Harnad, S., Steklis, H.D., Lancaster, J. (Eds.), Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, pp. 20–32.

Gärdenfors, P., 1998. Some tenets of cognitive semantics. In: Allwood, J., Gärdenfors, P. (Eds.), Cognitive Semantics: Meaning and Cognition. John Benjamins, Amsterdam, Philadelphia, pp. 19–36.

Haspelmath, M., 2007. Pre-established categories don't exist: consequences for language description and typology. Linguistic Typology 11, 119–132.

Hengeveld, K., Mackenzie, J.L., 2008. Functional Discourse Grammar: A Typologically-Based Theory of Language Structure. Oxford University Press, Oxford.

Hurford, J.R., 2003a. The neural basis of predicate-argument structure. Behavioral and Brain Sciences 26 (3), 261–283 discussion 283–316.

Hurford, J.R., 2003b. Ventral/dorsal, predicate/argument: the transformation from perception to meaning. Behavioral and Brain Sciences 26 (3), 301–311.

Hurford, J.R., 2004. Human uniqueness, learned symbols and recursive thought. European Review 12 (4), 551–565.

Jackendoff, R., 1999. Possible stages in the evolution of the language capacity. Trends in Cognitive Sciences 3 (7), 272–279.

Jackendoff, R., 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press, Oxford, New York.

Johansson, S., 2006. Working backwards from modern language to proto-grammar. In: Cangelosi, A., Smith, A.D.M., Smith, K. (Eds.), The Evolution of Language: Proceedings of the 6th International Conference on the Evolution of Language (EVOLANG6). World Scientific, Singapore, pp. 160–167.

Kako, E., 1997. Subcategorization semantics and the naturalness of verb-frame pairings. University of Pennsylvania Working Papers in Linguistics 4 (2), 155–167.

Kako, E., Wagner, L., 2001. The semantics of syntactic structures. Trends in Cognitive Sciences 5 (3), 102–108.

Langacker, R.W., 1990. Concept, Image, and Symbol: The Cognitive Basis of Grammar: Cognitive Linguistics Research. Mouton de Gruyter, Berlin, New York.

Lehmann, C., 1982. Universal and typological aspects of agreement. In: Seiler, H., Stachowiak, F.J. (Eds.), Apprehension: Das sprachliche Erfassen von Gegenständen. Narr, Tübingen, pp. 201–267.

Luuk, E., 2010. Nouns, verbs and flexibles: implications for typologies of word classes. Language Sciences 32 (3), 349–365.

Luuk, E., 2013a. The structure and evolution of symbol. New Ideas in Psychology 31, 87–97.

Luuk, E., 2013b. New methods for analyzing case and adposition meaning. Language Sciences 37, 136–146.

Luuk, E., Luuk, H., 2014. The evolution of syntax: signs, concatenation and embedding. Cognitive Systems Research 37, 136–146.

Montague, R., 1970. Universal grammar. Theoria 36, 373–398.

Pinker, S., Bloom, P., 1990. Natural language and natural selection. Behavioral and Brain Sciences 13 (4), 707–784.

Seuren, P.A.M., 1996. Semantic Syntax. Blackwell, Oxford.

Soames, S., 1992. Truth, meaning, and understanding. Philosophical Studies 65, 17–35.

Ujhelyi, M., 1998. Long call structure in apes as a possible precursor for language. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language: Social and Cognitive Bases. Cambridge University Press, Cambridge, UK.

Relevant Website

http://wals.info/.

Systematic Reviewing and Meta-Analysis

Jeffrey C Valentine, University of Louisville, Louisville, KY, USA
Therese D Pigott, Loyola University Chicago, Chicago, IL, USA
Timothy Lau, University of Louisville, Louisville, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The methodologies of systematic review and meta-analysis were developed to address the challenges of providing a transparent and replicable review of the research literature relevant to a particular question. In this article we provide a brief overview of systematic reviewing and meta-analysis and then discuss how to develop a good research question for a review, how to search for relevant studies, and how to extract information from the studies (including information that will help judge study quality). The second half of this article concerns meta-analysis specifically (i.e., the statistical synthesis of the results of the studies obtained through the systematic review process), including weighting, computing a mean effect size, conducting a homogeneity analysis, and carrying out moderator analyses. The article concludes by discussing new developments in meta-analysis.

Introduction

Imagine that you are responsible for finding programs that occur outside of the school day for students who are at risk of school failure. Among other options, you think about the possibility of mentoring – that is, connecting a responsible adult volunteer (the mentor) with an at-risk student (the mentee), with the goal of fostering positive development and well-being in the student (DuBois et al., 2011). But you might wonder if mentoring is an effective model for reducing school failure. It is relatively easy to find anecdotes on the Internet that attest to the power of mentoring. For example, the Web site for Big Brothers/Big Sisters of America relays an anecdote about a child, Deva, who was the first in her family to receive an academic scholarship (http://www.bbbs.org/site/apps/nlnet/content2.aspx?c=9iILI3NGKhK6F&b=6195821&ct=9103913&notoc=1). Anecdotes are useful ways of connecting readers to the experience of the participants, but it is difficult to discern the typical effects of an intervention from such stories. As such, you might look for studies that evaluate the effects of mentoring programs. The Web site for MENTOR, the National Mentoring Partnership, stated that a study by Herrera et al. (2013) found student "gains in social acceptance, academic attitudes and grades" (http://www.mentoring.org/about_mentor/value_of_mentoring). But one study, no matter how well conducted, is limited. You might therefore wish to find a literature review that examines a broader range of studies. For example, the review by Lawner et al. (2013) examined 19 studies of mentoring programs and concluded that the results of mentoring programs on educational outcomes are 'mixed.' But even when statements about effects are drawn from a range of studies as they were in the Lawner et al. study, it is often not clear how representative the studies are and how rigorously the findings from the individual studies have been synthesized.

The methodologies of systematic review and meta-analysis were developed to address these challenges by providing a transparent, replicable, and systematic review of the research literature relevant to a particular question. In the sections below we provide a brief overview of systematic reviewing and meta-analysis and then discuss how to develop a good research question for a review, how to search for relevant studies, and how to extract information from the studies (including information that will help judge study quality). The second half of this article concerns meta-analysis specifically, which refers to the statistical synthesis of the results of the studies obtained through the systematic review process.

Steps in the Systematic Review Research Process

A systematic review looks a lot like any other type of research (Cooper, 1982): Researchers need to formulate a good research question, develop a sampling plan and collect data, analyze the data, and report the results. Of course, the way that these steps are operationalized will differ if one is carrying out a primary study or a systematic review. For example, the sampling plan in a systematic review involves identifying potentially relevant research literature. The main point is that the framework should feel very familiar, and we just need to flesh out how these generic steps are implemented in the context of a systematic review.

Formulating a Good Research Question

Coming up with a good research question is a skill unto itself. Depending on the aims of the review, a question can be narrow (e.g., "What is the effect of the Big Brothers/Big Sisters mentoring model on academic achievement for urban girls aged 13–17?") or broad ("What is the effect of mentoring on outcomes for children aged 3–17?"). Narrow and broad questions typically lead to very different types of systematic reviews. DuBois et al., working with a broad research question, identified 73 studies for inclusion in the review. A scholar interested in synthesizing studies on a narrow question like the one posed above would find only a few studies. Indeed, a research question can be so narrowly tailored that no studies have ever been done that examine it. In general, narrow research questions are best when the interest is in examining the effects of a specific intervention (e.g., the Big Brothers/Big Sisters mentoring

model). Broad questions are usually best when the goal of the research is knowledge generation. For example, the scholar might be interested in identifying common elements in effective mentoring programs. A broad research question that includes many different models will be a better way of doing this than a narrow review will. For example, there might be interest in determining whether providing mentors with training matters. If a review just focuses on the Big Brothers/Big Sisters model, this question cannot be answered because that particular model mandates mentor training (so there is no variation on this dimension).

Setting Eligibility Criteria: PICOS

One way to think about a research question is that as much as possible it should specify the Participants, Intervention, Comparison, Outcomes, and Study types of interest, and thus go a long way of describing the characteristics of studies that will lead to study inclusion, and those that will lead to exclusion from the systematic review. Think about the narrow research question posed above. It explicitly specified the participants (urban girls), the intervention (mentoring as practiced by Big Brothers/Big Sisters), and the outcomes (academic achievement) of interest. The review question also implies that experiments will be examined (by asking "What is the effect of …"), and we can therefore assume that children who receive Big Brothers/Big Sisters mentoring will be compared to children who do not. Of course, researchers still need to make some decisions – for example, they need to operationally define what they mean by 'urban,' but many of these kinds of decisions will be relatively easy to make.

As the review question broadens, however, eligibility decisions are less likely to follow explicitly, and boundaries distinguishing between studies the reviewers would like to include and studies the reviewers would like to exclude are often harder to define. Consider the broad research question we posed above, which asks (in part) "What is the effect of mentoring …?" To carry out a systematic review of studies that examine mentoring (broadly defined), the reviewers have to decide what they mean by 'mentoring.' As an example, DuBois et al. (2011) defined a mentoring as one that pairs a youth in a one-to-one relationship with a volunteer adult, with the goal of fostering a relationship that enhances the youth's well-being. As such, studies that involve paid mentors, or have an explicit goal of helping the youth find a job, or involve a single mentor with multiple youths were not included in their review.

Logic Models Can Help with Problem Formulation and Eligibility Setting

Usually based on both theoretical and empirical considerations, logic models describe the connections between different determinants of an outcome, and how these determinants are related in a specific intervention (Anderson et al., 2011). A simplified version of the logic model used in DuBois et al. (2011, p. 61) looks like this: A strong and meaningful mentoring relationship develops the youth's mutuality, trust, and empathy. The development of these traits leads to positive socioemotional, cognitive, and identity development. These positive developments then lead to outcomes such as improved grades and social behavior. To see how the articulation of a logic model can shape the review question and eligibility criteria, consider the 'strong and meaningful mentoring relationship' component from the logic model developed by DuBois et al. A 'strong and meaningful' criterion implies that the intent of the mentoring program should be to foster deep, long-term relationships. Therefore, studies of mentoring programs that establish only short-term relationships were not included in DuBois et al.'s review (in essence, the type of mentoring examined in those studies is not the same type of mentoring that DuBois et al. were interested in researching). Decisions regarding where to draw fuzzy boundaries like these are facilitated by the articulation of a clear logic model.

Searching for Relevant Literature

Once a specific review question has been formulated and the inclusion criteria have been set, the next step involves locating studies that might be eligible for review. As a general rule, individuals carrying out a systematic review should attempt to find *every* study that meets their inclusion criteria. For most research question this is a daunting task, and in fact we will almost never know if we have attained our goal. A more reasonable goal might be to minimize the differences between the studies that we are able to collect and the studies that we could not locate. We try to minimize the differences between studies we identify and those we cannot identify by using multiple complementary literature search strategies. A search carried out in multiple electronic databases (like PsycINFO) is likely to be the backbone of the literature search strategy. However, systematic reviewers are advised to use other techniques to supplement this search. For example, in most cases multiple databases should be searched. And, it is almost always important to look beyond databases for potentially relevant studies. For example, literature retrieval experts often recommend sending requests to the listservs of relevant professional organizations, contacting authors of studies that are included in the review, and conducting targeted Internet searches (Campbell Collaboration policy brief). These strategies might help uncover studies that are unavailable through the usual routes – for example, studies that are currently under consideration for publication, or studies that authors have decided not to submit for (or have given up on) publication.

There are several reasons that individuals undertaking a systematic review are advised to go to such lengths to uncover potentially relevant studies. Having more studies generally increases statistical power, and will also usually serve to increase the generalizability of the results (because these are based on a wider variety of study characteristics). Just as important as statistical power and generalizability, it has been shown that studies that are not published tend to have smaller effects than studies that do get published, a phenomenon known as publication bias. Publication bias occurs because studies lacking statistically significant findings on their main outcomes are less likely to be submitted for publication, and if submitted less likely to be published, than studies with statistically significant results on their main outcomes. Statistical significance is a function of two things: (1) the size of the sample used to estimate the effect and (2) the magnitude of the effect. As a result, small studies that just happen to observe very large effects are more likely to get published than small studies that observe smaller effects. Because unpublished studies are

harder to find, small studies with small effects will tend to be missing from the meta-analytic database, creating a bias in the estimate of the effect. Therefore, individuals undertaking a systematic review should invest significant effort into finding studies that have not been published and evaluate them for potential inclusion in the review.

Screening and Coding Studies

Because literature searches have the goal of locating as many relevant studies as possible, they are necessarily broad. One side effect of a broad search is that it will likely identify many more potentially relevant than actually relevant studies. As a result, once the literature search has been carried out, the potentially relevant studies need to be screened for relevance. For studies identified through Internet and electronic database searches, screening is usually based on information available in the titles and abstracts of the studies identified as being potentially relevant. The goal of the screening process is to weed out any studies that are *clearly* not relevant. DuBois et al.'s (2011) search for mentoring studies turned up documents that were not studies (e.g., opinion and advocacy pieces), studies that did not involve mentoring as it was defined by DuBois et al. (e.g., tutoring), and studies that did not involve an adult mentor (e.g., peer mentoring). Studies that are not screened out at this stage are obtained, and a final relevance decision is made based on information available in the full text of the document.

Once a set of studies has been identified as being relevant to the review, they need to be coded. The coding process is analogous to collecting data on a survey questionnaire (Lipsey and Wilson, 2001). The researchers develop questions that they would like to obtain from each research report. Generally, the goal of coding is to describe the studies used in the meta-analysis and to represent qualitatively important dimensions of studies in data form (such as the research design and theoretically important variables that might affect the direction and/or strength of the relations observed in studies). Researchers will also code information related to the study's findings (e.g., sample size and effect size). Lipsey (2009) discusses how synthesists can go about identifying study features that might influence study outcomes.

Screening and coding are activities that are prone to human error, and as such, steps should be taken to increase the reliability and validity of the judgments made. Best practices in this regard are as follows. First, the individuals doing the screening and coding should be trained, and these individuals should meet regularly to discuss problems encountered and their potential solutions. Part of the goal of both the training and the ongoing meetings is to help coders develop and maintain normative understandings of the questions they are trying to answer. In addition, in most cases both the screening and coding should be done by at least two individuals who work independently. While a good reliability estimate requires more coders than most projects can afford (six coders is probably a bare minimum), systematic reviewers often used two independent coders and report their percentage agreement on key variables, and have the coders reconcile their disagreements (a third coder can be appealed to if the two coders cannot resolve their differences). Finally, the questions used to screen and to code studies should require as little inference as possible on the part of the individuals doing the work. Often this is best accomplished by (1) operationally defining key terms and (2) breaking complex judgments into several smaller, simpler questions. For example, asking coders to rate 'Study Quality' is almost certain to result in an unacceptably high rate of coder disagreement. As such, it is better to break this question into several smaller questions.

Judging Study Quality

Empirical studies vary in terms of the rigor with which they are conducted, and the quality of the studies comprising a systematic review can have an impact on the validity of conclusions arising from that synthesis. Therefore systematic reviewers have to pay close attention to issues related to study quality. Unfortunately, two intuitive and relatively easy to implement strategies turn out to be unhelpful. As mentioned, asking study coders to rate 'study quality' is likely to result in coder disagreement. The reason for this is that coders can differ in terms of the dimensions that they think are related to study quality, and even when they agree on the dimensions, they can disagree about the relative importance of any particular dimension.

Another appealing, simple, and ultimately unhelpful strategy is to use a study quality scale. A review of these instruments is beyond the scope of this article, but unfortunately quality scales have little to recommend them. They tend to (1) be insufficiently operational, (2) lack an empirical basis for the quality dimensions, and (3) use overall quality scores that are composed of several unrelated dimensions. As a result, quality scores tend to lack evidence of validity, and have performed poorly in empirical tests (e.g., see Jüni et al., 1999; see Valentine and Cooper, 2008; for a more thorough review of the issues involved in creating a quality scale). The lack of an empirical basis for the quality dimensions assessed is a key limitation. One of the reasons for this limitation is that what constitutes 'quality' is in fact highly dependent on the context of the research question (Valentine, 2009). This means that systematic reviewers need to think deeply and carefully about how various study quality dimensions might come into play for their review. As an example, attrition (i.e., the loss of participants from a study) is generally viewed as being problematic, and many study quality scales assess it. However, Valentine and McHugh (2007) demonstrated that in studies of educational processes conducted in classrooms with young children, typical levels of attrition might not be problematic at all. They reasoned that in many cases, students might not actually realize that they were in a study, and that attrition might therefore be related to process that are functionally random (e.g., children who happen to be ill on the day the posttest is given). Although this seems fairly obvious in retrospect, most quality scales are not constructed in a way that recognizes different study contexts might change the importance of study quality indicators. It is therefore critical that scholars undertaking a systematic review think carefully about the important study quality indicators for their research question and to code these for every study.

Summary of Systematic Reviewing

In conducting a systematic review, researchers move away from anecdote, the results of single studies, and the results of

incomplete collections of studies and instead approach a well-defined research question by attempting to assemble a collection of studies that strongly resembles the totality of the relevant research literature. Extensive steps are taken to obtain as much of the evidence as possible. Studies are screened by at least two trained individuals working independently. Information relating to the study's background, context, quality, and effects is then coded again by at least two individuals working independently.

Systematic reviewers next need to decide whether and how they will synthesize the results of the studies they have uncovered. The decision about whether or not to synthesize studies is an important one. While the minimum number of studies needed for a synthesis is the classical statistical tradition of two (Valentine et al., 2010), studies should be highly similar to one another for this exercise to be meaningful in such an information poor environment (Cooper, 2003). However, it is also important to consider the quality of the alternatives and the needs of the audience. Because we believe that most alternatives are poor (e.g., see Valentine et al., 2010) and because readers will often conduct the synthesis on their own if a synthesis is not provided (e.g., see Valentine et al., 2011), we believe that meta-analysis (the topic of the remainder of this article) is often the best option.

Meta-Analysis

The term meta-analysis refers to the statistical analysis portion of a systematic review. As we have seen, systematic reviews collect and synthesize the results of a representative sample of studies on a research question, and utilize transparent and systematic methods to conduct the review. In a systematic review, studies are the unit of analysis; the review focuses on making generalizations about a research question from a representative sample of studies on a given topic. The statistical analysis of those studies uses the techniques of meta-analysis. These techniques are analogous to the statistical analysis of any data such as those used to analyze measures collected from individual students or clients. We will introduce the techniques of meta-analysis using the typical stages of data analysis highlighting the differences necessary when the unit of analysis is the study itself.

Collecting Data from Each Study: The Effect Size

As in a primary study of individual students or clients, we systematically gather measures or assessments of outcomes, such as achievement or depression, which are the focus of the study. In a meta-analysis, we are interested in a measure of the study's outcome. This outcome relates back to the research question of the review, and could focus on the effectiveness of an intervention using experimental methods, or the relationship between two constructs such as achievement and depression. The data we gather from each study is called the effect size. In a review of mentoring programs, we might be interested in whether mentoring improves depressive symptoms. Unfortunately studies in the behavioral and social sciences often use different measures for the same construct, such as the Beck's inventory for depression or the Hamilton Depression Rating scale. Since not all studies use measures on the same scale, we need to gather an effect size from each study that is comparable across studies. The form of the effect size depends on the research question and study design. We will focus on three effect sizes that are most common in the behavioral and social science literature: the standardized mean difference, the odds ratio, and the correlation coefficient.

Measures of the Effectiveness of an Intervention

Systematic reviewers are often interested in understanding the effectiveness of an intervention such as a treatment for depression. Studies examining the effectiveness of an intervention employ an experimental design, in the simplest case comparing a treatment group with a control group. The natural measure of a study's effect for experimental designs is the difference between the treatment and control groups. When the outcome is a continuous measure such as a score on an achievement test or a depression assessment, the effect size used is the standardized mean difference. When the outcome is a dichotomous measure such as the number of high-school graduates in the treatment and control group, the effect size used is the odds ratio. Both measures compare the outcomes between a treatment and a control group.

The standardized mean difference is given by the following formula:

$$d = \frac{\overline{X}_T - \overline{X}_C}{s_p} \qquad [1]$$

where $\overline{X}_T$ and $\overline{X}_C$ are the means for the treatment group and the control group, respectively, on the study outcome of interest, such as an achievement score or depression assessment. The denominator of eqn [1], denoted by s_p is the pooled, sample standard deviation for the outcome measure for the treatment and control groups and is given by

$$s_p = \sqrt{\frac{(n_T - 1)s_T^2 - (n_C - 1)s_C^2}{n_T + n_C - 2}} \qquad [2]$$

where n_T and n_C are the sample sizes for the treatment and control group, and s_T^2 and s_C^2 are the sample variances for the outcomes for the treatment and control group.

Each effect size measure also has an associated standard error. Just as we compute the standard error of a treatment group mean, we also need to compute the standard error of our estimate of the effect size. The standard error of the standardized mean difference within a single study is

$$\mathrm{se}(d) = \sqrt{\frac{n_T + n_C}{n_T n_C} + \frac{(d)^2}{2(n_T + n_C)}} \qquad [3]$$

Note that the standard error of the standardized mean difference depends on the estimate of the effect size and (mostly) the sample size, so that studies with smaller sample sizes have larger standard errors. Because standard errors reflect the precision of the estimate (the smaller the standard error, the greater the precision of the estimate), we prefer estimates with smaller standard errors. In fact, in meta-analysis we weight studies differentially by the inverse of their squared standard errors, as will be demonstrated below.

When the experimental study examining the outcome of a treatment is a count of the number of patients who experience

an event such as graduation, the effect size of interest is the odds ratio. The odds ratio may be less familiar to behavioral and social scientists. To illustrate the odds ratio, let us assume that we have an experiment that collected the following data about the outcomes of a mentoring intervention, where the letters a, b, c, and d indicate the number of mentees in each category:

Experimental group	*Graduated (success)*	*Did not graduate (not successful)*
Treatment	*a*	*b*
Control	*c*	*d*

As can be seen in eqn [4], the odds ratio is literally the odds of success in the treatment group relative to the odds of success in the control group:

$$OR = \frac{a/b}{c/d} = \frac{a \times d}{b \times c} \quad [4]$$

We interpret the odds ratio as the ratio of the odds of treatment group members graduating and the odds of control group members graduating. A typical interpretation of the odds ratio uses the terms success and failure for the two conditions, so that we can say that the odds ratio is the ratio of the odds of success in the treatment group to the odds of success in the control group. Note that an odds ratio equal to 1 indicates that the treatment group members and the control group members have an equal odds of graduating (success). An odds ratio smaller than 1 indicates that treatment group members have a lower odds of not relapsing, and an odds ratio greater than 1 indicates that the treatment group has a higher odds of success than the control group.

One nonoptimal property of the odds ratio is that it ranges from 0 to positive infinity. This scale is not evenly spaced and thus the odds ratio does not have a distribution that allows commonly used statistical methods. Thus, we use the natural logarithm of the odds ratio when doing meta-analysis as given by $LOR = \ln(OR)$. The log odds ratio has a scale centered at 0, indicating equal odds of success for the treatment and control group, and ranges from $-\infty$ to $+\infty$. The standard error of the log odds ratio is given by

$$se(LOR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad [5]$$

Note that like the standardized mean difference, the standard error of the log odds ratio also depends on the sample size, with smaller sample sizes leading to larger standard errors, i.e., less precision in the estimation.

Measuring the Relationship between Two Constructs

We are often interested in understanding the strength and direction of an association between two constructs in a systematic review. For example, we may be interested in understanding the strength of the relationship between depression and self-concept in adolescents, or the association between mentor and mentee reports of relationship quality. In this situation, the natural estimate of the study's effect is the correlation coefficient, usually denoted by r. One problem with the correlation coefficient is similar to the problem with the odds ratio, the scale of measurement of r. As we know, the correlation coefficient ranges from -1.0 to 1.0 and thus has properties that do not allow the application of commonly used statistical methods. In meta-analysis, we use a transformation of the correlation coefficient, Fisher's z-transformation, given by

$$z = 0.51n\left(\frac{1+r}{1-r}\right) \quad [6]$$

where we are again taking the natural logarithm of a quantity as we did in the log odds ratio. The standard error of Fisher's z-transformation is

$$se(z) = \sqrt{\frac{1}{n-3}} \quad [7]$$

where n is the sample size in the study. As in all of the effect sizes discussed here, the standard error of Fisher's z-transformation also depends on the sample size, with smaller studies having larger standard errors.

Other Measures of Effect Size

Several other types of effect sizes may be relevant to reviewers. For example, if all the studies use the same measure, a reviewer might want to use the unstandardized mean as the effect size. Another area of interest for reviewers of observational studies is the synthesis of the results from regression studies. Researchers in school psychology and special education are also interested in the synthesis of studies that use single case designs. Methodological research is continuing in a number of these areas (references). In addition, reviewers might find that primary studies do not always provide the summary statistics needed to compute an effect size. There are a number of web-based effect size calculators that can assist reviewers in computing the most common effect sizes from a variety of different summary statistics.

Analyzing Effect Sizes across a Set of Studies

The first step in any statistical analysis centers on understanding the distribution of the observed measures by computing the mean and variance of a variable, and examining a graph of its distribution. In a meta-analysis, we are interested in examining the mean and spread of the effect sizes across studies. We are also interested in understanding how much effect sizes vary across studies, and if that variation is related to differences in the studies' methods, samples, or treatments. Thus, the stages of a meta-analysis conform to that of a typical statistical analysis for primary studies.

In a meta-analysis, however, we need to use methods adapted to the nature of effect sizes. We illustrated above that the standard errors of the study-level effect size depends on the sample size. To compute a mean effect size across studies, we need to take the differences across studies in their sample size, and more specifically in their standard errors of their estimates of effect size into account. Thus, we need to compute a weighted mean effect size across all studies. Imagine that we need to obtain the mean number of times that mentors and mentees met in a month completed across three different agencies. However, different numbers of mentors volunteer at

each agency. We would not want to take the mean number of meetings from each agency, add them up, and divide by three, as this 'average' would not take into account the fact that we might have 50% of the mentors volunteering at agency 1, 35% volunteering at agency 2, and the remaining 15% volunteering at agency 3. Instead, we would weight each agency mean by the proportion of mentors who volunteer at that agency.

In a similar way, we compute the weighted mean effect size across all studies, using the inverse of the standard error of the effect size as the weight. If we denote $w_i = 1/se(ES_i)$ for study i where ES_i is our effect size, we would compute our overall weighted mean effect size across $i = 1, \ldots, k$ studies as

$$\overline{ES} = \frac{\sum_{i=1}^{k} ES_i w_i}{\sum_{i=1}^{k} w_i} \quad [8]$$

The standard error of the mean effect size is given by

$$SE_{\overline{ES}} = \frac{1}{\sum_{i=1}^{k} w_i} \quad [9]$$

Given the standard error and the effect size, we can examine the 95% confidence interval around the overall effect size as

$$95\%CI : \left[\overline{ES} - 1.96SE_{\overline{ES}}, \overline{ES} + 1.96SE_{\overline{ES}}\right] \quad [10]$$

If the 95% confidence interval does not include 0, then the overall mean effect size is statistically significant.

Examining Heterogeneity of Effect Sizes

As is the case in any data analysis, the mean effect size only gives a reviewer a sense of the center of the distribution of effect sizes. We also need to understand how much the effect sizes vary across studies. If we are interested in the effectiveness of a treatment, we want to know not only if the average treatment effect is positive, we also want to know if that treatment effect varies a little or a lot. It could be that the average treatment effect may be positive, but there also may be several studies that find no effect of the treatment or a harmful effect. We examine the variation among effect sizes by computing measures of homogeneity. One test of homogeneity, designated by Q, tests whether the variation across effect sizes in a set of studies is greater than we would expect due to sampling error, i.e., due to the fact that each study uses an independent and different sample of participants. The formal definition of the Q test of homogeneity is given by

$$Q = \sum_{i=1}^{k} w_i \left(ES_i - \overline{ES}\right)^2 \quad [11]$$

Note that the computation of Q includes the difference between the effect size from each study i, and the mean effect size, weighted by the inverse of the standard error of the effect sizes. Thus, large values of Q indicate more variation of study effect sizes from the mean effect size. We compare the obtained value of Q to a chi-square distribution with $k - 1$ (k equal to the number of studies) degrees of freedom. The null hypothesis of Q is that the effect sizes are homogeneous; large values of Q lead to the rejection of the null and a conclusion that the effect sizes are not estimating a common mean effect size.

A second measure of the homogeneity of effect sizes is the index I^2. The index I^2 gives us the proportion or percentage of variation in the effect sizes that is not associated with sampling error, or differences in the samples of individuals in the study. We compute I^2 as

$$I^2 = \left(\frac{Q - k - 1}{Q}\right) \times 100\% \quad [12]$$

The index I^2 gives a reviewer a sense of how much of the variation across studies needs to be explored. For example, a small value of I^2 such as 20% indicates that only 20% of the variation in effect sizes is not due to sampling error. In contrast, a value of I^2 such as 75% indicates that much of the variation in effect size is not related to differences between the samples of participants and must be due to other differences among studies.

In a situation with a large I^2 and a statistically significant Q value, a reviewer should be hesitant to use the overall mean effect size as an indicator of the 'average' treatment effect or the 'average' strength of the relationship between two constructs. In the presence of heterogeneity among effect sizes, the reviewer should explore the possible reasons for that heterogeneity through the use of moderator analyses, a topic that will be covered below.

Moderator Analyses in Meta-Analysis

When a reviewer finds a large value of I^2 or a significant value of Q, these results are a signal that the studies may not be estimating a common effect size, and/or that the studies differ for reasons beyond the fact that they use different samples of participants. Reviewers can explore the heterogeneity across studies using models that are analogous to Analysis of Variance (ANOVA) and regression. Often a reviewer is most interested in potential variation among studies in their estimates of a treatment effect or the strength of a relationship between two constructs. For example, a mentoring intervention may have differential effects for boys and girls, or the effectiveness of mentoring may depend on how often the mentor meets with the mentee. Any variations among studies in the genders of participants or in the frequency of meetings could be associated with differences in estimates of the study's effect size. As we have seen, during coding reviewers will have gathered not only information about effect sizes from studies during the data collection phase of a systematic review, but will also have coded information about the studies' methods, participants, measures, and other characteristics. All of these study characteristics can be used as predictors in models that explore the potential moderators of effect size.

We use a statistical analysis comparable to ANOVA when we have a moderator variable that has a small number of categories. For example, we may want to compare the effect sizes in studies in which the typical relationship quality was rated as high to studies in which typical relationship quality was rated as average. The analysis context is similar to a one-way ANOVA where we want to compare the mean of one group (in this case, studies with typical relationship quality rated as high) versus another group (studies with typical relationship quality rated as average). As in ANOVA, the variation among effect sizes as represented by the Q statistic can be separated into two pieces, one associated with the variation among the group means, and one that is the variation within groups just as we compute the between and within components of variance in an ANOVA. Categorical models of effect size will compare the mean effect

size within each group using the statistic $Q_{between}$, similar in interpretation to an F-test for between groups. In addition, the within-group component of variation, Q_{within}, can be computed and is a measure of the homogeneity within groups in their estimates of effect size. Note that effect sizes have standard errors that depend on the sample size in the study, and thus do not share a common variance. Techniques of ANOVA assume heteroscedasticity, that the groups share a common within-group variance. The computations of the Q statistics in meta-analysis differ from the usual ANOVA calculations in that the inverse of the standard error of the effect sizes are used in a weighted analysis.

When the moderator variables of interest are continuous, such as the average age of the mentees in the sample, or when we have multiple variables of interest, regression techniques, commonly called meta-regression, can be used to explore the associations between study characteristics and effect size. In meta-analysis, we again use weighted regression techniques since our effect sizes violate the assumption of heteroscedasticity.

One caution with moderator analyses in meta-analysis concerns the types of generalizations that are possible. By definition, models of effect size in meta-analysis are observational studies. Reviewers do not have control over the types of studies that exist in the literature, and thus cannot randomly assign studies to different conditions as we can in a randomized trial. Given that a meta-analysis is an observational study design, we can only talk about associations between study characteristics and variation in study effects. In addition, the relationships we may discover between study characteristics and effect sizes are at the level of the study, and cannot be generalized to any other level. For example, if we find that studies with larger effect sizes tend to have more boys included in the study, this finding does not mean that the treatment is more effective for boys than girls. Our generalization would be subject to an ecological bias – assuming that a finding at the level of the study would hold at the level of the individual.

Random Effects Models of Effect Size

Up to this point, our presentation of meta-analysis has assumed what we call a fixed effect model, that the differences among studies in their effect size estimates is due only to sampling error and observed differences among study methods and characteristics. However, the more commonly used framework for meta-analysis is to assume a random effect model. In a random effects model, the reviewer assumes that the studies we have gathered for our systematic review are sampled from a population of studies and thus effect size estimates are essentially a random sample from an unknown population of effect sizes. Practically speaking, in a random effects model, the reviewer assumes that there may be underlying variation among study effect sizes that is unknowable, or just due to random variation that cannot be accounted for by differences among study methods.

In random effects models, we estimate not only the mean effect size, which represents the mean of the unknown population of effect sizes, but also the variance of that population, denoted by τ^2 and called the random effects variance component. The effect size estimated from each study has a standard error that is the sum of two components, the variance component, τ^2, and the usual sampling variance we detailed in our discussion above. Thus, the standard error for the standardized mean difference for a study is $\tau^2 + se(d)$ given in eqn [3]. Each study's estimate of the effect size will have a larger variance in a random effects model. In addition, adding τ^2 to each study's estimate of the standard error serves to make the weights (the inverse of the standard error of the effect sizes) more equal to each other (the ranks of the weights remain the same however, so the study with the most weight in the fixed effect analysis also has the most weight in the random effects analysis). The analysis of the overall mean effect size using random effects models focuses on estimating both the overall mean effect (using the inverse of the random effects standard errors as weights), and the variance component, τ^2. Confidence intervals around the overall mean effect size use the value of the variance component, and we test the homogeneity of the effect sizes through the statistical test of whether τ^2 is statistically different from zero.

Future Directions for Meta-Analysis

New meta-analysis work centers on two areas: analysis methods for study designs and research questions not addressed by current meta-analysis techniques, and methodological research on more-efficient and less-biased methods for computing effect size estimates and models of effect sizes.

Diagnostic Test Accuracy

Researchers in medicine have been developing methods for examining the accuracy of diagnostic tests. Often, the 'gold standard' test of a particular diagnosis is expensive or invasive, such as using tissue biopsy to diagnose cancer. Newer and less-invasive diagnosis procedures are often suggested, and researchers are interested in how accurate these tests are in comparison to the old gold standard test. There are a series of methods to synthesize the results of studies that compare the accuracy of a newer test to the gold standard for diagnosing a condition. Researchers in the social and behavioral sciences have not yet applied these methods for psychological or other types of tests. There are a number of areas, however, where students are diagnosed with a particular condition or tests are used to predict a student's later performance. The literature in these areas could be synthesized using techniques developed in medicine for examining diagnostic test accuracy.

Bayesian Meta-Analysis

As discussed earlier, random effects models of effect size have become the most common method for synthesizing a set of effect sizes. One difficulty, however, is the computation of the variance component, τ^2. There is much research on the drawbacks of our usual methods for estimating the variance component and the assumptions we make in using the current methods for computation. Bayesian methods, in general, have become more common in statistical analyses given innovations in computing and more readily available computer programs to conduct these analyses. In a Bayesian analysis, we are concerned with estimating the joint distribution of parameters of interest – in meta-analysis, the overall mean effect size, the effect sizes for each study, and the variance component – given the data we have in hand. More formally, we are interested in estimating the joint posterior distribution of the parameters of

interest in meta-analysis given the data. Bayesian analysis allows us to make probability statements about our parameters of interest, stating, for example, that the overall mean effect size has a 95% probability of lying between two points (this is how people often interpret 95% confidence intervals arising from more classical statistical traditions, but that interpretation is not strictly correct). In addition, Bayesian methods allow the use of more complex distributions of parameters so that we do not have to make simplifying assumptions such as normality that are not realistic. Bayesian methods are used more extensively in medicine especially in the synthesis of studies in genetics, and are likely to allow social and behavioral researchers more flexibility in applying more complex models of effect sizes.

See also: Logic Models; Meta-Analysis, History of.

Bibliography

Anderson, L.M., Petticrew, M., Rehfuess, E., Armstrong, R., Ueffing, E., Baker, P., Francis, D., Tugwell, P., 2011. Using logic models to capture complexity in systematic reviews. Research Synthesis Methods 2, 33–42.

Cooper, H., 1982. Scientific guidelines for conducting integrative research reviews. Review of Educational Research 52, 291–302.

Cooper, H., 2003. Psychological bulletin: editorial. Psychological Bulletin 129, 3–9.

DuBois, D.L., Portillo, N., Rhodes, J.E., Silverthorn, N., Valentine, J.C., 2011. How effective are mentoring programs of youth?: a systematic assessment of the evidence. Psychological Science in the Public Interest 12, 57–91.

Herrera, C., DuBois, D.L., Grossman, J.B., 2013. The Role of Risk: Mentoring Experiences and Outcomes for Youth with Varying Risk Profiles. A Public/Private Ventures Project Distributed by MDRC, New York.

Jüni, P., Witshci, A., Bloch, R., Egger, M., 1999. The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association 282, 1054–1060.

Lawner, E., Beltz, M., Moore, K.A., 2013. What Works for Mentoring Programs: Lessons from Experimental Evaluations of Programs and Interventions. Child Trends Research Brief #2013-14. Child Trends, Bethesda, MD.

Lipsey, M.W., 2009. Identifying interesting variables and analysis opportunities. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 147–158.

Lipsey, M.W., Wilson, D.B., 2001. Practical Meta-Analysis. Sage Publications, Inc., Thousand Oaks, CA.

Spiegelhalter, D.J., Abrams, K.R., Myles, J.P., 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. John Wiley and Sons Limited, Chichester.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., Song, F., 2000. Methods for Meta-Analysis in Medical Research. Wiley, New York.

Valentine, J.C., 2009. Judging the quality of primary research for research synthesis. In: Cooper, H., Hedges, L.V., Valentine, J.C. (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed. Russell Sage Foundation, New York, pp. 129–146.

Valentine, J.C., Cooper, H., 2008. A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the Study Design and Implementation Assessment Device (Study DIAD). Psychological Methods 13, 130–149.

Valentine, J.C., McHugh, C., 2007. The effects of attrition on baseline group comparability in randomized experiments in education: a meta-analytic review. Psychological Methods 12, 268–282.

Valentine, J.C., Pigott, T.D., Rothstein, H.R., 2010. How many studies do you need?: a primer on statistical power in meta-analysis. Journal of Educational and Behavioral Statistics 35, 215–247.

Valentine, J.C., Biglan, A., Boruch, R.F., Castro, F.G., Collins, L.M., Flay, B.R., Kellam, S., Mościcki, E.K., Schinke, S.P., 2011. Replication in prevention science. Prevention Science 12, 103–117.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 24

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 24

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 24 *xlv*

VOLUME 24

Taboo 1
Michael Lambek

Tacit Knowledge, Psychology of 4
Derrick McIver, Cynthia A Lengnick-Hall, and Indu Ramachandran

Tactile Maps in Geography 9
Regina Araujo de Almeida (Vasconcellos)

Tactile Processing 14
Alberto Gallace

Tangible User Interfaces in Learning and Education 20
Yuxia Zhou and Minjuan Wang

Taste and Smell, Psychology of 26
Edmund T Rolls

Tattoos and Body Modification 32
Mary Kosut

Taylor, Charles (1931–) 39
Mark Redhead

Teacher Behaviours and Student Outcomes 44
James H Stronge, Leslie W Grant, and Xianxuan Xu

Teacher Education 51
Ian Menter

Teacher Expertise 56
Julian G Elliott

Teacher Leadership 60
Alma Harris

Teacher Motivation 64
Helen MG Watt and Paul W Richardson

Teacher Stress and Burnout: Methodological Perspectives 72
Chris Kyriacou

Teaching as a Profession: United States 75
Ilana S Horn

Teaching Economics 83
Micheal Moos van Wyk

Teaching for Metacognition 89
Marcel VJ Veenman

Teamwork and Team Performance Measurement 96
Christopher W Wiese, Marissa L Shuffler, and Eduardo Salas

Technological Change and the Sociotechnical System, Applied Psychology of 104
Beryl Hesketh and Warwick Graco

Technological Determinism 109
Ronald R Kline

Technology and Organization 113
Tarek Tantoush

Technology and Social Control 117
Gary T Marx

Technology Assessment 125
Arie Rip

Technology, Evolution of 129
Stephen Shennan

Technology, Social Construction of 135
Wiebe E Bijker

Technology Transfer 141
Shiv Visvanathan

Teens, Gender, and Self-Presentation in Social Media 146
Susan C Herring and Sanja Kapidzic

Telegraph 153
David Hochfelder

Telephone 157
James E Katz

Television: General 163
Jostein Gripsrud

Television: History 169
James L Baughman

Temperament 173
Samuel P Putnam

Temperament and Human Development 180
MK Rothbart

Temperament and Motivation 184
Sari Mullola, Mirka Hintsanen, and Liisa Keltikangas-Järvinen

Temperament Development, Theories of 191
Xiaoxue Fu and Koraly Pérez-Edgar

Temporal Distinctiveness Models of Memory 199
Gordon DA Brown

Temporal Lobe 206
Bryan Kolb

Tense, Aspect, Mood and Evidentiality, Linguistics of 210
Östen Dahl

Territorial Conflicts 214
Niall Ó Dochartaigh

Territoriality: Geographical 221
David Storey

Terror and the Internet 227
Gabriel Weimann

Terrorism 237
Stephen Vertigans

Test Anxiety and Academic Achievement 244
Reinhard Pekrun and Elizabeth J Stephens

Test Theory: Applied Probabilistic Measurement Structures 250
Hartmann Scheiblechner

Textbooks 256
Richard Venezky

Thematic Maps in Geography 260
David Forrest

Theory of Mind During Infancy and Early Childhood Across Cultures, Development of 268
Beate Sodian and Susanne Kristen

Therapeutic Community 274
Rex Haigh

Therapeutic Effectiveness Assessment: Approaches in Contemporary Research 281
Mihaela Matei and François Lemaire

Therapist–Patient Relationship 287
Paul Crits-Christoph

Thick Description: Methodology 291
Tanya M Luhrmann

Think Tanks 294
Diane Stone

Third World 300
Maxwell Owusu

Thorndike, Edward Lee (1874–1949) 305
Robert M Thorndike

Time Perception Models 308
Lorraine G Allan

Time Series: Advanced Methods 311
Wai Keung Li and Howell Tong

Time Series: ARIMA Methods 316
George C Tiao

Time Series: Cointegration 322
Søren Johansen

Time Series: Cycles 331
Mark W Watson

Time Series: Economic Forecasting 337
James H Stock

Time Series: General 341
David R Brillinger

Time Series: Seasonal Adjustment 347
Estela B Dagum

Time Series: State Space Methods 354
Siem J Koopman and Jacques JF Commandeur

Time Series: Unit-Roots and Nonstandard Limiting Distributions 362
Ngai H Chan

Time Use and Gender 369
Tanja van der Lippe

Time Use and Resource Consumption 373
Henrike Rau

Time Use Research Methods 379
Jonathan Gershuny

Time Use Research: Recent Developments 383
John P Robinson and Teresa A Harms

Time Use Research: Subjective Time Use 398
John P Robinson and Jonathan Gershuny

Time, Anthropology of 406
Eric K Silverman

Time, Chronology, and Periodization in History 409
Masayuki Sato

Time-Space in Geography 415
Donald G Janelle

Tocqueville, Alexis de (1805–59) 421
Bernard Valade

Topographic Maps 426
Jon H Kaas

Torts in Law 429
Stephen D Sugarman

Torture and its Consequences, Psychology of 434
Danie Meyer-Parlapanis and Thomas Elbert

Torture: Legal Perspectives 442
Alexa Koenig and Jonathan Simon

Totalitarianism 448
Leslie Holmes

Totemism 453
Warren Shapiro

Touraine, Alain (1925–) 459
Howard Davis

Tourism 465
Shinji Yamashita

Tourism, Geography of 469
Allan M Williams and Gareth Shaw

Tourism, the Millennium Development Goals, and Gender Equality 474
Lucy Ferguson

Trade and Exchange, Archaeology of 481
Glenn R Summerhayes

Trade Policy 485
Jean-Baptiste Velut

Trade Unions and Social Work: Lessons from Canada 491
Winnie Ng

Trade Unions, Economic Behavior of 497
Alison Booth

Tradition, Anthropology of 503
Richard Bauman

Tradition: Social 508
Simon Langlois

Trafficking of Women and Children in Latin America 512
Mirna E Carranza

Training in Clinical Psychology in the United States: Practitioner Model 517
Roger L Peterson, Donald R Peterson, Jules C Abrams, George Stricker, and Kelly Ducheny

Training in Clinical Psychology in the United States: Scientist–Practitioner Model 524
Bonnie R Strickland

Transaction Costs and Property Rights 528
Oliver E Williamson

Transcranial Direct Current Stimulation and Language 533
Barry Gordon, James Livengood, Joseph J van Steenburgh, Rasika Jayatillake, Tracy D Vannorsdall, and David J Schretlen

Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts 545
Neil K Aggarwal

Transgender, Transexual, and Gender Variant Individuals 549
M Paz Galupo, JF Stuart, and Derek P Siegel

Transient Epileptic Amnesia 554
Serge Hoefeijzers and Adam Zeman

Transient Global Amnesia 562
Juliane Döhring, Karsten Witt, and Thorsten Bartsch

Transition to Adulthood 571
Maria Sironi

Transition to Adulthood, Cultural Differences in 576
Carol A Markstrom, Sera A Mathew, and Kalynn B Amick

Translating Sociology into Japanese, Chinese and Korean Languages 583
Jeremy CA Smith, Myoungkyu Park, and Albert Tzeng

Transnational Citizenship 588
Claus Leggewie

Transnational Organized Crime 596
Toine Spapens

Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns 602
Dale F Eickelman

Transportation Geography 607
Genevieve Giuliano, Sandip Chakrabarti, and Mohja Rhoads

Transportation Planning 616
Qing Shen

Transportation, Psychology of 623
Raymond W Novaco

Transportation, Supply, and Congestion 629
Robin Lindsey and André de Palma

Trials 635
Malcolm M Feeley

Triangulation: Methodology 639
Linda Heath

Tribe 645
Andre Gingrich

Trickster Ethnography 648
Jesse W Shipley

Trust, Political 658
Eric M Uslaner

Trust, Sociology of 664
Margaret Levi

Trust: Philosophical Aspects 668
Karen Jones

Truth and Credibility in Science 673
Steven Shapin

Tversky, Amos (1937–96) 679
Baruch Fischhoff

Twin Studies and Epigenetics 683
Tiphaine C Martin, Jordana T Bell, and Timothy D Spector

Twin Studies in Demography 703
Jere R Behrman

Tyranny 710
Stephen D Reicher and S Alexander Haslam

Uncertainty: History of the Concept 716
Joshua Wakeham

Unconscious: History of the Concept 722
Donald Levy

Underclass 728
Paul A Jargowsky and Lorraine C Minnite

Unemployment and Mental Health 733
David Fryer and Rose Stambe

Unemployment: Structural 738
Assar Lindbeck

Unintended Consequences: History of the Concept 744
Adriana Mica

United Nations and its Agencies 750
Janice W Wetzel and Nancy E Wallace

Universities and Science and Technology: Europe 758
Barbara Sporn

Universities and Science and Technology: United States 762
Michael A Dennis

Unobtrusive Methods, History of 767
Raymond M Lee

Urban Anthropology 774
Peter JM Nas

Urban Conflicts 783
Liam O'Dowd

Urban Design and Public Space 789
Ali Madanipour

Urban Entrepreneurialism 795
Andrew Wood and Timothy Brock

Urban Geography 800
Rita Schneider-Sliwa

Urban Growth Models 807
Britten Harris

Urban History 811
Rosemary Sweet

Urban Informality: The Production and Regulation of Space 818
Ananya Roy

Urban Internal Spatial Structure 823
Yan Song

Urban Life and Health 830
Roderick J Lawrence

Urban Life: Future Developments 836
MA Hughes

Urban Megaprojects 838
Fernando D Orueta

Urban Neoliberalism 846
Ugo Rossi and Alberto Vanolo

Urban Places: Statistical Definitions 854
Richard L Forstall and Kam Wing Chan

Urban Planning (Western), History of 862
Robert Freestone

Urban Planning in China 869
Tingwei Zhang

Urban Planning, Competitions and Exhibitions 882
Carola Hein

Urban Planning: Methods and Technologies 889
Richard E Klosterman

Urban Policies in Asia 894
Mee Kam Ng

Urban Policy in Europe 900
Patrick Le Galès

Urban Policy in North America 908
Michael B Teitz

Urban Poverty in Neighborhoods 915
JE Grigsby, III

Urban Regeneration 919
Claudio de Magalhães

Urban Sociology 926
Anne B Shlay and John Balzarini

Urban Sprawl 934
Robert Bruegmann

Urban Studies: Overview 940
Ronan Paddison

Urban System in Geography 945
Juan M Kanai

Urban Violence 949
Douglas S Massey

Urban Violence and Militarization 953
Jonathan Mendel

Urbanization 958
Gordon McGranahan

Urbanization in Africa 965
Deborah Potts

Urbanization in China 973
Hyun Bang Shin

Utilitarian Social Thought, History of 980
James E Crimmins

This page intentionally left blank

CONTRIBUTORS TO VOLUME 24

Jules C. Abrams
Widener University, Chester, PA, USA

Neil K. Aggarwal
New York State Psychiatric Institute, Columbia University Medical Center, New York, NY, USA

Lorraine G. Allan
McMaster University, Hamilton, ON, Canada

Regina Araujo de Almeida (Vasconcellos)
University of São Paulo, São Paulo, Brazil

Kalynn B. Amick
College of Education and Human Services, West Virginia University, Morgantown, WV, USA

John Balzarini
Temple University, Philadelphia, PA, USA

Thorsten Bartsch
University Hospital Schleswig-Holstein, University of Kiel, Kiel, Germany

James L. Baughman
University Avenue, Madison, WI, USA

Richard Bauman
Indiana University, Bloomington, IN, USA

Jere R. Behrman
University of Pennsylvania, Philadelphia, PA, USA

Jordana T. Bell
St Thomas' Hospital, Kings College London, London, UK

Wiebe E. Bijker
Maastricht University, Maastricht, The Netherlands

Alison Booth
Australian National University, Canberra, ACT, Australia

David R. Brillinger
University of California, Berkeley, CA, USA

Timothy Brock
University of Kentucky, Lexington, KY, USA

Gordon D.A. Brown
University of Warwick, Coventry, UK

Robert Bruegmann
University of Illinois at Chicago, Chicago, IL, USA

Mirna E. Carranza
School of Social Work, McMaster University, Hamilton, ON, Canada

Sandip Chakrabarti
Sol Price School of Public Policy, University of Southern California, Los Angeles, CA, USA

Kam Wing Chan
University of Washington, Seattle, WA, USA

Ngai H. Chan
Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Jacques J.F. Commandeur
Vrije Universiteit, Amsterdam, Netherlands

James E. Crimmins
The University of Western Ontario, London, ON, Canada

Paul Crits-Christoph
University of Pennsylvania, Philadelphia, PA, USA

Estela B. Dagum
University of Bologna, Bologna, Italy

Östen Dahl
Stockholm University, Stockholm, Sweden

Howard Davis
School of Social Sciences, Bangor University, Gwynedd, UK

Claudio de Magalhães
University College London, London, UK

André de Palma
Ecole Normale Supérieure de Cachan, Cachan, France

Michael A. Dennis
Georgetown University, Washington, DC, USA

Niall Ó. Dochartaigh
School of Political Science and Sociology, National University of Ireland Galway, Galway, Ireland

Juliane Döhring
University Hospital Schleswig-Holstein, University of Kiel, Kiel, Germany

Kelly Ducheny
Howard Brown Health Center, Chicago, IL, USA

Dale F. Eickelman
Dartmouth College, Hanover, NH, USA

Thomas Elbert
University of Konstanz, Konstanz, Germany

Julian G. Elliott
Durham University, Durham, UK

Malcolm M. Feeley
School of Law, University of California, Berkeley, Berkeley, CA, USA

Lucy Ferguson
Instituto Complutense de Estudios Internacionales, Universidad Complutense de Madrid, Madrid, Spain

Baruch Fischhoff
Carnegie Mellon University, Pittsburgh, PA, USA

David Forrest
School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK

Richard L. Forstall
Alexandria, VA, USA

Robert Freestone
University of New South Wales, Sydney, NSW, Australia

David Fryer
University of Queensland, Brisbane, QLD, Australia; University of South Africa, South Africa; and University of Stirling, Stirling, UK

Xiaoxue Fu
The Pennsylvania State University, University Park, PA, USA

Alberto Gallace
University of Milano-Bicocca, Milan, Italy

M. Paz Galupo
Towson University, Towson, MD, USA

Jonathan Gershuny
Nuffield College, University of Oxford, Oxford, UK

Andre Gingrich
Institute for Social Anthropology, Austrian Academy of Sciences, Vienna, Austria

Genevieve Giuliano
Sol Price School of Public Policy, University of Southern California, Los Angeles, CA, USA

Barry Gordon
The Johns Hopkins University, Baltimore, MD, USA

Warwick Graco
Australian Taxation Office, ACT, Australia

Leslie W. Grant
School of Education, College of William and Mary, Williamsburg, VA, USA

J.E. Grigsby
University of California, Los Angeles, CA, USA

Jostein Gripsrud
University of Bergen, Bergen, Norway

Rex Haigh
Berkshire Healthcare NHS Foundation Trust, UK; Royal College of Psychiatrists, UK; and Institute of Mental Health, University of Nottingham, Nottingham, UK

Teresa A. Harms
University of Oxford, Oxford, UK

Alma Harris
Institute of Education, University College, London, UK

Britten Harris
University of Pennsylvania, Philadelphia, PA, USA

S. Alexander Haslam
School of Psychology, The University of Queensland, St Lucia, QLD, Australia

Linda Heath
Loyola University Chicago, Chicago, IL, USA

Carola Hein
History of Architecture and Urban Planning, TU Delft, The Netherlands

Susan C. Herring
Indiana University Bloomington, Bloomington, IN, USA

Beryl Hesketh
University of Western Sydney, Sydney, NSW, Australia

Mirka Hintsanen
IBS, Unit of Personality, Work and Health Psychology, University of Helsinki, Helsinki, Finland

David Hochfelder
Rutgers University, New Brunswick, NJ, USA

Serge Hoefeijzers
University of Edinburgh, Edinburgh, UK

Leslie Holmes
School of Social and Political Sciences, University of Melbourne, VIC, Australia

Ilana S. Horn
Vanderbilt University, Nashville, TN, USA

M.A. Hughes
University of Pennsylvania, Philadelphia, PA, USA

Donald G. Janelle
University of California, Santa Barbara, CA, USA

Paul A. Jargowsky
Rutgers-Camden, Camden, NJ, USA

Rasika Jayatillake
The Johns Hopkins University, Baltimore, MD, USA

Søren Johansen
University of Copenhagen, Copenhagen, Denmark; and Aarhus University, Aarhus, Denmark

Karen Jones
The University of Melbourne, Parkville, VIC, Australia

Jon H. Kaas
Vanderbilt University, Nashville, TN, USA

Juan M. Kanai
University of Miami, Miami, FL, USA

Sanja Kapidzic
Ludwig-Maximilians-Universität, München, Germany

James E. Katz
Division of Emerging Media Studies, Center for Mobile Communication Studies, Boston University, Boston, MA, USA

Liisa Keltikangas-Järvinen
IBS, Unit of Personality, Work and Health Psychology, University of Helsinki, Helsinki, Finland

Ronald R. Kline
Cornell University, Ithaca, NY, USA

Richard E. Klosterman
What if?, Inc., Hudson, OH, USA

Alexa Koenig
U.C. Berkeley School of Law, Berkeley, CA, USA

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada

Siem J. Koopman
Vrije Universiteit, Amsterdam, Netherlands

Mary Kosut
State University of New York, Purchase, NY, USA

Susanne Kristen
Ludwig-Maximilians-University Munich, Munich, Germany

Chris Kyriacou
University of York, York, UK

Michael Lambek
University of Toronto, Toronto, ON, Canada

Simon Langlois
Université Laval, Québec, QC, Canada

Roderick J. Lawrence
Institute of Environmental Sciences, Faculty of Economic and Social Sciences, University of Geneva, Geneva, Switzerland

Raymond M. Lee
Royal Holloway University of London, Egham, UK

Patrick Le Galès
Sciences Po, CNRS, Centre d'études européennes, Paris, France.

Claus Leggewie
University of Giessen, Germany; and Kulturwissenschaftliches Institut (KWI), Institute for Advanced Study in the Humanities, Essen, Germany

François Lemaire
Eméritus at the University of Paris Est Créteil, France

Cynthia A. Lengnick-Hall
University of Texas at San Antonio, San Antonio, TX, USA

Margaret Levi
University of Washington, Seattle, WA, USA; and United States Studies Centre at University of Sydney, Sydney, NSW, Australia

Donald Levy
City University of New York, New York, NY, USA

Wai Keung Li
The University of Hong Kong, Hong Kong

Assar Lindbeck
Stockholm University, Stockholm, Sweden

Robin Lindsey
Sauder School of Business, University of British Columbia, Vancouver, BC, Canada

James Livengood
The Johns Hopkins University, Baltimore, MD, USA

Tanya M. Luhrmann
Stanford University, Stanford, CA, USA

Ali Madanipour
Newcastle University, Newcastle upon Tyne, UK

Carol A. Markstrom
College of Education and Human Services, West Virginia University, Morgantown, WV, USA

Tiphaine C. Martin
St Thomas' Hospital, Kings College London, London, UK

Gary T. Marx
Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

Douglas S. Massey
Office of Population Research, Princeton University, Princeton, NJ, USA

Mihaela Matei
Centre Européen d'Etudes et de Recherche Droit et Santé, Université Montpellier I, France

Sera A. Mathew
College of Education and Human Services, West Virginia University, Morgantown, WV, USA

Gordon McGranahan
International Institute for Environment and Development, London, UK

Derrick McIver
Western Michigan University, Kalamazoo, MI, USA

Jonathan Mendel
Geography, School of the Environment, University of Dundee, Dundee, UK

Ian Menter
University of Oxford, Oxford, UK

Danie Meyer-Parlapanis
University of Konstanz, Konstanz, Germany

Adriana Mica
Institute of Applied Social Sciences, University of Warsaw, Warsaw, Poland

Lorraine C. Minnite
Rutgers-Camden, Camden, NJ, USA

Micheal Moos van Wyk
College of Education, University of South Africa, Pretoria, South Africa

Sari Mullola
IBS, Unit of Personality, Work and Health Psychology, University of Helsinki, Helsinki, Finland

Peter J.M. Nas
Leiden University, Leiden, The Netherlands

Mee Kam Ng
The Chinese University of Hong Kong, Shatin, Hong Kong, China

Winnie Ng
CAW-Sam Gindin Chair in Social Justice and Democracy, Ryerson University, Toronto, ON, Canada

Raymond W. Novaco
University of California, Irvine, CA, USA

Liam O'Dowd
Queen's University, Belfast, UK

Fernando D. Orueta
Universidad de La Rioja, Logrono, La Rioja, Spain

Maxwell Owusu
University of Michigan, Ann Arbor, MI, USA

Ronan Paddison
School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK

Myoungkyu Park
Seoul National University, Seoul, Republic of Korea

Reinhard Pekrun
University of Munich, Munich, Germany

Koraly Pérez-Edgar
The Pennsylvania State University, University Park, PA, USA

Donald R. Peterson[†]
Rutgers University, NJ, USA

Roger L. Peterson
Antioch University New England, Keene, NH, USA

Deborah Potts
King's College London, London, UK

Samuel P. Putnam
Bowdoin College, Brunswick, ME, USA

Indu Ramachandran
Texas State University, San Marcos, TX, USA

Henrike Rau
School of Political Science and Sociology, National University of Ireland, Galway, Ireland

Mark Redhead
California State University, Fullerton, CA, USA

Stephen D. Reicher
School of Psychology and Neuroscience, University of St Andrews, St Andrews, UK

[†] Deceased.

Mohja Rhoads
Sol Price School of Public Policy, University of Southern California, Los Angeles, CA, USA

Paul W. Richardson
Faculty of Education, Monash University, Melbourne, Victoria, Australia

Arie Rip
University of Twente, Enschede, The Netherlands

John P. Robinson
University of Maryland, College Park, MD, USA

Edmund T. Rolls
Oxford Centre for Computational Neuroscience, Oxford, UK; and University of Warwick, Coventry, UK

Ugo Rossi
Università di Torino, Turin, Italy

M.K. Rothbart
University of Oregon, Eugene, OR, USA

Ananya Roy
University of California, Berkeley, CA, USA

Eduardo Salas
Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA

Masayuki Sato
University of Yamanashi, Kōfu, Yamanashi, Japan

Hartmann Scheiblechner†
University of Marburg, Marburg, Germany

Rita Schneider-Sliwa
University of Basle, Basle, Switzerland

David J. Schretlen
The Johns Hopkins University, Baltimore, MD, USA; and The Johns Hopkins University School of Medicine, Baltimore, MD, USA

Steven Shapin
University of California-San Diego, La Jolla, CA, USA

Warren Shapiro
Rutgers University, New Brunswick, NJ, USA

Gareth Shaw
University of Exeter Business School, Exeter, UK

Qing Shen
University of Washington, Seattle, WA, USA

Stephen Shennan
UCL Institute of Archaeology, Gordon Square, London, UK

Hyun Bang Shin
The London School of Economics and Political Science, London, UK

Jesse W. Shipley
Haverford College, Haverford, PA, USA

Anne B. Shlay
Temple University, Philadelphia, PA, USA

Marissa L. Shuffler
College of Business & Behavioral Science, Clemson University, Clemson, SC, USA

Derek P. Siegel
American University, Washington, DC, USA

Eric K. Silverman
De Pauw University, IN, USA

Jonathan Simon
U.C. Berkeley School of Law, Berkeley, CA, USA

Maria Sironi
University of Oxford, Oxford, UK

Jeremy C.A. Smith
Federation University Australia, VIC, Australia

Beate Sodian
Ludwig-Maximilians-University Munich, Munich, Germany

Yan Song
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Toine Spapens
Tilburg University, Tilburg, The Netherlands

Timothy D. Spector
St Thomas' Hospital, Kings College London, London, UK

Barbara Sporn
Vienna University of Economics and Business Administration, Vienna, Austria

Rose Stambe
University of Queensland, Brisbane, QLD, Australia

Elizabeth J. Stephens
University of Munich, Munich, Germany

James H. Stock
Harvard University, Cambridge, MA, USA

Diane Stone
Murdoch University, WA, Australia; and The University of Warwick, Coventry, UK

† Deceased.

David Storey
University of Worcester, Worcester, UK

George Stricker
American School of Professional Psychology at Argosy University Washington, DC, USA

Bonnie R. Strickland
University of Massachusetts, Amherst, MA, USA

James H. Stronge
School of Education, College of William and Mary, Williamsburg, VA, USA

J.F. Stuart
Towson University, Towson, MD, USA

Stephen D. Sugarman
UC Berkeley, Berkeley, CA, USA

Glenn R. Summerhayes
University of Otago, Dunedin, New Zealand

Rosemary Sweet
University of Leicester, Leicester, UK

Tarek Tantoush
Open University, Nottingham, UK

Michael B. Teitz
University of California, Berkeley, CA, USA

Robert M. Thorndike
Western Washington University, Bellingham, WA, USA

George C. Tiao
University of Chicago, Chicago, IL, USA

Howell Tong
The London School of Economics, London, UK

Albert Tzeng
International Institute for Asian Studies, Leiden, The Netherlands

Eric M. Uslaner
University of Maryland, College Park, MD, USA; Center for American Law and Political Science, Southwest University of Political Science and Law, Chongqing, China; and Aarhus University, Aarhus, Denmark

Bernard Valade
Paris Descartes University, Paris, France

Tanja van der Lippe
Utrecht University, Utrecht, The Netherlands

Tracy D. Vannorsdall
The Johns Hopkins University, Baltimore, MD, USA

Alberto Vanolo
Università di Torino, Turin, Italy

Joseph J. van Steenburgh
The Johns Hopkins University, Baltimore, MD, USA

Marcel V.J. Veenman
Institute for Metacognition Research (IMR), Hillegom, The Netherlands; and Leiden University, Leiden, The Netherlands

Jean-Baptiste Velut
Université Sorbonne Nouvelle, Paris, France

Richard Venezky
University of Delaware, Newark, DE, USA

Stephen Vertigans
Robert Gordon University, Aberdeen, UK

Shiv Visvanathan
Centre for the Study of Developing Societies, Delhi, India

Joshua Wakeham
University of Alabama, Tuscaloosa, AL, USA

Nancy E. Wallace
World Federation for Mental Health, United Nations

Minjuan Wang
Oriental Scholar of Shanghai International Studies University, Shanghai, China; and San Diego State University, San Diego, CA, USA

Mark W. Watson
Princeton University, Princeton, NJ, USA

Helen M.G. Watt
Faculty of Education, Monash University, Melbourne, Victoria, Australia

Gabriel Weimann
University of Haifa, Haifa, Israel

Janice W. Wetzel
International Association of Schools of Social Work, United Nations

Christopher W. Wiese
Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA

Allan M. Williams
School of Hospitality and Tourism Management, University of Surrey, Guildford, UK

Oliver E. Williamson
University of California, Berkeley, CA, USA

Karsten Witt
University Hospital Schleswig-Holstein, University of Kiel, Kiel, Germany

Andrew Wood
University of Kentucky, Lexington, KY, USA

Xianxuan Xu
School of Education, College of William and Mary, Williamsburg, VA, USA

Shinji Yamashita
The University of Tokyo, Tokyo, Japan

Adam Zeman
University of Exeter Medical School, Exeter, UK

Tingwei Zhang
University of Illinois at Chicago, Chicago, IL, USA

Yuxia Zhou
Yunnan Normal University, Kunming, China

This page intentionally left blank

Taboo

Michael Lambek, University of Toronto, Toronto, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15429–15431, © 2001, Elsevier Ltd.

Abstract

A concept assumed in the Victorian period to be central to the religions of supposedly less evolved societies, the term today is generally used to refer to proscriptions grounded in custom or religion rather than in bureaucratic law or common sense. Derived originally from Polynesia, the subsequent career of the concept has been paradigmatic of anthropological theory more generally. In the latter half of the nineteenth century taboo came to epitomize the essence of primitive thought. The reification of taboo was debunked soundly in the mid-twentieth century by Steiner. Freud and Lévi-Strauss each emphasized the structural significance of a universal incest taboo, while Douglas demonstrated the role of taboos in maintaining systems of classification and moral order. Others have examined taboos from the perspective of those who maintain them, drawing on phenomenology or practice theory to show the relevance of taboos as acts for understanding the constitution of the socially located moral person or, conversely, have explored the efflorescence of taboo violations or fantasies at sites and moments of moral vertigo, social dislocation, and state impunity or simply recognized the centrality to human experience of power and danger generated through negation, transgression, and excess.

A concept assumed in the Victorian period to be central to the religions of supposedly less evolved societies, the term taboo today generally is used to refer to prohibitions grounded in custom or religion rather than in bureaucratic law or common sense and hence bearing some moral weight. Taboos can be widespread, observed over the long term, internalized, and deeply felt, as in the case of incest taboos, or applied temporarily with respect to specific persons, acts, conditions, and objects. This article reviews the history of the concept which can be seen to reflect the development of anthropological theory more generally.

Taboos, Social Order, and Moral Persons

The word taboo is originally Polynesian. Via the writings of explorers like Cook and missionaries like Ellis it rapidly entered ordinary usage in Victorian England. In 1842 the 7th edition of the *Encyclopaedia Britannica* defined taboo succinctly as, 'A word used by the South Sea islanders, and nearly of the same import as prohibited or interdicted. It applies equally to persons and things, and is also expressive of anything sacred, devoted, or eminent' (quoted in Steiner, 1999: p. 164). In the hands of Robertson Smith and Frazer, taboo came to epitomize the essence of primitive thought, distinguishing it from civilized religion (i.e., the Judaeo–Christian tradition) in the way it combined holiness with defilement. In a deconstructive work equivalent to that which Lévi-Strauss performed for totemism, Steiner admirably summarized and critiqued the use of the term through the first half of the twentieth century, demonstrating the problems with applying a culturally particular concept in a comparative and evolutionary fashion. Steiner showed that it was impossible to separate defilement from the sacred in any consistent manner. Steiner's critique of the abstraction and reification of taboo has been pursued by MacIntyre, while his insights into its significance in Polynesia have been considerably enlarged by Gell, Sahlins, and Valeri.

The most profound reflections on taboo have concerned its universal social basis. Both Freud and Lévi-Strauss have seen the incest taboo as fundamental to the constitution of society. For Freud, internalizing prohibitions against intra-familial violence and sexual relations establishes the conditions for the emergence of kinship and moral conscience. Lévi-Strauss argued that the incest taboo generates inter-familial relations, providing the basis for the exchange of women (or marriage partners of either sex) that in effect constitutes society. Whereas Freud in *Totem and Taboo* emphasized the temporal significance of the incest taboo (whose phylogenetic origin in actual historical events is largely dismissed by anthropologists) and the implications for intrapsychic and intrafamilial relations (which finds broader though not unanimous support), Lévi-Strauss argued that the incest taboo forms the logical basis of society and is rooted in the universal mental capacity to make categorical binary distinctions, notably those between self and other, culture and nature. In an ambitious attempt to unify and critique the theory, Rubin added the injunction against homosexual relations.

It is Lévi-Strauss's emphasis on classification that has most influenced subsequent analyses of taboos and has proved fruitful in understanding their varied and often ostensibly trivial content. Drawing on their common heritage of Durkheim, both Douglas and Leach emphasized the role of taboos in protecting systems of social classification by banning anomalous or marginal words, objects, and practices that

might violate the order of the categories or blur their distinctions. The title of Douglas's major work, *Purity and Danger* (1966), captures the crux of her argument perfectly. Dirt, or that which is viewed with revulsion and hence tabooed, is 'matter out of place' and hence always relative to a system of classification but also dangerous for it. Building on Steiner, Douglas argued convincingly that individual taboos can only be understood as parts of larger social wholes; that taboos are not false or failed attempts at science; that Western lives are replete with taboos which are no more or less rational than those of other societies; and that without principles of restriction, conceptual and moral discrimination would be compromised and the ensuing freedom of both thought and action would be hopelessly chaotic.

Douglas's argument has inspired much subsequent work, especially on gender, the body, and food. Among the merits of her approach is the ability to show structural similarities in taboos, whatever their specific content, both within a given society, between societies, and across both ostensibly secular (e.g., hygiene, courtesy) and sacred realms. Somewhat more controversial have been attempts both to illustrate how taboos in various orders of phenomena (notably diet) reflect concerns with the maintenance of social boundaries and divisions and, at a greater level of abstraction, to plot the degree of general interest in maintaining the purity of classificatory orders against the underlying nature of the social relations (the type of society, relatively defined). Following one strain of Durkheim's thought, Douglas gives primacy to the form and integration of social groupings and thus, despite her radical demonstration of the contingent nature of particular taboos, the argument appears rather conservative. Purity and impurity are not only relative to a system of categories, but this system may also be arranged with respect to a hierarchy of sanctity within a holistic universe, as epitomized by Dumont's account of the South Asian caste system. Ensuing questions about the relationship of sanctity and encompassment to power have been much debated.

While the presence of taboos is universal, their substance and direction are not, giving rise to culturally specific social orders. In some parts of the world, like eastern Indonesia, prescriptive rules are emphasized, while in others, like Madagascar, proscriptions are pervasive (Lambek, 1992). Lévi-Strauss demonstrated that specific positive rules of marriage – made possible by the underlying incest taboo – generate certain patterns of relationship over time. Taboos, by contrast, in that they do not specify what must or should be done, allow greater degrees of freedom. Yet, as Fortes argued, they are also grasped more easily and internalized more deeply than positive injunctions.

While taboos usually are analyzed as discursive elements, it is important to understand that they become realized in practice. Often the most consequential taboos are the least explicit. If in a structural model the semantic contents of taboos distinguish who or what one is not, from a practice perspective it is the observance of the taboo that substantiates who one is. Van Gennep saw taboos as acts of obligation, creating both continuity and separation. Fortes emphasized the morally binding quality of taboos and the personal significance of observing them for the internalization of an enduring identity and the acknowledgment of dependence on, and obligation to, higher authorities or values. Moreover, taboos 'crystallize these abstract norms in concrete objects and precise rules of conduct which are the more effective because they are of no utilitarian or rational value' (Fortes, 1987: p. 129).

Fortes observed both the essentially inward orientation of taboos, especially with respect to eating, and their continuous nature. An important consequence of the fact that taboos are phrased in the negative is that their observance is a relatively continuous and private form of action, whereas carrying out a positive order is a discontinuous and relatively discrete and public act. Hence while the power entailed in observing a taboo may be less explicit and forthright than a direct command, the implications may be more profound.

Defiance of taboo may challenge subjection, identity, and value at quite a deep level.

Taboos, then, are not merely imposed, lifted, or carried out in discrete acts but observed continuously. As practical disposition the observance of taboos exemplifies Mauss's and Bourdieu's notion of the habitus and Aristotelian virtue that underlies it. Although not always accepted happily, taboos index personal dignity and (self-)respect, and insofar as they are embodied statements, they may be understood as acts which performatively establish and maintain moral states of social commitment and individual conscience. Moreover, in the visceral reactions to the violation of taboos (nausea, disgust, etc.), underlying social conventions are naturalized. Hence, as Douglas also intuits, taboos do not merely signify social or cosmological order but, in acts of separation, commitment, and principled refusal, they constitute it.

Taboos may be understood phenomenologically with respect to a dialectic of subject and object. Gell argues that it is through taboos, applied at varying levels of specificity across social categories, and operating as intentional, if negative, engagement with the world, that the ego or self is constituted. 'To observe a taboo is to establish an identifiable self by establishing a relationship (negative in this case – *not* eating) with an external reality such that the "self" only comes into existence in and through this relationship' (1979: p. 136). 'Taboos on eating, on killing, on sexual intercourse, on looking at, touching, etc., together circumscribe a "hole" in the texture of shareable intersubjective reality which, privileged with respect to a specific individual, constitutes his soul, or ego, or personality. Taboo does more than express the self: it constitutes the self" (1979: p. 137). When Gell's argument that taboos carve out the self is combined with Lévi-Strauss' argument that taboos generate exchange and hence society, it is clear that the self is relationally constituted, simultaneously held apart from and connected to others, simultaneously social and itself. There is a dynamic quality here: Umeda of Papua New Guinea alternate phases of taboo and consumption; in Gell's analysis the taboo phase restores the integrity of the ego by 'reconstituting it as a closed system' (1979: p. 146). Elsewhere Gell (1995) generalizes the logic of taboo in its original Polynesian locus where it produced differentiation in a cosmos in which sacredness was immanent rather than transcendental.

Transgression

One of Freud's central arguments concerned the essential ambivalence of taboos, understood as the coalescence of the conflict between a desire and its refusal, attraction, and

repression. It is clear that anticipation or suspicion of the violation of taboos often produces visceral reactions – whether laughter, lust, or queasiness – disgust rather than mere indignation. Moreover, because the embodied practice of powerful taboos goes without saying, their pervasive presence is only highlighted by the observance of violation. That is to say, taboos are often indicated through instances of their transgression, in effect, a double negation. As Douglas, Gluckman, and Turner showed, suitably framed transgressions may provide central episodes in religious experience, moral education, and social transition, but otherwise they are perceived as highly threatening. Transgressions may establish political dissent or domination, but they are also the source and substance of neurotic illness, social panic, and much fantasy.

Stallybrass and White (1986) conjoined these ideas with Bakhtin's notion of the carnivalesque and Elias's history of manners to explore the significance of taboos in the historical constitution of the European middle class. The rigorous expulsion of what was perceived as disorderly, low, and carnal has produced, in their view, various 'returns of the repressed.' These include nostalgia, longing, fascination, and even the academic preoccupation with the body.

While some people continue to argue for a biological basis to the incest taboo (and possibly certain other widespread prohibitions), the increased acknowledgment of the prevalence of actual incestuous relations supports those theories that find the source of taboos in culture or society and human ambivalence. Indeed, the question has been raised whether incest taboos are any longer necessary and whether they have been *de facto* lifted, or at least limited. On another front, the violence and rape characteristic of twentieth-century warfare, and the torture carried out by repressive political regimes, must be recognized. Indeed, these might be viewed as deliberate and state-sanctioned liftings of taboos in order to maximize personal violation. However, due in part to fears of contamination that accompany any discussion of transgression, research has focussed more on the condition of victims than the acts of perpetration.

Of course, the presence of taboos has always generated fantasies of transgression. These are often difficult to distinguish from actual transgressions and at the same time may produce transgressive acts of their own, as in the 'de-repression' characteristic of accusations of witchcraft and child abuse prevalent in recent years in North America and northern Europe but also globally (Comaroff, 1997; LaFontaine, 1998). The need for moral discernment is not always matched by epistemological clarity, yet the fantasies and acts, characteristic of sites and moments of moral vertigo and vacuum, underline the near universality of disgust in the face of certain kinds of physical and sexual brutality and invasiveness, the universal need for avenues for the exercise of dignified moral agency, and the place of taboos among such means.

Conversely, it could be argued that the preceding discussion suffers from an overly socialized, moralistic, and tame conception of taboos; that negation and excess are central to religious experience; and that transgression be recognized as 'the ultimate sacred act' (Taussig, 1998: p. 361).

Bibliography

Comaroff, J., 1997. Consuming passions: child abuse, fetishism, and the 'New World Order'. Culture 17, 7–19.

Douglas, M., 1966. Purity and Danger. Routledge & Kegan Paul, London.

Dumont, L., 1967. Homo Hierarchicus. 1970 University of Chicago Press, Chicago, Gallimard, Paris.

Fortes, M., 1987/1966. Totem and taboo. In: Goody, J. (Ed.), Religion, Morality, and the Person. Cambridge University Press, Cambridge, UK.

Freud, S., 1958. Totem and taboo. In: Strachey, J. (Ed. and Trans.), The Standard Edition of the Complete Psychological Works of Sigmund Freud. vol. 13. Hogarth, London.

Gell, A., 1979. Reflections on a cut finger: taboo in the Umeda conception of the self. In: Hook, R.H. (Ed.), Fantasy and Symbol. Academic Press, London.

Gell, A., 1995. Closure and multiplication: an essay on Polynesian cosmology and ritual. In: de Coppet, D., Iteanu, A. (Eds.), Cosmos and Society in Oceania. Berg, Oxford, UK.

LaFontaine, J.S., 1998. Speak of the Devil. Cambridge University Press, Cambridge, UK.

Lambek, M., 1992. Taboo as cultural practice among Malagasy speakers. Man 27 (2), 245–266.

Leach, E., 1964. Anthropological aspects of language: animal categories and verbal abuse. In: Lenneberg, E.H. (Ed.), New Directions in the Study of Language. MIT Press, Cambridge, MA.

Lévi-Strauss, C., 1949. Les structures élémentaires de la parenté, first ed. Presses Universitaires de France, Paris (1966 The Elementary Structures of Kinship. Beacon, Boston).

MacIntyre, A., 1990. Three Rival Versions of Moral Enquiry. Duckworth, London.

Rubin, G., 1975. The traffic in women: notes on the 'Political Economy' of sex. In: Reiter, R.R. (Ed.), Toward an Anthropology of Women. Monthly Review Press, New York.

Stallybrass, P., White, A., 1986. The Politics and Poetics of Transgression. Cornell University Press, Ithaca, NY.

Steiner, F., 1999/1956. Taboo. In: Adler, J., Fardon, R. (Eds.), Taboo, Truth, and Religion: Franz Baermann Steiner Selected Writings. vol 1. Berghahn, New York.

Taussig, M., 1998. Transgression. In: Taylor, M.C. (Ed.), Critical Terms for Religious Studies. University of Chicago Press, Chicago.

Van Gennep, A., 1904. Tabou et totémisme à Madagascar. Ernest Leroux, Paris.

Tacit Knowledge, Psychology of

Derrick McIver, Western Michigan University, Kalamazoo, MI, USA
Cynthia A Lengnick-Hall, University of Texas at San Antonio, San Antonio, TX, USA
Indu Ramachandran, Texas State University, San Marcos, TX, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.S. Reber, volume 23, pp. 15431–15435, © 2001, Elsevier Ltd.

Abstract

In his initial article in the *International Encyclopedia of Social and Behavioral Sciences* titled "Tacit Knowledge: Psychology of," Reber explored the ideas presented in Michael Polanyi's original work on tacit knowledge and discussed some of the trends, controversies, and disputes surrounding the topic. We build on this to show that as tacit knowledge has been diffused into management research, similar challenges, debates, and controversies about tacit knowledge continue.

Tacit knowledge is gaining increasing importance in a number of research domains because of its relation to value creation and a variety of organizational and individual processes, coupled with the challenges of leveraging and scaling the benefits of know-how that is embedded and difficult to articulate (Coff et al., 2006). In his initial article in the *International Encyclopedia of Social and Behavioral Sciences* titled "Tacit Knowledge: Psychology of," and in his previous work (1992, 1993), Reber explored the ideas presented in Michael Polanyi's (1958, 1966) original work that viewed tacit knowledge as information that was intensely personal, introspectively opaque, and communicatively unresponsive. Complex knowledge that is abstract in nature and deeply causal – typically referred to as tacit – has been a disputed and controversial domain for decades. Polanyi wrote that we know more about familiar conceptual domains than we can articulate and described how his own early life as a creative scientist was largely a struggle to make tacit knowledge more explicit and open to communication and articulation. The goal of converting tacit understanding to explicit and more easily shared knowledge dominated the attention of those conducting research in this arena.

Reber focused his article in the first volume of this encyclopedia on the psychology of tacit knowledge, which he acknowledged was often described in the psychology literature using a variety of other terms (e.g., implicit, procedural, covert, nonconscious). These terms describe knowledge that is acquired and used with little in the way of top–down control of consciousness. His article focused on four classical topics that have dominated psychological thought since the 1970s: perception, memory, learning, and motivation. He also outlined the methodological, functional, and representational disputes and controversies that have plagued and continue to plague literature on tacit knowledge.

The purpose of this article is not to update Reber's work and discuss the recent psychology-based arguments and directions of tacit knowledge research, but to show that as tacit knowledge has been applied to other domains and specifically as it has diffused into management research, similar challenges, debates, and controversies about tacit knowledge continue.

Why Is Tacit Knowledge So Important to Management Research?

During the last few decades, tacit knowledge has undergone a resurgence of attention in many disciplines and domains. Management research is one of the domains where knowledge and specifically tacit knowledge has become an important concept to grasp and to learn how to actively manage. Over the past two decades, the interest in and importance of understanding tacit knowledge has grown and proliferated into a variety of management research field areas such as organizational knowledge, the knowledge-based view, knowledge-creating organizations, the resource-based view of the firm, and strategic human resource management (Grant, 1996; Nonaka and Takeuchi, 1995; Orlikowski, 2002; Tsoukas, 1996). This attention has elevated the knowledge-based view of organizations to one of the main theories in management research that attempts to explain firm performance and more importantly to explain why some firms are able to outperform others and to sustain a competitive advantage (Eisenhardt and Santos, 2002). The foundation of the knowledge-based view is that one of an organization's main functions and what an organizational system can efficiently do is generate, combine, recombine, and exploit knowledge which leads to performance (Kogut and Zander, 1996). This has led many management researchers to identify knowledge as the most strategically significant resource of the firm (Dierickx and Cool, 1989; Grant, 1996; Kogut and Zander, 1992). However, researchers acknowledge that simply having strategically valuable knowledge within an organization is a necessary, but not sufficient, condition for superior performance. Knowledge must be effectively used, leveraged, and scaled within an organization and at the same time protected from imitation by rival firms to realize and sustain performance benefits.

Therefore, underlying most recent research is the need to uncover and understand characteristics of knowledge that relate to the generation, combination, recombination, and application or use of knowledge and/or knowledge workers. At the foundation of knowledge management theories, however, two fundamental questions identified by Grant

(1996: 110) remain unresolved: (1) What is knowledge? (2) What are the characteristics of knowledge which have critical implications for management? Answers to these two questions remain elusive and controversial. Within the knowledge-based view of organizations, finding answers to these two questions may provide direction, not only for future research, but for the creation of sustainable competitive advantages in organizations.

Two Schools of Thought on the Meaning of Tacit

The disputes and controversies surrounding organizational knowledge can be linked to the development of two distinct views of organizational knowledge. On the one hand, the 'commodity' school has focused on understanding the different types of 'knowledge' in an organization and then linking these types to other organizational phenomena such as innovation, alliance formation, and performance (e.g., Becerra et al., 2008; Nonaka and Takeuchi, 1995). On the other hand, the 'community' school has focused on 'knowing.' This second school has emphasized the lively act of knowing something and using that knowledge to achieve desirable outcomes (Brown and Duguid, 1991; Cook and Brown, 1999; Orlikowski, 2002; Pfeffer and Sutton, 1999; Tsoukas, 2003). Although potentially relevant to one another, the two schools of thought evolved almost independently with little theoretical integration. Among knowledge management theorists, however, a debate remains about what the important and meaningful characteristics of knowledge are and how they should be studied (Newell et al., 2002).

Commodity View

Advocates of the commodity or possession perspective stress two distinct kinds of knowledge: tacit and explicit. This distinction was followed with many other terms such as local versus universal, codified versus noncodified, procedural versus declarative, and know-how versus know-what to describe the different characteristics of distinct types of organizational knowledge (Orlikowski, 2002). Perhaps the most significant piece of work from the commodity perspective of organizational knowledge is the influential book *The Knowledge-Creating Company* by Nonaka and Takeuchi (1995). They suggested knowledge creation and management is an iterative process in which tacit knowledge and explicit knowledge are converted interchangeably in a continuous cycle through the four modes of socialization, externalization, combination, and internalization. Socialization involves the implicit sharing of tacit knowledge, externalization involves the transformation of tacit to explicit, combination involves conveying explicit knowledge from one person to another to make it a collective 'stock' of knowledge, and internalization involves the conversion of explicit knowledge back to its tacit form.

Knowledge management theorists apply the distinctions of tacit versus explicit, and similar taxonomies such as procedural versus declarative to focus on knowledge as a commodity that can be documented, stored, repackaged, and distributed. Research from this perspective has resulted in many important and useful theoretical contributions on how crucial organizational knowledge can be codified and transmitted in a formal, systematic way to influence or trigger certain organizational outcomes. Knowledge has commodity-like characteristics and can be accumulated, transferred, and stored independently of an individual (Buckley and Carter, 1999). In other words, knowledge is regarded as "an objective entity … as a thing out there that can be transmitted or exchanged without any intervention by a cognizing mind" (Ringberg and Reihlen, 2008: 914). An important emphasis from this perspective is developing processes and organizational mechanisms that enable the transformation of tacit understanding into explicit knowledge by encouraging individuals to put forth the required degree of effort.

The Community Perspective

The community view of knowledge management research is somewhat critical of a purely commodity or possession perspective. Scholars have argued that knowledge is not an 'object' that can be stockpiled or stored, but more of active phenomenon that cannot be separated from the very act of knowing something (Brown and Duguid, 1991, 2001; Cook and Brown, 1999; Orlikowski, 2002; Pfeffer and Sutton, 1999; Tsoukas, 2003). Central to the community view is Lave and Wenger's (1991) communities-of-practice learning theory which attempts to shift the focus to exploring the links between knowledge and the actions of organizations and the individuals within them. To communities-of-practice researchers, the notion that knowledge can be explicit and quantifiable draws a problematic distinction between knowledge as a tangible good and the use of that good in practice (Pfeffer and Sutton, 1999). Community-of-practice research argues that it is often impossible, or at least unproductive, to attempt to separate knowledge from how individuals act on what they know.

The difference between the two research streams is made clear by understanding the distinction between *knowledge* implying something that can be located and is independent of both the individual who has the knowledge and the context in which understanding takes place, and *knowing* which implies a process or action of knowers (Pfeffer and Sutton, 1999). A community-of-practice outlook explores the actions that people undertake while solving problems during their day-to-day interactions in the real world (Brown and Duguid, 2001; Cook and Brown, 1999; Orlikowski, 2002). In this research stream, organizational knowledge is seen through the lens of its manifestations in the activities that make tacit knowledge a part of every organizational action (Nag et al., 2007). Specifically, distinctions between explicit and tacit knowledge are not sufficient to understand all somebody knows; it is also necessary to understand the concept of knowing through action or experience-based know-how (Cook and Brown, 1999). Brown and Duguid (2001), and others (Blackler, 1993; Spender, 1996; Tsoukas, 1996) argue that a knowing perspective more accurately reflects Polanyi's conclusion that all knowledge has an inarticulate component (Brown and Duguid, 1991; Cook and Brown, 1999; Nag et al., 2007; Orlikowski, 2002). Community researchers argue that Polanyi did not write about two different types of knowledge

or describe two ends of a continuum. Instead, they emphasize Polanyi's proclamation that tacit knowledge is a vital component of all knowledge; in other words, all knowledge is tacit or at a minimum incorporates some elements of tacit knowledge (Botez, 1998). From this perspective, what matters most is the proportion of embedded or experience-based know-how reflected in what is known.

Knowledge-in-Practice

More recently within the knowledge management literature researchers have attempted to integrate the two disparate views of knowledge while trying to understand the critical characteristics of knowledge from a knowledge-in-practice (KIP) perspective (McIver and Lengnick-Hall, 2011; McIver et al., 2012, 2013, 2010). Building on the community-of-practice literature, McIver and colleagues focused on the practice aspects of understanding and using knowledge. A practice is an activity which is sustained and reproduced over time and is critical to the analysis of knowledge (Duguid, 2005; Lave and Wenger, 1991). According to Cook and Brown (1999), *practice* implies doing and refers to activities of organizational members informed by meanings. In other words, in knowledge management research practice refers to *the way in which work gets done and knowing how to do it* (Brown and Duguid, 2001: 200).

McIver et al. (2010, 2012, 2013) define KIP as the information and know-how involved in the sequences, routines, capabilities, or activity systems in an organization. They define information as the facts and data that can be understood, stored, and transferred. Information reflects the commodity perspective of knowledge. Know-how is defined as the skills and expertise that are action-based and reflect the 'knowing' view of knowledge. However, since KIP requires both information and know-how, it invariably incorporates the important tacit dimension or tacit knowledge. A KIP perspective underscores how all knowledge, or KIP, has a tacit component. While the conventional wisdom within the management literature has typically been that tacit knowledge is synonymous with difficult to learn, difficult to transfer, and hard to understand, the accuracy of this assumption is now being questioned. It seems that focusing solely on distinctions between tacit and explicit knowledge can lead to misrepresentation of the phenomenon.

Drawing from communities-of-practice research (McIver et al., 2013) added the dimension of learnability to help clarify the understanding of knowledge in organizations. They defined the tacitness of knowledge as *the degree to which the know-how involved in an organizational practice is unobservable, difficult to teach, unspecifiable, and/or highly embedded in the setting* (McIver et al., 2013). This definition distinguishes difficult to learn from difficult to teach, and separates the concept of hard to understand from challenging to articulate or specify and explicitly recognizes the embedded or situated nature of tacit knowledge. In addition, by defining tacitness as a characteristic of knowledge that varies in degree, this definition incorporates Polanyi's actual premise that all knowledge contains some measure of tacit understanding. This definition suggests that what varies is the proportion and role of tacit knowledge within different activities or practices within organizations. This view contrasts with perspectives that specify tacit and explicit knowledge as either two separate and independent dimensions or as opposite ends of a continuum and argue that tacit can be converted to explicit knowledge. The McIver et al. (2013) definition shows that all practices have a tacit component.

McIver et al. (2013) explain that there is a second essential dimension of KIP that is generally neglected: learnability. They define learnability as *the type and amount of effort, study, accumulated comprehension, and expertise that is involved in understanding the information and know-how for accomplishing work practices*. Learnability captures the ease (or difficulty) which new or occasional workers experience trying to perform certain work practices. Learnability distinguishes between KIP that is difficult to learn and KIP that is relatively easy to learn regardless of the degree of tacitness. Incorporating learnability into an understanding of tacit knowledge highlights two overlooked conditions – KIP with a high degree of tacitness that is easy to learn and KIP with a low level of tacitness that is quite difficult to learn. This distinction addresses the implicit, but arguably inaccurate, assumption noted previously that all tacit knowledge is difficult to learn. If all knowledge when used in the act of knowing has a tacit component, there must be a way to distinguish between knowledge or activities as practices that are difficult to grasp and knowledge in organizations that is not as difficult to master. Introducing the concept of learnability underscores important distinctions between learning and teaching, codification or observation and experience as the prerequisite for mastery, and seeing knowledge as independent rather than contextually embedded.

Interestingly the idea of knowledge learnability as well as the process of learning may change as the degree of tacitness in an activity or practice increases. Reber (1993) examined the concept of implicit learning in his work as the process whereby knowledge is acquired largely without the involvement of top–down, conscious control. He provided examples of language acquisition and the process of socialization to illustrate the operation of implicit learning. Reber's view of the socialization process is very close to the work done in communities-of-practice research (Lave and Wenger, 1991). Reber (1993) suggested implicit learning results in tacit knowledge and indicated that not only is the learner unaware of the process of learning, the very knowledge itself is highly resistant to explication. As Reber discussed, the notion that deep, abstract knowledge can be acquired and represented outside of consciousness is close to Polanyi's original use of the term. A particularly important observation, however, is that such deep, abstract knowledge is possible to acquire through action.

The Next Iteration

Current efforts to understand the dynamics of knowledge in organizations and to gain an enhanced appreciation for what knowledge can trigger and how knowledge can be used in organizations may require a step back to help us understand and rethink what knowledge is and what the important characteristics of knowledge are within practices or activities in

organizations. This presents an exciting extension of knowledge-based views of organizations and the potential contribution of tacit knowledge to achieving desirable organizational outcomes. A realization that all knowledge has a tacit component and that learning and the learnability of that knowledge is its own unique phenomena, only begins to scratch the surface of what must be understood.

What is missing are the links between Polanyi's original work and what psychology scholars examine as knowledge that is intensely personal, introspectively opaque, and communicatively unresponsive and is acquired and used with little in the way of top–down control of consciousness. Also missing are links between the seeing knowledge as personal to an individual and the tacit elements of KIP as it is studied in organizations. How can the connections between something that is identified as so important to the success of organizations, but yet so personal in nature to an individual be studied effectively? How is individual knowledge applied and transformed in organizations and how does it aggregate from individual knowledge to collective knowledge? How can the contradictory perspectives and assumptions about what tacit knowledge is and how it influences behavior best be reconciled? Research streams examining such issues as organizational routines (Feldman and Pentland, 2003; Nelson and Winter 1982), organizational learning (Crossan and Berdrow, 2003; Crossan et al., 1999), human capital (Coff, 1997; Lepak and Snell, 1999), social capital (Levin and Cross, 2004), and relational coordination (Gittell, 2001) may offer useful insights for both capitalizing on tacit individual knowledge in organizations and for developing mechanisms for generating collective tacit understanding.

Conclusion

Virtually no one in the social sciences and specifically management research doubts the importance of knowledge to society or the existence of and the importance of a tacit dimension. Yet, little agreement exists on the mechanisms that underlie tacit knowledge, or on how to best conceptualize, use, extend, and leverage this resource. Efforts to better understand how tacit knowledge shapes what employees need to be able to do on the job, what they need to know to be able to accomplish desired activities, and how tacit knowledge contributes to strategic success suggest a number of productive research agendas. While, as Reber noted in his initial article, the future looks interesting, progress will undoubtedly require cross-disciplinary teams of researchers to sort through the complex and challenging domain of tacit knowledge.

See also: Information and Knowledge, Organizational; Learning: Organizational; Organizational Decision Making; Polanyi, Karl (1886–1964).

Bibliography

Becerra, M., Lunnan, R., Huemer, L., 2008. Trustworthiness, risk, and the transfer of tacit and explicit knowledge between alliance partners. Journal of Management Studies 45, 691–713.

Blackler, F., 1993. Knowledge and the theory of organizations – organizations as activity systems and the reframing of management. Journal of Management Studies 30, 863–884.

Botez, A., 1998. Michael Polanyi and Lucian Blaga as philosophers of knowledge. Paper Presented at the Twentieth World Congress of Philosophy. Boston, MA.

Brown, J.S., Duguid, P., 1991. Organizational learning and communities of practice: toward a unified view of working, learning and innovation. Organization Science 2, 40–57.

Brown, J.S., Duguid, P., 2001. Knowledge and organization: a social perspective. Organization Science 12, 198–213.

Buckley, P.J., Carter, M.J., 1999. Managing cross-border complementary knowledge. International Studies of Management and Organization 29, 80–98.

Coff, R.W., 1997. Human assets and management dilemmas: coping with hazards on the road to resource-based theory. Academy of Management Review 22 (2), 374–402.

Coff, R.W., Coff, D.C., Eastvold, R., 2006. The knowledge-leveraging paradox: how to achieve scale without making knowledge imitable. Academy of Management Review 31 (2), 452–465.

Cook, S.D.N., Brown, J.S., 1999. Bridging epistemologies: the generative dance between organizational knowledge and organizational knowing. Organization Science 10, 381–400.

Crossan, M.M., Berdrow, I., 2003. Organizational learning and strategic renewal. Strategic Management Journal 24, 1087–1105.

Crossan, M.M., Lane, H.W., White, R.E., 1999. An organizational learning framework: from intuition to institution. Academy of Management Review 24 (3), 522–537.

Dierickx, I., Cool, K., 1989. Asset stock accumulation and sustainability of competitive advantage. Management Science 35, 1504–1511.

Duguid, P., 2005. "The art of knowing": social and tacit dimensions of knowledge and the limits of community of practice. The Information Society 21, 109–118.

Eisenhardt, K.M., Santos, F.M., 2002. Knowledge-based view: a new theory of strategy? In: Pettigrew, A., Thomas, H., Whittington, R. (Eds.), Handbook of Strategy and Management. Sage, Thousand Oaks, CA, pp. 139–164.

Feldman, M.S., Pentland, B.T., 2003. Reconceptualizing organization routines as a source of flexibility and change. Administrative Science Quarterly 48, 91–118.

Gittell, J.H., 2001. Supervisory span, relational coordination, and flight departure performance: a reassessment of postbureaucracy theory. Organization Science 12 (4), 468–483.

Grant, R.M., 1996. Toward a knowledge-based theory of the firm. Strategic Management Journal 17, 109–122.

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3, 383–397.

Kogut, B., Zander, U., 1996. What firms do? Coordination, identity, and learning. Organization Science 7, 502–523.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, New York.

Lepak, D.P., Snell, S.A., 1999. The human resource architecture: toward a theory of human capital allocation and development. Academy of Management Review 24 (1), 31–48.

Levin, D.Z., Cross, R., 2004. The strength of weak ties you can trust: the mediating role of trust in effective knowledge transfer. Management Science 50 (11), 1477–1490.

McIver, D., Lengnick-Hall, C.A., 2011. Knowledge management processes and unit performance: a contingency perspective. Academy of Management Best Paper Proceedings.

McIver, D., Lengnick-Hall, C.A., Lengnick-Hall, M.L., Ramachandran, I., 2012. Integrating knowledge and knowing: a framework for understanding knowledge-in-practice. Human Resource Management Review 22, 86–99.

McIver, D., Lengnick-Hall, C.A., Lengnick-Hall, M.L., Ramachandran, I., 2013. Understanding work and knowledge management from a knowledge-in-practice perspective. Academy of Management Review 38, 597–620.

McIver, D., Ramachandran, I., Lengnick-Hall, C.A., Lengnick-Hall, M.L., 2010. Rethinking the nature of knowledge management research: implications for strategic human resource management. Academy of Management Best Paper Proceedings.

Nag, R., Corley, K.G., Gioia, D.A., 2007. The intersection of organizational identity, knowledge, and practice: attempting strategic change via knowledge grafting. Organization Science 5, 821–847.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Belknap Press, Cambridge, MA.

Newell, S., Roberts, J., Scarbrough, H., Swan, J., 2002. Managing Knowledge Work. Palgrave, New York.

Nonaka, I., Takeuchi, H., 1995. The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York.

Orlikowski, W.J., 2002. Knowing in practice: enacting a collective capability in distributed organizing. Organization Science 13, 249–273.

Pfeffer, J., Sutton, R.I., 1999. Knowing "what" to do is not enough: turning knowledge into action. California Management Review 42, 83–108.
Polanyi, M., 1958. Personal Knowledge: Towards a Post Critical Philosophy. Routledge, London.
Polanyi, M., 1966. The Tacit Dimension. Routledge & Kegan Paul, London.
Reber, A.S., 1992. The cognitive unconscious: an evolutionary perspective. Consciousness and Cognition 1, 93–133.
Reber, A.S., 1993. Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious. Oxford University Press, New York.
Ringberg, T., Reihlen, M., 2008. Towards a socio-cognitive approach to knowledge transfer. Journal of Management Studies 45, 912–935.
Spender, J.C., 1996. Making knowledge the basis of a dynamic theory of the firm. Strategic Management Journal 17, 45–62.
Tsoukas, H., 1996. The firm as a distributed knowledge system: a constructionist approach. Strategic Management Journal 17, 11–25.
Tsoukas, H., 2003. Do we really understand tacit knowledge? In: Easterby-Smith, M., Lyles, M.A. (Eds.), Blackwell Handbook of Organizational Learning and Knowledge Management. Blackwell Publishing, Ltd, Malden, MA, pp. 410–427.

Tactile Maps in Geography

Regina Araujo de Almeida (Vasconcellos), University of São Paulo, São Paulo, Brazil

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Maps are graphic representations of space, and they are called tactile maps when in a format to be sensed by touch. Types of tactile maps are presented, and issues related to their design, production, and use are analyzed; the role of tactile cartography in geography is also discussed. The article highlights the relevance of new digital resources and assistive technologies that introduced great changes and possibilities combining tactile images, text, and sounds. Further development toward an inclusive multisensory and multimodal cartography direct to all persons is proposed, expanding the field of tactile maps.

Introduction

Maps are essential means to collect, process, represent, and communicate geographical data. For this reason, they should be available to all persons, including ones with visual impairment. Blind people have the right to access and use graphic representations and maps in a way that images can be 'seen' also by other senses, besides vision. Among the world's 700 million people at the end of 2012, 100 million live with some form of disability, and in this group, 285 million people have visual impairment (WHO, 2012a,b). These individuals need special attention within cartography and with images in general.

Maps are graphic representations of space, and they are called tactile maps when in a format to be sensed by touch. Tactile maps have been made since ancient times, as stones set in sand. Since then, maps with raised symbols have been constructed using a wide variety of materials such as embossed paper, molded plastic, and silk-screen printed fabric. Embossed signs and Braille text have been used for 150 years in maps. Many improvements have recently occurred in the area of tactile map production and use. New computer technologies have introduced great changes and possibilities combining tactile images, text, and sounds.

Tactile cartography helps to overcome informational barriers for those who cannot see, facilitating their access to a world dominated by images. Tactile maps can also help blind users to navigate through indoor and outdoor environments and to improve their geographical knowledge. All kinds of maps and graphics should become available in tactual form, preferably with sounds and visual graphic language to really promote graphic literacy and integration of the visually impaired. Types of tactile maps are presented in this article, and the various techniques used to create them are analyzed. The design, production, dissemination, and use of tactile maps are also discussed, including new computer technologies and current research. Future research and actions should move from tactile cartography to a broader field of inclusive cartography with maps reaching all people.

Tactile Map Concept, Types, and Relevance in Geography

Maps are graphic representations of space, and as abstractions of reality, they belong to the world of images. Most maps are not available for the visually impaired person and they need to be translated to a tactual graphic language using raised symbols. Maps are called tactile maps when they are in a format that lets them to be 'seen' by touch. People acquire and develop the concept of space mainly through the eye. Visual perception is fundamental for understanding the environment. People deprived of vision, either totally or partially, depend on the other senses to gather information on the real world and its geography. Blind people learn about geographical space through direct exploration, verbal explanations, sounds, smell, gestures, models, and tactile images.

Tactile maps have raised symbols on a flat surface; they differ from models, which are three-dimensional (3D), although both are perceptible by touch. Tactual perception works with movement (kinesthetic force), feeling raised images, surfaces, and objects by moving the fingers or other body parts. This is called haptic perception and it allows visually impaired persons to explore the tangible world. Besides tactile, other terms such as tangible and raised relief maps are currently applied to the same concept; haptic has been used in relation to computer mapping and visual impairment, although the word itself precedes the era of advanced technologies.

Tactile maps have been used since Egyptian times where they first appeared as stones set in sand. Since then, many materials have been used, including wood, plaster, and even embroidery. With the invention of Braille in the mid-nineteenth century, it became possible to describe map features using a font that blind people could understand. Embossed images and Braille text have been used for 150 years in maps. Recent technology has changed tactile maps and the entire cartographic communication process, from its creation to the use of images and graphic representations. Maps are essential to navigate, to develop space perception, and to gather geospatial data and knowledge. There are two main groups of tactile maps, based on their purpose, scale, and features: orientation and mobility maps, and geographical maps.

Orientation and Mobility Maps

These maps are directed only to blind and severe low-vision people who need to be aware of the physical environment in order to navigate in space. An orientation map gives an

overview of the area. Mobility maps, as the name suggests, provide the kind of detailed information that is needed if the person is to move safely through the environment. Mobility maps show obstacles to be avoided as well as elements to use such as sidewalks or elevators. They depict indoor and outdoor environments using large-scale representations of buildings, routes, streets, transportation networks, neighborhoods, and recreation areas. Verbal maps, also called sequential maps, can replace tactile mobility maps to explain routes to blind people. Often they are used with location devices such as global positioning system (GPS), but can be combined with audio–tactile maps with very good results for navigating in space.

Geographical Maps

Geographical maps represent larger areas in much smaller scale than mobility maps. General reference maps include political and physical tactile maps, ranging from cities, states, and countries to the whole globe. Tactile topographical maps are relief models, also produced in colored plastic for the sighted user, which need adaptation for the blind by having all symbols in raised format.

Thematic maps portray qualitative and quantitative geographical information in tactile form. They show the location, pattern, and structure of geospatial data on a wide variety of themes such as population, industry, climate, and distribution of vegetation, at all scales from local and regional to global. These types of maps are important teaching resources at schools and they should be directed to all students, not only those with visual impairment. At this time of innovative digital technologies, old-fashioned tactile maps still have an important role inside the classroom. Tactile maps can be used with 3D models, with or without sounds, as motivating and interesting strategies for geography teaching.

Tactile Map Creation and Use

Much work on tactile maps has been done in the past 30 years. The International Cartographic Association (ICA) has a Commission on Maps and Graphics for Blind and Partially Sighted People (see ICA Web site, also the Conferences Proceedings). Several authors have contributed significantly to the field of tactile cartography and results have been published extensively over time (Almeida and Tsuji, 2005; Barth, 1987; Edman, 1992; Heller, 1991; Jacobson, 2007; James, 1982; Kidwell and Greer, 1973; Perkins and Gardiner, 2003; Rowell and Ungar, 2003; Schiff and Foulke, 1982; Tatham and Dodds, 1988; Vasconcellos, 1996; Wiedel, 1983; Wiedel and Groves, 1972). Koch (2012) summarizes the outcomes of this research and presents an overview of tactile cartography, from design to new digital technologies, highlighting its chronological development.

Some concepts about visual impairment and tactual perception are important to map and graphics creation and use. Whether the person acquired visual impairment or was born blind, also type and degree of visual impairment may affect the way these people read images and maps. Information and publications about map design, production, and use can be found in the suggested websites and references, including issues like scale, standardization, degree of generalization, and choice of symbols.

Tactile Map Design

The design of a tactile map involves a cartographic language with tactile symbols such as raised lines, points, and textures. Visual variables studied by J. Bertin's graphic semiology (Almeida, 2011) must be translated into tactual form to construct tactile maps. The nature of geographical data, either quantitative or qualitative, has to be considered when selecting the correct variable, whether tactile, sonic or visual, to represent the particular data element. Graphic variables directed to vision and touch (e.g., size, value, texture, color, orientation, and shape) can be expressed in point, linear, and area modes. Color is not a part of the tactual graphic language when blind persons are involved, but elevation (height) can be used alone or combined with other variables in multiple ways to represent features in addition to height of the terrain.

Clarity and simplicity are important qualities in tactile mapping, as is the use of a map legend or key. A collection of maps or a multilayered tactile map can help avoid complexity and excessive clutter, making tactual perception easier. Tactile maps when designed for low-vision users should combine visual and tactile symbols, preferably in color, but with limited variation of hues and tones. Many studies have been done on the legibility and discriminative capability of symbols from the perspective of the visually impaired as map users. In general, choice of symbols, scale, and degree of generalization greatly depend on the production method to be used, including master construction and reproduction techniques. Concepts such as generalization and standardization can facilitate design and use, but they are not easy to deal with and are still a challenge for geographers and cartographers. Orientation and mobility maps are more likely to have a system of conventions established at the international level.

The author, in previous work (Vasconcellos, 1996: 99–100) discussed tactile map production and use, and presented a set of guidelines directed to design, reproduction, map use, and education. Several authors, mentioned before, contributed greatly to tactile map design, such as Barth, Edman, Kidwell, and Greer, Tatham. Rowell and Ungar (2003) presented the results of an interesting survey that collected data on production methods, techniques, outputs, and preferences related to tactile mapping design, such as the choice of map size and format, and choice of symbols. Perkins and Gardiner (2003) also contributed to tactile map design and use with several publications; they presented a manual with best practices and guidelines for the design, production, and presentation of vacuum-formed tactile maps, which can be seen at the Perkins School Web site.

Production Methods and Materials

Many techniques and materials are used in the construction and reproduction of conventional tactile maps. The advantages and

limitations of these production methods might include, for example, cost, available technology, number of copies needed, or the user's ability to read maps. The most common methods are

1. hand-made collage maps using a large variety of inexpensive materials, usually prepared by nonspecialists;
2. molded plastic maps done in vacuum-forming machines (e.g., Thermoform with Brailon plastic), and made from a master map prepared in a variety of materials such as embossed aluminum, collage, paper, or plaster carved manually or driven by computers;
3. embossed paper maps done by Braille printing machines, using conventional or digital methods;
4. silk-screen maps using puff ink, usually printed on paper; and
5. microcapsule paper maps compatible with conventional printers and raised by heat, largely used with digital maps and computers.

Collage has very good discriminatory properties because of its varied textures, low precision, and low-cost reproduction capability. On the other hand, silk-screen and microcapsule paper maps have lower discriminatory properties, but high reproduction capabilities and they are more pleasant for touching. Besides these more common techniques, there are others such as German film, virktyping, or thermocraft. Each has its advantages and disadvantages depending on a variety of factors including available technologies and user preferences.

Tactile Map Use: The Need for Education and Training

There are some rules to be followed regarding tactile map use. In general, more training is necessary for all persons involved in tactile mapping, both producers and users, than is the case for visual maps. In addition, many tactile map creators are not geographers or cartographers and therefore need even more education and training when responsible for mapping tasks.

Before using maps, it is advisable to teach basic geographical concepts such as scale, point of view, relative positioning, orientation, location, and distance. Tactual graphic language has to be introduced to the users before tactile map reading exercises that help them to understand physical space and its representation. For this purpose, a program to teach this language for children was designed, with six phases, each one presenting relevant geographical concepts: (1) the tactile graphic language (memory game with tactile graphic variables as proposed by Graphic Semiology); (2) concept of scale, reduction, and distance (activity using dolls and cars in several sizes, 3D models, and a classroom map); (3) point of view (graphic representation of objects in vertical and horizontal views, avoiding perspective that is too abstract for blind persons); (4) location and orientation (a game called Geographical Battle with coordinates and hemispheres); (5) map reading (the complete address of the school, from classroom to the universe); (6) map legend and the combination of all previous five steps (exercise of constructing a city using tactile/visual variables in a mat with sticky pieces). Finally, different tactile maps could be read by users. This program was tested with hundreds of students and proved to be very effective with blind and low-vision children and teachers (Almeida, 2011).

After research and evaluations with 200 visually impaired children and adults, and 80 school teachers, the author proposed a methodology and a set of guidelines as presented before (Vasconcellos, 1996). Just some ideas for map use are listed below:

1. it is essential to introduce geographical concepts and cartographic language prior to map use;
2. it is advisable to create and use conventions as much as possible to facilitate graphic information reading;
3. the map key is very important for the visually impaired; they are experts in decoding a legend and it should be used even for drawings;
4. proper reduction and generalization are vital; also size and distance between signs are important because tactual perception has a different resolution and is not global, as the blind user must put together pieces of information to form the image; and
5. enough relief and contrast together with the use of redundancy help in achieving better map design with the needed clarity. Accuracy and precision very often have a secondary role. Sometimes the solution is to make a collection of maps, facilitating map reading tasks.

It is important to consider the user's degree of sensory impairment, previous experience, knowledge about the environment, and cartographic skills. Tactual sense is coarser and it works by bringing pieces together to form a map. It is different from sight, which is synthetic with high resolution. Despite some limitations, blind people can have an excellent ability to decode map keys by using both hands simultaneously, one on the legend and the other on the map. In most cases, the visually impaired user needs personal assistance or good instructions to understand the message conveyed in a tactile map. Successful cartographic communication depends on innovative approaches and efficient training since many of the problems this group may encounter in map reading differ from those identified for sighted users.

Audio–Visual–Tactile Maps and the Multisensory Approach

Technology has changed the field of cartography, opening amazing new doors and possibilities in tactile mapping. Sound, touch and vision can be easily put together to depict the physical and social environment; current technologies can already add smell, and research will find ways to include taste in the future. Visualization techniques are improving and expanding to virtual reality and 3D images to map geographical space; also remote sensing techniques now collect, process, save, represent, and update information (Almeida and Tsuji, 2005). What was just tactile cartography aimed at map users who were blind and visually impaired can now take advantage of all five senses and many innovations in telecommunications and computer technology, as proposed by the new cybercartography paradigm (Taylor, 2003, 2005, 2013).

Assistive Technologies: Devices and Techniques in Digital Cartography

Information and communication technologies have transformed the creation and use of tactile maps. People with visual impairment, being blind or low vision in any degree, will need different devices (equipment, products, and software) to translate visual images to other senses, such as touch and hearing; these devices are called assistive technology. The distinction between computer and no-computer aids and products relates to the user's visual impairment in terms of sensory replacement, augmentation, or substitution. The Perkins School has very good information on these topics. Almost all sensory substitute technologies are now computer related, although several have their origins in predigital technology, such as tactile graphics, Braille displays, voice recorders, and video cameras (see Perkins Web site). Tactile cartography must take advantage of a large variety of assistive technologies, including conventional ones such as a Braille typewriter and computer products. High-tech devices can make a real difference in tactile cartography; examples include computer screen readers, voice synthesizers, a variety of software and hardware, adapted keyboards and mouses, special glasses and cameras, GPS, mobile phones, and so on. All these assistive computer technologies are vital for persons with visual impairment, helping them to 'see' images and graphics using tactile and sound resources. Accessibility to such technologies is still a problem, however, in that about 80% of the world's visually impaired population live in developing countries and might have no access to Information and Communications Technologies (ITCs) and products adapted to persons with visual impairment. At present, these new forms of tactile maps, using multisensory and multimodal interactive formats, can reach only a minority of users.

Innovative Methods and Products in Tactile Mapping

During the last three decades, many technologies were developed to support people with visual impairment to 'see' maps through touch and sound. A tactile–audio system called NOMAD was invented by Professor Don Parkes, in Australia, in 1986. This system combined touch and sound so that a tactile map, when touched, spoke the information under the finger; environmental sounds and music could also be added. NOMAD and all the Tactile Graphics Designer (TGD) components that followed the first version of this audio–tactile device had functions allowing location, navigation, distance, scale, and a variety of different features presented in tactile format, graphics and maps combined with sounds. The audio–tactile system had interactive capability and up to three layers of information could be embedded in every information point. Effectively, an entire encyclopedia could be embedded, but computer storage and speeds were far from what is available today. Maps, spatial guidance systems, a capacity to paint in sound with frequencies changing at user defined boundaries of any shape made a virtual 3D audio–tactile surface possible; to read Braille and interact with voice recognition to questions by users were among its many features. It was translated into Portuguese, Spanish, Japanese, French, Italian, German, Swedish, Malay, and Dutch.

From the original NOMAD system, the TGD components of AudioPIX, Audio CAD, and Audio TRIP were developed and presented in journals and scientific conferences (Parkes, 1995, 1998). In 2004, the system was passed to Duxbury Systems for their use of any of the software that suited their purposes in their established market for Braille translation systems and embossed dot graphics.

A few years ago, a Talking Tactile Tablet, TTT3, with the same functions of the previous NOMAD, came into the market, manufactured by Touch Graphics. For navigation, BrailleNote GPS and various Adapted GPS devices make independent travel possible for blind people. For computer use by a blind person a screen reader that shows images on the monitor through sounds and speech are essential as is a Braille display. Texts are easily translated to sound using any voice synthesizer software combined with the screen reader, and BrailleTouch allows people to use touch screen mobile phones (typing Braille). The American Foundation for the Blind has an extensive list of products and software, including all sorts of assistive technology that helps the visually impaired person to access computers, multimedia interactive maps, and images. Very interesting are the force feedback joystick and mouse, but in terms of products and software, problems with discontinuity and market restrictions inhibit production of those items at a scale that is affordable for all users.

Several specialists have dedicated time and efforts to develop and study audio–tactile and visual mapping devices and software directed to people with visual impairment. As a blind person and also a geographer, Reginald Golledge was one of the first to experiment navigating with GPS and a laptop; he has done extensive work on this area (Golledge et al., 2005). Jacobson (2007) discussed the future of tactile cartography and presented a comprehensive list of all tactile map formats from conventional static raised line to multimodal computer interfaces, highlighting digital methods related to audio–tactile, multimedia, map-based approaches.

Rowell (2007) conducted an interesting survey that compared conventional tactile maps with new wayfinding geotechnology, asking 30 visually impaired persons to evaluate various wayfinding methods (tactile maps, new mobile technologies, audio descriptions, and fixed-electronic navigation aids). The sample individuals reported that tactile maps and GPS devices were the most useful navigation aids for blind people; several comments were made about advantages and problems, discussing the potential of these new technologies. Multisensory interactive maps do work and can change the way visually impaired persons relate to cartography, as argued before (Almeida and Tsuji, 2005).

There are many recent developments in this field where computers, telecommunications, geographical position systems, mobile phones, and tablets connected to the Internet can work together, facilitating the production of multisensory maps combining sounds, images, and tactile graphics.

Future Directions: Barriers and Opportunities in a Digital World

New technologies have introduced great changes and improvements in the area of tactile map production and use.

Tactile maps will evolve to multimodal and multisensory maps, in an interactive multimedia format as proposed by cybercartography (Taylor, 2003). The challenge for cartography will be to provide these maps and innovative resources, as well as to make them available and accessible to all users, including those with visual impairment. Conventional and old-fashioned tactile maps with Braille text will be produced and used for a while yet, but new technologies will create a totally new area in mapping for people with visual impairment, bringing new possibilities for both tactile map creators and users.

In any case, it is important to prepare these new map users to deal with cartography and audio–tactile visual resources. On the other hand, there is a need to produce and disseminate tactile graphics, especially maps, not only for the blind, but also as teaching materials for all students and schools. Education and training in tactile cartography still needs attention and further action. Tactile cartography helps to overcome informational barriers for those who cannot see, facilitating their access to a world dominated by images. Therefore, all kinds of maps and graphics should be available in tactile and interactive multisensory forms with sound, promoting graphic literacy and integration for all visually impaired people, regardless of location or social class.

Future research should move from tactile cartography to a broader field of inclusive cartography with maps reaching more people, so that multimodal maps are accessible everywhere and by everyone. From an old thermoformed or collage map to a virtual map in mobile devices using location-based services or cloud computing, all maps are relevant and they are a unique means to achieve space perception and knowledge, to communicate spatial information, to navigate, and to learn geography.

See also: Cartographic Visualization; Cartography; Spatial Cognition.

Bibliography

Almeida (Vasconcellos), R., 2011. Bertin graphic semiology and its relevant contributions for research and teaching in Brazil. In: Proceedings, 25th International Cartographic Conference, Paris, France.

Almeida (Vasconcellos), R., 2013. Inclusive cartography: theoretical and applied issues in Brazil (chapter 8). In: Taylor, D. R. Fraser, Lauriault, Tracey P. (Eds.), Further Developments in the Theory and Practice of Cybercartography. Elsevier, Boston, MA.

Almeida (Vasconcellos), Regina, Tsuji, Bruce, 2005. Interactive mapping for people who are blind or visually impaired. In: Taylor, D.R.F. (Ed.), Cybercartography: Theory and Practice. Elsevier, pp. 411–431.

Barth, J.L., 1987. Tactile Graphics Guidebook. American Printing House for the Blind, Louisville, KY.

Edman, P.K., 1992. Tactile Graphics. American Foundation for the Blind, New York.

Golledge, R.G., Rice, M., Jacobson, R.D., 2005. A commentary on the use of touch for accessing on-screen spatial representations: the process of experiencing haptic maps and graphics. The Professional Geographer 57 (3), 339–349.

Heller, M.A., 1991. Haptic perception in blind people. In: Heller, M.A., Schiff, W. (Eds.), The Psychology of Touch. Erlbaum Associates, Lawrence, Hillsdale, NJ, pp. 239–261.

Ishido, Y. (Ed.), 1989. Proceedings of the Third International Symposium on Maps and Graphics for the Visually Handicapped People. Yokohama. Yokohama Convention Bureau, Japan.

Jacobson, D., 2007. The future of tactile cartography: from static raised lines to multimodal dynamic portable computer interfaces. In: Proceedings, XXIII International Cartographic Conference, Moscow, Russia, p. 358.

James, G.A., 1982. Mobility maps. In: Schiff, W., Foulke, E. (Eds.), Tactual Perception: A Source Book. Cambridge University Press, New York.

Kidwell, A.M., Greer, P.S., 1973. Sites Perception and the Nonvisual Experience: Designing and Manufacturing Mobility Maps. American Foundation for the Blind, New York.

Koch, Wolf G., 2012. State of the art of tactile maps for visually impaired people. In: Buchroithner, Manfred (Ed.), True-3D in Cartography – Autostereoscopic and Solid Visualization of Geodata. Springer-Verlag, Germany, pp. 137–151.

Parkes, Don N., 1995. Access to complex environments for blind people: multimedia maps, plans and virtual travel. In: Proceedings of the 17th International Conference, International Cartographic Association, vol. 2. Barcelona, Spain, pp. 2449–2460.

Parkes, D., 1998. Tactile audio tools for graphicacy and mobility: a circle is either a circle or it is not a circle. The British Journal of Visual Impairment 16 (3), 99–104.

Perkins, C., Gardiner, A., 2003. Real world map reading strategies. The Cartographic Journal 40 (3), 265–268.

Rowell, J., 2007. The end of tactile mapping or a new beginning: LBS for visually impaired people. In: Proceedings, XXIII International Cartographic Conference, 4–10 August, Moscow 2007, Russia, p. 363.

Rowell, J., Ungar, S., 2003. Feeling your way: a tactile map user survey. In: Proceedings, XXI International Cartographic Conference – ICA, Durban, SA, August 2003, pp. 652–659. http://icaci.org/files/documents/ICC_proceedings/ICC2003/Papers/078.pdf (accessed 27.01.13.).

Schiff, W., Foulke, E. (Eds.), 1982. Tactual Perception: A Source Book. Cambridge University Press, New York.

Tatham, A.F., 2003. Tactile mapping: yesterday, today and tomorrow. The Cartographic Journal 40 (3), 255–258.

Tatham, A.F., Dodds, A.G. (Eds.), 1988. Proceedings of the Second International Symposium on Maps and Graphics for Visually Handicapped People. University of Nottingham, King's College, London.

Taylor, D.R.F., 2003. The concept of cybercartography. In: Peterson, M. (Ed.), Maps and the Internet. Elsevier, Amsterdam, NL.

Taylor, D.R., Fraser, Lauriault, Tracey, P. (Eds.), 2013. Further Developments in the Theory and Practice of Cybercartography. Elsevier, Boston, MA.

Taylor, D.R.F. (Ed.), 2005. Cybercartography: Theory and Practice. Elsevier, Amsterdam, p. 574.

Vasconcellos, R., 1996. Tactile mapping design and the visually impaired user. In: Wood, C., Keller, P. (Eds.), Cartographic Design: Theoretical and Practical Perspectives. John Wiley, London, UK.

Vasconcellos, R., 2001. Tactile maps in geography. In: Hanson, S., Weinert, F.E. (Eds.), International Encyclopedia of Social and Behavioral Sciences. Elsevier, Boston, MA.

Wiedel, J.W., Groves, P., 1972. Tactual Mapping: Design, Reproduction, Reading and Interpretation. University of Maryland, College Park, MD.

Wiedel, J.W. (Ed.), 1983. Proceedings of the First International Symposium on Maps and Graphs for the Visually Handicapped. Association of American Geographers, Washington, DC.

World Health Organization – WHO, 2012a. Fact Sheet N°282. http://www.who.int/mediacentre/factsheets/fs282/en/ (accessed 01.02.13.).

WHO, 2012b. Global Data on Visual Impairment, p. 17. World Health Organization. http://www.who.int/blindness/GLOBALDATAFINALforweb.pdf (accessed 01.02.13.).

Relevant Websites

http://www.afb.org/ – American Foundation for the Blind (AFB) see also: Journal of Visual Impairment & Blindness.

http://icaci.org/ – International Cartographic Association (ICA) It has a Commission on *Maps and Graphics for Blind and Partially Sighted People*, Many Articles have been Published at the ICC Proceedings, Available at the Web site.

Perkins School; http://www.perkins.org/; It has a Large Amount of Information, Materials, Publications Related to Tactile Maps and Visual Impairment; see http://www.perkins.org/resources/scout/ and; http://www.perkins.org/resources/scout/education/geography/tactile-maps.html.

RNIB – Royal National Institute of Blind People; http://www.rnib.org.uk; It has Information and Publications on Tactile Maps, Also Offers Tactile Maps Workshops; see also RNIB Digital Accessibility Team at http://www.tiresias.org/index.htm.

University of São Paulo – USP and Laboratory of Geography Education – LEMADI; http://www.usp.br/ and http://www.geografia.fflch.usp.br/; Cartografia Tátil e Ensino de Geografia; It has Photos of Teaching Materials, Tactile Maps, Videos, Publications.

http://www.who.int/ – World Health Organization (WHO); Information and Data about Blindness and Visual Impairment.

Tactile Processing

Alberto Gallace, University of Milano-Bicocca, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Tactile information processing is of vital importance for differentiating and protecting our body from the external world. At the same time, it provides a powerful means of physical and social interaction with other individuals. Not surprisingly, then, tactile processing develops and reaches full functionality well before other sensory modalities. Despite the fact that the neurocognitive mechanisms of tactile information processing are understudied with respect to the processing occurring in other senses, researchers have just started to unveil their complexity. In particular, it has been shown that in order to process touch, our brain requires the integration of information coming from different classes of tactile sensory-specific receptors, with information arising from the processing of other sensory stimuli (e.g., visual or auditory). Cognitive factors, attention, beliefs, and the physiological parameters of the organisms also enter the computation that leads to tactile perception. Even more interestingly, research has shown that tactile processing is very effective in driving our behavior in a number of different contexts. This powerful effects of touch are, at least in part, due to the role played by newly discovered tactile fibers that would seem to be related to the most pleasant aspects of our social behavior.

Introduction

Far less attention has been given by researchers to the study of tactile processing as compared to the study of information processing in other sensory modalities, such as vision and audition. This has occurred for a number of reasons, comprising the difficulties experienced in the past in stimulating tactile receptors in a controlled fashion, as well as the common view that vision is our dominant sense. The last few years have seen a renewed interest in the study of tactile processing from a number of different perspectives (see Gallace, 2012; Gallace and Spence, 2014; Hertenstein and Weiss, 2011; McGlone and Spence, 2010). Neuroscientists, psychologists, anthropologists, engineers, designers, and clinicians are now more than ever before interested in understanding the mechanisms of tactile information processing, for both theoretical and applied reasons. This new wave of interest is likely driven by a number of factors, such as the availability of devices that allow for tactile stimuli to be precisely delivered, the realization of how important touch is for our well-being, the development of multisensory virtual reality technologies, the increased use of human–machine interfaces that require active manipulation, and the need to provide more complete access to information to visually impaired individuals, just to cite a few of them (Gallace and Spence, 2014; Gallace et al., 2011a).

It is now clear that tactile information processing is of great importance in order to offer protection to the body, from both physical and psychological points of view. That is, touch provides information about potentially dangerous stimuli that get in touch with our skin, and at the same time contributes to differentiating what is 'us' from what is the outside world. However, touch not only provides a protective function to our organism but also acts as a powerful interface that allows us to interact, both socially and physically, with external stimuli. Not surprisingly, then, touch is the very first sense to develop; and a 6-week-old embryo, who is still totally blind and deaf, has already the ability to perceive tactile stimuli (e.g., Atkinson and Braddick, 1982; Bernhardt, 1987; Bremner et al., 2012). Interestingly, early tactile interactions have also been reported to occur between twin fetuses within the womb (Castiello et al., 2010). Even after birth, when the sight of the newborn still has very poor resolution, touch offers the very first means to interact with the world. In fact, the automatic hand closure response of babies to objects that are put on the palm of their hands is likely a mechanism developed for helping the newborn to determine the nature of external stimuli.

Before discussing further the mechanisms of tactile processing, it is important to make a distinction between touch and haptic perception. In particular, researchers distinguish haptic from tactile perception because the former involves the active, serial exploration of a stimulus by means of different body parts (i.e., such as when we try to look for a certain object in our pocket or in a bag without looking at it). By contrast, tactile perception can occur regardless of our movements, even when an object lays or moves on our skin surface (Lederman and Klatzky, 2009).

Neural Mechanisms of Tactile Processing

Tactile processing starts from the activation of receptors distributed throughout our skin. Different classes of receptors respond to different forms of information or energies (see Gallace and Spence, 2014; Kandel et al., 2000). The Pacinian corpuscles, found mainly on the palms of the hands and soles of the feet, respond to the mechanical deformation of the skin. The Meissner corpuscles also respond to pressure stimuli and are responsible for the higher tactile sensitivity of certain body sites as compared to others. In fact, our sensitivity to tactile information is not the same across the entire body surface. Body areas such as the fingertips and the lips have far greater sensitivities than the back or the calf (e.g., Weinstein, 1968). Other receptors respond to thermal rather than mechanical information. That is, 'thermoreceptors' code absolute and relative changes in temperature within a nonpainful range (e.g., Darian-Smith et al., 1979; Schepers and Ringkamp, 2010). Certain receptors, such as the hair root plexuses, are wrapped around our hair roots and can detect

movements on the skin surface that disturb the hairs (e.g., Kandel et al., 2000; Purves et al., 2004). Importantly, receptors differ not only in terms of the form of energy that they are capable of transducing but also in terms of their pattern of response. In particular, fast-adapting receptors fire for only a few milliseconds after the presentation of a stimulus on the skin, and then they become silent. By contrast, slowly adapting receptors keep responding until the stimulus is removed from the skin (e.g., Birznieks et al., 2009).

It is important to mention here that even if tactile receptors specialize in transducing energies that are interpreted by our brain as tactile stimuli, their activation is not always necessary or sufficient to elicit specific tactile percepts. That is, it is now well established that people can sometimes mislocalize tactile sensation toward regions of space where there are no receptors or sometimes not even a body surface (see Gallace and Spence, 2014, for a review). Tactile sensory illusions as well as tactile hallucinations (e.g., 'delusional parasitosis,' where people perceive crawling sensations on their body) have also been extensively reported by researchers in this field (e.g., Hayward, 2008; Huber et al., 2008; Musalek et al., 1989). Moreover, what we commonly call 'touch' is actually the final product of the interactions occurring among information coming to the brain from different classes of peripheral receptors (see Berkley and Hubscher, 1995; Iggo, 1977; McGlone and Reilly, 2010). For example, certain sensations, such as those of wetness (given by the activation of both mechanoreceptors and thermoreceptors) or of tickle (produced by the common activation of the free nerve endings and the Pacinian corpuscles), are the consequence of the multisensory interactions occurring in the brain. That is, no dedicated receptor in our skin is specialized for coding wet or tickling stimuli.

Different brain areas are responsible for the processing of tactile information. In particular, the primary somatosensory cortex (SI) of the brain receives afferent stimuli from each part of our body, so that different areas within this structure represent tactile stimuli originating from specific body regions. By analyzing the responses of the somatosensory cortex to stimuli delivered to our skin, it is therefore possible to obtain a complete neural representation of the body, known as the 'homunculus man' (or, as some have renamed it, the 'hermunculus' when referring to the female body; Di Noto et al., 2012; Penfield and Rasmussen, 1950). This representation in our brain does not fully match the physical structure of the body, but it depends on the amount of neural fibers and receptors that innervate different body parts. For example, within the homunculus, the representation of the face is larger than the representation of the arm, the face being a body part that is innervated by a higher number of neural fibers (and thus more sensitive to touch) than the arm. The left primary somatosensory cortex responds only to stimuli presented on the right side of the body, while the right primary somatosensory cortex responds to stimuli coming from the left body side. Another important area in tactile processing is the secondary somatosensory cortex (SII), where information coming from different body parts is integrated. That is, there is now ample evidence showing that representations posterior to SI, such as those in SII, receive inputs from both body sides (e.g., Blatow et al., 2007; Ferretti et al., 2003).

Higher Order Processing of Tactile Information and Tactile Awareness

Most of the information that is processed by our brain does not enter our awareness, and this is certainly true also for tactile sensations. For example, tactile signals are continuously used by our neural system in order to control our posture when we sit or walk, but we are not constantly aware of the presence of the chair or the floor behind our body (at least if we do not explicitly attend to them). The relevance of this automatic, and mostly unaware, processing for our daily life becomes apparent only when we observe the behavior of those patients who lost their ability to perceive tactile and proprioceptive (information regarding the position of our body parts in space) signals coming from their body, due to certain forms of illness. As this quote clearly suggests, for these patients even apparently simple daily activities such as sitting, walking, or feeding become a huge challenge (e.g., Cole and Paillard, 1995):

> Speaking and chewing were almost impossible, and for several months she existed solely on pureed food. Then she became so exasperated by this that she began to experiment in eating. She learned to push food to one side of the mouth, chew a certain number of times, and then push the food to the back of her mouth and swallow it automatically. Facial expression was severely limited because of the absence of proprioception from the lower face. She refused to look in a mirror for two years after the illness. For a similar time, she had to remember to concentrate to keep her mouth from opening. (From Cole and Paillard, 1995: p. 249)

Despite the amount of implicit processing of tactile information occurring in our brain, there are situations where becoming aware of tactile stimuli becomes really important. This typically occurs when new or potentially relevant stimuli get in touch with (or are removed from) our body surface. That is, while it is not relevant for us to be constantly aware of the presence of the chair where we sit, perceiving that someone or something has suddenly touched our shoulder is often relevant. An unexpected change in the status in our tactile environment certainly requires awareness. Surprisingly, however, also potentially relevant information often goes unnoticed. Those who have been pickpocketed at least once in their life certainly know too well how easy it is to fail to perceive that something is happening somewhere on our body! Laboratory studies have shown that the presence of distractors, whether tactile or visual, as well as of a discontinuity in the flow of tactile information can easily lead to the failure to perceive tactile changes on our skin (a phenomenon that is known as 'change numbness' and that also has been shown to occur in visual perception – where it is known as 'change blindness'; Gallace et al., 2006; Simons and Rensink, 2005). This phenomenon is thought to be due to the competition between multiple sources of information for accessing processing (i.e., cognitive or attentional) resources in our brain (e.g., Simons and Rensink, 2005).

As far as the neural mechanisms of tactile awareness are concerned, ample evidence now suggests that tactile processing occurring in SI (discussed in this article) is necessary, but not sufficient, for the awareness of tactile sensations to occur (Gallace and Spence, 2010b). In particular, magnetic imaging

on human patients who reported a lack of somatosensory sensations revealed the presence of early neural activations in the primary somatosensory cortex, but the absence of any later activation in higher order brain areas (Preissl et al., 2001). Similarly, a study on neurologically normal participants showed that early brain activity generated in the somatosensory cortex contralateral to the stimulated region of the body was uncorrelated with the perception of the stimulus. By contrast, later activity in parietal and frontal areas was observed for consciously perceived target stimuli (Palva et al., 2005).

Interestingly, studies on patients affected by brain damage have shown that a correct localization of tactile stimuli can be performed despite a lack of awareness regarding the presence of the stimuli (e.g., Brochier et al., 1994; Imanaka et al., 2002; Rossetti et al., 2001). For example, Brochier et al. (1994) described a patient who, as a consequence of damage to his right primary somatosensory cortex, lost tactile sensitivity in his left arm but was able to point correctly to stimulated locations on his limb when blindfolded. That is, the ability to process spatial information regarding the position of tactile stimuli, carried on by higher order parietal brain areas, is not sufficient to lead to awareness of tactile information. On the basis of these different observations, it has been suggested that in order for awareness of touch to occur, tactile processing needs to be supported by complex brain networks likely involving sensory-specific areas (such as SI), as well as higher order associative brain areas (e.g., the posterior parietal cortex, the insular cortex, and the prefrontal cortex; e.g., Gallace and Spence, 2010b).

Social Aspects of Tactile Processing

Despite of the importance of social interactions in our life, their tactile aspects have been studied very little by researchers very little (see Gallace and Spence, 2010a; Hertenstein et al., 2006). However, the few studies that have addressed the role of tactile processing within this domain have found surprisingly interesting results. In particular, it has been shown that tactile social interactions, even if unnoticed or irrelevant to the context where they occur, have a very powerful effect on a person's evaluation of a given situation or in eliciting certain kinds of behavior. For example, people have been shown to be more compliant to a request when they are touched by another person than when they are not touched (a phenomenon known as the 'Midas touch effect'; e.g., Crusco and Wetzel, 1984). However, within certain contexts, touch can also have negative consequences on our behavior or judgments. For example, Martin (2012) reported that people who were touched by another customer when examining products in a store evaluated the brands more negatively than those who were not touched.

One of the most promising fields of investigation within research on social touch is likely related to the discovery in humans of a new class of tactile fibers, known as C-Tactile (CT) afferents (e.g., see Löken et al., 2009; McGlone and Spence, 2010). These fibers, found in the hairy but not in the glabrous skin, respond best when the skin is stroked at a speed of about 1–10 cm s^{-1}, a stimulation that resembles a caress (e.g., Bessou et al., 1971; Iggo, 1977; Olausson et al., 2008). That is, these fibers would seem to be associated with the perception of the more pleasant aspects of tactile stimulation. It has even been suggested that the CT fibers are part of a neural system responsible of the maintenance of physical and social well-being in humans (Björnsdotter et al., 2010).

The presence of CT fibers in humans is certainly relevant proof of the importance played by tactile processing in affiliative social interactions. This crucial role of tactile interactions in our social behavior is also supported by a number of studies on couples. In particular, it has been reported that women who received physical partner contact before a stressful situation exhibited significantly lower cortisol and heart rate responses to stress (e.g., Ditzen et al., 2007).

It is important to mention here that many effects of social touch on behavior and physiological responses are likely mediated by hormonal mechanisms. In fact, it has been demonstrated that tactile stimulation, whether sexual or nonsexual, induces the release of oxytocin (Shermer, 2004; Williams et al., 1992), a hormone implicated in human and animal bonding behavior (e.g., Bales and Carter, 2003). Interestingly, the relationship between tactile processing and hormone release would seem to be a two-way one. That is, touch certainly affects the release of hormones, but also the perception of touch can be affected by hormones. For example, during the refractory period (the period of time immediately after orgasm), men experience tactile hypersensitivity in the glans of the penis (Yilmaz and Aksu, 2000). This sudden change in sensitivity results in tactile stimuli being perceived as aversive, and seems to be related to the release of prolactin that follows the orgasm (Krüger et al., 2003).

The Contribution of Tactile Processing to the Representation of the Body

Tactile processing not only contributes to the perception of external stimuli but also plays a major role in differentiating our body from the rest of the world. Even if we often take for granted that we have a body and that this body is ours and nobody else's, our brain actually needs to activate specific mechanisms in order to support this perception of ownership. As a matter of fact, in order to correctly perform actions within the environment, at every instant our neural system needs to establish what is our body and where it is in space. This task is achieved mostly automatically, by combining tactile, proprioceptive, and visual information regarding the body and the space around it.

Evidence of the presence of a close relationship between our body and tactile processing comes from those studies that have investigated the role of body posture on tactile perception. For example, it has been shown that it takes more time to determine the order of two sequential stimuli presented on the hand when these are positioned far apart rather than close together (e.g., Gallace and Spence, 2005; Shore et al., 2002). Similarly, tactile (as well as noxious) stimuli are perceived as less intense and require more time to be localized when the arms are crossed over the midline, rather than when they are kept parallel (e.g., Gallace et al., 2011b). That is, the processing

of tactile information cannot be fully separated from the processing of information regarding the current status of the body (i.e., in terms of its spatial position).

A large and fast-growing body of evidence unequivocally shows that people can perceive tactile stimuli as originating from spatial positions where the body is not physically present, such as from a rubber hand or even from a mannequin (e.g., Botvinick and Cohen, 1998; Petkova and Ehrsson, 2008). This is achieved by delivering synchronous visuo-tactile signals that, after being interpreted by the brain, lead to a mislocalization of the veridical position of the body. For example, in the rubber hand illusion a fake hand is stroked synchronously with the participant's hidden real hand. After a few seconds, the participant starts to claim that tactile stimuli come from the position occupied by the rubber hand rather than from their real hand. If required to point toward their hidden hand, the participants point toward a position that lies between the real and the rubber hand (e.g., Botvinick and Cohen, 1998; Tsakiris and Haggard, 2005). It has also been shown that this illusion leads to a temporary thermoregulatory dysfunction (a drop in skin temperature) of the hand that is replaced by its rubber equivalent (Moseley et al., 2008). Importantly, the visual similarities between the rubber hand and the participant's real hand do not seem to play a role in terms of inducing the illusion (e.g., Hohwy and Paton, 2010; Farmer et al., 2012).

Multisensory Aspects of Tactile Processing

The phenomenon of the rubber hand illusion clearly suggests that the integration of visual and tactile information is vital for sustaining a coherent representation of our body. However, more in general, the presence of this illusion is also suggestive of the fact that tactile processing affects and is affected by the processing of information coming from other sensory modalities in our brain. As far as this point is concerned, it has often been reported that visual information dominates over tactile or auditory information (e.g., Bertelson and de Gelder, 2004; Rock and Harris, 1967). For example, Hartcher O'Brien et al. (2008) asked their participants to discriminate between unimodal visual, unimodal tactile, and bimodal (visual and tactile) stimuli. They found that the participants failed to respond to touch far more often than they failed to respond to vision on the bimodal trials. That is, the simultaneous presence of visual stimuli often extinguished the perception of tactile information.

Besides the studies that have shown visual dominance over touch, there are also examples suggesting that even touch, under certain conditions of stimulus presentation, can affect the perception of information coming from other sensory systems. For example, tactile cues often dominate over visual cues when judging attributes such as an object's temperature (e.g., Bushnell et al., 1985) or its fine surface texture (e.g., Guest et al., 2002). Similarly, Violentyev et al. (2005) found that participants report the presence of two flashes when only one was produced, if two tactile taps were delivered within a 60 ms interval. As a consequence of these observations, as well as of similar observations made on different couples of sensory modalities, it has been suggested that there is not a single sensory modality that always dominates over the others. By contrast, under conditions of multisensory stimulus presentation, the sensory modality that carries the lowest variance (i.e., noise) in a person's estimate regarding a given quality of a stimulus is the one that will likely 'drive' the processing of the inputs provided by the other senses (e.g., Ernst and Banks, 2002; Ernst and Bülthoff, 2004).

When one considers all of the factors that seem to affect tactile information processing, it should be mentioned that not only stimuli from other sensory modalities are able to drive our perception of touch, but also cognitive factors and beliefs are equally effective (e.g., Kress et al., 2012). For example, it has been demonstrated that people's evaluation of the very same tactile stimulation changed as a function of the gender of the person who was thought to deliver the stimuli (Gazzola et al., 2012). That is, somewhat unsurprisingly, it has been demonstrated that heterosexual males evaluate tactile stimuli as less pleasant when they believe that the stimulation was delivered by a man wearing a tracksuit, as compared to when they perceive that the stimuli are delivered by a woman wearing a black evening dress and high heels. However, the most interesting aspect of this study is perhaps that not only the evaluation of the stimuli but also the neural activations elicited by them were affected by the participant's beliefs. Specifically, the activity of SI, an area thought to be primarily involved in processing the more perceptual aspects of tactile stimuli (rather than their hedonic value), was modulated by the visually perceived gender of the caresser.

Conclusions

Tactile processing has certainly been less studied as compared to visual or auditory information processing. Despite this, scientific research within this domain is quickly moving and has just started to unveil how relevant tactile stimulations are for our survival and well-being. It is now clear that tactile processing contributes to supporting our sense of body ownership, as well as our ability to perform even the simplest movements. Tactile information processing is affected by other signals that are concurrently processed by the brain. In particular, stimuli from other sensory modalities, cognitive factors, beliefs, attention, and certain physiological parameters of the body (comprising the hormonal levels) are all able to drive our interpretation of (and response to) the stimuli that get in touch with our skin. Similarly, the processing of tactile stimuli can efficiently modulate our brain responses to other sources of information.

Tactile processing is likely to become an even more important research topic in the future, given its relevance to the development of modern technologies and product design. In fact, an increasing number of everyday objects, from touch screens to food packaging, in order to be fully appreciated by the consumer need to capitalize on our knowledge of the neurocognitive mechanisms supporting tactile information processing (e.g., Gallace and Spence, 2014; Spence and Gallace, 2011).

See also: Consumer Psychology; Neonatal Bonding; Social Cognition; Social Psychology.

Bibliography

Atkinson, J., Braddick, O., 1982. Sensory and perceptual capacities of the neonate. In: Stratton, P. (Ed.), Psychobiology of the Human Newborn. John Wiley, London, pp. 191–220.

Bales, K.L., Carter, C.S., 2003. Sex differences and developmental effects of oxytocin on aggression and social behavior in prairie voles (*Microtus ochrogaster*). Hormones and Behavior 44, 178–184.

Berkley, K.J., Hubscher, C.H., 1995. Are there separate central nervous system pathways for touch and pain? Nature Medicine 1, 766–773.

Bernhardt, J., 1987. Sensory capabilities of the fetus. MCN American Journal of Maternal Child Nursing 12, 44–46.

Bertelson, P., de Gelder, B., 2004. The psychology of multimodal perception. In: Spence, C., Driver, J. (Eds.), Crossmodal Space and Crossmodal Attention. Oxford University Press, Oxford, UK, pp. 141–178.

Bessou, P., Burgess, P.R., Perl, E.R., Taylor, C.B., 1971. Dynamic properties of mechanoreceptors with unmyelinated (C) fibers. Journal of Neurophysiology 34, 116–131.

Birznieks, I., Macefield, V.G., Westling, G., Johansson, R.S., 2009. Slowly adapting mechanoreceptors in the borders of the human fingernail encode fingertip forces. Journal of Neuroscience 29, 9370–9379.

Björnsdotter, M., Morrison, I., Olausson, H., 2010. Feeling good: on the role of C fiber mediated touch in interoception. Experimental Brain Research 207, 149–155.

Blatow, M., Nennig, E., Durst, A., Sartor, K., Stippich, C., 2007. fMRI reflects functional connectivity of human somatosensory cortex. Neuroimage 37, 927–936.

Botvinick, M., Cohen, J., 1998. Rubber hands 'feel' touch that eyes see. Nature 391, 756.

Bremner, A., Lewkowicz, D., Spence, C. (Eds.), 2012. Multisensory Development. Oxford University Press, Oxford.

Brochier, T., Habib, M., Brouchon, M., 1994. Covert processing of information in hemianesthesia: a case report. Cortex 30, 135–144.

Bushnell, E.W., Shaw, L., Strauss, D., 1985. Relationship between visual and tactual exploration by 6-month-olds. Developmental Psychology 21, 591–600.

Castiello, U., Becchio, C., Zoia, S., Nelini, C., Sartori, L., Blason, L., D'Ottavio, G., Bulgheroni, M., Gallese, V., 2010. Wired to be social: the ontogeny of human interaction. PLoS ONE 5 (10), e13199.

Crusco, A.H., Wetzel, C.G., 1984. The Midas touch: the effects of interpersonal touch on restaurant tipping. Personality and Social Psychology Bulletin 10, 512–517.

Darian-Smith, I., Johnson, K.O., LaMotte, C., Shigenaga, Y., Kenins, P., Champness, P., 1979. Warm fibers innervating palmar and digital skin of the monkey: responses to thermal stimuli. Journal of Neurophysiology 42, 1297–1315.

Di Noto, P., Newman, L., Wall, S., Einstein, G., 2012. The hermunculus: what is known about the representation of the female body in the brain? Cerebral Cortex. http://dx.doi.org/10.1093/cercor/bhs005.

Ditzen, B., Neumman, I., Bodenmann, G., vonDawans, B., Turner, R.A., Ehlert, U., Heinrichs, M., 2007. Effects of different kinds of couple interaction on cortisol and heart rate responses to stress in women. Psychoneuroendocrinology 32, 565–574.

Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415, 429–433.

Ernst, M.O., Bülthoff, H.H., 2004. Merging the senses into a robust percept. Trends in Cognitive Sciences 8, 162–169.

Farmer, H., Tajadura-Jiménez, A., Tsakiris, M., 2012. Beyond the colour of my skin: how skin colour affects the sense of body-ownership. Consciousness and Cognition 21, 1242–1256.

Ferretti, A., Babiloni, C., Gratta, C.D., Caulo, M., Tartaro, A., Bonomo, L., Rossini, P.M., Romani, G.L., 2003. Functional topography of the secondary somatosensory cortex for nonpainful and painful stimuli: an fMRI study. Neuroimage 20, 1625–1638.

Gallace, A., Spence, C., 2005. Visual capture of apparent limb position influences tactile temporal order judgments. Neuroscience Letters 379, 63–68.

Gallace, A., Spence, C., 2010a. The science of interpersonal touch: an overview. Neuroscience and Biobehavioral Reviews 34, 246–259.

Gallace, A., Spence, C., 2010b. Touch and the body: the role of the somatosensory cortex in tactile awareness. Psyche 16, 30–67. http://www.theassc.org/journal_psyche/archive/vol_16_no_1_2010.

Gallace, A., Spence, C., 2014. In Touch with the Future: The Sense of Touch from Cognitive Neuroscience to Virtual Reality. Oxford University Press, Oxford, UK.

Gallace, A., 2012. Living with touch: understanding tactile interactions. The Psychologist 25, 3–5.

Gallace, A., Ngo, M.K., Sulaitis, J., Spence, C., 2011a. Multisensory presence in virtual reality: possibilities & limitations. In: Ghinea, G., Andres, F., Gulliver, S. (Eds.), Multiple Sensorial Media Advances and Applications: New Developments in MulSeMedia. IGI Global.

Gallace, A., Tan, H.Z., Spence, C., 2006. The failure to detect tactile change: a tactile analogue of visual change blindness. Psychonomic Bulletin & Review 13, 300–303.

Gallace, A., Torta, D.M., Moseley, G.L., Iannetti, G., 2011b. The analgesic effect of crossing the arms. Pain 152, 1418–1423.

Gazzola, V., Spezio, M.L., Etzel, J.A., Castelli, F., Adolphs, R., Keysers, C., 2012. Primary somatosensory cortex discriminates affective significance in social touch. Proceedings of the National Academy of Sciences, USA 109, E1657–E1666.

Guest, S., Catmur, C., Lloyd, D., Spence, C., 2002. Audiotactile interactions in roughness perception. Experimental Brain Research 146, 161–171.

Hartcher O'Brien, J., Gallace, A., Krings, B., Koppen, C., Spence, C., 2008. When vision 'extinguishes' touch in neurologically-normal people: extending the Colavita effect. Experimental Brain Research 186, 643–658.

Hayward, V., 2008. A brief taxonomy of tactile illusions and demonstrations that can be done in a hardware store. Brain Research Bulletin 75, 742–752.

Hertenstein, M.J., Verkamp, J.M., Kerestes, A.M., Holmes, R.M., 2006. The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research. Genetic, Social, and General Psychology Monographs 132, 5–94.

Hertenstein, M., Weiss, S. (Eds.), 2011. The Handbook of Touch. Springer, Berlin.

Hohwy, J., Paton, B., 2010. Explaining away the body: experiences of supernaturally caused touch and touch on non-hand objects within the rubber hand illusion. PLoS ONE 5, e9416.

Huber, M., Karner, M., Kirchler, E., Lepping, P., Freudenmann, R.W., 2008. Striatal lesions in delusional parasitosis revealed by magnetic resonance imaging. Progress in Neuro-psychopharmacology and Biological Psychiatry 32, 1967–1971.

Iggo, A., 1977. Cutaneous and subcutaneous sense organs. British Medical Bulletin 33, 97–102.

Imanaka, K., Kita, I., Suzuki, K., 2002. Effects of nonconscious perception on motor response. Human Movement Science 21, 541–561.

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (Eds.), 2000. Principles of Neural Science, fourth ed. McGraw-Hill, New York.

Kress, I.U., Minati, L., Ferraro, S., Critchley, H.D., 2012. Direct skin-to-skin vs. indirect touch modulates neural responses to stroking vs. tapping. Neuroreport 22, 646–651.

Krüger, T.H., Haake, P., Haverkamp, J., Krämer, M., Exton, M.S., Saller, B., Leygraf, N., Hartmann, U., Schedlowski, M., 2003. Effects of acute prolactin manipulation on sexual drive and function in males. Journal of Endocrinology 179, 357–365.

Lederman, S.J., Klatzky, R.L., 2009. Haptic perception: a tutorial. Attention, Perception & Psychophysics 71, 1439–1459.

Löken, L.S., Wessberg, J., Morrison, I., McGlone, F., Olausson, H., 2009. Coding of pleasant touch by unmyelinated afferents in humans. Nature Neuroscience 12, 547–548.

Martin, B.A.S., 2012. A stranger's touch: effects of accidental interpersonal touch on consumer evaluations and shopping time. Journal of Consumer Research 39, 174–184.

McGlone, F.P., Spence, C., 2010. Editorial: the cutaneous senses: touch, temperature, pain/itch, and pleasure. Neuroscience and Biobehavioural Reviews 34, 145–147.

McGlone, F., Reilly, D., 2010. The cutaneous sensory system. Neuroscience and Biobehavioral Reviews 34, 148–159.

Moseley, G.L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A., Spence, C., 2008. Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. Proceedings of the National Academy of Sciences USA 105, 13168–13172.

Musalek, M., Podreka, I., Walter, H., Suess, E., Passweg, V., Nutzinger, D., Strobl, R., Lesch, O.M., 1989. Regional brain function in hallucinations: a study of regional cerebral blood flow with 99m-Tc-HMPAO-SPECT in patients with auditory hallucinations, tactile hallucinations and normal controls. Comprehensive Psychiatry 30, 99–108.

Olausson, H., Cole, J., Rylander, K., McGlone, F., Lamarre, Y., Wallin, B.G., Krämer, H., Wessberg, J., Elam, M., Bushnell, M.C., Vallbo, A., 2008. Functional role of unmyelinated tactile afferents in human hairy skin: sympathetic response and perceptual localization. Experimental Brain Research 184, 135–140.

Palva, S., Linkenkaer-Hansen, K., Näätänen, R., Palva, J.M., 2005. Early neural correlates of conscious somatosensory perception. Journal of Neuroscience 25, 5248–5258.

Penfield, W., Rasmussen, T.L., 1950. The Cerebral Cortex of Man: A Clinical Study of Localization of Function. Macmillan, London.

Petkova, V.I., Ehrsson, H.H., 2008. If I were you: perceptual illusion of body swapping. PLoS ONE 3, e3832.

Preissl, H., Flor, H., Lutzenberger, W., Duffner, F., Freudenstein, N., Grote, E., Birbaumer, N., 2001. Early activation of the primary somatosensory cortex without conscious awareness of somatosensory stimuli in tumor patients. Neuroscience Letters 308, 193–196.

Purves, D., Augustine, G.J., Fitzpatrick, D., Katz, L.C., LaMantia, A.-S., McNamara, J.O., Williams, S.M., 2004. Neuroscience. Sinauer Associates, Sunderland, MA.

Rock, I., Harris, C.S., 1967. Vision and touch. Scientific American 216 (5), 96–104.
Rossetti, Y., Rode, G., Boisson, D., 2001. Numbsense: a case study and implications. In: De Gelder, B., De Haan, E., Heywood, C. (Eds.), Out of Mind: Varieties of Unconscious Processing. Oxford University Press, pp. 265–292.
Schepers, R.J., Ringkamp, M., 2010. Thermoreceptors and thermosensitive afferents. Neuroscience and Biobehavioral Reviews 34, 177–184.
Shermer, M., 2004. A bounty of science. Scientific American 290 (2), 33.
Shore, D.I., Spry, E., Spence, C., 2002. Confusing the mind by crossing the hands. Cognitive Brain Research 14, 153–163.
Simons, D.J., Rensink, R.A., 2005. Change blindness: past, present, and future. Trends in Cognitive Sciences 9, 16–20.
Spence, C., Gallace, A., 2011. Multisensory design: reaching out to touch the consumer. Psychology & Marketing 28, 267–308.
Tsakiris, M., Haggard, P., 2005. The rubber hand illusion revisited: visuotactile integration and self-attribution. Journal of Experimental Psychology. Human Perception and Performance 31, 80–91.
Violentyev, A., Shimojo, S., Shams, L., 2005. Touch-induced visual illusion. Neuroreport 16, 1107–1110.
Weinstein, S., 1968. Intensive and extensive aspects of tactile sensitivity as a function of body part, sex, and laterality. In: Kenshalo, D.R. (Ed.), The Skin Senses. Thomas, Springfield, IL, pp. 195–222.
Williams, J.R., Catania, K.C., Carter, C.S., 1992. Development of partner preferences in female prairie voles (*Microtus ochrogaster*): the role of social and sexual experience. Hormones and Behavior 26, 339–349.
Yilmaz, U., Aksu, M., 2000. The postejaculatory refractory period: a neurophysiological study in the human male. British Journal of Urology International 85, 1093–1096.

Tangible User Interfaces in Learning and Education

Yuxia Zhou, Yunnan Normal University, Kunming, China
Minjuan Wang, Oriental Scholar of Shanghai International Studies University, Shanghai, China; and San Diego State University, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a replacement of the previous edition article by K. Opwis, volume 11, pp. 7573–7577, © 2001, Elsevier Ltd.

Abstract

Learning happens for human being mostly through physical, cognitive, and emotional interactions with the external environment. With the aid of information technology, interactions have gone beyond the limit of working on a desktop computer, using a mouse and keyboard to interact with windows, icons, menus, and pointers, in which tangible user interface (TUI) is gaining increasing attention. This article will provide an overview of TUI, discuss its functional characteristics, present some application cases, and discuss design and application issues for TUI in education.

Tangible user interface (TUI) is a user interface in which a person interacts with digital information through the physical environment (Roebuck, 2011). TUI aims to interlink the digital world and physical world seamlessly so as to allow humans to acquire knowledge of the world around them through holistic interactions with their external environment.

Tangible User Interface

How can human beings learn by taking the most advantage of interactions with the environment? In decades, with the fast development of information technology, the computer has become the critical part of what people interact with. However, sitting in front of the screen, hearing through earphone, moving mouse, and clicking keyboard have isolated humans from the physical world. In *Back to the Real World* (1993), the authors argued that both desktop computers and virtual reality are keeping human away from 'natural world'. Therefore people should augment and enrich the real world with digital technology, thereby transiting from the real world to the digital world fluidly without the cost of losing physical touch. Papert at the MIT Media Lab pioneered in making the use of computers for 'thinking about thinking' in embedded programs. Such project (e.g., LEGO/Logo) allows children to become the designers and creators of their physical interactive machines (e.g., robots, motorized cars, motion detectors, and various other simple digital transducers). Following development in an initiative interaction style is the notion of a graspable user interface (GUI) introduced by Fitzmaurice et al. (1995), where graspable handles made of wooded blocks are used to manipulate digital objects to increase the directness and manipulability of graphical user interfaces. However, GUIs fall short of embracing the richness of human senses and skills (Ishii and Ullmer, 1997). Transition from 'painted bits' to 'tangible bits' is further developed by Ishii and his research group. They use tangible objects rather than graphical blocks to manipulate and represent digital content by developing TUI prototypes named Geospace and Urp. TUI interaction model surpasses GUI's remote control of digital information and invites users to manipulate their physical world.

TUI Interaction Model

Input and output or control and representation are paired concepts in such interaction. Figure 1 describes the interactive relationship between human and digital information through inputting for control and getting output. According to the basic model, Ishii puts forward GUI and TUI models as shown in Figures 2 and 3.

As Figure 2 shows, in the GUI model, computer input and output are separated, because input is obtained from control 'peripherals' such as the keyboard and mouse and the output is delivered in the form of 'digital representations' such as screen-based graphics and text, or speakerphone. In other words, there is a separation between graphical representation and control. Whereas, TUI model (Figure 3) integrates physical representation with control by physically embodying the input and output in the interface that allows users to manually manipulate objects.

Characteristics of TUI

Representation is a keyword in tangible interaction and has a broad meaning. In general, there are internal representation and external representation: the former refers to a presentation to the mind in the form of an idea or image in cognitive process, or the programs and structures in computer process; the latter refers to the external manifestations of information

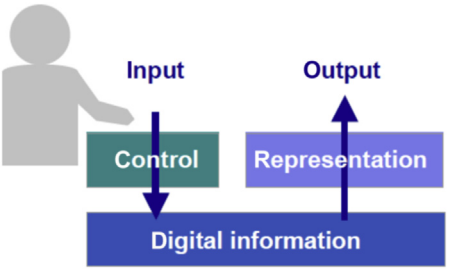

Figure 1 User interface model. Reproduced from Ullmer, B., Ishii, H., 2001. Emerging frameworks for tangible user interfaces. In: Carroll, J.M. (Ed.), Human-computer Interaction in the New Millennium. Addison-Wesley Publishing, Reading, MA, pp. 579–601.

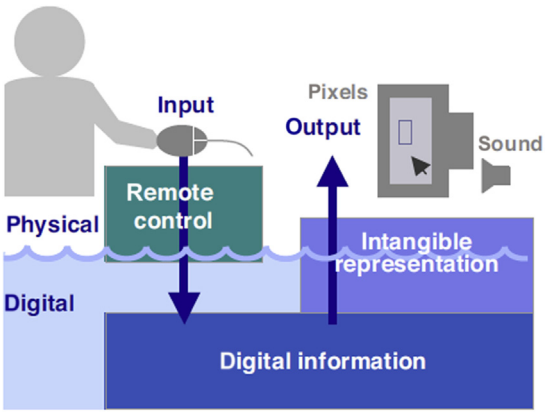

Figure 2 Graspable user interface interaction model. Reproduced from Ullmer, B., Ishii, H., 2001. Emerging frameworks for tangible user interfaces. In: Carroll, J.M. (Ed.), Human-computer Interaction in the New Millennium. Addison-Wesley Publishing, Reading, MA, pp. 579–601.

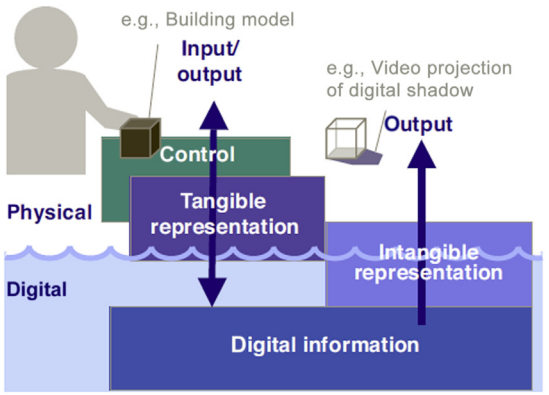

Figure 3 Tangible user interface interaction model. Reproduced from Ullmer, B., Ishii, H., 2001. Emerging frameworks for tangible user interfaces. In: Carroll, J.M. (Ed.), Human-computer Interaction in the New Millennium. Addison-Wesley Publishing, Reading, MA, pp. 579–601.

in fashions directly perceivable by the human senses. Ullmer and Ishii (2001) divide the space of external representations into two broad classes: physical representations and digital representation. Physical representation is physically embodied in concrete and tangible forms, while digital representation is digitally expressed to be computed and mediated in the computer system, thus intangible in form. For instance, the pictures on a screen or audio from a speaker are examples of digital representations, while physical abacus is an example of physical representation. The tangible interfaces are products of a careful balance between these two forms of representation.

Tangible interaction aims to seamlessly integrate the representation and control and thereby shows four properties:

- Tangible objects are coupled with digital data through computerized functionality.
- Manual interaction with tangible objects: the tangible objects embody the means of interactive control by allowing user to manipulate objects.
- The tangible objects are perceptually coupled with digitally produced representation.
- The state of the tangible objects embodies core aspects of the entire system's state (representational significance), and the system is thus at least partially legible if power is cut.

The Classification of TUIs

Ullmer et al. (2005) identify several main types of TUIs: interactive surfaces, constructive assembly, and token + constraint. Other researchers are now investigating additional kinds of systems. Slotta et al. (2013) recognized the following three types of TUI applications: immersive environments, embedded phenomena, and augmented reality (AR). Together, these perspectives cover most existing applications. Each is summarized in the list below:

- Interactive surface. The tangible objects are manipulated by users on planar surfaces in terms of arranging the location of objects and exploring their relations (e.g., the order of the placement).
- Constructive assembly. Users can handle modular and connectable factors to make meaningful objects. Unlike interactive surface, this tangible type enables users to spatially organize the modular and connectors and see the result of different orders.
- Token + Constraint system. Digital and physical objects are combined together. A token is a graspable object that embodies digital information and computation in the system that allows users to interact with either digital information or digital function. Constraint is an object that limits the properties and behavior of a token to which it is attached. It constrains a token's behavior in three ways: the way, limitation for interaction space, and reference frame for users to interact with a token + constraint system.
- Immersive virtual environments are virtual environment systems that amplify the effect of simulation by surrounding the user with numerous layers of sensory and perceptual information created by digital devices (Loomis et al., 1999). Not every immersive environment is an example of tangible interactions. Some, like immersive gaming environments, are purely digital; others, however, have been developed to promote an active, physical, or embodied interaction with the immersive environment. Below, we review the EvoRoom project, a room-sized immersive simulation environment where students and teachers interact physically with their surrounding media and their peers.
- Augmented reality. AR is a medium in which digital information is added to the physical world in registration with the world, and the information displayed depends on the location of the real world and the physical perspective of the users in physical world. In AR, users can interact with the mixed reality in different levels from simply changing the physical perspective (e.g., seeing it from a different point of view) to manipulating and even creating new information (Azuma, 1997).
- Embedded phenomenon is a learning technology framework in which simulated scientific phenomenon are mapped onto the physical space of classrooms. Students monitor

and control the local state of the simulation through distributed media positioned around the room, gathering and aggregating evidence to solve problems or answer questions related to those phenomena (Moher, 2006).

Above categories provide the continuum of TUI from manipulative to environment, and some have intertwined meaning, such as immersive environment and embedded phenomena, with commonality in using simulated reality. However, immersive environment emphasizes the involvement of user by seeing, hearing, sensing the environment, while embedded phenomenon makes innovative use of space and time in classroom for students by creating multiple natural representation in classroom and providing persistent experience for them to probe multiple vantage points.

Application Cases of TUI in Education

There is a lack of established theoretical or empirical base for how to design tangible interfaces for learning. Instead, designers can learn from successful cases in educational and related studies, as described in the following section.

Interactive Surface

For interactive surface, users are able to manipulate the tangible objects on a surface or move themselves on the surface to interact with the digitalized location and explore the relations of the order or place of the placement. Physical intervention in computational optimization (PICO) (Patten and Ishii, 2007) is such a digital surface TUI that can allow users to track and move small objects on the surface and learn from the changing variables represented by the positions of objects. A study on comparing PICO with two alternative interfaces without actuation showed that participants are better in solving a complex spatial layout problem using PICO system.

Constructive Assembly

In constructive assembly system, users can handle modular and connectable factors to make meaningful objects. Topobo is a 3D constructive assembly system for kinetic knowledge learning with model making and playing mechanics. By snapping together a combination of passive (static) and active (motorized) components, users can quickly assemble dynamic biomorphic forms such as animals or skeletons with Topobo, and activate the forms by operations such as pushing, pulling, and twisting the objects (Raffle et al., 2004). The evaluation of Topobo showed that it can help children understand certain physical principles affecting kinematic systems, and learn about the fields of modular robotics, system coordination, emergent dynamics, and locomotion. Topobo makes complex knowledge concrete by allowing users to assemble models embodying computing process and kinetic principles.

Token + Constraint System

Digital and physical objects are combined together by allowing users to manipulate tokens within the frame of constraint. Ullmer et al. (2005) developed two tangible query interface prototypes with tokens to represent database parameters upon physical constraints such as tracks and slots. User could learn database queries, views, and Boolean operation by manipulating the tokens on the constraints. However, the study showed that it was not better than traditional GUI in terms of database query learning achievement.

Immersive Environments

EvoRoom (http://encorelab.org/projects/evoroom/) is an immersive, smart classroom project where a room is turned into an 'immersive experience' – on all the walls, a simulation of a rain forest is projected, which makes users feel like standing in a rain forest environment. There are also rich audio and supplemental ambient features (like mosquito nets or potted plants) to achieve the feeling of immersivity. Students walk around the room, each with their own assignments, to achieve a scientific learning goal. They use tablet computers to make observations.

Augmented Reality

AR is an important type where information is mapped onto the physical environment, with extra layers of information allowing the learner to gain personally or contextually relevant information. Ambient Wood is an outdoor playful AR project in which pervasive technologies are used to digitally augment a woodland in a contextually relevant way, enhancing the 'usual' physical experience available to children exploring the outdoor world. Studies reveal that AR can be a highly engaging novel experience for learners. It can effectively support collaborative learning and provide preliminary guidelines for designing different ways of delivering digital information for learning.

Embedded Phenomenon

Embedded phenomenon is rather a learning technology framework than technology, in which simulated phenomena are projected onto the physical space of classrooms. Students monitor and control local state of the simulation through distributed media located around the classroom, so as to directly gather data for problem solving related to these phenomena (Moher, 2006).

A relatively mature project is RoomQuake developed by Moher and his colleagues. The classroom is a simulation system of active seismic field where a series of earthquakes is anticipated over the course of several weeks. Ambient media function as simulated seismographs that depict continuous strip chart recordings of local vibration where locality is decided by specific placement in the classroom. When simulated earthquake happens (increasing rumble), students move to the seismographic station to read the waveforms that signal the level of the vibration. The seismogram recorded provides two important pieces of information: the magnitude of the event and the distance of the event from the recording station. Students could collect several sets of data and then determine the epicenter. Two trial studies on RoomQuake show that

students' competence grew during both trials in terms of articulated skills related to explaining seismogram and understanding of event distributions. Although students could find solution by trilateration, the transformation from this physical skill to paper work is less successful, with only 40–60% of students being able to construct or select a proper rationale. Moreover, a pre–post comparison with a nontreatment group confirmed a significant learning effect associated with preunit conception.

While above application cases emphasize different tangible types (e.g., interactive surface, token manipulation, assembly, immersion, or AR), the commonality among them is to construct at-hand experience with external representation for body to regulate, constrain, and distribute cognition (Moher, 2006). With the ongoing manipulation of objects, students interact with the environment with their body, emotion, and cognition, so as to know the forthcoming and invisible principles that are not easy to grasp through abstract thinking.

Design of TUI in Education

Learners in the TUI environment are expected to engage physically, cognitively, and emotionally in learning tasks. However, how to design engaged learning in TUI is still new to many teachers. In order to fill in this gap, the Cybergogy model developed by Wang and Kang (2006) is a synthesis of current thinking of engaged learning from cognitive, social, and emotional aspects. Here we revise this model by adding the physical dimension (Figure 4).

Therefore, the TUI design model has four overlapping domains: cognitive, emotive, physical, and social. In order to effectively engage in learning, students must have the prior knowledge, be motivated to learn, and be comfortable with the learning environment in terms of social commitment, communication, involvement, and physical surroundings. For the physical dimension, learning environment is expected to provide learners with proper simulation, suitable tangible type, interactive scenario, and spatial arrangement. With this model, teachers or researchers could assess students' needs and attributes (Wang and Kang, 2006), detect students' emotional cues, and cultivate their positive feelings, so to increase their emotional engagement in learning. Similarly, to address students' cognitive need, they can also design related tangible type with specific physical interactive interface and simulation based on necessary spatial arrangement.

Challenge of TUIs in Education

While theory and application of TUI provide tentative evidence and possibility in learning, the application of TUI in education is still in its infancy. With the advantages of sensory engagement, accessibility to various students, and encouraging student to take part in group learning, comparative studies with traditional interactive learning environment also have positive

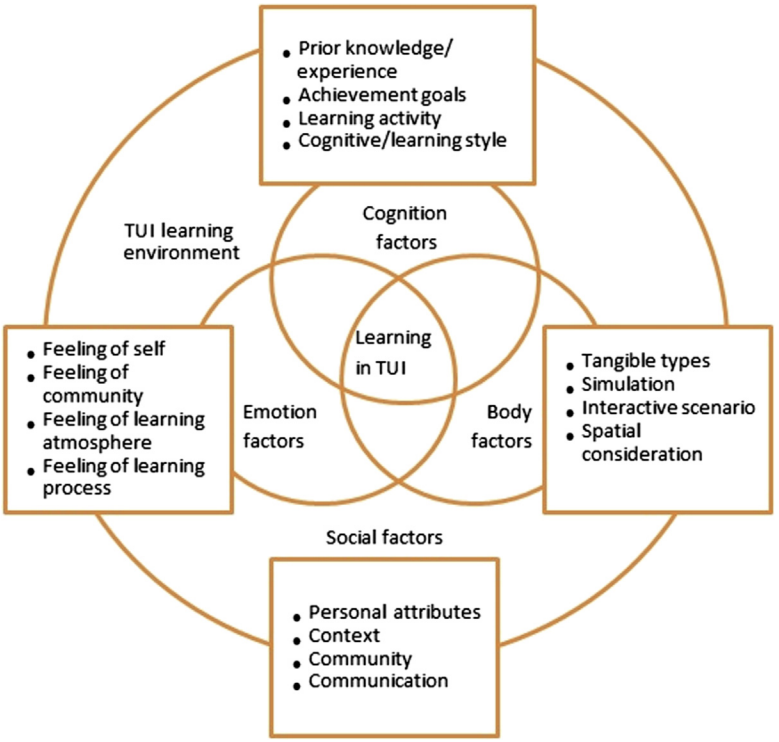

Figure 4 Learning design model in TUI. Revised from Cybergogy for Engaged Learning proposed by Wang and Kang.

learning effect at the aspects of cognition, motivation, and perception, and even abstract problem solving.

However, caution should be given to the limitation of TUI.

> Such systems require a careful balance between physical and graphical expression to avoid physical clutter, and to take advantage of the contrasting strengths of different representational forms. This balance between physical and digital representations stands as one of TUI's greatest design challenges.
> (Ullmer and Ishii, 2001: p. 600)

Scalability

Scalability is one of the biggest challenges for people who design and develop TUIs (Shaer and Hornecker, 2010). Digital representation may just have several lines code, while the mechanic to turn it into physical representation is not easy in terms of space, revision of model, and coupling the digital and physical representation in complex form.

Malleability

Malleability is another challenge because physical objects are rigid and static in comparison with digital objects that are malleable, easy to create, modify, replicate, and distribute (Shaer and Hornecker, 2010). Some technologies to meet the malleable input may be helpful, such as electrophoretic ink and shape changing materials that offer rapid form if combined with computational feedback. As for application in classroom, instructional context, participation, and affective impact on learners should be taken into consideration.

Instructional Context

In TUI, environment is a mixed context with virtual and physical representation, where there are no prescribed specific instructional sequences or materials, but students are expected to design, adapt, and construct the activity flow, maybe with scaffolding from teachers. To some degree, students are socialized in the context, and the appropriateness with proper balances should have multiple choices to meet individual need from students. In Moher's study, teachers helped much in exploring instructional contexts. For example, students' work should have the opportunity to be as visible as the phenomenon themselves. Large posters in subsequent classroom help to maintain the 'thread' of the ongoing activity and keep attention focused on the accumulating empirical evidence. Another contribution is that teachers' knowledge of individual students and how their learning activity should be organized and coordinated within the social fabric of the class (Moher, 2006). Teacher should make balances between choice and responsibility, skills and challenges, freedom and discipline, body and mind, and ensure productive team partnerships, whole class involvement, and individual scaffolding that are essential to learning.

Participation

Although TUI invites students to engage in interaction with the environment by allowing them to move and manipulate objects, the instructional decisions about the duration, space, and the challenges of tasks require careful planning and on-site negotiation with ongoing activities (Moher, 2006). The quality and the duration of engagement are determined by the design and scaffolding of the TUI application in classroom.

Student Adjustment

Although most of studies (e.g., Moher, 2006; Slotta et al., 2013) have positive feedback in terms of cognition and affection, interaction with TUI tends to require larger and different movements and manipulation than traditional mouse–keyboard interaction. Moreover, physical size of the students in relation to interaction surface determines reach. For students who are used to traditional interactions, both the interaction style and the effort needed to satisfy involvement are new challenges. Whether students feel comfortable and competent determines the learning process and result.

Integration with Curriculum

It is hard for designers and users to integrate TUI into the current fast-changing curriculum, since the physical representation, to some degree, is static and not easy to modify. Although the targets of a number of TUIs are to foster students' science literacy to investigate scientific knowledge with evidence-based learning, integration with frontline curriculum and practice is critical for its legitimacy.

Discussion

Because TUI is still an emerging and evolving field of research, its theory and practice are not sufficiently established for real applications. However, seamlessly integrating physical world and digital world is definitely a goal for digital natives or immigrants nowadays, when digital world tends to deprive people of interactions with the physical world. From emerging cognitive development (i.e., embodied cognition), TUIs gain the basic supporting points that mind is just partial for and determined by body which has richer sphere than mind to interact with the environment.

Although development in ubiquitous computing and AR has given promising technical supports for TUIs, another question is likely to arise: whether we will be smarter and closer to authentic environment, or we should tailor the environment to meet our infinite needs? The balance between adjustment and preservation, physical and digital, and human and nature is an old but unsolved problem.

How to use TUI in education? And how to engage student holistically in environment? In this article, we tend to address these questions through real cases. However, most studies of TUIs tend to be lab-based, having subjects interact with the system for a short time period, or take the form of field study. Such studies provide researchers with invaluable information about the positive affective feedback, and initial attractiveness of the system compared with alternatives. In addition, these studies will raise more questions that require further study. Based on the Cybergogy model, we propose a revised instructional design model to involve students physically, cognitively,

and emotionally in learning environments. Its application in education still needs further research, which will answer the following questions:

What will we expect students to learn in TUIs and how to assess it?
How will designers balance the body and mind in TUIs?
How will students adapt to TUIs?
How will teachers adjust themselves to the challenge?

Tackling these questions will be critical for showing the practical value and feasibility of TUIs.

See also: Computer Games and Academic Achievement; Computer Mediated Communication; Computer-Assisted Instruction; E-Communications Platforms and E-learning; Embodied Social Cognition; Human–Computer Interfaces; Virtual Reality and Spatial Cognition.

Bibliography

Azuma, R.T., 1997. A survey of augmented reality. Teleoperators and Virtual Environments 6 (4), 355–385.
Bai, Y., 2004. Discussion on stipulating curriculum goal of the curriculum standard draft. Curriculum Teaching Material and Method 9, 11–15.
Beer, R.D., 2000. Dynamical approaches to cognitive science. Trends in Cognitive Sciences 4 (3), 91–99.
Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. Harper & Row, New York.
Csikszentmihalyi, M., Rathunde, K., Whalen, S., 1997. Talented Teenagers: The Roots of Success and Failure. Cambridge University Press, Cambridge, MA.
Dreyfus, H.L., 2001. On the Internet, revised second ed. Routledge, London and New York.
Ferrari, P.F., Gallese, V., Rizzolatti, G., Fogassi, L., 2003. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. The European Journal of Neuroscience 17, 1703–1714.
Fitzmaurice, G.W., Ishii, H., Buxton, W., 1995. Bricks: laying the foundations for graspable user interfaces. In: Proceedings of CHI95. ACM, NY, pp. 442–449.
Gallese, V., 2000. The inner sense of action: agency and motor representations. Journal of Consciousness Studies 7, 23–40.
Grezes, J., Decety, J., 2001. Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. Human Brain Mapping, 1–19.
Heidegger, M., 1975. The Basic Problems of Phenomenology. Indiana University Press, Bloomington, IN.
Ishii, H., Ullmer, B., 1997. Tangible bits: towards seamless interfaces between people, bits and atoms. In: Proceedings of CHI97. ACM, NY, pp. 234–241.
Lakoff, G., Johnson, M., 1999. Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought. Basic Books, New York.
Loomis, J., Blascovich, J., Beall, A., 1999. Immersive virtual environment technology as a basic research tool in psychology. Behavior Research Methods 31, 557–564.
Moher, T., 2006. Embedded phenomena: supporting science learning with classroom-sized distributed simulations. In: Proc. CHI. ACM Press, pp. 691–700.
Patten, J., Ishii, H., 2007. Mechanical Constraints as Computational Constraints in Tabletop Tangible Interfaces. In: Proceedings of CHI'07. ACM, NY, pp. 809–818.
Raffle, H.S., Parkes, A.J., Ishii, H., 2004. Topobo: a constructive assembly system with kinetic memory. CHI '04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 647–654.
Roebuck, K., 2011. Tangible User Interfaces: High-impact Emerging Technology – What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Tebbo Publishing.
Shaer, O., Hornecker, E., 2010. Tangible user interfaces: past, present and future directions. Foundations and Trends in Human-computer Interaction 3 (1–2), 1–137.
Slotta, J., Cober, R., Fong, C., Lui, M., Tissenbaum, M., Shapiro, B., 2013. "Our world"...as Learning Scientists Who Engage with HCI. Paper for "Human-computer Interaction and the Learning Sciences" workshop. CSCL 2013, Madison, WI.
Ullmer, B., Ishii, H., 2001. Emerging frameworks for tangible user interfaces. In: Carroll, J.M. (Ed.), Human-computer Interaction in the New Millennium. Addison-Wesley Publishing, Reading, MA, pp. 579–601.
Ullmer, B., Ishii, H., Jacob, R., 2005. Token + constraint systems for tangible interaction with digital information. ACM Transactions on Computer-human Interaction 12 (1), 81–118.
Underkoffler, J., Ishii, H., 1998. Illuminating light: an optical design tool with a luminous-tangible interface. In: Proceedings of CHI98. ACM, NY, pp. 542–549.
Wang, M.J., Kang, J., 2006. Cybergogy of engaged learning through information and communication technology: a framework for creating learner engagement. In: Hung, D., Khine, M.S. (Eds.), Engaged Learning with Emerging Technologies. Springer Publishing, New York, pp. 225–253.
Wilson, M., 2002. Six views of embodied cognition. Psychonomic Bulletin & Review 9 (4), 625–636.
Yoon, C., Kim, S., 2007. Convenience and TAM in a ubiquitous computing environment: the case of wireless LAN. Electronic Commerce Research and Applications 6 (1), 102–112.
York, J., Pendharkar, P.C., 2004. Human–computer interaction issues for mobile computing in a variable work context. International Journal of Human–computer Studies 60, 771–797.
Zahavi, D., 2003. Husserl's Phenomenology. Stanford University Press, Stanford.
Zuckerman, O., Arida, S., Resnick, M., 2005. Extending tangible interfaces for education: digital montessori-inspired manipulatives. In: Proceedings of CHI '05. ACM, NY, pp. 859–868.

Taste and Smell, Psychology of

Edmund T Rolls, Oxford Centre for Computational Neuroscience, Oxford, UK; and University of Warwick, Coventry, UK

Published by Elsevier Ltd.
This article is a revision of the previous edition article by J.C. Stevens, volume 23, pp. 15440–15442, © 2001, Elsevier Ltd.

Abstract

There are five types of taste receptor cell, sweet, salt, bitter, sour, and umami (protein taste). There are 1000 olfactory receptor genes each specifying a different type of receptor each for a set of odors. Tastes are primary, unlearned, rewards and punishers, and are important in emotion. Pheromones and some other olfactory stimuli are primary reinforcers, but for many odors the reward value is learned by stimulus–reinforcer association learning. The primary taste cortex in the anterior insula provides separate and combined representations of the taste, temperature, and texture (including fat texture) of food in the mouth independently of hunger and thus of reward value and pleasantness. One synapse on, in the orbitofrontal cortex, these sensory inputs are for some neurons combined by learning with olfactory and visual inputs, and these neurons encode food reward value in that they only respond to food when hungry, and in that activations correlate with subjective pleasantness. Cognitive factors, including word-level descriptions, and attention, modulate the representation of the reward value of taste, odor, and flavor in the orbitofrontal cortex and a region to which it projects, the anterior cingulate cortex. Further, there are individual differences in the representation of the reward value of food in the orbitofrontal cortex. Overeating and obesity are related in many cases to an increased reward value of the sensory inputs produced by foods, and their modulation by cognition and attention that override existing satiety signals.

Rapid advances have been made recently in understanding the receptors for taste and smell, the neural systems for taste and smell, the separation of sensory from hedonic processing of taste and smell, and how taste and smell and also the texture of food are important in the palatability of food and appetite control. Emphasis is placed on these advances.

Taste and Olfactory Receptors

Taste receptors. There are receptors on the tongue for sweet, salt, bitter, sour, and the fifth taste, umami as exemplified by monosodium glutamate (Chandrashekar et al., 2006; Chaudhari and Roper, 2010). Umami taste is found in a diversity of foods rich in glutamate like fish, meat, human mothers' milk, tomatoes and some vegetables, and is enhanced by some ribonucleotides (including inosine and guanosine nucleotides), which are present in, for example, meat and some fish. The mixture of these umami components, which act synergistically at the receptor, underlies the rich taste characteristic of many cuisines (Rolls, 2009).

Olfactory receptors. There are approximately 1000 different types of odor receptor each coded for by a different gene (Buck and Bargmann, 2013). This is a significant proportion of the total number of genes in humans, which is in the order of 30 000. This is a simple but genetically expensive way to build specificity into a sensory system. In humans, only approximately one-third of these receptor types are used, with stop codes in the majority of the olfactory receptor genes. This may reflect the greater reliance of primates including humans on stimuli in other sensory modalities, especially vision and hearing.

In addition to this main olfactory system using the main olfactory epithelium and its connections to the main olfactory bulb, there is an accessory olfactory system with its peripheral receptors located in the vomeronasal organ (VNO), which projects to the accessory olfactory bulb. This system is involved in the detection of pheromones, present, for example, in urinary volatiles, which are used for mate attraction, territory marking, etc (Wyatt, 2014). The VNOs of humans and Old World monkeys are nonfunctional, although pheromones or pheromone-like substances acting through the main olfactory system may be involved in social and reproductive behavior in these species.

Taste, Smell, Reward, and Emotion

Emotions can be defined as states elicited by rewards and punishers (Rolls, 2014). A reward is any stimulus or event for which an animal (and that includes humans) will work for (i.e., learn and perform an action instrumental in obtaining the reward), and a punisher is any stimulus or event that an animal to work to escape from or avoid. All tastes are primary, unlearned, reinforcers. For example, animals show that a sweet taste (if hungry) and a salt taste (if salt deprived) are rewards the first time that taste is made available. Animals will similarly reject bitter tastes (which are generally produced by plants for exactly that function). Taste is thus very important in emotion, for it serves as an unlearned reward or punisher. Other primary reinforcers include soft pleasant touch, and pain, with many other examples provided in *Emotion and Decision-Making Explained* (Rolls, 2014).

Emotions, as states elicited by instrumental reinforcers (e.g., a warm pleasant feeling produced by a soft caressing touch), can thus be seen to provide an evolutionary adaptive value account of why genes, for example, taste receptor genes, code for the reward or punisher value of some stimuli. Genes do this

in their own selfish interests, and making a particular taste such as sweet rewarding is very adaptive, for this is much simpler and more flexible than a gene that tries to specify a response such as climbing a tree, picking an apple, and placing it in the mouth. By specifying rewards and punishers, the genes that provide a foundation for emotion are thus acting in their own selfish interests (Rolls, 2014). Tastes are thus important in many emotional and related states, such as the pleasantness of the taste of sweet.

Some olfactory stimuli are primary reinforcers, including pheromones, and some odors such as perhaps the smell of ripe fruit, and the smell of rotting food. An interesting example is that the odors of some individuals may be pleasant because of major histocompatibility complex genes, which specify olfactory receptors that signal reward produced be the smell of another individual with different immune system. The adaptive value is that offspring produced with an individual with a different immune system may have more diverse immune systems, and thereby greater resistance to disease. However, many odors only become reinforcers by a learned, stimulus–reinforcer association. An example is that the odor of the cheese brie may be initially unpleasant, but may become pleasant after learned association with its taste and fatty texture. (Fatty texture may be a primary reinforcer because it is an indicator of a high-energy value food.)

These concepts help to show how taste and smell are closely linked to reward value and emotion. A broad perspective on emotion and the pleasantness of sensory stimuli including taste and smell is provided in *Emotion and Decision-Making Explained* (Rolls, 2014).

We now turn to consider how taste and smell, and their functions related to reward value and subjective pleasantness, are implemented in the brain, again an area in which rapid progress is being made.

Taste Processing in the Brain

Pathways

A diagram of the taste and related olfactory, somatosensory, and visual pathways in primates is shown in **Figure 1**, as the focus of this article is on processing relevant to humans. (The pathways involved in taste in rodents are quite different, with subcortical connections to a pontine taste area, with no granular orbitofrontal cortex so that other areas are more involved in the processing, and with hunger modulating taste processing even in the brain stem (Rolls, 2014).) The multimodal convergence that enables single neurons to respond to different combinations of taste, olfactory, texture, temperature, and visual inputs to represent different flavors produced often by new combinations of sensory input is a theme of recent research that will be described, as is the separation in primates including humans of processing of what the taste is (in the primary taste cortex) from its reward value and pleasantness (in the orbitofrontal cortex and anterior cingulate cortex).

The Primary Taste Cortex

Rolls and colleagues have shown that the primary taste cortex in the primate anterior insula and adjoining frontal operculum contains not only taste neurons tuned to sweet, salt, bitter, sour, and umami as exemplified by monosodium glutamate, but also other neurons that encode oral somatosensory stimuli including viscosity, fat texture, temperature, and capsaicin. Some neurons in the primary taste cortex respond to particular combinations of taste and oral texture stimuli, but do not respond to olfactory stimuli or visual stimuli such as the sight of food. Neurons in the primary taste cortex do not represent the reward value of taste, that is the appetite for a food, in that their firing is not decreased to zero by feeding the taste to satiety (Rolls, 2012, 2014).

The Secondary Taste Cortex

A secondary cortical taste area in primates was discovered by Rolls et al. (1990) in the orbitofrontal cortex, extending several mm in front of the primary taste cortex. Different neurons in this region respond not only to each of the four classical prototypical tastes sweet, salt, bitter, and sour, but also to umami tastants such as glutamate (which is present in many natural foods such as tomatoes, mushrooms, milk) and inosine monophosphate (which is present in meat and some fish such as tuna). This evidence, taken together with the identification of glutamate taste receptors (Zhao et al., 2003; Maruyama et al., 2006), leads to the view that there are five prototypical types of taste information channels, with umami contributing, often in combination with corresponding olfactory inputs, to the flavor of protein. In addition, some neurons respond to water and others to somatosensory stimuli including astringency as exemplified by tannic acid, and capsaicin (Rolls, 2012, 2014).

The Pleasantness of the Taste of Food, Sensory-Specific Satiety, and the Effects of Variety on Food Intake

The modulation of the reward value of a sensory stimulus such as the taste of food by motivational state, for example, hunger is one important way in which motivational behavior is controlled. The subjective correlate of this modulation is that food tastes pleasant when hungry, and tastes hedonically neutral when it has been eaten to satiety. Following Edmund Rolls' discovery of sensory-specific satiety revealed by the selective reduction in the responses of lateral hypothalamic neurons to a food eaten to satiety (Rolls, 1981; Rolls et al., 1986), it has been shown that this is implemented in a region that projects to the hypothalamus, the orbitofrontal (secondary taste) cortex, for the taste, odor, and sight of food (Rolls et al., 1989; Critchley and Rolls, 1996).

This evidence shows that the reduced acceptance of food that occurs when food is eaten to satiety, the reduction in the pleasantness of its taste and flavor, and the effects of variety to increase food intake, are produced in the orbitofrontal cortex, but not at earlier stages of processing where the responses reflect factors such as the intensity of the taste, which is little affected by satiety (Rolls and Grabenhorst, 2008; Rolls, 2012, 2014). In addition to providing an implementation of sensory-specific satiety (probably by habituation of the synaptic afferents to orbitofrontal

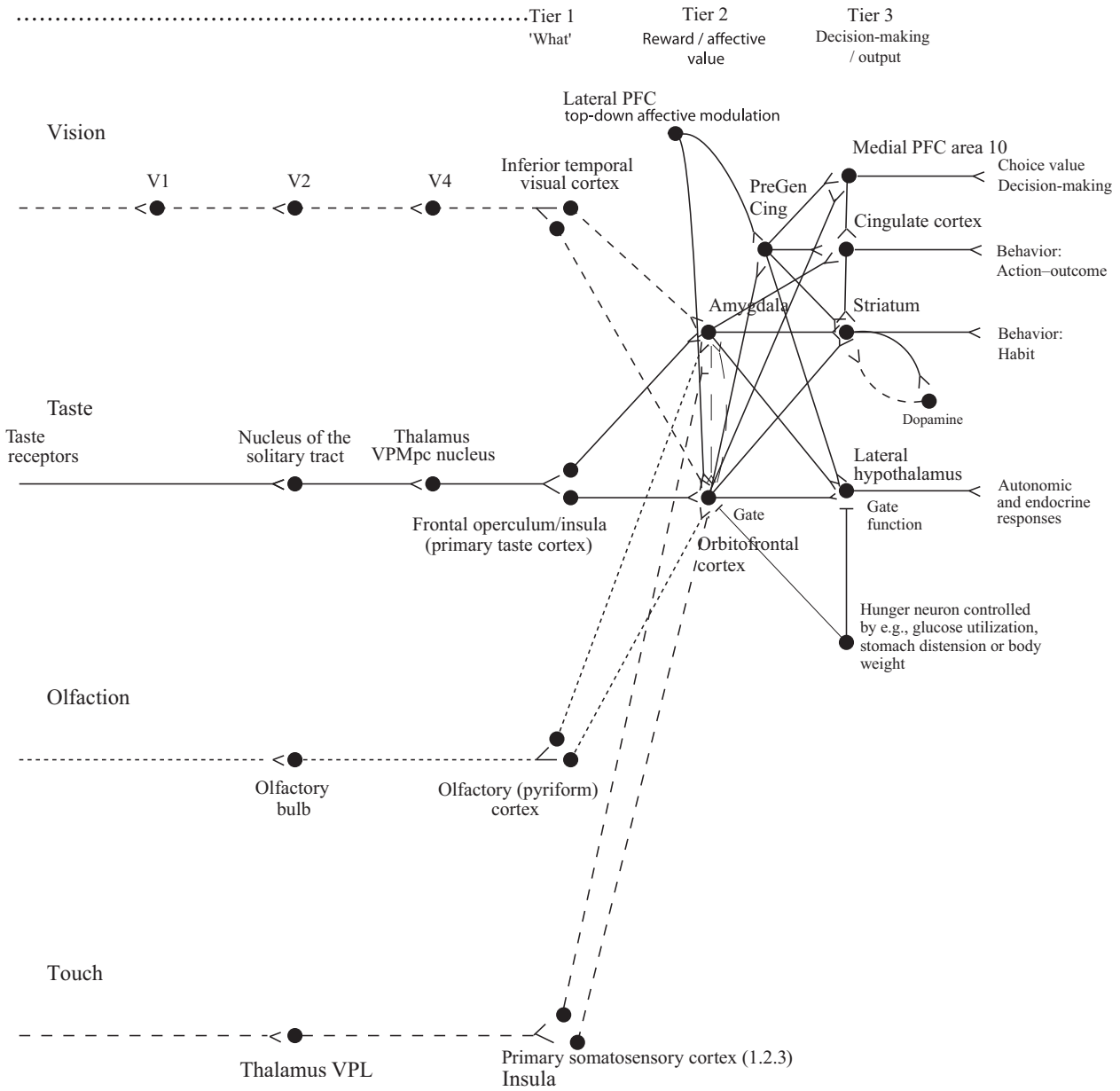

Figure 1 Schematic diagram showing some of the gustatory, olfactory, visual, and somatosensory pathways to the orbitofrontal cortex, and some of the outputs of the orbitofrontal cortex, in primates. The secondary taste cortex and the secondary olfactory cortex are within the orbitofrontal cortex. V1 – primary visual cortex. V4 – visual cortical area V4. PreGen Cing – pregenual cingulate cortex. 'Gate' refers to the finding that inputs such as the taste, smell, and sight of food in some brain regions only produce effects when hunger is present (Rolls, E.T., 2005. Emotion Explained. Oxford University Press, Oxford.). The column of brain regions including and below the inferior temporal visual cortex represents brain regions in which what stimulus is present is made explicit in the neuronal representation, but not its reward or affective value, which are represented in the next tier of brain regions, the orbitofrontal cortex and amygdala, and in the anterior cingulate cortex. In areas beyond these such as medial prefrontal cortex area 10, choices or decisions about reward value are taken, with the mechanisms described elsewhere (Rolls, E.T., 2008. Memory, Attention, and Decision-making: A Unifying Computational Neuroscience Approach. Oxford University Press, Oxford; Rolls, E.T., Deco, G., 2010. The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function. Oxford University Press, Oxford; Rolls, E.T., 2014. Emotion and Decision-making Explained. Oxford University Press, Oxford.). Medial PFC area 10 – medial prefrontal cortex area 10; VPMpc – ventralposteromedial thalamic nucleus (tasolfpaths6i.eps).

neurons with a time course of the order of the length of a course of a meal), it is likely that visceral and other satiety-related signals reach the orbitofrontal cortex (as indicated in **Figure 1**) (from the nucleus of the solitary tract, via thalamic and possibly hypothalamic nuclei) and there modulate the representation of food, resulting in an output that reflects the reward (or appetitive) value of each food.

The Representation of Flavor: Convergence of Olfactory, Taste, and Visual Inputs in the Orbitofrontal Cortex

Taste and olfactory pathways are brought together in the orbitofrontal cortex where flavor is formed by learned associations at the neuronal level between these inputs (see **Figure 1**). Visual inputs also become associated by learning in the orbitofrontal cortex with the taste of food to represent the sight of food and contribute to flavor. The visual and olfactory as well as the taste inputs represent the reward value of the food, as shown by sensory-specific satiety effects (Rolls, 2012, 2014).

The Texture of Food, Including Fat Texture

Some orbitofrontal cortex neurons have oral texture-related responses that encode parametrically the viscosity of food in the mouth (shown using a methyl cellulose series in the range 1–10 000 cP), others independently encode the particulate quality of food in the mouth, produced quantitatively, for example, by adding 20–100 μm microspheres to methyl cellulose (Rolls et al., 2003a), and others encode the oral texture of fat. Somatosensory signals that transmit information about capsaicin (chili) and astringency are also reflected in neuronal activity in these cortical areas.

Imaging Studies in Humans

Taste

In humans, it has been shown in neuroimaging studies using functional magnetic resonance imaging (fMRI) that taste activates an area of the anterior insula/frontal operculum, which is probably the primary taste cortex, and part of the orbitofrontal cortex, which is probably the secondary taste cortex. The primary taste cortex in the anterior insula of humans represents the identity and intensity of taste in that activations there correlate with the subjective intensity of the taste, and the orbitofrontal and anterior cingulate cortex represents the reward value of taste, in that activations there correlate with the subjective pleasantness of taste (Grabenhorst and Rolls, 2008; Grabenhorst et al., 2008).

We also found activation of the human amygdala by the taste of glucose. Indeed, the human amygdala is as much activated by the affectively pleasant taste of glucose as by the affectively negative taste of NaCl, and thus provided evidence that the human amygdala is not especially involved in processing aversive as compared to rewarding stimuli.

Odor

In humans, in addition to activation of the pyriform (olfactory) cortex, there is strong and consistent activation of the orbitofrontal cortex by olfactory stimuli, and this region appears to represent the pleasantness of odor. Further, pleasant odors tend to activate the medial, and unpleasant odors the more lateral, orbitofrontal cortex (Rolls et al., 2003b), adding to the evidence that it is a principle that there is a hedonic map in the orbitofrontal cortex, and also in the anterior cingulate cortex, which receives inputs from the orbitofrontal cortex (Rolls and Grabenhorst, 2008; Grabenhorst and Rolls, 2011). The primary olfactory (pyriform) cortex represents the identity and intensity of odor in that activations there correlate with the subjective intensity of the odor, and the orbitofrontal and anterior cingulate cortex represents the reward value of odor, in that activations there correlate with the subjective pleasantness of odor (Grabenhorst et al., 2007; Rolls and Grabenhorst, 2008; Grabenhorst and Rolls, 2011).

Olfactory-Taste Convergence to Represent Flavor, and the Influence of Satiety

Supradditive effects indicating convergence and interactions were found for taste (sucrose) and odor (strawberry) in the orbitofrontal and anterior cingulate cortex, and activations in these regions were correlated with the pleasantness ratings given by the participants (de Araujo et al., 2003). These results provide evidence on the neural substrate for the convergence of taste and olfactory stimuli to produce flavor in humans, and where the pleasantness of flavor is represented in the human brain.

McCabe and Rolls (2007) have shown that the convergence of taste and olfactory information appears to be important for the delicious flavor of umami. They showed that when glutamate is given in combination with a consonant, savory, odor (vegetable), the resulting flavor can be much more pleasant than the glutamate taste or vegetable odor alone, and that this reflected activations in the pregenual cingulate cortex and medial orbitofrontal cortex. The principle is that certain sensory combinations can produce very pleasant food stimuli, which may of course be important in driving food intake; and that these combinations are formed in the brain far beyond the taste or olfactory receptors (Rolls, 2009).

To assess how satiety influences the brain activations to a whole food, which produces taste, olfactory, and texture stimulation, we measured brain activation by whole foods before and after the food is eaten to satiety. The foods eaten to satiety were either chocolate milk, or tomato juice. A decrease in activation by the food eaten to satiety relative to the other food was found in the orbitofrontal cortex (Kringelbach et al., 2003) but not in the primary taste cortex. This study provided evidence that the pleasantness of the flavor of food and sensory-specific satiety, are represented in the orbitofrontal cortex.

Cognitive and Selective Attentional Effects on Representations of Smell, Taste, and Flavor

To what extent does cognition influence the hedonics of food-related stimuli, and How far down into the sensory system does the cognitive influence reach? To address this, we performed a fMRI investigation in which the delivery of a standard test odor (isovaleric acid combined with cheddar cheese odor, presented orthonasally using an olfactometer) was paired with a descriptor word on a screen, which on different trials was 'Cheddar cheese' or 'Body odor.' Participants rated the affective value of the test odor as significantly more pleasant when labeled 'Cheddar cheese' than when labeled 'Body odor,' and these effects reflected activations in the medial orbitofrontal cortex/rostral anterior cingulate cortex that had correlations with the pleasantness

ratings (de Araujo et al., 2005). The implication is that cognitive factors can have profound effects on our responses to the hedonic and sensory properties of food, in that these effects are manifest quite far down into sensory processing, so that hedonic representations of odors are affected (de Araujo et al., 2005). Similar cognitive effects and mechanisms have now been found for the taste and flavor of food, where the cognitive word-level descriptor was, for example, 'rich delicious flavor' and activations to flavor were increased in the orbitofrontal cortex and regions to which it projects including the pregenual cingulate cortex and ventral striatum, but were not influenced in the insular primary taste cortex where activations reflected the intensity (concentration) of the stimuli (Grabenhorst et al., 2008).

In addition, we have found that with taste, flavor, and olfactory food-related stimuli, selective attention to pleasantness modulates representations in the orbitofrontal cortex, whereas selective attention to intensity modulates activations in areas such as the primary taste cortex (Grabenhorst and Rolls, 2008; Rolls et al., 2008). Thus, depending on the context in which tastes and odors are presented and whether affect is relevant, the brain responds to taste and odor differently. These findings show that when attention is paid to affective value, the brain systems engaged to represent the stimulus are different from those engaged when attention is directed to the physical properties of a stimulus such as its intensity. This differential biasing by prefrontal cortex attentional mechanisms (Grabenhorst and Rolls, 2010; Ge et al., 2012) of brain regions engaged in processing a sensory stimulus depending on whether the cognitive demand is for affect-related versus more sensory-related processing may be an important aspect of cognition and attention, which have implications for how strongly the reward system is driven by food, and thus for eating and the control of appetite (Grabenhorst and Rolls, 2008, 2011; Rolls et al., 2008).

There are individual differences in the representation of the reward value of food, for example of chocolate, in the orbitofrontal cortex (Rolls and McCabe, 2007). Overeating and obesity may be related in many cases to an increased reward value of the sensory inputs produced by foods, and their modulation by cognition and attention, that override existing satiety signals. The findings described here thus have implications for understanding what causes obesity (Rolls, 2012, 2014).

Acknowledgments

This research was supported by the Medical Research Council. The participation of many colleagues in the studies cited is sincerely acknowledged.

See also: Emotion, Neural Basis of; Food Preferences, Psychology and Physiology of; Hunger and Eating, Neural Basis of; Olfactory System.

Bibliography

de Araujo, I.E.T., Rolls, E.T., Kringelbach, M.L., McGlone, F., Phillips, N., 2003. Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. European Journal of Neuroscience 18, 2059–2068.

de Araujo, I.E.T., Rolls, E.T., Velazco, M.I., Margot, C., Cayeux, I., 2005. Cognitive modulation of olfactory processing. Neuron 46, 671–679.

Buck, L., Bargmann, C.I., 2013. Smell and taste: the chemical senses. In: Kandel, E.R., Schwartz, J.H., Jessell, T.H., Siegelbaum, S.A., Hudspeth, A.J. (Eds.), Principles of Neural Science, fifth ed. McGraw-Hill, New York, pp. 712–742.

Chandrashekar, J., Hoon, M.A., Ryba, N.J., Zuker, C.S., 2006. The receptors and cells for mammalian taste. Nature 444, 288–294.

Chaudhari, N., Roper, S.D., 2010. The cell biology of taste. The Journal of Cell Biology 190, 285–296.

Critchley, H.D., Rolls, E.T., 1996. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. Journal of Neurophysiology 75, 1673–1686.

Ge, T., Feng, J., Grabenhorst, F., Rolls, E.T., 2012. Componential Granger causality, and its application to identifying the source and mechanisms of the top-down biased activation that controls attention to affective vs sensory processing. NeuroImage 59, 1846–1858.

Grabenhorst, F., Rolls, E.T., 2008. Selective attention to affective value alters how the brain processes taste stimuli. European Journal of Neuroscience 27, 723–729.

Grabenhorst, F., Rolls, E.T., 2010. Attentional modulation of affective vs sensory processing: functional connectivity and a top-down biased activation theory of selective attention. Journal of Neurophysiology 104, 1649–1660.

Grabenhorst, F., Rolls, E.T., 2011. Value, pleasure, and choice in the ventral prefrontal cortex. Trends in Cognitive Sciences 15, 56–67.

Grabenhorst, F., Rolls, E.T., Bilderbeck, A., 2008. How cognition modulates affective responses to taste and flavor: top-down influences on the orbitofrontal and pregenual cingulate cortices. Cerebral Cortex 18, 1549–1559.

Grabenhorst, F., Rolls, E.T., Margot, C., da Silva, M.A.A.P., Velazco, M.I., 2007. How pleasant and unpleasant stimuli combine in different brain regions: odor mixtures. Journal of Neuroscience 27, 13532–13540.

Kringelbach, M.L., O'Doherty, J., Rolls, E.T., Andrews, C., 2003. Activation of the human orbitofrontal cortex to a liquid food stimulus is correlated with its subjective pleasantness. Cerebral Cortex 13, 1064–1071.

Maruyama, Y., Pereira, E., Margolskee, R.F., Chaudhari, N., Roper, S.D., 2006. Umami responses in mouse taste cells indicate more than one receptor. Journal of Neuroscience 26, 2227–2234.

McCabe, C., Rolls, E.T., 2007. Umami: a delicious flavor formed by convergence of taste and olfactory pathways in the human brain. European Journal of Neuroscience 25, 1855–1864.

Rolls, E.T., 1981. Central nervous mechanisms related to feeding and appetite. British Medical Bulletin 37, 131–134.

Rolls, E.T., 2005. Emotion Explained. Oxford University Press, Oxford.

Rolls, E.T., 2008. Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach. Oxford University Press, Oxford.

Rolls, E.T., 2009. Functional neuroimaging of umami taste: what makes umami pleasant. American Journal of Clinical Nutrition 90, 803S–814S.

Rolls, E.T., 2012. Taste, olfactory, and food texture reward processing in the brain and the control of appetite. Proceedings of the Nutrition Society 71, 488–501.

Rolls, E.T., 2014. Emotion and Decision-Making Explained. Oxford University Press, Oxford.

Rolls, E.T., McCabe, C., 2007. Enhanced affective brain representations of chocolate in cravers vs non-cravers. European Journal of Neuroscience 26, 1067–1076.

Rolls, E.T., Grabenhorst, F., 2008. The orbitofrontal cortex and beyond: from affect to decision-making. Progress in Neurobiology 86, 216–244.

Rolls, E.T., Deco, G., 2010. The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function. Oxford University Press, Oxford.

Rolls, E.T., Sienkiewicz, Z.J., Yaxley, S., 1989. Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. European Journal of Neuroscience 1, 53–60.

Rolls, E.T., Yaxley, S., Sienkiewicz, Z.J., 1990. Gustatory responses of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. Journal of Neurophysiology 64, 1055–1066.

Rolls, E.T., Verhagen, J.V., Kadohisa, M., 2003a. Representations of the texture of food in the primate orbitofrontal cortex: neurons responding to viscosity, grittiness and capsaicin. Journal of Neurophysiology 90, 3711–3724.

Rolls, E.T., Kringelbach, M.L., de Araujo, I.E.T., 2003b. Different representations of pleasant and unpleasant odors in the human brain. European Journal of Neuroscience 18, 695–703.

Rolls, E.T., Murzi, E., Yaxley, S., Thorpe, S.J., Simpson, S.J., 1986. Sensory-specific satiety: food-specific reduction in responsiveness of ventral forebrain neurons after feeding in the monkey. Brain Research 368, 79–86.

Rolls, E.T., Grabenhorst, F., Margot, C., da Silva, M.A.A.P., Velazco, M.I., 2008. Selective attention to affective value alters how the brain processes olfactory stimuli. Journal of Cognitive Neuroscience 20, 1815–1826.

Wyatt, T.D., 2014. Pheromones and Animal Behavior, second ed. Cambridge University Press, Cambridge.

Zhao, G.Q., Zhang, Y., Hoon, M.A., Chandrashekar, J., Erlenbach, I., Ryba, N.J., Zuker, C.S., 2003. The receptors for mammalian sweet and umami taste. Cell 115, 255–266.

Tattoos and Body Modification

Mary Kosut, State University of New York, Purchase, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines tattooing as an ancient and contemporary form of body inscription and modification. It frames the popularity of tattooing in the Western World, in relation to other modern body modifications and within the landscape of consumer culture. Showing parallels between the motivations and meanings of contemporary tattoo practices, a historical overview of tattoos in premodern and non-Western cultures is presented. The divergent ways in which scholars have approached tattooing is examined, juxtaposing research in the behavioral sciences that links tattoos to deviant and antisocial behaviors with ethnographic and social–scientific studies that place tattoos within cultural, symbolic, esthetic, and prosocial frameworks.

Introduction

All humans undergo physical transformation as part of becoming and being with other bodies in the material and social world. The body is the intersection between self and society; it is a medium through which we negotiate life; and a conveyer of personal and social meanings. As a simultaneously biological and cultural entity, our body immediately succumbs to a process of modification at birth, ending when fleshy mortification finally betrays us. Throughout the life course, our bodies grow (for example, in height, muscle mass, and physical and mental dexterity) and they inevitably decline, suffering injuries, unforeseen breakdowns, and malfunctions. The human body is a vehicle of boundless pleasure and unthinkable pain and abjection. While bodily processes and experiences such as aging, illness, and dying are universal, our understanding of these processes and experiences depends upon the individual and the cultural contexts in which we live. Increasingly, in the neoliberal West we are seen as being responsible for what happens to our own bodies, especially with regard to health and beautification (which could be described as the visual evidence of well-being). To some degree, all of us are expected to engage in body inscription and modification, regardless of access to resources and levels of opportunities – particularly in affluent and technologically modernized countries.

Cross-culturally and throughout prehistory and history, people have participated in transitory or permanent corporal modification for sundry reasons. How humans modify their bodies – the where, what, and when – marks people as individuals and identifies individuals with large groups. While there are many public and immediately visible transitory markers that convey the body/self to others – hairstyle, clothing, and jewelry – humans are also drawn to permanent body modifications. Social science research reveals a range of ancient body modification customs, including earlobe and penis elongation, foot binding, head molding, neck stretching, branding, and scarification.

However, tattooing is arguably the oldest and most widely used form of permanent body modification. From a historical and anthropological perspective, it can be argued that where there is skin and opportunity, you will find tattooed bodies. Many social scientists are exploring the question of why tattoos have become so popular in the West in the late twentieth and early twenty-first centuries. While there are many answers – tatooos have become more socially acceptable, they are a means to visualize individuality, a way to display political or social affiliations – the most comprehensive answer is straightforward: tattoos are effective. Tattoos are tremendously functional on both micro- and macro-social levels.

Within different cultural contexts, tattoos may be viewed and experienced positively, but there are others who will likely always think of tattoos pejoratively, not as self-modification but as self-mutilation. This terminology is highly significant as the former term (modification) indicates that a person is engaged in body/self work that is reasonable to some extent; while the latter is a behavior that is linked to mental or physical illness and suffering. Modern tattooing is contradictory, currently visible, and popular in mainstream media platforms; however at the same time, it is considered unacceptable, if not categorically harmful. In contemporary Western cultures, tattoos have numerous meanings and are situated within an extensive range of social contexts. Individual motivations for getting a tattoo are extraordinarily varied; and social class, gender, race, sexuality, occupation, age, and geographic location may influence tattoo choices and experiences.

Current cross-cultural tattoo practices affirm the fluidity and malleability of the body in contemporary consumer culture, where it is taken for granted that those with a degree of economic capital will likely engage in some sort of work on the body/self in the name of maintenance, improvement, health, beauty, and the pursuit of youthfulness (see Shilling, 1993; Sweetman, 1999). The increase in the popularity of tattoos coincides with the increasing popularity of other forms of body modification. Ultimately, modern, privileged bodies are understood as a site of potential transformation and adjustment, whether through dietary regimes and beautification rituals or as more nonnormative and potentially stigmatizing visible large-format tattoos. In some cases, people engage in both acceptable (breast implants), and socially objectionable modifications (septum piercing). The body is not a static object that we are given, but rather, a flexible and changeable medium that we can rework and potentially recreate into a more attractive or functional form.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64027-8

Although tattooing is a permanent and decorative form of body modification with a lengthy history, it should be framed within a contemporary landscape of larger technologies and consumer practices including a sharp increase in semi-permanent and permanent cosmetic procedures like botox injections, laser-hair removal, breast implants, liposuction, and other invasive and potentially risky surgeries. However, tattoos should not be lumped with trendy transitory and sartorial forms of expressions. Because the act of getting tattooed typically encompasses three distinct stages of varying length (preplanning, receiving the tattoo, and an after-care regime and healing process) it cannot be compared to the purchase of a pair of sneakers or having your hair dyed blue, no matter how reflective and discriminating the consumer may be. This is because a tattoo is a consumer product whose consumption cannot be divorced from its mode of production. The idiosyncratic act of tattooing, permanently reinscribes the living body – a thinking, breathing, sweating, and wrinkling entity – with a type of symbolic agency that is ongoing and embodied.

Prehistoric and Non-Western Tattoo Practices in the Historical Context

The recent history of tattooing is directly influenced by globalization and the advent of new media technologies – tattoo blogs, websites, and online galleries are now ubiquitous across the Internet. Idiosyncratic tattoo traditions, philosophies, and styles are virtually shared and displayed, fertilizing new hybrid tattoos born from a culture of appropriation. However, ancient tattoo practices demonstrate that cultural forms have always migrated across continents and societies, albeit at a much slower pace. Tattoos clearly played a significant role within many premodern, non-Western societies. Anthropological studies document a wide range of practices chronicling the symbolic significance of tattoos and their functions. Significantly, the meanings and uses of ancient tattoo practices share many commonalities with contemporary tattoo customs. Modern-day tattoo narratives suggest that there are certain traditions, for example, using the tattoo to commemorate a rite of passage or a major accomplishment or to memorialize the death of a significant other, that transcends historical and cultural boundaries.

The earliest tattooed bodies date to roughly 6000 BCE – the oldest and arguably most well-known (and well-studied) is the mummified 'Iceman', discovered in the Ötztal Alps on the Italy–Austria border in 1991, dated 5300 BCE. The mummified body is covered in 57 tattoos including lines scored in groups on his lower back and ankles, a cross behind his right knee and two rings circling his left wrist. Anthropologist Lars Krutak speculates that the tattoos may have been connected to the Iceman's medical condition as 80% of the tattoos on his body overlap with Chinese acupuncture points. Krutak (1999) makes a strong case for the medicinal usage of tattoos by the indigenous Yupik peoples of St. Lawrence Island in Western Alaska. This research expands our search toward understanding the possible reasons why people tattooed themselves in early human history, and how tattoos may have functioned in relationship to health and embodiment. In addition to having spiritual and symbolic purposes, tattoos may have served in easing pain, promoting health, and curing illness. This connection between tattoos and curative medicinal procedures offers an approach that is at odds with studies that posit tattoos as a sign of pathology or deviant behavior (see Parry, 1933; Lombroso, 2006; Favazza, 1996; Stirn and Hinz, 2008; among others).

Archeologists and anthropologists have also found evidence of tattooing among the indigenous peoples of North and South America, Africa, and throughout the Pacific region. In fact, the oldest material evidence of tattoo in the world is from a mummy dated to 6000 BCE from the Chinchorro culture of South America. However, there has been much research centered on the prevalence of tattoos in Ancient Egypt, possibly due to the significant number of well-preserved material artifacts. There have been numerous tattooed female mummies dating backing to 2000 BCE, as well as female figurines and tomb scenes with tattooed bodies. Archeologist Joann Fletcher theorizes that although it has long been assumed that tattoos were the mark of prostitutes or were meant to protect women against sexually transmitted diseases, the tattooing of ancient Egyptian women "had a therapeutic role and functioned as a permanent form of amulet during the very difficult time of pregnancy and birth"(http://www.smithsonianmag.com/history-archaeology/tattoo.html, accessed 20.12.13). This linkage between tattoos, fertility, and reproductive processes is comparable to tattoos as a form of acupuncture, as the tattoos either literally or symbolically helped their bearers to come to terms with illness and the tremendous physical vulnerabilities that characterized premodern embodiment.

While the specific meanings of tattoo in preindustrialized cultures are based upon academic speculation, it is apparent that the tattoo was an integral aspect of collective social life for some cultures. From the perspective of sociologist Emile Durkheim (1995), tattooing is emblematic of collective consciousness – a tangible mark of society upon the individual body. For example, in his study of Polynesian tattooing, anthropologist Alfred Gell (1993) asserts that tattooing played an integral part in the organization and functioning of major institutions (politics, warfare, religion, and so on), and that tattoos were visual documents of the social systems in which tattooed people were embedded. In addition to larger macro-social functions, tattoos were believed to be a sign of beauty and individuality in Poylnesian cultures (Gell, 1993). They were also employed to mark rites of passage to adulthood, indicate political status, and commemorate major events like funerals and weddings (Kuwahara, 2005).

The Maori of New Zealand produced particularly sophisticated and elaborate designs referred to as 'Moko', which are differentiated from other Polynesian tattooing forms because of its symbolic significance, stylistic elements, and excruciatingly painful method of application by use of hand tools (Rubin, 1988; Sanders, 1989; Freidlander and King, 1999). The act of receiving a tattoo by this method required an ability to endure extend hours of physical discomfort. Enduring pain and pushing the body's physical threshold is a key element and a common practice in highly meaningful cultural and spiritual rituals. Some researchers have suggested that the purpose of Moko was to indicate status or rank (Brain, 1979; Sanders, 1989); while others have proposed that the meanings of Moko may incorporate existential characteristics, proposing

that it may have allowed individuals to achieve 'completeness' because it linked them to their ancestors and enabled an 'ideal social personality' (Gathercole, 1988).

Like other societies on the Pacific Rim, Japan has a long and rich tradition of tattooing (McCallum, 1988; Schildkrout, 2004). Artifacts such as stone tools and clay figurines suggest that tattooing may have been present as early as the Jomon period (10 000–300 BCE). While very early accounts of Japanese tattoo are highly speculative, texts from the Kofun period (CE 300–600) indicate that people of higher status had tattoos, as did criminals and thieves. Modern Japanese literary texts suggest parallels between American and Japanese tattoo cultures. Although extremely divergent in terms of esthetics (iconography and stylization), in both cultures tattoo went through lengthy periods of criminality and marginalization.

In Japanese, tattoo is known as 'irezumi' – 'ire' meaning to insert and 'zumi' referring to the pigments used. Irezumi, which emerged and flourished within the Edo period (AD 1600–1868), is distinguished by imagery that includes koi, dragons, warriors, and other mythical creatures surrounded by stylized aquatic and floral patterns such as waves and chrysanthemums (Kitamura, 2003). Still practised today and revered as an art form by many tattoo historians and aficionados, vividly elaborate irezumi designs often encompass large areas, and in some cases, the entire body. Regardless of its widely acclaimed esthetic qualities, just as tattooing was prohibited and condemned by government authorities in the United States during most of the twentieth century, during the Edo period the Tokugawa Bakufu (military government) prohibited tattooing. In both cultures the tattoo went underground, migrating to the bodies of the urban working classes. Despite its peripheral social location, it did not disappear completely in either case. Just as the tattoo was a macho emblem for some working-class men and women in the United States, irezumi was quite popular among firefighters and tradesmen who were known for their rowdy behavior.

By the nineteenth and twentieth centuries in Japan, tattoos were deemed immoral and a sign of barbarism and some Japanese criminals were punished by having their faces tattooed (Sanders, 1989; DeMello, 2000). Tattoos became commonly associated with people working in the underground economy who were involved in prostitution, extortion, and petty crime. Likewise, tattoos have a similar history in Russian prisons and within organized crime syndicates. Strictly forbidden legally and normatively, mainly gamblers and members of the Yakuza (Japanese organized crime syndicate) had tattoos until the end of World War II, at which time laws prohibiting tattooing were overturned. Interestingly, Western interest in tattooing revived the practice in Japan and the United States in the decades that followed. Many American and European tattoo artists continue to be greatly influenced by irezumi masters, and Japanese-inspired imagery remains popular among contemporary tattoo aficionados.

Nevertheless, some research assigns colonialism and the seafaring culture as giving rise to the birth of the modern Western tattoo (Sanders, 1989; DeMello, 2000). In the late 1700s, Captain James Cook encountered tattooing while exploring the islands of the South Pacific. He introduced the then exotic practice to European society, reinvesting tattooing with new meanings in the West. Cook was responsible for bringing the Tahitian word 'ta-tu' or 'tatau' (to strike or mark) into common parlance. In Europe and the United States, for most of the eighteenth and nineteenth centuries, tattooing was associated with marginalized populations; side show and circus workers, criminals, sailors, gang members, and the economically underprivileged. Nonetheless, there is evidence to suggest a short-lived tattoo fad within the circles of the cultural elite in Europe and the United States during the 1880s, because some members of polite society got tattoos as symbols of exoticism and distinction. The popularity of the tattoo among the 'leisured class' waned within a decade due to the invention of the electric tattoo machine. At this stage, tattoo became more affordable and accessible to the working- and middle-classes. However, members of the working class and the military continued to get tattooed before, during, and after this period.

It was not until the mid-1960s that a noticeable shift in tattoo culture, commonly referred to as the 'tattoo renaissance', revived the practice and signaled major changes in esthetics, population, and discourse (Rubin, 1988; Sanders, 1989; DeMello, 2000). Major social changes, including the civil rights and women's movements, resulted in shifts in American culture that produced fertile ground for a resurgence of subcultural, anarchistic, and artistic activities. The 1960s counterculture (the feminist, social justice, and antiwar movements) led to a renewed interest in tattoos within a climate of protest and a focus on the body as a site of symbolic dissention. Recognizable symbols such as peace signs, butterflies, marijuana leaves, and so on were used to distinguish rebellious bodies.

Tattoos as Deviant Behavior in Western Society

Western tattooing has always been a signifier of 'outsiderness', often overshadowed by symbolic associations with marginalized groups and subcultures. Most journalistic accounts of contemporary tattoo cultures begin with a familiar hook: tattoos were for 'them' (the disenfranchized, marginalized, incarcerated, impoverished, macho, crude, criminal) but now they are for 'us' (privileged, educated, mainstream, sophisticated, law abiding). Even though tattoos have been cleaned up – symbolically, esthetically, and culturally – and hold more positive cultural capital today than at any point in modern history, they are still tethered to forms of social marginality. Prisoners, gang members, and those with ties to illegal alternate economies are a part of tattoo's early history in the United States. Regardless of how tattoos may be commodified or esthetically elevated, they will never be sterilized and whitewashed to the point of being drained of all of their oppositional power. Even now tattoos semiotically provoke and disturb mainstream norms of body propriety and acceptability.

In the behavioral sciences especially, there is a history of equating the presence of tattoos to psychological disorders and socially deviant behavior (Copes and Forsyth, 1993; Grumet, 1983; Loimer and Werner, 1992). The literature is dominated with studies centering on patients in clinics and those in prison (Newman, 1982; Gittelson et al., 1969; Hamburger, 1966). This highly select group of individuals is typically described as being more likely to exhibit personality disorders than those found within the general population. As many tattoo scholars

have pointed out, the conclusions of reports such as these rely on populations found within these specialized institutions. The tattooed individuals under investigation were already diagnosed and labeled as psychopathological, or were convicted of committing a criminal act. Thus, the association made between tattoos and criminality has been at times spurious.

In the 1960s, studies purported that the presence of tattoos was indicative of homosexual inclinations and suppression of sexual desires (see Post, 1968; Yamamoto et al., 1963). In these cases, one innately and historically deviant outsider (the homosexual) was linked with another (the tattooed). The relationship between tattooing and forms of perceived sexual deviancy continue to attract behavioral scientists. An association between tattoos and certain sexual behaviors has recently received attention, as contemporary researchers have reported that tattoos may be connected with heightened sexually activity, becoming sexually active at a younger age, and having sex outside of committed relationships (Koch et al., 2005; Carroll et al., 2002). This type of framing works well with a more conservative political agenda, especially those advocating for family values and a desire to return to a traditional definition of heterosexual, monogamous marriage as the ideal. The individual body is always subjected to a matrix of powerful social and political discourses.

The motivations for getting a tattoo are also characterized as negative by certain researchers. Studies maintain that people may get tattoos because of antisocial tendencies and psychiatric disorders. For example, it has been suggested that the desire to tattoo oneself may be the result of extraversion, indicated by the presence of low levels of cortical arousal and a need for constant stimulation. Armando Favazza, a cultural psychiatrist, asserts that a psychiatric diagnosis cannot be presumed in persons who engage in tattooing (and also piercing and branding), although as a group, such persons 'probably' (emphasis added) have increased levels of pathology (1996). This diagnosis as it were, may be speculative but nonetheless aids in advancing the idea that there may be something inherently pathological about tattoos.

Within the field of psychology, some studies have focused on behavioral and individual differences between tattooed and nontattooed people (Tate and Shelton, 2008). A related body of research suggests that tattooed students are psychologically more inclined toward behavior that is impulsive, riskier, or sensation-seeking; including the use of illegal drugs and binge-drinking. Where notable differences have been found, the findings are often statistically insignificant and the extent to which such differences have real-world implications has been questioned. The inability to generalize findings may be related to sampling issues.

With few exceptions, in most quantitative studies in which differences between tattooed and nontattooed individuals have been examined, the population has been college students. Given the age cohort, and taking into account the cultural norms of college as a distinctive period of time before adulthood and its responsibilities, it is not surprising that these sample populations would report higher levels of perceived risky and reckless behavior. Sexual exploration, drug use, and alcohol consumption on college campuses is not automatically regarded as anomalous behavior within this institutional context.

Early criminologist, Cesare Lombroso (1836–1909), championed the concept of degeneracy as the central explanation of deviancy and linked tattooing to criminal behavior (1876). According to Lombroso, criminals were innately primitive and atavistic beings with recognizable physical deformities. Tattoo was a physical manifestation of a biologically criminal person – an easily discernable sign of pathology. Throughout much of the twentieth century, clinical psychiatrists and psychologists associated tattoo with mental illness, criminal behavior, self-mutilative behaviors, and sexual deviancy. The presence of tattoos was (and still is) being studied in relation to psychological abnormalities and unhealthful mental states. The image of a tattooed patient confirms that internal illness can manifest itself externally. In this case, the body of the other is constructed from inside and out.

Current Trends in Tattoos and Tattoo Research

In the United States, people tend to be divided as to whether tattoos are inherent mistakes, obdurate esthetic and social blemishes, or semiotic celebrations of self, embodiment, and connections to significant others. For example, a 2013 article about the visibility of tattoos in the academia in the *Chronicle of Higher Education* entitled 'The Inked Academic Body' discusses how American professors increasingly have visible ink (http://chronicle.com/blogs/conversation/2012/10/25/the-inked-academic-body/, accessed 10.01.14). The author, himself a tattooed professor, writes from a tattoo positive standpoint describing his own as a way to control his body and agentically modify it in a way that is nonnormative, but highly functional nonetheless. Tattoos may be read as a positive sign by some students; a kind of social can-opener that may mitigate the power imbalance between student and professor. While seemingly benign in its writing style and assertions, the article sparked a very contentious discussion thread. The online comments drew a sharp line between those who accept or appreciate tattoos as agentic self/body work, and those staunchly opposed to tattoos. One person stated that their heavily tattooed arm (or sleeve) had 'deep meaning'; however, another poster described tattoos as a 'tawdry act of self display'. This polarizing discussion illustrates the sharp divisions and the symbolic power tattoos can generate – prejudices and emotions often flare when tattoos become public.

Regardless of how they are framed, tattoos are today more popular in the United States than at any other time in history. Most polls indicate that roughly one in five American adults between the ages of 21–50 have tattoos, and that this statistic has continued to rise. According to a 2012 Harris Poll, adults aged 30–39 years are most likely to have a tattoo (38%) compared to both those younger (30% of those aged 25–29 years and 22% of those aged 18–24 years) and older (27% of those aged 40–49 years; 11% of those aged 50–64 years; and just 5% of those aged 65 years and older). Women are slightly more likely than men (23 vs 19%) to have one (one in five U.S. adults now has a tattoo. http://www.harrisinteractive.com/NewsRoom/HarrisPolls/tabid/447/mid/1508/articleId/970/ctl/ReadCustom%20Default/Default.aspx, accessed 08.01.14). Interestingly, most of the people polled (84%) said that they did not experience tattoo regret. The popularity of tattoos is certainly not

limited to the United States, as studies show that in England there is also an attraction to tattooing, again with one in five people reporting having inked bodies (do you have a tattoo too? Why Britons love being inked. http://metro.co.uk/2013/02/21/do-you-have-a-tattoo-too-why-britons-love-being-inked-3504642/, accessed 12.01.14).

Tattooing is a multifaceted, embodied phenomenon, not only in the concrete sense of literally inscribing layers of skin with outwardly visible pigment on areas of the body, but also in cultural, symbolic, and socio-psychological terms. Ethnographic and qualitative studies of modern tattooing explore the range of meanings that tattoos hold for those who have them; and how wider cultural understandings of tattoos have shifted in the context of radical global and technological shifts (see Schildkrout, 2004).

This research, conducted mainly by sociologists, anthropologists, and cultural studies scholars, focuses on populations in the postindustrial West, particularly in North American and Europe. Unlike psychological studies that have a tendency to conceptualize tattoos within frameworks of deviant behavior, risk-taking, and antisocial tendencies, qualitative research more often examines how tattoos relate to social scientific variables such as gender, social class, sexuality, and sense of self-identity. For example, Atkinson (2002a) offers a comprehensive treatment of contemporary tattooing practices. As a sociologist, ethnographer, and heavily tattooed person, he situates tattoo within a sophisticated micro- and macro-social frameworks, deliberately avoiding the rhetoric of deviance.

Ethnographic research from a phenomenologically driven frame explores how tattoos hold a multiplicity of meanings, and are related to self-construction and embodied knowledge. As many scholars have shown, tattoos are extensions of the self and serve as a means to visually communicate via the body (DeMello, 2000; Irwin, 2003; Kosut, 2000; Pitts, 2003; Sanders, 1989). The communication may be internal (as a form of self-reflexive body work) or external in the case of visible tattoos, or both. Visible tattoos do invite sociality, which can elicit both negative and positive responses.

The same image on the same body, however, may be read differently depending upon social environment. For example, tattoos on the bodies of gang members and prisoners that affiliate them with antimainstream and illegal activities can be interpreted simultaneously as beautiful and artistic, and also totally at odds with what is esthetically acceptable in larger society. As anthropologist Susan Phillips' (2013) ethnographic study of Chicano ex-gang members illustrates, the corporeal and material strategy of using a tattoo to survive in one social context can lead to stigmatization, exclusion, and discrimination in others. Tattoos may be creative expressions of the self at the intersection of art and embodiment, but the same tattoo can be read as both abomination and art.

Until recently, tattoo research tended to focus mainly on the perspectives of tattooed men. As tattooing has increased in popularity, and more women are getting tattooed, researchers are paying attention to the unique, embodied experiences of women with tattoos. Within the context of a patriarchy, the female body (or those identified visually as such) are subject to visual surveillance, commentary, harassment, and forms of public and private social control. Given embodied gender politics it is not surprising that women have gotten tattooed as a way to claim ownership of the body and as a form of personal empowerment.

Getting a tattoo is a visceral and direct way to both enact and announce a degree of agency and embodied authority to others. According to Mifflin (1997), the sexual revolution of the 1960s assisted the resurrection of tattoos from disrepute and situated them within feminist and counter-cultural movements. In the last decade, some women have also used tattooing as a way to reclaim the body after sexual abuse or trauma (see Atkinson, 2002a; Braunberger, 2000; DeMello, 2000; Pitts, 2003). In these examples, tattooing is experienced and conceptualized as a form of epidermal self help, rather than self-injury.

Scholars explore how some tattooed people process and interpret their changed identities and bodies after permanently modifying them. Sociologist D. Angus Vail, an ethnographer and tattoo enthusiast proposes that tattoo collectors (people who conceive of their bodies as canvases to be filled with tattoo artwork), begin to see their ink as art through a process of social interaction with key professionals in tattoo communities (1999). Such approaches contribute to the demystification of tattoos within the academia by establishing their relationship to functioning members of the mainstream; and they also semiotically elevate the tattoo's cultural status by classifying certain tattoos as esthetically significant.

Recently, scholars have called for 'prosocial' conceptualizations of tattooing that explore how tattoos constructively contribute to a sense of self and positive social interactions (see Strohecker, 2011). Ethnographic researchers assert that in many ways tattoos actually can, in essence, be good for us encouraging sociality. This burgeoning body of work demonstrates a paradigmatic shift in academic approaches to tattooing through the explicit rejection of deviance-driven stereotypes that prevailed for well over a century in the United States. Such a shift in interpreting the meaning and purpose of tattoos illustrates how theoretical conceptions of embodiment and the body (as an object of study) are as malleable as the human body itself. In fact, tattoo as a form of body modification is frequently framed under the heading of body art (Sanders, 1989; Kosut, 2013).

Tattoo designs have changed radically in the West since Samuel L. Riley first invented the electric tattoo machine in 1891, allowing for faster and less painful application. This led to the first spike in tattooing as a populist trend and set the stage for the beginning of a revolution in the esthetics of tattooing. While the ability to tattoo well requires skill and precision and a basic understanding of how a tattoo machine or tattoo gun works, framing tattoos as an art fundamentally changed the way that tattooists and clients viewed tattoos. Some professional tattoo artists have themselves inadvertently or purposefully advanced tattoo as an art, rather than a popular body modification or blue-collar craft. Clinton Sanders (1989) was one of the first sociologists to write about how tattooists and tattooed people understood tattoo as art in the 1980s. As such, it is more akin to drawing and painting rather than more craft-based mediums (Sanders, 1989; Kosut, 2013).

The formulation of professional tattoo organizations and the medicalization of tattooing have ushered in standard practices and procedures that have effectively changed the environment and act of getting tattooed. Once known as

tattoo shops, today tattoo studios prevail, some of which feature 'guest artists', long wait lists for appointments, and gallery-like environments where tattoo flash (drawings) and tattoo-inspired art work may be for sale on the walls. Tattoo studios are pleasant, highly estheticized, and professionally run. Just as tattoo shops have shifted into tattoo studios, tattooists are also commonly referred to as tattoo artists. A tattoo artist is expected to do original tattoo design work, be able to tattoo in a proficiency of styles, and possess a very high level of skill. Some tattoo artists have attended art school and earned fine arts degrees in classical mediums like drawing and painting. Even those who are not academically trained are often at least familiar with esthetic and art-historical discourse and frame and/or compare tattooing to styles of painting and couch them within esthetic-cultural theories (Kosut, 2013). The emergence of tattoo studios and a new generation of international tattoo artists has helped pave the way for the acceptance of the tattoo as body art – a quasi-legitimate esthetic practice that hovers between the landscapes of high, low, and popular culture.

The term body art is now ubiquitous in mass-mediated and academic discourses, and rarely challenged as a categorical referent in either milieu. Tattoo scholar Matt Lodder (2014) argues that art historians and social scientists alike have yet to explicitly examine tattoos from an art historical perspective (as body art). He argues that the methodologies and theories employed in critical analyses of other mediums, painting or sculpture, for example, can and should be applied to the genre of tattoo as an art form. This reframing of tattoo art within scholarly literature, positions tattoos on an art-historical scaffold. It gives credence to 'taken for granted' and popular conceptualizations of what body art may consist of (form), the way in which it is created (process), and how it can be identified according to formalist concerns such as line, balance, and perspective (content). This approach to tattooing is quite different from the previous studies mentioned that conceive of tattoo as an extension of the embodied self within social interaction. Here the tattoo and the tattooed body are framed as objects – media akin to paint and canvas. This approach takes the colloquial phrase of 'viewing skin as canvas' quite seriously and literally.

See also: Body, History of; Body: Anthropological Aspects; Collective Identity; Ethnic Identity and Ethnicity in Archaeology; Identity offline and Online; Identity and Identification; Integration and Multiculturalism: Focus on Western Europe; Mass Media and Cultural Identity; Personal Identity: Philosophical Aspects; Self and Identity Development During Adolescence across Cultures.

Bibliography

Atkinson, M., 2002a. Tattooed: The Sociogenesis of a Body Art. University of Toronto Press, Toronto.
Atkinson, M., 2002b. Pretty in ink: conformity, resistance, and negotiation in women's tattooing. Sex Roles 47, 219–235.
Brain, R., 1979. The Decorated Body. Harper and Row, New York.
Braunberger, C., 2000. Revolting bodies: the monster beauty of tattooed women. NWSA Journal 12, 1–23.
Burger, T., Finkel, D., 2002. Relationships between body modifications and very high risk behavior in a college population. College Student Journal 36, 203–214.
Caplan, J., 2000. Written on the Body: The Tattoo in European and American History. Princeton University Press, Princeton, NJ.
Carroll, S.T., Riffenburgh, R.H., Roberts, T.A., Myhre, E.B., 2002. Tattoos and body piercings as indicators of adolscent risk-taking behaviors. Pediatrics 109, 1021–1027.
Copes, J., Forsyth, C., 1993. The tattoo: a social psychological explanation. International Review of Modern Sociology 23, 83–89.
DeMello, M., 1995. 'Not just for bikers anymore': popular representations of American tattooing. Journal of Popular Culture 29, 37–52.
DeMello, M., 2000. Bodies of Inscription: A Cultural History of the Modern Tattoo Community. Duke University Press, Durham, NC.
Drews, D., Allison, C., Probst, P., 2000. Behavioral and self-concept differences in tattooed and non-tattooed college students. Psychological Reports 86, 475–481.
Durkheim, E., 1995. The Elementary Forms of Religious Life. The Free Press, New York.
Favazza, A., 1996. Bodies under Siege: Self-mutilation and Body Modification in Culture and Psychiatry. Johns Hopkins University Press, Baltimore and London.
Freidlander, M., King, M., 1999. Moko: Maori Tattooing in the Twentieth Century. David Bateman, Auckland, NZ.
Gathercole, P., 1988. Context of maori moko. In: Rubin (Ed.), Marks of Transformation: Artistic Transformations of the Human Body. University of California Press, Los Angeles, pp. 175–176.
Gell, A., 1993. Wrapping in Images: Tattoos in Polynesia. Oxford University Press, Oxford.
Gittelson, N., Wallen, G., Dawson-Bitterwork, K., 1969. The tattooed psychiatric patient. British Journal of Psychiatry 115, 1249–1253.
Grumet, G., 1983. Psychodynamic implications of tattoos. Journal of Orthopsychiatry 53, 482–492.
Hamburger, E., 1966. Tattooing as a psychic defense mechanism. International Journal of Social Psychiatry 12, 60–62.
Irwin, K., 2003. Saints and sinners: elite tattoo collectors as positive and negative deviants. Sociological Spectrum 23, 27–57.
Kitamura, T., 2003. Tattoos of the Floating World: Ukiyo-e Motifs in the Japanese Tattoo. Tattoo Museum, Amerstdam.
Koch, J., Roberts, A.E., Cannon, J.H., Armstrong, M.L., Owen, D.C., 2005. College students, tattooing, and the health belief model: extending social psychological perspectives on youth culture and deviance. Sociological Spectrum 25, 79–102.
Kosut, M., 2000. Tattoo narratives: the intersection of the body, self-identity, and society. Visual Sociology 15, 79–100.
Kosut, M., 2006a. An ironic fad: the commodification and consumption of tattoos. The Journal of Popular Culture 39, 1035–1048.
Kosut, M., 2006b. Mad artists and tattooed perverts: deviant discourse and the social construction of cultural categories. Deviant Behavior 27, 73–95.
Kosut, M., 2013. The artification of tattoo: transformations within a cultural field. Cultural Sociology. Published online.
Krutak, L., 1999. St. Lawrence island joint-tattooing: spiritual/medicinal functions and intercontinental possibilities. Études/Inuit/Studies 23 (1–2), 229–252.
Krutak, L., 2012. Spiritual Skin: Magical Tattoos and Scarification. Editions Reuss, Munich.
Kuwahara, M., 2005. Tattoo: An Anthropology. Berg, New York.
Loimer, Y., Werner, E., 1992. Tattooing and high-risk behavior amongst drug addicts. Medicine and Law 11, 167–174.
Lombroso, C., 1876[2006]. Criminal Man. Duke Universty Press, Durham, NC.
Matt, L., 2014. Tattoo: An Art History. IB Tauris, London.
McCallum, D., 1988. Historical and cultural dimensions of the tattoo in Japan. In: Rubin (Ed.), Marks of Transformation: Artistic Transformations of the Human Body. University of California Press, Los Angeles, CA, pp. 109–134.
Mifflin, M., 1997. Bodies of Subversion: The Secret History of Women and Tattoo. Juno, New York.
Newman, G., 1982. The implications of tattooing in prisoners. Journal of Clinical Psychiatry 43, 231–234.
Oksanen, Atte, Turtiainen, J., 2005. A life told in ink: tattoo narratives and the problem of self in late modern society. Auto/Biography 13, 111–130.
Parry, A., 1933. Tattoo: Secrets of a Strange Art as Practiced among the Natives of the United States. Simon and Schuster, New York.
Phillips, S.A., 2013. Gallo's body: decoration and damnation in the life of a Chicano gang member. Ethnography 2, 357–388.
Pitts, V., 2003. In the Flesh: The Cultural Politics of Body Modification. Palgrave Macmillan, New York.
Post, R., 1968. The relationship of tattoos to personality disorders. Journal of Criminal Law, Criminology and Police Science 59, 516–524.

Rosenblatt, D., 1997. The antisocial skin: structure, resistance, and modern primitive adornment in the United States. Cultural Anthropology 12, 287–334.
Rubin, A., 1988. Marks of Civilization. University of California Press, Los Angeles, CA.
Sanders, C., 1988. Marks of mischief: becoming and being tattooed. Journal of Contemporary Ethnography 16, 395–432.
Sanders, C., 1989. Customizing the Body: The Art and Culture of Tattooing. Temple University Press, Philadelphia, PA.
Schildkrout, E., 2004. Inscribing the body. Annual Review of Anthropology 33, 319–344.
Shilling, C., 1993. The Body and Social Theory. Sage, London.
Stirn, A., Hinz, A., 2008. Tattoos, body piercings, and self-injury: is there a connection investigations into a core group of participants practicing body modification. Psychotherapy Research 18, 326–333.
Strohecker, D., 2011. Towards a pro-social conception of contemporary tattooing: the psychological benefits of body modification. Rutgers Journal of Sociology 1, 10–36.
Sweetman, P., 1999. Anchoring the (postmodern) self? body modification, fashion and dentity. Body and Society 5, 51–76.
Tate, J.C., Shelton, B.L., 2008. Personality correlates of tattooing and body piercing in a collage sample: the kids are alright. Personality and Indvividual Differences 45, 281–285.
Vail, D.A., 1999. Tattoos are like potato chips … you can't have just one: the process of becoming a collector. Deviant Behavior 20, 252–273.
Wade, L., Sharp, G., 2010. Skull face and the self-fulfilling stereotype. Contexts 9, 80–81.
Yamamoto, J., Seeman, W., Lester, B., 1963. The tattooed man. Journal of Nervous and Mental Disease 136, 365–367.

Taylor, Charles (1931–)

Mark Redhead, California State University, Fullerton, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Charles Taylor is one of the world's leading social philosophers. He has made important contributions to debates in contemporary political theory, identity politics, and philosophy of the social sciences. He has also penned important studies of Hegel, the rise of Western modernity as well as Western secularity. In what follows, I briefly describe the various facets of his academic career before explicating the salient features of his work in the philosophy of the social sciences.

Professional Life

Charles Taylor was born in 1931 in Montreal and, with the exception of 2 years in which he attended a boarding school in Ontario, spent most of his childhood there. He then completed his undergraduate studies in History at McGill University. In 1953, Taylor left Montreal to pursue a PhD in Philosophy at Oxford. In 1961, he returned to Montreal and a professorship at the Universite de Montreal before moving on to an appointment at McGill. Except for a 5-year period (1976–81) when he returned to Oxford to hold the Chichele Chair of Political Philosophy, Taylor had been a notable member of McGill's faculty. After becoming an emeritus professor at McGill, Taylor taught for several years at Northwestern University. Besides his academic work, Taylor has been a prolific political actor and political commentator in both Quebec and Canada.

Overview of Taylor's Academic Writings

Taylor has penned some of the most influential pieces in the philosophy of the social sciences, contemporary political theory and Western intellectual history. In the first area, Taylor, along with thinkers such as Alasdair MacIntyre, Hans Georg Gadamer, and Paul Ricouer has been a leading critic of attempts to model the study of man in the social science on natural scientific models. Such models, particularly behaviorist and what calls atomist ones tend to marginalize the role of culture and the self-interpretation of cultural values in human agents self-understandings of their actions. Like Gadamer and Ricouer Taylor's work in this area has become synonymous with a hermeneutical approach to the social sciences, an approach which makes the self-understandings of agents the most conceptually significant element in one's explanation of a given social phenomena. Taylor's work has been highly influential and has drawn a wide array of rejoinders from mainstream social scientists who repeatedly have accused him of underplaying the predictive power of behaviorist and atomist models. These works have also placed him in debate with the German scholar Jurgen Habermas, who early in his career argued that Taylor's hermeneutical model left little room for engaging in critique of the false self-understandings of agents.

As a political theorist, Taylor, drawing of his lifelong involvement of the politics of native Quebec and Canada as well as his magisterial study of Hegel, has composed a number of highly influential essays on liberalism, multiculturalism, the fragmented nature of modern life, and the economic injustices of liberal capitalism. In the 1980s Taylor's work on liberalism was lumped together with that of MacIntyre as part of a supposed communitarian strand of political theory. Taylor (like MacIntyre) was never comfortable with the communitarian label. However, this did not stop his critiques of liberal theory from drawing waves of commentaries from liberal critics of communitarianism. The most trenchant was and still is fellow Canadian Will Kymlicka who argued that contra-Taylor (and MacIntyre) liberals do not rely on a context-free notion of rationality for the justification of their principles but in fact often root them in some appeal to common norms and understandings. Taylor's essay 'Multiculturalism and the Politics of Recognition' was one of the seminal texts in the 1990s through early 2000s identity politics debates transpiring across the humanities and social sciences of Anglo-American academia. The main source of controversy created by the piece had to do with his Hegelian understanding of recognition and its role in human life with defenders and fellow intellectual travelers like Axel Honneth pushing for a greater role for recognition in social theory and critics (particularly in the United States) feeling Taylor's usage of the term was overdone.

Finally, as an intellectual historian, Taylor is well known for not only his aforementioned work on Hegel but also for producing two of the most influential studies of Western modernity, *Sources of the Self* and *A Secular Age*. These texts are generally acknowledged as two of the more persuasive attempts at understanding the various ethical features of Western modernity or what Taylor calls the modern social imaginary, a moral ontology that, Taylor maintains, those living within contemporary Western liberal democracies are always already informed by when they attempt to publicly reason with other members of said societies. This work has been lauded by its defenders for its breadth and ambition and attacked by its critics for being too reductionist in its account of Western identity and too influenced by Taylor's own Catholic spirituality.

The scope and influence of Taylor's professional academic work has recently been recognized beyond academia. In 2007, he was awarded the Templeton prize for Progress Toward Research about Spiritual Realities. In 2008, he was awarded Japan's Kyoto Prize for his lifelong achievements in Philosophy.

Taylor's Social Scientific Thinking Briefly Explained

Taylor's Critique of Analytical Philosophy and Naturalist Approaches to the Social Sciences

Taylor has always been wary of analytical moral and political philosophy because its focus on the right as opposed to the good 'restricted our understanding of what morality is.' These models view "morality as concerned with questions of what we ought to do and to occlude or exclude questions about what it is good to be or what it is good to love. The focus is on obligatory action, which means that it turns away from issues in which obligation is not really the issue, as well as those where not just actions but ways of life or ways of being are what we have to weigh" (Taylor, 2011a: 3). The main culprits here are single term moralities like what one sees in John Rawls' *A Theory of Justice* that focus exclusively on issues of justice-benevolence and thus shorten our moral world by focusing only on what we are obligated to do (Taylor, 2011a: 10). Taylor, by contrast, insists that we also focus on ideals like the dignity of each human life that ethically move us into having a deeper concern with moral issues like the obligations justice entails (Taylor, 2011a: 11).

Taylor's antipathy toward analytical moral and political philosophy is an outgrowth of his longstanding criticisms of social scientific thinking modeled on the natural sciences. Taylor's hermeneutical optics emphasizes the manners in which man acts as a self-interpretative animal. To see man as a self-interpretative animal is to see that the meaning of an entity cannot be understood independently of the subject's interpretation of it. "What is interpreted is itself an interpretation; a self-interpretation which is embedded in a stream of action. It is an interpretation of experiential meaning which contributes to the constitution of the meaning" (Taylor, 1985a: 26).

As self-interpretative animals, we are also creatures whose self-identity is constituted through intersubjective or dialogical actions with others. These acts by which we share a sense of agency with those whom we act and interact with, are vital to our self-understanding of who we are and the world we inhabit since "we cannot understand human life merely in terms of individual subjects, who frame representations about and respond to others, because a great deal of human action happens only insofar as the agent understands and constitutes himself or herself as integrally part of a we" (Taylor, 1991: 314). This means that our identity is never simply defined in terms of our individual properties. It is situated within a social space making us partly define ourselves through our interactions with others.

Naturalist social science models underemphasize the features of man's self-interpretative and dialogical social existence, preferring to model the study of man on the natural sciences. These approaches valorize "the omnicompetence of post-Galilean science, (and) believes that it can cope with life-related values – survival, flourishing as a living being, including health, strength, energy, the desires which de facto accompany life, for sex, food, (in certain cases) dominance …. But it cannot see how values of an incommensurably higher range can have a place in post-Galilean nature" (Taylor, 2000: 245). Moreover, naturalist theories suggest that not only can "man be seen as a part of nature – in one sense or other this would surely be accepted by everyone – but that the nature of which he is a part is to be understood according to the cannons which emerged in the seventeenth-century revolution in natural science"(Taylor, 1985b: 2).

Naturalist social science models are the product of what Taylor in *A Secular Age* refers to as the modern need for 'bifocal' grasps of reality. One is an objectifying account of human nature and reality while the other is a creative act through which humans attempt (to use Arendtian language) to create durable worlds fit for living (Taylor, 2007: 182). This underlies a battle between objectivating social scientists and economists and their antipositivist critics like Taylor who chafe at the effects on freedom and authenticity that the former's work sanctions. Taylor's sympathies for the latter camp stem from the point that our selfhood, "cannot be treated scientifically in exactly the same way we approach our organic being. What it is to possess a liver or a heart is something I can define quite independently of the space of questions in which I exist for myself, but not what it is to have a self or be a person" (Taylor, 1985b: 4). Taylor's hermeneutic perspective takes one's self-interpretations of themselves and reality as constitutive of what and who they are. These interpretations cannot be reduced to the status of epiphenomenons or 'subjective' emotions and thus not so legitimate accounts of reality (Taylor, 1985b: 47). Humans are self-interpreting creatures who exist in a social space that is defined by 'distinctions of worth.' A being who exists only in self-interpretation can never be understood absolutely. This, though, is what naturalist theories claim to be able to do by giving a strictly causal and empirical account of man in a value-free language. The modern disengaged self "aspires after a kind of neutrality. Disengagement entails a kind of neutrality in relation to what is disengaged from" (Taylor, 1988: 312).

However, we can never escape the very human predilections that frame how we look at reality, least of all when we deceive ourselves into thinking we are reflecting upon ourselves from a position outside of ourselves. This is because any and every "given explanatory framework secretes a notion of the good, and a set of valuations, which cannot be done away with – though they can be over-ridden – unless we do away with the framework. Of course because values can be over-ridden, we can only say that the framework tends to support them, not that it establishes their validity" (Taylor, 1985a: 90). Hence "whatever we might identify as a fundamentally common human nature, the possible object of an ultimate experience-transcending science, is always and everywhere mediated in human life through culture, self-understanding, language" (Taylor, 2002: 11). A being "who can only be understood against the background of distinctions of worth cannot be captured by a scientific language which essentially aspires to neutrality" (Taylor, 1985b: 3).

The key phenomena underappreciated by naturalist models are collective meanings that we all draw on in creating our self-understandings of the world and ourselves. The most important dialogical actions humans engage in are the learning of languages of self-expression. A spoken language like English or French contains common and intersubjective meanings that we as individuals play off in our attempts to understand ourselves, others, and the world in which we and they reside. Common meanings are objects in a given social world that all

participants in that world share such as collective practices, collective feelings like patriotism and what Taylor defines elsewhere as common goods (such as the preservation of Francophone culture in Taylor's native Quebec) that all members of a society partake of. "Inter-subjective meaning gives a people a common language to talk about social reality and a common understanding of certain norms" (Taylor, 1985a: 39). Naturalist social science models are unable to account for common and intersubjective meanings because they lack a conception of meaning for a subject who can be part of a we as well as an I (Taylor, 1985a: 40). At the same time, these models rely on intersubjective meanings for their accounts to make sense. The result is situations like 1960s theories of development in comparative politics where scholars strove to create value neutral or objective understandings of politics that can apply universally to all societies but which in fact are based solely on historically specific Western-centric intersubjective meanings. This produced ethnocentric theories of development that predictably placed Western democracies at the apex of development (Taylor, 1985a: 42).

Good social science should go beyond the mere bounds of empirical verification to study changes in social reality at the level of its foundational intersubjective meanings (Taylor, 1985a: 52). A hermeneutician like Taylor strives to do this by providing a clearer interpretations of a given social phenomena than the immediate understanding of agents involved in it. Hence the hermeneutician's understandings must in an important sense be 'out of phase with the explicandum' (Taylor, 1985a: 27). The problem though is that this out of phase explanation will itself be a force for change in the meaning of a given phenomenon to an agent as well as a possible change of that entity and a given social practice. Entities only have "meaning in a field, that is, in relation to the meanings of other things. This means that there is no such thing as a single, unrelated meaningful element; and it means that changes in the other meanings in the field can involve changes in the given element" (Taylor, 1985a: 22). Intersubjective meanings define these fields as well as the parameters for change as they define the semantical resources from which subjects can construct their own self-interpretations of the phenomena.

Taylor's Account of the Strongly Evaluative Subject

Taylor is well aware that concepts such as human rights are dependent upon some ontology of value that cannot be reduced to a utilitarian or mechanist understanding of human social life (Taylor, 2003: 310). Such understandings derive their value and legitimacy from a larger ontology of meaning supplied by the cultures and traditions from which they emerged, in this case the modern West. Heidegger and Merleau-Ponty taught Taylor that contra a mediational or representationalist epistemology, "our grasp of things is not something that is in us, over against the world; it lies in the way we are in contact with the world" (Taylor, 2004a: 38). Hence the subject-centered picture of mediational epistemology that has held 'us' captive since Descartes can be escaped via a deeper understanding of our thinking, which Heidegger and Merleau-Ponty provide (Taylor, 2004a: 40). For Taylor, humans are both engaged and disengaged creatures at once at the effect of causes and under the spell of reasons for action. We are always already coping with the world via our bodily commerce with it. Yet central to our modern sense of self is the ability to disengage from this engaged mode of being so as to see things from a universal or 'objective' viewpoint (Taylor, 2004a: 46). This is because a central feature of the modern identity as it comes down to us through history is its buffered characteristics. The modern self is a buffered self in that things, which lie outside the boundary of 'the self' do not necessarily need to affect the self. "This self can see itself as invulnerable, as master of the meanings that things have for it" (Taylor, 2011a: 41). In our disenchanted world, "the physical world, outside the mind, must proceed by causal laws that in no way turn on the moral meanings things have for us"(Taylor, 2011b: 62). By contrast the premodern self is a porous one in which the boundary between self and external forces is porous. The latter lived in an enchanted world where the meaning of something existed already outside of us, prior to contact (Taylor, 2011b: 61).

A modern can learn that he feels depressed because of hunger, hormones, or whatever thereby allowing him to take a critical distance from and potentially devalue this feeling. The modern emphasis on a mind/body dichotomy allows her to relegate the feelings of the body to a contingent set of physical forces that do not necessarily need to have sway over the actions of said self. "But a pre-modern may not be helped by learning that his mood comes from black bile. Because this doesn't permit a distancing. Black bile *is* melancholy. Now he just knows that he is in the grips of the real thing" (Taylor, 2007: 37).

Hence we are in a moral sense made by ontologies of meaning even as we try to be disengaged and reflective upon them – the drive to disengage and produce out of phase explanations is itself part of the ontic framework that we try to reflect on. Moments of attempted disengagement provide clearer (but never clear) understandings of who we are but do not literally disengage us from the moral ontologies that we cannot help but draw upon in crafting our identities.

These ontologies help to create a social imaginary; a common or background set of understandings, which makes possible a society's common practices of politics, morality, religion, and economics as well provide a widely shared sense of the legitimacy (or lack thereof) of these practices (Taylor, 2004: 23). A social imaginary, denotes "that largely unstructured and inarticulate understanding of our whole situation, within which particular features of our world show up for us in the sense they have. It can never be adequately expressed in the form of explicit doctrines, because of its very unlimited and indefinite nature" (Taylor, 2010: 309). For example, far from being free standing, a form of public reasoning like Rawls' is an expression of some family of moral values and those values are indeed synonymous with certain metaphysical and ethical assumptions about human subjectivity, dignity, and the like. The model only becomes coherent by tacitly invoking a set of intersubjective and common meanings synonymous with the modern social imaginary.

Nevertheless these ontologies and the social imaginary they constitute are historical creations. Given the disengaged nature of modern existence, a crucial task for contemporary social scientists is to understand the diversity of moral goods that comprise the modern social imaginary and the ontologies of

meaning like the modern moral order that make us who we are even as we reflect upon and critically distancing ourselves from them. This entails freeing ourselves from philosophical systems that are either predicated upon a disengaged relationship between a subject and an object, or, like Marxism, that recognize only one genre of values and only one manner of interpreting reality (Taylor, 1985: 247). By doing so, we can open ourselves up to the diverse array of moral goods that we unavoidably invoke within our daily lives.

Taylor's Understanding of the Modern Social Imaginary

Taylor's most important works *Sources of the Self* and *A Secular Age* are attempts at articulating not only these goods but also the manners in which they combine to frame the most important moral and political conflicts at play beneath the modern social imaginary. *Sources of the Self* narrates the rise of our modern life goods and the moral sources that inform them as well as the subsequent development of the conflicts between these life goods. Taylor also distinguishes three aspects of (but what could more accurately be described as three movements in) the development of the modern social imaginary: "First, modern inwardness, the sense of ourselves as beings with inner depths and the connected notion that we are 'selves'; second, the affirmation of ordinary life which develops from the early modern period; third, the expressivist notion of nature as an inner moral source"(Taylor, 1989: x). The first movement extends from Augustine through Descartes and the original Christian turn inward. Taylor traces the second movement, the affirmation of ordinary life, "from the Reformation through the Enlightenment to its contemporary forms" (Taylor, 1989). The third Taylor depicts as originating with the Romantic Movement in the late-eighteenth century and then developing via modernist art, nineteenth-century thinkers like Schopenhauer, Baudelaire and Nietzsche, and later twentieth-century literature.

Taylor's depiction of these three movements leaves him with what he admits is an overly schematic map of the historical origins of our modern moral sources. The key point about this map is that the three domains are fluid as they are continually borrowing from and influenced by each other. Hence our life goods often draw from more than one set of moral sources and find their origins in multiple historical movements. For example, the ethic of benevolence, a product of the affirmation of ordinary life, draws on two sources: the ideal of disengaged reason and autonomous subjectivity and a conception of nature, which one would normally associate with the expressivist genre of constitutive goods. What Taylor calls the ethic of authenticity was a product of the romantic era's critique of Enlightenment forms of reasoning and its yearning for seemingly disavowed senses of community. However, the ethic also owes much of its development to the modern turn inward. Furthermore, most plausible notions of self-realization or authenticity presuppose – following the third spiritual set of constitutive goods – that some goals or aspirations that link the individual to something greater than self-interests.

A prominent criticism of *Sources of the Self* is that Taylor never actually shows us why or how the constitutive or life goods of modernity are actually embodied in the actions and lives of actual living human beings and hence why *Sources of the Self* has indeed provided us with the best account of 'our' moral order.

This problem is somewhat rectified in *A Secular Age* and the works that accompanied it. One of these related works *Modern Social Imaginaries* (a shorter version of which is a chapter in *A Secular Age*) provides a story of how a set of ideas initially wrapped up in the discursive practices of the seventeenth-century natural law tradition of Grotius and Locke becomes transformed by modern and contemporary intellectuals and along the way drives the development of the modern moral order with its practices of civility, the market economy, the public sphere, and the practice of democratic self-rule. These practices embody many of the life goods, such as autonomy, equality, disengaged reason, and the affirmation of ordinary life Taylor chronicles in *Sources* (Taylor, 2004b). In *A Secular Age* this story then complements other stories on Christian reform movements, the emergence of deism and eventually modern humanism and the growth of modern disciplined and disengaged notions of subjectivity to give us a picture of the modern social imaginary. More specifically, the stories collected in *A Secular Age* chronicle the development of contemporary secularity through a series of redactions that make it increasingly more inclusive, richer (in its contents) and demanding (in its requirements on modern Western subjects) as Western modernity progresses from its genesis in the seventeenth century to its present incarnations.

See also: Atomism and Holism: Philosophical Aspects; Foucault, Michel (1926–84); Hermeneutics; History and the Social Sciences; Identity and Identification, Social Psychology of; Kant, Immanuel (1724–1804); Multiculturalism; Naturalism; Practical Reasoning: Philosophical Aspects; Religion, History of; Self: Philosophical Aspects; Social Ontology; Social Science, The Idea of.

Bibliography

Dilemmas and Connections, 2011. Belknap Press, Cambridge, MA.
Explanation of Behavior, 1964. Humanities Press, New York.
The Ethics of Authenticity, 1995. Harvard University Press, Cambridge, MA.
Hegel and Modern Society, 1979. Cambridge University Press, Cambridge.
Hegel, 1975. Cambridge University Press, Cambridge.
Human Agency and Language: Philosophical Papers, vol. I, 1985. Cambridge University Press, Cambridge.
Philosophy and the Human Sciences: Philosophical Papers, vol. 2, 1985. Cambridge University Press, Cambridge.
Modern Social Imaginaries, 2004. Duke University Press, Durham, NC.
Philosophical Arguments, 1995. Harvard University Press, Cambridge, MA.
The Pattern of Politics, 1970. McClelland and Stewart, Toronto.
Reconciling the Solitudes: Essays on Canadian Federalism and Nationalism, 1994. McGill-Queen's University Press, Montreal.
A Secular Age, 2007. Harvard University Press, Cambridge, MA.
Sources of the Self, 1988. Harvard University Press, Cambridge, MA.
Taylor, Charles, 1985a. Philosophy and the Human Sciences: Philosophical Papers, vol. 2. Cambridge University Press, Cambridge.
Taylor, Charles, 1985b. Human Agency and Language: Philosophical Papers, vol. I. Cambridge University Press, Cambridge.
Taylor, Charles, 1988. The moral topography of the self. In: Messer, Stanley B., Sass, Louis A., Woolfolk, Robert (Eds.), Hermenutics and Psychological Theory. Rutgers University Press, New Brunswick, NJ, pp. 291–338.

Taylor, Charles, 1989. Sources of the Self. Harvard University Press, Cambridge, MA.
Taylor, Charles, 1991. The dialogical self. In: Hiley, David R., Bohman, James F., Shusterman, Richard (Eds.), The Interpretive Turn: Philosophy, Science and Culture. Cornell University Press, Ithaca, NY, pp. 302–318.
Taylor, Charles, 2000. McDowell on value and knowledge. Philosophical Quarterly 50 (199), 242–249.
Taylor, Charles, 2002. Remembering gadamer. IWM Newsletter 76, 9–14.
Taylor, Charles, 2003. Ethics and ontology. Journal of Philosophy C (6), 305–320.
Taylor, Charles, 2004a. Merleau-Ponty and the Epistemological Picture. The Cambridge Companion to Merleau-Ponty. Cambridge University Press, Cambridge, pp. 26–49.
Taylor, Charles, 2004b. Modern Social Imaginaries. Duke University Press, Durham, NC.
Taylor, Charles, 2007. A Secular Age. Harvard University Press, Cambridge, MA.
Taylor, Charles, 2010. Afterword: apologia pro Libro suo. In: Warner, Michael, VanAntwerpen, Jonathan, Calhoun, Craig (Eds.), Varieties of Secularism in a Secular Age. Harvard University Press, Cambridge, MA, pp. 300–324.
Taylor, Charles, 2011a. Dilemmas and Connections. Belknap Press, Cambridge, MA.
Taylor, Charles, 2011b. Western secularity. In: Calhoun, Craig, Juergensmeyer, Mark, VanAntwerpen, Jonathan (Eds.), Rethinking Secularism. Oxford University Press, Oxford, pp. 31–53.
Taylor, Charles, 2011c. Disenchantment-reenchantment. In: Levine, George (Ed.), The Joys of Secularism. Princeton University Press, Princeton, NJ, pp. 57–73.

Teacher Behaviours and Student Outcomes

James H Stronge, Leslie W Grant, and Xianxuan Xu, School of Education, College of William and Mary, Williamsburg, VA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Brophy, volume 23, pp. 15450–15454, © 2001, Elsevier Ltd.

Abstract

A multitude of empirical studies have documented that teacher effectiveness impacts student learning, and the impact is of magnitude that is larger than the effects associated with financial investment, class size, curriculum, school restructuring, and many other educational interventions. Teachers matter. Becoming an effective teacher is not simple. Teaching is a complex form of work. It involves a specialized, intricate, and constantly changing and renewing body of knowledge and skills. Nonetheless, there are common attributes that characterize effective teachers. This article summarizes the common background and identifies the qualities, dispositions, and behaviors of effective teachers that emerged in a careful review of the extant literature.

Teachers and Teaching

Teachers, as the most powerful school-related factor, matter massively to student learning outcomes and school performance. Indeed, among the factors within the control of school systems, teachers offer the greatest opportunity for improving the quality of life of students. As noted in *How the World's Best-Performing School Systems Come Out on Top*, an international study comparing data from the OECD's Programme for International Student Assessment, "The quality of an education system cannot exceed the quality of its teachers" (Barber and Mourshed, 2007: p. iii).

The research suggests that "while schools have powerful effects on student achievement differences, these effects appear to derive most importantly from variations in teacher quality" (Hanushek et al., 1998: p. 1). With the development of sophisticated statistical modeling involving longitudinal data, researchers have been able to isolate effects on student achievement and one theme emerges: the effect of the classroom teacher. A multitude of studies conducted in the United States have documented that teacher quality has a significant impact on student achievement (Stronge et al., 2011). Based on a review of the literature, qualities related to effective teaching include background characteristics of teachers, teacher dispositions, lesson planning, the classroom environment, lesson delivery, and classroom assessment.

Background Characteristics

Presage Teacher Characteristics

Presage variables are "properties of teachers that can be assumed to operate prior to, but also to have an influence on, the interactive phase of teaching" (Rowan et al., 2002: p. 1538). The most widely studied presage variables include certification status, teaching experience, content knowledge, degree, coursework, verbal skills, subject-matter knowledge, academic achievement, and the prestige ratings of teachers' undergraduate institutions. Collectively these presage – or background – qualities do appear to be important in terms of teacher effectiveness. Specific teacher presage qualities and their relationship to student achievement that will be explored briefly include: teacher degrees earned, coursework, certification status, content knowledge, teaching experience, and verbal ability.

Degrees Earned

Degree level and area of degree attained by the teacher have been widely examined in quantitative studies. Generally, the research has been inconclusive as to how degrees contribute to teacher effectiveness. Several studies found no statistically significant effect of teacher degree level on teacher effectiveness as measured by student achievement gains, while controlling for confounding effects of student background and other teacher-related characteristics (e.g., Wayne and Youngs, 2003). On the other hand, others have found statistically significant effects of teachers' degree status on growth at the elementary and secondary levels in reading and mathematics (e.g., Rowan et al., 2002).

Coursework

Another teacher attribute that is closely related with degree is teachers' course-taking. Over the last several decades, policy makers and researchers have used measures of the level and type of coursework taken by teachers as proxies for what teachers know and can do in their classrooms. Individual studies and syntheses of literature have found that coursework-taking has a positive effect on student achievement and these effects are seen in high school mathematics and science (Monk, 1994; Rice, 2003). However, in other subjects and grade levels, the impact of teacher content course-taking on student achievement was found to be indeterminate and inconsistent.

Certification Status

Historically, policy on teacher certification has been used as an important leverage to manipulate the quality of the teaching profession, and it is one of the most widely tested teacher attributes in educational research. The extant evidence indicates a relationship between certification in mathematics and students' high school mathematics achievement (Goldhaber and Brewer, 1997) but at the elementary level. Findings from

different value-added studies have yielded inconsistent results (Rowan et al., 2002; Palardy and Rumberger, 2008).

Types of teacher certification also have been studied at several levels. These studies have found consistent evidence that teachers with standard certification have more of a positive impact on student achievement than teachers with emergency or no certification (Darling-Hammond et al., 2005; Goldhaber and Brewer, 2000).

Content Knowledge

A teacher's understanding of subject facts, concepts, principles, and the methods through which they are integrated cognitively determine their pedagogical thinking and decision-making. Studies have found consistent results in that content knowledge and pedagogical content knowledge on the part of the teacher are crucial for student success in mathematics (Harris and Sass, 2007; Hill et al., 2005). Most studies related to content knowledge have focused on mathematics and therefore it is difficult to draw conclusions about the effects of content knowledge in other content areas.

Teaching Experience

Teacher experience is a crucial criterion for teacher compensation in most public school systems, and policy makers and researchers have shown increasing interest in investigating its impact on student achievement. Putting available evidences together, it seems that the relationship between years of teaching and teacher expertise is nonlinear. Teachers' experience matters only for the first few years of teaching – most definitely, the first 3 years. During these first few years, teachers appear to gain cumulatively in their contribution to student learning and, then, after three or so years, the contribution of experience to student learning levels off (Rivkin et al., 2005). Nonetheless, there are contradictory findings regarding the relationship between teaching experience and teacher effectiveness.

Teacher Verbal Ability

One key finding that has emerged from an analysis of teachers' background qualities is that students taught by teachers with greater verbal ability learn more than those taught by teachers with lower verbal ability (Rowan et al., 1997). Teaching is, among other things, explaining and is a verbal process. Thus, a discernable link appears to exist between effective teachers' verbal clarity and student academic achievement.

Teacher Dispositions

Although teachers vary significantly in their ability to improve student achievement gains, little of this variation can be attributed to observable characteristics such as degree and teaching experience (Rivkin et al., 2005). For instance, Goldhaber (2002) pointed out that only about 3% of the contribution teachers make to student learning is associated with teacher experience, educational level, certification status, and other readily observable characteristics. The remaining 97% of teachers' effects on student achievement are associated with intangible, unobserved aspects of teacher quality such as dispositions, attitudes, and classroom practices. Critical teacher dispositions that emerge in the literature include caring, a sense of efficacy, motivation and enthusiasm, and professional growth.

Caring

Caring about students and respecting them as individuals are prevalent in the literature descriptions of effective teachers (Cassidy and Bates, 2005; Noddings, 1992). Good teachers are often described as warm, friendly, and caring; conversely, ineffective teachers often are said to create a tense classroom and are described as cold, abusive, and uncaring (Walls et al., 2002). In classroom learning, when students are supported by a caring teacher, they are more likely to ask questions, take chances, and share their inner thoughts in creative writing and through other forms of expression (Cassidy and Bates, 2005).

Exemplary teachers regard the ethic of care and respect as a vital foundation for students' best learning and a prerequisite for effective teaching. They reach out to know their students by using multiple sources of knowledge (e.g., solicited critique, dialogues and questions, knowing students informally, knowing from colleagues, and knowing students' cultures) (Collinson et al., 1999).

Efficacy Beliefs

Self-efficacy refers to the belief in one's ability to accomplish intended outcomes. This belief influences an individual's behavior, motivation, effort, ultimately their success or failure. The reviews of research on teacher self-efficacy have summarized that teachers' self-efficacy was linked to teaching practices in their classrooms and student outcomes such as students' self-efficacy beliefs and student engagement, motivation, and achievement (Tschannen-Moran and McMaster, 2009). Compared to teachers with lower self-efficacy beliefs, teachers with stronger perceptions of self-efficacy tend to use more challenging teaching techniques, try innovative strategies, and employ classroom instruction that are more organized and better planned, student centered, and humanistic (Tschannen-Moran and Hoy, 2001). Efficacy is also a personal resource that can protect teachers from experiences of job stress and burnout (Schwarzer and Hallum, 2008).

Motivation and Enthusiasm to Teach

Motivation and enthusiasm are contagious in classrooms. Teachers who are more enthusiastic about teaching exhibit higher quality instructional behaviors, such as monitoring student learning, providing students with more autonomy support, offering more social support to students, and using higher levels of cognitive challenge. Teacher motivation also is expressed in a range of teacher behaviors that are perceived to be conducive to student learning, such as enthusiasm in content area taught, interest about students' personal and developmental needs, participation in content-related activities outside of class time, and displaying value and emotion for students (Long and Hoy, 2006).

Teachers who display enthusiasm and energy in the classroom often increase student interest and motivation to learn.

Among various teacher variables, enthusiasm was the most powerful unique predictor of students' intrinsic motivation and vitality. The students who received instruction from an enthusiastic teacher reported greater intrinsic motivation regarding the learning material and experienced higher levels of vitality (Kunter et al., 2008). Students who are under the tutelage of motivated and enthusiastic teachers are more likely to be enthusiastic toward the subject and exhibit higher rate of on-task behavior (Bettencourt et al., 1983).

Professionalism and Professional Growth

Another key attribute of effective teachers is a commitment to continuous improvement and perpetual learning. Evidence indicates that teachers who receive substantial professional development can help students achieve more. For example, based on the findings of one meta-analysis, teachers who receive substantial professional development (in this instance, 49 h) can boost their students' achievement by about 21 percentile points, and this effect size is fairly consistent across content areas (Yoon et al., 2007).

Lesson Planning

Teaching is a complex activity that involves careful preparation and planning, both for short-term and long-term learning purposes. Looking at national level teaching standards in the United States, such as the Interstate New Teacher Assessment and Support Consortium (InTASC) (2013) and the National Board for Professional Teaching Standards (NBPTS) (2012), instructional planning is a key aspect of any teacher's work in order to give students meaningful learning experiences. Among key attributes of effective teachers' planning are: making instruction a priority in teaching practices, holding high student expectations, developing planning routines and automaticity, providing clear learning objectives, creating quality assignments, and planning for instructional sequencing and curricular alignment. Each of these attributes in relation to the behaviors of effective teachers will be explored, in turn. (A significant portion of this article on effective teachers and instructional planning is adapted from a research review conducted by Jessica Straessle as part of her dissertation at the College of William and Mary with permission.)

Making Instruction a Priority

One of the best ways to maintain a focus on instruction is for teachers to have a written plan. Written lesson plans provide teachers with an opportunity to connect learning goals to activities, allocate time to complete those activities, and remove any foreseeable obstacles to instruction. Well-organized and thoughtfully constructed lesson plans help teachers become more effective leaders in the classroom and enable them to make more efficient use of class time (Panasuk et al., 2002). Effective use of class time helps teachers maintain focus on specific learning goals, since lesson plans provide a blueprint for instruction that allows teachers to allocate time for academic instruction.

Instructional Planning and Teacher Expectations

The beliefs that teachers have about their students' abilities can profoundly affect both teacher and student performance. Low expectations impact teacher efficacy, student behavior, and student achievement (Wentzel, 2002). Studies indicate that teachers can hold differing expectations based for wealthy versus poor students and poor students in urban settings (Auwarter and Aruguete, 2008; Warren, 2002).

Routines and Automaticity

In Yinger's (1980) study of teacher planning, he found that routines played a major role in a teacher's planning. Routines were used by the teacher to regulate activities and to simplify planning. In fact, he found that routines were so much a part of planning that he described planning "as decision-making about the selection, the organization, and the sequencing of routines" (p. 111). Hattie (2009) also discussed routines and the need for teachers to have routines in order to be experts. Listed under the dimension of *Monitoring and Providing Feedback*, Hattie (2009) determined that expert teachers are more automatic than nonexpert teachers.

Clear Lesson and Learning Objectives

Among the many decisions a teacher must make, research indicates that planning for effective teaching includes identifying clear lesson objectives (Stronge, 2007). Objectives provide a preliminary guide for planning (Marzano, 2007). Effective teachers provide a short statement of goals that provide focus for the instructional process.

Creating Quality Assignments

Developing clear learning objectives is merely an early step in instructional planning for effective teachers. Creating quality assignments is another characteristic of an effective teachers' lesson planning. In practice, they recognize that the assignment is just as important as the objective as it is the means to the end. The idea that clear objectives, followed by quality assignments, lead to student achievement, should make sense. Studies have found significant relationships between quality of classroom assignments and quality of student work (Clare, 2001; Koh and Luke, 2009).

Instructional Sequencing

Sequencing is another important aspect of logically structuring lessons for effective teachers. The teacher sets students up for success by identifying a sequence and ensuring that the knowledge gained is based in grounded ideas and building from that point. Good and Brophy (2008) found effective teaching requires that teachers plan sequences of lessons and do not plan in isolation.

Instructional Alignment

Alignment goes hand in hand with sequencing in enhancing teachers' planning practices. It is important for lesson planning

to be sequential, and within the sequence there must be alignment in all aspects – particularly the learning objective, activity, and evaluation. This alignment ensures that students are engaging in activities which support the learning objective and are being evaluated on the same learning objective.

Learning Environment

Caring, supportive, safe, challenging, academically robust: these attributes help define what it means to have a positive learning environment that is conducive to student success. The most prevalent criteria used to define learning environment are probably the physical arrangement of the classroom, discipline, and routines; organization of learning activities; and the engagement of students with tasks, among others (Cameron et al., 2008). Students need an engaging, stimulating, and enriching learning environment to grow and thrive. In order to achieve this type of rich environment, effective teachers establish and communicate guidelines for expected behavior; monitor student behavior; keep students on tasks; and infuse humor, care, and respect into the classroom interactions, so as to develop a climate that contribute to student learning (Kunter et al., 2007).

As a result, research has indicated that a positive learning environment can shape student outcomes in cognitive, motivational, emotional, and behavioral domains (Ludtke et al., 2009). Wang et al. (1994) analyzed a knowledge base comprising 11 000 statistical findings connecting a variety of variables and student achievement in order to answer the question: *What helps students learn?* Twenty-eight categories of factors, classified into six broad types of influences (i.e., student aptitude, classroom instruction and climate, context, program design, school organization, state and district characteristics), were scored based on their positive impact on learning. Of the 28 categories, classroom management ranked first and was the most influential variable, just ahead of student metacognitive processes and cognitive processes. In the researchers' view, "effective classroom management increases student engagement, decreases disruptive behaviours, and makes good use of instructional time" (p. 76).

Classroom Routines and Procedures

Classroom management includes actions taken by teachers to establish order, engage students, elicit student cooperation, with an ultimate purpose to establish and maintain a setting conducive to instruction and learning. Effective teachers were found to maintain their management system by monitoring and providing prompt feedback, and by consistently applying classroom procedures and consequences (Emmer and Stough, 2001). Stronge et al. (2008) found that compared with bottom-quartile teachers, top-quartile teachers (as determined by their ability to effect student achievement gains) were more organized than ineffective teachers with efficient routines and procedures for daily tasks. And they communicated higher behavioral expectations to students. The top teachers also were found to have less disruptive student behaviors (once every 2 h) than did the less-effective teachers (a disruption every 12 min). Researchers also noted that teachers who spend more time establishing instructional routines at the beginning of the school year did not need to exert as much effort on similar tasks later in the year (Cameron et al., 2008).

Physical Arrangement of the Classroom

Effective teachers develop functional floor plans with teacher and student work areas and furniture/materials placement for optimal benefit (Stronge, 2007). LePage et al. (2005) contended that an orderly classroom arrangement optimizes learning time, whereas inadequate planning causes disruptions and delays. Additionally, arranging the classroom carefully and thoughtfully supports orderly movement, minimizes distractions, and provides for efficient use of available space.

Positive Classroom Climate

Effective teachers build a classroom climate where error (i.e., risk-taking) is welcomed, student questioning is high, engagement is the norm, and students can gain reputations as effective learners (Hattie, 2009). Positive classroom climate has been found to be related to important educational outcomes such as enhanced academic achievement, constructive learning processes, and reduced emotional problems. Nevertheless, classroom climate can also be negative and toxic and related to undesirable outcomes, such as increased bullying and aggression and social and emotional maladjustment (Evans et al., 2009).

Classroom Talk

The interaction between teacher and students, and among students, is another significant indicator of learning environment. Authority is more distributed than centralized in a positive classroom environment. Additionally, the talk between teacher and student is personalized and personal. Exemplary teachers have been found to use authentic conversation to learn about students and encourage students to engage their peers' ideas (Allington and Johnston, 2000).

Instructional Delivery

Instructional delivery is a process in which teachers apply a repertoire of instructional strategies, to communicate and interact with students around academic content, and to support student engagement. An array of studies has found that the actual practice of teaching is a critical factor for student learning (e.g., Palardy and Rumberger, 2008). Teachers with the same background qualifications and same schooling resources do different things in their classrooms and, consequently, enable their students to achieve at different levels. To discover what makes a teacher effective, one needs to look into the black box of the classroom and see how teachers translate their content knowledge, pedagogical skills, and resources into opportunities for student learning.

Cohen et al. (2003) argued that the key causal agents for student achievement reside in instruction, and teachers' classroom teaching has the most immediate causal effects on student learning compared with other school-related factors. A

similar point was made by Palardy and Rumberger (2008). They noted that of three aspects of teachers discussed in the literature (i.e., background qualifications, personal dispositions such as attitudes, and instructional practices), instructional practices have the most proximal association with student learning. That is, "instructional practices are theorized to influence student learning directly, whereas teacher background qualifications and teacher attitudes are theorized to influence learning indirectly through their association with instructional practices" (p. 115). From a review of the literature, the following areas emerge as key instructional delivery practices.

Differentiation

Instead of using uniform strategies for all students, effective teachers design instruction that motivates each student and they communicate content in such a way that students are able to comprehend based on their individual prior learning and ability. Because students learn in a variety of ways and at a variety of rates, teachers should deliver their lessons with appropriate variety, also. A meta-analysis of the extant research suggests that instruction based on learning styles is positively related to student attitudes and achievement (Lovelace, 2005). Dunn et al. (2009) extended this finding to at-risk students, reporting that mean achievement increased nearly one standard deviation (i.e., approximately 84th percentile 50th percentile) when teachers accommodated for learning styles. Implementing a variety of classroom techniques and strategies also enhances student motivation and decreases discipline problems (Dolezal et al., 2003). Furthermore, differentiated instruction enables teachers to adjust their curriculum, materials, learning activities, and assessment techniques to ensure that all students in a mixed-ability classroom can have different avenues to process new knowledge and develop skills, while having equal access to high-quality learning (Tomlinson, 2001).

Cognitive Challenge

The teacher provides in-depth explanations of academic content and covers higher-order concepts and skills thoroughly (Wenglinsky, 2002). Kunter et al. (2008) found that teachers who engaged students in higher levels of cognitive challenge were more enthusiastic about teaching and therefore engaged in high-quality instructional behavior.

Student Engagement

In a more recent review of the literature, researchers found that student engagement is crucial for student learning, varies over time, and can be positively or negatively impacted (Finn and Zimmer, 2012). The teacher is supportive and persistent in keeping students on task and encouraging them to actively integrate new information with prior learning. In addition, the teacher uses effective questioning, smooth transition, and challenging but interesting activities to increase student engagement in learning and student accountability (Cotton, 2000; Wang et al., 1994).

Questioning

Questioning can be another highly effective instructional tool when used properly (Guo et al., 2007). In particular, the types of questions asked, wait time, and types of responses play a role in the propitious use of questioning (Walsh and Sattes, 2005). There are substantial differences in the adept use of questioning between effective teachers and ineffective teachers. On the negative side, in a study of mathematics classrooms Craig and Cairo (2005) found that teachers ask more than 99% of the questions. They also found that teachers tended to provide little wait time, asked recall and use questions, and designated a particular student to answer a question. On the positive side, one case study found that teachers deemed effective asked approximately seven times more higher cognitive-level questions than those considered ineffective (Stronge et al., 2008).

Relevance

Making instruction relevant to real-world problems is among the most powerful instructional practices a teacher can use to increase student learning (Schroeder et al., 2007; Wenglinsky, 2004). This kind of instruction allows students to explore, inquire, and meaningfully construct knowledge of real problems that are relevant to their life. Moreover, students are motivated and engaged when their learning is authentic, especially when the real-world tasks performed have personalized results.

Assessment of and for Learning

The practice of assessing student progress is essential for effective instruction and learning. High-quality assessment provides teachers with the information regarding the extent to which students have attained the intended learning outcomes, and it informs teachers' instructional decision-making (what to teach and how to teach) as well. Assessing student learning involves engaging in formative assessment practices, monitoring student progress and making decisions based on the progress, and providing meaningful feedback.

Formative Assessment

Over the past 20 years, research out of the Assessment Reform Group has informed and heightened a focus on the importance of assessment in the classroom. Research has indicated that teachers who introduce formative assessment into their classroom practice can affect substantial achievement gains. In their 1998 research review, Black and Wiliam (1998) examined a multitude of empirical studies to determine whether improvement in classroom assessments can lead to improvement in learning. They found that formative assessment has substantial positive effects on student achievement, with effect size ranging from 0.3 to 0.7 standard deviations. Particularly, they found that formative assessment is more effective for low achievers than other students, thus, reducing an achievement gap while raising achievement overall at the same time.

Progress Monitoring

Student progress monitoring is a practice that helps teachers use student performance data to continuously evaluate the effectiveness of their teaching and make more informed instructional decisions (Safer and Fleischman, 2005). Fuchs et al. (1984) used an experimental design to investigate the effects of frequent curriculum-based assessments. Analyses indicated that

- teachers in the experimental group, who adopted systematic assessment procedures, effected greater student achievement than those who used conventional monitoring methods;
- teachers in the experimental group had more improvement in their instructional structure; and
- experimental teachers' pedagogical decisions reflected greater realism and responsiveness to student progress.
- The students taught by experimental teachers were more knowledgeable of their own learning and more conscious of learning goals and progress.

Student progress monitoring facilitates teachers tracking students' academic growth on a regular basis. It can continually provide teachers with the data and evidence about students' performance to evaluate the effectiveness of their instruction and make adjustments in their pedagogical behavior.

Teachers who monitor their students' progress exhibit greater concerns about student learning and higher academic emphasis in their instruction. They also are better at supervising the adequacy of student learning, identifying students in need of additional or different forms of instruction, and determining what instructional modifications are necessary. Progress monitoring also can help teachers set meaningful student achievement goals to tap into greater student potential of learning. Empirical research found that when progress monitoring is combined with goal-setting, student learning profiles, and appropriate instructional modifications, it could help teachers build stronger instructional programs that are more varied and more responsive to students' learning needs, and effect better academic performance for students (Fuchs and Fuchs, 2003).

Feedback

Feedback is part of the formative assessment process and has emerged as a crucial part of a teacher's work. Hattie (2009) found that compared to their ineffective colleagues, effective teachers were adept at monitoring student problems and assessing their level of understanding and progress, and they provided much more relevant, useful feedback. Indeed, after reviewing meta-analyses of the myriad possible influences on student achievement, Hattie (2009) found that feedback was one of the most significant with an effect size of 0.73.

Conclusion

Studies on teacher effectiveness provide some insights into the qualities that connect teacher effectiveness and student achievement. The outcomes of these studies, while informative, have not led to a standard definition of teacher effectiveness. Neither have they generated a commonly agreed-upon list of effective teaching qualities. Generally, effective teachers plan carefully, use appropriate materials, communicate goals to students, maintain a brisk pace, assess student work regularly, and use a variety of teaching strategies. They use class time well and have coherent strategies for instruction. They hold the expectations that their students can learn and believe that they have a large responsibility to help (Cohen et al., 2003). This list is far from complete. In essence, teaching is a highly complex work with a multitude of teacher-related, as well as other, variables that affect student success. However, one clear message emerges from the literature: the impact of the teacher on student learning.

See also: Classroom Assessment; Classroom Climate; Classroom Level in Educational Effectiveness Research; Teacher Expertise; Teaching as a Profession: United States.

Bibliography

Allington, R.L., Johnston, P.H., 2000. What do we know about effective fourth-grade teachers and their classrooms? The National Research Center on English Leaning & Achievement, State University of New York, Albany, NY.

Auwarter, A.E., Aruguete, M.S., 2008. The effect of student gender and socioeconomic status on teacher perceptions. Journal of Educational Research 101, 243–246.

Barber, M., Mourshed, M., 2007. How the World's Best-Performing School Systems Come Out on Top. McKinsey & Company, London. http://www.mckinsey.com/locations/ukireland/publications/pdf/Education_report.pdf (accessed 10.11.15.).

Bettencourt, E.M., Gillett, M.H., Gall, M.D., Hull, R.E., 1983. Effects of teacher enthusiasm training on student on-task behavior and achievement. American Educational Research Journal 20 (3), 435–450.

Black, P.J., Wiliam, D., 1998. Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice 5 (1), 7–73.

Cameron, C.E., Connor, C.M., Morrison, F.J., Jewkes, A.M., 2008. Effects of classroom organization on letter-word reading in first grade. Journal of School Psychology 46, 173–192.

Cassidy, W., Bates, A., 2005. "Drop-outs" and "push-outs": finding hope at a school that actualized the ethic of care. American Journal of Education 112, 66–102.

Clare, L., 2001. Exploring the technical quality of using assignments and student work as indicators of classroom practice. Educational Assessment 7, 39–59.

Cohen, D.K., Raudenbush, S.W., Ball, D.L., 2003. Resources, instruction, and research. Educational Evaluation and Policy Analysis 25 (2), 119–142.

Collinson, V., Killeavy, M., Stephenson, H.J., 1999. Exemplary teachers: practicing an ethic of care in England, Ireland, and the United States. Journal for a Just and Caring Education 5 (4), 349–366.

Cotton, K., 2000. The schooling practices that matter most. Northwest Regional Educational Laboratory and Association for Supervision and Curriculum Development, Portland, OR and Alexandria, VA.

Craig, J., Cairo, L., December 2005. Assessing the Relationship between Questioning and Understanding to Improve Learning and Thinking (QUILT) and Student Achievement in Mathematics: A Pilot Study. Appalachia Educational Laboratory, Inc., Charleston, WV.

Darling-Hammond, L., Holtzman, D.J., Gatlin, S.J., Heilig, J.V., 2005. Does teacher preparation matter? Evidence about teacher certification, teach for America, and teacher effectiveness. Educational Policy Analysis Archives 13, 1–47.

Dolezal, S.E., Welsh, L.M., Pressley, M., Vincent, M.M., 2003. How third-grade teachers motivate student academic achievement. The Elementary School Journal 103, 239–267.

Dunn, R., Honigsfeld, A., Shea Doolan, L., Bostrom, L., Russo, K., Schiering, M.S., et al., 2009. Impact of learning-style instructional strategies on students' achievement and attitudes: perceptions of educators in diverse institutions. Clearing House: A Journal of Educational Strategies, Issues, and Ideas 82 (3), 135–140.

Emmer, E.T., Stough, L.M., 2001. Classroom management: a critical part of educational psychology, with implications for teacher education. Educational Psychologist 36 (2), 103–112.

Evans, I.M., Harvey, S.T., Buckley, L., Yan, E., 2009. Differentiating classroom climate concepts: academic, management, and emotional environments. New Zealand Journal of Social Sciences Online 4, 131–146.

Finn, J.D., Zimmer, K.S., 2012. Student engagement: what is it? Why does it matter? In: Christenson, S., Reschly, A., Wylie, C. (Eds.), Handbook of Research on Student Engagement. Springer, New York, pp. 97–131.

Fuchs, L.S., Deno, S.L., Mirkin, P.K., 1984. The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. American Educational Research Journal 21 (2), 449–460.

Fuchs, L.S., Fuchs, D., 2003. What Is Scientifically-Based Research on Progress Monitoring? National Center on Student Progress Monitoring, Washington, DC.

Goldhaber, D., 2002. The mystery of good teaching. Education Next 2 (1), 50–55. http://www.hoover.org (accessed 13.11.08.).

Goldhaber, D.D., Brewer, D.J., 1997. Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. Journal of Human Resources 32, 502–523.

Goldhaber, D.D., Brewer, D.J., 2000. Does teacher certification matter? High school certification status and student achievement. Educational Evaluation and Policy Analysis 22, 129–145.

Good, T.L., Brophy, J.E., 2008. Looking in Classrooms, tenth ed. Allyn & Bacon, Boston.

Guo, S., Tsai, C., Chang, F.M., Huang, H., 2007. The study of questioning skills on teaching improvement. The International Journal of Learning 14 (8), 141–145.

Hanushek, E.A., Kain, J.F., Rivkin, S.G., August 1998. Teachers, Schools, and Academic Achievement. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w6691 (accessed 13.11.05.).

Harris, D.N., Sass, T.R., 2007. Teacher Training, Teacher Quality and Student Achievement. National Center for Analysis of Longitudinal Data in Education Research, Washington, DC. www.caldercenter.org/PDF/1001059_Teacher_Training.pdf (accessed 13.11.05.).

Hattie, J., 2009. Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Student Achievement. Routledge, New York.

Hill, H.C., Rowan, B., Ball, D.L., 2005. Effects of teachers' mathematical knowledge for teaching on student achievement. American Educational Research Journal 42, 371–406.

Interstate New Teacher Assessment and Support Consortium (InTASC), 2013. InTASC Model Core Teaching Standards and Learning Progressions for Teachers 1.0. Author, Washington, DC.

Koh, K., Luke, A., 2009. Authentic and conventional assessment in Singapore schools: an empirical study of teacher assignments and student work. Assessment in Education: Principles, Policy & Practice 16, 291–318.

Kunter, M., Baumert, J., Koller, P., 2007. Effective classroom management and the development of subject-related interest. Learning and Instruction 17, 494–509.

Kunter, M., Tsai, Y., Klusmann, U., Brunner, M., Krauss, S., Baumert, J., 2008. Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. Learning and Instruction 18, 468–482.

LePage, P., Darling-Hammond, L., Akar, H., Guitierrez, C., Jenkins-Gunn, E., Rosebrock, K., 2005. Classroom management. In: Darling-Hammond, L., Bransford, J. (Eds.), Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do. Jossey-Bass, San Francisco, pp. 327–357.

Long, J.F., Hoy, A.W., 2006. Interested instructors: a composite portrait of individual differences and effectiveness. Teaching and Teacher Education 22 (3), 303–314.

Lovelace, M.K., 2005. Meta-analysis of experimental research based on the Dunn and Dunn model. Journal of Educational Research 98 (3), 176–183.

Ludtke, O., Robitzsch, A., Trautwein, U., Kunter, M., 2009. Assessing the impact of learning environments: how to use student ratings of classroom or school characteristics in multilevel modeling. Contemporary Educational Psychology 34, 120–131.

Marzano, R.J., 2007. The Art and Science of Teaching: A Comprehensive Framework for Effective Instruction. Association for Supervision and Curriculum Development, Alexandria, VA.

Monk, D.H., 1994. Subject area preparation of secondary mathematics and science teachers and student achievement. Economics of Education Review 13, 125–145.

National Board for Professional Teaching Standards, 2012. National Board Standards. Retrieved from: http://www.nbpts.org/ (accessed 13.11.14.).

Noddings, N., 1992. The Challenge to Care in Schools. Teachers College Press, New York.

Palardy, G.J., Rumberger, R.W., 2008. Teacher effectiveness in first grade: the importance of background qualifications, attitudes, and instructional practices for student learning. Educational Evaluation and Policy Analysis 30, 111–140.

Panasuk, R., Stone, W., Todd, J., 2002. Lesson planning strategy for effective mathematics teaching. Education 122 (714), 808–827.

Rice, J.K., 2003. Teacher Quality: Understanding the Effectiveness of Teacher Attributes. Economic Policy Institute, Washington, DC.

Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. Econometrica 73, 417–458.

Rowan, B., Chiang, F.S., Miller, R.J., 1997. Using research in employees' performance to study the effects of teachers on student achievement. Sociology of Education. 70, 256–284.

Rowan, B., Correnti, R., Miller, R.J., 2002. What large-scale, survey research tells us about teacher effects on student achievement: insights from the prospects study of elementary schools. Teachers College Record 104, 1525–1567.

Safer, N., Fleischman, S., 2005. How student progress monitoring improves instruction. Educational Leadership 62 (5), 81–83.

Schroeder, C.M., Scott, T.P., Tolson, H., Huang, T., Lee, Y., 2007. A meta-analysis of national research: effects of teaching strategies on student achievement in science in the United States. Journal of Research in Science Teaching 44, 1436–1460.

Schwarzer, R., Hallum, S., 2008. Perceived teacher self-efficacy as a predictor of job stress and burnout: mediation analyses. Applied Psychology: An International Review 57, 152–171.

Stronge, J.H., 2007. Qualities of Effective Teachers, second ed. Association for Supervision and Curriculum Development, Alexandria, VA.

Stronge, J.H., Ward, T.J., Grant, L.W., 2011. What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. Journal of Teacher Education 62, 339–355.

Stronge, J.H., Ward, T.J., Tucker, P.D., Hindman, J.L., 2008. What is the relationship between teacher quality and student achievement? An exploratory study. Journal of Personnel Evaluation in Education 20, 165–184.

Tomlinson, C.A., 2001. How to Differentiate Instruction in Mixed-Ability Classrooms, second ed. Association for Supervision and Curriculum Development, Alexandria, VA.

Tschannen-Moran, M., Hoy, W.A., 2001. Teacher efficacy: capturing an elusive construct. Teaching and Teacher Education 17 (7), 783–805.

Tschannen-Moran, M., McMaster, P., 2009. Sources of self-efficacy: four professional development formats and their relationship to self-efficacy and implementation of new teaching strategy. The Elementary School Journal 110 (2), 228–245.

Walls, R.T., Nardi, A.H., von Minden, A.M., Hoffman, N., 2002. The characteristics of effective and ineffective teachers. Teacher Education Quarterly 29 (1), 39–48.

Walsh, J.A., Sattes, B.D., 2005. Quality Questioning: Research-Based Practice to Engage Every Learner. Corwin Press, Thousand Oaks, CA.

Wang, M.C., Haertel, G.D., Walberg, H.J., 1994. What helps students learn? Educational Leadership 51 (4), 74–79.

Warren, S.R., 2002. Stories from the classrooms: How expectations and efficacy of diverse teachers affect the academic performance of children in poor urban schools. Educational Horizons 80 (3), 109–116.

Wayne, A.J., Youngs, P., 2003. Teacher characteristics and student achievement gains: a review. Review of Educational Research 73, 89–122.

Wenglinsky, H., 2002. How schools matter: the link between teacher classroom practices and student academic performance. Education Policy Analysis Archives 10 (12). http://epaa.asu.edu/epaa/v10n12/ (accessed 13.10.26.).

Wenglinsky, H., 2004. The link between instructional practice and the racial gap in middle schools. Research in Middle Level Education Online 28 (1), 1–13.

Wentzel, K.R., 2002. Are effective teachers like good parents? Teaching styles and student adjustment in early adolescence. Child Development 73 (1), 287–301.

Yinger, R.J., 1980. A study of teacher planning. Elementary School Journal 80, 107–127.

Yoon, K.S., Duncan, T., Lee, S.W., Scarloss, B., Shapley, K., 2007. Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Regional Educational Laboratory Southwest, Washington, DC.

Teacher Education

Ian Menter, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article traces the development of teacher education and in particular identifies a number of key debates and tensions that emerge from that history and are still relevant today. These include the nature of professional knowledge, the sites for learning, and the relationship between preservice, induction, and continuing professional development. The quality of teaching is now seen to be a key factor in the success of educational systems and the nature of teacher preparation and development has assumed great significance in the twenty-first century.

Introduction

In the early twenty-first century, teacher education has become a key focus within education systems across the world. This reflects growing international concern about the importance of the quality of teaching and recognition that this in turn is likely to be dependent on the quality of teacher preparation and teacher development.

This recent renewed interest in the importance of teaching and teacher education builds upon a long history of the development of institutional approaches to the preparation and ongoing development of teachers. This article therefore commences by offering a review of this history, exploring how the development of the teaching profession related to the arrangements for teacher education. Consideration is then given to a number of key issues in teacher education, as currently constructed, such as the variety of routes of entry into the profession, the significance of the contribution of different sites of learning for teachers, and the idea of teacher professional development and learning as a career-long continuum. Finally we reconsider the connections between teacher education and the globalization of education and the latest developments in 'integrated' approaches, which seek to develop teaching as a research-oriented profession.

History

There is no single history of teacher education around the world. There are notable differences between continents as well as within continents (e.g., Dent, 1970; Sedlak, 2008). Until the nineteenth century, however, it would be reasonable to suggest that the teaching profession was closely associated with universities and with religious organizations. It was the emergence of state provision for schooling in the late nineteenth century that led to the rapid expansion of systematic approaches to the preparation of teachers in many locations. In particular the establishment of 'normal schools' was a feature in Europe as well as in North America. These schools not only provided education for children but were also the sites for an apprenticeship style of training for prospective teachers.

As state education systems expanded, so Colleges of Education began to emerge, enabling the creation of larger classes of study for trainee teachers and sometimes the provision of experimental classrooms or schools where groups of trainees could observe teaching and be evaluated by others in their own teaching.

The twentieth century saw the rise of social science disciplines including psychology and sociology and the contribution that these subjects might make to the education of teachers was recognized. So we began to see the steady 'academization' of teacher education, the introduction of a body of specialized professional knowledge that was judged to be important. As this process continued, so the direct involvement of universities in teacher preparation increased (see Alexander et al., 1984). This was sometimes through the processes of validation or accreditation of teacher education programs but also increasingly through direct provision of programs by universities.

This in turn was associated with an increasing 'professionalization' of teaching and thus, for example, in England we saw by the early 1970s the call for teaching to become an 'all-graduate' profession (James, 1972). So, while many secondary schoolteachers had a first degree before they entered teacher training, many teachers working with younger children had entered teaching with some form of certificate that had not led to a degree as such. But this development saw the emergence of Bachelor of Education degrees as a major route into teaching for primary teachers, as well as for some secondary schoolteachers.

At the same time as this was happening, a large number of Colleges of Education were being amalgamated in a program of 'rationalization' and then were being incorporated into universities or polytechnics. The polytechnics themselves were converted into universities late in the twentieth century, so that by the end of the twentieth century, teacher education – especially initial or preservice teacher education – may be seen to have gone through a process of 'universitization' (Menter et al., 2006).

While the move toward the academy as a key contributor to the education of teachers was a common trend in most parts of the world, the late twentieth century saw some quite significant disruption in this trajectory. In the United States and in England in particular there has been a surge in the introduction of 'alternative' routes into teaching. Sometimes these have arisen in the context of teacher shortages, that is the supply of teachers failing to meet the numbers required by the schools, but quite often these innovations have been associated with ideological critiques of the nature of university-led teacher education. These themes will be picked up later in this article.

The focus on teacher education has been intensified by the publication of a number of transnational reports on teaching and attainment. The Programme for International Student

Assessment studies published by the Organisation of Economic Cooperation and Development (OECD) have heightened policymakers' concerns about the attainment of young people in core subjects across the world. The OECD has also been carrying out an international survey TALIS (Teaching and Learning International Survey), which has offered comparisons of the experiences and needs of teachers across many nations (OECD, 2009). Research from the school improvement and school effectiveness movement has indicated that the quality of teaching is a key factor in raising school level attainment and then transnational reviews, especially those carried out by the McKinsey corporation (McKinsey and Co., 2007; Barber and Mourshed, 2009) have suggested that the quality of teaching is one aspect where politicians may have a real impact on the quality of their schools.

Entry Routes into Teaching

The history of teacher education has seen a wide range of approaches to entry into the teaching profession. During the twentieth century the move to professionalize did lead to some consolidation of the association of teacher preparation with formal academic qualifications. So, in Britain and North America, the emergence of qualifications such as Bachelor of Education or Postgraduate Certificate in Education (PGCE) was seen as providing a recognized teaching qualification. However, the route through which these could be achieved did vary considerably.

The first distinction to be made is between the undergraduate and postgraduate route. Undergraduate entry is aimed at people who have not yet got a first degree and who have decided before doing so that they do want to become teachers. Whereas a postgraduate route is for people who have undertaken a first degree (typically a BA or BSc) and now wish to spend time preparing to become a teacher.

The undergraduate routes may be divided into two main types, described as concurrent and consecutive. A concurrent degree would involve the student undertaking study and practice of education throughout the program at the same time as studying a specialist subject at degree level. By contrast a consecutive degree would involve studying the specialist subject for a period of time – perhaps 2 years, followed by another period during which the emphasis is on professional preparation for teaching, this involving placement in school as well as study of education. The professional parts of such programs typically have three main components, educational studies, subject pedagogy, and the practicum.

In postgraduate programs, which are typically less than 1 year long, one would not expect to see separate study of the subject but only of the three professional components, as set out above.

As mentioned above, however, the recent diversification of routes has added much more variety to what is offered. So, we have seen the shortening of degree entry routes in some places. Where a BEd with Honors in the United Kingdom was traditionally 4 years long, in England there have been shortened routes of 3 or even 2 years. Some routes, especially postgraduate routes, have sometimes been available for part-time study – although there is usually a full-time practicum element at some stage.

However, even greater variation has arisen from school-based or employment-based routes. In the United States we have seen 'alternative certification' entry being offered in some states or in England we have seen 'School-Centred Initial Teacher Training Schemes' and more recently 'School Direct (salaried)' programs in which the school takes the lead and in the latter case at least, the trainee is actually an employee at the school during their training and may expect to continue to work in the school once they have qualified.

Another variation on employment-based routes is seen in the form of Teach for America (US) or Teach First (England), schemes that are sponsored by major corporations and seek to attract people into teaching, at least for a short period, who might otherwise not have entered. The English version of this scheme has a strong emphasis on training teachers in settings where there is significant social disadvantage. The scheme aims to challenge such disadvantage by recruiting the most able students from the best universities and bringing about change in the schools (see Wigdortz, 2012). In the case of Teach First, the training does involve some university provision, including a preparatory 6-week summer school as well as continuing support from university staff.

At the same time as there have been moves to shorten routes of entry and reduce the input from the university, there has been a contrary trend in some settings, to increase the academic-level entry requirements. In some countries, Finland being a notable example, teaching has moved beyond being a degree-level entry profession to being at Master's level. Teaching is a highly esteemed occupation in Finland and admission to teacher education programs is highly competitive. The program of training is led by universities, each of which works in association with university training schools. Programs of study incorporate a research study and lead after 4 years to a Master's award (Sahlberg, 2012).

In the United Kingdom, there were significant moves in this direction in the early part of this century. Many PGCE awards now include a proportion of Master's level work and so newly qualified teachers emerged with a certain number of 'M Level' credits that they can then use as the basis for further study toward a Master's award.

Sites of Learning in Teacher Education

Many of the debates about the form and structure of teacher education programs have centered on questions about the nature of teaching and what forms of knowledge, skills, and experience are required in order to fulfill this definition. In a review of literature of teacher education in the twenty-first century, Menter et al. (2010) suggested that it is possible to identify four paradigms of teaching, each of which will lead to rather different approaches to the formulation of pre-entry programs. These are effectively 'ideal types,' which may not exist in a pure form in any given context and indeed actual models of teaching deployed in teacher education programs are likely to draw on more than one of them:

1. The effective teacher – with an emphasis on technical skills.
2. The reflective teacher – with an emphasis on values and review.

3. The inquiring teacher – with the adoption of a research orientation.
4. The transformative teacher – with the adoption of a 'change agency' approach.

Moving from the first to the fourth, each paradigm incorporates those with a lower number but builds upon it. These might be seen as positions on a spectrum of professionalism, which using the terminology developed in the 1970s by Hoyle (1974) moves from 'restricted' professionalism at one end to 'extended' professionalism at the other.

At the restricted/effective end of the spectrum, there is a view that the best place to learn to teach is alongside an experienced and successful teacher, through an apprenticeship model. This is sometimes depicted as a 'craft' view of teaching.

The skills of teaching are learned by observation and by imitation and in turn by being observed and receiving feedback from the experienced teacher. On this model, knowledge of the subject content of the teaching is assumed to be present; in other words the trainee is already well versed in the subject and all they require is enthusiasm and an ability to learn from observation and feedback. If this is a limited view of becoming a teacher for a secondary schoolteacher of a particular subject it is even more challenging for the elementary or primary schoolteacher whose subject knowledge will need to range right across the school curriculum. This position has been well exemplified by a recent Secretary of State for Education in England, Michael Gove, in his foreword to a Government White Paper:

> Teaching is a craft and it is best learnt as an apprentice observing a master craftsman or woman. Watching others, and being rigorously observed yourself as you develop, is the best route to acquiring mastery in the classroom.
>
> DfE, 2010

The developments in teacher education curricula which took place over the twentieth century led to the proposition that teachers actually need a range of forms of professional knowledge, including pedagogical knowledge and subject-related pedagogical knowledge, in addition to 'pure' subject knowledge. That is, teachers need to have knowledge not only of their subject but of how their subject is best taught and learned as well as more general knowledge about teaching and learning. On this view teaching is a much more complex activity involving numerous challenging judgments and decision-making in planning, implementing, and evaluating classroom interaction. This position is actually closely aligned with that taken in a recent review of teacher education in Scotland, carried out by the former Chief Inspector, Graham Donaldson, who saw teachers as

> … reflective, accomplished and enquiring professionals who have the capacity to engage fully with the complexities of education and to be key actors in shaping and leading educational change.
>
> Donaldson, 2011: 4

The debates about the kinds of professional knowledge needed by teachers influence not just the content of the teacher education curriculum but the locations in which professional learning takes place. Over the twentieth century there was in all teacher education programs some balance of provision between academic study and professional experience. This was sometimes posited as a balance of theory and practice, although this we shall argue is a simplistic distinction as in the most mature programs what is achieved through an appropriate balance between and integration of learning in different sites is a form of 'practical theorizing' (Hagger and McIntyre, 2006). This debate in teaching is also informed by a wider sociological discussion about the nature of professionalism, that is, what are the distinctive features of a profession, as opposed to other occupational forms (see Evetts and Svensson, 2010).

In recent times there has also been development in our understanding of the nature of 'work-based learning.' The relationship between experience, understanding, and knowledge has been a complex one that is very context dependent (McNamara et al., 2014). This has heightened awareness of the importance of creating an appropriate learning environment in the practicum setting as well as in the university.

In order to achieve an effective balance of the academic and the professional, it has come to be recognized in many – but not all – quarters that a crucial element in teacher education is the partnership between the university and the schools in which students are experiencing their field practice or practicum. In different settings such partnerships have been developed in different ways. Furlong et al. (2000) distinguished between complementary and cooperative partnerships. The latter approach draws schoolteachers and university tutors closer together in sharing own leadership of the whole program, whereas in the complementary approach the key element is that the school-based and university staff have very distinctive and different roles. In cooperative programs school staff will experience training for their roles alongside university staff and may take such roles as 'subject mentor' or 'professional tutor.'

The Continuum of Teacher Education

Thus far we have tended to focus on the first stages of teacher education, that of the preservice or initial phase. However one of the important themes to emerge from research over many years now has been to recognize that beginning teachers can only learn the beginnings of what they should be capable of at this stage. Such is the complexity and scope of learning required that the teachers' professional career is increasingly seen as the basis for lifelong learning. This is frequently seen as a series of phases or stages in teacher learning. Each phase or stage will require somewhat different approaches as well as focusing on different aspects of teachers' knowledge and skills.

Usually, it is suggested – and sometimes required – that following initial qualification, teachers should then enter a phase of 'induction,' typically the first 1 or 2 years of being employed as a teacher in a school. During this period, they may have a timetable commitment that is less than the full one expected of experienced teachers and they may be assigned a mentor who will work with them to help them build upon what they have learned in their initial program. The usual

emphasis here is to ensure that they are learning to understand the particular approaches adopted in their own school and are deepening their understanding of the particular learning needs of students within that setting.

Having established this grounding in their school, the next phase – typically 2–5 years of the career – is typified as the Early Professional Development phase. For some teachers this may be an opportunity to deepen their theoretical understanding of learning and teaching, perhaps through a Masters' program, and/or to broaden their curriculum knowledge, and/or to develop new roles in addition to their class or subject teaching, such as pastoral roles. Such development may often be associated with taking an increasingly inquiry-oriented approach to their teaching, effectively carrying out small-scale school-based research (Forde et al., 2006; Campbell et al., 2003).

Subsequent professional development, often referred to as Continuing Professional Development, may take a wide variety of forms and may be more dependent on how the teacher seeks to develop their careers. Do they wish to become a specialist teacher within a subject or with a particular age range or are they aspiring to take on formal leadership roles within the school perhaps as a middle leader and eventually as a senior leader? In most settings there are particular programs available for such leadership development.

Teaching Quality and Teacher Education – A Global Agenda?

International reviews of current trends in teacher education very clearly demonstrate how much 'convergence' is taking place, at least in certain aspects (e.g., Darling-Hammond and Lieberman, 2012; Townsend, 2011). We have seen the impact of international assessment surveys and we have seen how many national reviews have been commissioned to that have been tasked with seeking improvements to provision in teacher education. For example, from the Holmes Group in the United States in the 1980s (Holmes, 1986), more recent reviews in Austria during the first decade of this century, and a string of reports in the Republic of Ireland (Sahlberg et al., 2012), Northern Ireland, Scotland, and Wales (see Beauchamp et al., 2013).

As ideas of 'the knowledge economy' and the competitiveness associated with 'globalization' have developed, so has the politicization of education generally (Rizvi and Lingard, 2009), but not least in teacher education. Politicians across the world emphasize the importance of teachers and teaching (the UK Government White Paper of 2010 is actually called *The Importance of Teaching* (DfE, 2010)).

The emphasis on performance has moved from a focus on pupils/learners to include teachers as well. Education systems worldwide are imbued with an ethos of 'performativity' (Jeffrey and Troman, 2012), which has tended to lead toward making sure that all performance can be measured. This has been a factor in the creation of sets of standards to define teaching. These typically amount to a list of competences that teachers should be capable of demonstrating, according to the particular phase of their career. Sometimes these standards are based on explicit value commitments that refer to the purposes of education and to principles such as justice, equity, and integrity.

In some settings, such standards and competence statements are developed largely by members of the profession and/or their organizations. Elsewhere governments have taken it upon themselves to define these matters, which has tended to lead to a loss of goodwill and a lack of trust in the relationship between policymakers and teachers (Mahony and Hextall, 2000).

One of the major difficulties for teachers arising from these trends is the tendency for politicians to assume that what works in one setting will work in their own and so the idea that policies developed successfully in one context can be simply borrowed and introduced in another context and be as successful there has led to some difficulties.

It has to be said that the work of the McKinsey Corporation in sponsoring major international reviews has tended to contribute to this. Similarly the OECD in their country studies as well as in transnational surveys have tended to create an assumption that what works there will work here – even the titles of some of the reports tend to demonstrate this, e.g., *How the world's best-performing education systems come out on top*.

Research and Development in Teacher Education

If there has been a growing tendency for convergence in approaches to teacher education especially in the developed world, that is to some extent countered by some quite local initiatives which seek to offer creative and innovative alternatives to this 'homogenization' and it is these alternatives that this article will conclude by considering.

From the groundbreaking insights of John Dewey onward and often incorporating developments in learning theory, there has been a growing but eclectic body of practice that has sought to address some of the key challenges within teacher education very directly. These challenges include the nature of professional knowledge and the best sites for learning as a teacher, matters covered earlier in this article.

In a paper commissioned for an Inquiry into Research and Teacher Education by the British Educational Research Association and the Royal Society for the Arts, Burn and Mutton (2013) consider what is meant by evidence-informed clinical practice in teacher education. They identify a number of programs in the United Kingdom, the Netherlands, Finland, and the United States, which have sought to integrate theoretical and practical knowledge in the learning of beginning teachers. Like many other reviews they come to the conclusion that it is very difficult to demonstrate the increased efficacy of such approaches in terms of pupil outcomes, but they do provide a very strong rationale for such approaches offering the most fully developed approach to teacher professionalism that is available.

They cite the Oxford internship scheme as an early example of these approaches. This was developed in the 1980s explicitly as a partnership between a university department of education and about 30 secondary schools. The whole scheme was developed and managed by the partnership and continues to this day to provide a well-integrated model of professional learning. In the United States, they refer to Professional

Development Schools as an attempt to do something similar, as well as to the Teachers for a New Era (TNE) initiative, an attempt to develop clinical practice in 11 settings. TNE provided the inspiration for Scottish Teachers for a New Era at Aberdeen, which along with the Glasgow West initiative, are two Scottish schemes referred to.

Schemes such as these provide a basis for teachers to set out on their careers as inquiring professionals, so that they can subsequently develop a wider range of knowledge, understanding, and skills that enable them to enhance their roles as teachers, so that they may see teaching as an activist profession and may adopt an inquiry orientation as a stance to their work (Cochran-Smith and Lytle, 2009). The systematic deployment of research skills within the work of a teacher is not a new idea however – in the United States, John Dewey established a teaching laboratory at the University of Chicago in 1896 (Mayhew and Edwards, 2007) and in the United Kingdom, Lawrence Stenhouse (1975) developed the ideas of teacher as research and as curriculum developer.

Conclusion

In short there are a number of key points of contention about the purpose, function, and nature of teacher education. These debates have been a characteristic of teacher education since its inception as an societal institution. Many of the debates though have been heightened in recent times through the increased competitiveness between nations and the increased awareness of the significance of education in a globalized knowledge economy. To a large extent the underlying debate is about who should decide on policy and practice in teacher education. In democratic societies there is undoubtedly a role for elected representatives, but what is the role of 'consumers,' whether young people in schools or teachers in training, and what should be the roles of those with a direct involvement in the processes, that is the teacher educators, whether they are based in universities or in schools?

Bibliography

Alexander, R., Craft, M., Lynch, J. (Eds.), 1984. Change in Teacher Education: Context and Provision Since Robbins. Holt, Rinehart and Winston, London.

Barber, M., Mourshed, M., July 6, 2009. Shaping the Future: How Good Education Systems Can Become Great in the Decades Ahead: Report on the International Education Roundtable. McKinsey & Company: Mckinsey Education, Singapore.

Beauchamp, G., Clarke, L., Hulme, M., Murray, J., 2013. Policy and Practice within the United Kingdom. Review Paper for the BERA/RSA Inquiry into Research and Teacher Education. BERA/RSA, London. Available at: http://www.bera.ac.uk/resources/research-and-teacher-education-bera-rsa-inquiry.

Burn, K., Mutton, T., 2013. Review of 'Research-Informed Clinical Practice' in Initial Teacher Education. BERA/RSA, London. Available at: http://www.bera.ac.uk/resources/research-and-teacher-education-bera-rsa-inquiry.

Campbell, A., McNamara, O., Gilroy, P., 2003. Practitioner Research and Professional Development in Education. Sage, London.

Cochran-Smith, M., Lytle, S., 2009. Inquiry as Stance. Practitioner Research for the Next Generation. Teachers' College Press, New York.

Darling-Hammond, L., Lieberman, A. (Eds.), 2012. Teacher Education around the World. Routledge, London.

Dent, H., 1970. The Training of Teachers in England and Wales 1800–1975. Hodder and Stoughton, London.

Department for Education, 2010. The Importance of Teaching. The Stationery Office, London.

Donaldson, G., 2011. Teaching – Scotland's Future. The Scottish Government, Edinburgh.

Evetts, J., Svensson, L. (Eds.), 2010. Sociology of Professions. Continental and Anglo-Saxon Traditions. Daidalos, Sweden.

Forde, C., McPhee, A., McMahon, M., Patrick, F., 2006. Professional Development, Reflection and Inquiry. Paul Chapman, London.

Furlong, J., Barton, L., Miles, S., Whiting, C., Whitty, G., 2000. Teacher Education in Transition. Open University Press, Buckingham.

Hagger, H., McIntyre, D., 2006. Learning Teaching from Teachers: Realising the Potential of School-Based Teacher Education. Open University Press, Maidenhead.

Holmes Group, The, 1986. Tomorrow's Teachers: A Report of the Holmes Group. The Holmes Group, East Lansing, MI.

Hoyle, E., 1974. Professionality, professionalism and control in teaching. London Education Review 3 (2), 13–19.

James, Lord, 1972. Teacher Education and Training (The James Report). Her Majesty's Stationery Office, London.

Jeffrey, B., Troman, G. (Eds.), 2012. Performativity in Education in the UK: Effects, Consequences and Agency. E&E Publishing, Stroud, Gloucestershire.

Mahony, P., Hextall, I., 2000. Reconstructing Teaching. Routledge, London.

Mayhew, K., Edwards, A., 2007. The Dewey School: The Laboratory School of the University of Chicago 1896–1903. Aldine Transaction, Piscataway, NJ.

McNamara, O., Murray, J., Jones, M. (Eds.), 2014. Workplace Learning in Teacher Education: International Practice and Policy. Springer, New York.

Menter, I., Brisard, E., Smith, I., 2006. Convergence or Divergence? Initial Teacher Education in Scotland and England. Dunedin Press, Edinburgh.

Menter, I., Hulme, M., Elliot, D., Lewin, J., 2010. Literature Review on Teacher Education in the 21st Century. Scottish Government Social Research, Edinburgh.

McKinsey & Co., (2007). How the World's Best-Performing Systems Come Out on Top. Available at: http://mckinseyonsociety.com/how-the-worlds-best-performing-scools-come-out-on-top/.

OECD, 2009. Creating Effective Teaching and Learning Environments – First Results from TALIS. Available at: http://www.oecd.org/edu/school/43023606.pdf.

Rizvi, F., Lingard, B., 2009. Globalizing Education Policy. Routledge, London.

Sahlberg, P., 2012. Teachers and teacher education in Finland. In: Darling-Hammond, L., Lieberman, A. (Eds.), Teacher Education around the World. Routledge, London.

Sahlberg, P., Furlong, J., Munn, P., 2012. Report of the International Review Panel on the Structure of Initial Teacher Education Provision in Ireland. Department of Education and Skills, Dublin.

Sedlak, M., 2008. Competing visions of purpose, practice and policy. The history of teacher certification in the United States. In: Cochran-Smith, M., Feiman-Nemser, S., McIntyre, D., Demers, K. (Eds.), Handbook of Research on Teacher Education, third ed. Routledge, New York.

Stenhouse, L., 1975. An Introduction to Curriculum Research and Development. Heinemann, London.

Townsend, T. (Ed.), 2011. Journal of Education for Teaching 37, 4 (special issue).

Wigdortz, B., 2012. Success Against the Odds. Five Lessons in How to Achieve the Impossible. The Story of Teach First. Short Books, London.

Teacher Expertise

Julian G Elliott, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Bromme, volume 23, pp. 15459–15465, © 2001, Elsevier Ltd.

Abstract

In considering the key domains of teacher expertise this article differentiates between teachers' knowledge of (1) academic subject content, (2) teaching and learning, and (3) the interpersonal dynamics of classrooms. In relation to teaching and learning, pedagogical content knowledge and associated conceptual difficulties are identified and discussed. In considering pedagogical knowledge and teachers' interpersonal skills, the tacit nature of much expert knowledge is highlighted and the implications of this for teacher professional development are considered.

Introduction

In an age dominated by high stakes testing, it might seem possible to conceptualize the expert teacher as one whose students will perform at a higher level on key academic measures than would anticipated on the basis of their prior performance. Leaving aside the many rejoinders to this assertion, such as the important impact upon student performance of multiple factors other than the teacher, and claims that there is more to high-quality education than test scores, this position tells us little about the special characteristics of the expert teacher. The difficulty of articulating a clear set of behaviors as an indicator of teaching expertise has often led researchers to draw upon easily measurable criteria, such as years of teaching experience or professional group membership, when selecting samples for investigation (Palmer et al., 2005). However, while expertise is typically the product of prolonged experience, it is possible to have many years of teaching experience but fail to learn from this (Sternberg and Hedlund, 2002).

Expertise has been defined as, "… the characteristics, skills and knowledge that distinguish experts from novices and less experienced people" (Ericsson, 2006: p.3). It is unique to a particular domain of activity and its acquisition typically takes many thousands of hours of focused effort and practice (Ericsson, 1996). Following the redundancy of early approaches to the study of expertise that examined teachers' personality traits or the nature and frequency of isolated classroom actions, the advent of cognitive approaches in psychology resulted in an emphasis upon the complex processes that underpin decision making about specific actions (Berliner, 1994), in particular the ways in which experts structure knowledge qualitatively differently to novices. Studies indicated that experts can access such knowledge in a more efficient, fluid, and faster manner to help solve novel problems, aided by superior pattern recognition enabling an ability to 'read the situation,' and are more aware of what it is that they know and do not know (Glaser and Chi, 1988).

Compared with describing complex cognitive processes, outlining the key domains of a teacher's expertise is seemingly not a difficult task. Expert teachers must have a sound understanding of the subject that they teach (subject knowledge). However, it is clear that a sound academic grasp of a subject is, in itself, no guarantee that an individual will be a competent teacher. Academic material needs to be structured, represented, and presented in a way that can make it most accessible to students with very different needs and understandings (pedagogical knowledge). While *meta*-analytic studies have offered teachers valuable information about the effectiveness of different types of teaching activity (Hattie, 2011; Higgins et al., 2013), the ability to amend day-to-day learning experiences to the particular needs-specific groups of students involves highly complex expert knowledge. Finally, teachers need to be capable of sustaining sound professional and interpersonal relationships with adults (parents and colleagues) and children (Torff and Sessions, 2005; Malm and Löfgrer, 2006), particularly for the purposes of managing and orchestrating children's classroom behavior. Indeed, this is so fundamental to create an environment in which learning can take place, one may take issue with Ball et al.'s (2008: p. 404) assertion that, "… there may be nothing more foundational to teacher competency" than subject knowledge. Even disciplinary specialists with sound insights into the relationship between their subject areas and student learning processes, and an advanced capacity to provide meaningful learning activities, will struggle if they are unable to achieve an orderly, disciplined, and engaged classroom environment.

The Pedagogical Skills of the Expert Teacher

Within the domain of teaching and learning, it is possible to differentiate between general expertise that operates across disciplinary fields and that relating to the learning of specific subject matter (pedagogical content knowledge, PCK). The conception of PCK was developed by Shulman, (1986, 1987) as one of seven core areas of teacher knowledge:

> a) Content knowledge
> b) General pedagogical knowledge
> c) Curriculum knowledge
> d) Pedagogical content knowledge
> e) Knowledge of learners and their characteristics
> f) Knowledge of educational contexts (ranging from group and classroom dynamics, governance and financing of school districts, to the character of communities and cultures
> g) Knowledge of educational ends, purposes and values and their historical and philosophical roots and influences.
>
> Shulman, 1987: p. 8

Shulman's conception concerned the 'special amalgam of content and pedagogy' (p.8) that was important for teaching a given subject or topic. His focus was upon the ways by which teachers transform subject matter knowledge to maximize student understandings. Knowledge of how this could be best achieved for a given topic in a given academic subject area in relation to the particular strengths and motivations of a given group of students, enables teachers to determine how best to structure the delivery of content, how to represent the material, and what forms of learning are best suited for the particular context. In this way, teacher expertise in relation to a given subject can be related to the number of representations and strategies they are able to generate, allied to their understanding of their students' particular learning needs.

In considering the forms of knowledge required to teach mathematics, Ball et al. (2008) built upon Shulman's conception to distinguish between two broad areas – subject matter knowledge and PCK. Subject matter knowledge consists of (1) disciplinary knowledge that is unrelated to teaching (common content knowledge), (2) that which is specific to the teaching process yet is clearly subject knowledge (for example, that which can provide understandings about particular student errors (specialized content knowledge)), and (3) knowledge of how specific topics fit within the wider span of the discipline (horizon knowledge). The value of horizon knowledge serves as a basis for claims that teachers of young children require advanced subject knowledge in key subject areas such as science and mathematics. Pedagogical content was similarly divided into three components, (1) knowledge of content and curriculum, (2) knowledge of content and students, and (3) knowledge of content and teaching. Sztajn et al. (2012) conceive of these in relation to student trajectories of learning whereby content and tasks are presented in ways that are closely aligned to students' increasingly sophisticated ways of thinking and understanding.

Currently there is no universally accepted understanding of the PCK term and, as is often the case with constructs in the social sciences that succeed in capturing the imagination, it has come to be used to describe a broader range of processes than was originally intended (for example, Ball et al.'s reformulation of Shulman's areas subsumes curriculum knowledge within PCK). There is a danger that, in some cases, the construct has broadened in scope to the point that it has become almost synonymous with teacher practices and beliefs in general (Hashweh, 2005). Van Driel and Berry (2010) note uncertainty as to whether PCK should be seen as wholly related to very specific topics or whether there are grounds for arguing that this can apply to broader disciplinary areas (for example, in relation to algebra or even the teaching of mathematics). They add that attempts to analyze PCK empirically may result in consideration of atomistic segments that may not adequately represent the complex and holistic nature of teaching.

A further difficulty of PCK, in relation to teacher expertise is that, despite much theorizing, there has been little research examining exactly how teachers actually transform subject knowledge in ways that reflect their views of their students' understandings (Van Driel and Berry, 2010). A related area of debate concerns the extent to which expert practices can be set out explicitly in ways that can inform the practice of others. Thus, Bausmith and Barry (2011) contend that PCK can be enhanced by the use of online videos of lessons taught by expert teachers. However, this does not sit easily with the emphasis upon teachers' developing understandings of their own particular contexts. As Van Driel and Berry (2012) point out, a topic such as 'force' may be taught differently, even to same-aged children, depending on the unique circumstances of a particular time and setting. The expert teacher not only modifies the approach they take according to such factors but can also make on the spot modifications in the light of changing circumstances.

The Interpersonal Skills of the Expert Teacher

A critical area of teacher expertise that is not highlighted in Shulman's seven categories concerns interpersonal dealings with students, parents, and colleagues. Creating and sustaining positive relationships with students is often perceived by teachers as the most difficult aspect of their work, particularly in schools with challenging circumstances (Johnson and Birkeland, 2003). Such a concern is particularly strong in many Western contexts where resistance to traditional authority figures has resulted in teachers feeling challenged both by students and parents (Maclure and Walker, 2000). Teachers often need to achieve disciplined environments by means of the exercise of their professional expertise rather than through the traditional sources of authority vested in them by the nation-state (Pace and Hemmings, 2007; Elliott and Tudge, 2007).

In relation to developing an orderly and engaged classroom environment, Elliott (2009) has argued that a fundamental skill of the expert teacher concerns their capacity to demonstrate authority. However, achieving this is often easier said than done and much of the relevant knowledge would seem to be tacit and context-bound. Despite this, it is possible to articulate some of the skills by which the teacher's expertise is communicated to their students.

It would seem that, when confronted by a challenging interpersonal problem, the expert teacher is more likely to identify a sound action from a range of possibilities. In an exploration of this, Elliott et al. (2011) presented novice and experienced teachers with a series of hypothetical school-related scenarios in which an interpersonal difficulty or challenge is presented. These scenarios, together with a list of possible responses, were derived from interviews with experienced teachers (see Stemler et al., 2006; Grigorenko et al., 2006). It was found that over time, novice teachers became more able to recognize undesirable responses as determined by their more experienced colleagues. However, there was little difference between the two groups in their capacity to identify 'good' responses. It appeared that the key learning for the novices centered upon avoiding poor choices. However, in considering the makeup of the samples it should be noted that experience should not be directly equated with expertise as tacit knowledge concerns the ability to learn from experience (Sternberg et al., 2000: Sternberg and Hedlund, 2002), a skill that would not appear to be the preserve of all teachers.

It is important to recognize that the expert teacher not only makes a wise choice about what steps to take in any given situation but should also carry these out with high levels of

skill. A decision for example, that one should confront a particular student presenting a problem in class requires the advanced interpersonal skill of achieving this satisfactorily. In addition, having knowledge of what might be the wisest step to take does not necessarily mean that the teacher will act in accordance with this. High levels of emotion, stress, or self-doubt may lead the teacher to respond in ways that they would not recommend if they were reflecting dispassionately and disinterestedly on possible courses of action. This may explain why there is little evidence that responses to teacher questionnaires actually predict the actual classroom behavior of teachers (Berliner, 2005).

A further difficulty with focusing upon reaction to challenging situations is that the expert teacher would appear to be not so much one who handles problems skillfully but, rather, one who prevents difficulties from emerging in the first place (Kounin, 1970). Those teachers who are presented with fewest challenges typically deploy a range of skills including an ability to: demonstrate an advanced awareness of everything taking place around them ('withitness'), orchestrate multiple events taking place at the same time ('overlapping'), operate clearly understood classroom routines, use their voice and body language in ways that signal authority and maintain student interest and alertness, (Van Tartwijk, 1993), and control spoken interactions in ways that ensure that the teacher is not overly responsive or subordinate to the child's agenda (Elliott and Place, 2012).

Teaching Expertise to Teachers and the Complexities of Tacit Knowledge

Given the vast reservoir of research that has been conducted on teacher skills, one might anticipate that it would be relatively straightforward to articulate a set of practices that is consonant with that of the expert teacher. Such a perspective is perhaps reflected by Bausmith and Barry's (2011) recommendation for the use of online videos of expert practice. Such a perspective has been criticized on the grounds that this reflects "… an outside-in, top-down, teacher-deficit model" (Anderson and Herr, 2011: p. 287) that has not been supported by research. To some extent this critique reflects a broader debate in education between those who advocate evidence-based approaches based upon models originating in the natural sciences, and others who advocate exploratory teacher action research involving reflection upon one's own professional practice, often in communion with colleagues (Feldman and Weiss, 2010).

A key difficulty for those seeking to help teachers develop classroom expertise is that the multiplicity of subtle factors, individual differences, and unique circumstances are such that attempts to offer relatively decontextualized guidance may lack sufficient meaningfulness, resonance, relevance, or value. Furthermore, unlike everyday routinized classroom practices and procedures (e.g., undertaking formal testing, managing student entry into the classroom, setting out equipment in the science laboratory) more complex professional understandings are often tacit and thus not easily made explicit as a set of guiding rules for action (Schön, 1983; Sternberg and Horvath, 1995).

Eraut (2000: p.133) has outlined a number of situations where tacit knowledge may be used and/or acquired in professional contexts:

1. Knowledge acquired by implicit learning of which the knower is unaware;
2. Knowledge constructed from the aggregation of episodes in long-term memory;
3. Knowledge inferred by observers to be capable of representation as implicit theories of action, personal constructs, schemas, etc.;
4. Knowledge which enables rapid, intuitive understanding or response;
5. Knowledge entailed in transferring knowledge from one situation to another; and
6. Knowledge embedded in taken-for-granted activities, perceptions and norms.

As this list demonstrates, rendering such knowledge formal and explicit is not an easily achievable undertaking. According to Sternberg and his colleagues (Sternberg and Horvath, 1995; Sternberg, 1997; Sternberg and Hedlund, 2002), and following the seminal work of Polanyi (1958), tacit knowledge is primarily acquired, not by direct instruction from others but, rather, from one's experience of operating within a given context. It typically takes the form of procedural knowledge that is concerned with how particular tasks are best undertaken in specific contexts. Furthermore, it is closely related to one's own goals and agendas and thus what might be understood as the most appropriate response to a given classroom incident will be influenced by the teacher's own beliefs and values.

Given these characteristics, it is hardly surprising that much of teachers' expert knowledge is not easily taught (Anderson, 1982) although, given that tacit knowledge is derived through experience, it seems that it can be facilitated by educational approaches that involve formal inputs, opportunities to experiment in practice, and the operation of teacher groups collectively thinking and reflecting upon their practice (Sternberg et al., 2000; Cianciolo et al., 2006; Van Driel and Berry, 2012).

In conclusion, while it is a truism that classroom learning will typically only be effective where the teacher seeks to exercise appropriate authority, demonstrates enthusiasm for the relevant academic material, and exhibits a strong desire to maximize the learning progress of the students, such dispositions are often insufficient. Additionally, it is the exercise of high-level professional expertise, not easily encapsulated by a set of straightforward practices that, while crucial, can prove elusive for some practitioners.

See also: Classroom Level in Educational Effectiveness Research; Continuing Professional Development; Teacher Behaviours and Student Outcomes; Teacher Education; Teacher Leadership; Teacher Stress and Burnout: Methodological Perspectives; Teaching as a Profession: United States.

Bibliography

Anderson, J.R., 1982. Acquisition of cognitive skill. Psychological Review 89, 369–406.

Anderson, G.L., Herr, K., 2011. Scaling up "evidence-based" practices for teachers is a profitable but discredited paradigm. Educational Researcher 40 (6), 287–289.

Ball, D.L., Thames, M.H., Phelps, G., 2008. Content knowledge for teaching: what makes it special? Journal of Teacher Education 59 (5), 389–407.

Bausmith, J.M., Barry, C., 2011. Revisiting professional learning communities to increase college readiness: the importance of pedagogical content knowledge. Educational Researcher 40, 175–178.

Berliner, D.C., 1994. Expertise: the wonder of exemplary performances. In: Mangieri, J.N., Block, C.C. (Eds.), Creating Powerful Thinking in Teachers and Students Fort Worth. Rinehart & Winston, TX, Holt, pp. 161–186.

Berliner, D.C., 2005. The near impossibility of testing for teacher quality. Journal of Teacher Education 56 (3), 205–213.

Cianciolo, A.T., Matthew, C., Sternberg, R.J., Wagner, R.K., 2006. Tacit knowledge, practical intelligence and expertise. In: Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R. (Eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press, Cambridge, pp. 613–632.

Elliott, J.G., 2009. The nature of teacher authority and teacher expertise. Support for Learning 24 (4), 197–203.

Elliott, J., Tudge, J., 2007. The impact of the West on post-Soviet Russian education: change and resistance to change. Comparative Education 43, 93–112.

Elliott, J.G., Place, M., 2012. Children in Difficulty, 3rd Edition. Routledge, London.

Elliott, J.G., Stemler, S.E., Grigorenko, E.L., Sternberg, R.J., Hoffman, N., 2011. The socially skilled teacher and the development of tacit knowledge. British Educational Research Journal 37 (1), 83–103.

Eraut, M., 2000. Non-formal learning and tacit knowledge in professional work. British Journal of Educational Psychology 70, 113–136.

Ericsson, K.A., 1996. The acquisition of expert performance: an introduction to some of the issues. In: Ericsson, K.A. (Ed.), The Road to Excellence: The Acquisition of Expert Performance in the Arts, Sciences, Sports, and Games. Laurence Erlbaum Associates Inc, Hillsdale, NJ, pp. 1–50.

Ericsson, K.A., 2006. An introduction to the Cambridge handbook of expertise and expert performance: Its development, organization, and content. In: Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R. (Eds.), The Cambridge Handbook of Expertise and Expert Performance. Cambridge University Press, Cambridge, pp. 3–19.

Feldman, A., Weiss, T., 2010. Understanding change in teachers' ways of being through collaborative action research: a cultural historical activity theory analysis. Educational Action Research 18 (1), 29–55.

Glaser, R., Chi, M.T.H., 1988. Overview. In: Chi, M.T.H., Glaser, R., Farr, M.J. (Eds.), The Nature of Expertise. Laurence Erlbaum Associates Inc, Hillsdale, NJ, pp. 15–27.

Grigorenko, E.L., Sternberg, R.J., Strauss, S., 2006. Practical intelligence and elementary-school teacher effectiveness in the United States and Israel: measuring the predictive power of tacit knowledge. Thinking Skills and Creativity 1, 14–33.

Hashweh, M.Z., 2005. Teacher pedagogical constructions: a reconfiguration of pedagogical content knowledge. Teachers and Teaching: Theory and Practice 11, 273–292.

Hattie, J., 2011. Visible Learning for Teachers: Maximizing Impact on Learning. Routledge, New York.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E., Coe, R., 2013. The Sutton Trust-education Endowment Foundation Teaching and Learning Toolkit. Education Endowment Foundation, London. http://educationendowmentfoundation.org.uk/toolkit/.

Johnson, S.M., Birkeland, S.E., 2003. Pursuing a sense of success: new teachers explain their career decisions. American Educational Research Journal 40 (3), 581–617.

Kounin, J.S., 1970. Discipline and Group Management in Classrooms. Holt, Rinehart & Winston, New York.

MaClure, M., Walker, B.M., 2000. Disenchanted evenings: the social organization of talk in parent–teacher consultations in UK secondary schools. British Journal of Sociology of Education 21 (1), 5–25.

Malm, B., Löfgrer, H., 2006. Teacher competence and students' handling strategies. Research in Education 76, 62–73.

Pace, J.L., Hemmings, A., 2007. Understanding authority in classrooms: a review of theory, ideology and research. Review of Educational Research 77 (1), 4–27.

Palmer, D.J., Stough, L.M., Burdenski, T.K., Gonzales, M., 2005. Identifying teacher expertise: an examination of researchers' decision-making. Educational Psychologist 40 (1), 13–25.

Polanyi, M., 1958. Personal Knowledge: Towards a Post-critical Philosophy. University of Chicago Press, Chicago, IL.

Schön, D.A., 1983. The Reflective Practitioner: How Professionals Think in Action. Basic Books, New York.

Shulman, L.S., 1986. Those who understand: knowledge growth in teaching. Educational Researcher 15 (2), 4–14.

Shulman, L.S., 1987. Knowledge and teaching: foundations of the new reform. Harvard Educational Review 57 (1), 1–22.

Stemler, S.E., Elliott, J.G., Grigorenko, E.L., Sternberg, R.J., 2006. There's more to teaching than instruction: seven strategies for dealing with the practical side of teaching. Educational Studies 32 (1), 101–118.

Sternberg, R.J., 1997. Successful Intelligence: How Practical and Creative Intelligence Determines Success in Life. Plume, New York.

Sternberg, R.J., Forsythe, G.B., Hedlund, J., Horvath, J., Snook, S., Williams, W.M., Wagner, R.K., Grigorenko, E.L., 2000. Practical Intelligence in Everyday Life. Cambridge University Press, New York.

Sternberg, R.J., Hedlund, J., 2002. Practical intelligence, *g*, and work psychology. Human Performance *15* (1/2), 143–160.

Sternberg, R.J., Horvath, J., 1995. A prototype view of expert teaching. Educational Researcher 24 (6), 9–17.

Sztajn, P., Confrey, J., Wilson, P.H., Edgington, C., 2012. Learning trajectory based instruction: toward a theory of teaching. Educational Researcher 41 (5), 147–156.

Torff, B., Sessions, D.N., 2005. Principals' perceptions of the causes of teacher ineffectiveness. Journal of Educational Psychology 97, 530–537.

Van Driel, J.H., Berry, A., 2010. Pedagogical content knowledge. In: Peterson, P., Baker, E., McGaw, B. (Eds.), International Encyclopedia of Education, third ed. Elsevier, New York, pp. 656–661.

Van Driel, J.H., Berry, A., 2012. Teacher professional development focusing on pedagogical content knowledge. Educational Researcher 41, 26–28.

Van Tartwijk, J., 1993. Sketches of Teacher Behavior: The Interpersonal Meaning of Nonverbal Behavior in the Classroom. W.C.C., Utrecht, The Netherlands [in Dutch].

Teacher Leadership

Alma Harris, Institute of Education, University College, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

No modern concept has been more powerfully received by those concerned with school improvement and system reform than leadership. This article highlights the contemporary literature showing the importance of teacher leadership in generating and sustaining school and system improvement. Effective leadership, in the guise of the school principal, has long been identified with school success. The sheer volume of literature concerning school leadership is testament to its importance and ensures that it remains center stage in contemporary discussions about organizational change and development.

Current theorizing reinforces the view that leadership is not the preserve of the individual but is a social influence process whereby intentional influence is exerted by one person (or group) over other people or groups to structure the activities and relationships in a group or organization (Yukl, 1990). High-performing organizations are characterized by the fact that leadership is not identified with the qualities of an individual but rather is understood as behavior that facilitates collective action toward a common goal Hargreaves, Boyle and Harris, (2014). Such collective action resonates closely with the theory of 'distributed leadership' which draws particularly upon distributed cognition and activity theory (Harris, 2008, 2009, 2014). Both theories underscore how social context influences human interaction and learning and reinforce that leadership is a form of social influence that goes beyond individuals and implies collective agency (Ogawa and Bossert, 1995).

The seminal work by Spillane et al. (2001) suggests that distributed leadership is a way of understanding leadership as a practice that is shared and realized within extended groupings and networks (Spillane et al., 2001; Harris, 2004; Gronn, 2000; Hulpia et al., 2011). In this sense, leadership is best understood as *practice distributed over leaders, followers, and their situation* (Spillane et al., 2001, p. 13). This distributed view of leadership *incorporates the activities of multiple groups of individuals in a school guiding and mobilizing staff in the process of instructional change* (Spillane et al., 2001, p. 13). Teacher agency, instruction, and collaboration are implicit within a model of distributed leadership. Although the distributed leadership literature has not explicitly acknowledged the idea of 'teacher leadership,' the same theoretical and conceptual terrain is clearly shared (Harris and Muijs, 2004).

Teacher Leadership

There is a substantial body of literature that focuses exclusively on the nature, form, and impact of teacher leadership (see York-Barr and Duke, 2004; Harris and Muijs, 2004). From this evidence base it is immediately evident that there are overlapping and competing definitions of the concept. York-Barr and Duke (2004, p. 260) note "in writing about teacher leadership, many authors readily assert its importance but usually fail to define it." Predictably, the lack of a consistent and precise definition in the literature has resulted in teacher leadership being loosely associated with a wide range of activities, roles, and behaviors. It has also meant that teacher leadership has become a convenient 'catch-all phrase' for diverse forms of teacher activity, engagement, and collaboration (York-Barr and Duke, 2004, p. 260).

To understand the variation of interpretation, it is important to look at the evolution of teacher leadership. The first phase of teacher leadership equates with recognition of the importance of the formal leadership roles that teachers undertake (e.g., department heads, pastoral leaders). In the second phase, teacher leadership became associated with the instructional expertise of teachers and their potential to develop the professional expertise of others. In the third phase, teacher leaders are viewed as central to the process of generating organizational development and change through their ability to influence others and their contribution to capacity building.

Katzenmeyer and Moller (2001, p. 17) define teacher leaders as teachers, who are leaders within and beyond the classroom, who identify with and contribute to a community of teacher learners and leaders, and influence others toward improved educational practice. Boles and Troen (1994, p. 11) characterize teacher leadership as a form of collective leadership in which teachers develop expertise by working collaboratively. Harris and Muijs (2004) suggest that there are four discernible dimensions of the teacher leadership role. The first dimension concerns the way in which teachers work with and across school boundaries and structures to establish social linkages within the community. This *brokering* role remains a central responsibility for the teacher as leader as it ensures that links within and across schools are in place and that opportunities for meaningful development among teachers are maximized. A second dimension of the teacher leader role focuses upon *participative* leadership where teachers work collegially with other teachers to encourage the examination of instructional practices.

A third dimension of teacher leadership is the *mediating role*. This suggests that teacher leaders are important sources of instructional expertise and information because they demonstrate high levels of instructional expertise. The final and possibly the most important dimension of the teacher leadership role, is that of *instructional leader*, i.e., forging close relationships with individual teachers so that reciprocal learning takes place.

In summary, there are a number of important things to highlight about teacher leadership from the literature. First,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92135-4

teacher leadership is associated with the creation of *collegial norms* among teachers that can contribute directly to school effectiveness, improvement, and development. Second, teacher leadership equates with giving teachers *opportunities to lead* which can have a positive influence upon individual self-efficacy and motivation. Third, at its most practical level, teacher leadership means that teachers are *working as instructional leaders* influencing curriculum, teaching, and the professional practice of others.

The next section considers the evidence about the impact of teacher leadership upon organizational and individual outcomes. It looks at the impact of teacher leadership at the school, teacher, and student level.

The Impact of Teacher Leadership: The School Level

The findings about the impact of teacher leadership at the school level are rather variable. Some studies indicate strong school level effects (Griffin, 1995) while others indicate the negligible influence of teacher leadership upon the school level. The greatest support for a relationship between teacher leadership and organizational change can be located within the school effectiveness, improvement, and reform literature. Here evidence suggests that generating teacher leadership, with its combination of increased teacher collaboration and increased responsibility, has positive effects on transforming schools and systems (Fullan, 2011; Harris, 2011; Harris and Jones, 2011).

Research has also reinforced that an organization's ability to improve and sustain improvement largely depends upon its ability to foster and nurture professional learning communities or 'communities of practice' (Morrissey, 2000; Holden, 2002). Linked to the idea of communities of practice, school improvement advocates have suggested that schools should operate as a professional learning community. Hargreaves (2000, p. 3) suggests "professional learning communities lead to strong and measurable improvements in students' learning. Instead of bringing about 'quick fixes' or superficial change, they create and support sustainable improvements that last over time because they build professional skill and the capacity to keep the school progressing." This idea of professional learning communities implies a commitment to teachers sharing learning and working collaboratively. It also embraces the notion of teacher leadership as it is assumed that teachers will be the main catalysts for change and development within a professional learning community (Louis and Stoll, 2007; Harris and Jones, 2011).

In summary, the literature confirms the positive effects of focused teacher collaboration upon organizational development and change. Where teacher leadership manifests itself in the development of trusting and collaborative relationships with colleagues, there is evidence that it does positively influence school culture and can generate the conditions where there are positive instructional and organizational outcomes.

The Impact of Teacher Leadership: The Teacher Level

By far the most discernible and powerful effect of teacher leadership is on the teacher leaders themselves. Research findings suggest that empowering teachers to lead enhances self-esteem and work satisfaction, which in turn leads to higher levels of performance and potentially much better outcomes for learners. Lieberman et al. (2000) reported that the teachers felt the experience of leading had improved their confidence in their own abilities, and had taught them to motivate, lead, and encourage other adults. Similarly, O'Connor and Boles (1992) reported improved self-confidence, increased knowledge, and an improved attitude to teaching among teachers who had been designated as leaders.

As highlighted earlier, the nature and quality of leadership within schools has been found to be an important condition for maximizing school effectiveness and improvement. Yet the relationship between teacher leadership and teacher effectiveness is generally implied in the literature rather than proven. The evidence shows that in schools where achievement is high, senior leaders tend to be more confident in allowing teachers to take on leadership roles. As one US study showed, improving student outcomes appeared as a condition for, rather than a result of, teacher leadership (Dickerson, 1992).

Teacher leadership has also reported to have effects on teacher practices at the classroom level. Smylie's (1994) review of redesignated teacher work and its effects on classroom practice drew two primary conclusions. First, that change in classroom practices was more likely to occur among those teachers whose work was redesigned, i.e., the teacher leaders and second, changes in classroom practice were more likely to occur when initiatives were collective, as opposed to individual.

In summary, the findings from the available studies suggest that teacher leadership is positively related to changes in teachers' classroom practice and their instructional effectiveness.

The Impact of Teacher Leadership: The Student Level

York-Barr and Duke (2004) found only five studies that directly examined the impact of teacher leadership on students. The first conducted by Ryan (1999) studied three schools where teacher leaders were found to be having a positive effect on students because they influenced the instructional practices of colleagues and were involved in school level decision-making. The second by Louis and Marks (1998) did not discern a direct relationship between teacher empowerment and student learning but strongly supported the argument that empowerment did positively influence teachers' efforts to improve instruction. The third by Taylor and Bogotch (1994) found no significant difference in terms of student attendance, achievement, or behavior between schools with high degrees of teacher participation in decision-making as opposed to schools with low degrees of teacher participation in decision-making.

The other two studies conducted by Leithwood and Jantzi (1999, 2000) explored the effects of school and teacher leadership on students' engagement with school. This study reported no statistically significant relationship between teacher leadership and student engagement. However it did conclude that teacher leadership outweighs principal leadership effects before taking into account the moderating effects of family educational culture.

In contrast, other studies have shown a more positive relationship between teacher leadership and student learning outcomes. Research by Silns and Mulford (2002) concluded

that student outcomes are more likely to improve where leadership sources are distributed throughout the school community and where teachers are empowered to lead. Similarly, a study of 86 US middle schools found that both teacher professionalism and collegial leadership were positively related to improved student outcomes (Hoy et al., 1998). Other work by Eddy- Spicer (2011) and Lemlech and Hertzog (1998) has suggested that encouraging teachers to take on leadership roles positively affects self-efficacy and behavior which subsequently influences student learning outcomes. Harris and Muijs (2004) found an indirect relationship between teacher involvement in decision-making and improved student outcomes.

In summary, the direct link between teacher leadership and improved learner outcomes is far from clear or established. The literature confirms that where instructional improvement is at the heart of the work of teacher leaders, there is a much greater chance of a positive impact upon student learning outcomes.

Conclusion

Within the literature there is substantial evidence to suggest that teacher leadership can have a positive impact upon teachers' self-efficacy, instructional knowledge, and professional relationships. The evidence also supports the positive benefits of teacher leadership upon the practice of other teachers and endorses a strong relationship between teacher leadership, in the form of greater involvement in decision-making, with positive organizational development. For those working toward school and system reform the implication is clear; if properly constructed and appropriately deployed, teacher leadership can be a powerful way to build the collective capacity so that school and system transformation is secured and sustained.

See also: Authentic Leadership; Leadership Education: Theory and Practice; Leadership in Organizations, Sociology of; Leadership; Teacher Education; Teacher Expertise.

Bibliography

Boles, K., Troen, V., 1994. Teacher Leadership in a Professional Development School. Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Day, C., Sammons, P., Leithwood, K., Harris, A., Hopkins, D., 2009. The Impact of Leadership on Pupil Outcomes. DCSF, London. Final Report.

Dickerson, C., 1992. The Lead Teacher Model: A Restructured School. Paper Presented at the Annual Meeting of the North Carolina Association of Research in Education, Chapel Hill, NC.

Edmonds, R., 1979. Effective schools for the urban poor. Educational Leadership 37, 15–24.

Eddy- Spicer, D.H., 2011. From artefact to tool: the development of teachers' collective agency in the enactment of reform. Pedagogies: An International Journal 6 (4), 359–380.

Fullan, M., 2011. The Moral Imperative Realized. Corwin Press, Thousand Oaks CA. Ontario Principals Council, Toronto.

Griffin, G.A., 1995. Influences of shared decision making on school and classroom activity: conversations with five teachers. The Elementary School Journal 96 (1), 29–45.

Gronn, P., 2000. Distributed properties: a new architecture for leadership. Educational Management and Administration 28 (3), 317–338.

Hallinger, P., Murphy, J., 1985. Assessing the instructional leadership behavior of principals. Elementary School Journal 86 (2), 217–248.

Harris, A., 2004. Distributed leadership: leading or misleading. Educational Management and Administration 32 (1).

Harris, A., Muijs, D., 2004. Improving Schools through Teacher Leadership. Open University Press, London.

Harris, A., Jones, M., 2011. Professional Learning Communities in Action. Leannta Press, London.

Harris, A., 2011. System improvement through collective capacity building. Journal of Educational Administration 49 (6), 624–636.

Harris, A., Hargreaves, A., 2013. Schools Performing beyond Expectations. Routledge, London.

Harris, A., 2008. Distributed Leadership: Developing Tomorrow's Leaders. Routledge, London.

Harris, A., 2009. Distributed Leadership: Different Perspectives. Springer Press, Netherlands.

Harris, A., 2014. Distributed Leadership Matters. Corwin Pr, California.

Hargreaves, A., 2000. Four Ages of Professionalism and Professional Learning. Teachers and Teaching 6 (2), 151–182.

Hoy, W.K., Hannum, J., Tschannen-Moran, M., 1998. Organizational climate and student achievement: a parsimonious and longitudinal view. Journal of School Leadership 8 (4), 336–359.

Hargreaves, A., Boyle, A., Harris, A., 2014. Uplifting Leadership. Wiley Business Press.

Holden, G., 2002. Towards a Learning Community: The Role of Teacher-led Development in School Improvement. Paper Presented at the CELSI British Council Leadership in Learning Conference, London.

Hulpia, H., Devos, G., Van Keer, H., 2011. The relation between school leadership from a distributed perspective and teacher's organizational commitment examining the source of the leadership function. Educational Administration Quarterly 47, 728.

Katzenmeyer, M., Moller, G., 2001. Awakening the Sleeping Giant. Helping Teachers Develop as Leaders, second ed. Corwin Press, Thousand Oaks, CA.

Leithwood, K., Jantzi, D., 1999. The relative effects of principal and teacher sources of leadership on student engagement with school. Educational Administration Quarterly 35 (Suppl), 679–706.

Leithwood, K., Jantzi, D., 2000. Distributed Leadership and Student Engagement in School. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Leithwood, K., Lieberman, A., 1988. Teachers and principals: turf, tension and new tasks. Phi Delta Kappan 69, 648–653.

Lieberman, A., Saxl, E.R., Miles, M.B., 2000. Teacher leadership: ideology and practice. In: The Jossey-Bass Reader on Educational Leadership. Jossey-Bass, Chicago, pp. 339–345.

Lemlech, J.K., Hertzog, H., April 1998. Preparing Teachers for Leadership Roles. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Little, J.W., 1990. The persistence of privacy: autonomy and initiative in teachers' professional relations. Teachers College Record 91, 50–536.

Louis, K.S., Marks, H., 1998. Does professional community affect the classroom? American Journal of Education 106 (4), 532–575.

Morrissey, M., 2000. Professional Learning Communities: An Ongoing Exploration. Unpublished Paper Southwest Educational Development Laboratory, Austin, TX.

Muijs, D., Harris, A., 2003. Teacher leadership: improvement through empowerment. Educational Management and Administration 31 (4), 437–449.

Murphy, J., 2002. Reculturing the profession of educational leadership: new blueprints. Educational Administration Quarterly 38 (2), 176–192.

O'Connor, K., Boles, K., 1992. Assessing the Needs of Teacher Leaders in Massachussetts. Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Ogawa, R.T., Bossert, S.T., 1995. Leadership as an organizational quality. Educational Administration Quarterly 31 (2), 224–243.

Ovando, M., 1996. Teacher leadership: opportunities and challenges. Planning and Changing 27 (1/2), 30–44.

Purkey, S., Smith, M., 1983. Effective schools: a review. Elementary School Journal 83, 427–452.

Ryan, S., April 1999. Principals and Teachers Leading Together. Paper Presented at the Annual Meeting of the American Educational Research Association. Montreal, Quebec, Canada.

Rosenholz, S., 1989. Teachers' Workplace: The Social Organisation of Schools. Teachers College Press, New York.

Stoll, L., Seashore Louis, K. (Eds.), 2007. Professional Learning Communities: Divergence, Depth and Dilemmas. Open University Press, Maidenhead, UK.

Smylie, M.A., 1994. Redesigning teachers' work: connections to the classroom. In: Darling-Hammond, L. (Ed.), Review of Research in Education, vol. 20. American Educational Research Association, Washington, DC, pp. 129–177.

Silns, H., Mulford, B., 2002. Leadership and School Results, Second International Handbook of Educational Leadership and Administration. Kluwer Press.

Spillane, J., Halverson, R., Diamond, J., 2001. Towards a Theory of Leadership Practice: A Distributed Perspective. Northwestern University, Institute for Policy Research Working Article.

Taylor, D.L., Bogotch, I.E., 1994. School-level effects of teachers' participation in decision making. Educational Evaluation and Policy Analysis 16 (3), 302–319.

York-Barr, J., Duke, K., 2004. What do we know about teacher leadership? Findings from Two Decades of Scholarship in Review of Educational Research 74 (3), 255–316.

Yukl, G., 1990. An evaluative essay on current conceptions of effective leadership European. Journal of Work and Organisational Psychology 8 (1), 33–48.

Teacher Motivation

Helen MG Watt and Paul W Richardson, Faculty of Education, Monash University, Melbourne, Victoria, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Motivational theories were developed in relation to students; over the last decade, researchers have turned their attention also to teachers. We overview each of three major motivation theories (expectancy-value, achievement goal, and self-determination) that have thus far been reinterpreted in relation to teachers, involving the adaptation of constructs and processes initially designed to understand students' motivations. Empirical findings pertaining to teachers' motivations are summarized, cultural differences contrasted, and methodological issues raised. We conclude with implications and suggestions for future research in the field.

Teacher Motivation: Theoretical and Empirical Developments

The last decade has seen something of a 'Zeitgeist' of research in the burgeoning field of teacher motivation. Theoretical developments have been concentrated to address pressing questions concerning teachers' career entry and commitment, professional effort exertion and instructional behaviors, and professional growth and personal well-being. Although the study of teacher motivation is not in itself new, until recently researchers had not drawn upon the motivation literature in an intensive or systematic way to develop theoretically grounded and psychometrically strong models. In the last decade, theories, constructs, and concepts from the well-established student motivation literature have been extrapolated to the study of teacher motivation. Transposing theoretical concepts to the previously neglected domain of teaching has required the adaptation of theories that were not originally developed to apply to teachers.

Motivation researchers have asked, first, what kinds of motivations are relevant for teachers; second, whether and how we can measure them; and third, whether and how they matter. Empirical findings have begun to demonstrate the importance of teachers' motivations for both teachers and their students. In this endeavor, the choice of theoretical lens has depended upon the outcomes under investigation. We began our program of research drawing upon expectancy-value theory (EVT) because we were interested in why individuals choose to pursue teaching as a career and consequences for their professional engagement, teaching style, and personal well-being (Watt and Richardson, 2007a, 2008). In other lines of research, achievement goal theory (AGT) has focused on how teachers strive to feel successful and how success is conceived in their daily work because the classroom is an achievement arena for teachers as well as students (Butler, 2007). Self-determination theory (SDT), which is more generally concerned with human growth and functioning, has explored teachers' controlled versus autonomous motivations, consequences for teachers' quality instructional behaviors versus burnout, and consequences for the quality of their students' motivation (see Fernet et al., 2012; Pelletier et al., 2002; Roth, 2014).

With the exception of self-efficacy theory, motivational theories have only recently begun to be adapted and applied to the study of teacher motivation. Teacher self-efficacy refers to the degree to which teachers feel efficacious to enact certain professional outcomes such as deploying effective instructional strategies, classroom management, and engaging students (Tschannen-Moran and Woolfolk Hoy, 2001). Studies of teachers' self-efficacy have shown that higher self-efficacy predicts many positive attitudes and behaviors for teachers and students, including innovative teaching strategies, task persistence, resilience and well-being, and student motivation and achievement (see Klassen et al., 2011; Tschannen-Moran et al., 1998).

In this article, we begin with an overview of each of the three major motivation theories influencing research on teachers, involving the adaptation of constructs and processes initially designed to understand students' motivations. We next review empirical findings pertaining to teachers' motivations, paying particular attention to EVT within which our work has concentrated. We conclude with implications and suggestions for future research in the field.

Theories of Teacher Motivation

Expectancy-Value Theory

The Eccles et al. model of EVT (see Eccles, 2009) proposes that one's educational, vocational, and other achievement-related choices are shaped, on the one hand, by one's ability beliefs and expectancies for success, and on the other, by the value attached to the task. The values component is comprised of how much a person enjoys the task (intrinsic value), whether it is seen to be useful (utility value), whether it is seen as being important for achieving a person's goals (attainment value), and how costly it will be to succeed in relation to what an individual must forego in undertaking a task (opportunity cost), entailing negative effects such as financial loss, psychological experiences (e.g., anxiety, stress), time commitment, or effort expenditure (see Perez et al., 2014).

Originally developed to examine gendered patterns of enrollment in senior high mathematics (Eccles (Parsons) et al., 1983), the model has been applied to the selection of various different types of careers (e.g., Watt et al., 2012b). Drawing on this work, the Factors Influencing Teaching Choice (FIT-Choice;

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26082-0

www.fitchoice.org) framework was advanced as a coherent theoretical model to propel systematic enquiry into key motivations for choosing teaching as a career (**Figure 1**; see Richardson and Watt, 2006, 2010; Watt and Richardson, 2007a, 2008).

What Motivates People to Choose a Teaching Career?

Researchers from widely different disciplinary approaches and using diverse methodologies have explored what motivates people to choose teaching as a career and, more specifically, why they sign up to be early childhood, elementary, or secondary school teachers of particular subjects. Given that teachers constitute a large, heterogeneous workforce, substantially funded from the public purse, it is not surprising that policymakers have turned their attention to what attracts people to teaching as a career, especially when faced with recruitment problems, shortages of specialist teachers, early attrition from the career, and more recently, insistent pressure to improve the quality of the teaching workforce.

Motivations for choosing teaching as a career identified in the 1960s and 1970s, particularly in large-scale American studies, focused on the desire for social mobility, the influence of parents and extended family, a desire to work with young children and adolescents, to work in a people-oriented profession, and job-related benefits such as security, pensions, and holidays (National Education Association, 1976). Even though there has been a steady stream of studies over the last five decades examining teaching choice and teachers' work and lives (e.g., Connell, 1985; Day et al., 2007; Huberman, 1993; Lortie, 2002), the approaches have been diverse, making comparisons of findings across studies difficult.

Until the last decade, the US accounted for the largest number of studies of teachers' career choice motivations. These mostly relied on questionnaires designed by researchers, including factors that frequently bore little relationship to one another, as well as numerous qualitative studies, precluding comparisons across samples or settings. Moreover, there was little agreement among researchers concerning underpinning theoretical frameworks or measures and a reliance on single-timepoint studies of participants who had self-selected into teaching. In reviewing this corpus of research up until the early 1990s, Brookhart and Freeman in their seminal analysis summarized, "altruistic, service-oriented goals and other intrinsic motivations are the source of the primary reasons entering teacher candidates report for why they chose teaching as a career" (1992: 46).

Since then, an OECD review of studies (2005), based on studies from France, Australia, Belgium (French Community), Canada (Québec), the Netherlands, the Slovak Republic, and the UK, reported that the most common motivations for teaching are the desire to work with youth, potential for intellectual fulfillment, and making a social contribution. The desire to work with children and adolescents has been identified as central in many studies (e.g., Fox, 1961; Joseph and Green, 1986; Kyriacou and Coulthard, 2000; Lortie, 2002; Tudhope, 1944), whereas extrinsic motives such as salary, job security, and career status have been more important in different sociocultural contexts such as Brunei (Yong, 1995), Zimbabwe (Chivore, 1988), Cameroon (Abangma, 1981), the Caribbean (Brown, 1992), Jamaica (Bastick, 1999), and Turkey (Kılınç et al., 2012). The absence of an agreed theoretical and analytical framework, however, has meant that researchers vary

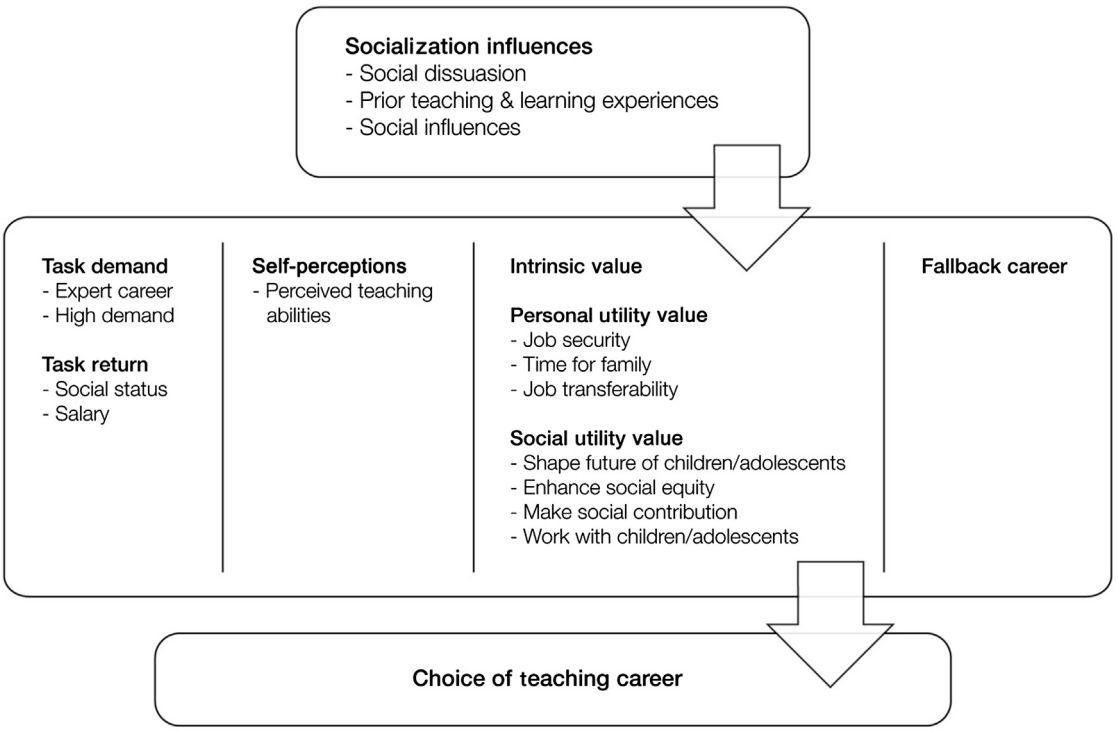

Figure 1 FIT-Choice theoretical model.

on what constitutes intrinsic, altruistic, extrinsic, or other motivations. For example, the desire to work with children has at times been referred to as an intrinsic (e.g., Young, 1995) and at other times as an altruistic motivation (e.g., Yong, 1995).

EVT provided a fruitful and coherent framework to guide our research concerning the motivation to choose a teaching career. We mapped motivations previously identified as influential in the teacher education literature to constructs in the expectancy-value framework, which allowed us to locate previously identified motivations within an integrative and comprehensive model, which additionally suggested other important motivations. Our FIT-Choice program of research began at its outset with the development of the FIT-Choice scale.

The FIT-Choice scale is psychometrically sound and yields reliable findings across settings (Watt and Richardson, 2007a, 2012; Watt et al., 2012a). It taps the altruistic-type motivations, often emphasized in the Western teacher education literature, such as to shape the future of children/adolescents, enhance social equity, make a social contribution, and work with children/adolescents (e.g., Book and Freeman, 1986; Brown, 1992; Lortie, 2002; Moran et al., 2001; Serow and Forrest, 1994), and ability-related beliefs that predominate in the career choice literature (Lent et al., 1993). The scale also elicits individuals' personally utilitarian motivations that tap quality of life issues such as job security, time for family, and job transferability; intrinsic motivations; social influences stemming from friends, family, or work colleagues who promote the idea that one is suited to be a teacher; positive prior teaching and learning experiences; and the negative motivation of having chosen teaching as a 'fallback' career (see Book et al., 1985; Haubrich, 1960; Robertson et al., 1983). The FIT-Choice framework also assesses perceptions of the demands (heavy workload, emotional demands, hard work), the rewards of teaching as a career (salary and social status), and career choice satisfaction. An imbalance between demands and rewards is conceptualized as a cost.

The FIT-Choice measurement platform allows for the identification of which motivations and task beliefs are more and less important for choosing teaching as a career. It also permits comparisons of the measured factors across settings, including Australia (Richardson and Watt, 2006), the United States (Lin et al., 2012; Watt et al., 2012a), Norway (Watt, et al., 2012a), Croatia (Jugović et al., 2012), China (Lin et al., 2012), Turkey (Kılınç et al., 2012), Germany (König and Rothland, 2012; Watt et al., 2012a), and Switzerland (Berger and D'Ascoli, 2012).

In general, teachers' motivations were more similar than they were different, although further studies in different sociocultural contexts are needed. Fallback career motivations were uniformly low, except in China and Turkey, likely reflecting the availability of career and work within those labor markets. Ability and intrinsic motivations were rated highly among all but the samples from China and Turkey, in which collectivist cultures career choices may be less based on individual interests and abilities, or more basic needs such as job security may have primacy in a newly industrialized country, such as Turkey, in which that sample indeed scored highest. Altruistic-type values (i.e., social utility values) were high in general, but notably lowest in the Chinese sample and in the middle for the German sample. The expectation of making a contribution to society may be taken more for granted in collectivistic Chinese culture, and the tracked school system in Germany could mean future teachers perceive lower agency to achieve social equity outcomes and youth opportunities through educational structures. Personal utility values were strikingly similar and rated moderately across samples, presumably reflecting basic needs in contemporary society, although, as mentioned, the Turkish sample rated job security somewhat higher.

In general, future teachers believed teaching is a highly demanding career (including heavy workload, emotional demand, and hard work), with low rewards in terms of salary and social status. The Chinese and especially the Turkish sample rated the demands of teaching substantially lower, possibly due to the collectivist approach to teacher development and group accountability in the Chinese sample, and the relative demandingness of alternative available work in the developing Turkish context. Perceptions of higher salary in the German setting reflected objective context differences. Values about teaching as a socially responsible and morally worthwhile career starkly contrast with fallback and personally utilitarian values, or beliefs that monetary rewards and status are important career outcomes.

EVT further posits that expectancies and values predict domain-specific achievement behaviors such as performance, efforts, and persistence. Yet, little is known about the long-term effects of initial teaching motivations. Can they have lasting effects on professional engagement and even on beginning teachers' subsequent teaching styles? Results from our longitudinal study highlight an enduring effect of initial motivations for choosing teaching, which, in turn, predict professional engagement and teaching styles up to 8 years later. We examined the extent to which initial motivations for teaching influenced professional engagement and career development aspirations (PECDA; Watt and Richardson, 2008) and self-reported teaching style (TSS; Watt and Richardson, 2007b), using longitudinal Australian FIT-Choice data spanning entry to teacher education (Time 1), exit from teacher education (Time 2), and after up 8 years of teaching experience (Time 3). Intrinsic and ability motivations to teach at Time 1 predicted Time 3 positive-reported teaching behaviors, as did social utility values through their influence on whether participants planned to remain in the profession at Time 2. Conversely, fallback career motivations predicted lowered PECDA and reduced positive teaching behaviors during early career. Social influences, such as being persuaded by family, friends, or others to become a teacher, predicted negative-reported teaching practices (see Richardson and Watt, 2014).

The most adaptive motivations for choosing teaching as a career were thus positive ability self-beliefs, intrinsic enjoyment of teaching, wanting to work with youth in order to help them shape their futures, and wanting to make a social contribution by enhancing social equity (social utility values). Problematic motivations were entering teaching as a fallback career and due to external social influences, both of which predicted later teaching negativity (including responding negatively or angrily to students' mistakes, use

of sarcasm, or deliberate embarrassment) (Richardson and Watt, 2014). We interpret the negative consequences from entering teaching due to external social influence in terms of SDT in that choosing a teaching career based on persuasion from others suggests a 'controlled' rather than an 'autonomous' motivation. These results have implications for teacher recruitment efforts; the negative effect of social influences needs to be carefully considered when persuading individuals into the teaching profession.

A comprehensive research strategy on teacher motivation needs to address several core issues: Why choose the career of being a teacher?; Why do people stay in the job, burnout, or leave?; How do motivations intersect with sociocultural factors to impact teachers' professional development and personal well-being?; and What types of profiles are evident in teachers' career trajectories? Such questions require following the same individuals over an extended period of time to address critical issues in the current climate of teacher shortages and concern regarding teacher quality with implications for policymakers, employers, and teacher educators. Longitudinal data allow the real and necessary opportunity to explore and test how processes unfold over time to produce outcomes. There are few high-quality longitudinal studies to address these questions.

Achievement Goal Theory

AGT originally distinguished a task (or mastery) from an ego (or performance) goal orientation (Dweck and Elliott, 1983; Nicholls, 1989). Since then, the trichotomous goal framework was proposed (Elliot and Church, 1997; Elliot and Harackiewicz, 1996; Nicholls, 1989), which distinguished *performance-approach* from *performance-avoidance* goals. Individuals who hold a performance-approach goal are motivated to demonstrate their abilities relative to others, in contrast to those who hold a performance-avoidance goal and are motivated to avoid demonstrating their relative lack of ability. A parallel distinction was subsequently proposed for mastery goals in the 2×2 achievement goal framework (see Bong, 2009; Elliot and McGregor, 2001; Harackiewicz et al., 2002), which introduced a mastery avoidance goal, defined as a focus to avoid misunderstanding, not learning, or not mastering a task. The two approach goal orientations predict positive outcomes, with mastery approach being the most positive. In contrast, performance-avoidance goals predict maladaptive outcomes, particularly when self-beliefs of competence are low (Law et al., 2012).

What Goals Are Teachers Motivated to Achieve?

Prompted by the notion that the classroom may function as an achievement arena for teachers as well as for students, Butler (2007) began to explore teachers' goals. Her program of research began with the development of the Goal Orientations for Teaching scale (2007), a self-report measure first used in a study with Israeli teachers and subsequently with samples of preservice and practising teachers in Germany (Dickhäuser et al., 2007; Retelsdorf et al., 2010). The scale identified four factors that reflect a mastery orientation (acquiring and developing professional competence), an ability-approach orientation (demonstrating superior teaching ability), an ability-avoidance orientation (avoiding displaying inferior teaching ability), and a work-avoidance orientation (to get through the day with as little effort as possible). Based on the logic that teaching is an inherently interpersonal activity, Butler (2012) subsequently identified a distinct new class of goal – relational goals – that describes teacher strivings to create caring relationships with their students.

Important links exist between teachers' achievement goals, patterns of communication and behavior in the classroom, and students' resultant learning and achievement outcomes (Butler, 2007, 2014; Butler and Shibaz, 2008). Similar to patterns of findings concerning students' achievement goals, teachers' mastery goals were associated with positive outcomes, including adaptive coping and engagement, mastery-oriented instruction, and their students' engagement. Performance-avoidance goals (to avoid the demonstration of poor teaching ability) predicted clear negative outcomes, including defensive avoidance of help, self-handicapping, burnout, career dissatisfaction, and surface approaches to instruction involving more competitive classroom climates. Patterns for performance-approach goals (to prove ability) were less clear-cut. Findings concerning the newly identified class of relational goals showed these teachers provided greater socioemotional support to students and cognitively stimulating instruction (see Butler, 2012, 2014). Similarly, Shim et al. (2013) found that teachers' achievement goals were associated with the classroom environments they created in the Midwestern United States: Mastery goals for teaching predicted the creation of classroom mastery environments and performance-approach goals predicted a classroom performance focus.

Using the Shim et al. (2013) data set, Cho and Shim (2013) examined the association of the work environment with teachers' goals for their teaching: Teachers who perceived a mastery-oriented school environment adopted mastery goals for teaching, whereas teachers working in school environments that they perceived as performance-oriented were more likely to adopt performance-approach goals for teaching. Interestingly, those teachers with the highest levels of self-efficacy endorsed both mastery and performance-approach goals and maintained high mastery goals for their teaching even when the school encouraged performance goals; teachers with lower self-efficacy conformed to the school goal environment. In another US study, highly performance-oriented school environments were associated with teachers' adoption of performance-approach goals for teaching, reduced motivations, and lowered sense of community (Ciani et al., 2008). These studies highlight the importance of school cultures and their impact on teachers' own goals for teaching.

Teachers' ideographic goals have been explored in order to determine whether additional goals should be assessed. Hagger and Malmberg (2011) identified 13 goals among a sample of preservice teachers, reflecting six self-related (self, teacher engagement, behavior management, teacher education course, collegial relationships, career), four task-related (teaching strategies, subject/curriculum knowledge, planning and organization, taking account of individual differences), and three student impact goals (student engagement, relationship with students, teacher impact). Goal contents were found to predict preservice teachers' concerns, and both self-esteem and depression.

In another study among preservice teachers, drawing upon Ford and Nichols' (1987) Taxonomy of Human Goals, Mansfield et al. (2012) identified within-person (affective and cognitive) and person–environment (social relationship and task) goals. It is noteworthy that the goals in these two studies encompassed a broad range of professional activities, aspirations, and conceptions of beginning teachers. Further research could explore ideographic goals among practising teachers, the role of teaching experience in goal contents and differentiation, and other correlates of teachers' goal contents for themselves and their students.

Self-Determination Theory

SDT (Deci and Ryan, 1985; Ryan and Deci, 2000) posits that fulfillment of three basic innate, human psychological needs (*autonomy*, *competence*, and *relatedness*) is necessary for optimal human functioning. When support for these needs is available, people will be intrinsically motivated to undertake tasks for their own sake because they are interesting, enjoyable, and inherently rewarding. Under these conditions, people are autonomously motivated. By contrast, when these psychological needs are unsupported or thwarted, wellness and growth will be negatively affected. In studies of student learning, autonomy relates to choice, competence refers to feeling efficacious during learning tasks, and relatedness refers to feeling connected to the teacher and class. The satisfaction of these needs is a necessary condition for learning.

SDT offers a potentially rich perspective on how individuals develop and internalize their motivations. *External regulation* is the least self-determined, least autonomous form of motivation, and located at the opposite end of the continuum from *intrinsic regulation*. Between the extremes of extrinsic and intrinsic motivation, differentiated levels of regulation occur; *introjected regulation* refers to forms of regulation that are taken in by the individual and provide for self-imposed pressures or 'should,' whereas *identified regulation* is experienced when individuals identify with the worth of an activity or behavior and engage in it accordingly (Ryan and Deci, 2000). Thus, intrinsic and extrinsic motivations are located at two ends of an internalization continuum with movement along that continuum being signaled by structural changes that progressively transform externally based regulations into a more internal personal regulatory system (cf Grolnick et al., 1997). Movement along this continuum means that an individual's experience of regulation is progressively more autonomous, integrated, and self-determined with initiated behaviors performed out of choice, simultaneously meeting an individual's innate need for a sense of competence, autonomy, and relatedness.

The key differentiation drawn by SDT theorists is between autonomous (or self-determined) and controlled motivations. Autonomous motivation involves volition and choice, and controlled motivation involves an external or internal sense of compulsion (Assor et al., 2004; Grolnick et al., 1997). It is possible that initially autonomous motivations could become controlled motivations, for example, when an initial decision (such as to become a teacher) is autonomous, but then actually doing the work entails a sense of compulsion or external responsibility. A large literature has examined predictors of students' autonomous motivation and benefits for their engagement and well-being (see Ryan and Deci, 2009).

What Are the Consequences of Teachers' Self-Determined Motivations?

The research based in SDT on teachers is quite scarce (see Roth, 2014). When teachers perform their job for the intrinsic value of working with and achieving positive outcomes for their students, they are said to be autonomously motivated. Yet, when teachers are externally regulated by invasive policies, curriculum requirements, principal supervision, and administrative practices, the theory posits that they experience low levels of choice in initiating behavior and a reduced sense of competence and autonomy (Fernet et al., 2012). In studies of teachers, autonomous motivations, such as to let children feel that teachers care about them or being in touch with children and adolescents (Roth et al., 2007), are associated with perceived accomplishment, teaching self-efficacy, autonomy-supportive teaching practices, and reduced burnout. Autonomous teacher motivation has additionally been found to be associated with autonomy-supportive teaching practices that furnish choice and relevance to students (Fernet et al., 2012; Pelletier et al., 2002; Roth et al., 2007; Taylor and Ntoumanis, 2007; Taylor et al., 2008).

Outlook and Future Directions

Over the last decade or so, there has been an intensification of interest in teacher motivation by researchers who have developed programs of research founded upon robust psychological theories of motivation to address questions such as those which structure our continuing FIT-Choice program: What motivates people to choose teaching as a career? How are teacher motivations constituted? Are teacher motivations stable or malleable, and to what degree? Do initial teacher motivations matter in the longer term and how? Why do many early-career teachers leave the profession? What are the factors that support teachers' confidence, motivation, and commitment to be an effective teacher over many years? What factors undermine teachers' motivation and commitment? Do teacher motivations relate to positive student learning outcomes? How are teacher motivations influenced by different contextual factors at the state, district, school, and classroom levels? How do teacher motivations differ across cultural and social contexts? Are there 'core' teacher motivations that do not vary substantially across social and cultural contexts?

Comparatively little attention has been given to teacher motivations at different points along a teacher's working life. A stage-based model of teachers' career development proposed by Huberman (1993) suggests that teacher motivations are not stable. While this model has heuristic appeal and has been reaffirmed in a cross-sectional study by Day et al. (2007), the question of whether teacher motivations are stable, change over time, respond to different contexts, or are malleable remains to be empirically tested using longitudinal designs. We are presently lacking a coherent developmental theoretical approach to the study of teachers' motivations throughout their career. In this pursuit, we may not wish

to keep measuring the same motivational factors over time; it may be that we need to invoke different theories as we seek to understand different developmental stages across the career life span. For example, EVT may be most relevant to the choice of teaching as a career, AGT for teachers' daily practices, and SDT to the promotion of self-determined teaching behaviors. It will be important to examine the motivations that sustain versus deter teachers in the profession, which may not be simply opposite sides of the same coin.

We expect that different kinds of school contexts will afford the realization or not of teachers' motivations, which, if left unfulfilled, could create a double-edged sword and lead to burnout and disappointment (see de Jesus and Lens, 2005). Teachers who are motivated to work with youth and enhance social equity, for example, find themselves frustrated and dispirited when their time is taken up by administrative and accountability work, which takes them away from what they believe are their core responsibilities. This creates a disjuncture between why teachers want to teach and the work they are required to spend their time doing. People who became teachers because they want to work with children and adolescents become less satisfied with their work if they have little time to engage in relational work on a daily basis. In this way, the same motivations can be a driving force for good or ill, dependent on the degree of match between a teacher and the teaching environment.

In determining which motivations are adaptive versus maladaptive, it is essential to consider what outcomes are being predicted and within what contexts. The decision of which outcomes ought to serve as outcome criteria in this endeavor is nontrivial before implications for policy and practice can be clarified. Stable and nonmalleable factors (for example, temperament) may be best considered at selection into teacher education, whereas changeable or malleable factors should be addressed during teacher education and early career induction. There is an urgent need for reliable, large-scale, long-term, cross-cultural data, incorporating extensive quantitative measures alongside targeted rich qualitative components, to examine the what and the why of teacher motivations and development over time and across settings.

Researchers are themselves experiencing something of a revolution as new methodological possibilities become available. There has been a reliance on self-report surveys in much of the motivational research literature, and it is increasingly recognized that multiple data sources and informants are required to understand the complexity of the antecedents, development, and outcomes of teacher motivations. Large-scale study designs that embrace and frame self-report qualitative interviews, and experience sampling methods that capture the dynamics of motivational systems, open up possibilities for illuminating processes at the individual and group level for different types of teachers. Structured observations (e.g., Pianta et al., 2008) can also provide more objective measurements concerning the expressed motivational elements of teachers' instruction and students' learning; student surveys provide complementary perspectives to, and outcomes from, teacher motivation. Teacher motivation research is increasingly drawing upon and informing other relevant domains of educational psychology, including self-efficacy, beliefs, future-time perspective, emotions, coping, responsibility, personality, and unconscious motivations (see Richardson et al., 2014), all of which link to the same outcomes with which teacher motivation researchers are ultimately concerned: teaching quality and student outcomes.

Acknowledgments

The FIT-Choice project (www.fitchoice.org) is supported by sequential Australian Research Council Discovery Projects DP140100402 (2014–16; Richardson and Watt), DP0987614 (2009–2012; Watt and Richardson) and DP0666253 (2006–09; Richardson, Watt, and Eccles).

See also: Avoidance and Approach Motivation: A Brief History; Expectancy-Value-Cost Model of Motivation; Interest, Psychology of; Mastery Learning; Motivation and Actions, Psychology of; Motivation, Learning, and Instruction; Passion and Motivation; Personal Projects; Self-Determination Theory; Self-Regulation in Adulthood; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Abangma, M.A., 1981. A Study of Primary Teachers' Attitudes towards Ruralisation of School Curriculum in English Speaking Cameroon. University of London, London.

Assor, A., Roth, G., Deci, E.L., 2004. The emotional costs of perceived parents' conditional regard: a self-determination theory analysis. Journal of Personality 72, 47–89. http://dx.doi.org/10.1111/j.0022-3506.2004.00256.x.

Bastick, T., 1999. A three factor model to resolve the controversies of why trainees are motivated to choose the teaching profession. In: Biennial Cross Campus Conference in Education, St Augustine, Trinidad.

Berger, J.-L., D'Ascoli, Y., 2012. Becoming a VET teacher as a second career: investigating the determinants of career choice and their relation to perceptions about prior occupation. Asia-Pacific Journal of Teacher Education 40, 317–341. http://dx.doi.org/10.1080/1359866X.2012.700046.

Book, C.L., Freeman, D.J., 1986. Differences in entry characteristics of elementary and secondary teacher candidates. Journal of Teacher Education 37 (2), 47–51.

Book, C., Freeman, D., Brousseau, B., (May/June) 1985. Comparing academic backgrounds and career aspirations of education and non-education majors. Journal of Teacher Education, 27–30.

Bong, M., 2009. Age-related differences in achievement goal differentiation. Journal of Educational Psychology 101, 879–896. http://dx.doi.org/10.1037/a0015945.

Brookhart, S.M., Freeman, D.J., 1992. Characteristics of entering teacher candidates. Review of Educational Research 62, 37–60.

Brown, M.M., 1992. Caribbean first-year teachers' reasons for choosing teaching as a career. Journal of Education for Teaching 18 (2), 185–195.

Butler, R., 2014. What teachers want to achieve and why it matters: an achievement goal approach to teacher motivation. In: Richardson, P.W., Karabenick, S.A., Watt, H.M.G. (Eds.), Teacher Motivation: Theory and Practice. Routledge, New York, pp. 20–35.

Butler, R., 2007. Teachers' achievement goal orientations and associations with teachers' help seeking: examination of a novel approach to teacher motivation. Journal of Educational Psychology 99, 241–252.

Butler, R., 2012. Striving to connect: extending an achievement goal approach to teacher motivation to include relational goals for teaching. Journal of Educational Psychology 104, 726–742. http://dx.doi.org/10.1037/a0028613.

Butler, R., Shibaz, L., 2008. Achievement goals for teaching as predictors of teacher support for and inhibition of question asking and help seeking and student help seeking and cheating. Learning and Instruction 18, 453–467. http://dx.doi.org/10.1016/j.learninstruc.2008.06.004.

Chivore, B.S.R., 1988. Factors determining the attractiveness of the teaching profession in Zimbabwe. International Review of Education 34 (1), 59–78. http://dx.doi.org/10.1007/BF00601918.

Cho, Y., Shim, S.S., 2013. Predicting teachers' achievement goals for teaching: the role of perceived school goal structure and teachers' sense of efficacy. Teaching and Teacher Education 32, 12–21. http://dx.doi.org/10.1016/j.tate.2012.12.003.

Ciani, K.D., Summers, J.J., Easter, M.A., 2008. A "top-down" analysis of high school teacher motivation. Contemporary Educational Psychology 33, 533–560. http://dx.doi.org/10.1016/j.cedpsych.2007.04.002.

Connell, R., 1985. Teachers' Work. Allen & Unwin, Sydney.

Day, C., Sammons, P., Stobart, G., Kington, A., Gu, Q., 2007. Teachers Matter: Connecting Lives, Work and Effectiveness. OUP, Berkshire.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-determination in Human Behavior. Plenum Press, New York.

Deci, E.L., Ryan, R.M., 2000. The 'what' and 'why' of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11, 227–268.

Dickhäuser, O., Butler, R., Tönjes, B., 2007. Das zeigt doch nur, dass ich's nicht kann – Zielorientierung und Einstellung gegenüber Hilfe bei Lehramtsanwärtern [That just shows I can't do it: goal orientation and attitudes concerning help among pre-service teachers]. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 39, 120–126.

Dweck, C.S., Elliott, E.S., 1983. Achievement motivation. In: Hetherington, E.M. (Ed.), Socialization, Personality, and Social Development. Wiley, New York, pp. 643–691.

Eccles, J., 2009. Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. Educational Psychologist 44 (2), 78–89. http://dx.doi.org/10.1080/00461520902832368.

Eccles (Parsons), J., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., Midgley, C., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivation. W.H. Freeman, San Francisco, CA, pp. 75–146.

Elliot, A.J., Church, M.A., 1997. A hierarchical model of approach and avoidance achievement motivation. Journal of Personality and Social Psychology 72, 218–232. http://dx.doi.org/10.1037/0022-3514.72.1.218.

Elliot, A.J., Harackiewicz, J.M., 1996. Approach and avoidance achievement goals and intrinsic motivation: a mediational analysis. Journal of Personality and Social Psychology 70, 461–475. http://dx.doi.org/10.1037/0022-3514.70.3.461.

Elliot, A.J., McGregor, H.A., 2001. A 2 × 2 achievement goal framework. Journal of Personality and Social Psychology 80, 501–519. http://dx.doi.org/10.1037/0022-3514.80.3.501.

Fernet, C., Guay, F., Senécal, C., Austin, S., 2012. Predicting intraindividual changes in teacher burnout: the role of perceived school environment and motivational factors. Teaching and Teacher Education 28, 514–525.

Ford, M.E., Nichols, C.W., 1987. A taxonomy of human goals and some possible applications. In: Ford, M.E., Ford, D.H. (Eds.), Humans as Self-constructing Living Systems: Putting the Framework to Work. Lawrence Erlbaum Associates Inc, Hillsdale, NJ, pp. 289–311.

Fox, R.B., 1961. Factors influencing the career choice of prospective teachers. Journal of Teacher Education 7 (4), 427–432. http://dx.doi.org/10.1177/002248716101200410.

Grolnick, W.S., Deci, E.L., Ryan, R.M., 1997. Internalization within the family: the self-determination theory perspective. In: Grusec, J.E., Kuczynski, L. (Eds.), Parenting and Children's Internalization of Values. Wiley, New York, pp. 135–161.

Hagger, H., Malmberg, L.-E., 2011. Pre-service teachers' goals and future-time extension, concerns, and well-being. Teaching and Teacher Education 27, 495–680. http://dx.doi.org/10.1016/j.tate.2010.10.014.

Harackiewicz, J.M., Barron, K.E., Pintrich, P.R., Elliot, A.J., Thrash, T.M., 2002. Revision of achievement goal theory: necessary and illuminating. Journal of Educational Psychology 94, 638–645. http://dx.doi.org/10.1037/0022-0663.94.3.638.

Haubrich, V.F., 1960. The motives of prospective teachers. Journal of Teacher Education 11 (3), 381–386.

Huberman, M., 1993. The Lives of Teachers. Teachers College Press, New York.

de Jesus, S.N., Lens, W., 2005. An integrated model for the study of teacher motivation. Applied Psychology – An International Review (Psychologie Appliquee – Revue Internationale) 54 (1), 119–134.

Joseph, P.B., Green, N., 1986. Perspectives on reasons for becoming teachers. Journal of Teacher Education 37 (6), 28–33. http://dx.doi.org/10.1177/002248718603700605.

Jugović, I., Marušić, I., Ivanec, T.P., Vidović, V.V., 2012. Motivation and personality of preservice teachers in Croatia. Asia-Pacific Journal of Teacher Education 40 (3), 271–287. http://dx.doi.org/10.1080/1359866X.2012.700044.

Kılınç, A., Watt, H.M.G., Richardson, P.W., 2012. Factors influencing teaching choice in Turkey. Asia-Pacific Journal of Teacher Education 40, 199–226. http://dx.doi.org/10.1080/1359866X.2012.700048.

Klassen, R., Tze, V., Betts, S., Gordon, K., 2011. Teacher efficacy research 1998–2009: signs of progress or unfulfilled promise? Educational Psychology Review 23 (1), 21–43. http://dx.doi.org/10.1007/s10648-010-9141-8.

König, Rothland, 2012. Motivations for choosing teaching as a career: effects on general pedagogical knowledge during initial teacher education. Asia-Pacific Journal of Teacher Education 40, 289–315. http://dx.doi.org/10.1080/1359866X.2012.700045.

Kyriacou, C., Coulthard, M., 2000. Undergraduates' views of teaching as a career choice. Journal of Education for Teaching 26 (2), 117–126. http://dx.doi.org/10.1080/02607470050127036.

Law, W., Elliot, A.J., Murayama, K., 2012. Perceived competence moderates the relation between performance-approach and performance-avoidance goals. Journal of Educational Psychology 104, 806–819. http://dx.doi.org/10.1037/a0027179.

Lent, R.W., Lopez, F.G., Bieschke, K.J., 1993. Predicting mathematics-related choice and success behaviors: test of an expanded social cognitive model. Journal of Vocational Behavior 42, 223–236.

Lin, E., Shi, Q., Wang, J., Zhang, S., Hui, L., 2012. Initial motivations for teaching: comparison between preservice teachers in the United States and China. Asia-Pacific Journal of Teacher Education 40 (3), 227–248. http://dx.doi.org/10.1080/1359866X.2012.700047.

Lortie, D.C., 2002. School-teacher: A Sociological Study. University of Chicago Press, Chicago.

Mansfield, C.F., Wosnitza, M., Beltman, S., 2012. Goals for teaching: towards a framework for examining motivation of graduating teachers. Australian Journal of Educational and Developmental Psychology 12, 21–34.

Moran, A., Kilpatrick, R., Abbott, L., Dallatt, J., McClune, B., 2001. Training to teach: motivating factors and implications for recruitment. Evaluation & Research in Education 15 (1), 17–32.

National Education Association, 1976. Status of the American Public School Teacher, 1975–1976. Washington, DC.

Nicholls, J.G., 1989. The Competitive Ethos and Democratic Education. Harvard University Press, Cambridge.

Organisation for Economic Co-operation and Development, 2005. Attracting, Developing and Retaining Effective Teachers. OECD, Paris.

Pelletier, L.G., Seguin-Levesque, C., Legault, L., 2002. Pressure from above and pressure from below as determinants of teachers' motivation and teaching behaviors. Journal of Educational Psychology 94, 186–196.

Perez, T., Cromley, J.G., Kaplan, A., 2014. The role of identity development, values, and costs in college STEM retention. Journal of Educational Psychology 106, 315–329. http://dx.doi.org/10.1037/a0034027.

Pianta, R.C., La Paro, K., Hamre, B.K., 2008. Classroom Assessment Scoring System (CLASS). Paul H. Brookes, Baltimore, MD.

Reeve, J., 2002. Self-determination theory applied to educational settings. In: Deci, E.L., Ryan, R.M. (Eds.), Handbook of Self-determination Research. University of Rochester Press, Rochester, NY, pp. 183–203.

Retelsdorf, J., Butler, R., Streblow, L., Schiefele, U., 2010. Teachers' goal orientations for teaching: associations with instructional practices, interest in teaching, and burnout. Learning and Instruction 20, 30–46.

Richardson, P.W., Watt, H.M.G., 2006. Who chooses teaching and why? Profiling characteristics and motivations across three Australian universities. Asia-Pacific Journal of Teacher Education 34 (1), 27–56. http://dx.doi.org/10.1080/13598660500480290.

Richardson, P.W., Watt, H.M.G., 2010. Current and future directions in teacher motivation research. In: Urdan, T.C., Karabenick, S.A. (Eds.), The Decade Ahead: Applications and Contexts of Motivation and Achievement, Advances in Motivation and Achievement, vol. 16B. Emerald, Bingley, UK, pp. 139–173.

Richardson, P.W., Watt, H.M.G., 2014. Why people choose teaching as a career: an expectancy-value approach to understanding teacher motivations. In: Richardson, P.W., Karabenick, S.A., Watt, H.M.G. (Eds.), Teacher Motivation: Theory and Practice. Routledge, New York, pp. 3–19.

Richardson, P.W., Watt, H.M.G., Karabenick, S.A. (Eds.), 2014. Teacher Motivation: Theory and Practice. Routledge, New York.

Robertson, S., Keith, T., Page, E., 1983. Now who aspires to teach? Educational Researcher 12 (6), 13–21.

Roth, G., 2014. Antecedents and outcomes of teachers' autonomous motivation: a self-determination theory analysis. In: Richardson, P.W., Karabenick, S.A., Watt, H.M.G. (Eds.), Teacher Motivation: Theory and Practice. Routledge, New York, pp. 36–51.

Roth, G., Assor, A., Kanat-Maymon, Y., Kaplan, H., 2007. Perceived autonomy in teaching: how self-determined teaching may lead to self-determined learning. Journal of Educational Psychology 99, 761–774.

Ryan, R.M., Deci, E.L., 2000. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemporary Educational Psychology 25, 54–67.

Ryan, R.M., Deci, E.L., 2009. Promoting self-determined school engagement: motivation, learning, and well-being. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook on Motivation at School. Routledge, New York, pp. 171–196.

Serow, R.C., Forrest, K.D., 1994. Motives and circumstances: occupational-change experiences of prospective late-entry teachers. Teaching and Teacher Education 10 (5), 555–563.

Shim, S.S., Cho, Y., Cassady, J.C., 2013. Goal structures: the role of teachers' achievement goals and theories of intelligence. Journal of Experimental Education 81, 83–103. http://dx.doi.org/10.1080/00220973.2011.635168.

Taylor, I., Ntoumanis, N., 2007. Teacher motivational strategies and student self-determination in physical education. Journal of Educational Psychology 99, 747–760. http://dx.doi.org/10.1037/0022-0663.99.4.747.

Taylor, I., Ntoumanis, N., Standage, M., 2008. A self-determination theory approach to understanding antecedents of teachers' motivational strategies in physical education. Journal of Sport and Exercise Psychology 30, 75–94.

Tschannen-Moran, M., Woolfolk Hoy, A., 2001. Teacher efficacy: capturing an elusive construct. Teaching and Teacher Education 17, 783–805. http://dx.doi.org/10.1016/S0742-051X(01)00036-1.

Tschannen-Moran, M., Woolfolk Hoy, A., Hoy, W.K., 1998. Teacher efficacy: its meaning and measure. Review of Educational Research 68, 202–248. http://dx.doi.org/10.3102/00346543068002202.

Tudhope, W.B., 1944. Motives for choice of the teaching profession by training college students. British Journal of Educational Psychology 14, 129–141. http://dx.doi.org/10.1111/j.2044-8279.1944.tb01550.x.

Watt, H.M.G., Richardson, P.W., 2007a. Motivational factors influencing teaching as a career choice: development and validation of the FIT-Choice scale. Journal of Experimental Education 75, 167–202. http://dx.doi.org/10.3200/JEXE.75.3.167-202.

Watt, H.M.G., Richardson, P.W., 2007b. The Teacher Style Scale. Unpublished document. Monash University, Melbourne.

Watt, H.M.G., Richardson, P.W., 2008. Motivations, perceptions, and aspirations concerning teaching as a career for different types of beginning teachers. Learning and Instruction 18, 408–428. http://dx.doi.org/10.1016/j.learninstruc.2008.06.002.

Watt, H.M.G., Richardson, P.W., 2012. An introduction to teaching motivations in different countries: comparisons using the FIT-Choice scale. Asia-Pacific Journal of Teacher Education 40, 185–197. http://dx.doi.org/10.1080/1359866X.2012.700049.

Watt, H.M.G., Richardson, P.W., Klusmann, U., Kunter, M., Beyer, B., Trautwein, U., Baumert, J., 2012a. Motivations for choosing teaching as a career: an international comparison using the FIT-Choice scale. Teaching and Teacher Education 28, 791–805. http://dx.doi.org/10.1016/j.tate.2012.03.003.

Watt, H.M.G., Shapka, J.D., Morris, Z.A., Durik, A.M., Keating, D.P., Eccles, J.S., 2012b. Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: a comparison of samples from Australia, Canada, and the United States. Developmental Psychology 48, 1594–1611. http://dx.doi.org/10.1037/a0027838.

Yong, B.C.S., 1995. Teacher trainees' motives for entering into a teaching career in Brunei Darussalam. Teaching and Teacher Education 11, 275–280. http://dx.doi.org/10.1016/0742-051X(94)00023-Y.

Young, B.J., 1995. Career plans and work perceptions of preservice teachers. Teaching and Teacher Education 11, 281–292. http://dx.doi.org/10.1016/0742-051X(94)00024-Z.

Relevant Websites

www.fitchoice.org.
http://www.psych.rochester.edu/research/apav/.
http://www.rcgd.isr.umich.edu/garp/.
http://www.selfdeterminationtheory.org/.

Teacher Stress and Burnout: Methodological Perspectives

Chris Kyriacou, University of York, York, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.S. Guglielmi, volume 23, pp. 15465–15468, © 2001, Elsevier Ltd.

Abstract

Teacher stress can be defined as the teacher's experience of negative emotions such as anxiety, frustration, tension, anger, or depression, in response to the demands and pressures they face in their work as a teacher. Teaching in schools is widely recognized as being one of the high-stress occupations. Our understanding of the nature of teacher stress and the processes involved provides a good basis for changes in schools that can help reduce teacher stress.

This article looks at the meaning and nature of teacher stress, what causes it, and what can be done about it. There is little doubt that teacher stress is an area of international concern both in terms of being a focus of research looking at the nature of teacher stress and the processes involved, and in terms of the development and evaluation of policies and intervention strategies designed to reduce it (Kyriacou, 2011; Lambert and McCarthy, 2006).

Defining Teacher Stress

'Teacher stress' originated as a short-hand term for occupational stress among schoolteachers, although it can also be applied to teachers working in other educational settings. The term 'stress' became widely used in occupational settings during the second half of the twentieth century. Unfortunately the term 'stress' can be used in a variety of ways, and this has led to a diversity of definitions of teacher stress.

Firstly, it can be used to refer to the level of demands and pressures that teachers face. Secondly, it can refer to the emotional, physiological, and/or behavioral response to such demands and pressures. Thirdly, it can refer to the degree of mismatch between the level of demands and pressures that teachers face on the one hand, and their capacity to deal with such demands and pressures on the other hand. Moreover, within each of these three main types of definition, researchers have embellished features of their definition in particular ways.

The second of these of types of definitions, in which teacher stress can be defined as "the teacher's experience of negative emotions such anxiety, frustration, tension, anger and depression, in response to the demands and pressures they face in their work as a teacher," is the most popular definition that has been adopted by researchers looking at teacher stress. It is this definition of teacher stress that is adopted in this article.

A number of theoretical models of stress in general and of occupational stress more specifically have been developed (Kyriacou, 2000; Langan-Fox and Cooper, 2011; Rogers, 2012). This has enabled us to get a clear picture of the nature of teacher stress. At the heart of applying such models to teacher stress is a focus on how a teacher perceives aspects of their working situation. All teachers face a range of demands. The nature and level of demands they face, how important it is to meet these demands, and the level of quality at which these demands need to be met are all partly influenced by external sources of pressure (from the school principal, students, parents, colleagues, and government agencies) and partly by the teacher's own expectations, beliefs, and values concerning their professional practice.

Teacher stress, in terms of the experience of negative emotions, is triggered by a perception that involves three key elements:

- The demands the teacher is facing are important (e.g., examination scripts need to be marked by the given deadline, a student misbehaving in the classroom needs to be controlled, and technical equipment needs to be checked before the lesson starts).
- Meeting these demands is difficult (e.g., this may be as a result of the teacher's view about their own ability, other aspects of their workload, and time pressures linked to other priorities).
- Not meeting these demands at a satisfactory level will have adverse consequences (e.g., impaired self-perception or how they are perceived by others, complaints by colleagues, students or parents, and impaired chances for a promotion).

If one of these three elements is missing from the situation, because the demands are not important, or the demands can be met easily, or not meeting the demands will have no adverse consequences, then teacher stress is not triggered.

There are several other terms in common use that have strong links with or conceptual overlap with teacher stress. These include teacher job satisfaction (or more specifically job dissatisfaction), teacher morale, teacher well-being, and teacher burnout. Of these, teacher burnout is the most important (Vandenberghe and Huberman, 1999). Teacher burnout is viewed as a syndrome of physical, emotional, and attitudinal exhaustion toward one's work, which results from experiencing teacher stress over a long period. In essence the teacher experiencing burnout no longer feels engaged with their work, as evidenced by a lack of energy, enthusiasm, and commitment. While early studies of teacher burnout focused on experienced teachers, more recent studies have identified that symptoms of burnout can be evident at an early stage of a teacher's career, including during the preservice period of initial training.

The Measurement of Teacher Stress

Measuring teacher stress has posed researchers with a number of problems. If teacher stress is essentially an emotional

experience, then we need to consider how well one teacher's report of how much stress they are experiencing equates with another's in a way that is both reliable and valid. If one teacher views being mildly annoyed as an indicator of a high level of stress and another views waking up each night in a state of panic as not a sign of stress, then the data collected will need to be regarded with suspicion. Fortunately, research studies using such measures appear to be fairly robust.

The most frequently adopted method of exploring teacher stress has been to ask teachers to complete a questionnaire. This will typically contain questions about their level of stress, the main sources of stress they experience (in terms of intensity and/or frequency), and what strategies they have adopted to reduce their level of stress. Over the years cross-referencing of research findings using different types of questionnaires and other methods of data collection (e.g., interviews, self-reported or observed behavioral indicators, and physiological measures) have provided a degree of confidence in the measures used.

Using such measures with data sets to compare groups (e.g., male vs female teachers), or to investigate a particular context (e.g., beginning teachers, special education teachers), or to drill down to the level of the individual teacher as part of a case study approach has proved to be very successful. In addition, such measures have allowed researchers to explore the relationship between teacher stress and a host of other variables that influence the processes involved in generating teacher stress (Helms-Lorenz et al., 2012; Klassen, 2010; Liu and Onwuegbuzie, 2012).

The Extent of Teacher Stress

Questionnaire data using a stem question such as "In general, how stressful do you find being a teacher?" coupled with a five-point rating scale labeled "not at all stressful, mildly stressful, moderately stressful, very stressful, and extremely stressful" has consistently indicated that between 20 and 30% of teachers report that being a teacher is either very stressful or extremely stressful. Some of the variability in the precise figure reported in different studies is a reflection of the setting of the study (e.g., the country or region within a country, the type of school, the type of students attending the school, and the time of year).

Comparisons have also been made between subgroups of teachers (e.g., beginning teachers vs experienced teachers, male teachers vs female teachers, well-qualified teachers vs less well-qualified teachers). While some studies have indicated differences between particular subgroups others have not, and there is no consistent pattern that has emerged regarding subgroup differences. Other data used as indicators of teacher stress have included levels of absence from work, mental health, physical symptoms such as aches and pains, heart rate, and levels of stress-related hormones.

Taken together, and comparing such data with other professional careers, being a teacher has been consistently identified as one of the high-stress professions, alongside being a nurse, a junior doctor, a social worker, a police officer, an air traffic controller, being in the clergy, working in journalism, and work as a theater artiste (Langan-Fox and Cooper, 2011).

The Main Sources of Stress Reported by Teachers

A wealth of research has been published that has enabled the main sources of stress reported by teachers to be identified (Carlyle and Woods, 2002; Kyriacou, 2011). These fall into the following 10 areas:

- administrative work;
- being appraised by others;
- career identity and development;
- classroom teaching;
- dealing with change;
- relationships with colleagues;
- relationships with pupils;
- targets and performance;
- working conditions; and
- workload.

In looking at the main sources of stress reported by teachers, we need to be aware that for a particular teacher their main source of stress may not figure as one that falls within one of the areas listed above. However, for that particular teacher their main source of stress may still contribute to an overwhelming sense of anxiety and frustration. As such, in dealing with teacher stress it is very important to consider each individual teacher's own unique pattern of sources of stress. Case studies based on a life-history approach, where teachers reflect on the ups and downs of being a teacher over several years or more, have been particularly fruitful in illustrating how the main sources of stress vary for each teacher over the course of their careers.

Action to Reduce Teacher Stress

There are two aspects to reducing teacher stress. The first concerns the strategies used by teachers to reduce their level of stress. These are referred to as 'coping actions.' The second concerns what others can do, be it as senior managers in the school, local, regional and national level advisers, policy makers, and those who design and deliver intervention programs, in order to reduce teachers' level of stress (Hayes, 2006; Kyriacou, 2000; Robins, 2010).

Coping actions fall into two main categories: direct action strategies and palliative strategies. Direct action strategies refer to things that a teacher can do either to reduce the likelihood of the source of stress occurring in the first place or to deal with it effectively so that its occurrence is short-lived. In contrast, palliative strategies do not deal with the source of stress, but rather focus on actions that can reduce the emotional experience of stress.

Direction action strategies are regarded as being the most effective as they deal with the source of stress directly. If a teacher experiences a high level of stress because they are finding it difficult to complete the required marking in time, there are several possible solutions – set less work that requires marking, engage students in marking each other's work, create longer gaps between when work is collected for marking and when it needs to be marked by, set aside more time at home to catch up on marking, manage one's time better so that less time is taken up doing things of lesser importance, systematize the

procedure for awarding marks better, share marking with a colleague, complete marking at a faster speed, and complete marking with less detail. If one of these strategies is effective in dealing with this source of stress, then its use as a coping action has been successful. Teachers often accept their situation as a given, rather than think carefully about how and why a particular source of stress occurs, and what changes in their practices can deal with this effectively. All sources of stress can be reduced by adopting practices that increase one's ability to deal with the demands involved more easily and effectively.

Palliative actions are sometimes also referred to as emotion control strategies. In essence, teacher stress involves experiencing negative emotions. There are a range of mental strategies and physical strategies that can lessen the experience of negative emotions. Mental strategies include putting the problem into perspective, looking at the humorous side of the situation, distracting oneself with more pleasant thoughts, not blaming oneself for the situation, being more realistic about what one can achieve, stopping oneself from overdwelling on the problem, and drawing on the emotional support of colleagues, friends, and family. Physical strategies include having a healthy lifestyle, adopting calming techniques such as deep breathing or relaxing muscles which have become tense, engaging in physical activities that are enjoyable (singing, swimming, jogging, eating, listening to music), and sleeping well.

In palliative coping strategies, the source of stress is still there – the teacher is still finding it difficult to complete the required marking in time. However, the teacher experiences less stress in this situation, because the feeling of anxiety and tension that was previously generated has now been mollified. While palliative coping strategies are important, they must not be seen as obviating the need to consider direct action strategies, as it is only the latter that can help generate better professional practice as a teacher.

A huge number of changes and initiatives have been introduced to reduce teacher stress (Rogers, 2012; Skovholt and Trotter-Mathison, 2010). These include preservice and in-service training courses and resources on how to deal with stress; stress reduction audits implemented by senior school managers to remove unnecessary sources of stress; and confidential advice and counseling for teachers experiencing stress. This can be contrasted with a situation a few decades ago where stress was not publicly spoken about, and a teacher experiencing stress would keep that information private, and probably view it as part of the job about which little can be done (other than changing to another job).

Conclusion

There is a wealth of research and other writings to draw upon which provides us with a clear understanding of teacher stress. This has created a basis for changes in schools both at an organizational level and at the level of the individual teacher's practice, for reducing stress. Unfortunately, there is little doubt that the bigger picture in most countries that frames teachers' working lives is one in which the demands made upon teachers and the degree to which their performance is held to account have been increasing. As such, the need to understand teacher stress and to apply this understanding to reducing teacher stress remains a major challenge for the education community.

See also: Environments for Education; Parents and Teachers as Partners in Education; Teacher Expertise; Teaching as a Profession: United States.

Bibliography

Carlyle, D., Woods, P., 2002. The Emotions of Teacher Stress. Trentham, Stoke-on-Trent, UK.

Hayes, C., 2006. Stress Relief for Teachers: The 'Coping Triangle'. Routledge, Abingdon, UK.

Helms-Lorenz, M., Slof, B., Vermue, C.E., Canrinus, E.T., 2012. Beginning teachers' self-efficacy and stress and the supposed effects of induction arrangements. Educational Studies 38, 189–207.

Klassen, R.M., 2010. Teacher stress: the mediating role of collective efficacy beliefs. Journal of Educational Research 103, 342–350.

Kyriacou, C., 2000. Stress-Busting for Teachers. Nelson Thornes, Cheltenham, UK.

Kyriacou, C., 2011. Teacher stress: from prevalence to resilience. In: Langan-Fox, J., Cooper, C.L. (Eds.), Handbook of Stress in the Occupations. Edward Elgar, Cheltenham, UK, pp. 161–173.

Lambert, G., McCarthy, C. (Eds.), 2006. Understanding Teacher Stress in an Age of Accountability: Research on Stress and Coping in Education. Information Age Publishing, Charlotte, NC.

Langan-Fox, J., Cooper, C.L. (Eds.), 2011. Handbook of Stress in the Occupations. Edward Elgar, Cheltenham, UK.

Liu, S., Onwuegbuzie, A.J., 2012. Chinese teachers' work stress and their turnover intention. International Journal of Educational Research 53, 160–170.

Robins, M.P. (Ed.), 2010. The Pressures of Teaching: How Teachers Cope with Classroom Stress. Kaplan, New York.

Rogers, B., 2012. The Essential Guide to Managing Teacher Stress: Practical Skills for Teachers. Pearson, Harlow, UK.

Skovholt, T.M., Trotter-Mathison, M.J., 2010. The Resilient Practitioner: Burnout Prevention and Self-Care Strategies for Counselors, Therapists, Teachers and Health Professionals, second ed. Routledge, New York.

Vandenberghe, R., Huberman, A.M. (Eds.), 1999. Understanding and Preventing Teacher Burnout: A Sourcebook of International Research and Practice. Cambridge University Press, Cambridge, UK.

Teaching as a Profession: United States

Ilana S Horn, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by E. Hoyle, volume 23, pp. 15472–15476, © 2001, Elsevier Ltd.

Abstract

The status of teaching as a profession in the United States has changed in the twenty-first century. The landmark accountability legislation *No Child Left Behind* has challenged efforts to professionalize teaching, pushing it further from semiprofessional status. This article examines issues of teacher certification, the role of professional organizations, workplace conditions (including attrition and pay), and efforts to create staged teaching careers in this new policy environment.

Introduction

Historically, the work of teaching in the United States has been viewed as a semiprofession. With its roots as pink-collar work, ideal for women who sought employment before starting families, its status has been what Lortie (1975) famously described as "special but shadowed" (p. 10). The continued lack of consensus on a critical knowledge base; the blue-collared organization of the workplace; the absence of a centralized professional organization; and relatively low salaries have maintained this status over the years, despite repeated efforts to professionalize the work.

Nonetheless, the teaching profession has seen some notable changes in recent decades. The strong interventionist federal education policy of *No Child Left Behind* (NCLB, 2002), coupled with the dramatic rise in alternative pathways into the work, have changed everything from notions of expertise in teaching to workplace conditions. As a result, this special but shadowed profession has shifted further into the dark than the light.

In its essence, NCLB is an accountability policy that also sought to address persistent inequalities in students' access to good teachers. However, the simultaneous increase in regulations of instruction and related decreases in teacher autonomy, coupled with substantially eased entry requirements, resulted in an overall devaluation of teachers as professionals.

To describe the current state of the teaching profession in the United States, I look at common indicators of professionalism. Specifically, I describe the training and licensing for entry into the profession; working conditions, including issues of attrition, retention, and pay; and the notion of career.

Training and Licensing

A contradictory set of conditions has emerged when it comes to training and licensing teachers. The text of the NCLB legislation calls for 'highly qualified' teachers in every classroom; however, never before have so many alternative pathways to certification existed. This environment sends mixed messages about what it means to be a good teacher. The discourse of rigor is effectively undermined by policies that make for ever easier entry into the profession.

Defining High-Quality Teachers

The conception of teacher competence animating NCLB can be traced to the landmark 1966 Coleman Report, which found that school funding did not impact educational outcomes as much as teacher quality did. Part of what Coleman and his colleagues identified is what has come to be known as the maldistribution of qualified teachers. In the United States, poor children and children from historically underrepresented groups are disproportionately assigned to the weakest teachers – a situation that persists to this day (Ingersoll, 2003). International comparisons reveal that the United States stands out in this maldistribution problem: our country's disparity in students' access to qualified teachers is among the largest in the world (Akiba et al., 2007).

Coleman's formulation has had a continuous impact on notions of teacher quality. Numerous subsequent studies have verified positive associations between student achievement and teachers' academic qualifications. This relationship has been confirmed through correlations between student outcomes with teachers' scores on various standardized exams (Ballou, 1996; Ehrenberg and Brewer, 1994, 1995; Ferguson and Ladd, 1996; Mosteller and Moynihan, 1972); level of teachers' content knowledge, usually proxied through course-taking counts (Darling-Hammond, 2000; Goldhaber and Brewer, 1997, 2000); and years of teaching experience (Darling-Hammond, 2000; Murnane and Phillips, 1981; Rivkin et al., 2005; U.S. Department of Education, 2001).

NCLB is animated by this legacy as it aims to address the maldistribution problem by legislating a definition of teacher quality and mandating highly qualified teachers in every classroom. True to the Coleman Report's logic, highly qualified teachers are defined as those with full certification or licensure, a college degree, and demonstrated content knowledge in the subject being taught.

However, this legal definition has left much open to debate. For instance, in the description of *highly qualified teachers*, the word 'student' only appears once, and then only as a moderator for 'academic achievement standards.' In other words, the relational work of teaching is completely ignored. Social, emotional, and cultural knowledge does not come through the policy's text, nor does the specialized content knowledge teachers require to cultivate students' understanding – what Shulman (1987) called pedagogical content knowledge.

In light of these omissions, the definition of highly qualified teachers may not reach deeply enough into the qualities teachers need to best serve students living in poverty, thus potentially undermining part of the law's intent.

Research on teachers of students historically underserved by schools emphasizes the particular knowledge and qualities teachers need, including cultural knowledge (Gay and Howard, 2000; Ladson-Billings, 1999; Nieto, 1999), a commitment for social justice (Cochran-Smith, 2004; Oakes and Lipton, 1999), alongside identifying organizational features that support more equitable outcomes (Bryk et al., 2010; Lee and Smith, 1996). Furthermore, the law's emphasis on *teachers* and not *teaching* conflates teacher qualifications with quality teaching, which are not interchangeable (Kennedy, 2010; Little and Bartlett, 2010).

In the end, the *highly qualified teacher* clause sets a minimum standard given the pragmatic demands of legislation, but it falls short of more professionalizing notions of teacher quality. As Cohen (2011) argued, there is a huge difference between competence in rote instruction and what has been termed 'ambitious teaching' (Lampert and Graziani, 2009) – teaching that strives to include children in rich forms of content. The mapping between the qualifications outlined in NCLB and ambitious instruction is not obvious. Although research on teacher quality following the Coleman Report related teacher characteristics to student outcomes, ambitious teaching sets the bar higher. In one study, certification and formal education within a subject did not predict the use of ambitious teaching methods; however, the combination of subject training and subject-specific pedagogical training did (Smith et al., 2005).

Quality Teaching and Licensure

In the past two decades, the contribution of teacher training to quality education has been heavily interrogated. In the most damning of these indictments, Secretary of Education Rod Paige's 2002 *Annual Report on Teacher Quality* argued to redefine teacher qualifications to minimize teacher preparation. He referred to certification systems as 'broken' and 'burdensome' (p. 8), urging credentials to restructure to emphasize content knowledge and making student teaching optional.

Alternative Certification Programs

In the wake of this report, alternative certification pathways flourished. The latest estimates are that 20–30% of new teachers come from alternative certification programs (National Research Council, 2010). That is significant shift in the past two decades, enough to impact U.S. schools. Unlike traditional training programs, alternative certification programs typically accelerate preparation, often to fill a specific content, demographic, or geographic need. Thus Teach for America (TFA) publicizes itself as seeking to 'end educational inequity'; Troops to Teachers recruits former military personnel into schools in low-income communities; the New York City Teaching Fellows program solicits prospective teachers for 'New York City classrooms that need them the most.' Increasingly, some programs explicitly strive to diversify the mostly white teacher workforce (Humphrey et al., 2008). The main difference between these programs and traditional programs is the requirement of only few weeks of training before participants teach full time (Kee, 2012).

TFA has become one of the most visible and controversial of these programs. Founded in 1990, TFA recruits college graduates, typically without an education background, to teach in low-income rural and urban schools in exchange for a 2-year commitment. TFA's supporters point to its role in bringing outstanding college graduates into teaching, while its detractors decry the temporary workforce it brings to high-needs schools, contributing to their instability. Indeed, more than 50% of TFA teachers leave after 2 years and more than 80% leave after 3 years (Vasquez Helig and Jez, 2014). Compared to the overall attrition rate in teaching, estimated among new teachers in the first 5 years to be between 30 (Darling-Hammond, 2010) and 50% (Ingersoll and Smith, 2003), TFA's rate is high. If TFA and similar programs aim to alleviate the teacher shortage, the comparatively high turnover rates work against this goal. In fact, teacher turnover has been found to be detrimental to student achievement, with this effect felt most strongly in schools with high numbers of low-performing students (Ronfeldt et al., 2013).

Aside from the disruption experienced by children and staff, teacher turnover brings immense fiscal burden to districts (Smith and Ingersoll, 2004; Texas Center for Educational Research, 2000). Are alternatively certified teachers worth the expense and churn? Peer-reviewed research on the quality of alternatively certified teachers is not yet conclusive, but it converges on the idea that certification – especially pedagogical training – is consequential for teachers' overall efficacy (Darling-Hammond et al., 2001; Foote et al., 2011; SRI International, 2002; Wilson et al., 2001b). The preponderance of evidence points to teacher preparation as enhancing the initial effectiveness of new teachers, as well as increasing the likelihood that they will teach long enough to grow in that efficacy, since teachers' effectiveness has been shown to improve significantly after their third year (Boyd et al., 2005; Clotfelter et al., 2007).

The Response of Traditional Teacher Education

At the same time that alternative routes into the teaching have proliferated, traditional teacher education programs have responded to numerous criticisms about their efficacy by trying to increase rigor. Several decades ago, in response to reports from groups like the Carnegie Task Force on the Future of Teaching (1986) and the Holmes Group (1986), many teacher education programs increased certification requirements in subject matter preparation, coursework in pedagogy and diversity, and strengthened clinical experiences. There is some evidence that these efforts help novices feel better prepared, decrease attrition, and allow them to teach more effectively (Andrew, 1990; Andrew and Schwab, 1995; Baker, 1993).

The teaching profession suffers from a lack of consensus on the nature of expertise. Unlike other professions, teaching has multiple governing bodies with differing views on this issue, making such consensus doubtful. Just as formal teacher training is optional on the U.S. education landscape, professional accreditation of teacher education programs is likewise optional. Until recently, the two different bodies – National Council for Accreditation in Teacher Education (NCATE), the

Teacher Education Accreditation Council (TEAC) – existed, representing distinctive values on teaching. In 2013, the two bodies came together as Council for the Accreditation of Educator Preparation (CAEP); the new, sole specialized accreditor for educator preparation.

The new governing body will have to reconcile distinct history of the two organizations. The states and the profession created NCATE, the oldest and largest body, over 50 years ago. NCATE has long taken a leading role in efforts to professionalize teaching. NCATE had governed the majority of teacher education programs, with more than two-thirds of traditionally prepared teachers coming from their accredited institutions (Wise, 2005). Their goal is to use accreditation as a mechanism for supporting professional teacher preparation by judging the degree to which programs meet professional standards based on research consensus about quality teaching. NCATE has participated in developing licensing standards (2002), aligning tests to professional standards in addition to their evaluative work.

TEAC, on the other hand, supported a looser stance on teacher quality. A relative newcomer, TEAC was formed in the 1990s out of a task force of the Council of Independent Colleges, constituted by college presidents who wanted to develop a less intensive approach to accreditation (Wise, 2005). Unlike NCATE's clear standards, TEAC took an agnostic stance, permitting institutions to decide their own standards, eroding the notion of professional knowledge in the process. On the whole, the CAEP standards look more aligned to NCATE than TEAC (Council for Accreditation of Educator Preparation, 2013).

Teachers unions are also sources of professional legitimacy. For instance, the National Education Association, the larger of the two, has a stance in line with NCATE's. In a recent policy brief (2011), they called for 'profession-ready teachers' – teachers who are ready to support student learning from their first day on the job. Among their qualities, profession-ready teachers demonstrate appropriate subject matter knowledge, pedagogical content knowledge, and have completed classroom-based performance assessments. These provide uniform tools that allow teacher candidates to connect theoretical ideas about teaching to the classroom. Typically, teacher educators score these portfolios using standardized scoring rubrics (Darling-Hammond, 2006), giving them feedback on their programs.

Although some of the professional organizations assert a vision of professional teaching through expanded training and articulated visions of expertise, they are competing on a landscape that increasingly marginalizes their value. In this way, the debate about professionalism in teaching continues.

Working Conditions

To understand teachers' working conditions, it is critical to understand norms that have shaped and been shaped by school as a workplace. First, there is a longstanding norm of *egalitarianism* (Lortie, 1975; Smylie, 1992), the notion all teachers deserve the same status and recognition. The standardized salary scale rewarding time-at-the-job over performance exemplifies this norm. In the United States, teachers unions typically negotiate schedules to establish salaries that are based on teachers' level of education and time served. Second, scholars find a commonplace norm of *seniority* (Little, 1990), wherein leadership was assumed to arise from experience and veteran teachers got the choicest students and teaching schedules. Finally, a prevalent norm of *autonomy*, which Lortie (1975) observed long ago, manifests in a widely held assumption that teachers choose the particulars of what and how they teach.

Numerous observers of the teacher workplace have commented on the primacy of autonomy among these norms (Donaldson et al., 2008; Johnson, 1990; Little, 1988; Mangin, 2005). Once teachers are in their classrooms, they are not typically beholden to colleagues. Thus enormous instructional variability can be found even in the same school. Teachers also commonly display caution in soliciting and engaging colleagues' opinions and feedback.

Teaching is shaped by the bureaucratic structure of schooling. Bureaucracy necessitates sorting students in ways that stand to interfere with the democratic goals of education, particularly for students from historically underserved populations. Teachers committed to educational equity often embrace another set of norms, those of *change agents* as they work against the bureaucratic organization. They use their knowledge of children and commitment to communities to manage the tensions that arise out of schooling's usual social reproduction functions (Cochran-Smith, 2004; Gay and Howard, 2000; Hollins and Guzman, 2005; Ladson-Billings, 1995; Nieto, 1999; Oakes and Lipton, 1999). Whether teachers work with them or against them, norms both reflect and shape the school as a workplace.

Inequity in Working Conditions

Working conditions for teaching are highly variable, as U.S. schools differ enormously in the resources they provide to students and educators. As Kozol (1991) documented, the practice of funding schools through local property tax revenues has led to grossly unequal resources, from buildings to textbooks, in schools that are walking distance from one another. More recently, investments from middle-class families, through fund-raising work and volunteering, have exacerbated inequities, as parent associations support everything from building upkeep, arts programs, and other resources to augment local school budgets (Posey, 2012). For these reasons, challenging conditions prevail in schools serving students in poverty, with larger class sizes and fewer materials (NCES, 1997).

Despite these inequalities, schools as workplaces share common features that contribute to the semiprofessional status of teaching. Classrooms are the most crowded institutional settings (Jackson, 1968), and compulsory attendance makes teachers work for clientele who have varying interest in their services (Cohen, 2011). Teachers' schedules exhibit the inflexibility typically associated with blue-collar jobs and little of the flexibility associated with the professions (Goodlad, 2004). Contracts are generally arrived at through collective bargaining, making consideration for individual circumstances rare. Teachers commonly work beyond the contracted workday – especially if they take on ambitious forms of instruction (Bartlett, 2004). Taken together, these circumstances too often set any individual's efforts to take professional stance on

teaching against the grain of the institutional setting (Cochran-Smith, 1991).

Class sizes, school buildings, and resources have been recurring themes in research on school as a workplace for some time. In the wake of NCLB, however, new concerns arise. NCLB mandates accountability systems emphasizing school- and district-level growth targets as measured by standardized tests. Unfortunately, because performance on standardized tests has historically been linked to student socioeconomic status and parents' levels of education (Reardon, 2011), this means of accountability only increase the salience of school demographics to teachers' decisions to stay or leave, increasing turnover in schools already considered 'hard-to-staff' (Opfer, 2011). As NCLB takes hold in U.S. teaching culture, schools deemed low performing are publicized – and often stigmatized – since test scores are commonly announced in newspapers, even with individual teachers' ratings. NCLB has perversely created disincentives for teachers to work in poor communities: schools labeled as "low performing" find it increasingly hard to retain or attract teachers (Clotfleter et al., 2003).

Teacher Retention and Attrition

National concerns about the teacher workforce often focus on the issue of teacher turnover. Soon after Richard Ingersoll (2002) found that as many as 46% of beginning teachers left the profession within 5 years, the National Commission on Teaching and America's Future called teacher retention a "national crisis" (p. 21), highlighting the particularly high attrition rates among both new teachers and those working in schools serving historically underserved student populations.

When teachers describe their decisions to change schools or leave the profession, they often cite work conditions – resources, administrative support, student demographics, and school climate (Darling-Hammond, 1997; Rivkin et al., 2005; Ingersoll, 2002). Darling-Hammond (1997) found that more than one-quarter of school leavers cited dissatisfaction with teaching as a reason for their decision, with teachers in high-poverty schools more than twice as likely to leave because of dissatisfaction than those in more affluent schools. Teacher turnover, especially among new teachers, is typically due to low salaries, lack of administrative support, and student discipline problems (Guarino et al., 2006; Ingersoll and Smith, 2003). Although a number of studies have related teacher attrition to student demographics (Carroll et al., 2000; Scafidi et al., 2007), other studies suggest that teachers in demographically similar schools were more likely to leave because of differences in resources, decoupling student characteristics from other aspects of the work environment (Harris, 2002). Facility quality is an important predictor in teachers' decisions to leave, even after controlling for other factors (Buckley et al., 2005).

Teacher Quality and Teacher Attrition

In general, attrition and retention studies focus on the characteristics of the teacher labor force, the working conditions in schools, or a combination of these. Most of these studies do not attend to issues of teacher quality, even though the most optimal case would be retaining the best teachers (Guarino et al., 2006; Johnson, 2004). Recently, proxies for teacher quality have been used to investigate this question by investigating relationships between attrition and value-added to students' test scores (Boyd et al., 2009; Goldhaber et al., 2007; Hanushek et al., 2004). Alarmingly, some studies have found that the most qualified teachers leave the profession at greater rates (Boyd et al., 2005; Podgursky et al., 2004).

Aside from teacher effectiveness measures, teacher traits like minority status have been examined as they relate to attrition. Using the large-scale *Schools and Staffing Survey and the Teacher Follow-up Survey*, Connor (2011) found that by 2004–05, minority teacher turnover in the United States had become significantly greater than turnover among white teachers – although minority teachers were less likely than white teachers to exit schools in high-poverty areas or with large populations of minority students. In fact, minority teachers may be more inclined to stay in high-poverty schools if they have explicitly humanistic missions (Achinstein et al., 2010), suggesting that the intrinsic rewards of teaching (Lortie, 1975) still matter.

Induction

One effort to retain teachers comes in the form of induction programs. These typically provide beginning teachers with different forms of support during their first years of teaching. In their review of research on induction programs, Ingersoll and Strong (2011) found that, on the whole, support programs for new teachers had positive impacts on three outcomes: teacher commitment and retention, classroom instruction, and student achievement. Literature on effective induction programs suggests certain features increase their efficacy. First, new teachers should have mentors to help them navigate typical problems they face (Feiman-Nemser, 1983, 2001; Gold, 1996; National Commission on Teaching and America's Future, 2003; Veenman, 1984). Ideally, the mentor–mentee relationship should be a trusting one (Daloz, 1986; Gold, 1996; Villani, 2002). Mentors should offer assistance around the particulars of teaching – planning, instruction, management – and opportunities to observe, provide feedback on instruction, and connect to resources for selecting and adapting materials (Darling-Hammond, 1999; Evertson and Smithey, 2000; Feiman-Nemser, 1983; Holloway, 2001; Humphrey et al., 2000; Wilson et al., 2001a). Finally, in keeping with the norms of autonomy, effective mentors also know when to grant mentees their privacy and independence (Little and McLaughlin, 1993). Although the qualities of a good mentor have been articulated in the literature, Kardos and Johnson (2007) found that these ideal arrangements are not typical of what new teachers experience in induction programs. Common shortcomings of induction programs include mismatches between mentors and mentees, inadequate classroom observation, infrequent opportunities for interaction – with these supports particularly absent in low-income schools or for teachers of mathematics or science.

Teacher Pay

Teacher pay is low compared to most college-educated professions. Flyer and Rosen (1997) maintained that between 1960 and 1990, the 'true' real wages of teachers, particularly for elementary teachers, declined in comparison with wages for other college graduates. The relative decline in pay compounds

staffing difficulties, since overall levels of compensation are associated with variations in teacher quality. Globally, Ladd (2007) found that, in higher salary OECD countries, the share of underqualified teachers is low (less than 4%) when compared to countries with lower salaries (more than 10% underqualified). Within the United States, this pattern also holds. For instance, Figlio (1997) used the Survey on School Staffing data to look at 19 large U.S. metropolitan districts to test the hypothesis that higher pay attracted more qualified teachers. When controlling for other district factors, higher paying districts more consistently attracted teachers from selective undergraduate institutions and with infield degrees. Other researchers have similarly found that teacher salaries are positively associated with retention (Podgursky et al., 2004; Hanushek et al., 2004; Kelly, 2004).

Educational economists have found that increases in pay result in higher student achievement. For instance, researchers found that increased teacher salaries produced larger effects in student achievement when compared to other expenditures (Greenwald et al., 1996; Ferguson, 1991). A longitudinal analysis across school districts found that students' educational attainment increased most in states and districts that invested in teacher wages (Loeb and Page, 2000).

However, bonus pay incentives seem less effective at addressing recruitment and retention issues, a strategy that numerous states have tried. Researchers found that money was less of an incentive to new hires than other work issues, such as opportunities for alternative certification or working conditions in schools (Liu et al., 2004).

Making a Staged Career

What has been called the 'career-less structure of teaching' is a long recognized obstacle to professionalization. As Lortie (1975) described, the limited opportunity for advancement makes teaching a flat career, with little differences in expectations between novices or veterans (Kardos and Johnson, 2007). In response, attempts to professionalize teaching often seek structures to signal professional advancement.

Second Stage Roles

In the 1980s, school districts created second stage roles for teacher leaders. These were intended to acknowledge expertise and create a mechanism for instructional improvement. Eventually, most of these second stage roles waned. Identifying expertise proved difficult, funding was short-term, and job responsibilities underspecified (Donaldson et al., 2008). With NCLB's demands for evidence for instructional improvement via measurable gains in student achievement scores, districts showed a renewed interest in school-level instructional expertise. A new wave of coaches and instructional leaders provided opportunities for second stage teachers (Boudett et al., 2005; Neufeld and Roper, 2003; Poglicino et al., 2003).

Advanced Certification

In response to the Carnegie *Task Force on Teaching as a Profession*, a nonprofit organization called the National Board for Professional Teaching Standards (NBPTS) was developed around the notion of a professional continuum in teaching. NBPTS describes its mission as "advanc[ing] the quality of teaching and learning by developing professional standards for accomplished teaching, creating a voluntary system to certify teachers who meet those standards and integrating Board-certified teachers into educational reform efforts" (NBPTS, retrieved 2014). By making advance certification, NBPTS institutionally recognizes the difference between novice and experienced teachers. Because NBPTS certification requires more rigorous performances than are typically expected in initial certification, some states and districts provided financial inducements for teachers to obtain them. These had varying aims. In some cases, teachers received bonuses for getting the credential while others specifically sought to bring NBPTS teachers into high-needs schools. With some variation in geography and content area, NBPTS certified supports increases in student achievement (National Research Council, 2008; Clotfelter et al., 2007; Harris and Sass, 2011). Also, on the whole, NBPTS teachers exhibit school-level leadership. In a social network analysis study, NBPTS certified teachers were sought out for instructional help more than their noncertified colleagues (Frank et al., 2008).

While the advanced certification may recognize advancement and expertise, it has done little to alleviate the maldistribution problem. In one study, researchers found that NBPTS teachers were more likely to be teaching in advantaged settings, suggesting that, once again monetary inducements to bring expert teachers into high-needs communities had not succeeded. With the NBPTS certified teachers, unequal distribution existed not only across school districts, but also within schools themselves (Goldhaber et al., 2007).

Workplace Authority

Teachers' agency in the workplace varies from school to school, depending, in part, on school size, administration, local norms, as well as the strength and character of the local teacher community (Talbert and McLaughlin, 1994). Historically, although teachers did not always have authority over their teaching loads or salaries, the norm of autonomy gave them enormous authority behind the closed doors of their classrooms.

The architects of NCLB may have underestimated the extent to which the increasing dependence on state and national standards has eroded this sense of autonomy. This has cut into authority in two ways. First, teachers have been told what to teach – and, increasingly, how to teach it (Gutiérrez et al., 2000). Second, the usually private work of teaching has been made public through the publication of student test scores. Although the details of instruction are still obscured, the results of instruction (student test scores disaggregated by teacher) are made highly visible to the outside world. Through *value-added modeling (VAM)*, teachers' job performance is estimated through statistical modeling of student performance gains based on achievement test scores (Corcoran and Goldhaber, 2013). These have caught the imagination of policymakers as an appealing alternative to more subjective teacher job performance ratings. Thus VAM has been at the heart of initiatives to link pay to performance. The promise of VAM has not been realized, however. In a randomized controlled trial, Springer et al. (2010) found that students whose teachers were eligible for

bonuses did not outperform students whose teachers were not. Aside from their effectiveness as incentives, VAM are not appropriate tools for the job (Haertel, 2013). Because of their instability within individuals over time, they may only be useful indicators of relative performance (very high or very low).

See also: Parents and Teachers as Partners in Education; Teacher Education; Teacher Expertise; Teacher Leadership; Teacher Motivation; Teacher Stress and Burnout: Methodological Perspectives.

Bibliography

Achinstein, B., Ogawa, R.T., Sexton, D., Freitas, C., 2010. Retaining teachers of color: a pressing problem and a potential strategy for "hard-to-staff" schools. Review of Educational Research 80 (1), 71–107.

Akiba, M., LeTendre, G., Scribner, J.P., 2007. Teacher quality, opportunity gap, and national achievement in 46 countries. Educational Researcher 36, 369–387.

Andrew, M., 1990. The differences between graduates of four-year and five-year teacher preparation programs. Journal of Teacher Education 41, 45–51.

Andrew, M., Schwab, R., 1995. Has reform in teacher education influenced teacher performance? An outcome assessment of graduates of eleven teacher education programs. Action in Teacher Education 17, 43–53.

Baker, T., 1993. A survey of four-year and five-year pro- gram graduates and their principals. Southeastern Regional Association of Teacher Educators Journal 2 (2), 28–33.

Ballou, D., 1996. Do public schools hire the best applicants? Quarterly Journal of Economics 111, 97–134.

Bartlett, L., 2004. Expanding teacher work roles: a resource for retention or a recipe for overwork? Journal of Education Policy 19 (5), 565–582.

Boudett, K.P., City, E.A., Murnane, R.J. (Eds.), 2005. Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning. Harvard Education Press, Cambridge, MA.

Boyd, D., Lankford, H., Loeb, S., Wyckoff, J., 2005. Explaining the short careers of high-achieving teachers in schools with low-performing students. American Economic Review 166–171.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., Wyckoff, J., 2009. Who Leaves? Teacher Attrition and Student Achievement (CALDER Institute Working Paper 23). CALDER, Washington, DC.

Buckley, J., Schneider, M., Shang, Y., 2005. Fix it and they might stay: school facility quality and teacher retention in Washington, DC. Teachers College Record 107, 1107–1123.

Bryk, A.S., Sebring, P.B., Allensworth, E., Luppescu, S., Easton, J.Q., 2010. Organizing Schools for Improvement. University of Chicago Press, Chicago.

Carnegie Task Force on the Future of Teaching, 1986. A Nation Prepared: Teachers for the 21st Century. Corporation of New York, New York: Carnegie.

Carroll, S., Reichardt, R., Guarino, C., 2000. The Distribution of Teachers Among California's School Districts and Schools. RAND Corporation, Santa Monica, CA.

Casey, L., 28 February 2012. The True Story of Pascuale Mauclair. Retrieved from: http://www.edwize.org/the-true-story-of-pascale-mauclair.

Clark, M.A., Chiang, H.S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., Puma, M., 2013. The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs (NCEE 2013-4015). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L., Diaz, R., 2003. Do school accountability systems make it more difficult for low performing schools to attract and retain high quality teachers? Paper prepared for the annual meeting of the American Economic Association, Washington, DC. February (revised).

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L., Wheeler, J., 2007. High-poverty Schools and the Distribution of Teachers and Principals (CALDER Working Paper 1). Urban Institute, Washington, DC.

Cochran-Smith, M., 1991. Learning to teach against the grain. Harvard Educational Review 61 (3), 279–310.

Cochran-Smith, M., 2004. Walking the Road: Race, Diversity and Social Justice in Teacher Education. Teachers College Press, New York.

Cohen, D.K., 2011. Teaching and Its Predicaments. Harvard University Press, Cambridge, MA.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPart- land, J., Mood, A.M., Weinfeld, F.C., York, R.I., 1966. Equality of Educational Opportunity. U.S. Department of Health, Education, and Welfare, Washington, DC.

Connor, R., 2011. Examining African-American Teacher Turnover, Past & Present (Unpublished Ph.D. dissertation), University of Pennsylvania, Philadelphia, PA.

Corcoran, S., Goldhaber, D., 2013. Value added and its uses: where you stand depends on where you sit. Education 8 (3), 418–434.

Council on Accreditation of Educator Preparation, 2013. CAEP Accreditation Standards. Author, Wasington, DC.

Daloz, L.A., 1986. Effective Teaching and Mentoring: Realizing the Transformational Power of Adult Learning Experiences. Jossey-Bass Publishers, San Francisco.

Darling-Hammond, L., 1997. Doing What Matters Most: Investing in Quality Teaching. National Commission on Teaching & America's Future (0-9654535-3-7), New York.

Darling-Hammond, L., 1999. Solving the Dilemmas of Teacher Supply, Demand, and Standards: How We Can Ensure a Competent, Caring, and Qualified Teacher for Every Child. National Commission on Teaching and America's Future, Washington, DC.

Darling-Hammond, L., 2000. Teacher quality and student achievement: a review of state policy evidence. Education Policy Analysis Archives 8 (1).

Darling-Hammond, L., 2006. Constructing 21st-century teacher education. Journal of teacher education 57 (3), 300–314.

Darling-Hammond, L., 2010. Teacher education and the American future. Journal of Teacher Education 61 (1–2), 35–47.

Darling-Hammond, L., Berry, B., Thoreson, A., 2001. Does teacher certification matter? Evaluating the evidence. Educational Evaluation and Policy Analysis 23 (1), 57–77.

Darling-Hammond, L., Chung, R., Frelow, F., 2002. Variation in teacher preparation how well do different pathways prepare teachers to teach? Journal of teacher education 53 (4), 286–302.

Donaldson, M., Johnson, S.M., Kirkpatrick, C., Marinell, W., Steele, J., Szczesiul, S., 2008. Angling for access, bartering for change: how second-stage teachers experience differentiated roles in schools. The Teachers College Record 110 (5), 1088–1114.

Ehrenberg, R.G., Brewer, D.J., 1994. Do school and teacher characteristics matter? Evidence from high school and beyond. Economics of Education Review 13, 1–17.

Ehrenberg, R.G., Brewer, D.J., 1995. Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. Economics of Education Review 14 (1), 1–21.

Evertson, C., Smithey, M., 2000. Mentoring effects on protégé s' classroom practice: an experimental field study. The Journal of Educational Research 93 (5), 294–304.

Feiman-Nemser, S., 1983. Learning to teach. In: Shulman, L.S., Sykes, G. (Eds.), Handbook of Teaching and Policy. Longman, New York, pp. 150–170.

Feiman-Nemser, S., 2001. From preparation to practice: designing a continuum to strengthen and sustain teaching. Teachers College Record 103 (6), 1013–1055.

Ferguson, R.F., 1991. Paying for public education: new evidence on how and why money matters. Harvard Journal on Legislation 28, 465–499.

Ferguson, R.F., Ladd, H.F., 1996. How and why money matters: an analysis of Alabama schools. In: Ladd, H.F. (Ed.), Holding Schools Accountable: Performance-based Reform in Education. Brookings Institution, Washington, DC, pp. 265–298.

Figlio, D.N., 1997. Teacher salaries and teacher quality. Economic Letters 55, 267–271.

Flyer, F., Rosen, S., 1997. The new economics of teachers and education. Journal of Labor Economics 15 (1), S104–S139.

Foote, M.Q., Brantlinger, A., Haydar, H.N., Smith, B., Gonzalez, L., 2011. Are we supporting teacher success: insights from an alternative route mathematics teacher certification program for urban public schools. Education and Urban Society 43 (3), 396–425.

Frank, K.A., Sykes, G., Anagnostopouloos, D., Cannata, M., Chard, L., Krause, A., McCrory, R., 2008. Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? Educational Evaluation and Policy Analysis 30 (1), 3–30.

Gay, G., Howard, T., 2000. Multicultural teacher education for the 21st century. The Teacher Educator 36 (1), 1–16.

Gold, Y., 1996. Beginning teacher support: attrition, mentoring, and induction. In: Sikula, J., Buttery, T.J., Guyton, E. (Eds.), Handbook of Research on Teacher Education, second ed. Simon & Schuster Macmillan, New York, pp. 548–594.

Goldhaber, D.D., Brewer, D.J., 1997. Evaluating the effect of teacher degree level on educational performance. In: Fowler, W. (Ed.), Developments in School Finance. National Center for Education Statistics, Washington, DC, pp. 197–210.

Goldhaber, D.D., Brewer, D.J., 2000. Does teacher certification matter? High school teacher certification status and student achievement. Educational Evaluation and Policy Analysis 22, 129–145.

Goldhaber, D., Gross, B., Player, D., 2007. Are Public Schools Really Losing Their ''best''? Assessing the Career Transitions of Teachers and Their Implications for the Quality of the Teacher Workforce. National Center for Analysis of Longitudinal Data in Education Research, Washington, DC.

Goodlad, J., 2004. A Place Called School. McGraw Hill, New York.

Greenwald, R., Hedges, L., Laine, R.D., 1996. The effect of school resources on student achievement. Review of Educational Research 66, 361–396.
Guarino, C., Santibanez, L., Daley, G., 2006. Teacher recruitment and retention: a review of the recent empirical literature. Review of Educational Research 76 (2), 173–208.
Gutiérrez, K.D., Baquedano-López, P., Asato, J., 2000. "English for the children": the new literacy of the old world order, language policy and educational reform. Bilingual Research Journal 24 (1–2), 87–112.
Haertel, E., 2013. Reliability and Validity of Inferences about Teachers Based on Student Test Scores. The 14th William H. Angoff Memorial Lecture. The National Press Club, Washington, DC.
Hanushek, E., Kain, J., Rivkin, S., 2004. Why public schools lose teachers. Journal of Human Resources 39 (2), 326–354.
Harris, P., 2002. Survey of California Teachers. Peter Harris Research Group, Rochester, NY.
Harris, D.N., Sass, T.R., 2011. Teacher training, teacher quality and student achievement. Journal of Public Economics 95 (7), 798–812.
Heilig, J.V., Jez, S.J., 2014. Teach For America: a return to the evidence. National Education Policy Center, Boulder, CO. Retrieved 24 October 2014 at http://nepc.colorado.edu/publication/teach-for-america-return.
Hollins, E., Guzman, M.T., 2005. Research on preparing teachers for diverse populations. In: Cochran-Smith, M., Zeichner, K. (Eds.), Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education. Lawrence Erlbaum, Mahwah, NJ, pp. 477–548.
Holloway, J.H., 2001. The benefits of mentoring. Educational Leadership 58 (8), 85–86.
Holmes Group, 1986. Tomorrow's Teachers. Holmes Group of Education Deans, East Lansing, MI.
Humphrey, D.C., Adelman, N., Esch, C.E., Riehl, L.M., Shields, P.M., Tiffany, J., 2000. Preparing and Supporting New Teachers: A Literature Review. SRI International, Menlo Park, CA.
Humphrey, D.C., Wechsler, M.E., Hough, H.J., 2008. Characteristics of effective alternative teacher certification programs. Teachers College Record 110 (1), 1–63.
Ingersoll, R., 2001. Teacher turnover and teacher shortages: an organizational analysis. American Educational Research Journal 38 (3), 499–534.
Ingersoll, R., 2002. The teacher shortage: a case of wrong diagnosis and wrong prescription. The NASSP Bulletin 86, 16–31.
Ingersoll, R., 2003. Is There Really a Teacher Shortage? a Report Co-sponsored by the Center for the Study of Teaching and Policy and the Center for Policy Research in Education. University of Washington, Center for the Study of Teaching and Policy, Seattle, WA.
Ingersoll, R.M., Perda, D., 2008. The status of teaching as a profession. In: Ballantine, J., Spade, J. (Eds.), Schools and Society: A Sociological Approach to Education. Pine Forge Press, Los Angeles, pp. 106–118.
Ingersoll, R.M., Smith, T.M., 2003. The wrong solution to the teacher shortage. Educational leadership 60 (8), 30–33.
Ingersoll, R.M., Strong, M., 2011. The impact of induction and mentoring programs for beginning teachers a critical review of the research. Review of educational research 81 (2), 201–233.
Jackson, P.W., 1968. Life in Classrooms. Teachers College Press, New York.
Johnson, S.M., 1990. Teachers at Work: Achieving Success in Our Schools. BasicBooks, New York.
Johnson, S.M., The Project on the Next Generation of Teachers, 2004. Finders and Keepers: Helping New Teachers Survive and Thrive in Our Schools. Jossey-Bass, San Francisco.
Kardos, S.M., Johnson, S.M., 2007. On their own and presumed expert: new teachers' experience with their colleagues. Teachers College Record 109 (9), 2083–2106.
Kee, A.N., 2012. Feelings of preparedness among alternatively certified teachers what is the role of program features? Journal of Teacher Education 63 (1), 23–38.
Kelly, S., 2004. An event history analysis of teacher attrition: salary, teach tracking, and socially disadvantaged schools. The Journal of Experimental Education 72 (3), 195–220.
Kennedy, M.M., 2010. Attribution error and the quest for teacher quality. Educational Researcher 39 (8), 591–598.
Kozol, J., 1991. Savage Inequalities: Children in America's Schools. Random House LLC.
Ladd, H.F., 2007. Teacher labor markets in developed countries. Future of Children 17 (1), 201–217.
Ladson-Billings, G., 1995. Toward a theory of culturally relevant pedagogy. American Educational Research Journal 32 (3), 465–491.
Ladson-Billings, G., 1999. Preparing teachers for diverse student populations: a critical race theory perspective. Review of Research in Education 24, 211–247.
Lampert, M., Graziani, F., 2009. Instructional activities as a tool for teachers' and teacher educators' learning. The Elementary School Journal 109 (5), 491–509.
Lee, V., Smith, J., February 1996. Collective responsibility for learning and its effects on gains in achievement for early secondary students. American Journal of Education 104, 103–147.
Little, J., 1990. The persistence of privacy: autonomy and initiative in teachers' professional relations. The Teachers College Record 91 (4), 509–536.
Little, J.W., 1988. Assessing the prospects for teacher leadership. In: Lieberman, A. (Ed.), Building a professional culture in schools. Teachers College Press, New York, pp. 78–106.
Little, J.W., Bartlett, L., 2010. The teacher workforce and problems of educational equity. Review of Research in Education 34 (1), 285–328.
Little, J.W., McLaughlin, M.W. (Eds.), 1993. Teachers' Work: Individuals, Colleagues, and Contexts. Teachers College Press, New York.
Liu, E., Johnson, S.M., Peske, H.G., 2004. New teachers and the Massachusetts signing bonus: the limits of inducements. Educational Evaluation and Policy Analysis 26 (3), 217–236.
Loeb, S., Page, M., 2000. Examining the link between teacher wages and student outcomes: the importance of alternative labor market opportunities and non-pecuniary variation. Review of Economics and Statistics 82, 393–408.
Lortie, D.L., 1975. Schoolteacher: A Sociological Study. University of Chicago Press, Chicago.
Lovett, I., 10 November 2010. Teacher's Death Exposes Tensions in Los Angeles. The New York Times. Retrieved from: http://www.nytimes.com/2010/11/10/education/10teacher.html#.
Mangin, M.M., 2005. Distributed leadership and the culture of schools: teacher leaders' strategies for gaining access to classrooms. Journal of School Leadership 15 (4).
Mosteller, F., Moynihan, D.P. (Eds.), 1972. On Equality of Educational Opportunity. Random House, New York.
Murnane, R.J., Phillips, B.R., 1981. Learning by doing, vintage, and selection: three pieces of the puzzle relating teaching experience and teaching performance. Economics of Education Review 1, 691–693.
National Board of Professional Teaching Standards, 2014. http://www.nbpts.org/ (retrieved 01.03.14.).
National Center for Education Statistics, 1997. America's Teachers: Profile of a Profession, 1993–94. U.S. Department of Education, Washington, DC.
National Commission on Teaching and America's Future, 2003. No Dream Denied: A Pledge to America's Children Summary Report. Author, Washington, DC.
National Council for Accreditation of Teacher Education, 2002. Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education. Author, Washington, DC.
National Education Association, 2011. Profession-Ready Teachers: An NEA Policy Brief. Author, Washington, DC.
National Research Council, 2010. Preparing Teachers: Building Evidence for Sound Policy. The National Academies Press, Washington, DC. Committee on the Study of Teacher Preparation Programs in the United States, Division of Behavioral and Social Sciences and Education.
Neufeld, B., Roper, D., 2003. Coaching: A Strategy for Developing Instructional Capacity: Promises and Practicalities. The Aspen Institute Program on Education and The Annenberg Institute for School Reform, Washington, DC.
Nieto, S., 1999. The Light in Their Eyes. Teachers College Press, New York.
No Child Left Behind Act, 2001. P.L. 107-110, 20 U.S.C. § 6319 (2002).
Oakes, J., Lipton, M., 1999. Teaching to Change the World. McGraw-Hill College, Boston, MA.
Opfer, D., 2011. Defining and identifying hard-to-staff schools the role of school demographics and conditions. Educational Administration Quarterly 47 (4), 582–619.
Paige, R., 2002. Meeting the Highly Qualified Teachers Challenge: The Secretary's Annual Report on Teacher Quality. US Department of Education, Washington, DC.
Podgursky, M., Monroe, R., Watson, D., 2004. The academic quality of public school teachers: an analysis of entry and exit behavior. Economics of Education Review 23, 507–518.
Poglicino, S.M., Bach, A.J., Hovde, K., Rosenblum, S., Saunders, M., Supovitz, J.A., 2003. The Heart of the Matter: The Coaching Model in America's Choice Schools. Center for Policy Research in Education, Philadelphia, PA.
Posey, L., 2012. Middle-and upper-middle-class parent action for urban public schools: promise or paradox? Teachers College Record 114 (1), 1–43.
Reardon, S.F., 2011. The widening academic achievement gap between the rich and the poor: new evidence and possible explanations. Whither Opportunity 91–116.
Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. Econometrica 73 (2), 417–458.
Roberts, G., 26 February 2012. Queen's Parents Demand Answers Following Teacher's Low Grade. Retrieved from: http://www.nypost.com/p/news/local/cursed_with_the_worst_in_queens_f5wLhEdDRN1Wl9h1GQgxAM.
Ronfeldt, M., Loeb, S., Wyckoff, J., 2013. How teacher turnover harms student achievement. American Educational Research Journal 50 (1), 4–36.
Scafidi, B., Sjoquist, D.L., Stinebrickner, T.R., 2007. Race, poverty, and teacher mobility. Economics of Education Review 26, 145–159.

Shulman, L.S., 1987. Knowledge and teaching: foundations of the new reform. Harvard Educational Review 57, 1–22.

Smith, T.M., Desimone, L.M., Ueno, K., 2005. "Highly qualified" to do what? The relationship between NCLB teacher quality mandates and the use of reform-oriented instruction in middle school mathematics. Educational Evaluation and Policy Analysis 27 (1), 75–109.

Smith, T., Ingersoll, R., 2004. What are the effects of induction and mentoring on beginning teacher turnover? American Educational Research Journal 41 (3), 681–714.

Smylie, M.A., 1992. Teachers' reports of their interactions with teacher leaders concerning classroom instruction. The Elementary School Journal 93 (1), 85–98.

Springer, M.G., Ballou, D., Hamilton, L.S., Le, V., Lockwood, J.R., McCaffrey, D.F., Pepper, M., Stecher, B.M., 2010. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. RAND Corporation, Santa Monica, CA. http://www.rand.org/pubs/reprints/RP1416.

SRI International, October 2002. Alternative Certification: Design for a National Study. Author, Menlo Park, CA.

Texas Center for Educational Research, 2000. The Cost of Teacher Turnover. Texas State Board for Educator Certification, Austin.

Talbert, J.E., McLaughlin, M.W., 1994. Teacher professionalism in local school contexts. American Journal of Education 102 (2), 123–153.

U.S. Department of Education, 2001. Monitoring School Quality: An Indicators Report. U.S. Government Printing Office, Washington, DC.

Vasquez Heilig, J., Jez, S.J., 2010. Teach for America: A Review of the Evidence. Education Policy Research Unit.

Veenman, S., 1984. Perceived problems of beginning teachers. Review of Educational Research 54 (2), 143–178.

Villani, S., 2002. Mentoring Programs for New Teachers: Models of Induction and Support. Corwin Press, Thousand Oaks, CA.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. The New Teacher Project, New York.

Wilson, S., Darling-Hammond, L., Berry, B., 2001a. A Case of Successful Teaching Policy: Connecticut's Long-term Efforts to Improve Teaching and Learning. Center for the Study of Teaching and Policy, Seattle, WA.

Wilson, S., Floden, R., Ferrini-Mundy, J., February 2001b. Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations (R-01–3). Center for the Study of Teaching and Policy, Seattle, WA.

Wise, A.E., 2005. Establishing teaching as a profession the essential role of professional accreditation. Journal of Teacher Education 56 (4), 318–331.

Teaching Economics

Micheal Moos van Wyk, College of Education, University of South Africa, Pretoria, South Africa

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We are living in a complex world which is faced with enormous economic challenges. The purpose of economic education is to create responsible citizens and effective decision makers. The importance of economic education goes far beyond the goal of improving an understanding of the basic principles of supply and demand and the workings of the economy. Economics can be taught by generating new knowledge with the help of exposing students to real-life learning environments and experiences. Teaching economics in this twenty-first century requires teachers to implement effective technology-integrated teaching and learning teaching, and assessment strategies not only to achieve cross critical outcomes, but providing ongoing support to the digital natives.

Introduction

Economic education focuses on the scholarship of teaching and learning economics as a subject. It encompasses the content to be taught, methods of teaching, evaluation of those methods, and information of general interest to teachers of economics in elementary through graduate school. Economic education is extremely important because it is vital to the future health of our nation's economy. The purpose of economic education is to create responsible citizens and effective decision makers. Economic education is a very crucial subject that many of our nation's schools tend to overlook. The importance of economic education goes far beyond the goal of improving an understanding of the basic principles of supply and demand and the workings of the economy. Economics can be taught by generating new knowledge with the help of exposing students to real-life learning environments and experiences. Students develop perceptions of their economic world at an early age, which, as they progress through the educational process, develop into attitudes and opinions about the subject of economics. Intended or not, teachers influence the direction of attitude development. By finding ways to teach students more about economics, teachers are contributing to improved attitudes toward the subject. By teaching basic economic concepts and applying them to classroom discussions of economic issues and institutions, teachers and lecturers are not indoctrinating students, but providing a knowledge foundation for more informed student opinions and decision making on vital economic issues. The more economic concepts students know, the more they like and value the subject and the more information they have about economic issues. Students who do not get the opportunity to learn economics and increase their economic understanding will probably never take much interest in the subject or in their economic world.

Organizations Promoting Economic Education

Council for Economic Education (CEE) – the United States

The Council for Economic Education (2009) (the new name, since January 2009, of the National Council on Economic Education) is a national nonprofit organization "committed to empowerment and opportunity through economic and financial literacy". The mission of the CEE (2009) is twofold: to advocate for better and greater school-based economic and personal finance education at the kindergarten through 12th-grade levels (K-12); and to educate young people in the United States and around the world, primarily through well-prepared teachers, so they may become empowered with economic and financial literacy. The National Council on Economic Education (NCEE) is a nationwide network in the United States that aims at promoting economic literacy among learners and their teachers. The NCEE's mission is to help learners develop the real-life skills they need to succeed: to be able to think and choose responsibly as consumers, savers, investors, citizens, members of the workforce, and effective participants in a global economy (NCEE, 1995, 2005). The NCEE *AmericaEconomics* and *InternationalEconomics* (Train-the-Trainers and Writers) programs prepare teachers/lecturers to conduct professional development for teachers of economics at school and university level. The *Train-the-Trainers* program consists of four one-week seminars, each of which focuses on strengthening participants' knowledge and skills of both economic concepts and pedagogy (NCEE, 2005). According to a survey compiled by the NCEE titled *Survey of the States. Economics and personal finance education in our nation's schools in 2007*, there is still much that needs to be done in terms of implementing economic education into our nation's schools. The CEE conducted a follow-up survey called *Survey of the States 2011: The state of economic and personal finance education in our nation's schools*. The 2011 survey reported several key findings:

- The number of states that now require students to take an economics course as a high-school graduation requirement increased from 21 in 2009 to 22 in 2011.
- However, only 16 states require the testing of student knowledge in economics, 3 fewer than in 2009.
- No improvement has been seen in the area of personal finance. The number of states that require students to take a personal finance course (or personal finance included in an economics course) as a high-school graduation requirement remains at 13.
- Only 2 more states now require that personal finance content standards be implemented, bringing the total to 36.

Visit www.surveyofthestates.com to explore the interactive Survey of the States.

Annamaria Lusardi, Denit Trust Professor of economics and accountancy at the George Washington University School of Business, concurred on the above statements and said:

> Just as it was not possible to live in an industrialised society without print literacy – the ability to read and write – so it is not possible to live in today's world without being financially literate. To fully participate in society today, financial literacy is critical.

The Survey of the States is a biennial report that brings attention to the critical importance of economics and personal finance education by documenting its status in the 50 states and the District of Columbia.

In the next paragraph, an example of an organization promoting economic education in a developing economy, the South African Foundation for Economic and Financial Education (SAFEFE), is discussed.

South African Foundation for Economic and Financial Education (SAFEFE)

The NCEE's (now CEE) entry point in South Africa in 2004 was through Professor Claudia Parliament's connection with the University of the Free State (UFS), Bloemfontein. During initial discussions in Lithuania in 2004 between the Council and UFS delegates, it was decided to develop partnerships to strengthen economic delivery systems firstly in the Free State province and soon expanding to the whole of South Africa and later also to the South African Development Communities (SADC) region. In 2008 the South African Foundation for Economic and Financial Education (SAFEFE, 2008) was established as a nongovernmental organization focusing on economic and financial education for K-12 teachers and university lecturers. SAFEFE is now legally organized and has a board of directors, a strategic plan, and a representative group of *Training of Trainers* and *Training of Writers* alumni to assist in the planning and delivery of professional development programs. SAFEFE focuses on seven activities which are seen as priorities: teacher development; material development; partnership building with the national Department of Education and higher education institutions; fundraising; networking with teachers; competitions for learners; and international networking. Currently SAFEFE provides seminars in professional development and other supported services to more than 260 000 economic teachers (grades 4–12), 1200 economics and management sciences and economics subject advisers in all nine provincial departments of education, and several economics lecturers at 18 universities in South Africa.

Integration of Teaching Methods and Media in Teaching Economics

Applying Creative and Innovative Teaching Methods

Economic education is a growing subject in secondary education. The current teaching strategies that are implemented by economic teachers at the high-school level are outdated and do not pertain to the ways in which students best comprehend economic content. Excellent and effective teaching demands a host of devices, techniques, and strategies not only to achieve cross critical outcomes, but because variety, itself, is a desideratum. Prior to looking at the many different teaching strategies, we must first look at why educators struggle with the teaching of economics through the many different strategies available. First and foremost, educators are not adequately educated themselves in terms of economic education. Social studies educators are generally the ones who are responsible for teaching high-school economics courses and the typical social studies educator on average has only completed four to eight credit hours of economics course work. This represents no more than 7% of the total social science hours required for certification (Joshi and Marri, 2006). How can we expect these educators to effectively teach our youth the knowledgeable skills and content of economics when they in fact do not have the requisite knowledge and skills themselves? The literature suggests several teaching strategies that actively involve students in the content, including collaborative problem solving, service learning, interactive role play and simulations, an inverted classroom, experiments/demonstrations/dramatizations, technology-based teaching strategies, and lecturing accompanied by student-centered activities. It is imperative that teachers utilize many teaching strategies and not just limit themselves to a specific one (Goffe and Sosin, 2005). This will allow them to target the many different learning styles their students have, as well as to create a learning environment that will keep their students actively involved as opposed to the same daily routine. Overall, effective teaching styles provide an effective learning environment in the classroom and enable students to apply classroom learning to real-life situations.

In 1995, 2000, 2005, and 2010 national surveys of US academic economists were conducted to investigate how economics is taught and assessed in four different types of undergraduate courses. In this paper we present the first basic results from the 2010 survey to identify persistent patterns and a few gradual changes in teaching and assessment methods. Despite calls for economists and other post-secondary instructors to make greater use of active, student-centered learning methods over the past two decades, 'chalk and talk' has remained the dominant teaching method in all types of undergraduate economics courses. But there is evidence of slow drops in mean (though not median) values for those variables, and of gradual growth in the use of some new teaching methods, including instructor-directed class discussions and computer-generated displays (such as PowerPoint). More instructors provided students with a prepared set of class notes; computer lab assignments were increasingly common in econometrics and statistics courses; and the Internet database searches were used by a growing (though still small) minority of instructors in all types of classes. Classroom experiments are used by a small share of instructors in introductory courses, but almost never in other kinds of courses (Bartlett, 2006). Calculus is still not viewed as important by a majority of economics instructors in any of the four types of courses, but for the 2010 survey these ratings were higher in intermediate theory and statistics and econometrics courses than in the

earlier surveys. Colander et al. (2004) compiled a book entitled *The changing face of economics* that gives a sense of the modern economics profession and how it is changing. The volume does so with a set of nine interviews with cutting-edge economists, followed by interviews with two Nobel Prize winners, Paul Samuelson and Kenneth Arrow, reflecting on the changes that are occurring. Emanating from this source is a clear picture of today's economics – and it is no longer standard neoclassical economics. The interviews and commentary together demonstrate that economics is currently undergoing a fundamental shift in method and is moving away from traditional neoclassical economics into a dynamic set of new methods and approaches. These new approaches include work in behavioral economics, experimental economics, evolutionary game theory and ecological approaches, complexity and nonlinear dynamics, methodological analysis, and agent-based modeling. One of the current problems with economic education is the strategies teachers use to teach the subject. The majority of teachers are stuck on traditional methods that have been in practice for many years. A specific example would be the chalk-and-talk method, which is defined as a traditional method of education in which the teacher addresses the students and uses the blackboard to provide examples or illustrations (Joshi and Marri, 2006). Another example that is very similar to chalk and talk is lecture, another method of direct instruction. Many students who take economics classes get very bored and uninterested in the material taught because of these outdated teaching methods. If teachers were able to use more effective methods in the economics courses they teach, their students would be able to better comprehend the information making them better citizens of not only the United States but the world as well. The issue is that academics cannot teach the same way they were trained a decade ago but need to make the paradigm shift to accommodate the diverse M-generation student population in their teaching styles which enters universities globally.

Service-Learning Approach in Teaching Economics

Service-learning (SL) is a well-established pedagogy used by college and university faculties for over two decades (Stanton et al., 1999). Economists have used SL to promote deep learning in economics classrooms for over a decade (McGoldrick, 1998; Simmons, 2003). Due to the flexibility of this pedagogical technique SL can be used in a multiple of ways in a variety of classroom settings. Scholars of the theory of SL reported that it is an effective tool for teaching economic theory (McGoldrick, 2002; McGoldrick and Ziegert, 2002). McGoldrick (2002) conducted a study by using SL in teaching economics and reported how participants learned to apply economics matter to real-life situations. SL is an approach to learning-by-doing that highlights the social relevance of the work performed by universities. It can be regarded as an augmentation of the concept of community service with the inclusion of a learning component that may carry academic credits or may be part of a specific course requirement (Rhoads, 1997; Mungaray et al., 2007). The fundamental idea is that students learn and develop themselves through active participation in communitarian experiences organized and led by faculty, where they have the opportunity to apply the concepts and instruments recently acquired to help their community (Yates and Yuniss, 1997). More formally, SL is a method of experiential learning that links the classroom with the local community, and requires students to spend time in volunteer service and relate their experiences to the educational knowledge they see in the classroom. For the economics discipline, SL provides a complementary teaching strategy that can motivate learning through experiences that make a social contribution. Furthermore, it is a method by which economics can be made more accessible to an increasingly diverse student body and improve the relationship between colleges and universities and their communities.

Technology-Integrated Media

Some circumstantial evidence suggests that declining numbers of economic majors at the tertiary levels in the 1990s led academic economists to devote more time and energy to teaching. Until the late 1999s, economists paid little attention to changing the way to create more participative teaching methods and strategies to enhance student learning in the classroom. Some academic economists spoke of 'exposing' students to concepts and 'imparting' knowledge, but what was the use for this stance? Most lecturer rooms were used for the transmittal of factual information through textbooks and demonstrations of mathematical and graphical constructions. The lecture, textbook, PowerPoint and blackboard were the simplest delivery modes and teachers of economics used little else (Becker and Watts, 1996). Arnold Katz and William Becker (1999) reported in an article entitled *Advancing the integration of new technologies into the undergraduate teaching of Economics* that the subject is lagging behind other disciplines in creating innovative ways that engaged students more actively in the learning process. In the 2000s a renewed impetus was injected into more research on the use of technology-integrated teaching and learning of economics to undergraduates. Another variation of online communication that is increasingly used in economic education is interactive software for groups working on the Internet. The *ComLabGames* (http://www.cmu.edu/comlabgames) are effective tools whereby students learn game theory with specialized software over the Internet. In some cases, students can play games with other students at remote sites. Such software is similar to 'massively multiplayer' games where hundreds of thousands can interact online. Currently, some of the leading online games are 'Sims Online' and 'Ultima.' More research studies focused on areas related to computer simulations and modeling, machine-based tutoring, laboratory supplements to lecture course, Web-based learning environments, multimedia teaching enhancements, collaborative learning techniques, classroom experiments, social media networking, evaluation and assessment, interdisciplinary instruction, and institutional resource sharing of open educational resources.

Dirk Mateer designed an online website, *Starting point: teaching and learning economics*, to assist economists in developing innovative teaching strategies. This e-learning platform illustrates the benefits of using multimedia to reinvigorate the way to teach economics. The use of media to enhance teaching and learning complements traditional approaches

to learning. Effective instruction builds bridges between students' knowledge and the learning objectives of the course. Using media engages students, aids student retention of knowledge, motivates interest in the subject matter, and illustrates the relevance of many concepts (Mateer, 2011). Students have different learning preferences. While those with a background in science or mathematics may have a comparative advantage in understanding the quantitative aspects of economics, research shows that people learn abstract, new, and novel concepts more easily when they are presented in both verbal and visual form (Salemi and Walstad, 2010). Visual media also make concepts more accessible to a person than text alone, promote deep learning rather than rote learning, and help with later recall (Willingham, 2009). The use of media is beneficial for teachers and students alike. Technology plays an important role in creating an interactive learning environment, which prompts two-way discussion in the form of student-created content (Goffe and Sosin, 2005; Bransford et al., 2000). An excellent list of technology activities for teaching economics can be accessed using the Journal of Economic Education's online section at http://www.indiana.edu/-econed/onlinehome.htm. Media are ideal for illustrating complex ideas in a short period of time, connecting learners with events that are culturally relevant and theories taught in the classroom with real-world events and policies (Mateer, 2011). Using social media of popular culture familiar to students is likely to (1) maintain attention and student interest in the theories and concepts being taught; (2) develop better analytical skills by analyzing media using the theories and concepts they are studying; and (3) break down the barrier between formal learning and understanding, enabling students to see concepts and new examples when they use these same media in their private life (Mateer, 2011). Some economics blogs are created to express the strong viewpoint of the owner, which may be a minority or even crank opinion on an economics issue. Yet, this diversity in viewpoints could be used as an excellent opportunity for teaching critical thinking, with a clear role for the instructor in helping students evaluate, compare, and summarize. A list of economics blogs can be found in the section of Resources for Economists on the Internet devoted to them at http://rfe.org/Blogs. Before using audiovisual materials in teaching, one should have an understanding of copyright, which by default applies to materials found on the web. The Economics Network's *The handbook for economics lecturers* provides useful instructions on this issue. Permission to use these materials might be requested directly or granted in advance using a license such as Creative Commons (Poulter, 2010). The biggest hurdle in using audiovisual materials – aside from a willingness to do so – is the workload involved and the skill in recognizing content that will enhance learning, instead of becoming a distraction (Mateer, 2011). Several sites are available such as *The mathematics for economics: enhancing teaching and learning* (METAL), *TED-talks* are well-produced video lectures on themes as diverse as business climate, developmental economics, environmental issues and globalization, Mateer created *Teaching economics with YouTube* to showcase in-class videos, *Economics in the movies*, *TV for economics*, *From Abba to Zeppelin*, *Led* (Hall et al., 2008) and *Flash music for economics*. The benefits of using multimedia in the classroom are well established in the academic literature but, as a discipline, economics has been slow to adopt innovative approaches to teaching (Becker, 2004; Becker and Watts, 1996, 2001a, b). Media can be used in almost any discipline to improve student learning in a variety of forms, both in class (lectures and tutorials) and out of class (homework, offline/online assignments, essays, and projects). Research shows that the use of multimedia or popular culture can stimulate discussion in introductory classes (Becker, 2004), illustrate basic concepts (Hall, 2005; Hall and Lawson, 2008), and explain abstract concepts like game theory (Dixit, 2005) at an advanced level.

Assessment Practices in Economic Education

Schaur et al. (2008) conducted a survey on *Assessment practices and trends in undergraduate economics courses*. Findings revealed that there was an increased use of multiple-choice questions in all types of undergraduate economics courses, after taking into account the greater use of most types of written assignments and student presentations, and the stable class sizes in these courses from 1995 to 2009. They further concluded that economists have been cutting back on the time and care they devote to assessment. Instead, the general picture is that economists are making greater use of assessment practices that have been more popular for some time in other academic fields. Lastly, it is reported that, although 'chalk and talk' is found to remain, by far, the most prevalent form of instruction in undergraduate economics classes, there is finally some evidence of movements in undergraduate economics courses toward greater use of student-centered teaching methods that are more commonly used in other subject areas. On the other hand, Myers et al. (2011) reported in a survey conducted on the *Assessment of the undergraduate economics major: a national survey* that more instructors did use performance in games, simulations, and classroom experiments for grading purposes, given the rising popularity of experimental economics during this period both as a research field and as a teaching method. Several commercial and academic websites and books on classroom experiments appeared during the period that the three surveys were conducted (http://veconlab.econ.virginia.edu/admin.htm), making it easier for instructors to use classroom experiments. But part of the explanation for this result may simply be that classroom experiments, simulations, and games are often used solely as motivational and pedagogical exercises, and not for assessment.

Another trend shows that economics departments are faced with growing demands to document what their graduates have learned on completion of the undergraduate major (Steven et al., 2011). The results of a national survey of economics department chairs in the United States reveal that nearly two-thirds of the departments have a formal assessment plan (Myers et al., 2011). There is substantial agreement on the most important student-learning outcomes, which are consistent with the Hansen proficiencies. The most common approaches that departments employ to measure learning outcomes are course-embedded assessments and senior exit surveys. Capstone courses and senior projects as program assessment methods are most common in departments that

are not in business schools and are without doctoral programs. Finally, more than half of the departments have adjusted their curriculum based on the results of their own assessment plans.

Conclusion

We are living in a complex world which is faced with enormous economic challenges. Economic education is the study of the scholarship of teaching economics effectively. It encompasses the *what, how and why* we teach economics in relation to the content, teaching methods, assessment in elementary schools and universities. Understanding this world requires education in basic economic concepts and their application to a multitude of economic issues. As students grow and learn, economic attitudes and opinions take shape and influence thoughts and actions over a lifetime. Consequently, academics and teachers need to know the nature of the relationship between attitudes and opinions in economics learning. Despite the increasing use of multiple-choice questions in all types of undergraduate economics courses, after taking into account the greater use of most types of written assignments and student presentations, and the stable class sizes in these courses from 1995 to 2011, we do not conclude that economists have been cutting back on the time and care they devote to assessment practices. Moreover, the general picture is economists and teachers of economics use more effective and applicable assessment tasks and practices that have been for some time employed in other subjects. Subsequently, economic education in schools and at universities is shifting from a predominantly academic approach toward more emphasis on teaching that meets the needs of students in their current and future lives. The challenges and opportunities in teaching economics in the twenty-first century and beyond are immense but we must not lose focus by empowering students to use economic knowledge in their daily life. Because of the global challenges, we must therefore educate students to be creative in becoming responsible citizens, effective change agents, and decision makers.

Bibliography

Allgood, S.W., Bosshardt, W., van der Klaauw, W., Watts, M., 2004. What students remember and say about college economics years later. American Economic Review: Papers and Proceedings 94 (5), 259–265.

Bartlett, R., 2006. The evolution of cooperative learning and economics instruction. In: Becker, W.E., Watts, M., & SR Becker (Eds.), Teaching Economics: More Alternatives to Chalk and Talk. Edward Elgar, Cheltenham, UK.

Becker, W.E., 2004. In: Becker, W.E., Andrews, M. (Eds.), Quantitative Research on Teaching Methods in Tertiary Education in the Scholarship of Teaching and Learning in Higher Education by. Indiana University Press, Bloomington.

Becker, W.E., 2000. Teaching economics in the 21st century. The Journal of Economic Perspectives 14 (1), 109–119.

Becker, W.E., Watts, M., 1996. Chalk and talk: a national survey on teaching undergraduate economics. American Economic Review 86 (2), 448–453.

Becker, W.E., Watts, M., 2001a. Teaching economics at the start of the 21st century: still chalk and talk. American Economic Review 91 (2), 446–451.

Becker, W.E., Watts, M., 2001b. Teaching methods in US undergraduate economics courses. Journal of Economic Education 32 (3), 269–279.

Bransford, J.A., Brown, L., Cocking, R.R., 2000. How People Learn: Brain, Mind, Experience, and School. National Academy, Washington, DC.

Colander, D.C., Holt, R.P.F., Rosser, J.B., 2004. The Changing Face of Economics: Conversations with Cutting Edge Economists. University of Michigan Press, MI.

Council on Economic Education (CEE), 2011. Survey of the States 2011: The State of Economic and Personal Finance Education in Our Nation's Schools. www.surveyofthestates.com (accessed 12.11.12.).

Council of Economic Education, 2008. Survey of the States: Economic and Personal Finance Education in Our Nation's Schools 2007, 2011. http://www.ncee.net (accessed 12.11.12.).

Dixit, A., 2005. Restoring fun to game theory. Journal of Economic Education 36 (3), 205–219.

Goffe, W.L., Sosin, K., 2005. Teaching with technology: may you live in interesting times. The Journal of Economic Education 36 (3), 278–291. http://www.jstor.org/stable/30042660 (accessed 24.01.13.).

Hall, J.C., 2005. Homer economicus: using *the Simpsons* to teach economics. Journal of Private Enterprise 30 (2), 165–176.

Hall, J.C., Lawson, R.A., 2008. Using music to teach microeconomics. Perspectives in Economic Education Research 4 (1), 23–36.

Hall, J.C., Lawson, R.A., Mateer, D.G., 2008. From Abba to Zeppelin, Led: using music to teach economics. Journal of Economic Education 39 (1), 107.

Hansen, W.L., 2004. Student course/instructor evaluations: assessing a proficiencies-oriented economics course. In: Paper Presented at the Annual Meeting of the Midwest Economics Association. Chicago, IL.

Joshi, P., Marri, R.M., 2006. An economics methods course? Challenges of teaching an economic education methods course for secondary social studies preservice teachers. The Social Studies 97 (5), 197–202.

Katz, A., Becker, W., 1999. Advancing the Integration of New Technologies into the Undergraduate Teaching of Economics.

Mateer, D.G., 2011. Using media to enhance teaching and learning. Starting Point: Teaching and Learning Economics. http://serc.carleton.edu/econ/media/index.html (accessed 16.05.11.).

McGoldrick, K., 2002. Using the theory of service-learning as a tool for teaching economic theory. In: McGoldrick, K., Ziegert, A. (Eds.), Putting the Invisible Hand to Work: Concepts and Models for Service-learning in Economics. Edward Zlotkowski.

McGoldrick, K., 1998. Service-learning in economics: a detailed application. Journal of Economic Education 29 (4), 365–376.

McGoldrick, K., Ziegert, A., 2002. Putting the Invisible Hand to Work: Concepts and Models of Service-Learning in Economics. The University of Michigan Press, Ann Arbor.

Mungaray, A., Ramírez-Urquidy, M., Texis, M., Ledezma, D., Ramírez-Angulo, N., 2007. Promoting learning in small entrepreneurs and higher education students through service-learning programs. International Journal of Business Research 3, 10–28.

Myers, M., Nelson, R., Stratton, D., 2011. Assessment of the undergraduate economics major: a national survey. Journal of Economic Education 42 (2), 195–179.

National Council on Economic Education, 2005. Train-the-Trainers' Program 2005–2007: Seminars A, B, C and D. Economics International, New York.

National Council on Economic Education, 1995. Train-the-Writers' Program 2000–2010: Seminars A, B, C and D. Economics International, New York.

Poulter, M., 2010. Creative Commons. Economics Network. http://www.economicsnetwork.ac.uk/themes/licensing.htm (accessed 16.11.12.).

Rhoads, R.A., 1997. Community Service and Higher Learning: Explorations of the Caring Self. State University of New York Press, Albany.

Saunders, P., 1998. Learning theory and instructional objectives. In: Walstad, W., William, B., Saunders, P. (Eds.), Teaching Undergraduate Economics: A Handbook for Instructors. Irwin/McGraw-Hill, Boston, London and Toronto, pp. 85–108.

Schaur, G., Watts, G., Becker, W.E., 2008. Assessment practices and trends in undergraduate economics courses. American Economic Review: Papers & Proceedings 98 (2), 552–556. http://www.aeaweb.org/articles.php?doi=10.1257/aer.98.2.552.

Salemi, M.K., Walstad, W.B., 2010. Teaching Innovations in Economics: Strategies and Innovations for Interactive Instruction. Edward Elgar Publishing Limited, MI.

Simmons, K., 2003. Teaching economics of poverty using a service-learning approach. Social Science Research Network (Electronic Paper Collection). http://ssrn.com/abstract=457201.

South African Foundation for Economic and Financial Education (SAFEFE) 2008. http://safefe.org.za/general/who-is-safefe-and-safefe-directors.tml (accessed 10.03.12.).

Stanton, T.K., Giles Jr., D.E., Cruz, N.I., 1999. Service-Learning: A Movement's Pioneers Reflect on Its Origins, Practice, and Future. Jossey-Bass, San Francisco.

Steven, C., Myers, M., Nelson, A., Stratton, W.R., 2011. Assessment of the undergraduate economics major: a national survey. Journal of Economic Education 42 (2), 195–199. http://dx.doi.org/10.1080/00220485.2011.555722 (accessed 12.03.12.).

Watts, M., Schaur, G., 2010. Teaching and Assessment Methods in Undergraduate Economics: A Fourth National Quinquennial Survey.

Walstad, W.B., 2006. Assessment of student learning in economics. In: Becker, W.E., Watts, M., SR Becker (Eds.), Teaching Economics: More Alternatives to Chalk and Talk. Edward Elgar, Cheltenham, UK.

Watts, M., Becker, W.E., Summer 2008. A little more than chalk and talk: results from a national survey. Journal of Economic Education 273–286.

Willingham, D.T., 2009. Why Don't Students Like School? Jossey-Bass, San Francisco.

Yamarik, S., 2007. Does cooperative learning improve student learning outcomes? Journal of Economic Education 38 (3), 259–277.

Youniss, J., Yates, M., 1997. Community Service and Social Responsibility in Youth. The University of Chicago Press, Chicago.

Teaching for Metacognition

Marcel VJ Veenman, Institute for Metacognition Research (IMR), Hillegom, The Netherlands; and Leiden University, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Metacognition concerns the descriptive knowledge of, and the operational control over one's cognitive processes. Together, these two components of metacognition constitute a student's repertoire of metacognitive self-instructions for the regulation of cognitive performance. Developmental trajectories of metacognition show that the repertoire of self-instructions becomes general by the age of 14, that is, across tasks and disciplines. Students, however, substantially vary in the adequacy of their metacognitive skills. In metacognitive-deficient students, both descriptive and operational components need to be explicitly taught in order to build up an organized repertoire of self-instructions. Conditions for effective training are being discussed.

Defining Metacognition

When Flavell coined the term in the 1970s, metacognition was conceptualized as cognition about cognition, or thinking about thinking (Flavell, 1979). This definition emphasized that metacognition takes one's own cognitive processes as the object of thought. Initially, the concept of metacognition was confined to descriptive knowledge about one's cognitive system, such as knowledge about how memory works. The assumption was that a better understanding of one's cognitive functioning would yield more adequate behavior. Subsequently, Brown and DeLoache (1978) added a self-regulation component to metacognition, referring to the control over one's cognitive behavior. According to this regulatory perspective, skillful planning and monitoring of one's behavior would be responsible for successful performance. Current definitions still reflect the contrast between a descriptive-knowledge component and a regulatory component of metacognition, respectively, referred to as metacognitive knowledge and metacognitive skills (Dinsmore et al., 2008; Veenman et al., 2006). Both components, however, are required for adequate metacognitive control over one's cognitive processes.

Wang et al. (1990) concluded from a literature review that metacognition is the most important predictor of learning performance. Moreover, in a recent overview of studies with students of different age, performing different tasks in various domains, metacognitive skillfulness accounted for 40% of variance in learning outcomes (Veenman, 2008). Students show substantial individual differences in metacognitive skillfulness. The undertow of this conclusion is that some students are deficient in metacognition, which deficiency could be relieved by teaching for metacognition. The ultimate question is what should be taught, and how.

Metacognitive Knowledge

Metacognitive knowledge concerns one's descriptive knowledge about the interplay between person, task, and strategy characteristics (Flavell, 1979). For instance, a person may think that he or she (person characteristic) is poor at recollecting names of people (task characteristic) and, therefore, that he or she should devote attention to rehearsing names (strategy characteristic). Yet, another person may positively appraise his or her capacity for retention of names, thus putting less effort into memorizing them. The second person, however, not necessarily has a better memory for names than the first person. Metacognitive knowledge can be either correct or incorrect as one may underestimate or overestimate the actual level of competency, relative to the complexity of the task (Veenman et al., 2006). One person may be embarrassed when addressing someone with the wrong name, whereas another person may erroneously disregard such memory failure. These two persons are likely to differ in estimations of their memory competency.

Metacognitive knowledge rests on a personal belief system that is built up from metacognitive experiences in the past and individual attributions of causes and consequences (Flavell, 1979). For instance, the metacognitive experience of forgetting another person's name evokes a negative attribution about one's capacity for recollecting names. As belief systems are tacit by nature, they are rather resistant to change (Veenman, 2011). Counter experiences may go unnoticed or they may be framed into existing attributions. Given the imperfection of attributions in the belief system, metacognitive knowledge may be inaccurate or flawed. Nevertheless, a person has to make behavioral decisions based on the available metacognitive knowledge, for better or worse.

Metamemory

Metamemory was discerned first as metacognitive knowledge (Flavell and Wellman, 1977). Initially, metamemory only referred to the descriptive knowledge of one's memory system and of strategies that affect memory processes. Acquisition of factual knowledge about memory processes was assumed to improve memory performance. Later, the focus of metamemory research shifted from the knowledge product to the process of metamemory (Nelson and Narens, 1990). Feeling of Knowing (FOK) and Judgment of Learning (JOL) refer to a student's predictions about future test performance, either on items that have been studied (JOL), or on novel items (FOK). A student's individual level of metamemory accuracy is expressed by the alignment of one's predictions with actual performance on a series of items (e.g., vocabulary items). The more accurate the

student is, the better he or she can decide whether or not further study of items is required. Thus, the monitoring or evaluation of memory contents is a key feature of metamemory. In fact, monitoring and evaluation are metacognitive skills that can be trained in order to facilitate metamemory accuracy.

Metacognitive Conditional Knowledge

Another specific type of metacognitive knowledge is conditional knowledge of strategy use (Schraw and Dennison, 1994). Conditional knowledge pertains to descriptive knowledge about when a particular strategy should be applied. Poor students often are ignorant of potential strategies. For instance, they do not know what goal setting and planning stand for, or when to apply these strategies. Knowing what to do when are the conditions for strategy use. Even when students have obtained adequate conditional knowledge, it does not guarantee the actual execution of a strategy as students may still lack the knowledge of how the strategy should be enacted. In fact, conditional knowledge is an overture to skill acquisition, where a metacognitive strategy has to be consciously applied step-by-step and gradually transformed into a skill through practice (Veenman, 2011). Thus, conditional knowledge is a prerequisite, but not sufficient condition for the acquisition of metacognitive skills.

Metacognitive Skills

Metacognitive skills refer to operational knowledge that is required for the fluent regulation of, and control over one's problem-solving and learning behavior. Orientation, goal setting, planning, monitoring, evaluation, recapitulation, and reflection are manifestations of metacognitive skills that directly affect learning behavior and, consequently, learning outcomes (Veenman, 2011). In metacognitive capable students, skills appear in an orderly sequence of preparatory activities at the onset of a task, systematical execution of activities during task performance, and evaluation and reflection at the end of the task.

At the onset of studying text, skilled students thoroughly read the assignment, pay attention to the title and activate prior knowledge about the topic of the text, scan the text for subheadings and structure, and set a reading goal and devise a reading plan related to the goal (e.g., the goal of understanding relations between main concepts may be effectuated in a plan to make schemas of main concepts and their relations). During reading, they monitor their understanding of words, phrases, and paragraphs. If necessary, they deliberately navigate forth and back through the text in search of information to resolve lack of understanding. At the end of the text, they summarize main conclusions, evaluate these conclusions against their original goal, and may reflect on their reading behavior in order to improve later text studying (Pressley and Afflerbach, 1995). Poor readers, on the other hand, skip the title and immediately start (technically) reading the text from start to end. They may misread words they do not know, without notice. They may bluntly recognize a lack of understanding, however, without doing something about it. At the end of the text, they simply stop reading and conclude that they are 'done.' A similar difference can be found in the problem solving of metacognitive skilled vs poor students (Van der Stel and Veenman, 2010). Good students read the entire assignment first, activate their prior knowledge about the problem content, determine what is given and asked for, make a sketch of the problem, set a clear goal for the outcome, and plan a course of action to reach that goal. They systematically follow their plan, or deliberately change their course of action when arriving at a deadlock position. They repeatedly check the correctness of intermediate results. Before submitting the answering, they evaluate their outcome against their goal and against the assignment. Poor problem solvers, on the other hand, only partially read the problem statement and tend to rush into the application of formulas without knowing what they are heading for. When they get stuck, one formula is simply exchanged for the other. This trial-and-error behavior makes it difficult to trace one's problem-solving steps and, thus, impede monitoring for mistakes and establishing whether progression is being made. When reaching an outcome, they are inclined to submit the answer without further considerations of its correctness or appropriateness. Obviously, many students can be positioned somewhere in between these two extremes. A major issue is how to account for these individual differences in metacognitive skillfulness.

Nelson's Model of Metacognition

Nelson (Nelson and Narens, 1990) distinguished an 'object level' from a 'meta level' in the cognitive system. At the object level lower-order cognitive activity takes place. For instance, lower-order processes needed for reading are decoding, parsing, and lexical access. Higher-order processes of evaluation and planning at the meta level govern the object level. Two flows of information between both levels are postulated. Information about the state of the object level is conveyed to the meta level through monitoring processes, while directions from the meta level are transmitted to the object level through control processes. Thus, if an error occurs on the object level, a metacognitive monitoring process will give notice of it to the meta level and control processes will be activated to resolve the problem. Metaphorically speaking, Nelson's model can be compared to a factory, in which the work floor represents the object level and the manager overlooking the work floor represents the meta level. The manager monitors activity on the work floor to detect errors that may obstruct the production process, in which case the manager's instructions to the work floor restore production. Basically, this is a bottom-up model because metacognitive control is triggered by anomalies in cognitive activity.

Metacognitive Skills as Self-Instructions

Veenman (2011) extended Nelson's bottom-up model with a top-down approach. Metacognitive skills are perceived as an acquired program of self-instructions for the control over and the regulation of cognitive activity at the object level (cf Gollwitzer and Schaal, 1998). This program of self-instructions is activated whenever the learner is faced with a task. Self-instructions can be perceived as condition-action rules that contain conditional knowledge about when to apply a skill and instructions for how to implement the skill on the object level (see **Figure 1**). For instance, activating prior knowledge can be represented as the

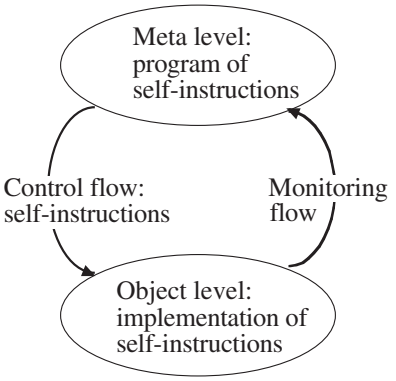

Figure 1 Metacognitive skills as a program of self-instructions. Adapted from Nelson, T.O., Narens, L., 1990. Metamemory: a theoretical framework and new findings. In: Bower, G. (Ed.), The Psychology of Learning and Motivation, vol. 26, pp. 125–173.

rule: IF you have read the task assignment, THEN retrieve all that you know about the topic from memory. In contrast with Nelson's view, self-instructions from the meta level are self-induced, that is, they need not necessarily be triggered by anomalies in task performance. The monitoring flow serves to identify which conditions are satisfied for activating self-instructions. Metacognitive skilled students have an orderly set of self-instructions available that will help them work through the task. The output of an implemented self-instruction will subsequently satisfy the conditions for the next self-instruction. Such a program of self-instructions is acquired through experience and training, much in the same way as cognitive skills are learned (Veenman, 2011). Some students may spontaneously pick up such a program of self-instructions from exemplary models in their environment (parents, teachers, peers), whereas others are devoid of, or not receptive to these models. Furthermore, the developmental trajectory of metacognition may impose limitations to metacognitive skillfulness.

Development of Metacognition

The development of metacognition has its early roots in Theory of Mind (ToM). ToM pertains to young children's knowledge about the mind and, in particular, to knowledge that the content of mental states, such as beliefs, desires, and intentions, may vary from one person to the other. Crucial to the development of ToM is the understanding of a child older than 4 years that another person may or may not know what the child knows. Longitudinal studies have shown that ToM at 4–5 years of age is a precursor of later metacognitive knowledge of metamemory at 5–6 years of age (Lockl and Schneider, 2006). Apparently, the notion that minds are separate entities is prerequisite to the development of metacognitive knowledge.

Development of Metacognitive Knowledge

From the age of 5 years on, children gain knowledge about their memory system. At first, metamemory knowledge pertains to the absolute limitations of their memory, but gradually children become more sensitive to the conditions of memory storage and retrieval. At the age of 6 years, children start to recognize the conditions for effective strategy use. This conditional knowledge becomes more accurate and fine-grained with the years. The adequacy of strategic behavior in young children, however, remains poor due to the immaturity of the prefrontal cortex in the brain (Crone et al., 2006). Young children are overly impulsive and they lack the cognitive capacity for planning their actions.

Development of Metacognitive Skills

At the age of 8 or 9 years, children spontaneously enact elementary planning activities and monitoring for mistakes that catch the eye. In the forthcoming years, planning and monitoring become more fine-tuned, and the metacognitive repertoire is extended with orientation activities, such as goal setting. It takes until 14 years of age before explicit evaluation at the end of the task is regularly displayed. Skills for reflection develop even later in adolescence. These developmental features impose constraints on the training of metacognitive skills. One should not expect the instruction of skills that emerge later in development, such as evaluation or reflection, to be successful in younger children.

Studies have provided evidence for a developmental increment of metacognitive skills. When presented with a series of inductive learning tasks, students in the age of 9, 11, 14, and 22 revealed an impressive growth in metacognitive skillfulness over the years (Veenman et al., 2004). Additional evidence in favor of metacognitive skill development has been obtained in longitudinal studies where students were followed for several years (Alexander et al., 1995; Van der Stel and Veenman, 2010). Despite developmental growth, a huge variation of metacognitive skillfulness within same-age students accounted for individual differences in learning performance (Veenman et al., 2004). The implication is that some students lag behind all along the developmental track.

Domain-Specific vs General Metacognitive Skills

Along the developmental trajectory, metacognitive skills evolve from being domain specific into being general by nature. At the age of 12 or 13, metacognitive skills are domain specific to a certain extent. Based on their study of students of 12 and 14 years of age, Veenman and Spaans (2005) argued that "metacognitive skills may initially develop on separate islands of tasks and domains that are very much alike," such as reading tasks vs problem-solving tasks. At the age of 14, however, metacognitive skills become an integrated, generalized repertoire of self-instructions that is applicable to a variety of tasks and domains, including novel ones (Van der Stel and Veenman, 2010; Veenman and Spaans, 2005; Veenman et al., 2004). This generality of metacognitive skills at the age of 14 has pertinent consequences for instruction and training. Metacognitive skills should be taught across tasks and school disciplines, so as to ensure transfer of skilled performance beyond the borders of a specific task or domain (Salomon and Perkins, 1989).

Assessment of Metacognitive Skills

Before teaching metacognitive skills, one needs to distinguish students with poor metacognitive skills from those with adequate skills. Instruction and training should focus on students with metacognitive deficiencies, rather than interfering with the smooth mastery of metacognitive skills in successful students. Two broad categories of instruments for assessing metacognitive skills pertain to *off-line* instruments, which are administered either prior or retrospective to task performance, and *online* instruments that are administered during actual task performance (Veenman et al., 2006). The essential difference between off-line and online instruments is that off-line methods merely rely on self-reports from the student, whereas online methods concern the coding of student behavior on externally defined criteria.

Off-Line Assessments

Off-line methods refer to questionnaires (e.g., Motivated Strategies for Learning Questionnaire (MSLQ) by Pintrich; Metacognitive Awareness Inventory by Schraw) and interviews. Off-line methods address students with questions about (the frequency of) their metacognitive strategy use and skill application. For instance, an item from the self-regulation scale of the MSLQ (Pintrich and De Groot, 1990) is: "I work on practice exercises and answer end of chapter questions even when I don't have to." Answers are given on a Likert scale, ranging from 1 ('not at all true to me') to 7 ('very true to me'). In interviews, similar questions are reframed as open-ended questions. When administered separate from performance on a target task, questions are formulated in terms of what students would do in general, or indicated as such during test instructions. The formulation of retrospective questions, on the other hand, refers to the specific task that was just completed.

The advantage of using questionnaires is that they can be easily administered to large groups of respondents, which obviously is the main reason for the omnipresence of questionnaires in metacognition research (Dinsmore et al., 2008; Veenman, 2005). Interviews have the disadvantage that they have to be individually administered and processed.

All off-line methods rely on self-reports by students, which entail major validity problems (Veenman, 2005, 2011). Students have to reconstruct their earlier metacognitive behavior from memory that makes their self-reports vulnerable to memory failure and distortions. Reconstructive interpretations may be elicited along with, or instead of correct recollections after task performance. Memory accuracy may be further distorted when off-line assessments are administered separate from actual task performance as students have to rely on their recollections of earlier learning experiences, further down the past. Another drawback of off-line methods is that questioning may interfere with the spontaneous self-report of metacognitive skills by students. Questions may prompt the recall of metacognitive behavior that never occurred. Not only may students be inclined to give socially desirable answers, they may also be triggered by questions to label their behavior accordingly. Students may take, for instance, their paraphrasing for summarizing.

Indeed, studies with a multimethod design have shown that different off-line measures often correlate poorly with one another, although they intend to assess the same metacognitive skills (Veenman, 2005). Furthermore, off-line measures hardly correlate with online measures obtained during task performance. Apparently, students do not actually enact what they have said beforehand, nor do they accurately recollect what they have recently done. In conclusion, off-line self-reports do not accurately reflect students' metacognitive skills because students cannot objectively review their own metacognitive behavior.

Online Assessments

Two frequently used online assessment methods are the analysis of think-aloud protocols and systematical observation. When thinking aloud, students are asked to merely verbalize their ongoing thoughts during task performance. Whenever they fall silent, students are encouraged to 'keep on thinking aloud' by the experimenter. Students are not incited to interpret their behavior, contrary to introspection. In fact, thinking aloud captures the content of working memory during task performance. Think-aloud protocols are recorded on tape and transcribed verbatim. In systematical observation, the student's task performance is video taped. Thus, observation protocols contain both behavioral information, as well as verbal utterings of students. Both think-aloud and observation protocols are rated by 'blind' judges according to specified criteria in a codebook. The application of such a standardized coding system by external agencies circumvents the validity problems mentioned before (Veenman, 2011). Moreover, the use of multiple judges allows for establishing an interrater reliability, indicating the rigor of a coding system.

Both methods of thinking aloud and systematical observation are time consuming. They need to be individually administered to students and the judgmental processing of data is labor intensive. Moreover, thinking aloud may not be appropriate to very young students with limited verbal capacities, or to tasks that are extremely difficult or demanding. Think-aloud protocols are then likely to be incomplete, which is known as the 'tip-of-the-iceberg' phenomenon (Veenman, 2011).

More recently, computerized assessments of metacognitive skills have become available (Veenman, 2013b; Veenman et al., 2014; Winne, 2010). A learning task is presented on a computer and all student activities in this computerized environment, such as button presses, scrolling, and electronic notes, are registered in a logfile on the computer. Frequencies and patterns of activities may be automatically computed and recorded as well. Logfile measures, however, represent activities of students on the cognitive object level without the metacognitive considerations provided by students when thinking aloud. A logfile indicates that a student scrolled back in an electronic text or pushed a particular button, but it does not disclose the reasons for why the student decided to do so. In order to determine whether logged activities are the consequence of control activities from the meta level, computerized assessments need to be initially validated against other online methods (Veenman, 2013b). Thus, computerized assessments should be able to distinguish between, for instance, button presses that occur at random and those deliberately chosen to pursue a particular metacognitive

goal. Once validated, computerized assessments can be administered to students at any time or place in a flexible way, or even to large groups of students at the same time (Veenman et al., 2014).

Studies with a multimethod design reveal substantial correlations among online measurements, indicating that different online methods converge in their assessments of metacognitive skills (Veenman, 2005). Moreover, online measures appear to be strong predictors of learning performance, in contrast with offline measures. These are compelling arguments for resorting to online methods when assessing metacognitive skills.

Instruction and Training of Metacognitive Skills

Students with learning disabilities or limited intellectual resources may suffer from metacognitive deficiencies (Pressley and Gaskins, 2006). However, many regular secondary school students and even intellectually gifted students may equally exhibit poor metacognitive behavior. Metacognitive skills can be successfully instructed and trained in all students (Dignath and Büttner, 2008). Like the acquisition of any other skill, however, it will take students' time and effort to arrange self-instructions in an organized set in order to attain a proficient level of skill application.

Three Principles for Instruction and Training

There are three principles for effective instruction and training of metacognitive skills (Veenman, 2013a). Any metacognitive instruction should abide by these three principles for successfully establishing a repertoire of metacognitive self-instructions in students.

Embedded Instruction

According to the first principle, metacognitive instruction should be embedded in a concrete learning context. Metacognitive skills cannot be executed in a 'void,' as they have to be operated upon cognitive processes at the object level. Embedded instruction enables the student to associate the conditional knowledge of *which* metacognitive skill is required *when* during task performance with the specific characteristics of the task. Metaphorically speaking, a learning context gives hands and feet to the metacognitive instruction of what to do when. Studies that failed to embed metacognitive instruction in a learning context invariably lacked training effects. In the same vein, study-skill lessons on school are fruitless if they are isolated from teaching content (Veenman, 2013a).

Informed Instruction

The second principle insists that students should be informed about the usefulness and benefits of metacognitive skills. When students are deficient in utilizing metacognitive skills, the execution of instructed skills initially requires extra effort and occupies working-memory space. This may result in a temporary cognitive overload of working memory, especially when the task at hand is demanding. In fact, task performance may be temporarily impeded by the effort expenditure needed for skill acquisition. Consequently, students may be inclined to abandon the instructed skills, unless they appreciate *why* the application of metacognitive skills facilitates their task performance in the end (Veenman, 2013a).

Prolonged Training

Following the third principle, metacognitive training should be spread over an extended period of time in order to allow for smooth and maintained application of metacognitive skills. When complex skills need to be built up from scratch, extensive modeling and practice of conditional and operational knowledge is required. When students are about to master a strategy, its execution still depends on conscious effort. Only after sufficient practice, the strategy is converted into a skill, that is, it becomes fluent and less demanding in its application (Veenman, 2013a). Duration of training, however, is not fixed. The time needed for metacognitive instruction and training depends on the quantity and complexity of skills that need to be acquired.

WWW&H Rule

The *WWW&H* rule (what to do, when, why, and how) captures the content of complete instruction of metacognitive skills in a given task context (Veenman et al., 2006). The four keywords of *what*, *when*, *why*, and *how* serve as the landmarks of adequate metacognitive functioning. Explicit instruction of *WWW&H* is crucial to students for linking task settings, which provide the conditions for action, to potential metacognitive actions in the initial phase of skill acquisition (Gollwitzer and Schaal, 1998; Veenman, 2011). For instance, when instructing students to apply comprehension monitoring during text studying, students should be explicitly taught what comprehension monitoring is ('checking your understanding of words and phrases'), when monitoring should be applied ('in case of misreading, or after each paragraph'), and why monitoring is necessary ('lack of understanding corrupts further reading'). Through modeling and scaffolding, the teacher facilitates the acquisition of monitoring processes in students. The teacher first demonstrates the *WWW&H* rule to the student while reading aloud a particular text. Next, the student enacts the monitoring skill, while receiving feedback from the teacher. Through further practice, the student should accomplish fluent mastery of the *WWW&H* rule.

When teaching multiple metacognitive skills, often step-by-step action plans are used that prescribe the order of metacognitive activities (Veenman, 2013a). Each step describes a concrete action, rather than an abstract skill. For instance, when solving science problems, orientation skills are represented by subsequent steps for reading the entire problem statement, discerning relevant from irrelevant information, making a sketch of the problem situation, determining what is given and what is asked for, and estimating the outcome. Such step-by-step action plans help students to integrate conditional knowledge with the actual enactment of skills in order to build up a repertoire of metacognitive self-instructions.

Metacognitive Training: An Example

Pressley and Gaskins (2006) designed a particularly successful training program at a special school for students with very low reading abilities. Throughout each school day, all teachers across disciplines incessantly addressed the students with metacognitive reading instructions, such as establishing a goal for reading, identifying the theme and main ideas in a text, relating new information to prior knowledge, monitoring understanding through self-questioning, dealing with incomprehension by rereading or looking for additional information sources, and summarizing the main conclusions. Instruction explicitly addressed when, why, and how strategies were to be used. After spending several years at this special school, most students returned to regular education with above-average scores on a national reading test. Similarly effective programs with step-by-step action plans are the IMPROVE program for solving math problems (Kramarski and Mevarech, 2003), or Systematical Approach for solving physics problems (Mettes et al., 1981).

Availability vs Production Deficiencies

Even though students may exhibit equally poor metacognitive behavior, they may differ in what causes their poor behavior. Students may suffer from either an availability deficiency or a production deficiency (Veenman et al., 2000). Students with an availability deficiency do not have metacognitive skills at their disposal, whereas students with a production deficiency are obstructed in the execution of available skills. Both deficiencies require differential approaches in the instruction of metacognitive skills.

Availability Deficiency

Students with an availability deficiency lack the essential metacognitive knowledge, the appropriate skills, or both as a rule. Take planning as an example. Students may not be aware of what planning is, when planning is required, or why planning is worthwhile. Even when this conditional knowledge is available, still students may not know how to do planning. For instance, they may not be able to transform an abstract goal into concrete subgoals with an orderly sequence of activities. Merely urging students to plan their work, as teachers do from time to time, does not help students to surmount an availability deficiency. Students, who are devoid of metacognitive skills, need to be instructed and trained from scratch, that is, fully instructed and modeled according to the three instructional principles and the *WWW&H* rule.

Production Deficiency

Students with a production deficiency are capable of skillful behavior under familiar conditions, but for some reason they do not spontaneously execute the available skills in a particular learning situation. For instance, they do not know when to plan or monitor their actions in the context of a specific task, they do not recognize the relevance of those skills for that particular task, or test anxiety may interfere with the deployment of skills. These students need not be fully trained in *how* to apply metacognitive skills. They merely should be reminded of applying the necessary skills by providing them with metacognitive cues during task performance (e.g., 'now is the time to draw up a plan'). A remedial teacher could deliver these cues or students could autonomously handle a set of cards on which the subsequent steps of an action plan are printed (Veenman, 2013a). Computerized learning environments could also provide individualized prompts that respond to (omitted) student actions on the computer (Veenman, 2013b; Winne, 2010). A key feature of cuing or prompting is that skills are not further instructed or modeled. Extended instruction and training could even interfere with common mastery by students with a production deficiency.

So far, the only valid method for distinguishing students with availability vs production deficiencies is to provide students with metacognitive cues during task performance (Veenman et al., 2000). Students are evidently suffering from a production deficiency when metacognitive cuing results in enhanced metacognitive behavior and learning performance. Additional cuing should suffice for these students, until they prevail over their production deficiency. Students who do not respond to cuing, on the other hand, are *assumed* to suffer from an availability deficiency. Most of them actually are available deficient, but others may have been insusceptible to the metacognitive cues given. When submitting these alleged available-deficient students to ample metacognitive instruction and training, instructors should be aware of this limitation. One student may pick up metacognitive instructions more rapidly than others, which requires a flexible approach by the instructor.

Conclusion

Effective metacognition is characterized by the application of metacognitive conditional knowledge, by knowing how to execute metacognitive skills, and by understanding the benefits of using those skills. Students should incorporate these principles of adequate metacognitive functioning in an organized set of self-instructions. Not all students, however, spontaneously develop such a general repertoire of metacognitive skills. Depending on whether conditional or operational knowledge is missing, metacognitive-deficient students should either be cued, or fully trained to enact metacognitive skills according to the three instructional principles and the *WWW&H* rule. Eventually, students should master a repertoire of metacognitive self-instructions that is broadly applicable to any new learning situation.

See also: Cognitive Development During Infancy and Early Childhood across Cultures; Cognitive Development, Educational Theories of; Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Metacognitive Development: Educational Implications; Mind, Theories of; Theory of Mind During Infancy and Early Childhood Across Cultures, Development of.

Bibliography

Alexander, J.M., Carr, M., Schwanenflugel, P.J., 1995. Development of metacognition in gifted children: directions for future research. Developmental Review 15, 1–37.

Brown, A.L., DeLoache, J.S., 1978. Skills, plans, and self-regulation. In: Siegel, R.S. (Ed.), Children's Thinking: What Develops? Erlbaum, Hillsdale, pp. 3–35.

Crone, E.A., Donohue, S.E., Honomichl, R., Wendelken, C., Bunge, S.A., 2006. Brain regions mediating flexible rule use during development. Journal of Neuroscience 26, 11239–11247.

Dignath, C., Büttner, G., 2008. Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. Metacognition and Learning 3, 231–264.

Dinsmore, D.L., Alexander, P.A., Loughlin, S.M., 2008. Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. Educational Psychology Review 20, 391–409.

Flavell, J.H., 1979. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. American Psychologist 34, 906–911.

Flavell, J.H., Wellman, H.M., 1977. Metamemory. In: Kail, R.V., Hagen, J.W. (Eds.), Perspectives on the Development of Memory and Cognition. Erlbaum, Hillsdale, pp. 3–33.

Gollwitzer, P.M., Schaal, B., 1998. Metacognition in action: the importance of implementation intentions. Personality and Social Psychology Review 2, 124–136.

Kramarski, B., Mevarech, Z.R., 2003. Enhancing mathematical reasoning in the classroom: the effects of cooperative learning and metacognitive training. American Educational Research Journal 40, 281–310.

Lockl, K., Schneider, W., 2006. Precursors of metamemory in young children: the role of theory of mind and metacognitive vocabulary. Metacognition and Learning 1, 15–31.

Mettes, C.T.C.W., Pilot, A., Roossink, H.J., 1981. Linking factual and procedural knowledge in solving science problems: a case study in a thermodynamics course. Instructional Science 10, 333–361.

Nelson, T.O., Narens, L., 1990. Metamemory: a theoretical framework and new findings. In: Bower, G. (Ed.), The Psychology of Learning and Motivation, vol. 26, pp. 125–173.

Pintrich, P.R., De Groot, E.V., 1990. Motivational and self-regulated leaning components of classroom academic performance. Journal of Educational Psychology 82, 33–40.

Pressley, M., Afflerbach, P., 1995. Verbal Protocols of Reading: The Nature of Constructively Responsive Reading. Erlbaum, Hillsdale.

Pressley, M., Gaskins, I., 2006. Metacognitive competent reading is constructively responsive reading: how can such reading be developed in students? Metacognition and Learning 1, 99–113.

Salomon, G., Perkins, D.N., 1989. Rocky roads to transfer: rethinking mechanisms of a neglected phenomenon. Educational Psychologist 24, 113–142.

Schraw, G., Dennison, R.S., 1994. Assessing metacognitive awareness. Contemporary Educational Psychology 19, 460–475.

Van der Stel, M., Veenman, M.V.J., 2010. Development of metacognitive skillfulness: a longitudinal study. Learning and Individual Differences 20, 220–224.

Veenman, M.V.J., 2005. The assessment of metacognitive skills: what can be learned from multi-method designs? In: Artelt, C., Moschner, B. (Eds.), Lernstrategien und Metakognition: Implikationen für Forschung und Praxis. Waxmann, Berlin, pp. 75–97.

Veenman, M.V.J., 2008. Giftedness: predicting the speed of expertise acquisition by intellectual ability and metacognitive skillfulness of novices. In: Shaughnessy, M.F., Veenman, M.V.J., Kleyn-Kennedy, C. (Eds.), Meta-cognition: A recent review of research, theory, and perspectives. Nova Science Publishers, Haupage, pp. 207–220.

Veenman, M.V.J., 2011. Learning to self-monitor and self-regulate. In: Mayer, R., Alexander, P. (Eds.), Handbook of Research on Learning and Instruction. Routledge, New York, pp. 197–218.

Veenman, M.V.J., 2013a. Training metacognitive skills in students with availability and production deficiencies. In: Bembenutty, H., Cleary, T., Kitsantas, A. (Eds.), Applications of Self-regulated Learning across Diverse Disciplines: A Tribute to Barry J. Zimmerman. Information Age Publishing, Charlotte, pp. 299–324.

Veenman, M.V.J., 2013b. Assessing metacognitive skills in computerized learning environments. In: Azevedo, R., Aleven, V. (Eds.), International Handbook of Metacognition and Learning Technologies. Springer, New York, pp. 157–168.

Veenman, M.V.J., Bavelaar, L., De Wolf, L., Van Haaren, M.P.G., 2014. The on-line assessment of metacognitive skills in a computerized environment. Learning and Individual Differences 29, 123–130.

Veenman, M.V.J., Kerseboom, L., Imthorn, C., 2000. Test anxiety and metacognitive skillfulness: availability versus production deficiencies. Anxiety, Stress, and Coping 13, 391–412.

Veenman, M.V.J., Spaans, M.A., 2005. Relation between intellectual and metacognitive skills: age and task differences. Learning and Individual Differences 15, 159–176.

Veenman, M.V.J., Van Hout-Wolters, B.H.A.M., Afflerbach, P., 2006. Metacognition and learning: conceptual and methodological considerations. Metacognition and Learning 1, 3–14.

Veenman, M.V.J., Wilhelm, P., Beishuizen, J.J., 2004. The relation between intellectual and metacognitive skills from a developmental perspective. Learning and Instruction 14, 89–109.

Wang, M.C., Haertel, G.D., Walberg, H.J., 1990. What influences learning? A content analysis of review literature. Journal of Educational Research 84, 30–43.

Winne, P.H., 2010. Improving measurements of self-regulated learning. Educational Psychologist 45, 267–276.

Teamwork and Team Performance Measurement

Christopher W Wiese, Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA
Marissa L Shuffler, College of Business & Behavioral Science, Clemson University, Clemson, SC, USA
Eduardo Salas, Institute for Simulation & Training, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Teams are essential in order to handle the complex problems faced by the contemporary organization. Consequently, a deeper understanding of what makes a team effective has been an ongoing goal of researchers. However, our understanding of effective teams is capped by the quality of measurement used to assess them. This article first provides theoretical basis for understanding the measurement of teams. Based on this information, we further offer guidelines and recommendations for those seeking measure team dynamics.

Introduction

The ever-increasing demands of the modern day organization are, at times, too much for an individual to handle. Contemporary technologies have increased the speed, the amount, and the variety of information organizations are expected to process. A consequence of this growth has been the increased use of teams as an appropriate means for handling complex problems and issues above and beyond individual capacities. As the research on teams have accumulated over the past 30 years – and the use of teams in organizational settings show no signs of slowing down – this article aims to fortify the assessment of teams in organizational settings by focusing on the measurement aspect of teamwork. Specifically, we highlight the importance of having a complete team measurement system, as a common issue arising during team assessment is the deficient measurement of multiple team components. A complete team measurement system gauges the individual characteristics of the team (e.g., knowledge, skills, abilities, and other characteristics), both taskwork and teamwork processes/emergent states (e.g., communication, collaboration, cohesion, conflict), and team outcomes. This article will briefly review a representative sample of what is known about each of these team measurement system components. First, we define teams and elaborate on these three components. Next, we identify six considerations for those developing/selecting team measurement systems. Third, we feature a representative sample of methods and tools that can be used in such systems. Finally, we end with a few remarks about the next frontier in teamwork and its measurement.

Overview of Team Measurement Systems

For our purposes, we define teams as a distinguishable set of two or more individuals who must perform distinct, complementary, or interdependent tasks en route to accomplishing a shared, common goal (Salas et al., 1992). This particular definition lends itself well to the discussion of team measurement systems. In the broadest of terms, a system of team measurement could reflect any battery of techniques used to observe and measure teams. However, a team measurement system is not complete unless it captures the individual characteristics of teams, team processes and emergent states, and team outcomes. One way to think of such systems is through the model displayed in **Figure 1**. This figure depicts two different levels of specification (i.e., individual level and team level) and three different components (i.e., inputs, mediators, and outcomes). In order to have a complete system, it is necessary to measure components in all six sections of the model. Though the measurement of individual-level factors is important, this article will focus on the measurement of team-level factors. In the following section, we will briefly define and discuss each of these team-level components.

Inputs

A team is made up of a collection of individuals whereby each individual has certain knowledge, skills, abilities, and other characteristics that have the potential to affect the work they do while on that team (Barrick et al., 1998). Research has indicated the importance of different types of knowledge (e.g., task knowledge), skills (e.g., technical skills), abilities (e.g., cognitive ability), and other (e.g., personality) characteristics in terms of their impact on performance at an individual level (Bunderson and Sutcliffe, 2003; Carson et al., 2007; Driskell et al., 2006; Earley and Mosakwoski, 2000). However, it has been suggested that the KSAOs related to team effectiveness may differ than those that are effective at the individual level (Morgeson et al., 2005). For instance, Stevens and Campion, (1994) suggest that rather than focusing on more individual-level KSAOs (e.g., technical skills, personality traits), there are specific interpersonal KSAs (e.g., conflict resolution, collaborative problem solving, and communication KSAs) and self-management KSAs (e.g., goal setting and performance management and planning/task coordination KSAs) that are more applicable to understanding teamwork. Other critical team-level input factors include, but are not limited to, social skills, teamwork knowledge, expertise diversity, task interdependence, and leadership skills (Morgeson et al., 2005; Chen et al., 2004; Burke et al., 2006; Mathieu et al., 2008).

Mediators

Mediators of the input–outcome relationship in regards to teamwork refer to those processes and other variables that serve

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22017-5

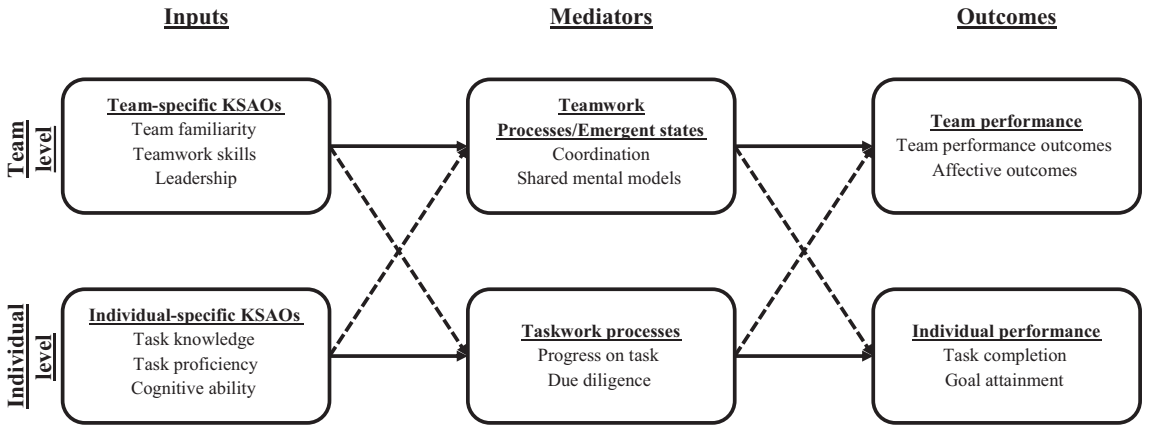

Figure 1 Multi-level team measurement system model. Solid arrows represent a stronger association between components. Dashed lines represent a weaker association between components. Influenced by McGrath, J.E. 1964. Social psychology: A brief introduction, New York, Holt, Rinehart, & Winston; Ilgen, D.R., Hollenbeck, J.R., Johnson, M., Jundt, D., 2005. Teams in organizations: from Input-Process-Output models to IMOI models, Annual Review of Psychology 56, 517–543; and Mathieu, J., Maynard, M., Rapp, T., Gilson, L., 2008. Team effectiveness 1997–2007: a review of recent advancements and a glimpse into the future. Journal of Management 34 (3), 410–476.

as the mechanisms through which individual and team level inputs influence team outcomes (Ilgen et al., 2005). When selecting important team-level mediators, there are two important distinctions to make. First is the difference between taskwork and teamwork. Taskwork is comprised of the actions that team members perform, which are not dependent on any of their fellow team members, whereas teamwork is comprised of those actions made necessary by interdependence and are used to facilitate coordination (Salas et al., 2008). From a measurement standpoint, one could consider actions that represent taskwork to be specific to the individual (e.g., behaviors an individual engages in to complete the team's goal), whereas teamwork needs to be measured at the team level (e.g., interactions between team members). Thus, the focus of team-level measurement is on teamwork mediators.

The second distinction regarding mediators relates to the different types of teamwork. Teamwork can be separated into two different classifications: processes and emergent states. Teamwork processes are the observable behaviors that occur when team members interact with one another, whereas emergent states are cognitive, motivational, and affective states of the team (Marks et al., 2001). Research has demonstrated that both team processes (e.g., strategy formulation and planning, mission analysis, affect management) and emergent states (e.g., team mental models, team cohesion, team empowerment) are crucial predictors of team performance (Edwards et al., 2006; Lepine et al., 2008; Smith-Jentsch et al., 2005).

Outcomes

Team outcomes can be separated into two distinct sets: performance and affective outcomes (Hackman and Morris, 1975). Team performance outcomes are typically denoted by the assessment of the team's accomplishment of assigned goals. The measurement of these outcomes can range from a simple checklist of predefined goals the team was assigned to accomplish to a supervisor's assessment of a team's accuracy and quality of work performed. The assessment of team performance outcomes typically is directly related to the accomplishment of task/team goals. Conversely, affective outcomes target how the team feels regarding their teamwork experience. Some prominent affective outcomes include the team's willingness to work together in the future, team satisfaction, and team member trust. Though some may consider affective outcomes are less important than performance outcomes, they may have critical implications for teams that plan to perform together in the future.

As mentioned earlier, this section served to briefly review some of the more prominent team-level factors crucial in the selection or development of any team measurement system. There are several other factors that are not represented here, but still may be needed in a system depending on the needs of the organization and context of the team. However, the selection of constructs is not the only consideration one should make in developing a team measurement system. In fact, the selection of these constructs is largely influenced by the purpose driving the need for measurement. The next section covers the significant considerations that we believe are necessary in the development or selection of any team measurement system.

Team Measurement System Selection/Development Considerations

Measurement is paramount for the understanding of human behaviors (Meister, 1985). This argument can be extrapolated to all levels of human behavior, an individual as well as a team. Therefore, as the development and use of a team measurement system is vital to understanding intricacies of an effective team, some basic assumptions should be covered regarding the development of such systems (Table 1). These assumptions form the basis for the design, development, and testing of team measurement systems. Understanding these underlying assumptions leads to the recognition that there are certain requirements that underlie effective measurement. This section outlines six different considerations that organizations and practitioners should contemplate when developing/selecting a system of measurement for teams. While developing/selecting a team measurement system, it is vital to recognize that

Table 1 Representative sample of basic team measurement assumptions

Basic measurement assumptions
1. The construction of team measures must be based on team-based theoretical models.
2. Team performance and processes are reflected through the actions and observable outcomes of individual team members
3. The goal of measurement is to allow observations of behavior and changes in behavior of team members, which represent underlying processes.
4. Certain team processes and outcomes are more readily available for measurement than others.
5. Processes not easily measured must be matched to corresponding observable behavior or quantified through feedback from team members.
6. Individual performance of team members can be considered within the context of the team and used, in combination with others factors, to assess team performance.

Based on Meister, D., 1985. Behavioral Analysis and Measurement Methods, Wiley-Interscience; Brannick, M.T., Salas, E., Prince, C.W., 1997. Team performance assessment and measurement: theory, methods, and applications, Psychology Press.

these considerations are not independent of one another and therefore should be deliberated in their entirety.

Measurement Purpose

The first consideration facing those who are developing or selecting a team measurement system is to specify the purpose of the measurement. Team measurement systems can be used to serve many purposes, such as team member selection, training feedback, performance evaluation, and promotion. The selection of the purpose will guide measurement content selection, decisions regarding the type of data needed, and appropriate data collection resources (Mesiter, 1985). For instance, if the purpose of the measurement is simply to gauge team members' reactions to training program, a paper-pencil measure specifically targeting their feelings about the training may suffice. However, if the purpose of the measurement is to provide feedback and develop a flight crew, a more labor-intensive, complex system may be necessary. In this latter instance, critical team process constructs would need to be specified, measured, and tied to specific team outcome variables so that detailed feedback could be given regarding the team's performance. Therefore, specifying the purpose of the measure is a critical first step in any team measurement system.

Construct Selection

Once the purpose of the team measurement system is specified, the next step involves selecting the constructs of interest. It is important for developers to measure all three factors previously discussed: team-level inputs, mediators, and outcomes. Incorporating all three elements into a team level system of measurement will provide the most holistic understanding of what factors may be impacting team performance, and how such factors may be interacting with one another. If the decision is made to only select a portion of these factors, the team measurement system may be deficient and unable to accomplish the overall purpose. For instance, only measuring team performance outcomes in a training program would not be conducive to providing feedback to team members, a critical aspect to effective training. In order to provide feedback, specific team processes must be measured. By giving feedback on specific team processes, team members will be able to learn specifically how they were performing improperly. Additionally, the selection of constructs identified as inputs, mediators, and outcomes will also impact when these constructs should be measured, as will further be elaborated upon later in this article.

Measure Referent

When measuring team-level variables, it is imperative that those developing measures understand the implications associated with specifying the focal point or referent of the measure. The focal point or referent of the measure refers to whom or what the measure is attempting to obtain information about; for teams, this could be regarding the entire team as a whole, or it could refer to the individuals of the team (Chan, 1998). This consideration encapsulates several questions: Do external raters need to observe the construct or can the construct be assessed through paper–pencil measures? Are the items on the measure specifically referring to the team instead of the individual? If the items on a measure are referring to the individual, can they simply be aggregated to a team level or is a more complex aggregation technique needed? Is the targeted construct observable? Does the task provide the opportunity for the targeted-construct to be performed? Such questions are critical to ask in order to ensure that the appropriate focal point is utilized for each of the constructs to be measured.

The relevance of these questions is partially determined by the constructs selected. For instance, when assessing team cohesion, it may be necessary to utilize survey items that specifically use the team as referent. Given that team cohesion is an emergent state, which can be very challenging to observe by outsiders (Mathieu et al., 2008), having team members report on their feelings with the team as the referent should provide the most appropriate measure of the construct. Alternatively, if one decides to measure teamwork processes, it may behoove the measurement system designer to choose a situation in which team members would have the opportunity to perform the behaviors associated with the construct prior to measurement in order to ensure that those behaviors have a chance to emerge at the team level. Much like the previous considerations, the decisions made here greatly influences the choices available in the forthcoming steps.

Measurement Source

There are five primary measurement sources typically used in team measurement systems: team members, supervisors, trained raters, customers, and objective sources. Not all of these sources are equally effective in assessing all different factors (i.e., inputs, mediators, and outcomes). For example, the measurement of input factors is more accurate when coming from the team members themselves. Supervisors may not be as accurate at assessing these input factors as they may not be aware of each team members' expertise. A summary of the specific advantages and disadvantages for each source is

Table 2 Pros and cons of measurement sources

Measurement source	*Pros*	*Cons*
Team members	● Observed team members on a daily basis ● Good understanding of what should be done ● Can access their own cognitions	● May be too lenient on team members if team rewards are tied to team behavior ● May be too severe on ratings if the climate is very competitive
Supervisors	● Expert understanding of what constitutes good behavior ● Witness to best behaviors	● May not be around on a daily basis ● Cannot assess team cognitions
Trained raters	● Trained understanding of the behaviors being assessed ● May be experienced in doing ratings	● Not privy to day-to-day operations ● Not familiar with organization/team culture
Customers	● Gives a different perspective compared to the aforementioned sources. ● Ratings come form experience with the team	● Ratings are very heavily influenced by experience ● Does not have an understanding of what constitutes good performance.
Objective sources	● Removes a great deal of bias from team measures. ● Can be concretely defined by the researcher	● Can only measure what it is programmed to measure ● Can typically only measure simple things (e.g., frequency, counts)

provided in Table 2. One must be selective in choosing the source of the measure. As some sources are better at assessing certain factors than others, the inappropriate selection of measurement source could result in inaccurate/biased measurement, which would consequentially lead to improper assumptions made about the team. Thus, it is highly recommended to use multiple sources in collecting team data. Some have even argued that it takes a team to effectively measure a team (Dickinson and McIntyre, 1997).

Measurement Timing

The fact that teams develop and evolve over time has implications on when constructs can and should be measured (Kozlowski and Klein, 2000). Certain constructs may be more appropriate to measure during certain points in a team's performance cycle than others. For instance, teamwork processes targeted at setting up teams for successful performance, such as mission analysis, goal specification, and strategy formulation and planning, may be more observable during the initial stages of team performance. Additionally, the choice of source may have implications on when a construct is measured. For example, externally trained raters can only make judgments regarding targeted constructs while the team is performing. Other team factors should only be measured at the end of team performance (e.g., objective performance, affective outcomes). Additionally, the timing of when constructs are measured may also be influenced by environmental characteristics that are not under the control of the team measurement system designer. If, for example, one of the conditions set forth by a union is that team member can only fill out research surveys when they are clocked in, the developer must try and create a measure that team members can fill out without disrupting the constructs they are trying to assess.

Measure Validation

The last consideration concerns the validation of the measures within a team measurement system. Depending on whether one is selecting or developing the measures in a system, these considerations will change slightly. However, the assurance that these measures have undergone proper validation processes is one of the most crucial steps in the development/selection of a team measurement system. Simply put, validity concerns whether the measure actually captures what it is supposed to measure (Messick, 1980). Though validity can never be absolutely proven, evidence can be provided supporting a measure's validity. There are three main types of validity used to support a measure's credibility: content validity, construct validity, and criterion-related validity. Each of these provides different types of evidence of validity and, consequently, one should look to support all three validity types.

Content validity typically refers to the extent that a measure represents a construct in its entirety. One way to provide evidence for content validity is by using subject matter experts (SMEs) to review a measure for any construct deficiency or contamination. Support for a measure's content validity does not necessarily have to come through an empirical process. That is, support for content validity can be established through a rationalized process of judgments rather than a quantitative investigation. Criterion-related validity simply concerns the degree to which a measure predicts an outcome of interest. Evidence for this type of validity is established when a measure correlates with the criterion. That is, criterion-related validity evidence is typically represented in the form of a correlation coefficient between a measure (predictor) and a criterion. Lastly, construct validity concerns whether or not a measure quantitatively represents the construct that it is designed to represent. There are several ways one can establish construct validity such as creating a nomological network or creating a multi-trait multi-method matrix (Cronbach and Meehl, 1955; Campbell and Fiske, 1959). Each type of validity is critical to the credibility of a team measurement system. As such, providing evidence for each type of validity is a fundamental consideration.

Available Measurement Tools

During the creation of a team measurement system, it is not necessary to 'reinvent the wheel,' or start completely from scratch in terms of developing tools. There are several tools/methods available to assess a team's inputs, mediators, and outcomes. In this section, we cover a representative sample of some of these tools/methods used to measure teams including rating scales, behaviorally anchored rating scales (BARS), event-based measurement scales, automated performance measurement, concept mapping, pathfinder, and probed protocol

analysis. There are numerous cases in recent literature that demonstrate the applicability of these tools for the collection of team data (Jones and Harrison, 1996; Lynn and Reilly, 2000). Moreover, though many of these tools can be used as a combination of inputs, mediators, and outputs, some are more apt at assessing certain constructs than others.

Rating Scales

One of the most common forms of assessment of team constructs is rating scales. Rating scales ask people to indicate their attitudes, opinions, beliefs, or feelings on a large number of items. Thus, rating scales can be used to assess reflections of team behaviors, opinions concerning how team members feel, or even beliefs concerning team member's cognitions. There are a number of ways to develop rating scales. If one desires to gauge opinions concerning team members' behavior, the first step would be to compile a description of the behaviors of interest. One approach to identifying and accumulating these behaviors would be to use a critical incident technique. A critical incident technique gathers information from SMEs regarding memorable events, both positive and negative, which occurred in a specific situation (Chell, 1998). These behaviors would then serve as the content for the questions being asked on the measure.

Once the content has been identified, there are several options on how to present the information. A few possibilities include: (1) scale points can simply be reflected by numbers, (2) descriptive words can be used at end scale points, or (3) verbal descriptions can be used at each scale point (Meister, 1985). Another choice comes with deciding how many scale points to be used. First, there must be enough scale points to fully represent the question being asked as a dichotomous scale might not represent the different degrees of the attribute being assessed. Another consideration is whether to use an odd number of scale points, which provides the rater with the option of choosing a neutral response. These options have their respective pros and cons, and it is completely up to the measure designer which option they prefer. An example of a ratings scale is provided in **Figure 2**.

Behaviorally Anchored Ratings Scales

BARS are a type of rating scale first introduced by Smith and Kendall (1963). Much like rating scales, BARS use verbal descriptors at each scale point. However, unlike rating scales, BARS provide very specific behavioral descriptors at each scale point. These anchors provide more concrete examples of what behaviors represent each scale point. That is, rather than one word representing each scale point (e.g., agree, never, always, disagree), BARS scale points has specific behavioral descriptions. Consequently, this approach lends itself well to the critical incident technique mentioned above. Instead of the behaviors becoming part of the item text (as alluded to above), they are transposed onto the scale points. Furthermore, BARS are especially suitable for assessing team behaviors. One distinct advantage of BARS over traditional rating scales is that it helps control for any individual idiosyncrasies of the raters (Salas et al., 2003). An example is provided in **Figure 3**.

Event-Based Measurement Scales

Another behaviorally oriented approach to team measure is the event-based measurement scales. Using this approach, the measure designer first identifies the team competencies they wish to measure. Following this, they insert events into behavioral scenarios designed to cue the specific behavioral competencies. In doing so, practitioners can tie measurement directly back to the targeted competences. Furthermore, this approach does provide some ease for raters. That is, since the raters know what cues are embedded into the scenario and when these cues take place, they are better able to focus on a certain behavioral competency when the time arises. By creating a situation in which team behavioral ratings are consistently drawn from the same event across teams, event-based measurement scales typically contain exceptional psychometric properties. An example of this type of measurement approach is provided in **Figure 4**.

Automated Performance Measurement

The type of information that automated performance measurement tools can collect is tremendously diverse. Automated performance measurement tools have the ability to collect data as simple as automatically counting the number of times team members say a certain word to data gathered from a highly complex jet flight simulator. One of the advantages of automated performance measurement tools is that they, to a certain degree, remove the errors associated with human decision making. Further, tools like this are not as intrusive as other aforementioned approaches. However, there are several downsides of using an automated performance measurement tool. First, they can be extremely expensive to design and maintain. Next, automated performance measures can only measure what it was programmed to measure. If you want to measure something else, you might not be able to get this information from data that you have already collected. Another

Question: How often did the team members offer information on other team members' performances?

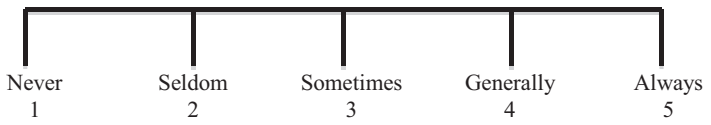

Figure 2 Example of a rating scale measuring feedback.

Question: How effective were team members in providing feedback?

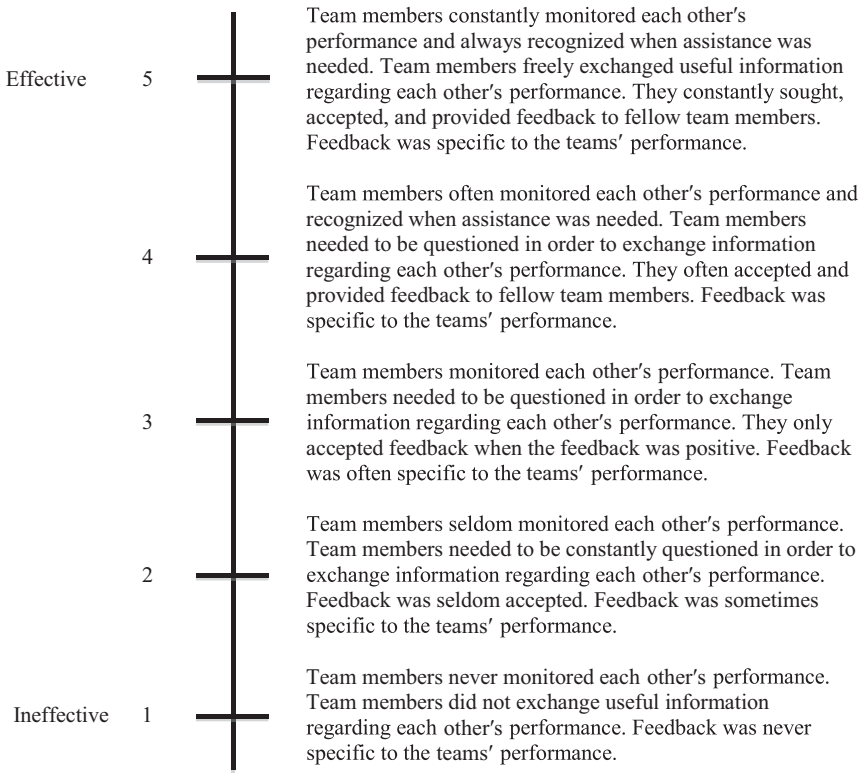

Figure 3 Example of a behavioral anchored ratings scale measuring feedback.

Flight segment	Event	Target	Hit
Approaching checkpoint	Navigation	Report distance to checkpoint.	
		Report outbound heading from checkpoint.	
At checkpoint	COMMUNICATION from base: SIGMET (SIGnificant METeorlogical information) over checkpoint.	Note implications of SIGMET (e.g., cannot land at checkpoint).	
		Obtain additional information (e.g., from flight service).	
		Make new plan – choose alternates from: • Return to base (best choice) • Pleasanton • Kelly AFB • Randolph AFB	

Figure 4 Events and associated TARGETS for a short-hop segment in which helicopter crew is on a training maneuver (Fowlkes et al., 1994).

downside is that it can only detect overt responses (Meister, 1985). Additionally, these types of tools are not especially well designed at measure team attitudes or cognitions.

Concept Mapping

Concept mapping produces a graphical representation of the knowledge content and structure of an individual (Swan, 1995). Concept mapping has more recently been adapted from the individual level to assess team cognitions. This method creates a map of cognitive concepts that are linked together based on their association with the other concepts. In order to generate the concept maps, the researcher can request the data from team members through interviews/surveys or by post-hoc examination of the data (Mohammed et al., 2000). There are several variations of concept maps, all with their unique purpose. For example, Pathfinder elicits similarity judgments from team members to produce a representation of the team's concept structure (Mohammed et al., 2000). This technique involves the researcher coming up with predefined concepts, usually using

a team task analysis. Participants are then instructed to make pairwise comparison ratings on each concept pairing. After these paired-comparison ratings are made, Pathfinder uses a specialized algorithm to determine the psychological proximity of the concepts as well as a conceptual structure (see Goldsmith and Kraiger, 1997; Cooke and Schvanevedlt, 1988).

Summary

Certain measurement tools are more apt at assessing attitudes (e.g., rating scales), behaviors (e.g., BARS, event-based measurement scales), and cognitions (e.g., concept mapping). When selecting measure to use in a team measurement system, it is recommended that one considers the strengths and weaknesses of each of these measurement tools. Though we emphasize the importance of having a complete team measurement system, if the proper tools are not utilized in appropriate ways, having such a system is meaningless.

The Future of Team Measurement

While much has been discussed thus far regarding the current state of the science of team measurement in terms of important considerations and potential approaches, there is still much to be done. Team measurement as a science is far from being perfect, especially as we continue to explore new and changing team contexts. The following provides a brief summary of two particularly important areas requiring future research, the increasing use of distributed teams communicating via virtual media, and the need to better represent the dynamic and temporal nature of teamwork.

Distributed Virtual Teams

The growing implementation of teams whose members are geographically distributed and communicated virtually is also an issue in need of additional research. Distribution of team members has become commonplace in organizations around the world as a means to deal with complex world problems, and is the focus of a growing body of academic research across multiple disciplines (Connaughton and Shuffler, 2007). Distributed virtual teams are defined as teams whose members are geographically separated and must utilize some type of technology in order to communicate, whether it be something as basic as e-mail or as complicated as a virtual (Driskell et al., 2003). This definition illustrates the major features of such teams: the use of technology to bring together team members that may be spatially distributed (i.e., not colocated), or temporally distributed (i.e., interacting asynchronously; Kozlowksi and Bell, 2003).

Distribution and virtuality are important factors to consider in terms of team measurement as they can change how team members interact with one another. Initial results of both lab and field research illustrate differences in distributed virtual teams when compared to face to face teams (Bos et al., 2006). However, these teams are often investigated using the traditional metric approaches applied to face-to-face teams, and do not take into consideration how distribution and virtuality may be influencing when and how metrics should be captured. Overall, the research regarding how to capture the nuances that may be important to distributed virtual teams is very limited, especially in terms of partial distribution as well as the combined influences of distribution and virtuality. Given the rise toward distributed teams relying upon technology to communicate, it is therefore critical that attention be paid to this issue from a measurement perspective.

Dynamic Assessment of Team Constructs

Although advances in measurement over recent years have led to increasingly effective techniques for capturing team level constructs, additional work is needed in order to effectively capture the dynamic nature of teams. The measurement techniques previously described in this article are effective, but do not always capture the degree to which teams can change rapidly over a period of time. For example, observers may only capture teamwork processes for a short period of time during the performance episode that they are observing, leaving out the richness of how the team may have developed up to that point or after that point. Furthermore, paper-based metrics provided by team members force members to summarize their thoughts and feelings over a period of time, essentially averaging rich data that could be critical to effectively understanding the nature of team performance. Thus, it is critical that future research regarding teamwork and team assessment focus on the dynamic, changing nature of teams over time.

Along these lines, increased attention regarding the implementation of unobtrusive measures in team settings has led to the proposed use of metrics such as physical closeness of team members over a period of time or biometric data (e.g., heart rate, perspiration, eye tracking) as proxies for team level constructs such as trust or cohesion (DeCostanza et al., 2012). Additionally, the need to reduce the reliance on expert observers is also critical to capturing this dynamic nature of teamwork. While metrics such as the BARS and TARGETS techniques described in this article allow less-expert observers to assess team variables, this still incorporates the use of humans in the measurement process, increasing the likelihood of error and mistranslation of information. Furthermore, such observers cannot be present constantly over the lifecycle of a team in every situation; therefore changes that the team may experience in the constructs of interest may not be fully captured. Thus, other ways to reduce this reliance on experts must be developed. In sum, team measurement is a critical concept for both research and practice, as only through effective measurement we will be able to more accurately understand the nuances of teamwork and team performance necessary for enhancing teams and their success in organizations.

See also: Behavior Analysis, Applied; Industrial–Organizational Psychology: Science and Practice; Pay, Compensation, and Performance, Psychology of; Personnel Selection, Psychology of; Program Evaluation.

Bibliography

Barrick, M.R., Stewart, G.L., Neubert, M.J., Mount, M.K., 1998. Relating member ability and personality to work-team processes and team effectiveness. Journal of Applied Psychology 83 (3), 377–391.

Bos, N., Olson, J., Nan, N., Shami, N., Hoch, S., Johnston, E., 2006. Collocation blindness in partially distributed groups: is there a downside to being collocated?. In: Conference on Human Factors in Computing Systems Proceedings, pp. 1313–1321.

Brannick, M.T., Salas, E., Prince, C.W., 1997. Team performance assessment and measurement: theory, methods, and applications. Psychology Press.

Bunderson, J.S., Sutcliffe, J.M., 2003. Management team learning orientation and business unit performance. Journal of Applied Psychology 88 (3), 552–560.

Burke, C., Stagl, K., Klein, C., Goodwin, G., Salas, E., Halpin, S., 2006. What type of leadership behaviors are functional in teams? A meta-analysis. Leadership Quarterly 17 (3), 288–307.

Campbell, D., Fiske, D., 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychological Bulletin 56 (2), 81–105.

Cannon-Bowers, J., Salas, E., 1997. A framework for developing team performance measures in training. In: Team Performance Assessment and Measurement: Theory, Methods, and Applications. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 45–62.

Carson, J.B., Tesluk, P.E., Marrone, J.A., 2007. Shared leadership in teams: an investigation of antecedent conditions and performance. Academy of Management Journal 50 (5), 1217–1234.

Chan, D., 1998. Functional relations among constructs in the same content domain at different levels of analysis: a typology of composition models. Journal of Applied Psychology 83 (2), 234–246.

Chell, E., 1998. Critical incident technique. In: Qualitative Methods and Analysis in Organizational Research: A Practical Guide. Sage Publications Ltd, Thousand Oaks, CA, pp. 51–72.

Chen, G., Donahue, L., Klimoski, R., 2004. Training undergraduates to work in organizational teams. Academy of Management Learning & Education 3 (1), 27–40.

Connaughton, S.L., Shuffler, M., 2007. Multinational multicultural distributed teams: a review and future agenda. Small Group Research 38 (3), 387–412.

Cooke, N.J., Schvaneveldt, R.W., 1988. Effects of computer programming experience on network representations of abstract programming concepts. International Journal of Man–Machine Studies 29 (4), 407–427.

Cooke, N.J., Salas, E., Cannon-Bowers, J.A., Stout, R.J., 2000. Measuring team knowledge. Human Factors 42 (1), 151–173.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychological Bulletin 52 (4), 281–302.

DeCostanza, A.H., DiRosa, G.A., Rogers, S.E., Slaughter, A.J., Estrada, A.X., 2012. Researching teams: nothing's going to change our world. Industrial and Organizational Psychology 5 (1), 36–39.

Dickinson, T.L., McIntyre, R.M., 1997. A conceptual framework for teamwork measurement. In: Brannick, M.T., Salas, E., Princem, C.W. (Eds.), Team Performance Assessment and Measurement: Theory, Methods, and Applications. Taylor and Francis, Mahwah, NJ, pp. 19–44.

Driskell, J.E., Radtke, P.H., Salas, E., 2003. Virtual teams: effects of technological mediation on team performance. Group Dynamics: Theory, Research, and Practice 7 (4), 297–323.

Driskell, J.E., Goodwin, G.F., Salas, E., O'Shea, P.G., 2006. What makes a good team player? Personality and team effectiveness. Group Dynamics: Theory Research and Practice 10 (4), 249–271.

Earley, C.P., Mosakowksi, E., 2000. Creating hybrid team cultures: an empirical test of transnational team functioning. Academy of Management Journal 43 (1), 26–49.

Edwards, B., Day, E., Arthur, W., Bell, S., 2006. Relationships among team ability composition, team mental models, and team performance. Journal of Applied Psychology 91 (3), 727–736.

Fowlkes, J.E., Lane, N.E., Salas, E., Franz, T., Oser, R., 1994. Improving the measurement of team performance: the TARGETs methodology. Military Psychology 6 (1), 47–61.

Goldsmith, T., Kraiger, K., 1997. Applications of structural assessment to training evaluation. In: Ford, J.K., Kozlowski, S.W.J., Salas, E., Teachout, M. (Eds.), Improving Training Effectiveness in Organizations. Erlbaum, Mahwah, NJ, pp. 73–96.

Hackman, J.R., Morris, C.G., 1975. Group tasks, group interaction process and group performance effectiveness: a review and partial integration. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 8. Academic Press, New York, NY, pp. 47–99.

Ilgen, D.R., Hollenbeck, J.R., Johnson, M., Jundt, D., 2005. Teams in organizations: from Input-Process-Output models to IMOI models. Annual Review of Psychology 56, 517–543.

Jones, M.C., Harrison, A.W., 1996. Is project team performance: an empirical assessment. Information & Management 31 (2), 57–65.

Kozlowski, S.W.J., Bell, B.S., 2003. Work groups and teams in organizations. In: Borman, W.C., Ilgen, D.R., Klimoski, R. (Eds.), Comprehensive Handbook of Psychology, Industrial and Organizational Psychology, vol. 12. Wiley, New York, pp. 353–376.

Kozlowski, S.W.J., Klein, K.J., 2000. A multilevel approach to theory and research in organizations: contextual, temporal, and emergent processes. In: Klein, K.J., Kozlowski, S.W.J. (Eds.), Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions. Jossey-Bass, San Francisco, CA, pp. 3–90.

LePine, J.A., Piccolo, R.E., Jackson, C.L., Mathieu, J.E., Saul, J.R., 2008. A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. Personnel Psychology 61 (2), 273–307.

Lynn, G.S., Reilly, R.R., 2000. Measuring team performance. Research-Technology Management 42 (2), 48–56.

Marks, M., Mathieu, J., Zaccaro, S., 2001. A temporally based framework and taxonomy of team processes. Academy of Management Review 26 (3), 356–376.

Mathieu, J., Maynard, M., Rapp, T., Gilson, L., 2008. Team effectiveness 1997–2007: a review of recent advancements and a glimpse into the future. Journal of Management 34 (3), 410–476.

Meister, D., 1985. Behavioral Analysis and Measurement Methods. Wiley-Interscience.

Messick, S., 1980. Test validity and the ethics of assessment. American Psychologist 35 (11), 1012–1027.

Mohammed, S., Klimoski, R., Rentsch, J.R., 2000. The measurement of team mental models: we have no shared schema. Organizational Research Methods 3 (2), 123–165.

Morgeson, F.P., Reider, M.H., Campion, M.A., 2005. Selecting individuals in team settings: the importance of social skills, personality characteristics, and teamwork knowledge. Personnel Psychology 58 (3), 583–611.

Mortensen, M., Hinds, P.J., 2001. Conflict and shared identity in geographically distributed teams. International Journal of Conflict Management 12 (3), 212–238.

Orasanu, J., 1990. Shared Mental Models and Crew Performance. No. 46. Princeton University, NJ. Laboratory of Cognitive Science, pp. 172–181.

Salas, E., Burke, C.S., Fowlkes, J.E., Priest, H.A., 2003. On measuring teamwork skills. In: Thomas, J.C., Hersen, M. (Eds.), Industrial and Organizational Assessment, vol. 4. John Wiley and Sons, Hoboken, NJ, pp. 427–442.

Salas, E., Cooke, N.J., Rosen, M.A., 2008. On teams, teamwork, and team performance: discoveries and developments. Human Factors: The Journal of the Human Factors and Ergonomics Society 50 (3), 540–547. http://dx.doi.org/10.1518/001872008X288457.

Salas, E., Dickinson, T.L., Converse, S.A., Tannenbaum, S.I., 1992. Toward an understanding of team performance and training. In: Swezey, R.W., Salas, E. (Eds.), Teams: Their Training and Performance. Ablex Publishing, Westport, CT, pp. 3–29.

Smith, P.C., Kendall, L.M., 1963. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology 47 (2), 149–155.

Smith-Jentsch, K.A., Mathieu, J.E., Kraiger, K., 2005. Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting. Journal of Applied Psychology 90 (3), 523–535.

Stevens, M.J., Campion, M.A., 1994. The knowledge, skill, and ability, requirements for teamwork: implications for human resource management. Journal of Management 20 (2), 503–530.

Swan, J.A., 1995. Exploring knowledge and cognitions in decisions about technological innovation: mapping managerial cognitions. Human Relations 48 (11), 1241–1270.

Technological Change and the Sociotechnical System, Applied Psychology of

Beryl Hesketh, University of Western Sydney, Sydney, NSW, Australia
Warwick Graco, Australian Taxation Office, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Technological change and its impact on work is explored within Emery's (1969) sociotechnical framework where changes in the technical system affect the cultural, social, and management fabric of organizations. The focus of this article is on the digital revolution, and its impact on employment practices, teams, and research opportunities in a global and culturally complex environment. Highlighted within this article are work relevant developments in areas such as robotics, the Internet, social media, and data analytics, all made possible through the abundance of computing and communication capacity. Areas for future research are outlined.

Each wave of technological change, including most recently the digital revolution, has had a major impact on the nature of work. This article highlights recent technological development and explores its impact on the social and cultural nature of work and organizations.

A useful framework for understanding the role of technology in work is Emery's (1969) sociotechnical systems model of organizations where the technological, social, and managerial components interact. The technological system includes all the equipment, infrastructure, and technology in the workplace. The social system includes cultural and other diverse groups and individuals, and the social processes and informal channels used to communicate and negotiate in the workplace. The managerial system is concerned with power and authority within the organization, including decision making and formal lines of communication. Technological change has an obvious and direct impact on the technical system but also on the social/cultural and management systems, especially in relation to modes of communicating and working, and the ways in which data are gathered, analyzed, and used.

Technological Change: Information Technology, the Internet, and the Digital Revolution

Each new wave of technological development has resulted in changes to the nature of work (Wilpert, 2000). Examples include the Industrial Revolution, developments in transport (air, land, and water), and most recently the information technology and communications revolution also known as the digital revolution. The uptake of new technologies has not been uniform across the globe, and it is therefore difficult to speak of this without acknowledging the diversity of stages of technological change in different countries and cultures.

Industrialization resulted in a particular form of work that emphasizes the division of labor, centralization of functions, and new forms of labor relations. The digital revolution has added another layer of change with implications for the location of work in time and space, and for the immediacy of communication irrespective of distance. Examples of the digital revolution include the abundance of computing capacity; the use of robots, smart devices, and sensors; and the rise of the Internet, video-based conferencing, social media, and smart phones in the home, in school, and at work.

Although computerization has been evident for decades, the recent growth in the sophistication, speed and availability of computing power, storage capacity, and communication bandwidth lies at the heart of the current digital revolution, providing the platform for developments in robotics and other areas that have and are radically changing work. As a consequence of faster and more economical computing and improvement in sensory capacity, planning control, actuation, and power systems, robots have become cheaper and more sophisticated (A Roadmap for US Robotics, 2009: pp. 17–19, 63–70, 74–78; Kozlowski, 2011). We illustrate the impact of recent digital advances using examples of robotic applications in manufacturing, transport and defense, while also highlighting the role of social robots, avatars, and the Internet in teamwork and other social roles.

Robots have been used for many decades on the factory floor to assist with the production of cars and other products and their impact is growing at 8% per year, representing a US $5 billion industry in manufacturing in 2009. Robots are seen to be important because of the aging population where optimistic forecasts point to two robot workers for each retired person by 2050, helping solve the need to sustain increases in productivity with a decrease in the number of human workers (A Roadmap for US Robotics, 2009: p. 9).

Robots are not only used in the manufacture of vehicles, but have led to the appearance of autonomous vehicles on highways and city roads (Dietsch, 2012). These are now being tested in traffic in the USA, demonstrating that they can move from one location to another without assistance of human beings. Autonomous vehicles also are expected to bestow many economic and social benefits with one being to transport those who are aged and/or handicapped and thereby provide sustained personal autonomy. This will be important with the expected doubling in the numbers of retirees as a percentage of the workforce by 2030 in countries such as the USA.

Robots have emerged as an important technology in the recent wars in Iraq and Afghanistan (Singer, 2009). Drones such as Predators, Global Hawks, and Ravens and mobile unmanned ground vehicles have replaced soldiers in performing some tasks especially those that are repetitive, dirty, distant, and dangerous such as mine clearance and bomb disposal.

In manufacturing, old school engineers who worked with lathes, drills, stamping presses, and molding machines have

been replaced by automated machines and robots with highly educated workers operating these via their computers. Three-dimensional printers can be used to build products layer by layer, allowing the manufacture of tailored products each customized to individual tastes and requirements (additive manufacturing). Factories are becoming vastly more efficient because of the digitization of manufacturing and the use of smart software (The Economist, 2012). Mass production is, in some areas, giving way to the production of tailored products, made economically in much smaller numbers and with less workers albeit highly skilled ones. These innovations are also influencing hospitals, schools, office buildings, and other places of work where smart technology is replacing workers in their traditional roles, and creating a range of new types of work.

These new technologies provide benefits such as reducing workplace accidents, shrinking manufacturing times, and leveraging the skills of human beings with robot strength, precision, and repeatability (Dietsch, 2012), but automation also creates challenges for the social and managerial systems as discussed later.

The Internet has revolutionized communication in organizations and society generally. Examples include the impact of Internet shopping on retailing, and the effect of e-mail and other electronic communication on traditional postal services. Secure digital mobile devices such as smart phones and tablets are becoming the mobile virtual offices of the future and are making it easier for people to communicate, cooperate, collaborate, and codesign (Gregg, 2011). Workers interact with others on the Internet to solve problems, tackle market opportunities and produce innovative products. Many organizations are also adopting social networking for marketing and other functions. Falls in technology costs and the digitizing of back-office functions such as clerical support and information technology services are expected to make virtual or 'bricks and mortar' organizations, smaller, leaner, and more efficient.

Technology has historically created more jobs than it has destroyed. Similarly the rise of the Internet displaced some jobs but continues to provide new ones from web site designers to professional bloggers (Pelissie du Rausas et al., 2011). However, it is likely that jobs requiring a 'high degree of flexibility in perception, motion and control that robots currently lack' (Chitta et al., 2012: p. 58) will not disappear in the short-term and hence nonroutine, low-paying jobs that are difficult to automate or outsource such as taxi drivers, hairdressers, and gardeners are likely to remain in demand.

Social, Cultural, and Managerial Systems

Organizations and workplaces are increasingly multicultural and diverse. This has arisen because of globalization of industry, education, and many other activities, and increased migration and portability of populations. Many large international firms have branches throughout the world, with a consequential requirement placed on their employees to deal with different ethnic groups (Hesketh and Bochner, 1994). These trends have led to the recognition of a need to understand different cultures as consumers and employees.

Triandis defines culture in terms of the group's characteristic way of perceiving its social environment (Triandis, 1972: p. 3). This should be extended to the group's way of perceiving the technological and managerial environment. Values are fundamental to these perceptions, and Smith and Schwartz (1997) have made a major contribution to thinking in the area through their empirically based typology for evaluating environments cross culturally. Cultures differ in terms of: (1) conservatism (collectivism), where the emphasis is on group goals, interdependence, and closely knit groups; (2) hierarchy, where resources and roles are allocated by the hierarchy; (3) mastery, where the emphasis is on self-enhancement and achievement; (4) autonomy (individualism), with independence emphasized in the pursuit of individual goals; (5) egalitarian commitment, where the welfare of others transcends self-interest; and (6) harmony, for the world as it is and for the preservation of the environment. Nations can be characterized in terms of distinct cultural value patterns or profiles that map on to a circumplex. Bochner and Hesketh (1994) present data that show how these types of value differences, particularly collectivism and individualism, can be found within one multicultural organization.

Cultural differences also exist among different age groups, such as between the Baby Boomers and Generation Y (Alsop, 2008). The aging population in the Western world has led to delayed retirement and a need for effective integration of older workers, including in the use of newer technologies (Hesketh et al., 2011).

Workplaces are diverse, and this diversity has both advantages and disadvantages as new waves of technology take effect. Cultural issues can arise with the use of different media. The use of videoconferencing and e-mails could exaggerate cultural and age differences where cues and jargon from those of a particular ethnic background or age group are misunderstood by others (Tannenbaum et al., 2012). Another problem that can cause difficulties is when staff from the same ethnic background communicate with each other in their native language, especially when there are tensions between managers and workers. There is a growing body of research addressing new digitally-based modes of communication and associated cultural issues (Kappas and Kramer, 2011). This is an area of research likely to increase in the next decade.

Despite the challenges, diverse workforces can result in better decision making, higher creativity, and innovation, more marketing opportunities to different ethnic communities and age groups and better and more economic opportunities. The challenge for management is to use technological developments positively to create a productive social and work climate. In the next section we use several examples to illustrate the implications of the digital revolution on the social and managerial systems with a focus on the diversity of the workforce.

Implications of Technological Change for the Culturally Diverse Social and Managerial Systems

Three themes will be used to illustrate the implications of technological change for the social and managerial system.

First, the ways in which social media is affecting employment practices in a transnational and multicultural context will be examined. Second, the impact of smart technologies and the digital revolution on individual and team work will be illustrated. Finally, there has been a significant increase in the amount of data in the world in the last few years. This potentially rich management tool and source for insights into organizations and individual workers is considered in the context of the new discipline of behavioral analytics, which is also referred to as behavioral informatics.

The Impact of Social Media on Employment Practices

Social media refers to web-based dialogues between communities, organizations, and individuals. It covers collaborative projects such as Wikipedia, blogs and microblogs such as Twitter, content communities such as YouTube, and social networking such as Facebook (Wilson et al., 2012). It assisted the Arab Spring in North Africa and in the Middle East in 2011 (Choudhary et al., 2012). Social media enabled people to exchange critical information and keep up with developments in these important events. Wilson et al. (2012) highlight the potential role of Facebook and other social media in researching cultural and other demographic differences.

In the employment context, it can be used for a variety of purposes such as finding passive information on potential employees. It can allow those seeking work to place resumes, photos, and postings thus giving recruiters easy access to a large talent pool around the world regardless of nation, culture, and race (see Brown and Daly, 2011, for comment). A complementary use of online services is to match candidates to jobs to improve retention levels and to reduce costs. Another use is microwork. This is a social media application using what is called crowdsourcing (Bohannon, 2011; Hoffman, 2012) to recruit temporary workers to do very small tasks for small amounts of money. An example of this is Amazon's Mechanical Turk that was originally created to have workers do simple tasks that computers find difficult to do (Hirth et al., 2011).

A new development with social media is that staff can market an organization's products and services through their social media connections and correspondence such as their tweets and blogs. It is reported that people are being scored for how influential they are on the Internet, and applicants for jobs are being eliminated because their scores on this construct are too low (Smith, 2012). It is now considered important for people to expand their social footprint regardless of their background and geographic location to enhance their employment and career prospects. Social media is transcending national and cultural boundaries and opening a market for talent to employers as well as giving many employees access to jobs previously denied because of geography, race, culture, and age.

Technological Change and the Nature of Individual and Team Work

The growth in communication technologies has made it possible to provide flexible work arrangements that are no longer restricted in time and space and that increasingly involve the integration of smart technologies. Although the uptake of flexible work arrangements has been slower than one might have expected, the new technologies do have implications for the way in which work is distributed and monitored.

Smart technologies in the form of smart agents in computers may replace human experts that individuals and team members can consult on a range of issues. Social robots (Fisher, 2011) that have personality are being integrated into teams and the work context. Avatars, which are computer representations of people and can act as their alter egos, are being employed. Social robots pose challenges including that they can invade workers' personal space and may be seen to be intrusive. They can tap people's need for affiliation and intimacy but the interactions are with machines without feelings and which can be socially clumsy (Midha and Nandekar, 2012). Feil-Seifer and Mataric (2011) outline a range of ethical issues needing further attention as socially assistive robotics increases.

Tannenbaum et al. (2012) point out that many teams now function in a more fluid, dynamic, and complex environment than in the past. They operate with looser boundaries and are more likely to be culturally heterogeneous, reliant on technology, geographically dispersed and in different time zones. These changes are being driven by the need for economic efficiencies including becoming leaner, quicker, and more responsive. Technological innovations coupled with costs of travel are enabling the use of virtual teams. Teams, as a result, are formed, reconfigured, and sometimes disbanded quickly. Movement is the new norm. Flash teams, such as emergency surgery teams, can form and disband quickly. A team may have an inner core or circle of permanent workers with an outer band of temporary workers often located in geographically dispersed areas. As noted above, through the use of social media, workers will compete for work in the global marketplace rather than where they live. Keeton et al. (2012) highlight space teams as the ultimate high technology, distributed, multinational, and multicultural team structure. They point to lessons learned from delayed communication between the spaceflight team and mission control that may be relevant to telecommuting. Selection of team members for technological and intercultural competencies in these situations is critical.

Managers will face challenges dealing with fluid teams and organizations (Galbraith, 2010) where individuals have membership of multiple teams all of whom have varying ties to the organization or the core of the team. The lack of clarity in boundaries creates difficulties in establishing team identity. Management may need help in selecting team members to optimize team cohesion, identity, shared purpose and performance (Tannenbaum et al., 2012), and more sophisticated methodologies will be needed to research and understand team performance (Kozlowski and Chao, 2012). Communicating to workers in different time zones is another challenge and managers will need to be cognizant of worker burnout and resentment because of the round-the-clock working day. Issues of trust will arise as managers will have to put in place measures that monitor individual and team activities to ensure that work is done and that privileges are not misused and abused (De Jong and Dirks, 2012).

How workers and work teams interact with social robots, avatars, and other smart agents will also test the capabilities of managers. Workers have interacted with robots on production lines for decades and they have recently learned to work with them on the battlefield. They will now have to learn to work with social robots in the office, at home, in hospitals, and in many other venues. Robots will soon be ubiquitous in that they will be found everywhere. Different cultural and age groups will react in various ways to these new forms of work involving social robots and avatars, providing important challenges in traditional areas of applied psychology such as selection, training, organizational change, and research (Kozlowski and Chao, 2012).

Use of Behavioral Analytics

Behavioral analytics can potentially help management make sense of the vast array of data available in digital form (Brown et al., 2011) in this digital age. Behavioral analytics emerged around 2005, as part of broader development of analytics which combines the digitally enabled upward surge in data with the use of statistical, machine learning, and operations research techniques to discover patterns and trends in data and to develop models that classify, predict or perform other functions (Davenport, 2010). Analytics as a term is generally used to refer to those quantitative and linguistic-based disciplines, such as natural language processing that extract knowledge from data to assist people to understand issues and to reach decisions (Benitez et al., 2012).

Breiman (2001) contrasts behavioral analytics and behavioral science in the context of the discipline of statistics, arguing that data-driven behavioral analytics complements the theory or model-based behavioral sciences. Those who do behavioral analytics discover, explain, and predict behavior based on knowledge extracted from data, while those who do behavioral science develop models or theories and then check if the data support them.

Behavioral analytics can be applied to a wide variety of issues including organizational change, consumer choices, regulatory behavior to understand compliance with government programs and to the analysis of performance to understand what gives athletes a competitive advantage. Behavioral analytics can be applied to the workplace to examine industrial production processes, how people and teams learn and evolve and the impact of generational change. Specific examples where behavioral analytics has been applied in the workplace include: fraud detection (Ngai et al., 2011), educational analytics (Baker and Yacef, 2009), and team dynamics (Murase et al., 2012). Murase et al. illustrate how network analytic methodologies can be used to capture the multilevel, distributed, longitudinal, and evolving nature of teams and team activity by leveraging 'the large volume of digital traces these teams generate' (p. 44). Research into complex dynamic behavior that has been comparatively neglected over decades because of limited data capture and analysis tools are now possible through digital capture of ongoing interactions with technology and new analytic approaches able to deal with voluminous data.

An example of behavioral analytics is a study of e-mail use (Fidock and Carroll, 2011). This is a social-technical study of how practitioners engage in developing, managing, and implementing systems that support cooperative work and communication. The researchers developed a model of technology appropriation to describe the processes through which users adopt, adapt, and then incorporate technology with their personal, social, or work practices. This approach sees people shaping and being shaped by systems they use. The researchers found from the study of e-mail usage that management must find a balance between being prescriptive about system use and allowing staff enough freedom to find innovative ways of using system and procedures to support their work. Participants do not simply adopt a technology and employ its features in a way consistent with the technology's intended use based on its design, functionality, and other characteristics. Instead users make choices about how to modify the technology and its characteristics to best meet their work preferences and circumstances. The researchers counseled that the design of new technologies should be an evolutionary process where design continues at the same time as a solution is being evolved and implemented and as users take ownership of it and adjust it to suit their needs.

Research Issues Requiring Further Research

There are many issues covered above that require further research. Several stand out including:

- Retraining of workers displaced by technological change and how they can quickly and effectively be reskilled to work in socially diverse and technologically changing environments.
- The social psychology of robots and workers and how they can interact effectively to achieve mutual outcomes. Social robots are reported to be awkward and socially clumsy. Research is required into how workers can develop trust in social robots and can work together to raise productivity (Kappas and Kramer, 2011).
- The most effective digitally enhanced modes of social and other communication across time, space, and culture.
- Social media and recruitment and how to prevent this process being misused and abused by employers where it discriminates against employees based on race, color and creed and based on what it reveals about applicants that is private and intrusive.
- Preparation of staff to manage at a distance or what is referred to as 'remote management'. There will be many challenges for managers coordinating and controlling workers with different qualifications and skills and from different cultural backgrounds and working in different locations in the world. An added difficulty will be how they manage smart technology as well as smart people.
- Exploration of new applications of behavioral analytics to a wide range of fields in applied psychology, especially as a sophisticated tool for research (see, for example, comment by Collins, 2012).

See also: Diversity in Organizations; Human Factors and Ergonomics; Industrial–Organizational Psychology: Science and Practice; Organizational Climate in the Work Setting; Personnel Selection, Psychology of.

Bibliography

A Roadmap for US Robotics - From Internet to Robotics, 2009. Computing Community Consortium, 21 May. [This report is from a workshop on emerging technologies and trends involving experts from academia and industry 14–15 August 2008.]

Alsop, R., 2008. The Trophy Kids Grow Up: How the Millennial Generation is Shaking Up the Workplace. Jossey-Bass, NY.

Baker, R.S.J.D., Yacef, K., 2009. The state of educational data mining in 2009: a review and future visions. Journal of Educational Data Mining 1 (1), 3–17.

Benitez, J.M., Garcia-Pedrajas, N., Herrera, F. (Eds.), 2012. New Trends in Data Mining. Knowledge-Based Systems 25 (1) (special issue).

Bochner, S., Hesketh, B., 1994. Power distance, individualism/collectivism, and job-related attitudes in a culturally diverse work group. Journal of Cross-Cultural Psychology 25 (2), 233–257.

Bohannon, J., 2011. Social science for pennies. Science 334 (6054), 307.

Breiman, L., 2001. Statistical modeling: the two cultures. Statistical Science 16 (3), 199–231.

Brown, B., Chui, M., Manyika, J., October 2011. Are you ready for the era of big data? McKinsey Quarterly, 1–12.

Brown, V.R., Daly, V.E., 2011. The writing on the (Facebook) wall: the use of social networking sites in hiring decisions. Journal of Business Psychology 26 (2), 219–225.

Chitta, S., Jones, G., Ciocarlie, M., Hsiao, K., 2012. Mobile manipulation in unstructured environments: perception, planning and execution. IEEE Robotics and Automation Magazine 19 (2), 58–71.

Choudhary, A., Hendrix, W., Lee, K., Palsetia, D., Liao, W., 2012. Social media evolution of the Egyptian revolution. Communications of the ACM 55 (5), 74–80.

Collins, F., 2012. How to fulfill the true promise of "mHealth". Scientific American 307 (1), 6.

Davenport, T.H., 2010. The New World of Business Analytics. IIA International Institute for Analytics.

De Jong, B.A., Dirks, K.T., 2012. Beyond shared perceptions of trust and monitoring in teams: Implications of asymmetry and dissensus. Journal of Applied Psychology 96, 391–406.

Dietsch, T., 2012. New robots: sidekicks to superheros. IEEE Robotics and Automation Magazine 19 (2), 7–8.

Emery, F.E., 1969. Systems Thinking. Penguin, Harmondsworth, UK.

Feil-Seifer, D., Mataric, M.J., 2011. Socially assistive robotics: ethical issues related to technology. IEEE Robotics and Automation Magazine 18 (1), 24–31.

Fidock, J., Carroll, J., 2011. Why do users employ the same systems in so many different ways? Intelligent Systems 24 (4), 32–29.

Fisher, K., 2011. How people talk with robots: designing dialogue to reduce uncertainty. Artificial Intelligence Magazine 32, 31–38.

Galbraith, J.R., 2010. The multidimensional and reconfigurable organization. Organizational Dynamics 39 (2), 115–125.

Gregg, M., 2011. Work's Intimacy. Wiley/Polity, Cambridge UK.

Hesketh, B., Bochner, S., 1994. Technological change in a multicultural context: implications for training and career planning. In: Triandis, H., Dunnette, M., Hough, L. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 4. Consulting Psychologist Press, Palo Alto, CA.

Hesketh, B., Griffin, B., Loh, V., 2011. A future-oriented retirement transition and adjustment framework. Journal of Vocational Behavior 79, 303–314.

Hirth, M., Hobfeld, T., Train-Gia, P., 2011. Human Crowd as Emerging Internet Application – Anatomy of Microworkers' Crowdsourcing Platform. Report Number 475. Institute of Computer Science, University of Wuerzburg, Germany.

Hoffman, J., 2012. Q & A: the data visualiser. Nature 486 (7401), 33.

Kappas, A., Kramer, N.C. (Eds.), 2011. Face to Face Communication over the Internet: Emotions in a Web of Culture, Language and Technology. Cambridge University Press, NY.

Keeton, K.E., Schmidt, L.L., Slack, K.J., Malka, A.A., 2012. The rocket science of teams. Industrial and Organizational Psychology 5 (1), 32–35.

Kozlowski, K.R. (Ed.), 2011. Robot Motion and Control. Springer-Verlag, London.

Kozlowski, S.W.J., Chao, G.T., 2012. Macrocognition, team learning and team knowledge: origins, emergence and measurement. In: Salas, E., Fiore, S., Letsky, M. (Eds.), Theories of Team Cognition: Cross Disciplinary Perspectives. Routledge Academic, NY, pp. 19–48.

Midha, V., Nandekar, A., 2012. Impact of similarity between avatar and their users or their perceived identifiability: evidence from virtual teams in second life platform. Computers in Human Behavior 28 (3), 929–932.

Murase, T., Doty, D., Wax, A., DeChurch, L.A., Contractor, N.S., 2012. Teams are changing: time to "think networks". Industrial and Organizational Psychology 5 (1), 41–44.

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X., 2011. The application of data mining techniques in financial fraud detection: a classification framework and review of literature. Decision Support Systems 50 (3), 559–569.

Pelissie du Rausas, M., Manyika, J., Hazan, E., Bughin, J., Chui, M., Said, R., May 2011. Internet matters: the Net's sweeping impact on growth, jobs and prosperity. McKinsey Global Institute Report, 55 pages.

Singer, P.W., 2009. Wired for War. Penguin, New York.

Smith, F., 22 May 2012. All you need to score is a little bit of 'klout'. Australian Financial Review, 62–63.

Smith, P.B., Schwartz, S.H., 1997. Values. In: Berry, J.W., Segall, M.H., Kagitcibasi, C. (Eds.), Handbook of cross-cultural psychology, second ed., vol. 3. Allyn & Bacon, Boston, pp. 77–118.

Tannenbaum, S.I., Mathieu, J.E., Salas, E., Cohen, D., 2012. Teams are changing: are research and practice evolving fast enough? Industrial and Organizational Psychology 5, 2–24.

The Economist, 4 May 2012. Farewell, traditional factories, hello tailor-made manufacturing. Australian Financial Review, 48–49.

Triandis, H.C., 1972. The Analysis of Subjective Culture. Wiley, NY.

Wilpert, B., 2000. Applied psychology: past and future societal and scientific challenges. Applied Psychology: An International Review 49 (1), 3–22.

Wilson, R.E., Gosling, S.E., Graham, L.T., 2012. A review of facebook research in the social sciences. American Psychological Science 7 (3), 203–220.

Technological Determinism

Ronald R Kline, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Technological determinism' is a term used to describe a set of claims made about the relationship between what we generally call 'technology' and 'society.' Two meanings have come into use: (1) an internal, technical logic determines the design of technological artifacts and systems; and (2) the development of technological artifacts and systems determines broad social changes. The second claim is much more common and is often associated with debates over Karl Marx's theory of history. But the two meanings are often conjoined in the claim that an autonomous technology (in both its development and use) shapes social relations. Other claims are less strong and express the belief that technology is a major cause, but not the sole determinant, of social change. Although scholars have argued for many years against the strong version of technological determinism, the general belief that technology is a major force shaping society, which dates to the early nineteenth century, still pervades popular culture in the United States and Europe. Ironically, critics of the harmful effects of technology tend to reinforce the strong claims of technological determinism. Methods developed to analyze the social construction of technology have moved the debate from questions like 'does technology drive history?' to arguments about a mutual relationship between technological and social change. New research recommends taking technological determinism seriously in order to understand its justificatory role in developing and using sociotechnical systems.

'Technological determinism' refers to a pervasive, yet controversial, theory about the relationship between technology and society. Although the term has had a variety of meanings, two related claims have been central to discussions of this topic: (1) the development of technology proceeds in an autonomous manner, determined by an internal logic independent of social influence; and (2) technological change determines social change in a prescribed manner (Staudenmaier, 1985; Misa, 1988; Bimber, 1994). The claims address two major questions about technology: how and why is technology developed and what is the relationship between technological change and social change? Arguments against the first claim have been a staple of research in the history and sociology of technology since the 1960s. The second claim has been at the center of debates about Karl Marx's theory of history since the early twentieth century. Weaker versions of technological determinism, sometimes called 'soft determinism,' maintain that technology is a major cause, but not the sole determinant, of social change.

History of the Concept

It is important to distinguish between widely held beliefs that an autonomous technology is the primary cause of social change and the use of the term technological determinism. In the United States, the Enlightenment doctrine that improvements in the mechanic and industrial arts fostered social progress was replaced over the course of the nineteenth and early twentieth centuries by a more technocratic concept of progress, in which improvements in technical and economic efficiency become ends in themselves, instead of the means for creating desired social ends. Through the writings of the economist and social theorist Thorstein Veblen and others, the word technology gradually displaced phrases like 'mechanic arts' in scholarly and then popular discourse in the twentieth century. In this period, the older meaning of 'technology' as systematic knowledge of crafts and industry was joined to the anthropological one of the artifacts and related practices of a culture (Schatzberg, 2006). The term's connotations of abstractness, scientific objectivity, and masculinization served to reinforce the idea that technology was an autonomous social force (Marx, 1994, 1997; Oldenziel, 1999).

These meanings entered popular culture through such media as utopian novels, exhibits at World's fairs, the Technocracy movement in the 1930s, and company advertisements that endlessly touted new products as the cure for social ills and the source of all happiness. The organizers of the 1933 Chicago World's Fair encapsulated the gender biased, applied science aspect of technological determinism with the motto, 'Science Finds, Industry Applies, Man Conforms.' Ironically, twentieth-century critics of technology's harmful effects, such as technological unemployment and environmental degradation, have tended to reify the notion that technology is an autonomous agent of change in their attempts to control it (Smith, 1994).

The term 'technological determinism' is more recent, with roots in the turn-of-the-century debates about Marx's theory of history. While Europeans used the phrases 'historical materialism' and 'economic determinism' to describe Marx's theory, many American social scientists employed the broader phrase, the 'economic interpretation of history,' popularized in a non-Marxist manner at the turn of the century by Edwin R.A. Seligman, editor-in-chief of the first edition of the *Encyclopedia of the Social Sciences*. Historian Werner Sombart in Germany and economist Alvin Hansen in the United States argued that Marx's theory should more properly be called a 'technological interpretation of history.' Although many scholars questioned this view, prominent Soviet theorists established a 'technicist' version of historical materialism as the dominant interpretation in orthodox Marxism during the first half of the century (Bober, 1927; MacKenzie, 1984; Miller, 1984).

During the early Cold War, social scientists and historians, who were increasingly concerned about how to control the

effects technology apparently was having on daily life, increasingly used the starker term, 'technological determinism' to criticize Marxist theories of technology and society. They also used the term to criticize the controversial views of such authors as Veblen, historian Lynn White, Jr., and the philosopher Jacques Ellul. White and Ellul said their views did not amount to a strict determinism. Only a few authors, such as economist Robert Heilbroner (1967), have labeled their own position a form of 'technological determinism' (soft determinism in Heilbroner's case). Philosopher Langdon Winner criticized scholars for dismissing the issue of technology's social effects because of the flaws in the concept of technological determinism. Saying that the "idea of determinism is not one that ought to be rejected out of hand," Winner argued that technology does constrain human activities (Winner, 1977: 77).

By the mid-1970s, the issue of technological determinism had become a central topic in both Marxist scholarship and in the interdisciplinary study of science, technology, and society (Cohen, 1978; Staudenmaier, 1985). In line with the more recent turn toward social constructivism in the latter field, the issue of technological determinism became (and remains) a focal point around which to criticize a theory of technology and society that is still prevalent in other disciplines and in popular culture.

Was Marx a Technological Determinist?

Scholars are sharply divided about whether or not Karl Marx's theory of history, the *locus classicus* of technological determinism, is indeed technological determinist. Much of the debate centers on the meaning of an often-quoted excerpt from the preface to Marx's *A Contribution to the Critique of Political Economy* (1859): humans enter into "relations of production [*Produktionsverhältnisse*] which correspond to a definite stage of development of their material productive forces [*Produktivkräfte*]. The sum total of these relations of production constitutes the economic structure [of society], the real basis on which arises a legal and political superstructure" (quoted in Cohen, 1978: 28).

Those who argue for a technological determinist interpretation of Marx usually equate "material productive forces" (or "forces of production") with technology and interpret the verb 'correspond' to mean that the relations of production evolve to maximize the productivity of the forces of production. William Shaw, who includes instruments, raw materials, and labor power (including knowledge) in the category of productive forces, defines Marx's "technological determinism" more properly as a "productive-force determinism." Shaw acknowledges that Marx allowed for interplay between forces and relations of production, but maintains that productive forces are the "long-run determinant of historical change" (Shaw, 1979: 160). Gerald Cohen defines productive forces in a similar, broad manner, but rejects the label of technological determinism in favor of a "'technological' interpretation of historical materialism." In this functionalist view, social structures adapt to technological change to increase productivity (Cohen, 1978: 29).

Those who argue that Marx was not a technological determinist typically have concentrated on the manner in which he wrote history. Harry Braverman, Nathan Rosenberg, and Donald MacKenzie point to the dialectical aspects of Marx's writings, especially where he attributed the development of productive forces to social relations. Indeed, MacKenzie characterizes Marx's treatment of the history of machinery in the British Industrial Revolution as an "attempt to develop a theory of the social causes of organizational and technical changes in the labor process" (MacKenzie, 1984: 492). Richard Miller argues that Marx was not a technological determinist because, in his historical work, "economic structures do not endure because they provide maximum productivity. Productive forces do not develop autonomously. Change in productive forces, in the narrowly technological sense that excludes work relations, is not the basic source of change in society at large" (Miller, 1984: 188). Bruce Bimber, who defines technology narrowly as artifacts, argues that, for Marx, human characteristics of accumulation and self-expression drive the forces of production. Accordingly, Marx was an economic, not a technological, determinist (Bimber, 1994). From a linguistic perspective, Raymond Williams argues that Marx's usage of the German verb *bestimmen*, usually translated as 'to determine,' in key passages denoting the relationship between the mode of production (base) and the realm of culture and ideology (superstructure), refers to a "notion of determination as setting limits, exerting pressures," rather than a prefigured determination (Williams, 1973: 4).

Varieties of Technological Determinism

Bimber and Thomas Misa effectively analyze varieties of technological determinism. Bimber (1994) distinguishes between normative, nomological, and unintended-consequences accounts. Normative accounts, evident in the writings of such critics of technology as Lewis Mumford, Jacques Ellul, and Jürgen Habermas, claim that society is relinquishing control over technology, or replacing political and ethical norms with those of the technologist's goals of efficiency and productivity. Nomological accounts, such as that by Heilbroner (1967), evoke the two-part definition of technological determinism given above: technology develops autonomously according to an internal logic and forces a prescribed social change. In unintended consequences accounts (e.g., Winner, 1977), technology produces unpredictable social change, a view that challenges determinism, but reinforces the idea that technology is out of control. Bimber argues that the term 'technological determinism' should be reserved for nomological accounts, which are the only ones that satisfy his strict definition of the term.

In his survey of scholarship about the relationship between technology and social change, Misa (1988) found that sweeping, macrolevel accounts were more likely to be technologically deterministic (using the two-part definition of the term) than microlevel accounts, which focused on specific practices. Philosophers, business historians, urban historians, and historians of physical sciences were generally more deterministic than technological historians and labor historians.

Another form of technological determinism is what Claude Fischer (1992) criticizes as the 'impact-imprint' model of technology. Rather than investigating how consumers use

a particular technology, scholars assume that a technology's capability (e.g., the ability of the telephone to enable people to talk at a distance) leads to a predictable 'impact' (e.g., new, long-distance communication patterns).

Political Artifacts and Systems

Winner and Thomas Hughes have proposed influential theories of technology and society that address the question of technological determinism. Winner (1986) argues that artifacts have politics (i.e., are associated with specific social changes) in two ways. First, technologists can design a (plastic) artifact to promote a certain type of politics, as when Robert Moses built low bridges over the Long Island Expressway in order to prevent inner-city buses (carrying poor minorities) from traveling to Jones Beach. Second, the design and building of some systems seems to require a certain type of politics. Nuclear power plants, for example, require a centralized form of government to run the reactor and prevent the proliferation of nuclear weapons. (For a criticism that the low-bridge example does not take the tradition of expressway building into account, see Joerges, 1999). Hughes (1994) states that his concept of technological systems allows for the social construction of a system when it is young, but technological determinism comes into play once it is established and has 'technological momentum.' From that point on it shapes society more than being shaped by society.

Social Construction of Technology

The advent in the mid-1980s of the approach known as the social construction of technology (SCOT) (Pinch and Bijker, 1987), along with a contextualist history of technology (Staudenmaier, 1985), has challenged the technical aspect of technological determinism (the first part of the definition given above). Sociologists and historians of technology have published numerous case studies of the development of technologies that argue that their design was the outcome of negotiations between several social groups (such as inventors, engineers, managers, salespersons, and users), rather than the product of an internal, technical logic (*see* Technology, Social Construction of). When Winner (1993) and others criticized this approach for leaving out power relations among social groups, constructivists responded by investigating the social aspect of technological determinism (the second part of the definition). Bijker (1995) employed a semiotic version of Michel Foucault's "micropolitics of power" to show the mutual construction of artifacts and social groups. Ronald Kline and Trevor Pinch (1996) utilized theories of gender in the original SCOT approach to show how gender relations and the automobile were mutually constructed when farm men in the United States before World War II adapted the car, as a stationary and mobile source of power, to a variety of novel uses, such as grinding grain, washing clothes, and plowing fields. Fischer (1992) employed a similar user-heuristic to argue that callers in the United States used the telephone before 1940 to widen and deepen existing communication patterns, rather than to create new ones. Kline (2000) extended that analysis to show how farm men and women wove several so-called 'urban' technologies (the telephone, automobile, radio, and electric light and power) into their lives, creating individual versions of 'rural modernity.'

Reconsidering Technological Determinism

The work by Bijker, Kline, and others on the mutual construction of technology and society resonates with recent calls to reconsider the dismissal of technological determinism in the history and sociology of technology. Rosalind Williams has asked historians of technology to drop their objection to the concept and become willing to "taste the forbidden fruit of technological determinism" by paying much more attention to the social and cultural changes involved in using technology (Williams, 2007: 104). Sally Wyatt has argued that science and technology studies (STS) scholars "must take technological determinism more seriously, disentangle the different types, and clarify the purposes for which it is used by social actors in specific circumstances." By recognizing four types of distinct rhetorical treatments of determinism (justificatory, descriptive, methodological, and normative) and reflecting on their personal attitudes toward technological determinism, Wyatt argues that STS scholars can deepen their "understanding of the dynamics of sociotechnical systems and the rhetorical devices of some decision makers" (Wyatt, 2008: 169, 176). Other sociologists of technology have taken up the concept of the social 'affordances' offered by technology, even when those are socially constructed (David and Pinch, 2006). These efforts are part of a larger movement to go beyond the original aims of SCOT in order to understand how using technology shapes society in a nondeterminist manner (Kline, 2013), in this case by studying what work the (academically discredited) belief in technological determinism does in promoting new technologies.

See also: Engineering Studies; Gender and Technology: From Exclusion to Inclusion?; Marx, Karl (1818–83); Science and Technology, Social Study of: Computers and Information Technology; Technology, Social Construction of; Wundt, Wilhelm Maximilian (1832–1920).

Bibliography

Bijker, W.E., 1995. Of Bicycles, Bakelite, and Bulbs: Towards a Theory of Sociotechnical Change. MIT Press, Cambridge, MA.

Bimber, B., 1994. Three faces of technological determinism. In: Smith, M.R., Marx, L. (Eds.), Does Technology Drive History? The Dilemma of Technological Determinism. MIT Press, Cambridge, MA, pp. 79–100.

Bober, M.M., 1927. Karl Marx's Interpretation of History. Harvard University Press, Cambridge, MA.

Cohen, G.A., 1978. Karl Marx's Theory of History: A Defence. Princeton University Press, Princeton, NJ.

David, S., Pinch, T.J., 2006. Six degrees of reputation: the use and abuse of online review and recommendation systems. First Monday 11 (6). http://firstmonday.org/ojs/index.php/fm/article/view/1590/1505 (accessed 02.02.14.).

Fischer, C.S., 1992. America Calling: A Social History of the Telephone to 1940. University of California Press, Berkeley.

Heilbroner, R.L., 1967. Do machines make history? Technology and Culture 8, 335–345.

Hughes, T.P., 1994. Technological momentum. In: Smith, M.R., Marx, L. (Eds.), Does Technology Drive History? The Dilemma of Technological Determinism. MIT Press, Cambridge, MA, pp. 101–114.

Joerges, B., 1999. Do politics have artefacts? Social Studies of Science 29, 411–431.
Kline, R.R., 2000. Consumers in the Country: Technology and Social Change in Rural America. Johns Hopkins University Press, Baltimore, MD.
Kline, R.R., 2013. Foundational stories. Technology and Culture 54, 117–129.
Kline, R.R., Pinch, T.J., 1996. Users as agents of technological change: the social construction of the automobile in the rural United States. Technology and Culture 37, 763–795.
MacKenzie, D., 1984. Marx and the machine. Technology and Culture 25, 473–503.
Marx, L., 1994. The idea of 'technology' and postmodern pessimism. In: Smith, M.R., Marx, L. (Eds.), Does Technology Drive History? The Dilemma of Technological Determinism. MIT Press, Cambridge, MA, pp. 237–258.
Marx, L., 1997. Technology: the emergence of a hazardous concept. Social Research 64, 965–988.
Miller, R.W., 1984. Analyzing Marx: Morality, Power, and History. Princeton University Press, Princeton, NJ.
Misa, T.J., 1988. How machines make history and how historians (and others) help them to do so. Science, Technology, & Human Values 13, 308–331.
Oldenziel, R., 1999. Making Technology Masculine: Men, Women, and Modern Machines in America, 1870–1945. Amsterdam University Press, Amsterdam.
Pinch, T.J., Bijker, W.E., 1987. The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other. In: Bijker, W.E., Hughes, T.P., Pinch, T.J. (Eds.), The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge, MA, pp. 17–50.
Schatzberg, E., 2006. Technik comes to America: changing meanings of technology before 1930. Technology and Culture 47, 486–512.
Shaw, W.H., 1979. 'The handmill gives you the feudal lord': Marx's technological determinism. History and Theory 18, 155–166.
Smith, M.R., 1994. Technological determinism in American culture. In: Smith, M.R., Marx, L. (Eds.), Does Technology Drive History? The Dilemma of Technological Determinism. MIT Press, Cambridge, MA, pp. 1–36.
Staudenmaier, J.M., 1985. Technology's Storytellers: Reweaving the Human Fabric. MIT Press, Cambridge, MA.
Williams, R., November–December 1973. Base and superstructure in Marxist cultural theory. New Left Review 82, 3–16.
Williams, R.H., 2007. Opening the big box. Technology and Culture 48, 104–116.
Winner, L., 1977. Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought. MIT Press, Cambridge, MA.
Winner, L., 1986. Do artifacts have politics? In: Winner, L. (Ed.), The Whale and the Reactor: A Search for Limits in an Age of High Technology. University of Chicago Press, Chicago, pp. 19–39.
Winner, L., 1993. Upon opening the black box and finding it empty: social constructivism and the philosophy of technology. Science, Technology, & Human Values 18, 362–378.
Wyatt, S., 2008. Technological determinism is dead; long live technological determinism. In: Hackett, E., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge, MA, pp. 165–180.

Technology and Organization

Tarek Tantoush, Open University, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 23, pp. 15502–15506, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

The term 'technology' is in common everyday use. It may be popularly visualized in purely physical terms as 'equipment,' 'apparatus,' 'pieces of hardware and software,' and it may be defined by different commentators in different ways. Whatever the case may be, what is much more serious than definitional problems are the practical implications of the development, implementation, and control of technology by *a few* for the use of *the many* in an organization and in wider society. This article addresses the issue of technology and the relationship between technology and 'organization.'

Organizing and Organization: An Overview

To initiate a discussion on the relationship between organization and technology, it is important first to define what is meant by each of the two terms. First, what is meant by the term 'organization'? Well, to organize something is to 'systemize' it for a certain desired end. 'Organizing' a group of people is a process that aims to give an orderly structure to that group with a view to achieving a goal or fulfilling an objective in mind. In such a process the group in question acquires, or is brought into, a state of being 'organized.' This definition is quite a broad one; it is meant to be so, since an encompassing conceptualization of the term is required to grasp the range of possible positions and influences of technology in and on organizations.

From an organizational studies perspective, organizations are defined as social arrangements for the controlled performance of collective goals (see Huczynski and Buchanan, 1991; Scott and Davis, 2007). In a broad sense, therefore, examples of an organization may include an African tribe, Juventus Football Club, a consumer electronic goods manufacturer, the European Union, or the University of Harvard.

Defining Technology

Technology has received increased attention since the late twentieth century because of the pace of technological advancement and the magnitude of its diffusion in society. Indeed, it is likely to impose itself as a pivotal organizational issue more forcefully than ever before because of the pervasiveness of 'new' technology and its effects on organizing and on the structure and functioning of organizations as well as the behavior of individuals and groups within them. Examples of new technologies include information technology, biotechnology, materials technology, space technology, and nuclear technology.

But what is technology? Well, again there is no single answer to this question, although to a layperson it is merely 'equipment,' 'apparatus,' or 'pieces of hardware and software.' Even among commentators in the field of organizational studies, differences can be found as to how technology is defined. Regardless of the way it is defined, however, technology is an important factor in organizing and organizational life. A suitable working definition for the purpose of this article is that, in essence, technology is a man-made creation of a *system* that is supposed to work logically and consistently, and is made operational to produce a desired accomplishment that is valued by its inventor or user. Such an accomplishment is often reproducible, through the system, to a given set of performance criteria or precision.

The creation of such a system involves translating some pure science knowledge into an *input–conversion–output* system format based on engineering principles and composed of elements that are functionally arranged or configured so as to yield a certain output as specified by the designer or user (see Clark et al., 1988). According to this definition, technology is defined not only in terms of its physical objects but also in terms of its methods of operation, use, or production and the technical know-how needed to apply them successfully. Human intervention in the operation of a system usually takes the form of monitoring the system's conversion process and controlling its output through a specific feedback mechanism.

Conceptualization and Meaning: From Neutral Rationality to Political Sensitivity

Conceptualization of technology merely in terms of 'equipment,' 'apparatus,' or 'pieces of hardware and software' carries with it shades of pure rationality in mind. This conceptualization depicts technology as nothing more than a physical construction of a proven scientific idea or inference. The emphasis here is on technology as a manifestation of a purely logical effort, which produces a neutral 'object' or 'pattern of operation' that helps mankind per se rationalize a given work task or improve the quality of a certain aspect of social or work life. This is often achieved by substituting human effort with machine effort, a phenomenon which began historically with physical effort (mechanization) and continues today with human coordination and intellectual efforts being replaced with inputs from 'smart machines' (computerization and automation). This takes place in an information age where information technology has become part of a global culture and knowledge-based artificial intelligence is likely to progress further. Thus far the notion of technology could be described as an objective description of an outer form of a disinterested, 'inanimate' object or mechanistic pattern of operation.

The above chosen *system* definition of technology, however, goes beyond this rather simplistic conceptualization as it takes

on board the role of the designer or user in *configuring* a system that gives an output in accordance with the designer's pre-specified performance criteria. Therefore, the proposition that technology could be explained merely in terms of presumed neutral rationality grounded in a specific scientific logic appears suddenly to be in question as a credible explanation because of the interference of the subjectivities of those associated with its instrumentalization and operation in the real world.

Just how politically sensitive technology is can be illustrated by explaining how significant it is to organizing and organizations. This is a turning point in conceptualizing technology as a serious organizational development involving actual and perceived impacts of a system's output upon organizing and within organizations. In this sense, technology is capable of being socially divisive (Leonardi and Barley, 2008). That is, a system's output can and does affect the status quo of any organization that accommodates it or comes to live under its influence. This will have implications for any long-established equilibrium in relationships, balances of power, and any sense of continuity within the organization. It may also have implications for the wider industrial sector or economy in which the organization operates and with which it interacts. Additionally, it may indeed have implications for the environment (e.g., pollution and health hazards in the case of a nuclear technology). Since differences of interpretation among individuals and groups will inevitably occur in the world of meaning, a system's output will signify and symbolize different things to different people; it will be highly valued by some but not by others; it may seem useful to some as something that improves the quality of working life, for example, and harmful to others who see in it a threat or deterioration. In short, it is liable to being controversial and bringing about a sense of dispute as to why it has been acquired in the first place and, now that it has, how it should be managed and controlled to minimize its potential damage.

Furthermore, what makes technology a politically sensitive issue in organizational life is the fact that invariably technology is decided upon, adopted, designed, implemented, and changed by a minority of people in an organization (e.g., managers) on behalf of the majority of the organization's members (e.g., employees) who have to live with the consequences of operating it as an integral part of their jobs. Nonusers are also likely to be affected, in that their jobs will now require different coordination skills to keep them in harmony with a changed rhythm and workflow throughout the workplace, for example, where new production methods are introduced by management. A common concern for users and nonusers alike may be that the introduction of a new *input–conversion–output* system format into the workplace is going to disrupt stability and that they have to change their working methods in order to suit the operational requirements of a new and somehow threatening technology-dictated work system.

By the same token, technology, at a societal level, is often invented by a single individual (e.g., a scientist) or an elitist minority in society (e.g., a technology research center or a team of top biochemists); it is commercialized, marketed, and diffused by another minority (e.g., entrepreneurs or manufacturing companies), who are driven by a different set of motives (e.g., profit-making); it is regulated and controlled centrally by yet another minority (e.g., a sovereign state's legislature or a federal government). Yet, this technology is highly likely to be consumed by the vast majority in society. Even nonconsumers are likely to be affected by it in either a positive or negative way, or both. Because of this minority–majority situation, technology acquires necessarily social, economic, and political implications as well as rational, logical, and technical significances.

It is no wonder, therefore, that the politics of organizational change are increasingly accomplished through technologies in a digital age. Many of the most radical redesign projects for organizations have been in and through the use of digital technologies: from computer-integrated manufacturing to process reengineering in manufacturing industry; from projects for 'virtuality' to extensive teleworking, organizations are being sold on projects of technical rationality as visions for the future. Yet, so many of these projects appear to be stillborn: the technologies work, but the organizational designs do not. Sometimes people sabotage the technologies; other times the technologies sabotage them. In a way, organizations are reduced to criteria of design from which most considerations of the politics of organizational change are absent. This is often the case even where the projects are explicitly political or require explicit politics to manage their changes. It is the politics of organizational change technologies that make the critical difference: How, and whether, they work?

The Relevance of Technology to Managing and Management

It would be misleading to suggest that management should constitute a homogeneous team making rational decisions. Different levels and functions of management within an organization tend not to be a single social or political decision-making group sharing a single rationality; these will therefore tend to hold different perceptions of technological change and are likely to differ among themselves as to how it should be managed (see Pettigrew, 1985). Yet, it is clear from the discussion so far that an organization includes people who, because of their membership of a collective social group, interact with one another in certain ways to achieve goals which if they act alone they cannot accomplish. Since it is management's responsibility to have work tasks carried out by employees, in the context of a business organization, for example, it follows that the impact of technology is of central concern to managers per se, particularly those who are in charge of implementing projects involving the introduction of a new technology. Planning, control, implementation, and evaluation activities relating to such projects require decisions to be made, ranging from individual motivation through job design to the management of change and coordination of a newly designed work organization. Such decisions are made by managers occupying different positions within an organization's management structure. Regardless of how successful or unsuccessful those decisions will come to be judged by others, it is *management* as an institution that is going to be judged in terms of its rationality and political sensitivity as well in terms of its dominant attitudes, values, and ideological persuasions.

Whatever the type of organization or management an external observer chooses to study, technology is relevant to

organization because the latter is necessarily a source of the following:

1. money and physical resources, including technology itself;
2. meaning and purpose, including how different individuals and groups within an organization perceive technology in relation to these variables;
3. order and stability, including how stabilizing or destabilizing it is in the context of a change project involving the introduction of a new technology;
4. security and protection, including the consequences of change for security of employment, and health and safety concerns;
5. prestige and self-confidence, including possible improvement or degradation of job design and satisfaction following the utilization of a certain technology; and
6. power and control, including technology-related assignment of authority and shifts in the *de facto* balance of power due to empowerment caused by, for example, expert knowledge of a key technology.

Therefore, management's role in, and responsibilities for, technological change should be emphasized as a legitimate concern for an organization's stakeholders who are not members of the *governing* institution of management, because of, among other things, the strategic, operational, and control choices available to managers in planning, design, and implementation (see Buchanan, 1983; Sorge et al., 1983; Child, 1987).

The Technology–Organization Link: Empirical Studies and Theorization

The relationship between technology and organization could be perceived, explained, and theorized upon differently by different people and groups, depending on many factors as explained above. This is in harmony with a central assumption of this article, which recognizes the legitimacy of a multiplicity of views as to what constitutes this relationship.

Such multiplicity is reflected in the management literature, including major empirical studies theorizing upon the organization–technology relationship. Examples of these studies include Joan Woodward's (1965) investigation of 100 firms in South East England, which used different unit and small batch, large batch and mass, and process production technologies. The main conclusion of this study was that *technical complexity* caused increases in the length of chain of command, in executives' span of control, in the proportion of indirect to direct labor, and in the proportion of managers to the total employed workforce. The technology used to manufacture a product placed certain requirements on those who operated it, which, in turn, were reflected in the organizational structure. Woodward's variable was further studied by Charles Perrow (1976), whose name is also associated with 'technological determinism,' since he argued that specific aspects of job discretion, coordination, and group interdependence were determined by the type of technology used, much more so than by other variables such as leadership style or small group dynamics.

The Aston research program in Birmingham, England (e.g., Pugh and Hickson, 1976; Pugh and Hinings, 1976), contributed to organization theory by blending some of the research methods and assumptions of psychology with conceptions of organizations and their workings from sociology and economics. The empirical work undertaken by the Aston group aimed to link organizational structure and functioning, group composition, and interaction, as well as individual personality and behavior. It did so by researching organizational structures in terms of their degrees of 'specialization,' 'standardization,' 'formalization,' 'centralization,' and 'configuration.' According to the Aston studies, an organization's size and its dependence upon other organizations were found to be much more significant variables than technology and ownership.

Other relevant empirical contributions to theory include that of Eric Trist and his colleagues, of London's Tavistock Institute, who studied the effects of mechanization in coal mining (e.g., Trist and Bamforth, 1951; Emery and Trist, 1960). Their 'sociotechnical perspective' contributed considerably to the Quality of Working Life movement, particularly in the USA and Canada, where Fredrick Taylor's (1947) 'scientific management' style dominated the industrial scene as many large organizations followed Henry Ford's mass production model, which caused degradation in working conditions in the context of a capitalist economy.

Classical theorizing about organizational technology has produced a set of typologies for categorizing technology, which are often used to develop and empirically test contingency theories, including Woodward's (1965) above mentioned work, giving rise to the notion of 'technological imperative.' Other important issues concerning social structure are associated with the position of technology in modernist organization theory, including task interdependence, uncertainty, complexity, and change in the environment of an organization. Linear models of technological innovation are increasingly being challenged, however, by postmodernist theorists and social constructionists (e.g., Clegg, 1990; Gergen, 1992; Pettigrew, 1985), who emphasize the need for researching the contexts in which technologies are introduced.

Conclusion

In summary, this article has addressed questions pertaining to the management of technology adoption and organizational adaptation in organizations. It defined the terms organization and technology, and discussed the relationship between them. One single point remains at issue: The desire for an 'optimum fit' between the two is a noble concern, but in practical situations involving many users and a few decision makers and, equally, in theoretical situations involving numerous typologies and competing interpretations, on whose terms?

See also: Cyclical Industrial Dynamics; Dominant Designs; Ecology: Organizations; Information Society, Geography of; Information and Knowledge, Organizational; Innovation; Organization: Overview; Organizations, Sociology of; Science and Technology, Social Study of: Computers and Information Technology; Supply Chain Management; Technology and Social Control.

Bibliography

Buchanan, D., 1983. Technological imperatives and strategic choice. In: Winch, G. (Ed.), Information Technology in Manufacturing Processes. Rossendale, London.

Child, J., 1987. Organizational design for advanced manufacturing technology. In: Wall, T., Clegg, C., Kemp, N. (Eds.), The Human Side of Advanced Manufacturing Technology. Wiley, Chichester, UK, pp. 101–133.

Clark, J., McLoughlin, I., Rose, H., King, R., 1988. The Process of Technological Change: New Technology and Social Change in the Workplace. Cambridge University Press, Cambridge, UK.

Clegg, S., 1990. Modern Organizations: Organization Studies in the Postmodern World. Sage, London.

Emery, F., Trist, E., 1960. Sociotechnical systems. In: Churchman, C., Verhulst, J. (Eds.), Management Science, Models and Techniques, vol. 2. Pergamon, Oxford, UK, pp. 93–99.

Gergen, K., 1992. Organization theory in the postmodern era. In: Reed, M., Hughes, M. (Eds.), Rethinking Organization: New Directions in Organization Theory and Analysis. Sage, London, pp. 207–226.

Huczynski, A., Buchanan, D., 1991. Organizational Behaviour: An Introductory Text. Prentice-Hall, London.

Leonardi, P.M., Barley, S.R., 2008. Materiality and change: Challenges to building better theory about technology and organizing. Information & Organization 18 (3), 159–176.

Perrow, C., 1967. Organizational Analysis: A Sociological View. Tavistock, London.

Pettigrew, A., 1985. The Awakening Giant: Continuity and Change in ICI. Blackwell, Oxford, UK.

Pugh, D., Hickson, D., 1976. Organizational structure in its context: the Aston programme I. Gower, Aldershot, UK.

Pugh, D., Hinings, C. (Eds.), 1976. Organizational Structure—Extensions and Replications; the Aston Programme II. Gower, Aldershot, UK.

Scarbrough, H., Corbett, J., 1992. Technology and Organization: Power, Meaning and Design. Routledge, London.

Scott, W.R., Davis, G.F., 2007. Organizations and Organizing: Rational, Natural and Open System Perspectives. Pearson Education, Upper Saddle River.

Sorge, A., Hartmann, G., Warner, M., Nicholas, I., 1983. Microelectronics and manpower in manufacturing. Gower, Aldershot, UK.

Taylor, F., 1947. Scientific Management. Harper & Row, New York.

Trist, E., Bamforth, K., 1951. Some social and psychological consequences of the longwall method of coal getting. Human Relations 4, 3–38.

Woodward, J., 1965. Industrial Organization: Theory and Practice. Oxford University Press, London, UK.

Technology and Social Control

Gary T Marx, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

One aspect of modernization is the use of science-based technology in rule enforcement. In the 'engineered society,' an ethos of rationalization is seen in the application of means to ends. The events of 11 September 2001, and the war on terror have brought increased attention to the question, but the increased use of technical means for control reflects continuity rather than disjuncture. Six social control strategies are discussed and illustrated: target removal, target devaluation, target insulation, offender incapacitation, offender exclusion, and identification of offenses and offenders. In complex settings in a democratic society, relying primarily on technology to control human behavior has clear social and ethical limitations.

The technology's narrowing of focus may come at a cost of failing to see larger systemic contexts, alternatives, and longer range consequences. The complexity and fluidity of human situations makes this a rich area for the study of trade-off, irony, and paradox. There are some parallels to iatrogenic medical practices in which one problem is cured, but at a cost of creating another. Technical efforts to ensure conformity may be hindered by conflicting goals, unintended consequences, displacement, lessened equity, complacency, neutralization, invalidity, escalation, system overload, a negative image of personal dignity and the danger of the means determining, or becoming ends.

Introduction

Since the last half of the twentieth century, there has been a significant expansion in the use of science and technology for purposes of social control. Control through technology is central to the working of modern society. The events of 11 September 2001, and the war on terror have brought increased attention to the issue, but reflect continuity rather than disjuncture. Consider the intensification of control seen in the United States such as the Patriot Act and other legislation, and the creation of the Department of Homeland Security; vastly augmented expenditures for research and development; the international spread of tools such as DNA analysis, RFID chips, drones, facial recognition systems, enhanced travel and border controls, data-based risk management and predictive tools; and the increased sharing of information and integration among public and private and local, national, and international agencies.

Many technologies developed for the military such as the Internet, satellites, and sensors have diffused to other institutions. Yet the use of tools to engineer behavior goes far beyond particular historical events and national security to everyday life – whether this involves work, consumption, health, recreation, school, or families and friends.

The engineering of social control is one of the defining characteristics of modern society. It is so prominent, ubiquitous, and transparent in daily life that it is often taken for granted. Our personal spatial, communication, social, cultural, and psychological environments and borders are increasingly subject to technological strategies designed to influence behavior, whether involving conformity with rules, safety, consumption, or attitudes.

In some ways, contemporary social control can be seen to reflect the ethos of the highly regulated prison in which conformity is sought by designing ever more features of the environment, rather than relying on trusting the individual or facing the uncertainty of human will and choice. Are new threats, technologies, expectations, and ways of living resulting in a move toward our becoming a 'maximum security society'? The engineering of control (in both its hard and soft forms) is an important component of our contemporary *surveillance society*. Marx (1985, 1988, forthcoming) expands on these developments. On surveillance more broadly, see Lyon (2007), Ball et al. (2012) and Staples (2000), and on technologies of crime control see Manning (1992) and Leman-Langlois (2013).

A major strand of social control involves the enforcement of rules and standards. This article provides a classification framework for technology-based efforts to prevent or reduce the harm from, or attractiveness of, rule violations and, when that is not possible, to increase the likelihood of discovering and apprehending violators. Such enforcement activities are distinct from other aspects of social control such as the creation of norms, processes of adjudication and sanctioning, or the broad societal guidance and integration that was of concern to early theorists of industrialization and urbanization (Gibbs, 1989; Janowitz, 1975).

The engineering of rule enforcement involves material artifacts (architectural and product design, sensors and alarms, access controls, and software), and also is distinct from forms of social control such as the creation and manipulation of culture, socialization, the redistributive rewards and penalties of the welfare state, and other interpersonal influences that are designed to influence behavior.

Contemporary efforts build upon, but go far beyond, a medieval fortification ethos. Of course, the inventors and builders of the first locks, safes, moats, and walled castles and the developers of early biometric identification systems (e.g., the Italian criminologist Cesare Lombroso, 1835–1909) were engaged in the engineering of social control. What is new is the scale and relatively greater scientific precision, power, omnipresence, continual invention and experimentation, and rapid global diffusion. Technical means of control saturate modern society, colonizing, documenting, and in some ways restricting ever more areas of life.

The roots of contemporary social control lie in the development of large organizations and standardized control technologies (Beniger, 1986). They are one strand of broad processes of rationalization, professionalization, and specialization occurring with modernization (Weber, 1958; Rule, 1973; Foucault, 1977; Cohen, 1985; Laudon, 1986; Gandy, 1993; Zuboff, 1988; Lyon, 1994; Shenhav, 1999). The heterogeneity, scale, mobility, and anonymity of mass society and the pursuit of efficiency and effectiveness encourage reliance on external, impersonal, distance-mediated, and technical means and database memories that locate, identify, register, record, classify, and direct individuals.

The perception of catastrophic risks in an interdependent, global world relying on complex technologies drives the search for definitive technical solutions. Consider issues such as terrorism, crime, obesity, drug abuse, acquired immunodeficiency syndrome, border controls, and the need for economic competitiveness and unprecedented flows of persons, information, and goods across borders. Developments in electronics, computerization, artificial intelligence, cognitive science, biochemistry, architecture, materials science, and many other areas promise new control possibilities. Entrepreneurial efforts (whether by private sector actors such as the security industry or governments such as the United States) have helped spread the technologies at home and abroad.

This use of contemporary technology contrasts with traditional approaches in which environments were less likely to be designed with rule enforcement in mind. Nor was there much probing beneath personal informational borders *before* untoward incidents occurred. Current preventive and anticipatory responses contrast with earlier reactive forms of control in which enforcement agents tended to become involved only *after* violations occurred and in a specific rather than a categorical fashion.

The engineering emphasis may be on environmental conditions, control agents, or actual or potential offenders and victims (sometimes the categories overlap). Consistent with the classical deterrence ideas that Thomas Hobbes (1588–1679) emphasized, some strategic efforts aim to create self-control and rational calculation on the subject's part others, however, seek greater certainty through applying hard engineering means believed to make the violation impossible. Rather than attend to the consciousness or will of subjects, the emphasis through automation is on eliminating, or at least significantly inhibiting, their ability to violate rules. Engineered efforts also seek to limit the ability of social control agents to demonstrate incompetence, mistakes, corruption, or discrimination, as the machine is geared to get 'the human out of the loop.'

Software programs aimed at prevention and early intervention have proliferated. There are many actuarial assessment protocols that profile and assign predictive scores to places and persons with respect to their presumed threat, vulnerability, and risk (Ericson and Haggerty, 1997). This has resulted in the expansion of various kinds of watch lists for at-risk populations such as the mentally ill, sex offenders, and those believed to be at risk of becoming homicide offenders or victims.

Any effort to *influence* persons or events can be seen as a form of engineered social control using tools exerted by an *agent* on a *subject* (or a population of potential subjects) or on an object. Engineering is present whenever a strategy based on beliefs about means–ends relations is applied and need not involve material factors. Prayers as well as magical practices would thus be examples, although they are far from the direct science-based efforts of interest here that involve rule enforcement. Failing to take any action as part of a strategy to avoid seeing a situation escalate is also an example.

Closer to the topic of the present article, although also distinct, is the design of presumably failsafe products *automated* for safety, such as the 'dead man's switch' on a train or the protective shield on a table saw – mechanisms that are intended to transcend human will. Engineering is also present in the *nonautomated* situations designed to persuade or guide behavior (but with choice remaining), such as advertisements for a negative political campaign or to discourage smoking, a waiter reciting the kinds of beers offered by moving from the least to the most expensive (anticipating that the customer will be prone to opt for the most recently mentioned beer), manufacturers increasing the nicotine or sugar content of cigarettes and food, or providing uncomfortable seating in fast-food restaurants to encourage rapid turnover. Also related are strategic efforts that permit, or even facilitate, *rule violations* through deception or reverse engineering. Consider police undercover operations, or online frauds that spoof computer security or create situations that permit blackmail. While sharing strategic logic, the above efforts are distinct from rule enforcement efforts.

Six Strategies

This section considers six engineering (primarily machine- or material-based) efforts to eliminate or limit violations by control of the physical and social environment, rather than by mere appeals to doing the right thing (or at least what an agent desires). This is followed by discussion of some social and ethical implications of such efforts.

The six ways of controlling persons and/or environments emphasize protection/alteration of the victim or the object of the violation (such as a car). Such actions make it impossible or at least more difficult or less inviting for the potential offender to act; or should prevention fail, tactics are intended to increase the likelihood of identification and apprehension. Some engineering efforts involve traditional notions of target hardening, but the engineering of control may also involve the idea of suspect softening or weakening.

Strategy 1 (and sometimes 4 and 5 below) includes *primary* direct prevention efforts. These are designed to eliminate the offense, or increase the difficulty of carrying it out. With the primary engineering strategy it is not necessary to affect the will or calculation of the potential rule breaker. The subjective orientations of the actor (whether based on calculation, a content-filled socialization, or a contentless discipline) are simply ignored. The emphasis is on altering opportunity structures and capabilities rather than the person's conscious choices. The social engineering example of castration as a device to control sexuality (whether literally or, as currently

may be done, chemically) clearly contrasts with appeals to virtue to accomplish the same end.

But primary strategies are not always available and may not live up to their promise. Hence, one sees a series of *secondary* engineering strategies (2, 3, 5, and 6 below) where concern with the will of the violator can be a factor. The more traditional goal of deterrence may be pursued by affecting the calculations of potential violators through devaluing and insulating targets, increasing the costs of nonconformity, and increasing the likelihood that violations and violators will be discovered and that evidence can be traced. What cannot, literally, be prevented may nonetheless be deterred by eliminating the gain, altering the cost-benefit ratio, or enhancing the chances for identification and apprehension.

As in other areas of social intervention such as public health, contemporary approaches place strong emphasis on preventing and minimizing harm. Rather than relying on persuasion (which assumes situations in which individuals have a choice about whether to respect a rule or to behave in a way that an agent wants them to), the emphasis is on direct technological interventions to design problems away. Ideally, problems are anticipated and eliminated (1, below); or where that is not possible, the goal is to create deterrence by reducing the gain (2), making violation more difficult and more costly (3, 4, 5), or by increasing the likelihood of identification and apprehension (6).

Target or Facility Removal

The logic of prevention is clearest and most effective here. Something that is not there cannot be taken or used. The move toward a cashless society is one example. Merchants who only accept credit or debit cards, or whose registers never have more than a modest amount of cash in them, are unlikely to be robbed by conventional means. Furniture built into the wall cannot be stolen. Subway cars and buses made with graffiti-resistant metals are hard to draw upon. Police stings that pass off a harmless white substance as cocaine or that substitute fake for real dynamite or provide inoperable weapons offer evidence of intent to commit a crime, but preclude harm resulting from it. Conversely, deception as disguise may work in the other direction: hiding the appearance of something valuable, as with painting gold ingots black.

Target Devaluation

Here, the goal is to reduce or eliminate the value of a potential target to anyone but authorized users. The target remains, but its uselessness makes it unattractive to predators. Examples include products that self-destruct, as with some car radios when stolen, or that leave clear proof of theft, as with exploding red dye packs that stain money taken in bank robberies. The broken tamper proof seal on food products is intended to deter those who would tamper, and serves as a warning. Another type of example is telephones, computers, automobiles, and even guns that require a unique, presumably nontransferable biometric means of identification (e.g., retinal, voice, or geometric hand pattern) before they can be used. Disposable access codes that can only be used once need not be protected. Encrypted messages can often be easily intercepted, but without the decryption code the data are useless. As an antiteenager congregating tool, some mall shops play classical music in front of their stores.

Target Insulation

With this ancient technique, the object of desire persists but is protected. Perimeter maintaining strategies such as fences, walls, moats, guards, and guard dogs can be distinguished from more specific protections surrounding an object such as safes, armor, chastity belts, and goods that are in locked cases or chained to immovable objects. Other examples involve the hiding or disguising of valuables or persons. Still other examples involve high security, such as in gated communities where access and egress is carefully controlled, or make use of networked sensors, alarms, internet video, and bulletproof barriers, such as in banks separating customers from tellers (a form in Europe controls both entry into and exit from a bank through an enclosed booth). The architectural development of 'skywalks' linking downtown private buildings creates 'sanitary zones' more subject to control than the potentially disorderly public streets below. Softer forms of border protection lie in the use of access codes based on pin numbers, passwords, and biometric measures. Schools and after school programs can be seen as 'safe houses' in which the young are protected from offenders (and not incidentally school officials can watch the contained children for appropriate behavior).

Offender Weakening or Incapacitation

This classic strategy seeks to render potential offenders harmless by disabling or weakening their will or ability to violate the norm in question, or, if the violation occurs, to escape. The means may act directly on the body by permanently altering it and making certain offenses impossible – cutting off the hands of thieves – or at least uninviting, as exemplified in the operant conditioning seen in the novel *A Clockwork Orange* (Burgess, 1962).

A negative association with a particular form of undesirable behavior may be created. After ingestion of a substance such as antabuse, an unpleasant physical reaction (gagging or vomiting) follows when alcohol is consumed. The morphine derivative trexan is used to detoxify methadone as a remedy for heroin addiction.

Drugs such as Depo-Provera may be used to reduce sex drive and birth control implants, as well as sterilization, are sometimes judicially mandated. Lowering serotonin levels and psychosurgery have been used to curb violence.

Passivity, disorientation, or the inability to flee may be created by sensory weapons. Various citizen protection devices that can be defensively used such as mace fit here, as do nonlethal crowd control devices such as electrical, chemical, strobe, and acoustical immobilizers that disorient, stop, restrain, or block individuals (e.g., Tasers, pepper spray, loud music, flash bang devices, beanbag shotguns, sticky foam released on a floor, straitjackets, and a cage dropped upon or a net fired upon those to be controlled). Spikes in the road may be used to stop vehicles and some cars can be remotely stopped.

Related efforts deal not with the body of the offender but with the instrumentalities involved in the offense. Examples include antidrunk driving interlock systems, which require passing a breathalyzer test attached to the automobile ignition system before a car will start; cell phone muzzling devices for automobiles that use global positioning system (GPS) sensors to prevent the phone from being used while the car is being driven; or mixing a bad smelling chemical into a product to prevent it from being inhaled for its hallucinatory effects. An antipaparazzi device identifies the presence of a digital camera and can then remotely disable it through projection of a light beam.

Note also efforts at redefining the meaning of offender behavior. Thus, treating drug addiction as a medical rather than a criminal problem is in one sense a 'crime' elimination program. In some countries such as the Netherlands, there are safe drug houses where registered heroin users are provided drugs under carefully controlled circumstances. This lessens their need to commit crime to obtain the drugs and also offers a safe environment where they are protected from victimization. A *de facto* policy of toleration and damage control can also be seen in the formal creation of (or winking at) adult entertainment zones ('red light districts'). This can mean greater citizen awareness of victimization risks and some self-policing by service providers.

Exclusion

Potential offenders have traditionally been kept away from targets or tempting environments by exile, prison, curfew, and place or activity exclusions (e.g., restricting the sale of alcohol and cigarettes for juveniles). A related form is the visible warning offered by a stigma such as the brand or clipped ear for offenders in medieval Europe, which encouraged others to stay away. Fear of being degraded and ostracized as a result of stigmatization might also serve as a source of deterrence. Electronic monitoring or location devices based on GPS are contemporary examples. In one form, an alarm goes off and a message is sent to authorities if an adjudicated person wearing a transmitter gets too close to a prohibited person (such as an abused spouse) or leaves (or enters) a restricted area.

Capital punishment is the ultimate form of exclusion. At the other extreme is prevention of birth. With the human genome project completed, neoeugenic modes of exclusion are likely to be entertained. For example, the belief (which ignores interactions with the environment and the socially crafted character of most rules) that DNA is linked to violence and other antisocial behavior could generate another ultimate form of exclusion – requiring a license indicating an 'acceptable' genetic pattern before a child is permitted to be born (*see* DNA, Eugenics). Perhaps a less fanciful example (for the present at least) would be requiring special programs of education and even surveillance for individuals carrying the identified genetic patterns presumed to be associated with the behavior of concern. And with DNA databases, convicted (or otherwise included) individuals, and even close family members can be monitored for future acts.

Offense/Offender/Target Identification

Where it is not actually possible to physically prevent the violation, or where it is too expensive to do so, it may be possible to at least know that it took place and perhaps who is responsible, and to locate where the suspects are and/or where contraband may be stashed (as with X-ray, metal, and chemical detectors).

The goal is to document the violation and identify the violator. A major goal of nineteenth-century forensic science was to develop reliable biometric measures of identity based on the analysis of fingerprints, facial measurements, and mug shots, and chemical properties (Thorwald, 1965). These have significantly expanded from involving a person's gait and voice to tracking their distinctive smell. One technique used by the former East Germany involved identifying individuals by their unique olfactory signature. Architectural designs emphasizing visibility as a deterrent fit here (Newman, 1972), as do video, audio, motion, and heat detection devices and access codes that are presumed to document who enters or leaves an area, or is using a resource such as a computer. Hand-activated personal alarm systems, luggage alarms that go off if a purse or suitcase is illegitimately moved or opened, gunshot detection and location devices, cameras to detect speeders and red light violations, and the electronic tagging of consumer items in stores or books in libraries are other examples.

Hidden identification marks left on paper by the manufacturer or by photocopying machines and identifiers, in material that can be used to make explosives, would be included here. The various 'see something, say something' campaigns, which are dependent on the ease and ubiquity of multipurpose cell phones and e-mail, are a related example. Citizens are encouraged to use hotlines or smartphone apps to report, for example, suspicious persons and objects, erratic highway drivers, drug dealing, poaching, and organizational malfeasance. The police in turn use mass communications media to help identify and locate wanted persons via amber alerts, posting warrant information on Websites and crime reenactments on television, and posing as friends on Facebook pages.

Enhancements to criminal history data systems involving new tools for crime mapping of hot spots and analysis (e.g., Computer Statistics (COMPSTAT)) have appeared and there is increased sharing of information within and between control agencies and between the public and private sectors. There has been a significant expansion of surveillance practices for which there are no, or minimal, specific reporting requirements, as with GPS data, stored communication, and subscriber records (e.g., social media sites, texting, and Web browsing and searching).

Some Other Social Control Dimensions

The above categories are based on combining several aspects, such as whether the focus is on the potential offender, victim, or a resource that is part of the violation, and whether the means literally prevent or simply increase the risks and costs or decreases the attractiveness of violation. Combining separate dimensions into such ideal types can be useful as a shorthand way to classify and compare. Yet this approach can also distort

by merging sources of variation that can be analyzed separately and by excluding other variables.

Another approach starts with single dimensions that may cut across the different forms. Among relevant dimensions are classifications based on visibility or invisibility; openness or secrecy regarding use of the tactic both generally and specifically; control of access into or out of a system; emphasis on an individual, an organization, or a network; a focus on the body, consciousness, or the environment; reliance on machines, humans, or both; normative or nonnormative influences and, if the former, whether they involve criminal or civil law, policies, or manners; relative costs of a tactic including errors and mistakes; reliability and validity and ways of determining such costs; the presence or absence of democratic decision making, and review processes regarding the adoption and application of a tactic; and whether the goals of subjects and agents are the same, or at least overlap, or are in conflict. While all technological control efforts have elements in common, those involving dissensus and inequality are particularly important because of the centrality of human rights issues and the prominence of control and countercontrol efforts.

Some Social and Ethical Implications

Whether dealing with questions of justice, quality of life, security, health, the environment, or other issues, the potential benefits of science and technology are hardly deniable. In the case of criminal justice, for example, Clarke (1997) documents a number of successes. Yet given the sense of urgency about many problems and what is often the self-justifying tunnel rhetoric of those offering solutions, caution is needed, as well as midterm corrections, limitations, or prohibitions. To argue for a yellow light is certainly not a call to cease innovation or the search for better solutions.

However ideal a technical control system may appear in the abstract under ideal laboratory conditions, or however successful it may be in the short run, the world of application is often much messier and more complicated than the public relations efforts claim. There is rarely a perfect, or cost-free, technical fix (if nothing else, a given choice is likely to involve using resources that might have gone elsewhere). The technology's narrowing of focus on a given problem may come at the cost of failing to see larger systemic contexts, alternatives, and longer range consequences. The complexity and fluidity of human situations makes this a rich area for the study of trade-offs, irony, and paradox. Goal conflicts, unintended consequences, and sources of neutralization are among the factors limiting technical efforts to ensure conformity, as discussed below.

Goal Conflicts

At an abstract level, consider the possible tension between values. In the case of the new supermaximum security prisons, there is an enduring tension between the values of custody and punishment, as against care and some form of rehabilitation (Rhodes, 2004). More intensive mechanical control, whether within the prison or the community, often comes with a diminution of human contact and help with efforts to overcome the social and personal deficits that contribute to violations. Short-term gains in control may come at a cost of longer term losses (Byrne et al., 2007).

Informing subjects that a technique is in use is intended to serve the goal of deterrence, but such information may in turn serve as a strategic support for clever rule breakers who will then seek ways around it.

The expectation that one should be judged as an individual and in context may conflict with the greater rationality and predictive success believed to be found in categorical responses based on aggregate data and models divorced from the richness of particular situations. An automatic process that eliminates the misuse of authority and appears to offer fair (as in universal treatment) can conflict with the need to respond to the uniqueness of particular contexts. The latter requires discretion to potentially override action by a machine unable to take account of the concrete variation of the situation.

More concretely, goal conflicts in immediate situations can be considered. Thus, barriers need to prevent uninvited parties from entering a private space, while enabling those within such a space to leave in case of emergency. For example, bars over windows designed to keep out thieves may also prevent occupants from escaping through the window in the event of a fire. Conversely, barriers intended to keep persons or animals contained within a facility may lead to their being unable to get out when there is a fire (e.g., in a prison or a horse stall).

In commercial settings where access to merchandise is important, attaching expensive clothes (e.g., leather jackets) to a rack with a locked cable reduces the likelihood that an item will be stolen, but it also complicates trying clothes on and impulse buying. Encryption of information offers security, but at a cost of increased expense, slowing down the time required for a transaction and the risk of losing the key.

Unintended Consequences

Situations involving unexpected and unwanted results offer a rich area for analysis (Merton, 1957; Sieber, 1982; Marx, 1981; Tenner, 1997). It may be difficult to limit the impact of a technology. Terms such as blowback, collateral damage, spillover, backfire, and overshooting the target capture this. Techniques that immobilize suspects may do the same for control agents (e.g., sound wave technology intended to cause suspects to lose control of their bowels and a slippery banana peel substance that makes it difficult to walk). Uncontrollable wind patterns may send tear gas to places where it is not directed (including, literally, blowback on users). Consider the fact that enhanced lighting and lines of visibility can help perpetrators identify victims or control agents, as well as the reverse. The roads used by the Roman legions venturing forth to conquer became equally available to other conquerors who later marched on Rome. President Nixon, in an effort to secretly tape others famously also taped himself, leading to his downfall. Conversely, a protective device can lock everyone out when the keys are lost. The removal of benches from public areas denies the homeless, as well as others, a place to sit.

Automatic processes can result in punishment without trial. For example, *The New York Times* (28 December 1989) reported that a thief in Mobile, Alabama, was killed in a trap set by a homeowner. The trap consisted of two hunting rifles placed in

separate locations. One pointed down a staircase. The rifle fired when the thief stepped on a wire rigged to the trigger. A neighbor called the police when he heard a shot fired and then entered himself. It is easy to imagine good Samaritan scenarios that end disastrously (e.g., a passerby who is shot by a home-made burglar alarm after seeing a fire and rushing in to help).

Second-order effects also may occur. For example, when initially used, barrier strips intended to stop fleeing cars almost instantly released the air in the tires, sometimes causing high-speed crashes. Persons publicly identified as sex offenders may face vigilante attacks. The death of 30 persons in a fire in the London King's Cross subway station was attributed to fumes from antigraffiti paint. Enhanced technical enforcement along the United States–Mexico border has led to a funnel effect in which immigrants seek to enter through more dangerous desert areas with an increase in mortality (Cornelius, 2001).

An intervention may interact with other conditions to produce an undesired outcome. Thus, pepper spray is intended as a nonlethal alternative, but may be fatal to those with severe asthma or other respiratory problems. Consider also the warnings to pacemaker users that electronic sensors are in operation in retail settings. Will a nonlethal antitheft device that delivers a 50 000-V shock be lethal to the driver of a stolen car with a weak heart?

Moreover, there may be longer term health consequences that are not immediately visible. Questions have been raised, for example, about the effect of repeated exposure to radiation from X-ray search machines, particularly for children, the pregnant, and agents with repeated exposure (as reported in *The New York Times*, 26 April 2012).

Displacement

Several forms of displacement can be noted involving place, time, type of offense, and offender (Reppetto, 1976; Norris et al., 1998). Issues of displacement are central to many control settings where there are conflicts of interest and where rule breakers who have some resources find ways to beat control efforts.

The commercialization of protection, such as with embedding a hidden transmitter in cars that permits it to be located by remote electronic activation or gated communities to keep out would-be thieves, may lead to greater victimization of others who are unable to afford such enhanced levels of security.

Derivative offences may appear. The discovery that a target has been rendered useless to an offender may increase violence, whether as a resource to gain the needed access, or out of frustration. For example, the appearance of 'carjacking' is related to more sophisticated antitheft devices on cars. The use of access codes to activate autos and appliances may mean that the crime of burglary is converted to robbery or kidnapping when thieves confront property owners and demand not only the property but also the code to make it work. A frustrated thief may respond to a self-destruct automobile radio by fire-bombing the car.

New secondary laws that criminalize the possession of artifacts or activities designed to thwart enforcement may develop. Consider laws prohibiting the production, distribution, and use of products intended to falsify drug tests or for the possession of a radar detector designed to thwart traffic enforcement. In such cases, the guilty face charges for the secondary offense of possession, even when they were not committing the primary offense.

Neutralization and Escalation

Whether out of self-interested rule breaking, principled rebellion, or human contrariness, individuals can be very creative in neutralizing systems of control. In a free market economy, new controls create incentives to develop means of neutralization. This may lead to a higher level of play without fundamentally altering the game. A dynamic struggle may escalate, for example, when police use body armor and offenders react by using more powerful weapons and wearing armor themselves.

The fact that locks open with keys and borders require access points means they are eternally vulnerable. Marx (forthcoming) identifies a number of *behavioral techniques of neutralization* – strategic moves through which subjects seek to subvert controls. Reverse engineering permits a *breaking* move. For example, not long after antitheft ignition protection systems appeared on automobiles, a device that permitted bypassing the lock became available. No special tool is needed to spray paint over the lens of a video camera. In a *distorting* move, the initial antidrunk driving car interlock systems can be beaten by saving air in a balloon or by having someone else blow into it to start the car. A variety of moves can be seen in efforts to defeat urine drug tests. These range from contaminating the sample by pouring bleach on one's hand to the *switching* move of using a catheter to insert drug free urine into the body. *Discovery* moves are illustrated by a driver's use of radar detectors to locate police radar. When systems cannot be defeated technically, such as with sophisticated encryption technology, then their human context may be compromised, whether through coercion, corruption, or deception. For example, a thief who was unable to break a manufacturer's sophisticated encryption code, nevertheless managed to embezzle millions of dollars through generating fake invoices. He did this by having an affair with the individual who had the decryption codes.

In many settings, subjects have some rights and are entitled to challenge specific technological control. One thus sees *explanatory* moves through which an unfavorable result is reframed in an acceptable way by offering alternative data and citing the claims of rival experts, or making claims about rights and procedural violations. An *empirically valid* result does not guarantee a *socially meaningful* result. Thus, a DNA match between material from a crime scene and a suspect cannot reveal if the death in question resulted from a homicide or self-defense. Even when there is no question that a homicide occurred, the sample might have been planted or the evidential chain of custody may have been compromised. A computer match between persons on welfare and those with bank accounts may reveal a person whose account exceeded the savings limit, but such data do not provide proof of *cheating* since funds may have been held in trust for a funeral – a contingency that is legally permitted, but not necessarily built into the computer program. Audio and video recordings may reflect what was done and said, but may not reveal why, or what a suspect intended. Seeing should not automatically mean believing. Thus, a suspect in an undercover scheme may

have been threatened or entrapped off-camera. Nor does a drug test, even if 'valid,' indicate the presence of drugs within a person's system, a necessary indication of a violation. Depending on the assessment used, if the standard is set low enough it is possible to have a 'false-positive' reading as a result of just being in a room where marijuana was smoked or the result may be as a result of medications or eating certain foods.

Some Additional Factors

The overselling of technical solutions may exaggerate the risks, engendering immobilizing fear and leading to an unduly suspicious and distrustful society as well as wasting resources. Alternatively, an unexamined faith in a tactic's supposedly failsafe nature can lead to complacency and a false sense of security in which individuals are lulled into not taking necessary precautions. Consider a Los Angeles case in which a man sentenced to house arrest and required to wear an electronic surveillance bracelet shot and killed his estranged wife. She had not reported his threats to police because she thought she was safe as long as he had the bracelet on.

The seemingly ever-greater ease and efficiency offered by technological means can conflict with traditional liberty-protecting ideas of reasonable suspicion, minimization, and impracticality. Of course, as a folk expression holds, "if you hang them all you will certainly get the guilty." Risk prediction technology, as in the case of profiling, may be statistically accurate across many cases, but inaccurate with respect to a given case (this is the issue of aggregate rationality vs. individual cases). Even if accurate in an individual case as a prediction of future behavior, it may conflict with the individual's expectation to be judged on the basis of actual behavior.

Improvements in the ability to identify rule breaking may vastly expand the pool of 'suspects' who are subject to social control. Control systems may become overloaded, which can lower morale among enforcers who feel overwhelmed or offer corruptible officials a resource (nonenforcement) to market. Since resources for acting on all the available information may not be at hand, authorities may face charges of discriminatory enforcement.

Even when adequate resources for full enforcement action are available, organizational effectiveness can be harmed. Automatic technical solutions developed without adequate appreciation of complexity and contingency run the risk of eliminating the discretion, negotiation, compromise, and informal understandings that are often central to the morale and effective working of organizations (Dalton, 1959; Goffman, 1961). By broadening the documented pool of violations and violators, authorities may feel compelled to take action in cases that they otherwise think should be ignored. The rigidity of the machine and limited possibilities for immediate innovation can be damaging. One strand of humor involves the automatic, unthinking, and repetitive quality of many mechanical devices (such as in the classic Charley Chaplin film *Modern Times*).

If technical solutions were somehow to be effective at eliminating all rule breaking (leaving aside the conflict between, and the ambiguity of and lack of consensus on, many rules), such systems would risk becoming unduly rigid. Much innovation is initially seen as deviance. Experimentation and risk taking require latitude and can be aided by anonymity and secrecy. A socially transparent, engineered society would be more orderly, but likely less creative or dynamic and less able to respond to changing conditions.

If order depended only on technical means of blocking infractions, rather than on legitimacy, how would people behave if such technical means failed (e.g., in the case of a power failure and a failure even of backup systems)? A social order based primarily on technical fixes is likely to be as fragile over time as one based primarily on overt repression.

Even if systems could somehow be made fool- and fail-proof, with ever more advanced technology, there is a danger of viewing humans as robots, rather than as creative beings capable of making choices about how to behave. The former image is inconsistent with belief in the dignity of the autonomous individual in a democratic society. Whatever a technology is capable of, the view of humans as volitional (and hence responsible for their behavior) and beliefs about the inviolability (absent clear justification) of the borders that protect private personal zones around the human body, mind, relationships, communications, physical space, and past history are central to ideas of respect for personhood. The tools we use to communicate say something about how we see human beings and what kind of a society we live in and seek to create. Symbolism matters. So do precedents.

With new and seemingly effective social control techniques it is important to ask, "where might this lead and what kind of a society is being created?" In the United States, a future radically at odds with the nation's higher ideals is not likely to come about through cataclysmic change, but gradually in a thousand little ways, each perhaps understandable (if often aided by fear or seduction). In totality, however, such small changes can create a very different world – a world arrived at through accretion under the radar, rather than through full public dialogue.

The search for stand-alone mechanical solutions also avoids the need to ask why some individuals and groups break rules, and it points away from examining the social conditions that may contribute to violations and to the possibility of changing those conditions, rather than only trying to change the individual. Technical solutions seek to bypass the need to create consensus in a community in which individuals act responsibly as a result of voluntary commitment to the rules, not because they have no choice or fear reprisals. This emphasis can further social neglect and subsequent problems, leading to calls for more intensive and extensive reliance on technology in a seemingly endless self-reenforcing spiral.

There is a magisterial, legitimacy-granting aura around both law and science (Ericson and Shearing, 1986). Technological controls, presumed to be science based, are justified as valid, objective, neutral, universal, consensual, and fair. Their legitimacy may be strengthened in free market societies where such tactics often can be used by citizens (e.g., video cameras to record police behavior or DNA analysis offered by criminal defendants) and internally by police managers for guarding the guards.

Yet tools are socially created and their results are socially interpreted and thus potentially disputable. They exist in dynamic interdependent systems where interests may conflict, inequalities are often present, and where full impacts may be difficult to envision. Critical inquiry and humility are as needed, as are innovation and experimentation. It is important to maintain, if not a doubtful attitude, at least a respectfully skeptical attitude toward claims for the new; at least until they are examined both empirically and morally.

Our age has two rather distinct fears of technology. One, a la George Orwell, is that it will work all too well, creating a manipulated, totalitarian society naively taking pride in how free it is. The other fear, reflective of Franz Kafka, is that it would not work well enough. This suggests a crazily complex, out-of-control, interdependent, opaque society steeped in technological errors and *Catch-22* absurdities.

A well-known, if often naïve, expression (given that individuals and groups do not start with equivalent resources) holds that where there is a will there is a way. This points to the role of human motivation in obtaining goals. However, in light of the control possibilities made available by science and technology, this slogan can be reversed to "where there is a way there is a will."

The myth of Frankenstein implies that one must be ever vigilant to be sure that one controls the technology rather than the reverse. As Jacques Ellul (1964) argues, there is a danger of self-amplifying technical means silently coming to determine the ends, or even to become ends in themselves divorced from a vision of, and the continual search for, the good society.

See also: Civil Rights; Control, Social; Crime, Sociology of; Deterrence Theory: Crime; Deterrence; Empirical Legal Studies; Eugenics, History of; Forensic Genetic Databases: Ethical and Social Dimensions; Genetics and Forensics; Human Sciences, History of; Law and Society: Development of the Field; Police: A Sociology of Knowledge Approach; Power in Society.

Bibliography

Ball, K., Haggerty, K., Lyon, D., 2012. Routledge Handbook of Surveillance Studies. Routledge, London.

Beniger, J., 1986. The Control Revolutions: The Technological and Economic Origins of the Information Society. Harvard University Press, Cambridge.

Burgess, A., 1962. A Clockwork Orange. W.W. Norton, New York.

Byrne, J., Marx, G.T., 2011. Technological innovations in crime prevention and policing: a review of the research on implementation and impact. Journal of Police Studies 20 (3), 17–38.

Byrne, J., Taxman, F., Hummer, D. (Eds.), 2007. The New Technology of Crime, Law, and Social Control. Criminal Justice Press, Monsey, NY.

Clarke, R. (Ed.), 1997. Situational Crime Prevention: Successful Case Studies. Harrow and Heston, New York.

Cohen, S., 1985. Visions of Social Control. Polity Press, Cambridge.

Cornelius, W., 2001. Death at the border: efficacy and unintended consequences of US immigration control policy. Population and Development Review 27 (4), 661–685.

Dalton, M., 1959. Men Who Manage. Wiley, New York.

Ellul, J., 1964. The Technological Society. Vintage Books, New York.

Ericson, R., Haggerty, K., 1997. Policing the Risk Society. University of Toronto Press, Toronto.

Ericson, R., Shearing, C., 1986. The scientification of police work. In: Bohme, G., Stehr, N. (Eds.), The Knowledge Society: The Growing Impact of Scientific Knowledge on Social Relations. Reidel, Dordrecht.

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Vintage, New York.

Gandy, O., 1993. The Panoptic Sort: Towards a Political Economy of Information. Westview Press, Boulder, CO.

Gibbs, J., 1989. Control: Sociology's Central Notion. University of Illinois Press, Urbana.

Goffman, E., 1961. Asylums. Anchor Books, Garden City, NJ.

Janowitz, M., 1975. Sociological theory and social control. American Journal of Sociology 81, 82–108.

Laudon, K., 1986. The Dossier Society: Value Choices in the Design of National Information Systems. Columbia University Press, New York.

Leman-Langlois, S., 2013. Technocrime, Technology, Crime and Social Control. Routledge, London.

Lyon, D., 1994. The Electronic Eye. Cambridge Polity Press, Cambridge.

Lyon, D., 2007. Surveillance Studies: An Overview. Polity Press, Cambridge.

Manning, P., 1992. Information technology and the police. In: Tonry, M., Morris, N. (Eds.), Modern Policing. University of Chicago Press, Chicago.

Marx, G.T., 1981. Ironies of social control: authorities as contributors to deviance through escalation, nonenforcement and covert facilitation. Social Problems 28, 221–246.

Marx, G.T., 1985. The Surveillance Society. The Futurist. June, 21–26.

Marx, G.T., 1988. Undercover: Police Surveillance in America. University of California Press, Berkeley.

Marx, G.T., forthcoming. Windows into the Soul: Surveillance and Society in an Age of High Technology. University of Chicago Press, Chicago.

Merton, R., 1957. Social Theory and Social Structure. Free Press, Glencoe, IL.

Newman, O., 1972. Defensible Space. MacMillan, New York.

Norris, O., Moran, J., Armstrong, G., 1998. Surveillance, Closed Circuit Television and Social Control. Ashgate, Aldershot.

Reppetto, T.A., 1976. Crime prevention and the displacement phenomenon. Crime and Delinquency 22, 166–177.

Rhodes, L., 2004. Total Confinement Madness and Reason in the Maximum Security Prison. University of California Press, Berkeley.

Rule, J., 1973. Private Lives, Public Surveillance. Allen-Lane, London.

Shenhav, Y., 1999. Manufacturing Rationality: The Engineering Foundations of the Modern Managerial Revolution. Oxford University Press, New York.

Sieber, S., 1982. Fatal Remedies. Plenum Press, New York.

Staples, W., 2000. Everyday Surveillance: Vigilance and Visibility in Postmodern Life. Rowman & Littlefield, Lanham, MD.

Tenner, E., 1997. Why Things Bite Back: Technology and the Revenge of Unintended Consequences. Vintage Books, New York.

Thorwald, J., 1965. The Century of the Detective. Harcourt Brace & World, New York.

Weber, M., 1958. In: Gerth, H., Mills, C.W. (Eds.), From Max Weber: Essays in Sociology. Oxford University Press, New York.

Zuboff, S., 1988. In the Age of the Smart Machine. Basic Books, New York.

Technology Assessment

Arie Rip, University of Twente, Enschede, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Technology assessment (TA) refers to the early identification and assessment of eventual impacts of technological change and applications, as a service to policy making and decision making more generally. TA can be done by (or for) business firms, or as a medical, social, or environmental impact assessment. Its main use is in the public domain, and includes parliamentary TA. Recent developments include public participation and involvement in innovation processes (Constructive TA). Since the 1980s and 1990s, public service TA has professionalized and articulated specific methodologies for that domain. But TA is also part of broader developments such as the emergence of a risk society.

The term 'technology assessment' (TA) has a variety of usages. The main meaning is the early identification and assessment of eventual impacts of technological change and applications, performed as a service to policy making and decision making more generally. The term TA also is used when firms, consultancies and technological institutes, agencies in health care and other social sectors, attempt to assess the promise, and potential profit, of new technological options, and/or perform a broad version of cost-benefit analysis, sometimes also including risks. In both cases, TA includes efforts to anticipate future developments (of technology, and its relations with markets and society), and it also involves the feedback of such anticipations into relevant decision arenas. This combination is the defining characteristic of TA exercises.

Context

Practical and political concerns with the uncertainties involved in developing and applying new technologies were the breeding ground for TA in the 1960s. One response to these concerns was an attempt to manage the uncertainties through a combination of technological forecasting and the development of techniques to address and evaluate impacts. US organizations like the RAND Corporation pioneered such TA techniques. This managerial response sometimes was broadened to include wider societal impacts; for example, the impacts of the US space program were explored by making comparisons with the impact of the railroad on American society.

The other response was policy-oriented and political, taking the emerging concern about the negative impacts of technology, for example, on the environment or in relation to defense and security issues, as a starting point for attempts to evaluate such impacts at an early stage. Ways to do so included the techniques developed as part of a managerial approach, but also were able to build on policy analysis and on action-oriented research.

Special agencies were made responsible, or even created, for such a task, as when in 1970 in the USA, a federal law established a Congressional Office of Technology Assessment (OTA), as "an aid in the identification of existing and possible impacts of technological application." Congress also was interested because such analytical support would increase its power vis-à-vis the Executive Branch. The example of OTA was noticed in European countries, and parliamentary TA agencies also were established there, after some delay, the specific format depending on the political tradition of each country.

By the late 1980s, TA was a recognized activity, professionally (cf Porter et al., 1980) and politically. Gradually, the earlier and almost exclusive emphasis on analytic approaches and the production of reports was complemented by an interest in linking up more closely with decision making, or at least contributing to setting an agenda. Public debates about energy and environmental issues helped give more prominence to this aspect of TA.

By the 1990s, a philosophy of TA had become widely accepted (even if it was not always formulated explicitly). This philosophy stressed the importance of anticipating impacts and giving feedback into choices and decisions, in order to reduce the human and social costs of learning how to handle technology in society, as compared to when this happens through trial and error. The experience of a series of chemical and nuclear accidents, as well as the uncertainties about the new biotechnology (and life sciences more generally), since the mid-1970s, contributed to the legitimacy of attempts to anticipate and avoid errors, up to disasters. This philosophy covers what would otherwise appear to be a wide variety of approaches, ranging from *technology forecasting* to *consensus conferences*. The philosophy is also visible in discussions about *technology policy*, when it is suggested that technology policy should include a strong TA component. Actually, in the 2000s, government policies on genomics and nanotechnology to an extent do incorporate TA, including attention to ethical, legal, and social aspects and so-called upstream public engagement.

Main Strands in the Variety of TA Approaches and Activities

At least since the 2000s, five strands of TA can be distinguished, each with its own style and addressed to different audiences:

1. TA in firms and in technological institutes, oriented toward mapping future technological developments and their value to the firm or institute, and used as an input in

strategy development (Hastbacka and Greenwald, 1994). 'Picking the winners' (or 'avoiding the losers') used to be the overriding orientation. This strand of TA has developed relatively independently of 'public domain' TA, but links are emerging because of the need of firms to take possible societal impacts and public acceptance into account. The *biotechnology* sector by the late 1990s recognized this because of the impasse about genetically modified organisms in agriculture. Thus, the 'insiders' are now interacting with the 'outsiders,' and the effect is that prudent firms are receptive to public-interest assessments.

2. TA in specific sectors, like environment or health care and regulation, as an instrument to determine acceptability of a project, new product, or technological option, in relation to the public interest in, and policies about, that sector. Environmental impact statements are now required in many countries before projects can be implemented (Vanclay and Bronstein, 1995). Medical and health care TA is a recognized specialty with its own international society and meetings, focusing on evidence-based evaluations of concrete medical and health care activities and options (cost-benefit analysis, risk analysis), not on wider societal impacts.
3. TA for policy development and political decision making about projects or broad programs with a strong technological component (for example, the electronic superhighway or modern agriculture) or important technologies (for example, genetic modification). One can call this 'public service' TA, and consider the US OTA (abolished in 1996) as the embodiment of this type of TA. OTA, during its lifetime, developed a robust approach to TA studies (see Wood, 1997). Other TA bodies that serve national parliaments and/or national governments now sometimes include participatory TA methods in addition to expert-based approaches.
4. TA exercises can be oriented to the public arena more generally, with a focus on articulating and building an agenda for handling new technology in society. This strand is influenced by the increasing calls for *participation* (at least by so-called new stakeholders like environmental groups). While it is particularly visible and more or less institutionalized in European countries like Denmark and the Netherlands, participatory methods such as consensus conferences have been taken up all over the world. *Agenda-building* TA has a longer history, however, given that controversies over new projects or new technologies (and studies and documents produced in the course of such controversies) induce learning about potential impacts and articulation of the value of the technology. Agenda-building TA thus merges with informed consultation processes to reach agreement on the value of new technology. The implication is that there is overlap between TA and more general political and policy approaches for articulation and learning.
5. A further strand of TA, which can be called bridging TA, aims to bridge the distance between ongoing innovation and anticipation of social aspects and impacts, and to do so in real time. In Constructive TA (CTA) (Schot and Rip, 1997), it is recognized that the impacts of new technology are jointly produced by the technology, the actors involved, and wider interactions. This strand of TA leads to exercises in coconstruction of technology development and its embedding in society with different stakeholders for older as well as newer technologies. Independent of CTA, such activities were already taking place when the introduction of new information and communication technologies, both within and between organizations, was broadened by having more user involvement in design. Real-time TA (RRTA) (Guston and Sarewitz, 2002) uses a variety of tools and approaches to anticipate and deliberate interactively. Both CTA and RRTA have developed sociotechnical scenario methods to address newly emerging technosciences like nanotechnology. Societal experiments with the introduction of new technologies, such as electric cars, which involve both private and public actors, are occasions for learning about new technologies, and can be seen as a further version of 'bridging' TA.

It is clear that, over time, the first three strands, all of which are strongly professionalized, have been complemented by broader strands focusing on interaction, deliberation, learning, and agenda building. Possible next steps will be discussed in the concluding section.

Methodological Issues

Earlier, methods of TA emphasized data collection and analysis about present impacts and trends. They also may involve counterfactual speculation, consultations of experts to assist with the refinement of the results, and/or the construction of sociotechnical scenarios. By now, the toolkit of TA contains a wide variety of tools, ranging from forecasting supplemented with social and sociotechnical mapping, multicriteria analysis and integrated assessment, to mediation and communication techniques. The selection of appropriate tools is often eclectic, related to the sponsor and domain of the exercise: private/promoter, sectorial/regulator, public policy, or the public domain more generally. While there are recognized TA professionals, TA exercises are carried out by a wide range of consultants, academics, members of dedicated agencies, their committees, and volunteers.

The combination of anticipation and feedback, characteristic of all varieties of TA, entails a methodology that combines writing a history of the future, supported by judgments of experts and by social science insights and data, and informing action or preparation for action. Relevant social science insights derive from the sociology and economics of technology and innovation, and organizational sociology, but also general sociology and political science. A key notion is the coevolution of technology and society. Estimating impacts is speculative, of course, but the speculation is controlled by data, scholarly insights and experience. Uptake into action remains precarious, exactly because of the fallibility of such estimates: who could have predicted the further development and impact of the Internet in the early 1990s? On the other hand, efforts at foresight can be productive even when they turn out not to be correct, because they stimulate joint learning.

In the literature, a fundamental dilemma has been identified, which is called the anticipation and control dilemma (Collingridge, 1980). At an early stage of technology development, the nature of the technology and the articulation of

interests in it are still malleable, but it is unclear at that stage what the effects and impacts will be. By the time these consequences become clear, the technology is entrenched and vested interests make it difficult to change it. If the technology turns out to be undesirable, as happened with the pesticide DDT, the only possible response then is to forbid further deployment of it. More is possible, however, than in this stark version of the control dilemma.

A generic strategy to avoid such decision regret is to keep flexible for as long as possible, at the risk of failing to pursue any of the various options for sufficiently long to see what they may be worth. An interesting example of such a flexible strategy is the 1991 French Law on nuclear waste handling, which requires three options to be maintained for a period of perhaps 15 years, and with annual public evaluation of progress and prospects. The Parliamentary *Office d'évaluation des choix scientifiques et techniques* is involved, and, over time, modifications can be made to the trajectories. (Further background considerations are offered by Callon et al., 2009.)

Recent research in the economics and sociology of technology has shown how path dependency occurs over time in technological development, and how coproduction of impacts is shaped in those paths. The QWERTY keyboard of typewriters, and now also of computers, is a well-known example of path dependency. Path dependencies can be seen as a problem for TA, because they reduce flexibility to go into more desirable directions. But understanding the dynamics that generate path dependency allows better anticipation of future developments, as well as identification of opportunities for them to be modulated to some extent (Garud and Karnøe, 2001).

Situating TA in Contexts, and Future Directions

Government technology policy most often focuses on promotion of (selected) technologies. The main question for such technology policy is to 'pick the winners' now for society as a whole. Their actual development requires anticipatory coordination, in which versions of TA (particularly 'bridging' TA) can contribute. This is what one can see happening in the European Technology Platforms stimulated by the European Commission. On the other hand, other agencies of the same government may be occupied with reducing the human and social costs associated with the introduction of new technology, for example, through occupational health, safety and environmental regulation. This conflict between promotion and control of new technology is characteristic of the way modern societies have come to handle technology. It is not only reflected in the division of labor between government agencies, but also in cultural and political views, as in the assumption that there will be proponents and opponents to any given new technology (Rip et al., 1995). Earlier, TA was liable to become a victim of such conflicting views when technology promoters condemned TA as harassment of progress. While contestation will continue, there is now more interest in constructive approaches. And there is a move, starting with nanotechnology, to ask promoters of new technology to become proactive in taking possible impacts into account. In European policy discourse, this move has acquired its own acronym RRI, for Responsible Research and Innovation (Fisher and Rip, 2013).

One key point, related to divisions of labor in modern society, is the asymmetry between 'impactors' (those at the source of impacts) and 'impactees.' The asymmetry can be due to a difference in power, but it also always involves a difference in timing. Initiators of technological development know more, and have invested more, at an early stage, and impactees and TA agents have to wait and, in a sense, follow their lead.

The asymmetry has another component: technology developers are 'insiders' and do not necessarily know very much about what happens on the 'outside' after a technology is developed. Adoption and diffusion, however, are up to 'outsiders' who typically have different knowledge, interests, and expectations than developers. The story of nuclear energy, since the 1960s, is in part a story of a struggle between insiders and outsiders. At a much smaller scale, the same storyline is visible in the development of cochlear implants for deaf people, where it turned out – unexpectedly for the insiders – that sizable factions of the deaf community were very negative about taking congenitally deaf children out of the sign-language culture of their deaf parents by providing them with implants (Reuzel, 2004).

Another key point is about cumulative effects of living in a technological society. Risk repertoires have emerged in which the impacts of new technological developments are treated similarly to those of earlier technologies. In the 1970s and early 1980s, the risks of recombinant DNA research were imagined in terms of scenarios of runaway organisms and addressed with probabilistic risk analyses, exactly as previously had been accepted for runaway accidents in nuclear power plants. Accordingly, the measures that were promoted focused on containment. A similar conception of risk is visible in concerns about engineered nanoparticles.

Anticipation and feedback occur all the time. Early warnings of hazards and controversies over new technologies are examples, and may be starting points for TA exercises (cf Harremoës, 2001). We live in a *risk society* and modernity has become reflexive (Adam et al., 2000). The rise and subsequent broadening of TA is then an indicator of *reflexive modernity*, and a conduit to enhance reflexivity – while also chaneling it in certain ways. There were criticisms of the instrumentalization of reflexivity and possible technocratic tendencies of TA in the 1970s and 1980s, as well as of the symbolic nature of participatory TA exercises in the 1990s. These tendencies are still present, but TA practitioners now are more conscious of them.

A further development is the increasing acceptance of the so-called *precautionary principle* (since the Rio Declaration of 1992): regulatory action is justified even if there is not yet conclusive evidence about risks, or negative impacts more generally. Precaution is actually predicated on speculative scenarios about what might happen, and TA can offer quality control of such scenarios.

Clearly, TA is not just a new professional and managerially oriented activity (since the late 1960s) and an approach to interactive agenda building (since the late 1980s). It is part and parcel of larger developments, and may need to develop further in order to contribute to the challenges involved (Rip, 2012). What is happening anyway is the increasing involvement of ethicists in TA, which is important but can also lead to a reduction in complexity if principle-based ethics and checklist approaches are followed. There is also the increasing interest in what has been called 'strategic intelligence' about technology in society, for

which TA would provide one of the inputs. The notion was developed in the context of technology policy (cf Kuhlmann, 2003), but its use is broader, since all sorts of actors, including civil society organizations, are interested in such strategic intelligence. TA will then become a pervasive feature of late modern, reflexive society, rather than a dedicated professional practice. The professional input will remain important, however, to allow for quality control.

See also: Innovation and Technological Change, Economics of; Reflexivity in Science and Technology Studies; Risk, Sociology and Politics of; Science and Technology Studies: Experts and Expertise; Technological Change and the Sociotechnical System, Applied Psychology of; Technological Determinism; Technology and Organization.

Bibliography

Adam, B., Beck, U., Van Loon, J. (Eds.), 2000. The Risk Society and beyond: Critical Issues for Social Theory. Sage, London.

Callon, M., Lascoumes, P., Barthe, Y., 2009. Acting in an Uncertain World: An Essay on Technical Democracy. MIT Press, Cambridge, MA.

Collingridge, D., 1980. The Social Control of Technology. Frances Pinter, London.

Fisher, E., Rip, A., 2013. Responsible innovation: multi-level dynamics and soft intervention practices. In: Owen, R., Bessant, J., Heintz, M. (Eds.), Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society. Wiley-Blackwell, Chichester, pp. 165–184.

Garud, R., Karnøe, P. (Eds.), 2001. Path Dependence and Creation. Lawrence Erlbaum Associates, Mahwah, NJ.

Guston, D.H., Sarewitz, D., 2002. Real-time technology assessment. Technology and Society 24, 93–109.

Harremoës, P. (Ed.), 2001. Late Lessons from Early Warnings: The Precautionary Principle, 1896–2000. Environmental Issue Report No. 22/2001. European Environment Agency, Copenhagen. http://www.eea.europa.eu/publications/environmental_issue_report_2001_22 (accessed 13.04.14.).

Hastbacka, M.A., Greenwald, C.G., 1994. Technology assessment – are you doing it right? Arthur D. Little – PRISM (Fourth Quarter), 35–45.

Kuhlmann, S., 2003. Evaluation as a source of "strategic intelligence". In: Shapira, P., Kuhlmann, S. (Eds.), Learning from Science and Technology Policy Evaluation: Experiences from the United States and Europe. Edward Elgar, Cheltenham, pp. 352–379.

Porter, A.L., Rossini, F.A., Carpenter, S.R., Roper, A.T., 1980. A Guidebook for Technology Assessment and Impact Analysis. North-Holland, New York.

Reuzel, R., 2004. Interactive technology assessment of paediatric cochlear implantation. Poiesis & Praxis 2 (2–3), 119–137.

Rip, A., 2012. Futures of technology assessment. In: Decker, M., Grunwald, A., Knapp, M. (Eds.), Der Systemblick auf Innovation: Technikfolgenabschätzung in der Technikgestaltung. Edition Sigma Verlag, Berlin, pp. 29–39.

Rip, A., Misa, T., Schot, J.W. (Eds.), 1995. Managing Technology in Society: The Approach of Constructive Technology Assessment. Pinter Publishers, London.

Schot, J., Rip, A., 1997. The past and the future of constructive technology assessment. Technological Forecasting and Social Change 54, 251–268.

Vanclay, F., Bronstein, D.A. (Eds.), 1995. Environmental and Social Impact Assessment. John Wiley & Sons, Chichester.

Wood, F.B., 1997. Lessons in technology assessment methodology and management at OTA. Technological Forecasting and Social Change 54, 145–162.

Technology, Evolution of

Stephen Shennan, UCL Institute of Archaeology, Gordon Square, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

An evolutionary approach to the history of technology starts from the position that technologies change as a result of processes of 'descent with modification' in Darwin's famous term; it is in this sense that technology is cumulative. Taking this perspective makes it possible to go beyond the tautology that superior technologies spread because they are superior. It introduces a range of transmission mechanisms with different consequences and an array of different forces that affect what is transmitted. These are described and their implications are illustrated by means of examples.

Introduction

The study of technology is a field in which evolutionary ideas have been current for many years, since the idea of a kind of Darwinian competition between different means of achieving the same practical ends is one that seems obvious, and indeed has been very fruitful. However, when we start trying to implement an evolutionary approach more rigorously it turns out to be more complex than we may initially suppose. Indeed, one of the important benefits of taking an evolutionary perspective is that it goes beyond relatively simple ideas of competition and technological improvement, introducing a range of other forces whose impact is not often considered. In this, it meets the requirements for any worthwhile approach to technology: first, to go beyond the tautology that superior technologies spread because they are superior; second, to account for the fact that while the history of technology is famously cumulative, in that innovations build on the continuing existence of features that have resulted from earlier innovations, the process can also go into reverse; and third to account for the fact that, "people in even the simplest human societies depend on tools that are beyond the inventive capacity of individuals" (Boyd et al., 2013: 120), the technologies of the Inuit in the Arctic being a famous example.

In what follows I will first outline the framework adopted here for viewing technology from a Darwinian evolutionary perspective as a domain characterized by the generation of variation, by mechanisms for its inheritance, and by selection and other mechanisms that affect differentially what is transmitted through time, leading to 'descent with modification'. I will then outline a number of studies that illustrate these different processes and address the question of whether an evolutionary theory of technology does indeed meet the requirements listed above.

What Evolves?

There are many different definitions of technology. For example, Lemonnier (1992: 1) proposes that technology "embraces all aspects of the process of action upon matter, whether it is scratching one's nose, planting sweet potatoes, or making jumbo jets", while Mesoudi et al. (2013) define it as "goal-oriented shared knowledge together with its external means. It is thus knowledge geared to, and organized around, solving specific problems faced by a group or society". The definition preferred here though is that of Arthur (2009), who defines a technology as "a phenomenon captured and put to use". This avoids the suggestion of teleology present in the previous definition. Inventions need not be made in the context of solving problems but may occur more or less by chance in the context of activities unrelated to the domain in which they subsequently turn out to have an application. Moreover, even in apparently very simple areas of activity there will often be a variety of different ways of putting a phenomenon to use. The French anthropologist Marcel Mauss noticed in World War I that French and British soldiers had different types of spades and shovels for digging trenches and were unable to use each other's tools effectively because they involved entirely different kinds of bodily movement (Mauss, 2006[1935]).

Whatever the definition though, the key question for any Darwinian approach to technology is, what are the things that are subject to 'descent with modification'? The answer in terms of the definition adopted here is that the entities that are the subject of variation, inheritance, and selection processes are technological lineages, recipes for techniques, routines, and practices linked by ancestor–descendant relationships (O'Brien et al., 2010), that capture and put to use specific phenomena in particular ways, based on specific areas of knowledge. Thus, ceramic vessels, for example, put to use one set of principles to make containers but barrels or plastic bowls make use of very different ones. A lineage, or rather several different initially independent lineages, of techniques or practices for the production of ceramic vessels can be traced for thousands of years up to the present. Plastic vessels put to use very different phenomena requiring the application of techniques and practices that have their own lineages in the history of the chemical industry. The replacement of ceramic vessels by plastic ones in some domain represents a replacement of one lineage by another, but we should not automatically assume that this means it is superior; such a claim requires testing in relation to alternatives. Indeed, it cannot be emphasized too strongly that an evolutionary theory of technical change does not require inevitable increases in superiority.

The Processes of Cultural Evolution

Variation in techniques and practices is generated by invention, which may involve the appearance of some novel traits or

novel combinations of traits that exist already (Lewis and Laland, in press, Enquist et al., 2011). It may simply represent unintended copying error but may also involve trial-and-error experimentation. However, there is no guarantee that goal-directed experimentation will result in a superior outcome, as the history of failed inventions shows. The concept of invention is common to all approaches to technology. What is distinctive of the evolutionary approach is the wide range of forces that are postulated to affect the transmission of the invention once it has occurred. In the case of technology, as with all aspects of culture, the transmission mechanism is social learning: people learn ways to think and act from others. Of course, the routes through which culture is transmitted are much more diverse than those for genes (Cavalli-Sforza and Feldman, 1981), and different routes, for example parent to offspring or teacher to class, have different consequences for the patterning of cultural change through time. It has also been shown that the processes of social learning themselves are varied, with some leading to greater fidelity in transmission than others, imitation and teaching in particular (Dean et al., 2012; Kline et al., 2013).

If no other forces are acting on what is transmitted ('unbiased transmission') then the relative frequencies of the different technological variants ('ways of doing') in the population after transmission will in principle be the same as before, except that in any finite population some variants will be copied slightly less often and some slightly more, purely by chance. Perhaps surprisingly, this 'unbiased' process will have major effects on the frequencies when it is repeated many times, leading to the disappearance of some variants and to others attaining high frequencies even when variants are functionally equivalent (see Bentley et al., 2004 for a worked example). This is the cultural equivalent of genetic drift. It can easily be appreciated that in these circumstances any invention appearing for the first time will have a high probability of disappearing almost immediately, even if it is potentially useful. However, in many cases there will be other forces acting. One is natural selection, in the sense that those individuals who inherit or acquire certain technological attributes may have a greater probability of surviving and/or reproducing than would those who do not; this process will continue, leading to their increased prevalence, so long as those possessing the attributes pass them onto their children. For example, in many parts of the world, adopting an agricultural rather than a hunting-and-gathering way of life clearly led to greater reproductive success (though not to better health (Lambert, 2009)), reflected in the fact that it led to increased population densities; as a result, the technical traits that characterize agriculture spread and, in some cases, subsequently influenced genetic evolution (e.g., the ability to digest lactose: Bersaglieri et al., 2004).

In addition to selection, a number of what have been termed 'bias processes' (Boyd and Richerson, 1985) can affect whether inventions are transmitted; these refer to factors that affect what and who people try to copy when they are learning from others. Thus, 'results bias' refers to the situation where people look at what other people do, for example, the crops they plant, compare the results with what they are doing themselves, and then change what they do because the other way of doing things seems to be more effective. 'Content biases' are affected by features of transmissible phenomena that make them intrinsically more or less memorable. 'Context biases' are aspects of the context of learning that affect what is transmitted; thus, something may be copied simply because the person initially doing it is prestigious ('prestige bias') or because it is what most people locally do ('conformist bias'). In these latter two cases, whether a particular technical attribute or practice that becomes more prevalent in a population has nothing to do with its intrinsic properties as such but only with the context of learning, though with the underlying assumption that adopting such a decision rule is likely to be a good one in practical terms where there is less clear evidence one way or the other about which of a set of variants is superior. It is worth adding that even technologies that are beneficial in cost–benefit terms are not immune to chance loss, drift effects, when populations are small, as shown by Rivers (1926) for canoes in one of the Pacific islands (and see below).

Finally, it is important to be aware that particular technical skills and practices that are favored by selection and bias processes at a given time will not necessarily continue to be favored if there are changes in aspects of the external environment, including humanly induced ones, even if they are not directly connected to the skills themselves. They are then likely to be vulnerable to loss. Thus, Reyes-García et al. (2013) have shown that the loss of indigenous knowledge of wild-plant use among the Tsimane, an indigenous group in Amazonia, is greater among men than women and among those who live nearer towns. They suggest that this may be because such knowledge is increasingly irrelevant in the new socioeconomic and cultural conditions in which people find themselves.

Technological Lineages and Human Populations

A key point to reiterate here is that the object of the evolutionary analysis of technology is technological lineages, not human populations, and because they are reproduced in different ways there is no reason in principle that they should run in parallel even though they are obviously linked. Thus, to understand trajectories of technological change we need to do two things: (1) we have to address histories of technologies; (2) we have to examine the histories of the human populations through which they are transmitted. With regard to the first, it is necessary to identify histories of technological transmission to show that an ancestor–descendant relationship exists, if indeed it does. Continuities or discontinuities through time may simply reflect contingently optimal responses to stable or unstable local ecological conditions. In any case, assuming that a real technological lineage has been identified, we must then attempt to understand the forces shaping it. Second, we need to independently understand the histories of the *relevant* populations (see below). The possible linkages or feedback loops between technologies and populations are many and varied. A technology may or may not have a positive effect on the survival and reproductive success of a human population while replicating successfully itself. Even if it does, its effects may be overwhelmed by other factors, such as the extinction of the population for other reasons. If a technology is vertically transmitted with respect to its human population (i.e., from parent to child) then it may be particularly vulnerable in this respect. It perhaps goes without saying that few if any studies

have taken all these different elements of the evolutionary analysis of technology into account (for recent proposals on how to approach this complex problem, see Beheim and Baldini, 2012; Mesoudi et al., 2013: 202–204).

Invention and Innovation

Some inventions are relatively obvious and have therefore been made again and again. For example, it seems that wherever in the world people came to depend to a significant extent on collected seeds for food, they invented and used grinding stones, since these are very effective tools for processing seeds into food and the cost of producing them is far outweighed by the subsequent benefits gained. But it is by no means necessarily the case that the appearance of new conditions will result in inventions that take advantage of them.

In this connection, it is worth referring to Perkins's (2000) contrast in the space of technological search between 'Homing spaces', where there is a clear gradient pointing toward a viable outcome, and 'Klondike spaces', where high-yield nuggets are sparsely distributed and there are few signs to indicate when you are near one of them. This idea may be linked to the metaphorical concept of 'fitness landscapes' (Kauffman, 1993; Kauffman et al., 2000), where peaks and troughs relate to areas of high and low fitness, respectively. O'Brien and Bentley (2011) see the example of spark arrestors on the smokestacks of the US locomotives in the nineteenth century in these terms. More than 1000 patents were taken out to solve the problem of trains causing fires while keeping the draft as unobstructed as possible. In this case there was no entirely successful solution, and this new adaptive space remained a rugged fitness landscape with a number of more or less equally fit local optima or peaks.

Previously empty or newly created niches offer new possibilities for innovation, either because there is no previous history from which to learn or because existing solutions may not work in the new niche. On Kaufmann's adaptive radiation model, whether biological or technological, there is an initial stage in which there are long jumps to new places in the fitness landscape, a second stage when the broad landscape has been explored and local 'hill climbing' occurs, because earlier choices constrain subsequent options in a process of so-called 'generative entrenchment' (Wimsatt, 1986), and a third stage of relatively little change where adaptations to local peaks cannot be improved but selection between different peaks begins to occur. Lake and Venti (2009) followed through the work of Van Nierop et al. (1997) and showed that the evolution of the bicycle did indeed follow this model.

When technologies involve complex and difficult production processes then they are far less likely to be reinvented if lost, simply because of that difficulty. Moreover, Roux (2010) makes the point that training in any production process that involves the acquisition of high levels of expertise that take a long time to acquire is likely to restrict innovation, because the whole process of learning is designed to ensure high-fidelity transmission, to fix particular sets of skills and knowledge (cf Martin, 2000); this may be particularly the case with physical expertise and motor skills. Only the most expert, those who have complete control of all aspects of a process and its associated knowledge, are likely to transcend the limits of what they have learned and invent something new. Roux cites (2010: 224) the example of inventions in the field of pyrotechnology by traditional craftsmen in India, which were only made by those who were most skillful, individuals 'exceptional as much for their skills as for their rarity'.

But Roux makes a further and more specific point relevant to the evolutionary trajectories of the more complex technologies, not least copper and iron production, that have emerged in the last 7000–8000 years. If the transmission of technologies based on novel principles involves a long apprenticeship that excludes the possibility of becoming correspondingly adept at other skills then it will almost inevitably result in a closed and more or less fragile technological system, at least to start with, because the effective population size of those involved in the activity in these circumstances will be very small. First, it will only be possible for relatively small numbers of individuals to undertake the apprenticeship. Second, not all societies will be able to sustain the required division of labor. Third, for various reasons, including people's desire to protect their livelihood, they are likely to keep their knowledge and expertise secret and either transmit it vertically to offspring or charge a significant entry fee. In these circumstances it may only take localized historical events, such as the destruction of local elites that supported specialists, to put an end to a technology altogether even if it was a functionally effective one, as Roux shows in the case of wheel-turned pottery in the ancient Near East.

A Case Study from Early Technology

How such processes work in the case of iron production has recently been examined by Charlton et al. (2010), who have taken an evolutionary approach to understanding preindustrial bloomery iron-smelting technology in a case study from north west Wales. Despite the technical complications of addressing this subject, it has the advantage that the problem can be clearly framed. There is no doubt about the goal (at least in general terms) and the conditions required to successfully smelt iron are well understood, arising as they do from universal properties of the materials involved. The bloomery method involves "the solid state reduction of iron oxides into a spongy mass of iron, called a bloom, and the production of a ferrosilicate slag" (Charlton et al., 2010: 353), based on the combination of iron ore and charcoal in a furnace supplied with air, where temperatures can reach up to 1400 °C, which is still below the melting point of iron. It was invented and diffused in several parts of the world and it seems safe to say that its spread and effective use did not involve any causal knowledge of the chemistry involved (cf Mokyr, 2000). Indeed, insofar as people formulated causal principles they were almost certainly completely mistaken.

Following the procedure above for the study of a technological lineage, the first issue is to identify patterns of cultural descent in the methods used and to distinguish variation arising from transmission from that relating, for example, to the local ore or fuel type; and then to characterize the forces affecting that variation, in a situation where, by the very nature of the process, there are only a limited number of successful solutions. The most informative source of information on the

processes involved in past episodes of early iron production is chemical variation in the slags produced as a waste product. In this case then the data are quantitative variations in the chemistry of chronologically ordered slag deposits and the problem is: can we establish whether or not there is a signal of cultural descent in the chemical variation? And if so, what can we infer about the factors affecting the transmission processes that produced it?

Charlton (2009) showed convincingly that a transmission signal could be identified in the slag chemistry, distinct from the effects of ore and fuel composition. In terms of the forces acting on the technical knowledge and practices passed on from one iron producer to another, it is easy to imagine that there might be some more or less random variation in exactly what was done each time. It is also likely that there would be strong selection for those practices that were successful, though, given the complexity of the process and its many stages, it would not necessarily be easy to identify precisely what produced a successful smelt on any given occasion. From the viewpoint of the smelters it is thus likely that transmission would be affected by results bias, albeit in a very noisy form, based on characteristics that they could observe and the connections they were able to make between variation in those characteristics and their techniques of ore preparation, furnace operation, etc. From the viewpoint of the smelting recipes as opposed to the smelters, incidentally, it could be argued that it represents a process of selection, since recipes would be differentially reproduced depending on their ability to successfully smelt iron. The results of Charlton et al. (2010) suggested that initially all changes related to furnace operation could be accounted for by a drift process but that at a certain point a second effective procedure was more or less accidentally discovered and a decision was taken to make use of the two distinct procedures, visible in consistently different slag signatures, one of lower yield than the other but producing a higher quality product in the form of steel. At the same time there were clear trends in the use of manganese-rich ores with better fluxing capabilities, and evidence of decreased variability in reducing conditions related to results bias: that is to say the iron makers consistently reproduced the airflow conditions that gave the best results for a given recipe.

However, this was a very small operation, using only a single furnace, and at present we do not know if the discovery of these Welsh iron smelters was also independently made by others, or if it was passed on outside this group. Roux's points regarding the closed and fragile nature of transmission networks in the context of preindustrial craft specialization would lead us to hypothesize that innovations made by this group of smelters could easily have been lost. It goes without saying that in this case we have no idea how these changes in the iron-production lineage impacted the survival and reproductive success of the smelters themselves since we have no evidence on this matter.

The subtle ways in which selection operated in the context of early complex technologies, often leading to optimal solutions, is also illustrated by Jackson and Smedley's work on medieval glass production in northern Europe (2008). Here bracken was used as a main source of plant ash for glass production. Jackson's analysis of the chemical composition of bracken ashes showed that it changes over the course of the growing season, with the concentration of SiO_2, CaO, Fe_2O_3, and MnO increasing throughout the season and K_2O and P_2O_5 peaking in mid-June and then declining. This makes mid-June the best time to harvest bracken since the alkali concentrations necessary to produce glass are high, and other components that produce less satisfactory outcomes, like coloring, are still low. The appreciation that this was the case, obviously without any knowledge of the relevant chemistry, was encapsulated in the recommendation to be found in contemporary documents about glass making that for the best results bracken should be harvested by the feast of St John the Baptist, 24th June.

Such examples show that we can realize the potential of cultural evolutionary theory for going beyond the tautology that superior technologies spread because they are superior. We can test hypotheses about the forces acting in any particular case. In this we are aided by the fact that there are independent engineering criteria for a technology's relative effectiveness, which can be established by experiments if they are not already known (cf Boyd et al., 2013); experiments can also help us to investigate the costs as well as the benefits by means of optimality studies. Time invested in producing complex technologies is time that could potentially be spent in other useful activities; they have an 'opportunity cost'. Bettinger et al. (2006) provide a productive model of technological intensification that addresses these cost–benefit issues and shows how they illuminate the use of different foraging technologies in prehistoric California. Of course, material costs are also relevant. An interesting example of where increasing costs have changed the direction of selection back to a simpler technology is the resurgence of animal traction for agriculture in parts of Africa as the costs of fuel and machinery have increased relative to local incomes (see www.atnesa.org).

The Cumulative Evolution of Technology

For cumulative evolution to occur, there needs to be some sort of balance between innovation and the rate of loss of existing technological attributes, determined at least partly by transmission fidelity. Obviously, if innovations are not retained then there is no possibility of accumulation and the same is true if fidelity is so high that innovations never occur. A simulation of a simple cultural transmission model by Lewis and Laland (in press) showed that the most significant factor affecting the build up of cumulative culture was cultural loss, and that a major difference was made by minor improvements in the fidelity of transmission, which led to increased longevity of cultural traits, while forms of innovation involving the combination of existing traits also played a significant role, much more so than the appearance of complete novelties.

In the case of genetic variation it is basically the case that mutations are random, though, of course, different genes have different mutation rates. In the case of cultural inventions it seems clear that rates of innovation in specific areas have varied over time. One tendency has been to take the 'necessity is the mother of invention' view of these, in which, when people are in a situation which they know is not sustainable, they are more likely to take risks and carry out trial-and-error experimentation (Fitzhugh, 2001). As Henrich (2010) points out, however, experimental evidence tends to be against this, and the risks involved are certainly very considerable, from devising

alternatives to embarking on them without prior knowledge of the costs and benefits. The contrasting view, not that they are mutually exclusive, is that they are more likely to arise simply as a result of the existence of larger effective populations, in terms of absolute size, degree of networking or both, since these are likely to increase the innovation rate per unit time and to increase the potential for the new combinations of existing elements that what Roux calls 'continuous inventions' generally represent. We have noted already that the collapse of populations and social systems leads to the loss of complex technological lineages that have only a few specialist bearers but it is now time to examine the issues involved in more detail.

The best-known evolutionary model that looks at the implications of specific innovation and transmission processes and their link to demography is that of Henrich (2004). This is based on a model of learning in which individuals attempt to imitate the most skilled individual available to them. Most individuals will fail to reach that level, so there will be a distance between the modal skill level in the population and the best. However, there will be a certain probability that an individual will exceed the current best and a new, higher, most skilled level will result. In these circumstances, given a constant gap between the best level and the modal level, the result will be an improvement in the modal level. The probability of improvement or decline in the population modal level depends on the ratio of the gap between the mode and the best, on the one hand, and the dispersion of the distribution of skill levels on the other. If the gap is small and the dispersion is large then improvement will be far more probable than in the opposite case. For a given ratio, though, Henrich shows that what happens will depend on population size, since improbable events will occur more often in larger populations. For very complex technologies where the gap between the modal level and the best is large and the variation between individuals is small, then a large population will be required to maintain or improve the technology. Thus, if external forces cause a decline in the population size the most complex technologies are likely to be lost. Simpler ones, even if they are lost, have a much higher probability of being reinvented. Henrich showed that this process could account for the well known and apparently deeply puzzling cultural impoverishment of the aboriginal Tasmanians after Tasmania was cut off from the Australian mainland at the end of the last Ice Age. More generally, "It is the selective transmission of lucky errors and occasional experiments that drives much of the evolution of adaptive technology, skills, beliefs, and practices" (Henrich, 2004: 202), and makes it strongly dependent on population size. Henrich's result has subsequently been confirmed by Kobayashi and Aoki (2012), who showed that a version of the model with the more satisfactory assumption of overlapping rather than discrete generations produced an even stronger population effect.

As we have seen already, the substantive assumption in Henrich's argument that real improvements are difficult to make because of the sheer amount of cultural knowledge embedded and embodied in human artifacts and cultural practices is supported by the ethnographic and archaeological work of Roux (2010), which also brings out the point that we always have to consider what is the relevant effective population size in any given case. In the example of Henrich's proposal for Tasmania the interacting population size decreases because of the loss of contact with mainland Australia.

As technologies get more complex, they include more elements. Lehmann and Wakano (2013) look at the effect of the multidimensionality of cultural traits on the process of cumulative evolution; in other words, where the best payoff depends on getting the right combination of a number of different features, for example, different aspects of size and shape of an artifact. With increasing numbers of dimensions there are far more ways of making things worse than making them better. Where the number of dimensions is very high then even a learner with very high computational ability going through different combinations and their effects in their imagination will not find the optimum, but a population of payoff-biased learners making uncorrelated errors will be able to reach it if it is sufficiently large. At least until the recent past effective population sizes may not have been large enough and for this reason it may be that many early technologies carried the equivalent of what in genetics is referred to as 'drift load', where deleterious mutations accumulate in the genome (cf Shennan, 2001). In other words, on this line of argument, cumulative technology does not necessarily result in improved performance. An excellent example of the kinds of complexity and multidimensionality involved is provided by Lombard and Haidle's (2012) construction of what they call the 'effective chain' of activities for the production of bows and arrows, based on evidence from Middle Stone Age Africa (see also Lemonnier, 1992 for other examples).

The general argument about the importance of population size for cultural complexity is supported by Kline and Boyd's (2010) study of marine fishing technology in Oceania. This showed a correlation between the number and complexity of marine fishing tools and the size and connectedness of island populations, with smaller and more isolated islands having fewer and less complex tools than larger and less isolated ones. Recently, experimental results have also demonstrated the cultural complexity – population size link (Derex et al., 2013; Muthukrishna et al., 2014).

These examples make the point that though the general human propensity for high-fidelity social learning (e.g., Dean et al., 2012; Haun and Over, 2013) is a prerequisite, specific instances of cumulative technological evolution are actually a property of specific populations or metapopulations linked by transmission processes. Selection will be operating on them through the cost–benefit dimension but, even in cases where the cost–benefit ratio is favorable, transmission failures, including the fact that early networks were often fragile, can overwhelm selection and lead to cultural loss, which is not necessarily reversed by reinvention.

Conclusion

The cultural evolution approach to technology is an extremely powerful one. Far from encouraging the tautologous view that what is prevalent is by definition what has been competitively successful, work in cultural evolution has demonstrated the complexity of making inferences about the forces at work while providing the tools to deal with it, including cost–benefit analyses and mathematical modeling. Moreover, the set of processes it

postulates, involving the generation of variation, its transmission, and the action of various sorting processes on what is transmitted provides a transparent basis for understanding how technology can be cumulative but also how cumulative evolution is specific to particular interacting 'effective' populations and therefore open to both devolution and optimality failures. Who you know is at least as important as what you know (cf Henrich, 2010). Early complex or specialist technologies had a number of intrinsic features that would have encouraged small effective population sizes, in terms of both absolute numbers of practitioners and the likelihood of them sharing information. Finally, the framework also makes clear what it is that is evolving – technological lineages – and thus directs us both to characterizing them and forcing us to think about the relations between them and populations of human agents.

See also: Cultural Evolution: Overview; Cultural Evolution: Theory and Models; Evolution: Diffusion of Innovations; Experiment in Science and Technology Studies; Indigenous Knowledge: Science and Technology Studies; Technological Determinism.

Bibliography

Animal Traction Network. www.atnesa.org (accessed 06.01.14.).

Arthur, W.B., 2009. The Nature of Technology. Allen Lane, London.

Beheim, B.A., Baldini, R., 2012. Evolutionary decomposition and the mechanisms of cultural change. Cliodynamics 3, 217–233.

Bentley, R.A., Hahn, M.W., Shennan, S.J., 2004. Random drift and culture change. Proceedings of the Royal Society Biological Sciences 271, 1443–1450.

Bettinger, R., Winterhalder, B., McElreath, R., 2006. A simple model of technological intensification. Journal of Archaeological Science 33, 538–545.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., Hirschhorn, J.N., 2004. Genetic signatures of strong recent positive selection at the lactase gene. American Society of Human Genetics 74, 1111–1120.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. Chicago UP, Chicago, IL.

Boyd, R., Richerson, P.J., Henrich, J., 2013. The cultural evolution of technology: facts and theories. In: Christiansen, M.H., Richerson, P.J. (Eds.), Cultural Evolution. MIT Press, Cambridge, MA, pp. 119–142.

Cavalli-Sforza, L., Feldman, M.W., 1981. Cultural Transmission and Evolution: A Quantitative Approach. Princeton UP, Princeton, NJ.

Charlton, M.F., 2009. Identifying iron production lineages: a case-study in north-west Wales. In: Shennan, S.J. (Ed.), Pattern and Process in Cultural Evolution. University of California Press, Berkeley, CA, pp. 133–144.

Charlton, M.F., Crew, P., Rehren, T., Shennan, S.J., 2010. Explaining the evolution of ironmaking recipes – an example from northwest Wales. Journal of Anthropological Archaeology 29, 352–367.

Dean, L.G., Kendal, R.L., Schapiro, S.J., Thierry, B., Laland, K.N., 2012. Identification of the social and cognitive processes underlying human cumulative culture. Science 335, 1114–1118.

Derex, M., Beugin, M.P., Godelle, B., Raymond, M., 2013. Experimental evidence for the influence of group size on cultural complexity. Nature 503, 389–391.

Enquist, M., Ghirlanda, S., Kimmo, E.K., 2011. Modelling the evolution and diversity of cumulative culture. Philosphical Transactions of the Royal Society B 366, 412–423. http://dx.doi.org/10.1098/rstb.2010.0132.

Fitzhugh, B., 2001. Risk and invention in human technological evolution. Journal of Anthropological Archaeology 20, 125–167.

Haun, D., Over, H., 2013. Like me: a homophily-based account of human culture. In: Christiansen, M.H., Richerson, P.J. (Eds.), Cultural Evolution. MIT Press, Cambridge, MA, pp. 75–86.

Henrich, J., 2004. Demography and cultural evolution: why adaptive cultural processes produced maladaptive losses in Tasmania. American Antiquity 69, 197–214.

Henrich, J., 2010. The evolution of innovation-enhancing institutions. In: O'Brien, M.J., Shennan, S.J. (Eds.), Innovations in Cultural Systems: Contributions from Evolutionary Anthropology. MIT Press, Cambridge, MA, pp. 99–120.

Jackson, C.M., Smedley, J.W., 2008. Medieval and post-medieval glass technology: the chemical composition of bracken from different habitats through a growing season. Glass Technology: European Journal of Glass Science and Technology Part A 49 (5), 240–245.

Kauffman, S., 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford UP, New York.

Kauffman, S., Lobo, J., Macready, W.G., 2000. Optimal search on a technology landscape. Journal of Economic Behavior & Organization 43, 141–166.

Kline, M.A., Boyd, R., 2010. Population size predicts technological complexity in Oceania. Proceedings of the Royal Society B 277, 2559–2564.

Kline, M.A., Boyd, R., Henrich, J., 2013. Teaching and the life history of cultural transmission in Fijian villages. Human Nature 24, 351–374.

Kobayashi, Y., Aoki, K., 2012. Innovativeness, population size and cumulative cultural evolution. Theoretical Population Biology 82, 38–47.

Lake, M.W., Venti, J., 2009. Quantitative analysis of macroevolutionary patterning in technological evolution. In: Shennan, S.J. (Ed.), Pattern and Process in Cultural Evolution. University of California Press, Berkeley, CA, pp. 147–174.

Lambert, P.M., 2009. Health vs. fitness: competing themes in the origins and spread of agriculture? Current Anthropology 50, 603–608.

Lehmann, L., Wakano, J.Y., 2013. The handaxe and the microscope: individual and social learning in a multidimensional model of adaptation. Evolution and Human Behavior 34, 109–117.

Lemonnier, P., 1992. Elements for an Anthropology of Technology. Anthropological Papers No.88. University of Michigan, Ann Arbor, MI.

Lewis, H.M., Laland, K.N., Transmission fidelity is the key to the build-up of cumulative culture. Philosophical Transactions of the Royal Society B (2012) 00, 1–10, in press.

Lombard, M., Haidle, M.N., 2012. Thinking a bow-and-arrow set: cognitive implications of Middle Stone Age bow and stone-tipped arrow technology. Cambridge Archaeological Journal 22, 237–264.

Martin, G., 2000. Stasis in complex artefacts. In: Ziman, J. (Ed.), Technological Innovation as an Evolutionary Process. Cambridge UP, Cambridge, pp. 90–100.

Mauss, M., 2006[1935]. Techniques of the body. edited by N. Schlanger. In: Mauss, M. (Ed.), Techniques, Technology and Civilisation. Berghahn Books, Oxford, pp. 77–96.

Mesoudi, A., Laland, K.N., Boyd, R., Buchanan, B., Flynn, E., McCauley, R.N., Renn, J., Reyes-García, V., Shennan, S., Stout, D., Tennie, C., 2013. The cultural evolution of technology and science. In: Christiansen, M.H., Richerson, P.J. (Eds.), Cultural Evolution. MIT Press, Cambridge, MA, pp. 193–218.

Mokyr, J., 2000. Evolutionary phenomena in technological change. In: Ziman, J. (Ed.), Technological Innovation as an Evolutionary Process. Cambridge UP, Cambridge, pp. 52–65.

Muthukrishna, M., Shulman, B.W., Vasilescu, V., Henrich, J., 2014. Sociality influences cultural complexity. Proceedings of the Royal Society of London. Series B, 20132511.

O'Brien, M.J., Bentley, R.A., 2011. Stimulated variation and cascades: two processes in the evolution of complex technological systems. Journal of Archaeological Method and Theory 18, 309–335.

O'Brien, M.J., Lyman, R.L., Mesoudi, A., VanPool, T.L., 2010. Cultural traits as units of analysis. Philosophical Transactions of the Royal Society B 365, 3797–3806.

Perkins, D., 2000. The evolution of adaptive form. In: Ziman, J. (Ed.), Technological Innovation as an Evolutionary Process. Cambridge UP, Cambridge, pp. 159–173.

Reyes-García, V., Guèze, M., Luz, A.C., Paneque-Gálvez, J., Macíc, M.J., Orta-Martínez, M., Pino, J., Rubio-Campillo, X., 2013. Evidence of traditional knowledge loss among a contemporary indigenous society. Evolution and Human Behavior 34, 249–257.

Rivers, W.H.R., 1926. Psychology and Ethnology. Kegan Paul, Trench, Trubner, London.

Roux, V., 2010. Technological innovations and developmental trajectories: social factors as evolutionary forces. In: O'Brien, M.J., Shennan, S.J. (Eds.), Innovations in Cultural Systems: Contributions from Evolutionary Anthropology. MIT Press, Cambridge, MA, pp. 217–234.

Shennan, S.J., 2001. Demography and cultural innovation: a model and some implications for the emergence of modern human culture. Cambridge Archaeological Journal 11, 5–16.

Van Nierop, O.A., Blankendaal, A.C.M., Overbeeke, C.J., 1997. The evolution of the bicycle: a dynamic systems approach. Journal of Design History 3, 253–267.

Wimsatt, W.C., 1986. Developmental constraints, generative entrenchment, and the innate-acquired distinction. In: Bechtel, W. (Ed.), Integrating Scientific Disciplines. Martinus-Nijhoff, Dordrecht, pp. 185–208.

Technology, Social Construction of

Wiebe E Bijker, Maastricht University, Maastricht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The 'social construction of technology' (SCOT) is one approach among several constructivist ways of studying science and technology that emerged in the 1980s. SCOT can be used to denote two different things. First, it is a research approach to study technical change in society, both in historical and in contemporaneous studies. And second, it is a theory about the relation between society and technology. This article sketches the genealogy of the social construction of technology by distinguishing mild and radical versions. Within the radical versions, broad and narrow forms are identified; the narrow one is denoted by the acronym 'SCOT.' Historically, the development of SCOT is closely linked to the sociology of scientific knowledge and to the science, technology, and society movement in the 1970s. An important starting point was to criticize technological determinism. In the 1980s and 1990s, it developed primarily as an academic enterprise, but then also found application in the policy domain, in innovation management, and in discussions about new forms of democracy.

Constructivist Studies of Science and Technology

The social construction of technology (SCOT) is one approach among several constructivist ways of studying science and technology that emerged in the 1980s. Here, 'constructivist' means a scholarly orientation that treats the truth of scientific facts and the working of technical artifacts as accomplishments – as being constructed – rather than as intrinsic properties of those facts and machines. The term 'social construction of technology' can be used to denote two different things. First, it refers to a research approach to study technical change in society, both in historical and in contemporaneous studies. Second, it signifies a theory about the development of technology. Historically, the development of SCOT is closely linked to the sociology of scientific knowledge (SSK) and to the science, technology, and society (STS – an acronym later used as much for science and technology studies) movement in the 1970s. In the 1980s and 1990s, it developed primarily as an academic enterprise, but then also found application in the policy domain, in innovation management, and in discussions about new forms of democracy.

The phrase 'social construction' was first used by Berger and Luckmann (1966) in their *Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Building on the phenomenological tradition in philosophy and social theory, and particularly on the work of Alfred Schutz, they argue that social reality is constructed through human actions and that the processes through which it is constructed should be the object of the sociology of knowledge. Berger and Luckmann focus on the social construction of ordinary knowledge of the sort that we use to make our way in a particular society. They are concerned with the reality of social institutions and their focus is on society at large, rather than on natural reality and the subcultures, such as fields of science and technology, that contend with it. Nevertheless, provocative lines of scholarship developed around such themes as the social construction of mental illness, deviance, gender, law, and class. In the 1970s, the theme of social construction was applied to scientific facts (Latour and Woolgar, 1986[1979]), and in the 1980s developed for an approach to artifacts (Bijker et al., 2012 [1987]).

Constructivist studies of science and technology come in a wide variety of forms, both mild and radical (Sismondo, 1993). The mild versions merely stress the importance of including the social context when describing the development of science and technology. Examples in the history of technology include Constant's (1980) account of the turbojet revolution and Nye's (1998) studies of the electrification of America. The radical versions argue that the content of science and technology is socially constructed. In other words, the radical claim is that the truth of scientific statements and the technical working of machines are not derived from nature alone, but are constituted in social processes. Radical constructivist studies of science and technology share the same background and have similar aims, and to an extent studies in both areas have been carried out by the same authors. The remainder of this article will focus on technology studies and more precisely on the radical versions of SCOT.

The Origin and Development of the Social Construction of Technology

SCOT grew out of the combination of three distinct bodies of work: the STS movement, SSK, and the history of technology. The first started in the 1970s, mainly in the Netherlands, Scandinavia, the United Kingdom, and the United States. Often led by science and engineering faculty members, its goal was to enrich the curricula of both universities and secondary schools by studying issues such as scientists' social responsibilities, the risks of nuclear energy, the proliferation of nuclear arms, and environmental pollution. The movement was quite successful, especially in science and engineering schools and faculties, and some of the STS courses were incorporated into degree requirements.

SSK emerged in the late 1970s in the United Kingdom on the basis of work in the sociology of knowledge, the philosophy of science, and the sociology of science. The central

methodological tenets of the 'strong program' (especially its symmetry principle) seemed equally applicable to technology. The symmetry principle called for a similar explanatory approach to scientific knowledge, regardless of whether it was now held to be true or false; in other words, sociological interests and group processes would be used to explain the development and acceptance of 'true' as well as 'false' knowledge. In studying technology, this should mean that working and nonworking machines were to be analyzed symmetrically and in the same terms.

In the history of technology, especially in the United States, an increasing number of scholars began to raise more theoretical and sociologically inspired questions (Constant, 1980; Hughes, 1983; Cowan, 1983). Path-breaking advocacy for this body of work in the history of technology was provided by the reader edited by MacKenzie and Wajcman (1999[1985]). Citing the title of this volume, the phrase 'social shaping of technology' is often used to denote a broader area of research, including work in industrial sociology, technology policy, and the economics of technical change.

Researchers from these three traditions convened an international workshop in 1984 in the Netherlands. The subsequent volume – *Social Construction of Technological Systems* – from that workshop, edited by an STSer, a historian of technology, and a sociologist of scientific knowledge, has been heralded as the starting point of SCOT. On the occasion of the Massachusetts Institute of Technology (MIT)'s 150th anniversary, this volume was included in the list of 30 most influential titles ever published by MIT Press, and it was republished in an anniversary edition 25 years later, with a new preface reviewing the development of the field since 1987 (Bijker et al., 2012[1987]). To understand the role of this workshop and volume, it is helpful to distinguish between a broad and narrow usage of the phrase 'social construction of technology' (but do note that both notions fall within the radical meaning of social constructivism).

When broadly used, 'SCOT' encompasses all the work represented in the 1987 volume, including the actor–network approach by Callon, Latour, and Law, and the technological systems approach advocated by Hughes. The author uses the acronym SCOTS, after the title of the 1987 volume, to indicate this broad meaning. Used more narrowly, it refers primarily to the program set out by Pinch and Bijker (1984). For this latter program of the social construction of technology, the acronym SCOT was introduced; but this acronym is now increasingly used in the broad sense too, without reference to one specific approach in constructivist studies of technology. The remainder of this article is primarily about SCOT in the narrow sense, although in the last sections, the author broadens the review by returning to SCOTS.

SCOT developed like any normal scientific program: its agenda, central concepts, and even unit of analysis shifted in response to research findings and discussions among contributing scholars. In that sense, one can distinguish early and late (or recent) versions of SCOT. An important, although negative, starting point for SCOT was a criticism of technological determinism. Technological determinism was taken to comprise two basic claims: (1) technology develops autonomously and (2) technology determines societal development to an important degree. These claims were seen as intellectually impoverished and politically debilitating. Proponents of SCOT argued that technological determinism involves a poor research strategy, because it entails a teleological, linear, and one-dimensional view of technological development. In addition, they considered it politically debilitating because technological determinism suggests that the paths taken by technological development are inevitable, thus making social and political interventions a futile endeavor. To bolster this critique, it was necessary to show that the development, stabilization, and even working of technology are socially constructed, with the emphasis on social. Key concepts in this program, as will be discussed in the next section, are 'relevant social group,' 'interpretive flexibility,' 'closure,' and 'stabilization.' The unit of analysis was the single artifact (that is, a tool, a device, or a machine). The choice of the artifact as unit of analysis was a choice of the 'hardest possible case,' similar to the choice of mathematics and natural sciences by the Strong Programme (David Bloor) and the Empirical Programme of Relativism (Harry Collins). To show that even the working of a bicycle or a lamp is socially constructed seemed a harder task, and thus – when successful – more convincing, than to argue that technology at a higher level of complexity and aggregation was socially shaped.

The agenda of demonstrating the social construction of artifacts by detailed analysis at the microlevel resulted in a wealth of case studies. A few years later, the program was broadened in two ways (Bijker and Law, 1992). First, new questions were raised at a meso- and macrolevel of aggregation: for example, about the political construction of radioactive waste, clinical budgeting in the British National Health Service, or technically mediated social order. Second, the agenda was broadened to include, once again, the issue of technology's impact on society, which had been bracketed for the sake of refuting technological determinism. Concepts developed for this agenda were 'technological frame,' and various conceptualizations of the obduracy of technology. The unit of analysis was broadened from the singular technical artifact to the more comprehensive and heterogeneous sociotechnical ensemble. The emphasis now was on 'construction' rather than on 'social' – a similar shift had occurred in the social construction of science, for example, when Latour and Woolgar dropped the word 'social' from the subtitle of their book *Laboratory Life* when it was reprinted in 1986. The aim became to study the mutual shaping, or 'coproduction,' of society and technology, rather than explaining how distinct 'social factors' shaped a separate domain of 'technology' (Oudshoorn and Pinch, 2003). Current research in SCOT continues to broaden the empirical, theoretical, and, indeed, political agendas of SCOT. The author returns to these agendas in the final two sections.

The Social Construction of Technology as a Heuristic for Research

As a heuristic for studying technology in society, research in SCOT can be set out as a progression through three steps (Bijker, 1995).

Key concepts in the first step are 'relevant social group' and 'interpretive flexibility.' An artifact is described through the eyes

of relevant social groups: groups that are relevant because the way they attribute a meaning to the artifact, shapes the artifact's development. Accordingly, relevant social groups can be identified as actors who refer to the artifact in the same basic way. For example, bicycle producers, young athletic ordinary users, women cyclists, and anticyclists each described the high-wheeled ordinary bicycle in the 1870s in a distinctive way. Because the description of an artifact through the eyes of different relevant social groups produces different descriptions – and, in a sense, different artifacts – the researcher is able to demonstrate the 'interpretive flexibility' of the artifact. Although it is common to speak of one artifact, in practice there are many. In the case of the ordinary bicycle, for example, it was an unsafe machine in the eyes of women, but a macho machine in the eyes of the young male ordinary users. For the typical woman of the era, the bicycle was a machine in which the skirt got entangled and from which the woman frequently made a steep fall; for the 'young man of means and nerve' riding the bicycle was a means to impress people (including young ladies).

In the second step, the researcher follows how interpretive flexibility diminishes over time, as some artifacts gain dominance over others and meanings converge, and in the end one artifact results from this process of social construction. Here, key concepts are 'closure' and 'stabilization.' Both concepts are meant to describe the outcome of the process of social construction. 'Stabilization' stresses process: a process of social construction can take several years in which the degree of stabilization slowly increases up to the moment of closure. 'Closure,' a concept stemming from SSK and describing the way scientific controversies end (Collins, 1985), highlights the irreversible end point of a discordant process in which several artifacts (e.g., bicycle designs) coexisted for awhile, before being succeeded by a single model.

In the third step, the processes of stabilization described in the second step are analyzed and explained by interpreting them with a broader theoretical framework. The explanation seeks to answer why a given social construction process follows one course, rather than another. The central concept here is 'technological frame.' A technological frame structures the interactions among the members of a relevant social group, and shapes their thinking and acting. It is similar to Kuhn's concept of 'paradigm' with one important difference: 'technological frame' is a concept that applies to all kinds of relevant social groups, while Kuhn intended 'paradigm' exclusively for scientific communities. (However, these days, the popularized 'paradigm' seems to apply to all sorts of communities and imagined communities.) A technological frame is built up when interaction 'around' an artifact begins. In this way, existing practice guides future practice, although without logical determination. The cyclical movement thus becomes artifact, technological frame, relevant social group, new artifact, new technological frame, new relevant social group, etc.

Typically, a person will be included in more than one social group, and thus also in more than one technological frame. For example, the members of the Women's Advisory Committees on Housing in the Netherlands are included in the technological frame of male builders, architects, and municipality civil servants. This allows them to interact with these men in shaping public housing designs. But at the same time, many of these women are included in the feminist technological frame, which enables them to formulate radical alternatives to the standard Dutch family house that dominates the male builders' technological frame (Bijker and Bijsterveld, 2000).

This three-step research process thus amounts to (1) sociological analysis of an artifact to demonstrate its interpretive flexibility, (2) description of the artifact's social construction, and (3) explanation of this construction process in terms of the technological frames of relevant social groups. The remainder of this section discusses some methodological issues and points of criticism.

It is important to appreciate that SCOT provides a research heuristic for interpretive sociology, not a research machine to churn out case studies as Russell (1986) and Winner (1993) seem to suggest. For example, the adage 'identify all relevant social groups by searching citations of the artifact by a variety of actors' does not diminish the researcher's obligation to decide which groups are important to include in the account, and which groups only obfuscate the picture by adding useless details.

As a research heuristic, SCOT shows its debt to the history of technology by emphasizing case studies and narrative, and to SSK by a preference for qualitative methods such as ethnography. Besides providing a theory of technology and society (as discussed in the next section), SCOT aims at helping researchers to give detailed and insightful accounts of the development of technology in society. Concepts such as 'relevant social group' and 'interpretive flexibility' primarily help to open up a case and to ask original and thought-provoking questions; they are less important for telling the story (see the debate between Angus Buchanan, John Law, and Philip Scranton (1991)). These case studies may be historical or contemporaneous. When doing the latter kind of study, the researcher will often find herself or himself drawn into the case and become one of the actors.

It is also a mistake to think that SCOT research focuses only on the early stages of an artifact's development, the design and prototype phases. If the key idea of SCOT – that all kinds of relevant social groups contribute to the social construction process – is taken seriously, then one should also accept that this construction process takes place in all contexts and all phases. Thus, for example, the automobile in the United States acquired a specific functionality in rural areas where it was used as a stationary power supply to washing machines, butter churns, corn shellers, water pumps, and wood saws, well after it had come off the production line as a transportation vehicle (Kline and Pinch, 1996).

Two concepts, seamless web and symmetry, may be fruitfully used to highlight one specific methodological characteristic of SCOT research. The phrase 'seamless web of technology and society' is often used as a reminder that nontechnical factors are important for understanding the development of technology, and that these factors are closely related to one another as well as the technology. A second, more sophisticated, meaning is that it is never clear a priori and independent of context whether an issue should be treated as technical or social. Was the Challenger accident a technical failure, an organizational mistake, or primarily a lack of adequate funding (Vaughan, 1996)? The activities of engineers and designers are best described as heterogeneous engineering, network building,

or system building, rather than as straightforward technical invention. The recognition that all kinds of social groups are relevant for the construction of technology supports this second way of using the seamless web metaphor. The third interpretation of that metaphor extends the symmetry principle from the strong program in the sociology of knowledge into the principle of general symmetry. Not only should one treat working and nonworking machines symmetrically, but also the very distinction between the social and the technical world can be treated in that way. The 'stuff' of the fluorescent lamp's invention is economics and politics as much as electricity and fluorescence (Bijker, 1995). Let us call this 'stuff' sociotechnology. The relations that play a role in the development of the fluorescent lamp are thus neither purely social nor purely technical; they are sociotechnical. The principle of general symmetry casts this into a methodological principle: human and nonhuman actors should be treated similarly, and the construction of society and the construction of technology should be explained in symmetrical terms (Callon, 1986). In other words, technical reductionism and social reductionism are both out of play as explanatory strategies. Instead, new forms of explanation need to be developed. The next section includes an example.

The Social Construction of Technology as a Theory of Technology in Society

SCOT also provides a theoretical perspective on the relation between technology and society, and on their joint developments or coevolution. As discussed earlier, the development of technology can be explained as a social process in which an open variety of relevant social groups participate. This social process does not stop when an artifact leaves the factory, but continues when users endow the technology with specific uses and meanings. Laws of physics and economics are relevant but insufficient to characterize a technology's development. Artifacts can be shown to have interpretive flexibility, due to the different constructions that various social groups make. When these relevant social groups interact, typically one meaning of the artifact will gain dominance and stabilize. At the end of that social process, closure occurs, and interpretive flexibility vanishes.

Analogously, the development of social institutions in modern society cannot be fully understood without taking into account the role of technology. Such institutions as the church, capitalism, the government, organized labor, communication systems, and education may keep the social fabric of society together; but where would all of these be without technology? Social order in modern society can only be explained by reference to technology (Latour, 1992). This requires a view of technology's impact on society. In SCOT, conceptualizing the hardness or obduracy of technology does this.

An artifact can be hard in two distinctly different ways. The first form, 'closed-in hardness,' occurs when the humans involved have a high inclusion in the associated technological frame. For example, upon meeting problems with printing a letter, a student with a high inclusion in the personal computer (PC) users' technological frame will check cabling, control panels, and printer definitions. Only after some time will it occur to her that she could also write the letter by hand: she was 'closed in' by the PC technology. Another student having the same problem, but with a low inclusion in the PC users' technological frame, has no clue about cables or control panels; he can barely locate the power switch. He experiences the second kind of impact by technology: 'closing-out obduracy.' He sees no alternative but to leave the technology aside and pick up his fountain pen. In both cases, the technology has an impact on these students, but in completely different ways.

These two forms of hardness of technology can also be seen on the societal level. Automobile technology, for example, exerts a 'closed-in hardness' on the inhabitants of Los Angeles, much differentiation within auto culture, but few alternatives outside it. The standardization of main power voltage and wall plugs means a 'closing-out obduracy' to most people; accept it by buying the right plug and apparatus, or refrain from using electricity.

This constructivist analysis provides a theoretical perspective on both SCOT and the technological impact on society. It offers a reconciliation of the previously opposed social constructivist and technological determinist views. This reconciliation needs some theoretical elaboration to extend the conclusion of the previous section: that the distinction between technology and society can be transcended and that the proper subject matter for analysis is sociotechnology. The concept 'technological frame' provides the theoretical linkage between the two views.

A technological frame describes the actions and interactions of actors, explaining how they socially construct a technology. But since a technological frame is built up around an artifact and thus incorporates the characteristics of that technology, it also explains the influence of the technical on the social. Part of the societal impact of the standard Dutch family house was that it dominated architectural thinking in the 1950s through the 1970s, and thus made it very difficult to conceive alternatives; architects and even members of the critical Women's Advisory Committees were 'closed in' by the builders' technological frame, only seeing the two-parent–two-child house.

Since the early 1990s, researchers in social constructivist studies of technology had indeed recognized that explaining the social shaping of technology could only mean half an understanding of the relations between technology and society (Bijker and Law, 1992). Technological determinism, as Wyatt (2008) convincingly argues, needs to be taken seriously. Research into the hardness of technology thus adds to a more balanced understanding of the relationship between the technical and the social; Hommels (2005) explicitly builds on SCOT and actor–network theory in her analysis of the relations between technology and society, while Leonardi (2012) also brings in organizational sociology to understand how the intertwinement (or 'imbrication' as he calls it) of technological and organizational changes results in new forms of sociomaterial stability. Work on the embeddedness of technology in society, partly in the context of policy-oriented European projects, or on the 'domestication' of technology further explores this agenda of an integrated understanding of SCOT and the technical impact on society (Lie and Sorensen, 1996). The earlier work of Don Ihde (1979) and postphenomenology approaches in the philosophy of

technology similarly give a new interpretation of the role of the material artifacts (Verbeek, 2005).

Broadening the Social Construction of Technology

To understand the recent developments in SCOT, it is helpful to widen our scope beyond the initial narrow interpretation of SCOT. In their new preface, Bijker and Pinch review the developments in SCOT at large, or SCOTS, since 1987 (Bijker et al., 2012[1987]).

One development is the investigation of a whole new range of empirical domains, with associated conceptual innovations. Internet and computer studies are the first example (Abbate, 1999; Edwards, 1996). In nanotechnologies, there is a broad recognition among scientists, engineers, and policy makers of the need to look beyond the laboratory (Bijker et al., 2009; Mody, 2011). Scholars such as MacKenzie (2006), Callon (1998), and Pinch and Swedberg (2008) now also extend SCOT, broadly conceived, into the study of economics and financial markets. The newly emerging field of sound studies is heavily indebted to SCOT too (Pinch and Bijsterveld, 2012).

Within organizational sociology, there has been a recent interest in SCOT by scholars such as Garud and Karnoe (2001), and Orlikowski (1992). These studies have a clear interventionist agenda and help to translate SCOT insights to a wider readership of managers and organization specialists.

A Turn to Politics and the Normative

We live in technological cultures; our modern, highly developed societies cannot be fully understood without taking into account the role of science and technology. SCOT and the broader SCOTS program both offer a conceptual framework for politicizing this technological culture. Although such a technological culture already is infused with politics, 'politicizing' here means showing hidden political dimensions, placing issues on the political agenda, and opening issues up for political debate. The SCOT approach not only gives an affirmative answer to Winner's (1980) question 'Do artifacts have politics?' but it also offers a way to analyze such politics. Technology is socially (and politically) constructed; society (including politics) is technically built, and technological culture consists of sociotechnical ensembles.

Studies that develop this agenda, again, typically draw on social constructivist technology studies in the broader sense. One of the most fruitful bodies of work is the analysis of gender and technology (Lerman et al., 2003). Works such as MacKenzie's (1990) history of guided missiles, Hecht's (2009 [1998]) history of nuclear power in France, Gottweis' (1998) study of genetic politics, and Vaughan's (1996) account of the Challenger disaster demonstrate that the new framework can also be productive for the analysis of classical political themes.

The issue of political decision making about technological projects acquires special guise in the light of social constructivism. If it is accepted that a variety of relevant social groups are involved in the social construction of technologies, and that the construction process continues through all phases of an artifact's life cycle, it makes sense to expand the set of groups involved in political deliberation about technological choices. Thus, several countries have experimented with consensus conferences, public debates, and citizens' juries. One of the key issues here is the role of expertise in public debates. The SCOT approach suggests that all relevant social groups have some form of expertise, but that no one group – for example, scientists or engineers – has a special and a priori superiority over the others.

A final development with a strong political and normative edge can be mentioned, which is the turn toward development and globalization. Work in SCOT has increasingly turned to a more global arena, rather than restricting attention to the northern and western hemispheres. This leads, for example, to a fresh approach to long-standing questions in other domains, compared with the typical SCOTS perspective, on the relations between science, technology, and society. One example is cold war history in connection to issues of development and colonization (Hecht, 2011), another is the role of innovation in public health (Engel and Bijker, 2012). In addition to investigating new topics at a global scale, studies are reconceiving traditional technology as a site for sociotechnical innovation (Mamidipudi et al., 2012). Possibly even more exciting than applying the received SCOTS perspectives to new geographies and questions, is the promise of truly new insights by extending the research into these domains, raising new questions, and developing collaborations with scholars and practitioners from Asia, Africa, and Latin America. Such new insights will, the author expects, range from the empirical to the theoretical, and from the epistemological to the political.

See also: Actor-Network Theory; Gender and Technology: From Exclusion to Inclusion?; Laboratory Studies: Historical Perspectives; Risk, Sociology and Politics of; Science and Technology Studies: Experts and Expertise; Scientific Controversies; Scientific Knowledge, Sociology of; Strong Program, Sociology of; Technological Determinism; Technology Assessment.

Bibliography

Abbate, J., 1999. Inventing the Internet. MIT Press, Cambridge, MA.

Berger, P., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday Anchor Books, New York.

Bijker, W.E., 1995. Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change. MIT Press, Cambridge, MA.

Bijker, W.E., Bal, R., Hendriks, R., 2009. The Paradox of Scientific Authority: The Role of Scientific Advice in Democracies. MIT Press, Cambridge, MA.

Bijker, W.E., Bijsterveld, K., 2000. Women walking through plans: technology, democracy and gender identity. Technology and Culture 41, 485–515.

Bijker, W.E., Hughes, T.P., Pinch, T.J. (Eds.), 2012[1987]. The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge, MA.

Bijker, W.E., Law, J. (Eds.), 1992. Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Law, J. (Ed.), Power, Action and Belief: A New Sociology of Knowledge? Routledge and Kegan Paul, London, pp. 196–233.

Callon, M. (Ed.), 1998. The Laws of the Markets. Blackwell, Oxford.

Collins, H.M., 1985. Changing Order: Replication and Induction in Scientific Practice. Sage, London.
Constant, E.W., 1980. The Origins of the Turbojet Revolution. Johns Hopkins University Press, Baltimore.
Cowan, R.S., 1983. More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave. Basic Books, New York.
Edwards, P.N., 1996. The Closed World: Computers and the Politics of Discourse in Cold War America. MIT Press, Cambridge, MA.
Engel, N., Bijker, W.E., 2012. Innovating tuberculosis control in India. Economic & Political Weekly 47, 111–118.
Garud, R., Karnoe, P., 2001. Path Dependence and Creation. Lawrence Erlbaum Associates, Mahwah, NJ.
Gottweis, H., 1998. Governing Molecules: The Discursive Politics of Genetic Engineering in Europe and the United States. MIT Press, Cambridge, MA.
Hecht, G., 2009[1998]. The Radiance of France: Nuclear Power and National Identity after World War II. MIT Press, Cambridge, MA.
Hecht, G., 2011. Entangled Geographies: Empire and Technopolitics in the Global Cold War. MIT Press, Cambridge, MA.
Hommels, A.M., 2005. Unbuilding Cities: Obduracy in Urban Sociotechnical Change. MIT Press, Cambridge, MA.
Hughes, T.P., 1983. Networks of Power: Electrification in Western Society, 1880–1930. The Johns Hopkins University Press, Baltimore, MD.
Ihde, D., 1979. Technics and Praxis. D. Reidel Publishing Company, Dordrecht and Boston.
Kline, R., Pinch, T., 1996. Users as agents of technological change: the social construction of the automobile in the rural United States. Technology and Culture 37, 763–795.
Latour, B., 1992. Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker, W.E., Law, J. (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge, MA, pp. 225–258.
Latour, B., Woolgar, S., 1986[1979]. Laboratory Life: The Social Construction of Scientific Facts. Princeton University Press, Princeton.
Leonardi, P.M., 2012. Car Crashes without Cars: Lessons about Simulation Technology and Organizational Change from Automotive Design. MIT Press, Cambridge, MA.
Lerman, N.E., Oldenziel, R., Mohun, A., 2003. Gender and Technology: A Reader. Johns Hopkins University Press, Baltimore.
Lie, M., Sorensen, K.H. (Eds.), 1996. Making Technology Our Own? Domesticating Technology into Everyday Life. Scandinavian University Press, Oslo.
Mackenzie, D., 1990. Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance. MIT Press, Cambridge, MA.
Mackenzie, D., 2006. An Engine, Not a Camera: How Finance Models Shape Markets. MIT Press, Cambridge, MA.
Mackenzie, D., Wajcman, J. (Eds.), 1999[1985]. The Social Shaping of Technology. Open University Press, Buckingham.
Mamidipudi, A., Syamasundari, B., Bijker, W., 2012. Mobilising discourses: handloom as sustainable socio-technology. Economic & Political Weekly 47, 41–51.
Mody, C.C.M., 2011. Instrumental Community: Probe Microscopy and the Path to Nanotechnology. MIT Press, Cambridge, MA.
Nye, D.E., 1998. Consuming Power: A Social History of American Energies. MIT Press, Cambridge, MA.
Orlikowski, W., 1992. The duality of technology: rethinking the concept of technology in organizations. Organization Science 3, 397–427.
Oudshoorn, N., Pinch, T.J. (Eds.), 2003. How Users Matter: The Co-construction of Users and Technologies. MIT Press, Cambridge, MA.
Pinch, T.J., Bijker, W.E., 1984. The social construction of facts and artifacts: or how the sociology of science and the sociology of technology might benefit each other. Social Studies of Science 14, 399–441.
Pinch, T.J., Bijsterveld, K. (Eds.), 2012. The Oxford Handbook of Sound Studies. Oxford University Press, New York.
Pinch, T.J., Swedberg, R. (Eds.), 2008. Living in a Material World: Economic Sociology Meets Science and Technology Studies. MIT Press, Cambridge, MA.
Russell, S., 1986. The social construction of artefacts: a response to Pinch and Bijker. Social Studies of Science 16, 331–346.
Scranton, P., 1991. Theory and narrative in the history of technology: response. Technology and Culture 32, 385–393.
Sismondo, S., 1993. Some social constructions. Social Studies of Science 23, 515–553.
Vaughan, D., 1996. The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press, Chicago.
Verbeek, P.P., 2005. What Things Do: Philosophical Reflections on Technology, Agency, and Design. Pennsylvania State University Press, University Park, PA.
Winner, L., 1980. Do artifacts have politics? Daedalus 109, 121–136. Reprinted in: Winner, L., 1986. The Whale and the Reactor, University of Chicago Press, Chicago, IL, pp. 19–39.
Winner, L., 1993. Upon opening the black box and finding it empty: social constructivism and the philosophy of technology. Science, Technology, & Human Values 18, 362–378.
Wyatt, S., 2008. Technological determinism is dead; long live technological determinism. In: Hacket, E.J., Amsterdamska, O., Lynch, M., Wajcman, J. (Eds.), The Handbook of Science and Technology Studies, third ed. MIT Press, Cambridge.

Technology Transfer

Shiv Visvanathan, Centre for the Study of Developing Societies, Delhi, India

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15532–15537, © 2001, Elsevier Ltd.

Abstract

Following a discussion of some basic concepts and paradigms in science and technology studies, the article traces how debates on technology transfer have resonated with wider debates on development. Technology transfer has produced a spate of critical discourses centering on intermediate technology, alternative technology, and alternative science. The transfer of military technology is also considered, suggesting proliferation is 'officially impermissible diffusion.' The article then examines the impacts and risks of development, the difficult questions raised by biotechnology, and the 'holy grail' of information technology. The author sees the concept of transfer of technology as one of the great myths of the twentieth century: the belief that the natural diffusion of science and technology can fulfill any democratic dream.

The concept of technology transfer (hereafter, transfer of technology or TOT) has diverse roots: in myth, in the nineteenth century preoccupation with the diffusion of technological artifacts, and in the economics of knowledge. The Promethean act of stealing fire from the gods could be considered a creation myth of technology transfer. Lynn White's classic essay on the origin and diffusion of the stirrup is paradigmatic of the studies on the diffusion of technological artifacts (1962). But TOT as an everyday concept could only come into being with the idea of what Whitehead dubbed the 'invention of invention.' In *Science and the Modern World*, he observed, 'to understand the modern epoch, we can neglect railways, telegraph, radios, spinning machines, and synthetic dyes. We must concentrate on the method itself' (1938: p. 118). Whitehead insisted that 'scientific knowledge is not an abundant storehouse from which one readily helped oneself. An intense period of imaginative design lies between them. One element in the new method is how to go about bridging the gap between a scientific idea and its ultimate product' (1938: p. 18).

The idea of the TOT reflects at once a model of governance, a civics we may say, and a paradigm for the interaction between science and technology (S&T). Also present in this concept is a recognition that science and technology operated as parallel streams throughout most of history. If anything connected them, it was the use of scientific instruments. Till the nineteenth century, science was not a systematic part of the productive process. Economic historians now agree that science played little role in the industrial revolution, which was more the work of craftsman inventors who had little systematic education in science or technology. The concept of technology transfer recognizes the problematic relation between S&T and seeks to increase the reciprocity and frequency of the interaction.

In its most rudimentary form, the TOT concept is founded on the idea of the innovation chain as a sequence of the three stages, namely invention, innovation, and diffusion, which are non-monotonic. Invention represents the conceptualization of a scientific idea conventionally in a laboratory and its concretization in a sample or prototype. Innovation represents the scaling up of the product or process for commercialization, and diffusion is the wider extension of the product in society. Innovation chains represent the lifecycle of a scientific idea within the technological and commercial domain. The nature of the schema disguises certain ideological and cultural biases. First, it assumes that science and technology axiomatically have separate spheres of operation, which the creativity and purity of science requires. Institutionally, science was seen as operative in laboratories and universities; technology in factories and engineering studios. The innovation chain was a linear sequence proceeding from pure science to applied science to development and production. TOT was conceived as following the same linear pathway.

Second, the schema is associated with a set of conceptual separations that tended to reinforce a ritual and institutional separation among the three stages. The relation between pure science, applied science, and development provided the stuff of science policy in the USA for years. If one seminal book embodied this model, it was Vannevar Bush's *Science, The Last Frontier*, which provided the paradigmatic formulation for the science policy of the Cold War years (1945). Bush's work was representative of the time and reflected the era's static understanding of the categories of research. Bush argued that basic research is the engine of technological progress and that it must be separated from applied research because applied research invariably drives out the pure.

Third, present in Bush's writings was not only an understanding of the dynamics of science but of nation building. Bush argued that nations which depend on others for their pure science would lose their competitive edge in world trade. The TOT model thus embodied the ideological assumptions of a particular subculture of science which equated invention with basic science and superiority in basic science with supremacy in world politics (see Stokes, 1997). This subculture was challenged both by the Japanese successes in competing with America in the 1980s and by a newly revisionist history which saw S&T as a more interactive field, in which scientists like Maxwell, Pasteur, and others derived their pure science questions from industrial and agricultural problems. The linearity was also challenged by the development of the new materials sciences which posited and institutionalized a reciprocal relation between science and technology (see Roy, 1970). But there were also ideological reasons for the collapse of the

conventional paradigm. If the quarrel of the categories constituted the stuff of the science policy discourse in the Anglo Saxon postwar world, the challenges to technology transfer models came from a different array of positions in the Third World.

Critical Discourses

The debates on technology transfer resonated with the wider debates on development. As development produced discursive epicycles of equal development, community development, participative development and sustainable development, technology transfer too produced a spate of critical discourses centering on intermediate technology, alternative technology, and alternative science. These responses were not merely driven by academic critiques but evolved as philosophies of grassroot movements.

Present in all these approaches was a geography of power. The innocent horizontality of TOT was now read across the center-periphery axis. The political economy of technology transfer saw the basic choices of S&T as emerging from the metropolis. The periphery was regarded not as a source of independent innovation but only as a recipient of it. The idea of TOT was embedded in the policy of import substitution; industrialization was seen as a passive civics that pulled its subjects into forms of governance over which they did not have active control.

The recognition of intellectual and economic inequality underlying TOT produced spectrum of responses. From Latin America came the work of the center-periphery theorists who argued through LAFTA and UNCTAD for better terms of trade. Here technology was absorbed in the wider discourse of political economy. Inspired by his work in Burma, E.F. Schumacher articulated the idea of intermediate technology (1973). The Vietnam War, the energy crisis and the peace movement produced a search for soft energy paths captured in the idea of appropriate or alternative technology (see Lovins and Price, 1975). All these critiques continued to see science as universal knowledge but felt that technology was local and adaptable. They believed that local knowledges, ethno sciences, could be brought into play to produce a more local and creative technology less discordant with culture. A fourth approach, however, sought to challenge the epistemological hegemony of science, seeking an alternative basis for science through critiques of Judeo-Christian cosmology and the ethnocentrism of modern science. The Indian chemist C.V. Seshadri argued that if the notions of efficiency implicit in the laws of thermodynamics were applied to Third World societies operating at ambient temperatures, such decisions would work against tribals in the Third World, or peasants opting for indigenous or traditional agricultural practices (Seshadri, 1993). The idea of intermediate technology and the emphasis on local knowledges was gradually adopted in World Bank models of technology transfer. The epistemic critique however faced a more ambiguous fate and was subjected to benign neglect.

All the above critiques tried to modify TOT as an antiseptic managerially limited discourse and to relocate it within the wider politics of knowledge and production. They reworked technology within an ethical worldview, arguing that the TOT was also a transfer of values (see George, 1984: pp. 75–76). Ivan Illich's work on education, medicine, and energy was a critique of the basic professionalism implicit in the TOT model (1992). He saw it as a disabling, iatrogenic program, a cultural code for disempowerment.

The effervescence of these critical movements dimmed by the 1990s, though they continue to inspire the struggle against large dams, the critique of forestry, and the logic of urban planning in Third World countries. Intermediate technology in its heyday produced a flowering of locally sensitive technologies such as biogas and modified kitchen stoves, and also revived traditional technologies for housing (see Warren et al., 1995). As dissenting imaginations, they were powerful, but their critical edge was blunted or absorbed by the UN sponsored idea of sustainable development. Sustainable development gave the TOT model a new lease on life by offering the possibility of a balance between ecology and development, renewability and productivity, the long run and the short run.

By the 1970 and 1980s, Science and Technology Studies (STS) had advanced beyond marginality to produce a more variegated set of discourses. The resurgence of the economics of scientific knowledge revitalized studies of the R&D system. Till the 1960s, economists thought that knowledge, like waste, was something external to the production system. Knowledge was not treated as a factor of production like land, labor, or capital. The work of Jewkes et al. (1958), Rosenberg (1976), Freeman (1974), and Machlup (1962) had a far-reaching impact in visualizing technology transfer as a policy discipline. They provided systematic studies exploring the relation between R&D expenditure and rates of productivity. They emphasized the systematic role of patents in evaluating the productivity of such systems and introduced methodologies of cost-benefit analysis to audit TOT processes. In all, they provided a fine-grained set of analytical constructs and empirical studies that helped professionalize science policy as a discipline. They also emphasized that the rationality of the firm was altered by the military-industrial complex.

Transfer of Military Technology

The conceptual world of technology transfer still operated at a split-level with a tendency to dichotomize civilian and military transfers of knowledge. One link between the two was the concept of 'spin-off,' the idea that investments in military research eventually had civilian potential and therefore legitimized the former. The other was the idea of conversion which suggested the transformation of swords into ploughshares, a notion which became popular after the Cold War. One of the finest critiques of military innovation chains was Mary Kaldor's *Baroque Arsenal* (1982). Kaldor argued that while military technology was once innovative, it had become decadent or baroque after World War II. She traced the baroquization of military technology to the interaction between 'the capitalist dynamic of arms manufacturers and the conservatism that characterizes armed forces and defense departments in peace time' (Kaldor, 1982: p. 4). Baroquization is a consequence of the paradoxical dynamic between, on the one hand, armies which cling to designs from the previous war and, on the other, the competition to stay in

business and win contracts that demands incessant technological effort.

The consequence is an involuted technology of perpetual improvements within a conservative framework. What we face is a narrowing of options, greater uniformity, and diminishing returns. More and more money is expended on small improvements ultimately to obtain military hardware which is difficult to maintain. Cost factors increase to such an extent that by 2020, Kaldor suggests, the defense department might be able to afford one plane. Built into such technology as well is planned obsolescence. What was created by Alfred Sloan of General Motors for the car becomes a part of weapons systems. The useful life of a weapon is much shorter than its physical life because of the inexorable power of obsolescence. The emphasis is more on luxuries, accessories, and minor improvements.

Kaldor warned that such military technology when transferred is dangerous both for industrialized and industrializing countries. She argued that the idea of 'Atlanticism' is not just a set of treaties, but a shared technological paradigm posited on a bipolar confrontation between Western and the (then) Eastern totalitarian systems (see Kaldor, 1987: p. 143). She showed how the security concerns of the UK, France, and Germany make them replicate the baroque technologies of the USA, resulting in similar cost overruns and duplication. Worse, alternative forms of defense, including Precision Guided Munitions (PGM) which implied decentralized and dispersed military organizations, could not be developed.

The impact of arms transfers is predictable in developing countries, as illustrated for instance in the regime of the Shah of Iran. The high incidence of arms transfers impairs the capacity to import what is needed for development and industrialization. Once a weapon system has been imported, a chain of supplementary imports is induced. Modern fighter aircraft, tanks, and naval units require an extensive network of support facilities including spare parts and presence of foreign experts. 'The installation of sophisticated military infrastructure is also suboptimal for the development process and often negative for existing traditional infrastructures. Simply put the stability of triage or a road for a tank has to be much higher than for a bullock cart' (Locke and Wulf, 1984: p. 114). The chain of demands appears endless. This pervasive logic of military technology independently suggests that Third World nations can arrive at a self-reliant position in military technology only if they abstain from imitating the military paradigms – both doctrines and weapons systems – developed in the context of the confrontation between industrialized nations. In this context, the Vietnam War was a model confrontation between two modes of technology transfer. Unfortunately the debate did not sustain itself into the new millennium.

Technology transfer as a discourse made incremental responses to new problems. It responded gradually to issues of obsolescence, pollution, proliferation, risk, spin-off on separate grids, while also keeping civilian and military processes on separate registers. There was a need to link concepts such as obsolescence and proliferation. If obsolescence raised problems of value and time in TOT models, proliferation emphasized the question of excess. Military technology represents the classic case, but all too often military technology transfer is restricted to a limited scenario with the legitimate nation-state as actor, without realizing that the smuggler, the mercenary, and eventually even 'rogue states' may be the agents of proliferation or illegitimate transfer.

Proliferation is officially impermissible diffusion. It is a cancerous threat to the ordered chaos of diffusion. As good civics, diffusion reinforces the system of innovation. Nuclear proliferation, however, threatens it, demanding a network of laws, treaties, alliances to restrict its flow. One is simultaneously caught in the debate between atoms for peace and atoms for war, unless one agreed with Robert Jungk (1976) that the distinction was a spurious one, difficult to maintain empirically. The question of proliferation was handled at two levels – technologically as well as legally and politically. Technological measures attempted to control latent proliferation through interventions in the nuclear fuel cycle. Scientists have proposed that proliferation resistance become a design criterion for nuclear energy systems through such techniques as isotopic denaturing. There was, however, a realization that such technical measures needed to be complemented with the political. Within the political domain, proliferation was met by two general strategies. First was the universalist argument in which, through the Nuclear Non-Proliferation Treaty and other legal measures, a full framework of integrated treaties worked for the elimination of nuclear regimes. The other, more restricted effort seeks to argue that a *controlled* spread of nuclear energy alters the behavior of adopting states, adding a dose of restraint and confidence to their behavior.

TOT and the Risks of Development

Throughout the 1960s to the 1980s, the TOT discourse operated between the axes of disarmament and development. It took four decades of development for experts to realize that development, the chief objective of TOT, might produce more refugees than war. India by some counts has over 40 million refugees from development projects; in fact it arguably has more refugees from development than from the wars it has fought after independence. Not unnaturally, some grassroots groups considered TOT as a continuation of war by other means (see for instance, Visvanathan, 1997: p. 41).

The accidents at Windscale, Chernobyl, and Bhopal introduced into the public sphere of TOT discourse the concept of risk and the more comprehensive idea of the risk society (Beck, 1992). The 1984 gas leak at Bhopal, India, the world's biggest industrial disaster, proved that international transfers of technology needed to be located within a wider ethics of responsibility than mere compensation. One needed not only a more systematic right to information and disclosure but also ideas of justice and rehabilitation beyond monetary compensation. The issues of pollution and the proliferation of hazardous waste facilities introduced into the TOT domain new and enlightened legal ideas, including the polluter pays principle. Actors within the legal system realized that as technology moves from certainty to uncertainty, and even to risk, one has to redefine both the rationality of technology and the notion of responsibility. With the rise of environmentalism, risk assessment became a significant element of the TOT discourse. It also

brought to the forefront the latent public aversion to science, as enacted in citizens protests. One had to ask whether citizens' worries about environmental hazards from carcinogenic pesticides, nuclear wastes, loss of global ozone was grounded, and then to develop methodologies for resolving such controversy. The work of Douglas and Wildavsky (1982), Brian Wynne (1987), Kristin Shrader-Frechette (1991) were all attempts in that direction, but with different sensitivities to the requirements of science and the demands of democracy.

The question of risk also prompted inquiry into the rationality behind technology. Ulrich Beck (1992) went so far as to institutionalize the problem. Beck argued that in advanced industrial society the social production of wealth is accompanied by the social production of risks. He claimed one needed a new concept of a risk society with which one not only acknowledged that modernization produces risks, but one also systematically and openly sought to answer how risks could be 'prevented, minimized, dramatized and extended' (Beck, 1992: p. 21). The notion of the risk society led to a range of debates in the social sciences, raising interesting questions about the role of experts in technology transfer, the question of trust in science, and the construction of scientific knowledge itself.

Meanwhile a revisionist history of the Green Revolution was altering the perception of one of the paradigmatic successes of the technology transfer discourse. The Green Revolution, a project which involved the transfer of hybrid, fertilizer-sensitive seeds to India, Pakistan, and other developing countries, was originally seen as a benign example of successful TOT which saved these countries from starvation and made India self-sufficient in food. Norman Borlaug, one of pioneers of the Green Revolution, received the Nobel Prize for peace. Robert Anderson challenged the received wisdom of the Green Revolution by arguing that it involved not merely high yielding packages of seed, pesticides and fertilizer but also incorporated social ideas of changing agricultural society. 'The broad economic links between modernized agricultural production and US economic interests were crucial elements in the strategy from the beginning' (Anderson et al., 1991: p. 361). Francesca Bray, arguing from a different perspective, showed that Green Revolution concentrated on only a few crops, 'wheat, and rice for human consumption, corn and soybean for animal feed and cotton for industry' (Bray, 1994: p. 30). The capital-intensive, highly mechanized model hastened soil erosion and environmental degradation. Bray and later Vandana Shiva (1992) particularly attacked the monocultural mindset behind such TOT, which ignored the productivity and sustainability of polycultural systems, which not only used several varieties of crop but had several uses for various parts of the crop. Such crops were also resistant to pests. It was the impoverishment of gestalts that constitutes the most devastating aspect of the critique of the Green Revolution as TOT.

TOT and Biotechnology

But narratives change and so do battle sites. By 1990, even its pioneers recognized the Green Revolution as a dated narrative. The productivity levels of wheat and rice had leveled off and the scene shifted to biotechnology as the site for the second Green Revolution. While critiques of the Green Revolution came after the event, the movements against biotechnology were spawned almost simultaneously.

The question of the transfer of biotechnology raises fundamental questions for the innovation process in the Third World countries. Pat Mooney (1983, 1998) documents how many crops developed by farmers over centuries became the property of a multinational at the stroke of a pen. Patenting, which was seen as a legal frame for recognizing original innovation, became a technique for appropriating the legacy of innovation in agricultural societies. Important controversies in the 1990s centered around *neem* and *basmati* rice. The growing opposition to chemical products led to a search for natural pesticides and to the *neem*. By 1985, US and Japanese firms had taken out a dozen patents on *neem* pesticides and *neem* toothpaste. The question one has to ask is whether *neem* patents constitute novel forms of knowledge or are such forms of novelty only based on the ignorance of a 2000-year-old tradition of biopesticides in India?

The controversy over *basmati* rice raised similar issues. In 1997, a patent was granted to Ricetech, Inc., a US company, on *basmati* lines and grains. The *basmati* variety was derived from the Indian *basmati* and crossed with semi-dwarf varieties. The question is whether Ricetech *basmati* is the Indian *basmati* and novel at the same time. If *basmati* patents were recognized in India, Indian farmers would ironically be forced to pay royalties to Ricetech.

A patent does not recognize the prior custody of farmers over their seeds. There is no recognition of prior art. Culturally, the patent defines agriculture as an activity performed in a laboratory rather than on a farm. It is a derecognition of an entire intellectual commons. The juridical framework of GATT and TRIPS has possibilities of becoming a new enclosure movement – comparable in spirit to the actions of fifteenth century British landowners – which turns a farmer who saves and replants the seed of patented or protected plant into a violator of the law.

There was a second, equally fundamental problem. Hybridization disables the seed from reproducing itself and forces the farmer to return to the breeder for further stock. Farmers who were custodians of the seed now were systematically deskilled both by technology and law. Finally, there is the ethical question of whether life forms can be patented at all as many of these communities locate the seed within the domain of the commons and the gift.

TOT and Information Technology

If biotechnology appears as a dismal science for the TOT process, information technology, as a narrative, was more celebratory. The rise of technologically innovative groups in Silicon Valley, California, created a new paradigm for industrialism outstripping in its impact Boston's legendary Route 128 (Castells, 1996). Silicon Valley created a false belief that distance and location did not matter in the TOT. Over the last decade, governments in India, China, Malaysia have pursued the holy grail of reinventing Silicon Valleys in their countries. Information technology and the legend of the innovative successes of Bill Gates and Sabir Bhatia, of Oracle, Sun Microsystems, and Intel created a myth that innovation

chains could reproduce themselves anywhere. As Manuel Castells remarked:

> The development of the information technology revolution contributed to the formation of the milieux of innovation where discoveries and applications would interact, and be tested, in a recurrent process of trial and error, of learning by doing; these milieus required (and still do in the 1990s, in spite of on-line networking) spatial concentrations of research centers, higher education institutions, advanced technology companies, a network of ancillary supplies of goods and services, and business networks of venture capital to finance start-ups.
>
> (Castells quoted in McGray, 1999: p. 153)

Information Technology is symptomatic of the cornucopian myths present in TOT. At its most reductive level, technology transfer is a methodology of hope combining the powers of science and technology. It appears as a discourse on management but, as it unravels, its reaches beyond management to the complexities of law, political economy, ethics and anthropology. What begins as a science policy issue transforms itself into a multidisciplinary exercise. Technology transfer is one of the great myths of the twentieth century, a belief that science and technology, if naturally diffused, can fulfill any democratic dream. It is the task of science and technology studies to show the lags, the gaps, the ironies that sustain the myth. What one needs to chronicle is these very lags, gaps, worries and possibilities as the myth of TOT acts itself out in the form of policy.

Bibliography

Anderson, R., Levy, E., Morrison, B.M. (Eds.), 1991. Rice Science and Development Politics. Clarendon Press, Oxford.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage, London.

Beer, J.J., 1959. The Emergence of the German Dye Industry. University of Illinois Press, Urbana, IL.

Bray, F., 1994. Agriculture for developing nations. Scientific American 271, 30–37.

Bush, V., 1945. Science, The Endless Frontier. Government Printing Office, Washington, DC.

Castells, M., 1996. The Rise of the Network Society. Blackwell Publishers, Oxford.

Donahue, D., 1973. Thieves of Fire. Faber and Faber, London.

Douglas, M., Wildavsky, A., 1982. Risk and Culture. University of California, Berkeley, CA.

Freeman, C., 1974. The Economics of Industrial Innovation. Penguin Books, Harmondsworth, UK.

George, S., 1984. Ill Fares the Land. Institute of Policy Studies, Washington, DC.

Illich, I., 1992. In the Mirror of the Past. Lectures and Addresses. Marion Boyars, New York.

Jewkes, J., Sawyers, D., Stillerman, R., 1958. The Sources of Invention. Macmillan, London.

Jungk, R., 1976. The Atom Staat (The New Tyranny). Warner Books, New York.

Kaldor, M., 1982. The Baroque Arsenal. Andre Deutsch, London.

Kaldor, M., 1987. The Atlantic technology culture. In: Kaldor, M., Falk, R. (Eds.), Dealignment. The United Nations University and Basil Blackwell, Oxford, UK, pp. 143–162.

Krohn, W., Layton Jr., E.J., Weingart, P. (Eds.), 1978. The Dynamics of Science and Technology. Yearbook of the Sociology of the Science. D. Reidel Publishing Company, Dordrecht, The Netherlands.

Locke, P., Wulf, H., 1984. Consequences of the transfer of military – oriented technology on the development process. In: Pradip, G. (Ed.), Disarmament and Development. Greenwood Press, Westport, CT, pp. 104–119.

Lovins, A., Price, J.H., 1975. Non-nuclear Futures: The Case for an Ethnical Energy Strategy. Harper and Row, New York.

Machlup, F., 1962. The Production and Distribution of Knowledge in the United States. Princeton University Press, Princeton, NJ.

McGray, D., Spring 1999. The silicon archipelago. Daedalus 128 (2), 147–176.

Mooney, P.R., 1983. The law of the seed: another development and plant genetic resources. Development Dialogue (1–2), 3–172.

Mooney, P.R., 1998. The parts of life: agricultural biodiversity, indigenous knowledge and the role of the third system. Development Dialogue (Special Issue), 6–181.

Rosenberg, N., 1976. Perspectives on Technology. Cambridge University Press, Cambridge, UK.

Roy, R. (Ed.), 1970. Materials Science and Engineering in the United States. Pennsylvania State University Press, University Park, PA.

Schmookler, J., 1966. Invention and Economic Growth. Harvard University Press, Cambridge, MA.

Schumacher, E.F., 1973. Small is Beautiful: A Study of Economics as if People Mattered. Blond and Briggs, London.

Shrader-Frechette, K., 1991. Risk and Rationality. University of California Press, Berkeley, CA.

Seshadri, C.V., 1993. Equity is Good Science. A. M. M. Murugappa Chettiar Research Centre, Madras, India.

Shiva, V., 1992. The Violence of the Green Revolution. The Other India Press, Goa, India.

Stokes, D., 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Institution Press, Washington, DC.

Visvanathan, S., 1985. Organizing for Science: The Making of an Industrial Research Laboratory. Oxford University Press, Delhi, India.

Visvanathan, S., 1997. A Carnival for Science. Oxford University Press, Delhi, India.

Warren, M.D., Slikkerveer, J.L., Brokensha, D. (Eds.), 1995. The Cultural Dimension of Development. Intermediate Technology Publications, London.

White, L., 1962. Medieval Technology and Social Change. Oxford University Press, Oxford, UK.

Whitehead, A.N., 1938. Science and the Modern World. Pelican Books, Harmondsworth, UK.

Wynne, B., 1987. Risk Management and Hazardous Waste. Springer-Verlag, New York.

Teens, Gender, and Self-Presentation in Social Media

Susan C Herring, Indiana University Bloomington, Bloomington, IN, USA
Sanja Kapidzic, Ludwig-Maximilians-Universität, München, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is concerned with how teenage boys and girls present themselves through online social media such as Facebook, Twitter, blogs, and chat forums. Based on research conducted mostly in the United States, it describes and considers the implications of social media use, profile construction, visual and textual self-presentation, profile visibility, truthfulness, and other facets of teens' self-presentation in relation to their gender.

Introduction

Teenagers – young people between the ages of 13 and 19 years – have been identified as the generation with the highest Internet use since the late 1990s (Kraut et al., 1998). As 'digital natives' (Prensky, 2001) who were born and raised in the age of computers and online communication, today's teens share self-created content, post their opinions, and link to other content online more than any other demographic group (Lenhart et al., 2010). The often-heralded democratizing potential of the Internet (e.g., Ess, 1996) has been realized especially strikingly for youth: at no other time in history have young people enjoyed such opportunity to make themselves visible to, and heard by, diverse audiences. At the same time, this exposure entails risks. Moreover, there are gender differences in teens' Internet use, as will be discussed below. Unless otherwise specified, the claims in this article pertain to English-speaking young people in the United States, about whom the bulk of research on online behavior has been conducted.

This article is concerned with how adolescent boys and girls present themselves to others through online social media. Social media are Web-based (and increasingly, mobile) services that allow users to connect and interact with friends, acquaintances, and strangers. Examples include social network sites such as Facebook and Twitter, media-sharing sites such as YouTube and Flickr, blogs, and other Web-based communication forums. Social media "build on the ideological and technological foundations of Web 2.0, and … allow the creation and exchange of User Generated Content" (Kaplan and Haenlein, 2010: p. 61).

Much of that content is photographs, links, and textual information that social media users post to present an online self. Self-presentation is generally considered to be motivated by a desire to make a favorable impression on others, or an impression that corresponds to one's ideals. As such, self-presentation is centrally involved in impression management and the projection of an online identity (cf Schlenker, 1980). Research has shown that teens' online self-presentations differ in various ways, projecting gendered identities.

The term gender is used here to refer to the socially constructed roles that individuals adopt and present to others. Gender normatively maps onto biological sex, and we use the terms *boys* and *girls* to refer to the normative mappings, albeit with the caveats that (1) it is often difficult to determine an Internet user's actual biological sex or offline gender and (2) gender and sex exist along a continuum, and intermediate realizations of both are possible. The following sections describe teenagers' social media use, profile construction, visual and textual self-presentation, profile visibility, truthfulness, and other facets of self-presentation in social media sites in relation to normative gender. The article concludes by discussing the implications of these findings and the future outlook they suggest.

Social Media Use

Teens as a demographic group are avid Internet and social media users in the United States. A recent survey found that almost all U.S. teens (95%) aged 12 through 17 years are online, compared to only 78% of adults. Of these teens, 80% have profiles on social media sites, as compared to only 64% of the online population aged 30 years and older (Lenhart et al., 2011). According to a study conducted by the Kaiser Family Foundation, individuals of 11–18 years of age spend on average over 1.5 h per day using a computer and 27 min per day visiting social network sites, more than one-fourth of their daily computer use (Rideout et al., 2010).

At the same time, because of their inexperience, limited capacity for self-regulation, and susceptibility to peer pressure, teens may not fully understand the possible repercussions of Internet use and are at some risk as they navigate and experiment with social media (O'Keeffe and Clarke-Pearson, 2011). Concerns have been expressed about the amount of time teens spend online (cf Gross, 2004), (lack of) parental control over teenage Internet use (Wang et al., 2005), privacy, risky behaviors such as sexting, cyberbullying, 'Facebook depression,' and exposure to inappropriate content (O'Keeffe and Clarke-Pearson, 2011). Yet other scholars have critiqued the tendency for the mass media and some scholars to fuel 'moral panics' regarding youth online (boyd, 2007; Herring, 2007), pointing out that the incidence of harm to teens resulting from Internet use is actually very low. Moreover, a survey conducted by Wang et al. (2005) suggests that parental awareness of and involvement with their children's Internet use are increasing, compared with earlier decades when Internet-illiterate adults often had little idea what their tech-savvy offspring were doing online. Finally, even given the risks, teens derive many benefits and gratifications from Internet use.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64108-9

Gender differences, as well as some similarities, are apparent in social media site preferences and amount of use. The sites most popular with teenagers and young adults of both genders (as of 2014) are Facebook and Twitter, which constitute social network sites according to the three criteria articulated by boyd and Ellison (2007): they have user profiles, allow for 'friending' (or 'following' on Twitter), and contain social networks that can be navigated to encounter friends of friends. A Pew study conducted in the United States found that 80% of online teens use social network sites, Facebook being the most popular, with 93% of those teens reporting its use (Lenhart et al., 2011). Teens have been leaving Facebook recently, however, and migrating to newer social media platforms such as Instagram, Snapchat, and WhatsApp, which they consider 'cooler' and where they are less likely to encounter their parents (Kiss, 2013). Girls on average spend more time on social network sites and use them more actively than boys do (Brenner, 2012; Rideout et al., 2010). More girls than boys use Facebook and Twitter; female users, including teens, also predominate on the online pinboard Pinterest. Conversely, more males use music-sharing sites such as last.fm, as well as Reddit, a social news Web site known for its sometimes misogynistic content (HuffPost Women, 2012; Williams, 2012). The gender breakdown for Instagram and Snapchat users, the majority of whom are teens, was roughly equal as of October 2013 (Duggan, 2013).

Gender differences are also present in the ways teens use the Internet and social media, although usage patterns have shifted over time. In 1999, teenage boys in the United Kingdom reported using computers more often than girls and feeling more comfortable doing so (Livingstone and Bovill, 1999). By 2004, however, both genders were embracing the Internet as a means of communicating with their friends: Gross (2004) found that the most common activity among American middle and high school students was chatting via instant messaging. In 2007, teenage girls in the United States were more active bloggers than boys – perhaps the first time that females were more active participants than males in a public mode of computer-mediated communication (CMC). Boys, meanwhile, were more likely to upload online videos and use video-sharing applications (Lenhart et al., 2007). In recent years, boys still spend more time using computers, especially playing video games and visiting video Web sites such as YouTube (Rideout et al., 2010). However, girls create and share more video (Lenhart, 2012) and also are more likely to video chat, in keeping with their more active texting and mobile communication behaviors (Lenhart et al., 2010). Regardless of gender, most teens in the United States today spend part of their leisure time online visiting social media sites (Rideout et al., 2010).

The main reason young adults in the United States give for visiting social network sites is to connect and communicate with others and to satisfy their curiosity about their online friends and acquaintances (Urista et al., 2009). Adolescent girls generally use them to communicate with peers and to reinforce preexisting relationships, while boys more often use the platforms to meet new people and make new friends. Boys are also more likely to identify with groups on social network sites that differ from their offline peer circles (Barker, 2009; Lenhart and Madden, 2007a).

A consequence of gender differences in social media use is that girls and boys frequent somewhat different sites and engage in different, albeit overlapping, activities on the sites they visit. However, most social media contexts involve a mix of both genders. This has implications for how teens self-present.

Social Media Profiles

Profile Content

Self-presentation online takes place primarily through social media profiles. Many social media sites allow users to create a profile and visually display connections to their social network (boyd and Ellison, 2007). In addition, many sites allow users to upload and share personal information, pictures, links, music, and other multimedia with their friends' or followers' networks. Profiles first attracted widespread attention on social network sites such as Friendster and MySpace in the middle of the last decade, and they have since been incorporated into many other social media platforms, including chat sites (Kapidzic and Herring, 2011), Wikipedia (Lam et al., 2011), and the mobile photo-sharing application Instagram (Ngak, 2012).

Profiles represent a change in the way Internet users self-present. In earlier textual interactive environments such as chat rooms and MOOs (Multiuser Dimensions, Object Oriented), it was not uncommon for users to invent nicknames and imagined personae (Bechar-Israeli, 1995; Nakamura, 1995). Now, popular platforms such as Facebook explicitly encourage users to provide truthful personal information. Thus teens tend to present their 'real identity' on these sites through their usernames, photographs, and other information that they provide about themselves (Zhao et al., 2008). Although this limits users' ability to experiment with their identity online – for example, it is more difficult to pretend to be a different gender, as was possible in text-based anonymous spaces (Danet, 1998) – young adults can still manipulate their profiles to create impressions that they consider favorable through various means, including number of friends (Ellison et al., 2007), lists of interests (Liu, 2007), and pictures that display particular tastes and preferences (Salimkhan et al., 2010).

Teenage girls and boys differ to some extent in the types of content they post to their profiles. In a study of profiles on several social networking sites, including Facebook, female participants from the United States reported that they post 'cute' pictures, while male participants were more likely to share pictures and comments that they described as self-promoting and that contained sexual content or references to alcohol (Peluchette and Karl, 2008). On a teenage dating site, however, teen girls' self-descriptions contained significantly more references to sex than boys' did (Pujazon-Zazik et al., 2012). Girls in both the United States and Sweden are more likely to display friendship ties on social media, for example, by posting photographs of themselves with their friends (Lenhart and Madden, 2007b; Sveningsson Elm, 2007). Boys, meanwhile, are more likely to orient toward technology, sports, and humor in the information they post to their profile (Sveningsson Elm, 2007) and to share their location and/or phone number (Lenhart and Madden, 2007b; Pujazon-Zazik

et al., 2012). In an earlier analysis of English-language teenage blogs, Huffaker and Calvert (2005) also found that boys provided more information about their location. Moreover, gay boys more often shared their sexual orientation, although there was no difference in how much personal information girls and boys shared in their blogs overall or in their mentions of romantic relationships, both of which were plentiful. Similarly, both boys and girls on a popular Swedish social networking site strongly emphasized their romantic relationships (Sveningsson Elm, 2007).

In a study conducted in the United States of young adults' perceptions of what constitutes appropriate content on social network sites, Peluchette and Karl (2008) found that the research participants expressed little concern about sharing updates and pictures on social network sites such as Facebook. Female participants expressed more concern about future employers seeing some of their pictures and comments, especially those related to alcohol, than males did; however, overall, young adults appear to utilize social media primarily as a way to attract and form relationships with peers, and they are unconcerned with maintaining a professional image (Peluchette and Karl, 2008). This practice can create problems for young adults when they go on the job market. A 2013 survey found that 1 out of 10 young job applicants were rejected because of the content they had posted on social media, including 'provocative or inappropriate photos or posts,' and 'content about drinking or using drugs' (Sherman, 2013).

Visual Self-Presentation

Visual content is a central resource for creating an online impression (Ellison et al., 2006). While in earlier environments such as chat rooms, girls and boys could represent themselves only through the use of textual descriptors (e.g., Nakamura, 1995) or cartoon avatars (e.g., Scheidt, 2004), recent platforms tend to promote the use of photographs in online profiles. Teens and young adults pay special attention to the photographs they select for their profiles (Salimkhan et al., 2010).

Teenagers' main criterion for choosing profile photographs is a belief that they look good in them. According to a study conducted in the United States among young adults, both genders perceive that girls place more emphasis than boys on selecting pictures in which they are attractive (Manago et al., 2008). Manago et al. (2008) interviewed young adults in focus groups about their self-presentation strategies on the social networking site MySpace and found that participants of both genders reported presenting stereotypically gendered images, e.g., of attractive women and strong men. In a survey of Estonian teenage social media users (Siibak, 2009), girls more than boys self-reported selecting pictures in which they thought they looked attractive. Similarly, in a study of teen chat sites, Kapidzic and Herring (2011) found that the vast majority of girls posted profile photographs that presented them in a seductive manner (looking up or sideways at the viewer) and suggestive dress (showing cleavage). Photo choices among male teens were more varied, including dominant, idealized, and affiliative behaviors in addition to seductive behaviors, although there was a tendency for boys to choose photos that showed them at a farther distance from the camera and looking away from the viewer. However, there were also some boys who posed seductively, for example in photos that showed them shirtless. Some young males in Manago et al.'s (2008) study also reported trying to look sexually attractive, a growing trend in social media, especially among Caucasian males in the United States (Kapidzic and Herring, 2014).

On the one hand, these behaviors reproduce cultural gender stereotypes and media portrayals of the sexually available woman and the strong, emotionally distant man (Kapidzic and Herring, 2011; Siibak, 2010). In particular, they reflect the prevalence of pornography dominated by masculine fantasy on the Internet: Studies show that nearly half of American teens surveyed have viewed sexually explicit Web sites, and those who have tend to have less progressive gender role attitudes, as well as more permissive sexual norms (Brown and L'Engle, 2009).

On the other hand, the trend in the early twenty-first century for both boys and girls to self-sexualize in their visual online presentations suggests a more general phenomenon of 'self-commodification' (Siibak, 2010). Most of the time, this is probably unconscious as a result of internalization of media images (Donnelly, 2011). In a study of 288 Facebook users in the United States, Kapidzic (2011) found that young adults who internalized media messages about stereotypical looks were more likely to select profile photos in which they were posing seductively and revealingly dressed. The study participants displayed relatively high levels of media internalization overall, with females scoring significantly higher on the scale. Self-commodification may also be intentional, as in the case of a teenage boy who operated a pornographic Web site featuring images of himself for several years (Leary, 2007). Other cases are less clear: Are 12-year-old-girls who post 'slutty' pictures of themselves on Facebook (Williams, 2012; see also Ringrose, 2011) intending to advertise themselves as sexually available, or are they just imitating the media and their peers? Explicitly pornographic images are prohibited on most social media sites; however, what constitutes 'pornographic' is not always clear.

Profile Visibility and Perceived Audience

Online self-presentation raises privacy issues. Accordingly, most social media sites allow users to adjust settings to control who has access to view their profiles. There is evidence that girls limit the visibility of their profiles more than boys do. Patchin and Hinduja (2010) conducted a content analysis of 2423 teenagers' profiles on a popular social media site to determine to what extent adolescents share information publicly; they found that girls were 1.5 times more likely than boys to restrict access to their profiles. Similar findings were reported by Thelwall (2008a) in an analysis of the profiles of young adults on MySpace, as well as in a recent Pew survey, in which 21% of profile-owning boys in the United States reported leaving their profiles fully visible to the public, as compared to only 12% of girls. Three-fourths of the girls who were interviewed restricted their profile visibility to their friends only, whereas only half of the boys reported doing so (Lenhart et al., 2011). Girls may be especially concerned with assuring the privacy of their online profiles because of the greater tendency for females to be harassed online based on their gender. Results from the Growing

Up With Media Survey, a national survey of more than 1500 American teens aged 10–15 years, indicate that girls are significantly more likely to have experienced sexual solicitation (e.g., requests to talk about sex) and harassment in the form of mean comments on social network sites than boys (Ybarra and Mitchell, 2008). Girls' greater concerns about privacy and identity disclosure on social media sites may predispose them to interact with individuals they already know and trust (Muscanell and Guadagno, 2012).

Privacy settings give one a measure of control over one's audience. However, many social media users do not understand how to adjust the settings, or they ignore them. Before social network sites existed, moreover, privacy settings were uncommon, and access to many social media, such as chat and blogs, was effectively open to the Internet-using public. Despite this reality, teen bloggers, for example, often imagined their audience to be limited: other teens – their friends and romantic interests – but not usually their parents, teachers, or employers (cf Qian and Scott, 2007; Viégas, 2005). This misperception still seems to affect some teen social network users, who reveal compromising information (e.g., about alcohol consumption) and display themselves provocatively attired in their profiles even when the profiles are public. Alternatively, it may be that teen Internet users today are not very concerned with privacy (Kiss, 2013; Peluchette and Karl, 2008).

Another, perhaps related, way in which audience influences self-presentation is the fact that many social media environments popular with young people are heterosexual (by default) marketplaces where flirting and sexual come-ons are common activities (Herring and Zelenkauskaite, 2009; Kapidzic and Herring, 2011). It seems reasonable to suppose that much of teens' online self-presentation in these social media has as a subtext attracting potential partners, be it for sex, dating, online chat, or just to get attention.

Truthfulness of Self-Presentation

For all that it may appear to be self-revealing, the information that girls and boys display about themselves in their profiles is not necessarily accurate. Both genders report experimenting with their online presentation and posting untruthful information to their profiles, such as lying about their age to make themselves older. Results from a Pew survey indicate that 56% of American adolescents with online profiles have posted false information on social media sites. Teenage boys posted fabricated information more often than girls (Lenhart and Madden, 2007b). Moreover, in a survey of more than 300 Dutch adolescents, Valkenburg et al. (2005) found gender differences in the kinds of information male and female teens misrepresented in online interactions. Boys pretended to be more macho, whereas girls pretended they were older and tried to give the impression of being more beautiful. Another study by the Girl Scouts of America found that girls who would describe themselves as 'smart' or 'kind' offline were more likely to post they were 'fun,' 'funny,' or 'social' on social network sites, and girls with low self-esteem were somewhat more likely than girls with high self-esteem to describe themselves as 'sexy' and 'crazy' (Carmon, 2010).

Other research points to a tendency for both adolescent girls' and boys' online self-presentations to mirror their real self in terms of personality traits. Back et al. (2010) asked 236 young adults from Germany and the United States to describe their ideal self and answer a questionnaire to assess personality traits such as openness and extroversion; in addition, research observers rated the participants' profiles. The authors found that the participants' personality scores reflected the observer ratings better than the idealized self-descriptions. Thus while teens may consciously distort the truth to appear more attractive, they have less control over how their personality subconsciously influences their profile descriptions.

Textual Communication and Interaction

Teens also give off textual cues consciously and unconsciously in their online self-presentation (cf Goffman, 1959). This occurs both in descriptions in profiles and, more generally, in textual interactions with other people through chat, instant messaging, discussion forums, blog comments, and the like. Social media sites are increasingly incorporating such CMC features into their platforms.

Numerous studies from English-speaking countries have analyzed textual CMC and have identified gender patterns on the discourse-pragmatic and stylistic levels of language use. Guiller and Durndell (2007), for example, analyzed young adults' language use in computer-mediated discussion groups in Scotland and found that although male and female users did not differ in how they employed linguistic variables such as first-person pronouns, interjections, and imperative verbs, significant gender differences were evident in the use of many stylistic variables: males were more likely to use authoritative language and to respond negatively in interactions, while females were more likely to agree explicitly, support others, and make more personal and emotional contributions. These findings are consistent with observations reported earlier for adult English-language CMC by Herring (1993) and others. Similarly, in an analysis of interactions on five teenage chat sites, most of them based in the United States, Kapidzic and Herring (2011) found significant gender differences. On the pragmatic level, boys used manipulative speech acts more than girls did, inviting and directing others, whereas girls produced more reactive acts. At the stylistic level, boys' messages were more aggressive and flirtatious than girls' messages, which were most often friendly. In an analysis of positive and negative message tone on English-language MySpace profiles, Thelwall et al. (2010) also found that female messages had a positive tone significantly more often than male messages. Kapidzic and Herring (2011) concluded that gendered patterns of communication online have not changed appreciably in the last 20 years.

Findings at the structural level of language are not as distinct. In an analysis of instant messaging communication between young American adults, Baron (2004) found that females employed uncontracted forms (more formal) twice as often as males; females also used emoticons (less formal) more frequently. In Huffaker and Calvert's (2005) analysis of teenagers' blogs, boys used flirty emoticons significantly more often than girls did, although there was no gender difference in amount of emoticon use. Furthermore, Huffaker and Calvert found that while the adolescent boys in their sample used words that were active and resolute more often than the girls

did, there was no difference in the use of cooperative words, which they expected to find more of in girls' blogs. These mixed results seem to support the conclusion that gender differences are most pronounced at broader, stylistic levels of language use.

Sexualized language is prevalent in teenagers' online interactions with their peers. Subrahmanyam et al. (2006) explored the use of adolescents' linguistic choices related to describing sexuality in an English-language teen chat environment. Male teens communicated more explicitly about sex, whereas female teens used more implicitly sexual language, choosing sexualized nicknames and writing about sexual themes. Smahel and Subrahmanyam (2007) studied the process of partner selection in the same chat environment, analyzing the linguistic means by which users invited others to chat. The authors found that in both monitored and unmonitored chat rooms, 16% of all partner requests were invitations of an implicitly or explicitly sexual nature, and these were produced in equal amounts by boys and girls. Girls produced somewhat more partner requests, however. Relatedly, Pujazon-Zazik et al. (2012) coded content on 752 publicly available profiles on an English-language teenage dating site for mentions of risky behavior, such as content related to drugs, violence, alcohol, cigarettes, or sex in the teenagers' self-descriptions. Almost one-third of the descriptions (28%) contained reference to risky behavior, and 16% of those references were related to sex. Girls were more likely to include sexual references in their descriptions, while boys mentioned drugs more often.

Few studies have yet examined gender and language use on social network sites. An exception is Thelwall (2008b), who analyzed gender in relation to the use of swearwords on the profile pages of close to 40 000 MySpace users in the United States and the United Kingdom. Automated language analysis of the pages of users who indicated their age revealed that 16- to 19-year-old boys in the United States used swearwords significantly more than girls of the same age in their self-presentations; in the United Kingdom, however, there were no significant gender differences. Thelwall suggested that the fact that girls in the United Kingdom are incorporating traditionally male language such as strong swearing into their linguistic choices is reflective of deeper changes in gender roles in British society, as evidenced, for example, by the increasing acceptance of binge drinking by U.K. females.

Conclusion

While some similarities exist in social media use by teenage girls and boys, online presentations differ in various respects. The studies described in this article indicate that girls often choose to limit the visibility of their profiles by restricting access by people they are not connected to, whereas boys often allow their profiles to be viewed publicly. Moreover, boys more often post false information on their online profiles. Male and female teens also differ in their textual self-presentation: Boys' linguistic choices reflect assertiveness in both style and tone, while girls seem to aim to please boys and facilitate social interaction. Similarly, in their visual presentations girls most often choose pictures that indicate a desire to appear attractive and sexually appealing, while for boys the patterns are less clear. Both genders' choices of pictures for self-presentation can be seen to reflect sexualized media portrayals.

While such findings might not be surprising for adults, some authors have suggested that teenagers growing up with the Internet and at ease with the many platforms that facilitate interaction might move beyond stereotypically gendered behaviors in online communication (e.g., Huffaker and Calvert, 2005). Nevertheless, much of the recent research conducted on adolescents' online self-presentation points to the persistence of traditional gender stereotypes, both in the use of language and in the selection of images for display. The sexualization of boys diverges from the traditional pattern but is consistent with an overall trend toward what Ringrose (2011) calls the 'pornification' of online self-presentation.

Taken cumulatively, the research described in this article reveals the importance that adolescents place on displaying gendered identities in social media. This finding is not surprising since, in the teenage years, identity and sexuality start to play major roles in the lives of young people as they seek to define and explore who they are (Erikson, 1968). Social media sites and the profiles they contain provide young adults with a reference point in the process of developing their social identity. There they interact with their peers, the most important reference group for teens. boyd (2007) posits that "by looking at others' profiles, teens get a sense of what types of presentations are socially appropriate; others' profiles provide critical cues about what to present on their own profile" (p. 10). This observation helps explain the high degree of normativity within gendered self-presentations in social media.

Minority gender identities – of gay and transgender teens, for example – are also expressed through social media, although, what research exists on the subject suggests that popular sites such as Facebook, MySpace, and Twitter are heteronormative environments (e.g., Carstensen, 2009). Moreover, the one-to-many nature of such sites presents challenges for identity negotiation and information control for queer teens who are not 'out' to all the members of their network, leading some to adopt strategies such as coded reference to their sexuality or maintaining two Facebook accounts, one in which they are closeted and another in which they are out (Cooper and Dzara, 2010; Young, 2012). At the same time, social network sites provide questioning teens an opportunity to explore and reconfigure their gender identities, and in the process, to clarify to themselves who they are (Cooper and Dzara, 2010).

More generally, social media sites provide a space where teens can explore the effects of their self-presented image on others (cf Schlenker, 1980). In many cases, posting pictures and comments will generate positive feedback and could have a positive impact on teenagers' self-esteem. Less-desirable reactions to online self-presentation include negative commenting, cyberbullying, and harassment (Li, 2006), which can have serious psychosocial consequences. Moreover, comparing one's own appearance with that of others in social media can lead to feelings of inadequacy and depression (O'Keeffe and Clarke-Pearson, 2011).

Implications and Future Outlook

The research described in this article has implications for the safety and well-being of teenagers, and thus should be of

interest to parents, educators, and Internet policy makers as well as scholars. It also has implications for the development of sex and gender roles in adolescence and how these are affected by exposure to online content. To the extent that sexualized images, including pornography, affect sexual attitudes and behavior, psychologists as well as health care professionals have a stake in understanding teens' online self-presentations. Finally, the research contributes to a growing body of scholarship on the effects of computer mediation on human social behavior, suggesting, among other things, that profiles constitute invitations to social interaction in much the same way as offline self-presentations do, and that they can be manipulated in equally subtle ways – even, and perhaps especially, in social media environments that encourage truthfulness (Kapidzic and Herring, 2014).

As today's young 'digital natives' grow into tomorrow's adults, we might predict that their practices will increasingly enter the mainstream, from the use of social media to conduct all kinds of business to a lesser concern with privacy to more open attitudes toward sex. Based on reports that young people in the United States today are more tolerant of diversity than their parents' generation (Jones, 2013), we might also predict that more nontraditional gender identities will be openly expressed online. At the same time, it is to be expected that with age, inexperienced youth will understand better the possible undesirable consequences of their online self-portrayals, for example, as regards employment, and make less of their personal information available to mass audiences.

See also: Computer Mediated Communication; Cyberbullying; Disposable Youth in America in the Age of Neoliberalism; Gender Differences in Personality and Social Behavior; Gender and Language: Cultural Concerns; Identity Offline and Online; Internet and Social Media: Anthropological Aspects; Oversharing: The Eclipse of Privacy in the Internet Age; Personality Changes During Adolescence Across Cultures; Self and Emotional Development in Adulthood and Later Life; Sexuality and Space; Social Media; Sociodigital Revolution: Digital Natives vs Digital Immigrants.

Bibliography

Back, M.D., Stopfer, J.M., Vazire, S., Gaddis, S., Schmukle, S.C., Egloff, B., Gosling, S.D., 2010. Facebook profiles reflect actual personality, not self-idealization. Psychological Science 21, 372–374.

Barker, V., 2009. Older adolescents' motivations for social network site use: the influence of gender, group identity, and collective self-esteem. CyberPsychology & Behavior 12 (2), 209–213.

Baron, N.S., 2004. See you online: gender issues in college student use of instant messaging. Journal of Language and Social Psychology 23, 397–423.

Bechar-Israeli, H., 1995. From "Bonehead" to "cLoNehEAd": nicknames, play and identity on Internet relay chat. Journal of Computer-Mediated Communication 1 (2). http://jcmc.indiana.edu/vol1/issue2/bechar.html (accessed 29.01.13.).

Brenner, J., 2012. Social Networking. Pew Internet and American Life Project. http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx (accessed 27.01.13.).

boyd, D., 2007. Why youth heart social network sites: the role of networked publics in teenage social life. In: Buckingham, D. (Ed.), Youth, Identity and Digital Media. MIT Press, Cambridge, MA, pp. 119–142.

boyd, D.M., Ellison, N.B., 2007. Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication 13 (1). http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html (accessed 29.01.13.).

Brown, J.D., L'Engle, K.L., 2009. X-rated. Sexual attitudes and behaviors associated with U.S. early adolescents' exposure to sexually explicit media. Communication Research 36 (1), 129–151.

Carmon, I., November 5, 2010. What Facebook Does to Teenage Girls. Jezebel. http://jezebel.com/5682488/what-does-facebook-do-to-teenage-girls (accessed 11.06.14.).

Carstensen, T., 2009. Gender trouble in Web 2.0: gender perspectives on social network sites, wikis and weblogs. International Journal of Gender, Science and Technology 1 (1), 106–127.

Cooper, M., Dzara, K., 2010. The Facebook revolution: LGBT identity and activism. In: Pullen, C., Cooper, M. (Eds.), LGBT Identity and Online New Media. Routledge, NY, pp. 100–112.

Danet, B., 1998. Text as mask: gender and identity on the Internet. In: Jones, S.G. (Ed.), Cybersociety 2.0. Sage, Thousand Oaks, CA, pp. 129–158.

Donnelly, A.M., 2011. Read my profile: Internet profile culture, young women, and the communication of power. In: Ames, M., Burcon, S.H. (Eds.), Women and Language: Essays on Gendered Communication across Media. McFarland & Co., Jefferson, NC, pp. 167–183.

Duggan, M., 28 October 2013. Photo and Video Sharing Grow Online. Pew Research Center. http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_Photos%20and%20videos%20online_102813.pdf (accessed 19.01.14.).

Ellison, N., Heino, R., Gibbs, J., 2006. Managing impressions online: self-presentation processes in the online dating environment. Journal of Computer-Mediated Communication 11 (2). http://jcmc.indiana.edu/vol11/issue2/ellison.html (accessed 28.01.13.).

Ellison, N.B., Steinfield, G., Lampe, C., 2007. The benefits of Facebook "friends": social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication 12 (4). http://jcmc.indiana.edu/vol12/issue4/ellison.html (accessed 29.01.13.).

Erikson, E.H., 1968. Identity: Youth and Crisis. Norton, Oxford, UK.

Ess, C., 1996. The political computer: democracy, CMC, and Habermas. In: Ess, C. (Ed.), Philosophical Perspectives on Computer-Mediated Communication. State University of New York Press, Albany, NY, pp. 197–230.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Penguin Books, New York.

Gross, E.F., 2004. Adolescent Internet use: what we expect, what teens report. Applied Developmental Psychology 25, 633–649.

Guiller, J., Durndell, A., 2007. Students' linguistic behavior in online discussion groups: does gender matter? Computers in Human Behavior 23 (5), 2240–2255.

Herring, S.C., 1993. Gender and democracy in computer-mediated communication. Electronic Journal of Communication 3 (2). http://ella.slis.indiana.edu/~herring/ejc.doc (accessed 27.01.13.).

Herring, S.C., 2007. Questioning the generational divide: technological exoticism and adult construction of online youth identity. In: Buckingham, D. (Ed.), Youth, Identity and Digital Media. MIT Press, Cambridge, MA, pp. 71–94.

Herring, S.C., Zelenkauskaite, A., 2009. Symbolic capital in a virtual heterosexual market: abbreviation and insertion in Italian iTV SMS. Written Communication 26 (1), 5–31.

Huffaker, D.A., Calvert, S.L., 2005. Gender, identity, and language use in teenage blogs. Journal of Computer-Mediated Communication 10 (2). http://jcmc.indiana.edu/vol10/issue2/huffaker.html (accessed 29.01.13.).

HuffPost Women, 21 June 2012. Social Media by Gender: Women Dominate Pinterest, Twitter, Men Dominate Reddit, YouTube. The Huffington Post. http://www.huffingtonpost.com/2012/06/20/social-media-by-gender-women-pinterest-men-reddit-infographic_n_1613812.html (accessed 27.01.13.).

Jones, J.M., 13 May 2013. Same-Sex Marriage Support Solidifies above 50% in U.S. Gallup Politics. http://www.gallup.com/poll/162398/sex-marriage-support-solidifies-above.aspx (accessed 14.01.2014.).

Kapidzic, S., 2011. The Influence of Personality and Internalization of Media Ideals on Facebook Image Selection. Indiana University Bloomington. Unpublished Master's thesis.

Kapidzic, S., Herring, S.C., 2011. Gender, communication, and self-presentation in teen chatrooms revisited: have patterns changed? Journal of Computer-Mediated Communication 17, 39–59.

Kapidzic, S., Herring, S.C., 2014. Race, gender, and self-presentation in teen profile photographs. New Media and Society. doi:10.1177/1461444813520301. Online first: http://dx.doi.org/10.1177/1461444813520301.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of social media. Business Horizons 53 (1), 61.

Kiss, J., December 27, 2013. Teenagers Migrate from Facebook as Parents Send Them Friend Requests. The Guardian. http://www.theguardian.com/technology/2013/dec/27/facebook-dead-and-buried-to-teens-research-finds (accessed 19.01.14.).

Kraut, R., Lundmark, V., Patterson, M., Kiesler, S., Mukhopadhyay, T., Scherlis, W., 1998. Internet paradox: a social technology that reduces social involvement and psychological well-being? American Psychologist 53 (9), 1017–1031.

Lam, S.K., Uduwage, A., Dong, Z., Sen, S., Musicant, D.R., Terveen, L., Riedl, J., October 2011. WP: clubhouse? An exploration of Wikipedia's gender imbalance. In: Proceedings of WikiSym, pp. 3–5. Mountain View, CA.

Leary, M.G., 2007. Self-produced child pornography: the appropriate societal response to juvenile self-sexual exploitation. Virginia Journal of Social Policy & the Law 15, 1–50.

Lenhart, A., 2012. Teens and Video. Pew Internet and American Life Project. http://pewinternet.org/Reports/2012/Teens-and-online-video/Findings.aspx?view=all (accessed 25.01.13.).

Lenhart, A., Madden, M., 2007a. Social Networking Websites and Teens. Pew Internet and American Life Project. http://www.pewinternet.org/~/media//Files/Reports/2007/PIP_SNS_Data_Memo_Jan_2007.pdf.pdf (accessed 07.11.12.).

Lenhart, A., Madden, M., 2007b. Teens, Privacy, & Online Social Networks. Pew Internet and American Life Project. http://www.pewinternet.org/~/media//Files/Reports/2007/PIP_Teens_Privacy_SNS_Report_Final.pdf.pdf (accessed 07.11.12.).

Lenhart, A., Madden, M., Macgill, A.R., Smith, A., 2007. Teens and Social Media. Pew Internet and American Life Project. http://www.pewinternet.org/~/media/Files/Reports/2007/PIP_Teens_Social_Media_Final.pdf.pdf (accessed 28.01.13.).

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., Rainie, L., 2011. Teens, Kindness and Cruelty on Social Network Sites. Pew Internet and American Life Project. http://pewinternet.org/~/media//Files/Reports/2011/PIP_Teens_Kindness_Cruelty_SNS_Report_Nov_2011_FINAL_110711.pdf (accessed 07.11.12.).

Lenhart, A., Purcell, K., Smith, A., Zickuhr, K., 2010. Social Media and Mobile Internet Use among Teens and Young Adults. Pew Internet and American Life Project. http://pewinternet.org/~/media//Files/Reports/2010/PIP_Social_Media_and_Young_Adults_Report_Final_with_toplines.pdf (accessed 12.11.12.).

Li, Q., 2006. Cyberbullying in schools: a research of gender differences. School Psychology International 27, 157–170.

Liu, H., 2007. Social network profiles as taste performances. Journal of Computer-Mediated Communication 13 (1). Article 13. http://jcmc.indiana.edu/vol13/issue1/liu.html (accessed 07.11.12.).

Livingstone, S., Bovill, M., 1999. Young People, New Media: Report of the Research Project Children Young People and the Changing Media Environment. Department of Media and Communications, London School of Economics and Political Science, London. http://eprints.lse.ac.uk/21177/1/Young_people_new_media_(LSERO).pdf (accessed 07.11.12.).

Manago, A.M., Graham, M.B., Greenfield, P.M., Salimkhan, G., 2008. Self-presentation and gender on MySpace. Journal of Applied Developmental Psychology 29 (6), 446–458.

Muscanell, N.L., Guadagno, R.E., 2012. Make new friends or keep the old: gender and personality in social networking use. Computers in Human Behavior 28, 107–112.

Nakamura, L., 1995. Race in/for cyberspace: identity tourism and racial passing on the Internet. Works and Days 25/26 13 (1&2), 181–193.

Ngak, C., November 5, 2012. Instagram Launches Online Profiles. CBS News. http://www.cbsnews.com/8301-205_162-57545371/instagram-launches-online-profiles/ (accessed 25.01.13.).

O'Keeffe, G.S., Clarke-Pearson, K., 2011. Clinical report – The impact of social media on children, adolescents and families. American Academy of Pediatrics, 800–804.

Patchin, J., Hinduja, S., 2010. Trends in online social networking: adolescent use of MySpace over time. New Media & Society 12 (2), 179–196.

Peluchette, J., Karl, K., 2008. Social networking profiles: an examination of student attitudes regarding use and appropriateness of content. Cyberpsychology & Behavior 11 (1), 95–97.

Prensky, M., 2001. Digital natives, digital immigrants. On the Horizon 9 (5), 1–6.

Pujazon-Zazik, M.A., Manasse, S.M., Orrell-Valente, J.K., 2012. Adolescents' self-presentation on a teen dating web site: a risk-content analysis. Journal of Adolescent Health 50 (5), 517–520.

Qian, H., Scott, C.R., 2007. Anonymity and self-disclosure on weblogs. Journal of Computer-Mediated Communication 12 (4). Article 14. http://jcmc.indiana.edu/vol12/issue4/qian.html (accessed 15.10.12.).

Rideout, V.J., Foehr, U.G., Roberts, D.F., 2010. Generation M: Media in the Lives of 8-to 18-year-olds. Henry J. Kaiser Family Foundation. http://www.eric.ed.gov/PDFS/ED527859.pdf (accessed 07.11.12.).

Ringrose, J., 2011. Are you sexy, flirty or a slut? Exploring "sexualisation" and how teen girls perform/negotiate digital sexual identity on social networking sites. In: Gill, R., Scharff, C. (Eds.), New Femininities: Postfeminism, Neoliberalism and Identity. Palgrave, London, pp. 99–116.

Salimkhan, G., Manago, A., Greenfield, P., 2010. The construction of the virtual self on MySpace. Cyberpsychology: Journal of Psychosocial Research on Cyberspace 4 (1). http://cyberpsychology.eu/view.php?cisloclanku=2010050203&article=1 (accessed 15.10.12.).

Scheidt, L.A., 2004. Buxom girls and boys in baseball hats: adolescent avatars in graphical chat spaces. In: Paper presented at the 54th Annual Conference of the International Communication Association. New Orleans, LA. http://www.academia.edu/214175/Buxom_girls_and_boys_in_baseball_hats_Adolescent_avatars_in_graphical_chat_spaces (accessed 29.01.13.).

Schlenker, B.R., 1980. Impression Management: The Self-concept, Social Identity, and Interpersonal Relations. Brooks/Cole, Monterey/California.

Sherman, E., 4 June 2013. 1 in 10 Young Job Hunters Rejected Because of Their Social Media. http://jobs.aol.com/articles/2013/06/04/applicants-rejected-social-media-on-device-research/ (accessed 19.01.14.).

Siibak, A., 2009. Constructing the self through the photo selection - visual impression management on social networking websites. Cyberpsychology: Journal of Psychosocial Research on Cyberspace 3 (1). http://cyberpsychology.eu/view.php?cisloclanku=2009061501&article=1 (accessed 15.10.12.).

Siibak, A., 2010. Constructing masculinity on a social networking website. The case-study of visual self-presentations of young men on the profile images of SNS Rate.ee. Young 18 (4), 403–425.

Smahel, D., Subrahmanyam, K., 2007. "Any girls want to chat press 911": partner selection in monitored and unmonitored teen chat rooms. CyberPsychology & Behavior 10 (3), 346–353.

Subrahmanyam, K., Smahel, D., Greenfield, P.M., 2006. Connecting developmental constructions to the Internet: identity presentation and sexual exploration in online teen chat rooms. Developmental Psychology 42 (3), 395–406.

Sveningsson Elm, M., 2007. Doing and undoing gender in a Swedish Internet community. In: Sveningsson Elm, M., Sundén, J. (Eds.), Cyberfeminism in Northern Lights. Gender and Digital Media in a Nordic Context. Cambridge University Press, Cambridge, UK. http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-18434 (accessed 19.01.14.).

Thelwall, M., 2008a. Social networks, gender, and friending. Journal of the American Society for Information Science and Technology 59 (8), 1321–1330.

Thelwall, M., 2008b. Fk yea I swear: cursing and gender in MySpace. Corpora 3 (1). http://www.scit.wlv.ac.uk/~cm1993/papers/MySpaceSwearing_online.doc (accessed 07.11.12.).

Thelwall, M., Wilkinson, D., Uppal, S., 2010. Data mining emotion in social network communication: gender differences in MySpace. Journal of the American Society for Information Science and Technology 61 (1), 190–199.

Urista, M.A., Dong, Q., Day, K.D., 2009. Explaining why young adults use MySpace and Facebook through uses and gratifications theory. Human Communication 12 (2), 215–229.

Valkenburg, P.M., Schouten, A.P., Peter, J., 2005. Adolescents' identity experiments on the internet. New Media & Society 7 (3), 383–402.

Viégas, F.B., 2005. Bloggers' expectations of privacy and accountability: an initial survey. Journal of Computer-Mediated Communication 10 (3). Article 12. http://jcmc.indiana.edu/vol10/issue3/viegas.html (accessed 07.11.12.).

Wang, R., Bianchi, S., Raley, S., 2005. Teenagers' internet use and family rules: a research note. Journal of Marriage and Family 67, 1249–1258.

Williams, M.A., 19 October 2012. The War on 12-Year-Old Girls. Salon. http://www.salon.com/2012/10/19/the_war_on_12_year_old_girls/ (accessed 28.01.13.).

Ybarra, M.L., Mitchell, K.H., 2008. How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. Pediatrics 121 (2), e350–e357.

Young, E.J., 2012. Alt-Shift: Queer Online Discourses on Co'ming Out in Serbia. Indiana University Bloomington. Unpublished Master's thesis.

Zhao, S., Grasmuck, S., Martin, J., 2008. Identity construction on Facebook: digital empowerment in anchored relationships. Computers in Human Behavior 24, 1816–1836.

Relevant Websites

http://www.commonsensemedia.org/research – Common Sense Media. Program for the Study of Children and Media.

http://growingupwithmedia.com/ – Growing Up with Media.

http://www.justleftthecloset.com/ – Just Left the Closet [LGBTQ Social Network Site].

http://www.macfound.org/programs/learning/ – Macarthur Foundation Digital Media and Learning.

http://www.pewinternet.org/topics/Teens.aspx?typeFilter=5 – Pew Research Center. Pew Internet and American Life Project.

Telegraph

David Hochfelder, Rutgers University, New Brunswick, NJ, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15555–15558, © 2001, Elsevier Ltd.

Abstract

The telegraph was an application of research in electricity and electromagnetism between 1800 and 1840. The first telegraph lines opened in Britain in 1837 and the USA in 1844. The telegraph remained the mainstay of rapid long-distance communication until practical long-distance telephony after 1900. After 1920, as long-distance telephone rates dropped, telegraphy's share of the long-distance communications market declined. By 1970 the overland telegraph industry in the industrialized world was dying. Submarine telegraphy survived longer; it remained the preferred medium for overseas communication until the installation of undersea telephone cables in the 1950s. After 1980 other technologies, such as facsimile and electronic mail, took over the role of long-distance record communications formerly supplied by the telegraph.

The telegraph was the outgrowth and application of advances in electrical science between about 1800 and 1840, particularly the discovery of galvanic electricity by Luigi Galvani and the invention of the electrical battery by Allesandro Volta around 1800, and research into electromagnetism by Hans Christian Oersted, Andre Marie Ampere, Joseph Henry, and Michael Faraday during the 1820s and 1830s. The first two telegraph lines were opened in Great Britain in 1837 and the USA in 1844. The telegraph was the only form of electrical communication until the invention of the telephone in 1876, and it remained the mainstay of rapid long-distance communication until the development of practical long-distance telephony after 1900. After 1920, as long-distance telephone rates dropped and many countries introduced airmail service, telegraphy's share of the long-distance communications market entered a long decline. By 1970 the overland telegraph industry in the industrialized world was dead or dying. Submarine telegraphy survived longer; it was the only medium for rapid overseas communications until the advent of transatlantic radiotelephony in 1927, and it remained the cheapest and preferred medium until the installation of undersea telephone cables in the 1950s. After 1980 other technologies, such as facsimile and electronic mail, took over the role of long-distance record communications formerly supplied by the telegraph.

In technological and scientific terms, the telegraph was important for three major reasons. First, telegraphy was the first major application of discoveries in electrical science, and it was one of the first technologies with a firm scientific foundation. Second, telegraph electricians during the mid-nineteenth century helped to establish the discipline of electrical engineering. Finally, technical problems, especially in submarine telegraphy, stimulated major scientific advances in physics and oceanography.

Telegraphy was the first technology to sever the connection between communication and transportation. Because of the telegraph's ability to transmit information almost instantly, it affected many aspects of society, culture, politics, international relations, and economics after 1840. It helped to create integrated national and international markets, sped the dissemination of news, provided a model for new literary forms, aided Western imperialism, and spurred national governments to develop telecommunications and technology policies.

The Telegraph in Great Britain

In the mid-1830s William Fothergill Cooke developed a telegraph which indicated letters by using the transmitted electrical current to deflect magnetized needles at the receiver. During the same period, Charles Wheatstone conducted several important experiments on the long-distance transmission of electricity. In February 1837, the two men pooled their efforts and in June they obtained a patent for an electromagnetic telegraph using five wires to deflect five needles in order to indicate letters of the alphabet. They built their first commercial line in 1838 and 1839. In late 1842, Cooke patented an improved telegraph which used two wires and needles, instead of five, to indicate letters; this halved construction and installation costs.

During the early 1840s, railways were the major customers of the telegraph. In 1846 Cooke and a group of investors founded the Electric Telegraph Company to build a commercial telegraph system for the use of the press and businessmen. By 1855 its network encompassed nearly every important town in England and Ireland. For the next quarter century, the Electric company dominated the industry but competed with four other companies for business. During this time the industry shifted from the Cooke and Wheatstone two-needle instruments to various forms of printing or recording telegraphs.

By the mid-1860s, much of the British public, Parliament, and the civil service supported the nationalization of the telegraph industry, claiming that Post Office operation would reduce rates, expand the network, and provide better service. In 1868, Parliament authorized the purchase of the telegraph lines, and the Post Office began operating them in February 1870. Despite initial successes, it was soon apparent that the British postal telegraph faced serious financial obstacles, including an inflated purchase price, unremunerative press and railway rates, commercial rates which barely broke even, and competition from the privately owned telephone industry.

At the turn of the century, telegraph traffic peaked at some 90 million messages. Throughout the following decades, traffic dropped and losses mounted because of competition from the

telephone and the letter post which delivered letters within the British Isles within 24 hours of posting. The telegraph's heyday was over by 1920, and by 1970 the telegraph had ceased to be an important communications medium in Britain.

The Telegraph in the USA

In 1837 and 1838, Samuel F.B. Morse publicly demonstrated his electromagnetic recording telegraph to prominent scientists and government officials. Morse's system at this time comprised an alphabetic code of dots and dashes, a telegraph key to encode messages, and a receiving register which recorded signals on to a moving strip of paper. Morse believed that the federal government should own and operate his telegraph, and hoped to convince the American government to purchase his system. In 1843, Congress gave Morse $30 000 to build an experimental line between Washington and Baltimore. This line began operation in May 1844, but the government refused to fund an extension of the line northward to New York. In 1846, the government turned over the Washington–Baltimore line to private investors.

Over the next 15 years, the telegraph network spread rapidly; by 1850, lines reached every important point east of the Mississippi River, and by 1861, a transcontinental line reached California. The major change to telegraph equipment in this period was the elimination of the recording register; operators shifted to sound reception by deciphering the clicks of the receiving magnet. During the late 1840s and early 1850s, a confusing welter of several dozen companies sprang up to build telegraph lines, many of them poorly organized and short-lived. During the late 1850s, a cartel of six companies coalesced and provided stability to the industry. The end of the Civil War in 1865 sparked a final wave of consolidation, and Western Union emerged in 1866 as the country's telegraph monopoly.

During the remainder of the nineteenth century telegraph technology remained relatively static. The only major advance arose in the 1870s, Thomas Edison's invention of the quadruplex which allowed the simultaneous transmission of four messages on a single wire. Between 1866 and 1910, Western Union faced two major challenges to its dominant position in the American communications market. During this period, a vocal and vigorous movement arose to nationalize the telegraph and place it under postal administration, and Western Union officials spent much time and energy defending their private monopoly to the public and to Congress. The telephone, independently invented in 1876 by Alexander Graham Bell and Elisha Gray, was Western Union's second major challenge. Bell offered to sell Western Union his patent, but Western Union refused and established a competing telephone system in the late 1870s using Gray's and Edison's patents. Because of Bell's superior patent position, Western Union withdrew from the telephone market in 1879 in exchange for Bell's promise not to compete with Western Union's long-distance telegraph business. In the 1870s and 1880s, the telephone could not operate over distances longer than a few dozen miles. However, by about 1890 telephone engineers expanded the range of audible conversations to a few hundred miles, culminating in the establishment of Bell's transcontinental telephone service in 1915. As Bell's long-distance network expanded and rates declined, the telephone steadily eroded the telegraph's share of the long-distance communications market.

Under the leadership of Theodore N. Vail, American Telephone and Telegraph (AT&T), Bell's parent company, was powerful enough to acquire working control of Western Union in 1910. AT&T's control of Western Union demonstrated that the telephone had eclipsed the telegraph as the nation's preferred long-distance communications medium. AT&T voluntarily relinquished its Western Union holdings in 1914 to forestall federal anti-trust proceedings. Although short-lived, Vail's administration modernized Western Union's antiquated accounting and management structure and spurred the company to replace its Morse instruments with automatic telegraph equipment.

Between 1915 and the end of World War II in 1945, the telegraph's share of the long-distance communications market continued to decline because of competition from the telephone, government air mail service, and AT&T's teletype service inaugurated in the 1930s. During the War, the Federal Communications Commission forced Western Union to divest itself of its international cable network and to consolidate with its sole remaining competitor, the bankrupt Postal Telegraph Company, two moves which weakened Western Union's long-term financial position.

Following World War II, Western Union's managers attempted to modernize the company's physical plant and to preserve a market niche for record communications. Between 1945 and 1980, the company undertook three major modernization programs, the development of analog facsimile technology, the construction of a microwave beam network to replace its wires and cables, and the launching of communications satellites. Although these systems were technologically successful, they failed to reverse Western Union's declining market share with respect to the telephone; by 1990, Western Union was defunct except for its money-transfer service.

Telegraphs in Asia and Africa

The first telegraph in India, constructed by two British army officers, opened in 1851. During the 1850s, the British government encouraged the construction of a government-owned telegraph network to link the major trading and political centers, and by 1856 India had some 7000 km of telegraph line and 45 offices.

The French telegraph network in Indo-China was also an instrument of colonial control. France began telegraph construction in 1861 and by the turn of the century the Indo-Chinese network comprised about 12 000 km of land lines and cables connecting over 200 offices. After 1880 the colonial powers occupying Africa built land and cable networks to connect their possessions with the home countries. The Indian, Indo-Chinese, and African telegraph networks were primarily instruments of imperial control and only secondarily commercial systems.

The Chinese government regarded telegraphy suspiciously, as a tool of the Western powers to gain control of their country. During the 1860s, Russian, British, and French entrepreneurs and government officials began pressuring China for telegraph

concessions. Chinese officials resisted until the 1870s, when they granted foreign companies limited franchises to build telegraph lines. In 1881, the government established an Imperial Telegraph Administration to build and to operate a national network, but officials continued to resist the wholesale construction of foreign-owned telegraph lines. However, several European countries used the Boxer Rebellion of 1900 as a pretext to extract political and economic concessions, including the right to build telegraphs, from the weakened Imperial Chinese government.

Whereas Chinese government officials sought to shield the country from foreign influences in the second half of the nineteenth century, in Japan the Meiji Restoration of 1868 brought to power an elite eager to modernize the country along Western European lines. Japan's new rulers embraced the telegraph and hired a British telegraph engineer in 1869 to build a line between Tokyo and Yokohama and gave landing rights to a Danish cable company. By 1872, Japan enjoyed direct telegraphic communication with Europe. During the next two decades, the Japanese telegraph network expanded rapidly. By 1891, Japan had over 400 telegraph offices connected by nearly 12 000 km of wire.

Submarine Telegraphy

The successful development of submarine telegraphy between 1851 and World War I established an international communications network and helped Western nations to pursue their imperialist and commercial ambitions in Africa, Asia, and Latin America. From 1850 to the early 1870s, engineers confronted and solved several problems in submarine telegraphy: insulation, cable-laying ships and equipment, signal distortion and attenuation, and sending and receiving apparatus. British scientists, engineers, and entrepreneurs played leading roles in all four areas.

In 1851, a British firm laid the first successful undersea cable between Dover, England, and Calais, France. British and American entrepreneurs and government officials soon began exploring the feasibility of a transatlantic cable. In 1856, the American Cyrus Field and the British telegraph engineers John Brett and Charles Tilson Bright headed a group of investors who started the Atlantic Telegraph Company. The company received subsidies and naval support from the American and British governments. After two failures in August 1857 and July 1858, the company successfully laid a cable between Ireland and Newfoundland in August 1858. The cable failed, however, after a few weeks of weak and intermittent signals. The 1858 cable failure, the straitened finances of the Atlantic Telegraph Company, and the American Civil War (1861–65) all combined to delay a renewed attempt until 1865. Under the leadership of John Pender, the newly formed Anglo-American Telegraph Company successfully laid two cables in July 1866, marking the start of uninterrupted electrical communication between the two hemispheres. During the last third of the nineteenth century, the Atlantic cables were instrumental in establishing global markets in news, agricultural commodities, and financial securities.

During the late 1850s and 1860s, British telegraph engineers and government officials sought to connect Britain with its possessions in India, southeast Asia, and Australia. In 1870, they successfully laid a cable from England through the Mediterranean, Red, and Arabian Seas to Bombay. In 1872, an extension from India provided direct telegraphic connection to Singapore and Australia. Between 1870 and 1914, several American and British cable companies competed for the telegraph business of the Caribbean and Latin America, a competition which mirrored and fueled the general commercial rivalry between the two countries in the region.

On the eve of World War I, about 500 000 km of undersea cables existed, over half under British control and one-fifth under American control. France, Denmark, Germany, and The Netherlands accounted for the remaining cables. This global submarine cable network helped Western nations to attain economic and political dominance of Asia, Africa, and Latin America. During World War I, British leadership in cable and wireless communications played an important role in the Allied victory over the Central Powers.

During the 1920s, demand for international communications boomed and the submarine telegraph industry reached its peak. Furthermore, AT&T introduced a major technological advance, inductively loaded cables which increased message-handling capacity by a factor of five. Western Union installed the first inductively loaded cable in the Atlantic Ocean in 1924, and the new cables ensured that submarine telegraphy would occupy a central place in global communications over the next several decades. However, economic depression and World War II between 1930 and 1945 reduced demand for international communications generally.

After World War II, two new technologies, telephone cables and satellite communications, converged to make international telephoning easier and cheaper, and hence to render submarine telegraphy obsolete by 1980. AT&T laid the first transatlantic telephone cable in 1956, and by 1983 seven transatlantic cables provided over 11 000 voice circuits. During the 1970s, satellite communications reached maturity, and by the early 1980s satellites provided about 50 000 voice circuits. The installation of digital fiber-optic cables during the 1980s increased the number of international telephone circuits by an order of magnitude and reduced greatly the cost of overseas telephone calls.

See also: Communication and Transportation, History of; Mass Communication: Normative Frameworks; Telephone; Television: General.

Bibliography

Ahvenainen, J., 1981. The Far Eastern Telegraphs: The History of Telegraphic Communications Between the Far East, Europe and America Before the First World War. Suomalainen Tiedeakatemia, Helsinki, Finland.

Blondheim, M., 1994. News Over the Wires: The Telegraph and the Flow of Public Information in America, 1844–1897. Harvard University Press, Cambridge, MA.

Carey, J.W., 1989. Communication as Culture: Essays on Media and Society. Routledge, New York.

Czitrom, D.J., 1982. Media and the American Mind: From Morse to McLuhan. University of North Carolina Press, Chapel Hill, NC.

Gabler, E., 1988. The American Telegrapher: A Social History, 1860–1900. Rutgers University Press, New Brunswick, NJ.
Harlow, A.F., 1936. Old Wires and New Waves: The History of the Telegraph, Telephone, and Wireless. Appleton-Century, New York.
Headrick, D.R., 1991. The Invisible Weapon: Telecommunications and International Politics, 1851–1945. Oxford University Press, New York.
Hugill, P.J., 1999. Global Communications Since 1844: Geopolitics and Technology. Johns Hopkins University Press, Baltimore.
Israel, P., 1992. From Machine Shop to Industrial Laboratory: Telegraphy and the Changing Context of American Invention, 1830–1920. Johns Hopkins University Press, Baltimore.
Kieve, J.L., 1973. The Electric Telegraph: A Social and Economic History. David and Charles, Newton Abbot, UK.
Oslin, G.P., 1992. The Story of Telecommunications. Mercer University Press, Macon, GA.
Perry, C.R., 1992. The Victorian Post Office: The Growth of a Bureaucracy. Royal Historical Society, London.
Reid, J.D., 1879. The Telegraph in America. Derby Brothers, New York.
Schwartzlose, R.A., 1989. The Nation's Newsbrokers. vols 2. Northwestern University Press, Evanston, IL.
Thompson, R.L., 1947. Wiring a Continent: The History of the Telegraph Industry in the United States, 1832–1866. Princeton University Press, Princeton, NJ.

Telephone

James E Katz, Division of Emerging Media Studies, Center for Mobile Communication Studies, Boston University, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the telephone's far-reaching impact on society, democracy, and geography from its inception in 1876 to the present day. The telephone has facilitated the rise of numerous occupations and industries ranging from secretaries and technicians to the distributed sales and telemarketing industries. The telephone has also influenced geographical dispersion, allowing people to physically isolate themselves while always being a phone call away from human contact. Telephones and mobile phones have also revolutionized the way people socialize setting the stage for a potential era of perpetual contact.

Since the invention in 1876, the telephone has become, literally, a fixture of everyday life throughout the industrialized world and much of the developing world as well. It is so commonplace that only its absence becomes noteworthy or even a topic of notoriety. Too easily forgotten by laypeople and scholars alike is the miracle of being able at any time to place a call to any other subscriber worldwide. Compounding the miracle is the fact that so many of these calls are done wirelessly and, increasingly, while multitasking. The enormous value of the telephone can be readily appreciated if one considers the plight of a villager who wants to know if there might be work available in a nearby town, or who needs to summon aid for a sick family member.

Also easily overlooked are the ramifications of this global connectivity. By allowing us to transcend a variety of physical and social barriers, the telephone has led to a complex set of dispersed personal and commercial relationships. By any measure, the telephone has dramatically altered the social landscape. Cars and airplanes were adumbrated respectively by horse-drawn vehicles and birds. Yet, in human imagination the power of real-time interactive oral communication over great distances had been a power reserved exclusively for divine beings. Given this godlike power of the telephone it is little wonder that it has been one of history's most successful inventions. Little wonder, too, that some of the earliest experiments with it were designed to ascertain whether the telephone could be used to communicate with the dead. (Perhaps not entirely regrettably, the answer is that it cannot, though efforts continue to be made to use it so.)

Global and Microcoordination

Today, for about 5 000 000 000 people, the act of making or receiving a telephone call is a normal, often daily, occurrence. (Perhaps another billion, at the other end of the development spectrum, have never used a telephone.) Nearly all settings become appropriate for using mobile telephones, and phone calls and messaging activities have invaded even sacrosanct occasions such as funerals. For rapidly growing numbers of people worldwide, the mobile telephone has become the data port through which text messages, microblogging services such as Twitter, e-mails, and documents can be exchanged. Add-on services such as location-aware monitoring accrete new layers of information on the quotidian physical environment. Given the commonplace quality of the ordinary telephone it remains hard for us to delineate the full spectrum of the telephone's ramifications throughout global societies and the interior world of individual psyches.

Although the telephone has numerous cumulative macrosocial consequences, most of them are predicated on its ability to support social interaction. An essential aspect of the telephone is its ability to allow coordination among geographically dispersed (or even locally concentrated) people. Numerous studies demonstrate that the telephone promotes business, sociality, democracy, and information flow. For those who are physically handicapped or otherwise constrained, it opens vast arrays of contacts. (It is interesting and perhaps surprising to note that the telephone has increased greatly the social contacts and informational resources for the deaf, via TeleTYpewriter (TTY) interfaces. This quality is particularly gratifying given Alexander Graham Bell's original interest in helping the deaf.)

The telephone allows convenient management over distance, and via messages, asynchronously. Upon its basis, extraordinarily large enterprises can be established and effectively controlled. As a corollary, those who would otherwise have greater independence find themselves circumscribed by those (sometimes literally) above them. The telephone has enabled the rapid conduct of large-scale financial transactions and has greatly assisted the formation of capital markets. The mobile quality of the cellular telephone has led to the phenomena of so-called 24 by 7 availability.

A collateral effect is that digital telephonic technology, including voice-mail systems, has been partly responsible for the elimination of huge numbers of middle managers and support staff throughout American industry. Ironically, many of these positions were created in the first place to deal with the communication flood created by the analog telephone. This picture is also influenced by the answering machines, which of course allow the conveyance of telephone messages without human intervention. In former times, self-respecting managers, professors, or other professionals would be reluctant to conduct their business without having a secretary on staff. (Indeed, much effort necessarily went into 'faking' the existence of a secretary to act as an intermediary.) Today, for many professionals, having a secretary would seem to be not only wasteful but also an indication to their clients that they are out of touch with best current practices.

The proliferation of digital messaging technology also affects the tenor of professional events. For instance, it used to

be the case that at a conference or symposium, the break-time would be used for social interaction and enjoying the 'here and now' aspects of the conference venue. Nowadays, however, attendees rush outside to use their mobile phones in an attempt to collect messages and return phone calls. In another context, the mobile phone and pager has led to a new form of social effrontery. While seemingly engaged in conversation, one coolly glances down and scrolls through messages of one's communication device, all the while nodding sagely as if heeding every word uttered by one's interlocutor. These are a few of the many faces of what Kenneth Gergen (2000) has termed 'absent presence.' People are seemingly at an event or socially engaged, but their attention and mental focus is elsewhere.

Telephones, especially mobile phones, increase the pace and efficiency of life. They also allow more flexibility and individual efficiency at business and professional levels and at those of family and personal life. Moreover, people can take action to allay the feeling they are not accomplishing enough. People can harness spare time, or time previously spent in tasks that do not require much attention (for instance, waiting on a shopping queue) to plan and coordinate with others, get information, or even shop remotely by phone while they shop in person. On the other hand, those who treasure respite may find themselves pressured to replace otherwise excusable isolation with productive tasks. Once upon a time, being aboard an airplane excused an executive from having to interact with colleagues. No more, for the satellite or Wi-Fi enabled telephone now follow even at 6 miles high; nor are the seashore and mountaintop beyond their reach. An age of perpetual contact, at least in potential terms, is dawning.

Despite the many advantages to having the telephone available any time and anywhere, there are drawbacks to it as well at both the individual and social levels. The telephone can be a minor irritant or a conduit for crime. In terms of the former, many social critics of the late nineteenth century expressed irritation at the way the telephone seemed to destroy the orderly pace of life, the considerable pleasures of social propriety and as well those of solitude and reflection; the target has remained lucrative, and the frustrations of telephone use and abuse remain a staple of the screen, both large and small. The telemarketing solicitation, of which there is a rising tide, is also an oft-remarked-upon irritant. Telemarketers themselves use market research, so it is no accident that their calls seem to arrive at dinnertime or other exquisitely inconvenient moments. More serious, though, is that there are a large variety of swindles – some of which even use computer-aided dialing schemes – that prey especially on those with the least social resources. Stalking by telephone and obscene phone calls to both men and women are also serious problems. On the other hand, with advances in technology that give call recipients greater power to control access to themselves (such as increasingly popular caller-identification services) people are better able to manage their communication environment to suit their individual tastes and needs.

The increasingly nuanced technology available to callers and callees demands ever greater sophistication to achieve a flawless performance of the choreography of communication. Think of the dilemma of a shy but sedulous teenage boy who wants to ask a popular girl to the school prom. Presumably he would not want to leave a trace on her caller-identification record of his 40 or so attempted calls before finally reaching her. Suffice it to say that, sadly, it appears that there is an iron law of telephone contact: people generally get the most calls from those whom they would least like to hear, and vice versa.

In all, though at times a source of irritation and adroit tool of criminal activity, the telephone has proven to be an important factor in increasing the pace, scope, and intensity of business and management operations while simultaneously lowering the cost of engaging in these activities. It has had a parallel effect on people's levels of social interaction.

Geographical Distribution

The telephone has influenced substantially the distribution of people across the physical landscape. On the one hand, the telephone has allowed the massive vertical downtown urban center to survive then thrive. Without the telephone, the skyscraper would have been most difficult to build and manage. Further, it would be difficult for the structure to maintain its usefulness to its denizens without the telephone.

Some claim that the telephone has enabled one particular form of geodemographic dispersion, the suburb. However, detailed studies of transportation system development indicate that the telephone was not an important factor in this remarkable internal migration. Still, the telephone does allow many in rural locations to participate in business and social relations in a viable way, which without the telephone would not be possible. At the same time, those in extremely low-density areas have persistent difficulty getting even minimal telephone service; the situation is exacerbated by the fact that this populace is generally poor and remote from economic opportunities. This, unfortunately, is a social problem that will endure for decades. The telephone, though, has allowed those who choose isolation – and such people are often drawn from the middle and upper strata of society – to enjoy it without having to forego ready communication. Some seek to 'get away from it all' but still want employment as consultants, or at least to stay in convenient contact with friends, family, and emergency services.

Economic Development

It has long been an article of faith that the telephone promotes economic development and social welfare. This belief has had powerful repercussions on national telecommunications policies worldwide. These have ranged from subsidized rates for the poor, subsidized rates for everyone, and special incentives for telephone companies to modernize their systems or physically extend their services to remote areas. Certainly there has been a strong association between telephone lines and economic growth, particularly, as Hudson (1984) has shown, in developing countries. Researchers have demonstrated a logarithmic relationship between main telephone lines and national income on a per capita basis, although the direction of the causal arrow has not yet been established.

What is not in doubt is that the telephone itself fosters employment growth, not the least to serve itself as an industry.

In the early days of mechanical switches, when telephone systems were growing dramatically, it was speculated that within a generation all young boys in America would have to be employed simply to connect phone calls. Although this scenario never materialized, phone companies have traditionally been enormous consumers of manpower, and are usually one of the largest employers of a region. Traditional telephone companies everywhere, though, are for the most part looking for ways to reduce their traditional workforces, heavily concentrated in jobs of installation or as operators. They are replacing vast numbers of these workers with smaller numbers of software engineers and technicians.

Even as the workers per 1000 telephone lines decreases worldwide, employment in the wireless telephone industry has mushroomed. In many cases, the growth in the mobile phone sector has far surpassed the cuts in the traditional wireline companies (many of which were formerly government ministries of Post, Telegraph, and Telephone services). As the expense of laying copper wire is far grater than erecting a radio tower, many people in developing countries are gravitating to mobile phones simply because they can get them more quickly than wireline ones. It is my sense, based on limited studies, that because the mobile phone is more flexible and can be used in more places, its economic impact will be still greater than that of the growth of its tethered counterparts.

Telephones can have other economic effects: they ease the job prospecting process directly, increasing economic efficiency by making it easier to change jobs. Also, by using the telephone as a sales tool, consumption is stimulated, which leads to job creation and economic growth. Staffing the sales force alone of course means that jobs will be created.

Although much of the prior discussion focused on macroeconomic effects, it is also worth mentioning the twofold microeconomic impact of the telephone. First, it allows individuals to run their own business, often from their home or voice mailbox. Thus, it encourages small enterprises. Yet, it also encourages large-scale enterprises as well since it allows firms to capture the benefits of large size without excessive coordination costs as discussed above.

Policy and Politics

Due to its ability to connect quickly to remote areas, the telephone has helped foster national integrations. At a structural level, it also serves to centralize exercise of political authority while concomitantly circumscribing regional centers of power. Notwithstanding this structural aspect, the telephone has something of a reverse consequence for individual political empowerment. In this regard, Ball (1968) and Aronson (1971) argue that the telephone has been a politically democratizing instrument. There is much to this viewpoint, and studies have shown that, controlling for levels of economic development, in Communist countries of Eastern Europe there were equivalent numbers of TV sets when compared with their Western European counterparts. However, there were drastically fewer telephones. The presumed reason for this maldistribution is that the telephone enables two-way communication, whereas the TV allows controlled information to be sent in a one-way direction from a centralized source to a passive mass audience.

The important role the telephone system could play in controlling the populace has not been lost on governments. In the late nineteenth century, France wanted a centralized, governmentally controlled telephone system because authorities anticipated that rebels might otherwise be able to seize control of it during an uprising. During the Soviet communist regime, the absence of telephone books in Moscow was legendary. Telephone tapping is a long practiced art that has been used to fight crime and enemies of the state.

Revolutionaries as much as police recognize that the telephone is the body politic's nervous system. Hence, it is a universal target for those seeking to subvert or protect the government. Lenin directed that the telephone and telegraph offices be initial targets of his Bolshevik seizure of power. As a contrary example, we point to the August 1991 coup attempt against Mikhail Gorbachev and his government. The plotters failed to seize the telephone system. Consequently, the anti-Communists were able to use it to rally themselves. Boris Yeltsin and his supporters, barricaded in the Parliament building, were able to make and receive calls from around the country and the world; they used this power to marshal and coordinate support. A report on the Russianet even claims the parliamentary defenders were able to phone in their food orders to the Moscow Pizza Hut. The telephone facilitates the management of political campaigns and the operation of political organizations.

For well-organized campaigns, the telephone bank has become an important element in any contest. They can also be used to generate otherwise lethargic public interest in issues. In the United States, a neologism has been created to describe the abuse of this process. The term is 'astroturfing,' in honor of the artificial grass of the same name. Astroturfing is the process of using telephone banks to create the incorrect impression among politicians that there is a 'grass-roots' swell of public concern about an issue when in fact there is little or none. Instead, the outpouring of opinion on the issue had been catalyzed by a small group of telephone operators who 'pipe through' a citizen to her representative's office to recite a prepared message. This example provides further demonstration of the principle that telephone technology has an impact that is governed largely by the motive of the user.

Societal Aspects

Traditional sociological variables of age, social class, socioeconomic status, gender, and race/ethnicity influence people's relationships with the telephone. Young people are generally much more likely to adopt new technology than their older counterparts. Material resources, as measured by income, also have a noteworthy effect. Predictably, those at the very bottom of a society's income distribution are least likely to have telephone service. More surprisingly, as shown by Katz's (1999) research, people in the third income quartile are disproportionately represented among purchasers of enhanced telecom services. This may be because these services are perceived as entertaining or of high prestige. This noteworthy trend occurs in both Europe and the US; whether it obtains in other regions is unknown. Subsequent studies have confirmed this

relationship, and even suggest that it is extensible to developing countries such as China (Katz, 2010).

Since its inception, there have been speculations about the small-group and social interactional consequences of the telephone. Marvin (1988), in her path-breaking study of the social reception of the telephone (among other nineteenth-century electrical technologies), demonstrated that concerns over the impact of the telephone were reproductive of those of the larger society. Among these were the erosion of family authority, race, and social class barriers, and sexual and social propriety. Also of substantial concern was disrespectful and deceitful behavior. Victorian-era Americans believed, justifiably, that the telephone reduced social cues ordinarily conveyed in the richer channels of face-to-face interaction and written correspondence. By masking cues, such as the location and social status of the caller, the telephone denied important cues to assist in making judgments about the interlocutor. Thus, it would be easier to violate social and legal codes. These concerns remain substantial even after a century of telephone use, and are born anew with contemporary communication technology such as the Internet.

The telephone plays a role in the conduct of gender relations and in the pursuit of sexual gratification. In contemporary, often suburbanized society, the telephone has become an important means by which teens socialize themselves in an historical period when they are defined as problematical non-adults (Wynn and Katz, 2000). Commonly, women run the 'social work' of maintaining family links and social schedules via the telephone. (All other things being equal, when both a man and a woman are in a room together, a woman is twice as likely to answer a ringing phone.) Rakow (1992) has documented the important role the telephone plays in women's relationships, and the ways it provides support for their independence despite various constraints and provides them with social resources they need to pursue their relationships and goals.

As in other areas, the telephone can enhance coordination in people's search for partners. The telephone's role in abetting prostitution has long been noted. (Indeed, a term-of-art, 'call-girls,' clearly refers to the role of the telephone.) While it is impossible to capture precisely the temporal, economic, or organizational magnitude of the telephone's role in the commercial sex industry, plainly large sections of the yellow pages and of urban newspapers are devoted to presenting this information in a slightly hidden argot to knowing patrons.

The telephone can also serve as a less direct conduit for sexual expression. Indeed, the conduit of remote interactive sex is an entirely *de novo* mode of expression enabled by the phone. The remote as well as the disembodied aspects of a telephone call can have its sexual appeal, and in this regard one hardly need mention the candlestick design that characterized phones for a half century. By relying on its aural and fantasy capabilities, it can also serve as a sanitary substitute for in-person sex. Recognizing this pent-up demand, an entire industry of 'pink services' has sprung up in any country where it is allowed. Definitive data as to its size are difficult to obtain, but an instructive data set can be derived from the Austrian experience. Austrian law had forbidden phone sex historically. However, in 1991, a revenue-sharing arrangement was established so that offshore companies could benefit from premium international telephone tariffs. The Dutch Antilles became a center of such 'pink services,' and numerous companies began placing ads in Austrian newspapers giving their Antillian numbers. The results were dramatic and instructive. For all of 1990, the total duration of minutes of calls to the Dutch Antilles from Austria was exactly 13 050. For the first 9 months alone of 1991, the number of minutes of calls from Austria to the Dutch Antilles exceeded 1 600 000.

Not all telephone sex is predicated on a financial arrangement. Many millions of obscene calls are reported annually to authorities. Both men and women make obscene phone calls to obtain erotic stimulation from the reaction of their shocked victims. These types of calls seem to occur throughout the world wherever a telephone system is present. As a case in point, within 6 months after the first automatic telephone network was installed in Papua New Guinea, an obscene phone call was reported to authorities. A third class of sexually oriented telephone calls is neither commercial nor inflicted on victims, but rather occurs by mutual consent. For decades, lovers have enjoyed erotic phone conversations, and the topic receives substantial treatment in literary and cinematic formats. Nevertheless, seldom has public attention to this private aspect of human behavior risen to the heights it did in 1998. As part of a larger relentless investigation, an incident of sexual gratification via telephone between US President Bill Clinton and his White House aide, Monica Lewinsky, was given protracted national media attention and related to Congress in an official report. Thus, presidents and perverts alike serve to demonstrate that the telephone, like every other communication technology – from the video camera to the car, and from the printed book to the web-enabled mobile phone – can be used by people as a context for sexual excitement and fulfillment.

Another striking aspect of the telephone is its use as a tool of personal empowerment. With emergency phone boxes set up on campuses, and especially with the mobile phone, people can feel safer in public areas. It has expanded the places and times that people, especially women, can go to a variety of places, thus increasing people's freedom and mobility. It also permits parents to expand the privileges they give their children in terms of when and where they can go. This has the intriguing result of both increasing and loosening parental control.

The telephone also enhances individual autonomy by allowing people to live more successfully by themselves. Yet it preserves instant access between those experiencing a medical or police emergency and a variety of support services. In addition, a substantial variety of community and psychological services are available at any hour and often with complete autonomy.

As with any technology, the use to which it is put is influenced by the motives of the human user. Just as not all people are good, neither are all purposes to which the telephone is put benign. The telephone has offered new venues for monitoring individuals, invading their privacy, and harming them materially or psychologically. In terms of monitoring, the wireless revolution, particularly as it intersects with the telephone, is increasingly affecting all aspects of social interaction, from the conduct of relationships to the receipt of medical care. These technologies will not only be used to deliver services but as well to monitor the individuals. In the past telephonic monitoring has targeted parolees and the elderly. The number of citizens who are regularly and passively tracked will be increasing; this

will be done through the data collected by systems used to regulate mobile phones and intelligent highways. Already mobile phone records are being used in civil (e.g., divorce actions) and criminal cases (e.g., the apprehending of kidnappers).

The telephone also gives individuals tremendous powers to affect others' lives at a distance. Though often used in pranks and practical jokes, the telephone has also been used to inflict protracted harassment and terror on others. Such practices include spreading malicious rumors, making harassing calls late at night, and even concocting identities in order to manipulate or frighten someone. In one case, a young girl pretended to be a nurse, and in this guise reported (falsely) to another girl that her lab test results indicated that she had AIDS and was pregnant. The girl victim attempted suicide and disaster was only narrowly averted. The culprit was caught using caller-ID.

Turning our attention to the underpinnings of social life, Fischer (1992) found that a social psychological effect of the telephone was that it seemed to accelerate the pace of social life. His classic study, *America Calling*, found that the spread of the telephone in the US reduced isolation and increased social contact. A concomitant finding is that a good proportion of telephone users experienced anxiety or 'communication apprehension.' Many citizens also feared excessive social contact, becoming the subject of telephone gossip, and being subjected to unwanted calls. However, he found no support for more extreme claims that it brought people to a constantly heightened sense of alertness and anxiety over the potential summoning outcry and the emergence of 'psychological modernity,' that is, the contemporary sensation of free-floating anxiety and dehumanization. Rather, he concludes that on the whole, people readily and comfortably assimilated the telephone into their lives. They used it as "the conscious product of people employing things, not of things controlling people."

Technological Convergence

It is impossible to speak of today's telephone exclusively in its formerly traditional wireline configuration. Mobile phones, discussed below, are but one aspect of this telephonic reconfiguration; it extends to all aspects of voice communication. One salient aspect is the way people manage their calls. For instance, in the US, the majority of telephone subscribers also have some form of answering machine. This device changes how people use their phones, e.g., to monitor before picking up the receiver, and which in turn affects the behavior of the caller ("I know you're there! Come on, pick up."). Having the telephone answered by a machine was until 1987 seen by the majority of the US population as an act of rudeness to callers. It was emblematic of the fear that 'robots would take over,' and that people were losing control over their daily lives. (This also ties into the above point about the importance of having a secretary to handle one's calls.)

Today opinion surveys show that the majority judges it an important courtesy that should be extended to callers. Some even go so far as to say that it is rude not to have an answering machine or voice mail. Still, the answering machine/voice mail is but one of the ever-multiplying ancillary services that are being added to the formerly black 'one-size-fits-all' telephone. Together they are changing what a telephone is and how it is used, as well as the subterfuges that are used to work around the telephone's expanded capabilities.

Even as the telephone becomes more powerful and capable, it is becoming tetherless. Not only is the home phone increasingly likely to be cordless with reception to a ½ km from base, but people are increasingly subscribing to mobile telephone services that will allow them to range across continents. Moreover, some people are abandoning their wireline phones entirely; in the US, more than 20% of subscribers have done so already and their numbers are increasing. Among other things, this trend means the loss of geographical meaning to area codes, something once deemed an important marker of physical identity, but now increasingly relegated to the category of trivia or arcana.

Computers themselves are increasingly encompassing the features of the telephone, becoming Internet phones and answering machines and voice chat messaging units. Personal digital assistants are beginning to act as both mobile phones and portable computers. For their part, cell phones are becoming more like computers. The age of the 'smart phone' could mean that the computer platform will for most people be subsumed within a portable personal device. Abilities are becoming intermixed, that is, telephone conversations can occur virtually anywhere with anybody via any machine. One consequence of this technological advance is that, without too much hyperbole, it appears that there is arising a new sense of 'perpetual contact.' This new state will allow better coordination, improved business operations, and enhanced kin and social relations. It will also introduce new conflict over the appropriated norms for, and possible constraints on, people's public behavior. To cite an instance, there are now mobile phone–free rail cars and restaurant sections.

Already, however, the proliferation of mobile phones has created a brouhaha over their use in public places, most especially while driving cars. Though some of the controversy revolves around manners and appropriate levels of sensitivity, the larger issue of the use of mobile phones and conduct in public places is only beginning to be addressed. It remains an unsolved puzzle for social scientists and the architectural profession to create public spaces that can amenably coexist with privatizing mobile telephonic behavior. However, the much more powerful issue is the question of safety. Mobile phones mean not only that lives can and will be saved but also that – because there are those who elect to use it while driving – lives will be lost. Sadly, there is a growing toll of people killed by mobile phone using drivers. Those who are affected by such tragedies will continue to produce legislation restricting the use of mobile phones in cars, perhaps in other areas as well.

The Telephone as Synecdoche for the Modern Condition

The telephone has had a substantial impact on the structure and quality of human existence, from business to sex, and on the conduct of war and peace from the international to the intrafamily level. Never has so small a device been used so easily by so many to do so much to so many.

The use and abuse of the telephone encapsulates and expresses the dimension of social scientific perspectives on modern life. What happens in the larger canvas of life is reproduced within the ambit of life on the telephone. The telephone's impact stems from the inherent qualities of a technology that allows easy, inexpensive real-time interaction by voice over both short and great distances. People have investigated aggressively and exploited this technology's potential. Through this unrelentinfg effort, there has been an isomorphism between what the technology can achieve – as its capabilities become more sophisticated – and the needs and desires that exist within mankind's biological and social framework.

The telephone system has been used in ways as broad as the human imagination, many times in ways that were never envisioned (or were even opposed) by telephone company system designers and marketers. Telephone systems have been used to broadcast operas and provide wakeup, dial-a-joke, and dial-a-prayer services. Psychological and emergency services, astrological advice, and interactive matchmaking services are among its staples. Users themselves have created a variety of 'work-arounds' to beat the telephone system (e.g., tone-generating boxes that allowed callers to bypass toll charging mechanisms). None of these uses or innovations was fore-ordained. Rather, as shown by Coopersmith (1991), far from a calculated manipulation of telephone company executives, some innovations as enormously successful as the fax and answering machine were at various points in their development opposed and thwarted by telephone company executives.

No dictator, the telephone has become a servant to those pursuing profit and pleasure. The picture after more than a century and a quarter is quite unlike those sketched originally by hand-wringing dystopians, such as Mark Twain. These critics saw the technology as depleting social capital, killing social relationships, and inducing depression (a view similar to today's critics of the Internet).

On balance, the telephone has clearly done far more to liberate humanity than enslave it. This observation is not meant to downplay the harm that can be and is inflicted telephonically. Like two other communication technologies – the airplane and the car – the telephone can be used for both good and evil. It can be used to wreak suffering and terror; it can be used to carry messages of love, hope, and redemption beyond the vast oceans. In this context, then, we should recall the lessons of the Bible and the Greek myths that even the existence of a powerful god is not without frustration.

See also: Immigration Policy; Telegraph.

Bibliography

Aronson, S.H., 1971. The sociology of the telephone. International Journal of Comparative Sociology 12, 153–167.

Ball, D.W., 1968. Towards a sociology of telephones and telephoners. In: Truzzi, M. (Ed.), Sociology and Everyday Life. Prentice-Hall, Englewood Cliffs, NJ, pp. 59–75.

Coopersmith, J., 1991. The failure of the fax: when a vision is not enough. Business and Economic History 23, 272–282.

Fischer, C., 1992. America Calling: A Social History of the Telephone to 1940. University of California Press, Berkeley, CA.

Gergen, K.J., 2000. A century of absent presence. In: Katz, J.E., Aakhus, M. (Eds.), Perpetual Contact: Mobile Communication, Private Talk, and Public Performance. Cambridge University Press, Cambridge, UK.

Hudson, H., 1984. When Telephones Reach the Village: The Role of Telecommunications in Rural Development. Ablex, Norwood, NJ.

Katz, J.E., 1999. Connections: Social and Cultural Studies of the Telephone in American Life. Transaction Publishers, New Brunswick, NJ.

Katz, J.E. (Ed.), 2010. Handbook of Mobile Communication Studies. MIT Press, Cambridge, MA.

Marvin, C., 1988. When Old Technologies Were New: Thinking about Electric Communication in the Late Nineteenth Century. Oxford University Press, New York.

Pool, I de sola, 1977. The Social Impact of the Telephone. MIT Press, Cambridge, MA.

Rakow, L.F., 1992. Gender on the Line: Women, the Telephone, and Community Life. University of Illinois Press, Champaign, IL.

Russianet. http://www.russianet.ru/~oldrn/history/coup.html.

Wynn, E., Katz, J.E., 2000. Teens on the telephone. Info 2 (4), 401–419.

Television: General

Jostein Gripsrud, University of Bergen, Bergen, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Television has been the dominant mass medium in most Western countries since the 1950s. It has been studied in a great variety of ways, in a number of scholarly disciplines, from psychology and the social sciences to film studies and art history. In the 1970s and 1980s, precisely because the medium invites cross- and interdisciplinary approaches, the study of television was central to the establishment of media studies and cultural studies as new forms of scholarly inquiry and teaching on the borders between the social sciences and the humanities. This article sketches the history of research on television by presenting and discussing main issues, positions, and findings within a selection of central themes. It starts out by looking at relations between the medium as a cultural form and fundamental features of modern society, and then proceeds to map areas of inquiry such as policy/regulation, television production, the structures and contents of programs, the study of audiences and their relation to the medium, and contributions to our understanding of the specificity of the medium. A final section briefly discusses the future of television research.

Television has been the dominant mass medium in most Western countries since the 1950s. It has been studied in a great variety of ways, in a number of scholarly disciplines, from psychology and the social sciences to film studies and art history. In the 1970s and 1980s, precisely because the medium invites cross- and interdisciplinary approaches, the study of television was central to the establishment of new forms and institutions of scholarly inquiry and teaching on the borders between the social sciences and the humanities, media studies, and cultural studies. So far, there are very few if indeed any university departments exclusively devoted to television studies.

The study of television is, thus, still organized in accordance with the general divisions between scholarly disciplines and between 'schools' within (mass) communication and media studies. But especially since the early 1980s, there has also been a trend toward a form of eclecticism where it is recognized that useful work is done from a number of starting points, employing a number of methodologies and a variety of theoretical frameworks. The digitalization of the medium and the rapid growth of various web-based services have further diversified television studies both thematically and methodologically.

The Medium and Modernity

Raymond Williams suggested in *Television: Technology and Cultural Form* (1975) an understanding of television that focused on its social organization as a broadcasting medium, arguing that there is a functional fit between this form of communication and the conditions of modern society. Processes of modernization led on the one hand to centralization of resources and power, on the other to increased social and geographical mobility. Williams termed this latter aspect of modernization 'mobile privatisation,' referring to the relative social isolation of individuals and nuclear families literally and metaphorically 'on the move.' Broadcasting was simply a highly adequate institution in this social situation. It makes immediate distribution of necessary information to all citizens possible, and efficiently contributes to the construction of the 'imagined community' (Anderson, 1983) of nation-states.

Historical research has demonstrated that television technology in itself had other possibilities as well. It was already envisaged in the nineteenth century as a new form of two-way communication. Companies such as AT&T experimented with 'picture-phones' in the 1920s, and amateurs started constructing simple transmitters and receivers. Also, the idea of video surveillance was explored at this time (Allen J., 1983). A more serious alternative to broadcast TV was television as a theatrical medium for collective, public reception. Nazi Germany developed such a system in the 1930s, with hundreds of public television halls throughout the country (Uricchio, 1989). Most people who experienced the first telecasts of the BBC in that decade did so in restaurants, at railways stations, and the like (Corrigan, 1990).

But already in the 1880s, there is evidence that popular imagination played with the idea of broadcast TV, that is, the transmission of sound and images from some central source directly to people's homes, organized in genres such as news, education, art/entertainment, and advertising (Wheen, 1985). In the 1890s, Thomas Alva Edison and others were striving to come up with a successful solution for broadcasting TV (Sklar, 1976: p. 11), and the idea circulated throughout the Western world. All of this lends support to the idea that broadcast television to private homes is related to fundamental social structures and interests in modern capitalist societies, such as the division between the domestic and the public realms on the one hand and the need for socially and culturally integrative efforts on the other. Furthermore, processes of democratization could obviously also be served by broadcasting media since they allow for a more equal access to resources of information, knowledge, and cultural experiences.

The development of multichannel systems of broadcasting, especially via satellites and cable systems since about 1980, may thus be thought of as 'divisive' since they constitute numerous separate viewer-communities within a given nation. So-called postmodernist theorizations of television consequently

saw multichannel television's abundance of programs on offer at any time as a manifestation of a new, presumably 'decentered' social and cultural situation (e.g., Collins, 1989). While the general decline in viewer numbers for the major, regular television channels (e.g., network TV in the United States, public broadcasters in Europe) is well documented, relatively little empirical work has been done on the larger social and cultural consequences of the transition from one or a few to dozens or hundreds of television channels.

There is now a growing number of scholarly publications on the variety of online audiovisual services on offer, from YouTube to Netflix, but few attempts at empirically grounded estimations of how these developments affect structures and functions of the public sphere, social coherence, distribution of knowledge, and so forth. However, most (adult) viewers still spend much or most of their viewing time with a handful of TV channels, free-to-air or subscription channels, and watch discreet programs, possibly prerecorded or streamed over the Internet, as much as they just zap around. Hence the continued existence of 'hit shows' and newscasts competing over ratings. Most 'generalist' channels still stick to well-known genres of programming, have fairly similar newscasts, and address the same or similar issues. It can, thus, be argued that television has retained some of the socially and culturally integrative functions ascribed to it by Williams. Clearly, though, these functions have been modified, first, by multichannel satellite and cable TV, and, second, by digitalization and web-based audiovisual services.

Policies

The history of television policies starts with the attempts to regulate radio broadcasting in the 1920s. The fact that the number of available frequencies was finite led to forms of public regulation unheard of for other mass media. While most European countries decided to organize broadcasting institutions as state-owned monopolies, financed by a license fee, the United States from the beginning regarded broadcasting as another business opportunity, where profits would come from advertising revenues. The United States also had to regulate broadcasting in ways that made it a quite particular area of private enterprise. Frequencies had to be allocated to broadcasting organizations by some public agency, operating in accordance with a set of politically determined principles. The first comprehensive broadcasting legislation in the United States, the 1927 Radio Act, made clear that those who were licensed to use radio waves, were required to serve 'the public interest, convenience, or necessity' (Hoynes, 1994: p. 38). The Federal Radio Commission, which in the 1930s became the Federal Communications Commission, was therefore established. It has ever since played a pivotal role in political struggles over the development of broadcasting in America, with agendas shifting as political administrations have come and gone. Key issues have been the ways in which competition is to be ensured with a necessarily limited number of actors in the market place, and whether or how space and resources are to be provided for noncommercial, public TV. Underlying both issues is the question of what the notion of the 'public interest' which broadcasting is to serve actually means. The scarcity of frequencies lost much of its importance as a basis for regulation with the introduction of satellite and cable television and even more when the, in principle, infinite space of the Internet became a channel for multiple TV-like audiovisual services. Trying to keep up with these changes, the European Union in 1989 adopted its Television without Frontiers directive, which after having been revised in 1997 was finally replaced by the Audiovisual Media Services Directive in 2007. The latter covers both linear (traditional broadcast) and nonlinear (on demand) audiovisual services, but regulations are lighter for on-demand than for linear services.

Regulations and the forms of funding clearly influence program output and thus the medium's social functions. Williams (1975) demonstrated in a simple quantitative comparison the differences both between programming in commercial and noncommercial channels and between the television systems in the United Kingdom and the United States. In the United States, public channels were without broadly popular drama or other entertainment, while network TV was lacking documentaries, one-off dramas, educational programming, and so forth. The result was a polarized system where the majority of viewers would in practice not encounter certain types of programs that might be important to them as citizens. In the British 'duopoly' system, the characteristic differences between commercial and public channels were also observable, but the differences were considerably less pronounced so that the programming menu available to all viewers was more varied.

The BBC generally functioned as a model for European public service television. The definition of public broadcasting's mission formulated by the BBC's first director general, Lord Reith, was thus programmatic for most or all of these institutions – they were to provide information, education, and entertainment to the entire population at a price affordable to all. Some critical scholars, especially in the 1970s, perceived the BBC as an elitist institution basically affirming the existing cultural and social hegemony in the United Kingdom. However, when commercial satellite television arrived in the 1980s such critique partly gave way to a view that emphasized the democratic qualities of the public service television tradition. Since the mid-1980s, the threats to public service television posed by so-called deregulation of broadcasting and the expanding commercial offerings have motivated lots of research into the traditions and conditions of this form of broadcast TV all over Europe, including comparative studies that clarify considerable differences between the various national systems (e.g., Syvertsen, 1992; Raboy, n.d.).

The European Union now decides much of the premises for broadcasting policies in Europe even if this is in principle an area where member states are to retain national sovereignty. On the other hand, the EU has officially decided that public service broadcasting is to be protected. The European Council of Ministers (which includes the EU member states but also all other European countries) has done the same and also defined 'public service television' in a way that very clearly sets it apart from and in opposition to purely commercial services. The distinction is of considerable social importance: Research seems to indicate that the democratic potential of broadcasting is enhanced when it operates as a public service, and diminishes with its degree of commercialization. This

has been argued in a large number of scholarly publications over the last few decades (e.g., Garnham, 1986; Murdock, 2005). But it has more recently also been argued on the basis of hard data from a large-scale, comparative study of media and broadcasting systems in four different countries (Curran et al., 2009). The study shows that (1) the overall levels of public knowledge are higher and (2) the difference in knowledge between people with different levels of education is very significantly smaller in countries where public service broadcasting is quite strictly regulated and not funded by advertising (Finland and Denmark) than in countries with very lax regulation of commercial forces in broadcasting (the United States – the United Kingdom is placed in an intermediary position.)

Production

Research on television production ranges from studies of political regulations, ownership, and other aspects of the economics of the television industry, via studies conducted on production processes in broadcasting institutions and independent production companies to theoretical discussions of the notion of 'authorship' in the context of the necessarily collective production processes in television.

The so-called 'political economy' tradition in mass communication research has always focused on the importance of economic forces and related social structures. Prominent representatives of this tradition are Schiller (1989) in the United States and Garnham (1990) in the United Kingdom, both regarding television as part of the larger media system of capitalist societies. Networks and production companies are parts of huge conglomerates encompassing not just media and leisure industries but also any number of other branches of business. These enormous concentrations of capital exert considerable political and cultural power, both domestically and internationally. British scholars of the 'political economy' tradition have, moreover, been exploring the implications of Jürgen Habermas's theory of the 'public sphere' for our understanding of television and the media in general (*see* Public Sphere and the Media; Mass Media, Political Economy of; Information Society).

In-depth studies of television production practices and personnel are still not very many, but there has been an increasing interest in this area – now due to digitalization often seen as part of a broader field of media production – over the last few years. Cantor's *The Hollywood TV Producer: His Work and His Audience* (1971) is a classic in the field, a piece of organizational sociology charting the space for creativity in television and describing three types of television producers – the 'filmmakers,' the 'old-line producers,' and the 'writer-producers.' The often well-educated writer-producers appeared as the only group that both take the medium seriously and expect the audience to accept more ambitious shows with social 'messages.' Newcomb and Alley (1983) represent a more traditional humanistic view of textual production in their *The Producer's Medium*, based on interviews with individuals they claim have managed to leave their personal imprint on the series they have produced. Gitlin's *Inside Prime Time* (1983) is a study also of top-ranking executives which demonstrates not least the space for personal agendas and preferences when decisions about programming are made. Executives may use results of audience research that support their intuitive preferences while setting aside those that do not. A major, more recent contribution, is John Caldwell's *Production Culture* (2008), a multiple-method study which more than previous work emphasizes the role of 'below the line' laborers in the Los Angeles–based film and video industry and how these as well as 'above the line' personnel make sense of their work. This focus on the industry as a whole has left its mark on later work where the situation of both more and less skilled labor and the industry's general *industrial* nature has been underlined.

Caldwell's book is also among the few studies where an analysis of production is combined with textual analyses of products. An early example of this is a study of the independent company MTM Enterprises, which contributed to a renewal of US television drama with shows such as *The Mary Tyler Moore Show* and *Hill Street Blues* (Feuer et al., 1984). In Britain, Tulloch and Alvarado (1983) did a similar study of the BBC serial *Doctor Who*. Buckingham (1987) and Gripsrud (1995) are examples of studies that also add an analysis of audiences to the study of production and textual analysis, thus covering the entire communicative process.

Programs

Williams (1975) was the first to name and point out the importance of the way in which television's (and radio's) programs are presented as parts of a continuous 'flow.' This means, according to Williams, that the impressions and meanings of individual programs will tend to overlap and merge in the viewer's experience. Flow remains a fundamental feature of broadcast programming. The ordering of programs in certain sequences, so-called scheduling, is of great importance to any television channel. It represents a channel identity and can be used to attract and not least keep viewers for the channel's programs and ads. As Ellis (1982) pointed out, most programs also consist of *segments*, that is, sequences of up to about 5–7 min that are more or less self-contained in terms of meaning. This structure results in part from its practical advantages from a production point of view, but it is also related to the need to make programs accessible to viewers who arrive in the middle of a program. The 'program' unit is thus in a sense a troubled category, threatened by dissolution both by the 'flow' and by internal segmentation.

The limits of 'the program' as an object of textual analysis have consequently been discussed, including the problem of separating the program text from the ads that are inserted into it and the question of the status of a single program or episode in soap operas that run an ongoing story for decades (Feuer, 1989; Allen R., 1985). In practice, viewers have few problems in identifying their favorite show and distinguishing it from the ads, and a soap opera (as well as newscasts and other long-running programs) may be analyzed by way of a more or less representative sample of episodes (Gripsrud, 1995).

It is probably correct to say that most textual analyses of television have been preoccupied with questions of representation and ideology, that is, the representation of gender and

sexuality, of race and ethnicity, class, social conflicts, political issues, and so forth and how they relate to larger ideological patterns and struggles. To these ends, all the theoretical, more or less philosophically grounded, approaches in textual analysis have been applied in a variety of blends – semiotics, psychoanalysis, phenomenology, hermeneutics, and so forth. While fiction and entertainment have provided most of the material for such work, also newscasts, talk shows, and documentaries have been analyzed, often interestingly demonstrating the ways in which such 'reality programming' is actually constructed in accordance with certain textual traditions and principles rather than relating directly the structures of external reality.

The most prominent of social science traditions in the field of program analysis is so-called quantitative content analysis, which defines certain elements of content and then proceeds to count their occurrences, most often across large samples of programs. Typically, the elements counted would be violent acts, sexual innuendo, women in powerful positions, and so forth. Such research may provide overviews of developments over time in certain dimensions of programming or suggest the extent to which certain portrayals of social categories or issues have a dominant position in the total program output or in a certain genre. A prominent example of large-scale quantitative content analysis of television programs is that conducted by Gerbner and coworkers in the late 1960s and early 1970s, the results of which were central to the formulation of the so-called 'cultivation theory,' that is, a theory of television's long-term consequences for people's world view (Gerbner, 1973). Their analyses indicated that (US) television, across genres and individual programs, presented a systematically distorted view of social realities in a number of areas such as family life, work and social roles, aging and dying, crime, and violence. Correlating content analysis with opinion surveys then seemed to suggest that viewers, especially heavy viewers, accepted this distorted view of social conditions as true.

Audiences

In line with the general history of research on modern mass media, worries over how the new medium affected especially children and adolescents were important starting points for the first studies of television in the 1950s. For example, Maccoby (1951) concluded that the medium undermined communication within the family, even if it made family members spend more time together. The same researcher later tried to show that television viewing had a negative 'escapist' function for children who were lacking real-life satisfactions (Maccoby, 1951). Studies such as these were fraught with both conceptual and methodological problems, and marked by a prejudiced, negative attitude to the medium. Any viewing of dramatic fiction would for instance tend to be counted as 'escapist,' reflecting a somewhat shallow understanding of the uses of fantasy and art in general. Around 1960, large empirical projects both in the United Kingdom and in the United States tried to estimate the social and cultural implications, especially for children, of the nationwide introduction of television (Himmelweit et al., 1958; Schramm et al., 1961). The everyday lives of children in communities with and without television were charted in surveys and compared. The findings of these studies were quite complex and did not give a totally negative view of the medium's social and cultural role.

The aforementioned 'cultivation theory' – or, rather, hypothesis – is basically in line with critical theories of the structural, long-term power of the media to define social reality in ways that harm various disadvantaged social groups. Throughout the 1970s such perspectives on television were theorized and significantly renewed in work conducted by researchers at the Centre for Contemporary Cultural Studies at the University of Birmingham, UK. Hall (1973) applied concepts from semiotic theory to the process of communication, speaking of production as a moment of 'encoding' and reception as a moment of 'decoding,' thus theoretically demonstrating the possibility of complications in the 'transfer of messages' that were not simply due to 'noise' or misunderstandings. Hall suggested three main types of 'readings' of television programs – 'dominant,' 'negotiated,' and 'oppositional,' that is, acceptance of the program's 'message,' partial agreement, and critical rejection. An empirical study of groups of viewers conducted by Morley and Brunsdon (Morley, 1980) demonstrated how social class and political convictions might shape audience attitudes to television programs, making some groups more critical than others. In the wake of this study, audience research within or close to the British cultural studies tradition has repeatedly demonstrated how varied interpretations and evaluations of programs might be, both within and between social classes, races, genders, and nations. They have also described how television viewing is inscribed in people's everyday lives and the routines and power struggles that prevail there (e.g., Morley, 1986; Seiter et al., 1989), how this came about historically in a wider social and cultural context (Spigel, 1992) and, somewhat more abstractly, how television viewing has deeper social functions related to its psychological functions for viewers (Silverstone, 1994; Ellis, 1999) (*see* Audiences, Media). More recent research on television audiences, employing a variety of quantitative and qualitative methodologies, has focused, e.g., the cultural diversity of audiences in increasingly multicultural societies and there has been, not least, increased attention to various forms of audience participation in programs and audience influence on program output that have become available particularly through digital technologies and online "fan" communities tied to particular genres and series. An important contribution to this direction of research was Henry Jenkins' *Convergence Culture* (2006).

Medium

Much scholarly writing and various sorts of research may be categorized as contributions to an understanding of television's specificity as a medium in terms of the ways in which it represents reality outside of itself and the ways in which it relates to its audiences. One of the first contributions of lasting significance on the latter theme was the article by social psychologists Horton and Wohl (1956) in which the authors coin the term 'para-social interaction' in a discussion of the 'intimacy at a distance' characteristic of the relations between viewers and television characters such as talk show hosts. The term refers to the way in which viewers tend to

establish pseudo-dialogical relations with television characters, 'talking back' to them, and so forth. These practices set television apart both from film and radio, and reflect the particular type of intimacy associated with this domestic audiovisual medium. They can, besides the domestic location of sets, be tied to the normal size of the screen, which renders close-ups of human faces close in size to those of interlocutors in the same room. Importantly, television's 'talking heads' often address viewers directly ('you'), simulating real conversation also in its preferred styles of speech. These features are shared by radio but have an even stronger effect when combined with moving images.

Furthering this line of thought, Meyrowitz (1985) argued that television has contributed significantly to a series of important changes both in the public sphere and in the relations between people in general. Television has moved politicians and other public figures closer to the rest of the population by presenting them in close-ups and settings that make them appear more 'ordinary.' A new way of appearing in public has thus developed, in which the public or private distinction has become blurred. Meyrowitz terms this 'middle-stage' behavior, referring to Goffman's (1959) distinction between a 'front stage' and a 'back stage' set of conventions for self presentation. This has radically altered the conditions for all sorts of rhetorical practices. Moreover, television programs have provided children with insights into the world of grown-ups and boys and girls now know more about each other's intimate lives. For better or worse, television has thus reduced distances between various social categories and weakened certain hierarchical and authoritarian structures.

The capacity for simultaneity between a real event and its transmission and reception as audiovisual representation, that is, 'liveness,' is central to television's specificity as a medium and its social role. It shares a form of distribution (broadcasting) with radio, it shares subject matter also with newspapers, magazines, and books, and it shares the audiovisual form of representation with film. Early television was all 'live,' but now nearly all television programming is prerecorded. Television producers and broadcasters actively seek the impression of immediacy or 'liveness'; it is simply a key aesthetic value in television. This is related to what Barthes (1981) termed the 'photograph effect': When we see a photograph we (used to) know that what is in the picture has actually once been in front of a camera. Television wants this, and more: not just 'this really happened,' but 'this really happens, right now!' Studio debates, interviews, and variety shows will either be live or try to appear to be. Pretaped shows like sitcoms will often inform us that they are 'taped in front of a live studio audience.' Liveness is particularly important to newscasts, since 'news' as a genre is based on getting as close to immediacy as possible. This is why reporters often talk to the camera – 'live,' 'on the spot' – at a dark place where nothing is happening (anymore). What is at work here is an equation of 'live' with 'real': liveness means reality, or *truth*.

A medium that can give us 'reality in the raw,' unfolding as it happens, cannot lie, it would seem, and this ability tends to lend a particular quality of immediacy, realism, or truth to much more than live programming (cf Heath and Skirrow, 1977). 'Liveness' as an aesthetic value is, therefore, still important even if new technologies and styles of presentation progressively make the medium *less* 'live' (cf Feuer, 1983). Caldwell (1995) has argued that television is more and more becoming a medium not about outside 'events' but about itself and its visual style. But there is still much live programming, and liveness as an effect remains important for the above reasons.

'Liveness' is furthermore a key to television's ability to establish experiential communities. It strengthens the general experience of collectivity in the act of viewing (cf Rath, 1989). Televised national events manifest and support national identity, and so the question is whether or to which extent globally televised events – rock concerts, Olympic games, royal weddings, or funerals – can contribute to the establishment of a new sense of transnational community. Dayan and Katz (1992) argue that television's live broadcasts of major events function as rituals that not only reinforce existing symbolical orders but also create new ones. This is just one of the ways in which television continues to be of great political, social, and cultural importance (*see* Media Events).

Already since the mid-1990s, television as we have known it has by many been said to be about to disappear. Digitalization leads to convergence with any and all other media of communication. The distributive model of broadcasting may appear to be replaced by a structure related to that of the World Wide Web, that is, a system without designated centers, presenting unlimited choice of contents to consumers. These consumers must now be thought of as active individuals and hence the term 'audience' is no longer adequate – they also produce audiovisual content and are also otherwise productive in relation to the medium, hence terms like "prosumers" have appeared. (cf e.g., Jenkins, 2006). A number of publications have dealt with these changes and often tried to prophesize about the medium's future, e.g., Banet-Weiser, Chris and Freitas (eds.): *Cable Visions: Television Beyond Broadcasting* (2007) and Amanda Lotz' *The Television Will be Revolutionized* (2007).

It may, however, also be argued that a total revolution is envisaged mostly on the basis of technological possibilities and that the outcome of an undoubtedly comprehensive process will be modified by social and cultural forces and interests that may sustain the further existence of broadcast television. Total fragmentation and individualization will be restrained by the need for community-forming, shared media experiences, and the need for competent, reliable editorial services or guidance in the confusing mass of contents on offer. Commercial interests will no doubt seek to exploit such needs and strive to establish control of gateways to and routes within the new landscapes of electronic information and entertainment. Governments may seek to find ways of securing the production of identity and distribution of cultural resources represented by the classic broadcasting model, particularly in its public service versions. Historical experience with introductions of new media indicates that already existing media survive alongside the newcomers. Theater was not killed by film; film was not killed by television. Even radio survived television, and books and newspapers are still around.

The dynamics of the ongoing process present television research with a number of exciting, difficult tasks, even if the profession of prophets should not be confused with that of serious scholars. The situation calls for studies of the relations between technological developments, economic forces, sociocultural structures, and, not least, the space for political

interventions. Such studies could well be motivated by an interest in how democracy is served by various media systems and structures. The link between theories of democracy and media is generally in need of sustained, philosophically qualified reflection. The process of convergence and increased interactivity also calls for in-depth analyses of the uses and functions of new and old versions of television, not least the ways in which new and old forms relate to each other. Finally, digitalization opens a whole set of new aesthetic and rhetorical possibilities; new forms of representation, new ways of telling stories, new ways of persuading consumers and citizens. While all of these new areas of inquiry will no doubt require new theoretical and methodological tools, the experiences of some 65 years of television research are well worth keeping in mind. Reinventing the wheel remains a waste of time.

See also: Audiences, Media; British Cultural Studies; Broadcasting: Regulation; Film and Video Industry; Mass Media, Political Economy of; Mass Media, Representations in; Television: History.

Bibliography

Allen, J., 1983. The social matrix of television: invention in the United States. In: Kaplan, E.A. (Ed.), Regarding Television: Critical Approaches—An Anthology. University Publications of America, The American Film Institute, Los Angeles, pp. 109–119.

Allen, R., 1985. Speaking of Soap Operas. University of North Carolina Press, Chapel Hill, NC.

Anderson, B., 1983. Imagined Communities: Reflections on the Origins and Spread of Nationalism. Verso, London.

Banet-Weiser, S., Chris, C., Freitas, A. (Eds.), 2007. Cable Visions: Television beyond Broadcasting. NYU Press, New York; London.

Barthes, R., 1981. Camera Lucida: Reflections on Photography. Hill and Wang, New York.

Buckingham, D., 1987. Public Secrets: 'East Enders' and Its Audience. British Film Institute, London.

Caldwell, J.T., 1995. Televisuality: Style, Crisis and Authority in American Television. Rutgers University Press, New Brunswick, NJ.

Caldwell, J.T., 2008. Production Culture: Industrial Reflexivity and Critical Practice in Film and Television. Duke University Press, Durham; London.

Cantor, M., 1971. The Hollywood TV Producer: His Work and His Audience. Basic Books, New York.

Collins, J., 1989. Watching ourselves watch television, or who's your agent. Cultural Studies 3 (3), 261–281.

Corrigan, P., 1990. On the difficulty of being sociological (historical, materialist) in the study of television: the 'moment' of English television, 1936–1939. In: Syvertsen, T. (Ed.), 1992 and after: Nordic Television in Transition, Report No.10. University of Bergen, Norway, pp. 130–160.

Curran, J., Iyengar, S., Lund, A.B., Salovaara-Moring, I., 2009. Media system, public knowledge and democracy: A comparative study. European Journal of Communication 24 (5), 5–26.

Dayan, D., Katz, E., 1992. Media Events: The Live Broadcasting of History. Harvard University Press, Cambridge, MA.

Ellis, J., 1982. Visible Fictions: Cinema, Television, Video. Routledge and Kegan Paul, London.

Ellis, J., 1999. Television as working through. In: Gripsrud, J. (Ed.), Television and Common Knowledge. Routledge, London, pp. 55–70.

Feuer, J., 1983. The concept of live television: ontology as ideology. In: Kaplan, E.A. (Ed.), Regarding Television: Critical Approaches—An Anthology. University Publications of America, The American Film Institute, Los Angeles, pp. 12–22.

Feuer, J., 1989. Reading dynasty: television and reception theory. South Atlantic Quarterly 88 (2), 443–460.

Feuer, J., Kerr, P., Vahimagi, T. (Eds.), 1984. MTM: 'Quality Television'. British Film Institute, London.

Garnham, N., 1986. The media and the public sphere. In: Golding, P., Murdock, G., Schlesinger, P. (Eds.), Communicating Politics. Mass Communication and the Political Process. Leicester University Press, Leicester, UK.

Garnham, N., 1990. Capitalism and Communication. Global Culture and the Economics of Information. Sage, Newbury Park, CA.

Gerbner, G., 1973. Cultural indicators—the third voice. In: Gerbner, G., Gross, L., Melody, W. (Eds.), Communications Technology and Social Policy. Wiley, New York, pp. 553–573.

Gitlin, T., 1983. Inside Prime Time, first ed. Pantheon Books, New York.

Goffman, E., 1959. The Presentation of Self in Everyday Life. Anchor, New York.

Gripsrud, J., 1995. The Dynasty Years: Hollywood Television and Critical Media Studies. Routledge, London; New York.

Hall, S., 1973. Encoding and Decoding in the Television Discourse. CCCS, Birmingham. Stencilled Occasional Paper no. 7.

Heath, S., Skirrow, G., 1977. Television: a world in action. Screen 18 (2), 7–59.

Himmelweit, H.T., Vince, P., Oppenheim, A.N., 1958. Television and the Child. Oxford University Press, London.

Horton, D., Wohl, R., 1956. Mass communication and para-social interaction. Observations on Intimacy at a Distance. Psychiatry 19 (3), 215–229.

Hoynes, W., 1994. Public Television for Sale: Media, the Market, and the Public Sphere. Westview Press, Boulder, CO.

Jenkins, H., 2006. Convergence Culture: Where Old and New Media Collide. NYU Press, New York; London.

Lotz, A., 2007. The Television Will Be Revolutionized. NYU Press, New York; London.

Maccoby, E., 1951. Television: its impact on schoolchildren. Public Opinion Quarterly 15 (3), 421–444.

Meyrowitz, J., 1985. No Sense of Place: The Impact of Electronic Media on Social Behavior. Oxford University Press, New York.

Morley, D., 1980. The 'Nationwide' Audience. British Film Institute, London.

Morley, D., 1986. Family Television: Cultural Power and Domestic Leisure. Comedia, London.

Murdock, 2005. Building the digital commons: Public broadcasting in the age of the internet. In: Jauert, P., Lowe, G.F. (Eds.), Cultural Dilemmas of Public Service Broadcasting. NORDICOM Gothenburg University, Gothenburg, pp. 213–230.

Newcomb, H., Alley, R.S., 1983. The Producer's Medium. Oxford University Press, New York.

Raboy, M. n.d. Public Broadcasting for the 21st Century. Acamedia Research Monograph No. 17, John Libbey Media, Luton, UK.

Rath, C.D., 1989. Live television and its audiences: challenges of media reality. In: Seiter, E., Borchers, H., Kreutzner, G., Warth, E.M. (Eds.), Remote Control: Television, Audiences and Cultural Power. Routledge, London, pp. 79–95.

Schiller, H.I., 1989. Culture Inc.: The Corporate Takeover of Public Expression. Oxford University Press, New York.

Schramm, W., Lyle, J., Parker, E.B., 1961. Television in the Lives of Our Children. Stanford University Press, Stanford, CA.

Seiter, E., Borchers, H., Kreutzner, G., Warth, E.M., 1989. Don't treat us like we're so stupid and naïve: towards an ethnography of soap opera viewers. In: Seiter, E., Borchers, H., Kreutzner, G., Warth, E.M. (Eds.), Remote Control: Television, Audiences and Cultural Power. Routledge, London, pp. 223–247.

Silverstone, R., 1994. Television and Everyday Life. Routledge, London; New York.

Sklar, R., 1976. Movie-made America: A Cultural History of American Movies. Vintage Books, New York.

Spigel, L., 1992. Make Room for TV: Television and the Family Ideal in Post-war America. Chicago University Press, Chicago.

Syvertsen, T., 1992. Public Television in Transition. University of Leicester. Ph.D. thesis.

Tulloch, J., Alvarado, M., 1983. Dr Who': The Unfolding Text. St Martin's Press, New York.

Uricchio, W., 1989. Rituals of reception, patterns of neglect: Nazi television and its postwar representation. Wide Angle 11 (1), 48–66.

Wheen, F., 1985. Television. Century Publishing, London.

Williams, R., 1975. Television: Technology and Cultural Form. Schocken, New York.

Television: History

James L Baughman, University Avenue, Madison, WI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15578–15583, © 2001, Elsevier Ltd.

Abstract

In the two decades after World War II, television became the most popular mass medium throughout most of the world. The USA developed the most successful commercial model, one that relied on advertisers to sponsor programming primarily intended for large audiences. With a vast comparative advantage in film production facilities, the USA was the largest exporter of TV entertainment shows. Most other nations, after long resisting the US model, not only aired many US series but had one or more commercial networks. Despite its global influence, US television underwent several transformations between 1950 and 2000. Initially, national networks, in combination with major advertisers, dominated the medium. Viewers chose from a limited number of channels, most of which were affiliated with a network. But beginning in the 1980s, cable and other new technologies began to undermine the networks' hegemony if not the very concept of mass audiences for TV programming.

Forecast in novels and science journals since the late nineteenth century, television came to the USA and other developed nations in the decade and a half following World War II. Major manufacturers of radio receivers, led by Radio Corporation of America (RCA), as well as individual entrepreneurs had labored in the 1930s in the USA, Germany, and the UK to perfect a device capable of receiving over the air clear images and sound. US companies had been especially anxious to perfect a receiver that could be mass produced. Military research during World War II had advanced their work.

In the late 1940s, the USA was best positioned to experience a 'television boom.' RCA and other companies were offering television sets that, though expensive, were not beyond the means of most consumers. At the same time, RCA established a nationwide network, National Broadcasting Company (NBC), that offered set owners regular programming. Another TV set manufacturer, Du Mont, similarly founded a TV network, as did the radio networks, ABC and Columbia Broadcasting System (CBS) (Du Mont ceased operations in 1955). The four chains provided programming to individual stations, most of which, because of federal regulations, were not owned by the networks. With TV stations operating in most of the nation's larger cities, consumers began purchasing sets. A strong economy allowed such consumption. So did the postwar boom in marriages and child-bearing, which left many young adults housebound, thus making television relatively cost-efficient home entertainment. Although high income was a strong predictor of set ownership in the first years of the television boom, family size soon became the stronger variable. By the late 1950s just under 90% of all US homes had one or more televisions. In time, owning a television became an expectation throughout the world. That said, television spread more slowly in poor countries because of the relatively high costs of receivers, the absence of electrical service and uneven TV production standards.

In contrast to what emerged in many other nations, the so-called 'American' system followed the US radio model of privately owned broadcast stations and networks carrying programs or series paid for and sometimes produced by advertisers. Oversight by the US Federal Communications Commission (FCC) was nominal. As the medium's patron, advertisers held enormous control over programs in television's first dozen years, especially because demand favored them over networks and stations. They normally sponsored entire shows, often including plugs of their wares. In most instances, advertisers sought the largest possible audience. This came to mean relying on a set of program 'genres,' most of which owed something to radio and motion pictures. That said, a few sponsors self-consciously cultivated a smaller, elite audience by underwriting musical, dramatic, or news programs. Many advertisers and broadcasters initially assumed that the TV viewer was better educated and well to do. Partly because of this misidentification of the audience, which also occurred during the early years of radio, many network programs were culturally ambitious. Original dramas, produced in New York, the nation's theatrical center, appeared every night. Critics later insisted that the USA had experienced a 'Golden Age of Television.' However, the fare most identified with the Golden Age could not, over time, compete for viewer favor with other, less pretentious efforts.

In the late 1950s, US television underwent a series of changes that established industry practice for the remainder of the century. As television spread into smaller communities, demand for dramatic and comedy programs that had followings in the larger cities collapsed. Individual comics like Milton Berle and Sid Caesar, whose humor resonated with eastern viewers, were unable to sustain their early success. The larger, more representative audience preferred filmed programs, including westerns and detective dramas, that were inexpensive imitations of motion pictures. This shift in demand encouraged the movement, already underway, of TV production from New York to California. Television increasingly took on the look of Hollywood as opposed to Broadway. At the same time, a sharp increase in advertiser demand gave the networks and larger stations the upper hand over advertisers in programming decisions. By the mid-1960s, most television programs had multiple sponsors.

Outside of Latin America and the Caribbean, much of the world actively sought to avoid taking up the US system. Much as they had disdained US motion pictures, political and intellectual elites abhorred what they construed to be the vulgarity of US television. As a result, few broadcasters enjoyed the autonomy of their US counterparts. Communist and many third-world countries tightly controlled television's development and production. The Soviet Union and other regimes regarded television, like radio, as state instruments. Other

countries, notably in Western Europe, followed the corporation model of the British Broadcasting Corporation (BBC). Under the British model, the government created an independent broadcasting authority that relied on receiver fees rather than advertising. In the mid-1980s, about two-thirds of Europe's television authorities derived their revenues from receiver fees, as did about half of all African and Asian systems. Most of the time, BBC was free of direct government interference. The UK did permit commercial television in creating the Independent Broadcasting Authority (IBA) in 1955. But commercial TV in the UK, as in many other nations, was subject to far more regulations than in the USA. In the UK, non-government TV endured an excess profits tax that had the effect of encouraging more elaborate cultural programming that rivaled the self-conscious, educational efforts of BBC. The Japanese Broadcasting System (NHK), in contrast, gave commercial TV far greater freedom. Although former British colonies initially embraced the BBC model, political anxieties caused the leaders of most newly independent countries to take up more direct, state control akin to that of the Soviet Union. Then, too, the relatively high cost of starting television service and the absence of domestic commercial investors partly explained the adoption of the statist model in most poor nations. Some third-world leaders sought to put off TV's development indefinitely, only to relent under pressure from urban elites who had encountered television in the West.

Although many if not most nations sought to avoid the US system, few could resist its programming. In the early 1980s, UNESCO reported that most nations, including those with totalitarian or Islamic fundamentalist regimes, offered more entertainment programming than any other type. This tendency created an enormous opportunity for US producers. US-made entertainment series, especially those that emphasized action or broad humor, proved immensely popular. As in the case of motion picture assembly, US television producers had a huge comparative advantage in the emerging global market for programming. The switch of US networks to filmed series created a substantial inventory of exportable entertainment. The USA possessed unequaled production facilities in southern California, as well as the movie industry's large reserve of on-camera and production talent. A large domestic audience and advertising demand, as well as the sale of syndication rights as 'reruns' to individual US stations (and eventually, cable outlets), meant that successful US programs were profitable before they were distributed overseas. They could thus be offered relatively inexpensively. British television production, with somewhat similar advantages, was a distant second as an exporter of TV programming. Most other nations simply could not produce programming of the same standard and appeal. This was especially true in less developed countries where, Elihu Katz and George Wedell estimated in the mid-1970s, imported series constituted just over half of all programming. Indeed, viewers in some countries, Katz and Wedell noted, frequently could select from different US programs aired by competing channels at the same time. One scholar of international communication concluded in 1988, 'A kind of sameness fills TV screens just about everywhere.' Revealingly, a 1999 NATO air strike against a communications center in Belgrade, Yugoslavia, destroyed 123 episodes of three US series.

The Americanization of television throughout much of the world had its critics. One of the first, Herbert I. Schiller, regarded the US's export of programming as part of a conscious attempt by the nation's political and business elites to extend the USA's influence via communication. 'The "have-not" nations,' he wrote, 'stand practically defenseless before a rampaging Western commercialism.' More likely, Katz and Wedell argue, market forces and the law of comparative advantage, and not sinister calculation, accounted for the US's impact on television globally.

In its first decade and a half, US commercial television presented relatively little news programming, very little of which elicited controversy. News programs lost money and drew few viewers. The networks considered informational programming a way of placating influential political and cultural leaders as well as of meeting the FCC's modest public service standard. In receiving TV frequencies or licenses for a fixed term, stations promised to devote a minority of its overall schedule to programs, including news, 'in the public interest.' In theory, license renewal hinged on stations meeting this 'minority interest' standard. In practice, the FCC very rarely scrutinized individual licensees' public service schedule. Nonetheless, station managers were not prepared to ignore the agency altogether. Not surprisingly, television news tended to be uncritical of those in power and, as Nancy Bernhard has argued, supportive of America's Cold War with the Soviet Union.

By the 1960s, US networks and many stations were profiting from certain types of informational programs, specifically early evening newscasts. These were not as popular as some contemporary observers assumed. In the late 1960s, one survey indicated, just over half of all US citizens watched no network evening newscasts during a 2-week period. Yet television news did command the attention of many influential US citizens, who assumed its primacy; TV newscasts also drew increasing advertiser interest. Indeed, nightly newscasts became stations' largest single source of revenue. In the 1970s, the network evening 'magazine,' CBS's *Sixty Minutes*, with its mix of exposes and features, started winning viewers For the first time, a network news program actually competed against evening entertainment shows in popularity rankings. Imitations on other networks followed. Although none could generate the following of *60 Minutes*, because they were less costly than entertainment series, the networks earned more money carrying such programs. All told, the realization that news shows could, in fact, be even more profitable, caused broadcasters to seek more viewers by adopting a less serious news agenda. The line separating news and entertainment blurred.

The first great test of US television news came with the US intervention in the Vietnam war. The medium earned at best a low pass. Popular mythology, perpetuated by surviving correspondents (and their critics), insists that TV brought the war's horrors home every evening to a people glued to their sets. Television reporters, according to this wisdom, criticized the conflict itself and thus helped to end US involvement. Subsequent scholarship by Daniel Hallin and others has largely discredited this self-serving history. As already noted, newscast audiences were never as vast as some imagined. Nor did TV news, for a variety of reasons, show much combat footage. Relatively few reports were carried live. And in the initial years of the US build-up, 1964–66, reporters and news anchors

(or readers) passively if not enthusiastically relayed the positions of the US military and civilian leadership. After 1966, some TV correspondents did challenge, ever so mildly, the progress of the war. Yet television was hardly, as some later asserted, leading public opinion or the news media generally. Other news organizations, specifically the hugely influential *New York Times*, had already started to express skepticism about the Vietnam commitment. And television news failed, Todd Gitlin has contended, to take the budding anti-war movement seriously.

As the USA withdrew forces from Southeast Asia, network television had to cover an even larger crisis. These involved criminal activities associated with the June 1972 break-in of the Watergate apartment complex in Washington DC, by figures tied to the reelection campaign of President Richard M. Nixon. The TV networks, especially CBS, covered the Watergate story. Still, they did not do so with nearly the rapidity and thoroughness of some newspapers. Herbert J. Gans' study of the evening network newscast production in the 1970s stressed the extent to which NBC and CBS depended on the *New York Times* for their news agenda.

US television remained primarily an entertainment medium, and TV's staple product began to change. A 1970 FCC regulation restricted the networks' ownership of entertainment programming. This gave producers more freedom, and some sought to convey the more relaxed cultural climate. To many critics, this shift fostered a second 'golden age.' Shows like *All in the Family* and *The Mary Tyler Moore Show*, in violation of industry practice, dealt with contemporary issues like race relations and the sexual revolution. Dramatic series in the 1980s and 1990s, usually involving law enforcement or medicine, similarly took a more realistic tone.

Despite renewed critical respect for some shows, network television's hold over the US viewer slipped in the latter decades of the century. In the 1950s and 1960s, most US citizens received their TV service over the air. The finite size of the radio band limited the number of on-air stations the typical set could receive. And until the late 1970s, the three networks in effect determined what US citizens watched. Most TV stations remained affiliated with a network. On a given evening, nine out of every ten televisions were tuned to a network program. In the 1980s, however, the networks began losing some of their audience. Although the presence of more independent stations cut into the networks' share, the spread of cable proved the far more decisive blow. By subscribing to a cable service, consumers could greatly augment their choices, originally from four to 12 to 30 or more channels. By 1997, the average US cable system provided 47 channels. This much larger array of choices proved too tempting for most consumers. The percentage of homes with cable rose from 7.6% in 1970 to 49.4% in 1988. By the late 1990s, up to 70% of all households had cable. Others had satellite signal receivers that afforded them even more services; such systems were more common elsewhere in the world. The US network audience shrank to about half of the total evening audience.

Cable's rise had several significant effects on the television. Providing many more channels greatly weakened an original justification – scarcity of service – for broadcast regulation, as did a larger disillusionment with the efficacy of business oversight. The FCC relaxed most rules regarding public affairs programming and eventually, network ownership of entertainment programs. The growth of all-news cable channels – along with the end of the Cold War – caused the network news divisions to reduce their coverage of more serious political and international news in favor of stories more likely to attract viewers. At the same, competition from cable channels caused the established networks to forsake their traditional observance of US prudishness. Trying to hold viewers, some network program producers took more risks in their handling of sexuality and crude language. Compared with their predecessors, broadcasters worried less about offending some viewers, who could always turn to a cable channel rerunning an innocuous 1960s situation comedy. In planning a racy police drama, *NYPD Blue*, producer Steven Bochco recalled being told by an ABC official in the early 1990s 'to come up with something that would at least begin to compete with cable.'

Cable channels offered a range of programs and formats. Even so, one of the ironies of cable's growth was that the most popular cable channels relied excessively on reruns of series originally telecast on the networks. These were accompanied on some channels by professional wrestling and variety shows that targeted young men who delighted in bad taste. Many cable channels practiced 'narrow-casting,' that is, they presented programming intended to engage only sports fans (ESPN) or pre-adolescents (MTV). Although the popularity of no one cable channel could compete with any of the networks, combined the cable outlets constituted a regiment of Lilliputians tying down the three Gullivers.

The networks were also wounded by the entry of new rivals in the 1980s and 1990s. None proved more successful than Fox, whose programs usually ignored industry practice and middle class proprieties. 'We were free from the hereditary constraints at the other networks, the decades of doing things a traditional way,' a Fox vice president explained in 1997. Along with still more networks, WB and United Paramount Network (UPN), Fox frequently violated another industry rule by playing to subgroups of the national audience, that is, younger or African – American viewers.

Although the advent of more TV channels mainly on cable complicated their routines, advertisers in many ways benefited from the more competitive television. Promoters of products more appealing to a specific subgroup, like adolescents or women, welcomed the efficiencies of cable narrow-casting. At the same time, cable's growth restored some power to advertisers in their dealings with the networks. Because younger viewers were thought to be more likely to try new brands and services, product makers in the 1990s were prepared to pay more – sometimes significantly more – to co-sponsor a TV program with the 'right' demographics. That is, they began basing more of their 'media buys' on who rather than how many watched a given show. All in all, this represented an unprecedented shift by the networks' underwriters, most of whom had heretofore prized the size rather than the composition of the audience.

Amid such industry upheaval came dramatic changes in industry proprietorship. After several generations of stable ownership, all three established networks changed hands in the 1980s. Capital Cities, owners of TV stations and newspapers, purchased ABC in 1985 and then sold the network to the Disney entertainment conglomerate in 1995. In 1986, General

Electric took over RCA and NBC; investor Laurence Tisch bought CBS. Westinghouse, an electronics company, secured CBS in 1995 only to allow itself to be combined with Viacom, another entertainment concern, in 2000. Meanwhile, individual stations, some of which had been controlled by the same company or family for several generations, were sold and resold through the 1980s and 1990s. Such franchise transfers increased all the more when Congress passed the Telecommunications Act of 1996, which lifted or weakened FCC regulations, including restrictions on the number of stations individual companies could own.

With only with a few exceptions did US television remain the mass medium that it had been in 1960 or 1975. Annual rituals like the National Football League's Super Bowl commanded huge audiences, as did momentarily popular shows like *Who Wants to Be a Millionaire* and *Survivor*. Otherwise, even top-ranked network programs like NBC's comedy *Seinfeld* had to settle for a much smaller percentage of the total television audience than had comparably ranked shows a generation earlier. The future offered more of the same. The spread of the Internet into more US homes at the century's end portended ever more individualized media services.

The Historiography of Television

The history of television has only begun to command serious scholarship. Until the 1990s, most research on the medium has consisted of highly specialized studies conducted by social scientists testing theories of media influence. The first wave of this research is ably summarized in Leo Bogart's *Age of Television* (1972). Historians, except for those tied to programs focusing on the film or journalism, have all but avoided the medium, despite its immense popularity and seeming influence. Their reluctance is understandable. Few are familiar with the constantly ballooning research on media effects. Moreover, evaluating mass culture itself, as opposed to its influence, has always been a tricky endeavor, much simpler for newspaper columnists than academics. It became all but impossible in the latter decades of the twentieth centuries, as scholars in the humanities and sociology dismissed cultural hierarchies altogether as hopelessly subjective, if not indicators of class and educational attainment. Historians who could pass these obstacles faced more logistical roadblocks. In the USA, corporations have been the major players in the medium's history, and relatively few have opened their records to scholars. Then, too, the sheer volume of TV programming over the years has all but compelled historians to limit their inquiries to case studies (see, e.g., Balio 1990, Spigel and Curtin 1997). Relying on middle-class sources, Lynn Spigel has sensitively examined how US citizens sought to adapt themselves and their families to television. There have been more ambitious histories, notably by Asa Briggs (1961–95) of the BBC and Erik Barnouw's more episodic treatment of US broadcasting (1970). Paul Rutherford has written a very competent history of Canadian television (1990, see also Levy 1999).

Meanwhile, many gaps will need to be closed, including far more examinations of ethnicity and gender. Virtually no systematic investigations have been conducted on advertisers' complex and changing relationships to the medium. Much more research needs to be attempted measuring the relationships between TV news and those in power.

Finally, the tendency of program producers to take up concepts first developed overseas needs to be explored. Successful US series had many imitators over the years. Still, the phenomenal success at the end of the twentieth century of ABC's *Who Wants to be a Millionaire* and CBS's *Survivor*, both modeled after European programs, demonstrated that even US program producers have occasionally engaged in plagiarism. CBS modeled its long-running *Sixty Minutes* after a Canadian program. British programs inspired such popular US series as *All in the Family* and *Sanford and Son* as well as some far less resilient endeavors.

See also: Globalization and World Culture; Media Effects on Children; Media Effects; Media Events; Media and History; News: General; Television: General.

Bibliography

Balio, T. (Ed.), 1990. Hollywood in the Age of Television. Unwin Hyman, Boston.
Barnouw, E., 1970. The Image Empire. Oxford University Press, New York.
Baughman, J.L., 1997. Republic of Mass Culture: Journalism, Filmmaking and Broadcasting in America since 1941, second ed. Johns Hopkins University Press, Baltimore.
Bernhard, N., 1999. US Television News and Cold War Propaganda, 1947–1960. Cambridge University Press, New York.
Bogart, L., 1972. The Age of Television, third ed. Frederick Ungar, New York.
Briggs, A., 1961–95. The History of Broadcasting in the United Kingdom, vol. 5. Oxford University Press, New York.
Gans, H., 1979. Deciding What's News. Pantheon, New York.
Gitlin, T., 1980. The Whole World Is Watching: Mass Media in the Making & Unmaking of the New Left. University of California Press, Berkeley, CA.
Hallin, D., 1986. The 'Uncensored War': The Media and Vietnam. Oxford University Press, New York.
Katz, E., Wedell, G., 1977. Broadcasting in the Third World: Promise and Performance. Harvard University Press, Cambridge, MA.
Levy, M. (Ed.), 1999. La Television dans la Republique, less annees 50. Complexe, Paris.
Rutherford, P., 1990. When Television Was Young: Primetime Canada 1952–1967. University of Toronto Press, Toronto.
Schiller, H.I., 1971. Mass Communications and American Empire. Beacon Press, Boston.
Spigel, L., 1992. Make Room for TV: Television and the Family Ideal in Postwar America. University of Chicago Press, Chicago.
Spigel, L., Curtin, M. (Eds.), 1997. The Revolution Wasn't Televised: Sixties Television and Social Conflict. Routledge, New York.

Temperament

Samuel P Putnam, Bowdoin College, Brunswick, ME, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Temperament represents the earliest manifestations of, and basic foundations for, the developing personality and thus is an essential starting point for the understanding of individual differences in humans and other animals. Coherent patterns of behavior are apparent early in life, are rooted in individuals' biological characteristics, and demonstrate consistency over time and situation, organizing the characteristic ways individuals respond to the environments they encounter. These early tendencies develop with maturation and in response to experience, contributing to adaptive and detrimental outcomes throughout the life course.

Definition

Although theorists differ on the specific boundaries of the concept, a degree of consensus has been achieved regarding defining features of temperament (Shiner et al., 2012). Temperament traits are relatively stable, constitutionally based individual differences in activity, reactivity, and regulation that emerge early in life and are shaped by complex interactions between genetic, environmental, and maturational factors. Stability of temperament may be modest in early years due to maturational processes that alter the ways in which underlying processes are manifest in behavior, and increases substantially at later ages. The term 'constitutional' refers to the biological bases of temperament traits, including activity at the genetic, epigenetic, neural, and physiological levels. Reactivity concerns differences in individuals' characteristic behavioral, attentional, and biological responses to change in the environment. Regulation refers to processes that act as modulators of reactivity, including effortful control of behavior and attention, as well as tendencies to approach or withdraw from stimuli (Rothbart and Derryberry, 1981).

Temperament is viewed as a subset of personality characteristics. Although the two constructs share an emphasis on the endurance of individual differences over time, personality is a wider term that encompasses goals, opinions, ideas, and conceptions of the self; cognitively based attributes that are not typically considered aspects of temperament. This focus on overt behaviors allows temperament to be studied in nonhuman animals and infants, allowing for important insight regarding the neurological and developmental origins of temperament, and recently proposed criteria for temperament includes counterparts in primates and other social animals (Zentner and Bates, 2008).

History

Concepts of temperament are apparent in ancient philosophical perspectives. Rothbart (2011) has drawn similarities between modern approaches to individual differences and Indian Rig Vedas and Chinese notions regarding chi. In Western tradition, Greco-Roman physicians, most notably Galen and Vindician, related aspects of behavior with individuals' physical constitution. The term 'temperament' is derived from the Latin 'temperamentum,' referring to proportional mixture, and the Greco-Roman model viewed temperament as being determined by the relative levels of different bodily fluids, resulting in four main types of persons. The melancholic person was believed to have an excess of black bile, predisposing them to negative moods such as fear and sadness. A predominance of yellow bile was associated with the choleric type, who were highly active, ambitious, and quick to anger. The sanguine individual was a friendly and outgoing pleasure seeker with excess levels of blood; and the phlegmatic person was calm and controlled, their even-tempered nature attributed to a predominance of phlegm. This model endured through the Middle Ages and Renaissance, and references to this typology can be found in the writings of Kant in the eighteenth century.

Pavlov's research on classical conditioning is a foundational pillar to the understanding of nurture in shaping behavior, but Pavlov's appreciation of the importance of nature is less well-known by Western scholars. In conducting his studies regarding conditioning of canines, Pavlov and his colleagues became aware of substantial and consistent differences in the responses of individual dogs to experimental paradigms, proposing differences in threshold for inhibitory and excitatory brain processes as underlying factors, and relating these differences to the ancient typology. For instance, sanguine dogs were those who were adaptive to changes in situation, easily acquiring conditioned responses, whereas melancholic animals demonstrated distress when encountering difficult discrimination tasks, sometimes losing conditioned responses they had previously acquired.

Early Western approaches to individual differences focused on questionnaire responses, and used the statistical technique of factor analysis, which empirically groups large numbers of variables to discern a smaller number of broad factors. The earliest effort in this vein was conducted in 1906 by Heymans and Wiersma. Using ratings made by doctors regarding their adult and child patients, three factors were identified: Activity (degree of energy expended in response to goals or stimuli), Emotivity (tendency to emotional reactions such as fear), and Primary–Secondary Function (tendencies to react impulsively rather than in a more slow and organized fashion). Subsequent factor analytic studies in Britain by Webb and Burt, and in the US by Cattell, revealed dimensions including introversion–extraversion, neuroticism, and will, establishing a tradition

linked to the Big Five or five-factor models commonly used in contemporary personality research.

Normative studies of infant and child development also contributed strongly to contemporary research on temperament. Gesell, through observations of large numbers of children, and Shirley, who intensively investigated small samples, aimed to establish timetables for normal sequences of motor and mental capabilities. In doing so, they became aware of notable variability in the children they studied. These researchers found these differences to be consistent across situation and proposed that the patterns of behavior formed the bases for emerging personality. A follow-up investigation of the infants studied by Shirley when they were adolescents provided support for this notion, as independent descriptions of the teens' personalities were predicted by sketches of their behavior during infancy (Neilon, 1948). These authors elucidated a number of principles regarding individual differences that foreshadowed current work on temperament, proposing that early appearance and consistency of traits, as well as similarity in traits among family members, strongly suggested genetic origins of personality.

The prominence of social learning theories emphasizing reward and punishment in the middle decades of the twentieth century resulted in a relative deficit of research on temperament. Notable exceptions from this period include Bergman and Escalona, who identified differences among children in their sensitivity to low levels of stimulation; Fries and Woolf, who documented differences in neonatal activity level; and Kagan and Moss, who demonstrated consistency of shy and timid behavior from age three through young adulthood (see Rothbart, 2011). Inspired by her findings, Escalona (1968) proposed the concept of 'effective experience,' suggesting that the unique ways that children's nervous systems respond to stimuli can cause identical environments to have very different meanings for different children.

Many temperament scholars would agree that the modern era of temperament research began with the publication of the initial report of their New York Longitudinal Study by Alexander Thomas, Stella Chess, and their colleagues (Thomas et al., 1963). Thomas and Chess were clinicians who noted substantial numbers of children who were experiencing adjustment difficulties despite the apparent absence of environmental challenge, as well as children who seemed to thrive in troubling circumstances, and argued that children differed from the earliest days of life, with these differences predisposing them to a wide range of adaptive and maladaptive outcomes. They initiated a study of infants between the ages of 3 and 6 months, interviewing their parents extensively regarding their children's responses in a variety of situations to quantify what they termed 'primary reaction patterns,' only adopting the term 'temperament' in their later publications. Thomas and Chess followed a sample of over 100 children through early adulthood, publishing their findings in multiple books that continue to influence clinicians, interventionists, parents, and researchers today.

Structure

An important beginning point for understanding temperament concerns identification of the underlying pattern of early individual differences. In addition to allowing for increased parsimony, identification of these basic elements of temperament may be useful in generating information regarding biological factors, as related behaviors may be controlled by common neural and physiological processes. An early organizational rubric was created by Thomas and Chess, who sorted parents' descriptions of their children on conceptual bases, resulting in nine behavioral dimensions: (1) Mood, (2) Approach versus Withdrawal, (3) Adaptability, (4) Intensity, (5) Rhythmicity, (6) Activity Level, (7) Threshold, (8) Distractibility, and (9) Attention Span/Persistence. Concerns, however, have been raised regarding this model. Some of the Thomas and Chess dimensions, such as Distractibility and Attention Span, overlap substantially. Conversely, analyses of scales measuring other dimensions suggested the existence of more than one factor. For instance, items in Mood scales assessing positive emotionality often did not correlate highly with those measuring negative affect.

Rothbart and colleagues have developed a battery of questionnaires that incorporate a revised set of constructs from the Thomas and Chess dimensions with those from other models of child temperament, adult personality, and animal research. Separate instruments designed for infants, toddlers, older children, adolescents, and adults contain upto 20 scales each. Factor analyses of these instruments suggest a common structure involving three broad dimensions across the lifespan (Putnam et al., 2001). A factor referred to as Surgency/Extraversion includes scales measuring activity level, approach, and impulsivity, as well as expressions of positive affect, especially during intense activities. A Negative Affectivity factor involves high levels of, and difficulties recovering from, frustration, sadness, and discomfort; fear is also associated primarily with this factor, but not as strongly as the others, and is also associated with low levels of Surgency. A third factor, labeled Orienting/Regulatory Capacity in infancy and Effortful Control thereafter, consistently contains scales measuring tendencies to focus attention. At younger ages, this characteristic is complemented by enjoyment of calm activities and close contact with others; at older ages, the ability to demonstrate inhibitory control and perform necessary but unpleasant tasks joins attentional control in defining this factor. These factors have been replicated in other cultures and are similar to those that have emerged from factor analyses of observational data gathered in the laboratory and from other measures of temperament and personality, as well as those identified among nonhuman animals, suggesting evolutionary origins of the biological systems implicated in these broad patterns of behavior (Rothbart and Bates, 2006).

An alternative perspective on the organization of temperament involves discernment of temperament types, often done through identification of common combinations of dimensions. In an early example of this 'person-centered' approach, Thomas and Chess (1977) proposed a categorical system, in which children could be characterized as 'difficult' (negative mood, withdrawing, unadaptable, intense, and irregular), 'easy' (positive mood, approachful, adaptive to change, low in intensity, and regular), or 'slow to warm up' (withdrawing, low activity level, low intensity). More recently, a number of researchers have identified a related scheme, in which children and adults can be classified as 'undercontrolled' (quick to

anger, disruptive, impulsive), 'overcontrolled' (shy, low activity level, compliant), and 'resilient' (positive, confident, emotionally stable). Zentner and Shiner (2012) have discussed the relative merits of the typological and dimensional approaches, concluding that both offer unique benefits to the understanding of individual differences.

Measurement

Although self-report questionnaires have been developed for older children and adults, teachers occasionally provide ratings of children, and temperament has been assessed in naturally occurring settings, the large majority of studies have measured temperament via questionnaires completed by parents or observations of children's responses to structured tasks. Questionnaires are advantageous in terms of both their depth and breadth. Parents see their children over long periods of time, so that general patterns of response can be reported upon, and observe their offspring in a wide variety of situations, including those that are logistically and ethically difficult to recreate in the lab setting, allowing for assessment of a large number of dimensions. Questionnaires are also relatively easy and inexpensive to administer, easing demands on participants, and enabling collection of large samples. Parent reports, however, have been criticized over concerns regarding parental bias, inaccurate recall, limited comparison bases, and other factors that may lead to inaccuracy. Structured observations are advantageous, in that they allow researchers to carefully control situations and objectively code specific behaviors. These situations, however, are often brief and somewhat artificial, and the laboratory situation itself may be stressful, influencing child behavior in ways that obscure their typical responses.

Acknowledging the pros and cons of differing measurement techniques, contemporary scholars have adopted a 'components-of-variance' interpretation, viewing parent reports as involving both subjective and objective components (Rothbart and Bates, 2006). Similarly, observational measurements will yield scores influenced by both the child's general tendencies and behavior unique to the specific situation and moment. Lending support to this view, convergence across methods has been demonstrated frequently, but not always (see Putnam et al., in press, for a recent review). Multimethod approaches are recommended whenever possible, but single method studies are still valuable when interpreted in light of their limitations. By comparing and contrasting findings generated with differing measures, a richer perspective can be gained regarding the phenomenon under consideration.

Biological Bases

Emphasis on constitutional origins is common to all definitions of temperament, and knowledge in this area has accelerated in recent years in correspondence with advances in methodological sophistication. A fundamental change concerns the presumed role of genes. Whereas early views conceptualized temperament as largely genetically determined at birth, with increasing influence of the environment over the lifespan, contemporary perspectives acknowledge the dynamic interplay of biological and environmental factors from the moment of conception (Shiner et al., 2012). It is now clear that pre- and postnatal influences alter the expression of genes, that some genetic influences do not emerge until later in development, and that genes are implicated in change, as well as stability, of behavioral patterns.

The findings of quantitative genetics research, which compares the behaviors of individuals differing in the similarity of their genetic profiles (e.g., monozygotic and dizygotic twins), are consistent with the suggestion that approximately 20–60% of variability in temperament traits is due to genetic differences (Saudino and Wang, 2012). In contrast to the intuitive expectation that the role of genes might be more profound at early ages, most studies suggest increases in genetic effects across the lifespan, as individuals select and elicit environments consistent with their genetic makeup. Longitudinal studies with genetically sensitive designs suggest an additional possibility that temperamental characteristics may be influenced by different genes at different times (e.g., before and after puberty).

Enthusiasm regarding molecular genetic studies, in which differing forms of individual genes are associated with temperament dimensions, has waned somewhat, due to inconsistencies in findings and frequent failures of replication studies. Despite the number of conflicting results in this burgeoning area of research, some general patterns are apparent (Saudino and Wang, 2012). Allelic variation in genes linked to dopamine systems has been connected in several studies to surgent and regulatory aspects of temperament, whereas genes associated with serotonin have been most frequently related to individual differences in temperament dimensions involving sociability and affect.

Genetic activity is translated to individual differences in behavior through processes involving the central and peripheral nervous systems. Research into the neural bases of temperament has frequently considered individual differences in the reactivity of neural systems implicated in emotional responses. The most influential studies in this area have been those conducted by Kagan (e.g., 1994) on children who are extreme in behavioral inhibition, demonstrating high levels of hesitance and negative affect in the face of novelty, in comparison to children who are relatively fearless. Relative to uninhibited children, inhibited children tend to exhibit high and stable heart rates, muscle tension, and high levels of norepinephrine; and react to stressful situations with exaggerated cardiac acceleration, pupil dilation, and elevated levels of cortisol. Kagan has argued that this profile is largely due to low thresholds for arousal in sections of the amygdala, and a follow-up study of adults that had been inhibited or uninhibited as young children have indeed revealed greater fMRI signal response in the amygdala in the former when viewing novel, in comparison to familiar, faces (Schwartz et al., 2003).

Regarding regulatory functioning, frontal neural networks, most notably the anterior cingulate cortex (ACC) and lateral prefrontal cortex play central roles (Posner and Rothbart, 2000), and correlations between anatomical aspects of the ACC and self-reported Effortful Control have been documented (Whittle et al., 2008). Regulatory aspects of temperament have also been related to peripheral measures. Variability in heart rate that occurs in response to respiration provides

an estimate of parasympathetic nervous system function, providing a route for regulation of emotional processes. Both high degrees of variability and large reductions in variability during situations requiring constraint have been associated with enhanced attentional and emotional regulation.

Development

Although several studies have now demonstrated consistency of individual differences from early childhood through adulthood (see Zentner and Shiner, 2012), this consistency is relatively modest and a developmental perspective requires the acknowledgment of change as well. Neural and motor systems underlying temperament undergo immense maturation during infancy and childhood, with regulatory processes increasingly laying influence upon more purely reactive systems (Rothbart and Derryberry, 1981). One important shift occurs in the second half of the first year of life, as fear-based inhibition emerges, functioning as a 'brake' on the 'accelerator' of more early developing approach tendencies. Across early childhood, this system is complemented by increases in more effortful regulation of attention and behavior. These later-developing characteristics not only represent traits that can only be assessed after the infancy period; they also allow for more flexible and intentional demonstration of other traits, leading to a degree of instability in these dimensions. Maturation of these capabilities continues after toddlerhood, but occurs at a slower pace. Correspondingly, estimates of stability increase sharply after the age of three, but do not reach a peak until late adulthood (Roberts and DelVecchio, 2000).

To take into account maturation, in addition to changes across development in the social acceptability of affective displays, researchers often consider heterotypic continuity. In contrast to homotypic continuity, which refers to consistency in identical or similar behaviors over time, heterotypic continuity involves connections between behaviors that may differ dramatically in form, but share a common biological or functional origin. For instance, Kagan (1994) found that infants who demonstrated a high motor activity and crying to novel stimuli at 4 months showed restricted activity in their failure to approach an unfamiliar adult when asked to do so as toddlers and were more serious than their peers when interviewed as adolescents. Heterotypic continuity has also been invoked to explain the influence of one trait on the developmental pathway of another, and the direction of influence may change with age. For example, high surgency in infancy has been shown to predict higher levels of effortful control in toddlerhood, but this pattern is later reversed, with high toddler surgency associated with decreased effortful control in older children (Putnam et al., 2008).

Context and Outcomes

The increase in volume of research on temperament in recent decades has been matched by breadth of investigation, and temperament has been integrated with an immense array of other constructs. Due to the correlational nature of much of this research, the causal direction of these relationships is frequently obscure. This is most apparent in studies of the proximal environment. The actions of parents and other persons are presumed to shape children's developmental trajectories, but children also shape the actions of those with whom they interact. Processes of dynamic reciprocity also take place between distinct aspects of the individuals themselves. That is, although phenomena such as adjustment or physical health are usually conceptualized as outcomes of temperament, the reverse is also plausible in many cases. Direction of causality is less ambiguous for certain types of variables. Biological sex and cultural forces are not open to influence from the behavior of individual children. Research on these overarching contexts is covered below, followed by description of research linking temperament with parenting and adjustment, before attention is turned to efforts to apply understanding of temperament to benefit individuals and their families.

Sex and Culture

Although reverse causality is logically improbable, the degree to which cross-cultural or sex differences in temperament are due to genetic or socialization factors is uncertain. Quantitative genetic designs are suited to discerning differences within populations, not between groups, so shed little light on this issue. The fact that some differences have been identified in early infancy suggests that the organizing role of the X/Y chromosomes and population-wide differences in the frequencies of alleles may explain a portion of the variance; alternately, the fact that many sex and cross-cultural differences emerge, increase, or decrease across childhood and adolescence suggests that environments shape these individual differences as well.

A recent meta-analysis of 191 studies by Else-Quest et al. (2006) revealed sex differences in several temperament traits. The largest effects were found among dimensions associated with surgency and effortful control, with males demonstrating higher activity level, impulsivity, and enjoyment of high-intensity stimuli; and females exhibiting a higher degree of inhibitory and attentional control. Girls also tended to express greater positive affect during low-intensity situations and were more sensitive to subtle changes in the environment. Fewer differences were found in aspects of negative affectivity. Whereas males were somewhat more likely to be rated as 'difficult,' girls expressed greater physical discomfort, fear, and sadness, with these differences becoming more pronounced during adolescence. Gender has also been investigated as a moderator of temperament effects. For instance, shyness appears to be more predictive of social difficulties in males than females, possibly due to social norms conferring greater acceptance of timidity for girls in North American cultures.

In general, shyness and social withdrawal are viewed more negatively in individualistic Western cultures than in group-oriented cultures, in which self-regulation is more strongly emphasized (Chen et al., 2012). These values are manifest in the behaviors of social partners, as Chinese parents and peers show more warmth and acceptance toward inhibited than uninhibited children, with the opposite pattern apparent among Canadian parents and youth. Correspondingly, several studies have found children in some Asian cultures to be more

fearful, less expressive, and demonstrate greater behavior control than their counterparts in Canada and the US Similar processes have been documented among individualist societies. For instance, Super and colleagues (2008) found low approach to be associated with parents' ratings of general difficultness in Italian, but not US, samples, which may explain recent findings of lower shyness in Italian than US children (Cozzi et al., 2013). Culture also alters contributions of temperament traits to outcomes. For example, whereas shyness is often linked to socioemotional problems in North America, studies carried out in Indonesia, Sweden, and China have not found this. Curiously, recent studies in China suggest that shyness has begun to relate to adjustment problems in urban parts of the country that have become increasingly market-oriented in the twenty-first century (see Chen et al., 2012). As the world becomes increasingly interconnected, other cultural effects may be similarly affected.

Parenting and Adjustment

Longitudinal studies that examine change in variables after accounting for earlier levels of the attributes have suggested that parenting both affects and is affected by temperament (see Bates et al., 2012, for a recent review). Positive emotionality and regulatory aspects of temperament demonstrate reciprocal effects, with positivity and self-regulation predicting higher warmth and positive forms of control from parents, whereas punitive control or low responsivity from parents forecasts diminished growth of self-regulation in children. Results regarding child negative emotionality influencing parenting are complex, with some studies suggesting decreased warmth in response to negativity, but others suggesting increased warmth, especially toward fearful children. Child fearfulness, in contrast, may decrease over time in response to parental warmth.

Temperament has been extensively investigated in relation to behavior problems. The general pattern of these findings suggests that general negative affectivity contributes to both internalizing and externalizing problems; fearfulness is more strongly connected to internalizing problems; and externalizing tends to be more closely related to high levels of surgency and low effortful control. A handful of studies have also documented interactions among traits. The most common result is for the detrimental effects of negative affectivity to be most profound among youth who are also characterized by low levels of effortful control.

In recent years, a classic perspective regarding temperament, parenting, and adjustment has been extended conceptually and empirically. Thomas and Chess (1977) introduced the term 'goodness of fit' in proposing that a key determinant of an individual's functioning was the degree to which their temperament was congruent with the demands of their environment. A similar perspective regarding temperament × environment interactions, the differential reactivity model (Wachs, 1992), suggests that temperament predisposes certain children to be more reactive than others to aspects of the environment. In this view, some temperament characteristics, such as negative emotionality, confer a vulnerability that leads some children to be particularly affected by stressors, whereas other characteristics, such as effortful control, serve as buffers that allow children to more readily cope with adversity. More recently, a differential susceptibility model has been proposed (e.g., Bakermans-Kranenburg and van IJzendoorn, 2011), in which individual differences in reactivity cause some children to be exceptionally sensitive to both positive and negative influences.

An elegant empirical example of temperament × parenting interaction was initially identified by Kochanska (e.g., 1997). Kochanska found that the development of morality in inhibited children was hampered by use of harsh control by mothers, but discipline style had little effect on the consciences of uninhibited children. For fearless, but not fearful, children, positive mother–child interactions and secure attachment were associated with more advanced morality. A second consistent pattern of findings suggests that children high in frustration or general negativity are more likely than less negative children to exhibit adjustment problems in response to negative parenting. Similarly, high levels of negative parenting (e.g., harsh discipline) or low levels of positive parenting (e.g., warmth) may have more detrimental effects on children low in effortful control. A recent meta-analysis applied the differential susceptibility perspective to molecular genetic findings, finding children who were carriers of 'risk alleles' in dopamine-related genes were more likely to actualize both positive outcomes in response to supportive environments and maladaptive outcomes in the presence of adverse contexts (Bakermans-Kranenburg and van IJzendoorn, 2011).

Transactional interactions between temperament and parenting have also been shown through experimental research. An influential study in this vein was conducted by van den Boom (1995). Building upon studies demonstrating correlations between difficult temperament and insecure attachment, van den Boom conducted an intervention aimed at promoting responsiveness to the emotional signals of difficult infants. In response to increased maternal sensitivity, infants in the experimental group were more likely to be securely attached and cooperative with mothers as toddlers than those in the control group. Furthermore, these children engaged in more responsive interaction with their fathers and were more frequently sought out by peers as play partners. In replications of this work, the effectiveness of interventions has been pronounced for infants who were highest on negative reactivity, but negligible for less irritable children.

Applications

Despite near-universal acceptance of temperament as an influence on development, progress has been slow in translating current knowledge regarding early appearing individual differences to practices designed to improve the human condition. Interventions incorporating temperament are often grounded in the goodness of fit framework, providing information on temperament and helpful strategies that encourage environments that are adaptive to the unique needs of individual children. A complementary goal is to alter child behavior, both by assisting children to use regulation techniques that are specific to problematic reactive tendencies and by enhancing more general capacities associated with effortful

control. Such programs can be designed and implemented at a variety of levels or tiers (McClowry and Collins, 2012), ranging from broad interventions that are aimed at fostering adaptive development of all children in a school or other setting; to more intensive efforts with children demonstrating risk for maladjustment; to strongly focused treatment for children with more advanced indication of behavior disorders.

The INSIGHTS program (see McClowry and Collins, 2012) provides a valuable model of a broad-based intervention. Over a 10-week period, the program proceeds through three steps during which parents and teachers are (1) taught to recognize, reframe, and consider their responses to children's temperament-relevant behaviors, (2) given strategies to elicit compliance that are matched to different traits, and (3) offered guidance on training children to use techniques that allow them to effectively self-regulate. A complementary program is administered to children in the classroom, using puppets to introduce children to the idea of temperament types and teaching coping strategies for challenges that arise for different individuals. The intervention has been shown to be effective, lowering rates of disruptive behavior and attentional difficulties in elementary school children from urban neighborhoods.

In contrast to the broad goals and audience addressed by programs such as INSIGHTS, higher-tier programs often emphasize single dimensions. One such intervention is the 'Cool Little Kids' program (e.g., Rapee et al., 2010), aimed at reducing risk for anxiety disorders among inhibited children. Parents of fearful children meet in small groups with a clinical psychologist to receive information regarding the effects of aspects of parenting such as anxiety modeling and overprotection, and intensive training in techniques such as gradual exposure to fear-eliciting situations. In comparison to a group who received no treatment, children whose parents completed the program had fewer anxiety symptoms when followed up 3 years later.

Once issues have led individuals, couples, or families to seek treatment, temperament can be a valuable tool for addressing a variety of clinical issues. Aron (2012) has articulated multiple ways in which temperament can lend a unique perspective to both therapists and their clients. A consideration of temperament can enhance assessment, enabling clinicians to interpret client behaviors in the context of the person's individual style. During the development of the therapeutic alliance, acknowledgment of a client's temperament may build trust and allow both members of the alliance to consider behaviors objectively, viewing traits in terms of their benefits as well as drawbacks. In addition, awareness of temperament can help provide realistic goals for the outcomes of therapy, so that "both therapist and client can develop a better sense of what can be changed and what cannot, and even should not, be changed" (Aron, 2012: 648).

Conclusion

Since the early 1970s, research on temperament has grown exponentially, with increases in publications on the topic dramatically surpassing growth in scholarship on similar topics such as attachment and personality (Zentner and Shiner, 2012). Empirical research has clearly supported ancient beliefs that the rich variety in the ways in which humans navigate their worlds have origins in biological processes and behavioral patterns that are apparent early in life. Given ongoing advances in technology and increased integration of temperament with other areas of the social and natural sciences, it is likely that growth in our understanding of the nature and implications of temperament will continue to accelerate.

Future advances in the field are sure to take the direction of increased complexity, with more studies of interactions across and within all levels of analysis (e.g., culture, social settings, family relationships, perceptions, behaviors, physiology, neurology, epigenetics, genetics). Across levels, the ways that culture affects temperament through parenting, the ways in which environments alter genetic activity, and the ways in which physiological reactions impact individuals' perceptions of their social partners are mere examples of the infinite questions that can be addressed. With levels, greater appreciation of fine-grained distinctions in biological, behavioral, and social processes will allow greater specificity in exploring, for instance, the ways in which certain traits alter the trajectory of others, or how neural systems react to the activity of one another.

Advances will also be salient in the applications of temperament to multiple settings. Implicitly, this is already occurring, as the majority of professionals in the teaching, counseling, medical, and related professions have received some training regarding temperament and the how it can inform their practices. In addition to enhancing this training, more explicit efforts, building from those recently developed by McClowry, Rapee and others, are needed, along with evaluations of exactly how, why, and for whom they are effective. Adaptations of interventions for use with different populations, as well as modifications to protocols that are based on efficacy studies, can combine with basic research in a circular fashion that holds promise for transformative change.

See also: Infancy and Childhood: Emotional Development; Nature-Nurture Controversy, History of; Neonatal Temperament: From Phenomenology to Mechanism; Neural Correlates of Social Development; Personality Development: Systems Theories; Personality Differences and Development: Genetic and Environmental Contributions; Personality, Biological Models of; Personality, Trait Models of; Temperament Development, Theories of; Temperament and Human Development; Temperament and Motivation.

Bibliography

Aron, E.N., 2012. Temperament in psychotherapy: reflections on clinical practice with the trait of sensitivity. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 645–670.

Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H., 2011. Differential susceptibility to rearing environment depending on dopamine-related genes: new evidence and a meta-analysis. Development and Psychopathology 23, 39–52.

Bates, J.E., Schermerhorn, A.C., Petersen, I.T., 2012. Temperament and parenting in developmental perspective. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 425–441.

van den Boom, D.C., 1995. Do first-year intervention effects endure? Follow-up during toddlerhood of a sample of Dutch irritable infants. Child Development 66, 1798–1816.

Cozzi, P., Putnam, S.P., Ersilia, M., Gartstein, M.A., Aureli, T., Montirosso, R., 2013. Cross-cultural differences in fine-grained aspects of toddler temperament: United States of America (US) and Italy. Infant Behavior and Development 36, 480–483.

Chen, X., Yang, F., Fu, R., 2012. Culture and temperament. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 462–478.

Escalona, S.K., 1968. The Roots of Individuality: Normal Patterns of Development in Infancy. Aldine, Chicago.

Else-Quest, N.M., Hyde, J.S., Goldsmith, H.H., Van Hulle, C., 2006. Gender differences in temperament: a meta-analysis. Psychological Bulletin 132, 33–72.

Kagan, J., 1994. Galen's prophesy. Basic Books, New York. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 315–346.

Kochanska, G., 1997. Multiple pathways to conscience for children with different temperaments: from toddlerhood to age 5. Developmental Psychology 33, 228–240.

McClowry, S.G., Collins, A., 2012. Temperament-based intervention: reconceptualized from a response-to-intervention framework. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 607–626.

Neilon, P., 1948. Shirley's babies after fifteen years: a personality study. The Pedagogical Seminary and Journal of Genetic Psychology 73, 175–186.

Posner, M.I., Rothbart, M.K., 2000. Developing mechanisms of self-regulation. Development & Psychopathology 12, 427–441.

Putnam, S.P., Ellis, L.K., Rothbart, M.K., 2001. The structure of temperament from infancy through adolescence. In: Eliasz, A., Angleitner, A. (Eds.), Advances/Proceedings in Research on Temperament. Pabst Scientist Publisher, Germany, pp. 165–182.

Putnam, S.P., Gartstein, M.A., Rothbart, M.K. Historical background of the study of temperament and new perspectives on assessment. In: DelCarmen-Wiggins, R., Carter, A.S. (Eds.), Oxford Handbook of Infant, Toddler, and Preschool Mental Health Assessment, second ed. Oxford University Press, New York, in press.

Putnam, S.P., Rothbart, M.K., Gartstein, M.A., 2008. Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. Infant and Child Development 17, 387–405.

Rapee, R.,M., Kennedy, S., Ingram, M., Edwards, S.L., Sweeney, L., 2010. Altering the trajectory of anxiety in at-risk young children. American Journal of Psychiatry 167, 1518–1525.

Roberts, B.W., DelVecchio, W.F., 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. Psychological Bulletin 126, 3–25.

Rothbart, M.K., 2011. Becoming Who We Are: Temperament and Personality in Development. Guilford Press, New York, NY.

Rothbart, M.K., Bates, J.E., 2006. Temperament. In: Damon, W., Lerner, R. (Series Eds.), Eisenberg, N. (Eds.), Handbook of Child Psychology, Social, Emotional, and Personality Development, sixth ed., vol. 3. Wiley, New York, pp. 99–166.

Rothbart, M.K., Derryberry, D., 1981. Development of individual differences in temperament. In: Lamb, M.E., Brown, A.L. (Eds.), Advances in Developmental Psychology, vol. 1. Erlbaum, Hillsdale, NJ, pp. 37–86.

Saudino, K.J., Wang, M., 2012. Quantitative and molecular genetic studies of temperament. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 315–346.

Schwartz, C.E., Wright, C.I., Shin, L.M., Kagan, J., Rauch, S.L., 2003. Inhibited and uninhibited infants "grown up": adult amygdalar response to novelty. Science 300, 1952–1953.

Shiner, R.L., Buss, K.A., McClowry, S., Putnam, S.P., Saudino, K.J., Zentner, M., 2012. What is temperament now? Assessing progress in temperament research on the twenty-fifth anniversary of Goldsmith et al. (1987). Child Development Perspectives 6, 436–444.

Super, C.M., Axia, G., Harkness, S., Welles-Nyström, B., Zylicz, P.O., Parmar, P., Bonichini, S., Bermúdez, U.M., Kolar, V., Palacios, J., Eliasz, A., McGurk, H., 2008. Culture, temperament, and the difficult child: a study in seven Western cultures. International Journal of Developmental Science 2, 136–157.

Thomas, A., Chess, S., 1977. Temperament and Development. Brunner/Mazzell, New York.

Thomas, A., Chess, S., Birch, H.G., Hertzig, M.E., Korn, S., 1963. Behavioral Individuality in Early Childhood. New York University Press, New York.

Wachs, T.D., 1992. The Nature of Nurture. Sage, Newbury Park, CA.

Whittle, S., Yucel, M., Fornito, A., Barrett, A., Wood, S.J., Lubman, D.I., et al., 2008. Neuroanatomical correlates of temperament in early adolescents. Journal of the American Academy of Child and Adolescent Psychiatry 47, 682–693.

Zentner, M., Bates, J.E., 2008. Child temperament: an integrative review of concepts, research programs, and measures. European Journal of Developmental Science 2, 7–37.

Zentner, M., Shiner, R.L., 2012. Fifty years of progress in temperament research: a synthesis of major themes, findings, and challenges and a look ahead. In: Zentner, M., Shiner, R.L. (Eds.), Handbook of Temperament. Guilford, New York, pp. 673–700.

Temperament and Human Development

MK Rothbart, University of Oregon, Eugene, OR, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 23, pp. 15586–15591, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Temperament refers to early-appearing dispositions upon which personality is based, and research on temperament offers an important place to begin in the study of human development. Rather than seeing individual differences solely as the result of socialization, temperament research identifies basic dispositions toward affect, arousal, and attention that will be further shaped by experience in development. In this article, temperament is defined, a brief history is presented, current research is reviewed, and directions for the future are outlined.

Because temperament refers to the early-appearing dispositions upon which personality is based, it can be studied in infants and nonhuman animals as well as older children and adults. It offers an important place to begin in understanding human development. Rather than seeing individual differences in personality and social behavior as solely the result of socialization, temperament research identifies basic dispositions toward affect, arousal, and attention that will be further shaped by experience in development.

A Definition of Temperament

Temperament has been defined as constitutionally based individual differences in emotional, motor, and attentional reactivity and self-regulation, demonstrating consistency across situations and relative stability over time (Shiner et al., 2013). The term 'constitutional' stresses links between temperament and biology. Over the long history of the study of temperament, individual differences in temperament have been linked to the constitution of the organism as it was understood at the time. The term 'reactivity' refers to the latency, rise time, intensity, and duration of responsiveness to stimulation. The term 'self-regulation' refers to processes that serve to modulate reactivity; these include behavioral approach, withdrawal, inhibition, and executive or effortful attention.

History of Temperament Research

Temperament study has an ancient history: individual differences in temperament were described in the fourfold typology of the Greco-Roman physician Vindician, who linked temperamental characteristics to Hippocrates' model of the humoral constitution of the body. The term 'temperament' derives from the Latin *temperamentum*, a proportionate mixture, denoting the relative preponderance of one or more of the body humors. In Vindician's typology, the melancholic person, quiet and moody, was seen as having a predominance of black bile; the choleric person, touchy, aggressive, and active, a predominance of yellow bile; the sanguine person, sociable and easygoing, a predominance of blood; and the phlegmatic person, calm and even-tempered, a predominance of phlegm. Ideas about the fourfold typology of temperament persisted throughout the Middle Ages and the Renaissance, and were found in the writings of Immanuel Kant.

Early in the twentieth century, major schools in Europe and the UK contributed to temperament research (Kagan and Fox, 2006). An Eastern European tradition began with I. P. Pavlov's observations of individual differences in his dogs' behavior in the laboratory. On the basis of the dogs' reactions, Pavlov and his followers posited characteristics of the nervous system, including strength or weakness of excitation and inhibition, and nervous system lability. This was nevertheless a conceptual nervous system; the nervous-system characteristics were based on behavioral observations. Work in Eastern Europe continued in Nebylitsyn's studies of individual differences in the laboratory, with adults as subjects. In the late twentieth century, due to a lack of consistency across stimulus modalities and response systems in the laboratory, questionnaire assays of individual differences have been undertaken.

In Central Europe, research on temperament in The Netherlands was to become highly influential. Heymans and Wiersma, in 1908, asked 3000 physicians to observe a family (parents and children) and to fill out a temperament/personality questionnaire on each person. When the more than 2500 responses were analyzed, three factors were identified: activity (the tendency to act out what is thought or desired), emotivity (the tendency to show body symptoms, to be fearful and shy), and primary versus secondary process (the tendency to react immediately or in a postponed and more organized way). This research, which involved a wide range of subjects, was followed by European research on temperament in early development, with intense longitudinal observation of infants. Wallon in France and Meili in Switzerland investigated infants' muscle tension and emotionality, and Meili reported stability across time in infants' irritable or anxious responses to standardized stimuli. Meili's contributions showed a similarity to later research in the USA on temperamental emotions and behavioral inhibition (G. Balleyguier, personal communication, 8 September 1996).

In the UK, factor-analytic studies of individual differences in temperament and personality were carried out using adults' self-reports. Early work yielded factors of introversion–extraversion,

emotional stability–instability (later called by Eysenck 'neuroticism'), and volition or will (Kagan and Fox, 2006). Eysenck added a physiological base to the temperament factors by relating them to cortical excitation and inhibition, and the functioning of the limbic system. Still later, Gray revised Eysenck's theory by positing individual differences in behavioral activation and inhibition, as well as tendencies to fight and flight. He also related these individual differences to an underlying neurophysiology. At the turn of the twenty-first century, Gray's theory was one of the major psychobiological models of temperament, along with models put forward by Panksepp et al. (2005).

In the USA, there were multiple contributions to the study of temperament. Allport used temperament in his trait-based theory of personality, and Cattell and Thurstone carried out early factor-analytic research on temperament and 'vectors of mind.' These would later be followed by research on the Big Five personality factors in adults (Biesanz et al., 2004). Nevertheless, in early research on child development, temperamental dispositions were rarely discussed. Instead, social-learning theories stressed the importance of reward and punishment in shaping behavior and accounting for individual differences. The study of cognitive development would also become much advanced before the study of emotional development and temperament became popular. Only in the late twentieth century did research grow and flourish in temperament and development. This renewed interest has been linked to the realization that the direction of parent–child influence is not only from parent to the child, but that the infant and the young child also bring much to shaping the interaction (Collins et al., 2000). Interest in temperament has also been linked to advances in neuroscience, with knowledge of individual differences in temperament providing links to genes and neural networks, as well as to social interaction.

Current Research

In the USA, the New York Longitudinal Study (NYLS) by Thomas, Chess, and colleagues formed the basis for much of the recent research on temperament in children. Parents of infants between the ages of 2 and 6 months were interviewed about their children's behavior across a wide variety of situations. A content analysis of interview information on the first 22 infants yielded nine dimensions of temperamental variability: activity level, rhythmicity, approach-withdrawal, adaptability, threshold, intensity, mood, distractibility, and attention span-persistence. The goals of the NYLS were chiefly clinical, and no attempt was made conceptually to differentiate these dimensions from one another. Nevertheless, numerous parent-report questionnaires were developed based on the nine NYLS dimensions.

As a result of more recent research, however, major revisions to the NYLS list have been proposed (Rothbart et al., 2000). Both factor-analytic research using NYLS items, and research following a more rational approach such as that of Rothbart, have identified a shorter list of temperament dimensions. These dimensions include positive affect and approach (extraversion or surgency), fearfulness, anger/frustration, attentional orienting, and executive attention (effortful control). These dimensions are particularly interesting because they are evolutionarily conserved affective-motivational and attentional systems. They serve adaptive functions, and can be seen in nonhuman animals as well as in the human species (Panksepp, 2005).

In factor-analytic studies of parent-reported temperament in childhood, three to four broad factors are frequently found (Rothbart et al., 2000). The first of these is surgency or extraversion, which includes activity level, sociability, impulsivity, and enjoyment of high-intensity pleasure. The second is negative affectivity, including fear, anger/frustration, discomfort, and sadness, and the third is effortful control, including attentional focusing and shifting, inhibitory control, perceptual sensitivity, and low-intensity pleasure.

Current research on temperament in childhood makes use of parent-report questionnaires, laboratory assessments of children's behavioral and psychophysiological responses to standardized stimuli, and observations of children's behavior in the home or school. With this information, it has been possible to study the development of temperamental systems, and to relate temperamental functioning to other important outcomes. These include the development of empathy and conscience, the development of behavior problems, children's adaptations to the classroom, their susceptibility to drug and alcohol abuse, and their accident-proneness.

Developmental research also indicates that the emotional systems of temperament and reactive attentional orienting are in place before the development of executive effortful attention (Rothbart et al., 2000). In the newborn, individual differences in irritability and orienting can be observed, and by 2 to 3 months, infants demonstrate clear positive responses to stimulation. Early forms of what will later be called extraversion or surgency are present in the smiling and laughter and rapid approach of infants, and measures of approach tendencies and smiling and laughter at this early age predict children's extraverted tendencies at 7 years. Throughout early development, children who are more extraverted also seem to be more susceptible to anger and frustration.

At 4 months, individual differences in infants' distress and body movement to stimulation predict later fear and behavioral inhibition. By 6 months, it is also possible to predict seven-year-old children's parent-reported frustration and anger by the infant's responses to toys out of reach or behind a plastic barrier. Infants' behavioral approach tendencies are also manifested at 6 months in their latency to reach and grasp objects.

The onset of fear or behavioral inhibition in the last quarter of the first year of life appears to work in opposition to the infant's approach tendencies, so that some infants who formerly rapidly approached novel objects are now slowed in their response to novel stimuli, and may not approach at all. They may also show distress to possibly threatening objects. As with approach tendencies, individual differences in fear-related behavioral inhibition show considerable stability across childhood and even into adolescence (Rothbart and Derryberry, 2002). Longitudinal research has reported stability of fearful inhibition from 2 to 8 years, and from the preschool period to age 18.

Fear-related control of behavior can be seen in the early development of conscience (Kochanska, 2002). Fearful

children are more likely to show early signs of conscience. In addition, fearful children whose mothers use gentle discipline, presumably capitalizing on the child's tendency to experience anxious states, develop highly internalized conscience. More fearful infants also later tend to be empathetic and susceptible to guilt reactions. Fear can thus be seen as a basic control mechanism that is importantly implicated in socialization.

Beyond the inhibitory control provided by fear, a later-developing attentional system makes a crucial contribution to socialization. As executive or effortful attention develops in the second or third year of life and beyond, individuals can voluntarily deploy their attention, allowing them to regulate their more reactive tendencies (Gartstein et al., 2013; Rothbart 2003). In situations where immediate approach is not allowed, for example, children can limit their attention to the rewarding properties of a stimulus, thereby resisting temptation and delaying gratification. Similarly, when faced with a threatening stimulus, children can constrain their fear by attending to environmental sources of safety as well as threat. In both of these examples, individual differences in attention allow children to suppress their more reactive tendencies, take in additional sources of information, and plan more efficient strategies for coping. These capacities have been referred to as effortful control, and they refer to temperamental characteristics that are distinct from extraversion/surgency and the negative emotions. Research indicates some stability of individual differences in effortful control during childhood. For example, the number of seconds delayed by preschool children while waiting for rewards that are physically present has predicted children's parent-reported attentiveness and ability to concentrate as adolescents.

In older children and adults, temperament is often studied through self-reports, allowing analysis of internally experienced feelings as well as behavior. The self-report method is also frequently used in the study of adult personality. In adult research, highly differentiated scales assessing temperament have been factor analyzed, yielding factors very similar to those found in Big Five or Five Factor Model research on adult personality. The Big Five personality factors have been derived from research using trait-descriptive adjectives, following the hypothesis that significant individual differences will be represented in the lexicon, that is, in the words people use to describe others and themselves. The Five Factor Model emerged from factor analyses of large numbers of personality self-report items. There is now considerable agreement for five broad factors of personality derived from this method, including extraverversion, agreeableness, conscientiousness, neuroticism, and openness to experience.

In research using adults' self-reports to scales assessing temperamental reactivity and self-regulation, three resulting factors are similar to those found in children and to measures of the Big Five personality factors. These include temperamental surgency or extraversion, positively related to personality extraversion; negative affectivity, related to neuroticism; and effortful control, related to conscientiousness. In addition, self-reported temperamental orienting to external and internal stimuli is related to personality openness to experience, and temperamental affiliativeness to personality agreeableness. Thus, the structures emerging from temperament research – using basic psychological processes of affect, arousal, and attention – and from personality research – using a lexical or personality-scale analysis – are closely related. Because temperamental individual differences are based upon fundamental psychological and neural processes and are present early in life, these findings suggest that early temperamental predispositions may form a core around which the later-developing personality will be built.

As noted above, temperamental systems are evolutionarily conserved. It is interesting that recent research has attempted to apply 'personality' constructs to studies of nonhuman species. A review of factors identified in studies of twelve nonhuman species has found support for factors of extraversion, including energy and enthusiasm; neuroticism, including negative affectivity and nervousness; agreeableness, including altruism and affection; and openness, including originality and open-mindedness (Gosling and Vazire, 2002). These are seen in the chimpanzee, monkey, hyena, dog, and cat. Several species tested did not give evidence of attentional openness, and evidence for individual differences in conscientiousness was reported only in chimpanzees. It seems likely that these studies, although labeled studies of personality, are getting at the evolutionarily conserved temperamental systems studied in human temperament research. In addition, since not all capacities appear to be shared across species, especially conscientiousness or effortful control, we can recognize important phylogenetic *differences* in temperament. For example, the capacity for effortful control, when linked with language, provides opportunities for self-regulation of reactive systems in humans that are not present in other species.

Advances in neuroscientific imaging technology and in our knowledge of brain networks underlying the emotions and attention have been an additional source of information for students of temperament and development. There is now detailed knowledge of networks subserving fear as well as both reactive orienting and effortful attention (Rothbart et al., 2000). Because imaging studies allow researchers to identify tasks that activate these brain networks, it is possible to adapt these tasks to children of different ages in order to study the development of a temperamental system (Rothbart, 2003). This marker-task approach has been used in the study of the development of attentional orienting and effortful control, but it is likely that marker tasks can be used to assess many additional individual differences. In late twentieth-century research, marker tasks of effortful attention in the laboratory were positively related to parents' reports about children's ability to control attention and emotion.

Methodological Issues or Problems

The usual issues and concerns about psychological measurement apply to the study of temperament. Because the focus of study in temperament is individual differences, it is important to design measures that will demonstrate variability. It is also important that the tests are reliable, giving consistent measurement, and valid, reflecting something meaningful about the child's dispositions. The latter is often demonstrated in a psychological measure's convergence with other measures.

In temperament research, researchers have often been skeptical about using parents as informants about their children's behavior. It has been felt that parental biases or lack of

knowledge will yield measures that are invalid, with direct observation seen as a preferable method. However, considerable evidence indicates convergence between parent report and observational measures (Rothbart et al., 2000). In addition, laboratory observations have their own limitations. For example, it is difficult to collect extensive information about children's emotionality in the laboratory when one emotional experience has carryover effects on another. Fearful children may also behave differently in a strange laboratory situation than less fearful children, because expressions of temperamental dimensions are not independent of one another.

Probable Future Directions of Theory and Research

Temperament is a basic integrative construct within psychology, and improvements in temperament concepts and methods will be critical to advances in understanding development. One goal will be a thorough understanding of the basic processes of temperament, including the components of positive affect/approach, fear, frustration, affiliation, and attention, and how they develop. Understanding the affects and attention will benefit the understanding of temperament, just as temperament analyses will benefit our understanding of the affects and attention. Many advances will come from areas of affective and cognitive neuroscience, where much progress has been made in understanding the emotions and attention. Progress in the study of brain structure and its relation to molecular genetics will also be of great importance.

At the same time, improved assessment methods are needed at multiple levels: measures of behavior in the laboratory, including marker tasks for the development of brain structures, behavior in naturalistic daily contexts, physiological measures, informant reports, and self-reports. Increased use of the laboratory, using marker tasks, brain imaging, and recording, is likely to provide further links with underlying neural systems. Questionnaire measures involving self-report will continue to be important.

To further understanding, it will also be necessary to continue to develop the theory of temperament and its development. Because temperament encompasses organized systems of emotional and attentional processes, rather than separate traits, studies of relationships among temperament variables allow a much richer view of development. In this approach, continuity and change within systems, as well as developmentally changing links among the systems, must be considered. Studies of developmental pathways between early temperament and later personality outcomes will be complex, because childhood and adult individuality unfolds in the context of social relationships, and continuity and change cannot be understood without considering the developmentally changing impact of social experience. To understand developmental pathways, it will be necessary to disentangle complex interaction effects among early temperament predispositions, socialization processes, relationships, and culture.

Because of those complexities, temperamentally different children may arrive at similar or equivalent outcomes via different pathways. Pathways of temperamentally different children may converge over development. For example, Kochanska (2002) has found that fearful toddlers whose parents used gentle discipline, and fearless toddlers whose parents capitalized on positive motivation in a close relationship, attained comparable levels of conscience. Temperamentally similar children's developmental pathways may also diverge as a result of different experiences within relationships or varying cultures.

We have entered a very exciting era in developmental psychology. With support from flourishing neuroscience, genetics, developmental psychopathology, and behavioral research in development, the twenty-first century promises to offer new understanding of the uniqueness of pathways of individual growth. Temperament research will make major contributions to this understanding.

See also: Extraversion; Neonatal Temperament: From Phenomenology to Mechanism; Neuroticism; Personality Development and Temperament; Personality and Adaptive Behaviors; Temperament Development, Theories of.

Bibliography

Biesanz, Jeremy C., West, Stephen G., 2004. Towards Understanding Assessments of the Big Five: Multitrait-Multimethod Analyses of Convergent and Discriminant Validity Across Measurement Occasion and Type of Observer. Journal of Personality 72 (4), 845–876.

Collins, W. Andrew, et al., 2000. Contemporary research on parenting: the case for nature and nurture. American Psychologist 55 (2), 218.

Gartstein, Maria A., et al., 2013. Cross-cultural Differences in the Development of Behavior Problems: Contributions of Infant Temperament in Russia and US. International Journal of Developmental Science 7 (2), 95–104.

Gosling, Samuel D., Vazire, Simine, 2002. Are we barking up the right tree? Evaluating a comparative approach to personality. Journal of Research in Personality 36 (6), 607–614.

Hirvonen, Riikka, et al., 2013. The role of temperament in children's affective and behavioral responses in achievement situations. Learning and Instruction 27, 21–30.

Kagan, Jerome, Fox, Nathan A., 2006. Biology, culture, and temperamental biases. Handbook of child psychology.

Kochanska, Grazyna, et al., 2002. Guilt in young children: Development, determinants, and relations with a broader system of standards. Child development 73 (2), 461–482.

Panksepp, Jaak, 2005. Affective consciousness: Core emotional feelings in animals and humans. Consciousness and cognition 14 (1), 30–80.

Rothbart, Mary K., et al., 2003. Developing mechanisms of temperamental effortful control. Journal of personality 71 (6), 1113–1144.

Rothbart, Mary K., Derryberry, Douglas, 2002. Temperament in children. Psychology at the turn of the millennium 2, 17–35.

Rothbart, Mary K., Derryberry, Douglas, Hershey, Karen, 2000. Stability of temperament in childhood: Laboratory infant assessment to parent report at seven years. Temperament and personality development across the life span 85–119.

Shiner, Rebecca L., DeYoung, Colin G., 2013. The Structure of Temperament and Personality Traits: A Developmental. In: The Oxford Handbook of Developmental Psychology, vol. 2: Self and Other 2, p. 113.

Temperament and Motivation

Sari Mullola, Mirka Hintsanen, and Liisa Keltikangas-Järvinen, IBS, Unit of Personality, Work and Health Psychology, University of Helsinki, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Both temperament and motivation have significant outcomes in somatic, societal, and mental well-being. The relationship between temperament and motivation is a two-level phenomenon; genetic/neurobiological and environmental/developmental. Both temperament and motivation are, in part, seen to develop from the reactivity of underlying neurobiological systems, which influence individual differences in attentional, motivational, and self-regulatory functions further. As a result of these core influences, individuals with different temperaments will be exposed to and become motivated by different types of information and environmental stimulation, leading to different content within their 'motivational and developmental representations' as defined by Derryberry and Rothbart.

The Concept of Temperament

There are several temperament theories and models that might differ even in the most essential points of view and define the characteristics of this concept differently. All prevalent and distinguished theories (for a review, see Strelau, 1998), however, agree that temperament refers to a biologically based innate behavioral style, which becomes evident in early childhood and is relatively stable across different situations and over the course of time. The nature of temperament and its differences from ability and motivation have been clarified by defining the temperament as the 'how of behavior,' while ability refers to the 'what and how well of behavior' and motivation reflects the 'why of behavior' (Thomas and Chess, 1977, p. 9).

Although disagreements exist, temperament is usually seen to be relatively independent of cognitions (Oliver et al., 2007). Only low to moderate correlations between traditional measures of IQ and temperament have been found (for reviews, see Guerin et al., 2003; Kristal, 2005), or IQ has been shown to moderate the relation between temperament and academic achievement (Oliver et al., 2007).

Personality and temperament have also been differentiated by some personality researchers (Allport, 1937; Cloninger, 2004). Then, biologically based temperament has been seen as raw material that forms an emotional basis for the later development of personality, which in turn includes cognitions about self and others as well as social factors like values, attitudes, and coping strategies (Rothbart, 2007).

Origins of Temperament

Genetics Background

Temperament has been demonstrated to have a strong biological basis. Although environmental factors, such as culture and education, for example, have significant influence in the manifestation of temperament traits, twin, family, and adoption studies have shown that approximately 50–65% of the variation between individual differences in temperament and personality is due to genetic variation (Gillespie et al., 2003). Research on molecular genetics of temperament traits was initiated in late 1990 by the findings of Ebstein et al. (1996) and Noble et al. (1998), who both revealed an association between Novelty Seeking and Dopamine receptor D4 (DRD4) polymorphism.

Novelty Seeking is one of four temperament traits of Cloninger's Psychobiological Model of Temperament and Character (Cloninger et al., 1993). Other traits of this theory are Reward Dependence, Harm Avoidance, and Persistence. Novelty Seeking is a tendency toward exploratory activity, impulsive decisions, and intense excitability in response to novel stimuli. Harm Avoidance refers to a tendency to inhibit behavior and avoid punishment and novelty, while Reward Dependence is a tendency to respond intensely to social rewards and to learn to maintain rewarded behavior. Persistence is a tendency toward perseverance of behavior despite frustration and exhaustion. These traits predispose a person to specific stimuli (Novelty Seeking to danger, Harm Avoidance to punishment, and Reward Dependence to social reward) and certain emotional reaction when faced with those stimuli. Most important in the context of biological basis of temperament, these traits are suggested to be associated with a biased secretion of specific neurotransmitters, Novelty Seeking with Dopamine, Harm Avoidance with Serotonin, and Reward Dependence with Norepinephrine (Cloninger et al., 1993).

In line with the theory and previous studies, associations of Novelty Seeking with DRD4 (Becker et al., 2005; Dan et al., 2004) and with Dopamine D2 receptor (DRD2) (Comings et al., 1993; Noble et al., 1998) have been supported in later studies while null findings also exist (e.g., Kluger et al., 2002; Light et al., 2007; Szekely et al., 2004). Evidence on associations of Harm Avoidance with Serotonin and Reward Dependence with Norepinephrine have been more conflicting (Becker et al., 2007; Sen et al., 2004). An association of Harm Avoidance with Serotonin transporter 5-HTTLPR, however, has been convincingly demonstrated (Van Gestel et al., 2002). Further, it has been shown that Harm Avoidance by the T102C Serotonin receptor 2A (HTR2A) gene interaction contributes to many adverse developmental paths (Jokela et al., 2007).

In addition to the links suggested by Cloninger's theory, other genetic associations have been found. Harm Avoidance,

for instance, has been shown to be associated with higher μ-opioid receptor availability in regions involved in the regulation of anxiety as well as in the control of emotions (Tuominen et al., 2012). Further, the interaction between catechol-*O*-methyltransferase (COMT) and Serotonin receptor 2A, in turn, might contribute to Novelty Seeking (Salo et al., 2010).

Until now, the findings on molecular genetics have not been consistent enough to definitely associate temperament to its suggested neurobiological origins, but rather, they are likely to generally refer to a neurobiological basis of temperament, and suggest that further research is needed. In addition, single gene associations are no more in the current interest of research, but instead, gene families and interactions of several genes are focused on.

Environmental Factors

The direct associations of various childhood environmental factors (such as parental depressive symptoms, hostile child-rearing attitudes, insecure attachment style, and alcohol intake) with the phenotype of temperament have been convincingly shown (e.g., Pesonen et al., 2004). Although direct environmental and genetics effects on the development of temperament have been demonstrated, the recent research has increasingly focused on interactions, that is, in gene-by-environment interactions predicting the development of temperament, and in temperament-by-environment interactions predicting the development of personality or adulthood societal and health outcomes of temperament.

Interactions between environmental and genetic factors predicting Novelty Seeking have been recently demonstrated in many studies, for example, by Keltikangas-Järvinen et al. (2009). They have shown that after 21 years of follow-up, the participants carrying any A1 allele of the DRD2 gene scored higher on Novelty Seeking than carriers of the A2/A2 genotype, when the childhood environment was strict and punitive. However, when the childhood environment was more encouraging, no genotype effect was found. Furthermore, the presence of the A/A haplotype of the TPH1 intron 7 A218A and A779C genes has predicted higher Harm Avoidance in adulthood when the childhood environment has been hostile (i.e., emotional rejection, maternal neglect, and harsh and inconsistent discipline) (Keltikangas-Järvinen et al., 2007).

Childhood temperament and environment may have direct or interactive effects on health. Anger of the child, for example, has been related to increased systolic blood pressure and high levels of high-density lipoprotein cholesterol (HDL-C) in serum (Ravaja et al., 2001). Childhood temperament and maternal child-rearing attitudes, in turn, have been found to contribute to the early development of insulin resistance syndrome (IRS) in a gender-specific way among children, thereby potentially increasing the risk of coronary heart disease later in life. Among boys, 'difficult temperament' (i.e., anger, activity, and low cooperativeness) increased the risk of having IRS risk-factor clustering, and hostile maternal child-rearing attitudes predicted increased HDL-C. Among girls, there were no significant cross-lagged effects between the behavioral factors and IRS, but child's low emotional significance to the mother increased the risk of IRS in the child (Ravaja et al., 2001).

Regarding a childhood temperament-by-environment interaction, a number of longitudinal studies have demonstrated a close interplay between the child's own, partly genetically determined temperament and environmental factors related to mothering, such as maternal role satisfaction, in the development of self-esteem and depressive tendencies in youth, and dispositional optimism–pessimism in adulthood (e.g., Heinonen et al., 2003).

Significance of Temperament in Academic Achievement

Student temperament has been shown to be a significant factor in studying situations and academic settings. For example, it plays a role in students' social functioning and peer relations in school environment (Hintsanen et al., 2010) and is associated with unemployment in young adults (Hintsanen et al., 2009). In daily classroom life, student temperament is associated with teacher–student interactions (e.g., Rudasill and Rimm-Kaufman, 2009), teacher expectations (e.g., Stipek, 2002), and teacher's educational decisions concerning a student (Rothbart and Jones, 1998). Consequently, temperament has been shown to predict academic achievement evaluated by both standardized achievement tests and teacher-assigned school grades (Guerin et al., 2003). In a series of recently published studies in a population-based sample of Finnish upper-comprehensive schools (e.g., Hintsanen et al., 2012; Mullola et al., 2014), approximately a proportion of one-third of teacher-assigned school grades in mathematics and mother language were explained by student's teacher-perceived temperament. The associations and their effect sizes were in line with the previous literature (for a review, see Guerin et al., 2003) and the given boundary values (Cohen, 1988), and thus suggested moderate to rather strong associations between student temperament and school grades (Mullola et al., 2014).

There are at least five mechanisms through which temperament may be associated with academic achievement. First, temperament may be mediated through an individual's natural working style (for a review, see Kristal, 2005), chosen problem-solving strategies (Davis and Carr, 2002), interests (Elliot and Thrash, 2002; Guerin et al., 2003), and willingness to approach learning tasks, especially the novel ones (Caspi and Shiner, 2006). Activity (i.e., tendency to show frequent and intense motor vigor and tempo), persistence (i.e., long attention span and a tendency to continue seeking a solution to difficult learning or performance problems), and distractibility (i.e., the ease with which a student's attention can be interrupted by low-level environmental stimuli) form a constellation called temperamental task orientation, which affects a student's attention span and tolerance to frustration when learning new skills as well as the ability to complete academic activities (e.g., homework) (Martin and Bridger, 1999). Therefore, high task orientation (i.e., low activity, high persistence, and low distractibility), might, for example, enable the student to have more assisting tools for self-regulating his/her learning process and for enhancing opportunities for the successful accomplishment of school tasks (for reviews, see Guerin et al., 2003; Kristal, 2005). On the contrary, the student with low task orientation (i.e., high activity, low persistence, and high

distractibility) is easily 'sidetracked' from the given learning task, particularly from a difficult one. It is challenging for him/her to maintain attention and to be seated still throughout long periods of time, which requires high concentration and involves occasions of frustration (Martin and Bridger, 1999).

The second mechanism that links temperament and academic achievement is that teachers experience certain temperament traits as more 'teachable' (i.e., a teacher's perception of the temperamental attributes of an ideal student) than other traits, reflecting the 'school appropriate behavior' (i.e., follows directions, is attentive and alert to classroom events, and finishes tasks on time) (for a review, see Keogh, 2003). This further affects teachers' teaching practices and grade giving toward the given student, which may even lead to the so-called 'self-fulfilling prophecies' (Jussim and Harber, 2005). In addition to task orientation mentioned above, other factors of teachability are personal–social flexibility (i.e., the combination of approach–withdrawal/inhibition, adaptability, and level of positive mood) and reactivity (i.e., the combination of threshold of response (which is the intensity and strength of the stimulation that is needed to arouse a reaction), level of negative mood, and intensity of response (which is the expressive and reactive energy of a reaction, whether happy, sad, or angry) (Keogh, 2003; Kristal, 2005). Personal–social flexibility contributes to how fluently changes and transitions go during the school day. Reactivity, in turn, contributes to teachers' opinions of a student (e.g., how much a teacher likes a student), and is also associated with a teacher's perception of a student's personal–social competencies (for a review, see Keogh, 2003) and social position in the class (Hintsanen et al., 2010).

Students with low task orientation, low personal–social flexibility, and high reactivity are consistently perceived by their teachers as less teachable and as having poor educational competence (i.e., teacher-perceived student cognitive ability, motivation, and maturity) (Keogh, 2003; Mullola et al., 2010). Teachers have been found to underestimate the intelligence, motivation, maturity, and cognitive ability of students with low teachability compared with the results of standardized cognitive tests (for a review, see Keogh, 2003). Student's lower teacher-perceived learning style and teachability (i.e., lower 'how of behavior' as defined by Thomas and Chess, 1977, p. 9) are seen as less 'school appropriate behavior.' Further, these students are perceived by their teachers as less motivated (i.e., lower 'why of behavior'), less capable, and less mature. Instead, students with high teacher-perceived teachability (i.e., high task orientation, high personal–social flexibility, and low reactivity) are experienced positively by teachers regarding their learning style. Consequently, they are perceived as more motivated, mature, and capable as well beyond their actual age and performance (Martin and Bridger, 1999).

Third, the association between student temperament and academic achievement may be mediated through teacher-perceived teachability. As teacher-rated temperament seems to be more strongly associated with school grades than students' self-rated temperament (Hintsanen et al., 2012), it is possible that the grade giving may be biased by teachers' perceptions of student temperament (the so-called 'halo-effect'). Low task orientation, in particular, correlates with lower grade point averages (Ham et al., 2006) and lower teacher-rated school grades (Mullola et al., 2010), even when students' self-rated temperament (Mullola et al., 2014), and cognitive ability and motivation (Hintsanen et al., 2012) have been controlled for during the analyses. Student's teacher-perceived high personal–social flexibility (positive mood and high adaptability), in particular, have been shown to predict higher grades than might be expected on the basis of standardized achievement tests (for a review, see Keogh, 2003). Teachers' ratings of student teachability and school grades may also vary by their gender, age, and by student gender (Mullola et al., 2011, 2012). Additional research is, however, needed in order to get more information about the possible bias in grade giving. It is also essential to examine the associations of teacher- and self-rated temperament traits with objective test results and to compare these associations with associations with the grades given by teachers.

The forth mechanism that links temperament and academic achievement is the possibility that the 'halo effect' runs in the opposite direction. This means that temperaments of gifted or high-ability students are evaluated more positively. Therefore, in future studies it is important to control students' standardized cognitive ability test performances during the analyses.

The Significance of Motivation in Education and Well-Being

Motivation is one of the most significant factors in educational and occupational settings (Hazrati-Viari et al., 2012; Komarraju et al., 2009; Tanaka et al., 2009) and also in general well-being related to important life goals, such as personal growth, relationships, community, health, and wealth (Gomez et al., 2012; Lüdtke et al., 2009). Salmela-Aro and colleagues have prospectively examined the effects of different goal contents (i.e., family-focused or self-focused) on subjective well-being in different age groups and life transitions (Salmela-Aro and Nurmi, 1997; Salmela-Aro et al., 2001; Salmela-Aro et al., 2009). The level of mother-perceived depressive symptoms during and after pregnancy, for example, has been predicted by family- and self-focused personal goals (Salmela-Aro et al., 2001). Subjective well-being in young adulthood have been predicted by an individual's interest in family-related goals, whereas low self-esteem have been predicted by interest in self-related goals and by low subjective well-being, further (Salmela-Aro and Nurmi, 1997). In another recent study, the importance of intrinsic goals has functioned as a mediator between extraversion and subjective well-being particularly in older adults, which also indicates some age-specificity in the dynamic interaction of temperament and motivation with general well-being (Gomez et al., 2012).

Research has shown that highly motivated individuals are more alert to their learning processes and outcomes (Bouffard-Bouchard et al., 1991), choose more challenging tasks (Chang and Burns, 2005; Zimmerman and Schunk, 2009), and set more energy and persistence to learn demanding tasks on their own when trying to achieve higher levels of skills and positions (Schunk and Hanson, 1985) than poorly motivated individuals. Further, individuals who are highly motivated experience deeper satisfaction and more positive affect than poorly

motivated individuals when given the occasion to learn and advance (Zimmerman and Kitsantas, 1999). Consequently, these aspects enhance an individuals' satisfaction in educational and occupational circumstances, which is widely demonstrated to be associated with various forms of general well-being and subjective gratification in life (Gomez et al., 2012; Lüdtke et al., 2009).

The Interaction between Temperament and Motivation

Previous literature nominates at least four central temperament and personality theories, whose concepts are suggested to be associated with each other and with motivation. These theories are (1) Five Factor theory (McCrae and Costa, 2003); (2) the Psychobiological Model of Temperament and Character (Cloninger et al., 1993); (3) The psychophysiological basis of introversion–extraversion (Gray, 1970); and (4) Reactive and effortful processes in the organization of temperament (Derryberry and Rothbart, 1997). The interactive biological processes between temperament and motivation influence behavioral functions, which manifest particularly in educational and occupational environment and also in subjective well-being. In most theories, temperament is suggested to be heritable, observable in early childhood, relatively stable across the life span and predictive of adolescent and adult unique behavior (Cloninger et al., 1993; Elliot and Thrash, 2002; Eysenck and Eysenck, 1987; Gray, 1982; McCrae and Costa, 2003). The above-mentioned four theories all contain an affective element, and the conceptualizers of the theories have used the term temperament to describe the nature of their structures (Cloninger et al., 1993; Elliot and Thrash, 2002; Eysenck and Eysenck, 1987; Gray, 1982; McCrae and Costa, 2003). Some of those theories make a distinct difference between temperament and personality, while others emphasize their overlapping and developmental paths between them. In the following text, a clear difference has not been made.

The first of the theories that link temperament and motivation focuses on trait adjectives (Eysenck and Eysenck, 1987; McCrae and Costa, 2003). The Big Five (McCrae and Costa, 2003) and the Big Three (Eysenck and Eysenck, 1987) are the most common models representing this line of research. The big five factors are neuroticism (i.e., tendency to be distressed, worried, nervous, and insecure and hold low opinion of oneself), extraversion (i.e., tendency to be outgoing, sociable, and enthusiastic but also somewhat impulsive and heedless), openness to experience (i.e., tendency to be imaginative, curious, analytical, and having insight and a broad perspective on life), conscientiousness (i.e., tendency to be hardworking, self-disciplined, responsible, reliable, and well-organized), and agreeableness (i.e., tendency to be warm, cooperative, helpful, patient, empathic, and sincere) (McCrae and Costa, 1987), whereas the big three is composed of neuroticism, extraversion, and psychoticism (i.e., tendency to be highly impulsive, aloof, nonconforming, aggressive, and lacking responsibility vs cooperative, empathic, tender-minded, and conventional) (Eysenck and Eysenck, 1987). Psychoticism has associations mainly with Agreeableness and Conscientiousness from the Big Five, but also with Extraversion and Neuroticism via its strong impulsivity component (McAdams, 2009). Conscientiousness and Openness have been found to explain 17% of the variance in intrinsic motivation (Komarraju et al., 2009), which is seen as the most significant motivation type particularly in educational and occupational settings. Of the variance in extrinsic motivation, Conscientiousness and Extraversion have been shown to explain 13%, whereas of the variance in amotivation, Conscientiousness and Agreeableness have been found to explain 11% (Komarraju et al., 2009).

The second theory that may link temperament and motivation is founded on Psychobiological Model of Temperament and Character developed by Cloninger et al. (1993), which has been mentioned before. The theory conceptualizes the variation between individuals by facilitative and inhibitory motivational systems that are basic to the structure of temperament and personality. Cloninger's temperament traits are suggested to be linked to brain activity with perceptual memory systems associated with different activities of neurotransmitters (Novelty Seeking with low basal dopaminergic activity, Harm avoidance with high serotonergic activity, Reward dependence with low basal noradrenergic activity, and Persistence with no special neural correlates, as told before) (Cloninger et al., 1993). Therefore, they are also indicated reflecting preconceptual or subconscious biases in the emergence of motivated behavior and in learning (Cloninger et al., 1993). The functional interaction of these dimensions (i.e., multidimensional temperament profiles) manifests in individual differences in information processing and associative learning (i.e., operant conditioning of habits and automatic and affective reactions to perceptual stimuli) in response to novelty, punishment, and reward (Cloninger et al., 1993). What comes to learning and working circumstances, getting into novel situations and receiving rewards and punishments are common experiences and therefore, temperament is highly relevant in these situations. Persistence, for example, has been associated with students' intrinsic academic motivation after adjustments for age and gender (Tanaka et al., 2009).

The third theory to link temperament and motivation is the theory developed by Gray (1970). This theory focuses on two basic motivational systems that shape and modify an individual's behavior and affect; (1) the behavioral activation system that is suggested to facilitate behavior and to create positive affect and (2) the behavioral inhibition system that is suggested to inhibit behavior and to create negative affect. Studies have shown that Behavioral activation system, Positive emotionality, and Extraversion load on Approach Temperament factor, whereas Behavioral inhibition system, Neuroticism, and Negative emotionality load on Avoidance Temperament factor (Elliot, 2006; Elliot and Thrash, 2002). Both approach and avoidance temperaments have been shown to be systematically linked to achievement goals (Elliot and Thrash, 2002), which in turn serve a directional function in approach-avoidance motivation and predict motivated behavior (Elliot, 2006). The hierarchical approach-avoidance motivation is proposed to come from two significant origins – temperaments and motives – which form a foundation for the achievement goals (Elliot, 2006).

The association of temperament with avoidance-approach motivation has also been examined in terms of personal goals related to avoidance/approach (Bjørnebekk and Diseth, 2010; Elliot and Thrash, 2002; Heimpel et al., 2006), achievement goals based on the hierarchical model of achievement motivation (Chen and Zhang, 2011), and in the context of the expectancy-value model (Tempelaar et al., 2011). Temperamental Activity, Flexibility, and Task orientation, as well as Neuroticism, Extroversion, and Conscientiousness, have been shown to predict mastery-approach, mastery-avoidance, performance-approach, and performance-avoidance goals (Chen and Zhang, 2011). Openness and Agreeableness, in turn, have been shown to predict the mastery-oriented and performance-oriented achievement goals, respectively (Chen and Zhang, 2011).

As the fourth theory linking temperament and motivation Derryberry and Rothbart have introduced a dynamic and interactive model of temperament, motivation, and attention. In this theory, temperament consists of three factors: surgency extraversion (i.e., dimensions of impulsivity, activity, sociability, approach, and high-intensity pleasure); negative affect (i.e., dimensions of discomfort, fear, anger/frustration, and sadness); and effortful control (i.e., dimensions of low-intensity pleasure, smiling/laughter, inhibitory control, perceptual sensitivity, and attentional focusing), the latest one being the most significant temperament dimension for attentional functions (Derryberry and Rothbart, 1997; Rothbart et al., 2007). Differences in a motivational affect, in turn, are associated with the direction of attention toward or away from a stimulus (Derryberry and Reed, 1994; Derryberry and Rothbart, 1997). The interaction between temperament, motivation, and attention manifests as individual differences in reactivity (i.e., the excitability of motor activity, affect, and endocrine and autonomic responses) and self-regulation (i.e., processes that either enhance or inhibit reactivity) (Derryberry and Rothbart, 1997).

The foregoing four theories share an agreement that individual differences in both temperament and motivation evolve partially from the reactivity of underlying neural systems, which influence individual differences in attentional, motivational, and self-regulatory functions (Chang and Burns, 2005; Deci and Ryan, 2000; Derryberry and Reed, 1994). As a result of these underlying influences, individuals with different temperaments will be exposed to and become motivated by different types of information and environmental stimulations, leading to different content within their "motivational and developmental representations" (Derryberry and Rothbart, 1997, p. 633).

Although temperament and motivation are relatively independent of cognitive ability (e.g., Guerin et al., 2003; Hintsanen et al., 2012; Oliver et al., 2007), temperament-based differences in emotional reactivity, energy level, motivation, and attention are significantly associated with the educational outcomes, in particular with the ability to persist and settle down to the task at hand, and show what one has learned (Rothbart and Jones, 1998). Mastery motivation is seen as an inborn and intrinsic disposition that refers to the positive versus negative affect/approach versus inhibition system, which is further strongly linked with an individual's tendency to experience reward and punishment in achievement and control-related situations (Barrett and Morgan, 1995; Rothbart and Jones, 1998).

Practically, mastery motivation is composed of task persistence, preference for challenge or novelty, and a tendency to control one's environment, but a tendency to show effortful control and a tendency to experience fear or frustration are also possible contributors (Rothbart and Jones, 1998). Rothbart and Jones have emphasized that teachers should receive thorough temperament-conscious teacher training in order to understand the differences of students' behavioral characteristics in emotional reactivity and attentional self-regulation, which are significant factors in educational and occupational settings (Rothbart and Jones, 1998).

Interaction between temperament and motivation on attention may be mediated via individual differences in reactions to stressful situations (Chang and Burns, 2005). Among children with performance-oriented motivation, the levels of effortful control have been suggested to regulate children's stress tolerance, and to further affect their attentional skills. However, among children with mastery-oriented motivation, the level of effortful control did not play a strong role in attentional abilities until the assignment was really stressful (Chang and Burns, 2005).

See also: Affect-Regulation Motivation; Avoidance and Approach Motivation: A Brief History; Gender and Academic Motivation; Leisure Activities Choices among Adolescents; Mastery Learning; Passion and Motivation; School Achievement: Motivational Determinants and Processes; Schooling: Impact on Cognitive and Motivational Development; Self-Determination Theory; Self-Regulated Learning: Theories, Measures, and Outcomes; Self-Regulation in Adulthood; Test Anxiety and Academic Achievement.

Bibliography

Allport, G.W., 1937. Personality: A Psychological Interpretation. Holt, New York.

Barrett, K.C., Morgan, G.A., 1995. Continuities and discontinuities in mastery motivation during infancy and toddlerhood: a conceptualization and review. In: MacTurk, R.H., Morgan, G.A. (Eds.), Mastery Motivation: Origins, Conceptualizations, and Applications, vol. 12. Ablex, Norwood, NJ, pp. 57–93.

Becker, K., El-Faddagh, M., Schmidt, M.H., Laucht, M., 2007. Is the serotonin transporter polymorphism (5-HTTLPR) associated with harm avoidance and internalising problems in childhood and adolescence? Journal of Neural Transmission 114 (3), 395–402. http://dx.doi.org/10.1007/s00702-006-0577-4.

Becker, K., Laucht, M., El-Faddagh, M., Schmidt, M.H., 2005. The dopamine D4 receptor gene exon III polymorphism is associated with novelty seeking in 15-year-old males from a high-risk community sample. Journal of Neural Transmission 112 (6), 847–858. http://dx.doi.org/10.1007/s00702-004-0223-y.

Bjørnebekk, G., Diseth, Å., 2010. Approach and avoidance temperaments and achievement goals among children. Personality and Individual Differences 49 (8), 938–943. http://dx.doi.org/10.1016/j.paid.2010.07.034.

Bouffard-Bouchard, T., Parent, S., Larivee, S., 1991. Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. International Journal of Behavioral Development 14 (2), 153–164.

Caspi, A., Shiner, R.L., 2006. Personality development. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology: Social, Emotional and Personal Development, sixth ed., vol. 3. John Wiley & Sons, Hoboken, NJ, pp. 300–365.

Chang, F., Burns, B.M., 2005. Attention in preschoolers: associations with effortful control and motivation. Child Development 76 (1), 247–263. http://dx.doi.org/10.1111/j.1467-8624.2005.00842.x.

Chen, C., Zhang, L.F., 2011. Temperament, personality and achievement goals among Chinese adolescent students. Educational Psychology 31 (3), 339–359. http://dx.doi.org/10.1080/01443410.2011.559310.

Cloninger, C.R., 2004. Feeling Good: The Science of Well-being. University Press, Oxford.

Cloninger, C.R., Svrakic, D.M., Przybeck, T.R., 1993. A psychobiological model of temperament and character. Archives of General Psychiatry 50 (12), 975–990.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Erlbaum, Hillsdale, NJ.

Comings, D.E., Flanagan, S.D., Dietz, G., Muhleman, D., Knell, E., Gysin, R., 1993. The dopamine D2 receptor (DRD2) as a major gene in obesity and height. Biochemical Medicine and Metabolic Biology 50 (2), 176–185. http://dx.doi.org/10.1006/bmmb.1993.1059.

Dan, B., Hagit, C., Lily, N., Ebstein, R.P., 2004. An association between fibromyalgia and the dopamine D4 receptor exon III repeat polymorphism and relationship to novelty seeking personality traits. Molecular Psychiatry 9 (8), 730–731.

Davis, H., Carr, M., 2002. Gender differences in mathematics strategy use: the influence of temperament. Learning and Individual Differences 13 (1), 83–95.

Deci, E.L., Ryan, R.M., 2000. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11 (4), 227–268. http://dx.doi.org/10.1207/s15327965pli1104_01.

Derryberry, D., Reed, M.A., 1994. Temperament and attention: orienting toward and away from positive and negative signals. Journal of Personality & Social Psychology 66 (6), 1128–1139.

Derryberry, D., Rothbart, M.K., 1997. Reactive and effortful processes in the organization of temperament. Development and Psychopathology 9 (4), 633–652.

Ebstein, R.P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Belmaker, R.H., 1996. Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. Nature Genetics 12 (1), 78–80.

Elliot, A.J., 2006. The hierarchical model of approach-avoidance motivation. Motivation and Emotion 30 (2), 111–116. http://dx.doi.org/10.1007/s11031-006-9028-7.

Elliot, A.J., Thrash, T.M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality & Social Psychology 82 (5), 804–818.

Eysenck, H.J., Eysenck, M.W., 1987. Personality and Individual Differences: A Natural Science Approach. Plenum Press, New York.

Gillespie, N.A., Cloninger, C.R., Heath, A.C., Martin, N.G., 2003. The genetic and environmental relationship between Cloninger's dimensions of temperament and character. Personality and Individual Differences 35 (8), 1931–1946. http://dx.doi.org/10.1016/S0191-8869(03)00042-4.

Gomez, V., Allemand, M., Grob, A., 2012. Neuroticism, extraversion, goals, and subjective well-being: exploring the relations in young, middle-aged, and older adults. Journal of Research in Personality 46 (3), 317–325. http://dx.doi.org/10.1016/j.jrp.2012.03.001.

Gray, J.A., 1970. The psychophysiological basis of introversion-extraversion. Behaviour Research and Therapy 8 (3), 249–266. http://dx.doi.org/10.1016/0005-7967(70)90069-0.

Gray, J.A., 1982. The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System. Clarendon Press, Oxford.

Guerin, D.W., Gottfried, A.W., Oliver, P.H., Thomas, C.W., 2003. Temperament: Infancy through Adolescence. The Fullerton Longitudinal Study. Kluwer Academic/Plenum Publishers, New York.

Ham, B.-J., Lee, Y.-M., Kim, M.-K., Lee, J., Ahn, D.-S., Choi, M.-J., Lee, M.-S., 2006. Personality, dopamine receptor D4 exon III polymorphisms, and academic achievement in medical students. Neuropsychobiology 53 (4), 203–209.

Hazrati-Viari, A., Rad, A.T., Torabi, S.S., 2012. The effect of personality traits on academic performance: the mediating role of academic motivation. Procedia – Social and Behavioral Sciences 32 (0), 367–371. http://dx.doi.org/10.1016/j.sbspro.2012.01.055.

Heimpel, S.A., Elliot, A.J., Wood, J.V., 2006. Basic personality dispositions, self-esteem, and personal goals: an approach-avoidance analysis. Journal of Personality 74 (5), 1293–1320. http://dx.doi.org/10.1111/j.1467-6494.2006.00410.x.

Heinonen, K., Räikkönen, K., Keltikangas-Järvinen, L., 2003. Maternal perceptions and adolescent self-esteem: a six-year longitudinal study. Adolescence 38 (152), 669–687.

Hintsanen, M., Alatupa, S., Jokela, M., Lipsanen, J., Hintsa, T., Leino, M., 2012. Associations of temperament traits and mathematics grades in adolescents are dependent on the rater but independent of motivation and cognitive ability. Learning & Individual Differences 22 (4), 490–497. http://dx.doi.org/10.1016/j.lindif.2012.03.006.

Hintsanen, M., Alatupa, S., Pullmann, H., Hirstiö-Snellman, P., Keltikangas-Järvinen, L., 2010. Associations of self-esteem and temperament traits to self- and teacher-reported social status among classmates. Scandinavian Journal of Psychology 51 (6), 488–494. http://dx.doi.org/10.1111/j.1467-9450.2010.00820.x.

Hintsanen, M., Lipsanen, J., Pulkki-Råback, L., Kivimäki, M., Hintsa, T., Keltikangas-Järvinen, L., 2009. EAS temperaments as predictors of unemployment in young adults: a 9-year follow-up of the cardiovascular risk in Young Finns Study. Journal of Research in Personality 43 (4), 618–623. http://dx.doi.org/10.1016/j.jrp.2009.03.013.

Jokela, M., Keltikangas-Järvinen, L., Kivimäki, M., et al., 2007. Serotonin receptor 2A gene and the influence of childhood maternal nurturance on adulthood depressive symptoms. Archives of General Psychiatry 64 (3), 356–360. http://dx.doi.org/10.1001/archpsyc.64.3.356.

Jussim, L., Harber, K.D., 2005. Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. Personality and Social Psychology Review 9 (2), 131–155.

Keltikangas-Järvinen, L., Pulkki-Råback, L., Elovainio, M., Raitakari, O.T., Viikari, J., Lehtimäki, T., 2009. DRD2 C32806T modifies the effect of child-rearing environment on adulthood novelty seeking. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 150B (3), 389–394. http://dx.doi.org/10.1002/ajmg.b.30830.

Keltikangas-Järvinen, L., Puttonen, S., Kivimäki, M., Elovainio, M., Rontu, R., Lehtimäki, T., 2007. Tryptophan hydroxylase 1 gene haplotypes modify the effect of a hostile childhood environment on adulthood harm avoidance. Genes, Brain and Behavior 6 (4), 305–313. http://dx.doi.org/10.1111/j.1601-183X.2006.00255.x.

Keogh, B.K., 2003. Temperament in the Classroom: Understanding Individual Differences. Paul H Brookes Publishing, Baltimore, MD.

Kluger, A., Siegfried, Z., Ebstein, R., 2002. A meta-analysis of the association between DRD4 polymorphism and novelty seeking. Molecular Psychiatry 7, 712–717.

Komarraju, M., Karau, S.J., Schmeck, R.R., 2009. Role of the big five personality traits in predicting college students' academic motivation and achievement. Learning and Individual Differences 19 (1), 47–52. http://dx.doi.org/10.1016/j.lindif.2008.07.001.

Kristal, J., 2005. The Temperament Perspective: Working with Children's Behavioral Styles. Paul H. Brookes Publishing Company, Baltimore, MD.

Light, K.J., Joyce, P.R., Luty, S.E., Mulder, R.T., Carter, J.D., Frampton, C.M., Kennedy, M.A., 2007. An association study of DRD2 and COMT polymorphisms with novelty seeking and harm avoidance scores, in two independent samples of depressed patients. Behavioral and Brain Functions 3 (3).

Lüdtke, O., Trautwein, U., Husemann, N., 2009. Goal and personality trait development in a transitional period: assessing change and stability in personality development. Personality and Social Psychology Bulletin 35 (4), 428–441. http://dx.doi.org/10.1177/0146167208329215.

Martin, R.P., Bridger, R.C., 1999. The Temperament Assessment Battery for Children-revised: A Tool for the Assessment of Temperamental Traits and Types of Young Children. Unpublished manual.

McAdams, D.P., 2009. The Person: An Introduction to the Science of Personality Psychology, fifth ed. Northwestern University: A John Wiley & Sons, Inc.

McCrae, R.R., Costa, P.T., 1987. Validation of the five-factor model of personality across instruments and observers. Journal of Personality & Social Psychology 52 (1), 81–90.

McCrae, R.R., Costa, P.T., 2003. Personality in Adulthood: A Five-factor Theory Perspective, second ed. Guilford Press, New York.

Mullola, S., Hintsanen, M., Jokela, M., Lipsanen, J., Alatupa, S., Ravaja, N., Keltikangas-Järvinen, L., 2014. Associations between teacher-rated versus self-rated student temperament and school achievement. Scandinavian Journal of Educational Research 58 (2), 147–172. http://dx.doi.org/10.1080/00313831.2012.725094.

Mullola, S., Jokela, M., Ravaja, N., Lipsanen, J., Hintsanen, M., Alatupa, S., Keltikangas-Järvinen, L., 2011. Associations of student temperament and educational competence with academic achievement: the role of teacher age and teacher and student gender. Teaching and Teacher Education 27 (5), 942–951. http://dx.doi.org/10.1016/j.tate.2011.03.005.

Mullola, S., Ravaja, N., Lipsanen, J., Alatupa, S., Hintsanen, M., Jokela, M., Keltikangas-Järvinen, L., 2012. Gender differences in teachers' perceptions of students' temperament, educational competence, and teachability. British Journal of Educational Psychology 82 (2), 185–206. http://dx.doi.org/10.1111/j.2044-8279.2010.02017.x.

Mullola, S., Ravaja, N., Lipsanen, J., Hirstiö-Snellman, P., Alatupa, S., Keltikangas-Järvinen, L., 2010. Teacher-perceived temperament and educational competence as predictors of school grades. Learning and Individual Differences 20 (3), 209–214. http://dx.doi.org/10.1016/j.lindif.2010.01.008.

Noble, E.P., Ozkaragoz, T.Z., Ritchie, T.L., Zhang, X., Belin, T.R., Sparkes, R.S., 1998. D2 and D4 dopamine receptor polymorphisms and personality. American Journal of Medical Genetics 81 (3), 257–267.

Oliver, P.H., Guerin, D.W., Gottfried, A.W., 2007. Temperamental task orientation: relation to high school and college educational accomplishments. Learning and Individual Differences 17 (3), 220–230.

Pesonen, A.-K., Räikkönen, K., Strandberg, T., Keltikangas-Järvinen, L., Järvenpää, A.-L., 2004. Insecure adult attachment style and depressive symptoms: implications for parental perceptions of infant temperament. Infant Mental Health Journal 25 (2), 99–116. http://dx.doi.org/10.1002/imhj.10092.

Ravaja, N., Katainen, S., Keltikangas-Järvinen, L., 2001. Perceived difficult temperament, hostile maternal child-rearing attitudes and insulin resistance syndrome precursors among children: a 3-year follow-up study. Psychotherapy and Psychosomatics 70 (2), 66–77.

Rothbart, M.K., 2007. Temperament, development, and personality. Current Directions in Psychological Science 16 (4), 207–212.

Rothbart, M.K., Jones, L.B., 1998. Temperament, self-regulation, and education. School Psychology Review 27 (4), 479.

Rothbart, M.K., Sheese, B.E., Posner, M.I., 2007. Executive attention and effortful control: linking temperament, brain networks, and genes. Child Development Perspectives 1 (1), 2–7. http://dx.doi.org/10.1111/j.1750-8606.2007.00002.x.

Rudasill, K.M., Rimm-Kaufman, S.E., 2009. Teacher–child relationship quality: the roles of child temperament and teacher–child interactions. Early Childhood Research Quarterly 24 (2), 107–120.

Salmela-Aro, K., Nurmi, J.-E., 1997. Goal contents, well-being, and life context during transition to university: a longitudinal study. International Journal of Behavioral Development 20 (3), 471–491. http://dx.doi.org/10.1080/016502597385234.

Salmela-Aro, K., Nurmi, J.-E., Saisto, T., Halmesmaki, E., 2001. Goal reconstruction and depressive symptoms during the transition to motherhood: evidence from two cross-lagged longitudinal studies. Journal of Personality & Social Psychology 81 (6), 1144–1159.

Salmela-Aro, K., Read, S., Nurmi, J.-E., Koskenvuo, M., Kaprio, J., Rantanen, T., 2009. Personal goals of older female twins: genetic and environmental effects. European Psychologist 14 (2), 160–167.

Salo, J., Pulkki-Råback, L., Hintsanen, M., Lehtimäki, T., Keltikangas-Järvinen, L., 2010. The interaction between serotonin receptor 2A and catechol-*O*-methyltransferase gene polymorphisms is associated with the novelty-seeking subscale impulsiveness. Psychiatric Genetics 20 (6), 273–281.

Schunk, D.H., Hanson, A.R., 1985. Peer models: influence on children's self-efficacy and achievement. Journal of Educational Psychology 77 (3), 313–322.

Sen, S., Burmeister, M., Ghosh, D., 2004. Meta-analysis of the association between a serotonin transporter promoter polymorphism (5-HTTLPR) and anxiety-related personality traits. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 127B (1), 85–89. http://dx.doi.org/10.1002/ajmg.b.20158.

Stipek, D., 2002. Communicating expectations. In: Stipek, D. (Ed.), Motivation to Learn: Integrating Theory and Practice, fourth ed. Allyn and Bacon, Boston, MA, pp. 210–229.

Strelau, J., 1998. Temperament: A Psychological Perspective. Plenum Publishing Corporation, New York.

Szekely, A., Ronai, Z., Nemoda, Z., Kolmann, G., Gervai, J., Sasvari-Szekely, M., 2004. Human personality dimensions of persistence and harm avoidance associated with DRD4 and 5-HTTLPR polymorphisms. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 126B (1), 106–110. http://dx.doi.org/10.1002/ajmg.b.20134.

Tanaka, M., Mizuno, K., Fukuda, S., Tajima, S., Watanabe, Y., 2009. Personality traits associated with intrinsic academic motivation in medical students. Medical Education 43 (4), 384–387. http://dx.doi.org/10.1111/j.1365-2923.2008.03279.x.

Tempelaar, D., Schim van der Loeff, S., Gijselaers, W., Nijhuis, J.H., 2011. On subject variations in achievement motivations: a study in business subjects. Research in Higher Education 52 (4), 395–419. http://dx.doi.org/10.1007/s11162-010-9199-7.

Thomas, A., Chess, S., 1977. Temperament and Development. Brunner/Mazel, New York.

Tuominen, L., Salo, J., Hirvonen, J., Någren, K., Laine, P., Melartin, T., Hietala, J., 2012. Temperament trait Harm Avoidance associates with μ-opioid receptor availability in frontal cortex: A PET study using [^{11}C]carfentanil. NeuroImage 61 (3), 670–676. http://dx.doi.org/10.1016/j.neuroimage.2012.03.063.

Van Gestel, S., Forsgren, T., Claes, S., Del-Favero, J., Van Duijn, C., Sluijs, S., Van Broeckhoven, C., 2002. Epistatic effect of genes from the dopamine and serotonin systems on the temperament traits of novelty seeking and harm avoidance. Molecular Psychiatry 7, 448–450.

Zimmerman, B.J., Kitsantas, A., 1999. Acquiring writing revision skill: shifting from process to outcome self-regulatory goals. Journal of Educational Psychology 91 (2), 241–250.

Zimmerman, B.J., Schunk, D.H., 2009. Motivation: an essential dimension of self-regulated learning. In: Schunk, D.H., Zimmerman, B.J. (Eds.), Motivation and Self-Regulated Learning. Routledge, New York, pp. 1–30.

Temperament Development, Theories of

Xiaoxue Fu and Koraly Pérez-Edgar, The Pennsylvania State University, University Park, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Temperament is defined as biologically rooted, early appearing dispositions that shape long-term patterns of socioemotional development. Much of the research has focused on broad dimensions of emotional reactivity, including positive and negative emotionality, and regulation, marked by effortful control. The article outlines five main research traditions that have approached temperament as both a continuous and a categorical construct. For each model, descriptions are provided for the dimensions/category proposed and their defining features, the underlying constitutional bases, and perspectives on temperamental continuity and discontinuity. Next, theoretical and measurement issues are discussed. Finally, we demonstrate how temperament may predict developmental psychopathology through temperament-by-temperament and temperament-by-environment moderations of early risk.

Introduction

Temperaments reflect biologically based individual differences in how children interact with, and adapt to, their environments across development, leading to marked interpersonal variations in both normative behavior and psychopathological profiles. The landmark round table discussion on temperament at the 1985 meeting of the Society for Research in Child Development (Goldsmith et al., 1987) presented four core models of temperament. The models each focused on early emerging, constitutionally based, multidimensional constructs, emphasizing the stability of temperament as a motivator of behavior, while recognizing that the expression of temperament may change across time (Rothbart and Derryberry, 1981) and contexts (Goldsmith and Campos, 1982). No consensus was reached regarding the conceptual 'boundaries' of temperament. For example, Thomas and Chess (1977) confined temperament to observable behavioral styles. Goldsmith and Campos (1982) defined temperament during infancy as the presence of primary emotions. Finally, Rothbart and Derryberry (1981) included cognitive processes as core components of temperament.

Despite these discrepancies, each model pursues a dimensional approach to temperament, quantifying temperamental variations along a continuous spectrum. In contract, Kagan (Kagan et al., 1984) takes a typological approach to temperament. He and colleagues identified a temperament category, behavioral inhibition (BI), characterized by hypervigilant and withdrawal behavior in response to unfamiliar people and situations in toddlerhood and childhood. Together, the five models have provided multidisciplinary (and complementary) assessment tools and theoretical frameworks for studying temperament and its development. These include parent- and self-report questionnaires, laboratory observations of behavior, psychophysiological measures, and behavioral genetic approaches.

Contemporary temperament research focuses on studying how intrinsic (neural and cognitive processes) and extrinsic (environmental) processes interact to moderate the continuity of temperament and influence the link between early temperament and subsequent socioemotional adjustment. This article will discuss (1) the five major approaches to temperament (Table 1); (2) current theoretical and methodological issues in research on temperament development; and (3) the mechanisms by which documented relations between temperament traits and maladjustment are moderated by other temperament traits (temperament × temperament moderation) and environmental factors (temperament × environment moderation). We take the perspective that early temperament does not determine developmental outcomes. Rather, temperament traits interact with internal and external factors to influence individuals' risks and resilience to events and contexts that shape developmental trajectories.

Theoretical Approaches to Temperament

The Thomas and Chess Approach

The New York Longitudinal Study stands as a landmark in temperament research. Thomas and Chess (1977) focused on behavioral styles characterized across nine dimensions: activity level, regularity, approach withdrawal, adaptability, threshold of responsiveness, intensity of reaction, quality of mood, attention span/persistence, and distractibility. Children are categorized into 'difficult,' 'easy,' and 'slow to warm' types based on their scores on each dimension. Moreover, Chess and Thomas introduced the concept of 'goodness of fit,' suggesting that optimal developmental outcomes are more likely to emerge when the child's temperament is closely matched to the demands, expectations, and opportunities of the environment (a good fit). Conversely, a mismatch between temperament and environmental characteristics is more likely to result in maladaptive outcomes (Thomas and Chess, 1977).

The Buss and Plomin Approach

Buss and Plomin (1975) proposed a behavior-genetics-oriented model, proposing that temperament traits have an early onset and are inherited, evolutionary adaptive, present in nonhuman animals, relatively stable during development and predictive of later behaviors. They identified three core dimensions: emotionality (E), activity (A), and sociability (S), each captured by the EAS Temperament Survey (Buss and Plomin, 1984). The approach utilized twin studies to

Table 1 Summaries of the five theoretical approaches to temperament

	Defining features	*Dimension(s)/Type(s)*	*Underlying biological mechanisms*	*Major instrument(s)*	*Key reading(s)*
Thomas and Chess	● Represent stylistic components of behavior	Dimensions: ● Activity level ● Regularity ● Approach withdrawal ● Adaptability ● Threshold of responsiveness ● Intensity of reaction ● Quality of mood ● Attention span/persistence ● Distractibility Types: ● Difficult ● Easy ● Slow to warm up	Not discussed	1–3 years old: The Toddler Temperament Scale (TTS; Fullard et al., 1984); 3–7 years old: The Behavioral Style Questionnaire (BSQ, McDevitt and Carey, 1978); 8–12 years old: The Middle Childhood Temperament Questionnaire (MCTQ, Hegvik et al., 1982).	Thomas and Chess (1977)
Buss and Plomin	● Inherited traits ● Evolutionary adaptive ● Present in our phylogenetically related species ● Early emerging and relatively stable in childhood ● Preserved in adulthood	Dimensions: ● Emotionality ● Activity ● Sociability ● Impulsivity (dropped)	Temperament dimensions have genetic bases.	EAS Temperament Survey (EAS, Buss and Plomin, 1984)	Buss and Plomin (1975) Buss and Plomin, (1984)
Goldsmith	● An emotional phenomenon; ● Related to individual differences; ● Behavioral tendencies rather than actual observable emotional behavior; ● Does not contain cognitive and perceptual components; ● Has cross-situational generality and short-term stability rather than being transitory.	Assessed in Lab-TAB: ● Activity level ● Positive affect ● Fear ● Anger ● Sadness ● Shyness ● Approach ● Persistence ● Inhibitory control	Temperament dimensions are genetically influenced	Toddler Behavior Assessment Questionnaire (TBAQ, Goldsmith, 1996); Laboratory Temperament Assessment Battery (Lab-TAB, Goldsmith and Rothbart, 1993)	Goldsmith and Campos (1982, 1986);
Rothbart	● Biologically based ● Contains reactive and regulatory aspects	Dimensions: ● Negative affect ● Surgency ● Effortful control	The attention network: Encompasses the anterior cingulate cortex and the lateral prefrontal regions	3–12 months: The Infant Behavior Questionnaire (IBQ, Rothbart, 1981); 18–36 months: The Early Childhood Behavior Questionnaire (ECBQ, Putnam et al., 2006); 3–7 years: The Children's Behavior Questionnaire (CBQ, Rothbart et al., 2001); 7–10 years: The Temperament in Middle Children Questionnaire (TMCQ, Simonds and Rothbart, 2004); 9–15 years: The Early Adolescent Temperament Questionnaire (EATQ, Capaldi and Rothbart, 1992)	Rothbart and Derryberry (1981) Rothbart and Bates (2007)
Kagan	● Biologically based and early emerging ● Defined a specific temperament category characterizes children who show elevated physiological reactivity, hypervigilance and withdrawal upon encountering novel stimuli (behavioral inhibition)	Type: Behavioral inhibition	Hyperresponsive amygdala	Not developed	Kagan (1994); Kagan and Fox (2007)

quantify genetic contributions to individual variance in temperament traits (Buss and Plomin, 1984). Monozygotic (MZ) twin pairs showed significantly larger correlations in all three EAS traits versus dizygotic (DZ) twins. Given that MZ twins have twofold more genetic similarity than DZ twins, the greater phenotypic similarity in MZ twins indicates genetic influences on trait variance (Saudino, 2009).

The Goldsmith Approach

Goldsmith and Campos (1982) defined temperament as individual differences in the propensity to express and experience primary emotions (e.g., joy, anger, and fear). Individual differences in temperament are reflected in variations in intensive and temporal parameters of behavior, including facial expressions, gestures, vocal and motor behaviors. The definition is restricted to infant temperament, in order to obtain a relatively 'pure' conceptualization of temperament expressions unmediated by socialization influences and cognitive processes. However, recent variants of the model include emotion regulation as a defining feature, as emotion expression and regulation might not be separable, even in infancy (Campos et al., 2004).

Goldsmith and colleagues developed the Toddler Behavior Assessment Questionnaire (Goldsmith, 1996), and the age-adapted Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith and Rothbart, 1993; **Table 1**). The Lab-TAB provides a series of standardized behavioral tasks and coding schemes that assess temperament dimensions in emotion-eliciting episodes in the laboratory or at home. It stresses that the context of behavior needs to be considered in measuring temperament, since the eliciting situations affect the expression of temperament dimensions (Goldsmith and Gagne, 2012).

Building on their initial conceptualization, Goldsmith and colleagues focused on the psychobiological processes underlying emotion reactivity and regulation. For example, Buss et al. (2004) combined both Lab-TAB and physiological measurements and found that heightened fearfulness in a mildly threatening situation was associated with heightened baseline sympathetic reactivity a week later. This implies a possible association between fearful temperament and reduced physiological regulation.

The Rothbart Approach

The Rothbart model (Rothbart and Derryberry, 1981) provides a more inclusive conceptualization, defining temperament as biologically rooted individual differences in reactivity and self-regulation in emotional, activational, and attentional processes. Reactivity refers to levels of biological arousal triggered by changes in internal and external stimulations, captured as dimensions of negative affect and surgency. Self-regulation refers to the processes that modulate reactivity (Rothbart and Derryberry, 1981) and is reflected in the temperament dimension, effortful control (EC). Taking a dimensional approach, each of the three higher-order dimensions is comprised of lower-order temperament traits (Rothbart and Bates, 2007).

Questionnaires from this approach assume heterotypic continuity in temperament development (Caspi, 1998). That is, age-appropriate questions or eliciting tasks take different forms of expression but are thought to evoke the same underlying bases throughout development (Fox and Henderson, 1999). As presented in Rothbart and colleagues' age-specific questionnaires, the broad dimensions remain stable, but the specific traits indexing each dimension changes across age groups. For example, EC during early infancy is measured as a form of involuntary attention orienting. As external controls become internalized, indices of EC come to also encompass attention shifting and focusing (Rueda, 2012).

The Rothbart model is unique in prominently placing self-regulation as a core component of temperament. EC has been studied with multiple-levels of analyses, including the aforementioned parent- and self-report questionnaires, as well as computerized cognitive tasks and laboratory observation tasks. Findings from computerized tasks demonstrate that EC capacities, including error detection and conflict resolution, first emerge at the end of first year of life and continue to improve during late childhood and into adolescence (Rothbart et al., 2007; Rueda, 2012). Neuroimaging research adopting these marker tasks suggests that the development of EC is subserved by the executive attention network, a neural system centered on the prefrontal cortex (Posner et al., 2014).

As EC develops during toddlerhood, it exerts 'brakes' on unregulated negative emotionality and motor activities (Rothbart and Derryberry, 2002), and enhances children's abilities to adapt to environmental demands. Longitudinal data indicate that higher EC predicts less negative emotional arousal. Conversely, elevated negative emotionality predicts lower EC efficiency (Eisenberg et al., 2010). At the neural level, efficiency in regulating reactivity is subserved by the coupling between regions associated with regulation (e.g., the prefrontal regions) and reactivity (e.g., the amygdala and the striatum; Dennis, 2010). The continued dynamic and reciprocal interactions between temperamental reactivity and regulation contribute to the observed instability of temperament expression over time (Rothbart and Bates, 2007).

The Kagan Approach

Distinct from dimensional approaches to temperament, Kagan and colleagues defined BI as a temperamental *category* characterized by discrete biological dispositions: high psychophysiological reactivity coupled with hypervigilance and behavioral withdrawal upon encountering novel people, objects, and events (Kagan et al., 1984). Kagan's approach emphasizes the use of behavioral observations and biological measures instead of parent-report questionnaires (Kagan, 2003). Based on laboratory observations, 4-month-old infants who displayed elevated motor reactivity and distress toward novel visual and auditory stimuli were categorized as 'high reactive' (Kagan and Snidman, 1991) or 'high negative' (Fox et al., 2001). BI in young children (14–48 months) is examined by coding their behavior toward unfamiliar people and objects. Assessment in older children (4 years to school age) focuses on their interactions with unfamiliar peers. Hence, Kagan's approach also focuses on heterotypic continuity of temperament development over time (Fox et al., 2001). Kagan and colleagues found that although not all high-reactive infants become behaviorally inhibited in childhood, the percentage of high-reactive infants

who are later categorized with BI was greater than those who were identified as uninhibited (Kagan, 2003). Fox et al. (2001) showed that at 14 months, infants who were classified as 'high negative' displayed higher levels of BI than those who were identified as 'low reactive' or 'high positive.'

Kagan proposed that the neurobiological foundation of BI is rooted in a highly excitable amygdala, which predisposes children to become hypervigilant toward unfamiliar stimuli (Kagan, 2012), and contributes to the moderate stability of BI (Fox et al., 2005). Adults who were identified as 'high reactive' at 4 months showed greater amygdala responses when presented with neutral unfamiliar faces compared to those who had been 'low reactive' (Schwartz et al., 2003, 2011). Moreover, Pérez-Edgar et al. (2007) found that, compared to non-inhibited adolescents, adolescents with sustained BI in childhood showed elevated amygdala response while rating how afraid they were of emotional and neutral faces. The amygdala has extensive connections to cortical areas that contribute to behavioral avoidance and deficient safety learning observed in behaviorally inhibited individuals (Schwartz et al., 2011).

Stable BI across childhood is a risk factor for anxiety disorders, especially social anxiety disorder (SAD, Pérez-Edgar and Fox, 2005a). BI and anxiety disorders have a range of overlapping behavioral, cognitive, and neurological features. Retrospective evidence suggests that adolescents and adults with anxiety symptoms also reported higher levels of social withdrawal in childhood (Degnan et al., 2010). Additionally, Clauss and Blackford's (2012) meta-analysis showed that the risk for SAD increases sevenfold for behaviorally inhibited children. However, only 43% of behaviorally inhibited children developed SAD. The central tenet of the Kagan's approach argues that early BI does not necessarily predict stability of temperament and later anxiety disorders. Rather, it constrains possible socioemotional development outcomes (Kagan and Fox, 2007).

Current Methodologies and Issues in Temperament Research

Contemporary temperament research has shifted its focus from debating the defining features and structure of temperament (Goldsmith et al., 1987) to studying developmental trajectories from early temperament to subsequent socioemotional adjustment outcomes. Summarizing the shared viewpoints and new perspectives on temperament since the round table discussion, Shiner et al. (2012) define temperament as early appearing traits in the domains of motor activity, emotion, attention, and self-regulation. They stress that temperament is the product of the continuous influences of genetic, biological, and environmental processes throughout development.

First, traits may have different developmental courses, and those that come online later during infancy and childhood (e.g., regulatory traits) may affect the expression of existing traits' (e.g., temperamental reactivity; Rothbart and Derryberry, 1981). Second, assessments of temperament, including the questionnaires developed under the Rothbart approach (Rueda, 2012) and laboratory observation tasks (Lab-TAB, Goldsmith and Rothbart, 1993; BI tasks, Fox et al., 2001), reflect a priori conceptualization of heterotypic continuity in temperament development (Caspi, 1998). Third, although temperament traits are biologically based, behavioral genetic research has found only moderate genetic influences. Rather, both genetic and environmental factors, especially nonshared environmental influences, contribute to the stability and change of temperament over time (Saudino and Wang, 2012).

There is an on-going debate regarding whether specific temperament traits are best captured as variations along a continuum or qualitative differences across distinct categories. Support for the typological approach has come from sophisticated modeling analyses (Loken, 2004), which replicated Kagan and Snidman (1991)'s categorization of high-reactive infants. Moreover, BI profile is associated with distinct profiles of neural reactivity evident in limbic (Pérez-Edgar et al., 2007; Schwartz et al., 2003, 2011) and striatal (Helfinstein et al., 2012) responses to salient stimuli. However, this association was not found when BI was treated as a continuous variable (Pérez-Edgar et al., 2013). Yet, longitudinal studies using discrete BI profiles (e.g., Chronis-Tuscano et al., 2009) and those adopting continuous measures of BI (e.g., Muris et al., 2001) have provided converging results, both suggesting that BI predicts greater risks for SADs. Hence, both categorical and dimensional conceptualizations could potentially provide useful frameworks to study the link between early temperament and later psychopathology (Clauss and Blackford, 2012).

Cultural Influences on Temperament

The influences of cultural values on temperament development are nonnegligible. Culture creates a bioecological environment (Bronfenbrenner and Morris, 2006) that encompasses children's physical and social environment. Parents' cultural belief system influences how they perceive, evaluate, and respond to their children's temperament, which, in turn, shape children's temperament development and adjustment (Super and Harkness, 2002). In studying the impacts of immigration and involvement in the host culture (i.e., acculturation) on parental perception of infant temperament, Gartstein et al. (2009) found that among Russian families who immigrated to the United States, greater adherence to Russian culture was correlated with higher ratings of child positive emotionality. In contrast, for immigrants to Israel, greater parental involvement in Russian culture was linked to lower child positive emotionality. Moreover, Russian parents with greater acculturation to Israeli society rated their children higher on orienting and focusing of attention.

Cultural norms may shift parental perception of what traits are adaptive, which, in turn, affect parental practice, how the child's temperament is manifested, and the parent's perception of their child's temperament (Gartstein et al., 2009). For example, BI is associated with greater maternal warmth and acceptance in Chinese children, whereas Canadian children with BI experience more maternal rejection (Chen et al., 1998). Furthermore, early BI predicted subsequent social and academic competence among Chinese children (Chen et al., 1999, 2009a). Greater social acceptance of BI in China, in contrast to findings from the West, may facilitate adaptive socioemotional development. However, Chen et al. (2009b)

also found increased maladjustment in inhibited Chinese children from urban areas, likely due to rapid Westernization. Taken together, the cross-cultural studies support the presence of cultural difference in temperamental reactivity and regulation and associated developmental outcomes.

Issues Relating to Measurement Approaches

Temperament assessments target naturally occurring behaviors reported by caregivers, behaviors elicited by laboratory controlled tasks, and neurophysiological correlates of these behaviors (Goldsmith and Gagne, 2012; Rothbart and Bates, 2007). Across approaches, temperament research stresses the adoption of a multimethod perspective (Kagan and Fox, 2007). Multiple measurement approaches complement each other to overcome limitations of a specific method, and together they provide insights into unique aspects of temperament. Caregiver reports have been criticized for being biased, unreliable, invalid, and limited in revealing the underlying mechanisms of overt traits (Kagan and Fox, 2007). However, Rothbart and Bates (2007) argue that parent reports provide unique information on children's general behavioral tendencies across different situations.

Laboratory observations allow for relatively objective assessments of microlevel information on latency, duration, and frequency of facial, vocal, and gestural responses to specific elicitors that are inaccessible to verbal reports and neurological measures. However, behavioral coding may be subject to experimenters' biases, observations may lack ecological validity, and laboratory triggers may elicit emotional states rather than measuring temperament traits (Goldsmith and Gagne, 2012). Despite these limitations, laboratory observations, particularly the Lab-TAB battery (Goldsmith and Rothbart, 1993), allow for flexibility in behavioral coding and data reduction compared to questionnaire assessments (Goldsmith and Gagne, 2012). This advantage is reflected in research on dysregulated fear (DF). Buss (2011) adopted a multilevel modeling approach to capture changes in fear behaviors across episodes low, moderate, and high in contextual threat. The analyses identified a DF profile in two-year-olds, characterized by high levels of fear in low threat and benign contexts (e.g., playing with puppets and a clown). This is different from the analytic method used to identify BI, where behavioral parameters are aggregated across episodes containing different levels of threat. Buss et al. (2013) found that children characterized with DF were four times more likely to show social anxiety symptoms at the transition to kindergarten than other fearful children. Children characterized with DF might be a more homogeneous group within BI that is associated with a developmental trajectory to social anxiety.

Behavioral genetics, as a theoretical and empirical approach, has made prominent contributions to evaluating the relative contributions of genetic, shared, and nonshared environmental effects on the continuity and instability of temperament (Saudino, 2009). It also provides a tool to examine the cross-method and cross-situation validity of temperament measures. For example, Saudino and Zapfe (2008) found different genetic influences for activity levels assessed at home versus the laboratory. These findings suggest that different instruments and measurement contexts may tap into distinct etiology of observed behaviors. Therefore, caution must be taken when making generalizations across data gathered from different measurement tools under different contexts.

Temperament and Developmental Psychopathology

A temperament trait may act as a risk or resilience factor that predisposes individuals to psychopathology or protects them from disorders, in part by influencing the child's experiences with the external environment and their exposure to additional risk or protective factors (Lengua and Wachs, 2012). The developmental psychopathology perspective stresses that temperament is an etiological factor that should be studied in interactive systems, in order to reflect the complex, transactional nature of developmental processes, in which adjustment outcomes are the result of successive adaptations that build upon one another (Sroufe, 2009). This includes both temperament-by-temperament interactions and temperament-by-environment interactions in predicting outcome (Zentner and Bates, 2008).

Temperament × Temperament Moderation

The framework of temperament-by-temperament moderation is rooted in the Rothbart approach, which highlights the interplay between reactive and self-regulatory components of temperament in shaping developmental trajectories. For example, components of EC moderate the link between BI (as negative reactivity) and anxiety. BI is associated with increased attention toward potentially threatening stimuli (Pérez-Edgar and Fox, 2005b), which is likely to elicit a cascade effect on subsequent information processing, leading individuals to encode and interpret relatively neutral social information as threatening, which in turn produces maladaptive behaviors such as social withdrawal. These processes influence each other through a feedback loop, contributing to the formation of a response repertoire that perpetuates and strengthens anxious behaviors over time. Indeed, longitudinal studies indicate that childhood BI strongly predicts subsequent social withdrawal only in adolescents and children displaying attention biases to threat (Pérez-Edgar et al., 2010a,b, 2011), suggesting that attention bias acts as a 'tether' to predispose children with BI to a developmental trajectory marked by elevated anxiety symptoms (Pérez-Edgar et al., 2014). In contrast, executive attention control, as an EC function, facilitates voluntary attention shifts from threat and a subsequent focus on more positive cues, which may, in turn, allow for better self-regulation of emotion and behavior (Lonigan and Vasey, 2009), decreasing the risk for anxiety in high-BI children (White et al., 2011).

Inhibitory control and response monitoring are two additional EC components that may impact the expression of temperamental reactivity. While better inhibitory control and response monitoring is associated with reduced externalizing and internalizing symptoms for most children (Eisenberg et al., 2010), high levels of these EC abilities in behaviorally

inhibited children increase their risks for the developing anxiety problems (Degnan and Fox, 2007; Lahat et al., 2014; McDermott et al., 2009; White et al., 2011). Presumably, when involuntary control functions, such as heightened reactivity, are combined with high voluntary control systems (i.e., EC functions), they may increase behavioral rigidity and inability to adjust to environmental demands (Derryberry and Rothbart, 1997).

Temperament × Environment Moderation

The conceptualization of temperament–environment interactions are in line with the mechanisms associated with gene-by-environment interplay. First, temperament traits influence how children perceive and react to positive and negative reinforcements from the environment (Shiner and Caspi, 2012). For example, threat-related attention biases may predispose behaviorally inhibited children to perceive negative aspects of the environment as more salient (Pérez-Edgar et al., 2014; Todd et al., 2012). This facilitates the learning of negative associations linked to their behavior (Shiner and Caspi, 2012), which may, in turn, lead to the emergence and maintenance of anxiety (Pérez-Edgar et al., 2014).

The second mechanism is environmental elicitation, suggesting that individual differences in children's temperament elicit different parenting behaviors. Social withdrawal in behaviorally inhibited children may evoke parental overcontrol (Rubin et al., 2009), which may limit children's opportunities to attain social competency. Indeed, parental overcontrol moderates the link between early BI and subsequent anxiety symptoms (Lewis-Morrarty et al., 2012).

In environmental selection, children actively select specific niches in their environments that are compatible with their temperament, which may, in turn, influence their exposure to risks (Shiner and Caspi, 2012). Rubin et al. (2006) found that behaviorally inhibited children are more likely to have best friends who are also more socially withdrawn and victimized compared to the best friends of noninhibited children. Such dyads may perpetuate social withdrawal, leading to continued rejection from larger peer groups and vice versa (Degnan et al., 2010).

Finally, environmental construal refers to the processes by which temperament shapes individuals' environmental exposure through influencing social cognition (Shiner and Caspi, 2012). From repeated experience of social rejections, children may develop an attributional schema characterized by self-blame for negative experiences (Rubin et al., 2008), as well as avoidant coping strategies (Wichmann et al., 2004). These attributional and coping styles may form a self-reinforcing cycle that leads to continued withdrawal and internalizing symptoms (Rubin et al., 2009).

A number of theoretical models have been proposed to explain how certain temperament traits make individuals more susceptible to aversive or supportive social environments (Lengua and Wachs, 2012; Reiss et al., 2013). The goodness-of-fit model originating from Thomas and Chess' (1977) approach emphasizes that children whose temperament matches the characteristics of their environment will have better adjustment outcomes than those who have a poorer fit. The traditional diathesis–stress model suggests that given the same amount of environmental adversity, temperamentally at-risk children will develop more adjustment problems compared to children without these risks. Negative emotionality, including anger/frustration and fearful temperament, is considered to be a direct risk factor for both externalizing and internalizing symptoms (Lengua, 2002) and may be shaped by parenting. For example, permissive parenting styles predict more internalizing symptoms in behaviorally inhibited children versus uninhibited children (Williams et al., 2009), while a rejecting parenting style is linked to increased externalizing symptoms in high-frustration but not low-frustration children (Lengua, 2008).

Extending from the diathesis–stress model, the differential susceptibility model (Belsky and Pluess, 2009) argues that temperamentally at-risk individuals may not only be more affected by negative experiences, but could also benefit the most from supportive environments, relatively to children with less temperamentally based sensitivity to the environment. Negative emotionality is likely to be a susceptibility marker (Belsky and Pluess, 2009). Compared to children with an 'easy' temperament, children with high negative emotionality display more behavioral problems during the transition to schools if they experienced low-quality day care, but they show fewer problems if they received high-quality day care (Pluess and Belsky, 2009). Moreover, temperamental irritability is associated with lower attachment security. At the same time, the beneficial effects of an intervention on attachment are significantly greater for highly irritable versus moderately irritable infants (Cassidy et al., 2011). The model underscores the importance of studying phenotypic and genetic markers of sensitivity to environmental influences, as well as both risk and resilient contextual factors for socioemotional development.

General Conclusions

Theoretical approaches to temperament have not reached a full consensus regarding the inclusive criteria and structure of temperament. Working definitions of temperament incorporate the domains of affectivity, self-regulatory processes, and motor activity in their conceptualization of the construct (Shiner et al., 2012). There is solid agreement that temperament is a multilevel construct, which is observable via behavioral and neurobiological measurements, and is continually influenced by genetic, biological, cognitive, and environmental factors. Developmental psychopathology and behavioral genetic perspectives provide theoretical and methodological frameworks to study temperament as a biologically based risk or resilience factor that influences individuals' trajectory to certain socioemotional adjustment outcomes through complex interactive processes involving individuals' own cognition, affect, and their external developmental contexts. Different temperament traits impose inherent vulnerability (e.g., high negative emotionality) or resilience (high EC) mechanisms that influence outcomes by moderating individuals' exposure to negative or enriching environments, constraining what environmental

characteristics are beneficial or harmful to specific temperament traits. Temperament traits also affect individuals' level of susceptibility to both aversive and favorable environmental influences. These dynamic interactive processes support the fundamental notion that temperament does not determine developmental outcomes. Rather, it sets the complex foundation on which multiple actors come together to shape developmental outcomes (Kagan, 2003).

Acknowledgment

Support for manuscript preparation was provided by grants from the National Institutes of Health (MH# 094633) and The Pennsylvania State University Social Science Research Institute (Level II Grant) to Koraly Pérez-Edgar.

See also: Neonatal Temperament: From Phenomenology to Mechanism; Temperament and Human Development; Temperament and Motivation; Temperament.

Bibliography

Belsky, J., Pluess, M., 2009. Beyond diathesis stress: differential susceptibility to environmental influences. Psychological Bulletin 135 (6), 885–908.

Bronfenbrenner, U., Morris, P.A., 2006. The bioecological model of human development. In: Lerner, R.M. (Ed.), Handbook of Child Development: Vol. 1. Theoretical Models of Human Development, sixth ed. Wiley, Hoboken, NJ, pp. 793–828.

Buss, K.A., 2011. Which fearful toddlers should we worry about? Context, fear regulation, and anxiety risk. Developmental Psychology 47 (3), 804–819.

Buss, K.A., Davidson, R.J., Kalin, N.H., Goldsmith, H.H., 2004. Context-specific freezing and associated physiological reactivity as a dysregulated fear response. Developmental Psychology 40 (4), 583–594.

Buss, K.A., Davis, E.L., Kiel, E.J., Brooker, R.J., Beekman, C., Early, M.C., 2013. Dysregulated fear predicts social wariness and social anxiety symptoms during kindergarten. Journal of Clinical Child & Adolescent Psychology, 603–616.

Buss, A.H., Plomin, R., 1975. A Temperament Theory of Personality Development. Wiley, New York.

Buss, A.H., Plomin, R., 1984. Temperament: Early Developing Personality Traits. Erlbaum, Hillsdale, NJ.

Campos, J.J., Frankel, C.B., Camras, L., 2004. On the nature of emotion regulation. Child Development 75 (2), 377–394.

Caspi, A., 1998. Personality development across the life course. In: Damon, W., Eisenberg, N. (Eds.), Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development. Wiley, New York, pp. 311–388.

Cassidy, J., Woodhouse, S.S., Sherman, L.J., Stupica, B., Lejuez, C.W., 2011. Enhancing infant attachment security: an examination of treatment efficacy and differential susceptibility. Development and Psychopathology 23 (1), 131–148.

Chen, X., Chen, H., Li, D., Wang, L., 2009a. Early childhood behavioral inhibition and social and school adjustment in Chinese children: a 5-year longitudinal study. Child Development 80 (6), 1692–1704.

Chen, X., Hastings, P.D., Rubin, K.H., Chen, H., Cen, G., Stewart, S.L., 1998. Child-rearing attitudes and behavioral inhibition in Chinese and Canadian toddlers: a cross-cultural study. Developmental Psychology 34 (4), 677–686.

Chen, X., Rubin, K.H., Li, B.S., Li, D., 1999. Adolescent outcomes of social functioning in Chinese children. International Journal of Behavioral Development 23 (1), 199–223.

Chen, X., Wang, L., Wang, Z., 2009b. Shyness-sensitivity and social, school, and psychological adjustment in rural migrant and urban children in China. Child Development 80 (5), 1499–1513.

Chronis-Tuscano, A., Degnan, K.A., Pine, D.S., Perez-Edgar, K., Henderson, H.A., Diaz, Y., Raggi, V.L., Fox, N.A., 2009. Stable early maternal report of behavioral inhibition predicts lifetime social anxiety disorder in adolescence. Journal of the American Academy of Child & Adolescent Psychiatry 48 (9), 928–935.

Clauss, J.A., Blackford, J.U., 2012. Behavioral inhibition and risk for developing social anxiety disorder: a meta-analytic study. Journal of the American Academy of Child & Adolescent Psychiatry 51 (10), 1066–1075.

Capaldi, D.M., Rothbart, M.K., 1992. Development and validation of an early adolescent temperament measure. The Journal of Early Adolescence 12 (2), 153–173.

Degnan, K.A., Fox, N.A., 2007. Behavioral inhibition and anxiety disorders: multiple levels of a resilience process. Development and Psychopathology 19 (3), 729–746.

Degnan, K.A., Almas, A.N., Fox, N.A., 2010. Temperament and the environment in the etiology of childhood anxiety. Journal of Child Psychology and Psychiatry 51 (4), 497–517.

Dennis, T.A., 2010. Neurophysiological markers for emotion regulation from the perspective of emotion-cognition integration: current directions and future challenges. Developmental Neuropsychology 35, 212–230.

Derryberry, D., Rothbart, M.K., 1997. Reactive and effortful processes in the organization of temperament. Developmental Psychopathology 9 (4), 633–652.

Eisenberg, N., Spinrad, T.L., Eggum, N.D., 2010. Emotion-related self-regulation and its relation to children's maladjustment. Annual Review of Clinical Psychology 6, 495–525.

Fox, N.A., Henderson, H.A., 1999. Does infancy matter? Predicting social behavior from infant temperament. Infant Behavior and Development 22 (4), 445–455.

Fox, N.A., Henderson, H.A., Marshall, P.J., Nichols, K.E., Ghera, M.M., 2005. Behavioral inhibition: linking biology and behavior within a developmental framework. Annual Review of Psychology 56, 235–262.

Fox, N.A., Henderson, H.A., Rubin, K.H., Calkins, S.D., Schmidt, L.A., 2001. Continuity and discontinuity of behavioral inhibition and exuberance: psychophysiological and behavioral influences across the first four years of life. Child Development 72 (1), 1–21.

Fullard, W., McDevitt, S.C., Carey, W.B., 1984. Assessing temperament in one-to three-year-old children. Journal of Pediatric Psychology 9 (2), 205–217.

Gartstein, M.A., Peleg, Y., Young, B.N., Slobodskaya, H.R., 2009. Infant temperament in Russia, United States of America, and Israel: differences and similarities between Russian-speaking families. Child Psychiatry and Human Development 40 (2), 241–256.

Goldsmith, H.H., 1996. Studying temperament via construction of the Toddler Behavior Assessment Questionnaire. Child Development 67 (1), 218–235.

Goldsmith, H.H., Buss, A.H., Plomin, R., Rothbart, M.K., Thomas, A., Chess, S., Hinde, R.A., McCall, R.B., 1987. Roundtable: what is temperament? Four approaches. Child Development, 505–529.

Goldsmith, H.H., Campos, J.J., 1982. Toward a theory of infant temperament. In: Emde, A.N., Harmon, R.J. (Eds.), The Development of Attachment and Affiliative Systems. Plenum, New York, pp. 161–193.

Goldsmith, H.H., Gagne, J.R., 2012. Behavioral assessment of temperament. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 209–228.

Goldsmith, H.H., Rothbart, M.K., 1993. The Laboratory Temperament Assessment Battery (LAB-TAB). University of Wisconsin Press, Madison.

Goldsmith, H.H., Campos, J.J., 1986. Fundamental issues in the study of early temperament: the Denver Twin Temperament Study. In: Lamb, M.E., Brown, A. (Eds.), Advances in Developmental Psychology. Erlbaum, Hillsdale, NJ, pp. 231–283.

Helfinstein, S.M., Fox, N.A., Pine, D.S., 2012. Approach–withdrawal and the role of the striatum in the temperament of behavioral inhibition. Developmental Psychology 48 (3), 815–826.

Hegvik, R.L., McDevitt, S.C., Carey, W.B., 1982. The middle childhood temperament questionnaire. Journal of Developmental & Behavioral Pediatrics 3 (4), 197–200.

Kagan, J., 2003. Behavioral inhibition as a temperamental category. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, New York, pp. 320–331.

Kagan, J., 2012. The biography of behavioral inhibition. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 69–82.

Kagan, J., Fox, N.A., 2007. Biology, culture, and temperamental issues. In: Damon, W., Lerner, R., Eisenberg, N. (Eds.), Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development, sixth ed. Wiley, New York.

Kagan, J., Reznick, J.S., Clarke, C., Snidman, N., Garcia-Coll, C., 1984. Behavioral inhibition to the unfamiliar. Child Development 55 (6), 2212–2225.

Kagan, J., Snidman, N., 1991. Temperamental factors in human development. American Psychologist 46 (8), 856–862.

Kagan, J., 1994. Galen's Prophecy. Basic Books, New York.

Lahat, A., Lamm, C., Chronis-Tuscano, A., Pine, D.S., Henderson, H.A., Fox, N.A., 2014. Early behavioral inhibition and increased error monitoring predict later social phobia symptoms in childhood. Journal of the American Academy of Child & Adolescent Psychiatry 53 (4), 447–455.

Lengua, L.J., 2002. The contribution of emotionality and self-regulation to the understanding of children's response to multiple risk. Child Development 73 (1), 144–161.
Lengua, L.J., 2008. Anxiousness, frustration, and effortful control as moderators of the relation between parenting and adjustment in middle-childhood. Social Development 17 (3), 554–577.
Lengua, L.J., Wachs, T.D., 2012. Temperament and risk: resilient and vulnerable responses to adversity. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 519–540.
Lewis-Morrarty, E., Degnan, K.A., Chronis-Tuscano, A., Rubin, K.H., Cheah, C.S., Pine, D.S., Heather, A.H., Fox, N.A., 2012. Maternal over-control moderates the association between early childhood behavioral inhibition and adolescent social anxiety symptoms. Journal of Abnormal Child Psychology 40 (8), 1363–1373.
Loken, E., 2004. Using latent class analysis to model temperament types. Multivariate Behavioral Research 39 (4), 625–652.
Lonigan, C.J., Vasey, M.W., 2009. Negative affectivity, effortful control, and attention to threat-relevant stimuli. Journal of Abnormal Child Psychology 37 (3), 387–399.
McDermott, J.M., Pérez-Edgar, K., Henderson, H.A., Chronis-Tuscano, A., Pine, D.S., Fox, N.A., 2009. A history of childhood behavioral inhibition and enhanced response monitoring in adolescence are linked to clinical anxiety. Biological Psychiatry 65 (5), 445–448.
Muris, P., Merckelbach, H., Schmidt, H., Gadet, B., Bogie, N., 2001. Anxiety and depression as correlates of self-reported behavioral inhibition in normal adolescents. Behavior Research and Therapy 39 (9), 1051–1061.
McDevitt, S.C., Carey, W.B., 1978. The measurement of temperament in 3–7 year old children. Journal of Child Psychology and Psychiatry 19 (3), 245–253.
Pérez-Edgar, K., Bar-Haim, Y., McDermott, J.N.M., Chronis-Tuscano, A., Pine, D.S., Fox, N.A., 2010a. Attention biases to threat and behavioral inhibition in early childhood shape adolescent social withdrawal. Emotion 10 (3), 349–357.
Pérez-Edgar, K., McDermott, J.N.M., Korelitz, K., Degnan, K.A., Curby, T.W., Pine, D.S., Fox, N.A., 2010b. Patterns of sustained attention in infancy shape the developmental trajectory of social behavior from toddlerhood through adolescence. Developmental Psychology 46, 1723–1730.
Pérez-Edgar, K., Fox, N.A., 2005a. Temperament and anxiety disorders. Child and adolescent psychiatric Clinics of North America 14 (4), 681–706.
Pérez-Edgar, K., Fox, N.A., 2005b. A behavioral and electrophysiological study of children's selective attention under neutral and affective conditions. Journal of Cognition and Development 6 (1), 89–118.
Pérez-Edgar, K., Hardee, J., Guyer, A.E., Benson, B., Nelson, E., Gorodetsky, E., Goldman, D., Fox, N.A., Pine, D.S., Ernst, M., 2013. DRD4 and striatal modulation of the link between childhood behavioral inhibition and adolescent anxiety. Social Cognitive and Affective Neuroscience 9 (4), 445–453.
Pérez-Edgar, K., Reeb-Sutherland, B.C., McDermott, J.M., White, L.K., Henderson, H.A., Degnan, K.A., Hane, A.A., Pine, D.S., Fox, N.A., 2011. Attention biases to threat link behavioral inhibition to social withdrawal over time in very young children. Journal of Abnormal Child Psychology 39, 885–895.
Pérez-Edgar, K., Roberson-Nay, R., Hardin, M.G., Poeth, K., Guyer, A.E., Nelson, E.E., McClure, E.B., Henderson, H.A., Fox, N.A., Pine, D.S., Ernst, M., 2007. Attention alters neural responses to evocative faces in behaviorally inhibited adolescents. Neuroimage 35 (4), 1538–1546.
Pérez-Edgar, K., Taber-Thomas, B., Auday, E., Morales, S., 2014. Temperament and attention as core mechanisms in the early emergence of anxiety. In: Lagattuta, K. (Ed.), Children and Emotion: New Insights into Developmental Affective Science, vol. 26. Karger Publishing, pp. 42–56.
Pluess, M., Belsky, J., 2009. Differential susceptibility to rearing experience: the case of childcare. Journal of Child Psychology and Psychiatry 50 (4), 396–404.
Posner, M.I., Rothbart, M.K., Sheese, B.E., Voelker, P., 2014. Developing attention: behavioral and brain mechanisms. Advances in Neuroscience Article ID 405094, 9. Hindawi Publishing Corporation.
Putnam, S.P., Gartstein, M.A., Rothbart, M.K., 2006. Measurement of fine-grained aspects of toddler temperament: the Early Childhood Behavior Questionnaire. Infant Behavior and Development 29 (3), 386–401.
Reiss, D., Leve, L.D., Neiderhiser, J.M., 2013. How genes and the social environment moderate each other. American Journal of Public Health 103 (Suppl. 1), S111–S121.
Rothbart, M.K. Temperament Questionnaires, 2006. http://www.bowdoin.edu/~sputnam/rothbart-temperament-questionnaires/instrument-descriptions/.
Rothbart, M.K., Bates, J.E., 2007. Temperament. In: Damon, W., Lerner, R., Eisenberg, N. (Eds.), Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development, sixth ed. Wiley, New York.
Rothbart, M.K., Derryberry, D., 1981. Development of individual differences in temperament. In: Lamb, M.E., Brown, A.L. (Eds.), Advances in Developmental Psychology, vol. 1. Erlbaum, Hillsdale, NJ, pp. 37–86.
Rothbart, M.K., Derryberry, D., 2002. Temperament in children. Psychology at the Turn of the Millennium 2, 17–35.
Rothbart, M.K., Sheese, B.E., Posner, M.I., 2007. Executive attention and effortful control: linking temperament, brain networks, and genes. Child Development Perspectives 1 (1), 2–7.
Rubin, K.H., Bowker, J., Kennedy, A., 2008. Avoiding and withdrawing from the peer group in middle childhood and early adolescence. In: Rubin, K.H., Bukowski, W., Laursen, B. (Eds.), Handbook of Peer Interactions, Relationships, and Groups. Guilford, New York.
Rubin, K.H., Coplan, R.J., Bowker, J.C., 2009. Social withdrawal in childhood. Annual Review of Psychology 60, 141–171.
Rubin, K.H., Wojslawowicz, J.C., Rose-Krasnor, L., Booth-LaForce, C., Burgess, K.B., 2006. The best friendships of shy/withdrawn children: prevalence, stability, and relationship quality. Journal of Abnormal Child Psychology 34 (2), 139–153.
Rueda, M.R., 2012. Effortful control. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 145–167.
Rothbart, M.K., 1981. Measurement of temperament in infancy. Child Development 52, 569–578.
Rothbart, M.K., Ahadi, S.A., Hershey, K.L., Fisher, P., 2001. Investigations of temperament at 3-7 years: the Children's Behavior Questionnaire. Child Development 72, 1394–1408.
Saudino, K.J., 2009. The development of temperament from a behavioral genetics perspective. Advances in Child Development and Behavior 37, 201–231.
Saudino, K.J., Wang, M., 2012. Quantitative and molecular genetic studies of temperament. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 315–346.
Saudino, K.J., Zapfe, J.A., 2008. Genetic influences on activity level in early childhood: do situations matter? Child Development 79 (4), 930–943.
Schwartz, C.E., Kunwar, P.S., Greve, D.N., Kagan, J., Snidman, N.C., Bloch, R.B., 2011. A phenotype of early infancy predicts reactivity of the amygdala in male adults. Molecular Psychiatry 17 (10), 1042–1050.
Schwartz, C.E., Wright, C.I., Shin, L.M., Kagan, J., Rauch, S.L., 2003. Inhibited and uninhibited infants "grown up": adult amygdalar response to novelty. Science 300 (5627), 1952–1953.
Shiner, R.L., Buss, K.A., McClowry, S.G., Putnam, S.P., Saudino, K.J., Zentner, M., 2012. What is temperament now? Assessing progress in temperament research on the twenty-fifth anniversary of Goldsmith et al. (1987). Child Development Perspectives 6 (4), 436–444.
Shiner, R.L., Caspi, A., 2012. Temperament and the development of personality traits, adaptations, and narratives. In: Zentner, M., Shiner, R. (Eds.), The Handbook of Temperament. Guilford Press, New York, pp. 497–518.
Sroufe, L.A., 2009. The concept of development in developmental psychopathology. Child Development Perspectives 3 (3), 178–183.
Super, C.M., Harkness, S., 2002. Culture structures the environment for development. Human Development 45, 270–274.
Simonds, J., Rothbart, M.K., October 2004. The Temperament in Middle Childhood Questionnaire (TMCQ): a computerized self-report measure of temperament for ages 7–10. In: Poster Session Presented at the Occasional Temperament Conference. Athens, GA.
Thomas, A., Chess, S., 1977. Temperament and Development. Brunner/Mazel, New York.
Todd, R.M., Cunningham, W.A., Anderson, A.K., Thompson, E., 2012. Affect-biased attention as emotion regulation. Trends in Cognitive Sciences 16 (7), 365–372.
Wichmann, C., Coplan, R.J., Daniels, T., 2004. The social cognitions of socially withdrawn children. Social Development 13 (3), 377–392.
White, L.K., McDermott, J.M., Degnan, K.A., Henderson, H.A., Fox, N.A., 2011. Behavioral inhibition and anxiety: the moderating roles of inhibitory control and attention shifting. Journal of Abnormal Child Psychology 39 (5), 735–747.
Williams, L.R., Degnan, K.A., Perez-Edgar, K.E., Henderson, H.A., Rubin, K.H., Pine, D.S., Steinberg, L., Fox, N.A., 2009. Impact of behavioral inhibition and parenting style on internalizing and externalizing problems from early childhood through adolescence. Journal of Abnormal Child Psychology 37 (8), 1063–1075.
Zentner, M., Bates, J.E., 2008. Child temperament: an integrative review of concepts, research programs, and measures. International Journal of Developmental Science 2 (1), 7–37.

Temporal Distinctiveness Models of Memory

Gordon DA Brown, University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Temporal distinctiveness models of memory are described and the history of their development is reviewed. Evidence that memory is temporally organized is discussed, and it is shown how temporal distinctiveness models account for a range of data from short-term and long-term memory tasks without assuming the existence of a separate short-term memory system.

Introduction

This article is about temporal distinctiveness models of memory. But what does it mean to say that something is distinctive? In everyday language, we say that something is distinctive to the extent that it is different from other things. A wine may have a distinctive aroma; a person may have a distinctive laugh. When thinking about memory, a similar concept of distinctiveness can be applied. Memories that are distinctive are memories that are different from, and hence unlikely to get confused with, other memories. Distinctiveness is seen as the inverse of similarity. Distinctiveness models of memory must specify exactly how to measure distinctiveness and also need to explain how the distinctiveness of memories makes them easy or difficult to recall.

We can introduce memory distinctiveness using a spatial analogy. It is now common to think of memory as being like a large space in which long-term memories – such as the meanings of words in a language – and also short-term memories occupy their own locations. Memory retrieval then requires finding the place in memory where the target is to be found. At an intuitive level, the process can be seen as analogous to picking out an object from an array of other objects on a table. Imagine a fairground stall in which an array of prizes to be won is scattered around on a large table some distance away from you. The prizes differ in value, and your task is to throw a counter onto the table. You receive whichever prize your counter has landed closest to. Of course, you want to receive one of the more valuable prizes, so you aim to throw your counter so that it settles closer to your target than to anything else. The more isolated your target item is (the more empty space it has around it), the easier this task will be. If there are many low-value prizes close to your high-value target, then your throw will have to be very accurate in order to win the prize. In this case, the location of the prize on the table is less distinctive – it occupies a similar location to many others, and hence you may by accident get one of the other prizes. Retrieving items from memory can be thought of as analogous. The counter that you throw onto the table is like a retrieval cue that you put into memory space. If the item that you are trying to retrieve from memory occupies a fairly isolated region of memory (i.e., there are relatively few very similar items that occupy nearby locations in memory, and hence interfering with your search), it will be easier to retrieve your target item.

Later we will show how recent models of memory quantify these notions of similarity, distance, and distinctiveness. First, however, we need to explain the idea of *temporal* distinctiveness. We explain temporal distinctiveness in terms of an analogy, introduced by Crowder (1976), which captures the intuition that underlies a number of distinctiveness approaches to memory, reviewed below, which were developed in the early 1970s. Crowder asked us to imagine a line of evenly spaced telephone poles that are being passed by a moving train. To an observer on the train, relatively nearby telephone poles (poles that have just been passed) are easy to discriminate from one another – they are distinctive. However, as the train moves on and the telephone poles recede into the distance, they become increasingly difficult to discriminate from one another – they become less distinctive. According to Crowder's analogy, we can think of memories receding into the past as time moves on just as the telephone poles recede into the spatial distance as the train moves on. Thus the central idea of temporal distinctiveness models of memory is that items become less discriminable from one another, and hence less distinctive and harder to retrieve, as they fade into the past in just the same way as telephone poles become less discriminable as they recede into the spatial distance. Next, we will discuss how this general idea led to a series of increasingly sophisticated temporal distinctiveness models of memory.

History

We first consider an approach to memory distinctiveness represented by what we term global distinctiveness models, then show how global distinctiveness models have been supplemented by local distinctiveness models. We then note a variety of more general temporal context models, which have developed in parallel and some of which can be seen as implementing the principles of temporal distinctiveness.

Early Global Distinctiveness Models

More than half a century ago, Murdock (1960) put forward an influential distinctiveness model of memory and identification. The underlying idea was that the distinctiveness of an item can be defined as the sum of its distances from other items in a context (the context could be something like a list of items to be retrieved or a set of objects differing in weight). For example, consider a five-item list of words to be recalled using their serial

positions within the list. The items to be remembered are (A B C D E), and their serial positions in the list are therefore [1 2 3 4 5]. Consider the second item, B. This position of B (i.e., 2) is one different from the position of A, one different from the position of C, two different from the position of D, and three different from the position of E. Therefore the summed difference of the position of B from the position of the other four items is $1 + 1 + 2 + 3 = 7$. If we calculate the summed difference for every item in this way, we end up with [10 7 6 7 10]. It is often then convenient to divide these numbers by the total, which gives us [.25 .175 .15 .175 .25] – a measure of the relative distinctiveness of each item. (For the purposes of this illustration we ignore possible transformations of items' values along a dimension, which will often be important.) These numbers make a plausible serial position curve – items near the end of the series are more distinctive than the mid-series items. This type of distinctiveness model – i.e., with distinctiveness calculated as the summed difference between an item and all other items – was extended by Neath (1993) to describe items' distinctiveness along a temporal dimension. Neath applied the model to primacy and recency effects in memory and their interactions with retention interval. We refer to this type of model as a *global distinctiveness* model, and note that it is quite different from the type of model that we outlined in the introduction. Going back to the fairground example, the difficulty of making your counter land close to your target prize is determined by the distance between the target and its nearby neighbors on the table. The distance between the target prize and a faraway prize, on the other side of the table, will not make the target any easier to win. We refer to models as *local distinctiveness* models (Neath et al., 2006) when it is mainly the similarities of nearby items, rather than the similarities of distant items as in the global distinctiveness model, that matter most. Considerable evidence, from both perceptual and identification tasks and memory research, supports the idea of local rather than global distinctiveness, and from this point on when we talk about 'distinctiveness' we will be referring solely to local distinctiveness.

Local Distinctiveness Models

The 1970s saw the development of a number of local distinctiveness models of memory, although they were not always labeled as such at the time. Most of the early models are often called *ratio-rule models*. Bjork and Whitten (1974) were interested in the explanation of recency effects in memory, which disappear after a filled retention interval but can reappear if the time that separates to-be-remembered items is increased (so-called long-term recency effects). They suggested that memory performance would depend on two amounts of time – the time separating items in a list to be remembered, and the time elapsed since those items were presented. This approach can be illustrated by the telephone pole analogy. Imagine that the telephone poles are separated by 5 m. Now consider the difficulty of discriminating two telephone poles that are 20 and 25 m away (analogous to two memory traces that were laid down 20 and 25 s in the past). The ratio of these two distances (or times) is $20/25 = 0.8$. If the difficulty in discriminating the telephone poles – or memories – depends on the ratio of their distances from the observer (of telephone poles) or recaller (of memories), then the distant items will be more confusable than the items that are just 5 and 10 (meters or seconds) away because in the latter case the relevant ratio is only $5/10 = 0.5$. Increasing the temporal gap separating items in a to-be-remembered list will reduce confusability even if temporal distance is constant.

A number of ratio-like models of memory emerged (e.g., Baddeley, 1976; Glenberg and Swanson, 1986, and several others). These early ratio-rule accounts can be seen as types of local temporal distinctiveness model. In stating that discriminability of a pair of items depends on the amount of time separating them, ratio models claim that the discriminability of an item is reduced just by the nearest of its closest temporal neighbors. Indeed it can be argued that ratio rule models are actually *too* local, and that the distinctiveness of an item in memory will likely be reduced not just by its single closest neighbor, but also (albeit perhaps to a lesser extent) by other nearby neighbors. The fact that the ratio-rule models we have described are so local means that – although they are good at accounting for recency effects (as with the telephone pole analogy)– they cannot without modification account for primacy effects – i.e., the tendency for several items near the start of a list to be better remembered than those in the middle portions of the list. A simple ratio rule model predicts that, if memories are evenly spaced in time, they become ever less distinctive as they recede into the past; there is no mechanism for primacy effects to emerge.

There are two ways in which primacy effects can be explained while preserving the basic idea of the ratio rule. One possibility is that rehearsal of early list items effectively increases their recency, by carrying the items forward through the list, with the result that a simple ratio-rule model can account for primacy effects, as well as recency effects, when rehearsal-induced recency is taken into account (Tan and Ward, 2000). Such an account offers the appealing prospect of being able to account for both primacy and recency effects, as well as differences between memory tasks, with a single simple model (Ward et al., 2010).

Another way of accounting for primacy effects within a temporal distinctiveness framework is given by a recent extension of temporal ratio models, called SIMPLE (Scale-Invariant Memory, Perception, and Learning: Brown et al., 2007b). SIMPLE states, as do some temporal ratio models, that the confusability of any two items in memory will depend on the ratio of the items' recall intervals (i.e., their temporal distances from the point of recall). However, this applies to every pairwise combination of items – although the confusability of an item with its nearest temporal neighbors will be most important, more distant temporal neighbors will also exert an influence, although their influence will be smaller. For example, consider four memory traces that were laid down 20, 19, 18, and 17 s in the past. The oldest item (that occurred 20 s ago) will be 19/20 confusable with the next oldest item, but also 18/20 confusable with the third item in the list and 17/20 confusable with the fourth item in the list. The distinctiveness of an item in memory is assumed to be inversely proportional to the sum of the item's confusabilities with all other items in a list. Brown et al. show that such an account, where all items in a target's temporal neighborhood can cause interference, can

account for primacy effects for the first few items in a list even without the assumption of rehearsal. This is because the items at the end of the list only have neighbors (from which they must be discriminated) on one side of them. Another difference between SIMPLE and earlier ratio-rule models is that SIMPLE contains a free parameter, which raises the temporal distance ratios to a power–this parameter effectively governs the rate at which confusability increases with temporal distance ratio. Brown et al. (2007a) and Neath and Brown (2006) show that this extended version of a ratio-like temporal distinctiveness model can account for a range of effects from both serial and free recall memory tasks.

Temporal Context Models

The temporal distinctiveness models outlined so far offer rather abstract descriptions of memory retrieval as a temporal discrimination process. The more general idea that constraining memory search using temporal/contextual cues might be important in memory retrieval also has a long history (e.g., Glenberg et al., 1983). More recently, the temporal/contextual model of working memory (Shipstead and Engle, 2013; Unsworth and Engle, 2006) suggests that low working memory capacity is associated with problems in locating sufficiently precise temporally defined regions of long-term memory.

What sorts of mechanisms, perhaps closer to a neural network level of description, might underlie such accounts? Several models have assumed that a temporal/contextual signal of some kind becomes associated with events at the time they are learned. To the extent that the relevant temporal/contextual signal can be used as a retrieval cue, items that were associated with it can be recalled. The idea that contextual elements, often ones assumed to change gradually over time, can become associated with events is not new (see e.g., Bower, 1972; Hintzman et al., 1973; Mensink and Raaijmakers, 1988). More recent models instantiate the idea of temporal context in various ways. For example, the OSCAR model (Brown et al., 2000) assumes the neural availability of an array of oscillators, similar to those used in independently developed models of timing (Church and Broadbent, 1990) which create a time-varying context signal to which items become associated. Reinstatement of the context signal allows it to wind forward, reconstructing the state at presentation and leading to serial recall (see also Houghton, 1996). A recent model by Farrell (2012) assumes that clusters of items are represented as episodes by being linked together through their associations to a common temporal context, and the model accounts for a wide range of findings from serial and free recall memory tasks.

Although space prevents a comprehensive review of temporal context models, not all such models are appropriately categorized as temporal distinctiveness models. For example, the influential temporal context model and its successors (e.g., Howard and Kahana, 2002; Polyn et al., 2009; Sederberg et al., 2008) has, as a key component, a temporal context that provides part of the learning context for a given item. The learning context is also given partly by preexperimental associations, and whenever an item is retrieved there is an associated recall of the learning contexts associated with the recalled item. The retrieved learning context will tend to cue recall of items that were studied nearby in time. This type of model may predict that temporally isolated items will be less well recalled than nearby items (i.e., the opposite prediction to the one normally made by a temporal distinctiveness account) because a retrieved cue will act as a less good probe for items that were studied at a greater temporal distance from the last recalled item (Polyn et al., 2014).

To summarize: Temporal distinctiveness models of memory have two core features. First, they assume that memory is chronologically organized – that is to say, memories of events that occurred nearby in time are located in nearby locations in memory. Second, they assume that memory traces that are more temporarily isolated will be more distinctive and easier to retrieve. Some, but not all, temporal distinctiveness and temporal context models also emphasize the continuity of memory over widely varying timescales – that is, they assume that the same principles of temporal distinctiveness underpin both short-term and long-term memory tasks, without assuming a distinction between underlying short-term and long-term stores. We now explore each of these assumptions in more detail.

The Temporal Organization of Memory

What does it mean to say that memory is temporally organized? Human episodic memory has long been viewed as a system for storing and retrieving temporally dated episodes (Tulving, 1972). A key claim of current temporal distinctiveness models of memory is the specific idea that human episodic memory is organized in such a way that the temporal relationship between the time an event occurred and the time of remembering is represented in the memory system. We should note that this is not the same as the claim that memories are somehow labeled with the calendar time they occur; there are considerable problems with such an account. The claim is that the temporal distances from the present, rather than some calendrical representation of the temporal location at which it occurred, is what is relevant.

Moving beyond the claim that memory is chronologically organized, an obvious question is *why* it might be so. Some researchers have argued that a 'rational analysis' of memory can be undertaken, such that the time course of forgetting reflects the organization of an optimally designed memory system in which the memories that are most likely to need to be retrieved are the ones that are easiest to retrieve. Anderson and Schooler (1991) demonstrated that items that have been encountered more recently are more likely to need to be retrieved, and that the decay of 'need probability' (the probability that an item will need to be recalled) occurs over time in a manner paralleled by the form of the forgetting curve. Thus time may be a basic organizing dimension in a memory system that adaptively reflects the statistical structure of the events in the world that it must remember.

Furthermore, as Gallistel (1990) has noted, it may be that the temporal (as well as the spatial) coordinates associated with an event form a crucial component of memory records in order to enable the unity of remembered experiences to be maintained. In addition, provided time can be represented as nonrepeating, the provision of temporal coordinates provides

a unique address for every item. Stored temporal information is used by many animals for both foraging and navigation (Clayton and Dickinson, 1988; Gallistel, 1990).

Thus there are several reasons why an implicit representation of temporal distance in the past may be an important component of memories. It should be borne in mind that claims of temporal distinctiveness models are, however, stronger than simply the claim that time is an important underlying dimension of memory; models may assume that time is important without necessarily predicting effects of temporal distinctiveness.

Serial Position Curves and Forgetting

What sorts of evidence are relevant to evaluating temporal distinctiveness models of memory, and how do the models account for key memory phenomena? We first give a brief explanation of how temporal distinctiveness models account for serial position curves and forgetting, and then examine sources of evidence that test specific predictions.

The intuitive explanation for recency effects in temporal distinctiveness models of memory is the one illustrated by the telephone pole analogy described earlier – items become less discriminable from one another as they recede into the temporal distance. We illustrate this, as well as how temporal distinctiveness can explain primacy effects, with a worked example based on the SIMPLE model.

Figure 1 shows a list of four to-be-remembered items (the filled squares), with the passage of time moving from left to right. Each item has written above it its temporal distance from the time of recall (i.e., its distance from the present). Consider the second item in the list. This has a temporal distance of 6 s, meaning that it occurred 6 s before the time of recall. The distinctiveness of the item in memory will depend on its confusability with other items along the temporal dimension – the more confusable it is, the lower will be its discriminability. These confusabilities will depend on the temporal distance ratios, so that the confusability of the item with the first list item will be 6/8, its confusability with the third list item will be 4/6, and its confusability with the final list item will be 2/6. We also have to consider the similarity of the item to itself, which is always 1.0. The sum of these confusabilities is then 2.75, and the discriminability of the item in memory will be inversely proportional to this summed confusability, i.e., it will be 0.36.

We can calculate the discriminabilities of the other three items in the list in exactly the same way, and they turn out to be 0.40, 0.38, and 0.48 for the first, third, and fourth items, respectively.

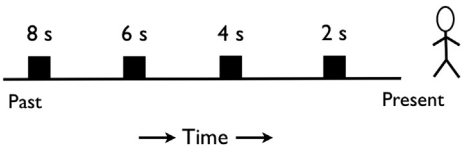

Figure 1 Illustration of the temporal distances from the point of retrieval of four items in a to-be-remembered list.

This worked example illustrates three key points. First, as shown earlier, there is a general trend for items further in the past to be less distinctive than more recent items – as in the telephone analogy and in earlier 'ratio-rule' models within a local temporal distinctiveness framework. Second, unlike in the earlier ratio-rule models, the first item on the list is more distinctive than the second item on the list (0.4 vs 0.36) – a primacy effect. The effect is only visible for the very first item in our example, because the list is very short. If we did the same calculation for a longer list, however, the primacy effect would extend over many more items. Third, we can see how forgetting and the reduction of recency effects occur. Consider, for example, the second and the fourth items (which have distinctiveness values of 0.36 and 0.48, respectively). Suppose we now increase the retention interval from 2 s to 32 s. This will lead to every item becoming more confusable with its neighbors. For example, the similarity of the first and second items, originally 0.75 (the ratio of their temporal distances; 6/8) will now be $36/38 = 0.95$. Because all the temporal distances are larger, all the ratios are increased and the distinctiveness of all items is reduced. In fact the distinctiveness values for all items after the delay, taking into account the confusability of each item with the other three, are 0.27, 0.26, 0.26, and 0.27. It is evident that, because of the ratio properties of the temporal distance confusability measure, the most recent item loses much or all of its relative advantage in distinctiveness. This can again be understood in terms of the telephone pole analogy – if one walks away from the end of the line of telephone poles, the poles that one was previously standing next to lose their relative advantage in discriminability. Although this account is only a simplified version of a complete model (which would also incorporate mechanisms to allow for items to be either recalled or not recalled as a function of their distinctiveness), it captures the essence of one temporal ratio model of memory.

In the remaining sections, we illustrate the types of evidence appropriate to evaluation of temporal distinctiveness models of memory, and consider how the models relate to wider issues of memory organization.

Evidence for Temporal Distinctiveness

Various types of experiments have tested the claims that memory is temporally organized and that temporal distinctiveness influences recall, and we review these in turn.

Memory as Temporal Discrimination

Distinctiveness models assume that memory retrieval is akin to temporal discrimination. One approach is to test this assumption directly by using temporal discrimination tasks and asking whether the discriminability of particular durations of time can predict the recallability of memory items that have corresponding distances in the past. Consider a task in which participants are required to identify tones in terms of their temporal duration (e.g., telling a 0.5 s tone apart from a 0.75 s tone; telling apart a 1.0 s tone from a 1.5 s tone). In a series of experiments examining both the discriminability of different temporal durations and memory for letters and words, Brown et al. (2009) found that the relative memorability of list

items mirrored the relative discriminability of temporal durations (the tones) that corresponded to the temporal distances of the to-be-remembered items. These results appear consistent with the idea that memory retrieval involves temporal discrimination.

Related work examines whether time, rather than the number of interfering events, influences memory judgments. Hintzman (2004) presented lists of words that varied in presentation rate, with some items being repeated. On encountering a repeated item, participants were asked to judge the number of items that had intervened since the last presentation of the item. Hintzman found that the amount of intervening time, rather than the number of intervening words, determined the judgment. If it is assumed that recency judgments are based on the difference between the learning contexts associated with the two occurrences of the relevant item, this result seems consistent with temporal accounts in which the contextual signal changes with the passage of time.

Time and forgetting

What causes forgetting? In particular, is it the passage of time since an item entered memory, or the number of interfering events (such as other items in a to-be-remembered list) that have occurred since the item entered memory, that determines forgetting? The debate about whether time or interference causes forgetting considerably predates temporal distinctiveness models but is relevant to their evaluation.

A temporal distinctiveness model of the type described above argues that forgetting can occur due to the passage of time alone even though interference alone is responsible for retrieval failure and no trace decay occurs (cf. the analogy of evenly spaced telephone poles receding into the distance and becoming less discriminable from one another). This no-decay assumption contrasts with the claims of many influential models of short-term and working memory, according to which, traces decay as time passes (e.g., Barrouillet et al., 2004; Page and Norris, 1998). Space does not permit a review of the large current literature on the causes of short-term forgetting. Although recently some authors have argued against forgetting due to time-based decay in short-term and working memory (Berman et al., 2009; Lewandowsky et al., 2004, 2009), the issue remains controversial both theoretically and empirically (Altmann and Schunn, 2012; Barrouillet and Camos, 2012; Campoy, 2012; Ricker and Cowan, 2010).

The key point here is that, evidence that forgetting occurs due to the passage of time alone does not in itself provide evidence for or against temporal distinctiveness models. Instead, deciding between trace decay models and temporal distinctiveness models requires examination of the *relative* temporal spacing between events in memory. In particular, temporal distinctiveness models predict that the temporal position of items prior to the to-be-remembered material will influence memory, while trace decay theory makes no such prediction. While some experiments have begun to look at this issue, evidence for temporal distinctiveness has been mixed, with some supportive evidence (Bright-Paul and Jarrold, 2012), but other studies finding no effect of distinctiveness at least for short-term auditory memory (McKeown and Mercer, 2012; Mercer and McKeown, 2014).

Temporal Isolation Effects

The final source of evidence that we consider for temporal distinctiveness models concerns temporal isolation effects. A key prediction of temporal distinctiveness models is that items that are temporally isolated at presentation (i.e., that are surrounded by large temporal gaps at the time of learning) should be better recalled. The prediction is that such items will be more easily recalled because they are less confusable with their immediate neighbors. While it is not possible to review the several dozen experiments that have examined the issue, the conclusions that have been reached to date depend on the nature of the memory task that is being used and on whether the presentation schedules are predictable or unpredictable.

The basic methodology involves varying the rate of presentation in such a way that some items are more temporally isolated than others. Presentations may be either predictable or unpredictable. Most early experiments examined memory for serial order, usually of verbal material, as the recall task. Several early studies using predictable presentation schedules found an advantage for items in less temporally crowded regions of the list, suggesting that the temporal isolation of items, not just their positions in the list, influence recall (e.g., Neath and Crowder, 1990). Several subsequent studies of forward serial recall found that the apparent effect of temporal isolation disappears when random schedules of presentation are used, but the most recent studies of serial recall have found that, even with random presentation schedules, effects of temporal isolation can be obtained in serial order memory tasks when the task demands encourage representation along a temporal dimension, as, for example, when order of report is unconstrained (Lewandowsky et al., 2008). Temporal isolation effects have been found in running memory tasks when list length is unpredictable (Geiger and Lewandowsky, 2008) and in recognition and free recall tasks (Morin et al., 2010), although Polyn et al. (2014) found, using longer temporal items than did Morin et al., that temporally isolated items were *less* well remembered in a free recall task and suggested that the Morin et al.'s finding represented stronger encoding, rather than enhanced retrieval of temporally isolated items. Most of these experiments have studied memory for verbal material; studies of memory for nonverbal material have sometimes found temporal isolation effects (e.g., Guérard et al., 2010; Shipstead and Engle, 2013) and sometimes not (e.g., Parmentier et al., 2006). Thus the evidence for temporal isolation effects is mixed, and those that have been observed in serial recall maybe due to effects of grouping rather than temporal isolation per se (Farrell et al., 2011). Indeed Farrell (2012) shows with a temporal clustering model that isolation effects in serial and free recall can emerge without the assumption of temporal distinctiveness.

In summary, while reliable temporal isolation effects do seem to be observed under some circumstances, their appearance across a range of memory tasks is far from robust and, when they do occur, alternative theoretical accounts may be

possible. Thus, while further research is needed, temporal isolation effects do not at present offer unequivocal support for temporal distinctiveness models.

Wider Theoretical Implications

We have already discussed how the claims of temporal distinctiveness models relate to the classic debate concerning the causes of forgetting – Is information loss from memory due to trace decay or interference? A closely related question is whether it is necessary to assume a distinction between separate short-term memory and long-term memory systems.

Long-Term Memory and Short-Term Memory

Some temporal distinctiveness approaches to memory, and indeed other temporal contextual models such as the temporal context model (TCM; Howard and Kahana, 2002), emphasize the similarity of memory processes that occur over many different timescales, both the short-term and the long-term. This perspective contrasts with the conventional assumption of a distinction between short-term memory and long-term memory. Specifically, both the retrieved context models, such as TCM, and temporal ratio models, such as SIMPLE, deny the need to assume a separate short-term memory system.

Although the claim that an short-term memory/long-term memory distinction is not required is controversial (e.g., Davelaar et al., 2005), a number of key sources of evidence for the distinction between short-term memory and long-term memory can be captured by some temporal distinctiveness models. We illustrated above how an extended ratio-rule model can simultaneously account for the primacy and delay-sensitive recency effects that have been taken as a signature of a separate short-term memory stem. Neuropsychological data (e.g., from patients with hippocampal amnesia), which have been taken as another source of evidence for an short-term memory/long-term memory distinction, can also be accounted for by a unitary temporal distinctiveness model when patterns of rehearsal are taken into account (Brown et al., 2007a; Dewar et al., 2011). As shown above, temporal distinctiveness models demonstrate that forgetting due to the passage of time alone need not in itself reflect a trace decay process of the type that might be taken to motivate separate memory stores each with their own characteristic cause of forgetting. As we also noted, however, there is still debate as to whether forgetting occurs due to the passage of time alone when temporal distinctiveness is taken into account.

Moreover, claims that a unitary account of memory may be possible fit well with a large body of emerging evidence suggesting that the same variables influence short-term memory and long-term memory tasks when factors such as rehearsal and methodological differences are taken into account (e.g., Tan and Ward, 2000; Ward et al., 2010). In summary, although much more research is necessary, the argument against separate short-term memory and long-term memory systems remains an important claim associated with temporal distinctiveness models.

Consolidation

A final theoretical issue concerns the importance of *consolidation* in memory. According to temporal distinctiveness models of memory, the sole cause of forgetting is interference. How does this claim sit with recent research arguing for the importance of a consolidation process in memory and forgetting? According to consolidation theories, memory representations become stronger, in the sense of being more resistant to forgetting, over time. As noted by Wixted (2004), consolidation processes have been central to theories of memory and forgetting within neuroscience but have seen relatively little application in cognitive models of memory. One of the key direct sources of evidence for consolidation, at least over the short-term, is the observation that an unfilled period of time following the formation of a memory leads to reduced forgetting. According to consolidation theories, the unfilled temporal interval leads to a stronger memory because consolidation processes are uninterrupted. However, the same effect is predicted by temporal distinctiveness models, but for a different reason. Specifically, the addition of an unfilled temporal interval following an event that must be remembered will make the item more temporally isolated, and hence superior memory for that item may reflect not consolidation but temporal isolation. Although there are many data (e.g., on processes happening during sleep) that have not yet been addressed by temporal distinctiveness models of memory, it has been argued that temporal distinctiveness offers an alternative explanation, not involving consolidation, on one of the key behavioral pieces of evidence for consolidation (Brown and Lewandowsky, 2010; Lewandowsky et al., 2012). Thus arguments for consolidation processes must ensure that alternative explanations in terms of temporal distinctiveness can be excluded.

Summary

In summary, we have shown how temporal distinctiveness models assume that memory is temporally organized and that items that occupy nearby temporal locations will tend to be confusable with one another and hence more difficult to retrieve from memory. We have also suggested that temporal distinctive models offer a new theoretical perspective on the causes of forgetting, showing how the fact of forgetting due to the passage of time alone does not necessarily implicate trace decay processes. Along with retrieved temporal content models, temporal distinctiveness models typically (although not always) deny the need to assume a separate short-term memory system.

See also: Amnesia: General; Context Dependent Memory; Episodic Memory; Proactive Interference; Retroactive Interference; Short Term Memories, Theories of; Short-Term Memory: Psychological and Neural Aspects; Working Memory, Psychology of.

Bibliography

Altmann, E.M., Schunn, C.D., 2012. Decay versus interference: a new look at an old interaction. Psychological Science 23 (11), 1435–1437.

Anderson, J.R., Schooler, L.J., 1991. Reflections of the environment in memory. Psychological Science 2, 396–408.
Baddeley, A.D., 1976. The Psychology of Memory. Harper & Row, London.
Barrouillet, P., Bernardin, S., Camos, V., 2004. Time constraints and resource sharing in adults' working memory spans. Journal of Experimental Psychology: General 133 (1), 83–100.
Barrouillet, P., Camos, V., 2012. As time goes by: temporal constraints in working memory. Current Directions in Psychological Science 21 (6), 413–419.
Berman, M.G., Jonides, J., Lewis, R.L., 2009. In search of decay in verbal short-term memory. Journal of Experimental Psychology: Learning Memory and Cognition 35 (2), 317–333.
Bjork, R.A., Whitten, W.B., 1974. Recency-sensitive retrieval processes in long-term free recall. Cognitive Psychology 6, 173–189.
Bower, G.H., 1972. Stimulus-sampling theory of encoding variability. In: Melton, A.W., Martin, E. (Eds.), Coding Processes in Human Memory. Winston, Washington, DC, pp. 85–123.
Bright-Paul, A., Jarrold, C., 2012. Children's eyewitness memory: repeating post-event misinformation reduces the distinctiveness of a witnessed event. Memory 20 (8), 818–835.
Brown, G.D.A., Lewandowsky, S., 2010. Forgetting in memory models Arguments against trace decay and consolidation failure. In: Della Sala, S. (Ed.), Forgetting, pp. 49–75.
Brown, G.D.A., Della Sala, S., Foster, J.K., Vousden, J.I., 2007a. Amnesia, rehearsal, and temporal distinctiveness models of recall. Psychonomic Bulletin & Review 14 (2), 256–260.
Brown, G.D.A., Neath, I., Chater, N., 2007b. A temporal ratio model of memory. Psychological Review 114 (3), 539–576.
Brown, G.D.A., Preece, T., Hulme, C., 2000. Oscillator-based memory for serial order. Psychological Review 107 (1), 127–181.
Brown, G.D.A., Vousden, J.I., McCormack, T., 2009. Memory retrieval as temporal discrimination. Journal of Memory and Language 60 (1), 194–208.
Campoy, G., 2012. Evidence for decay in verbal short-term memory: a commentary on Berman, Jonides, and Lewis (2009). Journal of Experimental Psychology: Learning Memory and Cognition 38 (4), 1129–1136.
Church, R.M., Broadbent, H.A., 1990. Alternative representations of time, number, and rate. Cognition 37 (1–2), 55–81.
Clayton, N.S., Dickinson, A., 1988. Episodic-like memory during cache recovery by scrub jays. Nature 395, 272–274.
Crowder, R.G., 1976. Principles of Learning and Memory. Erlbaum, Hillsdale, NJ.
Davelaar, E.J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H.J., Usher, M., 2005. The demise of short-term memory revisited: empirical and computational investigations of recency effects. Psychological Review 112 (1), 3–42.
Dewar, M., Brown, G.D.A., Della Sala, S., 2011. Restoring primacy in amnesic free recall: evidence for the recency theory of primacy. Cognitive Neuropsychology 28 (6), 386–396.
Farrell, S., 2012. Temporal clustering and sequencing in short-term memory and episodic memory. Psychological Review 119 (2), 223–271.
Farrell, S., Wise, V., Lelievre, A., 2011. Relations between timing, position, and grouping in short-term memory. Memory & Cognition 39 (4), 573–587.
Gallistel, C.R., 1990. The Organization of Learning. The MIT Press, Cambridge, MA, US.
Geiger, S.M., Lewandowsky, S., 2008. Temporal isolation does not facilitate forward serial recall – or does it? Memory & Cognition 36 (5), 957–967.
Glenberg, A.M., Bradley, M.M., Kraus, T.A., Renzaglia, G.J., 1983. Studies of the long-term recency effect: support for a contextually guided retrieval hypothesis. Journal of Experimental Psychology: Learning Memory and Cognition 9 (2), 231–255.
Glenberg, A.M., Swanson, N.G., 1986. A temporal distinctiveness theory of recency and modality effects. Journal of Experimental Psychology: Learning, Memory, & Cognition 12 (1), 3–15.
Guérard, K., Neath, I., Surprenant, A., Tremblay, S., 2010. Distinctiveness in serial memory for spatial information. Memory & Cognition 38 (1), 83–91.
Hintzman, D.L., 2004. Time versus items in judgment of recency. Memory & Cognition 32 (8), 1298–1304.
Hintzman, D.L., Block, R.A., Summers, J.J., 1973. Contextual associations and memory for serial position. Journal of Experimental Psychology 97, 220–229.
Houghton, G., 1996. The problem of serial order: a neural network model of sequence learning and recall. In: Dale, R., Mellish, C., Zock, M. (Eds.), Current Research in Natural Language Generation. Academic Press, San Diego, CA, pp. 287–319.
Howard, M.W., Kahana, M.J., 2002. A distributed representation of temporal context. Journal of Mathematical Psychology 46 (3), 269–299.
Lewandowsky, S., Duncan, M., Brown, G.D.A., 2004. Time does not cause forgetting in short-term serial recall. Psychonomic Bulletin & Review 11 (5), 771–790.
Lewandowsky, S., Ecker, U.K.H., Farrell, S., Brown, G.D.A., 2012. Models of cognition and constraints from neuroscience: a case study involving consolidation. Australian Journal of Psychology 64 (1), 37–45.
Lewandowsky, S., Nimmo, L.M., Brown, G.D.A., 2008. When temporal isolation benefits memory for serial order. Journal of Memory and Language 58 (2), 415–428.
Lewandowsky, S., Oberauer, K., Brown, G.D.A., 2009. No temporal decay in verbal short-term memory. Trends in Cognitive Sciences 13, 120–126.
McKeown, D., Mercer, T., 2012. Short-term forgetting without interference. Journal of Experimental Psychology: Learning Memory and Cognition 38 (4), 1057–1068.
Mensink, G.J., Raaijmakers, J.G.W., 1988. A model for interference and forgetting. Psychological Review 95 (4), 434–455.
Mercer, T., McKeown, D., 2014. Decay uncovered in nonverbal short-term memory. Psychonomic Bulletin & Review 21, 128–135.
Morin, C., Brown, G., Lewandowsky, S., 2010. Temporal isolation effects in recognition and serial recall. Memory & Cognition 38 (7), 849–859.
Murdock, B.B., 1960. The distinctiveness of stimuli. Psychological Review 67 (1), 16–31.
Neath, I., 1993. Distinctiveness and serial position effects in recognition. Memory & Cognition 21 (5), 689–698.
Neath, I., Brown, G.D.A., 2006. Simple: further applications of a local distinctiveness model of memory. In: Psychology of Learning and Motivation: Advances in Research and Theory, vol. 46, pp. 201–243.
Neath, I., Brown, G.D.A., McCormack, T., Chater, N., Freeman, R., 2006. Distinctiveness models of memory and absolute identification: evidence for local, not global, effects. Quarterly Journal of Experimental Psychology 59 (1), 121–135.
Neath, I., Crowder, R.G., 1990. Schedules of presentation and temporal distinctiveness in human memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 16 (2), 316–327.
Page, M.P.A., Norris, D., 1998. The primacy model: a new model of immediate serial recall. Psychological Review 105 (4), 761–781.
Parmentier, F.B.R., King, S., Dennis, I., 2006. Local temporal distinctiveness does not benefit auditory verbal and spatial serial recall. Psychonomic Bulletin & Review 13 (3), 458–465.
Polyn, S.M., Norman, K.A., Kahana, M.J., 2009. A context maintenance and retrieval model of organizational processes in free recall. Psychological Review 116 (1), 129–156.
Ricker, T.J., Cowan, N., 2010. Loss of visual working memory within seconds: the combined use of refreshable and non-refreshable features. Journal of Experimental Psychology: Learning Memory and Cognition 36 (6), 1355–1368.
Sederberg, P.B., Howard, M.W., Kahana, M.J., 2008. A context-based theory of recency and contiguity in free recall. Psychological Review 115 (4), 893–912.
Shipstead, Z., Engle, R.W., 2013. Interference within the focus of attention: working memory tasks reflect more than temporary maintenance. Journal of Experimental Psychology: Learning Memory and Cognition 39 (1), 277–289.
Tan, L., Ward, G., 2000. A recency-based account of the primacy effect in free recall. Journal of Experimental Psychology: Learning Memory and Cognition 26 (6), 1589–1625.
Tulving, E., 1972. Episodic and semantic memory. In: Tulving, E., Donaldson, W. (Eds.), Organization of Memory. Academic Press, New York, pp. 381–403.
Unsworth, N., Engle, R.W., 2006. A temporal-contextual retrieval account of complex span: an analysis of errors. Journal of Memory and Language 54 (3), 346–362.
Ward, G., Tan, L., Grenfell-Essam, R., 2010. Examining the relationship between free recall and immediate serial recall: the effects of list length and output order. Journal of Experimental Psychology: Learning Memory and Cognition 36 (5), 1207–1241.
Wixted, J.T., 2004. The psychology and neuroscience of forgetting. Annual Review of Psychology 55, 235–269.

Temporal Lobe

Bryan Kolb, University of Lethbridge, Lethbridge, AB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by K. Tanaka, volume 23, pp. 15595–15599, © 2001, Elsevier Ltd.

Abstract

The temporal lobe is one of the four major divisions of the cerebral cortex. The function of the temporal lobe is to transform visual, auditory, and gustatory sensory input into higher-order perceptions and memory. Temporal lobe injury leads to a complex set of symptoms of both basic perceptual processes as well as higher-order processes such as memory and emotion.

Introduction

The cerebral hemispheres can be grossly divided into four lobes (frontal, parietal, occipital, and temporal) that are largely defined by the cranial bone lying above them. The temporal lobes comprise all the tissue that lies below the lateral (Sylvian) fissure and anterior to the occipital lobe, including the limbic cortex on the medial surface of the temporal lobe (Figure 1). The temporal lobe also includes subcortical structures, including the amygdala and hippocampus and their associated cortical regions. This is a large volume of brain tissue that has multiple functions, including the processing of auditory, visual, and gustatory sensations, visuospatial information, aspects of emotion, and certain forms of memory.

Temporal Lobe Anatomy

The human temporal regions can be divided on the lateral surface into those that are auditory (Brodmann's areas 41, 42, and 22 in Figure 1(b)) and those that form the ventral visual stream on the lateral temporal lobe (areas 20, 21, 37, and 38 in Figure 1(b)). The visual regions are often referred to as *inferotemporal cortex* or by von Economo's designation, TE.

The temporal sulci enfold a lot of cortex. In particular, the lateral (Sylvian) fissure contains tissue forming the *insula*, which includes the gustatory cortex as well as the auditory cortex.

The superior temporal sulcus (STS) separates the superior and middle temporal gyri and houses a significant amount of neocortex as well, as illustrated for the rhesus monkey in Figure 2, including many subregions of multimodal cortex that receive input from auditory, visual, and somatic regions as well as from prefrontal and posterior parietal regions.

The medial temporal region includes the amygdala and adjacent cortex (uncus), the hippocampus and surrounding cortex (subiculum, entorhinal cortex, perirhinal cortex), and the fusiform gyrus (*see* Hippocampus and Related Structures).

The cortical region lying along the boundary of the temporal and parietal lobes is often called the *temporal–parietal junction* (*TPJ*). This abstract label refers roughly to the region at the end of the Sylvian fissure, including the ventral regions of the angular and supramarginal gyri (*see* Parietal Lobe) and adjacent temporal cortex. The TPJ is consistently shown to be

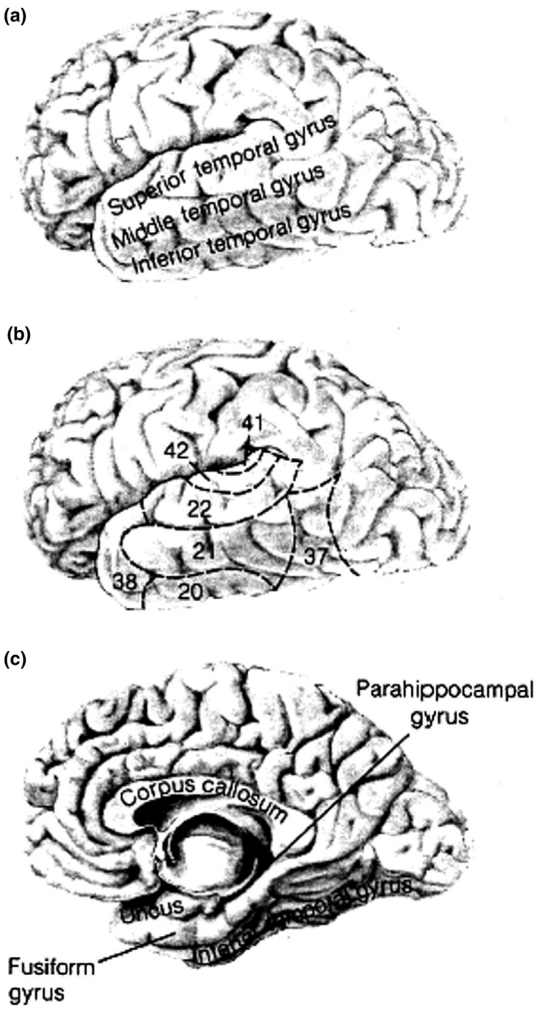

Figure 1 Gross anatomy of the temporal lobe. (a) Three major temporal lobe gyri are visible on its lateral surface. (b) Brodmann's areas on the lateral surface, where auditory areas are shown in yellow and visual areas in purple. Areas 20, 21, 37, and 38, the inferotemporal cortex, are often referred to as TE. (c) The temporal lobe gyri visible in a medial view. The uncus is the anterior extension of the hippocampal formation. Reproduced from Kolb, B., Whishaw, I.Q., 2015. Fundamentals of Human Neuropsychology, seventh ed. Worth, New York.

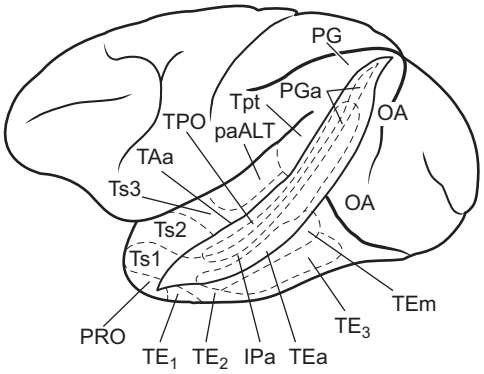

Figure 2 Lateral view of the rhesus monkey brain with the Superior Temporal Sulcus opened up to reveal multiple subareas as identified by Seltzer, B., Pandya, D., 1978. Afferent connections and architectonics of the superior temporal sulcus and surrounding cortex in rhesus monkey. Brain Research 149, 1–24. Adapted from Kolb, B., Whishaw, I.Q., 2009. Fundamentals of Human Neuropsychology, third ed. Freeman, New York.

active in neuroimaging studies investigating attention, memory, language, and social processing. Thus, the TPJ is proposed as central to decision-making in a social context.

Functions of the Temporal Lobe

We can identify five general functions of the temporal lobe: (1) processing of auditory input, (2) processing of visual input, (3) processing of gustatory input, (4) long-term storage of sensory input (memory), and (5) emotional regulation.

Processing of Auditory Input

Sound waves reaching the ear stimulate a cascade of mechanical and neural events – in the cochlea, brainstem, and, eventually, the auditory cortex – that result in a percept of sound. The auditory cortex has multiple regions, each having a tonotopic map. The precise functions of these maps are poorly understood, but the ultimate goal is to perceive sound-making objects, locate sound, and make movements in relation to sound.

Cells in the primary auditory cortex of mammals (Brodmann's area 41, **Figure 1**) respond to distinct frequencies of sound, whereas cells in higher auditory cortical regions respond more to combinations of frequencies. In humans, there is another type of distinction as well. Specifically, two fundamental characteristics of sound are its frequency (spectral feature of sound) and timing (temporal feature). Thus, sounds vary in their frequencies as well as in how quickly they change. In a seminal paper, Zatorre and Belin (2001) demonstrated using positron emission tomography imaging that the primary auditory cortex in both hemispheres responds to temporal features and the higher areas responded to spectral features, but the higher auditory regions showed an asymmetry: the left was weighted toward temporal processing and the right toward spectral processing. There is considerable evidence that the analysis of speech requires rapid auditory processing (i.e., temporal features), whereas analysis of tonal information, such as music, requires the identification of changes in pitch (i.e., spectral features).

Both speech and musical perception require more than just moment-to-moment analysis of temporal and spectral features, but there is also a memory component referred to as echoic memory. Echoic memory is crucial for auditory perception, because the meaning of any single sound depends upon the following sounds. Although the precise length of echoic memory is variable, it generally lasts about 3–4 s though some authors have shown it to last up to 20 s. Echoic memory appears to involve an extended network of cortical regions, but a key region is the superior temporal gyrus (e.g., Alain et al., 1998).

Processing of Visual Input

The occipital lobe sends a major projection to multiple regions in the temporal cortex both in the middle and ventral temporal gyri (inferotemporal areas labeled TE in **Figure 2**), as well as the anterior tip of the superior temporal gyrus and the STS. The posterior parietal lobe sends connections to the medial temporal regions (hippocampus and parahippocampal regions).

Neurons in the inferotemporal cortex are tuned for the features of objects and faces and many change their pattern of firing with experience, including how often a visual stimulus has been encountered and any association with reward. These properties make the ventral stream more tuned to the enduring characteristics of visual stimuli than with the moment-to-moment changes in the visual array, which provides a basis for object recognition.

Another characteristic of inferotemporal neurons is that they provide a mechanism for the internal representation of images of objects. Fuster and Jervey (1982) first demonstrated that, if monkeys are shown specific objects that are to be remembered, neurons in the monkey cortex continue to discharge during the 'memory' period. These selective neuronal discharges may provide the basis of working memory for the visual stimuli (sometimes referred to as iconic memory). Furthermore, these discharges may provide the basis for visual imagery. That is, the discharge of groups of neurons selective for characteristics of particular objects may provide mental images of those objects in their absence.

Regions of the ventral stream have surprisingly selective tuning to specific stimulus characteristics such as specific body parts (hands, bodies, faces), color, or scenes. One example is a region in the fusiform gyrus, the fusiform face area, which is selectively activated by faces (e.g., Kanwisher, 2010). Another is a more medial region in the hippocampal gyrus, the parahippocampal place area, which is preferentially active for scenes including buildings and places. This region is hypothesized to play a significant role in visuospatial navigation, reflected both in the specificity of activation to scenes as well as the input from the posterior parietal regions involved in directing movements in space (*see* Parietal Lobe; Cortical Areas Engaged in Movement: Neuroimaging Methods).

A different type of visual analysis is performed by the STS. The STS receives multimodal inputs that play a role in categorizing stimuli. A major category is social perception, which includes analyzing actual or implied body movements (biological motion) that provide socially relevant information. This information plays an important role in social cognition, a 'theory of mind' that allows us to develop hypotheses about other people's intentions. For example, the direction of a person's gaze provides us considerable information about what that person is – or is not – attending to. Many cells in the STS are especially tuned to faces, including the precise orientation (e.g., head-on, profile) as well as specific facial expressions. Some cells in the STS are maximally sensitive to primate bodies moving in a particular direction, which is remarkable because the basic configuration of the stimulus is identical as the body moves in different directions; only the direction changes (see review by Barraclough and Perrett, 2011).

Processing of Gustatory Input

The primary taste cortex is found in the insula (*see* Taste and Smell, Psychology of). The insula contains neurons tuned to the five basic tastes (sweet, salt, bitter, sour, and unami) as well as other neurons that encode oral somatosensory stimuli (viscosity, fat texture, temperature, and capsacin). Thus, our perception of taste is not just the basic tastes but also an interaction of the taste and texture of food. Human imaging studies using MRI have shown that insular activity correlates with the subjective intensity of tastes.

The insula projects to the orbitofrontal cortex, which is likely the secondary taste area. Activation of the orbitofrontal cortex correlates with the subjective pleasantness of taste rather than the intensity (*see* Taste and Smell, Psychology of).

Memory

Interest in the temporal lobes' function in memory was stimulated in the early 1950s by the discovery that bilateral removal of the medial temporal lobes, including the hippocampus and amygdala, results in amnesia for all events after the surgery (anterograde amnesia). It is now clear that both the medial temporal regions play a central role in the extended neural circuitry underlying explicit memory. The key structures are the hippocampus, entorhinal cortex, perirhinal cortex, and parahippocampal cortex (Petri and Mishkin, 1994). Although it is not possible to distinguish the roles of these regions in humans, studies in monkeys are providing some insights. The perirhinal cortex appears to support memory for objects (including factual knowledge about the world), whereas of the hippocampus and parahippocampal regions, the hippocampal regions are central to the circuits underlying contextual knowledge (autobiographic, or episodic, knowledge).

The amygdala also plays a role in memory, but is involved in a circuit underlying emotional memory. The amygdala has close connections with other medial temporal structures as well as the frontal lobe. It also sends projections to the brainstem structures that control autonomic responses such as blood pressure and heart rate, the hypothalamus that controls hormonal systems, and the periaqueductal gray matter that affects the perception of pain.

Emotional memory has been studied most thoroughly in fear conditioning (*see* Fear Conditioning: Overview). The amygdala seems to evoke our feelings of fear/anxiety toward stimuli that are threatening, but also can do the same toward stimuli that by themselves would not normally produce fear, such as is observed in anxiety disorders.

Emotional Regulation

An interconnected network of structures including the neocortex, thalamus, hippocampal formation, and amygdala forms the basis of emotional experience (*see* Emotion, Neural Basis of). The amygdala receives input from all of the sensory systems. The amygdala can therefore create a complex image of the sensory world. To be excited, cells in the amygdala require complex stimuli such as faces. Thus, like the STS, the amygdala plays a role in identifying facial expressions. As noted earlier, the amygdala also interacts with brainstem structures that control autonomic responses and thus can link external stimuli with emotional feelings.

Symptoms of Temporal Lobe Injury

Nine principal symptoms are associated with disease of the temporal lobes: (1) disturbance of auditory sensation and perception, (2) disorders of music perception, (3) disorders of visual perception, (4) disturbance in the selection of visual and auditory input, (5) impaired organization and categorization of sensory input, (6) inability to use contextual information, (7) impaired long-term memory, (8) altered personality and affective behavior, and (9) altered sexual behavior (Kolb and Whishaw, 2015).

In contrast to damage in the primary visual or somatic regions, bilateral damage to the primary auditory cortex does not produce a complex loss of sensation (auditory deafness). Rather, damage to auditory cortex impairs the processing of speech sounds (left hemisphere) or in making pitch discriminations (right hemisphere). These deficits are consistent with the observation of aphasia in patients with large injuries of the left temporal regions including both the primary and secondary auditory regions, including Wernicke's area. In contrast, damage to the superior temporal gyrus on the right disrupts the processing of musical sounds, the exact deficit depending upon the site of injury. For example, rhythm discrimination is most affected by right posterior superior temporal gyrus damage, whereas meter discrimination (for example, distinguishing a waltz and a march) is more affected by anterior damage to either temporal lobe.

People with inferotemporal cortex injuries do not have field defects such as those with occipital lesions, but they do have deficits in the perception of complex scenes, reflecting the role of the inferotemporal cortex in the ventral stream of visual processing. Recall that neurons in TE show selective tuning to specific object characteristics so it is not surprising that temporal lesions would interfere with perceiving these characteristics and may produce specific visual agnosias.

Persons with temporal lobe damage frequently report difficulties in attention to auditory or visual information when there are multiple competing inputs. For example, if two

people are talking at once we usually can shift from listening to one and then the other. Temporal lobe patients have great difficulty in doing this and as a result get little from either speaker. Similarly, one strategy that we unconsciously use for understanding visual and auditory material is to categorize it. When asked to learn a list of words such as 'dog, car, bus, apple, rat, lemon, cat, truck, orange,' most of us will organize the words into three different categories – animals, vehicles, and fruit. If the list is later recalled, the items are likely to be recalled by category, and recall of the categories is likely to be used as an aid in recalling the items. Patients with temporal lobe injuries do not do this, leading to less efficient perception and memory of the information.

Context plays an important role in normal perception and recall of sensory information. Seeing a familiar object or person in a novel context can often give us pause to identification. For example, encountering a clerk from your neighborhood store on a bus in Stockholm may provoke a feeling of familiarity but because the context is wrong, identification of the person is likely to be slow. It is clear that we use normally context as a cue in object identification. Patients with temporal lobe injuries are impaired at this ability, giving rise to both perceptual and memory difficulties.

But memory deficits are larger than just those related to context. Temporal lobe injuries, especially those involving medial temporal lobe structures, can result in amnesia both for events preceding the injury (retrograde amnesia) as well as for events following the injury (anterograde amnesia) (*see* Amnesia: General; Declarative Memory, Neural Basis of).

Finally, temporal lobe injury, especially when the amygdala is damaged, can produce profound changes in personality. Bilateral amygdala damage is associated with a loss of emotion, and especially fear. The general clinical impression is that temporal lobe patients have a clear personality change, sometimes referred to as a temporal lobe personality (e.g., Bear and Fedio, 1977). Symptoms include increased emotionality, depression, obsessionalism, religiosity, paranoia, and a reduced sense of humor.

Conclusion

In summary, the temporal lobe plays a central role in the higher-order analysis of auditory and visual information, and especially in context. As a result, temporal lobe injury leads to a complex set of symptoms of both basic perceptual processes as well as higher-order processes such as memory and emotion.

See also: Amnesia: General; Attention, Neural Basis of; Autobiographical Memory; Cortical Areas Engaged in Movement: Neuroimaging Methods; Declarative Memory, Neural Basis of; Emotion, Neural Basis of; Fear Conditioning: Overview; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Parietal Lobe; Spatial Navigation; Taste and Smell, Psychology of; Visual Streams: Dorsal and Ventral.

Bibliography

Alain, C., Woods, D.L., Knight, R.T., 1998. A distributed cortical network for auditory sensory memory in humans. Brain Research 812, 23–37.

Barraclough, N.E., Perrett, D.I., 2011. From single cells to social perception. Philosophical Transactions of the Royal Society B 366, 1739–1752.

Bear, D.M., Fedio, P., 1977. Quantitative analysis of interictal behavior in temporal lobe epilepsy. Archives of Neurology 34, 454–467.

Fuster, J.M., Jervey, J.P., 1982. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212, 952–955.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America 107, 11163–11170.

Kolb, B., Whishaw, I.Q., 2009. Fundamentals of Human Neuropsychology, third ed. Freeman, New York.

Kolb, B., Whishaw, I.Q., 2015. Fundamentals of Human Neuropsychology, seventh ed. Worth, New York.

Petri, H.L., Mishkin, M., 1994. Behaviorism, cognitivism, and the neuropsychology of memory. American Scientist 82, 30–37.

Seltzer, B., Pandya, D., 1978. Afferent connections and architectonics of the superior temporal sulcus and surrounding cortex in rhesus monkey. Brain Research 149, 1–24.

Zatorre, R.J., Belin, P., 2001. Spectral and temporal processing in human auditory cortex. Cerebral Cortex 11, 946–953.

Tense, Aspect, Mood and Evidentiality, Linguistics of

Östen Dahl, Stockholm University, Stockholm, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Together tense, aspect, mood, and evidentiality, or TAME for short, play a central role in the grammars of most languages in the world. They are typically marked inflectionally on verbs but can also be expressed by derivational morphology and by periphrastic constructions. Although tense, aspect, and mood, and sometimes also evidentiality are traditionally treated as separate grammatical categories, they are intertwined in TAME systems in ways that make it difficult to uphold such a treatment.

Introduction

The English sentences *He is asleep* and *He was asleep* differ in tense: the former is in the present and the latter in the past tense. In a similar way, *She opened the window* and *She was opening the window* are said to differ in aspect, and the verbs *wish* and *were* in the sentence *I wish I were dead* are said to differ in mood. In the grammatical tradition inherited from Latin grammar, verbs are said to possess the inflectional categories of tense and mood, and aspect was added to them later on, partly due to influence from the study of Slavic languages. In recent years, it has been proposed that evidentiality, defined as the grammatical marking of information source, should be recognized as a grammatical category distinct from tense, aspect, and mood. Together tense, aspect, mood, and evidentiality, or TAME for short, play a central role in the grammars of most languages in the world. It should be noted from the outset, however, that tense, aspect, mood, and evidentiality do not usually come neatly lined up as separate categories in grammars. Rather, the meanings of TAME forms often combine elements from more than one of them. Conversely, distinctions that would traditionally be thought of as belonging to the same category may be expressed in very different ways, for instance, by inflectional forms on the one hand and auxiliary constructions on the other. For this reason, many linguists prefer to see TAME as one large domain within which various grammatical distinctions may arise.

In traditional grammar, tenses and moods are inflections on the main verb of a sentence. Inflectional TAME can indeed be seen as the central and most typical case, but TAME can also be realized periphrastically, that is, by a construction involving more than one word, the grammatical markers being either auxiliaries or indeclinable particles.

General works on TAME include Comrie (1976, 1985) and Binnick (1991, 2011). Two early works that have had a particularly strong influence on the study of TAME, although their authors were philosophers rather than linguists, are those of Reichenbach (1947) and Vendler (1957). Important monographs are also those of Klein (1994) and Smith (1991).

TAME and Finiteness

TAME is intimately linked up with the notion of finiteness in grammar. Finiteness is a property that characterizes main clauses such as *She opened the window* and subordinate clauses such as (*I saw that*) *she opened the window* as opposed to other embedded, nonfinite constructions such as (*I saw*) *her open the window* or (*It is forbidden*) *to open the window*. In English and many other languages, TAME is fully realized only in finite clauses, while it is restricted or absent in nonfinite clauses. The term 'tensed clause' has sometimes been used in the same sense as 'finite clause.' For a collection of articles on finiteness, see Nikolaeva (2007).

The Typology and Diachrony of TAME

Typologically oriented works by Dahl (1985) and Bybee et al. (1994) have shown that although there is great variation among languages as to their TAME systems, the elements of those systems may be reduced to a relatively small number of types, each with its characteristic meaning and typical way of expression, developing historically by processes of grammaticalization. Such processes usually start with phrasal constructions involving lexical morphemes. Future markers may, e.g., develop out of constructions involving motion verbs such as 'come' and 'go,' or modal verbs with the meanings of obligation ('shall, must') and volition ('want'). Later on, inflections may arise through the coalescence of earlier free morphemes (auxiliaries or particles) with the stem of the main verb. The elements of a TAME system in a language may be at different stages in these grammaticalization processes; how far they have advanced determines both their form and their meaning. Thus, only a few of the types described below typically show up in inflectional distinctions, while the others tend to be expressed periphrastically.

Another finding of recent typologically oriented work is that there are some fairly strong areal tendencies in the makeup of TAME systems (Dahl and Velupillai 2011a,b,c,d,e; de Haan 2011a,b). Most notably perhaps, the distinction between past and nonpast is represented in virtually all languages in a large area stretching from Iceland in the northwest to the Horn of Africa in the south and Bangladesh in the southeast, including much of Central Eurasia (but excluding northeastern Siberia). It includes not only Indo-European languages but also languages from the Uralic, Altaic, and Afro-Asiatic families and some language isolates.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.52025-X

Tense

The basis for the distinctions between forms traditionally subsumed under the category of tense is the time dependence that characterizes the overwhelming majority of statements in natural languages. A sentence such as The *baby was asleep* does not make very much sense as long as we do not know what time we are talking about – that is, at what time the baby was asleep. In English, the tense of the verb gives us some indication about the time – we know that the speaker is referring to some point in the past. However, the exact location of that point is not indicated explicitly and has to be retrieved from the context. In many languages, e.g., Chinese, no distinction would be made between *was asleep* and *is asleep* here – in those languages, the identification of the 'topic time' or 'reference time' is left wholly implicit. Nevertheless, the identification has to be made, in order to make sense of the statement. Thus, the way in which sentences are interpreted is similar whether or not a language makes the grammatical distinction between present and past tense. Note also that tenses are not complementary to time adverbials, as is often assumed. Languages that lack tense distinctions do not, on the whole, use time adverbials more often than languages that do not. This is because time adverbials are used whenever the contextual temporal information is insufficient. Tenses, on the other hand, being more or less obligatory grammatical elements, are used whether or not the information they contain is necessary for the communication.

In the most straightforward cases, the choice of tense depends on the relationship between the topic time and the time of speech. Time as experienced by a person is naturally divided into present, past, and future; these are also the three basic tenses of traditional grammar, although it is common for languages to have only a simpler, two-way distinction in their grammar: either between past and nonpast or between future and nonfuture. In English, only the past tense (*he called*) differs from the present (*he calls*) morphologically; what is traditionally referred to as the future (*he will call*) is an auxiliary construction, although inflectional futures are far from uncommon in other languages (e.g., French *il dormira* 'he will sleep'). It is common for languages to have several ways of marking the future grammatically, often differing in subtle modal nuances (cf English *will* vs *is going to*). The common modal coloring of future tenses has led many scholars to deny the future status as a tense, but it is clear that temporal considerations are also part of the semantics of future tenses, and the mixture of temporal and modal criteria should rather be seen as typical of TAME in general.

Many languages distinguish grammatically degrees of remoteness in the past and sometimes also in the future. This means that different grammatical forms are used for recent and remote events. The distinctions may be more objective, that is, depend on units such as days or (less often) years, or more subjective, that is, depend on the speaker's assessment of an event as recent and remote. Commonly, objective and subjective criteria are combined in one and the same system. A typical system of remoteness distinctions in the past could include a 'hodiernal' past, for events earlier today, a 'hesternal' past, for events yesterday and possibly a few days back, and a 'remote' past, for events further back in time. Remoteness distinctions are found in TAME systems all over the world, but are most common in certain areas, notably sub-Saharan Africa, New Guinea, and Amazonia.

A distinction is customarily made between absolute and relative tense, depending on whether the point of departure is the time of speech or some other time point. The tenses of finite verbs in main clauses are usually absolute – past tense means that the topic time precedes the time of speech. Nonfinite verb forms such as participles, gerunds, and infinitives sometimes also show tense distinctions, but these predominantly mark relations of simultaneity or temporal sequence relative to the time of the next higher finite verb, as in the sentence *She lived in London, working for a large company* where the gerundial clause is understood as being about the same time as the main clause (which, in its turn, is understood as preceding the point of speech). Finite verbs in subordinate clauses may follow varying principles. In clauses expressing reported speech, the verb is in many languages in the same tense as it would have in a direct quote; in English and some other European languages, the tense of such clauses depends on the tense of the matrix main clause. This phenomenon, exemplified in a sentence such as *John said that he was ill* (cf *John said "I am ill"*), is traditionally referred to as 'sequence of tenses'; there is disagreement among scholars about its nature.

In some languages, special narrative tenses are used to express an event that takes place directly after another.

Aspect

Whereas the notions underlying the distinctions subsumed under tense are relatively straightforward, aspect is a more elusive notion. It is natural to see the sentences we utter as describing or evoking entities such as events, processes, and states. There is no word in everyday English that covers all these, but scholars use terms such as 'situations' and 'eventualities' when they generalize over them. Clearly, verbs and verb phrases (let us call them 'predicates' for short) differ in what kinds of entities they most naturally evoke. For instance, *Mary woke up* is about an event, whereas *Mary slept* is rather a state. Several ways of classifying predicates according to such criteria have been proposed. One such classification would distinguish stative predications from dynamic ones depending on whether they involve some kind of change or not. Thus, *sleep* would be stative, whereas *wake up* would be dynamic. Among dynamic predications one can distinguish telic and atelic predications, where telicity involves the presence of a boundary or the attainment of a specific result-state: *bake* would be atelic and *bake a cake* telic (since it involves the coming into existence of a cake). Telic predications can then be classified into punctual and nonpunctual or durative depending on whether the situations/eventualities are seen as having duration or not.

According to a popular view, grammatical aspect involves the choice of a perspective or viewpoint on a situation/eventuality, but an alternative analysis sees aspects as marking the conversion of a predicate from one type to another. The relationship between situation types and aspects is not easy to

disentangle, one reason being that there is seldom a one-to-one relationship between verbs as lexical items and situation types.

In languages that express aspectual distinctions in their inflectional systems, the most common distinction is between perfective and imperfective aspect. According to a common view, perfective aspect views the situation/eventuality in its totality. Imperfective aspect is used for what is going on at the topic time but also for habitually occurring events.

There is a connection between telicity and perfectivity in that telic predications are most naturally given a totality interpretation. But there is a further connection to temporal reference.

Referring to an event in its totality is usually only possible if it has been completed at the time of speech. As a consequence of this, perfective aspect tends to be restricted to past time reference, or at least to nonpresent time reference. Furthermore, telic predications that are not overtly marked for tense or aspect are interpreted as referring to the past in many languages.

These connections between temporal and aspectual notions lie behind the skewings that are typically found in TAME systems, where distinctions between perfective and imperfective aspects are only upheld in forms with past time reference. An example of such a skewed system is found in written French, in which perfective aspect is represented by the so-called passé simple or simple past (*il chanta* 'he sang'), while the present (*il chante* 'he sings/is singing'), the imparfait or imperfective past (*il chantait* 'he sang/was singing'), and the future (*il chantera* 'he will sing') represent combinations of the imperfective aspect with various tenses.

Aspect may also be expressed by periphrastic or derivational means. Progressives, as exemplified by the English construction *I am/was/will be singing*, are a common type in languages and tend to be expressed periphrastically. With regard to their meaning, they resemble imperfective aspect but have a more restricted domain of use in that they are only marginally used about states or habitual events. Historically, progressives are a common source of imperfectives.

Special habitual inflections or periphrastic habitual constructions are not uncommon; these may be used for events that take place regularly (*I used to visit her on Sundays*) but sometimes also for generic statements such as *Dogs bark*.

In many languages, the aspectual character of verbs may be modified by derivational processes. In contradistinction to inflectional aspect, these processes tend to be idiosyncratic and restricted to smaller sets of lexical items. They also more clearly involve conversion from one type of situation/eventuality to another. A typical distinction expressed by derivational aspect is that between semelfactive or punctual, on one hand, and durative on the other, as in Hungarian *köhintett* 'he coughed once': *köhögött* 'he coughed (for a longer or shorter time).' The famous aspectual systems of Russian and other Slavic languages are derivational in origin but have over time been regularized and received many features that make them seem more like inflectional aspect.

The Perfect

The English perfect, as in *He has sung*, represents a cross-linguistically common type the semantics of which defies the traditional classification in that it cannot be unambiguously subsumed under either tense or aspect. Perfects are used of events that took place in the past but have a more restricted domain of use than ordinary past tenses. A common way of formulating the meaning of perfects is to say that they imply 'current relevance' of the event related in the sentence. Another way of putting this is to make a distinction between the 'time of event' and the 'topic time' (or reference time): in the perfect, the time of event is in the past but the topic time is equal to the time of speech. In the pluperfect (e.g., *He had sung when I arrived*) the time of event also precedes the topic time but here both points are in the past. In a future perfect, the topic time is in the future and the time of event precedes it, as in *He will have sung when I arrive*.

Within the semantic domain typically covered by perfects, we also find more specialized forms and constructions. Resultatives express the resulting state of some event (e.g., *He is gone* in English); they are usually restricted to verbs that denote change. Historically, they are often extended in meaning to become perfects. Experientials, such as the Japanese *koto ga aru* construction (*Nihon de hataraita koto ga aru* "I have worked in Japan") are used to express that some type of event took place one or more times within a period in the past.

Mood

There is some terminological confusion about the terms 'mood' and 'modality,' but commonly, the latter is taken to refer to a broad semantic domain having to do with the relationship between a proposition and reality, involving e.g., notions such as necessity, possibility, and probability. In English, modality is typically expressed by modal auxiliaries such as *can*, *may*, and *must*, or adverbs such as *perhaps*. Mood, on the other hand, is traditionally a narrower notion, referring to grammatical distinctions, which, like TAME in general, tend to be expressed by inflections of verbs. Moods are commonly said to express the speaker's attitude to a proposition or to its truth-value. However, more often than not, the speaker is not free to choose between moods, but the choice is governed by the type of clause or by the syntactic environment. A typical grammatical mood, such as the subjunctive in English, French, or German, is thus used to mark certain nonasserted and grammatically subordinate clause types. Thus, in French, the subjunctive is obligatory in the complement clauses of verbs meaning 'want, desire, hope.' The subjunctive thus marks these clauses off against plain main clauses, which are the unmarked, indicative mood: *Il est parti* 'He has left' with the indicative versus *J'espère qu'il soit parti* "I hope that he has left," with the subjunctive. In German, both these examples would be translated with the indicative. On the other hand, the subjunctive is used in German in clauses expressing indirect speech, which is not the case in French. Thus, nonindicative moods vary cross-linguistically as to their specific functions, although they share the property that they occur in clause types that are typically not used to make an assertion. Furthermore, a language may have more than one nonindicative mood. Thus, Classical Greek had both subjunctive and optative moods with different nonassertive uses.

Another common function of nonindicative moods is to distinguish between possible and counterfactual situations in conditional constructions (cf *If he is ready, we can go* vs *If he*

were ready, we could go, where the latter sentence contains the obsolescent subjunctive form *were*). A form usually subsumed under mood is the imperative, whose primary use is to express speech acts such as commands, requests, etc. However, it can be argued that imperatives should be regarded as representing a separate dimension, not subsumable under mood. In many languages, particularly in the large Austronesian family, there are verb forms, which combine uses as a future tense with uses typical of nonindicative moods. These are commonly referred to as irrealis forms and are then opposed to realis forms, which correspond to present and past indicative forms in other languages. It should be noted, however, that similarly to non-indicative moods, irrealis forms vary widely with respect to the specific functions they have.

Evidentiality

Evidentiality is customarily defined as the grammatical marking of source of information (Aikhenvald, 2004). Grammaticalized evidentiality marking tends to always reflect a basic distinction between direct ('witnessed') and indirect ('non-witnessed') information sources, where the former notion is understood as meaning that the speaker has sensory evidence for the content of the sentence.

The types of indirect information that are most commonly relevant in TAME systems are (1) inference and (2) second-hand or quotative information (i.e., information obtained from another person). These can be marked separately or by a common marker. Many languages restrict evidentiality markings to past time reference. A standard example of a language with a grammaticalized evidentiality distinction is Turkish, where 'Ahmet came/has come' could be expressed either as *Ahmet gel-di* or as *Ahmet gel-miş*, where the suffix –di is used when the speaker has witnessed the coming, whereas the suffix *-miş* is used when the speaker has inferred or learnt from someone else that Ahmet has come. The TAME systems of most European languages are at most marginally sensitive to evidentiality distinctions, which may explain why evidentiality has not been given due attention in earlier work, although it is a common feature in the languages of all other continents except Africa. Complex evidentiality systems are found in many South American languages.

See also: Control, Syntactic; Evidentials in Linguistics; Grammatical Relations; Grammaticalization; Linguistic Typology; Logic and Linguistics; Markedness in Linguistics; Morphology in Linguistics; Negation, Linguistics of; Semantics; Space, Linguistic Expression of; Syntax (General); Valency and Argument Structure in Syntax.

Bibliography

Aikhenvald, A.Y., 2004. Evidentiality. Oxford University Press, Oxford.
Binnick, R.I., 1991. Time and the Verb: A Guide to Tense and Aspect. Oxford University Press, New York.
Binnick, R.I., 2011. The Oxford Handbook of Tense and Aspect. Oxford University Press, Oxford; New York.
Bybee, J.L., Perkins, R.D., Pagliuca, W., 1994. Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World. University of Chicago Press, Chicago.
Comrie, B., 1976. Aspect: An Introduction to the Study of Verbal Aspect and Related Problems, Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
Comrie, B., 1985. Tense, Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
Dahl, Ö., 1985. Tense and Aspect Systems. Blackwell, Oxford.
Dahl, Ö., Velupillai, V., 2011a. The future tense. In: Comrie, B., Dryer, M., Gil, D., Haspelmath, M. (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 270, 278–279.
Dahl, Ö., Velupillai, V., 2011b. The past tense. In: Comrie, B., Dryer, M., Gil, D., Haspelmath, M. (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 268–269, 276–277.
Dahl, Ö., Velupillai, V., 2011c. The perfect. In: Comrie, B., Dryer, M., Gil, D., Haspelmath, M. (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 271–272, 280–281.
Dahl, Ö., Velupillai, V., 2011d. Perfective/imperfective aspect. In: Comrie, B., Dryer, M., Gil, D., Haspelmath, M. (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 267–268, 274–275.
Dahl, Ö., Velupillai, V., 2011e. Tense and aspect. General introduction. In: Comrie, B., Dryer, M., Gil, D., Haspelmath, M. (Eds.), World Atlas of Language Structures. Oxford University Press, Oxford, pp. 266–267.
de Haan, F., 2011a. Coding of evidentiality. In: Dryer, M.S., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library.
de Haan, F., 2011b. Semantic distinctions of evidentiality. In: Dryer, M.S., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library.
Klein, W., 1994. Time in Language. Routledge, London.
Nikolaeva, I., 2007. Finiteness: Theoretical and Empirical Foundations, Oxford Linguistics. Oxford University Press, Oxford; New York.
Reichenbach, H., 1947. Elements of Symbolic Logic. Macmillan, New York.
Smith, C.S., 1991. The Parameter of Aspect. Kluwer Academic Publisher, Dordrecht.
Vendler, Z., 1957. Verbs and times. The Philosophical Review 66 (2), 143–160.

Territorial Conflicts

Niall Ó Dochartaigh, School of Political Science and Sociology, National University of Ireland Galway, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article begins with a review of historical sociological work on the linked origins of territory, organized violence, and political power. It then surveys the large-n studies on territorial conflict that search for correlations between territorial disputes and the incidence of interstate war. It examines recent disaggregated approaches to the study of territory and conflict before outlining the contributions of the rich case study literature to our understanding of territorial conflict. It analyzes the nexus of territory, technology, and conflict and concludes with a discussion of new theoretical approaches to territoriality, power, and conflict.

Introduction

The fact that almost all competitive team sports present minor variations on the same stylized contest over the control of territory is just one everyday reminder of the deeply rooted relationship between the social phenomena of territory and conflict. The penetration of the deepest part of the opponent's territory that is achieved when a goal or try or basket is scored is one of the simplest and most powerful symbols of victory yet devised. Territorial domination provides unambiguous evidence of domination in general. It has a simplicity that ensures that the symbol can travel easily across linguistic and political borders. Recent decades have seen major advances in social scientific theorizations that analyze territory at a fundamental level, as a core constitutive dimension of conflict of all kinds, a source of power, and a strategy for exerting power. Territory is characterized in this article as a socially constructed phenomenon that takes various historically specific forms but that nonetheless retains certain key characteristics that ensure that it remains a central focus for conflict and violent contestation. The article begins and ends with discussion of theoretical approaches to territory and conflict.

Much contemporary work on territorial conflict has a rather narrow focus however, concentrating on relations between sovereign states, and in particular, on disputes over interstate boundaries. This article reviews this strongly quantitative tradition in the study of territorial conflict, outlining the key questions asked, and the methodologies used. It outlines the key findings of the work on interstate conflict and looks at the way in which similar methodologies and datasets have been deployed to understand territorial contestation within states – by secessionist movements and ethnonational minorities.

More recently, there have been major innovations in the study of territory and conflict based on the geographical disaggregation of violent conflict. Recent studies have moved away from the dominant approaches in which the nation-state as a whole was treated as the key unit of measurement and effectively as the sole meaningful territorial context for explaining violent conflict. Large-scale databases of violence that have been geographically disaggregated to very small localized blocks of space, illustrate starkly the fact that violent conflict is very unevenly distributed within states. At the same time micro-level qualitative research on conflict, much of it by sociologists and anthropologists, has greatly advanced our understanding of the significance of boundaries and territories other than those of the nation-state. Zones of control, informal ethnonational residential boundaries, and administrative and linguistic territories, all help to shape and structure patterns of conflict and act often as a focus for violence.

The article examines the changing character of territorial conflict in an age of rapidly developing information and communication technologies. Remote-controlled technologies for territorial penetration and surveillance, technologies permitting direct communication with civilians in 'enemy' territory, and GPS technologies that permit the deployment of intensely spatially concentrated violence at huge distances unsettle long-established assumptions and legal frameworks for the management of warfare and violence. How then is territorial conflict changing? To what extent are we seeing a reconfiguration of the relationship between territory and conflict or a severing of the links between the two?

We conclude with a discussion of theories of territory and power that offer a way to analyze the connections between conflict and territory at all scales: at the level of interstate disputes; substate struggles over international borders by nationalist and secessionists; and the informal boundaries of everyday life.

The Sources of Territorial Conflict

Much recent research on territorial conflict has focused on disputes between sovereign states over international borders. However, both the sovereign nation-state and strongly institutionalized interstate boundaries are relatively recent phenomena. International border disputes are just one historically specific form of territorial conflict. The work of historical sociologists indicates that territory is not simply one among a multitude of variables that contribute to conflict. They emphasize that territory and organized violence are interdependent and interrelated at the most fundamental level.

In his work on the origins of power Michael Mann (1986) places territoriality at the heart of his explanation for the initial emergence of organized violence more than 5000 years

ago. He uses the metaphor of a gradually intensifying 'caging' of human activity to emphasize the importance of bounding processes in the initial emergence of both states and warfare (1986: 37–40, 45–49). Mann begins by outlining the way in which the initial shift to agriculture intensified the territorialization of social life, sowing the seeds of warfare. Once land had been cultivated and buildings constructed to store the grain surplus from year to year people were tied to specific spaces to a much greater extent than before, tied to their home territories. It became much more important to defend and maintain boundaries as the option of simply moving on elsewhere in the face of violent challenge became much less attractive and feasible. At the same time, the increased fixity of the population created both the opportunity and the need for more elaborate structures of rule and of coercion. These interlinked developments created the conditions for the first time for the emergence of states and the associated organization of violence. States, organized violence, and intensified territoriality emerged together – none of them being possible without the other. Territorially bounded spaces do not just form the backdrop to the organization of violence but are central to the phenomenon of organized violence of all kinds. Without territory there is no warfare.

The work of Tilly (1978) and Giddens (1996) similarly surveys patterns of conflict across the entire globe across huge swathes of time. Both authors illustrate the extent to which interstate boundary disputes in our current relatively stable global system are a historically specific phenomenon – just one of the forms that territorial conflict takes. As Giddens points out, warfare and political violence in the modern era take place within a broader framework of pacification at both national and geopolitical levels and must be understood in the context of long-run historical processes of increasing internal pacification, territorial centralization of power, the growth of state apparatuses, and increasingly clearly drawn international borders. This work places the changing character of territorial control at the heart of our explanations of all forms of organized violence.

Correlating Conflict and Territory

Are territorial disputes more likely to lead to war than other kinds of disputes? Is violence more intense and warfare more prolonged when territory is at stake? Is it really territory that causes war or does an apparent link between territorial disputes and war obscure the importance of adjacency and propinquity? Perhaps neighboring states are much more likely to go to war simply because they interact more and are more likely to have conflicting interests rather than because territory is uniquely contentious. These are the core questions addressed in the extensive large-n literature on territorial conflict that tests tightly defined hypotheses on the causation of war by examining correlations between territorial disputes and warfare in large international datasets.

While some of the work in this tradition argues that proximity or contiguity are the crucial causal factors in increasing the risk of war between two states, most work is dominated by what is called the 'territorial perspective': the argument that territorial disputes play a distinctive part in causing violence. The essential findings of this perspective are that "territorial disputes increase the probability of war and have a higher probability of going to war than other kinds of disputes" (Vasquez and Henehan, 2001: 123). Disputes over territory are more likely to lead to war than 'policy' or 'regime' disputes. The large-n studies treat territory as one variable among many, a variable whose effect can be isolated and whose influence on war, the primary object of inquiry in this literature, can be quantified and compared with that of other variables such as GDP or political system. One of the difficult questions raised by the theoretical literature on territory is whether in fact it is possible to treat such a fundamental constitutive aspect of the political as though it is just one more field to be isolated and defined. This correlative work on territorial conflict is deeply shaped by rational choice models of human behavior, modified to address the symbolic and emotional value that people attach to territory.

Why then should conflicts over territory be more likely to cause war than other conflicts? Surveying this literature, Hensel (2000) identifies three of the reasons why scholars working in this tradition regard territory as important. First, states have straightforwardly rational motivations to want territory because it is a resource valued for its strategic or economic utility. Secondly, territory is important as an idea, powerfully tied up with a sense of national identity and with important emotional aspects of political life. This symbolic dimension of territory often generates strong domestic resistance to compromise on territorial disputes with other states and regimes willing to compromise on territorial issues can experience difficulties in maintaining domestic support. This is what we might call the Field of Blackbirds or Jerusalem argument, given that it has become a commonplace in recent years to illustrate the point by reference to Serb reluctance to relinquish the sites of medieval battles in Kosovo or Israeli insistence on the indivisibility of the holy city of Jerusalem. It is primarily because of the meaning, the emotion, and the symbolism invested in territory that this most divisible of goods can be characterized from a rational choice perspective as indivisible. But this is not necessarily because individuals are hardwired to feel a deep attachment to the national homeland. Acknowledging that emotional and symbolic factors are centrally important in rendering Jerusalem indivisible and hindering a compromise, Goddard (2010) nonetheless argues that there is an instrumentally rational foundation for this indivisibility. She argues that political rhetoric invoking the sacred character of national territory, driven in turn by domestic interparty competition in Israel/Palestine, helped to render Jerusalem indivisible by making it a touchstone issue, thus trapping politicians in their own rhetoric. She contrasts this with the Irish case in which a modification of the Irish nationalist position and rhetoric on the indivisibility of Irish national territory rendered the territory divisible and facilitated territorial compromise.

The third factor identified in these studies is the effect on a state's reputation of yielding in a territorial dispute. States will refuse to compromise on territory in order to protect their reputation, to shape the expectations of other states about how they will behave when challenged and to ensure that they are not seen as weak and vulnerable to pressure. Here too the analysis is saturated with the assumptions of rational choice

theory about the mutually shaping power of both parties' expectations and understandings of the opponent (Schelling, 1960). Presumably territory is such an important factor in reputation both because it is constitutive of the state – a state that loses territory is no longer exactly the same state as it was – and because it provides one of the most unequivocal signs of submission and defeat. Unlike other concessions and compromises, a retreat from territory is particularly difficult to spin as a gain or an achievement of some kind.

While not engaging directly with debates on territory, game theorist Thomas Schelling has quite a bit to say about space and boundaries in his 1960 classic, *Strategy of Conflict*. His analysis of the shaping force of national borders in warfare provides a neat illustration of the way in which both the symbolic power of borders and their role in shaping expectations can powerfully influence the course of a war. Explaining why the United States decided not to cross the Chinese border during the Korean War despite Chinese involvement in the war Schelling writes:

> What makes the Soviet or Chinese border a pertinent or compelling place to draw a line in the event of war in that area is principally that there is usually no other plausible line to draw … there is just no other stopping place that can be tacitly acknowledged by both sides…It has a compelling power of suggestion, a claim to attention, the denial of which might seem … to be a denial of any limitation.
> Schelling 1960[1980]: 259

Schelling highlights the moderating and limiting power of clearly delineated and mutually recognized borders but the point has relevance too for territorial disputes. In highlighting ideas, understandings, and expectations, Schelling provides here a strongly rationalist argument for the importance of the symbolic dimension of boundaries in shaping the pattern and intensity of warfare.

While large-n studies of territorial conflict have focused overwhelmingly on interstate disputes over international borders, some scholars have turned their attention more recently to nationalist and secessionist disputes that are generally characterized as domestic or internal. Duffy-Toft (2003), for example, analyzes the correlations between the level of territorial concentration of minorities and the likelihood of political violence and finds that territorial concentration correlates strongly with high levels of violence.

One of the curious features of this large-n work is that much of it has very little concern with territory itself or even, in many cases, with territorial conflict. The relationship between territory and conflict is often investigated primarily because it provides one of the more straightforward ways in which to test the democratic peace thesis which holds that democracies are less likely to go to war. The likelihood of territorial disputes to cause conflict is investigated in order to determine whether political systems are the key determinant in warfare. In addition, even where this work is mainly focused on identifying the primary causes of conflict it does not tend to be very much concerned with investigating the nature of territory. It is rare for work in this tradition to cite theorists of power and territory such as Gottmann, Sack, Paasi, Raffestin, or Lefebvre. The result is that territory itself is curiously absent from much of this work.

Disaggregating Conflict

One of the weaknesses of the large-n work on territorial conflict is that it is restricted to just one kind of historically specific territory – the nation-state – and just one kind of territorial conflict, disputes over interstate borders, a weakness recognized by some of those working in the field. This is partly the product of the widespread methodological nationalism that treats the nation-state as the primary geographical unit of analysis and has locked much social scientific work into state-bounded frameworks for analysis. Using state territories as part of the framework for the coding and collection of data on territorial conflicts renders other kinds of territories invisible, reifies the distinction between internal and external, and steers analysis toward a relatively narrow set of questions focused on a small subset of territorial conflicts.

There is a radical spatial unevenness to violence in civil wars and other intrastate conflicts that defies explanation at the national level. This variation is almost invisible in datasets that use the state as the main unit of analysis. Civil wars can often be difficult to disentangle from interstate wars and the problem of unevenness also has implications for our understanding of the latter. In virtually all civil wars large swathes of territory remain largely unaffected by violence, and villages only a few miles apart can have sharply contrasting experiences of violence. This unevenness presents a challenge to explanations of conflict that treat states as the primary unit of analysis. If factors at the national-level generate civil war conflict why do they not generate it more evenly across the space of the states concerned? If conflict is radically uneven in its distribution and if contests over territory at the regional and local level are crucial in explaining these uneven patterns it follows that much of the explanation of the relationship between conflict and territory must be located at these same scales. If violence is unevenly distributed, perhaps the explanation for violence lies not with the characteristics of a state as a whole but the characteristics of particular regions or localities, or with spatially and temporally uneven processes in which local territorial frameworks of various kinds decisively shape the escalation and diffusion of violence.

In response to this problem of unexplained unevenness in violence, a range of databases of disaggregated data on incidences of violence have recently been established, and a lively publication program has begun to explore subnational patterns of distribution and diffusion of violence (O'Loughlin and Raleigh, 2008; Raleigh et al., 2010). This new approach is valuable in highlighting the exceptional unevenness of violence within states and showing the close relationships between interstate and internal violence. In the case of the Democratic Republic of Congo (DRC) for example, this research provides clear, graphic illustration of the fact that a huge proportion of violent conflict in the DRC is part of a trans-border zone of violence in the Great Lakes region. The clear implication is that it cannot be explained primarily in terms of the internal dynamics in the DRC. The lines of territorial control established by local combatants are just as important in understanding patterns of violence in this trans-border zone as the formal international boundaries across which many of them operate. These studies focus on the distribution of violent incidents in space and on correlations between variables at this

Table 1 Theoretical approaches to territory and conflict

Approach	*Method*	*Hypotheses*	*Findings*
Social caging (Michael Mann)	Sociological analysis of long-term historical processes	New economic and coercive structures lead to the 'caging' of the population into dominant power relations	States, organized violence and territorial caging emerge together and are co-constituted
The territorial perspective	Large-n analysis correlating incidence of warfare with interstate territorial disputes	Territorial disputes are more likely to lead to war than other kinds of disputes	Territory is more likely to lead to war because it is a resource and a potent symbol of national identity and its loss effects a nation's reputation
Spatial analysis of conflict	Geographically disaggregated large-n analysis of data on conflict	Local contexts decisively shape patterns of civil war violence	Areas distant from central control and with high population density are at greatest risk of civil war
Micro-level analysis (Kalyvas)	Correlating patterns of violence with levels of territorial control of protagonists	Territorial control decisively shapes patterns of violence	Violence is greatest in zones of disputed control where one group is more powerful, not in zones where their power is more equally balanced
Human territoriality (Sack)	Theorization of power and territory at multiple scales across a broad historical span	Territoriality is a uniquely effective strategy for exerting power	Three primary 'tendencies' in territoriality: enforcement, communication, and classification

disaggregated level. This permits the introduction of new territorially related variables such as distance from capital and from border. But in some senses they are almost as unconcerned with territory as are the state-level studies that preceded them. Despite the fact that they support the argument that state boundaries are not necessarily the boundaries of primary significance in conflicts, the studies tend to treat state borders as a relatively unproblematic variable.

However, other approaches to understanding this kind of local territoriality are possible, and some of the most important advances in understanding territory and conflict have come through rich, localized case-studies. One of the most important early contributions was Charles Tilly's study of counterrevolution in the Vendée in the late eighteenth century (Tilly, 1964). Covering just a few years in one region of France Tilly demonstrated that violence could not be adequately explained in terms of broad national patterns or structural categories such as class. He noted that the revolt was not simply localized to western France but that it was confined to certain subregions in the west and that even within those zones of rebellion, different villages and settlements responded quite differently. Boundaries between local villages and districts could take on intense meanings and importance in times of violent confrontation and territorial bases where sympathizers were densely concentrated were crucial in allowing the counterrevolution to gain momentum.

In 'The Logic of Violence in Civil War,' an analysis of civil war through a case study of violence in one Greek region, Kalyvas (2006) has placed local variation and the territorial unevenness of violence at the center of explanations of civil war, connecting it to distinctions between local and central actors and between the public and the private. He provides a deeply territorial explanation for patterns of violence, focusing not on formal boundaries of any kind, but on the local zones of control established by competing armed actors.

National-level methodologies and explanations have simultaneously been challenged from above, by analyses that emphasize the importance of geopolitical context and external influence in shaping patterns of 'internal' violence. There is a growing recognition across the social sciences that spatial variation, territorial boundaries, and bounding processes are properly located at the heart of any understanding of the causes of war and particularly of civil war. Subnational boundaries in their various forms, whether regional or local boundaries, lines of control established by rebels or no-go areas for state security forces, need to be analyzed in tandem with national borders and need to be understood in the geopolitical context in which all of these territories are ultimately nested (Table 1).

Conflict, Territory, and New Technologies

As new technologies emerge and weaponry develop, society changes; older modes of territoriality are superceded; and the territorial configuration of conflict is transformed. Just as agricultural technologies provided a foundation for the first city-states and the first wars, so the development of powerful artillery in the seventeenth and eighteenth centuries rendered the city walls of medieval Europe redundant. The new-found vulnerability of the walled city contributed in turn to the decline of the city-state as a political formation and the growth of the nation-state. If eighteenth century cannons removed the walls from the city-state, the aerial bombing of the Second World War 'removed the roof' from the nation-state. This new technology rendered the deepest part of the homeland penetrable and vulnerable. In so doing, it blurred the distinction between front line and homeland, and between military and civilian targets. From the siege warfare of the sixteenth century through the set-piece battles of the eighteenth, and the trench warfare of the early twentieth century to the smashing of the distinction between frontline and homeland sanctuary in the Second World War, the territoriality of violent conflict has undergone regular dramatic transformations made possible by new technologies.

The rapid acceleration in the development of new communication and information technologies since the advent

of the World Wide Web in the early 1990s seems to herald another deep historic shift in the territoriality of conflict. The 1990s saw much discussion of 'deterritorialization' and the fading of boundaries. But new technologies facilitated the hardening of boundaries in many cases, as new walls of concrete and wire mesh were put in place across the conflict zones of the Middle East and on crucial borders between first world and third, reinforced with electronic surveillance. Reterritorialization became the preferred catchphrase of the early 2000s as it became clear that shifts in the configuration of territory did not erode the power of territoriality as a strategy for exercising control. The struggle to control territory remained at the heart of violent confrontation.

Pilotless drones now allow technology-rich states to penetrate deeply into the territory of other sovereign states to carry out the kind of precisely targeted attacks that would previously have required a human presence on the ground. Communication technologies allow detailed, direct communication between states and the civilian population of enemy territories. Radio broadcasts into enemy territory during the Second World War were aimed at demoralizing enemy civilians by communicating directly the extent of the detailed local knowledge that the approaching bombers had. The use of mass texting to personal numbers provides a much more focused and effective version of the same phenomenon. The text messages sent in recent years by the Israeli Military to Lebanese and Palestinian civilians informing them that their homes or areas were about to be bombed served the dual aims of reducing civilian casualties and conveying penetration and control of enemy territory. At the same time, online videos, websites, and social networks all provide discussion spaces where civilians associated with opposing parties to conflict can engage in direct communication with the 'enemy.' It is a development that seems unlikely to have a moderating influence on conflict. It seems more likely that direct communication will reinforce and intensify hostility by providing people with direct personal experience of enemy 'aggression' online.

The widening gap in technological capability between protagonists ensures that territory can now be overrun and captured with extraordinary speed by small numbers of troops with superior technology facing a much larger number of opposing forces. The extent of this gap was evident in the lopsided death tolls of recent operations such as the 2003 invasion of Iraq and Israel's Operation Cast Lead in Gaza in 2008/9 – a ratio of a hundred to one in the latter case.

For those holding the new technological cards, territory is easier to penetrate and take, but it is still not easy to hold on to and to dominate in the long term. New technologies, combined with the old-fashioned walls and barriers built around some Iraqi towns and throughout Baghdad, provide crucial tools for maintaining control. Retina-scanning, face-recognition photography, and electronic surveillance combine with physical barriers to permit high levels of monitoring of the general population of certain districts. However,some of the new technologies, such as the mobile phones sometimes used to trigger Improvised Explosive Devices (IEDs) benefit challengers as well as incumbents. Local allies and enforcers remain vital in establishing and maintaining control on the ground. The territorial control that could be exerted at the neighborhood and district level by Shiite militias in Iraq and by groups such as the Northern Alliance in Afghanistan was essential to securing control on the ground – and even then it has proven very difficult to convert swift and overwhelming victory on the battlefield into secure control of conquered territory.

The changing character of territorial conflict in an age of new technologies is a central focus for a great deal of the recent work on space and violence by critical geographers (Elden, 2009; Graham, 2008; Gregory and Pred, 2006). This work is focused to a great extent on recent western military interventions and the geographies of the 'War on Terror.' It deals with issues such as the erosion of territorial sovereignty, the city as battlefield and the geographies of counterinsurgency and examines the imaginative geographies and geopolitical grand visions that shape action on the ground.

Theorizing Territoriality and Conflict

Robert Sack's theory of territoriality is just one of several innovative approaches to power and territory that have emerged since the mid-1980s. It serves here to illustrate the value of new theoretical approaches in advancing our understanding of territorial conflict at a fundamental level. Sack offers an analytical framework that can be applied to boundaries and territories of all kinds and that takes into account the transitory and historically specific character of territorial systems and borders. It also provides a framework for analyzing the role of ideas in the exertion of power through bounding and territoriality.

In his 'Human Territoriality: Its Theory and History' (1986) Sack analyzes territoriality as a uniquely effective strategy for exerting power. He defines it as "the attempts by an individual or group to affect, influence or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area." He describes territoriality as "a primary geographical expression of social power" (1986: 5, 19). These attempts "… to affect, influence or control people" by delimiting and controlling territory are evident first and foremost at the level of the global political system. The apparently natural and eternal division of the planet into sovereign states is maintained to a great extent by the cooperative efforts of nation-states which have a shared interest in elevating the status of existing states as far as possible above the claims of challengers. The value of delimitation as a means of exerting power is evident too at lower levels, in the manipulation of electoral and regional and municipal boundaries to bolster the power of a state or a dominant group or to provide political leverage to a geographically concentrated minority. It is clear too at the lowest scale, in the bounding of neighborhoods deemed to be problematic, in the building of localized barriers that steer movement and interaction and in the creation of special zones of control around key buildings. Nowhere is the power of territory more urgently and dramatically manifested than in situations of violent conflict where the location and strength of boundaries and barriers can make the difference between life and death.

Sack identifies three 'tendencies' that define territoriality. The first of these, enforcement, refers to the way in which control over a multitude of things can be secured by control

over one thing: access. You don't have to enforce a multitude of limitations on behavior in a particular space if you don't let people in to begin with. At the same time you obscure the extent of the control that is being exerted. Communication is the second tendency. The power of boundaries derives partly from their simplicity as a symbol that can nonetheless convey huge volumes of information. Finally Sack identifies classification as a key tendency. Territorial classifications provide a powerful and simple system of ordering that reduces the need to further classify and specify the contents of the delineated space.

Sack's theory is explicitly inter-scalar, theorizing territoriality in such a way that we can analyze the interlinked power of boundaries of different kinds. While Sack focuses primarily on the power exerted through the establishment of formal boundaries, the theory can usefully be applied to all kinds of boundaries.

Although the approach has weaknesses, it nonetheless provides a useful lens through which to view the relationship between territory and power, and between territory and organized violence. It is just one of a number of recent perspectives that can usefully be deployed in pursuit of a deeper understanding of the phenomenon of territorial conflict.

Territorial conflict involves much more than interstate boundary disputes. It involves many different kinds of interrelated boundaries: formal and informal; local, regional, and national; and political and cultural – and it transcends the distinction between internal and external. Indeed, it is the very distinction between inside and out that is at stake in many conflicts: while nationalists and secessionists seek to make the internal external, wars of unification or expansion pursue the goal of internalizing the external. There is accordingly a need for theoretical approaches that can address territorial conflict at these various interlinked levels, that can transcend the distinction between internal and external and avoid privileging the nation-state: an approach that can explain the territoriality of both interstate war and civil wars through the same theoretical lens. Adapting Brighenti's (2010) argument for a new approach to territory and law we might argue that the problem is not to distinguish territorial from nonterritorial conflict or to distinguish the internal from the external but to elaborate "… a concept of territory powerful enough to speak of [organized violence] in its entirety." Brighenti's original quote: "The problem is not separating and opposing territorial and nonterritorial law … rather, elaborating a concept of territory powerful enough to speak of law in its entirety; in other words, to found a territorology of law" (Brighenti, 2010).

Conclusion

There has been a broad shift in recent years from state-based to geographically disaggregated analyses of the causes of violence and from the study of interstate war to the study of civil war violence. This shift has helped to open up several new lines of inquiry in the study of territorial conflict. Theories of territory and power that provide analytical purchase on the relationship between territory and violence, and between territory and legitimacy and order, have much to contribute to the development of these lines of inquiry.

One of those lines of inquiry is the interrelatedness of boundaries of different kinds and at different scales in shaping patterns of violent conflict. Explaining how territories and boundaries are made and maintained in situations of violent conflict requires an approach that treats interstate borders as one more kind of boundary, albeit an exceptionally powerful kind of boundary. Interstate boundaries are related to, lend strength to, and draw strength from many other different kinds of boundaries. If boundaries are crucial in shaping expectations and informing behavior, understanding violent conflict requires a much better developed appreciation of the connections between violence and the making of territory.

There also remains much scope for work on the way in which the global state system shapes the forms that territorial disputes take, work that might help to explain why we have seen such a marked shift in recent decades from interstate war to civil war. A rigid global system in which there are immense pressures against the modification of interstate boundaries may well help to push territorial struggles down to lower levels, displacing them rather than eliminating them. In Bosnia and Herzegovina and in the Kurdish region of Northern Iraq, to take two of the clearest examples, strong internal borders confer almost as much autonomy and power on regionally concentrated minorities as international borders without presenting a direct challenge to the interstate system. In the case of Iraq, the struggle over the exact location of the internal boundary and over the control of local spaces, has taken on much of the character of an interstate boundary dispute. Here we have a clear example of a boundary dispute that could easily be converted into a secessionist and ultimately an interstate dispute, but for a political calculus in which the global state system weighs heavily.

Finally, we have the question of how new information and communication technologies are reshaping the territorial configuration of violent confrontation. Are the new technologies facilitating deep shifts in the processes by which borders and territories of all kinds are maintained and reproduced? How do new technologies of bounding and penetration impact on the territorial dynamics of violent conflict? If communication is a central dimension of territoriality, how do new forms of communication effect the deployment of territorial strategies? The work in this area is still in its infancy. It stands out as perhaps the most important and urgent topic in the understanding of contemporary territorial conflict.

See also: Ethnic Conflict, Geography of; Geopolitics; Military Geography; Nationalism, Sociology of; Postcolonial Geography; Rational Choice Explanation: Philosophical Aspects; Revolutions, Sociology of; Scale in Geography; Space and Social Theory in Geography; Territoriality: Geographical; Violence, History of; War, Sociology of; War: Anthropological Aspects, Historical Development of; Warfare in History.

Bibliography

Brighenti, Andrea, 2010. On territorology: towards a general science of territory. Theory, Culture & Society 27 (1), 52–72.

Duffy-Toft, Monica, 2003. The Geography of Ethnic Violence: Identity, Interests, and the Indivisibility of Territory. Princeton University Press, Princeton.

Elden, Stuart, 2009. Terror and Territory: The Spatial Extent of Sovereignty. U of Minnesota Press.
Giddens, Anthony, 1996. The Nation-state and Violence. Polity, Cambridge.
Goddard, Stacie, 2010. Indivisible Territory and the Politics of Legitimacy: Jerusalem and Northern Ireland. Cambridge University Press.
Graham, Stephen, 2008. Cities, War, and Terrorism: Towards an Urban Geopolitics. Wiley-Blackwell.
Gregory, Derek, Pred, Allen (Eds.), 2006. Violent Geographies: Fear, Terror, and Political Violence. Routledge, New York.
Hensel, Paul, 2000. Territory: theory and evidence on geography and conflict. In: Vasquez, John A. (Ed.), What Do We Know about War. Rowman and Littlefield, Boulder, CO, pp. 57–84.
Huth, Paul, Allee, Todd, 2003. The Democratic Peace and Territorial Conflict in the Twentieth Century. Cambridge University Press.
Kalyvas, Stathis, 2006. The Logic of Violence in Civil War. Cambridge University Press, Cambridge ; New York.
Mann, Michael, 1986. The Sources of Social Power. In: A History of Power from the Beginning to 1760 A.D, vol. I. Cambridge University Press, Cambridge; New York.
O'Loughlin, John, Raleigh, Clionadh, 2008. Spatial analysis of civil war violence. In: The Sage Handbook of Political Geography. Sage, London, pp. 493–508.
Raleigh, Clionadh, Linke, A., Hegre, H., Karlsen, J., 2010. Introducing ACLED: an armed conflict location and event dataset. Journal of Peace Research 47 (5), 651–660.
Sack, Robert D., 1986. Human Territoriality : Its Theory and History. Cambridge University Press, Cambridge.
Schelling, Thomas, 1980[1960]. The Strategy of Conflict. Harvard university press, Cambridge MA.
Tilly, Charles, 1964. The Vendée. Harvard University Press, Cambridge, MA.
Tilly, Charles, 1978. From Mobilization to Revolution. McGraw-Hill, New York.
Vasquez, John, Henehan, Marie, 2001. Territorial disputes and the probability of war, 1816–1992. Journal of Peace Research 38 (2), 123–138.

Territoriality: Geographical

David Storey, University of Worcester, Worcester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.D. Sack, volume 23, pp. 15601–15604, © 2001, Elsevier Ltd.

Abstract

Human geographers situate territoriality in a social and political context in which territories are seen to be socially produced and territorial strategies can be seen as mechanisms to achieve particular ends. The control of geographic space can be used to assert or to maintain power, or to resist the power of a dominant group. Rather than being natural entities, territories reflect specific ways of thinking about geographic space and they emerge from social practices and processes that link space and society. These geographic spaces convey messages of political power and control which are communicated through various means and the creation, maintenance, and patrolling of boundaries divides those inside from those outside. Control over territory is a key political motivating force and territoriality can be seen as a spatial manifestation of power. While some of the most obvious territorial constructions and visible expressions of territoriality occur at the level of the state, many more micro-scale examples may be observed. Territorial strategies can be used in order to express an identity while territorially transgressive acts can be employed to reclaim space and, hence, to assert basic rights.

Most people are regularly confronted with signs warning 'authorized personnel only,' 'no trespassing,' 'prohibido el paso,' and so on. Such regular warnings and admonishments are a reflection of attempts to impose forms of power (via sets of rules and regulations) through control of portions of geographic space. Space is claimed, boundaries are produced and patrolled, and territorial strategies are deployed across a range of spatial scales. Such strategies are not always successful and claims on space are often contested and boundaries transgressed. The various manifestations of human territoriality might lead to an assumption that it is an innate biological urge. While it is tempting to relate territorial behavior to some inbuilt feature of human beings, more nuanced perspectives suggest that much of our territorial behavior is a consequence of our broader sociopolitical conditioning. By and large human geographers tend to locate human territoriality in its social and political (rather than biological or genetic) context.

Whether explicit or implicit, control over territory is a key political motivating force and territorial practices are linked to the interplay of social and political forces. Territoriality can be seen as a spatial expression of power and the processes of control and contestation over portions of geographic space are among the central concerns of political geography (Storey, 2012). Any consideration of territories necessarily raises questions to do with boundaries. The borders of a territory may be quite clearly defined and formalized, as between most countries, or they may be somewhat informal or ill-defined, as is the case when considering the territories of rival urban gangs whose 'patch' or 'turf' may well have a core but the precise boundaries of which may be unclear. The demarcation of territories and the bounding of space obviously bring with them issues of exclusion and control and territories, and their boundaries, are often the center of disputes (in many instances quite violent) and the areal extent of particular territories (and even their right to exist) may be regularly called into question. Rather than being fixed immutable entities, human territories and their boundaries are subject to periodic or continuous contestation, modification, transformation, and destruction (Paasi, 2008). The global political map provides visible illustration of the most obvious manifestation of the territorial dimension to politics in the division of the earth into separate countries or states, but territories are also produced, controlled, bordered, and contested across a range of spatial scales and in a wide range of contexts. The following discussion explores territory and territoriality from a geographic perspective, it examines the concept of territory, and it explores the links between territorial formations and behavior, and senses of identity and belonging.

Territoriality

Crude deterministic perspectives suggest a territorial imperative that is seen as a natural and unchanging phenomenon. This viewpoint holds that the need for space is a characteristic innate to all species, including humans, and there is an associated impulse to defend this territory against others seeking to 'invade' it (Ardrey, 1967). Parallels between the behavior of humans and animals were popularized by the zoologist, writer, and broadcaster Desmond Morris (1994), among others, who characterized humans as basically territorial creatures with an innate need to defend territory, whether at the level of the nation or at the micro-scale level of the home. These sorts of analogies with animal behavior have a long history. The geographer Friedrich Ratzel developed organic theories of state formation in the late nineteenth century suggesting state expansion was a necessary means of ensuring a country's survival. Drawing parallels with evolutionary theory he suggested that states needed to adopt 'survival of the fittest' strategies in order to retain power, an argument serving as a convenient justification for aggressive territorial defense and acquisition (Bassin, 1987). These sorts of determinist argument attribute many human characteristics to nature and employ ideas from the natural sciences in attempting to explain human behavior. This is then seen as innate rather than emerging out of our social and cultural conditioning. However, while there may well be behavioral

parallels between animals and humans, it does not follow that the motivating factors are the same.

Contemporary geographical contributions to the study of human territoriality have tended to align with wider arguments within the social sciences which suggest human behavior is heavily conditioned by the economic, social, political, and cultural environment in which we exist. Central to this perspective is an emphasis on power relationships rather than biology or genetic inheritance. Soja (1971) suggests that, although characteristics such as aggression and territoriality are displayed by humans, this does not mean that there is anything inevitable about them or that they cannot be altered or avoided. Rather than adopting simplistic deterministic explanations, or getting entangled in the intricacies of nature–nurture debates, it seems reasonable to assume we are subject to complex sets of influences within our broader environment and these influences interact with our genetic potential.

Two geographers who have contributed significantly to debates on human territoriality are Jean Gottmann and Robert Sack. Gottmann argues that territory, and the assertion of control over it, represents an expression of power. In his view, territory confers security through the control of defensible space while also providing opportunities through facilitating the economic organization of that space. This emphasis on security and opportunity points to a political and economic basis for territorial formations which make for a more efficient means of political organization than systems of overlapping jurisdictions. Sack emphasizes the role of territoriality as a geographic and political strategy through which individuals or groups endeavor to "affect, influence, or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area" (1986: 19). He emphasizes the instrumental nature of territoriality and draws attention to the means through which territorial strategies may be used to achieve particular ends so that control of geographic space can be used to assert or to maintain power, or, importantly, to resist the power of a dominant group. What this means is that territoriality is deeply embedded in social relations so that territories, far from being natural entities, are the result of social practices and processes and are produced under particular conditions in order to serve specific ends (Delaney, 2005).

These perspectives are important in not only emphasizing the political context of territorial behavior but also in highlighting how territoriality as a strategy operates at all spatial scales from the geopolitical strategies of global superpowers down to the home and the workplace. Sack also draws attention to the significance of territoriality in facilitating classification, communication, and enforcement. Territoriality involves a classification by geographic area through which space is seen to be apportioned between states or between individuals. In this way, a room, or an office workstation becomes 'my' territory and others are only permitted to be there as 'guests.' Sack also argues that territoriality is easy to communicate via the use of boundaries which indicate territorial control and, hence, power over prescribed space (and of those within it). In this way, there is a distinct separation between those who are 'inside' and those who are 'outside,' thereby conveying clear messages about power and control. In this way, territory and border construction in classifying space and communicating power act as devices for the enforcement of authority. This may have important implications such as constraining, restricting, or limiting people's mobility, for example. Territories and territorial ways of thinking link space and society and serve to convey clear meanings relating to authority, power, and rights (Delaney, 2009).

Another tendency of territoriality identified by Sack is that it functions as a means through which power is reified. Through the visibility of land (or a room or a desk) power can be 'seen' in a way which suggests "territory appears as the agent doing the controlling" (Sack, 1986: 33). This reification is most obviously apparent in the deployment of the term 'the law of the land' whereby power appears to reside in the territory itself rather than in those who control. Attention is thereby diverted away from the power relationships, ideologies, and processes underpinning the maintenance of territories and their boundaries. In this way, "territory does much of our thinking for us and closes off or obscures questions of power and meaning, ideology and legitimacy, authority and obligation" (Delaney, 2005: 18). This interpretation of territorial behavior emphasizes the political functions of territoriality so that it can be seen as "a primary geographic expression of social power" (Sack, 1986: 5).

The Emergence of Territory

While Sack's functionalist emphasis is clearly important, a somewhat less instrumentalist view emerges in the work of the Swiss geographer Claude Raffestin (2012). He views territoriality in a somewhat broader way, seeing it as a process produced by the various relationships between individuals and groups and their wider social environment, not just something bound up with the attainment of particular goals. Raffestin's concern is with the emergence of territories out of social interaction, not solely with the operation of territorial strategies. In a connected vein, Stuart Elden has drawn attention to the concept of territory itself. He argues that within debates on territoriality, the idea of territory has been taken for granted so that "strategies and processes toward territory…conceptually presuppose the object that they practically produce" (2010). He suggests that territories themselves, and their boundaries, reflect a distinctive mode of social and spatial organization which is linked to particular ways of thinking about geographic space and these ways reflect notions of power and control. Elden argues that our contemporary ideas of territory emerged as a consequence of developments in cartography and spatial calculation; it is something that can be calculated, mapped, and controlled. The features of territoriality identified by Sack (1986), in allowing classification and differentiation, are thus products of the ways in which space is imagined and territories are in effect politicized space; mapped and claimed, ordered and bordered, measured and demarcated (Elden, 2007).

If territories reflect a particular way of thinking about space, then this points to the importance of maps in solidifying and legitimizing these spatial units. Rather than neutral depictions of supposed geographical realities, a more critical analysis suggests that maps have always been useful weapons in larger political projects associated with territorial claims and counterclaims (Wood, 1992). Mapping of territory itself functions so as to enhance power sending out messages signifying control

over portions of geographic space. Maps of the British Empire conventionally depicted Britain's overseas territorial possessions in pink which not only conveyed (often limited) information but more importantly broadcast a clear political message of power over roughly one quarter of the planet's land area. Advances in cartography and spatial calculation modified the ways in which space was considered and emerged alongside attempts to apportion and control it. The military and political underpinnings of cartographic developments should not be ignored and the consequent role of mapping (in both practical and symbolic terms) in the creation of colonial territory was a key element in the imposition and maintenance of political control.

The Territorial State

The creation of states and the construction of borders between them represent the most obvious political expression of territoriality and much geographical coverage of territory and territoriality focuses on these. The Westphalian state has become the world's dominant mode of political organization and it is difficult to conceive of a world without states and borders, so firmly embedded has it become. Territory is the functional space within which the state operates and wields power and can be seen as a key and necessary feature of the state. Borders are the dividing lines where state control both begins and ends. While the territorial state is at the center of the global political system, we need to be careful not to fall into what Agnew (1994) terms the territorial trap whereby we naturalize states and their borders. He argues that states are imagined as fixed sovereign entities with power over their territory, in turn suggesting a clear distinction between 'domestic' and 'foreign,' whereby states are characterized as territorial containers of economy and society. However, Brenner and Elden (2009) (drawing on the work of Henri Lefebvre) point to the ways in which territory is continually reproduced through the actions of states and through various forms of political struggle. They argue that territory, rather than simply being a feature of the state, both facilitates and emerges from state action. Clearly sovereignty and the state system place a heavy emphasis on geographical borders that separate one territorial political entity from another. Borders are often seen as taken for granted features serving as political–cultural dividing lines. However, they might more accurately be seen as discursive constructs that serve to legitimize the state (Fall, 2010; Newman, 2010). Not surprisingly, boundary disputes often lie at the heart of many conflicts and such disputes regularly emerge in regions with considerable ethnic complexity.

Territory and borders are necessary features of the contemporary state but these are closely linked to ideas of sovereignty or the right of the state to control its people and its territory. Agnew (2005) argues that sovereignty is highly complex and the sovereign power of an individual state is continually challenged, contested, and modified. Processes associated with globalization (or linked together under that broad umbrella term) are seen by many as signaling a diminishing sovereignty for the territorial state. Borders are commonly depicted as increasingly porous in a world of transnational flows of people, goods, finance, and ideas. The emergence of regional economic alliances, epitomized by the evolution of blocs such as the European Union and the Association of South East Asian Nations, is heralded by some as the successor to the state system in an increasingly globalized world where state boundaries are assumed to have diminishing significance. The spatial diffusion of the financial crisis in recent years points to a diminishing ability of the state to maintain economic sovereignty. However, predictions of the demise of the state and the advent of a borderless world seem somewhat exaggerated. While a neoliberal economic orthodoxy (and its attendant crises) centered on trans-global economic flows seems to point toward the irrelevancy of borders, it simultaneously utilizes and reconfigures them in attempts to increase the returns to capital. In addition, there is the apparent paradox that while capital flows relatively freely across international frontiers, many people (depending on who they are, where they are from, and where they are trying to go) are faced with the intimidating paraphernalia of the state manifested at border crossing points.

Geopolitically, Elden (2010) and others suggest that, far from disappearing, territory is continually being reconfigured and spatial political relations are constantly being reshaped in ways suggestive of reterritorialization rather than deterritorialization. Power is unequally distributed and some political–territorial formations exert significantly more than others. The US detention center of Guantánamo Bay provides one example of the increasingly complex nature of territorial sovereignty with the highly strategic use of a geographically bounded space that is both simultaneously under US control while also outwith the ordinary political jurisdiction of the country (Reid-Henry, 2007). Sovereignty, for some, is contingent on adhering to particular ground rules laid out by the more powerful. External interventions by Western powers in a range of regions (including Iraq, Afghanistan, Libya, Syria, Mali) serve to demonstrate that territorial sovereignty is highly contingent on the behavior of a state with major powers appearing to reserve the right to intervention. While such actions are often clothed in the language of civil rights and humanitarianism, in reality they are usually bound up with the advancement of the strategic interests (economic and political) of powerful elites. Some states display a willingness and ability to exert power well beyond their formal territorial boundaries while others (sometimes labeled 'failed' or 'rogue' states) become arenas for external intervention. The linkages between state sovereignty and territorial control are far from straightforward and lead to a consideration of effective sovereignty wielded by both states and quasi-state actors and deployed across a range of territorial contexts (Elden, 2009; McConnell, 2010).

Of course these contemporary geopolitical strategies are nothing new. Empire-building by European powers represented a territorial expression of power on the part of those countries which led to much of the world's surface area being carved up between them and out of which there has been an enduring territorial legacy (together with a range of economic, political, and cultural consequences) throughout Africa, Asia, and the Americas. More recently, the Cold War era of the middle years of the twentieth century saw the United States and the Soviet Union seeking to establish geographical spheres of influence. While this might not have been formalized territorial

control, it was a process whereby 'friendly' governments were encouraged, or sometimes forcibly installed, in countries which the superpowers saw as vital to their strategic interests. Some would argue that the ongoing creation and maintenance of military bases in other countries and the economic and political pressures exerted (both directly and indirectly) by major powers further highlights the contingent nature of state sovereignty while reflecting a contemporary version of colonialism through which subject peoples and territories are maintained in a subordinate position (Harvey, 2003; Gregory, 2004).

Territory, Place, and Nation

While drawing attention to the emergence and functions of territories, it is important to note that the phenomenon should not be reduced to a mere commodity or territories seen simply as (sometimes leaky) spatial containers. Territories, and the ways in which they are imagined, can play an important role in the formation of peoples' self-identity and contribute to feelings of belonging or exclusion. The intrinsic geographic importance of territory is stressed by Soja who suggests it "provides an essential link between society and the space it occupies primarily through its impact on human interaction and the development of group spatial identities" (1971: 33). Once created, territories can become the spatial containers in which people are socialized through various social practices and discourses. Paasi (2008) suggests that a number of important dimensions of social life and social power are brought together in territory. These include a material component such as land, a functional element reflected in control of (or attempts to control) space and a symbolic component associated with people's social identity. The social, cultural, and political are brought together so that people identify with territories in such a way that they can be seen "to satisfy both the material requirements of life and the emotional requirements of belonging" (Penrose, 2002: 282). Although Sack (1986) suggests that territoriality forms a backcloth to human spatial relations and conceptions of space, territory is much more than this, and more also than simply the material manifestation of sets of social relations. As Sack himself also intimated territory is a device through which we experience the world and give it meaning. The spatial is not simply the outcome of the social but the two are intrinsically interwoven so that territories, and territorial strategies associated with them, can be seen as powerful influences shaping social life (Delaney, 2009).

This is very evident in the context of national space where there is a significance seen to be attached to a national territory, reflected in territorial claims and various expressions of ethnonational identity and belonging. Nations (as distinct from states) can be seen as groups of people linked through a sense of a common identity bound together by feelings of a shared history and an attachment to an identifiable territory or homeland. While nations might be seen as comparatively 'modern' constructions they rely heavily on history, myth, and territory. Nationalism is both a political and territorial ideology in which a 'national' territory is defended, claimed, or sought in the name of the 'nation,' though the construction of nationhood is almost always a highly contested process in which competing narratives are pitted against each other. Territory is quite central in many nationalist discourses and is habitually invoked in national conflicts where both generic territory and specific places may acquire enormous symbolism. References to the national soil abound so that fighting for, or even dying for, the land are often seen as supreme acts of patriotism in times of national conflict. A second element in the connections between land and identity is the 'production' and veneration of generic landscapes. Particular parts of the national territory may acquire a significance as the presumed heartland or 'core' of the nation. Nationalist conflicts periodically give rise to attempts to create ethnically pure spaces linked to the mythologizing of certain places and associated boundary-making, as apparent in the Balkan disputes of the 1990s (Kolstø, 2005). Mobilization around ethnic identities provides one example of the forces that may give rise to pressures for territorially based entities to secede and to create new states.

Territory, Space, and Society

While some of the most obvious (and contested) expressions of territoriality are manifested at the level of the state, many more micro-level examples of territorial control and territorial strategies may be observed and this is reflected in the creation and sustaining of a vast array of spatial enclaves (Sidaway, 2007). Social practices associated with race, class, gender, and sexuality are spatialized with boundaries constructed and contested in people's everyday lives while territorial strategies are used to control and to resist (Storey, 2012). Issues of identity frequently have a spatial expression as social fractures are given material form through spatial divisions. While these spaces might not necessarily be regarded as territories (and may be less obvious, more vaguely defined, and with less clear-cut boundaries) these more micro-scale contexts provide insights into the ways in which geographic space is politicized through attempts to achieve particular outcomes, to exert (or resist) control, and to convey quite clear meanings to those concerned (Delaney, 2005). One manifestation is through the geographies of residential housing. Class divisions are reflected in patterns of residential segregation and processes of gentrification so that poorer or working-class neighborhoods are distinguished from middle-class or more affluent areas. The emergence of gated communities is perhaps the most obvious and clearly demarcated manifestation of this wider and long-standing phenomenon (Le Goix and Webster, 2008). In some contexts, this is accompanied by clustering (or separation) along ethnic lines with the emergence of ghettos and other areas with high concentrations of specific ethnic groups. Discriminatory ideologies and class divisions, often reinforced through planning policies, may contribute toward the exclusion of people from particular areas and the negative stereotyping of such places and their residents. In these areas, segments of the population, often the young, may feel particularly marginalized and one manifestation of such problems of exclusion, marginalization, and victimization may be outbreaks of civil unrest, as in many French cities in recent years (Dikeç, 2007). Some urban areas are sites of territorially based gang behavior (often at the level of the

street or postcode demarcation), a phenomenon that itself feeds into further cycles of stigmatization of places and those who live there (Ralphs et al., 2009).

Alongside residential segregation, recreational spaces such as golf courses, wider tourist complexes, or beach resorts in various parts of the world can be seen as further examples of spatial enclaving designed to enclose the relatively privileged consumers from the excluded local 'other' (Bunnell et al., 2006). These increasingly privatized spaces of consumption have developed in tandem with other controlled arenas such as shopping malls which reflect an erosion of shared public space and its replacement with privatized, exclusionary, and highly regulated environments (Staeheli and Mitchell, 2006). Alongside the creation of various other economically driven territorializations such as export processing zones, a wide array of spaces of exclusion continue to emerge which might be viewed as the outcome of an inherent tendency within capitalist social relations leading toward the ever-increasing privatization of the public domain (Vasudevan et al., 2008).

Feminist perspectives in geography have drawn attention to the nature, construction, and legitimacy of a public–private divide. The 'naturalizing' of specific roles has had a spatial corollary. Long-standing stereotypes of women as home makers whose rightful place is in the domestic realm have given rise to a gendering of space and the use of a territorial frame to restrict women's access to a wider public domain. Such divisions are reinforced through various geographies of fear so that women may be afraid to walk through some areas or frequent specific streets due to fear of assault. These gendered constructions also prompt broader questions about the nature and meanings attaching to 'home' and the social construction of domestic space (Blunt and Dowling, 2006). Similar issues may impact on sexual minorities who may lead more spatially restricted lives due to fear of physical or verbal homophobic abuse. At a yet more micro-level, territorial behavior occurs in all sorts of different settings including the workplace, the office, or the school playground (Thomson, 2005). For example, the demarcation of particular rooms within a house is also a form of territorial behavior through which power (to include or to exclude) is expressed. Parental power is often expressed territorially through, for example, discouraging children entering some rooms in the house such as the kitchen or the parents' bedroom. Similarly, the prohibition of employees from certain rooms or areas within their place of work is an assertion of managerial power. These territorially based constructions of difference can have profound social and spatial impacts raising important questions about power, dominance, and control.

While territorial constructions and strategies are used to impose forms of control they are also resisted. From the activities of revolutionary groups such as the Zapatistas in Chiapas in Mexico to 'reclaim the night' marches, constant struggles are engaged in as attempts are made to control space or to wrest space from other groups. Resistance may be enabled and solidified through senses of place identity which help facilitate forms of action aimed at obtaining particular outcomes. The formation of community or residence groups may reflect feelings of belonging or attachment to a particular place and this territorial affinity may play a mobilizing role. Territorially transgressive acts, whether protest marches or the painting of graffiti, can be employed to reclaim space and to assert basic rights and identities. Responses to homophobia may be reflected spatially through the emergence of spaces where gay people may feel more at ease through being accepted rather than rejected, scorned, or ignored (or worse) by their neighbors (Browne et al., 2007). Just as dominant ideologies can be reinforced through territorial practices, they can also be resisted. Territorial strategies are useful mechanisms in the assertion of identity and spatial concentrations within particular geographic areas make visible people and issues that might otherwise remain unseen. They can be used to draw attention to exclusionary practices and to assert the right to be equal citizens.

Summary

Territories are political geographic entities that have emerged out of particular ways of conceiving geographic space. Territory often constitutes a major component in shaping our everyday lives and in contributing to our senses of place, identity, and belonging. Particular ideologies and social practices are manifested spatially and are reflected in struggles and territorial claims over the use and control of space. Territorial strategies are utilized in conflicts concerned with social power and identity at a range of spatial scales from the global down to the very local. These strategies may be to do with maintaining power or with resisting the imposition of power by a dominant group. Exclusionary ideologies may be consolidated and reinforced through territorial practices, yet they can also be resisted through similar means. The desire to exert territorial control needs to be viewed in its broader social and political context, and not seen simply as a response to an elementary biological urge. Territoriality is a mechanism of power and territorial formation, control, resistance, and transgression are all political phenomena. While it might be supposed that places and territorial boundaries are becoming less and less significant in an increasingly globalized world, the (re)production of territories and the associated creation and contestation of boundaries seem likely to continue to be important for the foreseeable future.

See also: Geography: Overview; Geopolitics; Globalization: Geographical Aspects; Political Geography; Social Geography; Space and Social Theory in Geography; Spatial Thinking in the Social Sciences; State, Geography of; Urban Geography.

Bibliography

Agnew, J., 1994. The territorial trap: the geographical assumptions of international relations theory. Review of International Political Economy 1 (1), 53–80.

Agnew, J., 2005. Sovereignty regimes: territoriality and state authority in contemporary world politics. Annals of the Association of American Geographers 95 (2), 437–461.

Ardrey, R., 1967. The Territorial Imperative. A Personal Inquiry into the Animal Origins of Property and Nations. Collins, London.

Bassin, M., 1987. Imperialism and the nation-state in Friedrich Ratzel's political geography. Progress in Human Geography 11 (4), 473–495.

Blunt, A., Dowling, R., 2006. Home. Routledge, London.

Brenner, N., Elden, S., 2009. Henri Lefebvre on state, space, territory. International Political Sociology 3 (4), 353–377.
Browne, K., Lim, J., Brown, G. (Eds.), 2007. Geographies of Sexualities: Theory, Politics and Practice. Ashgate, Aldershot.
Bunnell, T., Muzaini, H., Sidaway, J.D., 2006. Global city frontiers: Singapore's hinterland and the contested socio-political geographies of Bintan, Indonesia. International Journal of Urban and Regional Research 30 (1), 3–22.
Delaney, D., 2005. Territory. A Short Introduction. Blackwell, Malden.
Delaney, D., 2009. Territory and territoriality. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, Oxford.
Dikeç, M., 2007. Revolting geographies: urban unrest in France. Geography Compass 1 (5), 1190–1206.
Elden, S., 2007. Governmentality, calculation, territory. Environment and Planning D: Society and Space 25 (3), 562–580.
Elden, S., 2009. Terror and Territory. The Spatial Extent of Sovereignty. University of Minnesota Press, Minneapolis.
Elden, S., 2010. Land, terrain, territory. Progress in Human Geography 34 (6), 799–817.
Fall, J.J., 2010. Artificial states? On the enduring geographical myth of natural borders. Political Geography 29 (3), 140–156.
Gottman, J., 1973. The Significance of Territory. University Press of Virginia, Charlottesville.
Gregory, D., 2004. The Colonial Present. Afghanistan, Palestine, Iraq. Blackwell, Malden.
Harvey, D., 2003. The New Imperialism. Oxford University Press, Oxford.
Kolstø, P. (Ed.), 2005. Myths and Boundaries in South-Eastern Europe. Hurst & Company, London.
Le Goix, R., Webster, C., 2008. Gated communities. Geography Compass 2 (4), 1189–1214.
McConnell, F., 2010. The fallacy and the promise of the territorial trap: sovereign articulations of geopolitical anomalies. Geopolitics 15 (4), 762–768.
Morris, D., 1994. The Naked Ape. A Zoologist's Study of the Human Animal. Vintage, London.
Newman, D., 2010. Territory, compartments and borders: avoiding the trap of the territorial trap. Geopolitics 15 (4), 773–778.
Paasi, A., 2008. Territory. In: Agnew, J., Mitchell, K., Toal, G. (Eds.), A Companion to Political Geography. Blackwell, Malden, pp. 109–122.
Penrose, J., 2002. Nations, states and homelands: territory and territoriality in nationalist thought. Nations and Nationalism 8 (3), 277–297.
Raffestin, C., 2012. Space, territory and territoriality. Environment and Planning D: Society and Space 30 (1), 121–141.
Ralphs, R., Medina, J., Aldridge, J., 2009. Who needs enemies with friends like these? The importance of place for young people living in known gang areas. Journal of Youth Studies 12 (5), 483–500.
Reid-Henry, S.M., 2007. Exceptional sovereignty? Guantanamo Bay and the re-colonial present. Antipode 39 (4), 627–648.
Sack, R., 1986. Human Territoriality: Its Theory and History. Cambridge University Press, Cambridge.
Sidaway, J.D., 2007. Enclave space: a new metageography of development? Area 39 (3), 331–339.
Soja, E.W., 1971. The Political Organization of Space. Association of American Geographers, Washington.
Staeheli, L.A., Mitchell, D., 2006. USA's destiny? Regulating space and creating community in American shopping malls. Urban Studies 43 (5–6), 977–992.
Storey, D., 2012. Territories. The Claiming of Space, second ed. Routledge, London.
Thomson, S., 2005. 'Territorialising' the primary school playground: deconstructing the geography of playtime. Children's Geographies 3 (1), 63–78.
Vasudevan, A., McFarlane, C., Jeffrey, A., 2008. Spaces of enclosure. Geoforum 39 (5), 1641–1646.
Wood, D., 1992. The Power of Maps. Routledge, London.

Terror and the Internet

Gabriel Weimann, University of Haifa, Haifa, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Paradoxically, the most innovative network of communication developed by modern Western societies – the Internet with its numerous online networking platforms – now serves the interests of the greatest foe of the West, namely, international terrorism. The growing presence of modern terrorism on the Internet is at the nexus of two key trends: (1) the democratization of communications driven by user-generated content on the Internet and (2) the growing awareness of modern terrorists of the potential of the Internet for their purposes. The Internet has long been a favorite tool of the terrorists. Decentralized and providing almost perfect anonymity, it cannot be subjected to control or restriction, and allows access to anyone who wants it. Large or small, terrorist groups have their own Web sites, using this medium to spread propaganda, raise funds, seduce, radicalize, recruit and train members, communicate and conspire, and plan and launch attacks. Besides thousands of terrorist Web sites, modern terrorists rely on chat rooms, e-groups, forums, social networking, and online platforms like YouTube, Facebook, Twitter, and more. Thus, media-oriented terrorism has presented a tough challenge to democratic societies and their liberal values.

Mass-Mediated Terrorism

> When one says 'terrorism' in a democratic society, one also says 'media.' For terrorism by its very nature is a psychological weapon which depends upon communicating a threat to a wider society. This, in essence, is why terrorism and the media enjoy a symbiotic relationship.
>
> (Wilkinson, 2001: p. 17)

From its early days, terror has combined communicative and psychological aspects. Specifically, the word 'terror' comes from the Latin word 'terrere,' which means 'to frighten' or 'to scare.' The first use of large-scale terrorism was in fact state terrorism, emerging during the popular phase of the French Revolution. The period from June 1793 to July 1794 in France is known as the Reign of Terror or simply 'the Terror.' The upheaval following the overthrow of the monarchy threw the nation into chaos and the government into frenzied paranoia. The Revolutionary Tribunal sentenced thousands to the guillotine. Nobility and commoners, intellectuals, and politicians and prostitutes, all were marched to 'Madame Guillotine' or 'The National Razor.' Estimates of the death toll is placed between 16 000 and 40 000. The executions were conducted before large audiences and were accompanied by sensational publicity thus spreading the intended fear.

Modern terrorists have become exposed to new opportunities for exerting mass psychological impacts as a result of technological advances in communication technologies. They replaced the public executions in Parisian squares with spectacular violent productions performed on the global stages of the mass media. Government officials have tended to blame the media with terrorists' success, as the then British Prime Minister Margaret Thatcher argued: "democracies must find ways to starve the terrorists and hijackers of the oxygen of publicity on which they depend" (as cited by Picard, 1991: p. 50). More recently, Kratcoski has wryly commented that "if one of the elements of terrorism is the wish to obtain publicity for a cause and create propaganda, the media has obviously overreacted in responding to this desire" (Kratcoski, 2001: p. 468).

The emergence of media-oriented terrorism led several communication and terrorism scholars to reconceptualize modern terrorism within the framework of symbolic communication theory. As Jenkins concluded in his analysis of international terrorism, "Terrorist attacks are often carefully choreographed to attract the attention of the electronic media [radio, television] and the international press. Taking and holding hostages increases the drama. The hostages themselves often mean nothing to the terrorists. Terrorism is aimed at the people watching, not at the actual victims. Terrorism is a theater" (Jenkins, 1975: p. 4). Weimann adopted the 'Theater of Terror' metaphor to examine modern terrorism as an attempt to communicate messages through the use of orchestrated violence (Weimann, 1983, 1987, 1988, 1990, 2004, 2008; Weimann and Winn, 1994).

Modern terrorism can be understood in terms of the production requirements of theatrical engagements. Terrorists pay attention to script preparation, cast selection, sets, props, role playing, and minute-by-minute stage management. Just like stage plays or ballet performances, the media orientation in terrorism requires full attention to detail to be effective. One of the most influential strategist of modern terrorism was the Brazilian Carlos Marighela, whose *Minimanual of the Urban Guerrilla* became a sourcebook for many terrorist movements all over the world. In his publications, Marighela (1971a,b) outlined the various uses that can be made of the media:

> The coordination of urban guerrilla action, including each armed action, is the principal way of making propaganda. These actions, carried out with specific and determined objectives, inevitably become propaganda material for the mass communication system. Bank assaults, ambushes, desertion and diverting of arms, the rescue of prisoners, executions, kidnapping, sabotage, terrorism, and the war of nerves, are all cases in point. Airplanes diverted in flight, ships and trains assaulted and seized by guerrillas, can also be solely for propaganda effects.
>
> (1971b: p. 103)

The most powerful and violent performance of modern terrorism was the 11 September 2001, attack on American targets. In November 2001, shortly after the 9/11 attacks, Osama bin Laden discussed the twin attacks. Referring to the kamikaze pilots whom he called 'vanguards of Islam,' bin Laden marveled, "Those young men said in deeds, in New York and Washington, speeches that overshadowed other speeches made everywhere else in the world. The speeches are understood by both Arabs and non-Arabs, even Chinese."(Quote from the translated transcript of a videotape; cit. B. Nacos (2003b: p. 41).) With these remarks, argues Nacos (2003a), bin Laden revealed that he considered terrorism first and foremost as a vehicle to dispatch messages – 'speeches' in his words – and, with respect to the events of 11 September 2001, he concluded that Americans in particular had heard and reacted to the intended communication.

When New Terrorism Met the New Media

The emergence of media-oriented terrorism and its growing sophistication in manipulating the media have led governments and media organizations to react by regulating terrorists' access to the conventional mass media, reducing and censoring news coverage of terrorist acts and their perpetrators, and minimizing the terrorists' capacity for manipulating the media. However, the new media technologies, especially computer-mediated communication and the Internet, allow terrorist organizations to send messages more freely and easily than through other means of communication. al-Qaeda, for example, is using online platforms and advanced communication technologies in a most sophisticated way (Weimann, 2004, 2006). In his assessment of al-Qaeda, Hoffman (2003: p. 4) noted this "long-established sophistication of bin Laden and al-Qaeda's propaganda efforts – employing multimedia vehicles, including prerecorded video and audiotapes, CD-ROMs, DVDs, and the Internet; dramatically choreographed and staged dissemination opportunities; and other mass outreach techniques (e.g., via al Qaeda's phantom Alneda website and those of sympathetic, above-ground radical Islamic sites and organizations)." With the Internet, terrorists can bypass the regulations, limitations, and censorship imposed by traditional media and directly reach hundreds of millions of people, globally.

Terrorist organizations have generated thousands of Web sites that support psychological warfare, propaganda, fund raising, recruitment, coordination, and inner communication. From these terrorist sites, supporters and followers can download multimedia training materials; buy games, T-shirts, and music CDs; and access forums, social networking sites (such as Facebook), and chat services (Weimann, 2006). Most of these sites are extremely dynamic in that they emerge overnight, frequently modify their contents, and then swiftly 'disappear' by changing their URLs, which are later announced via online forums (Weimann, 2006). They are often hosted on free space providers or by unsecured and poorly maintained commercial servers. The level of technical sophistication of the terrorist Web sites has increased dramatically in recent years.

The great virtues of the Internet – ease of access, lack of regulation, vast potential audiences, fast flow of information, and the availability of multimedia platforms – have been converted into the advantage of groups committed to terrorizing societies to achieve their goals. The anonymity offered by the Internet is very attractive for modern terrorists (Rogers, 2003): due to their extremist beliefs and values, terrorists require anonymity to exist and operate in social environments that may not agree with their particular ideology or activities. The Internet provides this anonymity as well as easy access from everywhere with the option to post messages, e-mail, upload or download information – and to disappear into the dark.

Terrorist Web Sites and Forums

It is easy for terrorists to set up a Web site. Take for example, the al-Qaeda's original Web site, alneda.com. In the case of alneda.com, al-Qaeda members created a fictitious organization, 'The Center for Islamic Studies and Research,' a bogus street address in Venezuela, and a free Hotmail e-mail account to contact Emerge Systems, a Web hosting company in Malaysia, US intelligence officials say. The group then wired $87 to a Malaysian bank to pay for the cost of the Web site for a year. Just before alneda.com was pulled off its server (in July 2002), it warned its members that the Federal Bureau of Investigation (FBI), Central Intelligence Agency, and Customs Service were probably monitoring the site. It promised to e-mail members the new address of the site once it was in operation. It also told them they could find the address in chat rooms on other terror sites, such as Hamas' qassam.net. Within years, al-Qaeda launched (either directly or indirectly) thousands of Web sites, forums, chat rooms, and so on. These Web sites are used to disseminate lectures, announcements, videos of terrorist activities, guidebooks, and instruction manuals.

The rise of networked terrorist groups is part of a broader shift to what Arquilla and Ronfeldt (2001a,b) called 'Netwar.' Netwar refers to an emerging mode of conflict and crime at societal levels, involving measures short of traditional war in which the protagonists are likely to consist of small, dispersed groups who communicate, coordinate, and conduct their campaigns in an 'internetted' manner and without a precise central command. Netwar differs from modes of conflict in which the actors prefer formal stand-alone hierarchical organizations, doctrines, and strategies, for example, as in past efforts to build centralized revolutionary movements along Marxist lines. Netwar can be launched by various actors, ranging from NGOs and social activists to violent terrorists.

Terrorist Online Networking

In addition to launching their own Web sites, terrorists can harness the interactive capabilities of the Internet using chat rooms, instant messengers, blogs, video-sharing Web sites, and thousands of postings on Facebook, MySpace, Twitter, and YouTube (Kennedy and Weimann, 2011; Weimann, 2011). As Kohlmann concluded his monitoring of online terrorist presence: "Today, 90% of terrorist activity on the Internet takes place using social networking tools, be it independent bulletin boards, Paltalk, or Yahoo! eGroups. These forums act as a virtual firewall

to help safeguard the identities of those who participate, and they offer subscribers a chance to make direct contact with terrorist representatives, to ask questions, and even to contribute and help out the cyberjihad" (Kohlmann, as cited by Noguchi, 2006; web page).

Online social networking is another mean of attracting potential members and followers. These types of virtual communities are growing increasingly popular all over the world, especially among younger demographics. Youths are especially targeted for propaganda, incitement, and recruitment purposes by jihadist terrorist groups. Predominately Western online communities like Facebook, MySpace, and Second Life and their Arabic equivalents are being used more and more by terrorist groups and their sympathizers.

Social networking Web sites allow terrorists to disseminate propaganda to an impressionable age bracket that might empathize with their cause and possibly agree to join. Many users join interest groups that may help terrorists target users whom they might be able to manipulate. Many social network users accept people as friends whether or not they know them, thereby giving perfect strangers access to personal information and photos. Some people even communicate with the strangers and establish virtual friendships. Terrorists apply the narrow-casting strategy to social networking sites as well. The name, accompanying default image, and information on a group message board are all tailored to fit the profile of a particular social group. The groups also provide terrorists with a list of predisposed recruits or sympathizers. In the same way that marketing groups can view a member's information to decide which products to target to your Web page, terrorist groups can view people's profiles to decide whom they are going to target and how they should configure the message.

As a strategy to distribute jihadist propaganda to a wide range of Muslims and overcome countermeasures, a jihadist posting suggests that administrators of jihadist forums and media organizations create e-mailing groups. This suggestion was posted on the al-Fallujah jihadist forum on 16 March 2009. The mailing group is patterned after the Ansar Mailing Group, an inactive jihadist media distributor that dispatched news of the mujahedeen to users via e-mail. He suggests that other jihadists, too, create such groups to reach the largest possible number of users and that they should remove any obstacle in the registration process that hinders distribution. To this end, the jihadist, in a later posting, provides instructions for creating groups on Google. Another forum participant, pleased with the suggestion, instructs how to create a user account on Yahoo and adds that groups may be created on that service.

A recent report by the Student Rights and the Henry Jackson Society shows that British students are being targeted online by Islamist extremists using social media (Sutton and Stuart, 2012). The study reveals how online platforms are used to promote fear of a Western war against Islam, support for paramilitary violence in Israel, intolerance of nonbelievers and Islam as an obligatory political system for law and governance. Chilling videos of armed insurgents, accompanied by hate-filled speeches from leading al-Qaeda figures, have been posted on Web sites linked to Islamic societies at several leading British universities. The social media were used to share extremist material including video and audio recordings of Abu Ibrahim, designated by the US government as an al-Qaeda linked fighter, recruiter, facilitator, and propagandist; Abdul Rahman Saleem, convicted in 2008 of inciting terrorism overseas during a speech at Regent's Park mosque; and of deceased senior al-Qaeda cleric and propagandist Anwar al-Awlaki.

Terrorist Chat Rooms

Chat rooms and electronic forums enable terrorist groups to communicate with members and supporters all over the world, to recruit new followers, and to share information at little risk of identification by authorities. The free chat room service PalTalk, which includes voice and video capabilities, has become particularly popular with terrorist cells. In one PalTalk chat room, British Islamic militants were found to have set up support forums for the killed leader of the insurgents in Iraq, Abu Musab al-Zarqawi. In one chat room, Arabic-speaking users are sharing personal experiences of fighting Arab-Afghans. In another, relatives of Iraqi insurgents were praising the 'martyrdom' of the terrorists (Kohlmann, 2005). On the alneda.com forum, al-Qaeda members posted comments praising Osama bin Laden, such as "Oh Allah! support your fighting slave Osama bin Laden." Other message boards included threats against global security and reference to the 2005 London bombings. The Web site and forum were hacked into and closed down. But that did not stop al-Qaeda members. They simply started a new forum.

In addition to being used to generate support, chat rooms are used to share tactical information. Jihadist message boards and chat rooms have been known to have 'experts' directly answer questions about how to mix poisons for chemical attacks, how to ambush soldiers, how to carry out suicide attacks, and how to hack into computer systems. One chat room on the PalTalk index, with a name that is slightly altered each time but still identifiable, has been routinely advertised on Jihadi web forums and has been used on a daily basis to post the links for al-Qaeda propaganda videos and terrorist instruction manuals. The forums Qalah, al-Shamikh, Majahden, and al-Faloja are especially popular among terrorist cells, and new recruits are encouraged to refer to the sites to read the jihadist literature. These chat rooms also aim to convince prospective members to join or to stage personal suicide attacks.

According to the Search for International Terrorist Entities' (SITE) special report on Western Jihadist Forums, during 2009 several notable technical changes occurred in many of the jihadist forums (SITE, 2009). Thus, for example, the long offline, prominent English-language jihadist forum, al-Firdaws English, returned on 24 May 2009. The newly resurrected al-Firdaws English forum is open to the public, rather than being password protected. Permitting nonmembers to view discussion and content on the forum is a significant departure for al-Firdaws style, as previous iterations of the forum have been both completely password protected and not open to new membership. The forum administration's decision to open the forum to public observation suggests that they may envision the forum containing less-sensitive information in the future. Despite allowing forum visitors to access threads and read content, al-Firdaws English remains closed to new and prospective members.

The case of Younes Tsouli is especially demonstrative of the resourceful uses of the Internet by Internet-savvy terrorists. Tsouli, more commonly known by his Internet pseudonym *Irhabi 007*, illustrated perfectly how terrorists are using the Internet not just to spread propaganda, but to organize attacks. Between 2003 and the time of his arrest in December 2007, *Irhabi 007* engaged in several instrumental activities on the Internet. In 2003, he began joining various terrorist Internet forums, where he uploaded and published pictures, videos, and instruction manuals on computer hacking. Shortly thereafter his skills were sought out by al-Qaeda leaders who wanted him to provide logistical support for their online operations, and in 2005, Tsouli became the administrator of the extremist Internet forum, al-Ansar, where he began publishing bomb-making instruction manuals and details related to suicide bombing operations. He helped Zarqawi's al-Qaeda faction in Iraq and became a central figure in enabling Zarqawi to reestablish the links between al-Qaeda-affiliated groups after the fall of the Taliban. *Irhabi 007* eventually hacked his way into an unprotected file directory on an Arkansas state government Web site. He then posted propaganda and beheading videos. Cyber-tracking Intelligence immediately noticed *Irhabi 007*'s perfect English and questioned the cybercriminal's nationality. Younis Tsouli was caught in 2006. On his home computer, British investigators found photos of locations in Washington D.C. that had been e-mailed to him by colleagues, which suggested that he was helping to organize a terrorist attack on Capitol Hill. Of course, after Tsouli was caught, other cyberterrorists learned from his mistake.

When Terrorists Tweet

A report released in 2008, by the US Army's 304th Military Intelligence team, examines the possible ways terrorists could use mobile and Web technologies such as the Global Positioning System, digital maps, and Twitter to plan and execute terrorist attacks. The report includes a chapter entitled the 'Potential for Terrorist Use of Twitter,' which expressed the Army's concern over the use of the blogging services. The report says that "Twitter is already used by some members to post and/or support extremist ideologies and pserspectives. Extremist and terrorist use of Twitter could evolve over time to reflect tactics that are already evolving in use by hacktivists and activists for surveillance. This could theoretically be combined with targeting" (The Army's 304th Military Intelligence OSINT Team, 2008: p. 8).

The report lays out three possible scenarios in which Twitter could become a terrorist tool: (1) using Twitter, terrorists can send and receive near real-time updates on the logistics of troops' movements in order to conduct more successful ambushes; (2) one operative with an explosive device or suicide belt uses Twitter to send images of his or her location to a second operative who can use the near actual-time imagery to time the precise moment to detonate the explosive device; and (3) a cyberterrorist operative finds a soldier's Twitter account, hacks into this account, and communicates with other soldiers under the stolen identity.

In his study on 'Online Jihad on Twitter,' Stalinsky (2011a) describes how Twitter is increasingly being used by terrorist organizations and their media outlets, and how their online followers are growing in number. One of the many terrorist groups now utilizing Twitter is al-Qaeda and its followers. After the deaths of Anwar al-Awlaki and Samir Khan were confirmed by al-Qaeda, online jihadi activists reacted by calling for flooding American Web sites and social networking sites including Twitter with their speeches, writings, and videos. A member of the Shumoukh al-Islam forum declared in English: "Together for Islamizing [the] U.S.A. I want to see the American forums, websites, YouTube channels, and Twitter full of Sheikh Anwar al-Awlaki's lectures and videos. It will be a curse chasing the Americans and their dogs." He added in Arabic: "To the lions of uploading and [online] distribution, to the students of Anwar al-Awlaki … I entrust you with supporting your brothers, the *mujahideen*, with this knowledge … we shall spread their actions and the truth … You are the hope of the *ummah*, you are the hope of jihad … Let it be a vengeance raid for our sheikh, al-Awlaki. We want a raid of every American forum, every Facebook page, every Twitter account …" (MEMRI, 2011).

There are multiple ways in which terrorist groups are using Twitter. Among them are sharing videos, 'tweeting' news flashes about attacks and battles, spreading propaganda, sharing operational instructions, and calls for action. One example of such use of Twitter is the British jihadist Anjem Choudary, spokesman for the banned Islam4UK organization, cofounder of al-Muhajiroun. On 19 March 2011, Choudary launched a Twitter account and began using Twitter for disseminating and linking to his videos on YouTube, for announcing collaboration with other jihadi groups, and for calling on his followers in the UK and the US for a return to the Caliphate; for *shari'a* law to replace the US Constitution; for Islam to conquer the White House and the UK; for Islam to take over the world; for Libyans to declare jihad and annex other countries; and for Allah to destroy the enemies of Islam, as well as praise for Anwar al-Awlaki and Osama bin Laden, and many more. Another example is the tweet of a Jihadi Twitter user, which called for readers to contribute and to make efforts to bring about the kidnapping of a US soldier. The item, tweeted on November 2011, followed a $1 million offer by a member of the Saudi royal family and by influential Saudi cleric Awadh al-Qarni for kidnapping Israeli soldiers. It should be noted that al-Qarni tweets prolifically (at http://twitter.com/#!/Dr_alqarnee) and has over 257 000 followers. As noted by Stalinsky (2011a), al-Qarni is only one of many Islamists and jihadists, who have begun using Twitter to spread their ideology. Some of the more notable groups and figures now actively Tweeting include the Taliban, Afghanistan, Abdulqahar Balkhui, Sheikh Yusuf al-Qaradhawi, the Muslim Brotherhood, Ikhwan Syria, Hamas, the Al Qassam Brigade in Palestine, Jihad al-Ansar Media, Nukhbat alI'lam al-Jihadi, Ribat Media Center, Hezbollah, the Somali al-Qaeda-affiliated al-Shabaab al-Mujahedeen, and many others.

Hezbollah, for example, began tweeting on 16 November 2011. In the first month since it launched its Twitter account, Hezbollah has posted 15 254 tweets – that is, over 400 tweets per day – and has 7339 followers. The tweets are in Arabic, French, and English. The tweets included statements by Hezbollah officials as well as news stories from the Middle East. Hezbollah uses Twitter tools such as AddThis, which bills itself as "the world's largest sharing and social insights platform. AddThis helps you integrate sharing tools into your

site, spread your content, and drive social traffic. In-depth analytics provide insight into your audience and their activity."

The Taliban commander Abdul Sattar Maiwandi, who is in charge of the Taliban's official web presence, gave in February 2011, a detailed interview to the Taliban's Arabic-language magazine, al-Somood. In the interview, he boasted of defeating the US in the media field and discussed the Taliban's online media campaign, including its presence on Twitter. As part of an apparent strategy by the Afghan Taliban to broaden their online reach by using social media, the organization has created several accounts on Twitter. The Taliban maintains several official Twitter accounts. The first, the Islamic Emirate of Afghanistan Twitter account, http://twitter.com/#!/alemarahweb, was created on 19 December 2010, and posted 1998 'tweets,' many of them, announcements of attacks on US and coalition forces. Thus, for example, among the tweets posted to this account were those dated 13 September 2011, when Taliban insurgents attacked two symbols of the US diplomatic and military presence in Kabul – the US embassy and NATO headquarters. They included minute-by-minute updates on the attack, detailing the killing of US soldiers. The other Taliban Twitter account, http://twitter.com/#!/ABalkhi, was created on 12 May 2011; it has posted 537 tweets and has 1663 followers. It describes itself as 'The Islamic Emirate of Afghanistan, News, Interviews, Analysis, Statements' and links to the official Taliban Web site. The Taliban's use of Twitter is rather sophisticated; it uses services and applications such as FeedTwit, which sends RSS feeds to Twitter by direct message. This can be used by the Taliban fighters in the field to communicate with each other directly using their phones, rather than via the Internet or e-mail.

al-Qaeda's most important English-language forum Ansar al-Mujahideen English Forum (AMEF) is also very active on Twitter; its Twitter efforts are supported by an entire network of online activists and groups (Stalinsky, 2012). AMEF is considered the primary English-language jihadi forum, disseminating the majority of al-Qaeda's propaganda for the English-speaking West. On 27 October 2011, AMEF opened a Twitter account, and has since then posted over thousand tweets, approximately a dozen per day. The AMEF tweets focus on jihad news, from fronts in Afghanistan, Pakistan, Iraq, Russia, Kazakhstan, Dagestan, Syria, Somalia, Kenya, Libya, Uzbekistan, Albania, Bosnia, and the West. They include links to speeches and videos of al-Qaeda leader Ayman al-Zawahiri, and of Abu Yahya al-Libi, Anwar al-Awlaki, Abu Muhammad al-Maqdisi, and other jihadist leaders. Tweeted by AMEF include: al-Qaeda Central, al-Qaeda in the Arabian Peninsula (AQAP), al-Shabab, al-Qaeda in the Islamic Maghreb (AQIM), Taliban, Tehreek e-Taliban Pakistan, Islamic State of Iraq (ISI), Jund al-Khilafa, Abdullah Azzam Brigades, Boko Haram, Fath al-Islam, Ansar al-Jihad in the Sinai Peninsula, Abu Sayyaf Brigades in the Philippines, Islamic Party of Turkestan, and Western Islamists such as Anjem Choudary of Sharia4UK, Muslims Against Crusaders, Shariah 4 Belguim, Hani Sibai, and others.

Finally, 'Asad al-Jihad 2,' a prominent member of the jihadist online community, confirmed in June 2012 his new Twitter account. This message details the benefits of using Twitter to reach followers and present material without a media spin: "As everyone knows, Twitter is a very important arena for public and private media, and this arena holds the segment of people with whom we can deal and keep in touch. This is an important step in breaking the media barriers that were placed by the enemies of Islam against the people of the truth, and were made to stand between them and the public in the Ummah with their different inclinations" (SITE, 2012).

Terrorists on Facebook

Membership within the international Facebook community has boomed in recent years. Facebook is currently the world's most popular social networking Web site with an estimated 900 million users worldwide. Terrorists have taken note of the trend and have set up profiles as well. There are numerous Facebook groups declaring support for paramilitary and nationalist groups that were designated as terrorist organizations by most Western nations. Before the widespread adoption of social networking platforms like Facebook, individuals had to visit dedicated jihadist Web sites or forums in order to learn about the state of jihad in the world, the status of jihadist groups, and tools and tactics for carrying out jihad. Today, however, one can instead simply go to a Facebook page or a YouTube video to consume jihadist content. Facebook is clearly filling niches once reserved for jihadist official Web sites and forums. The US Department of Homeland Security (DHS) report, 'Terrorist Use of Social Networking Sites: Facebook Case Study,' notes that although terrorists have been using social networking sites for quite some time, their strategies for exploiting Facebook have evolved and that they have learned 'the inherent value in exploiting a nonideological medium' (DHS, 2010). According to the DHS report, terrorists and jihadists use Facebook for numerous purposes: as a way to share operational and tactical information, including bomb recipes and weapons maintenance; as a gateway to extremist forums; as a media outlet for propaganda; and as a source of remote reconnaissance for targeting purposes.

In December 2008, jihadist started online calls for 'Facebook Invasion.' The first call was posted on al-Fallujah, a password-protected jihadist forum. The author hopes to reach Muslims and Americans, and 'fight the media offensive on jihadist media, its forums, and its Web sites.' He provides a synopsis of Facebook and its growth since founding, noting the number of its users and reach. The posting also explains how jihadists can use Facebook and presents images detailing the registration process, adding friends, and setting up groups. The goals of the 'Facebook Invasion' according to the text are "(1) Reaching the vast base of Muslims who subscribe to Facebook; (2) Encouraging the brothers to create new media tools on the Internet in order to support jihadist media; (3) Creating a solid foundation on Facebook and shedding light on it as a media tool to reach people; (4) Moving from the elite community (jihadist forums and sites) to Muslims in general, participating and interacting with them; (5) Developing media work and encouraging its creativity, renewal, flexibility, and change; (6) Reaching big portions among the Crusaders, broadcasting losses of their armies, revealing the lies of their commanders, and urging Muslims to fight them." Later that month, a group calling itself, 'Knights of Support Invasion 2,' appeared on Facebook. The group is comprised of seven

brigades each tasked with distributing and promoting Salafist doctrine and mujahideen propaganda. The seven brigades include a Shariah Brigade, English Media Brigade, Martyrs and Prisoners Brigade, and Media and Design Brigade.

A year later, in December 2009, determined to strengthen the 'Facebook Invasion' and ensure its success, a posting on al-Fallujah jihadist explained the next stage of the 'Facebook Invasion' and how to fix errors discovered in prior stages. The user explaining the third stage, 'Omar Abdul Hakim,' suggests how a jihadist can start a group and recruit members, such as finding Muslims who share similar beliefs and interests, and slowly introducing them to jihadist propaganda. He gives specific advice how to communicate with Facebook users, providing a form letter that they may send, and also suggests steps to take while forming the group and what to do should the group be deleted, so as to resume activity.

Perhaps more than any other jihadist group, Hezbollah has learned to capitalize on Facebook's accessibility and global usage to reach a wider audience with its propaganda. As early as early 2008, Hezbollah had established a presence on Facebook, which has since ballooned into hundreds of pages, available in no fewer than nine languages, and with fans numbering in the tens of thousands. This network of Hezbollah pages on Facebook includes groups dedicated to Hezbollah itself as well as its leaders like Secretary General Hassan Nasrallah and helps to spread Hezbollah propaganda across the world. SITE Intelligence Group (2011a) documented the strong presence of Hezbollah on Facebook with dozens of pages dedicated to the group, reaching out to more than 100 000 fans. The international reach of Hezbollah on Facebook is evident in pages like the Indonesian-language 'Hizbullah,' founded 29 March 2010, which has attracted hundreds of fans, the Turkish-language 'Hizbullah,' founded 10 November 2010, and many more Hezbollah Facebook pages in French, Hebrew, Albanian, Spanish, Urdu, and Persian. The wide majority of Hezbollah Facebook pages, however, are in Arabic or English. Pages dedicated to Hezbollah leaders are numerous and very popular. An Arabic-language page dedicated to the deceased Mohammed Hussein Fadlallah, considered the spiritual mentor of Hezbollah, was founded in March 2008 and has gathered 23 400 fans till 2011. An English-language equivalent called 'Grand Ayatollah Sayed Fadlullah,' also founded in early 2008, has attracted 26 600 fans. Hezbollah's Secretary General Hassan Nasrallah has inspired many Facebook users. For instance, the English-language page 'Sayyed Hassan Nasrallah' has accumulated almost 1700 fans since its launching on 9 January 2011, while the Arabic-language 'Mercy Upon Sayyed Hassan Nasrallah' has nearly 1200 fans and was founded 23 February 2011.

The Palestinian Islamic Jihad (PIJ) is another designated terrorist organization that at least since early 2009 has been busy building a Facebook presence. PIJ has spawned a large network of Arabic-language Facebook pages, which has attracted fans across nearly 50 Facebook pages. PIJ's Facebook pages include those dedicated to the group itself, its militant wing – al-Quds Brigade, and its founder, Fat'hi al-Shaqaqi. The pages feature official PIJ propaganda, such as videos and flyers, as well as incitements to jihad, military communiqués, and claims of attacks. The pages described here are "In Memory of the Martyrdom of Dr Fat'hi al-Shaqaqi," "Who Lives the Martyrdom, It is al-Quds Brigade," "Lovers of the Movement of Islamic Jihad – Lebanon," "The Islamic Jihad Movement (Nasirat Neighborhood)," "The Islamic Jihad Movement in Palestine," and "Lovers of the Martyrs of al-Quds Brigade."

In response to the killing of bin Laden on 1 May 2011, a popular and active 'We Are All Usama Bin Laden' fan page on the Facebook emerged. Before the page was removed from Facebook, the group contained nearly 30 000 members who posted hundreds of messages each hour. The page featured videos, poems, and reactions commemorating Osama bin Laden, as well as incitements for retaliation. Facebook users active on the page shared their reactions to bin Laden's death, giving brief eulogies, and encouraging readers to activism against Israel. Such networking on Facebook facilitates connections between groups and causes through linked content. Yet, while the original 'We Are All Usama bin Laden' page has been removed, hundreds of new pages replaced it, calling for attacks in retaliation for the death of the leader. Supporters of bin Laden and al-Qaeda have flooded Facebook with hundreds of fan pages and groups praising Osama bin Laden, nearly two hundred of which share the name 'We Are All Usama bin Laden.' These pages range in membership from less than ten up to several thousand Facebook users.

Facebook pages are also used to mirror terrorist Web sites: in 2011, a Facebook page was used for mirroring content found on the Arabic-language Ansar al-Mujahideen jihadist forum (Ansar). Ansar is a distribution point for al-Qaeda and affiliated jihadists. The Facebook page, 'Ansar al-Mujahideen Network,' thus also hosts the most recent releases from jihadist groups such as the Somali Shabaab al-Mujahideen Movement and the Afghan Taliban, and media groups including the Global Islamic Media Front and al-Nokhba Jihadi Media. Since its founding nearly 3 months ago, the page has amassed more than 3000 fans and has thousands of posts. To make sure that a new forum post is mirrored on Facebook, the page administrator set up an automated system using the 'RSS Graffiti' tool to link content automatically from Ansar. In this way, the Facebook page is directly updated from Ansar. According to the SITE report (2011b), the 'Ansar al-Mujahideen Network' Facebook page has ballooned in popularity, with thousands becoming fans of the page. The material on the group is exclusively comprised of jihadist material, with official communiqués and al-Qaeda releases directing readers to material published on the affiliated jihadist forum. Therefore, the Facebook page reflects al-Qaeda's ideology of global jihad and shares this propaganda with a larger audience, allowing them to reach potential new forum members as well as individuals who may be unable or unwilling to go to a jihadist forum.

The use of Facebook by terrorists is rather effective: in 2009, Deputy Director for Intelligence at the National Counterterrorism Center, Andrew Liepman, reported to Congress that the FBI is tracking a few Somali Americans from the Minneapolis area, who were reportedly recruited for the purpose of starting a US terrorist cell of the al-Shabaab faction through Facebook. Analysis of all recent cases of lone-wolf terrorists reveals that all these individuals were using new social media (e.g., Facebook, Twitter, YouTube, and more). The accused Fort Hood shooter, Nidal Hasan, exchanged Internet messages with Awlaki, while the Frankfurt terrorist Arid Uka, who confessed to the 2011

killing of two US airmen at the Frankfurt airport, had links with radical groups online using Facebook. For lone wolves, Facebook communication provides the needed social bonding, a (virtual) community, and a source of guidance, support and moral backing. The 2010 FBI arrest of Baltimore resident Antonio Martinez, aka Muhammad Hussain, who expressed his wish to die as a martyr on his Facebook page, highlights the use of Facebook as a tool for effective recruitment and indoctrination by jihad. "When are these crusaders gonna realize they can't win?" Martinez boasted on his Facebook page on 4 August 2010 "How many more lives are they willing to sacrifice. Allahhuakbar."

Terrorists can use this online social networking to monitor military personnel. In 2008, the Canadian Defense Department and the British Secret Service, M15, requested troops to remove personal details from social networking sites because of alleged monitoring by al-Qaeda operatives. US personnel are also warned about posting certain details or photos on their profile pages. Even if the information does not give details about the logistics of troop movements, it could potentially endanger the friends and relatives of the military and security personnel. Many soldiers unwittingly post detailed information about themselves, their careers, family members, date of birth, present locations, and photos of colleagues and weaponry. Canadian troops have been asked to exclude any information from their profiles that might even link them with the military. A report from the Lebanese capital of Beirut later that year stated that Hezbollah had been monitoring Facebook to find potentially sensitive information about Israeli military movements and intelligence that could be harmful to the national security of Israel. The report quoted an Israeli intelligence official saying that "Facebook is a major resource for terrorists, seeking to gather information on soldiers and IDF [Israel Defense Forces] units and the fear is soldiers might even unknowingly arrange to meet an Internet companion who in reality is a terrorist" (cited by Weimann, 2010: p. 25).

As terrorist groups increasingly use Facebook and other social networking Web sites to distribute propaganda and promulgate their ideology, they are also looking for ways to maintain their presence on these sites and not have their pages closed and accounts suspended: on 22 April 2011, a guide on how to 'effectively post' on Facebook appeared in Shumukh al-Islam forum. The 23-page guide opens with the following introduction:

> The Facebook website has become an important platform for spreading the true Islamic call that is represented in establishing the soul of jihad, raising up the Ummah, and getting out of weakness to strength and from loving the world and hating death to loving martyrdom and sacrifice in the Cause of Allah. However, the developments that have been integrated into the website have made posting on it to be more difficult than before. Accounts are being frozen after posting on one or two pages, thus letting our efforts go to waste. But thanks to Allah, no matter what cunning the owners of the site may try, Allah's cunning wins, and its result comes after relying on Allah and taking the means. Therefore, praise be to Allah, your brothers in the Shumukh al-Islam Network were able to overcome these obstacles, and this is what has been explained in this guidebook. We ask Allah on your behalf and ours to grant acceptance and martyrdom in His Cause.

The author, 'Abu Musab al-Dahik,' wrote that the problem is Facebook's antispam filter, and to circumvent the script, which will freeze an account that posts the same material to multiple pages, he suggested the user change the text of what they post. For his examples, he posted a video from the al-Qaeda-affiliated ISI to Facebook pages for Egypt and Gaza.

Yet, terrorists are well aware of the risks involved. A member of a Jihadi forum in English issued a warning, reminding readers that a Facebook network would allow security agencies to trace entire groups of jihadists, arguing:

> Don't make a network in Facebook … Then Kuffar will know every friend you have or had in the past. They will know location, how you look, what you like, they will know everything! Join Facebook if you want and use it to keep in touch with friends and brothers far away but not as a network.
> (SITE, 2009)

YouTube and They Tube

YouTube was established in February 2005 as an online repository facilitating the sharing of video content. YouTube claims to be the 'the world's most popular online video community.' According to Web site monitoring, YouTube's viewership in 2012 exceeded 4 billion video views per day and 1 h of video uploads per second. 70% of the traffic on YouTube comes from outside the US. While the first video on YouTube enjoyed 1.96 million views, the current most popular video (Justin Bieber's 'Baby') exceeded 521 million views. The exposure rates are even higher among younger people who spend many hours every day on YouTube watching, downloading, and uploading video clips. YouTube has faced criticism over the offensive content in some of its videos. All YouTube users can upload videos of up to 15 min duration each. Users who have a good track record of complying with the site's guidelines may be offered the ability to upload videos up to 12 h length. The uploading of videos containing defamation, pornography, and material encouraging criminal conduct is prohibited by YouTube's terms of service. In December 2010, YouTube added 'promotes terrorism' to the list of reasons that users can give when flagging a video as inappropriate. In October 2010, US Congressman Anthony Weiner urged YouTube to take down from its Web site videos of imam Anwar al-Awlaki, the Jihadi propagandist, saying that by hosting al-Awlaki's messages, "We are facilitating the recruitment of homegrown terror." British security minister Pauline Neville-Jones commented: "These Web sites would categorically not be allowed in the U.K. They incite cold-blooded murder, and as such are surely contrary to the public good."

Terrorist groups realized the potential of this easily accessed platform for the dissemination of their propaganda and radicalization videos. Terrorists themselves praised the usefulness of this new online apparatus; for example, convicted terrorist Younis Tsouli (so-called *Irhabi007*) testified (cited by S. O'Neil, 2007): "A lot of the funding that the brothers are getting is coming because of the videos. Imagine how many have gone after seeing the videos. Imagine how many have become shahid [martyrs]". In 2008, jihadists suggested a 'YouTube Invasion' to

support jihadist media and the administrators of al-Fajr-affiliated forums. This suggestion was posted on al-Faloja, a password-protected jihadist forum, on 25 November 2008. The posting provides a synopsis of the YouTube site and its founding, and notes its use by the US President-elect Barack Obama during his presidential campaign, and by others. YouTube is argued to be an alternative to television and a medium that allows jihadists to reach massive, global audiences. The message even instructs how jihadists should cut mujahedeen videos into 10-min chunks, as per YouTube's requirements, and upload them sequentially to the site. "I ask you, by Allah, as soon as you read this subject, to start recording on YouTube, and to start cutting and uploading and posting clips on the jihadist, Islamic, and general forums," said the posting on the jihadist Web site. "Shame the Crusaders by publishing videos showing their losses, which they hid for a long time."

Today, all terrorist groups have propaganda videos on YouTube. In 2008, Hamas allegedly launched its own video-sharing Web site (AqsaTube), although the group denied ownership of the site. AqsaTube, in addition to choosing a similar name, was designed to look just like YouTube and even copied its logo. Once certain Internet providers refused to host the Web site, Hamas launched a PaluTube and TubeZik. These videos are not meant to target local audiences alone but also to reach worldwide publics. More recent videos posted on these video-sharing Web sites are dubbed in English or have English subtitles. Hundreds of terrorist videos are uploaded to YouTube each day and because of the enormous volume of content, YouTube says it is impossible to prescreen each video before it is posted on the site. This caused Senator Lieberman to send, in May 2008, an official complaint to Google, the owner of YouTube, saying, "Google continues to allow the posting of videos by organizations the State Department has designated as Foreign Terrorist Organizations. No matter what their content, videos produced by terrorist organizations like al Qaeda, that are committed to attacking America and killing Americans, should not be tolerated. Google must reconsider its policy." The protest was not very effective: in 2010, Representative Anthony Weiner sent a letter to YouTube CEO Chad Hurley, calling for the removal of videos featuring al-Awlaki, dubbed the 'bin Laden of the Internet,' whose videos aim to find new recruits and promote radical Islamic extremism and violent jihad against Americans. al-Awlaki, has appeared in over 700 videos on YouTube with a combined 3.5 million views. Although YouTube, following government requests, removed Awlaki's videos, a simple keyword search for 'al-Awlaki' conducted in June 2012, brought up about 8000 video clips, most of them the audio and video recordings of Awlaki.

Propaganda videos from various al-Qaeda spinoffs and franchises are easily available, posted by AQAP, Al Qaeda in Yemen, AQIM, pro-Taliban groups, and more. The process is quite simple: the terrorists and their sympathizers or propagandists first obtain new videos from file-sharing sites. The videos are then uploaded to jihadist bulletin boards and web forums; users of these forums then typically upload the videos to YouTube and other commercial sites for mainstream consumption. Even when YouTube does remove videos after complaints, they are often reuploaded. Thomas Hegghammer, senior research fellow at the Norwegian Defense Research Establishment and editor of the Jihadica blog, notes that

> YouTube is extremely important for the dissemination of propaganda videos. I think YouTube is much more influential than bulletin boards because it reaches a much wider audience. Websites that specialise in jihadi propaganda are self-selecting in the sense that the people who access them are interested in jihadi propaganda to begin with. On YouTube, on the other hand, you can come across jihadi videos by accident, for example while searching for more moderate sermons. Basically, having jihadi propaganda on YouTube dramatically lowers the barrier of access.
>
> (Cited in The Java Report, 2010)

A study conducted by Conway and McInerney analyzed the online supporters of jihad-promoting video content on YouTube, focusing on those posting and commenting on martyr-promoting material from Iraq (Conway and McInerney, 2008). The findings suggest that the majority are younger than 35 years of age and reside outside the Middle East and North Africa, with the largest percentage of supporters located in the United States. As the researchers concluded (2008: p. 117): "What is clearly evident, however, is that jihadist content is spreading far beyond traditional jihadist websites or even dedicated forums to embrace, in particular, video sharing and social networking – both hallmarks of Web 2.0 – and thus extending their reach far beyond what may be conceived as their core support base in the Middle East and North Africa region to Diaspora populations, converts, and political sympathizers."

Terrorists and their followers are very resourceful about bypassing YouTube rules and bans. In April 2001, members of the password-protected English-language AMEF shared strategies for evading YouTube censorship of materials promoting jihad. Online responses provided techniques for rendering jihadist content less conspicuous to YouTube administrators and individuals who search for jihadist content to report, and also gave strategies to make creating new accounts easier and minimize the loss of contacts and information after accounts are removed. One suggestion was to keep two separate accounts, one for collecting previously uploaded material for friends and a second for uploading jihadist videos.

In recent years, Jihadi groups have begun free web hosting, where content, including large videos, can be uploaded anonymously and at no cost. One such 'invasion' was to the Internet Archive. Founded in 1996, the Internet Archive nonprofit organization whose stated mission is 'universal access to all knowledge.' The Internet Archive is recognized by the state of California as a library, and is a member of the American Library Association. As Stalinsky (2011b) revealed, "Al-Qaeda and other groups are continually and increasingly using the Internet Archive to spread their propaganda/recruitment messages, including video and audio recordings of speeches, attacks, publications, and much more. They often include content encouraging, and urging, terrorist attacks against America … Content is uploaded to and downloaded from the Internet Archive by jihadists daily."

Conclusion

Paradoxically, the most innovative network of communication developed by the West with its numerous online networking platforms – now serves the interests of the greatest foe of the West, namely, international terrorism. Much of the original online terrorist content used to be one-directional, text-based, in the form of traditional Web sites with a heavy reliance on text and messages posted on forums. However, technological advances, particularly the increased availability of sophisticated, but cheap and easy to use, video capturing hardware (e.g., hand-held digital video cameras, mobile telephones, etc.,) and interactive online networking platforms (e.g., Facebook) changed terrorist online communications. The global community created by social networks and interactive forums on the Internet is advancing cultural awareness and reconciliation efforts, but it is also advancing terrorists' goal to share its extremist messages to global audiences. By using these online communities to their advantage, not only can terrorists promote global paranoia, share their messages with sympathizers and obtain donations, but also create more terrorists. The Internet has provided terrorists with a whole new virtual realm to conduct their most sinister transactions. Terrorist groups are no longer confined to specific regional boundaries. Now terrorist networks can recruit the members located in any part of the globe. A person in the United States can literally take a terrorist training course within the privacy of their bedroom.

The interactive capabilities of the Internet, like chat rooms, social networking sites, video-sharing sites, and online communities, allow terrorists to assume an offensive position. Instead of waiting for web-surfers to come across their Web sites and propaganda materials, terrorists can now lure targeted individuals to the sites. The best way to persuade individuals to support your cause is to create an emotional, psychological, or intellectual bond with the targeted individual. The online social networking provides terrorists with an ideal platform to attract, seduce, teach, radicalize, train, and activate individuals all over the world.

Media-oriented terrorism has presented a tough challenge to democratic societies and their liberal values. The threat is not limited to media manipulation and psychological warfare launched by terrorists; it also includes the danger of restrictions imposed on the freedom of the press and freedom of expression by those who try to fight terrorism. These developments could wind up having profound implications for democracies and their values, adding heavy prices in terms of civil liberties to the destructive effects of terrorism.

See also: Internet and Social Media: Anthropological Aspects; Law and Terrorism in the US and Beyond; Social Movements and Political Violence; Terrorism; War and Terrorism, Geography of.

Bibliography

Alexander, Y., 1978. Terrorism, the media and the police. Journal of International Affair 32 (1), 101–113.

The Army's 304th Military Intelligence OSINT Team, 2008. Al Qaida-like Mobile Discussions & Potential Creative Uses. http://www.fas.org/irp/eprint/mobile.pdf.

Arquilla, J., Ronfledt, D., 2001a. The advent of netwar (revisited). In: Arquilla, J., Ronfeldt, D. (Eds.), Networks and Netwars: The Future of Terror, Crime, and Militancy. RAND Corporation, Washington, DC, pp. 1–25.

Arquilla, J., Ronfledt, D., 2001b. Networks, netwars, and the fight for the future. First Monday 6. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/889/798.

Bassiouni, M.C., 1981. Terrorism, law enforcement, and the mass media: perspectives, problems, and proposals. Journal of Criminal Law and Criminology 72 (1), 1–51.

Bell, J.B., 1975. Transnational Terror. American Enterprise Institute for Public Policy Research, Washington DC.

Bell, J.B., 1978. Terrorist scripts and live-action spectaculars. Columbia Journalism Review 17 (1), 47–50.

Barnhurst, K.G., 1991. The literature of terrorism. In: Alali, A.O., Eke, K. (Eds.), Media coverage of Terrorism: Methods of Diffusion. Sage, Newbury Park, CA, pp. 112–137.

Conway, M., 2006. Terrorism and the internet: new media – new threat? Parliamentary Affairs 59, 283–298.

Conway, M., McInerney, L., 2008. Jihadi video & auto-radicalization: evidence from an exploratory YouTube study. Intelligence and Security Informatics 5376, 108–118.

DHS, 2010. Terrorist use of social networking: facebook case study. Public Intelligence. http://publicintelligence.net/ufouoles-dhs-terrorist-use-of-social-networking-facebook-case-study/.

Dobson, C., Paine, R., 1977. The Carlos Complex: A Pattern of Violence. Hodder & Stoughton, London.

Dowling, R.E., 1986. Terrorism and the media: a rhetorical genre. Journal of Communication 36 (1), 12–24.

Galtung, J., Ruge, M.H., 1970. The structure of foreign news. In: Tunstall, J. (Ed.), Media Sociology. Constable, London, pp. 259–298.

Hoffman, B., 2003. Al Qaeda, trends in terrorism, and future potentialities: an assessment. Studies in Conflict and Terrorism 26 (6), 429–442.

The Java Report, 13 November 2010. YouTube: Where Anwar al-Awlaki Promotes Terror 24/7. http://mypetjawa.mu.nu/archives/204930.php.

Jenkins, B., 1975. International Terrorism. Crescent Publication, Los Angeles, CA.

Jenkins, B., 1983. Research in terrorism: areas of consensus, areas of ignorance. In: Eichelman, B., Soskis, D., Reid, W. (Eds.), Terrorism: Interdisciplinary Perspectives. American Psychiatric Association, Washington, DC, pp. 153–177.

Karber, P., 1971. Urban terrorism: baseline data and conceptual framework. Social Science Quarterly 52, 527–533.

Kennedy, J., Weimann, G., 2011. The strength of weak terrorist ties. Terrorism and Political Violence 23 (2), 201–212.

Kohlmann, E., 2 August 2005. Al Qaeda and the Internet. Washingtonpost.com video report. The Washington Post.

Kratcoski, P.C., 2001. Terrorist victimization: prevention, control, and recovery. Studies in Conflict and Terrorism 24 (7), 467–473.

Laqueur, W., 1976. The futility of terrorism. Harper's Magazine 252 (1510), 99–105.

Liepman, A., March 11, 2009. Violent islamist extremism: al-shabaab recruitment in america. Hearing Before the Senate Homeland Security and Governmental Affairs Committee.

Marighela, C., 1971a. For the Liberation of Brazil. Pelican, Harmondsworth.

Marighela, C., 1971b. Minimanual of the urban guerrilla. In: Mallin, J. (Ed.), Terror and the Urban Guerrilla: A Study of Tactics and Documents (In Appendix). University of Miami Press, Coral Gables, FL.

MEMRI, 11 October 2011. English-Language Jihadi Forum Publishes Eulogy for Anwar al-awlaki, Samir Khan; Online Jihadi Calls for 'Vengeance Raid' on American Websites. Special Dispatch No. 4197. http://www.memri.org/report/en/0/0/0/0/0/0/5716.htm.

Nacos, B., 1994. Terrorism and the Media: From the Iran Hostage Crisis to the World Trade Center Bombing. Columbia University Press, New York.

Nacos, B., 2003a. The terrorist calculus behind 9-11: a model for future terrorism? Studies in Conflict & Terrorism 26, 1–16.

Nacos, B., 2003b. Mass-mediated Terrorism. The Central Role of the Media in Terrorism and Counterterrorism. Rowman & Littlefield Publ., Lanham, MD.

Nacos, B., 2006. Terrorism and media in the age of global communication. In: Hamilton, D.S. (Ed.), Terrorism and International Relations. Center for Transatlantic Relations, Washington, DC, pp. 81–102.

Noguchi, Y., 19 April 2006. Tracking Terrorists Online, Washingtonpost.Com Video Report. http://www.washingtonpost.com/wp-dyn/content/discussion/2006/04/11/DI2006041100626.html.

O'Neil, S., 26 April 2007. Terrorist 007 'was internet propagandist for al-Qaeda'. The Times. http://business.timesonline.co.uk/tol/business/law/article1706787.ece.

Picard, R.G., 1991. News coverage as the contagion of terrorism: dangerous charges backed by dubious science. In: Alali, A.O., Eke, K. (Eds.), Media coverage of Terrorism: Methods of Diffusion. Sage, Newbury Park, CA, pp. 49–62.

Rogers, M., 2003. The psychology of cyber-terrorism. In: Silke, A. (Ed.), Terrorist, Victims and Society. John Wiley & Sons, Chichester, pp. 77–92.

Rodrigo, M., 1991. Los medios de comunicación ante el terrorismo. The Media of Communication Confront Terrorism. Icariia, Barcelona.

Shactman, N., 18 December 2008. Online jihadists plan for 'Invading facebook,'. Wired. http://www.wired.com/dangerroom/2008/12/online-jihadist/.

Signorielli, N., Gerbner, G., 1988. Violence and Terror in the Mass Media: An Annotated Bibliography. Greenwood Press, New York.

SITE Intelligence Group, June 2009. Western jihadist forums. InSITE Magazine.

SITE Intelligence Group, 2011a. Social Network Jihad: Hezbollah's Capitalization of Facebook. http://news.siteintelgroup.com/component/customproperties/show/showcp_post_type[0]=social_network_jihad&cp_groups=&cp_leaders=&cp_text_search=&submit_search=Search.

SITE Intelligence Group, 2011b. Social Network Jihad: Facebook Used to Mirror Ansar al-Mujahideen Jihadist Forum. http://news.siteintelgroup.com/component/customproperties/tag/Post%20Type-Social%20Network%20Jihad.

SITE Intelligence Group, 2012. Prominent Jihadist 'Asad al-Jihad 2' Joins Twitter. http://news.siteintelgroup.com/component/customproperties/show/showcp_post_type[0]=jihadist_news&cp_groups=&cp_leaders=&cp_text_search=&submit_search=Search.

Stalinsky, S., November 2011a. Deleting Online Jihad on Twitter: The Case of British Jihadi Anjem Choudary – Tweeting for the Caliphate and the Conquest of the White House. MEMRI's Jihad and Terrorism Threat Monitor. http://www.memri.org/report/en/0/0/0/0/0/0/5779.htm.

Stalinsky, S., August 2011b. Al-qaeda, Jihadist Infest the San Francisco, California-Based 'Internet Archive' Library. MEMRI's Inquiry & Analysis Series Report No. 724. http://www.memri.org/report/en/print5576.htm.

Stalinsky, S., 20 January 2012. Now Tweeting: Ansar al-mujahideen English Forum (AMEF) – the Main al-qaeda English-Language Forum – Spreading Online Jihad, Inciting and Instructing Activists to Carry Out Terrorist Acts, Obstructing Counter-terrorism Measures in the West. MEMRI's Inquiry & Analysis Series Report No.787. http://www.memri.org/report/en/0/0/0/0/0/0/6008.htm.

Sutton, R., Stuart, H., 2012. Challenging Extremists. The Henry Jackson Society, London.

Weimann, G., 1983. The theater of terror: the effects of press coverage. Journal of Communication 33 (1), 38–45.

Weimann, G., 1987. Media events: the case of international terrorism. Journal of Broadcasting & Electronic Media 31 (1), 21–39.

Weimann, G., 1988. Mass mediated theater of terror: must the show go on? In: Bruck, P. (Ed.), The News Media and Terrorism. Carleton University Press, Ottawa, pp. 1–22.

Weimann, G., 1990. Redefinition of image: the impact of mass-mediated terrorism. International Journal of Public Opinion Research 2 (1), 16–29.

Weimann, G., 2004. The theater of terror: the psychology of terrorism and the mass media. In: Danieli, Y., Brom, D., Waizer, J. (Eds.), The Trauma of Terror: Sharing Knowledge and Shared Care. The Haworth Maltreatment & Trauma Press, New York, pp. 379–390.

Weimann, G., 2006. Terror on the Internet: The New Arena. The New Challenges. United States Institute for Peace (USIP) Press, Washington, DC.

Weimann, G., 2008. The psychology of mass-mediated terrorism. American Behavioral Scientist 52 (1), 69–86.

Weimann, G., 2010. Terrorist facebook: terrorist and online social networking. In: Last, M., Kandel, A. (Eds.), Web Intelligence and Security. NATO Science for Peace and Security Series, pp. 19–30.

Weimann, G., April 2011. Al Qaeda has sent you a friend request: terrorists using online social networking. In: Paper Presented at annual conference of the Israeli Communication Association, Haifa, Israel.

Weimann, G., Brosius, H.B., 1991. The newsworthiness of international terrorism. Communication Research 18 (3), 333–354.

Weimann, G., Winn, C., 1994. The Theater of Terror: Mass Media and International Terrorism. Longman, New York.

Wilkinson, P., 2001. Terrorism versus Democracy. Frank Cass, London.

Terrorism

Stephen Vertigans, Robert Gordon University, Aberdeen, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Crenshaw, volume 23, pp. 15604–15606, © 2001, Elsevier Ltd.

Abstract

Terrorism can be considered to be a violent strategy which aims to provoke public and political reactions in order to bring about political change. Such politically aggressive approaches have a long history that predates today's experiences. Throughout this history, the beliefs that terror groups hold, the locations where they are formed, the reasons that people join and remain within groups, and ultimately why they leave have been diverse. Yet contemporary public, academic, and political attention and understanding of terrorism is frequently viewed through the prism of the September 2001 attacks on America. In turn the American-led response under the umbrella of the 'war on terror' has been instrumental in changing how counterterrorism is considered and the legitimacy of government-led attacks.

Introduction

Terrorism is an intentional form of violence that involves targeted attacks which are designed to intentionally scare, intimidate, to 'terrorize' for a political purpose. Attacks are usually directed at immediate targets such as people, events, or institutions and a wider audience whom protagonists aim to intimidate, shock or threaten. By so doing, terror groups intend to arouse support for their political discourse while undermining rival, usually dominant, ideologies as part of an overarching strategy to win concessions or defeat the declared enemy. These forms of political violence can be relatively inexpensive and undertaken by small numbers of activists. Hence acts of terrorism are considered to be a form of asymmetrical warfare, committed by marginalized groups who often resort to such tactics because they lack more conventional political resources and avenues for engagement. However, this opinion overlooks the strategic, clandestine use of violence by national governments or their complicit supporters. Such government actions are usually classified under the reactive heading of 'counterterrorism.'

Since September 2001, perceptions of terrorism have been dominated by impressions of contemporary 'Islamic' groups. Yet terrorism is a well established strategy that has been implemented by groups across the ideological spectrum and historical periods.

Defining Terrorism

Achieving a universally recognized definition of terrorism has proved impossible and has led to multiple interpretations. Schmid's (1984) seminal study identified 109 different definitions and distinct arenas of classifications such as state, public, and academic. He concluded that none of the definitions would satisfy the multitude of interested parties. There are a number of reasons for this morass. The main factors revolve around moral and power impositions which, Crenshaw (2011) notes, have pejorative connotations. The normative struggle over semantics has correlated with the shift in the struggle over propaganda and the wider perception of terrorism as an illegitimate strategy and the appropriateness of targets. Weinberg et al. (2012) trace this to Menachem Begin's decision, when leader of the Zionist Irgun, to describe followers as 'freedom fighters.' In the process he unintentionally contributed to one of the unwritten rules that articles on defining terrorism must refer to the adage that "one man's terrorist is another's freedom fighter." The credibility of such an application is generally easier if the cause is connected to a struggle for independence and/or is against authoritarian regimes and the denial of liberties.

Consequently the authority and ability to attribute actions as terrorism, counterterrorism, or tactic of war helps to shape the extent to which those actions are viewed as justifiable. The successful application of the label of terrorism can then be integral into how those groups are considered. In essence, the strategy and the negative connotations would become the defining characteristic of the group. This places groups, and conversely the labeling governments, in ambiguous positions. At one level, being classified according to the act of violence strengthens the impression of groups' capabilities and helps in their attempts to 'terrorize.' Yet the inflation of the violent, unacceptable threat isolates the incident from the longer term political aims. The focus is upon the emotive reactions of the targeted audience rather than their considered reflection. In so doing, groups are defined by the perceived irrational outburst of violence and not the rationality of their discourse. The wider empathetic interest in the groups' agenda, which is required if they are to extend their opposition and achieve long-term aims, is therefore compromised. By comparison, during periods when the term was much less value-laden, earlier protagonists were less reticent when describing their actions as terrorism.

History of Terrorism

The term terrorism first came to prominence during the French Revolution in reference to the 'Reign of Terror.' However behavior that fits within definitions of terrorism can be identified throughout human history. Particular examples with semantic legacies include the 'Hindu' Thugees (Thugs), the Muslim Ismailis-Nizari or Hashashins (Assassins), and 'Jewish'

Zealot-Sicarii (Zealots). The Thugees are widely believed to have been involved in strangling anywhere between five lakh and one million people between seventh and nineteenth centuries (Rapoport, 2012). By comparison with the Thugs, the Assassins (1090–1275) had a political agenda that relied upon assassination. Daggers were their weapon of choice. Assassins aimed to purify Islam and reintegrate political and religious institutions between the eleventh and thirteenth centuries. Targeted killing allied to the expected capture and death of the perpetrator aroused considerable interest which at times threatened to overwhelm several governments. The utilization by both the Assassins and Zealots of religious doctrine was to have longer lasting impacts. Although the Zealots were only active for around 25 years in the first century AD their impact continues to resonate. Attempts to instigate Jewish rebellion against the large Greek population and the governing Romans failed spectacularly and ultimately contributed to exile. These experiences have become embedded within Jewish consciousness and continue to have a profound impact.

Despite the long history, the tracing of contemporary terrorism tends to stop at nineteenth century influential theorists such as German migrants Karl Heinzen and Johann Most, Russian anarchists, and Irish republicans. Tactics ranged from selective assassinations to indiscriminate bombing and what became known as 'propaganda of the deed.' Hence the action would both scare opponents and mobilize support. Terrorism became a strategy to challenge and overthrow governments. Ideas and tactics from this era were to influence subsequent generations and to inform the approaches of other groups. Rapoport (2012) categorized these forebears within an anarchist wave. Subsequent waves incorporated an anticolonial phase during which liberation movements in colonies such as Algeria, Cyprus, Kenya, and Palestine sought independence from European dominance. He argues that this was followed by the 'New Left wave' and the 'red' groups from Italy, Japan, the United States, and West Germany. A fourth religious phase followed to account for the rise of terror groups such as the Aum Supreme Truth who sought to kill thousands in the 1995 Tokyo underground sarin gas attack and militant Islamic groups such as al-Qa'ida. The precision within the wave classification is not without problems. In particular, the phases overlook earlier secessionists in Ireland and India, continuing nationalist inspired attacks in places such as Armenia, the Basque region, Bengal, Philippines, and Sri Lanka, preceding religious attacks in Saudi Arabia, Kashmir, Khalistan, and ideological actions in the name of left and right wing discourses in South America. Nevertheless emphasis upon a historical trajectory challenges the tendency to stress the novelty of contemporary forms of terrorism. In other words, terrorism did not begin on 11 September 2001.

Methods and Targets

Forms of terrorism tend to be partly determined by location, availability, histories of violence, and calculations over means and ends. Jenkins (1985: 17) remarks that tactics fall into three categories, namely "Terrorists blow up things, kill people, or seize hostages." Groups can adopt a range of tactics or become specialists, associated with particular methods. Nineteenth century Russian anarchists tended to rely on assassinations and gained considerable notoriety through killing prominent figures such as Czar Alexander II in 1881 and throughout the oncoming 'decade of regicide' (1882–1901). Subsequently assassinations were to become associated with the Italian revolutionaries who also became known for indiscriminate thrown bombs into cafes and later the Greek 17 November group who utilized assassinations through three-people scooter teams. Latin American revolutionaries and Palestinian nationalists were recognized proponents of hostage taking in the late 1960s; a tactic revived and adapted by Chechen secessionists, Colombian Fuerzas Armadas Revolucionarias de Colombia (FARC), and Filipino Abu Sayyaf. Palestinian groups also became renowned for hijacking, particularly, planes. Arguably the most famous hijacker of the era was Leila Khaled whose fame proved detrimental to the role's clandestine activities. Ideological framing also influences the selection of targets. Webb and Cutter (2009) explore shifts in the U.S. attacks from left-wing attacks in the 1970s against more overtly capitalist sites to fundamentalist Christian targeting of abortion and women's health clinics during the 1980s and environmental and animal rights targets in the 2000s.

The evolution of tactics requires predecessors from whom to learn. Moreover strategic approaches are also influenced by the attitudes and reactions of targets and intended supporters. Over the last 30 years, considerable attention has been allotted to terrorists who willingly die for their cause. Muslim and Tamil groups have been most associated with suicide bombers. Stemming from the 1983 bombing of the U.S. embassy Pape (2005) documents the extensive use of suicide bombers. For Pape, such attacks are rational approaches in that they are strategic and bombers become intelligent devices. Cognition enhances the bombs' reactive capability to situations, creating the most devastation in comparison with more conventional forms of explosives.

Another form of terrorists' 'suicidal tactic' has been the hunger strike, most associated with Irish republicans during the 1980s. The histories of the struggles, which groups connected into, provided examples of political violence which continue to resonate and enabled subsequent actions to be positioned as justifiable and part of a longer term trajectory of opposition. Irish republicans were able to draw upon the history of hunger strikes and the interwoven symbolism of sacrifice within Catholicism. Similar connections can be noticed in other forms of self sacrifice. For groups associated with al-Qa'ida, the utilization of what was ostensibly the taking of one's own life, through explosion, also connected into a credible religious legacy that stemmed from the belief that death in the name of Allah was not suicide but martyrdom.

Because impressions are dominated by outcomes of attacks and the asymmetric and limited form of political violence, which terrorism tends to take, careful consideration has to be allocated to targets. If a small number of members are to arouse widespread fear, the impact has to resonate beyond the immediate target. Hence sites that have significant symbolism are more likely to have a wider effect if attacked. The Twin Towers are an obvious example of locations which were symbolic of American economic, cultural, and political strength. At an individual level the Russian revolutionary Sergei Nechaev argued in 1868 that when selecting targets, "the guiding principle must be the measure of service the person's

death will necessarily render the revolutionary cause" (cited in Vertigans, 2011: 113).

Following the Tokyo sarin and plane attacks on America in 2001, concerns have grown that a new stage of terror methods and mass casualties has been reached. Despite these fears, most groups remain bounded by the restraints of impact. In crude terms, if groups aim to attract support and achieve sustainable longer term goals, there is a limit on the extent of devastation that they can cause.

Purpose of Attacks

The immediate intention behind acts of terrorism is exemplified by the famous 2500-year-old proverb attributed to Sun Tzu, a Chinese military strategist, who declared 'Kill one, frighten 10 000.' Following the subsequent development of extensive, rapid forms of communication, the number of potential people to frighten has magnified considerably. Beyond this instant aim, Crenshaw (2011: 9) proposes attacks by terror groups are committed for five interrelated reasons:

1. to set the political agenda;
2. to undermine the opposing government's authority;
3. to provoke an overreaction;
4. to mobilize potential supporters; and
5. to coerce compliance.

Without acts of political violence, groups would struggle to achieve these goals and to this end terrorism can be effective. Issues are raised and discussed in public and political arenas which might otherwise be ignored. Moreover attacks undermine economic confidence and the legitimacy of governments' who are observed to be failing to meet their primary duty of protecting citizens. Facing a crisis of credibility and anxious populations, governments can indiscriminately overreact in haste to defeat the threat.

Conversely the outcomes of attacks create difficulties for terror groups and in particular how to justify associated death and devastation. A common tactic is to emphasize the reactionary nature of the strategy and to apportion blame to enemies who have forced the situation to arise. Heinzen helped to set up a template explanation which continues to be recycled, albeit usually less eloquently, when proposing in 1849 that

> The path to Humanity will pass through the zenith of Barbarity. Our enemies have made this principle a law of politics and we shall either have to observe this 'law,' follow this 'constitutional path' or be buried, and our freedom with us.
>
> cited in Vertigans, 2011: 28

The likelihood of the tactics being effective depends upon multiple factors. First is the extent to which groups' aims resonate with wider populations who have the potential to be mobilized into widespread opposition to governing forces. The second and interrelated factor is the government's counterterror response and the extent to which opposition has the freedom to organize.

Causes of Terrorism

A number of explanations are put forward for why people become terrorists. These can be broadly categorized into individualist and socioeconomic criteria.

Individualist explanations tend to be prevalent across populist swathes of populations which struggle to reconcile acts of terrorism with mainstream norms and values regarding the sanctity of life and the killing of the 'defenseless.' Hence the destructive impact of attacks informs impressions and leads to a strong association with terror groups and violence. Such forms of aggression can become the dominant, possibly, only association people have with the group. Against this framing, terrorists are associated with personal failings which range emotional vulnerability through to mental illness and onto irrational, bloodthirsty individuals. Emotional explanations tend to be particularly prominent within accounts for why females become involved in terrorism and which differs from analysis of male engagement. Common reasoning includes women's humiliations, trauma, and shame often allied to manipulation by male recruiters.

Within psychology the emphasis on individual traits led to a long-standing commitment to identify a terrorist 'personality.' Yet dating back from the period of the Russian anarchist Narodnaya Volya, as Venturi (cited in Crenshaw, 2011: 44) reports, "the policy of terrorism united many different characters and mentalities." Moreover, the emphasis on the bloodthirsty and the mentally ill loses sight of the controls that terrorists have to possess and conditions which are placed upon targets. As a member of the Italian 'red' Prima Linea explains, "it was only permissible to kill certain persons in accordance with particular rules" (cited in Vertigans, 2011:133).

At the end of the explanatory spectrum is causal reasoning which revolves around structural conditions such as poverty, education, and repression. Rationality tends to underpin such explanations which provide pragmatic, materialistic accounts for involvement. Calculated criterion is usually attributed to males rather than females. Crenshaw (2011) points out that these generic explanations are problematic, not least because the impacts of such factors are not restricted to individuals who become terrorists. Instead large sections of populations may experience the same conditions and yet do not become terrorists. The obvious example would be to point to the underrepresentation of women in groups yet who are disproportionately poorer and politically weaker than the overrepresented males. There are a number of reasons for this inverse correlation. These include the patterns of inequality and restricted opportunities that characterize particular societies that are replicated within terror groups. Hence females in locations as diverse as Saudi Arabia, Northern Ireland, and parts of the United States often do not learn the capabilities and attitudes that appeal to skilled occupations and terror groups. Nor do they possess the social relationships that facilitate entry into groups. And when females are part of groups, they are often reported to be in support, marginalized roles. Even within egalitarian red groups which included prominent female members such as the (West German) Red Army Faction's Ulrike Meinhof and Gudrun Ensslin, the Japanese Red Army's Shigenobu Fusako and the (American) Weather's Bernardine Dohrn, there have been reports of machismo and

sexism (MacDonald, 1991). Portrayals within media and academic coverage will often draw upon physical reference points and sexuality in the analysis.

When experiencing the wider structural factors, reactions can take numerous nonviolent forms. Individuals may become involved in other forms of protest, could be apathetic or might seek alternative forms of explanation or escapism. Attempts to generalize based upon such generic reasoning cannot explain why these people do not become terrorists. However as the post-September 2001 American conflagration of the 'war on terror' with a 'war on poverty' highlighted, these oversimplified causal connections continue to resonate and influence approaches to terrorism. Conversely although Western 'red' groups tended to position themselves as agitators on behalf of the dispossessed and exploited, members tended to originate from well-educated, relatively privileged backgrounds.

Within the response to the 'war on terror,' emphasis has also been placed upon the introduction of democratic arrangements as an antidote to terrorism. Such a stance may overestimate the passive consequences of democracy. Indeed a quick review of a number of the prominent groups over the last 40 years highlights activities which originated in (West) Germany, Italy, Japan, the United Kingdom, and the United States; in other words democratic nations are not immune to terrorism. Conversely repressive regimes and particular measures contribute to the formation of terror groups. Crenshaw (2011) outlines how severe government actions were instrumental in the development of terrorism in diverse locations such as Tzarist Russia and post-1960s Northern Ireland. The latter example was also notable because government constraints were imposed within a democratic nation. Hence the denial or obvious futility of other forms of political channels for change contributes to radicalization. Consequently both democracy and repression can both be contributory causes and potential solutions. Much depends upon preconditions and underlying processes.

Trigger Points

Particular events can often be described as 'triggers,' incidents which contribute to individuals becoming involved with terror groups. Examples include the deaths of like-minded protesters by government security forces, which led to members joining the (West German) Red Army Faction (Aust, 1987; Bougereau, 1981; Neidhardt, 1982; Varon, 2004) or the Waco tragedy, which influenced the Oklahoma bomber Timothy McVeigh (Michel and Herbeck, 2001; Serrano, 1998), and killing of family members by rival groups, which was a factor behind recruitment both to loyalist and republican groups, in Northern Ireland (McDonald and Cusack, 2005; Taylor, 2000; Toolis, 1995). Interrelated to these events can be the suppression of the wider social movement or ethnic group which tends to be associated with nationalist struggles in locations such as Chechnya, Sri Lanka, and Uighur struggles in East Turkestan, China. Other triggers include unfulfilled expectations where governments fail to meet hopes around policies such as Islamification in Sadat's Egypt and Saudi Arabia and vehemently opposed government decisions such as the American-led coalition's invasion of Iraq. Alongside these politicized events can be more mundane, personalized, or heightened emotions such as boredom, excitement, honor, self gratification, and comradeship. These emotions formulate within social processes which some groups utilize in order to enhance recruitment and often resonate within existing preconditions. For Horgan (2005) the likelihood of terrorism as a strategy emerging and the subsequent direction is heightened when triggering events combine with preconditions.

Preconditions

The likelihood that triggers will lead to terrorism seems to be influenced by societal preconditions (Crenshaw, 2011; Horgan, 2005, 2009). Examples include historical development, rapid structural transformations, weak governance, foreign interference, and levels of pacification and legitimacy of asymmetrical political violence. In essence, in some locations violence remains embedded within dispositions as a credible option for instigating change.

Locations with a history of political violence that continue to resonate will often provide the preconditional basis which combine with particular triggers. Hence the incorporation of political aggressive struggles within Chechen, Columbian, Palestinian territories, and American mid-West provides credible templates on which subsequent generations can draw. In these contexts common memories incorporate symbols, sites, victories, betrayals, and narratives provide frameworks through which triggers are interpreted and processed (Vertigans, 2008, 2011).

Wider Social Movements

Terror groups tend to emerge out of wider movements and particular ideological, ethnic, religious, and national issues or combinations thereof. Consequently there are multiple issues which lead both to mobilization of the broader levels of activism and ultimately the decision to shift toward a violent strategy.

The demographic representativeness of that movement and the extent to which groups maintain ties and share commonalities will have a huge impact on the life expectancy and possibility of successful outcomes. Mobilization can originate in wider movements such as civil rights in Northern Ireland, student activism in Japan, Christian Identity in the United States, and the first *Intifada* in the Palestinian territories. These movements raise the profiles of social, economic, and political problems and engender acceptance over the right to challenge the status quo. For reasons identified above, such as perceived failure and government repression, factions within the movement breakaway and formulate alternative strategies in order to overcome the frustrations, disillusionment, anger or failings of the wider movement. Violence can become a viable alternative. Hence individuals must progress through evolving processes. Members shift from participating in passive challenging movements to being advocates of more radical forms of activism including attending demonstrations, and then onto political violence. However, once the breach is formulated, the process can become condensed for subsequent members who

are informed by the perceived failings of the original wider movement and the example of group founders. Moreover, second generations of members onward are, as Crenshaw (2011) outlines, also reacting to government counterterror reactions to the initial burst of terrorism. These responses are often instrumental in contributing to a surge in membership, particularly when government behavior resembles the authoritarian or partial roles outlined within radical discourse.

Processes into Terrorism

The intent to scare has contributed to individual terrorists being transformed into societal and, through globalization, international folk devils. Osama bin Laden being the recent most notable example. However the stereotypes which such images help to inform detract from the social processes which led to the formation, maintenance, and ending of terrorism. Similarly it is important that the inclination to isolate terrorists according to individualist criteria and psychological weaknesses does not contribute to the failure to locate terrorism as generally being a group activity. Lone wolves of terrorism, who operate by themselves, maybe prominent within populist media but in reality tend to be much rarer.

Consequently rather than search for common individual dispositions, much more can be learned from exploring the social processes which contribute to terrorism. For instance, Crenshaw (2011: 46) argues "Interactions among members of the group may be more important in determining behavior than the psychological predispositions of individual members." Within the group, a "pattern of mutual reassurance, solidarity, and comradeship develops, in which the members of the group reinforce each other's self-righteousness, image of a hostile world, and sense of mission" (Crenshaw, 2011: 47). In locations where there is widespread acceptance of the terror groups discourse, if not their methods, such as within popular nationalist struggles the transformations of individuals who are marginalized by residence into violent behavior may not be unexpected. By comparison, when becoming involved in left-wing violent activism, the shifts which relatively privileged members undertook required substantial changes both in attitude and in behavior.

The transition into terrorism is also a two way process. Individuals must usually be willing to join and groups must want their membership. The latter distinction is often lost within a general assumption that terror groups will accept anyone. On the contrary, groups must select carefully. Recruits must be able to operate clandestinely, be incorporated within groups, maintain secrecy, be capable, to offer utilizable skills, and be trusted. Consequently individuals must be known to the group or to reliable gatekeepers who can vouch for them. In nationalist communities, there are channels of communication which contribute to controlled and informed access to terror groups through points of contact. By comparison, individuals lacking existing networks which extend into terror groups and restricted opportunities for interaction would struggle to join without a credible link. It is partly for this reason that the emphasis upon recruitment through the Internet is overstated because few groups would accept members without a form of trusted reference.

These social bonds help to generate forms of consciousness that enable individuals to shift from mainstream normative behavior to challenging norms and values that may well entail, the intended or otherwise, seclusion from previous dominant relationships. And in order to make this transition, individuals must learn both about political ideology and associated forms of violence and become sufficiently detached from previous conflicting loyalties and affective ties. During this process, individuals or groups targeted for attack can be depersonalized and dehumanized or existing perceptions reemphasized. Hence terror groups will refer to targets in animalistic terms such as pig, monster and governments will tend to reciprocate through their emphasis upon perpetrators. For the terror groups, nonhuman characteristics help to mobilize support and commitment to attacks. In turn governments explain attacks through reference to innate and ultimately deterministic evilness. Malešević (2010) draws upon the role of ideologies which legitimize killing in the name of truth, justice, and equality. Powerful narratives within discourses create demarcations which include and exclude around religious identity, political allegiance, national consciousness, and racial features. These divisions provide the parameters to enable individuals to overcome the tensions that exist within societies over the reverence of life and the necessity of creating death. However, the narrative has to resonate with contemporary experiences and/or historical memory. Individuals who shift toward political violence will do so for numerous reasons including their belief that the reasoning for attack is justified. Hence some Islamic group members will consider that the United States is trying to destroy Islam. By comparison, American white supremacists are convinced that Jews have taken over the country. Generally members of the different discursive groups tend to share a belief that their actions are altruistic and are implemented as part of a wider strategy in order to create a better society for their followers, both real and potential.

Contrary to the populist view of the innately evil and abnormal terrorist, such individuals are not 'natural' terrorists. Instead individuals have to learn about the radical discourse and how to be involved in acts of political violence. Within contemporary impressions of terrorism and the perceived dominance of 'Islamic' groups, education and in particular religious schools have been identified as primary causes of radicalization. Certainly a variety of schools in diverse locations such as Northern Ireland, the United States, and Saudi Arabia have contributed to perceptions of superiority, injustice, and separation. Again though as with the poverty causal explanation, millions of children are educated yet only a minority proceeds into terrorism. Education seems to be contributing to particular social dispositions and a wider precondition. Nevertheless in isolation, schools and universities are not conveyor belts for terrorism. Information is also acquired through a variety of media sources which can contribute to processes of radicalization. And within the media, considerable emphasis has been placed upon the role of the Internet in extending the terror threat. The Internet has become a useful tool of communication for terror groups and a potential source of recruitment. However there is a danger that the risks are overstated and the Internet becomes incorporated within the populist view that terrorists are brainwashed into joining groups. News items on a variety of medium do

influence views and values. However they only do so in the form of triggers, combining with preexisting conditions which provide the framework through which the event is interpreted.

Within nationalist settings, family members who share similar life experiences can often be instrumental in the formation of radical ideas. With more discursively orientated groups, peers can be more fundamental, particularly in societies in which they are more demarcated boundaries between generations. The formation and subsequent membership of left-wing groups in the late 1970s are particular examples of the prominence of peer groups. Within Italy, this happened during a period when modernization and secularization were contributing to a diminution of Catholicism, family constraints, and ties. Friendships became more integral to identity formation as younger people sought answers from like-minded people who were sharing similar experiences. And once potential members are part of groups which either develop into terror organizations or lead into such organizations, the social dynamics are integral to the decision-making with relation to joining, remaining, and ultimately whether to leave alive or to continue to risk imprisonment and death. For instance, members of paramilitary groups in Northern Ireland would often socialize together, meeting in pubs and clubs and members of Jemaah Islamiyah ate, played football, and did business together. Socializing within groups had two main consequences: First, it strengthens the bonds between members; second, it contributes to a weakening of ties outwith the group. Longer term outcomes can vary, depending in part upon the extent to which members become separated from the people they claim to represent. Groups became the emotional focal point of satisfactions and frustrations. In nationalist communities this is less evident. By comparison, some of the 'red' groups' clandestine and transient existence contributed to long periods of inactivity, boredom, isolation, and ultimately a critical distancing between their policy members and what the wider movement and intended supporters wanted.

Processes Which End Terrorism

Terrorism ends within particular locations for numerous reasons. Groups can be defeated through forms of counterterrorism which incorporate physical destruction, repression, imprisonment, or conversely through offers of amnesty for repentant terrorists. Alternatively groups may achieve their aims or at least obtain sufficient concessions to justify the end of the campaigns. Otherwise members may disengage voluntarily after becoming disillusioned with the tactics, lifestyle, or lack of success. For reasons of security and for the wider impressions of commitment levels that may arise, in some groups there is a reluctance to allow members to disengage of their own volition. And as with processes into terrorism, group dynamics are instrumental in decisions to leave, particularly if personal relationships form within groups that members want to prolong within more risk averse environments.

Often the intense relationships and loyalties to other members, alive or dead, can mean that there is an individual reluctance to leave. Solidarity with the group can exceed that of loyalty to the cause. Hence more widespread approaches to peace in civil wars can be adopted which incorporate disarmament, demobilization and reintegration. To be successful, related approaches have to be developed which encourage individuals to leave terrorism through the delegitimization of the radical discourse and/or show how life outside the group would be better than staying in the group. Education programs and employment prospects are integral to prospects for rehabilitation (Bjørgo and Horgan, 2009). Moreover, feelings of hatred and suspicion have to dissipate and be replaced by heightened levels of empathy and moderation. Arguably this is harder than more pragmatic learning and occupationally based approaches at reintegration. Shifts in emotions and the concomitant need for compromise and consensus require stability and security which generate over periods of time. And at this point, a cyclical process can become noticeable because the perceived lack of stability and security were factors in the original decision to join or form groups. Consequently repentance and commitment to pacification cannot be assumed to endure indefinitely if deep-rooted problems are left unresolved.

Conclusion

Contrary to the populist tendency to generalize terrorism and terrorists, groups form and different people join in different locations, for different reasons and with different intentions. They are united by their strategic use of terrorism as a short-term tactic to scare and intimidate and as a means to an end in order to achieve fundamental societal or global change. The routes and processes into terrorism are also diverse and often involve elements of mis/fortune that revolve around the locations in which individuals reside and the people who are integral to their social networks. Locations and networks allied to preconditions and triggers are instrumental in the likelihood that groups form and people join.

Beyond these similarities, there is striking heterogeneity between, and even, within groups. The multitude of discursive groups, regions where they form, issues that provide justification, and complex processes in and out of terrorism restrict the formation of generic explanations. Consequently approaches to prevent terror groups from forming, attacks from being committed, the defeat of groups, or to encourage member disengagement have to be nuanced and adapted to reflect the contextual, ideological, and discursive variations. As the unintended, terrorist reinforcing, consequences of counterterror approaches can indicate many governments struggle to achieve this degree of knowledge and understanding.

See also: Law and Terrorism in the US and Beyond; Terror and the Internet; Violence: Public; War and Terrorism, Geography of.

Bibliography

Aust, S., 1987. The Baader-meinhof Gang: The inside Story of a Phenomenon. The Bodley Head, London.

Bjørgo, T., Horgan, J. (Eds.), 2009. Leaving Terrorism behind: Individual and Collective Disengagement. Routledge, Abingdon, UK.

Bougereau, J.M., 1981. Memoirs of an International Terrorist: Conversations with Hans Joachim Klein. Cienfuegos Press, Sanday, Orkney.
Crenshaw, M., 2011. Explaining Terrorism: Causes, Processes and Consequences. Routledge, Abingdon.
Horgan, J., 2005. The Psychology of Terrorism. Routledge, London.
Horgan, J., 2009. Walking Away from Terrorism: Accounts of Disengagement from Radical and Extremist Movements. Routledge, Abingdon.
Jenkins, B., 1985. Future Trends in International Terrorism. RAND Corporation. Available at: http://www.rand.org/pubs/papers/2006/P7176.pdf.
Malešević, S., 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.
MacDonald, E., 1991. Shoot the Women First. First Estate, London.
McDonald, H., Cusack, J., 2005. UDA: Inside the Heart of Loyalist Terror. Penguin Books, Dublin.
Michel, L., Herbeck, D., 2001. American Terrorist: Timothy McVeigh and the Oklahoma City Bombing. Regan Books, New York.
Neidhardt, F., 1982. Soziale Bedingungen Terroristischen Handelns: Das Beispiel der 'Baader-Meinhof-Gruppe.' In: von-Baeyer-Katte, W., Classens, D., Fuger, H., Neidhardt, F. (Eds.), Gruppenprozesse. Westdeutscher, Opladen.
Pape, R., 2005. Dying to Win. Random House, New York.
Rapoport, D., 2012. The four waves of modern terrorism. In: Horgan, J., Braddock, K. (Eds.), Terrorism Studies: A Reader. Routledge, Abingdon.
Schmid, A., 1984. Political Terrorism: A Research Guide. Transaction Books, New Brunswick, NJ.
Serrano, R., 1998. One of Ours: Timothy McVeigh and the Oklahoma City Bombing. W.W. Norton & Co., New York.
Taylor, P., 2000. Loyalists. Bloomsbury Publishing, London.
Toolis, K., 1995. Rebel Hearts. Picador, London.
Varon, J., 2004. Bringing the War Home. University of California Press, Berkeley and LA.
Vertigans, S., 2008. Terrorism and Societies. Ashgate, Aldershot.
Vertigans, S., 2011. Sociology of Terrorism: People, Places and Processes. Routledge, Abingdon.
Webb, J., Cutter, S., 2009. The geography of U.S. Terrorist incidents, 1970–2004. Terrorism and Political Violence 21 (3), 428–449.
Weinberg, L., Pedahzur, A., Hirsch-Hoefler, S., 2012. The challenges of conceptualising terrorism. In: Horgan, J., Braddock, K. (Eds.), Terrorism Studies: A Reader. Routledge, Abingdon.

Test Anxiety and Academic Achievement

Reinhard Pekrun and Elizabeth J Stephens, University of Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Theories and findings on test anxiety are reviewed herein. Test anxiety has been researched extensively over the past 60 years. Findings show that test anxiety (1) comprises multiple affective, cognitive, physiological, and behavioral components; (2) is frequently experienced by students; (3) can negatively impact performance on complex and difficult tasks; and (4) is negatively related to students' academic achievement. Individual failure expectancies, overly high perceived importance of achievement, excessively high task demands, competitive educational structures, normative grading, and high-stakes testing contribute to test anxiety arousal. Psychotherapy and learning environments that strengthen students' competencies and self-confidence can alleviate test anxiety.

Test anxiety is a ubiquitous phenomenon in modern societies. Given that educational and occupational career opportunities are made contingent on individual achievement, attaining success and avoiding failure is vital for individuals across the life span. As a consequence, anxiety of failure is one of the most frequently experienced emotions around the globe today – an emotion that surfaces in various life domains, including education, work, and athletics. Anxiety of failure (or evaluative anxiety) is at the heart of achievement anxiety. Specific variants of achievement anxiety include test anxiety, mathematics anxiety, performance anxiety, competitive sports anxiety, and social–evaluative anxiety. Among these, test anxiety is certainly the most important construct in academic contexts (Cassady, 2010).

The relevance of test anxiety was recognized early in the twentieth century. At that time, psychoanalytical writers began to theorize about fear related to examinations (e.g., Stengel, 1936). Some of the first empirical investigations, including the development of a questionnaire measure of test anxiety, were conducted by J. Brown at the University of Chicago in the 1930s (Brown, 1938). The construct became popular after G. Mandler and S.B. Sarason (1952) published their seminal paper entitled 'A study of anxiety and learning.' Since then, over 1000 studies examining test anxiety have been published. Thanks to these studies, we have not only garnered cumulative evidence on the nature, functions, origins, and development of this emotion, but we have also learned how to successfully alleviate test anxiety by individual psychotherapy.

Concepts of Test Anxiety

Test anxiety is defined as the anxiety related to taking tests that is experienced before and during test taking (here, the term 'test' is broadly defined, including written and oral examinations). At the heart of this emotion is the fear of failing a test and of the consequences associated with failure. Similar to other emotions, test anxiety is comprised of a complex array of component processes. These processes can include (1) affective components in terms of uneasy, nervous feelings; (2) physiological components including physiological activation and symptoms such as increased heart rate, increased respiration rate, or sweating; (3) cognitive components involving worries about possible failure and its consequences; (4) motivational components in terms of motivation to avoid failure and impulses to escape from the situation; and (5) behavioral components in terms of facial and postural expression of anxiety. Although a prototypical case of intense test anxiety comprises all of these components, they are only loosely coupled, and not all of them need to be present in all episodes of test anxiety.

At the onset of research into this emotion, test anxiety was considered a unidimensional construct. Progress toward a more differentiated construct was made when Liebert and Morris (1967) proposed to distinguish the affective and physiological components of test anxiety (referred to as 'emotionality') from the cognitive components (referred to as 'worry'). Later, a number of additional distinctions were introduced, including I.G. Sarason's (1984) differentiation between affective tension, bodily symptoms, worry, and task-irrelevant thinking. Although scholars today agree that test anxiety is a multifaceted construct, opinions vary as to which facets best comprise this emotion.

Another important distinction within the test anxiety construct centers on distinguishing between habitual propensities to experience anxiety related to test taking and momentary episodes of anxiety experienced in a specific test taking situation. Using the distinction between trait and state emotions originally introduced by Cattell (1950), the former variant refers to trait test anxiety, and the latter variant to state test anxiety. Trait test anxiety is an individual's disposition to experience anxiety related to tests. Whenever a trait test-anxious individual is confronted with a test or examination, this disposition can manifest itself in the experience of momentary state anxiety. As such, trait test anxiety is best viewed as a contextualized personality trait.

Traditionally, it was assumed that trait test anxiety generalizes across various types of achievement situations. This assumption is challenged, however, by recent evidence showing that students' anxiety is more domain-specific than previously thought (Goetz et al., 2007). Similar to students' self-concepts of ability, correlations between students' anxiety tend to be low across domains such as math versus languages. Moreover, the available evidence shows that the domain

 http://dx.doi.org/10.1016/B978-0-08-097086-8.26064-9

specificity of students' anxiety increases with age. As such, dispositions to experience test anxiety may best be conceptualized as a hierarchy in which more generalized dispositions to experience this emotion are located at higher levels and domain-related dispositions (such as math test anxiety) are located at lower levels.

Assessment

The multicomponential nature of test anxiety implies that this emotion can be assessed using various methods. These include self-report, report by others, observation of facial and postural expression, analysis of peripheral physiological activation, and neuropsychological assessment of activation in central brain areas such as the amygdala. Among these methods, self-report techniques quickly gained popularity because of the ease with which they are administered in the classroom and because of their excellent psychometric properties.

Whereas J. Brown's (1938) early self-report questionnaire did not gain widespread attention, the Test Anxiety Questionnaire (TAQ) developed by G. Mandler and S.B. Sarason (1952) became the progenitor of the numerous self-report scales developed since then. Early instruments such as the TAQ measured test anxiety as a unidimensional construct. From early on, however, factor analyses showed that test anxiety scales do in fact represent more than one dimension. Following the distinction between the emotionality and worry components proposed by Liebert and Morris (1967), multidimensional instruments were developed. Examples are the Reactions-to-Tests instrument developed by I.G. Sarason (1984), Spielberger's (1980) Test Anxiety Inventory, and the integrative test anxiety scale proposed by Hodapp and Benson (1997).

As the assessment of achievement emotions has developed, it has been recognized that test anxiety is but one of several important test emotions deserving of empirical attention. Specifically, the Achievement Emotions Questionnaire (AEQ; Pekrun et al., 2004, 2011) makes it possible to examine test anxiety in the context of other test-related emotions including enjoyment, hope, pride, relief, anger, shame, and hopelessness. Confirmatory factor analyses of the AEQ scales have shown that test anxiety, although a unique emotion, is closely related to students' test-related shame and hopelessness (Pekrun et al., 2011).

Impact on Learning and Academic Achievement

Empirical Evidence

The link between test anxiety and learning has been studied in a large number of laboratory experiments and nonexperimental field studies (Hembree, 1988; Zeidner, 1998, 2014). Experimental studies have examined the effects of anxiety on performance on various types of cognitive tasks, including working memory tasks, mathematical problems, text comprehension, intelligence test items, complex problem solving, music performance, and sports performance. In general, the findings show the following pattern.

(1) Trait test-anxious students tend to perform less well on difficult or complex cognitive tasks than their less anxious counterparts. They tend to perform equally well or even better on tasks that are easy and less complex. (2) In experimental work, the detrimental effects of trait test anxiety are more pronounced when a task is introduced as a 'test' and involves ego-threat (e.g., when individuals believe they are working on an intelligence test). (3) State test anxiety experienced when performing complex and difficult tasks relates negatively to performance on these tasks. The relation is stronger for cognitive components of test anxiety (such as worries about failure) than for affective–physiological components.

Taken together, these findings indicate that anxiety impairs performance on difficult and complex tasks, but has null or even positive effects on performance on easy and less complex tasks. In line with the negative link between anxiety and complex task performance, (trait) test anxiety also tends to correlate negatively with students' academic achievement. The average correlations found in two meta-analyses were $r = -0.29$ (in grade 4 through postsecondary student samples; Hembree, 1988) and, for mathematics achievement specifically, $r = -0.23$ (Schwarzer et al., 1989). In addition, meta-analyses found that test anxiety is also negatively correlated with performance on intelligence tests (average $r = -0.33$; Ackerman and Heggestad, 1997), and that general trait anxiety is negatively correlated with performance on creativity tasks (Byron and Khazanchi, 2011).

It is important to note, however, that the range of correlations across studies also included zero and positive coefficients. Moreover, although correlations did not differ significantly across genders (Schwarzer et al., 1989), they did vary across grade levels. In the *meta*-analysis by Hembree (1988), the link between anxiety and academic achievement was near zero at the beginning of elementary school and increased across the elementary school years.

Theoretical Explanations, Mediating Mechanisms, and Reciprocal Causation

Most accounts of test anxiety posit that the cognitive components of this emotion explain the extant empirical evidence. Rumination about failure is thought to consume cognitive resources in terms of attention and working memory capacity, thereby reducing the resources available for task performance – worries about possibly failing an examination prevent one from fully concentrating on the task. Such a reduction in resources in turn leads to deteriorated performance on complex or difficult tasks that put higher demands on these resources. This mechanism explains why anxiety reduces performance on these tasks, whereas performance on easy tasks remains unaffected. Attention-deficit and resource allocation models targeting this kind of mechanism were part of early theories of test anxiety (e.g., Mandler and Sarason, 1952) and continue to be one dominant theoretical paradigm.

More recent approaches to the study of test anxiety call attention to other mediating mechanisms and to alternative causal pathways. For instance, as addressed in Pekrun's (1992a, 2006) cognitive–motivational model of emotion effects, beyond impacting cognitive resources, anxiety can also affect students' motivation to learn, information processing, use of learning strategies, and self-regulation of learning. With

regard to *motivation*, anxiety is likely to reduce students' interest and intrinsic motivation to learn because it is not compatible with the enjoyment needed to develop interest-based motivation. On the other hand, anxiety can arouse strong motivation to invest effort to avoid failure. By implication, the overall effects of anxiety on motivation and effort can be variable.

The impact of anxiety on *learning strategies* and *information processing* can be complex as well. In mood research, positive mood generally promotes top-down, heuristic, and creative styles of thinking and a relational, global mode of processing information, whereas negative mood promotes bottom-up, detail-oriented, analytic styles of thinking that focus on item-specific, local details of information. Similar to negative mood, anxiety likely promotes more rigid learning strategies that focus on rehearsal and the details of learning materials, and undermines the use of flexible cognitive strategies such as elaboration and organization of learning material and the use of metacognitive strategies such as flexible time management. In addition, anxiety can foster students' reliance on *external guidance for learning* and reduce their self-regulation of learning (Pekrun et al., 2011). Depending on the type of task and context, these effects can either undermine or promote academic task performance.

Moreover, for interpreting correlational links between test anxiety and academic performance, reciprocal causation of anxiety and performance should be taken into account. Though the correlations between anxiety and achievement can be due to negative effects of anxiety on students' performance, the reverse causal direction can play a role as well – low academic performance can trigger subsequent anxiety before and during examinations (*developmental model*; Meece et al., 1990; Pekrun, 1992b). In addition, performance deficits in test-anxious students can be caused by a lack of ability or a lack of study skills, implying that anxiety would reflect a lack of competences rather than serve as the cause of performance decrements (*skills deficit model*; Everson et al., 1989).

These various explanations should be viewed as complementary rather than mutually exclusive. In many students, lack of ability and/or learning skills and resulting academic failure may initially incite test anxiety. Subsequently, anxiety and lack of skills jointly impact students' performance trajectories by consuming cognitive resources, reducing intrinsic motivation, and deteriorating the quality of learning (e.g., use of suboptimal cognitive and metacognitive strategies). Overall, such feedback loops between anxiety and performance involve vicious cycles in which anxiety deteriorates performance and low performance further increases the level of anxiety.

A number of studies have shown that the effects of anxiety on achievement can indeed be due to a number of cognitive and motivational mechanisms (Pekrun et al., 2011; Zeidner, 1998, 2014). Multiple mechanisms involving differential effects on performance can also explain why some studies find zero or positive correlations with students' academic achievement instead of negative relations (Hembree, 1988; Schwarzer et al., 1989). Finally, the findings from a few longitudinal studies corroborate that anxiety and achievement are linked by reciprocal causation over time (Meece et al., 1990; Pekrun, 1992b; Schnabel, 1998).

Developmental Trajectories

For many students, the development of test anxiety can be traced back to the preschool years. Psychodynamic approaches suggest that overly high parental expectations and negative judgments, which produce feelings of dependency, guilt, aggression, and anxiety, can play a major role at this age (S.B. Sarason et al., 1960). Nevertheless, in early elementary school the average levels of test anxiety are still relatively low. Mean values increase over the elementary years, which is likely due to many students' emerging experiences with cumulative academic failures (Hembree, 1988; Wigfield and Eccles, 1989; Zeidner, 1998). This development is congruent with a decline in self-concepts of ability (Hembree, 1988). Moreover, correlations between test anxiety and achievement tend to become stronger at this stage, indicating that students' performance and their affective experiences begin to be linked more closely, as suggested by developmental models of the anxiety–achievement relationship.

After elementary school, average levels of test anxiety remain fairly stable throughout secondary school and college. Nevertheless, the individual student can still experience a substantial increase or decrease in test anxiety, indicating developmental plasticity. Specifically, levels of test anxiety can change due to transitions between schools or classrooms. For instance, transitions to a new group can involve changes in one's relative academic status in the group. Social comparison processes can enhance or diminish student's self-concepts of ability and achievement expectations, and thus affect their test anxiety (Zeidner, 1998).

Individual and Social Origins

Biological Factors and Individual Appraisals

In general, emotions can be caused and modulated by numerous individual factors. These include genetic dispositions, neurohormonal processes, cognitive appraisals, situational perceptions, and sensory feedback from facial, gestural, and postural expression (Davidson et al., 2003; Scherer et al., 2001). All of these factors are important for anxiety as well.

With regard to more distal, biological variables, individual differences in general trait anxiety can in part be explained by genetic factors. Up to 50% of the variance in trait anxiety may be due to genetic variance (Eley et al., 2003; van Beijsterveldt et al., 2004), suggesting that biologically determined dispositions also partially define students' propensity to react with worries and emotional arousal in stressful academic situations. More proximal variables that cause and modulate emotions include students' appraisals of their competencies and prospects to achieve. At school, students have to learn how to adapt to academic demands while preserving individual autonomy – a process inevitably guided by appraisals. These appraisals in turn shape students' affect and motivation, including their test anxiety.

Appraisal theories of test anxiety focus on the impact of threat-related appraisals. From the perspective of R.S. Lazarus's *transactional model of stress and anxiety* (Lazarus and Folkman, 1984), test anxiety is based on two kinds of appraisals. The *primary appraisal* pertains to the likelihood and subjective

importance of failure, and the *secondary appraisal* involves cognitively exploring possibilities to cope with the situation. Depending on the combined result of the two appraisals, different emotions can be aroused. Should one perceive threat of failure and insufficient means to control the threat, then anxiety is assumed to be instigated. Lazarus's analysis implies that test anxiety is aroused when two conditions are met: First, the individual must anticipate possible failure on a subjectively important, evaluative task, and second, the individual must doubt whether potential failure can be avoided.

In Pekrun's (1984, 1992b, 2006) *expectancy-value model of test anxiety*, these two assumptions are reconceptualized in terms of expectancies and values related to failure. The basic proposition of this model is that test anxiety is a function of the subjective likelihood (expectation) of failure and the subjective value of failure. Both components are posited to be necessary for the instigation of test anxiety. As such, if failure is avoidable or if the student is indifferent to the consequences of failure, then anxiety should not arise. Anxiety is proposed to be a curvilinear function of expectancy, and is expected to be replaced by hopelessness if failure is subjectively certain.

Typically, situational appraisals of these kinds are based on objective characteristics of the setting, such as the relative difficulty of examination material, but they are also influenced by students' achievement beliefs. These beliefs can take 'irrational' forms (Ellis, 1962), for example, when failure is appraised as likely despite high individual ability, or when failure on subjectively unimportant outcomes is perceived as undermining self-worth. Irrational beliefs can make students not only highly vulnerable to anxiety, but also to other, related negative emotions such as shame ("I am not allowed to fail. If I fail, I am a worthless person").

The available empirical evidence regarding appraisal antecedents of test anxiety is in line with these propositions. Test anxiety has been found to correlate positively with students' expectancies of failure, and negatively with their self-concepts of ability, academic self-efficacy expectations, and academic control beliefs (Hembree, 1988; Pekrun et al., 2004; Zeidner, 1998, 2007).

In addition, students' appraisals are shaped by their achievement goals. Mastery goals relate to task mastery and developing competence, and performance goals relate to one's performance relative to others. In addition, both types of goals can focus either on approaching positive outcomes (i.e., success) or avoiding negative outcomes (i.e., failure; Elliot and McGregor, 2001; Elliot et al., 2011). Specifically, performance-approach goals pertain to outperforming others, and performance-avoidance goals pertain to not doing worse than others. Given that performance-approach and performance-avoidance goals can focus students' attention on the controllability and values of normative academic success versus failure, respectively, it stands to reason that these goals also influence students' anxiety (Pekrun et al., 2006, 2009). In line with these hypotheses, research examining the relationship between achievement goals and anxiety consistently finds that students' performance-avoidance goals are positively related to their test anxiety scores (Linnenbrink and Pintrich, 2002; Pekrun et al., 2006, 2009).

The Role of Education

Students' appraisals are influenced by the social environment. Factors in the environment that impact failure expectancies and the perceived importance of tests and examinations likely contribute to students' test anxiety (Pekrun, 2006). These factors include learning environments in the classroom and the home, and the organization of tests and examinations.

Learning Environments

Lack of structure and clarity in classroom instruction, as well as excessively high task demands, relate to students' test anxiety (Zeidner, 1998). The effects of these factors are likely mediated by students' perceived lack of control and expectancies of failure (Pekrun, 1992b). Overly high achievement expectancies from important others, negative evaluation feedback, and negative consequences of failure also correlate positively with students' anxiety (Hembree, 1988).

Additionally, competition within the classroom is positively related to test anxiety – likely because competition reduces expectancies for success and increases the importance of avoiding failure (Zeidner, 1998). In contrast, in K-12 research, social support from parents and teachers and a cooperative classroom climate are uncorrelated with students' test anxiety scores (Hembree, 1988). Two explanations may account for these findings. First, coercive components of supportive behavior may counteract the beneficial effects of support (Helmke, 1983). Second, support and anxiety may be linked by negative feedback loops: Social support can alleviate anxiety (negative effect of support on anxiety), but anxiety can engender support (positive effect of support on anxiety), thus yielding an overall zero correlation.

Tests and Examinations

Lack of structure and transparency within examinations has been shown to also contribute to students' anxiety (Zeidner, 1998). The findings point to factors such as clarity concerning demands, materials, examination procedures, and grading standards as critical in the elicitation of test anxiety. Furthermore, open-ended test items are found to induce more anxiety than multiple-choice items. Because open-ended formats require more working memory capacity, the worry and task-irrelevant thinking associated with anxiety can consume cognitive resources needed for the task, in turn eliciting more threat and further debilitating performance in anxious students. Methods effective at reducing these effects include using multiple-choice formats and providing external aids, such as books and computers, because such practices reduce working memory load and the threat of failure.

Furthermore, classroom practices such as allowing students to choose the questions they respond to on a test, relaxing time constraints, and giving second chances (e.g., providing the option to retake a test or correct incorrect responses for make-up points) can also reduce students' test anxiety. These methods likely function to minimize anxiety by increasing students' perceived control and alleviating expectations of failure. Finally, high-stakes testing that includes long-term and often far-reaching consequences of failure (e.g., impeded educational and career trajectories) is also likely to contribute to students' anxiety. This implies that educators who aim to

prevent the development of test anxiety should indeed avoid these types of evaluations.

Emotion Regulation, Coping, and Therapy

Given that test anxiety can hinder well-being and achievement, students can and should attempt to regulate this emotion. In the emotions literature, the regulation of negative emotions, and of stressful situations which tax or exceed individual capabilities, is referred to as *coping* (Zeidner and Endler, 1996). More severe anxiety can be successfully managed by psychotherapy.

Emotion Regulation and Coping

Emotion regulation implies strengthening or maintaining adaptive emotions, and decreasing or preventing maladaptive emotions. Basic components include recognizing and understanding one's emotions, managing one's emotions, and using emotions for action (e.g., for studying). More specifically, emotion management can be achieved by targeting the symptoms of the emotion (*emotion-oriented regulation*), by changing underlying appraisals (*reappraisal*), by acquiring competences (*competence-oriented regulation*), and by selecting and shaping emotion-eliciting tasks and learning environments (*situation-oriented regulation*; Pekrun, 2006). Beyond regulatory competences pertaining to one's own emotions, emotional competences also comprise abilities to recognize, understand, manage, and use others' emotions. Overall, cognitive competences to regulate one's own and others' emotions constitute students' emotional intelligence (Matthews et al., 2002).

Coping with stress and anxiety is addressed by Lazarus' transactional stress model (Lazarus and Folkman, 1984). In this model, anxiety is thought to lead to attempts to cope with this emotion and/or the stress that caused it. Many taxonomies related to coping with anxiety have been proposed, and most of these conceptions differentiate between problem-oriented, emotion-oriented, and avoidant coping (Rost and Schermer, 1987; Zeidner and Endler, 1996; Skinner et al., 2003). Clearly, these categories overlap with the above-mentioned categories of emotion regulation.

Problem-oriented coping entails active attempts to change the situation that causes anxiety, such as employing learning strategies to acquire competencies. Problem-oriented coping related to examinations (e.g., preparing for an examination) can itself have adverse effects. For instance, merely dealing with exam material can arouse thoughts about an upcoming exam, which in turn can increase situational anxiety (Bolger, 1990). For most students, however, the beneficial long-term effects of improving competencies via problem-oriented coping likely outweigh any negative situational effects.

Emotion-oriented coping is aimed toward directly changing the experience of anxiety. Emotion-oriented strategies include (1) anxiety reduction by means of alcohol, nicotine, or pharmaceutical consumption, or by means of relaxation techniques; (2) reduction of emotional tension by simply accepting anxiety and the possibility of failure ('secondary control'; Morling and Evered, 2006); (3) induction of positive, anxiety-incompatible emotions (e.g., by using humor, music, or emotional support from others); and (4) cognitively reappraising the situation as more controllable or less subjectively important. Many of these strategies are effective at reducing anxiety, but some of them clearly have negative side effects.

Avoidance-oriented coping implies behaviorally or mentally escaping from an anxiety-inducing situation. This can be achieved by (1) searching for mental distraction by focusing attention on task-irrelevant contents; (2) procrastination, taking excessively long breaks, and cutting one's preparation short; and (3) truancy, examination-avoidance, and school withdrawal. Like emotion-oriented coping, these strategies can also lead to an immediate reduction in anxiety. However, avoiding the experience of anxiety can produce a detrimental increase in a less conscious form of emotional arousal occurring at the physiological level (Spangler et al., 2002). Moreover, such strategies leave the underlying factors contributing to the experience of anxiety untreated. In this way, these strategies can be detrimental for students' learning and future career prospects.

Psychotherapy

Individual test anxiety is treatable. Some of the treatments for test anxiety are among the most successful psychological therapies available, with effect sizes above $d = 1$ (Hembree, 1988). Similar to the various coping strategies, different treatments for test anxiety focus on different manifestations and antecedents of this emotion. These include: affective–physiological symptoms (emotion-oriented therapy), cognitive appraisals (cognitive therapy), and competence deficits caused by lack of strategies for learning (skills training; Zeidner, 1998).

Emotion-oriented therapy includes anxiety induction (e.g., flooding), biofeedback procedures, relaxation techniques (e.g., progressive muscle relaxation; Jacobson, 1938), and systematic desensitization. *Cognitive therapies* aim to modify anxiety-inducing control beliefs, values, and styles of self-related thinking. Examples are cognitive–attentional training, cognitive restructuring therapy, and stress-inoculation training. *Study-skills training* teaches students to understand and use learning strategies and problem-solving skills that promote academic success and thus decrease anxiety. Finally, *multimodal therapies* integrate different procedures to address different symptoms and antecedents of anxiety within one treatment.

Cognitive and multimodal therapies have proven especially effective at both reducing test anxiety and enhancing academic performance (Zeidner, 1998). Study-skills training can successfully reduce test anxiety in students with deficits in their learning strategies. Consistent with the arguments above, therapy focusing exclusively on emotion-oriented procedures has been shown to successfully reduce anxiety, but has proven less effective at improving academic achievement. These kinds of therapy address the affective and physiological components of anxiety, but not the underlying cognitive components of anxiety that are primarily responsible for the performance-debilitating effects of this emotion.

See also: Academic Achievement Motivation, Development of; Coping across the Lifespan; Emotional Regulation; Stress in Childhood: Effects on Development; Stress, Coping and Health; Stress: Psychological Perspectives.

Bibliography

Ackerman, Paul L., Heggestad, E.D., 1997. Intelligence, personality and interests: evidence for overlapping traits. Psychological Bulletin 121, 219–245.

Bolger, N., 1990. Coping as a personality process: a prospective study. Journal of Personality and Social Psychology 59, 525–537.

Brown, C.H., 1938. Emotional reactions before examinations: II. Results of a questionnaire. Journal of Psychology 5, 11–26.

Byron, K., Khazanchi, S., 2011. A meta-analytic investigation of the relationship of state and trait anxiety to performance on figural and verbal creative tasks. Personality and Social Psychology Bulletin 37, 269–283.

van Beijsterveldt, C.E.M., Verhulst, F.C., Molenaar, P.C.M., Boomsma, D.I., 2004. Genetic basis of problem behavior in 5-year old Dutch twin pairs. Behavior Genetics 34, 229–242.

Cattell, Raymond B., 1950. Personality: A Systematic Theoretical and Factual Study. McGraw Hill, New York.

Cassady, J.C. (Ed.), 2010. Anxiety in Schools: The Causes, Consequences, and Solution for Academic Anxieties. Peter Lang, New York.

Davidson, Richard J., Scherer, Klaus R., Goldsmith, H.H. (Eds.), 2003. Handbook of Affective Sciences. Oxford University Press, Oxford, UK.

Eley, T.C., Bolton, D., O'Connor, T.G., Perrin, S., Smith, P., Plomin, R., 2003. A twin study of anxiety-related behaviours in pre-school children. Journal of Child Psychology and Psychiatry 44, 945–960.

Elliot, Andrew J., McGregor, H., 2001. A 2 × 2 achievement goal framework. Journal of Personality and Social Psychology 80, 501–519.

Elliot, Andrew J., Murayama, Kou, Pekrun, Reinhard, 2011. A 3 × 2 achievement goal model. Journal of Educational Psychology 103, 632–648.

Ellis, Albert, 1962. Reason and emotion in psychotherapy. Lyle Stuart, New York.

Everson, H., Millsap, R.E., Browne, J., 1989. Cognitive interference or skills deficit: an empirical test of two competing theories of test anxiety. Anxiety Research 1, 313–325.

Goetz, Thomas, Frenzel, Anne C., Pekrun, Reinhard, Hall, Nathan C., Lüdtke, Oliver, 2007. Between- and within-domain relations of students' academic emotions. Journal of Educational Psychology 99, 715–733.

Helmke, Andreas, 1983. Schulische Leistungsangst (Academic Achievement Anxiety). Lang, Frankfurt, Germany.

Hembree, Ray, 1988. Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research 58, 47–77.

Hodapp, Volker, Benson, Jeri, 1997. The multidimensionality of test anxiety: a test of different models. Anxiety, Stress and Coping 10, 219–244.

Jacobson, E., 1938. Progressive Relaxation. University of Chicago Press, Chicago.

Lazarus, Richard S., Folkman, Susan, 1984. Stress, Appraisal, and Coping. Springer, New York.

Liebert, R.M., Morris, L.W., 1967. Cognitive and emotional components of test anxiety: a distinction and some initial data. Psychological Reports 20, 975–978.

Linnenbrink, Elizabeth A., Pintrich, Paul R., 2002. Achievement goal theory and affect: an asymmetrical bidirectional model. Educational Psychologist 37, 69–78.

Mandler, George, Sarason, Seymor B., 1952. A study of anxiety and learning. Journal of Abnormal and Social Psychology 47, 166–173.

Matthews, G., Zeidner, Moshe, Roberts, Richard D., 2002. Emotional Intelligence: Science and Myth. MIT Press, Cambridge, MA.

Meece, Judith L., Wigfield, Allan, Eccles, Jacquelynne S., 1990. Predictors of math anxiety and its influence on young adolescents course enrollment intentions and performance in mathematics. Journal of Educational Psychology 82, 60–70.

Morling, B., Evered, S., 2006. Secondary control reviewed and defined. Psychological Bulletin 13, 269–296.

Pekrun, Reinhard, 1984. An expectancy-value model of anxiety. In: Schwarzer, Ralf, van der Ploeg, Henk, Spielberger, Charles D. (Eds.), Advances in Test Anxiety Research., vol. 3. Swets & Zeitlinger, Lisse, The Netherlands, pp. 53–72.

Pekrun, Reinhard, 1992a. The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. Applied Psychology 41, 359–376.

Pekrun, Reinhard, 1992b. Expectancy-value theory of anxiety: overview and implications. In: Forgays, D.G., Sosnowski, T., Wrzesniewski, K. (Eds.), Anxiety: Recent Developments in Self-appraisal, Psychophysiological and Health Research. Hemisphere, Washington, DC, pp. 23–41.

Pekrun, Reinhard, 2006. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18, 315–341.

Pekrun, Reinhard, Elliot, Andrew J., Maier, Markus A., 2006. Achievement goals and discrete achievement emotions: a theoretical model and prospective test. Journal of Educational Psychology 98, 583–597.

Pekrun, Reinhard, Elliot, Andrew J., Maier, Markus A., 2009. Achievement goals and achievement emotions: testing a model of their joint relations with academic performance. Journal of Educational Psychology 101, 115–135.

Pekrun, Reinhard, Goetz, Thomas, Frenzel, Anne C., Barchfeld, Petra, Perry, Raymond P., 2011. Measuring emotions in students' learning and performance: the Achievement Emotions Questionnaire (AEQ). Contemporary Educational Psychology 36, 36–48.

Pekrun, Reinhard, Goetz, Thomas, Perry, Raymond P., Kramer, Klaudia, Hochstadt, Michaela, 2004. Beyond test anxiety: development and validation of the Test Emotions Questionnaire (TEQ). Anxiety, Stress and Coping 17, 287–316.

Rost, Detlev H., Schermer, F.J., 1987. Emotion and cognition in coping with test anxiety. Communication and Cognition 20, 225–244.

Sarason, I.G., 1984. Stress, anxiety, and cognitive interference: reactions to tests. Journal of Personality and Social Psychology 44, 929–938.

Sarason, Seymor B., Davidson, K.S., Lighthall, F.F., Waite, R., Ruebush, B.K., 1960. Anxiety in Elementary School Children. Wiley, New York.

Scherer, Klaus R., Schorr, Angela, Johnstone, T. (Eds.), 2001. Appraisal Processes in Emotion. Oxford University Press, Oxford, UK.

Schnabel, Kai, 1998. Prüfungsangst und schulisches Lernen (Test Anxiety and Learning at School). Waxmann, Münster, Germany.

Schwarzer, Ralf, Seipp, Bettina, Schwarzer, Christine, 1989. Mathematics anxiety and performance: a meta-analysis. In: Schwarzer, Ralf, van der Ploeg, Henk, Spielberger, Charles D. (Eds.), Advances in Test Anxiety Research, vol. 6. Swets & Zeitlinger, Lisse, The Netherlands, pp. 105–119.

Skinner, Ellen A., Edge, K., Altman, J., Sherwood, H., 2003. Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. Psychological Bulletin 129, 216–269.

Spangler, Gottfried, Pekrun, Reinhard, Kramer, Klaudia, Hofmann, Hubert, 2002. Students' emotions, physiological reactions, and coping in academic exams. Anxiety, Stress and Coping 15, 413–432.

Spielberger, Charles D., 1980. Test Anxiety Inventory: Preliminary Professional Manual. Consulting Psychologist Press, Palo Alto, CA.

Stengel, E., 1936. Prüfungsangst und Prüfungsneurose (Exam anxiety and exam neurosis). Zeitschrift für psychoanalytische Pädagogik 10, 300–320.

Wigfield, A., Eccles, J.S., 1989. Test anxiety in elementary and secondary school students. Educational Psychologist 24, 159–183.

Zeidner, Moshe, 1998. Test Anxiety: The State of the Art. Plenum, New York.

Zeidner, Moshe, 2014. Anxiety in education. In: Pekrun, Reinhard, Linnenbrink-Garcia, Lisa (Eds.), International Handbook of Emotions in Education. Francis & Taylor, New York, pp. 265–288.

Zeidner, Moshe, Endler, Norman (Eds.), 1996. Handbook of Coping: Theory, Research, Applications. Wiley, New York.

Test Theory: Applied Probabilistic Measurement Structures

Hartmann Scheiblechner[†], University of Marburg, Marburg, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15619–15625, © 2001, Elsevier Ltd.

Abstract

Psychological test data for unidimensional psychological constructs (abilities, traits, …) have the form of data matrices with the rows corresponding to subjects and the columns corresponding to items and the entries being single graded codings (ratings) of reactions of subjects to items. The reactions are governed by response probabilities in the sense of probabilistic IRT (item response theory). The product set of subjects and items can be weakly ordered by the concept of stochastic dominance. A hierarchy of probabilistic measurement models (in the sense of representational measurement theory) is reported for those data. The response vectors of subjects (and items) determine a coordinate wise partial order. Two independence axioms (ISOP, isotonic probabilistic measurement structure) grant the construction of separate ordinal scales for subjects and items. An additional cancellation axiom (ADISOP) gives interval scales for subjects and items. An extension of cancellation to another independence axiom (CADISOP) gives an additional interval scale for response categories. All structures can be extended to linear structures for items and/or subjects. It is the essential advantage of the hierarchical organization of the models that each step of complexity can be tested separately and independently.

Psychological test theory invites the formulation and testing of methods of assigning numbers to individual's attributes, such as degrees of skills or personality traits. Measurement theory is the formal and logical theory of necessary and sufficient conditions of attributing measures to objects or events. It is astonishing that test theory is not usually treated as a special area of measurement theory. Test theory (modern test theory, item response theory, IRT) is probabilistic and concerned with the relation of response probabilities to latent traits. It usually neglects the measurement principles (relations and operations, homomorphisms) necessary to represent observations by numbers and their relations (with the exception of the principle of specific objectivity, see below). Usually a representation by an analytical model (see below) is postulated and statistical tests are given to test the adequacy of the representation of empirical data by the model (item characteristic function, parametric IRT). Measurement theory usually is deterministic and axiomatic and concentrates on relations and operations. The relations typically are qualitative or ordinal. How can measurement theory be made probabilistic? How can test theory be made concerned about logical relations and homomorphisms (structure preserving representations)?

The more recent development of nonparametric IRT (Hemker et al., 1997; Holland and Rosenbaum, 1986; Irtel and Schmalhofer, 1982; Junker, 1993; Junker and Ellis, 1997; Mokken, 1971; Rosenbaum, 1984; Scheiblechner, 1995, 1998, 1999; Stout, 1987, 1990) is closer to measurement theory. The authors specify empirically (statistically) testable criteria necessary for unidimensional representations of psychological test data.

Typical psychological tests consist of a set of items ($i = 1, 2, \ldots, k$) which are supposed to 'measure' a common (unidimensional) psychological trait (e.g., verbal, numerical, spatial … intelligence; neuroticism, depression, …), and to which the subject ($v = 1, 2, \ldots, n$) chooses one answer out of several ordered categories ($h = 0, 1, 2, \ldots, m$). In most cases the test performance of a subject is summarized by the Likert score (simple integer score) or the trivial scoring function: all answers are weighted equally with constant weights for categories and the weights are summed over items. A major advance over this naive procedure was made by parametric rating scale models inspired by Rasch's measurement theory. A well-known parametric IRT model for ordered response categories is the rating scale model (RSM).

$$P(x = h|v, i) = \frac{\exp(h(\theta_v - \beta_i)\varpi_h)}{\sum_{x=0}^{m} \exp(x(\theta_v - \beta_i)\varpi_i)}$$

$$\beta_i = \sum_l a_{il}\eta_l + c$$

(Andersen, 1977; Andrich, 1978; Masters, 1982) where x is the response category of the answer ($x = 0, 1, 2, \ldots, h, \ldots, m$), θ_v the subject (ability) parameter, β_i the item (difficulty) parameter (with possibly a linear model with respect to basic components η_l) and ϖ_h the response category (attractiveness) parameter (with the normalizations for uniqueness of parameterization $\sum_i \beta_i = 0, \varpi_0 = \varpi_l = 0, c = 0$).

Models like this are an improvement over simple integer scoring because the fit of the model can be tested by a conditional likelihood ratio statistic (and the fit of the linear item parameter model can be tested by an additional likelihood ratio test). Thus we may detect, for instance, that no such psychological trait as conjectured does exist (in the sense of the model) or that not all items belong to the same psychological trait, or that the scoring of the categories may depend on the item and so on. However, the logistic form of the category characteristic curve (CCC), the function describing the probability of choosing category x; Hemker et al., 1996), the existence of three scales (subjects, items, categories) and their additive concatenation, the response category as scoring function (factor h or x in the exponent) are postulated all at once, and if the model does not fit there is no (or little) evidence as to which assumptions are failing (the situation is much better concerning the linear model for the item parameter assuming that the basic model fits).

An essential advantage of the probabilistic measurement theoretic formulation of such models is that no such risky

[†] Deceased.

assumptions are necessary simultaneously in advance, that each step can be tested separately, and that those axioms can be identified which are possibly responsible for lack of fit. The test of the logistic form of the CCC and of the adequacy of the scoring function are rendered unnecessary by a measurement theoretic test model because no analytic CCC is assumed and there is no need of a scoring function (beyond an ordinal scale). The existence of ordinal scales for subjects and items can be tested separately from each other and independently of the postulation of an additive conjunction of interval scales for subject and item parameters (see for instance the exponent of the CCC of the RSM). Should the test of additive item and subject parameters be successful, we could then test an additive conjunction with an appropriately constructed scoring function (an interval scale for the response category scale). Independently of the outcome of the latter test, and separately from each other, we can test additive models for the item and subject parameters. In short, the advantages of measurement theoretic test models are the stepwise construction of a hierarchy of gradually more complex models, and an exact specification of the level at which the construction of the models possibly might fail.

Making Deterministic Axiomatic Structures Probabilistic

A successful translation of the deterministic axioms of a measurement structure into a probabilistic version grants the formal rigor of measurement structures while being at the same time probabilistic, i.e., provides a statistical theory for testing the validity of the model. Some possibilities of probabilization of axiomatic measurement structures are expounded by Heyer and Nideree. A version of a probabilistic additive conjoint structure will be given below (ADISOP, CADISOP). The most basic proposition of most measurement structures is that an object a is in some respect larger, more attractive, etc., than an object b, symbolically, $a \gtrsim b$, where the objects and the relation are assumed to be empirically observable and deterministic (for example, two weights and the turning of the scale of an idealized balance). How could this be translated into the single, inconsistent qualitative reactions (right–wrong, strongly refuse–strongly prefer) of subjects to single items? Falmagne (1979) suggested replacing $a \gtrsim b$ with the probability $P(a \gtrsim b)$ of observing a 'larger reaction' (for example, the judgment a is louder than b) upon repeated presentation of the pair (a, b) (or of observing a larger median reaction time, etc.). This is feasible for some psychophysical tasks, but totally impractical for usual psychological intelligence and personality tests where individually identifiable items reasonably can be presented only once to subjects (a repetition would change the meaning of the item). For these I suggest replacing the larger relation $a \gtrsim b$ or the probability of a larger relation $P(a \gtrsim b)$ with the relation of two probabilities, $P(X \leq x \mid a) \leq P(X \leq x \mid b)$, where $P(X \leq x \mid a)$ is the probability of observing a specified or 'smaller' reaction than x (for example, a predominantly right answer) on presentation of a. If there are more than two graded responses (rating scales, polytomous ordered response categories) available for items I suggest using the concept of stochastic dominance (see below). The probabilities are an intermediate layer intervening between the empirical and numerical relational system of a measurement structure (Roberts, 1979: p. 274) and even estimates from single observations turn out to be sufficient for statistical testing of larger restrictive measurement structures (for example, if a subject 'moderately agrees' to an item his cumulative probabilities of 'disagreeing' or 'moderately disagreeing' are set to 0 and his probabilities to 'moderately agree' or 'agree' are set to 1).

A stochastic variable W dominates a stochastic variable V, if the probability that W is smaller/equal than an arbitrary value x is smaller/equal to the probability that V is smaller/equal to x for all x, that is:

Definition. Stochastic dominance $\gtrsim$.

$$W \gtrsim V \Leftrightarrow F_w(x) \leq F_v(x) \quad \text{for all } x$$

where $F(x) = P(X \leq x)$

Definition. Strict stochastic dominance $\succ$.

$$W \succ V \Leftrightarrow F_w(x) \leq F_v(x) \quad \text{for all } x$$

and

$$F_w(x) < F_v(x) \text{ for at least one } x.$$

If $W \gtrsim V$ we also write $F_w(x) \gtrsim F_v(x)$ for the cumulative distribution functions (CDFS) and similarly for $\succ$.

As an example, the students of a lecture on statistics rate their attitude (preference/aversion) toward the usefulness of statistics for the understanding of psychology (U) (high/low), the importance of statistics for professional success (P) (important/unimportant), the time needed to learn statistics (T) (acceptable/annoying), and the work with formulae (F) (fascinating/disgusting) on a four point scale. The students reject formulae most, the time requirements for learning statistics next, and think more positively about the importance of statistics for professional success, and most for understanding psychology (Table 1).

If the preference (aversion) for one option stochastically dominates the preference (aversion) for another option, then the probability that the first option is accepted (rejected) is always greater than or equal to the probability that the second option is accepted (rejected) if the ordinal scale is dichotomised (transformed into a binary scale) at an arbitrary position. This invariance makes the stochastic order relation independent of a great variety of different empirical operationalizations (experimental arrangements for the determination of larger relations).

If the deterministic order relations are made probabilistic by the concept of stochastic dominance, then the concept of empirical verification and falsification of single relations by

Table 1 Stochastic dominance of variables, $F \gtrsim T \gtrsim P \gtrsim U$ (cumulative distribution functions: 1, positive; 2, moderately positive; 3, moderately negative; 4, negative)

	Rating			
Category	*1*	*2*	*3*	*4*
U	0.417	0.778	0.972	1.000
P	0.183	0.563	0.817	1.000
T	0.064	0.278	0.714	1.000
F	0.063	0.175	0.588	1.000

single observations must be replaced by statistical testing of order relations (Bartholomew, 1983; Robertson et al., 1988) and by the likelihood principle of model testing. Suppose we want to test the validity of the above order relation $F \gtrsim T \gtrsim P \gtrsim U$ for all subjects of a specified population. Then there are many nonparametric tests to test the null hypothesis of no differences and we can test the hypothesis of the same (similar) order of items for all subjects by testing samples from all relevant subpopulations of subjects; we can also calculate the likelihood of all possible orders and choose the maximum likelihood order.

The Construction of Ordinal Scales for Subjects and Items: the Isotonic Probabilistic (ISOP) Measurement Structure

The most basic level of the hierarchy of models for the construction of unidimensional scales for psychological traits is the construction of ordinal scales for subjects and/or items. The construction of two scales simultaneously for two sets of objects (e.g., subjects and items) has been referred to elsewhere as biscalability or biorder (Guttman scale). Because these concepts are not probabilistic or appropriate only in psychophysical contexts they are not reported here (but they presumably could be made probabilistic).

In the example of rating different aspects (items i) of a negative attitude toward statistics each subject v generates a response vector $\boldsymbol{x}_v = (x_{v1}, x_{v2}, x_{v3}, \ldots)$ of ratings of aversions for items 1, 2, 3, …. The subjects possibly are characterized by a unidimensional latent trait 'aversion against statistics.' The response distributions of the response variables X_{vi} of subjects $v = 1, 2, \ldots, n$ and items $i = 1, 2, \ldots, k$ can be ordered by stochastic dominance – but how can the subjects (items) be ordered? The response vectors of the subjects give no simple order of the subjects with respect to aversion against statistics. The order relation on the set of subjects (items) has to be defined by a new concept. We can say that one subject dominates another (with respect to aversion against statistics) if all his/her ratings are larger than or equal to the ratings of the other subject. Of course, many pairs of subjects will not be ordered by this relation (the order is partial). The subjects cannot be compared directly, but by means of their response vectors, i.e., by means of the items as instruments; and therefore the measurement structure is called an instrumental probabilistic pair comparison system (Irtel and Schmalhofer, 1982; bioorder).

The set of observed multidimensional polytomous response vectors $\boldsymbol{x} = (x_1, x_2, x_3, \ldots)$ ordered by the relation.

$$\boldsymbol{x} \text{ dominates } \boldsymbol{y} = x \gtrsim y \text{ iff } x_i \geq y_i \text{ for all } i$$

is a multidimensional coordinatewise partial order.

If several variables are stochastically ordered and their order coincides with a scalar parameter of their distributions then they are said to be *isotonic* (similarly ordered) in that parameter. That parameter may represent a known attribute of the variables or a scale value on a latest dimension (for example, aversion in **Table 1**).

Definition. Isotonicity. A family of variables X_i or distribution functions F_i depending on a parameter p_i is isotonic in p_i if $X_i \prec X_j \Leftrightarrow F_i(x) \geq F_j(x)$ for all x of $p_i \leq p_j$.

In order to be able to construct ordinal scales for the set of subjects and items we now extend the concept of isotonicity to a two-dimensional array of variables X_{vi}, for example the aversions of several subjects for a set of attitude items or the preferences of visitors for various holiday arrangements. This is achieved by the axioms of the ISOP measurement structure.

Let A be a set of subjects $v, w, \ldots$, and Q a set of items $i, j, \ldots$. We associate with the Cartesian product $A \times Q$ a family of random variables X_{vi} with distribution functions $F_{vi}(x) = P[X \leq x \mid v, i]$ defined on the set $X = \{x, x', x'', \ldots\}$ of categories (values) of an ordered dependent ('achievement') variable X.

Definition. A system $\langle A \times Q, F\rangle$ where F is a family of distribution functions on $A \times Q$ is weakly, independently ordered (an ISOP model) iff the following axioms are valid.

Axiom W1. *Weak subject independence* holds if the response distribution of subject v dominates the response distribution of subject w at item i, then the response distributions of v are not dominated by w at any other item j.

Axiom W2. *Weak item independence* holds if the response distribution of item i dominates the response distribution of item j for some subject v, then the response distributions of i are not dominated by j for any other subject w.

Axiom LI. *Local independence* holds if the probability of the response vector of a subject v is the product of the probabilities of the single responses

$$F(x_{v1}, x_{v2}, x_{v3}, \ldots | v) = F(x_{v1}|v)F(x_{v2}|v)F(x_{v3}|v)\ldots$$

(Scheiblechner, 1995, 1998).

We can write W1 in the form if $F(X_{vi}) \prec F(X_{wi})$ for some i, then $F(X_{vj}) \gtrsim F(X_{wj})$ for all j meaning that the stochastic dominance relations between subjects are similar (isotonic) for all items. Analogously for W2, the stochastic order of items is similar for all subjects and therefore independent of subjects (except for ties: Scheiblechner, 1995, 1998). The pair $\langle A \times Q, F\rangle$ is called an instrumental probabilistic pair comparison system. Axiom W1 orders subjects by the stochastic order of their distribution functions of reactions to items and states that the orders are independent from the specific items used, except for possible ties (subjects are similarly ordered by items). This independence of the orders of subjects of the items is the basis for the construction of the latent unidimensional subject ordinal scale or person parameter. Axiom W2 similarly states that the stochastic order of items is independent of subjects and this is the base of the item ordinal scale or item parameter. The axioms are independent and can be statistically tested independently of each other.

If the system of variables is an ISOP, then the distribution functions can be ordered, and there exist ordinal scales $\varphi_A(v)$ on the set of subjects, and $\varphi_Q(i)$ on the set of items, such that the distribution functions of the family X_{vi} of random variables are functions of the two ordinal scales (Scheiblechner, 1995):

$$F_{vi}(x) = F(x|\varphi_A(v), \varphi_Q(i))$$

where $\varphi_A(v)$ and $\varphi_Q(i)$ are ordinal scales.

The two ordinal scales are 'order independent relative to subsystems' (Scheiblechner, 1995), that is, the ordinal scale for subjects does not depend on the set of items and the ordinal scale for items does not depend on the set of subjects, both except for ties (the corresponding property is called specific

objectivity or sample independence in the logistic Rasch model; Andersen, 1995; Rasch, 1960).

The following implications of ISOP can be tested:

1. the validity of axiom W1 for each item and for the total test (= set of items);
2. the validity of axiom W2 for each response vector (subject) and for the total sample;
3. the validity of the coordinatewise partial order of response vectors for each subset of items of the test and for the total test;
4. the validity of the coordinatewise partial order implied by one part of the test for another part of the test; and
5. the validity of a large number of manifest partial and linear isotonicities.

Especially the facility to test separately the single axioms of the measurement structure and the separate tests for different parts and aspects (formal relations) of the data are a measurement theoretical advantage of the ISOP measurement structure.

The Construction of Interval Scales for Subjects and Items: the Additive Conjoint ISOP (ADISOP) and the Complete ADISOP (CADISOP)

The axioms of the ISOP model are sufficient for the construction of two separate unidimensional ordinal scales on the sets of subjects and items. We now strengthen the ISOP model by additional axioms in order to derive ordered metric/interval scales for subjects and items. First the ADISOP model is considered which is an ISOP model with a cancellation axiom added that relates the subject and item factors.

Definition. A probabilistic pair comparison system $\langle A \times Q, F_{vi}(x)\rangle$, where $F_{vi}(x)$ is a family of cumulative distribution functions defined on $A \times Q$, is an ADISOP iff the axioms of the ISOP model and the following axiom are satisfied.

Axiom Co. Validity of cancellation of subjects and items up to order o.

(Similar structures are the PACOME models, Falmagne, 1979). An explanation of cancellation up to order o is given in Scheiblechner (1999) and Scott (1964). It means that a reduction of the item parameter by an amount δ can be compensated by an increase in the subject parameter by the same amount δ. The subject and item parameters 'cancel.' Or if we attribute real values to subjects and items then the order of their sum is isotonic in the order of their distribution functions. For example, if the 'ability' parameter of subject v plus the 'easiness' parameter of item i is larger than or equal to the sum of the ability and easiness parameter of subject w and item j, then the distribution functions are isotonic in the sum, i.e., $F_{vi}(x) \precsim F_{wj}(x)$ and the reverse.

If the system $\langle \boldsymbol{A} \times \boldsymbol{Q}, \boldsymbol{P}\rangle$ is an ADISOP model then the order of the CDFs is identical to the order of the sums of the subject and item parameters, i.e., there exist real functions ϕ_A, ϕ_Q on A and Q and $\psi_{A\times Q}(v,i) = \phi_A(v) + \phi_Q(i)$ on $A \times Q$ such that for all $(v, i), (w, j) \in A \times Q$

$$\phi_A(v) + \phi_Q(i) < \phi_A(w) + \phi_Q(j) \Leftrightarrow F_{vi}(x) > F_{wj}(x)$$

for some x or $\Leftrightarrow$ $F_{vi}(x) \prec F_{wj}(x)$

The scales ϕ_A, ϕ_Q are unique up to positive linear transformations, i.e., they are interval scales if Axiom Co is valid up to arbitrary order o, and they are ordered metric scales if Co is valid up to the finite order $(\min\{n, k\} - 1)$ where n and k are the number of elements of A and Q, respectively. Ordered metric scales are unique up to monotonic transformations which leave invariant the order of the (finite number of) intervals. Ordered metric scales are approximations to interval scales. The cumulative response distribution functions on the product set of subjects and items

$$F_{vi}(x) = F(\varphi(x); \varphi_A(v) + \phi_Q(i))$$

are antitonic (do not increase) in $\psi_{A\times Q}(v,i) = \phi_A(v) + \phi_Q(i)$ and isotonic (do not decrease) in $\varphi(x)$, where $\varphi(x)$ is an arbitrary strictly isotonic function of x (i.e., is an ordinal scale). The functions $\phi_A(v)$ and $\phi_Q(i)$ are interval scales (ordered metric scales) with an arbitrary origin (additive constant) and a common unit (scale factor).

The ADISOP model can further be strengthened to a model where the response category parameter also is additive to the item and subject parameter. The axiom W3 strengthens Co and extends cancellation from subjects and items to include subjects and instrumental variable (response category) and items and response variable.

Definition. A complete ADISOP (CADISOP) model is an ADISOP with axiom (W3) (and RS) replacing axiom Co:

Axiom W3. Weak instrumental variable independence holds if:

$F_{vi}(x) < F_{vi}(x')$ for some v and some i, then $F_{wj}(x) \leq F_{wj}(x')$ for all w and all j.

Axiom RS. Restricted solvability. Certain restrictions apply to the solution of stochastic dominances.

The weak instrumental variable independence axiom W3 is an independence axiom like W1 and W2 and makes the factor of the coding of the response variable X independent of the experimental factor of subjects and of the factor of items (for each value x of the instrumental variable the biorder of subjects and items is the same). It makes the coding of the instrumental variable another additive factor. Cancellation is no longer restricted to subjects and items but includes the response variable factor (an increase of the response category parameter by an amount c can be compensated by an equal increase in the subject or item parameter). Axiom RS grants that the system of equations and inequalities representing stochastic dominances has a solution under certain restrictions (e.g., if the CDF of subject v at item i and category x is between the CDFs of subjects u and w at item j and category x, then there is a subject s whose CDF at j and x' is equal to the CDF at (v, i, x); similarly for the item and category factors (Roberts, 1979: p. 218).

If the pair comparison system $\langle A \times Q, P\rangle$ is CADISOP then there exist scales ϕ_A on A, ϕ_Q on Q and ϕ_X on X such that for all $(v, i, x), (w, j, x') \in A \times Q \times X$

$$\phi_X(x) - (\phi_A(v) + \phi_Q(i)) \leq \phi_X(x') - (\phi_A(w) + \phi(j)) \Leftrightarrow F_{vi}(x) \leq F_{wj}(x').$$

The scales are interval scales with a common unit (scale factor). The 'system characteristic function'

$$F_{vi}(x) = F(\phi_X(x) - (\phi_A(v) + \phi_Q(i)))$$

is strictly isotonic in its argument (Scheiblechner, 1999).

The Integration of Linear Models for Item and Subject Parameters in Additive Models: Linear ISOP, ADISOP, and CADISOP Models

The ADISOP can be combined easily with an additive model for the item parameter:

$$\varphi_Q(i) = \sum_l a_{il}\eta_l$$

where a_{il} are known coefficients specifying the frequency or amount of 'basic components' l present in item i, and η_l are the (easiness, attractiveness) parameters of the basic components.

The matrix **A** with the coefficients a_{il} is termed the (item) model matrix. The estimation of the linear item model can easily be integrated into the algorithm for fitting an ADISOP model (Scheiblechner, 1999).

It is interesting to note that the coordinatewise partial order of items implied by the model matrix **A** can be tested in ISOP model with only ordinal scales for items (and subjects) available. The full linear model for items with interval estimates for the basic parameters, however, requires interval scales for items (and subjects) and thus an ADISOP. Thus in a model with latent scales and latent concatenations the question of additivity of item components cannot be separated completely form the question of additivity of item and subject parameters.

It is possible also to include an additive model for the subject parameter

$$\varphi_A = \mathbf{B}\cdot\delta$$

where φ_A is the column vector of subject parameters, **B** the (n, k) matrix of 'dosages' of (basic) treatments which were given to subjects and δ the column vector of 'gains' (increments of the subject parameter) caused by the treatments (the first column of **B** may be a column of 1s for an additive constant). The necessary supplements of the estimation algorithm are similar to the inclusion of the item model. Both extensions can be performed in a CADISOP model, of course.

The linear ADISOP/CADISOP models with the inclusion of linear models for items and/or subjects can be tested versus the respective unrestricted models by likelihood ratio tests.

Discussion and Future Directions

Probabilistic measurement models are known in psychophysics and utility theory, for choice behavior and pair comparisons, for decision making under uncertainty, and measurement of subjective probability. Not much attention is usually placed on estimation of parameters (probabilities and scale values) and statistical testing of fit (a rare but well-known and unsatisfactory (biased) example is Mosteller's test of fit for pair-comparison data). The probabilistic version of the additive conjoint measurement structure by Falmagne (1979) is a noteworthy psychophysical example for the statistical treatment of measurement models. Some statistical principles of estimation and testing could be borrowed and adapted from isotonic regression (Dykstra, 1983; Robertson et al., 1988). The tests concern stochastic (probabilistic) transitivity or consistency. The ISOP model is a weak stochastic transitivity model for a product structure $(A \times Q)$ and satisfies the statistical standards of probabilistic test theory. There are many statistical tests for various assumptions (axioms) in various parts of the data (see above, end of Introduction). The tests are based on generalizations of Goodman and Kruskal's index for ordinal association (1954, 1959, 1963, 1972). The U-model by Irtel and Schmalhofer (1982) is ISOP for dichotomous items and is its immediate precursor. The authors suggest an optimal contrast test for the bi-isotonic model (based on Robertson et al., 1988).

It is possible to specialize the ISOP model so as to pair comparison data (and to compare it with other probabilistic measurement models; Fishburn, 1973). It is possible also to generate probabilistic polynominal conjoint measurement structures (Roberts, 1979) by exactly the same steps as with ISOP structures (there is little use for those structures at present because no interesting empirical procedures for gathering corresponding data are known). Finally, it is possible to apply ISOP structures to geometrical measurement structures (multidimensional scaling, Kruskal, 1964a,b).

A permanent problem in psychological test theory is the modeling of the speed–accuracy trade-off for items where the speed and accuracy of the response are recorded at the same time. This problem can be attacked along the lines of isotonic probabilistic measurement structures whereby some interesting new problems arise.

See also: Classical (Psychometric) Test Theory; Factor Analysis and Latent Structure: IRT and Rasch Models; Psychometrics.

Bibliography

Andersen, E.B., 1977. Sufficient statistics and latent trait models. Psychometrika 42, 69–81.

Andersen, E.B., 1995. Polytomous Rasch models and their estimation. In: Fischer, G.H., Molenaar, I.W. (Eds.), Rasch Models: Foundations, Recent Developments and Applications. Springer, New York, pp. 271–292.

Andrich, D., 1978. A rating formulation for ordered response categories. Psychometrika 43, 561–573.

Bartholomew, D.J., 1983. Isotonic inference. In: Kotz, S., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences, vol. 4. Wiley, New York, pp. 260–265.

Dykstra, R.L., 1983. An algorithm for restricted least squares regression. Journal of the American Statistical Association 78 (384), 837–842.

Falmagne, J.C., 1979. On a class of probabilistic conjoint measurement models: some diagnostic properties. Journal of Mathematical Psychology 19, 73–88.

Fishburn, P.C., 1973. Binary choice probabilities: on the varieties of stochastic transitivity. Journal of Mathematical Psychology 10, 327–352.

Goodman, L.A., Kruskal, W.H., 1954. Measures of association for cross classifications, I. Journal of the American Statistical Association 49, 732–764.

Goodman, L.A., Kruskal, W.H., 1959. Measures of association for cross classifications, II. Journal of the American Statistical Association 54, 123–163.

Goodman, L.A., Kruskal, W.H., 1963. Measures of association for cross classifications, III. Journal of the American Statistical Association 58, 310–364.

Goodman, L.A., Kruskal, W.H., 1972. Measures of association for cross classifications, IV. Journal of the American Statistical Association 67, 415–421.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., Junker, B.W., 1996. Polytomous IRT models and monotone likelihood ratio of the total score. Psychometrika 61, 679–693.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., Junker, B.W., 1997. Stochastic ordering using the latent trait and the sum score in polytomous IRT models. Psychometrika 62, 331–347.

Holland, P.W., Rosenbaum, P.R., 1986. Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics 14, 1523–1543.

Irtel, H., Schmalhofer, F., 1982. Psychodiagnostik auf Ordinal skalenniveau: Meßtheoretische Grundlagen, Modelltest und Parameterschätzung. Ordinal psychological diagnostics: measurement theoretic foundation, tests of the model and estimation. Archiv für Psychologie 134, 197–218.

Irtel, H., 1995. An extension of the concept of specific objectivity. Psychometrika 60 (1), 115–118.
Junker, B.W., 1993. Conditional association, essential independence and monotone unidimensional item response models. Annals of Statistics 21, 1359–1378.
Junker, B.W., Ellis, J.L., 1997. A characterization of monotone unidimensional latent variable models. Annals of Statistics 25, 1327–1343.
Kruskal, J.B., 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 115–129.
Kruskal, J.B., 1964b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29, 115–129.
Masters, G.N., 1982. A Rasch model for partial credit scoring. Psychometrika 47, 149–174.
Mokken, R.J., 1971. A Theory and Procedure of Scale Analysis. Mouton, Paris.
Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen.
Rasch, G., 1961. On General Laws and the Meaning of Measurement in Psychology. In: Proceedings of the IVth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4. University of California Press, Berkeley, pp. 321–333.
Roberts, F.S., 1979. Measurement theory. In: Rota, G.-C. (Ed.), Encyclopedia of Mathematics and Its Applications, vol. 7. Addison-Wesley, London.
Robertson, T., Wright, F.T., Dykstra, R.L., 1988. Order Restricted Statistical Inference. Wiley, New York.
Rosenbaum, P.R., 1984. Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika 49 (3), 425–435.
Scheiblechner, H., 1995. Isotonic ordinal probabilistic models (ISOP). Psychometrika 60 (2), 281–304.
Scheiblechner, H., 1998. Corrections of theorems in Scheiblechner's treatment of ISOP models and comments on Junker's remarks. Psychometrika 63, 87–91.
Scheiblechner, H., 1999. Additive conjoint isotonic probabilistic models (ADISOP). Psychometrika 64 (3), 295–316.
Scott, D., 1964. Measurement models and linear inequalities. Journal of Mathematical Psychology 1, 233–247.
Stout, W.F., 1987. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika 52, 589–617.
Stout, W.F., 1990. A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. Psychometrika 55, 293–325.

Textbooks

Richard Venezky, University of Delaware, Newark, DE, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15631–15636, © 2001, Elsevier Ltd.

Abstract

Textbooks are simultaneously surrogate curricula, cultural artifacts, and commercial products. As surrogate curricula, they define areas of study and normal pathways through these areas. As cultural artifacts they reveal, particularly through their latent or hidden curricula, particular societal values and norms. As commercial products, they reflect particular states of print production, distribution, and use. Textbooks evolved over the past 2000 years from original works and compendia of such works, to illustrated and tightly structured presentations of subject areas, accompanied quite often by pedagogical directions, assessment materials, and practice exercises. Across countries, subject areas, and grade levels, the position of textbooks in the educational process varies. Equally variable are the locus of control for textbook content and the nature of the groups that vie for this control. Due to the high demands that tertiary level textbooks make on capital and expertise, many developing countries, particularly in Africa, have become dependent upon the larger, wealthier countries for these products. The special role that textbooks play in education has been threatened in the past by programmed learning and computer-assisted instruction. In the future the threat will most likely derive from the digitized world of wireless communications and multimedia computers.

The textbook remains today what Altbach calls a 'virtual icon of education' (Altbach, 1991, p. 242), yet controversy surrounds nearly every aspect of its production, dissemination, and use, and its prospects for survival are uncertain in a high bandwidth, digitized, and wireless world. Textbooks in modern society function simultaneously as surrogate curricula, cultural artifacts, and commercial objects. Their development over the past two millennia has been driven by the evolution of formal schooling and of the printing arts, as wells as by the development, particularly since the start of the twentieth century, of educational psychology. Important differences in textbook form and use, and of the issues surrounding these areas, exist across the schooling levels – elementary, secondary, and tertiary; curricular areas; and countries. The form that textbooks will assume in the future is projected to change with the fuller integration of information and computer technologies in schooling.

Definition and Evolution

Definition

Textbook usually is interpreted in a general sense, referring to texts that are especially prepared for the study of a subject. The prototypic textbook has as its core a sequence of topics that are arranged to lead a student through a subject or course of study, accompanied on the periphery by occasional guiding statements, summaries, or the like, exercises for practice, questions for self-assessment, and assistance for the instructor. All textbooks possess the core element but differ in representation of the periphery. Primary level textbooks tend to possess all of the peripheral attributes; secondary level ones, most, and tertiary level ones, fewer. Nevertheless, all textbooks are viewed as surrogate curricula, defining a prescribed codification of knowledge.

Historical Development

Until at least the Enlightenment, textbooks for advanced study were either original works, such as the grammars of Donatus and Pricisian, Boethius's *Geometry*, and Ptolemy's *Mechanisms of the Heavens*, or abridged compendiums of these works, such as Martinus Capella's *The Marriage of Mercury and Philology* (Cubberley, 1920). For children, specially developed materials originate with the *Orbis Pictus* of Comenius (1658/1887), which was the first illustrated schoolbook published for introductory learning. By the end of the seventeenth century primers for teaching reading were published in Europe with the alphabet, lists of syllables and words, as well as religious material: the Lord's Prayer, the Creed, and so on. From the influence of Rousseau and Pestalozzi came, by the second decade of the nineteenth century, carefully sequenced materials for learning, along with occasional suggestions for the instructor. With the adoption of tax supported schooling, particularly in Europe and North America, both the market and the need for textbooks expanded dramatically, leading to changes in both size and content. Coincidentally, improvements in printing technology and paper making throughout the nineteenth century were also reflected in textbooks. With the rise in educational psychology, beginning at the end of the nineteenth century, instructional approaches became more differentiated and more scripted, leading by the post-World War II period, to separate teacher manuals to accompany many elementary and secondary level textbooks.

Surrogate Curriculum

According to Altbach and Kelly (1988, p. 3), '(Textbooks) constitute the base of school knowledge, particularly in third world countries where there is a chronic shortage of qualified teachers.' But even in industrialized countries textbooks are the primary codification of school knowledge and the principal guide for teachers. Teachers at the elementary levels, who are responsible for teaching a broad curriculum in self-contained classrooms, choose out of necessity to use textbooks, along with workbooks and teacher guides. At higher levels of schooling the textbook is as much a prop for students, a *vade mecum* for a course, as it is a guide for the instructor.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.95064-5

Grade Level Differences

Although textbooks are used at all levels of education and in all subjects, their importance varies not only across countries but also across grade levels and curricular areas. At the primary levels their use is limited primarily to literacy, at least until the students, on average, can read well enough to gain meaning from other content area texts. In many school systems language arts and mathematics texts may be the only ones commonly used prior to the fourth or fifth year of schooling. Then, through the first years of graduate school, textbooks tend to be the basis of instruction for most subject areas.

Subject Area Differences

Thomas Kuhn has pointed out in *The Structure of Scientific Revolutions* that at the college level textbook use varies dramatically by field of study. Music, the graphic arts, and literature typically are taught through direct exposure of the student to artists' works, textbooks playing only a secondary role. In history, philosophy, and the social sciences, textbooks have a more significant role, but students usually are also assigned readings from the classics and from contemporary research reports. In contrast, the natural sciences are taught almost exclusively from textbooks (Kuhn, 1970). Within the natural sciences, textbooks are rewritten after each paradigm shift (or 'revolution,' in Kuhn's terms), with the revolution obscured and the past reinterpreted to show only its connection to the issues of the new paradigm.

Use of Textbooks by Teachers

How teachers use textbooks is a matter of concern for educational administrators. In developing countries where teachers may not be well trained, close adherence to the sequence and directions of the textbook is desired. In other countries, reliance upon textbooks, particularly at the primary levels, is not viewed positively. Studies of teaching practices, however, have tended to show that high percentages of elementary teachers do use textbooks as their primary guide for instruction (e.g., Shaver et al., 1979; Weiss, 1987). But other studies have shown that teachers, in general, do not follow their textbooks slavishly (e.g., Stodolsky, 1989).

Control of Textbook Content

In most countries a central ministry of education determines the general content of textbooks but the actually writing of the texts may be performed by persons outside the ministry, as in Russia, or within the ministry, as in Singapore. In the US, Canada, and Germany control over textbook content is vested at a provincial or state level, although the actual texts are produced by the private sector. In many developing countries, where indigenous textbook industries generally are not yet developed, multinational textbook companies exercise some control over content. French produced textbooks, for example, are used in many West African secondary schools (e.g., Togo); some of these texts portray sex roles and other culturally sensitive matters that are alien to West African cultures (Biraimah, 1988).

Imported Textbooks

At a postsecondary level and especially in technical subjects, textbooks are relatively expensive to produce, due to the cost of illustrations, complex diagrams, formulas, and the like. Many countries, including the smaller industrialized ones such as The Netherlands, Finland, and Luxembourg, must import textbooks for these levels, and the majority of the these texts are written in English. A further factor that affects the content of textbooks worldwide, but particularly at a secondary and tertiary level, is that a small number of highly industrialized countries dominate knowledge production and dissemination. Countries like the US invest large amounts of money in all fields of research and development, including educational research, and tend, therefore, to set many of the tertiary level standards worldwide for curriculum and assessment. Through academic institutions, research laboratories, libraries, and professional journals, countries like Britain, Germany, and the US exert both direct and indirect influence over textbooks in most of the rest of the world. That this knowledge is shared freely is offset partially by the reduction in cultural autonomy that results.

International Concerns over Content

Even within a world that recognizes, at least in principle, the right of each country to design and teach its own curriculum, cross-national conflicts over curriculum content have occurred. One example is the concern expressed by a number of countries to specific national treatments of twentieth-century history and particularly of World War II and the Holocaust. In 1982 The People's Republic of China and South Korea protested changes in Japanese history books that removed references to the Japanese massacre in Nanking and watered down the language previously used to describe other Japanese behavior toward conquered territories (Ketcham, 1986). More recently, an Austrian publisher of a high school history series was forced, through pressure from other countries, to stop referring to Nazi concentration camps as 'punishment camps.' History texts in the US have been attacked for over a half century for lack of fairness and objectivity in describing wars (Walworth, 1938), for being bland and condescending (Sewall, 1987), and for failing to attribute blame to individuals (FitzGerald, 1979). Similarly, science texts have been criticized for their handling of evolution (Skoog, 1984).

Cultural Artifact

The Latent Curriculum

The table of contents and index of a typical textbook define, to a large degree, the manifest curriculum offered by the text. But on top of this curriculum are a secondary set of messages, reflected by commission and omission, that constitute the latent or hidden curriculum (Cronbach, 1955). In part, this is what Williams calls the selective tradition of society (Williams, 1980). A mathematics textbook that constructs its word problems around interests of a consumer society – goods to be purchased, interest on loans, fees paid for services – conveys a different message from one that presents the same mathematical challenges built around revolutionary themes – distribution of land to the poor, triumph in sports over capitalistic societies, and so on. In China, for

example, mathematics textbooks during the Cultural Revolution conformed to the prevailing ideology; with a shift in politics, mathematics textbooks by 1980 were stripped of all praise for the Gang of Four and their policies (Kwong, 1988).

Class Structure

Texts for teaching reading have been studied worldwide for different aspects of their latent curricula, including gender roles (e.g., Lorimer and Long, 1979–1980; Zimet, 1972; Biraimah, 1988), need for achievement (e.g., McClelland 1961), moral/ethical values (e.g., de Charms and Moeller, 1962), and sociocultural ideologies (e.g., Ahier, 1988; Apple and Christian-Smith, 1991; Luke 1988). Drawing on Paulo Freire's claim that pedagogy for literacy, by definition, requires an educational ideology, some of these latter authors argue that in countries such as Britain, the US, and Canada, literacy textbooks are as much concerned with maintaining class structures and political hierarchies as they are with imparting reading and writing. Although there is a surface validity to these claims, as demonstrated particularly by textbooks from the 1950s and 1960s, they are difficult hypotheses to test on modern textbooks in industrialized countries, given the pressures on publishers to produce textbooks that are unbiased toward any sex, race/ethnicity, age, country of origin, or political belief. Texts today could be attacked as much for failing to promote cultural continuity as for over-representing any particular class or belief in society.

Commercial Object

In almost every country that has formal schooling, textbooks have some relationship to the commercial sector. In Singapore, for example, textbooks are designed and written by a Ministry of Education, but copublished with private sector firms. In the USSR, prior to its breakup, the Ministry of Education or a State Committee for Public Education designed the content of textbooks but authors were selected through public competitions and paid royalties for their efforts, which were published by state controlled printing houses. In Mexico, where textbooks have been provided free to the six grades of public and private elementary school, government support has been provided to develop what World Bank researchers have called 'the most dynamic private sector publishing industry in Latin America' (Neumann and Cunningham, 1982).

Concentrations of Expertise and Production Capacity

Textbook publishing and marketing require expertise in subject matter and pedagogy, along with capital, production and distribution capacity, and access to paper. These resources are found primarily in the major metropolitan centers of North America, Asia, and Europe, and are increasingly being dominated by the larger European conglomerates – Pergamon and Pearson in the UK, Bertelsmann in Germany, and Hachette in France. Even in the US, few major textbook publishing houses have remained independent. The rest of the world lies in the periphery of textbook publishing, even though many other large countries such as India, China, Mexico, and Egypt produce many if not most of their own textbooks and the latter two export regionally to smaller developing countries (Altbach, 1991).

Physical Characteristics

The appearance of textbooks – the quality of paper, typography, illustrations, bindings, and cover stock – are a function of printing technology, different national economies, grade level, and national (or regional/local) educational policies, such as the adoption cycle. In all school systems primary level textbooks usually are done on the cheapest paper, often with paper covers and limited use of color for illustrations. Russia textbooks, which the former Soviet government subsidized, were produced to last 2–3 years and were done with cheap paper, cover stock, and bindings. In contrast, textbooks in the US are produced to last 5–7 years and due to market factors have far more expensive components. But they also cost almost 10 times as much as their Russian counterparts (Mehlinger, 1991).

Authorship

Authors of secondary and tertiary textbooks generally write the works that bear their names. At the elementary level, on the other hand, writing houses or other anonymous writers often do the majority of the development. The authors usually are involved in planning and review, and may even draft a few model lessons, but their main function is to impart validity to the resulting product, through their professional reputations, affiliations, or their race/ethnicity, age, or sex. In the US, most elementary reading textbooks in the eighteenth and nineteenth centuries were authored by single individuals, with a few having two authors, or (rarely) three. By the early part of the twentieth century, most readers had two or three authors, by the 1950s, four, and by the 1970s, as many as eight authors and instructional consultants (Woodward, 1986).

Marketing and Adoption

In many countries, including some with national ministries of education, textbooks compete in a partially open marketplace. The selection of textbooks for schools, however, is often enshrouded in a complexity of national, regional, and local politics. In the US, where local control of schools and lay control of the school curriculum are fundamental tenets of the educational process, school board prerogatives for textbook selection often clash with particular citizen interests as well as with state and national politics. State-wide adoption practices, initiated originally for both economic and political reasons, have led to several heavily populated states (e.g., Texas, California) exerting unwarranted influence over the entire elementary and high school textbook trade.

The Future of Textbooks

In a science fiction classic entitled *A Round Trip to the Year 2000 or a Flight through Time*, written by Cook (1903), news and-novels in the year 2000 are no longer printed. Instead central 'bureaus' transmit voice recordings upon request from phonographic records. However, textbooks still exist and

specific mention is made of a 'text-book on Applied Psychology which has a place in every school in the nation' (Cook, 1903/1974, p. 109). Whether the projected survival of textbooks was considered as a positive or negative factor by the author is not known; however, while daily newspaper reading has declined steadily since the early 1950s and the novel has been declared dead at various times over the past decades, textbooks have continued to thrive at all levels of education and in all countries where formal education exists.

In the 1960s the main threat to the existing form of textbooks was the programmed textbook, a dull, tedious approach to sequencing subject matter in baby-sized bites, based on the behavioral conditioning studies that the psychologist B. F. Skinner performed with pigeons and rats (Reid, 1969). In the 1980s computer-assisted instruction was offered as a hightech alternative to textbooks and to teacher directed instruction, with the promise that each student could progress at his or her own pace – a goal that was seldom challenged by educators. Today the World Wide Web (WWW), hypertext, and inexpensive, book-sized computers that can display texts on high quality displays, are the assumed replacements.

For the developing world, which is still struggling to recruit and train sufficient numbers of teachers and to provide adequate buildings for schooling, the provision of textbooks written in the national languages and appropriate for their cultures remains a challenge. Similarly, smaller industrialized countries would like to find inexpensive approaches for producing or adapting tertiary textbooks to their languages rather than importing them. Finally, for the historian, the preservation of textbooks remains an issue. Few, if any major research libraries are willing to invest resources in preserving and cataloging textbooks, particularly elementary and secondary ones. Fortunately, a number of newer research centers and projects for the study of textbooks have been funded, including the Japan Textbook Research Center, The Teaching Resources and Textbook Research Unit in the Faculty of Education at the University of Sydney, Australia, the UNESCO International Textbook Research Network (in cooperation with the Georg Eckert Institute in Germany), and the Textbook Colloquium in the Open University Faculty of Arts in the UK.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Environments for Education; Learning Theories and Educational Paradigms; Learning to Learn; Literacy and Illiteracy, History of; Literacy and Orality; Marxist Approaches to Social Work; Mass Communication: Normative Frameworks; Mass Media: Introduction and Schools of Thought; Publishing as Medium; School Learning for Transfer.

Bibliography

Ahier, J., 1988. Industry, Children, and the Nation: An Analysis of National Identity in School Textbooks. Falmer, London & New York.

Altbach, P.G., 1987. Textbooks in comparative context. In: Thomas, R.M., Kobayashi, V.N. (Eds.), Educational Technology—Its Creation, Development and Cross-cultural Transfer. Pergamon Press, Oxford, UK, pp. 159–175.

Altbach, P.G., 1991. Textbooks: the international dimension. In: Apple, M.W., Christian-Smith, L.K. (Eds.), The Politics of the Textbook. Routledge, New York, pp. 242–258.

Altbach, P.G., Kelly, G.P. (Eds.), 1988. Textbooks in the Third World: Policy, Content and Context. Garland, New York & London.

Apple, M.W., 1986. Teachers and Texts: A Political Economy of Class and Gender Relations in Education. Routledge and Kegan Paul, New York.

Apple, M.W., Christian-Smith, L.K. (Eds.), 1991. The Politics of the Textbook. Routledge, New York & London.

Biraimah, K.L., 1988. Gender and textbooks: an African case study. In: Altbach, P.G., Kelly, G.P. (Eds.), Textbooks in the Third World: Policy, Content and Context. Garland, New York & London, pp. 117–143.

Comenius, J.A., 1658/1887. Orbis Sensualium Pictus. Bardeen, Syracuse, NY.

Cook, W.W., 1903/1974. A Round Trip to the Year 2000: Or a Flight through Time. Hyperion, Westport, CN.

Cronbach, L.J. (Ed.), 1955. Text Materials in Modern Education. University of Illinois Press, Champaign, IL.

Cubberley, E.P., 1920. The History of Education. Houghton Mifflin, Boston.

De Charms, R., Moeller, G.H., 1962. Values expressed in American children's readers: 1800–1950. Journal of Abnormal and Social Psychology 64, 136–142.

FitzGerald, F., 1979. America Revised: History Schoolbooks in the Twentieth Century. Little, Brown, Boston.

Ketcham, A.F., 1986. World War Two as viewed through secondary school textbooks: a report on research in process. Internationale Schulbuchforschung 8, 61–65.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. University of Chicago Press, Chicago.

Kwong, J., 1988. Curriculum in action: Mathematics in China's elementary schools. In: Altbach, P.G., Kelly, G.P. (Eds.), Textbooks in the Third World: Policy, Content and Context. Garland, New York & London, pp. 227–245.

Lorimer, R., Long, M., 1979–80. Sex-role stereotyping in elementary readers. Interchange 10, 25–45.

Luke, A., 1988. Literacy, Textbooks and Ideology: Postwar Literacy Instruction and the Mythology of Dick and Jane. Falmer, London.

McClelland, D.C., 1961. The Achieving Society. Van Nostrand, Princeton, NJ.

Mehlinger, H., 1991. American textbook reform: can we learn from the Soviet experience? In: Altbach, P.G., Kelly, G.P., Petrie, H.G., Weis, L. (Eds.), Textbooks in American Society: Politics, Policy, and Pedagogy. State University of New York Press, Albany, NY, pp. 145–160.

Neumann, P.H., Cunningham, M.A., 1982. Mexico's Free Textbooks: Nationalism and the Urgency to Educate, World Bank Staff Working Papers No. 541. The World Bank, Washington, DC.

Reid, J.M., 1969. An Adventure in Textbooks: 1924–1960. Bowker, New York.

Sewall, G.T., 1987. American History Textbooks: An Assessment of Quality. Educational Excellence Network, New York.

Shaver, J.P., Davis Jr., O.L., Helburn, S.W., 1979. The status of social studies education: impressions from three NSF studies. Social Education 43, 150–153.

Skoog, G., 1984. The coverage of evolution in high school biology textbooks published in the 1980s. Science Education 68, 117–128.

Stodolsky, S., 1989. Is teaching really by the book? In: Jackson, P.W., Haroutunian-Gordon, S. (Eds.), From Socrates to Software: The Teacher as Text and Text as Teacher (Eighty-Ninth Yearbook of the National Society for the Study of Education, Pt. 1). National Society for the Study of Education, Chicago, pp. 159–184.

Walworth, A., 1938. School Histories at War. Harvard University Press, Cambridge, MA.

Weiss, I.R., 1987. Report of the 1985–86 National Survey of Science and Mathematics Education. National Science Foundation. Research Triangle Park, NC.

Williams, R., 1980. Problems in Materialism and Culture. Verso, London.

Woodward, A., 1986. Taking teaching out of teaching and reading out of learning to read: a historical study of reading textbook teachers'guides, 1920–1980. Book Research Quarterly 2, 53–73.

Woodward, A., Elliott, D.L., 1990. Textbooks: consensus and controversy. In: Elliott, D.L., Woodward, A. (Eds.), Textbooks and Schooling in the United States (Eighty-ninth Yearbook of the National Society for the Study of Education, Pt. 1). National Society for the Study of Education, Chicago, pp. 146–161.

Zimet, S. (Ed.), 1972. What Children Learn in School. Grume & Stratton, New York.

Thematic Maps in Geography

David Forrest, School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Monmonier, volume 23, pp. 15636–15641, © 2001, Elsevier Ltd.

Abstract

Thematic maps represent information about a specific topic or about the relationship between topics. Producing effective thematic maps relies on selecting an appropriate base map and an understanding of the nature of the data being mapped. Once the representation method is selected, the individual symbols need to be designed, using variations in the visual or graphic variables. Having introduced these concepts, I go on to discuss basic strategies for representing data using point, line, and area symbols.

Introduction

Maps can be divided into three groups: topographic or reference maps; special purpose maps; and special topic or thematic maps. These three groups are closely related to three major types of map-use tasks (Keates, 1996). Topographic or reference maps are used to find out information such as the location of a place and, in theory at least, represent a balanced view of the physical and cultural landscape, including coastlines, roads, rivers, settlements, boundaries, etc. Special purpose maps are mostly associated with wayfinding, some being associated with trained users, such as hydrographic and aeronautical charts, and others with broader user groups, such as road and tourist maps. Sometimes referred to as 'special topic' maps, as the name implies, thematic maps emphasize a topic or theme, such as soils or population, and are used where there is a desire to learn about a particular topic. The topic can be relatively broad, such as global population distribution, or very narrowly focused, as for a map showing infant mortality for a specific year by districts within a city. It is also possible to show the relationship between phenomena, for example, a map showing how educational attainment and income levels compare and relate across a region. Comparatively rare at the outset of the nineteenth century, thematic maps as a group have become more numerous, varied, and intellectually richer as data have become available and representation techniques have developed (Robinson, 1982), and the current popularity of map 'mash-ups' on the World Wide Web has brought them to a much wider audience. This article examines the distinctive elements of thematic maps, including principal strategies and pitfalls.

Elements of Thematic Mapping

Thematic maps have two main components: a base map and the thematic overlay. Typically the base information comes from topographic or general purpose mapping, most likely today in the form of a cartographic database. The special topic information comes from a much wider range of sources. Specific field surveys may have been carried out on features such as geology, or vegetation, or nonmap sources such as government census, and other databases of attribute information that have some spatial reference may be used. In terms of map design, two approaches can be taken. In some cases, an integrated approach is used where reference and topic information is considered together and specific symbols chosen to represent each feature appropriately. Alternatively, in many cases, the base map is used with little or no modification, or perhaps all just reduced to gray, and the special topic information is overlaid on this base, typically using stronger colors. Many examples of the latter approach can be seen today using Google Maps as the base, overlaid by markers of thematic information. Regardless of which approach is taken, the aim of map design is to create a visual hierarchy of information, with the base information less visually prominent than the thematic component, so that the user may focus on the specific distribution or relationship, as shown in Figure 1.

An Appropriate Map Projection is Essential

Because the Earth is spherical and most maps are flat, the base map distorts the geometry of boundaries and other reference features (Maling, 1992). The amount of distortion depends in a large measure on the map's scale, defined as the ratio of distance on the map to distance on the Earth. On a 'large-scale' map covering a small area, such as a neighborhood or archaeological excavation, distortion will be minimal. By contrast, a 'small-scale' map covering a large area, such as a continent, hemisphere, or the entire World, must distort large shapes, distances, and directions, and may also distort angles and areas. By selecting an appropriate projection, the map author can avoid distortions that are likely to mislead readers. At larger scale, certainly at 1:250 000 scale and larger, it is likely that national topographic mapping will be used as the base, without any change of projection. At smaller scales, for most purposes in thematic mapping, an 'equivalent' or equal-area projection should be selected, but frequently authors simply use what is available, which can lead to concerns about fair representation of information. Often a distinct disadvantage of using readily available Web mapping (such as Google Maps or Bing Maps) as a basis for thematic maps is that the projections used for display are not equal area.

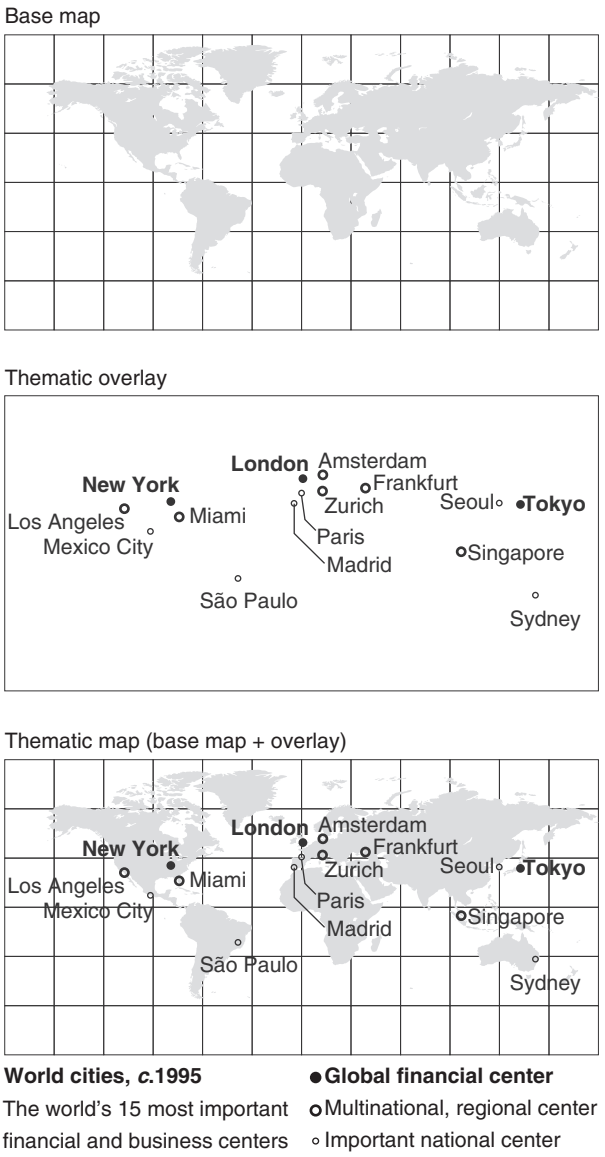

Figure 1 A thematic map consists of a base map and a thematic overlay.

Geocoding

In order to map data, there must be a way to plot the data on a base map. Where data have been expressly collected for mapping, this is likely to be straightforward. In other cases the map maker may have to create a link between the topic information and the base map. For census and many other socioeconomic data sets, the data have been gathered based on some form of political or administrative units, such as countries, states, wards, etc., depending on the scale of the data. As long as there is a field (attribute) in the data set that matches one in an appropriate boundary file, then the geocoding process is simply a matter of linking the attribute data to the spatial data (Goldberg, 2008). However, often there can be imprecise definitions of location in thematic data sets which may require manual effort to resolve any nonmatched data. It is also quite likely that boundaries will have altered over time, so care is required to ensure an appropriate set of boundaries is used. For some data sets it may be possible to use postal or zip codes to relate the topic information to a spatial data set. The Internet has greatly increased the availability of suitable base data sets, but beyond national- or state-level boundaries it may be difficult to find an appropriate data set for many parts of the world, and often there are quite high prices for detailed boundary data.

Understanding the Data to be Mapped

The data will likely require some processing and/or classification in order to be in a suitable form for mapping. The author must have an understanding of the nature of the data so that appropriate representation methods and symbols can be specified. First to consider is the spatial nature of the phenomena. Generally speaking, we can identify four types of spatial distribution: phenomena can be found a specific point location, or at least a location considered to be a point at the scale of mapping (such as a city on a very small-scale map); they can be distributed along lines; they can fill an area defined by some form of boundaries; or they can vary continuously over the surface. In most cases, the spatial data we have about the phenomena will be points, lines, or area boundaries. The relationship between the phenomena and the spatial data may not always be simple. For example, people live at specific locations, but typically the data we have is collected for enumeration areas, often to protect individual anonymity. So, although we know the data are about a point phenomenon, what we have for mapping relates to bounded areas. Similarly, it is not possible to collect data for every possible point over a surface, so data may be a sample of points, or lines on the surface. This relationship between the phenomenon and the spatial data is often an indirect one as frequently we are dealing with processed data or secondary sources. In choosing an appropriate mapping technique, it is important to think about what we are trying to map, not just the data we have.

The second aspect to consider is the nature of the attribute we are mapping. Often this is discussed in terms of the 'level of measurement' of the attribute. Many texts recognize four levels of measurement: nominal, ordinal, interval, and ratio. Nominal or qualitative data are where the distinction between members of the data set is by some form of naming or categorization; such as crop type or ethnic origin. Ordinal data are where there is some ordering or ranking of the data, such as, for settlements: National Capital, State Capital, County town, etc. Interval and ratio measurements are both quantitative and are often considered together, the key difference being the origin of the scale and how appropriate various mathematical operators are on the scale. This simple classification is satisfactory in many cases, but Chrisman (2002) and Forrest (1999) considerably extend the classification to help determine appropriate operations on and representations of the data. For example, it is frequently necessary to distinguish between count data, where the total number present in an area is mapped, and intensity data where the relative density or proportion of a phenomenon is mapped, even though both types are numerical and measured on a ratio measurement scale.

Issues relating to this difference between count data and intensity data are well illustrated by Figure 2. All three maps essentially illustrate the same phenomena, infant mortality in the eastern half of the United States. At the left, proportional circles describe differences in the number of infant deaths; a bigger circle unambiguously representing more deaths for that state than a smaller circle. In the center, a gray-tone treatment of the same data yields a potentially misleading map with dark symbols for states with relatively large numbers of deaths and comparatively light shades for states with fewer deaths. The dark tones suggest a health crisis in New York, Florida, and other relatively populous states and comparatively benign conditions in less-populous states such as Alabama and Mississippi. By contrast, the gray tones of the right-hand map offer a far more realistic portrait of the geographic variations in infant mortality which adjusts infant deaths for the number of live births. This recognizes that the size and population of the state will have an impact on our perception of the data. By relating the infant deaths to a more meaningful variable and mapping the proportion, we are in effect normalizing the data and removing the artifact of variable size states. If all states were the same size and population evenly distributed, we would not have to worry about such issues, but this is rarely the case.

Once the relationship among phenomenon, spatial data, and attribute measurement is understood, the map maker can choose an appropriate representation method. In many cases there is little choice, but, as we have seen in the example above, in some cases there may be several options (see Section 'Principle Cartographic Strategies').

The Visual or Graphic Variables

Having chosen the representation method, the individual symbols representing each class of information need to be designed or chosen. These symbols are the language of the map – the way the map maker communicates information to the reader.

Jacques Bertin (1983), a French semiologist who proposed a formal theory of graphic language, recognized eight visual variables: the two variables of location in the visual field (x- and y-coordinates) and six 'retinal' variables portraying a feature's character. The six retinal variables described by Bertin are size, shape, orientation, texture, color hue, and color lightness. Other cartographers have since developed Bertin's theory to include additional variables, such as the generally recognized third variable of color – color saturation, pattern, focus (relating to the sharpness of the symbol), and others (e.g., MacEachren, 1995). To some extent, how much Bertin's original set of variables should be extended depends on the interpretation of the variables. For example, some authors do not consider shape to be a possible variable for area symbols as the external shape of the area cannot normally change, preferring to introduce the variable 'pattern' for variation in the shape of the component elements covering the area, whereas others accept that the use of 'shape' refers directly to such components (as Bertin did), making 'pattern' unnecessary.

More recently, consideration has been given to variables that are possible on interactive computer displays, such as motion, duration, and synchronization which are not possible on paper maps (MacEachren, 1995; Slocum et al., 2009). While such dynamic variables must be viewed over time, their intention is to represent some other variable, not temporal change.

Table 1 summarizes the suitability of Bertin's six retinal variables for representing attributes with different levels of measurement. Generally speaking, it becomes obvious that those variables suitable for representing qualitative data are less suitable for representing quantitative data, and vice versa. If even such basic guidelines were followed by map authors, we would see far fewer maps representing data in inappropriate ways.

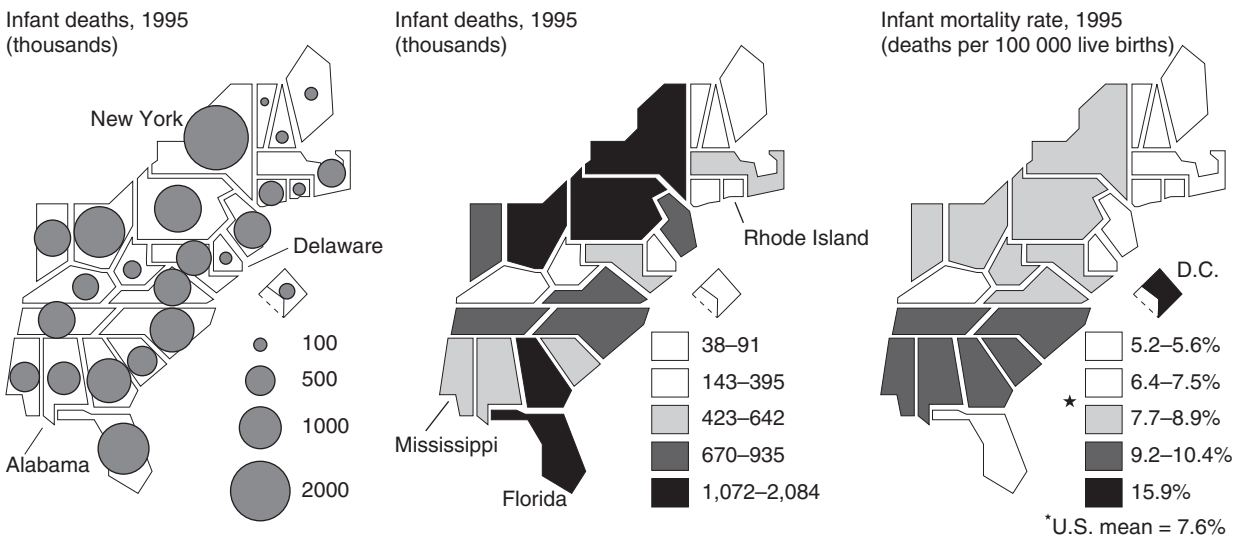

Figure 2 Comparative treatment of count (left and center) and intensity (right) data with symbols varying in size (left) and colour lightness (centre and right). Misuse of intensity symbols for count data on the center map yields a misleading pattern.

Table 1 Use of visual variables for different types of attribute measurement

Measurement	*Nominal*	*Ordinal*	*Interval/Ratio*
Visual variable			
Size	Poor	Usable	Extensive use
Shape	Extensive use	Some use, generally combined with size	Poor
Orientation	Good	Poor	Poor
Texture	Some use	Usable	Limited use
Color hue	Extensive use	Some use	Limited, but can use spectral ordering
Color lightness	Poor	Usable	Extensive use

While Bertin's visual variables have had significant influence on cartographic textbooks, further aspects of his theories of graphics are often overlooked although they do provide a key to what we are often trying to achieve with our symbols. Bertin also discussed the perception of the image and identified four possibilities. In *associative* perception, the impression is of similarity; we can see there are differences, but association between symbols is emphasized by having relatively low contrast between them. On the other hand, with *selective* perception we immediately separate the visual field into two (or more) groups, but none is perceived as more important than the other; typically this process involves greater contrast. *Ordered* perception is where we can see a clear ranking, in that one feature is seen as more important than another. Finally *quantitative* perception is where we get some impression of the relative numeric value of the phenomena. Figure 3 illustrates these four properties with the examples of point symbols, but this could also be done for lines or areas and use different graphic variables.

Principle Cartographic Strategies

Although the range of possible map topics is limitless and interactive graphics has greatly expanded the possible methods for visualizing geospatial data, for static mapping, an examination of printed atlases and cartographic textbooks will confirm that relatively few commonly used cartographic representation methods are in use. These methods (sometimes called different types of maps, rather than methods) are sufficient to portray most combinations of phenomena and mappable data. Although there are different ways to classify them, they will be treated here by way of the point, line, area, and surface representation methods. These representation methods are summarized in Figure 4.

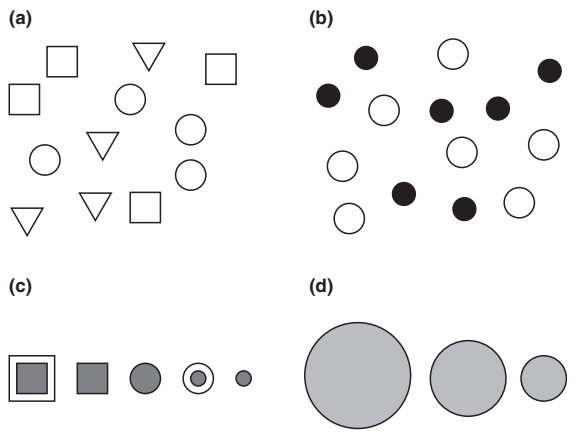

Figure 3 Bertin's classification of image perception: (a) associative, (b) selective, (c) ordered, (d) quantitative.

Representations Using Point Symbols

Seemingly a simple method, the dot distribution method represents information by a series of small dots, all of the same size and shape. It is possible, but uncommon for each dot to represent one individual; generally each dot represents a number of occurrences of the phenomena. For example, a dot may represent 100 people. Three issues need to be resolved: what is the ratio of features per dot, what size should the dots be, and where should they be placed. The first and second are interrelated. The dots need to be large enough for isolated dots to be noticed, but not too large, because it is the distribution pattern that is important, not the individual dot. Ideally in the densest areas dots should coalesce to become almost solid. Where to place the dots is often the biggest challenge. Although many computer systems can produce 'dot maps,' generally they use a random number generator to scatter dots within an enumeration area, which is not effective. Further information, such as a topographic map, is required to decide where the phenomenon most likely occurs. A good dot map is an excellent way of representing distributions, but the effort required to logically place the dots limits their use.

A more common way of representing quantitative data by point symbols is the proportional or graduated symbol. Here the size of the symbol varies in relation to the value of the topic, as shown on the left in Figure 2. With proportional symbols, this variation in size is in direct proportion to the value, such as the area of a circle or square increasing in direct relation to the value. Unfortunately, where there is a great data range, it can be impossible to create a scale where the smallest symbols are visible, but the largest are not too big to dominate the map. The solution is to divide the data into classes (much like choropleth maps) and have a fixed size symbols for each class, creating a set of graduated symbols. The most common symbol is the circle, which has a number of advantages, but other shapes are also used. Our perception of the relative size of squares is better than circles, so their use should be

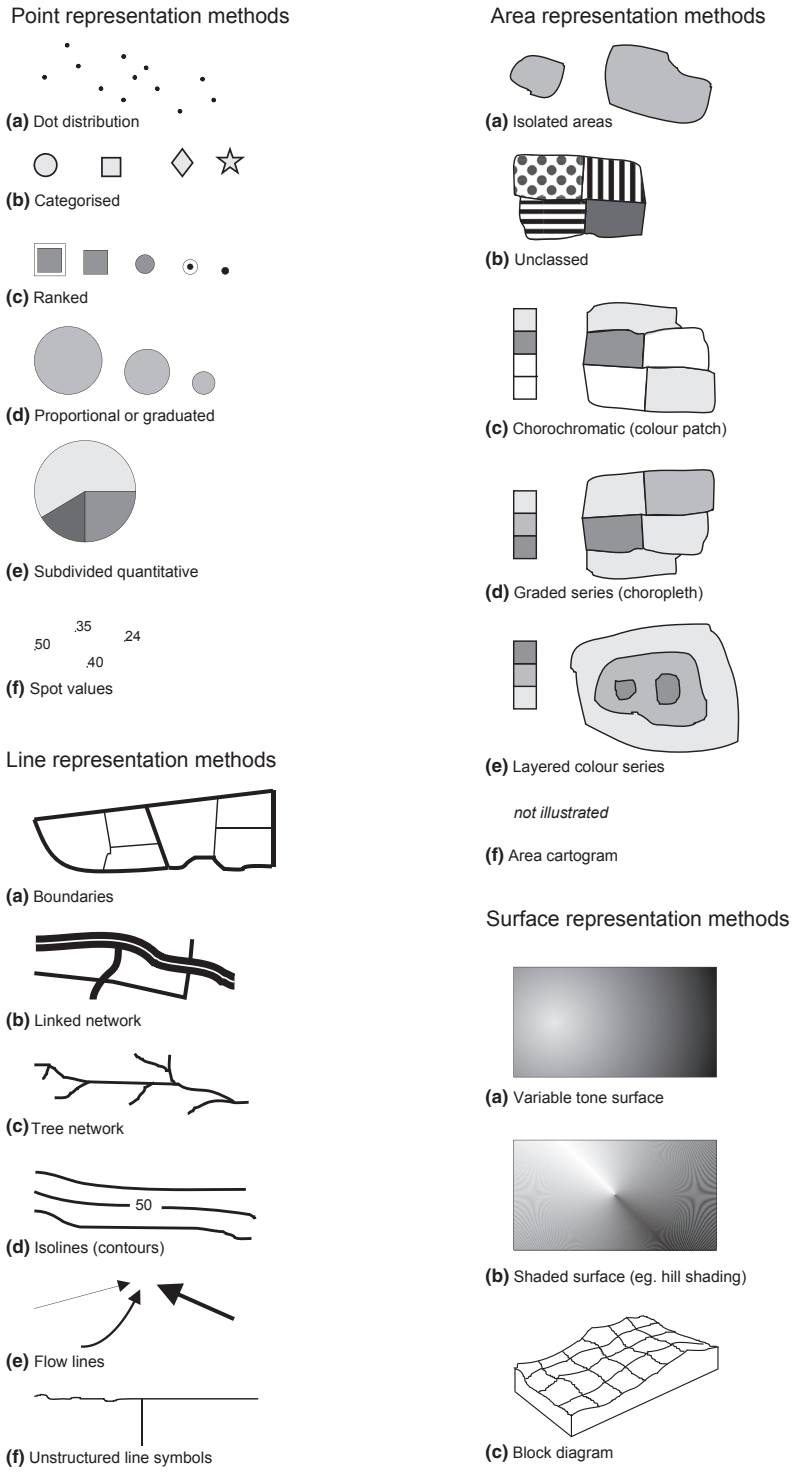

Figure 4 Basic cartographic representation methods (after Forrest, D., 1999. Geographic Information: its nature, classification and cartographic representation. Cartographica 36 (2), 31–53.).

more popular than it is. There is some use of pictographic symbols, particularly in journalistic maps, but our ability to estimate the relative value of more complex shapes is poor, so they are probably best avoided.

Where the data for the location can be broken down into components, a number of options are possible, such as pie charts, or various other graphic devices located at points. These can either be fixed in size, with the focus being on the relative proportions of the components, or varied in size like proportional and graduated symbols to convey both the overall total distribution and the breakdown of components.

For nonnumerical data, we simply show each category by varying the shape and/or hue of the symbol. The key design aspect here is to make the symbols appear different, but equivalent. When appropriate, pictographic symbols can be exploited here as the user can understand the meaning of the symbols without referring to the legend.

Representations Using Line Symbols

Boundaries are a common feature represented by lines. Generally the data about these are hierarchical, so the symbols chosen should emphasize the rank of the individual boundary. This could be done by simply varying the thickness of the lines, but most often a combination of thickness and line pattern (dashes) is used, with shorter, thinner dashes for lower-ranked boundaries.

Representing quantities by line symbols can either involve a network of lines or flow lines simply connecting pairs of points or areas. This tends to get little coverage in cartographic texts, perhaps because, in the writer's experience, they are often challenging to both design and produce. The obvious graphic solution is to vary the width of the lines in proportion to the variable being mapped, but there is a limit to the range of data that can be treated in this manner. One problem is that if two lines have the same value, but differ in length, the longer one may give the impression of having greater value. With the flow type of map, crossing lines pose further problems for both design and perception.

Another way of representing route information is a linear cartogram, the most famous example being the London Underground map, but a similar style has been used for many metropolitan subway systems, rail and bus networks. Space is distorted for clarity and simplicity purposes, often with the angle of lines limited to a few directions (usually 6 or 8). Scale is varied to allow space in more detailed areas, often city centers, where routes converge and stops are closer. Scale can be reduced where there is less to show, such as in the outer suburbs. Typically each route is shown by a different color. Often there is no other information, but major geographical features, such as the River Thames in London, may be included in simplified form.

Representations Using Area Symbols

Many maps represent qualitative data about areas, such as those about geology, soils, and vegetation. There is no generally accepted term for this form of cartographic representation. In some U.K. textbooks, the term 'chorochromatic' has been used and some North American texts use terms such as 'color-patch' or 'mosaic' maps, both of which adequately describe their appearance. Unfortunately some use the term 'choropleth map' (see below), but should only be applied to quantitative maps. At larger scales, the design of this type of mapping is often demanding due to the large number of classes and subclasses that can be present. Using many classes requires imaginative use of the graphic variables to ensure that a balanced representation is achieved where no class should appear more important than another. With a classification hierarchy, a major concern is to ensure that subclasses are clearly associated with others within the same overall category and have a greater difference to those in other categories. Often this is solved by using a main color (hue) to distinguish the top-level categories and use of patterns or texture to differentiate subclasses.

Probably one of the most widely used techniques for thematic mapping is the choropleth map, as shown on the right of Figure 2. This represents quantitative data for areas, usually some form of administrative unit, such as country, state, wards, census tract, or some other areal unit. They are popular partly because there are enormous amounts of numerical data available for areal units and, assuming the boundary data are available, they are easy to produce using desktop mapping systems and Geographical Information System (GIS). This ease of production has also led to great misuse, typically using data that are affected by the variation in zone sizes as discussed above in terms of representation, but also by poor selection of the area symbols. The typical approach is to represent the data values by light-to-dark series of ordered gray tones or varying the lightness of a single hue, such as shades of green from light to dark. With a single hue, the number of classes that the user can easily identify is generally four or five; if more classes are required, then a part spectral scheme can be adopted, perhaps going from yellow, through oranges, to reds. For showing data such as rates of change where there can be both positive and negative values, two contrasting hues can be used, such as red and blue, becoming darker as the value increases in each direction. Sometimes a neutral color can separate the two scales, representing no, or minimal, change. Such a scheme can also be applied to emphasize variation from the mean, such as divorce rates, where the regional average could be used as the dividing point. In recent times, a 'traffic light' color scheme has become popular, using green for good/above average, yellow for neutral, and red for poor/below average, particularly when applied to performance data, or business data such as sales figures. An excellent resource for investigating color choices for choropleth maps (and some other types) is the ColorBrewer Web site (Brewer and Harrower, 2009).

Apart from the choice of color scheme, the two key decisions are the number of classes to depict and the choice of class intervals. The fewer the classes the more generalized the representation; too many classes and the user starts to have some difficulty in distinguishing among them. It is also possible, but less common, to produce choropleth maps without classes (see Kennedy, 1994). Because different classifications of the data can produce radically different spatial patterns, the efficiency and effectiveness of a choropleth map depends heavily on its categories (Evans, 1997). Common approaches include equal intervals, where the data range is divided equally, and quantiles, where the same number of zones is included in each class. The choice would not be difficult if data were evenly or normally distributed, but often it is not, and highly skewed data sets present significant difficulties in selection of appropriate classes. Mechanical selection of classes will often result in rather meaningless class boundaries. A wise practice, often ignored, is the use of inherently meaningful category breaks, such as zero on a map of migration rates, or the national mean on a map of per capita income (Monmonier, 1993).

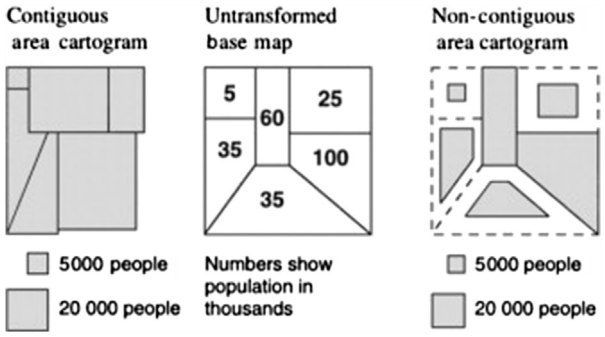

Figure 5 Area cartograms adjust the size of areal units according to a transforming variable such as population. Contiguous cartograms (left) may distort shape severely, whereas non-contiguous solutions (right) preserve shape but lose contiguity.

A major disadvantage of choropleth maps is that the boundaries often bear no relation to the actual spatial distribution of the phenomenon. Dasymetric maps appear graphically similar to choropleth maps, in that the data are represented in a similar manner, but in this case the boundaries are drawn to reflect uniform areas in the distribution of the specific phenomenon. Clearly, producing such boundaries requires additional information and effort compared to the choropleth map, which goes a long way to explain their limited use, especially when maps of many variables are to be produced.

On an 'area cartogram,' the size of the areal units being mapped varies according to population or some other transforming variable. As Figure 5 illustrates, noncontiguous cartograms sacrifice contiguity for shape whereas contiguous area cartograms may require severe distortions of shape in order to retain contiguity. Unless the reader has a good grasp of the original geography, the distorted shapes may make the size differences meaningless. This can effectively be countered by showing both a geographically correct map (typically choropleth) adjacent to the cartogram, so that an easy comparison can be made. Assuming that the reader appreciates the distortion in geography, cartograms make a dramatic demonstration of disparities in aspects such as population, wealth, education, or military spending. An isodemographic map – a population-based area cartogram – can also provide a useful base map for representation of other socioeconomic data, election outcomes, and public health (Dorling, 1994). Adjusting the base map for population in this way diminishes the exaggerated impact of large areas with small populations and ensures small, densely populated areas are visible. The availability of software to generate effective contiguous cartograms has significantly increased their use.

Representations of Continually Varying Surfaces

Sometimes referred to as contour maps, isarithmic maps refer to maps that represent continuously varying quantities over the surface by lines of equal values. Strictly, the term contour should be used only when the variable mapped is terrain height. Common terms are used for some variables, such as isobars for pressure. Terrain contour maps typically use an equal vertical interval between contour lines. In this case, the user can appreciate slopes by variations in the spacing of the contours. In many thematic mapping situations, nonequal intervals are used, so the nature of the 'surface' can be harder to visualize. This approach is often adopted because the data values are highly skewed or because the map maker wants to emphasize some particular values.

An extension of the isarithmic map is to apply a scale of colors to the zones between isarithms. The choice of colors is similar to that for choropleth maps, but more classes, with smaller color difference is successful here due to the nested nature of adjacent zones. GIS and surface modeling software make extensive use of a chromatic scale with large numbers of classes. Generally the data range is automatically divided by the number of steps in the scale, often resulting in unwieldy values in the legend, rather than selecting a simpler, more rounded interval.

Like the surface being represented, there is some value in adopting a representation method that also varies continuously. In terrain mapping, hill shading uses a continuously varying gray tone to represent relief. In its simplest form, the steeper the slope the darker the shading. To give a better impression of the three-dimensional nature of relief, such shading is often modified to have an assumed light source coming from the top left of the map, but this is probably not appropriate for other surfaces.

Other techniques of representing continuously varying surfaces depend on a nonorthogonal view of the mapped area. This includes fishnet plots and fully rendered block diagrams.

Conclusions

The power of the computer and the development of GIS and scientific visualization software have lead to a significant development of representation methods in recent years. Some of these possibilities have been known for a long time but have been too costly or impractical to carry out manually. Many of the newer methods rely heavily on interactivity and so are beyond the scope of this article. With modern systems, there is more scope to experiment with representation, as a change of color, or even representation method can happen in an instant with the click of a button. Despite these advances, the basic principles of thematic mapping, understanding data to be mapped, and proper use of the visual variables remain as important as ever.

See also: Cartographic Visualization; Cartography; Dynamic Mapping in Geography; Geographic Information Systems; Spatial Data; Tactile Maps in Geography.

Bibliography

Bertin, J., 1983. Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press, Madison, WI. Reprinted 2010 by ESRI Press, Redland, CA.

Brewer, C.A., 2005. Designing Better Maps: A Guide for GIS Users. ESRI Press, Redlands, CA.

Brewer, C.A., Harrower, M., 2009. ColorBrewer2. http://colorbrewer2.org/.

Chrisman, N., 2002. Exploring Geographic Information, second ed. Wiley, New York.
Dent, B.D., 1999. Cartography: Thematic Map Design, fifth ed. William C. Brown, Boston.
Dorling, D., 1994. Cartograms for visualising human geography. In: Hearnshaw, H.M., Unwin, D.J. (Eds.), Visualization in Geographical Information Systems. Wiley, Chichester.
Evans, I.S., 1997. The selection of class intervals. Transactions, Institute of British Geographers. New Series 2, 98–124.
Forrest, D., 1999. Geographic Information: its nature, classification and cartographic representation. Cartographica 36 (2), 31–53.
Goldberg, D.W., 2008. A Geocoding Best Practice Guide. North American Association of Central Cancer Registries, Inc.
Keates, J.S., 1996. Understanding Maps. Addison Wesley Longman Ltd, Harlow.
Kennedy, S., 1994. Unclassed choropleth maps revisited: some guidelines for the construction of unclassed and classed choropleth maps. Cartographica 31 (1), 16–25.
MacEachren, A.M., 1995. How Maps Work: Representation, Visualization, and Design. Guildford, New York.
Maling, D.H., 1992. Coordinate Systems and Map Projections, second ed. Pergamon, Oxford.
Monmonier, M., 1993. Mapping It Out: Expository Cartography for the Humanities and Social Sciences. University of Chicago Press, Chicago.
Robinson, A.H., 1982. Early Thematic Mapping in the History of Cartography. University of Chicago Press, Chicago.
Slocum, T.A., McMaster, R.B., Kessler, F.C., Howard, H.H., 2009. Thematic Cartography and Geovisualization, third ed. Pearson Prentice Hall, Upper Saddle River, NJ.

Theory of Mind During Infancy and Early Childhood Across Cultures, Development of

Beate Sodian and Susanne Kristen, Ludwig-Maximilians-University Munich, Munich, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article presents a summary of Theory of Mind (ToM) development across cultures. It presents cross-linguistic and cross-cultural research on implicit and explicit ToM skills, ranging from ToM in infancy through children's emerging mental state talk to children's explicit comprehension that others might hold false beliefs and that their actions are guided by them. It is argued that more research is needed, especially research contrasting the multiple facets (parental practices, beliefs, language) through which culture might contribute to ToM development. Finally, cultural effects on the relations between ToM and its developmental factors are discussed.

Theory of Mind

A Theory of Mind (ToM) is the ability to attribute mental states to oneself and others and to use these attributions for the prediction and explanation of action. ToM is sometimes called a belief–desire psychology, since the attribution of beliefs and desires is central to our commonsense mentalism. Children acquire a ToM in two steps: By the age of 2.5 to 3 years, they understand the subjectivity of desires and predict an agent's action based on their desires, and by the age of about 4 years, they understand that beliefs may differ from reality and that agents' beliefs, rather than the state of reality, will determine their actions (Perner, 1991). For example, when shown a typical container for sweets (a Smarties tube), and informed about its rather unusual content (a pencil), 4-year-olds will predict that their friend, who has not been able to look into the container, will believe that there are Smarties in the container, whereas 3-year-olds will typically answer that an uninformed other will say that there is a pencil in the container. The concept of belief is crucial to an understanding of the mind, since it entails the distinction between reality and a mental representation of reality. Since the first systematic investigation by Wimmer and Perner (1983), a large body of evidence on the development of false-belief understanding in preschoolers has been gathered, mostly using story format tasks such as the location-change task in which a story figure (Maxi) places an object (chocolate) in location A, then leaves; during his absence another story figure (Mother) transfers the object from A to B; then Maxi returns and looks for the chocolate; the child is asked the test question: "Where will Maxi look for the chocolate?" or "Where does Maxi think the chocolate is?" In a meta-analysis of more than 500 studies of false-belief understanding by Wellman et al. (2001), a robust developmental trend was observed, across tasks and situational factors, between 3 and 5 years of age; under some task conditions, improved performance was observed in 3-year-olds, but a marked age-trend persisted. Children with autism show a severe deficit in false-belief tasks and acquire a ToM only with a gross delay and persisting deficiencies even in adulthood (Hill and Frith, 2003). Taken together, these findings indicate a conceptual deficit in typically developing 3-year-olds and in children with autism. Conceptual coherence in ToM development is indicated by close interrelations between a whole set of conceptual distinctions (e.g., appearance and reality, different visual perspectives on the same entity) and causal relations (e.g., seeing-knowing) that are acquired around the same age as the concept of belief and by the findings of training interventions, indicating transfer effects (Slaughter and Gopnik, 1996). ToM development in preschool age is closely related to language acquisition and to the development of executive functions. However, 3-year-olds' difficulty with ToM tasks cannot be accounted for by the linguistic and executive demands of the tasks (Perner et al., 2002).

Implicit ToM in Infancy

An implicit understanding of false belief precedes a full or explicit one. In early studies (Clements and Perner, 1994), young 3-year-olds were shown to take an agent's false belief into account in their anticipatory looking, while they failed when asked for an explicit judgment of the agent's belief or a prediction of his action. More recently, rich evidence for an implicit understanding of belief in infants as young as 13–18 months has been gathered with looking-time and anticipatory-looking methods, but also in interactive tasks requiring inferences about an agent's goals based on his or her beliefs (see Baillargeon et al. (2010) for a review). For example, Onishi and Baillargeon (2005) showed in a habituation task that 15-month-olds expected an agent who had or had not witnessed an unexpected transfer of an object from one location to another to reach into the location where the agent truly or falsely believed the object to be hidden, that is, the infants looked reliably longer when the agent reached to the other location. Longitudinal evidence indicates continuity between an implicit understanding of belief at 18 months of age and an explicit one at 48 months of age on a task-specific level (Thoermer et al., 2012). There is an ongoing debate about whether infants' success on implicit false-belief tasks is evidence for an implicit ToM or whether the findings can be accounted for by lower-level mechanisms such as statistical

learning (Ruffman et al., 2012) and the application of behavioral rules (see Sodian (2011) for a review).

The findings on infants' implicit understanding of false belief are part of a larger body of recent research on psychological reasoning in infancy. Two distinct conceptual systems have been claimed to support infants' reasoning about intentional action: one allowing the attribution of motivational states such as goals and dispositions and permitting the correct prediction of an agent's action when the agent's informational state corresponds to reality, and other allowing action prediction when the agent's informational state is incongruent with reality, as in the case of the agent holding a false belief. There is ample evidence for infants' propensity to conceive of human action in terms of action goals, even in the first year of life (see Woodward (2009) for a review): In Woodward's paradigm, 5-month-old infants were habituated to a hand repeatedly reaching toward and grasping one of two objects placed on the opposite ends of a small stage; in the test phase, the positions of the objects were reversed, and the hand now reached either for the previously grasped object on a new path or for the old object on the same path; infants showed stronger dishabituation to the new-object event than to the new-path event, indicating that they had represented the action in terms of its goal, not in terms of spatiotemporal properties. Further, infants as young as 6 months are able to keep track of what an agent can and cannot see when representing the agent's goal (Luo and Johnson, 2009). The ability to infer an agent's goals when observing a failed attempt and to analyze the goals of incomplete actions, as well as infants' understanding of action roles in social interaction and their sensitivity to others' motives provide evidence for the attribution of intentions to agents (see Poulin-Dubois et al. (2009a) for a review).

Joint attention, emerging toward the end of the first year of life, has been interpreted as an implicit ToM, since it involves the sharing of an experience with another person and therefore a sharing of perspectives. There is increasing evidence for this view from studies showing that 12- to 18-month-old infants use an adult's gaze direction to obtain information about objects outside their visual field, that they use the pointing gesture specifically to convey information, and that they represent what others have previously experienced (see Moll and Meltzoff (2011) for a review).

Infants' ability to communicate preverbally with others and to attribute goals and intentions to agents is a specific developmental precursor of later full-fledged mental state attribution: Several independent studies indicate significant predictive relations of social information-processing skills toward the end of the first year of life and ToM competencies in preschool age, independently of language abilities and IQ (e.g., Wellman et al., 2008). Similarly, joint attentional skills around the first birthday are predictively related to later explicit talk about the mind (Kristen et al., 2011). In sum, the longitudinal evidence, albeit still scarce, indicates conceptual continuity from an implicit to an explicit ToM in early childhood development.

Cultural Variation

A ToM is fundamental for human cognition. Therefore, it appears plausible that it should be universal. However, explicit ToM reasoning develops only in the preschool years and is closely related to language acquisition. Social influences, such as individual variation in maternal interaction styles, have been shown to be related to children's ToM development (Meins et al., 2002). Thus, there is reason to expect cultural variation in the timing of the acquisition of mental-state concepts. Still, if this variation is not very large, if false-belief understanding is acquired by all typically developing children of a culture within the preschool years, and if the sequence of acquisition of mental-state concepts is the same across cultures, we would take this as evidence for the universality assumption. Cross-cultural ToM research has addressed children's performance in ToM tasks across cultures, as well as their spontaneous construal of the mental domain in their developing mental state language. In contrast, to date very little research has addressed cultural variation in preverbal infants' psychological understanding.

Infancy

Recently, the first systematic study on cultural variation in early social-cognitive skills was carried out by Callaghan et al. (2011). In a series of eight studies, they assessed the social-cognitive skills of 1- to 3-year-old children in rural Canada, Peru, and India. Children's performance on basic joint-attention and intention-reading tasks such as imitation, helping, gaze following, and communicative pointing was similar across the three cultural settings. Similarly, children displayed much the same competencies in interactive episodes of collaboration and joint attention. Differences emerged only in symbolic skills, with pretense and pictorial skills emerging much earlier in the Canadian children than in the other two cultural groups. The authors conclude from these findings that children get enough social experience in all cultural settings to develop the most basic social-cognitive skills, but that the acquisition of more specific symbolic skills may be dependent on specific learning experiences that Western cultures amply provide.

Mental State Language

Toddlers' emerging spontaneous talk about the mind is an important source of evidence for their nascent *explicit* understanding of mental states. Based on an extensive literature search of mental state words and from natural speech samples, Bretherton and Beeghly (1982) created a 78-item parental checklist of mental state words. Overall, 28-month-olds produced terms from six categories. Thereby, their volition/ability vocabulary was most varied, followed by their physiological vocabulary and perceptual vocabulary. Children's emotion and affect and moral judgment/obligation vocabulary were more varied than children's cognition vocabulary, which significantly lagged behind all other categories. A longitudinal study examining mental state language in the home environment (Bartsch and Wellman, 1995) showed that children begin to use desire terms at around 18–24 months of age. Genuine references to cognitive states (e.g., "I know that the cat is in the garage") develop at around 3 years of age (e.g., Taumopeau and Ruffman, 2006, 2008). Concerning their relative frequency, beliefs and thoughts only begin to match desires at around 5 years of age. Across ethnic groups, the developmental pattern

applying to Caucasian-American children was found to be identical in African-American preschoolers (Curenton and Justice, 2004). The result that children acquire desire terms before they acquire belief terms can be generalized not only to different ethnic groups living in the same culture but also to other languages spoken within Western cultures such as German, Spanish, and French (Kristen et al., 2012; Pascual et al., 2008; Poulin-Dubois et al., 2009b).

Moving outside the boundaries of Western cultures, studies in Chinese children (Mandarin- and Cantonese-speaking) (Tardif and Wellman, 2000) demonstrated variation in the timing of the beginning and the end points of this developmental sequence as a function of linguistic as well as cultural factors. Carefully choosing languages (Mandarin and Cantonese) which produce a great number of contrasts when compared with English, a study by Tardif and Wellman (2000) assessed the development of mental state language in 10 Mandarin-speaking (21–27 months of age) and 8 Cantonese-speaking (8–44 months of age) children. The results suggest a pattern of development which is similar to the one observed in English, with early use of desire terms being followed by other mental state references. However, results also revealed cross-cultural variations. First, Chinese-speaking children were found to use desire terms much earlier than Anglo-American children. These findings were corroborated by studies showing that in Mandarin the term 'want' is among the first words that children use (Tardif et al., 1999). The authors conclude that Chinese children's overall attention to verbs and their early verb learning might account for their accelerated reference to desires as compared with English-speaking children (Tardif, 1996). Further, the term 'know' was referred to relatively early in Mandarin- and Cantonese-speaking children (before their second birthday), whereas Bartsch and Wellman (1995) found that American-English children acquired 'knowing that' closer to their third birthday. The authors discuss that the lack of a finite distinction and the lack of a complementizer after the Mandarine and Catonese verbs for 'know' make it easier for Mandarin- and Cantonese-speaking children to understand and express notions of 'knowing that' than for English-speaking children. Finally, the authors concluded that linguistic as well as cultural differences result in even later reference to 'thinking' in Chinese-speaking children than in English-speaking children. In contrast to English, containing mainly neutral terms to talk about beliefs, Chinese languages contain neutral as well as nonneutral belief terms (e.g., *yi3wei2* in Mandarin and *ji5wai4* in Cantonese) specifically marking someone's belief as false (Tardif et al., 2004; cited in Liu et al., 2008). In regard to possible cultural explanations for these differences, folk psychologies, societal expectations, and parental practices differ significantly between Eastern and Western cultures all of which could potentially affect children's developing talk of mental states. In specific, Chinese parents' disinclination to talk about 'thinking' with young children might account for them referring to 'thinking' later than English-speaking children (Liu et al., 2008).

In turn, internal state language might give insight into folk psychology in different cultures. In some languages, the contextual use of mental state terms might reflect cultural beliefs. For instance, the checklist developed by Bretherton and Beeghly (1982) has recently been translated into Dakota. The Dakota language represents one of the major three dialects (the other two being Lakota and Nakota) spoken by the Sioux, a Native American people in North America and is known to embody a value system about how to live and how to relate to each other, as well as describing the Dakota people's connection to nature, animals, the Great Spirit, and the order of things (Knudson et al., 2010). In the Dakota psychology, all creations (humans, animals, and the universe), while belonging to different building blocks of creation, are mentally connected. These relations are also reflected in small subtleties in the contextual use of mental state terms. For instance, while in English and other Indo-European languages the moral term 'mean' is mainly used in reference to people, in the Dakota language the term refers to both people and animals. *Wicaste kin canze* means the 'man is mean,' while *Sunka wan canze* means 'a dog is mean' (Canku and Kristen, 2013).

Mental State Reasoning in ToM Tasks

Two main issues have been addressed in cross-cultural ToM research: the universality of the acquisition of false-belief understanding and the universality of the order in which mental state concepts are acquired, with desire reasoning preceding belief reasoning.

Cross-cultural comparisons have been limited by small sample sizes and sample specificities. Therefore, Liu et al. (2008) conducted a meta-analysis of studies contrasting Chinese (i.e., Mandarin- and Cantonese-speaking) and North American samples. The meta-analysis consisted of 196 Chinese conditions (from Mainland China and from Hong Kong), representing responses from more than 3000 children, and 155 North American conditions. The findings showed parallel developmental trajectories of false-belief understanding for children in China and North America coupled with significant differences in the timing of development, with performances varying across different locales by as much as 2 or more years. More specifically, Liu and colleagues report that Canadian children show above-chance performance on false-belief tasks around the age of 38 months, whereas Hong Kong children do so around the age of 64 months. The authors conclude that these data show the importance of both universal trajectories and specific experiential factors in the development of ToM.

A similar variation in the timing of false-belief acquisition was found in non-Western cultures other than China. An early study by Avis and Harris (1991) in rural Cameroon yielded findings remarkably similar to those reported for European and North American samples. In a carefully adapted task, children saw an adolescent member of the community cooking kernels in a hut. After that, the adolescent placed them in a bowl and then left the room. In his absence, an adult entered the hut and conspired with a child to trick the adolescent by taking the kernels out of the bowl and putting them into a closed pot. The child was then asked where the adolescent would look for the kernels when he would come back. The authors found that preliterate Baka children in southeast Cameroon passed at an age comparable to that seen in European and North American studies. Consistent with the original experiment, 5- to 6-year-olds answered

correctly that the adolescent would mistakenly look for the kernels in the bowl. Consistently, Oberle (2009), who did research on the Micronesian islands of Yap and Fais and used a classical surprise-content task with culturally adjusted research materials, found that 3-year-olds did not display false-belief understanding, while 5-year-olds did.

In contrast, Vinden (1996) conducted a study of false-belief understanding and understanding of representational change with carefully adapted materials in 4- to 8-year-old Junin Quechua children in the Peruvian highlands. Performance on these tasks was poor even in 8-year-olds. One reason for this deviant developmental pattern may be the relative absence of explicit mental state vocabulary among the Junin Quechua. Surprisingly, some studies have also reported a marked delay in false-belief understanding in Japanese children. Naito and Koyama (2006) found that Japanese children understand false belief only at the age of about 6–7 years. This delay may be due to cultural attribution styles: children in some Asian cultures may be led to attribute behaviors (especially mistaken action based on a false belief) to social roles rather than to internal causes. Note, however, that the study by Naito and Koyama (2006) used careful translations but did not adapt the material to the cultural context. In any case, there is no generalized delay in Japanese children's grasp of epistemic cues: Matsui et al. (2009) found that Japanese 3-year-olds benefitted from the explicit statement of a speaker's false belief in a way that German 3-year-olds did not, when it was marked with a Japanese grammaticalized particle of speaker certainty. Nor is there an across-the-board delay in Asian children's performance on standard false-belief tasks: When comparing preschoolers from Korea with preschoolers from the United States on a set of false-belief problems (inanimate object, voluntary person movement, involuntary person movement), children from the Korean culture, which fosters a relatively interdependent conception of the self, were shown to outperform American children (Ahn and Miller, 2012). Overall, variations in children's relatively interdependent self-responses were shown to be predictive of good false-belief performance.

The most serious deviations from developmental synchrony as postulated by the universality assumption were found in studies of ToM development in Samoan children. Callaghan et al. (2005) investigated mental state reasoning in Canada, Peru, Thailand, India, and Samoa with a false-belief paradigm, which was held constant across cultures involving a change of location. In all the five countries, the majority of 3-year-olds significantly failed. In the group of 4-year-olds, the difference between those who passed and those who failed did not reach significance in all of the countries, with the exception of Samoa, where 18 out of 25 children still failed ($p < .05$). While the majority of the 5-year-old children clearly passed the task in Canada and Peru ($p < .001$), as well as in Thailand and India ($p < .01$), only 13 out of 18 children in Samoa passed the task ($p < .10$). To further explore this difference, Mayer and Träuble (2012) provided the first continuous survey of false-belief understanding beyond 5 years of age and tested more than 300 Samoan children using a change-of-location task. Samoan children improved gradually but rather slowly, with no succeeding majority before 8 years of age. One-third of the 10- to 13-year-olds still failed. Using a different translation among 55 children from 4 to 8 years of age resulted in the same results. The authors interpret their findings as supporting the cultural variability of ToM development: Samoans believe that it is impossible or very difficult to know what others think or feel (Robbins and Rumsey, 2008; cited in Mayer and Träuble, 2012). Due to this opacity-of-the-mind doctrine, mental states constitute no suitable object for conjecture in the Samoan culture and thus children may lack the learning experiences necessary to construct an explicit understanding of other minds.

To study cultural variation in the *sequence* of mental state comprehension, several studies used the ToM scale developed by Wellman and Liu (2004). This scale has previously been well validated and shown, via both Guttman and Rasch analyses, to identify a reliable developmental sequence of mental state concepts in which progress to any later scale step was contingent on mastery of all earlier ones. Preschoolers in the United States, Australia, and Germany mastered the steps in the following order: diverse desires, then diverse beliefs, then knowledge access, then false beliefs, and finally hidden emotions (Kristen et al., 2006; Peterson and Wellman, 2009; Peterson et al., 2005). When comparing Chinese-speaking preschoolers in Beijing and Iranian preschoolers with English-speaking preschoolers in the United States, the authors observed cross-cultural differences (Shahaeian et al., 2011; Wellman et al., 2006, 2011). Beijing and Iranian preschoolers reversed the diverse beliefs and knowledge access steps to reveal a scale sequence placing expertise with knowledge access (people who witness an event know about it) reliably ahead of diverse beliefs (different people have different ideas and opinions about the same thing). The authors concluded that these differences might be due to the emphasis on mutual respect and avoidance of dispute in collectivist cultures. Importantly, across all studies employing the ToM scale, the diverse-desire task was mastered before children mastered the false-belief task. Thus, the desire–belief sequence seems to follow a universal pattern.

Summing up, consistent with the findings of studies in early *mental state talk*, studies of ToM reasoning in non-English-speaking children indicate that the overall developmental sequence of *mental state comprehension*, with desire understanding preceding understanding of beliefs, is consistent across cultures. However, rather specific cross-cultural variation seems to exist in the onset of mental state comprehension. These temporal variations have been interpreted as an effect of cross-cultural input differences (e.g., emphasis of filial respect in a given culture).

Cross-cultural ToM research is, however, still limited by design problems. It seems warranted to conduct cross-cultural studies that allow us to investigate the effects of language, family input, and cultural beliefs in the same sample. While most studies interpret their findings in these terms, few studies have actually compared parental input and cultural representations of a ToM across Western and non-Western cultures, and related them systematically to interindividual differences in children's mental state production and comprehension across a broad range of categories within the same sample.

Furthermore, cultural variation in the relationship between ToM and its most important developmental correlates (e.g., inhibitory control) has rarely been studied. Existing research draws a differentiated picture. Chasiotis et al. (2006) found that conflict inhibition, but not delay inhibition, predicts false-belief understanding across three cultural contexts: Germany, Costa Rica, and Cameroon. However, another study found that Korean children display ceiling effects in inhibitory-control tasks before they pass false-belief tasks and that the relationship between inhibitory control and false-belief understanding is weaker in Korea than in Britain (Oh and Lewis, 2008). Thus, more research is needed to fully understand the effects of culture on the complex developmental mechanisms leading to the emergence of an explicit ToM in preschool age and to inform theories on the precursors of a ToM.

Acknowledgments

The writing of this article was supported by a grant from the German Research Council to Beate Sodian (so 213/27-2).

See also: Cognitive Development During Infancy and Early Childhood across Cultures; First Language Acquisition, Developmental Psychology of; Infancy and Human Development; Intentionality During Infancy and Early Childhood, Development of; Joint Attention During Infancy and Early Childhood Across Cultures; Language Development, Theories of; Social Referencing During Infancy and Early Childhood across Cultures.

Bibliography

Ahn, S., Miller, S.A., 2012. Theory of mind and self-concept: a comparison of Korean and American children. Journal of Cross-cultural Psychology 43 (5), 671–686.

Avis, J., Harris, P.L., 1991. Belief-desire reasoning among Baka children: evidence for a universal conception of mind. Child Development 62 (3), 460–467.

Baillargeon, R., Scott, R.M., He, Z., 2010. False-belief understanding in infants. Trends in Cognitive Sciences 14 (3), 110–118.

Bartsch, K., Wellman, H.M., 1995. Children Talk about the Mind. Oxford University Press, New York.

Bretherton, I., Beeghly, M., 1982. Talking about internal states: the acquisition of an explicit theory of mind. Developmental Psychology 18 (6), 906–921.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., et al., 2011. Early social cognition in three cultural contexts. Monographs of the Society for Research in Child Development 76 (2), 1–142.

Callaghan, T., Rochat, P., Lillard, A., Claux, M.L., Odden, H., Itakura, S., Singh, S., 2005. Synchrony in the onset of mental-state reasoning evidence from five cultures. Psychological Science 16 (5), 378–384.

Canku, C., Kristen, S., 2013. The ISLQ in the Dakota Language. Unpublished Manuscript.

Chasiotis, A., Kiessling, F., Hofer, J., Campos, D., 2006. Theory of mind and inhibitory control in three cultures: conflict inhibition predicts false belief understanding in Germany, Costa Rica and Cameroon. International Journal of Behavioral Development 30 (3), 249–260.

Clements, W.A., Perner, J., 1994. Implicit understanding of belief. Cognitive Development 9 (4), 377–395.

Curenton, S.M., Justice, L.M., 2004. African American and Caucasian preschoolers' use of decontextualized language: literate language features in oral narratives. Language, Speech, and Hearing Services in Schools 35 (3), 240–253.

Hill, E.L., Frith, U., 2003. Understanding autism: insights from mind and brain. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358 (1430), 281–289.

Knudson, N., Snow, J., Canku, C., 2010. Beginning Dakota/Tokaheya Dakota Iyapi Kin: 24 Language and Grammar Lessons with Glossaries. Minnesota Historical Society.

Kristen, S., Sodian, B., Thoermer, C., Perst, H., 2011. Infants' joint attention skills predict toddlers' emerging mental state language. Developmental Psychology 47 (5), 1207–1219.

Kristen, S., Sodian, B., Licata, M., Thoermer, C., Poulin-Dubois, D., 2012. The development of internal state language during the third year of life: a longitudinal parent report study. Infant and Child Development 21 (6), 634–645.

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., Sodian, B., 2006. Skalierung von "Theory of Mind"-Aufgaben. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 38 (4), 186–195.

Liu, D., Wellman, H.M., Tardif, T., Sabbagh, M.A., 2008. Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. Developmental Psychology 44 (2), 523–531.

Luo, Y., Johnson, S.C., 2009. Recognizing the role of perception in action at 6 months. Developmental Science 12 (1), 142–149.

Matsui, T., Rakoczy, H., Miura, Y., Tomasello, M., 2009. Understanding of speaker certainty and false-belief reasoning: a comparison of Japanese and German preschoolers. Developmental Science 12 (4), 602–613.

Mayer, A., Träuble, B., 2012. Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. International Journal of Behavioral Development 37, 21–28.

Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., Tuckey, M., 2002. Maternal mind–mindedness and attachment security as predictors of theory of mind understanding. Child Development 73 (6), 1715–1726.

Moll, H., Meltzoff, A., 2011. Perspective-taking and its foundation in joint attention. In: Roessler, J. (Ed.), Perception, Causation, and Objectivity. Issues in Philosophy and Psychology. Oxford University Press, Oxford, pp. 286–304.

Naito, M., Koyama, K., 2006. The development of false-belief understanding in Japanese children: delay and difference? International Journal of Behavioral Development 30 (4), 290–304.

Oberle, E., 2009. The development of theory of mind reasoning in Micronesian children. Journal of Cognition and Culture 9, 39–56.

Oh, S., Lewis, C., 2008. Korean preschoolers' advanced inhibitory control and its relation to other executive skills and mental state understanding. Child Development 79 (1), 80–99.

Onishi, K.H., Baillargeon, R., 2005. Do 15-month-old infants understand false beliefs? Science 308 (5719), 255–258.

Pascual, B., Aguado, G., Sotillo, M., Masdeu, J.C., 2008. Acquisition of mental state language in Spanish children: a longitudinal study of the relationship between the production of mental verbs and linguistic development. Developmental Science 11 (4), 454–466.

Perner, J., 1991. Understanding the Representational Mind. The MIT Press, Cambridge.

Perner, J., Lang, B., Kloo, D., 2002. Theory of mind and self-control: more than a common problem of inhibition. Child Development 73 (3), 752–767.

Peterson, C.C., Wellman, H.M., 2009. From fancy to reason: scaling deaf and hearing children's understanding of theory of mind and pretence. British Journal of Developmental Psychology 27 (2), 297–310.

Peterson, C.C., Wellman, H.M., Liu, D., 2005. Steps in theory-of-mind development for children with deafness or autism. Child Development 76 (2), 502–517.

Poulin-Dubois, D., Brooker, I., Chow, V., 2009a. The developmental origins of naïve psychology in infancy. Advances in Child Development and Behavior 37, 55–104.

Poulin-Dubois, D., Chiarella, S.S., Polonia, A., 2009b. Toddlers' vocabulary about the mind: a cross-linguistic study using the Internal State Language Questionnaire. Rivista Di Psicolinguistica Applicata 3, 35–49.

Robbins, J., Rumsey, A., 2008. Introduction: cultural and linguistic anthropology and the opacity of other minds. Anthropological Quarterly 81 (2), 407–420.

Ruffman, T., Taumoepeau, M., Perkins, C., 2012. Statistical learning as a basis for social understanding in children. British Journal of Developmental Psychology 30 (1), 87–104.

Shahaeian, A., Peterson, C.C., Slaughter, V., Wellman, H.M., 2011. Culture and the sequence of steps in theory of mind development. Developmental Psychology 47 (5), 1239–1247.

Slaughter, V., Gopnik, A., 1996. Conceptual coherence in the child's theory of mind: training children to understand belief. Child Development 67 (6), 2967–2988.

Sodian, B., 2011. Theory of mind in infancy. Child Development Perspectives 5 (1), 39–43.

Tardif, T., 1996. Nouns are not always learned before verbs: evidence from Mandarin speakers' early vocabularies. Developmental Psychology 32 (3), 492–504.

Tardif, T., Gelman, S.A., Xu, F., 1999. Putting the "noun bias" in context: a comparison of English and Mandarin. Child Development 70 (3), 620–635.

Tardif, T., Wellman, H.M., 2000. Acquisition of mental state language in Mandarin- and Cantonese-speaking children. Developmental Psychology 36 (1), 25–43.

Tardif, T., Wellman, H.M., Cheung, K.M., 2004. False belief understanding in Cantonese-speaking children. Journal of Child Language 31 (4), 779–800.

Taumoepeau, M., Ruffman, T., 2006. Mother and infant talk about mental states relates to desire language and emotion understanding. Child Development 77 (2), 465–481.

Taumoepeau, M., Ruffman, T., 2008. Stepping stones to others' minds: maternal talk relates to child mental state language and emotion understanding at 15, 24, and 33 months. Child Development 79 (2), 284–302.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., Kristen, S., 2012. Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. British Journal of Developmental Psychology 30 (1), 172–187.

Vinden, P.G., 1996. Junin Quechua children's understanding of mind. Child Development 67 (4), 1707–1716.

Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child Development 72 (3), 655–684.

Wellman, H.M., Fang, F., Liu, D., Zhu, L., Liu, G., 2006. Scaling of theory-of-mind understandings in Chinese children. Psychological Science 17 (12), 1075–1081.

Wellman, H.M., Fang, F., Peterson, C.C., 2011. Sequential progressions in a theory-of-mind scale: longitudinal perspectives. Child Development 82 (3), 780–792.

Wellman, H.M., Liu, D., 2004. Scaling of theory-of-mind tasks. Child Development 75 (2), 523–541.

Wellman, H.M., Lopez-Duran, S., LaBounty, J., Hamilton, B., 2008. Infant attention to intentional action predicts preschool theory of mind. Developmental Psychology 44 (2), 618–623.

Wimmer, H., Perner, J., 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13 (1), 103–128.

Woodward, A.L., 2009. Infants' grasp of others' intentions. Current Directions in Psychological Science 18 (1), 53–57.

Therapeutic Community

Rex Haigh, Berkshire Healthcare NHS Foundation Trust, UK; Royal College of Psychiatrists, UK; and Institute of Mental Health, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Therapeutic Communities are both a radical group-based therapeutic method, and a complex therapeutic philosophy. This article traces their sociopolitical history and describes the numerous influences which have been influential in the field's development. It distinguishes between the 'British Model' (mostly used in mental health settings) and the 'American Model' (mostly for addictions) and describes how they now share much of the same territory. The phases of treatment are explained, as are the indications and contraindications for their use. Evidence of effectiveness, together with suitable sources of further information, are given.

Defining Beliefs

Certain beliefs about human relationships and the nature of therapy are central to therapeutic communities.

- Staff are not completely 'well' and residents are not completely 'sick.' There is a basic equality as human beings between staff and residents, who share many of the same psychological processes and experiences.
- Whatever the symptoms or behavior problems, the individual's difficulties are primarily in his or her relationships with other people.
- Therapy is essentially a learning process, both in the sense of learning new skills – how to relate to others or deal more appropriately with distress – and learning to understand oneself and others.

The British Model

In the United Kingdom the term 'therapeutic community' is most commonly ascribed to Tom Main (1946), who used it in his description of the psychiatric hospital systems which emerged during the second World War for treating battle-shocked soldiers. David Clark (1964, 1965) later refined the term to distinguish the general 'therapeutic community approach,' which was then usually applied to the whole hospital, and 'the therapeutic community proper' – units which see themselves with a distinct identity and which have clear membership. The modern day equivalent term for the 'therapeutic community approach' is 'therapeutic environments,' which is becoming recognized – at least in the United Kingdom – as something that has been lost in various settings, such as hospital wards, day services, and community facilities. The therapeutic community proper is now more commonly called the 'democratic therapeutic community' in Britain, often abbreviated to 'DTC.'

Roberts defines a therapeutic community as

> a consciously designed social environment and programme within a residential or day unit in which the social and group process is harnessed with therapeutic intent. In the therapeutic community the community is the primary therapeutic instrument.
>
> Roberts, 1977: 4

These therapeutic communities are sometimes known as the 'Maxwell Jones Model,' 'mental health,' 'psychiatric,' 'flattened hierarchy' as well as 'democratic' type.

History

Despite the recent identification of therapeutic communities as a specific entity, their origins go back much earlier. 'Mentally afflicted pilgrims' attended the shrine of St Dymphna at Geel in Flanders in the mid-thirteenth century, and some commentators identified this as the first recognizable format reflecting the principles and practices of therapeutic communities (Parry-Jones, 1981). The pilgrims were taken in by the residents of the village, and cared for communally; their outcome criteria was the more weight pilgrims gained, the better the result! The next documented and identifiable strand of therapeutic community ideas emerged in 'moral treatment' in the eighteenth century. William Tuke founded the Retreat hospital in York in 1796, based on the development of his ideas on moral treatment. Tuke's particular approach was to treat the insane as closely as possible to how one would treat 'normal' people. Moral treatment is also particularly identified with Pinel, who referred to 'treatment through the emotions,' and practiced at the Salpêtrière in Paris in the early nineteenth century. These ideas crossed the Atlantic, and in 1817 a hospital was founded in Pennsylvania modeled on the Retreat by the Pennsylvanian Quakers (Tuke, 1813; Rothman, 1971; Scull, 1979).

Another therapeutic environment tradition which evolved independently is the field of therapeutic education for children. Bridgeland, in 1971, described residential therapeutic education as "a first attempt at combining psychotherapeutic ideas with participatory democracy" (Bridgeland, 1971). Homer Lane set up the 'Boys Republic' between 1907 and 1912 in Chicago for deprived urban children. He imported the idea into the United Kingdom in 1913 and set up the 'Little Commonwealth' in Dorset. This was also for disturbed youngsters (Bazeley, 1928). In 1924, A.S. Neill set up Summerhill School which continues to this day as an institution specializing in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.27074-8

progressive education (Neill, 1937, 1967). These ideas were taken up by others, notably George Lyward with Finchden Manor (Burn, 1956), and David Wills and Marjorie Franklin with the Hawkspur and Q camp experiments in the 1930s (Wills, 1969; Franklin, 1966). Melvyn Rose adapted the ideas in the 1970s for the 'approved school' model for very disturbed children at Peper Harow (Rose, 1990). He ran this strictly as a therapeutic community, and the emerging ideas were also taken up by other approved schools, borstals, and probation hostels. A descendant organization from the Peper Harow Foundation, 'Childhood First,' continues to run residential therapeutic communities for children. Charterhouse Group is a similar organization (www.charterhousegroup.org.uk).

As well as therapeutic communities for disturbed children in the education field, there are also a number of communities for people with learning disabilities. The best known of these are the Camphill communities founded by König – which are based on the ideas of Steiner and his 'anthroposophy,' from which Montessori schools also owe an acknowledgment of their own theoretical base (Brendtro and Ness, 1983; Bloor et al., 1988). A similar, Catholic-based movement called L'Arche was founded by Jean Vanier in 1964 (Vanier, 1982). These two groups are closer in intention to the thirteenth-century caring communities in Geel, than to hospital or community-based treatment programs.

Another group of therapeutic communities was founded in the 1960s based on the radical ideas of the antipsychiatrists, particularly Ronnie (R.D.) Laing and David Cooper. The best known of these are the Philadelphia Association and Arbours Association in London, and Soteria House in the United States (Berke, 1982; Gunderson et al., 1983). Kingsley Hall, where the celebrated patient Mary Barnes was treated, was led by R.D. Laing between 1965 and 1970 (Cooper, 1967; Laing, 1971; Laing et al., 1965; Berke, 1980; Barnes and Berke, 1971). Both continue to this day: the Philadelphia Association as a housing organization; Arbours as a reputable psychoanalytic psychotherapy training institution.

However, most of the early therapeutic communities were found in the state psychiatric hospitals. The origin of democratic therapeutic communities is normally traced back to three experiments during the Second World War: two at Northfield Military Hospital in Birmingham and one at Mill Hill Hospital in North London (Harrison, 2000; Jones, 1946, 1968, 1979a). At Northfield, Bion's experiment with Rickman (Bion and Rickman, 1943; Bion, 1961) was stopped after 6 weeks when his work was more challenging than the military authorities would permit, as Tom Main later described

> neither the commanding officer nor his staff was able to tolerate the early weeks of chaos, and both were condemning and rancorous about Bion's refusal to own total responsibility for the disorder of others.
>
> Main, 1977

Bion was replaced by Main, Foulkes, and Bridger. This 'Second Northfield Experiment' continued with a less provocative and more integrated therapeutic community program until the end of the war (Bridger, 1946, 1985). Tom Main went on to be director of the Cassel Hospital at Richmond (Ham) in London, which for many years was a renowned intensive inpatient psychotherapy unit and only in the last decade has identified itself as a type of therapeutic community. S.H. Foulkes was another army officer running the second Northfield experiment, and he went on to found group analytic psychotherapy, culminating in the foundation of the Institute of Group Analysis in 1972 (Foulkes, 1948, 1964; Main, 1977). Harold Bridger died in 2005 after a long lifetime's work in organizational consultancy with the Tavistock Institute of Human Relations.

In 1941 at Mill Hill Hospital, Maxwell Jones (then an award-winning respiratory physiologist) saw that soldiers suffering from what was called 'effort syndrome' were more helpful to each other than the staff, by sharing the emotional meaning of their experiences in discussion groups (Jones and Lewis, 1941). After developing a successful short-term program based on this principle during the war, for a 100 or so men at a time, Jones later went on to set up the 'Industrial Neurosis Unit' in South London to get 'vagabonds and scroungers' back to work. This became the Social Rehabilitation Unit and later the Henderson Hospital, which was the flagship 'British Model' democratic therapeutic community proper from the 1950s (Jones, 1946, 1953, 1979b; Whiteley, 1980) until its closure in 2008.

Although Tom Main coined the term Therapeutic Community, it was Maxwell Jones whose name became most closely identified with the democratic therapeutic community movement – through the rapid growth and popularization of 'social psychiatry' in the 1950s. Between then and the 1970s, all the leading British psychiatric hospitals had wards operating as therapeutic communities. These include an early example of the fusion principle with Bertram Mandelbrote – an Oxford psychiatrist whose admission ward was Phoenix, but he also was responsible for founding the Ley Community, an addiction therapeutic community (Mandelbrote, 1965; Wilson, 1978). The therapeutic communities also moved out of the wards into the community and began to be applied to day units. Other nonhospital residential units also developed, for example Richmond Fellowship Hostels (Jansen, 1980).

Another development was in 1962 when HMP Grendon opened in Buckinghamshire, as a therapeutic community prison – not a drug rehabilitation wing, but as a mental health unit (Gunn et al., 1978; Morris, 2004). At the same time as HMP Grendon was established, Dennie Briggs opened Chino, a prison in California (Whiteley et al., 1972). Both were based on the Henderson model.

British-model therapeutic communities' fortunes in the prisons have waxed and waned for the last 40 years; HMP Grendon still exists with its six therapeutic community programs, as do several others in the British prison system: but as many others have come and gone over the same period, unable to adapt and survive sufficiently to a fundamentally authoritarian environment (Parker, 2007).

In 2000, two replicas of Henderson were set up: but they did not survive the first decade of the twenty-first century, as the expense of providing hospital beds became increasingly unacceptable to the funding authorities. Over the same period, a number of nonresidential units were established, including day units (most commonly operating 3 days per week), and

a few others using a 'low dose' treatment of 1 or 2 days per week.

Theoretical Foundations

In terms of theoretical underpinnings, Tom Main's approach was more explicitly based on psychoanalytic theory; the Henderson Hospital and other therapeutic communities were founded on a social learning/sociotherapy model. Group analytic theory offers a theoretical bridge between them, and many modern therapeutic community practitioners are trained in group analytic psychotherapy. The modern British therapeutic communities also incorporate and use a range of other therapeutic approaches, including cognitive and behavioral therapy; humanistic and integrative psychotherapy; art, music, movement, and the creative therapies; psychoeducation; systemic work with families; and other less mainstream ones (see Campling and Haigh, 1999). So a 'British Model Therapeutic Community' now describes the format of therapy – in groups and with a particular nature of relationships and democratic procedures – rather than a specific theoretical orientation.

The group-analytic based therapeutic communities have reached a very high level of sophistication and complexity which is perhaps best demonstrated by their extensive development in Greece, in organizations such as the treatment, training, and research communities of the Open Psychotherapeutic Centre and Kypseli in Athens, as well as the Department of Preventative Psychological Medicine in the Greek Navy (Kokkinidis, 1988).

The original definition of the British model comes from anthropological research at Henderson in the 1950s, to produce what have been called Rapoport's four 'Articles of Faith': *permissiveness*, *reality confrontation*, *democratisation*, and *communalism*. Permissiveness means that all members should tolerate one another; communalism suggests there should be tight-knit, intimate sets of relationships; democratization implies every member of the community (clients and staff) should share equally in decision making about community affairs; and reality confrontation suggests residents should be continually presented with interpretations of their behavior as it is seen by others, in order to counteract their tendency to distort, deny, or withdraw from their difficulties in getting on with others (Rapoport, 1960).

Haigh (1999) brought together various psychological and sociological theories to explore a necessary sequence of experiences for 'secondary emotional development' (as in a therapeutic community), and how this might relate to 'primary emotional development' (as in all successful human growth). The five sequential 'principles' are attachment, containment, communication, inclusion, and agency, with corresponding 'cultures' of belonging, safety, openness, involvement, and empowerment.

The American Model

In 1958, a new kind of therapeutic community was created in America, by the forceful personality of its originator: Chuck Dederich. He started by setting up a weekly group for helping ex-Alcoholics Anonymous members and ex-addicts in his own flat, based around free association, 'attack' therapy groups or 'Synanons,' and educational seminars, particularly based on philosophical ideas. This was the beginning of Synanon, and it was the first of what came to be known as 'concept-based,' 'hierarchical,' 'behavioral,' or 'programmatic,' and more recently 'addiction,' therapeutic communities. They were self-help communities, created by ex-alcoholics and drug addicts, which were called 'concept-based' because they referred to a set of explicit concepts about the nature of drug addiction and its treatment, based on psychological explanations of the causes of drug addiction and how a therapeutic community could help people live drug-free lives. The concepts are based on three main themes: those which explain the nature of the addict's problems; those which demonstrate how therapy works; and those which underpin how the addict should behave in the therapeutic community (Kennard, 1998). Another of the major characteristics of Synanon was its profound antiprofessionalism, with emphasis on self-help, including taking responsibility for one's situation, and the inclusion of nonaddict community members ('straights,' 'lifestylers,' and 'squares') (Casriel, 1963; Yablonsky, 1965).

Although the original group developed out of contacts through AA, Synanon eventually became concerned mainly with the rehabilitation of ex-drug addicts. It developed a very explicit program revolving around the highly confrontative encounter group called the 'game' or the 'stew,' based on the notion of 'no-holds-barred' honest confrontation, and particularly the challenging of emotional defenses, or 'cop-outs.' Communities had a very strict hierarchical structure, up which newcomers had to slowly work their way. The communities also functioned to a model of an autocratic family structure.

Synanon communities proliferated, but in 1963 a derivative of Synanon also developed – the Daytop village ('DAYTOP' stands for Drug Addicts Treated On Probation), which was started by a team of psychiatrists and probation officers who had been impressed by a visit to Synanon (Casriel, 1971; Sugarman, 1974). In 1968, another major therapeutic community organization was set up – Phoenix House, New York (Biase and Rosenthal, 1969). It became the largest and best known concept house, with places for about 400 residents at any one time (De Leon, 1973). In 1970, three graduates from Phoenix House came to London and set up a concept house with Griffith Edwards, at the same time as Ian Christie was setting up Alpha House in Portsmouth (Kooyman, 2001); others were later set up throughout the United Kingdom.

The spread of addictions therapeutic communities across the world has been considerable in the last two decades. There are now six organizations in world regions to coordinate activities. The Federation of Latin American Therapeutic Communities covers South America, and they now have more than 600 programs in over 20 countries. The Asian Federation of Therapeutic Communities covers the whole of Asia, and gives its main current preoccupations as war, the easy availability of very pure heroin, and very little coordination of responses to the problems. The European Federation of Therapeutic Communities covers Western Europe and has an extra 'European ingredient,' probably related to its separate roots in

psychoanalysis, critical theory, and sociology – for example, the work of Foucault and Lacan; it is also closely related to the European Working Organisation for Drugs-Orientated Research. The Federation of Therapeutic Communities of Central and Eastern Europe now covers communities in Eastern Europe, and has a flavor more of 'action learning' – by communities undertaking vigorous outdoor activities in their programs. Geographically most remote, there is the Australasian Therapeutic Communities Association, which was founded in 1986 and covers Australia, New Zealand, and Papua New Guinea. Therapeutic Communities of America covers the United States, Canada, and Mexico.

The American addiction therapeutic community treatment model is based on a number of concepts:

- Self-help
- The use of community as method
- Safety in acceptance
- Hierarchical structure
- Verbal reprimands and learning experiences
- Encounter groups
- Family work
- Drug abstinence
- Social confidence training
- Relating

The last few years have brought about some changes. These 'changed' therapeutic communities are referred to by, for example, Broekaert et al. (2006) as 'the new therapeutic communities' and have largely developed in terms of complex networks dealing with special needs and specific target groups. They have included treating different types of people, such as people with dual diagnosis, children of abusers, homeless people, those with posttraumatic stress disorders, immigrants, prostitutes, HIV sufferers, and street children in Latin America.

The 'Fusion' Model

The history of the two traditions of the main types of therapeutic community shows that very different models are now coming together, in the face of similar challenges (Haigh and Lees, 2008). In the past, the differences were more stark, but changes in the political and social context have influenced the programs in both main types of therapeutic community: some former therapeutic community practices are no longer regarded as acceptable; client rights and the health consumerism have changed treatment practice; there is increasing professionalization of staff, together with increased employment of ex-service users in both types of therapeutic community. Both types are also increasingly finding their target client groups that are ever more similar: more people entering democratic therapeutic communities have addiction issues, and more people entering addiction therapeutic communities have mental health issues. Both types have more structured programs, treatment times, and clear career progressions through treatment for clients, as well as extensive after-care provision. Both emphasize the importance of the therapeutic community as the mode of treatment (De Leon's 'community as method'), and the importance within this of the emancipatory use of groups and peer therapy. The successful growth in Europe and Australasia of therapeutic communities which deliberately combine the democratic and addiction models has led to a strong hybrid, with strengths from both and without some of the problems associated with thoroughbreds. They offer a model for the future development, and possible protection, of therapeutic communities.

Pathways and Process: Phases in the Therapeutic Journey

There are usually four distinct but overlapping phases in a member's journey to and through a therapeutic community.

Engagement Phase

Referral, preparation, and selection procedures are an integral part of therapeutic community practice, involving both the prospective member and existing residents as active participants in the process, which start with referral or self-referral. Many prospective members of therapeutic communities are wary or fearful of the forthcoming therapy, and need support and encouragement to persist.

Assessment and Preparation Phase

When a decision to proceed toward formal treatment has been made, several arrangements may be needed. These include formal assessment processes, practical planning, and agreeing a treatment contract. This work is frequently arranged through the use of an 'assessment and preparation group,' which is also designed to be a time-limited foretaste of what the treatment phase entails.

Treatment Phase

This usually begins with a formal 'case conference' or 'selection panel' including current community members with a decision made by voting. Subsequent therapy programs vary considerably: from 1 day per week to whole-time residential; from predominantly sociotherapy to a range of psychoanalytic, cognitive-behavioral, humanistic and interpersonal, and systemic groups; from a few weeks duration to several years, either time-limited or open-ended; and from group size of less than six to more than 50. Some communities include individual therapy, but others consider this inimical to the group dynamic process. A modal or typical program would be for between 3 and 5 days per week, with all interventions in groups comprising a mixture of community meetings, small therapy groups, shared lunch, and informal time together.

Recovery Phase

Until recently, many therapeutic communities had a 'cliff-edge' ending, where one day members are able to have the community's full emotional and practical support, and the next day are not allowed to contact any other members. Although this

has some theoretical justification in terms of 'coming to term with endings,' and has strong advocates among ex-members of therapeutic communities, it is now generally considered better practice to support members over the leaving process, and then into re-establishing mainstream social networks. This can be done with a specific 'leavers group,' that members join while in the full treatment phase and continue afterward, for either a fixed or indefinite period. These groups normally include a practical focus, and are social and supportive rather than exploratory and therapeutic. For those who are ready and able, they can have objectives of securing employment or education for members.

Evidence

Personality Disorders

While efficacy in the treatment of personality disorders (PD) is still questioned, the picture has recently become clearer with the publication of the first systematic review of therapeutic community treatment for people with personality disorders (Lees et al., 1999). A meta-analysis found that 19 studies showed a positive effect within the 95% level of confidence while the remaining 10 straddled the neutral score. The authors concluded that there is strong evidence for the effectiveness of therapeutic communities. A more recent systematic review assessed evidence for interventions for people with PD in general and for dangerous and severe personality disordered offenders and made clear recommendations about the most promising treatment interventions for PD in use or currently in development. They concluded that "the TC model currently has the most promising evidence base in this poor field" (Warren et al., 2003).

Offending Behavior

Therapeutic communities have been established in prisons to deal with disruptive, violent inmates and also with the underlying problems of antisocial PD. Results at Barlinnie Special Unit in Scotland demonstrated substantial reductions in violent incidents within prison. Reconviction studies carried out at Grendon, a prison run entirely on therapeutic community lines, found that prisoners had lower rates of reconviction, fewer custodial sentences, and fewer reconvictions for violent offenses than prisoners on the Grendon waiting list who never went there. Those who stayed at Grendon longer than 18 months showed the greatest reductions in reconviction rates. Reoffending rates have also been found to be lower in the former Federal Republic of Germany for prisoners in Social Therapeutic Institutions than those receiving standard prison sentences (Rawlings, 1999). In 2004 the 'Democratic Therapeutic Community Core Model' was accredited as a treatment program for use in prisons in England and Wales. Such programs must show evidence of their capacity to impact on dynamic risk factors known to be associated with reoffending. Risk factors that therapeutic communities have been found to have a positive influence on include negative attitudes toward authority, identification with antisocial role models, and acceptance of responsibility for offending behavior.

Drug Dependence

Therapeutic communities for drug dependence use the hierarchical or concept-based model. These form one part of the range of treatments for drug abuse, which includes other residential models such as Christian communities and the Minnesota model, as well as methadone maintenance programs, and psychotherapy. Therapeutic communities for drug dependence, often modified to suit different local cultures, can now be found in many European and international countries as well as in the United States where they originated. Several national studies have evaluated the outcome of these programs. Randomized controlled trials show that no-treatment groups have a higher level of recidivism than those who complete treatment in a prison therapeutic community, and that recidivism is further reduced by participation in a aftercare therapeutic community (Wexler, 1997). Since 1995 a number of these therapeutic communities have also been established in English prisons. The general conclusion, for both secure and nonsecure therapeutic communities, is that residents who stay in programs for longer periods have lower rates of drug use and criminal behavior and higher rates of employment than those who stay for shorter times.

First Episode Psychosis

A small number of therapeutic communities have been developed in the United Kingdom, the United States, Switzerland, and Germany on the principle that first episodes of psychosis can be effectively treated in low-stress family-like settings providing round the clock personal support, with no or minimal use of neuroleptics. This has become known as the Soteria model. Two Soteria houses, in California and Berne, have been subjected to randomized or matched control trials comparing them with usual hospital treatment. One study found that completing subjects with schizophrenia exhibited a large effect size benefit with Soteria treatment, especially in the areas of psychopathology, work, and social functioning. Length of stay in Soteria Berne was initially longer but this was subsequently reduced to less than the admission ward. In both studies the 2-year outcomes were at least as good in the Soteria group and less antipsychotics were prescribed for the Soteria group (Ciompi and Hoffman, 2004). A 20-year study of an acute psychiatric ward in Finland found that people with acute psychotic episodes and borderline conditions seemed to benefit from the therapeutic community model with a high level of support, negotiation, order, and organization.

Severe and Enduring Mental Illness

The therapeutic community approach (Clark, 1965) has been widely used in large mental hospitals to counter the effects of institutionalization and to mobilize the residual capacity of those suffering from chronic mental illness for social relationships, purposeful employment, and personal responsibility. With the reprovision of services for people with enduring mental illness in the community, therapeutic community principles have been found to be an effective way of structuring staffed hostels and homes. In one version of this, the 'ward in

a house,' the model is close to the original practice of the York Retreat, an antecedent of therapeutic communities (Leff and Trieman, 1997). The challenges presented by the severely mentally ill chemical user have also been addressed using a modified therapeutic community with some evidence of success. Three main modifications required to the TC structure were increased flexibility, decreased intensity, and greater individualization (Sacks, 2000).

Children and Adolescents

Therapeutic communities for children and adolescents exist for a variety of needs: learning disability, delinquency, and emotional disturbance. Little systematic evaluation has been carried out. A survey of 186 children in 9 therapeutic communities for emotionally disturbed children found evidence of increased stability and hopeful outcomes for those who stayed. A 20-year follow-up of 28 children in one community reported evidence of long-term improvement (Rose, 1997).

Wright and Richardson, reviewing the current state of research for therapeutic communities for children and young people, conclude that, "when rigorous quality controls are introduced there are as yet too few studies to draw any aggregated conclusions. Perhaps the clearest qualified statement would be that there is low level evidence that some residential therapeutic placements produce changes in the mental and social functioning of some young people who have been unable to cope with family life" (Wright and Richardson, 2003).

Learning Disabilities

Long-term residential communities for adults with learning disabilities exist in the charitable sector in many countries around the world. These are value based rather than evidence based. There is some evidence that many families express a strong preference for village-style communities such as the Camphill for their mentally handicapped relatives (Cox, 1995).

See also: Alcohol Interventions: Disease Models vs. Harm Reduction; Borderline (Patient) Personality; Deinstitutionalization, Psychiatry of; Group Processes, Social Psychology of; Military and Disaster Psychiatry; Personality Disorders.

Bibliography

Barnes, M., Berke, J., 1971. Mary Barnes: Two Accounts of a Journey through Madness. Hart-Davis MacGibbon, London.

Bazeley, E.T., 1928. Homer Lane and the Little Commonwealth. Allen & Unwin, London.

Berke, J.H., 1980. Kingsley hall. In: Jansen, E. (Ed.), The Therapeutic Community. Croom Helm, London, pp. 95–101.

Berke, J.H., 1982. The Arbours Centre. International Journal of Therapeutic Communities 3 (4), 248–261.

Biase, D.V., Rosenthal, M., 1969. Phoenix house: therapeutic community for drug addicts. Hospital and Community Psychiatry 20, 27–30.

Bion, W.R., 1961. Experiences in Groups and Other Papers. Tavistock, London.

Bion, W.R., Rickman, J., 1943. Intra-group tensions in therapy. Lancet 2, 678–681.

Bloor, M.J., McKeganey, N.P., Fonkert, J.D., 1988. One Foot in Eden: A Sociological Study of the Range of Therapeutic Community Practice. Routledge, London.

Brendtro, L.K., Ness, A.E., 1983. Re-educating Troubled Youth: Environments for Teaching and Treatment: Modern Application in Social Work Series. Aldine de Gruyter, Birmingham.

Bridgeland, M., 1971. Pioneer Work with Maladjusted Children. Staples Press, London.

Bridger, H., 1946. The Northfield experiment. Bulletin of the Menninger Clinic 10, 71–76.

Bridger, H., 1985. Groups in open and closed systems. In: Ottenberg, D.J. (Ed.), The Therapeutic Community Today. A Moment of Reflexion in its Evolution. Proceedings of the 1st World Institute of Therapeutic Communities. Contro Italiano do solidarietà, Rome, pp. 54–70.

Broekaert, E., Vandevelde, S., Soyez, V., Yates, R., Slater, A., 2006. The third generation of therapeutic communities: the early development of the TC for addictions in Europe. European Addiction Research 12, 1–11.

Burn, M., 1956. Mr Lyward's Answer. Hamish Hamilton, London.

Campling, P., Haigh, R. (Eds.), 1999. Therapeutic Communities: Past, Present and Future. Jessica Kingsley Publishers, London.

Casriel, D., 1963. So Fair a House: The Story of Synanon. Prentice Hall Inc., New York.

Casriel, D., 1971. The Daytop story and the Casriel method. In: Blank, L., Gottsegen, G.G., Gottsegen, M.G. (Eds.), Confrontation: Encounters in Self and Interpersonal Awareness. MacMillan, New York, pp. 179–193.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Boston.

Ciompi, L., Hoffman, H., 2004. Soteria Berne: an innovative milieu therapeutic approach to acute schizophrenia based on the concept of affect-logic. World Psychiatry 3 (3), 140–146.

Clark, D.H., 1964. Administrative Therapy. Tavistock, London.

Clark, D.H., 1965. The therapeutic community – concept, practice and future. British Journal of Psychiatry 3, 947–954.

Cooper, D., 1967. Psychiatry and Anti-psychiatry. Tavistock, London.

Cox, C., 1995. The case for village communities for people with learning disabilities. British Journal of Nursing 4 (19), 1130–1134.

De Leon, G., 1973. The Phoenix house therapeutic community: changes in psychopathological signs. Archives of General Psychiatry 28, 131–135.

De Leon, G., 1983. The next therapeutic community: autocracy and other notes toward integrating old and new therapeutic communities. International Journal of Therapeutic Communities 4 (4), 249–261.

DeLeon, G. (Ed.), 1977. Community as Method: Therapeutic Communities for Special Populations and Special Settings. Greenwood Publishing Group, Inc., Westport, CT.

De Leon, G., Sacks, S., McKendrick, K., 1999. Modified therapeutic communities for homeless mentally ill chemical abusers; emerging subtypes. American Journal of Drug and Alcohol Abuse 25 (3), 493–513.

De Leon, G., Sacks, S., Staines, G., McKendrick, K., 2000. Modified therapeutic communities for homeless mentally ill chemical abusers: treatment outcomes. American Journal of Drug and Alcohol Abuse 26 (3), 461–480.

Foulkes, S.R., 1948. Introduction to Group-Analytic Psychotherapy. Heinemann, London.

Foulkes, S.R., 1964. Therapeutic Group Analysis. George Allen & Unwin, London.

Franklin, M.E., 1966. Q Camp. Planned Environment Therapy Trust, London.

Glatt, M.M., 1974. An addicts' therapeutic community in a London prison. Lecture presented at International Council on Alcohol and Addictions, Copenhagen Institute.

Glatt, M., 1985. Reflections on the working and functioning of an addicts' therapeutic community within a prison: the Wormwood Scrubs Annexe. In: Prison Reform Trust (Ed.), Prison Medicine: Ideas on Health Care in Penal Establishments. Prison Reform Trust, London, pp. 83–98.

Gunn, J., Robertson, G., Dell, S., Way, C., 1978. Psychiatric Aspects of Imprisonment. Academic Press, London.

Gunderson, J.G., Will Jr., O.A., Mosher, L.R. (Eds.), 1983. Principles and Practice of Milieu Therapy. Jason Aronson, Northvale, NJ.

Haigh, R., 1999. The quintessence of a therapeutic environment. In: Campling, P., Haigh, R. (Eds.), Therapeutic Communities: Past, Present and Future. Jessica Kingsley Publishers, London, pp. 246–257.

Haigh, R., Tucker, S., 2004. Democratic development of standards: a quality network of therapeutic communities. Psychiatric Quarterly 75, 263–277.

Haigh, R., 2005. Charismatic ideas: coming or going? Therapeutic Communities 26 (4), 367–382, 7th Annual "Maxwell Jones Lecture", RIBA, London, 10 September 2004.

Haigh, R., 2007a. New kids on the block: the government-funded English personality disorder services. Therapeutic Communities 28 (3), 300–310.

Haigh, R., 2007b. The 16 personality disorder pilot projects. Mental Health Review Journal 12 (4), 29–39.

Haigh, R., Lees, J., 2008. Fusion TCs: divergent histories, converging challenges. Therapeutic Communities 29 (4), 347–374.

Haigh, R., Tucker, S., 2004. Democratic development of standards: a quality network of therapeutic communities. Psychiatric Quarterly 75, 263–277.
Harrison, T., 2000. Bion, Rickman, Foulkes and the Northfield Experiments. Jessica Kingsley Publishers, London.
Jansen, E., 1980. The Therapeutic Community. Croom Helm, London.
Jones, M., 1946. Rehabilitation of forces neurosis patient to civilian life. British Medical Journal 1, 533–535.
Jones, M., 1953. The Therapeutic Community: A New Treatment Method in Psychiatry. Basic Books, New York.
Jones, M., 1979a. Therapeutic communities, old and new. American Journal of Drug and Alcohol Abuse 6 (2), 137–149.
Jones, M., 1979b. The therapeutic community, social learning and social change. In: Hinshelwood, R.D., Manning, N. (Eds.), Therapeutic Community – Reflections and Progress. Routledge & Kegan Paul, London.
Jones, M., 1984. Why two therapeutic communities. Journal of Psychoactive Drugs 16 (1), 23–26.
Jones, M., 1968. Social Psychiatry in Practice. Penguin, Harmondsworth.
Jones, M., Lewis, A., 1941. Effort syndrome. British Medical Journal 2, 338–340.
Kennard, D., 1998. An Introduction to Therapeutic Communities. Jessica Kingsley, London.
Kennard, D., Haigh, R., 2009. In: Gelder, Michael (Ed.), The New Oxford Textbook of Psychiatry, second ed. Oxford University Press, Oxford, pp. 142–148, (Chapter 6.3.9).
Kokkinidis, A., 1988. A therapeutic community in the Greek Navy. International Journal of Therapeutic Communities 9 (4), 299–306.
Kooyman, M., 2001. The history of therapeutic communities: a view from Europe. In: Rawlings, B., Yates, R. (Eds.), Therapeutic Communities for the Treatment of Drug Users. Jessica Kingsley Publishers, London.
Laing, R.D., 1971. The Politics of the Family and Other Essays. Tavistock, London.
Laing, R.D., Esterton, A., Cooper, D., 1965. Results of family oriented therapy with hospitalized schizophrenics. British Medical Journal II, 1462–1465.
Lees, J., Manning, N., Rawlings, B., 1999. Therapeutic Community Effectiveness: A Systematic International Review of Therapeutic Community Treatment for People with Personality Disorders and Mentally Disordered Offenders. CRD Report 17, University of York, NHS Centre for Reviews and Dissemination.
Leff, J., Trieman, N., 1997. Providing a comprehensive community psychiatric service. In: Leff, J. (Ed.), Care in the Community: Illusion or Reality. Wiley, Chichester, pp. 189–201.
Main, T.F., 1946. The hospital as a therapeutic institution. Bulletin of the Menninger Clinic 10, 66–70.
Main, T.F., 1977. The concept of the therapeutic community, variations and vicissitudes. Group Analysis 10, 2–16.
Mandelbrote, B.M., 1965. The use of psychodynamic and sociodynamic principles in the treatment of psychotics. Comprehensive Psychiatry 6 (6), 381–387.
Manning, N.P., 1989. The Therapeutic Community Movement: Charisma and Routinization. Routledge, London.
Morris, M., 2004. Dangerous and Severe – Process, Programme and Person. Grendon's Work. Jessica Kingsley Publishers, London.
Moschonas, D., 2008. Therapeutic community principles and inpatient care in the Greek army psychiatric service. Therapeutic Communities 29 (2), 160–179.
Neill, A.S., 1937. That Dreadful School. Herbert Jenkins, London.
Neill, A.S., 1967. Summerhill. Gollancz, London.
NIMHE, 2003. Personality Disorder: No Longer a Diagnosis of Exclusion: Policy Implementation Guidance for the Development of Services for People with Personality Disorder. Department of Health, London.
Norton, K., 2006. Setting up New Services in the NHS. 'Just Add Water!'. Jessica Kingsley Publishers, London.
Parker, M. (Ed.), 2007. Dynamic Security. The Democratic Therapeutic Community in Prison. Jessica Kingsley Publishers, London.
Parry-Jones, W., 1981. The model of the Geel Lunatic Colony and its influence on the nineteenth century asylum system in Britain. In: Scull, A. (Ed.), Madness, Mad Doctors and Madmen. Athlene Press, London.
Pearce, S., Pickard, H., 2013. How therapeutic communities work: specific factors related to positive outcome. International Journal of Social Psychiatry 59 (7), 636–645.
Rapoport, R.N., 1960. Community as Doctor. New Perspectives on a Therapeutic Community. Tavistock Publications, London.
Rawlings, B., 1999. Therapeutic communities in prisons: a research review. Therapeutic Communities 20 (3), 177–193.
Roberts, J., 1977. How to recognise a therapeutic community. Prison Service Journal 111, 4–7.
Rose, M., 1997. Transforming Hate to Love: An Outcome Study of the Peper Harow Treatment Process for Adolescents. Routledge, London.
Rose, M., 1990. Healing Hurt Minds: The Peper Harow Experience. Tavistock/Routledge, London.
Rothman, D., 1971. The Discovery of the Asylum. Brown, Boston, Little.
Royal College of Psychiatrists, 2006. Service Standards for Therapeutic Communities. http://rcpsych.ac.uk/PDF/Service%20Standards%20for%20Therapeutic%20Communities%205th%20Edition.pdf (accessed 29.12.13.).
Rubel, J.G., Baker, K.G., Bratten, T.E., Hartwig-Thomson, L., Smirnoff, A.M., 1982. The role of structure in the professional model and the self-help concept of the therapeutic community: different strokes for different folks. International Journal of Therapeutic Communities 3 (4), 218–232.
Sacks, S., 2000. Co-occurring mental and substance use disorders: promising approaches and research issues. Substance Use and Misuse 35 (12–14), 2061–2093.
Scull, A., 1979. Museums of Madness. Allen Lane, London.
Sugarman, B., 1974. Daytop Village – a Therapeutic Community. Holt, Rinehart & Winston, New York.
Sugarman, B., 1984. Towards a new, common model of the therapeutic community: structural components, learning processes and outcomes. International Journal of Therapeutic Communities 5 (2), 77–98.
Tucker, S. (Ed.), 2000. A Therapeutic Community Approach to Care in the Community. Dialogue and Dwelling. Jessica Kingsley Publishers, London.
Tuke, S., 1813. Description of the Retreat. Republished in 1996. Process Books, London.
Vanier, J., 1982. The Challenge of L'Arche. Darton, Longman & Todd, London.
Warren, F., Preedy-Fayers, K., McGauley, G., Pickering, A., Norton, K., Geddes, J., Dolan, B., 2003. Review of Treatments for Severe Personality Disorder. Home Office Online Report 30/03. Home Office, London.
Wexler, H., 1997. Therapeutic communities in American prisons. In: Cullen, E., Jones, L., Woodward, R. (Eds.), Therapeutic Communities for Offenders. Wiley, Chichester, pp. 161–179.
Whiteley, J.S., 1980. The Henderson hospital. International Journal of Therapeutic Communities 1 (1), 38–58.
Whiteley, J.S., Briggs, D., Turner, M., 1972. Dealing with Deviants. The Hogarth Press, London.
Williams, R., 19 December 2002. The Dimbleby Lecture. BBC Online.
Wills, D.W., 1969. The Hawkspur Experiment, second ed. Allen & Unwin, London.
Wilson, S., 1978. The effect of treatment in a therapeutic community on intravenous drug abuse. British Journal of Addiction 73 (4), 407–411.
Wright, J.C., Richardson, P., 2003. The challenge of research. In: Ward, A., Kasinski, K., Pooley, J., Worthington, A. (Eds.), Therapeutic Communities for Children and Young People. Jessica Kingsley, London, pp. 244–253.
Yablonsky, L., 1965. Synanon. The Tunnel Back. Basic Books, New York.
Yates, R., Wilson, J., 2001. The modern therapeutic community: dual diagnosis and the problem of change. In: Rawlings, B., Yates, R. (Eds.), Therapeutic Communities for the Treatment of Drug Users. Jessica Kingsley Publishers, London and Philadelphia.

Relevant Websites

www.communityofcommunities.org.uk.
www.enablingenvironments.com.
www.indtc.org.
www.livinglearningexperience.com.
www.pettarchiv.org.uk.
www.personalitydisorder.org.uk.
www.soterianetwork.org.uk.
www.therapeuticcommunities.org.
www.wftc.org.
www.winship.info.

Therapeutic Effectiveness Assessment: Approaches in Contemporary Research

Mihaela Matei, Centre Européen d'Etudes et de Recherche Droit et Santé, Université Montpellier I, France
François Lemaire, Emméritus at the University of Paris Est Créteil, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Today, the way in which therapeutic effectiveness is assessed is once again subject to criticism. Burdensome bureaucratic regulatory procedures in traditional clinical research, the costs of health care systems, and the uncertainness as to the most effective treatment options available have raised significant questions about whether traditional research methods, such as randomized controlled trials (RCTs), can adequately address the contemporary issues posed by the health care systems. In particular, the recent development of health economics-related research demonstrated the limitations of RCTs as research methods (Kumar, 2013), widely criticized for its incapacity to reduce uncertainty about relative effects of treatments in individual patients. In order to better inform decision-makers clinicians as well as patients on the best practices available, the health economic studies pointed out the need for developing new research methods and designs more likely to provide evidence-based information on the effectiveness, benefits, harms, and costs of different treatments. These changes in clinical research designs and methods of evaluation of therapeutic effectiveness raise new ethical concerns and regulatory questions that need to be addressed properly. Recent initiatives to revise the legislative and regulatory framework for biomedical research (USA, EU, and France) directly tackle some of these issues. After describing contemporary advances in types of research on therapeutic effectiveness, this article will give an overview of the ongoing revision of the regulatory and legislative framework in the USA, EU, and France with a special focus on the way these initiatives address the legal and regulatory issues raised by the therapeutic effectiveness evaluation.

Currently new types of research that focus on the evaluation of treatment effectiveness – such as comparative effectiveness research (CER), patient-led research (PLR), or Internet research – are offered alongside classical randomized controlled trials (RCTs). These approaches open up new possibilities of assessing the therapeutic effectiveness of patient care in conditions that are more representative of the 'real world' (Chang and Winkelmayer, 2012). First of all, in contrast to traditional clinical trials which remain the gold standard for showing the safety and efficacy of a new medicinal product 'under ideal conditions' (Kumar, 2013), these new types of research are intended to more accurately reflect the effectiveness of therapies delivered in typical clinical practice. For example, in the case of CER, the evidence is generated through studies that compare two or more alternative tests, therapies, or procedures, most of which are routinely used in clinical settings. In particular, there is a growing interest in development of research on usual care, which is intended to assess which treatments are most effective in routine care. Subsequently, the new models attach utmost importance to the role of practice in the development of generalizable knowledge: "practice is a continuous source of data for the production of generalizable knowledge and the knowledge that is produced is used to continuously change and improve practice" (Kass et al., 2013). In addition, in order to fill in gaps in patient health information, researchers claim access to all of the available evidence: data from existing clinical trials along with data available in patient health records and patient-generated data outside of clinical settings. The development of a field of research such as Internet research may facilitate the collection of this type of health data and help researchers to fill in these gaps.

These changes in clinical research designs and methods involve a more comprehensive clinical data-sharing system, interventions at the cluster level, and direct patient involvement in the conduct of research projects, which in turn raise new ethical and regulatory questions that need to be addressed properly.

In particular, these changes challenge the capacity of the current research ethical and regulatory paradigm, which is dominated by the sharp distinction between research and clinical practice, to foster new types of research which links the two activities (Solomon MZ and Bonham AC, 2013). Subsequently, many authors argue that a new ethical foundation needs to be developed in order to facilitate *both care and research likely to benefit patients* ("Hastings Center Special Report of Ethical Oversight of Learning Health Care Systems, pp. S2 to S44," 2013). The new paradigm should abandon the 'faulty distinction' between clinical research and clinical care activities as well as the research ethical oversight based on this distinction in favor of a new framework integrating medical research and practice (Kass et al., 2013).

In spite of these arguments, recent initiatives to revise the legislative and regulatory framework for biomedical research (USA, EU, and France) tend to take a less radical position by proposing a new regulatory paradigm based on a risk-based approach without abandoning the distinction between research and practice. The creation of a minimal risk research category to which applies a risk-adapted oversight is intended to facilitate the assessment of therapeutic effectiveness.

After describing some of the contemporary advances in types of research on therapeutic effectiveness, this article will give an overview of the ongoing revision of the biomedical research regulatory and legislative framework in the USA, EU, and France with a special focus on the way these initiatives address the legal and regulatory issues raised by the therapeutic effectiveness evaluation.

Comparative Effectiveness Research

CER is largely recognized as a prominent method for the assessment of therapeutic effectiveness, especially when

standard clinical trials have important methodological limitations. In contrast to traditional RCTs, which focus on novel interventions, CERs often target existing therapies and widely adopted clinical practices in which there is clinical uncertainty about best treatments (Kass et al., 2013). Even though the concept of CER is not new, it only became a topic of priority in 2009, when President Barack Obama enacted into law the American Recovery and Reinvestment Act, which allocated 1.1 billion US dollars to CER studies.

The US Institute of Medicine (IOM) provided a comprehensive definition of CER: *CER is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care*. According to IOM, the key goal of CER is assisting consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels. In addition, CER aims to inform policy decisions about health care utilization in terms of cost-effectiveness and how to prioritize the use of public funds in particular with regard to "areas where private incentives are low relative to social benefits" (Meltzer et al., 2010).

The Agency for Healthcare Research and Quality defines two ways in which evidence can be found through CER ("Methods Guide for Effectiveness and Comparative Effectiveness Review. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Chapters available at: www.effectivehealthcare.ahrq.gov," 2014):

> *research reviews* consist in the systematic reviewing of existing evidence (e.g. analyses of the available evidence about the benefits and harms of each choice for different groups of people from existing clinical trials, clinical studies, and other research).

> *studies* that generate, prospectively, *new* evidence of effectiveness or comparative effectiveness of a test, treatment, procedure, or healthcare service.

An example of CER is the comparison of AMD treatments trials (CATT) study (Rosenfeld, 2011) which compared two anti-VEGF (vascular endothelial growth factor) drugs used in the treatment of age-related macular degeneration, a disease causing irreversible blindness. The trial showed ultimately that both drugs had the same effects on benefits and harms. This trial could have a great economical impact, as the cost of these two drugs is remarkably different (US $2000 versus $50).

The study by Hejblum et al. (2009) – the comparison of the use of daily routine chest radiographs in 21 ICUs to on-demand x-rays for mechanically ventilated patients – could also have consequences on health economics. Results showed that the number of x-rays that led to or contributed to therapeutic or diagnostic interventions were the same in both groups, as well as mortality and length of stay. Despite that there was a reduction of 30% of the number of x-rays in the group on demand.

Prospective observational cohorts are another example of CER. Generally they are easier to run, cost efficient, and closer to 'real life' than classical randomized studies. Despite their methodological limitations, well-done observational studies proved, in many instances, to be as efficient as RCTs (Benson and Hartz, 2000; Concato et al., 2000).

In the US, the IOM set up a list of priority research topics that should be addressed in order to improve health care through CER. Among these topics, therapeutic effectiveness assessment is central. For instance, the IOM's list includes comparisons of the effectiveness of the different assistive listening devices like cochlear implants, electric acoustic devices as well as various treatment interventions in eradicating methicillin-resistant *Staphylococcus aureus* in communities, institutions, and hospitals. It includes also different strategies for localized prostate cancer and comprehensive care coordination programs, such as usual care in managing children and adults with severe chronic disease, especially in populations with known health disparities, etc. (IOM, n.d.).

In view of these examples, CER can take many forms: retrospective data analysis, systematic reviews of literature, as well as prospective studies, at both individual and cluster levels. For example, cluster randomized trials (CRTs) became an increasingly important method of assessment of treatment effectiveness, since the research interventions in cluster trials may be directed at social units such as hospitals wards, classrooms, primary care practices, and communities (Weijer et al., 2011). As Weijer and colleagues explained, the CRT is more appropriate than an individually randomized trial in many circumstances – when the nature of intervention targets entire communities, in the case of training or education of health professionals, and when investigators desire to reduce the effect of the so-called 'treatment contamination' (i.e., for instance, the interferences between tested patients and controls). A fourth circumstance identified by the authors is when investigators wish to study both individual and group effects of an intervention.

Despite the growing interest in CER, studies carried out on the feasibility of this type of research have shown that the current ethical and legal framework discourages its implementation. Subsequently, there is a significant debate about appropriate oversight required by CER. It is commonly argued that the ethical regulatory oversight should be refocused on specific features of CER (Platt et al., 2014).

For example, research on usual care is not easy to perform since its evaluation is subject to the same level of regulatory and administrative burdens as experimental interventions. Another major question is whether informed consent should always be required for this type of research, since in most cases the intervention added (e.g., randomization) by research bears only minimal risk and has little effect on the patient's usual care. Is it ethical to perform usual care research without written consent when patients are randomly assigned (Faden et al., 2014)?

Another example is CRTs. Monica Taljaard and colleagues (Taljaard et al., 2013) argue that these trials have unique features due to substantial methodological differences between CRTs and conventional randomized trials: *differing units of allocation (for example, hospitals), intervention (for example, health professionals), and outcome measurement (for example, patients) within a single study*. As a consequence, standard ethics guidelines for research are not adapted to CRTs. In the absence of

specific ethical guidelines, a research team comprising 15 investigators from Canada, the UK, and the US developed international guidelines for CRTs, known as the 'Ottawa Statement' (Taljaard et al., 2013). It sets out 15 recommendations in the form of a checklist that may be used by researchers in ethics applications: justification of the CRT design, research ethics review, identification of research participants, informed consent, role of gatekeepers, evaluation of benefits and harms, and protection of vulnerable patients. With regard to informed consent, the recommendations define the conditions under which the waiver of consent is ethically acceptable: *Waivers or alterations of consent can be justified only when the study would not otherwise be feasible and when the risks involved are minimal* (recommendations 6 and 7).

From a legal and regulatory standpoint, it has been suggested that the best way out would be to increment a risk-based regulatory approach according to which ethical review is determined based upon the level of risk posed to the studied participants. In the light of this paradigm, CER involving no more than low risk should be reviewed according to a simplified procedure. For instance, in France, some of these legal and regulatory issues raised by research on usual care were first addressed in 2004. The current French legislation ("Loi no 2004- 806 du 9 août 2004 relative à la politique de santé publique," 2004 codified in the French Public Health Code) authorizes investigators to carry out clinical research on usual care involving negligible risks without requiring patients' consent. Even though the written consent is not legally required in this case, patients have to be informed and have the right to object to their inclusion in the study as well as the right to withdraw at any time. Currently, in France there is a legislative revision on the way which is intended to adapt the review of this type of research to the level of risk added by research to usual care.

Despite its practical benefits, CER also has limitations. For instance, ethical concerns are raised by the use of CER results to determine how to prioritize the public spending on health care. In addition, CER also faces some methodological challenges due to its attempt to produce evidence from secondary data sources or nonrandomized data: *it relies heavily on unverifiable assumptions that are required for the method to give a correct answer* (Edward L. Korn and Boris Freidlin, Methodology for Comparative Effectiveness Research: Potential and Limitations) (Marc L. Berger et al., Good Research Practices for Comparative Effectiveness Research).

Internet Research and Data-Sharing Initiatives

The development of 'Internet research' (Markham and Buchanan, 2012) not only as a tool but also as a field of research also contributes to the evaluation of therapeutic effectiveness in a more rapid and easier way. Therefore, compared to other modes of data collection, the Internet is considered by some researchers to be a democratized way of data collection, giving access to large databases of health data, produced and collected in both clinical and research contexts.

On a purely clinical level, the implementation of electronic medical records has helped clinicians develop programs that collect data on clinical outcomes and to use this information in order to improve quality and safety for their patients. In terms of research as well, many data-sharing initiatives tend to expand public access to raw (participant-level) data from drugs and devices trials (Peloquin et al., 2013). One of the potential benefits of accessing large health databases and data sharing is allowing independent researchers to check on clinical trial results by replicating analyses with the use of raw data. Other benefits would be to *increase the precision of estimates of treatment efficacy, detect safety problems unobservable in smaller samples, allow exploration of subgroup effects, and permit analysis of how therapeutic effects vary in different geographic settings because of such factors as population genetics and health care delivery systems* (Mello et al., 2013). Because of the therapeutic benefits of analyses based on the existing data, and the will not to expose patients to any further physical risks, data-sharing initiatives are regarded as ethically acceptable. Nevertheless, these initiatives raise additional practical and ethical issues: the extent to which considerations of informed consent would be taken into account when individual-level data are released, concerns about the risk of patients reidentification, questions about the ethical oversight required by data-sharing models, etc. (Peloquin et al., 2013).

The Patient-Led Research

The traditional research field centered on the investigator/sponsor team has expanded to 'patient-centered' (Kaye et al., 2012) or 'patient-led' (Vayena and Tasioulas, 2013) research methods, allowing practitioners access to new types of data on relative health outcomes and clinical effectiveness of medical treatments. Defined as *tools, programs, and projects that empower participants to engage in the research process* (Kaye et al., 2012), PLR relies heavily on patient involvement in the conduct of research projects. This emerging model is defined as a complementary approach to traditional research since PLR allows individuals to share information on their conditions (self-reporting data) and investigators to *gather data that standard research systematically overlooks (e.g., drug side-effects or off-label uses)* (Vayena and Tasioulas, 2013). It is also intended to make patients' consent more readily obtainable, *as participants can be directly approached for consent to the use of their identifiable information for new research purposes* (Vayena and Tasioulas, 2013). However, the PLR method raises important questions about its capacity to achieve the scientific rigor needed to complement conventional investigator-led research.

The 'Learning Health Care System'

In 2007, the US IOM called on health care providers to change the health system into a 'learning health care system' (LHS) (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007). This new framework is intended to provide constantly updated evidence about health care interventions and their effectiveness. The LHS is mainly defined as theoretical and ethical alternatives to the sharp research-treatment distinction, and is justified by the growing number of activities that cannot easily be classified as either research or therapy: quality improvement activities, CER, practice-based research,

building of networks, and large databases of aggregated health care data, etc. Faden et al. argue that it is useless in most cases to try and distinguish research activity from clinical practice: *all such activities are learning activities* (Faden R.R., et al., 2013). By contrast, the LHS model sets a 'moral' priority on 'learning' health care activities that integrate research and clinical practice. More specifically, the LHS involves an interconnection of learning activities that combine observational research, CER, reviews of medical records, as well as the creation of registries alongside delivery of medical care.

The scientific evidence generated by the interconnection of these activities requires a greater access to health data, including to patient-level clinical trials data. Accordingly, the LHS model *urges a new ethical foundation, based on evidence-based practice decisions guided by the general knowledge produced from data collection activities essential to learning health care systems rather than on the research–practice distinction* (Largent et al., 2013).

From a legal and regulatory standpoint, one of the major impediments to the LHS model is the legal and regulatory framework pertaining to the use and reuse of clinical data. The current frameworks allow clinicians to share patients' health data with other clinicians, only under strict conditions, because of data protection rules and the risk of identification. Similar concerns around patient consent, privacy, and legitimate reuse need to be addressed with regard to the greater access and the sharing of clinical trials data, in particular of raw data. The impact on the sponsors' ability to patent inventions is another concern raised by the public disclosure of clinical trials. There is currently a worldwide effort at addressing these issues (Mello et al., 2013). Special attention is paid to the issues raised by public access to raw data, such as patient privacy, commercial interests, standardization, and quality. At the European level, the new EU clinical trials regulation imposes upon the sponsor the obligation to publish a summary of the results from clinical trials within a year of the trial's end. In addition, the EU regulation states that access to raw data should be permitted by the sponsor on a voluntary basis. The European Commission shall produce guidelines for the formatting and sharing of those data. In addition, the European Medicines Agency (EMA), which approves drugs for marketing within the EU, has proposed the expansion of access to data submitted in clinical trials' regulatory applications. The EMA announced in a draft policy in June 2014 that it will provide public access to some of the data in marketing dossiers, including participant-level data. Some of the data and documents will be published online excluding the raw data. The access to raw data will be controlled. Similarly, in the US, the Food and Drug Administration published in June 2013 a proposal to release deidentified participant-level data requesting public comments on its proposal.

In the light of these initiatives, the move toward greater transparency in the area of clinical data appears irrevocable.

The Necessity of Revising Existing Legislation and Regulation

Finally, these contemporary advances in therapeutic effectiveness assessment raise significant questions as to whether or not the current research legislation and regulation have kept pace with this evolution.

There is currently a worldwide effort at revising the biomedical legislation by using a new paradigm, which is commonly called the 'risk-based approach.' The issues raised by the evaluation of treatment effectiveness represent, to a great extent, the starting point for these legislative reforms.

For example, many academic sponsors have pointed out that the current legal and regulatory frameworks are not adapted to the research on existing therapies. One of their major critics is that the current legal provisions *apply largely irrespectively to risk to subject safety and without matching practical considerations and requirements* (European Commission, 2012a), in a 'one size fits all' manner (OECD, 2013). A report from the OECD (OECD, 2013) recommended that legislations on research should first establish a categorization of risks by taking account of the status of the medical product (authorized, off-label, innovative) as well as the invasiveness of additional procedures, the vulnerability of participants, and the deviation from usual practice. Secondly, they should adapt the regulatory requirements to that level of risks.

At the European level, the EU Commission has likewise advocated a risk-adapted approach to regulating research in its proposal of 17 July 2012 for a Clinical Trials Regulation (Reith et al., 2013; Markham A and Buchanan E, 2012). Although the scope of the proposal is limited to clinical research on medicinal products, the Commission faced multiples challenges. For instance, in its "Impact assessment report on the revision of the 'Clinical Trials Directive' 2001/20/EC" (DIRECTIVE, 2001/20/EC, 2001) the EU Commission recognized that the directive has been responsible of hindering therapeutic effectiveness assessment, both in terms of regulatory burdens and costs for administrative requirements. There has been a particular criticism of academic sponsors who are more likely than 'commercial sponsors' to carry out clinical trials on authorized drugs in order to improve the effectiveness of current treatments. In response to this, on 17 July 2012, the Commission adopted a Proposal for a Regulation (European Commission, 2012b) of the European Parliament and of the Council on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC, in which the Commission recognized that clinical trials on medicinal product that are covered by a marketing authorization or used according to published scientific evidence *are often of crucial importance for assessing standard treatments and diagnoses, thereby optimizing the use of medicinal products and thus contributing to a high level of public health.* The proposal was submitted to the European Parliament and Council who engaged in ordinary legislative procedure. After being largely modified, the final text of the new Clinical Trials Regulation, once voted, was published in the Official Journal on 27 May 2014 (European Parliament and Council, 2014).

The new European regulation is intended to facilitate the clinical trials on medicines that pose only minimal risks to subject safety compared to normal clinical practice. Accordingly, the regulation establishes a risk-based clinical trials typology and calibrates regulatory burdens to the level of risk involved by the research. For instance, specific provisions are intended to ease the monitoring, vigilance, traceability, and damage compensation, in clinical trials with 'low intervention.' The 'low intervention clinical trial' category covers clinical trial which fulfills all of the following conditions: *(a) the investigational*

medicinal products, excluding placebos, are authorised; (b) according to the protocol of the clinical trial, (i)the investigational medicinal products are used in accordance with the terms of the marketing authorisation; or (ii) the use of the investigational medicinal products is evidence-based and supported by published scientific evidence on the safety and efficacy of those investigational medicinal products in any of the Member States concerned; and (c) the additional diagnostic or monitoring procedures do not pose more than minimal additional risk or burden to the safety of the subjects compared to normal clinical practice in any Member State concerned. With regard to the informed consent requirements, the regulation authorizes investigators to obtain informed consent by 'simplified means' in the case of cluster trials involving only minimal risk, when the investigational medicinal products are authorized and used in accordance with the terms of the marketing authorization.

The French legislature enacted in 2012 a new Law on Research involving human subjects, commonly referred to as the Jardé law (Jardé law, n.d.). One of the major purposes of this law was to facilitate all biomedical research – including research on usual care – that involves only minimal risk for patients. It is important to recall that the French Parliament had already established in 2004 a minimal risk research category, under the name of 'research on usual care.' However, these provisions were never fully enforced, as the legal concept of 'research on usual care' was not properly defined. In addition, it was difficult for researchers to provide the evidence required for qualifying an intervention as usual care intervention (Matei et al., 2011). As a consequence, ethics committees were reluctant to authorize such clinical research. Most often these protocols were qualified as experimental research. Subsequently, the Jardé law is intended to match regulatory provisions to the risks posed to participants by research interventions. Accordingly, three categories of research have been identified according to the level of risks involved by research interventions: (1) interventional research involving higher than minimal risk, (2) interventional research involving minimal risk (covers only research on nonmedicinal products), and (3) noninterventional research. The law establishes specific provisions for the research category that involves only minimal risk for human subjects. In particular, only the review by an ethics committee is required, without authorization by the competent authority, and consent may be given orally.

The move toward a risk-adapted regulatory approach raises further key theoretical and practical issues that need to be addressed at both national and international levels. The defenders of learning health care models have already underlined one of these issues: how to draw the line between 'usual care' and 'research' under the risk-based approach. For example, the Jardé law defines interventional research as studies involving interventions *not justified by the person's usual care.* However, the line between 'usual care' and 'research activities' is often hard to discern. Another major question is the lack of definition of minimal risk. In France, the new law does not define the 'minimal risk.' The lack of legal definition will be overcome by the publication of a detailed list of interventions involving only minimal risk (such as blood sampling, external monitoring of blood pressure, EKG, etc.). This list will be established on empirical evidence. Similarly, the European regulation does not provide a definition of 'low intervention.' Instead, the difficulty has been circumvented by limiting the low intervention category to research on approved medicines, within and without their marketing authorization (and in this latter case, only when their use is evidenced based). The underlying assumption is that the existing knowledge and prior experience with the tested medicinal product prevents any unknown risk, which can be subject to debate.

To sum up, the recent advances in clinical research call for a radical change in attitudes toward therapeutic effectiveness assessment, based on stronger involvement of the participants and their families, translation of new information into medical evidence, and networks of shared clinical and research data (Navathe and Conway, 2010). The risk-based approach as a legal and regulatory solution to the legal issues raised by these changes in clinical research has its limitations. For instance, issues related to data protection legislation or intellectual property rights generated by the use of large health databases are only partially addressed. Also the involvement of patients in the assessment of clinical effectiveness has important implications in terms of responsibility and liability. Even though the regulatory risk-based approach neither addresses nor solves all the issues surrounding the contemporary development of research on therapeutic effectiveness, the enforcement of these regulations will permit us to appreciate its impact on these new approaches.

See also: Common Rule: US Regulatory Protections for Human Research Participants; Informed Consent of Research Subjects; Internet Research: Ethical Concerns; Privacy of Individuals in Social Research: Confidentiality.

Bibliography

Benson, K., Hartz, A.J., 2000. A comparison of observational studies and randomized, controlled trials. New England Journal of Medicine 342, 1878–1886. http://dx.doi.org/10.1056/NEJM200006223422506.

Chang, T.I., Winkelmayer, W.C., 2012. Comparative effectiveness research: what is it and why do we need it in nephrology? Nephrology Dialysis Transplantation: Official Publication of the European Dialysis and Transplant Association – European Renal Association 27, 2156–2161. http://dx.doi.org/10.1093/ndt/gfs154.

Concato, J., Shah, N., Horwitz, R.I., 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine 342, 1887–1892. http://dx.doi.org/10.1056/NEJM200006223422507.

Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the Approximation of the Laws, Regulations and Administrative Provisions of the Member States Relating to the Implementation of Good Clinical Practice in the Conduct of Clinical Trials on Medicinal Products for Human Use, 2001.

Emanuel, E.J., Menikoff, J., 2011. Reforming the regulations governing research with human subjects. New England Journal of Medicine 365, 1145–1150. http://dx.doi.org/10.1056/NEJMsb1106942.

Etheredge, L.M., 2010. Creating a high-performance system for comparative effectiveness research. Health Affairs Project Hope 29, 1761–1767. http://dx.doi.org/10.1377/hlthaff.2010.0608.

European Commission, 2012a. Impact Assessment Report on the Revision of the "Clinical Trials Directive" 2001/20/EC.

European Commission, 2012b. Regulation (Proposal for a) of the European Parliament and of the Council on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/EC.

European Parliament and Coucil, 2014. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/EC.

Faden, R.R., Beauchamp, T.L., Kass, N.E., 2014. Informed consent, comparative effectiveness, and learning health care. New England Journal of Medicine 370, 766–768. http://dx.doi.org/10.1056/NEJMhle1313674.

Faden, R.R., Kass, N.E., Goodman, N.E., Pronovost, P., Sh, Tunis, Beauchamp, T.L., 2013. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Centre Report S16.

Hastings Center Special Report of Ethical Oversight of Learning Health Care Systems, 2013. pp. S2–S44.

Hejblum, G., Chalumeau-Lemoine, L., loos, V., Boëlle, P.-Y., Salomon, L., Simon, T., Vibert, J.-F., Guidet, B., 2009. Comparison of routine and on-demand prescription of chest radiographs in mechanically ventilated adults: a multicentre, cluster-randomised, two-period crossover study. Lancet 374, 1687–1693. http://dx.doi.org/10.1016/S0140-6736(09)61459-8.

Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007. The Learning Healthcare System: Workshop Summary, the National Academies Collection: Reports Funded by National Institutes of Health. National Academies Press (US), Washington, DC.

IOM, n.d. Initial National Priorities for Comparative Effectiveness Research.

Jardé law, n.d. LOI n 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine (1), JO du 6 mars 2012.

Kass, N., Faden, R., Tunis, S., 2012. Addressing low-risk comparative effectiveness research in proposed changes to US federal regulations governing research. Journal of American Medical Association. 307, 1589–1590. http://dx.doi.org/10.1001/jama.2012.491.

Kass, N.E., Faden, R.R., Goodman, S.N., Pronovost, P., Tunis, S., Beauchamp, T.L., 2013. The research-treatment distinction: a problematic approach for determining which activities should have ethical oversight. Hastings Center Report Spec No. S4–S15. http://dx.doi.org/10.1002/hast.133.

Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S.M., Kanellopoulou, N., Lund, D., MacArthur, D.G., Mascalzoni, D., Shepherd, J., Taylor, P.L., Terry, S.F., Winter, S.F., 2012. From patients to partners: participant-centric initiatives in biomedical research. Nature Reviews Genetics 13, 371–376. http://dx.doi.org/10.1038/nrg3218.

Kumar, R., 2013. Health economics and cost-effectiveness research with special reference to hemato-oncology. Medical Journal Armed Forces India 69, 273–277. http://dx.doi.org/10.1016/j.mjafi.2013.06.003.

Largent, E.A., Miller, F.G., Joffe, S., 2013. A prescription for ethical learning. Commentary. Hastings Center Report Spec No. S28–S29. http://dx.doi.org/10.1002/hast.135.

Lidz, C.W., Garverich, S., 2013. What the ANPRM missed: additional needs for IRB reform. Journal of Law, Medicine and Ethics: A Journal of the American Society of Law, Medicine and Ethics 41, 390–396. http://dx.doi.org/10.1111/jlme.12050.

LOI no 2004-806 du 9 août 2004 relative à la politique de santé publique, 2004.

Markham, A., Buchanan, E., 2012. Ethical Decision Making and Internet Research.

Matei, M., Thalabard, J.-C., Laude, A., Misse, C., Chassany, O., 2011. Research on usual care: assessing the qualification criteria of the current legal framework. Presse Médicale Paris France 1983 (40), e189–196. http://dx.doi.org/10.1016/j.lpm.2010.10.022.

Mello, M.M., Francer, J.K., Wilenzick, M., Teden, P., Bierer, B.E., Barnes, M., 2013. Preparing for responsible sharing of clinical trial data. New England Journal of Medicine 369, 1651–1658. http://dx.doi.org/10.1056/NEJMhle1309073.

Meltzer, D., Basu, A., Conti, R., 2010. The economics of comparative effectiveness studies: societal and private perspectives and their implications for prioritizing public investments in comparative effectiveness research. PharmacoEconomics 28, 843–853. http://dx.doi.org/10.2165/11539400-000000000-00000.

Methods Guide for Effectiveness and Comparative Effectiveness Review. AHRQ Publication No. 10(14)-EHC063- EF, January 2014. Agency for Healthcare Research and Quality, Rockville, MD. Chapters available at: www.effectivehealthcare.ahrq.gov.

Navathe, A.S., Conway, P.H., 2010. Optimizing health information technology's role in enabling comparative effectiveness research. American Journal of Managed Care 16, SP44–47.

OECD, 2013. Recommendation on the Governance of Clinical Trials.

Peloquin, D., Combs, K., Bieber, N., Barnes, M., 2013. Draft Policy on Publication and Access to Clinical Trials Data Provides Broad Researcher Access to Participant-level Data.

Platt, R., Kass, N.E., McGraw, D., 2014. Ethics, regulation, and comparative effectiveness research: time for a change. Journal of the American Medical Association 311, 1497–1498. http://dx.doi.org/10.1001/jama.2014.2144.

Reith, C., Landray, M., Devereaux, P.J., Bosch, J., Granger, C.B., Baigent, C., Califf, R.M., Collins, R., Yusuf, S., 2013. Randomized clinical trials — removing unnecessary obstacles. New England Journal of Medicine 369, 1061–1065. http://dx.doi.org/10.1056/NEJMsb1300760.

Rosenfeld, P.J., 2011. Bevacizumab versus Ranibizumab for AMD. New England Journal of Medicine 364, 1966–1967. http://dx.doi.org/10.1056/NEJMe1103334.

Solomon, M.Z., Bonham, A.C., 2013. Ethical Oversight of Learning Health Care Systems. pp. S2–S44.

Taljaard, M., Weijer, C., Grimshaw, J.M., Eccles, M.P., Ottawa Ethics of Cluster Randomised Trials Consensus Group, 2013. The Ottawa Statement on the ethical design and conduct of cluster randomised trials: precis for researchers and research ethics committees. British Medical Journal 346, f2838.

Vayena, E., Tasioulas, J., 2013. The ethics of participant-led biomedical research. Nature Biotechnology 31, 786–787. http://dx.doi.org/10.1038/nbt.2692.

Weijer, C., Grimshaw, J.M., Taljaard, M., Binik, A., Boruch, R., Brehaut, J.C., Donner, A., Eccles, M.P., Gallo, A., McRae, A.D., Saginur, R., Zwarenstein, M., 2011. Ethical issues posed by cluster randomized trials in health research. Trials 12, 100. http://dx.doi.org/10.1186/1745-6215-12-100.

Therapist–Patient Relationship

Paul Crits-Christoph, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The therapist–patient relationship is accepted widely as a central aspect of psychotherapy. This article will examine the varying perspectives on the therapist–patient relationship in both historical writings on psychotherapy and in current research and theory. Methodological issues relevant to studies attempting to investigate the therapist–patient relationship will also be delineated, and speculations on future directions for research and theory will be offered. But, the first section will attempt to define the therapist–patient relationship.

Definition of the Therapist–Patient Relationship

A 'relationship' is something that emerges when two or more people interact. However, with the therapeutic relationship, and perhaps any human relationship, it becomes difficult to think about or observe this 'emergent' property of the interaction independent of the contributions of each of the participants. The thoughts, feelings, attitudes, and behaviors of each of the participants, as well as the technical interventions of the therapist, all potentially play a role in the formation and maintenance of the therapist–patient relationship and need to be considered in any conceptualization of this facet of psychotherapy. In the broadest sense, then, the therapist–patient relationship is the emergent entity that is produced by the expectations, attitudes, thoughts, feelings of the participants about each other, as well as their behaviors toward each other.

Despite this broad and elusive definition of the construct, we know a good therapist–patient relationship, and a bad one, when we see it. A good therapist–patient relationship is usually evident when patient and therapist are acting in a warm, caring, empathic, and productive way toward each other. A bad therapist–patient relationship is characterized by distance, hostility, lack of respect, manipulation, or poor capacity to work effectively together.

As mentioned, the positive, 'reality'-based aspects of the therapist–patient relationship are subsumed under the concept of the 'therapeutic alliance.' Bordin's (1979) definition of the therapeutic alliance in particular has had considerable influence in theoretical and research writings. The therapeutic alliance, according to Bordin (1979), consists of three elements: (1) agreement between patient and therapist on the goals of therapy, (2) agreement on the tasks of therapy, and (3) the affective bond between patient and therapist. Although the bond incorporates the emotional aspects of the patient–therapist attachment, agreement on goals and tasks refers to expectations or cognitive elements of the relationship. To understand how these aspects of the therapist–patient relationship have come to be accepted widely and how they apply to the full range of different psychotherapies requires a historical perspective, given in the next section.

History of Concepts of the Therapist–Patient Relationship

Modern views of the therapist–patient relationship can be traced directly to Freud's writings. Freud considered the therapist–patient relationship to be a unique laboratory for observing aspects of patient's personalities that may have relevance to the patient's symptoms or psychopathology. Moreover, he utilized the relationship to effect change in the patient (*see* Psychoanalysis in Clinical Psychology). Freud described several components of the therapist–patient relationship. In terms of the patient's contributions to the relationship, Freud (1958) described both positive and negative transference reactions. Transference refers to the expectations, desires, thoughts, and feelings that are 'transferred' from a previous relationship on to a new relationship (e.g., the therapist). Positive transference was further divided into the reality-based "friendly and affectionate aspects of the transference which are admissible to consciousness and which are the vehicle of success" (Freud, 1958) vs other positive feelings and perceptions (e.g., sexual feelings, strong dependency) that were not reality-based, originating instead from the patient's unconscious linking of the therapist with significant past relationships. Similarly, negative transference included negative feelings and perceptions that originated from past relationships and were not evident in the therapist–patient relationship. Finally, Freud acknowledged that distortions in the therapist–patient relationship might result from the therapist unconsciously linking the patient to significant people in the therapist's past ('countertransference') (*see* Transference in Classical, Contemporary, and Cultural Contexts). Although reality-based friendly feelings motivated a patient to stay in treatment and engage in therapeutic work, distorted positive and negative transference reactions served as the basis for the therapist's technical interventions. By illustrating these distortions to the patient, previously unconscious conflicts are brought to the conscious level and such insights lead to patient improvements.

A number of subsequent psychodynamic theorists expanded upon Freud's discussion of the therapeutic relationship. The term 'ego alliance' was used by Sterba (1934) to refer to the patient's capacity to work with the therapist in treatment,

alternating between experiencing and observing. Subsequently, Greenson (1965) described this capacity as the 'working alliance,' and theorized that it stemmed from the patient's mature ego functioning in conjunction with identification with the work orientation of the therapist. Zetzel (1956) discussed the patient's attachment to the therapist (labeling it the 'therapeutic alliance'), and postulated that it originated from positive aspects of the mother–child relationship.

Because the analysis of transference became the defining feature of many approaches to psychodynamically oriented psychotherapy, a relatively large clinical literature has developed on this aspect of the therapist–patient relationship. Esman (1990) has collected a sample of influential papers on transference that illustrates the perspectives of a range of psychoanalytic writers including Freud, Klein, Winnicott, Gill, Kohut, Kernberg, Lacan, and others.

Although there are historical disagreements about whether the positive aspects of the therapist–patient relationship represent patient distortions or realistic appraisals of the treatment situation, the writings of Bordin (1979) served to mark the beginning of consensus on the importance of the reality-based, collaborative therapist–patient relationship (the 'alliance'). This is not to say that other schools of psychotherapy had been silent on the importance of the therapist–patient relationship. In many ways, client-centered psychotherapy, developed by Rogers (1951), can be seen as an approach that relied exclusively on a positive therapist–patient relationship for inducing change. Unlike the psychodynamic emphasis on reactions of the patient to the therapist, Rogers' (1951) focus was on the therapist's contribution to the relationship. Rogers claimed that if certain 'facilitative conditions' (therapist empathy, genuineness, and unconditional positive regard) were provided to a patient, the patient's natural tendencies for growth and healing would be activated.

The behavioral school of psychotherapy, however, has often been characterized as minimizing the role of the therapist–patient relationship. This school of therapy describes behavior change in terms of principles of learning (i.e., classical or operant conditioning; social learning theory), with no focus on the context in which therapy takes place – i.e., the therapist–patient relationship. Practicing behavior therapists came to realize that this was an oversimplification. Goldfried and Davison (1976), for example, devoted an entire chapter of their book on clinical behavior therapy to the therapist–patient relationship, describing how behavior therapists can use the therapeutic relationship to directly sample aspects of the patient's problematic behavior. Moreover, a positive relationship was seen as crucial to facilitating favorable patient expectations for change and receptiveness to the behavioral approach, enlisting patient's active cooperation in treatment, and motivating the patient to attempt new behaviors outside of therapy.

Beck's cognitive therapy approach has long acknowledged that the therapist–patient relationship was important to successful outcome, but it was largely considered a 'given,' with only brief explicit attention to the relationship in the original writings on cognitive therapy as a treatment for depression (Beck et al., 1979). However, as cognitive therapy expanded its scope to the treatment of other disorders, especially personality disorders, the therapist–patient relationship has come into focus increasingly as a key element of the process of treatment (Newman, 1998).

Emphases in Current Theory

Different schools of psychotherapy have largely converged in their agreement that the therapist–patient relationship is of central importance to successful psychotherapy. Even psychodynamic approaches to treatment, which historically have focused attention on patient distortions about the therapist, now highlight the role of the reality-based aspects of the therapist–patient relationship. The consensus about the importance of the therapeutic alliance is evident in an edited volume by Safran and Muran (1998) presenting chapters examining the alliance in psychodynamic, cognitive, experiential, strategic, family, and group psychotherapies.

Safran and Muran (1998) have identified a number of principles on which there is basic agreement across therapeutic schools regarding what the therapist can do to facilitate a positive alliance. Among others, these include establishing a bond with the patient by conveying warmth, respect, and interest, outlining the tasks and goals of treatment at the beginning of therapy, focusing on realistic goals, maintaining a clear therapeutic focus, keeping a balance between being active (which helps maintain the treatment focus and the patient's involvement) vs receptive to patient needs and experiences, and addressing ruptures in the alliances when they occur.

Despite the agreement that has emerged regarding the importance of the alliance, there remains considerable disagreement about how a positive alliance results in more favorable treatment outcome. One view is that the alliance in itself is curative (e.g., Rogers, 1951; Luborsky, 1984). This perspective proposes that a positive relationship between therapist and patient provides a corrective emotional experience for the patient. Through their relationship with the therapist, the patient emotionally learns over time that other people are not abusive, controlling, manipulative, or otherwise destructive in relationships, as the patient has experienced others to be in the past. Thus, the patient's core expectations or beliefs about people change, and this improves relationships outside of therapy.

The second view on the role of the therapist–patient relationship posits that a good alliance sets the stage for psychotherapy techniques to have their impact (e.g., Beck et al., 1979). Whether the techniques are cognitive behavioral, such as homework assignments, or psychodynamic, such as interpretations of interpersonal patterns, the patient's receptivity to such therapist techniques will depend upon the extent to which the patient has a good bond with the therapist and basically agrees with the therapist on the goals and tasks of therapy. In this model, it is neither a good relationship nor good techniques that lead to better outcome, but rather the interaction of the two – a combination of good techniques in the context of a good relationship.

Both of the above models assume that the alliance is a relatively stable aspect of the therapist–patient relationship.

A contrasting perspective has been offered that views the alliance as a relatively fluid process with ongoing ruptures occurring frequently during treatment sessions (Safran et al., 2011). It is unclear at this point whether such moment-to-moment fluctuations in the alliance are the norm or whether they occur primarily in the treatments of certain types of patients (e.g., those with severe problems in interpersonal relationships).

Research on the Therapist–Patient Relationship

Beginning in the 1970s, a variety of studies reported on the development of scales for assessing the alliance and examined its relation to psychotherapy outcome. Several reliable self-report measures and observer ratings scales (applied to tapes of psychotherapy sessions) were constructed. Many of the scales are linked to the theoretical model described by Bordin (1979).

Horvath et al. (2011) have summarized the literature on the relation of the alliance to psychotherapy outcome. In a metaanalysis of 190 independent data sources covering more than 14 000 treatments, these authors reported that, on average, there was a small-to-moderate relationship between the quality of the alliance during treatment and eventual treatment outcome. The alliance has even been found to be associated with positive outcome in studies of psychosocial treatments for children/adolescents (e.g., Kazdin et al., 2005) and elderly adults (Gaston et al., 1998). Moreover, the alliance was found by Horvath et al. (2011) to be related to treatment outcome across diverse psychotherapies. The size of the statistical relationship, however, suggests that the alliance is not the only factor in determining psychotherapy outcome. In brief, while the therapist–patient relationship appears critical, there is no reason to suggest that a good therapist–patient relationship is the only curative factor in psychotherapy.

Several studies have examined the common clinical hypotheses that good psychotherapeutic technique has its greatest impact in the context of a positive alliance, i.e., the alliance and technique interact to produce better outcomes. No consistent findings, however, have emerged across studies, with some investigations reporting evidence for an interaction (Owen and Hilsenroth, 2011), other studies yielding inconclusive but suggestive findings with small samples (Gaston et al., 1998), and some work suggesting that any interaction between techniques and alliance depends upon characteristics of the patient, particularly the patient's typical level of interpersonal functioning (Høglend et al., 2011).

Methodological Issues in Research

A host of methodological issues need to be considered when conducting research on the relation of the alliance to treatment outcome. Perhaps the most crucial issue relates to whether a positive alliance is actually a causal factor in producing better treatment outcomes. The reverse causal direction needs to be entertained as well: as symptoms improve, patients are more likely to report a positive attitude about the therapist and treatment. Thus, significant correlations between the alliance (measured during treatment) and outcome may be spurious – early improvement in symptom lead to both a positive alliance and good eventual treatment outcome. This potential confounding between the alliance and prior improvement has now been evaluated in 11 different studies (reviewed in Crits-Christoph et al., 2013). On average across studies, there is evidence that the alliance remains moderately associated with treatment outcome even when prior improvement was controlled; however, substantial variability in this effect across studies is apparent. Thus, this important issue of the direction of causality remains to be addressed in future research.

Another methodological issue of importance is the fact that psychotherapy is a dyad consisting of a patient and a therapist, with either (or both) potentially contributing to a particular process. It may be, for example, that the connection between the alliance and treatment outcome has more to do with differences among therapists in their abilities to form positive relationship with patients and less to do with differences among patients in their capacity to form a positive patient–therapist relationship. Modern statistical techniques such as multilevel modeling have permitted studies to decompose the alliance–outcome association into the component that is related to differences among therapists and the component that is related to differences among patients. To date, the evidence indicates that it is indeed the differences between therapists, not patients, that are most responsible for the association between the alliance and outcome (Baldwin et al., 2007; Crits-Christoph et al., 2009). A further consideration is that most research linking the alliance to treatment outcome has examined the alliance in only one or two sessions of psychotherapy (usually early in treatment sessions). But if the alliance varies from session to session depending on ruptures in the alliance triggered by events occurring inside, or outside, the therapeutic relationship, it may be necessary to measure the alliance over many sessions to characterize the patient–therapist relationship as generally 'good' or 'poor.' Indeed, a recent study demonstrated that the relation between the alliance and treatment outcome increases substantially as the alliance is measured in progressively more sessions, with at least four sessions needed to adequately capture the general quality of the alliance for a psychotherapy treatment as a whole (Crits-Christoph et al., 2011).

Future Directions

The almost universal acceptance of the therapeutic alliance as an essential part of psychotherapy points the way to some obvious future directions for scientific study. If a positive therapist–patient relationship is key to successful outcome, and the therapist contribution to the alliance is what explains the connection between the alliance and outcome, can therapists be trained to maximize the quality of the alliance? If not, is this because good therapists are 'born' not 'made'? Resolution of this issue could have profound implications for the training of psychotherapists. If therapists cannot be taught how to establish and maintain a good alliance, perhaps greater effort should

be placed on the selection process for training programs in the disciplines (psychology, psychiatry, social work) which conduct psychotherapy. Priority would be given to applicants who have innate interpersonal skills (e.g., empathy) that are associated with positive therapist–patient relationships. If, however, therapists can be trained to maximize the quality of the alliance, such training should become a core element of all training programs, regardless of which treatment modalities are emphasized.

Rather than focusing on the broad question of whether the alliance predicts treatment outcome, future research is likely to continue to explore more refined and specific questions that attempt to unravel the contexts under which a positive alliance is especially important. Larger and more definitive studies, following up on intriguing preliminary findings with smaller samples, can identify the treatment variables and patient dispositional variables that interact with the therapeutic alliance to produce optimal outcome. Research on human relationships, emerging from basic studies in social cognition and interpersonal processes, is likely to provide direction for understanding the nuances of the therapist–patient relationship. Similarly, as intensive empirical examination of the therapist–patient relationship increases our knowledge about the influence of this relationship, we may find applications of these findings to other helping relationships in medicine and elsewhere in society.

See also: Agreeableness; Cognitive Behavioral Therapy; Coping, Personality And; Core Self-evaluations; Emotion, Perception and Expression of; Emotional Regulation; Emotions and Aging; Epistemological Beliefs and Motivation; Expectancy-Value-Cost Model of Motivation; Gender Differences in Personality and Social Behavior; Health Self-Regulation, Motivational and Volitional Aspects of; Interpersonal Psychotherapy; Moral Reasoning in Psychology; Motivation in Australian Aboriginal Populations; Motivation in Clinical Interventions; Person Perception, Accuracy of; Personality Assessment: Overview; Personality Theory and Psychopathology; Psychotherapy Integration; Reflexivity; Self-Efficacy; Selfconscious Emotions, Psychology of; Temperament and Motivation; Temperament; Values, Social Psychology of.

Bibliography

Baldwin, S.A., Wampold, B.E., Imel, Z.E., 2007. Untangling the alliance-outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. Journal of Consulting and Clinical Psychology 75, 842–852.

Beck, A.T., Rush, A.J., Shaw, B.F., Emery, G., 1979. Cognitive Therapy of Depression. Guilford Press, New York.

Bordin, E.S., 1979. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research, and Practice 16, 252–260.

Crits-Christoph, P., Gallop, R., Temes, C.M., Woody, G., Ball, S.A., Martino, S., Carroll, K.M., 2009. The alliance in motivational enhancement therapy and counseling as usual for substance use problems. Journal of Consulting and Clinical Psychology 77, 1125–1135.

Crits-Christoph, P., Connolly Gibbons, M.B., Hamilton, J., Ring-Kurtz, S., Gallop, R., 2011. The dependability of alliance assessments: the alliance-outcome correlation is larger than you might think. Journal of Consulting and Clinical Psychology 79, 267–278.

Crits-Christoph, P., Connolly Gibbons, M.B., Mukherjee, D., 2013. Psychotherapy process-outcome research. In: Lambert, M. (Ed.), Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, sixth ed. John Wiley & Sons, Inc., Hoboken, NJ, pp. 298–340.

Esman, A.H. (Ed.), 1990. Essential Papers on Transference. New York University Press, New York.

Freud, S., 1958. The dynamics of transference. In: Strachey, J. (Ed.), The Standard Edition of the Complete Psychological Works of Sigmund Freud (Original Work Published 1912), vol. 12. Hogarth Press, London.

Gaston, L., Thompson, L., Gallagher, D., Cournoyer, L., Gagnon, R., 1998. Alliance, technique, and their interaction in predicting outcome of behavioral, cognitive, and brief dynamic therapy. Psychotherapy Research 8, 190–209.

Goldfried, M.R., Davison, G.C., 1976. Clinical Behavior Therapy. Holt, Rinehart, and Winston, New York.

Greenson, R.R., 1965. The working alliance and the transference neuroses. Psychoanalytic Quarterly 34, 155–181.

Høglend, P., Hersoug, A.G., Bøgwald, K.P., Amlo, S., Marble, A., Sørbye, Ø., Røssberg, J.I., Ulberg, R., Gabbard, G.O., Crits-Christoph, P., 2011. Effects of transference work in the context of therapeutic alliance and quality of object relations. Journal of Consulting and Clinical Psychology 79, 697–706.

Horvath, A.O., Del Re, A.C., Flückiger, C., Symonds, D., 2011. Alliance in individual psychotherapy. Psychotherapy 48, 9–16.

Kazdin, A.E., Marciano, P.L., Whitley, M.K., 2005. The therapeutic alliance in cognitive–behavioral treatment of children referred for oppositional, aggressive, and antisocial behavior. Journal of Consulting and Clinical Psychology 73, 726–730.

Luborsky, L., 1984. Principles of Psychoanalytic Psychotherapy: A Manual for Supportive–Expressive Treatment. Basic Books, New York.

Newman, C., 1998. The therapeutic relationship and alliance in short-term cognitive therapy. In: Safran, J.D., Muran, J.C. (Eds.), The Therapeutic Alliance in Brief Psychotherapy. American Psychological Association Press, Washington, D.C.

Owen, J., Hilsenroth, M.J., 2011. Interaction between alliance and technique in predicting patient outcome during psychodynamic psychotherapy. The Journal of Nervous and Mental Disease 199, 384–389.

Rogers, C.R., 1951. Client-Centered Therapy. Houghton-Mifflin, Boston.

Safran, J.D., Muran, J.C. (Eds.), 1998. The Therapeutic Alliance in Brief Psychotherapy. American Psychological Association Press, Washington, D.C.

Safran, J.D., Muran, J.C., Eubanks-Carter, C., 2011. Repairing alliance ruptures. Psychotherapy 48, 80–87.

Sterba, R., 1934. The fate of the ego in analytic therapy. International Journal of Psychoanalysis 15, 117–125.

Zetzel, E.R., 1956. Current concepts of transference. International Journal of Psychoanalysis 37, 369–376.

Thick Description: Methodology

Tanya M Luhrmann, Stanford University, Stanford, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Thick description' is the term that Geertz used to describe ethnography in one of the most famous and influential anthropology texts in the second half of the twentieth century, *The Interpretation of Cultures* (1973). It has come to imply two different things: on the one hand, good ethnography – what all anthropologists have done and do when they write up the results of solid, careful ethnography – and, on the other hand, what Geertz understood to be the intellectual entailments of this particular methodology.

'Thick description' is the term that Geertz used to describe ethnography in one of the most famous and influential anthropology texts in the second half of the twentieth century, *The Interpretation of Cultures* (1973). The phrase was the title of the first essay in the volume of the collected essays, and Geertz explained at the time that he had written it to state his theoretical position as generally as he could. As a result, thick description has come to imply two different things: on the one hand, good ethnography – what all anthropologists have done and do when they write up the results of solid, careful ethnography – and, on the other hand, what Geertz understood to be the intellectual entailments of this particular methodology.

As Methodology

Geertz borrowed the phrase 'thick description' from the philosopher Ryle (1949), who used it to contrast two different ways of describing the same piece of behavior. Ryle's example was a boy's twitch and his wink. Both could be described as 'rapidly contracting his right eyelids.' That kind of description Ryle called 'thin.' He used the term 'thick description' to characterize the intentional, communicative, interpretative, meaning of the behavior: why it was done, how it was read, and using which social codes. That, Geertz argued, was the object of ethnography, "a stratified hierarchy of meaningful structures in terms of which twitches, winks, fake-winks, parody, rehearsals of parodies are produced, perceived and interpreted, and without which they would not … in fact exist, no matter what anyone did or didn't do with his eyelids" (1973, p. 7).

Ethnography is the written description of long-term participant observation. Professional anthropologists now typically expect that a person doing serious fieldwork (of the sort required for an anthropological dissertation, for example) will spend at least a year living in the 'field,' in the social setting that they have chosen to study. They expect that fieldworkers will speak the dominant language spoken in that setting, and they expect that fieldworkers will learn about that world not only or even primarily by interviewing its members, but also by learning how to participate as a member themselves: how to joke, to be in relationship, to be an appropriate person of their type (e.g., a woman in her 20s) in this social setting. If you are studying Londoners who practice witchcraft and magic, you learn how to write rituals, to read tarot cards, and to feel the magic running in your blood. If you study Malagasy villagers who plant rice and perform cattle sacrifice, you work in the rice paddies and offer up a cow. Particularly in recent decades, professional anthropologists have decried their pith-helmeted, informant-paying forebears. They are likely to evaluate the student fieldworker by asking whether the student hung back from human contact or really got to know the people he or she worked among and came to be accepted by them. The point of this emphasis is to make sure that the fieldworker has grasped the realities and the rules of engagement of the world he or she has come to study. As Stephen Hugh-Jones once remarked, there is little point of having an abstract debate about what it means when the natives say that they 'see' their gods during rituals if you have not participated in those rituals and learned that they involve hallucinogens.

Professional ethnographers use at least three criteria to help them judge the reliability and validity of the ethnographic report.

First, they look at the richness of observational detail. Ethnography can be at risk of being dismissed as hearsay or anecdotal. But when the observational detail, reliably reported, is detailed enough – 'thick' enough – these objections no longer seem plausible. Ethnographic monographs are often quite lengthy for that reason: the ethnographer supports an argument on a foundation of thousands of specific observations about thousands of specific events.

Second, they look for external corroboration of the ethnography. In part, this is provided by prepublication peer review by experts in the specific area of the world the ethnography concerns. However, the primary responsibility for external corroboration lies with the ethnographer, who is expected to refer to published materials by field subjects (when relevant) and by other ethnographers to support his or her conclusions. At this point in world history, there are few societies remote enough to be completely unknown.

Third, they look for evidence that the ethnographer knows the ethnographic area well. As in any discipline, the conventions to judge competence are somewhat arbitrary, but still powerful. A year of full-time fieldwork is considered minimal; more years provide more reassurance about the quality of the data. Ethnographies thus present the specific dates and conditions of the fieldwork. To be able to converse in the language is minimal; to command idioms, whether in the spoken language itself or in the metaphorical language of local practice, again provides more reassurance. Ethnographies should display mastery of the local idioms of practice (the intricacies of the sheep trade, for instance). To be accepted in the target social world well enough to have conversational and observation

access to field subjects is also minimal, and so ethnographies typically describe how that access was acquired and maintained; sometimes these accounts are called 'arrival stories.'

There is, however, a hard-to-specify distinction between the competent ethnography and the really good one. Ethnographies do much more than give an account of who did what and when. An ethnographer can sit in a courtyard for one full day and describe exactly what he sees, and his observations will fill a monograph, and the monograph will be unreadable and unintelligible. By contrast, good ethnographies are accounts of a social world that make sense of that world to those who are alien to it. "What the ethnographer is in fact faced with," Geertz explained, "is a multiplicity of complex conceptual structures, many of them superimposed upon or knitted into one another, which are at once strange, irregular, and inexplicit, and which he must contrive somehow first to grasp and then to render" (1973, p. 10). Good ethnography describes what is really important to understand about this world, both to the participants and to its observers; it conveys what people mean when they say, 'you really get it.' When anthropologists use the phrase 'thick description' to refer to the ethnographic method, they mean to imply that the anthropologist does serious, engaged fieldwork; that he really grasps the social process of the world being studied; and that he writes an ethnography so detailed and so observant that it is utterly persuasive.

As an Epistemology

"If you want to understand what a science is," Geertz remarks early on in his famous essay, "you should look in the first instance not at its theories or its findings, and certainly not at what its apologists say about it; you should look at what the practitioners of it do" (1973, p. 5). Ethnographers immerse themselves in another society for long periods of time, and collect far more bits of information than they could ever report or code or convey. Then they write books that attempt to describe the foreignness of that other world in ways that make it comprehensible to their readers.

As a consequence, Geertz argued, ethnographies are by their nature interpretations. They are not hypothesis driven and predictive. This does not mean that they are unconstrained and undisciplined. "That is like saying that as a perfectly aseptic environment is impossible, one might as well conduct surgery in a sewer" (1973, p. 30). But they are not normal science – normal social science – in the style of psychological experiments and economics theories. The anthropologist, from this perspective, is more like a literary critic who writes about a novel that anyone could in principle read. They may write about the novel in many ways – explain it to those who do not understand, praise or critique it, argue that other readers have misunderstood it, suggest that it captures certain human experiences that help its readers to see those experiences differently from before, and insist that one way of thinking about it is more illuminating than another. But whatever they do, that novel exists as an art form independent of their presence, and their job is to help others to appreciate it more deeply. The claim to attention of an ethnographic account does not rest on its author's ability to capture primitive facts in faraway places and carry them home like a mask or a carving, but on the degree to which he is able to clarify what goes on in these places, to reduce the puzzlement – what manner of men are these? – to which unfamiliar acts emerging out of unknown backgrounds naturally arise (1973, p. 16).

What can be interpreted, however, is only what can be 'read.' The power in the comparison of ethnography and literary criticism lay not only in the reasonableness of the general claim – fieldwork is indeed different from laboratory science – but in the pragmatic usefulness of its central metaphor. Following Paul Ricouer, Geertz not only compared the anthropologist to a literary critic, but described culture and social interaction as a text – the metaphorical 'literature' that the anthropological critic interpreted. More loosely, he described culture as discourse and as language. He argued that culture – "this acted document" (1973, p. 10) – was public because meaning was public. As an ethnographer, a watcher of others, you have no access to people's minds. What you can see is that they communicate, and you can interpret what they communicate, and what you are after is in effect the language or discourse that they are using, not the ideas they have in their mind. Geertz used Wittgenstein's private language argument to support the claim that culture, like language, was a public, social phenomenon, but his primary concern at the time was that you did not need to understand psychological process in order to be a good ethnographer. Understanding the mind was a psychologist's job; understanding culture was the anthropologist's.

This central metaphor, of culture as a text, gave the ethnographer something specific to do: to identify symbols that somehow represented the particular social process of that group, and to then explain their meaning. In a much quoted definition Geertz stated, "Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning" (1973, p. 5). Geertz went on to argue that the kinds of things that anthropologists interpret – rituals, myths, cockfights – are themselves interpretations of the society in which they were found, and so the anthropologist is producing second- and third-order interpretations for a different purpose. "The culture of a people is an ensemble of texts, themselves ensembles, which the anthropologist strains to read over the shoulders of those to whom they properly belong" (1973, p. 452).

Because (from this perspective) culture is a semiotic system, so that all representation is in effect interdependent, and because the social behaviors that despite everything anthropologists tended to think of as cultural – rituals, stories, songs; the noninstrumental side of life – are also understood as interpretations of their own society, the anthropologist could focus on some practice that on the surface seemed trivial or irrational, and use that practice to understand and reflect upon the nature of the social whole. Geertz did this most famously in his essay on the Balinese cockfight. The cockfight is a seeming unimportant and irrational piece of Balinese culture in which grown men bet money they cannot afford to watch birds kick each other to death in illegal matches. Geertz suggested that Balinese men were gripped by their cockfights because the matches flung away the restrained, polished, social hierarchy

that otherwise defined Balinese daily life, and through his discussion, this small social practice became the prism through which you came to understand fundamental features of Balinese society.

In the decade after these two essays on thick description and the cockfight were published, Geertz became one of the leading figures of sociocultural anthropology, if not its dominant presence, and his work precipitated a paradigm shift in the field and influenced other fields as well, most notably history. The new approach was called 'symbolic' anthropology and contrasted both with functionalism (a primarily British approach which sought to understand how all elements of social practice contributed to the effective functioning of that social world) and structuralism (an initially French approach that sought to identify the formal logic underlying cultural representation). David Schneider and the midcareer Marshal Sahlins were also associated with this new emphasis on the way cultural practices represented social process, and symbolic anthropology tended to be associated with the department of anthropology at the University of Chicago, where Geertz had been before he left for the Institute of Advanced Study at Princeton. Rosaldo's *Knowledge and Passion* (1980), an ethnography that explores the society of the Filipino Ilongot headhunters through a dominant emotion term, is a good representative of symbolic anthropology.

Now, however, the phrase that is associated with thick description is 'interpretive anthropology.' That phrase tends to mean that the ethnographer values serious (in depth, detailed) fieldwork and understands the ethnography as an attempt to make the foreign world comprehensible to the reader. The phrase also often implies that the anthropologist is self-consciously aware of the literary nature of the ethnographic enterprise: that ethnography is an act of writing, and that successful anthropologists are first and foremost writers. Finally, the phrase often (but not always) implies that this writerly, interpretive act is the primary goal of ethnography, and that anthropology cannot be a hypothesis-testing science, because its methods will not allow it to be so. As a result, the phrase is sometimes used to imply more about what ethnography is not, than about what it is, and the phrase can be used as a battle cry by those who wish to see anthropology associated with the humanities. Those battles, however, take the phrase beyond the domain that Geertz first mapped out.

See also: Ethnography; Field Observational Research in Anthropology and Sociology; Interpretation in Anthropology; Participant Observation.

Bibliography

Geertz, C., 1973. The Interpretation of Cultures: Selected Essays. Basic Books, New York.

Rosaldo, M., 1980. Knowledge and Passion: Ilongot Notions of Self and Social Life. Cambridge University Press, New York.

Ryle, G., 1949. The Concept of Mind. Hutchinson, London.

Think Tanks

Diane Stone, Murdoch University, WA, Australia; and The University of Warwick, Coventry, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Think tanks first proliferated in Anglo-American countries during the mid-twentieth century but by the twenty-first century had become prevalent throughout the world. These organizations vary dramatically in size, legal status, policy concerns, and organizational structure but share a mission to inform policy, educate political elites, and shape decisions. The political influence and social impact of policy institutes waxes and wanes across time and policy issue, as well as from one political culture to the next. They have become key informants to policy making at local, national, and transnational levels of governance.

With the label 'brain trust' or 'policy institute,' the idea of what constitutes a 'think tank' has changed dramatically since the last century. The term 'think tank' originated in the United States during World War II. It was used to describe secure and 'sealed environments' for expert strategists preoccupied with military planning, most notably the RAND Corporation. By the 1960s, the term was entrenched in the lexicon and was being applied to independent policy research institutes throughout the English-speaking world.

Think Tank Diversity and the Dilemmas of Definition

During the twentieth century, social science definitions of think tanks were shaped by Anglo-American experience. The proliferation of think tanks was most pronounced in the United States and remains so. In this country, generous foundation and corporate support, as well as extensive government contracting out of research boosted their development. The constitutional architecture, federal structure, and relatively weak party system provided a conducive environment for well over 1500 think tanks to engage in policy debates and interact with government (Weidenbaum, 2011). By contrast, the stronger party systems, corporatist modes of decision making, strong, and relatively closed bureaucracies, and sometimes weak philanthropic sectors are considered to have dampened think tank development in the parliamentary systems of Great Britain, its former dominions of Canada, Australia, and New Zealand, and most of Western Europe (Stone, 1996). Nevertheless, think tanks were present in these systems throughout the last century although in considerably smaller number and size than their US counterparts.

The degree of legal, financial, and intellectual autonomy and independence differs substantially from one nation to another; a sociopolitical context that inevitably shapes the channels for, and modalities of, communicating research and analysis on contemporary issues. Think tanks are (usually) relatively autonomous from government, political parties, and pressure groups as well as business although they are often in resource-dependent relationships with these patrons. Nevertheless, institutes attempt to maintain their research freedom and usually claim not to be beholden to specific interests. Think tanks seek policy influence primarily through intellectual argument and policy analysis rather than direct lobbying; they are concerned with the development and application of ideas and concepts that underpin policy, although not all their research and activity need be policy relevant. Think tanks collect, synthesize, and create a range of information products, often directed toward a political or bureaucratic audience, but also for the benefit of the media, civil society groups, and the general public. They are concerned with knowledge creation although political communication and effective knowledge utilization may be their primary purpose.

Think tanks defy exact definition. In both the scholarly and professional field of governance, think tanks can be defined as policy research centers engaged in policy analysis and advocacy that is directed toward government but also undertaken for international organizations, the business and nonprofit sectors. Think tanks vary considerably in size, legal form, policy ambit, longevity, organizational structure, standards of inquiry, and political significance. Not only is there considerable scholarly difference over how to identify these organizations, but the directors and senior scholars of these organizations often make subtle distinctions between 'research institute' and 'think tank.' Such disputes usually revolve around the role of advocacy on the one hand, and organizational capacity for high-quality policy research on the other, with think tanks deemed to do the former and research institutes the latter. Often, this distinction is false and points to a deeper ontological distinction made between 'knowledge and power.' Generally, think tanks are represented as a bridging institution, linking the scholarly world and the separate political domain by translating jargon ridden, densely technical or highly theoretical research into policy palatable recommendations. As a UNDP (2003: p. 6) publication put it: "They are the bridge between knowledge and power in modern democracies."

Some organizations claim to adopt a 'scientific' or technical approach to social and economic problems; others are overtly partisan or ideologically motivated. While some institutes are routinely engaged in intellectual brokerage and the marketing of ideas, whether in simplified policy-relevant form or in sound bites for the media, others are more academic. Many institutes are discipline-based – economic policy think tanks, foreign policy institutes, and social policy units. Specialized environmental think tanks, regionally focused operations, and those that reflect the communal interests of ethnic groups appeared

from the 1990s. While most display a high level of social scientific expertise or familiarity with governmental structures and processes, there is considerable diversity in style and output of think tanks. The intellectually engaged 'ink tank' can be poised against the action-oriented 'think-and-do tank': that is, there are significant differences between think tanks that are scholarly in focus and geared toward publication of books and reports, and think tanks that are more activist and engage in grassroots activity. However, the impact of information technology has leveled the playing field somewhat, as all think tanks are heavily engaged in sophisticated Web site development as well as forms of social networking for targeted dissemination of output. As a consequence, a sound reputation for quality research and analysis, and credibility in political and policy communities, while intangible, has increasingly become the standard of success.

The Western view that a think tank requires independence and autonomy from the state, corporate, or other interests in order to be free-thinking and proffer dispassionate or nonpartisan policy advice does not accord with experiences in other cultures. In many nation states, the line between policy intellectuals and the state is blurred to such an extent that to talk of independence as a defining characteristic of think tanks makes little sense. Many organizations now called think tanks operate within government; this is evident in China (Zhu, 2011). Elsewhere, some institutes have been founded or incubated in government and subsequently 'spun off' as independent entities; it is not unusual to find institutes with political patrons or formal links to political parties. Many of the German political foundations, for instance, have been established by political parties or have strong ties to the Länder and some German institutes have grown to such a large size that they rival some of the major American institutes. In other circumstances, research institutes are attached to corporations, as is evident in Japan, South Korea, and Taiwan. Notwithstanding funding dependence or their political affiliation, high-quality research and analysis along with critical advice is feasible.

Accordingly, financial independence can be construed as developing an endowment or having numerous sponsors and a diverse funding base, so that an organization is not dependent on any one benefactor. Scholarly independence is reliant upon certain norms and practices within an institute: the processes of peer review and a commitment to open inquiry rather than directed research along with the willingness of directors to protect and champion such norms. Legal or administrative independence comes from constitution as a private organization. Administrative and financial independence may be a facade under authoritarian or illiberal regimes where censorship and control prevail. Alternatively, some institutes may exhibit a high degree of research freedom despite being found within the state architecture. Appreciation of the political culture of a nation is essential to interpreting think tank independence and intellectual leverage.

A further problem of definition is that it is often difficult to establish a precise distinction between think tanks and other organizations with similar objectives, for example, operating foundations. Think tanks could also be regarded as more intellectual variants of pressure groups or nongovernmental organizations (NGOs): Many now base their arguments on relatively sophisticated research that is sometimes conducted in-house. Nevertheless, there are important differences: think tanks do not engage in public demonstrations and they tend to address multiple policy areas (notwithstanding specialization) rather than focus on a single issue as do many pressure groups. There are similarities with consultancy firms, financial institutions, or legal offices that provide contract research and analysis, although these organizations engage in a wider range of activity than research; yet many think tanks only survive through their consultancy services and contract research. Some think tanks have been described as 'universities without students' (Weaver, 1989: p. 564). While the relationship between think tanks and universities has been close in most political systems, the vast majority of think tanks are not degree-granting institutions. There are a few exceptions, however, notably RAND in the United States and FLACSO in Latin America. In terms of future think tank development, the blurring of boundaries and the overlap of objectives and activities between institutes and other organizations in our 'knowledge societies' may mean that these organizations lose organizational distinctiveness and their monopoly on the claim that they 'bridge' science and the state.

Worldwide Expansion and Transnational Networking

In the last two decades of the twentieth century, the number of think tanks proliferated. Estimates vary along the lines of definition, but 5000 think tanks worldwide is not an unrealistic figure (McGann and Sabatini, 2011). Countries where think tanks were already present such as the United States, Britain, Sweden, Canada, Japan, Austria, and Germany witnessed further organizational growth. Increased competition in the think tank industry has often encouraged policy advocacy and the politicization of institutes, most particularly in the United States. The Heritage Foundation in Washington DC is usually cited as the exemplar but similar advocacy styles can be observed elsewhere. Democratic consolidation and economic development in Latin America, the Indian subcontinent, and Asia provided fertile conditions for think tank development, particularly as higher education systems expanded in these regions creating new generations of educated professionals and opinion leaders. The demise of the Soviet Union, the color revolutions, and the Arab Spring opened new political spaces for donor agencies and policy entrepreneurs to support think tanks as both a driver and a symptom of civil society capacity building (Stone and Maxwell, 2005).

On the one hand, the global think tank boom has been prompted by philanthropic foundations, corporations, and other nonstate actors demanding high-quality research, policy analysis, and ideological argumentation as advisory input within their organizations. On the other hand, growth has been propelled by grants, gifts, and other funding from governments and international organizations seeking to extend policy-analytic capacities or promote evidence-based policy formulation in developing and transition countries. Well-known programs include the 'Think Tank Initiative' supported by Hewlett Packard and administered by Canada's International Development Research Center, as well as the Open Society Foundation's Think Tank Fund that supports new institutes

across Central and South Eastern Europe and the southern Caucasus.

As a corollary, donor organizations often regard think tanks as NGOs and as entrepreneurs for the promotion of civil society development. Similarly, think tanks are often viewed by donors as an agent of democratization and a vehicle for public participation in policy debate. However, think tanks are an elite and relatively technocratic organizational expression of civil society when constituted as NGOs or other private forms of organization. The legitimacy for their input into policy making is claimed on the basis of their expert status and epistemic authority. This basis for policy engagement is in tension with rationales of think tank activity being a mode of democratic deliberation or representing a pluralistic counterweight to the state monopoly of decision making.

The think tank model has been exported around the world. The Anglo-American term and institutional template of 'think tank,' with all its cultural connotations, has penetrated the political systems of developing and transition countries. While the term is ubiquitous, its translation into organizational formats in quite different sociopolitical environments has been varied, taking hybrid form in countries as diverse as Vietnam or Croatia. After two and a half decades of transition, think tanks have occasionally come to be seen as an unwelcome foreign influence in some of the states of Southeastern Europe. Donor-funded think tanks sometimes supported uncritically many of the market liberal policy priorities emerging from the power centers of Brussels and Washington DC. Along with the high social costs of such programs, and in the wake of the Euro-crisis, the analytical support seen to be given by think tanks is one possible contributing factor to the rise of political populism in the region and increasing disdain for expert bodies. Even so, through other parts of Central and Eastern Europe as well as the former Soviet Union, think tank initiation remains a popular start-up option for young professionals.

International organizations such as the United Nations agencies, the World Bank, the International Monetary Fund, and the World Trade Organization have drawn think tanks into their ambit via conferences, joint research, and fellowship programs (for example, UNDP, 2003). The World Bank launched the Global Development Network (GDN) to promote think tanks in developing countries (Plehwe, 2007). The European Community provides yet another institutional forum for think tank activity with the emergence of EU-wide think tanks such as Notre Europe or the Centre for European Policy Studies seeking to transcend specific national identities with research and policy proposals for regional integration. Furthermore, with ongoing advances in information and communication technology, policy research disconnected from specific organizational settings has become increasingly feasible and fashionable. It has also made international research exchange and collaboration between think tanks commonplace.

Think tank transnationalization has been fuelled by the growth of global knowledge networks and this development can be seen as another manifestation of the 'bridging' function of think tanks (Stone and Maxwell, 2005). These networks have promoted transnational policy analysis and scientific expertise. They create the overlapping personal and communications infrastructure that allows the fast transfer of new ideas and policy learning across borders.

Networking ranges from the very informal, socializing or the 'thin' element of a virtual network through to formal international associations with a secretariat and large membership. Global and regional think tank networks like GDN and PASOS (a network of 30+ think tanks from the former Soviet Union and Central and Eastern Europe) have become extensive. International networks have coalesced around common areas of interest and policy themes as well as around ideology. For example, the Institute for European Environmental Policy has offices in London and Brussels and a network of partners throughout Europe. Tied by ideology, the Atlas Economic Research Foundation connects a global network of more than 400 free-market organizations in over 80 countries to the ideas and resources needed to advance the cause of liberty, providing start-up funds and technical assistance. Since 1997, the Japan Center for International Exchange has convened 'Global ThinkNet' meetings to promote policy-oriented dialogues within Asia and between Asia and other regions. Additionally, think tanks have become transnational by opening branch offices in other countries, as does RAND, the Carnegie Endowment for International Peace and the Urban Institute.

Networks have also become a form of governance. The typical target of the think tank has been government. In an era of privatization, contracting-out, and the entrenched logic of the 'new public management,' the growing importance of the private and voluntary sectors in the delivery of public goods and services has de-centered policy dialogue and service delivery away from the state. This is more pronounced above the nation state where international organizations, corporations, and large international NGOs promote global policy programs and public–private partnerships to deal with transnational policy problems. These governance networks above the nation state tend to be more weakly institutionalized than at the national level where sovereign power is clearly defined in administrative institutions and state agencies. The more fragmented and fluid transnational policy space provides numerous opportunities for think tanks individually and in coalition to proffer their policy analysis services.

Policy Impact and Influence

Following this extensive organizational growth, scholarly interest has expanded. Where discussion of think tanks was once limited almost exclusively to Anglo-American systems (Abelson, 2006; Weaver, 1989; Stone, 1996), interest has expanded to comparative political analysis of think tanks in other OECD nations (for example, Boucher et al., 2004) and beyond to developing or transition countries (inter alia Stone and Maxwell, 2005; UNDP, 2003). Wider disciplinary interest has emerged to include international relations and security studies specialists as well as a greater array of anthropologists, economists, and sociologists writing in this field.

Analysis of think tanks has diverged into two branches. Early research focused on the organizational form where writers were interested in explaining why and how think tanks have emerged and why some think tanks are influential. They distinguished independent public policy institutes from academic research centers, government research units, or

lobbyists. They focused on the organizational ingredients of what makes a think tank successful, how think tanks are managed, who funds them, who quotes them, and whom they try to influence (inter alia Struyk, 2007; Weaver, 1989; Weidenbaum, 2011). Initially, the growth and increasing diversity was represented as sign of a healthy pluralist democracy where think tanks contributed to a 'marketplace of ideas.' However, as competition for resources and policy attention became more acute in the 1980s, a new breed of ideologically aggressive think tank emerged in the United States, and emulated elsewhere, whereby policy partisanship lead observers to speak of a 'battlefield of ideas' with winners and losers in policy debate.

Writers in the second branch analyzed think tanks as a vehicle for broader questions about the policy process and the role of ideas and expertise in governance (inter alia Abelson, 2006; Campbell and Pedersen, 2011; the contributions in Stone and Maxwell, 2005). They employed network concepts such as 'policy community,' 'advocacy coalitions,' and 'epistemic communities' to explain why ideas matter. Others addressed think tank roles from a neo-Marxist or Gramscian perspective (Pautz, 2011; Plehwe, 2007) to argue that policy research is free neither of ideological content nor of material (class) interests. Yet others have utilized Bourdieu's field analysis (Medvetz, 2012) or adopted discourse frameworks (Fischer, 2003) to unveil the argumentative power of research and analysis and the capacity of ideas to shape public and policy maker perceptions of reality.

These studies cut to a central issue in the study of think tanks: Do think tanks exert policy influence, and if so, how? Notwithstanding their extensive growth, think tanks do not enjoy automatic political access or unhindered bureaucratic routes to decision makers. Think tanks need to have some kind of engagement with government if they are to have policy relevance; however, their desire to preserve intellectual autonomy means that most think tanks need to strike a delicate balance between dependence on government and political parties as clients, and isolation from such audiences that comes from policy research that is either too critical and uncompromising or too esoteric and irrelevant to pressing political timetables. Conducting policy-relevant research and attempting to broker it to decision makers does not equate with policy influence. This conundrum has been the major methodological problem in the study of think tanks.

Proof of influence is elusive. There is no clear causal nexus between policy research or an idea espoused by an institute, and political decisions and policy change. It is a rare occurrence that a policy outcome can be attributed to one organization, as there are numerous intermediary forces: political parties, interest groups, and the media also shape the content and reception of ideas. Furthermore, the cognitive order of politicians and bureaucrats cannot be accessed by 'think tank' researchers, and interviewing has its own problems of faded memories and selective recall in establishing the veracity of perceptions of think tank policy input. This hiatus is compounded as the political stature of many institutes waxes and wanes over time. Moreover, the criteria by which influence is assessed vary considerably.

Nevertheless, there are many countries where overstretched bureaucracies and limited capacities for in-house governmental policy analysis provide the opportunity for think tanks to serve government needs; however, the degree of incorporation and cooption varies from think tank to think tank and from one country to another. Latin American institutes, particularly the liberal institutes, were often marginal to the political system until the demise of authoritarian regimes. By contrast, a group of elite think tanks in Southeast Asia have enjoyed a much closer relationship with their governments where the trade-off was accommodation with government concerns and controls. The line between policy 'engagement' and 'entrapment' becomes unclear in such circumstances (Acharya, 2011). A large number of neoliberal or free-market research institutes eschew government funding in order to lay claim to untainted policy analysis. Their critics argue that such institutes can be beholden to other interests such as corporate donors. The organizational choices made by think tanks about their relationship to centers of power and authority must be understood not only by reference to the ideological disposition or mission of the think tank, but also by acknowledging environmental conditions such as legal and economic constraints of the political culture in which they are located.

Outside the political sphere and state sector, think tanks have cultivated other audiences. Students and academics in colleges and universities regularly use think tank publications. Foundation officials, business executives, bureaucrats from various international organizations, university researchers, journalists, and, for want of a better term, the 'educated public' are often engaged by think tank pursuits. Similarly, third-sector intellectuals such as are found among trade unionists, faith-based spokespeople, NGO leaders, and social movement activists can be attracted by the intellectual and ideological sustenance to be found in the think tank community. Indeed, think tanks have became more accessible to a wider range of publics via blogs, YouTube, and other media as they move into the social networking sphere 'twittering' their events and publications. As they do so, think tanks provide an organizational link and communication bridge between their different audiences. They connect disparate groups by providing a forum for the exchange of views, by translating academic or scientific research into policy-relevant 'sound bites,' and by spreading policy lessons internationally. Again, this is a pluralistic conception of a 'marketplace of ideas,' but one that allows a wider societal realm for understanding think tank impacts and influence.

Much of the preceding discussion has focused on individual think tanks seeking impact in short political time horizon or in medium-term policy planning. However, it may be more fruitful to think of influence in the longer term. Paradigmatic change usually occurs over generations. Rather than influence resting in the scientific work of one or more significant scholars, it rests in the aggregate contributions of the research community as whole, of which think tanks are just one type of knowledge actor. It is the collective weight and longer term structural impact of scientific consensus or orthodoxies that institutionally embed certain knowledge as dominant or hegemonic within international organizations, government programs, and other evolving architectures of governance or what has been described as a 'knowledge regime' (Campbell and Pedersen, 2011). The paradigmatic influence of knowledge in governance is much more incremental and 'atmospheric.' As John Maynard Keynes

(1936: p. 383) noted last century, knowledge can be rather more powerful and pervasive in the long term.

> The ideas of economists and political philosophers… are more powerful than is commonly understood. Indeed the world is ruled by little else. Practical men who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist. Madmen in authority who hear voices in the air are distilling their frenzy from some academic scribbler of a few years back.

Yet, due to the passage of time and other intermediary factors, it is difficult methodologically to demonstrate, in positivist terms, ideational cause and subsequent governance effect.

This kind of long-term structural influence is different from the more mundane policy uses of knowledge agents on the one hand, and the symbolic or ritualistic use of knowledge artifacts for legitimacy in governance on the other. In short, the relevance of think tanks may be considered in terms of unintended consequences and negative impact. Rather than conceiving a one-way supply process of think tanks pushing policy knowledge, data, and analysis *into* government circles, it may be more pertinent to consider when, why, and how politicians and bureaucrats find these bodies useful. Rather than seeking sound advice for 'rational' policy making, the demand for think tank products and services can be much more instrumental. Think tanks represent an indirect form of legitimacy for the commissioning government department or international agency in having sought an 'evidence base' or scientific foundation for decisions. Think tanks also have symbolic value: state agencies can legitimize their policy position by arguing that they are consulting independent experts or civil society organizations. For instance, they are useful as 'independent' evaluators monitoring progress on adherence to legislation or international treaties and agreements.

They contribute to the 'everyday affairs' of governance and institution building by facilitating exchange between official and other private actors via networks as well as through various 'low-key' training and professional development courses many institutes offer to the public service. Other mundane roles include organizing seminars and hosting foreign delegations. Because of their professional image, they are viewed as more benign or cooperative partners when compared to the relatively more critical stance and occasionally disruptive lobbying of many NGOs. Think tank work can also be used as an ex post facto justification of policy that was decided upon without reference to independent sources outside government. Suggestions that think tanks are influential when their research is taken on board by government can by undermined if that research is politicized, especially when policy actors are dealing with sensitive social or economic issues. Via selective use, decontextualization, or misquotation, research findings are easy to abuse. Decision makers face incentives to do so to reinforce their policy positions (or prejudices). They will look to institutes that are generally supportive of their policy goals and ideology rather than engage in a time-consuming comprehensive search for the most reliable, 'scientific' or 'evidence-based' policy advice. Not all think tanks can be influential.

A recent development, and one that reflects think tank institutional embedding in the policy milieu, is the rise of support services for think tanks. A cottage industry of advisors and consultants has emerged providing tool kits and guide books or management expertise on how to manage a think tank (Struyk, 2007), how to network a think tank (Weyrauch, 2007), or how to engage in best-practice research (Buldioski, 2009). The Atlas Foundation offers programs such as a Think Tank MBA and Think Tank Leadership Training. Since 2001, *Prospect* magazine has organized an annual competition now described as 'the Oscars of the think tank world' by the BBC, to recognize important and influential work done by think tanks across the globe. There is a Facebook community – On Think Tanks! These initiatives and actors orbiting think tanks testify to the strength and tenacity of the think tank in contemporary policy processes and the political imagination.

In conclusion, the think tank, in all its diversity, has seen a century of growth in virtually all countries around the world. Moreover, these organizations have spread into international governance with developments like the *Think20* network seeking to inform and influence the G20. Yet, with rapid access to analysis and advocacy afforded by the internet, this century may witness less growth in the face of greater competition from NGOs, universities, and consultancy firms for policy influence and political attention.

See also: Agenda Settting, Public Policy In; Global Public Policy; Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks; Policy Analysis: Evidence Based Policy-making; Policy Knowledge: Foundations; Policy Networks; Policy Process: Nonprofit Institutions in the United States.

Bibliography

Abelson, D., 2006. A Capitol Idea: Think Tanks and US Foreign Policy. McGill-Queens University Press, Montreal.

Acharya, Amitav, 2011. Engagement or entrapment? Scholarship and policymaking on Asian regionalism. International Studies Review 13 (1), 12–17.

Boucher, S., Cattaneo, D., Ebélé, J., et al., October 2004. Europe and Its Think Tanks; A Promise to Be Fulfilled. An Analysis of Think Tanks Specialised in European Policy Issues in the Enlarged European Union. Studies and Research No. 35. Notre Europe, Paris.

Buldioski, Goran, 2009. Think tanks in Central and Eastern Europe in urgent need of a code of ethics. The International Journal of Not-for-Profit Law 11 (3).

Campbell, John, Pedersen, Ove, 2011. Knowledge regimes and comparative political economy. In: Beland, Daniel, Cox, Robert Henry (Eds.), Ideas and Politics in Social Science Research. Oxford University Press, New York.

Fischer, Frank, 2003. Reframing Public Policy: Discursive Politics and Deliberative Practices. Oxford University Press, Oxford.

Keynes, J.M., 1936. The General Theory of Employment, Interest, and Money. Harcourt, Brace & World, New York.

McGann, James. G., Sabatini, Richard, 2011. Global Think Tanks. Routledge, Abingdon, Oxon.

Medvetz, Tom, 2012. The Rise of Think Tanks in America: Merchants of Policy and Power. University of Chicago Press.

Pautz, Hartwig, 2011. Revisiting the think tank phenomenon. Public Policy and Administration 26 (4), 419–435.

Plehwe, D., 2007. A global knowledge bank? The World Bank and bottom-up efforts to reinforce the neoliberal developments perspectives in the post Washington consensus era. Globalizations 4 (4), 514–528.

Stone, D., 1996. Capturing the Political Imagination: Think Tanks and the Policy Process. Frank Cass, London.

Stone, Diane, Maxwell, Simon (Eds.), 2005. Global Knowledge Networks and International Development. Routledge, London.

Struyk, R.J., 2007. Managing Think Tanks: Practical Guidance for Maturing Organizations, second ed. Local Government and Public Service Reform Initiative, Budapest; Urban Institute, Washington, DC.

UNDP (United Nations Development Program), 2003. Thinking the Unthinkable. UNDP Regional Bureau for Europe and the Commonwealth of Independent States, Bratislava.

Weaver, R. Kent, September 1989. The changing world of think tanks. PS: Political Science and Politics, 563–578.

Weidenbaum, Murray, 2011. The Competition of Ideas: The World of Washington Think Tanks. Transaction Publishers, New Brunswick, New Jersey.

Weyrauch, Vanesa, 2007. Weaving Global Networks: A Handbook. CIPPEC, and CSGR Working Paper, University of Warwick, Argentina. http://www2.warwick.ac.uk/fac/soc/csgr/research/workingpapers/2007/wp21907.pdf.

Zhu, Xufeng, 2011. Government advisors or public advocates? Roles of think tanks in China from the perspective of regional variations. The China Quarterly. http://dx.doi.org/10.1017/S0305741011000701.

Relevant Websites

http://www.ebpdn.org/ – Evidence Based Policy In Development Network.

www.gdnet.org – Global Development Network (GDN).

http://www.ndri.ned.org/ – National Endowment for Democracy (NED) Network of Democracy Research Institutes.

http://www.opensocietyfoundations.org/about/programs/think-tank-fund – Think Tank Fund of the Open Society Foundations.

http://www.thinktankinitiative.org/ – Think Tank Initiative (co-ordinated by IDRC Canada).

Third World

Maxwell Owusu, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article retraces definitions of the term *Third World* (*Tiers Monde*). The genesis, varied meanings, and popular uses of the concept show historical connections between globalizing capitalism and the emergence of colonial empires. The usefulness and limitations of Third World as an analytical, ideological, or political category are discussed. After World War II, the position of Third World nations changed again, and organizations such as the United Nations fostered enhanced status, sense of purpose, and solidarity among Third World nations. Understandings of what constitutes the Third World came about in the international system after the end of the Cold War. The article concludes with the relevance of the concept in the era of the United Nations Millennium Development Goals.

Defining the Third World

The *Third World* has become a convenient catch phrase that refers collectively to the large number of heterogeneous societies, cultures, civilizations, countries, territories, states, and nations of the non-Western world, most of which lie in the tropics, have experienced European conquest or colonial rule, whose economies are less industrialized or not completely modernized, whose political orders are often less democratic, more autocratic or authoritarian and prone to political instability including coups d'état and civil wars, and a majority of whose population has the lowest standards of living or are considered or consider themselves poor relative to those of the First and Second Worlds. The Third World includes all countries of Africa (except South Africa), Asia (except Japan), Latin America and the Caribbean, and most states and territories of Oceania (except Australia and New Zealand).

The Origin of the Term and Its Varied Meanings

The term *Third World* (*Le Tier Monde*) is of French origin. It was first used in France in the early 1950s by a number of French writers like Alfred Sauvy (politician, economist who is credited with coining the term in 1952) and Claude Boudel in the context of apparently increasing polarization of the international system into rival blocs. The term conveys a conscious echo of the older concept of *le tiers état* that is, *the third estate* (the common people, the poor and underprivileged) of the prerevolutionary *ancien régime*. Some claim that it was Frantz Fanon, a black French Caribbean writer and psychiatrist from Martinique (author of *The Wretched of the Earth*, 1963) who first used the term to refer to the poor decolonized and decolonizing societies and territories of Africa, Asia, Latin America, and the Caribbean. Educated in France, Fanon was a political activist who worked in Algeria, supported the Algerian nationalist struggle for independence from French occupation in the 1950s, and developed revolutionary theories for the decolonization of Africa. A weasel word, the term *Third World* has been used in several related senses, substituting for or succeeding other terms used by Europeans to refer to non-European peoples and cultures: *aboriginal*, *indigenous savage*, *barbarous*, *traditional*, (*primitive*, *tribal*, *peasant*) *native*, *Indian*, *Mohammedan*, *pagan*, *precapitalist*, *preindustrial*, *preliterate*, *non-Western*, *nonwhite*, *premodern*, and *black*. Such terms have been used sometimes with deliberate contempt and condescension, with pejorative connotations, or even racist overtones.

These terms have also been adopted by the educated, more privileged urban dwellers in the Third World itself.

Third World has also been employed as a shorthand euphemism for the social condition or plight of the very poor or powerless or underdeveloped society anywhere in the world, including the First and Second Worlds, even though a vast majority of the people fitting the description live in the Third World. In this sense, the phrase *America's Third World* refers to America's rural and urban poor: the black underclass, Native Americans, and rural whites in Appalachia. In a similar usage, the former Soviet Union was once described by the *Times* of London as having a Third World economy with First World weapons. South Africa is described as a First World society in which the Third World is ever present.

The term *Third World* has also been extended to the poorer Eastern European countries whose economies were subjected to the needs of and exploited by Northern Europe in the early modern period. In everyday use, the term has thus acquired loose, but frequently negative, social, cultural, political, economic, ideological, and racial meanings, often with a fine disregard for geography. From 1945 to 1989 a capitalist block (First World) led by the United States and a communist one (Second World) led by the Soviet Union fought a Cold War that ended with the collapse of the Soviet Union in 1991. In the Cold War context, the Third World consisted mostly of newly decolonized, uncommitted countries not aligned diplomatically with either the capitalist world or the communist world. But the two dominant historic models of development, capitalism and socialism, especially the former, continued to broadly shape the rate and direction of change in the Third World.

Third World as an Analytical Concept

Since World War II and the invention of the term, the conditions of Third World, and the relationship among the First (the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12154-3

rich Western industrialized capitalist countries), Second (the industrialized communist countries of the Soviet Union and Eastern Europe), and Third Worlds have been the focus of theory, intensive research, and extensive documentation aimed at explaining and understanding the nature, origin, and consequences of poverty and underdevelopment in the Third World, and of policy concerned with appropriate strategies for development and poverty alleviation. The leading theories of Third World development and underdevelopment include 'modernization theory,' 'dependency theory,' and 'world-system theory.' Some Marxist scholars have rejected the term *Third World* as a false concept masking the contradictions of capitalist development in the former European colonies (of conquest, exploitation, and settlement). Third World countries often see themselves as old or ancient societies 'lagging behind' economically, technologically, and even politically and thus trying to 'catch up' with (and even overtake) the more advanced industrialized countries of the First and Second Worlds.

The rise and spread of capitalism globally (1500s–1900s) was intimately linked to the growth of large, powerful oppressive and exploitative European colonial empires, inspired by the prevailing mercantilist philosophy of the early modern period dedicated to the unbridled pursuit of economic power and wealth of nation-states backed and secured by strong military presence.

Because of its vast natural resources and potential as sources of readily available cheap or unpaid labor (slave, indentured, semi-proletarian, etc.), which were largely untapped, its considerable potential commercial value and importance of some areas to white settlement and occupation, the Third World has been for over five centuries a principal focus of European territorial ambition and rivalry.

Most of the states and territories of the Third World took their present form in the nineteenth and early twentieth centuries when large parts of Asia and Africa were colonized by European imperial powers. By the 1880s the world had been carved up among a handful of European industrial powers. By 1914, European economic and political control covered most of the global land surface, claiming nearly the entire non-European world. Japan successfully resisted European colonization, and China was never formally colonized but for the highly lucrative treaty ports along the coasts, e.g., Hong Kong following the Opium Wars of the 1830s.

Latin America, conquered, colonized, and settled by the Portuguese and Spanish crowns three centuries before Asian and African colonization, was transformed into independent republics, along with formerly French Haiti in the Caribbean (the first black republic anywhere) in the early nineteenth century, after long, bloody, and bitter nationalist struggles.

Despite wide differences in local cultures, historical traditions, and responses to colonialism, colonies came to share several common features: political control from Europe; unchecked economic exploitation; Euro centrism; incipient Westernization; religious intolerance, especially against the so-called 'pagan' and other indigenous beliefs and practices (including Islam and Hinduism); and particularly in the black world, the construction and reproduction of a racist social order of which apartheid of South Africa was the most notorious recent example. Colonialism that followed the transatlantic slave trade in Africa, during which an estimated 10 million or more Africans were exported to provide unpaid labor for European plantations and mines of the Americas, did incalculable damage to black Africa, notwithstanding some lasting beneficial by-products: Christianity, schools, hospitals, and the start of modern administration. The colonial borders that were drawn across Africa divided African peoples and cultures, and along with the authoritarian and repressive colonial political tradition are largely responsible for much of the chronic poverty, cultural identity crisis, and political instability of independent African countries. One enduring legacy of European colonial empires is a global economic order with an asymmetrical North–South division between a rich industrial North of mostly former imperial powers and poor, less industrialized South of former colonies. The North–South divide is exacerbated by neocolonialist attempts to influence or exercise de facto control over the policies of the South through formal and informal alliances and agreements calculated to ensure the perpetuation of the hegemonic economic power of the North.

The United Nations and the Rise of the Third World

The United Nations (UN) charter endorsed the principle of self-determination and sovereign equality of all nations. After the late 1940s, the large ex-colonies like India, Pakistan, Egypt, and Indonesia, supported by Latin American countries, used the UN as a forum to press for an end to colonialism and rapid development of former colonies. The term *Third World* received formal endorsement following the Bandung conference of Asian and African states in Indonesia in April 1955, which affirmed the solidarity of the Third World and prepared for the birth of the nonaligned movement (NAM). At the conference, China declared herself a Third World country. As an expression of China's solidarity with the Third World, the Chinese leader Chou En Lai on a state visit to Conakry, Guinea, West Africa, spelled out the eight principles of Chinese aid to developing countries, to emphasize an obvious and sharp contrast between the Chinese approach to overseas aid and that of the dependence-creating objectives of Western economic, military, and technical assistance to developing countries of the Third World. The principles embodied China's commitment to making Chinese overseas aid mutually beneficial to both China and the recipient country with no political strings whatsoever attached. The term *Third World* was also adopted in the UN classification system to distinguish between the developed and underdeveloped or developing countries.

The conferences of NAM, the first one held in Belgrade (former Yugoslavia) in 1961 and the founding in June 1964 of the UN Group of 77, at the conclusion of the first UN conference on Trade and Development (UNCTAD), which was convened at the initiative of Third World countries, marked the growth of the Third World as a potent force in world affairs.

With approximately two-thirds of the total UN membership, the 1960s and 1970s witnessed the rapid rise in the power and sense of purpose of the Third World. The G-77 spearheaded the UN call for a New International Economic Order (NIEO) to reform the economic relations between the North and the South and to rectify the inequality evident in the

relative access of the rich and poor countries to markets, to supplies, to science and technology, and to credit, which implied a redistribution of wealth in favor of the South. The issue of North–South relations was made the subject of an international commission that met from 1977 to 1979 under the chairmanship of Willy Brandt.

By the 1990s, with the collapse of the Soviet Union and the end of bipolar politics, NAM and the G-77 which have a joint coordinating committee, shifted their focus to problems of underdevelopment, the Third World debt burden, trade liberalization, and aid.

The G-77, whose membership increased to 130 with the admission of postapartheid South Africa, believes that global shared prosperity through North–South cooperation on a fair and equitable basis is the only long-term remedy for Third World poverty and human misery.

Diversity, Division, and Integration

Depending on the definition and context, different countries or populations will make up the Third World. In 1989 before communism in the former Soviet Union and Eastern bloc countries (Second World) suddenly collapsed, the World Bank – the only global multinational development organization grouped countries by regions and by gross national product (GNP) per capita into three main categories as follows: (1) high income countries – 25 countries, 19 of which consisted mostly of the Western industrial developed countries, many of them former colonial powers, and Japan (First World). The remaining six countries – Hong Kong, Israel, Kuwait, Saudi Arabia, Singapore, and United Arab Emirates – had high GNP per capita but not necessarily at the same level of development as the Western industrial countries, with a total population of 0.8 billion; (2) developing countries (Third World) divided into (a) 42 low-income countries – with the lowest per capita income and a population of 2.8 billion; and (b) 53 middle-income countries, with per capita income slightly higher than that of (a) with a population of 1.1 billion. The total population of the 95 developing countries was 3.9 billion; and (3) nonreporting, nonmember countries – a mixed group of Second and Third World communist/socialist countries – e.g., USSR, Cuba, Albania, German Democratic Republic, Peoples Republic of Korea, Angola and Namibia, and countries (mostly islands) with populations of less than 1 million each. The total population of this group was 0.4 billion. The goal of economic development, for the World Bank, is to fight poverty and raise the living standards of low- and middle-income countries of the Third World. Clearly, more than three-quarters of all humanity live in developing countries of the Third World. By 2000, the Third World population was estimated to have risen to four-fifths of mankind. The developing countries as a whole, a very diverse group, accounted for more than two-thirds of the earth's land surface and 80% of the world's poor. Today (in 2012) of the estimated global population of 7 billion, only about 20% live in the rich, developed North – the overwhelming majority still living in the mostly impoverished but rapidly developing emergent market countries of the South, such as China, India, Brazil, Indonesia, Mexico, and South Africa. Accordingly, the Third World may be further divided into those states that have access to rich natural resources, faster growth rates, or better development potential, and those that do not. While the latter are sometimes described as the *Fourth World*, that term is probably more frequently used to refer to marginalized indigenous or tribal peoples. Despite internal divisions and differences, Third World countries share some common features as well as problems.

Most of the peoples of the Third World still suffer from hunger, disease, and lack of shelter, potable water, and opportunities (especially for women) to realize their potential as human beings. Third World countries are faced with crippling external debt burdens, poor social and physical infrastructure, and inadequate technology, making it difficult or impossible to recover from or respond effectively (without massive external help mostly from the First World) to natural disasters and civil conflicts to which many are prone. Because of the chronic and deepening poverty in the midst of burgeoning populations, diminishing natural resources, and increasingly degraded environments, these countries have also become major producers of massive illegal political and economic migrations to the more prosperous industrial capitalist North, and of economic and political refugees and internally displaced persons within the Third World itself.

The Third World and the Post-Cold War International System

The Third World as an ideological concept has been undermined by the end of the Cold War, the collapse of Soviet-style communism, and the increasing cooperation between the leading industrial capitalist nations – the G-7 and Russia. Its economic meaning has also been challenged by the fundamental changes in the world economy, brought about by the forces of globalization. The transition of the former Soviet Union and Eastern bloc countries and some Third World nations to pluralist democracy and market economies has made it increasingly difficult to distinguish between some First and Second World countries, and in some cases between Third World and Second World countries.

Given the wide range in the levels of economic development and standards of living within and between Third World countries, placing prosperous Hong Kong and impoverished Haiti, or industrializing Brazil and hunger-prone Bangladesh, in the same category is clearly questionable. Again, by the end of the twentieth century, new economic groupings made up of selected countries of the First, Second, and Third Worlds, some of them regional (e.g., NAFTA), were formed to promote regional prosperity; and others, cross-regional (e.g., Group of 20), emerged to foster dialogue if not unity of action between the rich industrial countries and countries of the emerging markets. The growing convergence of national and transnational economic and political interests has added a new challenge to the usefulness of the Third World as an economic or ideological category.

In the new strategic global environment, NAM also seems increasingly anachronistic. But the overall objectives of NAM, which include the pursuit of world peace, the elimination of disease, hunger, and ignorance; the conduct of international relations on the basis of respect for and understanding of other

cultures, systems, and values; and the elimination of all forms of neocolonialism and racism, make it still relevant.

The Third World, the Third Millennium, and the WORLD WAR ON WANT

In September 2000, at a special session of the UN General Assembly, 189 member states of the UN adopted an historic Millennium Declaration committing countries rich and poor to take concerted action to eradicate poverty, promote human dignity and equality, and achieve peace, democracy, and sustainable environment. The world leaders agreed on a set benchmark of eight Millennium Development Goals (MDGs) and related targets to guide developing countries to pursue programs along the agreed goals with the start date fixed at 1990 and the end date for achieving the goals of 2015 or earlier. The MDGs are as follows: eradicating extreme poverty and hunger; achieving universal primary education; promoting gender equality and empowering women; reducing child mortality; improving maternal health; combating HIV/AIDS, malaria, and other diseases; ensuring environmental sustainability; and developing a Global Partnership for Development. Indeed many developing countries were pursuing programs aimed at tackling the development challenges facing their countries and peoples. What the adoption of the MDGs did was to galvanize most leaders of the developing countries to undertake and aggressively pursue unprecedented measures aimed at meeting the needs of their most downtrodden, disadvantaged, and discriminated-against populations with the target dates in mind. Since 2001, the implementation of the MDGs has been transforming development and human lives in the Third World for the better. Governments, aid agencies, nongovernmental organizations (NGOs), and civil society organizations worldwide have been reorienting their development programs and work around the MDGs, including the construction of model Millennium Development villages to stimulate holistic and sustainable rural development.

The World Bank's Development Research Group in its estimates for global poverty updated from 2005 (its figures for people living in absolute poverty – that is both the number and proportion of the population living on less than the equivalent of $1.25 (US) a day (at 2005 prices) – the most commonly accepted poverty line) shows that in 2008, the first year of the global financial and food crises, absolute poverty was falling in every part of the world. This was the first instance of declines across the board since the World Bank started collecting the figures in 1981. The estimates for 2010, though incomplete, show global poverty that year was half its 1990 level. The world thus reached the UN's MDG of halving world poverty between 1990 and 2015, 5 years early.

Significantly, much of the credit for this unusual reduction in poverty goes to the emerging powers/markets, especially the so-called BRICS countries – Brazil, India, Russia, China, South Africa – and led by China. Fifty percent of the long-term rate of world poverty decline is attributable to China alone, which has pulled more than half a billion (660 million) people out of poverty since 1981. Of the roughly 1.3 billion people worldwide, living on (the equivalent of) less than $1.25 (US) a day in 2008, 1.1 billion of them were outside China. China also accounts for most of the rapid progress in East Asia, which in the early 1980s had the highest poverty rate in the world with 77% of the population living below the poverty line.

Africa has also seen the largest recent long-term improvement in poverty reduction. The number of people in poverty in sub-Saharan Africa almost doubled from 205 million in 1981 to 395 million in 2005. But in 2008 it fell by 12 million to 47%, the first time less than half of the African population has been below the poverty line. The number of poor people in Latin America had also been rising (from lower levels) as in Eastern Europe and Central Asia. But these regions have reversed the trend since 2000.

The positive turn in economic and social progress in many Third World countries reflects the long-run high rates of sustainable growth in China, which has replaced Japan as the world's second largest economy after the United States. China trade and investment abroad is reshaping the twenty-first century world. Equally important in poverty reduction in Africa and Latin America are good governance; political stability; the global spread of access to credit through microfinance, especially by women; the rising prices for Third World principal export commodities; and the social programs pushed by democratically elected populist or leftist governments in Brazil, Venezuela, and Bolivia that have greatly boosted social spending to help the poorest of the poor in urban slums and rural farms making less than $1.25 (US) a day.

Among the developing countries that have either halved poverty already or are expected to do so by 2015 are Cambodia and Ghana. Overall, 88% of developing countries should meet the Millennium Development poverty goal. Indeed, the world is on track to halve extreme hunger in many of the same countries where poverty is declining. About 45% of Third World nations have already reached this goal, and another 33% are close to reaching it. Again 83% of developing countries are either close to reaching the goal of providing universal primary education or have reached it. Achievements of MDGs have been helped by the big rise in inter–Third World aid and by improvements in systems of technical assistance and aid delivery, including e-health diagnosing by direct teleconferencing. Between 2005 and 2010 Brazil, which has replaced the United Kingdom as the world's sixth largest economy, and India grew their foreign aid spending by more than 20%. This type of aid using low-cost resources and strategies has proven more efficient and effective in reaching large populations in nonurban settings in the global South.

The number of people living below the poverty line worldwide is projected to be 883 million in 2015 compared with 1.4 billion in 2005 and 1.8 billion in 1990. But within developing countries that are making progress toward MDGs, there still exist persistent pockets of deprivation with concentrations in urban slums among society's excluded populations and in the countryside. In most of the Third World there is a disturbing contrast between a growing, yet still relatively small middle-class elite and the large majority of people left behind. There is also a lower tier of countries, mostly in sub-Saharan Africa, that remains trapped in extreme poverty and misery. Many of these are either landlocked states or states with few natural resources, degraded

environments, or ineffective or corrupt and/or unstable governments.

Taking a long-range view of the Third World the experts estimate, based on the developing world's high growth rates from the middle of the last decade and its enormous development potential, that by 2050, 80% of future global growth will be from the emerging markets in Asia, Africa, and South America, and the future prosperity of the developed West will be linked to the success of those countries led by China, India, Brazil, Indonesia, South Africa, and the rise of the rest. The major players in the developing world are already catching up to or even surpassing their counterparts in the developed world. The convergence of the economies of the Third World (with Russia in solidarity with the Third World) and those of the First World would seem to be the order of the day. This could usher in a new world order, not dominated by any one superpower, but a world more united, more peaceful, more just, and more equal, that is a truly one-world order.

See also: Center–Periphery Relationships; Colonialism, Anthropology of; Development: Social-Anthropological Aspects; Global South; Globalization, Anthropology of; Modernity: Anthropological Aspects; Nationalism: General; Orientalism, Oppression, and Contemporary Social Work; Political Economy in Anthropology; Postcoloniality.

Bibliography

Amin, S., 1976. Unequal Development. Monthly Review Press, New York.
Brandt, W., 1980. North-South: A Program for Survival. The Report of the "Independent Commission on International Development Issue Under the Chairmanship of Willy Brandt". MIT Press, Cambridge, MA.
The Economist. Special Report World Economy, 13 October 2012.
Fanon, F., 1963. The Wretched of the Earth. Grove Press, New York.
Forson, A., 1998. The non-aligned movement and the third world. Daily Graphic (Accra) August 26 (7).
McConnick, J., August 1988. America's third world. Newsweek 8, 21–24.
Owusu, M., 1992. Democracy and Africa – a view from the village. The Journal of Modern African Studies 30 (3), 369–396.
Owusu, M. (Ed.), 1979. Colonialism and Change. Mouton Publishers, The Hague, The Netherlands.
Owusu, M., 1975. Policy studies, development, and political anthropology. The Journal of Modern African Studies 13 (3), 367–381.
Rodney, W., 1972. How Europe Underdeveloped Africa. Bogle L'Ouverture Publications, London.
Sauvy, A., 1952. General Theory of Population (C. Campos, Trans.). Methuen, London.
Singham, A.W. (Ed.), 1977. The Non-Aligned Movement in World Politics. Lawrence Hill, Westport, CT.
Stavrianos, L.S., 1981. Global Rift. The Third World Comes of Age. William Morrow, New York.
The Times. 22 April 1983.
Wolf, E.R., 1982. Europe and the People without History. University of California Press, Berkeley, CA.
World Bank Global Monitoring Report, 2011. Improving the Odds of Advancing the MDGs. Population Living below 1.25 per Day by Region and Year Estimates.
Worsley, P., 1984. The Three Worlds. Culture and World Development. University of Chicago Press, Chicago.
Yapp, M.E., 1985. What is third world history? History Today 35, 37.

Thorndike, Edward Lee (1874–1949)

Robert M Thorndike, Western Washington University, Bellingham, WA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 22, pp. 15681–15683, © 2001, Elsevier Ltd.

Abstract

'Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality' (Thorndike, 1918, p. 16). Few careers as diverse as that of Edward Lee Thorndike could be summarized so succinctly. Yet the hallmark of all the work of this extremely prolific psychologist was his abiding faith in the quantifiability of all aspects of human experience. From his first abortive attempts in 1896 to study mind reading through his last publication on shifts of interest with age, he was committed to quantification and, where possible, the controlled experiment. He applied empirical methods to topics as diverse as animal experimental psychology, the scientific study of education, and lexicography. 'Professor Thorndike's major life-purpose was … to demonstrate the unrivalled fruitfulness of the scientific method in the solution of social problems and to introduce it in exacting, quantitative forms which had been so productive in the physical fields' (Gates, 1949, p. 31).

'Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality' (Thorndike, 1918, p. 16). Few careers as diverse as that of Edward Lee Thorndike could be summarized so succinctly. Yet the hallmark of all the work of this extremely prolific psychologist was his abiding faith in the quantifiability of all aspects of human experience. From his first abortive attempts in 1896 to study mind reading through his last publication on shifts of interest with age, he was committed to quantification and, where possible, the controlled experiment. 'Professor Thorndike's major life-purpose was to demonstrate the unrivaled fruitfulness of the scientific method in the solution of social problems and to introduce it in exacting, quantitative forms which had been so productive in the physical fields' (Gates, 1949, p. 31).

Life Events

E. L. Thorndike was born August 31, 1874 in Williamsburg, Massachusetts, the second son of Edward Roberts and Abbie (Ladd) Thorndike. The senior Thorndike was a Methodist minister who served churches in Maine and Massachusetts. The family often spent parts of their summers during E. L. Thorndike's childhood at an evangelical Methodist camp meeting on Martha's Vineyard. Perhaps in part as a reaction to the intensely religious nature of his upbringing, E. L. Thorndike rejected religious practices almost as soon as he left home for study at Wesleyan University. Throughout his life, he maintained a posture that favored science over religion.

After graduating with highest honors from Wesleyan in 1895, Thorndike went to Harvard University to study with William James (see *James, William* (*1842–1910*)). It was under James that he began his psychological studies, first of mind reading (a failure), then of learning (a success). The course of his studies for the next two years took him from William James at Harvard to James McKeen Cattell at Columbia, and from maze learning in chicks to his famous puzzle box studies with cats, in which he traced for the first time the trial-and-error learning process as his hungry subjects attempted to escape from crude cages to reach food. His doctoral dissertation describing these experiments, *Animal Intelligence: An Experimental Study of the Associative Processes in Animals* (1898), is one of the most well-known and influential studies in all of psychology.

Upon completion of his doctorate, Thorndike took a position at Western Reserve University. After one year he gladly accepted an invitation to return to New York City and become a member of the founding faculty at Teachers College in the fall of 1899. Thorndike remained at Teachers College until his retirement in 1940, a period that saw the publication of approximately 450 books, monographs, and papers. Another 70 were published in retirement.

Thorndike married Elizabeth Moulton on August 29, 1900. They had four children who survived infancy. Feeling his own training had been very deficient in mathematics, Thorndike insisted that all his children develop a solid background in mathematics. All of the children earned doctorates in scientific fields, and the second son, Robert L., followed in his father's footsteps so closely that he eventually occupied E. L.'s office at Teachers College.

E. L. Thorndike's work was widely recognized within his own lifetime. He was elected President of the American Psychological Association in 1912, President of the American Association for the Advancement of Science in 1933, and served as Second President of the Psychometric Society in 1936. In 1925 he was awarded the Butler medal in gold by Columbia University. Cattell found him ranked first among American psychologists in a 1921 poll for American Men of Science. Many of Thorndike's friends and colleagues said he was a 'workaholic,' but what he enjoyed more than anything was simply messing around with data. Other than tennis (he built his own court) and golf (he used a putter anywhere within 50 yards of the green), his recreation was analyzing data. Retirement was not a happy time for Thorndike. His health was not good: he had been a heavy smoker and quite overweight much of his adult life, and he suffered from arteriosclerosis. But what troubled him most in his old age was an acute awareness of his own declining cognitive abilities. His mind was no longer capable of the feats of reasoning that had been central to his professional life (R. L. Thorndike, personal communication). He died of a massive cerebral hemorrhage, at his home in Montrose,

New York, on August 9, 1949, the day on which plans to celebrate his 75th birthday were to be finalized.

Thorndike's Research and Theory

As a result of his experiments with learning in cats, E. L. Thorndike formulated a theory of mental associations which he called connectionism. He believed that the central nervous system was the basis of all behavior, that learning and other behavioral modifications took place when connections were formed between neurons, and that there was a continuity among species and a similar continuity in human development. 'The intellectual evolution of the race consists in the number, delicacy, complexity, permanence, and speed of formation of associations. In man this increase reaches such a point that an apparently new type of mind results, which conceals the continuity of the process' (Thorndike, 1901, p. 65).

It has been suggested that Thorndike was driven to applied work with humans by employment pressures. This may have been true for Robert Yerkes (Thorndike, 1990) but in E. L. Thorndike's case his first interest was humans and he resorted to animal research 'because they wouldn't let him work with children' (R. L. Thorndike, personal communication).

Upon returning to Columbia, Thorndike quickly turned his attention to human problems. He was fascinated by individual differences, trying to measure them, explain them, and use them for prediction. He believed, as did many of his generation, that science could solve all human problems, and he set out to apply scientific method to problems of education. Much of the remainder of his career was spent trying to develop methods to measure individual differences and to use these measurements to make education more efficient and successful.

Thorndike set out immediately to place education on a sound scientific footing. Educational practice of the time, the doctrine of mental discipline, was guided by 'the teachers' faith that mental abilities are so little specialized that improvement in any one of them will help all the rest; manual training is often introduced into schools on the strength of somebody's confidence that skill in movement is intimately connected with efficiency in thinking' (Thorndike, 1903, p. 2).

Curriculum was founded on opinion, not evidence. Based on his transfer of training studies with Woodworth (Thorndike and Woodworth, 1901) and similar studies he and his students conducted, he concluded that 'since those who succeed in the study of Latin are better at general discrimination and judgment than those who fail, we conclude that learning Latin vastly improves general discrimination and judgment ... It suits the vanity of educational theory to fancy that the changes [that result from study] are wholly due to discipline' (Thorndike, 1903, p. 93).

Thorndike continued his pursuit of quantification and applying scientific method in education with his publication of one of the first books on psychological and educational measurement and statistics in 1904. Here he noted the difficulty of defining interval scales for measuring cognitive activities, advocated the use of norm-referenced measurement as pioneered by Galton, and sounded a strong warning concerning the reliability and validity of measurements.

From the beginning, Thorndike adopted the view that inherited differences in cognitive abilities were important. 'There is no reason to suppose that the brain is less influenced by [immediate ancestry] than the tissues that cause height. ... Immediate ancestry is thus the probable cause for original mental nature' (Thorndike, 1903, p. 53). 'The schools always have and always will work to create a caste, to emphasize inequalities. Our care should be that they emphasize inequalities, not of adventitious circumstances, but of intellect, energy, idealism, and achievement' (p. 96). Thorndike was a thoroughgoing intellectual meritocrat, but he opposed discrimination on any other basis.

The rise of behaviorism in the 1920s shifted psychological research away from an interest in 'original nature' for almost 60 years. Many of the findings of modern behavioral genetics and cognitive psychology support positions Thorndike proposed at about the time genes were being rediscovered or the mind was being banished from psychology. For example, he argued that there was an interaction between genotype and environment. Anticipating current thought, he suggested that 'Each nature in some measure selects its own environment, and each nature may get from an environment a different influence' (Thorndike, 1911, p. 48).

In addition to his efforts to make education more efficient and scientifically sound, Thorndike expended considerable effort from 1900 to 1925 in developing measures of what he referred to as intellect. He rejected the notion proposed by Charles Spearman that there was a single mental trait of intelligence, arguing instead that each test must specify what aspects of intellect it was designed to measure (see Thorndike, 1990 for a description of the Spearman–Thorndike debate). Beginning in 1903, and anticipating more popular work by Alfred Binet (see Binet, Alfred (1857–1911)), Thorndike and his students developed a wide variety of measures of human abilities. The work culminated in the CAVD, a 'test constructed (from 1922 to 1925) as a sample of what a measuring instrument for a mental ability should be' (Thorndike, 1949, p. v). With the CAVD, Thorndike (Thorndike et al., 1927) produced a test that would foreshadow item response theory and the test design methods of the 1990s.

In the 1930s, Thorndike turned his attention to issues of social science and social value. He pioneered the scientific study of environmental quality with his research on urban environments in Your City (1939) and he subjected the domain of values to scientific scrutiny in *Science and Values* (1936). But perhaps Thorndike had his greatest impact during this period through his contributions to lexicography. In the 1930s Thorndike revolutionized the dictionary by placing the design of these reference works on an empirical footing.

For years Thorndike had been collecting data on the relative frequency of words in published English. A list of the most frequently appearing 20,000 words, along with their relative frequencies, was published in 1931. The list was expanded to 30,000 entries in 1944. Thorndike used word counts to determine which words should appear in his dictionary, and semantic counts (the most common meanings of each word) were used to select the definitions and their order. In addition, Thorndike insisted that each word be defined only in terms of words that would be more familiar than the one being defined. The Thorndike Century Junior Dictionary, published in 1935,

was the first dictionary to use these principles and to focus its content on the needs of its most probable users. By 1950, the three Thorndike Century dictionaries dominated the field of dictionaries in education, and every serious competitor was copying his approach. All modern dictionaries follow his pattern.

E. L. Thorndike's influence is so pervasive in modern education and psychology that it is almost impossible to detect. If his contributions to lexicography were not enough, he pioneered the scientific study of education, made major contributions to psychological and educational measurement and testing, and put comparative psychology on a firm experimental footing. There is hardly a specialty in either field, with the possible exception of some clinical areas, that has not been significantly touched by his work or that of his hundreds of students, a list of which would read like a Who's Who of psychology and education in the middle and latter half of the twentieth century, including Abraham Maslow (who rejected his empiricism) and Raymond Cattell (see Cattell, Raymond Bernard (1905–98)) (who embraced it). A fairly complete list of the publications of E. L. Thorndike is available in Gates (1949); a full biography is given in Joncich (1968).

See also: Cognitive Psychology: History; Connectionist Models of Concept Learning; Educational Learning Theory; Lexicology and Lexicography; Pavlov, Ivan Petrovich (1849–1936); Skinner, Burrhus Frederick (1904–90); Watson, John Broadus (1878–1958).

Bibliography

Gates, A.I., 1949. The writings of Edward L. Thorndike. Teachers College Record 51, 28–31.

Joncich, G., 1968. The Sane Positivist: A Biography of Edward Lee Thorndike, first ed. Wesleyan University Press, Middletown, CT.

Thorndike, E.L., 1898. Animal Intelligence: An Experimental Study of the Associative Processes in Animals. Columbia University, Ph.D. thesis.

Thorndike, E.L., 1901. Evolution of human intellect. Popular Science Monthly 60, 58–65.

Thorndike, E.L., 1903. Educational Psychology. The Science Press, New York.

Thorndike, E.L., 1911. Individuality. Houghton Mifflin, New York.

Thorndike, E.L., 1918. The nature, purposes, and general methods of measurement of educational products. In: Courtis, S.A. (Ed.), The Measurement of Educational Products (17th Yearbook of the National Society for the Study of Education, Pt. 2. pp. 16–24). Public School, Bloomington, IL.

Thorndike, E.L., 1949. Selected Writings from a Connectionist's Psychology. Appleton-Century-Crofts, New York.

Thorndike, E.L., Woodworth, R.S., 1901. The influence of improvement in one mental function on the efficiency of other functions. Psychological Review 8, 247–261, 384–95, 556–64.

Thorndike, E.L., Bregman, E.O., Cobb, M.V., Woodyard, E., 1927. The Measurement of Intelligence. Teachers College Bureau, New York.

Thorndike, R.M., 1990. A Century of Ability Testing. Riverside, Chicago, IL.

Time Perception Models

Lorraine G Allan, McMaster University, Hamilton, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15696–15699, © 2001, Elsevier Ltd.

Abstract

This article focuses on signal detection and information-processing models for time perception. The pseudologistic model and the scalar timing model are presented as exemplars of these two classes of models. For both models Weber's law describes the relationship between variability in perceived time and mean perceived time, and a power function with an exponent close to 1.0 describes the relationship between mean perceived time and clock time. The application of the two models to the temporal bisection task is considered.

Time perception is a distinct area of study with its own psychophysical methods designed for assessing the perceived duration of a temporal interval. There is now an extensive database to guide the development of models of time perception. Most timing models are firmly rooted in an *internal clock*. Internal clock models postulate a timing-specific mechanism which produces perceived temporal values that bear orderly relations to real time. This timing-specific mechanism, or internal clock, is an interval timer, rather than a periodic or oscillatory mechanism. A periodic clock, such as the kind responsible for circadian rhythms, is continuously running and self-sustaining. In contrast, interval clocks need to be initiated with a signal. Once started, the system goes to completion, comes to a rest on its own, and must be restarted.

The first psychophysical models for time perception were proposed at about the same time by Creelman (1962) and Treisman (1963). Creelman (1962) introduced the signal-detection approach, which was continued by Allan and Kristofferson (1974). The most sophisticated signal detection model is Killeen et al.'s (1997) *Pseudo-logistic model* (PLM). The information processing approach was introduced by Treisman (1963) and continued by many others, including Eisler (1975), Thomas and Cantor (1978), and Zakay and Block (1994). The most developed of the information-processing models is Gibbon et al.'s (1984) *Scalar timing* (ST), which has been reformulated as a connectionist model by Church and Broadbent (1991). In direct contrast to information-processing models is Killeen and Fetterman's (1988) behavioral theory of timing (BeT).

Because signal-detection models and information-processing models have been applied to both human and nonhuman timing data, we shall focus on such models in this article, presenting ST and PLM as exemplars of these two classes of models. Specific versions of ST and PLM have been derived for a variety of timing tasks, such as time-left, temporal generalization, and *temporal bisection*. We shall examine ST and PLM in their application to temporal bisection.

In the prototypic temporal bisection task (e.g., Allan and Gibbon, 1991), the subject is familiarized with a constant pair of referents, one short (S) and the other long (L). On probe trials, a temporal interval t, $S \leq t \leq L$, is presented, and the subject is required to indicate whether t is more similar to $S(R_S)$ or to $L(R_l)$. Temporal bisection data are summarized as a psychometric function relating the proportion of long responses, $P(R_L)$, to probe duration t. The *bisection point* ($T_{1/2}$) is the value of t at which R_S and R_L occur with equal frequency, $P(R_L) = 0.5$. One interpretation of $T_{1/2}$ is that it is the point of subjective equality (PSE), identifying the perceived value of t that is midway between S and L. Empirically, it has been found that the location of $T_{1/2}$ is influenced by the size of the L-to-S ratio (r) and by the spacing of the probes (see Allan, 1998). For $r \leq 2$, $T_{1/2}$ is around the geometric mean (GM) of the S and L referents, but moves toward their arithmetic mean (AM) as r increases. For $r > 2$, $T_{1/2}$ for linearly spaced probes is greater than that for logarithmically spaced probes, with the probe spacing effect becoming negligible for $r \leq 2$.

Scalar Property

Both ST and PLM incorporate *Weber's law* for the duration values used in most psychophysical experiments. According to Weber's law, for a given clock time $t > 0$, the variability in perceived time (σ_t) is *scalar* with mean perceived time (μ_t); that is, there is a constant $\gamma > 0$ such that

$$\frac{\sigma_t}{\mu_t} = \gamma \qquad [1]$$

Also, for both models, mean perceived time is a power function of clock time with an exponent close to 1.0, i.e.,

$$\mu_t = t \qquad [2]$$

and therefore eqn [1] can be rewritten as

$$\frac{\sigma_t}{t} = \gamma \qquad [3]$$

The scalar property in eqn [3] results in distributions that superpose when the temporal axis is normalized with respect to the mean of the distribution. This superposition in relative time reflects a rescaling in time, a scale-invariant error distribution for perceived time.

Scalar Timing

ST has its origins in scalar expectancy theory (Gibbon, 1977), and made its appearance as a timing model in Gibbon et al. (1984). The model consists of three, interrelated information

processing stages – clock, memory, and decision – which are associated with timing, storage, and responding, respectively. The clock is responsible for transforming clock time into perceived time. It consists of a pacemaker that emits pulses at some mean rate, a switch that is controlled by a timing signal, and an accumulator that sums the pulses. There are two memory registers for the storage of clock information, *working memory* and *reference memory*. Working memory is loaded from the accumulator and serves as an extended buffer for temporal information from the current trial. Reference memory stores critical temporal information from past trials. The comparator determines a response on the basis of a decision rule which involves a comparison of current time stored in working memory with remembered time sampled from reference memory.

According to ST, the decision to respond R_S or R_L in temporal bisection is made by comparing the similarity of the perceived value of t with memories of the two referents, S and L. This comparison is based on a ratio of the similarity of t to S relative to the similarity of t to L. If that ratio is less than some fixed value β, the response is R_L. In his mathematical derivation of the *bisection function*, Gibbon (1981) combined the similarity ratio decision rule with two locations for the scalar variability, perception or memory. In the referent known exactly (RKE) model, he assumed no variability in the memory for the referents and placed the scalar variability in the perception of the probes. In contrast, in the stimulus known exactly (SKE) model, he assumed no variability in the perception of the probes and placed the scalar variability in the memory for the referents. There are two parameters in these models. The noise parameter γ is the coefficient of variation of time. For SKE, γ is the scalar variability in the memory representations of the S and L referents; for RKE, γ is the scalar variability in the perception of t. For both models, β is a response bias parameter. It follows from the assumptions of the model that the SKE bisection function is approximately

$$P(R_L) = 1 - \phi\left[\frac{1 - \beta\left[\frac{t}{\sqrt{SL}}\right]^2}{\gamma\sqrt{\gamma^2 + 2}}\right] \quad [4]$$

and the RKE function is approximately

$$P(R_L) = 1 - \phi\left[\sqrt{\frac{SL}{\beta\gamma^2 t^2}} - \frac{1}{\gamma}\right] \quad [5]$$

where ϕ is the cumulative normal distribution function. For both models, $T_{1/2}$ is proportional to the geometric mean (GM) of the S and L referents:

$$T_{1/2} = \sqrt{\frac{SL}{\beta}} = \frac{GM}{\sqrt{\beta}} \quad [6a]$$

and for unbiased responding ($\beta = 1$), $T_{1/2}$ is at the GM of the S and L referents:

$$T_{1/2} = \sqrt{SL} = GM \quad [6b]$$

Equation [6a] provides for variables such as ratio and probe spacing to have their effect on the location of $T_{1/2}$ through β. For example, with a large r and linear probe spacing, $T_{1/2}$ located at the GM would yield more R_L than R_S responses. To balance the frequency of the two response categories, it would be necessary to set $\beta < 1$, moving $T_{1/2}$ towards the AM. With log probe spacing, however, the two response categories would be equal with $\beta = 1$ and $T_{1/2}$ at the GM.

Pseudologistic Model

In addition to deriving the bisection function for the similarity ratio rule, Gibbon (1981) also derived the function for the signal detection likelihood ratio rule. For the likelihood rule, the decision to respond R_S or R_L is based on the odds ratio that the perceived value of t is generated by S or L. For this rule, Gibbon (1981) placed the scalar variability in the perception of the probes, resulting in a signal-detection model for temporal bisection. Killeen et al. (1997) simplified the derivation of the bisection function (and the resulting equation) by approximating the normal distribution with the logistic. The pseudologistic bisection function is

$$P(R_L) = \left[1 + \exp\left[\frac{C - t}{0.55\gamma t}\right]\right]^{-1} \quad [7]$$

where C is the signal detection criterion. According to PLM, $T_{1/2}$ is the value of t where the S distribution crosses the L distribution. Killeen et al. (1997) showed that $T_{1/2}$ is proportional to the harmonic mean (HM) of S and L; i.e.,

$$T_{1/2} = (1 + k)\frac{2}{\frac{1}{S} + \frac{1}{L}} = (1 + k)HM \quad [8]$$

where

$$k = \left[\frac{\gamma \ln\left[\frac{L}{S}\right]}{2}\right] \quad [9]$$

PLM predicts that the L-to-S ratio will affect the location of $T_{1/2}$.

PLM is presented schematically in **Figure 1** for $r = 4$ ($L = 8$ and $S = 2$). The spacing of t is linear, $\gamma = 0.15$, and C is located at $(1 + k)HM$. It is clear from **Figure 1** that because of the scalar property a percept located at $C = (1 + k)HM$ is more likely to have been generated by L than by S. Killeen et al. (1997) suggested that in order to balance the frequency of the two response categories, C would move toward the AM with increasing r rather than being set at $(1 + k)HM$.

Pseudologistic Model and the BeT

Killeen et al. (1997) emphasized that PLM is a descriptive rather than a process model in that it resides close to the level of the data and does not postulate particular underlying processes that generate the data. They do describe, however, how in their view PLM dovetails with BeT.

Other Models

Allan and Kristofferson (1974: 26) concluded that "There are few quantitative theories of duration discrimination and few

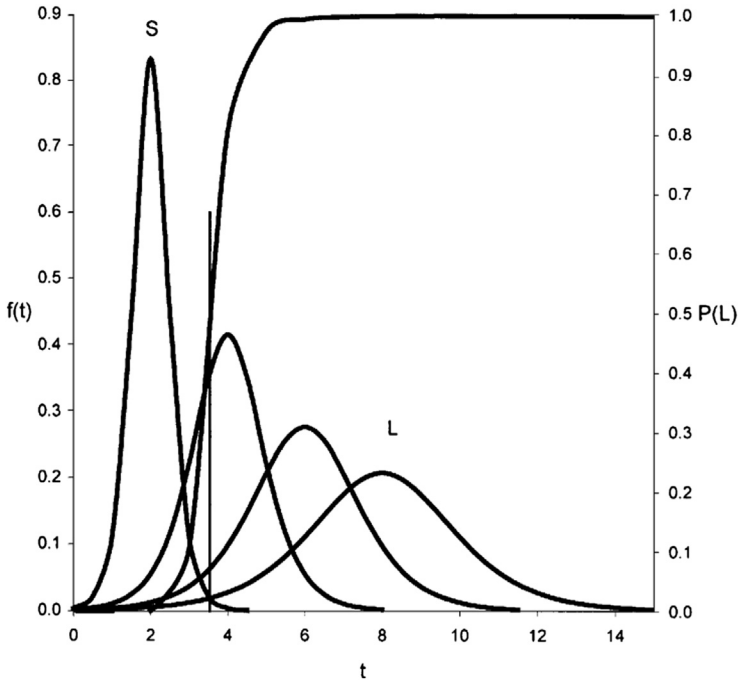

Figure 1 Schematic of the PLM model for $r = 4$ ($L = 8$ and $S = 2$). The logistic density functions [f (t)] are for $t = 2$, 4, 6, and 8. The psychometric function [$P(R_L)$] is for C located at $T_{1/2} = (1 + k)HM$ and $\gamma = 0.15$.

established empirical phenomena to guide theorizing." Five years later, Allan (1979: 340), referring back to Allan and Kristofferson (1974), stated that "Since that time, there has been a dramatic change. Many articles on the discrimination of brief temporal intervals, and several new quantitative models, have appeared in a period of a few years." Now in 2000 we have an abundance of quantitative models for time perception. In this short article, I have merely provided the reader with a few gems. To obtain a broader exposure to and appreciation of the current status of quantitative models for time perception, the reader is referred to two special issues of *Behavioural Processes* in 1998 and 1999 edited by Jennifer Higa.

Bibliography

Allan, L.G., 1979. The perception of time. Perception and Psychophysics 26, 340–354.

Allan, L.G., 1998. The influence of the scalar timing model on human timing research. Behavioural Processes 44, 101–107.

Allan, L.G., Gibbon, J., 1991. Human bisection at the geometric mean. Learning and Motivation 22, 39–58.

Allan, L.G., Kristofferson, A.B., 1974. Psychophysical theories of duration discrimination. Perception and Psychophysics 16, 26–34.

Church, R.M., Broadbent, H.A., 1991. A connectionist model of timing. In: Commons, M.L., Grossberg, S., Staddon, J.E.R. (Eds.), Neural Network Models of Conditioning and Action. Erlbaum, Hillsdale, NJ, pp. 225–240.

Creelman, C.D., 1962. Human discrimination of auditory duration. Journal of the Acoustical Society of America 34, 582–593.

Eisler, H., 1975. Subjective duration and psychophysics. Psychological Review 82, 429–450.

Gibbon, J., 1977. Scalar expectancy theory and Weber's law in animal timing. Psychological Review 84, 279–325.

Gibbon, J., 1981. On the form and location of the bisection function for time. Journal of Mathematical Psychology 24, 58–87.

Gibbon, J., Church, R.M., Meck, W.H., 1984. Scalar timing in memory. In: Gibbon, J., Allan, L.G. (Eds.), Annals of the New York Academy of Sciences: Timing and Time Perception, vol. 423. New York Academy of Sciences, New York, pp. 52–77.

Killeen, P.R., Fetterman, J.G., 1988. A behavioral theory of timing. Psychological Review 95, 274–295.

Killeen, P.R., Fetterman, J.G., Bizo, L.A., 1997. Time's causes. In: Bradshaw, C.M., Szabadi, E. (Eds.), Time and Behaviour: Psychological and Neurobehavioural Analyses. Elsevier, Amsterdam, pp. 79–139.

Thomas, E.A.C., Cantor, N.E., 1978. Interdependence between the processing of temporal and non-temporal information. In: Requin, J. (Ed.), Attention and Performance VII. Erlbaum, Hillsdale, NJ, pp. 43–62.

Treisman, M., 1963. Temporal discriminations and the indifference interval: implications for a model of the internal clock. Psychological Monographs 576, 1–31.

Zakay, D., Block, R.A., 1994. An attentional-gate model of prospective time estimation. In: Richelle, M., DeKeyser, V., d'Ydewalle, G., Vandierendonck, A. (Eds.), Time and the Dynamic Control of Behavior. Université de Liège, Liège, Belgium, pp. 167–178.

Time Series: Advanced Methods

Wai Keung Li, The University of Hong Kong, Hong Kong
Howell Tong, The London School of Economics, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15699–15704, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Recent developments in the time-domain analysis of time series are reviewed. The concept of dynamical systems serves as a unifying theme of the review. We consider first methods for the modelling of the drift component or the conditional mean of a time series. This includes the class of threshold models and its variants. We then consider methods for the modelling of the diffusion component or the conditional variance of a time series which includes the popular generalized autoregressive conditional heteroscedastic G(ARCH) models and the stochastic volatility (SV) models. Hybrid models for the modelling of both the drift and the diffusion are then introduced. Long memory and discrete-valued time series models are also included. The main focus is on univariate series although multivariate series are also mentioned where appropriate.

Introduction

Many social phenomena can be described, at least approximately, by a discrete-time dynamical system, which can be expressed as a difference equation,

$$Y_t = F(X_{t-1}) \qquad [1]$$

where $\{Y_t\}$ is a sequence of m dimensional observations and X_t is a d' dimensional state vector, with $t = 1, 2, \ldots$ and $1 \le m' \le m$. In many applications, X_t stands for the past values of Y_t. Depending on the functional form $F(\cdot)$, the difference eqn [1] can exhibit a variety of dynamical features such as limit cycles, time irreversibility, and periodic changes (Tong, 1983, 1990). However, in the real world, stochastic noise is often present and therefore it is more realistic to consider a dynamical system with noise such as:

$$Y_t = F(X_{t-1}, a_t; \theta) \qquad [2]$$

where $\{a_t\}$ is often assumed to be a sequence of independent identically distributed (*i.i.d*) (possibly vector valued) random variables and θ is a vector of parameters. For instance, if $Y_t = X_t$ with $m = 1$ and $F(\cdot)$ is a scalar valued function, say, $f(X_{t-1}; \theta) + a_t$, then (2) takes the form:

$$X_t = f(X_{t-1}; \theta) + a_t \qquad [3]$$

which is a nonlinear autoregressive process of order one. Clearly eqn [3] can be generalized to include more past values of X_t and then we have

$$X_t = f\left(X_{t-1}, X_{t-2}, \ldots, X_{t-p}; \theta\right) + a_t \qquad [4]$$

which is a general nonlinear autoregressive process of order p. An important special case of eqn [4] is when $f(\cdot)$ is linear in the past values of X_t and we have then

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + a_t \qquad [5]$$

which is the usual linear autoregressive model of order p. Depending on $f(\cdot)$ or more generally $F(\cdot)$, the number of nonlinear models is clearly unlimited and the choice of a model is typically dictated by the situation and the ability of the model to describe/predict a social phenomenon.

It is clear that the dynamical system formulations eqn [2] and eqn [4] can be extended by considering a specification of the variance of the noise term a_t. Thus,

$$X_t = f(X_{t-1}; \theta) + g(X_{t-1}; \theta) a_t \qquad [6]$$

where $g(\cdot)$ is positive. Now, X_t is conditionally heteroscedastic. In other words, model eqn [6] now consists of both specifications for the mean (the drift) and for the conditional variance (the diffusion). Below we consider some advanced models in time series which have been developed and found to be useful since around 1980. For simplicity, we let $Y_t = X_t$ and a_t be a sequence of *i.i.d.* scalar random variables.

Nonlinear Time Series for the Mean

The Class of Threshold Models

Proposed by Tong in the later 1970s, the threshold models are a natural generalization of the linear autoregression eqn [5] by allowing different linear autoregressive specification over different parts of the state space. For example, a two-regime threshold autoregressive model of order p_1 and p_2 may be defined as follows:

$$X_t = \begin{cases} \phi_0^{(1)} + \phi_1^{(1)} X_{t-1} + \ldots + \phi_{p_1}^{(1)} X_{t-p_1} + a_t^{(1)}, & \text{if } X_{t-d} \le c \\ \phi_0^{(2)} + \phi_1^{(2)} X_{t-1} + \ldots + \phi_{p_2}^{(1)} X_{t-p_2} + a_t^{(2)}, & \text{otherwise} \end{cases} \qquad [7]$$

where $1 \le d \le \max(p_1, p_2)$, $\{a_t^{(i)}\}$ are two *i.i.d.* noise sequences with mean zero and variance σ_i^2, $i = 1, 2$, $\{a_t^{(1)}\}$ and $\{a_t^{(2)}\}$ are also independent of each other. We call c the threshold parameter and d the delay parameter. Equation [7] is called a self-exciting threshold autoregressive (SETAR $(2; p_1, p_2)$) model. Generalizations to more than two regimes are immediate. Stationarity and ergodicity conditions for eqn [7] when $p_1 = p_2 = 1$ and $\phi_0^{(i)} = 0$, $i = 1, 2$ have been obtained while a sufficient condition for the general SETAR $(2; p, p)$ model is available (Tong, 1990).

Estimation of eqn [7] can be easily done using the conditional least squares method given the parameters p_1, p_2, c, and d. Identification of p_1, p_2, c, and d can be done by the minimum Akaike information criterion (AIC) (Tong, 1990). For finite samples the corrected AIC or AICC is recommended (Wong and Li, 1998). The computer programme STAR 3 accompanying Tong (1990) provides a comprehensive set of modeling tools for threshold models. Diagnostic checking for model adequacy can be done using residual autocorrelations. The appropriate asymptotic distribution was derived in Li (1992). Tsay (1989) suggested an approach in the detection and modeling of threshold structures which is based on explicitly rearranging the least squares estimating equations using the order statistics of X_t, $t = 1,\ldots, n$, where n is the length of realization. (see Tong, 1990 for references.) A similar rearrangement was incorporated in the software STAR 3. Chen and Tsay (1993) considered a functional-coefficient autoregression model which has a very general threshold structure.

An easy-to-use statistic for detecting departure from linearity is the port-manteau test based on squared residual autocorrelations, the residuals being obtained from an appropriate linear autoregressive moving-average model fitted to the data (McLeod and Li, 1983). Of course, a general test statistic may not be optimal in terms of power when specific alternative hypotheses are considered. Tong (1990) has described other tests for nonlinearity due to Davies and Petruccelli, Keenan, Tsay and Saikkonen and Luukkonen, Chan and Tong. Petruccelli (1990) considered a comparison for some of these tests.

As a textbook-like example (albeit outside the social sciences), we consider the annual Canadian lynx trapping data in the MacKenzie River for the period 1821–1934. The nonlinearity of the data has been extensively documented by Tong (1990). Threshold nonlinearity was confirmed by applying the likelihood ratio test of Chan and Tong (1986) at the 1% percent level. Following other authors we transform the data by taking common log. We use the AICC as a criterion in selecting the best SETAR (2; p_1, p_2) model. The maximum possible value for p_1 and p_2 is 10, and the maximum possible value for the delay parameter d is 6. It is recommended that possible candidates of the threshold parameter can be chosen from a subset of the order statistics of the data. Following Wong (1998) we use 2.4378, 2.6074, 2.7769, 2.9464, 3.1160, 3.2855, and 3.4550, as potential values of the threshold parameter. The best fitting model using the minimum AICC criterion is the following SETAR (2; 4, 2) model,

$$\begin{aligned} X_t &= 1.0303 + 0.8959X_{t-1} - 0.0313X_{t-2} + 0.0406X_{t-3} \\ &\quad - 0.2001X_{t-4} + a_t^{(1)} \\ &\quad \text{if } X_{t-3} \le 2.9464, \\ X_t &= 1.6092 + 1.3487X_{t-1} - 0.8937X_{t-2} + a_t^{(2)}, \\ &\quad \text{if } X_{t-3} > 2.9464 \end{aligned}$$

where $a_t^{(1)}$ and $a_t^{(2)}$ have estimated variance equal to 0.0164 and 0.0642, respectively. The residual autocorrelation and squared residual autocorrelation show no significant values suggesting that the above model is adequate.

Instead of adrupt jumps between regimes in eqn [7] a smooth transition threshold autoregression was proposed by Chan and Tong (1986). Teräsvirta (1994) considered some further work in this direction. Continuous time threshold model was considered by Tong and Yeung (1991) with applications to water pollution data. Brockwell (1994) and others considered further work in the continuous time. Multivariate (mainly bivariate) threshold models were included in the seminal work of Tong in the 1980s and further developed by Tsay (1998). The relation between chaos and nonlinear time series is also treated in some detail in Tong (1990).

The Class of Bilinear Time Series Models

Bilinear models featured for a while in the early days of nonlinear time series modeling. In its simplest form, a single term bilinear model is given by

$$X_t = \beta X_{t-k} a_{t-l} + a_t \qquad [8]$$

where k, $l \ge 1$. Note that eqn [8] is linear in $X_{t-k}(a_{t-l})$ given $a_{t-l}(X_{t-k})$.

Granger and Andersen (1978) provided the first comprehensive introduction to these models. By deriving the second order autocovariances of eqn [8], they demonstrated that if $l < k$ then the time series cannot be distinguished from an uncorrelated noise process. The general bilinear autoregressive moving average model of order (p, q, P, Q) has the form

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{i=0}^{q} b_i a_{t-1} + \sum_{k=1}^{P} \sum_{l=1}^{Q} \beta_{kl} X_{t-k} a_{t-l} \qquad [9]$$

Granger and Andersen (1978: 31) denoted this by BARM (p, q, P, Q) and Subba Rao (1981) denoted this by $BL(p, q, P, Q)$. Subba Rao (1981) and Subba Rao and Gabr (1984) gave a general theory for model (9). Further results on causality and stationarity were given by Liu and Brockwell (1988). Identification (in the sense of estimation) of model (9) is in general difficult because of the enormous number of possibilities.

Models with Conditional Heteroscedasticity

One very popular class of such models is the autoregressive conditional heteroscedastic (ARCH) models of Engle (1982). The ARCH model of order one (ARCH(1)) is given by

$$X_t = \sqrt{h_t} a_t \qquad [10]$$

and $\sqrt{h_t} = g(X_{t-1}; \theta) = (\alpha_0 + \alpha_1 \alpha_{t-1}^2)^{\frac{1}{2}}$. Here $\alpha_0 > 0$, $\alpha_1 \ge 0$ and a_t is again *i.i.d.* with mean zero and unit variance. It can be shown that X_t is covariance stationary if $0 \le \alpha_1 < 1$. The unconditional variance of (6) is then $\alpha_0/(1-\alpha_1)$. The fourth moment of (6) exists if $\alpha_1^2 < \frac{1}{3}$. It is then greater than 3. For $r > 1$, the ARCH(r) model is defined by including more past values of X_t^2 so that X_t^2 X_t^2 X_t^2 $h_t \equiv E_t^2 \mid$ *past information* $= \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha$, X_{t-r}^2 ($x_0 > 0$, $\alpha_i \ge 0$ X_t^2 X_t^2 X_t^2 X_t^2). Again, X_t is covariance stationary if $\alpha_1 + \alpha_2 + \cdots + \alpha_r < 1$. It can be shown that the autocorrelation function of X_t^2 behaves like that of an AR(r) process. Milhøj (1985) derived the condition for the existence of the $2m$-th moments of X_t. Estimation of model 10 can easily be done by assuming a_t to be

standard normal so that X_t is conditionally normal with mean zero and variance h_t. The conditional log likelihood l is then given by

$$l = \sum_{t=1}^{n} l_t = \sum_{t=1}^{n} -\frac{1}{2}\log h_t - \frac{1}{2}\frac{X_t^2}{h_t}. \qquad [11]$$

Maximum likelihood (ML) estimation can be performed using Newton–Raphson or a method that does not require explicit derivatives. Assuming finite fourth moments, Weiss (1986) showed that the ML estimates are consistent and asymptotically normally disturbed. Many commercial statistical softwares such as the SAS now provide capabilities for ARCH modeling.

Diagnostic checking of the fit of ARCH models using standardized squared residual autocorrelations was considered in Li and Mak (1994). A fast method of detecting ARCH is by means of the port-manteau test of McLeod and Li (1983) using residuals from an autoregressive moving average model. As an illustration, we consider the daily closing price of the Hong Kong Hang Seng Index in the year 1980. The data are transformed by taking natural logarithm and differenced once to achieve stationarity. The transformed data R_t are called the returns of the index. There are 245 observations. The sample autocorrelations of R_t in **Figure 1** show basically no noticeable dependence structure in R_t itself. However, the sample autocorrelations of R_t^2 in **Figure 2** show a very strong dependence structure for R_t^2 suggesting the presence of an ARCH structure. Based on the partial autocorrelations of R_t^2 we consider an ARCH model of order 7 for R_t. Using the method of ML we have the following result:

$$R_t = \sqrt{h_t}a_t$$

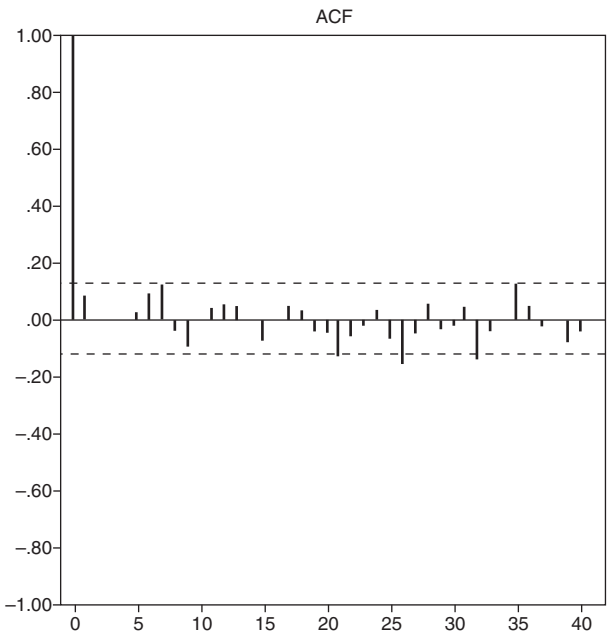

Figure 1 Sample autocorrelations of R_t.

with

$$h_t = 0.00012 + 0.3997R_{t-1}^2 + 0.13506R_{t-2}^2 + 0.12798R_{t-3}^2 + 0.15475R_{t-6}^2 + 0.28445R_{t-7}^2.$$

As reported in Li and Mak (1994) the sample autocorrelations of R_t^2 standardized by h_t show no further significant values and the above ARCH model is thus considered adequate.

Bollerslev (1986) generalized the ARCH(r) model to the GARCH(r, s) model, where

$$X_t = \sqrt{h_t}a_t$$

and

$$h_t = \alpha_0 + \sum_{i=1}^{r}\alpha_i X_{t-i}^2 + \sum_{i=1}^{s}\beta_i h_{t-i} \qquad [12]$$

with $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$.

The condition for covariance stationarity is $\sum_{i=1}^{r}\alpha_i + \sum_{i=1}^{s}\beta_i < 1$. Estimation of model 12 is similar to that of the ARCH models and exogeneous variables can be easily included. The second order moment properties of X_t^2 are like those of an ARMA(p, s) where $p = \max(r, s)$. The GARCH models have been very popular in the modeling of volatility in the finance literature. Bollerslev et al., (1992, 1994) provided two extensive reviews of the GARCH literature. Gourieroux (1997) and Bera and Higgins (1995) gave a more advanced coverage including multivariate ARCH models. The work edited by Engle (1995) has included many of the important contributions to the ARCH literature.

Models with Long Memory

Many stationary dynamical systems under model 4 in fact lead to $\{X_t\}$ which are short memory in the sense that the moduli of the autocovariances $C(k) = \mathrm{cov}(X_t, X_{t-k})$, $k = 1, 2, \ldots$ sum to a finite value. However, in some applications it is possible that $\sum_{k=0}^{\infty}|C(k)| = \infty$. A time series with this property is said to have the long memory property. One approach to modeling the long memory property is by means of fractional differencing. Suppose that B is the backshift operator, $BX_t = X_{t-1}$. Then a long memory dynamical system can be constructed by considering the process satisfying (Granger and Joyeux, 1980),

$$(1-B)^d X_t = a_t \qquad [13]$$

where d is a non-integer and

$$(1-B)^d = \sum_{k=0}^{\infty}\binom{d}{k}(-1)^k B^k$$

The process is stationary and invertible if $-1/2 < d < 1/2$ and possesses the long memory property if $0 \leq d < 1/2$. The fractionally integrated autoregressive moving average (FARIMA) process of order (p, d, q) satisfies Hosking (1981), and Granger and Joyeux (1980),

$$\phi(B)(1-B)^d X_t = \theta(B)a_t \qquad [14]$$

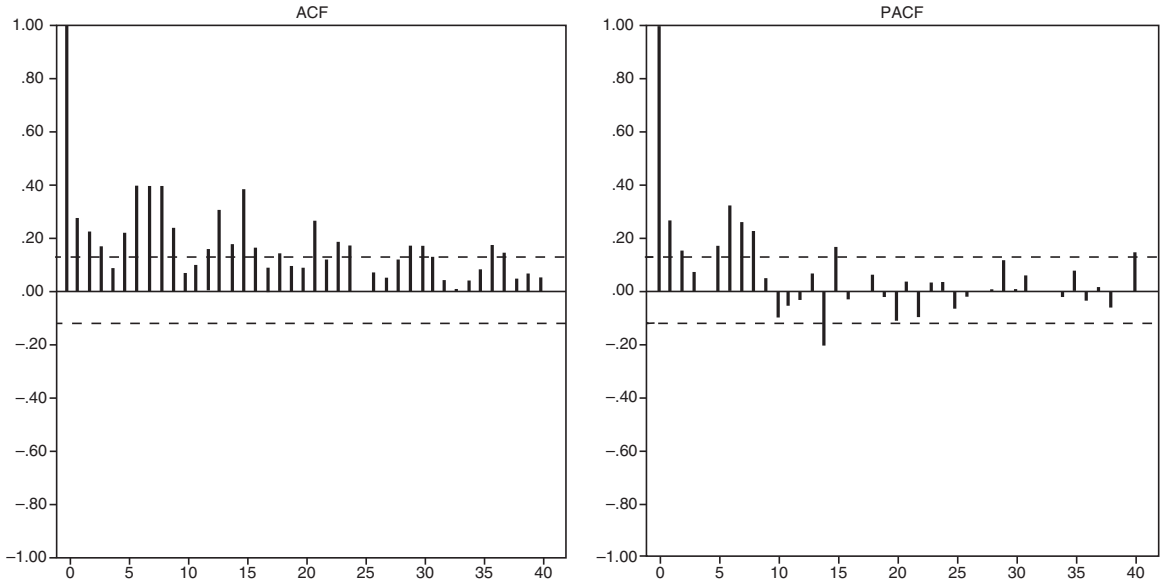

Figure 2 Sample autocorrelations of R_t^2.

where

$$\phi(B) = 1 - \phi_1 B \ldots - \phi_p B^p, \theta(B) = 1 - \theta_1 B \ldots - \theta_q B^q$$

and d is as in eqn [13]. It is further assumed that ϕ(B) and θ(B) have all roots outside the unit circle, which implies that X_t is stationary and invertible. Hosking (1981) showed that the autocovariances $C(k)$ of X_t are of the order $|k|^{2d-1}$ so that unlike ARMA process, $C(k)$ decays hyperbolically rather than exponentially. The spectral density $f(\lambda)$ of the process can be written in the form

$$f(\lambda) = \frac{\sigma_a^2}{2\pi}|1 - e^{i\lambda}|^{-2d} . \frac{\left|\theta\left(e^{i\lambda}\right)\right|^2}{\left|\phi\left(e^{i\lambda}\right)\right|^2}, \qquad [15]$$

where σ_a^2 is the variance of a_t. Beran (1994) contained a detailed exposition of the subject while Brockwell and Davis (1991) gave a good introduction including least squares and ML estimation.

Second Generation Models

Tong (1990) suggested that many of the aforementioned models can be combined to form new nonlinear models. He called these second generation models. For example, Tong (1990) suggested the possibility of a threshold AR model with ARCH. His idea was applied to real data by Li and Lam (1995). Li and Li (1996) extended this further to the double threshold ARCH or DTARCH models. These models allow for a threshold structure in both the conditional mean and the conditional variance. Clearly the number of possibilities is unlimited and one can consider models like double threshold long memory models with GARCH.

Discrete Valued Time Series

In many applications, the time series under consideration may be integer valued or may possess a discrete marginal distribution. An obvious approach to generate discrete random variables is by means of a possibly higher order Markov chain. Such is the approach of Raftery (1985) and Pegram (1980). A more elegant approach is that of McKenzie (1986) which is based on the method of thinning. Böckenholt (1999) considered a mixed integer-valued autoregressive Poisson regression with an application to category purchase data. Discrete-valued time series may also be generated using a conditional generalized linear model as in Zeger and Qaqish (1988). Li (1991) considered the goodness of fit problem for this approach. The recent work of MacDonald and Zucchini (1997) gave a valuable and timely survey of this vast area.

Conclusion

There is certainly a myriad of nonlinear models to consider when one decides to leave the restriction of the linear domain. What models to consider clearly depends on the goal and nature of the investigation. In many cases, flexibility in describing a variety of features such as time irreversibility, limit cycles, and autocorrelation structure is an important criterion. The availability of handy softwares is another important factor as illustrated by the examples considered. In many cases, automatic model selection criteria such as AIC and BIC, appropriately used, help to reduce the number of choices to a manageable few and so do model checking methods. The final test of a model in time series, in many applications is, of course, its ability to predict the future accurately.

Bibliography

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5 (1), 31–56.

Bera, A.K., Higgins, M.L., 1995. A survey of ARCH models: properties, estimation and testing. In: Oxley, L., George, D.A.R., Roberts, C.J., Sayer, S. (Eds.), Surveys in Econometrics. Blackwell, Oxford, UK, pp. 215–272.

Beran, J., 1994. Statistics for Long-Memory Processes. Chapman and Hall, London.
Böckenholt, U., 1999. Mixed INAR(1) poisson regression models: analyzing heterogeneity and serial dependencies in longitudinal count data. Journal of Econometrics 89, 317–338.
Bollerslev, T., Chou, R.Y., Kroner, K.F., 1992. ARCH modeling in finance: a review of the theory and empirical evidence. Journal of Econometrics 52, 5–59.
Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, 4. North-Holland, Amsterdam.
Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.
Brockwell, P.J., Davis, R.A., 1991. Time Series Theory and Methods, second ed. Springer-Verlag, New York.
Brockwell, P.J., 1994. On continuous time threshold ARMA processes. Journal of Statistical Planning and Inference 39, 291–304.
Chan, K.S., Tong, H., 1986. On estimating thresholds in autoregressive models. Journal of Time Series Analysis 7, 178–190.
Chan, K.S., Tong, H., 1990. On likelihood ratio tests for threshold autoregression. Journal of the Royal Statistical Society 52, 469–476.
Chen, R., Tsay, S., 1993. Functional-coefficient autoregressive models. Journal of the American Statistical Association 88, 298–308.
Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrika 50, 987–1008.
Engle, R.F., 1995. ARCH: Selected Readings. Oxford University Press, Oxford, UK.
Fan, J., Yao, Q., 2002. Nonlinear Time Series. Springer, New York, NY.
Ghil, M., Allen, M.R., Dettinger, M.D., Ide, K., Kondrashov, D., Mann, M.E., Robertson, Andrew W., Saunders, A., Tian, Y., Varadi, F., 2002. Advanced spectral methods for climatic time series. Reviews of Geophysics 40 (1), 3-1–3-41.
Gourieroux, C., 1997. ARCH: Models and Financial Applications. Springer, New York.
Granger, C.W.J., Andersen, A.P., 1978. An Introduction to Bilinear Time Series Models. Göttingen. Vandenhoeck and Ruprecht, Germany.
Granger, C.W.J., Joyeux, R., 1980. An introduction to long-range time series models and fractional differencing. Journal of Time Series Analysis 1, 15–30.
Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68, 165–176.
Li, C.W., Li, W.K., 1996. On a double threshold autoregressive heteroscedastic time series model. Journal of Applied Econometrics 11, 253–274.
Li, W.K., Lam, K., 1995. Modeling asymmetry in stock returns by a threshold autoregressive conditional heteroscedastic model. Journal of Royal Statistical Society D 44, 333–341.
Li, W.K., Mak, T.K., 1994. On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. Journal of Times Series Analysis 15, 627–636.
Li, W.K., 1991. Testing model adequacy for some Markov regression models for time series. Biometrika 78, 83–89.
Li, W.K., 1992. On the asymptotic standard errors of residual autocorrelations in nonlinear time series modeling. Biometrika 79, 435–437.
Liu, J., Brockwell, P.J., 1988. On the general bilinear time series model. Journal of Applied Probability 25, 553–564.
MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-valued Time Series. Chapman and Hall, London.
McKenzie, E., 1986. Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. Advances in Applied Probability 18, 679–705.
McLeod, A.I., Li, W.K., 1983. Diagnostic checking ARMA time series models using squared residual autocorrelations. Journal of Time Series Analysis 4, 269–273.
Milhøj, A., 1985. The moment structure of ARCH processes. Scandinavian Journal of Statistics 12, 281–292.
Pegram, G.G.S., 1980. An autoregressive model for multilag Markov chains. Journal of Applied Probability 17, 360–362.
Percival, D.B., Walden, A.T., 2006. Wavelet methods for time series analysis. Cambridge University Press, Cambridge, UK.
Petruccelli, J.D., 1990. A comparison of tests for SETAR-type non-linearity in time series. Journal of Forecasting 9, 25–36.
Raftery, A.Z., 1985. A model for high-order Markov chains. Journal of Royal Statistical Society B 47, 528–539.
Subba Rao, T., Gabr, M.M., 1984. An introduction to bispectral analysis and bilinear time series models. In: Lecture Notes in Statistics, 24. Springer, New York.
Subba Rao, T., 1981. On the theory of bilinear time series models. Journal of Royal Statistical Society B 43, 244–255.
Teräsvirta, T., 1994. Specification, estimation and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association 89, 208–218.
Tong, H., Yeung, I., 1991. Threshold autoregressive modeling in continuous time. Statistica Sinica 1, 411–430.
Tong, H., 1983. Threshold models in non-linear time series analysis. In: Lecture Notes in Statistics, 21. Springer-Verlag, New York.
Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Clarendon Press, Oxford, UK.
Tsay, R.S., 1989. Testing and modeling threshold autoregressive processes. Journal of American Statistical Association 84, 231–240.
Tsay, R.S., 1998. Testing and modeling multivariate threshold models. Journal of the American Statistical Association 93, 1188–1202.
Weiss, A.A., 1986. Asymptotic theory for ARCH models: estimation and testing. Econometric Theory 2, 107–131.
Wong, C.S., Li, W.K., 1998. A note on the corrected Akaike information criterion for the threshold autoregressive models. Journal of Time Series Analysis 19, 113–124.
Wong, C.S., 1998. Statistical inference for some nonlinear time series model. Ph.D. thesis. University of Hong Kong, Hong Kong.
Zeger, S.L., Qaqish, B., 1988. Markov regression models for time series: a quasi-likelihood approach. Biometrics 44, 1019–1031.

Time Series: ARIMA Methods

George C Tiao, University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 23, pp. 15704–15709, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Time series data in business, economics, environment, medicine, and other scientific fields tend to exhibit patterns such as trends, seasonal fluctuations, irregular cycles, and occasional shifts in level or variability. The objectives of analyzing such series are often to extrapolate the dynamic pattern in the data for forecasting future observations, to estimate the effect of known exogenous interventions, and to detect unsuspected interventions. This article sketches some useful methods for achieving these objectives. Serial dependence in time series data can often be adequately approximated by a linear dynamic model. We focus on the development of a class of parsimonious linear models, the mixed autoregressive moving average (ARMA) models, which are widely used to represent stationary and nonstationary series. For unit root nonstationarity, this leads to the class of autoregressive integrated moving average (ARIMA) models. An iterative model building approach proposed by Box and Jenkins, consisting of tentative model specification, efficient estimation, and diagnostic checking, is discussed and illustrated by an actual example. The ARIMA models are extended to represent seasonal patterns in the data, and to include an additive component for the dynamic effect of interventions. The correspondence between the ARMA models and the GARCH models commonly used in analyzing financial time series data with heterogeneous variances is illustrated. Finally, some extensions to nonlinear time series models, in particular the threshold autoregressive models, are presented.

Objectives of Time Series Analysis

Data in business, economics, engineering, environment, medicine, and other fields of scientific investigations are often collected in regular time intervals. Such data are called time series. Examples are hourly temperature readings, daily stock prices, weekly traffic volume, monthly money supply, quarterly GNP (gross national product), annual growth rates, etc. Time series data tend to exhibit patterns such as trends, seasonal fluctuations, irregular cycles, and occasional shifts in level or variability. The data are, therefore, dynamically or serially correlated.

In practice, the main objectives of analyzing time series are: (1) to ascertain and extrapolate the dynamic pattern in the data for forecasting future observations, (2) to assess the effect of exogenous interventions known to have occurred at given time points, and (3) to detect unknown and unsuspected interventions. For example, one may use the available weekly sales of frozen orange juice over the last 3 years to forecast the sales for the next 6 weeks and to assess the impact of a major promotion campaign initiated 4 weeks ago. As another example, an air pollution data analyst may discover a sudden drop in the time series of monthly ambient carbon monoxide concentration level and find out that the drop can be attributed to an unsuspected change in the measuring instrument. We here sketch some useful methods for achieving these objectives.

Stochastic Structure of Time Series Observations

Let $\ldots, t-1, t, t+1, \ldots$ denote equally spaced time points and y_t the observation at time t. The probability model generating the series $\{y_t\}$ is called a *stochastic process*, and the observed time series data are then a realization of the process. It is important to note a key distinction between time series observations and observations generated from designed experiments. For time series, there is usually one and only one realization at a given time point t, while because experiments can be repeated, there can be many realizations under the 'same' experimental conditions. Thus, statistical inference on time series depends critically on the nature of the serial dependence of the stochastic process under study.

Stationary versus Nonstationary Processes

When the joint probability distribution of any k observations $y_{t+1}, \ldots, y_{t+k}$ from a stochastic process remains the same for different values of t, the process is said to be 'stationary.' In practice, this means that there is a state of equilibrium in which the overall behavior of the time series stays roughly the same over time. When this condition does not hold, the series is said to be nonstationary. The theory of stationary processes has been well developed and plays an important role in time series analysis. Although many of the time series encountered in practice exhibit a nonstationary behavior, it is frequently the case that a suitable transformation such as differencing the data will render them stationary so that the theory of stationary processes can be applied.

To illustrate, Figure 1 shows the time series of monthly interest rates of 90-day Treasury bills from January 1985 to December 1993, and Figure 2 the series of month to month changes of the rates (first differences). The interest rate series exhibits a drifting or wandering behavior without a stable level. On the other hand, the series of the first differences seems to fluctuate about a fixed mean level with constant variance over the observational period. This example shows that a drifting nonstationary series can be transformed into a stationary one

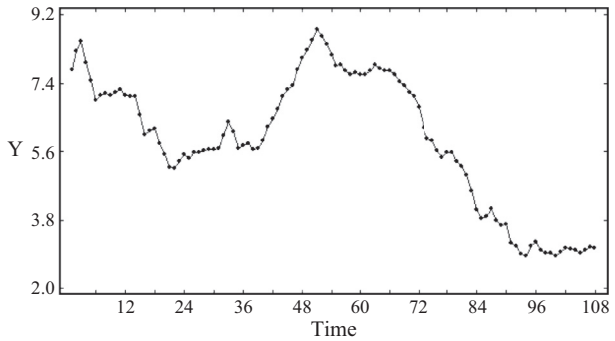

Figure 1 Monthly interest rates of 90-day Treasury bills, 1985–93.

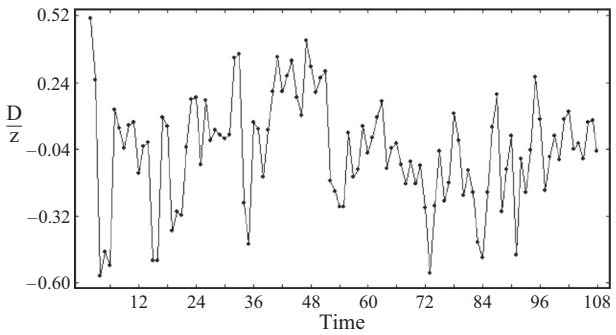

Figure 2 Month to month changes of the interest rates.

by the differencing operation. In fact, financial time series such as stock prices, prices of derivatives, and exchange rates often behave in this manner.

Linear Dynamic Models

The serial dependence in the time series $\{y_t\}$ can often be adequately approximated by a linear model of the form.

$$y_t = c + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + \pi_j y_{t-j} + \ldots + a_t \qquad [1]$$

where c is a constant, π_1, π_2,… are parameters relating y_t linearly to its past values, and a_t is the innovation or shock hitting the system at time t. It is commonly assumed that $\{a_t\}$ is a series of identically and independently distributed normal random variables with zero mean and common variance σ^2. The linear dynamic model in eqn [1] can also be expressed in the alternative form

$$y_t = c^* + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} \ldots + \psi_j a_{t-j} + \ldots \qquad [2]$$

where c^* is a constant, and the two parameter sets (π_1, π_2,…) and (ψ_1, ψ_2,…) are functionally related. In eqn [2] ψ_j measures the effect of the shock a_{t-j} hitting the system j periods ago on the observation y_t. The ψ_j's are called impulse responses or the memory of the system. The two representations in eqns [1] and [2], although equivalent, highlight two distinctive features of the linear system.

For forecasting future observations at a given time origin T, let us consider first the next observation y_{T+1}. By putting $t = T + 1$ in eqn [1], it is easy to show that the best (minimum mean squared error) predictor of y_{T+1} is a linear combination of the available observations $(y_T, y_{T-1}, \ldots)$, $\widehat{y}_{T+1} = c + \pi_1 y_T + \pi_2 y_{T-1} + \ldots$, and the associated forecast error is simply $y_{T+1} - \widehat{y}_{T+1} = a_{T+1}$ having variance σ^2. It is straightforward to deduce from eqn [1] best predictors of observations y_{T+l}, $\ell = 2, 3, \ldots$ periods ahead as linear functions of (y_T, y_{T-1},…) and their associated forecast errors as linear combinations of future shocks (a_{T+1},…, $a_{T+\ell}$).

ARMA and ARIMA Models

In reality, the model in eqn [1] or [2] is not known and has to be built from available data. Since y_t can in principle relate to its remote past, or to the remote past of a_t, for parsimony it is useful to make the π's or ψ's functions of a small set of parameters which can then be accurately estimated from the data. A class of such models originally proposed by Yule and Slutsky in the 1920s is the (mixed) autoregressive moving average (ARMA) models of orders (P, Q).

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_P y_{t-P} = c + a_t - \theta_1 a_{t-1} - \ldots - \theta_Q a_{t-Q} \qquad [3]$$

This class of models has been found useful to represent a variety of stationary, nonstationary, and seasonal time series in diverse scientific fields. By straightforward substitutions, one can readily deduce the π's and ψ's in eqns [1] and [2] as functions of the finite parameter set (ϕ_1,…, ϕ_P; θ_1,…, θ_Q) in eqn [3].

To illustrate, consider first the case ($P = 0$, $Q = 1$) so that $y_t = c + a_t - \theta_1 a_{t-1}$, which is the first-order moving average model. This model says that when the shock a_t hits the system at time t it will affect not only the observation y_t but also the next observation y_{t+1} (since $y_{t+1} = c + a_{t+1} - \theta_1 a_t$). Thus, the effect of the shock a_t will last over the two periods t and $t + 1$ and dissipate afterward. This short memory stationary model has been found useful to represent time series of various financial returns.

Consider next the case ($P = 1$, $Q = 0$) so that the model is $y_t = c + \phi_1 y_{t-1} + a_t$, which is the first-order autoregressive model frequently discussed in the economics literature. This model is stationary when $|\phi_1| < 1$. It can be expressed in the alternative memory form $y_t = c^* + a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \mathrm{L}$ showing that the effect or memory of the shock a_t will persist over many periods but will gradually dissipate when the model is stationary. In this case, the l-step ahead forecast will converge to the mean level of the series as the horizon l increases. For ($c = 0$, $\phi_1 = 1$), this model is a nonstationary 'random walk' widely believed to be appropriate for stock prices in an 'efficient market.'

As a final illustration, for ($c = 0$, $P = Q = 1$, $\phi_1 = 1$) we have the model $y_t - y_{t-1} = a_t - \theta_1 a_{t-1}$. For this model, the left-hand side is the first difference of y_t implying that the series $\{y_t\}$ is nonstationary but the first difference series is stationary and follows the first-order moving average model. At the time origin T, the l-step ahead predictor is, for all $\ell \geq 1$, $\widehat{y}_{T+\ell} = (1 - \theta_1) y_T + \theta_1 (1 - \theta_1) y_{T-1} + \theta_1^2 (1 - \theta_1) y_{T-2} + \ldots$, which is an exponentially weighted average of current and past observations. Predictors of this kind which give more weight to recent information and less and less weight to information from the remote past seem intuitively

appealing because time series data, especially in business and economics, tend to exhibit nonstationary behavior over time. As a result, the 'exponential smoothing' forecast discussed here has been the most widely applied forecasting technique in practice. It will be seen later that the Treasury bills data shown in Figure 1 can be adequately represented by this model.

To sharpen the distinction between stationary and nonstationary models, we can express the ARMA (P, Q) models in eqn [3] alternatively as.

$$\begin{aligned}\nabla^d y_t - \varphi_1 \nabla^d y_{t-1} - \ldots - \varphi_p \nabla^d y_{t-p} \\ = c + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}\end{aligned} \quad [4]$$

where $d = 0, 1, 2, \ldots, \nabla^d y_t$ is the dth difference of y_t (i.e., $\nabla^d y_t = y_t$ if $d = 0$, $\nabla^d y_t = y_t - y_{t-1}$ if $d = 1$, $\nabla^d y_t = y_t - 2y_{t-1} + y_{t-2}$ if $d = 2$, etc.), $P = p + d, Q = q$, the two parameter sets $(\phi_1, \ldots, \phi_P)$ and $(\varphi_1, \ldots, \varphi_p)$ are functionally related, and we shall assume that the ARMA model for $\nabla^d y_t$ is stationary. This is saying that the time series $\{y_t\}$ is such that its dth difference $\nabla^d y_t$ is stationary and follows a mixed autoregressive moving average model. For $d = 0$, y_t itself is stationary; and for $d > 0$, y_t is nonstationary. The models in eqn [4] are called autoregressive integrated moving average (ARIMA) models of orders (p, d, q) for the time series y_t. Here d signifies the order of differencing to make the transformed series stationary, p is the number of lagged values of $\nabla^d y_t$, and q is the number of lagged terms of a_t in the model.

It should be noted that for the ARIMA models in eqn [4], the nonstationarity of the series y_t is solely accounted for by the difference operation $\nabla^d y_t$. While this is not the only kind of nonstationarity in the ARMA models in eqn [3], we shall, however, focus primarily on this class of models.

Model Building

To build an ARMA or ARIMA model for the data at hand, Box and Jenkins (1976) have proposed an iterative approach consisting of (a) tentative model specification, (b) efficient estimation, and (c) diagnostic checking. This approach has been widely adopted and in fact has revolutionized the use of time series models for forecasting in practice.

In the tentative specification phase, the goal is to employ computationally simple techniques to narrow down the range of parsimonious models worthy of more detailed further study. Techniques such as the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF) have been widely used. The SACF can help determine the order of differencing needed to make the transformed series stationary and pin down the order of a moving average model for stationary series. The SPACF, on the other hand, is useful to help decide the order of a stationary autoregressive model. The more recently proposed sample extended autocorrelation function (SEACF, see Tsay and Tiao (1984)) makes it possible to specify directly the orders of an ARMA (P, Q) model, stationary or nonstationary, and hence simplifies the procedure for model specification. In this case, the question of stationarity will be answered at the estimation stage.

Once a model is tentatively entertained, the parameters can then be efficiently estimated by the method of maximum likelihood. In addition to standard tests on the statistical significance of parameter estimates, unit root tests may provide a useful tool here to help determine the order of nonstationarity in the series if needed. Finally, residuals calculated from the fitted model are then subject to various diagnostic checks to determine the overall adequacy of the entertained model and to search for directions of improvement. Of particular importance are procedures to detect the existence of autocorrelations in the residuals, outliers and structural changes such as level shifts, variance changes, and nonlinearities.

To illustrate this model building approach, we consider the monthly interest rate data introduced earlier. Table 1 shows the SACF of the series y_t, the SACF and SPACF of the differenced series ∇y_t, and the SACF of the residuals $\hat{a}_t$ for lags 1–12.

In each of the four panels in Table 1, the first row gives the sample correlation values and the second their associated estimated standard errors. We now discuss tentative model specification. First, the persistently high values of the SACF of y_t over long lags suggest that the series is nonstationary, leading to considering the first difference of the series. Second, the SACF of ∇y_t has only one large value at lag 1, and this 'cutting off' property strongly indicates that a simple first-order moving average model would be adequate for the differenced series. On the other hand, the pattern of the SPACF of ∇y_t, showing two large values at lag 1 and lag 3, would suggest a more complex model.

Table 1 SACF and SPACF for the interest rate data

1.	SACF of y_t												
	Lags 1–12	0.97	0.94	0.89	0.85	0.81	0.77	0.73	0.68	0.64	0.59	0.53	0.48
	ST.E.	0.10	0.16	0.21	0.24	0.27	0.29	0.31	0.32	0.34	0.35	0.36	0.36
2.	SACF of ∇y_t												
	Lags 1–12	0.46	0.06	0.12	0.12	0.07	0.09	0.07	0.09	0.13	0.08	0.05	−0.02
	ST.E	0.10	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
3.	SPACF of ∇y_t												
	Lags 1–12	0.46	−0.18	0.22	−0.05	0.07	0.03	0.01	0.08	0.06	−0.02	0.03	−0.11
	ST.E.	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	SACF of $\hat{a}_t$												
	Lags 1–12	0.00	0.03	0.07	0.09	−0.00	0.08	0.03	0.03	0.11	0.02	0.03	0.02
	ST.E.	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

Table 2 Simplified SEACF for the interest rate data

($Q \rightarrow$)	0	1	2	3	4	5	6	7	8	9	10	11	12
($P = 0$)	X	X	X	X	X	X	X	X	X	X	X	O	O
($P = 1$)	X	O	O	O	O	O	O	O	O	O	O	O	O
($P = 2$)	X	X	O	O	O	O	O	O	O	O	O	O	O
($P = 3$)	X	X	O	O	O	O	O	O	O	O	O	O	O
($P = 4$)	O	X	O	O	O	O	O	O	O	O	O	O	O
($P = 5$)	X	O	X	O	O	O	O	O	O	O	O	O	O
($P = 6$)	X	O	O	O	O	O	O	O	O	O	O	O	O

Consequently we are led to consider the ARIMA (0, 1, 1) model.

$$\nabla y_t = a_t - \theta_1 a_{t-1} \qquad [5]$$

The above mode of model specification, while widely adopted in practice, has several drawbacks. First, judgment of whether to difference the data based on 'persistently high values of SACF' can often be difficult in practice. Second, for $d \geq 2$, at least three cycles of SACF and SPACF are needed before a model can be determined. Third, both SACF and SPACF are not designed to help specify stationary mixed ARMA (P, Q) models. The SEACF procedure is designed to address these difficulties. Table 2 gives the simplified SEACF for the interest rate data. In this two-way table, the rows correspond to the AR orders P and the columns the MA orders Q. An entry 'X' signifies that the sample estimate is over twice its estimated standard error in magnitude and a '0' otherwise. The goal is to search for a triangle of 0's just like the one shown in Table 2, and the coordinates of the vertex give the AR and MA orders. In our case, we have clear indication of an ARMA (1, 1) model.

$$y_t - \phi_1 y_{t-1} = c + a_t - \theta_1 a_{t-1} \qquad [6]$$

which reduces to the model in eqn [5] when $\phi_1 = 1$ and $c = 0$. The estimation results for both models in eqns [5] and [6] are summarized in Table 3. The values in parentheses are the associated estimated standard errors of the estimates. The results $\widehat{\phi}_1 \cong 1$ and $\widehat{c} \cong 0$ clearly show that the fitted models eqns [5] and [6] are essentially identical. Thus, instead of making the decision to difference the data series at the tentative model specification stage, the same result can be obtained at the estimation phase which seems more appealing. To provide a partial check of the adequacy of the model, the bottom panel of Table 1 shows the SACF of the residuals $\hat{a}_t$'s from the model in eqn [5]. All the sample autocorrelations are very small, indicating no dynamic pattern left in the residuals. Some additional residual diagnostics including the Box–Ljung statistic and an iterative outlier test have also been performed without revealing any serious model inadequacy. The fitted model can then be used to predict future observations if desired.

Table 3 Estimation results for the interest rate data

	$\widehat{\phi}_1$	$\widehat{c}$	$\widehat{\theta}_1$	$\widehat{\sigma}$
Model (5)			−0.597 (0.077)	0.198
Model (6)	0.993 (0.017)	−0.002 (0.107)	−0.592 (0.078)	0.196

Seasonal ARIMA Models

Business, economic, and atmospheric time series data often exhibit a strong, but not necessarily deterministic, seasonal behavior. It seems natural to think that the seasonal pattern arises simply because observations at the same time point of the year are dynamically related across the years. For monthly data, this means that observations 12, 24,… periods apart are correlated. Just as the class of ARIMA (p, d, q) models represent the dependence of observations in successive periods (months), we may construct similar models to account for dynamic dependence of observations in the same month of successive years. As an illustration, corresponding to the regular ARIMA (0, 1, 1) model in eqn [5], we may consider a seasonal model of the from $\nabla_{12} y_t = e_t - \Theta_{12} e_{t-12}$, where $\nabla_{12} y_t = y_t - y_{t-12}$ is the seasonal difference and e_t is the noise term. To account for both seasonal dependence as well as month to month dependence, we may make e_t to follow the ARIMA (0, 1, 1) model $\nabla e_t = a_t - \theta_1 a_{t-1}$. Together, we have that.

$$\nabla\nabla_{12} y_t = a_t - \theta_1 a_{t-1} - \Theta_{12}(a_{t-12} - \theta_1 a_{t-13}) \qquad [7]$$

where $\nabla\nabla_{12} y_t = y_t - y_{t-1} - (y_{t-12} - y_{t-13})$ is the double, regular and seasonal, difference of y_t. This 'doubly' exponential smoothing model, proposed by Box and Jenkins and motivated by their analysis of airline passenger data, has been widely used to present a variety of seasonal time series in practice. It is known as the 'airline model.' Similar extensions are readily obtainable for other ARIMA models.

Intervention Analysis and Outlier Detection

Time series observations are often subject to the influence of special events, or interventions, that occurred at known time points. Examples are the outbreak of armed conflicts, natural disasters, changes in fiscal policy, and major sales promotions. It has been found useful to employ linear dynamic models of the form.

$$Y_t - \delta_1 Y_{t-1} - \ldots - \delta_r Y_{t-r} = \omega_0 x_t - \ldots - \omega_s x_{t-s} \qquad [8]$$

where x_t is an indicator (dummy) variable for an intervention and Y_t represents its dynamic impact on the time series observations. This class of models can accommodate a very wide variety of dynamic patterns including isolated large values, level shifts, trend, oscillations, and cycles. We can write the model of the observations as the sum of the intervention component Y_t and a noise component N_t, $y_t = Y_t + N_t$, where

N_t follows an ARMA model of the form in eqn [3]. The parameters in the intervention and the noise components can then be jointly estimated to quantify the effects of these two components.

When the time of occurrence of interventions is not known or the interventions themselves are unsuspected, we have the problem of detecting unusual events or 'outliers' in time series. Estimation procedures developed for interventions can be readily adapted to become iterative procedures for outlier detection, and these methods, especially those for detecting isolated large values and level shifts, have been commonly adopted as part of diagnostic checking.

Time Series with Heterogeneous Variances

The ARMA model in eqn [3] supposes that the variance σ^2 of the innovation a_t stays constant over time. In reality, this variance homogeneity assumption is often violated, especially in highly disaggregated financial data such as hourly stock returns or exchange rates where clusters of high volatility tend to occur. Several models have been developed to describe this kind of phenomenon. Denoting a_t^2 as the variance of a_t, one may employ a random step-change model in which there is a nonzero probability λ that σ_t^2 may change at any given point of time. Alternatively, one may relate σ_t^2 to past values of a_t^2 and itself. In particular, the model.

$$\sigma_t^2 = \alpha_0 + \alpha_1 \alpha_{t-1}^2 + \beta_1 \alpha_{t-1}^2 \qquad [9]$$

which is called the generalized autoregressive conditional heteroscedastic (GARCH) model of orders (1, 1) has been extensively used in the analysis of financial time series. Now writing $a_t^2 = \sigma_t^2 + u_t$, the above model can be alternatively expressed as

$$a_t^2 = \alpha_0 + (\alpha_1 + \beta_1) a_{t-1}^2 + u_t - \beta_1 u_{t-1} \qquad [10]$$

It can be readily shown that the u_t's have zero mean and common variance, and are uncorrelated. Thus the model in eqn [10] has exactly the same autocovariance structure as an ARMA (1, 1) model in a_t^2. The above correspondence is readily extended to higher order models. Thus the GARCH models can be viewed as an interesting extension of the ARMA structure to the squared innovations a_t^2's.

Nonlinear Time Series Models

There is an extensive literature on nonlinear time series models where y_t is related to its past values more generally by a nonlinear function. Such models are necessarily more complex, but they can represent phenomena such as time irreversibility and limit cycles that cannot be adequately described by linear models. For example, it is well known that the dynamic behavior of economic indicators such as the GNP or the unemployment rate during a recession is quite different from that when the economy is expanding. To describe this kind of nonhomogeneous behavior, one may employ a piecewise linear model for different regimes of economic activity. To illustrate, let y_t represent the quarterly change in GNP. We may postulate a model such that.

$$y_t = \begin{cases} c^{(1)} + \phi^{(1)} y_{t-1} + a_t^{(1)}, & \text{if} \quad y_{t-1} < 0 \text{ and } y_{t-2} < 0 \\ c^{(2)} + \phi^{(2)} y_{t-1} + a_t^{(2)}, & \text{if} \quad \text{otherwise} \end{cases} \qquad [11]$$

where $c^{(1)} \neq c^{(2)}$, $\phi^{(1)} \neq \phi^{(2)}$, and $a_t^{(1)}$ and $a_t^{(2)}$ have unequal variances. In eqn [11] the regime ($y_{t-1} < 0$ and $y_{t-2} < 0$) represents a period of economic decline, and the model says that the dynamic behavior of y_t is different for the two regimes. This model is a special case of a class, known as the threshold autoregressive models, which have been used for time series data in atmospheric sciences, economics, and other scientific fields.

Further Reading

The reader interested in a deeper understanding of the basic concepts and methods of time series analysis may consult the books by Box et al. (1994), Brockwell and Davis (1996), Shumway and Stoffer (2000), and Wei (1990), and the book edited by Pena et al. (2001) which also contains several chapters on multivariate ARMA models. For further details on the GARCH models and the general topic of heterogeneous variances and volatility clustering in economic time series, see Hamilton (1994). Nonlinear time series models are discussed in Priestley (1988) and Tong (1990). Finally, for multivariate extension of linear time series models, the reader may study the books by Lutkepohl (1993), Reinsel (1993) and Reinsel and Velu (1998).

See also: Time Series: Advanced Methods; Time Series: Cointegration; Time Series: Economic Forecasting; Time Series: Seasonal Adjustment.

Bibliography

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.
Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis, Forecasting and Control, rev ed. Holden-Day, San Franscisco.
Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis, Forecasting and Control, third ed. Prentice-Hall, Englewood Cliffs, NJ.
Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70–79.
Brockwell, P.J., Davis, R.A., 1996. An Introduction to Time Series: Theory and Methods. Springer-Verlag, New York.
Chang, I., Tiao, G.C., Chen, C., 1988. Estimation of time series parameters in the presence of outliers. Technometrics 30, 193–204.
Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimates for autoregressive time series with a unit root. Journal of the American Statistical Association 74, 427–431.
Durbin, J., Koopman, S.J., 2012. Time Series Analysis by State Space Methods. Oxford University Press.
Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. Econometrica 50, 987–1008.
Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.
Kalpakis, K., Gada, D., Puttagunta, V., 2001. Distance measures for effective clustering of ARIMA time-series. In: Proceedings IEEE International Conference on Data Mining, 2001 (ICDM 2001). IEEE, San Jose, CA, pp. 273–280.

Lutkepohl, H., 1993. Introduction to Multiple Time-Series Analysis. Springer-Verlag, Berlin.
Pena, D., Tiao, G.C., Tsay, R.S. (Eds.), 2001. A Course in Time Series Analysis. Wiley, New York.
Priestley, M.B., 1988. Non-Linear and Non-Stationary Time Series Analysis. Academic Press, Orlando, FL.
Reinsel, G.C., 1993. Elements of Multivariate Time-Series Analysis. Springer-Verlag, New York.
Reinsel, G.C., Velu, R.P., 1998. Multivariate Reduced Rank Regression. Springer-Verlag, New York.
Shumway, R.H., Stoffer, D.A., 2000. Time Series and its Applications. Springer-Verlag, New York.
Tong, H., 1990. Nonlinear Time Series: A Dynamical System Approach. Oxford Science Publications, New York.
Tong, H., Lim, K.S., 1980. Threshold autoregression, limit cycles, and cyclical data (with discussion). Journal of the Royal Statistical Society, Series B 42, 245–292.
Tsay, R.S., Tiao, G.C., 1984. Consistent estimates of autoregressive parameters and extended autocorrelation function for stationary and nonstationary ARMA models. Journal of the American Statistical Association 79, 84–96.
Wei, W.W.S., 1990. Time Series Analysis. Addison-Wesley, Reading, MA.
Zhang, G.P., 2003. Time series forecasting using a hybrid Arima and neural network model. Neurocomputing 50, 159–175.

Time Series: Cointegration

Søren Johansen, University of Copenhagen, Copenhagen, Denmark; and Aarhus University, Aarhus, Denmark

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by N.H. Chan, volume 23, pp. 15709–15714, © 2001, Elsevier Ltd.

Abstract

An overview of results for the cointegrated vector autoregressive model for nonstationary $I(1)$ variables is given. The emphasis is on the analysis of the model and the tools for asymptotic inference. These include: formulation of criteria on the parameters, for the process to be nonstationary and $I(1)$, formulation of hypotheses of interest on the rank, the cointegrating relations and the adjustment coefficients. A discussion of the asymptotic distribution results that are used for inference. The results are illustrated by a few examples. A number of extensions of the theory are pointed out.

Introduction

The term cointegration was defined by Granger (1983) as a formulation of the phenomenon that nonstationary processes can have linear combinations that are stationary. It was his investigations of the relation between cointegration and error correction that brought modeling of vector autoregressions with unit roots and cointegration to the center of attention in applied and theoretical econometrics; see Engle and Granger (1987).

During the last 30 years, many have contributed to the development of theory and applications of cointegration. The account given here focuses on theory, more precisely on likelihood-based theory for the vector autoregressive model; see Johansen (1996). By building a statistical model as a framework for inference, one has to make explicit assumptions about the model used and hence has a possibility of checking the assumptions made before conducting inference.

We start with some examples of cointegration.

Example 1: As a simple economic example of the main idea in cointegration, consider 229 observations of US monthly interest rates in the period 1987:1 to 2006:1, which defines the period when Greenspan was the chairperson of the Federal Reserve System. The data are taken from IMF's financial database and consist of the 6 months treasury bill rate and the 3 years bond rate, denoted by i_{6m} and i_{3y}, respectively. In **Figure 1** we plot the two interests rates and their spread $i_{6m} - i_{3y}$. The expectations hypothesis implies that these interest rates should be equal up to a constant, $i_{3y} = i_{6m} + c$, and such a relation is not found in data. We can formulate it instead as their spread being stationary around a constant, possibly zero. This is an example of the formulation of an economic regularity as a cointegrating relation and we want below to analyze a statistical model, which allows such a formulation.

As simple examples of models for processes of this nature, we first consider a model for a random walk and a stationary process. Throughout we consider the sequence of p-dimensional errors ε_t, $t = 1,\ldots, T$, which are independent and identically distributed with mean zero and variance matrix, Ω.

Example 2: (Autoregressive processes) (**Figure 2**) Let $x_t = (y_t, z_t)'$ be given by the equations for $t = 1,\ldots,T$

$$y_t = y_{t-1} + \varepsilon_{yt} \quad [1]$$

$$z_t = \rho z_{t-1} + \varepsilon_{zt} \quad [2]$$

here $-1 < \rho < 1$. It is seen that $y_t = y_0 + \varepsilon_{y1} + \ldots + \varepsilon_{yt}$ and that $E(y_t|y_0) = y_0$ and $Var(y_t|y_0) = t\Omega$, so the variance is increasing and the process is nonstationary. We also find $z_t = \rho^t z_0 + \sum_{i=0}^{t-1}\rho^i\varepsilon_{z,t-i}$ which implies that $E(z_t|z_0) = \rho^t z_0$ and $Var(z_t|z_0) = \sigma_z^2\sum_{i=0}^{t-1}\rho^{2i} \to \sigma_z^2\sum_{i=0}^{\infty}\rho^{2i} = \sigma_z^2/(1-\rho^2)$. We can make z_t stationary by choosing $z_0 = \sum_{i=0}^{\infty}\rho^i\varepsilon_{z,-i}$ and then $z_t = \sum_{i=0}^{\infty}\rho^i\varepsilon_{z,t-i}$. We call y_t an $I(1)$ process and z_t an $I(0)$ process, see section Integration and Cointegration.

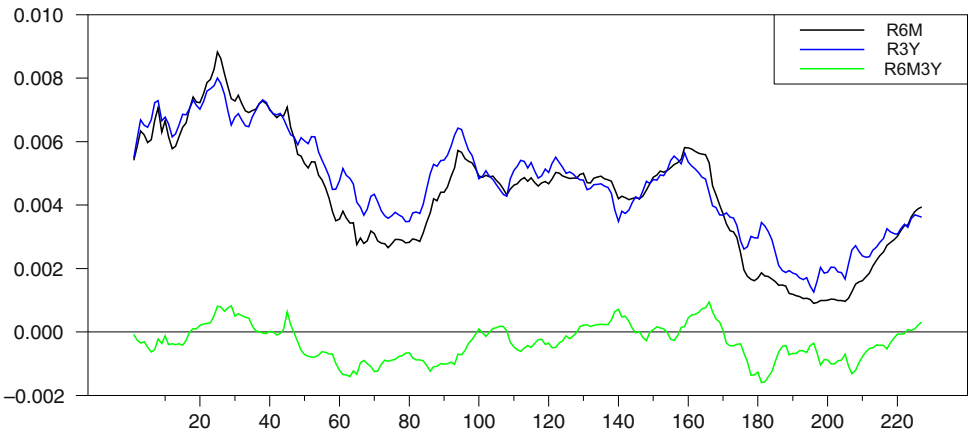

Figure 1 A plot of 229 monthly observations of the 6 months treasury bill rate and the 3 years bond rate and their spread $i_{6m} - i_{3y}$. Note the nonstationary behavior of the interest rates and the much more stationary spread, see Example 1.

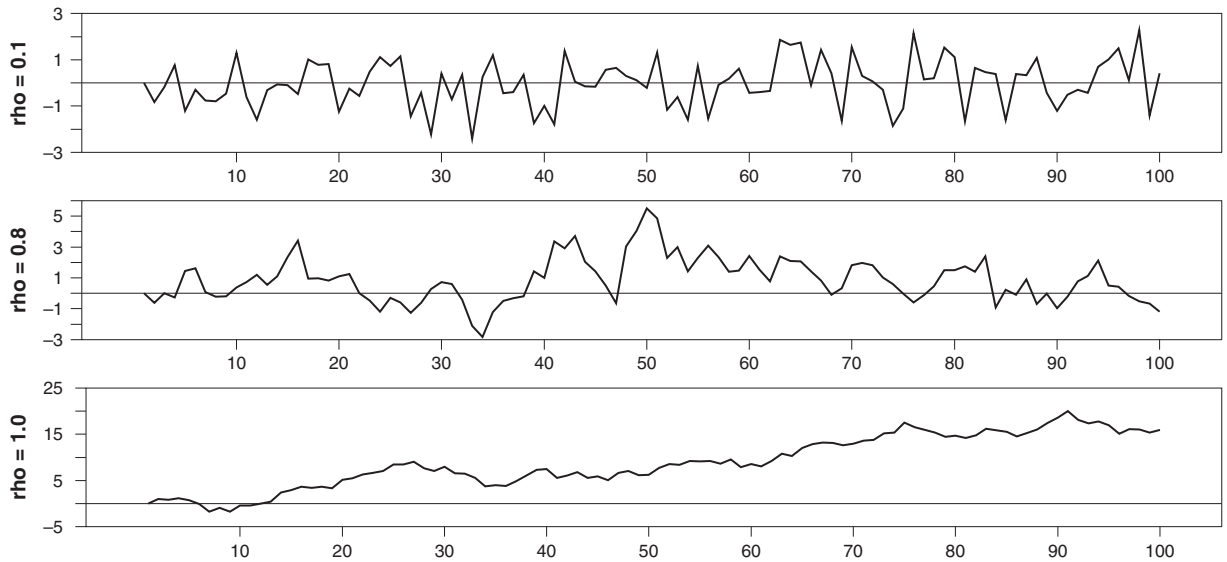

Figure 2 Three examples of $AR(1)$ processes. Each plot has 100 observations. The first two are stationary with $\rho = 0.1$ and $\rho = 0.5$, and the last is a random walk with $\rho = 1$, see Example 2.

Next we give a model for nonstationary variables that are cointegrated, using the notation, $\Delta x_t = x_t - x_{t-1}$.

Example 3: (Cointegrated processes) A bivariate process is given for $t = 1,\ldots,T$ by the equations

$$\begin{aligned}\Delta x_{1t} &= \alpha_1(x_{1t-1} - x_{2t-1}) + \varepsilon_{1t} \\ \Delta x_{2t} &= \alpha_2(x_{1t-1} - x_{2t-1}) + \varepsilon_{2t}\end{aligned} \qquad [3]$$

Subtracting the equations, we find that the process $y_t = x_{1t} - x_{2t}$ is autoregressive, see [2] and stationary if $\rho = 1 + \alpha_1 - \alpha_2$ satisfies $|\rho| < 1$, and the initial value is given by its invariant distribution. Similarly we find that $s_t = \alpha_2 x_{1t} - \alpha_1 x_{2t}$ is a random walk, see [1], so that the process $x_t = (x_{1t}, x_{2t})'$ is given by

$$x_{1t} = \frac{s_t - \alpha_1 y_t}{\alpha_2 - \alpha_1} \quad \text{and} \quad x_{2t} = \frac{s_t - \alpha_2 y_t}{\alpha_2 - \alpha_1}$$

This shows that when $|1 + \alpha_1 - \alpha_2| < 1$, x_t is $I(1)$, $x_{1t} - x_{2t}$ is stationary, and $\alpha_2 x_{1t} - \alpha_1 x_{2t}$ is a random walk $\sum_{i=1}^{t}(\alpha_2\varepsilon_{1i} - \alpha_1\varepsilon_{2i})$, so that x_t is a cointegrated $I(1)$ process with cointegration vector $\beta' = (1, -1)$. We call s_t a common stochastic trend and α, the adjustment coefficients (Figure 3).

Three Approaches to Cointegration

There are at present three different ways of modeling the linear cointegration idea in a parametric statistical framework. To illustrate the ideas they are formulated in the simplest possible cases, leaving out deterministic terms.

Regression Formulation

The multivariate process, $x_t = (x'_{1t}, x'_{2t})'$ of dimension $p = p_1 + p_2$ is given by the regression equations

$$\begin{aligned}x_{1t} &= \gamma' x_{2t} + \varepsilon_t \\ \Delta x_{2t} &= \varepsilon_{2t}\end{aligned}$$

This model implies that x_{2t} is a nonstationary random walk, and $x_{1t} - \gamma' x_{2t}$ gives p_1 stationary linear combinations. Hence, in this case, the cointegration rank of x_t is p_1, see section Integration and Cointegration. The first estimation method used in this model is least squares regression, Engle and Granger (1987), which is shown to give a superconsistent estimator by Stock (1987). This estimation method gives rise to residual based tests for cointegration. It was shown by Phillips and Hansen (1990) that, for a more general error term, a modification of the regression estimator gives useful methods for inference on coefficients of cointegration relations; see also Phillips (1991).

Autoregressive Formulation

In the rest of this contribution we focus on the autoregressive formulation of the p-dimensional process, x_t, defined by the equations

$$\Delta x_t = \alpha\beta' x_{t-1} + \varepsilon_t$$

where α and β are $p \times r$ matrices of rank r. Under the condition that Δx_t is stationary, the solution is

$$x_t = C\sum_{i=1}^{t}\varepsilon_i + \sum_{i=0}^{\infty} C_i\varepsilon_{t-i} + A \qquad [4]$$

where $C = \beta_\perp(\alpha'_\perp\beta_\perp)^{-1}\alpha'_\perp$ and $\beta' A = 0$. Here $\beta_\perp$ is a full rank, $p \times (p - r)$, matrix so that $\beta'\beta_\perp = 0$. This formulation allows for modeling of both the long-run relations, $\beta' x$, and the adjustment, or feedback coefficient α, toward the attractor set $\{x : \beta' x = 0\}$ defined by the long-run relations. Models for different cointegration ranks are nested and the smallest, for $\alpha = \beta = 0$, corresponds to p random walks. The rank can be analyzed by likelihood ratio tests. Methods usually applied for this analysis are derived from the Gaussian likelihood function, which is discussed here; see also Johansen (1988, 1996), and Ahn and Reinsel (1990).

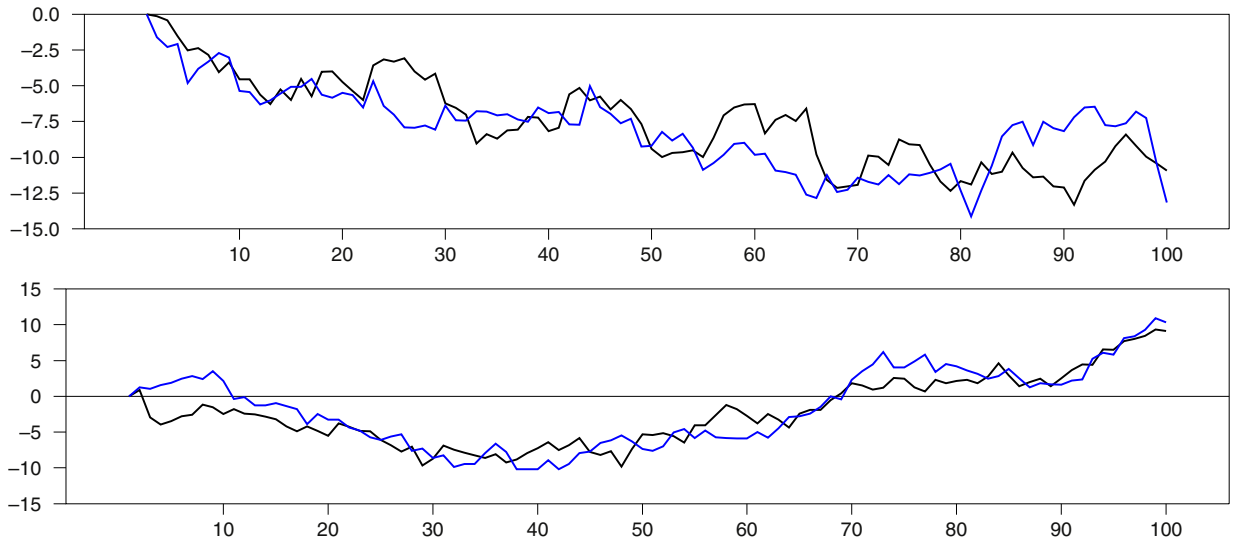

Figure 3 Two simulations of model [3] for cointegrated variables, see Example 3.

Unobserved Components Formulation

Let x_t be given by

$$x_t = \xi\eta' \sum_{i=1}^{t} \varepsilon_{1i} + \varepsilon_{2t}$$

where ε_{2t}, typically, is independent of ε_{1t}.

In this formulation too, hypotheses of different ranks are nested but in the opposite direction, and the smallest, for $\xi = \eta = 0$, corresponds to stationary processes. The parameters are linked to the autoregressive formulation by $\xi = \beta_\perp$ and $\eta = \alpha_\perp$, even though the linear process, $\sum_{i=0}^{\infty} C_i\varepsilon_{t-i}$, in [4] depends on the random walk part, so the unobserved components model and the autoregressive model are not the same. However, both adjustment and cointegration can be discussed in this formulation, and hypotheses on the rank can be tested. Rather than testing for unit roots, one tests for stationarity, which is sometimes a more natural formulation. Estimation is performed by the Kalman filter, and asymptotic theory of the rank tests has been worked out by Nyblom and Harvey (2000), see also Durbin and Koopman (2012).

The Model Analyzed in This Contribution

In the following we consider cointegration modeled by the cointegrated vector autoregressive (CVAR) model, $H(r)$, for the p-dimensional process x_t,

$$H(r) : \Delta x_t = \alpha(\beta' x_{t-1} + \gamma D_t) + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \Phi d_t + \varepsilon_t \quad [5]$$

The terms D_t and d_t are deterministic terms like constant, trend, seasonal or intervention dummies. The matrices α and β are $p \times r$ where $0 \le r \le p$. In section Integration and Cointegration, conditions for the processes $\beta' x_t$ and Δx_t to be stationary around their means are given, and model [5] can then be formulated as

$$\begin{aligned}\Delta x_t - E(\Delta x_t) &= \alpha(\beta' x_{t-1} - E(\beta' x_{t-1})) \\ &\quad + \sum_{i=1}^{k-1} \Gamma_i(\Delta x_{t-i} - E(\Delta x_{t-i})) + \varepsilon_t\end{aligned}$$

This shows how the change of the process reacts to feedback from disequilibrium errors $\beta' x_{t-1} - E(\beta' x_{t-1})$ and $\Delta x_{t-i} - E(\Delta x_{t-i})$, via the short-run adjustment coefficients, α and Γ_i. The equation $\beta' x_t - E(\beta' x_t) = 0$ defines the long-run relations between the processes.

There are many surveys of the theory of cointegration; see for instance Watson (1994) and Stock (1994) or Johansen (2006, 2009), where the last two form the basis for the presentation here. The topic has become part of most textbooks in econometrics; see among others Banerjee et al. (1993), Hamilton (1994), Hendry (1995), and Lütkepohl (2006). For a general account of the methodology of the CVAR model with applications to the analysis of economic data, see Juselius (2006).

Linear Stationary Processes

We consider p-dimensional linear stationary processes

$$z_t = \sum_{i=0}^{\infty} C_i \varepsilon_{t-i}$$

which are well defined if the coefficient matrices satisfy the condition that $\sum_{i=0}^{\infty} tr^{1/2}(C_i' C_i) < \infty$, where the trace of a matrix, C, is $tr(C) = \sum_{i=1}^{p} C_{ii}$. If in [5] we consider $r = p$ we define the matrix $\Pi = \alpha\beta'$ and the matrix valued characteristic polynomial

$$\Psi(z) = (1-z)I_p - \Pi z - (1-z)\sum_{i=1}^{k-1} \Gamma_i z^i \quad [6]$$

with determinant $|\Psi(z)|$. The properties of the solution of [5] are determined by $\Psi(z)$. We define the roots $z_i, i = 1,\ldots,n$, as the

solutions of $|\Psi(z)| = 0$, and get, because $|\Psi(0)| = 1$, that $|\Psi(z)| = \prod_{i=1}^{n}(1 - z/z_i)$. The inverse characteristic polynomial is given by

$$\Psi(z)^{-1} = \frac{\mathrm{adj}(\Psi(z))}{\det\Psi(z)}, z \neq z_i$$

that is, the adjoint of $\Psi(z)$ divided by the determinant of $\Psi(z)$.

The function, $C(z) = \Psi(z)^{-1}$ has poles at the roots of the polynomial $|\Psi(z)|$, and the position of the poles determine the stochastic properties of the solution of [5]. We first mention a well-known result; see Anderson (1971).

Theorem 1: If the roots satisfy $|z_i| > 1$, then α and β have full rank $r = p$, and the coefficients of $\Psi^{-1}(z) = \sum_{i=0}^{\infty} C_i z^i$ are exponentially decreasing. Let $\mu_t = \sum_{i=0}^{\infty} C_i(\alpha\Upsilon D_{t-i} + \Phi d_{t-i})$. Then the distribution of the initial values of x_t can be chosen so that $x_t - \mu_t$ is stationary with moving average representation

$$x_t - \mu_t = \sum_{i=0}^{\infty} C_i \varepsilon_{t-i}$$

Thus the exponentially decreasing coefficients are found by simply inverting the characteristic polynomial if the roots are outside the unit disk. The matrices, C_i contain the impulse response coefficient of the process in the sense that a shock at time zero to variable k will have the effect, $(C_t)_{ik}$ at time t to variable i.

Integration and Cointegration

The basic definitions of integration and cointegration are given together with a moving average representation of the solution of the error correction model [5]. This solution reveals the stochastic properties of the solution, see Example 3.

If the roots of $|\Psi(z)| = 0$ are not greater than 1, the equations generate nonstationary processes of various types, and the coefficients are not exponentially decreasing. Still, the coefficients of $C(z) = \Psi^{-1}(z)$ determine the stochastic properties of the solution of [5].

Definition 1: If $\sum_{i=0}^{\infty} tr^{1/2}(C_i'C_i) < \infty$, the linear process $x_t - E(x_t) = \sum_{i=0}^{\infty} C_i \varepsilon_{t-i}$ is called $I(0)$ if $C(1) = \sum_{i=0}^{\infty} C_i \neq 0$ The process, x_t is called integrated of order 1, $I(1)$, if $\Delta x_t - E(\Delta x_t)$ is $I(0)$. If there is a vector $\beta \neq 0$ so that $\beta' x_t$ is stationary around its mean, then x_t is cointegrated with cointegration vector, β. The number of linearly independent cointegration vectors is the cointegration rank.

We consider the process given by [5] and the characteristic polynomial, $\Psi(z)$ defined in [6]. This has a unit root, if $\Psi(1) = -\Pi$ is singular, and by Theorem 1, the process is not stationary. A singular matrix Π of rank r can be expressed as $\Pi = \alpha\beta'$, where α and β are $p \times r$. We next formulate a condition for the process to be $I(1)$. We define $\Gamma = I_p - \sum_{i=1}^{k-1} \Gamma_i$.

Assumption 1: (The $I(1)$ condition) The $I(1)$ condition is satisfied if the roots, $|\Psi(z_i)| = 0$ satisfy $|z_i| > 1$ or $z_i = 1$ and it holds that

$$|\alpha_\perp' \Gamma \beta_\perp| \neq 0 \qquad [7]$$

Condition [7] is needed to avoid solutions that are integrated of order 2 or higher; see section Further Topics in Cointegration for references. For a process with one lag $\Gamma = I_p$, and [5] implies

$$\beta' x_t = (I_r + \beta'\alpha)\beta' x_{t-1} + \beta'\varepsilon_t$$

In this case the $I(1)$ condition is equivalent to the condition that $\beta' x_t$ is stationary, that is, the absolute value of the eigenvalues of $I_r + \beta'\alpha$ are less than 1, and in Example 2 this condition reduces to $|1 + \alpha_1 - \alpha_2| < 1$.

Example 3 presents a special case of the Granger Representation Theorem, which gives the moving average representation of the solution of the error correction model.

Theorem 2: (The Granger Representation Theorem) If $\Psi(z)$ has unit roots and the $I(1)$ condition is satisfied, then

$$(1 - z)\Psi(z)^{-1} = C(z) = \sum_{i=0}^{\infty} C_i z^i = C + (1 - z)C^*(z)$$

converges for $|z| \leq 1 + \delta$ for some $\delta > 0$. The matrix C is defined by

$$C = \beta_\perp (\alpha_\perp' \Gamma \beta_\perp)^{-1} \alpha_\perp'$$

The solution x_t of [5] has the moving average representation

$$x_t = C\sum_{i=1}^{t}(\varepsilon_i + \Phi d_i) + \sum_{i=0}^{\infty} C_i^*(\varepsilon_{t-i} + \Phi d_{t-i} + \alpha\Upsilon D_{t-i}) + A \qquad [8]$$

where A depends on initial values, so that $\beta' A = 0$.

This result implies that Δx_t and $\beta' x_t$ are stationary around their mean, so that x_t is a cointegrated $I(1)$ process with r cointegration vectors β and $p - r$ common stochastic trends $\alpha_\perp' \sum_{i=1}^{t} \varepsilon_i$.

One of the useful applications of the representation [8] is to investigate the role of the deterministic terms. Note that d_t cumulates in the process with a coefficient, $C\Phi$, but that D_t does not, because $C\alpha\Upsilon = 0$. A leading special case is the model with $D_t = t$, and $d_t = 1$, which ensures that any linear combination of the components of x_t is allowed to have a linear trend. Note that if $D_t = t$ is not allowed in the model, that is $\Upsilon = 0$, then x_t has a trend given by $C\Phi t$, but the cointegration relation, $\beta' x_t$ has no trend because $\beta' C\Phi = 0$.

Interpretation of the *I*(1) Model

In this section, model $H(r)$ defined by [5] is discussed. The parameters in $H(r)$ are

$$(\alpha, \beta, \Gamma_1, \ldots, \Gamma_{k-1}, \Upsilon, \Phi, \Omega)$$

All parameters vary freely and α and β are $p \times r$ matrices. The normalization and identification of α and β are discussed, and some examples of hypotheses on α and β are given.

The Relation between the Models *H*(*r*)

The models $H(r)$ are nested

$$H(0) \subset \ldots \subset H(r) \subset \ldots \subset H(p)$$

Here $H(p)$ is the unrestricted vector autoregressive model, so that α and β are unrestricted $p \times p$ matrices. The model $H(0)$ corresponds to the restriction, $\alpha = \beta = 0$, which is the vector autoregressive model for the process in

differences. Note that in order to have nested models, we allow in $H(r)$ for all processes with rank of α and β less than or equal to r.

The formulation allows us to derive likelihood ratio tests for the hypothesis, $H(r)$ in the unrestricted model, $H(p)$. These tests can be applied to check if one's prior knowledge of the number of cointegration relations is consistent with the data, or alternatively to construct an estimator of the cointegration rank.

Note that when the cointegration rank is r, the number of common trends is $p-r$. Thus, if one can interpret the presence of r cointegration relations, one should also interpret the presence of $p-$ r independent stochastic trends or $p-r$ driving forces in the data.

Normalization of Parameters

The parameters α and β in [5] are not uniquely identified, because given any choice of α and β and any nonsingular $r \times r$ matrix ξ, the choice $\alpha\xi$ and $\beta\xi'^{-1}$ gives the same matrix $\alpha\xi(\beta\xi'^{-1})' = \alpha\xi\xi^{-1}\beta' = \alpha\beta'$.

If $x_t = (x_{1t}', x_{2t}')'$ where x_{1t} is $r \times 1$ and x_{2t} is $(p-r) \times 1$, and $\beta = (\beta_1', \beta_2')'$, with β_1, $r \times r$, and $|\beta_1| \neq 0$, we can solve the cointegration relations as

$$x_{1t} = \gamma' x_{2t} + u_t$$

where u_t is stationary and $\gamma' = -(\beta_1')^{-1}\beta_2'$. This represents cointegration as a regression equation, see Section Regression Formulation. A normalization of this type is sometimes convenient for estimation and calculation of 'standard errors' of the estimator, see Section Asymptotic Distribution of Estimators, but many hypotheses are invariant with respect to a normalization of β, and thus, in a discussion of a test of such a hypothesis, β does not require normalization.

Similarly $\alpha_\perp$ and $\beta_\perp$ are not uniquely defined. From the Granger Representation Theorem we see that the $p-r$ common trends are the nonstationary random walks in $C\sum_{i=1}^t \varepsilon_i$, that is, can be chosen as $\alpha_\perp' \sum_{i=1}^t \varepsilon_i$. For any full rank $(p-r) \times (p-r)$ matrix η, the processes, $\eta\alpha_\perp' \sum_{i=1}^t \varepsilon_i$ could also be used as common trends because

$$C = \beta_\perp(\alpha_\perp' \Gamma \beta_\perp)^{-1}\alpha_\perp' = \beta_\perp(\eta\alpha_\perp' \Gamma \beta_\perp)^{-1}\eta\alpha_\perp'$$

Thus identifying restrictions on the coefficients in $\alpha_\perp$ are needed to find their estimates and standard errors, and a similar result holds for $\beta_\perp$.

In the cointegration model there are therefore four separate identification problems: one for the cointegration relations, one for the common trends, one for $\beta_\perp$, and finally one for the short-run dynamics, if the model has simultaneous effects.

Hypotheses on Long-Run Coefficients

One purpose of modeling economic data is to test hypotheses on the coefficients, thereby investigating whether the data support an economic hypothesis or reject it. As an example consider the series $x_t = (e_t, p_t, p_t^*)'$, where p_t and p_t^* are the log price indices in two countries and e_t the exchange rate. The hypothesis of the law of one price, *PPP*, is that $e_t = p_t - p_t^*$. We formulate that as the hypothesis that (1, −1, 1) is a cointegration vector so that $e_t - p_t + p_t^*$ becomes stationary. Similarly, the hypothesis of price homogeneity is formulated as the restriction

$$R'\beta = (0,1,1)\beta = 0$$
$$\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = H\varphi$$

where $H = R_\perp$. A general formulation of restrictions on each of r cointegration vectors, including a normalization, is

$$\beta = (h_1 + H_1\varphi_1, \ldots, h_r + H_r\varphi_r). \qquad [9]$$

Here h_i is $p \times 1$ and orthogonal to H_i, which is $p \times (s_i - 1)$ of rank $s_i - 1$, so that $p - s_i$ restrictions are imposed on the vector, β_i. Let the restrictions be $R_i = (h_i, H_i)_\perp$ then β_i satisfies $R_i'\beta_i = 0$, and the normalization $(h_i'h_i)^{-1}h_i'\beta_i = 1$. The usual rank condition for identification is that β_i is identified by $R_i'\beta_i = 0$, if

$$\operatorname{rank}\left(R_i'(\beta_1, \ldots, \beta_r)\right) = r - 1$$

Hypotheses on Adjustment Coefficients

The coefficients in α measure how the process adjusts to disequilibrium errors. The hypothesis of weak exogeneity is the hypothesis that some rows of α are zero; see Engle et al. (1983). We decompose the process x_t as $x_t = (x_{1t}', x_{2t}')'$ and the matrices are decomposed similarly so that the model equations (without deterministic terms and $k = 2$) become

$$\Delta x_{1t} = \alpha_1\beta' x_{t-1} + \Gamma_{11}\Delta x_{t-1} + \varepsilon_{1t}$$
$$\Delta x_{2t} = \alpha_2\beta' x_{t-1} + \Gamma_{21}\Delta x_{t-1} + \varepsilon_{2t}$$

The conditional model for Δx_{1t} given Δx_{2t} and the past is

$$\Delta x_{1t} = \omega\Delta x_{2t} + \alpha_1\beta' x_{t-1} + (\Gamma_{11} - \omega\Gamma_{21})\Delta x_{t-1} + \varepsilon_{1t} - \omega\varepsilon_{2t} \qquad [10]$$

where $\omega = \Omega_{12}\Omega_{22}^{-1}$. If $\alpha_2 = 0$, there is no levels feedback from $\beta' x_{t-1}$ to Δx_{2t}, and if the errors are Gaussian, x_{2t} is called weakly exogenous for α_1 and β. In this case the likelihood is a product of two factors depending on $(\Gamma_{21}, \Omega_{22})$ and $(\alpha_1, \beta, \Gamma_{11}, \omega, \Omega_{11.2})$, respectively. Because the parameters are unrestricted (variation independent), likelihood inference on β and α_1 can be conducted in the conditional model alone.

If the hypothesis of weak exogeneity is not satisfied, inference of the conditional model is complicated because limit distributions contain nuisance parameters, and asymptotic inference is not Gaussian.

If x_{2t} is weakly exogenous, $\alpha_\perp$ contains the columns of $(0, I_{p-r})'$, so that $\sum_{i=1}^t \varepsilon_{2i}$ are common trends. Thus the errors in the equations for Δx_{2t} cumulate in the system and give rise to nonstationarity.

Likelihood Analysis

This section contains first some comments on what aspects of the data are important for checking for model misspecification, and then describes the calculation of reduced rank regression, introduced by Anderson (1951). Then reduced rank regression

and modifications thereof are applied to estimate the parameters of the $I(1)$ model [5] and various submodels defined by restrictions on β, see Johansen and Juselius (1990).

Checking for Specification

In order to apply Gaussian maximum likelihood methods, the assumptions behind the model have to be checked carefully, so that one is convinced that the statistical model contains the density that generated the data. If this is not the case, the asymptotic results available from the Gaussian analysis need not hold. Methods for checking vector autoregressive models include choice of lag length, test for normality of residuals, tests for autocorrelation, and test for heteroscedasticity in errors. Asymptotic results for estimators and tests derived from the Gaussian likelihood turn out to be robust to some types of deviations from the above assumptions. Thus the limit results hold for independent identically distributed (iid) errors with finite variance, and not just for Gaussian errors, but autocorrelated errors violate the asymptotic results, so autocorrelation has to be checked carefully.

Finally and perhaps most importantly, the assumption of constant parameters is crucial. In practice it is important to model outliers by suitable dummies, but it is also important to model breaks in the dynamics, breaks in the cointegration properties, breaks in the stationarity properties, etc. The papers by Seo (1998) and Hansen and Johansen (1999) contain some results on recursive tests in the cointegration model, and Doornik and Hendry (2013) contains a description of a general algorithm (*Autometrics*) for finding a model that describes the data.

Reduced Rank Regression

Let u_t, w_t, and z_t be three multivariate time series of dimensions p_u, p_w, and p_z respectively. The algorithm of reduced rank regression, see Anderson (1951), can be described in the regression model

$$u_t = \alpha\beta' w_t + \Xi z_t + \varepsilon_t$$

where ε_t are iid $(0,\Omega)$. The product moments are

$$S_{uw} = T^{-1}\sum_{t=1}^{T} u_t w_t'$$

and the residuals, which we get by regressing u_t on w_t, are

$$(u_t|w_t) = u_t - S_{uw}S_{ww}^{-1}w_t$$

so that the conditional product moments are

$$S_{uw.z} = S_{uw} - S_{uz}S_{zz}^{-1}S_{zw} = T^{-1}\sum_{t=1}^{T}(u_t|z_t)(w_t|z_t)'$$

$$S_{uu.w,z} = T^{-1}\sum_{t=1}^{T}(u_t|w_t,z_t)(u_t|w_t,z_t)' = S_{uu.w} - S_{uz.w}S_{zz.w}^{-1}S_{zu.w}$$

For fixed β the regression estimates are

$$\widehat{\alpha}(\beta) = S_{uw.z}\beta(\beta' S_{ww.z}\beta)^{-1}$$
$$\widehat{\Omega}(\beta) = S_{uu.z} - S_{uw.z}\beta(\beta' S_{ww.z}\beta)^{-1}\beta' S_{wu.z}$$

so that

$$|\widehat{\Omega}(\beta)| = |S_{uu.w}|\frac{|\beta' S_{ww.uz}\beta|}{|\beta' S_{ww.z}\beta|}$$

Minimizing over β gives the reduced rank estimators. This minimization problem is solved as follows. First we solve the eigenvalue problem

$$\left|\lambda S_{ww.z} - S_{wu.z}S_{uu.z}^{-1}S_{uw.z}\right| = 0$$

The eigenvalues are ordered $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_{p_w}$ with corresponding eigenvectors $\widehat{v}_1, \ldots \widehat{v}_{p_w}$. The reduced rank estimate of β is

$$\widehat{\beta} = (\widehat{v}_1, \ldots, \widehat{v}_r) \qquad [11]$$

and the other estimators are found by regression of u_t on $\widehat{\beta}' x_{t-1}$ and z_t. Finally we find

$$|\widehat{\Omega}| = |S_{uu.z}|\prod_{i=1}^{r}\left(1 - \widehat{\lambda}_i\right)$$

The eigenvectors are orthogonal with respect to $S_{ww.z}$, that is, $\widehat{v}_i' S_{ww.z}\widehat{v}_j = 0$ for $i \neq j$, and they are normalized by $\widehat{v}_i' S_{ww.z}\widehat{v}_i = 1$. The calculations described here are called a reduced rank regression and are denoted by $RRR(u_t, w_t \mid z_t)$.

Likelihood Analysis

It is assumed for the likelihood analysis that ε_t is iid $N_p(0, \Omega)$, but for asymptotic results the Gaussian assumption is not needed. The Gaussian likelihood function shows that the maximum likelihood estimator can be found by the reduced rank regression of Δx_t on $(x'_{t-1}, D'_t)'$ correcting for $\mathbb{X}_t = (\Delta x'_{t-1}, \ldots, \Delta x'_{t-k+1}, d'_t)'$

$$RRR\left(\Delta x_t, \begin{pmatrix} x_{t-1} \\ D_t \end{pmatrix} | \mathbb{X}_t\right)$$

The estimates are given by [11], and the maximized likelihood is, apart from a constant, given by

$$L_{\max}^{-2/T} = |\widehat{\Omega}| = |S_{00}|\prod_{i=1}^{r}\left(1 - \widehat{\lambda}_i\right) \qquad [12]$$

where $S_{00} = T^{-1}\sum_{t=1}^{T}(\Delta X_t|\mathbb{X}_t)(\Delta X_t|\mathbb{X}_t)'$.

Note that all the models, $H(r)$, $r = 0, \ldots, p$, have been solved by the same eigenvalue calculation. The maximized likelihood is given for each r by [12] and by dividing the maximized likelihood function for r with the corresponding expression for $r = p$, the likelihood ratio test for cointegration rank is obtained:

$$-2\log LR(H(r)|H(p)) = -T\sum_{i=r+1}^{p}\log\left(1 - \widehat{\lambda}_i\right)$$

The asymptotic distribution of this test statistic and the estimators are discussed in section Asymptotic Analysis.

The model obtained under the hypothesis, $\beta = H\varphi$ is analyzed by

$$RRR\left(\Delta x_t, \begin{pmatrix} H'X_{t-1} \\ D_t \end{pmatrix} | \mathbb{X}_t\right)$$

and a number of hypotheses of this type for β and α can be solved in the same way, but the more general hypothesis

$$\beta = (h_1 + H_1\varphi_1, \ldots, h_r + H_r\varphi_r)$$

see [9], cannot be solved by reduced rank regression.

Asymptotic Analysis

A discussion of the most important aspects of the asymptotic analysis of the cointegration model is given. This includes the result that the rank test requires a family of Dickey–Fuller type distributions, depending on the specification of the deterministic terms of the model. The asymptotic distribution of $\widehat{\beta}$ is mixed Gaussian and that of the remaining parameters is Gaussian, so that tests for hypotheses on the parameters are asymptotically distributed as χ^2.

Asymptotic Distribution of the Rank Test

The asymptotic distribution of the rank test is given in case the process has a linear trend.

Theorem 3: Let ε_t be iid $(0, \Omega)$ and assume that $D_t = t$ and $d_t = 1$, in model [5]. Under the assumptions that the cointegration rank is r, the asymptotic distribution of the likelihood ratio test statistic $-2\log LR(H(r)|H(p))$, [13], is

$$tr\left\{\int_0^1 (dB)F'\left(\int_0^1 FF'du\right)^{-1}\int_0^1 F(dB)'\right\} \qquad [13]$$

where F is defined by

$$F(u) = \left(\begin{array}{c|c} B(u) & \\ u & 1 \end{array}\right)$$

and $B(u)$ is the $p - r$ dimensional standard Brownian motion.

The limit distribution is tabulated by simulation. Note that it does not depend on the parameters $(\Gamma_1,\ldots,\Gamma_{k-1},\Upsilon,\Phi,\Omega)$, but only on $p - r$, the number of common trends, and the presence of the linear trend. For finite samples, however, the dependence on the parameters can be quite pronounced. A small sample correction for the test has been given in Johansen (2002), and the bootstrap has been investigated by Rahbek et al. (2012).

In the model without deterministics, the same result holds, but with $F(u) = B(u)$. A special case of this, for $p = 1$, is the Dickey–Fuller test and the asymptotic distributions given in Theorem 3 are called the Dickey–Fuller distributions with $p - r$ degrees of freedom; see Dickey and Fuller (1981).

Asymptotic Distribution of Estimators

The main result here is that the estimator of β, suitably normalized, converges to a mixed Gaussian distribution; see Johansen (1988). This result implies that likelihood ratio tests on β are asymptotically χ^2 distributed. Furthermore the estimators of the adjustment parameters α and the short-run parameters, Γ_i are asymptotically Gaussian and asymptotically independent of the estimator for β.

To illustrate how to conduct inference on a cointegrating coefficient, and why it becomes asymptotic χ^2 despite the asymptotic mixed Gaussian limit of $\widehat{\beta}$, we consider an example.

Example 4: (Mixed Gaussian distribution) Let x_t be a bivariate process with one lag for which $\alpha = (-1, 0)'$ and $\beta = (1, \theta)'$. The equations become

$$\begin{aligned} x_{1t} &= \theta x_{2t-1} + \varepsilon_{1t} \\ \Delta x_{2t} &= \varepsilon_{2t} \end{aligned} \qquad [14]$$

This model as a special case of [5] with $\alpha' = (-1, 0), \beta' = (1, -\theta), p = 2, k = 1$. If we add the assumption that ε_t is Gaussian with mean zero and variance $\Omega = diag(\sigma_1^2, \sigma_2^2)$ the maximum likelihood estimator simplifies to a regression estimator, and becomes

$$\widehat{\theta} = \frac{\sum_{t=1}^T x_{1t}x_{2t-1}}{\sum_{t=1}^T x_{2t-1}^2} = \theta + \frac{\sum_{t=1}^T \varepsilon_{1t}x_{2t-1}}{\sum_{t=1}^T x_{2t-1}^2}$$

The distribution of $\widehat{\theta}$ conditional on the whole process $\{x_{2t}\}_{t=1}^T$ is clearly Gaussian:

$$\widehat{\theta}\big|\{x_{2t}\}\text{is distributed as } N\left(\theta, \sigma_1 \Big/ \sum_{t=1}^T x_{2t-1}^2\right)$$

By integrating out the process x_{2t} we get a distribution which we call mixed Gaussian with mixing parameter $1/\sum_{t=1}^T x_{2t-1}^2$, and hence

$$E\left(\widehat{\theta}\right) = \theta \quad \text{and} \quad Var\left(\widehat{\theta}\right) = \sigma_1^2 E\left(1 \Big/ \sum_{t=1}^T x_{2t-1}^2\right)$$

When constructing a test for $\theta = \theta_0$ we do not use

$$\left(\widehat{\theta} - \theta\right) \Big/ Var\left(\widehat{\theta}\right)^{1/2}$$

but instead expand the likelihood function and find the Wald test, which is based on the observed information:

$$\widehat{\sigma}_1^{-1}\left(\sum_{t=1}^T x_{2t-1}^2\right)^{1/2}\left(\widehat{\theta} - \theta\right)$$

This statistic is asymptotically distributed as $N(0, 1)$: Thus we normalize by the observed information, $\sum_{t=1}^T x_{2t-1}^2/\sigma_1^2$, not the expected information often used when analyzing stationary processes.

Figure 4 shows a scatter diagram of 1000 simulations of $(\widehat{\theta}, \sum_{t=1}^T x_{2t-1}^2/\sigma_1^2)$. That is, the estimator and the information about the parameter. We note that when the information is large, the variation of $\widehat{\theta}$ is small, and when the information is small, the variation of $\widehat{\theta}$ is much larger. Thus the variation of $\widehat{\theta}$ should be measured by its conditional variance, which is the reciprocal information in the data. This has the further advantage that if we only consider those estimates with a given information we see that $\widehat{\theta}$ is approximately Gaussian.

The main result is that tests on β are asymptotically distributed as χ^2, and we formulate that as.

Theorem 4: Let εt be iid $(0, \Omega)$. The asymptotic distribution of the likelihood ratio test statistic for the restrictions [9] in model [5] is distributed as χ^2 with degrees of freedom given by $\sum_{i=1}^r (p - r - s_i + 1)$. A small sample correction for some tests on β has been developed in Johansen (2000).

Further Topics in Cointegration

The basic model for $I(1)$ processes has been extended to other types of nonstationarity. In particular models for seasonal roots, Ahn and Reinsel (1994) and Johansen and Schaumburg (1998), explosive processes, Nielsen (2010), $I(2)$ processes, Johansen (1997), fractional processes, Johansen and Nielsen (2012), nonlinear processes, Lange and Rahbek (2009), panel data cointegration, Larsson et al. (2001) and Pesaran et al.

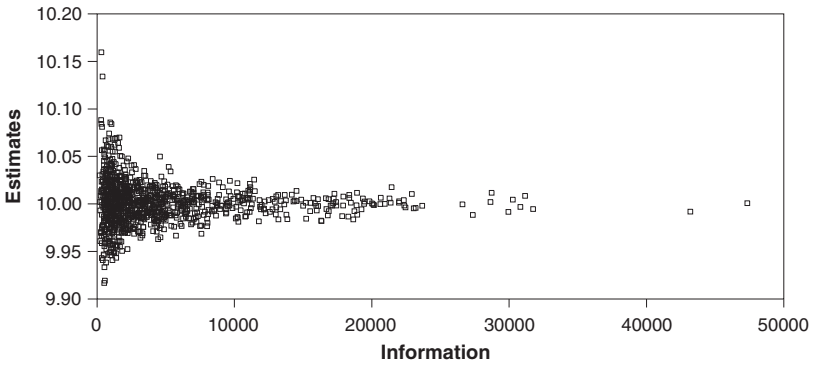

Figure 4 A scatter plot of $\widehat{\theta} = \sum_{i=1}^{T} x_{1t} x_{2t-1} / \sum_{i=1}^{T} x_{2,t-1}^2$ and the information. $\sum_{i=1}^{T} x_{2,t-1}^2 / \sigma_1^2$ for model [14]. The number of observations is $T = 100$, and the number of simulations is 1000.

(2004), and finally applications to rational expectation models, Johansen and Swensen (2011).

Concluding Remarks

In summary one can say, that what has been developed for the CVAR model is a set of useful tools for the analysis of many types of economic time series. The theory is now part of many textbooks, and software for the analysis of data has been implemented in several packages, e.g., in CATS, in RATS, Givewin, Eviews, Microfit, Shazam, R, Gauss, GRETL, etc.

We have given a brief tour in the cointegration landscape showing some of the major sights without indicating, except by examples, how the formal analysis is conducted. We concluded with a list of extensions of the basic model, which shows that the ideas behind the CVAR extend to a large number of other models.

Acknowledgment

The author acknowledges the support of the Center for Research in Econometric Analysis of Time Series (CREATES – DNRF78, funded by the Danish National Research Foundation).

See also: Probability Theory: Formal; Time Series: General; Time Series: Unit-Roots and Nonstandard Limiting Distributions.

Bibliography

Ahn, S.K., Reinsel, G.C., 1990. Estimation for partially nonstationary multivariate autoregressive models. Journal of the American Statistical Association 85, 813–823.

Ahn, S.K., Reinsel, G.C., 1994. Estimation of partially non-stationary vector autoregressive models with seasonal behavior. Journal of Econometrics 62, 317–350.

Anderson, T.W., 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. Annals of Mathematical Statistics 22, 327–351.

Anderson, T.W., 1971. The Statistical Analysis of Time Series. Wiley, New York.

Banerjee, A., Dolado, J.J., Galbraith, J.W., Hendry, D.F., 1993. Co-integration Error-correction and the Econometric Analysis of Nonstationary Data. Oxford University Press, Oxford.

Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.

Doornik, J., Hendry, D.F., 2013. Empirical Model Discovery and Theory Evaluation. MIT Press, Cambridge.

Durbin, J., Koopman, S.J., 2012. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55, 251–276.

Engle, R.F., Hendry, D.F., Richard, J.-F., 1983. Exogeneity. Econometrica 51, 277–304.

Granger, C.W.J., 1983. Cointegrated variables and error correction models. UCSD Discussion Paper, 83–13a.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, New Jersey, NJ.

Hansen, H., Johansen, S., 1999. Some tests for parameter constancy in the cointegrated VAR. The Econometrics Journal 2, 306–333.

Hendry, D.F., 1995. Dynamic Econometrics. Oxford University Press, Oxford.

Johansen, S., 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12, 231–254.

Johansen, S., 1996. Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press, Oxford.

Johansen, S., 1997. Likelihood analysis of the *I*(2) model. Scandinavian Journal of Statistics 24, 433–462.

Johansen, S., 2000. A Bartlett correction factor for tests on the cointegration relations. Econometric Theory 16, 740–778.

Johansen, S., 2002. A small sample correction of the test for cointegration rank in the vector autoregressive model. Econometrica 70, 1929–1961.

Johansen, S., 2006. Cointegration, a survey. In: Mills, T.C., Patterson, K. (Eds.), Palgrave Handbook of Econometrics, vol. 1. Econometric Theory. Palgrave Macmillan, Basingstoke, pp. 671–693.

Johansen, S., 2009. Cointegration. Overview and development. In: Andersen, T.G., Davies, R., Kress, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer, New York, pp. 671–694.

Johansen, S., Juselius, K., 1990. Maximum likelihood estimation and inference on cointegration — with applications to the demand for money. Oxford Bulletin of Economics and Statistics 52, 169–210.

Johansen, S., Nielsen, M.Ø., 2012. Likelihood inference for a fractionally cointegrated vector autoregressive model. Econometrica 80, 2667–2732.

Johansen, S., Schaumburg, E., 1998. Likelihood analysis of seasonal cointegration. Journal of Econometrics 88, 301–339.

Johansen, S., Swensen, A.R., 2011. On a numerical and graphical technique for evaluating some models involving rational expectations. Journal of Time Series Econometrics 3, Article 9. http://dx.doi.org/10.2202/1941-1928.1089.

Juselius, K., 2006. The Cointegrated VAR Model,Econometric Methodology and Macroeconomic Applications. Oxford University Press, Oxford.

Lange, T., Rahbek, A., 2009. An Introduction to regime switching time series models. In: Andersen, T.G., Davies, R., Kress, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer, New York, pp. 871–888.

Larsson, R.L., Lyhagen, J., Löthgren, M., 2001. Likelihood based cointegration tests in heterogenous panels. The Econometric Journal 4, 109–142.

Lütkepohl, H., 2006. New Introduction to Multiple Time Series Analysis. Springer, New York.

Nielsen, B., 2010. Analysis of explosive processes. Econometric Theory 26, 882–915.

Nyblom, J., Harvey, A., 2000. Tests of common stochastic trends. Econometric Theory 16, 176–199.

Pesaran, M.H., Schuermann, T., Weiner, S.M., 2004. Modeling regional interdependencies using a global error-correcting macroeconometric model. Journal of Business and Economic Statistics 22, 169–172.

Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. Econometrica 59, 283–306.

Phillips, P.C.B., Hansen, B.E., 1990. Statistical inference on instrumental variables regression with *I*(1) processes. Review of Economic Studies 57, 99–124.

Rahbek, A., Cavaliere, G., Taylor, A.M.R., 2012. Bootstrap determination of the cointegration rank in VAR models. Econometrica 80, 1721–1740.

Seo, B., 1998. Tests for structural change in cointegrated systems. Econometric Theory 14, 222–259.

Stock, J.H., 1987. Asymptotic properties of least squares estimates of cointegration vectors. Econometrica 55, 1035–1056.

Stock, J.H., 1994. Unit roots, structural breaks and trends. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Netherlands, pp. 2739–2841.

Watson, M., 1994. Vector autoregressions and cointegration. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Netherlands, pp. 871–888.

Time Series: Cycles

Mark W Watson, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 23, pp. 15714–15721, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

This article discusses cyclical (or periodic) properties of stochastic processes. The spectrum (or spectral density function) is introduced as a method for characterizing the approximate periodic behavior of realizations from stochastic processes. The spectral representation is introduced as a way to decompose a stochastic process into its periodic components. Moving average filters change the cyclical properties of time series, and formulae characterizing these changes are developed. Specific moving average filters are discussed, and general formulae for extracting components associated with specific frequencies (bandpass filters) are presented. Spectral properties of ARMA process are derived. Methods for estimating spectra are outlined.

Time series data in economics and other fields of social science often exhibit cyclical behavior. For example, aggregate retail sales are high in November and December and follow a seasonal cycle; voter registrations are high before each presidential election and follow an election cycle; and aggregate macro-economic activity falls into recession every six to eight years and follows a business cycle. In spite of this cyclicality, these series are not perfectly predictable, and the cycles are not strictly periodic. That is, each cycle is unique. Quite generally, it is useful to think of time series as realizations of stochastic processes, and this raises the question of how to represent cyclicality in stochastic processes. The major descriptive tool for this is the 'spectral density function' (or 'spectrum'), which is the subject of this article.

Cycles in a Typical Time Series

Figure 1 shows a time series plot of new housing authorizations ('building permits') issued by communities in the USA, monthly, from 1960 through 1999. This plot has characteristics that are typical of many economic time series. First, the plot shows a clear seasonal pattern: permits are low in the late fall and early winter, and rise markedly in the spring and summer. This seasonal pattern is persistent throughout the sample period, but does not repeat itself exactly year-after-year. In addition to a seasonal pattern, there is a slow moving change in the level of the series associated with the business cycle. For example, local minima of the series are evident in 1967, the mid-1970s, the early 1980s, and in 1990, which correspond to periods of macroeconomic slowdown or recession in the US economy.

How can the cyclical variability in the series be represented? One approach is to use a periodic function like a sine or cosine function. But this is inadequate in at least two ways. First, a deterministic periodic function doesn't capture the randomness in the series. Second, since several different periodicities (seasonal, business cycle) are evident in the plot, several periodic functions will be required. The next section presents a representation of a stochastic process that has these ingredients.

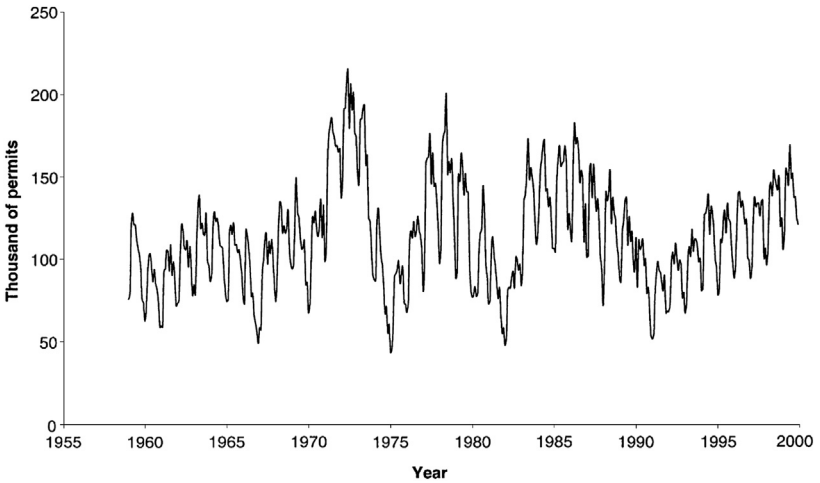

Figure 1 Building permits in the USA.

Representing Cyclical Behavior in Time Series: The Spectrum

Spectral Representation of a Covariance Stationary Process

Consider a sequence of scalar random variables $\{Y_t\}$ where $t = 0, \pm 1, \pm 2, \ldots$, and $E(Y_t) = \mu$, $Cov(Y_t Y_{t-k}) = \lambda_k$. That is, assume that Y_t is observed at regular intervals, like a week, a month, a year, etc. Assume that the first and second moments of the process do not depend on time, that is μ and λ_k are time invariant. To simplify notation, assume that $\mu = 0$.

To motivate the spectral representation of Y_t, it is useful to begin with a very special and simple stochastic process:

$$Y_t = \alpha \cos(\omega t) + \delta \sin(\omega t) \qquad [1]$$

where α and δ are random variables with $E(\alpha) = E(\delta) = 0$; $E(\alpha\delta) = 0$; and $E(\alpha^2) = E(\delta^2) = \sigma^2$. While simple, this process has three attractive characteristics. First, it is periodic: since $\cos(a + 2\pi) = \cos(a)$ and $\sin(a + 2\pi) = \sin(a)$, then $Y_{t+j2\pi/\omega} = Y_t$ for $|j| = 1, 2, \ldots$. So Y_t repeats itself with a period of $2\pi/\omega$. Second, the random components α and δ give Y a random amplitude (value of at its peak) and a random phase (value at $t = 0$). Thus, two realizations of Y will have different amplitudes and different timing of their peaks and troughs. Finally, the process is covariance stationary with $E(Y_t) = 0$ and $\lambda_k = \sigma^2 \cos(\omega k)$.

Adding together several components like eqn [1] produces a more interesting stochastic process:

$$Y_t = \sum_{j=1}^{n} \left[\alpha_j \cos(\omega_j t) + \delta_j \sin(\omega_j t)\right] \qquad [2]$$

with $E(\alpha_j) = E(\delta_j) = 0$, for all j; $E(\alpha_j \delta_k) = 0$, for all j, k; $E(\alpha_j, \alpha_k) = E(\delta_j \delta_k) = 0$, for $j \neq k$; and $E(\alpha_j^2) = E(\delta_j^2) = \sigma_j^2$. For this process, Y_t is the sum of n different uncorrelated components, each corresponding to a different frequency, and each with its own variance. (The variance of $\alpha_j \cos(\omega_j t) + \delta_j \sin(\omega_j t)$ is σ_j^2.) A calculation shows $E(Y_t) = 0$, $\lambda_k = \sum_{j=1}^{n} \sigma_j^2 \cos(\omega_j k)$, and the variance of Y_t is given by $var(Y_t) = \sum_{j=1}^{n} \sigma_j^2$.

So far, this has all been special. That is, two very special processes have been constructed. However, a fundamental result (Cramér, 1942; Kolmorgorov, 1940) shows that a generalization of this decomposition can be used to represent any covariance stationary process. The result, known as the 'Spectral Representation Theorem,' says that if Y_t is covariance stationary, then it can be represented as

$$Y_t = \int_0^{\pi} \cos(\omega t) d\alpha(\omega) + \int_0^{\pi} \sin(\omega t)\, d\delta(\omega)$$

where $d\alpha(\omega)$ and $d\delta(\omega)$ are zero mean random variables that are mutually uncorrelated, uncorrelated across frequency, and have variances that depend on frequency. Like the special example shown in eqn [2], this general representation decomposes Y_t into a set of strictly periodic components – each uncorrelated with the others, and each with its own variance. For example, processes with important seasonal components will have large values of the components corresponding to the seasonal frequency, processes for series with strong 'business cycle' components will have large variances at business cycle frequencies, etc.

Sometimes the spectral representation is written using complex numbers as:

$$Y_t = \int_{-\pi}^{\pi} e^{i\omega t} d\gamma(\omega) \qquad [3]$$

where $d\gamma(\omega) = (1/\sqrt{2})\,[d\alpha(\omega) - i d\delta(\omega)]$ for $\omega \geq 0$ and $d\gamma(\omega) = d\gamma(-\omega)$ for $\omega < 0$. This representation simplifies notation for some calculations.

The spectral representation eqn [3] is a generalization of eqn [2] where the 'increments' $d\gamma(\omega)$ have orthogonality properties like those of α_j and δ_j. The variance of each component is frequency specific (like σ_j^2), and is summarized by a density function $S(\omega)$. Specifically, the variance of $d\gamma(\omega)$ is $E(d\gamma(\omega)\, d\gamma(\omega)) = S(\omega) d\omega$. (The use of a density function to summarize the variance of $d\gamma(\omega)$ is not completely general and rules out processes with deterministic (or perfectly predictable) components. The references listed in the last section provide this more general result.)

You may wonder why the spectral representation doesn't use frequencies larger than π. The answer is that these are not needed to describe a process measured at discrete intervals. This discreteness means that periodic components associated with frequencies larger than π will look just like components associated with frequencies less than π. For example, a component with frequency 2π will have period of 1 and, because the series is measured only once every period, this component will appear to be constant – it will look just like (be 'aliased' with) a component that has a frequency of $\omega = 0$.

Relationship between the Spectral Density and the Autocovariances

There is a one-to-one relationship between the spectral density function and the autocovariances of the process. The autocovariances follow directly from the spectral representation:

$$\begin{aligned} \lambda_k &= E(Y_t Y_{t-k}) = E(Y_t \overline{Y}_{t-k}) \\ &= E\left[\int e^{i\omega t} d\gamma(\omega) \int e^{-i\omega(t-k)} \overline{d\gamma(\omega)}\right] \\ &= \int e^{i\omega k} S(\omega) d\omega. \end{aligned} \qquad [4]$$

The spectrum can be determining by inverting this relation:

$$S(\omega) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \lambda_j e^{-i\omega j} \qquad [5]$$

(To verify that eqn [5] is the inverse of eqn [4], use.

$$\int_{-\pi}^{\pi} e^{iwk} dw = \begin{Bmatrix} 2\pi & \text{for} & k = 0 \\ 0 & \text{for} & k \neq 0 \end{Bmatrix} \qquad [6]$$

so that $(2\pi)^{-1}\{\int_{-\pi}^{\pi} \sum_{j=-\infty}^{\infty} \lambda_j e^{-i\omega(j-k)} d\omega\} = \lambda_k$.

A simpler formula for the spectrum is.

$$S(\omega) = (2\pi)^{-1} \left[\lambda_0 + 2 \sum_{j=1}^{\infty} \lambda_j \cos(j\omega)\right]. \qquad [7]$$

which follows from eqn [5], since $\lambda_j = Cov(Y_t Y_{t-j}) = Cov(Y_{t+j} Y_t) = \lambda_{-j}$.

A Summary of the Properties of the Spectrum

Summarizing the results presented in Section Relationship between the Spectral Density and the Autocovariances, the spectrum (or spectral density function) has 4 important properties.

1. $S(\omega)d\omega$ can be interpreted as the variance of the cyclical component of Y corresponding to the frequency ω. The period of this component is $2\pi/\omega$.
2. $S(\omega) \geq 0$. This follows, because $S(\omega)$ is a variance function.
3. $S(\omega) = S(-\omega)$. This follows from the definition of $\gamma(\omega)$ in the spectral representation eqn [3], or from eqn [7], since $\cos(a) = \cos(-a)$. Because of this symmetry, plots of the spectrum are presented $0 \leq \omega \leq \pi$.
4. $Var(Y) = \int_{-\pi}^{\pi} S(\omega)\mathrm{d}\omega$.

The Spectrum of Building Permits

The time series plot of building permits (**Figure 1**) shows that most of the variability in the series comes from two sources: seasonal variation over the year and business-cycle variability. This is evident in the estimated spectrum for the series, shown in **Figure 2**. Most of the mass in the spectrum is concentrated around the seven peaks evident in the plot. (These peaks are sufficiently large that spectrum is plotted on a log scale.) The first peak occurs at frequency $\omega = 0.07$ corresponding to a period of 90 months. This represents the business cycle variability in the series. The other peaks occur at frequencies $2\pi/12$, $4\pi/12$, $6\pi/12$, $8\pi/12$, and π. These are peaks for the seasonal frequencies: the first corresponds to a period of 12 months, and the others are the seasonal 'harmonics' 6, 4, 3, and 2 months.

Spectral Properties of Moving Average Filters

Some General Results

Often, one time series (Y) is converted into another time series (X) through a moving average function such as:

$$X_t = \sum_{j=-r}^{s} c_j Y_{t-j}$$

For example, X_t might be the first difference of Y_t, i.e., $X_t = Y_t - Y_{t-1}$, or an annual moving average $X_t = \frac{1}{12}\sum_{j=0}^{11} Y_{t-j}$, or one of the more complicated moving averages that approximate official seasonal adjustment procedures (*see* Time Series: Seasonal Adjustment).

Let the 'lag' operator L shift time series back by one period (so that $LY_t = Y_{t-1}$, $L^2Y_t = L(LY_t) = Y_{t-2}$, $L^{-1}Y_t = Y_{t+1}$, etc.), then this moving average can be represented as:

$$X_t = c(L)Y_t$$

where,

$$c(L) = c_{-r}L^{-r} + \cdots + c_{-1}L^{-1} + c_0L^0 + c_1L + \cdots + c_sL^s$$

The operator $c(L)$ is sometimes called a 'linear filter.'

How does $c(L)$ change the spectral properties of Y? If the spectral representation of Y_t is written as:

$$Y_t = \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t}\mathrm{d}\gamma(\omega)$$

then X_t can be written as,

$$\begin{aligned} X_t &= \sum_{j=-r}^{s} c_j Y_{t-j} = \sum_{j=-r}^{s} c_j \int_{-\pi}^{\pi} \mathrm{e}^{i\omega(t-j)}\mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t}\left(\sum_{j=-r}^{s} c_j \mathrm{e}^{-i\omega j}\right)\mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t} c(\mathrm{e}^{i\omega})\mathrm{d}\gamma(\omega) \\ &= \int_{-\pi}^{\pi} \mathrm{e}^{i\omega\left(t - \frac{p(\omega)}{\omega}\right)} g(\omega)\mathrm{d}\gamma(\omega) \end{aligned} \qquad [8]$$

where the last line uses the polar form of $c(\mathrm{e}^{-i\omega})$: $c(\mathrm{e}^{-i\omega}) = g(\omega)\mathrm{e}^{-ip(\omega)}$ where

$$g(\omega) = \left|c(\mathrm{e}^{i\omega})\right| \qquad [9]$$

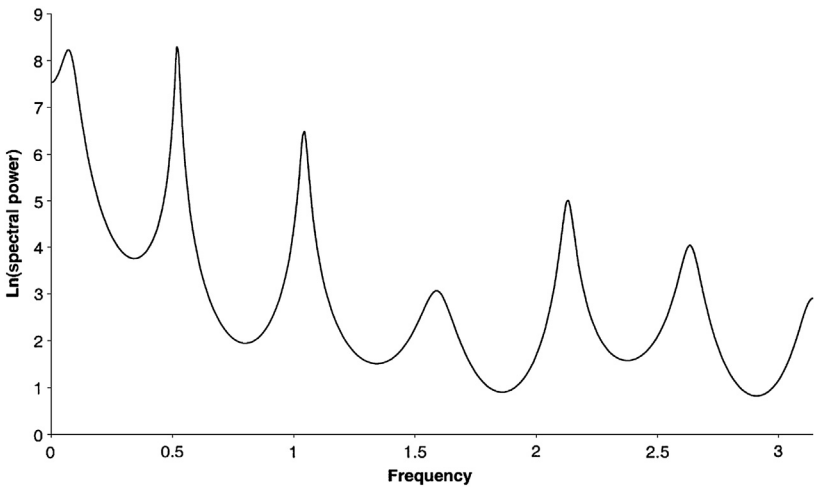

Figure 2 Spectrum of building permits.

and

$$\rho(\omega) = \tan^{-1}\left[\frac{-\mathrm{Im}\left[c\left(e^{-i\omega}\right)\right]}{\mathrm{Re}\left[c\left(e^{-i\omega}\right)\right]}\right] \qquad [10]$$

This representation for X_t shows that the filter $c(L)$ 'amplifies' each component of Y_t by the factor $g(\omega)$ and shifts it back in time by $\rho(\omega)/\omega$ time units.

The function $g(\omega)$ is called the filter *gain* (or sometimes the 'amplitude gain.' The function $\rho(\omega)$ is called the filter 'phase' and $g(\omega)^2 = [c(e^{-i\omega})c(e^{-i\omega})]$ is called the 'power transfer function' of the filter.

Three Examples

Example One: Lag Filter

Suppose that $c(L) = L^2$. Then $c(e^{-i\omega}) = e^{-2i\omega}$.

$$g(\omega) = \left|c\left(e^{i\omega}\right)\right| = 1$$

and

$$\rho(\omega) = \tan^{-1}\left[\frac{\sin 2\omega}{\cos 2\omega}\right] = 2\omega$$

Thus, this filter does not alter the amplitude of any of the components, $g(\omega) = 1$, but shifts each component back in time by $\rho(\omega)/\omega = 2$ time periods.

Example Two: First Difference Filter

Suppose that $c(L) = (1 - L)$ so that X_t is the first difference of Y_t. Then $c(e^{-i\omega}) = 1 - e^{-i\omega}$,

$$g(\omega) = \left(-e^{-i\omega} + 2 - e^{i\omega}\right)^{\frac{1}{2}} = \sqrt{2(1-\cos(\omega))}$$

and

$$\rho(\omega) = \tan^{-1}\left[\frac{\sin(\omega)}{1-\cos(\omega)}\right]$$

Thus, both the gain and phase are frequency specific. For example, $g(0) = 0$ so that the first difference filter eliminates the lowest frequency component of the series (the level of the series is 'differenced out'), and $g(\pi) = \sqrt{2}$ so that the high frequency components are amplified (X_t is 'choppier' than Y_t).

Example Three: Census X-12 Seasonal Adjustment Filter

The official monthly seasonal adjustment procedure in the USA and several other countries (Census X-12) can be linear filter $c(L) = \sum_{j=-84}^{84} c_{|j|}L^j$ where the coefficients c_i are given in Wallis (1974). Since the filter is symmetric, $c_j = c_{-j}$, $c(e^{i\omega})$ is real and so $\rho(\omega) = 0$. The function $g(\omega)$ is plotted in Figure 3. The gain is nearly unity except near the seasonal frequencies, and so the filter can be interpreted as producing a new series that leaves the non-seasonal components unaltered, but eliminates the seasonal components. For a detailed discussion of seasonal adjustment procedures *see* Time Series: Seasonal Adjustment.

Band-Pass Filters

The seasonal adjustment filter has the desirable properties of (essentially) zeroing out certain frequency components (the 'seasonal') and leaving the other components (the non-seasonals) unaltered. Filters with this characteristic are called 'band-pass filters,' since they 'pass' components associated with certain frequencies.

Band-pass filters for prespecified frequencies are easy to construct. For example, suppose that you want a filter with no phase shift that passes components between $0 \le \omega \le \underline{\omega}$. That is, you want to construct a filter, say $c(L)$ with gain $|g(\omega)| = 1$ for $0 \le \omega \le \underline{\omega}$ and $|g(\omega)| = 0$ for $\underline{\omega} \le \omega \le \pi$. (Assume that the corresponding negative frequencies are also passed.) Choosing a symmetric filter, $c_j = c_{-j}$, will make the phase 0 for all ω. In this case, $c(e^{i\omega})$ is a real number, and so the gain is given by:

$$g(\omega) = \sum_{j=-\infty}^{\infty} c_{|j|}e^{-i\omega j}$$

Now, using eqn [6], it is straightforward to verify the identity.

$$c_j = (2\pi)^{-1}\int_{-\pi}^{\pi} e^{-i\omega j}\sum_{j=-\infty}^{\infty} c_j e^{-i\omega j}\mathrm{d}\omega$$
$$= (2\pi)^{-1}\int_{-\pi}^{\pi} e^{i\omega j}\, c\left(e^{i\omega}\right)\mathrm{d}\omega$$

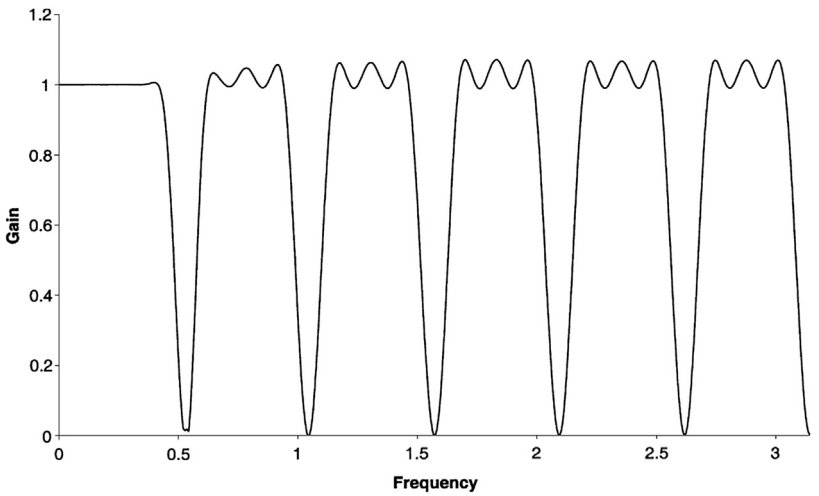

Figure 3 Gain of linear approximation to monthly × 12 seasonal adjustment filter.

Replacing $c(e^{i\omega}) = g(\omega)$ with the target value of the gain for the band-pass filter and carrying out the integration yield.

$$c_j = (2\pi)^{-1}\frac{1}{ij}e^{i\omega j}\Big|_{-\underline{\omega}}^{\underline{\omega}} = \left\{\begin{array}{ll} \frac{1}{j\pi}\sin\left(\underline{\omega}j\right) & \text{for} \quad j \neq 0 \\ \underline{\omega}/\pi & \text{for} \quad j = 0 \end{array}\right\}$$

The resulting filter $c(L)$ passes all components with frequencies between $0 \leq \omega \leq \underline{\omega}$. This means that the filter $1\text{–}c(L)$ passes everything except these frequencies. The filter $c(L)$ is called a low-pass filter (it passes low frequencies), and $1\text{–}c(L)$ is called a high-pass filter (it passes the high frequencies). Combinations of high- and low-pass filters can be used to construct band-pass filters for any set of frequencies. For example, a filter that passes components with frequencies between ω_1 to ω_2 (with $\omega_1 \geq \omega_2$) can be constructed as the difference of the low-pass filter for ω_2 and the low-pass filter for ω_1.

One practical problem with exact band-pass filters is that the coefficients c_j die out very slowly (at the rate $1/j$). This introduces important 'endpoint' problems when applying these filters to finite realizations of data, say $\{Y_t\}_{t=1}^{T}$. One approach is simply to truncate the filter at some point, for example to use $c_k(L) = \sum_{j=-k}^{k} c_j L^j$ and apply this filter to $\{Y_t\}_{t=k+1}^{T-k}$. An alternative procedure is to construct a minimum mean square error estimate of the infeasible band-pass values:

$$X_t = c(L)Y_t = \sum_{i=-\infty}^{\infty} c_i Y_{t-i},$$

as

$$E\left(X_t \middle| \{Y_j\}_{j=1}^{T}\right) = \sum_{i=-\infty}^{\infty} c_i E\left(Y_{t-i} \middle| \{Y_j\}_{j=1}^{T}\right)$$

This can be accomplished by using backcasts and forecasts for the missing pre-sample and post-samples values of Y_t. The relative merits of these two approaches for band-pass filters is discussed in Baxter and King (1999), Geweke (1978), and Dagum (1980) discuss this problem in the context of seasonal adjustment.

Spectra of Commonly Used Stochastic Processes

Suppose that Y has spectrum $S_Y(\omega)$ and $X_t = c(L)\,Y_t$. What is the spectrum of X? As was shown in eqn [8], the frequency components of X are the frequency components of Y scaled by the factor $g(\omega)e^{i\rho(\omega)}$, where $g(\omega)$ is the gain and $\rho(\omega)$ is the phase of $c(L)$. This means that spectra of X and Y are related by:

$$S_X(\omega) = g(\omega)^2 S_Y(\omega) = \left|c\left(e^{-i\omega}\right)\right|^2 S_Y(\omega)$$

which follows from $|e^{i\rho(\omega)}| = 1$ and the definition of $g(\omega)$.

Now, suppose that ε_t is a 'white noise' process, defined by the properties $E(\varepsilon_t) = 0$, $\lambda_0 = \sigma^2$, and $\lambda_k = 0$ for $k \neq 0$. The spectrum of ε is then easily calculated from eqn [7]:

$$S_\varepsilon(\omega) = (2\pi)^{-1}\sigma^2$$

So, the spectrum of white noise is constant. (Which is why the process is called 'white' noise.) Now suppose that $Y_t = c(L)\varepsilon_t$. Then:

$$S_Y(\omega) = \left|c\left(e^{-i\omega}\right)\right|^2 S_\varepsilon(\omega) = \left|c\left(e^{-i\omega}\right)\right|^2 (2\pi)^{-1}\sigma^2$$

This result can be used to determine the spectrum of any stationary ARMA process (*see* Time Series: ARIMA Methods) for a detailed discussion of these models). If Y_t follows an ARMA process, then it can be represented as:

$$\phi(L)Y_t = \theta(L)\varepsilon_t$$

The autoregressive operator, $\phi(L)$ can be inverted to yield $Y_t = c(L)\varepsilon_t$ with $c(L) = \phi(L)^{-1}\theta(L)$. This means that:

$$S_Y(\omega) = \left|c\left(e^{-i\omega}\right)\right|^2 (2\pi)^{-1}\sigma^2 = \frac{2(\pi)^{-1}\sigma^2\left|\theta\left(e^{-i\omega}\right)\right|^2}{\left|\phi\left(e^{-i\omega}\right)\right|^2} \qquad [11]$$

As an example, consider the AR(1) model:

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

equivalently written as,

$$(1 - \phi L)Y_t = \varepsilon_t.$$

Applying eqn [11] yields.

$$S_Y(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{\left|1 - \phi e^{-i\omega}\right|^2} = \frac{\sigma^2}{2\pi}\frac{1}{\left(1 + \phi^2 - 2\phi\cos(\omega)\right)}$$

This spectrum is equal to $\sigma^2[2\pi(1 + \phi^2 - 2\phi\cos(\omega))]^{-1}$ at $\omega = 0$. When $0 < \phi < 1$, it falls steadily as ω increases from 0 to π. This means that, relative to white noise, the low frequency components of the AR(1) are more important than the high frequency components. Thus, realizations of the series appear smoother than white noise.

Spectral Estimation

There are two general approaches to estimating spectra. Perhaps the simplest is to estimate an ARMA model for the series (as explained in Time Series: ARIMA Methods) and then compute the implied spectrum for this estimated model. Asymptotic approximations can then be used for statistical inference. Since the estimators of the ARMA parameters are asymptotically normally distributed and the spectrum is a smooth function of these parameters (see eqn [11]), the estimated spectrum at any point ω will be asymptotically normal with a variance that can be computed using the δ-method (a mean-value expansion).

An interesting extension of this procedure is discussed in Berk (1974). He shows that, quite generally, a suitably long autoregression can be specified and the spectrum estimated using the estimated AR coefficients. The precise definition of 'suitably long' depends on the size of the available sample, with more terms included in larger samples. These 'autoregressive spectral estimators' are shown by Berk to be consistent and approximately normally distributed in large samples.

An alternative set of estimators are based on non-parametric methods. These estimators can be motivated by eqn [7] which shows the spectrum as a weighted sum of the autocovariances. This suggests an estimator of the form:

$$S(\omega) = (2\pi)^{-1}\sum_{k=-l_T}^{l_T} w_T(k)\widehat{\lambda}_k$$

where $\widehat{\lambda}_k$ is an estimator of λ_k, l_T are truncation values that depend on the sample size T, and $w_T(k)$ is a weight function that

also depends on T. These estimators can be studied using nonparametric methods (*see* Nonparametric Statistics: Asymptotics).

A closely related set of nonparametric estimators are based on the 'periodogram'. The periodgram provides a frequency decomposition of the sample variance of a partial realization of the process, say $\{Y_t\}_{t=1}^{T}$. The periodgram is computed at frequencies $\omega_j = 2\pi j/T, j = 1, 2, \ldots, T/2$ (assuming T is even), and the value of the periodgram at frequency ω_j, is:

$$p_j = \frac{2}{T}\left|\sum_{t=1}^{T}(y_t - \bar{y})e^{-i\omega_j t}\right|^2$$

The periodgram has three interesting asymptotic sampling properties: (1) $E(p_j)/4\pi \rightarrow S_Y(\omega_j)$; (2) $var(p_j/2\pi) \rightarrow S_Y(\omega_j)^2$; and (3) $cov(p_j, p_k) = 0$ for $j \neq k$. Property (1) shows that a scaled version of the periodgram provides an asymptotically unbiased estimator of the spectrum, but since the variance doesn't approach zero (property (2)), the estimator is not consistent. The third property shows that averaging a set of periodgram ordinates around a given frequency reduces the variance of the estimator, although (from property (1)) it may introduce bias. By carefully selecting the averaging weights, consistent estimators with good finite sample mean square error properties can be constructed.

Some recent advances for nonparametric spectral estimators are developed in Andrews (1991).

See also: Nonparametric Statistics: Asymptotics; Time Series: ARIMA Methods; Time Series: Seasonal Adjustment.

Bibliography

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Baxter, M.B., King, R.G., 1999. Measuring business cycles: approximate band-pass filters for economic time series. Review of Economics and Statistics 81 (4), 575–593.

Berk, K.N., 1974. Consistent autoregressive spectral estimates. Annals of Statistics 2 (3), 489–502.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.

Cramér, H., 1942. On harmonic analysis of certain function spaces. Arkiv fur Matematik. Astronomi Och Fysik 28B (12), 1–7.

Dagum, E.B., 1980. The X-11 ARIMA Seasonal Adjustment Method. Research Paper, Statistics, Canada.

Geweke, J.F., 1978. The revision of seasonally adjusted time series. Proceedings of the Business and Economic Statistics Section – American Statistical Association, 320–325.

Granger, C.W.J., Newbold, P., 1977. Forecasting Economic Time Series. Academic Press, New York.

Hamilton, J., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Harvey, A., 1993. Time Series Models, second ed. MIT Press, Cambridge, UK.

Haydon, D.T., Shaw, D.J., Cattadori, I.M., Hudson, P.J., Thirgood, S.J., 2002. Analysing noisy time–series: describing regional variation in the cyclic dynamics of red grouse. Proceedings of the Royal Society of London. Series B: Biological Sciences 269 (1500), 1609–1617.

Kolmorgorov, A.N., 1940. Kurven in Hilbertschen Raum die gegenüber eine einparametrigen Gruppe von Bewegungen invariant sind. C.R. (Doklady) de L'Academie des Sciences de l'IRSS. New Series 26, 6–9.

Nerlove, M., Grether, D.M., Carvalho, J.L., 1979. Analysis of Economic Time Series. Academic Press, New York.

Niemira, M.P., Saaty, T.L., 2004. An analytic network process model for financial-crisis forecasting. International Journal of Forecasting 20 (4), 573–587.

Priestly, M.B., 1981. Spectral Analysis and Time Series. Academic Press, New York.

Peña, D., Tiao, G.C., Tsay, R.S., 2011. A Course in Time Series Analysis. John Wiley & Sons, New York.

Wallis, K.F., 1974. Seasonal adjustment and relations between variables. Journal of the American Statistical Association 69, 18–31.

Time Series: Economic Forecasting

James H Stock, Harvard University, Cambridge, MA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15721–15724, © 2001, Elsevier Ltd.

Abstract

Time-series forecasting uses historical empirical regularities, guided by a theoretical understanding of economic processes, to make future projections. Economic forecasts are used in a wide range of activities, including setting monetary and fiscal policies, state and local budgeting, financial management, and financial engineering. Key elements of economic forecasting include selecting the forecasting model(s) appropriate for the problem at hand, assessing and communicating the uncertainty associated with a forecast, and guarding against model instability.

Time-series forecasts are used in a wide range of economic activities, including setting monetary and fiscal policies, state and local budgeting, financial management, and financial engineering. Key elements of economic forecasting include selecting the forecasting model(s) appropriate for the problem at hand, assessing and communicating the uncertainty associated with a forecast, and guarding against model instability.

Time-Series Models for Economic Forecasting

Broadly speaking, statistical approaches to economic forecasting fall into two categories: time-series methods and structural economic models. Time-series methods use economic theory mainly as a guide to variable selection, and rely on past patterns in the data to predict the future. In contrast, structural economic models take as a starting point formal economic theory and attempt to translate this theory into empirical relations, with parameter values either suggested by theory or estimated using historical data. In practice, time-series models tend to be small with at most a handful of variables, while structural models tend to be large, simultaneous equation systems, which sometimes incorporate hundreds of variables (*see* Simultaneous Equation Estimation: Overview). Time-series models typically forecast the variable(s) of interest by implicitly extrapolating past policies into the future, while structural models, because they rely on economic theory, can evaluate hypothetical policy changes. In this light, perhaps it is not surprising that time-series models typically produce forecasts as good as, or better than, far more complicated structural models. Still, it was an intellectual watershed when several studies in the 1970s (reviewed in Granger and Newbold, 1986) showed that simple univariate time-series models could outforecast the large structural models of the day, a result that continues to be true (see McNees, 1990). This good forecasting performance, plus the relatively low cost of developing and maintaining time-series forecasting models, makes time-series modeling an attractive way to produce baseline economic forecasts.

At a general level, time-series forecasting models can be written,

$$y_{t+h} = g(X_t, \theta) + \varepsilon_{t+h} \qquad [1]$$

where y_t denotes the variable or variables to be forecast, t denotes the date at which the forecast is made, h is the forecast horizon, X_t denotes the variables used at date t to make the forecast, θ is a vector of parameters of the function g, and ε_{t+h} denotes the forecast error. The variables in X_t usually include current and lagged values of y_t. It is useful to define the forecast error in eqn [1] such that it has conditional mean zero, that is, $E(\varepsilon_{t+h}|X_t) = 0$. Thus, given the predictor variables X_t, under mean squared error loss the optimal forecast of y_{t+h} is its conditional mean, $g(X_t, \theta)$. Of course, this forecast is infeasible because in practice neither g nor θ is known. The task of the time-series forecaster therefore is to select the predictors X_t, to approximate g, and to estimate θ in such a way that the resulting forecasts are reliable and have mean squared forecast errors as close as possible to that of the optimal infeasible forecast.

Time-series models are usefully separated into univariate and multivariate models. In univariate models, X_t consists solely of current and past values of y_t. In multivariate models, this is augmented by data on other time series observed at date t. The next subsections provide a brief survey of some leading time-series models used in economic forecasting. For simplicity attention is restricted to one-step ahead forecasts ($h = 1$). Here, the focus is on forecasting in a stationary environment; the issue of nonstationarity in the form of structural breaks or time varying parameters is returned to below.

Univariate Models

Univariate models can be either linear, so that g is linear in X_t, or nonlinear. All linear time-series models can be interpreted as devices for modeling the covariance structure of the data. If $\{y_t\}$, $t = 1, \ldots, T$ has a Gaussian distribution and if this distribution is stationary (does not depend on time), then the optimal forecast is a linear combination of past values of the data with constant weights. Different linear time-series models provide different parametric approximations to this optimal linear combination.

The leading linear models are autoregressive models, autoregressive integrated moving average (ARIMA) models, and unobserved components models.

An autoregressive model of order p [AR(p)] is written,

$$y_{t+1} = \mu + \alpha_1 y_t + \cdots + \alpha_p y_{t-p+1} + \varepsilon_{t+1} = \mu + \alpha(L) y_t + \varepsilon_{t+1}$$

where (μ, α_1, …, α_p) are unknown parameters, L is the lag operator, and $\alpha(L)$ is a lag polynomial. If y_t is observed at all dates, $t = 1, \ldots, T$, these parameters are readily estimated by ordinary least squares.

ARIMA models extend autoregressive models to include a moving average term in the error, which has the effect of inducing long lags in the forecast when it is written as a function of current and past values of y_t. These are discussed in Time Series: ARIMA Methods.

In an unobserved components model, y_t is represented as the sum of two or more different stochastic components. These components are then modeled, and the combined model provides a parametric approximation to the autocovariances; see Harvey (1990) for a detailed modern treatment. The unobserved components framework can provide a useful framework for extracting cyclical components of economic time series and for seasonal adjustment (*see* Time Series: Seasonal Adjustment).

If $\{y_t\}$ is not Gaussian, in general, the optimal forecast will not be linear, which suggests the use of forecasts based on nonlinear time-series models. Parametric nonlinear time-series models posit a functional form g, and take this as known up to the finite dimensional parameter vector θ. Two leading examples of parametric nonlinear time-series models, currently popular for economic forecasting, are threshold autoregressive models and Markov switching models. Both are similar in the sense that they posit two (or more) regimes. In the threshold autoregression, switches between the regimes occur based on past values of the observed data; in Markov switching models, the switches occur based on an unobserved or latent variable.

The threshold autoregressive family of models can be written as,

$$y_{t+1} = d_t\alpha(L)y_t + (1 - d_t)\beta(L)y_t + \varepsilon_{t+1} \qquad [2]$$

where the mean is suppressed, $\alpha(L)$ and $\beta(L)$ are lag polynomials, and d_t is a nonlinear function of past data that switches between the 'regimes' $\alpha(L)$ and $\beta(L)$. Variations of eqn [2] interact d_t with only some autoregressive coefficients, the intercept, and/or the error variance. Various functions are available for d_t, either a sharp indicator function (the threshold autoregressive model) or a smooth function (smooth transition autoregression). For example, setting $d_t = [1 + \exp(\gamma_0 + \gamma_1'\zeta_t)]^{-1}$, yields the logistic smooth transition autoregression (LSTAR) model, where ζ_t denotes current or past data, say, ζ_t might equal y_{t-k} for some k. See Granger and Teräsvirta (1993) for additional details. For an application of threshold autoregressions (and other models) to forecasting the US unemployment, see Montgomery et al. (1998).

The Markov switching models are conceptually similar, except that the regime switch depends on an unobserved series. In the context of eqn [2], d_t is modeled as unobserved and following a two-state Markov process; see Hamilton (1994).

An alternative approach is to treat g itself as unknown, which leads to nonparametric methods. These include nearest neighbor, kernel, and artificial neural network models for the conditional expectation. In practice these nonparametric approaches to economic forecasting have met with success, which is mixed at best. There are several possible reasons for this, including that there are insufficiently many observations in typical economic data sets to support the use of these data-intensive methods.

Multivariate Models

In multivariate time-series models, X_t includes multiple time-series that can usefully contribute to forecasting y_{t+1}. The choice of these series is typically guided by both empirical experience and economic theory, for example, the theory of the term structure of interest rates suggests that the spread between long- and short-term interest rates might be a useful predictor of future inflation.

The multivariate extension of the univariate autoregression is the vector autoregression (VAR), in which a vector of time-series variables, Y_{t+1}, is represented as a linear function of $Y_t, \ldots, Y_{t-p+1}$, perhaps with deterministic terms (an intercept or trends). An interesting possibility arises in VARs that is not present in univariate autoregressions, specifically, it might be that the time series are cointegrated (that is, the individual series are nonstationary in the sense that they are integrated, but linear combinations of the series are integrated or order zero); *see* Times Series: Cointegration and Watson (1994).

Multivariate time-series models involve a large number of unknown parameters, a problem, which is greatly exacerbated when nonlinearities are introduced. Conceptually, the extension of univariate nonlinear models to the multivariate setting is straightforward. In practice, however, because of the relatively small number of time-series observations available to economic forecasters, it is unclear how best to implement nonlinear multivariate models and there are currently no definite conclusions in this area.

Models for Forecasting Higher Moments

The foregoing categorization of forecasting models has focused on prediction of conditional means. In some economic applications, especially in financial economics, interest is also in forecasting conditional higher moments. The most developed tools in this area are models of conditional variances. Two families of such models are autoregressive conditional heteroskedasticity models (Engle, 1982; see Bollerslev et al., 1994) and stochastic volatility models. Both approaches provide parametric models for the volatility of the series, which can then be forecasted given parameters estimated from historical data.

When Do Time-Series Forecasts Fail?

Regrettably, only by luck will economic forecasts be 'right': there are inevitable sources of uncertainty that make perfectly accurate predictions of the future impossible.

Shock, Model, and Estimation Uncertainty

Sources of uncertainty that comprise the error term ε_{t+h} in eqn [1] ('shock uncertainty') consist of unknowable future events, random measurement error in the series being forecast, and unforeseeable data revisions. Additional forecast

uncertainty arises from model approximation error (that is, error in the choice of g) and from error in the estimation of the model parameters θ.

Sensible use of economic forecasts requires that this forecast uncertainty be accurately communicated. Prediction intervals can be computed either directly based on historically estimated parameters and residuals, or by simulation for more complicated nonlinear models. Ideally these prediction intervals should incorporate parameter estimation uncertainty. There are several ways to do this, but a natural one is to construct forecast error distributions using a simulated out of sample prediction methodology. In this approach, the model is estimated using data through some date, say t', then a forecast is made for the next period ($t' + h$, for h-step ahead forecasts). The model is then reestimated using data through date $t' + 1$, and the forecast is made for $t' + h + 1$. This is repeated until a series of simulated out-of-sample forecast errors is constructed, which, in turn, can be used to construct prediction intervals around the point forecast of interest.

Overfitting

Overfitting poses a particular threat in economic forecasting given the impossibility of creating new data sets through experiments and given the relatively short number of observations on many economic time series. Overfitting can be addressed in part by relying on automatic methods for pruning a model. Leading such methods is model selection using information criteria such as the Akaike or Bayes (*see* Model Testing and Selection, Theory of). Alternatives include Bayes or Empirical Bayes estimation. Combining forecasts by taking averages or weighted averages can also be interpreted as a method for addressing the overfitting that arises in model selection.

Difficulties Modeling Trends

An important characteristic of economic data is that many economic time series contain considerable persistence, sometimes resembling a trend. Starting with the seminal work of Nelson and Plosser (1982), there has been considerable debate about whether these trends are best modeled as deterministic or stochastic (that is, arising from unit autoregressive roots), see Stock (1994). If a forecasting model incorrectly incorporates deterministic trends, but the series is highly persistent, spuriously large weight can be placed on an estimated trend component and out-of-sample forecasts can suffer. Such considerations lead to the use of methods that explicitly address the possibilities of stochastic trends, an example being preliminarily testing for a unit root in a univariate time series and constructing a univariate model in levels or in differences depending on the outcome of the pretest.

Model Instability

Clements and Hendry (1999) argue that many major failures of economic forecasts arise because of structural breaks to which conventional forecasting models fail to adapt. In some cases, model instability can arise because of slow changes in parameter values, perhaps arising from evolution in underlying technologies in the economy. In other, more dramatic cases, model instability appears as a sudden break, such as a change in monetary policy or the collapse of an exchange rate regime.

Whether structural change is gradual or abrupt, it is critical to monitor forecast performance and to test for model instability. Various statistical methods are available for detecting structural change; these typically involve testing for evidence of instability either in estimated coefficients or in simulated forecast errors. In practice, these methods need to be augmented by expert judgment that draws on information not incorporated in the model, such as knowledge of particularly severe weather (which could lead to a structural break being spuriously detected) or of a known change in a policy regime (so a break might be imposed even if it is not detected statistically). Unfortunately, expert judgment can be wrong, and knowing when to intervene in an economic forecast is a rather murky art.

Which Methods Work Best?

It is perhaps overly ambitious to hope for a simple prescription that provides a good forecasting method for all economic time series. Nonetheless, a number of studies have attempted to provide some general guidance on this question. Two general lessons emerge: among univariate models, simple linear models often forecast as well or better than more complicated nonlinear models; and the gains from moving from univariate to multivariate models are often (but not always) small.

This is illustrated by forecasts of the rate of price inflation over 6 months in the United States, as measured by the percentage rate of inflation of the Consumer Price Index (this example is taken from Stock and Watson's (1999) comparison of linear and nonlinear forecasting models for 215 US macroeconomic time series). Simulated out of sample, 6-month-ahead forecasts over the period March 1971–June 1996 (the full sample begins January 1959) produce a root mean square forecast error (RMSFE) of 2.44 percentage points (at an annual rate) for a univariate AR(4) fit recursively to the annualized rate of inflation. If the autoregressive lag length is selected recursively using the Bayes information criterion and a recursive unit root protest is used to determine whether the autoregression is fit to the rate of inflation or the change in the rate, the RMSFE drops to 2.05. Forecasts based on artificial neural networks and LSTAR models have RMSFEs exceeding 2.15 and in some cases exceeding 2.5, depending on the specification. A VAR that adds the unemployment rate and short-term interest rate (recursive Bayes information criterion lag selection) produces an RMSFE of 2.31. For these models, then, introducing nonlinearity or adding these other predictors does not improve upon a simple autoregression with a unit root pretest and data-dependent lag length selection.

The future of economic forecasting promises the development of new, computer-intensive methods and the real-time availability of very large data sets with complex patterns of historical dependence. These developments could prove invaluable for advancing economic forecasting, and the lessons from the past four decades of economic forecasting might not extend to these new data sets. Yet those lessons do underscore

the importance of having a simple and reliable benchmark forecast against which alternative forecasts can be assessed.

See also: Econometrics, History of; Explanation: Conceptions in the Social Sciences; Quantification in the History of the Social Sciences; Time Series: Advanced Methods; Time Series: General.

Bibliography

Bollerslev, T., Engle, R.F., Nelson, D.B., 1994. ARCH models. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Amsterdam, pp. 2959–3038.

Clements, M.P., Hendry, D.F., 1999. Forecasting Non-stationary Economic Time Series. MIT Press, Cambridge, MA.

Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50, 987–1008.

Granger, C.W.J., Newbold, P., 1986. Forecasting Economic Time Series, second ed. Academic Press, Orlando, FL.

Granger, C.W.J., Terasvirta, T., 1993. Modelling Non-linear Economic Relationships. Oxford University Press, Oxford, UK.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Harvey, A.C., 1990. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge, UK.

McNees, S.K., 1990. The role of judgment in macroeconomic forecasting accuracy. International Journal of Forecasting 6, 287–299.

Montgomery, A.L., Zarnowitz, V., Tsay, R.S., Tiao, G.C., 1998. Forecasting the US unemployment rate. Journal of the American Statistical Association 93, 478–493.

Nelson, C.R., Plosser, C.I., 1982. Trends and random walks in macroeconomic time series – some evidence and implications. Journal of Monetary Economics 10, 139–162.

Stock, J.H., 1994. Unit roots, structural breaks, and trends. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Amsterdam, pp. 2740–2843.

Stock, J.H., Watson, M.W., 1999. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, R.F., White, H. (Eds.), Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger. Cambridge University Press, Cambridge, UK, pp. 1–44.

Watson, M.W., 1994. Vector autoregressions and cointegration. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Amsterdam, pp. 2844–2915.

Time Series: General

David R Brillinger, University of California, Berkeley, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15724–15731, © 2001, Elsevier Ltd.

Abstract

A time series is a stretch of values on the same scale indexed by a time-like parameter. The basic data and parameters are functions. Time series take on a dazzling variety of shapes and forms, indeed there are as many time series as there are functions of real numbers.

A *time series* is a stretch of values on the same scale indexed by a time-like parameter. The basic data and parameters are functions.

Time series take on a dazzling variety of shapes and forms, indeed there are as many time series as there are functions of real numbers. Some common examples of time series forms are provided in **Figure 1**. One notes periods, trends, wandering, and integer values. The time series such as those in the figure may be contemporaneous and a goal may be to understand the interrelationships.

Concepts and fields related to time series include longitudinal data, growth curves, repeated measures, econometric models, multivariate analysis, signal processing, and systems analysis.

The field, time series analysis, consists of the techniques, which when applied to time series lead to improved knowledge. The purposes include summary, decision, description, and prediction.

The field has a theoretical side and an applied side. The former is part of the theory of stochastic processes (e.g., representations, prediction, information, limit theorems) while applications often involve extensions of techniques of 'ordinary' statistics, for example, regression, analysis of variance, multivariate analysis, and sampling. The field is renowned for jargon and acronyms – white noise, cepstrum, *autoregressive moving average* (ARMA), ARCH (see Granger, 1982).

Importance

> … but time and chance happeneth to them all.
> *Ecclesiastes*

Time series ideas appear basic to virtually all activities. Time series are used by nature and humans alike for communication, description, and visualization. Because *time* is a physical concept, parameters and other characteristics are mathematical models for time series can have real-world interpretations. This is of great assistance in the analysis and synthesis of time series.

Time series are basic to scientific investigations. There are: circadian rhythms, seasonal behaviors, trends, changes, and evolving behavior to be studied and understood. Basic questions of scientific concern are formulated in terms of time series concepts – Predicted value? Leading? Lagging? Causal connection? Description? Association? Autocorrelation? Signal? Seasonal effect? New phenomenon? Control? Periodic? Changing? Trending? Hidden period? Cycles?

Because of the tremendous variety of possibilities, substantial simplifications are needed in many time series analyses. These may include assumptions of stationarity, mixing or asymptotic independence, normality, and linearity. Luckily such assumptions often appear plausible in practice.

The subject of time series analysis would be important if for no other reason than that it provides means of examining the basic assumption of statistical independence invariably made in ordinary statistics. One of the first commonly used procedures for this problem was the Durbin–Watson test (Durbin and Watson, 1950). The autocovariance and spectrum functions, see below, are now often used in this context.

History

Contemporary time series analysis has substantial beginnings in both the physical and social sciences. Basic concepts have appeared in each subject and made their way to the other with consequent transfer of technology. Historical researchers important in the development of the field include the following: Lauritzen (1981), Hooker (1901), Einstein (1914), Wiener (1949), Yule (1927), Fisher (1929), Tukey (1984), Whittle (1951), and Bartlett (1966). Books particularly influential in the social sciences include Moore (1914) and Davis (1941). Nowadays many early analyses appear naive. For example, Beveridge in 1920 listed some 19 periods for a wheat price index running from 1500 to 1869 (Beveridge, 1922). It is hard to imagine the presence of so much regular behavior in such a series. When statistical uncertainty is estimated using latter-day techniques most of these periods appear insignificant; see the periodogram with 95% error bounds in **Figure 2**. Many historical references are included in Wold (1965).

Historians of science have made some surprising discoveries concerning early work with time series. An example is presented in Tufte (1983). He shows a time series plot from the tenth- or eleventh-century AD. This graph is speculated to provide movements of the planets and sun. It is remarkable for its cartesian type character. More generally Tufte remarks, following a study of newspapers and magazines, that "The time-series plot is the most frequently used form of graphic design." Casual observation certainly supports Tufte's study.

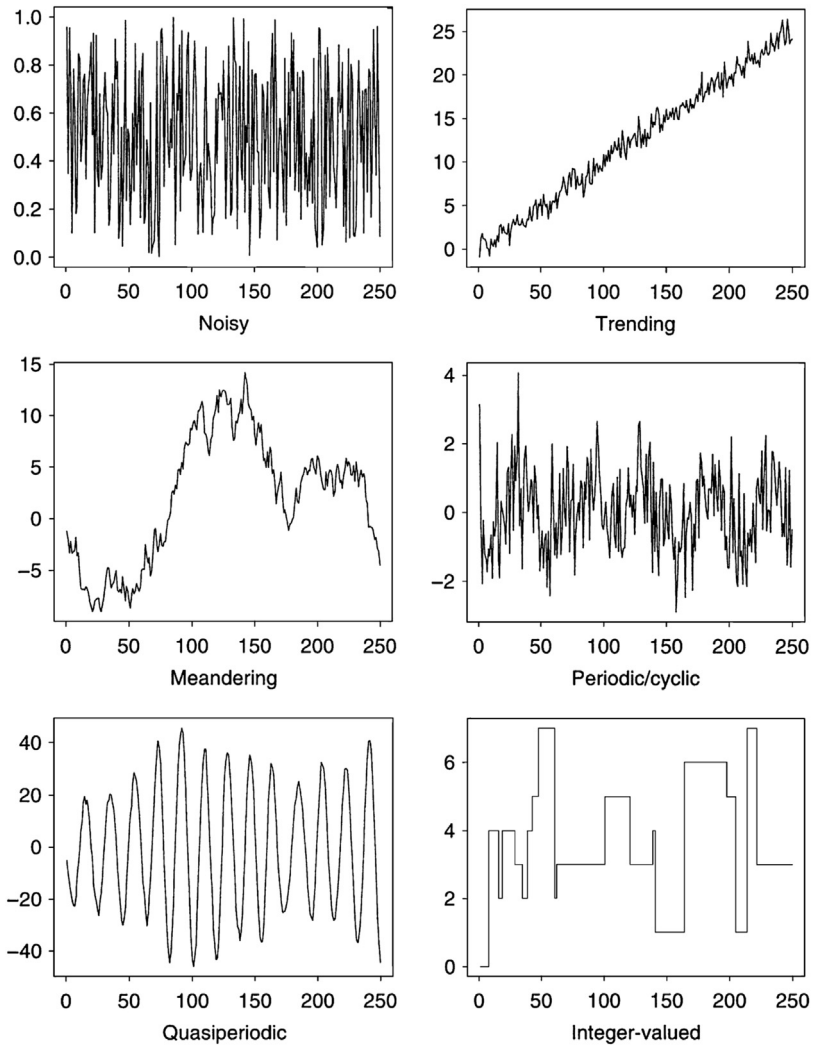

Figure 1 Some different types of time series.

Important problems that were addressed in the twentieth century include seasonal adjustment, leading and lagging indicators, and index numbers. Paradigms that were developed include:

$$\text{series} = \text{signal} + \text{noise}$$

$$\text{series} = \text{trend} + \text{seasonal} + \text{noise}$$

$$\text{series} = \text{sum of cosines} + \text{noise}$$

$$\text{series} = \text{regression function} + \text{noise}$$

These conceptualizations have been used for forecasting, seasonal adjustment, and description among other things. There are surprises, for example, ordinary least squares are surprisingly efficient in the time series case (see Grenander and Rosenblatt, 1957). Other books from the 1950s and 1960s that proved important are Blackman and Tukey (1959), Granger and Hatanaka (1964), and Box and Jenkins (1970). Important papers from the era include Akaike (1965), Hannan (1963), and Parzen (1961).

An important development that began in the 1950s and continues today is the use of recursive computation in the state space model; see Kalman (1963), Harvey (1990), Shumway and Stoffer (2000), and Durbin and Koopman (2001).

Basics

Concepts

There are a number of concepts that recur in time series work. Already defined is the *time series*, a stretch of values on the same scale indexed by a time parameter. The time parameter may range over the positive and negative integers or all real numbers or subsets of these. *Time series data* refer to a segment of a time series. A time series is said to have a *trend* when there is a slowly evolving change. It has a *seasonal* component when some cyclic movement of period 1 year is present. (The *period* of a cyclic phenomenon is the amount of time for it to repeat itself. Its *frequency* is the reciprocal of the period.)

There is an algebra of mathematical operations that either nature or an analyst may apply to a time series $\{X(t)\}$ to produce another series $\{Y(t)\}$. Foremost is the *filter*, or *linear*

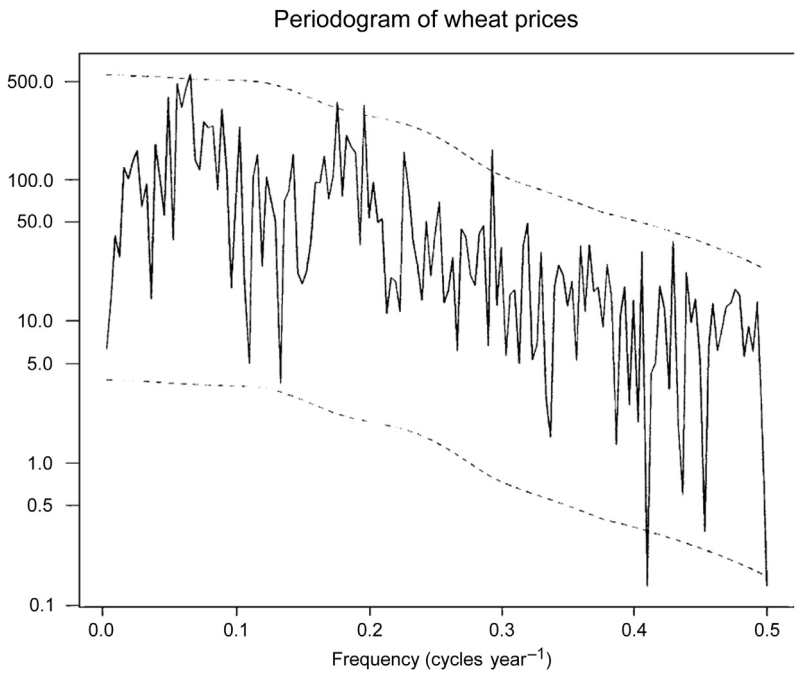

Figure 2 The dashed lines provide 95% bounds about a central curve.

time invariant operation or *system*. In the case of discrete equispaced time this may be represented as

$$Y(t) = \sum_u a(u)X(t-u)$$

t, $u = 0, \pm 1, \ldots$. An example is the running mean used to smooth a series. The functions $X(.)$, $Y(.)$ may be vector-valued and $a(.)$ may be matrix-valued. In the vector-valued case feedback may be present. The sequence $\{a(u)\}$ is called the *impulse response function*. The operation has the surprising property of taking a series of period P into a series of the same period P. The filter is called *realizable* when $a(u) = 0$ for $u < 0$. Such filters appear in causal systems and when the goal is prediction.

The above ideas concern both deterministic and random series. The latter prove to be important in developing solutions to important problems. Specifically it often proves useful to view the subject of time series as part of the general theory of stochastic processes, that is, indexed families of random variables. One writes $\{Y(t, \omega), t \text{ in } V\}$, with ω a random variable and V a set of times. Time series data are then viewed as a segment, $\{Y(t, \omega_0), t = 0, \ldots, T-1\}$ of the realization labeled by ω_0, the obtained value of ω. Stochastic time series $\{Y(t), t = 0, \pm 1, \pm 2, \ldots\}$ sometimes are conveniently described by *finite dimensional distributions* such as

$$\text{Prob}\{Y(t) \le y\} \text{ and } \text{Prob}\{Y(t_1) \le y_1, Y(t_2) \le y_2\}$$

This is particularly the case for time series values with joint normal distributions.

Time series may also be usefully described or generated by stochastic models involving the independent identically distributed random variables of ordinary statistics.

Stochastic models may be distinguished as parametric or nonparametric. Basic parameters of the nonparametric approach include moments, joint probability and density functions, mean functions, autocovariance and cross-covariance functions, and power spectra.

One basic assumption of time series analysis is that of *stationarity*. Here the choice of time origin does not affect the statistical properties of the process. For example, the mean level of a stationary series is constant. Basic to time series analysis is handling temporal dependence. To this end one can define the *cross-covariance function* of the series X and Y at lag u as the covariance of the values $X(t+u)$ and $Y(t)$. In the stationary case this function does not depend on t. In an early paper, Hooker (1901) computed an estimate of this quantity. Another useful parameter is the *power spectrum*, a display of the intensity or variability of the phenomenon versus period or frequency. It may be defined as the Fourier transform of the autocovariance function. The power spectrum proves useful in displaying the serial dependence present, in discovering periodic phenomena and in diagnosing possible models for a series.

In the parametric case there are the ARMA series. These are regression-type models expressing the value $Y(t)$ as a linear function of a finite number of past values $Y(t-1)$, $Y(t-2)$, … of the series and the values $\varepsilon(t)$, $\varepsilon(t-1)$, … of a sequence of independent identically distributed random variables. ARMAs have proved particularly useful in problems of forecasting.

Contemporary models for time series are often based upon the idea of *state*. This may be defined as the smallest entity summarizing the past history of the process. There are two parts to the state space model. First, a state equation that describes how the value of the state evolves in time. This equation may contain a purely random element. Second there is an equation indicating how the available measurements at that time t come from the state value at that time. It too may involve a purely random element. The concept of the state of a system has been basic to

physics for many years in classical dynamics and quantum mechanics. The idea was seized upon by control engineers, for example, Kalman (1963) in the late 1950s. The econometricians realized its usefulness in the early 1980s (see Harvey, 1990).

A number of specific probability models have been studied in depth, including the Gaussian, the ARMA, the bilinear (Granger and Andersen, 1978), various other nonlinear (Tong, 1990), long and short memory, ARMAX, ARCH (Einstein, 1914), hidden Markov (MacDonald and Zucchini, 1997), random walk, stochastic differential equations (Guttorp, 1995; Prakasa Rao, 1999), and the periodically stationary.

A list of journals where these processes are often discussed is included at the end of this article.

Problems

There are scientific problems and there are associated statistical problems that arise. Methods have been devised for handling many of these. The scientific problems include smoothing, prediction, association, index numbers, feedback, and control. Specific statistical problems arise directly. Among these are including explanatories in a model, estimation of parameters such as hidden frequencies, uncertainty computation, goodness of fit, and testing.

Special difficulties arise. These include missing values, censoring, measurement error, irregular sampling, feedback, outliers, shocks, signal-generated noise, trading days, festivals, changing seasonal pattern, aliasing, data observed in two series at different time points.

Particularly important are the problems of association and prediction. The former asks the questions of whether two series are somehow related and what is the strength of any association. Measures of association include the cross-correlation and the coherence functions. The prediction problem concerns the forecasting of future values. There are useful mathematical formulations of this problem but because of unpredictable human intervention there are situations where guesswork seems just as good.

Theoretical tools employed to address the problems of time series analysis include mathematical models, asymptotic methods, functional analysis, and transforms.

Methods

Descriptive

Descriptive methods are particularly useful for exploratory and summary purposes. They involve graphs and other displays, simple statistics such as means, medians, and percentiles, and the techniques of exploratory data analysis (Tukey, 1977).

The most common method of describing a time series is by presenting a graph, see **Figure 1**. Such graphs are basic for communication and assessing a situation. There are different types. Cleveland (1985) mentions the connected, symbol, connected-symbol, and vertical-line displays in particular. **Figure 1** presents connected graphs.

Descriptive values derived from time series data include extremes, turning points, level crossings, and the periodogram (see **Figure 2** for an example of the latter). Descriptive methods typically involve generating displays using manipulations of time series data via operations such as differencing, smoothing, and narrowband filtering.

A display with a long history (Laplace, 1825; Wold, 1965) is the Buys Ballot table. Among other things it is useful for studying the presence and character of a phenomenon of period P such as a circadian rhythm. One creates a matrix with entry $Y[(i-1)P+j]$ in row i, column $j = 1, \ldots, P$ and then, for example, one computes column averages. These values provide an estimate of the effect of period P. The graphs of the individual rows may be stacked beneath each other in a display. This is useful for discerning slowly evolving behavior.

Descriptive methods may be time-side (as when a running mean is computed), frequency-side (as when a periodogram is computed), or hybrid (as when a sliding window periodogram analysis is employed).

Parameter Estimation

The way to a solution of many time series problems is via the setting down of a stochastic model. Parameters are constants of unknown values included in the model. They are important because substantial advantages arise when one works within the context of a model. These advantages include estimated standard errors, efficiency, and communication of results. Often parameter estimates are important because they are fed into later procedures, for example, formulas for forecasting.

General methods of estimating parameters include method of moments, least squares, and maximum likelihood. An important time series case is that of harmonic regression. It was developed in Fisher (1929) and Whittle (1951).

There are parametric and nonparametric approaches to estimation. The parametric has the advantage that if the model is correct, then the estimated coefficients have considerably smaller standard errors.

Uncertainty Estimation

Estimates without some indication of their uncertainty are not particularly useful in practice. There are a variety of methods used in time series analysis to develop uncertainty measures. If maximum likelihood estimation has been used there are classic (asymptotic) formulas. The delta method or method of linearization is useful if the quantity of interest can be recognized to be a smooth function of other quantities whose variability can be estimated directly. Methods of splitting the data into segments, such as the jackknife, can have time series variants. A method currently enjoying considerable investigation is the bootstrap (Davison and Hinkley, 1997). The assumptions made in validating these methods are typically that the series involved is stationary and mixing.

Seasonal Adjustment

Seasonal adjustment may be defined as the attempted removal of obscuring unobservable annual components. There are many methods of seasonal adjustment (National Bureau of Economic Research, 1976), including state space approaches (Hannan, 1963).

The power spectrum provides one means of assessing the effects of various suggested adjustment procedures.

System Identification

System identification refers to the problem of obtaining a description or model of a system on the basis of a stretch of input to and the corresponding output from a system of interest. The system may be assumed to be linear time invariant as defined above.

In designed experiments the input may be a series of pulses, a sinusoid, or white noise. In a natural experiment the input is not under the control of the investigator and this leads to complications in the interpretation of the results. System identification relates to the issue of causality. In some systems one can turn the input off and on and things are clearer.

Computing

Important computer packages for time series analysis include Splus, Matlab, Mathematica, SAS, SPSS, TSP, and STAMP. Some surprising algorithms have been found: the fast Fourier transform (FFT), fast algorithms from computational geometry, Monte Carlo methods, and the Kalman–Bucy filter. Amazingly a variant of the latter was employed by Thiele in 1880 (Lauritzen, 1981), while the FFT was known to Gauss in the early 1800s (Heideman et al., 1984).

Current Theory and Research

Much of what is being developed in current theory and research is driven by what goes on in practice in time series analyses. What is involved in a time series analysis? The elements include the question, the experiment, the data, plotting the data, the model, model validation, and model use. The importance of recognizing and assessing the basic assumptions is fundamental.

The approach adopted in practice often depends on how much data are available. In the case that there are a lot of data even procedures that are, in some sense, inefficient are often used effectively. A change from the past is that contemporary analysis often results from the appearance of very large fine data sets. The amount of data can seem limitless as, for example, in the case of records of computer tasks. There are many hot research topics. One can mention: exploratory data analysis techniques for very large data sets, so-called improved estimates, testing (association? cycle present?), goodness of fit diagnostics. There are the classical and Bayesian approaches, the parametric, semiparametric, and nonparametric analyses, the problem of dimension estimation and that of reexpression of variables.

Current efforts include research into bootstrap variants, long-memory processes, long-tailed distributions, non-Gaussian processes, unit roots, nonlinearities, better approximations to distributions, demonstrations of efficiency, self-similar processes, scaling relationships, irregularly observed data, doubly stochastic processes as in hidden Markov, cointegration, disaggregation, cyclic stationarity, wavelets, and particularly inference for stochastic differential equations.

Today's time series data values may be general linear model type (Fahrmein and Tutz, 1994), for example, ordinal, state-valued, counts, proportions, angles, and ranks. They may be vectors. They may be functions. The time label t may be location in space or even a function. The series may be vector-valued. The data may have been collected in an experimental design.

There are some surprises: the necessity of tapering and prewhitening to reduce bias, the occurrence of aliasing, the high efficiency of ordinary least squares estimates in the correlated case, and the approximate independence of empirical Fourier transform values at different frequencies (Brillinger, 1975).

Future Directions

It seems clear that time series research will continue on all the existing topics as the assumptions on which any existing problem solution has been based appear inadequate. Further, it can be anticipated that more researchers from nontraditional areas will become interested in the area of time series as they realize that the data they have collected, or will be collecting, are correlated in time.

Researchers can be expected to be even more concerned with the topics of nonlinearity, conditional heteroscedasticity, inverse problems, long memory, long tails, uncertainty estimation, inclusion of explanatories, new analytic models, and properties of the estimates when the model is not true. The motivation for the latter is that time series with unusual structure seem to appear steadily. An example is the data being collected automatically via the World Wide Web. Researchers can be anticipated to be seeking extensions of existing time series methods to processes with more general definitions of the time label – spatial, spatial-temporal, functional, and angular. At the same time they will work on processes whose values are more general, even abstract.

More efficient, more robust, and more applicable solutions will be found for existing problems. Techniques will be developed for dealing with special difficulties such as missing data, nonstationarity, and outliers. Better approximations to the distributions of time series–based statistics will be developed.

Many have stressed the advantages of linear system identification via white noise input. Wiener (1958) stressed the benefits of using Gaussian white noise input. This idea has not been fully developed. Indeed data sets obtained with this input will continually yield to novel analytic methods as they make their appearance.

The principal journals in which newly developed statistical methods for time series are presented and studied include *Journal of Time Series Analysis*, *Annals of Statistics*, *Stochastic Processes and their Applications*, *Journal of the American Statistical Association*, *Econometrica*, and *IEEE Transactions in Signal Processing*.

See also: Statistics, History of; Time Series: Advanced Methods; Time Series: Seasonal Adjustment.

Bibliography

Akaike, H., 1965. On the statistical estimation of the frequency response function of a system having multiple input. Annals of the Institute of Statistics and Mathematics 17, 185–210.

Bartlett, M.S., 1966. Stochastic Processes. Cambridge University Press, Cambridge, UK.

Beveridge, W.H., 1922. Wheat prices and rainfall in western Europe. Journal of the Royal Statistical Society 85, 412–459.

Blackman, R.B., Tukey, J.W., 1959. The Measurement of Power Spectra from the Point of View of Communications Engineering. Dover, New York.
Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: *Forecasting* and Control. Holden-Day, San Francisco.
Brillinger, D.R., 1975. Time Series: Data Analysis and Theory. Holt, Rinehart, and Winston, New York.
Cleveland, W.S., 1985. The Elements of Graphing Data. Wadsworth, Belmont, CA.
Davis, H.T., 1941. The Analysis of Economic Time Series. Principia, Bloomington, IN.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge, UK.
Durbin, J., Koopman, S.J., 2001. Time Series Analysis by State Space Methods. Oxford University Press, Oxford, UK.
Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression. Biometrika 37, 409–428.
Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression. Biometrika 38, 159–177.
Einstein, A., 1914. Méthode pour la détermination des valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. Archives of Sciences Physiques et Naturelles 37 (4), 254–255.
Engle, R.F. (Ed.), 1995. ARCH Selected Readings. Oxford University Press, Oxford, UK.
Fahrmein, L., Tutz, G., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York.
Fisher, R.A., 1929. Tests of significance in harmonic analysis. Proceedings of the Royal Society of London A 125, 54–59.
Granger, C.W.J., 1982. Acronyms in time series analysis. Journal of Time Series Analysis 2, 103–108.
Granger, C.W.J., Andersen, A.P., 1978. An Introduction to Bilinear Time Series Models. Vandenhoeck und Ruprecht, Gottingen.
Granger, C.W.J., Hatanaka, M., 1964. Spectral Analysis of Economic Time Series. Princeton University Press, Princeton, NJ.
Grenander, U., Rosenblatt, M., 1957. Statistical Analysis of Stationary Time Series. Wiley, New York.
Guttorp, P., 1995. Stochastic Modelling of Scientific Data. Chapman & Hall, London.
Hannan, E.J., 1960. Time Series Analysis. Methuen, London.
Hannan, E.J., 1963. Regression for time series. In: Rosenblatt, M. (Ed.), Time Series Analysis. Wiley, New York, pp. 17–37.
Harvey, A.C., 1990. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge, UK.
Harvey, A.C., Koopman, S.J., Penzer, J., 1998. Messy time series: a unified approach. Advances in Econometrics 13, 103–143.
Heideman, M.T., Jondon, D.H., Burrus, C.S., 1984. Gauss and the history of the fast Fourier transform. Archives in History and Exact Science 34, 265–277.
Hooker, R.H., 1901. Correlation of the marriage-rate with trade. Journal of the Royal Statistical Society 64, 696–703.
Kalman, R.E., 1963. Mathematical description of linear dynamical systems. SIAM Journal of Control 1, 152–192.
Laplace, P.S., 1825. Traité de Méchanique Céleste. Bachelier, Paris.
Lauritzen, S.L., 1981. Time series analysis in 1880: a discussion of contributions made by T.N. Thiele. International Statistical Review 49, 319–331.
MacDonald, I.L., Zucchini, W., 1997. Hidden Markov and Other Models for Discrete-Valued Time Series. Chapman and Hall, London.
Moore, H.L., 1914. Economic Cycles: Their Law and Cause. Macmillan, New York.
National Bureau of Economic Research/Bureau of the Census, 1976. Conference on Seasonal Analysis of Economic Time Series. Dept. of Commerce, Bureau of the Census, Washington, DC.
Parzen, E., 1961. An approach to time series analysis. Annals of Mathematical Statistics 32, 951–989.
Prakasa Rao, B.L.S., 1999. Statistical Inference for Diffusion Type Processes. Arnold, London.
Shumway, R.H., Stoffer, D.S., 2000. Time Series Analysis and Its Applications. Springer, New York.
Tong, H., 1990. Non-Linear Time Series: A Dynamical System Approach. Oxford University Press, New York.
Tufte, E.R., 1983. The Elements of Graphing Data. Graphics Press, Cheshire, CT.
Tukey, J.W., 1977. Exploratory Data Analysis. Addision-Wesley, Reading, MA.
Tukey, J.W., 1984. The Collected Works of John W. Tukey I–II Time Series. Wadsworth, Belmont, CA.
Whittle, P., 1951. Hypothesis Testing in Time Series Analysis. Almqvist and Wiksell, Uppsala, Sweden.
Wiener, N., 1949. Time Series. MIT Press, Cambridge, MA.
Wiener, N., 1958. Nonlinear Problems in Random Theory. MIT Press, Cambridge, MA.
Wold, H.O.A., 1965. Bibliography on Time Series and Stochastic Processes. MIT Press, Cambridge, MA.
Yule, G.U., 1927. On the method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. Philosophical Transactions A 226, 267–298.

Time Series: Seasonal Adjustment

Estela B Dagum, University of Bologna, Bologna, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article discusses the main characteristics of seasonal variations, common statistical seasonal models and the seasonal adjustment methods widely applied by statistical agencies. Seasonal adjustment means the removal of seasonal variations and plays a significant role in standardizing time series for the analysis of current social and economic conditions, particularly in determining the stages of the business cycle at which the economy stands. Such knowledge provides the bases for decision making to control the level of economic activity and it is of major importance around cyclical turning points.

The best known *seasonal adjustment methods* belong to the following types: (1) regression methods which assume *global or local* simple functions of time, (2) *stochastic-model-based* methods which assume simple autoregressive integrated moving average (ARIMA) models, and (3) moving average methods which are based on linear filtering and hence do not have *explicit parametric* models. Moving average and ARIMA model based methods are those applied by statistical agencies to produce officially seasonally adjusted series.

Introduction

A time series consists of a set of observations ordered in time, on a given phenomenon (target variable). Usually the measurements are equally spaced, e.g., by year, quarter, month, week, day. The most important property of a time series is that the ordered observations are dependent through time, and the nature of this dependence is of interest in itself. Examples of time series are the gross national product, the unemployment rate, or the daily closing value of the Dow Jones index.

Time series of sub-yearly observations, e.g., monthly, quarterly, weekly, are often affected by seasonal variations. The presence of such variations in socio–economic activities has been recognized for a long time. Indeed seasonality usually accounts for most of the total variation within the year.

Seasonality is due to the fact that some months or quarters of the year are more important in terms of activity or level. This originates from climate and conventional seasons, like religious, social, and civic events, which repeat from year to year.

The climatic seasons influence trade, agriculture, the consumption patterns of energy, fishing, mining, and related activities. For example, in North America, the level of unemployment is generally higher during the winter and spring months and lower in the other months. Similarly, the consumption of heating oil increases in winter, and the consumption of electricity increases in the summer months because of air conditioning.

Institutional seasons like Christmas, Easter, civic holidays, the school and academic year have a large impact on employment, retail trade and on the consumption of certain goods and services, namely travel by plane, hotel occupancy, consumption of gasoline.

Yearly series cannot contain seasonality because the component is specified to cancel out over 12 consecutive months or 4 consecutive quarters.

Flow series may also be affected by other kind of seasonal variations associated with the composition of the calendar. The most important calendar variations are the trading-day variations, which are due to the fact that some days of the week are more important than others. Trading-day variations imply the existence of a daily pattern analogous to the seasonal pattern but these daily factors are usually referred to as *daily coefficients*.

Similarly, the *moving-holiday* or *moving-festival* component is attributed to calendar variations, namely due to the fact that some holidays span consecutive months from year to year. For example, Easter can fall between 23 March and 25 April. These festivals affect flow and stock variables and may cause a displacement of activity from 1 month to the previous or the following month. For example, an early date of Easter in March or early April can cause an important excess of activity in March and a corresponding short-fall in April, in variables associated to imports, exports, and tourism. When the Christian Easter falls late in April (e.g., beyond the 10th), the effect is captured by the seasonal factor of April. Note that in the long run, Easter falls in April 11 times out of 14.

Seasonal adjustment means the removal of seasonal variations in the original series jointly with trading day variations and moving holiday effects. The main reason for seasonal adjustment is the need of standardizing socioeconomic series because seasonality affects them with different timing and intensity. Hence, the seasonally adjusted data reflect variations due only to the remaining components, namely trend-cycle and irregulars.

The information given by seasonally adjusted series plays a crucial role in the analysis of current economic conditions and provides the basis for decision making and forecasting, being of major importance around cyclical turning points.

Seasonal methods can be classified as *deterministic* or *stochastic* depending on the assumptions made concerning how seasonality evolves through time. *Deterministic* methods assume that the seasonal pattern can be predicted with no error or with zero variance of the prediction error. On the contrary, *stochastic* methods assume that seasonality can be represented by a *stochastic process*, a process governed by a probability law and, consequently, the variance of the prediction error is not null. The best known *seasonal adjustment methods* belong to the following types: (1) regression methods which assume *global or local* simple functions of time, (2) *stochastic-model-based*

methods which assume simple autoregressive integrated moving average (ARIMA) models, and (3) moving average methods which are based on linear filtering and hence do not have *explicit parametric* models. Only methods (1), which assume *global* simple functions for each component, are *deterministic*; the others are considered *stochastic*. Moving average or linear filtering methods are those applied mostly by statistical agencies to produce officially seasonally adjusted series.

This article discusses the main characteristics of seasonal variations, common statistical seasonal models and the seasonal adjustment methods widely applied by statistical agencies.

The Main Characteristics of Seasonality

The impact of seasonality in socioeconomic activities has long been recognized. Seasonal variations in agriculture, the low level of winter construction and high pre-Christmas retail sales are all well known.

The main causes of seasonality are mainly exogenous to economic systems but human intervention can modify their extent and nature. Thus, seasonal variations in the automobile industry are affected by manufacturers' decisions regarding the extent of model changeover each year; seasonality in the money supply and related series can be controlled by central banks' decisions on the interest rate. Another characteristic of seasonal variations is that although they repeat annually with certain regularity, they may evolve. Many causes can produce an evolving seasonality, for example, changes in technology alter the importance of climatic factors; customs and habits are affected by different distributions of income and thus, can modify the seasonal demand pattern for certain goods. For most time series, moving seasonality is more the rule than the exception. Depending on the main causes of the seasonal variations, these can change slowly or rapidly, in a deterministic or a stochastic manner or a mixture of both. A third important feature of seasonality is the feasibility of its identification even if it is a latent (non-directly observable) variable. The identification and estimation of seasonality, however, is not done independently of the other components affecting the series under study. Traditionally, three other components have been distinguished, namely; the trend, the cycle, and the irregulars. The trend reflects long term movements lasting many years. It is generally associated with structural causes, for example, institutional events, demographic and technological changes, new ways of organization. The cycle, usually referred to as business cycle, is a quasi-periodic oscillation characterized by periods of expansion and contraction. It lasts from 3 to 5 years on average. In much analytical work, the trend and the cycle are combined because, for series covering a short period of time, the long-term trend cannot be estimated adequately. Finally, the irregulars represent unforeseeable movements related to events of all kinds. In general, the irregular component has a stable random appearance but, in some cases, it may include extreme variations or outliers with well identifiable causes, for example, floods, severe winters, strikes.

The seasonal variations can be distinguished from the trend by their oscillatory character, from the business cycle by having annual periodicity and from the irregulars by being systematic. The four main causes of seasonality are attributed to the weather, composition of the calendar, major institutional deadlines and expectations. The changing industrial mix of an economy transforms the seasonal pattern because some industries are more seasonal than others. In particular, economies which diversify and depend less on 'primary' industries (e.g., fishing, agriculture) typically become less seasonal. As for institutional change, the extension of the academic year to the summer months in the 1970s affected the seasonal pattern of Canada unemployment for the population of 15–25 years of age. Similarly the practice of spreading holidays over the whole year impacted on seasonality.

The seasonal pattern basically repeats from year to year. Merely repeating the seasonal pattern of the last 12 months usually provides a reasonable forecast but in most situations, seasonality evolves even though slowly and gradually.

Another important characteristic of seasonality is the feasibility of its identification even if it is a latent (not directly observable) variable. The identification and estimation of seasonality, however, is not done independently of the other components affecting the time series under study, trend, business cycle, trading day variations, moving holidays and irregulars.

In order to determine whether a series contains seasonality, it is sufficient to identify at least 1 month (or quarter) which tends to be systematically higher or lower than other month.

For illustrative purposes, Figure 1 shows the pattern of an observed time series (US Unemployed Males, from January 1990 till December 2013) which is affected strongly by both seasonality and the business cycle, but their effects cannot be clearly inferred by simply looking at the graph.

The various components are shown in Figures 2–4. Figure 2 exhibits the seasonal variations that are characterized by two peaks and one trough. A large peak is observed in January when unemployment is highest and a small secondary one in June when young people have left school and look for work whereas the trough corresponds to the month of September.

Seasonality entails large costs to society and businesses. One cost is the necessity to build warehouses to store inventories of goods to be sold as consumers require them, for example grain elevators. Another cost is the under-use and over-use of the factors of production: capital and labor.

Capital in the form of un-used equipment, buildings and land during part of the year has to be financed regardless. For example, this is the case in farming, food processing, tourism, electrical generation. The cold climate increases the cost of buildings and infrastructure, e.g., roads, transportation systems, water and sewage systems, schools, hospitals; not to mention the damage to the same caused by the action of ice.

The labor force is over-used during the peak seasons of agriculture and construction for example; and, under-used in trough seasons sometimes leading to social problems.

The seasonal pattern measures the relative importance of the months of the year. The constant 100% represents an average month or a non-seasonal month. The *peak* month in January 1990 is 117.5 or 17.5% larger than on an average month; the *trough* months is September with 89.3 in 1990, almost 11% lower than on an average month. The *seasonal amplitude*, the difference between the peak and trough months of the seasonal pattern, has changed from 28.5% in 1990 to

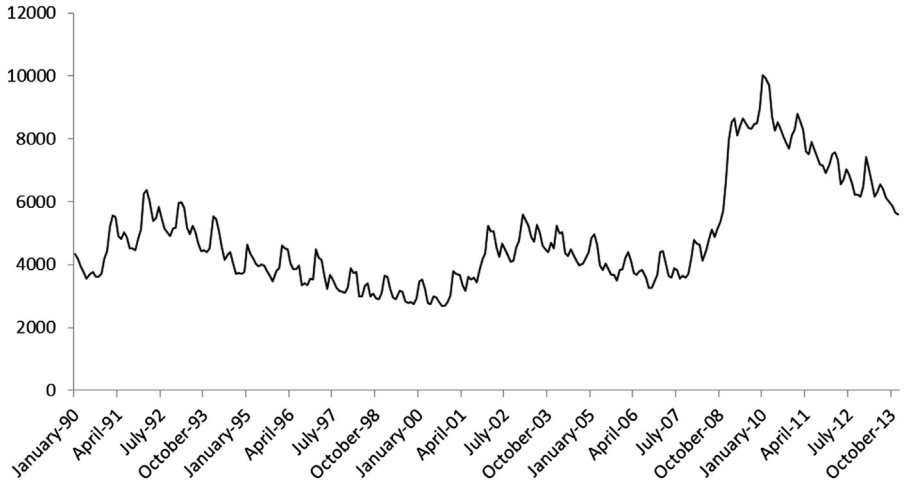

Figure 1 Original series (thousands) US unemployment level – Male.

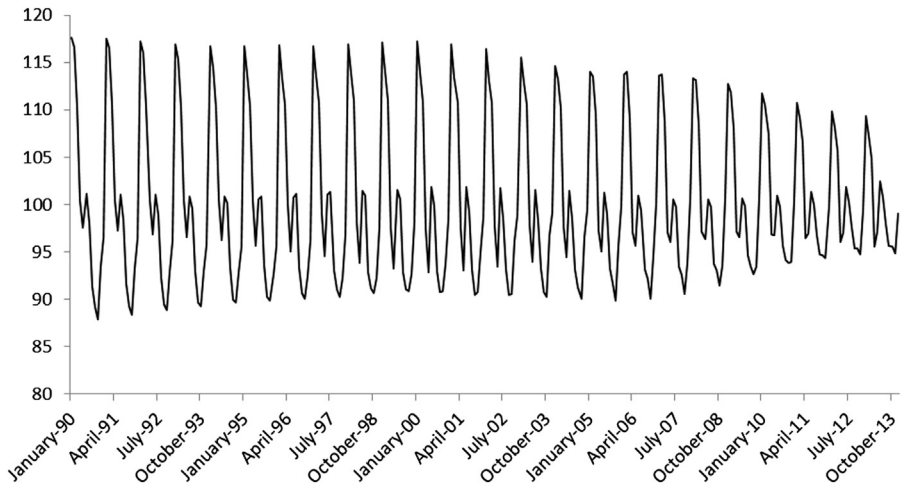

Figure 2 Seasonal factors (%) US unemployment level – Male.

15% in 2013. It is apparent that the seasonal amplitude is decreasing since 2003.

The effects of the trend-cycle are shown in Figure 3 with three large jumps in 1992, 2003, and 2009 which are recession years. The irregulars exhibited in Figure 4 are generally small except for a few cases where extreme values are present. Finally, the seasonally adjusted data are shown in Figure 5 clearly indicating that the impact of seasonality has been removed from the original series. The seasonally adjusted series contains the effects of the trend-cycle and those of the irregular component and, therefore, it is less smooth than the trend-cycle shown in Figure 3.

Seasonal Adjustment Methods

The seasonal adjustment methods developed so far are based on univariate time series decomposition models with no causal explanation. It is difficult to classify existing methods into mutually exclusive categories. However, it is possible to group the majority of seasonal adjustment methods into two main classes, one based on moving averages or linear filters and the other on explicit models with a small number of parameters for each component.

Moving Average Methods

The best known and most often applied seasonal adjustment methods are based on moving averages or smoothing linear filters applied sequentially by adding (and subtracting) one observation at a time. These methods assume that the time series components change through time in a stochastic manner. Given a time series, x_t, $t = 1, \ldots, T$, for any t far removed from both ends, say $m + 1 \leq t \leq T - m$, the seasonally adjusted value x_t^a, is obtained by application of a symmetric moving average h_m (B) applied to $2m + 1$ observations surrounding the target point t, x_{t+j}, $j = -m, \ldots, m$

$$x_t^a = h_m(\mathrm{B})x_t = \sum_{j=-m}^{m} h_{m,j} x_{t-j} \qquad [1a]$$

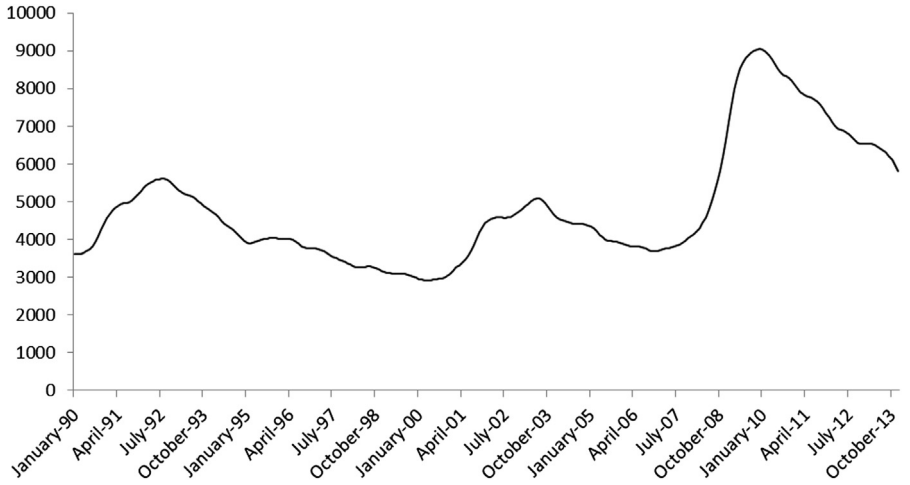

Figure 3 Trend-cycle component (thousands) US unemployment level – Male.

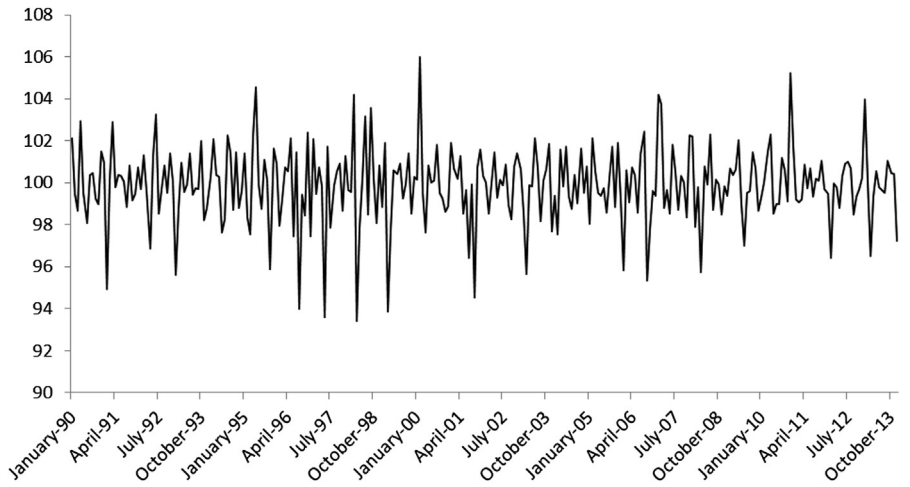

Figure 4 Irregular component (%) US unemployment level – Male.

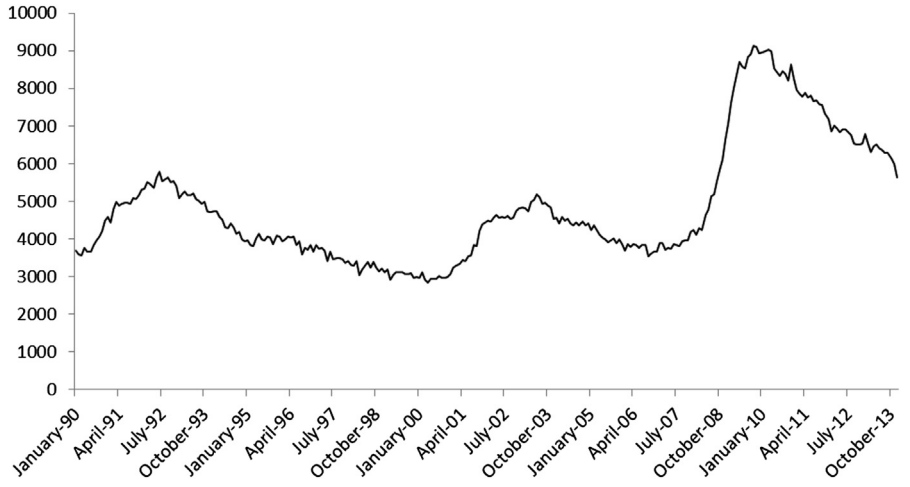

Figure 5 Seasonally adjusted series (thousands) US unemployment level – Male.

where the weights $h_{m,j}$ are symmetric, that is $h_{m,j} = h_{m,-j}$, and the length of the average is $2m+1$.

For current and recent data $(T-m < t \leq T)$ a symmetric linear filter cannot be applied, and therefore truncated asymmetric filters are used. For example, for the last available observation x_T, the seasonally adjusted value is given by

$$x_T^a = h_0(B)x_T = \sum_{j=0}^{m} h_{0j}x_{T-j} \qquad [1b]$$

The asymmetric filters are time varying in the sense that different filters are applied for the $m+1$ first and last observations. The end estimates revised as new observations are added because of: (1) the new innovations, and (2) the differences between the symmetric and asymmetric filters. The estimates obtained with symmetric filters are often called 'final.'

The development of electronic computers contributed to major improvements in seasonal adjustment based on moving averages and facilitated their massive application.

In 1954, Julius Shiskin of the US Bureau of Census developed software called Method I, based mainly on the works of Macaulay (1931) already being used by the US Federal Reserve Board. Census Method I was followed by Census Method II and 11 more experimental versions (X1, X2, …, X11). The best known and widely applied was the Census X11 variant (Shiskin et al., 1967) but produced poor seasonally adjusted data at the end of the series which is of crucial importance to assess the direction of the short-term trend and the identification of turning points in the economy.

Dagum (1978) developed the X11ARIMA to correct for this serious limitation. The X11ARIMA method consists of: (1) modeling the original series with an ARIMA model of the Box and Jenkins (1970) type, (2) extending the original series 1–3 years with forecasts from the ARIMA model that fits and extrapolate well according to well defined acceptance criteria, and (3) estimating each component with moving averages that are symmetric for middle observations and asymmetric for both end years. The latter are obtained via the convolution of Census X11 variant and the ARIMA model extrapolations. For flow series, deterministic components such as trading-day variations and Easter holiday effects are estimated with dummy variable regression models and removed from the series, so that only the remainder is subject to steps (1) to (3) above. The X11ARIMA was extended again by Dagum (1988) and, later Findley et al. in 1998 developed X12ARIMA that offers a RegARIMA option to estimate deterministic components such as trading day variations and Moving Holiday effects simultaneously with the ARIMA model for extrapolation. It also includes new diagnostic tests and spectral techniques to assess the goodness of the results. The X12ARIMA method is today the one most often applied by statistical agencies in the world. The US Bureau of Census continued research on the development of seasonal adjustment methods and recently produced a beta version called X13-ARIMA-SEATS which enables the estimation of the seasonal component either via linear filters as those available in X12ARIMA or based on an ARIMA decomposition model.

All the moving average methods here discussed are nonlinear; hence, the seasonally adjusted total of aggregated series is not equal to the algebraic sum of the seasonally adjusted series that enter into the aggregation. The main causes of nonlinearity are: (1) a multiplicative decomposition model for the unobserved components, (2) the identification and replacement of extreme values, (3) the ARIMA extrapolations, and (4) automatic selection of moving average length for the estimation of the trend-cycle and seasonality.

The properties of the combined linear filters applied to estimate the various components were originally calculated by Young (1968) for a standard option. Later, Dagum et al. (1990) calculated and analyzed all possible filter combination of Census X11 and X11ARIMA. In the same context, Cleveland and Tiao (1976) and Burridge and Wallis (1984) found ARIMA models that approximated well some of the linear filters used for the trend-cycle and seasonal component of Census X11.

Model-Based Seasonal Adjustment Methods

The best known model-based methods are: (1) the regression methods with global or locally deterministic models for each component, and (2) stochastic model based methods that use ARIMA models.

Regression Methods

Seasonal adjustment by regression methods is based on the assumption that the systematic components of time series can be closely approximated by simple function of times over the *entire* span of the raw series. In general, two types of mathematical functions are considered. One is a polynomial of fairly low degree to represent the trend component; the other, linear combinations of sine and cosine functions with different periodicity and fixed amplitude and phase, to represent business cycles and seasonality. To overcome the limitation of using global deterministic representations for the trend, cycle, and seasonality, regression methods were extended to incorporate stochastic representations by means of local polynomials (spline functions) for successive short segments of the series and introducing changing seasonal amplitudes. A major breakthrough in this direction was made by Akaike (1980) who introduced prior constraints to the degree of smoothness of the various components and solved the problem with the introduction of a Bayesian model. Another important contribution is the regression method with locally deterministic models estimated by the LOESS (locally weighted regression) smoother developed by Cleveland et al. (1990).

The simplest and often studied seasonal adjustment regression method assumes that the generating process of seasonality can be represented by strictly periodic functions of time of annual periodicity. The problem is to estimate s seasonal coefficients (s being the seasonal periodicity, e.g., 12 for monthly series, 4 for quarterly series) subject to the seasonal constraint that they sum to zero.

This regression seasonal model can be written as

$$\begin{aligned} Y_t &= S_t + \varepsilon_t, \quad t = 1, \ldots, T; \\ S_t &= \sum_{j=1}^{s} \gamma_j d_{jt} \quad \text{subject to} \quad \sum_{j=1}^{s} \gamma_j = 0 \end{aligned} \qquad [2]$$

here the d_j's are dummy variables taking a value of unity in season j and a value of zero otherwise; $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma_\varepsilon^2)$ is a white noise process with zero mean, homoscedastic, and nonautocorrelated. The seasonal parameters γ_js can be interpreted

as the expectation of Y_t in each season and represent the seasonal effects of the series under question. Since the parameters are constant, the seasonal pattern repeats exactly over the years, that is,

$$S_t = S_{t-s}, \text{ for all } t > s \quad [3]$$

Equation [2], often included in econometric models, assumes that the effects of the seasonal variations are deterministic because they can be predicted with no error. However, for most economic and social time series, seasonality is not deterministic but changes gradually in a stochastic manner. One representation of stochastic seasonality is to assume in regression eqn [2] that the γ_js are random variables instead of constant. Hence, the relationship, eqn [3], becomes

$$S_t = S_{t-s} + \omega_t, \text{ for all } t > s \quad [4]$$

where $\{\omega\} \sim \mathrm{WN}(0, \sigma_\omega^2)$ and $\mathrm{E}(\omega_t \varepsilon_t) = 0$, for all t. The stochastic seasonal balance constraint is here given by $\sum_{j=0}^{s-1} S_{t-j} = \omega_t$ with expected value equal to zero.

Equation [4] assumes seasonality to be generated by a nonstationary stochastic process. In fact, it is a random walk process and can be made stationary by seasonal differencing $\nabla_s = (1 - B^s)$ where B denotes the backward shift operator. In reduced form, the corresponding regression model is a linear combination of white noise processes of maximum lag s. Since $(1 - B^s) = (1 - B)(1 + B + \cdots + B^{s-1})$, model-based seasonal adjustment methods attribute only $S(B) = \sum_{j=0}^{s-1} B^j$ to the seasonal component leaving $(1 - B)$ as part of a nonstationary trend. Thus, the corresponding seasonal stochastic model can be written as

$$S(B)S_t = \omega_t, \quad [5]$$

Equation [6] represents a seasonality that evolves in an erratic manner and it was included in seasonal adjustment methods based on structural model decomposition (see, e.g., Harvey, 1981; Kitagawa and Gersch, 1984). This seasonal behavior is seldom found in observed time series where seasonality evolves smoothly through time. To represent this latter, Hillmer and Tiao (1982) introduced a major modification

$$S(B)S_t = \eta_s(B)b_t \quad [6]$$

where the left-hand side of eqn [6] follows an invertible moving average process of maximum order $s - 1$, denoted by MA($s - 1$). Equation [6] is discussed extensively in Bell and Hillmer (1984) and can be generalized easily by replacing the MA part with stationary invertible autoregressive moving average (ARMA) processes of the Box and Jenkins (1970) type. Depending on the order of the ARMA process a large variety of evolving stochastic seasonal variations can be modeled.

Other seasonal models used in regression methods where the components are assumed to be local polynomials of time, are based on a smoothness criterion. For example, given $Y_t = S_t + \varepsilon_t$, $t = 1, \ldots, T$, restrictions are imposed on S_t, such that, $\sum_t (S_t - S_{t-s})^2$ and $\sum_t [S(B)S_t]^2$ be small. The solution is given by minimizing a weighted linear combination of both criteria as follows

$$\sum_{t=1}^{T} (Y_t - S_t)^2 + \alpha \sum_{t=1}^{T} (S_t - S_{t-s})^2 + \beta \sum_{t=1}^{T} [S(B)S_t]^2 \quad [7]$$

where the ill-posed problem of choosing the α and β parameters was solved by Akaike (1980) using a Bayesian model.

Mixed seasonal models with deterministic and stochastic effects are given in Pierce (1978).

Stochastic Model Based Methods

Stochastic model based methods were mainly developed during the 1980 following two different approaches. One, originally known as seasonal adjustment by signal extraction (SIGEX) was developed by Burman (1980) and as ARIMA model based seasonal adjustment (Hillmer and Tiao, 1982) largely discussed in Bell and Hillmer (1984), the other is referred to as structural model decomposition method (see e.g., Harvey, 1981; Kitagawa and Gersch, 1984). The main difference between these two approaches is that in the latter simple ARIMA models are directly specified for each unobserved component whereas in the former an overall ARIMA model is obtained from observable data, and by imposing certain restrictions, models for each component are derived. Since the components are unknown, to obtain a unique decomposition Hillmer and Tiao (1982) proposed a canonical decomposition which has the properties, among others, of maximizing the variance of the irregulars and minimizing the variance of the stochastic seasonal models.

ARIMA models identification and estimation are sensitive to outliers or extreme values and cannot deal with deterministic components such as trading days and moving holidays. Therefore further developments were made by combining dummy variables regression models with ARIMA models to deal with these cases. In this regard, Gomez and Maravall (1996) developed at the Bank of Spain a seasonal adjustment software called TRAMO-SEATS which is currently applied mainly by European statistical agencies. TRAMO stands for "Time Series Regression with ARIMA noise, Missing observations and Outliers" and SEATS for "Signal Extraction in ARIMA Time Series." Firstly, TRAMO first estimates via regression dummy variables and direct ARIMA modeling, the deterministic components, which are removed from the input data. In a second round, SEATS estimates the seasonality and trend-cycle components from an ARIMA model fitted to the data where the deterministic components are removed. SEATS uses the filters derived from the linearized ARIMA model that describes the behavior of the time series. It should be mentioned that Eurostat, with the collaboration of the Bank of Belgium, the US Bureau of the Census, the Bank of Spain, and the European Central Bank, is developing an interface of TRAMO-SEATS and X12-ARIMA called DEMETRA+, entirely reprogrammed in "dot.com" following the "TRAMO-SEATS components" application developed at the National Bank of Belgium (DEMETRA+ will replace the old DEMETRA). At the Bank of Spain and Eurostat web site it is also possible to find a considerable number of papers relevant to TRAMO-SEATS as well as in the European Statistical System (ESS) Guidelines.

On the other hand the structural model decomposition method starts directly with an observation equation (sometimes called measurement equation) consisting of the unobserved components. Simple ARIMA or stochastic trigonometric models are assumed for each component. The structural model

is cast in a state-space form and estimated with Kalman filtering and smoothing. Koopman et al. (1998) developed STAMP that stands for Structural Time Series Analyser, Modeller and Predictor and includes several types of models for each component. STAMP is set up in an easy-to-use form which enables the user to concentrate on model selection and interpretation. STAMP 8 is an integrated part of the OxMetrics modular software system for time series data analysis, forecasting, financial econometric modeling, and the statistical analysis of cross-section and panel data. STAMP is not used officially by statistical agencies but mainly by econometricians, academics, and researchers involved with time series analysis, forecasting, financial econometric modeling, and statistical analysis of cross-section and panel data.

Conclusion

Seasonal adjustment plays a significant role in standardizing time series for the analysis of current social and economic conditions, particularly in determining the stages of the business cycle at which the economy stands. Such knowledge provides the bases for decision making to control the level of economic activity and it is of major importance around turning points.

The fact that seasonality and the other time series components, trend and cycle, change through time leads to major modifications of the traditional methods based on moving averages and to the development of new ones mainly based on ARIMA models.

After the 1990s, major financial, economic, and ecological changes of a global character have introduced more variability in time series and, consequently, in their seasonally adjusted values, making it very difficult to determine the direction of the short-term trend for an earlier detection of a turning point. As a consequence, producers and users of seasonally adjusted data have centered their attention on the need for further smoothing. In this regard, new research has been oriented toward a better estimation of the trend-cycle for current economic analysis, including parametric and non-parametric approaches as well as linear and nonlinear smoothers. For information on trend-cycle estimation the reader may refer to Kenny and Durbin (1982), Maravall (1993), Dagum (1996), Dagum and Luati (2009, 2012), Dagum and Bianconcini (2008, 2013), and references given therein. The presence of many outlying observations have also stimulated research on nonlinear and non-Gaussian models for seasonal adjustment. More information on seasonal adjustment theory and practice can be found in Hylleberg (1992), Zellner (1978, 1983), and Bell et al. (2012).

See also: Government Statistics; Nonparametric Regression; Time Series: Advanced Methods; Time Series: ARIMA Methods; Time Series: Cycles; Time Series: Economic Forecasting; Time Series: General; Time Series: State Space Methods; Time Series: Unit Roots and Nonstandard Limiting Distributions; Times Series: Cointegration.

Bibliography

Akaike, H., 1980. Seasonal adjustment by a Bayesian modelling. Journal of Time Series Analysis 1, 1–13.

Bell, W.R., Hillmer, S.C., 1984. Issues involved with the seasonal adjustment of economic time series. Journal of Business and Economic Statistics 2 (3), 291–320.

Bell, W.R., Holan, S.H., McElroy, T.S., 2012. Economic Time Series: Modeling and Seasonality. Taylor and Francis Group, Boca Raton, USA.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Burman, J.P., 1980. Seasonal adjustment by signal extraction. Journal of the Royal Statistical Society Series A 143, 321–337.

Burridge, P., Wallis, K.F., 1984. Unobserved-components models for seasonal adjustment filters. Journal of Business and Economic Statistics 2 (4), 350–359.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: a seasonal trend decomposition procedure based on LOESS. Journal of Official Statistics 6 (1), 3–33.

Cleveland, W.P., Tiao, G.C., 1976. Decomposition of seasonal time series: a model for the census X-11 program. Journal of the American Statistical Association 71, 581–587.

Dagum, E.B., 1978. Modelling, forecasting and seasonally adjusting economic time series with the X11ARIMA method. The Statistician 27 (3), 203–216.

Dagum, E.B., 1988. The X11ARIMA/88 Seasonal Adjustment Method. Foundations and User's Manual. Statistics Canada, Ottawa.

Dagum, E.B., 1996. A new method to reduce unwanted ripples and revisions in trend-cycle estimates from X11ARIMA. Survey Methodology 22 (1), 77–83.

Dagum, E.B., Bianconcini, S., 2008. The Henderson smoother in reproducing kernel Hilbert space. Journal of Business and Economic Statistics 26 (4), 536–545.

Dagum, E.B., Bianconcini, S., 2013. A unified probabilistic view of nonparametric predictors via reproducing kernel Hilbert spaces. Econometric Reviews 32 (7), 848–867.

Dagum, E.B., Chhab, N., Chiu, K., 1990. Derivation and properties of the X11ARIMA and census X11 linear filters. Journal of Official Statistics 12 (4), 329–347.

Dagum, E.B., Luati, A., 2009. A cascade linear filter to reduce revisions and false turning points in real time trend-cycle estimation. Econometric Reviews 28 (1–3), 40–59.

Dagum, E.B., Luati, A., 2012. Asymmetric filters for trend-cycle estimation. In: Bell, W.R., Holan, S.H., McElroy, T.S. (Eds.), Economic Time Series: Modeling and Seasonality. Taylor and Francis Group, Boca Raton, USA, pp. 213–230.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B., 1998. New capabilities and methods of the X12ARIMA seasonal adjustment program. Journal of Business and Economic Statistics 16 (2), 127–152.

Gomez, V., Maravall, A., 1996. Program TRAMO and SEATS. Bank of Spain, Madrid.

Harvey, A.C., 1981. Time Series Models. Allan, Oxford, UK.

Hillmer, S.C., Tiao, G.C., 1982. An ARIMA-model-based approach to seasonal adjustment. Journal of the American Statistical Association 77, 63–70.

Hylleberg, S. (Ed.), 1992. Modelling Seasonality. Oxford University Press, New York.

Kenny, P.B., Durbin, J., 1982. Local trend estimation and seasonal adjustment of economic and social time series. Journal of the Royal Statistical Society Series A 145, 1–41.

Kitagawa, G., Gersch, W., 1984. A smoothness priors – state space modeling of time series with trend and seasonality. Journal of the American Statistical Association 79 (386), 378–389.

Koopman, S.J., Harvey, A.C., Doornik, J.A., Shephard, N., 1998. Structural Time Series Analysis STAMP (5.0). Thomson Business Press, London.

Macaulay, F.R., 1931. The Smoothing of Time Series. National Bureau of Economic Research, New York.

Maravall, A., 1993. Stochastic linear trends: models and estimators. Journal of Econometrics 56, 5–37.

Shiskin, J., Young, A.H., Musgrave, J.C., 1967. The X-11 Variant of the Census Method II Seasonal Adjustment Program. Technical Paper 15 (revised). US Department of Commerce, Bureau of the Census, Washington, DC.

Young, A.H., 1968. A linear approximation to the census and BLS seasonal adjustment methods. Journal of the American Statistical Association 63, 445–457.

Zellner, A. (Ed.), 1978. Seasonal Analysis of Economic Time Series. US Bureau of the Census, Washington, DC.

Zellner, A. (Ed.), 1983. Applied Time Series Analysis of Economic Data. US Bureau of the Census, Washington, DC.

Time Series: State Space Methods

Siem J Koopman and Jacques JF Commandeur, Vrije Universiteit, Amsterdam, Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

State space modeling provides a unified methodology for treating a wide range of problems in time series analysis. The Kalman filter and its related methods have become key tools in the analysis of time series in economics, finance, and in many other fields as well. In an increasingly more complex world, static and dynamic models have proven to be too limited in empirical and relevant policy studies. The modeling of time-varying features in a time series has been given much attention recently. In this article we review and provide some adequate details and guidance for the adaptation of state space methods in univariate and multivariate time series analysis. We provide more detailed discussions for linear Gaussian model formulations and more concise reviews for nonlinear and non-Gaussian departures.

Introduction

State space methods originated in the field of control engineering, starting with the groundbreaking paper of Kalman (1960). They were initially (and still are) deployed for the purpose of accurately tracking the position and velocity of moving objects such as ships, airplanes, missiles, and rockets. Around the 1980s it was recognized in other fields than control engineering that these ideas could be applied to time series analysis generally as well. Since then state space methods have been applied in a wide range of subjects, including economics, finance, political science, environmental science, road safety, and medicine. The monographs of Harvey (1989) and West and Harrison (1997) have been instrumental in the dissemination of state space methods outside the field of control engineering.

Nowadays state space methods are widely used for a range of purposes and applications in the statistical analysis of time series. It has been recognized that many different models can be represented in state space form and be analyzed using the Kalman filter and related methods. The applications range from the fitting of autoregressive integrated moving average (ARIMA) models of Box and Jenkins (1976) when missing observations are present, to the forecasting of time series using high-dimensional time-varying parameter models as formulated by Watson and Engle (1983), and to the decomposition of time series into trend, seasonal, and irregular components using the structural time series model of Harvey (1989).

Linear Gaussian State Space Models

A time series is a set of observations which are sequentially ordered over time. In a state space analysis, the time series observations are assumed to depend linearly on a *state vector* that is unobserved and is generated by a stochastically time-varying process (a dynamic system). The observations are further assumed to be subject to measurement error that is independent of the state vector. The state vector can be estimated or identified once a sufficient set of observations become available. In this section we concentrate on the state space model and its special cases. In Section State Space Analysis we will discuss methods for estimation, residual analysis, and forecasting on the basis of state space models. The expositions below rely on the treatments provided by the introductory textbook of Commandeur and Koopman (2007), the introductory paper of Commandeur et al. (2011), and the more advanced textbooks of Harvey (1989) and Durbin and Koopman (2012).

The general linear Gaussian state space model for the n-dimensional observation sequence $y_1, \ldots, y_n$ can be written as

$$y_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{NID}(0, H_t) \qquad [1]$$

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim \mathrm{NID}(0, Q_t), \quad t = 1, \ldots, n \qquad [2]$$

where α_t is the state vector; ε_t and η_t are disturbance vectors; and the system matrices Z_t, T_t, R_t, H_t, and Q_t are fixed and known but a selection of elements may depend on an unknown parameter vector. Equation [1] is called the *observation* or *measurement equation*, while eqn [2] is called the *state* or *transition equation*. The $p \times 1$ vector, y_t, represents p dependent variables for which we assume to have observations available for all time periods $t = 1, \ldots, n$. All methods reviewed below can treat the case where ranges of observations are missing for the p variables. The $m \times 1$ state vector, α_t, represents unobserved dynamic processes but may also contain fixed effects such as unknown constants and regression coefficients. The $p \times 1$ irregular vector, ε_t, has zero mean and $p \times p$ variance matrix H_t.

The $p \times m$ matrix, Z_t, links the observation vector, y_t, with the state vector, α_t, and may consist of regression variables. It is therefore necessary to assume that Z_t is time-varying but as a deterministic or exogenous process. The $m \times m$ transition matrix, T_t, in eqn [2] determines the dynamic evolution of the unobserved processes in the state vector. The $r \times 1$ disturbance vector, η_t, for the state vector update has zero mean and $r \times r$ variance matrix, Q_t. The observation and state disturbances, ε_t and η_t, are assumed to be serially independent and independent of each other at all time points. In many standard cases, $r = m$ and matrix R_t is the identity matrix I_m. In other cases, matrix R_t is an $m \times r$ selection matrix with $r < m$. Although matrix R_t can be specified freely, it is often composed of a selection from the first r columns of the identity matrix I_m.

The initial state vector α_1 is assumed to be generated as

$$\alpha_1 \sim \mathrm{NID}(a_1, P_1)$$

independently of the observation and state disturbances ε_t and η_t. Mean a_1 and variance P_1 can be treated as given and known in almost all stationary processes for the state vector. For nonstationary processes and regression effects in the state vector, the associated elements in the initial mean a_1 can be treated as unknown and need to be estimated. For an extensive discussion of *initialization* in state space analysis, we refer to Durbin and Koopman (2012: Chapter 5).

Local Level Model and Other Unobserved Component Models

By appropriate choices of the vectors α_t, ε_t, and η_t, and of the matrices Z_t, T_t, H_t, R_t, and Q_t, a wide range of different time series models can be derived from eqns [1] and [2]. Here we discuss the class of *unobserved components time series models*. A number of special cases will be discussed in some detail. Special attention is given to the univariate *local level model*.

By having

$$\alpha_t = \mu_t, \quad \eta_t = \xi_t, \quad Z_t = T_t = R_t = 1, \quad H_t = \sigma_\varepsilon^2, \quad Q_t = \sigma_\xi^2$$

(all of order 1×1) for $t = 1, \ldots, n$, models [1] and [2] reduce to the local level model as given by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{NID}\left(0, \sigma_\varepsilon^2\right) \\ \mu_{t+1} &= \mu_t + \xi_t, \quad \xi_t \sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \end{aligned} \qquad [3]$$

for $t = 1, \ldots, n$. The level component, μ_t, can be regarded as the time-varying equivalent of the intercept in the classical linear regression model $y_t = \mu + \varepsilon_t$ which is obtained by setting all the level disturbances ξ_t in eqn [3] equal to zero and with $\mu = \mu_1$. The key difference is that the intercept μ in a regression model is fixed whereas the level component, μ_t, in eqn [3] is allowed to change from time point to time point.

Since the second equation in eqn [3] defines a random walk, the local level model is also referred to as the random walk plus noise model (where the noise refers to the irregular component ε_t). It can be shown that the dynamic process for $x_t = y_{t+1} - y_t = \eta_t + \varepsilon_{t+1} - \varepsilon_t$, when $t = 1, \ldots, n$, reduces to the moving average process $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$ where θ relates to the *signal-to-noise ratio* $q = \sigma_\xi^2/\sigma_\varepsilon^2$ via a quadratic function. A key notion is that the moving average coefficient, θ, in the x_t process is restricted to $-0.5 < \theta < 0$ for $q > 0$. Furthermore, the forecasting function of the local level model eqn [3] is equivalent to the *exponentially weighted moving average* (or *exponential smoothing*) scheme where the smoothing coefficient is directly related to q as well.

By having

$$\alpha_t = \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix}, \quad \eta_t = \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix}, \quad T_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad Z_t = (1 \quad 0),$$

$$H_t = \sigma_\varepsilon^2, \quad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}, \text{ and } R_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and by representing the resulting eqns [1] and [2] in scalar form, we obtain the local linear trend model as given by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{NID}\left(0, \sigma_\varepsilon^2\right) \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{NID}\left(0, \sigma_\zeta^2\right) \end{aligned} \qquad [4]$$

for $t = 1, \ldots, n$.

The local linear trend model requires a 2×1 state vector α_t: one element for the level component, μ_t, and one element for the slope component, ν_t. The slope or gradient component represents the linear growth of the trend component. It can be conceived of as the time-varying equivalent of the regression coefficient in the classical regression model where the observed time series y_t is regressed on the 'exogenous variable' time index t: $y_t = \mu + \nu t + \varepsilon_t$ with $\mu = \mu_1$ and $\nu = \nu_1$. The key difference of the models is that the regression coefficient or weight ν is fixed in the classical linear regression model, whereas the slope ν_t in the local linear trend model is allowed to stochastically change over time.

In situations where the observed time series consists of quarterly or monthly data, for example, the local level and the local linear trend model can be extended with a stochastically evolving seasonal dummy component denoted here by γ_t. In the case of a quarterly time series (the seasonal length s equals 4), by having

$$\alpha_t = \begin{pmatrix} \mu_t \\ \gamma_{1,t} \\ \gamma_{2,t} \\ \gamma_{3,t} \end{pmatrix}, \quad \eta_t = \begin{pmatrix} \xi_t \\ \omega_t \end{pmatrix},$$

$$T_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad Z_t = (1 \quad 1 \quad 0 \quad 0),$$

$$H_t = \sigma_\varepsilon^2, \quad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 & 0 \\ 0 & \sigma_\omega^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \text{ and } R_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and by formulating eqns [1] and [2] in scalar notation, we obtain

$$\begin{aligned} y_t &= \mu_t + \gamma_{1,t} + \varepsilon_t, & \varepsilon_t &\sim \mathrm{NID}\left(0, \sigma_\varepsilon^2\right) \\ \mu_{t+1} &= \mu_t + \xi_t, & \xi_t &\sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \\ \gamma_{1,t+1} &= -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} + \omega_t, & \omega_t &\sim \mathrm{NID}\left(0, \sigma_\omega^2\right) \end{aligned} \qquad [5]$$

with $\gamma_{2,t+1} = \gamma_{1,t}$ and $\gamma_{3,t+1} = \gamma_{2,t}$, for $t = 1, \ldots, n$. It is the formulation of a level plus seasonal model for a quarterly time series with stochastically time-varying level and seasonal components. The seasonal dummy model is not the only specification for incorporating a time-varying seasonal effect in an unobserved components time series model; see Proietti (2000) for a review of different seasonal models and their properties. When the slope component, ν_t, is included in the level component, μ_t, in eqn [5] as well, we obtain the *basic structural time series model* of Harvey (1989). A typical application of this model is for the *seasonal adjustment* of time series. A seasonally adjusted time series is defined in this context simply by the estimate of $y_t - \gamma_t = \mu_t + \varepsilon_t$ for $t = 1, \ldots, n$.

Another and final extension is to include one or more cycles to any of the unobserved components time series models as formulated above. For example, by having

$$\alpha_t = \begin{pmatrix} \mu_t \\ c_t \\ c_t^* \end{pmatrix}, \quad \eta_t = \begin{pmatrix} \xi_t \\ \kappa_t \\ \kappa_t^* \end{pmatrix},$$

$$T_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho\cos(\lambda_c) & \rho\sin(\lambda_c) \\ 0 & -\rho\sin(\lambda_c) & \rho\cos(\lambda_c) \end{bmatrix}, \quad Z_t = (1 \quad 1 \quad 0),$$

$$H_t = \sigma_\varepsilon^2, \quad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_c^2(1-\rho^2) & 0 \\ 0 & 0 & \sigma_c^2(1-\rho^2) \end{bmatrix},$$

$$\text{and } R_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

in eqns [1] and [2], we obtain the level plus cycle model as given by

$$\begin{aligned} y_t &= \mu_t + c_t + \varepsilon_t, \\ \mu_{t+1} &= \mu_t + \xi_t, \\ c_{t+1} &= \rho\left[\cos(\lambda_c)c_t + \sin(\lambda_c)c_t^*\right] + \kappa_t, \\ c_{t+1}^* &= \rho\left[-\sin(\lambda_c)c_t + \cos(\lambda_c)c_t^*\right] + \kappa_t^*, \\ \xi_t &\sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \\ \xi_t &\sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \\ \kappa_t &\sim \mathrm{NID}\left[0, \sigma_c^2(1-\rho^2)\right] \\ \kappa_t^* &\sim \mathrm{NID}\left[0, \sigma_c^2(1-\rho^2)\right] \end{aligned} \qquad [6]$$

for $t = 1, \ldots, n$, where $0 < \rho \leq 1$ is the *damping factor* and λ_c is the frequency of the cycle measured in radians so that $2\pi/\lambda_c$ is the overall *period* of the cycle. In case $\rho = 1$, the cycle reduces to a fixed sine–cosine wave but the component is still stochastic since the initial values c_1 and c_1^* are stochastic variables with mean zero and variance σ_c^2. A typical application of this model is for the signal extraction of *business cycles* from economic time series.

Regression and Intervention Effects

Time series variables can be subject to specific effects and events that occur within a time period. We can capture these by including fixed explanatory and intervention variables in the model. The explanatory variable typically varies for each time period, while the intervention variable is designed to capture a specific effect in the dependent variable such as an outlier due to an extreme event or a structural break due to a policy or regime change. The intervention variable represents an outlier when all values are zero except at one time point where it equals unity. It represents a structural break when all values are zero up to a time point and all values are unity after this time point. Other types of intervention effects can also be specified; see Box and Tiao (1975).

Suppose a single regression variable, x_t, and a specific intervention variable, w_t, need to be included in the local level model, we have $y_t = \mu_t + \beta x_t + \lambda w_t + \varepsilon_t$, for $t = 1, \ldots, n$, as our model of interest, where β and λ are fixed unknown regression coefficients. The state space model representation of this model requires a state vector of three elements: one for the level component μ_t, one for the regression coefficient β, and one for the intervention coefficient λ. The substitution of

$$\alpha_t = \begin{pmatrix} \mu_t \\ \beta_t \\ \lambda_t \end{pmatrix}, \quad \eta_t = \xi_t, \quad T_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Z_t = (1 \quad x_t \quad w_t),$$

$$H_t = \sigma_\varepsilon^2, \quad Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \text{ and } R_t = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

in eqns [1] and [2] leads to

$$\begin{aligned} y_t &= \mu_t + \beta_t x_t + \lambda_t w_t + \varepsilon_t, & \varepsilon_t &\sim \mathrm{NID}\left(0, \sigma_\varepsilon^2\right) \\ \mu_{t+1} &= \mu_t + \xi_t, & \xi_t &\sim \mathrm{NID}\left(0, \sigma_\xi^2\right) \end{aligned}$$

for $t = 1, \ldots, n$, with $\beta_{t+1} = \beta_t = \beta_1 = \beta$ and $\lambda_{t+1} = \lambda_t = \lambda_1 = \lambda$. This is the local level model with an explanatory variable x_t and an intervention variable w_t. By adding disturbance terms to the state vector element with β_t, the regression coefficient effectively becomes a random walk process. Hence, this specification allows easily for the estimation of time-varying regression effects; see Harrison and Stevens (1976) for an early application.

Structural Time Series Analysis

From the previous sections, a structural approach to time series analysis emerges which is facilitated by the state space framework. In this approach, different unobserved components or building blocks responsible for the dynamics of the series such as trend, seasonal, cycle, and the effects of explanatory and intervention variables are identified separately before being put together in a state space model. It is the responsibility of the researcher to decide what components are required in a specific situation, and then to consider whether they apply to the time series under consideration. This explains why state space models are often referred to as *structural time series models*.

ARIMA and ARIMA Components Models

A more traditional approach to time series analysis is proposed in Box and Jenkins (1976); see also the discussions in Tiao (2014). We show how the *ARIMA* model can be put into state space form. An essential feature of the Box–Jenkins approach to time series analysis is that the observed time series is assumed to be *stationary*, meaning that its means and covariances should be invariant when the series is shifted or translated through time. If an observed time series is stationary, then its analysis can be based on an *autoregressive moving average* (ARMA) model. However, if nonstationary features are present, the Box–Jenkins analysis suggests to enforce stationarity by *differencing* the time series. The key features of the series such as trend and seasonal will be removed from the series y_t by transforming it into

$$\Delta y_t = y_t - y_{t-1}$$

to remove the trend from the series, and

$$\Delta_s y_t = y_t - y_{t-s}$$

to remove the seasonal effect with periodicity s from the series.

To illustrate how an ARMA model can be formulated in state space form, we consider the ARMA(2, 1) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \zeta_t + \theta_1 \zeta_{t-1}, \quad \zeta_t \sim \mathrm{NID}\left(0, \sigma_\zeta^2\right)$$

for $t = 1, \ldots, n$, where we assume that the observed series already satisfies stationarity; see Tiao (2014) for further details. By considering the state space models [1] and [2], and by defining

$$\alpha_t = \begin{pmatrix} y_{t+1} \\ \phi_2 y_t + \theta_1 \zeta_{t+1} \end{pmatrix}, \quad \eta_t = \zeta_{t+1}, \quad T_t = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix},$$

$$Z_t = (1 \quad 0), \quad H_t = 0, \quad Q_t = \sigma_\zeta^2, \text{ and } R_t = \begin{bmatrix} 1 \\ \theta_1 \end{bmatrix}$$

for $t = 1, \ldots, n$, we have represented the ARMA(2, 1) model in the state space framework. For a general overview of how any of the ARIMA(p, d, q) models can be represented in the state space framework, we refer to Durbin and Koopman (2012: Section 3.4).

A more general approach to time series analysis can be based on the ARIMA components model as given by

$$y_t = \sum_{j=1}^{m} \mu_t^{(j)}, \quad \mu_t^{(j)} \sim \mathrm{ARIMA}, \quad j = 1, \ldots, m \qquad [7]$$

for $t = 1, \ldots, n$, where the $\mu_t^{(j)}$'s follow mutually independent ARIMA processes with different levels of differencing and different values for p and q for $j = 1, \ldots, m$. The dynamic properties of y_t can be derived from the different ARIMA processes, while the time series y_t can be expressed as an ARIMA process itself. Unobserved components time series models can be regarded as special cases of the ARIMA components model. For example, the local level model [3] can be expressed as an ARIMA components model with $m = 2$, $\mu_t^{(1)} = \mu_t$, and $\mu_t^{(2)} = \varepsilon_t$. We take $\mu_t^{(1)}$ as an ARMA(0, 0) process after first differencing (with $p = 0$ and $q = 0$), and we take $\mu_t^{(2)}$ as an ARMA(0, 0) process without differencing. A more general and more detailed discussion of ARIMA components models can be found in Bell (2004).

Multivariate Models

The structural approach to time series analysis is easily extended to multivariate time series. For example, when we denote y_t as a $p \times 1$ vector of observations, a multivariate local level model can be adopted for modeling the p time series simultaneously:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{NID}(0, \Sigma_\varepsilon) \\ \mu_{t+1} &= \mu_t + \xi_t, \quad \xi_t \sim \mathrm{NID}(0, \Sigma_\xi) \end{aligned} \qquad [8]$$

for $t = 1, \ldots, n$, where μ_t, ε_t, and ξ_t are $p \times 1$ vectors and Σ_ε and Σ_ξ are $p \times p$ variance matrices. In what is known as the *seemingly unrelated time series equations model* [8], the series are modeled as in the univariate situation, but the disturbances driving the level components are allowed to be instantaneously correlated across the p series. When slope, seasonal, or cycle components are also included in the model, each of these three components has an associated $p \times p$ variance matrix for their disturbance vectors, allowing for correlation between the series.

If the rank r of Σ_ξ in eqn [8] is assumed smaller than p, then the model implies that the p series have r *common levels*. Such common factors may not only have a nice and interesting interpretation, but may also result in more efficient inferences and forecasts since the number of parameters reduces.

The Latent Risk Model

Risk is at the center of many policy decisions in many kinds of organizations. In risk analysis, measures of exposure to risk, outcomes (or events), and, possibly, losses are analyzed. Risk itself is not observed. For example, in case of road safety research, exposure is the number of cars (or the number of kilometers traveled), outcome is the number of accidents, and loss is the cost of damages (or the number of fatalities). From these measures we can learn about risk and its associated severity. In a time series context, we typically observe the measures in totals for a country (or a region, or a specific group) and for a specific period (month, quarter, or other). The time series for exposure y_{1t}, outcome y_{2t}, and loss y_{3t} are typically observed with error and are subject to, possibly, trend, seasonal, cycle, and regression effects. We can carry out a time series analysis with directly interpretable variables via the multiplicative latent risk model

$$y_{1t} = \mu_{1t} \times \xi_{1t}, \quad y_{2t} = \mu_{1t} \times \mu_{2t} \times \xi_{2t},$$

$$y_{3t} = \mu_{1t} \times \mu_{2t} \times \mu_{3t} \times \xi_{3t}$$

where μ_{1t} is exposure corrected for observation error, μ_{2t} is risk, and μ_{3t} is severity, while ξ_{jt} are the multiplicative errors with their means equal to 1, for $j = 1, 2, 3$. The model implies that expected outcome is the product of exposure and risk, while expected loss is the product of outcome and severity. By taking logs we obtain the trivariate latent risk model in additive form

$$\begin{pmatrix} \log y_{1t} \\ \log y_{2t} \\ \log y_{3t} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \theta_t + \varepsilon_t$$

with the signal vector $\theta_t = (\log \mu_{1t}, \log \mu_{2t}, \log \mu_{3t})'$ and with the disturbance vector $\varepsilon_t = (\log \xi_{1t}, \log \xi_{2t}, \log \xi_{3t})'$. The signals for exposure, risk, and severity (all three variables in logs) can be modeled simultaneously as linear functions of the state vector. Letting

$$\theta_t = S_t \alpha_t, \quad t = 1, \ldots, n$$

denote the signal vector where S_t is a $3 \times m$ system matrix that relates the signal θ_t with the state vector α_t, the model is placed in state space form [1] and [2] with

$$Z_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} S_t$$

Multiple measures for exposure, outcome, and/or loss can be incorporated in this framework in a straightforward manner. Typical applications for the latent risk model are for studies on insurance claims, credit card purchases, and road safety. Bijleveld et al. (2008) provide further discussions on the latent risk model and show how the general methodology can be effectively used in the assessment of risk.

Dynamic Factor Models

Principal components and factor analysis are widely used statistical methods in the applied social and behavioral sciences. These methods aim to identify commonalities in the covariance structure of high-dimensional data sets. A factor analysis model can be based on the form given by Lawley and Maxwell (1971), that is

$$y_i = \Lambda f_i + u_i, \quad f_i \sim N(0, \Sigma_f), \quad u_i \sim N(0, \Sigma_u)$$

where y_i represents a zero-mean data vector containing measured characteristics of subject i, for $i = 1, \ldots, n$, while Λ is the coefficient matrix for the low-dimensional vector f_j of latent variables. The latent vector f_i with its variance matrix Σ_f and the disturbance vector u_j with its variance matrix Σ_u are assumed to be mutually and serially independent for $i, j = 1, \ldots, n$. The factor analysis model implies the decomposition of the variance matrix of y_i into

$$\Sigma_y = \Lambda \Sigma_f \Lambda' + \Sigma_u$$

Estimation of Λ, Σ_f, and Σ_u can be carried out by maximum likelihood procedures; the loading coefficients in Λ and the variance matrices are subject to a set of linear restrictions necessary for identification. A detailed discussion of the maximum likelihood approach to factor analysis is given by Lawley and Maxwell (1971). In the application of factor analysis in a time series context, the measurements in y_i correspond to a time period t rather than a subject i, we therefore have y_t instead of y_i. The time dependence of the measurements can be accounted for by replacing the serially independence assumption for f_t by a serial dependence assumption. For example, we can assume that f_t is modeled by a vector autoregressive process. We can also let f_t depend on the state vector α_t in a linear way, that is, $f_t = U_t \alpha_t$ where U_t is typically a known selection matrix.

In applications to economics, y_t may consist of a large set of macroeconomic indicators associated with variables such as income, consumption, investment, and unemployment. These variables are all subject to economic activity that can often be related to the business cycle. The dynamic features of the business cycle may be disentangled into a set of factors with possibly different dynamic characteristics. In the context of finance, y_t may consist of a large set of daily prices or returns from individual stocks that make up the indices such as Standard & Poor's 500 and Dow Jones. A set of underlying factors may represent returns from particular portfolio strategies. In the context of marketing, y_t may contain market shares of groups or subgroups of product brands. The factors may indicate whether marketing strategies in certain periods affect all market shares or only a selection of market shares. In all these cases it is not realistic to assume that the factors are independent over time so we require to formulate a dynamic process for the factors. The wide variety of applications of factor analysis in a time series context have led to many contributions in the statistics and econometrics literature on the inference of dynamic factor models; see, for example, Geweke (1977), Engle and Watson (1981), Watson and Engle (1983), Litterman and Scheinkman (1991), Quah and Sargent (1993), Stock and Watson (2002), Doz et al. (2006), and Diebold et al. (2006).

The dynamic factor model can be regarded as a state space model in which the state vector consists of latent factors where dynamic properties are formulated in the state equation of the state space models [1] and [2]. The size of the observation vector is typically large, while the dimension of the state vector is small so we have $p \gg m$. In the example

$$y_t = \Lambda f_t + \varepsilon_t, \quad f_t = U_t \alpha_t, \quad \varepsilon_t \sim N(0, H_t)$$

the dynamic factor model is a special case of the state space models [1] and [2] for which the observation equation has $Z_t = \Lambda U_t$.

State Space Analysis

For given values of all system matrices – and for given initial conditions a_1 and P_1 – the state vector can be estimated in three different ways, yielding what are known as the *filtered*, the *predicted*, and the *smoothed* estimates of the state vector. Depending on the types of state estimates required in the analysis, the estimates of the state vector can be obtained by performing one or two passes through the observed time series:

1. a *forward* pass, from $t = 1, \ldots, n$, using a recursive algorithm known as the *Kalman filter* enables the computation of predicted states, based on $y_1, \ldots, y_{t-1}$; filtered states, based on $y_1, \ldots, y_t$; and observation prediction errors;
2. a *backward* pass, from $t = n, \ldots, 1$, using output of the Kalman filter and using recursive algorithms known as *state and disturbance smoothers* enable the computation of smoothed estimates of states and disturbances that are based on all observations $y_1, \ldots, y_n$.

Kalman Filter for Prediction, Filtering, and Forecasting

The forward pass through the data with the well-known Kalman (1960) filter provides all estimates that are relevant for the filtered and the predicted state. The main purpose of the Kalman filter is to obtain optimal estimates of the state at time point t, *only* considering the observations $\{y_1, y_2, \ldots, y_{t-1}\}$. A key property of the predicted state and its related estimates is therefore that they are only based on *past values* of the observed time series. The recursive formulas for the Kalman filter are

$$\begin{aligned} v_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + H_t \\ K_t &= T_t P_t Z_t' F_t^{-1}, & L_t &= T_t - K_t Z_t \\ a_{t+1} &= T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t L_t' + R_t Q_t R_t' \end{aligned} \qquad [9]$$

for $t = 1, \ldots, n$. The values of a_t in eqn [9] represent the predicted state, while the values of P_t quantify the estimation error variance matrix of the predicted state a_t. Under the assumption of normality, the latter variances are useful for the construction of *confidence intervals* for the predicted state, which – assuming that we are interested in their 90% confidence limits, for example – can be calculated as

$$a_t \pm 1.64\sqrt{P_t}$$

for $t = 1, \ldots, n$. A modification of the Kalman filter also allows the computation of the filtered estimate of the state vectors, that is

$$a_{t|t} = a_t + P_t Z_t' F_t^{-1} v_t, \quad P_{t|t} = P_t - P_t Z_t' F_t^{-1} Z_t P_t,$$
$$t = 1, \ldots, n$$

where $a_{t|t}$ is the optimal estimate of the state at time point t by considering the observations $\{y_1, y_2, \ldots, y_t\}$, while $P_{t|t}$ is the state-filtered estimation error variance matrix. The values of v_t in eqn [9] are called the *one-step ahead prediction* or *forecast* errors, since they quantify the lack of accuracy of a_t in predicting the observed value of y_t at time point t; the values of F_t are the variances of these one-step ahead prediction errors v_t.

One of the convenient features of state space methods is the ease with which they deal with two important aspects of time series analysis – forecasting and missing observations: by treating them in exactly the same way. Missing observations are handled by setting K_t and v_t in eqn [9] equal to zero. Forecasts for $y_{n+1}, \ldots, y_{n+k}$ given $y_1, \ldots, y_n$ are simply obtained by applying the Kalman filter for $t = 1, \ldots, n, n+1, \ldots, n+k$ and by treating $y_{n+1}, \ldots, y_{n+k}$ as missing observations.

State and Disturbance Smoothing

The backward pass through the data is only required for smoothing that leads to estimates such as the smoothed states and smoothed disturbances. The main purpose of state and disturbance smoothing is to obtain estimated values of the state and disturbance vectors at time point t, considering *all* available observations $\{y_1, y_2, \ldots, y_n\}$. The recursive formulas for state smoothing are

$$r_{t-1} = Z_t' F_t^{-1} v_t + Z_t' r_t, \quad N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t \qquad [10]$$

$$\widehat{\alpha}_t = a_t + P_t r_{t-1}, \quad V_t = P_t - P_t N_{t-1} P_t \qquad [11]$$

for $t = n, \ldots, 1$. The recursive formulas for smoothing in eqn [10] are initialized with $r_n = 0$ and $N_n = 0$. The state smoothing eqn [11] yields the smoothed state estimate $\widehat{\alpha}_t$ and is defined as the optimal estimate of α_t using the full set of observations $\{y_1, y_2, \ldots, y_n\}$; the state smoothing equations also yield the corresponding smoothed state estimation error variance matrix V_t.

Analogous to the predicted state, under the assumption of normality the smoothed state estimation error variance matrix V_t is useful for the construction of confidence intervals for the smoothed state components, which – should we happen to be interested in their 90% confidence limits for example – can be calculated as

$$\widehat{\alpha}_t \pm 1.64\sqrt{V_t}$$

for $t = 1, \ldots, n$.

The recursions for r_{t-1} and N_{t-1} in eqn [10] also enable the computation of the smoothed estimates of the disturbances ε_t and η_t in the following way:

$$\widehat{\varepsilon}_t = H_t\left(F_t^{-1} v_t - K_t' r_t\right), \quad \mathrm{Var}(\widehat{\varepsilon}_t) = H_t\left(F_t^{-1} + K_t' N_t K_t\right) H_t \qquad [12]$$

$$\widehat{\eta}_t = Q_t R_t' r_t, \quad \mathrm{Var}(\widehat{\eta}_t) = Q_t R_t' N_t R_t Q_t \qquad [13]$$

for $t = n, \ldots, 1$. The eqns [12] and [13] compute the smoothed observation disturbances $\widehat{\varepsilon}_t$, the smoothed state disturbances $\widehat{\eta}_t$, and their corresponding smoothed estimation error variance matrices $\mathrm{Var}(\widehat{\varepsilon}_t)$ and $\mathrm{Var}(\widehat{\eta}_t)$, respectively.

Diagnostic Checking

All significance tests in linear Gaussian state space models and the construction of confidence intervals are based on three assumptions concerning the residuals of the analysis: the residuals should satisfy independence, homoscedasticity, and normality, in this order of importance. Whether the residuals satisfy these three assumptions can be established by diagnosing what are known as the *standardized prediction errors*. They are defined as

$$v_t^* = \frac{v_t}{\sqrt{F_t}} \qquad [14]$$

for $t = 1, \ldots, n$. For the computations of the one-step ahead prediction errors v_t and their variances F_t in eqn [14], we refer to the recursive formulas for the Kalman filter given in eqn [9]. In case y_t is a vector of time series, the prediction error v_t^* is defined by $v_t^* = G_t v_t$, where G_t is constructed such that $G_t G_t' = F_t^{-1}$.

The assumptions of independence and normality of the residuals can be diagnosed using the Box–Ljung test statistic and the Bowman and Shenton test statistic, respectively. The assumption of homoscedasticity can be checked by testing whether the variance of the standardized prediction errors in the first third part of the series is equal to the variance of the errors corresponding to the last third part of the series. For further details concerning these diagnostic tests, we refer to Harvey (1989), Durbin and Koopman (2012), and Commandeur and Koopman (2007).

A second diagnostic tool for determining the appropriateness of a model is provided by inspection of what are known as the *auxiliary residuals*. As already mentioned above, the disturbance smoothing filters applied in the backward pass through the data yield, among others, estimates of the smoothed observation and state disturbances, and of their variances. The auxiliary residuals are obtained by dividing the smoothed observation and state disturbances with the square root of their corresponding variances, as follows:

$$e_t^* = \frac{\widehat{\varepsilon}_t}{\sqrt{\mathrm{Var}(\widehat{\varepsilon}_t)}} \text{ and } r_t^* = \frac{\widehat{\eta}_t}{\sqrt{\mathrm{Var}(\widehat{\eta}_t)}} \qquad [15]$$

for $t = 1, \ldots, n$, resulting in *standardized* smoothed disturbances. Visual inspection of the standardized smoothed observation disturbances, e_t^*, allows for the detection of possible *outlier* observations in a time series, while inspection of the standardized smoothed state disturbances, r_t^*, enables the detection of *structural breaks* in the underlying development of a time series.

Each auxiliary residuals can be considered as a t-test for the null hypothesis that there was no outlier observation (when inspecting the auxiliary residuals at the left of eqn [15]) or as a t-test for the null hypothesis that there was no structural break in the corresponding unobserved component of the observed time series (when inspecting the auxiliary residuals at the right of eqn [15]). Applying the usual 95% confidence limits of ± 1.96 corresponding to a two-tailed t-test, possible outlier observations or structural breaks in the unobserved components making up the state vector are thus easily detected. A more detailed discussion on outlier and break detection in a state space analysis is provided by Harvey and Koopman (1992) and de Jong and Penzer (1998).

Parameter Estimation

So far, we have presented all of the results that can be obtained with state space methods as if the disturbance variances, the fixed regression effects, the parameters ρ and λ_c associated with cycles, etc., are given and known. In practice, of course, these parameters are unknown, and have to be estimated.

It can be shown that the Kalman filter presented in eqn [9] also provides the necessary ingredients required for evaluating the log-likelihood function, which is given by

$$\log L(y|\psi) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\log\left|F_t\right| + v_t' F_t^{-1} v_t\right) \quad [16]$$

where the v_t are the one-step ahead prediction errors, the F_t are their variances for $t = 1, \ldots, n$ defined in eqn [9], and ψ denotes the vector of unknown parameters. The log-likelihood eqn [16] is maximized with respect to ψ numerically using the score vector or the EM algorithm; see Durbin and Koopman (2012) and Commandeur and Koopman (2007) for details. In this way we obtain the maximum likelihood estimate of ψ.

Computing

State space methods have been implemented on many computer platforms and in many software packages including STAMP, R, MATLAB, REGCMPNT, SAS, EViews, GAUSS, Stata, RATS, gretl, and SsfPack with links established to S-PLUS and Ox. For more details and an overview of all these software packages, including many interesting applications to time series analysis, we refer to Commandeur et al. (2011).

Non-Gaussian and Nonlinear Time Series

The state space model eqns [1] and [2] are for Gaussian observations y_t that are modeled as linear functions of the state vector. In case of observations y_t that are discrete, are nonlinear in the state vector, or have other distributional properties than Gaussian, we can consider the generalized state space model

$$y_t \sim p(y_t|Z_t\alpha_t) \quad [17]$$

for $t = 1, \ldots, n$, where $p(\cdot)$ is a particular non-Gaussian density function, possibly for discrete observations, and with the state vector α_t and system matrix Z_t as defined in the previous section. The observation density $p(\cdot)$ may be Gaussian but implies a nonlinear relation with α_t. Conditional on the state vector α_t, we assume that the observations are serially independent. This assumption is similar to the linear Gaussian model where ε_t is assumed to be serially uncorrelated. In all cases, the state vector α_t evolves stochastically over time according to the linear Gaussian state eqn [2]. This condition may also be relaxed but in economics and finance we can mostly assume a linear Gaussian process for α_t.

Time series observations which come from the exponential family distribution require density function

$$p(y_t|\theta_t) = \exp\left[y_t'\theta_t - b_t(\theta_t) + c_t(y_t)\right] \quad [18]$$

where $\theta_t = Z_t\alpha_t$, $b_t(\cdot)$ is a twice differentiable function and $c_t(\cdot)$ is a function of y_t only. The modeling framework [17, 2, 18] makes it possible to analyze time series observations coming from, among others, Poisson, binary, binomial, and multinomial distributions. For example, in case a Poisson density is considered, we have $b_t(\theta_t) = \exp\theta_t$ and $c_t(y_t) = \log(y_t!)$.

The generalized state space models [17, 2, 18] are more general than the exponential family only. We can consider continuous density functions for $p(\cdot)$ such as those related to the heavy-tailed distributions Student's t, mixture of normals, and general error. In these cases, we can often return to the linear observation eqn [17] but written as

$$y_t = \theta_t + \varepsilon_t, \quad \varepsilon_t \sim p(\varepsilon_t)$$

where $p(\varepsilon_t) = p(y_t|\theta_t)$. When we take $p(\cdot)$ as the Gaussian density with mean zero and variance H_t, the generalized model reduces to the Gaussian state space models [1] and [2].

A particular example of a non-Gaussian model of interest for financial markets is the stochastic volatility model where the observations are financial returns with a constant mean and a stochastically time-varying variance. The univariate stochastic volatility model is given by

$$y_t = \exp\left(\frac{1}{2}\theta_t\right)\varepsilon_t, \quad \varepsilon_t \sim p(\varepsilon_t) \quad [19]$$

where $p(\varepsilon_t)$ can be Gaussian, Student's t, or mixture of normals. The log variance θ_t can be generally specified as $\theta_t = Z_t\alpha_t$ but it is usually considered as a stationary autoregressive process. For more detailed discussions concerning the stochastic volatility model [19], see Harvey et al. (1994) and Shephard (2005).

The statistical analysis based on nonlinear and non-Gaussian extensions of the state space model can still rely on linear methods such as the Kalman filter and smoothing recursions of Section State Space Analysis but the results may be highly inaccurate. A number of statistical approaches can be adopted to treat non-Gaussian and nonlinear features of the model more adequately and more accurately. For example, Durbin and Koopman (2000) have developed an estimation methodology based on importance sampling and simulated maximum likelihood methods. Bayesian estimation methods for the more general classes of state space models rely typically on Markov chain Monte Carlo methods which are originally developed by Frühwirth-Schnatter (1994), Carter and Kohn (1994), Shephard and Pitt (1997), and others.

Recent Developments

More recent developments on the analysis of generalized or nonlinear state space models are reported in the engineering literature. These developments are based on methods collectively known as particle filtering and originate from the work by Gordon et al. (1993), Kitagawa (1996), and with a noteworthy contribution by Pitt and Shephard (1999). Particle filtering and related Monte Carlo methods are still in development as can be learned from the recent review article of Creal (2012).

Textbook treatments of generalized state space models from both classical and Bayesian perspectives are presented by, among others, Harvey (1989: Section 6.6), West and Harrison (1997), Doucet et al. (2000), and Durbin and Koopman (2012).

See also: Time Series: ARIMA Methods; Time Series: Cycles; Time Series: Economic Forecasting; Time Series: Factor Models; Time Series: General; Time Series: Seasonal Adjustment; Time Series: Volatility Models.

Bibliography

Bell, W.R., 2004. On RegComponents time series models and their applications. In: Harvey, A.C., Koopman, S.J., Shephard, N. (Eds.), State Space and Unobserved Components Models. Cambridge University Press, Cambridge.

Bijleveld, F., Commandeur, J., Gould, P., Koopman, S.J., 2008. Model-based measurement of latent risk in time series with applications. Journal of the Royal Statistical Society A 171, 265–277.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis. Holden-Day, San Francisco.

Box, G.E.P., Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70, 70–79.

Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.

Commandeur, J.J.F., Koopman, S.J., 2007. An Introduction to State Space Time Series Analysis. Oxford University Press, Oxford.

Commandeur, J.J.F., Koopman, S.J., Ooms, M., 2011. Statistical software for state space methods. Journal of Statistical Software 41 (1), 1–18.

Creal, D.D., 2012. A survey of sequential Monte Carlo methods for economics and finance. Econometric Reviews 31, 245–296.

Diebold, F., Rudebusch, S., Aruoba, S., 2006. The macroeconomy and the yield curve: a dynamic latent factor approach. Journal of Econometrics 131, 309–338.

Doucet, A., deFreitas, J.F.G., Gordon, N.J. (Eds.), 2000. Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York.

Doz, C., Giannone, D., Reichlin, L., 2006. A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models. Discussion paper, CEPR.

Durbin, J., Koopman, S.J., 2000. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). Journal of the Royal Statistical Society B 62, 3–56.

Durbin, J., Koopman, S.J., 2012. Time Series Analysis by State Space Methods, second ed. Oxford University Press, Oxford.

Engle, R.F., Watson, M.W., 1981. A one-factor multivariate time series model of metropolitan wage rates. Journal of the American Statistical Association 76, 774–781.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of Time Series Analysis 15, 183–202.

Geweke, J., 1977. The dynamic factor analysis of economic time series. In: Aigner, D.J., Goldberger, A.S. (Eds.), Latent Variables in Socio-economic Models. North-Holland, Amsterdam.

Gordon, N.J., Salmond, D.J., Smith, A.F.M., 1993. A novel approach to non-linear and non-Gaussian Bayesian state estimation. IEE-Proceedings F 140, 107–113.

Harrison, P.J., Stevens, C.F., 1976. Bayesian forecasting. Journal of the Royal Statistical Society B 38, 205–247.

Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Harvey, A.C., Koopman, S.J., 1992. Diagnostic checking of unobserved components time series models. Journal of Business and Economic Statistics 10, 377–389.

Harvey, A.C., Ruiz, E., Shephard, N., 1994. Multivariate stochastic variance models. Review of Economic Studies 61, 247–264.

de Jong, P., Penzer, J., 1998. Diagnosing shocks in time series. Journal of the American Statistical Association 93, 796–806.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, Transactions ASMA Series D 82, 35–45.

Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. Journal of Computational and Graphical Statistics 5, 1–25.

Lawley, D.N., Maxwell, A.E., 1971. Factor Analysis as a Statistical Method, second ed. Butterworths, London.

Litterman, R., Scheinkman, J., 1991. Common factors affecting bond returns. Journal of Fixed Income 1, 54–61.

Pitt, M.K., Shephard, N., 1999. Filtering via simulation: auxiliary particle filter. Journal of the American Statistical Association 94, 590–599.

Proietti, T., 2000. Comparing seasonal components for structural time series models. International Journal of Forecasting 16, 247–260.

Quah, D., Sargent, T.J., 1993. A dynamic index model for large cross sections. In: Stock, J.H., Watson, M. (Eds.), Business Cycles, Indicators and Forecasting. University of Chicago Press, Chicago, pp. 285–306.

Shephard, N., 2005. Stochastic Volatility: Selected Readings. Oxford University Press, Oxford.

Shephard, N., Pitt, M.K., 1997. Likelihood analysis of non-Gaussian measurement time series. Biometrika 84, 653–667.

Stock, J.H., Watson, M., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Tiao, G.C., 2014. Time series: ARIMA methods. In: Wright, J.D. (Ed.), The International Encyclopedia of the Social and Behavioral Sciences, second ed., vol. IV. Elsevier, Amsterdam, pp. 15704–15709.

Watson, M.W., Engle, R.F., 1983. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression. Journal of Econometrics 23, 385–400.

West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models, second ed. Springer-Verlag, New York.

Time Series: Unit-Roots and Nonstandard Limiting Distributions

Ngai H Chan, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.A. Dickey, volume 23, pp. 15731–15735, © 2001, Elsevier Ltd.

Abstract

This article reviews some of the recent developments of the unit root time series and nonstandard limiting distributions. It gives an overview of this important topic, describes the impact of some of the recent progress, and briefly points to open challenges on subsequent research of this field.

Introduction

Unit root time series are ubiquitous in science, engineering, social sciences, and many other disciplines. Many scientific fields, especially in the social sciences, have used nonstationary time series for their problems. Research under the umbrellas of unit root, near unit root, nonstationary, nearly nonstationary, integrated, and near-integrated processes have been pursued actively. In essence, all these titles referred to time series with autoregressive roots on or near the unit circle. Since the seminal paper of White (1958), numerous attempts have been devoted to studying the asymptotic behavior of the least squares estimate (LSE) of the autoregressive coefficient of a unit root auto-regressive (AR) model. In this article, we focus on some of the strategic developments in this area related to statistics and offer some future directions and briefly point to open challenges. Succinct reviews on early developments on this subject can be found in Chan (2006, 2009) and the references therein.

This article is organized as follows. In Sections First-Order Model, Fisher's Information Number, and AR(p) Models, we review the developments of the unit root problem according to the limiting distributions of the LSE for AR(1) and AR(p) models. In Sections Negative Moment Bounds and Mean-Squared Prediction Error, we review the recent developments of negative moment bounds and mean-squared error predictions for nonstationary time series. This article also provides a road map for certain portions of the extensive literature on these topics, and the impact of some of the seminal work in related areas. Section Open Challenges contains ideas on new developments and some concluding remarks.

Unit Root Models

Developments of the unit root problem can be classified into three stages: estimation, inference, and model selection. Throughout these stages, there are some underlying common themes. The seminal paper of Lai et al. (1978) considered the strong consistency property of the LSE of a multiple regression model. Specifically, consider the model

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad i = 1, 2, \ldots \qquad [1]$$

where the ε_i are unobservable errors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are unknown parameters, and y_i is the observable response corresponding to the design vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Then

$$\mathbf{b}_n = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i \qquad [2]$$

is the LSE of the unknown parameter vector $\boldsymbol{\beta}$ based on the observations $\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n$. Herein, the unobservable sequence $\{\varepsilon_n\}$ is assumed to be a martingale difference sequence with respect to an increasing sequence of sigma fields $\mathscr{F}_n$ satisfying a Lyapunov condition

$$\sup_n E(|\varepsilon_n|^\gamma | \mathscr{F}_{n-1}) < \infty \text{ almost surely (a.s.)} \quad \text{for some } \gamma > 2 \qquad [3]$$

By assuming that the design vector at stage n is adaptable, i.e., $\mathbf{x}_n$ is $\mathscr{F}_{n-1}$ measurable, Lai and Wei (1982a) proved the following.

Theorem 2.1. $\mathbf{b}_n \to \boldsymbol{\beta}$ *almost surely if*

$$\lambda_{\min}(n) \to \infty \text{ a.s. and } \log \lambda_{\max}(n) = o[\lambda_{\min}(n)] \text{ a.s.} \qquad [4]$$

where $\lambda_{\min}(n)$ and $\lambda_{\max}(n)$ denote respectively the minimum and the maximum eigenvalues of the design matrix $\sum_{i=1}^{n} x_i x_i^T$ at stage n.

This result extended the earlier work of Anderson and Taylor (1979) and constituted the turning point of the study of strong consistency in the regression literature. Based on this groundbreaking result, Lai and Wei (1982b, 1985) established the strong consistency of the LSE for a general AR model, irrespective to the location of its characteristic roots. This is one of the most general results concerning the strong consistency of the LSE of a general nonstationary AR(p) model. Tiao and Tsay (1983) subsequently considered the weak consistency property of the LSE for a nonstationary ARMA model. Recently, Nielsen (2005) extended Lai and Wei's result to the general VAR model with deterministic components.

First-Order Model

Back in 1958, White among others, first showed that the LSE of the autoregressive coefficient of a nonstationary AR(1) model, $y_t = \beta y_{t-1} + \varepsilon_t$, i.e., when the autoregressive coefficient β equals to one, converges in distribution to a function of

a stochastic integral of a standard Brownian motion. It turns out that this model has a strong bearing in the econometric literature in testing whether or not a time series is a random walk, the so-called unit root testing problem. After Dickey and Fuller (1979) established the form of this limiting distribution as a ratio of sums of identical and independently distributed (iid) random variables, the unit root testing problem became a topical issue in econometrics. Numerous articles were written on this topic and two elementary surveys on the statistical and econometric literature were given in Dickey et al. (1986) and Stock and Watson (1988), respectively.

With the strong consistency results, the next natural question is asymptotic inference. One is interested in the limiting distributions of the LSEs of the parameters of a general nonstationary autoregressive model. For the inference of the LSE $\widehat{\beta}_n$ of β, early developments can be summarized into the next theorem.

Theorem 2.2

1. $|\beta| < 1$, $\sqrt{n}(\widehat{\beta}_n - \beta) \Rightarrow N(0, 1-\beta^2)$, ε_t iid, finite variance, Anderson (1959).
2. $\beta = 1$, $n(\widehat{\beta}_n - \beta) \Rightarrow \frac{\int_0^1 W_t \, dW_t}{\int_0^1 W_t^2 \, dt}$, ε_t iid normal, White (1958).
3. $|\beta| > 1$, $\frac{\beta^n}{\beta^2-1}(\widehat{\beta}_n - \beta) \Rightarrow$ Cauchy, ε_t iid normal, Anderson (1959).

Likewise, if we consider the self-normalized t-statistic, then the preceding theorem can be rewritten as follows.

Theorem 2.3

$$\left(\sum_1^n y_t^2\right)^{1/2}\left(\widehat{\beta}_n - \beta\right) \Rightarrow \begin{cases} Z, & |\beta| < 1 \\ \mathscr{L}(0), & \beta = 1 \\ Z, & |\beta| > 1 \end{cases}$$

Herein, Z denotes a standard normal random variable, $\mathscr{L}(0) = \frac{\int_0^1 W_t \, dW_t}{\left(\int_0^1 W_t^2 \, dt\right)^{1/2}}$, W_t is a standard Brownian motion, and $\Rightarrow$ denotes weak convergence as $n \to \infty$. As will be seen in the next section, the self-normailzed factor $\sum_1^n y_t^2$ is the observed Fisher's information number and the order of its expected value plays a crucial role.

Fisher's Information Number

It turns out that the breakdown of asymptotic normality in the transition of $|\beta| < 1$ to $\beta = 1$ can be inferred through the order of the observed Fisher's information number. Let $I_n = \sum_{t=1}^n y_{t-1}^2$. Then direct calculation shows that

$$E(I_n) \approx \begin{cases} \dfrac{n}{1-\beta^2}, & |\beta| < 1 \\ n^2, & \beta = 1 \end{cases}$$

Based on this fact, Lai and Siegmund (1983) established the following result under a sequential sampling scheme.

Theorem 2.4. Lai and Siegmund (1983). Let $N_c = \inf\{n : I_n \geq c\}$. Then

$$(I_{N_c})^{1/2}\left(\widehat{\beta}_n - \beta\right) \Rightarrow Z$$

uniformly in β as $c \to \infty$.

This result indicates that when the Fisher's information number reaches a certain horizon c and when the horizon c tends to infinity, then uniform asymptotic normality for the LSE can be attained. Although this result carries a lot of theoretical gravity, most time series were not sampled sequentially with c tending to infinity. In practice, most of the time series would only be sampled with n tending to infinity. Nevertheless, Theorem 2.4 reveals that a key quantity to establishing the asymptotic result is given by the magnitude of the observed Fisher's information number I_n. If we need to know the asymptotic behavior of the LSE $\widehat{\beta}_n$, then we can take advantage of the behavior of I_n in a neighborhood of one. We quantify the closeness of β to one by maintaining the order of $E(I_n)$ in the vicinity of one. To this end, let $\beta_n = 1 - \gamma/a_n$. Direct calculation shows that

$$E(I_n) = \frac{n}{1-\beta_n^2} = \frac{na_n}{\gamma(2-\gamma/a_n)}$$

To maintain $E(I_n) \approx n^2$, we need to take $a_n = n$ and hence the nearly unstable model is characterized by using

$$\beta_n = 1 - \frac{\gamma}{n}$$

where γ is a nuisance parameter. In short, the near unstable AR(1) model becomes

$$y_t = \beta_n y_{t-1} + \varepsilon_t, \quad t = 1, \ldots, n, \; y_0 = 0, \; \beta_n = 1 - \frac{\gamma}{n} \quad [5]$$

Chan and Wei (1987) considered such an AR(1) model when the autoregressive coefficient converges to one asymptotically. Instead of a Brownian motion, Chan and Wei (1987) showed that the limiting distribution of the LSE of a nearly nonstationary AR(1) model converges weakly to a functional of an Ornstein–Uhlenbeck process. Specifically, considering a triangular array of first-order autoregressive processes as given in eqn [5], Chan and Wei derived the following theorem.

Theorem 2.5. Chan and Wei (1987). Let y_t follow a nearly unstable AR(1) model and consider the statistics $\widehat{\tau}_n = (\sum_1^n y_{t-1}^2)^{1/2}(\widehat{\beta}_n - \beta_n)$. Then

1. $\widehat{\tau}_n \Rightarrow \frac{\int_0^1 W_{\gamma,t} \, dW_t}{\left(\int_0^1 W_{\gamma,t}^2 \, dt\right)^{1/2}} := \mathscr{L}(\gamma)$, where $W_{\gamma,t}$ is the Ornstein–Uhlenbeck process. Furthermore,
2. $\mathscr{L}(\gamma) \Rightarrow Z$ as $|\gamma| \to \infty$.

This particular result encompasses the unit root case of White (1958) when $\gamma = 0$. In this case, the autoregressive coefficient $\beta_n = 1$ and $\mathscr{L}(0) = \int_0^1 W(t) \; dW(t)/(\int_0^1 W^2(t) \, dt)^{1/2}$, which is the limiting distribution of the Dickey–Fuller statistic. Using reproducing kernels, Chan (1988) further developed this limiting form as sums of iid random variables. The near-integrated notion was formulated with reference to the work of LeCam about limiting experiments in terms of contiguous alternatives. This idea was later explored by Jeganathan (1991, 1995), who generalized the near-integrated notion to a general AR(p) case and introduced the idea of Local Asymptotic Brownian Functional in studying optimality issues. Furthermore, part (2) of Theorem 2.5 indicates the limiting behavior of an explosive model, which was used by Phillips and Yu (2009) to study the issues of financial bubbles.

More recently, Buchmann and Chan (2007) combined the notions of long-memory and near nonstationarity and established the limiting distributions of the LSE under such a general setting, in which they made use of Malliavin calculus.

In a spectral setting, Dahlhaus (1985) considered both tapered and nontapered Yule–Walker estimates for near-integrated models. Since then, numerous extensions have been carried out by econometricians and statisticians. For example, on the statistical front, Cox and Llatas (1991) considered the M-estimation of a near nonstationary process, Pham (1992) and Ing (2001) studied the bias and prediction mean square error expansion of the LSE of a near unit root model, Basawa et al. (1991) investigated the bootstrap estimate of a unit root model, Larsson (1998) considered the Bartlett correction property of the unit root test statistics and Chan (1988) extended the notion of near unit root to a seasonal model. On the econometric front, the issues of testing for the unit root hypothesis for econometric series was considered in Phillips (1987) and Phillips and Perron (1988), among others. Many of these results were explored further by various people under the topics of trend breaks and cointegration. Interested readers can consult the review articles of Stock (1994) and Watson (1994) on these topics. Finally, it should also be pointed out that while extensive studies were pursued by statisticians and econometricians alike, one of the earlier developments in this problem was first given in the monograph of Arató (1982) and later in an unpublished thesis of Bobkoski (1983).

A closely related problem to the unit root AR(1) model is a unit root moving average order one (MA(1)) model given by

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1} \qquad [6]$$

with $\{\varepsilon_t\}$ satisfying eqn [3]. Asymptotic properties of the maximum likelihood estimate (MLE) $\widehat{\theta}_n$ of the MA coefficient when $\theta = 1$ constitute an actively pursued area. Cryer and Ledolter (1981) first examined the consistency property of $\widehat{\theta}_n$. Davis and Dunsmuir (1996) derived the limiting distribution of the maximum likelihood estimator of θ under the unit root $\theta = 1$ and $\theta = 1 - c/n$ (for a constant $c > 0$) setups. Interestingly, there is a significant pile-up probability, i.e., $P(\widehat{\theta}_n = 1)$ has a nonzero limit for all values of $c > 0$. The limiting pile-up probabilities can be quite large especially for small values of c. Saikkonen and Luukkonen (1993) examined the testing for a moving average unit root issue. Using the idea of the so-called derived process, Chan and Tsay (1996) studied the limiting distribution of the MLE for a general unit root MA(q) model. A systematic account on the unit root moving average model is given in Chapter 8 of Tanaka (1996).

AR(p) Models

Chan and Wei (1988) considered the limiting distributions of the LSE of a general nonstationary AR(p) model when the characteristic roots lie on or outside the unit circle, each of which may have different multiplicities. This was the first comprehensive treatment of the LSE for a general nonstationary AR(p) model, and it was shown in Chan and Wei (1988) that the locations of the roots of the time series played an important role in characterizing the limiting distributions. Specifically, they considered a general nonstationary AR(p) model

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t \qquad [7]$$

In eqn [7], the autoregressive polynomial $\beta(z) = 1 - \beta_1 z - \cdots - \beta_p z^p$ has roots lying on or outside the unit circle. That is,

$$\beta(z) = (1-z)^a (1+z)^b \Pi_{k=1}^{\ell} \left(1 - 2\cos\theta_k z + z^2\right)^{d_k} \psi(z) \qquad [8]$$

where a,b,ℓ,d_k are nonnegative integers, θ_k belongs to $(0,\pi)$ and $\psi(z)$ is a polynomial of order $r = p - (a + b + 2d_1 + \cdots + 2d_k)$ that has all roots outside the unit disk.

When the underlying model is stationary with all roots lying outside the unit circle, classical central limit theorem type results can be obtained. But when the roots are of unit modulus, it turns out that the asymptotic distributions are characterized in terms of the integrated integrals of Brownian motions. The key idea in obtaining these results lies in analyzing the order of magnitude of the observed Fisher's information matrix. Note that the LSE of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ can be expressed as

$$\mathbf{b}_n = \left(\sum_{t=1}^{n} \mathbf{y}_{t-1}\mathbf{y}_{t-1}^T\right)^{-1} \sum_{t=1}^{n} \mathbf{y}_{t-1} y_t \qquad [9]$$

where $\mathbf{y}_t = (y_t, \ldots, y_{t-p+1})^T$ and $\mathbf{y}_0 = (0, \ldots, 0)^T$. Similar to the estimation problem, different characteristic roots carry different information. By transforming the original nonstationary AR model into components according to their characteristic roots, Chan and Wei (1988) were able to derive the precise form of the limiting distributions. During the course of this investigation, they also obtained an important result about the weak convergence of stochastic integrals (Theorem 2.4 in Chan and Wei, 1988), which is of independent interest and has many applications in different areas, see for example, Kurtz and Protter (1991). In addition, Chan and Wei also showed that different components are asymptotically uncorrelated and, as a result, a joint limiting law can be established. Specifically, using the notation of Chan and Wei (1988), the following theorem was established.

Theorem 2.6. Assume that $\{y_t\}$ follows eqn [7] with the characteristic polynomial satisfying eqn [8] and the innovation sequence $\{\varepsilon_t\}$ satisfying eqn [3]. Then as $n \to \infty$

$$Q^T G_n^T \left(b_n - \beta\right) \to_D \Big(\left(F^{-1}\xi\right)^T, \left(\tilde{F}^{-1}\eta\right)^T, \left(H_1^{-1}\zeta_1\right)^T, \ldots, \left(H_\ell^{-1}\zeta_\ell\right)^T, N^T\Big)^T$$

where $(F,\xi),(\tilde{F},\boldsymbol{\eta}),(H_1,\zeta_1), \ldots, (H_\ell,\zeta_\ell),N,G_n,Q$ are independent and defined in eqns (3.2), (3.3), and Theorem 2.2 of Chan and Wei (1988).

This result of Chan and Wei paved the way to the analysis of nonstationary processes and since then, numerous extensions have been conducted. Jeganathan (1991) generalized this idea to the near-integrated situations where the limiting distributions of the LSE are expressed in terms of iterated integrals of Ornstein–Uhlenbeck processes. When the underlying model has long-memory, Chan and Terrin (1995) extended this result to

functionals of fractional Brownian motions, while Chan and Zhang (2012) considered the case when the innovations are heavy-tailed. In a recent paper, Buchmann and Chan (2013) introduced a Jordan canonical decomposition idea to combine the notions of near nonstationarity and long-memory behavior for such a general AR(p) model. Finally, extensions of this result to vector AR processes are given in Tsay and Tiao (1990) and to processes with deterministic trends in Chan (1989b). On the econometric front, Theorem 2.4 provides a fundamental tool in analyzing cointegrated systems. Comprehensive reviews on cointegrated vector autoregressions are given in Johansen (2009). A useful introduction of using cointegration ideas in statistical arbitrage is given in Chapter 15 of Chan (2010).

Model Selection

Consider the stochastic regression model eqn [1] again. A natural problem is to study the performance of the LSEs in prediction and formulated the so-called predictive principle for model selection. By analyzing the order of the cumulative predictive errors

$$C_n = \sum_{k=1}^{n} \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_k - \mathbf{b}_{k-1}^{\mathrm{T}}\mathbf{x}_k\right)^2 = \sum_{k=1}^{n} \left(\widehat{\varepsilon}_k - \varepsilon_k\right)^2$$

where $\widehat{\varepsilon}_k = y_k - \widehat{y}_k = y_k - \mathbf{b}_{k-1}^{\mathrm{T}}\mathbf{x}_k$ is the one-step prediction error, Wei (1987, 1992) observed that the term C_n plays a crucial role for model selection. For a succinct review on this issue, we refer the readers to Chan (2009) and the references therein.

Negative Moment Bounds

Subsequent to Wei's prediction error principle, further developments can be found in Ing (2007) and Ing et al. (2010). In this section, we review some of the recent developments about prediction errors calculations for unit root time series. It turns out that this important problem is related to the negative moment bounds of the underlying time series. Specifically, let $y_1, \ldots, y_n$ be generated from the stochastic regression model,

$$y_t = g_t(\boldsymbol{\eta}_0) + \varepsilon_t, \quad t = 1, \ldots, n \qquad [10]$$

where $\boldsymbol{\eta}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,p1}, \beta_{0,1}, \ldots, \beta_{0,p2})^T$ is an unknown coefficient vector and $g_t(\boldsymbol{\eta}_0)$ has the auto-regressive moving average (ARMA) representation

$$g_t(\boldsymbol{\eta}_0) = \alpha_{0,1}y_{t-1} + \cdots + \alpha_{0,p_1}y_{t-p_1} - \beta_{0,1}\varepsilon_{t-1} - \cdots - \beta_{0,p_2}\varepsilon_{t-p_2} \qquad [11]$$

with the initial conditions $y_t = \varepsilon_t = 0$ for all $t \le 0$. Define

$$\widehat{\boldsymbol{\eta}}_n = \arg\min_{\eta\in\Pi} \sum_{t=1}^{n} [y_t - g_t(\boldsymbol{\eta})]^2$$

where $\prod \subset R^{p_1+p_2}$ is a compact set that includes $\boldsymbol{\eta}_0$ as an interior point and whose elements $\boldsymbol{\eta} = (\alpha_1, \ldots, \alpha_{p1}, \beta_1, \ldots, \beta_{p2})^T$ satisfy the following properties: for all $|z| \le 1$,

$$A_{1,\boldsymbol{\eta}}(z) = 1 - \sum_{j=1}^{p_1} \alpha_j z^j \neq 0, \quad A_{2,\boldsymbol{\eta}}(z) = 1 - \sum_{j=1}^{p_2} \beta_j z^j \neq 0$$

$A_{1,\boldsymbol{\eta}}(z)$ and $A_{2,\boldsymbol{\eta}}(z)$ have no common zeros with $\left|\alpha_{p_1}\right| + \left|\beta_{p_2}\right| > 0$

An important issue is to investigate the moment convergence of the LSE, in particular the order of

$$\mathrm{E}\left\|n^{1/2}(\widehat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\right\|^q, q \ge 1 \qquad [12]$$

It turns out that the uniform negative moment bound for $n^{-1}\sum_{t=1}^{n}[g_t'(\boldsymbol{\eta})]^2$, that is,

$$\mathrm{E}\left\{\sup_{\boldsymbol{\eta}\in B_{\delta_1}(\boldsymbol{\eta}_0)} \left(n^{-1}\sum_{t=1}^{n}\left[g_t'(\boldsymbol{\eta})\right]^2\right)^{-q}\right\} = O(1), q \ge 1 \qquad [13]$$

plays a crucial role in establishing eqn [12], where $B_{\delta 1}(\boldsymbol{\eta}_0) = \{\boldsymbol{\eta}:|\boldsymbol{\eta} - \boldsymbol{\eta}_0| < \delta_1\}$ for some $\delta_1 > 0$.

For the ARMA model $\{y_t\}$ defined in eqn [10], we have the following theorem.

Theorem 3.1. Chan and Ing (2011). Assume model [10], with $g_t(\cdot)$ as defined in eqn [11] and ε_t being independent random variables satisfying $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^2) = \sigma^2$ for all $t \ge 1$. Moreover, assume that there exist positive constants α_1, ξ and M_1 such that for any $0 < s_2 - s_1 \le \xi$

$$\sup_{1\le m\le t<\infty, \|\mathbf{v}\|=1} \left|F_{t,m,\mathbf{v}}(s_2) - F_{t,m,\mathbf{v}}(s_1)\right| \le M_1(s_2 - s_1)^{\alpha_1}$$

where $\mathbf{v} \in R^m$ and the function $F_{t,m,\mathrm{v}}(\cdot)$ denotes the distribution function of the scalar product $\mathbf{v}^T(\varepsilon_t, \cdots, \varepsilon_{t+1-m})^T$. Then,

$$\mathrm{E}\left\|n^{1/2}(\widehat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\right\|^q = O(1), q \ge 1 \qquad [14]$$

As an application of this Theorem, an asymptotic expression for the mean-squared prediction error (MSPE) of $\widehat{\boldsymbol{\eta}}_n$, $\mathrm{E}\{y_{n+1} - g_{n+1}(\widehat{\boldsymbol{\eta}}_n)\}^2$, is given in the next Theorem.

Theorem 3.2. Chan and Ing (2011). Assume that the assumptions of the preceding Theorem hold. Moreover, let ε_t be iid random variables satisfying for some $q_1 > 18$,

$$\mathrm{E}|\varepsilon_1|^{q_1} < \infty$$

Then,

$$\lim_{n\to\infty} n\left[\mathrm{E}\{y_{n+1} - g_{n+1}(\widehat{\boldsymbol{\eta}}_n)\}^2 - \sigma^2\right] = (p_1 + p_2)\sigma^2 := \overline{p}\sigma^2$$

Subsequently, this result has been extended to the long-memory situation. In particular, when $\{y_t\}$ follows an auto-regressive fractionally integrated moving average (ARFIMA) model such as

$$\alpha(B)(1 - B)^{d_0}y_t = \beta(B)\varepsilon_t \qquad [15]$$

where $\alpha(B)$ and $\beta(B)$ are defined as before and $d_0 \in R$, we arrive at the next Theorem. Unlike the traditional setting, note that d_0 is no longer restricted to $[-1/2, 1/2]$.

Theorem 3.3. Chan et al., (2013). Assume that $\{y_t\}$ satisfies an ARFIMA model as given in eqn [15]. Further, let $d_0 \in$, $D = [L, U]$ with $-\infty < L < U < \infty$. Under certain regularity conditions as those given in Theorem 3.1, the negative moment bound eqn [14] holds, that is

$$\mathrm{E}\left\|n^{1/2}(\widehat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}_0)\right\|^q = O(1), q \ge 1$$

Mean-Squared Prediction Error

In this section, we apply the result of the Section Negative Moment Bounds to study the prediction problem. To predict y_{n+h}, $h \geq 1$, based on $y_1, \ldots, y_n$, consider

$$\begin{aligned} p_t(\boldsymbol{\eta}, y_{t-1}, \ldots, y_1) &:= y_t - \varepsilon_t(\boldsymbol{\eta}) \\ &= \left[1 - (1-B)^d A_{1,\eta}(B) A_{2,\eta}^{-1}(B)\right] y_t \end{aligned}$$

The one-step conditional sum of squares (CSS) predictor of y_{n+1} is given by

$$\begin{aligned} \widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n) &= p_{n+1}(\widehat{\boldsymbol{\eta}}_n, y_n, \ldots, y_1) \\ &= y_{n+1} - \varepsilon_{n+1}(\widehat{\boldsymbol{\eta}}_n) = \widehat{\boldsymbol{\eta}}_n^T \mathbf{y}_n \end{aligned}$$

Note that $y_{n+1}(\widehat{\boldsymbol{\eta}}_n)$ depends solely on $y_1, \ldots, y_n$. Further, y_{n+h}, $h \geq 2$, can be predicted recursively by the h-step CSS predictor,

$$\widehat{y}_{n+h}(\widehat{\boldsymbol{\eta}}_n) := p_{n+h}(\widehat{\boldsymbol{\eta}}_n, \widehat{y}_{n+h-1}(\widehat{\boldsymbol{\eta}}_n), \ldots, \widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n), y_n, \ldots, y_1)$$

which is sometimes known as the plug-in predictor. For the one-step MSPE, it is shown in Chan et al. (2013) that

$$\lim_{n\to\infty} n\left[\mathrm{E}\left\{y_{n+1} - \widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n)\right\}^2 - \sigma^2\right] = \bar{p}\sigma^2 := (p_1 + p_2 + 1)\sigma^2$$

This result shows that the second-order MSPE of $\widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n)$ only depends on the number of estimated parameters; the dependent structure of the underlying process is irrelevant as far as the second-order MSPE is concerned. This result not only is interesting, but it is also intriguing and revealing when compared to the second-order MSPE of the least squares (LS) predictor for integrated AR models. Assume that a forecaster thinks that the true model is possibly an integrated AR(p_1) model,

$$\begin{aligned} &\left(1 - \tilde{\alpha}_1 B - \cdots - \tilde{\alpha}_{p_1} B^{p_1}\right) y_t = \varepsilon_t \\ &(1-B)^{v_0}\left(1 - \theta_1 B - \cdots - \theta_{p_1 - v_0} B^{p_1 - v_0}\right) y_t = \varepsilon_t \end{aligned} \qquad [16]$$

where $1 - \theta_1 z - \cdots - \theta_{p_1 - d} z^{p_1 - v_0} \neq 0$ for all $|z| \leq 1$ and $v_0 \in \{0, 1, \ldots, p_1\}$ is unknown. Then it is natural for this forecaster to predict y_{n+1} using the LS predictor $\tilde{y}_{n+1}$, in which $\tilde{y}_{n+1} = y_n^T(p_1)\tilde{\boldsymbol{\alpha}}_n(p_1)$ with $\mathbf{y}_t^T(p_1) = (y_t, \ldots, y_{t-p_1+1})$ and $\tilde{\boldsymbol{\alpha}}_n(p_1)$ satisfying

$$\sum_{t=p_1}^{n-1} \mathbf{y}_t(p_1)\mathbf{y}_t^T(p_1)\tilde{\boldsymbol{\alpha}}_n(p_1) = \sum_{t=p_1}^{n-1} \mathbf{y}_t(p_1) y_{t+1}$$

Suppose that there is another forecaster having doubts on whether the v_0 in eqn [16] is really an integer. This forecaster chooses the more general alternative,

$$\left(1 - \alpha_1 B - \cdots - \alpha_{p_1} B^{p_1}\right)(1-B)^{d_0} y_t = \varepsilon_t$$

where $L_1 \leq 0 \leq d_0 \leq p_1 \leq U_1$ with $-\infty < L_1 < U_1 < \infty$ being some prescribed numbers. This model includes eqn [16] as a special case and the forecaster naturally chooses CSS predictor.

If v_0 is an integer, according to Ing et al. (2010), the MSPE of the LS predictor satisfies

$$\lim_{n\to\infty} n\left[\mathrm{E}\left\{y_{n+1} - \tilde{y}_{n+1}\right\}^2 - \sigma^2\right] = (p_1 + v_0^2)\sigma^2$$

On the other hand, according to the aforementioned result of Chan et al. (2013), the MSPE of the CSS predictor satisfies

$$\lim_{n\to\infty} n\left[\mathrm{E}\left\{y_{n+1} - \widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n)\right\}^2 - \sigma^2\right] = (p_1 + 1)\sigma^2$$

While the second-order MSPE of the LS predictor $\tilde{y}_{n+1}$ increases with the dependence structure (i.e., v_0 increases), the second-order MSPE of the CSS $\widehat{y}_{n+1}(\widehat{\boldsymbol{\eta}}_n)$ does not change with v_0. This observation reveals the interesting fact that for integrated AR models, the popular LS predictor is inferior to the CSS predictor.

For multistep MSPE, Chan et al. (2013) established the following result.

$$\begin{aligned} &\lim_{n\to\infty} n\left\{\mathrm{E}[y_{n+h} - \widehat{y}_{n+h}(\widehat{\boldsymbol{\eta}}_n)]^2 - \sigma_h^2(\boldsymbol{\eta}_0)\right\}/\sigma^2 \\ &\quad = \underline{f}_h(p_1, p_2) + \underline{g}_h(\boldsymbol{\eta}_0) + 2J_h(\boldsymbol{\eta}_0) \\ &\quad = \text{ARMA} + I(d) + \text{ARMA and } I(d) \text{ (joint effect)} \end{aligned}$$

Unlike the one-step MSPE situation, the dependence structure has a substantial effect on the multistep MSPE. Consider the pure $I(d)$ case. It follows that for any $d_0 \in R$, there exists $0 < C_{1,d_0} \leq C_{2,d_0} < \infty$ such that

$$\begin{aligned} C_{1,d_0} h^{-1+2d_0} &\leq n\mathrm{E}[y_{n+h} - \widehat{y}_{n+h}(\widehat{\boldsymbol{\eta}}_n)]^2 - \sigma_h^2\left(\boldsymbol{\eta}_0\right)\Big\}\Big/\sigma^2 \\ &\leq C_{2,d_0} h^{-1+2d_0} \end{aligned}$$

which shows that for $h > 1$, a larger d_0 (stronger dependence) gives rise to a larger second-order MSPE.

What about the least squares method? Suppose that a random walk model, $y_t = \alpha y_{t-1} + \varepsilon_t$ for $\alpha = 1$ is modeled by an integrated order d (I(d)) process, $(1-B)^d y_t = \varepsilon_t$, in which $d = 1$ corresponds to the true model. Then

$$\lim_{n\to\infty} n\left\{\mathrm{E}\left[y_{n+h} - \widehat{y}_{n+h}\left(\widehat{d}_n\right)\right]^2 - h\sigma^2\right\} = \sigma^2 \quad \text{for } h = 1$$

and the same limit is smaller than

$$\left[4.87h + (1 + \log h)^2 - 2(1 + \log 2h)\right]\sigma^2 \quad \text{for } h \geq 2$$

Let $\tilde{y}_{n+h}$ be the h-step LS predictor of this AR(1) model. It follows from Ing et al. (2010) that

$$\lim_{n\to\infty} n\left\{\mathrm{E}(y_{n+h} - \tilde{y}_{n+h})^2 - h\sigma^2\right\} = 2h^2\sigma^2$$

which is larger than σ^2 when $h = 1$, and larger than $[4.87h + (1 + \log h)^2 - 2(1 + \log 2h)]\sigma^2$ when $h \geq 2$. Hence, $\widehat{y}_{n+h}(\widehat{d}_n)$ is always better than $\tilde{y}_{n+h}$ in terms of the MSPE.

As far as estimations are concerned, the situation is completely reversed, however. The LSE converges much faster to 1 than $\widehat{d}_n$ under the random walk model. This example reminds us that when the true model belongs simultaneously to several different parametric families, the optimal choice of parametric families can vary according to different purposes of statistical inferences, e.g., estimation, testing, and prediction, etc. Is there a general rule? The above result suggests the following phenomena.

- More Fisher information, less estimation risk.
- Less Fisher information, less prediction risk.

This general principle seems to indicate that as $n \to \infty$,

$$n \times (\text{One-stepMSPE} - \sigma^2) \to \frac{1}{\log n}[\log\det(\text{Fisher Information Matrix})]$$

in probability.

Open Challenges

Although the field of time series in general and the subject of unit root time series in particular have been going through rigorous developments for the last few decades, there are still many open and interesting areas to be explored. In nonstationary time series, the recent popularity of long-memory models remains an open and challenging field, see Chan and Terrin (1995) and Buchmann and Chan (2007, 2013). Equally important is the area of model selection and prediction errors. In particular, we are faced with the challenges of:

- How to develop consistent model selection criteria for ARFIMA(p_1,d,p_2) with $d \in R$? What about Bayesian Information Criterion (BIC)-type criteria?
- How to develop asymptotically efficient criteria for ARFIMA(∞,d,0) with $d \in R$? Should we still use Akaike Information Criterion (AIC)?

The recent development of Chan and Ing (2011) and Chan et al. (2013) offer some insights into solving these issues. In summary, this article only offers a modest account for some of the developments of unit root time series. As witnessed by the aforementioned discussions, the subject of unit root time series has been manifested into different forms and found applications in areas such as engineering and econometrics. Through asymptotic inference, the scope of unit root time series has been greatly broadened and many profound challenges still remain to be resolved, which in turn will push the subject into new frontiers.

Acknowledgment

This research was supported in part by HKSAR-RGC-GRF Grants 400313, 400410 and HKSAR-RGC-CRF: CityU8/CRG/12G.

See also: Model Selection and Model Averaging; Time Series: Cointegration; Time Series: General.

Bibliography

Anderson, T.W., 1959. On asymptotic distributions of estimates of parameters of stochastic difference equations. Annals of Mathematical Statistics 30, 676–687.
Anderson, T.W., Taylor, J., 1979. Strong consistency of least squares estimators in dynamic models. Annals of Statistics 7, 484–489.
Arató, M., 1982. Linear Stochastic Systems with Constant Coefficients: A Statistical Approach. Springer-Verlag, Berlin. Lecture Notes in Control and Information Sciences 45.
Basawa, I.V., Mallik, A.K., McCormick, W.P., Reeves, J.H., Taylor, R.L., 1991. Bootstrapping unstable first-order autoregressive processes. Annals of Statistics 19, 1098–1101.
Bobkoski, M.J., 1983. Hypothesis Testing in Nonstationary Time Series (Unpublished Ph.D. thesis). Department of Statistics, University of Wisconsin, Madison, USA.
Buchmann, B., Chan, N.H., 2007. Inference for nearly unstable processes under strong dependence. Annals of Statistics 35, 2001–2017.
Buchmann, B., Chan, N.H., 2013. Unified asymptotic theory for nearly unstable AR(p) processes. Stochastic Processes and Their Applications 123, 952–985.
Chan, N.H., 1988. The parameter inference for nearly nonstationary time series. Journal of the American Statistical Association 83, 857–862.
Chan, N.H., 1989a. On the nearly nonstationary seasonal time series. Canadian Journal of Statistics 17, 279–284.
Chan, N.H., 1989b. Asymptotic inference for unstable autoregressive time series with drifts. Journal Statistical Planning and Inference 23, 301–312.
Chan, N.H., 2006. Inference for time series and stochastic processes. Statistica Sinica 16, 683–696.
Chan, N.H., 2009. Time series with roots on or near the unit circle. In: Andersen, T.G., Davis, R.A., Krei, J.P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer-Verlag, New York, pp. 695–707.
Chan, N.H., 2010. Time Series: Applications to Finance with R and S-Plus, second ed. Wiley, New York.
Chan, N.H., Ing, C.K., 2011. Uniform moment bounds of Fisher's information with applications to time series. Annals of Statistics 39, 1526–1550.
Chan, N.H., Terrin, N.C., 1995. Inference for unstable long-memory processes with applications to fractional unit-root autoregressions. Annals of Statistics 23, 1662–1683.
Chan, N.H., Tsay, R.S., 1996. Asymptotic inference for non-invertible moving average time series. Journal of Time Series Analysis 17, 1–17.
Chan, N.H., Wei, C.Z., 1987. Asymptotic inference for nearly nonstationary AR(1) processes. Annals of Statistics 15, 1050–1063.
Chan, N.H., Wei, C.Z., 1988. Limiting distributions of least squares estimates of unstable autoregressive processes. Annals of Statistics 16, 367–401.
Chan, N.H., Zhang, R.M., 2012. Nonstationary autoregressive processes with infinite variance. Journal of Time Series Analysis 33, 916–934.
Chan, N.H., Huang, S.F., Ing, C.K., 2013. Moment bounds and mean squared prediction errors of long-memory time series. Annals of Statistics 41, 1268–1298.
Cox, D.D., Llatas, I., 1991. Maximum likelihood type estimation for nearly nonstationary autoregressive time series. Annals of Statistics 19, 1109–1128.
Cryer, J.D., Ledolter, J., 1981. Small sample properties of the maximum likelihood estimator in the first-order moving average model. Biometrika 68, 191–194.
Dahlhaus, R., 1985. Data Tapers in Time Series Analysis (Habilitation thesis). Universität-GHS, Essen, Essen, Germany.
Davis, R.A., Dunsmuir, W.T.M., 1996. Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. Econometric Theory 12, 1–29.
Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit-root. Journal of the American Statistical Association 74, 427–431.
Dickey, D.A., Bell, W.R., Miller, R.B., 1986. Unit-roots in time series models: tests and implications. The American Statistician 40, 12–26.
Ing, C.K., 2001. A note on mean-squared prediction errors of the least squares predictors in random walk models. Journal of Time Series Analysis 22, 711–724.
Ing, C.K., 2007. Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. Annals of Statistics 35, 1238–1277.
Ing, C.K., Sin, C.Y., Yu, S.H., 2010. Prediction errors in nonstationary autoregressions of infinite order. Econometric Theory 26, 774–803.
Ing, C.K., Wei, C.Z., 2003. On same-realization prediction in an infinite-order autoregressive process. Journal of Multivariate Analysis 85, 130–155.
Ing, C.K., Wei, C.Z., 2005. Order selection for the same-realization prediction in autoregressive processes. Annals of Statistics 33, 2423–2474.
Jeganathan, P., 1991. On the asymptotic behavior of least-squares estimators in AR time series with roots neat the unit circle. Econometric Theory 7, 269–306.
Jeganathan, P., 1995. Some aspects of asymptotic theory with applications to time series models. Econometric Theory 11, 818–887.
Johansen, S., 2009. Cointegration: overview and development. In: Andersen, T.G., Davis, R.A., Kreiss, J.P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer-Verlag, New York, pp. 671–693.
Kurtz, T.G., Protter, P., 1991. Weak limit theorems for stochastic integrals and stochastic differential equations. Annals of Probability 19, 1035–1070.
Lai, T.L., Robbins, H., Wei, C.Z., 1978. Strong consistency of least squares estimates in multiple regression. Proceedings of the National Academy of Sciences of the United States of America 75, 3034–3036.
Lai, T.L., Siegmund, D., 1983. Fixed accuracy estimation of an autoregressive parameter. Annals of Statistics 11, 478–485.
Lai, T.L., Wei, C.Z., 1982a. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics 10, 154–166.
Lai, T.L., Wei, C.Z., 1982b. Asymptotic properties of general autoregressive models and strong consistency of least squares estimates of their parameters. Journal of Multivariate Analysis 13, 1–23.
Lai, T.L., Wei, C.Z., 1985. Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis VI. North-Holland, Amsterdam, pp. 375–393.
Larsson, R., 1998. Bartlett corrections for unit-root test statistics. Journal of Time Series Analysis 19, 425–438.
Nielsen, B., 2005. Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. Econometric Theory 21, 534–561.
Pham, D.T., 1992. Approximate distribution of parameter estimators for first-order autoregressive models. Journal of Time Series Analysis 13, 147–170.

Phillips, P.C.B., 1987. Toward a unified asymptotic theory of autoregressions. Biometrika 74, 535–547.

Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.

Phillips, P.C.B., Yu, J., 2009. Limit Theory for Dating the Origination and Collapse of Mildly Explosive Periods in Time Series Data. Singapore Management University. Technical Report.

Saikkonen, P., Luukkonen, R., 1993. Testing for a moving average unit-root in autoregressive integrated moving average models. Journal of the American Statistical Association 88, 596–601.

Stock, J.H., Watson, M.W., 1988. Variable trends in economic time series. Journal of Economic Perspective 2, 147–174.

Stock, J.H., 1994. Unit-roots, structural breaks and trends. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Amsterdam, pp. 2739–2841.

Tanaka, K., 1996. Time Series Analysis: Nonstationary and Noninvertible Distribution Theory. Wiley, New York.

Tiao, G.C., Tsay, R.S., 1983. Consistency properties of least squares estimates of autoregressive parameters in ARMA models. Annals of Statistics 11, 856–871.

Tsay, R.S., Tiao, G.C., 1990. Asymptotic properties of multivariate nonstationary processes with applications to autoregressions. Annals of Statistics 18, 220–250.

Watson, M.W., 1994. Vector autoregressions and cointegration. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier, Amsterdam, pp. 2843–2915.

Wei, C.Z., 1985. Asymptotic properties of least squares estimates in stochastic regression models. Annals of Statistics 13, 1498–1508.

Wei, C.Z., 1987. Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. Annals of Statistics 15, 1667–1682.

Wei, C.Z., 1992. On predictive least squares principles. Annals of Statistics 20, 1–42.

White, J.S., 1958. The limiting distribution of the serial correlation coefficient in the explosive case. Annals of Mathematical Statistics 29, 1188–1197.

Time Use and Gender

Tanja van der Lippe, Utrecht University, Utrecht, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Men and women have to make decisions on how to use their time, as a day has only 24 h, or rather time is limited. It is best measured by using the time diary method. The traditional gender division of time use has changed over the last decades: more women have paid jobs and the division of household labor has become more equal due to the decline in domestic work of women. Explanations of gender differences in time use are given by new home economics, exchange theory, and gender role theory. In addition to individual factors, differences in the institutional context are important as well for time use decisions of men and women.

Time Use

Men and women have to make decisions on how to use their time, as a day has only 24 h, or rather time is limited. Activities men and women make decisions about can be divided into paid work, domestic work and child care, personal obligations, and leisure. Time has become an increasingly scarce good; newspapers and other media abound with stories suggesting that time pressure is becoming a real epidemic. This article discusses how time use is measured and the patterns of time use of men and women and provides explanations of gender differences.

The Measurement of Time Use

To understand time use of men and women, it can be categorized in several ways. If one is interested in paid work and leisure, a distinction is made between work and nonwork activities. With respect to the division of labor between husbands and wives, the most frequently used division is between paid work, domestic work and child care. It is evident what is meant by paid work, and most authors use more or less the same description. However, household labor can be defined in a variety of ways. Naturally, cooking and cleaning are domestic tasks, but it is not clear how to classify activities that are part leisure and part work, such as vegetable gardening, gourmet cooking, or playing with the children. Leisure is a broad category, including indoor activities like reading books, watching television and gaming, outdoor activities like sporting, but also social contacts and going out for dinner. Personal obligations often form a kind of residual category, including activities like sleeping and eating at home.

The most accurate and reliable way to measure the time use of people is the time diary method (Andorka, 1987; Gershuny, 2000; Van der Lippe et al., 2011). In this approach, respondents are asked to keep records of what they were doing at all times for a day or more. In some studies the main activity is noted in fixed interval units (e.g., in quarters of an hour) while in others intervals are left free (i.e., starting and finishing times are indicated). The period of time when the diary is kept also differs. In some cases respondents are only required to keep a diary for a single day, while in others they are asked to do so for a whole week. The time diary method prevents respondents from exaggerating the durations of their favorite activities while neglecting the time devoted to day-to-day concerns and trivialities. Thus, the actual subject being studied is not overstated, which sometimes happens when specialized questionnaires are used. A serious limitation of time diaries is the difficulty in dealing with more than one activity performed simultaneously. For example, people may cook dinner while supervising children, or sew while watching television. Another method to measure patterns of time use is asking direct questions that require men and women to estimate their usual time spent on a list of activities. In order to help the respondent, in these questionnaires often domestic work is distinguished into some clear categories like cooking, cleaning, doing the laundry, and shopping. Analyses of the data from such questionnaires and the Experience Sampling Method show that husbands overestimate their own time, but that their wives make accurate estimates of husbands' time on domestic work. With respect to the amount of time wives spend on domestic work, Lee and Waite (2005) conclude that both wives and husbands may substantially overestimate this.

Time Use Patterns

There has been a traditional gender division of time consuming activities within Western societies: women used to do most of the domestic work and child care, whereas men worked for pay the whole day and were responsible for providing family income. This has changed substantially in recent decades. At the beginning of the twenty-first century, women are more likely to have a paid job than to be housewives in almost all industrialized countries, a statement, one could not imagine just after World War II (Van der Lippe, 2010). Despite the overall increase in levels of women's employment, large differences exist between countries in the northern continents. Nordic countries, such as Sweden and Denmark, have had higher levels of female labor force participation than other Western countries for a long time. The Anglo-Saxon countries, such as the United States and the UK were runners up, while in Italy still only half of the active female population has a paid job. The Netherlands has had a low level of women's

employment for a long time, but recently the number of working women increased sharply. For eastern Europe, the picture is different. The communist ideology, established after the World War II in eastern Europe with the exception of Russia, forced many women to enter the labor market. In the 1960s already three quarters of the women had a paid job. Activity rates of women decreased after the transition from socialism in these countries and they are now among the lowest in Europe (Eurostat, 2012).

Despite the overall increase in the level of women's employment, there are considerable differences between countries in the number of hours per week that women work for pay. While part-time work is common in Norway, United Kingdom, and the Netherlands, employed women in Italy, Greece, Spain, and Portugal almost all work full-time. Also about 80% of US employed women work full-time (Van der Lippe, 2010). In eastern Europe part-time work simply did not exist before the political turnover. After 1989, full-time work is still the rule in the former socialist countries (Van der Lippe and Fodor, 1998).

The growth in paid work of women is linked to the often noted decline in domestic work on the part of women in Western countries (Gershuny, 2000). Compared to this large change in women's time use in the west in general, there has been hardly any change in men's time use. Although there has been a small increase in men's domestic work, these changes are trivial compared to the increase in women's employment and the reduction in their domestic work. The more equal division of domestic work between husbands and wives nowadays is less a result of an increase in men's domestic work and more because of the reduced time women spend on household duties (Treas and Drobnic, 2010). When men perform domestic tasks, they disproportionately do what are often considered more 'pleasant' tasks (e.g., child care and cooking rather than cleaning the toilet).

Although the amount of leisure time differs less between the sexes as between those in paid and domestic work, the nature of leisure time is different (Mattingly and Bianchi, 2003). Women are more disturbed in their leisure time by other time activities, probably due to their responsibility for children during their leisure time. Some authors show that the amount of leisure time for both men and women has increased over the last decades due to a decline in market work hours and home production work hours (Aguiar and Hurst, 2007). Others show a decline in leisure time among parents, which is sharper for mothers than fathers suggesting that mothers give up leisure time so they can spend more time to paid work ánd child care (Sayer, 2005). Results show that partners now spend more time together engaged in various leisure activities than they did in the past, not only in percentages but also in absolute number of minutes (Voorpostel et al., 2010).

Explanations for the Different Use of Time by Gender

A rich body of literature is available on men and women's allocation of time (Gershuny, 2000; Bianchi et al., 2000; Shelton and John, 1996). Many study explanations of men's and women's division of paid and domestic work (Bianchi and Milkie, 2010; Hook, 2006; Van der Lippe et al., 2011) and some of these studies compare thereby several countries (Georgas et al., 2006).

The Individual Level

An important theoretical explanation of time allocation has been provided by economic theory (Becker, 1965). Given their constraints or resources, human beings strive for goals and make choices in their time use. An obvious resource within economic theory is human capital. Human capital constitutes the knowledge and skills that increase an individual's productivity. Education is often considered the most important form of human capital. So, for highly educated women and men, it is appealing to devote a great deal of their time to working in the labor market, since the benefits in terms of money and status can be high. This implies less time for domestic work and nonwork activities. To understand the division of paid and domestic work between husbands and wives, economic theory is valuable as well. Within the 'new home economics,' explicit attention is given to the importance of household activities (Becker, 1981). In this model of household production, households are viewed as small factories that, within the time and income limitations, make every effort to maximize their output. If the husband can earn more money and is less productive in household labor than his wife, the husband will specialize in paid labor and the wife in household labor (Becker, 1981). In Becker's view, which does not highlight any conflicts of interest between spouses but assumes altruism, each spouse specializes in what will make his or her best contribution to the family.

Sociological exchange theory (Blood and Wolfe, 1960; Brines, 1994) provides a different account of spousal specialization. Differences in resources, such as education and income, are considered to be a main cause for the different time use by men and women. Assuming that individuals do not like to do domestic work, this theory predicts that the spouse with the greater resources will have more power and will use it to minimize his or her domestic work. Thus the predictions are similar to those of the new home economics theory that spouses with higher relative earnings will do less household work, but the reasoning is different. Whereas in economic theory, each spouse chooses between market and household work in order to maximize contributions to the family, exchange theory typically treats employment and earnings as exogenous and examines how they affect bargaining power, and sees doing household work as an indication of less power. Research supports the notion that men's contribution to housework reduces the higher their earnings relative to female spouses (Bittman et al., 2003; Greenstein, 2000; Presser, 1994), although American men who are economically dependent on their wives compensate by adopting traditional behaviors in the household (Greenstein, 2000; Brines, 1994). Relative education is related to division of housework in a same way, implying that men of higher educated women do a higher proportion in housework (Bianchi et al., 2000). Results with respect to the link between the participation of women in paid labor and the division of household labor show that hours of female paid work have a strong negative effect on women's

domestic work but hardly any effect on men's time spent on domestic duties (Van der Lippe, 2010).

Another important theoretical perspective emphasizes gender specific socialization (Presser, 1994; Greenstein, 2000). It is not such a long time ago that women were socialized to be mothers and housewives and men to be wage-earners and support the family. More recently there is a great deal of variation in the degree of gender traditionality of socialization. Research suggests that men who adhere to traditional sex roles contribute less to household tasks than men whose sex roles are nontraditional (Kamo, 1988; Fuwa, 2004) and it has also been noted that women with modern role expectations do more paid and less domestic work (Presser, 1994). However, doing gender perspectives argue that household duties are perceived as 'women's work,' so that the effect of full-time paid work of the women on male participation in household work might be less than expected.

These theories can also partly explain the change in time use of men and women in recent decades. Given that the educational level of women all over the Western world has increased and is nowadays comparable to the level of men, according to the resource or productivity argument it has become attractive for women to spend more time on paid work. Consequently, they have less time available for domestic work and leisure. Furthermore, attitudes in the Western world have changed. Nowadays, working women are accepted in society and fathering has become one of the opportunities for men. According to the sex-specific socialization, it has become more likely as well that men carry more of the domestic workload than some decades ago, and women make a career. Indeed we do see an increase in time spent in child care by men, but rather unexpectedly also by women. Given the sociological and economic explanations, one would expect the opposite for women. It might be that the psychological meaning of children has changed. Children in our modern society are a highly valued 'good,' in which one wishes to invest a great deal of time. Moreover, the growing threat of the environment might play an important factor. The number of cars has increased continuously with the consequential danger of teaching children to play unattended or travel to activities alone. Finally, it might be an artifact of the methodology, due to the fact that only first activities are reported. Full-time housewives in unmechanized households used to do several household and caring tasks simultaneously, but reported only the first activity.

The Institutional Level

In addition to individual factors, differences in the institutional context are important for time use decisions of men and women. Although time use decisions are made on a micro level, the institutional context of a country can facilitate or restrict a certain time-use decision. Increasingly it is studied to what extent individual-level explanations hold in different contexts. There are differences among Western countries, which become clear using Esping-Anderson's typology (1990) of Welfare State regimes (liberal, conservative, social democratic), or related more gender-friendly typologies. In social democratic countries, such as the Scandinavian ones, public child care facilities are more extensive and women are stimulated more to have a paid job. In conservative countries, for a long time, women were not encouraged to have a paid job, and consequently child care was not available. In countries with less well-developed child care arrangements, women with young children are less likely to be employed than women with older children (Gornick et al., 1998). Based on a review of studies between 2000 and 2009 into household labor, Lachance-Grzela and Bouchard (2010) show that there is only some evidence supporting the idea that political efforts and social policies can modify the behaviors that occur in the home, but results are inconsistent. National contexts in which female empowerment is greater and in which structures advocate gender equity tend to favor egalitarian allocations of household labor within couples more (Lachance-Grzela and Bouchard, 2010). Moreover, lagged institutional accomodations to female employment is related to more gender equality in household division of labor (Treas and Tai, 2012).

In addition, macro indicators interact in a more significant way with micro indicators for the explanation of paid work than for the explanation of housework. In countries with a large economic development, higher educated women do more paid work – it seems rewarding for them in terms of money and status to spend much time on the labor market. In these economies, the family phase is important as well: for women having children implies that less time is spent on the labor market. For men, this is only true when having young children. Married women spend less time on paid labor, and unexpectedly this is also true for married men (Van der Lippe et al., 2011). Is this an emancipation effect of highly developed economies? More research is needed to investigate the relation between highly developed economies and part-time work of both men and women. In line with the disappearing boundaries between countries and states, due to the increase in communication and information technology, comparative studies on time use become increasingly important. At this moment, information is available on many Western industrialized countries, but one can only hope to compare their figures with those of Asian, South American, and African countries.

Time Use and Time Pressure

As balancing work with family life has become important nowadays, increasingly attention is also paid to feelings of time pressure. Many people feel down between work and family, not only because their households increasingly juggle competing responsibilities, but also because job expectation and parenting standards have becoming more demanding (Van der Lippe, 2007). In two-earner families, both partners are exposed to the greed associated with the jobs in which they work, and thus both will bring the restrictions imposed by their work situation to the time allocation in the household. Consequences range from feelings of stress, work–home interference, time pressure, and fatigue to burnout and other health problems. Spending much time on paid work and domestic duties implies more time pressure for both men and women, although women in general experience more time pressure.

See also: Family and Gender; Family Bargaining; Household Production; Motherhood: Economic Aspects.

Bibliography

Andorka, R., 1987. Time budgets and their uses. Annual Review of Sociology 13, 149–164.

Aguiar, M., Hurst, M., 2007. Measuring trends in leisure: the allocation of time over five decades. The Quarterly Journal of Economics 969–1006.

Becker, G.S., 1965. A theory of the allocation of time. The Economic Journal 80, 493–517.

Becker, G.S., 1981. A Treatise on the Family. Harvard University Press, Cambridge, MA.

Bianchi, S.M., Milkie, M.A., Sayer, L.C., Robinson, J.P., 2000. Is anyone doing the housework? Trends in the gender division of household labor. Social Forces 79, 191–228.

Bianchi, S.M., Milkie, M.A., 2010. Work and family research in the first decade of the 21st Century. Journal of Marriage and Family 72, 705–725.

Bittman, M., England, P., Folbre, N., Sayer, L., Matheson, G., 2003. When does gender trump money? Bargaining and time in household work. American Journal of Sociology 109, 186–214.

Blood, R.O., Wolfe, D.M., 1960. Husbands and Wives: The Dynamics of Married Living. The Free Press, New York.

Brines, J., 1994. Economic dependency, gender, and the division of labor at home. American Journal of Sociology 100, 652–688.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton, NJ.

Eurostat, 2012. Employment Statistics, 2012 – Statistics Explained (2013/6/1). http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Employment_statistics#.

Fuwa, M., 2004. Macro-level gender inequality and the division of household labor in 22 countries. American Sociological Review 69, 751–757.

Georgas, J., Berry, J.W., de Vijver, van, Kagitcibasi, C., Poortinga, Y.H. (Eds.), 2006. Family Structure and Function: A 30 Nation Psychological Study. Cambridge University Press, Cambridge.

Gershuny, J., 2000. Changing Times. Work and Leisure in Postindustrial Society. Oxford University Press, Oxford.

Gornick, J.C., Meyers, M.K., Ross, K., 1998. Public policies and the employment of mothers: a cross-national study. Social Science Quarterly 79, 35–54.

Greenstein, T., 2000. Economic dependence, gender, and the division of labor in the home: a replication and extension. Journal of Marriage and the Family 62, 322–335.

Hook, J.L., 2006. Care in context: men's unpaid work in 20 countries, 1965–2003. American Sociological Review 71, 639–660.

Kamo, Y., 1988. Determinants of household division of labor: resources, power, and ideology. Journal of Family Issues 9, 177–200.

Lachance-Grzela, M., Bouchard, G., 2010. Why do women do the lion's share of housework? A decade of research. Sex Roles 63, 767–780.

Lee, Y., Waite, L., 2005. Husbands' and wives' time spent on housework: a comparison of measures. Journal of Marriage and the Family 67, 328–336.

Mattingly, M., Bianchi, S.M., 2003. Gender differences in the quantity and quality of free time: the U.S. experience. Social Forces 81, 999–1030.

Presser, H.B., 1994. Employment schedules, gender, and household labor. American Sociological Review 59, 348–364.

Robinson, J.P., 1977. How Americans Use Time: A Social–Psychological Analysis of Everyday Behavior. Praeger, New York.

Sayer, L., 2005. Gender, time and inequality: trends in women's and men's paid work, unpaid work and free time. Social Forces 81, 285–303.

Shelton, B.A., John, D., 1996. The division of household labor. Annual Review of Sociology 22, 299–322.

Treas, J., Drobnic, S., 2010. Household Division of Labor in Cross National Perspective. Stanford University Press, Stanford.

Treas, J., Tai, T., 2012. Apron strings of working mothers: maternal employment and housework in cross-national perspective. Social Science Research 41, 833–842.

Van der Lippe, T., Fodor, E., 1998. Changes in gender inequality in six eastern European countries. Acta Sociologica 41, 131–149.

Van der Lippe, T., 2007. Dutch workers and time pressure: household and workplace characteristics. Work, Employment and Society 21, 693–711.

Van der Lippe, T., 2010. Women's employment and household work. In: Treas, J., Drobnic, S. (Eds.), Household Division of Labor in Cross National Perspective. Stanford University Press, Stanford, pp. 41–58.

Van der Lippe, T., van der, De Ruijter, E., De Ruijter, J., Raub, W., 2011. Persistent inequalities in time use between men and women: a detailed look at the influence of economic circumstances, policies, and culture. European Sociological Review 27, 164–179.

Voorpostel, M., Van der Lippe, T., Gershuny, J., 2010. Spending time together: changes over four decades in leisure time spent with a spouse. Journal of Leisure Research 42, 243–365.

Time Use and Resource Consumption

Henrike Rau, School of Political Science and Sociology, National University of Ireland, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The ways in which people view and use time have very real material consequences that both shape and reflect the relationship between human society and the physical environment, thereby affecting sustainability transitions. More or less resource-intensive time cultures, that is, shared repertoires of time-related meanings and solutions to everyday problems, coexist both within and between countries. Moreover, they often cut across other socioeconomic and cultural fault lines such as class, gender, and rural–urban divisions. Having remained largely underresearched in the past, the recent 'temporal turn' in the social sciences has highlighted these time-cultural differences and their social and ecological impacts.

Introduction

The ways in which people use their time have very real material consequences that both shape and reflect the relationship between human society and the physical environment. For example, novel production and consumption practices afforded by the discovery of fossil fuels and parallel changes in the socioeconomic structures of societies such as Britain and Germany in the eighteenth and nineteenth-century are claimed to have initiated an acceleration and desynchronization of everyday life (Rosa, 2005). On the other hand, predictions by advocates of modern time management that time-saving technologies will radically reduce working hours and enhance people's quality of life did not materialize, partly because of the intensification of work and consumption whereby material and time-related rebound effects cancel out many technology-aided efficiency gains. Interestingly, the fundamental question of how people spent the time they saved and how this impacts society and the environment has never been satisfactorily answered.

Recent sustainability debates among academics, policymakers, and practitioners have highlighted this close connection between material consumption and human time use, albeit often in an unsystematic way, resulting in calls for a radical transformation of how people view and use their time (Jackson, 2009; Schor, 2010; cf Rau and Edmondson, 2013 for an overview). Bottom-up sustainability initiatives in many western societies such as efforts toward downshifting, voluntary simplicity movements, and collaborative consumption networks incorporate more or less radical changes in human time use. Many of these proposals imply a transformation of people's time culture, that is, their shared repertoires of time-related meanings and solutions to everyday problems (Rau, 2004), although this is not always made explicit. Proposals abound for changes in people's time horizon, a shift away from short-term thinking toward long-term planning, a deceleration of everyday life, and a reallocation of temporal resources away from consumption and toward nonconsumerist activities such as voluntary work or 'quality time' with others (cf Rau and Edmondson, 2013).

Yet others argue for the temporal reorganization of society through changes in the political conditions and power relations that underpin the temporal-spatial ordering of human social life. For example, Mückenberger's research on urban temporal policies and their influence on citizens' quality of life captures historical and contemporary links between governance, citizenship, and time (e.g., Boulin and Mückenberger, 1999; Mückenberger, 2010, 2011; Läpple et al., 2010). Recognizing the centrality and significance of time organization and 'time-welfare' for social cohesion and quality of life, he advocates time policies, that is, deliberate political efforts to enhance people's everyday time-related conditions (Mückenberger, 2011). In addition, there is evidence of an emergent temporal-spatial transformation of towns and cities in Europe and beyond. Recent work by Pink and Servon (2013) records the growing popularity of the Slow City (Cittaslow) movement in Europe as both a deceleration counternarrative and a novel approach to urban development. They argue that Cittaslow successfully connects (experiences of) the time-related qualities of a town or city and the quality of life of its citizens. Cittaslow exemplifies the increasingly diverse range of proposals for the creation of a new (urban) time culture, to counteract the negative implications of unfettered acceleration. The success, or otherwise, of such efforts to influence the temporal topography of cities requires coordinated social, institutional, and material transformations, including changes in urban form that foster more sustainable time use.

The remainder of this article is divided into four principal parts. Initially, it examines important connections between people's time use and their consumption of material and nonmaterial resources. Subsequently, it presents evidence for the continued coexistence of distinct time cultures both within and across countries, which in turn require different material resources and which produce more or less sustainable ecological footprints. Thirdly, the chapter critically examines proposals for a more just and sustainable society beyond the 'growth logic' of capitalism that call for a radical shift in time use and time culture. It concludes with a general outlook regarding the link between time use and resource consumption in theory, research, and practice.

Connections between Time Use and Resource Consumption

The complexity of the relationship between people's time use and their consumption of material and nonmaterial resources and its potential implications for society and the environment

cannot be overestimated. Records of both historical and recent transformations in the organization of time and associated changes in its allocation to different activities reveal this. Many developed and developing countries now rely on clock time as their main source of temporal reference and societal synchronization, replacing alternative modes of time reckoning that rely on natural pacers such as diurnal and lunar cycles, seasons and cycles of growth, and decay (Adam, 1996, 1998). The resulting discrepancy between natural cycles and social pacers in many modern capitalist societies, made possible through the availability of cheap fossil fuels, more efficient industrial production methods, in particular with regard to food, as well as the reorganization of labor, has been identified as a key sustainability challenge. The growing availability of electricity and lighting throughout the twentieth century produced conditions for a '24-h society' that never sleeps, thus decoupling human activities from naturally occurring diurnal cycles and increasing society's ecological footprint in the process. A shift toward renewable energy sources, which is frequently advocated as part of a sustainability transition, is unlikely to reverse this disconnect between societal and natural rhythms.

Sociopolitical and cultural transformations and parallel changes in infrastructure and systems of provision that facilitate individualized activities, as opposed to collectively shared ones, have initiated the desynchronization of human activities and associated rises in resource consumption. Automobility afforded by pro-car policies, the construction of road networks, the availability of (relatively) cheap car and fossil fuels, and the acquisition of driving skills by a large proportion of the population serves as a prime example. Even though the amount of time spent on mobility has changed relatively little over the past 100 years, the level and quality of people's 'consumption of distance' has increased significantly due to the availability of technology such as cars and planes and associated infrastructure (Rau, 2014). Moreover, expectations regarding how much or how little time people are prepared to spend traveling have been transformed in the process, with commuting being seen as 'normal' by many today. The resulting implications for the environment are enormous, with the transport sector consistently featuring among the top three greenhouse gas emitters globally.

That said, the individualization of time use and resource consumption is neither inevitable nor all-encompassing. As Higginson et al. (2013) were able to show, the persistent concomitance of particular activities (e.g., making tea, watching TV, cooking dinner) and associated peaks and troughs in energy demand in UK households present significant barriers to time-shifting demand for energy to match more volatile supply from renewables such as wind energy. While this may be seen as 'bad news' by those who advocate a flexibilization of energy demand in response to the volatility of renewable energy supplies, the apparent persistence of these collectively shared and synchronized time use patterns could play an equally significant role in sustainability transitions if their ecological footprint was to be radically reduced. In other words, the modification of existing mainstream time use practices could fruitfully complement the introduction and mainstreaming of less resource-intensive niche activities frequently associated in the literature with sustainability transitions.

Systematic efforts to theorize and conceptualize the role of time in society–nature interactions remain the exception (cf Rau et al., 2014). Nevertheless, the significance of time use patterns for both the organization of human society and its interactions with the biophysical environment has increasingly been recognized, especially in the field of interdisciplinary sustainability research. Marina Fischer-Kowalski's seminal social-ecological analysis of the relationship between productive work, societal metabolism, and (un)sustainable development points toward the significance of a temporal reorganization of societies for their material resource base (e.g., Fischer-Kowalski and Schaffartzik, 2008). Other authors have identified inherent conflicts between natural and cultural time regimes, pointing to the longevity of nuclear waste beyond the human time-scale or the 'unnatural' rhythms of the 24-h society (Adam, 1996, 1998; Adam et al., 1997). Yet others emphasize the persistence of forms of entrainment whereby the timing of social processes in modern societies continues to be influenced by naturally occurring cycles. Trentmann's (2009) examination of the impact of extreme weather events such as drought, floods, or heat waves on both 'normal' social practices and shared infrastructural systems offers unique insights into the continuities and discontinuities of society–nature interactions in modern societies.

The extent to which different forms of time use incur different ecological costs presents an interesting sustainability challenge. What work and leisure activities require the most energy, water, food, or mobility? In this context, time use patterns associated with particular lifestyles have been shown to be more resource-intensive than others. For example, the ecological footprint of an executive from a Western industrialized country whose work involves numerous long-haul flights per annum, who consumes mostly processed foods, and who engages in resource-intensive leisure pursuits contrasts with that of a small-scale farmer in a developing country who spends most of his/her time to grow food for himself/herself and him/her family and whose everyday life is closely linked to the locality. In addition, the global spread of resource-intensive 'Western' lifestyles, which almost always coincides with adjustments in how people view and use time, has increased resource consumption in areas such as transport and energy. Finally, it has been shown that major life events or 'turning points' in people's biography can impact their consumption patterns in significant ways, clearly demonstrating the need for systematic longitudinal monitoring of time use and resource consumption across the life course (Schäfer et al., 2012; Rau et al., 2014).

To quantify time-use-related differences in ecological footprint has proven to be challenging, although some progress has been made of late. For example, Jalas' (2002) study of two-person households in Finland offers an interesting example of the opportunities and challenges that arise when measuring the energy intensity of selected domestic activities, including food preparation and consumption, washing and cleaning, reading or watching TV. While Jalas himself recognizes the methodological limitations of his approach such as difficulties demarcating activities and the subsequent allocation of energy demand to each activity, his work nevertheless provided a very important starting point

for empirically connecting time use and resource consumption. Similar efforts have since been made. For example, Druckman et al. (2012) use UK data to explore how 'nonwork' activities such as household chores and leisure pursuits give rise to varying amounts of household carbon emissions. By doing so they establish that leisure activities are generally associated with lower carbon emission than nonleisure activities, a fact that has significant implications for gendered analyses of time use and resource intensity.

On an international scale, the availability of harmonized international time use data sets such as the Multinational Time Use Study (MTUS) based at the Centre for Time Use Research in Oxford (UK) offers new opportunities for linking people's time use and their resource consumption and for comparing temporal-ecological footprints across countries. Improved data availability, for instance, through 'smart' metering of household energy use, can provide important insights. These could be complemented by informed estimates of (in)direct resource impacts of different time use categories, for example, through expert rankings of different activities according to how much energy, water, or mobility they require, or how much waste or greenhouse gas emissions they generate. Finally, in-depth qualitative research that captures the social significance of different time use categories could help identify activities that are more or less open to transformation. The relevance of such research efforts for the development of sustainability policies cannot be overestimated.

There is considerable evidence that people's working time has decreased since the Industrial Revolution, which can be attributed to time-saving technologies as well as successful political campaigns for a reduction in working hours and greater regulation of working time. However, measurable reductions in average working hours frequently contradict perceptions of growing time pressure across many societies. In this context, the intensification of time use through multitasking deserves attention as a potential source of stress and harriedness (Shove et al., 2009). Claims that people's days have 40 hours or more if all parallel activities are taken into account demonstrate the significance of what more conventional time use studies label as secondary or tertiary activities. Importantly, these additional activities also incur environmental costs that need to be considered in sustainability assessments of different time use patterns. Is multitasking bad for the environment? This and related questions deserve greater attention in the future.

To summarize, existing work linking time use and resource consumption clearly shows how the social organization of time constitutes an important link between social and natural processes. Greater engagement with the topic of time in social-scientific and interdisciplinary sustainability research, as demonstrated by research discussed in this section, also demonstrates the need to develop alternative theoretical and conceptual approaches that recognize the cultural specificities of human time use both within and across societies.

The Continuous Diversity of Time Cultures

Arguments abound that the worldwide expansion of a modern, clock-based time regime has contributed to the marginalization of alternative time cultures. There is ample evidence of homogenizing influences that impact on the diversity of time cultures today, many of which are associated with the global spread of capitalism and its dependence on economic growth (cf Geißler et al., 2006). For example, Geißler (2002) observes how the spread of a globalized temporal monoculture has drastically reduced time-cultural diversity around the world.

> Modern industrial society has maltreated temporal diversity. It has absorbed and negated it. Time has been mapped monoculturally in line with the abstract measurement of clock time. The aim of this has been to make anything that does not follow unified clock time as compatible with the latter as possible.
>
> Geißler, 2002, p. 132

Interestingly, the taming of human nature through time discipline was part of diverse religious, modernist, and nationalist projects in the past, which frequently included the establishment of powerful institutional time regimes to synchronize social life. For example, attempts by economic and political elites to standardize the measurement and use of time accompanied the establishment of many Western nation-states since the seventeenth century (Zerubavel, 1982; Elias, 1984). Here, Norbert Elias's (1984) work on time and the 'civilizing process' details the gradual internalization of temporal rules set by powerful institutions and elites, including court societies, the church, and the landed gentry.

However, claims of an apparent loss of time-cultural diversity are by no means universally accepted. According to Zerubavel (1997, p. 109), "our standard system of reckoning the time is still not absolutely universal today." There are many alternative time reckoning strategies which continue to coexist worldwide, despite the dominance of Western calendars and clock time. In fact, it is possible to detect a rich tapestry of different time cultures both in more traditional societies that are largely reliant on subsistence farming determined by seasonal cycles and in developed countries whose economies depend on rigorous time-keeping using clocks and Information and Communications Technologies (Rau, 2004, 2014). In fact, most countries harbor significant time-cultural differences. Sociological and anthropological research has revealed a great deal of intranational time-cultural diversity, frequently intersecting with other sociocultural, economic, and material gradients such as class, gender, or rural–urban variations in everyday practices and associated (infrastructural) systems of provision. Moreover, observable changes in time regimes are often context-specific, such as in the case of intraorganizational transformations in working time regimes in a particular company that impact on workers and outside suppliers as well as their families and social networks. For example, Ladner (2009) observes that a strong emphasis on calculating and managing time experienced by employees in the Internet advertising sector significantly influences their domestic lives and that these 'spillovers' from the workplace into the home affect men and women very differently.

Attempts by powerful religious and secular elites to control people's physical environment and standardize their use of time have also remained incomplete because individuals and social

groups have successfully resisted their imposition. Efforts to protect 'nature' and traditional ways of life against the homogenizing pressures exerted by modern capitalism are central to many social movements worldwide, most notably those committed to environmental protection and indigenous rights. Interestingly, the protection of traditional time cultures from modernity has not attracted the same level of collective action, although historical accounts of labor movements in England, Germany, and other countries detail many acts of resistance to the commoditization and calculation of time in the work place (Thompson, 1967). Deliberate efforts to protect this time-cultural diversity such as the *Society for the Slowing Down of Time* (*Verein zur Verzögerung der Zeit)* established in 1990 in Austria remain the exception, partly because the taken-for-granted nature of many time cultures appears to complicate their protection. (http://www.zeitverein.com/framesets/fs_zeitverein.html (last accessed 08.12.13.).) That said, the increasing urgency of many contemporary social-ecological problems has sparked a new wave of interest in alternative and potentially more sustainable time cultures, albeit largely on the margins of the sustainability debate.

The extent to which the diversity of time cultures is under threat because of economic globalization and associated forms of cultural homogenization is unknown. There is ample evidence that a rich diversity of time cultures persists in both (semi-)traditional or modern societies. On the other hand, many alternatives to clock-time thinking and the quantification of time have come under significant pressure. UNESCO's efforts to protect intangible forms of cultural heritage, including cooking rituals and festivals, clearly point toward the possibilities and challenges of sustaining time-cultural diversity into the future.

That said, possibilities exist for protecting socially and environmentally beneficial time cultures against the pressures of a rapidly expanding economic system based on accelerating flows of goods, money, ideas, and people. Using qualitative scenario building, recent sustainable consumption research identified promising practices that require a more or less radical reallocation of temporal and material resources and that could be rolled out within households or organizations. For example, the Spaces for Sustainable Eating scenario for 2050 that emerged from CONSENSUS research envisages more time to engage in cooking, growing food, and eating together, partly because of a reduction and flexibilization of working time (www.consensus.ie; cf Davies et al., 2014 for details of all visioning and backcasting work carried out as part of CONSENSUS). In fact, the growing application of visioning and backcasting techniques in sustainability research and practice, including urban planning and local energy transitions, presents opportunities for refocusing people's time horizon toward the future. Importantly, this type of work demonstrates the need for a long-term approach to sustainability problems that lack an immediate and clear solution, including the transformation of resource-intensive time use patterns in households and organizations and associated shifts in time culture.

Finally, time-cultural differences influence shared understandings of what constitutes a sustainable society and how to achieve it, although this is often not sufficiently recognized (Rau and Edmondson, 2013). For example, variations in time culture relating to how (much) people plan and what pace of life they prefer are likely to shape their sustainability thinking and practice. Are time cultures that embrace long-term planning and distal goals more likely to make the transition to sustainability? Will fast-paced societies find it more difficult to 'slow down' to reduce their ecological footprint and enhance people's quality of life? An in-depth understanding of variations in time cultures between countries, regions, or communities is crucial to their respective sustainability efforts.

Proposals for a Transformation of Time Cultures toward Greater Sustainability

Given that time represents a central aspect of sustainability thinking and practice, it seems little surprising that both theoretical debates regarding sustainability and proposals for concrete measures to promote greater sustainability have covered the topic of time, albeit not always with sufficient depth and rigor. Most definitions of sustainability advocate a long-term approach to development that involves the fair, just, and viable distribution of material resources and opportunities for human flourishing both within *and* across generations. According to Hueting and Reijnders (1998, p. 139), "[t]he notion of sustainability has a long intellectual history, going back to the concept of a 'stationary' or 'steady state' economy employed by nineteenth-century economists […] preference for intergenerational equity has always remained a core element of the concept." Many mainstream indicators of (sustainable) human development also incorporate time-related elements, including life expectancy (cf Rau and Edmondson, 2013).

Claims that a sustainability transition will require a radical transformation of existing time culture which may or may not be compatible with existing systems of time-space regulation have gained momentum. For example, key contributors to sustainable consumption debates have argued for fundamental shift in how individuals, households, and societies view and use time (e.g., Jackson, 2009; Schor, 2010). Juliet Schor (2010) advocates major changes in the allocation of time to address today's pressing social and ecological problems. This also implies a radical shift in how individuals and households produce and consume goods and services. In her view, 'true wealth' depends on the diversification of economic activity and a move away from 'business-as-usual' (BAU) markets and their temporal structures. She observes the emergence of a new social stratum of 'pioneers' whose microlevel activities are geared toward 'plenitude' rather than increased material prosperity. Schor's vision of an alternative social order based on a radical reallocation of temporal *and* material resources by individual actors clearly differs from existing models of (time) poverty and wealth that call for collective action to change the material conditions in society. Instead she advocates a shift toward 'true materialism' that recognizes the environmental costs of production and consumption.

Similarly, Tim Jackson (2009) advocates a redefining of prosperity that embraces human flourishing within ecological limits and that challenges the current uncritical acceptance of economic growth as engine of human development. Importantly, he criticizes inherently unsustainable consumption and spending habits of individuals, organizations, and governments

that result from people's preferences for goods and actions purely because of their novelty, or newness. "[…T]wo interrelated features of modern economic life […] drive the growth dynamic: the production and consumption of novelty" (Jackson, 2009, p. 9). This overemphasis on novelty and its association with human and societal progress has amplified the use of materials and the production of pollution and waste, thus gradually undermining the foundations of human existence on earth. Efforts to stop human-induced climate change and to address other pressing ecological problems thus require a fundamental change in people's time-cultural values and practices that challenge the uncritical acceptance of novelty as inherently positive.

Interestingly, existing efforts to tackle the social and ecological (un)sustainability of consumerism and 'throwaway thinking' in many capitalist societies have focused on the potential (re)introduction of time-cultural alternatives that incorporate both novel and traditional time-related values such as an appreciation of slowness, durability, and 'things that last,' alongside greater recognition of the intrinsically social nature of time and its important synchronizing effects on society (e.g., Grober, 2002). Countercultural forms of resistance to the perceived acceleration and intensification of social and economic life in late modern societies, including the Slow Food movement, newly emerging forms of association that focus on recycling and repairing, or groups engaged in explicitly slow leisure pursuits such as walking, deserve particular attention in this context. It remains to be seen whether or not these niche practices will become mainstream and potentially displace existing 'fast' practices as part of a society-wide sustainability transition. Given that the modernization of many societies has coincided with a largely uncritical rejection of slower, more traditional ways of life as outdated and backward, the future success of initiatives to slow down everyday life is by no means guaranteed.

The threat of accelerating climate change has drawn renewed attention to socially and ecologically unsustainable everyday practices such as long-distance, car-dependent commuting. Proposed solutions include calls for a radical transformation of how people view and use time, focusing in particular on reductions in working hours (and resulting spending power) and an expansion of unpaid activities that enhance people's relationships and well-being. Recent prominent proposals for an overall reduction in working hours (Schor, 2010) or a 21-hour working week (Coote et al., 2010) reflect this debate.

Proposals for a radical reduction in working hours cannot be viewed in isolation from related debates on the causes and consequences of (un)sustainable consumption. The question as to whether current consumption levels need to be drastically reduced (=degrowth) or whether a change in the quality of consumption (='greening' of existing practices) is sufficient to reign in socially and ecologically harmful forms of (over) consumption remains a key focal point of sustainable consumption debates. Many contributors to this expanding field have acknowledged the inherently social nature of everyday practices such as cooking, eating, and traveling and have connected consumption to socially negotiated notions of 'the good life' (e.g., Jackson, 2009; Shove, 2010; Davies et al., 2014). Others have cautioned against business-as-usual approaches that promise continued economic growth through a change in the quality of consumption. For them, calls to consume differently (rather than not to consume at all) simply prolong rather than fundamentally challenge the dominant growth-based economic system that threatens the social and biophysical foundations of human life (e.g., Latouche, 2010; Murphy and Kirby, 2013).

Recent critical debates regarding the relationship between consumption, development, and quality of life have also provided opportunities for exploring more or less radical alternatives to current growth-based models of production and consumption and associated views and uses of time. For example, some authors have put forward very convincing arguments, complemented by detailed empirical evidence, that consumption beyond a certain level does not significantly enhance people's quality of life (e.g., Jackson, 2009; Wilkinson and Pickett, 2010). Yet others see the current economic crisis in Europe and beyond as an opportunity to "imagine a very different future and establish an 'ecological or ethical socialist model,' with less focus on (quantitative) growth and consumption, and more focus on (qualitative) living well" (Murphy and Kirby, 2013). Interestingly, many of these calls for a greater focus on quality of life imply a more or less radical transformation of people's time culture, although this is not always made explicit by their advocates. Moreover, there is currently little evidence of a fundamental shift in thinking away from quantitative growth and toward qualitative changes in how societies organize themselves and their economies, including people's time-related views and practices. Instead, many countries in the European Union and beyond have concentrated on efforts to return to pre-recession levels of economic growth. Existing proposals for a more sustainable, quality of life focused use of time as part of a concerted challenge to the prevailing growth logic have not yet received the level of recognition they require to be successfully implemented on a larger scale.

Arguments that a deceleration of everyday life forms an essential part of any sustainability transition have also not been without their critics. At the conceptual level, Shove et al. (2009) challenge simplistic fast/slow dichotomies of time use and narrow causal models that attribute the modern 'time problem' solely to the work-and-spend logic of consumerism. They also critique concepts of time as a resource that can be easily quantified, stored, released, and used up. Instead, these authors argue for a more nuanced analysis of the complexity of everyday temporal practices, emphasizing the creative (re) production and consumption of multiple temporalities (Shove et al., (2009), p. 3).

Jalas (2002) draws attention to the possibility of time rebound effects whereby time that has been freed up due to efficiency gains may be filled with new and potentially more resource-intensive activities. For example, a reduction in working hours would most likely increase rather than reduce people's ecological footprint if time poverty (rather than financial constraints) had prevented their engagement in resource-intensive acts of consumption in the past. This clearly challenges more simplistic 'time equals money' views of human time use and associated resource consumption. Whether or not a reduction in working hours will increase or reduce people's ecological footprint thus depends on what types of activities people adopt *in lieu* of these working hours.

Conclusions and Outlook

Changes in the temporal fabric of society, including the acceleration of production and consumption, have intensified the use of natural resources to an unsustainable level. This has sparked debates about an alternative future for people and the planet that reconciles economic development with social justice and ecological integrity and that promotes the adoption of new temporal practices that work *with* nature rather than against it. Arguments for a radical shift in time culture away from short-termism and toward a long-term perspective that recognizes the needs of future generations are central to recent sustainability debates. Similarly, calls abound for a shift in the allocation of temporal resources, including a reduction in working hours, an improved work-life balance, and a general deceleration of everyday life. However, many of these proposals tend to remain too focused on measurable aspects of (working) time. While these time-related changes have the potential to transform how people live and work, they are unlikely to suffice in reversing current trends in material consumption, emissions, and waste, partly because of emerging time rebound effects. What is instead required is a critical examination of qualitative changes in people's time culture that has accompanied the transition to late capitalism, culminating in unsustainably high levels of consumption. Their potential to systematically undermine alternative sociocultural practices that adopt a long-term, intergenerational perspective and that are less resource-intensive as a result deserves particular attention. Overall, the necessity to take time seriously presents a major challenge for future sustainability research and practice, with promising starting points already in place.

See also: Degrowth; Ecological Economics; Energy and Society; Environmental Movements; Environmental Policy: Protection and Regulation; Leisure and Cultural Consumption: The European Perspective; Social Ecology; Sustainable Lifestyles; Sustainable Practices; Sustainable Production and Consumption; Time Use Research Methods; Time Use Research: Recent Developments.

Bibliography

Adam, B., 1996. Re-vision: the centrality of time for an ecological social science perspective. In: Lash, S., Szerszynski, B., Wynne, B. (Eds.), Risk, Environment and Modernity: Towards a New Ecology. Sage, London, pp. 84–103.

Adam, B., 1998. Timescapes of Modernity. Routledge, London.

Adam, B., Geißler, K., Held, M., 1997. Die Non-stop Gesellschaft und ihr Preis: Vom Zeitmißbrauch zur Zeitkultur. Hirzel, Stuttgart.

Boulin, J., Mückenberger, U., 1999. Times in the City and Quality of Life: BEST European Studies on Time. European Foundation for the Improvement of Living and Working Conditions, Dublin. http://www.eurofound.europa.eu/pubdocs/1999/40/en/1/ef9940en.pdf (accessed 02.11.12.).

Coote, A., Franklin, J., Simms, A., 2010. 21 Hours: Why a Shorter Working Week Can Help Us All to Flourish in the 21st Century. New Economics Foundation, London.

Davies, A., Fahy, F., Rau, H. (Eds.), 2014. Challenging Consumption: Pathways to a More Sustainable Future. Routledge, London.

Druckman, A., Buck, I., Hayward, B., Jackson, T., 2012. Time, gender and carbon: a study of the carbon implications of British adults' use of time. Ecological Economics 84, 153–163.

Elias, N., 1984. Über die Zeit. Suhrkamp, Frankfurt am Main.

Fischer-Kowalski, M., Schaffartzik, A., 2008. Arbeit, gesellschaftlicher Stoffwechsel und nachhaltige Entwicklung. In: Füllsack, M. (Ed.), Verwerfungen moderner Arbeit. Zum Formwandel des Produktiven. Bielefeld, Transcript, pp. 65–82.

Geiβler, K., 2002. A culture of temporal diversity. Time and Society 11 (1), 121–140.

Geiβler, K., Kümmerer, K., Sabelis, I. (Eds.), 2006. Zeitvielfalt: Wider das Diktat der Uhr. Hirzel, Stuttgart.

Grober, U., 2002. Konstruktives braucht Zeit: Über die langsame Entdeckung der Nachhaltigkeit. Politik und Zeitgeschichte B31-32, 3–7.

Higginson, S., Thomson, M., Bhamra, T., 2013. "For the times they are a-changin": the impact of shifting energy-use practices in time and space. Local Environment 19 (5), 520–538.

Hueting, R., Reijnders, L., 1998. Sustainability is an objective concept. Ecological Economics 27 (2), 139–147.

Jackson, T., 2009. Prosperity without Growth? Sustainable Development Commission, London.

Jalas, M., 2002. A time use perspective on the materials intensity of consumption. Ecological Economics 41, 109–123.

Ladner, S., 2009. 'Agency time': a case study of the postindustrial timescape and its impact on the domestic sphere. Time and Society 18 (2/3), 284–305.

Läpple, D., Mückenberger, U., Oßenbrügge, J. (Eds.), 2010. Zeiten und Räume der Stadt. Theorie und Praxis. Oplanden, Barbara Budrich.

Latouche, S., 2010. Farewell to Growth. Wiley, London.

Mückenberger, U., 2010. Was ist Zeitpolitik? Und welche Politik brauchen wir, damit wir mehr Zeit haben? In: Bechmann, U., Friedl, C. (Eds.), Montagsakademie ZEIT. Universitätsverlag, Graz, pp. 151–172.

Mückenberger, U., 2011. Local time policies in europe. Time and Society 20 (2), 241–273.

Murphy, M.P., Kirby, P., August 2013. The solution: degrow – emphasise the collective, the simple and the environmental. Village: Ireland's Political Magazine (online) 6. http://www.villagemagazine.ie/index.php/2013/08/the-solution/ (accessed 07.08.13.).

Pink, S., Servon, L.J., 2013. Sensory global towns: an experiential approach to the growth of the slow city movement. Environment and Planning A 45 (2), 451–466.

Rau, H., Edmondson, R., 2013. Time and sustainability. In: Fahy, F., Rau, H. (Eds.), Methods of Sustainability Research in the Social Sciences. Sage, London, pp. 173–190.

Rau, H., 2004. Time Perspectives and Temporal Practices: A Cross-Cultural, Comparative Study of Time Cultures in Ireland and Germany (PhD thesis). National University of Ireland, Galway.

Rau, H., 2014. Part II: Moving – section introduction and summary. In: Davies, A., Fahy, F., Rau, H. (Eds.), Challenging Consumption: Pathways to a More Sustainable Future. Routledge, London, pp. 57–61, 99–101.

Rau, H., Davies, A., Fahy, F., 2014. Conclusion: moving on: promising pathways to a more sustainable future. In: Davies, A., Fahy, F., Rau, H. (Eds.), Challenging Consumption: Pathways to a More Sustainable Future. Routledge, London, pp. 187–205.

Rosa, H., 2005. Beschleunigung. Die Veränderungen der Zeitstrukturen in der Moderne. Suhrkamp, Frankfurt am Main.

Schäfer, M., Jaeger-Erben, M., Bamberg, S., 2012. Life events as windows of opportunity for changing towards sustainable consumption patterns? Journal of Consumer Policy 35, 65–84.

Schor, J., 2010. Plenitude: The New Economics of True Wealth. Penguin Press, New York.

Shove, E., Trentmann, F., Wilk, R. (Eds.), 2009. Time, Consumption and Everyday Life: Practice, Materiality and Culture. Berg, Oxford.

Shove, E., 2010. Beyond the ABC: Climate change policy and theories of social change. Environment and Planning A 42, 1273–1285.

Thompson, E.P., 1967. Time, work-discipline and industrial capitalism. Past and Present 38, 56–97.

Trentmann, F., 2009. Disruption is normal: blackouts, breakdowns and the elasticity of everyday life. In: Shove, E., Trentmann, F., Wilk, R. (Eds.), Time, Consumption and Everyday Life: Practice, Materiality and Culture. Berg, Oxford, pp. 67–84.

Wilkinson, R., Pickett, K., 2010. The Spirit Level: Why Equality Is Better for Everyone. Penguin, London.

Zerubavel, E., 1982. The standardization of time: a sociohistorical perspective. American Journal of Sociology 88 (1), 1–23.

Zerubavel, E., 1997. Social Mindscape: An Invitation to Cognitive Sociology. Harvard University Press, Cambridge/London.

Relevant Websites

www.consensus.ie.
http://www.zeitverein.com/framesets/fs_zeitverein.html.
http://www.cittaslow.org/.
http://www.timeuse.org/mtus/.

Time Use Research Methods

Jonathan Gershuny, Nuffield College, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Individuals' time use is a key social indicator, directly reflecting both production and consumption as well as embodied capital investment. Various methods are used to measure time use, of which the time-use diary sample study is the most widely used. Time-diary survey methods vary widely both in design and mode of administration, though all instruments have in common a requirement of complete and continuous registration of activities of sample members on designated days. This article discusses the range of design options, and as well as major studies conducted over the last 50 years.

Why Study Time Allocation?

We all have 1440 min in our days. While we can buy goods, and the services of other people, we cannot directly purchase more minutes for ourselves. However, *as* we buy goods and services, and sell our own labor, we are trading our own time, and other people's. Furthermore our position in this market for time reflects aspects of our own past histories – how much time we spent in interaction with pa and siblings in our households of origin, how much time in schools and training and in various sorts of work and leisure activities. The work experience accumulates through the life-course to form that aspect of human capital that determines prospective earnings rates. Those whose hourly wage is greater than the average may wish to sell a relatively large part of their day as paid labor, since their earnings can, in effect, buy a still larger amount of time indirectly, as embodied in goods and services.

So the apparently equally distributed resource, is, paradoxically, both a source of, and a fundamental expression of, social differentiation. How we spent our time in the past accumulates to provide the embodied social, cultural and economic capitals (or capabilities) that help to determine our current time allocation, and not just between work and leisure, but also, through mechanisms set out by Becker (1965) and Bourdieu (1985), among different sorts of paid work and of leisure consumption. And employment-related human (or economic) capital is in effect a mechanism for redistributing time, giving individuals differential levels of effective control over time, either below or above the 1440 min of the day.

Those whose expected earnings rates are below the a society's average wage (and who have no financial capital) *must necessarily* exchange each hour of their paid work for goods and services that embody on average *less than* an hour of the work of others. Those with higher wage rates may exchange multiple hours of embodied labor for each hour of their own work. Once we subtract paid work time from individuals' days, and add back the time embodied in purchased goods and services, the poorer members of a modern developed society (with lower levels of embodied human capital) might effectively control only 15 or fewer hours each day, while the richer members (those with higher human capital) might control 50 h or more per day.

Yet despite this crucial role in the emergence of social structure, time allocation evidence does not at present form part of the core research material of any social science discipline. Why?

One contributing explanation is the paucity of adequate individual-level time-use data. We have limited knowledge of our own time use. Like fish in water, we live in time; we are *in* it, but not aware *of* it. As a consequence, we are not well equipped to mark its passage or duration and in any case, the social scientist is interested not in time itself but in the particular activities to which successive periods are devoted. So, for example, if we ask survey respondents directly how much time they devote to a particular activity over a given period, the answers we get reflect little more than conventional expectations ("I work 40 h per week" our respondent may tell us, because her employment contract states this – but her *actual* hours worked are only accurately known if there is some externally imposed mechanism for registering this). For many categories of activity (housework, leisure activities, travel, etc.) respondents are unable to provide any straightforward answers at all (unless they report complete nonparticipation). There is no particular reason why we should need to know such an abstraction as the amount of time devoted to each element in a general classification of activities in the course of daily living; so we cannot always give accurate answers to researchers when they ask us about them.

Methodological Alternatives

There are four distinct approaches to the measurement of time use in a population: 'stylized estimates' as part of questionnaire enquiries, *observational methods*, using *administrative records*, and *diary methods*.

Despite the fact that people are often unaware of their time allocation, much of what currently passes for scientific knowledge of time allocation nevertheless relies on direct questions to survey respondents. In particular, the standard models for the Labor Force Surveys collected by most developed countries routinely ask for estimates of amounts of time, on average, or over a typical week, or during the previous week, spent working in paid employment. Other surveys ask similar questions about time devoted to, for example, childcare and domestic chores. These produce 'stylized estimate' measures of time use.

This approach is the simplest and cheapest way to collect information about how people use their time, but it has

significant drawbacks. Research on the paid work time estimates deriving from this methodology suggests substantial and systematic biases (e.g., Robinson et al. 2011). Those who claim to work in excess of 40 h per week progressively overstate the length of their work-week. The reason for this bias seems to be that as work time increases it becomes more difficult to accommodate other necessary or desirable activities outside scheduled work times. The longer the actual work time, the more likely it becomes that activities that might otherwise be located in 'free' time (e.g., medical visits, shopping, sports participation) are inserted at times that the respondent normally expects to be at work, and hence the greater the proportional overestimate. These exceptional events between conventional work-start and -finish times are often forgotten, whereas the exceptional events that lengthen the working week (e.g., overtime) are more likely to be remembered in the context of a question about paid work time.

Ethnographic research has often involved direct observation of people and the recording of their activities by researchers. New technology now extends the role of direct observation, including such unobtrusive devices as individual location monitoring systems and closed-circuit or internet camera systems.

Observational methods have certain advantages, in that the researcher can ensure that respondents do not forget or suppress particular activities of interest. In the case of new technology, the option of reviewing audio tapes and video images can ensure accurate recording of people's activities; ambiguities can be reassessed until a maximally precise determination can be reached. Such methods may be useful for analyzing small samples of special groups (e.g., residents of a retirement home or a small rural community) or activities in a specific location (such as the numbers of people who pass by a shop display and react or who engage in various activities in a town center).

On the other hand, direct observation studies are costly, they are generally labor-intensive, and they may compromise personal privacy. Moreover, people can be expected to censor or change their normal patterns when they are aware of continuous monitoring (the more so if they are conscious of an observer of differing nationality, class, sex, ethnicity, age group, or religion). However, continuous monitoring of behavior via continuous recording of position using GPS, physical activity using accelerometers, and individuals' visual fields via continuously recording cameras worn on clothing or as headgear, are gaining acceptability. And some extension of these methods, perhaps in conjunction with 'beeper' techniques (more formally 'experience sample methods' or ESM, used for example in a 2010–11 add-on module to the annual American Time Use Survey) allow respondents to record affective responses to current activities on random occasions during the day.

But, for cost and other reasons, notably the low and highly selective samples resulting from the intrusive and demanding nature of these methods, these approaches are not at present widely used to provide representative evidence of societies' time-use patterns.

Less direct observational methods ('unobtrusive measures') are becoming increasingly prevalent. These include programs tracking the use made of computers at work and school (measuring the time people spend word processing, performing data analysis, checking e-mail, browsing the web, and so forth), phone call and retail transaction records and the like, recording what people do at particular times and in particular places.

Administrative records, some of which cover considerable periods of time, chiefly concern the official records of hours during which people are in engaged within institutions. Most typically, this involves employers' records of employees' work hours, but it could in principle now involve any of those recreational facilities, hospitals, prisons that use electronic methods to monitor individuals' entries and exits. Administrative records typically provide large numbers of cases for analysis. However, they only cover the population which has contact with the institutions keeping the records, and do not indicate if people have simultaneous contact with related institutions at the same time (e.g., working in two jobs). They tend to contain a bias against short spells, and the contextual information about personal and socioeconomic characteristics typically needed for academic or policy research is sparse.

Many of the problems with the foregoing methods are solved if respondents are asked to provide a complete account of the sequence of events, during one or more days, within a framework of clock-times. This 'diary' approach, unlike the stylized estimate method, involves a narrative skill that is often called for in the course of daily life. Diary methods are by far the most widely used techniques for the estimation of time-use patterns: There are at least 500 such studies, from at least 80 countries. Approximately half of these countries can provide at least one full national random sample of diary-based time-use evidence.

Diary Methods

The diary method provides a number of clear advantages, which stem from the fact that the narratives mirror, in a straightforward way, the structure of people's daily activities. Diaries require the respondent to engage in an active process of recall, in which they reconstruct the successive events of the day, so that the particular occurrences (such as a trip to the doctor that disturbs the normal passage of the working day) emerge naturally as the respondent recollects 'what happened next.' Diary accounts are complete and continuous, so any respondent intending to suppress some activity must make a conscious effort to substitute some other activity. This will also tend to limit exaggeration of the duration of particular activities, since the potential for double counting (e.g., of a period as both work and leisure) that is entirely possible using the stylized estimates, is precluded in diary accounts. There are other advantages for the analyst in the diary-reporting format. Increasingly researchers are interested in investigating the activity sequences themselves, or the times of the day at which activities occur; simultaneous activities can be properly investigated, and if multiple diaries are collected within a single household, researchers can use them to investigate patterns of copresence, interdependence and cooperation.

The time-diary method itself is nevertheless not unproblematic. In particular, it produces some "under-reporting of socially undesirable, embarrassing, and sensitive" behaviors,

including personal hygiene, sexual acts, and bathing (Robinson and Godbey, 1997: 115). Techniques for persuading people to account completely for their days need some refinement.

Diary Data Collection

There is wide variation in the design of time-use diaries. The summary results seem quite stable with respect to the different methods, and comparisons of different methodologies do not suggest any unequivocally dominant methodological choices (although Harvey, 1999 provides general purpose recommendations). Choices are problematic (cost issues apart) because particular design characteristics affect respondent burden and hence response rates. There is a straightforward trade-off between a 'heavy' design, which yields a fuller picture of individuals' activities but relatively low response, or a 'light' design providing a less detailed picture, but with a higher response rate.

For example, diaries may cover longer or shorter periods. The problem, which is unusual in survey research terms insofar as most personal characteristics are fixed at any historical point, is that individuals' activities differ from day to day. Suppose all respondents do a weekly clothes wash on one randomly chosen day of the week: a single day diary will suggest that one-seventh of the sample spend a lot of time in laundry and six-sevenths do none, where the reality is that all do, on average, a small amount. A short diary gives a spurious impression of interpersonal variability which is in fact wholly intrapersonal. A week-long or even multiple-day diary would reveal the real situation, but is of course very demanding for respondents.

There are two alternative modes of diary administration, conventionally referred to as 'yesterday' and 'tomorrow' studies, the former involving an interviewer-led and recorded process, the latter, a self-completion instrument. The 'yesterday' diary normally does cover the previous day, as recall of events decays quite rapidly, but it is practicable to collect useable information after a gap of 2 or even 3 days, and difficulties in recruiting respondents over the weekend may mean that this is necessary so as to ensure even coverage of the days of the week. The interviewer leaves the 'tomorrow' self-completion diary with the respondent with the instruction to make records as frequently as possible through the diary period so as to maximize the recall quality. Tomorrow diaries have normally covered either 1 or 2 days, but experience in the Netherlands and the UK shows that periods up to a full week are possible, though with reduced response rates.

A variety of interview technologies have been employed for yesterday diaries. In addition to conventional pen and paper personal interviews, there are examples of computer-aided personal interviews (CAPI) and computer aided telephone interview (CATI) methods. Studies sometimes use a combination of yesterday and tomorrow diaries with the self-completion instrument being posted back by the respondent. In this case the tomorrow diary shows lower and somewhat biased responses. For tomorrow diaries, response is substantially improved if the interviewer returns to collect it.

A second substantial design issue concerns the temporal framework within which the activities are recorded. A key decision here is whether to adopt an 'event' basis of recording: 'When did you wake up?,' 'What did you do first?,' 'When did you stop?,' 'What did you do next?' This style is appropriate for a 'yesterday' interview, but not for a tomorrow diary. For self-recording, it is more appropriate to use a schedule with regular time slots, of 5–15 min (although slots of up to 30 min have been used, and the length of the slots may vary through the day and night). With this approach, respondents indicate the duration of each activity on the schedule using lines, arrows or ditto marks.

Activities may be reported using either a 'closed' vocabulary, in which a predetermined set of activity codes is provided to cover the full range of possibilities, or an 'open' approach in which respondents use their own words to describe the activities, these being coded subsequently. Respondents often report that they find the 'closed' approach constraining, however well-designed the categories, and the choice of categories may, to a limited degree, influence the level of reporting (e.g., a more detailed classification of travel types leads to a small increase in estimates of travel time). But the open approach has symmetrical shortcomings: respondents report at varying levels of detail, which are difficult to disentangle from genuine interpersonal differences.

For much of the day it is difficult to classify the current state as a single activity. For example, 'writing at the computer' may be accompanied by 'drinking coffee' or 'talking to my child.' The normal practice is to ask respondents to record activities in a hierarchical manner: What are/were you doing? Were you doing anything else at the same time? Some studies ask respondents to record all activities in a nonhierarchical manner, but respondents tend to prefer the opportunity to use a narrative mode ("First I did this, and then I did that …") which provides a hierarchy of primary and secondary activities. The hierarchy also simplifies the process of analysis, though there are ways of analyzing the data that are nonhierarchical but still conserve the 1440 min of the day.

Other ancillary information included within the rows or columns of diary forms include variables such as: 'Who else is present/engaged in the activity?' and 'Where are you?' (the latter often requesting information on modes of transport). National income accountants' interest in nonmonetized economic activity has led to attempts to acquire 'Who for?' information, but these have not so far been successful. Specialized diary studies also include interpretative or affective information (e.g., 'Is this work?' or 'How much did you enjoy this?') attached to each event or activity.

Beyond the design of the diary instrument, there are also various specialized sampling issues. Seasonal effects may be of importance; representative samples of time use require an even coverage across the year. There are serious practical difficulties but most national-scale samples attempt multiple sweeps at different parts of the year. The most developed approach to temporal sample design is found in Poland, whose 1970s and 1980s studies achieved approximately even numbers of diaries for each *day* of the year. Samples that include simultaneous diaries for all household members greatly extend the analytical usefulness of the evidence. Conventional 'tomorrow' studies include children from the ages of eight and Italian national studies have, by a combination of simplified forms for younger children, and parental proxies for preliterate children, achieved total household coverage.

Time-Diary Research Studies

Large-scale national random sample studies have been conducted in each of three or more decades in more than 20 countries. The first cross-national comparative time-diary study was supported by United Nations Educational, Scientific and Cultural Organization (UNESCO) in 1965 (Szalai, 1972). Gershuny (2003) reports on diary studies from showing 16 countries' historically changing time-use patterns. Seventeen countries participated in Eurostat's 1996–2004 European Harmonised Time Use Study (HETUS) project, and a new tranche of HETUS studies (covering 2008–15) is currently in the field. Results from the Multinational Time Use Study, providing evidence of change in time use over the period 1961–2010 from 23 countries, are reported by Fisher and Robinson (2011). And an increasing number of Pacific Rim, Developing and Less Developed Countries are now collecting time-diary studies.

See also: Economic Sociology; Everyday Life, Anthropology of; Gender, Economics of; Household Economics; Informal Sector; Informal and Underground Economics; Shadow Work: Measurement; Time Use Research: Recent Developments; Time Use and Gender; Time Use and Resource Consumption; Time, Anthropology of.

Bibliography

Becker, Gary, 1965. A theory of the allocation of time. Economic Journal 80, 493–517.

Bourdieu, Pierre, 1985. Distinction. RKP, London.

Fisher, Kimberly, Robinson, John, 2011. Daily life in 23 countries. Social Indicators Research 101 (2), 295–304.

Gershuny, Jonathan, 2003. Changing Times, second ed. Oxford University Press.

Harvey, Andrew S., 1999. Guidelines. In: Pentland, W.E., Harvey, A.S., Lawton, M.P., McColl, M.A. (Eds.), Time Use Research in the Social Sciences. Kluwer, New York, pp. 19–45.

Robinson, John P., Godbey, Geoffrey, 1997. Time for Life: The Surprising Ways Americans Use Their Time. Pennsylvania State University, Ann Arbor, MI.

Robinson, John, Martin, Steve, Glorieux, Ignace, Mirren, Joeri, 2011. The overestimated workweek revisited. Monthly Labor Review xx, 1–13.

Szalai, Alexander (Ed.), 1972. The Use of Time. Mouton, The Hague.

Time Use Research: Recent Developments

John P Robinson, University of Maryland, College Park, MD, USA
Teresa A Harms, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Academic and government researchers from most industrialized countries and an increasing number of developing countries are collecting and analyzing full 24-h time-diary data. These diary data provide a comprehensive and accurate basis for generating national accounts of time spent working, sleeping, and spending free time, and how daily life varies across demographic groups, across decades and across countries. Diaries can also provide data on multitasking, interacting with others, and the location and timing of daily activities in the American Time Use Survey and in more than 25 other (mainly European) counties. These diary data are being standardized and archived via the Multinational Time Use Survey at the University of Oxford.

Considerable progress has been made in producing tables of comparative time use in and across these countries and in identifying and visualizing patterns of multinational time use using Multidimensional Scaling. Major attention has been focused on uses of time in families, with special attention to the implications for our understanding of the division of paid and unpaid work within households. Given the richness of diary data available, less attention is now given to trends from time-estimate questions, although brief mention is made here to useful estimates from the US General Social Survey on important free-time activities and subjective time measures.

In order to provide greater insight into the sociological meaning of these activities for national accounts, supplemental data have also been collected on subjective and qualitative characteristics of diary activities, such as enjoyment, stress, and outcomes. Both in the United States and other countries, diarists generally rate activities related to social life, religion, and interactive childcare higher in enjoyment, and activities related to household upkeep and health care at the bottom. In more recent surveys, paid work activities and TV have received less positive ratings. Furthermore, recent US surveys indicate no increase in time pressure, a subjective factor that has been consistently related to lower life satisfaction. Fuller coverage of these ratings is covered in a separate entry in this volume titled Subjective Time.

Introduction

Time figures prominently in the structure and quality of people's lives. Sociologically, time-use research focuses on the lives of individuals – usually in the form of self-reports from nationally representative samples – a simple documentation and accounting of the hours and minutes spent on various daily activities, like paid work, TV viewing, and time alone. Time use has become prominent as more government agencies conduct time-diary surveys, particularly with the creation of the American Time Use Survey (ATUS), described below.

Impressive advances have been made as well in measuring *subjective time* (covered in more detail in a separate article) using surveys in questions that ask people how they perceive their time use and how they feel about and value the way they spend their time. Simple diary figures by themselves generally do not have immediate quality-of-life (QOL) implications, in that survey analysts cannot assume that a person's QOL would rise or fall as they spent more time eating or traveling. In contrast, analysts generally seem more ready to draw such QOL implications when parents spend more time with their children or when people have more available free time.

Infrastructural Developments

The main two diary data sources to available researchers studying time use are the Multinational Time Use Survey (MTUS) for multinational data and the ATUS for US data. Both are multiyear and national in scope, the MTUS covering from the 1960s to today and the ATUS continually from 2003 to the present.

The MTUS Data Archive

A major stumbling block to comprehensive scientific study of time use has been the wide variety of methods and field procedures used by researchers in different disciplines and different countries to address diverse social issues. This Tower-of-Babel problem had been largely sidestepped in the 1965 pioneering multinational data collection – the multinational time-budget study, as initiated and supervised by the Hungarian mathematician and sociologist, Alexander Szalai (1972). Based on his extensive experience of trying to make sense of the fragmented series of early 'pilot' and historical studies in Eastern Europe, Szalai organized and enforced a strict set of methodological guidelines and procedures in this monumental 12-country effort.

In the ensuing 30 years, diary researchers from a number of countries and disciplines have organized a series of individual advances in diary methods and procedures, but their successes were undermined by a failure to achieve comparability across countries and even studies within countries. This disparity was formally recognized and addressed by the efforts of EUROSTAT's Harmonised Time Use Study (HETUS) in the 1990s. However, there remained the tedious tasks of standardizing, documenting, and disseminating these data.

The Oxford Centre for Time Use Research (CTUR) has created a standardized data archive from both the HETUS, and other disparate individual country diary studies (Fisher and Gershuny, 2014). Their harmonized data files are now available to social science scholars around the world. The MTUS Web site contains complete data files for a rapidly growing group of studies from more than 20 countries (some dating back to the 1960s). It also houses the American Heritage Time Use Study (AHTUS), a complete standardized formatting of all US diary studies dating back to 1965, with diary codes and background variables harmonized for comparability across time and including the ambitious diary collection in the ATUS described next. CTUR also produces a number of valuable resources including newsletters identifying the latest studies, publications, and conferences of interest.

An illustration of the value of the MTUS archive is shown in Table 1, as compiled by Fisher and Robinson (2011). It summarizes, in a simple single comparative table, the contrasting time use patterns across 23 countries around the world to identify the major similarities and differences in daily life between 1998 and 2005. An example of how the complex set of relations in Table 1 can be simplified even further and distilled into a plausible two-dimensional map, is shown in Figure 1 below.

CTUR's role in the MTUS is now supported over the period 2014–19 by a grant from the UK Economic and Social Research Council. This funding will allow the Center to update the UK studies of diary trends; innovate new diary methods involving the Web and smart phone applications, as well as deploy monitoring devices such as GPS, wearable cameras, and accelerometers; expand studies of subjective time; engage in collaborative diary studies with other HETUS countries; and recruit and train new generations of time use scholars from other countries.

The American Time Use Study

The first US national diary study was conducted in 1965 using personal interviewers and was then replicated every decade afterward, in 1975, 1985, 1993–95, and 1998–2001, mainly by telephone. All diaries were collected by academic survey firms, first at the University of Michigan and since 1985 at the University of Maryland. Most of the surveys achieved response rates over 65% and used an activity classification derived from that used in Szalai's pioneering 1965 study.

Since 2003, the US Census Bureau has been collecting ATUS 'yesterday' diary data continuously on behalf of the Bureau of Labor Statistics (BLS). With samples of around 12 000 respondents per year, it now has a cumulative sample base of more than 140 000 respondents aged 15 years and older. Respondents are selected from the US Census Bureau's Current Population Survey (CPS) and are interviewed by telephone (after the 8th CPS wave). The ATUS maintains a response rate of about 55% of CPS respondents. Despite this relatively low-response rate for a Census Bureau survey, Abraham et al. (2006) present considerable evidence that the ATUS sample remains representative of the US population.

The ATUS has an unusually long list of diary activity categories, but more than 400 entries are not directly comparable to the Szalai-derived schema used in previous US academic diary surveys. The detailed set of ATUS activity categories are described at www.bls/tus.gov and the data are archived for direct access at www.atus-x. A condensed list of 25 summary activities is shown in Table 2. The ATUS has also enhanced the interpretability the diary records with new modules of questions related to eating and health and subjective well-being, discussed in more detail in various sections below. In its AHTUS standardized archive, the CTUR has generated a single harmonized set of data files that merge the early surveys with the more recent ATUS file.

Table 2 presents a trend summary of ATUS results for each year in the decade since 2003 for the total US population aged 15+, as reported in BLS news releases (Robinson, 2013c). It can be seen that the activity totals are remarkably consistent across years. This is particularly surprising because in 2007, the US underwent its most serious economic recession since the 1930s. Yet, as Table 2 shows, there was only a 1 h (3%) overall decline in work-related hours in the ATUS across the decade (or 1.6 h from its high of 26.3 h in 2006). Table 2 also shows that the ATUS recorded a decrease of almost an hour per week in care given to nonhousehold members (which could signal a serious hole in the nation's social 'safety net'). thus decreases were offset mainly by a 1.5 h increase in sleep and 1.7 h increase in TV viewing, two activities that have been strongly linked to decreased work hours.

Unexpectedly, there was less than an hour of extra time spent with new information technologies (IT), a decade so obviously marked by their rapid diffusion. However, a major problem in measuring IT use in the ATUS data is that IT use for other noncommunication functions (like shopping and games) are counted as time for those activities, rather than for communication. Similarly, low weekly hours of IT use were found in time diaries from other Western countries (Robinson, 2011). Again, cross-national comparisons with US IT usage is complicated by the different activity coding schemes employed. However, the continuous annual nature of the ATUS data compares favorably with the 5–10-year intervals for the other national data series in the MTUS.

Visualization Maps of International Activity Differences

Time diaries contain a wealth of data on almost every facet of everyday life, and the detailed breakouts of more than the simple disaggregation of 400+ activities by demographic variables can generate thousands of relevant data tables, which may become unmanageable for time use researchers. Factor analysis and related tools are useful methods of data reduction in the social sciences more generally. But the single-day data collection period common in time-diary studies is too short to reveal meaningful within-country patterns of activity correlations for factor analytic purposes.

However, one of the notable methodological innovations that became available to social scientists during the 1960s was the multivariate data-reduction technique of Smallest Space Analysis (SSA) (Guttman, 1968; Kruskal, 1964). Also referred to as multidimensional scaling (MDS), it has become one of the standard analytic tools available in SPSS and other statistical packages (Bloombaum, 1970).

Table 1 Multinational average weekly hours across 23 countries for whole population, 18–64 h (from Fisher, K., Robinson, J., 2011. Daily life in 23 countries. Social Indicators Research 101, 295–304.)

	Oceania north/south America				*Central Europe*					*Northern Europe Nordic*			*Eastern Europe*						*South Europe/ Mediterranean*			*Asia*	
Population aged 18–64 h per week	*Australia*	*Brazil*	*Canada*	*The United States*	*Belgium*	*France*	*Germany*	*The Netherlands*	*UK*	*Finland*	*Norway*	*Sweden*	*Bulgaria*	*Estonia*	*Latvia*	*Lithuania*	*Poland*	*Sloven*	*Italy*	*Spain*	*Turkey*	*Japan*	*Korea*
Paid work away from home	26.1	25.8	28.7	28.6	18.8	22.1	20.4	18.7	23.0	22.2	24.5	26.7	23.7	27.1	29.3	24.9	20.1	23.6	23.6	24.6	20.8	32.2	26.0
Paid work at home	2.0	2.6	NA	1.6	1.1	1.3	1.2	1.1	2.1	2.1	1.2	1.2	0.2	1.5	2.6	5.6	3.5	1.1	0.5	0.7	NA	1.9	0.2
Study and job or skill training	0.7	2.1	1.0	1.1	2.0	1.9	1.6	1.6	0.9	1.8	1.4	1.4	0.6	1.1	1.9	2.0	2.0	1.5	1.1	2.0	2.8	4.1	2.0
Homework	0.9	1.2	1.2	0.9	1.2	1.2	0.7	0.9	0.4	0.7	0.8	0.7	0.5	0.5	0.7	0.7	1.3	1.6	1.4	1.2	NA	1.1	0.4
Commuting job/study/travel	3.2	5.8	3.0	2.5	3.2	2.8	3.0	2.8	3.2	2.5	3.2	2.9	2.8	3.3	4.3	3.4	2.9	2.9	3.5	3.6	NA	4.7	6.0
Cooking and food related	6.1	5.0	4.7	3.5	5.8	6.0	4.9	6.4	6.0	5.1	5.6	5.8	8.6	7.4	5.7	7.0	8.2	7.2	7.1	7.1	8.9	5.7	5.8
All other housework/repairs, garden	7.2	6.2	8.9	7.8	8.8	7.9	8.4	7.1	6.9	7.7	6.3	6.8	11.6	9.5	7.7	9.6	8.1	11.9	8.9	6.7	7.5	6.0	6.4
Shop, service, other domestic	4.6	3.2	3.6	3.7	4.2	4.7	4.8	4.3	4.8	3.9	4.1	4.2	1.8	3.5	2.6	2.0	2.9	2.5	3.6	4.3	1.6	3.0	3.9
Housework/personal care travel	2.7	1.4	2.4	4.3	1.5	0.1	2.5	2.1	2.2	1.8	1.6	2.2	1.9	2.1	2.5	2.2	2.1	1.9	1.8	1.3	NA	1.4	1.8
Physical medical childcare	2.2	1.4	1.9	2.0	1.3	1.9	1.3	2.1	2.3	1.9	2.3	2.0	1.1	0.0	1.1	1.4	1.8	1.4	1.5	2.1	3.4	0.8	1.5
Interactive other childcare	3.2	0.7	1.0	2.0	0.7	0.9	0.9	1.9	1.4	0.8	0.8	0.9	1.1	0.9	0.6	0.8	1.6	1.1	1.2	0.6	NA	1.1	1.3
Childcare-related travel	0.8	0.5	0.6	0.6	0.4	0.5	0.4	0.7	0.9	0.2	0.4	0.6	0.1	0.2	0.2	0.1	0.2	0.2	0.5	0.6	NA	0.4	0.8
Pet care	0.6	0.1	0.5	0.4	0.4	0.8	0.4	1.2	0.4	0.4	0.1	0.2	0.1	0.2	0.1	0.1	0.2	0.2	0.1	0.1	NA	0.1	NA
Organizational voluntary activity	1.3	3.2	3.2	3.6	0.8	1.3	2.2	3.2	1.5	2.0	1.5	1.6	1.1	1.8	1.4	1.9	2.9	1.4	1.8	1.4	4.4	0.6	1.7
Sleep and naps	58.7	56.4	58.4	58.6	58.3	61.1	57.3	59.5	58.8	59.0	56.2	56.4	62.4	59.5	59.9	58.9	58.7	58.1	57.3	59.0	59.3	53.7	54.4
Wash, dress, other personal care	6.2	7.2	4.5	5.6	5.1	5.0	6.1	6.1	5.4	4.9	5.5	5.3	4.4	6.2	4.7	6.4	6.1	4.7	7.1	5.6	18.8	7.7	8.8
Meals (at home/pack lunch)	6.7	7.1	6.8	5.8	11.0	12.4	10.9	9.0	8.8	8.4	8.5	10.3	12.6	8.4	9.8	10.0	10.4	9.6	11.7	11.3	NA	13.1	11.6
Walking including walking dogs	0.7	0.7	0.6	0.5	1.8	1.9	1.9	NA	1.9	2.0	1.8	2.0	2.1	1.6	1.9	1.2	2.1	2.5	2.3	3.9	NA	2.1	1.4
Sports and other exercises	1.9	0.9	2.5	1.6	1.5	1.1	1.6	1.8	1.3	2.3	2.1	2.0	0.9	1.1	1.5	1.1	1.1	1.6	1.3	1.3	0.8	0.0	1.4
Restaurants, bars, pub, cafe	1.3	2.6	2.3	1.8	1.5	3.2	0.8	1.9	1.1	0.7	0.9	0.4	1.8	0	0.5	0.1	0.2	0.6	1.5	0.9	NA	2.6	NA
Party visits socialize away	2.2	2.4	1.5	0.5	4.4	3.2	4.6	8.2	5.3	3.7	5.6	4.1	2.6	2.3	2.7	2.5	3.4	4.1	4.6	5.1	0.4	0.0	3.5
Party visits socialize at home	0.4	2.9	4.0	6.1	2.5	1.8	3.3	2.9	1.9	2.6	6.5	3.2	1.9	1.4	1.4	1.5	2.8	2.9	1.9	1.4	8.3	0.0	1.4
Leisure away from home	2.5	0.6	3.5	1.1	1.3	1.4	1.6	1.1	0.9	0.7	0.9	0.7	0.1	0.6	0.7	0.2	0.4	0.6	0.7	0.8	NA	0.0	2.7
Other travel	2.0	2.0	1.0	2.2	5.0	3.6	4.2	3.0	3.3	4.2	4.1	4.6	2.1	2.6	3.1	2.8	3.1	3.2	4.7	3.3	9.3	2.7	2.7
Relax do nothing	1.5	1.6	1.9	1.9	3.0	0.7	1.8	1.4	2.2	2.1	1.3	2.6	0.9	1.8	2.1	1.2	1.3	3.4	3.3	2.7	4.0	0.9	1.2
Line Computing/Internet games	0.4	0.5	1.7	1.2	2.8	0.6	2.0	1.8	1.2	0.9	1.3	1.4	0.1	0.4	0.5	0.8	1.1	0.7	0.7	1.1	NA	1.2	2.8
Television	12.3	14.3	13.6	15.4	15.4	13.2	12.1	2.1	15.6	14.7	12.6	11.9	16.6	15.4	13.8	15.3	15.3	13.2	10.6	12.0	13.8	13.9	14.6
Radio, ipod, other audio	2.3	0.4	2.1	0.7	0.5	0.4	0.6	4.0	0.7	0.9	0.7	0.5	0.5	0.7	0.5	0.6	0.9	0.6	0.4	0.4	0.5	0.9	0.2
Read	2.2	0.7	0.2	1.9	2.5	2.2	3.9	3.7	2.6	4.9	3.7	3.3	2.0	4.1	2.8	2.5	2.6	2.5	2.0	1.6	1.3	2.0	1.8
Other leisure and hobbies	4.6	1.3	2.2	0.3	1.1	3.0	2.2	1.4	0.7	1.8	2.1	1.6	1.3	1.3	1.0	0.9	0.5	1.0	1.0	1.1	2.1	3.5	1.9
Unrecorded time	0.5	5.8	0.2	0.1	0.1	NA	0.4	0.0	0.4	1.1	0.4	0.5	0.2	1.5	0.4	0.3	0.2	0.2	0.3	0.2	NA	0.6	NA
Free time	34.8	NA	37.3	36.3	43.4	36.3	41.0	39.3	39.0	42.6	44.0	38.8	33.4	34.8	32.9	31.0	35.0	37.1	35.3	35.8	NA	30.4	35.6
Total	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168	168

NA: no data available.

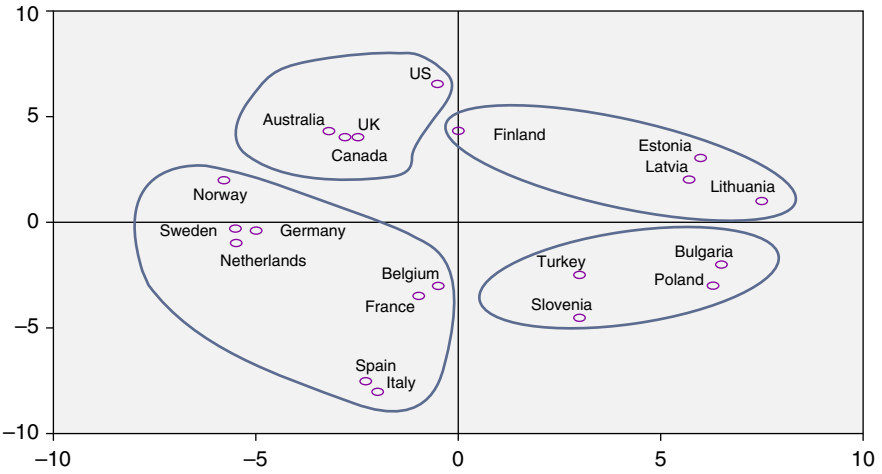

Figure 1 MDS Map for 20-country MTUS data.

Among the wide variety of other complex social phenomena that have been reduced to simpler spatial visualizations via SSA include occupations (Laumann and Guttman, 1966), work values (Elizur, 1984), and public perceptual maps of the world (Robinson and Hefner, 1968). The technique has more recently been applied to summarize results from the World Values Survey, using the respondent's agreement with various value statements in different countries. Inglehart and Welzel (2011) have generated a map that reduces these complex responses to a simple two-dimensional space that can be viewed at worldvaluessurvey.org.

Table 2 Hours per week on different activities (ATUS)

2003	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	
25.8	*25.6*	*25.9*	*26.3*	*26.0*	*26.1*	*24.7*	*24.5*	*25.0*	*24.7*	*Paid work*
23.3	23.3	23.5	23.8	24.3	23.7	22.3	22.0	22.4	22.3	Work
2.5	2.3	2.4	2.5	2.4	2.5	2.5	2.5	2.6	2.4	Commute+
3.3	*3.4*	*2.9*	*3.4*	*3.0*	*3.3*	*3.2*	*3.3*	*3.3*	*3.5*	*Education*
2.0	2.0	1.8	2.1	1.8	1.9	1.8	2.0	2.0	2.0	Classes
1.3	1.3	1.1	1.3	1.2	1.4	1.4	1.3	1.3	1.5	Homework+
24.4	*24.2*	*23.9*	*23.4*	*23.5*	*22.8*	*23.2*	*22.8*	*22.5*	*22.1*	*Family*
4.3	4.1	4.3	4.3	4.5	4.1	4.2	4.0	4.1	4.2	Housework
3.7	3.6	3.6	3.7	3.6	3.6	3.8	3.9	3.8	3.7	Cook
1.4	1.4	1.4	1.4	1.5	1.3	1.4	1.5	1.3	1.4	Lawn, etc.
0.9	1.0	1.1	0.9	1.0	0.9	0.9	0.8	0.8	0.9	Manage
2.5	2.5	2.5	2.2	2.3	2.2	2.3	2.3	2.3	2.0	Other housework
2.8	2.9	2.9	2.8	2.7	2.7	2.7	2.6	2.6	2.5	Shopping
2.9	2.9	2.7	2.9	2.7	2.7	2.7	2.7	2.5	2.5	Services
2.9	3.0	2.9	2.9	2.9	2.9	3.0	2.8	2.7	2.8	Household child care
1.0	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	Household adult care
2.0	1.9	1.7	1.5	1.4	1.6	1.5	1.5	1.5	1.3	Nonhousehold care
73.9	*73.9*	*74.7*	*74.5*	*74.0*	*74.3*	*74.7*	*75.5*	*75.1*	*75.4*	*Personal*
60.0	59.9	60.3	60.4	60.0	60.2	60.7	60.7	61.0	61.1	Sleep
8.5	8.6	8.7	8.6	8.7	8.6	8.5	8.8	8.7	8.8	Eat
5.4	5.5	5.7	5.5	5.3	5.5	5.5	5.6	5.5	5.5	Groom
40.7	*40.9*	*40.7*	*40.5*	*40.9*	*41.4*	*42.2*	*42.4*	*42.1*	42.3	*Free time*
1.0	0.8	0.9	0.8	1.1	1.0	1.1	1.1	1.1	1.1	Religion
1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	1.2	Club, org
5.5	5.3	5.3	5.3	5.1	5.0	4.9	4.9	4.9	5.2	Socialize
1.3	1.3	1.3	1.3	1.3	1.5	1.4	1.3	1.1	1.1	Telephone
2.0	2.0	2.0	2.0	2.2	2.1	2.2	2.2	2.1	2.2	Fitness
18.1	18.6	18.1	18.1	18.3	19.4	19.7	19.1	19.3	19.8	TV
10.2	10.4	10.6	10.3	10.1	9.8	9.9	10.2	10.2	10.0	Other free
1.3	1.1	1.3	1.4	1.5	1.4	1.7	2.3	2.1	1.7	Other
168	*168*	*168*	*168*	*168*	*168*	*168*	*168*	*168*	*168*	*Total*

MDS Applied to Szalai's (1965) Multinational Data

The multinational time-diary data from the pioneering time-diary study of Szalai (1972) became available soon after the SSA/MDS technique was developed. Converse (1972) applied MDS to a matrix of differences in the daily times spent by people in each country on a range of activities as shown in Table 1. His resulting two-dimensional MDS visualizations provided immediate and plausible insights into how similar life was in the different national settings involved in the study. Converse (1972) succinctly described the resulting MDS diagram as a picture that "bears a substantial resemblance to a map of the western world…It is remarkable that statistical compression of these raw data yields anything resembling a physical map" (p. 150).

Updated 1998–2005 MTUS Mappings

Figure 1 illustrates the application of MDS to the 1994–2005 MTUS time use data in Table 1. Following the same procedures that Converse (1972) employed, a half-matrix of Euclidian distances between the time use patterns in 23 countries between 1998 and 2005 (as calculated directly from Table 1) was used to generate a map representing the major differences between countries, again in a combination of cultural and geographic terms (Robinson and Gershuny, 2013). Again, Figure 1 clearly shows the influence of geography, although now more along language/culture lines than pure geographical proximity – mainly because, unlike the earlier Szalai data set, the widely dispersed Anglophone countries of UK, Canada, and Australia are now included in the survey.

The MDS has also been applied to examine parallel country differences among immigrant groups in the ATUS, which by 2010 included diaries from more than 10 000 immigrant respondents. Table 3 reports the time US immigrants from 26 categories of national origin devoted to various broad time use categories.

Initial analysis showed that the time use of these immigrants did not differ markedly from their US-born counterparts. However, there are still marked origin-country differences within Table 3. Immigrants from India, Africa, and certain Asian and Central American countries spend notably higher averages of time in paid work, for example, while immigrants from Canada and European countries spend less time in paid employment than other immigrants and native-born Americans.

Some of these country-of-origin differences are suggestive of more general regional or cultural patterns. Time spent in housework, for example, is higher among migrants from western European countries and lower from eastern European countries. Sleep time is higher among immigrants from Latin American countries compared with Japan and Korea. IT use is higher among Asian and eastern European immigrants and lower among those from Latin American countries. Clearly, it is almost impossible to comprehensively summarize these differences shown in the 664 entries in Table 3 in a few sentences or paragraphs. When subjected to an MDS analysis, however, Figure 2 again captures the distinct cultural and geographic patterns that separate the activity differences among immigrant groups in the United States.

Among US residents with different countries of origin, US-born citizens lie almost in the center of its two dimensional mapping, with Latin American origin countries generally to the 'south' of the United States, Asian countries generally to the 'east' and European countries to the 'north.' African and Indian immigrants cluster to the 'south' and 'east' of the United States, about equidistant from Asian and Latin countries. Russian and eastern European migrants appear closer to Asians than to other Europeans, and closer to south American (mainly from Brazil and Argentina), who are rather distant from other Latin countries. The closest lifestyles to the US-born citizens are from former residents of the Philippines and Japan.

Trends and Patterns in Specific Activities

Paid Work

The availability of long-term historical comparative time use materials produces dramatic research findings. Gershuny (2009) reported a 'Veblen in Reverse' trend. Members of the 'dominant' class at the end of the nineteenth century displayed their high status (and wealth) through conspicuous consumption during plentiful free time. By contrast, today's dominant classes spend more time at work, while the earlier 'working' class has fewer hours of paid work. Mestdag and Glorieux (2009) challenged the model of a new global '24/7-economy' by observing no greater proportions of Belgians in paid work across the day than in diary studies in the 1960s.

Unpaid Work

In an influential *Scientific American* article Vanek (1974) suggested that, despite the spread of labor-saving home appliances over previous decades, women's housework had basically not declined since the 1920s. It is now being widely reported that the ensuing 1970s gender revolution, and associated declining housework since these 1970s, has hit a wall, or at least 'stalled.' However, using recent data from the MTUS, Kan et al. (2011) show a striking downward trend in women's share of the unpaid work in the United States, and three other Anglophone countries (Canada, Australia, and the UK) suggesting a regular but slowing shift toward gender equality (Figure 3).

In the United States, for example, women's proportion of unpaid work declined from 77% in 1965 to 62% in 2010 – in both cases a little ahead of the other Anglophone countries. Contrary to the negative view of the US record on gender equality in the workplace, American men in fact carry out a larger share of the unpaid work than do men in other Anglophone states – 38% of the total in 2010 and showing continuing, if slowed, progress toward the 50% point of gender equality. Moreover, they present parallel graphs that show Anglophone countries performing somewhat better than in continental Europe, and *much better* than in southern European countries, but with countries from northern Europe Nordic doing the most (Swedish men at 41%).

This follows the earlier findings of Kan et al. (2011), Bianchi et al. (2006) and Fisher et al. (2006) historical gender convergence in the United States. At the same time, Bianchi et al. (2000) produced tables that showed a leveling off of that trend in the United States since the

Table 3 2003–10 ATUS activity times among US immigrant groups

	US BORN (74 036)	MEXICO (3439)	CENTRAL AMERICA (817)	CUBA (370)	HAITI/DR (381)	CARIBEAN (310)	SOUTH AMERICA (878)	PUERTO RICO (487)	CANADA (323)	UK (330)	WEST EUROPE (148)	ITALY (122)	MEDITER-RANEAN (83)	GER-MANY (444)	EAST EUROPE (204)	RUSSIA (146)	CHINA (252)	TAIWAN (153)	JAPAN (133)	KOREA (131)	VIETNAM (270)	PHILIPP-INES (161)	INDIA (546)	IRAN (79)	AFRICA (277)	OTHERS (730)
WORK	23.6	26.7	29.2	27.2	25.2	24.3	28.3	19.7	20.7	22.8	25.5	21.4	25.5	21.2	25.5	26.7	27.6	24.1	23.9	24.4	27.5	25.7	31.9	28.7	30.0	27.8
COMMUTE	1.9	2.7	3.7	2.6	3.1	2.9	3.0	2.0	1.4	1.8	1.8	2.0	2.1	1.8	2.2	2.6	2.9	2.6	2.4	2.4	2.2	2.2	2.9	2.6	3.0	2.7
EDUCA-TION	3.2	1.9	1.8	1.6	4.1	2.5	4.3	1.0	4.9	0.7	3.4	0.2	0.5	2.2	3.7	5.1	7.3	4.7	3.3	11.8	12.3	3.6	2.6	6.1	1.8	1.6
HOUSE-WORK	12.7	14.7	13.0	13.8	9.9	11.0	13.1	14.2	13.5	16.6	14.7	17.7	17.9	16.6	10.7	10.5	11.8	11.0	14.6	9.7	12.4	15.9	10.5	14.8	10.5	12.1
CHILD	4.5	6.0	3.5	3.4	4.4	6.3	5.1	4.2	5.0	4.1	4.0	3.0	5.5	4.5	4.1	6.3	6.9	5.2	6.6	4.6	6.0	5.0	7.2	2.8	6.6	5.6
SHOP	5.6	5.9	4.8	5.3	5.9	5.3	5.7	6.1	7.4	6.8	5.0	5.3	3.8	6.1	6.7	6.1	5.5	6.8	6.3	6.4	3.7	5.1	5.7	5.7	6.4	5.2
SLEEP	59.8	64.2	64.1	59.8	64.1	62.1	60.8	61.6	59.7	58.6	59.7	59.9	62.0	60.8	62.2	61.3	60.2	61.8	59.4	58.9	61.3	60.6	60.8	54.3	60.7	61.2
EAT	7.8	7.7	6.9	7.2	5.3	5.9	7.7	6.6	8.8	9.2	10.0	8.7	8.0	8.2	7.9	8.3	10.2	10.1	9.6	8.5	8.4	8.5	9.2	10.0	6.5	8.5
PERSONAL	5.5	4.5	4.5	5.3	6.2	7.5	5.3	6.4	5.7	4.4	4.0	4.9	4.3	5.9	4.8	4.4	5.1	4.6	5.4	6.0	3.8	5.1	5.3	4.1	4.5	4.9
RELIGION	0.9	0.9	1.7	0.6	1.7	2.1	0.8	2.0	0.4	0.7	1.2	0.4	1.6	0.7	0.5	0.4	0.2	1.3	0.4	2.8	0.6	1.2	1.2	0.3	2.0	1.0
CLUB	1.0	0.4	0.6	0.2	0.2	0.6	0.4	1.0	0.6	1.1	0.5	0.6	0.7	0.8	0.4	0.9	0.9	0.4	1.3	1.2	0.1	0.4	0.6	1.1	1.1	0.4
EVENT	1.0	0.4	0.5	0.9	0.4	0.5	0.6	0.7	0.9	0.4	0.6	1.0	0.6	0.4	0.8	1.0	0.9	1.1	0.7	0.8	0.0	0.7	0.6	1.2	0.7	0.6
SOC	5.3	5.2	4.4	5.0	4.9	4.0	5.5	4.9	4.4	4.0	5.0	5.6	4.4	4.9	4.7	5.0	2.8	3.9	3.8	4.3	4.3	3.0	4.6	4.1	5.1	4.7
PHOF	0.3	0.1	0.3	0.2	0.1	0.3	0.3	0.2	0.2	0.5	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.4	0.1	0.0	0.2	0.3	0.1	0.3	0.2
PHOO	0.6	0.2	0.2	0.4	0.9	0.5	0.4	0.3	0.8	0.6	0.6	0.7	0.2	0.6	0.5	0.8	0.3	1.0	0.3	0.4	0.1	0.3	0.4	0.2	0.6	0.4
FIT	2.1	1.5	1.6	1.7	1.6	1.0	1.4	1.1	2.6	2.9	2.6	2.3	1.5	2.1	1.8	2.3	1.9	1.9	2.7	2.1	1.5	1.6	2.0	2.2	1.8	1.6
HOBBY	0.3	0.1	0.0	0.3	0.0	0.1	0.1	0.0	0.1	0.2	0.1	0.3	0.1	0.8	0.1	0.1	0.0	0.0	0.1	0.1	1.5	1.6	2.0	2.2	0.1	0.1
TV	18.6	18.1	18.5	22.2	20.0	17.8	15.7	24.2	17.1	17.8	16.8	21.6	17.0	17.8	16.1	13.7	9.8	11.8	13.3	12.5	15.0	16.7	12.6	14.4	13.0	14.8
AUDIO	0.4	0.3	0.4	0.7	0.4	0.9	0.4	0.3	0.4	0.2	0.5	0.6	0.1	0.4	0.5	0.4	0.3	1.0	1.3	0.2	0.6	0.1	0.2	0.9	0.3	0.3
ONLINE	1.0	0.3	0.3	0.5	0.3	0.5	0.3	0.5	1.0	1.4	0.8	0.5	1.5	1.0	1.0	2.2	2.4	2.3	2.0	3.4	0.9	0.7	1.3	4.1	1.4	1.3
READ	2.7	0.6	0.4	1.6	0.7	1.7	1.4	1.1	4.3	4.2	4.9	2.3	5.4	3.4	3.3	3.9	3.4	2.4	3.1	2.0	1.3	1.0	2.0	2.8	2.4	1.9
RELAX	2.3	1.1	2.3	1.0	3.0	4.1	1.7	1.9	1.3	2.2	0.7	2.2	2.4	1.3	3.1	1.5	1.5	1.1	2.5	1.4	1.4	3.1	1.9	1.0	2.5	2.0
FTRAVEL	1.9	1.5	1.7	1.7	1.8	1.6	1.5	1.7	1.4	1.4	1.9	1.9	1.6	1.4	2.0	2.0	1.9	2.1	1.3	2.2	1.1	1.4	1.7	2.3	1.9	1.8
OTHER	5.0	3.0	3.6	4.8	3.8	4.5	4.2	6.3	5.4	5.4	3.5	4.6	1.2	4.9	5.1	2.4	4.0	6.5	3.3	1.8	2.0	4.3	0.5	2.0	4.8	7.3
TOTAL	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0	168.0

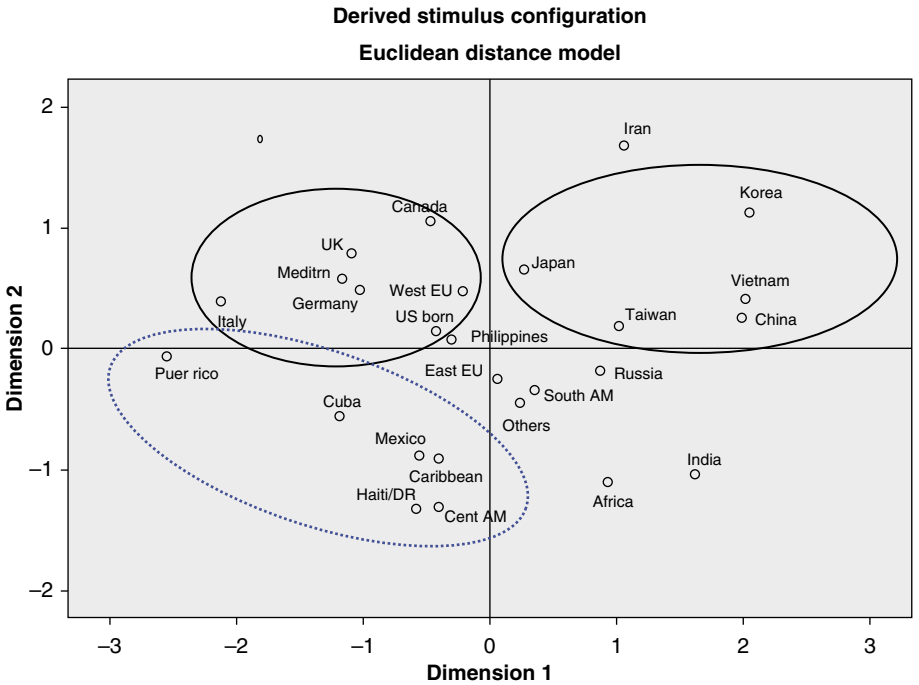

Figure 2 MDS map of time use of US immigrant groups.

1990s. Sayer and Bianchi (2004) and Gupta et al. (2009) found important differences in housework times among different ethnic groups in the United States and in Scandinavian countries.

Childcare

Bianchi et al. (2006) found that American mothers and fathers were spending consistently more time with their children and that this trend could be found in other Western countries. This is much as Gauthier et al. (2004) had determined earlier. At the same time, Bianchi et al. (2006) and Milkie et al. (2004) found that US parents *felt* they were spending too little time in childcare, even though their children did not share such feelings. Gauthier et al. (2004) also found parents investing more time in children in selected industrialized countries.

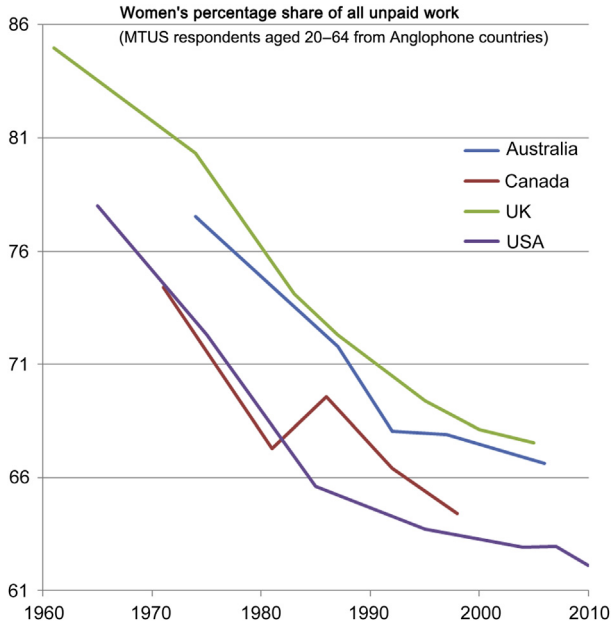

Figure 3 Multinational trends in the gender division of unpaid work.

Shopping and Services

As shown in Table 2, there has been relatively small change in the weekly hours US women (6 h) and men (4 h) spend shopping, which is much the same difference that was found in 1965. This is in marked contrast to the converging trends in the revolutionary changes found for the other two aspects of family care noted above.

Personal Care

Sleep

Contrary to the popular stereotype of the United States and other Western countries as increasingly sleep-deprived, Robinson and Michelson (2010) provided evidence of increases of sleep over the last 40 years in the United States and Canada, as well as other Western countries. Moreover, Michelson (2011) found no support for the argument that 'unhealthy decreases' in sleep patterns resulted from the imposition of daylight savings time in North America. It appears that the ATUS data systematically recorded an hour or more a week more sleep than in earlier studies, possibly related to the 4 a.m. diary starting time.

In terms of the related personal care activity of sex, one that is grossly underreported or disregarded in the ATUS and other time diaries, Robinson (2010) found that less than an estimated hour a week was spent in such activity and that time

had remained quite steady since 1989. Moreover, there was considerable support for the accuracy and consistency of these stylized estimates.

Eating

Substantial use has been made of the new 'eating and health' data, now a regular supplement to the ATUS, in which questions related to Americans' obesity problem were added. Both Janes et al. (2011) and Zick and Stevens (2011) found no direct link between eating time reported in diaries and obesity, although there was evidence relating obesity to less exercising as well as eating as a secondary activity across time. However, Brown et al. (2011) found a strong link between snack eating and obesity (and TV viewing) among Australian teens.

Grooming

Some evidence of decreased grooming time in the ATUS, which may be related to the 'dressing down' movement, can be found in the US diaries.

Free Time

Aguiar and Hurst (2007, 2009) provided evidence that American free time had increased since 1965, much as Robinson and Godbey (1999) reported earlier. However, some reversals have been apparent in the ATUS data since 2006 (see **Table 2**) due to the slight decrease in paid work activity.

Organizations and Volunteering

Abraham et al. (2009) noted how respondents who agreed to complete a time diary for the ATUS reported more volunteering activity in a prior interview than those who did not complete a diary. This is consistent with a view that survey participation is itself a form of volunteering activity.

Social Life

Robinson and Lee (in press) and Robinson (2011) found there were no notable declines in socializing apparent in more recent GSS estimates and ATUS diary times in social life. Gershuny (2003b) reports similar findings.

TV viewing and IT Use

Robinson and Martin (2009) noted no notable declines in TV viewing apparent in more recent MTUS diary times, much as had been reported in earlier studies of Internet impact (DiMaggio et al., 2001).

Differences across Special Demographic Groups

Children's Time Use

Hofferth and Sandberg (2001) and Hofferth (2010) have continued to document important changes in how American children spend their time, as has Raley (2006). Brown et al. (2011) found that higher TV viewing time by school-age children was associated with lower physical activity, higher snack food consumption, and unhealthy weight.

The Child Development Supplement (CDS) of the Panel Study of Income Dynamics (PSID) surveyed over 2500 children through health and achievement test surveys, primary caregiver and child interviews, and a 2-day time use diary (1 week and one weekend day (Hofferth and Sandberg, 2001). Data were collected in 1997, 2002/2003, and 2007/2008 for children in this cohort who remained under 18. The information collected includes a broad array of developmental outcomes including physical health, emotional well-being, intellectual and academic achievement, cognitive ability, social relationships with family and peers, time diaries, and much more.

Hofferth and Sandberg (2001) examined how American children under age 13 spend their time, sources of variation in time use, and associations with achievement and behavior. Their two-day diaries revealed information about the children's time spent in school settings, free play versus organized activities, out-of-school learning activities, and family activities. The data for the analysis were drawn from the 1997 CDS to the PSID.

Seniors

Robinson and Caporaso (2009) found that most differences among retired US seniors centered on fewer active activities, with increases in TV viewing, reading, and relaxing.

Time Changes and New Communication Technologies (IT)

Three early studies of the impact of IT on American society suggested that it was having a negative impact on social life, as well as mass media use. For example, initial findings from Kraut et al. (1998) suggested a negative Internet impact on social life, which was later reversed in Kraut et al. (2002) as more people went online.

However, this was not supported by the results from two large multiyear US national surveys that have been monitoring social change in US daily life with high response rates. These include the 1974–2012 General Social Survey (GSS) with more than 55 000 adult respondents aged 18+ years, which collected time-estimate questions on particular social and media activities, as shown in **Figures 4(**a–c).

As in the ATUS, there seemed little if any evidence of significant impact of IT media on social/media time, even though the GSS showed that IT had already become a predominant source of information and social contact by 2004. **Figure 4(a–c)** shows that GSS respondents in general have not reported lower levels of social or media contact since the 1990s, and while those GSS respondents who spent more time on the Internet reported fewer social visits with relatives, they also reported more visits with friends compared to those who did not use the Internet.

The main difference between users and nonusers in the ATUS was with time at paid work, and that was only partially explained by higher Internet use by teens and on days off from work. For reading and certain other behaviors, Internet use was sometimes associated with *increased* media use in these surveys. Moreover, no consistent decline in either social or media activities was found in either survey across this period of Internet diffusion, much in line with other national studies (Robinson and deHann, 2006; Robinson and Martin, 2009; Wellman et al., 2001). Nielsen/NetRatings (1999) found no reduction of TV viewing in Internet households.

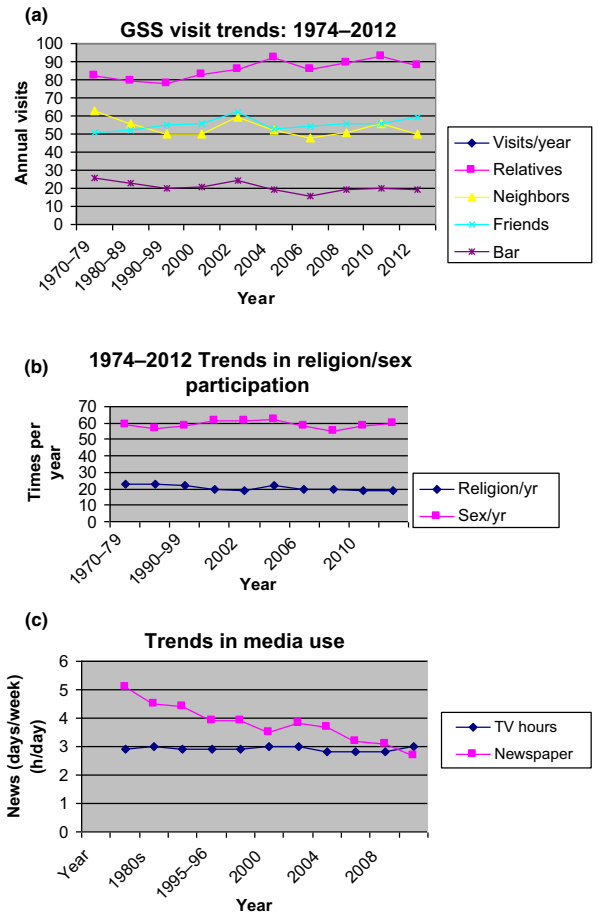

Figure 4 (a): 1972–2012 GSS trends in annual social visits. (b): 1974–2012 GSS trends in annual religion/sex. (c): 1974–2012: GSS trends in daily TV hours and newspaper days.

Longer Historical Perspectives

The long-term ATUS and HETUS data collections have renewed interest in comparisons with much older time use data sets to establish possible longer term historical trends in various daily activities, particularly changes in the division of labor in unpaid work and childcare. Of particular importance are the very detailed United States Department of Agriculture (USDA) 'homemaker studies' and the UK Mass Observation records.

USDA 'Homemaker' Studies' 1923–35

The 1925 *Purnell Act* enabled US State Experiment Stations to use federal funds for research into the economic and social problems of agriculture by undertaking studies in rural sociology, and agricultural and home economics (Kunze, 1988; Pinkett, 1984). Studies of rural and town homemakers' use of time were carried out in a number of states including: Nebraska (Clark and Gray, 1930); Idaho (Crawford, 1927); Washington (Arnquist and Roberts, 1929); South Dakota (Wasson, 1930); Oregon (Wilson, 1929); Rhode Island (Whittemore and Neil, 1929); and Montana (Richardson, 1933). The primary data from all these studies have been missing for a number of years, although the findings are published in Agricultural Experiment Station and USDA Bulletins.

Over the periods 1924–28 and 1930–31, the USDA, with Hildegarde Kneeland as lead researcher, collected approximately 1500 week-long time use diaries from US 'homemakers' in rural and urban areas. The 1924–28 studies included both farm households (559, 69% of the total) and nonfarm homes (249, 31% of the total) in open country or in towns and villages of fewer than 2500 people, while the 1930–31 sample of 692 homemakers resided mainly in cities and large towns (USDA, 1944, p. 1). It is still unclear whether these were separate samples or part of the state-conducted studies listed above.

Recently 643 weekly time use 'summary records' (time budgets) containing totals of time allocated to 58 everyday activities across the 168 h of the diarists' weeks were rediscovered in the US National Archives; 566 (of possibly 808) of these are derived from the diaries kept by homemakers in the 1924–28 sample. Although the 'summary records' do not include biographical information about the diarists, they provide their names and postal addresses. Research at the Centre for Time Use Research (CTUR) has identified more than 95% of these diarists in one or more of the 1920, 1930, or 1940 US Federal censuses, allowing accurate estimation of ages and reconstruction of respondents' occupational and family circumstances.

Only 77 of up to 692 records survive from the 1930-31 'College Women' sample, but they are complete and include: verbatim 7-day diaries completed on each of the 77 respondents; 'supplementary information' questionnaires providing detailed information on household composition and characteristics, appliances and equipment, paid and unpaid help and some open-ended questions on enjoyment of homemaking and summary records for each homemaker's week and; 'summary records' for each diarist.

Harms and Gershuny (2014) contend that Kneeland's (1929) *Annals* article – which sets out what they suspect to be the first ever statement of the modern economists' case for extensions to National Accounts to recognize and include the monetary value of women's unpaid domestic work – is based in part on these studies and believe that these data derive from her own working materials.

Vanek (1974, 1978), Ramey (2009) and others have cited figures from tabular results published in various USDA Bulletins (e.g., Wilson, 1929) in their papers on the impact of household technology. However, Harms and Gershuny (2014) are the first researchers to use the original microdata since the USDA Bulletins were published. They used the resulting individual-level data set to extend the modeling of rural women's time use trends backward from the modern American Time Use Study and the 1975 Michigan Time Use Survey to the new 1920s evidence (the 1930–31 data await analysis). They find, after controlling for demographic and economic circumstances, consistent trends through the 85-year period: substantial decreases of time in routine household operations, and substantial growth in childcare and shopping time (**Figure 5**).

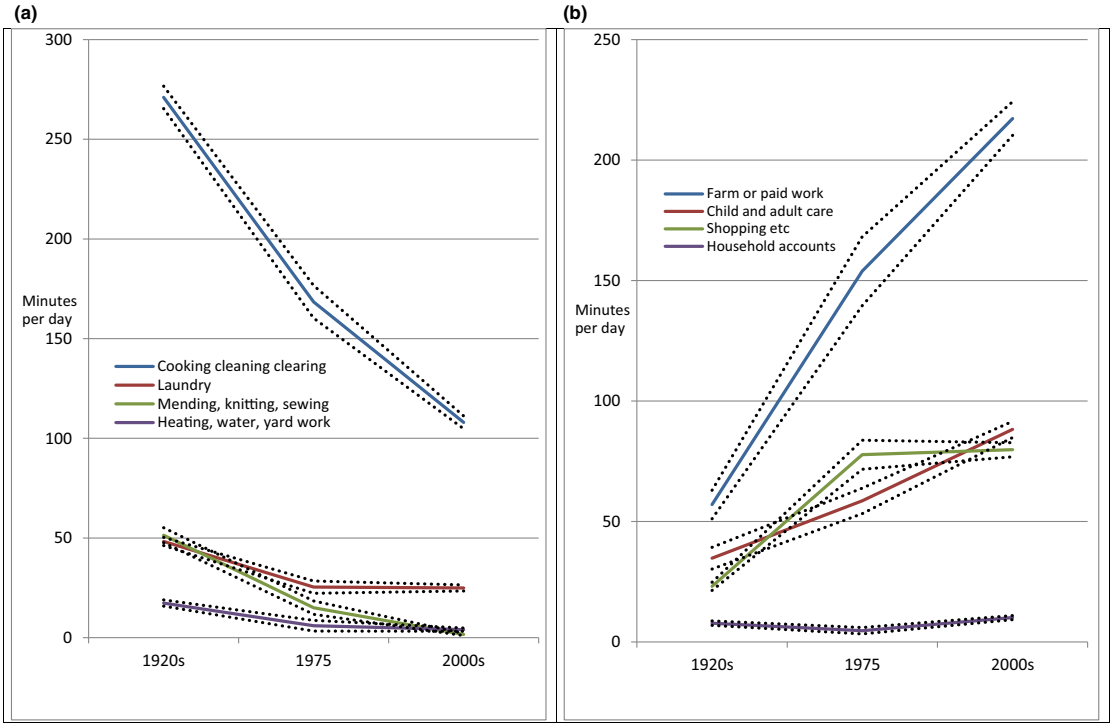

Figure 5 (a): US farm and small town married women from 1920s to 2000s: Decreasing work-time categories. (b): US farm and small town married women from 1920s to 2000s: Increasing work-time categories. Harms, Teresa, Gershuny, Jonathan, 2014. 85 Years of US Rural Women's Time Use. Sociology Working Papers 2014-15. University of Oxford.

Mass Observation

The UK Mass Observation project was an independent British social research organization that documented the attitudes, opinions, and everyday lives of the British people from 1937 to the early 1950s. Madge and Harrison (1937) and others recruited more than 500 volunteers (termed Mass Observers) to report on monthly observation tasks, including complete single-day diaries across an entire year (Gershuny, 1991). The Mass Observation Archive is a charitable trust in the care of the University of Sussex and contains papers generated by the original Mass Observation social research organization and newer material collected continuously since 1981 (www.massobs.org.uk). Only a small proportion of the original diary materials have been coded.

New Data Collection Methods

Diary Supplements: IT

The conventional four aspects of time (what, what else, where, and with whom) became a *de facto* standard as a result of the Szalai et al. 1965 study, and have remained the standard dimensions for time use reporting (e.g., Harvey, 2002) in the HETUS Guidelines (1998, 2008). Newer reporting aspects now included in more recent time use surveys include the emerging 'fifth' field registering whether or not the respondent was using a smart phone, tablet, or computer (e.g., UK National Time Use Study, 2014).

Wearable Cameras

In addition to amateur enthusiasts engaging in 'life-logging' activities, professional researchers are also using wearable cameras (e.g., MeCam, Narrative Clip) or applications on their smart phones and tablets (e.g., DayOne, Momento). Wearable cameras are hands-free (worn on a lanyard or pinned to clothing) and capture about three images per minute of everyday activities from the wearer's point of view; although they do not collect sound, they can log location using GPS. Special ethical considerations need to be taken into account, particularly issues associated with the informed consent of people whose images are captured by participants wearing the camera (Kelly et al., 2013).

Wearable cameras have been used mainly in small-scale, specialized health, and medical studies examining topics such as physical activity (Doherty et al., 2013), sedentary behavior (Kerr et al., 2013), dietary analysis/food diaries (O'Loughlin et al., 2013), travel behavior, (Kelly et al., 2013) and memory loss (Hodges et al., 2011). Researchers have also been experimenting with activity categorization and autocoding (Byrne et al., 2008; Wang and Smeaton, 2013).

It is only very recently that diary methods have been used in conjunction with wearable cameras. Currently underway is a UK study (150 participants) matching data from wearable cameras, time use diaries, and accelerometers to clarify and classify daily activities. One of the goals of the project is to validate and calibrate self-report diaries as a data collection instrument by coding 'streams' of continual day-long images from the wearable camera.

Holmes and Bloxham (2009) describe promising results from a US study using an 'electronic shadow' device that tracks a subject's daily activity by means of an observer who records each activity on that device every 10 s during the waking day.

Internet and Mobile Applications

Smart phones and tablets, with their wide range of 'applications,' offer researchers new approaches to capturing people's everyday time use. The smart phone applications provide time-tracks of location and personal energy expenditure, using GPS and motion sensors, as well as provide opportunities for respondents to register activities, purposes and other descriptive materials. Many people carry and use their smart phone (or tablet) with them throughout the day and use it regularly, enabling researchers to easily ask respondents follow-up subjective and objective questions.

Web-Based Data Collection

Researchers at the Vrije University in Belgium have designed a Modular Online Time Use Survey (MOTUS) as a pilot project to collect time use data using online registration methods (Minnen et al., 2013). A main goal of the study was to evaluate the comparability between online time use data and traditional paper-and-pencil diary data. The data collection took place in 2013, with 40 000 respondents (aged 18–75 years). The researchers believe that MOTUS will allow low cost, time-efficient, and high-quality time use data to be gathered (e.g., coding done at point of data collection, inconsistencies identified during collection via prompting, etc.). The online diary is based on the HETUS Guidelines (2008).

Smartphone Applications

Researchers from the Netherlands Institute for Social Research (Fernee et al., 2013) carried out a pilot time-diary study with 100 respondents using a smart phone application they designed. Before the pilot phase, they conducted several experiments to test how well participants with and without their own smart phones were able to use the appication. Results suggest that both groups were able to participate effectively, with some exceptions. A larger time use study with 2000 respondents is underway to allow comparison between smart phone and traditional paper diaries, both based on the 2008 HETUS Guidelines (2009).

Subjective Time

The QOL or positive benefits of engaging in particular daily activities is not always obvious, although social observers seem to concur that increases in daily activities like childcare, volunteering, and other potentially altruistic behaviors (as well as free time) represent improvements in a person's or society's QOL. In contrast, increases in time spent on routine housework, repair activities, and TV viewing are seen as less desirable. Nonetheless, it is important to have empirical support for such assumptions, and this is directly provided by the subjective time measures (Robinson, 2013a,b; Robinson and Gershuny 2013).

Subjective time measures fall into three categories, which are as follows: (1) *Affect* relates to enjoyment and satisfaction, with questions about how much enjoyment, pleasure, or happiness respondents experience from specific activities. (2) *Time pressure* concerns perceptions of time pressure and stress, with questions about feeling rushed, time crunch, and personal stress. (3) *Output* relates to questions about outputs from time inputs, such as being more satisfied with home meals or cleanliness. These have supplemented more 'objective' measures of output, like grades in school, house cleanliness, or obesity.

Fuller coverage of the subjective measures is provided in a separate topic section in this volume.

General Trends and Patterns in Time Use

Cross-National Results

Finally, multinational research reveals surprising activity convergences across most of the more than 30 countries studied (some identified below), although most of them are developed Western societies in Europe (that can afford to conduct expensive time-diary surveys). For example, Bittman (2000) found similar increases in free time since the 1960s in other countries, much as in the United States. In an analysis of all productive activity (contracted and committed time together) in 12 countries, Goldschmidt-Clairmont and Pagnossin-Aligisakis (1995) found the same basic equality of men and women in overall hours spent on such activities. A study of parental time with children found the same pattern of increased father childcare across six Western countries since 1990 (France being a notable exception), much as was reported in a 2004 study (Gauthier et al., 2004) for many other countries over this period. Robinson and Godbey (1999) found many similarities in both the trends and predictors of time use in Japan, Russia, and Canada. Gershuny (2003a) has extended these results to more than 15 other countries.

US Trends in Time Use

As noted above, diary data have been collected at roughly 10-year intervals since 1965 and have shown some consistent trends since then. The major overall trend has been in the form of a gender switch; paid work has been declining for men in contrast to paid work increasing for women, coupled with increased housework for men, and a large decrease in housework for women. Moreover, there is more general gender convergence evident in many other personal and free-time activities as well (Robinson and Godbey, 1999). Overall, there seems little evidence of a more hectic lifestyle, say in terms of less free time or sleep or declining sleep.

Other major general increases (for both men and women) are evident for childcare, TV viewing, fitness activities, and relaxing (and free time in general), in contrast to notable decreases in newspaper reading, hobbies, and work breaks. Many of these changes for those aged 18–64 years had been reported in Robinson and Godbey (1999). Data on activities for which estimate trend data are also available, mainly from

the GSS, the Survey of Public Participation in the Arts (SPPS), the PSID, and organizations interested in particular activities (like sleep or volunteering).

The General Social Survey

As illustrated in **Figure 4**, there are several GSS activities for which estimate data are available: work, TV and newspaper reading, religious attendance, sex, and several other types of social interaction. Analyses in Robinson and Martin (2008) and Marsden (2012) show the direction of 1972–2000 trends for many of these behaviors, particularly the declines in newspaper reading and church attendance (Presser and Stinson, 1998) and little consistent change in social activity. However, the GSS data show no change in TV viewing hours over the period 1972–2004, in contrast to the 70% increase in viewing found in the 2005 diaries compared to 1965. The PSID housework decline for women and increase for men is quite consistent with that shown in the diaries (Juster et al., 2003).

Summary and Conclusions

While the time diary is one of many options for measuring how people spend time, its general adoption across most Western countries thus far has produced a rich body of insightful results. This promises to increase in the future given the increased availability of continuous, large-sample, and high-quality government data, as in the many and varied applications of the ATUS (Robinson, 2009).

Diary data have raised serious issues about the accuracy of data obtained by more traditional 'time-estimate' questions, both in terms of simple amounts levels of time spent and in trends in these amounts across recent decades. Research undertaken with diaries supports the conclusion that the time estimates are open to the common survey problem of social desirability (Krosnick, 1998). While more definitive methodological research is needed as these ATUS data are more widely used, the diary method has showed promising results in terms of the major criteria of reliability and validity. It also needs to be noted that, although most of the data reviewed here are for those aged 18–64 years, the ATUS diary data are available for senior citizens aged 65+ years (Robinson and Godbey, 1999; Robinson, 2010). The ATUS now also provides activity data for large samples of US teenagers aged 15–17 years, and PSID diary data provide supplemental diary data for even younger children (Raley, 2006; Robinson and Godbey, 1999), often revealing the few hours American children apparently spend doing school homework, paid work, or household work.

Diary trends since the first national US diary study in 1965 indicate a number of unexpected societal changes, particularly in the converging lifestyles of men and women, who are now doing more equal amounts of not just paid work and housework, but most personal and free-time activities as well. Diary trends also indicate more TV viewing, childcare, and fitness activity, which has been offset by decreases in housework, eating meals, and newspaper reading. There also appears to be a notable increase in secondary activities, with evidence of increased multitasking (mainly involving TV, radio, and other newer media). This is reflected as well in the 20% point increase in the proportion of respondents who say they are often doing two or more activities at once (Bianchi et al., 2006).

Based on these subjective data, it could be taken as an encouraging sign that the trends in US diary hours do follow a generally hedonistic direction, in that Robinson and Godbey (1999) find people report spending increased time on activities they find more enjoyable (like childcare and free-time activities) and less time on the things they do not like to do (especially housework). Along the same line, Bianchi et al. (2006) report increased time spent on childcare, especially interactive or 'fun' activities with one's child. Moreover, these same long-term *time* trends are found in other countries as well (Bittman, 1998; Gershuny, 2003a). Conversely, time spent on the most enjoyable daily activity of sex has not been increasing since 1989, in spite of the increase in Viagra, Internet services and pornography, and other marital aids (Robinson, 2010).

In terms of public policy, time diaries provide clear evidence of several social inequalities. In the case of the division of housework between men and women, **Table 4** shows that both men and women rate housework as quite low in terms of enjoyment. Yet women continue to do twice as much of it.

The 'Take Back Your Time' movement (De Graaf, 2003) proposed that American workers stop all paid work after mid-October, so that their total annual work hours would be equivalent to those of European workers. Yet, Hochschild (1997) found in one company that it was other workers themselves who sabotaged innovative family friendly policies to reduce work hours. The National Sleep Foundation has raised alarms about sleep deprivation as a national crisis, although time diaries show no decline in sleep over the last 40 years (which holds steady at about the recommended 8 h a night). The 'turn off your TV' movement has designated 1 week a year for people to leave their sets off and to participate in more potentially gratifying activities instead. However, **Table 2** suggests that, although people rate TV in general as relatively low on the fun meter, the programs they watched on the diary day rated far higher in enjoyment.

Of all the public potential policy issues related to time use, possibly the most impressive is the lack of solid data on the marginal outputs that workers produce and obtain for their inputs of work time. Some data suggest that the same work productivity can be achieved with 20–50% less

Table 4 SWB item intercorrelations and factor loadings

	Sad	*Pain*	*Tired*	*Stress*	*Happy*[a]	*Factor loading*
Sad		0.37	0.27	0.49	0.32	*0.75*
Pain			0.34	0.34	0.17	*0.67*
Tired				0.37	0.17	*0.62*
Stress					0.32	*0.77*
Happy						*0.54*

Table 5 ATUS SWB scores for different types of activities (6.0 = highest SWB to 0.0 lowest SWBL)

ATUS CODE	WORK	HOUSEWORK	CHILDCARE	SHOP	PERSONAL	EDUCATION	ORGANIZ'N	SOCIAL	RECREATION	MEDIA	TRAVEL	ATUS CODE
	5	2	3	9 10 11	1 4 11	6	14 15	12 13	12 13	12 16	18	
6.0							INSTRUCT	PARTY				6.0
5.9								OTHER ARTS				5.9
5.8			ART									5.8
5.7							SERVICES					5.7
5.6									SWIM			5.6
5.5				PERS CARE								5.5
5.4			PLAY									5.4
5.3								MOVIES	BASKETBALL			5.3
5.2			PHYS OKID						FISH			5.2
5.1									GOLF		RELIGION	5.1
5.0			READ TO				VOLUNTEER				PARTY	5.0
4.9									GAMES			4.9
4.8			PICK OTH KID								RECREATION	4.8
4.7		WALK PET		BANK						AUDIO		4.7
4.6												4.6
4.5										RADIO	EAT	4.5
4.4			WAIT KID						CRAFTS			4.4
4.3								SOCIAL			VOLUNTEER	4.3
4.2							PRACTICE			READ	ARTS	4.2
4.1									BIKE		SOCIAL	4.1
4.0			SUPERVISE						WALK		PETS	4.0
3.9									WORK OUT		GAS	3.9
3.8					EAT						OTHER SHOP	3.8
3.7								PHONE OTHER			GROCERY	3.7
3.6			PICK UP								OTHER ADULTS	3.6
3.5		SHOW FOOD		OTH FOOD					RUN		OTHER KID	3.5
3.4				OTH SHOP							HH MGT	3.4
3.3			SPORT	GAS			VOL-CMPTR				FOOD	3.3
3.2			PHYSICAL								OTH KIDS	3.2
3.1		COOK	TALK/PLAN									3.1
3.0		REPAIR						PHONING	COMPUTER		OTHER	3.0
2.9		CAR REPR	HOMEWORK								KIDS	2.9
2.8		MAIL		GROCERY				PHONE			BANK	2.8
2.7		FEED PET								TV		2.7
2.6		EMAIL										2.6
2.5		DISHES							LIFT	SMOKE		2.5
2.4		PLAN									CLASS	2.4
2.3		OTH HOUSE									COMMUTE	2.3
2.2		CLEAN								RELAX		2.2
2.1						CLASS						2.1
2.0												2.0
1.9												1.9
1.8												1.8
1.7									CARDIO			1.7
1.6	MAIN JOB											1.6
1.5											PERSONAL	1.5
1.4												1.4
1.3												1.3
1.2												1.2
1.1		HEAT REPAIR										1.1
1.0												1.0
0.9												0.9
0.8												0.8
0.7												0.7
0.6												0.6
0.5												0.5
0.4												0.4
0.3												0.3
0.2												0.2
0.1	2nd JOB	FINANCES										0.1
0.0	JOB SEARCH			MEDICAL							MEDICAL	0.0

time spent on the job (Parkinson, 1957; Schor, 1991; Schuman et al., 1985) and that many workers put in simple 'face time' to impress their employers.

The findings reported in **Tables 4 and 5** represent an important first step toward gaining more nuanced insights into time use statistics, but more holistic study designs – such as a 'representative ethnography' that combines the insights of observation with the generalizability of representative sampling of respondents could provide more enlightened meanings (Robinson and Meadow, 1982). Although confined to northern Californian middle-class parents, the observational study of Darrah and Murphy (2009) provides an impressive model. Diary data allow researchers to document what people are doing, but we need to find out more about why, with whom, for whom, and to what ends people are using their time.

See also: Cultural Participation, Trends in; Gender, Time-Use, and Fertility Recovery in Industrialized Countries; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective; Time Use Research Methods; Time Use Research: Subjective Time Use; Time Use and Gender; Time Use and Resource Consumption.

Bibliography

Abraham, K., Helms, S., Presser, S., 2009. How social processes distort measurement: the impact of nonresponse on estimates of volunteer work in the United States. American Journal of Sociology 114 (4), 1129–1165.

Abraham, K., Maitland, K., Bianchi, S., 2006. Nonresponse in the American time-use survey. Public Opinion Quarterly 70 (5), 676–703.

Aguiar, Mark, Hurst, Erik, 2009. A summary of trends in American time allocation. Social Indicators Research 92, 70–80.

Aguiar, Mark, Hurst, Erik, 2007. Measuring trends in leisure: the allocation of time over five decades. Quarterly Journal of Economics 122 (3), 969–1006.

Arnquist, Inez, Roberts, Evelyn H., July 1929. The Present Use of Work Time of Farm Homemakers. State College of Washington Agricultural Experiment Station, Division of Home Economics, Bulletin No. 234.

Bianchi, S., Milkie, M., Sayer, L., Robinson, J., 2000. Is anyone doing the housework? Trends in the gender division of household labor. Social forces 79 (1), 191–228.

Bianchi, S., Robinson, J., Milkie, M., 2006. Changing Rhythms of American Family Life. Russell Sage, New York.

Bianchi, S., Sayer, L., Milkie, M., Robinson, J., 2012. Housework: who did, does or will do it, and how much does it matter? Social Forces 91 (1), 55–65.

Bittman, M., 2000. Now it's 2000: "trends in doing and being in the new Millennium". Journal of Occupational Science 7, 108–117.

Bittman, Michael, 1998. The land of the lost long weekend? trends in free time among working age Australians, 1974-1992. Society and Leisure 21 (2), 353–378.

Bloombaum, M., 1970. Doing smallest space analysis. Journal of Conflict Resolution 14, 409–416.

Brown, J., Nicholson, J., Broom, D., Bittman, M., 2011. Television viewing by school-age children: associations with physical activity, snack food consumption and unhealthy weight. Social Indicators Research 101, 221–225.

Byrne, Daragh, Doherty, Aiden, Cees, G., Snoek, M., Gareth, G., Jones, F., Smeaton, Alan, 2008. Validating the detection of everyday concepts in visual lifelogs. In: Duke, David, Hardman, Lynda, Hauptmann, Alex, Paulus, Dietrich, Staab, Steffen (Eds.), Semantic Multimedia. Springer, Berlin Heidelberg, pp. 15–30. Lecture Notes in Computer Science.

Clark, Ruth, Gray, Greta, January 1930. The Routine and Seasonal Work of Nebraska Farm Women. University of Nebraska College of Agriculture Experiment Station, Home Economics, Bulletin No. 238.

Converse, P., 1972. Country differences in time use. In: Szalai, A. (Ed.), The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries. Mouton, The Hague, pp. 145–177.

Crawford, Ina, January 1927. The Use of Time by Farm Women. University of Idaho Agricultural Experiment Station, Department of Home Economics, Bulletin No. 146.

Darrah, Charles, Murphy, Alicia, 2009. Ethnographic Methods and Work-Family Research. Entry in Sloan Network Encyclopedia.

DiMaggio, P., Hargittai, E., Neuman, W., Robinson, J., 2001. Social implications of the internet. Annual Review of Sociology 27, 307–336.

Doherty, A., Kelly, P., Kerr, J., Marshall, S., Oliver, M., Badland, H., Hamilton, A., Foster, C., 2013. Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. International Journal of Behavioral Nutrition and Physical Activity 10, 1–11.

Elizur, D., 1984. Facets of work values: a structural analysis of work outcomes. Journal of Applied Psychology 69 (3), 379–389.

Eurostat, 2009. Harmonised European Time Use Surveys: Guidelines 2008. Eurostat, Luxembourg.

Fernee, H., Sonck, N., Scherpenzeel, A., 2013. Data Collection with Smartphones: Experiences in a Time Use Survey. Tilburg University.

Fisher, K., Egerton, M., Gershuny, J., Robinson, J., 2006. Gender convergence in the American heritage time use survey. Social Indicators Research 70, 349–368.

Fisher, K., Gershuny, J., 2014. Multinational Time Use Study, Versions World 5.5.3, 5.80 and 6.0 (released October 2012). Created by Jonathan Gershuny and Kimberly Fisher, with Evrim Altintas, Alyssa Borkosky, Anita Bortnik, Donna Dosman, Cara Fedick, Tyler Frederick, Anne H. Gauthier, Sally Jones, Jiweon Jun, Aaron Lai, Qianhan Lin, Tingting Lu, Fiona Lui, Leslie MacRae, Berenice Monna, José Ignacio Giménez Nadal, Monica Pauls, Cori Pawlak, Andrew Shipley, Cecilia Tinonin, Nuno Torres, Charlemaigne Victorino, and Oiching Yeung. Centre for Time Use Research, University of Oxford, United Kingdom.

Fisher, K., Robinson, J., 2011. Daily life in 23 countries. Social Indicators Research 101, 295–304.

De Graaf, J. (Ed.), 2003. Take Back Your Time. Berrett-Koehler Publishers, San Francisco.

Gauthier, A.H., Smeeding, T.M., Furstenberg Jr., F.F., 2004. Are parents investing less time in children? Trends in selected industrialized countries. Population and Development Review 30, 647–671.

Gershuny, J., 1991. Are we running out of time? Futures 24 (1), 3–22.

Gershuny, J., 2003a. Changing Times: Work and Leisure in Postindustrial Society, second ed. Oxford University Press, Oxford.

Gershuny, J., 2003b. Web-use and net-nerds: a neo-functionalist analysis of the impact of information technology in the home. Social Forces 82 (1), 141–168.

Gershuny, J., 2009. Veblen in reverse: evidence from the multinational time use archive. Social Indicators Research 93 (1), 37–45.

Goldschmidt-Clairmont, L., Pagnossin-Aligisakis, E., 1995. Measures of Unrecorded Economic Activities in Fourteen Countries. United Nations Development Programme, Human Development Report Office, Occasional Paper No. 20.

Gupta, S., Sayer, L., Cohen, P., 2009. Earnings and the stratification of unpaid time among U.S. women. Social Indicators Research 93, 153–157.

Guttman, Louis, 1968. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika 33, 469–506.

Harms, Teresa, Gershuny, Jonathan, 2014. 85 Years of US Rural Women's Time Use. Sociology Working Papers 2014-15. University of Oxford.

Harvey, Andrew, 2002. Guidelines for time use data collection and analysis. In: Time Use Research in the Social Sciences. Springer, US, pp. 19–46.

Hochschild, Arlie Russell, 1997. The Time Bind: When Work Becomes Home and Home Becomes Work. Metropolitan Books, New York.

Hodges, S., Berry, E., Wood, K., 2011. Sensecam: a wearable camera that stimulates and rehabilitates autobiographical memory. Memory 19 (7), 685–696.

Hofferth, S., 2010. Changes in American children's time: 1997-2003. Electronic International Journal of Time Use Research 6 (1), 26–47.

Hofferth, S., Sandberg, J., 2001. How American children spend their time. Journal of Marriage and the Family 63, 295–308.

Holmes, M., Bloxham, M., 2009. An observational method for time use research: lessons learned from the Middletown Media Studies. Social Indicators Research 99, 186–194.

Inglehart, Ronald, Welzel, Christian, 2011. The WVS Cultural Map of the World. World Values Survey. http://www.worldvaluessurvey.org/wvs/articles/folder_published/article_base_54 (accessed 10.08.14.).

Janes, M., Hamrick, K., Lacey, J., 2011. Exercise, eating patterns, and obesity: evidence from the ATUS and its eating & health module. Social Indicators Research 101 (2), 215–219.

Juster, T., Ono, H., Stafford, F., 2003. An assessment of alternative measures of time use. Sociological Methodology 33, 19–54.

Kan, M.Y., Sullivan, O., Gershuny, J., 2011. Gender convergence in domestic work: discerning the effects of interactional and institutional barriers from large-scale data. Sociology 45 (2), 234–251.

Kelly, P., Doherty, A., Berry, E., Hodges, S., Batterham, A., Foster, C., 2011. Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study. International Journal of Behavioral Nutrition and Physical Activity 8, 1–9.

Kelly, P., Marshall, S., Badland, H., Kerr, J., Oliver, M., Doherty, A.R., Foster, C., 2013. An ethical framework for automated, wearable cameras in health behavior research. American Journal of Preventative Medicine 44 (3), 314–319.

Kerr, J., Marshall, S.J., Godbole, S., Chen, J., Legge, A., Doherty, A.R., Kelly, P., et al., 2013. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. American Journal of Preventative Medicine 44 (3), 290–296.

Kneeland, Hildegarde, 1929. Woman's economic contribution in the home. Annals of the American Academy of Political and Social Science 143, 33–40.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., Crawford, A., 2002. Internet paradox revisited. Journal of Social Issues 58, 49–74.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., Scherlis, W., 1998. Internet paradox. American Psychologist 53 (9), 1017–1031.

Kunze, Joel, 1988. The purnell act and agricultural economics. Agricultural History 62 (2), 131–149.
Krosnick, J., 1998. Maximizing questionnaire quality. In: Robinson, J., Shaver, P., Wrightsman, L. (Eds.), Measures of Personality and Social Psychological Attitudes. Academic Press, New York, NY.
Kruskal, J., 1964. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29, 115–129.
Laumann, Edward, Guttman, Louis, 1966. The relative associational contiguity of occupations in an urban setting. American Sociological Review 3, 169–178.
Minnen, Joeri, Glorieux, Ignace, van Tienoven, Theun Pieter, Djiwo, Weenas, 2013. MOTUS: Modular Online Time-Use Survey. Vrije Universiteit, Brussel.
Madge, C., Harrison, T., 1937. Mass Observation by London. Frederick Muller Ltd, London.
Marsden, P. (Ed.), 2012. Social Trends in American Life Princeton. Princeton University Press, Mass Observation Archive, University of Sussex, UK, Princeton, NJ. www.massobs.org.uk.
Mestdag, Inge, Glorieux, Ignace, 2009. Change and stability in commensality patterns: a comparative analysis of Belgian time-use data from 1966, 1999 and 2004. The Sociological Review 57 (4), 703–726.
Michelson, W., 2011. Sleep time, media hype vs diary data. Social Indicators Research 101, 277–280.
Milkie, Melissa M., Nomaguchi, Kei, Bianchi, Suzanne, 2004. The time squeeze: parental statuses and parents' feelings about time with children. Journal of Marriage and Family 66, 739–761.
Nielsen/NetRatings, 1999. TV Viewing in Internet Households. A report by Nielsen Media Research. Retrieved from: http://www.nielsen-netratings.com/.
O'Loughlin, G., Cullen, S., McGoldrick, A., O'Connor, S., Blain, R., O'Malley, S., Warrington, G., 2013. Using a wearable camera to increase the accuracy of dietary analysis. American Journal of Preventative Medicine 44 (3), 297–301.
Parkinson, C., 1957. Parkinson's Law and Other Studies in Administration. The Riverside Press, Massachusetts.
Pinkett, H., 1984. Government research concerning problems of American rural society. Agricultural History 58 (3), 365–372.
Presser, S., Stinson, L., 1998. Data collection mode and social desirability bias in self-reported religious attendance. American Sociological Review 63, 137–145.
Raley, S., 2006. Children's time use. In: Bianchi, S., Robinson, J., Milkie, M. (Eds.), Changing Rhythms of American Family Life. Russell Sage Foundation, New York.
Ramey, Valerie, 2009. Time spent in home production in the twentieth-century United States: new estimates from old data. The Journal of Economic History 69 (1), 1–47.
Records of the U.S. Department of Agriculture Bureau of Human Nutrition and Home Economics, Division of Economics, Record Group 176, Boxes 640–654, Records Relating to a 'Use of Time on Farms' Study, National Archives and Records Administration, College Park, Maryland.
Richardson, Jessie, 1933. The Use of Time by Rural Homemakers in Montana. Montana State College Agricultural Experiment Station, Department of Home Economics, Bulletin No. 271.
Robinson, J., 2009. Time use and qualities of life. Social Indicators Research 99, 1–254.
Robinson, J., 2010. Sex, arts and verbal abilities: three further indicators of how American life has not improved. Social Indicators Research 99 (1), 1–12.
Robinson, J., 2011. IT, TV and time displacement: what Alexander Szalai anticipated but couldn't know. Social Indicators Research 101 (2), 193–206.
Robinson, J., 2013a. Americans less rushed, but no happier: 1965–2010 trends in subjective time and happiness. Social Indicators Research 113 (3), 1091–1104.
Robinson, J., 2013b. Happiness is being just rushed enough. Scientific American Online February.
Robinson, J., Spring–Summer 2013c. Time use as a social indicator. SINET 1 (2), 114–115.
Robinson, John, Caporaso, A., 2009. Senioritis in repose. Electronic International Journal of Time Use Research 7 (1), 61–72.
Robinson, J., deHaan, J., 2006. Information technology and family time displacement. In: Kraut, R., Brynan, M., Kiesler, S. (Eds.), Computers, Phones and the Internet. Oxford University Press, Oxford, pp. 51–69.
Robinson, J.P., Godbey, G., 1999. Time for Life: The Surprising Ways Americans Use Their Time. Pennsylvania State University Press, University Park.
Robinson, John, Hefner, Robert, 1968. Perceptual maps of the world. Public Opinion Quarterly 32 (2), 273–280.
Robinson, J., Lee, C. Society's (virtually) time-less transition into the digital age. Social Indicators Research, in press.
Robinson, J., Gershuny, J., 2013a. Visualizing multinational daily life via multidimensional scaling (MDS). Electronic International Journal of Time-Use Research 10, 76–92.
Robinson, J., Gershuny, J., 2013b. Life-style conclusions from synthetic weekly time diaries and stylized questions. Electronic International Journal of Time-Use Research 10, 102–106.
Robinson, J., Martin, S., 2008. IT and activity displacement: evidence from the general social survey. Social Indicators Research 91, 115–139.
Robinson, J., Martin, S., 2009. Of time and television. The Annals of Social and Political Science (The End of Television) 265, 74–87.
Robinson, J., Martin, S., 2010. IT use and declining social capital: more cold water from the GSS and ATUS. Social Science Computer Review 29 (1), 45–63.
Robinson, J., Michelson, W., 2010. Sleep as a victim of the time crunch – a multinational analysis. Electronic International Journal of Time-Use Research 7 (1), 61.
Robinson, J., Meadow, R., 1982. Polls Apart. Seven Locks Press, Potomac, MD.
Sayer, Liana C., Bianchi, Suzanne M., 2004. Are parents investing less in children? Trends in mothers' and fathers' time with children. American Journal of Sociology 110, 1–43.
Schor, J., 1991. The Overworked American. Basic Books, New York.
Schuman, Howard, Walsh, Edward, Olson, Camille, Etheridge, Barbara, 1985. Effort and reward: the assumption that college grades are affected by quantity of study. Social Forces 63 (4), 945–966.
Szalai, A., 1972. The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries. Mouton, The Hague.
United States Department of Agriculture, Agricultural Research Administration, Bureau of Human Nutrition and Home Economics, 1944. The Time Costs of Homemaking – a Study of 1,500 Rural and Urban Households. US Government Printing Office, Washington, DC.
Vanek, J., 1974. Time spent in housework. Scientific American 231 (5), 116–120.
Vanek, J., 1978. Household technology and social status: rising living standards and status and residence differences in housework. Technology and Culture 19 (3), 361–375.
Wang, Peng, Smeaton, Alan, 2013. Using visual lifelogs to automatically characterize everyday activities. Information Sciences 230, 147–161.
Wasson, Grace, 1930. The Use of Time by South Dakota Farm Homemakers. South Dakota State College Agricultural Experiment Station, Department of Home Economics, Bulletin No. 247.
Wellman, B., Haase, A., Witte, J., Hampton, K., et al., 2001. Does the internet increase, decrease or supplement social capital? social network, participation and community commitment. American Behavioral Scientist 45, 437–456.
Whittemore, Margaret, Neil, Berniece, 1929. Time Factors in the Business of Homemaking in Rural Rhode Island. Rhode Island State College Agricultural Experiment Station, Bulletin No. 221.
Wilson, Maud, 1929. Use of Time by Oregon Farm Homemakers. Oregon State Agricultural College Agricultural Experiment Station, Bulletin No. 256.
Zick, K., Stevens, R., 2011. Time spent eating and its implications for Americans' Energy balance. Social Indicators Research 101 (2), 267–273.

Time Use Research: Subjective Time Use

John P Robinson, University of Maryland, College Park, MD, USA
Jonathan Gershuny, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The quality of life (QOL) or positive subjective benefits of the time people spent engaging in particular daily activities is often not obvious, although social observers seem to concur that increases in daily activities like child care, volunteering, and other potentially altruistic behaviors (as well as overall free time) represent improvements in a person's or society's QOL. In contrast, increases in time spent on routine housework, repair activities, and TV viewing are seen as less desirable. Nonetheless, it is helpful to have empirical support for many of these assumptions, and this is directly provided by the subjective time measures reviewed here.

Introduction

Subjective time refers to attitudes and feelings related to various activities or uses of time. Subjective time measures in this article fall into three categories:

1. *Affect* relates to enjoyment and satisfaction, with questions about how much enjoyment, pleasure, or happiness respondents obtain from specific activities (or from their lives more generally), as shown in Tables 1 through 4. These mainly include how people feel momentarily, or in 'real time' as in a time diary, but can be extended to include more general quality of life (QOL) feelings, and there is now evidence in the American Time-Use Survey (ATUS) that the two are connected and may predict each other.
2. *Time pressure* concerns perceptions of felt time pressure and stress, with questions about feeling rushed, time crunch, and personal stress.
3. *Outcome* concerns questions related to outputs from time inputs, such as being more satisfied with one's job, house cleanliness, or how well one's children are doing in life. These can be supplemented with more 'objective' measures of output, like grades in school or health

Affect

Specific attempts to identify how time-diary keepers feel about their daily activities began in the first U.S. diary study in 1965, when respondents were asked about how much satisfaction they generally derived from different daily activities, and about what were their most and least liked activities during the diary day (work was listed relatively often in response to both questions).

In contrast, respondents in waves 2 and 4 of the subsequent 1975 diary interviews *were* asked to give specific enjoyment ratings to the 19 common daily activities shown in Table 1, using a 0–10 scale, where 0 meant 'Dislike a great deal' and 10 meant 'Enjoy a great deal.' Average scores on these 1975 general activity scales from Juster and Stafford (1985) are ordered by level of enjoyment on the right side of Table 1 (alongside the parallel diary-based ratings collected in 1985 on the left side, as described next).

In their 1985 telephone interviews, then, diary respondents were asked to rate each activity in their time diary on that same 0–10 scale employed in the 1975 general ratings, as averaged on the left side of Table 1. In other words, in addition to the usual diary follow-up questions on secondary activities, location, and with whom, 1985 diary respondents were also asked to rate how much they enjoyed each such activity 'in real time' rather than generally. Some differences by the two Table 1 rating contexts are discussed below, although it can be seen that most differences are not notable.

Indeed, Robinson and Godbey (1999) found that the rank order of activity affect for these ratings in the diary correlated about 0.70 with Juster and Stafford's (1985) ratings of activities in general. But there were important exceptions. First, 1985 workers rated their daily *paid work* activities as less enjoyable than 1975 workers rated their work in general. Second, 1985 diary respondents rated their *TV viewing* episodes as more enjoyable in their diaries than 1975 respondents did in general (perhaps explaining why people may inadvertently watch more TV than they would like). It was also found that these 1985 diary enjoyment ratings were highly predictive of their TV viewing times in diaries collected 3 months later, a case of attitudes predicting behavior, and one that held for other diary behaviors as well.

Social Well-Being Ratings in ATUS 2010

One important distinction between the 1985 enjoyment ratings in Table 1 and the 2010 Social Well-Being (SWB) diary ratings described next is that the 1985 enjoyment ratings used only a single 0–10 rating scale for all activities, as opposed to the five descriptive ratings in Table 2, each of which was rated using 0–6 unipolar scales. This greater breadth and coverage meant, however, that they could only be asked for three of the activities in each respondent's diary, thus missing the opportunity to examine the full ebb and flow of daily emotions (to test Gershuny's (2012) intriguing 'diminishing utility of an activity across-the-day hypothesis,' for example).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32204-8

Table 1 Enjoyment ratings from 1985 diary activities versus 1975 in general: national data

10 ENJOY A GREAT DEAL *1985 Diary (average = 7.0)*	*1975 General (average = 6.8)*
9.3 Sex	
9.2 Play sports	
8.7 Playing/read with children	8.9 Child care
8.5 Church, religion	8.6 Play with children
8.5 Sleep	
8.2 Meals away	
8.2 Socialize, visit others	
8.0 Socialize with family	8.0 Socializing, talking
8.0 Work breaks	*8.0 Work*
7.9 Reading	
7.8 Meals at home	
7.8 TV	**7**.5 Sleep
7.4 Hobbies, crafts	7.4 Eating
7.2 Exercise	7.4 Washing, dressing
7.2 Baby care	7.3 Church, religion
7.2 Organizations	
7.0 Work	7.0 Reading
7.0 Bathing	
6.6 Cooking	6.8 Hobbies
6.6 Other shopping	6.5 Play sports
6.4 Child care	6.5 Cultural events
6.4 Help others	
6.3 Work commute	6.2 Cooking
6.1 Dressing	
5.8 Other housework	*5.9 TV*
5.5 Grocery shopping	
5.5 Home repairs	
5.2 Pay bills, financial, etc.	5.1 Home repairs
5.0 Yardwork	5.0 Organizations
4.9 Clean house	4.6 Grocery shopping
4.9 Laundry	
4.8 Health care, doctor	
4.7 Car repair	
	4.3 Other shopping
	4.2 Clean house
0 DISLIKE A GREAT DEAL	

From Robinson, J., Godbey, G., 1999. Time for Life, rev. ed. Penn State Press, State College, PA.

These 2010 SWB ratings were based on the promising pilot-study results of a Princeton University-led team of prominent social scientists (e.g., Krueger et al. 2009; Kahnemann et al. 2004). Encouraged by these results, the 2010 ATUS began enhancing respondent time-diary accounts by asking their diary-keepers how they felt, using these five Princeton-based (SWB) feelings (sad, pain, tired, stress, and happiness), in reference to while they were engaged in these specific diary activities. Thus, ATUS respondents were thus asked how they felt while doing *three* such activities (chosen at random), using scales running from 0 (no such feelings) to 6 (maximum such feeling). The distribution of their responses to these five scales is shown in Table 2, along with their overall average scores on each of the five 0–6 scales (a sixth adjective item ('meaningful') was disregarded because it did not correlate with the other five).

While Table 2 shows how differently these ATUS respondents generally felt on each of the five major SWB adjective scales, it also shows that these five disparate feelings/adjectives are still significantly related to each other, despite their covering a varying range of emotions, both positive and negative. Indeed, in the last column of Table 3, it can be seen that the five can be reduced to form a single dimension of SWB based on a factor analysis. This provides a justification for summing the five items into a single overall SWB index, potentially running from 0 (most positive on all five) to 30 (most negative on all five). These average scores were then subtracted from 13.0 to generate a higher positive score for each activity, with higher scores now reflecting a higher SWB to ease interpretation, with this average score of 3.0 in Table 4 still indicating how strongly positively these diary activities were rated. These scores are then ranked by SWB activity score in Table 4, as described in more detail below.

Table 2 Percent of 2010 ATUS respondents giving each SWB scale score

Score	*Sad (%)*	*Pain (%)*	*Tired (%)*	*Stress (%)*	*Happy (%)**	
0	75	70	30	51	5	
1	7	7	9	11	2	
2	5	6	14	12	6	
3	5	6	17	11	16	
4	3	5	19	7	19	
5	3	3	4	4	23	
6	2	3	7	4	29	
	100	100	100	100	100	
Average	0.65	0.92	2.30	1.42	5.29	
*Reverse score					1.71*	Sum = 7.0

At the same time, Table 3 indicates that these SWB intercorrelations are not that strong, being below 0.20 for two of them with (reverse-scored) happiness. Even the negative correlation between the virtual antonyms sad and happy is only −0.32, indicating the aversion most respondents felt about rating daily activities negatively (as evident in Table 2). Thus, more than 70% of respondents gave absolute zero (0) ratings to the adjectives 'sad' and 'tired' in rating their daily activities. Nonetheless, the five items do intercorrelate as well as many other accepted scales in the attitude literature (such as trust in people or anomia/alienation in the General Social Survey (GSS)). That should not discourage researchers from examining 'contradictory' activities, such as those that bring people happiness but are also stressful or tiring (e.g., Wang, 2013)).

The overall ratings shown in Table 4 thus outline and rank almost 100 diary activities according to the type of activity and from top to bottom in how they scored on the five-item SWB

Table 3 SWB Item Intercorrelations in ATUS 2010–12

	Sad	*Pain*	*Tired*	*Stress*	*Happy**	*Factor loading*	*2012 Ladder correlation*
Sad		0.37	0.27	0.49	0.32	0.75	0.26
Pain			0.34	0.34	0.17	0.67	0.23
Tired				0.37	0.17	0.62	0.20
Stress					0.32	0.77	0.26
Happy						0.54	0.31
							0.37

Table 4 ATUS positive feelings by activity

ATUS CODE	WORK	HOUSEWORK	CHILD CARE	SHOP	PERSONAL	EDUCATION	ORGANIZ'N	SOCIAL	RECREATION	MEDIA	TRAVEL	ATUS CODE
	5	2	3	9 10 11	1 4 11	6	14 15	12 13	12 13	12 16	18	
6.0							REL EDUC	PARTY				6.0
5.9								OTH ARTS				5.9
5.8			ART									5.8
5.7							REL SERVICE					5.7
5.6									SWIM			5.6
5.5				PERS CARE								5.5
5.4			PLAY									5.4
5.3								MOVIES	BASKETBALL			5.3
5.2			PHYS OKID						FISH			5.2
5.1									GOLF		RELIGION	5.1
5.0			READ TO				VOLUNTEER				PARTY	5.0
4.9									GAMES			4.9
4.8			PICK OTH KID								RECREATION	4.8
4.7		WALK PET		BANK						AUDIO		4.7
4.6												4.6
4.5										RADIO	EAT	4.5
4.4			WAIT KID						CRAFTS			4.4
4.3								SOCIAL			VOLUNTEER	4.3
4.2							REL PRACTICE			READ	ARTS	4.2
4.1									BIKE		SOCIAL	4.1
4.0			SUPERVISE						WALK		PETS	4.0
3.9									WORK OUT		GAS	3.9
3.8					EAT						OTHER SHOP	3.8
3.7								PHONE OTH			GROCERY	3.7
3.6			PICK UP								OTH ADULTS	3.6
3.5		SHOW FOOD		OTH FOOD					RUN		OTH KIDS	3.5
3.4				OTH SHOP							HH MGT	3.4
3.3			SPORT	GAS			VOL-CMPTR				FOOD	3.3
3.2			PHYSICAL								OTH KIDS	3.2
3.1		COOK	TALK/PLAN									3.1
3.0		LAWN/REPR						PHONE FAM	COMPUTER		OTHER	3.0
2.9		CAR REPR	HOMEWORK								KIDS	2.9
2.8		MAIL		GROCERY				PHONE			BANK	2.8
2.7		FEED PET								TV		2.7
2.6		EMAIL										2.6
2.5		DISHES							LIFT	SMOKE		2.5
2.4		PLAN									CLASS	2.4
2.3		OTH HOUSE									COMMUTE	2.3
2.2		CLEAN								RELAX		2.2
2.1						CLASS						2.1
2.0												2.0
1.9												1.9
1.8												1.8
1.7									CARDIO			1.7
1.6	MAIN JOB											1.6
1.5											PERS CARE	1.5
1.4												1.4
1.3												1.3
1.2												1.2
1.1		HEAT REPAIR										1.1
1.0												1.0
0.9												0.9
0.8												0.8
0.7												0.7
0.6												0.6
0.5												0.5
0.4												0.4
0.3												0.3
0.2												0.2
0.1	2nd JOB	HH FINANCE										0.1
0.0	JOB SEARCH			MEDICAL							MEDICAL	0.0

measure. This is shown horizontally across columns (from work to personal care to free time to travel activities) and vertically in ordered rows, with highest SWB ratings at the top (about 6.0) to lowest ratings at the bottom (0). More detailed interpretation of the origin and meaning of these ratings is reported in Robinson (2014).

This newly derived five-item SBW scale from ATUS ratings in Table 4 does largely replicate the general feelings about various daily activities identified in a 1985 U.S. national time study (Robinson and Godbey, 1996; Robinson, 1993) and in a 1996 U.K. national survey (Gershuny, 2012), also using a single enjoyment scale as its QOL measure. Thus, social

and visiting activity, interactive activities with children, and religious/volunteer activities top the list of favorite daily activities as rated in 'real time' as well as in general. Routine household tasks and attending to medical and other personal needs rate at the bottom. (That also tends to be the ranking reported from the U.S. and French urban community studies reported in **Table 3** of Krueger et al. 2009).

What is most surprising, important (and troubling) in these new ATUS ratings, however, is the notably lower SWB ratings given to paid work activities. This perhaps reflects a disturbing trend in life at work, jobs which many workers may endure simply to keep their basic personal and household needs afloat. These again are momentary diary ratings, but finding that these ratings are far below those in 1985 remains a concern. These low ratings are also at odds with responses U.S. workers give to general survey questions about the role and importance of work in their lives (Robinson and Godbey, 1999).

Of further concern is the below-average ratings given to Americans' most prevalent free-time activity of TV viewing. TV may serve to offset the even-lower rated activity of paid work, but it still rates notably lower than almost all other choices of a free-time activity. Earlier results from Kubey and Csikszentmihalyi (1990) suggested TV was not that much lower in enjoyment ratings than other activities, but that it was more deficient on adjectives 'challenging' and 'alertness' (not included in the present SWB ratings). Thus, these findings again suggest that engaging in more active free-time activities could be one road to improving America's collective QOL.

Many of the same conclusions about favorite activities are reached from the single question used in the Statistics Canada diary studies, as analyzed by Michelson (2011). He also found Canadian respondents unable to identify a single favorite activity on the diary day to rate themselves as lower in overall life satisfaction, a diary-overall difference discussed as our next topic.

Overall Life Satisfaction and Happiness

The above diary-based measures of momentary activity affect are rooted in a much larger and earlier survey tradition that has been asking about the public's personal QOL issues more generally, as highlighted and strengthened during the 'social indicators movement' of the 1960s, and reinvigorated by more recent attention from leading economists. Perhaps the most widely cited measure in these efforts has been the 'happiness question' first asked in the 1930s but now a feature of QOL time-series monitors like the GSS (e.g., Smith, 2011; Marsden, 2012). Here, the economists, like many other social observers, are quick to contrast the lack of increased happiness (or other positive affect) in the GSS or other trend surveys across periods of marked increases since the 1950s in economic prosperity and standards of living.

That self-rated happiness question had been intensively investigated by Bradburn and Caplovitz (1966) prior to its being asked continuously by the GSS since 1972. It simply asks respondents whether they feel 'Very, Somewhat, or Not' happy. The question has consistently been shown to correlate positively with several demographic factors, such as higher income, being married, being more educated, older and White – as well as with other subjective QOL variables, such as alienation, self-esteem, locus of control, optimism and life satisfaction, among others (Andrews and Robinson, 1991).

In one general overview study using long-term time-estimate (not diary) data from the GSS since 1972, Robinson and Martin (2008) found that happier people reported more frequent social visits, attendance at religious services, and reading of newspapers, again after demographic predictors of these activities were taken into account. In contrast, happier GSS people reported watching much less TV. These findings are generally reinforced by diary-based ratings in **Table 4**. Further clues to time or life factors more important to higher long-term QOL can be found in the pioneering surveys of Campbell, Converse ad Rodgers (1978) and Andrews and Withey (1976).

Thus, a welcome recent addition to the ATUS is the inclusion of a long-term QOL question to relate to the momentary diary-based affect measures in **Tables 1–4** This is a 'ladder' question developed by Cantril (1965), on which ATUS respondents are first asked to think of their best possible life as being a 10 (at the top of the ladder) in contrast to a 0 as the bottom rung of the ladder for their worst possible life. ATUS 2012 respondents gave an average score of 7.0 on this measure, and when their ladder score was correlated with each of the five feeling items in **Table 3**, the highest value was with the happy item ($r = 0.31$), indicating this to be the central feeling implicit in their ladder rating (although the other four adjectives are significant as well, not much predictive power is added by including the five-item SWB score in **Table 4**, which correlates 0.37 with the ladder; moreover, the disregarded 'meaningful' item correlated only 0.03 with the ladder, providing further justification for disregarding it). The important point here however is that there is a clear connection between diary and overall QOL ratings, one that does not evaporate when they are adjusted for other QOL predictors.

These studies provide evidence of a, perhaps not surprising but reassuring, connection between the general and the momentary ratings, in that generally happy people also report being happier during their diary activities, and vice versa. That may be a finding that could be simply expected or explained as a result of a survey 'response set,' but it probably has wider social and scientific implications. It is not clear which has precedent, but it suggests that engaging more in the higher rated activities in **Table 4** on a daily basis may be associated with a more satisfying life in the longer run.

Time Pressure

A general consensus seems to have emerged that the pace of U.S. daily life is speeding up (e.g., Rosa, 2005; Rosa and Scheuerman, 2010), authors who joined and expanded on conclusions from several other researchers in echoing or amplifying on the faster-paced lifestyle theme, including *Work without End* (Hunnicutt, 1998), *The Overworked American* (Schor, 1991), *Busy Bodies* (Burns, 1993), *Faster* (Gleik, 1999), *Fighting for Time* (Epstein and Kalleberg, 2004), *Work to Live* (Robinson (2003), *Take Back Your Time* (de Graff, 2003), and *Busier Than Ever* (Darrah et al., 2007). Several other prominent social scientists contributed essays further echoing or amplifying on this theme in a special issue of the journal *Social Research* on busyness edited by Mack (2005).

Despite this broad consensus about an increasing time scarcity, little empirical documentation is in evidence of its presence, let alone its increase. The American Psychological Association (APA 2012) has begun to monitor stress levels in the U.S. public on an annual basis since 2007, although with self-selected samples not generalizable to the U.S. public as a whole. The APA's 2012 report noted that extreme stress levels in their samples have steadily *declined* from 32% in 2007 to 22% in 2011. (At the same time, more than 40% of these same respondents also estimated that their stress levels had *increased* in the previous 5 years, even though that is not reflected in their personal stress ratings per se).

Earlier, Robinson and Godbey (1999) had reviewed three short measures of subjective time pressure: (1) stress questions developed and measured by the U.S. National Institutes of Health in the 1990s; (2) a 'time crunch' scale of 10 Yes–No questions; and (3) two earlier questions asking respondents how often they felt 'rushed' or had excess time on their hands. The latter had the advantage of having been asked across several decades starting in 1965, asking the two main questions in **Tables 5** and **6** to capture these feelings. One asked how rushed respondents generally felt, and the others how much excess time they felt they had (Robinson and Godbey, 1999). These have been replicated in several subsequent national surveys, with **Table 5** showing almost a 10-point increase in feeling rushed during the 1990s, but a return to almost the same 1965 levels of feeling rushed in 2010. This decrease was found despite the recent proliferation of new time-intensive IT technologies (Robinson, 2013). While Americans aged 18–64 years had increased their feelings of being 'always' rushed, from about 25% in 1965 to about 35% in the 1980s and 1990s, this trend came to a decided halt in two surveys in 2009–10 (Robinson, 2013).

This article centers on the final two columns in **Table 5**, updated readings from: (1) the 2010 GSS survey and (2) a national 2009 omnibus survey conducted at the University of North Florida (UNF). These updated **Table 5** data are summarized graphically in **Figure 1**.

This **Table 5** trend is thus quite counter to the societal consensus that the pace of daily life is speeding up, as proposed by Rosa and other authors. Contrary to such expectations, both of these most recent surveys now show decreases in percents feeling always rushed among those in the 18–64 age-group, a 6-point drop to 28% in the 2009 UNF survey and a larger 9-point drop to 25% in the 2010 GSS survey. **Figure 1** and **Table 5** also show that responses to the second subjective time question on having excess time (in terms of 'never' having time on one's hand and not knowing what to do with it) had remained fairly constant. These declines thus put the 'always rushed' levels in 2009–10 at about where they were in the 1970s. The declines were not simply due to unemployment, as they were found among employed respondents as well.

Overall, then, the two sets of results are rather surprising, given the plateau of readings at about 35% 'always rushed' levels since the 1980s, and in light of all the new time-intensive IT technologies and the 2007 economic crises. In

Table 5 TRENDS in subjective time responses by year

Year–>		*1965*	*1971*	*1975*	*1976*	*1982*	*1985*	*1995*	*1996*	*1998*	*2004*	*2009*	*2010*	*2009*	*2010*
Age 18–64		*UMich*	*UMich*	*UMich*	*UMich*	*GSS*	*UMD*	*UMD*	*GSS*	*UMD*	*GSS*	*UNF*	*GSS*	*UNF*	*GSS*
A. Rushed	*n*=	1134	8888	1234	3102	1581	2424	988	1260	920	855	739	788	739	788
1. Always		24%	27	27	20	27	35	34	33	35	34	28	25	27	25
2. Sometimes		53%	50	52	61	56	50	56	56	56	55	59	60	58	60
3. Almost never		23%	23	21	19	17	15	10	11	9	11	13	15	15	15
B. Time on hands	*n*=870		1999	1111	2479	1150	NA	658	NA	NA	569	634	NA	639	NA
1. Quite often		15%	14	12	8	14	NA	7	NA	NA	10	9	NA	9	NA
2. Now and then		37%	41	42	44	42	NA	38	NA	NA	41	46	NA	46	NA
3. Almost never		48%	45	46	48	44	NA	55	NA	NA	49	45	NA	45	NA
All adults (18+)			UMich 1971	UMich 1975	UMich 1976	GSS 1982	UMD 1985	UMD 1995	GSS 1996	UMD 1998	GSS 2004	UNF 2009	GSS 2010	UNF 2009	GSS 2010
A. Rushed	*n*=		9519	1513	3590	1852	2976	1199	1452	975	986	994	997	994	997
1. Always		NA	22	24	18	25	32	29	30	35	31	27	22	27	22
2. Sometimes		NA	51	50	58	53	48	54	53	55	54	58	57	5 8	57
3. Almost never		NA	27	26	24	22	20	17	17	10	15	15	21	15	21
B*. Time on hands	*n*=		2261	2222	2927	1404	NA	848	NA	NA	682	729	NA	729	NA
1. Quite often		NA	15	13	9	16	NA	9	NA	NA	11	9	NA	9	NA
2. Now and then		NA	37	40	43	42	NA	35	NA	NA	41	46	NA	46	NA
3. Almost never		NA	48	47	48	42	NA	56	NA	NA	48	45	NA	45	NA

Rushed question: "Would you say you always feel rushed, even to do the things you have to do, only sometimes feel rushed, or almost never feel rushed?"

Time on hands question: "How often would you say that you have time on your hands you do not know what to do with quite often, only now and then, or almost never?"

*As asked only of those not 'always' rushed.

Table 6 Happiness by subjective time (% very happy)

		1976 UM	*1982 GSS*	*1996 GSS*	*2009 UNF*	*2010 GSS*
Rushed	*n*=	[3662]	[1848]	[1450]	[990]	[982]
1. Always		28	29	27	25	24
2. Sometimes		33	35	33	38	24
3. Almost never		41	39	37	41	35
DIFF 3−1		+13	+10	+10	+16	+11
Time on hands	*n*=	[3002]	[1401]	NA	[990]	NA
1. Quite often		20	24	NA	24	NA
2. Now and then		33	34	NA	31	NA
3. Almost never		40	48	NA	39	NA
DIFF 3−1		+20	+24	NA	+15	NA

addition to these decreases in societal subjective time pressure, some of their other QOL implications are surprising as well.

QOL Implications of Subjective Time

These two subjective time attitudes have also been found to have implications for more direct measures of a person's QOL, with the percent of 'very happy' responses used as the major indicator of one's QOL. The 1972–2012 trends in this very happy category have proved remarkably steady over this 40-year period, averaging about 33% across the decades. Table 6 shows that these 'very happy' responses have also been consistently and significantly related to both subjective time questions, both in previous GSS surveys (1982, 1996, and 2004), as well as in the 1976 pioneering study of Campbell, Converse and Rodgers (1978). Thus, in the 1982 GSS, only 29% of those always feeling rushed said they were very happy, compared to 35% of those sometimes feeling rushed and 39% who felt 'almost never rushed,' a regular increase of 10 percentage points. Similarly in 1996, the difference was from 27% to 33% and to 37%, again a regular progression of 10 points. As shown in Table 6, in Campbell, Converse ad Rodgers (1978) the progression was a slightly greater 13 points (from 28% to 33% to 41%). Much the same pattern is found in both 2009–10 surveys in Table 6. Both data sets were poststratified to match Census Bureau demographic figures on gender, age, race, marital status, education, and income. (Figure 2)

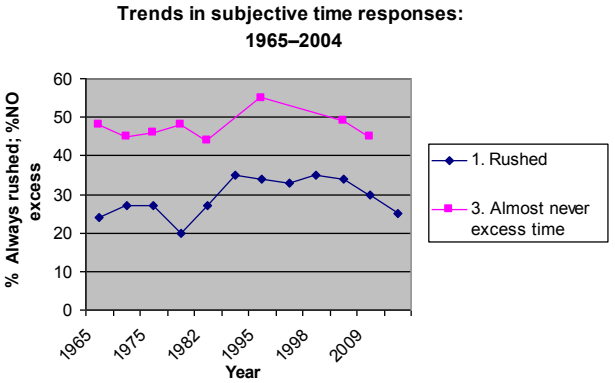

Figure 1 Trends in 'always rushed' responses: 1965–2010.

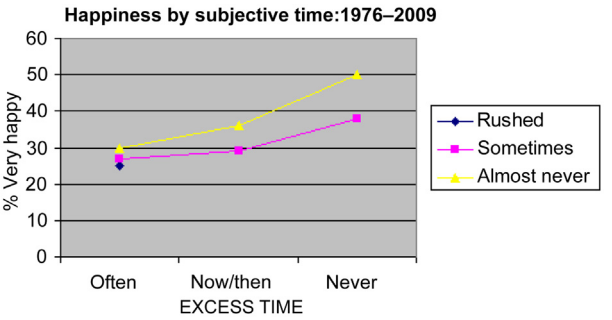

Figure 2 Happiness levels for combined time attitudes 1976–2010.

Further support for Table 5 comes from a question concerning whether Americans feel more pressure from this factor of time or from its competing resource of money. This question has shown a marked recent swing since 1998, when the majority (52%) of Americans said that lack of time was of more concern to them than lack of money. That dropped to 33% in 2013, a decline that was reported across all age, education, and income groups, not just the unemployed. As expected, and as in earlier surveys, it was the case that those feeling lack of time more than lack of money did have notably higher incomes, but even higher income groups reported the recent increase in choosing lack of money over lack of time.

Stress

Given the great medical concern and societal attention paid to the stressful lifestyles in America, it is surprising that there are so little data using standardized trend questions on stress from representative national samples. The National Center for Health Statistics did include a direct question on perceived stress in its very large 1985 to 1998 surveys, and it did find a significant rise from 50% to 56% between 1986 and 1993, but then a surprising decline to 48% in 1995. Unfortunately, this question, that had the same pattern of correlates as those in Table 5, appears not to have been asked since then to track more recent levels and correlates of stress.

Time Crunch Scale

Robinson and Godbey (1999) reported results from a 1993 nonprobability national survey that developed a 10-item measure of feelings of time pressure, with an average of about 33% of respondents agreeing with these items (about the same proportion as who describe themselves as always rushed). As with the rushed and stress questions, women and those aged 25–44 years were most likely to respond being under more time crunch. The scale has not been replicated in the United States, but it has been repeated in Canadian national diary studies, as well as in other countries (Michelson, 2011).

Other Questions

Another single item asked respondents in that 1993 survey whether they would be willing to give up a day's pay to get a day off from work, to which nearly 50% said 'Yes,' and that

proportion has remained at about that level in subsequent U.S. surveys. On the other hand, using a more balanced question, only 13% of respondents said they would like to work fewer hours (compared to 35% who wanted more work – and 52% who said the same work hours). That latter result is in keeping with the small minority of 18% of workers in both 1975 and in 1998 who reported being 'very tired' at the end of a workday, here suggesting few workers face little overall time problem with their workload.

Time pressure concerns are also evident in questions asking respondents whether they have more or less free time than 5 years previously, and here far more respondents say 'less' free time than 'more.' In family surveys, when asked whether they spend enough time with their children, most parents say 'not enough,' even though mothers (but especially fathers) are spending more time with their children than 10 or 20 years ago (Bianchi et al., 2006). Parallel questions about time with one's spouse or by oneself drew similar or higher levels of 'not enough time' responses. Other GSS questions have asked perceptions of how well balanced one's work and family lives were. Robinson and Godbey (1999) found that less than 30% of wives said they wished their husbands would help them more with the housework or with child care, and that has not changed much since then.

We turn next to a different type of data, one that attempts to provide greater insight into the results of these time measures, and their QOL implications.

Outcomes from Time Inputs

Here, the concerns are about the productivity or consequences of various ways of spending time. For example, does one feel that the hours devoted to an activity have any parallel in increased output? Thus, one set of survey questions has asked parents about how well their children are doing in life, about their satisfaction with the amount of time the family spends together, or about awareness of where their children are at times when they are not with them (Bianchi et al., 2006). Other survey items have asked how satisfied respondents are with the cleanliness of their homes (Robinson and Milkie, 1998) or the quality of the family meals (Juster and Stafford, 1985).

A few studies have examined time's relations with more objective measures of output. Schuman et al. (1985) found little relationship between study time and grades in school. Robinson and Godbey (1999) similarly found little relation between time spent on housework and respondent satisfaction with household cleanliness, or on ratings of that cleanliness either by the respondent, or by the interviewer who observed the dwelling. Nor did they find that presence of household appliances helped reduce housework time, perhaps because it is the busier and more ambitious families that buy the technology in the first place, as found earlier in Morgan et al. (1966).

Some more promising economic advances in linking time inputs and outputs or societal welfare have been recently proposed by Bittman and Ironmonger (2011), who presented a typology outlining several options to valuing time by using a person's simple wage rate. Folbre (2009) elaborated ways in which time-use data can improve both subjective and objective measures of living standards, pointing out fundamental flaws in the ways conventional economics puts values on different activities. Hamermesh (2009, 2013) has proposed other new ways economists need to use more economic theory to guide their analyses of time.

Alternatively, Goodin et al. (2008) asserted how time is an inherently superior metric of welfare than money, reversing the usual process of valuing time in dollars and instead expressing money in terms of the time required to earn it. Their arguments are built around the central idea that the *ability to choose* how to allocate one's time lies at the core of a positive notion of freedom. The authors developed a new standard of welfare from the idea of escape from the 'realm of necessity' by defining 'discretionary time' as what is left after necessary minimum time that must be devoted to personal care, paid work, and nonmarket work. Discretionary time thus represents an advance over the more conventional category of 'free time' as a new metric of welfare.

On the negative end of the welfare scale, there is the concern about the possible negative impacts of time spent eating. Zick and Stevens (2011) report that, while time spent eating as a primary activity has remained relatively unchanged over previous decades, eating time as a secondary activity has risen significantly – particularly as an accompaniment to TV viewing. In Australia, Brown et al. (2011) presented evidence that adolescent TV viewing is accompanied by increased consumption of high-calorie snack foods. The ATUS has begun to ask new questions relating time use with food consumption and fitness activity to examine further linkages between daily activity and obesity (Janes et al. 2011).

See also: Cultural Participation, Trends in; Family and Work; Gender, Time-Use, and Fertility Recovery in Industrialized Countries; Leisure and Cultural Consumption: The European Perspective; Leisure and Cultural Consumption: US Perspective; Time Use Research Methods; Time Use Research: Recent Developments; Time Use and Gender; Time Use and Resource Consumption.

Bibliography

American Psychological Association, 2012. Stress in America: 2011. American Psychological Association, Washington, DC.

Andrews, F., Robinson, J., 1991. Happiness and life satisfaction. In: Robinson, J., Shaver, P., Wrightsman, L. (Eds.), Measures of Personality and Social Psychological Attitudes. Academic Press, New York.

Andrews, F., Withey, S., 1976. Social Indicators of Well Being. Plenum, New York.

American Time-Use Survey (ATUS) http://www.bls.gov/atus *www.bls.gov/atus* ATUS-X.

Bianchi, S., Robinson, J., Milkie, M., 2006. Changing Rhythms of American Family Life. Russell Sage, New York.

Bittman, M., 2011. Goodin et al.'s new measure of freedom: authors meet critic. Social Indicators Research 101 (2), 185–187.

Bittman, M., Ironmonger, D., 2011. Valuing time: a conference overview. Social Indicators Research 101 (2), 173–183.

Bradburn, N., Caplovitz, D., 1966. Reports on Happiness. Aldine, Chicago.

Brown, J., Nicholson, J., Broom, D., Bittman, M., 2011. Television viewing by school-age children: associations with physical activity, snack food consumption and unhealthy weight. Social Indicators Research 101, 221–225.

Burns, L., 1993. Busy Bodies: Why Our Time Obsessed Society Keeps Us Running in Place. W.W. Norton, New York.

Campbell, A., Converse, P., Rodgers, W., 1978. The Quality of American Life. Russell Sage, New York.
Cantril, H., 1965. The Pattern of Human Concerns. Rutgers University Press, New Brunswick, NJ.
Darrah, C., Freeman, J., English-Lueck, J., 2007. Busier than Ever! Why American Families Can't Slow Down. Stanford University Press, Stanford, CA.
Diener, E., Tay, L., 2014. Review of the Day Reconstruction Method (DRM). Social Indicators Research 116, 255–267.
Economist, 2010. The Great American Slowdown. April 10, 2008.
Epstein, C., Kalleberg, A. (Eds.), 2004. Fighting for Time: Shifting Boundaries of Work and Social Life. Russell Sage Foundation, London.
Folbre, N., 2009. Time use and living standards. Social Indicators Research 93 (1), 77–83.
Frey, W., 2010. The Great American Migration Slowdown. Brookings Research Report, Washington, DC, 12/9.
Gershuny, J., 2009. Comment on Krueger and colleagues. Social Indicators Research 93, 23–26.
Gershuny, J., 2012. National utility: measuring the enjoyment of activities. European Sociological Review 10 (1093), 1–17.
Gleik, J., 1999. Faster. Vintage Books, New York.
Goodin, R.E., Rice, J.M., Parpo, A., Eriksson, L., 2008. Discretionary Time: A New Measure of Freedom. Cambridge University Press, Cambridge.
de Graff, J., 2003. Take Back Your Time. Berrett-Koehler Publishers, San Francisco.
Hamermesh, D., 2009. It's time to do economics with time-use data. Social Indicators Research 93 (1), 65–68.
Hamermesh, D., 2013. How do immigrants spend time? Journal of Population Economics 26, 137–153.
Hunnicutt, B., 1998. Work without End: Abandoning Shorter Hours for the Right to Work. Temple University Press, Philadelphia.
Jacobs, J., Gerson, K., 2004. The Time Divide: Work, Family and Gender Inequality. Harvard University Press, Cambridge.
Janes, M., Hamrick, K., Lacey, J., 2011. Exercise, eating patterns, and obesity: evidence from the ATUS and its eating & health module. Social Indicators Research 101 (2), 215–219.
Juster, T., 2009. Some perspectives on the study of time use. Social Indicators Research 93, 19–21.
Juster, T., Stafford, F., 1985. Time, Goods, and Well-Being. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.
Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: the day reconstruction method. Science 306 (5702), 1776–1780.
Kreuger, A., et al., 2009. Comparing time use and subjective well being in two countries. Social Indicators Research 93, 1–17.
Kubey, R., Czikszentmihalyi, M., 1990. Television and the Quality of Life. Earlbaum, Hillsdale N.J.
Mack, A., 2005. Editor's introduction. Social Research 72 (2), 263.
Marsden, P., 2012. Social Trends in American Life. Princeton University Press, Princeton, NJ.
Michelson, W., 2009. Adding affect to time diary accounts. Social Indicators Research 93, 31–32.
Michelson, W., 2011. What makes an activity most enjoyable? Social Indicators Research 103 (1), 77–91.
Morgan, J., Sirageldin, I., Baerwaldt, N., 1966. Productive Americans: A Study of How Individuals Contribute to Economic Progress. Institute for Social Research, University of Michigan, Ann Arbor.
Parkinson, C., 1957. Parkinson's Law, and Other Studies in Administration, vol. 24. Houghton Mifflin, Boston.
Robinson, J., 2003. Work to Live: The Guide to Getting a Life. Perigee Books, New York.
Robinson, J., 2013a. Americans less rushed, but no happier: 1965–2010 trends in subjective time and happiness. Social Indicators Research 113 (3), 1091–1104.
Robinson, J., 2013b. Happiness is being just rushed enough. Scientific American. Online February.
Robinson, J., February 1993. As we like it. American Demographics 44–48.
Robinson, J., 2014. As We (Still) Like It: Socializing, Religion, Kids Remain Our Favorite Daily Activities. Working paper No. PWP-MPRC-2013-024. Maryland Population Research Center. http://papers.ccpr.ucla.edu/papers/PWP-MPRC-2013-024/PWP-MPRC-2013-024.pdf (accessed 31.01.14.).
Robinson, J., Godbey, G., 1999. Time for Life, rev. ed. Penn State Press, State College, PA.
Robinson, J., Godbey, G., 1996. The great American slowdown. American Demographics 18 (6), 42–47.
Robinson, J., Martin, S., 2010. Sex, arts and verbal abilities: three further indicators of how American life has not improved. Social Indicators Research 99 (1), 1–12.
Robinson, J., Martin, S., 2008. What do happy people do? Social Indicators Research 89, 565–571.
Robinson, J., Milkie, M., 1998. Back to the basics. Journal of Marriage and the Family 98, 205–218.
Rosa, H., 2005. Acceleration. Die Veränderung der Zeitstrukturen in der Moderne. Suhrkamp, Frankfurt am Main. Zugleich Habil.-Schrift, Univ. The Change of Time Structures in the Modern Era. Suhrkamp, Frankfurt am Main.
Rosa, H., Scheuerman, W., 2010. High-Speed Society: Social Acceleration, Power, and Modernity. Penn State Press, State College, PA.
Schor, J., 1991. The Overworked American: The Unexpected Decline of Leisure. Basic Books, New York.
Schuman, H., Walsh, E., Olson, C., Etheridge, B., 1985. Effort and reward: the assumption that college grades are affected by quantity of study. Social Forces 63 (4), 945–966.
Smith, T., 2011. Trends in Well-Being: 1972–2010. General Social Survey. National Opinion Research Center, University of Chicago.
Wang, W., Oct. 8 2013. Parents' Time with Kids More Rewarding. Pew Research Center, Washington, DC.
Zick, K., Stevens, R., 2011. Time spent eating and its implications for Americans' energy balance. Social Indicators Research 101 (2), 267.

Time, Anthropology of

Eric K Silverman, De Pauw University, IN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15683–15686, © 2001, Elsevier Ltd.

Abstract

Sociocultural constructions of time and time-reckoning are central topics of anthrological inquiry. Many cultures lack an explicit category 'time.' Still, time is everywhere present, in numerous local figurations that may or may not evidence a few universal forms. This article offers a representative review of time cross-culturally that focuses on some of the important concepts, debates, and intriguing new areas of investigation. The article, by discussing possible universal configurations of time – e.g., human lifecycle and bodily processes, seasonality, celestial patterns, binary constructions, and the opposition between incremental/linear/irreversible and episodic/cyclical/repetitive forms of time. Next, some traditional anthropological ideas about time are discussed: time categories and conventions are 'social representations' that mirror the organization of society (e.g., kinship groups), the political manipulation of time, the existence of multiple or layered forms of time which are not necessarily integrated into a single unifying system, ritual versus mundane time, and the linguistic relativity of time. Finally, some recent topics of concern are addressed, specifically, the spatialization of time, embodied time, gendered time, and the 'end' of time.

Sociocultural constructions of time and time reckoning are now central topics of anthropological inquiry. Many cultures lack an explicit category 'time.' Still, time is everywhere present, in numerous local figurations that may or may not evidence a few universal forms, and thus the anthropological spectrum of research, theories, and contentions is vast (Gell, 1992; Munn, 1992). I offer here a representative, nonprogrammatic review of the topic that focuses on some of the important concepts, debates, and intriguing new areas of investigation.

Universal Time

Certain natural phenomena appear to be universally temporalized in terms of periodicity and, in some but not all cultures, progressive quantification: the human lifecycle and bodily processes (e.g., menstruation), seasonality, celestial patterns, day/night, and so forth. The same is true for the reproduction of social order by generation and various social and kinship groups. All societies regularly coordinate labor, occasionally punctuate the tempo of everyday life with ritual, and envision some type of past, present, and future. Calendrical arrangements of days, weeks, months, and years are also universal – but, again, not everywhere counted. In anthropological theory, time is implicitly understood to be binary. For example, all cultures accommodate conceptual categories and linguistic terms for duration and sequence, which are often said to be the basic forms of time. Leach (1961) proposed that time is a social construction rooted in two basic human experiences: the repetitions of nature, and the irreversibility of life. Many anthropologists contend that all societies possess both incremental/linear/irreversible and episodic/cyclical/repetitive forms of time – which are not necessarily antithetical since time cycles can return to the 'same logical, not temporal, point' (Howe, 1981). While Farriss (1987) claims that any one mode can incorporate the other, Leach proposes that all religions deny the finality of human mortality by subsuming linear time under a cyclical framework. Birth and death become two phases of an eternal sequence. Leach also claimed that time is everywhere a 'sequence of oscillations between polar opposites.' This temporal pendulum is related to basic sociological processes such as reciprocal gift-exchange and marriage. More recently, Gell (1992) argued that all culture-specific patterns of temporality are variations of two cognitive modes: an A-series which orders events according to relative notions of pastness, presentness, and futurity, and a B-series of absolute before/after.

Social Time

Universal conjectures notwithstanding, most anthropologists focus on local notions of time and subscribe to the Durkheimian thesis that time categories and conventions are 'social representations' that mirror the organization of society. This sociological perspective is best exemplified by Evans-Pritchard's (1939) elegant study of the East African Nuer, which is perhaps the most widely-known anthropological account of a distinctly non-Western style of time. Nuer 'oecological time,' which is cyclical or seasonal, reflects the environmental rhythms of cattle pastoralism. More importantly, Nuer indicate time but they lack an abstract chronological system that is numerical, cumulative, uniformly incremental, and measurable. Beyond the year, Nuer speak about temporal duration, intervals, and sequences with sole reference to 'structural time' or significant genealogical events in the social system – say, the division of a lineage, or the birth of an age set. Since Nuer genealogies are fixed at a depth of five generations, too, the passage of time is contained within a static framework. Similarly, there is considerable debate over whether or not the terminological groupings and repetitions in many kinship systems freeze, reverse, or emphasize the passage of time (Silverman, 1997). Here, too, the opposition between cyclical and linear time is a leitmotif of anthropological analysis.

Cultural Time

The Durkheimian approach to time, enshrined in British social anthropology, anchors temporal conceptions to social organization. An alternative approach to time, represented by the 'interpretivist' tradition of American cultural anthropology, highlights the symbolic dimensions of time. Here, analysis seeks to discern the relationship between general temporal attitudes and the overall emotional emphases and focal symbols of a cultural worldview. Geertz (1973) suggested that the cyclical structure of Balinese kinship, naming, and calendrical systems 'don't tell you what time it is; they tell you what kind of time it is.' In a celebrated passage, Geertz concluded that 'Balinese social life takes place in a motionless present, a vectorless now' in which 'time lacks motion because Balinese social life lacks climax.' Moreover, the Balinese 'immobilization of time' is linked to the 'anonymization of persons.' Ultimately, the formal complexity of Balinese calendars pertains not to social order but to the unique, somewhat aesthetic themes and tones of Balinese culture and personhood. Whereas Evans-Pritchard pioneered the study of how others think about temporality, Geertz has been an inspiration to anthropologists who seek to understand how others experience time.

Political Time

One key debate in the anthropology of time, then, concerns the locus of analysis – broadly speaking, society vs. culture – and the question of whether or not time configurations can be reduced to a more elemental, causal factor such as human ecology. Another quarrel arises from the materialist division between infrastructure and superstructure. Bloch (1977) criticized Geertz for failing to distinguish between ritual time, which does differ significantly crossculturally, and universally cognized 'normal' time that organizes practical activities. Hierarchical societies, Bloch continued, tend to elaborate a ceremonial view of timeless time. Consequently, ritual concepts of time are ideologies that sustain everyday productive inequalities. But while Geertz, in this view, may be guilty of foregrounding difference or 'radical temporal alterity,' and thus neglects the politics of time, Bloch himself assumes that people everywhere subscribe to the same 'practical reason,' which remarkably resembles the economic concerns of Western Marxism. Many anthropologists emphasize the lability of time concepts and, more recently, the political manipulation of time. The Northern Thai calendar, which is oriented around a 'fatal quality of time,' enables post hoc explanations of misfortune that exonerate individuals from personal blame (Davis, 1976). The Kodi of Indonesia, for whom calendrical knowledge is prestigious, manipulate temporal sequences and durations, question the timing of ceremonial feasts and other public events, and strategically stage disruptions (Hoskins, 1993). The Kodi calendar is also a 'debating ground' as traditional forms of time now clash with exoteric calendars associated with the state, Christianity, and modernity. Multiple, or layered, forms of time which are not, moreover, necessarily integrated into a single unifying system are common. One temporal series may regulate markets, while another system organizes lunation. Time is often contextual – as well as competing and seemingly contradictory. Kabyle precapitalist peasants of Algeria submit to the present but exalt foresight within the temporal horizon of the 'forthcoming.' They also view the future – which capitalist hubris, in this view, seeks to predict – as a void known only to God (Bourdieu, 1963). A final note on the politics of time: when competing groups anchor their prestige in a politico-ritual system to ancestral primacy, the mere assertion that a temporal series, or time itself, has a specific beginning point or location is inevitably political since it denies the validity of rival times. This is true for small-scale societies as well as the contemporary United States, which remains embroiled in debate between evolution and creationism.

Ritual Time

Not only is the differentiation of ritual from mundane time important to many anthropologists, but so is the construction of time within ritual itself. Humans tend, wrote Leach (1961), to conceptualize time as a repeating series of different intervals. Each year is distinctive but they all begin and end with an identical festival. Significant calendric events are everywhere marked by ceremonial antics and attire that are censored from everyday life. These periodic ritual inversions or 'rites of passage' actually create time by introducing intervals in social life. Yet Barnes (1974: pp. 128–129) offered the opposite view for the Kédang of Indonesia: ritual insures precisely that time is not cyclical – that the dead, for example, are irreversibly dead, and that calendric cyclicity is unable to prevail over linear temporality. From a different perspective, Gell (1992: pp. 37–53) recently evoked Lévi-Strauss and argued that the ritual reversal or regression of normal time among the Umeda of Papua New Guinea resolves problems that arise from the imposition of 'cultural' conventions such as temporal classifications onto the 'real' or 'natural' world.

Language of Time

Unlike spatial deixis, the linguistic study of time remains sparse in anthropology. Yet the famous Sapir–Whorf hypothesis of linguistic relativity was initially popularized through the alleged absence of Hopi verb tenses (now disproved). Still, language shapes our experience of the world, including time. Hence, the Trobriand Islanders of Papua New Guinea are unable to discuss and conceptualize linear, chronological sequences of events (Lee, 1950). They have no time line but only temporal points organized as noncausal patterns with 'at most a swelling in value.' Devoid of linguistic temporal connections, Trobriand events are self-contained: 'a series of beings but not becoming.' Related studies of language and time include past vs. future-oriented locutions (e.g., Dundes, 1969), orality vs. literacy (Goody, 1991), and cognitive 'maps' of spatiotemporal passage (Frake, 1985).

Spatial and Bodily Time

The spatialization oftime and the temporality of space are topics of growing anthropological interest (see Munn, 1992).

Many cultures – e.g., Melanesia and Aboriginal Australia (Silverman, 1997) – understand the primordial creation of the universe through a series of ancestral migrations, radiating outwards from an origin place, that simultaneously generated space and time. But other cultures possessing a temporal topology, such as the Iraqw of Tanzania, lack an origin: 'there is no zero point against which to set a regularized chronology' (Thornton, 1980: p. 184). Spatiotemporal mythologies occur across lowland South America, where time and space are also embodied. The passage of time in the external world – the flow of a river, the movement of the sun – represents the human life-cycle. Conversely, the temporal unfolding of physiological processes such as sexuality and alimentation are projected onto the local topography (e.g., Hugh-Jones, 1979: Chap. 7). The experience of time is thus understood through a bodily schema, specifically, the circulation of substances into and out of the body, and wider cycles of birth and death.

Gendered Time

The relationship between gender and time is another promising area of anthropological study. The riverine Eastern Iatmul of Papua New Guinea associate water with uterine fertility, death, and the inexorable erosion of land and trees that threatens to regress the cosmos back to the timeless and feminine era that preceded male creation (Silverman, 1997). Like the river, female time flows to uncertainty. Male time, often conceptualized as the planting of trees and the establishment of villages, progresses forward against the flow of female temporality. But the arboreal ascendance of masculine structure is illusory. The river and rainy season floods ultimately prevail. All men can do to stem the feminine tide of regressive time is plant trees, build houses, and enact rituals. Weiner (1991) also discusses the gender of time in a Melanesian society, and links temporality to linguistic iconicity and motions through the landscape and body.

The Future and End of Time

Anthropologists are no longer content merely to bemoan the end of the 'time of the primitive.' Nor do we continue to define non-Western 'Otherness' as an essentially temporal construct, communities who are removed from us in both space and time (Fabian, 1983). Instead, anthropologists are increasingly interested in historical conjunctions and contemporary practices that blur the distinctions between local and global times. Through temporal ambiguities, local people exercise creativity and resistance. While all societies allocate time between different activities (Gross, 1984), the spread of global capitalism and wage-labor has imposed on people everywhere computational, monetary, and aggressive notions of temporality such as saving, wasting, and fighting against time. But the future of time, to invoke a common idiom, is not confined to 'time is money.' The millennial turn encouraged many communities to confront the historical dissonances of culture contact by rethinking conventional notions of past, present, and future, cyclical and linear time, tradition and modernity, and the relationship between the biblical apocalypse and ancestral spirits. The millennium was a time to ponder the end of time (Scaglion, 2000) – as well as, in the context of Y2K, the opposition between 'virtual time' and 'real time.' Thus the opening decades of this millennium, including its end-of-time scenarios, will doubtless prove to be a propitious era for the anthropological study of time both globally, locally, and in the newest temporal zone: cybertime.

Bibliography

Barnes, R.H., 1974. Kédang: A Study of the Collective Thought of an Eastern Indonesian People. Clarendon, Oxford, UK.

Bloch, M., 1977. The past in the present in the past. Man (N.S) 12, 278–292.

Bourdieu, P., 1963. The attitude of the Algerian peasant toward time. In: Pitt-Rivers, J. (Ed.), Mediterranean Countrymen: Essays in the Social Anthropology of the Mediterranean. Mouton, Paris, pp. 55–72.

Davis, R., 1976. The Northern Thai calendar and its uses. Anthropos 71, 3–32.

Dundes, A., 1969. Thinking ahead: a folkloristic reflection on the future orientation in American worldview. Anthropological Quarterly 42, 53–72.

Evans-Pritchard, E.E., 1939. Nuer time reckoning. Africa 12, 189–216.

Fabian, J., 1983. Time and the Other: How Anthropology Makes Its Object. Columbia University Press, New York.

Farriss, N.M., 1987. Remembering the future, anticipating the past: history, time, and cosmology among the Maya of Yucatan. Comparative Studies in Society and History 29, 566–593.

Frake, C.O., 1985. Cognitive maps of time and tide among medieval seafarers. Man (N.S.) 20, 254–270.

Geertz, C., 1973. Person, time, and conduct in Bali. In: The Interpretation of Cultures. Basic Book, New York, pp. 360–411 (Originally 1966).

Gell, A., 1992. The Anthropology of Time: Cultural Constructions of Temporal Maps and Images. Berg, Oxford, UK.

Goody, J., 1991. The time of telling and the telling of time in written and oral cultures. In: Bender, J., Wellbery, D.E. (Eds.), Chronotypes: The Construction of Time. Stanford University Press, Stanford, CA, pp. 77–96.

Gross, D.R., 1984. Time allocation: a tool for the study of cultural behavior. Annual Review of Anthropology 13, 519–558.

Hoskins, J., 1993. The Play of Time: Kodi Perspectives on Calendars, History, and Exchange. University of California Press, Berkeley, CA.

Howe, L.E.A., 1981. The social determination of knowledge: Maurice Bloch and Balinese time. Man 16, 220–234.

Hugh-Jones, C., 1979. From the Milk River: Spatial and Temporal Processes in Northwest Amazonia. Cambridge University Press, Cambridge, UK.

Leach, E., 1961. Two essays concerning the symbolic representation of time. In: Rethinking Anthropology. Athlone, London, pp. 124–136.

Lee, D., 1950. Lineal and nonlineal codifications of reality. Psychosomatic Medicine 12, 89–97.

Munn, N.D., 1992. The cultural anthropology of time: a critical essay. Annual Review of Anthropology 21, 93–123.

Scaglion, R., 2000. Zero hour: reflecting backward, looking forward. Ethnohistory 47, 227–240.

Silverman, E.K., 1997. Politics, gender, and time in Melanesia and Aboriginal Australia. Ethnology 36, 101–121.

Thornton, R.J., 1980. Space, Time, and Culture Among the Iraqw of Tanzania. Academic Press, New York.

Weiner, J.F., 1991. The Empty Place: Poetry, Space, and Being Among the Foi of Papua New Guinea. Indiana University Press, Bloomington, IN.

Time, Chronology, and Periodization in History

Masayuki Sato, University of Yamanashi, Kōfu, Yamanashi, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Homo sapiens is the only species on Earth that is equipped to look back on its past consciously. The past, however, is chaotic. Nevertheless, it is the peculiar ability of humankind, as *Homo historicus*, to put some kind of order on this chaos, making it understandable. It is this conscious looking back to the past that identifies *H. sapiens* as human. The most important element in bringing order to the chaos of the past is a concept of time. Time is a human invention rendering chaos intelligible, rather than something that exists in itself. Among various notions of time, chronology and periodization are the most closely connected with organizing of our past. Without these conceptual inventions, there would be no way of organizing our past as history.

Homo sapiens is the only species on Earth that is equipped to look back on its past consciously. The past, however, is chaotic. Nevertheless, it is the peculiar ability of humankind, as *Homo historicus*, to order this chaos, making it understandable. It is this conscious looking back to the past that identifies *H. sapiens* as human.

The most important element in bringing order to the past is a concept of time. Rather than positing time as existing in itself, it may profitably be viewed as a transcendental condition that renders the largely unstructured flux of the past intelligible. Among the various notions of time, chronology, and periodization are the most closely connected with the organization of our past. Without these conceptual inventions, there would be no way of organizing our past as history.

Chronology is a dating system, or a system of reckoning a series of years from a starting point (year), which gives an era a theoretical definition. In this respect, chronology should be distinguished from a calendar, which is defined as a system of fixing the year, month, and date (Blackburn and Holford-Strevens, 1999).

Periodization is a form of historiological cognition proposed by the human mind for making the past intelligible and meaningful by dividing the collective past into compartments along time (Croce, 1916(1960)).

On the one hand, chronology is different from periodization in that it is used by a group of people for practical purposes, mainly for counting years. Those who try to make sense of history through rendering the past retrospectively intelligible, on the other hand, mostly use periodization. It is of no practical use for counting time.

Chronology, therefore, exists in these terms mainly for constructing time, and periodization for dividing time. However, further inquiry reveals that chronology also acts to divide time by creating a starting point for a new era, e.g., regnal years, Christian chronology, and East Asian era names. On the other hand, the very act of periodizing of the past effectively constructs a new time by inventing an arbitrary starting point, though these times are not used by people in everyday life, as for example the 'Victorian period' or the 'Middle Ages'.

Four Aspects of Time in History

Time in history will be considered from the following four viewpoints: the three types of time; the relation between linear time and cyclical time; the relation between what I call the stream of time and the stack of time; and finally, what governs time.

Three Types of Time

Time can be classified into three categories: natural time, human time, and metaphysical time. Natural time is a reflection of natural cycles as seasons, human life, etc. Human time is what *H. sapiens* has invented to make the remembered past intelligible and comprehensible, and consists of such schemes as chronology and periodization. Metaphysical time, as proposed by philosophers of history, is what *H. sapiens* has created to make the remembered past meaningful beyond the experience of a single human.

Linear Time and Cyclical Time

Among Western historians, time in history has been discussed in terms of the established dichotomy of linear and cyclical time. There is a consensus among them that a culture based on nonlinear time would not be able to construct historical thinking. India, for example, has sometimes been described as being without a 'historical mind', because it was a culture that lacked linear time and lived in cyclical time. That linear time that is necessary for the rise of the historical mind has been accepted by a large number of European philosophers, although it was contested by Giambattista Vico, Friedrich Nietzsche, Oswald Spengler, and others. This assumption is also contradicted by the historiographical accomplishments of East Asia, which led the premodern world both in the quality and quantity of its historical production, despite the fact that linear conceptions of time were relatively weak. They were, however, not lacking altogether. Even ancient Indian genealogies inscribed on tombstones and monuments show a richness of linear conceptions of time. Archeological remains discovered in ancient Mesoamerican civilizations also testify to the employment of linear time (Feldherr and Hardy, 2011).

Stream of Time and Stack of Time

The examination of what constitutes linear time will lead us to a comprehensive understanding of the nature of historical time.

Linear time proceeds from the viewpoint that time is ordered sequentially like a stream of water in the river. This time is recorded by a sequence of numbers, as is characteristic in the Western (Christian) way of counting years, and it has given us a view of historical time as proceeding in a linear fashion.

Along with this idea, a frame of regnal years, or East Asian era names (a metamorphosis of regnal years) has been widely used. This implies a view of time as being stacked like a pile of bricks, rather than flowing like water in the river. This image of time reflects a fundamental stratum of our consciousness, as when we look back into the past, and remember 'my school days', 'our Tokyo period', or 'between the wars'.

The Western cultures that now dominate the modern world extended their mode of historical thinking based on linear time across the globe, often imposing their sense of time on others as if such a scheme were something universal. Recent studies have shown, however, that a sense of linear time is culturally comparatively limited, while compartmentalized time is in many ways more fundamental and widespread (Sato, 1991). Cross-cultural observations seem to indicate that 'stack time' actually constitutes the most common sense of historical time. In the East Asian context, in particular, varied combinations of compartmentalized time have been developed. That linear time played such an important role in European circumstances is thus probably the result of a necessity for periodization in the development of historical studies.

Time Belongs to the Supreme Nominal Power

Throughout history conceptions of time were regulated and authorized by those who possessed the supreme nominal power. Time was set and controlled by the central authority.

Historically, the powers that governed human time can be divided into two different types: religious power and political power. Religious powers in the Western Hemisphere, in particular Christianity, Judaism, and Islam have controlled time, in broad terms. However, time in the Eastern Hemisphere has been governed by political power.

In China, time belonged to the emperor. A new dynasty and a new emperor declared the enthronement of a new power by proclaiming a new era name and a new calendar. The Chinese emperor confirmed the establishment of his power by the introduction of his time. In Japan, since the seventh century, time also belonged to the emperor. It was the emperor's duty and prerogative to declare a new era's name, a privilege retained even during military rule. Even the Tokugawa shogunate, despite its military and political might, could not deprive the Emperor of this privilege. After World War II, the Christian chronology–based time came into use in East Asia, along with Greenwich Mean Time. However, Japan and Taiwan still employ the era name system officially. North Korea introduced an era name system in 1997, and the new era name was 'Juche' (Autonomy), to commemorate the late Kim Il-Sung. The era started retroactively from his birth in 1921.

Western time based on Christian chronology and universal time now occupies the role of a common time in the world. The introduction of Christian chronology made the practice of a world historiography possible by giving a synchronicity to the history of the world. The time that now governs the world belongs to the supreme world power called 'globalization'. The expansion of the Internet communications necessitates a World common time. A similar situation was found in early modern Europe. Problems arose for merchants in setting contract dates between different countries that used various regnal years and dating systems. As a result merchants were among the first to promote the use of a common Christian chronology strongly (Lippincott et al., 2000).

Chronology

Chronology can be classified into the three types according to their features: linear chronology, stratified chronology, and cyclical chronology. The major issue of chronology focuses on the question of why, in spite of its defects, the Christian chronology was able to become a common system.

Linear Chronology

Linear chronology is a system of reckoning a series of years in a straight line from a starting point (year) of an era. Jewish chronology, Christian chronology, Islamic chronology, Buddhist chronology, etc., all belong to this type of chronology.

Linear chronology consists of two types: one counts the years serially from the starting year toward the future, and the other counts years both backward and forward from the starting year. The latter idea of reckoning years both to the future and to the past endlessly is only found in Christian chronology.

This idea of a single-era system of counting years is found all over the world. In ancient times, Mesopotamian chronology began with the accession of Nabonassar in 747 BC, and Roman chronology reckoned years from the foundation of Rome in 753 BC. Jews have counted years from the foundation of the world (*Anno Mundi*) since the third century.

Dionysius Exiguus, an abbot in Rome in 525, first introduced Christian chronology, as we know it, on the grounds that years should be noted from the incarnation of Jesus Christ, instead of the era of Diocletian, the persecutor of Christians in ancient Rome. Dionysius calculated the birth of Christ as the 753rd year from the foundation of Rome, and made this year the first year of *Anno Domini* (AD). However, it was not until the eleventh century that this chronology began to circulate widely in Europe (Finegan, 1964).

Islamic chronology (the era of the *Hijra*) was instituted by Caliph 'Umar I in 638 by introducing the system of counting years from the *Hijra*, that is the flight of Mohammed from Mecca to Medina on 16 July, 622, the first day of the lunar year. The Islamic chronology is unique among the other linear chronologies in that it is a purely lunar-based chronology, i.e., the year is counted as the cycle of 12 lunar months, not as the revolution of the Earth around the Sun.

Buddhist chronology is now used in some Southeast Asian countries. In Thailand and Cambodia, years are counted from the putative date of the death of the Buddha in 543 BC, while Myanmar employs the date of the introduction of Buddhism in 638 as the first year of their chronology. In India, the government reformed the calendar in 1957 based on the Saka era (starting from 78), and it is now widely used.

At least 10 other single-era linear chronologies have been proposed, including the era of the emperor Huang and the era of Confucius in China, and the era of Tangun in Korea.

Stratified Chronology

Stratified chronology is a system of reckoning years by successively piling eras (sets of years) upon one another in sequence, an idea that originated in the regnal year system. The method of counting years by royal succession has been used throughout the world since ancient times, and it was also used widely in Europe even after Christian chronology was circulated in the sixteenth century. The method of dating by regnal years, though troublesome, satisfies the theoretical requirements of specifying years, and this chronology finds its most significant expression in the era name system of East Asia. The era name system derives its origin from the idea of conferring a significant title to a certain series of years, instead of using simple regnal years. It was first introduced by the emperor Wu of the Han dynasty in China, in 114 BC. The years of his reign were named according to heavenly good omens as 'Founding the First Era', 'The First Light', 'The First Hunting', etc.

This idea of chronology later rippled out in all directions to the countries on China's periphery. It was employed in Vietnam and Korea by the middle of the sixth century, and by the middle of the seventh century, at the latest, it had reached Japan. These East Asian countries on the Chinese periphery took the idea of the era name system but made little use of the Chinese names themselves. When used in tandem with the Chinese sexagesimal cyclical system, the era name system fulfilled its function as a method of reckoning years in East Asia.

East Asia was chronologically under the exclusive dominance of the era name and sexagesimal cycle systems. The era name system was premised on the very existence of Chinese letters as ideographic characters. Ideographic characters made possible the development of the East Asian philosophy of chronology, i.e., that the counting of years is not an act of assigning an ordinal number to a year, but rather that of giving a year a meaningful name.

It originates in the *Weltanschauung* inherent in ideographic characters: that Chinese characters by their very nature assign a meaningful name to every aspect of the world and the era name system could be seen as an extension of this. Assigning a meaningful name to a month was common throughout the world, but outside of East Asia designating a set of years by a single name, was, when it occurred at all, usually a matter of a historian's conception and not of an official or public designation.

For more than 2000 years, until the middle of the twentieth century, this era name system was employed in East Asia, to the virtual exclusion of all other modes of reckoning. The Christian era system was brought into East Asia during the latter half of the nineteenth century as part of Western culture. The system took the place of era names in the subsequent century. Japan, Taiwan, and North Korea still employ the era name system officially, though the Christian era system is widely used as well. It tells us that the idea of reckoning years within the context of meaningful names is deeply rooted and alive in the collective consciousness of East Asian people (Prusek, 1963).

Cyclical Chronology

In addition to the above two types of chronology, cyclical chronologies have long been used all over the world. Well-known examples of these are the chronology by indiction (*indictio*) in early Medieval Europe, and the sexagesimal cycle in East Asia. In this type of chronology, years were given according to their place in the cycle of 15 years (*indictio*), or 60 years (the sexagesimal cycle), the number of the *indictio* or the sexagesimal cycle itself being ignored. They could be called chronologies in the sense that they turned into absolute chronology by giving ordinal numbers to the cycle, in the same way that the Olympic dating system operates. It was also possible to apply such an ordering to the sexagesimal cycle, but it was rarely used in practice. This method may also be called a relative chronology in the sense that it was not used to designate specified years in comparison with an absolute chronology.

The Hindu idea of chronology is cyclical, but it is unique in its philosophy of time in that cosmic cycles of creation and destruction continue indefinitely. In Hinduism, time belongs to the supreme God, and one chronological cycle (*yugas*) is 4 320 000 years (Eliade, 1954).

The Christian Era System as the Common World Chronology

Since the Christian chronology starts from the AD 1, the twenty-first century begins in 2001, not 2000. The concept of zero was not yet known in the West when the Christian chronology was created in the sixth century. Zero came to Europe around the tenth century, from India, via the Islamic countries. In addition it is now believed that there was an error in assigning the year of Jesus' birth, and he was in fact born a few years before the year AD 1.

In spite of these imperfections, we live with a Christian chronology that is now so dominant in worldwide usage that there is no possibility of it being replaced with a new or revised method at any time in the foreseeable future. Christian chronology now occupies the role of a common or neutral chronology in the world. This was theoretically brought into effect by introducing the concept of 'BC' and deciding on the first day of January as New Year's Day; and it had been politically determined by the dominance of Western culture since the nineteenth century.

It is important to inquire here why the invention of 'BC' made the Christian era system the universal chronology. It is because it has the great advantage of locating every historical event on a single time base.

Venerable Bede, in his *Historia ecclesiastica gentis Anglorum* (731), first proposed the idea as *anno igitur ante incarnationem Dominicam* (in the year, therefore, before the incarnation of the Lord). The expression became widely used after the late sixteenth century in a period known as the dispute over Christian chronology. This dispute arose because of disagreements between scholars concerning the period between the Creation and the birth of Christ, and because of inconsistencies between Christian and non-Christian ancient histories, as for example in the differences between the Chinese records and the

Bible. However, since the seventeenth century, the idea of BC has been used widely in Europe.

It should be remembered that BC was first used, not independently, but as an auxiliary chronology to *Anno Mundi*. The secularization of the Christian idea of history was necessary for the exclusive use of BC, which started in the eighteenth century. An early example of the transformation of the Biblical chronology into a secular chronology can be found in Voltaire's *Essai sur les moeurs et l'esprit des nations* (1756), where the concept of BC became an indispensable part of modern chronology.

As a result of the developments in archaeology and geology in the scientific revolution of early modern Europe, the method of counting serially backwards from the birth of Christ proved extremely useful in dealing with the debate over 'the creation of the world'. It was, perhaps, accidental that BC came into such use, since it was originally conceived in order to describe the years between the Biblical Creation and the birth of Christ. It was beyond the imagination of its founders to think of a time 'before the Creation'.

It is a curious irony that this Christian chronology was so suited to tackle the scientific revolution, and it spread itself beyond the boundaries of Christendom. The invention of BC transformed the Christian era system from being a specifically Christian chronology to being one with a universal potential. It is now widely used in non-Christian countries as the main chronology. For example, China officially introduced the Christian chronology as the Common Era (*Kungyuan*) in 1949, in order to become synchronized with the rest of the world. Joseph Needham proposed the idea of chronological expression avoiding the writing of dates as 'BC' and 'AD' by the use of '−' and '+' respectively in his *Science and Civilization in China*, because "such a convention seems also more suitable for civilizations which were never part of Christendom" (Needham, 1966).

The idea of a common chronology is premised on the notion that all peoples, cultures, and nations have their own 'year', which allows the common man to create a balanced sense of history, and acquire the possibility of a free and varied sense of time. Because the reckoning of years is not a mere assignment of numbers in order, but the creation of a cognitive framework for experience and memory, it is in itself an advanced intellectual exercise peculiar to the human species. It is also a political, social, and cultural act of human intelligence (Sato, 1991).

Periodization

Periodization has been practiced since ancient times. Well-known examples are the four metallic ages (gold, silver, bronze, and iron) as introduced by Hesiod in the seventh century BC, Daniel's dream of the four kingdoms and its interpretation (Babylonian, Medo-Persian, Macedonian, and Roman monarchies), the Buddhist idea of the threefold periods (the true law, the reflected law, and the end of the law), and Saint Augustine's proposition of six periods (*infantia*, *pueritia*, *adulescentia*, *iuventus*, *gravitas*, and *senectus*, the last named corresponding to the period from the Christ to the end of time). Each had a fundamental influence on the premodern people in its respective civilization. In the East, it was a matter of primary concern to know when the period of the end of the law started. Similarly, in the West people were eager to know the advent of Doomsday.

An examination of these periodizations shows us that (1) at a very fundamental historical level people employ compartmentalized time and not linear time for their historical thinking, and (2) characteristically a people acquire their historical *raison d'être* using their past to place their present in history.

This article, however, focuses on the periodization that appeared with the rise of modern historical scholarship in the seventeenth century, and has become a major historiographical asset.

Two Types of Periodizing Terms

Time has been worked within two ways, using both continuous time and compartmentalized time.

It is natural and inherent for us to employ compartmentalized time when we look back into the past. In the same way historians in their research invent compartmentalized time in order to give the past a certain unity. Thus compartmentalized time is a natural and indispensable historiographical tool.

Historians have employed compartmentalized time in two fundamentally different ways; one was to use the historically existing compartments, and the other was to invent and retrospectively apply a structure. Dynastic periodizations widely used in the East and the West are typical of the former; however, the latter represented, for example, by Christophus Cellarius, Karl Marx, Fernand Braudel have almost always been created in the West (Braudel, 1980).

Periodization as a Historiological Tool

Periodization is a form of historical understanding, designed as a historiological tool for making the past understandable, intelligible, and meaningful by dividing it into compartments. It was developed largely in the West in the early modern period under the dominance of Christian chronology. More generally speaking, it was to be found in civilizations where linear chronology played a central role, for example Ibn Khaldun's distinction between the *umran badawi* (rural culture) and the *umran hahari* (urban culture).

Periodization became a key feature of Western historiography, which developed into a cognitive discipline by proposing new perspectives and varied interpretations. This contrasts strongly with the East Asian experience. Even though a rich historiographical culture flourished in East Asia for more than 2000 years as a major field of learning, it did not propose the universally applicable type of terminologies used by Western historians. However, East Asian historians employed the combination of many different types of periodization in their written history, such as dynasties, era names, sexagesimal cycles, and periods based on changes in the site of the capital. Modern European history, on the other hand, drew on a long and deep-rooted tradition of annals and chronicles, whose style depended solely on a concept of continuous time.

The various periodizations were essential for their establishing history. Even now, a list of titles of recent works by Western historians reads like an inventory of variously compartmentalized time.

The Christian era system had been in use in the West long before the birth of historiography in modern Europe, at which time the historians took an interest in compartmentalizing the past in order to analyze history. In one sense, it could be viewed as a declaration of independence by modern historiography from such traditional forms as annals and chronicles.

In the 1700s, the word 'century' came into use, signifying that the modern Western mind needed compartmentalized time for thinking about history. In Western historiography such statements as 'the first half of a century', 'the first quarter of a century', or '1930s' are used quite freely and this is a sign that compartmentalized time has become indispensable for looking back upon the past, which cannot be grasped simply by employing absolute years.

To summarize, in the West the idea of a single-era count formed the basis of chronology, i.e., the Christian era system. Given this basis historians employed many different kinds of periodization to compartmentalize time to think about history. By contrast, in East Asia only the compartmentalized time could be found in historiography. Their historiographical style was developed on a stratified or stack-shaped chronology and various other types of compartmentalized time. The expansion of modern European historiography, paralleling the global reach of Euro-American military and economic power in the late nineteenth century meant that historians everywhere are now using the technique of using periodization to understand history. *Pensare la storia e certamente periodizzarla* (to think history is certainly to periodize) (Croce, 1916(1960)) is a view now shared by all world historians.

Later development of periodization to incorporate such concepts as the 'feudal period', the 'medieval period', etc. has given rise to the reinterpretation of human history on an international scale. The idea of dividing history into three periods was the first success of the method.

Universality of the Tripartite Periodization

The universal acceptance of a tripartite periodization: ancient, medieval, and modern has had a fundamental influence on the development of world historiographies.

This threefold periodization was first employed by Georg Hornius, a Dutch historian in his *Arca Noae* (1666), and became widely known later through Christophorus Cellarius, a German historian, in his *Historia nova, hoc est XVI et XVII saeculorum* (1696). With the transmission of modern German historical scholarship it spread across the world two centuries later and has been taken for granted as a global standard since then.

This threefold periodization found universal acceptance for two reasons: (1) the terminologies were almost free from religious connotations and Euro Christian signs, even though *medium Aevum* was originally a theological notion, and (2) it included a universal attitude toward the past by instituting a negative period between the glorious past and the contemporary period.

There is still no agreement in deciding on the exact border years among the ancient, the medieval, and the modern. Professional historians have concentrated later on subdivisions of the three periods. This lack of consensus encouraged disputes among the historians of the world. As a result of this debate and the development of social and economic histories it became apparent that the very idea of border years was out-of-date.

Periodization as a way of historical thinking was developed still further by new proposals for dividing the human past. The most influential of these ideas of periodization were the 'feudal period', pioneered by Henri Pirenne, and the 'capitalist period', widely used by Marxist historians. These periodizations were discussed in the historical contexts not only of Europe, but also of the non-Western world, especially China, India, and Japan.

After World War II, Fernand Braudel published a new idea of periodization in 1949, by proposing the notion of three different rhythms of history. The *long dureé* (the long duration) became the main concern of postwar historians. This structural periodization spread over the world on the wind of social history, and became the shared perspective of the historians of the world.

Visualization of Historical Time, Chronology, and Periodization

History is the science of making our chaotic past intelligible and meaningful through a variety of interpretive and structural techniques; among these, the imposition of a systematic notion of time has proved to be one of the most important. The construction of time lines, royal lines, genealogies, chronicles, annals, and historical tables, which have developed since ancient times, helps to form our historical knowledge and mind. The further necessity of knowing the synchronicity of the plural lines was the first step toward reaching a historical consciousness. Peoples have proposed a variety of ideas to synchronize the different times of human history.

The synchronizing of historical events that happened in different places and times can be viewed as the singular most distinguished achievement in the historical consciousness of the human mind. This is reflected in the construction of chronological tables or historical tables.

It is necessary here to distinguish chronological tables from annals and chronicles; they have very different purposes. Annals and chronicles attach importance to the content they record; on the other hand, chronological tables stress the date or sequence of historical events.

The chronological table gives us a visualized temporal image of our past, in the same way that a world atlas shows us a visualized spatial image of the world. It has two important functions. One is to give a bird's eye view of the stream of time in history, the other is to make certain of the synchronicity of historical events in various parts of the world.

In East Asia, the use of chronological tables as a reference for comprehending history was first proposed by Sima Qiam (135–93 BC) in his *Shiji* (Records of the Historian). They can be viewed as the beginning of a consciousness of historical time. He arranged 'the chronological table of the twelve dukes' by separating lines for the events of each of the 12 states. Reading down the columns, one sees what happened in each state in

any given year by using the sexagesimal cycle and regnal years of that state. Reading across the columns, one can see the sequence of historical events for any given state. We could say that he made the table to show us history in both its temporal and spatial contexts.

The development of the chronological table in Europe was effected by choosing Anno Mundi, or Christian chronology, as the key time scale, around which the various historical events were listed.

The historical time-based Christian chronology was already established in Europe by the sixteenth century. However, at that time it was primarily employed in the field of ecclesiastical history. It was not until the eighteenth century that these two chronological methods were unified into Christian chronology. At that time historical events were rewritten in terms of the Christian chronology and thus were synchronized with events in the rest of the world. This synchronicity laid the foundations of a time consciousness capable of giving birth to the modern method of historical thinking. Benjamin Marshall's *Chronological Tables* (1713) is one of the earliest works of this genre. John Blair published *The Chronology and History* (London, 1754), in which he proposed a table of various regnal years in Europe, which was constructed around a time line of years since the birth of Christ. James Bell established the earliest style of modern historical tables in his *A View of Universal History* (London, 1842). Historical tables were disseminated throughout the world through the works of Karl Ploets (1819–81), a German historian (Sato, 1991).

Since the latter half of the nineteenth century, Christian chronology as the historiographical 'standard time' has spread throughout the non-Christian world. This chronology has been established as the pivotal time by which the world creates and maintains the possibility of a synchronicity of the various local historical times. It is only with this temporal foundation that it is possible to write a history of the world.

See also: Civilization, Concept and History of; Comparative History; Cultural History; History: Overview; History: Representation Within and Functions of; Intellectual History; Middle Ages, The; Modernization and Modernity in History; Religion, History of.

Bibliography

Blackburn, Bonnie, Holford-Strevens, Leofranc, 1999. The Oxford Companion of the Year. Oxford University Press, Oxford.
Braudel, Fernand, 1980. On History. The University of Chicago Press, Chicago.
Croce, Benedetto, 1916(1960). History: Its Theory and Practice (D. Ainslie, Trans.). Russell & Russell, New York.
Eliade, Mircea, 1954. The Myth of the Eternal Return. Pantheon Books, New York.
Feldherr, Andrew, Hardy, Grant (Eds.), 2011. The Oxford History of Historical Writing, Volume 1: Beginnings to AD 600. Oxford University Press, London.
Finegan, Jack, 1964. Handbook of Biblical Chronology. Princeton University Press, Princeton.
Fraser, Julius T. (Ed.), 1972. The Study of Time. Springer-Verlag, Berlin.
Hall, Edward T., 1983. The Dance of Life: The Other Dimension of Time. Anchor Press, New York.
Lippincott, Kristen, Eco, Umberto, Gombrich, Ernst H., National Maritime Museum (Great Britain) (Eds.), 2000. The Story of Time. Merrell Holberton, London.
Needham, Joseph, 1966. Time and knowledge in China and the West. In: Fraser, J.T. (Ed.), The Voice of Time: A Cooperative Survey of Man's Views of Time as Expressed by the Sciences and by the Humanities. George Braziller, New York, pp. 92–135.
Prusek, Jaroslav, 1963. History and epics in China and in the West. Diogenes 42, 20–43.
Ricoeur, Paul, 1984–1985. Time and Narrative, 2 vols. The University of Chicago Press, Chicago.
Sato, Masayuki, 1991. Comparative ideas of chronology. History and Theory 30, 26–50.
Whitrow, Gerald J., 1988. Time in History. Oxford University Press, Oxford.

Relevant Websites

http://www.historyandtheory.org/ – History and Theory.
http://www.historiographyinternational.org/organization.htm – The International Commission for the History and Theory of Historiography.
http://www.cisi.unito.it/stor/home.htm – Storia della Storiografia (History of Historiography).

Time-Space in Geography

Donald G Janelle, University of California, Santa Barbara, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the development of time-space perspectives in geography and exposes linkages between these perspectives and society's prevailing technologies for travel and communication. Special attention is given to the role of information, computation, and visualization technologies that shape the research practices that advance the potential to understand social organization and human activity behavior in a time-space context.

Geographers have always regarded space and time as fundamental attributes of reality. However, significant variations have existed in the meaning associated with these concepts. The independence of space and time – where each concept provides a separate frame of reference for positioning objects and events – is seen in both regional studies and in attempts to cast geography as a purely spatial science. In contrast, the merger of time and space into the singular notion of time-space has for many researchers become the ideal, at least in concept, if not in practice. In large measure, the technologies available at any given time have conditioned the historical transition in philosophical perspectives on time, space, and time-space in geography. Recent developments in data gathering methods linked through computational and information technologies and advanced systems for analysis and visualization (often referred to collectively as cyberinfrastructure) have dramatically altered the prospects for understanding and estimating the impact of space-time processes on human-environment relationships and human behavior (Dijst, 2013; Goodchild, 2013; Richardson, 2013). This raises an intriguing question; How might a truly time-space perspective with complementary supportive tools alter the outlook for geography as an integrative and relevant discipline in the twenty-first century? The following sections explore the transition in time-space thinking in geography and address some of the philosophical and technical challenges in their application to the study of human behavior.

Absolute vs Relative Views of Time and Space

Immanuel Kant (1724–1804), who was less noted for instruction in physical geography at the University of Königsberg than for his broader theory of knowledge, was instrumental in laying a philosophical basis for treating time and space as a priori separate from each other, with absolute dimensions validated through human experience. In contrast to Kant's view, the inseparability of space and time is embraced in the concept of time-space as a single definable dimension of reality, where one cannot have space without time, and vice versa. This perspective, one of the foundations of relativity theory, conceives time-space in relationship to the movement of material objects and to events, an idea dependent for confirmation on significant advances in measurement technologies since the late nineteenth century. Conceptually, the coincidence of time and space is easy to grasp – any object or event simultaneously embraces positions in time and in space (e.g., tossing a ball and observing its flight) – easy to say in the realm of what is observable with our unaided human senses but more challenging as distances increase in geographical space and move beyond our direct senses in nano and cosmic spaces.

From an operational and measurement perspective, researchers often return to the Kantian view, treating the separation of time and space as a practical necessity. Geographers have attempted to work around this difficulty in two principal ways. First, explanatory and descriptive models (e.g., Walter Christaller's central place theory) may be used to consider the human settlement system and human activity patterns at different points in time as a means of detecting historical changes in the spatial character of landscapes and human behavior. Second, researchers may abstract the time dimension of change into periods or stages, sometimes in arbitrary correspondence to the processes being investigated. An example of the latter is the use of census data, where changes might be assessed over a 10-year period, such as population growth within a region between 2000 and 2010. The easy availability of a national census has rendered census data as a pragmatic, although less than ideal, choice for describing and mapping demographic and other patterns of change that shape human behavioral landscapes. However, the implicit assumptions and the limitations of this approach are problematic, especially the representation of continuous and often complex processes as disrupted sequences of snapshots at regular intervals in time and the chunking of space into counties, census tracks, block groups, and blocks. Researchers often skirt this issue altogether by conceptualizing and defining periods of time or stages according to the dominance of certain social circumstances (e.g., the baby-boom period) or of prevailing technologies (e.g., the automobile era). This approach is attractive for description of general processes but in most cases falls short of rigorous analysis in a strictly scientific context.

From Spatial to Time-Spatial Perspectives

Attempts in the 1950s and 1960s to recast geography from a descriptive to a predictive science offer a clear demarcation in the progression of thinking about time-space in geography. In the traditional paradigm of regional descriptive geography, time and space were treated as separate even though attention was given to documenting landscape conditions at different

points in time. Theoretical quantitative geography emerged to counter the descriptive focus of this tradition; nonetheless many researchers retained the separation of time from space in what was termed the exceptionalism of the spatial paradigm. This was most pronounced in the advocacy of Schaefer (1953) and Bunge (1962) for whom geography was a theoretical science in search of spatial morphological laws. However, the merger of geography's theoretical thrust with the practical issues of society necessitated a predictive perspective, one focused more specifically on the mapping of events and related movements across space through time.

From Snapshots to Dynamic Time-Space Conceptualizations of Behavioral Landscapes

Interest in prediction resulted in a shift of emphasis from the depiction and explanation of geographical patterns to that of process identification for the understanding of the dynamics of regional economies and the nature of human spatial behavior. Thus, from a fundamental focus on pure spatial form, attention turned to the movements of objects and to the passage of events through space. Models to simulate urban growth and other phenomena as processes of spatial diffusion represent explicit attempts to mesh the time and space dimensions in geographical research.

Significant statements advancing time-space perspectives are found in Hägerstrand's (1970, 1973) conceptualizations of spatial diffusion and time geography, Janelle's 1968 formulation of time-space convergence (TSC) (see Janelle, 2014), and Harvey's (1990) focus on time-space compression. The time geography, convergence, and compression concepts are discussed in the following sections, especially in relation to rapid changes in the availability of geo-referenced data, live data feeds from remote and embedded environmental sensors, and a plethora of ever-more-sophisticated tools for information gathering, processing, analysis, visualization, and communication.

Time Geography

At the behavioral level, Torsten Hägerstrand's time geography model of society stands out in its attempt to capture the movement and activity participation of people through conceptualization and graphic representation of their space-time locations. The concept of an individual's time-space path captures the continuity of human existence as a directed line (a trajectory) that occupies space and time simultaneously and continuously, depicting a person's activities in space through the temporal cycles of the day or through a lifetime. The inherent Euclidean nature of this space is a consequence of treating time-space as a homogeneous medium for movement, broken up in response to social, environmental, and other constraints that determine who has access to specific spaces at given times and under what circumstances.

Hägerstrand's motivation for a time geography of social activity was to provide tools to assist planners and policy makers in conceptualizing how their decisions influence the spacing and timing of landscapes and services so as to minimize constraints on access to essential and desired resources (e.g., jobs, day care services, health care, schools, and recreation opportunities). His identification of capability, coupling, and authority constraints represents an insightful means of delimiting the geographical limits to an individual's activity patterns over any given period of time. In Hägerstrand's conceptualization, people's access to opportunities is based on their capabilities, such as: their means for moving from one location to another, their ability to free up time and finances for traveling and engaging in activities within a specified time span; their need to interact (couple) with other individuals or physical resources to carry out economic, social, and other transactions; and their ability to perform activities that ascribe to the authority rules, physical boundaries, and fees that govern access to resources and places. For example, the operational hours of public and private facilities (e.g., the library or clinic), entry fees for a concert or sporting event, and ownership rights that control for trespass or public accessibility are instances of authority constraints, often essential to rationalizing investments but restrictive to open participation. Thus, through time geography, the space-time autonomy of individuals or groups is defined by the behavioral options that are accessible in time and space within a local context after consideration of capability, coupling, and authority constraints.

Through graphic representation, Hägerstrand provided a unique means of representing time geography problems and solutions. Figure 1(a) depicts the core elements of the model: a representation of a time-space (shown here in three

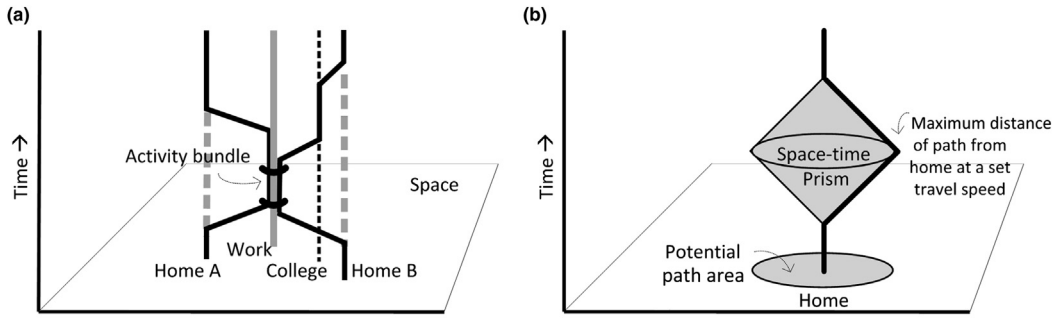

Figure 1 Graphic representation of some of the basic elements of time geography: (a) space-time activity paths of individuals A and B; (b) the space-time prism defines the temporal and spatial limits of an individual's activity options for a single day, projected in 2D space as a potential path area (PPA).

dimensions), the space-time paths of individuals A and B over the course of a 24-hour day, and the linking of their separate paths to form an activity bundle (coupling) at a work location. The home base is allocated special significance for allowance of regular sleep, sustenance, and fulfillment of individual and family obligations, whereas employment or schooling activities may be seen as obligatory for acquiring the income and knowledge to function within a society. In Figure 1(b), the path is constrained by the necessity to be at home in the earlier and later parts of the day. At a given travel speed (designated by the slope of the path, and subject to capability and authority constraints), the outer margins of travel distance define the bounds of a space-time prism in which all activities must take place. In geographical space, the confines of activity participation are shown as a potential path area (PPA). More complex patterns could be represented by composite prisms linked by travel paths that result in a series of PPAs. Such graphic representations maybe used to depict a broad range of problems at varying levels of complexity to add clarity to social and planning issues and to possible solutions and outcomes. Examples include documenting temporal patterns in the gender differentiation of urban space, disease contagion, measuring and predicting the likelihood of successful trip outcomes based on the means of travel, equity in accessibility to services in cities, multiscale experiences in physical and virtual worlds, time-geographic strategies of police surveillance to protect vulnerable neighborhoods from burglary activity, trade-offs in telecommuting at household and societal levels, and the implementation of flexible work hours on activity choices, among many others.

Since its formulation in the early 1960s, the adoption of time geography has been constrained for lack of accessible data resources and computational models, but the past decade has seen a significant surge of interest and refinement in time geography theory, with the development of new tools for data acquisition, analysis, and display, especially in facilitating the integration of data from activity diaries and from social networks within geographic information systems (GISs). Some highlights of these developments include:

- Miller's (2005: 17) time geographic measurement theory has contributed "analytical formulations for basic time geography entities and relations, specifically, the space–time path, prism, composite path-prisms, stations, bundling, and intersections" that increase the likelihood of time geography's use within formal computational display systems (e.g., GIS);
- The mapping of opportunity spaces for individual space-time paths draws on refined conceptualizations that allow for variations in travel speeds and for isotropic and anisotropic conditions for travel by Miller and Bridwell (2009); and
- Kwan and Lee (2004) and Chen et al. (2011) provide examples of successful integration of time geography within GIS and other advanced data visualization systems.

These technical achievements help ground time geography as a supportive framework for thinking about and visualizing the role of constraints and facilitators of social and behavioral activities in family life and in other diverse socioeconomic settings tied to the character and conception of home, place, and other behavioral environments related to expressions of human agency. Critiques of time geography have focused on its structural determinism, its overreliance on absolute physical dimensions, and on difficulties for incorporating the emotions and diversity of meanings that accompany human behavior in real geographical spaces and places.

The time geography of everyday life and lifetime transitions at individual and societal levels are subject to changes in the physical frameworks in which human activity occur. The following section gives specific attention to technological transitions in modes of transportation and communication. These are seen as space-adjusting technologies that directly affect the authority, coupling, and capability constraints that link human behavior to the socioeconomic circumstances and opportunities of local, regional, and global environments.

TSC and Human Extensibility

The TSC concept has its origins in the mid-1960s (see Janelle, 2014), in a period of significant transition in transportation and communication technologies – the coming of age of limited access highways in Europe and North America, growth in automobile access, European and Japanese investments in high-speed passenger rail systems, the transition from propeller aircraft to jet and supersonic passenger planes for transcontinental and transoceanic flights, and expanding access to television and telephones – all contributing to a discussion of life in a shrinking world. TSC was an attempt to measure this shrinkage and to assess its impact on settlement systems and on the competitive prospects of places in regional and global economies. It is a measure of change in the effort required to overcome distance. Generally, TSC is defined as the average rate of decline in travel time between two places or among sets of places over a specific time period. For example, owing to innovations that span stagecoach travel to automobiles on limited access highways or to express rail, Boston and New York have approached each other at an average annual rate of about 20 min per year between the years 1800 and 2010. Interestingly, these cities converged at an average of about 80 min per year in the period 1800–1830 as a result of innovation in stagecoach technologies and logistics but only 1.2 min per year via automobiles from 1920 to 2010. In contrast to convergence, the parallel concept of time-space divergence provides for the tangible measurement of increases in separation between places that result from traffic congestion and failure to maintain infrastructure, especially in urban regions.

Implicit in the convergence–divergence viewpoint is the recognition that physical points (places on the earth) are in relative motion with respect to one another whenever functional measures (such as travel time and cost) are used as distance metrics. Because levels of investment in transportation and communication technologies are not uniform across space, some places converge much more rapidly with the remainder of the settlement system than do others. Some places, bypassed by modern space-adjusting technologies, will diverge; others will converge or diverge,

contingent on the alterations of transport infrastructure and service provision (for instance, abandonment of air and bus services on selected routes can counter the advantages of a new highway that provides convergence gains only for those with automobiles). Variations in the set of rates that define convergence among all possible pairs of places create a nonhomogeneous time-space that is relative and dependent on the inferred movement of places with respect to one another. In this time-space, places are seen to compete in their quest for comparative gains in accessibility over other places, thereby deriving advantages for attracting economic and other opportunities.

TSC is analogous to the physical concept of velocity and purports to assess the earth-shrinking impact of technological investments to overcoming distance. As a time rate of change in travel time over historical time, it is a metaconcept, abstracted and knowable only through measurement. The structure of the resulting space maybe seen as the frame of reference for human activities in a time geography context, but this framework is constantly in flux and is not describable in terms of Euclidean geometry. Even though this abstraction of reality may not be recognized intuitively, the time-space convergence–divergence process embraces explicit recognition of its social origins as the result of deliberative innovation through investment decisions. It is this human behavioral response to TSC that reshapes the space of human geography. Yet, there is also reciprocity in this process, whereby the existing spatial structure of society also influences emergent patterns of investment to encourage TSC (a form of positive reinforcement). Such investments often reflect the hierarchical positioning of places and their comparative power to influence patterns of convergence among places in regional and global economic systems.

From the perspective of individuals and social groupings, opportunities to take advantage of transportation and telecommunication innovations are dependent on income, education, and social or professional networks that contribute to variations in the potential extensibility of people and places. Human extensibility is an affordance made possible by TSC (Janelle, 1973). It refers to the ability of individuals or agencies to extend their influence and presence over space and through time to distant locations via media and other means of influence over human resources. Such capabilities depend on one's socioeconomic situation and connectivity to the broader world.

Human extensibility parallels the construct of time-space distanciation as advanced in Giddens's (1984) structuration theory. Together, these concepts pose questions regarding the closure or spatial boundedness of human societies and systems. Adams (1995), Kwan (2000), Shaw and Yu (2009), and Buliung (2011) have illustrated how time geography offers a useful means of conceptualizing and documenting the role of communication and information technologies in differentiating processes of human extensibility for various subpopulations at local through global scales. They also bring forward issues concerning the differential abilities of individuals, firms, nations, and other forms of human agency to reach beyond their physical locations to shape possibilities for human livelihood elsewhere. Such discrepancies led Harvey (1990) to advance yet another conception of time-space – time-space compression – a concept rooted in the historical materialism of the political economy.

Time-Space Compression

David Harvey's notion of time-space compression provides a compelling rationale for imbuing time-space with social content and experiential meaning. This is brought about by attempts within Capitalism to extract as much surplus value per unit of time as possible. Although annihilation of space through time is analogous to TSC, the broad implications of compression may be less amenable to direct measurement. Use of the term compression captures the general acceleration of all kinds of events that intensify the turnover of capital and that implode on human existence as substantive levels of stress. Furthermore, in Harvey's interpretation, the transcendence of global capital over local domains thrives on shortened time horizons and on the ability to eradicate distance as an inhibitor to exchange and dominance. The ability of capital to shift resources to different places easily, and often with impunity to the disruption of life at the local scale, suggests that time-space compression may denigrate the importance of place in human society. It is in the inherent mobility and speed of such possible actions that time-space compression challenges any attempts to define the time-space nature of society too rigidly, a theme that Warf (2008) expands upon.

Spatiotemporal Analysis and Cyberinfrastructure

Empirical documentation and hypothesis testing, as well as more humanistic social interpretation, have entered a new stage of early twenty-first century technological prowess for acquiring, generating, searching, integrating, and distributing information about human activity behavior in a time-space context. This cyber-spatial infrastructure incorporates the georeferencing capabilities of global positioning systems (GPSs) and cell tower triangulation to log locations continuously to a meter resolution or less, the ability to capture and/or disseminate information collected through mobile and fixed-location sensors via ICT-based smart handheld phones and wearable instruments, and to draw upon an explosion in the sets of services offered through the Internet. Such a suite of tools is already within the possession of a sizable proportion of the human population and is expanding rapidly; these technologies are the backbone of location-based services – usually offered through private firms and governments that capture and disseminate information based on an individual's location at a given time. Related services include near real-time traffic reports based on the aggregation of cell phone GPS signals, strategically positioned cameras, and other sensors of traffic streams on highways; navigation aids for vehicles, pedestrians, and cyclists; access to maps and satellite imagery based on where one is located, and the ability to know if friends and associates or specific kinds of facilities are located nearby; the means to share correspondence and photographs with specific individuals or networks of friends and associates

via Internet-based social networks; the opportunity to store digital resources of all kinds on large computer servers (the cloud) for safekeeping and future accessibility and for sharing selectively in a secure environment; and, through big-data initiatives, the ability to link data (massive amounts of data – think petabytes) from diverse data providers (Goodchild, 2013; Richardson, 2013).

In the context of research use by scholars and planners, this cyber-spatial infrastructure provides capabilities for monitoring behavioral spaces, for analyzing the social ecologies of place, and for creating dynamic spatial demographies of local regions. Such infrastructure can be used to promote communities of interest at local, regional, and global scales; enhance emergency responses through timely and detailed locationally sensitive information; and accommodate the creation of significant new data resources through volunteered geographic information (Goodchild, 2007). Places are of particular interest to geographers and provide a convenient base for linking other data sources (e.g., social media). However, understanding the origins and reliability of such sources and managing their storage and access entail daunting technical challenges to ensure security, interoperability across software and hardware platforms, and ease of use.

Aside from data, an issue that has hindered broad application of time-space perspectives in the social and behavioral sciences has been the difficulty inherent in representing the underlying complexity of human organization of space through time. But increasingly, integration of space-time analyses and modeling within the frameworks of dynamic mapping, cartographic visualization, and GIS suggest that display technologies are catching up with the long-standing needs of social and behavioral scientists to make effective use of new data resources to provide empirical assessment of alternative theoretical constructs. Advances in agent-based modeling and microsimulation, network analysis, and digital animation are offering new analytical and display methodologies for conveying socioeconomic and behavioral processes for both research and education. The development of virtual globes, such as Google Earth (which features both reasonably current and historical map and imagery coverage), offers opportunities for exploratory learning activities through time-line-based views of current and past landscapes – locally, domestically, and internationally. News media have jumped on the bandwagon of using maps and animated spatial and time-spatial representations to inform the public about events and processes of general public interest – storm tracks, election projections, disease transmission, migration, demographic change, and urban growth, to name a few. A new era has emerged for temporal-spatial thinking for social science researchers and for society.

Summation: Time-Space in Geography and the Humanist Challenge

Transformations in how time and space are viewed and integrated into research and education have been closely linked to the prevailing technologies for acquiring, processing, and disseminating information. This is evident in how people and societies have employed ICT and transportation advances to interact at expanded geographical scales. These technological shifts have always posed opportunities and fundamental challenges for how geographers and other scholars interpret the meaning of place in everyday life and how they integrate these interpretations within theoretical and decision-making frameworks. For example, do the meanings of space and the meanings of time suggest anything beyond their relationship to one another as seen through technologically and socially enabled and continually changing perspectives of time-space?

The theoretical grounding for answering the above question rests, in part, on the ability of researchers to use the emerging tools of analytical representation in their descriptions, analyses, and interpretations. At the same time, answers rest fundamentally on being able to recognize humanist qualities in the nature of time-space. The humanization of time-space acknowledges its relativity to different value systems, recognizes its manifold possible metrics, and is able to relate these to human thought processes and experiences. Casting time-space in this manner draws on past formulations and opens the realm of possible investigations to include analytical time-space prediction as well as research formulations grounded in the approaches of critical science and critical social theory.

This takes us back to the question posed at the outset of this article. How might a truly time-space perspective with complementary supportive tools alter the outlook for geography as an integrative and relevant discipline in the twenty-first century? A definitive answer would be risky, but maybe it is safe to speculate that widespread space-time representation of human and natural processes will be tempered by adjudication over fundamental questions regarding the value of privacy and confidentiality. Geography is clearly part of and engaged in such a debate. For geographers and other social and behavioral scientists, the promise of cyberinfrastructure technologies requires equal attention to the important issues regarding the sensitivity of information, protection of confidentiality, and privacy rights (including locational privacy) that cast concern over the social costs associated with advances in space-time information access and search capabilities.

Geography as a discipline is positioned in both the technical worlds of GIScience and cyber-spatial infrastructure and in the humanistic worlds that critically question the implications of these developments for human well-being. Through the integrative roles of theory that address social matters in a time-space perspective and through research tools that are inherently well suited to integrate information spatially (e.g., GIS and cartography), a geography that takes advantage of new technologies to focus on movement and transition, data linkage, interactivity, and animation is well positioned to build connections across knowledge domains to enhance our understanding of the complex space-time transformations that impact human interactions at local through global scales.

See also: Access: Geographical; Automobility; Cartographic Visualization; Dynamic Mapping in Geography; Geographic Information Systems; Geography: Overview; Information Society, Geography of; Location, Absolute and Relative;

Mobility and Travel Activity Patterns; Networks, Geography of; Place in Geography; Space and Social Theory in Geography; Spatial Data; Spatial Thinking in the Social Sciences; Spatial–Temporal Modeling; Technology and Organization.

Bibliography

Adams, P.C., 1995. A reconsideration of personal boundaries in space-time. Annals of the Association of American Geographers 85, 267–285.

Buliung, R.N., 2011. Wired people in wired places: stories about machines and the geography of activity. Annals of the Association of American Geographers 101 (6), 1365–1381.

Bunge, W., 1962. Theoretical Geography. Gleerup, Lund, Sweden.

Chen, J., Shaw, S.-L., Yu, H., Lu, F., Chai, Y., Jia, Q., 2011. Exploratory data analysis of activity diary data: a space-time GIS approach. Journal of Transport Geography 19 (3), 394–404.

Dijst, M., 2013. Space-time integration in a dynamic urbanizing world: current status and future prospects in geography and GIScience. Annals of the Association of American Geographers 103 (5), 1058–1061.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. Polity Press, Cambridge, UK.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Goodchild, M.F., 2013. Prospects for a space-time GIS. Annals of the Association of American Geographers 103 (5), 1072–1077.

Hägerstrand, T., 1970. What about people in regional science? Papers of the Regional Science Association 24, 1–12.

Hägerstrand, T., 1973. The domain of human geography. In: Chorley, R.J. (Ed.), Directions in Geography. Methuen, London, pp. 67–87.

Harvey, D., 1990. Between space and time: reflections on the geographical imagination. Annals of the Association of American Geographers 80, 418–434.

Janelle, D.G., 1973. Measuring human extensibility in a shrinking world. Journal of Geography 72, 8–15.

Janelle, D.G., 2014. Time-space convergence. In: Stimson, R.J., (Ed.), Handbook of Research Methods and Applications in Spatially Integrated Social Science. Edward Elgar Publishing, Cheltenham, UK. http://www.e-elgar.co.uk/bookentry_main.lasso?id=14407

Kwan, M.-P., 2000. Human extensibility and individual hybrid-accessibility in space-time: a multi-scale representation using GIS. In: Janelle, D.G., Hodge, D.C. (Eds.), Information, Place, and Cyberspace: Issues in Accessibility. Springer-Verlag, Berlin, pp. 217–239.

Kwan, M.-P., Lee, J., 2004. Geovisualization of human activity patterns using 3D GIS: a time-geographic approach. In: Goodchild, M.F., Janelle, D.J. (Eds.), Spatially Integrated Social Science. Oxford University Press, New York, pp. 48–66.

Miller, H.J., 2005. A measurement theory for time geography. Geographical Analysis 37 (1), 17–45.

Miller, H.J., Bridwell, S.A., 2009. A field-based theory for time geography. Annals of the Association of American Geographers 99, 49–75.

Richardson, D.B., 2013. Real-time space-time integration in GIScience and geography. Annals of the Association of American Geographers 103 (5), 1062–1071.

Schaefer, F.K., 1953. Exceptionalism in geography: a methodological examination. Annals of the Association of American Geographers 43, 226–249.

Shaw, S.-L., Yu, H., 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. Journal of Transport Geography 17 (2), 141–149.

Warf, B., 2008. Time-space Compression: Historical Geographies. Routledge, Abingdon, UK.

Tocqueville, Alexis de (1805–59)

Bernard Valade, Paris Descartes University, Paris, France

Published by Elsevier Ltd.

Abstract

This article presents the life and work of the French writer and politician – mainly focusing on his two most important works, *Democracy in America* and *The Ancient Regime and the Revolution*. B. Valade underscores the originality and richness of Tocqueville's approach and contribution. From a methodological point of view, Tocqueville innovatively used the comparative method and laid the ground for what later was called the methodological individualism. From a theoretical point of view, he laid the foundations of the sociology of intellectuals, of social movements, and of social change. His thought on the democratic mentality, political passions, and tensions between freedom and equality makes proof of a remarkable modernity.

Tocqueville, Alexis de, a French writer and politician, born in Paris on 29 July 1805; died in Cannes on 16 April 1859. Born into an aristocratic background, but free of any resentment, he never ceased to examine the revolutionary process which had been set in motion in 1789. A great traveler, an observer of American democracy, and an analyst of the Ancien Régime in France, he was also a deputy under the Second Republic. His name is associated with a contemplation of the origin and future of democratic society, as well as with a method – comparativism – which was applied spontaneously in his two major works, *De la Démocratie en Amérique* (Democracy in America) (1835–40), and *L'Ancien Régime et la Révolution* (The Ancien Régime and the Revolution) (1856), both of which made their author into a central figure in social science.

The Unity of a Process

Born into a family of very old Norman nobility, Tocqueville first studied philosophy and rhetoric at the Collège Royal in Metz, then law in Paris. In 1827, he was appointed to the judiciary and became a probationary judge at Versailles. Throughout the restoration, he took part in the bitter struggle which set the old aristocracy, enclosed in its traditions, against the new France, led by the middle classes. After the revolution of 1830, and having sworn – not without hesitation – the oath of allegiance to Louis-Philippe, he gave himself plenty of room as far as events were concerned by requesting a commission to conduct a survey into the American penitentiary system. This trip (April 1831–February 1832), undertaken with his friend Gustave de Beaumont, resulted first in a report entitled *Du Système pénitentiaire aux Etats-Unis et de son application en France (On the penitentiary system in the United States and its application in France)* (1833), then the first two volumes of *De la Démocratie en Amérique (Democracy in America)* (1835), and finally what is known as the *Seconde Démocratie (Second Democracy)* (1840). The success of the first part of this work was enough to establish his reputation; in 1838 he was elected to the Academy of Moral and Political Sciences and, 3 years later, he entered the Académie française.

Tocqueville's liking for philosophical reflection did not cancel out his sense of political action. Elected to the Chamber of Deputies in 1839, and reelected several times in his Valognes constituency, he committed himself resolutely to the parliamentary path on the side of the constitutional opposition. On 27 January, in a speech with a prophetic note, he announced the revolution which was to break out 1 month later. After the February Days, he was witness to the clashes of 23–26 June 1848 which put 'the haves' at odds with the 'have-nots.' His *Souvenirs (Recollections)* (1951–to date, vol. XII) relate what this 'great battle waged in Paris' was like. A short-lived Minister for Foreign Affairs under the Prince President of the Second Republic, Tocqueville was returned to his studies for good with the coup d'État of 2 December 1851. *L'Ancien Régime et la Révolution (The Ancien Regime and the Revolution)* appeared in 1856, the second great work which its author was anticipating expanding when tuberculosis cut short his life. He died on 16 April 1859 in Cannes.

Tocqueville's work – which comprises the two great texts of 1835/40 and 1856, the *Recollections*, various papers, speeches, and reports, as well as a vast correspondence – draws its unity from the constant questioning which inspired it: why since 1789, in 1830, and then in 1848, was it always the *same* revolution which starts up again? The unity of the work also stems from the method adopted in order to understand the repetition – in Europe, and in France in particular – of the revolutionary process. Its comparative method, not systematized, but rather applied spontaneously, allowed it to avoid the perils that Eurocentricity and sociocentricity are denounced for today. "Whoever has only studied and seen France will understand nothing of the French Revolution" (1951–to date, vol. II, 1, p. 94). It was therefore without being obsessed with one singular case, and being free of all aristocratic prejudice, that Alexis de Tocqueville set about penetrating the essence of democratic society; by combining 'holism' and 'individualism,' he posited a diagnosis along with a prognosis on modern society whose value and validity were only recognized much later.

In Search of Democracy

"It is not only in order to satisfy curiosity – however legitimate – that I examined America," Tocqueville wrote "I wanted to find any lessons there from which we might profit […] I confess that in America I saw more than just America; I sought an image of democracy itself, its tendencies, its character, its prejudices, its

passions. I wanted to become acquainted with it, if only at least to know what we should hope or fear from it" (1951–to date, vol. I, 1, p. 11). With this definition, the Tocquevillian project was set out as an effort to understand a 'social state' and, secondarily, a political regime. If the first readers of the 1835 work were attracted to the evocation of exotic locations, they were rather more invited to turn the American mirror upon themselves. Three years later, they were to see a different image of democracy which was less suited to charming them.

Tocqueville's confessed search was not, in fact, separable from the recognition of a tendency – considered from the outset irresistible and irreversible – which was specific to modern societies: the march toward the equalization of conditions. Behind the analyses which were put forward, the contrast between aristocratic society and democratic society operated in a virtually permanent state, and it was the transition from the traditional order to the new arrangements in human relations that Tocqueville undertook to study. He approached this by going back to the principles which govern social states and to the institutions which arise out of them. From an early stage, he realized that one principle had to dominate and that the government which was called mixed, that is, which united an aristocratic element with a democratic one, was merely an idle fantasy. The English monarchy, with its strongly dominant aristocratic character, was still able to pass for the model of government; but it was the American system which was called upon to become the model for modern nations. Equally, Tocqueville also developed an 'ideal type' for democratic society, based on the equality of conditions and set against an aristocratic society founded on the hierarchy of status.

If the paths leading to democracy are marked out in different ways, "the great advantage of Americans is to have arrived [there] without having to undergo democratic revolutions, and to have been born equal rather than having to become it." While examining the customs – defined as "the moral and intellectual state of a people" – of Americans, Tocqueville assigned their foundation to a fully accepted equality of conditions (1951–to date, vol. 1, p. 300). In the United States, democracy was in some way 'natural'; its inhabitants had not had to struggle against privileged groups, such as the ones which had resisted in France; they made up an immense middle class within which there was active mobility. Whatever was set down in custom was fostered by institutions – the commune, the jury, and numerous associations. Decentralization and communal democracy allowed citizens to take part in public affairs, and allowed the body social to regulate itself. "Each individual makes up an equal portion of the ruling body, and plays an equal part in the government of the State." The close bond between the political and the social was actually realized in this society, a society which acted upon itself and through itself, since power only existed within it. In this way, the negative effects of individualism were limited. As for the ills which could arise out of equality, liberty provided the cure. In this society where one endeavored to bring people together rather than to set them against each other, and where the most serious social antagonisms were excluded (one found rich people, not a rich class), religion made a powerful contribution to social regulation. Tocqueville clearly highlighted the role of religious beliefs in the bolstering of customs and the upholding of republican institutions. Religion prevented the American citizen 'from imagining everything and forbids him from daring to do everything.' It ordered him to confine himself to the representation of that which was real and to the conception of that which was useful. Through this, the American mentality took shape – pragmatic, not speculative; with a firm grasp of the facts, not prey to ideology.

The general characteristics of the American model are set out clearly in *Democracy in America II*, along with its potential variations such as liberal democracy and revolutionary or democratic despotism; it also includes the amendment to the relationship between the passion for liberty and the passion for equality. Individualism – the childhood illness of democracy – is defined in this work as "a deliberate and peaceful feeling which inclines each citizen to isolate himself from the mass of his fellow men and to withdraw from his family and friends. In this way, after having created for himself a small-scale society for his own benefit, he readily leaves wider society to itself" (1951–to date, vol. I, 2, p. 105). In *The Ancien Regime and the Revolution* one again finds the threat this represents illustrated in the case of France, and the dangers signaled in 1840 which were brought up to date by the comparative analysis of the passions of democracy and of revolution. Another model, this time continental, was in place at the time and was in a way the negative of the American model.

The Analysis of Social Change

"What strikes the European who travels through the United States," Tocqueville noted in *Democracy in America I*, "is the absence of what we call government or administration" (1951–to date, vol. I, 1, p. 69). In 1840, he further drew attention to his initial observation: "Wherever, at the head of a new venture, you see the government in France and a nobleman in England, in America you can be sure to find an association" (1951–to date, I, 2, p. 113). If it was true that a good government was one which taught its citizens to manage without it, this was not working in France. The backdrop to the 1856 work, which links the interpretation of an historical crisis with a theory of social change, was provided by feudalism and its networks of dependence and solidarity where masters and servants were united by reciprocal services. It was this system, social and political, with its intermediary groups, corporations, village communities, provincial assemblies, and so on, that the French monarchy had set about destroying by means of administrative centralization. No one benefited from it; neither the nobleman whose preeminence was affected, nor the bourgeois whose vanity was wounded, nor the ordinary man who still remained the underling. Discontent was universal.

By suppressing the intermediary bodies, by condemning the nobility – stripped of its function and devitalized – to appearing merely as the ornament of despotism, while at the same time encouraging the emergence of new elites but without satisfying their statutory ambitions, royalty had converted an organic whole where each individual was bound to a group into an homogenous mass oscillating between apathy and demand, submission and subversion. Such was the new social state: "men, no longer being bound to each other by any bond of caste, class, corporation, or family, are only too inclined to concern themselves with their own individual interests and too prone to considering only themselves and to withdrawing into a state of narrow individualism where every political virtue is stifled." Thus, centralization, the restriction of liberties, and

the deterioration of public life combine forces. In this way, equality leads to despotism, since it is no longer built on liberty or on participation in communal affairs.

The explanatory variations that Tocqueville identifies therefore relate, as in the 1835/40 work, to characteristic customs, institutions, and circumstances. In the process which they deal with, the revolution no longer appears as a rupture; the terror had only prolonged the process of state centralization which had been initiated earlier. There was, in reality, a continuity between royal absolutism, revolutionary Jacobinism, and imperial politics. In the process in question, religion had also played an important role by virtue of the universalist dimension presented by Catholicism. At the same time, the revolution – itself preaching the abolition of privilege and the equalization of conditions, had "itself become a kind of new religion, an imperfect religion" (1951–to date, vol. II, 1, p. 89). As a result of this, Tocqueville was in no doubt that "all the Revolution did would have been done without it; it was only a violent and hasty process with whose help the political state was adapted to the social state, the deeds to the ideas, and the laws to the customs" (1951–to date, vol. II, 1, p. 66).

One should take note of the manner in which Tocqueville analyzes the social conflicts and their origins and exacerbation. They were born in eighteenth century France in a short period of economic recession following a long phase of general prosperity. It was therefore not increased poverty which triggered off the conflicts, nor was it the weight of former coercive institutions which were the cause. The revolution "did not break out in the regions where these institutions were better preserved and their discomfort and rigor felt most acutely by the people, but on the contrary in those regions where they were felt the least, in such a way that their yoke had seemed all the more unbearable precisely where it was the least heavy" (1951–to date, vol. II, 1, p. 99). It is a question here of situations that one analyzes today in terms of relative frustrations. Tocqueville recognized them perfectly in the case of the emancipation of the peasantry, just as he had in the case of the ascendancy of the bourgeoisie. They made two strongly felt passions explosive: one was "the violent and inextinguishable hatred of inequality"; the other "more recent and less deep-rooted" led individuals "to want to live not only equally, but also freely."

These were the passions which the *philosophes* of the enlightenment had stirred up. They provided a theoretical framework for social protest by showing the necessity of replacing traditional habits with new rules. Far removed from the affairs of state, with no practical experience and therefore inclined to conceive of 'general and abstract theories in government matters,' men of letters undertook to found cloud cuckoo land. Vain debates of ideas took the place of public life: "Political life was violently driven back into literature, and writers, taking the direction of public opinion in hand, found themselves from one moment to the next holding those positions which the leaders of political parties usually occupied in free countries" (1951–to date, vol. II, 1, p. 196). With this observation, Tocqueville paved the way to a sociology of intellectuals.

The Originality of the Contribution

The author of *Democracy in America* was hailed by his contemporaries as a new Montesquieu. The historians of the nineteenth century saw in *The Ancien Regime and the Revolution* a major contribution to historiography. If the glory of Tocqueville knew no eclipse in the United States, his work did in France, however, fall into a kind of oblivion from which it did not emerge until after World War II. It is to Raymond Aron (1967) and André Jardin (1984) – who actively worked on the publication of the *Oeuvres Complètes (Complete Works)* – as well as to Jean-Claude Lamberti (1970, 1983) – the author of the first state-commissioned thesis on Tocqueville – that one principally owes the renewal of interest in a work which from then on took its place among the classic works of sociology. There are many reasons for this long purgatory. Tocqueville was not a maker of systems as one so likes in France; he appeared as the champion of liberalism at a time when this current of thought was being strongly denounced; nor did his interpretation of the French Revolution seem ideologically sound; and then his way of thinking seemed to come under a mixed genre of headings – political philosophy, philosophy of history, and general sociology.

The reasons for Tocqueville's return are by no means separable from the turnaround in the intellectual climate which took effect in the last two decades of the twentieth century. The new tendencies are well known – the decline of hypersocialized conceptions of the individual; the freeing of collective memory previously marked by the rupture of revolution; and the loosening in particular of the vice of monocausal determinism. On the latter, the *Recollections* give this opinion which is consistent with the orientation of the social sciences at the beginning of the twenty-first century: "I, for my part, hate these absolute systems, which make all the events of history dependent on great first causes, each one being linked to the other by a chain of inevitability, and which, as it were, remove men from the history of mankind" (1951–to date, vol. XII, p. 84). The social protagonists are restored with a power of deliberation and decision, the choice of strategies, and a certain amount of room for maneuver through methodological individualism. They have 'good reasons,' as Raymond Boudon (1977) definitively emphasized, to act as they do, and it was these reasons which, in concrete situations, Tocqueville attempted to bring to light.

This writer was original and modern first and foremost through his method of implementation; it brings together observations, comparisons in space and time, and generalizations; it retains the significant fact without losing sight of the point of view of the collective whole; and the investigation which it guides is situated both on the macrosociological as well as the microsociological level. The rational reconstruction of the significance, in the end ascribed to human actions and historical events, is carried out from objective and subjective elements. The protagonists' reasons are integrated into the explanatory schema along with their passions. Tocqueville, in fact, invites his reader to weigh up the importance of the passion factors; he urges him thus to 'imagine' through a sympathetic understanding the fate of the French peasant in the eighteenth century: "Imagine this man's condition and calculate if you can the wealth of hatred and envy which has amassed in his heart" (1951–to date, vol. II, 1, p. 106).

Tocqueville's contribution to social science is also original in the way that it poses problems and formulates questions. Why, it is asked in the *Mémoire sur le pauperisme (On pauperism)*

(1835), are the countries which appear to be the most poverty-stricken the ones which, in reality, have the lowest number of inhabitants? And why, inversely, is a section of the population of the opulent nations forced to have recourse to the help of other people in order to live? A comparison between Portugal – where poverty was hardly to be seen because the responsibility was assumed by the networks of traditional solidarity – and England – whose social regulations were broken by the Industrial Revolution which created its wealth – provided the answer. It is therefore not simply a question of observing and describing – *Democracy in America* is not a simple monograph resulting from banal field research – but of understanding the social processes arising from 'good questions.'

Tocqueville's way of thinking was modern, in particular because of the subject it was applied to – that is, democracy. A type of entirely new society was establishing itself in the nineteenth century. Emphasizing its principal traits – notably the reign of public opinion – and the dangers it comprised – in particular the tyranny of the majority – Tocqueville assigned its foundation to a dependent work force and the equality of conditions. The development of the egalitarian mentality which accompanies the establishment of democratic society is a source of dissatisfaction that the hope of mobility is not up to appeasing. Inequalities, reduced in one sphere, spring up again elsewhere – there will never be an end to them. Individualism which causes social completeness to disintegrate brings about a displacement of the boundary which separates public space from the private sphere; the latter extends like a zone of retreat where one concern rules: that of material well-being. At the same time, the author of *Democracy in America* 'saw,' in a splendid passage where prophecy and prediction merge the "type of despotism that democratic nations have to fear." This was a vision of a "vast mass of people all alike and all equal," seeking trifling pleasures, all separated from each other, and above whom "rises up an immense guardian power, absolute, intricate, regulated, provident, and gentle" (1951–to date, vol. I, 2, p. 324), which reduces men to the level of children by making them dependent – this tableau really is the Welfare State.

Tocqueville: Past and Present

Current works touching upon different aspects of Tocqueville's work tend to question some of his acknowledged theses in sociology and history of political ideas, as these allow multiple interpretations, in some cases divergent.

In a substantial volume on Tocqueville's work, Lucien Jaume, theoretician of political ideas, elaborates theses contrary to the ones proposed by Jean-Claude Lamberti and Raymond Boudon (Jaume, 2008). Expanding his demonstration around the concepts of authority and sovereignty, Lucien Jaume argues that there is no 'conflict of perspectives' between the two volumes of *Democracy in America*. In this context, the considerations from 1835 on the 'tyranny of the majority' are not to be opposed to the 'soft despotism' of the 'immense and tutelary power' brought up in 1840. This second volume only continued, in a different form, the same thesis on authority, finding the same solution for different problems: the intermediary bodies. Furthermore, these conclusions are different from the ones of Raymond Boudon (2005), who emphasizes the theoretical and heuristic value of the sociology of Tocqueville. According to L. Jaume, there is a place for 'social constrain' in Tocqueville's sociology, which would be fundamentally Durkheimian.

Numerous studies look at the differences and the similarities between Tocqueville's approaches and the ones of the French liberal thinkers, primarily Benjamin Constant and Madame de Staël, but also the doctrinaire school, represented by Pierre-Paul Royer-Collard, Victor de Broglie, François Guizot, and Charles de Rémusat. Lucien Jaume, as well as Aurelian Craiutu worked toward stressing the converging points rather than the differences between Tocqueville and the French liberal thinkers. They did so by comparing mainly the works of Tocqueville and François Guizot, and rather neglecting the integral liberalism of Benjamin Constant, who belongs to a moral and political philosophy not very popular in France.

As mentioned above, the sociological analysis of Raymond Boudon (*Tocqueville for today*, 2005) highlights the modern aspect of Tocqueville's methodology: his objectivism, his insisting on the usefulness of explanations based on comparisons rather than on interpreting social facts, the use of a variety of conditional rules, and the explanation of social facts based on actions meaningful to the social actors. All these elements make Tocqueville a pioneer of the methodological individualism and a visionary, due to the modern touch of his approach, regarding current social phenomena.

In a similar direction as the one approach by Raymond Boudon, the recent work of Jon Elster on Tocqueville (*Alexis de Tocqueville. The first social scientist*, 2009) stresses the scientific aspects of Tocqueville's work relevant for various areas of study such as causal explanations, individual passions and beliefs and interactions on opinion, etc. Jon Elster does not situate Tocqueville's work in its historical moment, but argues, based on a methodological perspective, why Tocqueville is a pioneer in modern game theory, in rational choice theory and in the analysis of collective action.

To conclude, it is important to acknowledge the role of *Revue Tocqueville/Tocqueville Review*, a bilingual journal published since 1979 and coordinated by Henri Mendras for a long period. This journal, currently biannual, dedicates a considerate number of pages to erudite works, to analyses of social change, to comparative studies, mainly between France and the United States, in an interdisciplinary perspective which reflects the spirit and fruitfulness of Tocqueville's oeuvre.

See also: Atomism and Holism: Philosophical Aspects; Bureaucratization and Bureaucracy, History of; Causation: Physical, Mental, and Social; Democracy, History of; Enlightenment; Feudalism; French Revolution, The; Individualism versus Collectivism: Philosophical Aspects; Intellectual History; Intentionality and Rationality; Liberalism: Historical Aspects; Methodological Individualism: Philosophical Aspects; Religious Nationalism and the Secular State; Social Inequality in History (Stratification and Classes).

Bibliography

Abbott, J., 2007. Whither Tocqueville in American Sociology? American Sociologist 38, 60–77.

Aron, R., 1967. Les Étapes de la pensée sociologique. Gallimard, Paris.

Audier, S., 2004. Tocqueville retrouvé. Genèse et enjeux du renouveau tocquevillien français. Vrin/EHESS, Paris.

Benoît, J.-L., 2004. Tocqueville moraliste. Honoré Champion, Paris.

Birnbaum, P., 1970. Sociologie de Tocqueville. Presses universaires de France, Paris.

Boudon, R., 1977. Effets pervers et ordre social. Presses universaires de France, Paris.

Boudon, R., 2005. Tocqueville aujourd'hui. O. Jacob, Paris (Boudon, R. (2006). Tocqueville for today. Oxford: The Bardwell Press).

Brogan, H., 2006. Alexis de Tocqueville : a biography. Profile, London.

Camus, A., Storme, T., 2011. Carl Schmitt, lecteur de Tocqueville. La démocratie en question. In: Revue européenne des sciences sociales, pp. 1–49.

Coenen-Huther, J., 1997. Tocqueville. Presses universaires de France, Paris.

Craiutu, A., 2003. Liberalism under Siege: The Political Thought of the French Doctrinaires. Lexington Books, Lanham, Md; Oxford.

Drescher, S., 1968. Dilemmas of Democracy – Tocqueville and Modernization. University of Pittsburgh Press, Pittsburgh, PA.

Elster, J., 2009. Alexis de Tocqueville. The First Social Scientist. Cambridge University Press, Cambridge.

Jardin, A., 1984. Alexis de Tocqueville 1805–1859. Hachette, Paris.

Jaume, L., 2008. Tocqueville: les sources aristocratiques de la liberté. Biographie intellectuelle. Fayard, Paris.

La revue Tocqueville/Tocqueville Review (1979–), published biannually by The Tocqueville Society.

Lamberti, J.-C., 1970. La Notion d'individualisme chez Tocqueville. Presses universaires de France, Paris.

Lamberti, J.-C., 1983. Tocqueville et les deux démocraties. Presses universaires de France, Paris.

Laval, C., 2002. L'ambition sociologique: Saint-Simon, Comte, Tocqueville, Marx, Durkheim. Weber. Éd. la Découverte-MAUSS, Paris.

Lively, J., 1962. The Social and Political Thought of Alexis de Tocqueville. Clarendon Press, Oxford.

Mancini, M.J., 1994. Alexis de Tocqueville. Maxwell Macmillan International, New York.

Manent, P., 1982. Tocqueville et la nature de la démocratie. Julliard, Paris.

Melonio, F., 1993. Tocqueville et les Français. Aubier, Paris.

Nolla, E. (Ed.), 1990. De la Démocratie en Amérique Alexis de Tocqueville, 2 vols. Vrin, Paris.

Pierson, G.W., 1969. Tocqueville in America. Peter Smith, Gloucester, UK (Originally published as Tocqueville and Beaumont in America. Oxford University Press, New York, 1938).

Schleifer, J.T., 1980. The Making of Tocqueville's Democracy in America. University of North Carolina Press, Chapel Hill, NC.

Tocqueville, A., 1951–to date. Oeuvres complètes (OC), vols. 29. Gallimard, Paris (edition still in progress).

Tocqueville, A., 1986. De la démocratie en Amérique; Souvenirs; L'Ancien régime et la Révolution. Robert Laffont, Bouqins, Paris (A collection of the main works with studies by J.-C. Lamberti, F. Mélonio, J. Schleifer).

Tocqueville, A., 1991. Oeuvres, Bibliothèque de la Pléiade. In: L'Ancien Régime et la Révolution, vol. 1. Gallimard, Paris.

Tocqueville, A., 1992. Oeuvres, Bibliothèque de la Pléiade. In: De la Démocratie en Amérique, vol. 2. Gallimard, Paris.

Tocqueville, A., 2003. In: Mélonio, Françoise, Guellec, Laurence (Eds.), Lettres choisies; Souvenirs: 1814–1859. Gallimard, Paris.

Valade, B., 1985. Tocqueville le contemporain. In: Universalia, 1985. Encyclopaedia Universalis, Paris.

Valade, B., 2009. Guizot and democracy. In: Cherkaoui, M., Hamilton, P. (Eds.), Raymond Boudon: A Life in Sociology: essays in Honour of Raymond Boudon. Bardwell Press, Oxford.

Topographic Maps

Jon H Kaas, Vanderbilt University, Nashville, TN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Topographic maps in the brain are organized collections of neurons within subcortical nuclei or cortical areas that isomorphically represent the receptors of the eye, skin, or internal ear or the muscles of the body. The order of the representation reflects the order of the activating inputs to the nucleus or cortical area. Such maps are important in mediating the functions of the brain.

Topographic Maps in the Brain

Topographic maps (a representation in a nucleus or cortical area in the brain that is activated by inputs in a pattern that reflects the pattern of the source of the input such as the receptors of the skin or retina) exist throughout brains (Merzenich and Kaas, 1980). They consist of relatively faithful representations of one level of organization at the next level of organization in a hierarchical system (**Figure 1**). Thus, for example, when the arrangement of sensory receptors in the skin, retina, or cochlea projects to a structure in the central nervous system in a manner that preserves the arrangement of the receptors, the activation pattern in the central structure contains a topographic map of the sensory surface. When the first structure activated in the central nervous system with a topographic map of a sensory surface projects to a second structure in a pattern that continues to preserve the orderly arrangement of the receptors on the sensory surface, the second structure is said to also contain a topographic map, and so on over a series of structures (nuclei and then cortical areas in mammals). As structures at one level may project to more than one structure at the next level, the number of topographic maps at subsequent levels of processing may increase. Topographic maps become less precise from each level as termination of axonal arbors spread out and overlap. In part, this tendency is limited by inhibitory neurons at each level, which suppress weak inputs and thereby limit the functional spread. Strictly topographic representations may be further degraded as neurons with similar response properties grouped at higher levels of processing to subdivide nuclei and cortical areas into repeating sets of modules (a smaller subdivision of neocortex within a cortical area that is anatomically and functionally distinct from bordering cortex; they repeat many times within an area, and they are of many functional types) that locally represent different sensory aspects of the same regions of the receptor sheet. Thus, the second visual area, V2, in visual cortex of primates, represents a series of 40–50 small sectors of visual space (matching parts of the two retinas) three times for three different types of processing. Thus, the topographic map is subdivided and becomes more complicated, and less precisely represents the sensory surface. In addition, topographic maps change from one level of processing to the next level, so that they less directly reflect the sensory surface and reflect a computed parameter or dimension, a so-called computational map (Knudsen et al., 1987). For example, auditory receptors are arranged so that they are activated by sound frequencies in an orderly arrangement from high to

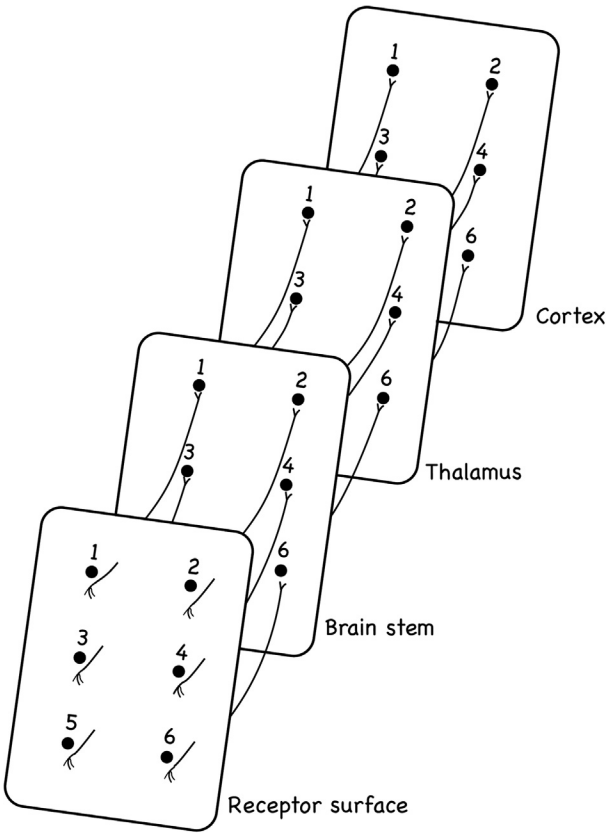

Figure 1 A simplified schematic of how axon projections from the receptor distribution on a sensory surface terminate in a topographic pattern in brain stem nuclei, and how this order is maintained over several steps of processing by similar topographic patterns of connections between subsequent nuclei and cortical areas. Due to a divergence of projections from one level to the next, and other factors, neurons at higher levels of processing generally have larger receptive fields, and the topographic pattern is less precise, and it may be altered in other ways.

low frequencies along the line of receptors, but maps of the locations of sounds relative to the listener exist in the brain that are the result of neural computations based on how sounds are differently activated by frequency-matched receptors in the two ears.

In some sense, when the organization of any nucleus (a subcortical structure that has the features of a cortical area) or cortical area (a division of neocortex that is identified by a pattern of activating input and outputs to other parts of the brain, specific functional roles, and structural specializations; they have been considered 'the organs of the brain') globally reflects the organization of the preceding area that provides the activating input, the nucleus or area is considered to have a topographic map of its activating input. In this broader sense, most nuclei and areas are topographically organized.

Topographic patterns also exist between cortical motor areas and the arrangement of nuclei that move body parts. As an example, primary motor cortex in mammals relates in a medial-to-lateral pattern across cortex to a pattern of movements from hindlimb to forelimb to face and then tongue movements, as revealed by electrical stimulation of neurons in cortex to produce movements. Yet, at a local level, the arrangement is not very topographic, as neurons that involve finger movements might alternate with those that are involved in wrist, forearm, or shoulder movements. Thus, for poorly understood functional reasons, the topography of the motor map in regard to produced movements is fractured to form a mosaic of different movement zones (Kaas, 2012).

The Discovery of Sensory and Motor Topographic Maps in the Brain

The existence of topographic maps in the brain was inferred or revealed in various ways. One of the first signs of the existence of such maps came from careful clinical observations of positions with limited small lesions of the brain or a kind of epilepsy that spread over a short period of time. Thus, in 1860s, John Hughlings Jackson, the neurologist, inferred that only the existence of orderly motor and sensory maps in the brain could explain the experiences of patients with epilepsy when a sensation or movement would start at one location on the body, and then expand systematically to include successively more distant body parts. More recently, compelling experimental evidence for such maps came from studies where investigators directly stimulated the brain electrically through electrodes placed at various locations on the surface of the brain of dogs and monkeys, and later human patients during neurosurgery (Penfield and Boldrey, 1937). As human patients were often awake during such procedures, they could also report where sensations were localized during stimulation. With the advent of microelectrodes and recording systems that could record the responsiveness of neurons to sensory stimuli, maps of sensory surfaces were commonly revealed in nonhuman mammals, and in humans during surgery. Now such maps in humans and monkeys can be revealed noninvasively with magnetic resonance imaging of activity-dependent blood flow responses in the brain.

Brain Maps Today

After years of intensive study, a number of conclusions about topographic maps seem justified. (1) The maps are based on topographic patterns of feed-forward axonal connections. The arrangement of these connections can be very precise, but some scatter that is partly suppressed by inhibitory neurons within the maps is usual. (2) Maps with direct inputs from the receptor sheet are highly topographic, but subsequent higher-order maps become progressively less topographic, so that they are only globally topographic. This means that lower-order maps preserve location and other information from the receptors while higher-order areas have more integrative functions. (3) Topographic maps within systems may differ in many ways. For example, the superior colliculus of most mammals represents the complete visual field of the contralateral, but only the contralateral hemifield of that eye, in conjunction with substantial inputs from the ipsilateral eye, in primates. As another example, the massive sensory input from the hand and resulting large representation of the hand in primary somatosensory cortex of most primates disrupts the continuity of the map such that the representation of the thumb adjoins that of the chin. To conform to its elongated shape along the outer border of primary visual cortex, the second visual area, V2, has an almost complete split of its representation of the contralateral visual hemifield into upper and lower field halves. Other disruptions of topography occur as different types of modular organizations occur within representations. Thus, V2 in primates has four fragmented, interdigitated maps of the contralateral visual hemifield. Primary somatosensory cortex in primates isolates small patches of cortex devoted to slowly adapting cutaneous afferents from patches activated by rapidly adapting afferents such that sectors of the same skin are represented twice within the overall topographic map. (4) Higher-order maps may be derived or computed from topographic maps. Thus, a map of auditory space for the location of sounds exists in the midbrain. This map is derived from earlier topographic maps. Most higher-order cortical maps appear to reflect both topographic and computational components. (5) Vertebrates and especially mammals vary in the number of topographic maps in their brains. Mammals have greatly varying amounts of neocortex providing varying amounts of space for maps. All mammals have some topographic maps in cortex, but mammals with large amounts of cortex, especially humans, have more maps. This increases the ways sensory information can be manipulated.

The Development of Maps

Sensory inputs to brain stem nuclei form orderly arrangements during development, and these nuclei form orderly connections with other nuclei, which in turn project to cortical areas in an orderly fashion and so on. The rough order in at least the early stations of this sequential process seems to depend on intrinsic factors in the brain (Sanes, 1993; Udin and Fawcett, 1988). More specifically, it has long been hypothesized that afferents are chemically labeled according to some gradient, and target neurons are also chemically labeled with a comparable gradient across the target structure. Thus, the neurons find

chemical matches and a topographic map is created. However, chemical or molecular matching is not the only factor in the development of brain maps, since there is much evidence that the development of such maps can be altered by changes in the sensory environment that induce changes in neural activity patterns. The details of sensory maps appear to be fine-tuned by an adjustment and sorting of local connections in largely adaptive ways in response to variations in patterns of sensory activation. The claim is that neurons that fire together wire together. Coactive neurons with overlapping connections maintain and strengthen their synapses on shared target neurons, while neurons that do not respond at the same time tend to weaken and lose synapses on their shared target neurons. Thus, neurons with uncorrelated activity patterns tend to segregate their terminations so as to activate separate groups of neurons in the next structure. This allows maps to adjust and compensate for losses or additions of inputs, and maps in a sequence to adjust to each other (Kaas and Catania, 2002).

There are many types of evidence that maps adjust to sensory changes during development. Usually, this evidence comes from some sort of experiment where some sensory input is removed or altered. Normal inputs expand to activate neurons formerly devoted to the removed or weakened inputs. Other evidence comes from the variations observed in maps in nature. Rats and mice have morphological maps of the whiskers of the face in S1. Each whisker activates neurons in a small oval of cortical tissue that resembles a barrel, and is commonly called a cortical barrel (a barrel-shaped module or column in somatosensory cortex of rats and mice that is anatomically distinct and responds to a specific whisker on the face. The 'barrel field' has a barrel for each whisker of the whisker pad of the face). Rats and mice usually have the same number of whiskers, and there is one barrel in S1 for each whisker. Likewise, there is one barrel-like structure for each whisker for the termination of afferents from the face in the brain stem, and in the ventroposterior nucleus of the thalamus. Thus, maps at three levels of the somatosensory system precisely match the arrangement and number of whiskers on the side of the face. This match could be due to innate developmental mechanisms, such as the matching of connections between levels as a result of a chemical code, but this appears to be unlikely.

Mice have been found that differ in the number of whiskers by having one or two more or less than the normal number, and in these mice the maps at all three levels match exactly the increase or decrease in whiskers. Thus, somehow the receptors in the skin are informing the central maps about their organizational state. In a similar way, the star-nosed mole has 11 sensory rays or protrusions from each side of its nose (Kaas and Catania, 2002). Processed sections of somatosensory cortex reveal 11 corresponding bands of tissue in primary somatosensory cortex, S1, one for each ray. Eleven bands are also seen in the second somatosensory area, S2, the ventroposterior nucleus, and the somatosensory brain stem. Moles with 10 or 12 rays have been found, and 10 or 12 bands are then seen in all four brain structures in these moles. This matching of maps from one level to the next and with the peripheral receptor sheet would be most easily accomplished by a system that uses natural activity patterns to fine-tune its maps.

Why Have Topographic Maps?

Much of the brain is devoted to topographic maps suggesting that they are a useful way of organizing brain tissue. One cannot argue that such maps are simply preserving the order of afferents from the periphery, since afferents become mixed and disorderly, but sort themselves out again as they terminate. Instead, topographic maps appear to be adaptive in that they provide a substrate for making the most common types of computations in the nervous system over short, local connections (Kaas, 1997; Chklovskii and Koulakov, 2004). Neurons in topographic representations are well arranged to make center-surround comparisons between stimuli anywhere on the receptive field sheet, and stimuli in the immediate surround. This type of comparison provides biologically useful information. Neurons rapidly determine if a stimulus at some location is brighter, greener, harder, louder, or moving differently than stimuli in the immediate surround. While it may be possible to make such computations in nontopographic structures and systems, neurons would need to be massively interconnected and over larger distances. This would add unreasonable bulk to the brain, and would be metabolically costly. In complex brains with large amounts of neocortex, it is probably more efficient to have multiple interconnected maps, rather than just making maps larger. Large maps are most suitable for extremely local comparisons of stimuli, while more global comparisons can be mediated over short connections in small maps. Thus, it seems to be more productive to subdivide tasks, and process information in parallel over sequences of interrelated small maps, and use long connections to coordinate the activities of maps within and across processing sequences. Maps vary in size in the same individual because large maps are better at fine-grain local comparisons while small maps are better at more global comparisons.

See also: Multisensory Integration, Principles of; Neural Plasticity; Parietal Lobe; Visual Streams: Dorsal and Ventral.

Bibliography

Chklovskii, D.B., Koulakov, A.A., 2004. Maps in the brain: what can we learn. Annual Review of Neuroscience 27, 369–392.

Kaas, J.H., 1997. Topographic maps are fundamental to sensory processing. Brain Research Bulletin 44, 107–112.

Kaas, J.H., 2012. Evolution of columns, modules, and domains in the neocortex of primates. Proceedings of the National Academy of Sciences of the United States of America 109, 10655–10660.

Kaas, J.H., Catania, K.C., 2002. How do features of sensory representations develop? Bioessays 24, 334–343.

Knudsen, E.I., Du Lac, S., Esterly, S.D., 1987. Computational maps in the brain. Annual Review of Neuroscience 10, 41–65.

Merzenich, M.M., Kaas, J.H., 1980. Principles of organization of sensory-perceptual systems in mammals. In: Sprague, J.M., Epstein, A.N. (Eds.), Progress in Psychobiology and Physiological Psychology. Academic Press, New York.

Penfield, W., Boldrey, E., 1937. Somatic motor and sensory representations in the cerebral cortex of man as studied by electrical stimulation. Brain 37, 389–443.

Sanes, J.R., 1993. Topographic maps and molecular gradients. Current Biology 3, 67–74.

Udin, S.B., Fawcett, J.W., 1988. Formation of topographic maps. Annual Review of Neuroscience 11, 289–327.

Torts in Law

Stephen D Sugarman, UC Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article replaces the previous edition article by T. Tanase, volume 23, pp. 15775–15779, © 2001, Elsevier Ltd.

Abstract

Lawsuits in tort allow victims of other people's wrongful and sometimes merely harmful behavior to claim monetary damages for their injuries. Defenders of tort law extol it as a vehicle for achieving justice, promoting safety, internalizing accident costs, compensating victims, and more. Critics doubt tort law's effectiveness on all these criteria and would prefer legal jurisdictions to follow the lead of New Zealand, which has replaced tort law for personal injuries with a comprehensive accident compensation scheme.

A *tort* is a 'civil' or personal wrong to a private individual or enterprise, and the legal remedy for a tort victim lies in the injured party bringing a private lawsuit against the injurer.

This is in contrast to a 'crime,' which is wrong against society and which can lead to the prosecution of the wrongdoing criminal by public authorities. The conventional social remedy against those who commit crimes is incarceration and/or a financial punishment (a fine) that is paid to the government. By contrast, when tort victims make successful claims against their injurers they usually recover money damages from their injurer (or the injurer's insurer).

Torts are like crimes in the sense that liability is often based upon the violation of standards of conduct set by the society. Yet not all torts are crimes, and violations of many criminal statutes do not injure private actors in ways that would give rise to tort claims.

Tort claims should be distinguished from claims based on breach of promises, claims arising from contract law. Speaking generally, in the latter the victim seeks to hold the injurer to the promise that was privately made, rather than to a socially created standard (although contemporary contract law often reads implied promises into contracts in part to satisfy public goals as well as consumer expectations).

The money damages that tort victims seek are understood to fall into two categories (Luntz, 2008). The first concerns what are typically called pecuniary (or perhaps economic) losses. These include things such as lost income and profits both past (from the time of the tort until the time of the trial or settlement) and future (projected out for so long as those losses are expected to occur). The other major sort of pecuniary loss for which victims sue is the recovery of expenses that have already been incurred and are likely in the future to be incurred because of the tort. The best examples of these losses are medical and related expenses.

The second category of damages is typically termed nonpecuniary or noneconomic. These are damages for what tort law in the United States calls pain and suffering, although that label is generally meant to also capture loss of pleasure, loss of amenities of life, loss of dignity, the upset one feels as a result of an impairment or a disfigurement, and the like. Other legal systems frequently label these nonpecuniary awards using different terminology, such as moral damages.

These sorts of nonpecuniary damages are primarily awarded to those who suffer personal physical injuries, and physical harm to the person will be the focus here. Quite obviously, at the time of a trial or settlement of a case, someone who has suffered a grave disability as a result of another's tortious conduct might still be suffering and be expected to suffer until the end of life. Hence in this respect pain and suffering damages, like pecuniary damages, may be awarded for both past and projected future losses.

Nations differ enormously, however, in the actual amount of nonpecuniary losses that are awarded for similar serious injuries (Sugarman, 2006). Within Europe, victims with similarly grave injuries might win 10 times as much in nonpecuniary damages in some nations as compared with others. And some victims in the United States might well be awarded 10 times as much as would be awarded in the most generous European country. Some of these differences undoubtedly are a product of national wealth and the level of average citizen material well-being which influence judgments about appropriate sums to award. But attitudes toward the disabled and existing social services arrangements for dealing with the disabled in general surely also come into play, as probably do national cultural attitudes toward money in general – because tort law can hardly put victims back into their former state of health but can only offer money as a kind of solace.

Other sorts of indignities – such as invasions of privacy, defamation, false imprisonment, and so on – where recognized as giving rise to tort claims, can also result in the award of nonpecuniary damages, although those sorts of claims will be put aside here. Moreover, although tort claims frequently arise on behalf of those seeking compensation for financial losses, damage to real or tangible personal property, injuries to reputation, and so on, as noted already the focus here will be on personal injury and death. This focus is in accord with US practice; most American lawyers who bring tort claims are typically called 'personal injury lawyers' because bodily injury claims form the core of their legal practice. Things are quite different in New Zealand, however, where tort claims for personal injury have been almost entirely eliminated and replaced by a government-created compensation scheme to deal with all accident victims (Palmer, 1979). There, tort lawyers tend to focus on financial and property injuries.

Concentrating on personal injuries, tort claims arise from a wide range of human interactions. The most important categories of accidents that can lead to tort claims include (1) transportation injuries (accidents involving occupants of autos, planes, trains, buses, and motorbikes, as well as pedestrians who are struck by such vehicles); (2) medical injuries (arising from the failed treatment by physicians and other health professionals, often termed 'malpractice'); (3) product injuries (when a specific item is defective or when an entire line of products is defective with respect to its design or the adequacy of its warnings); (4) injuries that occur on the property of others (such as slips and falls, dangerous conditions that result in accidents, and dangerous activities that lead to harm), (5) recreational injuries (suffered by participants in organized or disorganized sporting and related activities); and (6) work injuries.

Most industrialized nations have separate work-injury compensation schemes (sometimes called workers' compensation or industrial accident schemes) that pay money to those injured in the course of their employment. Nations often allow those suffering work injuries to sue in tort to recover losses not already covered by the industrial accident scheme (although the amount of compensable additional damages is often sufficiently minor that few victims elect to pursue their rights and sue their employers). But elsewhere, such as in all of the US states, for ordinary work injuries the victim's exclusive remedy is though the workers' compensation system; in return for funding this benefit employers are exempt from tort claims (Kramer and Briffault, 1991).

There are pragmatic reasons for focusing a discussion of tort law on personal injury and death cases. These injuries can give rise to devastating harms and very large losses (and sometimes to very large financial recoveries). Personal injury claims for *accidental* injuries tend to be the most numerous sorts of tort claims because all too often those who commit *intentional* torts (which are also frequently crimes, such as murder, rape, and armed robbery) have no financial means to draw on to provide financial compensation to their victims. That is to say, even if they are captured, they are typically judgment proof. Hence society tends to deal with these sorts of wrongdoers through the criminal justice system.

One reason that criminals are frequently judgment proof is that, even if they were to have liability insurance to stand behind them if they were to commit a tort, that sort of insurance does not cover intentional harms (Abraham, 1986). Rather, it only applies to accidental harm, even if tortuously imposed. Hence, you can buy insurance that would pay what you owe in tort if you carelessly drive your auto into someone or carelessly allow someone to slip and fall on your property, but you cannot buy insurance that would cover your tort liability if you wrongfully punch someone in the face. However, sometimes a person or enterprise is sued in tort for carelessly failing to prevent the victim from being injured by someone else acting in a criminal way – for example, a psychiatrist who fails to warn a victim that the therapist's patient is coming to kill her, or a vehicle owner who gives his car keys to a drunk, or an apartment building owner who is aware of criminal attacks on tenants and fails to take reasonable precautions that would have prevented an attack (Franklin et al., 2011). Liability insurance will cover the legal obligation of the noncriminal defendant in these sorts of cases.

It is also true for many countries that the most intellectually interesting and controversial tort cases arise out of accident settings and have resulted in the sorts of judicial rulings that have gained the most sustained scholarly attention (Rabin and Sugarman, 2003).

The tort law of Great Britain, the United States, and many other former English colonies (like Canada and Australia) is said to be 'common law' (Baudouin and Linden, 2010; Dobbs, 2000; Fleming, 2011) and to be distinguished from the tort law of other places, like France, Germany, Japan, and China, where tort law is said to be 'civil law' or 'code law' (Markesinis and Unberath, 2002). Although there are important theoretical and practical differences between the two approaches, not too much should be made of the difference.

Common law is law that is announced by judges in the course of giving their opinions in individual cases that come before them for resolution. In these opinions, judges explain not only what the law is but also state what they believe are the relevant facts in the matter before them and give the reasons why they reached the conclusion they did. This practice helps assure that the judges are at least formally providing reasoned elaborations in justification of their conclusions (although some believe that judges first decide how they feel the case should come out and then construct arguments designed to support that feeling). It also provides analysis for lawyers on both sides of new conflicts to argue about and draw on in trying to predict how judges are going to decide new disputes involving arguably importantly different facts. And these past written opinions provide a trail with which judges must contend when deciding new matters, either following along in what they see as the relevant track or distinguishing or overruling past opinions.

Although it was once imagined that judges would find the law of torts through logic, somewhat analogously to the way that scientists find the laws of physics, this is not a realistic way of looking at things. To be sure, common law judges tend to give great deference to the decisions of other judges in prior similar cases, following the principle captured in the Latin phrase *stare decisis* (the doctrine of precedent). Yet this too should not be overemphasized. First, the obligation to follow precedent is said to apply only to 'lower' or 'inferior' courts with respect to prior decisions of higher courts. Second, in torts cases especially, the facts are always different if one digs deeply enough, and judges quickly became expert in distinguishing prior decisions they do not like by emphasizing different facts in the matters before them. Third, the top-level courts in all common law jurisdictions reserve the right to change the law if they see fit, and they do so. Many reasons are employed to justify such changes, including the argument that social circumstances and/or expectations have changed, contemporary morals are different, new situations have developed to which it makes little sense to apply old principles adopted for other times, the prior decision was (alas) never persuasively reached in the first place, and so on.

Hence, on the ground, so long as the relevant legislative bodies keep out of the way, the common law of tort in each country is perhaps best seen as an organic system that evolves under the direction of judges in furtherance of the values and

other factors they believe best suit the contemporary world. And if the judges are seen to have gotten it wrong, then the relevant legislature can and sometimes does intervene and overrules what the judges have decided. So, for example, when the California Supreme Court decided, as a common law matter, that a social host has a legal duty of care to take reasonable steps to assure that his or her social guest who has drunk too much alcohol at the host's home does not leave and, as a drunk driver, carelessly run someone else down, the California legislature decided that this was inappropriate and by statute changed the law (Franklin et al., 2011).

In the United States, this debate usually occurs at the state level, because in the United States tort law is state law. In fact there is no national body of tort law in the United States, but rather 50 different bodies of tort law in the 50 different states (and although on most issues the laws of all the states are similar, on many controversial matters they are not).

In Canada and Australia, unlike the United States, there is a national common law declared by the highest court of each nation. But there too, like the United States, provincial and state legislatures may and do sometimes act to change by statute the tort law in their jurisdiction. For example, not long ago in New South Wales, Australia, the government restricted the amount of a victim's lost income that tort law would replace to three times the average wage in the state (Fleming, 2011). In England and Wales, it is normally the national Parliament that would change the common law as it did decades ago, for example, when adopting the principle that the victim's contributory negligence would no longer be a complete defense but rather would serve only to reduce appropriately the damages a victim could receive from his or her careless injurer (Arvind and Steele, 2012).

In civil law nations, it is said that judges simply apply the law as laid down by the government, which in many places (unlike the United States) is, in effect, a combination of the executive and legislative branches, because in parliamentary democracies (or even in monarchies or tyrannies) the political leadership commonly simultaneously controls both the legislative and executive functions. Yet it is not our worldwide experience that governments in civil law nations are routinely altering their law of torts in substantial ways. To the contrary, they generally tend to have adopted a few basic statutory legal principles that form the core of their civil code. In many places these codes may be ultimately traced back to the Napoleonic Code or even earlier, say, to Roman law. To be sure, just as there are statutory changes made in the common law of torts, so too civil code countries update, alter, and make more complex their respective bodies of tort law. But in the end many of the key decisions of courts in civil law nations reflect the efforts of judges, as best they can, to apply long-established simple principles of justice to new and changing events. In this respect, they are, as a practical matter, often functioning very much like common law judges. The tradition of writing opinions in civil code jurisdictions varies, however, so that it is not infrequently the role of academics to try to rationalize new decisional outcomes with the pattern of past interpretations of code provisions.

Interestingly enough, until lately noncommon law nations avoided use of the word 'tort,' with countries like France being traditionally said to have a private or civil law that covered both what the English would term tort law and contract law. This is now changing, and many now speak of European tort law, German tort law, and French tort law, which is perhaps a welcome relief because the word 'tort' itself has a French origin. When the common law of torts was first evolving, the elites of England, at least in the legal world, wrote in what is sometimes termed 'law French,' since the evolving English language was left more to the common man (not the common law).

The curious thing here is that even today the word 'tort' in contemporary French is ambiguous in its meaning and brings us to a fundamental issue in the law of accidents. Many would say that 'tort' means 'wrong,' but others would say that it also, or perhaps even better, means 'harm.' This distinction is critical. If tort law is about wrongs, then it rests on the principle that you may shift your losses from yourself to another through tort law when that other has wronged you – that is, when the one you sue has misbehaved in a way that resulted in your injury. If all that is required is harm, however, then one who innocently injures you may also be held legally responsible for the consequences of his or her behavior.

In contemporary parlance, this raises the question of whether tort law does or should rest on the 'fault' principle or whether people should be 'strictly liable' (that is, liability in the absence of fault) for harms they cause (Calabresi, 1970). Of course, many defendants act in a faulty way and hence would be liable in tort whether the system rests on the fault principle or on a strict liability principle. The difficulty comes when you or your activities generate injuries for which one would say you are not to blame in this specific sense: the victim is not saying that you should have acted differently, in a manner that would have avoided the injury; rather, the victim is saying that it was not socially objectionable for you to have acted as you did, but nonetheless you should compensate me for my loss that your behavior caused.

As will be discussed later in the article, there are good reasons that can be put forward in support of a system of strict liability – at the least in situations in which the injurers are enterprises and the victims are ordinary people. But in fact, at least on the books, the law of torts around the world is primarily a fault-based system. The victim must show that he or she was wronged, that he or she should not have been injured, and that the defendant in the lawsuit acted, as they say in the United States, 'unreasonably.' More precisely this means that the defendant failed to act in the way that a prudent person would have acted, taking into account the riskiness of the defendant's act and the burden that would have been born by the defendant to have avoided the harm.

In many legal systems, there are islands of strict liability in this sea of fault liability (Dobbs, 2000; Fleming, 2011). Some places, for example, have effectively made the law governing auto accidents one of strict liability (although in many places the law governing auto accidents is not really tort law anymore, but rather a separate statutory scheme for compensating victims of auto accidents on a no-fault basis). In many places, when a manufacturer provides a product that is defective in the sense that something is wrong with it (it is dangerous in ways that the product maker did not intend and the victim did not expect), there is strict liability in the sense that it is no excuse that the manufacturer employed the best

available inspection system and no such system could have detected the defective item before it caused injury; proof of lack of due care, in short, is not required. And in some places like the United States, if the actor is carrying on certain abnormally dangerous activities, like dynamite blasting, then he or she is liable for harm done even if the activity was carried out in an altogether reasonable manner and simply unavoidably went astray. Although there are other examples, the instances in which liability is 'strict' are but a few of the many ways in which people are accidentally injured, and for those many ways proof of fault is typically required.

In trying to untangle this fault versus harm conundrum, we should attend to the functions of tort law and appraise how well tort law in fact achieves the goals that many have set for it.

For some, tort law has an almost magical character (Geistfeld, 2008; Portat and Stein, 2001). Its threat, they believe, prompts most people and businesses to act properly. Hence, we achieve socially desirable levels of safety (accident prevention) without the cost and heavy hand of bureaucratic regulations. People who violate community standards and injure others are held accountable through a privately initiated system that delivers victim compensation and yet involves only minor public costs of administration. Holding wrongdoers accountable in this way achieves corrective justice in the sense that victims are 'made whole' to the extent that money can do that, and injurers are punished for their sins. Moreover, the costs of injuries that should have been avoided are now internalized into the cost of the activities that caused the harm (and are thus either reflected in the price of those activities or in reduced profits for shareholders). Because the government makes available this form of redress, people who are wronged (and their families) are less likely to seek vengeance against wrongdoers through private violence.

Notice how the social goals that tort law is said to serve by these lines of argument might be separated into social welfare (or utilitarian) goals and fairness (or moral) goals – perhaps counting compensation of victims, cost internalization, pre-empting violent revenge, and accident reduction as utilitarian and then counting justly putting victims back in the position they were in before the injury, as well as punishing wrongdoers and giving satisfaction to victims on the fairness side. Note also that this package of social goals said to be served by tort law rests on the assumption that tort law is fault based.

Others, sometimes termed 'legal realists,' are rather doubtful about the actual achievements of tort law in action with respect to these goals (Sugarman, 1989). They may scoff, for example, at the actual deterrence impact of tort law on unsafe conduct. After all, private morality, market pressures, self-protection instincts, and existing governmental regulation already promote safer conduct. For that reason, it is not clear just how much more tort can achieve, especially once it is appreciated that many individuals from time to time are just not up to behaving with appropriate caution, that executives can only have so much success in training staff to take due care when those employees simultaneously are being pressed to maximize firm profits, and especially when the costs of precaution are born now and the costs of failing to do so might not be incurred until later, when the existing executive leadership is long gone. Furthermore, as a practical matter, both individuals and firms that carelessly cause injuries are generally insured against their tort liability, so the actual monetary price they pay is only secondarily reflected, if at all, in higher insurance premiums after the fact.

Empirical studies of tort law's role in promoting greater safety (Dewees et al., 1996) are mixed, although some have found tort to have a positive, if perhaps modest, impact in the right direction. Others researchers have found that tort law can have negative impacts in terms of overdeterrence (for example, excessive investment in safety through doctors wasting health resources on too many medical tests) or the withdrawal from the society of socially desirable activities (such as removing equipment for children to climb from school playgrounds).

Skeptics also question the sensibility of viewing tort as a way to appropriately compensate victims. First, the administrative costs of the system are enormous as compared with typical private insurance or public social insurance schemes. In the United States, for example, costs of administration (especially legal fees on both sides) account for about half of the cost of the system, so that benefits to victims are only about half of what is spent. Besides that, a significant share of the benefits that are provided by tort law duplicate benefits already provided by other sources such as health insurance and Social Security, leading either to double recovery or yet more administrative expense as the double payment sum is recouped by the collateral source. Second, as compared with social insurance that generally limits the lost wages it replaces, tort law is very generous to the rich, providing full compensation for lost income (in most jurisdictions) that makes the system rather 'regressive' in its impact (and explains why New South Wales, as noted earlier, decided to cap the level of income replacement). Third, not all accident victims are compensated via tort law. This is both because some injurers are financially unable to pay what they owe and because some victims are injured through no one's fault or perhaps theirs alone (and when both parties are at fault, then in most systems compensation is reduced or perhaps even eliminated). The existence of liability insurance also blunts the punishment function of tort law, and even more so when victims are injured by employees of enterprises, because it is the firm that pays – not the actual worker who committed the tort.

Although tort liability does internalize some accident costs into activities that cause them, it does not fully do so unless strict liability is employed, but that is rarely the case at present. A much more sweeping embrace of 'enterprise' liability (strict liability for business defendants, especially in claims brought by individuals) would more fully internalize accident costs into the relevant enterprise activity. It would as well provide compensation to many more victims. Yet on the other hand, if tort law were to become mostly a matter of strict liability, then its current functions of identifying wrongdoers and punishing them would be undermined. This reality shows the deep tension that exists when trying to use tort law to simultaneously serve multiple functions.

Turning to yet more tort law goals, one can also be skeptical about how well tort law actually satisfies victims' sense of justice. In practice many tort victims do not sue. Some are unable to find a solvent defendant who would be liable for their injury. Others are unable to prove that they were victims of someone's fault even though that was the case, and in some

situations that is because injurers have hidden or even destroyed incriminating evidence. A number of victims – for example, in medical injury cases – do not even realize they were victims of negligence. And some who have theoretically valid claims cannot find lawyers to represent them because the costs of litigation will outweigh the small amount of damages they could recover (a problem especially severe in medical injury cases in the United States, where each side bears its own legal fees). When victims do sue, their cases normally settle without trial, and settlement agreements typically include a 'no liability' disclaimer. This can be highly frustrating for victims who want to formally hold their injurer accountable. Worse, many find the litigation and settlement process highly unpleasant. Their own lawyer often ignores them, the other side subjects them to harsh treatment, cases seem to go on endlessly for no discernible reason, and so on. These realities often undermine the idea that tort law brings moral satisfaction to victims.

Finally, today's organized and reasonably effective police forces, social norms, and fears of criminal prosecution are likely to be more effective than the promise of tort recovery in discouraging violent vengeance especially against those who have merely accidentally, even if negligently, committed torts.

Tort law is far more robust in the twenty-first century than it was in the past (Sugarman, 2000). Take lawsuits against physicians, for example. Although these have long been theoretically available to patients who have been injured by medical malpractice, in the past the idea of suing one's family doctor was sharply counter to social norms, especially when the doctor often provided helpful service over the years. Moreover, it was traditionally very difficult to find a doctor who would testify against another doctor, and yet winning a malpractice case almost always requires relying on an expert witness. Nowadays, at least in some places, medicine is seen as more of a business, medical specialization means that a doctor who injures you may well be someone you have never encountered before, the public is far more aware that doctors sometimes make careless mistakes that lead to catastrophic consequences, and getting a third-party doctor to stand up for the victim is no longer quite so difficult (certainly in US jurisdictions that now permit testimony from any qualified expert instead of requiring a local expert).

So, too, product liability litigation is far more promising for victims than in the past. At one time manufacturers of defective products could only be sued by their direct buyers, a rule that effectively immunized them from liability once sales of products began to be carried out through retailers. Later, manufacturers could be sued when their negligence caused a product defect that injured someone. Nowadays in many nations, product makers are strictly liable for defective products they produce, and a subset of expert lawyers specializes in these sorts of cases, leveling the playing field in litigation against large-scale enterprises.

One aspect of tort law that remains peculiarly American is the widespread use of lay juries to decide liability in most tort cases (Fleming, 1988). Elsewhere juries are either never or rarely used in civil cases. Although there is a strong ideological commitment to juries in America, it is probably fair to say that their use introduces uncertainty and variability into legal disputes over accidents. In turn, although 'tort reform' is a mantra of certain defense interests in other nations as well, it is in the United States that business interests and some conservative political players have most actively sought to 'rein in' tort law. Because conservatives are committed to the idea of personal responsibility and certainly are not eager to replace tort with more government regulation of safety, it is perhaps not surprising that most 'tort reform' has been in the form of either reducing the amount of money that tort law may award victims or changing tort procedures, rather than moving away from fault liability. These changes, where enacted (and in the United States this mostly means on a state-by-state basis), have somewhat lowered the stakes of tort litigation. Still, personal injury law remains a frequently wielded tool in the United States whenever plausible connections between solvent injurers and victims can be made (Shapo, 2012).

Indeed, many today defend tort law precisely as a mechanism for uncovering and then disclosing misconduct by those in charge of large institutions that all too often seems overlooked by government officials. In short, the 'ombudsman' or 'private attorney general' function of personal injury lawyers is one of their strongest claims for promoting the public interest (Linden, 2003; Lytton, 2008).

See also: Civil Law; Common Law; Constitutional Courts; Constitutional Courts; Insurance and the Law; Legal Realism; Legal Systems, Classification of; Litigation; Remedies and Damages: Legal Aspects; Social Insurance: Legal Aspects; Supreme Courts.

Bibliography

Abraham, K., 1986. Distributing Risk: Insurance, Legal Theory and Public Policy. Yale University Press.
Arvind, T., Steele, J. (Eds.), 2012. Tort Law and the Legislature: Common Law, Statute and the Dynamics of Legal Change. Hart Publishing.
Baudouin, J.-L., Linden, A., 2010. Tort Law in Canada. Wolters Kluwer.
Calabresi, G., 1970. The Cost of Accidents. Yale University Press.
Dewees, D., Duff, D., Trebilcock, M., 1996. Exploring the Domain of Accident Law. Oxford University Press.
Dobbs, D., 2000. The Law of Torts. West Group.
Fleming, J., 1988. The American Tort Process. Clarendon Press.
Fleming, J., 2011. In: Sappideen, Vines (Eds.), Fleming's Law of Torts, tenth ed. LawBook Co. Thompson Reuters.
Franklin, M., Rabin, R., Green, M., 2011. Tort Law and Alternatives, ninth ed. Foundation Press.
Geistfeld, M., 2008. Tort Law. Wolters Kluwer.
Kramer, O., Briffault, R., 1991. Workers Compensation: Strengthening the Social Compact. Insurance Information Institute Press.
Linden, A., 2003. The joy of torts. In: Beaulac, S., Pitel, S., Schulz, J. (Eds.), The Joy of Torts. Lexis-Nexis Butterworths.
Luntz, H., 2008. Assessment of Damages for Personal Injury and Death, fifth ed. Lexis-Nexis Butterworths.
Lytton, T., 2008. Holding Bishops Accountable. Harvard University Press.
Markesinis, B., Unberath, H., 2002. German Law of Torts: A Comparative Treatise, fourth ed. Hart Publishing.
Palmer, G., 1979. Compensation for Incapacity. Oxford University Press.
Portat, A., Stein, A., 2001. Tort Liability Under Uncertainty. Oxford University Press.
Rabin, R., Sugarman, S. (Eds.), 2003. Torts Stories. Foundation Press.
Shapo, M., 2012. An Injury Law Constitution. Oxford University Press.
Sugarman, S., 1989. Doing Away with Personal Injury Law. Quorum.
Sugarman, S., 2000. A century of change in personal injury law. California Law Review 88, 2403.
Sugarman, S., 2006. A comparative look at pain and suffering awards. DePaul Law Review 55, 399.

Torture and its Consequences, Psychology of

Danie Meyer-Parlapanis and Thomas Elbert, University of Konstanz, Konstanz, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Basoglu, volume 23, pp. 15779–15783, © 2001, Elsevier Ltd.

Abstract

Torture is utilized around the world to intimidate and punish individuals. The psychological damage created by torture has far-reaching effects on both the victim and the perpetrator. The extreme trauma leads to numerous mental health and behavioral complications, such as post-traumatic stress disorder, complex post-traumatic stress disorder, depression, anxiety, and addiction. Some individuals even report being fascinated by the violence, exhibiting appetitively aggressive behavior. narrative exposure therapy, and narrative exposure therapy for forensic offender rehabilitation are two mental health interventions designed to treat both torture victims and perpetrators, respectively, and assist in the reintegration process into society.

Introduction: Definition and Direction

Internationally, the currently utilized, institutional definition of torture provided by the United Nations Convention Against Torture delineates a top-down politicized act in which one person, under the acquiescence of governing authorities, intentionally inflicts severe pain or suffering on another with the objective of obtaining information, inflicting punishment or discriminating via intimidation. The definition concludes by emphasizing in Article 1, Section 1 that torture "does not include pain or suffering arising only from, inherent in or incidental to lawful sanctions" (United Nations, 1984). Specifically, demarcating torture as an event that occurs only outside of 'lawful sanctions' perilously politicizes the human experience of torture as an issue isolated to distant lands and cultures beyond the scope of Western legal standards. And when it does occur in the West, for instance, in prisons such as Guantanamo Bay, the torture practices utilized are neatly renamed interrogation techniques, and laws are shifted to redefine such practices within the requirements of legal sanction. In juxtaposition to popular thought, torture is far from being isolated to remote crisis regions or Western prisons detaining the terrorists stemming from those crisis regions. Torture is reported in numerous countries around the world (Amnesty International, 2013) and effectively establishes a feeling of fear, horror, and intimidation in the population at large.

Conceived in 1984, this outdated, sterilized definition is structured politically and omits the wide array of experiences and individual motivations of both the victim and the torturer. Unfortunately, torture is an event that is neither so simple nor so distant. From an educated, highly trained top-level government interrogator to an orphaned, hungry, child soldier in a rebel group to a gang member having caught a member from the rival gang, the varieties in the purpose of torture and how it is administered go as far as the limits of human creativity. Torture is as ubiquitous as people and as complex as the labyrinth of physiological and psychological mechanisms that constitute the human experience. A more comprehensive understanding of torture involves investigating both the neurological and the mental health underpinnings at the individual level of both the victim and perpetrator.

The Torture Victim: Experience and Consequences

A victim of torture can be forced to endure a wide array of terrifyingly painful physical and mental suffering. From extreme isolation, sleep deprivation, and highly uncomfortable sitting and standing positions to starvation, waterboarding, electric shock, limb removal, and fake executions, the victim never knows when the current treatment will end or what to expect next. Unable to escape, the victim reacts to the threatening stress of torture with a series of functionally distinct psychophysiological responses that operate as a defense cascade. This defense cascade follows the order of freeze-flight-fight-fright-flag-faint (see **Figure 1**; Schauer and Elbert, 2010). Initially freezing, the individual senses a threat, stops moving in an attempt to remain undetected, and increases attention. If the threat increases, the individual moves into flight or fight mode to protect him- or herself. Fight or flight mode increases sympathetic activation, thus enabling physiological reactions, such as tachycardia and peripheral vasoconstriction to minimize bleeding. Fear, but also anger and rage, enable the individual to fight back. The final stages of the defense cascade function to cope with situations in which the threat is overwhelming, fighting is useless, and the organism can only minimize damage. The individual becomes motionless, with motor paralysis, sensory deafferentation, and dissociative psychological conditions that range all the way up to a complete shutdown (Schauer and Elbert, 2010; Bracha, 2004; Bovin et al., 2008; Fusé et al., 2007). In trauma situations involving physical fighting and self-defense, the earlier phases of the defense cascade may be activated, while in situations in which the victim is overpowered, such as with torture or rape, the later phases are activated.

Typically, for nontraumatic memories, the initial sensory, emotional, physiological, and cognitive concomitants of an event, the 'hot memories,' gradually become integrated into declarative or 'cold' memory temporal context, can be recalled consciously and consequently are perceived as memories (Brewin et al., 1996, 2010; Brewin, 2011; Elbert and Schauer, 2002; Schauer et al., 2011). For example, when an adult recalls happily holding a puppy, the individual simultaneously remembers that the puppy was a gift, it was summer time, and family members except for the unborn

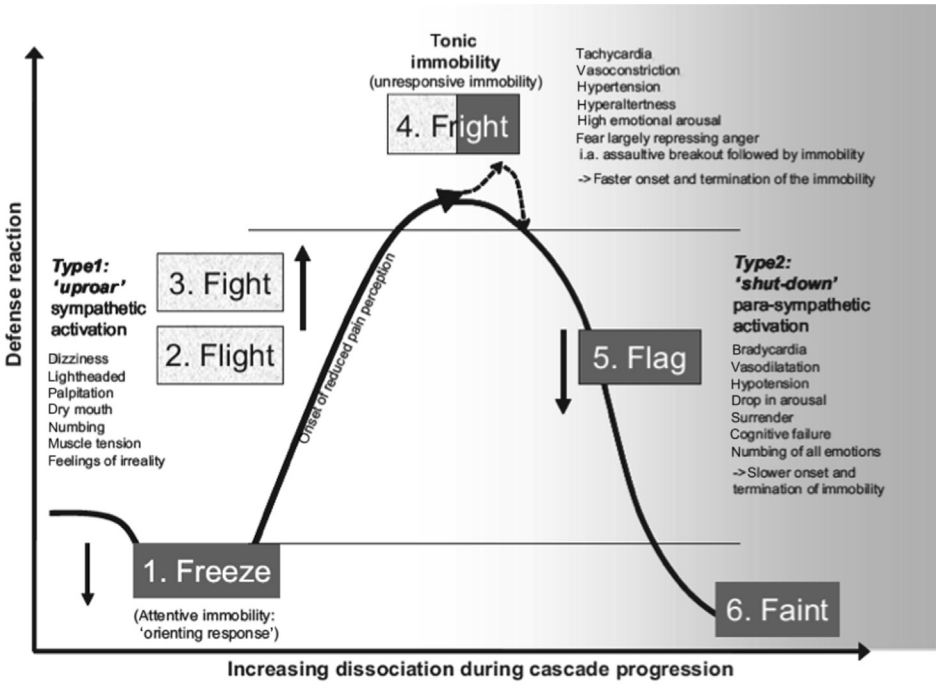

Figure 1 Schematic illustration of the defense cascade as it progresses along the 6-F course of action. The "uproar" sympathetic arousal reaches a maximum at the fright stage, eventually superseded by the onset of dissociative "shutdown" (gray area).

younger brother were present, and thereby it occurred at the person's fifth birthday party. Traumatic memories, however, function differently. The *hot memories* of traumatic events are so overwhelming that they do not consolidate with the *cold memory* temporal and spatial context (Brewin et al., 2010; Brewin, 2011). These experiences remain separate from autobiographical memory but still manifest themselves as vivid intrusions and nightmares. They seem to be lurking in the mind, and on the cue of the slightest, perhaps even unrecognizable, reminder of the trauma, they spring to the surface of the survivor's awareness. The sight of an officer's uniform, the sound of extremely loud heavy metal music, or the sensation of being under water at a swimming pool can reactivate the chain of connected associations that compose the hot memory, and the person perceives that he or she is once again in the torture situation. In such situations, although the survivor physically is located in the present, he or she remains mentally and psychologically situated in the past, in a network of experiences that never have been assigned the proper place and time (Brewin, 2011; Schauer et al., 2011).

The most characteristic symptom of trauma disorders is this intense reliving of these experiences as though they currently are happening, rather than being a part of a memory from the past (see Figure 2). When the torture survivor is suddenly in a situation with hot memory cues and the body begins to react as though the traumatic event is occurring in that moment, panic, horror, or dissociative responses set in (Elbert et al., 2011). At the initial stages of the defense cascade, the events are vividly present in all their components: sensory (smell of blood), physiological (heart racing, shortness of breath),

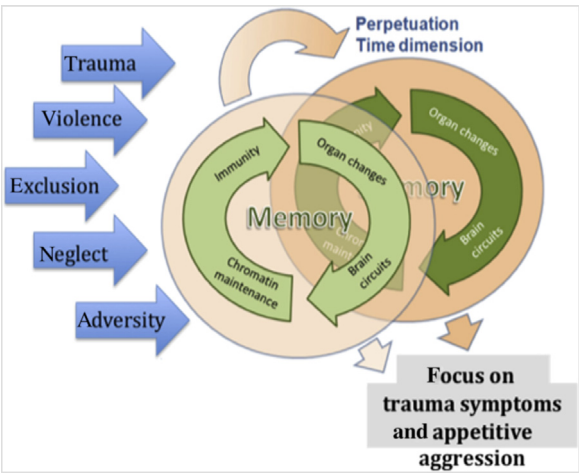

Figure 2 Multiple adversities affect a dynamic, self-organizing system that may continue to change, i.e., driven to a different system attractor with altered processing dynamics.

emotional (terror), and cognitions ('I'm going to die') (Schauer et al., 2011). The cold memory historical context placing the prior traumatic events in the correct time and space is missing, rendering the survivor seemingly defenseless to the physiological reactions to the perceived stress (see Figure 3). Torture survivors may have learned to rather rapidly switch to the later stages of the defense cascade, where dissociation sets in and current sensory information is gated and perceived as distant and unreal. Intrusions

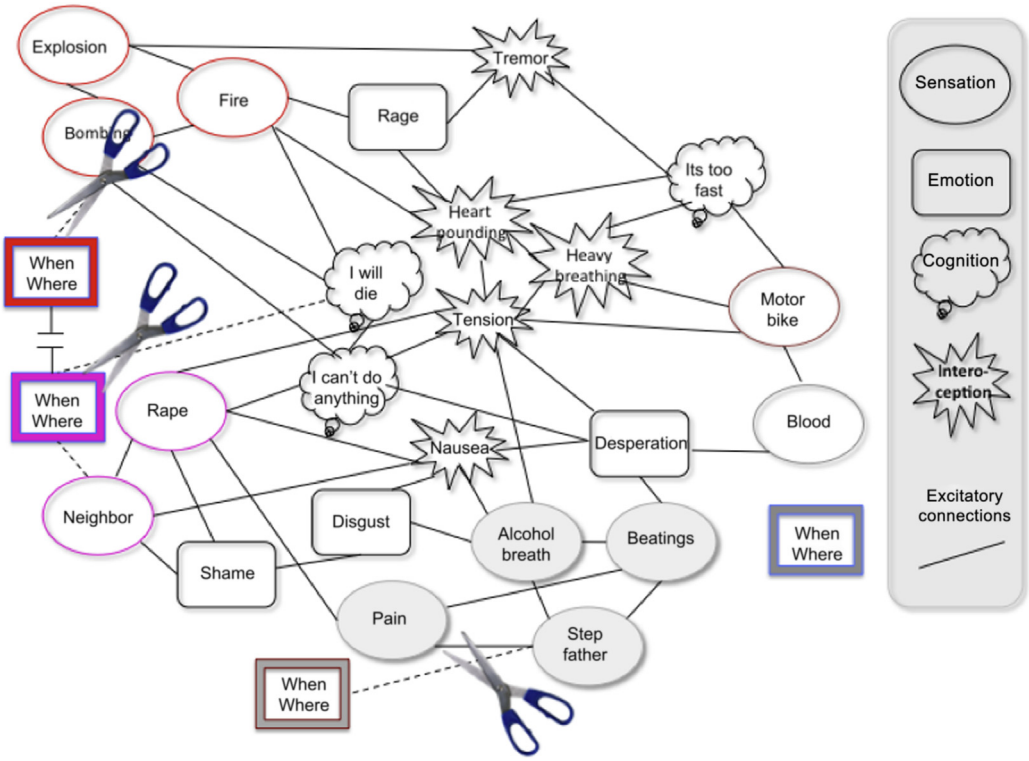

Figure 3 A fear/trauma network results from multiple fearful experiences: the representation of a single event may well connect to the context and episode. If, however, the perception cues an already existing trauma network, the context is lost while sensory, cognitive, emotional, and physiological representations interconnect with increasingly mutual excitatory power. The codes of "where" and "when," however, are not consistently co-activated as the brain's architecture does not support the simultaneous activation of two different places coded in the hippocampus. Thus the fear/trauma network becomes disconnected (symbolized by the scissors) from time and place, and the fear generalizes, giving rise to feelings of impending threat. Narrative Exposure Therapy is thought to reverse this process by contextualizing the trauma network. Adapted from Schauer, M., Elbert, T., 2010. Dissociation following traumatic stress. Zeitschrift für Psychologie/Journal of Psychology 218 (2), 109–127.

therefore can be understood as repetitive displays of parts or fragments of the event, which then elicit a corresponding combination of hyperarousal and dissociation, depending on the dominant physiological responding during the torture situations. Subsequent trauma-related disorders, such as post-traumatic stress disorder (PTSD) and other forms of disorders of extreme stress, represent the chronic reactivation of these defense cascade reactions (Elbert et al., 2011).

PTSD is the most frequently reported consequence of trauma, especially in cases of atrocities perpetrated by human beings (De Jong et al., 2001). PTSD is correlated linearly with traumatic experiences and outlines a type of building block effect for trauma in which each additional traumatic experience is accompanied by an increased likelihood that the individual eventually will develop PTSD (Neuner et al., 2004). It is a condition that occurs when an individual experiences an event that threatened survival or physical integrity. PTSD involves recurrent, disjointed, fragmental memories of the traumatic event, often in the form of flashbacks, that the sufferer strives to avoid, resulting in a hyperaroused, easily irritated, distracted state. It persists over longer periods of time and can be extremely debilitating in many areas of social and occupational functioning (American Psychiatric Association (DSM-V), 2013). The devastating effects of PTSD dramatically affect the transition of many torture survivors to their homes, and the individuals are markedly different upon return than when they left. Because of coverage by the mainstream media, the public acknowledges that many returning soldiers might contend with PTSD. Research is now beginning to acknowledge that soldiers, especially if they were taken as prisoners, might be dealing with trauma symptomology that affects the very sense of self with an altered personality. The changes are related to macroscopically altered brain function in torture survivors (Elbert et al., 2006, 2011; Ray et al., 2006) and structure (Eckart et al., 2011). Lateral prefrontal, parietal, and posterior midline structures have been implicated in the pathophysiology inflicted by torture (Eckart et al., 2011). As these regions are involved specifically in episodic memory, emotional processing, and executive control, this might have important implications for the understanding of the trauma-related symptoms. Long-term effects on the immune system (Sommershof et al., 2009) and overall physical health also have been noted in those individuals exposed to torture-related stress.

Recent work with torture survivors has demonstrated that prolonged adulthood trauma also can result in complex post-traumatic stress disorder (CPTSD) symptoms (McDonnell et al., 2013). CPTSD encompasses all of the symptoms

of PTSD and also incorporates the severe personality disorganization that can occur in some individuals following long periods of threat stress (Kolb, 1989). Conventionally, CPTSD has been associated with childhood trauma, however, researchers have shown that repeated and multiple traumatic events during adulthood lead to the breakdown of self-regulatory abilities – even in the absence of childhood trauma (Cloitre et al., 2009). Moreover, CPTSD symptoms are more closely related to the nature of the traumatic event rather than to the developmental stage of the individual at the time of the event (McDonnell et al., 2013). Perhaps the direct link of why CPTSD can occur in adults without childhood abuse owes to the captive nature of torture. Unable to fight or flee, the dissociative states associated with the latter half of the defense cascade, such as voluntary thought suppression, minimization, and denial, engage. These protective mechanisms can result in dissociation, depersonalization, and even changed identity as associated with CPTSD (Hárdi and Kroó, 2011). These responses, over a prolonged period of time and under various levels of static duress, can manifest in three disorders that are not part of PTSD diagnostic criteria: dissociative identity disorder, borderline personality disorder, and somatization disorder (Herman, 1992). Continued research with survivors of torture demonstrates that we have only begun to recognize the vast array of mental health complications with which torture survivors will have to contend.

Alongside PTSD and CPTSD, numerous comorbid conditions abound. Major Depression, Dysthymia, Phobias, Alcohol and Drug Abuse, and Conduct Disorder are among the most frequently observed disorders comorbid with PTSD (Kessler et al., 1995). Although it is possible for comorbid disorders to occur independently from the traumatic event, the comorbid conditions often are extensions of the PTSD symptomology (Schauer et al., 2011). For example, a torture survivor may just want to avoid thinking about past events by avoiding friends and family and drinking, behaviors possibly leading to a clinical depression or substance abuse diagnosis. Such an individual battling a trauma disorder, depression, and addiction among the various other socioeconomic and political obstacles involved with reintegrating into society would require specific, comprehensive treatment if reintegration to normal life is to be even remotely successful.

The Torture Victim: Narrative Exposure Therapy as a Treatment Option

Numerous treatment modalities are involved in treating trauma disorders, but few interventions are developed specifically to address the diverse complications of the trauma involved in torture. Narrative exposure therapy (NET) is an evidence-based treatment approach designed to treat survivors of organized violence who have endured multiple or continuous traumatic stressors (Schauer et al., 2011) and has proven effectiveness in torture survivors (Hensel-Dittmann et al., 2011; Neuner et al., 2010) (see **Figure 4**). NET addresses both the personal consequences and the political aspects of organized violence and torture. First, to reduce trauma symptoms, NET utilizes exposure therapy. Patients are confronted with their hot memories of traumatic events as means to connect their current emotional responses to the appropriate place in time from which they originate. Second, in an attempt to organize the fragmented, inconsistent autobiographical cold memories, the patient and therapist work together to reconstruct a consistent narrative of the survivor. The end goal is stabilized trauma symptomology, improved overall functioning, and a narrative that provides both an autobiographical structure for the survivor and testimonies that prove instrumental in human rights legislation. Evaluations by outside researchers demonstrate that NET is a promising, evidence-based intervention ideal for war-related trauma disorders (Robjant

Restructuring traumatic memories: **Narrative Exposure Therapy (NET)**

Based on the model presented in Figure 3, Schauer, Neuner and Elbert have developed the psychotherapy NET (2nd edition of the manual, Schauer, Neuner & Elbert, 2011).

In NET, the patient, with the assistance of the therapist, constructs a chronological narrative of her/his life story with a focus on the traumatic experiences. Fragmented reports of the traumatic experiences will be transformed into a coherent narrative. Empathic understanding, active listening, congruency and unconditional positive regard are key components of the therapist's behavior. For traumatic stress experiences, the therapist asks in detail for emotions, cognitions, sensory information and physiological responses and probes for respective observations. The patient is encouraged to relive these experiences while narrating without losing the connection to the "here and now." Using permanent reminders that the feelings and physiological responses result from memories, the therapist links these memories to episodic facts, i.e. time and place. The imagined exposure to the traumatic memories is not terminated until the related fear presented by the patient demonstrably reduces. In this way, the therapist is supportive yet directive in the elicitation of the narrative in order to counter avoidance and recover the full, implicit information of the trauma. As the end of treatment, the recorded autobiography may be used for human rights advocacy.

NET has several advantages compared to other treatment approaches. Although the duration of treatment is short (varies between 4 and 12 sessions have been tested), the method of narrating the entire life story does not require the patient to select a single traumatic event from their trauma history. Additionally, the fact that the survivor receives a written biography as a result of the treatment has turned out to be a major incentive especially for victims of human rights violations to complete treatment.

Figure 4 The therapeutic agents of NET: (1) active chronological reconstruction of all important events of the autobiographic/episodic memory - 'life line approach'; (2) prolonged exposure to the 'hot spots' and full activation of the fear memory in order to modify the emotional network; (3) meaningful linkage and integration of psycho-physiological and somato-sensory responses to the time-, space-, and life-context; (4) the cognitive reevaluation of behavior and patterns, as well as reinterpretation of the meaning-content through reprocessing of negative, fearful, and traumatic events; (5) regaining of survivor's dignity and satisfaction of the need for acknowledgment. Explicit human rights orientation of 'testifying.' Schauer, M., Neuner, F., Elbert, T., 2011. Narrative Exposure Therapy, second ed. Hogrefe & Huber, Göttingen.

and Fazel, 2010; Crumlish and O'Rourke, 2010; McPherson, 2012). Moreover, NET is a low-cost, relatively short-term intervention and is designed to be taught to local counselors, making its implementation feasible in most settings, regardless of infrastructure and access to resources (Jacob et al., in press).

After considering the mental health implications and suitable treatment options for the victim of torture, the journey to alleviate the consequences of and even to prevent torture is far from over. To prevent torture, we need to better understand not only the systems behind it but also the individuals perpetrating torture (i.e., their motivations, mental health, and rehabilitation options). Most research on torture largely focuses on understanding and treating the victim, all the while neglecting the mental health needs and rehabilitation of the perpetrator. Unfortunately, there is very little quantitative research on torturers themselves. Because all torturers are humans with brains and minds, and many are combatants with the specific task of torturing captives, combatant research, and general neurological studies are the primary available data sources from which to draw.

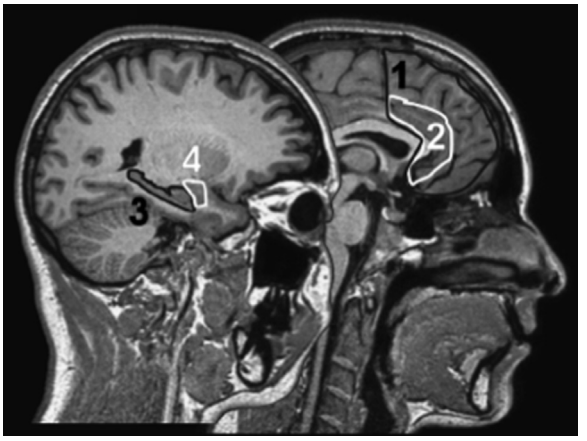

Figure 5 Major brain regions associated with PTSD: (1) prefrontal cortex; (2) including the anterior cingulated cortex; (3) hippocampus; (4) amygdala (from Kolassa, I., Elbert, T., 2007. Structural and functional neuroplasticity in relation to traumatic stress. Current Directions in Psychological Science 16 (6), 321–325).

The Torturer: Motivation

What motivates an individual to participate in torture? Perhaps some of the most informative clues about what motivates individuals lie in the overarching systems of torture training and implementation. One source of such training is in military organizations. Governments and militaries bear the burden of maintaining some semblance of transparency and accountability for their actions – for example, when a regime is accused of crimes against humanity, such as torture, they are required to provide a defense or explanation for such actions. One of the most prevailing rationales motivating the perpetration of torture is based on the belief system that torture yields accurate intelligence.

As mentioned, the United Nations publically lists the obtaining of information or confessions as a goal of torture. However, the objective of gathering intelligence using advanced interrogation techniques, that is, torture, raises red flags due to its stark inconsistency with current neurological research on memory function. O'Mara (2011) asserts that there is a popular belief that extreme stress or torture is an effective way to elicit intentionally withheld information and memories. There is little to no public research on the accuracy of the information that results from torture. If anything, interrogation techniques that involve extreme, consistent, and prolonged stress will achieve compromises in both the memory function and executive function (O'Mara, 2009). In other words, not only is the factual content of the desired intelligence contaminated but also the intention to withhold information is confounded. If stress is to be understood as the foreboding perception that present or future events will be very unpleasant and uncontrollable, and this perception results in heightened excitability in the brain and body (Kim and Diamond, 2002), torture can be categorized as an ultimate form of stress – both in intensity (consistent and overwhelming life threat) and duration (often 6 months or longer (O'Mara, 2011)).

The ultimate form of stress involved in torture has a significant neurological impact. The hippocampus and prefrontal cortex are the signature regions of the brain essential for normal memory function, and when they function improperly, as when under extreme stress, memory can be impaired (see **Figure 5**). When the stress condition persists over a prolonged period of time, normal physiological functioning of the brain regions governing memory is disrupted and can lead to tissue loss (Arnsten, 2009; O'Mara, 2011; Ulrich-Lai and Herman, 2009). For example, the prefrontal cortex becomes hyporesponsive, showing signs of tissue loss, while the hippocampus loses volume and slows in neurogenesis (Roozendaal et al., 2009). These reductive changes in brain structure and function result in conditions less than ideal for an interrogation. Furthermore, a common consequence of frontal lobe disorders like the ones listed earlier is confabulation, the pathological production of false memories (Turner et al., 2007). Confabulation likely would be one of the least desired effects for advanced interrogation sessions as the torturer tries to navigate around intentional misinformation (lies) and unintentional mistruths (confabulation) to uncover the intentionally withheld truths.

In summary, the notion that torture productively elicits accurate information unravels in the face of research on the neurological anomalies that result from prolonged stress. These anomalies evince that prolonged stress would counterproductively yield an indiscernible mix of accurate and inaccurate information. As there is no evidence that torture elicits significant amounts of accurate information and that the extreme stress involved in torture actually could produce false memories and misinformation, torture must persist for reasons other than collecting intelligence.

Perhaps the motivation behind the practice of torture is simple: power and control. The obtainment and maintenance of power and control more clearly pertain to the overarching authorities typically seeking to induce horror and intimidation in certain populations like ethnic or religious minorities. At the individual level, the mechanisms behind the transformation of a seemingly average civilian into a torturer can extend beyond exerting power and control over others so far as to develop a fascination in violent or aggressive behavior.

The Trained Torturer: Experience and Consequences

During a conflict, one group wants to dominate the enemy group, and torture is a clear way to demonstrate power and exert control by instilling fear. The torturer, however, is not a generic embodiment of groupthink. The torturer is a human being, very likely composed of all the same emotions and desires for fulfillment and survival as anyone else. In the same way that the victim adapts to the conditioning involved with torture, most torturers (excluding those who might be considered psychopaths and are not acting as part of a group) are similarly conditioned (Elbert et al., 2010). The individuals who torture others during conflicts most often are not renegade soldiers fulfilling their own malicious desires. The torture they commit is part of the broader goal of the group they represent, that is, a goal to dominate and intimidate. These individuals are carefully selected and trained. They often are given specific orders of what to do and how to do it. They are punished if they do not follow orders. They are cut off from their families and communities, made dependent on their superiors, and given promises of reward and honor. Some combatants who engaged in torture reported being physically abused by their commanders and witnessed public displays of horrifying acts of brutality and cruelty. Others were abused as children and adolescents. The terror of these experiences evolved into submission and desensitization, and cruelty became a passing, trivial act. The elaborate training process of torturers involves deliberately desensitizing the individuals to violence, diminishing the revulsion associated with gruesome acts through violent chants and songs to build camaraderie, and dehumanizing the target through group-specific language to make it easier to maim and kill the enemy (Gibson and Haritos-Fatouros, 1986). These combatants often are not in a position to easily escape their commitments to the group without fear of a fatal reprisal. One of the most terrifying reported development of some combatants is the gradual shift from violence as not only normal and acceptable but also exciting and a signal of skill (Maclure and Denov, 2006). They are swept up in a cycle of violence in which they have no alternative other than to continue participating in even more atrocities. Trapped, they have to choose between death and a brutal existence in which they will undoubtedly experience a change of their own personality.

The torturer is a combatant, sometimes with the same general warfare tasks as other soldiers and sometimes with the unique role of torturing captives. Sometimes torture is part of the role for all combatants in a group. For example, combatants from Sierra Leone reported fighting at the frontline, looting, burning villages, and the deliberate killing, maiming, and torturing of civilians (Maclure and Denov, 2006). In the chaos of armed conflict, combatants are often simultaneously both perpetrators and victims of violence (Betancourt et al., 2010; Maclure and Denov, 2006; Medeiros, 2007; Schauer and Elbert, 2010) and, as a result, can be left to contend with various mental health complications. Substance abuse commonly is documented among torturers (Rayfield, 2007). Additionally, perpetrators are vulnerable to developing PTSD, whether it be due to the traumatic experiences of being a victim (Annan et al., 2009; Elbert et al., 2006) or possibly stemming from the act of killing (MacNair, 2002). In addition to substance abuse and the broad spectrum of complications related to PTSD, combatants also are confronted with new attitudes about violence that could lead to aggressive antisocial behavior (Elbert et al., 2006; Weierstall and Elbert, 2011).

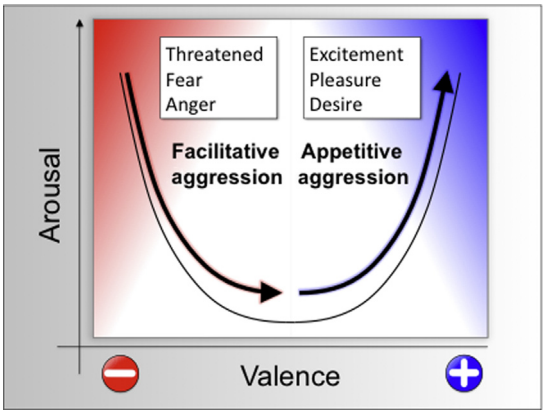

Figure 6 Facilitative aggression (reactive) usually associated with anger, encompasses any form of behavior that is directed to fighting another person and intends harm in order to reduce threat and fear, thus reducing the aversive arousing state. Appetitive aggression (predatory) also aims at harming others, however, the goal is to achieve a rewarding state which can result from the gains, but also simply from a 'desire' to hunt (from Weierstall, R., Elbert, T., 2011. The appetitive aggression scale. European Journal of Psychotraumatology 2, 8430. http://dx.doi.org/10.3402/ejpt.v2i0.8430).

As mentioned, former combatants report not only acclimating to perpetrating violence but also beginning to perceive violence as fascinating and appealing (Elbert et al., 2010; Maclure and Denov, 2006). They develop the fascination of a hunter – a human hunter (see **Figure 6**). This fascination to violence and aggressive behavior without the benefit of any immediate reward, appetitive aggression, is related to higher rates of perpetrated violence and demonstrates functioning as a buffer against PTSD under certain conditions (Elbert et al., 2010; Hecker et al., 2012; Weierstall et al., 2011, 2012). Although they initially might have been motivated to torture others to fulfill their duties or to preserve their own lives, some torturers have now grown to relish the violence and crave to repeat it for purely hedonistic purposes.

The combination of the instability of substance abuse and trauma symptomatology with the relishing of violence in highly trained and experienced combatants sets up an ominous scenario for the communities to which these perpetrators return. Regardless of any public opinion denouncing combatants and torturers as mere animals who had a moral choice and chose wrongly, these individuals are in dire need of mental health rehabilitation and reintegration assistance. How do we untrain the trained torturer?

Untraining the Trained Torturer: FORNET as a Treatment Option

Following a conflict, perpetrators face numerous challenges to successful reintegration. Effective programs typically offer

formal education and vocational training and try to instill hope for the future (Mädl et al., 2010). Unfortunately, trauma symptoms and aggressive behavior can thwart successful attendance, participation, and completion of even the most comprehensive reintegration programs. The effectiveness of NET is not limited to rehabilitating the victims of organized violence. We thus are developing NET for forensic offender rehabilitation (Elbert et al., 2012) that aims to reduce both PTSD symptoms and appetitive aggressive behavior by recalling the experiences, as both victim and perpetrator, through narrative exposure. It helps the former combatant to anchor not only fearful and traumatic experiences but also positive feelings that might have been linked to various forms of aggressive behavior in the past. The role change from a combatant to a civilian is addressed specifically and reinforced. Additionally, visions for the future are developed to foster successful integration into society. Although there are early FORNET pilot trials (Hermenau et al., 2013), future studies need to focus on whether survivors on both sides of this cycle of violence can profit from evidence-based methods.

Conclusion and Future Directions

The concept of torture needs to be updated continually by neurological and psychological research and broadened to encompass the myriad of motivations and consequences of torture for all affected parties – victim and perpetrator alike. Torture utilizes tactics on a wide spectrum of violence: from directly aggressive techniques, such as beatings, limb removal, and waterboarding, to more indirect forms of aggression, including threatening or kidnapping family members and extreme periods of isolation. Regardless of the technique utilized, neurocognitively speaking, pain is pain, whether it is physical or social pain (Eisenberger et al., 2003).

One common, current justification for subjecting individuals to such pain involves the reasoning that accurate intelligence is procured through torture. Further neurological and psychological empirical investigations will help to expose the sinister and hedonistic motivations behind why torture persists: motivations that likely have little to do with gathering intelligence and a great deal to do with punishment and intimidation. Using research to inform and update international standards is a solid and necessary means of generating human right's improvements on the topic of torture. Furthermore, expanding the understanding of tyranny of torture beyond an event situationally and geographically specific to soldiers and prisoners in military conflicts tucked away in war-torn regions will acknowledge the needs of countless individuals coming off of less obvious battlefields. The sooner there is more widespread, official recognition of the intricate mosaic of factors that both motivate and result from torture, the sooner clinical treatment can be further empirically researched, funded, and implemented.

See also: Mass Trauma: Psychopathological Effects across the Life Span.

Bibliography

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders: DSM-V. American Psychiatric Association, Washington, DC.

Amnesty International, 2013. Amnesty International Annual Report 2013. The State of the World's Human Rights. Amnesty International Ltd, London. ISBN: 978-0-86210-480-1.

Annan, J., Brier, M., Aryemo, F., 2009. From "rebel" to "returnee" daily life and reintegration for young soldiers in Northern Uganda. Journal of Adolescent Research 24 (6), 639–667.

Arnsten, A., 2009. Stress signaling pathways that impair prefrontal cortex structure and function. Nature Reviews Neuroscience 10 (6), 410–422.

Betancourt, T., et al., 2010. Sierra Leone's former child soldiers: a follow-up study of psychosocial adjustment and community reintegration. Child Development 81 (4), 1077–1095.

Bovin, M.J., Jager-Hyman, S., Gold, S.D., Marx, B.P., Sloan, D.M., 2008. Tonic immobility mediates the influence of peritraumatic fear and perceived inescapability on posttraumatic stress symptom severity among sexual assault survivors. Journal of Traumatic Stress 21, 402–409. http://dx.doi.org/10.1002/jts.20354.

Bracha, H.S., 2004. Freeze, flight, fight, fright, faint: adaptationist perspectives on the acute stress response spectrum. CNS Spectrums 9, 679–685.

Brewin, C.R., 2011. The nature and significance of memory disturbance in post traumatic stress disorder. Annual Review of Clinical Psychology 7, 203–227.

Brewin, C.R., Dalgleish, T., Joseph, S., 1996. A dual representation theory of posttraumatic stress disorder. Psychological Review 103 (4), 670–686.

Brewin, C.R., Gregory, J.D., Lipton, M., Burgess, N., 2010. Intrusive images in psychological disorders: characteristics, neural mechanisms, and treatment implications. Psychological Review 117 (1), 210–232. http://dx.doi.org/10.1037/a0018113.

Cloitre, M., Stolbach, B.C., Harman, J.K., Kolk, B., Pynoos, R., Wang, J., Petkova, E., 2009. A developmental approach to complex PTSD: childhood and adult cumulative trauma as predictors of symptom complexity. Journal of Traumatic Stress 22 (5), 399–408.

Crumlish, N., O'Rourke, K., 2010. A systematic review of treatments for post-traumatic stress disorder among refugees and asylum-seekers. The Journal of Nervous and Mental Disease 198, 237–251.

De Jong, J., Komproe, I., Van Ommeren, M., El Masri, M., Araya, M., Khaled, N., Somasundaram, D., 2001. Lifetime events and posttraumatic stress disorder in 4 postconflict settings. JAMA: The Journal of the American Medical Association 286 (5), 555–562.

Eckart, C., Stoppel, C., Kaufmann, J., Tempelmann, C., Hinrichs, H., Elbert, T., Kolassa, I., 2011. Structural alterations in lateral prefrontal, parietal and posterior midline regions of men with chronic posttraumatic stress disorder. Journal of Psychiatry and Neuroscience 36 (3), 176.

Eisenberger, N.I., Lieberman, M.D., Williams, K.D., 2003. Does rejection hurt? An fMRI study of social exclusion. Science 302 (5643), 290–292.

Elbert, T., Schauer, M., 2002. Burnt into memory. Nature 419 (6910), 883. http://dx.doi.org/10.1038/419883a.

Elbert, T., Weierstall, R., Schauer, M., 2010. Fascination violence: on mind and brain of man hunters. European Archives of Psychiatry and Clinical Neuroscience 260, 100–105.

Elbert, T., Rockstroh, B., Kolassa, I., Schauer, M., Neuener, F., 2006. The Influence of Organized Violence and Terror on Brain and Mind: A Co-constructive Perspective. Bibliothek der Universität Konstanz, Konstanz.

Elbert, T., Schauer, M., Ruf, M., Weierstall, R., Neuner, F., Rockstroh, B., Junghöfer, M., 2011. The tortured brain. Zeitschrift für Psychologie 219 (3), 167–174.

Elbert, T., Hermenau, K., Hecker, T., Weierstall, R., Schauer-Elbert, M., 2012. FORNET: Behandlung von traumatisierten und nicht-traumatisierten Gewalttätern mittels Narrativer Expositionstherapie. Interventionen bei Gewalt-und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie (S. 255–276). Medizinisch Wissenschaftliche Verlagsgesellschaft.

Fusé, T., Forsyth, J.P., Marx, B., Gallup, G.G., Weaver, S., 2007. Factor structure of the tonic immobility scale in female sexual assault survivors: an exploratory and confirmatory factor analysis. Journal of Anxiety Disorders 21, 265–283. http://dx.doi.org/10.1016/j.janxdis.2006.05.004.

Gibson, J., Haritos-Fatouros, M., 1986. The education of a torturer. Psychology Today 20 (11), 50–58.

Hárdi, L., Kroó, A., 2011. The trauma of torture and the rehabilitation of torture survivors. Zeitschrift für Psychologie/Journal of Psychology 219 (3), 133.

Hecker, T., Hermenau, K., Maedl, A., Elbert, T., Schauer, M., 2012. Appetitive aggression in former combatants – derived from the ongoing conflict in DR Congo. International Journal of Law and Psychiatry 35 (3), 244–249.

Hensel-Dittmann, D., Schauer, M., Ruf, M., Catani, C., Odenwald, M., Elbert, T., Neuner, F., 2011. Treatment of traumatized victims of war and torture: a randomized controlled comparison of narrative exposure therapy and stress inoculation training. Psychotherapy and Psychosomatics 80 (6), 345–352.

Herman, J.L., 1992. Complex PTSD: a syndrome in survivors of prolonged and repeated trauma. Journal of Traumatic Stress 5 (3), 377–391.

Hermenau, K., Hecker, T., Schaal, S., Maedl, A., Elbert, T., 2013. Addressing post-traumatic stress and aggression by means of narrative exposure: A randomized controlled trial with ex-combatants in the Eastern DRC. Journal of Aggression, Maltreatment & Trauma 22 (8), 916–934.

Jacob, N., Neuner, F., Schaal, S., Maedl, A., Elbert, T. Dissemination of psychotherapy for trauma-spectrum disorders in resource-poor countries: a randomized controlled trial in Rwanda. Psychotherapy and Psychosomatics, in press.

Kessler, R., Sonnega, A., Bromet, E., Hughes, M., Nelson, C., 1995. Posttraumatic stress disorder in the National Comorbidity Survey. Archives of General Psychiatry 52 (12), 1048.

Kim, J.J., Diamond, D.M., 2002. The stressed hippocampus, synaptic plasticity and lost memories. Nature Reviews Neuroscience 3 (6), 453–462.

Kolassa, I., Elbert, T., 2007. Structural and functional neuroplasticity in relation to traumatic stress. Current Directions in Psychological Science 16 (6), 321–325.

Kolb, L.C., 1989. Heterogeneity of PTSD. The American Journal of Psychiatry 146 (6), 811.

Maclure, R., Denov, M., 2006. "I didn't want to die, so I joined them": structuration and the process of becoming boy soldiers in Sierra Leone. Terrorism and Political Violence 18 (1), 73–85.

MacNair, R., 2002. Perpetration-induced traumatic stress in combat veterans. Peace and Conflict: Journal of Peace Psychology 8 (1), 63–72.

Mädl, A., Schauer, E., Odenwald, M., Elbert, T., 2010. Psychological rehabilitation of ex-combatants in non-western post-conflict settings. In: Matz, E. (Ed.), Trauma Rehabilitation after War and Conflict. Springer, New York, pp. 177–214.

McDonnell, M., Robjant, K., Katona, C., 2013. Complex posttraumatic stress disorder and survivors of human rights violations. Current Opinion in Psychiatry 26 (1), 1–6.

McPherson, J., 2012. Does Narrative exposure therapy reduce PTSD in survivors of mass violence? Research on Social Work Practice 22 (1), 29–42.

Medeiros, E., 2007. Integrating mental health into post-conflict rehabilitation: the case of Sierra Leonean and Liberian child soldiers. Journal of Health Psychology 12 (3), 498–504.

Neuner, F., Schauer, M., Karunakara, U., Klaschik, C., Robert, C., Elbert, T., 2004. Psychological trauma and evidence for enhanced vulnerability for posttraumatic stress disorder through previous trauma among West Nile refugees. BMC Psychiatry 4 (1), 34.

Neuner, F., Kurreck, S., Ruf, M., Odenwald, M., Elbert, T., Schauer, M., 2010. Can asylum-seekers with posttraumatic stress disorder be successfully treated? A randomized controlled pilot study. Cognitive Behavior Therapy 39 (2), 81–91.

O'Mara, S., 2009. Torturing the brain. Trends in Cognitive Sciences 13 (12), 497–500.

O'Mara, S., 2011. On the imposition of torture, an extreme stressor state, to extract information from memory: a baleful consequence of folk cognitive neurobiology. Zeitschrift für Psychologie/Journal of Psychology 219 (3), 159–166.

Ray, W., Odenwald, M., Neuner, F., Schauer, M., Ruf, M., Wienbruch, C., Elbert, T., 2006. Decoupling neural networks from reality dissociative experiences in torture victims are reflected in abnormal brain waves in left frontal cortex. Psychological Science 17 (10), 825–829.

Rayfield, D., 2007. Stalin and His Hangmen: The Tyrant and Those Who Killed for Him. Random House Digital, Inc., New York.

Robjant, K., Fazel, M., 2010. The emerging evidence for narrative exposure therapy: a review. Clinical Psychology Review 30 (8), 1030–1039.

Roozendaal, B., McEwen, B., Chattarji, S., 2009. Stress, memory and the amygdala. Nature Reviews Neuroscience 10 (6), 423–433.

Schauer, M., Elbert, T., 2010. Dissociation following traumatic stress. Zeitschrift für Psychologie/Journal of Psychology 218 (2), 109–127.

Schauer, M., Neuner, F., Elbert, T., 2011. Narrative Exposure Therapy, second ed. Hogrefe & Huber, Göttingen.

Sommershof, A., Aichinger, H., Engler, H., Adenauer, H., Catani, C., Boneberg, E., Kolassa, I., 2009. Substantial reduction of naïve and regulatory T cells following traumatic stress. Brain, Behavior, and Immunity 23 (8), 1117–1124.

Turner, M., Cipolotti, L., Yousry, T., Shallice, T., 2007. Confabulation: damage to a specific inferior medial prefrontal system. Cortex 44 (6), 637–648.

Ulrich-Lai, Y., Herman, J., 2009. Neural regulation of endocrine and autonomic stress responses. Nature Reviews Neuroscience 10 (6), 397–409.

UN General Assembly, 10 December, 1984. Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment. Treaty Series, vol. 1465. United Nations, p. 85. http://www.refworld.org/docid/3ae6b3a94.html (accessed 31.07.13.).

Weierstall, R., Elbert, T., 2011. The appetitive aggression scale. European Journal of Psychotraumatology 2, 8430. http://dx.doi.org/10.3402/ejpt.v2i0.8430.

Weierstall, R., Schaal, S., Schalinski, I., Dusingizemungu, J., Elbert, T., 2011. The thrill of being violent as an antidote of posttraumatic stress disorder in Rwandese genocide perpetrators. European Journal of Psychotraumatology 2, 6345. http://dx.doi.org/10.3402/ejpt.v2i0.6345.

Weierstall, R., Schalinski, I., Crombach, A., Hecker, T., Elbert, T., 2012. When combat prevents PTSD symptoms – results from a survey with former child soldiers in Northern Uganda. BMC Psychiatry 12 (1), 41.

Torture: Legal Perspectives

Alexa Koenig and Jonathan Simon, U.C. Berkeley School of Law, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a replacement of the previous edition article by G. Kaiser, volume 23, pp. 15783–15786, © 2001, Elsevier Ltd.

Abstract

Torture is considered one of the most egregious of domestic and international crimes. Yet throughout history, state authorities have used torture and other forms of cruel, inhuman, and degrading treatment as means of interrogation and punishment. Accordingly, a number of legal frameworks have been developed to discourage physical and psychological torture and to encourage accountability for perpetrators. This article addresses the legal phenomenon of torture, including definitions, historical understandings, key cases, and emerging issues.

Introduction

This article addresses the legal phenomenon of torture, including definitions, historical understandings, key cases, and emerging issues.

Definition

At its most basic, torture can be thought of as the use of pain to elicit information from or to punish individuals in state custody. However, various legal approaches expand on that concept. Today, there are three frameworks that broadly address the legal phenomenon of torture: international humanitarian law, which considers torture in the context of armed conflicts; human rights law, which is concerned with the dignity and protection of individuals subject to state control; and criminal law, both international and domestic.

The human rights and international humanitarian law frameworks that address torture were largely developed and expanded in the aftermath of World Wars I and II. Shocked by revelation of the 'hidden horrors' of those wars – which ranged from forced disappearances to gassing to medical experimentation on live subjects – the international community came together in 1945 to create a global legal framework to prevent such horrors from reoccurring, and to assert a universal consensus on 'the supreme value of the human person' (Universal Declaration of Human Rights website). An intricate system of treaties, commissions, covenants, and protocols emerged, outlining a universal body of fundamental rights and freedoms designed to protect against future abuse, to provide for legal accountability, and (in some cases) to offer remedies for any wrongs.

With regard to international humanitarian law, also known as the law of armed conflict, the most notable prohibition against torture can be found in the Geneva Conventions of 1949. Torture is recognized as a 'grave breach' of the Conventions, and thus as among the most serious of war crimes. Within the human rights framework, the 1948 Declaration of Human Rights confirmed the existence of a foundational norm against torture, the prohibition of which is recognized as part of customary international law (custom being the primary source of international law).

Today, a growing number of international and domestic laws prohibit torture. The Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT) provides the most comprehensive definition of torture. Article 1 defines torture as "any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of, or with the consent or acquiescence of, a public official or other person acting in an official capacity" (CAT, art. 1). Notably, the Convention never defines cruel, inhuman or degrading treatment: it states only that such acts 'which do not amount to torture' are also prohibited when "committed by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity" (CAT, art. 16, para. 1).

Drafted during the late 1970s and early 1980s and opened for signature on 4 February 1985, the Convention was signed by 25 countries within the first two months. Today, more than 150 states are parties to the Convention, reflecting near-universal condemnation of state-sanctioned abuse of individuals in state custody. According to Article 16 of the Convention, States Parties are obliged to prevent (and, presumably, to prosecute) acts of torture, as well as to prevent 'acts of cruel, inhuman or degrading treatment or punishment which do not amount to torture' committed by, or at the behest of, a public official (CAT, art. 16). Torture has been recognized by the world community as so despicable that it is categorized as one of only a few *jus cogens* crimes, meaning those crimes that are absolutely prohibited under all circumstances – no deviation from the prohibition against torture is permitted, even for states of emergency.

Despite the broad consensus against torture, how courts should interpret the crime's various elements has engendered significant debate. Issues have commonly focused on (1) whether the perpetrator was a state actor and/or was acting in an official capacity, (2) what is needed to demonstrate the perpetrator's 'intent' (e.g., whether all that is needed is evidence of an intent to perform the underlying act, versus evidence of intent to torture), and (3) the purpose for which

the offending act or acts occurred. Courts have especially grappled with how narrowly to construe each of these elements, and whether to focus strictly on the mental culpability of the perpetrator or to incorporate the subjective experiences of the victim. How these various elements are interpreted has a significant impact on the scope of the prohibition and whether anyone will be held legally accountable for its violation.

Historical and Theoretical Underpinnings

In order to fully understand what torture is from a legal perspective, it is important to know what the prohibition is meant to protect. Many scholars agree that the protection of human dignity is the underlying motivation not only for the prohibition against torture and other laws that prohibit ill treatment (Amnesty International USA, 2000), but also for the entire field of human rights and international humanitarian law (Kaufmann et al., 2011), and for domestic laws that undergird related harms, such as the US constitutional prohibition against cruel and unusual punishment (Simon, 2014). This concern with the protection of human dignity is evident in both case law and scholarship that considers the role of dignity in human rights protections. Thus, whenever one is considering possible interpretations of legal prohibitions against torture, it is critical to discuss the relationship between dignity and ill treatment, precisely because it is from the notion of dignity that all such prohibitions commence.

Arnd Pollmann provides a particularly helpful historical overview of the philosophy of human dignity. According to Pollman, human dignity was first acknowledged as an important concept during Roman Antiquity; this conception equated dignity with an exalted social status, as exemplified by the then-emergent term 'dignitary' (Pollmann, 2011, 246). The Middle Ages birthed a second conception, one intimately linked to Judeo-Christian theology, which reflected an idea of humanity as the 'pride of God's creation' and people as made in 'God's image' (Pollman, 2011, 246). Critically, while classical dignity resided in multiple forms of status as well as in virtuous conduct, the dignity that arises from the Hebrew concept of *kvod habriot*, or divine creation, applies to all human beings, regardless of their status or conduct, because of their creation in the image of God. Human dignity would become secularized during early modernity and the Italian Renaissance, as reflected in the writings of Immanuel Kant and Pico della Mirandola. For them, "As human beings we do not owe our dignity to the fact that we reflect a divine splendor but because we *ourselves* are capable of something great: of 'reason' and 'moral autonomy'" (Pollmann, 2011, 246). Yet an undercurrent of religiosity remained: René Cassin, a major contributor to the Universal Declaration of Human Rights, notably acknowledged the scriptural significance of dignity in his 1968 Nobel Prize acceptance speech (see Friedman, 2008).

The latest conception of human dignity, which is arguably the basis for the modern-day human rights regime, emerged in the aftermath of World War II. According to this conception, dignity is conditional; a person's particular circumstances dictate the extent to which he or she can live a dignified life. Accordingly, this conception birthed a perceived need for legally recognized and enforced rights to protect this critical human characteristic.

Pollman explains that per this fourth view, "although every human being—as a member of humankind—has an individual 'potential' to live a life in human dignity, the question whether we really are able to fully unfold this potential or capability is dependent upon the concrete circumstances of our life" (Pollmann, 2011, 250). Thus, dignity must be preserved and facilitated through law and related social norms.

Changing Tactics; Changing Laws

The ancient Greeks and Romans believed that physical torture was necessary to elicit truthful testimony for legal purposes. They also used torture as a form of punishment and as a means to discipline slaves. Depending on the location and era, practices included burning people alive, throwing people into pits, crucifixion, starvation, and more.

Such practices continued largely unabated and unquestioned for centuries. They also appeared across the world. Early European visitors to Asia reported witnessing torture, which was condoned by early Japanese and Chinese criminal codes. Spanish conquerors in the Americas tortured Indians in their search for gold, while English raiders in Africa tortured village leaders as a means to solicit the location of villagers, whom they would capture and sell as slaves. In the New World, 'witches' were also tortured, both as punishment and to encourage recantation of their presumed beliefs.

In response to the cruelties of the Inquisition and the barbarities of the Dark Ages, the eighteenth and nineteenth centuries witnessed a growing dissatisfaction with torture as a means of both punishment and information gathering. Foucault has credited two phenomena for this change. The first is the decline of punishment as a public spectacle. Whereas legal punishments used to be carried out in public, during the late 1700s and early 1800s punishment was increasingly driven indoors and/or underground; corporal practices ultimately gave way to an ever-growing use of imprisonment as a means to effectuate punishment. The second factor was a shift away from pain as a necessary component of punishment. Instead, the focus of punishment became one of control: control over the body, the mind, and the soul of the convicted (Foucault, 1975). These shifts resulted in many societies eventually outlawing torture.

Torture found an especially vocal critic in Cesare Beccaria, whose eighteenth-century treatise *On Crimes and Punishments* became a seminal work underlying the then-burgeoning anti-torture movement. But even his condemnation of the practice made little immediate impact on its use as a tool of inquisition and interrogation. Even as torture was increasingly criminalized, it remained a tactic used by police and other agents of the state; the practice just became more and more hidden. Accordingly, relatively unseen spaces within police stations and prisons – as opposed to public squares – became the predominate sites of modern-day torture.

In tandem with these changes, torture evolved from a practice focused almost exclusively on the body to one increasingly focused on the psyche. As has been pointed out by Darius Rejali

(taking an international perspective) and Richard Leo (domestically in the United States), western democracies in particular have increasingly employed psychological torture in place of its physical counterpart – engaging in practices that 'leave no marks' – in order to avoid international criticism for the torture of prisoners (Rejali, 2007; Leo, 2008). While both international and domestic laws expressly condemn psychological torture, the difficulty in gathering evidence of practices that leave few lasting and/or physical marks has meant that modern-day torture can be especially difficult to prove; because of this, accountability suffers, despite numerous studies that reveal that the psychological sequelae of torture often last longer than physical harms.

Torture has also evolved from a problem of active abuse – direct assaults on the body or mind – to a problem of deliberate indifference to the medical and mental health needs of prisoners and detainees. As US Supreme Court Justice Anthony Kennedy wrote in the 2011 Supreme Court case *Brown v. Plata*, borrowing from a Supreme Court case decided more than a century earlier, "A prison's failure to provide sustenance for inmates may … produce physical torture or a lingering death."

Key Cases

Despite such recognitions, the law has struggled to keep up with changing tactics. Several key cases help illuminate how the postmodern world conceptualizes torture.

The foundation for a modern-day, international jurisprudence of torture was laid in 1969 in the *Greek Case* (1969 Y.B. Eur. Conv. on H.R.), the first case ever brought for torture under the authority of the European Convention of Human Rights. In that case, the European Commission on Human Rights had to determine whether particular practices qualified as torture or as the lesser crimes of inhuman or degrading treatment. Ultimately, they declared that 'all torture must be inhuman and degrading,' but not all inhuman and degrading treatment is torture (Reidy, 2002, 11). Instead, they reserved the term torture for those extreme forms of inhuman treatment that have, as their 'purpose,' obtaining information or confessions, or inflicting punishment.

This seminal case is critical for establishing two of three dominant ways of conceptualizing the relationship between torture and cruel, inhuman, and degrading treatment. In the decision, a first understanding of the relationship – as a hierarchy that ranges from degrading treatment at one extreme, through inhuman treatment, and ultimately to torture – was emphasized. In this way, torture was conceived as an 'extreme' form of inhuman treatment. However, it was also conceived as one that requires an additional element to inhuman treatment, the proving of one of several prohibited 'purposes.' As the Commission explained in the *Greek Case*, "The word 'torture' is often used to describe inhuman treatment, which has a purpose, such as the obtaining of information or confessions, or the infliction of punishment, and it is generally an aggravated form of inhuman treatment."

This first, hierarchical approach was largely adopted by the European Court of Human Rights when, in 1978, it decided what would arguably become the most influential and foundational international case to address the prohibition against torture and cruel, inhuman, and degrading treatment. In that case, *Ireland v. United Kingdom*, the European Court of Human Rights debated whether the British authorities had tortured detainees in Northern Ireland. There, the court embraced the notion that the difference between torture and cruel, inhuman, and degrading treatment is merely one of degree when it concluded that "th[e] difference derives principally from a difference in the intensity of the suffering inflicted" (*Ireland v. United Kingdom*, para. 167). The Court ultimately held that the institutional treatment of detainees (which included wall standing, hooding, and depriving detainees of food and sleep) did not descend into torture. However, they concluded the techniques used on the prisoners were cruel, and therefore still illegal (*Ireland v. United Kingdom*, para. 167).

The European Court of Human Rights has been heavily criticized for its determination in the *Ireland* case that these practices were not torture, and especially for the reasoning upon which it relied to make its determination. As written by one critic, "[t]he concept of torture [embraced by the court] is one that requires extreme intensity of pain or suffering. It is significant in this regard that a subjective test, assessing the ill-treatment from the point of view of the victim, was not applied by the Court, which chose not to call witnesses. … By failing to engage with the victims' point of view, the Court overlooked a fundamental aspect of how to assess the severity of ill-treatment. It also merits attention that by defining torture primarily in terms of its physical intensity, the practice of torture is narrowly conceived by the Court so that it does not include acts that are not exceptionally [physically] brutal" (Cullen, 2008).

As Justice Evrigenis, who drafted a dissenting opinion in the case, further warned, such an objective perspective fails to encompass modern, predominately psychological forms, especially inventive techniques that have yet to emerge. As he noted, "Torture no longer presupposes [physical] violence, a notion to which the judgment refers expressly and generically. Torture can be practiced—and indeed is practiced—by using subtle techniques. … By means of new forms of suffering that have little in common with the physical pain caused by conventional torture it aims to bring about, even if only temporarily, the disintegration of an individual's personality, the shattering of his mental and psychological equilibrium and the crushing of his will" (*Ireland v. United Kingdom*, dissenting opinion of Judge Evrigenis).

Indeed, as has been written about in the contemporary era by law professor Richard Leo, it was precisely during this post–World War II period that interrogators increasingly favored psychological, 'rationalized,' and 'scientific' forms of torture that left no marks on victims and thus could more readily evade detection than their physical counterparts (Leo, 2008).

The tendency of the *Ireland* decision to overemphasize stereotypical, historic modes of physical torture and underemphasize psychologically based wrongs is especially prescient in the warning offered by Judge Evrigenis when the decision was released: "The notion of torture which emerges from the judgment is in fact too limited. By adding to the notion of torture the notions of inhuman and degrading treatment, those

who drew up the Convention wished, following Article 5 of the Universal Declaration of Human Rights, to extend the prohibition in Article 3 of the Convention [Against Torture]—in principle directed against torture—[to] other categories of acts causing intolerable suffering to individuals or affecting their dignity, rather than to exclude from the traditional notion of torture certain apparently less serious forms of torture and to place them in the category of inhuman treatment which carries less of a 'stigma'—to use the word appearing in the judgment" (*Ireland v. United Kingdom*, separate opinion of Judge Evrigenis).

Two decades later, in 1997, the United Nations Committee Against Torture (the body of experts that reviews state compliance with the Torture Convention's obligations) found that the especially harsh treatment of prisoners by Israel had violated the Convention Against Torture's prohibitions against cruelty *and* torture (Committee against Torture, Concluding observations of the Committee against Torture: Israel, para. 257). There, the practices under investigation included the 'standard' use of multiple prison-based practices, especially in combination: restraining individuals in painful positions, hooding them, exposing them to loud music, depriving them of sleep for prolonged periods, threatening them, shaking them, and using cold air to chill them. Two years later, Israel's High Court similarly ruled that certain interrogation methods used by Israel's General Security Service – including shaking, sleep deprivation, and use of the shabach position (the shackling of detainees in a painful configuration, typically while hooded and bombarded with loud music) – were illegal, although they avoided saying these practices amounted to torture (*Public Committee Against Torture in Israel v. Israel*, 1999, HCJ 5100/94). Instead, the Israeli High Court referenced *Ireland v. United Kingdom* to note that, in Ireland, similar interrogation techniques were found inhumane and degrading but not torturous and, thus, were barred only on that lesser ground. On that basis, the High Court ruled that the 'combination' of techniques gave rise to pain and suffering such that they should be prohibited as cruel and inhuman, although not as torture.

While the Israeli High Court largely reiterated the reasoning in *Ireland*, this case underscores a third major approach that courts have come to take when interpreting the relationship between torture and cruel, inhuman, and degrading treatment: a contextual approach, one more akin to a 'totality of the circumstances' analysis than a sliding scale.

The decade from 1990 to 2000, when these latter cases were decided, ultimately proved one of the most prolific with regard to interpreting international prohibitions against torture and other forms of cruel, inhuman, and degrading treatment. As these cases were proceeding through treaty-based courts and the court of public opinion, two new international tribunals were being formed. The first since Nuremberg and Tokyo to determine accountability for horrors committed in the context of specific conflicts, the International Tribunals for the former Yugoslavia (ICTY) and Rwanda (ICTR) emerged in the mid-1990s. Their chambers had to quickly establish a jurisprudence to expound upon the crimes articulated in their statutory law, which includes prohibitions against torture.

Through these tribunals, legal interpretations of torture expanded considerably. One of the ICTY's most notable torture-related cases was the *Furundzija* case (IT-95-17/1). There, the trial chamber defined torture as the "intentional infliction, by act or omission, of severe pain or suffering, whether physical or mental, for the purpose of obtaining information or a confession or of punishing, intimidating, humiliating or coercing the victim or a third person, or of discriminating on any ground against the victim or a third person. For such an act to constitute torture, one of the parties thereto must be a public official or must, at any rate, act in a non-private capacity, e.g., as a de facto organ of a State or any other authority wielding entity."

One of the most significant advancements to emerge from this tribunal – in both the *Celebici* (IT-96-21) and *Furundzija* cases – was recognition that rape can be a form of torture. The ICTR, in the *Akayesu* case (ICTR-96-4-T), also acknowledged that rape can constitute torture. Such findings went a long way to further international case law related to both sexual violence and torture, increasing the possibility for recognizing and ensuring accountability for harms committed against women and children.

The International Criminal Court, which came into being in 2002, has benefited from these earlier attempts to adjudicate torture cases. As the world's first permanent international criminal tribunal, the Court's stated purpose is to ensure accountability for the most serious crimes of international concern. Torture is expressly criminalized in the Rome Statute, the Court's founding treaty, as a war crime and crime against humanity. This relatively nascent court provides yet another forum for adjudicating claims of torture when domestic jurisdictions are unable or unwilling to act.

Emerging Issues and Practices

Despite these advancements, significant gaps remain in the law's ability to ensure accountability for torture, especially as abusive tactics continue to evolve. For example, while torture has traditionally been conceptualized as practices by state actors against individuals, modern warfare and imprisonment have increasingly been outsourced to private parties. Increased privatization has blurred the public–private divide and proven problematic, especially when private companies or individuals act only indirectly on a state's behalf.

There are also troubling signs that torture is re-emerging as a form of punishment. While the renewed use of torture for interrogation purposes has received significant attention following the 11 September 2001 attacks on the United States by al Qaeda operatives, the use of torture as a means of prison-based punishment has received less attention.

Prior to the eighteenth century, torturous punishments were common in Western justice systems (Foucault, 1975). However, the adoption of the Eighth Amendment's prohibition on cruel and unusual punishment was intended to forbid punishments that approximated torture in their degree of pain. The rejection of torture, along with a significant reduction in the use of capital punishment, was a primary motive underscoring the development of imprisonment as the predominate punishment

for serious crimes in the West beginning in the early nineteenth century. However, by the late twentieth century, a politicized fear of crime led to unprecedented increases in prison populations in the United States and in other Western countries (Garland, 2001; Simon, 2007), and the adoption of exceedingly long and harsh prison sentences. Chronic overcrowding (where prison populations sometimes exceeded 200% of capacity), combined with lack of adequate health care, led the US Supreme Court in *Brown v. Plata* (2011) to affirm a massive prisoner population reduction order under the Eighth Amendment and to warn that a prison's failure to provide adequate sustenance for inmates can produce physical torture.

Because no individual state actor is deliberately inflicting pain or damage in such cases, some adjudicators would reject overcrowding and neglect as examples of torture. However, since the results are a predictable product of deliberate incarceration policies, including a deliberate emphasis on security over health care and rehabilitation, many would argue otherwise. Thus, while torture has historically been an intentional instrument, contemporary torture can often be the product of 'deliberate indifference.'

The United States' prison-based use of extreme isolation has also been cited as a possible form of torture. Toward the end of the twentieth century, particularly harsh penal policies led most US states and the federal prison system to adopt a severe form of solitary confinement known as 'supermax' or a 'secure housing unit' in which prisoners are kept in near-total isolation – either alone or with one other inmate – for months, years, or even decades. Because sustained solitary confinement can lead to mental illness and suicide, many believe such extreme forms of incarceration should be considered cruel, inhuman, and degrading at a minimum, and possibly even torture.

Finally, empirical and theoretical research increasingly suggests the importance of taking into consideration victims' experiences in addition to a perpetrator's intent when defining torture. For example, there is growing recognition of the cumulative effect of abusive prison practices; empirical observations of prisoners' experiences strongly suggest that even when a single practice is not experienced as torture on its own, the collective use of abusive practices can add up to an overall experience of torture (Stover and Nightingale, 1985; Basoglu, 2007; Koenig et al., 2009).

The potential cumulative effect of ill treatment became particularly salient in the aftermath of evidence that the United States tortured detainees in the context of its post-2001 'war on terror.' For example, in 2009, a Pentagon official who was responsible for deciding whether detainees should be brought before military commissions declared that the practices used on a Saudi national held at the Guantanamo Bay Detainment Facility had amounted to torture. After reviewing his file, she chose not to refer the detainee's case for prosecution, declaring that "the techniques they used were all authorized, but the manner in which they applied them was overly aggressive and too persistent. You think of torture, you think of some horrendous physical act done to an individual. This was not any one particular act; this was just a combination of things that had a medical impact on him, that hurt his health. It was abusive and uncalled for. And … clearly coercive. It was the medical impact that pushed me over the edge [to call it torture]" (quoted in Woodward, 2009).

Despite several government and nongovernment reports documenting the torture of detainees in the twenty-first century, very few individuals have been brought to task for such abuse. The lack of successful prosecutions confirms the extent to which political will and diplomatic pressure remain critical stumbling blocks to ensuring accountability for this especially heinous crime.

See also: Human Rights and Social Work; International Law and Treaties; Mass Imprisonment and Its Consequences; Nazi Law; Prisons and Imprisonment; Punishment: Social and Legal Aspects; Socialist Law; Torture and its Consequences, Psychology of.

Bibliography

Amnesty International USA, 2000. Torture Worldwide: An Affront to Human Dignity. Amnesty International, New York.

Basoglu, M., 2007. Torture vs other cruel, inhuman and degrading treatment: is the distinction real or apparent? Archives of General Psychiatry 64, 277–285.

Beccaria, C., 1764. On Crimes and Punishments. Empire Books, North Carolina (reprinted 2012).

Cullen, A., 2008. Defining torture in international law: a critique of the concept employed by the European Court of Human Rights. California Western International Law Journal 34, 29–46.

Donahue, D.E., 1980. Human rights in Northern Ireland: Ireland v. the United Kingdom. Boston College International and Comparative Law Review 3, 377–432.

Foucault, M., 1975. Discipline and Punish: The Birth of the Prison. Vintage Books, New York (reprinted 1995).

Friedman, H.H., 2008. Human dignity and the Jewish tradition. Jewish Law Blog. http://www.jlaw.com.

Garland, D., 2001. The Culture of Control: Crime and Social Order in Contemporary Society. University of Chicago Press, Chicago.

Kaufmann, P., Kuch, H., Neuhauser, C., Webster, E. (Eds.), 2011. Humiliation, Degradation, Dehumanization: Human Dignity Violated. Springer Science + Business Media, New York.

Koenig, K.A., Stover, E., Fletcher, L.E., 2009. The cumulative effect: a medico-legal approach to United States torture law and policy. Essex Human Rights Review 6, 145–168.

Leo, R.A., 2008. Police Interrogation and American Justice. Harvard University Press, Boston, Massachusetts.

Maier, A., 2011. Torture: how denying moral standing violates human dignity. In: Kaufmann, P., Kuch, H., Neuhauser, C., Webster, E. (Eds.), Humiliation, Degradation, Dehumanization: Human Dignity Violated. Springer Science + Business Media, New York.

Pollmann, A., 2011. Embodied self-respect and the fragility of human dignity: a human rights approach. In: Kaufmann, P., Kuch, H., Neuhauser, C., Webster, E. (Eds.), Humiliation, Degradation, Dehumanization: Human Dignity Violated. Springer Science + Business Media, New York.

Reidy, A., 2002. The Prohibition of Torture: A Guide to the Implementation of Article 3 of the European Convention on Human Rights, Human Rights Handbooks No. 6. Council of Europe.

Rejali, D., 2007. Torture and Democracy. Princeton University Press, Princeton, New Jersey.

Scarry, E., 1985. The Body in Pain: The Making and Unmaking of the World. Oxford University Press, Oxford.

Simon, J., 2007. Governing through Crime: How the War on Crime Transformed American Democracy and Created a Culture of Fear. Oxford University Press, New York.

Simon, J., 2014. Mass Incarceration on Trial: America's Courts and the Future of Imprisonment. New Press, New York.

Stover, E., Nightingale, E., 1985. The Breaking of Bodies and Minds. W.H. Freeman & Co., Ltd, London.

Tanaka, Y., 1996. Hidden Horrors: Japanese War Crimes in World War II. Westview Press, Boulder, Colorado.

Woodward, B., 2009. Guantanamo Detainee was Tortured, Says Official Overseeing Military Trials. Washington Post, Washington, DC.

Relevant Websites

http://www.apt.ch – Association for the Prevention of Torture.
http://www2.ohchr.org/english/bodies/cat – Office of the United Nations High Commissioner for Human Rights: Committee Against Torture.
http://www2.ohchr.org/english/law/cat.htm – Office of the United Nations High Commissioner for Human Rights: Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment.
http://untreaty.un.org/cod/icc/statute/romefra.htm – Rome Statute of the International Criminal Court.
http://www.un.org/en/documents/udhr/index.shtml – Universal Declaration of Human Rights.

Totalitarianism

Leslie Holmes, School of Social and Political Sciences, University of Melbourne, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Totalitarianism was very much a twentieth century concept, and was applied primarily to Communist and fascist systems. This article examines the diverse interpretations of totalitarianism, and, in considering its defining features, explores the ways in which it is seen to differ from authoritarianism. Having analyzed the development of the concept, the concluding section considers its contemporary relevance and applicability to political developments in the twenty-first century.

Introduction

Totalitarianism was a major concept of the twentieth century. Yet both its meaning and its application were contested. While most political scientists believed that the term summarized well the key features of fascism (inc. Nazism) and Communism, historians, such as Alan Bullock (1998), Ian Kershaw (2000) and to some extent Richard Overy (2005) argued that the differences between Nazi Germany and the Soviet Union, even under Stalin, were sufficient for it to be misleading to use the same word to describe both. (In this article, and following a common convention, Fascism (upper case) refers explicitly to Mussolini's Italy, whereas fascism is a more general term to describe right-wing totalitarian systems. The term Communist (upper case) refers here to actual systems, to distinguish it from the ultimate ideal or goal of communism. The present tense is used for conceptual analysis, the past when referring to actual historical phenomena.) Although there unquestionably were significant distinctions between fascism and Communism, particularly in terms of the role and nature of ideology and the end goal, the dominant view nowadays is that the term is a useful shorthand way of describing a range of systems that, on balance, had more commonalities than differences, particularly in terms of their style and means. As will be demonstrated, while having much in common and in practice often overlapping with authoritarianism, totalitarianism is conceptually distinct.

This article begins by considering the defining features of totalitarianism, and then outlines the development of the concept in terms of its historical and intellectual contexts. Following this is a brief evaluation of the term. The final section concerns its relevance today and what this might be in the future.

Defining Features of Totalitarianism

As suggested by the root of the word, totalitarianism refers to a system in which control over the population is greater – closer to 'total' – than in any other type of political system. It is the most extreme version of subordination of the individual to the state, and to a questionable notion of the collective. It represents the politicization of all aspects of life. Totalitarian and authoritarian dictatorships have much in common, including a lack of respect for citizens' rights, for the public accountability of political elites, and for the rule of law. But a totalitarian dictatorship is distinguishable from an authoritarian one in several ways (see Linz, 1975, 2000 for a more detailed analysis).

First, ideology plays a far more significant role in the former than in the latter, even though many observers agree that ideology was less important in Fascist Italy and Nazi Germany than in Stalin's USSR and other Communist states. While totalitarian ideology often makes use of the past, particularly national traditions, it is fundamentally radical and future-oriented ('chiliastic' or 'teleological'), whereas authoritarian regimes typically focus more on the past and continuity.

Closely related to the first point is that the role of propaganda in totalitarian states is much greater than is typical in authoritarian states. Whereas the term propaganda has a negative connotation in liberal democracies, it has a positive one in totalitarian states. In the USSR, 'agitprop' (agitation and propaganda) was an important aspect of the Communist Party's role, while Hitler appointed one of his closest and most trusted colleagues, Joseph Goebbels, to head the Ministry of Propaganda and Enlightenment. Simple slogans – such as 'Credere, Obbedire, Combattere' (Believe, Obey, Fight) in Fascist Italy; 'Arbeit macht Frei' (Work Liberates) and 'Ein Volk, ein Reich, ein Führer' (One People, One Empire, One Leader) in Nazi Germany; or 'Za Rodinu! Za Stalina! Za Mir! Za Kommunizm!' (For the Motherland! For Stalin! For Peace! For Communism!) in Stalin's Russia – were intended to convey the state's message straightforwardly but powerfully to the populace.

Totalitarianism is also often overtly exclusionary, and arbitrarily identifies particular minority groups as 'enemies' of the state and/or the people (e.g., Jews, gays, and communists in Nazi Germany; 'the bourgeoisie' and kulaks – rich peasants – in Stalin's USSR; Roma, gays, socialists, and communists in Fascist Italy), for whom life could become literally unbearable. In some cases, citizens could, if they chose to, renounce their allegiance; this applied to membership of a political party that was considered by the authorities to be hostile to the totalitarian system, but if one were Jewish or Roma or gay/lesbian, such options were not available. And sometimes, as in Stalin's USSR, social class labeling by the state could be highly arbitrary anyway.

This high level of apparent arbitrariness is related to another distinguishing feature of totalitarianism: its use of terror. Thus, the Nazis are remembered for the holocaust and concentration

camps and Stalin for the labor camps administered by the GULag. The exact number of deaths that resulted from such state terror is disputed, but undoubtedly runs into many millions. In terms of the *proportion* of the population who died at the hands of a totalitarian system, however, Pol Pot's Communist Cambodia was the worst of all.

In theory at least, the single political party plays a more prominent role in a totalitarian regime than in an authoritarian one. In the latter, it is often the military or a small clique that controls the state. There were unquestionably military elements integral to the rule in totalitarian states. For example, Hitler, Mussolini, and Stalin sometimes appeared in military uniform, while their more familiar party uniforms were clearly influenced by military ones. However, the military was ultimately subservient to the party.

Totalitarian systems are often held to be distinguished from authoritarian ones in terms of the image of the leader. The leading totalitarian systems all had putatively charismatic leaders – a feature found far less often in authoritarian systems – although some have challenged the notion that the charisma was real, rather than merely the result of artificial personality cults promoted by the state authorities, especially in the case of Stalin. But it must be acknowledged that many Soviets, Germans, and Italians did hold their leaders in awe and mourned their passing. The image of strong leaders was reflected in the terms used to describe them in official propaganda; *der Führer*, *Il Duce*, and *Vozhd'* all connote *the* ultimate leader or chief in German, Italian, and Russian, respectively. Citizens were often expected to show respect for the leader in the form of salutes, as in 'Heil Hitler' (Hail Hitler) in Germany; in Italy, the Fascists adopted the 'Roman Salute,' and by the 1930s were requiring citizens to use this instead of the 'bourgeois' and 'unhygienic' handshake. The leader was often portrayed as a kindly father figure to the nation; good examples include China's Mao Zedong, Vietnam's Ho Chi Minh, and North Korea's Kim II Sung.

Authoritarian systems typically permit a limited amount of pluralism. This might be in the form of (often manipulated) competitive elections. In contrast, totalitarian systems do not permit pluralism, and even forbid overt differences of opinion within the party.

It has been argued by Linz (1975, 2000) that citizen mobilization is typically low – neither extensive nor intensive – in authoritarian systems, but high in totalitarian ones. The totalitarian state wants its citizens to demonstrate their loyalty in a much more public and overt way than does the authoritarian state. It was because Communist systems expected their citizens to express devotion to the state that what might appear to an outsider to be a symbolically trivial gesture made by the Hungarian leader in 1961 caused such a stir. Until then, like other Communist states, Hungary claimed to assume that 'if you're not with us, you're against us'; but in that year, János Kádár reversed this, saying that in future the assumption would be that 'if you're not against us, you're with us.' In terms of citizen devotion to the cause expected by the authorities, both fascist and most Communist states even encouraged women to bear more children, so as to increase the population and hence the strength and size of the totalitarian state.

While most authoritarian systems do not seek to control citizens' personal beliefs, notably their religion, Communist totalitarian states did. Indeed, James Gregor (2012) has argued that the totalitarian ideology was itself a form of religion – even where there was official commitment to secularization – with its focus on selfless sacrifice, messianism, and a glorious future. This said, it should be noted that the fascist and Communist approaches to the organized church were quite different, and there needs to be nuancing on this variable. In Japan, the fascists promoted 'State Shintoism,' which essentially linked the state with the country's traditional religion and nationalism. Mussolini knew that he needed to reach a *modus vivendi* with the powerful Roman Catholic Church in Italy, and so designated it as the sole state religion (i.e., other religions were banned). Hitler had a complex approach to religion, wanting to ban the Old Testament because of its Jewish connotations, treating Catholicism more harshly than Protestantism, but advocating a 'positive Christianity.' The Communist approach to religion was also complex in practice. As Marxists, Communists were atheists who saw religion as 'the opium of the people.' But only one Communist state – Albania, in 1967 – ever formally banned religion. Stalin was initially hostile to the Church. But after the Soviet Union was compelled by the Nazis to enter the Second World War (1941), he realized that he needed to maximize the nation's support for what the Communists called the 'Great Patriotic War,' and so became far more tolerant at least of the main Russian Church (Russian Orthodoxy). In return, the Church was to encourage patriotism and support for the Stalinist system and war effort.

Another area that the totalitarian state seeks to control and harness for its own use far more than does the authoritarian state is art and architecture. In Stalin's USSR, for instance, artists were to abide by the tenets of 'socialist realism,' while architects in Fascist Italy, Nazi Germany, and Stalin's Soviet Union were expected to design state buildings that were often on a grand scale and very solid, yet largely devoid of ornamentation. This approach was intended to symbolize a combination of strength, durability, functionality, optimism, and a commitment to the common people (see e.g., Affron and Antliff, 1997; Lahusen and Dobrenko, 1997; Krier, 2013).

Authoritarianism relates primarily to a political system, and there is no necessary implication that the state is heavily involved in the economy. In contrast, most theorists of totalitarianism maintain that, while Communist states generally controlled the economy much more than did the fascist states and introduced almost complete state ownership, the overall level of state interference in and interaction with the economy was much higher in all kinds of totalitarian states than in authoritarian ones.

The leading totalitarian systems were expansionist. Thus, the Nazis advocated more *Lebensraum* (living space) for Germans, and invaded Czechoslovakia and other European states in this context; the Italian Fascists made various irredentist claims, and invaded Abyssinia (now essentially Ethiopia), partly on the grounds that a superior nation needs to demonstrate its strength if it is not to fall into decline; while the Japanese fascists invaded Indochina and parts of China. Although Stalin theoretically rejected overt communist expansionism in the 1920s, when he promoted the concept of

'socialism in one country,' he took advantage of the Second World War and its aftermath to extend Soviet power not only into the Baltic States (Estonia, Latvia, and Lithuania) and Moldavia (now Moldova), which were incorporated into the USSR, but also into Central and Eastern Europe. It should be noted that most of the smaller countries that have also been described as totalitarian did not have such expansionist policies, however; examples include Laos, and Albania.

Finally, but importantly, totalitarianism is sometimes seen as a particular and distorted version of advanced modernity, and as a phenomenon that did not and could not emerge until the twentieth century – points that do not apply to authoritarianism. This observation about totalitarianism's relative newness is made mainly because it was not until the last century that the technological means for very high levels of control of populations existed. With the rapid development of the means of mass communication, notably the radio and cinema, citizens could be subjected to a constant barrage of state propaganda in a manner that was not possible in earlier kinds of dictatorship. This, together with the great emphasis placed by totalitarian regimes on state-run purposive education for all, meant that the state's capacity for indoctrination ('thought control,' pejoratively known as brainwashing) reached qualitatively new levels.

Although there have been several comparative analyses of totalitarianism, the most widely cited remains that of Carl Friedrich and Zbigniew Brzezinski (1956; amended version 1965). In the earlier edition, these authors identified what they considered to be the six salient features of totalitarian regimes. These were an official, all-embracing and chiliastic ideology; a single mass party, typically led by one man (the dictator); a system of terroristic police control (high levels of apparently arbitrary coercion); a near-complete monopoly of the means of communication; a near-complete monopoly of the means of armed combat; and central control of the entire economy. In the second edition of their book, this list was supplemented to include a commitment to expansionism and the administrative control of justice.

Another well-known analysis is that of Hannah Arendt (1951), a political philosopher and social theorist after whom a research institute on totalitarianism, founded in 1993 and based in Dresden, is named. Arendt argued that totalitarian systems often commanded widespread support in their early stages, since they offered security, comfort, and employment for most citizens in post-war and crisis contexts, as well as national pride following what was perceived to be their country's national humiliation (at least in the cases of Nazi Germany and, albeit to a lesser extent, Fascist Italy). But over time, totalitarian systems also lead to increasing atomization of the citizenry, to such an extent that even members of one's family may not be trusted. One of the saddest aspects of Stalinist Russia was that the authorities held up as a model citizen a teenage boy, Pavel Morozov, who allegedly reported his own father for opposing Stalin, and was then killed by his family – several members of which were then themselves executed by the state. While recent research suggests that much of the Pavel Morozov incident was a fabrication by the Soviet authorities, there is no question that the state used the story to encourage Soviet citizens to denounce each other; this is quintessential totalitarian atomization. For Arendt, one of the worst aspects of atomization was that it stifled all independent citizen initiatives.

It must be noted that some authoritarian regimes have displayed some of the features of totalitarianism outlined above. The use of state terror against alleged enemies of the state was a significant feature of General Pinochet's rule in Chile, for example. Moreover, some of the alleged differences are highly disputed, such as that totalitarian systems have charismatic leaders, which authoritarian systems do not. Nevertheless, if all the variables are combined, there really remain some important differences between authoritarian and totalitarian systems. This is not to deny that they are related, however; they clearly are, and some see totalitarianism as simply an extreme version of authoritarianism. In many ways, the approach adopted by Linz already in the 1960s, which is to see authoritarianism as a midway point along a continuum between democracy and totalitarianism, is the most useful way of conceptualizing these types of system.

The Development of the Concept

The concept of totalitarianism can be traced back to 1920s' Italy; from 1925, Mussolini himself used it in a positive way to describe the Italian Fascist system. But the term was not in common usage in English until after the Second World War. As Archie Brown (1999) points out, one of the first to alert the public to the differences between traditional authoritarian and the new totalitarian regimes was not a social scientist, but the novelist George Orwell. In *Animal Farm* (1945), Orwell focused on the hypocrisy and undemocratic nature of Communism, while in *1984* (published 1949), he highlighted the horrors of a society in which the notion of the ends justifying the means was taken to new extremes, and in which political elites could manipulate mass consciousness through high technology indoctrination. In this nightmare world, the goal of achieving a better society in the future was seen to override the concerns of ordinary individuals in the present.

Within the social sciences, it was primarily a number of scholars of German or East European origin who elaborated the concept and produced comparative analyses of its application in practice. While Arendt and Friedrich were of German origin, Brzezinski had Polish roots. These analysts, or members of their families, had often experienced the horrors of totalitarianism first hand. While fascism appeared to have been defeated by the end of the Second World War, Stalin's USSR became stronger, with much of Eastern Europe coming under direct Soviet influence for the first time. One of the objectives of many scholars of totalitarianism was thus a normative one. They sought not only to explain what it was and how it had arisen, but also to demonstrate that it was inherently evil, and why the West should resist it. In the climate of the early Cold War between liberal democracy and Communism, such normative analyses found considerable resonance and support, especially in the USA.

But the term began to lose some of its popularity by the 1960s, even in the United States (see Friedrich et al., 1969). Many now felt that it was inappropriate to use the same term to describe the USSR under both Stalin and Khrushchev, for

instance, especially since the latter was by the early 1960s referring to his predecessor's crimes against the Soviet population. The Hungarian leadership's liberalization in the 1960s, briefly alluded to above, was another reason why many began to question the continuing validity of the concept of totalitarianism to describe Communist systems. The term became even less fashionable in the early-1970s, when the relationship between the Soviet bloc and the West entered a more constructive phase, known as *détente* (lit. a reduction in tension). At about the same time, scholars such as Jerry Hough in the USA and Gordon Skilling in Canada began publishing articles and books, based on solid research, which suggested that there might be a limited form of pluralism emerging in the Soviet Union.

While the popularity of the concept was declining in the West, Jacques Rupnik (1988) maintains that, in the aftermath of the Warsaw Pact invasion of Czechoslovakia of 1968, many intellectuals in what is nowadays known as Central and Eastern Europe began to use the term totalitarianism in a different way from those that had earlier been favored by so many Western observers. These East Europeans tended to focus less on political institutions, and more on the political use of the language – not merely to indoctrinate citizens, but to limit the scope of discussion (discourse) and hence the parameters within which thought could roam. They also criticized the methods used by totalitarian – here, basically meaning Communist – regimes for suppressing knowledge of the past that did not suit those regimes. Totalitarianism was thus seen to be intent on destroying those historical memories, and hence the aspects of culture that did not suit the rulers.

The Usefulness of the Concept

As noted above, the concept of totalitarianism has been criticized for over-generalization about the regimes that some analysts consider too diverse to be grouped together under one heading. To Italy and Germany, for instance, could be added several other countries that were fascist or quasi-fascist at some point during the 1930s and 1940s, including Croatia, Hungary, Japan, and Romania. Among Communist states, there were substantial differences between, for instance, the *relatively* liberal and outward-looking Yugoslavia or Hungary and the far more orthodox and isolationist Albania or North Korea (it is worth noting that countries such as Croatia, Hungary, and Romania have been classified by analysts as both Communist and fascist totalitarian systems at different points in their histories). Certain aspects of the Friedrich and Brzezinski model, such as expansionism, did not apply to these smaller countries – although it must be acknowledged that Brzezinski distanced himself from the second edition of their book. The point also pertains in some cases to significant changes within one country over time; this lack of dynamism in Friedrich and Brzezinski's approach was strongly criticized by Karl Dietrich Bracher (1981). Moreover Slavoj Žižek has used the term totalitarianism so broadly, to include inter alia neoliberal democracies, that there is a danger of the term losing precision. Is the concept therefore too amorphous and too contentious to be useful?

There is considerable variety among systems usually labeled 'liberal democracies' – presidentialism, semi-presidentialism, and parliamentarism; federalism and unitarism; and so on – yet few would advocate abandoning the term liberal democracy; most accept that there is more in common between these systems than there are differences. Much the same can be said of totalitarian systems. Discarding Žižek's absurdly broad usage, which represents extreme conceptual stretching, the term totalitarianism is particularly useful if it is used as an ideal type, rather than as a model. The former is taken here to mean a pure form of an idea, against which actual regimes are measured. While the concept of a model has diverse meanings, it refers here to an attempt to encapsulate the salient features of a given system or type of system in a structured, schematic way. If totalitarianism is used as an ideal type, it is possible to say that a given country is more totalitarian or less totalitarian than another, or more totalitarian at one time than at another time. Using totalitarianism in this way overcomes the criticism that it is statically descriptive and lacks any internal dynamism.

Totalitarianism Today and Tomorrow

Most totalitarian systems have long since collapsed, for complex and diverse reasons that cannot be explored here. The major post-totalitarian systems soon adopted laws designed to ensure that totalitarianism could never emerge again in these countries. For example, it is illegal in Germany to display the swastika or deny the Holocaust; the 1947 Italian Constitution made it illegal to attempt to reorganize the Fascist Party; while Yeltsin banned the Communist Party in Russia in 1991 (although the Russian Constitutional Court later rescinded this decision and the Communists continue to perform reasonably well in Russian elections).

The concept of totalitarianism has experienced a limited resurgence of popularity since the collapse of Communist power in 1989–91, especially among Central and East Europeans. When the Communists were in power in that region, it was generally forbidden to suggest publicly that they were totalitarian dictatorships. Indeed, most of these countries officially claimed to be 'socialist democracies.' Since the early-1990s – and in some more liberal Communist states even before that period – many critics of the old regime have regularly and publicly referred to their countries' pasts (until the late-1980s) as totalitarian. Although this usage of the word is often loose and little more than a term of abuse, it has led some scholarly analysts to argue that the concept may have been abandoned prematurely during the later stages of the Cold War (see Siegel, 1998), and that, liberalization phases and certain changes notwithstanding, most Communist states remained fundamentally totalitarian until their collapse.

But does the term still enjoy currency elsewhere? Post-modernists such as Jean-François Lyotard were claiming in the late-twentieth century that the world is now moving on from the universalistic or grand theories – 'metanarratives' in his terminology – such as Marxism. Since 'grand theory' was the basis of Communist – and to a lesser extent fascist – ideology, which has been seen as such a distinguishing feature of

totalitarianism, it might be assumed that totalitarianism is a thing of the past. Although most commentators agree that there are no overtly fascist regimes at present (2013), some analysts (e.g., Christopher Hitchens) and politicians have refered to the rise of Islamic fundamentalism and various highly authoritarian regimes in predominantly Muslim states as Islamofascism. In 2005, US Secretary of State Condoleezza Rice – a Soviet and Russian specialist, thus well aware of the debates on totalitarianism – described Iran as a totalitarian state. Other Muslim states that have at various times been labeled totalitarian include Afghanistan (under the Taliban), Iraq (under Saddam Hussein), Libya (under Muammar al-Gaddafi), and Syria (under Bashar al-Assad). Moreover, there remain five Communist states (PR China, Cuba, DPR Korea, Laos, and Vietnam), and these are sometimes still described as totalitarian. With the exception of North Korea, however, there have been such significant changes in these Communist states in recent years that the label is misleading and inappropriate. In line with changes in the contemporary world, the Economist Intelligence Unit's *Democracy Index*, which was first published in 2007, does not even include totalitarianism as a regime type (its four categories are full democracy, flawed democracy, hybrid, and authoritarian).

Yet, we cannot be certain that totalitarianism is all but dead. It would be reassuring to feel confident that humanity learns from past mistakes; however, to be so would be naïve. Moreover, while it is true that globalization and the internet are rendering it ever more difficult at present for individual governments to control their populations in ways they sometimes did in the twentieth century, it is quite possible that governments will catch up with the technological revolution and harness it to new totalitarian objectives and styles. A number of states already censor the internet, and this practice could well increase in the future. Although many analysts point to the rise of social media such as Facebook and Twitter as developments that render state control of citizens more problematic, and cite the role of such media in rapidly inciting mass unrest in countries in which regimes have been either threatened or actually toppled (e.g., Myanmar and several Middle Eastern states during the so-called 'Arab Spring'), it would be unduly optimistic to assume that states will not attempt to control these too in the future. And, while the almost exponential increase in the use of CCTVs, in accessing individuals' bank accounts, and in listening to private telephone conversations may well be primarily a function of states' attempts to reduce crime rates and increase the public's security, the acclimatization of citizens to such developments also opens up possibilities for state elites so inclined to tend toward more totalitarian control.

Moreover, the rise of the extreme right (neo-fascism) in several European countries – including Austria, France, Germany, and the UK – since the 1990s means that it would be ingenuous to consign totalitarianism to the rubbish bin of history. Even state agencies in some countries have recently in some ways softened their approach toward totalitarianism. Thus, the Hungarian Constitutional Court revoked a ban on Communist and fascist symbols in February 2013. While this led to considerable public outcry, the Court defended its position in part by reference to November 2011 and April 2012 decisions by the European Court of Human Rights to permit the Red Star, on the grounds that it did not represent a symbol of tyranny, and that banning its use represented an unjust limitation on freedom of expression.

Given the points made in the previous two paragraphs, and while it may well be in a new form, since it would have to engage with a world that is very different from that of the early-to-mid-twentieth century, totalitarianism may yet rear its ugly head again. If it does, it is to be hoped that the collapse of both fascist and Communist systems in the past prove that the power of the state to control and mold the human spirit and suppress the desire for freedom can only ever be limited and temporary. Unfortunately, temporary does not necessarily mean short-lived.

See also: Attitudes, Political: Authoritarianism and Tolerance; Political Parties: Extreme Right.

Bibliography

Affron, M., Antliff, M. (Eds.), 1997. Fascist Visions. Princeton University Press, Princeton, NJ.
Arendt, H., 1951. The Origins of Totalitarianism. Harcourt Brace Jovanovich, New York.
Brown, A., 1999. The study of totalitarianism and authoritarianism. In: Hayward, J., Barry, B., Brown, A. (Eds.), The British Study of Politics in the Twentieth Century. Oxford UP, Oxford, pp. 345–394.
Bracher, K.-D., 1981. The disputed concept of totalitarianism. In: Menze, E. (Ed.), Totalitarianism Reconsidered. Kennikat Press, Port Washington, NY, pp. 11–33.
Bullock, A., 1998. Hitler and Stalin: Parallel Lives, second ed. Fontana.
Friedrich, C.J., Brzezinski, Z., 1956. Totalitarian Dictatorship and Autocracy, second ed. 1965. Harvard UP, Cambridge, MA.
Friedrich, C.J., Curtis, M., Barber, B., 1969. Totalitarianism in Perspective. Pall Mall, London.
Gregor, A., 2012. Totalitarianism and Political Religion: An Intellectual History. Stanford UP, Stanford, CA.
Kershaw, I., 2000. The Nazi Dictatorship: Problems and Perspectives of Interpretation, fourth ed. Bloomsbury, London.
Krier, L., 2013. Albert Speer: Architecture 1932–1942, second ed. Monacelli, New York.
Lahusen, T., Dobrenko, E. (Eds.), 1997. Socialist Realism without Shores. Duke University Press, Durham, NC.
Linz, J., 1975. Totalitarian and authoritarian regimes. In: Greenstein, F., Polsby, N. (Eds.), Handbook of Political Science, vol. 3. Addison-Wesley, Reading, MA.
Linz, J., 2000. Totalitarian and Authoritarian Regimes. Lynne Rienner, Boulder, CO.
Overy, R., 2005. The Dictators: Hitler's Germany and Stalin's Russia. Penguin, London.
Rupnik, J., 1988. Totalitarianism revisited. In: Keane, J. (Ed.), Civil Society and the State. Verso, London, pp. 263–290.
Schapiro, L., 1972. Totalitarianism. Macmillan, London.
Siegel, A. (Ed.), 1998. The Totalitarian Paradigm after the End of Communism. Rodopi, Amsterdam.
Žižek, S., 2011. Did Somebody Say Totalitarianism?: Five Interventions in the (Mis)Use of a Notion, second ed. Verso, London.

Totemism

Warren Shapiro, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Totemism' is a word widely used, not only by anthropologists but also in popular discourse. The arguments are made here that if the word points to anything in the real world, it is to a relationship between systems of social divisions and their analogous objectifications; the objectifications may or may not be animal forms; and when they are, a relationship of consubstantiality may or may not be posited between individuals and their objectification.

A book could be written about the myriad ways in which the term 'totemism' has been bandied about in the history of anthropology. The famous French scholar Claude Lévi-Strauss did just this in his *Totemism* (Lévi-Strauss, 1963), only to resurrect the term's initial significance as a religious association with animal forms. What Lévi-Strauss, without due acknowledgment to Alexander Goldenweiser (Shapiro, 1991), added was an analogical link between these forms and certain social divisions found widely in the tribal world, such that

> animal species A:social division 1::animal species B:social division 2::animal species C:social division 3 …

Current thinking on totemism has mostly accepted the Goldenweiser/Lévi-Strauss theory. But the main problem with this position is that it has become unmoored from the ethnographic bases from which the term originated, with the counterproductive result that this term has taken on a life of its own. It needs to return to its roots, which lie in native North and South American and especially aboriginal Australian ethnography – more particularly with respect to social divisions that anthropologists conventionally call 'clans.' This label too has become essentialized, pointing, allegedly, to units in which kinship links among members are assumed but not demonstrable.

Recent reexaminations of the data from all the three geographical areas noted, which are associated with clans, indicate that kin connections are neither demonstrable nor assumed; that the social divisions in question sometimes have relationships with their objectifications – which, as we shall see, are not accurately rendered as animal species – but are of an entirely different order. Thus Tooker (1971) tells us that in native North America "there is often evidence that the members of the clan consider themselves not linked by reason of the relationship through their mother or father, as the case may be, but by reason of their relationship to a clan bundle or fetish and its associated rights." As for South America, more particularly Amazonia, Murphy (1979) has noted "a kind of descent that incorporates within itself qualities of timelessness." This needs exemplification. In Northwest Amazonia there are local clans whose male members live permanently on landed estates supposedly created in a mythical period, and whose sisters move to their husbands' estates with marriage. This sounds like textbook 'patrilineal descent,' except for the salient consideration that the mythical period is not linked lineally to the present time. Rather, it is held to be recreated in all-male ritual, wherein the participants are said to be consubstantial with Creative Beings and whose purpose is to recreate the world (Goldman, 2004; C. Hugh-Jones, 1979; S. Hugh-Jones, 1979; Jackson, 1996). Thus, far from being a relationship of descent, the tie among clansmen might more accurately be described as a relationship of *anti-descent*, a denial of the decaying influence of time – in short, an assertion of immortality through consubstantiation with Creative Beings. Such an assertion, far from being arcane, is nothing less than the central tenet of Christianity. Northwest Amazonian belief differs primarily from Christianity, in that consubstantiality is segmented by clanship, not unified through a monotheistic doctrine.

An especially remarkable illustration of all this in Amazonia is provided by the Bororo of Central Brazil. Here the number of clans is fixed at eight; and if a clan is in danger of becoming extinct, its membership is replenished by changing the affiliation of a sufficient number of individuals of other clans – again, as if the influence of time must be denied. The same denial is apparent in the principle of clan affiliation, which is not through descent but through naming: an individual is a member of the clan that owns his/her name. This is normally the mother's clan, but the Bororo deny that clan members are all descended from a single ancestor or ancestress, i.e., that their relationship is carnal. As if to clinch the argument, an individual gets the names of certain kin, and for several days before the conferral of the name, these kin are expected to refrain from sexual relations (Crocker, 1985). Again, there are Christian parallels: Jesus is sometimes referred to as The Word (i.e., the name), his life said to be 'without sin' (i.e., sexless), and his quintessentially creative act was performed not through sex but through male violence.

This opposition of creation through sex and creation through violence is, in fact, widespread. This is particularly clear in ideologies of sacrifice, nearly always a male monopoly, in which the man's ability to 'give death' is very frequently contrasted with the woman's ability to give birth and the former is held to be superior (Jay, 1992). In many parts of native America, war captives were held to be the 'children' of their captors, only to be put to death by their 'fathers' (Conklin, 2001; Rosaldo and Atkinson, 1975; Santos-Granero, 2009). In the case of the Aztecs, such sacrifice was held to be necessary to renew the world itself (Read, 1998). There is more than a passing resemblance between all this and the New Testament narrative of the crucifixion.

Moreover, there is a link between violence and naming: both are assertive acts. Hence in the Old Testament, God creates the world by the latter method but in the New Testament he recreates it through the former. Hence we should not be surprised that both procedures are widely considered to be antithetical to, and superior to, sexual generation (Noble, 1992; Shapiro, 1995). Nor should we be surprised at the blatant sexism of religious ideologies around the world – as if men, by one or another kind of assertion, can provide enduring life, as contrasted with the ephemeral life provided by women.

The same elements can be seen in aboriginal Australian clan totemism. Here, as in Northwest Amazonia, there is no notion of descent in the strict sense but rather consubstantiality with mythical beings, who supposedly created the world. Here too is the all-male cult, in supposed imitation of the very acts of creation. And these acts usually employ one or another sort of assertion. Creation occurred, in native theory, when mythical creatures brought something into being by calling its name, or by their bodily movements, or by voluntarily divesting themselves of parts or all of their bodies that are said to have become features of the landscape. They sometimes have sex – there are female and hermaphroditic as well as male beings – but it is the assertive acts, not the erotic ones, that are held to be creative (Morton, 1987a,b; Munn, 1970). This is so much so that, when talking about human reproduction, aboriginal men often emphasize the dream encounter – usually by the father – with a Creative Being, who is then held to enter the mother and be reinstantiated in the child. Several generations of anthropologists have concluded – quite wrongly – from this sort of testimony that the native Australians are ignorant of physical paternity (Shapiro, 1996). One might just as well argue – with equal absurdity – that Roman Catholics or Orthodox Christians are similarly ignorant, for in these traditions too there are ideologies of rebirth in which a 'father' acts as intermediary between the world of the flesh and that of the spirit.

The degree to which theories of totemism were founded upon classic studies of aboriginal people in the Central Desert of Australia can hardly be overstated. Freud (1918), seizing upon (1) the fact that in some aboriginal Australian populations clans are *exogamous* i.e., a person may not marry his/her clansmate – upon the supposed descendant/ancestor relationship that the individual has with his/her totem; (2) that in some of these populations, a man may not kill a present-day animal representation of his totem; and (3) the view, widely held at the beginning of the twentieth century, that aboriginal Australians represent the 'earliest stage' in human history, argued that at this earliest stage, the central human concerns were the denial of the murder of the 'primal' father and the renunciation of the desire to mate with one's mother.

This Oedipal theory of the origins of human sociality does not farewell in the light of present knowledge. Although in some parts of aboriginal Australia clans are exogamous, this is not, as noted, the case everywhere on the continent; and in other parts of the world this may or may not be the case. In any case, with a *patrifilial* (father/child) principle of recruitment, plus exogamy, a man's mother will be a member of another clan. The argument could be made – and was made by scholars writing around the same time – that the earliest form of clanship was based on a *matrifilial* (mother/child) link, but there is no real evidence for this. And even if such evidence could be found, the question could still be raised as to why the sexual prohibition pertained to a number of women and not just to the mother. In any case, there is now unmistakable evidence that people, far from desiring sex with close kin, are instead neurologically inhibited from copulating with others raised from early childhood in the same locale (Wolf, 1993).

A further objection to Freud's theory has already been considered, viz., the relationship of consubstantiality that exists between an individual and his/her totem. This is especially clear in the most famous case, that of the Aranda (now rendered as Arrernte), where an individual is said to be a reincarnation of a Mythical Period being, with the father, as noted, only an intermediary in the act of spiritual transmission (Spencer and Gillen, 1899). One can see here, incidentally, how western observers were misled into thinking that aboriginal Australians are ignorant of physical paternity: the father's dream experience, which has to do with the generation of the spiritual person, logically ignores it. But we now know that the Arrernte also hold, as we do, that the physical person is a product of sexual relations (Austin-Broos, 2009).

A remarkable consideration here is that six male anthropologists have studied the Arrernte, and all six have been given an account of spiritual generation, a male monopoly. By contrast, two female anthropologists have interviewed Arrernte men, who proceeded to restrict their discourse to sexual creation, probably because ideas concerning spiritual creation are secreted from their own women.

As for not killing a current representative of the totem animal, this is a mixed bag. The prohibition has been reported for some parts of aboriginal Australia but not for others; and when killed, the totem animal may or may not be accorded special respect, such as an apology to its spirit. Still another variable has to do with which animals may or may not be killed. Thus in my own fieldwork, a 1000 miles to the north of the Arerrnte in Australia's monsoonal tropics, I found a serious ban on killing applying only to rock pythons, whose archetypical form has a special significance. Yet one man I talked with, apparently extending this sort of prohibition, suggested that men associated with other totems might be upset if one of the current representatives of any of the totems were killed.

Another classic attempt to find 'original' humanity in aboriginal Australian materials was made by Durkheim (1915). Durkheim argued that clan totemism is a projection onto the external world of social relations, of what he, and we, call 'society.' Note that this label, suggesting as it does an entity apart from human beings and their relationships, is itself a projection – what has been called a 'reification' (Berger and Luckmann, 1967). The idea in itself that religious notions stem from human affairs is reasonable enough, and it can lead to empirically testable propositions (Swanson, 1964). But Durkheim took as the elements of sociality in aboriginal Australia the clans themselves – a position true of a very great deal of theorizing both before and since he wrote, but which is utterly untrue to the facts. We now know that although these units are linked to territories, this linkage

is purely ideological, rather like the attention paid in Judaism, Christianity, and Islam to Jerusalem and its environs; the effective local groups on the continent are networks of bilateral kin, with special emphasis on the tie between a man and his in-laws (Hiatt, 1965; Peterson, 1970; Shapiro, 1973). The clans, in short, are not localized, despite their iconic link to locality.

Attention should be paid to the link between aboriginal Australian clanship and male bloodletting rites – more generally, to a man's progress in spiritual life. The most startling of these rites to a western observer is what has been called *subincision*, practiced by the Arrernte and many of their neighbors. Here a youth is secreted from the main camp and the eyes of women and children, and his penis is split lengthwise, so as to expose the urethra. Although the significance of subincision has been debated by anthropologists, it is clear that one intention is to endow the boy with a pseudovagina. He is thus a hermaphrodite of sorts, capable of a kind of reproduction that is independent of women. More specifically, he is now able to participate with men already subincised, in rites designed to increase the supply of animals associated with the pertinent clan, and in ceremonies in which younger boys are reborn as men. The reproductive symbolism of both kinds of rituals is plain as day. In some rites of the former sort, a subincised man or several men enter a large crevice which has been carved into the ground and which is likened to the female reproductive area. The man or men next re-open their subincisions, deliberately shedding blood, which is often regarded as the female contribution to the fetus. Then, having adorned their heads with white clay and having allowed it to congeal, they cause their bodies to quiver, one result of which is to spread particles of dried white clay into the crevice (Róheim, 1945, 1971; Spencer and Gillen, 1899; Strehlow, 1947). One hardly has to be a doctrinaire Freudian to argue that the crevice is (seen as) a vagina, the quivering body as a penis or penes in the course of copulation, and the particles of white clay as seminal emissions. Similarly, Hiatt (1971) has called attention to the widespread occurrence of what he calls 'uterine rites' in transforming aboriginal Australian boys into men. In these, men secreted from women, form objects which are likened to the female reproductive area. In the most remarkable instance, men whose penes have already been cut, form a line and spread their legs as widely as possible. The boy then moves through the line, to emerge from it as a man, newly born from a collective male vagina.

Although, for obvious reasons, it is subincision that has gained the most attention from anthropologists, there are comparable hermaphroditic rites elsewhere in aboriginal Australia – and, indeed, in other parts of the world. Further north on the continent, boys are first circumcised and the blood that flows from the womb is expressly likened to vaginal blood, which is as noted, the supposed female contribution to the fetus. Then, a few years later, they are scarified in the upper arm to create a pseudovagina, and this wound, like the Arrernte subincision, is periodically reopened, not with a view to increasing the supply of animals as this is a much better-watered area, but to enable the process of transformation – the rebirth – from boyhood to manhood. East of New Guinea much the same ideology is effected when men ritually gash their penes with the claw of a crab (Hogbin, 1970). And it appears that scalping among the Plains Indians and the blood of warfare in parts of Africa are associated with a similar set of notions, though here the blood is taken not from oneself but from another (de Heusch, 1982). What is remarkable about this last set of cases is the coupling of ritual hermaphrodism with violence, as noted above, with both being considered higher forms of creation than sexual relations.

That what is involved here is a distinction between physical birth and spiritual rebirth is made even clearer by another procedure which occurs both among the Arrernte and further north, in connection with all-male ritual. After a boy is introduced to ritual life by circumcision and before he is otherwise cut (via subincision or gashing of the upper arm), he is shown his own clan totems, objectified in rock or wooden form, and he is told that *these* constitute his body, what he *really* is, with the implication that his physical body is of decidedly secondary significance. And this real body is generated by the sacrificial and naming acts of the Creative Period, not, as the secondary body is, by sexual relations, which, it is implied, have no sustained creative power whatsoever (Fry, 1933). Hence, again, it is unsurprising that anthropologists witnessing all-male rituals among aboriginal Australians have reported that these people are unaware of a connection between sex and generation – this despite the blatant erotic symbolism of the male pseudo-reproductive cult.

Although we do nothing quite so astonishing as subincision, there are clear parallels in western religions. Jewish males are circumcised on the eighth day after their emergence from their mothers to signify their identification with the biblical quasi-sacrifice of Isaac and perforce with all-male ritual. Christianity originally opened its ceremonial doors to women as well, as do the more liberalized forms of Judaism. But baptism was derived from the Jewish ritual bath, designed to remove the supposed decaying effects of sex from men and of menstruation and childbirth from women, i.e., to recreate an individual in unsullied form. And some modern versions of that faith – particularly in the Roman Catholic and Orthodox churches – have explicitly designated rebirth rituals in which the baptismal basin is likened to a womb and holy water to the blood of parturition, while the connection of the child to a new set of parents linked to him not through sex but through sacrifice is made by a 'father' who acts independently of women (Chock, 1974; Gudeman, 1972).

There are even approximations to ritual gender egalitarianism in aboriginal Australia. Thus Bell (1983) has made much of the fact that among the Walbiri of the Central Desert, not too-distant neighbors of the Arrernte, there are all-female secret cults; but she also tells us that these are held to be less sacred than their male equivalents, and that in order to engage in them, women need male permission, though the reverse is not the case. What is most commonly celebrated in female ritual in aboriginal Australia and elsewhere, is a girl's coming of reproductive age, but male ritual is a horse of a different color. The key enactments in the latter case are a boy's separation from his mother and his coming under the control of a male collectivity. Hence the collective nature of male pseudo-reproductive ritual. Hence too, the emphasis on

a young male's participation in the secret rituals of several clans – first his own (usually his father's as well), then those of other close kin, then those of less closely related men: indeed, a man's reputation, and this is by no means confined to aboriginal Australia, is gauged largely by the number and reputation of other men with whom he has some sort of relationship, including relationships of the 'blood brother' sort. Aboriginal Australian women, by contrast, have a much more limited sphere of sociality: most of their time is spent with their mothers, sisters, and other close kin, and when they begin to menstruate the celebration is comparably confined (Munn, 1969). This contrast seems to be widespread: indeed, Carnes (1989) has suggested a parallel in human thought between a male's progress in 'primitive' ritual life and notions of spiritual advancement in the so-called 'higher' religions.

It has been noted than an aboriginal Australian individual usually is a member of his/her father's clan. This 'patrilineal' principle is very often supplemented by another one, viz. that an individual is a member of the clan on whose territory the father had the dream experience already noted. Hence Maddock (1974) writes of 'conception clans,' i.e., units composed of individuals linked to the same territory through this sort of experience. But this is misleading, for two reasons. First, it is becoming increasingly clear that the dream experience occurs not at conception but at quickening, i.e., when the fetus shows obvious signs of movement (Scheffler, 1978). Second, as noted, affiliation via the experience is probably always supplementary to, rather than competitive with, affiliation directly through the father's membership. In fact, Myers (1986), writing of the Pintupi of Australia's Western Desert, lists no fewer than 36 principles, each apparently complementary with the others, by which individuals can become affiliated to clans. These include (1) conception (quickening?) at the associated locale; (2) conception (quickening?) at another locale created by the same Creative Being who created the first locale; (3) conception (quickening?) at another locale mythically traversed, though not created, by the Being who created the first locale; (4) subincision at the associated locale (for a male); (5) birth at the associated locale; (6) conception (quickening?) of one's father at the associated locale; (8–11) conditions two to five true for one's father; (12) conception (quickening?) of mother at the associated locale; (13–16) conditions two, three, or five being true for one's mother; (17–20) conception (quickening?) of one or another grandparent at the associated locale; (21–34) conditions two, three, five or (in the case of grandfathers) four being true; (35) residence in or near the associated locale; and (36) death of a close relative in or near the associated locale. The Pintupi case more than any other study shows that neat and established labels like 'patrilineal' and 'matrilineal' cannot be applied to totemic groups.

Then there is the supposedly invariant animal connection. Even in the classic Arrernte case, the animals are represented as beings who are partly human, and in the course of myths these beings change shape back and forth and sometimes have no animal aspect (Morton, 1987b; Spencer and Gillen, 1899). The Rock Python, as noted, is an important totem in the monsoonal north, but so are two human sisters with whom it is mythically associated (Berndt, 1951). Aboriginal people in this area had long contact with Indonesian fishermen until about a century ago. So it is not too surprising that the mast of an Indonesian ship is a totem of some of the clans (Warner, 1937; Worsely, 1955). Even when the totems have animal form, they do not figure, as they do among the Arrernte, in increase rites. What *is* remarkable is that, regardless of the form of the totem, its historicity is never admitted; rather, it is presumed always to have existed. There is good evidence that over time clans sometimes die out or split into two or more such units (Keen, 1995); and, as just noted, they adopt new totems, some but not all of which may be shared with other clans; but such historical change is not acknowledged. This primitivism, or, as one scholar has put it, 'the will to devalue time' (Eliade, 1959), is a regular feature of religious thought. Thus although we know that Jesus was an historical figure who emerged in fleshly form in what was then Roman-occupied Judaea in the first century, his advent was supposedly prefigured in Old Testament prophecy, and indeed the very first verse of the Book of John places him "In the Beginning" of things. And supposedly secular Marxist theory posits a Return to an Age of Primitive Communism.

A good example of (something like) totemism is provided by North American baseball teams. Each is associated with a city or state, and each has a nickname. Sometimes this nickname is the name of an animal, as with the Detroit Tigers and the Arizona Diamondbacks. But then there are the Los Angeles Dodgers. How on earth did this happen? Here are the facts. The Los Angeles Dodgers used to be the Brooklyn Dodgers, who did not move west until 1958. Brooklyn, however, is neither a city nor a state but one of the five boroughs that make up New York City. So the 'totemic' plan is doubly challenged. We can account for one challenge by the consideration that, until 1897, when professional baseball was already nearly three decades old, Brooklyn was a city of its own; but the other challenge – the lack of an animal name – remains. Aha! The Dodgers were once called the Robins. The name was changed because the quickest way to get to the ballpark in which they played their home games, was to cross a certain thoroughfare. This thoroughfare was circumlocuted by trolley cars, which baseball fans had to dodge to get to the game. Hence the name Trolley-dodgers, soon shortened to Dodgers. So, it might be supposed, the original plan of the National Baseball League involved only animal names. But this is not true either. The first team to join the League was called the Cincinnati Red Stockings; but the Red Stockings became the Reds (probably when the word 'stockings' became applicable exclusively to women's hosiery). The Reds became the Redlegs during the McCarthy witch hunt period, when no good American could be a Red. However, the team is now once more called the Cincinnati Reds.

There is a system here, but unlike the Bororo and aboriginal Australian clans, no attempt is made to disguise its historicity, which can readily be gleaned from the rapidly increasing library on the history of baseball. There is no notion of lineal descent, but this is also true of the clans. A fan of the Cincinnati Reds may marry another such fan but he may also marry a Los Angeles Dodger fan (though this might cause some marital strife), just as an Arrernte may marry someone of his own totem

or someone of another totem. Some animal association exists in both sets of cases, but it is limited, and, more importantly, subordinated to a more general notion which may be called *objectification*. So Lévi-Strauss needs to be revised as follows:

> objectification A:social division 1::objectification B:social division 2::objectification C:social division 3 …

Totemism is not a distinctive religion, as Freud, Durkheim, and many others (including a remarkable number of Internet 'totemites' in search of their Inner Wolf or Bear or Lion) have thought. Similarities with western religions have been noted in the course of this article. It should be added here that the Cree Indians of the Great Lakes area in Canada, like the Arrernte, have rituals to increase the supply of animals, but, unlike the Arrernte, have no clans whatsoever (Brightman, 1993; Tanner, 1979).

In researching this article I have been struck by the number of other articles – in books and in electronic form – beginning glibly with "Totemism is …." It seems as if, because there is a word out there, it must pertain to something real; and what we regard as real is just the instantiation of definitions. This, again, is reification, and it is the stuff of which religions and introductory anthropology texts are made, but it is not based on empirical inquiry. A consideration of pertinent ethnographic materials suggests, at best, only the limited set of analogies just noted. It is in such a limited sense that I have used the term 'totemism' in the course of this article, and it is the only sense which, by my lights, does not violate the ethnographic data.

Bibliography

Austin-Broos, D., 2009. Arrernte Present, Arrernte Past: Invasion, Violence, and Imagination in Indigenous Central Australia. The University of Chicago Press, Chicago.

Bell, D., 1983. Daughters of the Dreaming. McPhee Gribble, Melbourne.

Berger, P., Luckmann, T., 1967. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Penguin, New York.

Berndt, R.M., 1951. Kunapipi: A Study of an Australian Aboriginal Religious Cult. F W Cheshire, Melbourne.

Brightman, R., 1993. Grateful Prey: Rock Cree Human-Animal Relationships. University of California Press, Berkeley, CA.

Carnes, M.C., 1989. Secret Ritual and Manhood in Victorian America. Yale University Press, New Haven, CT.

Chock, P.P., 1974. Time, nature, and spirit: a symbolic analysis of Greek-American spiritual kinship. American Ethnologist 1, 33–47.

Conklin, B.A., 2001. Consuming Grief: Compassionate Cannibalism in an Amazonian Society. University of Texas Press, Austin, TX.

Crocker, J.C., 1985. Vital Souls: Bororo Cosmology, Natural Symbolism, and Shamanism. The University of Arizona Press, Tucson, AZ.

de Heusch, L., 1982. The Drunken King, or, The Origin of the State. Indiana University Press, Bloomington, IN.

Durkheim, E., 1915. The Elementary Forms of the Religious Life. Allen & Unwin, London.

Eliade, M., 1959. Cosmos and History: The Myth of the Eternal Return. Harper & Brothers, New York.

Freud, S., 1918. Totem and Taboo: Resemblances between the Psychic Lives of Savages and Neurotics. Moffat Yard & Co., New York.

Fry, H.K., 1933. Body and soul: a study from western Central Australia. Oceania 3, 247–256.

Goldman, I., 2004. Cubeo Hehénewa Religious Thought: Metaphysics of a Northwestern Amazonian People. Columbia University Press, New York.

Gudeman, S., 1972. The compadrazgo as a reflection of the natural and spiritual person. Proceedings of the Royal Anthropological Institute of Great Britain and Ireland for 1971, 45–71.

Hiatt, L.R., 1965. Kinship and Conflict: A Study of an Aboriginal Community in Northern Arnhem Land. The Australian National University Press, Canberra.

Hiatt, L.R., 1971. Secret pseudo-procreation rite among the Australian Aborigines. In: Hiatt, L.R., Jayawardena, C. (Eds.), Anthropology in Oceania: Essays Presented to Ian Hogbin. Chandler Publishing Company, San Francisco.

Hogbin, I., 1970. The Island of Menstruating Men: Religion in Wogeo. Chandler Publishing Company, London, New Guinea.

Hugh-Jones, C., 1979. From the Milk River: Spatial and Temporal Processes in Northwest Amazonia. Cambridge University Press, Cambridge.

Hugh-Jones, S., 1979. The Palm and the Pleiades: Initiation and Cosmology in Northwest Amazonia. Cambridge University Press, Cambridge.

Jackson, J.E., 1996. Coping with the dilemmas of affinity and female sexuality: male rebirth in the central Northwest Amazon. In: Shapiro, W., Linke, U. (Eds.), Denying Biology: Essays on Gender and Pseudo-Procreation. University Press of America, Lanham, MD, pp. 89–127.

Jay, N., 1992. Throughout Your Generations Forever: Sacrifice, Religion, and Paternity. The University of Chicago Press, Chicago.

Keen, I., 1995. Metaphor and metalanguage: "groups" in northeast Arnhem Land. American Ethnologist 22, 502–527.

Lévi-Strauss, C., 1963. Totemism. Beacon Press, Boston.

Maddock, K., 1974. The Australian Aborigines: A Portrait of Their Society. Penguin, Harmondsworth, UK.

Morton, J., 1987a. The effectiveness of totemism: 'increase ritual' and resource control in Central Australia. Man 22, 453–474.

Morton, J., 1987b. Singing subjects and sacred objects: more on Munn's 'transformation of subjects into objects' in Central Australian myth. Oceania 58, 100–118.

Munn, N.D., 1969. The effectiveness of symbols in Murngin rite and myth. In: Spencer, R.F. (Ed.), Forms of Symbolic Action. University of Washington Press, Seattle, WA, pp. 178–207.

Munn, N.D., 1970. The transformation of subjects into objects in Walbiri and Pitjantjatjara myth. In: Berndt, R.M. (Ed.), Australian Aboriginal Anthropology: Modern Studies in the Social Anthropology of the Australian Aborigines. The University of Western Australia, Nedlands, Australia, pp. 141–163.

Murphy, R.F., 1979. Lineage and lineality in Lowland South America. In: Margolis, M.L., Carter, W.E. (Eds.), Brazil: Anthropological Perspectives; Essays in Honor of Charles Wagley. Columbia University Press, New York, pp. 217–224.

Myers, F.R., 1986. Pintupi Country, Pintupi Self: Sentiment, Place, and Politics among Western Desert Aborigines. Smithsonian Institution Press, Washington.

Noble, D.F., 1992. A World without Women: The Christian Clerical Culture of Western Science. Oxford University Press, New York.

Peterson, N., 1970. The importance of women in determining the composition of residential groups in Aboriginal Australia. In: Gale, F. (Ed.), Women's Role in Aboriginal Society. Australian Institute of Aboriginal Studies, Canberra, pp. 9–16.

Read, K.A., 1998. Time and Sacrifice in the Aztec Cosmos. Indiana University Press, Bloomington, IN.

Róheim, G., 1945. The Eternal Ones of the Dream. International Universities Press, New York.

Róheim, G., 1971. Australian Totemism: A Psycho-Analytic Study in Anthropology. Humanities Press, New York.

Rosaldo, M.Z., Atkinson, J.M., 1975. Man the hunter and woman: metaphors for the sexes in Ilongot magical spells. In: Willis, R. (Ed.), The Interpretation of Symbolism. J M Dent & Sons, London, pp. 43–75.

Santos-Granero, F., 2009. Vital Enemies: Slavery, Predation, and the Amerindian Political Economy of Life. University of Texas Press, Austin, TX.

Scheffler, H.W., 1978. Australian Kin Classification. University of Cambridge Press, Cambridge.

Shapiro, W., 1973. Residential grouping in northeast Arnhem Land. Man 8, 365–383.

Shapiro, W., 1991. Claude Lévi-Strauss meets Alexander Goldenweiser: Boasian anthropology and the study of totemism. American Anthropologist 93, 599–610.

Shapiro, W., 1995. Sex, violence, and 'cultural constructionism'. Journal of the Royal Anthropological Institute of Great Britain and Ireland 1, 625–626.

Shapiro, W., 1996. The quest for purity in anthropological inquiry. In: Shapiro, W., Linke, U. (Eds.), Denying Biology: Essays on Gender and Pseudo-Procreation. University Press of America, Lanham, MD, pp. 167–189.

Spencer, B., Gillen, F.J., 1899. The Native Tribes of Central Australia. MacMillan, London.

Strehlow, T.G.H., 1947. Aranda Traditions. Melbourne University Press, Melbourne.

Swanson, G.E., 1964. The Birth of the Gods: The Origin of Primitive Beliefs. The University of Michigan Press, Ann Arbor, MI.

Tanner, A., 1979. Bringing Home Animals: Religious Ideology and Mode of Production of the Mistassini Cree Hunters. St. Martin's Press, New York.
Tooker, E., 1971. Clans and moieties in North America. Current Anthropology 12, 357–376.
Warner, W.L., 1937. A Black Civilization: A Social Study of an Australian Tribe. Harper, New York.
Wolf, A.P., 1993. Westermarck redivivus. Annual Review of Anthropology 22, 157–175.
Worsely, P.M., 1955. Totemism in a changing society. American Anthropologist 57, 851–861.

Touraine, Alain (1925–)

Howard Davis, School of Social Sciences, Bangor University, Gwynedd, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alain Touraine's distinctive contributions to sociology include his sociological theory of action, the method of 'sociological intervention' in the study of social movements, and his critique of modernity which offers a 'post-social' approach to contemporary societies. This article explains the key stages of Touraine's work in the context of post-war social theory in France. It considers some key criticisms of his theoretical and empirical work as well as his influence as a sociologist and public intellectual.

Introduction

Alain Touraine, the French sociologist, is most renowned for three major contributions to the discipline: his sociological theory of action; the method of 'sociological intervention' in the study of social movements; and his critique of modernity which offers a 'post-social' approach to contemporary societies. The unifying theme is the sociology of the subject and social actors, which he understands in terms of the forces which are tearing apart modern experience and the actions which can help to rebuild it. Like other influential sociologists of the post-1945 generation in France, Touraine pursued his career through the elite Ecole Normale Supérieur (ENS) and the institutions of the French higher education and research system centered in Paris. In 1958 he created the Laboratoire de Sociologie Industrielle, which in l970 became the Centre d'Etude des Mouvements Sociaux of the Ecole Pratique des Hautes Etudes. He was founder in l981, and director until l993, of the Centre d'Analyse et d'Intervention Sociologiques (CADIS) of the Ecole des Hautes Etudes en Sciences Sociales (EHESS). From l960 onwards he was professor and director of studies at the EHESS, and from 1966 to 1969 was professor at the University of Paris X Nanterre.

Biography

Touraine was born in 1925 in Hermanville-sur-Mer in the department of Calvados in north-west France. He has described his family background as middle class ('bourgeois' because his father was a doctor) and his life as orderly, focused on study, and insulated from the social and political turbulence of the interwar years (Touraine, 1977: 14–16). He was educated in Paris and was strongly supported to succeed in his studies. In 1940 he obtained his baccalaureate from the Lycée Louis-le-Grand. The 'great debacle' of the German invasion and the occupation saw the first sociological stirrings in Touraine's thinking. "The society was turned upside down, its principles crumbled to dust; France, character of almost divine importance, was nothing but occupied territory" (1977: 27). But the war years were spent in more studies, the *khâgne*, in preparation for entry to the ENS. He joined the ENS as a student of history in 1945, qualifying as *agrégé* in history in 1950. The narrowness of this intellectual environment, where at this time sociology was not established as a university discipline and where there was no empirical research tradition, encouraged Touraine to interrupt his studies in 1947 to travel to Hungary and then to Yugoslavia, which was already under Soviet occupation. On his return, he spent several months working in the mines in the north of France. This first-hand experience of the living and working conditions of the world of coal mining provoked a series of sociological questions about labor, class, industrial organization, and social change which were to occupy Touraine for many years. His first sociological mentor was Georges Friedmann, the industrial sociologist, who appointed him as research assistant to study work and technological change in the Renault car factory at Billancourt. This established him as a researcher at the Centre National pour la Recherche Scientifique (CNRS), where he worked until 1958. In another important interlude, he was Rockefeller Fellow at Harvard, Columbia, and Chicago universities between 1952 and 1953. These travels exposed him to the US intellectual elite as well as to the inequalities, conflicts, and experience of social transformation among both white and black populations in Chicago. His encounters with the leading figures of American sociology including Talcott Parsons confirmed his determination to resist normative notions of 'society' and any assumption that the core of sociology is social integration, adaption, or equilibrium. An invitation to the University of Chile in 1956 was to have far-reaching consequences. It led to his marriage to Adriana Arénas (d.1990) as well as to an ongoing research interest in Latin America. As the concepts of action and social movement became central to Touraine's sociology in the late 1950s and 1960s, he was one of the most recognized figures in French sociology, which had rapidly grown into a popular field. An established scholar and public intellectual, both nationally and internationally, Touraine is a man of the left and an outspoken critic of neoliberalism. Key intellectual influences were Marxist theory (though he was never a communist, as many of his intellectual contemporaries were) and existentialism (where he draws on the antiessentialism of Sartre and his concepts of subjectivity). His thinking has some affinities with the antistructuralist position of the political philosopher Cornelius Castoriadis (1975). Touraine has continued to develop his own distinct contributions to theory and empirical analysis in a continuous, integrated fashion. These are now presented in summary form.

Touraine's Sociology

There are three recognizable phases of development in Touraine's research and publications and a fourth is still in the making. The following sections describe the main publications from each period and their significance in the development of his sociology.

1. The first phase was devoted to *sociologie du travail* or the sociology of labor (in contrast to what was usually called industrial sociology in English, which was defined more by its object of study rather than a theoretical purpose). The interest in labor reflects Touraine's direct encounters with the mining and auto manufacturing industries and the research program of the Laboratoire de Sociologie Industrielle. It is also the foundation for his early contributions to theory construction. Starting with detailed observation of the uses of automation, his research focused on the relationships, often conflictual, between the world of work, the world of organization, collective consciousness, and class movements. Broadly, the position was critical of scientific management and Taylorism but positive toward the opportunities afforded by automated processes. In *La conscience ouvrière* (1965a), Touraine characterizes the issues as a collective struggle for the control of the 'society' of industrial production (Touraine, 1977: 55). This society is not a determinate system of relations of economic production, but rather the process of genesis of values in an advanced technological milieu. In contrast to the system of craft production and its successor, mass manufacturing, the automated system represents a mutually dependent system with new forms of meaning and action, with the opportunity for a new kind of workers' consciousness. While the study was criticized for its technical flaws (see Rose, 1996) it represented an important departure from orthodox Marxist analyses of class consciousness as well as managerialist thinking. Also in this period Touraine published his first works carried out in Latin America, on the Chilean mineworkers and foundry workers.

The theoretical aspects of the argument concerning workers' consciousness were elaborated in Touraine's doctoral thesis published as *Sociologie de l'action* (The Sociology of Action) (1965b). The book builds on the historical understanding of industrial society as a system which exists as 'the child of its own creative activities,' in which labor is the paradigmatic concept (not action in the sense used by Parsons). It goes beyond interpretivism as well as functionalism and structuralism. The meaning of action is not conceptualized as the property of an individual human subject but rather as the activity of the historical subject, which is expressed in the capacities that a society has to act upon itself. It is an alternative way to describe 'society' or 'mode of production' without resorting to concepts of system, social regime, or even nation. The action approach (*actionnalisme*) is the mode of sociological analysis which aims to reconstruct the normative orientation of the action of a historical subject without reifying individuals or institutions. Workers' movements occupy the most important part of Touraine's discussion at this stage but other examples include civil disobedience campaigns and movements among ethnic minorities or religious groups. Touraine acknowledges the ambiguity between the subject defined in terms of an historical notion of labor (human–nature relations, technical divisions of labor, relations of production, and so on) and the philosophical notion of the normative orientation of conduct of the subject and its contribution to planning, programming, and controlling. However, some of these ideas become clearer when applied to the specific issue of the emerging forms of social conflict discussed in *La société post-industrielle* (1969) (The Post-Industrial Society). Despite the title, Touraine's preferred term is 'the programmed society,' which better conveys the way in which the society acts upon itself through advanced technical systems, education, and a cultural model of constant change. These are the very points at which an 'explosion of contestation' occurred in the 1960s (Touraine, 1969: 232). For Touraine, these vigorous but only partly coherent new movements (not yet capable of grasping the totality of the cultural model of the new society) represent the reconfiguration of the historical subject, 'tomorrow's social history.'

The first phase of Touraine's work in which he aimed to redefine the core concepts and tasks of sociology culminated in the book *Production de la société* (1973a) (The Self-Production of Society). Here he continues to strongly insist that sociology must reject all assumptions that try to represent society as an actor: "the social order has no metasocial warrant for its existence" (1973a: 8). He presents the alternative in terms of three central concepts:

> "*historicity*, which defines the instruments of society's self-production; the *system of historical action*, meaning that totality composed of social and cultural orientations, by means of which historicity exercises its dominion over society's functioning; and *class relations*, struggles for the control of historicity and of the system of historical action" (1973a: 13).

The first part explains and defines each of these concepts in turn. The final chapter discusses the role of sociology, while the other chapters offer detailed theoretical analysis in political sociology, social organization, and social change which prepare the ground for Touraine's later sociological practice, especially in the study of social movements. Society differs from an organism which seeks to ensure its stability by homeostasis. Its capability is not limited to adapting itself to its environment but is society's capacity to act upon itself and to modify the rules of its own functioning and self-representation. This is 'historicity.' It possesses a symbolic capacity that enables it to construct a system of *knowledge* together with technical tools which it can use to intervene in its own functioning.

The components of historicity are the model of knowledge (image of society and nature); accumulation (the economic formation of a surplus for work on work); and the cultural model (which constitutes societal recognition of the gap between historicity and functioning and includes the image of creativity). When this insistence on historicity is linked with concepts like industrial or postindustrial society, it verges on historicism but the theoretical foundations were set for the next stage of Touraine's work which had already begun with

his theoretical and practical engagement with the manifestations of social change both in France and internationally.

2. At the time of the May 1968 events in France, Touraine was professor and head of department of sociology at Nanterre where, from a leftist position, he mediated between the educational authorities, his fellow faculty, and student factions, making few friends along the way. These experiences were a laboratory for Touraine's sociology of action and he took the cultural revolt to be a new kind of struggle to take over control of social change in the programmed society, as he explained in *Le Mouvement de Mai ou le Communisme Utopique* 1968 (The May Movement: Revolt and Reform). Equally important for Touraine's thinking during the years which followed were the military coups in Latin America, including the 1973 Chilean coup which installed Pinochet as dictator (*Vie et mort du Chili populaire* 1973b). With colleagues in the Centre d'Etude des Mouvements Sociaux he developed a distinctive research program for the study of social movements, especially those he considered to be the carriers of historicity in the new type of society. With the relative decline of labor movements, the idea of 'new' social movements was becoming widely used to describe key conflicts of postindustrial society which appeared to transcend old class divisions, such as the student movement, feminism, antinuclear, and environmental movements. If the movement of labor was the central social conflict of industrial societies, it was thought that there should be a comparable central axis of conflict in the system of historical action of postindustrial societies: the struggle against political and social, not just economic, domination (*Pour la sociologie* 1974: 187–188).

A short definition of Touraine's concept of a social movement is that it is "the organized collective behaviour of an actor struggling against an adversary for the social direction of historicity in the concrete collectivity" (*La voix et le regard*: p.103). A movement is defined by the combination of three principles: identity, opposition, and totality. Without all these elements it will fall short of influencing a new cultural model of society. Actors have to define themselves as such, so identity (I) refers to the self-consciousness of social actors, their awareness of themselves as subject to domination, and capable of acting. Opposition (O) refers to conflicts between opposing sets of interests, the definition of the adversary. Totality (T) is the principle which combines these elements. It defines a common sphere of action in which the stakes of social change are recognized by both the movement and its adversary (in the case of industrial society, the model of industrialization). This IOT analysis is now widely identified with Touraine's contribution to social movement theory and its value is that it can be used to analyze the relationship between the elements of a movement (I–O, O–T, I–T) and thus explain its trajectory, impact, or self-limiting features.

While these concepts demarcate a particular approach to the study of social movements, Touraine's most distinctive contribution in this field is methodological: the development of the method of 'sociological intervention,' explained in *La voix et le regard* 1978 (The Voice and the Eye 1981) and demonstrated through a series of empirical studies including the antinuclear movement (Touraine et al., 1980) and the Occitanist movement (1981). Intervention is consistent with the theoretical definition of social movement as the struggle to reconstitute the social, not simply exert pressure on institutions or express an identity. It consists in the sociologist entering into the activities of a social movement to create controlled and experimental situations to bring to the surface the relationships, the conflicts, or the agreements of which it is made up. The role of the researchers – who have roles both as 'interpreters' and 'analysts' – is to interact with activists in meetings constituted rather like focus groups (avoiding the movement leaders who will tend to speak for the movement). The method promotes a self-analysis of the movement, to reinterpret its past, to critique its assumptions and ultimately if 'conversion' occurs, to reorient its action.

In 1981, Touraine spent a year in Poland with a French and Polish research team studying Solidarity, which was the most significant opposition movement in Eastern Europe in the late Soviet period. It was built on a trade union and working class base but in the context of the Polish party-state it developed national and democratic ambitions for the whole society as well as rights for workers. In Touraine's terms this was a 'total' social movement. At the same time, its character was defined by internal differences between two strands: the movement among workers to defend their rights and the movement for liberation of society (involving intellectuals as well as workers). In a review of this work, Martell and Stammers (in Clark and Diani, 1996: 127 ff.) commend the results of the intervention method for its detailed understanding of the self-definitions of actors in Solidarity. However, they point out that the research pays less attention to the organizational and political dynamics within the movement, or to the more shadowy aspects of populist nationalism and neoliberal thinking contained within Solidarity. As a counter to the criticism, it is clear that Touraine's studies in Latin America (e.g., on Chile in *La Parole et le Sang* 1988) and his later writings on French politics are rich in detail on such questions.

Touraine's numerous studies of movements demonstrated a problem with applying the assumption of a central conflict associated with a given system of historical action or type of society (e.g., an antitechnocratic movement to transform the programmed society). Actual social movements are an unstable combination of elements in concrete national and local settings. "Like all the elements of a modern society, social movements are both more autonomous and more fragile than in societies where overall principles of – political, social and cultural – integration existed" (Touraine in Clark and Diani, 1996: 311) Yet Touraine insists on the unity of a general definition of social movements even when the characteristic form of new social movements is fragmented rather than united around a common principle. This helps to explain why the main focus of Touraine's work shifts away from movements as such (although he has continued to publish in this area and places increasing emphasis on feminism and social transformations associated with gender, e.g., *Le monde des femmes*, 2006).

3. The beginning of the third period of Touraine's work is marked by *Le retour de l'acteur* (1984) (Return of the Actor), a collection of essays and articles offering a restatement of

his theoretical position in the light of his studies of social movements and political developments. It also responds to contemporary developments in social theory, notably the growing influence of post-structuralism and post-modernism. The key concepts from earlier work remain: he continues to demystify the illusion of social order and coherence, he insists on the capacity of society to make history (historicity) and, predicably, the actor remains the central concept. He questions and updates each of the themes. The direction of travel is toward a more extended analysis of the meaning of modernity and Touraine's diagnosis of the crisis which develops from the inherent tension between rationalization and 'subjectivation,' which he capitalizes as Reason and Subject to indicate their status as historical principles. The fullest statement of these themes is *Critique de la modernité* (1992) (Critique of Modernity). The aim of this work is to seek a new definition of modernity distinct from rationalization and secularization, which are typically seen as the all-absorbing, linear processes of the last three centuries. Touraine's vision of the transition to modern society is one of a creative, turbulent, interrupted process involving cultural fields, social actors, and forms of political action. The period from the 1960s to the 1980s was the period of crisis in the transition from industrial to programmed society when the conflict between the technocratic vision and the new movements to liberate the subject came to a head. Touraine's thesis rejects the myth of integration associated with concepts of tradition and community. Instead, from a resolutely modernist position he explores the societal conflicts that are characteristic of modernity and their associated social philosophies and cultural models.

The key theme in *Critique de la modernité* is the dissociation between means and ends, the divergence between the system and actors, instruments and meaning. The failure of dialogue between the two principles is responsible for the dictatorship of reason on the one hand and an obsession with the subject (e.g., narcissistic individualism) on the other. Cultural identity has ceased to be about community and is a matter of personal conviction without social roots. While there are some clear similarities between this distinction and other categories in contemporary social theory (system and life world in Habermas, or structure and agency in Giddens, for example), Touraine has a unique understanding of the crisis that is produced by the opposition between growing system rationalization and its opposite, the retreat into subjectivity. His credo is that "we do not belong to a society; we attempt to construct our life histories, to be the actors of our own histories by mobilizing our cultural resources to enter the world of markets and technologies" (Touraine in Clark and Diani, 1996: 300). The individual is neither a victim of the system nor an autonomous entity. But some of Touraine's critics are right when they point out that subjectivation holds more fascination for him than the workings of institutions and power (Martuccelli in Clark and Diani, 1996: 244).

Touraine gives an explicit account of how the processes of modernity are played out by specific forces of modernization at work in nationalism, commercial competition and consumption, and sexuality (1992: Chapter 3). These are the 'actors of modernization,' whose impact cannot be reduced to instrumental logic. The national state is modernity's political form and connections between modernization and nation-state formation have been central to the story of progress. But the nation relies on cultural particularity as well as universalizing, legal-rational principles. Similarly, the organization of firms for economic production for exchange and profit sees a disjunction between the rationalization of the technical system and the ends embodied in 'strategy.' The system of consumption presents a corresponding problem: previously it could be understood as the language of taste determined by social status or the actor's position in the economic system. As mass consumption develops, the correspondence between levels of status is undermined and the actor loses points of reference and becomes an "actor in search of himself" (p. 144). In relation to gender, the critique recognizes the system of gender domination and control of reproduction as well as the commercial exploitation of sexuality. The new element is the study of the conditions in which women become subjects, indeed principal actors in modern societies.

The crisis of modernity is therefore a long-developing crisis of fragmentation between the four worlds of the nation, the firm, consumption, and self (the subject able to reflect on his or her own identity) (1995: 274). Echoing Marx, Touraine says

> Human beings make their own history, but they do so through social conflicts, and on the basis of cultural choices. We have not left industrial society and we have not entered post-modernity; we are building a programmed society in which the production of symbolic goods has the central role that once devolved upon the production of material goods in industrial society. It is possible that this society may see a complete divorce between the economy and culture, just as the forces of economic or scientific development created islands of rationality in a world of tradition and community at the beginning of modernity. Yet such a rupture must be seen as pathological, and it can be analysed only as a divorce between complementary domains. The political system must establish mediations between the two.
>
> 1995: 358–359

Touraine developed the political and ethical aspects of this critique at greater length in a series of related publications including *Qu'est-ce que la démocratie?* (1994), *Pourrons-nous vivre ensemble? Egaux et différents* (1997), *Comment sortir du libéralisme?* (1998), *Un nouveau paradigme: Pour comprendre le monde aujourd'hui* (2005) which advocate a renewal of democratic politics on the basis of newly formulated demands for rights, especially cultural rights. These publications span the period when center-left and coalition parties dominated European politics, as well as the later turn to the right and the years of economic crisis. They not only defend liberal principles of equality, majority rule, and the limitation of state power but they also articulate notions of the subject, social movements, and rights that come directly from Touraine's theory of action – or more precisely, his theory of the formation of new social actors in the context of developments that threaten social decay and decomposition. While some of the content addresses specifically French manifestations of more widespread problems (e.g., *laïcité* and multiculturalism), most of it is

pertinent to any society experiencing a decline in democracy and the effects of globalization.

The global financial crisis of 2007–09 presented an opportunity for Touraine to apply his thinking to a context in which subjects or 'actors' appeared to have become very remote from the workings of the economic system. In *Après la crise* (2010) (After the Crisis) he analyzes the growing contradiction between actors and systems in what he calls the 'post-social' era and the implications that this has for the capacity for collective action by actors. The crisis is a "theatre without actors" (2010: 57) and "the crisis prevents us from transforming ourselves into subjects" (2010: 125) but he does not counsel despair. Although the processes of subjectivation have enfeebled the world of actors, the crisis has created the possibility for new, moral subjects to become cognizant of rights they have been denied and therefore to become collective actors.

4. There is no clear break between the critique of modernity and Touraine's work since 2001. The opening sentence of *La société invisible* (1997) 'Society has disappeared' already conveyed a challenge to concepts of society and the social rooted in the past. His *La fin des sociétés* (2013), 'The End of Societies,' brings together the fruits of a lifetime's work and offers a diagnosis of unregulated capitalism, the financial crisis, the Arab Spring and its aftermath, and other epochal changes. His target of criticism is all kinds of relationships between human subjects and social actors in a post-industrial, 'hypermodern' society where the logic and meanings of 'the social' are decomposing along with politics, institutions, and social movements. He deploys key concepts in a mixture of conventional and novel ways. If *Society* describes the use of resources according to a particular cultural pattern or orientation shaped by social institutions such as the welfare state (2013: 38), the 'end of society' refers to the hollowing out of key institutions, a process that is marked by the loss of capacity for integration and control of the relationship between the system and actors. *Modernization* refers to the multiple pathways (not only western) by which societies achieve a state of continuous self-transformation through the application of technical and intellectual means while *Modernity*, a singular concept, indicates the primacy of reason and universalistic criteria of judgment (2013: 413 ff.). Modernity does not refer to a particular type of society or form of social relation but rather the elevation of universal principles: "no [society] can be called modern if it refuses to place scientific reason above other modes of thinking … [or] rejects the idea of universal human rights" (2013: 420). Touraine sees the relationship between modernization and modernity as inherently complex, being manifested in the whole range of societies and political regimes as well as unfinished or interrupted projects of modernization. At the center of his thinking is the *Subject*. The subject is not the individual person (let alone a social actor) but that which gives human beings the capacity to create and to transform nature and the self. In the post-social era the subject is liberated from the constraints of religion and society which prevailed in former times. His subject is the carrier of rights that can be invoked to challenge any group or individual which denies universalist principles. The related concept of *subjectivation* refers to a convergence between individuality and universality, the highest expression of the subject (2013: 17).

The potential for subjectivation increases in the post-social context when the individual is compelled to claim the right to be an actor in his or her own world. Rights are not a natural 'given' (whether by religious mandate or natural law). They are the outcome of individual and collective historical struggles that reinforce the individual's consciousness of being a subject defined by universal rights, express the creative power of the subject, and lead to the growth of freedom. Touraine underlines the difference between this conception and the dominant view of ethics as the measure of the capacity for individuals to pursue their interests and impose their personal choices. His insistence on 'the right to have rights' (echoing Arendt) links him directly to the classical liberal democratic tradition while he simultaneously offers a distinctive post-social interpretation of ethics. Despite its apocalyptic-sounding title, 'The End of Societies' is not a pessimistic work, nor does it envisage the end of sociology or the demise of critique.

Commentary and Critiques

It is noticeable that Touraine's sociology is less well known in the English speaking world than the work of some of his contemporaries. For example, in his study of French post-war social theory, Robbins (2012) chooses five theorists including Althusser, Lyotard, and Bourdieu but not Touraine, despite the quantity and breadth of his output. His influence has been significant at various times in the sociology of work, the sociology of social movements, and in political life in France (not to mention his international profile in professional associations in sociology), but his theoretical thinking has generally been less widely applauded. It has a dialectical quality noted by Michel Crozier, his early contemporary in the CNRS, as both brilliant and stylistically opaque (in Clark and Diani, 1996: 14). That is by no means a unique difficulty in French theoretical writing and Touraine has spoken of his fear of being trapped by an ideology, including one of his own making. However, commentators have regularly drawn attention to a number of more significant difficulties in the logic and empirical substantiation of Touraine's arguments. Three can be highlighted as particularly important. First, there is the problem of the relationship between sociology and history. The early emphasis on historicity against the backdrop of the transition from industrial to postindustrial society led him to seek one central conflict in every form of society. For Alexander (in Clark and Diani, 1996: 224), this was a 'debilitating historicism' that undermined his critique of the classical approach to social movements and that prevented him from taking a more rounded view of the culture and institutions of civil society. There is a second, related, criticism which points to an empirical failure to recognize the power of institutions and incorporate them into his analysis. His sociology of action does not embrace all forms of action. Joas and Knöbl (2009: 427), for example, point out that Touraine does not examine the machinery of the state, institutions, or markets: "often, he merely characterizes them by deploying the highly imprecise term 'anti-subject.'" Similarly, Martuccelli (in Clark and Diani, 1996: 244) identifies a failure to theorize

power and the ruling class. This alleged weakness needs to be seen in context, which is Touraine's aim to provide sociology of the subject and a corrective to 'institutional sociology.' A third type of criticism is directed toward Touraine's theory of democracy, which is to all appearances like a conventional defense of liberalism. Some objections are misplaced because Touraine's writing on the subject is often polemical and linked to specific circumstances. It is more fruitful to read his political writings since the late 1990s in the light of his theoretical statement in *La fin des sociétés*, which offers an interpretation of the concept of a universal 'right to have rights as subject,' able to incorporate particularities, without the need for recourse to historical, metaphysical, or religious foundations.

For Touraine "the starting point for social science will always be a distrust of 'the social,' and it must distance itself from anything that reduces the workings of society to a technico-administrative operation" (1992: 127). In this sense, he is a critical sociologist who seeks categories that can capture the self-transforming character of social relations, avoiding determinism and invalid assumptions of social coherence. As an active public intellectual, he has been equally committed to social and political action that can serve to increase the capacity of subjects to act with freedom and solidarity as they make their own history.

See also: Action, Theories of Social; Collective Action; Collective Behavior, Sociology of; Historicism; Liberalism: Political Doctrine and Impact on Social Science; Marxism in Contemporary Sociology; Modernity; Movements, Social; Postindustrial Societies; Sociological Theory; Work and Industry, Sociology of.

Bibliography

Castoriadis, Cornelius, 1975/1998. The Imaginary Institution of Society (K. Blamey, Trans.). MIT Press, Cambridge, MA.
Clark, Jon, Diani, Marco (Eds.), 1996. Alain Touraine. Falmer, London.
Joas, Hans, Knöbl, Wolfgang, 2009. Social Theory. Cambridge University Press, Cambridge.
Robbins, Derek, 2012. French Post-War Social Theory. Sage, London.
Rose, Michael, 1996. Alain Touraine: *Sociologue du Travail*, Proudhonian, Pessimist. In: Clark, Jon, Diani, Marco (Eds.), Alain Touraine. Falmer, London, pp. 17–31.
Touraine, Alain, 1965a. La conscience ouvrière. Seuil, Paris.
Touraine, Alain, 1965b. Sociologie de l'action. Trans. The Sociology of Action. Seuil, Paris.
Touraine, Alain, 1968. Le Mouvement de Mai ou le Communisme Utopique. Trans. The May Movement: Revolt and Reform. 1971. Random House, New York.
Touraine, Alain, 1969. La société post-industrielle. Trans. The Post-Industrial Society. 1974. Wildwood House, London.
Touraine, Alain, 1973a. Production de la société. Trans. The Self-Production of Society. 1977. Chicago University Press, Chicago.
Touraine, Alain, 1973b. Vie et mort du Chili populaire. Seuil, Paris.
Touraine, Alain, 1974. Pour la sociologie. Seuil, Paris.
Touraine, Alain, 1977. Un désir d'Histoire. Stock, Paris.
Touraine, Alain, 1978. La voix et le regard. Seuil, Paris. Trans. The Voice and the Eye. 1981. Cambridge University Press, Cambridge.
Touraine, Alain, Hegedus, Z., Dubet, F., Wieviorka, M., 1980. La Prophétie antinucleaire. Seuil, Paris. Trans. Anti Nuclear Protest: The Opposition to Nuclear Energy in France. 1981. CUP, Cambridge.
Touraine, Alain, 1984. Le retour de l'acteur. Fayard, Paris. Trans. Return of the Actor. 1988. University of Minnesota Press.
Touraine, Alain, 1988. La parole et le sang. Odile Jacob, Paris.
Touraine, Alain, 1992. Critique de la modernité. Trans. Critique of Modernity. 1995. Blackwell, Oxford.
Touraine, Alain, 1994. Qu'est-ce que la démocratie? Fayard, Paris.
Touraine, Alain, 1997a. La société invisible. Seuil, Paris.
Touraine, Alain, 1997b. Pourrons-nous vivre ensemble? Egaux et différents. Fayard, Paris. Trans. Can We Live Together? 2000. Polity, Cambridge.
Touraine, Alain, 1998. Comment sortir du libéralisme? Fayard, Paris. Beyond Neoliberalism. 2001. Polity, Cambridge.
Touraine, Alain, 2005. Un nouveau paradigme. Pour comprendre le monde d'aujourd'hui. Fayard, Paris. Trans. A New Paradigm for Understanding Today's World. 2007. Polity, Cambridge.
Touraine, Alain, 2006. Le monde des femmes. Fayard, Paris.
Touraine, Alain, 2010. Après la crise. Seuil, Paris. After the Crisis.
Touraine, Alain, 2013. La fin des sociétés. Seuil, Paris.

Tourism

Shinji Yamashita, The University of Tokyo, Tokyo, Japan

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Tourism has become a tremendously important phenomenon in the contemporary world. This article deals with the sociocultural implications of tourism from an anthropological point of view. After presenting the general framework of the anthropology of tourism, the article examines various aspects of the phenomenon of tourism: the tourist, the tourist experience, the host society, 'touristic culture' or the culture produced by tourism, and sustainability in tourist development. The article concludes by suggesting the future.

The Age of Tourism

Humankind has traveled since ancient times, but the phenomenon of tourism is a new kind of travel born of modernity. The word 'tourism' began to be used only in the nineteenth century due to the development of modern transportation technology – such as the steamship and train – and to socioeconomic changes caused by the industrial revolution. Prior to these developments, travel was difficult, requiring patience and effort. In fact, the English word 'travel' is etymologically related to the French word 'travail' meaning 'labor.'

In the West in the second half of the twentieth century, the prevailing work ethic was transformed as people's desire for making life worthwhile through leisure and enjoyment became increasingly evident. The length of the working week decreased and paid holidays lengthened. As a result, tourism as a form of leisure activity became widespread in developed countries and also among the newly emerging middle classes of developing countries.

According to the United Nations World Tourism Organization (UNWTO), 980 million people traveled across their respective national borders in 2011. The figure is expected to grow to 1.6 billion in 2020. We are undeniably in the age of global human mobility. Against this background of a massive expansion in tourism, many countries, both the developed and the developing, are now pouring resources into the tourism industry, which they see as an important option for economic development. The World Tourism and Travel Council (WTTC) estimates that the total economic contribution of tourism industry in 2011 was 9% of the world gross domestic product and it concerned 8.3% of the gainfully employed people around the world. Tourism has become, thus, tremendously important in the contemporary world.

The Anthropology of Tourism

In the social sciences, however, tourism was not taken seriously as a subject for research, as we see in the personal histories of pioneer tourism researchers (Nash, 2007; Crick, 1989). Among anthropologists, this area of research made its debut quite recently. The first academic symposium on tourism was held in conjunction with the annual meeting of the American Anthropological Association in Mexico City in 1974. Up until then, there were a few pioneering efforts, but nothing systematic had been carried out. The results of the symposium were published in the book, *Hosts and Guests: The Anthropology of Tourism* (Smith, 1989). Focusing on the relationships between 'hosts' (society receiving the tourists) and 'guests' (tourists themselves), this book was the first edited volume to deal with the anthropology of tourism. In the 1980s, tourism became established as an item on the menu of anthropology in the West, both for research and teaching. Outside Western academic circles, in Japan for instance, the anthropological study of tourism was initiated in the late 1980s by Shuzo Ishimori at the National Museum of Ethnology in Osaka. The first panel on the anthropology of tourism was organized at the 1990 annual meeting of the Japanese Society of Ethnology (called the Japanese Society of Cultural Anthropology since 2004) held at the University of Tokyo.

The anthropology of tourism investigates the following three fundamental research areas. The first deals with aspects of the political economy of tourism, which include tourist policy, tourist development, the tourist industry, and power relations between guest tourists and host society. The second examines tourism as a sociocultural phenomenon. The main corpus of anthropological research focuses on this area. The third pursues the meanings of tourism and tourist experiences. The central task of these three areas of focus in the anthropological research of tourism is to investigate the sociocultural implications of tourism. In fact, in many parts of the world it has become obvious that tourism is playing a culturally dynamic role. Amidst the 'global cultural flows' (Appadurai, 1996; Greenblatt et al., 2010), social boundaries are weakening and people are enjoying the benefits of cross-cultural contact. Consequently, while expressions of traditional culture are increasingly fragmented, culture is also being reinvented, especially in the context of tourism. Therefore, MacCannell (1992) observed that 'tourism is the primary ground for the production of new cultural forms on a global basis.'

These circumstances present a challenge to the student of culture. It is impossible to treat culture in the way that it once was, as a closed system of meanings and symbols. It is also increasingly difficult to use the old-fashioned ethnographic method of recording coherent, patterned cultural systems enclosed within discrete territories and peoples, as Hannerz (1996), Gupta and Ferguson (1997), Marcus (1998), and others have discussed.

The Tourist

Smith (1989) defines a tourist in the following way: "a temporarily leisured person who voluntarily visits a place away from home for the purpose of experiencing a change." Here, 'work' is contrasted with 'leisure' and time is structured as follows: tourists are released from work, this leisure time they devote to themselves, and they return to work refreshed once more. This structure of time is very similar to that of ritual as analyzed by Leach (1961) who suggested that ritual puts a stop to profane time, and creates sacred time. The end of the ritual signifies reintegration with secular time. Building on this analogy of ritual, Graburn (1989) treats tourism as a descendant of the 'sacred journey' or some version of a pilgrimage. In fact, in Europe, Japan, and many other parts of the world, pilgrimage, which provided the model for tourism, was simply a ritual-oriented journey. Tourism, therefore, may be a form of ritual in the modern world, as Graburn suggested.

Just as rites of passage bring about a change in status or circumstances through the ritual experience, what the tourist is looking for is the experience of a change. The essence of tourism always consists of experiences that are different. As a result, a trip is a success if a change can be seen in the self. People refresh themselves by traveling: they may feel they are 'reborn' or recharged with a new lease on life.

A difference in perceptions of time may also apply to the tourist space, in what Urry (1990) has called the 'tourist gaze.' Applying Michel Foucault's concept of the 'gaze,' Urry states that the tourist experience is 'to gaze upon or view a set of different scenes, of landscapes or townscapes which are out of the ordinary.' Just like tourist time, then, tourist space is marked and 'framed' to use MacCannell's term. Tourism is, as MacCannell (1992) put it, an ideological framing of history, nature, and tradition; this framing has the power to reshape culture and nature to its own needs.

However, in terms of the recent developments in tourism, one should note that there are perceptual changes about tourist time and space. In the acceleration of contemporary global human mobility, the distinction between 'home' and 'away' sometimes blurs. For example, Bali has become 'another home' for Japanese women tourists who are married to Balinese men whom they met after repeated tours to Bali (Yamashita, 2003). Since 2002, the Malaysian government has been promoting Malaysia as a 'second home' for foreign retirees. In these examples, Smith's definition of the tourist as a temporarily leisured person who is away from home cannot be easily applied; for these lifestyle migrants/tourists, 'leisure' and 'work,' 'home' and 'away,' and 'dwelling' and 'traveling' are not necessarily opposed. Therefore, we need a new definition of tourism in accordance with the changing perception of tourist time and space (Yamashita, 2012).

Like lifestyle migrants/tourists, anthropologists too are either traveling dwellers or dwelling travelers in their ethnographic journey. Some have noted similarities between tourists and anthropologists. Crick (1995), in discussing 'the anthropologist as tourist,' suggests there is a partial overlap in identities among tourists and anthropologists, and tourists may be regarded as 'distant relatives' of anthropologists. However, based on his 'ethnographer-as-tour guide' experiences in Bali, Indonesia, Bruner (1995, 2005) observes that there are apparent disparities between touristic and ethnographic ways of viewing such events as festivals. 'Anthropologist gaze,' in their professional viewing of the world is, after all, different from 'tourist gaze.'

The Tourist Experience

There are two important theories of the tourist experience. The first is the theory developed by MacCannell (1976) that tourism itself consists of the search for 'authenticity.' This starts from Marx's observation that living in the modern world produces a state of alienation in which we cannot realize our real selves. The reason why people travel is, therefore, to escape from the real world in order to experience a different world, and to discover the 'authentic.'

On the other hand, Boorstin (1962), writing before MacCannell, observes that the authentic tourist experience is impossible in the contemporary world. Rather than being a journey in search of authenticity, travel is increasingly based on artificial images. Before setting out on a trip, tourists read the guidebook. They then go to the places written up in the guidebook, and take photographs similar to those in the guidebook. This is an experience that Boorstin has called a 'pseudo-event.' Thus, the tourist experience ends up oscillating between being a 'pursuit of authenticity' and a 'pseudo-event.' Pseudo-events, simulations and fabrications, and staged and scripted events that are a kind of counterfeit version of actual happenings have become the dominant feature of tourism space and time. However, not all tourists seek authenticity; nor are all tourist experiences pseudo-events. Cohen (1979), therefore, does not regard the tourist experience as either of these two ends. He suggests instead that there are a number of different modes of tourist experience ranging from the recreational mode in search of mere pleasure to the existential mode in search of one's real self. There are, indeed, many different types of tourism (e.g., 'recreational tourism,' 'cultural tourism,' 'ethnic tourism,' 'ecotourism,' 'heritage tourism,' and so forth), and many different types of tourists, from mass tourists to backpackers. Tourist experiences also vary depending on 'who travels, and where' (Moon, 1997).

Furthermore, today, as the boundary between 'the original' and 'the copy' is becoming increasingly blurred, the very basis of the question of whether tourism is a search for authenticity or a pseudo-event has lost its meaning. *Cannibal Tours*, a film by O'Rouke (1987), illustrates this point. The film depicts a tour to the villages on the Sepik River in Papua New Guinea. As Bruner (1989) says, the tourists already know that in New Guinea things like cannibalism no longer exist. If cannibalism were to exist, or if there were to be no air-conditioned cruiser, they would probably not have gone to New Guinea. What the tourists encounter is not the real 'primitive,' but a contrived and reconstructed primitiveness. The substance of this kind of 'postmodern tourism' is, then, the experience of a copy, fabrication, simulation or *simulacrum*, and the illusory copy becomes more important than the real thing.

Cannibal Tours also raises the critical question of how this type of tourism is based upon a neocolonial relationship between the North and the South. In fact, the majority of the world's tourists are from the 'northern' countries: 57% from Europe and 16% from North America. Eighty percent of all international travelers are nationals of just 20 countries (Survival International, 1995; quoted in Mowforth and Munt, 2003). In terms of expenditure on tourism, the world's top 10 countries on the basis of spending on tourism, in 2009 – Germany, the United States, the United Kingdom, China, France, Italy, Japan, Canada, Russia, and the Netherlands – were responsible for 47.3% of the world total (UNWTO statistics). However, one should note that this Western-dominated tourist map will be changing, as the tourist market is growing rapidly among the middle classes in developing countries in Asia. China, for example, is predicted to produce one billion international tourists by 2020.

The Host Society

The anthropology of tourism investigates the 'contact zone,' or interactions between hosts and guests. The central question had been how the development of tourism influences the culture of the host society. Much of the debate has focused especially on whether the effects of tourism are beneficial or not to the host society. In discussions about the relationship between tourism and culture, tourism was often seen in a negative light, the general tenor being that it contributes to the destruction of traditional culture. However, this line of reasoning is not necessarily correct. First of all, tourism is not the only influence on host societies. Local cultures are also transformed by other factors, including industrialization, urbanization, migration, the mass media, pollution, poverty, and civil war. Therefore, it is simplistic to consider only the 'impact' of tourism as an exogenous force, assessing its effects as either 'good' or 'bad.' What is at stake is not simply the impact of tourism on local cultures, but rather how the local cultures develop and make use of the dynamic process of tourism to redefine their own identities (Lanfant et al., 1995).

It is useful at this point to recall Clifford's distinction between two metanarratives of culture: metanarratives of homogenization and loss, on the one hand, and metanarratives of emergence and invention, on the other (Clifford, 1988). A narrative of homogenization becomes an 'entropic' narrative or one of loss, where modernization or Westernization means a loss of indigenous tradition. But such narratives that regard tourism as precipitating the disappearance of an 'innocent' tradition unaffected by modern Western civilization are often incorrect, because they presuppose the idea that cultural traditions date back to primordial times, which is seldom the case in actuality (Hobsbawm and Ranger, 1983).

Regarding the creation of culture in tourist contexts, therefore, special attention should be paid to Clifford's notion of narratives of emergence and invention. Culture in this perspective is observed as always emerging; it is not something static, but a dynamic, evolving process. An example may be taken from Bali, Indonesia. In Bali, traditional culture, and particularly the traditional dances such as the *kecak* dance, have been elaborated and refined since the 1930s under the gaze of Western artists and tourists. 'Traditional culture' reconstructed in this way not only has become 'symbolic capital' for developing tourism but also has been appropriated by the local people to assert their own identity (Yamashita, 2003).

Touristic Culture

Anthropology was conventionally concerned with the view of culture as customs enacted unconsciously. Given the discussion above, culture is rather an object of conscious manipulation within a broader social, economic, and political context. In other words, culture, in the context of tourism, is not the 'authentic culture,' but a newly created, scripted 'staged culture.' Staged culture in the tourist context, or what Picard (1996) has called 'touristic culture' from his study in Bali, such as dance performances, cultural exhibitions, or souvenir production for tourists, is usually disparaged, because it is not 'authentic.' However, in the case of Bali, the 'tourist gaze' has stimulated local creativity in the arts. What, then, is the mechanism for the emergence of touristic culture, particularly tourist art?

As an example, the Sepik carvings, the souvenirs that appeared in *Cannibal Tours*, originally were used in ritual contexts. At the end of the nineteenth century, however, they were designated 'primitive art' by Western explorers, and were seen as having objective, 'authentic' value by Western scholars. Now they have been transformed into 'touristic art' to be sold to tourists. This kind of transformation occurs in the scheme of what Clifford (1988) called the 'modern art-culture system.' The Sepik carvings have, in this way, been transformed from cultural (ritual) objects to artistic objects, and from artistic objects to tourist (commercialized) objects (see also Graburn, 1976).

These phenomena do not comprise an 'entropic narrative' of traditional cultures that are vanishing. Rather, they have the makings of new cultures emerging in the context of tourism. Tourism, then, has become a catalyst for contemporary cultural production.

Tourism and Sustainability

Sustainability became something of a buzzword from the 1980s, and particular kinds of tourism could lead to increased awareness of environmental and other problems. In 1989 in Zakopane, Poland, the theme of the first general meeting of the International Academy for the Study of Tourism was 'alternative forms of tourism' such as 'appropriate tourism,' 'responsible tourism,' and 'ecotourism' (Smith and Eadington, 1992).

Sustainable tourism can be defined as development that will benefit people today without harming the resources for future generations. Sustainable tourism seeks development that maintains a balance between the ecosystem and the society and culture concerned. According to de Kadt (1992), sustainable development involves the following traits: ecological soundness; small-scale production; recognition of needs other than those of material consumption; equal consideration

of the needs of all (including future generations); and political involvement from below. Ecotourism may be a fundamental model of sustainable tourism, though it is a challenging business.

In addition to the natural environment, culture needs sustainable developments as well. Let us take an example again from Bali, which is often regarded as a successful case of cultural tourism. Until the 1980s, Balinese tourism had been confined to the southern part of Bali, but in recent years it has spread to places in all parts of Bali. Furthermore, commercialization has accelerated. Through this process, culture has become an asset, a form of 'cultural capital,' which can now be owned, managed, and controlled by outsiders – Jakartan or otherwise. Culture is thus slipping away from Balinese hands. Cultural tourism was originally encouraged so that the Balinese could control their own culture, but that culture has begun moving beyond their control. This is the paradox that Balinese cultural tourism faces today (Yamashita, 2010).

The Future of Tourism: Toward a Reflexive Tourism

As mentioned at the beginning of this article, tourism was born out of the development of the modern society. Urry (1995) states: "Central to the idea of modernity is that of movement, that modern societies have brought about striking changes in the nature and experience of motion or travel." However, the situations observed in contemporary Bali indicates that tourism, like other institutions of modernity, stands at a crossroads in the path of what Ulrich Beck et al. (2005) have called 'reflexive modernization.'

'Reflexive modernization' refers to the creative (self-) destruction of an entire epoch, that of industrial society. Beck writes: "This new stage, in which progress can turn into self-destruction, in which one kind of modernization undercuts and changes another, is what I call the stage of reflexive modernization." 'Reflexive modernization,' then, is taken as a process by which one modernity is replaced with another.

Further, as Beck points out, the other side of the obsolescence of the modern society is the emergence of 'risk society.' Tourism also must contend with the management of risk which modern society has produced. Examining the case of Bali, Yamashita has proposed a 'reflexive tourism,' which seeks an alternative modernity in the field of tourism (Yamashita, 2010). Ecotourism and heritage tourism, for example, may be regarded as typical forms of reflexive tourism, because they attempt to manage the contemporary environmental destruction and cultural disruption from the point of view of tourism. The future of tourism should be in the direction of this reflexive vein in which we attempt to manage 'risk society.'

See also: Environmental Sociology; Gay and Lesbian Tourism; Leisure, Sociology of; Modernity; Tourism, Geography of.

Bibliography

Appadurai, A., 1996. Modernity at Large: Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Beck, U., Giddens, A., Lash, S., 2005. Reflexive Modernization: Politics, Tradition and Aesthetics in the Modern Social Order. Polity Press, Cambridge.

Boorstin, D.J., 1962. The Image: Or what Happened to the American Dream. Atheneum, New York.

Bruner, E.M., 1989. Of cannibals, tourists, and ethnographers. Cultural Anthropology 4, 438–445.

Bruner, E.M., 1995. In: Lanfant, M.F., Allcock, J.B., Bruner, E.M. (Eds.), The Ethnographer/Tourist in Indonesia, pp. 224–241.

Bruner, E.M., 2005. Culture on Tour: Ethnographies of Travel. The University of Chicago Press.

Clifford, J., 1988. The Predicament of Culture: Twentieth-century Ethnography, Literature, and Art. Harvard University Press, Cambridge, MA.

Cohen, E., 1979. A phenomenology of tourist experiences. Sociology 13, 179–201.

Crick, M., 1989. Representations of international tourism in the social sciences: sun, sex, sights, savings and servility. Annual Review of Anthropology 18, 307–344.

Crick, M., 1995. In: Lanfant, M.F., Allcock, J.B., Bruner, E.M. (Eds.), The Anthropologist as Tourist: An Identity in Question, pp. 205–223.

de Kadt, E., 1992. In: Smith, V.L., Eadington, W.R. (Eds.), Making the Alternative Sustainable: Lesson from Development for Tourism, pp. 47–75.

Graburn, N.H.H., 1989. In: Smith, V.L. (Ed.), Tourism: The Sacred Journey, pp. 21–36.

Graburn, N.H.H. (Ed.), 1976. Ethnic and Tourist Arts: Cultural Expressions from the Fourth World. University of California Press, Berkeley, CA.

Greenblatt, S., et al., 2010. Cultural Mobility: A Manifesto. Cambridge University Press, Cambridge.

Gupta, A., Ferguson, J. (Eds.), 1997. Anthropological Locations: Boundaries and Grounds of a Field Science. University of California Press, Berkeley, CA.

Hannerz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, London and New York.

Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge.

Lanfant, M.F., Allcock, J.B., Bruner, E.M. (Eds.), 1995. International Tourism: Identity and Change. Sage, London.

Leach, E.R., 1961. Rethinking Anthropology. Athlone, London.

MacCannell, D., 1976. The Tourist: A New Theory of the Leisure Class. Schocken, New York.

MacCannell, D., 1992. Empty Meeting Grounds: The Tourist Papers. Routledge, London and New York.

Marcus, G., 1998. Ethnography through Thick and Thin. Princeton University Press, Princeton, NJ.

Moon, O., 1997. Tourism and cultural development: Japanese and Korean contexts. In: Yamashita, S., Din, K.H., Eades, J.S. (Eds.), Tourism and Cultural Development in Asia and Oceania. Penrbit Universiti Kebangsaan Malaysia, Bangi, Malaysia, pp. 178–193.

Mowforth, M., Munt, I., 2003. Tourism and Sustainability: New Tourism in the Third World, second ed. Routledge, London and New York.

Nash, D. (Ed.), 2007. The Study of Tourism: Anthropological and Sociological Beginnings. Elsevier Science Publishing Company.

O'Rouke, D., 1987. Cannibal Tours (Film). Distributed by Direct Cinema, Santa Monica, CA.

Picard, M., 1996. Bali: Cultural Tourism and Touristic Culture. Archipelago Press, Singapore.

Smith, V.L. (Ed.), 1989. Hosts and Guests: The Anthropology of Tourism, second ed. University of Pennsylvania Press, Philadelphia, PA.

Smith, V.L., Eadington, W.R. (Eds.), 1992. Tourism Alternatives: Potentials and Problems in the Development of Tourism. University of Pennsylvania Press, Philadelphia, PA.

Survival International, 1995. Tourism and Tribal Peoples. Background Sheet. Survival International, London.

Urry, J., 1990. The Tourist Gaze: Leisure and Travel in Contemporary Societies. Sage, London.

Urry, J., 1995. Consuming Places. Routledge, London and New York.

Yamashita, S., 2003. Bali and Beyond: The Exploration of the Anthropology of Tourism (Translated by J. Eades). Berghahn Books, New York and Oxford.

Yamashita, S., 2010. A 20–20 vision of tourism research in Bali: towards reflexive tourism studies. In: Pearce, D.G., Butler, R.W. (Eds.), Tourism Research: A 20–20 Vision. Goodfellow Publishing, Oxford, pp. 161–173.

Yamashita, S., 2012. Here, there, and in-between: lifestyle migrants from Japan. In: Haines, D.W., Yamanaka, K., Yamashita, S. (Eds.), Wind over Water: Migration in an East Asian Context. Berghahn Books, New York and Oxford, pp. 161–172.

Tourism, Geography of

Allan M Williams, School of Hospitality and Tourism Management, University of Surrey, Guildford, UK
Gareth Shaw, University of Exeter Business School, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The field of tourism geography is seen in terms of broad, and sometimes overlapping, concerns with the space, place, and environmental dimensions of geography. Despite broad consensus around these dominant themes, the definition of tourism geography remains problematic, because of the blurred understandings of both the temporalities and motivations of tourism and tourists. Methodologically and theoretically, there have been significant shifts over time from more empirical to more theoretical approaches, and from spatial models, to political economy and cultural perspectives.

Introduction

The field of tourism geography has mostly been seen in terms of concerns with the space, place, and environmental dimensions of tourism, reflecting a long tradition in the larger discipline of geography (Wilson, 2012). The definition of tourism geography, however, remains problematic, with debates focusing on both its temporality – where, for example, does it fit in relation to temporary migration and recreational studies – and motivations, The rise of the mobilities approach has added to these challenges in recent years, while also providing new perspectives that have enriched tourism geography (Duncan, 2012).

There have been a number of shifts in the dominant theoretical and methodological approaches in tourism geography in recent years, again reflecting broader shifts in geography and the social sciences, and this article maps out three of these. The positivist approach has largely been inductive and has involved the construction of spatial models, particularly of tourism flows, an approach that has been fortified recently by technological innovations such as GPS and GIS. The political–economy approach has drawn on capital accumulation, commodification, and regulation theories, while cultural interpretations have researched the discursive, representational, and performative realms of tourism and tourist experiences. The way in which tourism geography is increasingly engaging with other social sciences is exemplified by the work on environmental issues, especially sustainable tourism, and the role of tourism in making and remaking spaces and places. Tourism geography has, however, largely been dominated by human geographers, and the study of environmental issues, for example, has mostly been in the domain of social science, rather than a meeting ground of natural and social science – or physical and human geography.

While there is a distinctive field of tourism geography studies, with its own specialist journal, *Tourism Geographies: An International Journal of Tourism Space, Place and the Environment* (first published by Routledge in 1999), there are considerable overlaps with the interests of economic and cultural geographers, as well as mobilities researchers (Urry, 2007). These have infused tourism geography with greater theoretical content in recent years, counterbalancing a traditional emphasis on empirical and applied research. Nevertheless, this has raised concerns for some tourism geographers that their field has become a theoretical borrower, leaning heavily on cognate disciplines, rather than generating its own distinctive theoretical contributions. However, other commentators point to the creative and leading role of tourism geographers in many cutting-edge discourses on subjects such as globalization, place making, and visual representation.

Definition

A recurring issue in tourism geography has been the generic definition of tourism. Most definitions emphasize that tourism is characterized by nonpermanent moves, an intention to return home within a relatively short time period, and a travel purpose other than taking up permanent residence or employment. Two difficulties arise from this apparently simple definition. First, there is no clear conceptualization of 'permanence,' an issue that has gained in importance because of the increasing diversity of human mobility and the blurring of tourism and migration (Williams and Hall, 2002). The response of leading international tourism bodies, such as the United Nations World Tourism Organization, has been to rely on the pragmatic definition of absence from home for at least one night but for less than 12 months – a device that signally fails to engage with the overlap with noneconomic temporary migration (Bell and Ward, 2000). Second, the definition chaotically bundles together diverse motivations and practices, ranging from business tourism, to visiting friends and relations, to education to pleasure. There is considerable debate as to whether tourism should only be seen as concerned with pleasure and leisure trips. However, this approach is problematic if only because many overnight trips are multipurpose, a blurring that further obfuscates the picture for those seeking simple definitions.

Another important question relating to tourism geography is whether it constitutes a field of study, which is distinctive from tourism studies. The response to this draws, first, on the traditional concerns of geography being with space, place, and the environment (Hall and Page, 2002). And, second, it draws not so much on contentious definitions, as on the implications of overnight stays, compared to journeys from home, in terms of the potential to deepen the interactions between the destination and host areas, and their respective populations. There are three main dimensions to this. First, there are

economic implications arising from the expenditures related to the overnight stay (long distance travel, commodified accommodation unless staying with friends and relatives, meals, and other services) that can lead to significant interregional and international income transfers. Some of the spending on hospitality services necessarily is not incurred in home-based leisure and recreation, which will also have significantly less impact on interregional welfare or income transfers. Second, travel – particularly with the greater distances implied by overnight stays – potentially brings tourists and local residents from different cultures into contact with each other, generating a range of social and cultural costs and benefits (Wall and Mathieson, 2006). Third, tourism brings a distinctive dimension to the environmental impacts of human activity, resulting from the scale and intensity of travel, the polarized redistribution of human activity across space, and difficult-to-change attitudes to the associated climate change and sustainability issues – captured by the notion of 'binge flying' (Cohen et al., 2011).

Changing Perspectives in Tourism Geography

The changing theoretical and methodological approaches to tourism geography research have generally mirrored wider shifts in other social sciences and particularly within human geography. In broad terms, this has been characterized by a shift from a largely inductive positivist framework to approaches grounded in political economy, and cultural theories and or postmodernism (Hannam and Knox, 2009). Tourism geography, however, remains methodologically diverse, as is evident in the title and contents of its leading journal, *Tourism Geographies: An International Journal of Tourism Space, Place and the Environment*.

Spatial Models

Positivist approaches have mostly been inductive, but there are some notable deductive contributions. Foremost among these is a relatively neglected early paper by Christaller (1963), who is otherwise best known for his work on central place theory. This article sought to conceptualize tourism flows from 'central' to 'peripheral' locations in Europe in terms of basic principles of accessibility, and – at least indirectly – was the forerunner of an array of models concerned with the spatial interactions between areas of tourist origin and destination. More recently attention has also focused on intradestination movements, mostly empirically but with some theoretical contributions (Lew and McKercher, 2006), assisted by technological innovations that are discussed below.

Pearce (1995) proposed a fourfold typology of such models. Models of tourist travel analyze the alternative routes, which link places, and the volume and character of tourist flows along these; they are often informed by a distance decay assumption. Origin–destination models are more sophisticated, examining flows to and from a multiple set of areas, and identifying spatial hierarchies of tourist destinations. Structural models are concerned with the evolution of distinctive spatial structures in destination areas, reflecting the asymmetrical power relationships between tourists, the tourist industry, and the destination society. These relationships are epitomized by the development of resort enclaves, which are spatially, culturally, and economically weakly embedded in their local milieu. Finally, evolutionary models – reflecting geographers' time–space concerns – have examined the development of tourism spaces, especially resorts.

The most widely known evolutionary model is Butler's (1980) resort life cycle model, which contends that resorts pass through a series of stages: exploration, involvement, development, consolidation, and stagnation, culminating in a number of alternative trajectories, ranging from immediate decline to rejuvenation. The key factors shaping the evolution of individual resorts are visitor numbers, the roles of local/external entrepreneurs, competition, and the pressures on carrying capacity. The model is implicitly linear and deterministic, but does acknowledge the role of human agency in the latter stages when alternative futures are possible, depending on the responses of the tourism industry complex. Although the model is more descriptive than explanatory, its imaginative insights have influenced a generation of subsequent researchers such as Agarwal (2002) who have sought to operationalize the model in a variety of resort settings. It has also spawned a considerable debate as to the shape of the resort development curve, the determinants of change, and the criteria for identifying turning points between the different stages (Butler, 2006).

Positivist approaches to spatial analysis have been refreshed in recent years, at least in terms of methods, notably by developments in GPS tracking, assisted by the application of GIS systems for analytical purposes. Collecting detailed itinerary data for tourists, as opposed to home-destination flows, has always been a challenge for spatial analysis, and has traditionally been addressed through primary data collection, namely the use of trip diaries. Van de Knapp (1999), for example, demonstrated how GIS could be used to analyze the data captured by diaries, and to present these visually, so as to simulate dynamic systems. There were, however, still limitations in the amount of data that tourists were willing, or able, to capture in their diaries. These tended to focus only on the main nodes and some of the principal if generalized flows, and this deficiency was not remedied until the application of GPS systems, whereby the movements of individuals could be identified through the use of tracker devices (Shoval and Isaacson, 2007) using a range of technologies to monitor the spatial and temporal movement of visitors. Shoval (2012) has provided a technologically based link to the early work of Hagerstrand's (1970) concepts of space time geography. The collection of more sophisticated data has also led to new methods of data analysis using so-called sequence alignment methods (Shoval and Isaacson, 2007). These have the advantages of being low cost and precise, although there are significant challenges in managing the enormous volumes of data produced. These advances have also tended to reinforce an essentially empirical orientation in much of the work on spatial models, with theoretical advances tending to lag behind.

The Political Economy of Tourism

Since the 1970s, there has been a more critical approach to tourism geography, which initially at least, drew on the

experiences of the less-developed countries and on the political economy of development. De Kadt's (1979) *Tourism: Passport to Development?* argued for a more critical approach to understanding tourism impacts, and questioned the uncritical view of tourism as a vehicle of economic growth, but failed to offer a coherent alternative conceptual framework. Britton (1991), however, did set out such a framework, drawing on notions of capital accumulation, cultural capital, and the representation of place; the work of David Harvey, among others, was influential in his thinking. This key paper had two main legacies in terms of tourism geography research: the commodification of culture and place, and the particularities of tourism production within capitalist systems. Britton's work did inspire a more critical approach in tourism geography (see the overview in Sharpley and Telfer, 2002), although there has been relatively limited further development of a political economy framework.

One major exception to this has been the work of Ioannides and Debbage (1998), whose book sought to integrate studies of tourism production with economic geography. Their own contribution to this volume analyzed changes in tourism production in terms of a shift from Fordism to postFordism, drawing on regulation theory. Inevitably, they concluded that there are sectoral differences between, for example: airlines and tourism retailing, and those different forms of tourism production coexist at any one time. Moreover, Debbage and Ioannides (2012) have identified the 'critical turns' in tourism geography, which reflects similar shifts in main stream human geography. These cultural shifts engage with political economy to varying degrees. The application of regulation theory was subsequently extended by Shaw and Williams (2004) who demonstrated how the notions of regime of accumulation and mode of regulation can be used to analyze the interaction of production, consumption, and various forms of regulation in changing and diverse market economies. More recently, Bianchi (2009) has reasserted the importance of a political economy perspective in the face of what he sees as the limitations of the recent shift to cultural approaches in tourism geography, namely the failure to analyze the economic and political relations of power, material inequalities, working conditions, and social polarization. Another strand of research that is part of the legacy of the work of Britton and others is a concern with commodification, understood as the differentiation between use and exchange values. This is driven in part by debates around neoliberalism and privatization, although these discourses are still weakly represented in the tourism literature, despite the significant contradictions that exist around the use of public or free goods in tourism, and the scope for commodifying the tourism experience. Tourism geographers have also been concerned with the relationships between commodification and authenticity, exemplified in research on slum tourism (Frenzel and Koens, 2012), as well as the way in which host–guest relationships (Smith, 1977) have been mediated by changes in the nature of tourism.

Cultural Interpretations

Tourism geographers have a long-standing interest in the notion of place, and in how tourism behavior is both shaped by and shapes places. They have drawn extensively on the work of sociologists, such as Cohen (1984), who emphasized that all tourists are seeking some degree of novelty and difference, while at the same time wanting to retain an element of familiarity. The way in which tourists balance these differing needs was used by Cohen to construct a typology of tourists ranging from mass to more individualistic travelers, a theme that has run through much of the subsequent work on tourism and tourism places.

The representation of place is also of central importance in tourist practices. One of the most influential concepts in this respect was the notion of the tourist gaze (Urry, 1990: p. 2), which "presupposes a system of social activities and signs which locate the particular tourist practices." In terms of image, the tourist gaze is constructed through signs and signifiers in the landscape, with tourists being collectors of such signs. Tourist practices do not simply entail the purchase of specific goods and services but their practices involve the consumption of signs, which are constructed around and help to define notions of place. One of the most interesting aspects of Urry's thesis is the contention that, in the late-twentieth century, there was a change in the values held by tourists, and in the signposting of the objects to be gazed upon. The contrast between tourism as a separate sphere of social activity and the practices of every day life has greatly diminished. Tourism has become dedifferentiated or merged with other activities, and this has contributed – at least in the more developed economies – to a shift in interest from traditional seaside resorts to urban/industrial heritage sites, along with valued rural landscapes and cultures.

While Urry's seminal work, *The Tourist Gaze*, continues to have almost iconic importance in tourism, more recent postmodern interpretations have been characterized by broader concerns with how tourism is performed, represented, and experienced. There has been recognition of how tourism is not jut gazed on, but also smelled, heard, tasted, and embodied in a variety of sensory ways (Crouch, 1999). Molz (2007), for example, has explored the importance of eating familiar foods as being an encounter with otherness, emphasizing that culinary tourism is not necessarily about experiencing the other culture, but rather about displaying openness, and performing cosmopolitanism. Another important strand of research has been concerned with how tourism is represented and shaped in various literature, including travel writings (Robinson and Anderson, 2002).

Despite the call for a broader remit, some of the most significant contributions of tourism geographers have been in respect of the visual. Photographs play a particularly important role in the interchanges between tourists and guide books, tourists and local populations, and tourists' representations and narratives about their tourism experiences (Edensor, 1998). While much of the debate in this area has focused on the creation of spaces of exclusion, dependency and disempowerment, there is also evidence that tourist photography creates opportunities for empowerment through self-direction (Scarles, 2012). This also resonates with the emphasis on seeing practices as scripted and as places being made and performed by tourists, who themselves are transformed in this process: the outcomes are not deterministic, and there is room for contestation, struggle, and surprises (Crang, 2004).

Tourism Geography and the Challenges of a Changing World

Tourism geographers make an important contribution to debates about the role of tourism in a changing world. Two themes are of particular note in their work: the environment, including aspects of sustainability and climate change, and the restructuring of places or economies.

In recent decades, there has been considerable research on the environmental effects of tourism, and geographers have played a key role in this. Mathieson and Wall (1982) provided an early and enduring review of the impacts of tourism. Most research, subsequently, has involved a refinement of the measurement of these impacts, through such techniques as carrying capacity, rather than advances in conceptualization per se. Most of the research has been in wilderness or other protected areas and there has been surprisingly less work on mass tourist environments. In the 1990s tourism geography engaged enthusiastically with the notion of sustainability, and a specialist literature developed around sustainable tourism (Weaver, 2006). Tourism geographers have argued for a more holistic approach to sustainability, which recognizes the links between the economic, social, and environmental dimensions. They have also become more critical of idealistic attempts to build partnerships of stakeholders, as evidence emerges of unequal power relationships within these (Bramwell, 2011). The focus in recent years has increasingly shifted to the implications of climate change for tourism, the contribution of tourism to climate change, and how tourists respond to the environmental consequences of their air travel in particular (Becken, 2007; Cohen et al., 2011). Within this context, interest has also focused on 'sites of practice' and the way consumers' attitudes toward sustainable behavior contrast between home and holiday environments. The emphasis placed on practices and context reveal that comfortable notions of environmental responsibility and sustainable consumption in the home often conflict with the discourses of consumption reduction associated with climate change in tourism contexts (Barr et al., 2011).

Tourism geography is also engaging with economic restructuring, and this is particularly evident in urban areas. In the face of major declines in their traditional economies, many urban localities have looked to tourism as one, and often the leading, component in their regeneration strategies. Tourism has the advantages of being labor intensive, requiring relatively limited investment, and having the capacity to refashion place images so as to reposition localities in the wider map of economic opportunities. Tourism geographers have played a role in analyzing the social and spatial impacts of such strategies, while demonstrating that their success is in part contingent on place characteristics. Similar research also exists on the role of tourism in other types of place environments, including the rural and the coastal resort, as well as on mega events such as Olympic games. Such analyses often present tourism as part of the cultural industries, and emphasize its contribution to creativity, making links to both the experience economy and to the skills of the postmodern tourist.

In terms of the economic restructuring of tourism destinations, an area of greater interest concerns the consequences of globalization processes. These processes have been represented in a variety of ways but center on the nature of the connections that operate across national boundaries. Such global flows of capital and knowledge in tourism may be associated with the activities of transnational companies. In addition there are a series of key drivers that shape some resort destinations. These include the power exercised by dominant firms who control flows of knowledge and competencies in the global market. Of particular interest are those global agents of coastal resort developments in the so-called pleasure-periphery (King, 2001), which mainly refers to the tourism economies of less-developed countries of Southeast Asia, Latin America, and the Pacific, although increasingly also includes Africa. These agents wield significant economic power that often produces unequal relationships in the tourism destinations. As yet, there is only limited understanding of the role of resort developers and the associated key sources of finance via global investment companies, with such agencies providing important areas for research in the geographies of tourism.

See also: Conspicuous consumption and distinction, history of; Consumption, Geography of; Dependency Theory; Ecotourism; Environmental Policy: Protection and Regulation; Leisure, Sociology of; Mobility and Travel Activity Patterns; Social Geography; Tourism.

Bibliography

Agarwal, S., 2002. Restructuring seaside tourism: the resort life cycle. Annals of Tourism Research 29, 25–55.

Barr, S., Gilg, A., Shaw, G., 2011. Helping people make better choices: exploring the behaviour change agenda for environmental sustainability. Applied Geography 31, 712–720.

Becken, S., 2007. Tourists' perception of international air travel's impact on the global climate and potential climate change policies. Journal of Sustainable Tourism 15, 351–368.

Bell, M., Ward, G., 2000. Comparing temporary mobility with permanent migration. Tourism Geographies: International Journal of Place, Space and the Environment 2, 87–107.

Bianchi, R.V., 2009. The 'critical turn' in tourism studies: a radical critique. Tourism Geographies: An International Journal of Tourism Space, Place and Environment 11, 484–504.

Bramwell, B., 2011. Governance, the state and sustainable tourism: a political economy approach. Journal of Sustainable Tourism 19, 459–477.

Britton, S., 1991. Tourism, capital and place: towards a critical geography of tourism. Environment and Planning D: Society and Space 9, 451–478.

Butler, R.W., 1980. The concept of a tourist area cycle of evolution: implications for management of resources. Canadian Geographer 24, 5–12.

Butler, R.W., 2006. The Tourism Area Life Cycle, vols. 1 and 2. Channel View Publications, Clevedon.

Christaller, W., 1963. Some considerations of tourism location in Europe: the peripheral regions—underdeveloped countries—recreation areas. Regional Science Association Papers 12, 95–105.

Cohen, E., 1984. The sociology of tourism: approaches, issues and findings. Annual Review of Sociology 10, 373–392.

Cohen, S., Higham, J.E.S., Cavaliere, C.T., 2011. Binge flying: behavioural addiction and climate change. Annals of Tourism Research 38, 1070–1089.

Crang, M., 2004. Cultural geographies of tourism. In: Lew, A.M., Hall, C.M., Williams, A.M. (Eds.), Companion to Tourism. Blackwells, Oxford, pp. 74–84.

Crouch, D., 1999. Leisure/Tourism Geographies. Routledge, London.

Debbage, K., Ioannides, D., 2012. The economy of tourism spaces in a multiplicity of critical turns? In: Wilson, J. (Ed.), The Routledge Handbook of Tourism Geographies. Routledge, Abingdon, pp. 149–156.

Duncan, T., 2012. The "mobilities turn" and the geography of tourism. In: Wilson, J. (Ed.), The Routledge Handbook of Tourism Geographies. Routledge, Abingdon, pp. 113–119.

Edensor, T., 1998. Tourists at the Taj: Performance and Meaning at a Symbolic Site. Routledge, London.
Frenzel, F., Koens, K., 2012. Slum tourism: developments in a young field of interdisciplinary tourism research. Tourism Geographies: An International Journal of Tourism Space, Place and Environment 14 (2), 195–212.
Hagerstrand, T., 1970. What about people in regional science ? Papers, Regional Science Association 24, 7–21.
Hall, C.M., Lew, A. (Eds.), 1998. Sustainable Tourism: A Geographical Perspective. Publisher, Harlow.
Hall, C.M., Page, S.J., 2002. The Geography of Tourism and Recreation: Environment, Place and Space, second ed. Routledge, London.
Hannam, K., Knox, D., 2009. Understanding Tourism: A Critical Introduction. Sage, London.
Ioannides, D., Debbage, K. (Eds.), 1998. The Economic Geography of the Tourist Industry: A Supply Side Analysis. Routledge, London.
Kadt, E.de, 1979. Tourism: Passport to Development? Oxford University Press, Oxford.
King, B., 2001. Resort-based tourism on the pleasure-periphery. In: Harrison, D. (Ed.), Tourism and the Less Developed: Issues and Cases. CABI Books, Wallingford, pp. 175–190.
van der Knaapp, W., 1999. GIS-oriented analysis of tourist time-space patterns, to support sustainable tourism development. Tourism Geographies 1, 56–69.
Lew, A., McKercher, B., 2006. Modelling tourist movements: a local destination analysis. Annals of Tourism Research 33, 403–423.
Mathieson, A., Wall, G., 1982. Tourism: Economic, Physical and Social Impacts. Longman, Harlow.
Molz, J.G., 2007. Eating difference: the cosmopolitan mobilities of culinary tourism. Space and Culture 10, 77–93.
Pearce, D., 1995. Tourism Today: A Geographical Analysis. Longman, Harlow.
Robinson, M., Anderson, H.C. (Eds.), 2002. Literature and Tourism. Thomson, London.
Scarles, C., 2012. The photographed other: interplays of agency in tourist photography in Cusco, Peru. Annals of Tourism Research 39, 928–950.
Sharpley, R., Telfer, D.J., 2002. Tourism and Development: Concepts and Issues. Channel View Publications, Clevedon.
Shaw, Williams, A.M., 2004. Tourism and Tourism Spaces. Sage, London.
Shoval, N., 2012. Time geography and tourism. In: Wilson, J. (Ed.), The Routledge Handbook of Tourism Geographies. Routledge, Abingdon, pp. 174–179.
Shoval, N., Isaacson, M., 2007. Tourist Mobility and Advanced Tracking Technologies. Taylor & Francis, New York.
Smith, V. (Ed.), 1977. Hosts and Guest: The Anthropology of Tourism. University of Pennsylvania Press, Philadelphia.
Urry, J., 1990. The Tourist Gaze. Sage, London.
Urry, J., 2007. Mobilities. Polity Press, Cambridge.
Wall, G., Mathieson, A., 2006. Tourism: Change, Impacts and Opportunities. Pearson, Harlow.
Weaver, D.B., 2006. Sustainable Tourism. Elsevier, Oxford.
Williams, A.M., Hall, M., 2002. Tourism, migration, circulation and mobility: the contingencies of time and place. In: Hall, C.M., Williams, A.M. (Eds.), Tourism and Migration: New Relationships between Production and Consumption. Kluwer, Dordrecht, pp. 1–52.
Wilson, J. (Ed.), 2012. The Routledge Handbook of Tourism Geographies. Routledge, Abingdon.

Tourism, the Millennium Development Goals, and Gender Equality

Lucy Ferguson, Instituto Complutense de Estudios Internacionales, Universidad Complutense de Madrid, Madrid, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

While institutions representing the tourism industry are keen to claim that tourism contributes to the Millennium Development Goals (MDGs), there is little evidence to suggest that this is the case. Indeed, in many ways tourism perpetuates underdevelopment and the violation of human rights. This article provides a detailed analysis of the relationship between tourism and the MDGs, with a focus on MDG3: promoting gender equality and empowering women. Tourism offers the potential to contribute to development goals including women's empowerment, but a significant restructuring of the sector would have to take place in order to achieve such goals.

Introduction

Tourism accounts for 5% of the world's Gross Domestic Product (GDP), 6–7% of global employment, and 30% of the world's export of commercial services. Tourism is the main source of foreign exchange for one-third of developing countries and half of the least developed countries. In 25 out of 49 least developed countries, tourism is a strategic priority in national development plans (UNWTO, 2011: p. 1). This article critically analyzes the role of the tourism sector in global development, using the original Millennium Development Goals (MDGs) as a framework for the analysis. While the MDGs are set to expire in 2015, they nevertheless remain the primary reference point for global development milestones at the time of writing. This article begins with an overview of debates on the relationship between tourism and development, setting out two contrasting positions. It then goes on to outline the policy context for tourism and the MDGs, focusing in particular on the United Nations World Tourism Organization (UNWTO), a UN specialized agency and the main international institution focusing on the tourism sector.

This article provides an overview of which goals are included and excluded from the current tourism agenda. It argues that three out of eight goals are to some extent included, while the remaining are predominantly excluded from debates. Although several MDGs are included in the tourism policy agenda, the extent to which the tourism sector has the capacity to address these issues is much in question. In order to explore this in more detail, this article goes on to discuss one particular goal – MDG3: promote gender equality and empower women. This detailed analysis argues that the ways in which gender issues are currently framed in tourism policy and practice limit the potential of the sector to contribute to this goal. The concluding comments offer reflections on the potential of tourism to contribute to the MDGs and, more broadly, the post-2015 development agenda.

Approaches to Tourism and Development

Claims about the link between tourism and global development range from the most optimistic – tourism is a force for good, poverty reduction and greater opportunities – to the most pessimistic – that tourism is destructive of the natural environment and promotes primarily exploitative relations between employers and employees as well as between 'hosts' and 'guests.' Early critical work on tourism and development focused on how the tourism industry often exploited colonial associations and turned to colonial power structures to promote and construct the industry in developing countries (Perez, 1974, 1975; Britton, 1982; Nash, 1989). In response, tourism studies turned its attention to 'alternative' forms of tourism, suggesting that these were more likely to overcome the exploitative dimensions of mass tourism in developing countries (Lea, 1993; Brohman, 1996; Khan, 1997).

In contemporary tourism, responsible tourism is an umbrella term for a range of approaches from both academia and civil society. Responsible tourism aims to increase the positive impact of tourism and reduce the negative consequences of tourism development. A particular strand of responsible tourism – 'propoor tourism' – has developed tools for aiming to improve livelihoods and increase well-being in tourism destinations. This approach targets the benefits of tourism more directly toward poverty reduction (Brown and Hall, 2008). Propoor tourism and responsible tourism, more broadly, are premised on the fundamental understanding that a more equal sharing of the benefits generated by tourism will lead to improvement for those living in tourism communities. This confidence in the potential of tourism to contribute to development is reflected in the discourse and policy of institutions such as UNWTO, where tourism is framed as an overwhelming force for social and economic progress, as discussed in more detail below.

In contrast, more radical critiques of tourism development take a fundamentally different standpoint. They argue that tourism tends to be primarily destructive, displacing communities and violating human rights. The campaigning organization *Tourism Concern*, for example, argues that countless numbers of poor and vulnerable people worldwide 'have their basic human rights violated as a direct result of tourism's growth, exacerbating their poverty and trapping them in a cycle of deprivation' (Tourism Concern, 2009). Research into the distribution of benefits from alternative forms of tourism suggests that there is no automatic connection between 'new tourism' and human development in tourism destinations (Mowforth and Munt, 2008). Further critiques come from

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64080-1

those who see tourism as an essentially exploitative industry from both an environmental and social perspective. Rosaleen Duffy, for example, has extensively criticized the environmental impact of tourism on local ecosystems (Duffy, 2004). Such research has also demonstrated the ways in which tourism promotes the sexual exploitation of women and children (Pettman, 1997; Jeffreys, 1999). Other studies detail the exploitative working practices and poor labor conditions often involved in the tourism industry (ILO, 2001).

Two distinct approaches to tourism and development can therefore be identified. The first conceptualizes tourism as a vehicle for global development with the potential to contribute to the MDGs and other development objectives. While such an approach acknowledges the weaknesses and challenges involved, it is argued that these can be overcome with small changes in policy and practice. In contrast, in the more critical framing tourism is not considered so much a tool for development as a potential threat to the livelihoods and rights of people in developing countries. In this understanding, tourism is premised on the inequalities embedded in global development, and as an industry draws on these inequalities. Tourism and the MDGs are therefore perceived to be fundamentally incompatible. This article aims to draw together insights from both sets of literature in order to critically analyze the relationship between tourism and the MDGs. In order to do this, we now turn to an overview of how the MDGs feature in tourism policy and practice.

The MDGs in Tourism Policy and Practice

The tourism sector is best understood as a collection of industries operating at a number of different levels – public, private, and community – from national tourism boards to multinational hotel chains to small-scale local tourism enterprises. At the same time, there are a wide range of forms of tourism – mass tourism, ecotourism, adventure tourism, volunteer tourism, etc. As such, it is a challenging sector to analyze from the perspective of international development. While national tourism administrations are responsible for all aspects of the tourism sector in their country, nevertheless, the power and influence of multinational travel and tourism companies often means that such bodies have little control over tourism planning and processes. The main international institution with responsibility for tourism policy is UNWTO, a UN specialized agency based in Madrid. UNWTO is responsible for the coordination of tourism policy worldwide and 'provides leadership and support to the tourism sector in the advancement of sustainable policies and practices' (UNWTO, 2011: p. 1).

UNWTO is a good starting point for analyzing the relationship between tourism and the MDGs, as it acts as a reference point for all levels of the tourism sector – public, private, and community. Development is enshrined in the organization's mission statement, which asserts that "UNWTO is committed to the United Nations Millennium Development Goals, geared toward reducing poverty and fostering sustainable development through tourism" (UNWTO, 2011: p. 1). In the 2010 publication, entitled *Tourism and the Millennium Goals: Sustainable, Competitive, Responsible*, UNWTO outlines in detail the ways in which the tourism sector can contribute to the MDGs. These points are set out within the framework of Corporate Social Responsibility (CSR) and the UN Global Compact.

The first MDG to consider here is MDG1 – eradicate extreme poverty and hunger. A key claim made about tourism's contribution to development is the creation of direct and indirect employment (Sharpley and Telfer, 2002). In practice, most aspects of tourism work are dominated by informality, through high staff turnover, long working hours, subcontracting, 'flexible' working conditions, the prevalence of 'casual workers,' and seasonal variations in employment (ILO, 2001: 56–63). Multinational hotel chains have been criticized as setting a precedent for flexibilized, low-skill labor with little room for mobility and promotion of staff (ILO, 2001). However, in many ways this is also true of smaller organizations such as 'boutique' hotels employing between 20 and 100 staff. Likewise in small, 'family-run' enterprises those employees who make up the main body of the workforce do not tend to see much social mobility in their jobs (Sinclair, 1997a). As such, the claim that tourism can contribute to poverty reduction through employment creation should be tempered with a discussion of the nature and conditions of such employment and interrogated through a lens of decent work.

In terms of extreme poverty – or the 'poorest of the poor' – it is unclear that what impact tourism has in this area. Initiatives such as UNWTO's Sustainable Tourism – Eliminating Poverty (ST-EP) Program fund a total of 49 projects worldwide. This program promotes poverty elimination through the provision of assistance to sustainable tourism development projects targeted toward people living on less than a dollar a day. In terms of CSR, UNWTO suggests that companies consider the following actions to achieve MDG1: 'Local recruitment, local enterprise investment, local supply purchase, fair-trade, poverty awareness campaigns, in-kind funds, staff participation in humanitarian projects' (UNWTO, 2011: p. 12). As set out above, the 'propoor tourism' framework aims to ensure that the economic benefits of the tourism industry contribute to poverty reduction in local communities (see, e.g., Ashley and Roe, 2002). However, the concept of propoor tourism has not been debated in any substantive way in the tourism literature, and critics have argued that this approach does little to alleviate inequalities in local communities, therefore failing to alleviate extreme poverty for the poorest of the poor (Hall, 2007; Harrison, 2008). As such, it is difficult to substantiate the claim that tourism contributes to eradicating extreme poverty and hunger.

The second goal with most relevance to contemporary tourism policy and practice is MDG7 – ensure environmental stability. A wealth of literature debates the relationship between tourism and environmental protection/degradation. As outlined above, extensive research has questioned the capacity of tourism to ensure environmental stability. Recent research, for example, suggests that tourism is one of the least prepared industries for addressing the risks of climate change (Scott, 2011). Moreover, the paradigm of sustainable tourism has been subjected to intense scrutiny by academics and activists (see, e.g., Duffy, 2002). Nevertheless, UNWTO continues to promote tourism as a means of achieving

environmental protection. For the Rio + 20 event in June 2012, UNWTO organized a session entitled 'Tourism for a Sustainable Future,' outlining the relationship between tourism and environmental sustainability. Here it was suggested that tourism can be a driver of the green economy, and that sustainable development and tourism can be complementary if the right mechanisms are put in place. Within the framework of CSR, UNWTO suggests a number of actions tourism companies can take to promote environmental stability, such as 'compliance to advanced norms and standards (e.g., ISO 14001, Green Globe Certificate), resource efficiency, renewable energies, recycling and ecomaterials, wildlife, and ecosystem preservation actions (i.e., offsetting, monitoring, cleanups, etc.)' (See http://www.uncsd2012.org/content/documents/448unwto_cln.pdf. http://www.uncsd2012.org/index.php?page=view&nr=108&type=1000&menu=126 for more on this event). However, research has shown the extent of 'greenwashing' in the hotel sector, where tourism companies have recognized the value of the green consumer market. In response, promoting 'ecotourism' has become a way of gaining a competitive edge without any fundamental changes being made in business practices (Connell and Page, 2008). As such, any claim that tourism can contribute to MDG7 needs to be analyzed in the context of contemporary debates on climate change, corporate sustainability, and the dangers of greenwashing.

Little attention has been paid to the relationship between tourism and MDGs 2, 4, 5, 6, and 8 – both from a research and policy perspective. Indeed, it is difficult to find any reference to these issues in tourism policy and practice. As such, the only source of information on tourism and the other goals comes from UNWTO's 2010 MDGs publication. An overview of the suggestions made for each goal in this publication is provided here in order to understand how these development issues are framed in tourism discourse:

- *MDG2*: achieve universal primary education – Funding for schools, offering teachers grants, delivering school supplies and computers, in-company technical and language training, scholarships, and mentorship programs for staff and local community
- *MDG4*: Reduce child mortality – Social investment in health facilities, free medical checkups in company and community, nutrition advice and programs, free dispatch of medical equipment and staff
- *MDG5*: Improve maternal health – Free screenings and medical care, free supply of maternity-related complements, education programs, investment in research and development, funding of health facilities, child care facilities in company
- *MDG6*: Combat HIV/AIDS, malaria, and other diseases – In-company prevention training and awareness raising activities, host community education workshops (i.e., use of mosquito nets), free medical checkups, social inclusion programs
- *MDG8*: Develop a global partnership for development – Collaboration between public and private sectors, independent associations and NGOs to provide funding and assistance to host communities, enhance staff capacities, clients' well-being, and engage consumers (Adapted from UNWTO, 2010).

This summary of UNWTO's approach to the MDGs demonstrates that such issues are considered appropriate to be tackled by the tourism private sector, but not by the public sector or UNWTO. This demonstrates the limits of the CSR framework for tackling development issues. It is clear that these activities are unlikely to address the structural issues underlying underdevelopment, and their potential to contribute to the achievement of the MDGs is questionable at best.

This brief overview demonstrates that some MDGs are taken more seriously and others in tourism policy and practice. While MDGs 1 and 7 – and 3, as discussed below – are addressed to some extent, those addressing more complex development challenges are not effectively dealt with. This raises a number of questions about the capacity of tourism to contribute to the MDGs.

Tourism and MDG3: Promote Gender Equality and Empower Women

In order to explore the issues identified above in more detail, we now go on to analyze more extensively the relationship between tourism and MDG3 – promote gender equality and empower women (This analysis draws on an article in *Current Issues in Tourism* dealing with these questions (Ferguson, 2011)). In terms of policy, UNWTO has been markedly slow on the uptake with gender equality and women's empowerment compared with other UN institutions. Gender as a concept was not widely visible in the organization until 2007, when World Tourism Day was celebrated with the theme 'Tourism Opens Doors for Women.' As part of this event, a round table involving a broad range of participants was held with the express purpose of exploring the relationship between this theme and the MDGs, setting up a work agenda for the future. Following on from this, a further UNWTO Forum in Berlin in March 2008 explored the theme of Women in Tourism, leading to a Memorandum of Understanding with the United Nations Development Fund for Women, now UN Women.

Since these initial steps, a number of milestones have been reached in UNWTO's approach to gender equality and women's empowerment. In 2011, the first *Global Report on Women in Tourism* was published, detailing women's status and participation in the tourism industry in developing regions (UNWTO/UN WOMEN, 2011). A second Memorandum of Understanding was signed with UN WOMEN in 2011, committing both institutions to work together to:

Build on the findings of the Global Report to generate momentum for tourism's active contribution to equality and women's empowerment, as well as to develop a gender mainstreaming strategy for the public and private members of the tourism sector (http://www.unwomen.org/2011/11/unwto-and-un-women-commit-further-to-an-ongoing-collaboration/).

In addition, the 2012 meeting of the UNWTO Commission for Africa – the annual meeting of the ministers of tourism of African countries – was held under the theme *Responsible Tourism: Opportunities for Women and Youth*. This

seminar examined the ways in which tourism can be used to promote gender equality and women's empowerment in African tourism (http://africa.unwto.org/en/event/regional-seminar-responsible-tourism-opportunities-women-and-youth). Ongoing activities at UNWTO include the development of a Women in Tourism Empowerment Program and a gender mainstreaming strategy for national tourism administrations. In terms of the private sector and CSR framework, UNWTO's MDG publication suggests that companies support gender equality and women's empowerment through:

Scholarship programs for managerial coaching, professional mobility workshops, family support structures for working mothers (i.e., in-company child care programs, flexible hours for working mothers), technical training and craft learning for women local communities, etc. (UNWTO, 2011: p. 12).

In spite of these emerging initiatives on women's empowerment, UNWTO has not yet taken a global leadership stance on gender equality issues. Within the institution, gender currently falls within the remit of Ethics and Social Dimensions of Tourism Program, where it is just one of several areas the small team are concerned with, along with migration, human rights, nondiscrimination, and the protection of children. As such, there are limited resources with which to develop and promote a strong gender agenda. In addition, gender is largely compartmentalized into one department, leaving these issues outside the concern of the senior management and other departments. If the institution is serious about gender, sufficient resources and expertise will need to be committed in order for these policy initiatives to be effective and influential, as well as broad-based support across the senior management team.

In addition to UNWTO, some bilateral donors have demonstrated a commitment to promoting gender equality and women's empowerment through tourism. The Japan International Cooperation Agency funds a number of projects, which integrate gender and tourism development, along with other aspects such as reproductive health. Also in Honduras, Norwegian funding and expertise was directed toward a feasibility study on the potential for tourism to contribute to gender equality (World Bank, 2003). In 2010, The Spanish International Cooperation and Development Agency approved two projects directed toward gender equality and tourism development in Mali and Namibia in collaboration with the UNWTO ST-EP program. The World Bank has funded several projects with links between tourism and gender equality or women's empowerment – in countries such as Mozambique, Bolivia, Ethiopia, and Honduras. The 'Regional Development in the Copán Valley' project in Honduras included 'gender' as a key component from the outset (see Ferguson, 2010a). In each of these projects, tourism is framed as an opportunity to generate the conditions for 'marginal groups' (women, young people, and old people) to *access the opportunities* afforded by tourism development.

With the exception of the Honduras World Bank project, no research has been carried out on these tourism and gender projects, so they cannot be analyzed here. This is an issue in itself, as there is currently limited funding for gathering evidence on the impact of tourism projects on gender equality and women's empowerment. Moreover, such examples are sporadic at best, suggesting that the links between gender and tourism have not been explicitly made in bilateral development institutions. This again demonstrates a missed opportunity by UNWTO to generate a global dialog and partnership on tourism and gender. The critique of the relationship between tourism and MDG3 now proceeds in two ways. First, we draw on extensive research to analyze the ways in which tourism impacts on gender equality and women's empowerment in practice. Second, we explore the ways in which gender equality and women's empowerment are framed in the policies and projects set out above, and interrogate their potential to contribute to MDG3.

The Impact of Tourism on Gender Equality and Women's Empowerment

In spite of the assertions of UNWTO and UN Women (UNWTO/UN WOMEN, 2011), there is a wealth of evidence that demonstrates the ways in which tourism reinforces gender inequality and impedes women's empowerment. There is a clear segmentation of men's and women's work in tourism, whereby the majority of women's work is concentrated in seasonal, part time, and low paid activities, such as retail, hospitality, and cleaning (Sinclair, 1997; Chant, 1997). Employment conditions in both global hotel chains and smaller, locally owned enterprises have been extensively criticized in the tourism studies literature. However, these conditions become even starker when looked at through the lens of gender. The tourism industry is characterized by 'the seasonal nature of many of its activities and (by) important fluctuations even in normal periods.' These features of the industry create a 'generic tendency to operate on the basis of a core staff and to employ the labor needed for day-to-day operations under atypical contractual arrangements' (ILO, 2001: p. 48). This involves maintaining a large pool of temporary labor to be drawn upon in times of high demand, made up of predominantly young and/or female workers. The International Labour Office reports that women account for 46% of workers in wage employment in tourism globally. However, expanding the definition to include catering and accommodation brings the proportion of female labor up to 90%.

M. Thea Sinclair (1997) argues that work in tourism should be understood as a reflection of wider inequalities in the tourism industry, as the majority of women's work is concentrated in seasonal, part time, and low paid activities, such as retail, hospitality, and cleaning (Sinclair, 1997). Sylvia Chant (1997) offers a useful summary of the gendered nature tourism work:

> Female recruitment in formal sector enterprises catering for international tourists tends to draw heavily on male-constructed and male-biased gender stereotypes and to place women in occupations, which in many respects crystallize and intensify their subordinate positions in society, whether through their assignation to low level, behind-the-scenes domestic work as laundrywomen or chambermaids, or to jobs where their physical attributes are used to attract men or to gratify male sexual needs, as in front-line hotel, commercial and restaurant posts, and in entertainment establishments. (Chant, 1997: p. 161)

More recent research exploring the informal dimensions of tourism work, such as handicraft production for tourist consumption reveals similarly gendered dynamics. Hazel Tucker, for example, in her research on Turkey, argues that small and microbusiness development was not as beneficial for the women of Göreme (as it was for the men) (Tucker, 2007: p. 101). In many ways, therefore, the evidence would suggest that tourism employment has a detrimental effect on gender equality and women's empowerment, and indeed may exacerbate inequalities.

While tourism work tends to be highly segregated by gender, there are nevertheless cases in which tourism has challenged gender inequality in destination countries. As Chant (1997) has argued, despite the structural inequalities of women's participation in tourism production, such work has some benefits for the empowerment of women workers. Sinclair (1997a) similarly points to some of the increased benefits of work in tourism for women – for example, she argues that such work can lead to greater status in the household and consequently increased bargaining power in the household context. Tucker's longitudinal research in Turkey shows how women can craft out space within tourism and that this can in the long-term contribute to the 'undoing of shame' (Tucker, 2007: p. 101). In Belize, Gentry found that the alternative tourism sector 'is notable for allowing women to use stereotypes regarding their work to their advantage' (Gentry, 2007: p. 491). In research conducted in Central America, many women workers in the tourism industry discussed how they felt working in tourism contributed to their personal and economic empowerment. They suggested that male control over household income had been diminished by the many opportunities for women to earn money in tourism communities. Others enjoyed the interaction with people from around the world, in particular being exposed to different gender norms, such as women traveling alone or men taking primary responsibility for child care (Ferguson, 2010a).

To date, however, these limited benefits of tourism have tended not to be accompanied by wider influence in society. As argued below, there is no automatic correlation between women's economic empowerment through income-generating activities in the tourism industry and broader political and social empowerment. One of the main reasons for this is the lack of attention paid to women's unpaid labor in the domestic sphere, particularly as contributing family workers in tourism enterprises (Ferguson, 2013). Research on the impact of tourism in Central America shows how child care and parenting are being neglected in tourism communities due to the demands of employment in the tourism industry, increasing women's burden for social reproduction without any decrease in their assumed responsibility for such work (Ferguson, 2010b). As Darcie Vandegrift argues, "the benefits of more highly rewarded feminized work are dampened by the limits of a tourism economy that does not provide an infrastructure to support working mothers" (Vandegrift, 2008: p. 794). These findings are echoed in the majority of research on gender and tourism in contexts as diverse as China (Morais et al., 2005), Belize (Gentry, 2007), and Turkey (Tucker, 2007). These issues of women's unpaid work in tourism need to be addressed as a matter of priority in tourism policy. As the above analysis suggests, the relationship between tourism and MDG3 is far from straight forward.

Framing MDG3 in Tourism Policy and Discourse

While tourism work is highly stratified by gender due to the kinds of labor it requires and the ways in which such labor is to be performed, to some extent it can be argued that tourism contributes to economic and personal empowerment for women. In order to explore this tension in more detail I explore how gender is framed in tourism policy and practice, arguing that three key issues need to be addressed. First, a serious discussion of empowerment is required in tourism policy and discourse. Contemporary analyses have highlighted the problems of narrowly framing women's empowerment as economic empowerment, in line with neoliberal development policies (Kabeer, 2005; Keating et al., 2010). The problem for some feminists with the overarching interest in women's economic empowerment is that it is characterized by a lack of any discussion concerning the links between economic growth, development and macroeconomic policy changes, and the perpetuation of gender inequality (Roberts and Soederberg, 2012). As argued above, attempts to promote women's empowerment through income-generating activities in the tourism industry have not usually lead to broader social and political empowerment. The primary understanding of gender policy within development projects is that it relates to women in their roles as producers of tourism goods and services. In general, support and assistance for these productive activities is what constitutes the gender component of policy. As such, issues of gender equality are left out of the picture as equality of opportunity and participation are the primary policy objectives (Ferguson, 2011). Tourism policy would benefit from a more open and honest discussion of what kind of 'empowerment' is desirable and feasible. As the research set out above suggests, economic empowerment through tourism work does not automatically translate into a meaningful redress of power relations beyond a relative improvement in economic conditions.

The second issue for the framing of MDG3 in tourism policy and practice is the issue of unpaid work.

As set out above, a large body of evidence on social reproduction in tourism development communities demonstrates the ways in which the lack of child care provision and other social services is a fundamental stumbling block to promoting women's empowerment and gender equality through tourism. Policy makers would need to engage with debates around social reproduction in a serious and committed way in order to devise practical and sustainable solutions to this issue. This might include new ways of bringing in the private sector in order to encourage tourism enterprises to take more responsibility for the welfare of their employees, for example, by providing free or low-cost child care. Finally, it is important to acknowledge the limitations of presenting discussions on gender equality and women's empowerment only in a 'business case' framework. Extensive contemporary feminist research has argued that the attempt to combine a 'gender-friendly' approach to

poverty reduction and development while maintaining a commitment to neoliberal structural adjustment is ultimately limiting, and is unlikely to lead to gender equality outcomes (Bergeron, 2001; Perrons, 2005; Griffin, 2006). In order to be able to contribute to MDG3 in any substantive way, tourism policy and practice need to move beyond a CSR or 'business case' framework. A more open debate would allow the tourism sector to move beyond narrow, market-based conceptualizations and to present more creative and innovative ways of achieving gender equality and women's empowerment.

Conclusions

While the MDGs are taken into consideration in tourism policy and practice to some extent, it is primarily in terms of a CSR framework in which the private sector is expected to take the lead in ensuring that tourism makes a contribution to the MDGs. In terms of the largest tourism institution – UNWTO – certain MDGs are incorporated into policy and operations. MDG1 and MDG7 are dealt with by the ST-EP Program and the Sustainable Development Program respectively, and to a lesser extent MDG3 is tackle by the Ethics and Social Dimension Program. In practice, however, there are a number of tensions over the ways in which tourism interacts with global development issues such as extreme poverty and environmental sustainability, with many critics questioning the potential of tourism to address these challenges. In order to explore these contradictions, this article provided an in-depth analysis of MDG3, highlighting some of the problems involved in considering tourism as a vehicle for the MDGs.

Some evidence suggests that tourism has the potential to contribute to MDG3, but policy makers need to take a broad range of issues into account if they are serious about promoting gender equality and women's empowerment through tourism development. This article argues that much more needs to be done before a solid link between tourism and the MDGs can be established. It is also important to bear in mind and critically examine the ways in which tourism is harmful to global development, in environmental, social, and economic terms. These points have been elaborated here in the analysis of MDG3. While some potential benefits to gender equality and women's empowerment from tourism can be identified, the current policy framework and operations of the tourism sector are overwhelmingly in favor of reinforcing gender inequality and preventing more lasting forms of women's empowerment. The year 2015 offers a milestone for reframing commitments to global development. In order to meet their obligations, the tourism public and private sectors need to find ways to engage more seriously with international commitments, or will find themselves increasingly isolated from human progress and global development.

See also: Boserup, Ester (1910–99); Child Labor; Children, Rights of: Cultural Concerns; Climate Change and Global Poverty; Consumption, History of; Empowerment Evaluation; Female Power at Work; Gender History; Gender and Feminist Studies in Geography; Gender and Language: Cultural Concerns; Hochschild, Arlie (1940–); Human Rights, History of; Innovation and Employment; Poverty in History.

Bibliography

Ashley, C., Roe, D., 2002. Making tourism work for the poor: strategies and challenges in southern Africa. Development Southern Africa 19 (1), 61–82.
Bergeron, S., 2001. Political economy discourses of globalization and feminist politics. Signs 26 (4), 983–1006.
Britton, S.G., 1982. The political economy of tourism in the third world. Annals of Tourism Research 9, 331–358.
Brohman, J., 1996. New directions in tourism for third world development. Annals of Tourism Research 23 (1), 48–70.
Brown, F., Hall, D., 2008. Tourism and development in the global south: the issues. Third World Quarterly 29 (5), 839–849.
Chant, S., 1997. Gender and tourism employment in Mexico and the Philippines. In: Sinclair, M.T. (Ed.), Gender, Work and Tourism. Routledge, London.
Connell, J., Page, S.J. (Eds.), 2008. Critical Concepts in Sustainable Tourism. Routledge, London.
Duffy, R., 2002. A Trip Too Far: Ecotourism, Politics and Exploitation. Earthscan, London.
Duffy, R., 2004. Ecotourists on the beach. In: Sheller, M., Urry, J. (Eds.), Tourism Mobilities: Places to Play, Places in Play. Routledge, London.
Ferguson, L., 2010a. Interrogating 'gender' in development policy and practice: the World Bank, tourism and microenterprise in Honduras. International Feminist Journal of Politics 12 (1), 3–24.
Ferguson, L., 2010b. Tourism development and the restructuring of social reproduction in Central America. Review of International Political Economy 17 (5).
Ferguson, L., April 2011. Promoting gender equality and empowering women? Tourism and the third Millennium Development Goal. Current Issues in Tourism 14 (3), 235–249.
Ferguson, L., March 2013. Social reproduction and 'contributing family workers' in global political economy: the case of tourism'. paper presented at. In: European Conference on Politics and Gender, Barcelona.
Gentry, K.M., 2007. Belizean women and tourism work: opportunity or impediment? Annals of Tourism Research 34 (2), 477–496.
Griffin, P., 2006. The World Bank. New Political Economy 11 (4), 571–581.
Hall, C.M., 2007. Pro-poor tourism: do 'tourism exchanges benefit primarily the countries of the South'? Current Issues in Tourism 10 (2/3), 111–118.
Harrison, D., 2008. Pro-poor tourism: a critique. Third World Quarterly 29 (5), 851–868.
ILO, 2001. 'Human Resources Development, Employment and Globalization in the Hotel, Catering and Tourism Sector'. Report for discussion at the Tripartite Meeting on the Human Resources Development, Employment and Globalization in the Hotel, Catering and Tourism Sector. International Labour Office, Geneva.
Jeffreys, S., 1999. Globalizing sexual exploitation: sex tourism and the traffic in women. Leisure Studies 18 (3), 179–196.
Kabeer, N., 2005. Gender equality and women's empowerment: a critical analysis of the third millennium development goal. Gender & Development 13 (1), 13–24.
Keating, C., Rasmussen, C., Rishi, P., 2010. The rationality of empowerment: microcredit, accumulation by dispossession, and the gendered economy. Signs 36 (1), 153–176.
Khan, M.M., 1997. Tourism development and dependency theory: mass tourism vs ecotourism. Annals of Tourism Research 24 (4), 988–991.
Lea, J.P., 1993. Tourism development ethics in the third world. Annals of Tourism Research 20, 701–715.
Morais, D.B., Yarnal, C., Dong, E., Dowler, L., 2005. The impact of ethnic tourism on gender roles: a comparison between the Bai and the Mosuo of Yunnan province, PRC. Asia Pacific Journal of Tourism Research 10 (4), 361–367.
Mowforth, M., Munt, I., 2008. Tourism and Sustainability: Development and New Tourism in the Third World. Routledge, London.
Nash, D., 1989. Tourism as a form of imperialism. In: Smith, V.L. (Ed.), Hosts and Guests: The Anthropology of Tourism. University of Pennsylvania Press, Philadelphia.
Perez, L., 1974. Aspects of underdevelopment in the West Indies. Science and Society 37, 473–480.
Perez, L., 1975. Tourism in the West Indies. Journal of Communications 25, 136–143.

Perrons, D., 2005. Gender mainstreaming and gender equality in the new (market) economy: an analysis of contradictions. Social Politics 12 (3), 389–411.
Pettman, J.J., 1997. Body politics: international sex tourism. Third World Quarterly 18 (1), 93–108.
Roberts, A., Soederberg, S., 2012. Gender equality as *smart economics*? A critique of the 2012 world development report. Third World Quarterly 33 (5), 949–968.
Scott, D., 2011. Why sustainable tourism must address climate change. Journal of Sustainable Tourism 19 (1), 17–34.
Sharpley, R., Telfer, D.J., 2002. Tourism and Development: Concepts and Issues. Channel View Publications, Clevedon.
Sinclair, M.T., 1997. Issues and theories of gender and work in tourism. In: Sinclair, M.T. (Ed.), Gender, Work and Tourism. Routledge, London.
Sinclair, M.T., 1997a. Gendered work in tourism: comparative perspectives. In: Sinclair (Ed.), Gender, Work and Tourism. Routledge, London.
Tourism Concern, 2009. Putting Tourism to Rights: A challenge to human rights abuses in the tourism industry. Available at http://www.tourismconcern.org.uk/uploads/file/campaigns/LowRes_Putting%20Tourism%20to%20Rights_A%20report%20by%20TourismConcern2.pdf.
Tucker, H., 2007. 'Undoing shame: tourism and women's work in Turkey'. Journal of Tourism and Cultural Change 5 (2), 87–105.
UNWTO, 2005. Tourism's Potential as a Sustainable Development Strategy. UNWTO, Madrid.
UNWTO, 2011. Tourism for Development: UNWTO Submission to the OECD-dac. UNWTO, Madrid.
UNWTO/UN WOMEN, 2011. Global Report on Women in Tourism (UNWTO: Madrid/UN Women: New York).
Vandegrift, D., 2008. "This isn't paradise I work here": global restructuring, the tourism industry and women workers in Caribbean Costa Rica. Gender and Society 22, 778–798.
World Bank, 2003. Project Appraisal Document: Regional Development in the Copán Valley. Report No. 25798. World Bank, Washington, DC.

Relevant Websites

tourism.org – Equality in Tourism Equality.
tourismconcern.org.uk – Tourism Concern.
unwomen.org – UN Women.
unwto.org – UNWTO.

Trade and Exchange, Archaeology of

Glenn R Summerhayes, University of Otago, Dunedin, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Studies into the archaeology of trade and exchange have surged over the last few years, mostly as a result of technological innovations providing easier access to the appropriate technology. This has unfortunately not kept pace with advances in theoretical developments into the role trade and exchange plays in understanding past societies. This article will provide an overview into the prominent role that trade and exchange has played for archaeologists, and where that role may proceed in the future.

The term 'trade' can invoke images of professionals making profits in marketplaces, while 'exchange' implies the reciprocity of goods, such as food, between nonkin, or the movement of valuables within a social group itself. In this article trade and exchange will be treated equally and defined as the movement of materials or goods between individuals, social groups, or organizations. Ancient trade and exchange are often associated with rare and exotic valuable materials known to have restricted distributions, such as lapis lazuli, ostrich egg shell (Africa), amber, spondylus shell, jade, obsidian, copper, and other precious material items. Yet most past trade and exchange would have involved materials or goods used for daily existence, such as food. These, however, would be rarely seen in the archaeological record, being perishable. Thus the archaeology of trade and exchange deals mostly in those resilient artifacts hardy enough to be left in the archaeological record, such as obsidian, stone, pottery, or jade. Even then, these goods are residues "resulting from a succession of repeated exchange transactions" (Kirch, 1997: p. 228). How and why an archaeologist identifies prehistoric trade and exchange is the subject of this article.

The Development of Trade and Exchange in Archaeology

Since the 1980s there has been an upsurge in studies on the archaeology of trade and exchange. The reasons for this are twofold. First, advances in science and technology have allowed for, and made easily available, the sourcing and characterization of many materials. This provides evidence for the physical identification of their movement. Second, archaeologists began to see the usefulness in using trade and exchange in modeling the distribution of materials over space, and in explaining changes in societies over time.

Technological Developments

Developments in technology since the 1960s have led to an upsurge in chemical 'sourcing' or characterization studies. Yet the identification of traded archaeological material is not a recent phenomenon. Perhaps the most famous early study was by Anna Shepard, who in the 1930s and 1940s postulated the exchange of pottery over wide areas of the Southwest USA by analyzing mineral inclusions in thin sections to pinpoint the origins of the minerals (Shepard, 1965). Chemical techniques to identify the elemental composition of archaeological materials as a means of characterization are also not new. Neutron activation analysis (NAA) was applied on coins from the Louvre in 1952, while in 1957 oriental ceramics were analyzed using nondestructive methods of X-ray fluorescence spectrometry (XRF) and X-ray diffraction (XRD) analysis (Young and Whitmore, 1957). However, access to these techniques was restricted and costly.

A major advance in the chemical analysis of archaeological material came from the initiative of Dr R. Oppenheimer. On 8 March 1956, he assembled a group of archaeologists and chemists at the Institute of Advanced Studies, Princeton, to discuss the possibility of applying methods of nuclear research to the study of archaeology (Sayre and Dodson, 1957). As a result of this meeting, work was undertaken at two laboratories – the Brookhaven National Laboratory in the USA and the Research Laboratory for Archaeology and the History of Art at Oxford, UK. Techniques used included NAA and spectrographic methods. These early studies were reasonably successful, being able to separate pottery wares from Asia Minor, Greece, and Italy, and different factories of Samian ware. These studies laid the foundations for chemical analyses for the next three decades, in which thousands of analyses using varying techniques were carried out on many types of object including pottery, stone (obsidian, marble, chert, volcanic rocks), amber, and metals including coins. Apart from NAA, XRF, XRD, and spectrographic methods, techniques currently in use include proton induced X-ray emission and proton induced gamma ray emission, inductively coupled plasma emission spectrometry, lead isotope analysis, and electron microscopy, to mention a few. Major changes in the instrumentation of these techniques since the 1970s have meant that more elements can be analyzed with a higher precision. See Harbottle (1982) for a list of materials that can be characterized.

The choice of technique depends on availability to the archaeologist and cost. Over the last 5 years there has been a marked proliferation in sourcing studies directly as a result of easier access by archaeologists to new hand held portable X-ray fluorescence spectrometry (PXRF) machines. Such machines are affordable for anthropology departments, proving popular among archaeologists. Testimony to this is seen is in sessions

devoted to PXRF in major international conferences (see Shackley, 2010a,b).

Theoretical Developments

The importance of trade and exchange in archaeology is the result of a reorientation in the ways archaeologists have perceived trade/exchange operating in the past. The realization that migrationist/invasion models could not adequately explain the distribution of material across the landscape, nor change in societies over time, led archaeologists to look at other mechanisms. Inspired by the works of economic anthropologists such as Polanyi (1957), who defined three types of trade and exchange (reciprocity, redistribution, and market exchange), and anthropologists such as Malinowski (1922), who studied trade and exchange in a Melanesian society, archaeologists realized that trade and exchange were embedded within the social and economic structures of society. By studying trade and exchange, the intention was to reconstruct the economies and organizations of past societies and their changes over time. The access to and control over prestige goods through trade and exchange were seen as prime movers for change leading to ranked societies.

A major proponent for the study of ancient trade and exchange was Colin Renfrew. In a classic paper on trade and culture process in European prehistory, Renfrew (1969) argued that trade was a causal factor for cultural change leading to the beginning of urbanization in the Aegean. An increase in trade with new wealth, craft specialization, and communication, led to change from a 'village subsistence economy into an urban society'. This was a marked shift away from inferring invasions or diffusion from a 'higher culture' to account for the Aegean Early Bronze Age. See Rathje (1971) for a similar case study involving the Maya.

Current Theory and Research

From these earlier studies a number of influential edited volumes on the archaeology of trade and exchange appeared: Sabloff and Lamberg-Karlovsky (1975), Ericson and Earle (1977), Earle and Ericson (1982), Renfrew and Shennan (1982), and Renfrew and Cherry (1986). These proceedings covered an eclectic array of topics from the Old World and New World, and provided the benchmark for trade and exchange studies for the next decade. These studies were important not only in identifying the movement of materials in the past, but also in cementing links with the chemists, geologists, or physicists who undertook the physicochemical analyses which demonstrated that trade and exchange took place. These studies are also important in initiating innovative research into the archaeology of trade and exchange, such as correlating the distribution of materials with modes of exchange and types of social organization, or exploring the relationship between trade and exchange with production and consumption patterns. These innovations will be discussed below.

Within this research a distinction has been made between two schools of thought about the way in which trade and exchange should be modeled: formalist and substantivist. The formalist approach works within the paradigm of market economics. Here, all exchange transactions can be viewed by economic rationalism, based on cost efficiency. Falloff models presented in Section Distribution Patterns are based on cost efficiencies. The substantivist school, on the other hand, views exchange as being embedded in social and economic processes. This approach allows an investigation into social groups, and social change. In reality, archaeologists draw from both approaches to model ancient trade and exchange.

Research mostly followed the following steps:

1. identifying that exchange/trade took place;
2. modeling the distribution patterns of the exchanged/traded material.

Identifying What Is Moving and What Is Not

Trade or exchange *must be demonstrated* rather than assumed. Archaeologists often view stylistic similarities over large distances as a consequence of exchange. Stylistic similarities in pottery, for instance, have been equated with the trade or exchange of pots from one or a restricted number of centers. Yet, in many instances this is not the case. One pottery tradition, Lapita, for instance, is spread from Papua New Guinea to Tonga, a distance of over 4000 km. It is stylistically homogeneous, having a suite of shared decorated motifs made by a dentate stamping technique. Before the 1970s it was thought that this pottery was made in one or a handful of manufacturing centers and then exchanged to other communities, thus accounting for its spread within Melanesia. Yet, after extensive petrographic and chemical analysis, it was established that with the exception of a couple of assemblages, most Lapita pottery is locally made. It was people and ideas that were exchanged, not the pots. Thus, if local production of an object is demonstrated then models other than the trade/exchange of the object must be developed to account for stylistic similarity.

Identifying the source of an object and studying the context of production is not easy. The use of a 'resource zone concept of source' is useful here in defining a geographical region from which the object was produced (Clark et al., 1992). For instance, the exact spot where the firing of a pot took place, or where obsidian was extracted, is not crucial here, only whether the pottery is of local origin or not, or whether the obsidian originated from a defined area. Attributing an object to a general regional zone (or procurement zone) is often more important than trying to identify a precise spot for manufacture.

The aim of many physicochemical analyses is to discern the geographical origin of the object under study. Ideally, the sources of material with restricted distribution could be characterized and 'fingerprinted' chemically or petrographically. Obsidian, for instance, is chemically homogeneous, and its sources can be chemically fingerprinted. Obsidian found in archaeological contexts can be analyzed and related back to the source. Identifying the potential source areas of a material requires an intensive search within a known region. For instance, in a recent study on obsidian exchange within the western Pacific over a 20 000-year time span, archaeologists have concentrated on the source areas within West New Britain,

Papua New Guinea. Over a number of years, obsidian flows were mapped and samples taken for characterization, with the result that a finer discrimination of sources was possible with five source localities chemically defined (Summerhayes et al., 1998). By analyzing obsidian found in archaeological sites (some up to 3700 km distant), it was possible not only to trace obsidian to its source localities, but also to define changes in the selection of different sources over time. For an overview of archaeological approaches to obsidian studies, see Shackley (1998).

Pottery, on the other hand, is more difficult to characterize or fingerprint (*see* Ceramics in Archaeology). Often archaeologists use what is called a 'criterion of relative abundance', which infers that most pottery from a production center is to be found locally. Yet if pottery was produced for exchange only, then little would be found locally, thus limiting the application of this concept. The 'criterion of relative abundance' is useful, however, when used with a physicochemical analysis for provenance studies; that is, relating the pottery under study back to its raw materials: nonplastic inclusions (minerals, sands, etc.) and clay. Mineral inclusions found within pottery can be related to local geology and both river and beach sands. Clays, on the other hand, vary greatly in respect of the underlying geology, and a study of the elemental composition of pottery may give a clue to its origin. The chemical analysis of clay is complex. Bishop et al. (1982) note that clay minerals not only depend on their parent material, but are also affected by weathering or hydrothermal activity, climate, and geomorphology. They state that, "even if a region is considered to be homogeneous in its gross geologic characteristics, significant mineralogical and chemical differences may be discerned between clay deposits" (Bishop et al., 1982).

In any provenance study of archaeological objects a few words of caution from Garman Harbottle, a nuclear physicist, should best be heeded:

> archaeologists love the term sourcing, with its upbeat, positive thrust—that you analyze or examine an artifact and, by comparison with the material of known origin, 'source' it. In point of fact, with very few exceptions, you cannot unequivocally source anything. What you can do is characterize the object, or better, groups of similar objects found in a site or archaeological zone by mineralogical, thermoluminescent, density, hardness, chemical and other tests, and also characterize the equivalent source materials, if they are available, and look for the similarities to generate attributions. A careful job of chemical characterization, plus a little numerical taxonomy and some auxiliary archaeological and/or stylistic information, will often do something as useful: It will produce groupings of artifacts that make archaeological sense. This, rather than absolute proof of origin, will often necessarily be the goal.
>
> (Harbottle, 1982)

For an overview of archaeological approaches to pottery characterization studies, see Neff (1992).

Distribution Patterns

Archaeologists analyze the distribution of an object from its source or production area to its consumption site in order to model the type of trade/exchange that took place. In some instances the nature of the society participating in the trade/exchange system is inferred, such as egalitarian vs chiefly, or nonhierarchical vs hierarchical.

Popular among archaeologists is the use of falloff curves and modes of trade/exchange. By plotting the frequency of an item against distance from its source (falloff curve), Renfrew (1975) observed different distribution patterns which he attempted to relate to modes of exchange or access. Some of these modes included reciprocity, down-the-line, and central-place exchange.

How were falloff curves and modes of exchange related? Down-the-line exchange and reciprocity led to an exponential falloff in abundance away from the source or production center. The term 'law of monotonic decrement' was coined to account for this falloff in the frequency of an item away from the source (Renfrew, 1977). Attempts were also made to define an object's value by the slope of the curve. A small slope equates with high-value goods, such as spondylus shells in Europe, Maya jade, and Oaxaca iron ore mirrors. These modes were equated with an egalitarian society (semisedentary agriculturalists and hunter-gatherers; nonranked; nonhierarchical).

Central-place exchange, on the other hand, was seen as directional (no exponential linear falloff), should lead to multinodal falloff curves, and is indicative of redistribution. Central place relates to the movement of goods toward one place or center, from which they are then distributed. Redistribution has been related to complex chiefly societies (hierarchical, ranked, and/or stratified societies). Examples of such societies given by archaeologists include early states such as the Maya, Egypt, the Near East, or chiefdoms such as Hawaii.

Although problems occur with equifinality – that is, different modes of exchange produce similar distributions – this approach is important as it relates the measurement of archaeological material away from the source to trade and exchange modes and the structures of societies.

Future Directions

Approaches to the archaeology of trade and exchange will continue to change on three fronts. First, change need to occur in tune with developments in archaeological theory. As archaeologists will be asking different questions about the roles of trade and exchange and society, changes in approaches and methodologies will take place. Second, at the same time, it is reasonable to expect advances in characterizing techniques, making them yet more accessible and cheaper to use. Unfortunately, over the last two decades theoretical developments have lagged behind advances and accessibility of characterizing techniques (see Dillian and White, 2010). Lastly, archaeologists will consolidate knowledge on the sources and distribution of a variety of goods, giving a more complete perspective on the roles of trade and exchange within a society. As an example, attention is drawn to the Pacific, where the production and distribution of a number of materials, such as pottery, obsidian, chert, volcanic adzes, and shell artifacts, are used to show that each class of material has different production and distribution patterns. If any class of artifact were taken by itself, a 'myopic' view would be seen of the roles of trade and exchange within societies. Only by comparing the production

and distribution patterns of the complete range of material available can a proper picture of trade and exchange be given.

See also: Archaeometry; Ceramics in Archaeology; Economic Anthropology; Economic Sociology; Exchange in Anthropology; Geoarchaeology; Intensification and Specialization, Archaeology of; Malinowski, Bronislaw (1884–1942); Pacific Islands, Archaeology of; Polanyi, Karl (1886–1964).

Bibliography

Bishop, R., Rands, R., Holley, G., 1982. Ceramic compositional analyses in archaeological perspective. Advances in Archaeological Method and Theory 5, 275–320.

Clark, C.P., Neff, H., Glascock, M., 1992. Neutron activation analysis of the late Woodland ceramics from the Lake Superior Basin. In: Neff, H. (Ed.), Chemical Characterization of Ceramic Pastes in Archaeology. Prehistory Press, Madison, WI, pp. 255–267.

Dillian, C., White, C.L. (Eds.), 2010. Trade and Exchange: Archaeological Studies from History and Prehistory. Springer, New York.

Ericson, J.E., Earle, T.K. (Eds.), 1977. Exchange Systems in Prehistory. Academic Press, New York.

Earle, T.K., Ericson, J.E. (Eds.), 1982. Contexts for Prehistoric Exchange. Academic Press, New York.

Harbottle, G., 1982. Chemical characterization in archaeology. In: Ericson, J.E., Earle, T.K. (Eds.), Contexts for Prehistoric Exchange. Academic Press, New York, pp. 13–51.

Kirch, P.V., 1997. The Lapita Peoples: Ancestors of the Oceanic World. Blackwell, Oxford, UK.

Malinowski, B., 1922. Argonauts of the Western Pacific. Routledge and Kegan Paul, London.

Neff, H. (Ed.), 1992. Chemical Characterization of Ceramic Pastes in Archaeology. Prehistory Press, Madison, WI.

Polanyi, K., 1957. The economy as instituted process. In: Polanyi, K., Arensberg, C.M., Pearson, H.W. (Eds.), Trade and Market in the Early Empires. Free Press, Glencoe, IL, pp. 243–270.

Rathje, W., 1971. The origin and development of classic Maya civilization. American Antiquity 36, 275–286.

Renfrew, C., 1969. Trade and cultural process in European prehistory. Current Anthropology 10, 151–169.

Renfrew, C., 1975. Trade as action at a distance. In: Sabloff, J.A., Lamberg-Karlovsky, C.C. (Eds.), Ancient Civilization and Trade. University of New Mexico Press, Albuquerque, NM, pp. 3–59.

Renfrew, C., 1977. Alternative models for exchange and spatial distribution. In: Earle, T.K., Ericson, J.E. (Eds.), Exchange Systems in Prehistory. Academic Press, New York, pp. 71–90.

Renfrew, C., Cherry, J.F., 1986. Peer Polity Interaction and Socio-Political Change. Cambridge University Press, Cambridge, UK.

Renfrew, C., Shennan, S. (Eds.), 1982. Ranking, Resource and Exchange: Aspects of the Archaeology of Early European Society. Cambridge University Press, Cambridge, UK.

Sabloff, J.A., Lamberg-Karlovsky, C.C., 1975. Ancient Civilization and Trade. University of New Mexico Press, Albuquerque, NM.

Sayre, E.V., Dodson, R., 1957. Neutron activation study of Mediterranean potsherds. American Journal of Archaeology 61, 35–41.

Shackley, S., 1998. Advances in Archaeological Volcanic Glass Studies. Plenum Press, New York.

Shackley, S., 2010a. Is there reliability and validity in portable X-ray fluorescence spectrometry (PXRF)? SAA Archaeological Record 10, 17–20.

Shackley, S. (Ed.), 2010b. X-Ray Fluorescence Spectrometry (XRF) in Geoarchaeology. Springer, New York.

Shepard, A., 1965. Rio Grande glaze-paint pottery: a test of petrographic analysis. In: Matson, F.R. (Ed.), Ceramics and Man, Viking Fund Publications in Anthropology, vol. 41. Aldine, Chicago, pp. 62–87.

Summerhayes, G.R., Bird, R., Fullagar, R., Gosden, C., Specht, J., Torrence, R., 1998. Application of PIXE-PIGME to archaeological analysis of changing patterns of obsidian use in West New Britain, Papua New Guinea. In: Shackley, S. (Ed.), Advances in Archaeological Volcanic Glass Studies. Plenum Press, New York, pp. 129–158.

Young, W., Whitmore, F., 1957. Analysis of oriental ceramics by non-destructive methods. Far Eastern Ceramic Bulletin 9, 1–27.

Trade Policy

Jean-Baptiste Velut, Université Sorbonne Nouvelle, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The notion of trade policy is a transdisciplinary object that has received considerable attention among economists, political scientists, international relations scholars, and sociologists seeking to understand its origins, nature, and outcomes. Its conceptualization has changed according to theoretical developments in each of its nurturing disciplines – the rise of neo-institutionalism and international political economy being only two of the main subfields contributing to enrich theories of trade policy. At the empirical level, the very definition of trade policy has always been contingent upon the history of the world economy. Thus, the changing nature and ever-greater significance of international trade over the past 50 years has renewed controversies surrounding the economic, social, and environmental ramifications of trade policy, and more broadly surrounding the evolving relationship between markets and states under the new constraints of globalization.

What is Trade Policy?

At its core, trade policy consists of managing the international exchanges of goods and services between national and regional economies. In the Westphalian era, this realm of public policy has traditionally been the prerogative of national governments, although the behavior of private actors has shaped international trade flows from the monopoly of the East India Company to Apple's global supply chains. The conduct of trade policy falls into two main categories: (1) the regulation of imports and (2) the management of exports, which itself is divided into export promotion and export controls. The growing significance of international trade – from 22% of world GDP in the early 1970s to more than 50% in 2012 – and the intensification of economic competition in the second half of the twentieth century prompted decision makers to develop an ever-wider array of policy tools to manage the circulation of goods and services.

Long confined to tariff duties and import quotas, trade policy has mutated into more sophisticated forms of import restrictions designed to protect domestic interests threatened by foreign competition, among which are rules of origins, sanitary and phytosanitary standards, and other technical barriers to trade. Conversely, national governments have resorted to a growing range of measures aimed at supporting both small and large exporting companies, whether through preferential loans, insurance programs, technical assistance, or trade missions. Finally, export controls remain an integral part of trade policy, although they have been generally subordinate to security interests (economic sanctions, arms controls).

Trade policy reflects a country's openness to, or isolation from, the rest of the world and is, therefore, implemented through regional or multilateral trade agreements. Multilateral agreements have been negotiated under the aegis of the General Agreement on Trade and Tariffs (GATT) since 1947, before it was replaced by the World Trade Organization (WTO) in 1995. Today, the WTO continues to promote trade liberalization (under the nondiscriminatory principle of the 'Most Favored Nation') and has played an important role in settling trade conflicts among its 150 members, thanks to its dispute settlement body.

Unlike multilateral agreements, regional accords – whether bilateral or plurilateral – are by definition preferential trade agreements (PTAs) insofar as they grant concessions to specific trading partners. PTAs have proliferated over the past 20 years as countries compete for foreign markets and/or favor regional integration over the more competitive route of multilateralism. In 2014, the WTO recorded no fewer than 583 notifications of regional trade agreements throughout the world. Trade policymakers can resort to three types of PTAs according to the degree of economic integration that they seek. Partial scope agreements such as the European Steel and Coal Community (1951) or the US–Canada Auto Pact (1965) are conducted on a sectoral basis and are often preliminary to greater economic cooperation. Free trade areas like the North American Free Trade Agreement (NAFTA) aim to dismantle trade barriers among a set of economic partners while leaving them free to pursue independent policies with nonmembers. In contrast, customs unions like the European Union imply not only trade liberalization within a certain geographic zone but also impose common external tariffs to its members.

The growing complexity of both regional and multilateral agreements reflects the above-mentioned diversification of trade policy tools. At the multilateral level, trade negotiators began to focus on 'nontariff barriers' during the Tokyo Round (1973–79). Since then, the GATT/WTO's agenda has embraced issues that go far beyond trade, such as intellectual property rights, investment, procurement, or regulatory standards (Maskus, 2000; Muzaka, 2009; Weiss, 2006). The so-called 'behind-the-border' rules have raised sensitive questions about national sovereignty and, as a result, involved an ever-greater number of trade policy stakeholders. This has made the design of trade policy increasingly complex.

Trade Policy Formation

The study of trade policy formation can be classified into four categories, each of which has its own subdivisions and variations: system-centered, society-centered, and state-centered approaches, as well as 'two-level' or 'multilevel' games. These interpretations should not be conceived as mutually exclusive

but as the different pieces of the trade policy puzzle, a complex process at the intersection of the domestic and international sphere. 'Two-level games' are particularly well equipped to depict the complexity of trade policy models insofar as they incorporate multiple variables at each level of analysis: the role of societal forces, decision makers, and institutional structures in the domestic sphere; and state actors, intergovernmental organizations, and transnational advocacy networks in the international sphere.

Structural Theories of Trade Policy

Hegemonic Stability Theory

System-centered theories of trade policy are primarily concerned with the international level of analysis. They focus on the relation between the world's distribution of power and the structure of international trade. This type of literature can be subdivided into two principal schools of thoughts: hegemonic stability theory and (neo)Marxian models.

Kindleberger's *World in Depression* (1973) is commonly viewed as the earliest formulation of hegemonic stability theory. In his study of the interwar period, Kindleberger identified Britain's inability and America's reluctance to stabilize the international economic system as the main causes of the Great Depression. Two years later, amid fears about the declining US power, Gilpin (1975) drew the lessons from Kindleberger's seminal contribution to predict that America's increasingly contested hegemony would lead to a renewal of mercantilist policies. Krasner (1976) reformulated these empirical studies in a coherent framework that would later be called hegemonic stability theory. Krasner's systemic theory of trade policy can be summarized as follows: "a hegemonic distribution of power is likely to result in an open trading structure" (Krasner, 1976). His rationale rests upon the idea that trade openness, while raising national economic income, tends to generate social dislocations as labor shifts from one sector to another. Consequently, only a large and dynamic economy, like Britain in the nineteenth century or the United States after World War II, can bear the brunt of promoting trade liberalization.

In contrast with Gilpin and Krasner's realist perspectives, Keohane renewed with Kindleberger's 'benevolent' version of hegemonic stability theory to stress the correlation between the 'leadership lag' of the 1970s and changes in international economic regimes, namely in the oil, monetary, and trade areas (Keohane, 1980; Keohane and Nye, 1977). In a reformulation of his argument, Keohane (1984) argued that cooperation can in fact persist in a nonhegemonic international system under the tutelage of international regimes. His emphasis on the 'independent life' of regimes distinguishes Keohane from Gilpin, who has continued to argue that the absence of a hegemon renders cooperation exceptionally difficult (Gilpin, 2001).

(Neo)Marxian Approaches to Trade Policy

The first prominent authors to apply Marxian theories to the international trade system were Lenin and his contemporaries. Along with the fellow Marxists Bukharin, Hilderling, and Luxemburg, Lenin interpreted imperialism as a new era dominated by *financial* capital. In this new phase of capitalism, developed countries occupied a monopolistic position and struggled for the control of foreign markets and industries. Their foreign economic policies served the interests of financial capital and were by definition the result of "the struggle of the Great Powers for the economic and political division of the world" – (Lenin, 1939). What was idiosyncratic of this period was the fact that Marxists like Lenin and Kautsky associated imperialism with mercantilist trade policies (Etherington, 1982), eventhough Rosa Luxemburg stressed the inherent contradictions between the expansionist proclivities of financial capital and the nationalist scope of protective tariffs (1968).

Conversely, the next wave of Marxian interpretations of trade policy emerged during an era dominated by trade-liberalizing policies. In the late 1960s to early 1970s, *dependencia* theorists interpreted foreign penetration in the political economies of Latin America as a source of underdevelopment. For *dependentistas*, as well as world system theorist Wallerstein, the development of capitalism divided the world economy into metropolitan centers and peripheral satellites, the former expropriating the latter's economic surplus (Chilcote, 1974; Frank, 1967, 1969; Wallerstein, 1979). Like Leninism, this form of neo-Marxism interpreted trade policy as the result of structural economic forces. As Frank wrote, "trade relations are based on monopolistic control of the market" (1970: p. 231). Within this framework, capital interests drove trade policy.

Society-Centered Theories of Trade Policy

While system-centered theories focus on the international level of analysis, society-centered or 'market' perspectives on trade policy are concerned with domestic politics. Within this framework, the policy process is the result of a struggle for influence among various interest groups, a political competition arbitrated by decision makers (Ikenberry et al., 1988). In political science, this theoretical model borrows from the American pluralist tradition and has dominated analyses on the US trade politics since Schattschneider's seminal study of the Smoot–Hawley Tariff Act of 1930 (Schattschneider, 1935; Gourevitch, 1977; Baldwin, 1985).

In the sphere of economics, theories of trade policy have developed along two standard economic paradigms: the factor model and the sector model. Based on the Stolper–Samuelson theorem, the first claims that trade will reward abundant factors (e.g., capital in the United States or labor in China), but punish or reduce the income of scarce factors (American workers or Chinese capital-owners), leading eventually to 'factor price equalization,' i.e., the convergence of factor costs (Oatley, 2006: pp 70–74). Under this assumption (also referred to as the Hecksher–Ohlin model), trade policy is driven by the competition between mobile factors of production. From these principles of neo-classical economic theory, Rogowski (1989) derived political implications: scarce factors will likely lobby for protectionism, while abundant factors will pressure the state for increased openness. 'Losers' from trade liberalization will experience declining political leverage, while economic winners will become increasingly powerful.

The second and dominant society-centered paradigm rejects the mobility of production factors and the monolithic vision of class interests upon which the factor model is based, pointing not only to the problematic conversion of low-skilled workers into high-skilled technicians, but also to the lack of

fungibility of capital from one sector (e.g., mining industry) to another (e.g., biotechnology). Instead, the Ricardo–Viner model holds factors as 'specific' to industries and focuses on the conflict of interests between sectors, and more precisely on the competition between import-competing industries and export-oriented companies (Schattschneider, 1935; Bauer et al., 1972; Baldwin, 1985; Magee and Young, 1987; Alt et al., 1996). While these two economic theories are generally considered to be mutually exclusive, Hiscox (2001, 2002) has sought to bridge the gap between factor and sector models. He argues that coalition patterns depend on interindustry factor mobility. Thus, broad class-based coalitions are more likely where factor mobility is high, while narrow industry-based coalitions are more common where mobility is low.

While conventional society-centered interpretations of trade policy provide useful insights into the domestic dynamics of trade policy, they tend to neglect or downplay the importance of two major developments in the international political economy. First, the often missing items in societal trade models are the transformative effects that the geographic mobility of capital has had on trade politics over the past three decades. Indeed, since the 1970s, multinational corporations have developed complex global supply chains leading to a dramatic expansion of intrafirm trade. The rapid liberalization of international trade and investment flows contrasts with the enduring immobility of labor. This asymmetry has generated intrasectoral class conflicts that conventional sector-specific models can hardly explain. On the one hand, multinational corporations can benefit from the cost reductions resulting from outsourcing and are, therefore, prone to support laws that liberalize trade and investment flows. On the other hand, domestic workers are the first victims of offshoring and will logically oppose the very same trade agreements – as witnessed by the protectionist retrenchment of American labor since the 1970s (Helleiner, 1977; Moody, 1997; Darves and Dreiling, 2002). The growing class cleavages of American trade politics have been particularly visible during the recent debates on NAFTA, which, for instance, pitted autoworkers against their employers (Chase, 2003; Cox, 1995).

A second lacuna of both factor- and sector-based models is related to another feature of contemporary trade politics that cannot be captured by class-based analyses: the expanding scope of trade policy beyond tariff barriers. The growing social and environmental ramifications of trade agreements logically triggered the political mobilization of new trade policy stakeholders among who are environmentalists, consumer advocates, or human rights groups. From the NAFTA debates to the Seattle protests of 1999, nongovernmental organizations in the United States and elsewhere have invited themselves to the negotiating table (Hinojosa-Ojeda, 2002; Mayer, 1998; Vogel, 1997; Destler and Balint, 1999; Esty, 1998). They have struggled to broaden the terms of trade debates in a policy sphere long reserved to trade associations. Some scholars have focused on the challenges and opportunities inherent to integrating trade policy and environmental regulation (Gallagher, 2004; Fickling and Schott, 2011). Others have examined the rising prominence of labor standards in the trade policy sphere and pondered whether globalization deters or encourages the enforcement of labor and/or human rights (Elliott and Freeman, 2003; Flanagan, 2006).

Globalization scholars are often better equipped to study the dynamics of these new trade-related debates than conventional trade economists. The latter tend to reduce trade policy outcomes to either free trade or protectionism, thereby leaving aside important moral questions on the social terms under which international trade takes place.

Finally, and despite their dominance in the field, society-centered studies of trade policy have been criticized for their disregard for the role of decision makers and institutions. In particular, the market-based approaches described above often reduce the policymaker to "an auctioneer in a product market" (Magee and Young, 1987), a lacuna that has been challenged by neo-institutionalist scholars.

State-Centered Approaches

Over the past two decades, political science has departed from pluralistic, economics-based models of policymaking to 'bring the state back in' and (re)emphasize the conditional autonomy of government on the hand, and the 'policy legacies' of institutions on the other (Evans et al., 1985; March and Olsen, 1998; Immergut, 1998). This renewed interest in institutional forces gave rise to state-centered approaches to trade policy, which fall into two categories: 'state-as-an-actor' theories, which highlight the independence of decision makers, and institutionalist models, which stress the impact of domestic structures.

Within the former literature, the emphasis is on how governments shape trade policy to defend their national interest, of which international competitiveness is a crucial component. Under this category of argument falls the perennial 'infant-industry' argument according to which government must shield its emerging domestic industries from foreign competition until they become fully competitive. This argument can be traced back as far as 1645 (Irwin, 1996), but was most famously enunciated by Alexander Hamilton (1791), who called for the adoption of tariff duties to protect its nascent manufactures of America. This challenge to the free trade orthodoxy of classical economists like Adam Smith and David Ricardo reemerged in the 1980s under the banner of strategic trade theory. Often associated with the work of Brander and Spencer (1983) and Krugman (1986), strategic trade theory took stock of the imperfection of markets, the logic of economies of scale, scope, and experience, as well as the role of technology spillovers to explain how government could protect and invest in national firms in order to maintain a competitive advantage in strategic sectors of the economy (Gilpin, 2001). This theoretical conception of trade policy has broad empirical implications to the extent that it tends to legitimize interventionist policies in the name of economic competitiveness. The rise of China has renewed interest in the so-called 'developmental states' and their ability to shape growth through industrial and trade policies (Rodrik, 2007; Haley and Haley, 2013). Today, as in the 1980s, when Japan's economic surge threatened American competitiveness, sector-specific policies ranging from government subsidies, nontariff barriers, and export promotion are once again a source of both envy and anxiety toward the so-called 'unfair trade.' Amid these

political debates, political economists have, however, shattered conventional dichotomies between strong and weak states, revealing, for instance, that the United States has long played an active role in supporting the private sector, not least with protectionist or interventionist trade policies (Block, 2008; Chang, 2003; Weiss, 1998).

While this approach confers great responsibility to decision makers, a second school of thought has devoted greater attention to the institutional constraints under which they must operate. For institutionalist scholars of foreign economic policy, the formation of domestic structures has a durable impact on politics as it restricts future policy choices and potentially shapes ideas and coalitions (Katzenstein, 1978; Gourevitch, 1978; Goldstein, 1994). As Goldstein writes, "policy is path-dependent: the world does not start anew at each decision point. Rather, history builds an ever-more-complex web of interests, ideas and policies" (1994: p. 22). In short, institutions shape the orientation of trade policy, while institutional changes occur primarily in exceptional circumstances such as periods of crises (major wars, deep recessions, etc.) (Ikenberry et al., 1988).

For instance, in the United States, the Reciprocal Trade Agreements Act of 1934 has been largely held as a catalyst for the adoption of liberal trade policies. Indeed, by transferring the tariff-making authority from the legislative to the executive, it redefined the relationship between business and government. This power shift was renewed with the creation of fast track authority in 1974, which delegates trade-negotiating authority to the executive branch, and has arguably helped to shield the decision-making process from pressure groups (Destler, 1986; Haggard, 1988; Goldstein, 1994). On the other hand, free trade institutions have coexisted not only with older protectionist institutions such as the US escape clause, but also with the resurgence of protectionist remedies in the last quarter of the twentieth century, giving rise to a 'hybrid' trade policy that promotes both trade liberalization and protective tariffs in certain sectors (Goldstein, 1986).

A common grievance addressed to institutionalism is that it is a poor predictor of change. Indeed, by themselves, institutional factors are nothing more than intervening variables. They are only relevant because they can affect the mobilization of political actors on behalf of policy innovation. On the other hand, and as some statist theorists acknowledge, institutions are also 'malleable' and can be instrumentalized by decision makers (Haggard, 1988; Ikenberry, 1988).

As Ikenberry notes, "in terms of explaining foreign economic policy, the institutional approach can provide only some of the answers" (Ikenberry, 1988). To be useful, institutional and societal forces must, therefore, be combined. Finally, the study of trade policy formation must also account for the interplay between the domestic and international levels of analysis, an intellectual undertaking best encapsulated in two-level games.

Two-Level and Multiple-Level Games

While economists and institutionalists provided crucial insights into the internal dynamics of policymaking, international relations scholars have encouraged students of trade policy formation to consider both domestic and international factors. The seminal contributions of Bruck et al. (1962) and Gourevitch's *Second Image Reversed* (1978) represented two early attempts to analyze the interaction between domestic politics and the international environment. Gourevitch posited that international events were not only the reflection of domestic structures – 'a second image' – but also a source of political change at the state level: a 'second image reversed.' Both Bruck et al. and Gourevitch's analyses included a wealth of variables: structural, societal, and even ideational factors at the domestic level; institutional and power-related determinants at the international level.

Putnam (1988) built upon the work of his predecessors to formulate a more parsimonious theory. His analysis of international economic negotiations at the Bonn Summit of 1978 has since become a classic in international relations. For Putnam, international politics can be conceived as a two-level game: at the national level (Level II), domestic actors pursue their interests by pressuring the government, while politicians seek political support by building coalitions; at the international level (Level I), national governments strive to satisfy their domestic constituencies, while minimizing the adverse effects of foreign developments.

Within his framework, the decision-making process hinges on a 'win-set,' defined as the set of all possible international agreements that would obtain the necessary majority among domestic constituencies. For Putnam, the size of this 'win-set' is crucial to the outcome of negotiations. Larger win-sets are more likely to be successful, i.e., to lead to international cooperation. Conversely, smaller win-sets, whether real or simulated by negotiators, are more difficult to overcome and, thus, likely to confer more bargaining power to state actors. What forces shape win-sets? Here, Putnam points to three factors: domestic politics, state institutions, and the strategy of international negotiators. Despite his emphasis on the significance of domestic politics, Putnam does not exclude the possibility of a 'second image reversed.' In fact, his concept of 'reverberation' posits that international pressures can impact the domestic sphere. Putnam provides a robust theoretical toolbox to understand the complex dynamics of trade politics, insofar as it synthesizes many elements of the society-centered and statist approaches.

As a result, it has become a popular framework for the study of trade policy formation, particularly congenial with intergovernmental negotiations. For instance, two separate studies of the negotiations on NAFTA show that two-level games are particularly well equipped to analyze North American trade politics (Cameron and Tomlin, 2002; Mayer, 1998). However, Putnam's analytical framework shows less explanatory power when applied to intergovernmental negotiations within well-established international organizations like the European Union. Indeed, not only do 'two-level games' ignore the autonomous role of supranational institutions in agenda setting, but they also disregard the hypothesis that national governments might instrumentalize international negotiations to pursue unpopular domestic reforms that fall outside their win-sets. This is all the more important in the trade policy sphere, where national governments have delegated their negotiating authority to the European Union. To remedy this problem, scholars have expanded Putnam's framework into

a 'three-level game' that incorporates the logic and dynamics of supranational institutions (Moyer, 1993; Patterson, 1997). In doing so, they renewed with the neo-functionalist tradition of Ernst Haas (1958), who studied the role of transnational actors and supranational officials to understand the self-reinforcing process of European integration. Others have developed multilevel analyses to account for multiple factors including subnational groups, supranational actors, or transnational advocacy networks (Smith and Ray, 1993; Hinojosa-Ojeda, 2002).

The Future of Trade Policy

These sophisticated models illustrate the growing complexity of the trade policy process in the globalized economy, a policy sphere that is no longer confined to tariff duties. The challenge of negotiating far-reaching agreements under the unanimity rule of the WTO – illustrated by the failure of the Doha Round – has led some countries to pursue trade initiatives at the bilateral and regional levels that may be more attuned to their economic interests and worldviews (Krist, 2013; Lynch, 2010). Amid the proliferation of regional trade agreements, two competing varieties of regionalism have emerged. One model, defended aggressively by the United States and Europe, aims to promote comprehensive agreements that go far beyond conventional tariffs to tackle complex issues such as regulatory coherence, digital trade, and state-owned enterprises. This deeper form of integration contrasts with China's policy of noninterference and the more circumscribed nature of its multiple trade initiatives. Without a revival of multilateral trade negotiations, the future of trade policies is likely to be shaped by these competing models. More than a complex bureaucratic apparatus, the debates surrounding trade policy are in essence a battle over the norms and rules of economic globalization that will be shaped by the dynamics of the world economy; the geopolitical tactics of competing nation-states; and the social and economic interests of producers, workers, and consumers.

See also: Globalization: Geographical Aspects; International Business; International Business: Geographic Aspects; International Trade: Commercial Policy and Trade Negotiations; International Trade: Economic Integration; Multinational Corporations.

Bibliography

Alt, J.E., Frieden, J., Gilligan, M.J., Rodrik, D., Rogowski, R., 1996. The political economy of international trade: enduring puzzles and an agenda for inquiry. Comparative Political Studies 29, 689–717.

Baldwin, R.E., 1985. The Political Economy of U.S. Import Policy. The MIT Press, Cambridge, MA/London.

Bauer, R.A., Pool, I.D.S., Dexter, L.A., 1972. American Business and Public Policy: The Politics of Foreign Trade. Aldine-Atherton, Chicago.

Block, F., 2008. Swimming against the current: the rise of a hidden developmental state in the United States. Politics & Society 36 (2), 169–206.

Brander, J.A., Spencer, B.J., 1983. International R&D rivalry and industrial strategy. Review of Economic Studies 50, 707–722.

Bruck, H.W., Sapin, Burton M., Snyder, Richard C., 1962. Foreign Policy Decision-Making: An Approach to the Study of International Politics. Free Press of Glencoe, New York.

Cameron, M.A., Tomlin, B.W., 2002. The Making of NAFTA. How the Deal Was Done. Cornell University Press, Ithaca/London.

Chang, H.J., 2003. Kicking Away the Ladder. Development Strategy in Historical Perspective. Anthem Press, London.

Chase, K.A., 2003. Economic interests and regional trading arrangements: the case of NAFTA. International Organization 57, 137–174.

Chilcote, R.H., 1974. Dependency: a critical synthesis of the literature. Latin American Perspectives 1, 4–29.

Cox, R.W., 1995. Explaining business support for regional trade agreements. In: Frieden, J., Lake, D. (Eds.), International Political Economy. Perspectives on Global Power and Wealth. Wadsworth, Belmont, CA.

Darves, D., Dreiling, M., 2002. Corporate political networks and trade policy formation. Humanity & Society 26, 5–27.

Destler, I.M., 1986. Protecting congress or protecting trade? Foreign Policy 62, 96–107.

Destler, I.M., Balint, P.J., 1999. The New Politics of American Trade: Trade, Labor, and the Environment. Institute for International Economics, Washington.

Elliott, K., Freeman, R.B., 2003. Can Labor Standards Improve under Globalization? Peterson Institute, Washington, DC.

Esty, D.C., 1998. Environmentalists and trade policymaking. In: Deardoff, A., Stern, R. (Eds.), Representation of Constituent Interests in the Design and Implementation of US Trade Policies. University of Michigan Press, Ann Arbor, pp. 201–219.

Etherington, N., 1982. Reconsidering theories of imperialism. History and Theory 21, 1–36.

Evans, P.B., Rueschemeyer, D., Skocpol, T. (Eds.), 1985. Bringing the State Back in. Cambridge University Press, New York.

Fickling, M., Schott, J.J., 2011. NAFTA and Climate Change. Peterson Institute, Washington.

Flanagan, R.J., 2006. Globalization and Labor Conditions: Working Conditions and Labor Rights in a Global Economy. Oxford University Press, New York/Oxford.

Frank, A.G., 1967. Capitalism and Underdevelopment in Latin America. Historical Studies of Chile and Brazil. Monthly Review Press, New York/London.

Frank, A.G., 1969. Latin America: Underdevelopment or Revolution. Essays on the Development of Underdevelopment and the Immediate Enemy. Monthly Review Press, New York/London.

Gallagher, K.P., 2004. Free Trade and the Environment. Mexico, NAFTA and beyond. Stanford University Press, Stanford.

Gilpin, R., 1975. U.S. Power and the Multinational Corporation. Basic Books, New York.

Gilpin, R., 2001. Global Political Economy. Understanding the International Economic Order. Princeton University Press, Princeton/Oxford.

Goldstein, J., 1986. The political economy of trade: institutions of protection. American Political Science Review 80, 161–184.

Goldstein, J., 1994. Ideas, Interests, and American Trade Policy. Cornell University Press, Ithaca/London.

Gourevitch, P., 1977. International trade, domestic coalitions, and liberty: comparative responses to the crisis of 1873–1896. The Journal of Interdisciplinary History VIII, 281–313.

Gourevitch, P., 1978. The second image reversed: the international sources of domestic politics. International Organization 32, 881–912.

Haley, U.C.V., Haley, G.T., 2013. Subsidies to Chinese Industry: State Capitalism, Business Strategy, and Trade Policy. Oxford University Press, New York.

Haggard, S., 1988. The institutional foundations of hegemony: explaining the reciprocal trade agreements act of 1934. International Organization 42, 91–119.

Hamilton, A., 1791. Report on manufactures. In: Letwin, W. (Ed.), A Documentary History of American Economic Policy since 1789. Aldine Publishing Company, Chicago.

Haas, E., 1958. The Uniting of Europe: Political, Social and Economic Forces, 1950–1957. Stanford University Press, Stanford.

Helleiner, G.K., 1977. Transnational enterprises and the new political economy of U.S. trade. Oxford Economic Papers 29, 102–116.

Hinojosa-Ojeda, R., 2002. Integration policy from the grassroots up: transnational implications of Latino, labor, and environmental NGO strategies. In: Brooks, D., Fox, Jonathan (Eds.), Cross-Border Dialogues. U.S.-Mexico Social Movement Networking. Center for U.S.-Mexican Studies, La Jolla, CA. University of California, San Diego.

Hiscox, M.J., 2001. Class versus industry cleavages: inter-industry factor mobility and the politics of trade. International Organization 55, 1–46.

Hiscox, M.J., 2002. Commerce, coalitions, and factor mobility: evidence from congressional votes on trade legislation. The American Political Science Review 96, 593–608.

Ikenberry, J.G., Lake, D.A., Mastanduno, M., 1988. Introduction: approaches to explaining american foreign economic policy. International Organization 42, 1–14.

Ikenberry, J.G., 1988. Conclusion: an institutional approach to American foreign economic policy. International Organization 42 (1), 219–243.
Immergut, E.M., 1998. The theoretical core of the new institutionalism. Politics and Society 26, 5–34.
Irwin, D., 1996. Against the Tide. An Intellectual History of Free Trade. Princeton University Press, Princeton.
Katzenstein, P.J. (Ed.), 1978. Between Power and Plenty. Foreign Economic Policies of Advanced States. The University of Wisconsin Press, Madison.
Keohane, R.O., Nye, J.S., 1977. Power and Interdependence: World Politics in Transition. Brown, Boston, Little.
Keohane, R.O., 1980. The theory of hegemonic stability and changes in international economic regimes, 1967–1977. In: Holsti, O.R., Siverson, Randolph M., George, Alexander L. (Eds.), Change in the International System. Westview Press, Boulder, pp. 131–162.
Keohane, R.O., 1984. After Hegemony: Cooperation and Discord in the World Political Economy. Princeton University Press, Princeton.
Kindleberger, C.P., 1973. The World in Depression 1929–1939. Allen Lane The Penguin Press, London.
Krasner, S.D., 1976. State Power and the Structure of International Trade. World Politics 28 (3), 317–347.
Krist, W., 2013. Globalization and America's Trade Agreements. Woodrow Wilson Center Press, Washington, DC.
Krugman, P.R. (Ed.), 1986. Strategic Trade Policy and the New International Economics. MIT Press, Cambridge.
Lenin, V.I., 1939. Imperialism. The Highest Stage of Capitalism. A Popular Outline. International Publishers, New York.
Luxemburg, R., 1968. The Accumulation of Capital. Modern Reader Paperbacks, New York/London.
Lynch, D.A., 2010. Trade and Globalization: An Introduction to Regional Trade Agreements. Rowman & Littlefield, Lanham, MD.
Magee, S.P., Young, L., 1987. Endogenous protection in the United States. In: Stern, R.M. (Ed.), U.S. Trade Policies in a Changing World Economy. MIT Press, Cambridge.
March, J.G., Olsen, J.P., 1998. The institutional dynamics of international political orders. International Organization 52, 943–969.
Maskus, K., 2000. Intellectual Property Rights in the Global Economy. Institute for International Economics, Washington, DC.
Mayer, F.M., 1998. Interpreting NAFTA. The Science and Art of Political Analysis. Columbia University Press, New York.
Moody, K., 1997. Workers in a Lean World. Verso, New York.
Moyer, H.W., 1993. The European community in the GATT Uruguay round: preserving the common agriculture policy at all costs. In: Avery, W. (Ed.), World Agriculture and the GATT. Lynne Rienner Publishers, Boulder, pp. 95–117.
Muzaka, V., 2009. Shaping global rules: proprietary pharmaceutical companies as global political actors. New Political Economy 14 (2), 289–301.
Oatley, T., 2006. International Political Economy. Interests and Institutions in the Global Economy. Pearson Longman, New York/San Francisco/Boston.
Patterson, L.A., 1997. Agricultural policy reform in the European community: a three-level game analysis. International Organization 51, 135–165.
Putnam, R.D., 1988. Diplomacy and domestic politics: the logic of two-level games. International Organization 42, 427–460.
Rodrik, D., 2007. One Economics. Many Recipes. Globalization, Institutions, and Economic Growth. Princeton University Press, Princeton, Oxford.
Rogowski, R., 1989. Commerce and Coalitions. Princeton University Press, Princeton.
Schattschneider, E.E., 1935. Politics, Pressures and the Tariff. Archon Books, Hamden.
Smith, D., Ray, J.L., 1993. The 1992 project. In: Smith, D., Ray, James L. (Eds.), The 1992 Project and the Future of Integration in Europe. M.E. Sharpe, Armonk.
Vogel, D., 1997. International trade and environmental regulation. In: Vig, N.J., Kraft, Michael E. (Eds.), Environmental Policy in the 1990s. CQ Press, Washington.
Wallerstein, I., 1979. The Capitalist World-Economy. Cambridge University Press, London/New York/New Rochelle.
Weiss, L., 1998. The Myth of the Powerless State: Governing the Economy of a Global Era. Polity Press, Oxford.
Weiss, L., 2006. The business of buying American: public procurement as trade strategy in the USA. Review of International Political Economy 13 (5), 701–724.

Trade Unions and Social Work: Lessons from Canada

Winnie Ng, CAW-Sam Gindin Chair in Social Justice and Democracy, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article argues that the rise of new managerialism and new public management has taken the 'worker' out of social work, to the extent that the field has lost its historic connection to social justice. The need for solidarity between social workers and unions is urgent in the face of rising neoliberalism and austerity agendas that threaten to undo fundamental social reforms. Through historical examples and case studies, and with an eye toward the importance of equity, this article proposes means for building new community–labor alliances through a critical linkage between social workers as professionals and trade union activists.

Introduction

Setting the Broader Context

Globally, trade unions, public services, and the public at large are bearing the brunt of the deep cuts in social spending, job reductions, outsourcing, and related deficit reduction measures. In the present economic climate (termed 'gloves-off' by Bernhardt et al., 2008), where managerial flexibility and corporate profit margins take precedence over the needs of working people, global labor market restructuring threatens to have serious impacts on workers, their families, and their communities (Berberoglu, 2011).

The dominant political narrative of the austerity agenda has been a 'blame game' – blaming the victims, blaming others who are more vulnerable; blaming anyone (such as union bosses and environmentalists) who can be construed as threatening profit maximization or making business less 'efficient.'

At the same time, the serious funding reductions in social services have ushered in a new era of commodification of services (Dominelli, 2004). Social work services are increasingly under threat and undermined. As a result, people in poverty end up having less access to public services. In the context of a gendered and racialized labor market (Galabuzi and Block, 2011), the intersectionality of race, class, gender, and other forms of systemic oppression is reflected in who bears the brunt of this global race to bottom.

Given the adverse impact of the austerity agenda on both the labor movement and social work, this article explores what trade unions and social work share, and reimagines the conditions for possible convergences between the two. Finally, this article considers the strategies that social workers in the dual roles of social work practitioners and trade union activists can carry out to create a collective narrative of resistance.

The Need for Renewal and Convergence between Trade Unions and Social Work Practice

Within the Trade Union Context

There has been a massive erosion of public sector employment. In Canada alone, the Federal government has eliminated 25 000 jobs in past year. The other pressing challenge is the legislative attack on union recognition such as the Rand formula in Canada and right to work legislation in the United States. In November 2012, the American State of Michigan became the 24th state to pass 'right to work' legislation which weakens union membership and provides a pretext to lower wages.

Over the past two decades, with the decline in union density, labor scholars have raised the urgent need to reverse the trend and proposed new organizing strategies to reach women, youth, racialized, and immigrant workers (Kumar and Schenk, 2003; Briskin, 2006). However, studies have also shown that recruiting new membership without meaningful engagement can be problematic and there is a need for labor to democratize the internal union practices for stronger member participation (Murray and Levesque, 2005; Bernard, 2006).

Trade unions, as a microcosm of a society where sexism, racism, and other forms of systemic discrimination are reflected in everyday practice, also find themselves being challenged on their policies and practices of exclusion by their own membership. The vision of building an interracial working-class movement for social justice and solidarity has also been put forth as part of the critical dialogue on labor renewal (Gasper and Fletcher, 2008; Ng, 2012).

Within Social Work Context

The austerity agenda has social workers juggling the ideals of social justice and inclusion with the harsh realities of reduced funding. Government funders have imposed restrictions on the amount of advocacy work that agencies can do, as well as further restrictions on welfare benefits, unemployment insurance, and other 'safety net' features of the so-called welfare state. These austerity measures have been imposed throughout the West – in the United States and Canada, to be sure, but also in Spain, Portugal, Greece, Cyprus, and Italy, among others. Such restrictions have caused agencies to self-censor or even retreat entirely from advocacy work, which is now deemed 'too political' (Carniol, 2010).

The new approach to service provision means clients are 'case managed' rather than empowered and social workers are 'micro-managed' as if they were piece-workers in the manufacturing sector. Service providers are pressured to process more cases, rather than do more good, to justify their

funding. At the end of the helping process, clients are never engaged as autonomous agents, only as passive recipients of services.

Very often, the clients of service providers are workers who have lost their jobs due to plant closures and are now seeking help through a traumatic period of transition. Not only is there a total erasure of the former identities and experiences as employees, union members, or activists, there is an absence of collective work that probes the root cause of these problems. Fisher and Shragge (2008) have been critical of such social work practice and urge community organizations to embrace the model of social action organizing that will deal with underlying issues, create opposition, and challenge the status quo. On the other hand, one wonders how many social workers will identify with Donna Baines' stance that unions are an effective vehicle for antioppressive resistance (2011) and suggest union organizing to their clients as a possible course of action.

Edward Scanlon and Scott Harding (2005) suggest the potential for social work to 'reconnect' with organized labor, based on shared fundamental values and the history of the early days of social work. The social work profession needs a political ally if it is to be successful in advancing the historical progressive vision of the profession, and protect social programs from current cuts or threats. The same sentiment is poignantly expressed in the statement of the Social Worker Action Network (SWAN) based in United Kingdom (http://www.socialworkfuture.org/):

> We believe the role of the unions is crucial given the attacks that are looming in the public sector where no area of social work is likely to go unscathed. We will all be blamed for the crisis rather than the target setting, caseloads, staff shortages, bureaucracy and IT pressures that are currently dominating all areas of our practice.

The Deep Historical Roots of Collaboration and Shared Values

The labor movement has always acknowledged unequal power relations in the workplace and in society at large and recognized that countervailing power can come from organizing, mobilizing, and agitation. Employers have been unwilling to relinquish or share their control and power in their workplaces. In North American labor history, worker-oriented legislation and social programs were won as a result of protest, civil disobedience, and working-class solidarity (see, e.g., Hicks and Swank, 1992).

May First as International Workers Day originated as a way to commemorate the workers killed in agitating for an 8-h work day in Chicago in May 1886. In Canada, the General strike of Winnipeg in 1919 is a testimony to workers' courage when skilled trade workers protested about unemployment after WWI. The Trek to Ottawa in the 1930s was initiated by unemployed workers as a response to the intolerable conditions in the work camp set up by the Federal government during the recession. As a result, unemployment insurance benefits were legislated in 1936. Through the years, other key social reforms such as pensions, worker's compensation, universal health care, occupational health and safety, employment standards have all come partly as a result of labor and community mobilizing.

A social justice orientation within social work has deep historical roots in mobilizing and organizing for progressive change (Benjamin, 2011). Throughout these and other milestones in workers' struggles, social workers as community advocates walked side by side with organized labor, a tradition that lives on today, for example, in the 'radical social work' movement in the United Kingdom (Lavalette, 2011).

During the Great Depression, a 'rank and file' movement of social workers emerged. Swelling to as many as 15 000 members by 1936, the movement was a force to be reckoned with. Rooted in the Marxist class analysis, the rank and filers identified themselves as white-collar workers and were strong supporters for the rights of white-collar workers to organize unions. Similarly, founders of the settlement house movement such as Jane Addams, Lillian Wald, and Florence Kelley established the National Women's Trade Union League in 1903 and worked closely with organized labor in advocating for strong, progressive social and public policy and services. As a result of these formidable alliances, far-reaching social and labor legislation – the Social Security Act and the National Labor Relations Act – were passed in the United States during this period (Scanlon and Harding, 2005).

However, after the post-WWII era, the relationship between organized labor and social work grew tenuous as unions partially retreated from activism during McCarthyism. And social workers also retreated for fear of being targeted politically. Many moved along the path of professionalization and 'cultivating' a middle-class professional identity. Instead of pursuing union organizing strategies, priorities were placed on the professionalization of social work, such as advanced clinical training, the development of professional journals, and the formation of professional association of social workers (Walkowitz, 1999). The resurgence of activism in the 1960s and 1970s only partially rekindled social work's reformist zeal. Today's social workers are, in general, more engaged in mainstreaming the social work profession and have left the notion of the 'worker' in social work behind. Others, of course, have retained a social justice orientation and got involved in the civil rights movement, the women's movement, antiwar protests, and, more recently, environmentalism and the movement for marriage equality and LGBT rights.

Community–Labor Alliances and Best Practices in the Mobilization for Social Justice

For the past three decades, some of the more progressive trade unions have recognized the need to build broad-based community coalitions as part of the resistance to the austerity agenda. Coalitions are able to provide a shield for community agencies participating on broader shared goals, without being targeted or running the risk of defunding. Peter Leonard refers to such coalitions as a 'confederation of diversity' where the politics of difference are incorporated with the politics of solidarity (Leonard, 1997; Mullaly, 2002). Four examples of successful coalition building are discussed below: the women's movement's involvement in trade unionism, the mobilization

for employment equity legislation, the minimum wage campaign, and the Solidarity City Campaign.

One key example of community–labor alliances has been through women who are active in both the labor and feminist movements. The influx of women into the public sector in the 1970s enabled activists in the women's movement to use their trade unions as another site to challenge gender discrimination in workplaces, unions, and the broader society. As a result of their commitment, progressive and inclusive policies such as childcare, pay equity, and initiatives to end violence against women were given voice and in many cases then legislated into broader social change.

In the late 1980s, trade unionists in Ontario, Canada, came together with women's organizations, people of color, people with disabilities and Indigenous Peoples to create the Alliance for Employment Equity. The advocacy work of this Alliance, which addressed systemic and discriminatory barriers to employment, led to the passing of the first ever mandatory employment equity legislation in 1993.

The successful mobilization around the minimum wage campaign in Ontario is another example of community mobilizing and political bargaining. The Toronto Workers Action Centre, a community-based agency providing advocacy services for low-wage, nonunionized workers, began the mobilizing and campaigning for a minimum wage raise after a 10-year freeze under a conservative regime. Their campaign was strengthened by a targeted mobilizing effort by the Toronto and York Region Labor Council in forging a broad-based coalition and groundswell of support (Vosko, 2012–13).

On 21 February 2013, the Toronto City Council passed a historic motion to provide services to residents without full immigration status or documents. This work was organized by members of Solidarity City Network which is composed of many community workers and activists from many frontline service providers, legal clinics and antipoverty organizations and immigrant rights advocacy groups.

The above examples of community and labor collaboration speak to the power in numbers. However, the mobilization of these alliances is not without tensions and contradictions when coalition partners have to work outside their comfort zones, being challenged to work across differences. Sometimes, partners come to the table for the sake of appearances while others (who come with financial contributions) can be overzealous in taking up space and claiming credit. Resolving some of these moments of tensions through critical antioppression social work practice has provided invaluable lessons of deepening relationships and solidarity building.

Lessons from the Los Angeles Immigrant Labor Activism

The successful collaboration between labor and community can be best demonstrated by the mobilization of Latino immigrants in Los Angeles (LA) County. Due to deep organizing over the past 15 years, LA has gained over 60 000 new, predominately Latino, unionized workers, bringing new energy and vision to the American labor movement (Milkman, 2012).

Ruth Milkman's (2012) study argues that the LA success in organizing immigrant labor is explained by three overlapping approaches to organizing. The first involves traditional trade union organizing. Several leading U.S. unions have organized Latino immigrants employed in low-wage janitorial, retail, and hospitality work. The Justice for Janitor campaign launched by Service Employees International Union (SEIU) is a prime example of a more innovative organizing approach of direct action and sectorial city wide coverage rather than being restricted to a single workplace.

The second approach within immigrant labor activism is the advocacy and organizing efforts of labor-oriented worker centers. These centers, which number more than a 100 nationwide, provide advocacy, support, and mobilization to workers in precarious employment or in the informal sector (Fine, 2006).

Finally there is the broader immigrant and migrant rights movement and its organizing. This too involves a broad-based coalition including partners from churches, ethnic organizations, community agencies, and organized labor. Foregrounded by the vision of human rights, the coalition has worked for legal status as the panacea to improve immigrant workers' employment and working conditions (Milkman, 2011).

The synergy of the three interlocking strategies has bonded immigrant workers into a cohesive group with a class identity grounded in their common experiences of exploitation and exclusion. It has dispelled the myth commonly held by unions that immigrant workers cannot be organized (Tattersall, 2009). Mobilization has also politicized workers and transformed members of the Latino community into a powerful political and voting bloc that U.S. political parties have had to take notice of.

Social Workers as Professionals and Union Members

The ideal of social work practice as a caring profession that fosters social justice has been greatly undermined by the rise of the new managerialism where service standardization and performance management models are now repacked and introduced into the social services sector as New Public Management (NPM) (Baines, 2011). With multitasking and casework overload, social workers are either being pushed to have more qualifications, or they experience deskilling and devaluing of their work at the other extreme. Dominelli (2004) urges for social workers to break from "the shackles of a government-imposed bureau-rationality that has turned them into bureau-technocrats."

Unionization provides an alternate forum for social workers to exercise their commitment not only to improve their own working conditions, but also to advocate broad public policy change. Collective bargaining has been effective in countering the pace of restructuring in the social services sector. For example, one community agency in Toronto has secured a no-contracting-out clause in their collective agreement and set a precedent in the sector.

Others have used the contract negotiation process to raise work restructuring issues that are detrimental to the quality of care and well being of their clients. Tara La Rose's (2009) study of the strike of a child welfare agency describes how the workers, who were members of the Canadian Union of Public Employees (CUPE), went on strike for 6 weeks after the management refused to address the issue of heavy caseloads

and high levels of stress experienced by their members. Through the bargaining and strike process, workers developed a strong bond, exercised their collective power as union members, challenged state managerial policies, and created positive social change (La Rose, 2009). The hard-won gains such as caps on caseloads and improved workload contract language have set a precedent for child welfare agencies across Ontario (Baines, 2011). The empowerment of these workers through the strike exemplifies the spirit of the union slogan – "collectively we bargain, individually we beg!"

Another successful example of collective resistance comes from a group of social workers who work for a social service agency in Toronto and belong to the Ontario Public Services Employees Union (OPSEU). Ben Carniol (2010) provides a vivid account of how Lisa Quarta, a clinical social worker and the President of her local, organized to stop management from dismissing housekeeping and clerical staff who were predominately racialized women, to cut costs. Quarta first called a union meeting where all affected workers shared their stories and reactions about the pending dismissals with other union members, who were mainly white professionals. As a result, staff members showed up to the subsequent meeting with management wearing black as a way of protesting. Two days later, management backed off the restructuring scheme. Not only did this collective action succeed in saving jobs, it also built solidarity and bridged the divide between racialized members who were in the lower, entry-level positions and white professional members (Carniol, 2010).

The challenges for trade unions representing social workers are twofold. Firstly, as a female-dominated sector, how can unions engage more social workers, when women social workers are often juggling the responsibility of caring for their families with overfull caseloads? Secondly, how do we create transformation at the local union level, where it is closest to the membership? There is great need for rank and file membership to experience transformative moments as union members (Bernard, 2006).

In labor organizing, a key goal is to persuade workers to think less about 'I' and more about 'We.' Organizing does not stop when the workers sign their union cards. They must also develop a sense of engagement and belonging. The following example from the Chicago Teachers Union offers insights on the practice of organizing from the inside outward to involve the wider public.

The Work Behind the Successful Chicago Teachers' Strike

On 10 September 2012, the Chicago teachers went out on strike for 9 days and were able to win public support to stop the austerity moves proposed by the Mayor. It took everyone, including the leadership of the labor movement, by surprise, but the seeds had been planted 2 years earlier when Karen Lewis was elected President of the Chicago Teachers' Strike (CTU) on a platform to "change this into a democratic union responsive to its members" (Sokolower, 2012).

Instead of employing a model of traditional leadership that stifles opposition, the leadership reached out and listened to the diverse opinions of the membership. The CTU rank and file activists were (and are) involved in their local school community, working with parents and other community members on specific issues. Through such involvement, a strong bond between parents and teachers was built. When the strike came, this bond was reflected in widespread parents' support for the teachers' action. This mobilizing approach is very much reflected in Karen Lewis' stance (Sokolower, 2012):

> It all comes down to how you teach people to fight with the tools they have. We have been fighting with the bosses' tools. We can spend a lot of time doing legislation. I think that's fine – have a legislative approach. But understand that you don't control that process….You need to have good relationships with legislators; you need to have members get in touch and let them know what's important to you. That's one tool. But it's not the only tool. Our best tool is our ability to put 20 000 people in the street. I don't care if one rich guy buys up all the ad space. The tool that we have is a mass movement. We have the pressure of mass mobilization and organizing.

What is interesting about this case is that community involvement and mobilization extended long past the resolution of the strike. Six months later, CTU and a coalition of students and parents staged a mass community protest against the closure of 54 schools affecting more than 30 000 students, primarily in low-income black and Latino areas. Networks of Faith, Community and Teachers for Students (FACTS) had sprung up throughout the city to stop the closures.

The Union's leadership also vowed to launch a voter education campaign and send organizers to neighborhoods where schools were slated to close and register a minimum 100 000 new voters. Such action speaks to what Mullaly has referred to as a political strategy to 'relegitimate the state' and "build more rather than less government involvement in social, economic and human affairs" (2002: p. 202). It underscores the importance of moving from the traditional practices of electoral politics to building broad-based coalitions where the inclusion of marginalized groups is meaningful and real. These activities allow a deeper relationship building whereby trust and commitment to each other come to be based on shared values and struggles.

Translating Progressive Core Values into Meaningful Practice

The progressive core values that trade unions and social work share is grounded in a critical analysis of the various ways that social structures, power dominance, and privilege are constructed and in learning to question who benefits from any particular set of social relations. It is also about finding new ways of challenging unjust social relations (Carniol and Del Valle, 2010) and transforming them into equitable processes based on economic, political, environmental, and social justice (Baines, 2011; Benjamin, 2011; Carniol and Del Valle, 2010; Dominelli, 2004).

In reimagining possible new convergences to resist the austerity agenda, it is critical to translate some progressive core values into meaningful practice. For further collaboration and bridging between trade unions and social work, I offer the following three ideas for consideration.

First is advice for trade unions about handling complaints and grievances. Very often, in the expediency of time and resources, union representatives end up resorting to the shortest route to resolve or mediate individual cases, without making deeper connections between the incidents and the larger structural context. Potential cases of harassment and/or differential treatment based on sexism, racism, and other grounds of discrimination are often complicated and difficult to prove. These cases require empathy, skills, and time for the local leadership to take such complaints seriously and respectfully. For a union member from one of the equity-seeking groups, who has finally mustered enough courage to step forward, the support and encouragement offered by a shop steward will make a world of difference and bring them closer to the union family. These can be powerful and transformative moments of union building.

A second recommendation is for the social work profession to reinforce the clarity of social justice and put the notion of *worker* back into social work. The primary entry point for this transformation is through social work education – by integrating social action organizing as a major element in the curriculum, and having a stronger presence of workers' struggles and trade unions in it. I would argue that social work student placements in unions should be expanded, to create a critical mass of such students, dispelling some of the myths about unions, and facilitating a deeper understanding and relationship building for future joint opportunities.

Last is the need to glean lessons from past community and labor collaboration. The conditions to deepen the roots of these community and labor alliances are more horizontal democratic practices that are bottom-up; a transformative leadership that has the political will and courage to challenge the status quo; a commitment to empowerment; and mobilizing strategies that will move people from the comfort of their own silos to the larger space of building the 'We.'

It is in the vision of rebuilding the 'We' that there is a need to probe deeper into the nuanced notion of community unionism and the complexity of movement building for social justice. If the end goal for community unionism is only to broaden the unions' reach to potential membership or to 'rebrand' unions in a more contemporary image, the effort will rapidly dissolve into self-interest and competition. Trade unions in social movements need to go beyond a narrow economistic focus and define themselves by demonstrating their capacity to connect with life experiences under capitalism (Fairbrother, 2008).

In the new norm of workplaces with precarious and informal labor, where workers are no longer confined to traditional workplace settings, it is important to build a repertoire of practice that will reach out to marginalized workers. Dae Oup Chang's study (2012) of the organizing efforts of workers in precarious employment in four East Asian countries proposes an alternate vision of the labor–social movement, where the anchoring is shifted from unions as the site of growth to a community where precarious and marginalized workers, regardless of status and workplace, all come together as an united force. In these community settings, social work practitioners can play a pivotal role in mobilizing to overcome fragmentation in the social sector. They can potentially represent a different path of coalition building that offers a more genuine partnership, empowerment, and innovative practice for social transformation.

Acknowledgments

The author thanks Ben Carniol, Pramila Aggarwal, Deena Ladd, Judy Chow, and David Kidd for their invaluable insights for the writing of this article.

See also: Critical Social Work Practice; Labor Unions; Marxist Approaches to Social Work; Neoliberalism; New Managerialism and New Public Sector Management; Public Sector Organizations; Radical Social Work; Social Action and Social Justice; Social Justice: Historical and Theoretical Considerations; Social Work Ethics; Trade Unions, Economic Behavior of; Urban Neoliberalism.

Bibliography

Baines, D. (Ed.), 2011. Doing Anti-oppressive Practice: Social Justice Social Work. Fernwood Publishing, Black Point, NS.

Benjamin, A., 2011. Doing anti-oppressive social work: the importance of resistance, history, and strategy. In: Baines, D. (Ed.), Doing Anti-oppressive Practice: Social Justice Social Work. Fernwood Publishing, Black Point, NS, pp. 289–297.

Berberoglu, Berch, 2011. The global capitalist crisis: its origins, dynamics, and impact on the United States. International Review of Modern Sociology 37 (2), 159–184.

Bernard, E., October/November 2006. Organizing the Organized. Our Times Magazine.

Bernhardt, A., Boushey, H., Bresser, L., Tilly, C. (Eds.), 2008. The Gloves-Off Economy: Problems and Possibilities at the Bottom of America's Labor Market. Labor and Employment Relations Association, Champaign, IL.

Briskin, L., 2006. Union leadership and equity representation. Gender and Work Database (online). http://www.genderwork.ca/Briskin_Leadership_Paper_April_2006.pdf.

Carniol, B. (Ed.), 2010. Case Critical: Social Services and Social Justice in Canada, sixth ed. Between the Lines, Toronto, ON.

Carniol, B., Del Valle, V., 2010. "We have a voice": helping immigrant women challenge in, abuse. In: Hick, S., Peters, H.I., Corner, T., London, T. (Eds.), Structural Social Work in Action: Examples from Practice. Canadian Scholars' Press Inc, Toronto, ON.

Chang, D., 2012. The neoliberal rise of East Asia and social movements of labour: four moments and a challenge. Interface: A Journal for and about Social Movements 4 (2), 22–51.

Dominelli, L., 2004. Social Work: Theory and Practice for a Changing Profession. Polity Press, Cambridge, UK.

Fairbrother, P., 2008. Social movement unionism or trade unions as social movements. Employee Responsibilities and Rights Journal 20, 213–220.

Fine, Janice, 2006. Worker Centers: Organizing Communities at the Edge of the Dream. Cornell University Press, Ithaca, NY.

Fisher, J., 1980. The Response of Social Work to the Depression. G.K. Hall, Boston, MA.

Fisher, R., Shragge, E., 2008. Challenging community organizing. Journal of Community Practice 8 (3), 1–9.

Fletcher Jr., B., Gapasin, F., 2008. Solidarity Divided: The Crisis in Organized Labor and a New Path toward Social Justice. University of California Press, Berkeley, CA.

Galabuzi, G.E., Block, S., 2011. Canada's Colour-Coded Labour Market: The Gap for Racialized Workers. Wellesley Institute and Canadian Centre for Policy Alternatives, Toronto, ON.

Hicks, A.M., Swank, D.H., 1992. Politics, institutions and welfare spending in industrialized democracies, 1960–1982. American Political Science Review 86 (3), 658–674.

Kumar, P., Schenk, C. (Eds.), 2003. Paths to Union Renewal: Canadian Experiences. Broadview Press, Peterborough, ON.

La Rose, T., 2009. One small revolution: unionization, community practice, and workload in child welfare. Journal of Community Practice 17 (1).

Lavalette, M. (Ed.), 2011. Radical Social Work Today: Social Work at the Crossroads. Policy Press, Bristol, UK.

Levesque, C., Murray, G., Le Queux, S., 2005. Union disaffection and social identity: democracy as a source of union revitalization. Work and Occupations 32 (4), 400–422.

Milkman, R., 2011. Immigrant workers, precarious work, and the US labor movement. Globalizations 8 (3), 361–372.

Milkman, R., 2012. Immigrant workers, precarious work, and the US labour movement. In: Munck, R., Schierup, C., Wise, R.D. (Eds.), Migration, Work and Citizenship in the New Global Order. Routledge, London and New York, pp. 113–124.

Mullaly, B., 2002. Challenging Oppression: A Critical Social Work Approach. Oxford University Press, Toronto, ON.

Ng, W., November 2012. Pedagogy of solidarity: educating for an interracial working class movement. Journal of Workplace Learning 24.

Scanlon, E., Harding, S., 2005. Social work and labor unions: historical and contemporary alliances. Journal of Community Practice 13 (1), 9–30.

Sokolower, J., 2012. Lessons in social justice unionism: an interview with Chicago teachers union president Karen Lewis. Rethinking Schools (online). http://www.rethinkingschools.org/archive/27_02/27_02_sokolower.shtml.

Tattersall, A., 2009. Using their sword of justice: the NSW teachers 161 federation and its campaign for public education between 2001 and 2004. In: McBride, J., Greendwood, I. (Eds.), Community Unionism: A Comparative Analysis of Concepts and Contexts. Palgrave Macmillan, Basingstoke.

Vosko, Leah F., 2012–13. "Rights without remedies": enforcing employment standards in Ontario by maximizing voice among workers in precarious jobs. Osgoode Hall Law Journal 50, 845ff.

Walkowitz, D.J., 1999. Working with Class: Social Workers and the Politics of Middle Class Identity. UNC Press, New York.

Trade Unions, Economic Behavior of

Alison Booth, Australian National University, Canberra, ACT, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

A trade union is an organized association of workers formed for the protection and promotion of their common interests. The standard view of unions is that they are monopoly organizations that improve the welfare of members, principally by raising wages above the competitive level. For a union to be able to increase wage rates above the competitive level, there must be some surplus that can be shared between the firm and the union, and the union must have some bargaining power to induce the firm to share this surplus. This article investigates the conditions under which a union can increase wages and explores ways of modeling the competing preferences of unions and management. The article also notes the arguments suggesting that, in the presence of imperfect information and uncertainty, unions may enhance efficiency. To the extent that unions reduce labor turnover and negotiating costs, they may increase the available surplus to be shared between parties.

Trade Unions

A trade union is an organized association of workers formed for the protection and promotion of their common interests. The standard view of trade unions is that they are monopoly organizations that improve the welfare of members, principally by raising wages above the competitive level. This article investigates the conditions under which a union can increase wages, and explores ways of modeling the competing preferences of unions and management.

Before beginning, we briefly discuss trends in two measures of union power – union density and union coverage – over the period 1990 to the most recent year for which these data are available. Table 1 shows the figures for the OECD countries.

While in 1990 trade union density averaged 40.1% across OECD countries, by 2009 it had declined to 28.5%. (These figures come from the OECD database on trade unions and Visser, 2011.) (Trade union density refers to the number of trade union members as a percentage of wage and salary earners.) While this is a large drop, the averages conceal considerable heterogeneity across countries. For many countries, the influence of trade unions at the macroeconomic level is better indicated by the extent of union coverage of the workforce, rather than by union density. (The definition of the collective bargaining coverage rate is the number of workers covered by wage bargaining agreements as a proportion of all wage and salary earners.) Across OECD countries, union coverage averaged 70% in 1990 and declined to 62% two decades later. Nonetheless, there are still some countries in which both density and coverage are low. For example, Table 1 shows that in the US only 11% of the workforce belong to a union and only 14% are covered by union collective bargaining.

Given that union density is declining in most industrialized countries, and that in many countries only a minority of the workforce is covered by unions, do we really need to worry about providing appropriate models of trade union behavior? One answer is that union bargaining models are generalizable and hence more widely applicable. Indeed, union contracts may be viewed simply as an explicit formulation of a wider variety of labor contracts that are found in many labor markets. Where nonunion firms face an incumbent workforce with a degree of bargaining power, management and workers may be in a situation of bilateral monopoly that can be

Table 1 Trade union density and collective bargaining coverage

	Trade union density		*Collective bargaining coverage*	
	1990	*Latest year*[a]	*1990*	*Latest year*[a]
Australia	39.5	18.0	80	40
Austria	46.9	28.1	98	99
Belgium	53.9	52.0	96	96
Canada	34.0	31.6	38	31.6
Czech Republic	43.5	17.3	–	42.5
Denmark	75.3	68.8	84	80
Estonia	–	8.0	–	19
Finland	72.5	69.9	81	90
France	9.9	7.6	92	90
Germany	31.2	18.5	72	62
Great Britain	38.2	25.8	54	32.7
Greece	34.1	24.0	70	65
Hungary	49.1	16.8	–	33.5
Iceland	92.9	79.3	96.4	88
Ireland	48.5	33.4	60	44
Italy	38.8	35.1	83	80
Japan	25.4	19.0	23	16
Luxembourg	46.4	37.3	60	58
Mexico	22.4	13.2	–	7
Netherlands	24.3	18.2	82	82.3
New Zealand	49.5	20.8	61	17
Norway	58.6	54.6	70	74
Portugal	27.9	19.3	79	45
Slovakia	64.2	17.2	–	40
Spain	12.5	15.9	82.2	84.4
Sweden	80.0	67.7	89	91
Switzerland	22.7	17.8	48	48
United States	15.5	11.3	18.3	13.6
OECD	40.1	28.0	70.3	62.1

[a]'Latest year' refers to the latest year for which data were available when searched May 2013.
Data obtained from: http://www.keepeek.com/Digital-Asset-Management/oecd/economics/oecd-economic-surveys-sweden-2012/trade-union-density-and-collective-bargaining-coverage_eco_surveys-swe-2012-graph36-en.

characterized by a union-firm bargaining model. Moreover, the *threat* of union organization of a nonunion sector may provoke management to provide wages and working conditions that mimic those negotiated in union firms.

Conditions Under Which a Union Can Improve Workers' Welfare

For a trade union to be able to increase wage rates above the competitive level, there must be some surplus that can be shared between the firm and the union, and the union must have some bargaining power to induce the firm to share this surplus.

Bargaining Power

Union bargaining power derives from the ability of the union to inflict damage on the firm by withdrawing labor. This works when there is no substitute pool of nonunion workers, for example, if there are high labor turnover costs or if the union controls labor supply through entry restrictions or through having high levels of membership relative to the size of the sector. Union membership density may play a part in determining the power of the union to extract from the firm a share of any rents. As membership density increases, there is a shrinking pool of workers available to the firm should it choose not to employ union labor. Thus one would expect to find a minimum critical level of membership below which firms would refuse to recognize and bargain with the union (Osborne, 1984). Where membership density is crucial to the union's monopoly power as, for example, with the historical emergence of the general unions, the union must consider how to attract workers at the margin. The basic problem for economists in explaining why workers join a union relates to the free-rider problem, arising because of the collective nature of many of the goods provided by the union (Olson, 1965; Booth, 1985). Why join the union when union dues are costly, and when the benefits are not immediately available or apply to all workers regardless of their union status? Of course established unions may also provide incentive private goods that encourage workers to unionize. Indeed, a union can act as the workers' agent, since it can reduce the costs of negotiation and processing of grievances, a role which may also be useful to employers (Faith and Reid, 1987).

Is There as Surplus to Be Shared?

Even if a union controls labor supply, it will not necessarily be able to negotiate a large wage increase relative to the competitive level. The magnitude of the union wage depends crucially on the elasticity of labor demand in that sector. There can be little doubt that, if unions emerge in competitive markets, high union wages introduce allocative inefficiencies into the economy through the distortion of factor prices. Without unions, allocative efficiency is associated with the equalization of the marginal products of identical factor inputs across sectors. With higher wages in the union sector, union firms employ fewer workers. Displaced union workers crowd into the nonunion sector, lowering wages there. As a result, too few workers are employed in the union sector where output falls, while too many workers are employed, and too many goods produced, in the nonunion sector. There is a deadweight efficiency loss, because the value of marginal products in the two sectors is not identical. In addition, there are distributional issues to consider, and longer run effects due to substitution of capital for labor. Furthermore, in unionized sectors there may be under-investment in capital through the hold-up problem (see, for example, Grout, 1984).

This discussion of allocative inefficiency assumes that the union has emerged in an economy characterized by competitive product and labor markets. But there is considerable evidence that, in modern industrialized countries, many product markets are characterized by imperfect competition. Moreover, nonunion firms may also face an incumbent workforce with a degree of bargaining power (Lindbeck and Snower, 1988). Thus, even in the absence of unionization, management and workers may be in a situation of bilateral monopoly. The majority of formal trade union models in the literature assume a perfectly competitive product market to allow the models to focus on wage and employment determination in the simplest environment (exceptions are, inter alia, Dowrick, 1989; Dobbelaere and Mairesse, 2013). Nonetheless, it is an empirical regularity that imperfections in product and labor markets are correlated (Stewart, 1990).

Even within the competitive model, there are arguments suggesting that, in the presence of imperfect information and uncertainty, unions may enhance efficiency. To the extent that unions reduce labor turnover and negotiating costs, they may increase the available surplus to be shared between parties (Freeman and Medoff, 1984). Of course, there may be interdependence between the monopoly and efficiency roles of trade unions: unless the union has some bargaining power, it may be unable to increase efficiency.

Formal Modeling of Union Behavior

To investigate union behavior, the economist needs to model the objectives of the relevant actors involved in wage setting, as well as the means of resolving their potentially conflicting objectives. Most union models consider the objectives of two agents – the union and management. Typically, the government's behavior is assumed exogenous. Because the objectives of union and management may be conflicting, models need to consider how conflicts might be resolved within the constraints provided by the institutional environment. The range of issues covered by collective bargaining in Europe is narrower than in the USA, which probably reflects a relative lack of US social welfare legislation. (In Europe, legislation typically covers issues such as health insurance, pensions, redundancy pay, and unfair dismissal that are often covered by collective agreements in the US.)

While bargaining may take place over a number of issues, for tractability researchers typically concatenate wages and other forms of compensation, such as fringe benefits, under the single heading of wages. Union preferences appear to include employment as well as wages, but it is rare to find instances of the union and the firm bargaining over both wages and employment. Of course, unions affect employment

indirectly, not only through increasing wages but also through negotiation over staffing levels, recruitment, and redeployment.

While unions typically have a number of goals (from increasing remuneration to broad political aims concerning the labor movement as a whole), the orthodox union model assumes unions are concerned only with the unionized sector. But a union typically comprises heterogeneous individuals with varying and possibly conflicting preferences as to the union's strategy. For example, there may be principal–agent problems arising because union leaders (agents) have objectives that may conflict with the membership (the principal). The leadership may be better informed than its membership, or perhaps interested in self-aggrandizement or in wider labor goals. Unions are usually large, with complex organizational structures concerned with representation of members' interests and with administration. Because the internal organization of the union is typically hierarchical in structure, the question arises as to how well members' preferences are transmitted through the various levels, each of which may have conflicting aims. Moreover, there is heterogeneity across unions in terms of their structure and constitution (ranging from the democratic through to the totalitarian) that may determine a union's maximizing strategy.

While there is a small literature allowing for union objectives to be affected by membership heterogeneity or a conflict of interests between leaders and members (Farber, 1978; Blair and Crawford, 1984), the widely used models of union behavior side-step these problems by supposing that all workers in the union sector are identical and are union members. They also assume that the union cares only about the economic welfare of the union sector. For simplicity, union-negotiated benefits are assumed to be encapsulated in the wages variable. Because the union is also hypothesized to care about the employment of union workers, union objectives can be specified as an increasing function of wages w and employment n.

In most union literature, the union is assumed to maximize either a utilitarian objective function, or an expected utility objective function. The utilitarian objective function assumes that the union maximizes the sum of the individuals' utilities. The union comprises identical members treated identically by the union, and therefore in this context no distributional or normative judgments are being made by the utilitarian objective. However, if workers were assumed heterogeneous, the use of this as the union objective would introduce normative judgments (since if individuals were not identical, it would be necessary to consider why the utilitarian social welfare should be adopted instead of some alternatives such as the Rawlsian or maxi–min criterion, or a more egalitarian approach). If the union's wage-setting behavior results in some union members being unemployed, the utilitarian objective function is:

$$U = nu(w) + (t-n)u(b) \quad w > b;\ 0 < n \le t \qquad [1]$$

where $u(.)$ is a concave (indirect) utility function, w denotes wages, n denotes employment in the union sector, t is the total number of union workers in the sector (and membership by assumption), and b is the alternative sector wage or the unemployment benefit.

The second popular union objective function is the expected utility function. If membership is fixed, eqn [1] can be written equivalently in terms of expected utility. Since the union raises wages above the competitive level, each member faces some probability $(1-n/t)$ of being unemployed. If unemployed, a worker receives the alternative sector wage or the unemployment benefit. The expected utility of a representative union member is given by:

$$E(U) = \frac{n}{t}u(w) + \left(1-\frac{n}{t}\right)u(b) \quad u'(w) > 0;\ u''(w) \le 0 \qquad [2]$$

where $u'(w) = \mathrm{d}u/\mathrm{d}w$. The slope of the union indifference curve in (w, n) space is found by totally differentiating either eqn [2] or [3] with respect to w and n, holding (expected) utility fixed, yielding $\mathrm{d}w/\mathrm{d}n = -[u(w) - u(b)]/nu'(w)$.

Both the utilitarian and the expected utility objective functions have the implication that identical workers receive different outcomes *ex post*. Employed members receive the union wage, while the unemployed workers receive b. However, if unemployed members effectively leave the union, the leadership will be concerned about future union viability, since there will be a larger pool of nonunion labor that might undercut the union wage, and union income from dues would fall. This is likely to affect union outcomes, and is an argument for inclusion of the level of membership in the union objective function as an endogenous variable. Moreover, b may not be exogenous – the union may affect remuneration for unemployed members through redistribution by benefits or redundancy pay, suitably adjusted for the disutility of work for those in employment. Indeed, historically, many unions paid unemployment benefits to out-of-work members, and some commentators argue that this was important for union growth and survival.

Now consider the behavior of the firm. The necessary conditions for a union to be able to appropriate any surplus in a competitive product market, without driving the firm out of existence, are very high levels of unionization in a competitive industry, no nonunion foreign competition, entry barriers, and inelastic product demand. Without the first three conditions, domestic firms, foreign firms, or new firms entering the industry could simply hire nonunion labor and produce at a lower cost than unionized competitors, thereby driving them out of business. Assuming the firm maximizes profits and has no power in the labor market (which means it will take wages as given), it has just one choice variable – employment. The optimization problem facing the firm is therefore:

$$\max_n \pi = pq(n) - wn\ q'(n) > 0;\ q''(n) < 0;\ q(0) = 0 \qquad [3]$$

where π denotes profits, p denotes product price, and $q(n)$ is the production function characterized by diminishing returns to labor. The firm's wage bill is wn and capital is held fixed. The slope of the firm's isoprofit curve in (w, n) space is found by totally differentiating eqn [3] with respect to wages and employment keeping profits fixed, to obtain $\mathrm{d}w/\mathrm{d}n = [pq'(n) - w]n$. The locus of the turning point of each isoprofit curve traces out the labor demand curve where $pq'(n) = w$, as illustrated in **Figure 1**. Two isoprofit curves IP_0 and IP_1 are illustrated in **Figure 2**, together with two union indifference curves IC_0 and IC_1.

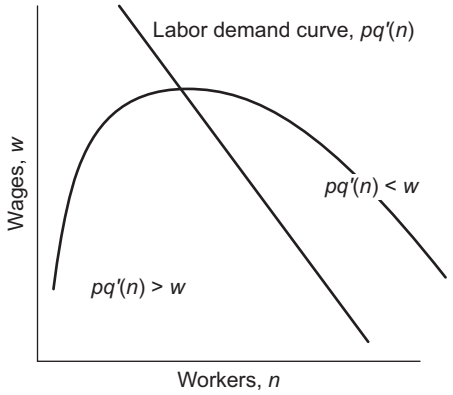

Figure 1 The isoprofit curve.

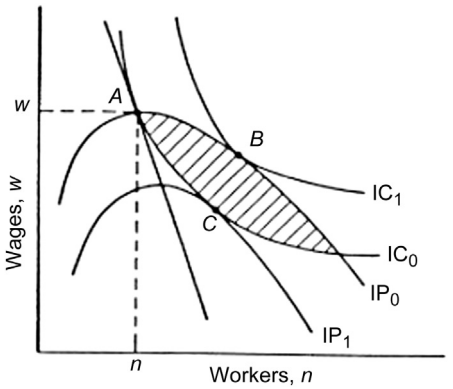

Figure 2 Inefficiency of the monopoly union model.

The Monopoly Union Model

In the monopoly union model, there is no bargaining between the two parties. Instead, the union acts as a monopolist in the supply of labor and imposes a wage rate on the firm. The firm, however, retains its managerial prerogative to determine the number of workers employed at the union wage. Once the union sets the wage, the firm reads off from the labor demand curve the number of workers to hire at that wage. Thus, a union maximizing the expected utility of its representative worker sets wages as follows:

$$\max_n E(U) = \frac{n}{t}u(w) + \left(1 - \frac{n}{t}\right)u(b) \quad \text{subject to } pq'(n) = w \qquad [4]$$

The first-order condition can be found directly by equating the slopes of the union indifference curve (giving the union's marginal rate of substitution of employment for wages) with the slope of the labor demand curve, illustrated as point A in Figure 2. The monopoly union seeks its highest level of utility, subject to the constraint of the firm's labor demand curve, where w^* and n^* are the optimal wage and employment levels, respectively. It can be seen, from the first-order condition in eqn [5], that wages will be set by the union such that the percentage increase in a member's utility due to a percentage increase in wages is exactly equal to the wage elasticity of labor demand ε. Intuitively, this result arises because an increase in wages reduces union employment, and thus each worker faces a greater probability of being unemployed. On the other hand, any member lucky enough to be employed receives a higher wage. Thus, in equilibrium, the percentage marginal benefit is equal to the percentage marginal cost:

$$\varepsilon \equiv -\frac{n'(w)w}{n} = \frac{wu'(w)}{[u(w) - u(b)]} \qquad [5]$$

This equation simply characterizes the equilibrium situation. To obtain a specific value for wages, it is necessary to choose explicit functional forms for the utility function and the labor demand function that can then be inserted into the generalized characterization and solved for the wage rate. For evaluation of empirical work attempting to test this model, see Pencavel (1991) and Booth (1995). For a more recent empirically relevant formulation of the union objective function see Pencavel (2009).

Using this model, it is straightforward to show how the wage set by the monopoly union alters due to a change in each of the explanatory variables. An isoelastic increase in labor demand has no effect on wages, but employment increases. An improvement in alternative opportunities increases the union wage, but employment is reduced as we move up the labor demand curve. An increase in membership has no effect on wages or the numbers employed. The empirically testable predictions of the monopoly union model are first that the alternative wage and unemployment benefit should affect wages directly. But employment should be determined by the wage rate, and unaffected by variables measuring alternative opportunities. Second, membership changes should have no impact on the union wage.

The Bargaining Approach

Next consider more sophisticated methods of reconciling the different preferences of union and management, using a bargaining approach. It turns out that the monopoly union model is nested as a special case within a simple bargaining framework. While some researchers argue that the monopoly union model is a reasonable approximation for particular industries at particular times, observation suggests that wages are more often bargained over by unions and firms. When union and firm bargain over wages, they are in a bilateral monopoly situation, in which a single seller confronts a single buyer of labor. A vital element in this bargaining is the ability of each agent to impose costs on the other agent if it refuses to bargain and engages in a strike or lockout. Although strikes are relatively infrequent, this does not mean that the strike or lockout is unimportant. It is the threat to withdraw that counts. What induces the two parties to bargain is the threat of facing costs imposed by the other party.

Since the union and the firm are in a situation of bilateral monopoly, additional structure must be imposed on the model to generate a determinate solution of the division of the available surplus between the two parties. There are two broad approaches to modeling bargaining behavior – the axiomatic

or Nash bargaining approach, and the more recent game theoretical approach (Nash, 1950). Under certain conditions, these generate identical predictions in spite of the fact that they proceed from very different assumptions (Binmore et al., 1986).

The solution concept widely used in the union literature is the generalized Nash bargaining solution, according to which wages are determined by maximization of the product of each agent's gains from reaching a bargain, weighted by their respective bargaining strengths. Since management is assumed to retain the right to determine employment unilaterally, this is referred to frequently as the 'right-to-manage' model. Define a status quo or fall-back point for each party as its position if no bargain is reached. For the firm, the status quo is zero, because if it does not reach a bargain with the union workforce, it cannot obtain any other workers since there are no nonunion workers in the sector. So the firm's net gain is simply its profits function. For the union, its status quo for a representative member is $u(b)$, since that is what a member will receive if no bargain is reached. Thus, the net gain to the union is simply the expected utility of a representative member (see eqn [2]) less fall-back utility, given by $[u(w) - u(b)]n/t$.

The generalized Nash bargaining solution (the product of the weighted net gains to each party) can be written as:

$$\max_{w} B = \left\{\frac{n}{t}[u(w) - u(b)]\right\}^{\beta}\{pq(n) - wn\}^{1-\beta} \qquad [6]$$

where β measures the bargaining strength of the union and $(1 - \beta)$ is the bargaining strength of the firm. The solution to this bargaining problem is obtained from the first-order condition of eqn [6], and is given by:

$$\frac{\beta w u'(w)}{u(w) - u(b)} = \beta\varepsilon + \frac{(1-\beta)wn}{pq(n) - wn} \qquad [7]$$

where ε is the wage elasticity of labor demand. The left-hand side of eqn [7] represents the proportional marginal benefit to the bargain from a proportional increase in the wage. Since only the union receives this benefit, it is weighted by the union's bargaining power β. The first term on the right-hand side is the union's proportional marginal cost (the percentage reduction in employment due to the proportionate wage increase) weighted by union power β. The second term on the right-hand side represents the firm's proportional marginal cost, weighted by the firm's power $(1 - \beta)$. If the union has all the power, $\beta = 1$, and the union extracts the entire surplus, the right-to-manage model collapses to the monopoly union model. For $0 < \beta < 1$, the employment and wage outcome will lie on the labor demand curve anywhere between the monopoly union extreme and the competitive outcome. If the firm has all the power, $\beta = 0$ and the wage is then simply the alternative wage b.

The empirically testable predictions of the right-to-manage model are the same as a monopoly union model with the addition of the prediction that variables proxying union strength will affect wages directly.

Is the right-to-manage model Pareto-efficient? (Pareto efficiency is defined by economists as a situation where no one can be made better off without making someone else worse off.) Or can at least one party be made better off by shifting to some other (w, n) pair? In **Figure 2**, the isoprofit curves IP represent different levels of profits for the firm, with higher curves denoting lower profits since they are associated with higher wages. IC denotes the union indifference curves, with expected utility increasing as the union moves northeast. Suppose the equilibrium of the right-to-manage model is at point A. The union could be made better off while keeping the firm on the same isoprofit curve, by shifting to point B. Alternatively, the firm could shift to a higher isoprofit curve, while keeping the union on the original indifference curve, by moving to point C. Both parties could be made better off by shifting from A to any point in the shaded area. Positions on the labor demand curve are thus Pareto-inefficient. Either the union or the firm could be made better off by bargaining over employment as well as wages, allowing them to shift from a point such as A to B or C or anywhere in the shaded area. Efficiency in the context of this model is where the marginal rates of substitution of employment for wages, for both union and firm, are equal.

The Efficient Bargaining Model

Now suppose the union and firm simultaneously determine wages and employment. While evidence suggests that unions and firms do typically not bargain over wages and employment, this approach has been important in the union literature because of its efficiency property beloved by economists. But the fact that this does not appear to happen in practice suggests that there may be something economists are missing in their modeling. (Of course the wage and employment outcome predicted by the efficient bargaining model is inefficient from the viewpoint of society as a whole.)

Efficiency requires only that one party's welfare be maximized subject to an arbitrarily fixed level of welfare of the other. This makes it possible to characterize an efficient outcome simply by considering union and firm preferences, without imposing a particular bargaining structure. While this does not yield a unique outcome, it illustrates the necessary conditions for efficiency. An efficient bargain over (w, n) solves:

$$\begin{aligned} &\max_{w,n} pq(n) - wn \quad n \geq 0;\ w \geq b \\ &\text{subject to } \frac{n}{t}[u(w) - u(b)] + u(b) = U \end{aligned} \qquad [8]$$

From the first-order conditions is obtained the result that an efficient (w, n) pair is given by equating the marginal rates of substitution of employment and wages for the union and the firm. This means equating the slopes of the isoprofit and union indifference curves, illustrated as point C in **Figure 2**, and shown below:

$$pq'(n) - w = -\frac{[u(w) - u(b)]}{u'(w)} \qquad [9]$$

The contract curve is defined as the locus of all points satisfying eqn [9]. Two points on the contract curve are B and C in **Figure 2**. The contract curve is positively sloped for risk-averse workers (but vertical if workers are risk-neutral). From eqn [9], note that the value of the marginal product of labor $pq'(n)$ is less than the wage rate, by an amount that is equal to the marginal rate of substitution of employment for wages. Hence, the contract curve lies to the right of the labor

demand curve for $w > b$. If $w = b$, then from eqn [9] $pq'(n) = w$, the perfectly competitive result. Thus, the contract curve and the labor demand curve coincide where $w = b$. Since for efficiency we only require that one party's welfare be maximized subject to any arbitrarily fixed level of welfare of the other, any point on the contract curve represents an efficient outcome for a particular arbitrarily fixed level of welfare of one agent.

To determine a unique (w, n) pair, a means of obtaining a particular solution to the bargaining procedure must be added to the model. A common approach is to use the generalized Nash bargaining model of eqn [6] modified to allow for bargaining over both w and n. The solution to this maximization problem generates a determinate outcome for wages and employment, which lies somewhere on the contract curve (Leontief, 1946; McDonald and Solow, 1981). In the 1980s considerable attention was devoted to devising ways of testing which model was appropriate for the union sector (Pencavel, 1991).

Many unions are unable to bargain over employment directly, and thus the relevance of this fully efficient model might be questioned. However, unions can often affect employment indirectly, for example, through bargaining over the labor-to-capital ratio, work intensity, or redundancy payments. Johnson (1990) found that, in models in which there is bargaining over wages and the labor-to-capital ratio or work intensity, the wage–employment outcome lies somewhere between the fully efficient and the right-to-manage model.

Conclusions

This article can only give a flavor of the economics behavior of the trade union, and many important topics have been omitted; for example, the impact of unionization on the macroeconomy (Layard et al., 1991; Boeri et al., 2001). Some countries (such as many European economies) have centralized wage bargaining arrangements, while others such as the US are decentralized. It is important that in future work these institutional differences are incorporated into the modeling framework, as discussed in more depth in Boeri et al. (2001) and Cahuc and Zylberberg (2004).

Other important topics omitted from this survey (but covered in Booth, 1995) include the economics of strikes (*see* Dispute Resolutions in Economics), the relationship between unions and hours of work, and the modeling of endogenous union membership, an important topic in countries such as the UK, in which compulsory union membership is illegal. There is also a vast empirical literature estimating the impact of unions on economic performance.

While union membership and power have declined in many industrialized countries from the 1980s (for example, Britain and the US), it is unlikely that unions will vanish all together, especially if they adopt creative coping strategies (Boeri et al., 2001: Chapter 6). Moreover, the evidence suggests that as male membership declines, more women are unionizing. It is also possible that in the future unions may become stronger in Eastern European, Asian, and African countries. It is therefore important for researchers working on the economics of trade unions to consider what form of unionism might be least distortionary in a modern economy. Furthermore, while many models of union behavior refer to explicit labor contracts between a union and management, the models are also relevant to a wider class of situations than those in which a union explicitly represents workers. The models are potentially applicable, with only a few minor modifications, to nonunion labor markets where incumbent workers have some market power and where the firm has economic rents that might be shared.

See also: Anthropological Approaches to the Study of Education: The United States and Beyond; Economic Geography; Employment and Labor, Regulation of; Labor, Division of; Unemployment: Structural; Work, Sociology of.

Bibliography

Binmore, K., Rubinstein, A., Wolinsky, A., 1986. The Nash bargaining solution in economic modeling. RAND Journal of Economics 17 (2), 176–188.
Blair, D.H., Crawford, D.L., 1984. Labor union objectives and collective bargaining. Quarterly Journal of Economics 99 (3), 547–566.
Boeri, T., Brugiavini, A., Calmfors, L., 2001. The Role of Unions in the Twenty-first Century. Oxford University Press, Oxford.
Booth, A.L., 1985. The free rider problem and a social custom theory of trade union membership. Quarterly Journal of Economics 100 (February), 253–261.
Booth, A.L., 1995. The Economics of the Trade Union. Cambridge University Press, Cambridge.
Cahuc, P., Zylberberg, A., 2004. Chapter 7: Collective bargaining. In: Labor Economics. MIT Press, Cambridge, MA.
Dobbelaere, S., Mairesse, J., 2013. Panel data estimates of the production function and product and labor market imperfections. Journal of Applied Econometrics 28, 1–46.
Dowrick, S., 1989. Union-oligopoly bargaining. Economic Journal 99, 1123–1142.
Faith, R.L., Reid, J.D., 1987. An agency theory of unionism. Journal of Economic Behavior and Organization 8, 39–60.
Farber, H.S., 1978. Individual preferences and union wage determination: the case of the united mine workers. Journal of Political Economy 86 (5), 932–942.
Freeman, R.B., Medoff, J.L., 1984. What Do Unions Do? Basic Books, New York.
Grout, P.A., 1984. Investment and wages in the absence of binding contracts. Econometrica 52 (2), 449–460.
Johnson, G.E., 1990. Work rules, featherbedding and pareto-optimal union–management bargaining. Journal of Labor Economics 8, S237–S259.
Layard, R., Nickell, S.J., Jackman, R., 1991. Unemployment Macroeconomic Performance and the Labor Market. Oxford University Press, Oxford.
Leontief, W., 1946. The pure theory of the guaranteed annual wage contract. Journal of Political Economy 54 (1), 76–79.
Lindbeck, A., Snower, D.J., 1988. The Insider–Outsider Theory of Employment and Unemployment. MIT Press, Cambridge, MA.
McDonald, I.M., Solow, R.M., 1981. Wage bargaining and employment. American Economic Review 71, 896–908.
Nash, J.F., 1950. The bargaining problem. Econometrica 18 (2), 155–162.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, MA.
Osborne, M.J., 1984. Capitalist–worker conflict and involuntary unemployment. Review of Economic Studies 51, 111–127.
Pencavel, J., 1991. Labor Markets under Trade Unionism. Blackwell, Oxford, UK.
Pencavel, J., 2009. How successful have trade unions been? A utility-based indicator of union well-being. Industrial and Labor Relations Review 62 (2), 147–156.
Stewart, M.B., 1990. Union wage differentials, product market influences and the division of rents. Economic Journal 100, 1122–1137.
Visser, J., 2011. http://www.uva-aias.net/208.

Tradition, Anthropology of

Richard Bauman, Indiana University, Bloomington, IN, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15819–15824, © 2001, Elsevier Ltd.

Abstract

Few concepts have played a more central role in the development and practice of anthropology than tradition. This article opens with an outline of the principal senses in which tradition figures in anthropological thought: as cultural substance, dynamic process, social practice, and authorizing force. The second section traces the historical foundations of these contemporary usage's in enlightenment philosophy, Protestant reformism, antiquarianism, classical and biblical philology, the Herderian synthesis, Romanticism, and the nineteenth- and early twentieth-century typological schemes of Tönnies, Durkheim, Weber, etc. The final section sets out the principal foci of empirical, conceptual, and critical interest in tradition in twentieth-century anthropology: the social organization of tradition; the dynamic inter-relationship between tradition and creativity (especially in the anthropology of art); the relationship between tradition and communicative technologies (especially orality and literacy); and the politics of tradition.

Few concepts have played a more central role in the development and practice of anthropology than tradition. We will first outline the principal senses in which tradition figures in anthropological thought, then trace the historical foundations of these contemporary usages, and conclude with the major lines of inquiry centering on tradition in contemporary anthropology.

Tradition as Substance, Process, Practice

In its most general usage, tradition may be defined as the element of historical continuity or social inheritance in culture, or the social process by which such continuity is achieved. More substantively, tradition refers to the collective social inheritance of particular people, culture, society, group, or collectivity, and as such, stands as the referent of its collective identity. In this sense, tradition is generally conceived of as the aggregate mass of cultural forms – cast as knowledge, beliefs, practical routines, or texts – that are the constituent elements of a culture, handed down from one generation to the next. One may speak also of 'a tradition' as any discrete element or trait within the general mass. Here, tradition takes on the sense of an inherent quality of the past which is diagnostic of traditionality; hence, 'a traditional practice,' 'a traditional belief,' 'a traditional tale,' and so on. So conceived, a tradition assumes the guise of a durable natural object, passed down through time by successive generations of 'culture bearers.'

A more concrete and practice-centered understanding of tradition recognizes that the temporal continuity of tradition is instantiated in successive enactments of cultural forms that are guided by precedent and convention and construed as replicating, in some essential sense, what has been done before. When the focus is on the dynamics of tradition, the process of handing down, or intergenerational transmission, the element of social learning is often emphasized, whether under the rubric of enculturation, which foregrounds the acquisition of cultural knowledge, or the now preferred socialization, which foregrounds the social matrix of the acquisition process.

The conjoint consideration of socialization and tradition tends to foster a conception of intergenerational transmission as the replication of uniformity, in Anthony F. C. Wallace's (1970) apt phrase. This orientation points to what is perhaps the most common theme in anthropological considerations of tradition, that is, its normative force, the notion that tradition enjoins conformity. The binding force of tradition is variously attributed to the force of precedent or habit inherent in the routinized quality of replicated practice, the psychological pressure attendant upon deviating from accustomed modes of thought and action, the moral weight of public opinion, and – most strongly – the power of a coercive apparatus of external sanctions sustained by sacred mandate (as from the ancestors) and intensified by ritual enactment. Where such discriminations are made systematically, they are generally marked by terminological distinctions, as between custom and tradition, with tradition reserved for the most strongly sanctioned and ritualized elements of the social inheritance. In general, however, there is little definitional rigor or regularity of terminological usage in the field; as often as not, terms such as tradition, custom, or folkways are employed as loosely equivalent. In such systematization, again, tradition may be distinguished from law, the former deriving its force from 'the sanctity of immemorial traditions,' the latter from 'the legality of enacted rules,' in Max Weber's classic formulation. Others, however, such as Bronislaw Malinowski, take custom and tradition as more or less equivalent, but argue that certain traditions, those surrounded by supernatural sanctions and sustained by coercive machinery, are the functional equivalent of law in 'primitive' societies.

Perspectives also vary in regard to how they frame the locus of traditional authority. One line of inquiry focuses on epistemological issues, emphasizing the binding force of traditional modes of thought, characterized by a purported resistance to or incapacity for rational, reflective, critical, or skeptical stances toward knowledge and belief. In a related vein, the emphasis may fall on the behavioral correlates of such epis-temological conservatism, manifested in traditional patterns of 'tried and true,' naturalized modes of action and in ritual. A third line maintains a discursive focus, concentrating

on authoritative discourse; here, authority is vested in the weight of ancestral and collective truth expressed in oft-repeated, socially resonant texts, such as proverbs, myths, or legends. Finally, we may identify a sociopolitical frame of reference, which is primarily concerned with traditional authority as vested in particular statuses or roles, such as elders (by virtue of their accumulated knowledge and the historical reach of their experience into the past) or ritual specialists (who control the coercive machinery of supernatural sanctions), whose claims to authority rest upon their connection to an authorizing past. These distinctions are analytical; much work, both theoretical and substantive, explores dimensions of interrelationship among these loci of traditional authority.

Tradition and the Construction of Modernity

The roots of the anthropological preoccupation with tradition lie in the seventeenth and eighteenth centuries, in the epistemological challenges of enlightenment empiricism, the religious overturning of the Protestant Reformation, and the interpretive explorations of classical and biblical philology. Enlightenment philosophers, such as Bacon and Locke in the UK and Condorcet in France, insisted that the rational, empirical pursuit of true knowledge required a throwing off of traditional authority, which demands the surrender of one's capacity for independent thought and renders one incapable of making new contributions to knowledge.

Baconian natural philosophy and the concomitant enlightenment ideology of progress were foundational to the development of antiquarianism in the late seventeenth and eighteenth centuries. Antiquities were the etiolated remnants of past ways of thinking, acting, and conveying knowledge that were 'derived down' uncritically into the present, shorn of their original meaning and functional coherence. In this sense, the notion of antiquities anticipates E. B. Tylor's concept of survivals. Only by rigorous historical investigation of their origins could they be fully revealed as irrational survivals of 'old customs' and expunged from modern life. For Protestant reformers, the remnants of old 'heathen' and 'Romish' 'superstitions,' handed down to the credulous members of the populace by 'custom,' 'tradition,' and 'hearsay,' were likewise to be revealed by historical investigation and extirpated from modern life. There was a clear sociological component to the antiquaries' and reformers' antiquarian programs as well, identifying the 'vulgar' classes, country people, the elderly, the uneducated, and women (as in 'old wives' tales') as the segments of the population among whom the survivals of past error were most likely to persist, thus, in effect, constructing a set of social 'Others' within contemporary society who were to be the special targets of tradition-oriented antiquarian investigation.

The development of eighteenth-century classical and biblical philology was stimulated by the literary battles of the Ancients and the Moderns, contesting the proper basis of contemporary literary production. This line of inquiry had a textual and interpretive focus: to render ancient and exotic texts, like the Homeric epics and the Hebrew Old Testament, intelligible, coherent, and meaningful to contemporary readers and to secure their canonical status and value. This interpretive approach combined a kind of protorelativism, reading the texts as shaped by the social, cultural, and linguistic milieu of ancient Greece or Israel – that is, as the inspired expressions of a nation – with a species of conjectural history, in terms of which the ancient texts were characteristic expressions of an early stage in the development of human society. One of the most significant methodological consequences of this approach was to motivate the use of data from living cultures – chiefly peoples of 'the Orient,' such as Arabs, Turks, East Indians – viewed as developmentally behind their modern European contemporaries, to illuminate the archaic cultures reflected in the Homeric epics and the Old Testament. These peoples of the East, then, took on the guise of exotic 'Others,' living exemplars of a 'rude' and 'primeval' age, bound by tradition rather than by an ideology of progress, reliant on oral communication rather than literacy, credulous and emotional rather than rational, and so on.

In a sweeping synthesis and extension of the philologists' program and a critical counterstatement to the social contract theories of enlightenment political philosophers, the German political philosopher, Johann Gottfried Herder argued that the authentic foundation of a viable polity is the poetic tradition of its people, made up of the inspired expressions of its poets – who are also its 'lawgivers' – who give voice to the spirit of the Volk, rooted in its particular time and place. This intertextually constituted tradition became 'the archive of the people,' the principal expression of national identity, the mechanism of its cultural cohesion and continuity, and the charter for its social and political life. In the most general sense of the term, Volk designates a nation, a people in its organic wholeness, but it may also designate that portion of a more complex, stratified society that remains grounded in its inherited language and traditions and still open to poetic feeling, as distinct from those who have been distanced from their traditions and feelings by over-rational refinement, the alienation of writing, or the adoption of foreign languages and alien ways. In Herder's (see Bauman and Briggs, 2000) intellectual synthesis, then, a precursor to Boasian anthropology, tradition is most powerfully operative in simpler, premodern, community-based societies and in those sectors of complex modern society in which a sense of community cohesion, the emotionally resonant ties of locality, kinship, and attunement to the national spirit still prevail.

Herder's thought was influential in the emergence of Romanticism, which was also inflected by the violent upheavals of the French Revolution. One response to the Revolution took the form of a conservative reaction, in which the sanctity of tradition and the cohesive force of kinship, religion, customary law, rootedness in place, and traditionally founded hierarchy were valorized as the essence of an authentic national character and a viable and moral polity. What is noteworthy, however, is that traditionalist conservatives tended to see the trajectory of change in the same terms as the rational reformers, framing an oppositional contrast between tradition and modernity. This contrast, convergent with jurisprudential investigations and symbolic constructions of 'customary law,' became the core of the classic typological tradition in social thought out of which emerged the modern social disciplines, including anthropology, in which the force of tradition served as a defining criterion of social types.

By the latter half of the nineteenth century, the foundational ideas we have summarized coalesced in a series of typological syntheses aimed at comprehending the broad historical trajectory of human social and cultural development and at providing a classificatory framework of social formations based upon stages of evolutionary progress. Whether framed in terms of ideal types, developmental continua, or evolutionary stages, these magisterial syntheses all accorded a defining or diagnostic role to tradition in the construction of contrastive social types. E. B. Tylor (1960 [1881]), for example, in contrasting 'savage' and 'civilized' man, ascribes the former's intellectual conservatism and aversion to change to his 'tendency … to consider his ancestors as having handed down to him the perfection of wisdom.' Ferdinand Tönnies (1988[1887]), in basing his classic distinction between Gemeinschaft and Gesellschaft on contrastive forms of common will, characterizes the 'natural will' that predominates in Gemeinschaft societies as becoming naturalized 'through practice and tradition.' Emile Durkheim, in outlining the contrast between mechanical and organic solidarity, attributes the authority of the collective conscience in mechanically solidarity societies to 'the authority of tradition.' Max Weber, in constructing his three pure types of legitimate domination, maintains that 'the validity of a social order by virtue of the sacredness of tradition is the oldest and most universal type of legitimacy.' These and related formulations served to crystallize the notion of 'traditional society' as a typological category, intellectually and behaviorally conservative, bound by an unreflective and unquestioning adherence to traditional ways and submission to traditional authority, and slow to change, in contrast to 'modern society,' dynamic, impelled by the exercise of reason and rationally motivated innovation, and open to change. This typological contrast has proven remarkably durable, persisting to the present day as an orienting framework in the social disciplines, not least in anthropology. *Mutatis mutandis,* it underlies Robert Redfield's folk-urban continuum, Karl Popper's closed versus open society, Talcott Parsons' pattern variables, Lévi-Strauss' cold vs. hot society, and a host of other variations on the theme, persisting in the face of trenchant criticisms of its resistance to operational specificity, its lack of analytical utility in the investigation of empirical cases, its masking of the dynamics of tradition in so-called 'modern' society, and its role in sustaining structures of social inequality and policies of domination.

Tradition in Contemporary Anthropology

The conceptual and critical reach of tradition in contemporary anthropology is pervasive, especially as it implicates the element of continuity in culture. We may, however, outline several foci of empirical, conceptual, and critical interest in which the notion of tradition figures relatively explicitly.

The Social Organization of Tradition

One prominent line of inquiry concentrates on what we might call the social organization of tradition, addressing those aspects of the typological model that cast traditional societies as homogeneous or undifferentiated, with the implication that tradition is collectively shared within the group. This understanding may be called into question in several ways, most directly by the recognition within the group of cultural specialists, who are charged with or have an individual bent toward the cultivation, formulation, or communication of tradition, as ritual specialists, elders, reciters, or other custodians of expert or esoteric knowledge. The foci of investigation here include the definition and organization of specialist roles; the recruitment, socialization and careers of cultural specialists; and the institutional and situational context in which cultural specialists act in their expert capacities.

This latter focus may direct attention to cultural performances or display events, such as rituals, festivals, contests, or dramatic performances in which the knowledge, symbols, and values of the group are embodied, enacted, and placed on public display. The production of such events typically requires the efforts of specialists in various capacities, and an organization of production by which the performance is accomplished. These lines of inquiry have been especially prominent in studies framed in terms of the distinction between the Great Tradition (cf 'high,' 'classic,' 'cultivated') and the Little Tradition (cf 'low,' 'folk,' 'popular') developed by Robert Redfield.

Investigation of the social organization of tradition may converge as well with the study of socialization and enculturation: the roles, contexts, and practices by which knowledge and behavioral patterns are conveyed to, and acquired by, succeeding generations. Finally, it is worth noting that the social organization of tradition has critical implications for anthropological method, insofar as cultural specialists figure prominently in ethnographic work as favored sources of data, and that ethnographic encounters, organized around the getting and giving of knowledge, are frequently framed as pedagogical events.

Tradition and Creativity

A further variation on the categorical opposition between tradition and modernity is the contrast between tradition and creativity, especially prominent in the anthropology of the arts. Classic work in this area, such as Franz Boas's *Primitive Art* (1955[1927]), tended to view tradition as a constraint on the 'primitive' artist, consistent with the typological conception of traditional society as inherently conservative and Herderian notions of artistic forms as the collective expression of a national spirit. Thus the convention of collective attribution that has dominated the presentation (with a few notable exceptions) of verbal and material art forms from 'primitive' or 'traditional,' or 'folk' societies (a Navajo chant, a French folktale, a Yoruba staff).

Since the 1960s, however, there has been an accelerating shift in the anthropology of the arts toward greater attention to the role of the individual artist/performer, and more nuanced investigation of the interplay between convention and creativity in artistic practice. In the anthropological study of the material arts, the distinctive styles and virtuosic achievements of individual artists and the situated practice of artistic production, use, and evaluation have aided significantly in the recalibration of the balance between tradition and creativity, and even further to a reevaluation of what constitutes creativity, no longer restricted to novelty but understood to include the use of conventional forms in new ways.

There has been a parallel development in the study of the verbal arts, stimulated by Milman Parry and Albert Lord's pioneering studies of the composition in performance of Serbo-Croatian oral epic and by the correlative turn to the ethnography of performance by linguistic anthropologists and folklorists. Central to these investigations is a focus on the situated nature of performance as a mode of communicative practice and recognition of the emergent quality of performance. In linguistic anthropology especially, the ethnography of performance has incorporated the close analysis of formal, functional, and intertextual relationships as a vantage point on the calibration in performance of conventional, 'traditional' orienting frameworks for the production and interpretation of discourse on the one hand, and the exigent and emergent 'creative' qualities of situated performance on the other. Both qualities are always present, and the task becomes one of determining the relative weight and the dynamics of interplay of the two in particular social and historical circumstances.

Tradition and Communicative Technologies

Closely related to the typological contrast between traditional and modern societies is the distinction between oral and literate societies, which foregrounds specific technologies of communication as diagnostic – even determinative – of the contrastive types. The typological uses of orality and literacy have their roots in eighteenth-century philology, which in turn had a shaping influence on anthropology. The twentieth century revival of this framework by classical scholars has once again infiltrated anthropology and re-energized debates concerning the validity of 'Great Divide' typologies. Stated in very general terms, the 'literacy hypothesis' attributes to the advent of writing, the invention of print technologies, and the associated spread of literacy, a concomitant transformation of modes of thought, specifically the objectification and decontextualization of knowledge, and the encouragement of a critical or skeptical orientation toward received truth. These intellectual (or cognitive) orientations, according to proponents of the literacy hypothesis, had a transformative influence on society and culture, effecting an epochal shift from a blind and unreflective adherence to traditional modes of thought and practice to the rational, scientific openness to change that is diagnostic of modernity.

While some anthropologists have endeavored to find ethnographic corroboration of the typological contrast between traditional and modern modes of thought, the weight of ethnographic evidence has tended to call the literacy hypothesis into question. The most effective and persuasive critiques have come from linguistic anthropology and related work in the anthropology of teaching and learning. A large and growing body of ethnographic studies, focusing on speaking, writing, reading, and the acquisition of communicative competence as social practices embedded in situational and institutional contexts of use, has provided a range of critical correctives to unitary conceptions of orality and literacy studied in relation to abstract systems of thought. These lines of research, illuminating the role of communicative technologies, forms, and practices in the production and reproduction of social life, offer as well an increasingly ramified and nuance understanding of the dynamics of continuity and change in human society, to supplant the classic categorical distinction between tradition and modernity.

The Politics of Tradition

The political force of tradition has been a consistent theme in anthropology, built into the very foundations of the discipline and epitomized in Weber's identification of traditional authority as one of the three pure types of legitimate domination. Oriented by classic conceptions of traditional authority and 'traditional' society more generally, foundational studies of traditional authority in anthropology have tended largely to view it as a unifying force in the service of social system maintenance, sustained by the mechanisms of socialization, the mythological charter of institutions, and the rigid formalization of ritual and traditional oratory.

Here again, however, the ethnographic turn to situated practice, and especially to discursive interaction and performance, has led to recognition of the ways that traditional tales, for example, may be instruments of divisiveness rather than unity, insofar as they may be performed in defense of competing claims to status, or that the formal regimentation that marks the performance of oratory as 'the words of the ancestors' may serve to enhance the rhetorical power of disputants in political conflict. That is to say that rhetorical appeals to tradition and the public ideologies of traditional authority that motivate them do not necessarily have the binding force in practice that participants and observers alike may claim for them. Ideologies of traditional authority – like all ideologies – are positioned, interested, and strategically invoked, and tradition as a rhetorical resource may serve as well for conflict as for the 'binding force' of acquiescence.

The ideology and rhetoric of tradition are understandably more apparent under conditions of marked social and cultural change; indeed, the capacity of 'traditional' peoples consciously to intervene in the management of tradition was recognized earliest in anthropology in the study of nativistic movements, 'any conscious, organized attempt on the part of a society's members to revive or perpetuate selected aspects of its culture.' (see Linton, 1943) This work anticipated a vigorous line of anthropological inquiry that has burgeoned since the 1980s, devoted to the investigation of tradition and related notions of authenticity as symbolic, interpretive constructions in and of the present through the invocation of links to a meaningful past.

The process of traditionalization – under the rubric of tradition, heritage, patrimony, custom, or the like – is a prominent resource in the service of nationalism, whether cultural nationalism or nation-state formation, and it is a conspicuous aspect of the politics of culture in many parts of the world in which there is an extended history of anthropological research but which are now increasingly determined to shed the vestiges of a colonial past and assert the integrity and authenticity of their own 'cultural traditions.' The readiness of anthropologists to identify the tradition or custom, or patrimony on which such claims stand as selective, strategic, symbolic constructions is a source of tension both within the discipline and between anthropologists and their ethnographic interlocutors: whose interpretations and constructions are to be accepted as authoritative?

This question and the further problems it raises afford an illuminating critical vantagepoint on the historical formation of anthropological theory and practice. The modern conception of tradition was shaped in an intellectual arena that was defined largely in political terms. What are the foundations of a viable polity? A legitimate authority? A national culture? What are the bases of social inequality? Who is authorized to intervene in the management of these formations? Tradition and allied conceptions of culture proved to be foundational to the emergent discipline of anthropology and the definition of its purview. A measure of the success of anthropology may be strength with which its core concepts have come to be accepted by the formerly 'traditional' peoples who have been its central objects of study. One of the challenges that now confronts anthropology is that in asserting their own subjectivity, these peoples should base their claims on those foundational concepts just as anthropologists themselves are increasingly inclined to deconstruct them, to recognize traditionalization as an interpretive process, to challenge the reification of tradition, or to unmask primordial traditions as recent inventions.

Bibliography

Bauman, R., Briggs, C.L., 2000. Language philosophy as language ideology: John Locke and Johann Gottfried Herder. In: Kroskrity, P.V. (Ed.), Regimes of Language: Ideologies, Polities, and Identities. School of American Research Press, Santa Fe, NM.

Ben-Amos, D., 1984. Seven strands of tradition: varieties in its meaning in American folklore studies. Journal of Folklore Research 21, 97–131.

Biebuyck, D.P. (Ed.), 1969. Tradition and Creativity in Tribal Art. University of California Press, Berkeley, CA.

Briggs, C.L., 1996. The politics of discursive authority in research on the 'invention of tradition'. Cultural Anthropology 11, 435–469.

Boas, F., 1955[1927] Primitive Art. Dover Publications, New York.

Collins, J., 1995. Literacy and literacies. Annual Review of Anthropology 24, 75–93.

Finnegan, R., 1988. Literacy and Orality: Studies in the Technology of Communication. Blackwell, Oxford, UK.

Finnegan, R., 1990. Tradition, but what tradition and for whom? Oral Tradition 6, 104–124.

Firth, R., 1961. History and Traditions of Tikopia. Polynesia Society, Wellington, New Zealand.

Handler, R., Linnekin, J., 1984. Tradition, genuine or spurious. Journal of American Folklore 97, 273–290.

Hobsbawm, E., Ranger, T. (Eds.), 1983. The Invention of Tradition. Cambridge University Press, Cambridge, UK.

Jolly, M., Thomas, N. (Eds.), 1992. The politics of tradition in the Pacific Oceania 62 (4) (special issue).

Kelley, D.R., 1990. 'Second nature': the idea of custom in European law, society, and culture. In: Grafton, A., Blair, A. (Eds.), The Transmission of Culture in Early Modern Europe. University of Pennsylvania Press, Philadelphia, PA.

Linton, R., 1943. Nativistic movements. American Anthropologist 45, 230–240.

Shils, E., 1981. Tradition. University of Chicago Press, Chicago.

Singer, M.B. (Ed.), 1959. Traditional India: Structure and Change. American Folklore Society, Philadelphia, PA.

Tönnies, F., 1988[1887] Community and Society. Transaction Press, New Brunswick, NJ.

Tylor, E.B., 1960[1881] Anthropology. University of Michigan Press, Ann Arbor, MI.

Wallace, A.F.C., 1970. Culture and Personality, second ed. Random House, New York.

Tradition: Social

Simon Langlois, Université Laval, Québec, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Contemporary sociology considers tradition as an invented reference developed in function of the demands of the present and as a form of rationality among others that orients action. To be part of a tradition, objects, norms, or ideas have to be *living* in the sense of never having been forgotten. Tradition is a transformed heritage which survives and orients contemporary actions as seen in many fields of empirical research. The market consumption society offers standardized products but cultural diversity continues to be reaffirmed by living cultural traditions. New technologies have a great impact on the character of work, but work is also embedded in culture and traditions.

While the concept of tradition has been important in sociological thought since the foundation of the discipline, its signification has completely changed in contemporary works. At the turn of the twentieth century, sociologists and anthropologists utilized this concept to outline a type of society defined by certain traits placed in opposition to the traits characterizing industrial society, and in a well-known work Max Weber distinguished three types of authority: legal-rational, charismatic, and traditional. One hundred years later, sociology considers tradition as an invented reference developed by societies in function of the demands of the present (Hobsbawm and Ranger, 2012) and as a form of rationality among others that orients action (Boudon, 1992).

The tradition–modernity dichotomy dates back to Saint-Simon who compared the characteristics of the emerging industrial society to those of the *Ancien régime*. This dichotomy became a key element in explaining the social transformations that were occurring in modernizing societies during the first part of the twentieth century. The defining traits of the traditional society are well known: the fusion of family and business, the division of labor dictated by social structure (age, gender, relatives), reference to customs as the dominant value, the privileged attention accorded to saving and accumulation to the detriment of productive investment, the importance of the local community, relative autarky, and self-reliance. In a work that has been read by generations of American students and remains a classic in the social sciences to this day, the American anthropologist Horace Miner (1939) found a kind of archetype of preindustrial traditional rural life in Saint-Denis-de-Kamouraska, Quebec. At the opposite end of the spectrum, the industrial society described by Raymond Aron, Georges Friedman, Daniel Bell, Krishan Kumar, and others is marked by the growing importance of technology, economic growth, calculation, and management, productive investment and the accumulation of capital, the separation of family and business, the establishment of a labor market, and urbanization.

The transition from traditional to industrial society paradigm and the folk–urban continuum formalized by Redfield (1956) were criticized by some authors as descriptive approaches of social change and as paradigms of limited explanatory power. Henri Mendras (1976) argued that these perspectives did not recognize the proper originality of nonindustrial societies as well as their diversity. Furthermore, as a concept, traditional society is used to refer to several types of societies that are in fact very different from each other (e.g., primitive societies and peasant societies) and one should not confuse them. If hunter–gatherer societies are totally independent, peasant societies are relatively autonomous but embedded in larger, global societies having more diverse characteristics. The paradigm of the transition from traditional to industrial society is based upon an evolutionist postulate (heavily criticized) that claims developed societies constitute the norm to thrive for. In the same vein, we can criticize the thesis of the resistance to social change by tradition because it supposes clear, obtainable goals and also the existence of bad faith on the part of traditional individuals attached to the past and incapable of recognizing the merits of anything new (Boudon, 2007).

Consider another manner of conceptualizing the opposition between the traditional and the modern. Levi-Strauss (1962) compares the knowledge and technique of the handyman, which are traits of traditional society, with the knowledge and technique of the engineer, which are attributes of industrial society. The first type of knowledge is empirical and inductive. It relies upon a long tradition of precision which can be complex and efficient but it is also filled with errors, notably as to the explanation of its efficiency. The knowledge of the engineer is of a different order. This type is based upon experimental science, the accumulation of scientific knowledge, verification, technology, the manipulation of objects, operationalization, and organization. The classical opposition in economics between protectionism and free trade also illustrates the opposition between the ancestral or traditional ways of doing business and the modern methods. The authorities in the Republic of Venice jealously protected the ancestral traditions of their glass makers, sources of important benefits. Contemporary entrepreneurs on the other hand count on innovation, discovery, free circulation, investment, and strategic alliances to support the growth of their activities and profits.

The examples of the engineer and the entrepreneur clearly demonstrate that tradition and modernization are differentiated in two respects. First, the relationship to time is different. Tradition is oriented toward a legitimate reference to the past while modernization is oriented toward the mastery of the

future, the discovery of the new. Second, tradition and modernization are marked by a strongly different manner of intervening in the world, of producing goods and services; an empirical manner of learning over time and transmission across social relations that are embedded within the family or small communities, and a scientific manner that asserts itself as the result of research and calculation, embedded within complex social networks and formal organizations.

Shils (1981) proposed a definition of the concept of tradition that was largely accepted. He defined tradition as anything that is transmitted or handed down from the past to the present. Traditions include beliefs, norms, institutions, as well as material objects which all individuals will find when they grow up; traditions refer to transmitted things, ideas, values. It is now generally accepted that the scope of tradition should be narrowed down. We will retain two aspects – which are hidden in the folk–urban continuum or in the community-society ideal type of Tonnies – which contribute to define tradition in contemporary analyses.

First, to be part of a tradition, objects, norms, or ideas have to be *living* in the sense of never having been destroyed, abandoned, or forgotten. Specifically qualified fragments of the heritage must maintain a strong intimate link with the present to remain a tradition. Second, it is generally accepted that traditions change with the passage of time. While traditions are often thought of as immutable and fixed, they are subject to be reexamined and questioned through the passage of time (Dumont, 1968). Tradition is a transmitted heritage which survives and orients contemporary actions, but this heritage is often transformed (Halbwachs, 1994/1936; Eisenstadt, 1973; Sylvestre, 2012). In this regard, Sztompka (1993) stated that a special part from the whole of the historical heritage is elevated to the status of tradition by contemporary people for their own usage. From time to time, nations are re-founded on a new basis in order to recognize and incorporate emerging elements, for example, a growing proportion of immigrants (Dumont, 1993; Taylor, 1989). In the United States, national history is being rewritten to take into account the role of the indigenous peoples, women, and ethnic minorities. Accordingly, Michael Lind (1995) distinguished four types of 'American nations': Anglo-America, Euro-America, multicultural America, and finally, 'the new American ethnicity.'

In 1998, the Canadian government established a new territory populated by an Inuit majority – Nunavut – that took into account the testimonies of the oral traditions of the elders. Scientific cartography and satellite images were the privileged work instruments but the life histories and experiences of those who inhabited the land were taken into account, especially to establish its Eastern frontier. Similarly, the oral traditions of Canadian Aboriginals and the wampums of their material culture were also taken into account as elements of proof in the Canadian system of justice that examined their grievances and their ancestral rights. Canadian judges are now obliged to interpret the rights of Aboriginal people in light of ancestral traditions, on one side, and the orientations of modern ways of living (equality between men and women, for example). The resort to tradition inspires other courts of justice throughout the world where we find indigenous cultures, in Latin America or in Oceania, for example. Here, tradition appears also as a legitimate form of knowledge, but different than the scientific knowledge and cartographic techniques of historians or jurists, and it is susceptible to applications in the same way as the latter. Tradition no longer appears as the only principal source that orients behavior and decisions but as one element among others that is susceptible to being criticized.

Hobsbawm and Ranger (2012) have contributed to the new perspectives on tradition by stressing its invented character. They have distinguished three groups of invented characteristics. First, there are traditions which express the social cohesion of nations and communities. Second, traditions can legitimize existing order, status, institutions, or authority. Third, traditions socialize individuals to certain values or rules of behavior. The traditions of the British crown, for example, leave the impression that they are rooted deep into the past. The work of historians has now demonstrated that these traditions are of more recent origins and were invented quite recently (in fact in the nineteenth century) in order to give greater legitimacy to the institution of the monarchy. The celebration of the new Beaujolais in France is a recent creation, the fruit of a new approach in the marketing of the wine of this region that has been successful in countries where they appreciate French wines. Another example, some Australians and Canadians have questioned whether or not their country should abandon the royal symbols from their colonial period. Abandoning those traditions and symbols would probably oblige them to invent other symbolic tradition to mark the birth of their society peopled in large part by immigrants. A new tradition would, therefore, be created for use in the present.

The protection of the environment, fauna and flora, became a daily theme at the end of the twentieth century. In North America, the Aboriginals have been characterized as environmentalists before the advent of the environmental movement. Research by anthropologists and historians has demonstrated that the reality is much more complex. Krech (2000) recalls that Aboriginals had overexploited the beaver in the St. Lawrence Valley to the point of extinction after the arrival of Europeans in sixteenth century New France. The notions of environmental protection and endangered animals were foreign to the traditional culture of the Aboriginals, because the way of life of hunter-gatherers is based upon the maximal exploitation of the immediate environment. On the other hand, Aboriginals have an animist conception of the animal world. Thus, the Plains Indians believed that they could not allow an animal to escape under the pretext that it would warn the others to not venture in that area, they therefore killed all the animals they saw. Because they believed in reincarnation, the majority of Aboriginals thought that the dead animals would, at any rate, return the following season. For Krech, the myth of the ecologist Indian essentially reflects contemporary preoccupations for the environment rather than the traditional and historical culture of the indigenous peoples.

This analysis leads us to the following conclusion: sociology has abandoned the transition paradigm (from traditional to industrial societies) and integrates tradition into analyses of ongoing social change. Tradition is used to define a type of society (be it theoretical or empirical), and it also defines a subjective orientation toward action. We will further discuss these different approaches.

In his general theory of action, Freitag (1986) defines different types of societies based on forms of definition of the norms of social action. The first form he proposes – culture – defines mythic and archaic societies. The second one – power – defines two other types of societies: traditional societies and modern societies, the later ones built against tradition. The third form of regulation – control – characterizes postmodern society, in which a large number of organizations contribute to the structuration of society.

Tradition is also used to describe empirically different societies. Huntington (1996) argued that the world is divided into eight major cultural zones which reveal persistent cultural differences caused by different religious traditions which are still powerful despite industrialization and the same forces of economic development. Inglehart and Baker (2000) studied values by surveys made in 65 countries which cover more than 75% of the world population and they observed the persistence of distinctive traditional values.

In contemporary sociology, tradition is considered as a normative orientation toward action, a perspective underlying Tocqueville's books (1856, for example). In his works, Raymond Boudon (1992) formulates a general theory of action based upon the actor's search for meaning. He proposes to enlarge the concept of the rationality of the actor by distinguishing five types of rationality: utilitarian, teleological, axiological, cognitive, and traditional. All of these types of rationality describe the good reasons why the actor is driven to act in a certain way. Economists and rational choice theorists privilege the utilitarian rationality based upon the pursuit of self-interest. However, for Boudon, there are other kinds of reasons/motives that can also influence the way individuals act, and limiting the concept of rationality to the single dimension of utility is an exercise in reductionism. Custom and habit can also incite individuals to act in a certain way. Therefore, when one speaks of traditional rationality – "it has always been that way" – he or she is speaking of one type of rationality among others.

Sociological analysis has to define the context and circumstances in which one type of rationality or another is at work, keeping in mind that one type is no more important than another. This approach permits us to understand the role of tradition in sociological explanations. In peasant societies, the rhythm of the seasons is of major importance and technology is not very advanced (Weber, 1976). Here, the reference to custom and tradition is strong and individuals do not have good reasons to question what already works. They are led to value stability and the family that provides the necessary labor to exploit the land. However, when technological inventions appear to menace the traditional ways of life, tradition remains important for understanding the implantation of these transformations or the difficulties in accepting them. The introduction of inventions in peasant societies cannot be understood independently of the host culture because these innovations have to be first integrated into the traditional culture before being fully accepted.

The celebrated monograph by Thomas and Znaniecki (1984) on immigrant Polish peasants in Chicago gives a good example of the role of tradition in ongoing social change at the turn of the twentieth century. The immigrants that adapted the most rapidly to the host society were those that manifested the highest degree of traditionalism and attachment to their original milieu. This integration in a familiar milieu gave them support and a frame of reference that slowed down the process of social disorganization that was typical of immigrant communities when faced with a radical rupture with their traditional original community. The Polish immigrants adopted a rational action that allowed them to integrate to their new society with the traditional collectivity serving as a relay.

To illustrate the richness of this approach, here is an example taken this time from French sociology. In his studies of religious sociology, Le Bras (1955) observed the following paradox. The deeply religious Bretons who migrated to Paris in the 1950s rapidly stopped practicing their religion. A superficial analysis gave the impression that they had simply abandoned a strict and conformist religion as soon as they found themselves away from the traditional village life. Le Bras proposed another explanation by demonstrating that the underlying logic of behavior was the same in the village and in the city. In the village, behavior was regulated by the social group. Going to mass was not only a way of practicing one's faith, it was also participating in a valued social activity which strongly regulated village life. In the city, a pluralist environment, going to mass was seen as expressing an opinion and a behavior that did not attract the adherence of all, and in a way, that was contrary to majority opinion. Thus, it was difficult for the newcomers, who wished to integrate to their new environment, to oppose something which was the norm in their original milieu. In this case, one could say that going to mass in the village and not going once in the city follow the same social logic.

All of these examples point out different ways to refer to the concept of tradition. Tradition describes a type of society but it also defines an orientation of action. This perspective can be found in numerous works of contemporary sociology and demonstrates how different types of rational behavior can syncretize. To illustrate this perspective, we will recall how the reference to tradition is posed in three major fields of empirical research.

The market consumption society in the developed world offered standardized products – fast food, coca-cola, computers, cars, jeans, CDs, films – and globalization tends to homogenize commercial practices and modes of production. However, sociologists have observed the multitude of innovative responses in the national cultures of consumption and have noted that local contexts remain extremely diversified and, in each case, outside material influences are reinterpreted. Cultural diversity continues to be reaffirmed and supported by living national and cultural traditions. Coca-cola does not occupy the same place in the dietary culture of the French and the connoisseurs know that American pizza is quite distant from its Italian cousin (Ritzer, 2006). New technologies have a great impact on the character of work, but work is also embedded in culture and traditions. People work, and supervise other's work, within stringent institutional limits marked by both national traditions and organizational innovation.

The empirical work on the integration of immigrants in developed societies demonstrates that they do not form totalities with contours that are well defined by their original traditions. In fact, they are quite diverse and in reconstructing their traditions in the new country, they combine innovative

elements with the adopted elements of their new home. Tradition and modernity combine to produce new cosmopolitan identities that can be different from one country to another as Kymlicka (1995) suggested. The reference to tradition, therefore, remains quite alive in sociological analyses.

See also: Collective Beliefs, Sociological Explanation of; Collective Memory, Psychology of; Conformity: Sociological Aspects; Control, Social; Social System; Societies, Types of; Trust, Sociology of.

Bibliography

Boudon, R., 1992. Action. In: Boudon, R. (Ed.), Traité de Sociologie. Presses universitaires de France, Paris, pp. 21–55.
Boudon, R., 2007. Essai sur la théorie générale de la rationalité. Presses universitaires de France, Paris.
Dumont, F., 1993. Genèse de la société québécoise. Boréal, Montréal.
Dumont, F., 1968. Le lieu de l'homme. La culture comme distance et mémoire. HMH, Montréal.
de Tocqueville, A., 1986/1856. L'Ancien régime et la Révolution. Robert Laffont, coll. Bouquins, Paris.
Eisenstadt, S.M., 1973. Tradition, Change, and Modernity. J. Wiley, New York.
Freitag, M., 1986. Dialectique et société. Introduction à une théorie générale du symbolique. Genève. Éditions L'âge d'homme and Montréal: Éditions Saint-Martin.
Halbwachs, M., 1994/1936. Les cadres sociaux de la mémoire. Albin Michel, Paris.
Hohsbawm, E., Ranger, T. (Eds.), 2012. The Invention of Tradition. Cambridge University Press, Cambridge.
Huntington, S.P., 1996. The Clash of Civilization and the Remaking of World Order. Simon and Schuster, New York.
Inglehart, R., Baker, W.E., 2000. Modernization, cultural change, and the persistence of traditional values. American Sociological Review 65 (1), 19–51.
Krech, S., 2000. The Ecological Indian. Myth and History. W. W. Norton and Co, New York.
Kymlicka, W., 1995. Multicultural Citizenship. Oxford University Press, Oxford.
Le Bras, G., 1955. Études de sociologie religieuse, tomes I et II. Presses universitaires de France, Paris.
Lévi-Strauss, C., 1962. La Pensée Sauvage. Plon, Paris.
Lind, M., 1995. The Next American Nation. The New Nationalism and the Fourth American Revolution. Free Press, New York.
Mendras, H., 1976. Les sociétés paysannes. Éléments pour une théorie de la paysannerie. Armand Colin, Paris.
Miner, H., 1939. St. Denis. A French-Canadian Parish. The University of Chicago Press, Chicago.
Redfield, R., 1956. Peasant Society and Culture. University of Chicago Press, Chicago.
Ritzer, G., 2006. Madonaldization. The Reader, second ed. Sage Publication, Thousand Oaks.
Shils, E., 1981. Tradition. Chicago University Press, Chicago.
Sylvestre, J.-P., 2012. La Tradition. Hermann, Paris.
Sztompka, P., 1993. The Sociology of Social Change. Blackwell, Oxford and Cambridge (USA).
Taylor, C., 1989. Sources of the Self. The Making of Modern Identity. Harvard University Press, Cambridge, MA.
Thomas, W.I., Znaniecki, F., 1984. The Polish Peasant in Europe and America. New York [1927], rééd. University of Illinois Press, Chicago.
Weber, E., 1976. Peasants into Frenchmen. The Modernization of Modern France 1870–1914. Stanford University Press, Stanford.

Trafficking of Women and Children in Latin America

Mirna E Carranza, School of Social Work, McMaster University, Hamilton, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes emergent themes related to social work and the trafficking of women and children in Latin America. A literature review revealed that empirical and scholarly work in this field is very limited, so a systematization of the literature – in English and Spanish, including internal community and organizational reports – is introduced here. Key themes include the recruitment, purpose(s), risk factors, consequences of trafficking on survivors' health, barriers to addressing trafficking, and the role of social work in this area. Women and children in Central America appear to be at particularly high risk due to geopolitical, historic, and economic issues.

Introduction

Human trafficking is a pervasive global social phenomenon and is a multimillion dollar business (Estés and Weiner, 2003). According to the United Nations Children's Fund (UNICEF), 1.2 million children are trafficked every year (UNICEF-Child Protection, 2010). In 2000, the United Nations (UN) introduced two optional protocols to the UN Convention on the Rights of the Child to address some of the worst forms of human rights abuses perpetrated against children and women globally, including sex trafficking. The protocols prohibit the sale of children, child prostitution, and child pornography; require that a state's criminal or penal law regulate particular banned acts; set forth various methods for enforcing jurisdiction and extradition; and specify the protection of children victimized by the prohibited practices (Revaz, 2009). To date, 18 of 23 Latin American countries have signed or ratified such protocol (Guinn, 2008). Signing countries have developed legislation, public policies, and preventive strategies to deal with this social issue.

Trafficking of women and children is very difficult to detect and quantify because of its underground nature (CONAPINA, 2002). Therefore, numbers and statistics are often estimated (Wolf-Branigin et al., 2010). Most of the available data come from institutional reports. To date, little research has focused on the nature and extent of trafficking of women and children in Latin America (Sumner et al., 2004), but nongovernment organizations (NGOs) and human rights advocates report that Latin American women and children continue to be subjected to violence, abuse, and trafficking for the purpose of commercial sexual exploitation (CSE) (Scarpa, 2005). From 2007 to 2010, an estimated 6000 individuals were trafficked in Latin America: 1600 (27%) children and 4400 (73%) adults, almost entirely women (UNODC, 2012). In Latin America, the trafficking of children and women is rooted in its history of colonization. Women, especially Indigenous and African women, have been trafficked for the purposes of: (1) providing free labor, (2) bearing more slaves, and (3) sexual objectification (Chiarotti, 2003). Research also indicates that this region is currently a destination for accessing trafficked women and children (Phinney, 2001). According to the Human Rights Watch (HRW), weak government institutions, the absence of a formal regional legal framework for penal, economic, and financial interventions (2012), and a lack of political endorsement and/or commitment to addressing this complex issue make Latin America fertile ground for human trafficking (III-LACHT, 2012).

Recruitment and Trafficking

Recruiters use various techniques to bring women and children into the sex trade: kidnapping, bondage, deceit about possible employment, fraud, fear, physical violence, confiscation of travel documents, forced substance use, threats of violence toward family and friends, debt, and threatened deportation (Potocky, 2010). Recruiters are often family members or friends of victims, and many traffickers are women (Azola, 2001). A recent study conducted in the Dominican Republic (Carranza et al., 2010) reported that children who are forced to grow up in this type of environment are used as recruiters when they become older (17–19 years). They befriend younger children, gain their trust, and convince them to join what appears to be a 'fun' life. Another study conducted in Nicaragua (Carranza et al., 2011a) reported that children and young women are recruited for CSE within schools in impoverished communities. The Latin American region is very distinct in which 50% of the individuals prosecuted for human trafficking are women (UNODC, 2012). However, in this region, "women who are prosecuted for trafficking in persons are less likely than men to be convicted in comparison with Eastern Europe and Central Asia" (UNODC, 2012: 61–62). The reasons for this require further exploration.

Trafficking in Humans: For What Purpose?

According to the United Nations Office on Drugs and Crimes' (UNODC, 2012) report on global human trafficking, 51% of the trafficking in Latin America is for the purposes of sexual exploitation, with forced labor a close second at 44%. 'Sexual exploitation' includes: (1) child prostitution, (2) pornography, and (3) sex tourism (Covenant House, 2001). Other reasons for trafficking may include: (1) to be part of a militia, (2) forced marriage, (3) organ harvesting, and (4) illegal adoptions (UNODC, 2012).

Forced labor can be considered a particular form of slavery that is difficult to identify and estimate. It uses a different criminal network from that of sex trafficking, and instead involves involuntary servitude through coerced household,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28114-2

factory, or fieldwork (TIP, 2009). According to Molina (2008), trafficked women and children may also be involved in the entertainment business and/or drug industry, e.g., they may be used as 'mules' to transport drugs across borders. These individuals often risk their lives by ingesting drugs to avoid detection by border security. A recent study conducted in the Dominican Republic (UNICEF-Dominican Republic, forthcoming) reported that Haitian children and adolescents migrating alone to the Dominican Republic are at great risk of being captured by traffickers. Traffickers force boys to beg on the streets, and force girls to work as maids. A growing transnational and underground network is involved in moving women and children from place to place (Fontela, 2007), and in 2010, UNODC reported that of the 140 000 individuals trafficked to Western and Central Europe for sexual exploitation, 13% were from Latin America.

Conflict reproduces distinctive kinds of power relations (Bukachi et al., 2010). Recruitment of young women and children into militias is common in countries in political turmoil (Escorcia, 2007), including El Salvador, Guatemala, and Colombia. According to the HRW (2012), the Fuerzas Armadas de Colombia has reportedly kidnapped children for use as child soldiers. Molina (2008) added that more than 2500 Venezuelan children have been abducted to fight in the Colombian war.

Other purposes for trafficking children and women include forced marriages or servitude. Captors often confiscate documents and force victims to commit petty crimes under the threat of reporting them to the police if they disobey (Chiarotti, 2003). Organ harvesting is another common reason for human trafficking. Reports of missing women and children are often dismissed by authorities, who assume the individual has simply run away, so family members are often left to search for their loved ones alone (Molina, 2008).

In Latin America, the trafficking of children for the purpose of illegal adoption is a US$20 million per year industry, and children in Honduras, El Salvador, Ecuador, Venezuela, and Guatemala appear to be most at risk (Molina, 2008). According to a 2000 United Nations Economic and Social Council report, Guatemala is the fourth largest exporter of babies worldwide; the report also describes how mothers are tricked, paid, or robbed of their children.

Risk Factors

Most of the factors associated with trafficking of women and children for the purpose of CSE in the Latin American region are similar to those in the rest of the world: economic disparities, gender violence, and gender discrimination (Pop, 2012). Geography, class, and ethnicity are also important factors: Indigenous women from Amazonian and rural communities (Kempadoo, 2007), and Black Caribbean women (Bullock, 2010) are highly sought out and marketed to Caucasian males using colonial and racist descriptors such as 'dependent,' 'erotic,' and 'sex-crazed' (Bullock, 2010). A study conducted in the Dominican Republic reported that customers from European and North American countries particularly seek Black children and women, but prefer them to be well-groomed (i.e., clean, good teeth, well-groomed feet, straight hair; see Carranza et al., 2010).

The trafficking of women and children is a manifestation of gender violence (Robles and Pérez, 2012) and is the ultimate act of domination, both sexual and economic, over its victims (Aguilar Pérez, 2009). It is rooted in patriarchy and masculine gender ideology (Acharya, 2009), in which gender relationships are socially constructed (see e.g., Godoy, 2013). Societal attitudes (Gómez and García, 2006) of gender exclusion and violence underpin the violation of human dignity against women and children (Covenant House, 2004). However, the general population in Latin America tends to take these acts for granted as part of the 'natural world' (Centeno, 2006).

Researchers have identified other specific vulnerabilities in Latin America, including gang affiliation, ongoing substance abuse issues, having children before the age of 18 (Sumner et al., 2004), political instability (Phinney, 2001), and migration (Covenant House, 2001, 2004). Carranza (2012) found that children whose parents have migrated tend to experience isolation, abandonment, nonbelonging, and social stigma. They also tend to grow up with less supervision, and are therefore highly vulnerable to being recruited for CSE.

Childhood sexual abuse and intrafamilial violence reportedly play a significant role in the trafficking of women and children for the purpose of CSE (Martínez-Sotelo and Zúñiga, 2007; Roby, 2005). A study conducted in Nicaragua (Proyecto Miriam, 2008) reported that of 83 participants who had endured child CSE, 28% were survivors of sexual assault as children prior to entering the sex trade. Estés and Weiner (2001) studied three countries (Mexico, the US, and Canada) and found that most of the children involved in the sex trade had previously been exposed to childhood sexual abuse.

Central America: A Unique Region

Central America is a unique region in terms of trafficking women and children for CSE. Of its 41.3 million residents, approximately 33 million live in poverty due to historical influences (Solís and Solano, 2001). Its geopolitical position has been linked with organized delinquency (Heinemann and Verner, 2006), drug trafficking (Call, 2000), and trafficking in persons for sexual exploitation (ILO, 2009).

A 2002 study by the International Human Rights Law Institute of Central America reported that trafficking women and children for CSE is widespread in the region and occurs both internally (from poorer to wealthier or more developed areas) and across international borders. Internal trafficking is common in Guatemala, Honduras, and the Dominican Republic, with women and children trafficked from rural villages to tourist destinations along the coast. Crimes related to CSE are not generally considered illegal by either state institutions or citizens; rather, they are understood as part and parcel of transnational organized crime (Save the Children, 2011). Together, these illegal activities have eroded the social fabric, generated private and public corruption, and intensified institutional fragility.

In Central America, most victims of CSE are moved from Honduras to Guatemala (TIP, 2012). Nicaragua also receives trafficking inflows from South America and is a point of origin for traffic flowing south (Costa Rica) and north (Honduras).

The Free Transit Agreement among Nicaragua, Honduras, El Salvador, and Guatemala has strengthened the international route toward Northern countries (i.e., Canada and the US; see Acharya, 2009).

Consequences of Human Trafficking on Children and Women

Children and women survivors of sex trafficking endure life-threatening health consequences (Sexually transmitted diseases, HIV/AIDS, unwanted pregnancies), trauma (see Carranza et al., 2011a), and chemical dependency, which prevent them from achieving optimum mental and social well-being (UNFPA, 2007). Furthermore, when survivors of human trafficking attempt to access the judicial system in search of recognition and/or validation of the crimes committed against them, they are often met with acts and omissions aimed at delaying, preventing, or persuading them against exercising their human right to justice (Flores Chang, 2007). Furthermore, lack of education and poor health (Chiarotti, 2003), coupled with unemployable skills, weaken the social integration of women and children who have been subjected to trafficking (Martínez-Sotelo and Zúñiga, 2007). Children, in particular, face discrimination as they attempt to reintegrate with schools, friends, and communities of origin (Carranza et al., 2010). Political issues of repatriation continue, as victims of trafficking are often enmeshed in political bureaucracy (Carranza et al., 2011b). With a few exceptions, holistic initiatives related to prevention, intervention, and/or reparation are scarce (see Arredondo, 2011 for Chile; García, 2012 for the Dominican Republic). It is vital to develop spaces and trained personnel to support women and children who have endured trafficking.

Barriers to Addressing Human Trafficking

A major challenge related to addressing the trafficking of women and children is the 'invisibility' of perpetrators. For example, South American tabloids often portray police forces raiding brothels and nightclubs, detaining women and children who have been sexually exploited, rather than focusing on the traffickers (Rodríguez, 1999). Although the focus should be on the demand for CSE and the perpetrators of CSE, blame tends to fall on victims (Chiarotti, 2003).

The demand for CSE in Latin America is increasingly global (Altman, 2001). Hence a critical examination of how global forces intersect with local realities is necessary to clarify the actors and power dynamics involved in CSE (Carranza and Parada, 2010). In a qualitative study conducted in the Dominican Republic, Carranza et al. (2010) found that national Dominicans are perpetrators of CSE, but Asian, European, and North Americans (the US and Canada) tend to be the major consumers. In another study, Carranza et al.(2011c) found that women and children along Nicaragua's Atlantic Coast are particularly vulnerable to CSE from cargo ships' crew arriving from North American, Asia, and Europe, and from truck drivers moving cargo across the borders in Costa Rica and Honduras. They also found that in Granada, European men openly parade with a female child on their arm (Carranza et al., 2011b).

These crimes frequently go unpunished due to the silence of the civil population and corruption of government officials at all levels (Aguilar, 2009). The International Labour Organisation (2005, 2009) found that public acceptance of trafficking and sexual exploitation of women and children increased between 2005 and 2009. UNICEF (2005) found that many accusations made, to perpetrators of trafficking and/or CSE, by survivors or bystanders are dismissed as false. A recent study conducted in the Dominican Republic (Carranza et al., 2010) reported that children were being subjected to CSE mostly in tourist areas. However, this social issue is rapidly expanding to other public places such as shopping malls strongly suggesting that CSE is becoming more accepted throughout all segments of society and not just relegated to tourist areas. In another qualitative study conducted in Nicaragua (Carranza et al., 2011b), the participants noted that public apathy can actually protect survivors and their families by keeping them out of the legal system; the judicial system tends to revictimize survivors by attacking their character rather than focusing on perpetrators. This particular study also found that despite the work of human rights and child advocates, the demand in Granada, Nicaragua is so powerful that sexual relations between White men and local children are being constructed as 'normal relations' and part of the entitlement of White men.

Besides public apathy, perhaps the biggest challenge in addressing the trafficking of women and children for CSE is the lack of trust in authorities (i.e., police, government). A government's role is to represent and protect its citizens, particularly vulnerable groups, but women and children who have been trafficked report that they cannot trust authorities to assist them (Amar, 2009). According to participants in recent studies (Carranza et al., 2010 for the Dominican Republic; Carranza et al., 2011b for Nicaragua), most cases of trafficking of women and children involve government officials and police as consumers and direct beneficiaries, as they often accept bribes to ignore illicit activities and not pursue and prosecute individuals accused of these types of crimes. Thus, participants not only reported fear of reprisals from perpetrators, but indicated that reporting CSE is a waste of time because the authorities are often involved in the phenomenon.

The Role of Social Work

A few social work interventions have focused on CSE in Latin America, as in most Latin American countries the articulation of social work and CSE is emerging – i.e., Colombia (Mendez Villamizar et al., 2011), and other Caribbean (see Parade et al., 2007 for Dominican Republic), and Central American countries (see Carranza et al., 2011a,b,c for Nicaragua). Although Musto and Trajtenberg (2011) noted the lack of social work research in this particular area, social work is being actively conducted in several countries (e.g., Brazil, Argentina, Costa Rica, Chile). Social workers tend to work alongside other professions to address CSE but, according to the III Latin American Congress on Human Trafficking (2012), regional initiatives undertaken by academics, NGOs,

and the private sector are still in the early stages of raising awareness about CSE and understanding, preventing, and penalizing it. Interventions may include harm reduction, with victim-centered holistic approaches focusing on basic needs (e.g., food, shelter, clothing). Human rights activists continue to advocate for health care, counseling, education, and legal rights (Androff, 2010). Pressure by international NGOs and human rights activists has also increased the perception of human trafficking as a crime against humanity (e.g., Covenant House, Salvation Army, Save the Children). Laypersons and professionals (sociologists, teachers, lawyers) have been charged with performing interventions related to CSE, so international NGOs (Save the Children, UNICEF) have been involved in providing competency training for local employees working with the victims of trafficking (Guajardo, 2010).

Professional Social work in Latin America faces many challenges that impede worker's ability to effectively address CSE. Some countries, for example, still face the legitimation of the profession (e.g., Dominican Republic) and consequent limited funding (Sevilla, 2006); in other countries, social work has been marginalized due to its involvement in revolutionary and resistance movements (e.g., El Salvador, Honduras, Guatemala, and Chile). Social workers have been persecuted, tortured, or 'disappeared' throughout decades of government repression. In some cases they continue to be a watched group from the perspective of oppressive governments, which severely limits their ability to respond to or prevent CSE.

Conclusions

This article has identified several important challenges related to CSE in Latin America: (1) the complex issue of the trafficking of women and children is not fully understood, (2) only few preventive strategies have been designed to address this problem, (3) no guarantees are available to ensure that the human rights of victims are protected in relation to accessing justice; (4) perpetrators and traffickers are invisibilized as legal actions remain underdeveloped; and (5) aid and service provision for survivors of CSE is limited. CSE tends to be labeled and treated only in terms of 'woman and/or child abuse,' thereby contributing to its invisibility and to the revictimization of women and children that have endured trafficking if they disclose.

At the core of human trafficking – particularly of children and women – is the failure to recognize that children are entitled to human rights, and the denial of basic human rights for women. The trafficking of women and children within the Latin American region is closely linked with sociocultural and historic tenants including gender inequality, gender violence, social exclusion, expressions of masculinity, discrimination, and prejudice.

The role of the international community is pivotal. Involvement by the UN and international NGOs is crucial, as their 'international' status provides them with the authority and power to pressure government officials to address CSE. Moreover, international social workers play a pivotal role by providing technical support to social workers working with survivors of CSE. Reliable and ethical research protocols and social work practices are imperative. Moreover, the trafficking of women and children for the purpose of CSE must be included in academic curricula and research programs for social workers.

See also: Critical Social Work Practice; Globalization and Human Trafficking; Social Work Theory.

Bibliography

Acharya, A., 2009. Un análisis del trafico de mujeres y su tipología de origen. Andamios 6 (12), 299–322.

Aguilar Pérez, M., 2009. Análisis de genero de la explotación sexual comercial contra niños, niñas y adolescentes. In: Foro Nacional de la Explotación Sexual Comercial de niños, niñas y adolescentes. Universidad Autónoma de Nicaragua, Estelí, Nicaragua.

Altman, D., 2001. Global Sex. University of Chicago Press, Chicago.

Amar, P., 2009. Operation Princess in Rio de Janeiro: policing 'sex trafficking', strengthening worker citizenship, and the urban geopolitics of security in Brazil. Security Dialogue 40 (4–5), 513–541. http://dx.doi.org/10.1177/0967010609343300.

Androff, D.K., 2010. The problem of contemporary slavery: an international human rights challenge for social work. International Social Work 54 (2), 209–222. http://dx.doi.org/10.1177/0020872810368395.

Arredondo, O., 2011. La victima de explotación sexual comercial: Análisis desde el espacio de intervención especializada V región – Chile. Rayuela. Revista Iberoamericana sobre Niñez y Juventud en Lucha por sus Derechos 2 (4), 208–216.

Azola, E., 2001. La Explotación Sexual de Niños en las Fronteras. UNICEF, México City.

Bukachi, S., Juma, F., Africa, P., 2010. Baseline Survey on Gender & Long-term Forced Migration. Report prepared for Akina Mama wa Africa.

Bullock, B., 2010. Sex trafficking and trafficking in the Caribbean. In: McCabe, k, Mainia, S. (Eds.), Sex Trafficking: A Global Perspective. Lexington Books, New York, pp. 121–129.

Call, C.T., 2000. Sustainable Development in Central America: The Challenges of Violence Injustice and Insecurity. Institute for Iberoamerica-Kunde, Hamburg. CA 2020 Working Paper #8.

Carranza, M.E., Parada, H., 2010. Child commercial sexual exploitation in Nicaragua: A critical globalisation perspective. In: McCabe, K.A., Manian, S. (Eds.), Sex Trafficking: A Global Perspective. Lexington Books, New York, pp. 131–141.

Carranza, M.E., Parada, H., Hasbun, J., 2010. The Trafficking of Women and Children in Dominican Republic. A Qualitative Study. SSRCH-International Funding Opportunities.

Carranza, M.E., Parada, H., López, L.A., Jiménez, I., 2011a. Children and Women Trafficking within the Central American Forced Migrations – Nicaragua's Experience. Arts Research Board (ARB), McMaster University. Major International Research Seed Grant.

Carranza, M.E., Parada, H., López, L.A., Jiménez, I., 2011b. Forced Migration: Children and Women Trafficking within Central America – Nicaragua's Experience. Latin American Canadian Research Educational Grants (LACREG).

Carranza, M.E., Parada, H., López, L.A., Jiménez, I., 2011c. Forced Migration: Children and Women Trafficking within Central America – Nicaragua's Atlantic Cost Experience II. HIVOS – International.

Carranza, M.E., 2012. "No es lo mismo - ni es igual": Las experiencias transnacionales de adolescentes Salvadoreños. Rayuela Revista Iberoamericana sobre Niñez y Juventud en Lucha por sus Derechos 3 (5), 121–126.

Centeno, R., 2006. Preguntas pendientes a la descentralización: "Esta conduciendo hacia la igualdad de genero". Encuentro. Revista Académica de la Universidad Centro Americana 39 (76), 67–86.

Chiarotti, S., 2003. La Trata de Mujeres: Sus Conexiones y Desconexiones con la Migración y los Derechos Humanos. Población y Desarrollo, Serie 39. Centro Latinoamericano y Caribeño de la Demografía (CELADE) – División de Población. Banco Interamericano de Desarrollo (BID).

CONAPINA, 2002. Política Publica Contra la Explotación Sexual Comercial de Niños, Niñas y Adolescentes. CONAPINA/UNICEF, Managua.

Covenant House/Casa Alianza, 2001. Regional Investigation on Trafficking, Prostitution, Child Pornography and Sex Tourism with Children in Central America and Mexico.

Covenant House/Casa Alianza, 2004. La Explotación Sexual-comercial a Niñas, Niños Y Adolescentes Centro Americanos en Guatemala.

Escorcia Núñez, V., 2007. Derechos Humanos en Nicaragua. Informe Anual. Centro Nicaragüense de Derechos Humanos (CENIDH), Managua, Nicaragua.

Estés, R., Weiner, N., 2001. The Commercial Sexual Exploitation of Children in the US, Canada and Mexico. Full Report on the US National Study. University of Pennsylvania.
Estés, R., Weiner, N., 2003. La Explotación sexual comercial de niños en Estados Unidos. In: Estés, R., Azola, R. (Eds.), La Infancia Como Mercancía Sexual: México, Canadá Y Estados Unidos, pp. 44–90. Siglo Veintiuno Eds, México City, DF.
Flores Chang, M., 2007. Ruta Critica de las Dominicanas Sobrevivientes de Violencia de Genero. ProFamilia, Santo Domingo, Republica Dominicana.
Fondo de Población de la Naciones Unidas (UNFPA), 2007. Vidas vividas en riesgo. Caracterizando el vinculo entre violencia contra la mujer y el VIH/SIDA en Republica Dominicana. Resumen de Investigación.
Fontela, M., 2007. La prostitución, la trata de mujeres, niñas y la ley: ¿Derechos de los humanos o seguridad del estado? Cátedra Libre de los Derechos Humanos. Facultad de Filosofía y Letras.
García, D.M., 2012. Consejo Nacional para la Niñez y Adolescencia (CONANI). Rayuela: Revista Iberoamericana sobre Niñez y. Juventud en Lucha por sus Derechos 3 (5), 9–16.
Godoy, E., 2013. Women's Rights Still Denied in Latin America. Inter Press Service (IPS) News Agency, Mexico City.
Gómez, A.F.H., García, C.I., 2006. Masculinity in Colombia: deconstructing the conventional way of becoming a man. In: Bunnon, I., Correira, M.C. (Eds.), The Other Half of Gender: Men's Issues in Development. The World Bank, Washington, pp. 93–110.
Guajardo, M., 2010. Unaccompanied migrant children in transit: Challenges and opportunities for Mexico. In: Global Eyes on Human Trafficking. A Bulletin of News, Information and Analysis on Trafficking in Persons. International Organisation for Migration.
Guinn, D.E., 2008. Defining the problem of trafficking: the interplay of US law, donor, and NGO engagement and the local context in Latin America. Human Rights Quarterly 30 (1), 119–145.
Heinemann, A., Verner, D., 2006. Crime and Violence in Development: A Literature Review of Latin America and the Caribbean. World Bank Policy Research Paper #4041.
Human Rights Watch (HRW), 2012. In: World Report 2012 Americas. World Report 2012 Events of 2011, pp. 205–285. United States of America.
III-Latin American Congress on Human Trafficking (LACHT), 2012. Globalisation, Human Trafficking and Access to Justice: Articulation of Regional Dialogues. Centro de Estudios Internacionales (CEI) Universidad de los Andes.
International Human Rights Law Institute (IHRLI), 2002. In Modern Bondage: Sex Trafficking in the Americas – Central America and the Caribbean. De Paul University College of Law, Chicago. Available at: http://www.law.depaul.edu/ihrli.
Kempadoo, K., 2007. The war on human trafficking in the Caribbean. Race & Class 49 (79), 79–85. http://dx.doi.org/10.1177/0306396807049002060 2.
Martínez-Sotelo, D., Zúñiga, F.M., 2007. No Mas Vidas Truncadas. Documento de Delitos de Violencia Sexual Contra niñas y Adolecentes en Cinco Departamentos de Nicaragua. IXCHEN, Nicaragua.
Méndez Villamizar, R., Montero Torres, L., Aguilar Rueda, N., Linares Abad, M., Grande, M.L., 2001. Explotación sexual commercial infantil (ESCI): Ambito emergente de intervención professional del trabajador social. Universidad de Jaen, Colombia.
Molina, M.L., 2004. La Cuestión Social y la Formación Profesional en Trabajo Social en el Contexto de las Nuevas Relaciones de Poder y la Diversidad Latinoamericana. Espacio Ed, Argentina.
Molina, A., 2008. Explotación Sexual, Evaluación Y Tratamiento. Editorial Duncan, Buenos Aires.
Musto, C., Trajtenberg, N., 2011. Prostitución y trabajo infantil: El estado de arte de la investigación en Uruguay. Revista de Ciencias Sociales. DS. FCS 24 (29), 138–156.
Organización Internacional del Trabajo [ILO in English], 2004. Explotación Sexual Comercial y Masculinidad: Un Estudio Regional Cualitativo con Hombres de la Población General. IPEC. WEM.
Organización Internacional del Trabajo [ILO in English], 2005. El Comercio Sexual con Menores de Edad en Centro América, Panamá y Republica Dominicana. Segundo Estudio de Tolerancia Social. Análisis y Resultados, Desafíos y Recomendaciones. Programa Internacional para la Erradicación del Trabajo Infantil (IPEC).
Organización Internacional del Trabajo [ILO in English], 2009. El Comercio Sexual con Personas Menores de Edad: En Centro América, Panamá y Republica Dominicana. Segundo estudio de Tolerancia social Análisis de Resultados, Desafíos y Recomendaciones. Programa Internacional para la Erradicación del Trabajo Infantil, Costa Rica.
Parada, H., Moffat, K., Duval, M., 2007. Emerging practices of social work in the Dominican Republic: the protection of women's and children's safety. International Social Work 50 (6), 755–769.
Phinney, A., 2001. Trafficking of Women and Children for Sexual Exploitation in the Americas. Women in Development Program. Pan-American Health Organisation (PAHO). Retrieved from: http://www.oas.org/en/cim/docs/Trafficking-Paper[EN].pdf.
Pop, L., 2012. Focus: Trafficking of humans beings – An academic attempt to support the EU actions in the fight against it. Journal of Identity and Migration Studies 6 (1), 138–145.
Potocky, M., 2010. Effectiveness of services for victims of international human trafficking: an exploratory evaluation. Journal of Immigrant & Refugee Studies 8, 359–385.
Proyecto Miriam, 2008. Taller de Conceptualización y Manifestaciones de la Violencia Sexual y Explotación Sexual Comercial con Miembros de la Comisión de la Niñez (p. 14.) Estelí, Nicaragua.
Revaz, C., 2009. The Optical Protocols to the UN Convention on the Rights of the Child on Sex Trafficking and Child Soldier, pp.13–16. www.wcl.american.edu/hrbrief/09/1revaz.pdf (retrieved from 12.08.13.).
Robles, M.A.L., Pérez, R.M., 2012. Cuerpos deseados…cuerpos castrados: La explotación sexual desde un enfoque de genero. Rayuela: Revista Iberoamericana sobre Niñez y Juventud en Lucha por sus Derechos 3 (5), 43–49.
Roby, J.L., 2005. Women and children in the global sex trade: toward more effective policy. International Social Work 48 (2), 136–147. http://dx.doi.org/10.1177/0020872805050206.
Rodríguez, Z.R., 1999. Making harm visible: global sexual exploitation of women and girls speaking out and providing services. Report from Latin America (February). Retrieved from: http://www.uri.edu/artsci/wms/hughes/mhvla.htm.
Save the Children, 2007. Guía de Intervención Psicosocial para la Asistencia Directa con Personas Victimas de Trata. Internacional Organisation for Migration (OIM). Save the Children, Suecia.
Save the Children, 2011. Diagnostico de Comunicación sobre Violencia y Trata de Personas. Save the Children.
Scarpa, S., 2005. Child trafficking: the worst face of the world. Global Migration Perspectives 40. Global Commission on International Migration.
Sevilla, A., 2006. La prostitución infantil: Trabajo social y las necesidades sociales. Estudio Social, México. www.monografias.com (retrieved 20.08.13.).
Solís, L., Solano, P., 2001. Central América: The Difficult Road towards Integration and the Role of Canada. Focal Policy Paper. http://www.focal.ca/summary_central.htm.
Sumner, C.P., Vallance, P.A., Renaud, T., 2004. Literature Review of Trafficking in Persons in Latin American and the Caribbean. Development Alternatives, Inc. USAID.
Trafficking in Persons Report (TIP), 2009. Department of State. United States of America.
Trafficking in Persons Report (TIP), 2012. Department of State. United States of America.
UNICEF-America Latina & the Caribbean, 2005. Información por Países Sobre Explotación Sexual Comercial y Trata de Niñas, Niños y Adolescentes. Cooperación Italiana. SICA.
UNICEF-Dominican Republic, forthcoming. Estudio Cualitativo de niños, niñas y adolescentes Haitianos no acompañados en Republica Dominicana. Unicef-DR, forthcoming.
UNICEF Child Protection. UNICEF, 2010. http://www.unicef.org/protection/index_exploitation.html (retrieved 10.08.13.).
United Nations Economic and Social Council (ECOSOC), 2000. Report of the Special Rapporteur on the Sale of Children, Child Prostitution and Child Pornography. Ms. Ofelia Calcetas-Santos. Addendum. Report on the Mission to Guatemala. Commission on Human Rights. Fifty-Sixth Session.
United Nations Office on Drugs and Crimes (UNODC), 2010. The Globalisation of Crime a Transnational Organized Crime Threat Assessment. United Nations Publication, Sales No. E.10.IV.6.
United Nations Office on Drugs and Crimes (UNODC), 2012. Global Report on Trafficking in Persons. United Nations publication, Sales No. E.13.IV.1.
Wolf-Branigin, M., Garza, S., Smith, M.A., 2010. Reducing demand for human trafficking: a non-linear approach for developing capacity. Social Work & Christianity 37 (4), 424–436.

Relevant Websites

www.fas.org.
www.globalsecuritystudies.com.
www.globalsocialwork.org.
www.ilo.org.
www.resdal.org.
www.state.gov.com.
www.undp.org/protection/57929_58005.html.
www.unfpa.org.

Training in Clinical Psychology in the United States: Practitioner Model

Roger L Peterson, Antioch University New England, Keene, NH, USA
Donald R Peterson[†]**,** Rutgers University, NJ, USA
Jules C Abrams, Widener University, Chester, PA, USA
George Stricker, American School of Professional Psychology at Argosy University Washington, DC, USA
Kelly Ducheny, Howard Brown Health Center, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The National Council of Schools and Programs of Professional Psychology's practitioner-scholar model for education and training in professional psychology appeared in 1997 and 2010 (Peterson et al.) with five central components. Many elements of the five components are updated from their original presentations: (1) broadened view of psychology with a flexible epistemology, multiple ways of knowing, and an articulation of how practitioners doing practice remain 'local clinical scientists' doing 'disciplined inquiry'; (2) integrative pedagogy; (3) competency-based core curriculum integrating practical and scientific knowledge, skills, and attitudes, now including diversity; (4) elements of practice, including multiple roles, the self of the psychologist and reflective practice, practicum and internship training, and systematic evaluation; and (5) the social nature of professional psychology and its social responsibility, including an appreciation of ethnic and racial diversity, and gender. The impetus for the practitioner-scholar model and the critiques of the professional psychology movement are briefly mentioned.

The practitioner-scholar model of professional psychology education along with the scientist-practitioner model and the clinical scientist model are the three most influential models in doctoral-level professional psychology education. This summary of the practitioner-scholar model abbreviates the Peterson et al. (2010) version, which itself updated the Peterson et al. (1997) National Council of Schools and Programs of Professional Psychology (NCSPP; see NCSPP.info).

Founded in 1976, NCSPP is an organization of professional schools and programs in psychology whose central mission is the progressive improvement, enhancement, and enrichment of professional psychology education and training. In 2012, NCSPP consisted of 66 members with full accreditation by the American Psychological Association (APA) and 20 associate members who were at various stages of the accreditation process. The description of NCSPP's model is consistent with the requirement by APA's Committee on Accreditation (APA Office of Program Consultation and Accreditation, 2011) that programs identify an educational model that has a "clearly specified philosophy of education and training…appropriate to the science and practice of psychology" (p. 5); and that the chosen model "may be one identified through a national conference of psychologists, from which guidelines for professional education and training have been approved by the delegates" (p. 6).

The model began with the University of Illinois's short-lived Psy.D. program developed by Donald Peterson. In general, the practitioner-scholar model was a response to the view that the scientist-practitioner model overemphasized education for the production of university research and underemphasized preparation for the provision of psychological services.

The practitioner-scholar model in itself with the exception of one area has not received much criticism. McFall (2006, 2012), for example, sees the science training received by many Psy.D. students as inadequate for the production of the scientifically oriented professional psychologists that he sees as serving the field best. Most of the criticisms by him and others, including Donald Peterson, have been of the professional psychology movement, not the model (which is the focus of this article). They point to lower selection ratios (and therefore weaker students), larger classes, weaker faculties, and the location of programs in institutions outside of traditional universities.

The five central components that characterize the practitioner-scholar model will be addressed: (1) a broadened view of psychology, with a flexible epistemology, multiple ways of knowing, and a delineation of how practitioners *doing practice* remain *local clinical scientists* doing *disciplined inquiry*; (2) a description of the pedagogy of integrative experiences; (3) the professional core competencies and the requirements of a competency-based core curriculum in which practical and scientific knowledge, skills, and attitudes are integrated; (4) elements of practice, including multiple roles, the self of the professional psychologist and reflective practice, practicum and internship training, and systematic evaluation; and (5) the social nature of professional psychology and the public responsibility of the profession to serve the larger society, with special consideration of issues related to ethnic and racial diversity, and gender.

Psychological Science and Education for Professional Psychology

The most serious controversies and misunderstandings surrounding the practitioner-scholar model have concerned the role of science in the education of professional psychologists.

[†] Deceased.

All professional psychologists, whether PhD, or PsyD, are to be scientifically educated and trained in all aspects of their education, not just when they are conducting research or evaluation. Rigorous professional activity has traditionally been conceived as the application of scientific knowledge developed through laboratory experiments or controlled field research to the understanding and solution of human problems. In the practitioner-scholar model, practitioners engage the immediate challenge of providing services to people. Starting with the needs of each client, the practitioner-scholar model requires that practitioners bring the best available theoretical conceptions, the most useful available research, and their individual and collective professional experience to bear in studying and improving the functional condition of the client. Professional activity is not the application of knowledge derived from a separate scientific research process; it is a form of science and, indeed, a form of research in and of itself. The systematic process of scientific professional work has been characterized as 'disciplined inquiry' by Donald. R. Peterson (1991, 1995, 1996a,b). The role of the professional psychologist is that of the 'local clinical scientist' (Stricker and Trierweiler, 1995; Trierweiler and Stricker, 1992, 1998). These books and papers provide very detailed technical and epistemological arguments that go far beyond the brief summaries in this article. The properly trained professional psychologist is a scientist in the same sense that the skilled physician is a local clinical biological scientist and the skilled engineer a local physical scientist.

Nevertheless, the shift in viewpoint toward the local clinical scientist carries profound implications for the education of professional psychologists. Though most programs mandate dissertation research that contributes to general scientific knowledge as one central requirement of scientific education and training for all practitioners, the local clinical scientist model demands that situations and issues encountered by practicing professional psychologists to be examined in the course of graduate study as exercises in disciplined inquiry.

Within NCSPP, programs have taken an array of positions regarding the type of scholarly products integral to scientific education, ranging from requiring relatively small, clinically oriented doctoral projects to dissertations of the level and quality that might be found in traditional PhD programs. Those programs that require dissertations usually emphasize an applied focus that embodies a broader array of investigative approaches and a wider range of dissertation topics; all of which demonstrate an omnipresent emphasis on disciplined inquiry as basic to clinical education. Types of dissertations may include the following: (1) theoretical analyses; (2) surveys; (3) analyses of archival data; (4) outcome research, including program development and evaluation; (5) systematic qualitative investigations (6) public policy and legislative analysis; (7) case studies; and (8) group-based nomothetic investigations.

Throughout their careers, all practitioners are expected to retain a pervasive, deeply ingrained scientific outlook. A substantial number of graduates will take on the sorts of applied research listed above. Some will conduct professionally relevant, generalizable applied research, and a few will devote themselves to the traditional forms of research that preoccupy the faculties of most research universities. In a general way, these outcomes are not substantially different from those produced by traditional training models. The difference is that career expectations and educational experiences in the practitioner-scholar model are designed to match the realities of professional work in communities – as closely as they are now known and as clearly as they can be foreseen for the years ahead. These realities only rarely arise from traditional versions of the disciplinary core or from preparation for careers in academic research – careers that are unlikely to be available to the great majority of graduates, regardless of the program type.

The Practitioner as Local Clinical Scientist

NCSPP adopted Trierweiler and Stricker's (1992, 1998; Stricker and Trierweiler, 1995) vision of local clinical scientists who are "critical investigators of local (as opposed to universal) realities: (1) who are knowledgeable of research, scholarship, personal experience, and scientific methodology; and (2) who are able to develop plausible, communicable formulations for understanding essentially local phenomena using theory, general world knowledge including scientific research, and most important, their own abilities as skeptical scientific observers" (Trierweiler and Stricker, 1992: p. 104).

Scientific training in professional psychology is viewed as essential for developing and enhancing critical thinking in students, and it should be integrated throughout the curriculum. "Skills in local investigation and in problem solving (thinking on one's feet) assume unusual importance" (Trierweiler, 1992: p. 10). According to Trierweiler (1992), "the guiding metaphor becomes a Sherlock Holmes or a Jane Marple standing in direct confrontation with the constraints, mysteries, banalities, and surprises of unique realities, rather than the distant, conservative, skeptical, and abstractly speculative university-based scientist most of us have struggled with in our professional identities "(10–11).

Though the goals and forms of scientific training in professional programs differ from those in traditional programs, rigorous scientific training is equally important in both models of education.

Multiple Ways of Knowing

The broadened educational domain of professional psychology (e.g., its theories, research methods, and applications) is characterized by scholarly, disciplined thought that is grounded in science, the humanities, and personal and professional experiences, and is enhanced by interdisciplinary perspectives and a broad-based array of continually developing methods.

Reflexive professional psychology requires critical analysis of the theories that guide disciplined inquiry and the methods through which investigations are conducted. Reflective practice was first described by the Schoen Study in the early 1980s (see also D.R. Peterson, 1995; R.L. Peterson, 1992). These include the philosophical foundations of inquiry including, for example, epistemological and theoretical assumptions and implicit values (e.g., D.R. Peterson and R.L. Peterson, 1997) and the associated ethical issues, are therefore important in the education of professional psychologists.

The epistemological basis of disciplined inquiry in psychology must be comprehensive; responsive to wide-ranging, diverse, and fluid social contexts; and cognizant of invariably embedded values. This condition requires multiple ways of knowing that inform and enrich each other and that are appropriate and sensitive to the diverse populations to which they are applied. These ways of knowing include an enhanced array of methods drawn from related fields of inquiry – quantitative and qualitative, objective and subjective. In this context, research methods and theoretical formulations that increase understanding of women's issues and of diversity are stressed more strongly than they have been in the past.

In a typical case, to participate in disciplined inquiry, the necessary multiple ways of knowing require integrating material from the traditional scientific psychology knowledge base (including psychopathology, treatment approaches and outcomes, development, gender issues) with some relevant medical knowledge, the local economic circumstances and available resources, some history and relevant local historical narratives, knowledge about a local ethnic culture, a reasonably sophisticated understanding of a relevant religion, and, of course, the economic and service delivery issues, all working together in the context of a psychologist–client relationship. Though any particular requirements differ from situation to situation and client to client, the work of professional psychologists inevitably requires an array of multiple sources of data, collected in ways consistent with a variety of epistemologies.

Professional psychologists, in their roles as researchers, need explicit educational experiences to become self-critical with respect to the methodological, sociopolitical, and philosophical implications of inquiry. The integrative, reflective, educational experiences, described next as a central element of the NCSPP pedagogy, are just as relevant for the role of researcher as they are for the role of clinician. Conclusions should be carefully consistent with the limits of research designs, and particular consideration should be given to the likelihood of negative impact on underserved and diverse populations.

These positions, which may seem on the surface to be no more than bland ecumenical clichés, have general as well as clinical consequences. The areas of knowledge seen as relevant to professional psychology are greatly expanded. There is a strong movement away from narrow and doctrinaire conceptualizations of psychological science toward perspectives that oblige methodological multiplicity. Epistemology and critical thinking become more central to professional psychology training. An appropriate scientific basis for professional psychology demands an openness to the other groups and disciplines who share psychologists' concerns in a manner that emerges as an increased emphasis on ecological relevance (Peterson and Peterson, 1997). All of these ideas should be seen in contrast to the traditional scientific epistemology for psychology that appears in introductory texts and books on basic research.

Integrative Pedagogy

Even at this point in the portrayal of the NCSPP model, it is clear that the model necessitates a pedagogy of integrative experiences that is designed to educate students as local clinical scientists. To be more explicit, in this section we characterize the educational experiences and the contexts in which a conversation about a case becomes part of a student's learning. The integrative pedagogy of the NCSPP model can be seen as having the following seven aspects, including material described later as student development as a reflective process and explicit attention to relevant social issues such as marginalization, power, and authority.

Academic-scientific materials, both research and theory, including general information. Academic-scientific materials are used in combination with concepts and a process that facilitates local application (Stricker and Trierweiler, 1995; Trierweiler and Stricker, 1998). In a single learning experience, for example, materials might include critical discussions of information on a certain disorder, on possible treatments, on multiproblem clients and families, on ethical issues, on relevant issues of diversity, and on the economic realities of treatment, including the differences and contradictions that regularly appear in the literature.

Real examples and real experiences. These include the instructor's cases, cases seen under supervision, and live demonstrations and videos of students, instructors, and other practitioners. Students have the opportunity to engage in the real phenomena of professional psychology practice, so that disciplined inquiry can arise initially out of attention to the phenomena, not to a research problem.

The local, unique elements relevant to a particular client or professional situation (e.g., D.R. Peterson and R.L. Peterson, 1997; Trierweiler and Stricker, 1998).

Faculty and supervisory role models. Role models enthusiastically engage in the professional activities of the sorts that mirror the experiences students will encounter in clinical training and in their postgraduate practice.

Appropriate attitudes, including explicit ways of thinking like a psychologist. Thinking like a psychologist is integrated such that it will permeate all professional work.

Of course, depending on the program and the nature of the experience (whether, for example, a basic course with 30 students, a professional seminar of 8, or a supervision group of 4), different aspects of the pedagogy are emphasized. To take an example typical of many professional programs, an advanced student in a small group situation might present a diagnostic, conceptual, therapeutic, or ethical issue in a practicum case in light of the psychological literature for the feedback of the faculty leader and student colleagues. In such situations, the faculty person, and ultimately the students, should continually ask, "What have you observed?," "Why did you do that?," "What were the client's reactions?," "What is it like for you to sit with this client?," "How are these issues viewed by people of the same gender, ethnicity, and culture as the client?," "What, therefore, does this mean?," "What are your assumptions?," "How does this fit with your deeply held beliefs?," and, of course, "What is the evidence for your position?" or, "How do you know that?." As a presented case continues to unfold over time, different psychological literature may become relevant along with an evolving focus on the practice dilemmas that arise when a client's situation goes beyond the available readings. In doing so, participants become attuned to reflection and disciplined inquiry.

These integrative pedagogies have developed in an interesting variety of ways. All are designed to provide the best attainable professional service for each individual, group, or organizational case with which the students and faculty are concerned. Also, all are designed to be firmly embedded in the relevant scholarly knowledge, to reflexively educate the professional psychology graduate student, and to open doors that should lead to improved technologies and generalizable knowledge of relevance to the profession.

Core Curriculum and the Professional Core Competency Areas

The practitioner-scholar model has long been based on competencies. NCSPP articulated its initial work in the late 1980s and identified six competencies. A comprehensive statement was made in the original NCSPP model paper (Peterson et al., 1997). The descriptions below summarize much of the original material below plus diversity. Though not referenced, the language is drawn as closely as possible from the integrated resolutions and from papers describing these areas.

Relationship

Relationship is the capacity to develop and maintain a constructive working alliance with peers, colleagues, supervisors, members of other disciplines, consumers of services, and community organizations. The relationship competency is foundational to, and supportive of, all other competencies and is greatly impacted by students' awareness and connection to their self-identity. It involves the whole person and thus may include intellectual, emotional, cognitive, physical, cultural, and spiritual aspects, as well as the context or environment surrounding the person. Attitudes essential for the relationship competency, include, but are not limited to, (1) intellectual curiosity and flexibility, (2) open-mindedness, (3) belief in the capacity for change, (4) appreciation of individual and cultural difference (ICD), (5) personal integrity and honesty, and (6) belief in the value of self-awareness. These attitudes can be seen as embedded in the pedagogy of integrative experiences described earlier.

Assessment

Psychological assessment is a complex, integrative, and conceptual activity that involves deriving inferences from multiple sources of information to achieve a comprehensive understanding of a client or client system. It involves the ability to measure and formulate degree of need and mental status, develop psychological profiles in response to particular referral problems, and evaluate outcome with tests, measures, and diagnostic interviewing across a range of client populations. The assessment competency requires students to integrate data from multiple sources, to effectively answer referral questions, and to communicate their inferences and recommendations clearly.

Intervention

Intervention includes activities that promote, restore, sustain, and/or enhance positive functioning and a sense of well-being in clients through preventive, developmental, and/or remedial services. Development of competence in intervention requires understanding of theory and its application in personality, psychopathology, change processes, and the interactions and influences of social, environmental, cultural, and physiological factors.

Research and Evaluation

To give particular emphasis to the practitioner-scholar position on science and practice issues, the majority of the material on the research and evaluation competency has been presented earlier in the section titled *Psychological Science and Education for Professional Psychology*.

Consultation and Education

Consultation refers to the planned collaborative interaction between the professional psychologist and one or more clients or colleagues, in relation to an identified problem area or program. Psychological consultation is an explicit intervention process that is based on principles and procedures found within psychology and related disciplines, in which the professional psychologist has no *direct* control of the actual change process. Education is the directed facilitation by the professional psychologist of the growth of knowledge, skills, and attitudes in the learner. Competency in education involves knowledge about models of learning and pedagogy, as well as the foundations of, and innovations in, instructional design and evaluation. It also involves skill building in facilitating student knowledge acquisition and scholarly and personal development.

Management and Supervision

Management includes activities that organize, direct, or control services offered to the public. These activities require effective functioning within organizations, an understanding of financial management and strategic planning, and the ability to influence public and organizational policy. Supervision is a form of management and represents an intervention provided by a more senior member of a profession to a more junior member or members of the same profession. The relationship is evaluative, extends over time, and has the simultaneous purposes of enhancing the professional functioning of the more junior person, monitoring the quality of professional services offered to the client, and serving as a gatekeeper of those who are to enter the particular profession.

Because the majority of graduates of professional psychology programs are ultimately employed in positions requiring management and supervisory skills, NCSPP is convinced that this competency should occupy an enhanced status in the core curriculum. Going further, professional psychology programs

should support advanced preparation for leadership, advocacy, and public and social policy planning roles.

Diversity

The diversity competency requires the ability to identify and understand issues of Individual and Cultural differences (ICD), and issues of power, privilege, and oppression. This understanding informs and influences all professional functions and activities. Competency in diversity requires an affirmation of the richness of human differences, ideas, and beliefs. Important dimensions of diversity include (but are not limited to) age, disability and health, ethnicity, gender, language, national origin, race, religion and spirituality, sexual orientation, and social economic status, as well as the intersection of these multiple identities and multiple statuses. Exploration of power differentials, power dynamics, and privilege is at the core of understanding diversity issues and their impact on social structures and oppression.

In the last few years, NCSPP spelled out developmental achievement levels for each of the competencies. This work explains what is to be expected of students at each level as they move through their programs.

Knowledge Base

From the competency perspective, professional education is more inclusive but is, in fact, no less scientific. The professional competencies continue to be related to an evolving and developing knowledge base that should include the following areas: (1) biological bases of behavior, (2) cognitive-affective bases of behavior, (3) cultural bases of behavior, (4) dysfunctional behavior and psychopathology, (5) the historical and philosophical context of psychology, (6) life span development, (7) professional ethics and standards, (8) psychological measurement, (9) social bases of behavior, and (10) theories of individual and systems functioning and change. This has come to be known as the core curriculum and/or broad and general education.

Elements of Practice

In a context that reflects (1) the core competencies, (2) the wide range of activities in which professional psychologists are currently engaged and will be in the future, and (3) the likelihood of an expanded scope of practice in the future, the practitioner-scholar model has long emphasized the need to prepare psychologists for multiple roles. The educational process is not seen as one in which the student is pumped full of knowledge and skills that are then dispensed to the public for a price, but as an experience through which the personal and professional self of each student is systematically developed, consistent with ideas about reflective practice.

Multiple and Expanding Roles

The primary task of education in professional psychology is preparation for effective functioning in the multiple roles graduates will fill during the course of their careers. Conceptually, the idea of competencies is closely related to the idea of professional roles. Preparation for the wide array of roles evident in professional life is dependent on an education with a very broad understanding of each of the professional competencies.

The Self of the Professional Psychologist and Reflective Practice

Preparation in professional psychology involves the education of the personal and professional selves of students so as to develop the habits of reflective practice and lifelong learning. Awareness of self and self-other relatedness is central.

Both academic and interpersonal material relevant to inequalities of power and authority are key educational topics. All of these issues underscore the importance of training in both reflective practice and person-focused social and interpersonal sorts of professional development experiences that provide a context in which such conversations can occur. Different aspects of self and different ethical principles become relevant as professional psychologists move through multiple roles. Programs that educate professional psychologists carry an obligation to systematically attend to, develop, and evaluate the personal and interpersonal fitness of graduates.

Practicum and Internship Training

Recognizing that, for supervised practice in psychology, there is no one correct training model, the practitioner-scholar model strongly supports a diversity of practicum and internship models within professional psychology. The desirability of integrative service, inquiry, and teaching models in clinical training was reaffirmed. A strong relationship between educational coursework and field training is particularly valued.

Systematic Evaluation

NCSPP professional psychology educators have committed themselves to outcome and evaluation research to assess traditional and alternative models of training and service. There continues to be general affirmation of the need for systematic evaluation of students, programs (including faculty and supervisors), and services. Evaluation of professional competence in field experiences is based on a three-way learning contract among the student, the practicum or internship center staff, and the program faculty. Critical aspects of the evaluation process involve supervision; direct observation; and instruments and procedures incorporating self-evaluation, student peer evaluation, and program field faculty evaluation. Critical to an institution's ability to vouch for the quality of its graduates, competency-based clinical examinations are viewed as central to the training process. Such examinations should assess competence relevant to professional practice using

multimethod evaluation techniques (that might include live evaluation or other direct scrutiny of students' work by means of audiotapes and videotapes, written case presentations, examinations by faculty, and presentations to peers) as generally exemplified in the American Board of Professional Psychology (ABPP) examination. Professional psychology programs should not attend only to academic progress but are also responsible for evaluating students with regard to those personal attitudes, aptitudes, and values that appear likely to predict future professional competence. As the competency language (with knowledge, skills, and attitudes as the basic definition) has come to dominate psychology and medicine, these elements can be seen as part of *attitudes*.

Social Responsibility, Diversity, and Gender

The model has taken leadership in emphasizing the social nature of professional psychology and the social responsibilities of both programs and graduates. Against the backdrop of the history of education and training in psychology, NCSPP has believed that issues related to diversity and gender needed additional, particular attention as exemplified by developing its seventh competency of diversity.

Social Nature of Professional Psychology

Education must be socially responsive and responsible. A profession of psychology can be justified only if it meets the fundamental needs of the larger society. Professional practice in psychology and the education for that practice are fundamentally social, not private or individual. This social vision and understanding are apparent in a number of areas: (1) an examination of the importance of the socially situated role of psychologists, (2) the basic and underlying aspect of the relationship competency, (3) historical analyses of the social influences on the core curriculum in professional, (4) the development of the competency areas, (5) the necessity of directing explicit attention to the organizational contexts and cultures in which education occurs, (6) the interpersonal nature of reflective education, and (7) the development of a broad-based definition of social responsibility.

Implications for Professional Psychology

This work continues to have a number of implications for the training and practice of professional psychologists:

The competencies as a heuristic. The competencies and the competency model can be seen as a heuristic for the field. The overall model has and will continue to help psychology understand the growth of competencies from the beginning student to the graduate pursuing ABPP certification. The competency model organizes conversations about both current and expanded roles for the future, educational experiences leading to those roles, and credentialing for them.

Broad professional psychology education. Students are being educated in the broadest sense to be professional psychologists, not, to mention some stereotypes, psychotherapists with a few science courses or researchers with a few clinical courses.

Preparation for actual roles as practitioners. Graduates explicitly educated to become practitioners are likely to be better prepared to do what they actually do after graduation, as clearly as the realities can be known or the future foreseen. Often, practitioners trained in earlier times with narrower models had much to learn on their own about the professional world after receiving their doctorates.

Education for service delivery. Students who seek a career in psychology as a helping profession are educated in a model that is committed to service delivery as a primary value. Practitioners, specifically trained to serve in that capacity, may be more likely to retain connections with their programs and institutions in a way that will facilitate continuing education, enhance graduates' pride and acknowledgment of value in the work they are doing, and enhance mutual respect among differently trained psychologists.

Education relevant to social issues and diversity. Beyond the problems of individuals, the practitioner-scholar model provides the focus and background for dealing with social issues, with both sensitivity to and education about diversity.

More attention to consultation and education and to supervision and management. Expanded curricular attention still needs to be given to consultation and education and to supervision and management to reinforce that these activities are critical aspects of professional psychology practice.

Attitudes critical to professional development. This model values attention to attitudes as a critical aspect of professional development. These are the same attitudes, which along with knowledge and skills, comprise competencies. The competency movement now provides the dominant organizing vision in the education of professional psychologists as well as physicians. Further, the model articulates how the self of the student professional psychologist should be involved in reflective professional education and practice. Without substituting self for science, this reflexive model is consistent with, not divergent from, mainstream professional values. Education and practice involve the person of the professional psychologist, not only the narrow technical application of some scientific knowledge.

Science as implemented in psychological practice. Over the long term, on the basis of concepts like the local clinical scientist and disciplined inquiry, the most significant impact of this model may be to focus attention on how science is actually – and could be more effectively – applied in psychological practice. Though some of these conceptual pairs do denote bona fide and important differences, the practice of psychology as characterized in the concepts and examples of the local clinical scientist may bring these groups together because they identify common, desirable processes for practice that are based on disciplined inquiry and are in a language that is familiar across subcultures.

In many ways, there is a great degree of overlap between the practitioner-scholar model and the scientist-practitioner model. In best examples of each, there might be little disagreement, if there were agreement on which are the best examples. In less ideal cases, practitioner-scholar model advocates still might say that the Boulder model includes too much emphasis on the examination and production of career-irrelevant research, whereas the scientist-practitioner advocates might say that the science disappears in the focus

on clinical work. At bottom, though, there is a common interest; psychological science is an integrative, disciplined, reflexive way of thinking that helps all psychologists to work together to solve great and important human problems.

See also: Diagnostic Processes in Clinical Psychology and Psychotherapy; Multimodal Therapy in Clinical Psychology; Psychological Treatments Delivered Nontraditionally; Therapist–Patient Relationship; Training in Clinical Psychology in the United States: Scientist–Practitioner Model.

Bibliography

American Psychological Association Office of Program Consultation and Accreditation, 2011. Book 1: Guidelines and Principles for Accreditation of Programs in Professional Psychology; Book 2: Accreditation Operating Procedures. American Psychological Association, Washington, DC.

Kenkel, Mary Beth, Peterson, Roger L. (Eds.), 2010. Competency-Based Education for Professional Psychology. American Psychological Association, Washington, DC.

McFall, Richard M., 2006. Doctoral training in clinical psychology. Annual Review of Clinical Psychology 2, 9.1–9.29.

McFall, Richard M., 2012. Psychological clinical science accreditation system: FAQs and facts. The Behavior Therapist 55, 11–15.

Peterson, R.L., 1992. The social, relational, and intellectual context of the core curriculum and the San Antonio conference. In: Peterson, R.L., McHolland, J.D., Bent, R.J., Davis-Russell, E., Edwall, G.E., Magidson, E., Polite, K., Singer, D.L., Stricker, G. (Eds.), The Core Curriculum in Professional Psychology. American Psychological Association and National Council of Schools of Professional Psychology. Washington, DC, pp. 3–12.

Peterson, Donald R., 1991. Connection and disconnection of research and practice in the education of professional psychologists. American Psychologist 46, 422–429.

Peterson, Donald R., 1995. The reflective educator. American Psychologist 50, 975–983.

Peterson, Donald R., 1997. Educating Professional Psychologists: History and Guiding Conception. APA Books, Washington, DC.

Peterson, Donald R., 1996a. Making conversation possible. Applied and Preventive Psychology 5, 17–18.

Peterson, Donald R., 1996b. Making psychology indispensable. Applied and Preventive Psychology 5, 1–8.

Peterson, Roger L., Peterson, Donald R., Abrams, Jules C., Stricker, George, 1997. The National Council of Schools and Programs of Professional Psychology educational model. Professional Psychology: Research and Practice 28, 373–386.

Peterson, Roger L., Peterson, Donald R., Abrams, Jules C., Stricker, George, Ducheny, Kelly, 2010. The National Council of Schools and Programs of Professional Psychology education model 2009. In: Kenkel, Mary Beth, Peterson, Roger L. (Eds.), Competency-Based Education for Professional Psychology. American Psychological Association, Washington, DC, pp. 13–42.

Peterson, Donald R., Peterson, Roger L., 1997. Ways of knowing in a profession: toward an epistemology for the education of professional psychologists. In: Peterson, Donald R. (Ed.), Educating Professional Psychologists: History and Guiding Conception. APA Books, Washington, DC, pp. 191–228.

Stricker, George, Trierweiler, Steven J., 1995. The local clinical scientist: a bridge between science and practice. American Psychologist 46, 995–1002.

Trierweiler, S.J., Stricker, G., 1992. Research and evaluation competency: training the local clinical scientist. In: Peterson, R.L., McHolland, J., Bent, R.J., Davis-Russell, E., Edwall, G.E., Magidson, E., Polite, K., Singer, D.L., Stricker, G. (Eds.), The Core Curriculum in Professional Psychology. American Psychological Association and National Council of Schools of Professional Psychology. Washington, DC, pp. 103–113.

Trierweiler, Steven J., Stricker, George, 1992. Research and evaluation competency: training the local clinical scientist. In: Peterson, Roger L., McHolland, James D., Bent, Russell J., Russell, Elizabeth Davis-, Edwall, Glenace, Magidson, Ethel, Polite, Kenneth, Singer, David L., Stricker, George (Eds.), The Core Curriculum in Professional Psychology. American Psychological Association and National Council of Schools of Professional Psychology, Washington, DC, pp. 103–113.

Trierweiler, Steven J., Stricker, George, 1998. The Scientific Practice of Professional Psychology. Plenum, New York.

Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Bonnie R Strickland, University of Massachusetts, Amherst, MA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15836–15840, © 2001, Elsevier Ltd., with an updated Bibliography section supplied by the Editor.

Abstract

Although some argue that the science and practice of psychology is as old as recorded history, contemporary psychology, as an evolving discipline, emerged at the end of the nineteenth century. From the beginning, the science and practice (or application) of psychology were inextricably woven together although few early psychologists embraced this model. This article reviews the history of clinical psychology training in the United States, with an emphasis on what is known as the "scientist–practitioner model."

Since the mid-twentieth century, scientist-practitioner training in clinical psychology has been the preeminent educational model for clinical psychologists. In 1949, 71 psychologists and educators held a 2-week meeting in Boulder, Colorado to define clinical psychology and propose a model of education and training. They recommended that clinical psychologists complete a doctor of philosophy degree in general psychology, marked by an empirical base and a grounding in the scientific method. They further recommended that such training be accompanied by practical experience including at least a year of a predoctoral internship. The report of the Boulder Conference set the stage for clinical psychologists to become both scientists and practitioners, a unique model among the major disciplines and professions. The scientist–practitioner model was embraced by and continues in most university-based, PhD-granting programs that educate and train clinical psychologists.

Although some argue that the science and practice of psychology is as old as recorded history, contemporary psychology, as an evolving discipline, emerged at the end of the nineteenth century. From the beginning, the science and practice (or application) of psychology were inextricably woven together although few early psychologists embraced this model. Experimental psychology was born in 1879 in Willhelm Wundt's laboratory in Leipzig, Germany and brought to the USA by numerous students who worked with him. They pursued rigorous scientific investigations of animal and human subjects specifically around learning, memory, and sensory processes. At about the same time in Vienna, Austria, Sigmund Freud was developing his 'talking cure' psychoanalysis, a clinical approach to the influence of trauma or past events on disruptive emotional symptoms often having to do with cognitive processes such as 'repression' or 'transference' (*see* Psychoanalysis, History of). But whatever the overlapping interest in learning, memory, and relearning, proponents of these two approaches had little to do with each other and a reapproachment between the scientific investigations of behavior and the clinical techniques of psychoanalysis was not offered until some 50 years later (Dollard and Miller, 1950).

The late nineteenth century was also a time of considerable debate about the etiology of mental illness. Physicians, such as Emil Kraepelin who had also studied with Wundt, embraced an organic model which proposed that manic-depressive psychosis and dementia-praecox (schizophrenia) would be found to have a biological basis similar to syphilis, for which the spirochetes causing the brain degeneration had just been found. Others assumed that psychological processes composed the underlying etiology of emotional and mental disorders. These debates were further couched within discussions about whether investigations of general behavior or a concern with individual differences would lead to new understandings about the nature of human behavior.

Lightner Witmer, the father of clinical psychology, was an extraordinary exemplar of a scientist–practitioner (**Table 1**). In his attempts to use psychology to help children who were having difficulty in school, Witmer was the first to write about the need to integrate the basic science of psychology and its

Table 1 A comparison of practitioner–researcher and scientist–practitioner models

	Practitioner	*Scientist*
Training	Practice and research oriented	Research oriented
Coursework oriented	Application and practice oriented	Research and application
Research	Applied/consumer	Basic and applied
Department culture	Practice oriented	Research oriented
Training goals	Practitioner–researcher	Scientist–practitioner
Career settings	Community mental health centers, colleges counseling centers, clinics, hospitals private practice	Colleges and universities, research agencies, medical schools
Admissions criteria	Human service interest and/or experience, interpersonal skills, intellectual abilities	Research interest and/or experience, interpersonal skills, intellectual abilities

application. In fact, he believed that there was no valid distinction between a pure science and an applied science.

Witmer founded the first Psychological Clinic at the University of Pennsylvania in 1896 and the first clinical journal, *The Psychological Clinic*, in 1907 (*see* Clinical Psychology in North America, History of). In the opening article of the journal, Witmer writes movingly of the difficulties he had finding ways to make psychology useful in his treatment of clients. One of his first cases was a 14-year-old boy referred by his teacher because he could not seem to learn to spell. Witmer (1996) looked to basic psychology but could not find "that the science of psychology had ever addressed itself to the ascertainment of the causes and treatment of a deficiency in spelling. Yet, here was a simple developmental defect of memory; and memory is a mental process of which the science of psychology is supposed to furnish the only authoritative knowledge. It appeared to me that if psychology was worth anything to me or to others it should be able to assist the efforts of a teacher in a retarded case of this kind" (p. 2). Witmer goes on to quote the President of the American Association for the Advancement of Science noting that "the final test of the value of what is called science is in its applicability…the pure and the applied sciences advance in a single front. What retards the progress of one, retards the progress of the other; what fosters one fosters the other" (Witmer, 1996: p. 2).

In an effort to bridge the science and practice of psychology, Witmer linked his psychological clinic to the psychology laboratory at the University of Pennsylvania. His clinical cases also comprised part of the curriculum for the child development courses. Witmer proposed a 'scheme of practical work in psychology' that laid the foundation of clinical, counseling, and school psychology. This consisted of the investigation of the phenomena of mental development by means of statistical and clinical methods, a clinic supplemented by a training school, practical experience for those working with children, and the training of students for a new profession – the psychological expert. Forty years later these guidelines were adopted almost in their entirety by the Boulder Conference on education and training in clinical psychology.

Numerous other children's psychological clinics were developed during the first few decades of the twentieth century. Usually these were associated with universities but some were established to provide clinical intervention for juvenile delinquents. Clinical psychologists were also working in facilities for the mentally retarded and became heavily involved in intelligence testing. Both basic and applied psychologists extended psychology's expertise in 'mental' testing to the assessment of recruits into the military during World War I. Unfortunately, this began one of the most tragic episodes in psychology's history – the use of intelligence testing to keep thousands upon thousands of immigrants from relocating in the USA.

Debates about appropriate educational and training models for the proper roles and functions of psychology began in the earliest days. Most psychologists, especially those who had trained with Wundt, envisioned psychology as an evolving science which would explore the intricate processes of consciousness and memory, unlock the mysteries of sensory processes, and answer questions about how people perceive and make sense of the world. Psychological findings would be based on the most rigorous of experimental methods. Other psychologists, faced with practical and poignant human problems, such as a child's failure to learn, believed that psychology could advance knowledge through a consideration of practical issues and problems and be immediately applied to help people in need.

While the parallels are not perfect, these differences in epistemological styles reflect a philosophical conflict of some 2500 years. A Socratic view of education places emphasis on curiosity and the accumulation of knowledge for its own sake while the Sophists emphasize human welfare and believe "man is the measure of all things" (Kimble, 1984). Within early psychology, William James (1907) described the history of philosophy as a constant battle between these two sharply contradictory intellectual styles. He described the 'tough-minded' skeptical, empirical scientific psychologists who depend only on facts versus the 'tender-minded' rational and idealistic sort who base their efforts on principles. James commented on the antagonisms between the 'unrefined, callous, or brutal' tough-minded and the 'sentimental, soft-headed,' tender-minded (James, 1907). Later, Snow (1964) wrote extensively about the clash of the 'two cultures' of science and humanism and Kimble (1984) extended this observation into psychology. Kimble suggests that contemporary psychology continues to have an identity problem in that the two cultures of the scientist and practitioner cannot agree on a coherent set of values.

This clash of cultures, as exemplified by the different values embraced by the scientific versus the practitioner psychologist has always been with us. As early as the turn of the twentieth century, applied and clinical psychologists were complaining that the newly formed American Psychological Association (APA) was not meeting their needs. Practicing psychologists were particularly concerned that they be recognized through some formal mechanism as having been trained in psychology. In 1917, an American Association of Clinical Psychology was proposed as a vehicle whereby psychologists who were qualified to perform psychological testing could be identified. With some reluctance, APA agreed to certify some members as 'consulting psychologists' and formed a Clinical Section. The primary mission of APA, however, was to promote psychology as a science.

As practitioners involved themselves in psychological testing, child guidance, and helping individuals with 'adjustment' problems, issues of ethics, licensing, and standardization of training became paramount. In the early 1930s, the New York Psychological Association in an effort to meet practitioner needs at a national level became the Association of Counseling Psychologists. They became so strong that in 1937 the Clinical Section of APA disbanded to join them in the new American Association for Applied Psychology (AAAP).

AAAP began to move across the national scene to meet the needs of practitioners and state affiliates in regard to uniform training, certification, public relations, and ethical conduct. The group proposed a model certification act to gain licensure or registration for qualified psychologists and published a membership directory to identify such practitioners for the public. They proposed an ethical code and pushed for a curriculum, both undergraduate and graduate, that would include applied and clinical psychology. Further, Carl Rodgers (1939) proposed a supervised, 1-year internship training for

clinical psychology graduate students and discussed the idea of a doctor of psychology degree, rather than the PhD (Napoli, 1981).

The advent of World War II had a powerful influence on organized psychology. In an effort to present a unified front for responding to the war efforts and the needs of the military, an Emergency Committee in Psychology was established that included representatives from all of the national psychology organizations (now up to six). A subcommittee on Survey and Planning for Psychology, chaired by Robert Yerkes, was formed to begin making long-range plans for psychology's future. Their first report deplored the splintering of psychology into various groups and proposed an increased emphasis on professional activities within the APA. They further proposed a new national organization that would serve both science and practice. When the war was over, the stage was set for the APA to reorganize to serve the needs of psychologists, whether their interests be basic or applied psychology. The two cultures of science and practice in psychology would be subsumed, once again, within one organization.

Psychologists had served in a number of capacities during World War II, ranging from the regular duties of the infantry to preparing propaganda and disseminating disinformation. Basic psychological scientists became involved in human factors designing user-friendly instruments for vehicles and planes. Clinical psychologists found themselves providing counseling and psychotherapy for soldiers and by the end of the war, some 450 clinical psychologists were serving in the US Army.

At the urging of the National Research Council, the disparate national groups of psychologists began to meet together at the end of the war. Acting on recommendations from the Yerkes committee, the APA revised its by-laws to "advance psychology as science, a profession, and a means of promoting human welfare" and incorporate other groups and representatives from state associations. The AAAP became the Clinical Division of APA and professional psychologists began to feel more at home in the then 4000 member organization.

The National Institute of Mental Health and the Veteran's Administration were eager for clinical psychologists to continue to provide services for veterans who were suffering psychological and emotional problems. In fact, they began to subsidize training to increase the numbers of psychiatrists and social workers. They were willing to do the same for clinical psychology if appropriate training programs could be identified. APA provided the names of 22 such university programs and appointed a Committee of Training in Clinical Psychology to determine how best clinical psychologists might be trained (American Psychological Association, 1947). The report of this Committee became the basis for a 1949 conference held in Boulder, Colorado which would recommend a revolutionary program of education and training. Clinical psychologists would be trained as both scientists and practitioners. Such an integrative model, unique among the major sciences and professions, would bridge scientific and humanistic traditions.

The Boulder conference was supported by the National Institute of Mental Health and the United States Public Health Service. Seventy-one representatives from universities, mental health service agencies, and allied professions met daily for 2 weeks to define clinical psychology and develop consensual standards for education and training. A 252-page book, *Training in Clinical Psychology*, reported the recommendations of the conference which were both comprehensive and specific (Raimy, 1950). Clinical students' attributes were enumerated, faculty described, and curriculum proposed. Clinical psychologists were to be trained in general psychology and demonstrate research competence culminating in a doctoral (PhD) dissertation. Further, students would engage in field training and a predoctoral internship so that they would be skilled in clinical assessment, diagnosis, and treatment. Such education and training was to take place in universities and internship facilities accredited under the auspices of the APA.

During the second half of the twentieth century, the Boulder or scientist–practitioner model became the preeminent mode of education and training in clinical psychology. Admissions to university-based, PhD-granting programs were, and continue to be, highly selective. Academic requirements include general areas of psychology such as history and systems and psychopathology as well as cognitive, developmental, experimental, personality, physiological, and social psychology. Students are expected to learn to conduct research and be familiar with the various statistical methods necessary to evaluate results. Further requirements include practical training in assessment and intervention under close supervision including a 1-year predoctoral internship. The scientist–practitioner model has been reaffirmed at each major graduate educational conference held since Boulder although a professional model was also recommended at a 1973 conference held in Vail, Colorado (Strickland, 1985, 1987, 1998).

The scientist–practitioner model is not reserved solely for clinical psychology. Other applied areas in psychology such as counseling and industrial–organizational psychology also embrace this paradigm as do the specialty areas of aging and health. For those who believe that psychology is a continually evolving discipline and field, basic training in science would seem essential in producing new knowledge. Further, practical training within such a model should lead to state-of-the-art applications. An integration of science and practice allows the continued interplay between scientific knowledge and clinical application where each informs the other.

Many would maintain that the development of the scientist–practitioner model has been one of psychology's finest achievements. Certainly for a newly evolving science which has clear implications for application, this novel educational and training approach gives the field vitality and direction. Embracing the two cultures of science and humanism, the scientist–practitioner model has not been an easy one for individual psychologists to achieve. But, scientist–practitioners have served psychology and society well both through advancing knowledge and applying that knowledge for the greater good.

See also: Clinical Psychology in Europe, History of; Clinical Psychology in North America, History of; Diagnostic Processes in Clinical Psychology and Psychotherapy; Mental Disorders, Epidemiology of; Multimodal Therapy in Clinical Psychology; Training in Clinical Psychology in the United States: Practitioner Model.

Bibliography

American Psychological Association Committee on Training in Clinical Psychology (D. Shakow, Chairman), 1947. Recommended graduate training programs in clinical psychology. American Psychologist 2, 539–558.

Dollard, J., Miller, N.E., 1950. Personality and Psychotherapy. McGraw-Hill, New York.

James, W., 1907. Pragmatism: A New Name for Some Old Ways of Thinking. Longmans, New York.

Kimble, G., 1984. Psychology's two cultures. American Psychologist 39 (8), 833–839.

Lilienfeld, S.O., Lynn, S.J., Lohr, J.M. (Eds.), 2003. Science and Pseudoscience in Clinical Psychology. Guilford Press, New York.

Napoli, D.S., 1981. Architects of Adjustment. Kennikat Press, Port Washington, NY.

Raimy, V.C., 1950. Training in Clinical Psychology. Prentice-Hall, New York.

Rodgers, C.R., 1939. The Clinical Treatment of the Problem Child. Houghton Mifflin, Boston, MA.

Snow, C.P., 1964. The Two Cultures: And a Second Look. The New American Library, New York.

Strickland, B.R., 1985. Over the boulder and through the vail. The Clinical Psychologist 38, 52–56.

Strickland, B.R., 1987. On the threshold of the second century of psychology. American Psychologist 42, 1055–1056.

Strickland, B.R., 1988. Clinical psychology comes of age. American Psychologist 43, 104–107.

Strickland, B.R., 1998. History and introduction to clinical psychology. In: Cullari, S. (Ed.), Foundations of Clinical Psychology. Allyn and Bacon, Boston, pp. 1–26.

Trull, T.J., Prinstein, M.J., 2013. Clinical Psychology, eighth ed. Wadsworth, Belmont, CA.

Witmer, L., 1996. Clinical psychology. American Psychologist 51 (3), 248–251. Reprinted from: The Psychological Clinic 1, 1–9 (1907).

Transaction Costs and Property Rights

Oliver E Williamson, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 23, pp. 15840–15845, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

Property rights concepts and thinking go back to antiquity, but it was only in the 1960s, when the concept of transaction costs was introduced into the analysis of property rights by Ronald Coase, that the modern treatment of property rights got underway. Applications to externalities, common pool problems, variable access, the theory of the firm, rent seeking, economic development and reform, and *de facto* property rights are sketched. Although property rights concepts and property rights thinking have gone a long way to reshape our understanding of economic institutions and effect public policy reform, the early ambitions of the economics of property rights have been realized incompletely. Here, as elsewhere, there can be too much of a good thing. Examples include unduly sanguine recommendations to jetison regulation, over-reaching interpretations of the modern corporation, and undue reliance on mass and rapid privatization as the way to implement economic reform.

Transaction costs and property rights are related, important, and expansive concepts. Both move beyond technology to introduce law and organization. Both deal with issues of first-order economic importance (as against nuances). Because both have been difficult to operationalize, both have been used expansively and tautologically.

Property rights is the older concept, going back to antiquity. The use of transaction cost reasoning in the 1960s brought new life to the property rights movement. Not only is the economics of property rights basic, enduring, and important in its own right, but it also helped to usher in and remains an important part of the new institutional economics: 'Modern institutional economics focuses on the institution of property and on the systems of norms governing the acquisition or transfer of property rights' (Furubotn and Richter, 1991, p. 7). Of the various categories of property rights, the right of ownership – which consists of the right to use an asset, the right to appropriate the returns from an asset, and the right to change its form, substance, or location (Furubotn and Richter, 1991, p. 6) – is the most important.

The article begins with a sketch of the antecedents, then some of the leading property rights successes are examined, followed by the limitations and the conclusion.

Antecedents

Frank Michelman (1967) traces property rights concepts and thinking from Aristotle through the Christian Fathers, John Locke, David Hume, Jeremy Bentham, and Oliver Wendell Holmes. Despite this early and continuing interest, the economics of property rights 'has not been discussed analytically until very recently' (Demsetz, 1998, p. 144). The introduction of transaction cost reasoning into the study of property rights in the 1960s is responsible for much of the analytical advance.

Modern Property Rights

Externalities

Work on the economics for property rights flourished in the 1960s. Leading contributors included Ronald Coase (1959, 1960), Armen Alchian (1961, 1965), and Harold Demsetz (1967). Much of this work contested the orthodox interpretation of externalities as 'market failures' for which government intervention 'in the form of taxes, or subsidies, or regulation' was presumptively warranted. Rather than take property rights assignments as given, Coase moved the analysis back a step to examine the ramifications of reassigning property rights. His treatment brought out 'the reciprocal nature of the relationship which tends to be ignored by economists who, following Pigou, approach the problem in terms of a difference between private and social producers but fail to make clear that the suppression of the harm which A inflicts on B inevitably inflicts harm on A. The problem is to avoid the more serious harm' (Coase, 1959, p. 24). Assessing the more serious harm entailed introducing transaction costs.

Coase proceeded in a two-part way. Starting with the assumption of zero transaction costs, Coase advanced the (then) revolutionary proposition that 'the delimitation of rights is an essential prelude to market transactions; but the ultimate result (which maximizes the value of production) is independent of the legal [award of property rights to one party rather than the other]' (Coase, 1959, p. 25). This is because, distributional considerations aside, the parties will bargain to the same efficient result whichever way property rights are awarded.

For Coase, however, this was an expositional device. The real world is one of positive transaction costs, whereupon the assignment of property rights one way or another does matter. Efficiency now turns on the differential transaction costs of assigning property rights one way rather than another. Coase furthermore contested the idea that a 'benign government' could be used to correct market failures, it being the case that all feasible forms of organization are flawed (Coase, 1964). He also advised that total as well as marginal conditions needed to be included in a comparative assessment: there is a need to 'compare the total product yielded by alternative social arrangements' (Coase, 1960, p. 42). This comparative approach in which transaction costs are featured reshaped the study of property rights, market failures, and economic organization in the ensuing years, up to and including the present.

Common Pool Problems

Common pool problems arise in a variety of contexts, of which common grazing land (Hardin, 1968), fisheries (Turvey, 1964), petroleum extraction (Libecap and Wiggins, 1985), and water and air pollution are familiar examples. Whereas the externalities examined by Coase were predominantly of a bilateral kind, the common pool problem poses the need to coordinate many actors. Additional problems arise because (1) different actors are differently situated, (2) the relevant information is dispersed and costly to verify, (3) strategizing and politicking often arise, and consent is needed to, (4) agree on an overarching plan, and (5) craft an organization to implement whatever plan is agreed to. The need for political action and regulation often arises in efforts to reach and enforce collective agreements in the multiactor context.

Recourse to public ordering, is not, however, the only or always the best response. Private ordering can and has successfully mitigated many common pool problems (Furubotn and Richter, 1997, p. 102). Government assistance – for unitizing common pools (such as oil) or for creating markets (as in the creation of tradeable property rights for air pollution) – is often the cost-effective way to proceed.

Variable Access

The simplest transaction is an outright sale. But even that is complicated by the question of how the property right is defined. As Demsetz observes, 'To have the right to till a parcel of land is not equivalent to having all the rights to the land. Others may have rights both to minerals contained in the land and to trespass across the parcel. The identification of the owner of the asset is made more difficult by the virtually infinite number of different rights that can be associated with use of a given asset' (1998, p.145).

Access arrangements other than sale include leasing, tenant farming, inside contracting, and franchising. These can vary in length and protections. Incentives for due care (where the concern is over negligence by the lessee) and assured compensation for unexpired improvements (where the lessor stands to gain by refusing compensation for improvements made by the lessee that extend beyond the length of the lease) are common concerns. Access terms is also an area where the risk of changes in the political rules of the game is a chronic concern, the imposition of rent control being one example and the arbitrary exercise of imminent domain and 'takings' by the state being another (Michelman, 1967).

Theory of the Firm

The theory of the firm poses numerous property rights issues, of which the most famous is the separation of ownership from control to which Adolf Berle and Gardiner Means (1932) made early reference in the context of the large corporation where shareholding is widely dispersed. Thus although shareholders remain *de jure* owners in the large corporation, effective or *de facto* ownership purportedly passes to the management. Berle and Means thereupon questioned the presumption that firms are reliably given to profit maximization.

Analytical reactions to this condition of separation have varied. One response has been to dismiss the issue by assuming that competition in product and capital markets is relentlessly taut. As Tjalling Koopmans has observed, however, 'if this is the basis of our belief in profit maximization, then we should postulate that basis itself and not the profit maximization which it implies in certain circumstances …. It would lead us to expect profit maximization to be most clearly exhibited in industries where entry is easiest and where the struggle for survival is keenest, and would present us with the further challenge to analyze what circumstances give to an industry that character' (1957, pp. 140–1).

Another response to the separation of ownership from control is to reformulate the objective function to reflect *de facto* control rights. The sales maximization hypothesis (Baume, 1959), growth maximization hypothesis (Marris, 1964), and managerial discretion hypothesis (Williamson, 1964) are in this spirit. Still another response is to reconceptualize the problem of organization in contracting terms. The 'nexus of contract' approach (Jensen and Meckling, 1976) brings out agency cost features to which monitoring is a response, while the governance structure approach (Williamson, 1975, 1985) treats organization form as a decision variable. According to the latter, hierarchical decomposition principles – especially between centralized and decentralized modes of organization – matter. Organization is thus viewed both as a source of the managerial discretion problem and a partial solution, in that the large corporation can be reconfigured to mitigate some of its own failures.

A 'modern' property rights theory of the firm, of which Oliver Hart (1995) is the principal architect, has taken shape more recently. This theory works out of an incomplete contracting setup. The central issue with which this theory is concerned is that of vertical integration: who should own the assets of two related stages of production? This theory illustrates both the strengths and limitations of property rights reasoning. To its credit, the analysis makes clear how the ownership of property rights influences *ex ante* investments. But this exclusive focus on property rights comes at a cost, in that problems of *ex post* maladaptation are suppressed (by assuming that the parties will costly bargain to an efficient result, whatever the assignment of property rights). The idea that organization matters – that the management of some transactions is better done by markets and of others by hierarchy – is sacrificed in the process.

Rent Seeking

The *de facto* ownership of property rights poses numerous intertemporal issues, of which rent seeking is one. The basic rent seeking proposition is this: the creation of 'rights' to which supernormal returns potentially accrue invites *ex ante* prepositioning, thereby to better qualify the applicant/bidder for the job or contract in question. For example, the award of a natural monopoly franchise or of a favored class of employment (Civil Service) invites parties to politick to receive the award and to make other efforts to improve their eligibility. Rent protective efforts also ensue (Tollison, 1998).

In the degree to which such responses went unnoticed or were ignored by economists, such costs were mistakenly omitted from the original calculus. Upon 'discovering' such costs, most economists declared them to be wasteful. That, however, assumes that economics trumps politics (Stigler, 1992). If political choices are to be respected, then the best course of action is to recognize that prepositioning is a predictable consequence of efforts to create new classes of rights, whereupon prepositioning effects are folded into the project design calculus. In that event, rent seeking, and its attenuation, are included in the 'approved plan.' Those who condemn rent seeking are loathe to come to terms with this proposition.

Economic Development and Reform

Early development theory emphasized 'macroeconomic accounting aggregates such as savings and the balance of payments, and the relative balance between broadly defined "sectors" such as "industry" and "agriculture"' (Lal, 1983, p. 5). When that prescription proved disappointing, the pendulum swung in the opposite direction. The new imperative was to activate the market and 'get the prices right' (Lal, 1983, p. 107), but that too was overly simple.

'Getting the property rights right' seemed more responsive to the pressing needs for reform in Eastern Europe and the former Soviet Union. But while 'an essential part of development policy is the creation of polities that will create and enforce efficient property rights, … we know very little about how to create such polities' (North, 1994, p. 366). As discussed in Section Russian privatization, privatization is important but is not a panacea.

De Facto Property Rights

Just as the study of contract has benefited by going beyond a legal rules concept of contract to make prominent provision for private ordering (Llewellyn, 1931; Galanter, 1981), so too does the study of property rights benefit by going beyond *de jure* property rights to consider *de facto*. The recent application of *de facto* reasoning to explain economic development and reform in China (Qian and Weingast, 1996) is illustrative. The dilemma was this: *de jure* privatization was politically unacceptable, yet the state-owned enterprises in China had weak incentives and poor performance. What to do?

Decentralization, by creating local township and village enterprises in which local government remained prominent, turned out to be a productive move. As Qian and Weingast put it (1996, pp. 87–8):

> These local government-owned enterprises have remarkably different governance structures from (central) state-owned enterprises and thus face better positive and negative incentives…. Politicians in every political system tend to bail out inefficient firms or spend on wasteful consumption. In a federal system, however, the mobility of resources across regions raises the opportunity costs to local governments of bailing out inefficient firms or wasteful public expenditures.

Not only do local governments share the successes of local enterprises, but they incur the risk that investments will relocate if governments interfere. Federalism is thus an instrument for realizing quasi-privatization and the incentive benefits that accrue thereto. *De facto* property rights thus vary with the political environment, *de jure* property rights unchanged.

There are nevertheless limits to the risks that parties will incur in a regime where *de jure* credibility is lacking. I conjecture that China will experience serious incentive obstacles in its attempts to develop leading edge technologies unless it provides more secure *de jure* supports for intellectual property rights.

Limitations

Although property rights concepts and property rights reasoning have gone a long way to reshape our understanding of economic institutions and effect public policy reform, the early ambitions of the economics of property rights have been incompletely realized. Here, as elsewhere, there can be too much of a good thing.

Oversimplification

According to Coase, 'a private-enterprise system cannot function properly unless property rights are created in resources, and, when this is done, someone wishing to use a resource has to pay the owner to obtain it. Chaos disappears; and so does the government, except that a legal system to define property rights and to arbitrate disputes is, of course, necessary' (Coase, 1959, p. 12).

This property rights solution applied not merely in general but even to such apparently intractable problems as allocating electromagnetic spectrum to radio broadcasters: 'there can be little doubt that, left to themselves, the courts would have solved the problems of the radio industry in much the same way [by defining and enforcing property rights] as they had solved similar problems in other industries' (Coase, 1959, p. 30). Other interested parties were less sanguine. Chief Justice Taft, for example, resolutely 'dodged the radio question. I have refused to grant writs and have told the other justices that I hope to avoid passing on this subject as long as possible' (Coase, 1959, p. 30).

Indeed, what would the courts, left to themselves, have done? Taft was baffled by the radio question because property rights were ill-defined and the courts did not know how to define them. Awaiting a workable definition of property rights in electromagnetic spectrum – with the help of economists, lawyers, engineers, or managers – a property rights 'solution' remained elusive.

As Arthur DeVany put it, 'the real stumbling-block to a market system of spectrum allocation is the interference problem. The spectrum is a multidimensional space; signals propagate indefinitely in the frequency, spatial, and time domain' (1998, p. 169). The team of DeVany, Ross Eckhart, Dan O'Hara, Charles Meyers, and Richard Scott (three economists, a lawyer, and a physicist) turned their attention to this matter in 1969. Their best efforts to define property rights that would be 'exclusive, predictable, capable of subdivision, enforceable and flexible' (1998, p. 169)

notwithstanding, DeVany now concedes that their 1969 definition was merely a template and needed further refinements (1998, p. 169).

Indeed, going beyond the definition of property rights, there is a further problem: how to devise an auction? Although DeVany avers 'that our model has been used to privatize the spectrum in New Zealand, Australia, and Latin America' (1998, p. 168), Preston McAfee and John McMillan maintain that 'simultaneous single-round, sealed-bid auctions were used in New Zealand for spectrum licenses and in Australia for satellite-television licenses with disappointing results: low revenues and inefficient license allocations' (1996, p. 162). New developments in auction theory, which did not take shape until the 1980s and 1990s, were also needed. The simultaneously ascending auction was adopted by the FCC in the 1990s to auction spectrum for personal communications services: mobile telephones, two-way paging, portable fax machines, and wireless computer networks. Some judge that auction to be a 'success' (McAfee and McMillan, 1996, p. 163), but the actual sale of spectrum turned out to be 'more complicated that anything in auction theory' (McAfee and McMillan, 1996, p. 171).

Merely to conduct an auction ought not to be confused, moreover, with full privatization. In the US the FCC remains an active player in the management of spectrum. The upshot is that defining and enforcing property rights for some resources is a daunting exercise. Politics often continue, usually with the active participation of the key players. Realpolitik and partial privatization are evidently bedfellows.

Overreaching

The modern corporation

Might the salient features of the modern corporation be explained in property rights terms? Harold Demsetz took this position. Upon observing that there are 'significant economies of scale in the operation of large corporations … [and] that large requirements for equity capital can be satisfied more cheaply by acquiring the capital from many purchasers of equity shares' (1967, p. 358), Demsetz described the modern corporation in the following property rights terms: (1) efficiency is realized by delegating effective ownership to the management; (2) shareholders are essentially lenders of equity capital; and (3) limited liability relieves shareholders of the need to carefully examine corporate liabilities and the assets of other shareholders (as they would need to under partnership law).

This property rights interpretation of the modern corporation over-reaches in the first two respects. First, to suggest that diffuse ownership is responsible for delegation, whereupon the 'management group becomes the *de facto* owners' and that ownership is 'legally concentrated in management's hands' (Demsetz, 1967, p. 358) is incorrect. For one thing, delegation is observed in large corporations whether there are many or few owners of equity shares. For that matter, delegation is observed in every large organization – public, private, nonprofit. Well-known limits on the span of control (from organization theory) rather than the economics of property rights are responsible for that result. Additionally, of the three rights of ownership – the right to use, transform, and appropriate the income from an asset – the management is never delegated this last. Equity is the residual claimant.

Second, to describe shareholders as 'lenders' of equity capital is misleading. Debt and equity are not merely instruments of finance; they are also instruments of governance. If shareholders are merely lenders, why not just sell collateralized bonds? Or why not loans from the banks? It is not for naught that equity investors, who invest for the life of the firm and are residual claimants, are awarded control of the board of directors (Fama and Michael, 1983). The efficiency logic of debt and equity is usefully viewed as a governance issue (Williamson, 1988).

Russian privatization

The architects of privatizing Russian industry in the early 1990s appealed to the modern property rights theory of Sanford Grossman and Oliver Hart (1986) to urge mass and rapid privatization of all sectors of Russian industry (Boycko et al., 1995). The overarching theory was this: 'When property rights over a productive asset are clearly specified, and the person who decides how to employ his asset bears full costs and enjoys the full benefits of employment, he puts the asset to its most productive use' (Boycko et al., 1995, p. 19). Upon privatizing two-thirds of Russian industry in June 1994, Boycko et al. boldly declared that privatization had been brought to a 'triumphant completion' (1995, p. 8).

Subsequent events have been unkind. Although it may be unduly pessimistic to view Russian privatization as a failure (Braguinsky, 1999; Black et al., 1999), it is undisputed that Russian privatization has not been a triumphant success.

Errors of two kinds explain the premature pronouncement of success. First, the property rights theory of the firm suppresses problems of *ex post* governance by assuming costless renegotiation. That has analytical advantages, but it is also disconnected from reality. Had Boycko et al. inquired into contract implementation, they would have discovered that the efficacy of privatization varies, working much better in some industries than others (Williamson, 1976; Goldberg, 1976; Priest, 1993). Second, Boycko et al. never faced up to lapses in the laws and, even more, their efficacious enforcement in Russia. Failures to look beyond the immediate award of property rights and examine the needs for *ex post* governance support and effective legal recourse turned out to be fateful.

As with the allocation of electromagnetic spectrum and interpreting the modern corporation, so too with Russian privatization. Thus although property rights reasoning is pertinent to all, the economics of property rights is a self-limiting perspective. Awaiting a unified theory of economic organization, which is not in prospect, the application of several lenses – to include an examination of hazards that beset *ex post* governance and the condition of the institutional environment – has a lot to recommend it.

Conclusions

The economics of property rights is unarguably important. Were it that the Socialist Controversy in the 1930s was less preoccupied with technicalities (such as marginal cost pricing) but addressed itself more to an assessment of the

property rights differences between capitalism and socialism, the field of comparative economic systems would have developed very differently (Demsetz, 1998). But while property rights are crucial, they are not fully determinative of economic organization. To claim that efficiency will be realized upon defining and enforcing property rights simply overreaches. Not only can property rights be difficult to define, but best efforts at court enforcement can be costly. Also, property operates in combination with contract and organization.

Kenneth Arrow's remarks about externalities, market failures, and transaction costs help to restore perspectives: 'I contend that market failure is a more general category than externality [and, because] market failure is not absolute … , it is better to consider a [still] broader category, that of transaction costs, which in general impede and in particular cases completely block the formation of markets' (1969, p. 48). Transaction cost is an inclusive concept that applies to property, contract, and governance.

Editor's Note

Transaction costs (basically, the costs of entering a market) continue to figure prominently in economic thinking. Some recent exemplars of research in this area include: Argyres and Zenger, 2012; Williamson, 2010; Coggan et al., 2010; Tadelis and Williamson, 2012.

See also: Property Rights and Urban Planning.

Bibliography

Alchian, A.A., 1961. Some Economics of Property, RAND D-2316. RAND Corporation, Santa Monica, CA.

Alchian, A.A., December 1965. The basis of some recent advances in the theory of management of the film. Journal of Industrial Economics 14, 30–41.

Argyres, N.S., Zenger, T.R., 2012. Capabilities, transaction costs, and firm boundaries. Organization Science 23 (6), 1643–1657.

Arrow K., 1969. The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In: The Analysis and Evaluation of Public Expenditure: The PPB System. US Joint Economic Committee, 91st Congress, 1st Session. US Government Printing Office, vol. 1, Washington, DC, pp. 59–73.

Baume, W.J., 1959. Business Behavior, Value and Growth. Macmillan, New York.

Berle, A.A., Means Jr., G.O., 1932. The Modern Corporation and Private Property. Commerce Clearing House, New York.

Black, B, Kraakman, R, Tarassova, A 1999. Russian privatization and corporate governance: What went wrong? Unpublished manuscript. Stanford Law School.

Boycko, M., Shleifer, A., Vishny, R., 1995. Privatizing Russia. MIT Press, Cambridge, MA.

Braguinsky, S., June 1999. Enforcement of property rights during the Russian Transition: Problems and some approaches to a New Liberal solution. Journal of Legal Studies 28, 515–544.

Coase, R., October 1959. The federal communications commission. Journal of Law and Economics 2, 1–40.

Coase, R., October 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Coase, R., May 1964. The regulated industries: Discussion. American Economic Review 54, 194–197.

Coggan, A., Whitten, S.M., Bennett, J., 2010. Influences of transaction costs in environmental policy. Ecological Economics 69 (9), 1777–1784.

Demsetz, H., May 1967. Toward a theory of property rights. American Economic Review 57, 347–359.

Demsetz, H., April 1969. Information and efficiency: Another viewpoint. Journal of Law and Economics 12, 1–22.

Demsetz, H., 1998. Property Rights. In: Newman, P. (Ed.), In New Palgrave Dictionary of Economics and Law, Vol. 3. Stockton Press, New York, pp. 144–155.

DeVany, A., Eekert, R., Meyers, C., O'Hara, D., Scott, R., June 1969. A property system for the allocation of the electromagnetic spectrum. Stanford Law Review 21, 1499–1561.

DeVany, A., 1998. Property rights in electromagnetic spectrum. In: Newman, P. (Ed.), In New Palgrave Dictionary of Economics and Law, vol. 3. Stockton Press, New York, pp. 167–171.

Fama, E.F., Michael, C.J., June 1983. Separation of ownership and control. Journal of Law and Economics 26, 301–326.

Furubotn, E., Richter, R., 1991. The new institutional economics: An assessment. In: Furubotn, E.G., Richter, R. (Eds.), The New Institutional Economics. Texas A&M Press, College Station, TX, pp. 1–32.

Furubotn, E., Richter, R., 1997. Institutions and Economic Theory. University of Michigan Press, Ann Arbon, MI.

Galanter, M., 1981. Justice in many rooms: Courts, private ordering, and indigenous law. Journal of Legal Pluralism 19, 1–47.

Goldberg, V., 1976. Regulation and administered contracts. Bell Journal of Economics 7, 426–452.

Grossman, S.J., Hart, O.D., August 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. Journal of Political Economy 94, 691–719.

Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.

Hart, O., 1995. Firms, Contracts, and Financial Structure. Oxford University Press, New York.

Jensen, M., Meckling, W., October 1976. Theory of the firm: Managerial behavior, agency costs, and capital structure. Journal of Financial Economics 3, 305–360.

Koopmans, T., 1957. Three Essays on the State of Economic Science. McGraw-Hill, New York.

Lal, D., 1983. The Poverty of Development Economics. Institute of Economic Affairs, London.

Libecap, G.D., Wiggins, S.N., June 1985. The influence of private contractual failure on regulation: The case of oil field unitization. Journal of Political Economy 93, 690–714.

Llewellyn, K.N., 1931. What price contract? An essay in perspective. Yale Law Journal 40, 704–751.

Marris, R.L., 1964. The Economic Theory of Managerial Capitalism. Free Press of Glencoe, New York.

McAfee, R.P., McMillan, J., Winter 1996. Analyzing the airwaves auction. Journal of Economic Perspectives 10, 159–175.

Michelman, F., April 1967. Property, utility and fairness: Comments on the ethical foundations of 'Just compensation' law. Harvard Law Review 80, 1165–1257.

North, D.C., June 1994. Economic performance through time. American Economic Review 84, 359–368.

Priest, G.L., 1993. The origins of utility regulation and the 'Theories of Regulation' debate. Journal of Law and Economics 36, 289–323.

Qian, Y., Weingast, B., 1996. China's transition to markets: Market preserving federalism, Chinese style. Journal of Policy Reform 1, 149–185.

Stigler, G., October 1992. Law or Economics? Journal of Law and Economics 35, 455–468.

Tadelis, S., Williamson, O.E., 12 March 2012. Transaction Cost Economics. Available at SSRN: http://ssrn.com/abstract=2020176 or http://dx.doi.org/10.2139/ssrn.2020176.

Tollison, R., 1998. Rent seeking. In: Newman, P. (Ed.), In The New Palgrave Dictionary of Economics and the Law, vol. 3. Stockton Press, New York, pp. 315–322.

Turvey, R., March 1964. Optimization and suboptimization in fishery regulation. American Economic Review 54, 64–76.

Williamson, O.E., 1964. The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm. Prentice-Hall, Englewood Cliffs, NJ.

Williamson, O.E., 1975. Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.

Williamson, O.E., Spring 1976. Franchise bidding for natural monopolies—In general and with respect to CATV. Bell Journal of Economics 7, 73–104.

Williamson, O.E., 1985. The Economic Institutions of Capitalism. Free Press, New York.

Williamson, O.E., July 1988. Corporate finance and corporate governance. Journal of Finance 43, 567–591.

Williamson, Oliver E., 2010. Transaction cost economics: the natural progression. Journal of Retailing 86 (3), 215–226.

Transcranial Direct Current Stimulation and Language

Barry Gordon, James Livengood, Joseph J van Steenburgh, Rasika Jayatillake, and Tracy D Vannorsdall, The Johns Hopkins University, Baltimore, MD, USA
David J Schretlen, The Johns Hopkins University, Baltimore, MD, USA; and The Johns Hopkins University School of Medicine, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Transcranial direct current stimulation (tDCS) may be able to affect language functions, over both shorter (minutes) and longer (hours to perhaps months) intervals. Individual differences in psychological capacities, neurophysiology, and neuroanatomy make within-subjects, repeated-measures designs ideal for tDCS studies using present-day methods. True 'sham' stimulation may not be possible at higher levels of current. Designs that take these factors into account, coupled with improved methods of focusing current and orchestrating when and where it is applied, will likely strengthen the usefulness of tDCS for investigating language functions and improving them.

Introduction

There have been several recent excellent reviews of the use of transcranial direct current stimulation (tDCS) to alter various aspects of language (Chrysikou and Hamilton, 2011; Floel, 2012; Fridriksson et al., 2012; Monti et al., 2012; Schlaug et al., 2011). Rather than simply update these reviews, this article will provide a broader perspective on current and potential applications of tDCS to language functions.

Ideally, the rational use of tDCS would be guided by a thorough knowledge of its effects on neural tissue and by knowledge of how language and related processes are generated by the brain. But we are far from having such knowledge. The best we can do at present is make educated guesses and advance by trial and error. This article is an attempt to outline what knowledge we do have and how we might best proceed to test for the effects of tDCS on language. In this article, a reference to 'language' will generally mean not just language itself, but also speech and all the processes associated with language comprehension, usage, and production.

Neural Effects of tDCS

To a first approximation, tDCS as it usually is employed (with 1–2 mA stimulation levels) generates electric fields that are thought to range from 0.2 to 1.5 volts per meter ($V\,m^{-1}$) in regions of the cortex and subcortical structures (Parazzini et al., 2011). Initial studies in animals and humans suggested a straightforward consequence: Tissue underneath the anodal electrode was stimulated, whereas tissue underneath the cathodal electrode was inhibited (Bindman et al., 1964; Creutzfeldt et al., 1962; Matsunaga et al., 2004; Nitsche and Paulus, 2000, 2001; Purpura and McMurtry, 1965). However, with increasing knowledge of how such relatively weak electric fields might affect neural tissue has come an appreciation that the "anodal stimulatory/cathodal inhibitory" rule likely is a somewhat misleading simplification. This section, on the possible neural bases for tDCS effects, is organized primarily by spatial scale.

Molecular and Membrane Effects

Weak electric fields are thought to exert their initial effects at the molecular level by displacing intracellular ions and thereby altering neuronal resting membrane potentials (Ruffini et al., 2012). Typical currents delivered by tDCS at the scalp (1–2 mA) are severely attenuated by the time they reach the brain (Miranda et al., 2009, 2013), and the low field intensities they produce probably cannot directly stimulate cerebral structures (Bikson et al., 2004; Roth, 2009), although see (Feurra et al., 2011). Indeed, each volt per meter of field intensity can produce on average only around 0.3 mV of membrane polarization (Bikson et al., 2004; Roth, 2009). This is much less than the tens of millivolts required to produce action potentials. Thus, tDCS generally is regarded as a modulatory technique that requires ongoing neural activity to produce effects. However, cortical morphology may cause current to be distributed unevenly beneath the electrodes, with cortical surface density in some areas 10- to 100-fold higher than the average density beneath the electrode (R. Sadleir, personal communication, April 2013). Thus, we cannot rule out the possibility that tDCS might trigger action potentials in some instances.

The immediate 'online' effects (during stimulation) are thought to depend on transient changes in resting membrane potential. This is borne out by studies in which pharmacological blockade of ion channels, but not *N*-Methyl-D-aspartate (NMDA) or γ-Aminobutyric acid (GABA) receptors, abolished changes in cortical excitability that coincide with stimulation (Nitsche et al., 2003, 2004a). Aftereffects (or off-line effects) of tDCS can also persist for minutes to hours (Bindman et al., 1964; Nitsche and Paulus, 2001), and these long-term changes likely also involve molecular-level mechanisms (Stagg and Nitsche, 2011). These may include dependence on NMDA receptors (Liebetanz et al., 2002; Nitsche et al., 2004b); protein synthesis (Gartside, 1968a,b); changes in intra- and extracellular calcium levels (Islam et al., 1995a); early gene expression (Ranieri et al., 2012); alterations of dendritic spines (Jiang et al., 2012); and the modulation of intracellular messengers implicated in calcium signaling and neuroplasticity, such as protein kinase C (Islam et al., 1995b), cyclic adenosine

monophosphate (cAMP) (Hattori et al., 1990; Moriwaki et al., 1994), myoinositol (Rango et al., 2008), and brain-derived neurotrophic factor (BDNF) (Fritsch et al., 2010). A growing body of evidence, therefore, supports the notion that Long Term Potentiation (LTP)-like mechanisms, long considered important for learning and memory (Malenka and Bear, 2004; Owen and Brenner, 2012), may be responsible for effects that begin a few minutes after the onset of tDCS and persist after stimulation ends.

Despite strong evidence that the aftereffects of tDCS involve synaptic mechanisms, alternatives have been proposed (Ardolino et al., 2005). These include the nonsynaptic modulation of axonal excitability, intra/extracellular pH changes induced by water electrolysis, and even the direct alteration of transmembrane protein conformation and migration within the cell membrane, analogous to gel electrophoresis. This mechanism is of particular interest in light of evidence that during synaptic transmission, ion fluxes through postsynaptic receptors produce electric fields that may clear neurotransmitters from or draw them toward the synaptic cleft via electrodiffusion (Savtchenko et al., 2004; Sylantyev et al., 2008, 2013). They may even cluster postsynaptic α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors across from vesicular release sites (Savtchenko et al., 2000).

Cellular Effects

At the cellular level, changes in membrane potential are usually not distributed evenly across a single neuron. Anodal fields oriented radially to the cortical surface, and therefore parallel to the axodendritic axes of pyramidal cells, tend to depolarize somas and will actually hyperpolarize apical dendrites (Radman et al., 2009). The reverse is true for cathodal stimulation. However, the fields produced by tDCS also tend to have a significant tangential component (Miranda et al., 2013) that does not polarize somas but, instead, hyperpolarizes or depolarizes fibers running parallel to the cortical surface, which then facilitates or depresses synaptic efficacy (Rahman et al., 2013).

This raises the question of why differential polarization of neuronal compartments does not have opposing effects that cancel out net modulation of the cell. For anodal stimulation at least, the answer may be due to the possibility that, in some instances, hyperpolarized dendrites can depolarize neurons. Hyperpolarization-activated cyclic nucleotide-gated (HCN) channels are richly expressed in apical dendrites and counteract membrane hyperpolarization by admitting a sodium and potassium current (Kase and Imoto, 2012), which may explain why differential polarization within neurons enhances, rather than interferes with, neural excitability under tDCS. Moreover, these channels are believed to help induce neuronal oscillations and to facilitate firing-pattern transitions from single spiking to bursting (Benarroch, 2013) when, for instance, neuronal action potentials backpropagate into dendrites. This may be an important way for seemingly negligible changes in membrane potential to amplify and synchronize ongoing neural activity (Tse, 2013: pp. 89–103).

While effects on pyramidal cells have received a great deal of attention, tDCS likely affects other cell types. Recent calculations using cable theory suggest that tDCS may also act on glial cells (Ruohonen and Karhu, 2012), although this has not been confirmed experimentally (Roth, 2012). Moreover, weak electric fields are known to enhance cell motility (Heo et al., 2012) and neurite outgrowth (Koppes et al., 2011). Field orientation may even alter neuronal precursor migration and differentiation during development (Ariza et al., 2010; Yao et al., 2009).

Both excitatory and inhibitory interneurons may also be affected by tDCS (Nitsche et al., 2005). As with pyramidal cells, their stimulation thresholds and polarizability can vary as a function of morphology (Radman et al., 2009) and cortical depth (Purpura and McMurtry, 1965). Additionally, pharmacological and magnetic resonance spectroscopy (MRS) studies suggest that tDCS alters cortical and even subcortical neurotransmitter levels of dopamine (Kuo et al., 2008; Nitsche et al., 2006; Monte-Silva et al., 2009; Nitsche et al., 2009; Tanaka et al., 2013), acetylcholine (Kuo et al., 2007), and GABA (Stagg et al., 2009). For instance, tDCS effects on cortical GABA levels may enhance cortical excitability via pyramidal cell disinhibition. To that effect, a recent neural modeling study showed that decreasing extrasynaptic GABA levels induced continuous subthreshold membrane depolarization that did not change spontaneous activity but did enhance pyramidal responses to subsequent inputs (Hoshino, 2011). Since different classes of GABAergic interneurons target specific cell compartments (e.g., dendrites, somas) (Royer et al., 2012), tDCS may alter how they regulate distinct aspects of neural activity (e.g., pyramidal bursting, phase synchronization).

Network Effects

Evidence suggests that weak electric fields produce small changes in the spike timing and firing patterns of individual neurons, which may then be amplified by the dynamic activity and mass effects of the neural networks in which they are embedded (Reato et al., 2010). Although it is well established that neurons have subthreshold sensitivity to very low voltage gradients of around 1 V m^{-1}, networks might be even more sensitive to field modulation than single neurons (Francis et al., 2003). This likely results from recurrent excitatory loops within networks that contain locally coupled inhibitory activity (Reato et al., 2010; Weiss and Faber, 2010). In fact, some networks, by virtue of their uncommonly uniform cell orientation and tight cell packing, can have especially low stimulation thresholds. Hippocampal region CA1, for instance, responded to fields as small as 0.140 μV mm^{-1}, an order of magnitude lower than what was previously thought possible. The adjacent but somewhat sparser and less cytoarchitecturally homogeneous CA3 region did not show this enhanced susceptibility (Francis et al., 2003). Regional discrepancies of this kind are likely to figure prominently in the myriad interindividual differences in tDCS responsiveness that we consider further below.

Many of these effects occur endogenously through ephaptic interactions, by which spatially proximate groups of neurons produce extracellular potentials that then feedback and synchronize their own activity (Anastassiou et al., 2011). Recent work shows that (at least in ferret visual cortex) weak DC fields of similar strength to *in vivo* ephaptic potentials can enhance, while both sinusoidal and naturalistic waveforms can entrain, the activity of cortical networks (Frohlich and McCormick, 2010). By showing how waveforms that closely resemble those produced *in vivo* can modulate network activity, this work raises

the possibility that tDCS is engaging a very important component of neural processing (Ruffini et al., 2012; Weiss and Faber, 2010).

These findings seem to lend credence to reports that tDCS and alternating current (tACS) can modulate and perhaps entrain cortical oscillations in humans (Antal et al., 2004; Garcia et al., 2011; Jacobson et al., 2011; Kanai et al., 2010; Siebner and Ziemann, 2010; Zaehle et al., 2010), although see Brignani et al. (2013), especially when the stimulation matches the endogenous frequency (Merlet et al., 2013; Neuling et al., 2013). Cortical oscillations may have an important relation to the substrates of speech and language processing (Giraud and Poeppel, 2012), and so these emerging strategies may prove extremely influential.

Although not necessarily as a result of modulating oscillatory activity (Stagg and Nitsche, 2011), tDCS can affect neural processing within both local and more distributed, higher-level networks. Many studies have shown perceptual, behavioral, and excitability modulations in motor (Bergmann et al., 2009; Cambiaghi et al., 2010; Di Lazzaro et al., 2012; Galea and Celnik, 2009; Kwon et al., 2008; Monte-Silva et al., 2012; Polania et al., 2012), visual (Antal and Paulus, 2008; Chi et al., 2010), auditory (Ladeira et al., 2011; Zaehle et al., 2011), and somatosensory cortices (Dieckhofer et al., 2006; Marquez-Ruiz et al., 2011, 2012; Matsunaga et al., 2004; Ragert et al., 2008; Rogalewski et al., 2004). Various cross-modal interactions have also been affected following, for instance, stimulation in areas peripheral to those being directly engaged by a given task (Bachmann et al., 2010; Bolognini et al., 2011; Lapenta et al., 2012). This suggests that the effects of tDCS are less localized than might be hoped, despite recent attempts to optimize focality with various high-definition models and montages (Datta et al., 2013; Edwards et al., 2013; Huang et al., 2012; for recent reviews, see Bai et al., 2013; Bikson and Datta, 2012; Ruffini et al., 2012).

Indeed, stimulation of what was assumed to be relatively discrete cortical areas can induce changes in cortical blood flow across distant brain regions (Venkatakrishnan and Sandrini, 2012; Zheng et al., 2011). However, the widely distributed nature of many higher cognitive functions, such as language (Friederici and Gierhan, 2013; Price, 2012), may render this lack of focality more of a strength than a weakness.

Functional and Neuroanatomic Organization of Language Processes

The functional and neuroanatomic architectures of language introduce their own complexities into the possible effects of tDCS.

Functional Organization

Network Organization

The notion that language (in common with most if not all higher functions) is generated through a set of more-or-less independent subprocesses (or stages) is now both an empirical demonstration and, in some respects, an article of faith. In the current context, we must consider how such a complex network might be affected by an external perturbation such as tDCS. One complexity to bear in mind is that neither the subfunctions nor their products are always easily defined. Phonologic perception, orthographic identification, lexical semantics, phonologic production, and articulation have been relatively well identified and teased apart. But pragmatic intent, prelinguistic language (Jackendoff, 2002), and many processes encompassed by the term syntax may be less modular, or have more pervasive representations. Moreover, we should expect a high degree of rapid interconnectivity among various stages, not just feed-forwards, but also feed-backwards (Gordon, 1997).

In addition, it is generally appreciated that language as a whole is a subsystem embedded within, and responsive to, other systems of the mind. There is independent evidence, based mainly on function Magnetic Resonance Imaging (fMRI), for a "default network," an arousal/attentional network, and a cognitive control network (Bressler and Menon, 2010). All three are believed to underlie cognitive, and especially executive, functions considered important for language (Acheson and MacDonald, 2009; Gruber and Goschke, 2004; Ye and Zhou, 2009). These include working memory (Fuster and Bressler, 2012) and selective attention (Miller and Buschman, 2013) as well as potential interactions between automatic and controlled processing, for instance, during lexical–semantic retrieval (Hills et al., 2012; Hurks et al., 2004; Vannorsdall et al., 2012a,b) or language switching in bilinguals (Hervais-Adelman et al., 2011; Magezi et al., 2012). It is also possible that network hubs like the anterior insula and larger salience network may be critical for effectively switching between resting and executive networks (Menon and Uddin, 2010). Recent evidence suggests tDCS can modulate the functional connectivity of these networks (Keeser et al., 2011; Meinzer et al., 2012; Park et al., 2013; Pena-Gomez et al., 2012; Pereira et al., 2012; Polania et al., 2011; Sehm et al., 2012), which makes them attractive targets for study. And, regardless of whether language systems can be considered logically independent from these networks, they often have a close anatomic proximity if not dense interweaving and, perhaps even, physiologic interdependencies.

As is true of any complex system with multiple responsibilities, we might also expect to find trade-offs and conflicts. Strengthening a subfunction might paradoxically worsen overall network performance; inhibiting a subfunction might improve it. Moreover, the impairment or improvement at the network level may occur in subtle but important aspects of language processing. Anodal stimulation over right anterior temporal lobe and cathodal stimulation over the left anterior temporal lobe, for instance, have been shown to enhance creative problem solving (Chi and Snyder, 2012), insight (Chi and Snyder, 2011; Metuki et al., 2012), and mediated semantic priming (Weltman and Lavidor, 2013). Presumed interference with cognitive control processes via cathodal stimulation of the left dorsolateral prefrontal cortex (DLPFC) was found to facilitate the adoption of unconventional object uses (Chrysikou et al., 2013), enhance convergent thinking (Cerruti and Schlaug, 2009), and alter the cognitive strategies used to learn novel grammars (Friederici et al., 2013). Inhibition can also be used to enhance performance when stimulation of a given region is desired (e.g., DLPFC to facilitate working memory), but direct stimulation may produce complex interactions, for instance, due to the region's complex cell morphology or dynamic, reciprocal network

connectivity (Arnsten et al., 2012). For instance, Pope and Miall (2012) showed that cerebellar cathodal stimulation improved working memory performance, presumably by disinhibiting the prefrontal cortex (PFC). Via a potentially similar disinhibitory technique, cathodal stimulation of right Wernicke's area was used to enhance language comprehension in subacute stroke patients (You et al., 2011). Other studies have shown that cathodal stimulation can suppress competing channels and improve the signal-to-noise ratio during dichotic listening paradigms (Alexander et al., 2012) and attentional tasks (Weiss and Lavidor, 2012). Thus, the disinhibition of a cross-hemispheric, homologous structure or via two distant structures may offer an innovative way to explore the efficacy of indirect approaches to stimulation.

Within-Stage Processing

The situation is even more complex when we consider how tDCS might affect the operations within a processing stage. While no one can pretend to know the exact operations performed in any stage involved in language, there are empirical and theoretical hints as to the approximate kind of behaviors we can expect (Gordon, 1997).

At a fairly general level, it seems reasonable to distinguish between the dynamic patterns of electrochemical activity in a neural network, and the underlying neuroanatomic connections and synaptic weights that generate and support those patterns. It is tempting to imagine that psychological entities such as abstract letter identities, orthographic word forms, semantic features, and phonologic bundles are encoded as more-or-less discrete patterns of neuronal activity. We can also speculate more generally about the ability of a particular neural network to produce these patterns. One way of better formalizing and specifying these notions is to represent these psychological states as attractor states of neural activity (in some ways a more neurophysiologically grounded interpretation of earlier notions of nodes or resonant states, such as in Ratcliff (1978); see Gordon (1997) for a discussion). A specific word form or semantic feature would then be a specific attractor state that the network may achieve with appropriate inputs and internal connectivity. These attractor states, and the ghostly meaning that they convey, are not static. It is reasonable to imagine the concepts or entities that they represent gradually coalescing out of neural activity, and then dissipating when the stimulus is gone and intrinsic activity no longer supports the particular pattern (or gets replaced by a new one). Externally, this coalescence and dissipation would seem to correspond to the growth and decay of mental entities that has been delineated by other means (Hart et al., 1998).

From abundant psychological evidence, we can put some temporal boundaries on processing at these individual code levels. Phonemic and phonologic entities may be able to stabilize within tens of milliseconds. Orthographic and semantic ones probably take more on the order of hundreds of milliseconds (cf Hart et al., 1998).

In this conception, the neural network responsible for a particular code has many latent representations of specific, learned elements or features of that particular code. Active patterns of activity that are essentially orthogonal to each other might be expected to coexist to some extent within a single neural network. There are many potential reasons for real or virtual mutual inhibition and constraint (e.g., Reggia et al., 1992). Evidence for mutual competition between within-level representations might even serve as an important clue to the nature of the internal mechanics of these more elemental levels of processing.

To the extent this general understanding captures and explains the dynamic characteristics of the elements that make up language functions, it suggests ways in which the effects of tDCS and other manipulations might be observed. tDCS might alter such parameters as the depth (or height) of attractor states, the rate at which patterns evolve toward their end points, or the noisiness or meandering of the pattern-seeking process. It might also affect interactions among patterns as they evolve.

Although the myriad parameters at the single-neuron and network level seem daunting, it is also possible that they may blend into only a relatively few variables of global relevance. Profound differences in behavioral functions might then be affected by relatively simple alterations, perhaps analogous to the way differences between normal and aphasic naming have been attributed to quantitative rather than qualitative differences (Dell et al., 1997).

Neuroanatomic Organization

Some processing stages seem to correspond to relatively distinct neuroanatomic regions; for others, the correspondence is far from clear (Price, 2012). In the clearest functional–neuroanatomic correlations, such as lesions that affect orthographic word codes (the visual word form), the cortical areas involved can be as small as a few square centimeters (Dehaene and Cohen, 2011). But even these dependencies may not be as absolute as classical anatomy assumed. There may be latent representations of functions in the tissue adjacent to the most critical areas, as evidenced by the relatively rapid restitution of functions that can occur after focal lesions (cf Gordon et al., 1999).

What complicates the functional–neuroanatomic correlation of language is evidence that many if not most human beings have a variable degree of duplication of language and other functions across the hemispheres (Hugdahl and Westerhausen, 2010). And this brings us to the even broader question of individual differences.

Individual Differences

Individual Differences in Language Capabilities

There are vast differences in the language capabilities of different individuals, even of putatively 'normal' or 'neurotypical' people (Chandrasekaran et al., 2010; Shao et al., 2012; Sontam et al., 2009; Szenkovits et al., 2012; Yap et al., 2012). Interindividual differences can be two- to threefold or more in magnitude. In addition, many psychological functions are notoriously variable over time, even within the same individual (Dykiert et al., 2012; Evans and Demuth, 2012; Hayiou-Thomas et al., 2012; Heyanka et al., 2013; Niemela et al., 2013; Ode et al., 2011; Pnevmatikos and Trikkaliotis, 2013; Small et al., 1995; Vandermorris et al., 2013). Furthermore, the level of particular abilities within a single individual is not necessarily predictive of the level of other abilities (Schretlen et al., 2008; Selnes and Gordon,

1999). The basis or bases of inter- and intra-individual variability could offer important clues to the origin of these functions. However, they are also a major experimental barrier. As a result, many language functions that are fine-grained enough to be interesting can only be studied with adequate precision using within-subjects designs.

It is also likely that there are differences in the routes or mechanisms through which different individuals process language information (Caplan, 1988) or, probably for some functions, even in the same individual at different times. Given that interindividual variability in psycholinguistic processing mechanisms has been found in college undergraduates (Chiarello et al., 2012; Prat and Just, 2011; Welcome et al., 2011), one can reasonably expect that the broader population is likely to show greater differences (cf (Henrich et al., 2010; Gutchess et al., 2010)), especially persons with developmental or acquired disorders (Abbeduto et al., 2007; Bishop and McArthur, 2005; Borella et al., 2011; Budden, 2012; Elsabbagh et al., 2011; Gooch et al., 2012; Haley and Martin, 2011; Marinus and de Jong, 2010; Pellicano, 2010; Staiger et al., 2012; Venker et al., 2013).

Individual Differences in Neuroanatomy and Neurophysiology

We are only just beginning to appreciate that individuals can show remarkable heterogeneity of even basic neurophysiologic parameters, such as NMDA receptor allotypes (Sanz-Clemente et al., 2013) and dopamine-related genes (Wong et al., 2012), which may well be important in tDCS effects. What has long been appreciated are individual differences in regional anatomy as defined by cytoarchitectonics, with surface area differences of up to threefold, in regions including visual cortex (Amunts et al., 1999; Dougherty et al., 2003; Van Essen et al., 1984; Van Essen et al., 2001), Broca's area (Amunts et al., 1999; Keller et al., 2009; Sherwood et al., 2003), the inferior parietal (Bogolepova and Malofeeva, 2004; Caspers et al., 2006) and the prefrontal cortices (Fornito et al., 2008), and even subcortical limbic areas (Amunts et al., 2005; Bogolepova and Malofeeva, 2008). The basis and significance of most of these differences are still unknown, although some have been correlated with individual abilities (see Kanai and Rees (2011) for a review), and some may be a product of practice, not inheritance (Li et al., 2010).

Furthermore, there are great interindividual differences in the pattern of cortical folding with respect to the skull (Derrfuss et al., 2009; Keller et al., 2007; Mangin et al., 2004). Since tDCS must penetrate scalp, skull, and CSF to affect brain tissue, the variability across individuals in the amount of hair, the thickness of the scalp and subcutaneous tissues, the thickness of the skull, the presence of sutures and penetrating veins, and the amount of CSF are also relevant. Even montages that deliver relatively focal current are at times unable to produce the same effect across individuals without more than doubling the electric field strength (Edwards et al., 2013). Moreover, since the current in tDCS is volume-conducted and can be shunted by other structures, any given individual presents a very complicated situation in terms of just where current applied to the scalp will wind up in the brain, how much current will reach any one location, and how effective it will be upon arrival (Datta et al., 2009; Sadleir et al., 2012).

How Reliable Are Reported tDCS Effects?

One reason we have not tried to draw broad conclusions in this article about the effects of tDCS on language functions is our concern about the reliability of results reported in the literature. This is not because of evidence specific to tDCS itself. To our knowledge, there are no published efforts to directly replicate any language-related tDCS studies, although aspects of some results have failed to replicate (Huey et al., 2007; Koenigs et al., 2009). Rather, there simply is no reason to believe that tDCS studies of language will be more reliable and reproducible than those in other areas of psychological and cognitive neuroscience (Button et al., 2013; Uttal, 2013; Yong, 2012; Makel et al., 2012), or of science in general (Ioannidis, 2012; Prinz et al., 2011).

Considerations for Future tDCS Studies

This article leads us to suggest the following prescriptions and cautions for future tDCS studies of language:

1. While studies of tDCS have appropriately focused on possible modulatory effects, what is known of tDCS neurophysiology suggests that a search for possible effects on item representations, interitem competition, and network-level parameters and dynamics may also be fruitful.
2. When designing studies using tDCS to alter behavior, investigators must always consider the influence of homeostatic effects on neuroplasticity. There has been little exploration of state-dependent effects when using tDCS to alter language function, but tDCS effects on the motor cortex might provide guidance. There, prestimulation neural network activity powerfully modulates the strength and direction of electrophysiological responsiveness to stimulation (Fricke et al., 2011; Monte-Silva et al., 2010). For instance, neural excitability that has been altered by noninvasive brain stimulation, or even by mental states or behaviors, may determine subsequent responsiveness to tDCS (Ridding and Ziemann, 2010; Siebner et al., 2004). Cerebral tissue has a number of metaplastic mechanisms (Hulme et al., 2013; Stagg and Nitsche, 2011). In general, neural tissue is less excitable following excitatory stimulation, and less likely to be depressed after being inhibited. Such mechanisms demonstrate the importance of controlling prestimulation activity, and they suggest that the timing of stimulation will be crucial for maximizing treatment effects. For example, somatosensory paired association was enhanced following anodal tDCS but inhibited by concurrent stimulation (Nitsche et al., 2007). When anodal stimulation of the DLPFC precedes working memory tasks, sometimes only reaction times improve (Mulquiney et al., 2011; Teo et al., 2011), whereas accuracy improves with concurrent anodal stimulation of the same area (Ohn et al., 2008). Moreover, the type of current waveform (e.g., direct, randomized, or regularly alternating current) may also determine whether stimulation should

precede or coincide with a task (Pirulli et al., 2013). Such mechanisms may also determine the appropriate duration of stimulation. Longer stimulation periods do not necessarily add to the effects; with longer periods, effects may fluctuate (Kwon et al., 2011). Attending to and possibly exploiting homeostatic mechanisms will be instrumental in the design of effective interventions that use repeated stimulation (Fricke et al., 2011; Ohn et al., 2008; Reis et al., 2009).

Through similar mechanisms, the behavior in which tDCS recipients engage during stimulation can also moderate response. Even a single application of a specific waveform for a given duration can dramatically alter how neural tissue responds to other inputs. For instance, low-frequency stimulation of mouse motor cortex, which typically produces LTD, instead produced LTP when done in conjunction with anodal tDCS (Fritsch et al., 2010). In human motor cortex, short-lasting voluntary muscle contraction (which is not neuromodulatory by itself) in conjunction with tDCS applied at M1 reduced or reversed the typical neuromodulatory effects of both anodal and cathodal tDCS (Thirugnanasambandam et al., 2011).

Throughout a task, however, the relevant processing and neuroplastic parameters may not be uniform. Consequently, the effects of tDCS might vary over different phases of learning. Interestingly, concurrent cathodal stimulation of V1 during a visual discrimination task was recently found to be crucial for overnight consolidation (Peters et al., 2013). In a more cognitive domain, improved planning ability as measured by a Tower of London task required cathodal stimulation of the DLPFC during acquisition and early consolidation that was also followed in later sessions by anodal stimulation. These "training-phase-specific effects" (Dockery et al., 2009) were still present a year later. Since task-specific factors likely determine the time course of homeostatic plasticity, efforts to classify and standardize them may lend invaluable insights into how and when tDCS can be optimally deployed to produce robust and lasting effects.

3. Current application through focused arrays, possibly informed by modeling of anticipated current flows (Bikson et al., 2012), or even by direct *in vivo* measurements of current flow via MRI (R. Sadlier, personal communication, April 2013), might increase the anatomic specificity of tDCS.
4. There is growing concern about blinding in tDCS experiments. Subjects given 2 mA anodal stimulation were able to identify 'active' stimulation 72% of the time and 'sham' 56% of the time during their first session. After a second session, they identified both active and sham correctly more than 88% of the time (O'Connell et al., 2012). A large group of subjects reported more tingling, itching, burning, pain, and difficulty concentrating after 1.5 mA of active stimulation compared to sham, and they reported that tingling, itching, and burning persisted longer after active than sham stimulation (Kessler et al., 2012). Although blinding by sham procedures has been considered effective at 1 mA (Gandiga et al., 2006), recent reports suggest that subjects experience more discomfort during active than sham stimulation even at 1.0 mA (Ambrus et al., 2012). In the same study, investigators (i.e., experienced tDCS recipients) were able to reliably distinguish 1.0 mA active from sham stimulation. Thus, tDCS researchers will have to either develop more effective sham stimulation procedures, or circumvent the problem by altering cutaneous perceptions. Topical anesthetics have been shown to reduce cutaneous sensations associated with tDCS at 2 mA (McFadden et al., 2010) and at high current densities (Bikson and Datta, 2012; Borckardt et al., 2012). Alternatively, sensations associated with sham stimulation could be intensified by increasing the salinity of the solution in which the sponges are soaked (Dundas et al., 2007; Minhas et al., 2011). Of course, reporting side effects of stimulation might be less important than whether tDCS recipients can distinguish between active and sham stimulation. In this respect, it is as yet unclear whether conscious or unconscious awareness of the difference between active and sham stimulation will alter the impact of tDCS in any appreciable way.
5. Experimenters have not been blinded in most studies reported to date. With the introduction of preprogrammable or remotely controllable stimulators, experimenter blinding will become a feasible standard, as long as precautions are taken to blind the experimenter to redness at the electrode site that is more apparent after active stimulation than sham (Ambrus et al., 2010; Boggio et al., 2008). Acetylsalicylic acid or skin cream can be applied immediately after taking off the electrodes to obscure this difference (Palm et al., 2013), or such masking agents could be applied by a third party, or the subject could wear a head covering.
6. Statistical designs used for tDCS studies will have to address variability within and across subjects. Traditionally, randomized block and matched-pair designs have attempted to minimize between-subject variability by blocking or matching subjects based on what are thought to be the most influential sources of variance (e.g., age, gender, education level). However, accounting for individual variability is more difficult in tDCS experiments, because individual differences in tDCS response may be larger than the variability of within-subjects measurements. So tDCS studies are often best suited to within-subjects designs, such as cross-over, factorial, and Latin square designs. Such designs also require fewer subjects than parallel designs (Chow et al., 2008; Hills and Armitage, 1979; Jones and Kenward, 2003; Senn, 1999). Cross-over designs have, in fact, been the most frequent form used in tDCS studies to date.

In cross-over designs, each subject receives the same treatments over a series of sessions, but the treatment sequence varies across subjects. Cross-over designs introduce a problem in making direct treatment comparisons because of practice effects, session effects, and carry-over effects. Counterbalancing stimulation sequences and allowing sufficient wash-out periods between sessions can minimize possible carry-over effects. Some investigators have tried to minimize practice effects by giving subjects practice before the critical experimental tasks (Ohn et al., 2008). It may be important to have participants practice tasks well before stimulation in order to avoid homeostatic effects. Experimenters have also used different but comparable stimuli for different sessions (Cattaneo et al., 2011).

There is little research about the duration of cognitive effects to inform decision-making regarding wash-out periods. The duration of alterations in motor-evoked potentials following tDCS of the motor cortex provide coarse guidelines, with effects of 20 min of anodal stimulation at 2.0 mA having

lasted for at least 120 min (Alonzo et al., 2012). Similar findings have been discovered when visual-evoked potentials were measured after stimulation of the occipital cortex (Accornero et al., 2007; Antal and Paulus, 2008), although the duration of effects was somewhat different. Wash-out periods for such observable effects seem to be the minimum basic requirements to minimize carry-over. Perhaps a better guideline would be at least 24 h after a single session of stimulation. But so little is known about the persistence of effects of repeated stimulation that it would be premature to speculate on what intervals will prove to be appropriate wash-out periods.

In short, despite taking measures to minimize period, carry-over, and possible interactions, these effects need to be considered and reported in statistical analyses. There are several choices of statistical methods that can be used in the analysis. For the simple two-treatment, two-period, cross-over trial, Hills and Armitage (1979) proposed a series of two sample *t*-tests to test the significance of direct treatment difference and period or session differences. The disadvantage of this method is that the previous effects can only be tested if the carry-over effects of the two treatments are the same.

The other choices of analysis methods include repeated measures Analysis of Variance (ANOVA) and mixed models. Jones and Kenward (2003) give a very basic statistical model for cross-over trials. In this model, subject effects can be considered as either fixed or random. The majority of reported studies have treated subjects as fixed effects or carried out analyses using two-way repeated measures Analysis of Variance (ANOVA). However, there are reasons to consider subjects as random effects (drawn randomly from a population). Furthermore, in circumstances where the subject totals contain relevant information, then these can only be recovered if the subject effects are treated as random (Jones and Kenward, 2003). In this case, mixed models should be used; See (Putt and Chinchilli, 1999) and (Brown and Prescott, 2006).

The limiting assumptions of statistical analyses are also critical for analyzing tDCS studies (as they are in general). *T*-tests assume normal distributions. The assumption of normality of data, as is assumed in *t*-tests, is usually overlooked or not reported in findings. Mixed models and repeated measures ANOVA are robust to small deviations from normality. But when statistical examination of model residuals (e.g., the Shapiro–Wilk test and visual inspection of q–q plots) reveals significant deviations from normality, then nonparametric methods must be used.

7. We join in the recent observations that many reports in the literature have been inadequately documented (Raising Standards, 2013). In fact, one finding of this article is that it was often difficult to understand exactly what the investigators did and what they actually found. In reporting tDCS experiments, it will be critical to include all relevant or possibly relevant details of the experimental situation. Description of the stimulation parameters must include the current strength and waveform, current density, total dose of current, total number of active sessions, 10–20 electrode positioning, electrode treatments including solution salinity and topical treatments, and a complete description of sham stimulation including waveform and methods to verify blinding of both subjects and assessors. Description of subjects' activities before and during stimulation is essential, as is specification as to whether data were collected during (online) or after (off-line) stimulation. Including frequently omitted information, such as the number and nature of dropouts from the study, is also important.

Finally, when reporting findings, it will be important to include effect sizes, in terms such as Cohen's d or partial eta squared (Cohen, 1988). These data will give readers a better understanding of the effectiveness of tDCS as an intervention, and they allow for possible meta-analyses. In studies with small samples (e.g., 10–20 participants), presenting profile plots of the participants or corresponding mean plots would also provide readers with a better understanding of the responses to tDCS at subject level. In larger studies, attempts to characterize responders can inform future investigations.

In this regard, we also urge experimenters to submit negative results for consideration for publication, and we urge editors and reviewers to weigh negative results as much as they do putatively positive ones.

Looking forward, increasing the range, rigor, and integration of both experimental and computational approaches to tDCS can be expected to yield rich dividends, as efforts to apply it to the understanding, enhancement, and restoration of language and cognition mature.

Acknowledgments

Preparation of this article was aided in part by the Therapeutic Cognitive Neuroscience Professorship (BG); the Therapeutic Cognitive Neuroscience Fund (BG); the Benjamin and Adith Miller Family Endowment on Aging, Alzheimer's, and Autism (BG); the William and Mary Ann Wockenfuss Research Fund Endowment (BG); and by NICHD grant HD053312 (DJS). We thank all those who have been involved in our tDCS efforts, including Drs. Elizabeth Lacey, Kerry Ledoux, Rosalind Sadleir, Richard Skolasky, Kyongje Sung, and Jacqueline R. Weaver, and Megan A. Andrejczuk, Laura V. Bosley, Patrick Chambers, Mariya Chernenok, Julia Hernandez, Kara Reese, and Mark Varvaris. James Livengood and Dr. Rasika Jayatillake are no longer with The Johns Hopkins University.

See also: Agraphia (Acquired Dysgraphia); Aphasia; Bayesian Models in Neuroscience; Cognitive Control (Executive Function): Role of Prefrontal Cortex; Cognitive Neuropsychology, Methods of; Cognitive Neuroscience; Dyslexia, Developmental; Dyslexias, Acquired and Agraphia; Functional Brain Imaging of Language processes; Handedness and Language, Relationship between; Interhemispheric Interaction in Language and Cognitive Processes; Language Disorders, Treatment and Remediation of; Language Processing, Functional Magnetic Resonance Imaging of; Lateralization of Language as Demonstrated by Brain Imaging Procedures; Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects; Lexical Semantics; Lexicon; Long-Term Potentiation and Long-Term Depression; Neural Basis of Language Development; Neural Basis of Reading; Neural Plasticity; Neurological Approaches to Agrammatism; Prefrontal Cortex; Proper and Common Names: Impairments of Anomia; Sentence Comprehension, Psychology

of; Speech Errors, Psychology of; Speech Perception; Speech Processing, The Cortical Organization of; Speech Production, Neural Basis of; Speech Production, Psychology of; Split Brains and Human Consciousness; Syntax–Semantics Interface; Visual Word Recognition, Neurocognitive Psychology of.

Bibliography

Abbeduto, I., Brady, N., Kover, S.T., 2007. Language development and fragile X syndrome: profiles, syndrome-specificity, and within-syndrome differences. Mental Retardation and Developmental Disabilities Research Reviews 13, 36–46.

Accornero, N., Li Voti, P., La Riccia, M., Gregori, B., 2007. Visual evoked potentials modulation during direct current cortical polarization. Experimental Brain Research 178, 261–266.

Acheson, D.J., MacDonald, M.C., 2009. Verbal working memory and language production: common approaches to the serial ordering of verbal information. Psychological Bulletin 135, 50–68.

Alexander, T., Avirame, K., Lavidor, M., 2012. Improving emotional prosody detection in the attending ear by cathodal tDCS suppression of the competing channel. Neuroscience Letters 508, 52–55.

Alonzo, A., Brassil, J., Taylor, J.L., Martin, D., Loo, C.K., 2012. Daily transcranial direct current stimulation (tDCS) leads to greater increases in cortical excitability than second daily transcranial direct current stimulation. Brain Stimulation 5, 208–213.

Ambrus, G.G., Al-Moyed, H., Chaieb, L., Sarp, L., Antal, A., Paulus, W., 2012. The fade-in – short stimulation – fade out approach to sham tDCS – reliable at 1 mA for naive and experienced subjects, but not investigators. Brain Stimulation 5, 499–504.

Ambrus, G.G., Paulus, W., Antal, A., 2010. Cutaneous perception thresholds of electrical stimulation methods: comparison of tDCS and tRNS. Clinical Neurophysiology 121, 1908–1914.

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N.J., Habel, U., Schneider, F., Zilles, K., 2005. Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. Anatomy and Embryology (Berlin) 210, 343–352.

Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H.B., Zilles, K., 1999. Broca's region revisited: cytoarchitecture and intersubject variability. Journal of Comparative Neurology 412, 319–341.

Anastassiou, C.A., Perin, R., Markram, H., Koch, C., 2011. Ephaptic coupling of cortical neurons. Nature Neuroscience 14, 217–223.

Antal, A., Paulus, W., 2008. Transcranial direct current stimulation and visual perception. Perception 37, 367–374.

Antal, A., Varga, E.T., Kincses, T.Z., Nitsche, M.A., Paulus, W., 2004. Oscillatory brain activity and transcranial direct current stimulation in humans. Neuroreport 15, 1307–1310.

Ardolino, G., Bossi, B., Barbieri, S., Priori, A., 2005. Non-synaptic mechanisms underlie the after-effects of cathodal transcutaneous direct current stimulation of the human brain. The Journal of Physiology 568, 653–663.

Ariza, C.A., Fleury, A.T., Tormos, C.J., Petruk, V., Chawla, S., OH, J., Sakaguchi, D.S., Mallapragada, S.K., 2010. The influence of electric fields on hippocampal neural progenitor cells. Stem Cell Reviews 6, 585–600.

Arnsten, A.F., Wang, M.J., Paspalas, C.D., 2012. Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. Neuron 76, 223–239.

Bachmann, C.G., Muschinsky, S., Nitsche, M.A., Rolke, R., Magerl, W., Treede, R.D., Paulus, W., Happe, S., 2010. Transcranial direct current stimulation of the motor cortex induces distinct changes in thermal and mechanical sensory percepts. Clinical Neurophysiology 121, 2083–2089.

Bai, S., Loo, C., Dokos, S., 2013. A review of computational models of transcranial electrical stimulation. Critical Reviews in Biomedical Engineering 41, 21–35.

Benarroch, E.E., 2013. HCN channels: function and clinical implications. Neurology 80, 304–310.

Bergmann, T.O., Groppa, S., Seeger, M., Molle, M., Marshall, L., Siebner, H.R., 2009. Acute changes in motor cortical excitability during slow oscillatory and constant anodal transcranial direct current stimulation. Journal of Neurophysiology 102, 2303–2311.

Bikson, M., Datta, A., 2012. Guidelines for precise and accurate computational models of tDCS. Brain Stimulation 5, 430–431.

Bikson, M., Inoue, M., Akiyama, H., Deans, J.K., Fox, J.E., Miyakawa, H., Jefferys, J.G., 2004. Effects of uniform extracellular DC electric fields on excitability in rat hippocampal slices in vitro. The Journal of Physiology 557, 175–190.

Bikson, M., Rahman, A., Datta, A., 2012. Computational models of transcranial direct current stimulation. Clinical EEG and Neuroscience 43, 176–183.

Bindman, L.J., Lippold, O.C., Redfearn, J.W., 1964. The action of brief polarizing currents on the cerebral cortex of the rat (1) during current flow and (2) in the production of long-lasting after-effects. The Journal of Physiology 172, 369–382.

Bishop, D.V., McArthur, G.M., 2005. Individual differences in auditory processing in specific language impairment: a follow-up study using event-related potentials and behavioural thresholds. Cortex 41, 327–341.

Boggio, P.S., Zaghi, S., Lopes, M., Fregni, F., 2008. Modulatory effects of anodal transcranial direct current stimulation on perception and pain thresholds in healthy volunteers. European Journal of Neurology 15, 1124–1130.

Bogolepova, I.N., Malofeeva, L.I., 2004. Variability in the structure of field 39 of the lower parietal area of the cortex in the left and right hemispheres of adult human brains. Neuroscience and Behavioral Physiology 34, 363–367.

Bogolepova, I.N., Malofeeva, L.I., 2008. Individual variability in the cytoarchitectonics of anterior limbic field 24 in the human brain. Neuroscience and Behavioral Physiology 38, 737–741.

Bolognini, N., Rossetti, A., Casati, C., Mancini, F., Vallar, G., 2011. Neuromodulation of multisensory perception: a tDCS study of the sound-induced flash illusion. Neuropsychologia 49, 231–237.

Borckardt, J.J., Bikson, M., Frohman, H., Reeves, S.T., Datta, A., Bansal, V., Madan, A., Barth, K., George, M.S., 2012. A pilot study of the tolerability and effects of high-definition transcranial direct current stimulation (HD-tDCS) on pain perception. The Journal of Pain: Official Journal of the American Pain Society 13, 112–120.

Borella, E., Chicherio, C., Re, A.M., Sensini, V., Cornoldi, C., 2011. Increased intra-individual variability is a marker of ADHD but also of dyslexia: a study on handwriting. Brain and Cognition 77, 33–39.

Bressler, S.L., Menon, V., 2010. Large-scale brain networks in cognition: emerging methods and principles. Trends in Cognitive Sciences 14, 277–290.

Brignani, D., Ruzzoli, M., Mauri, P., Miniussi, C., 2013. Is transcranial alternating current stimulation effective in modulating brain oscillations? PLoS One 8, e56589.

Brown, H., Prescott, R., 2006. Applied Mixed Models in Medicine. John Wiley & Sons Ltd, UK.

Budden, S., 2012. Clinical variability in early speech-language development in females with Rett syndrome. Developmental Medicine and Child Neurology 54, 392–393.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafo, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14, 365–376.

Cambiaghi, M., Velikova, S., Gonzalez-Rosa, J.J., Cursi, M., Comi, G., Leocani, L., 2010. Brain transcranial direct current stimulation modulates motor excitability in mice. European Journal of Neuroscience 31, 704–709.

Caplan, D., 1988. On the role of group studies in neuropsychological and pathopsychological research. Cognitive Neuropsychology 5, 535–547.

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., Zilles, K., 2006. The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. Neuroimage 33, 430–448.

Cattaneo, Z., Pisoni, A., Papagno, C., 2011. Transcranial direct current stimulation over Broca's region improves phonemic and semantic fluency in healthy individuals. Neuroscience 183, 64–70.

Cerruti, C., Schlaug, G., 2009. Anodal transcranial direct current stimulation of the prefrontal cortex enhances complex verbal associative thought. Journal of Cognitive Neuroscience 21, 1980–1987.

Chandrasekaran, B., Sampath, P.D., Wong, P.C., 2010. Individual variability in cue-weighting and lexical tone learning. Journal of the Acoustical Society of America 128, 456–465.

Chi, R.P., Fregni, F., Snyder, A.W., 2010. Visual memory improved by non-invasive brain stimulation. Brain Research 1353, 168–175.

Chi, R.P., Snyder, A.W., 2011. Facilitate insight by non-invasive brain stimulation. PLoS One 6, e16655.

Chi, R.P., Snyder, A.W., 2012. Brain stimulation enables the solution of an inherently difficult problem. Neuroscience Letters 515, 121–124.

Chiarello, C., Welcome, S.E., Leonard, C.M., 2012. Individual differences in reading skill and language lateralisation: a cluster analysis. Laterality 17, 225–251.

Chow, S.-C., Wang, H., Shao, J. (Eds.), 2008. Sample Size Calculations in Clinical Research. Chapman and Hall/CRC Biostatistics Series, Boca Raton.

Chrysikou, E.G., Hamilton, R.H., 2011. Noninvasive brain stimulation in the treatment of aphasia: exploring interhemispheric relationships and their implications for neurorehabilitation. Restorative Neurology and Neuroscience 29, 375–394.

Chrysikou, E.G., Hamilton, R.H., Coslett, H.B., Datta, A., Bikson, M., Thompson-Schill, S.L., 2013. Noninvasive transcranial direct current stimulation over the left prefrontal cortex facilitates cognitive flexibility in tool use. Cognitive Neuroscience 4 (2), 81–89.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Inc, USA.

Creutzfeldt, O.D., Fromm, G.H., Kapp, H., 1962. Influence of transcortical d-c currents on cortical neuronal activity. Experimental Neurology 5, 436–452.

Datta, A., Bansal, V., Diaz, J., Patel, J., Reato, D., Bikson, M., 2009. Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad. Brain Stimulation 2, 201–207, 207 e1.

Datta, A., Dmochowski, J.P., Guleyupoglu, B., Bikson, M., Fregni, F., 2013. Cranial electrotherapy stimulation and transcranial pulsed current stimulation: a computer based high-resolution modeling study. Neuroimage 65, 280–287.

Dehaene, S., Cohen, L., 2011. The unique role of the visual word form area in reading. Trends in Cognitive Sciences 15, 254–262.

Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., Gagnon, D.A., 1997. Lexical access in aphasic and nonaphasic speakers. Psychological Review 104, 801–838.

Derrfuss, J., Brass, M., Von Cramon, D.Y., Lohmann, G., Amunts, K., 2009. Neural activations at the junction of the inferior frontal sulcus and the inferior precentral sulcus: interindividual variability, reliability, and association with sulcal morphology. Human Brain Mapping 30, 299–311.

Di Lazzaro, V., Manganelli, F., Dileone, M., Notturno, F., Esposito, M., Capasso, M., Dubbioso, R., Pace, M., Ranieri, F., Minicuci, G., Santoro, L., Uncini, A., 2012. The effects of prolonged cathodal direct current stimulation on the excitatory and inhibitory circuits of the ipsilateral and contralateral motor cortex. Journal of Neural Transmission 119, 1499–1506.

Dieckhofer, A., Waberski, T.D., Nitsche, M., Paulus, W., Buchner, H., Gobbele, R., 2006. Transcranial direct current stimulation applied over the somatosensory cortex – differential effect on low and high frequency SEPs. Clinical Neurophysiology 117, 2221–2227.

Dockery, C.A., Hueckel-Weng, R., Birbaumer, N., Plewnia, C., 2009. Enhancement of planning ability by transcranial direct current stimulation. Journal of Neuroscience 29, 7271–7277.

Dougherty, R.F., Koch, V.M., Brewer, A.A., Fischer, B., Modersitzki, J., Wandell, B.A., 2003. Visual field representations and locations of visual areas V1/2/3 in human visual cortex. Journal of Vision 3, 586–598.

Dundas, J.E., Thickbroom, G.W., Mastaglia, F.L., 2007. Perception of comfort during transcranial DC stimulation: effect of NaCl solution concentration applied to sponge electrodes. Clinical Neurophysiology 118, 1166–1170.

Dykiert, D., Der, G., Starr, J.M., Deary, I.J., 2012. Age differences in intra-individual variability in simple and choice reaction time: systematic review and meta-analysis. PLoS One 7, e45759.

Edwards, D., Cortes, M., Datta, A., Minhas, P., Wassermann, E.M., Bikson, M., 2013. Physiological and modeling evidence for focal transcranial electrical brain stimulation in humans: a basis for high-definition tDCS. Neuroimage 74, 266–275.

Elsabbagh, M., Cohen, H., Cohen, M., Rosen, S., Karmiloff-Smith, A., 2011. Severity of hyperacusis predicts individual differences in speech perception in Williams Syndrome. Journal of Intellectual Disability Research 55, 563–571.

Evans, K.E., Demuth, K., 2012. Individual differences in pronoun reversal: evidence from two longitudinal case studies. Journal of Child Language 39, 162–191.

Feurra, M., Paulus, W., Walsh, V., Kanai, R., 2011. Frequency specific modulation of human somatosensory cortex. Frontiers in Psychology 2, 13.

Floel, A., 2012. Non-invasive brain stimulation and language processing in the healthy brain. Aphasiology 26, 1082–1102.

Fornito, A., Wood, S.J., Whittle, S., Fuller, J., Adamson, C., Saling, M.M., Velakoulis, D., Pantelis, C., Yucel, M., 2008. Variability of the paracingulate sulcus and morphometry of the medial frontal cortex: associations with cortical thickness, surface area, volume, and sulcal depth. Human Brain Mapping 29, 222–236.

Francis, J.T., Gluckman, B.J., Schiff, S.J., 2003. Sensitivity of neurons to weak electric fields. Journal of Neuroscience 23, 7255–7261.

Fricke, K., Seeber, A.A., Thirugnanasambandam, N., Paulus, W., Nitsche, M.A., Rothwell, J.C., 2011. Time course of the induction of homeostatic plasticity generated by repeated transcranial direct current stimulation of the human motor cortex. Journal of Neurophysiology 105, 1141–1149.

Fridriksson, J., Hubbard, H.I., Hudspeth, S.G., 2012. Transcranial brain stimulation to treat aphasia: a clinical perspective. Seminars in Speech and Language 33, 188–202.

Friederici, A.D., Gierhan, S.M., 2013. The language network. Current Opinion in Neurobiology 23, 250–254.

Friederici, A.D., Mueller, J.L., Sehm, B., Ragert, P., 2013. Language learning without control: the role of the PFC. Journal of Cognitive Neuroscience 25 (5), 814–821.

Fritsch, B., Reis, J., Martinowich, K., Schambra, H.M., Ji, Y., Cohen, L.G., Lu, B., 2010. Direct current stimulation promotes BDNF-dependent synaptic plasticity: potential implications for motor learning. Neuron 66, 198–204.

Frohlich, F., McCormick, D.A., 2010. Endogenous electric fields may guide neocortical network activity. Neuron 67, 129–143.

Fuster, J.M., Bressler, S.L., 2012. Cognit activation: a mechanism enabling temporal integration in working memory. Trends in Cognitive Sciences 16, 207–218.

Galea, J.M., Celnik, P., 2009. Brain polarization enhances the formation and retention of motor memories. Journal of Neurophysiology 102, 294–301.

Gandiga, P.C., Hummel, F.C., Cohen, L.G., 2006. Transcranial DC stimulation (tDCS): a tool for double-blind sham-controlled clinical studies in brain stimulation. Clinical Neurophysiology 117, 845–850.

Garcia, C., Schoebel, C., Von Mengden, I., Glos, M., Fietze, I., Penzel, T., 2011. Influence of slow oscillating transcranial direct current stimulation (so-tDCS) on electroencephalogram (EEG) in general and sleep related parameters in healthy subjects. Clinical Neurophysiology 122, S121.

Gartside, I.B., 1968a. Mechanisms of sustained increases of firing rate of neurones in the rat cerebral cortex after polarization: role of protein synthesis. Nature 220, 383–384.

Gartside, I.B., 1968b. Mechanisms of sustained increases of firing rate of neurons in the rat cerebral cortex after polarization: reverberating circuits or modification of synaptic conductance? Nature 220, 382–383.

Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience 15, 511–517.

Gooch, D., Snowling, M.J., Hulme, C., 2012. Reaction time variability in children with ADHD symptoms and/or dyslexia. Developmental Neuropsychology 37, 453–472.

Gordon, B., 1997. Models of naming. In: Goodglass, H., Wingfield, A. (Eds.), Anomia: Neuroanatomical and Cognitive Correlates, Foundations of Neuropsychology. Academic Press, San Diego, CA, pp. 31–64.

Gordon, B., Boatman, D., Lesser, R.P., Hart, J.J., 1999. Recovery from acquired reading and naming deficits following focal cerebral lesions. In: Broman, S.H., Fletcher, J.M. (Eds.), The Changing Nervous System: Neurobehavioral Consequences of Early Brain Disorders. Oxford University Press, New York, pp. 199–213.

Gruber, O., Goschke, T., 2004. Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. Acta Psychologica (Amsterdam) 115, 105–121.

Gutchess, A.H., Hedden, T., Ketay, S., Aron, A., Gabrieli, J.D., 2010. Neural differences in the processing of semantic relationships across cultures. Social Cognitive and Affective Neuroscience 5, 254–263.

Haley, K.L., Martin, G., 2011. Production variability and single word intelligibility in aphasia and apraxia of speech. Journal of Communication Disorders 44, 103–115.

Hart Jr., J., Crone, N.E., Lesser, R.P., Sieracki, J., Miglioretti, D.L., Hall, C., Sherman, D., Gordon, B., 1998. Temporal dynamics of verbal object comprehension. Proceedings of the National Academy of Sciences of the United States of America 95, 6498–6503.

Hattori, Y., Moriwaki, A., Hori, Y., 1990. Biphasic effects of polarizing current on adenosine-sensitive generation of cyclic AMP in rat cerebral cortex. Neuroscience Letters 116, 320–324.

Hayiou-Thomas, M.E., Dale, P.S., Plomin, R., 2012. The etiology of variation in language skills changes with development: a longitudinal twin study of language from 2 to 12 years. Developmental Science 15, 233–249.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? The Behavioral and Brain Sciences 33, 61–83 discussion 83–135.

Heo, C., Yoo, J., Lee, S.Y., Lee, S., Joo, E.Y., Hong, S.B., Lee, Y.H., Suh, M., 2012. Enhanced mobility of neural cells with a transparent electric field stimulator. Journal of Nanoscience and Nanotechnology 12, 5222–5227.

Hervais-Adelman, A.G., Moser-Mercer, B., Golestani, N., 2011. Executive control of language in the bilingual brain: integrating the evidence from neuroimaging to neuropsychology. Frontiers in Psychology 2, 234.

Heyanka, D.J., Holster, J.L., Golden, C.J., 2013. Intraindividual neuropsychological test variability in healthy individuals with high average intelligence and educational attainment. International Journal of Neuroscience 123 (8), 526–531.

Hills, M., Armitage, P., 1979. The two-period cross-over clinical trial. British Journal of Clinical Pharmacology 8, 7–20.

Hills, T.T., Jones, M.N., Todd, P.M., 2012. Optimal foraging in semantic memory. Psychological Review 119, 431–440.

Hoshino, O., 2011. Subthreshold membrane depolarization as memory trace for perceptual learning. Neural Computation 23, 3205–3231.

Huang, Y., Su, Y., Rorden, C., Dmochowski, J., Datta, A., Parra, L.C., 2012. An Automated Method for High-Definition Transcranial Direct Current Stimulation Modeling. San Diego, CA, pp. 5376–5379.

Huey, E.D., Probasco, J.C., Moll, J., Stocking, J., Ko, M.H., Grafman, J., Wassermann, E.M., 2007. No effect of DC brain polarization on verbal fluency in patients with advanced frontotemporal dementia. Clinical Neurophysiology 118, 1417–1418.

Hugdahl, K., Westerhausen, R. (Eds.), 2010. The Two Halves of the Brain: Information Processing in the Cerebral Hemispheres. MIT Press, Cambridge.

Hulme, S.R., Jones, O.D., Abraham, W.C., 2013. Emerging roles of metaplasticity in behaviour and disease. Trends in Neurosciences 36 (6), 353–362.

Hurks, P.P., Hendriksen, J.G., Vles, J.S., Kalff, A.C., Feron, F.J., Kroes, M., Van Zeben, T.M., Steyaert, J., Jolles, J., 2004. Verbal fluency over time as a measure of automatic and controlled processing in children with ADHD. Brain and Cognition 55, 535–544.

Ioannidis, J.P.A., 2012. Why science is not necessarily self-correcting. Perspectives on Psychological Science 7, 645–654.

Islam, N., Aftabuddin, M., Moriwaki, A., Hattori, Y., Hori, Y., 1995a. Increase in the calcium level following anodal polarization in the rat brain. Brain Research 684, 206–208.

Islam, N., Moriwaki, A., Hori, Y., 1995b. Co-localization of c-fos protein and protein kinase C gamma in the rat brain following anodal polarization. Indian Journal of Physiology and Pharmacology 39, 209–215.

Jackendoff, R., 2002. Foundations of Language: Brain Meaning, Grammar, Evolution. Oxford University Press, New York.

Jacobson, L., Javitt, D.C., Lavidor, M., 2011. Activation of inhibition: diminishing impulsive behavior by direct current stimulation over the inferior frontal gyrus. Journal of Cognitive Neuroscience 23, 3380–3387.

Jiang, T., Xu, R.X., Zhang, A.W., Di, W., Xiao, Z.J., Miao, J.Y., Luo, N., Fang, Y.N., 2012. Effects of transcranial direct current stimulation on hemichannel pannexin-1 and neural plasticity in rat model of cerebral infarction. Neuroscience 226, 421–426.

Jones, B., Kenward, M.G., 2003. Design and Analysis of Cross-Over Trials. Chapman and Hall/CRC, Boca-Raton.

Kanai, R., Paulus, W., Walsh, V., 2010. Transcranial alternating current stimulation (tACS) modulates cortical excitability as assessed by TMS-induced phosphene thresholds. Clinical Neurophysiology 121, 1551–1554.

Kanai, R., Rees, G., 2011. The structural basis of inter-individual differences in human behaviour and cognition. Nature Reviews Neuroscience 12, 231–242.

Kase, D., Imoto, K., 2012. The role of HCN channels on membrane excitability in the nervous system. Journal of Signal Transduction 2012, 619747.

Keeser, D., Meindl, T., Bor, J., Palm, U., Pogarell, O., Mulert, C., Brunelin, J., Moller, H.J., Reiser, M., Padberg, F., 2011. Prefrontal transcranial direct current stimulation changes connectivity of resting-state networks during fMRI. Journal of Neuroscience 31, 15284–15293.

Keller, S.S., Highley, J.R., Garcia-Finana, M., Sluming, V., Rezaie, R., Roberts, N., 2007. Sulcal variability, stereological measurement and asymmetry of Broca's area on MR images. Journal of Anatomy 211, 534–555.

Keller, S.S., Roberts, N., Hopkins, W., 2009. A comparative magnetic resonance imaging study of the anatomy, variability, and asymmetry of Broca's area in the human and chimpanzee brain. Journal of Neuroscience 29, 14607–14616.

Kessler, S.K., Turkeltaub, P.E., Benson, J.G., Hamilton, R.H., 2012. Differences in the experience of active and sham transcranial direct current stimulation. Brain Stimulation 5, 155–162.

Koenigs, M., Ukueberuwa, D., Campion, P., Grafman, J., Wassermann, E., 2009. Bilateral frontal transcranial direct current stimulation: failure to replicate classic findings in healthy subjects. Clinical Neurophysiology 120, 80–84.

Koppes, A.N., Seggio, A.M., Thompson, D.M., 2011. Neurite outgrowth is significantly increased by the simultaneous presentation of Schwann cells and moderate exogenous electric fields. Journal of Neural Engineering 8, 046023.

Kuo, M.F., Grosch, J., Fregni, F., Paulus, W., Nitsche, M.A., 2007. Focusing effect of acetylcholine on neuroplasticity in the human motor cortex. Journal of Neuroscience 27, 14442–14447.

Kuo, M.F., Paulus, W., Nitsche, M.A., 2008. Boosting focally-induced brain plasticity by dopamine. Cerebral Cortex 18, 648–651.

Kwon, Y.H., Ko, M.H., Ahn, S.H., Kim, Y.H., Song, J.C., Lee, C.H., Chang, M.C., Jang, S.H., 2008. Primary motor cortex activation by transcranial direct current stimulation in the human brain. Neuroscience Letters 435, 56–59.

Kwon, Y.H., Nam, K.S., Lee, M.Y., Park, J.W., Jang, S.H., 2011. The temporal change of cortical activation induced by the ongoing effects of transcranial direct current stimulation. Journal of Physical Therapy Science 23, 65–69.

Ladeira, A., Fregni, F., Campanhã, C., Valasek, C.A., De Ridder, D., Brunoni, A.R., Boggio, P.S., 2011. Polarity-dependent transcranial direct current stimulation effects on central auditory processing. PLoS One 6.

Lapenta, O.M., Fregni, F., Oberman, L.M., Boggio, P.S., 2012. Bilateral temporal cortex transcranial direct current stimulation worsens male performance in a multisensory integration task. Neuroscience Letters 527, 105–109.

Li, S., Han, Y., Wang, D., Yang, H., Fan, Y., LV, Y., Tang, H., Gong, Q., Zang, Y., He, Y., 2010. Mapping surface variability of the central sulcus in musicians. Cerebral Cortex 20, 25–33.

Liebetanz, D., Nitsche, M.A., Tergau, F., Paulus, W., 2002. Pharmacological approach to the mechanisms of transcranial DC-stimulation-induced after-effects of human motor cortex excitability. Brain 125, 2238–2247.

Magezi, D.A., Khateb, A., Mouthon, M., Spierer, L., Annoni, J.M., 2012. Cognitive control of language production in bilinguals involves a partly independent process within the domain-general cognitive control network: evidence from task-switching and electrical brain activity. Brain and Language 122, 55–63.

Makel, M.C., Plucker, J.A., Hegarty, B., 2012. Replications in psychology research: how often do they really occur? Perspectives on Psychological Science 7, 537–542.

Malenka, R.C., Bear, M.F., 2004. LTP and LTD: an embarrassment of riches. Neuron 44, 5–21.

Mangin, J.F., Riviere, D., Cachia, A., Duchesnay, E., Cointepas, Y., Papadopoulos-Orfanos, D., Scifo, P., Ochiai, T., Brunelle, F., Regis, J., 2004. A framework to study the cortical folding patterns. Neuroimage 23 (Suppl. 1), S129–S138.

Marinus, E., de Jong, P.F., 2010. Variability in the word-reading performance of dyslexic readers: effects of letter length, phoneme length and digraph presence. Cortex 46, 1259–1271.

Marquez-Ruiz, J., Leal-Campanario, R., Molaee-Ardekani, B., Gruart, A., Sanchez-Campusano, R., Ruffini, G., Wendling, F., Delgado-Garcia, J.M., 2011. Transcranial direct current stimulation (tDCS) effects on somatosensory local field potential of alert rabbits. Clinical Neurophysiology 122, S125.

Marquez-Ruiz, J., Leal-Campanario, R., Sanchez-Campusano, R., Molaee-Ardekani, B., Wendling, F., Miranda, P.C., Ruffini, G., Gruart, A., Delgado-Garcia, J.M., 2012. Transcranial direct-current stimulation modulates synaptic mechanisms involved in associative learning in behaving rabbits. Proceedings of the National Academy of Sciences of the United States of America 109, 6710–6715.

Matsunaga, K., Nitsche, M.A., Tsuji, S., Rothwell, J.C., 2004. Effect of transcranial DC sensorimotor cortex stimulation on somatosensory evoked potentials in humans. Clinical Neurophysiology 115, 456–460.

McFadden, J.L., Borckardt, J.J., George, M.S., Beam, W., 2010. Reducing procedural pain and discomfort associated with transcranial direct current stimulation. Brain Stimulation 4, 38–42.

Meinzer, M., Antonenko, D., Lindenberg, R., Hetzer, S., Ulm, L., Avirame, K., Flaisch, T., Floel, A., 2012. Electrical brain stimulation improves cognitive performance by modulating functional connectivity and task-specific activation. Journal of Neuroscience 32, 1859–1866.

Menon, V., Uddin, L.Q., 2010. Saliency, switching, attention and control: a network model of insula function. Brain Structure & Function 214, 655–667.

Merlet, I., Birot, G., Salvador, R., Molaee-Ardekani, B., Mekonnen, A., Soria-Frish, A., Ruffini, G., Miranda, P.C., Wendling, F., 2013. From oscillatory transcranial current stimulation to scalp EEG changes: a biophysical and physiological modeling study. PLoS One 8.

Metuki, N., Sela, T., Lavidor, M., 2012. Enhancing cognitive control components of insight problems solving by anodal tDCS of the left dorsolateral prefrontal cortex. Brain Stimulation 5, 110–115.

Miller, E.K., Buschman, T.J., 2013. Cortical circuits for the control of attention. Current Opinion in Neurobiology 23 (21), 216–222.

Minhas, P., Datta, A., Bikson, M., 2011. Cutaneous perception during tDCS: role of electrode shape and sponge salinity. Clinical Neurophysiology 122, 637–638.

Miranda, P.C., Faria, P., Hallett, M., 2009. What does the ratio of injected current to electrode area tell us about current density in the brain during tDCS? Clinical Neurophysiology 120, 1183–1187.

Miranda, P.C., Mekonnen, A., Salvador, R., Ruffini, G., 2013. The electric field in the cortex during transcranial current stimulation. Neuroimage 70, 48–58.

Monte-Silva, K., Kuo, M.F., Hessenthaler, S., Fresnoza, S., Liebetanz, D., Paulus, W., Nitsche, M.A., 2012. Induction of late LTP-like plasticity in the human motor cortex by repeated non-invasive brain stimulation. Brain Stimulation 6 (3), 424–432.

Monte-Silva, K., Kuo, M.F., Liebetanz, D., Paulus, W., Nitsche, M.A., 2010. Shaping the optimal repetition interval for cathodal transcranial direct current stimulation (tDCS). Journal of Neurophysiology 103, 1735–1740.

Monte-Silva, K., Kuo, M.F., Thirugnanasambandam, N., Liebetanz, D., Paulus, W., Nitsche, M.A., 2009. Dose-dependent inverted U-shaped effect of dopamine (D2-like) receptor activation on focal and nonfocal plasticity in humans. Journal of Neuroscience 29, 6124–6131.

Monti, A., Ferrucci, R., Fumagalli, M., Mameli, F., Cogiamanian, F., Ardolino, G., Priori, A., 2012. Transcranial direct current stimulation (tDCS) and language. Journal of Neurology, Neurosurgery, and Psychiatry.

Moriwaki, A., Hattori, Y., Hayashi, Y., Lu, Y.F., Islam, N., Hori, Y., 1994. Repeated application of anodal direct current produces regional dominance in histamine-elicited cyclic AMP accumulation in rabbit cerebral cortex. Acta Medica Okayama 48, 323–326.

Mulquiney, P.G., Hoy, K.E., Daskalakis, Z.J., Fitzgerald, P.B., 2011. Improving working memory: exploring the effect of transcranial random noise stimulation and transcranial direct current stimulation on the dorsolateral prefrontal cortex. Clinical Neurophysiology 122, 2384–2389.

Neuling, T., Rach, S., Herrmann, C.S., 2013. Orchestrating neuronal networks: sustained after-effects of transcranial alternating current stimulation depend upon brain states. Frontiers in Human Neuroscience 7, 161.

Niemela, P.T., Vainikka, A., Forsman, J.T., Loukola, O.J., Kortet, R., 2013. How does variation in the environment and individual cognition explain the existence of consistent behavioral differences? Ecology and Evolution 3, 457–464.

Nitsche, M.A., Fricke, K., Henschke, U., Schlitterlau, A., Liebetanz, D., Lang, N., Henning, S., Tergau, F., Paulus, W., 2003. Pharmacological modulation of cortical excitability shifts induced by transcranial direct current stimulation in humans. The Journal of Physiology 553, 293–301.

Nitsche, M.A., Grundey, J., Liebetanz, D., Lang, N., Tergau, F., Paulus, W., 2004a. Catecholaminergic consolidation of motor cortical neuroplasticity in humans. Cerebral Cortex 14, 1240–1245.

Nitsche, M.A., Jaussi, W., Liebetanz, D., Lang, N., Tergau, F., Paulus, W., 2004b. Consolidation of human motor cortical neuroplasticity by D-cycloserine. Neuropsychopharmacology 29, 1573–1578.

Nitsche, M.A., Kuo, M.F., Grosch, J., Bergner, C., Monte-Silva, K., Paulus, W., 2009. D1-receptor impact on neuroplasticity in humans. Journal of Neuroscience 29, 2648–2653.

Nitsche, M.A., Lampe, C., Antal, A., Liebetanz, D., Lang, N., Tergau, F., Paulus, W., 2006. Dopaminergic modulation of long-lasting direct current-induced cortical excitability changes in the human motor cortex. European Journal of Neuroscience 23, 1651–1657.

Nitsche, M.A., Paulus, W., 2000. Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. The Journal of Physiology 527 (Pt 3), 633–639.

Nitsche, M.A., Paulus, W., 2001. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. Neurology 57, 1899–1901.

Nitsche, M.A., Roth, A., Kuo, M.F., Fischer, A.K., Liebetanz, D., Lang, N., Tergau, F., Paulus, W., 2007. Timing-dependent modulation of associative plasticity by general network excitability in the human motor cortex. Journal of Neuroscience 27, 3807–3812.

Nitsche, M.A., Seeber, A., Frommann, K., Mein, C.C., Rochford, C., Nitsche, M.S., Fricke, K., Liebetanz, D., Lang, N., Antal, A., Paulus, W., Tergau, F., 2005. Modulating parameters of excitability during and after transcranial direct current stimulation of the human motor cortex. Journal of Physiology (London) 568, 291–303.

O'Connell, N.E., Cossar, J., Marston, L., Wand, B.M., Bunce, D., Moseley, G.L., De Souza, L.H., 2012. Rethinking clinical trials of transcranial direct current stimulation: participant and assessor blinding is inadequate at intensities of 2mA. PLoS One 7, e47514.

Ode, S., Robinson, M.D., Hanson, D.M., 2011. Cognitive-emotional dysfunction among noisy minds: predictions from individual differences in reaction time variability. Cognition & Emotion 25, 307–327.

Ohn, S.H., Park, C.I., Yoo, W.K., Ko, M.H., Choi, K.P., Kim, G.M., Lee, Y.T., Kim, Y.H., 2008. Time-dependent effect of transcranial direct current stimulation on the enhancement of working memory. Neuroreport 19, 43–47.

Owen, G.R., Brenner, E.A., 2012. Mapping molecular memory: navigating the cellular pathways of learning. Cellular and Molecular Neurobiology 32, 919–941.

Palm, U., Reisinger, E., Keeser, D., Kuo, M.F., Pogarell, O., Leicht, G., Mulert, C., Nitsche, M.A., Padberg, F., 2013. Evaluation of sham transcranial direct current stimulation for randomized, placebo-controlled clinical trials. Brain Stimulation 6, 690–695.

Parazzini, M., Fiocchi, S., Rossi, E., Paglialonga, A., Ravazzani, P., 2011. Transcranial direct current stimulation: estimation of the electric field and of the current density in an anatomical human head model. IEEE Transactions on Biomedical Engineering 58, 1773–1780.

Park, C.H., Chang, W.H., Park, J.Y., Shin, Y.I., Kim, S.T., Kim, Y.H., 2013. Transcranial direct current stimulation increases resting state interhemispheric connectivity. Neuroscience Letters 539, 7–10.

Pellicano, E., 2010. Individual differences in executive function and central coherence predict developmental changes in theory of mind in autism. Developmental Psychology 46, 530–544.

Pena-Gomez, C., Sala-Lonch, R., Junque, C., Clemente, I.C., Vidal, D., Bargallo, N., Falcon, C., Valls-Sole, J., Pascual-Leone, A., Bartres-Faz, D., 2012. Modulation of large-scale brain networks by transcranial direct current stimulation evidenced by resting-state functional MRI. Brain Stimulation 5, 252–263.

Pereira, J.B., Junqué, C., Bartrés-Faz, D., Martí, M.J., Sala-Llonch, R., Compta, Y., Falcón, C., Vendrell, P., Pascual-Leone, Á., Valls-Solé, J., Tolosa, E., 2012. Modulation of verbal fluency networks by transcranial direct current stimulation (tDCS) in parkinson's disease. Brain Stimulation 6 (1), 16–24.

Peters, M.A., Thompson, B., Merabet, L.B., Wu, A.D., Shams, L., 2013. Anodal tDCS to V1 blocks visual perceptual learning consolidation. Neuropsychologia 51 (7), 1234–1239.

Pirulli, C., Fertonani, A., Miniussi, C., 2013. The role of timing in the induction of neuromodulation in perceptual learning by transcranial electric stimulation. Brain Stimulation 6, 683–689.

Pnevmatikos, D., Trikkaliotis, I., 2013. Intraindividual differences in executive functions during childhood: the role of emotions. The Journal of Experimental Child Psychology 115, 245–261.

Polania, R., Nitsche, M.A., Paulus, W., 2011. Modulating functional connectivity patterns and topological functional organization of the human brain with transcranial direct current stimulation. Human Brain Mapping 32, 1236–1249.

Polania, R., Paulus, W., Nitsche, M.A., 2012. Reorganizing the intrinsic functional architecture of the human primary motor cortex during rest with non-invasive cortical stimulation. PLoS One 7, e30971.

Pope, P.A., Miall, R.C., 2012. Task-specific facilitation of cognition by cathodal transcranial direct current stimulation of the cerebellum. Brain Stimulation 5, 84–94.

Prat, C.S., Just, M.A., 2011. Exploring the neural dynamics underpinning individual differences in sentence comprehension. Cerebral Cortex 21, 1747–1760.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847.

Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews. Drug Discovery 10, 712.

Purpura, D.P., McMurtry, J.G., 1965. Intracellular activities and evoked potential changes during polarization of motor cortex. Journal of Neurophysiology 28, 166–185.

Putt, M., Chinchilli, V.M., 1999. A mixed effects model for the analysis of repeated measures cross-over studies. Stat. Med. 18, 3037–3058.

Radman, T., Ramos, R.L., Brumberg, J.C., Bikson, M., 2009. Role of cortical cell type and morphology in subthreshold and suprathreshold uniform electric field stimulation in vitro. Brain Stimulation 2, 215–228.e3.

Ragert, P., Vandermeeren, Y., Camus, M., Cohen, L.G., 2008. Improvement of spatial tactile acuity by transcranial direct current stimulation. Clinical Neurophysiology 119, 805–811.

Rahman, A., Reato, D., Arlotti, M., Gasca, F., Datta, A., Parra, L.C., Bikson, M., 2013. Cellular effects of acute direct current stimulation: somatic and synaptic terminal effects. The Journal of Physiology 591 (10), 2563–2578.

Raising Standards, 2013. Nature journals' updated editorial policies aim to improve transparency and reproducibility. Nature Neuroscience 16, 517.

Rango, M., Cogiamanian, F., Marceglia, S., Barberis, B., Arighi, A., Biondetti, P., Priori, A., 2008. Myoinositol content in the human brain is modified by transcranial direct current stimulation in a matter of minutes: a 1H-MRS study. Magnetic Resonance in Medicine 60, 782–789.

Ranieri, F., Podda, M.V., Riccardi, E., Frisullo, G., Dileone, M., Profice, P., Pilato, F., Di Lazzaro, V., Grassi, C., 2012. Modulation of LTP at rat hippocampal CA3-CA1 synapses by direct current stimulation. Journal of Neurophysiology 107, 1868–1880.

Ratcliff, R., 1978. A theory of memory retrieval. Psychological Review 85, 59–108.

Reato, D., Rahman, A., Bikson, M., Parra, L.C., 2010. Low-intensity electrical stimulation affects network dynamics by modulating population rate and spike timing. Journal of Neuroscience 30, 15067–15079.

Reggia, J.A., D'autrechy, C.L., Sutton III, G.G., Weinrich, M., 1992. A competitive distribution theory of neocortical dynamics. Neural Computation 4, 287–317.

Reis, J., Schambra, H.M., Cohen, L.G., Buch, E.R., Fritsch, B., Zarahn, E., Celnik, P.A., Krakauer, J.W., 2009. Noninvasive cortical stimulation enhances motor skill acquisition over multiple days through an effect on consolidation. Proceedings of the National Academy of Sciences of the United States of America 106, 1590–1595.

Ridding, M.C., Ziemann, U., 2010. Determinants of the induction of cortical plasticity by non-invasive brain stimulation in healthy subjects. The Journal of Physiology 588, 2291–2304.

Rogalewski, A., Breitenstein, C., Nitsche, M.A., Paulus, W., Knecht, S., 2004. Transcranial direct current stimulation disrupts tactile perception. European Journal of Neuroscience 20, 313–316.

Roth, B.J., 2009. What does the ratio of injected current to electrode area not tell us about tDCS? Clinical Neurophysiology 120, 1037–1038.

Roth, B.J., 2012. Are glial cells responsible for transcranial direct current stimulation? Clinical Neurophysiology 123, 1901.

Royer, S., Zemelman, B.V., Losonczy, A., Kim, J., Chance, F., Magee, J.C., Buzsaki, G., 2012. Control of timing, rate and bursts of hippocampal place cells by dendritic and somatic inhibition. Nature Neuroscience 15, 769–775.

Ruffini, G., Wendling, F., Merlet, I., Molaee-Ardekani, B., Mekkonen, A., Salvador, R., Soria-Frisch, A., Grau, C., Dunne, S., Miranda, P., 2012. Transcranial current brain stimulation (tCS): models and technologies. IEEE Transactions on Neural Systems and Rehabilitation Engineering 21 (3), 333–345.

Ruohonen, J., Karhu, J., 2012. tDCS possibly stimulates glial cells. Clinical Neurophysiology 123, 2006–2009.

Sadleir, R.J., Vannorsdall, T.D., Schretlen, D.J., Gordon, B., 2012. Target optimization in transcranial direct current stimulation. Frontiers in Psychiatry/Frontiers Research Foundation 3, 90.

Sanz-Clemente, A., Nicoll, R.A., Roche, K.W., 2013. Diversity in NMDA receptor composition: many regulators, many consequences. Neuroscientist 19, 62–75.

Savtchenko, L.P., Korogod, S.M., Rusakov, D.A., 2000. Electrodiffusion of synaptic receptors: a mechanism to modify synaptic efficacy? Synapse 35, 26–38.

Savtchenko, L.P., Kulahin, N., Korogod, S.M., Rusakov, D.A., 2004. Electric fields of synaptic currents could influence diffusion of charged neurotransmitter molecules. Synapse 51, 270–278.

Schlaug, G., Marchina, S., Wan, C.Y., 2011. The use of non-invasive brain stimulation techniques to facilitate recovery from post-stroke aphasia. Neuropsychological Review 21, 288–301.

Schretlen, D.J., Testa, S.M., Winicki, J.M., Pearlson, G.D., Gordon, B., 2008. Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. Journal of the International Neuropsychological Society 14, 436–445.

Sehm, B., Schafer, A., Kipping, J., Margulies, D., Conde, V., Taubert, M., Villringer, A., Ragert, P., 2012. Dynamic modulation of intrinsic functional connectivity by transcranial direct current stimulation. Journal of Neurophysiology 108, 3253–3263.

Selnes, O.A., Gordon, B., 1999. Neuropsychologic testing. In: Evans, R.W. (Ed.), Diagnostic Testing in Neurology. W.B. Saunders, Philadelphia, pp. 151–167.

Senn, S., 1999. Cross-over trials. In: Clinical Research. John Wiley & Sons, New York.

Shao, Z., Roelofs, A., Meyer, A.S., 2012. Sources of individual differences in the speed of naming objects and actions: the contribution of executive control. Quarterly Journal of Experimental Psychology (Hove) 65, 1927–1944.

Sherwood, C.C., Broadfield, D.C., Holloway, R.L., Gannon, P.J., Hof, P.R., 2003. Variability of Broca's area homologue in African great apes: implications for language evolution. The Anatomical Record. Part A, Discoveries in Molecular, Cellular, and Evolutionary Biology 271, 276–285.

Siebner, H.R., Lang, N., Rizzo, V., Nitsche, M.A., Paulus, W., Lemon, R.N., Rothwell, J.C., 2004. Preconditioning of low-frequency repetitive transcranial magnetic stimulation with transcranial direct current stimulation: evidence for homeostatic plasticity in the human motor cortex. Journal of Neuroscience 24, 3379–3385.

Siebner, H.R., Ziemann, U., 2010. Rippling the cortex with high-frequency (>100 Hz) alternating current stimulation. The Journal of Physiology 588, 4851–4852.

Small, S.L., Holland, A.L., Hart Jr., J., Forbes, M.M., Gordon, B., 1995. Response variability in naming: a computational study. Clinical Aphasiology 23, 25–38.

Sontam, V., Christman, S.D., Jasper, J.D., 2009. Individual differences in semantic switching flexibility: effects of handedness. Journal of the International Neuropsychological Society 15, 1023–1027.

Stagg, C.J., Best, J.G., Stephenson, M.C., O'shea, J., Wylezinska, M., Kincses, Z.T., Morris, P.G., Matthews, P.M., Johansen-Berg, H., 2009. Polarity-sensitive modulation of cortical neurotransmitters by transcranial stimulation. Journal of Neuroscience 29, 5202–5206.

Stagg, C.J., Nitsche, M.A., 2011. Physiological basis of transcranial direct current stimulation. Neuroscientist 17, 37–53.

Staiger, A., Finger-Berg, W., Aichert, I., Ziegler, W., 2012. Error variability in apraxia of speech: a matter of controversy. Journal of Speech Language and Hearing Research 55, S1544–S1561.

Sylantyev, S., Savtchenko, L.P., Ermolyuk, Y., Michaluk, P., Rusakov, D.A., 2013. Spike-driven glutamate electrodiffusion triggers synaptic potentiation via a homer-dependent mGluR-NMDAR link. Neuron 77, 528–541.

Sylantyev, S., Savtchenko, L.P., Niu, Y.P., Ivanov, A.I., Jensen, T.P., Kullmann, D.M., Xiao, M.Y., Rusakov, D.A., 2008. Electric fields due to synaptic currents sharpen excitatory transmission. Science 319, 1845–1849.

Szenkovits, G., Peelle, J.E., Norris, D., Davis, M.H., 2012. Individual differences in premotor and motor recruitment during speech perception. Neuropsychologia 50, 1380–1392.

Tanaka, T., Takano, Y., Tanaka, S., Hironaka, N., Kobayashi, K., Hanakawa, T., Watanabe, K., Honda, M., 2013. Transcranial direct-current stimulation increases extracellular dopamine levels in the rat striatum. Frontiers in Systems Neuroscience 7, 6.

Teo, F., Hoy, K.E., Daskalakis, Z., Fitzgerald, P.B., 2011. Investigating the role of current strength in tdcs modulation of working memory performance in healthy controls. Frontiers in Psychiatry 2, 45.

Thirugnanasambandam, N., Sparing, R., Dafotakis, M., Meister, I.G., Paulus, W., NITSCHE, M.A., FINK, G.R., 2011. Isometric contraction interferes with transcranial direct current stimulation (tDCS) induced plasticity: evidence of state-dependent neuromodulation in human motor cortex. Restorative Neurology and Neuroscience 29, 311–320.

Tse, P.U., 2013. The Neural Basis of Free Will: Criterial Causation. MIT Press, Cambridge.

Uttal, W.R., 2013. Reliability in Cognitive Neuroscience: A Meta-Meta-Analysis. MIT Press, Cambridge.

Van Essen, D.C., Lewis, J.W., Drury, H.A., Hadjikhani, N., Tootell, R.B., Bakircioglu, M., Miller, M.I., 2001. Mapping visual cortex in monkeys and humans using surface-based atlases. Vision Research 41, 1359–1378.

Van Essen, D.C., Newsome, W.T., Maunsell, J.H., 1984. The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. Vision Research 24, 429–448.

Vandermorris, S., Murphy, K.J., Troyer, A.K., 2013. Age-related elevations in intra-individual variability on associative memory tasks. Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition 20 (6), 722–734.

Vannorsdall, T.D., Maroof, D.A., Gordon, B., Schretlen, D.J., 2012a. Ideational fluency as a domain of human cognition. Neuropsychology 26, 400–405.

Vannorsdall, T.D., Schretlen, D.J., Andrejczuk, M., Ledoux, K., Bosley, L.V., Weaver, J.R., Skolasky, R.L., Gordon, B., 2012b. Altering automatic verbal processes with transcranial direct current stimulation. Frontiers in Psychiatry/Frontiers Research Foundation 3, 73.

Venkatakrishnan, A., Sandrini, M., 2012. Combining transcranial direct current stimulation and neuroimaging: novel insights in understanding neuroplasticity. Journal of Neurophysiology 107, 1–4.

Venker, C.E., Eernisse, E.R., Saffran, J.R., Weismer, S.E., 2013. Individual differences in the real-time comprehension of children with ASD. Autism Research: Official Journal of the International Society for Autism Research 6, 417–432.

Weiss, M., Lavidor, M., 2012. When less is more: evidence for a facilitative cathodal tDCS effect in attentional abilities. Journal of Cognitive Neuroscience 24, 1826–1833.

Weiss, S.A., Faber, D.S., 2010. Field effects in the CNS play functional roles. Frontiers in Neural Circuits 4, 15.

Welcome, S.E., Chiarello, C., Thompson, P.M., Sowell, E.R., 2011. Reading skill is related to individual differences in brain structure in college students. Human Brain Mapping 32, 1194–1205.

Weltman, K., Lavidor, M., 2013. Modulating lexical and semantic processing by transcranial direct current stimulation. Experimental Brain Research 226, 121–135.

Wong, P.C., Morgan-Short, K., Ettlinger, M., Zheng, J., 2012. Linking neurogenetics and individual differences in language learning: the dopamine hypothesis. Cortex 48, 1091–1102.

Yao, L., Mccaig, C.D., Zhao, M., 2009. Electrical signals polarize neuronal organelles, direct neuron migration, and orient cell division. Hippocampus 19, 855–868.

Yap, M.J., Balota, D.A., Sibley, D.E., Ratcliff, R., 2012. Individual differences in visual word recognition: insights from the English Lexicon Project. The Journal of Experimental Psychology: Human Perception and Performance 38, 53–79.

Ye, Z., Zhou, X., 2009. Executive control in language processing. Neuroscience & Biobehavioral Reviews 33, 1168–1177.

Yong, E., 2012. Replication studies: bad copy. Nature 485, 298–300.

You, D.S., Kim, D.Y., Chun, M.H., Jung, S.E., Park, S.J., 2011. Cathodal transcranial direct current stimulation of the right Wernicke's area improves comprehension in subacute stroke patients. Brain and Language 119, 1–5.

Zaehle, T., Beretta, M., Jancke, L., Herrmann, C.S., Sandmann, P., 2011. Excitability changes induced in the human auditory cortex by transcranial direct current stimulation: direct electrophysiological evidence. Experimental Brain Research 215, 135–140.

Zaehle, T., Rach, S., Herrmann, C.S., 2010. Transcranial alternating current stimulation enhances individual alpha activity in human EEG. PLoS One 5, e13766.

Zheng, X., Alsop, D.C., Schlaug, G., 2011. Effects of transcranial direct current stimulation (tDCS) on human regional cerebral blood flow. Neuroimage 58, 26–33.

Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts

Neil K Aggarwal, New York State Psychiatric Institute, Columbia University Medical Center, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.S. Wallerstein, volume 23, pp. 15851–15856, © 2001, Elsevier Ltd.

Abstract

Transference and countertransference constitute foundational principles in psychoanalytic and psychodynamic thoughts. Freud first conceptualized transference as the patient 'transferring,' onto the analyst, reactions to major figures from early life, primarily parents. This was initially seen as an obstacle to the therapeutic work of recovering repressed pathogenic experiences of childhood into consciousness. Subsequently, Freud understood transference as essential, since it represented the reliving of repressed early experiences in the present moment and rendered such experiences accessible to interpretation. Neurotic symptomatology expressed in transference is known as transference neurosis. Psychotic symptomatology expressed in transference is known as transference psychosis. The clinician's display of neurotic transference onto the patient is known as countertransference. As with transference, countertransference was originally seen by Freud as an obstacle, but it is now regarded as an avenue to understanding the types of pressures exerted by the patient onto the analyst and an opportunity to better understand the patient. Newer traditions in psychoanalytic and psychodynamic thoughts have moved beyond models of transference and countertransference as psychic phenomena projected by individuals to consider their co-construction and cultural contexts.

Definition

Transference and countertransference constitute foundational principles in psychoanalytic and psychodynamic thoughts. Transference is the conception that our past experiences endure within us and play a formative role in our present interpersonal behaviors. In accord with his original conception that neurotic symptoms and behaviors represented earlier mental conflicts that patients had been impelled to repress from conscious memory, Freud devised the method of 'free association,' requiring the patient to verbalize every thought as it arose. This verbalization was to occur without attention to usual considerations of logical and tactful expression, i.e., no matter how seemingly nonsensical, offensive, or embarrassing. To his dismay, he found that despite the patient's conscious intent, the process of sequential narration would occasionally be blocked as the patient lapsed into shameful silence or claimed that nothing came to mind. When pressed, the patient could often acknowledge that unwanted thoughts or feelings about the analyst had come to mind, displacing the historical recounting.

Freud initially saw this as an unwelcome impediment to the joint therapeutic task of uncovering original traumatic events and tried to disengage the patient from such distractions. However, he realized the regularity of this occurrence, that it was not accidental or inadvertent, but a regular 'transferring' onto the analyst of reactions to early major figures in the patient's life, primarily parents and other significant figures from the past. In Freud's words, "They (transferences) are new editions or facsimiles of the impulses and phantasies which are aroused and made conscious during the progress of the analysis; but they have this peculiarity … that they replace some earlier person by the person of the physician. To put it another way: a whole series of psychological experiences are revived, not as belonging to the past, but as applying to the person of the physician at the present moment" (Freud, 1905/1953).

To state it simply: a feeling of anger toward the analyst, perceived in the moment as an overbearing and arbitrary authority, would be a representation of childhood interactions with the patient's father, experienced often in that light. Consequently, Freud designated transference reactions as father or mother transferences (or sibling, or mentor, or whomever), and as positive or negative transferences, depending on their emotional valences.

Freud thus reversed his reaction to transference manifestations, from seeing them as impediments to viewing them as essential. By 'analyzing' the transferences, i.e., unraveling their meanings during analysis, and tracing them as past representations of pathogenic interpersonal interactions relevant in the present, the patient's repressed traumas would be brought to light. This uncovered, their formative role in the maintenance of neurotic behaviors, would be progressively dissipated. For example, a patient's chronically damaging interactions with a work superior, perceived as arbitrary, overbearing, and impeding job satisfaction, would be progressively eased, and even eliminated as the patient interpreted them as unwitting reenactments of earlier experiences with his father.

In fact, Freud believed that because of the emotional intensity experienced in the transference interaction, compared to the remoteness of describing historical interactions in the distant past, the patient was most susceptible to analyzing the emotional impact in order to disrupt established, automatized behavioral patterns, leading to insight and therapeutic change. This effect is described in one of Freud's most famous sentences as, "For when all is said and done, it is impossible to destroy anyone *in absentia* or *in effigie*" (Freud, 1912/1958). In a well-known paper by Strachey, one of Freud's early adherents, interpretation of the transference was described as the only truly 'mutative' interpretation, all other interpretive efforts only leading up to, and supportive of, interpretation in and of the transference (Strachey, 1934). Freud's technical advice from the start was to allow analytic transferences to unfold gradually, as they inevitably would, and only to interpret them when they

crystallized in therapy as resistances to ongoing work. Such resistances could manifest either as negative transferences of hateful feelings directed at the analyst or as erotized positive transferences of loving feelings. Freud made an exception for what he called "unobjectionable positive transferences," the benign feelings that led the patient to best cooperate with the analyst, experienced by the patient as expert, helpful, and interested in effecting therapeutic change and cure. In modern times, some analysts have separated this so-called 'unobjectionable positive transference' as 'therapeutic alliance' (Zetzel, 1956) or the 'working alliance' (Greenson, 1965). Other analytic authors have, however, doubts about this unobjectionable transference, seeing it as a defensive cover that hides other objectionable transference manifestation, which would ultimately need to be interpretively undone to complete therapy.

As transference unfolded over the course of therapy, Freud further conceived of what he called the 'regressive transference neurosis' in which the presenting neurotic symptoms and behaviors stopped in patients' outer and inner lives though neurotic components came to full expression in therapy. To Freud, a fully evolved transference neurosis, focused completely on the analyst, recapitulated what he called the 'infantile neurosis,' the set of maladaptive interactions in childhood that established the template for later neurotic illness that brought the patient to treatment. The interpretive resolution of this full-fledged regressive transference neurosis marked, for Freud and many of his followers, the central task of treatment. Treatment was only successful when the transference neurosis was fully resolved and various transference distortions were transformed into an objectively realistic relationship of the patient with the therapist.

Transference Psychosis

A related concept to transference neurosis developed in the 1950s known as transference psychosis. This concept developed along two distinct trajectories. In one trajectory, transference psychosis was an analogue to the concept of transference neurosis; just as neurotic patients display 'neurotic' transference reactions to the analyst, culminating in the full-fledged transference neurosis, so also do psychotic patients display 'psychotic' transference reactions, which can coalesce into a fully formed transference psychosis. In another conceptual trajectory, a neurotic patient, seemingly well-suited for a classical psychoanalytic treatment, could develop a psychotic transference response with loss of reality testing that maintains transference as an 'as-if' phenomenon (Wallerstein, 1967).

Understanding of Transference at the End of the Twentieth Century

Since Freud's time, our understanding of transference has been transformed significantly. For example, Gill and Hoffman (1982) squarely challenged the notion that transference manifestations consist only of the projections of the patient's past relationship patterns upon the neutral objective analyst who would be the sole arbiter of reality. Given the inherent ambiguity in all human relationships, Gill saw transference as the patient's plausible construction of his or her experience of the analyst and of the analytic interaction. Analytic work would then be understood as the co-construction of how the patient and analysis each differently perceive the same interaction, each perception being the product of the individual life experience of the perceiver, with individually different predispositions toward preferred ways of endowing meaning to the encounter. Therapeutic focus would be on the psychology of the patient, what life experience disposed him or her to construe the therapeutic interaction in the manner manifested, and to link the lived past to the present transference manifestations. The past is still present in the transference as an underlying template of interpersonal expectations, but focus has shifted to a present-day interaction in the here-and-now of the therapeutic encounter as immediately experienced by the two participants. In other words, focus has shifted from the discovery of the past emergent in the present to the interaction in the present as framed by the past.

Along with that has been the concomitant shift from transference as a purely intrapsychic phenomenon within the patient, projected onto the neutral, objective, and uninvolved analyst (what is now called one-person psychology), to transference as an interaction of two subjectivities, each bringing their own built-in predispositions to that interaction. The focus remains the patient's construal of the situation, embedded in the life experience and consequent neurotic structuring of the patient. This is now called two-person psychology, the intersubjective interaction of two participants, though focused on the difficulties of the patient. These altered conceptions of the transference have been developed further within the object-relational perspective in psychoanalysis as distinct from Freud's perspective, called now drive/structural, or put differently, two-person versus one-person psychologies (Stolorow et al., 1978 or Greenberg and Mitchell, 1983). Finally, Neil Altman (1995) has contended that culture and society appear within the therapeutic dyad in what is termed as the third-person model. In this formulation, the significance of racial, cultural, and socioeconomic class differences of the patient and analyst are analyzed as providing the content of the transference.

Analytical Techniques

These modern interpretations in ideas of transference have led to alterations in analytical technique. The focus is now on the early interpretation of transference and on seeking out implicit allusions to the transference to render them explicit. The idea of unfolding full-fledged transference neurosis as a necessary recapitulation of the infantile neurotic experience has yielded to a variety of transference manifestations in the present in unique relationship patterns, with past and present relationship experiences in constant dialectical interplay.

This entire history of transference development has been summarized in an article by Cooper (1987), in which he contrasted 'historical' and 'modernist' models of transference. In Cooper's words, "The first idea, close to Freud, is that the transference is an enactment of an earlier relationship, and the task of transference interpretation is to gain insight into

the ways that the early infantile relationships are distorting and disturbing the relationship to the analyst, a relationship which is, in turn, a model for the patient's life relationships. I shall refer to this as the *historical* model of transference, implying both that it is older and that it is based on an idea of the centrality of history. The second view regards the transference as a new experience rather than as an enactment of an old one. The purpose of transference interpretation is to bring to consciousness all aspects of this new experience including its colorings from the past. I shall refer to this as the *modernist* model of the transference, implying both that it is newer, in fact still at an early stage of evolution, and that it is based on an idea of the immediacy of experience." In the historical concept, transference is an intrapsychic phenomenon of one person, the patient; and in the modernist concept, it is an interpersonal phenomenon, involving two people.

Analysts working within three-person models have also contributed to techniques that uncover the cultural and social bases of transference. Using Altman's (1995) conception of transference as the co-constructed relationship between the patient and analyst based on race, culture, and social status, Aggarwal (2011, 2012) has contended that transference and countertransference can be seen as a 'cultural third,' in which cultural values, meanings, and expectations must be analyzed to understand what is at stake for the patient during analytic work. This perspective recognizes that patient and analyst must find a way to bridge social and cultural differences embedded in their separate life experiences to negotiate a way toward therapeutic work. One way of uncovering the cultural sources of the patient's transference reaction is to use the Cultural Formulation Interview published in Diagnostic and Statistical Manual of Mental Disorders (DSM-5) that includes questions to the patient on how his or her cultural identity relates to their presenting problems and their expectations of care from the analyst (American Psychiatric Association, 2013).

Countertransference

The counterpoint to transference is known as countertransference. Originally, countertransference was conceived by Freud as the analyst's inappropriate and nonobjective responses to the patient's transferences, stemming from unresolved issues that were elicited by the patient's transferences. This would cloud the analyst's proper evaluation of transference. Countertransference, therefore, paralleled Freud's original view of transference as an impediment to analytic work. In this conception, the remedy was constant watchfulness by the analyst, to monitor inadvertent and inappropriate responses, and to control facial expression and prevent contaminating transference display. If such monitoring proved ineffective, an additional remedy would be rigorous self-analysis or the need for the analyst to return to his or her own therapy. This view later came to be called the 'narrow' conception of the countertransference. Others, especially those working with sicker patients, widened the conception of countertransference as the totality of the analyst's emotional responses to the patient; this would include countertransference, all emotional reactions by the analyst to the patient as a whole, and to the activity of analysis itself. This wider conception was called 'totalistic,' and some analysts (Winnicott, 1949) divided countertransference into 'neurotic' and 'objective' components.

Though Freud came to understand transference as essential to therapeutic work, he failed to take the same step in considering countertransference. Only in the early 1950s did psychoanalysts, first in the UK, begin to view countertransference not as an occasionally inadvertent and inappropriate intrusion upon therapeutic work, but as inevitable and as a constant expression of the analyst's involvement with the patient in interactive treatment. This moment reflected a shift to the totalistic conception of the countertransference as the entirety of the analyst's emotional responses to the patient, whether in specific reaction to the patient's transferences, or whether as stemming from the analyst's developmental history and overall responses to the patient and to therapy between them.

In this time, it was further theorized that through careful monitoring of countertransference thoughts and feelings, the analyst could obtain useful insights into what the patient was trying to get the analyst to think or feel. These insights would lead to a wider knowledge of the patient's psychological and developmental history. This process of arousing emotional responses within the analyst has been variously called 'projective identification' (by the Kleinian school of psychoanalysis) or 'role responsiveness' (by the classical Freudian school; cf Sandler, 1976). In this conception, the perception of countertransference has likewise shifted from an initial view of an occasional and inappropriate intrusion upon therapy (to be overcome by the analyst), to an inevitable and necessary vehicle toward understanding the patient. The conceptual evolution of countertransference has therefore mirrored the conceptual evolution of transference. In the same way that content of transference has also been seen to be inherently cultural (Altman, 1995; Aggarwal, 2011), countertransference has also been shown to emerge in relation to similarities and differences in race, culture, and ethnicity with the patient that can yield essential insights into underlying psychic conflicts (Comas-Díaz and Jacobsen, 1991).

Therefore, transference and countertransference are completely equivalent, neither being just a response to the other, but each stemming from the psychological development of the respective parties to the therapeutic process. Some analysts (cf McLaughlin, 1981) have therefore even called for the elimination of the word countertransference, calling both sides of the interaction as transference and without privileging either.

Countertransference evocations tend to be more subtle and contained among psychologically mature analyst, and with classically typical neurotic patients. These evocations become more severe and difficult to contain with sicker patients, typically narcissistic or borderline in their character organization and seen with increasing frequency in psychoanalytic and psychotherapeutic practices. The intense transference–countertransference interplay results in what are often called 'role enactments' such that the analyst feels impelled to take on the role (punitive authority, benign rescuer, neglectful parent, etc.) that the patient, through conscious or unconscious speech and behavior, elicits from the analyst. By

interpreting situations in which the analyst inadvertently adopted or fell into the assigned role, the analyst can become aware of his or her countertransference involvement and use this to clarify psychic pressures from the patient. When identified, these pressures become insights into the patient's character predispositions and can be traced to the patient's developmental history.

A major issue, and a matter of professional controversy, is when the analyst can communicate to the patient's aspects of understanding countertransference. There are some who utilize countertransference as additional avenues of understanding the patient and would avoid any countertransference 'confession' to the patient to prevent self-disclosure or to spare the patient from an additional psychic load to bear. Others advocate for analysts to share their countertransference responses with the patient under certain circumstances as ways of obtaining an understanding of what kind of reaction the patient wanted to evoke in the analyst, how this reaction reflects relationships outside of therapy, and the kinds of responses that are elicited among others. There are other instances where the patient detects the countertransference response before the analyst is aware of it ("You are angry at me and taking it out on me"), and the analyst should deal with these instances truthfully and tactfully rather than to deny or avoid it.

A final point about transference and countertransference: Therapy is a highly contrived arena in which emotional reactions are elicited in an intimate and intensely meaningful that can be subjected to mutual scrutiny. Such scrutiny can enhance psychological understanding of neurotic issues (and their developmental history) in a patient seeking relief from mental and emotional difficulties. But it should be evident by now that we all live by transferences in all of our relationships. We always bring ourselves, our personalities and characters, as shaped and colored by our pasts, into our interpersonal encounters, and we see our present relationships and expectations largely through dispositions cultivated by past experiences. In fact, it can be said that our entire capacity for relationships, for empathy with others, even for creative endeavor in any arena, rests on the transferences that we bring and live by (cf Loewald, 1980).

Summary

All of the issues covered in this article are summarized in Aggarwal (2011), which explains conceptions of transference and countertransference in one-, two-, and three-person models of psychology, with subsequent recommendations for clinical practice.

See also: Freud, Sigmund (1856–1939); Interpersonal Psychotherapy; Psychoanalysis, History of; Psychoanalysis: Current Status; Therapist–Patient Relationship.

Bibliography

Aggarwal, N.K., 2011. Intersubjectivity, transference, and the cultural third. Contemporary Psychoanalysis 47, 204–223.
Aggarwal, N.K., 2012. Hybridity and intersubjectivity in the clinical encounter: impact on the cultural formulation. Transcultural Psychiatry 49, 121–139.
Altman, N., 1995. The Analyst in the Inner City: Race, Class, and Culture through a Psychoanalytic Lens. The Analytic Press, Hillsdale.
American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Press, Arlington.
Comas-Díaz, L., Jacobsen, F.M., 1991. Ethnocultural transference and countertransference in the therapeutic dyad. American Journal of Orthopsychiatry 61, 392–402.
Cooper, A.M., 1987. Changes in psychoanalytic ideas: transference interpretation. Journal of the American Psychoanalysis Association 35, 77–98.
Freud, S., 1905/1953. Fragment of an Analysis of a Case of Hysteria, standard ed., vol. 7 1–122.
Freud, S., 1912/1958. The Dynamics of Transference, standard ed., vol. 12 97–108.
Gill, M.M., Hoffman, I.Z., 1982. Analysis of Transference. In: Theory and Technique. Psychological Issues Monograph 53, vol. I. International University Press, New York.
Greenberg, J.R., Mitchell, S.A., 1983. Object Relations in Psychoanalytic Theory. Harvard University Press, Cambridge, MA, and London.
Greenson, R.R., 1965. The working alliance and the transference neurosis. Psychoanalysis Quarterly 34, 155–181.
Loewald, H.W., 1980. Papers on Psychoanalysis. Yale University Press, New Haven, CT and London.
McLaughlin, J.T., 1981. Transference, psychic reality, and countertransference. Psychoanalysis Quarterly 50, 639–664.
Sandler, J., 1976. Countertransference and role-responsiveness. International Review of Psycho-Analysis 3, 43–47.
Stolorow, R.D., Atwood, G.E., Ross, J.M., 1978. The representational world in psychoanalytic therapy. International Review of Psycho-Analysis 5, 247–256.
Strachey, J., 1934. The nature of the therapeutic action of psycho-analysis. International Journal of Psycho-Analysis 15, 127–159.
Wallerstein, R.S., 1967. Reconstruction and mastery in the transference psychosis. Journal of American Psychoanalysis Association 15, 551–583.
Wallerstein, R.S., 1995. The Talking Cures: The Psychoanalyses and the Psychotherapies. Yale University Press, New Haven, CT and London.
Winnicott, D.W., 1949. Hate in the counter-transference. International Journal of Psycho-Analysis 30, 69–74.
Zetzel, E.R., 1956. Current concepts of transference. International Journal of Psycho-Analysis 37, 369–375.

Transgender, Transexual, and Gender Variant Individuals

M Paz Galupo and JF Stuart, Towson University, Towson, MD, USA
Derek P Siegel, American University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Valentine, D. Kulick, volume 23, pp. 15888–15893, © 2001, Elsevier Ltd.

Abstract

Gender identity refers to an individual's experience or identity in relation to gender. Transgender, transsexual, and gender variant individuals (often considered under the umbrella term of transgender or trans*) have gender identities that are not congruent with the gender assigned to them at birth. Trans* identity and experiences are framed with regard to medical perspectives, social and legal perspectives, and community perspectives.

Gender identity refers to an individual's experience or identity in relation to gender. For transgender, transsexual, and gender variant individuals, gender identity is not congruent with the gender assigned at birth. In contrast, *cisgender* is a term that refers to individuals whose gender identity is congruent with the gender assigned at birth. Although the experiences of transgender, transsexual, and gender variant individuals may be distinct, they are often considered under a common umbrella term of *transgender* or *trans**. Individuals who fall under the trans* umbrella may include cross-dressers, transvestites, transsexuals, drag kings and queens, as well as individuals who identify as transwomen, transmen, bigender, two-spirit, androgynous, genderqueer, gender nonconforming, and gender variant.

Medical Perspectives

Current understandings of gender identity have been greatly impacted by the pervasive medical framework that has sought to define trans* experience. Although there is great variation in how trans* individuals relate to gender identity terminology, the term transsexual has been particularly impacted by its use in a medical context. The psychiatric literature in particular distinguishes between transvestites and transsexuals. *Transvestism* is characterized by a focus on clothing of the opposite gender. In contrast, the focus for *transsexualism* is on gender identity, where an individual's gender identity is not congruent with the gender assigned at birth. Thus, the medical literature focuses on individuals who transition (or are in the processing of transitioning) their bodies to better match their gender identity. Medical perspectives on trans* experience are focused on providing treatment, often in the form of *gender reassignment surgery* (GRS) or other gender affirming treatments inclusive of hormone therapy, surgical alteration of secondary sex characteristics, or otherwise altering physical appearance or presentation to better match the target gender.

The focus on medical transition of transsexual individuals creates a focus on the body (specifically the genitals) that obscures the fact that many trans* individuals do not seek medical transition (Grant et al., 2011; Bockting, 2014). For example, for many gender nonconforming trans* individuals, there may be no transition process at all. Alternatively, transition may describe their journey from the gender they were assigned at birth to their current gender identity (which may or may not be male or female). Trans* individuals may not medically transition for various reasons. Some trans* individuals do not feel that their body needs to be altered in order to socially transition while some trans* individuals do not see their bodies as incongruent with their gender identity. Many other trans* individuals do not medically transition because they do not have access to the financial resources, insurance, or social support network necessary to facilitate such a transition. In addition, while some individuals view their trans* experience as an important part of their identity, others view their trans* experience as a status or a medical history. There is great variability with regard to disclosure as not every individual who is trans* feels the need or desire to disclosure their trans* identity or status (Zimman, 2009).

Diagnostic Categories: Gender Identity Disorder and Gender Dysphoria

The formal diagnoses and treatment of trans* individuals in the medical establishment came in 1980 with the dual emergence of diagnostic categories of Gender identity disorder (GID) and the establishment of the Standards of Care (SOC) for the Health of Transsexual, Transgender, and Gender Nonconforming People. GID is conceptualized as the experience of significant discontent with the sex assigned at birth and/or with the gender roles associated with that sex. Formal diagnoses of GID became possible with the publication of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders III (DSM-III) in 1980 and its associated inclusion in the International Classification of Diseases (ICD-10).

Initial diagnoses of GID rely heavily upon an individual's cross-sex identification, rejection of sex-'appropriate' behavior, and dissatisfaction with the body. With the 2013 reclassification of GID to gender dysphoria in the DSM-5, modifications were made to address earlier criticisms. In particular, diagnostic criteria for gender dysphoria focuses less on gender identity, and instead centers on the experience of distress. This allows for an end point for the diagnoses. In addition, diagnostic criterion in the DSM-5 shifts the language from sex to gender in order to encompass individuals who are intersex (have biological characteristics at birth of both males and female) and those whose have gender nonconforming trans* identities.

Standards of Care

The World Professional Association for Transgender Health, Inc. (WPATH), formerly the Harry Benjamin International Gender Dysphoria Association, Inc., is a professional organization devoted to the understanding and treatment of gender dysphoria. Originally founded in 1979, the organization was named after Dr Harry Benjamin, one of the first physicians in the United States to begin work with transsexuals. In 2007, the organization changed its name to WPATH. WPATH is best known for publishing the SOC for the Health of Transsexual, Transgender, and Gender Nonconforming People. Prior to the publication of the SOC there was no medical consensus for procedures or approaches in providing safe, legal medical options for trans* individuals. The SOC, first drafted in 1979 has been periodically revised (1980, 1981, 1990, 1998, 2001, and 2011). The SOC established a process and timeline of therapy before medical treatment for trans* individuals. The original SOC recommended that prior to GRS trans* individuals should obtain recommendations ('letters') from at least two psychologists/psychiatrists, should be established on hormonal treatment for at least 6 months, and have lived exclusively in the role of the chosen gender for at least a year. In addition, these guidelines suggested that a 'good candidate' for GRS included having a lack of sexual activity and no history of prostitution.

There are a number of criticisms of the medical establishment's framing of trans* experience. Through the diagnostic categories and the SOC, trans* individuals' gender identity and presentation is scrutinized and regulated by medical professionals. In addition, trans* individuals are placed in a position to 'prove' their gender identity in order to receive treatment (Serano, 2007; Stryker, 2008). Medical providers, then, often serve as 'gatekeepers' to medical and gender affirming treatments. Some trans* individuals advocate for the declassification of gender dysphoria as a psychiatric category and criticize the way in which the medical establishment pathologizes trans* individuals and reinforces false notions of sex and gender as binary. In addition, the prior classification of GID under Sexual Dysfunctions and Paraphilic Disorders in the DSM has been scrutinized as contributing to the conflation of sexual orientation and gender identity.

The current change from GID to gender dysphoria in the DSM-5 and the publication of the most recent version of the SOC signals a shift toward providing comprehensive trans* care that is more inclusive of gender nonconforming identities and nonmedically transitioning trans* identities (Coleman et al., 2012). These changes reflect the larger shift within the medical and psychological communities to expand trans* care and research. Although these changes are significant, the authority of the medical community to define trans* bodies and identities continues to serve to pathologize experience within larger society.

Social and Legal Perspectives

Despite greater medical and social recognition of trans* identities, individuals who are (or are perceived as) trans* continue to face discrimination, harassment, and bias-motivated violence in many areas of their lives due to the irrational fear and/or hatred and negative cultural ideology embedded in *transphobia*. Violent behavior and hostility is estimated to be four times higher for trans* individuals when compared with the national average; 26% of respondents in the National Transgender Discrimination Survey (Grant et al., 2011) reported having been physically assaulted on at least one occasion because of anti-trans bias. Trans* women (transition from male to female) are disproportionately affected, with half of lethal anti-lesbian, gay, bisexual (LGB), and transgender hate crimes committed against trans* women and girls. At the same time, 50% of transgender people report they are uncomfortable seeking police assistance (Grant et al., 2011).

This negative climate fosters hate crimes against trans* people and contributes to their 'invisibility.' In a society invested in gender dimorphism, the emotional expression of *genderism* (the ideology that there are, and should be, only two genders that are inevitably tied to the gender assigned at birth) and the violent manifestation of these emotions occurs through discriminatory behavior in the form of *gender bashing*. *Such gender bashing* perpetuates the ways that gender operates as a vector of regulation and distribution, describing a coercive system that orders the world through gender-based violence (Beemyn and Rankin, 2011).

In addition to high rates of violence and imprisonment, trans* individuals' experience severe economic insecurity and exclusion from social services. They are more likely to experience hostile or aggressive familial interaction, more likely to be kicked out of their homes by parents, more likely to become homeless or live below the poverty line, and are less likely to be employed. Ninety-seven percent of trans* individuals reported having experienced some form of workplace harassment or discrimination on the basis of gender identity or expression, with more than one in four trans* adults having lost at least one job due to bias (Grant et al., 2011). Trans* people of color, trans* youth, trans* people with disabilities, homeless, and undocumented trans* populations are particularly vulnerable. These compounding factors have resulted in limited health care access, familial aid, and other resources, further increasing the risk of harassment, discrimination, and violence.

Legal Recognition and Protection

There has been no broad law or legal strategy that offers full protection for members of the trans* community, but rather protections have been made in a more piecemeal approach that has left trans* individuals legally vulnerable. Historically both disability protections and sex discrimination laws have been utilized in select cases involving gender identity, although there has been criticism by the medical community that marking transgenderism as a disability only further stigmatizes the population, while still specifically excluding gender variant individuals who do not fit the Americans with Disabilities Act guidelines (ACLU, 2013). Regarding the limitations of law reform as intervention, trans* scholar, Dean Spade (2011) explains that a key challenge facing trans* politics is how to properly account for the uneven distribution of vulnerability and violence across trans* populations due to varying levels of vulnerability within the community at large. Policy language is a vital detail in providing protections. Each policy must be

broad enough to cover all the variation within the population but specific enough to provide actual protection. Currently, the framework of studying, describing, and addressing the discrimination faced by trans* persons through the single vector of transphobia often causes the erasure of the reality of conditions facing those most vulnerable, such as trans* people of color or gender variant individuals (Spade, 2011).

Advocates of trans* equality have primarily pursued two law reform interventions: antidiscrimination laws that list gender identity as a category of nondiscrimination, and hate crime laws that include crimes motivated by the gender identity of the victim as triggering the application of a jurisdiction's hate crime statute. In 2009, the *Matthew Shepard and James Byrd, Jr.* Hate Crimes Prevention Act became the first and only federal hate crime law to include gender identity; however, at the time of this writing, there was still legislative debate regarding the inclusion of gender identity in the Employment Non-Discrimination Act (Transgender Law and Policy Institute, 2013). The passage of federal antidiscrimination laws would create a nation-wide basis for legal claims against discriminating employers, housing providers, restaurants, hotels, stores, and the like. Despite 13 states and 108 counties/cities having such laws, the National Gay and Lesbian Task Force (2009) analysis of jurisdictions with explicitly trans*-inclusive nondiscrimination laws posits that only 39% of the population is actually protected.

Gender identity is the most common phrase used for policies or laws that protect trans* rights, conceivably for its power of immutability (GenderPAC, 2007). Most nondiscrimination laws do not explicitly include gender identity, but prohibit discrimination based on perceived, as well as actual, sexual orientation, limiting a trans* individual's protection to the perpetrator's belief (accurate or not) that the victim was gay. While trans*-inclusive nondiscrimination laws can offer legal protections based on gender identity/expression, sexual orientation, sex, disability, and citizenship, accurate and up-to-date documentation like necessary identity documents (i.e., driver's license, passport) and records (i.e., birth certificate, social security) is usually a prerequisite for access to education, jobs, community services, and safe spaces (ACLU, 2013). According the National Survey (2009), only one-fifth of trans* individuals are actually able to update all of their legal credentials to match their current gender identity.

Trans* Case Law

Although federal laws banning sex discrimination (Title VII of the 1964 Civil Rights Act; CRA) do not prohibit gender identity discrimination, federal courts that have considered the issue more recently have found some protections in the CRA for trans*-specific cases (ACLU, 2013). The landmark case of *Smith v. City of Salem* held that a trans* firefighter suspended from his job after complaints that his appearance was not 'masculine enough' had an actionable claim for sex stereotyping based on his gender nonconforming conduct and his gender identity. Smith's original allegation that his suspension was 'based on sex' was dismissed at both the state and federal levels, ruling that Title VII prohibition of sex discrimination does not include gender identity (TL and PI, 2013). In his appeal, Smith relied on the US Supreme Court's decision in *Price Waterhouse v. Hopkins* regarding a woman's 'overly masculine' dress/appearance to support his claim that discrimination based on sex stereotyping warrants Title VII protection, successfully drawing an analogy with Smith's claim to the *Price Waterhouse* claim based on sex stereotyping (TL and PI, 2013).

Future of Trans* Rights

Even with the increasing number of states adopting more inclusive antidiscrimination laws, including the legal recognition of same-sex marriage, the legality of these contracts relies on the definable and legally recognized sex. Additional current trans* discrimination topics being addressed at the grassroots level include trans* participation in collegiate and professional sports, trans* protections in the military, the use of public spaces like homeless shelters and gender-specific resource centers, gendered prison assignment, adoption/child custody rights, among others. However, legislation has yet to officially tackle many of these issues, because the failure to cohesively articulate oppression and discrimination results in the inability to claim retribution, since one must be able to name the act of harm/exclusion in order to seek legal justice (Lair, 2013). While all these topics are considerably important to the growth of trans* inclusion and recognition, determining the most pressing issues is contingent upon which intersection is being challenged within and outside of the trans* community.

Community Perspectives

Creating a Modern Transgender Movement

Until the introduction of a conscious 'transgender politic' in the 1990s, trans* communities in the United States historically organized around issues rather than shared identity.

'Transvestite' groups from the 1950s, for example, catered exclusively to heterosexual cross-dressing males by supporting homophobic and antitranssexual policies. These early communities did not see themselves as marginalized. In fact, they sought acceptance for the clothes that they wore – attempting to expand dress norms for 'proper' men – rather than for the act of cross-dressing itself. Isolating these experiences from a deeper examination of self, participation in these groups provided the opportunity to explore gender expression without threatening one's reputation (Stryker, 2008). From elite cross-dressing groups to lobbyists on Capitol Hill, there has never been a single transgender community with a cohesive set of goals. The transgender movement has developed its contemporary structure, rather, as particular communities address their needs within changing political climates.

While issue-based advocacy in the 1950s had little to do with identity, the 1960s saw coalitions of marginalized people organizing around shared discrimination. In the 1969 Stonewall Riots, often considered the origin of the 'homophile' or gay rights movement, transgender women joined with drag queens ('transvestites'), gay men, and lesbians in a violent reaction to routine police harassment. Three years earlier, economically segregated transgender women in the Tenderloin District of San Francisco rioted at Compton's Cafeteria. They

allied themselves with communities of color to protest harmful urban land-use and domestic war policies (Stryker, 2008). Although these riots did not create immediate structural change, they succeeded in altering broader perceptions of transgender citizenship. Through the shared political interests underlying these protests, transgender communities – such as 'Vanguard,' a street organization for gay and trans* youth in Tenderloin (an historically economically depressed neighborhood in San Francisco) – began to take shape.

These emerging transgender communities, however, complicated the identity politics characteristic of the homophile and feminist movements in the 1970s and 1980s and there was tension among these different movements from the outset. The homophile movement's assimilationist strategies were particularly vulnerable to transphobia through the act of 'gender baiting,' by which LGB activists sought to differentiate themselves from more subversive gender presentations among transgender individuals (Feinberg, 1996). Likewise, transgender identities challenged feminist ideology based on a monolithic womanhood set apart from anything masculine. Transgender women were explicitly excluded from the yearly Michigan Womyn's Festival and subject to belligerently transphobic rhetoric from many feminist scholars. Facing antagonism from all ranges of the political spectrum, transgender communities through the 1980s tended to be small and locally funded (Stryker, 2008). In New York, Stonewall veterans Sylvia Rivera and Marsha P. Johnson founded Street Transvestite Action Revolutionaries, while the Queens' Liberation Front organized political demonstrations and the popular publication *Drag*.

Transgender subgroups individually developed their own political identities, not converging in a structured fashion until the 1990s. Unlike male-to-female (trans* women) communities, who had been openly identified and active for the past two decades, transgender men only began to organize in the 1970s. Historians attribute this time lapse to female-to-males' ability to present themselves more convincingly as young men, therefore not needing to ban together for mutual support in the same way as transgender women (Stryker, 2008). The line between transgender friendships and organizations are not always clear, however, meaning that many communities serve both social and political functions. Just as each transgender population has its own political goals, these communities also serve different social functions. They can be characterized generally as networks that provide shared experiences, tools to reduce social stigma, and information lacking within traditional resources (Hines, 2007).

Uniting disparate communities under the political identity 'transgender' and merging – at least in name – with LGB activists, the contemporary transgender movement surfaced in the 1990s. The first indication that local, issue-based advocacy would give way to a national network of organizations was the acquired immunodeficiency syndrome (AIDS) crisis. Responding to AIDS brought together a wide coalition of communities affected by the disease, as well as offering financial resources to transgender individuals who were seen as particularly vulnerable to contracting the human immunodeficiency virus. AIDS activism relied on 'queer politics,' which gave birth to the academic discipline of Queer Theory. Queer, a reclaimed slur that embodies any challenge to traditionally oppressive sexual and gender norms, provided the language and tools for transgender communities to articulate their own political interests. Although transgender issues were marginalized in the Queer Nation movement, activists formed a suborganization called Transgender Nation that pressured LGB groups to promote transgender inclusion. Queer theory created new language to discuss gender, helping to spark a movement of nonbinary trans* identities (Stryker, 2008), and the arrival of the Internet was a means to connect transgender voices and experiences from across the country (Beemyn and Rankin, 2011).

Trans* Identity Politics

Individual trans* activists and communities' political goals correspond profoundly with how they draw the boundaries of 'transgender' (Davidson, 2007); in other words, the process of deciding who belongs under the trans* umbrella further implies whose interests are being advocated. The question of defining trans* identity politics owes its roots to the history of inclusion and exclusion within broader LGB and feminist movements. One school of thought promotes a borderless trans* community that avoids issues of hierarchy that have plagued other social movements, perpetuating inequalities among subgroups. Feminists, for instance, clashed over who could legitimately identify as women and therefore 'deserved' access to movement resources – often marginalizing transwomen, lesbians, women of color, and working-class women (Serano, 2007). Kate Bornstein introduces the equalizing concept of 'gender outlaws' who violate the rules of gender, implying that there is an essential fluidity that characterizes trans* identities (1994). Julia Serano, however, sharply critiques this 'antiidentity movement' that ignores the lived differences between various populations. She prefers to imagine transgender as a "set of alliances rather than monolithic communities" (2007: p. 142). The umbrella model of trans* identity can fail to express the ideological and political tensions between subcommunities who may disagree as to who belongs within the broader coalition.

A great point of controversy between transsexual and gender variant communities, for example, is the degree to which abolishing the gender binary is seen as a fundamentally transgender issue. Routine academic critiques of sex and gender from both feminist and queer theorists have been described as undermining transsexual identities, just as the political movement to remove gender dysphoria from the DSM often ignores the medical and legal realities of physical transition (Davidson, 2007: p. 68). Despite the architecture of umbrella politics, several hierarchies have emerged that reflect ideological boundary drawing. There are tropes that stereotype transsexuals as 'assimilationists' for typically presenting themselves within the binary. Such 'subversivism' can rank those who present as gender variant as somehow superior to more traditional presentations (Serano, 2007). On the other hand, genderqueer and nonbinary communities are often assumed to be white, young, educated, and socioeconomically privileged.

Boundary drawing reflects power inequalities in that more powerful subgroups are able to set political agendas that either include or exclude particular populations. Intersex communities are often included under the trans* umbrella, usually without consultation, which in itself reveals power disparities. Because intersex bodies disrupt the sex binary between male

and female, activists describe this co-optation as "trans people using intersex to validate [their own identities]" (Davidson, 2007). A more implicit type of boundary drawing occurs when individuals adopt 'mainstreaming' or assimilatory politics with the goal of their transgender identity/status no longer being an issue. Most sociologists agree that it is easiest for those with social privilege to reject identity as an organizing principle (Davidson, 2007). In other words, those who want to mainstream transgender experiences are not people who also experience racism, classism, or other forms of inequality. By not seeing how trans*-related discrimination intersects with these systems of inequality, privileged trans* folks promote a political agenda that does not address the oppression of less privileged individuals.

Debate over the role of cisgender allies in the modern trans* movement can also reflect power inequalities. Leslie Feinberg's strategy for allyship – the recognition that "each individual has a stake in trans liberation" (1996: p. 92) – does not always account for cisgender privilege. Transgender rights articulated within a broader gender rights movement often invoke unequal power dynamics, due to the fact that critical examinations of 'gender' usually address binary identities while marginalizing trans* experiences. When GenderPAC, famous for popularizing genderqueer identities, changed its mission in 2000, activists observed how this reduced the work GenderPAC did in trans* communities (Davidson, 2007). When cisgender and trans* individuals face different political realities, failure to recognize privilege perpetuates inequalities.

Julia Serano's vision of allyship more explicitly addresses issues of privilege between trans* subpopulations and cisgender allies. She pivots away from Feinberg's strategy by advocating for a trans* movement that would allow boundaries to grow naturally but focuses primarily on building alliances among various communities (2007). This vision of trans* identity politics acknowledges the intersection between gender rights and transgender issues. It also implies that effective allyship must include the mutual recognition of our distinct privileges and lived experiences. While we each have a stake in issues such as employment, housing, violence, and health care, allyship recognizes that these individual issues affect each trans* subpopulation and cisgender allies differently. Contemporary organizations such as the National Center for Transgender Equality and the National Transgender Advocacy Coalition are structured as coalitions, representing the trans* movement as Serano suggests: "a set of alliances rather than monolithic communities." This ideological return to issue-based politics draws the boundaries of transgender identity quite extensively while also respecting the importance of individualized identity.

See also: Feminist Theory: Radical Lesbian; Gay and Lesbian Movements; Gay, Lesbian, and Bisexual Youth; Gender History; Gender, Economics of; Identity Movements; Queer Anthropology; Queer Theory and Intersectionality; Sexual Orientation: Biological Influences; Sexual Orientation: Historical and Social Construction; Sexuality and Social Work; Sexuality, Theories of.

Bibliography

American Civil Liberties Union, 2013. Know Your Rights: Transgender People and the Law. LGBT Basic Rights and Liberties. https://www.aclu.org/translaw.

Beemyn, G., Rankin, S., 2011. The Lives of Transgender People. Columbia University Press, New York, NY.

Bockting, W., 2014. Transgender identity development. In: APA Handbook of Sexuality and Psychology, vol. 1. American Psychological Association, Washington, DC, pp. 739–758.

Bradley, O., 2010. They Need Labels: Contemporary Institutional and Popular Frameworks for Gender Variance. In: Anthropology Theses. Paper, vol. 35. Georgia State University, 81–90.

Coleman, E., Bockting, W., Botzer, M., Cohen-Kettenis, P., DeCuypere, G., Feldman, J., et al., 2012. Standards of care for the health of transsexual, transgender, and gender nonconforming people, 7th version. International Journal of Transgenderism 13 (4), 165–232.

Davidson, M., 2007. Seeking refuge under the umbrella: inclusion, exclusion, and organizing within the category transgender. Sexuality Research & Social Policy 4 (4), 60–80.

Feinberg, L., 1996. Transgender Warriors: Making History from Joan of Arc to Dennis Rodman. Beacon Press, Boston, MA.

GenderPAC, 2007. "Definition of Terms." Gender Public Advocacy Coalition Work – Place Fairness n.d. http://www.gpac.org/workplace/terms.html.

Grant, J.M., Mottet, L.A., Tanis, J., Harrison, J., Herman, J.L., Keisling, M., 2011. Injustice at Every Turn: A Report of the National Transgender Discrimination Survey. National Center for Transgender Equality and National Gay and Lesbian Task Force, Washington, DC. http://endtransdiscrimination.org/PDFs/NTDS_Report.pdf.

Hines, S., 2007. TransForming Gender: Transgender Practices of Identity, Intimacy, and Care. Policy Press, Bristol, UK.

Lair, L., 2013. Compulsory Cisgendering and the Denial of Trans* Experience. Paper Presented at National Women's Studies Conference. Cincinatti, OH.

Lev, A.I., 2004. Transgender Emergence: Therapeutic Guidelines for Working with Gender-Variant Individuals and Their Families. Haworth Press, New York.

Serano, J., 2007. Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity. Seal Press, Emeryville, CA.

Spade, D., 2011. Normal Life: Administrative Violence, Critical Trans Politics, and the Law. South End Press, New York.

Stryker, S., 2008. Transgender History. Seal Press, Berkeley, CA.

Transgender Law and Policy Institute, (TL and PI) 2013. Litigation: Case Law. http://www.transgenderlaw.org/college/index.htm.

Zimman, L., 2009. 'The other kind of coming out': transgender people and the coming out narrative genre. Gender and Language 3 (1), 53–80. http://dx.doi.org/10.1558/genl.v3il.533il.53.

Relevant Websites

http://www.gpac.org – GenderPAC.
http://transequality.org/ – National Center for Transgender Equality (NCTE).
http://www.ntac.org – National Transgender Advocacy Coalition (NTAC).
http://www.transgenderlaw.org – Transgender Law & Policy Institute.

Transient Epileptic Amnesia

Serge Hoefeijzers, University of Edinburgh, Edinburgh, UK
Adam Zeman, University of Exeter Medical School, Exeter, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Transient epileptic amnesia (TEA) is characterized by brief, recurrent episodes of transient amnesia occurring as a result of epilepsy. During these episodes, declarative memory is impaired while other cognitive functions remain intact. TEA is a syndrome of mesial temporal lobe epilepsy (TLE), which typically affects middle-aged people, particularly men. The amnesic attacks often occur on waking and can be accompanied by other symptoms of TLE such as automatisms, olfactory hallucinations, and short periods of unresponsiveness. EEG may reveal interictal epileptiform activity arising from the temporal lobes. Gross structural abnormalities are rarely observed, but group studies reveal subtle volume loss in hippocampal, perirhinal, and orbitofrontal cortices. The amnesic seizures respond very well to anticonvulsant medication, though most patients (around 80%) report interictal memory problems such as accelerated long-term forgetting (ALF), autobiographical, and topographical memory loss. These often persist despite successful treatment of the epilepsy. ALF, which is reported by around 50% of patients with TEA, involves the rapid forgetting of newly acquired information, occurs within 24 h of acquisition, and is likely to reflect a deficit in early memory consolidation or subtle impairments in memory acquisition. The autobiographical memory loss, seen in around two-thirds of patients with TEA, appears to be due to a degradation of established memories, though impairments of acquisition, consolidation, and retrieval may also play a part. These interictal memory difficulties are likely to result from either subtle structural damage in the medial temporal lobes or physiological disruption of memory processing by subclinical epileptiform activity.

Introduction

Transient epileptic amnesia (TEA, a term coined by Kapur, 1990) is a form of temporal lobe epilepsy (TLE) characterized by recurrent episodes of transient amnesia. During these episodes, the sufferer is unable to form new memories (anterograde amnesia) and/or unable to retrieve established memories (retrograde amnesia). These brief (usually less than 1 h) attacks of amnesia occur in middle-aged or elderly people and are sometimes the sole manifestation of epilepsy (in around one of three cases). Besides these ictal features, TEA is associated with two striking forms of persistent interictal memory impairment, namely accelerated long-term forgetting (ALF) and remote memory deficits (especially affecting autobiographical and topographical memory). ALF is the accelerated loss of newly acquired information that usually takes place over subsequent days to weeks (Butler et al., 2007; Zeman and Butler, 2010). This chapter discusses the background of the syndrome of TEA and reviews the ictal and interictal features of this unusual form of epilepsy of which the main manifestation is amnesia.

Clinical Features and Demography

TEA typically begins in late-middle to old age (mean age at onset is around 60 years) and is predominantly found among men (male:female ratio is 2:1) (Butler et al., 2007). The amnesic episodes usually occur around once a month. The exact number of attacks is likely to be higher as episodes sometimes go unnoticed (if the patient has no recollection of them and the attack occurs unwitnessed).

Amnesia is the sole manifestation of TEA in around one-third of patients. In other cases, the amnesic event is associated with other more familiar manifestations of epilepsy like olfactory hallucinations, automatisms such as lip smacking and chewing, or a brief period of unresponsiveness. Patients occasionally report other types of seizure, usually complex partial seizures, in which amnesia is not the most prominent feature. Tonic–clonic seizures occur rarely. Attacks often occur upon waking (around 70% have at least some of their attacks on waking) and typically last 30 min to 1 h, though, shorter (several minutes) and longer (several hours) episodes also occur. During the attack, 50% of patients ask repetitive questions such as 'Where am I?,' 'What day is it?' These questions reflect both an element of retrograde amnesia (patients are unable to access information laid down shortly before the episode which would allow them to orient themselves) and an element of anterograde amnesia, as they are often unable to remember the answers given to their questions. Anterograde amnesia is often partial: after the attack, in 44% of the cases, patients 'remember not being able to remember.' Apart from amnesia, cognitive functioning remains largely or wholly intact (Butler et al., 2007; Zeman and Butler, 2010).

The medial temporal lobes (MTLs) are likely to be the seizure source. Indeed, bilateral seizure activity in the temporal lobes is the most common ictal EEG-abnormality (detected in 8 out of 10 cases monitored during an attack). Interictal epileptiform abnormalities (where present) have been observed both bilaterally (56%) and unilaterally (32% left-sided, 12% right-sided) (Butler and Zeman, 2008b). Major structural abnormalities are rarely detected. However, mild bilateral hippocampal, perirhinal, and orbitofrontal atrophy has been detected in TEA patients as a group in two recent studies (Butler et al., 2009, 2013).

The amnesic episodes of TEA respond very well to anticonvulsant medication, with cessation of attacks on a modest dose of a single anticonvulsant drug in almost all cases. Despite this

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51067-8

clear-cut response to low doses of antiepileptic drugs (AEDs), most patients (around 80%) report interictal memory problems that persist to some degree after successful treatment of the attacks: 70% report a patchy loss of remote memories of salient autobiographical events such as weddings or holidays (i.e., autobiographical amnesia); 44% of the patients complain that newly acquired information fades abnormally rapidly from memory over days to weeks: this phenomenon is referred to as ALF. Moreover, 36% have difficulties in finding their way around areas they used to be very familiar with (i.e., topographical amnesia) (Butler et al., 2007; Zeman and Butler, 2010). These persistent interictal memory impairments are discussed in more detail further below.

The characteristics of the syndrome (sudden and late onset) and nature of the ictal and interictal memory complaints sometimes make patients concerned that they are in the early stages of dementia. However, despite raising understandable doubts in patients' minds about their memory performance, TEA does not elevate levels of anxiety or depression markedly (Butler and Zeman, 2008b).

Diagnostic Criteria for TEA and Differential Diagnosis from Other Transient Amnesic Syndromes

Episodes of transient amnesia, in which the patient experiences an acute onset of dense memory loss, are striking clinical events that present a diagnostic challenge to the physician. TEA is by no means the only cause for such episodes: they can be induced by disorders including transient global amnesia (TGA), head injuries, transient ischemic attacks, drug intoxication (e.g., alcohol, benzodiazepines), migraine, hypoglycemia, psychogenic fugue, and electroconvulsive therapy (Bartsch and Butler, 2013). In TEA, acute amnesia is, in one-third of cases, the sole 'visible' manifestation of the seizure. Not surprisingly, TEA can be misdiagnosed for other syndromes of transient amnesia – it is most often confused with TGA, transient ischemic amnesia (TIA) due to a transient ischemic attack or minor stroke, and psychogenic amnesia. Dementia is sometimes suggested as a possible cause, although it has not, in fact, been established that dementia presents in this way. The distinctive features of the four main causes of transient amnesia are summarized in **Table 1**.

TGA gives rise to a dense anterograde and more variable retrograde amnesia lasting for several hours, often provoked by physical exertion or psychological stress, most often occurring on only a single occasion. TIA typically manifests itself in combination with additional neurological symptoms or signs suggestive of dysfunction of the posterior cerebral circulation and is often accompanied by somnolence. Psychogenic amnesia is often distinguished by the presence of a psychiatric history, psychiatric precipitants, and profound retrograde memory loss, sometimes including loss of personal identity, in the presence of normal anterograde memory.

The diagnostic criteria for TEA itself (Zeman et al., 1998) are

- A history of recurrent, witnessed episodes of transient amnesia.
- Other cognitive functions besides memory, judged by a reliable witness, are intact during typical episodes.
- Evidence for a diagnosis of epilepsy based on one or more of the following:
  - epileptiform abnormalities on EEG,
  - concurrent onset of other clinical features of epilepsy (such as lip smacking or olfactory hallucinations),
  - clear-cut response to anticonvulsant therapy.

The diagnosis of TEA is primarily clinical. An epileptiform EEG supports the diagnosis, but as in other forms of epilepsy a single routine EEG is a relatively insensitive test (positive in around one of three cases) and a normal EEG does not, therefore, exclude a diagnosis of epilepsy. Sleep deprivation can be used to increase the sensitivity of EEG. EEG is expected to give essentially normal results in TGA, TIA, and psychogenic amnesia.

MRI will generally be normal in TEA, but may reveal evidence of cerebrovascular disease in TIA. Diffusion-weighted imaging will often show punctate changes in the hippocampi 48–72 h after an episode of TGA (Bartsch and Butler, 2013).

Living with TEA: Treatment and Prognosis

Treatment

Treatment of TEA is often delayed owing to misdiagnosis of patients as having TGA, TIA, psychogenic attacks, or dementia (Ioannidis et al., 2011; Bartsch and Butler, 2013). As a consequence, patients with TEA typically experience several episodes of amnesia before the correct diagnosis is established. AEDs have been found to be extremely effective at stopping the seizures. A low dose of a single anticonvulsant drug is usually sufficient (Butler et al., 2007). Effective AEDs in TEA include lamotrigine, levetiracetam, carbamazepine, and sodium valproate.

Prognosis

In contrast to dementia, TEA does not appear to cause a progressive decline in memory, although this has not yet been studied formally. Anticonvulsant medication usually prevents the occurrence of further amnesic attacks. Although patients sometimes report a subjective improvement in memory generally when anticonvulsant treatment is started, it has not been formally established that the persistent memory problems associated with TEA (including ALF, autobiographical amnesia, or topographical amnesia) improve on treatment with AEDs (Bartsch and Butler, 2013). This issue is discussed further below. Despite the persistent memory problems caused by TEA, patients generally maintain an active, independent life. Driving may be restricted for a period following diagnosis.

TEA: Case Histories

As mentioned earlier, TEA is characterized by the frequent occurrence of relatively short episodes of amnesia as a result of TLE. During these attacks, anterograde and/or retrograde memory are impaired while other cognitive functions remain intact. We provide two typical examples of TEA, before outlining what is known of the pathology and pathophysiology.

Table 1 Key diagnostic features of transient epileptic amnesia (TEA), transient global amnesia (TGA), transient ischemic amnesia (TIA), and functional/psychogenic amnesia

	TEA	*TGA*	*TIA*	*Functional*
Typical age	50 + years	50 + years	50 + years	Any age
Duration	Typically <1 h	Typically 4–6 h	Variable,<24 h	Typically days or months
Acute memory deficits	Variable mix of anterograde and retrograde amnesia – may later remember attack	Dense anterograde amnesia; variable retrograde amnesia	Not well characterized	Variable, but often dense retrograde amnesia (+/− loss of personal identity) with preserved anterograde memory
Other features (sometimes present)	Olfactory hallucinations, automatisms, brief loss of awareness	Headache, nausea	Focal neurological symptoms and signs, usually arising from posterior circulation	Other functional symptoms and signs may be present; mood disturbance
Precipitants	Waking	Physical (esp. immersion in cold water) and psychological stresses	–	Stressful life events, minor head injury, mood disturbance
Recurrence	Once a month on average	6–10% Recurrence rate/year	Recognized but not well characterized	Unusual but occurs in, e.g., recurrent psychogenic fugue
Medical history	No established risk factors	Migraine	Cerebrovascular risk factors	Psychiatric illness, 'organic' transient amnesia, substance abuse
Interictal/postictal memory	+/− Mild impairment on standard tests; accelerated long-term forgetting, remote memory loss, topographical amnesia	Subtle subclinical impairment may persist for days–months, but no permanent deficit	Risk of lasting memory impairment from completed stroke	Variable

Case 1

L.S., a 62-year-old retired woman, experienced nine episodes of amnesia, until successful treatment with levetiracetam. About half of the amnesic episodes occurred on awakening and some were accompanied by olfactory hallucinations (a 'burning' smell) or unresponsiveness with tongue biting. The episodes of amnesia generally lasted around 20 min to half an hour and were occasionally accompanied by repetitive questioning. According to her husband, she communicated normally during these episodes and, besides the evident memory problems (anterograde and retrograde amnesia), other cognitive abilities remained intact. He described a particular occasion where they had gone to buy bread together. While paying for the bread, she abruptly wandered out of the shop and entered an amnesic state until arriving back home about 15 min later. She 'remembered not being able to remember' in several of the episodes, suggesting that her anterograde amnesia was incomplete. Moreover, on some occasions, memory loss for past events (retrograde amnesia) had been dramatic. For example, in the very first attack, she did not remember that her son and his girlfriend were staying at her home. Around the time that her seizures began, she noticed a deterioration in her memory for familiar places and routes, and a tendency for memories for events to fade more rapidly than she would expect.

Case 2

Mr. B.K. experienced his first amnesic attack at the age of 67 while living in Cyprus. The episode of amnesia came on at the end of a morning spent chopping logs on a hot day. When his wife first noticed something was wrong, he initially failed to recognize her, and then asked questions such as 'What am I doing here?,' 'Where are we?' His cognitive functioning appeared otherwise normal. The amnesic state persisted for about 90 min, and then gradually recovered. He was taken to hospital where he was told initially that he had suffered a stroke. However, on returning to England, he was told that the episode was likely to be one of TGA. Mr. B.K. then began to have briefer episodes of amnesia, often, but not always, occurring on waking. If they occurred on waking, he would ask such questions as "Where are we?," "What day is it?" His amnesic state would gradually improve over about 30 min. These episodes were preceded by nonspecific warnings of 'funny feelings in the head.' Over the year, 12–15 attacks occurred before diagnosis of TEA was established. Moreover, amnesic episodes of a different kind were also witnessed. These short episodes, lasting only for several minutes, were accompanied by lip smacking and humming. Recovery was rapid and memory was not affected by these attacks. During a 24-h ambulatory EEG, nonspecific slow wave activity was observed over the anterior temporal lobes (more marked on the left side), together with rare sharp waves over the left temporal lobe indicating a liability to a partial onset seizure disorder. In conjunction with his seizures, he noticed a tendency for memories to fade more rapidly than he would expect.

Not all cases of TEA are as clear cut as the ones described. The amnesic episodes described in both cases are typical for TEA. Attacks on waking provide a valuable diagnostic clue, but do not always occur. Features such as olfactory hallucinations

or automatisms, suggesting epilepsy, can be absent. As in other focal epilepsies, the interictal EEG may be normal and structural imaging is typically normal. However, brief, recurrent amnesic attacks, especially when combined with interictal complaints of autobiographical, topographical, and/or anterograde memory loss are strongly suggestive of TEA (Butler and Zeman, 2008b; Zeman and Butler, 2010). The diagnostic criteria for TEA mentioned earlier provide a useful summary of key diagnostic features, though as with most such criteria some patients who suffer from the disorder will fail to satisfy them.

Biology of TEA

Several lines of evidence indicate that TEA is a medial TLE. These include clinical features such as the prominence of amnesia in attacks of TEA, the high frequency of olfactory hallucinations, and the associated oroalimentary automatisms (Zeman and Butler, 2010). Structural abnormalities, though rare, and epileptiform discharges, when present, point to the temporal lobes (Butler and Zeman, 2008a, 2008b; Zeman et al., 2013). Recent studies have identified subtle, bilateral hippocampal atrophy across the whole patient group (~8% volume decrease compared with controls) (Butler et al., 2009) with additional bilateral atrophy in the perirhinal and orbitofrontal cortices.

The Microscopic Neuropathology

While there is little information about the neuropathology of TEA, postmortem investigations in a patient with TLE who showed ALF, together with autobiographical memory loss, yielded evidence of neuronal loss and gliosis in regions of both the right and the left hippocampus. The neuronal loss and gliosis were more evident in anterior than posterior hippocampus (Butler et al., 2012).

The Underlying Cause

It is unclear why TEA should mainly occur later in life. Further research on clarifying the relationship between TEA and other forms of late-onset TLE might shed more light on this issue. Suggested explanations include underlying cerebrovascular disease, a neurodegenerative disorder, and a form of immune-mediated limbic encephalitis. These options could explain the age distribution, though at present, there is no compelling support for any of these possibilities (Zeman et al., 2013): no evidence is found for increased risk or burden of vascular disease in TEA (Butler and Zeman, 2011) and an underlying neurodegenerative or immune-mediated disorder is also not observed in the great majority of cases (Razavi et al., 2010).

Accelerated Long-Term Forgetting

Even once the amnesic attacks have been successfully treated with low doses of antiepileptic medication, a substantial number of TEA patients (44%) report difficulty in retrieving memories from the recent past (from around a day to a few weeks before) despite apparently normal initial memory (Butler and Zeman, 2008b). This phenomenon of accelerated forgetting is not unique to TEA and has also been described in patients with other forms of TLE (Blake et al., 2000; Kapur et al., 1997; Kemp et al., 2012; Lucchelli and Spinnler, 1998; Martin et al., 1991; Mayes et al., 2003; O'Connor et al., 1997; for an exception, see Giovagnoli et al., 1995).

Despite often complaining of memory impairment for day-to-day events, TEA patients usually perform normally – that is, within one standard deviation of the mean – on memory tests at standard retention intervals (~30 min) (Butler et al., 2009). This might be taken to suggest that the patients' subjective long-term memory complaints are not in fact the result of a memory disorder at all, but instead, for example, are related to mood disorders (Corcoran and Thompson, 1993). Recent research, however, found no evidence of severe mood changes in TEA: anxiety and depression scores were normal (Butler and Zeman, 2008b). Moreover, tests designed to examine memory at longer intervals have objectively demonstrated an abnormal decrement in long-term memory performance. This phenomenon of accelerated forgetting, occurring over hours, days, or weeks, despite normal retention at the standard 30-min delay interval, provides an alternative explanation for the patients' subjective memory complaints. Indeed, scores on subjective memory measures in TEA correlate with objective measures of ALF but not with standard memory measures (Butler et al., 2009; Zeman et al., 2013). Memory tests with extended retention intervals are likely to be a valuable addition to the neuropsychologists' tool kit in patients with epilepsy (Witt et al., 2012).

The Clinical and Neuropsychological Features of ALF

Experimental tasks using extended retention intervals have revealed increased forgetting rates over intervals varying from 24 h to 6 weeks for verbal and nonverbal laboratory material (Butler et al., 2007, 2009; Manes et al., 2005) and real-life events (Muhlert et al., 2010). When ALF first becomes apparent is not yet known. However, recent studies in patients with TEA and TLE have found that ALF becomes apparent within the first 24 h after learning (see **Figure 1**; for word lists and real-life events: Muhlert et al., 2010; for stories: Jansari et al., 2010 for word pairs: McGibbon and Jansari, 2013).

Undoubtedly, the timescale of ALF, measured under laboratory settings, depend on the test material used (e.g., word lists, stories, word pairing, designs, real-life events) and the experimental protocol applied (e.g., number of learning trials or exposure times, frequency of recall and recall interval) (Butler et al., 2007; Jansari et al., 2010; Kemp et al., 2012; Manes et al., 2005; McGibbon and Jansari, 2013; Muhlert et al., 2010). For example, spaced retrieval of material can prevent or reduce the occurrence of ALF (Jansari et al., 2010; McGibbon and Jansari, 2013) while repetition of immediate retrieval during word list learning does not appear to be effective (Hoefeijzers et al., 2013).

The impact of anticonvulsant medication on long-term forgetting is still an unresolved issue. Many TEA patients show ALF despite effective control of the amnesic episodes by

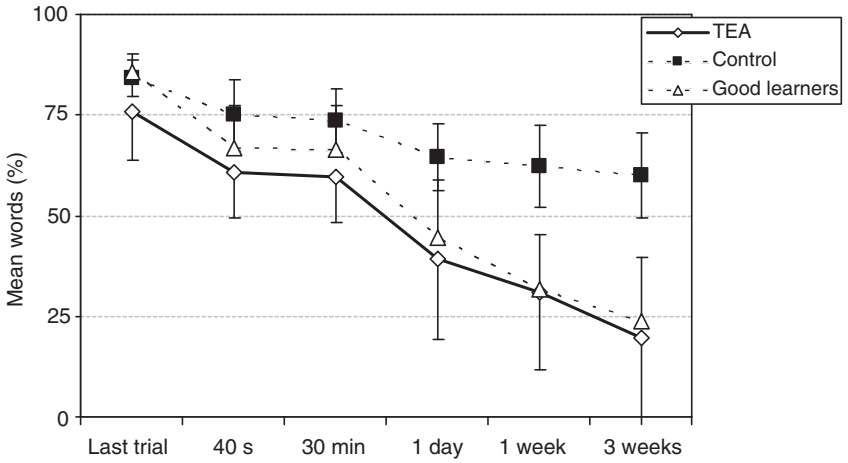

Figure 1 Mean recall performance of TEA and control groups on the list learning test at the last trial and after delays of 40 s, 30 min, 1 day, 1 week, and 3 weeks. TEA = All patients with TEA; Good Learners = TEA patients who were good learners (only those meeting the learning criterion). Error bars show 95% confidence intervals. Reprinted from, Muhlert, N., Milton, F., Butler, C., Kapur, N., Zeman, A. Accelerated forgetting of real-life events in Transient Epileptic Amnesia. Neuropsychologia 48 (11), 3235–3244, © 2010, with permission from Elsevier.

antiepileptic medication (Butler et al., 2007; Manes et al., 2005; Muhlert et al., 2010), indicating that ALF is not typically prevented by such treatment. Pre- and post-treatment studies on ALF show contradictory results: no substantial changes in long-term forgetting were observed in a TEA patient after 6 months of AED treatment (Jansari et al., 2010), while AED treatment did prevent accelerated forgetting in other patients (Midorikawa and Kawamura, 2007; O'Connor et al., 1997). There is, at present, no clear explanation for these conflicting results, but it could be that in some patients ALF is due to subclinical epileptiform discharges, which may be sensitive to anticonvulsant treatment, while in others structural factors, which are not sensitive to treatment, are responsible. This question is discussed further in the next section.

The Biology of ALF

The pathology underlying ALF remains unclear. Given that (1) to date the phenomenon has been reported almost exclusively in association with epilepsy (though see Manes et al., 2005 for an exception) but (2) ALF remains observable, despite abolition of seizures by AEDs, it has been proposed that clinical and subclinical seizure activity may both play a causative role in ALF (Zeman and Butler, 2010). Indeed, several studies have found a correlation between seizure frequency and long-term forgetting rates in epilepsy (Fitzgerald et al., 2013; Mameniskiene et al., 2006; Tramoni et al., 2011; but see Blake et al., 2000 for an exception). Moreover, some studies report improvement of long-term memory performance with treatment of seizures (Gallassi et al., 2011; Midorikawa and Kawamura, 2007). Alternatively, given the persistence of ALF in many patients despite abolition of seizures, ALF could be related to structural brain abnormalities (Zeman and Butler, 2010). Narayanan et al. (2012) observed that the presence or absence of hippocampal damage in TLE determined whether patients showed ALF. However, among patients with TEA, no relationship was found between long-term forgetting rates and regional brain volumes (Butler et al., 2009).

The Cognitive Architecture of ALF

The underlying cognitive basis of ALF also remains uncertain. The pattern of relatively intact retention over short delays, coupled with poor retention after longer delays is suggestive of a deficit in memory consolidation (Blake et al., 2000; Butler et al., 2007; Hoefeijzers et al., 2013; Muhlert et al., 2010). There is evidence for ALF at 24 h in TEA (Muhlert et al., 2010). In TLE, a single case report indicates that it can be detected within 1 h (McGibbon and Jansari, 2013). While it has been suggested that ALF might result from impairment of memory consolidation during sleep, a preliminary study indicates that the benefit of sleep for memory is preserved in patients with TLE (Deak et al., 2011). An impairment of a relatively early stage of memory consolidation is thus a strong candidate for the cognitive basis of ALF, and it may be that the precise timing of the impairment varies between patients. Nevertheless, a subtle impairment of memory encoding, weakening the memories at the time of acquisition, without affecting memory retrieval at the standard test interval of 30 min, remains a possible explanation (Hoefeijzers et al., 2013; Kopelman, 2000; Muhlert et al., 2010).

Remote Memory Impairment

Remote memory impairment, especially affecting autobiographical and topographical memory, is another persistent interictal memory deficit frequently observed in TEA (Butler et al., 2007; Zeman and Butler, 2010). Seventy percent of patients with TEA report remote loss of memory for salient autobiographical events such as weddings or holidays (i.e., autobiographical amnesia). Patients are usually able to remember that the events took place and they had

participated in them. They are, however, unable to re-experience these autobiographical memories vividly and cannot remember details of the events. Moreover, 36% of patients with TEA complain of having difficulties in finding their way around areas they used to be very familiar with (i.e., topographical amnesia) (Butler et al., 2007; Zeman and Butler, 2010). These complaints of topographical memory loss in TEA have, however, not yet been quantified in detail. As in the case of ALF, similar remote memory problems have been described in patients with other types of TLE (Addis et al., 2007; Butler and Zeman, 2008b; Noulhiane et al., 2007; St-Laurent et al., 2009, 2011; Viskontas et al., 2000).

The Clinical Features of Remote Memory Impairment

In keeping with patients' subjective autobiographical memory complaints, Butler et al. (2007) observed autobiographical memory loss across the four most recent life decades in 22 TEA patients, using a modified version of the Autobiographical Memory Interview (AMI) (Kopelman et al., 1989). Moreover, personal semantic memory was also slightly impaired over the most recent decades. Similar results were observed by Manes et al. (2005) using the Galton–Crovitz test of cued autobiographical memory. Here, seven TEA patients showed autobiographical memories loss for events that happened two decades earlier (period of 1985–94) but not for autobiographical memories of the five most recent years (period 1995–99). More recently, Milton et al. (2010) tested 14 TEA patients and 11 age- and IQ-matched controls using the Autobiographical Interview (Levine et al., 2002). In contrast to the results of Butler et al. (2007), autobiographical memory loss was observed across the entire lifespan (see **Figure 2**) and was present for all the different types of contextual information (event, time, place, perceptual, and thought/emotion details). This test difference may well be due to the use of the Autobiographical Interview, which is thought to be more sensitive than the AMI (Zeman et al., 2013). Again these patients showed mild impairments of personal semantic memory and, to a lesser extent, of public semantic memory (Milton et al., 2010).

The inability to retrieve autobiographical memories acquired many years before the onset of TEA, in the absence of any deficit on standard tests of anterograde memory, indicates that TEA is a form of focal retrograde amnesia as this term is usually understood (Butler et al., 2007; Manes et al., 2001, 2005; Milton et al., 2010; Zeman et al., 2013). Notably, the patients' subjective memory complaints correlate with performance on remote autobiographical memory tests, but not with standard test of anterograde memory (Butler et al., 2009). Therefore, this remote memory deficit seems to be, just like ALF, a nonstandard form of memory impairment that contributes to the patient's everyday memory complaints, while not being detected on standard memory tests (Zeman et al., 2012).

The Biology of Remote Memory Impairment

Unlike standard anterograde memory test results, measures of autobiographical amnesia and ALF are not related to focal atrophy in the temporal lobes (Butler et al., 2009) and orbitofrontal regions (Butler et al., 2013). These two unusual forms of selective long-term memory loss have no gross anatomical correlate. It is more likely that subtle structural damage or physiological disruption, presumably by subclinical epileptic activity, affect the function and connectivity of memory networks to cause these two forms of epilepsy-related long-term amnesia (Butler et al., 2012; Milton et al., 2010). Indeed, Milton et al. (2012) found that during retrieval of autobiographical memories, activation of the autobiographical memory network was less extensive in TEA patients than in control participants. Hypoactivation was detected particularly in regions of the right hemisphere: activation was reduced in the posterior parahippocampal gyrus (pPHG) (especially for mid-life and recent memories), the temporoparietal junction, and the cerebellum. Moreover, the connectivity between the right pPHG and right middle temporal gyrus was also diminished.

The Cognitive Architecture of Remote Memory Impairment

TEA patients can lose the ability to vividly re-experience autobiographical memories acquired many years before the onset of TEA. Whether this remote memory loss reflects a deficit in memory acquisition, consolidation or retrieval is still not known with certainty, but the available evidence suggests that a degradation of existing engrams plays an important role (Dudai, 2004).

A subtle anterograde memory deficit seems unlikely to explain this remote memory impairment. First, in many cases, there is strong evidence that the lost autobiographical memories had previously been accessible over long periods of time. Second, despite the late onset of the epilepsy and its associated memory complaints, the autobiographical memory loss affects the entire lifespan. While a subtle anterograde memory deficit such as ALF would be expected to give rise to loss of episodic memories acquired after the onset of TEA, it is unlikely to explain the lifelong depletion of autobiographical memory loss seen in patients with TEA (Zeman et al., 2012).

Failure to maintain stored memories provides the best explanation for the autobiographical memory loss seen in TEA (Tramoni et al., 2011). This could reflect damage to stored engrams or, relatedly, a deficit in very long-term consolidation processes or reconsolidation mechanisms (Dudai, 2006; Wang and Morris, 2010).

A retrieval deficit could also hinder the retrieval of remote memories. However, it is unlikely to be the main cause for the autobiographical memory loss seen in TEA. First, it would be expected to complicate the retrieval of very recently acquired memories as well, but these often appear to be spared in patients with TEA. Second, extensive cueing, both verbally and visually, fails to improve recollection of autobiographical memories (Manes et al., 2001, 2005; Muhlert et al., 2010; Milton et al., 2010). However, Milton et al. (2011) report a single TEA patient who recovered some 'lost' autobiographical memories following a series of episodes of *déjà vu*, suggesting that indeed some autobiographical memory 'loss' might be due to impairments in memory retrieval (Zeman et al., 2012).

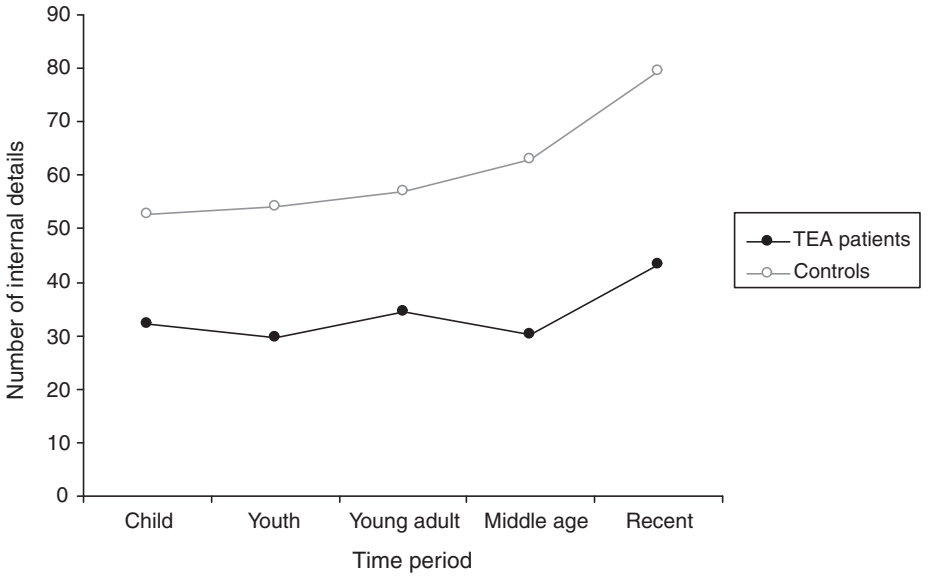

Figure 2 Mean number of details recalled for events from five life periods by patients with TEA compared with control participants. Reproduced from Adam Zeman, Chris Butler, Nils Muhlert, and Fraser Milton. Novel forms of forgetting in temporal lobe epilepsy. Epilepsy & Behavior 26, 335–342, © 2013, Elsevier with permission.

Autobiographical Memory Loss in TEA and Theoretical Models of Memory

There are currently two principle theories of remote autobiographical memory. The 'standard' theory of consolidation proposes that autobiographical memory traces are transferred over time, via 'systems consolidation,' into neocortical locations in which they become independent of the hippocampi (e.g., Squire and Bayley, 2007). In contrast, the multiple trace theory (MTT) suggests the hippocampal complex is involved life-long in the evocation of rich autobiographical memories (e.g., Moscovitch et al., 2005; Rosenbaum et al., 2001). At first sight, the uniform impairment of autobiographical memories extending across the entire lifespan appears more in line with the predictions of MTT. However, the standard consolidation theory could explain these results if epileptiform activity spreading from the MTLs disrupted the cortical memory network. Therefore, current findings fail to adjudicate between these theories (Zeman et al., 2013).

Conclusion

TEA is a form of TLE characterized by recurrent episodes of transient amnesia caused by epilepsy. This epileptic syndrome usually occurs in middle age and is often accompanied by ALF and remote memory impairment. Both interictal complaints are nonstandard forms of memory deficits, which are at least partially resistant to AEDs and can occur in the face of normal performance on standard memory tests (Zeman et al., 2013).

Recent work has elucidated the clinical features, pathology, and pathophysiology of the transient amnesic attacks occurring in TEA, and the associated ALF and remote memory loss. Further research on the relationship between epilepsy and memory impairments in TEA will deepen our understanding of the morphological and physiological processes involved in the formation, storage, and retrieval of episodic memories. Understanding how epileptic activity disturbs these processes may allow more effective treatment of epilepsy-related memory disorders in the future (Zeman et al., 2013).

See also: Aging and Memory; Amnesia: General; Amnesia: Psychogenic; Epilepsy; Episodic Memory; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of; Post Traumatic Amnesia; Reconstructive Memory, Psychology of; Retrieval; Retroactive Interference; Sleep and Memory; Transient Global Amnesia.

Bibliography

Addis, D.R., Moscovitch, M., McAndrews, M.P., 2007. Consequences of hippocampal damage across the autobiographical memory network in left temporal lobe epilepsy. Brain 130 (9), 2327–2342.

Bartsch, T., Butler, C., 2013. Transient amnesic syndromes. Nature Review Neurology. 9, 86–97.

Blake, R.V., Wroe, S.J., Breen, E.K., McCarthy, R.A., 2000. Accelerated forgetting in patients with epilepsy: evidence for an impairment in memory consolidation. Brain 123 (3), 472–483.

Butler, C., Bhaduri, A., Acosta-Cabronero, J., Nestor, P.J., Kapur, N., Graham, K.S., Zeman, A., 2009. Transient epileptic amnesia: regional brain atrophy and its relationship to memory deficits. Brain 132 (2), 357–368.

Butler, C., van Erp, W., Bhaduri, A., Hammers, A., Heckemann, R., Zeman, A., 2013. Magnetic resonance volumetry reveals focal brain atrophy in transient epileptic amnesia. Epilepsy & Behavior 28 (3), 363–369.

Butler, C., Graham, K.S., Hodges, J.R., Kapur, N., Wardlaw, J.M., Zeman, A., 2007. The syndrome of transient epileptic amnesia. Annals of Neurology 61, 587–598.

Butler, C., Kapur, N., Zeman, A., Weller, R., Connelly, A., 2012. Epilepsy-related long-term amnesia: anatomical perspectives. Neuropsychologia 50, 2973–2980.

Butler, C., Zeman, A., 2008a. A case of transient epileptic amnesia with radiological localization. Nature Clinical Practice Neurology 4 (9), 516–521.
Butler, C., Zeman, A., 2008b. Recent insights into the impairment of memory in epilepsy: transient epileptic amnesia, accelerated long-term forgetting and remote memory impairment. Brain 131, 2243–2263.
Butler, C., Zeman, A., 2011. The causes and consequences of transient epileptic amnesia. Behavioural Neurology 24, 299–305.
Corcoran, R., Thompson, P., 1993. Epilepsy and poor memory: who complains and what do they mean? British Journal of Clinical Psychology 32 (2), 199–208.
Deak, M.C., Stickgold, R., Pietras, A.C., Nelson, A.P., Bubrick, E.J., 2011. The role of sleep in forgetting in temporal lobe epilepsy: a pilot study. Epilepsy & Behavior 21, 462–466.
Dudai, Y., 2004. The neurobiology of consolidations, or, how stable is the engram? Annual Review of Psychology 55, 51–86.
Dudai, Y., 2006. Reconsolidation: the advantage of being refocused. Current Opinion in Neurobiology 16 (2), 174–178.
Fitzgerald, Z., Thayer, Z., Mohemed, A., Miller, L.A., 2013. Examining factors related to accelerated long-term forgetting in epilepsy using ambulant EEG monitoring. Epilepsia 54, 819–827.
Gallassi, R., Sambati, L., Poda, R., Stanzani Maserati, M., Oppi, F., Giulioni, M., Tinuper, P., 2011. Accelerated long-term forgetting in temporal lobe epilepsy: evidence of improvement after left temporal pole lobectomy. Epilepsy & Behavior 22 (4), 793–795.
Giovagnoli, A.R., Casazza, M., Avanzini, G., 1995. Visual learning on a selective reminding procedure and delayed recall in patients with temporal lobe epilepsy. Epilepsia 36 (7), 704–711.
Hoefeijzers, S., Dewar, M., Della Sala, S., Zeman, A., Butler, C., 2013. Accelerated long-term forgetting in transient epileptic amnesia: an acquisition or consolidation deficit? Neuropsychologia 51 (8), 1549–1555.
Ioannidis, P., Balamoutsos, G., Karabela, O., Kosmidis, M.H., Karacostas, D., 2011. Transient epileptic amnesia in a memory clinic setting: a report of three cases. Epilepsy & Behavior 20 (2), 414–417.
Jansari, A.S., Davis, K., McGibbon, T., Firminger, S., Kapur, N., 2010. When "long-term memory" no longer means "forever": analysis of accelerated long-term forgetting in a patient with temporal lobe epilepsy. Neuropsychologia 48 (6), 1707–1715.
Kapur, N., 1990. Transient epileptic amnesia: a clinically distinct form of neurological memory disorder. In: Markowitsch, H. (Ed.), Transient Global Amnesia and Related Disorders. Hogrefe and Huber, New York, pp. 140–151.
Kapur, N., Millar, J., Colbourn, C., Abbott, P., Kennedy, P., Docherty, T., 1997. Very long-term amnesia in association with temporal lobe epilepsy: evidence for multiple-stage consolidation processes. Brain and Cognition 35 (1), 58–70.
Kemp, S., Illman, N.A., Moulin, C.J., Baddeley, A.D., 2012. Accelerated long-term forgetting (ALF) and transient epileptic amnesia (TEA): two cases of epilepsy-related memory disorder. Epilepsy & Behavior 24 (3), 382–388.
Kopelman, M.D., 2000. Focal retrograde amnesia and the attribution of causality: an exceptionally critical view. Cognitive Neuropsychology 17 (7), 585–621.
Kopelman, M.D., Wilson, B.A., Baddeley, A.D., 1989. The autobiographical memory interview: a new assessment of autobiographical and personal semantic memory in amnesic patients. Journal of Clinical and Experimental Neuropsychology 11 (5), 724–744.
Levine, B., Svoboda, E., Hay, J.F., Winocur, G., Moscovitch, M., 2002. Aging and autobiographical memory: dissociating episodic from semantic retrieval. Psychology and Aging 17 (4), 677–689.
Lucchelli, F., Spinnler, H., 1998. Ephemeral new traces and evaporated remote engrams: a form of neocortical temporal lobe amnesia? A preliminary case report. Neurocase 4 (6), 447–459.
Mameniskiene, R., Jatuzis, D., Kaubrys, G., Budrys, V., 2006. The decay of memory between delayed and long-term recall in patients with temporal lobe epilepsy. Epilepsy & Behavior 8 (1), 278–288.
Manes, F., Graham, K.S., Zeman, A., de Lujan Calcagno, M., Hodges, J.R., 2005. Autobiographical amnesia and accelerated forgetting in transient epileptic amnesia. Journal of Neurology, Neurosurgery & Psychiatry 76 (10), 1387–1391.
Manes, F., Hodges, J.R., Graham, K.S., Zeman, A., 2001. Focal autobiographical amnesia in association with transient epileptic amnesia. Brain 124 (3), 499–509.
Martin, R.C., Loring, D.W., Meador, K.J., Lee, G.P., Thrash, N., Arena, J.G., 1991. Impaired long-term retention despite normal verbal learning in patients with temporal lobe dysfunction. Neuropsychology 5 (1), 3–12.
Mayes, A.R., Isaac, C.L., Holdstock, J.S., Cariga, P., Gummer, A., Roberts, N., 2003. Long-term amnesia: a review and detailed illustrative case study. Cortex 39 (4), 567–603.
McGibbon, T., Jansari, A.S., 2013. Detecting the onset of accelerated long-term forgetting: evidence from temporal lobe epilepsy. Neuropsychologia 51 (1), 114–122.
Milton, F., Butler, C., Benattayallah, A., Zeman, A., 2012. The neural basis of autobiographical memory deficits in transient epileptic amnesia. Neuropsychologia 50 (14), 3528–3541.
Milton, F., Butler, C., Zeman, A., 2011. Transient epileptic amnesia: déjà vu heralding recovery of lost memories. Journal of Neurology, Neurosurgery & Psychiatry 82 (10), 1178–1179.
Milton, F., Muhlert, N., Pindus, D.M., Butler, C., Kapur, N., Graham, K.S., Zeman, A., 2010. Remote memory deficits in transient epileptic amnesia. Brain 133 (5), 1368–1379.
Midorikawa, A., Kawamura, M., 2008. Recovery of long-term anterograde amnesia, but not retrograde amnesia, after initiation of an anti-epileptic drug in a case of transient epileptic amnesia. Neurocase 13 (5–6), 385–389.
Moscovitch, M., 2001. New views on old memories: Re-evaluating the role of the hippocampal complex. Behavioural Brain Research 127 (1), 183–197.
Moscovitch, M., Rosenbaum, R.S., Gilboa, A., Addis, D.R., Westmacott, R., Grady, C., Nadel, L., 2005. Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. Journal of Anatomy 207 (1), 35–66.
Muhlert, N., Milton, F., Butler, C., Kapur, N., Zeman, A., 2010. Accelerated forgetting of real-life events in Transient Epileptic Amnesia. Neuropsychologia 48 (11), 3235–3244.
Narayanan, J., Duncan, R., Greene, J., Leach, J.P., Razvi, S., McLean, J., Evans, J.J., 2012. Accelerated long-term forgetting in temporal lobe epilepsy: verbal, nonverbal and autobiographical memory. Epilepsy & Behavior 25 (4), 622–630.
Noulhiane, M., Piolino, P., Hasboun, D., Clemenceau, S., Baulac, M., Samson, S., 2007. Autobiographical memory after temporal lobe resection: neuropsychological and MRI volumetric findings. Brain 130 (12), 3184–3199.
O'Connor, M., Sieggreen, M.A., Ahern, G., Schomer, D., Mesulam, M., 1997. Accelerated forgetting in association with temporal lobe epilepsy and paraneoplastic encephalitis. Brain and Cognition 35 (1), 71–84.
Razavi, M., Barrash, J., Paradiso, S., 2010. A longitudinal study of transient epileptic amnesia. Cognitive and Behavioral Neurology 23, 142–145.
Rosenbaum, R.S., Winocur, G., Moscovitch, M., 2001. New views on old memories: Re-evaluating the role of the hippocampal complex. Behavioural Brain Research 127 (1), 183–197.
St-Laurent, M., Moscovitch, M., Levine, B., McAndrews, M.P., 2009. Determinants of autobiographical memory in patients with unilateral temporal lobe epilepsy or excisions. Neuropsychologia 47 (11), 2211–2221.
St-Laurent, M., Moscovitch, M., Tau, M., McAndrews, M.P., 2011. The temporal unraveling of autobiographical memory narratives in patients with temporal lobe epilepsy or excisions. Hippocampus 21 (4), 409–421.
Squire, L.R., Bayley, P.J., 2007. The neuroscience of remote memory. Current Opinion in Neurobiology 17 (2), 185–196.
Tramoni, E., Felician, O., Barbeau, E.J., Guedj, E., Guye, M., Bartolomei, F., Ceccaldi, M., 2011. Long-term consolidation of declarative memory: insight from temporal lobe epilepsy. Brain 134 (3), 816–831.
Viskontas, I.V., McAndrews, M.P., Moscovitch, M., 2000. Remote episodic memory deficits in patients with unilateral temporal lobe epilepsy and excisions. The Journal of Neuroscience 20 (15), 5853–5857.
Wang, S.H., Morris, R.G., 2010. Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. Annual Review of Psychology 61, 49–79.
Witt, J.A., Glockner, C., Helmstaedter, C., 2012. Extended retention intervals can help to bridge the gap between subjective and objective memory impairment. Seizure 21, 134–140.
Zeman, A., Boniface, S.J., Hodges, J.R., 1998. Transient epileptic amnesia: a description of the clinical and neuropsychological features in 10 cases and a review of the literature. Journal of Neurology, Neurosurgery & Psychiatry 64 (4), 435–443.
Zeman, A., Butler, C., 2010. Transient epileptic amnesia. Current Opinion in Neurology 23, 610–616.
Zeman, A., Butler, C., Hodges, J., Kapur, N., 2012. The syndrome of transient epileptic amnesia. In: Zeman, A., Kapur, N., Jones-Gotman, M. (Eds.), Epilepsy and Memory. Oxford University Press, Oxford, pp. 139–159.
Zeman, A., Butler, C., Muhlert, N., Milton, F., 2013. Novel forms of forgetting in temporal lobe epilepsy. Epilepsy & Behavior 26, 335–342.

Relevant Website

http://projects.exeter.ac.uk/time/ – TIME Project.

Transient Global Amnesia

Juliane Döhring, Karsten Witt, and Thorsten Bartsch, University Hospital Schleswig-Holstein, University of Kiel, Kiel, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Transient global amnesia is an acute onset amnesia characterized by an anterograde and retrograde amnesia of declarative memory lasting less than 24 h without concomitant neurological deficits. The prognosis of this enigmatic syndrome is benign. Although the exact etiology and pathogenesis remain cryptic, focal lesions confined to the CA1 field of the hippocampus can be detected in magnetic resonance imaging. These findings indicate a perturbation of hippocampal dependent memory circuits as the correlate of the amnesia. Here, we will describe the clinical presentation of transient global amnesia and illustrate associated pathogenetic mechanisms.

Introduction

Short-lasting amnesic syndromes were recognized already in the nineteenth century (Ribot, 1882; Sollier, 1892). In the classic French neurological literature of the nineteenth century, cases of transient amnesia have been reported that occurred as a consequence of an emotional event (Gil et al., 2010). Guyotat and Courjon (1956) provided the first systematic phenotypical description of the syndrome by characterizing its clinical and epidemiological properties in 17 patients. Fisher and Adams (1964) referred to this phenomenon as 'transient global amnesia' (TGA) and suggested hypotheses about its possible etiology.

Nowadays, TGA is a defined clinical syndrome that is characterized by a sudden onset of a severe amnesic syndrome that extends to all sensory modalities (therefore 'global'). An amnesic syndrome is defined by the inability to learn new, consciously accessible – or explicit – information (anterograde amnesia). This deficit is accompanied by difficulty recalling events that occurred prior to the damage (retrograde amnesia). Retrograde amnesia is often temporally graded in that remote memories, acquired well before the lesion, are better preserved than recent memories. By definition, nonmemory cognitive domains such as perception, cognition, general intelligence, and praxis are intact.

In the following, an overview of the present state of TGA research will be provided, beginning with its epidemiological properties, its anatomical underpinning, the diagnosis, and possible differential diagnosis. The majority of TGA patients report certain precipitating events before TGA onset that will be characterized in the next section. Further, the main hypotheses of TGA pathology will be presented. Finally, some findings from imaging studies will be summarized.

Clinical Presentation

The clinical presentation of TGA is temporally limited, ranging from a few hours up to (per definitionem) 24 h (Bartsch and Butler, 2013). The onset is sudden and manifests in an inability to recollect recent experiences or form new memories. The amnesia affects all modalities of senses (visual, verbal, tactile). The retention span of the anterograde deficit at its peak may last only 2–3 min. When the attention shifts, the retention span may be even shorter. Stereotype repetition of statements or repeated questioning concerning situational circumstances or objects that changed in the recent past is typical. Because of the inability to reconstruct recent events, patients are frequently confused and disorientated in time and place (Owen et al., 2007). The retrograde amnesia varies between individuals and can extend to weeks or decades into the past. Most patients cannot recollect what they had been doing at the beginning of the attack.

Apart from the amnesia, TGA patients do not show impairments of cognition, vigilance, or procedural knowledge. Their intellectual functions are preserved and they are in fact able to accomplish complex operations like driving a car, cooking, or gardening. Working memory is largely maintained as well as procedural memory, priming, conceptual, and public knowledge. During the acute state social skills, awareness of the own identity or the identity of close family members should be retained. A loss of self-awareness and consciousness excludes the diagnosis of a TGA. However, mostly patients are not explicitly aware of their memory failure. Instead they 'feel' something is not quite right without being able to name it explicitly.

Beyond amnestic symptoms, an altered emotional state can sometimes be observed: some patients exhibit a certain helplessness, perplexity, irritation, and/or anxiety or fear to die. Some patients cry and exhibit symptoms close to a panic attack (Inzitari et al., 1997). An altered emotional state and signs of depressed mood can aggravate the amnesic deficit. Deteriorations of the emotional state improve following the dynamic of the episode (Noel et al., 2008).

TGA can be accompanied by mild vegetative symptoms such as headache, vertigo, nausea, cold extremities, chills, or flushes and may be accompanied by a moderately elevated blood pressure (Quinette et al., 2006).

After a few hours, the retrograde amnesic deficit gradually resolves from remote to recent memories. Likewise, the short-term memory recovers gradually, reflected by an extending retention span. Typically after recovery an amnesic gap for the acute TGA episode remains (Sander and Sander, 2005).

Case Example

Mrs. E. is a 71-year-old biologist who had an open-heart surgery several years ago and since then, thoughts about this episode stirs up

 http://dx.doi.org/10.1016/B978-0-08-097086-8.51064-2

her emotions. In general, she describes herself as a rather anxious person with a past history of migraine. In April she watched a medical documentary in TV, dealing with the subject 'heart surgery.' She asked her husband to switch off the TV program because she did 'not want to be confronted with it again.' Minutes after this event she repeatedly asked the same questions ("Who put the suitcase there?" "Where does the flower vase come from?"). Both objects were quite recently arranged on that place. On the way to the hospital she was alert, but her repeated questioning indicated a profound anterograde amnesia with a retention span of 2–3 min. After several hours, she fully recovered but a lack of memory covering the acute phase remained. Neurological examination, EEG, and cerebrovascular ultrasound were normal. Subsequent magnetic resonance imaging (MRI) 2 days later showed a diffusion lesion in her left hippocampus.

Diagnosis and Differential Diagnosis

Diagnosis

Diagnosis of TGA is based entirely on the clinical history and examination. Caplan (1985) proposed diagnostic criteria that were supplemented by Hodges and Warlow (1990a,b). The diagnosis of TGA relies primarily on clinical findings and comprises the criteria in Table 1. Caplan and Hodges added the exclusion of other causes of amnesia. Although not explicitly mentioned, there is a temporally graded retrograde amnesia present during the attack. Certain mild vegetative symptoms may occur additionally.

Differential Diagnosis

If the clinical symptoms do not match the diagnostic criteria (due to an ambiguous symptomatology, no reliable witness, presence of vascular risk factors), certain differential diagnoses have to be excluded (Bartsch and Deuschl, 2010) (Table 2). The first category of differential diagnosis that may rarely mimic a TGA includes (1) vascular diseases: stroke in hippocampus and thalamus (Berli et al., 2009), transient ischemic attack (TIA) in the vascular distribution of the posterior cerebral artery, infarction in the territory of the anterior choroidal artery (Bartsch and Deuschl, 2010) as well as strategic diencephalic lacunar syndromes. If stroke risk factors have been detected, an MRI and/or CT, an echocardiography, as well as a cardiac check-up are required. (2) The second category, which can mimic a TGA refers to epileptic activity including prolonged complex partial seizures or nonconvulsive status epilepticus as well as a transient epileptic amnesia (TEA) (Table 3). TEA may be difficult to differentiate from a TGA, but usually attacks are shorter (~1 h) but more frequent (~3–4 per year) and respond well to anticonvulsive medication (Berli et al., 2009). In contrast to TGA, TEA carries a risk of developing persistent memory impairment, which can be mistaken for dementia (Zeman and Butler, 2010). Epilepsy can be ascertained by means of an electroencephalogram. In young patients suffering from diabetes, hypoglycemic seizures or hypoglycemia itself can lead to TGA-like symptoms. (3) Subacute prodromal encephalitis such as herpes simplex encephalitis or limbic encephalitis can lead to an amnesic syndrome. (4) Head injuries like cerebral contusion sometimes produce posttraumatic retrograde amnesia. (5) Amnesic syndromes caused by drug intake (hypnotics, benzodiazepines, opioids, antidepressants, marijuana, alcohol) have to be excluded. Side effects of psychotropic medication such as psychiatric pharmacotherapy should be considered (Bartsch and Deuschl, 2010). Hence, a drug history should be obtained. (6) Also psychogenic syndromes might elicit TGA-like presentations: functional dysfunctions after emotionally affecting events in general as well as psychiatric disorders (e.g., dissociative episodes, psychogenic fugue, acute episode of depression). Psychogenic amnesia typically involves profound retrograde amnesia, often with loss of personal identity, in the context of preserved new learning.

Table 1 Diagnostic criteria for transient global amnesia

Diagnostic criteria for transient global amnesia after Caplan and Hodges (Caplan, 1985; Hodges and Warlow, 1990a,b):
● The presence of an anterograde amnesia, that is witnessed by an observer
● No clouding of consciousness or loss of personal identity
● Cognitive impairment limited to amnesia
● No focal neurological or epileptic signs
● No recent history of head trauma or seizures
● Resolution of symptoms within 24 h

Mild vegetative symptoms (headache, nausea, dizziness) might be present during the acute phase. Temporally graded retrograde amnesia is usually present during the acute attack, although this feature is not included in the current diagnostic criteria.

Table 2 Differential diagnostic considerations of transient global amnesia

Differential diagnostic considerations of transient global amnesia
● Vascular diseases: ischemia in the posterior cerebral circulation
● Intoxication, adverse drug side effects
● Complex focal seizures, 'transient epileptic amnesia,' postictual conditions
● Psychogenic syndromes: functional amnesia, psychogenic fugue, dissociative conditions, acute episode of depression
● Posttraumatic amnesia after head injury
● Hypoglycemia

Table 3 Diagnostic criteria of transient epileptic amnesia

Diagnostic criteria of *transient epileptic amnesia* after Zeman et al. (1998) and Butler et al. (2007)
● A history of recurrent witnessed episodes of transient amnesia
● Cognitive functions other than memory judged to be intact during typical episodes by a reliable witness
● Evidence for a diagnosis of epilepsy based on one or more of the following:
• Epileptiform abnormalities on electroencephalography
• The concurrent onset of other clinical features of epilepsy (e.g., lip-smacking, olfactory hallucinations)
• A clear-cut response to anticonvulsant therapy

Epidemiology

The incidence of TGA per year ranges between 3 and 8 per hundred thousand (Miller et al., 1987; Hodges and Warlow, 1990a,b; Berli et al., 2009). A Finnish study found a higher incidence (10 per hundred thousand per year in the general population, 32 per hundred thousand for patients older than 50 years (Koski and Marttila, 1990)). The peak of incidence is located in the 7th decade of age (mean = 60.3 years of age, based on 246 patients described in the literature (Quinette et al., 2006)). The majority of attacks (75%) occur between the age of 50 and 70 years. An occurrence in persons younger than 40 years and above 80 years is rare.

Gender Distribution

Gender seems almost uniformly distributed in TGA. In their metaanalysis, Quinette et al. (2006) did not find a significant gender difference (46.4% men, 53.6% women) among 1333 patients collected from 86 published studies. However, men and women probably differ in terms of the risk factor profile. Women more frequently report a history of migraine and tend to develop a TGA after an emotionally arousing event more often than men. Men were more frequently engaged in physical exerting activities before TGA.

Time of Onset

In the literature, an unequal seasonal distribution of TGA is reported, peaking in spring and summer (61%). In autumn and winter only 33% of the cases have been described (Quinette et al., 2006). With respect to the time of day, a TGA attack most frequently occurs in the morning (59% from 6.00 to 11.30 a.m.), the remaining part is distributed equally between afternoon and evening. There is no reported case of TGA starting in the night (between 0.00 and 6.00 a.m.) or waking up with a TGA.

Precipitating Events

The majority of patients (about a half up to 90%, depending on subjective criteria) report precipitating events that happened immediately before TGA onset. About half of these patients experienced emotional stress or physical effort preceding the attack. About a third was exposed to water contact, temperature change or acute pain, or had sexual intercourse (Figure 1).

Typical precipitating situations of the 'emotionally induced TGA' (Merriam et al., 1992) are stressful events like quarrels, the receipt of bad news, a heated argument, attending a funeral, family reunions, overwork, money worries, or reliving of traumatic situations. Quinette et al. (2006) emphasize the importance of remote factors regarding the personality profile of TGA patients. Persons with an anxious profile and phobic personality traits seem more vulnerable to develop a TGA (Figure 2).

Men are more likely to experience a TGA after physical effort like sports, housework, gardening, and arousing manual activities like coating a boat or sawing wood. Some precipitants contain both psychic and physical arousal like sexual intercourse, health problems in general, medical diagnostic procedures and acute pain. Examples for 'merely somatic' precipitants are Valsalva-associated maneuvers (e.g., coughing, emesis, straining). TGA may be elicited by sudden changes in body temperature. TGA provoked by sudden immersion in cold or hot water tends to cluster in the summer months. Some patients were particularly exposed to temperature changes (washing their hair). Rarely, changes in the atmospheric environment (at high altitudes) are reported. Caplan (1990) suggested that the main categories of precipitants (stress, effort, water) can be considered as abrupt changes in physical activity, temperature, or atmospheric environment, which may evoke an emotional, hormonal, or autonomic body response. The circumstances of TGA occurrence can be also summarized under psychic and physical stress reactions, changes of bodily homeostasis, an allostatic overload, and changes in emotional state (Merriam et al., 1992; Kessler et al., 2001).

Cognitive Outcome

Although the recovery seems complete within 24 h, it is controversially debated whether TGA results in longer-lasting, but more subtle neuropsychological deficits (Quinette et al., 2003; Bartsch et al., 2006; Jager et al., 2009a) as subtle effects persisting for several months following an acute episode has been described (Guillery-Girard et al., 2006). Kessler et al. (2001) report slight disturbances in verbal and nonverbal long-term memory as well as in verbal fluency a few days after the TGA attack. Some patients exhibit a persistent, but subclinical impairment of memory functions (Caffarra et al., 1981). However, in a metaanalysis comprising 25 studies, Jager et al. (2009a) did not identify significant differences in long-term performance between TGA patients and healthy controls. Uttner et al. (2012) did not determine significant cognitive differences between TGA patients and healthy controls in a 2-year follow up. This was aimed at patients with and without detected diffusion-weighted imaging (DWI) lesions. Also complaints concerning the emotional state (anxiety, irritability, fatigue) resolve, though in some cases they persist for some days after the episode. Neri et al. (1995) report a higher rate of mood disorders several months after the acute state.

Recurrence

In principle, TGA is considered a benign condition. The rate of recurrence varies widely across studies, depending on the length of the follow-up duration. Melo et al. (1992) reported a recurrence rate of 2.9%, whereas Fredericks (1990) registered a recurrence rate of 26.3%. In the series of Quinette et al. (2006), 3.5% of 142 TGA patients had a second episode between 1994 and 2004. Gandolfo et al. (1992) compared the survival curve of TGA patients with the analogous curve of sex- and age-matched healthy controls. The curves did not differ significantly.

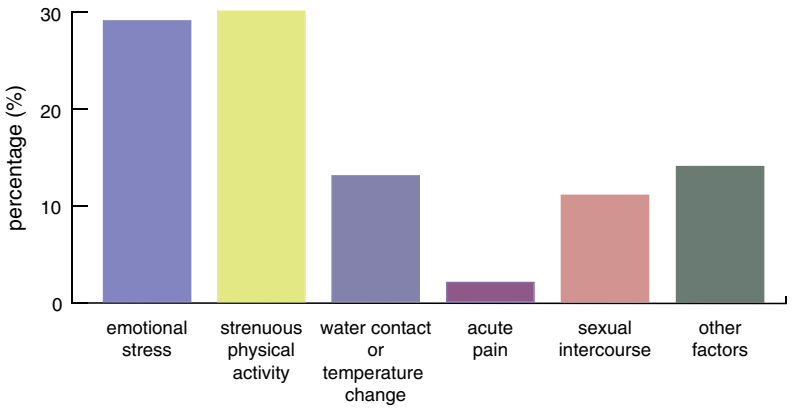

Figure 1 Frequency graphs of various precipitating events preceding the onset of an acute TGA (Quinette et al., 2006). Used with permission from Oxford University Press. Modified after Bartsch and Deuschl (2010), used with permission from Elsevier.

Neuropsychology

The central neuropsychological feature of TGA is the retrograde as well as the anterograde impairment of declarative memory with the anterograde component being more profound (Eustache et al., 1999; Guillery-Girard et al., 2004). The formation of declarative memory contents is primarily based on the hippocampus and the temporal lobe system. Declarative memory comprises semantic and episodic memory. Episodic memory represents memory of personal experiences and specific events in time and space that can be explicitly stated. Semantic memory includes facts, meanings, concepts, and knowledge about the external world and refers to general factual knowledge. The hippocampus is a central node in a large bidirectional network that is essential for declarative memory processing of encoding, consolidation, and retrieval. It includes neocortical association regions, subcortical nuclei, the medial temporal lobe (MTL), parahippocampal areas, and the hippocampus itself (Bartsch and Butler, 2013). It receives input from almost all neocortical association areas and the entorhinal cortex. The severe impairment of episodic memory is thought to be caused by hippocampal dysfunction during TGA. It is thought that the hippocampus is involved in the representation of recent episodic memory contents to remote ones by a gradual information transfer from the hippocampus to neocortical structures. Accordingly, the hippocampal dysfunction in the acute TGA state produces a retrograde amnesia as defined by Ribot (1881) who states that remote memories are more robust than more recent memories. Such a temporal gradient in TGA could be shown by Bartsch et al. (2011). Autobiographical memories are considered a part of episodic memory and refer to the recollection of episodes of a person's history, whose temporal sequence, space, and setting can be vividly relived. These aspects are subsumed under the term 'recollection.' Familiarity-based recognition memory (the feeling that the event was previously experienced, typically with spared

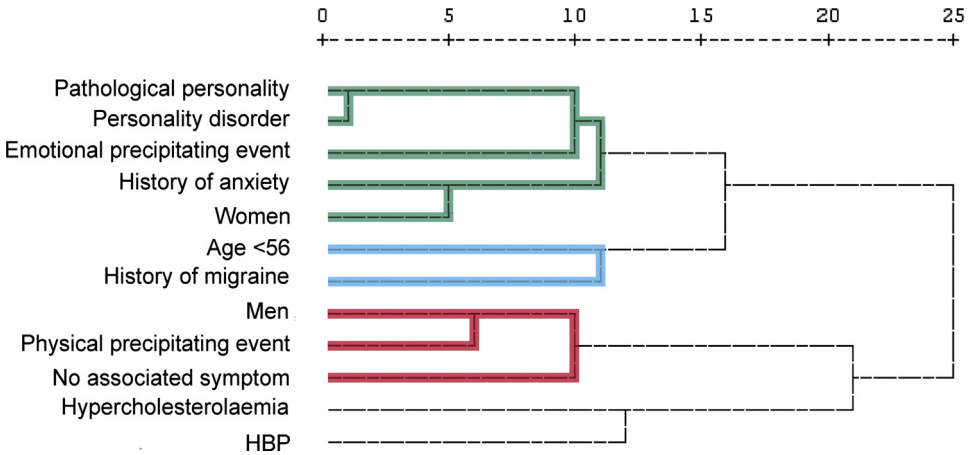

Figure 2 Cluster analysis of 63 TGA patients characterizing the three different classes of patients according to their clinical and epidemiological phenotype (Quinette et al., 2006). Used with permission from Oxford University Press. Modified after Bartsch and Deuschl (2010), used with permission from Elsevier.

contextual information) is thought to be supported by extrahippocampal brain regions. It appears to be preserved during the acute state (Jager et al., 2009b). This differentiation between recollection and familiarity is explained by dual-process models that postulate two distinct memory phenomena, which contribute to the recognition of personal experiences; however, alternative models have been suggested as well (Mandler, 1980).

Spatial memory is a further subtype of declarative memory and can also be impaired during TGA. Patients navigating through a virtual water maze performed more poorly than follow-up or healthy controls. Place and reversal learning and retrieval were impaired (Bartsch et al., 2010). Hainselin et al. (2011) further revealed an impairment of prospective memory during TGA, which describes the ability to remember a previously determined intention to act at the right time.

Several memory functions in TGA are preserved during the acute phase, such as priming, sensitization, habituation, and Pavlovian (classical) conditioning. These implicit, non-declarative memory functions are not essentially dependent on hippocampal functioning. Findings concerning affection of working memory during TGA are mixed due to the diversity of paradigms. Working memory is responsible for temporary information storage. It is necessary for complex cognitive tasks such as semantic speech, comprehension, forming a mental representation of the actual environment, learning, and reasoning and is thought to rely on prefrontal processing (Ngo et al., 2012). Working memory is considered preserved in the acute TGA state but there is evidence for impairment in specific executive functions regarding the limits of short-term memory (Quinette et al., 2003).

Taken together, the neuropsychological profile in TGA is a reflection of a hippocampal impairment.

Pathophysiological Mechanisms

The exact pathophysiological mechanisms of TGA are still unclear. The precipitant events, albeit heterogenous, can be subsumed under three higher order clusters that might shed light on the underlying pathophysiological mechanisms: emotional stress, physical effort as well as abrupt temperature change. Three pathogenic main hypotheses have been formulated: a link to migraine-related mechanisms, TGA as an epileptic phenomenon and a relation to hypoxic-ischemic events. Additionally, TGA has been linked to venous flow abnormalities and psychically induced mechanisms (Bartsch, 2012).

TGA and Migraine

Due to its ictal nature, migraine-associated mechanisms have been suggested to be involved in the pathophysiology of TGA (Olivarius and Jensen, 1979; Crowell et al., 1984; Schmidtke and Ehmsen, 1998). Several studies found a higher incidence of migraine in TGA patients (Olesen and Jorgensen, 1986; Schmidtke and Ehmsen, 1998). A prior history of migraine is assumed to be a risk factor for TGA patients younger than 56 years of age (Quinette et al., 2006). Nevertheless, TGA does not seem to reflect an episode of acute migraine because very few patients experience an acute migraine attack before or during a TGA. A key mechanism in the pathophysiology of migraine with aura is a transient wavelike spreading decrease of neuronal activity called 'cortical spreading depression' (CSD). This is a glutamate-mediated depolarization propagating across the cortex at a rate of 3–5 mm min^{-1} in which hyperperfusion is followed by a cortical hypoperfusion. This reversible process can be observed in several pathological conditions such as epilepsy and stroke. In animals, a CSD can also originate in the hippocampus, leading to an altered excitability of CA1 neurons amongst others (Wernsmann et al., 2006). These phenomena could theoretically act as a neurophysiological trigger evoking cellular metabolic changes that eventually result in a TGA, although this has not been shown in humans (Olesen and Jorgensen, 1986). Bartsch and Deuschl (2010) argue that the findings from Wernsmann et al. (2006) based on rodent studies cannot easily be translated to humans given the substantially higher threshold for eliciting a CSD in the hippocampus compared to the neocortex. Moreover, a hippocampal-elicited CSD has not been shown so far in humans. Additionally, patients during an acute TGA state do not exhibit symptoms typical for an acute migraine attack. In summary, the link between TGA and migraine-related mechanisms is controversial. In a comparison between TGA patients and healthy controls, Schmidtke and Ehmsen (1998) did not find an association between TGA and migraine.

TGA and Epilepsy

Certain similarities between TGA and epilepsy suggest a TGA being an epileptic phenomenon. In patients suffering from transient epileptic amnesia anterograde ictal amnesia and persistent retrograde amnesia can be detected. However, most EEG studies in TGA do not confirm this assumption. Epileptic activity was detected neither during the acute TGA state (Bartsch and Butler, 2013) nor after TGA (Jacome, 1989). In addition, relevant symptoms of epilepsy during the acute TGA phase are lacking, such as impaired consciousness and other cortical dysfunctions, automatisms, staring, and motor signs. Also temporal factors (the comparatively short duration of epileptic seizures) and its high rate of recurrence deviate from TGA. Considering the evidence, an epileptic origin of TGA seems rather unlikely.

TGA and Hypoxic-Ischemic Events

Some studies suggested an association between TGA and stroke in terms of a hemodynamic or thromboembolic event (Logan and Sherman, 1983; Felix et al., 2005; and Di Filippo and Calabresi, 2007). Accordingly, TGA came into consideration as a manifestation of a transient cerebral ischemia in memory-relevant structures (Jensen and De Fine Olivarius, 1980). It has also been suggested that in a subgroup of TGA patients, Valsalva-like activities may elicit a cerebral venous congestion (Winbeck et al., 2005). TGA patients show a low incidence of cerebrovascular events after the acute TGA state (Miller et al., 1987; Zorzon et al., 1995; Pantoni et al., 2005). A comparison between TGA patients and patients with reversible ischemic

attacks or lacunar syndrome revealed a statistically significant better prognosis for TGA patients (Haberman, 1984). Enzinger et al. (2008) compared vascular risk factors, magnetic resonance imaging markers of cerebral small-vessel disease, and other evidence of a cerebrovascular disorder between TGA patients and sex- and age-matched healthy controls. Because of a comparable risk factor profile and concomitant changes on brain MRI a cerebrovascular etiology of TGA could not be supported. In terms of the development of dementia, the findings from Gandolfo suggest no enhanced risk for TGA patients. Winbeck et al. (2005) detected an increased intima-media thickness in the common carotid arteries and a higher prevalence of carotid plaques in patients with hippocampal diffusion lesions combining duplex sonography and MRI. This outcome may indicate a higher incidence of atherosclerosis. In contrast, Pantoni et al. (2000) point out that laboratory investigations in TGA patients usually reveal slight abnormalities that presumably correspond to normal aging effects. Several studies compared the cerebrovascular risk factor profile of TGA patients with patients who had a TIA (Hodges and Warlow, 1990a,b; Zorzon et al., 1995) and failed to provide evidence of a connection of high stroke risk factors and TGA. In their metaanalysis, Quinette et al. (2006) confirmed this good outcome.

Stress and TGA

Already in the 1950s (Bender, 1956) a psychogenic origin of TGA was considered. Emotional arousal and stress was identified as a precipitant event in a relevant proportion of patients ranging from 14 (Hodges and Warlow, 1990a,b) to 29% (Quinette et al., 2006). Some patients exhibit emotional changes such as anxiety, helplessness, and depressed mood during the acute phase. Moreover, symptoms of a panic attack (dizziness, palpitations, chest pains, and nausea) have been observed (Inzitari et al., 1997). About 40% of the patients in Pantoni et al.'s (2000) study had symptoms that are typical for panic attacks. They state these patients could be particularly predisposed to hyperventilation during emotionally arousing events, leading to transient changes in cerebral blood flow. Situational circumstances seem to play a role as in about a half of the patients experience a TGA relapse under similar local/situational conditions as during the first TGA attack (Pantoni et al., 2000).

Some authors suggest that states of vulnerability in terms of personality traits like anxiety, phobic attitudes, pathological avoidance behavior, or depressive symptoms as predisposing factors for a TGA (Yesavage, 1983; Neri et al., 1995; Inzitari et al., 1997). Pantoni et al. (2000) point out that psychiatric disorders are more common in the family history of TGA patients compared to controls. It is worth mentioning that more women are affected by an emotionally induced TGA than men, possibly due to their increased vulnerability to affective disorders (Quinette et al., 2006).

Various studies about the impact of emotional changes (depression, anxiety, and stress) on the hippocampus revealed anatomical and functional modifications, leading to certain memory deficits. The hippocampus plays a central role in the glucocorticoid feedback-loop. Acute emotional and behavioral stress enhances the glutamatergic transmission and increases the intracellular calcium influx in hippocampal CA1 neurons. Joels et al. (2009) considers the increased calcium exposure as a potential risk factor for CA1 neurons that elicits a metabolic vulnerability and leads to an impairment of its structural cellular integrity. In the rodent hippocampus, acute stress causes a reduction of long-term potentiation as well as an enhancement of long-term depression, thus affecting hippocampus-dependent memory circuits (Howland and Wang, 2008).

Neuroimaging in TGA

DWI and other imaging techniques (positron emission tomography (PET), single photon emission computed tomography (SPECT)) provide the tool for studying hemodynamic and metabolic changes in mesiotemporal regions that are affected in TGA. The number of reported changes in these studies, however, varies considerably probably due to the sensitivity of the imaging technique and the study time relative to the TGA attack. Several SPECT studies conducted during TGA revealed cerebral blood flow reductions in MTL structures (Stillhard et al., 1990; Takeuchi et al., 1998; Venneri and Caffarra, 1998), however but also in (pre)frontal, occipital, pre- and postcentral areas, and other anatomical structures including the thalamus, cerebellum, amygdala, nucleus lentiformis, and striatum (Lin et al., 1993; Sakashita et al., 1997; Schmidtke et al., 1998; Eustache et al., 2000). Generally, in follow-up analyses the cerebral hemodynamics have returned to a normal state (Lampl et al., 2004).

However, the advent of high-resolution imaging techniques such as magnetic resonance imaging provides a step forward in the anatomical analysis of TGA. Two to three days after the episode, punctuate lesions (1–5 mm) can be detected in the CA1 region of the hippocampal cornu ammonis (Bartsch and Deuschl, 2010) (Figure 3). The detectability peaks between day 2 and 3 after the acute phase. On magnetic resonance spectroscopy, the emergence of lactate in these lesions could be shown that are interpreted as metabolic cellular stress (Bartsch et al., 2008). Lesions resolve spontaneously in the postacute phase within 14 days. Usually lesions outside the hippocampus are not detected.

Due to the fact that neurons of the CA1 region of the cornu ammonis play a key role in the formation and consolidation of memory, these lesions constitute the neuronal correlate of the memory deficit in terms of a transient perturbation of CA1 networks, presumably leading to a functional disconnection of the hippocampus. This assumption is supported by two single case MRI studies during the acute TGA state that uncovered reduced or no activation in temporal lobe structures during encoding and recognition of visual scenes (LaBar et al., 2002; Westmacott et al., 2008).

Summary

TGA is a rare amnesic syndrome known since half a century. The clinical onset is subacute, and mainly characterized by a severe anterograde and retrograde amnesia. After a few hours, symptoms typically resolve. It is diagnosed on clinical grounds and its

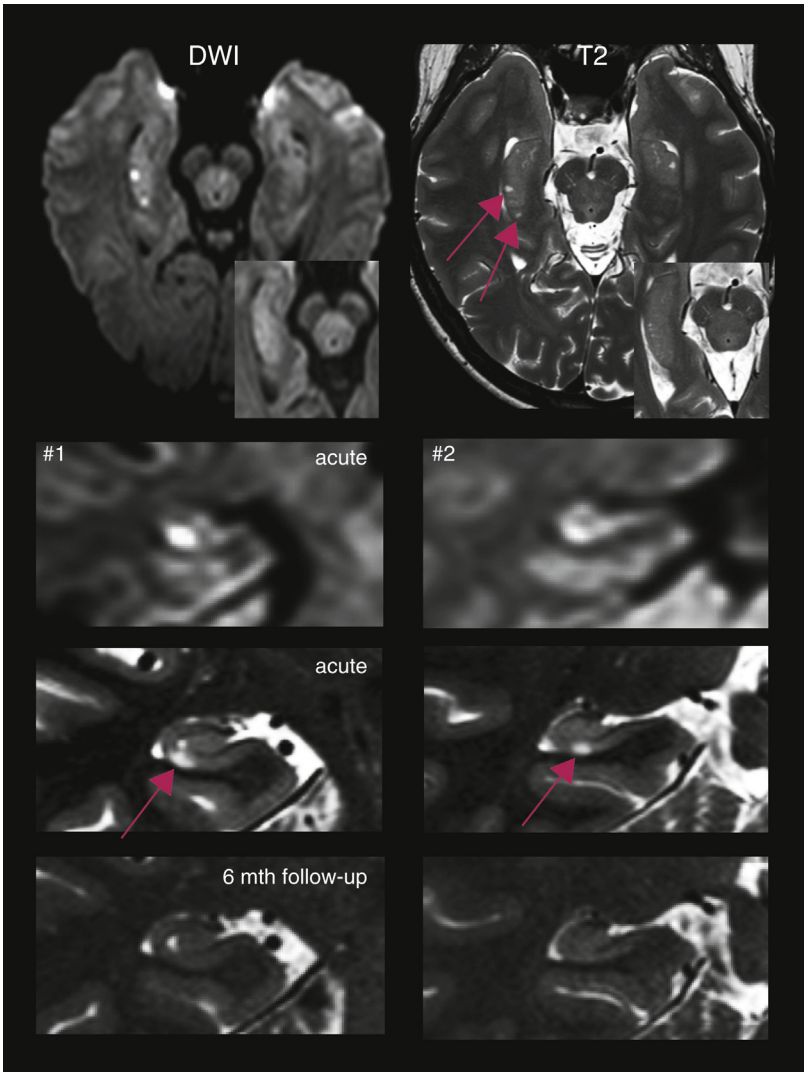

Figure 3 High-resolution MRI (3T) in TGA showing typical hippocampal lesions in the CA1 region of the cornu ammonis. Follow-up 6 months after the TGA shows a complete resolution of lesions. Modified after Bartsch and Deuschl (2010), used with permission from Elsevier.

outcome is considered benign. Intellect, personality, and other neuropsychological and cognitive functions remain unaffected. In many cases, certain precipitant events are reported that form clusters referring to physical effort, emotional arousal, or changes in body homeostasis. The etiology of TGA is still cryptic, but hypoxic-ischemic events, migrainelike mechanisms, epileptic activity, and stresslike mechanisms were taken into consideration. Using high-resolution imaging techniques functional correlates of the transient amnesia have been identified: in many patients, focal lesions confined to the CA1 sector of the hippocampus could be detected, a region that plays a major role in the functioning of episodic and autobiographical memory, place memory, prospective memory, and others. However, this region is also vulnerable to oxidative and metabolic stress.

See also: Amnesia: General; Amnesia: Psychogenic; Declarative Memory, Neural Basis of; Episodic Memory; Hippocampal Amnesia; Hippocampus and Related Structures; Learning and Memory, Neural Basis of.

Bibliography

Bartsch, T., 2012. The hippocampus in neurological disease. In: Bartsch, T. (Ed.), The Clinical Neurobiology of the Hippocampus. Oxford University Press, Oxford.

Bartsch, T., Alfke, K., Stingele, R., Rohr, A., Freitag-Wolf, S., Jansen, O., Deuschl, G., 2006. Selective affection of hippocampal CA-1 neurons in patients with transient global amnesia without long-term sequelae. Brain 129 (Pt 11), 2874–2884.

Bartsch, T., Alfke, K., Wolff, S., Rohr, A., Jansen, O., Deuschl, G., 2008. Focal MR spectroscopy of hippocampal CA-1 lesions in transient global amnesia. Neurology 70 (13), 1030–1035.

Bartsch, T., Butler, C., 2013. Transient amnesic syndromes. Nature Reviews Neurology 9 (2), 86–97.

Bartsch, T., Deuschl, G., 2010. Transient global amnesia: functional anatomy and clinical implications. Lancet Neurology 9, 205–214.

Bartsch, T., Dohring, J., Rohr, A., Jansen, O., Deuschl, G., 2011. CA1 neurons in the human hippocampus are critical for autobiographical memory, mental time travel,

and autonoetic consciousness. Proceedings of the National Academy of Sciences of the United States of America 108 (42), 17562–17567.
Bartsch, T., Schonfeld, R., Muller, F.J., Alfke, K., Leplow, B., Aldenhoff, J., Deuschl, G., Koch, J.M., 2010. Focal lesions of human hippocampal CA1 neurons in transient global amnesia impair place memory. Science 328 (5984), 1412–1415.
Bender, M.B., 1956. Syndrome of isolated episode of confusion with amnesia. Journal of the Hillside Hospital 5, 212–215.
Berli, R., Hutter, A., Waespe, W., Bachli, E.B., 2009. Transient global amnesia – not so rare after all. Swiss Medical Weekly 139 (19–20), 288–292.
Butler, C.R., Graham, K.S., Hodges, J.R., Kapur, N., Wardlaw, J.M., Zeman, A.Z., 2007. The syndrome of transient epileptic amnesia. Annals of Neurology 61 (6), 587–598.
Caffarra, P., Moretti, G., Mazzucchi, A., Parma, M., 1981. Neuropsychological testing during a transient global amnesia episode and its follow-up. Acta Neurologica Scandinavica 63 (1), 44–50.
Caplan, L.R., 1985. Transient global amnesia. In: Vinken, P.J., Bruyn, G.W. (Eds.), Handbook of Clinical Neurology. Elsevier Science, Amsterdam, pp. 205–218.
Caplan, L.R., 1990. Characteristic features and overview. In: Markowitsch, H.J. (Ed.), Transient Global Amnesia and Related Disorders. Hogrefe & Huber, Toronto, pp. 15–27.
Crowell, G.F., Stump, D.A., Biller, J., McHenry Jr., L.C., Toole, J.F., 1984. The transient global amnesia-migraine connection. Archives of Neurology 41 (1), 75–79.
Di Filippo, M., Calabresi, P., 2007. Ischemic bilateral hippocampal dysfunction during transient global amnesia. Neurology 69 (5), 493.
Enzinger, C., Thimary, F., Kapeller, P., Ropele, S., Schmidt, R., Ebner, F., Fazekas, F., 2008. Transient global amnesia: diffusion-weighted imaging lesions and cerebrovascular disease. Stroke 39 (8), 2219–2225.
Eustache, F., Desgranges, B., Aupee, A.M., Guillery, B., Baron, J.C., 2000. Functional neuroanatomy of amnesia: positron emission tomography studies. Microscopy Research and Techniques 51 (1), 94–100.
Eustache, F., Desgranges, B., Laville, P., Guillery, B., Lalevee, C., Schaeffer, S., de la Sayette, V., Iglesias, S., Baron, J.C., Viader, F., 1999. Episodic memory in transient global amnesia: encoding, storage, or retrieval deficit? Journal of Neurology Neurosurgery and Psychiatry 66 (2), 148–154.
Felix, M.M., Castro, L.H., Maia Jr., A.C., da Rocha, A.J., 2005. Evidence of acute ischemic tissue change in transient global amnesia in magnetic resonance imaging: case report and literature review. Journal of Neuroimaging 15 (2), 203–205.
Fisher, C.M., Adams, R.D., 1964. Transient global amnesia. Acta Neurologica Scandinavica Supplementum 40 (Suppl. 9), 1–83.
Fredericks, J.A.M., 1990. Transient global amnesia: an amnesic TIA. In: Markowitsch, H.J. (Ed.), Transient Global Amnesia and Related Disorders. Hogrefe & Huber, Toronto, pp. 28–47.
Gandolfo, C., Caponnetto, C., Conti, M., Dagnino, N., Del Sette, M., Primavera, A., 1992. Prognosis of transient global amnesia: a long-term follow-up study. European Neurology 32 (1), 52–57.
Gil, R., Abdul-Samad, F., Mathis, S., Neau, J.P., 2010. Was there a confusion before 1950 between global transient global amnesia and psychogenic amnesia? Revue Neurologique (Paris) 166 (8–9), 699–703.
Guillery-Girard, B., Desgranges, B., Urban, C., Piolino, P., de la Sayette, V., Eustache, F., 2004. The dynamic time course of memory recovery in transient global amnesia. Journal of Neurology Neurosurgery and Psychiatry 75 (11), 1532–1540.
Guillery-Girard, B., Quinette, P., Desgranges, B., Piolino, P., Viader, F., de la Sayette, V., Eustache, F., 2006. Long-term memory following transient global amnesia: an investigation of episodic and semantic memory. Acta Neurologica Scandinavica 114 (5), 329–333.
Guyotat, M., Courjon, J., 1956. Les ictus amnésiques. Journal de Medicine de Lyon 37, 697–701.
Haberman, S., 1984. Long-term prognosis after transient ischaemic attacks. Neuroepidemiology 3 (2–3), 108–122.
Hainselin, M., Quinette, P., Desgranges, B., Martinaud, O., Hannequin, D., de La Sayette, V., Viader, F., Eustache, F., 2011. Can we remember future actions yet forget the last two minutes? Study in transient global amnesia. Journal of Cognitive Neuroscience 23 (12), 4138–4149.
Hodges, J.R., Warlow, C.P., 1990a. The aetiology of transient global amnesia. A case-control study of 114 cases with prospective follow-up. Brain 113 (Pt 3), 639–657.
Hodges, J.R., Warlow, C.P., 1990b. Syndromes of transient amnesia: towards a classification. A study of 153 cases. Journal of Neurology Neurosurgery and Psychiatry 53, 834–843.
Howland, J.G., Wang, Y.T., 2008. Synaptic plasticity in learning and memory: stress effects in the hippocampus. Progress in brain research 169, 145–158.
Inzitari, D., Pantoni, L., Lamassa, M., Pallanti, S., Pracucci, G., Marini, P., 1997. Emotional arousal and phobia in transient global amnesia. Archives of Neurology 54 (7), 866–873.
Jacome, D.E., 1989. EEG features in transient global amnesia. Clinical Electroencephalography 20 (3), 183–192.
Jager, T., Bazner, H., Kliegel, M., Szabo, K., Hennerici, M.G., 2009a. The transience and nature of cognitive impairments in transient global amnesia: a meta-analysis. Journal of Clinical and Experimental Neuropsychology 31 (1), 8–19.
Jager, T., Szabo, K., Griebe, M., Bazner, H., Moller, J., Hennerici, M.G., 2009b. Selective disruption of hippocampus-mediated recognition memory processes after episodes of transient global amnesia. Neuropsychologia 47 (1), 70–76.
Jensen, T.S., De Fine Olivarius, B., 1980. Transient global amnesia as a manifestation of transient cerebral ischemia. Acta Neurologica Scandinavica 61 (2), 115–124.
Joels, M., Krugers, H.J., Lucassen, P.J., Karst, H., 2009. Corticosteroid effects on cellular physiology of limbic cells. Brain Research 1293, 91–100.
Kessler, J., Markowitsch, H.J., Rudolf, J., Heiss, W.D., 2001. Continuing cognitive impairment after isolated transient global amnesia. International Journal of Neuroscience 106 (3–4), 159–168.
Koski, K.J., Marttila, R.J., 1990. Transient global amnesia: incidence in an urban population. Acta Neurologica Scandinavica 81 (4), 358–360.
LaBar, K.S., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., 2002. Functional changes in temporal lobe activity during transient global amnesia. Neurology 58 (4), 638–641.
Lampl, Y., Sadeh, M., Lorberboym, M., 2004. Transient global amnesia – not always a benign process. Acta Neurologica Scandinavica 110 (2), 75–79.
Lin, K.N., Liu, R.S., Yeh, T.P., Wang, S.J., Liu, H.C., 1993. Posterior ischemia during an attack of transient global amnesia. Stroke 24 (7), 1093–1095.
Logan, W., Sherman, D.G., 1983. Transient global amnesia. Stroke 14 (6), 1005–1007.
Mandler, G., 1980. Recognizing: the judgment of previous occurrence. Psychological Review 87 (3), 251–271.
Melo, T.P., Ferro, J.M., Ferro, H., 1992. Transient global amnesia. A case control study. Brain 115 (Pt 1), 261–270.
Merriam, A.E., Wyszynski, B., Betzler, T., 1992. Emotional arousal-induced transient global amnesia. A clue to the neural transcription of emotion? Psychosomatics 33 (1), 109–113.
Miller, J.W., Petersen, R.C., Metter, E.J., Millikan, C.H., Yanagihara, T., 1987. Transient global amnesia: clinical characteristics and prognosis. Neurology 37 (5), 733–737.
Neri, M., Andermarcher, E., De Vreese, L.P., Rubichi, S., Sacchet, C., Cipolli, C., 1995. Transient global amnesia: memory and metamemory. Aging (Milano) 7 (6), 423–429.
Ngo, H.V., Claussen, J.C., Born, J., Molle, M., 2012. Induction of slow oscillations by rhythmic acoustic stimulation. Journal of Sleep Research.
Noel, A., Quinette, P., Guillery-Girard, B., Dayan, J., Piolino, P., Marquis, S., de la Sayette, V., Viader, F., Desgranges, B., Eustache, F., 2008. Psychopathological factors, memory disorders and transient global amnesia. British Journal of Psychiatry 193 (2), 145–151.
Olesen, J., Jorgensen, M.B., 1986. Leao's spreading depression in the hippocampus explains transient global amnesia. A hypothesis. Acta Neurologica Scandinavica 73 (2), 219–220.
Olivarius, B.D., Jensen, T.S., 1979. Transient global amnesia in migraine. Headache 19 (6), 335–338.
Owen, D., Paranandi, B., Sivakumar, R., Seevaratnam, M., 2007. Classical diseases revisited: transient global amnesia. Postgraduate Medical Journal 83 (978), 236–239.
Pantoni, L., Bertini, E., Lamassa, M., Pracucci, G., Inzitari, D., 2005. Clinical features, risk factors, and prognosis in transient global amnesia: a follow-up study. European Journal of Neurology 12 (5), 350–356.
Pantoni, L., Lamassa, M., Inzitari, D., 2000. Transient global amnesia: a review emphasizing pathogenic aspects. Acta Neurologica Scandinavica 102 (5), 275–283.
Quinette, P., Guillery, B., Desgranges, B., de la Sayette, V., Viader, F., Eustache, F., 2003. Working memory and executive functions in transient global amnesia. Brain 126 (Pt 9), 1917–1934.
Quinette, P., Guillery-Girard, B., Dayan, J., de la Sayette, V., Marquis, S., Viader, F., Desgranges, B., Eustache, F., 2006. What does transient global amnesia really mean? Review of the literature and thorough study of 142 cases. Brain 129 (Pt 7), 1640–1658.
Ribot, T., 1881. Les maladies de la mémoire. Bailliére, Paris.
Ribot, T., 1882. Diseases of Memory. Appleton, New York.
Sakashita, Y., Kanai, M., Sugimoto, T., Taki, S., Takamori, M., 1997. Changes in cerebral blood flow and vasoreactivity in response to acetazolamide in patients with transient global amnesia. Journal of Neurology Neurosurgery and Psychiatry 63 (5), 605–610.
Sander, K., Sander, D., 2005. New insights into transient global amnesia: recent imaging and clinical findings. Lancet Neurology 4 (7), 437–444.
Schmidtke, K., Ehmsen, L., 1998. Transient global amnesia and migraine. A case control study. European Neurology 40 (1), 9–14.
Schmidtke, K., Reinhardt, M., Krause, T., 1998. Cerebral perfusion during transient global amnesia: findings with HMPAO SPECT. Journal of Nuclear Medicine 39 (1), 155–159.

Sollier, P., 1892. Les troubles de la mémoire. Rueff, Paris.
Stillhard, G., Landis, T., Schiess, R., Regard, M., Sialer, G., 1990. Bitemporal hypoperfusion in transient global amnesia: 99m-Tc-HM-PAO SPECT and neuropsychological findings during and after an attack. Journal of Neurology, Neurosurgery and Psychiatry 53 (4), 339–342.
Takeuchi, R., Yonekura, Y., Matsuda, H., Nishimura, Y., Tanaka, H., Ohta, H., Sakahara, H., Konishi, J., 1998. Resting and acetazolamide-challenged technetium-99m-ECD SPECT in transient global amnesia. Journal of Nuclear Medicine 39 (8), 1360–1362.
Uttner, I., Prexl, S., Freund, W., Unrath, A., Bengel, D., Huber, R., 2012. Long-term outcome in transient global amnesia patients with and without focal hyperintensities in the CA1 region of the hippocampus. European Neurology 67 (3), 155–160.
Venneri, A., Caffarra, P., 1998. Transient autobiographic amnesia: EEG and single-photon emission CT evidence of an organic etiology. Neurology 50 (1), 186–191.
Wernsmann, B., Pape, H.C., Speckmann, E.J., Gorji, A., 2006. Effect of cortical spreading depression on synaptic transmission of rat hippocampal tissues. European Journal of Neuroscience 23 (5), 1103–1110.
Westmacott, R., Silver, F.L., McAndrews, M.P., 2008. Understanding medial temporal activation in memory tasks: evidence from fMRI of encoding and recognition in a case of transient global amnesia. Hippocampus 18 (3), 317–325.
Winbeck, K., Etgen, T., von Einsiedel, H.G., Rottinger, M., Sander, D., 2005. DWI in transient global amnesia and TIA: proposal for an ischaemic origin of TGA. Journal of Neurology, Neurosurgery and Psychiatry 76 (3), 438–441.
Yesavage, J.A., 1983. Direct and indirect hostility and self-destructive behavior by hospitalized depressives. Acta Psychiatrica Scandinavica 68 (5), 345–350.
Zeman, A., Butler, C., 2010. Transient epileptic amnesia. Current Opinion in Neurology 23 (6), 610–616.
Zeman, A.Z., Boniface, S.J., Hodges, J.R., 1998. Transient epileptic amnesia: a description of the clinical and neuropsychological features in 10 cases and a review of the literature. Journal of Neurology, Neurosurgery and Psychiatry 64 (4), 435–443.
Zorzon, M., Antonutti, L., Mase, G., Biasutti, E., Vitrani, B., Cazzato, G., 1995. Transient global amnesia and transient ischemic attack. Natural history, vascular risk factors, and associated conditions. Stroke 26 (9), 1536–1542.

Transition to Adulthood

Maria Sironi, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The process by which adolescents and teenagers face the transition to adulthood has changed substantially in the Western developed world. This is one of the demographic transformations that have occurred in the last 50 years, and can be framed in the second demographic transition theory. In this work the main changes in the life course trajectories of young adults are highlighted, and the differences across gender and social class are described. Also, the role of institutions and culture in shaping the transition to adulthood is taken into account.

Introduction

The second half of the twentieth century has been characterized by substantial changes in demographic behaviors. Developed countries have witnessed large decreases in fertility, with a very recent reversal of the decline (Myrskyla et al., 2009), a strong reduction in marriage rates or at least a postponement of family formation, and an increase in the prevalence of cohabitation and divorces (*see* Partnership Formation and Dissolution in Western Societies). Among these important transformations, also the process by which adolescents and teenagers transition to adulthood has changed greatly in many – if not all – countries in the Western developed world. Typically, the demographic events of adult transitions have included such markers as leaving home, finishing education, securing a job, marrying or cohabiting, and having children. In the postwar years, all the events in the transition occurred quickly (usually by 25 years of age for both men and women) and in an orderly sequence beginning with school completion, full-time work, and home leaving. Since then, the transition has lengthened and become more complex.

All the events of the transition to adulthood have been delayed (Aassve et al., 2002; Furstenberg, 2010), and it is important to take into account their interdependence and strong interactions: there has been a substantial expansion of education, especially among women, leading to an increase in the age at school completion. This increase may be correlated to a later entrance in the labor market and a delay in leaving the parental home – at least in some contexts (e.g., southern Europe). A postponement in the achievement of independence (monetary and nonmonetary) from the family of origin can be to some extent associated to or be responsible for a later age at family formation, i.e., cohabitation, marriage, and parenthood.

Moreover, individuals had to adapt their life course trajectories because of profound structural and cultural changes (such as the globalization process and a decline in the normative imperative to marry) occurred in developed societies in the last decades. As a consequence, these trajectories have become more diverse. The 'second demographic transition' theory would use the term *individualization* to characterize changes in the life course (Lesthaeghe, 1995), which includes many different aspects, such as the deinstitutionalization, the destandardization, and the differentiation in the life trajectories of young adults. The order and the timing of events may vary considerably between individuals, as well as the duration in different states (*see* Life Course: Sociological Aspects). When studying all these trends it is necessary to consider that changes are not uniform across countries (and welfare states), across social classes and there are also large differences by gender (Buchmann and Kriesi, 2011; Sironi and Furstenberg, 2012).

There is much public debate about why the transition to adulthood changed so suddenly, and social scientists have not yet reached consensus on this question. An increase in the years of schooling may cause a delayed entry in the labor market; over time the structure of the labor market itself could have changed so that well-paid jobs require more education and are therefore more difficult for young adults to find (Danziger and Ratner, 2010). Increases in earnings inequality, employment instability, and shortening of job tenure due to higher turnover are some of the most relevant changes that have occurred since the 1970s. Together with these structural changes, youths' expectations about what it takes to live independently have risen, along with the actual costs of living. Moreover, the normative and cultural shifts that have taken place in the second half of the twentieth century may also contribute to an increase in the heterogeneity characterizing individual life courses (Arnett, 2000; Lesthaeghe, 1995; Van de Kaa, 2001). The post-1960s shift in sexual attitudes and practices, for example, has also slowed what was once a rush into adult roles. Fifty years ago, premarital sex was still highly stigmatized. Although the stigma did not deter many young couples from breaching the norms, marriage served as a safety net in the event of a premarital pregnancy. Also, although today's young adults and their parents' value independence highly, both tolerate and even endorse a slower schedule for attaining economic and social maturity (Arnett, 2000).

The transition process has traditionally been considered as complete when an individual has experienced all events – completion of education, labor market entry, leaving the parental home, forming a union, and entry into parenthood. This idea has been questioned in recent years, for example, by Arnett's (2000) concept of emerging adulthood, a new developmental stage in the life course. Between adolescence and adulthood, this stage is characterized by long-lasting explorations of identity, lifestyles, and career possibilities (Arnett, 2000; Buchmann and Kriesi, 2011).

Recent Findings: Have the Changes Been Equal for Everyone?

Overview

Recent studies on the transition to adulthood in Europe and, more generally, in developed countries, found some common trends. First of all, the process leading to adulthood has been prolonged. This means that not only most of the events have been delayed, but also that the time span from the first event (usually the completion of education) to the last (usually childbearing) is now much longer (Aassve et al., 2002; Billari, 2004). Secondly, all countries have experienced increasing diversity at the individual level, the so-called trend of 'converging divergences' (Blossfeld et al., 2012; Mills et al., 2008).

The expansion of education, that has characterized the last five decades in developed countries, has raised the median age for completing schooling and contributed to the delay in the age at which young adults start working (Schizzerotto and Lucchini, 2004). The median age of completing their education among cohorts born in the 1950s in Great Britain, Germany, and Italy, for example, was 16 years, but it increased to 17 years for cohorts born in the 1970s in Great Britain and Germany and to almost 20 years in Italy (Buchmann and Kriesi, 2011; Schizzerotto and Lucchini, 2004). The proportion completing upper secondary (90%) and tertiary education (40%) is high in the United States among those born in the late 1960s and 1970s. The figures are even higher in Scandinavian countries, where the completion rate of tertiary education approached 50% in 2007 (OECD, 2009).

In addition, the structure of the labor market itself has changed significantly amid global competitive pressure, leading to instability of labor markets and greater difficulties in finding a job – another cause of the delay (Brzinsky-Fay, 2008). Since the 1970s, earnings inequality and employment instability have grown and job tenure has declined with growing turnover (Sironi and Furstenberg, 2012). Among people born in the 1950s, the median age at first job was between 16 and 19 years. Among those born in the 1970s, the median age was between 19 and 24 years in developed Western countries (Schizzerotto and Lucchini, 2004). Young adults in southern Europe often experience a period of career instability when they exit education, and youth unemployment is highest in those countries. The entry into the labor market is easier and faster in Continental countries, while liberal and Scandinavian countries occupy an intermediate position (Quintini et al., 2007). In fact, in northern Europe young people may experience high unemployment and uncertainty immediately after school completion, but only for a short time.

The departure from the parental home marks the beginning of social independence (if not financial). The age of leaving has increased with time, but the postponement has been modest in most countries. Southern Europe is the exception. There the median age of leaving has risen substantially. In northern Europe, young adults leave their family of origin very early, before 20 years of age. In continental and liberal welfare states, individuals leave on average in their early 20s, while in countries such as Italy, Spain, Portugal, and Greece, young people in their late 20s or early 30s (Aassve et al., 2002; Billari, 2004).

One of the main steps leading to independence is financial self-sufficiency. In many countries, in the 1990s, only a minority of young people is able to support themselves in their late teens and early 20s with their earnings alone. Even when social transfers were taken into account, a significant proportion of young people are unable to support themselves, much less a family, before their mid-to late 20s (Smeeding and Phillips, 2002). This issue has worsened over time, and some studies found a declining economic self-sufficiency in the last two decades. Only women in their late 20s and early 30s saw somewhat improved prospects, even though their earnings remain below men's (Sironi and Furstenberg, 2012).

Finally, a strong postponement and eventually a total foregoing, have been observed in family formation. The mean and median age at first marriage have increased in all developed countries, and more frequently marriage has been substituted by cohabitation (Billari, 2004). Age at first birth has increased by 3–5 years in the last four decades (Sobotka, 2008), and also the total fertility rate has decreased, even though not equally in all countries and with a recent reversal of the decline (Myrskyla et al., 2009).

Differences across Gender and Social Classes

Notwithstanding this general trend of postponement, the changes in the process leading to adulthood have not been uniform across gender and social class (*see* Class: Social; Social Stratification).

The relevance of family social class on the subsequent life course starts before birth, it continues throughout adolescence, and it is able to shape the course of young adult transitions and psychological development in the third and fourth decades of life (*see* Socialization in Infancy and Childhood). Youth from affluent and well-educated families marry and have children later in life because of a longer education, much more extended search for a permanent partner in life, and a lower incidence of unintended pregnancy (Furstenberg, 2010). Young adults from disadvantaged families, even if they go to college, do it with fewer resources and therefore face more difficulties in completing the degree. Moreover, they do not want and cannot afford to remain unemployed for too long, and consider education as a way to get a job. So they are more likely to drop out of school if they are able to find an occupation. Finally, the economic difficulties linked to the housing costs, may hinder the independent living before family formation. Family background can influence not only the timing of events in the transition to adulthood but also the sequencing of these events thus modifying the propensity to experience traditional or innovative patterns. For example, children from higher family social status, on the one hand, tend to postpone their first union and their first child birth (Rijken and Liefbroer, 2009); on the other hand, tend to acquire their housing autonomy earlier, without directly making the transition to live with a partner (Blaauboer and Mulder, 2010).

The employment and earnings changes of the past four decades have affected the transition to adulthood in ways that vary sharply by gender. Education expansion has been more substantial among women, who started from a lower level and caught up with men in earning college degrees and enrolling in graduate studies. Increases in the labor force participation rate have also been much more pronounced

among women (Danziger and Ratner, 2010; Sironi and Furstenberg, 2012). The changing economic situation of young men and women has implications for the second phase of adult transitions: union formation and parenthood. Career opportunities and earnings, in fact, affect men's and women's decisions about marriage and parenthood differently. Among men, steady employment and earnings are positively associated with marriage and childbearing. For women, the picture is less clear. The new home economics hypothesis (Becker, 1965) states that women's economic independence is the main reason behind delayed marriage and motherhood in industrialized countries. Gains from marriage and role specialization within marriage dissipate with women's growing investments in human capital and careers. The higher their level of education and the better their job opportunities, the more likely women are to postpone or even avoid marriage and motherhood. An alternative hypothesis, with different implications, is that women's earnings contribute to a couple's higher standard of living, which encourages marriage (income effect). Which scenario dominates is not clear yet and appears to depend on other factors, such as local and country context, birth cohort, and educational attainment (Harknett and Kuperberg, 2011).

Cross-National Differences: the Role of Institutions and Culture

The main issue concerning changes in the transition to adulthood over time is the lack of convergence in life course patterns in different countries (Liefbroer and Goldscheider, 2006). It has been argued that the lack of convergence may stem from differences in labor market regulations, welfare regimes (and so in the diversity underlying institutional structures), and cultural aspects such as family ties. Both institutions and culture influence life course trajectories, creating prospects and constraints to which people adapt (Breen and Buchmann, 2002).

Different types of welfare regimes can influence individuals' behavior; especially their 'demographic' decisions (Esping-Andersen, 1990) (*see* Welfare State, History of). The four regimes in the West include the *social democratic* welfare state (Scandinavian countries), characterized by high levels of state support and a focus on the individuals. The *conservative* welfare state provides generous support for the family rather than the individual (for example, in France and Germany) and thus reinforces family ties. The *liberal* welfare regime provides a modest level of means-tested benefits, applicable only in case of inefficient markets, and oriented to individuals (examples include the United Kingdom and the United States). The fourth category includes the southern European countries with their strong reliance on the family as the locus of support and low levels of welfare provisions from the state (Trifiletti, 1999). Eastern European (former communist) countries have not yet been systematically included in comparative research, and could be considered as a new and interesting research field.

Each regime is associated with a different transition to adulthood. In social democratic countries, for example, labor market entry and leaving the parental home should take place at relatively younger ages owing to generous social benefits, which provide more certainty about future economic conditions. In conservative and liberal countries, it is difficult to predict a typical life course trajectory, but one would expect a lower destandardization and variability of life courses than in Scandinavian countries. In contrast, southern European countries should show late exit from parental home, late marriage, and childbearing, given the lower level of benefits and resulting stronger dependence on the family of origin (Aassve et al., 2002; Billari, 2004; Esping-Andersen, 1990; Mills et al., 2008).

As far as family support is concerned and cultural differences more generally, the initial conditions from which modern societies have developed are fundamental for the present differences in behavior. More precisely, in southern Europe, the influence of Islam raised the importance of kinship and vertical relationship between generations, so that the prolonged stay of children in their parents' home and the caring work of children toward their parents are two faces of the same coin, a 'strong' family. In the north, Germanic tradition and the reformation contributed to the development of a 'weak' family (Reher, 2005). Empirical evidence supports these views, finding that the share of young adults who are financially dependent on their parents and family members is much larger in southern Europe (Billari, 2004). The cultural inheritance model (as the welfare regimes explanation) predicts that convergence will not take place at a national level: factors, which lead to homogenization of behaviors throughout the world, like those implied by modernization and economic globalization, will not have the same impact on societies with different cultural and historical origins, thus leading to diverse and possibly divergent coping behaviors and timing of transitions.

Macro Typologies of Life Course Trajectories

Based on the findings of the current literature on the transition to adulthood and on the possible determinants of cross-national differences just described, it is possible to identify four different macrocategories of life course trajectories. Each one includes different countries of the developed Western world. In order to make the distinctions easier to understand for the reader the characteristics of each typology are reported in **Table 1**.

It is important to note that these categories are not exhaustive of all the experiences reported in developed societies, and that part of the empirical evidence does not match well within these groups. This may be partly due to the influence of short-term economic fluctuations, recent policy changes, and the increase of migration flows that might alter the composition of national populations. Also, it is possible that the way in which welfare regimes and culture impact demographic behavior is not exactly as predicted by the theoretical arguments. Hence, further research is required to better understand the role of macro characteristics, as institutions and family ties.

Conclusions

The transition to adulthood is a complex concept, and the complexity has certainly increased over time with the emergence

Table 1 Typologies of the transition to adulthood

Group of countries	*Welfare state*	*Culture and family ties*	*Independence transition*	*Family formation*
Scandinavian countries	Welfare rights and benefits granted as individual entitlements; unemployment benefits available for young people.	Weak family ties that imply individual's priority over the family.	High completion rates of tertiary education and widespread vocational training promote easier labor market access. High importance of young people's autonomy, very early home leaving.	Early nontraditional union formation (mostly cohabitation); late age at marriage (if any). Great heterogeneity in the age at entry into parenthood.
Liberal countries	Strong market orientation, loosely regulated labor market; means-tested benefits oriented to individuals.	Weak family ties, strong focus on self-reliance and autonomy	Weak regulation of the labor market supports an early completion of schooling and easy access to the labor market (even with high instability); early home leaving, early school completion, and entry into the labor market.	The early 'independence' transition facilitates early family formation. Possible delay in family formation in times of greater instability of the labor market.
Continental/ Conservative countries	Welfare transfers oriented to families.	Stronger family ties than in liberal and Scandinavian countries that are reinforced by family-oriented welfare transfers.	Strong vocational education system; easy access to the labor market even if at later ages. Later age at home leaving and achievement of economic independence, but with more stability.	Family formation shortly after the achievement of financial self-sufficiency, given the greater stability in the labor market. The postponement of economic and employment stability may be a cause for delayed home leaving
Southern Europe	Weak welfare state transfers, mainly family-oriented.	Strong family ties; family is the locus of support for young people.	Late age at school completion and no vocational training hinder an easy access to stable jobs; strong senior worker protection; Very late age at home leaving.	Having a stable job and economic self-sufficiency are important conditions to form a family, so also age at union formation and parenthood are higher than in other countries.

of new and heterogeneous behaviors. What is clear is that a common trend is taking place in all developed societies: a deterioration in economic and employment conditions among young adults. This is true in all countries among men, while in some context there has been an improvement in women's participation to the labor market with some positive influences on their attainment of financial self-sufficiency (Sironi and Furstenberg, 2012).

This phenomenon, that started in the second half of twentieth century as part of the second demographic transition, is bringing to a convergence in life course trajectories across gender that may have important implications for the subsequent steps like family formation. The consequences of these changes in the first steps of the transition – that somewhere are not the same across gender – should be examined in future research, especially their effects on the timing of first unions and childbearing. A delay in the achievement of independence among men is likely to produce a delay in union formation and parenthood, while an increased education and labor force participation among women could have positive effects on this life domain. The positive effects are strictly dependent on the ability of policies to adapt to these changes, and to be flexible enough and allow women to combine work and family life (McDonald, 2000).

This issue is strongly linked to the methodologies that are usually used to study the transition to adulthood. Most studies, both comparative and noncomparative ones, focus on a specific phase of the life course trajectories, if not on one event only. Very often the departure from the parental home is the key event of the analysis, some research focus on the transition from school to work, and some others only on family formation. However, the life course paradigm assumes that individuals, as human agents, build their future on the basis of the constraints and opportunities experienced in the past. The process is iterative and cumulative, since initial advantages or disadvantages often are amplified with time (Giele and Elder, 1998). Life courses are embedded in different time and location and are affected by the social context in which individuals live. In addition, different life domains are strongly interdependent. Hence, it is important that life course analysis looks at the whole life trajectory. Under this perspective, characteristics such as type, number, duration, and order of events should be all taken into account. Only a limited number of studies adopt a complete life course perspective. Future research should exploit the already existing methodologies (e.g., sequence and latent class analysis), and also refine them or develop new ones that allow to capture the complexity that is inherent in a phenomenon like the transition to adulthood (Barban and Billari, 2012).

See also: Age Stratification; Age, Sociology of; Class: Social; Identity in Childhood and Adolescence; Life Course and Event History Analysis; Partnership Formation and Dissolution in Western Societies; Social Stratification; Socialization in Infancy and Childhood.

Bibliography

Aassve, Arnstein, Billari, Francesco C., Mazzucco, Stefano, Ongaro, Fausta, 2002. Leaving home: a comparative analysis of ECHP data. Journal of European Social Policy 12 (4), 259–275.

Arnett, Jeffrey J., 2000. Emerging adulthood. A theory of development from the late teens through the twenties. The American Psychologist 55 (5), 469–480.

Barban, Nicola, Billari, Francesco C., 2012. Classifying life course trajectories: a comparison of latent class and sequence analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics) 61 (5), 765–784.

Becker, Gary S., 1965. A theory of the allocation of time. The Economic Journal 75 (299), 493–517.

Billari, F.C., 2004. Becoming an adult in europe: a (macro/micro)-demographic perspective. Demographic Research Special Collections 3 (2), 15–44.

Blaauboer, Marjolein, Mulder, Clara H., 2010. Gender differences in the impact of family background on leaving the parental home. Journal of Housing and the Built Environment 25 (1), 53–71.

Blossfeld, Hans-Peter, Klijzing, Erik, Mills, Melinda, Kurz, Karin, 2012. Globalization, Uncertainty and Youth in Society: The Losers in a Globalizing World. Routledge.

Breen, Richard, Buchmann, Marlis, 2002. Institutional variation and the position of young people: a comparative perspective. The Annals of the American Academy of Political and Social Science 580 (1), 288–305.

Brzinsky-Fay, Christian, 2008. 11. Get the balance right: risk and flexibility in school-to-work transition sequences. Innovating European Labour Markets: Dynamics and Perspectives 275.

Buchmann, Marlis C., Kriesi, Irene, 2011. Transition to adulthood in europe. Annual Review of Sociology 37 (1), 481–503.

Danziger, Sheldon, Ratner, David, 2010. Labor market outcomes and the transition to adulthood. The Future of Children/Center for the Future of Children, the David and Lucile Packard Foundation 20 (1), 133–158.

Esping-Andersen, Gøsta, 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton, NJ.

Furstenberg, Frank F., 2010. On a new schedule: transitions to adulthood and family change. The Future of Children/Center for the Future of Children, the David and Lucile Packard Foundation 20 (1), 67–87.

Giele, J.Z., Elder, G.H., 1998. Methods of Life Course Research: Qualitative and Quantitative Approaches. SAGE Publications, Los Angeles, CA.

Harknett, Kristen, Kuperberg, Arielle, 2011. Education, labor markets, and the retreat from marriage. Social Forces 90 (1), 41–63.

Lesthaeghe, Ron J., 1995. In: Mason, K.O., Jensen, A.-M. (Eds.), The Second Demographic Transition in Western Countries: An Interpretation. Clarendon, Oxford.

Liefbroer, Aart, Goldscheider, Frances, 2006. Transitions to adulthood: how unique is Sweden the European context? In: Bernhardt, E.M., Goldscheider, C., Goldscheider, F., Bjerén, G. (Eds.), Immigration, Gender and Family Transitions to Adulthood in Sweden. University Press of America, Lanham, MD.

McDonald, Peter, 2000. Gender equity in theories of fertility transition. Population and Development Review 26 (3), 427–439.

Mills, Melinda, Blossfeld, Hans-Peter, Buchholz, Sandra, Hofäcker, Dirk, Bernardi, Fabrizio, Hofmeister, Heather, 2008. Converging divergences?: An international comparison of the impact of globalization on industrial relations and employment careers. International Sociology 23 (4), 561–595.

Myrskyla, Mikko, Kohler, Hans-Peter, Billari, Francesco C., 2009. Advances in development reverse fertility declines. Nature 460 (7256), 741–743.

OECD, 2009. Education at a Glance. OECD Indicators. OECD, Paris.

Quintini, Glenda, Martin, John P., Martin, Sébastien, 2007. The Changing Nature of the School-to-Work Transition Process in OECD Countries. WDA-HSG Discussion Paper No. 2007-2.

Reher, David, 2005. In: Zuanna, Gianpiero, Micheli, Giuseppe (Eds.), Family Ties in Western Europe Strong Family and Low Fertility: A Paradox? Springer, Netherlands, pp. 45–76.

Rijken, Arieke J., Liefbroer, Aart, 2009. Influences of the family of origin on the timing and quantum of fertility in the Netherlands. Population Studies 63 (1), 71–85.

Schizzerotto, Antonio, Lucchini, Mario, 2004. Transition to adulthood. In: Iacovou, M., Berthoud, R. (Eds.), Social Europe: Living Standards and Welfare States. Edward Elgar, Cheltenham, UK, pp. 46–68.

Sironi, Maria, Furstenberg, Frank F., 2012. Trends in economic independence of young adults in the United States: 1973–2007. Population and Development Review 38 (4), 609–630.

Smeeding, Timothy M., Phillips, Katherin R., 2002. Cross-national differences in employment and economic sufficiency. Annals of the American Academy of Political and Social Science 580, 103–133.

Sobotka, Tomas, 2008. Overview chapter 6: the diverse faces of the second demographic transition in europe. Demographic Research 19 (8), 171–224.

Trifiletti, Rossana, 1999. Southern European welfare regimes and the worsening position of women. Journal of European Social Policy 9 (1), 49–64.

Van de Kaa, D.J., 2001. Postmodern fertility preferences: from changing value orientation to new behavior. Population and Development Review 27, 290–331.

Transition to Adulthood, Cultural Differences in

Carol A Markstrom, Sera A Mathew, and Kalynn B Amick, College of Education and Human Services, West Virginia University, Morgantown, WV, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Of interest is the transition to adulthood in non-Western cultures according to culture-based mechanisms including rites of passage. Cultural markers of transition are reviewed according to those that are biologically based, those that include ceremonies and celebrations, and those that involve instruction. The changing nature of these markers is discussed in light of globalization.

Introduction

It is commonly stated that adolescence begins with biology and ends in culture. The biological mandates of physical growth and sexual maturation at the onset of adolescence are universal, but what is less uniform is the cultural management and meaning of biological pubescence. Of additional variability and ambiguity are the sociocultural indicators that signify the transition from adolescence to adulthood. The social experience of adolescence as a life stage has been shown to be nearly universal in Schlegel and Barry's (1991) landmark analysis using the standard cross-cultural sample. Further, a stage of youth exists in many cultures – a somewhat comparable time period applicable to later adolescence which today is called emerging adulthood (Arnett, 2007; Schlegel, 2009). The focus of this article is on the culture-based markers of the transition from adolescence to adulthood in current non-Western societies. Of particular interest is the degree to which traditional rites of passage and culture-based rituals are changing and modified due to globalization.

A useful starting point is to define what is meant by 'transition to adulthood.' The National Research Council and Institute of Medicine (NRCIM, 2005) assessed the sociodemographics of this transition in developing countries and pointedly concluded that the transition is presently experienced by more young people than ever. The social markers of adulthood are distinctive features accomplishing this transition compared to the biological markers of transitions of earlier stages of development. The shift from being dependent upon to becoming independent from parents is a key overall indicator of successful transition to adulthood as well as providing for one's own needs as well as the needs of others (Hamilton and Hamilton, 2009). It is through a society's institutions (i.e., education, employment, civil society, and family) that the transition from adolescence to adulthood occurs (Hamilton and Hamilton, 2009). Jeffrey and McDowell (2004) identify the changing global order – particularly related to economic factors – that makes today's transition to adulthood particularly complex.

Of further consideration is how the transition to adulthood varies between Western and non-Western cultures. In Western cultures, there are various chronological age markers conferring adult privileges, such as freedoms and responsibilities associated with driving, voting, consuming alcohol, and being accountable as an adult relative to the criminal justice system. The current trend is for adolescence to morph into emerging adulthood (approximately ages 18–25) during which time individuals acquire all societally derived adult privileges, but do not yet demonstrate complete adult autonomy in all areas (Arnett, 2007). Hence, while legal privileges are conferred early in this time frame, adult attainment relative to completing education, financial independence, and marriage and long-term emotional commitments, such as parenting, vary greatly. Cross-culturally, delays in the age of marriage are a key indicator that the transition to adulthood is now later than times past (Hamilton and Hamilton, 2009). This factor is heavily linked to economic viability which is a facilitator of marriage.

Emerging adulthood may not be an applicable stage concept in impoverished and developing countries' cultures due to constraints in time and resources for adolescent experimentation in relationships, lifestyles, and careers that limit opportunities to have and enact choices (Markstrom, 2011a). Further, emerging adulthood implies a quality of Western individualism which may not be consistent with communal/collectivistic values of some cultures considered in this article. Nonetheless, adolescents worldwide are becoming increasingly exposed to diverse pathways of careers and lifestyles. In cases of favorable economic conditions, emerging adulthood may acquire more relevance in some non-Western cultures.

Globalization is a weighty consideration in this article because most of the current writings on the transition to adulthood in non-Western cultures consider cultural changes due to globalization. Globalization and Westernization (one domain of influence of globalization) are advanced through present-day media exposure including the internet and social networking sites, as well as increased travel. Certainly, there is variability in access to global ideas due to socioeconomic status, rural/urban living status, educational opportunities, and related factors. However, with greater global awareness, losses relative to local language and cultural traditions may occur including those attached to the transition to adulthood.

Youth are active participants in globalization (Larson et al., 2009). Due to greater maturity and autonomy, unrestricted access to global ideas and products, social opportunities outside of the family, identity exploration, and the quest to position themselves in the world around them, adolescents are keenly interested in new ideas and opportunities (Jensen and

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23214-5

Arnett, 2012). The global media is a prime source for adolescents across cultures providing direct, unrestricted access to information (Schlegel, 2000). Due to trends and ideas in fashion, music, language, and ideas, the notion of a more pervasive and somewhat homogenized global youth culture is not far-fetched. Hence, identifications to local culture may be supplemented, transformed, or in some way altered leading to bicultural, multicultural, and hybrid forms of identity. Certainly, there is the potential for confusion as identity formation at the local level is managed vis-à-vis different and potentially conflicting information from global sources. Changes to the performances of traditional rites of passage are inevitable – we are not assuming that these rituals are simply relinquished but rather managed and transformed in various manners.

The next section reviews classic models of rites of passage within the topic of the transition to adulthood. Then, several illustrations of changing cultural practices or markers of this life transition are reviewed, i.e., traditional ceremonies and practices that serve to usher in the transition into adulthood. This article closes with overall discussion on the changing themes of transition to adulthood in non-Western cultures according to globalization.

Rites of Passage and the Transition to Adulthood

Of primary interest in this article are cultural markers of the transition to adulthood – specifically as occurring in non-Western cultures or subcultures within Western cultures. Cultural markers encompass traditions, ceremonies, and rites of passage that facilitate change at particular times in the life span. The term 'initiation' also applies and, particular to adolescence, Schlegel and Barry (1980) defined initiation as "some social recognition, in ceremonial form, of the transition from childhood into either adolescence or full adulthood" (p. 698). In traditional societies of the past, more often small societies, such ceremonies occurred anytime during late childhood or adolescence (8 and 18 years of age) with most occurring around puberty. At adolescence, the more specific functions of rites of passage surround the idea of growing to adulthood (see **Table 1** for a list of functions).

Initiations may take the form of rites of passage which are ritually linked activities that occur in conjunction with major life events, such as birth, coming-of-age, marriage, and death. They serve many functions that span psychological (e.g., altered self-perception), social (enhanced connections to the social group), and ethological (e.g., assignment and classification of gender roles) explanations (Schlegel and Barry, 1980). The classic van Gennep (1908/1960) rites of passage model includes three sequences of separation, transition, and incorporation. With separation, the initiate is symbolically or physically removed from the mundane rhythms of his or her society (either alone or with other initiates depending on the culture). Separation does not necessarily equate isolation, but rather it is some form of distancing from the everyday or ordinary in order to progress into the extraordinary of the transitional sequence of the rite of passage. Transition can be a psychologically discomforting experience because classification according to previous life identifications is not available nor has the individual been conferred to the next stage of life (Turner, 1967). A rite of passage in many North American Indigenous cultures was the vision quest or the Walkabout practiced among Aborigine Australians. Initiates were required to be physically isolated (separated) and engaged in fasting and prayer during the time of transition to facilitate spiritual transcendence through the acquisition of a vision. In some cultures, rituals entailing bodily adornments, such as special clothing or tattooing, prayers, and instruction and mentorship also occur during transition (see **Table 2** for illustrations of rituals of transition). With successful accomplishment of the rituals of transition, the initiate becomes incorporated into the stability of everyday life but at a new social stage with accompanying responsibilities. Ultimately, rites of passage serve a strong social conveyance function that informs the initiate and his/her social group that an important social change is occurring.

Table 1 Examples of functions of adolescent initiations and rites of passage in various cultures of the world

- To compel the initiate to accept desired social roles.
- To announce marital eligibility of the initiate.
- To ascribe a socially valued identity on to the initiate.
- To encourage the initiate to exhibit discipline, maturity, and social responsibility.
- To compel the initiate to assume expected social and behavioral changes in accordance with physical maturation.
- To facilitate the initiate's inculcation of cultural values, beliefs, and practices.
- To compel the initiate's adoption of desired character traits related to disposition and temperament.
- To promote the initiate's longevity.
- To recognize and celebrate change of status in the initiate's life.
- To advance reproductive viability or fertility of the initiate.
- To solidify and strengthen kinship and other social bonds between the initiate and others.
- To facilitate spiritual enhancement or transcendence of the initiate.
- To build physical strength and endurance of the initiate.

Adapted from Markstrom, C.A., 2011b. Initiation ceremonies and rites of passage. In: Brown, B.B., Prinstein, M.J. (Eds.), Encyclopedia of Adolescence, vol. 2. Academic Press, San Diego, pp. 152–159.

Table 2 Examples of rituals occurring during the transition phase of rites of passage in various cultures of the world

- Apprenticeships
- Bathing
- Capping
- Changing of clothes
- Dancing/singing
- Ear piercing
- Fasting
- Feasting/lavish parties
- Genital mutilation
- Hair cutting/head shaving
- Hair pinning
- Hand washing
- Henna painting rituals
- Tattooing
- Inducting into a secret society
- Praying
- Wearing turbans/head scarves

Culture-based rites of passage practiced in the transition to adulthood in Western societies are limited. Certainly ceremonies such as graduation from high school or higher education do confer a measure of maturity with marriage bringing even further adult distinctiveness. Some adolescents may undergo religion-based rites of passage such as Jewish *Bar Mitvahs* and *Bat Mitvahs* and Protestant and Catholic religious confirmation ceremonies. These ceremonies of commitment, confirmation, and initiation teach and elicit commitments of fidelity to a belief system (Markstrom, 2008). In contrast, the remainder of this article addresses rites of passage and other markers of the transition to adulthood in non-Western cultures where such practices may represent more culture-embedded practices. Of further consideration are changing customs in an increasingly globalized world.

Cultural Markers of the Transition to Adulthood

Examples of cultural markers of transition to adulthood are organized according to those that are biologically based, those that include ceremonies and celebrations, and those that involve instruction. The reinvention and revisions of these markers are considered according to exposure to global ideas.

Biologically Linked Markers

Genital Mutilation

Genital mutilation is a marker of a new status, albeit a highly controversial marker. Female genital mutilation (FGM), also termed as female cutting or female circumcision, is one type of genital mutilation and is defined by the World Health Organization (WHO, 2008) as "all procedures involving partial or total removal of the external female genitalia or other injury to the female genital organs for non medical reasons" (p. 1). UNICEF (2013) estimated that 125 million girls and women in 29 African and Middle Eastern counties have undergone FGM. Although there have been decreases in the rate of FGM, UNICEF suggests that about 30 million girls are still at risk of experiencing it in the coming decade. Reasons for the persistence of FGM include desires for the protection of ethnic and gender identity, female sanctity and virginity, family prestige, hygiene, and ensuring matrimony (UNICEF, 2013). FGM practices have complex etiologies and are rooted in ideological, material, and spiritual reasons for those that practice them (Berg and Denison, 2013).

FGM is a divisive issue and has sparked international outcry asking for its abandonment as a practice. As a result, these rituals have been modified and eliminated in many traditional cultures due to social and health concerns. For instance, rising rates of HIV/AIDS in Africa is forcing communities to reshape the goals and social capital rooted in cultural rituals such as circumcision (Kang'ethe, 2013). There have also been reports of sterile procedures being introduced in both male and female circumcisions (Kang'ethe, 2013; Prazak, 2007). African practicing communities, in attempts to hold on to their cultural values, norms, and belief systems, have added alternative approaches and rituals. For instance, Oloo et al. (2011) discuss 'alternative rituals' used in Kenya with varying degree of success. Components of these rituals included periods of teaching, frequently occurring in isolation, as well as communal festivities and sometimes a certificate to indicate the completion of the rituals. The alternative practices focus on retaining the essential components of female circumcision, such as instruction about duties and roles in the home, gift exchange, festivities, and public acknowledgment of the change of status, in the absence of the cutting ceremony. These alternative rituals help girls acquire knowledge about reproductive health and promote perspectives on gender equality in the family and among community members.

Cross-national migration has also contributed to the changing face of genital mutilation across the globe. Berg and Denison (2013) stated that the legal frameworks and national discourses in Western countries have served as barriers to immigrants' practices of FGM. Some from these immigrant communities now question the practice in their new social, political, and cultural contexts while some argue for its necessity as a cultural tradition (e.g., Isman et al., 2013).

Menarche and Pubertal Development

Rituals of genital mutilation often are associated with the changing pubescent, but there are other types of cultural practices during this phase of transition. Albeit these markers occur early in the developmental transition to adulthood and, today, do not necessarily confer full adult status. Nonetheless, cultural markers at pubescence advance recognition and celebration that the individual is moving toward adulthood. Markstrom (2008) conducted extensive research on contemporary practices of female coming-of-age ceremonies at menarche among various North American Indian tribes. The Western Apache Sunrise Dance and the Navajo Kinaaldá are two ceremonies with continuity in practice over eons. Adolescent initiates are similar to their counterparts throughout the present-day United States – they attend school, they engage with social media, they listen to popular music, they wear the same trends in fashion, and so. Yet, for a sequence of several days, contemporary life is suspended and a physical context is constructed with resemblance of earlier ways of their ancestors. These ceremonies reflect a remarkable coexistence of present-day life suspended for a brief time while ancient rituals are implemented with seemingly minor changes from the past (Markstrom, 2008). Transition is ushered in through complex and physically and emotionally challenging rituals richly imbued with meaning and reflective of important cultural values. Traditional beliefs, values, and practices are reinforced through the retelling of origin stories, speaking of the language, preparing and consuming traditional foods, incorporating traditional sacred objects into rituals (e.g., feathers), and wearing of traditional clothing (Markstrom, 2008). Apache, Navajo, and similar ceremonies of other tribes facilitate spiritual transcendence so that the initiate is empowered and prepared for challenges of contemporary life that will require her management. Ultimately, the initiate returns to her everyday life but with a new social status and altered self as well as other perceptions.

Social treatment of pubescence has also been examined within the fluidity of the ever-changing political and socioeconomic contexts of African cultures. In indigenous Africa, the community takes responsibility for socialization of youth and engages in ceremonial rituals to support the youth's transition

from adolescence to adulthood. With Western influences, many pubertal rites of passage are changing or being eliminated completely (Adegoke, 2001). For example, in Kikuyu culture, adolescent girls only had contact with young men other than their brothers at the Ngweko dance which was monitored through peer supervision. The dance also functioned to formally transfer from parents to youth the responsibilities of premarital sex and mate selection all within a supportive context. Today, adolescent girls interact with males within school contexts in the absence of norms and instructions for their behavior with the opposite sex (Adegoke, 2001).

Pubescence is an obvious biological marker that has been the focus of intensive and extensive ceremonial rituals and rites of passage for mostly girls and also for boys in certain cultures (Schlegel and Barry, 1980). These events at pubescence and other time frames during adolescence are accompanied by rituals and frequent celebrations, of which some examples are given in the next section.

Rituals and Celebrations

Bodily Adornment

van Gennep (1908/1960) situated bodily transformations such as tattooing, scarification, or painting within rites of passage. Separation can be heightened by rituals of bathing and hair washing, changing of clothes, tattooing, and other physical alterations. Margret Mead (1928) described adolescence in her book, *Coming of Age in Samoa*, drawing attention to the Samoan tattooing ritual to signify the commencement of adolescence. However, this tattooing ritual in Samoa has diminished over the years with the presence of Christian missionaries, rise of education, and immigration for jobs. But a resurgence of tattooing has been seen among those who remain in Samoa as a separation response to the forces of globalization in Samoan society (Jensen and Arnett, 2012).

Revival of bodily adornment rituals has also been seen among immigrants who display bicultural identities. This is often an effort to preserve culture of origin as individuals assimilate into their new cultures after immigration (Phinney et al., 2001). For example, the bride's henna ritual practiced by Jews in Yemen and its subsequent transformation after their immigration to Israel (Sharaby, 2006) which illustrates the negotiation of bicultural identities among immigrants. The henna ritual was seen as a seminal rite of passage for women in Yemen and as a way of conserving societal institutions which promoted gender separation and non-egalitarian structures in society. In the past, only women participated in the ceremonies and singing and dancing gave women a space to challenge the rules of modesty and social norms that governed their lives. Moving to Israel, coupled with shifting from extended to the nuclear family system, exposed Yemenite women to new economic and social activities. There have been major alterations to the henna ritual as a way to renovate a traditional ritual to fit the shifting social and cultural context for young immigrants; for example, the wedding henna ritual has been reinvented to focus more so on the couple's change of status, instead of just on the woman's rite of passage as she transitions into a man's wife (Sharaby, 2006).

Youth Galas

The glamorization of the prom in the Bahamas is an interesting example of the influence of globalization on the youth of a non-western culture according to a hybrid ritual that combines elements of traditional culture, postcolonial reactions, and influences from the United States (Thompson, 2011). The celebration of proms in the Bahamas was initiated as a response to their independence from the British in order to find new ideals of modernity – it serves as a coming-of-age practice for Bahama's own youth. Proms were initially celebrated in hotels as low-key events; however, they have evolved into something much more dramatic among Bahamian youth culture. The popularity of American television, the American expatriate community, and even its close proximity to America allows for an easy flow of hip-hop culture to Bahamian youth. This influence has led proms to transform into something of a spectacle, especially prom entrances. The desire of youth attending the prom is to be seen in the most spectacular way possible with tricked out cars, flashy bling, and elaborate prom wear (Thompson, 2011). However, components of cultural heritage also are evident. For example, while Western style prom dresses are worn by some prom-goers, others choose to dress in more elaborate costumes reminiscent of Caribbean stylings. Overall, this hybrid celebration provides space for expression of meaningful cultural traditions as well as youth customs imported from elsewhere.

Quinceañera

Another example of an elaborate youth celebration is the Latina quinceañera at the time of a girl's 15th birthday. The young female wears an elaborate or elegant dress, traditionally white, and has a 14-member court that serves as a symbolic representation of her 14 years of life. In the past, the quinceañera has been practiced in Latina/Hispanic Catholic communities as a symbol of the girl's readiness to marry. Within the Catholic Church, this right is very important and is considered an extension of baptism for the girls. The quinceañera emphasizes the importance of virginity and community service, but it has been changing (Banuelas, 1981 as cited by Davalos, 1996). Today, quinceañeras signify the parents' acknowledgment of their daughter's readiness to date, and has been transformed into something more similar to a 'sweet-sixteen' party rather than the entrance into womanhood function of the past (Davis et al., 2009). In Ecuador, quinceañeras are losing their celebratory extravagance in favor of other types of extravagances such as expensive gifts or foreign vacations for the girls (World Trade Press, 2010). Among Protestant Mexican-Americans in the United States, the quinceañera ceremony has been redefined as a cultural rite of passage rather than a Catholic rite (Barton, 2006). With the influence of American culture, the quinceañera has been expanded into something akin to a wedding with the associated high costs, advanced planning, and encompassing a cultural significance rather than being defined solely on religious grounds (Dieter, 2002). Further, in some cases, the quinceañera has begun to be identified as a debut or sweet-sixteen celebration. Although the extravagance and religious and cultural significance are changing, the quinceañera is still an important part of a girl's transition into womanhood for many Latina/Hispanic cultures in the United States.

Marriage

Marriage is a formalized event that brings recognition, meaning, and significance to life transitions as a rite of incorporation. It has universally been regarded as a rite of passage to adulthood. However, the United Nation's World Marriage Data (2008) shows that since the 1980s, marriage rates have decreased in countries that encompass a total of 87% of the world's population. This trend suggests a global decline in the rates of marriage, which further affirms the changing face of coming-of-age in the twenty-first century. Emerging adulthood has been often cited as a reason for postponing marriage and parenthood in industrialized nations (Jensen and Arnett, 2012), where it has been primarily studied. There is a further need to study marriage and emerging adulthood in collectivist cultures like India or China as together they represent 42% of the population of young people in the developing world (NRCIM, 2005), which increases their influence on global youth trends.

An interesting case study of Indian youth reveals how young urban professionals navigate two identities: one that participates in an international economy and the other that seeks to adhere to Indian traditions such as arranged marriages and caring for their parents in old age (Verma and Saraswathi, 2002). However, it is important to consider that these social norms surrounding community and kinship, coupled with the structural constraints of poverty, shape the lives of youth in rural India very differently than their urban counterparts. Disparities in employment, education, and healthcare between urban and rural youth continue to be a persistent problem in present-day India (Patel, 2013). Additionally, India also struggles with child marriages, which UNICEF (2007) defines as marriage prior to age 18. Child marriages show considerable variability across countries; nonetheless, about half of those experienced are by girls in South Asia (UNICEF, 2007).

Marriage and family are also highly valued in China, where influences from Confucianism along with tumultuous economic and social reform movements contribute to an understanding of the role of these institutions in young people's lives today (Yeung and Hu, 2013). Even in the face of globalization, marriage and childbearing within marriage continue as important features of social and personal life and son preference remains strong (Yeung and Hu, 2013). These features indicate that personal conduct and compliance with social norms are strong among Chinese young people (Nelson et al., 2004). However, there is evidence signifying a rise in the age of first marriage and childbearing, along with cohabitation and premarital sex in urban youth in modern China (Yeung and Hu, 2013).

Marriage is one facet of signifying transition to adulthood, but the acquisition of useful skills and employment are necessary for becoming economically self-sufficient and viable. This is addressed in the next section along with the role of instruction in socializing young people toward adulthood.

Skill Building and Learning

Apprenticeship

Apprenticeship is linked with learning and is a socially situated practice deriving importance and meaning as supported within the structure of the surrounding culture and society. Learning is a part of an ongoing, sometimes subtle, transition from unskilled to skilled workmanship (Lave and Wenger, 1991). For example, the Oku carvers in the Cameroon Grassfields utilize apprenticeships not only as a means of passing down knowledge of carving but also to shape the apprentice's new identity. The leaving ceremony that occurred in the past in Oku illustrates the conferral of adult status once the apprentice had learned all that he could from his master. At that point, the apprentice would bring an offering of food which the master would consume. This act served many functions including validation of the apprentice, celebration of the apprentice's success, and a marker of the apprentice's independence. Tools were traditionally limited to apprenticed carvers, who made the tools with the master gifting them to the apprentice for use in his own workshop. As tools became more available to non-apprenticed carvers, free market entrepreneurs emerged who carved and sold to dealers who then sold to foreigners and tourists (Argenti, 2002). In recent years, business with foreigner consumers has reduced, but the nonapprenticed Oku carvers continue to produce work. This dynamic suggests that carvers are not solely interested in pleasing the Western consumer but in creating a culture-based identity within the global political economy.

Informal Instruction

Universally, informal instruction is an implicit component of culture-based mechanisms for the transition to adulthood. For instance, puberty rituals were discussed in an earlier section but it is important to note that instruction of initiates is an integral component of many coming-of-age practices (Markstrom, 2008). The content of instruction includes socialization toward expected adult roles and responsibilities that frequently include gender-specific content. In Mozambique culture, informal instruction and counseling by community and family elders prepares the initiate for transition into a new phase of her life (Kotanyi and Krings-Ney, 2009). During initiation rituals, girls are taught about sexuality, moral adult behavior, feminine hygiene, traits of a good wife, and preparing the body for sexual activity. Female instruction during initiation rituals in Mozambique occurs in four stages over years with content specific to the age of the initiate ranging from a procedure of stretching the inner labia at ages 8–10 years to the birth of the first child (Kotanyi and Krings-Ney, 2009). A woman is said to be socially accepted as a complete individual after motherhood. These initiation rituals have also taken the form of HIV/AIDS education platforms in some of these communities as discussed earlier (Kang'ethe 2013). Informal instruction and counseling of this type can have far-reaching contributions providing an opportunity for women, particularly the integration of elderly women, into HIV/AIDS counseling (Kotanyi and Krings-Ney, 2009).

Similarly, in Malawi, informal instruction and counseling for young girls at the onset of their first menstruation is commonly practiced (Munthali and Zulu, 2007). An older woman serves as an authority on sexual and reproductive health and serves as the head counselor to young initiates. For boys, there is limited counseling on sexuality in some regions of Malawi because of the lack of characteristic markers of puberty such as menarche. Nonetheless, male relatives and friends instruct boys about sexuality (Munthali and Zulu,

2007). In other cultures, as well, for boys, the lack of a specific biological marker may account for the less prominent role of formal instruction and socialization toward adulthood – informal socialization may very well occur.

Discussion

This article has focused on culture-based practices facilitating the transition to adulthood in non-Western cultures with a particular eye toward our present-day global, interconnected, and changing world. Culture-based rituals can be sources of happiness, accomplishment, and build a sense of community among those that practice them, for whom rituals convey social capital (Kang'ethe, 2013). Globalization has brought with it a distinctive challenge to the youth around the world. It demands that they transform and reinvent traditional social markers while trying to maintain the embedded social capital that these rituals bring with them. Larson et al. (2009) categorizes societal changes in response to globalization as 'top-down' and 'bottom-up.' Top-down societal changes are driven by forces such as large corporations working in developing countries. The risk of such change is disruption of traditional subsistence economies and livelihoods, which may necessitate relocation to urban regions where cultural resources are absent. Bottom-up changes are driven by everyday people in a community or society who bring their resources to bear on the global community, such as marketing traditional wares of a culture to a broader consumer base. In response to bottom-up changes, youth often take on the role of 'social shifters' (Durham, 2000) and are active agents in processes of globalization. However, it is important to remember that the agencies and capacities of young adults across the globe must be contextualized in their unique social, demographic, economic, and political context. Customs may change more rapidly in urban and affluent settings where youth have more resources and economic mobility to be coconstructors of these new identities.

Given the fact that there is cross-cultural variability in the transition to adulthood (e.g., urban and rural distinctions as previously discussed), it is difficult to provide generalizing statements. However, the *Growing Up Global* project of the National Research Council attempted to define cross-cultural factors applicable in signifying success in this transition. They identified six factors:

- Good mental and physical health, including reproductive health, and the knowledge and means to sustain health during adulthood.
- An appropriate stock of human and social capital to enable an individual to be a productive adult member of society.
- The acquisition of prosocial values and the ability to contribute to the collective well-being as citizen and community participant.
- Adequate preparation for the assumption of adult social roles and obligations, including the roles of spouse or partner, parent, and household and family manager.
- The capability to make choices through the acquisition of a sense of self and a sense of personal competence.
- A sense of general well-being (NRCIM, 2005: p. 26).

Implicit to these statements are universal mandates that individual and societal maintenance require health, social capital, and prosocial values. Further, socialization of younger members of the society should include needed skills, values, and behaviors that contribute to self-, other-, and societal health and well-being.

In summary, the mechanisms and practices signifying the transition from adolescence to adulthood demonstrate cross-cultural variability as shown in this article. Some markers are biologically based (e.g., at pubescence) and these tend to occur earlier in adolescence, yet reflect societal perceptions that the individual is growing into adulthood and now should mirror more mature, responsible qualities. Other facets in the transition to adulthood encompass ceremonies and celebrations that may involve bodily modifications signifying adult status. Still other celebrations mimic the behaviors of adult such as dressing up in a more formalized manner, as in special formals, dances, or celebrations. Weddings are universal celebratory rites of passage in the transition to adulthood. However, in case of child marriage, such status is conferred prior to attaining the physical and socioemotional changes of adulthood. The final category of transition to adulthood delineated in this article encompasses formal and informal education and socialization of the young person for assumption of the roles and responsibilities of adulthood. Instruction may very well be a universal action associated with preparing youth for adulthood.

See also: Acculturation; Collectivism and Individualism: Cultural and Psychological Concerns; Ethnic Identity, Psychology of; Globalization: Welfare and Social Work; Immigration: Social Psychological Aspects; Self and Identity Development During Adolescence across Cultures; Social Identity in Social Psychology; Social and Economic Change: Psychological Challenges for Individuals.

Bibliography

Adegoke, A., 2001. Pubertal development and traditional support systems in Africa: an overview. African Journal of Reproductive Health 5 (1), 20–30.

Argenti, N., 2002. People of the chisel: apprenticeship, youth, and elites in Oku (Cameroon). American Ethnologist 29 (3), 497–533.

Arnett, J.J., 2007. Emerging adulthood: what is it, and what is it good for? Child Development Perspectives 1 (2), 68–73.

Banuelas, A., 1981. La tradicion de la Quinceanera. Liturgy 80 (12 (7)), 5.

Barton, P., 2006. Hispanic Methodists, Presbyterians, and Baptists in Texas. Press, Texas, University of Texas.

Berg, R., Denison, E., 2013. A tradition in transition: factors perpetuating and hindering the continuance of female genital mutilation/cutting (FGM/C) summarized in a systematic review. Health Care for Women International 34 (10), 837–859.

Davalos, K.M., 1996. "La Quinceanera:" making gender and ethnic identities. Frontiers: A Journal of Women Studies 16, 101–127.

Davis, D.E., Boyle, D.P., Deaton, T.M., 2009. Voices from the Nueva Frontera: Latino Immigration in Dalton, Georgia. University of Tennessee Press, Tennessee.

Dieter, K., 2002. From church blessing to quinceañera Barbie: America as "spiritual benefactor" in la quinceañera. Christian Scholars Review 32 (1), 31–48.

Durham, D., 2000. Youth and the social imagination in Africa: introduction to parts 1 and 2. Anthropological Quarterly 73 (3), 113–120.

van Gennep, A., 1908/1960. The Rites of Passage (M. Vizedom, G. Caffe, Trans.). Chicago University Press, Chicago (Original work published 1908).

Hamilton, S.F., Hamilton, M.A., 2009. The transition to adulthood: challenges of poverty and structural lag. Handbook of Adolescent Psychology 2, 492–526.

Isman, E., Ekéus, C., Berggren, V., 2013. Perceptions and experiences of female genital mutilation after immigration to Sweden: an explorative study. Sexual & Reproductive Healthcare 4, 93–98.
Jeffrey, C., McDowell, L., 2004. Youth in a comparative perspective global change, local lives. Youth & Society 36 (2), 131–142.
Jensen, L.A., Arnett, J.J., 2012. Going global: new pathways for adolescents and emerging adults in a changing world. Journal of Social Issues 68 (3), 473–492.
Kang'ethe, S.M., 2013. The panacea and perfidy of cultural rites of circumcision in African countries: examples from Kenya, Botswana and South Africa. Eastern Africa Social Science Research Review 29 (1), 107–123.
Kotanyi, S., Krings-Ney, B., 2009. Introduction of culturally sensitive HIV prevention in the context of female initiation rites: an applied anthropological approach in Mozambique. African Journal of AIDS Research 8 (4), 491–502.
Larson, R.W., Wilson, S., Rickman, A., 2009. Globalization, societal change, and adolescence across the world. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, Contextual Influences on Adolescent Development, third ed., vol. 2. Wiley, Hoboken, NJ, pp. 590–622.
Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, New York.
Markstrom, C.A., 2008. Empowerment of North American Indian Girls: Ritual Expressions, at Puberty. University of Nebraska Press, Lincoln.
Markstrom, C.A., 2011a. Identity formation of American Indian adolescents: local, national, and global considerations. Journal of Research on Adolescence 21, 519–535.
Markstrom, C.A., 2011b. Initiation ceremonies and rites of passage. In: Brown, B.B., Prinstein, M.J. (Eds.), Encyclopedia of adolescence, vol. 2. Academic Press, San Diego, pp. 152–159.
Mead, M., 1928. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation. Morrow, New York.
Munthali, A.C., Zulu, E.M., 2007. The timing and role of initiation rites in preparing young people for adolescence and responsible sexual and reproductive behaviour in Malawi. African Journal of Reproductive Health 11 (3), 150.
National Research Council and Institute of Medicine (NRCIM), 2005. Growing up global: the changing transitions to adulthood in developing countries. In: Lloyd, Cynthia B. (Ed.), Panel on Transitions to Adulthood in Developing Countries. Committee on Population and Board on Children, Youth, and Families. Division of Behavioral and Social Sciences and Education, Washington, DC. http://www.nap.edu/ (accessed 11.05.13.).
Nelson, L.J., Badger, S., Wu, B., 2004. The influence of culture in emerging adulthood: perspectives of Chinese college students. International Journal of Behavioral Development 28 (1), 26–36.
Oloo, H., Wanjiru, M., Newell-Jones, K., 2011. Female Genital Mutilation Practices in Kenya: The Role of Alternative Rites of Passage: A Case Study of Kisii and Kuria Districts. Feed the Minds, Kenya. http://www.feedtheminds.org/downloads/FGM%20Report_March2011.pdf (accessed 28.11.13.).
Patel, V., 2013. State of the Urban Youth, India 2012: Employment, Livelihoods, Skills. IRIS Business Services, Mumbai, India. http://www.esocialsciences.org/General/A201341118517_19.pdf.
Phinney, J.S., Horenczyk, G., Liebkind, K., Vedder, P., 2001. Ethnic identity, immigration, and well-being: an interactional perspective. Journal of social issues 57 (3), 493–510.
Prazak, M., 2007. Introducing alternative rites of passage. Africa Today 53 (4), 19–40.
Schlegel, A., Barry III, H., 1980. The evolutionary significance of adolescent initiation ceremonies. American Ethnologist 7, 696–714.
Schlegel, A., 2000. The global spread of adolescent culture. In: Crockett, L.J., Silbereisen, R.K. (Eds.), Negotiating Adolescence in Times of Social Change. Cambridge University, Press, New York, pp. 71–88.
Schlegel, A., 2009. Cross-cultural issues in the study of adolescent development. In: Lerner, R.M., Steinberg, L. (Eds.), Handbook of Adolescent Psychology, Contextual, Influences on Adolescent Development, vol. 2. Wiley, Hoboken, NJ, pp. 570–589.
Schlegel, A., Barry III, H., 1991. Adolescence: An Anthropological Inquiry. The Free Press, New York.
Sharaby, R., 2006. The bride's henna ritual: symbols, meanings and changes. Nashim: A Journal of Jewish Women's Studies & Gender Issues 11 (1), 11–42.
Thompson, K.A., 2011. Youth culture, diasporic aesthetics, and the art of being seen in the Bahamas. African Arts 44 (1), 26–39.
Turner, V., 1967. The Forest of Symbols: Aspects of Ndembu Ritual. Cornell University Press, Ithaca, NY.
United Nations, 2009. Department of Economic and Social Affairs, Population Division, World Marriage Data 2008 (POP/DB/Marr/Rev2008). http://www.un.org/esa/population/publications/WMD2008/Main.html (accessed 11.07.13.).
UNICEF, 2007. Progress for Children. Protecting against Abuse, Exploitation and Violence. Child Marriage. http://www.unicef.org/progressforchildren/2007n6/index_41848.htm (accessed 12.02.13.).
UNICEF, 2013. Female Genital Mutilation/Cutting: A Statistical Overview and Exploration of the Dynamics of Change. http://www.unicef.org/media/files/FGCM_Lo_res.pdf (accessed 12.01.13.).
Verma, S., Saraswathi, T.S., 2002. Adolescents in India: street urchins or Silicon Valley millionaires? In: Larson, R., Saraswathi, T.S. (Eds.), The World's Youth: Adolescence in Eight Regions of the Globe. Cambridge University Press, New York, pp. 105–140.
WHO, 2008. Eliminating Female Genital Mutilation: An Interagency Statement OHCHR, UNAIDS, UNDP, UNECA, UNESCO, UNFPA, UNICEF, UNIFEM, WHO. The World Health Organization, Geneva. http://www.un.org/womenwatch/daw/csw/csw52/statements_missions/Interagency_Statement_on_Eliminating_FGM.pdf (accessed 11.02.13.).
World Trade Press, 2010. Ecuador Society and Culture Complete Report. World Trade Press, Petaluma, California.
Yeung, W.J.J., Hu, S., 2013. Coming of age in times of change the transition to adulthood in China. The Annals of the American Academy of Political And Social Science 646 (1), 149–171.

Relevant Websites

http://www.yale.edu/hraf/ – Human Relations Area File. Cultural Information for Education and Research.
http://www.nap.edu/catalog/11174.html – National Academies Press. Growing up Global: The Changing Transitions to Adulthood in Developing Countries.
http://www.nap.edu/catalog/11524.html – National Academies Press. The Changing Transitions to Adulthood in Developing Countries: Selected Studies (National Academies Press).
http://www.unicef.org/media/files/FGCM_Lo_res.pdf – UNICEF Female Genital Mutilation/Cutting: A Statistical Overview and Exploration of the Dynamics of Change.
http://www.un.org/esa/population/publications/WMD2008/Main.html – United Nations, Department of Economic and Social Affairs, Population Division. World Marriage Data 2008.
http://www.who.int/reproductivehealth/publications/fgm/en/ – World Health Organization. Female Genital Mutilation (FGM) and Harmful Practices.

Translating Sociology into Japanese, Chinese and Korean Languages

Jeremy CA Smith, Federation University Australia, VIC, Australia
Myoungkyu Park, Seoul National University, Seoul, Republic of Korea
Albert Tzeng, International Institute for Asian Studies, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

East Asia's encounters with the West include an interesting era of translation of sociological concepts into Japanese, Chinese, and Korean. This includes the concept of 'society' itself. This article explores how the concept of 'society' and sociology as a discipline were translated from European languages into East Asian languages in significantly different contexts to that of Europe. The authors focus on the role of intellectuals, organizations, and social movements in the process of development of Japanese, Chinese, and Korean vocabularies of sociology.

The intriguing history of East Asia's encounters with the West includes unusual cases of translation of sociology and sociological concepts into Japanese, Chinese, and Korean. This is simultaneously a regional process and three distinct (though connected) national processes associated with East Asian modernities. It is part of a phase of intercivilizational encounters of an exceptional kind.

East Asian Sociology: Western Modernity and Japanese Translation

Development of East Asian sociology is one of the civilization processes in which Western modernity was expanded to the far eastern region. In encounters with other civilizations, or the world religions or universal ideologies, Japanese intellectuals actively interpreted and relativized those civilizations, religions, and ideologies against a native cultural core. The Japanese case of translation of 'society' and 'sociology' is a case in point and an instructive case because of Japan's sudden encounter with Western sociology. 'Society' can be difficult to translate and would be doubly so with transliteration from one set of languages to another language family. Japan's acquisition of sociological concepts provides a rich case study in engagement with the unfamiliar in two respects. Firstly, the commitment to interpretation of Western sociology and the conceptual apparatus behind it produced a rapid change of understanding. Secondly, the change of understanding took place in a different language family and involved transliteration into kanji. That made the process of adaptation of sociology more complicated.

Meiji Japan's Engagement of Western Sociology

The young generation of converted samurai, which led the Meiji Restoration (1868) prioritized the acquisition and comprehension of Western sociology. The quest for knowledge stimulated the 1871 Iwakura Mission, an large-scale international tour of hundreds of Japan's leaders to the United States and much of Europe, a tour which must rate as one of the most systematic fact-finding missions in world history (Nish, 1998). The cultural learning that took place after nearly 400 days of travel incorporated many of the main disciplines in the physical and social sciences and humanities. Subsequently, the knowledge of engineering techniques, systems of government, and political philosophy and social science acquired was spread to a literate and intelligent populace. Within this wider engagement with global knowledge, three features of Western sociology attracted Japanese interest:

1. Comtean social thought encountered first by the Japanese appeared coherent and persuasive. Nishi Amane (1829–97) – a pivotal figure in early translation of sociological concepts – adapted it. He studied Comte in the Netherlands for 4 years in the 1860s. Comte's *Cours de philosophie positive* greatly impressed him and, like Comte, he introduced the term and idea of 'sociology' to the Japanese universe of learning. Comte's comprehensive and encyclopedic sociology provided one entire system of knowledge that could frame intercultural learning for the Japanese. The receptivity of the Meiji elite to the encyclopedic nature of the Western social sciences made Comte's evolutionary sequencing appealing in the 1870s.
2. For a new program of modernization, the Meiji elite needed an ideological vehicle. Comte's sociology – and succeeding it, Herbert Spencer's social evolutionism (Yamashita, 1984) – presented a theory of history, which validated the Meiji transformation and provided further cultural resources for it. The world-historical sweep of Spencerian and Comtean evolutionism was pliable enough to the climate of debate about the Meiji program of modernization (Nozomu, 1994: pp. 4–6). Ideologically, evolutionism made sense to a new regime striving to realize a place in the world order of empires. Comtean and Spencerian sociologies cast a world of modernizing states that were empire building, large-scale industrial capitalist economies.
3. Sociology as a body of thought presented itself as the science of social change, which was precisely the historical process that Japan was undergoing. As the transformation was led by a highly reflexive social force it had to involve an activist disposition toward the process of transformation itself. Sociology appeared to the Meiji elite to accommodate such a disposition. Learning and translating its conceptual apparatus became an urgent endeavor.

The translation and interpretation of social sciences associated with the project of 'civil society' separate from the state in a society that did not generate any such civil society was quite difficult. Even a word to connote the core concept of 'society' was hard to formulate.

Concepts in Translation

From the 1860s to the 80s, there were at least three Japanese words that could act as a translation of 'society': *nakama*, *setai*, and the part-neologism *shakai*.

Nakama (仲間) connotes a voluntarily instituted group of townspeople, in which members share a certain aim or belief and communicate as equals. Some were guildlike or corporate in nature. There were countless numbers of informal *nakama*, such as religious groups, hobby clubs, and alumni organizations. The Meiji intelligentsia experimented with the term *nakama* as an equivalent to 'society,' because it inferred qualities of civil society such as voluntary association and the institution of groups outside the range of kinship.

The second option, *setai* (世帯), has Buddhist origins and denotes the tacit norms of the earthly world. In the Buddhist world view invoked by *setai*, we have to obey tacit rules in ordinary life, even if we find them irrational and detestable. In contrast to a heavenly existence the mundane world appears a transient and provisional place (Sako, 2006: pp. 5–6). This phenomenological term for 'society' distinguishes the transcendental from the mundane. *Setai* is normative in effect as it implies tacit rules for ordinary life. On the occasion when sociology was taught for the first time at the University of Tokyo *setaigaku* was the translation for 'sociology' chosen.

Shakai (社 會) has the most telling history. Nishi Amane (西 周 1829–97) was the vital figure in the emergence of this term as the main one to translate 'society.' He operated in a rich intellectual milieu. Together with Arinori Mori (1847–89), Nishi founded the famous *Meirokusha* (明六社) publishing house and the journal *Meiroku Zasshi*. The *Meirokusha* ('Meiji Six Society') worked at the interstices of Western and Japanese perspectives on societal transformation. The work of the group's members couched the project of modern reconstruction in sociological terms for Japan (Braisted, 1976). They were translators also. Nakamura Masanao translated Smiles' *Self Help* and then Mill's *On Liberty*. Amane too was involved with translation of whole texts, including Mill's *Utilitarianism*. These translations and others presented touchstone works of Western thought to the Japanese. The thinking of the *Meirokusha* took place at the interstices of civilizational engagement laying out for the Japanese urban public "description of a received world and construction of a new one" (Howland, 2002: p. 5). They were not merely translators or conveyors of Western ideas. Their translations were an exercise in interpretative work that began the development of new sociological and philosophical ideas in Japan.

Translation across languages and language groups of key words featured in their early discussions. Nishi was the first to use the term *shakai* to denote 'society' in 1874. He translated 'philosophy' as *tetsugaku* and a number of other crucial terms. With *shakai*, Nishi intended to convey 'society' in the sense of 'associations' or rather 'societies.' Nishi's use is part neologism and part adaptation. *Shakai* was derived from twelfth to thirteenth-century Chinese literature and referred to "peasants festivals" (Sako, 2006: pp. 6–7). In Japan, *shakai* had a different use. In Nishi's hands, and in wider use by the *Meirokusha* and university authorities, *shakai* could be effective as a term for 'society.'

Despite its novelty, the careful recombination of *kanji* based on the Chinese original proved salient. Etymologically, the word *shakai* is compounded from *sha* (社, *she* in Chinese, *sa* in Korean) and *kai* (會, *hui* in Chinese, *hoe* in Korean). *Sha* (社) indicates shrine. *Kai* (會) denotes meeting or assembly. In Chinese, as noted, it signified 'festival.' The term *shakai* inherited by the Meiji intelligentsia carried the connotation of a meeting at a temple. The connotation of gathering in a public place of an open crowd was the closest equivalent to 'society' in the sense of constellation of 'associations' that Nishi was seeking. He was not alone in his use of this term. Spencer's work *Education: Intellectual, Moral, and Physical* was first translated in 1880 by Shinpachi Seki (尺振八). He used *shakai-gaku* for 'sociology.' Seki's translation was followed, in 1884, by Koutaro Noritake's (乗竹孝太郎) translation of Spencer's *Principle of Sociology* as *Shakai-gaku no genre* (社会学原理).

Nishi's role was more general than that of other Meiji intellectuals who specialized in large but particular problematics: the national spirit (Yukichi Fukuzawa), constitutional arrangements (Kato Hiroyuki), and the adoption of foreign customs (Tsuda Mamichi) (Motoyama et al., 1997: pp. 269–270). His intellectual authority helped consolidate recognition of *shakai* as the legitimate term in educational and government circles. *Shakai* and *shakai-gaku* had successfully emerged as translations of 'society' and 'sociology' in Japanese.

Chinese Development of Sociology

The idea of sociology entered the Chinese Empire in the 1890s, roughly two decades behind Japan. The rapid intake of Western knowledge caused a pressing need for Chinese translations of Western concepts. There were two approaches. The first, adopted mainly by reformists who were familiar with Chinese scholarship, was to create new phrases with characters carefully chosen to link the Western concepts with elements in Chinese heritage. The second was to borrow from the readily available stock of Japanese-made Chinese phrases. In many cases, there were parallel translations from an identical Western concept competing for recognition and acceptance (Huang, 2011). The term 'sociology,' too, had been associated with two Chinese terms. The first was *qunxue* (群學), a term that was related by its user to the classical Confucius doctrine. The second was *shehuixue* (社會學), a term of identical Chinese characters of *shakai-gaku* in Japanese.

Qunxue literally means the study (*xue* 學) of *qun* (群 or 羣), a character that consist of a figurative part meaning *sheep* (羊) and the phonetic part reads *jun* (君). Etymologically *qun* means 'sheep flock' or 'to drive sheep into flock.' It was later appropriated to refer to 'the action that leads to the formation of a group.' The Confucius philosopher Xunzi (荀子312–230 BC) wrote a paragraph that became widely cited in the elaboration *qun*. In *Wang Chi* (王制, literally 'king's institution'), an essay on the art of governance, Xunzi asked: why men, who 'are not

as strong as cow, nor can run as fast as horse,' could master the natural? The answer he suggested was 'the ability to *qun*.' He pointed out that this ability relied on 'division' (分) (of roles or resources); and division relied on *yi* (義, righteous and appropriate ethic). The 'division according to *yi*,' he argued, was the foundation for harmony, unity, and strength (Zhang, 1992: pp. 5–6). This 'ability to *qun*' became a central concern for many nineteenth-century Chinese intellectuals who attributed the series of defeat by the Western powers to the lack of national unity and cooperation (Yao, 2006: p. 19).

There is dispute on who first coined the term *qunxue*. Its first appearance in existing literature is in the 1895 article 'yuan qian' (原強, literally "original strength"), in which the author Yen Fu (嚴復), a renowned translator, illustrated the theories of Darwin and Spencer and translated sociology into *qunxue*. Yen Fu also translated Spencer's *The Study of Sociology* and published it under the title *Qunxue Yiyan* (群學肄言) in 1903. However, it was alleged by the political reformist Liang Q-Chao (梁啟超) that his mentor Kang You-Wei (康有為) had in 1891 included *qunxue* as a subject in the teaching curriculum of Changxing School (長興學舍) in Wanmu Cottage (萬木草堂), Guangzhou (Liang 1928). Some scholars accept the account while tracing the origin of *qunxue* (e.g., Zhang, 1992: p. 5), while others suggest that Liang might have borrowed the term from Yen's writing and labeled their earlier agenda as such (Chen, 1984; Yao, 2006: p. 24).

It is important to note that the version of *qunxue* described by Liang is of very different nature from Yen Fu's. The reformists' *qunxue* might have been inspired by Western ideas, but their understanding of Western sociology was superficial at the best and the project was largely a reinterpretation of Chinese classics (Yao, 2006: p. 24). By contrast, Yen employed the term *qunxue* only to connect the translated Spencerian sociology with the legacy of Xunzi. Yen's use of *qunxue* was imbued with Spencerian sociology.

Shehuixue, the alternative translation, was introduced in the aftermath of the First Sino-Japanese War (1894–95). The bitter defeats by the small Asian neighbor cast a huge impact. Eager to discover the secrets of Japan's rapid modernization, Chinese students started to flood into this Eastern neighbor to study a wide range of disciplines that include social sciences. Those students were exposed to Western scholarship through literature in Japanese language, which was linguistically more accessible to them than Western languages. Many started to introduce the modern sciences to China via introductory texts published on press, or even via translations of Japanese textbooks (Harrell, 1992).

This flow of modern knowledge inevitably caused the proliferation of "Japanese made Chinese language" (和製漢語) – the modern phrases translated by Japanese scholars (notably Nishi Amane) using Chinese characters. These imported phrases triggered strong criticism and resistance among the conservative elites in China, who considered those phrases a contamination of classical aesthetics of writing. Policies to ban their usage were seen in some schools or examinations. However, those imported phrases still grew in their popularity and continuously marginalized the translations coined by Chinese scholars. Four reasons are suggested. First, the inflow of modern knowledge was just explosive, highlighting gaps in Chinese language where the Japanese made phrases were the only communicable terminology. Second, even in places where a domestic translation existed, the texts translated from Japan just outnumbered and gave their terminology more popularity. Third, the translation coined by Chinese scholars generally showed a preference for classical aesthetics (e.g., fewer characters-per-phrase and links with Chinese classics), which restricted their acceptance among those who demand more comprehensible language. Fourth, Japan had already constructed a consistent system of terminology that makes its understanding more easily. By contrast, the translation efforts in China remain individual and fragmented (Huang, 2011).

The term *Shehuixue* first appeared in Chinese in 1896 when Tang Si-Tong (譚嗣同, 1865–98) listed *shehuixue* as the Western scholarship anyone devoted to *renxue* (仁學, "study of benevolence") should learn. He did not elaborate what *shehuixue* meant. Read in context, the term might be adopted to loosely refer to a general idea of "social studies" (Han, 1987: pp. 25–26). Ding Yi (1988) identified Han Tan-Shou as the first Chinese scholar who used *shehuixue* to translate sociology.

The systematic introduction of *shehuixue* into China started in 1902. Zhang Tai-yan (章太炎, 1869–1936), a philosopher and revolutionary in exile in Japan, published the first sociology textbook in China. This book was based on his translation of *Sociology* by Kishimoto Nomuta (岸本能武太). The translation of *Ethnic Evolutionism* by Ariga Nagao (有賀長雄) was published in the same year too. 1903 marked the publication of Wu Jian-Chang's (吳健長) translation of Giddings' *Theory of Socialisation* based on its Japanese translation by Ichikawa Genzu (市川源三) and Ma Jun-Wu's (馬君武) translation of Spencer's *The Introduction of Sociology* (Yen, 2004: p. 2).

Concepts of Society and Sociology in Korea

There were also early encounters between Koreans and Western sociology via Japanese translation. The earliest instance of translation of 'society' is found in 1896 when *sahoe* (사회) was first used as a newly translated version of *shakai* by Korean students sent to Japan by the government to learn 'civilization and enlightenment' (Park, 2000). The students formed the *Daejoseon Ilbon Yuhaksaeng Chinmokhoe* (Fraternal Society of Korean Students in Japan) and published the bulletin '*Chinmokhoe hoebo*.' In its second issue in March 1896, Sin Hae-yeong, vice-president of the society, and others wrote articles that utilized the term 'society.' The impact of this bulletin was limited but it took a relatively short time for the newly imported concept to become a commonly used word in Korea. By the early twentieth century, *sahoe* was in use in the articles of journalists and intellectuals educated in Japan.

As with Japanese adaptation of the conceptual apparatus of sociology, the process of translation was not a simple one-way expansion of Western thought, but a deep entanglement of Western, Japanese, and Korean conceptual universes. *Sahoe* was linguistically and conceptually different to *shakai*, even though both originated with the same Chinese characters as oral Korean was the real medium of communication in the Korean sphere of the public opinion not the written Chinese form understood in all three countries of East Asia.

This entanglement became more complex with Japan's colonization of Korea in 1910. The sharp divergence of spoken Japanese and Korean proved a flashpoint of conflict with colonial authorities. Koreans understood *sahoe* from the vantage point of their own semantic paradigm and experiences. Semantic independence could easily be mobilized as a cultural resource for political movement for national sovereignty. In order to suppress this sphere of independence, the Japanese colonizers tried to enforce the policy of using Japanese language in official life-provoking huge resistance.

By the 1920s, the notion of 'Korean society' (*Joseon Sahoe*) was in widespread usage. A number of conditions contributed to this enlarged meaning of society: emergence of a public sphere with consolidation of two famous vernacular daily newspapers, *Dong-A Ilbo* and *Chosun* Ilbo; an increase of scholarly and political debates in journals; the rapid growth of newly formed social organizations and social movements; and the increasing influence of socialism. Those conditions were all accepted as an interrelated phenomenon in that 'social' groups were engaged in a changing 'society,' which had to be analyzed by 'social' ideas. Socialism was translated into '*sahoe*-ism,' which was easily understood an ideology related with '*sahoe*.' Many collective movements both for national liberation and for class struggle were called '*sahoe*-movements.' The ideas and thoughts that supported these collective movements were named as '*sahoe*-thought.'

Under colonial rule, sociology generally served different and conflicting purposes. On one side, it was an instrument of Japanese colonial scholars who designed programs of sociological research to legitimatize colonial rule. Thus diverse research projects mapping Korean traditional folk culture, kinship structure, traditional organization, local markets, and religious life were carried out by scholars trained in sociology and anthropology. On the other side, the development of sociological knowledge and sociology's conceptual apparatus was utilized broadly by oppositional movements. Through this, sociology helped establish a terrain of conceptual tension, albeit within a tightly controlled public sphere.

The very meaning of 'society' was disputed. *Sahoe* carried three diverse connotations of 'society': a contract of independent individuals, the public character of social associations, and morality of integration. Developments in the late-nineteenth century created fertile conditions for the development of social contract theory. The abolition of traditional status estates in 1894 and rapid weakening of Confucianism provided a receptive environment for accepting the idea of individualism. The new concept of society was welcomed since it gave a chance to develop individuality against traditional conceptions of humanity. With the circulation of *sahoe*, a paradigm of three structures, 'individual-society-state,' came to be widely accepted. There was a plethora of essays written about the difference between state and society and the difference between individual and society. Society was regarded as a medium, buffer, linkage, and center between state and individual. It allowed imagination of a new space where independent individuals found their own position.

Another characteristic of *sahoe* was the legalization of voluntary establishment of associations and groups. This formalized an embryonic associationalism to which the concept of society could be applied. 'Society' suited the set of organizations that emerged. These applications in usages of the concept of society resulted in the dismantling of traditional corporate boundaries delimiting the public. Up until 1894 – when the social status system was abolished – the public was monopolized by the *yangban* class centering on the king, and any public authority or activities to criticize or refuse this monopoly were suppressed. Consolidation of the concept of society provided an occasion to understand a new public sphere that was different from the state and the ruling class. It was at this time that a word for public opinion was also introduced. *Yeoron* displaced the word of *kongron* (공론). Even though there were some disagreements regarding the effects of self-interest-oriented groups, the idea of 'the public' had changed as a result of increased interest and discussion about the idea of society.

A third important connotation of *sahoe* was a perspective that social community was an organic entity with moral solidarity. In textbooks strongly influenced by the governing Japanese, 'society' was explained in relation to the goal of moral consensus. This understanding of society as an organism carried a completely different implication from the social contract theory. In place of a tension between the social and the political, a general and fundamental institutional integration of individual, family, and state was emphasized. In this third connotation, 'society' was conceptualized with an uncritical set of depoliticized moral principles. Most Korean intellectuals could not accept that depoliticized notion of the public. Nationalists argued that public opinion had to be found within the people who had kept strong orientation to independence. For them, Korean society meant Korean national identity and hope. With increasing influence of socialism, intellectuals of the Left argued that society was a historical community characterized by class conflict, not by moral integration. In addition they argued that such struggle rather than moral harmony was the real driving force of historical development including socio-political independence.

During almost four decades of colonial rule, Japan tried to depoliticize the concept of society in colonial Korea, while Koreans continuously stressed the political implication of social elements for the national independence. It was not only a political conflict but also a cultural struggle for hegemony. The concept of society was freely used from the 1920s onward, but its meaning was not unitary, and its political implication also contained heterogeneous elements in many instances of its usage.

Conclusion

Taken as a whole, the circulation of translated sociological concepts at this time in East Asia can be seen as regional process made up of three separate (though linked) national episodes. As a case of intercivilizational encounters the translation of sociology and sociological concepts into Japanese, Chinese, and Korean is exceptional in that Western colonialism was not directly involved, though Japanese colonialism had a hand in Korea's case. East Asia has been a distinct site state of intercultural exchange in this period where the context of colonial power is different to other world regions.

See also: Asia, Sociocultural Aspects: China; Civilization, Concept and History of; Colonization and Colonialism, History of; Japan: Sociocultural Aspects; Spencer, Herbert (1820–1903).

Bibliography

Braisted, W.R., 1976. Meiroku Zasshi: Journal of the Japanese Enlightenment. Harvard University Press, Cambridge, MA.

Chen, S.L.(陈旭麓), 1984. 'Wuxu shiqi weixinpai de shehui guan – qunxue'(戊戌时期维新派的社会观—群学 the sociological view of the reformists in wuxu era – qunxue). Studies of Modern History (近代史研究).

Ding, Yi, 1988. An investigation into the first introduction of western sociology into China. Study of Sociology 6, 87.

Han, M.M.(韩明模), 1987. Zhongguo Shehuixue Shi (中国社会学史 A History of Chinese Sociology). People's Publihser, Tiangjing.

Harrell, 1992. Sowing the Seeds of Change: Chinese Students, Japanese Teachers. Stanford University Press, Stanford, CA.

Howland, D., 2002. Translating the West: Language and Political Reason in Nineteenth Century Japan. University of Hawaii, Honolulu, HI.

Huang, M.K.W.(黄克武), 2011. The war of neologisms: the competition between the newly translated terms invented by Yan Fu and by the Japanese in the late qing. Approaches to Translation Studies 34.

Motoyama, Y., Elisonas, J.S.A., Rubinger, R. (Eds.), 1997. Proliferating Talent: Essays on Politics, Thought and Education in the Meiji Era. University of Hawaii Press, Honolulu, HI.

Nish, I. (Ed.), 1998. The Iwakura Mission in America and Europe, a New Assessment. Curzon Press, Surrey.

Nozomu, K., 1994. Sociology and Society in Japan. Kegan Paul International, London.

Park, Myoung-Kyu (Bak, Myeong-gyu), 2000. "Hanmal sahoe gaenyeom-ui suyong-gwa geu uimi chegye-e gwanhan yeongu" (On the acceptance of the concept of society and its system of meaning in the Late Great Han Empire period). Sahoe-wa Yeoksa (Society and History) 58.

Sako, T., 2006. Japanese early importation of the Western concepts 'society' and 'sociology'. Bulletin of Seikei University 38 (5).

Sako, T., 2007. The founding fathers of Japanese sociology. In: Sako, T., Steinmetz, S.K. (Eds.), Japanese Family and Society. Words from Tongo Takebe, a Meiji Era Sociologist. The Hamworth Press, New York.

Yamashita, S., 1984. Herbert spencer and Meiji Japan. In: Conroy, H., et al. (Eds.), Japan in Transition: Thought and Action in the Meiji Era 1868–1912. Associated University Press, London.

Yang, Y.B.(杨雅彬), 1987. Zhongguo Shehuixue Shi (中国社会学史 A History of Chinese Sociology). People's Publishing, Shangdong.

Yao, C.A.(姚纯安), 2006. Shehuixue zai jindai zhongguo de jincheng 1895–1919 (社会学在近代中国的进程 1895–1919 The Progress of Sociology in Contemporary China). SDX Joint Publishing, Beijing.

Yen, M.(阎明), 2004. Yimen Xueke Yu Yige Shidai (一门学科与一个时代：社会学在中国 a Discipline and a Time: Sociology in China). Tsinghua University Press, Beijing.

Zhang, Z.(張琢), 1992. Zhongguo Shehui He Shuhuixue Bainian Shi (中國社會和社會學百年史 a Centurial History of Chinese Society and Sociology). Chung Hwa Book Co., Hong Kong.

Zheng, H.S.(郑杭生), Li, Y.S.(李迎生), 2003. A History of Chinese Sociology (Zhang Chun, Trans. English (张春)). China Remin University Press, Beijing.

Transnational Citizenship

Claus Leggewie, University of Giessen, Germany; and Kulturwissenschaftliches Institut (KWI), Institute for Advanced Study in the Humanities, Essen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article summarizes the growing literature on citizenship rules which have been conceived during the era of the nation-state and are now under pressure through transnational developments (globalization), by the multiple citizenship adherences within, for example, the European Union (EU), and through migratory as well as refugee movements of all kinds. Concepts to regularize these developments by multinational citizenship rights and/or a European citizenship as such are in debate. First attempts to institutionalize supranational citizen rights within the EU are discussed as approaches to a broader and more inclusive citizenship regime.

Citizenship: Legal and Cultural Dimensions

Citizenship can mean a lot of things, as described by Bauböck:

> 1. a *formal legal status* that links individuals to a state or another established polity (such as the European Union or a federal province); 2. a *bundle of legal rights and duties* associated with this status, including civil liberties, rights to democratic representation, and *social rights* to education, health care, and protection from poverty risks; 3. a *set of responsibilities, virtues, and practices* that support democratic self-government; 4. a *collective identity* that can be shared across distinctions of class, race, gender, religion, ethnic origin, or way of life.
>
> 2008: 3

Bauböck connects legal and political aspects of status and participation with cultural and identitarian phenomena that have always intermingled within the framework of modern nation-states (Bader, 1997; Fox, 2005). Indeed, various efforts were made to resolve the constant tension between the legal and the political features of citizenship and their cultural underpinnings via nation-building: some states were based on linking political status to ethnic descent, whereas most republics pretended to be 'culture blind.' However, immigration and transnational cultural interconnections have relativized these typologies, and economic and cultural globalization has definitely transcended the legal and political boundaries of sovereign nation-states.

Even before the formation of the modern nation-state, cosmopolitan aspirations were already emerging: *Me velle civis totius mundi non civis oppidi* was the answer of Erasmus of Rotterdam in the sixteenth century, when Swiss Reformation leader Ulrich Zwingli offered him citizenship in Zurich. The humanist philosopher is said to have replied that he did not want to become a citizen of a single city, but of the 'entire world.' World citizenship in this sense was always one of humanity's dreams, one considered as honorable as it was unfulfillable. Now globalization has not only razed the fortress of the nation-state (as progress in ballistics did the fortified walls of medieval city-states), it has also created the opportunity for what is called *transnational citizenship*. Historically, nations provided the framework of modern community and society; they defined (and limited) the space of communication within which political parties and interest groups of all kinds operated, thus creating the prerequisites for equal representation and civic participation. In the welfare state, too, who should be included in and who should be excluded from the system of mutual support was defined within this national framework. But the universalistic principle of inclusion itself helped migrants gain a highly valued status as residents, without being formal citizens. A person's national citizenship was thus uncoupled from his or her right to have rights, insofar as international laws, for example, to protect migrant labor, and human rights conventions became plausible as independent sources of individual and collective rights. The question at hand is: Between local patriotism and global markets, where is *homo politicus* located?

World Citizenship: Facets and Drivers of Transnationalization

The *Oxford Dictionary* dates the emergence of the term 'transnational' to about 1920, as documented by a quotation from an economic text that described Europe after World War I as being characterized by its 'international or more correctly transnational economy.' Interestingly, the dictionary provided another source identifying the Christian Church as the sole power that could create the conditions for a 'transnational, nonracial democratic polity.' One author, who had 'transnational America' in view as early as 1916, is not mentioned: Randolph S. Bourne. While the nations of Europe fought a savage war, this literary critic and pacifist from New York emphasized the special potential of the USA, something he thought could provide the bases for a future world society. The USA, as the 'first new nation' of immigrants, clearly did not have recourse to the European bases – either ethnic-cultural or state-bureaucratic – for nation-building and collective identity. On the other hand, Bourne rejected the then current idea of a 'melting pot': "We are all foreign-born abroad or the descendants of foreign-born, and if distinctions are to be made between us they should rightly be on some other ground than indigenousness." One should not seek the foundations of American collective identity in a mystified past, as was the case with European nationalism, instead "we must perpetrate the paradox that our American cultural tradition lies in the future." Bourne drew the conclusion, "America is coming to be, not a nationality but a trans-nationality, a weaving back and forth

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64014-X

with other lands, of many threads of all sizes and colors" (Bourne, 1916: 87, 92, 96).

At the beginning of the twenty-first century, 'transnational' has become a technical, but also a fashionable term. It is used especially in connection with international organizations and the 'multis' (transnational corporations). In accordance with the original meaning of the Latin prefix 'trans,' it points beyond the usual diplomatic dealings among sovereign nation-states and the customary division of labor among 'national economies'; the unit of analysis, not only in international relations, has become *Weltgesellschaft* (world society, see Luhmann, 2000). This did not imply the 'end of nation-states'; however, state sovereignty as we knew it has become 'anachronistic' (Czempiel, 1969). From the macroperspective or bird's-eye view of the world economy, the glance drifts to the microlevel of *Lebenswelt* (lifeworld), where we can trace the cross-boundary paths via which people and objects, metaphors and symbols, individual life histories, and collective biographies are transferred. It is time for social science and political actors to acknowledge this 'paradigm shift' from international to transnational relations, which is also to recognize that a new form of world politics is emerging: citizenship (and governance) beyond the nation-state.

Economic and Cultural Preconditions of Transnationality

From the beginning, the most powerful and conspicuous relativizers of the nation-state were the transnational corporations, with their world-spanning trade and finance transactions. There has been a world market since sailing and techniques of navigation first permitted intercontinental trade, currency exchange, and investment; the Hanseatic League or the spread of Chinese family households throughout the Pacific demonstrates that the postmodern subversion of national borders has its premodern precursors. Yet, capital's inherent 'lack of patriotism' (*Vaterlandslosigkeit*, Karl Marx) has meanwhile developed beyond the customary relations between parent companies and foreign subsidiaries; economic and financial conglomerates have merged so that, aside from having a legal seat, they are no longer anchored in one particular nation. E-commerce, including business-to-business deals, has further radicalized this despatialization.

Certainly, the prospects of the new economy must be viewed with caution, but state and cultural boundaries already play hardly their former rule in its leading branches – telecommunications, health care, and biotechnology. Elimination of boundaries through 'digital capitalism' already has dramatic consequences for taxes and customs – which, after all, are the primary sources of revenue for nation-states and even supranational economic communities like the European Union (EU). Global standards of management and services are joined by the harmonization of trade and economic law through legal firms, auditing companies, and consultancies. These need not really have global reaches; what we observe is rather a regionalization (in the sense of 'continentalization') of economic domains (North American Free Trade Association, EU, AEAN, Mercorsur, ECOWAS) or a geographic subdivision, as with the executive board of Internet Corporation for Assigned Names and Numbers and other transnational private–public institutions.

Migrant Communities

These examples show that transnational spaces have emerged since 1945. Operating in the interstices is a 'business class,' a loose conglomerate of managers, consultants, and service providers more closely associated with companies and concepts (e.g., the 'Apple culture' and the 'Linux movement') than with native countries and native tongues. Border-crossing mobility is not limited to the upper echelons of the labor market; even unskilled workers with precarious residence status have meanwhile become highly mobile and multicultural. This *Völkerwanderung* (mass migration), joined by masses of foreign travelers, commuters, and retirees in the Sunny South, has stimulated much thought about 'transnational social spaces' (Faist, 2000). Today's nomads differ from their long-familiar precursors, especially those caught up in the streams of immigration to the USA, Canada, Australia, and France in the nineteenth and early twentieth centuries. Today, 'homeland' and 'host country' are more closely connected and networks between them are much denser. With cheap transportation and easy communication, not to mention the worldwide reception of 'home' television and radio broadcasts via cable, satellite, and Internet, permanent settlement no longer seems necessary. Bags were always packed and ready in the Diaspora, but usually they remained in the attic; the back and forth of those times – for example, between New York's Lower East Side and southern Italy or between Chicago and Warsaw – also lacked the emotional routine that now characterizes transnational mobility, and above all it did not have the effects on the structure of entire societies that transnational communities are likely to have today (Ong, 1999; Schmidtke and Ozcurumez, 2007).

We can now define more precisely as transnational a social field, transcending national affiliation, in which a growing number of people lead a kind of double life. The expropriation of farmers, which tore the agrarian population from the 'soil' and drove them into the industrial cities, now finds its counterpart in a kind of 'expropriation of borders.' This casts doubt on sociology's 'container vision'; its spatial metaphors referring to closed national societies now seem outmoded. For many transmigrants live long term in two or more places, constantly speak two or more languages, possess *en masse* two or more passports, and pass continually in both directions through makeshift households, networks of relationships, and spaces of communication. From this perspective, migration 'routes' count more for cultural studies than the 'roots' of a migrant's personal identity in national collectivities (Clifford, 1997; Hannertz, 1996). All of which, perhaps combined with and crowned by the World Wide Web, a medium of transnationalization *par excellence*, leads to a despatialization that, unlike classic emigration, permits 'virtual proximity.' This allows communities to maintain themselves even without constant face-to-face encounters, as the poles – 'at home' and 'abroad' – of such transitory existence finally become almost interchangeable.

Religious Pluralism

Wanderers between the worlds generally bring their native gods and rites in their bags, for people who are supposed to have lost

their ties are especially likely to seek 'ties back.' The Latin of 'ties back' (*re-ligere*) is in fact the literal translation of 'religion.' Religious communities have always transcended national borders; spiritual belief is especially suitable for 'stretching' and reconstructing in what, following the scattering of Jewish and early Christian congregations, became known as 'the Diaspora.' While the supralocal spread of religious ideas and communities is not new in principle (the Catholic Church can be seen as the first ever agent of globalization during the Roman Empire in the Mediterranean world), the transnationalization of religion is no longer identical to Christianization, nor has secularization, which became typical for the Christian societies of Europe, prevailed as a global model.

Though religious life, along with the dominant cultures based on it, is relativized by transnational migration, emigration can revive religious feelings of belonging. In regard to world society, religion pushes globalization forward precisely by resisting the profane effects of economic globalization and forming particularist congregations. Diaspora, once experienced as catastrophic, is no longer an exception in the religious pluralism of our time; at the same time, protection of religious freedom has improved all over the world, so that imported religious symbols are becoming visible again in the public space of secularized societies. Globalization is thus not limited to corporate mergers, Internet communication, and financial transactions. Religious communities, too, travel around the globe, and not only with the organizational and functional support of world churches, but also and especially as decentralized grassroots movements, as an unofficial, heterodox, self-founding, religious civil society.

Hybrid Popular Mass Cultures

That which can be said about religious coexistence is also true for the diffusion and mixing of multicultural societies in general. A characteristic of this world of hyphens is provided by 'world music' (e.g., in the work of Peter Gabriel or the Putumayo label) – a collage of disparate materials that often fuses completely contrary styles. In such mixtures, which are described as 'hybrid' in allusion to biological crosses, things grow together that would not sound compatible to any purist's ear. But proponents of musical and other cultural mixtures point to more or less successful combinations dating all the way back to the origins of art. Indeed, on the whole, the creativity of cultures is never based on keeping the foreign at a distance. Rather, it is based on 'extraordinary' borrowings, on mimetic appropriation and the constant exchange of inventions, on innovation at the margins, and on the acquisition of the seemingly unassimilable. Seen in this way, phenomena of 'world culture' merely signal a further stage in the hybridization of hybrid cultures. All efforts toward canonical restoration are futile and the barriers between high and low culture have been razed too (Appadurai, 1996; Tomlinson, 1999).

The 'wild' combining of artifacts, symbols, and identities began as the primary means of expression and communication; language itself underwent 'creolization' at the points of contact between two or more groups of speakers. The process of assimilation was usually asymmetrical and often violent, but, at the same time, hybrid constellations mirror the growing interdependence of world society, which of course excludes as often as it includes. In this process, components of all cultures are uncoupled from local roots and contexts; this has often been interpreted as 'standardization' and, since the origins of hegemonial popular culture can often be traced to the United States, as Americanization. There is much to be said for this assumption, if we observe the selectivity and modes of influence of the culture industry, the effects of which sometimes amount to a kind of cultural extinction; on the other hand, fulfilling the world society's structural demands does not lead to complete cultural uniformity.

In the terminology of cultural anthropology, culture must be conceived as bounded and embedded, and thus as a field that can be cultivated only in a particular time and place. The global economy and technology require and produce universal phenomena like money and standardized systems of experts, which are of necessity oblivious to local origins. Abstraction is the only thing that permits communication and creates trust. And although cultural enclaves do resist standardization, it is now rare for this aloofness to achieve the power of a fundamentalist resistance aimed at preserving one's own culture. For however much desires for distinctions can be agitated and pitted against each other in the 'culture wars,' in the end, they remain oriented toward a structure of common differences. Popular mass culture permits travel through time and space without requiring viewers and listeners to leave their homes. For the first time, 'major events' – including warlike conflicts, athletic competitions, and appearances by entertainment stars – reach global audiences via electronic mass media such as CNN and MTV. There are topics that enthrall literally the entire world, and there are growing audiences for global stagings. But this thematic concentration is accompanied by a fragmentation of local audiences, which again illustrates the dialectics of globalization and localization. Social media are the latest proof of this glocalization of communication.

Transnational Citizenship: Two Challenges of Constructing a Demos beyond Borders

These tendencies have all relativized the traditional concept of citizenship as being bound to the entitlements and duties determined by a nation-state. If the nation, in combination with government (incorporating the mechanisms of bureaucratic agencies and popular representation), has been the fulcrum of personal identity and the condition for social affiliation and membership since the nineteenth century, at the beginning of the twenty-first century, flexible forms of belonging and community transcending the nation-state challenge the representivity and legitimation of democratic rule. Now the question is: Between local cultures and global markets, how can a cross-border demos be constructed?

Transnationalization of Citizenship via Migration

Emigration and immigration have been the main causes of a growing incongruence between cultural and political collective identities: Emigrants from country A turned their back to their former homeland and forged new affinities in country B even if many of them retained some emotional ties to their country of origin (A) and tended to assimilate to their

country of choice (B) via compatriot communities from country A. The same is true for immigrants: While they centered their lives on a new location (B) they retained more or less strong loyalties to their 'heimat' (A), sending money home and keeping up-to-date on what was going on there while continuing to speak their mother tongue (van Bochove et al., 2010; Kaya, 2012).

As already mentioned, cheap airfares, mass-produced communication devices with a global reach, and personal contacts via social media have further relativized the dichotomy between country A and B; transnational communities have emerged in the lifeworld. Thus, expatriates may want to participate in the political life of their former homeland and may even be inclined to vote in their former constituencies or as a particular voting bloc of nonresident citizens from abroad, and immigrants may stipulate voting and other participatory rights to overcome their status as noncitizen residents (Bauböck, 2005; Lopez-Guerra, 2005; Rubio-Marín, 2006; Owen, 2011). As immigrants here (A) are expatriates there (B), they might even have the idea of being a part of two 'demoi' and contributing to the political agenda of two states. It was exactly this prospect that conjured up the evil of 'illoyality' in the minds of the administrative agencies in the Westphalian system of distinct and incongruous nation-states (Benhabib, 2006; Schmidtke and Ozcurumez, 2007).

Nevertheless, nation-states and political elites have allowed not only for the electoral inclusion of citizens living abroad (which is supported by ethnic conceptions of nationhood beyond a territorial state), but also for electoral rights of noncitizen residents. One reason is the desirability of the political integration of immigrants, underpinned by civil and social rights; other explanations can be drawn from historical links (e.g., a common, immemorial past, the Commonwealth, etc.), cultural affinities, and linguistic community. Bauböck (2005) has sketched four ideal-typical positions speaking for or against this undermining of traditional concepts of national citizenship:

1. Civic republicans would argue that "only citizens who are present in the polity can govern themselves by participating in making laws" (685); a republic may be open to newcomers but those should apply for full citizen rights by naturalization and giving up former citizenship(s).
2. Ethnic nationalists would "support the inclusion of expatriates but reject political rights for noncitizen residents" because they conceive the nation as a community of culture.
3. Liberal democrats would, on the contrary, hold that voting rights for expatriates might undermine "the integrity of the democratic process since those who live permanently abroad should not be able to influence the making of laws to which only internal residents will be subjected" (686, see also Lopez-Guerra, 2005); however, they would favor voting rights for noncitizens just because long-term residents are effectively subjected to the local political authority and its laws, so noncitizens should therefore have "equal rights to representation and participation in the making of these laws" (686).
4. Cosmopolitans would claim for voting rights and participation on the grounds of an all-affected principle: *quod omnes tangit ab omnibus approbetur* (what concerns all must be approved by all). In this case, the demos is not a given entity of people but is constructed by way of those decisions that have a "profound impact on the interests of another country's population" (686).

As one can object to all the four concepts, Bauböck introduces a fifth principle he calls 'stakeholder citizenship' (2008). This involves a mixture of republican and liberal cosmopolitan principles. With regard to the former, it retains the idea that citizenship guarantees the status of full membership in a self-governing polity and that voting rights should generally be attached to such status. In accordance with liberal cosmopolitan principles, stakeholder citizenship would give stakeholders a subjective claim to membership and electoral rights. Bauböck provides a balanced conclusion: "Stakeholdership should … be less vague and overinclusive than affected interests. […] [It] would require the political inclusion of immigrants, but – different from inclusion derived from mere territorial subjection – it could justify the condition of long-term residence and the common requirement that immigrants have to apply for naturalization instead of being automatically turned into citizens. Stakeholdership would also permit (although probably not require) extending the vote to expatriates, but it would exclude those who have never lived in the country and would not give access to citizenship to persons whose interests lie in economic investment or tax evasion but who do not take up permanent residence" (2005: 686).

Transnationalization of Citizenship via Stakeholdership and Participation

Emigration and immigration have given rise to a revision of common understandings of how a people is constituted and what democracy as rule of the people means. The literature distinguishes two principles – the all-subjected principle, where the guiding idea is "that those subject to a rule also should be its authors"; and the aforementioned all-affected principle "that all those affected by a political decision ought, directly or indirectly, to have a say in its making" (Näsström, 2011: 120, 117). The first rule starts from a given people, be it the (very exclusive) demos of the Greek polis or the (more inclusive) demos of a modern republic. The second rule shifts the focus from a given people within polis and nation-state boundaries to (the justifying of) people-making within a democracy whose political agenda becomes transnational as well as transcultural. This does not only mean that "the demos needs to extend to everyone in a particular territory" (Gould, 2006: 49) but also enlarges the citizenry beyond the city walls and state borders. The all-affected principle negotiates and politicizes the boundaries of decision making.

With regard to the watering down of national sovereignties, all citizenship regulations tend to be conservative and exclusive, and the naturalization processes within most states continue to remain rigid and dismissive. As already shown, mass immigration has put the accepted status of national citizenship in doubt and the number of people with multiple citizenships in Europe has seen considerable growth. This growth in the possession of more than one nationality is, however, viewed with suspicion by many states, as they still believe it to promote conflicts of loyalty. The children of immigrants are, therefore, often forced as adults to decide

between their cultural and geographical origins. For example, second generation Turkish settlers in Germany will be made to decide between becoming Turkish or German; a decision which could then ultimately lead to a loss of their EU citizenship. Something which of course does not happen in families of mixed European citizenship, where, should the offspring decide to become French, Dutch, German, or British, they would nonetheless retain their EU citizenship.

Nevertheless, the national monopoly on citizenship is gradually beginning to weaken. A great many practical and normative arguments exist for the granting of transnational citizenships and the development of voting rights for foreign nationals. The economic and cultural impacts of globalization are, for example, gradually creating scenarios, where, within many areas, long-term foreigners are able to claim a greater say in what goes on around them. Meanwhile, in law scholarship and within the social sciences, the term 'global rights' is gaining in influence. Global rights are rights granted to people independent of their locations and are supported by wide-ranging human rights standards. Some of these standards themselves contradict national laws, and yet their impacts can, as in the case of the German Asylum Reform of 1992, often only be reduced through procedural limitations.

Such international developments suggest that reforms of the state citizenship regime are overdue and that these reforms should lead to more inclusive and flexible judicial statuses for people worldwide. Bauböck has repeatedly proposed the 'stakeholder principle' as a broader successor concept to the idea of citizenship: "Stakeholders are those who have a stake in the polity's future because of the circumstances of their lives." The so-called 'all-affected principle' goes a step further. It suggests that we can all be affected by global catastrophes and that therefore, we are indeed all 'all affected.' And if the role of nature as protagonist can be recognized in this sense, then why not also consider the proposal, that, not only humans but also nonhuman sentients and other objects or phenomena should be formed into a 'parliament of the things' (Latour)? Furthermore, where the rights of future generations play a role in contemporary decision making, there has been debate regarding the consideration of specific voting rights for children, to be exercised *in absentia* by their parents (Goerres and Thieman, 2009).

One note on spatial inclusion: Suppose we take the term 'global society' as being not just a metaphor (as do the majority of sociologists), but a reality forged by communications, transport, and free exchange within a globalized social sphere. It can then be argued that the space within which to exclude someone consequently ceases to exist and that societal exclusions become completely unjustifiable (Stichweh, 1997). Stichweh came to the conclusion that exclusions nowadays are only internal and temporary and that a dynamic exists, within which an act of inclusion occurs with every (attempted) act of exclusion. This is best displayed in total institutions. For example, in a jail, criminals are taken out of the society within which they would normally exist, yet in most circumstances this occurs with the explicit aim of then ensuring their ultimate reintegration into the society from which they have temporarily been removed.

Another note on generational inclusion: The potential heteronomy of future generations will ultimately be determined by their own future circumstances. People today simply cannot judge what coming generations will do and are equally unable to preordain what they themselves should do or allow to be done to them. The concept of sustainability (and indeed that of budget balancing as recently framed within the fiscal pact in many OEEC states) has, however, inaugurated an important proviso. Namely, it is advisable for people living within a dynamic, growing, global society to not only not use more of their resources than can then be realistically regenerated, but also abstain from impinging so much upon future generations, that dangerous, existence-threatening tipping points are both reached and surpassed.

A graphic example of unacceptable levels of heteronomy for future peer groups is the program for the 'temporary' storage of atomic waste, an issue for most EU member states. As part of this program, waste is temporarily placed in a location while its final resting place is discussed. These discussions, however, tend to last at least two generations, despite the fact that the radioactive outfall from this temporary placement can have a lasting impact upon an area for thousands of years. In this example, the temporal and spatial elements of concern are therefore united, as the risks of atomic energy hardly allow themselves to be confined to either national territories or specific periods of time. Right now, numerous atomic waste shipments are traveling across national boundaries throughout Europe. An international supervisory commission, however, only exists in the most rudimentary of forms and this despite the fact that a recent EU Commission report stated that the majority of European reactors are insufficiently secured. This example illustrates how political stakeholders can be on the receiving end of international decisions, regardless of state and/or cultural boundaries (Mason, 2009; Bauböck, 2009b; Owen, 2011).

It would, however, likely be difficult to reconcile stakeholder citizenship with the present system of state citizenship. 'Cultural' grounds will doubtless be raised. By which I mean the ethnic and religious differences between nations, which, from a traditionalist perspective, often preclude supranational integration. It is said, for example, that supranational bodies pay too little attention to the respective particularities of individual groups (such reasoning is presently being used by various groups in Spain, Belgium, and Scotland to demand greater regional independence from the central governments in these countries). Naturally, such objections tend to ignore the binding relationships between various peoples across national borders. Historically, these cross-border affinities have, in fact, been strongly emphasized, and never more so than in postwar societies where long-standing enmities are gradually reduced through transcultural exchanges and political cooperation. An exemplary model for this can be seen in the case of Germany and France, while moves are also being made in this direction in, for example, the ethnic and religious communities of the countries of the former Yugoslavia.

Prospects of Supranational Citizenship: The Case of the EU

The challenge then is to develop a concept of citizenship that is wide enough to address the ways in which national citizens are affected by global problems of all kinds (e.g., climate change) and concrete enough to represent the identitarian and communitarian bases of being a citizen in the context of

a particular *Lebenswelt*. The EU, until now, provides only a rather embryonic form of citizenship, limited to a formal legal status. The existence or even possibility of a European demos is disputed and contested. An EU citizenship per se does not yet exist as the individual EU states continue to defend their right to give and take citizenship as a national privilege. Behind this fiercely defended national right lurk the ancient, never fully applied, and increasingly undermined myths of the national union of land, culture, and ethnology. Whether German, Polish, or French, citizenship is determined either through birth or through a demanding process of naturalization. In the case of official national identity, it would appear that there is, as yet, no third way.

For the success of such supranational ventures, high levels of empathy and well-thought-out routines for cooperation are prerequisite, as well as the ability and desire to understand your counterparts. What is, however, uncertain is whether a transnational citizenship then subsequently develops out of such a process, bestowing the citizens concerned with institutionally secured and enforceable rights and duties. In other words, rights that are not just morally demanded but are also successfully exercised. A citizen of the EU is, according to Article 20 of the treaty on the functioning of the EU, an individual who possesses the nationality of an EU member state. Citizenship of the EU is therefore an extension, rather than a replacement, of national citizenship. European citizenship rights include the right to appeal to EU representatives, to instigate new legal proposals, and to participate, both actively and passively, in elections relating to the various European institutions. Additional rights bestowed upon European citizens include the freedom of movement within and between EU territories and certain residential rights for both them and their dependents. These residential rights have also been streamlined in order to reduce to a minimum the administrative effort involved in establishing one's entitlement to them. Limitations have also been placed on the circumstances within which these rights can be refused. In fact, every citizen of an EU nation has the right to permanent residency within any other EU state so long as he or she has spent at least 5 years uninterrupted within the said state's territory without committing any offenses for which action has had to be taken against him or her. This right comes without further requirements and also applies to any familial dependents of the individual concerned – regardless of their own nationality – so far as they themselves have also spent 5 years uninterrupted within a member state. Both EU citizens and their dependents also have a right to the diplomatic and consular services of all EU member states when in a country not of the EU.

How then would the application of a regional 'all-affected' principle build upon the status quo? As regards a statement of transnational citizenship, the participational rights of residents within EU states could be improved upon by adopting the proposals suggested at the Arhus Convention: for example, enabling citizens to participate in national votes, citizen initiatives, or class actions (Wallrabenstein, 2012).

The following European programs are already in place to promote similar initiatives and should lead to changes in this direction:

- Council Decision 2010/37/EC of 27 November 2009 on the European Year of Voluntary Activities Promoting Active Citizenship (2011) aimed at creating conditions that enhance the *participation of civil society in voluntary activities*, as well as increasing the visibility of volunteering.
- Council Decision 2007/252/EC of 19 April 2007 has established for the period 2007–13 the programme 'Fundamental Rights and Citizenship' as part of the General Programme 'Fundamental Rights and Justice'; it aims to promote the *development of a European society based on respect for fundamental rights*, to strengthen civil society and encourage an open and transparent dialogue, to fight racism and xenophobia, and to improve mutual understanding between the judicial and administrative authorities and the legal professions.
- Decision No 1904/2006/EC of the European Parliament and of the Council of 12 December 2006 established for the period 2007–13 the programme 'Europe for Citizens' to promote active European citizenship; it may create the *preconditions for a European demos by developing a European identity among European citizens* based on recognized common values, history, and culture, and by fostering a sense of ownership of the EU among its citizens.

More concrete was the Charter of Fundamental Rights that recognizes a range of personal, civil, political, economic, and social rights of EU citizens and residents, enshrining them into EU law. It is grouped into six major fields which I would like to quote extensively because they are almost unknown to the most European citizens:

Chapter I: *dignity* (human dignity, the right to life, the right to the integrity of the person, prohibition of torture and inhuman or degrading treatment or punishment, prohibition of slavery and forced labour);

Chapter II: *freedoms* (the right to liberty and security, respect for private and family life, protection of personal data, the right to marry and found a family, freedom of thought, conscience and religion, freedom of expression and information, freedom of assembly and association, freedom of the arts and sciences, the right to education, freedom to choose an occupation and the right to engage in work, freedom to conduct a business, the right to property, the right to asylum, protection in the event of removal, expulsion or extradition);

Chapter III: *equality* (equality before the law, non-discrimination, cultural, religious and linguistic diversity, equality between men and women, the rights of the child, the rights of the elderly, integration of persons with disabilities);

Chapter IV: *solidarity* (workers' right to information and consultation within the undertaking, the right of collective bargaining and action, the right of access to placement services, protection in the event of unjustified dismissal, fair and just working conditions, prohibition of child labour and protection of young people at work, family and professional life, social security and social assistance, health care, access to services of general economic interest, environmental protection, consumer protection);

Chapter V: *citizens' rights* (the right to vote and stand as a candidate at elections to the European Parliament and at municipal elections, the right to good administration, the

right of access to documents, European Ombudsman, the right to petition, freedom of movement and residence, diplomatic and consular protection);

Chapter VI: *justice* (the right to an effective remedy and a fair trial, presumption of innocence and the right of defence, principles of legality and proportionality of criminal offences and penalties, the right not to be tried or punished twice in criminal proceedings for the same criminal offence).

Since 1995, the European Ombudsman, on behalf of Europe's citizens, has been investigating cases of alleged maladministration by EU institutions or bodies, that is, the European Commission, the Council of the European Union, the European Parliament, etc. The EU citizenship report of 27 October 2010 'Dismantling the obstacles to EU citizens' rights' (COM (2010) 603 final) has identified the main obstacles that EU citizens may still confront in their private, academic, or professional life, when consuming goods and services or in their role as political actors.

Conclusions

1. There is a continuous tension between (inclusionary) citizenship and (exclusionary) cultural identity. By means of exclusion, ethnonationalism wishes to resolve this tension in a destructive and often militant fashion. As a result of its cultural blindness, republicanism dispels the tension by ignoring it. In most cases, the behavior of transmigrants who have created truly transcultural spaces must not be mistrusted by state authorities at all. They predominantly identify with the local community of their compatriots and with the wider urban community as well as with families and relatives at home – to whom they pay visits, make phone calls, and eventually send remittances. However, they rarely maintain 'thick' political loyalties relating to their country of origin. This is the case in ties to the home country based on political religions, for example, of communities from India and Pakistan. Localism and transnationalism challenge the perceptions of the established domestic populations only insofar as their ethnic and/or religious identities are concerned.
2. Global communication and cultural industries, transnational migration, and international market relations have led to the relativization of state borders, rendering purely national affiliations a fiction, and paving the way for a significant number of instances of dual or multiple nationalities. Awarding dual and multiple citizenships irrespective of cultural, ethnic, and religious background should therefore be seen as a more effective integration tool than assimilationist policies that operate on a legal level.
3. Global problems, such as climate change or other environmental risks, the crisis of financial markets, and the abuse of human rights, have led to the political demands of participation extending beyond nation-state borders and the opening up of a cosmopolitan arena beyond the nation-state and the United Nations system. Being 'all affected' is not sufficient in and of itself to guarantee effective participation; the stakeholder principle must be supported by effective citizenship rights (voting, access to the public sphere, recognition of NGOs, etc.).
4. The EU serves as an advanced example of an institution that provides a basis for supranational citizenship. Even though the EU is the result of national citizenship, it is 'overdetermined' by numerous cosmopolitan elements. Through supranational legal title, rights of participation of a moral nature are secure and can be effectively exercised. The EU thus provides a replicable model of a regional union that supports supranational citizenship.

Bibliography

Andersson, B.

Appadurai, A., 1996. Modernity at Large. Cultural Dimensions of Globalization. University of Minnesota Press, Minneapolis, MN.

Bader, V., 1997. The cultural conditions of transnational citizenship. On the interpenetration of political and ethnic cultures. Political Theory 25 (6), 771–813.

Bauböck, R., 1994. Transnational Citizenship. Membership and Rights in International Migration. Elgar, Aldershot, UK.

Bauböck, R., 2005. Expansive citizenship – voting beyond territory and membership. PS: Political Science and Politics 38 (4), 683–687.

Bauböck, R., 2008. Stakeholder Citizenship: An Idea Whose Time Has Come? MPI, Washington, DC.

Bauböck, R., 2009a. The rights and duties of external citizenship. Citizenship Studies 13 (5), 475–499.

Bauböck, R., 2009b. Global justice, freedom of movement and democratic citizenship. European Journal of Sociology 50 (1), 1–31.

Benhabib, S., 2006. Another Cosmopolitanism. Oxford University Press, New York.

van Bochove, M., Rusinovic, K., Engbersen, G., 2010. The multiplicity of citizenship: transnational and local practices and identifications of middle-class migrants. Global Networks 10 (3), 344–364.

Bourne, R., 1916. Trans-national America. Atlantic Monthly 118, 86–97.

Braidotti, R., Hanafin, P. (Eds.), 2012. After Cosmopolitanism. Routledge, London.

Clifford, J., 1997. Routes, Travel, and Translation in the Late 20th Century. Harvard University Press, Cambridge, MA.

Faist, T., 2000. The Volume and Dynamics of International Migration and Transnational Social Spaces. Clarendon Press, Oxford.

Fox, J., 2005. Unpacking transnational citizenship. Annual Review of Political Science 8, 171–201.

Goerres, A., Thieman, G., 2009. Kinder an die Macht? Mögliche Folgen eines Treuhänder-Wahlrechts für Eltern. Politische Vierteljahresschrift 50, 50–74.

Hannertz, U., 1996. Transnational Connections: Culture, People, Places. Routledge, London.

Held, D., 1995. Democracy and the Global Order. From the Modern State to Cosmopolitan Governance. Stanford University Press, Stanford, CA.

Kaya, A., 2012. Transnational citizenship: German-Turks and liberalizing citizenship regimes. Citizenship Studies 16 (2), 153–172.

Lopez-Guerra, C., 2005. Should expatriates vote? Journal of Political Philosophy 13 (2), 216–234.

Luhmann, N., 2000. Die Politik der Gesellschaft. Frankfurt/Main. Suhrkamp, Germany.

Mason, A., 2009. Environmental obligations and the limits of transnational citizenship. Political Studies 57, 280–297.

Näsström, S., 2011. The challenge of the all-affected principle. Political Studies 59, 116–134.

Ong, A., 1999. Flexible Citizenship. The Cultural Logics of Transnationality. Duke University Press, Durham, NC.

Owen, D., 2011. Transnational Citizenship and Rights of Political Participation. Normative Orders Working Paper 06/2011.

O'Brien, R., Goete, A.M., Scholte, J.A., Williams, M., 2000. Contesting Global Governance. Multilateral Economic Institutions and Global Social Movements. Cambridge University Press, Cambridge, UK.

Rosenau, J.N., Czempiel, E.O. (Eds.), 1992. Governance without Government: Order and Change in World Politics. Cambridge University Press, Cambridge, UK.

Rovisco, M., Nowicka, M. (Eds.), 2001. The Ashgate Research Companion to Cosmopolitanism. Ashgate, London.

Rubio-Marín, R., 2006. Transnational politics and the democratic nation-state: normative challenges of expatriate voting and nationality retention of emigrants. New York University Law Review 81 (1), 117–147.

Schmidtke, O., Ozcurumez, S. (Eds.), 2007. Of States, Rights, and Social Closure: Governing Migration and Citizenship. Palgrave, London.

Snel, E., Engbersen, G., Leerkes, A., 2006. Transnational involvement and social integration. Global Networks 6 (3), 285–308.
Stichweh, R., 1997. Inklusion/Exklusion, funktionale Differenzierung und die Theorie der Weltgesellschaft. Soziale Systeme 3, 123–136.
Stoker, G., et al., 2011. Prospects of Citizenship. Bloomsbury Academic, London, pp. 110–133. http://dx.doi.org/10.5040/9781849662444.
Tomlinson, J., 1999. Globalization and Culture. Chicago University Press, Chicago.
Wallrabenstein, A., 2012. In: Bäuerle, M., et al. (Eds.), Zum Vergassungsrecht der Unionsbürgerschaft. Demokrarie-Perspektiven, Tübingen, pp. 741–763.

Relevant Website

http://www.getcited.org/pub/103146517 – Benedict Anderson on Long Distance Nationalism.

Transnational Organized Crime

Toine Spapens, Tilburg University, Tilburg, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Even in history, criminals crossed borders to commit crimes, traffic in illicit goods, and provide illegal services. More recently, however, opportunities for transnational (organized) crime have increased, in particular because of the growing level of mobility, the opening of borders and the advent of the Internet. So far, however, there is little proof of the emergence of 'crime multinationals.' Although a complex and growing legal and organizational framework is in place, law enforcement cooperation is still largely traditional and it remains advantageous for criminals to operate across different jurisdictions.

Introduction

This article focuses on transnational (organized) crime and the instruments developed by the international community to combat it. The fact that criminals cross the borders of jurisdictions to commit crimes, avoid arrest and conviction, and traffic in illegal goods is of course nothing new. Section Transnational Crime in History: Some Examples begins by addressing some historical examples of transnational crime. In recent decades, however, cross-border crime has expanded and diversified, in particular because of the increasing level of mobility, the opening of borders and the emergence of the Internet. Section Developments in Transnational Crime since World War II discusses these developments. Next, Section Transnational Crime Networks focuses on whether some crime syndicates, such as the Italian Mafia and the Chinese Triads, are developing into crime multinationals that span the world. The subsequent sections address how law enforcement agencies from different countries cooperate in the fight against transnational (organized) crime. Section The Legal Framework discusses the development of legal frameworks, such as the bilateral and multilateral conventions necessary for law enforcement cooperation. Section The Organizational Framework then describes the institutions, such as Interpol, that play an important part in law enforcement cooperation. Section Conclusion concludes this brief overview.

Transnational Crime in History: Some Examples

Transnational crime is in itself not a new phenomenon and there are many historical examples of criminals crossing the borders of jurisdictions to commit crimes, traffic in illegal goods and offer outlawed services such as prostitution. Banditry, for example, flourished in Europe from the seventeenth to the early nineteenth centuries. Bands of brigands committed crimes in Northwestern Europe, in the Balkans, the United Kingdom, Italy, and France (Hobsbawm, 1972; Chesney, 1979; Henry, 1984). Although some of these groups limited their activities to the local level, others were highly mobile and operated across different jurisdictions. A Dutch example is the White Feather gang. Its members resided in Kaatsheuvel in the south of the country and totaled almost a 100 members, including women and children (Grootswagers, 1983). Such gangs committed burglaries and violent robberies in the countryside and extorted money from farmers by threatening to set fire to their farmhouses unless they paid them off (Egmond, 2004). Dutch gangs also operated in what is now Belgium and in the German border areas. Their demise finally came after 1798, when the government, under French influence, introduced a unified penal code that put an end to the independent powers of several 100 local jurisdictions. Furthermore, the State assembled Gendarmerie units to operate in the countryside. In only a few short years, the gangs were history. New opportunities arose for criminals in the Netherlands when Belgium gained its independence in 1830 and it became lucrative to smuggle all sorts of goods back and forth across the new border. One of the hamlets that had once been the lair of an itinerant gang was located close to the new border, and it rapidly turned into a notorious smugglers' den. Large-scale smuggling continued until 1960, when the Benelux Customs Union came into effect (Spapens, 2012a).

Another interesting example is the international trafficking in women, which was acknowledged as a growing problem at the end of the nineteenth century. The problem of 'white slave traffic,' however, must be viewed in a broader context. From the eighteenth century onward, the Industrial Revolution brought about fundamental changes in the lives of Europeans and North Americans and sparked rapid urbanization. The social disorganization and poverty in the cities promoted large-scale alcohol abuse, prostitution, and gambling. One of the responses was a revival of moral values from the second half of the nineteenth century onward, known as 'Victorianism.' Starting in the United Kingdom, the battle against alcohol, prostitution, and gambling spread to other countries.

In the late nineteenth and early twentieth centuries, large numbers of women began to migrate to other countries for shorter or longer periods to work, for example, as servants for well-to-do families. The debate about moral values, combined with this migration (which occurred across Europe), triggered the discussion concerning white slavery (Doezema, 1999). Scholars now believe that the actual number of trafficking cases was very small, although the problem did exist, notably in connection with sexual exploitation. In Europe, white slavery was discussed at a conference organized in Paris in 1895, followed by similar conferences in London and

Budapest in 1899. It led to the International Agreement for the Suppression of the 'White Slave Traffic,' signed in Paris in 1904, in which 13 countries agreed to take measures to protect women and girls effectively against this type of crime.

In the course of the nineteenth century, the need for international cooperation in the field of cross-border crime and terrorism also increased (Fijnaut, 2010). In 1898, a number of European countries convened in Rome and agreed to join forces to combat violent anarchism (Bach Jensen, 1981). This meeting was followed by a symposium in Monaco in 1914 on how to tackle cross-border crime. After the First World War, other types of transnational crime appeared on the international agenda. In the 1920s, for example, inflation was endemic in Germany and led to the counterfeiting of money, particularly of hard currencies. Another emerging type of crime was the trafficking in stolen vehicles.

These developments in turn triggered the founding of the International Criminal Police Organization – Interpol – in Vienna in 1923. The Nazification of Interpol by the Third Reich brought the organization into disrepute, among other things because information on the Nazi's political opponents stored in its records fell into the hands of the regime. However, the need for an organization for cross-border police cooperation was evident and led to Interpol's reestablishment in as early as 1946 (Bresler, 1992).

Developments in Transnational Crime since World War II

After the Second World War, the opportunities for committing crimes across borders increased. First, from the 1960s onward, the demand for illicit drugs grew and became more diversified. Second, the fall of the Iron Curtain in 1989 created new markets for European criminals, and the 'open borders' of the Schengen zone in Europe not only made it easier for them to move around but also sparked the phenomenon of cross-border petty crime in border areas (Spapens, 2010a). Third, the steady increase in mobility had profound effects on the nature and scope of transnational organized crime. Finally, new types of cross-border crime developed along with advances in information and communication technology, such as cybercrime, but also because environmental crime emerged as a new category of crime, for example. Below, these developments are discussed in more detail.

Youth culture developed rapidly in Western countries in the 1950s and 1960s. Part of this trend was the increasing use of illicit drugs, particularly hashish and marijuana. Users and dealers in colder climates had to import the drugs from abroad, and Pakistan, Lebanon and Morocco soon became important source countries for European users. At first, the drugs were smuggled in small quantities by young people, often users themselves, who had traveled to the East. As demand increased, traditional criminals also started to discover the market, and some of the 'hippies' managed to expand their businesses as well. Demand for other types of illicit drugs, such as heroin and cocaine, increased in the 1970s and 1980s, and in the 1990s ecstasy became fashionable among large numbers of users.

The global drug market gradually expanded and now involves a considerable number of people. The United Nations Office on Drugs and Crime estimates the number of people who have used illicit drugs at least once in 2010 at about 153–300 million worldwide (UNODC, 2012). Most of these used cannabis (119–224 million), followed by synthetic drugs (14–53 million), opiates (13–21 million), and cocaine (13–20 million). About 60% of all drug users live in Asia. Heroin, cocaine, and synthetic drug production is concentrated in a relatively small number of countries, and the products must usually be trafficked over long distances to reach customers. Afghanistan is the source country of most of the poppies used to manufacture heroin. Colombia and Peru are the main producers of cocaine. Important producers of synthetic drugs are Myanmar and the United States for methamphetamine, Canada, the Netherlands, and Belgium for ecstasy, and the latter two along with the Baltic States for amphetamine (Paoli et al., 2010). By comparison, cannabis cultivation is far more widespread, particularly since the introduction of indoor cultivation methods in the early 1990s. This implies that smuggling lines for hashish and cannabis are relatively short.

It is clear that the intricate distribution network for illicit drugs involves thousands of persons in different roles. Of course, not all members of this network are globally connected. Instead, the network is 'glocal' and composed of local subnetworks of producers and consumers, linked by traders (Hobbs, 1998). There are people who grow cannabis, poppies, and coca leaves, for example, groups that manufacture cocaine, heroin, and synthetic drugs, and sell them to wholesalers, who export from several hundred up to several thousand of kilos of illicit drugs. Of course, there are also small entrepreneurs who smuggle kilos or even less. At the end of the distribution chain, we find the small dealers who are in direct contact with the customers.

The fall of the Iron Curtain in 1989, followed by the disintegration of the former Soviet Union, are the next events important for the development of transnational crime. The opening up of the former Warsaw Pact countries created new crime markets. Demand for illicit drugs increased in the East, and new smuggling routes emerged, for instance for heroin via Tajikistan to Russia (Paoli et al., 2010). In the former communist world, demand for luxury goods rose, for example, for Western-built cars. Itinerant crime groups emerged, some of whom specialized in stealing vehicles and vehicle parts in Western Europe to sell at home. Other groups, for instance originating from Southeastern Europe, temporarily based themselves in the West in order to commit all sorts of predatory crimes, such as burglary, shoplifting, armed robbery, and the skimming of debit and credit cards (Spapens, 2008).

In turn, Eastern European and former Soviet countries themselves became the sources of victims of human trafficking, who worked in Western Europe either in appalling conditions or as prostitutes. Traders in former communist countries sold their surplus of small arms and light weapons to traffickers, who delivered them to combatants in the conflict zones of Africa, for example (cf United Nations, 2001). Finally, the regime changes in, for example, Russia and Albania led to the emergence of indigenous organized crime groups who also became involved in the trafficking of illicit drugs, for example.

In Europe, the abolishment of fixed border controls in the Schengen zone made it easier for itinerant groups to move around, but it also simplified the trafficking of stolen goods

and illicit drugs across borders, particularly after the expansion of the zone in 1999. Furthermore, increasing economic and social integration in the European Union (EU) border areas after Schengen promoted a rise in cross-border petty crime, especially in densely populated border areas (Spapens, 2010a).

Generally speaking, mobility and migration increased substantially in the past two decades. In the 1990s, criminals smuggled growing numbers of Chinese migrants to Europe and North America, for example (Zhang and Chin, 2001). These trends also enabled itinerant gangs to operate across longer distances. The notorious 'Pink Panther gang,' a network originating in Serbia, which is suspected of having conducted over 100 robberies of jewelry stores in over 20 countries, is a noteworthy example. Finally, it is now far easier than in the past for migrants from different parts of the world to visit relatives in their home countries, and those who are willing and able to engage in illegal activity have less difficulty finding and exploiting illegal market opportunities.

The revolution in information and communication technology and, in particular, the rise of the Internet-created opportunities to commit new types of cross-border crime and modernized traditional forms. An example of this, documented a few years ago, is a Chinese crime group that had bought credit card information hacked by a group of Ukrainians and used it to falsify credit cards to buy designer goods in Paris and Brussels (Spapens and Fijnaut, 2005). Another example is the emergence of online gambling in the mid-1990s. It is usually illegal to operate a gambling website without a license issued by the State in which one is offering the service. The Internet, however, knows no borders and has thus disrupted the regulatory regimes based on sovereignty. It is especially problematic that most of the ultimate beneficial owners of gambling websites are unknown. Online gambling also globalized traditional types of crime, such as match fixing (Spapens, 2012b).

Finally, environmental crimes also emerged as a global concern. Of course, pollution problems plagued humans, flora and fauna in the past as well, but it was not until the 1960s that it became a crime to inflict damage on the environment. Environmental crimes may be transnational in their effect, such as the polluting of air and water. They may also concern trafficking, for example in hazardous waste from an industrialized nation to a Third World country. Mobile environmental crime may, for example, involve the captain of a ship constantly dumping waste, such as bilge water, instead of offering it to a specialized harbor facility for processing. Finally, cross-border environmental crime may even take place in cyberspace, for example, by committing fraud in emission trading schemes (Spapens, 2010c).

Transnational Crime Networks

Traditional diasporas such as the Chinese and Jewish have always played an important role in international trade. Bonds of trust between communities spread out across the globe offered a secure basis for business agreements (cf Weidenbaum and Hughes, 1996). Understandably, traders in illicit goods also made use of such networks and trust relationships. The 'Kosher connection,' for instance, played an important role in the trafficking in ecstasy from the Netherlands to the United States in the 1990s (Spapens, 2006). Of course, transnational banditry depends in the first instance on mobility, but itinerant gangs also misuse immigrant communities, for instance to obtain temporary lodgings. One example is Albanian criminals in the Belgian city of Liège, who forced members of the immigrant community to assist them by threatening to harm family members still living in Albania (Farcy, 2002). Nowadays, migrant communities from all over the world reside in the cities of the industrialized West, and (illicit) trade can easily follow from connections with associates or family members still living abroad. Whether crime groups can actually exploit such links will depend on whether there is a market, for example, for specific illicit products and services, stolen goods, and dumping waste.

Particularly since the fall of the Iron Curtain, authorities worry about the risk of organized crime networks turning into global crime multinationals (cf Strange, 1996; United Nations, 2004; Williams, 2003). Such networks may spread their activities across the world, operating from safe havens where governments and law enforcement institutions are weak (cf Godson, 2003).

Theoretically, foreign crime groups can undermine other societies in three basic ways. First, such groups may move to another country to provide additional illegal goods and services or take over the markets of existing providers. Second, crime networks may set up shop abroad to commit predatory crimes, including extortion. Third, they may invest their criminal gains in legitimate economic activities in other countries (van Dijk and Spapens, 2013).

Although it is not difficult to present examples, foreign crime groups almost never operate parallel to existing local crime networks. Instead, they are always connected through personal or business relationships. The only notable exception is truly mobile groups such as the Pink Panther gang mentioned above. Seeking cooperation is far more rewarding than trying to outcompete established local crime groups and thus sparking violent responses. Furthermore, in countries with homogeneous populations, foreign criminals stand out, and their illegal activities may more readily attract the attention of law enforcement. Finally, it is difficult to introduce new types of organized crime, such as extortion and racketeering, in countries with functioning legal systems. Ironically, it is also complicated in places where government is weak, because chances are that powerful crime groups already exist in such circumstances (van Dijk and Spapens, 2013).

The question remains whether 'masterminds' control the activities of foreign criminals who have connections with crime syndicates such as the Italian Mafia, the Chinese Triads, or the Russian Mafiya. The police often assume that this is the case, particularly if they observe a suspect who is residing in their country regularly flying back to China, Russia, or Italy. Although foreign criminals may use such connections to gain support, for instance financial assistance, there is no proof that others run the illegal operations from abroad (van Dijk and Spapens, 2013).

It is clearly easier now to establish and maintain cross-border criminal relations, to commit predatory crimes over longer distances, and to carry out certain illegal activities on the Internet than just a few decades ago. However, criminals must still operate without any legal protection in violent

surroundings and with the constant risk of colleagues robbing them, and they are therefore especially vulnerable when trying to expand their business to other countries or even within their own country (Varese, 2011).

The Legal Framework

Operating across borders presents specific difficulties for organized crime groups but it also poses a challenge to law enforcement. When talking to police practitioners, one can often hear the complaint that international law enforcement cooperation is in many respects 'not working.' Cooperation is often slow and requires all sorts of formalities. The core principle of the criminal justice system is the sovereignty of the State. As a rule, it will not accept foreign police officers operating in its territory or other countries prosecuting suspects for acts committed within its borders, with the exception of certain limited and well-defined cases. Cross-border operational law enforcement cooperation thus depends on a system of requests for mutual legal assistance.

This section briefly addresses the legal framework for cooperation in criminal investigations. It leaves aside extradition and other important instruments, such as capacity building and conventions in which States agree to take specific action against certain forms of crime. An example of the latter is the Convention on International Trade in Endangered Species of Wild Fauna (CITES), which prohibits the trade in certain species of animals and flora. However, CITES does not provide for the exchange of requests for mutual legal assistance following an actual criminal investigation.

There are conventions on mutual legal assistance at different levels. The most comprehensive in terms of participating States are the United Nations Convention against traffic in illicit drugs and psychotropic substances of 1988 and the United Nations Convention against transnational organized crime and the protocols thereto of 2004.

The first enables the parties to afford one another the widest measure of mutual legal assistance in investigations, prosecutions, and judicial proceedings in relation to criminal offenses involving illicit drugs and psychotropic substances listed by the UN. The forms of assistance available include taking evidence or statements from persons, facilitating the service of judicial documents, executing searches and seizures, examining objects and sites, and providing information and evidence. It also allows for the transfer of proceedings and controlled deliveries of shipments of illicit drugs. The United Nations adopted the Convention against transnational organized crime on 15 November 2000. This convention covers only transnational serious crime, which it defines as "conduct constituting an offence punishable by a maximum deprivation of liberty of at least 4 years or a more serious penalty."

Another convention to which a substantial number of countries are party is the Commonwealth Scheme for Mutual Assistance in Criminal Matters of 1990 (Harare Scheme). The Commonwealth of Nations now includes 53 Member States, with 1.9 billion inhabitants. It provides for assistance similar to the UN Conventions described above but also includes provisions for facilitating the personal appearance of witnesses; temporarily transferring persons in custody to appear as a witness; tracing, seizing, and confiscating the proceeds or instrumentalities of crime; and preserving computer data such as traffic data.

In Europe, the European Convention on Mutual Assistance in Criminal Matters of 1959 was the first important multilateral treaty concerning police and judicial cooperation in Europe. Drawn up by the Council of Europe, it provides for the exchange of formal requests for mutual legal assistance (letters rogatory). Such requests describe the offense or offenses under investigation and, if known, the identities of the suspects involved as well as the assistance required. The requested party can only comply if the offense is also punishable under its own domestic penal law. The Convention includes specific provisions for interrogating suspects, interviewing witnesses, searching premises, and exchanging information from criminal and judicial records. A party may also request another to institute proceedings in its own territory. The 1959 Convention has entered into force in all of the 47 member states of the Council of Europe, and in three nonmember states (Chile, Israel, and Korea). It also served as a template for the UN Conventions and the Harare Scheme mentioned above, as well as for many bilateral treaties.

The Maastricht Treaty established the European Union in 1992 and gave the EU competence in the field of justice and home affairs. It led to many new provisions for operational law enforcement cooperation. In 1998, the Treaty of Amsterdam incorporated the Schengen Implementation Convention into the legal framework of the EU. 'Schengen,' originally the initiative of Belgium, France, Germany, Luxemburg, and the Netherlands, introduced the free movement of persons, goods, capital, and services between these countries. However, 'open borders' also called for measures to enhance police cooperation, to prevent criminals from taking advantage. Thus, the Schengen Implementation Convention provides for the direct exchange of information between the police forces and for cross-border pursuit, cross-border surveillance, and controlled delivery of illicit drugs.

In May 2000, EU Member States signed the Convention on Mutual Assistance in Criminal Matters between the Member States of the European Union. It includes an extensive framework for intercepting telephone conversations and contains provisions with regard to cross-border covert operations. Moreover, it formally introduced joint investigation teams as a means of enhancing operational police cooperation. The convention also extended existing provisions, for instance by enabling telephone or videoconference interviewing of witnesses and experts.

More recent instruments are the 'Framework Decision on simplifying the exchange of information and intelligence between law enforcement authorities of the Member States' (2006) and the 'Framework Decision on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime' (2008). The first expands and simplifies information exchange between the EU member states. The second introduces, among other things, online access to databases of DNA profiles and fingerprints on a hit/no-hit basis. In other words, the police can check online if a match is available abroad. If this is the case, the competent authority can send a request for mutual legal assistance to ask

for further details. It also contains articles specifically concerning cross-border public-order policing.

Apart from engaging in multilateral conventions, countries can also agree to bilateral arrangements on mutual legal assistance. The United States, for example, has concluded such treaties with over 40 countries. Finally, if no treaties or applicable provisions exist, the national authorities of the countries involved can decide on providing mutual assistance in individual cases subject to specific agreed terms and conditions.

The Organizational Framework

Effective international law enforcement cooperation requires an organizational infrastructure for handling requests for mutual legal assistance (Spapens, 2010b). This section briefly discusses the most important multilateral arrangements at the international and EU levels. It includes Interpol, Europol, police-liaison networks, and, finally, the European Judicial Network and Eurojust.

Interpol, now with 188 Member States worldwide, is the best-known global institution for police cooperation. One of its main tasks is to maintain a network of National Central Bureaus that can exchange information on a 24/7 basis. In addition, Interpol provides data services, maintains databases of police information, and offers operational support to national and international investigations through its specialized crime units and 24/7 command and coordination center. Interpol also provides training and support to the police with regard to fighting cross-border crime. However, contrary to what the public may think, Interpol itself has no operational investigative powers. Its role in the Second World War resulted in its powers being severely restricted with respect to ideological crimes. Article 3 of the Interpol Constitution specifically forbids the organization 'to undertake any intervention or activities of a political, military, religious or racial character.' However, after the attacks on the World Trade Center in 2001, Interpol did become involved in the fight against terrorism.

Networks of police-liaison officers also have an important role to play in law enforcement cooperation. Their primary task is to assist in information sharing for the purpose of combating crime through both prevention and law enforcement and by responding to requests for mutual police and judicial assistance in criminal matters. Liaison officers have no executive powers and they cannot investigate crimes independently on foreign territory. The exchange of police-liaison officers is usually the result of a bilateral agreement and often mirrors lines of cross-border crime. The Netherlands, for instance, already had a liaison officer in Thailand in the 1970s because, at that time, Southeast Asia was a major source of heroin. Large-scale trafficking in ecstasy from the Netherlands to Australia after the turn of the millennium led to the posting of a Dutch police liaison in Australia and vice-versa.

Finally, the European Union established three institutions to support law enforcement cooperation in the fields of serious crime and terrorism in particular. First, Europol was founded in 1999, although the idea for creating a European criminal investigation department had already been raised in the 1970s (Oberleitner, 1998). Although originally envisioned as a Federal Bureau of Investigation for the EU, Europol still does not have executive powers and remains a 'clearing house for information.' Europol's office is in The Hague in the Netherlands. Europol can also handle information requests with regard to terrorism.

In the field of judicial cooperation, the European Judicial Network (EJN) began its work in 1998. This network consists of the national judicial authorities of the EU member states, each of which has assigned one experienced public prosecutor to the EJN. The primary aim of the network is to support the exchange of requests for mutual legal assistance. For instance, the EJN can assist public prosecutors who want to contact an authority in another Member State and assist in formulating requests for mutual legal assistance.

The final important international institution is Eurojust, also with its offices in The Hague. This organization was founded in 2002 in order to facilitate law enforcement cooperation between the EU member states, especially in the field of organized crime. If necessary, Eurojust can host meetings between criminal investigators from two or more Member States working on cross-border cases and furnish facilities for simultaneous interpretation.

Conclusion

This article has shown that criminals have always tried to take advantage of borders between jurisdictions and States, to avoid apprehension and punishment, to find targets for predatory crimes, and to exploit market opportunities for illicit goods and services arising from differences in supply and demand as well as differences in laws and regulations. From the 1960s onward, the growing demand for illicit drugs promoted the development of a 'glocal' network of producers, traffickers, and dealers. Particularly since the late 1980s, the increasing level of mobility, the fall of the Iron Curtain, and the popularization of the Internet have presented new opportunities for cross-border crime. Specific features of the 'underworld,' however, seem to prevent the emergence of multinational criminal organizations. Notwithstanding these problems, criminal networks appear to be more effective in operating across borders than their counterparts in law enforcement. Despite the development of an impressive legal and organizational framework to aid cross-border investigation, sovereignty remains the basic principle. Cooperation through requests for mutual legal assistance is notoriously slow and inefficient. Mechanisms that could be more effective, such as joint investigation, are still in their infancy, however, even in the EU.

See also: Crime, Organized; Crime Trends and Debates; Cyber Crime: Identity Theft; Drug Control: International Aspects; Economics and Crime; Immigration and Crime in the United States; Mafias, Criminology of; Policing; Property Crime: Victimization and Trends.

Bibliography

Bach Jensen, R., 1981. The international anti-anarchist conference of 1898 and the origins of Interpol. Journal of Contemporary History 16 (2), 323–347.

Bresler, F., 1992. Interpol. Sinclair-Stevenson, London.

Chesney, K., 1979. The Victorian Underworld. Pelican Books, Harmondsworth.
Doezema, J., 1999. Loose women or lost women? Gender Issues 18 (1), 23–50.
Egmond, F., 2004. Multiple underworlds in the dutch republic of the seventeenth and eighteenth centuries. In: Fijnaut, C., Paoli, L. (Eds.), Organised Crime in Europe. Concepts, Patterns and Control Policies in the European Union and Beyond. Springer, Dordrecht, pp. 77–107.
Farcy, F., 2002. How Is the Albanian Mafia Setting up? Locally, Nationally and in Europe. Departement de recherche sur les menaces criminelles contemporaines, Paris.
Fijnaut, C., 2010. The past, present and future of general police cooperation within the European Union. Journal of Police Studies 2010-3 (16), 19–38.
Godson, R., 2003. Menace to Society: Political Criminal Cooperation around the World. Transaction, New Brunswick.
Grootswagers, K., 1983. De historie van het Loonse kerkdorp Kaatsheuvel. Uitgeverij Brabantia Nostra, Breda.
Henry, G., 1984. Cartouche. Tallandier, Paris.
Hobbs, D., 1998. Going down the glocal: the local context of organized crime. Howard Journal 37 (4), 407–422.
Hobsbawm, E., 1972. Bandits. Penguin Books, Harmondsworth.
Oberleitner, R., 1998. Schengen und Europol. Kriminalitätsbekämpfung in einem Europa der inneren sicherheit. Manzsche Verlags- und Universitätsbuchhandlung, Vienna.
Paoli, L., Spapens, T., Fijnaut, C., 2010. Drug trafficking. In: Brookman, F., Maguire, M., Pierpoint, H., Bennet, T. (Eds.), Handbook on Crime. Willan Publishing, Cullompton, pp. 626–650.
Spapens, T., 2006. Interactie tussen criminaliteit en opsporing. Intersentia, Antwerp/Oxford.
Spapens, T., 2008. Georganiseerde misdaad en strafrechtelijke samenwerking in de Nederlandse grensgebieden. Intersentia, Antwerp/Oxford.
Spapens, T., 2010a. Cross-border public order policing in the Dutch border areas. Journal of Police Studies 2010-3 (16), 163–179.
Spapens, T., 2010b. Cross-border cooperation in criminal investigations. In: Botterill, D., Jones, T. (Eds.), Tourism and Crime. Key Themes. Goodfellow Publishers Ltd, Oxford, pp. 129–147.
Spapens, T., 2010c. Cross-border Police Cooperation in Tackling Environmental Crime. www.inece.org (accessed 03.03.2013.).
Spapens, T., 2012a. Netwerken op niveau. Criminele micro-, meso- en macro-netwerken. Prisma Print, Tilburg.
Spapens, T., 2012b. Gambling, illegal. In: Beare, M. (Ed.), Encyclopedia of Transnational Crime and Justice. Sage Publications, Thousand Oaks, pp. 156–157.
Spapens, T., Fijnaut, C., 2005. Criminaliteit en rechtshandhaving in de Euregio Maas-Rijn. Intersentia, Antwerp/Oxford.
Strange, S., 1996. The Retreat of the State. The Diffusion of Power in the World Economy. Cambridge University Press, Cambridge.
United Nations, 2001. Report of the Panel of Experts Pursuant to Security Council Resolution 1343 (2001), paragraph 19, concerning Liberia. s/2001/1015.
United Nations, 2004. Challenges and Change, a More Secure World: Our Shared Responsibility. Report of the Secretary-general's High-level Panel on Threats. United Nations, New York.
van Dijk, J., Spapens, T., 2013. Transnational organized crime networks. In: Reichel, P., Albanese, J. (Eds.), Handbook of Transnational Crime and Justice, second ed. Sage Publications, Thousand Oaks, pp. 213–226.
Varese, F., 2011. Mafias on the Move. Princeton University Press, Princeton.
Weidenbaum, M., Hughes, S., 1996. The Bamboo Network. How Expatriate Chinese Entrepreneurs Are Creating a New Economic Superpower in Asia. The Free Press, New York.
Williams, P., 2003. Transnational organized crime and the state. In: Hall, R., Bierstecker, T. (Eds.), The Emergence of Private Authority in Global Governance. Cambridge University Press, Cambridge, pp. 161–182.
Zhang, S., Chin, K., 2001. Chinese human smuggling in the United States of America. Forum on Crime and Society 1, 31–52.

Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

Dale F Eickelman, Dartmouth College, Hanover, NH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Transnational religious identity' refers to personal or collective religious identities that transcend enclaves, localities, regions, nations, and states to achieve a wider unity of belief, practice, and community. The term implies the existence of the nation-state as the principal worldwide form of political sovereignty. Transnational religious identities are continuously reconstructed in differing contexts and ongoing struggles over ideology and authority in which shifting coalitions of elites, counterelites, ethnic groups, and those excluded from local authority structures compete for voice and recognition.

The subject of 'transnational religious identity' is unexpectedly complex. It refers to personal or collective religious identities that transcend enclaves, localities, regions, nations, and states to attain a wider unity of belief, practice, or community. Transnational religious identities also imply an inherent tension between claims to universality and the specifics of language, ethnicity, place of origin, locally shared culture, values, and history. The term also implies the existence of the nation-state system – with its own competing claims for personal and collective identity – as the principal worldwide form of political sovereignty. Transnational religious identities are continuously reconstructed in different contexts and ongoing struggles over ideology and authority in which shifting coalitions of elites, counterelites, ethnic groups, or those excluded from local authority structures compete for voice and recognition.

Constructing Religious Identities

Transnational religious identities, like other personal and collective identities, are socially constructed. They are not given or 'natural,' although the advocates of some religious identities, like secular nationalists, often make such claims. In the nineteenth century, William Robertson Smith's influential *Lectures on the Religion of the Semites* (Smith, 1956) traced the development of Semitic religious practices and institutions, stressing how a people's social and political institutions, and the shifting boundaries between their own and 'other' communities, contributed to shaping their religious and ethical conceptions. Thus, the prophecies of Ezekiel and Samuel differed not because of a contradiction in the Bible, but because each prophet addressed himself to different groups and historical contexts.

For Smith and his successors, a key element in the growth and development of religions was the growing awareness that ideas and practices of community are socially constructed even if they are firmly rooted in divine revelation. In this view, a more 'modern' understanding of religious belief and practice implies a self-consciousness of ritual and tradition, places a high value on individual responsibility and commitment, and recognizes the existence of other religions and accords them some measure of legitimacy. In Christianity, Islam, and Judaism, this individual 'taking charge' of faith and practice also led to the greater role of a nonclerical 'public' in determining the role of religion in society (Salvatore, 2007). In this respect, Quranic verses such as the following, revealed in seventh century Arabia, are distinctly modern: "To each among you, we have ordained a law and assigned a path for each of you. Had God pleased, he could have made you one nation, but His will is to test you by what He has given you; so compete in goodness" (Quran 5: 48). From its beginnings in seventh-century Arabia, Islam as religious tradition and practice has been remarkably open to the outside world (Ernst, 2011). The vast expanse of the Muslim majority world, ranging from Morocco to Indonesia, inevitably meant that it came to encompass a variety of civilizations and cultural forms. In Arabic, the Quran exhorted its initial seventh-century audience to break through parochial identities of tribe and kinship and aspire to a wider, more inclusive community of belief and practice.

Other religious traditions, including Judaism and Christianity, show similar diversity of form and interpretation, commingling local and universal elements. Pilgrimages in all faiths serve to project believers across lines of immediate locality or group. Pilgrims – and visitors to shrines and holy places – set off from home in the hope of reaffirming contact with a spiritual center, encountering 'others' and returning home with a sharpened awareness of difference and similarity (Eickelman and Piscatori, 2003; Sumption, 1975). Whether to Mecca, Jerusalem, Ayodhya, Rome, the holy places of Shi'a Muslims, or to the shrines of Sephardic saints (*zaddiqim*) in North Africa and Israel, pilgrimages and visitations create unity as they highlight differences of locality. In the Mediterranean (for Christians, Muslims, and Jews), as for South Asia (for Hindus and Muslims), pilgrimages and visitations often involve shared sacred spaces (Albera and Couroucli, 2012; Pemberton and Nijhawan, 2008).

Catholic missionaries and the adepts of Sufi religious brotherhoods also carried the word and practice of their respective faiths across vast spaces before they became nation-states (Rudolph and Piscatori, 1997). The missionary activities of the Catholic Church in Latin America date from the Spanish conquests of the sixteenth century. Missionary orders such as the Jesuits and Franciscans implanted complex networks of churches, missions, and schools, transferring people and

resources between Spain and the colonies. In the Middle East in the nineteenth century, French Catholic and American Protestant missionaries opened religious schools in the Levant, and from the 1880s onward, the Alliance Israélite Universelle sent emissaries throughout North Africa and the Ottoman Empire. Although not primarily religious in orientation, Alliance missionaries had a profound impact on the self-perception of the Jewish communities where they worked. The work of these missionary enterprises was aimed, respectively, at Christians and Jews. These activities occurred within a framework of economic hegemony and political domination in which foreign powers actively sought to protect the interests of Middle Eastern Jews and Christians. However, these 'minority' communities were widely distributed geographically throughout the Middle East in the nineteenth century, so that missionary activities aimed at them had a significant indirect impact on the region's Muslim majority. For example, the founder of the Muslim Brotherhood, Hasan al-Banna (1906–49), a strong opponent of Christian missionaries in Egypt and elsewhere, may have been strongly influenced by what he saw of their activities and organization when he was posted as a primary school teacher to Isma'iliya, a town in the Suez Canal Zone, in 1927. Such 'learning' from other religious traditions is an important element in transnational religious movements and not confined to the last two centuries (Höfert and Salvatore, 2004).

Sufi orders had a major role in propagating Islam, both in the Islamization of North Africa and in the subsequent consolidation of Islamic domination. Some orders, such as the Bektashiya, attracted as adepts the political elite of the Ottoman Empire. Others, such as the Tijaniyya, had a major role in the spread of Islam to West Africa. Some orders were highly structured, but most were only loosely linked through informal but pervasive networks of affinity and trust. Their forms of association were widely understood and respected throughout the Muslim world, and their organizational forms continue to serve as a template for many contemporary religious movements.

Religious Identity and Nationalism in the Nineteenth Century

In nineteenth century Europe, the idea took hold that peoples could be divided into nations sharing a common language, religion, ethnicity, and culture, and therefore should form separate political units. This notion was complemented by a heightened self-consciousness, systematization, and explicitness of religious traditions. By the 1830s, the plural 'religions' came into common currency, with a similar development in Arabic, to signify the idea of 'religion' as a system of beliefs and practices that could be set alongside nonreligious elements of the human condition and apart from other religions (Smith, 1963). This idea added impetus to the erosion of clerical authority over public discussions about religion in all faiths, an idea that often crossed from one faith to another (Höfert and Salvatore, 2004), as in the heated discussions between Ernest Renan and Jamal al-Din al-Afghani.

The linkage between nationalism and religion was implicit and the boundaries of religious and national identities rarely coincided. Partially in response to political and ideological developments in nineteenth-century Europe, a pan-Islamic movement developed in the late nineteenth century. Although the pan-Islamic movements of the 1880s bore similarities to the pan-Hellenic and pan-German movements of the time, only pan-Islam was based formally on a religious commonality (Landau, 1994). The pan-Islamic movements were initially confined to the educated elite, although this elite claimed to speak for all Muslims. The combined French and Spanish effort to crush the rebellion of Abd al-Karim al-Khattabi in 1927 generated protests throughout the Muslim world and appeals for pan-Islamic unity, but the results were transient and minimal. Equally profound is the accelerated nonelite transformation of world religions as they incrementally are influenced by the diverse national contexts in which they develop. For example, some observers write of the birth of a 'French Islam' (Kepel, 2011, p. 11).

The Reemergence of Religion in Public Life

Prevailing theories of modernity and modernization in the mid-twentieth century assumed that religious movements, identities, and practice had become increasingly marginal to modern or modernizing societies and that only those religious intellectuals and leaders who attached themselves to the nation-state would continue to play a significant role in public life. As Casanova (1994) notes, by the late 1970s this prevalent view rapidly eroded through the impact of several nearly simultaneous developments: the Iranian revolution, the rise of the Solidarity movement in Poland, the role of liberation theology in political movements throughout Latin America, and the return of Protestant fundamentalism as a force in American politics. Religion continues to play a significant role in politics and society worldwide. Even locally rooted movements benefit significantly with transnational sponsors or supporters, whether emigrants, states, or others sharing a group's objectives. Opposition to the Shah's regime was firmly rooted in Iran, but the safe haven provided for opposition leaders in exile and their transnational ties and freedom of movement and action were integral to the revolution's success, just as such transnational ties contribute strongly to reshaping how Iranians think of Islam and politics in the era of the post-1979 Islamic Republic (Adelkhah, 2012).

Contemporary religious transnationalism and the identities it supports take many forms. Some are highly structured, while others are only informally organized and related primarily by affinity of goals. Informality and the lack of formal structure can be a strength in contexts of political and religious oppression.

At one end of the spectrum of organizational structure is the vertical transnationalism of the Roman Catholic Church. In its encounter with the emerging system of sovereign nation-states in Europe, the Church incrementally reached an accommodation with nation-states so that the Church hierarchy, although transcending the boundaries of nation-states, also worked within them. The close ties formed with sovereign states often preserved a large measure of local freedom of action for the Church, but they also checked its ability to serve as a voice for social and political reform and for human rights.

The Second Vatican Council (1962–65) significantly altered the nature of Catholic transnationalism. With the Vatican's Declaration of Religious Freedom, *Dignitatis Humanae*, it unambiguously supported the doctrine of universal human rights and used its voice internationally to secure these rights for all peoples and not just Catholics. Vatican II also intensified interfaith dialogue and formally recognized the diversity of Catholic belief and practice. It spurred the translation of liturgies into local languages and promoted lay participation in Church activities, further adapting the Church to local political and social circumstances.

The Church's overall hierarchical structure remained intact after Vatican II, but it was complemented and sometimes overshadowed in importance by decentralized grassroots networks and organizations created for purposes as diverse as Bible study, evangelism, and social and political actions. Many organizations and networks had direct ties to the Church hierarchy, while others operated autonomously. In Latin America, as elsewhere, both types of networks benefited from transnational support. Dutch, Italian, French, Swiss, the United States, Spanish, and other organizations acting transnationally contribute funds and personnel to sustain local developmental programs and action networks. Church-related human rights networks (along with groups with no religious affiliation, such as Amnesty International and Human Rights Watch) have played a significant role in sustaining resistance to authoritarianism and repression, giving concrete form to the idea and practice of transnational civil society.

Protestant transnationalism lacks the time depth of the Catholic presence in Latin America, Africa, and elsewhere, but missionaries from mainline Protestant denominations established a stronghold in Africa by the last part of the nineteenth century, often informally dividing spheres of influence with competing Catholic and Protestant groups, establishing schools and hospitals that often had no state counterparts. Schools have been especially important, and often formed the cadres of early nationalist movements (Beidelman, 1982). Education, intended to inculcate Christian identity and practice, ironically undermined colonial authority and provided the cadres for African independence movements.

The first Protestant churches in Latin America were established primarily for European immigrants. By the 1960s, however, the mainline denominations were increasingly displaced by evangelical, Pentecostal, and fundamentalist churches and missionaries, trained and supported primarily from North America. Levine and Stoll (1997, p. 71) describe the growth of these groups, especially the 'Religious Right,' mixing evangelism with nationalism and anticommunism, as 'spectacular.' They suggest that these groups, in addition to providing needed community services, hospitals, and schools in areas where state facilities are minimal, appeal primarily to people experiencing intense social change and aspiring to improve their status and lives.

The ease of travel and the greatly enhanced role of the mass media have also enhanced the scope and ability of nongovernmental associations and networks to infuse an inherent transnational dimension to contemporary religious identity and practice. Once established, the 'horizontal' networks of civic and religious engagement for both Catholic and Protestant groups became durable and self-sustaining. The transnational ties engendered through travel for labor migration, missionary activity, civic action, and human rights work creates both institutional ties and a sense of moral community for all participants, not just for the elite. Unlike the hierarchical transnationalism of the Church in earlier periods, the loose network organization that characterizes many contemporary groups is more difficult for states to regulate or control. Moreover, the ease with which participants in such networks can spread information across state boundaries and secure external support serves as a brake on the state abuse of authority. This is as much the case for Latin America as elsewhere. Della Cava (1993), for example, notes the influential role of Catholic nongovernmental philanthropic and service organizations in Central and Eastern Europe as a catalyst for change and an alternative, although not a replacement, for attachment to the nation-state.

Transnational Islamic Identities

Islamic transnational movements occur in a significantly different context than Christian ones because Islam lacks a central organizational hierarchy. Although some Muslim intellectuals call for a spiritual or political unity, the Muslim world has almost always been characterized by a *de facto* pluralism of belief and practice. Even radical Islamic groups such as Lebanon's Hizbullah use the universal language of Islam but in practice seek only to establish an Islamic state or an Islamic order within their respective national borders even if they secure external support (Norton, 2009).

Some Islamic transnationalism is formal, organized, and linked to state organizations. After an arsonist's attack on the al-Aqsa mosque in Jerusalem in 1969, Muslim majority states founded the Organization of the Islamic Conference (OIC), which consists of 56 member states. However, this group has never advocated political unity and some member states, such as Nigeria, are anxious not to appear too closely identified with the organization. Nor has the OIC proved effective in the Iran–Iraq war (1980–88), in defusing the Iraqi–Kuwaiti tension that led to the 1990–91 Gulf War, or in other conflicts among Muslim majority states. However, its Islamic Development Bank, established in Saudi Arabia in 1974, has become one of the most effective aid agencies in the developing world, and various charitable funds associated with it have had a significant impact in its regions of activity.

Unlike Catholicism, in which vertical and horizontal transnationalism coexist, Islamic transnationalism is entirely horizontal. In addition to formal organizations such as the OIC, some clandestine or 'deniable' transnationalism is state supported. After the Soviet invasion of Afghanistan in 1979, for example, the United States, with the cooperation of Pakistani and Saudi Arabian security services, supported the recruitment and training of mercenaries from other countries, including Egypt, Algeria, the Yemen, and Sudan to operate alongside Afghan refugees from bases in Pakistan. This 'joint venture' subsequently turned against its initial sponsors. Relationships among like-minded groups are highly fluid, with loosely organized networks of ideological support, financing, and practical training (Sageman, 2008).

Islamic transnationalism is also manifested in nongovernmental movements such as the Muslim World League (*Rabitat al-'alam al Islami*). It sponsors educational programs, religious scholarships for students to study at Saudi and Muslim world universities, and a vigorous publication program throughout the world. Such nonstate organizations, particularly through their formal presentation of Islamic issues and standardization of language and approach, help create common ideological communities that transcend state and national frontiers. Such organizations also show an increasing concern over the situation of Muslim minorities throughout the world, bringing their plight to the wider Islamic community worldwide. The concern of these groups with Muslims living in North America, Europe, and Australia is reciprocated by the emigrants' active concern with the situation of coreligionists in their countries of origin or heritage and with the situation of Muslims in general. Other movements, such as those associated with the Turkish religious leader, Fetullah Gülen, are based on loosely connected clusters of supporters who lack formal central direction. Gülen's movement sponsors schools throughout Turkey, Central Asia, and Eastern Europe, as well as radio, newspaper, and television stations; notwithstanding the lack of central organization, the movement's educational and media programs have emerged as one of the most enduring forms of Turkish presence in Central Asia and Turkish immigrant communities in Europe, North America, and Australia (Turam, 2006).

An even less formal and structured form of transnationalism is represented by *da'wa* (call to Islam) movements such as the Tablighi Jama'at (Masud, 2000). Founded in India in 1927, the movement expanded worldwide after 1947 and currently operates in 80 countries. One principle of the movement is that the practice of Islam does not depend on educated elites alone. The movement's main efforts are directed at deepening the faith of other Muslims and, to promote unity, it avoids theological controversies and politics. South Asian in origin, the Tablighi Jama'at is also active in Europe, North America, and the Arab world. The movement offers no formal training institutes, although some institutions of religious learning are attached to Tablighi centers. Unlike other Muslim organizations, the Tablighi Jama'at lacks welfare or educational programs. Since the movement eschews the broadcast and print media as a matter of doctrine, its spread is entirely through personal contact. With its emphasis on strict adherence to rituals and a concern with the development of individual morality and renewal of personal faith rather than collective action and community change, the Tablighi movement has limited appeal to educated Muslims.

Religious Identity and Globalization

Since the last third of the twentieth century, modern forms of travel, communication, and mass education have accelerated religious transnationalism – the flow of ideologies, access to information on organizational forms and tactics, and the transformation of formerly elite movements to mass ones – rendering obsolete earlier notions of frontier as defined primarily by geographical boundaries and accelerating in pace the earlier spread of ideas already speeded through the introduction of print technologies. Thus, the devotees of the North African Jewish saint, Baba Sali (Rabbi Israel Abouhatzeira) meet not only in Israel and Morocco, but also in New York and Europe.

Especially in the Muslim majority world, the advent of mass higher education since the mid-twentieth century, combined with the proliferation of mass media and the greater ease of communication and travel, has led to a significant reimagining of religious and political identities. One change is that patterns of authority have become more diffuse and open. Religiously committed activists increasingly replace traditionally educated scholars as sources of authority. Second, religion has implicitly been systematized and objectified in the popular imagination, making it self-contained and facilitating innovation. Questions such as 'What is my religion?', 'Why is it important to my life?', and 'How do my beliefs guide my conduct?' have become foregrounded in the lives of large numbers of believers, facilitating individual questioning and encouraging individuals to assume responsibility for their conduct and to participate in movements of social and political action. These transformations also mean that 'authentic' religious tradition and identity are foregrounded, questioned, and constructed rather than taken for granted. Finally, the greater access to education, like access to the mass media, has broadened senses of language and community. Just as the rise of written vernacular languages in Europe since the sixteenth century and the spread of print technology created language communities wider than those of face-to-face interactions, mass education and mass communications play a role in the contemporary world, transcending the narrow bases of local communities and dialects and creating new, if fluctuating, bases of religious identity.

See also: Buddhism; Christianity, Origins of; Citizenship, Historical Development of; Judaism; Religion: Nationalism and Identity; Transnational Citizenship.

Bibliography

Adelkhah, F., 2012. Les mille et une frontières de l'Iran: Quand les voyages forment la nation. Karthala, Paris.

Albera, D., Couroucli, M. (Eds.), 2012. Sharing Sacred Spaces in the Mediterranean: Christians, Muslims, and Jews at Shrines and Sanctuaries. Indiana University Press, Bloomington, IN.

Beidelman, T.O., 1982. Colonial Evangelism. Indiana University Press, Bloomington, IN.

Casanova, J., 1994. Public Religions in the Modern World. University of Chicago Press, Chicago, IL.

Della Cava, R., 1993. 'Financing the faith': the case of Roman Catholicism. Journal of Church and State 35, 37–59.

Eickelman, D.F., Piscatori, J., 1990. Social theory in the study of Muslim societies. In: Eickelman, D.F., Piscatori, J. (Eds.), Muslim Travellers: Pilgrimage, Migration, and the Religious Imagination. Routledge, London, pp. 3–25.

Eickelman, D.F., Piscatori, J., 2003. Muslim Politics, new ed. Princeton University Press, Princeton, NJ.

Ernst, C.W., 2011. How to Read the Qur'an. University of North Carolina Press, Chapel Hill, NC.

Höfert, A., Salvatore, A., 2004. Between Europe and Islam: Shaping Modernity in a Transcultural Space, second ed. P.I.E.-Peter Lang, Brussels, Belgium.

Kepel, G., 2011. Banlieue de la république. Gallimard, Paris.

Landau, J.M., 1994. The Politics of Pan-Islam: Ideology and Organization, second ed. Clarendon Press, Oxford, UK.

Levine, D.H., Stoll, D., 1997. Bridging the gap between empowerment and power in Latin America. In: Rudolph, S.H., Piscatori, J. (Eds.), Transnational Religion and Fading States. Westview Press, Boulder, CO, pp. 63–103.
Masud, M.K. (Ed.), 2000. Travellers in Faith: Studies of the Tablighi Jama'at as a Transnational Islamic Movement for Faith Renewal. Brill, Leiden, The Netherlands.
Norton, A.R., 2009. Hizbollah: A Short Introduction, with New Afterword. Princeton University Press, Princeton, NJ.
Pemberton, K., Nijhawan, M. (Eds.), 2008. Shared Idioms, Sacred Symbols, and the Articulation of Identities in South Asia. Routledge, London, UK.
Rudolph, S.H., Piscatori, J. (Eds.), 1997. Transnational Religion and Fading States. Westview Press, Boulder, CO.
Sageman, M., 2008. Leaderless Jihad: Terrorist Networks in the Twenty-First Century. University of Pennsylvania Press, Philadelphia, PA.
Salvatore, A., 2007. The Public Sphere: Liberal Modernity, Catholicism, Islam. Palgrave Macmillan, New York, NY.
Smith, W.C., 1963. The Meaning and End of Religion: A New Approach to the Religious Traditions of Mankind. Macmillan, New York, NY.
Smith, W.R., 1956. The Religion of the Semites. Meridian Books, New York, NY.
Sumption, J., 1975. Pilgrimage: An Image of Mediaeval Religion. Faber & Faber, London, UK.
Turam, B., 2006. Between Islam and the State: The Politics of Engagement. Stanford University Press, Stanford, CA.

Transportation Geography

Genevieve Giuliano, Sandip Chakrabarti, and Mohja Rhoads, Sol Price School of Public Policy, University of Southern California, Los Angeles, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Transportation geography describes, explains, and predicts flows of people, goods, and information across space. Flows occur as a result of some form of exchange: people offer their services to employers in exchange for wages, friends exchange information about spouses and children, manufacturers produce and sell consumer goods to wholesalers and retailers. Flows occur within a network: the journey to work in vehicles on roads or rail, information via telephone lines or the Internet, and goods across roads, rail, shipping lanes, or air route. Other subfields of geography are concerned with places and their spatial organization. Places and flows are interdependent; hence the boundaries of transportation geography are quite fluid. This article describes the fundamental concepts of transportation geography, discusses major theories, and highlights the current state of research.

Understanding Flows

Transportation is critical to contemporary society. In the United States (US), the transportation sector represents around 9% of the gross domestic product (GDP), and transportation is the second largest expenditure category for households. Interestingly, its share of GDP is declining as transport industries become more efficient. Economic growth is closely linked with transportation: movement of people and goods is positively associated with GDP. Why is transportation so critical, and what motivates these flows? Transport is costly, both in terms of time and money. Hence there must be some benefit to justify incurring these costs. Flows occur when the benefits derived from transport exceed transport costs. Flows can be understood in terms of a few basic principles.

Flow Depends on Transport Costs and Gains from Trade

The first principle of flows can be illustrated with the journey to work. Consider a worker: the cost of commuting is incurred in order to remain employed. How much is a worker willing to pay in order to receive a given wage rate? The average US commute is estimated at around 28 min (Santos et al., 2011); and just 7.5% are longer than 1 h (McKenzie and Rapino, 2011). Since commute costs reduce the effective wage (wages net of commute costs), it follows that as commute costs increase, fewer people will be willing to accept employment at a given wage level. The converse also holds: as commute costs decrease, more workers will be willing to accept employment at a given wage level. The likelihood of flow taking place is a function of transport costs and benefits from trade. This principle applies to goods and information as well as people.

Declining Transport Costs Are Associated with Increased Flows

Transport cost includes both time and money costs. Technological change and infrastructure investment have increased travel speeds and reduced unit costs. The late nineteenth century transcontinental railroad offered trips from New York to San Francisco in 6 days; the trip by air today takes about 6 h. Similarly, France's high-speed train offers service from Paris to Lyon in 2 h. Around 1900 the same trip would have taken 6 or 7 days. These improvements also apply to intrametropolitan travel, but to a lesser extent, since intrametropolitan travel has so far been limited to surface transport modes and involves shorter distances. Transport cost has also greatly declined. For example, it is estimated that the average cost of commuting by streetcar in the US at the turn of the twentieth century was about 20% of the average daily wage (Giuliano, 1989). Today the average cost for automobiles commuters is about 4% (Roberto, 2008).

Increases in speed result in 'time/space convergence,' a reduction in friction of both distance and time. Time/space convergence is illustrated in Figures 1(a) and 1(b) for all of Europe, excluding Russia and the Ukraine (Spiekermann and Wegener, 1994). **Figure 1(a)** provides actual spatial relationships and is calibrated on a homogeneous travel speed of 60 km/h **Figure 1(b)** provides a picture of temporal relationships, based on the 1993 rail transport network. Note that Paris is 'closer' to major cities of Switzerland, The Netherlands, and Amsterdam, while in East Europe, Bucharest and Sofia are both 'farther' from one another and from Central Europe. The 1993 rail network does not account for the high-speed rail links that have been built in Europe during the past two decades. These links have further changed time/space relationships.

Time/space convergence is associated with globalization and the emergence of the time-sensitive economy. Dicken (2011) argues that improvements in shipping (containerization, air transport, and information and communication technology (ICT)), together with trade and investment liberalization policies, explain the rapid globalization observed over the past several decades.

Reductions in transport costs, in absolute terms and relative to per capita income, along with increased demand due to population growth have resulted in large increases in passenger and freight travel. **Figure 2** shows the increase in the US passenger-miles (private vehicle and domestic air service) between 1960 and 2005, and US domestic surface freight transport ton-miles between 1980 and 2004. **Figure 3** shows the increase in total inland passenger-miles (all modes) and total inland freight ton-miles (all modes) for five select major industrialized nations (OECD members) over the 1980–2005 period.

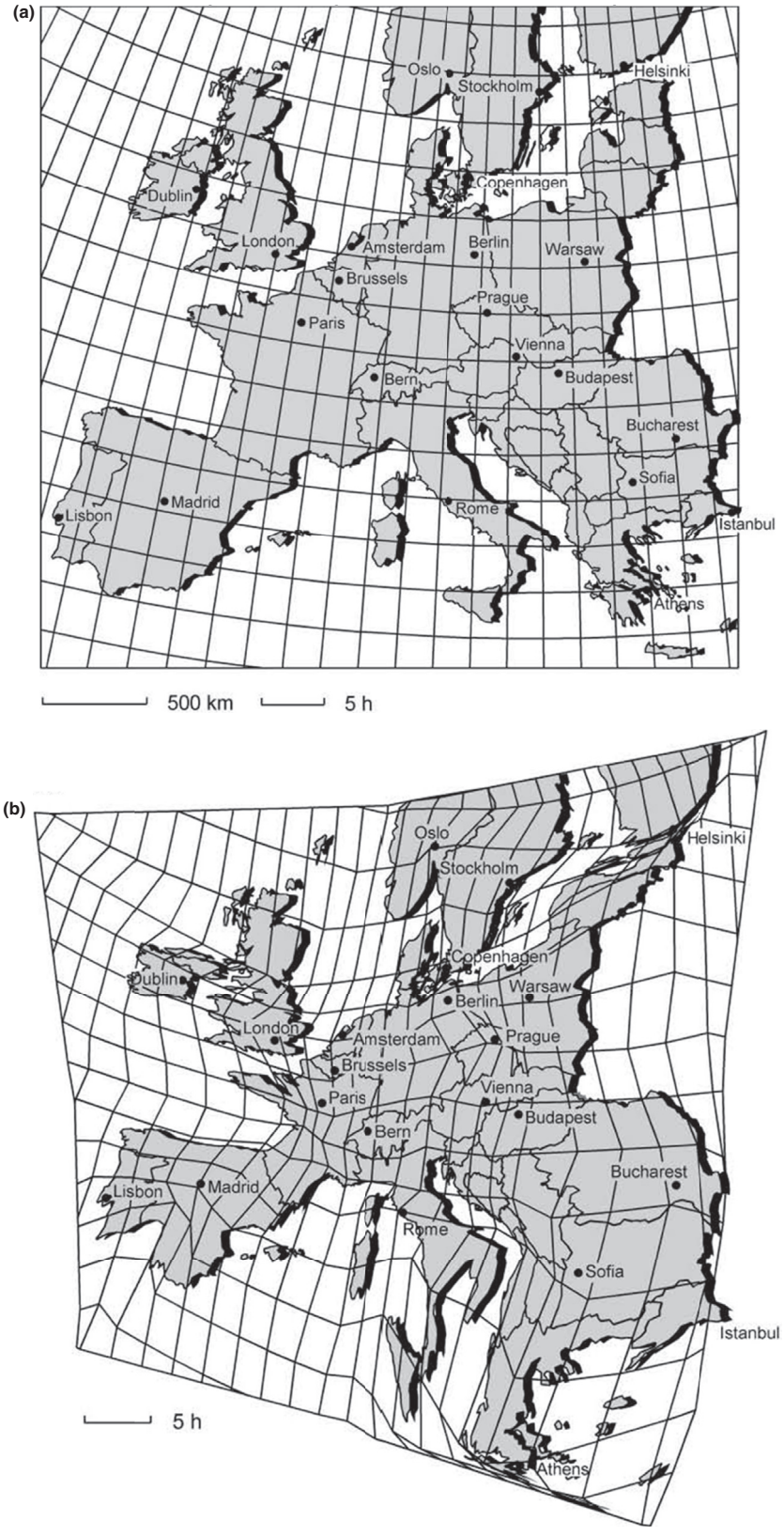

Figure 1 Time/space convergence. (a) Panel A: Map of Europe, constant travel speed. (b) Panel B: Map of Europe, 1993 rail transport network.

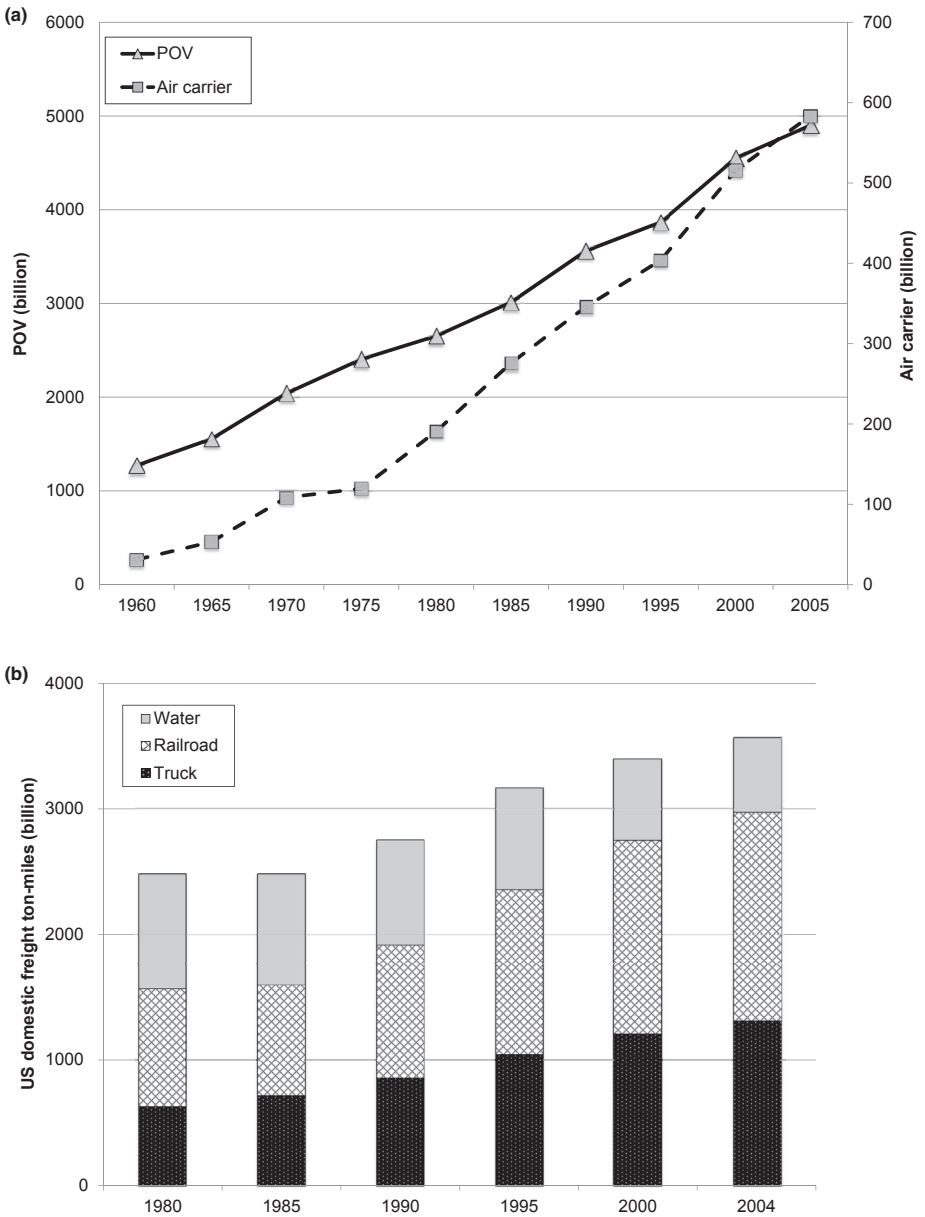

Figure 2 US transportation trends. (a) Panel A: Passenger-miles of travel for privately occupied vehicle (POV) and air carrier. (Data from US DOT Bureau of Transportation Statistics (http://www.rita.dot.gov/bts/) Note: 1 mile = 1.61 km.) (b) Panel B: Domestic freight ton-miles. (Data from US DOT Bureau of Transportation Statistics (http://www.rita.dot.gov/bts/) Note: 1 mile = 1.61 km and 1 Ton = 1000 kg.)

Declining Transport Costs and Increased Flows Are Associated with Development of the Space Economy

Historical development of the space economy is explained in part by declining transport costs. In the preindustrial era, nonmotorized transport was slow and difficult. There was little exchange between cities, because transport costs exceeded benefits at very short distances. Technological changes that reduce transport cost, e.g., 'space-adjusting technologies,' allow regions to exploit comparative advantage, engage in regional specialization, and expand their areas of economic influence (hinterlands).

Nonmotorized modes of the preindustrial era were replaced with motorized modes in the nineteenth century: steamships and railroads. These technological advances greatly reduced the cost of intercity travel and hence led to large increases in both passenger and goods transport. The railroads in the US connected vast regions, resulting in the emergence of an integrated national economy. Railroads also contributed to the development of the national economies in Europe. Change accelerated in the twentieth century with the emergence of paved roads, automobiles and trucks, and the development of commercial air transport.

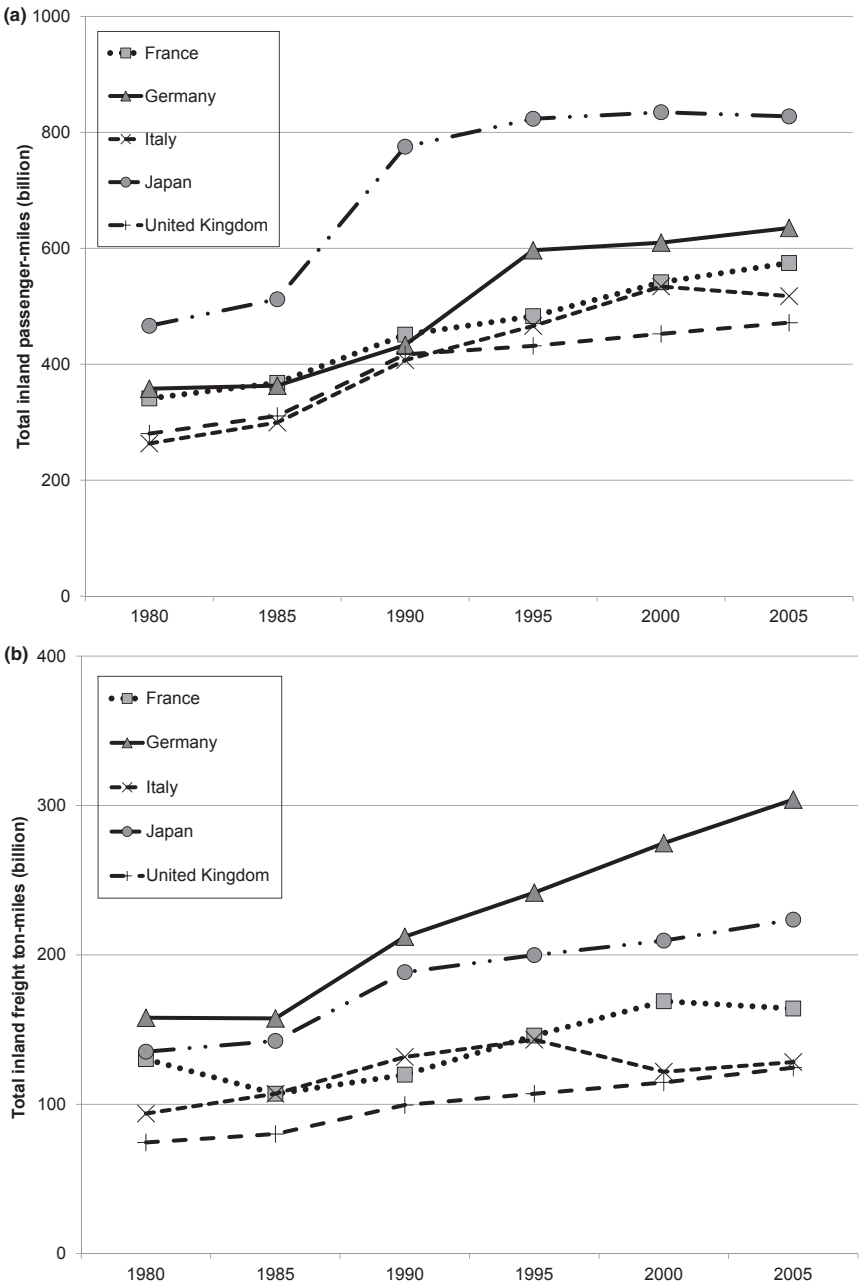

Figure 3 Transportation trends in select major global economies. (a) Panel A: Total inland passenger-miles of travel (all modes). (Data from OECD (http://stats.oecd.org/) Note: 1 mile = 1.61 km.) (b) Panel B: Inland freight ton-miles (all modes). (Data from OECD (http://stats.oecd.org/) Note: 1 mile = 1.61 km and 1 Ton = 1000 kg.)

A famous example of the impact of reduced transport costs is the Erie Canal, built in 1825 to open a waterway between Lake Erie and New York by connecting Lake Erie to the Hudson River. The reduction in transport cost was dramatic: over-the-road transport costs 20 times as much per ton-mile as shipping via canal; travel time was 20 days over the road and 6 days via canal. Goods movement greatly increased. New York's hinterland was extended through the entire northeast region, and its position as the US center of international trade and the dominant city in the US urban hierarchy was consolidated.

Empirical research on the relationship between transport and economic development is mixed. Research on both developed and less developed countries indicates that transport infrastructure is a necessary but not sufficient condition for economic development (Hoyle et al., 1998). The first links of major networks (rail, highway) have generally had significant impact, while the impact of additional links is more uncertain due to the incremental nature of even large investments when networks are more fully developed. In the case of the US Interstate Highway System, significant impacts of the initial investments are documented, while studies of more recent

investments are mixed (Giuliano and Agarwal, forthcoming). There are also many failed examples of transport infrastructure investments made to promote trade and economic development, as for example, the US efforts to promote economic development in the Appalachian region.

Transport and Urban Spatial Structure Are Interdependent

Urban spatial structure is the outcome of economic, social, political, and cultural forces. Within the complex process of urban evolution, however, transportation has played a major role. The historical relationship between transport and urban form is well established (Muller, 2014). With faster and cheaper transport has come more decentralized and dispersed spatial structure, and as activities have become more spread out, travel has increased. Decreased accessibility has generally been offset by increased travel speeds. But the explosion in travel demand has outpaced capacity expansions; and complexity of travel patterns has made transportation management challenging. The consequence is peak-period congestion on roads and public transport systems. With transport costs therefore increasing in some metropolitan areas, the question is whether a corresponding decrease in travel will take place. One possible indicator is the stability and in some cases decline in private vehicle travel observed in several countries since 2008 (e.g., Boarnet, 2013).

US cities in the latter half of the nineteenth century were crowded and congested. Industrialization had generated a surge of urbanization. The lack of motorized transport forced population and businesses to colocate in close proximity to one another. Electricity-powered rail transport proved to be the breakthrough technology. Starting in the late 1880s, urban railways allowed cities to grow outward as workers took advantage of the new mode to seek less crowded residential locations outside the city core, and gave the city center special dominance as the center of the intracity transport network. Urban rail was followed in the 1920s by the automobile and truck. Development of the highway network proceeded rapidly, increasing travel speeds and allowing ever more rapid decentralization. The post-World War II freeway era continued these trends, with both jobs and population decentralizing, and with the emergence of suburban activity centers. European cities have followed a similar trajectory, but under quite different social, economic, and political conditions, and with – to date – less dramatic spatial change.

The relationship between transportation and urban form has been the subject of much theoretical research. The familiar 'standard model' of urban location and spatial structure as developed by Alonso (1964) and others, serves to illustrate the close relationship between transport costs and urban spatial structure. Consider the simple case of a city where all employment is located at the center, transport cost is a linear function of distance, and households trade-off housing and commuting costs in making their residential location decisions. How will workers locate themselves with respect to the city center? Clearly wages, net of transport costs, will be highest at the city center, where transport costs are near zero; hence workers will prefer location near the center, and locations near the center will be valued higher than locations away from the center. If workers could bid on locations, the result would be a gradient of both land prices and population density that declines with distance from the center. Declining transport cost results in less steep density gradients, as the relative value of living near the city center declines. Decreases in the density gradient have been documented for metropolitan areas in the US and around the world (e.g., Kim, 2007). Declining transport cost (both for goods and passengers) has affected urban spatial structure in complex ways. Employment and population are now concentrated (and growing) in a few metropolitan areas, and simultaneously decentralizing within metropolitan areas (Glaeser and Kohlhase, 2004). Decentralization is characterized by polycentricity and spatial dispersion but impacts on travel are less clear.

Research Issues

Increasing concerns regarding the social and environmental consequences of urban transport, structural changes resulting from development of ICTs, and a changing institutional context in transportation policy decision-making are affecting the research agenda of transport geographers (e.g., Hanson, 2000; Black, 2010). This section discusses three areas of current research: travel behavior; transportation, land use, and environmental costs; and transportation and information technology.

Travel Behavior

Transport geographers have examined personal travel at both aggregate and disaggregate levels of analysis.

Aggregate Models

Various forms of spatial interaction models have been applied in aggregate analysis, most commonly the gravity model. The gravity model incorporates two basic factors that affect the level of flow between places: the population of each place (or some measure of potential for flow), and the distance between them.

The model states that flow from origin i to destination j is a positive function of (1) the potential of the origin zone for generating trips and (2) the relative attractiveness of the destination zone, and is a negative function of the distance between them. Gravity models have been applied to many topics, from intercity transport to local shopping flows.

They have been extended to account for alternative travel destinations (intervening opportunities), differences in nodal characteristics, and differences in travel costs across modes or links. The gravity model has also been widely applied for the estimation and analysis of bilateral trade flows.

Geographic information science has made possible the use of gravity formulations to measure accessibility patterns of different modes or population segments and hence analyze such policy issues as the distributional consequences of urban transport investment. The gravity model plays an integral role in state-of-practice urban transportation planning models: trip distribution (the allocation of trips to specific zonal destinations) is accomplished using a model that accounts for the relative attractiveness of zones as destinations and expresses distance as a generalized cost.

Disaggregate Models

Understanding the travel patterns of individuals is a more complex task. Geographers and economists began to address the issue in the 1970s with the application of utility theory. In this case, travel decisions are discrete choices made on the basis of the utility of each alternative available. The utility of an alternative is a function of its attributes and the attributes of the individual making the choice. Since utility cannot be observed directly, utility-based models estimate the probability that a given alternative will be chosen by observing how people actually behave, or by capturing their stated choice in hypothetical scenarios.

The first applications of discrete choice were made to mode choice problems; later applications included destination choice, route choice, departure time choice, and car ownership. Mode choice models are now routinely used in transportation planning applications. Models have been extended to account for conditional or sequential choices, as for example, mode choice given car ownership. Multilayered models are one approach to developing integrated transportation planning models.

A major theoretical issue in travel behavior analysis is what should be predicted: the trip or the underlying behavior that motivates the trip. Given that travel constitutes a disutility, it is argued that understanding travel requires understanding activity patterns. Termed activity analysis, this approach to travel behavior predicts travel choices as a function of activity choices (Kitamura, 1988).

The roots of the activity analysis approach to travel behavior lie with the Swedish school's time–space geography (Hägerstraand, 1970). In this case the individual has a set of resources (time, money, vehicles) and a set of constraints (schedule, physical and coupling constraints) that define the feasible set of activities they may accomplish over the course of a day. Hagerstrand terms this feasibility area the space–time prism. Activity-based models seek to capture the complex process of choosing a set of daily activities and their location, timing, and sequencing, given feasibility constraints.

Utility-based models assume rational behavior. Travel choices are a function of resources, preferences, and supply characteristics. Some geographers and planners argue that people act as much on attitudes and perceptions as on resources and constraints. Also, travel choices are joint choices (resources and tasks are allocated among household members) which reflect (and affect) underlying attitudes and perceptions. They therefore borrow somewhat from the disciplines of sociology and psychology. Many recent studies have demonstrated the role of personal attitudes and preferences in determining, for example, automobile (Gardner and Abraham, 2010) and public transport (Popuri et al., 2011) mode choice.

Travel as Having Positive Utility

Travel has historically been considered a derived demand. However, some recent research suggests that this may not be the case (Mokhtarian and Salomon, 2001). People may derive pleasure from travel to a far greater extent than assumed. They may conduct activities outside the home in order to travel, or they may travel to more distant destinations because the travel itself is part of the activity experience (take a drive and go out to dinner). Transportation planners have consistently underpredicted the growth in travel demand, even accounting for unanticipated changes in demographics, income levels, fuel prices, and other key factors. Positive utility of travel may explain people's greater-than-predicted propensity to travel.

A related question is whether use of certain modes is associated with greater satisfaction or positive perceptions. Morris and Guerra (2014) found that automobile travelers are happier on average compared to transit riders. Proponents of active transportation (walking, biking, skateboarding, etc.) argue that a shift from motorized to nonmotorized mode of travel can have positive impacts on health and the environment. People who walk and bike have been found to derive relatively greater pleasure out of travel (Morris and Guerra, 2014).

The concept of travel as having positive rather than negative utility has significant policy implications. On the one hand, it helps to explain why travel demand is highly inelastic. On the other, it suggests that efforts to reduce travel for environmental or other reasons will be challenging to accomplish. Planners suspect that those with positive attitudes toward travel (particularly private vehicle travel) may actually be less likely to telecommute or live in dense mixed-use neighborhoods (Ory and Mokhtarian, 2005).

Travel Budgets

Finally, many geographers argue that time budgets determine travel. People make about four (one-way) trips per day and spend an average of about 1.3 h each day traveling. These averages have been observed consistently both over time and across places, ranging from preindustrial agrarian villages to the major developed nations of the world. This consistency has led to the concept of a time budget: people allocate a given amount of time to traveling (Zahavi and Talvitie, 1980). Hence distance traveled is a function of transportation technology and infrastructure. As travel speeds increase (say by construction of a new highway or rail line), people will use travel time savings to travel longer distances. There are two main criticism of the time budget. First, taken to its logical extreme, it conflicts with notions of rationality and utility-maximizing behavior. Second, the empirical observation may be an artifact of central tendency, and therefore may have little value as an explanation for human behavior.

Transportation, Land Use, and Environmental Costs

Transportation provides many benefits to society. As noted earlier, an efficient transportation system is critical to economic growth. People have benefitted greatly from cheap and efficient mobility. The value placed on mobility is revealed in willingness-to-pay: people devote a significant portion of their expenditure to transportation, about 18% in the US (US Bureau of Labor Statistics, http://www.bls.gov/cex/). However, mobility generates an array of environmental costs, notably air and noise pollution, traffic congestion, accidents, damage to the natural environment, and most seriously global warming. The transport sector accounts for 28% of US and 13% of global-green house gas emissions (US Environmental Protection Agency, http://www.epa.gov).

In addition, the benefits of mobility have not been realized by everyone. Those without access to a private vehicle, for example, have suffered a relative decline in accessibility as

activities have dispersed and public transport has become an increasingly poor substitute for the car. Concerns about these external costs are intensified by observation of current trends (Figures 2 and 3) and forecasts of global car ownership. Given current understanding of global climate change impacts, it is evident that policy actions must be taken to move toward a sustainable transport system. However, the question is what constitutes a sustainable transport system. Sustainable transport has many dimensions. Strategies range from improving transport technology to smooth traffic flow to reducing private vehicle travel via regulation, pricing, or land-use policy.

Transport geographers have focused on land use. The relationship between accessibility and travel is well known. Some geographers make a distinction between accessibility and mobility (e.g., Hanson, forthcoming). Accessibility measures potential – the relative location of opportunities – and is a function of spatial form and transportation supply. Mobility measures outcomes – how much a person travels. Mobility is a function of both accessibility and travel resources. It follows that as accessibility increases, activities can be conducted with less travel, all else equal. It is therefore argued that one promising strategy for sustainable transport is to promote higher-density development. Proponents argue that such a strategy preserves accessibility while reducing overall travel: activities can be accomplished closer to home, and shorter distances may allow more nonmotorized travel.

Increasing development density as a travel reduction strategy may have limited effectiveness. Total distance traveled is inversely related to density, but the relationship is of small magnitude: a unit increase in density yields a much smaller decrease in travel. For example, a recent US study estimates that if 75% of all new developments are built in a compact fashion (double existing density), and if residents drive 25% less on average in such developments, vehicle miles of travel, energy use, and carbon emissions would decrease about 8–11% by 2050 (TRB, 2009).

Higher density is also associated with congestion and slower travel speed, which may have a negative effect on energy use/emissions. Therefore, the net positive effect of density is unclear. Some have argued, however, that the proximity effect dominates (Levine et al., 2012), but it is important simultaneously to reduce traffic generated at or attracted to dense areas to counter the congestion effect (Melia et al., 2011). The way to reduce traffic is to provide viable alternatives and/or reduce inefficient auto travel by pricing or regulatory mechanisms.

Expansion of public transport infrastructure has become a key policy priority, since transit is seen as less damaging to the natural environment and as contributing to urban quality of life. Effective public transport depends upon concentration. Many argue that the presence of transit will promote higher-density development (e.g., Newman and Kenworthy 1998; in the presence of a favorable policy environment.

Transit and nonmotorized travel are facilitated by appropriate infrastructure. It is argued that sidewalks and bike lanes along with supportive urban designs and integration with public transport can complement existing pro-environment initiatives (e.g., Pucher et al., 2010). Interestingly, the extent to which such interventions can reduce transportation-related energy use and air pollution remains uncertain (Brand et al., 2014).

Long-term economic and technological trends suggest continued decentralization, dispersion, and increasing demand for private vehicle travel as per capita income grows in currently less developed countries (e.g., Muller, 2014). Hence some argue that environmental costs must be addressed via technology (e.g., better fuel efficiency, electric- or solar-powered cars, etc.), pricing, and regulation. Pricing and regulatory policies that directly affect private vehicle owners and users are often considered to be more equitable and economically efficient. Implementation of and reforms in car-ownership taxes, fuel taxes, parking charges, tolls, and congestion or road user charges are most common (Bonsall and Willumsen, 2014). However, such interventions are often politically unpopular. The challenge is to find the most effective method for achieving environmental goals. The social and environmental consequences of not reversing current trends are enormous, and it is assumed that the political will necessary to achieve major change will be found.

Transportation and Information Technology

Rapid advances in ICTs, through affordable mobile devices, connections, and applications, are an increasingly important theme of transport research. ICTs are changing travel patterns, firm locations, and the movement of freight. People can now access multiple activities at the same time (shopping while working and working while shopping) and access activities, goods, and services without travel. Transport geographers have been investigating how ICTs modify or promote travel through concepts of fragmentation, fixity, and flexible workplace practices.

The combination of increased fragmentation (breaking activities into subtasks), reduced fixity (tying activities to specific times and places), and flexible work practices (such as temporary work or self-employment) has enormous implications for travel behavior. It is possible, for example, that trip frequency will increase (fragmentation effect), and peak travel will decrease (flexibility effect). Figure 4 below plots the average frequency of trips taken by time of day for workers who have the option to work at home versus those who do not. On average, workers who have the option to work at home start their work trips later than the rest of the population. They also make more personal trips throughout the day, while those without flexible workplace options tend to cluster personal trips at the end of the workday. Effects on travel distance or mode choice are uncertain. Research on these issues is just beginning.

ICTs and their networking capabilities are also creating a new generation of transportation alternatives. Within the past decade Web-based taxi services, car sharing and ride-matching services, bicycle-sharing systems, and point-to-point intercity bus services have emerged. It is too early to observe how these services may affect travel, but such alternatives may reduce car ownership demand as travel by other means becomes more convenient.

The impacts of ICT on firm location and on the structure of production also have enormous implications for transport. Restructuring is characterized by outsourcing, customized production, flexible production, and just-in-time management of production, and distribution. The emergence of spatial

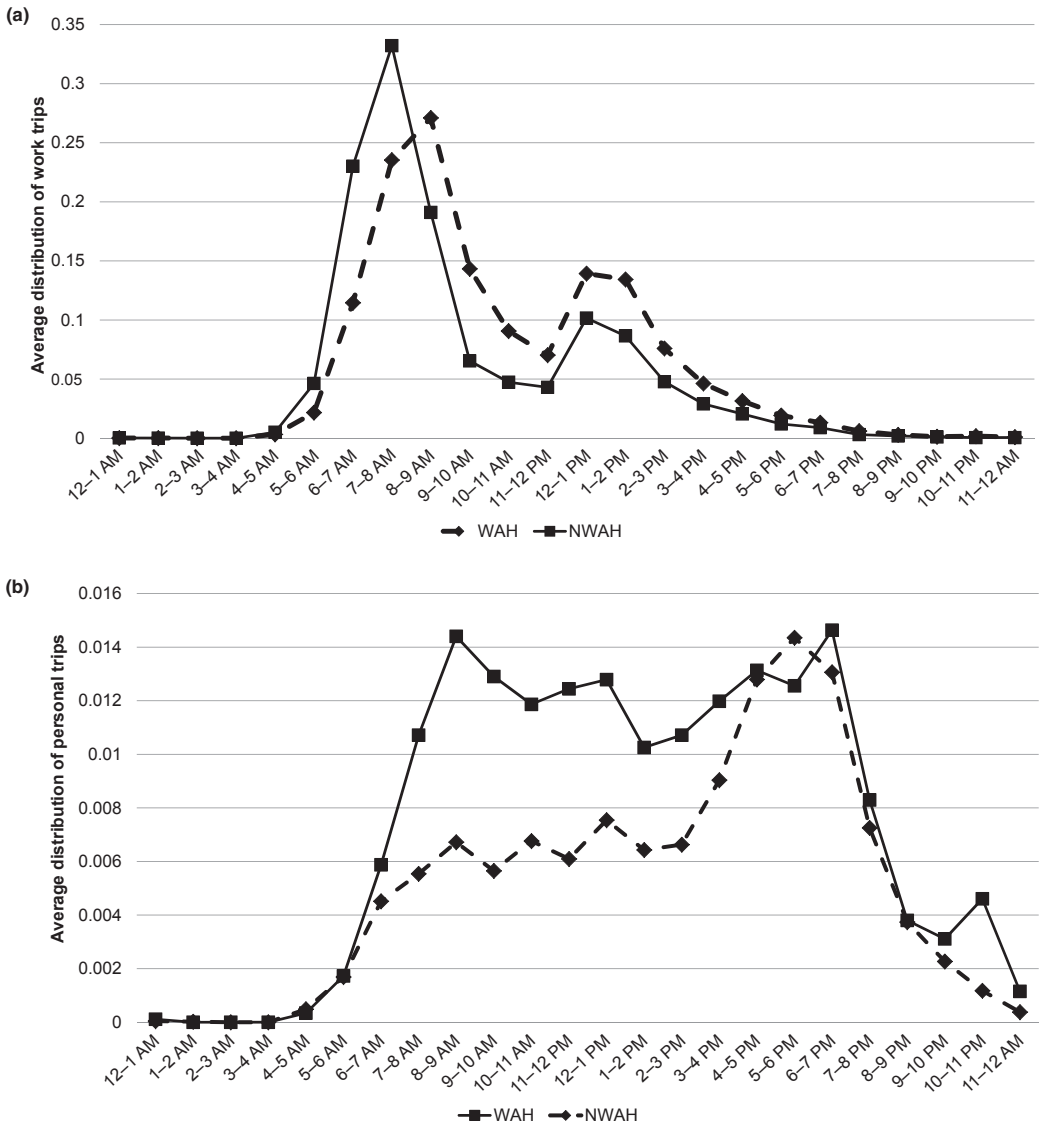

Figure 4 Distribution of trips by time of day and option to work at home. (a) Panel A: Average distribution of work trips individuals who are allowed the option to work at home (WAH) versus individuals who are not allowed the option (NWAH). (National Household Transportation Survey, calculated by Rhoads.) (b) Panel B: Average distribution of personal trips by individuals who are allowed the option to work at home (WAH) versus individuals who are not allowed the option (NWAH). (Source: National Household Transportation Survey, calculated by Rhoads.)

production networks implies increased goods movement and growing dependence on an efficient global transport network.

ICTs have large implications for the freight industry. Globalization has increased the complexity and competition of freight operations, and ICTs provide tools to enhance the effectiveness and efficiency of freight operations; empirical research on adoption and performance is limited, however (Perego et al., 2011).

Time/space compression is bound to increase as the boundaries between transportation and communication erode. Indeed, ICT calls into question the very notion of accessibility as an essentially physical concept (Hodge and Janelle, 2000). The direct and indirect effects of ICT will be a major area of future research in transport geography.

See also: Access: Geographical; Automobility; Economic Geography; Location Theory; Mobility and Travel Activity Patterns; Spatial Analysis in Geography; Spatial Interaction; Time-Space in Geography.

Bibliography

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.

Black, W.R., 2010. Sustainable Transportation: Problems and Solutions. Guilford Press, New York.

Boarnet, M.G., 2013. The declining role of the automobile and the re-emergence of place in urban transportation: the past will be prologue. Regional Science Policy and Practice 5 (2), 237–254.

Bonsall, P., Willumsen, L., 2014. Pricing methods to influence car use. In: Handbook of Sustainable Travel. Springer, Netherlands, pp. 95–111.

Brand, C., Goodman, A., Ogilvie, D., 2014. Evaluating the impacts of new walking and cycling infrastructure on carbon dioxide emissions from motorized travel: a controlled longitudinal study. Applied Energy 128, 284–295.

Dicken, P., 2011. The global economy (chapter 1). In: Hoel, L., Giuliano, G., Meyers, M. (Eds.), Intermodal Freight Transportation: Moving Freight in a Global Economy. Eno Transportation Foundation, Washington, DC.

Gardner, B., Abraham, C., 2010. Going green? Modeling the impact of environmental concerns and perceptions of transportation alternatives on decisions to drive. Journal of Applied Social Psychology 40 (4), 831–849.

Giuliano, G., 1989. New directions for understanding transportation and land use. Environment and Planning A 21, 145–159.

Giuliano, G., Agarwal, A. Land use impacts of transportation investments. In: Giuliano, G., Hanson, S. (Eds.), The Geography of Transportation, fourth ed. Guilford Press, New York, forthcoming.

Glaeser, E.L., Kohlhase, J.E., 2004. Cities, regions and the decline of transport costs. Papers in Regional Science 83, 197–228.

Hägerstraand, T., 1970. What about people in regional science? Papers in Regional Science 24 (1), 7–24.

Hanson, S., 2000. Transportation. In: Sheppard, E., Barnes, T. (Eds.), A Companion to Economic Geography. Blackwell, Oxford.

Hanson, S. The context of urban travel (chapter 1). In: Giuliano, G., Hanson, S. (Eds.), The Geography of Transportation, fourth ed. Guilford Press, New York, in press.

Hodge, D., Janelle, D., 2000. Measuring and Representing Accessibility in an Information Age. Springer-Verlag, Berlin.

Hoyle, B., Leinbach, T., Smith, J., Spencer, A., 1998. The role of transport in the development process. In: Hoyle, B., Knowles, R. (Eds.), Modern Transport Geography, second ed. John Wiley and Sons, Chichester.

Kim, S., 2007. Changes in the nature of urban spatial structure in the United States, 1890–2000*. Journal of Regional Science 47 (2), 273–287.

Kitamura, R., 1988. An evaluation of activity-based travel analysis. Transportation 15, 9–34.

Levine, J., Grengs, J., Shen, Q., Shen, Q., 2012. Does accessibility require density or speed? A comparison of fast versus close in getting where you want to go in US metropolitan regions. Journal of the American Planning Association 78 (2), 157–172.

McKenzie, B., Rapino, M., 2011. Commuting in the United States: 2009, American community survey reports. Available at: http://www.census.gov/prod/2011pubs/acs-15.pdf (accessed 14.07.14.).

Melia, S., Parkhurst, G., Barton, H., 2011. The paradox of intensification. Transport Policy 18 (1), 46–52.

Mokhtarian, P., 1997. The transportation impacts of telecommuting: recent empirical findings. In: Understanding Travel Behavior in an Era of Change. Pergamon Press, Oxford, UK.

Mokhtarian, P.L., Salomon, I., 2001. How derived is the demand for travel? Some conceptual and measurement considerations. Transportation Research Part A: Policy and Practice 35 (8), 695–719.

Morris, E.A., Guerra, E., 2014. Mood and mode: does how we travel affect how we feel? Transportation 1–19.

Muller, P. Transportation and urban form: stages in the spatial evolution of the American metropolis. In: Giuliano, G., Hanson, S. (Eds.), The Geography of Transportation, fourth ed. Guilford Press, New York, in press.

Newman, P., Kenworthy, J., 1998. Sustainability and Cities: Overcoming Automobile Dependence. Island Press, Washington, DC.

Ory, D.T., Mokhtarian, P.L., 2005. When is getting there half the fun? Modeling the liking for travel. Transportation Research Part A: Policy and Practice 39 (2), 97–123.

Perego, A., Perotti, S., Mangiaracina, R., 2011. ICT for logistics and freight transportation: a literature review and agenda. Logistics and Freight Management 41, 457–483.

Popuri, Y., Proussaloglou, K., Ayvalik, C., Koppelman, F., Lee, A., 2011. Importance of traveler attitudes in the choice of public transportation to work: findings from the Regional Transportation Authority Attitudinal Survey. Transportation 38 (4), 643–661.

Pucher, J., Dill, J., Handy, S., 2010. Infrastructure, programs, and policies to increase bicycling: an international review. Preventive Medicine 50, S106–S125.

Roberto, E., 2008. Commuting to Opportunity: The Working Poor and Commuting in the United States. Brookings Institute, Washington DC.

Santos, A., McGuckin, N., Nakamoto, H.Y., Gray, D., Liss, S., 2011. Summary of Travel Trends: 2009 National Household Travel Survey. US Department of Transportation. Available at: http://trid.trb.org/view.aspx?id=1107370 (accessed 14.07.14.).

Spiekermann, K., Wegener, M., 1994. The shrinking continent: new time – space maps of Europe. Environment and Planning B 21, 653–673.

TRB, 2009. Driving and the Built Environment: The Effects of Compact Development on Motorized Travel, Energy Use, and CO_2 Emissions. Transportation Research Board, National Research Council of the National Academies, Washington, DC. Retrieved July 11, 2014, from: http://onlinepubs.trb.org/Onlinepubs/sr/sr298.pdf.

Zahavi, Y., Talvitie, A., 1980. Regularities in travel time and money expenditures. Transportation Research Record 750, 13–19.

Transportation Planning

Qing Shen, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R. Cervero, volume 23, pp. 15873–15878, © 2001, Elsevier Ltd.

Abstract

This article provides an overview of transportation planning – an interdisciplinary field of research and practice that informs and supports decision making for improving the movement of people and goods that is essential for efficient economic production and effective social interaction. It introduces the readers to the basic concepts, frameworks, and analytical tools. Drawing primarily upon experience from the United States, it also briefly describes the field's evolving focus from mobility to accessibility, livability, and sustainability. The article ends with a discussion of the coming challenges and opportunities for transportation planning.

Transportation planning is a field of research and practice that informs and supports decision making – at both strategic and operational levels – for mobility and accessibility improvements. Building national highway and rail networks to stimulate economic growth, developing high-density cities with convenient public transit to reduce automobile dependence, promoting flexible work schedules and telecommuting to relieve peak-hour traffic congestion, and adopting fuel and congestion taxes to lower the vehicle-miles traveled (VMT) are examples of strategic decisions. Siting airports, railway stations, freeway interchanges, and bus lines and terminals; providing additional park-and-ride facilities; extending bike lanes and expanding bike share programs; and setting user fees for high-occupancy toll (HOT) lanes are examples of decisions at the operational level.

The main functions of transportation planning are to establish the linkages between transportation and the goals of a community (such as a city or a region) or an organization (such as a government agency or a private firm); forecast future movements of people, goods, and ideas between geographic locations; evaluate alternative approaches to facilitate these movements; and inform stakeholders of the likely economic, social, and environmental consequences of the choices. The analysis and evaluation often employ empirical data and quantitative methods, and are guided by the fundamental principles of efficiency, equity, and, increasingly, livability and sustainability.

The growth and evolution of transportation planning have reflected the changing demographic, economic, social, institutional, and technological characteristics of society. In the 1950s and much of the 1960s, transportation planning placed its primary emphasis on developing conceptual frameworks and analytical tools to support the construction of highways to accommodate the fast growth of populations, economies, and metropolitan areas and the rapid increase in automobile use (Meyer and Miller, 2001). Since the late 1960s, various social and environmental issues have become major concerns in transportation planning, and the efficient use of existing infrastructure, reduction of environmental impacts, and creation of livable cities have become important societal goals (Meyer and Miller, 2001). In particular, since the late 1990s, increasing awareness of global climate change and its strong connection with greenhouse gas (GHG) emissions from automobiles has placed sustainability at the center stage of transportation planning (Cervero, 1998; Newman and Kenworthy, 1999; Banister, 2002). Reflecting the diverse concerns and shifting emphasis, many new ideas and approaches have emerged in this field over the last four decades.

As the world becomes increasingly urbanized, new challenges and opportunities for transportation planners are mostly found in cities. Furthermore, transportation planning is especially needed in large urban areas where human activities are highly concentrated, and where spatial interactions among individuals and organizations are highly intensive and complex. Therefore, urban transportation planning has been, and will continue to be, the most active area of this field.

Basic Concepts and Frameworks

Transportation planning is an *interdisciplinary* field that has multiple intellectual roots. Like any other field, transportation planning employs many concepts and frameworks. These concepts and frameworks are from diverse sources, including the academic disciplines of planning, engineering, economics, and geography.

Transportation is closely tied to the concept of *mobility* – the ability to move people and goods from one location to another quickly, safely, and cost-effectively. Travel speed, road capacity, transit service frequency and reliability, the resulting time savings, and the resulting volumes of people and freight transported are among the common measures of mobility. High mobility is generally desirable, but should not be seen as the ultimate goal for planning and decision making. This is because the movement of people and goods is rarely an end in itself; rather, it is a means to achieve certain goals of individuals, organizations, and communities. Going to work, meeting friends, increasing labor productivity, manufacturing products, supporting economic development, and improving livability are examples of goals for which transportation serves as a means.

Once the goals associated with movement are taken into consideration, transportation is most fundamentally about

providing *accessibility*, which refers to the ease of reaching spatially distributed economic and social opportunities. The most basic mathematical expression of accessibility is the summation of spatially distributed opportunities weighted by the corresponding impedances, which measure the effect of travel distance, time, or monetary cost. Accessibility is a function of both *the transportation system* and *the activity system* that is usually characterized by land use. Increasing mobility is only one possible way to improve accessibility; shortening travel distances to destinations (i.e., proximity) and enabling remote interactions via the Internet or other electronic means are other possible ways (Levine et al., 2012; Shen, 2000). Focusing narrowly on the transportation system would be a partial and unsatisfactory approach to transportation planning.

Closely related to the notion of transportation as a means for participation in economic and social activities, transportation is *derived demand* in that its utilities come from the activities in various locations accessed via travel, not from the travel itself. In fact, with the exception of occasional driving for fun and biking and walking for recreation, transportation involves disutility in the form of time spent, in addition to monetary costs, safety risks, and discomfort. The *generalized travel cost* represents the sum of these various forms of disutility. Conceptually, transportation demand can shift because of changes in the activity system or in travel cost. The former is fundamentally determined by the level, location, and other characteristics of economic production and social interaction. The latter is primarily a function of transportation service quality, capacity, and price. Technology is often the key driving force behind dynamic changes in transportation demand and supply.

For participation in a given set of activities, individuals (and organizations, such as firms) will choose the *transportation options* that minimize their generalized travel costs. A transportation option is some combination of travel modes, routes, and schedules. Individuals with different socioeconomic, demographic, and lifestyle characteristics tend to find different options to be the best choice. For example, high-income professionals tend to place higher opportunity costs on travel time and hence prefer driving alone, whereas low-income workers and senior citizens tend to take public transit for, respectively, financial and health reasons. Similarly, different firms often choose different transportation options for shipping inputs and products. This diversity in demand provides the fundamental economic rationale for developing *multimodal transportation*. It also relates to many social equity issues that must be addressed in transportation decision making. Historically, advances in technology and restructurings of the economy have continuously redefined the available transportation options, which have recently added various combinations of telecommuting, teleshopping, telesocializing, and travel in physical space.

Individuals' rational transportation options, however, are not always preferable from the viewpoint of the society as a whole. This is because the *private cost of transportation* is often different from the *social cost of transportation* due to market and policy failures. In the United States and many other parts of the world, excessively large percentages of people choose to travel by private automobiles because the costs borne by them do not fully account for the *environmental and social externalities*. The discrepancy between private and social costs of transportation fundamentally justifies transportation planning in the public domain.

Transportation planning is an integral part of transportation decision making that takes place within an *institutional framework* consisting of special government agencies, service providers, participating organizations, community and interest groups, formal and informal interactions among these entities and people, funding mechanisms, and relevant laws, regulations, and norms. The composition and characteristics of the institutional environment largely define the transportation planning and decision-making process. Because the institutional environment is dynamic, the transportation planning and decision-making process also changes over time.

In the United States, transportation decision making in the 1950s and 1960s was based on the rational planning model (Meyer and Miller, 2001). The *rational planning process* typically included the key steps of defining goals and objectives, generating alternative approaches, evaluating alternatives, selecting an optimal alternative, and implementing the chosen alternative. Since the 1970s, *alternative planning models*, such as incremental planning, advocacy planning, and communicative planning, have been adopted in various contexts of transportation decision making. The application of these alternative planning models, often to complement rather than replace the rational model, reflects increasingly pluralistic values, diverse interests, complex organizational and power structures, acute awareness of environmental and financial constraints, and better understanding of uncertainties. It also reflects the growing sophistication of transportation planning as a field.

A systems perspective is effective for conceptualizing transportation planning problems and examining transportation impacts (Meyer and Miller, 2001). Transportation can be viewed as one system that is a component of numerous systems, such as the economic system and the ecological system. It interacts with many other systems, such as land use, energy, telecommunications, demographics, and social relationships. The transportation system itself is composed of subsystems, including users, modes, networks, facilities, and services, which are interconnected. Key concepts of systems theory have found useful applications in transportation planning (Meyer and Miller, 2001).

Analytical Tools

Transportation planning is also an *applied* field that relies on some standard tools to perform its functions. Many of these tools are analytical methods adopted from other fields. Examples are (1) statistical analysis of travel survey data, which applies methods developed in mathematics and statistics; (2) visualization and spatial analysis of transportation networks, facilities, built environments, peak and off-peak distributions of trip origins and destinations, and traffic flows using geographic information systems (GIS) and associated tools that originated from computer science and geography; (3) evaluation of alternative transportation projects and plans by conducting cost–benefit analysis or cost-effectiveness analysis, which are commonly applied techniques of economic

analysis; and (4) assessment of air quality, noise, and other environmental impacts of transportation by employing models built by scientists and engineers.

Some analytical tools have been developed specifically for transportation planning, even though many researchers from other disciplinary backgrounds have contributed to the efforts. The most fundamental tool of this kind is the urban transportation modeling system (UTMS), also known as the *four-step travel demand model*, which was originally developed in the 1950s and 1960s to generate forecasts of future volumes of travel on highways and transit lines (Meyer and Miller, 2001). The four steps refer to (1) trip generation, (2) trip distribution, (3) mode split, and (4) trip assignment. In the first step, the region is subdivided into a large number of areal units called traffic analysis zones (TAZs). Then, trips by travel purpose (e.g., home-based work, home-based other, and nonhome-based) are generated for each zone. Trips are classified as either a production (the home end or origin of a trip) or an attraction (the nonhome end or destination of a trip). Linear regression models are typically used to predict trips generated in a zone. Models for productions often include the number of households in the zone, household size, income, and automobile ownership as independent variables, whereas models for attractions often include numbers of jobs and households in the zone as independent variables.

In the second step, a mathematical procedure – most commonly a gravity model – is used to link the productions and attractions among all zones, and distribute the trips from each production zone to each attraction zone. These zone-to-zone flows include commuting trips to workplaces and other trips for various purposes such as shopping, visiting friends, and going to entertainment and sporting events. In an analogous form of the physics law of gravitation, the model computes the flow of trips between each zone pair based on the total number of trips produced in the origin zone, the total number of trips attracted to the destination zone, and travel impedance for going from the origin to the destination. Everything else being equal, the flow of trips from an origin zone to a destination zone increases with the total numbers of productions and attractions but decreases with the travel impedance.

In the third step, the predicted flows of trips are divided among available modes, such as drive alone, carpool, public transit, bike, and walk. Modal split is determined by the characteristics of competing modes such as speeds and costs; the characteristics of travelers such as income, automobile ownership, and age; as well as the characteristics of the trips themselves such as the trip purpose and the built environments of the origin and destination. Alternative methods have been developed to predict the percentage of trips using each available mode. Today, this procedure is typically accomplished by employing some kind of discrete choice model – most often the multinomial logit model – that computes the probability of choosing each mode.

In the final step, the predicted flows by mode are assigned to highway and transit networks. There are alternative computer algorithms for trip assignment, including minimum path assignment, various forms of user equilibrium assignment, and system optimum assignment. These alternative methods are based on different assumptions and, therefore, are suitable for different application contexts. Minimum path assignment, which loads all trips from any origin zone to any destination zone onto a single lowest cost path, is unreasonable except in the rare case where traffic is sparse, networks are uncongested, and alternative routes are highly differentiated in terms of travel cost. User equilibrium assignment, which generally assumes that each traveler chooses the perceived optimal route, is the more realistic approach. In contrast, system optimum assignment, which aims to minimize the overall congestion by informing and influencing travelers' route choice, is not behaviorally realistic but useful for improving the efficiency of existing networks.

For over half a century, the four-step travel demand model has been widely applied, especially to urban transportation planning in large metropolitan regions. While robust and long lasting, the model has been criticized for its various shortcomings (Banister, 2002). The four-step technical process is based on the rational model of planning and decision making, which has its limitations in dealing with problems involving diverse values and major uncertainties. In addition, it is built upon a weak behavioral foundation and relies primarily on empirical relationships to predict changes. Furthermore, it has a strong orientation toward promoting more and faster travel as the solution to transportation problems.

Researchers have effectively responded to some of the criticisms of the model. In particular, since the 1970s, *discrete choice models* based on the random utility theory have been incorporated into the modal split procedure of the traditional four-step model. Discrete choice models, developed by Domencich and McFadden (1975), Ben-Akiva and Lerman (1985), and others, model the decision making of individual travelers. The underlying assumption is that an individual traveler will choose an alternative that appears to maximize the utility. Recognizing that people do not always choose the objectively best option or make consistent choices over time, the utility function consists of two components: (1) the systematic and observable component that is identical to the conventional economic utility function, and (2) a random component that is used to capture variations in perceptions among individual travelers, as well as model specification and data measurement errors. The systematic component of the utility for an individual is a function of the attributes of each alternative and the characteristics of the individual.

With the random component included, the utility function does not determine which alternative is definitely the best for an individual, but enables the calculation of the probability that a given alternative will be chosen from the available choice set by an individual who has certain characteristics. The probability can have different functional forms based on different assumptions about the statistical distribution of the random component. The commonly used logit model results from an assumed Gumbel distribution of the random term.

The development of the discrete choice models has significantly advanced travel demand analysis by strengthening its behavioral foundation. Although modal choice has been the main area of application, discrete choice models have been employed to examine many other transportation decisions, including automobile ownership, trip destination choice, and residential location choice.

Almost in parallel, over three decades of research efforts have been made to develop *activity-based models*, which predict travel demand by focusing on people's participation in activities, such as working, shopping, and recreation. These models incorporate the time–space geography framework originated by Hägerstrand (1970), which conceptualizes the opportunities for, and constraints on, activity participation by individuals with various locational, demographic, and socioeconomic characteristics. Travel is derived from daily activities undertaken by households and individuals, and these activities are spatially distributed but temporarily sequenced and bounded. In addition to the activity-based framework, these models typically incorporate the tour instead of the trip as the basic unit for travel demand analysis, and apply microsimulation techniques at the disaggregate level of households and persons to convert activity and travel choices into a series of discrete decisions (Davidson et al., 2007).

The activity-based approach has several distinctive strengths that do not exist in the four-step modeling framework. It is highly disaggregate in representing the travel behavior of each individual and household, which helps it avoid the errors associated with generalization and lend itself to a more realistic and accurate portrayal of travel behavior. It explicitly models the relationships between an individual's different activities (e.g., work and nonwork), as well as the interactions between household members. These capacities, together with detailed information about activities and travel segmentation, substantially broaden the range of transportation planning and policy issues that the models can help address. For example, activity-based models can more realistically assess the effect of implementing high-occupancy vehicle (HOV) lanes by accounting for intrahousehold travel decision making. Another example of a useful application of activity-based models is to examine the amount of travel reduction resulting from telecommuting, which may redistribute some component trips (e.g., dropping off the kids at school on the way to one's workplace in the morning and shopping for groceries in a supermarket on the way back in the afternoon) of the preexisting commuting tour to other time windows.

Activity-based models also have shortcomings: they are more complicated and, therefore, the development cost is much higher; they require the collection of disaggregate travel diary data that are more complete and accurate in identifying household interactions as well as individuals' activities, which is a high expectation not easily met through home interview surveys. Furthermore, there are still some challenging theoretical and technical issues that need to be addressed in future research (Davidson et al., 2007).

Activity-based models have recently become viable alternatives to four-step models for a variety of applications. In the United States, a number of metropolitan areas, including New York, San Francisco, Denver, and Portland, are now using activity-based models for transportation planning. Several other metropolitan areas, including Chicago, Seattle, and Houston, are currently developing models of this type. Activity-based models have also been developed and applied in several European cities. However, the practical – not just the theoretical – advantages of the activity-based approach must be more clearly demonstrated before it can be widely adopted for travel demand modeling.

Another important technical tool, which produces required data for long-range travel demand analysis, is the *land use model*. Since the movement of people and goods is derived from the economic and social activities in which they participate, or to which they contribute, the quantities and locations of residential, employment, commercial, service, and recreational activities must first be measured or predicted before the demand for transportation is analyzed. Understanding and modeling the activity system are especially critical for the long term because it cannot be assumed to remain the same. Land use models serve as an essential tool for depicting likely future changes in the activity system as a basic input for forecasting future travel demand. At the same time, these models also provide the capacity for assessing the likely short-term and long-term impacts of proposed changes in the transportation system on the activity system.

Urban modeling, a field of research focusing on building land use models for urban areas, has been active for over half a century, although its development certainly did not follow a smooth upward curve. In general, these models make conditional predictions of land use changes in a city or metropolitan region in accordance with changes in the economy, employment, and population. The models either explicitly or implicitly incorporate both the market mechanisms for, and the policy constraints on, urban land allocation. Accessibility and other attributes are evaluated for each land unit (e.g., a parcel or a grid cell), and available land units are allocated to competing users (e.g., firms in different economic sectors and households with different socioeconomic characteristics) who place different utilities on these attributes.

Despite the fundamental importance of accessibility and spatial interaction in land use models, in the United States the analysis of the activity system proceeded essentially independent of the analysis of the transportation system for decades. That was primarily because, institutionally and professionally, different agencies and people were responsible for the respective tasks (Meyer and Miller, 2001). Since the early 1990s, however, the connection between land use and travel demand analyses has been greatly strengthened. Recognizing the effects of transportation on land use and the environment, the 1990 Clean Air Act Amendments and the 1991 Intermodal Surface Transportation Efficiency Act both required metropolitan planning agencies to integrate land use and transportation planning (Meyer and Miller, 2001; Waddell, 2002). Additional pressure for this integration has come from state programs for growth management that promote a stronger linkage of land use and transportation planning, as well as from concerned researchers, practitioners, and environmental advocates (Waddell, 2002).

Today, in various parts of the world, operational models are employed to produce land use and travel demand forecasts by accounting for the interactions between the two systems. Some of these models are *integrated land use and transportation models*, which represent the interactions between land use and transportation within a unified model of activity and transportation systems. Others are separate but loosely coupled models of

land use and travel demand that feed each other with outputs (e.g., jobs and households from the land use model, and modal shares, travel times, and travel costs from the travel demand model) for future years. In either case, the dynamic relationship between land use and transportation – for instance, improved accessibility in an area attracts more firms and households to locate there, which in turn can cause traffic congestion, lower accessibility, and discourage further growth – can be captured.

Evolving Focus from Mobility to Accessibility, Livability, and Sustainability

The most basic function of a transportation system is to increase mobility, but the true economic value of mobility is not realized until it effectively connects people and goods to their desired destinations. Therefore, while mobility is the most distinctive property of a transportation system, accessibility is conceptually the more appropriate measure of its performance (Levine et al., 2012). As livability and sustainability become primary goals for a growing number of countries, regions, and cities, accessibility will increasingly become the main focus of transportation planning. In a world with finite resources but growing needs, transportation planners are responsible for helping improve quality of life by increasing accessibility while avoiding irreparable damage to the environment by restraining travel-related energy consumption and emissions.

Despite the clear advantage of emphasizing accessibility over mobility, in most parts of the world enhancing mobility used to be, and in many cases continues to be, the primary goal for transportation planning. In the Unites States, for almost half a century after World War II, transportation planning aimed principally at building highways to accommodate private automobiles. While it certainly achieved many positive results, including higher economic productivity and speedier travel for the great majority of the population, it also generated various undesirable consequences. Traffic congestion, lack of adequate public transit service and bike and pedestrian facilities, and degradation of old urban neighborhoods have been some of the negative outcomes (Cervero, 1998). In addition, the automobile-oriented mobility enhancement has resulted in a huge accessibility gap for those who depend on public transit (Shen, 1998).

The United States has become the world's most automobile-dependent country: the number of vehicles per capita is about 0.8; and 76% of commuters drive alone, whereas barely 4.9% of them take public transportation (Davis et al., 2012). This high level of automobile dependency poses major challenges for energy security and environmental sustainability: with less than 5% of the world's population, the United States consumes about 22% of global petroleum production and generates about 20% of total carbon dioxide emissions, and the transportation sector accounts for approximately one-third of its carbon emissions (Davis et al., 2012).

Continuation of mobility-focused transportation development would contradict the global trend of sustainable development. Although the awareness of certain negative environmental and social consequences of automobile dependency has existed for decades, the real wake-up call came in the 1990s with the understanding of climate change and its strong connection to transportation. The adoption of the Kyoto Protocol in 1997 marked the beginning of a new era of transportation planning in the United States, as well as in many other countries. Reducing GHG emissions has become a key objective not only for automobile manufacturers and energy suppliers, but also for transportation planners. This is because, for the foreseeable future, more efficient vehicles and lower carbon fuels will not effectively mitigate GHG emissions unless people reduce their motorized travel (Hankey and Marshall, 2010).

By shifting focus from mobility to accessibility, and by setting livability and sustainability as the ultimate goals, transportation planning has opened up to new ideas and incorporated new approaches. Most importantly, land use is no longer taken as a given, but seen as a means for achieving the goals by modifying the patterns of activity participation and travel. New urbanism design guidelines, smart growth principles, and transit-oriented developments (TOD) have become widely accepted land use measures for reducing automobile use and encouraging transit use, biking, and walking. There have been many other significant changes, including a stronger emphasis on transportation system management and demand management, wider applications of technology (e.g., telecommuting and real-time transit information), more complete cost accounting in transportation pricing, and deepened concerns for social equity and environmental impact (Meyer and Miller, 2001).

More opportunities have been created for new and exciting research in transportation planning. For example, a large number of publications have been added to the literature on the travel behavioral effects of land use initiatives (Ewing and Cervero, 2010). Vigorous research activities have also been seen in important topic areas of travel demand management, including parking charges, congestion pricing, and mileage-based fuel taxes. In addition, solid efforts have been made to shed new light on the viability and values of transportation alternatives. For example, real-time transit information accessible through mobile phones creates new possibilities to make public transportation more attractive and competitive (Watkins et al., 2011). Another example is the sudden surge of research interest in nonmotorized transportation, which has obtained encouraging findings of measurable health benefits from biking and walking.

The new era has also witnessed significant changes in the practice of transportation planning. Travel demand models and land use models have become more integrated (Waddell, 2002). Numerous development projects that follow the guidelines of new urbanisms and TOD have been implemented, and smart growth principles have been adopted by many planning agencies in the United States and beyond. More carpool and bus lanes have been created, and more bicycle facilities have been built, in many cities. Parking regulations have been modified by many local governments. These are just a few examples of the encouraging changes.

Challenges and Opportunities for Future Research

Transportation planning, which is essential for well-informed decision making on infrastructure, services, and policies for overcoming spatial barriers to economic and social activities,

will continue to be a vibrant field in the future. To play its role effectively – in particular, to help cities and regions build appropriate connections between their mobility and accessibility improvement strategies and their livability and sustainability goals – transportation planning must continue to advance knowledge. There are major knowledge gaps, including the following, which present both great challenges and opportunities for future research.

First, there is *confusion over the effect of land use change on travel behavior*. A large volume of research on this topic has so far generated inconsistent findings. Most of the empirical studies based on intrametropolitan data have identified a statistically significant relationship between land use and travel demand, but suggested a rather modest effect as shown by the low elasticity values (Ewing and Cervero, 2010). Much greater effects, however, are seen by comparing cities or metropolitan areas using aggregate data on density and per capita energy consumption (Newman and Kenworthy, 1999; Glaeser, 2011). What do these inconsistent findings imply regarding the relationship between the activity system and the transportation system? Conceptually, land use change, such as densification or diversification, directly alters only some physical aspect of the activity system. It may have little impact on individuals' current pattern of activity participation, and hence their short-term travel behavior remains unchanged. The challenge is to reconcile the generally weak short-term effects and the much stronger cumulative effects. In addition to land use change, what other factors may trigger substantial adjustments in people's spatial and temporal arrangements of activities? Implementation of complementary governmental policies such as congestion pricing and mileage-based fuel tax, provision of convenient and safe transit and nonmotorized options, formation of new societal norms, and change of individual values are all possible factors for future investigation.

Second, there is currently *a lack of systematic research effort to support planning for alternative transportation modes*. Although there are recent empirical studies on various issues concerning public transit and nomotorized travel, they do not serve as a consolidated foundation for reenvisioning the alternative modes. For example, several studies have identified important economic and psychological benefits brought to transit riders by real-time bus arrival information systems (Zhang et al., 2008; Watkins et al., 2011), and many research projects have recently been undertaken to show the health benefits associated with walking and biking. These efforts, although offering useful insights, do not provide sufficient information to address transportation planning questions such as whether or not public transit will become a viable option for suburban residents, and whether or not adding hundreds of miles of bicycle lanes to the existing street network will produce enough benefits to offset the costs.

Third, there is *rather limited understanding of the likely impacts of technologies* on the transportation system, the activity system, and their interrelation in the foreseeable future. Major breakthroughs in automotive and fuel technologies could dramatically reduce oil dependence and GHG emissions, and significantly improve the sustainability prospect of the transportation system. Although future advances in these technologies are uncertain, transportation planning will be more informative if it can draw from careful analyses of alternative scenarios and implications.

Similarly, information and communications technologies (ICT) have started to demonstrate their huge capacity to modify the patterns of activity and transportation of individuals and organizations, but the implications for transportation planning are not clear. Transportation researchers tend to concentrate their analysis on the direct and short-term effects of ICT in the forms of travel substitution, generation, and reconfiguration (Mokhtarian, 1990). The longer term secondary impacts on the transportation system through relocations of residence and activities are insufficiently investigated (Shen, 2000). Perhaps more critically for livability and sustainability, timely efforts should be made to understand how future developments in ICT may provide individuals and organizations with an adaptive ability, so that they can maintain their basic economic and social functions while adjusting their spatial and temporal configurations of activities at a relatively low transaction cost.

Fourth, transportation planners *must learn to incorporate uncertainty explicitly into the planning process*. The future is always unknown, but the level of uncertainty, and hence the risk, is especially high when the decision making is about a major project or policy that has long-term livability and sustainability consequences. High-level uncertainty is quite common in planning and decision making, but for decades transportation planning managed to largely follow the rational planning process (Banister, 2002). Given its strong relationships with the future well-being of human society, on one hand, and with dynamic technological forces, on the other hand, the transportation system should be planned with a clear understanding of the assumptions, uncertainties, and complexities involved. In addition, it will be critical to devise mechanisms to support 'incremental but informed planning' by continuously monitoring technological, economic, social, and behavioral changes over time, and by promptly veering onto another course if future events make that necessary (Wachs, 2012).

Finally, with a rapidly growing population and economy, and rising levels of automobile ownership, *developing countries are the new frontier for transportation planning research and innovation*. Because the majority of the world's population reside in these countries, future changes in their travel behavior will have a profound implication for sustainability. The transportation sector currently accounts for a lower percentage of the total GHG emissions in developing countries, but it is increasing rapidly with the expansion of the car-owning middle-class population (Dulal et al., 2011). For example, China has recently become the world's top country in vehicle sales, and its annual vehicle growth rates for the next 10 years are projected to be around 10%. Fast-paced motorization and urbanization in developing countries present a highly challenging yet rewarding task for transportation planners: to help achieve livability and sustainability goals by helping reshape the mobility and activity systems. What do they need to study to prepare themselves for this huge responsibility?

The above list is far from exhaustive; one could add many more future challenges and opportunities, ranging from improving travel demand models to reforming transportation planning institutions. But the list is sufficient to illustrate the

variety of important problems that will continue to make transportation planning an exciting, and hopefully fruitful, field for the foreseeable future.

See also: Urban Internal Spatial Structure; Urban Planning: Methods and Technologies; Urban Policy in North America; Urban Sprawl.

Bibliography

Banister, David, 2002. Transport Planning, second ed. Spon Press, London.

Ben-Akiva, Moshe, Lerman, Steven R., 1985. Discrete Choice Analysis. MIT Press, Cambridge, MA.

Cervero, Robert, 1998. The Transit Metropolis: A Global Inquiry. Island Press, Washington, DC.

Davidson, William, Donnelly, Robert, Vovsha, Peter, Freedman, Joel, Ruegg, Steve, Hicks, Jim, Castiglione, Joe, Picado, Rosella, 2007. Synthesis of first practices and operational research approaches in activity-based travel demand modeling. Transportation Research Part A 41, 464–488.

Davis, Stacy C., Diegel, Susan W., Boundy, Robert G., 2012. Transportation Energy Data Book, thirty first ed. Oak Ridge National Laboratory, Knoxville, TN.

Domencich, Tom A., McFadden, Daniel L., 1975. Urban Travel Demand: A Behavioral Analysis. North-Holland, Amsterdam, NL.

Dulal, Hari Bansha, Brodnig, Gernot, Onoriose, Charity G., 2011. Climate change mitigation in the transport sector through urban planning: a review. Habitat International 35, 494–500.

Ewing, Reid, Cervero, Robert, 2010. Travel and the built environment: a meta-analysis. Journal of the American Planning Association 76, 265–294.

Glaeser, Edward L., 2011. Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier. Penguin Books, New York.

Hägerstrand, Torsten, 1970. What about people in regional science? Papers of the Regional Science Association 24, 6–21.

Hankey, Steve, Marshall, Julian D., 2010. Impacts of urban form on future US passenger-vehicle greenhouse gas emissions. Energy Policy 38, 4880–4887.

Levine, Jonathan, Grengs, Joe, Shen, Qingyun, Shen, Qing, 2012. Does accessibility require density or speed? Journal of the American Planning Association 78, 157–172.

Meyer, Michael D., Miller, Eric J., 2001. Urban Transportation Planning: A Decision-oriented Approach, second ed. McGraw-Hill, New York.

Mokhtarian, Patricia L., 1990. A typology of relationships between telecommunications and transportation. Transportation Research Part A 24, 231–242.

Newman, Peter, Kenworthy, Jeffrey, 1999. Sustainability and Cities: Overcoming Automobile Dependence. Island Press, Washington, DC.

Shen, Qing, 1998. Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers. Environment and Planning B 25, 345–365.

Shen, Qing, 2000. New telecommunications and residential location flexibility. Environment and Planning A 32, 1445–1463.

Wachs, Martin, 2012. Planning for high speed rail. Access 41, 38–40.

Waddell, Paul, 2002. Modeling urban development for land use, transportation, and environmental planning. Journal of the American Planning Association 68, 297–314.

Watkins, Kari, Ferris, Brian, Borning, Alan, Rutherford, G. Scott, Layton, David, 2011. Where is my bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. Transportation Research Part A 45, 839–848.

Zhang, Feng, Shen, Qing, Clifton, Kelly, 2008. Examination of traveler responses to real-time information about bus arrivals using panel data. Transportation Research Record 2082, 107–115.

Transportation, Psychology of

Raymond W Novaco, University of California, Irvine, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psychology of transportation entails the study of human behavior and well-being in conjunction with the regularized movement of people, goods, and services regarding work, personal life, and community activity systems. It includes the interface of humans with transport system components, including their design and operation, provisions for human needs, forms of utilization, and social impacts, including public safety.

The psychology of transportation entails the study of human behavior and well-being in conjunction with the regularized movement of people, goods, and services regarding work, personal life, and community activity systems. The information age has radically modified both work products and the means of transportation, yet physical mobility on transport systems remains fundamental to our commerce, recreation, and other life-sustaining activities. The psychology of transportation also includes the interface of humans with transport vehicles and facilities, such as their design for optimal operation, provisions for human needs, forms of utilization, and social impacts, including public safety.

Overview and Scope

The domain of transportation has great relevance for psychology, precisely because of its salience in the routines of daily life and its importance for meeting personal, family, and organizational goals. Correspondingly, because components of transport systems are operated by people, psychology has an important role in system development, efficiency, and safety. Human cognitive, personality, and performance factors affect vehicle operation. Recognition of the intrinsic value of psychology produced the launching of a new section of the Transportation *Research* journal devoted to psychology and behavior, which opened with the article by Groeger and Rothengatter (1998).

Transport systems shape the structure of our communities and have a significant impact on both personal and community well-being. For example, urban sprawl, with its low-density land use and automobile commuting dependence, has public health impacts associated with air pollution, motor vehicle crashes, and psychological stress (Frumkin, 2002). Workers' average daily commuting time varies substantially across countries, ranging from about 15 min in Finland and New Zealand to about 40 min in Korea and Japan, with the United States and United Kingdom being around 25 min (Organization for Economic Co-Operation and Development, 2013; Pisarski, 2006). Within countries, there is, of course, much regional variation, such as in the United States where metropolitan area commute time averages range from 14 to 35 min (United States Census Bureau, 2011).

Personal well-being and the progress of human societies have always been associated with mobility and transportation. People respond to environmental constraints on mobility and strive to enhance well-being. However, the mobility and well-being relationship is not uniform, as higher levels of travel may be reflective of constraints on opportunities, depletion of resources, and impairments to personal health and job performance. What happens on a particular journey depends on many intertwined psychological elements, including personality disposition, attitudes about the origin and destination of the trip, and resources for choosing alternative travel modes and schedules, including telecommuting. Working at home has been the fastest growing commuting mode from 1980 to 2010, while carpooling has steadily declined (Pisarski and Polzin, 2013). There is thus a vast terrain for psychological research and application of psychological knowledge. The creation and use of transportation systems, the experience and behavior of people in transportation conditions, and the effects of mobility constraints on personal and community health are important areas in the psychology of transportation.

Human factors have long been recognized as being important to vehicle operation, most certainly heightened by the application of science to warfare. However, the psychology of transportation in civil society is much broader in scope than the interface between the human operator, the vehicle, and the road design or rail system (Rothengatter and Vaya, 1997). It entails travel mode decisions, driver attitudes and performance, the interplay between work and family, emotional well-being, and major societal issues, such as the needs of special populations, the definition of the workplace, the provision of transportation facilities, and travel safety.

The field of transportation has been dominated by engineers, urban planners, and economists; but the contributions of psychologists and sociologists integrated into this arena for decades (e.g., Altman et al., 1981; Stokols and Novaco, 1981; Wachs and Crawford, 1992). Discussions of the social costs of transportation are now common, and psychological elements are given weight in such considerations. Because individuals, families, and work organizations recognize that they absorb hidden costs associated with traffic congestion, air pollution, and traffic noise, psychological science is important for understanding how to optimize transportation systems and to assess the consequences of travel strains.

One area where psychology's impact on the transportation field has grown substantially is public safety, particularly with regard to driving behavior. Interest in this domain has expanded from traditional work on unsafe driving associated

with fatigue, substance use, driving histories, aging, and risk-taking tendencies to include developmental issues associated with younger and older drivers, specialized attention to personality styles, motor vehicle accident trauma, antisocial behavior on public transport, driver education, and problem-targeted intervention programs. Gossling (2013) has suggested that 'clinical transport psychology' be developed as an area. There is now bountiful literature on 'aggressive driving' (e.g., Galovski and Blanchard, 2004; Shinar and Compton, 2004) and various instruments to assess 'driving anger' (e.g., Deffenbacher et al., 2002; Sullman and Stephens, 2013). Although far less extensive as that in the personality/emotion domain, the study of psychological trauma associated with motor vehicles accidents (e.g., Mayou et al., 2002; Wiesenthal and Vingilis, 2013) is an important and rich area for research, particularly given the high prevalence of vehicle crashes.

Historical Development of the Field

Involvement by psychologists in the field of transportation began in the 1920s when transport companies and government agencies sought to improve the selection of drivers of public transport vehicles, and psychometric testing became used in choosing the operators of buses, trams, and trains. A central concern was to minimize accident risk, and psychological tests were used to screen applicants who could be anticipated to be accident prone. Wartime studies of adjustment among military personnel and the early psychological formulations of stress, adaptation, and coping that resulted from observations made by psychological researchers working with the military extended quite nicely to the subject of driver performance demands and safety risk. As well, industrial psychology research on fatigue at the workstation transferred to the issue of vigilance in professional train and road drivers (Barjonet, 1997).

As the motor car progressively came to dominate travel in Western societies in the 1950s and 1960s, car ownership grew dramatically; accident rates and road deaths became serious public health problems. In most European countries, the number of road deaths doubled between 1955 and 1970, much worse than in the US Public safety policies, such as speed limits, penalties for drunk driving, and ultimately seat belts were instituted to curtail casualties, and research into the human causes of accidents came into play. Social psychological knowledge was used in public persuasion efforts, and theories and techniques of learning found fertile ground in driver education and rehabilitation. The advent of the computer augmented the capacity to analyze and integrate data, and understanding driver-based safety risk in the 1970s and 1980s often took the form of psychological modeling incorporating personal background, attitudinal, motivational, and emotion variables on top of ergonomics. That period is now overshadowed by twenty-first century technology of traffic management, collision warning, and collision mitigation systems.

The growth in automobile commuting that occurred between 1960 and 1980, especially in the United States, prompted interest in modeling how trips were generated, how they were distributed across travel zones, and how trips split across available transportation modes (car, bus, train). The boom in automobile commuting followed from substantial growth in the number of jobs, in the working age population, and in the number of women in the labor force. In addition, homes and jobs shifted to suburban locations. Because neither highway nor transit systems were designed for suburb-to-suburb commuting, traffic congestion in metropolitan areas became a salient issue. In the United States, the use of private vehicles between home and work nearly doubled from 43 million users in 1960 to 83 million in 1980. At the same time, alternative modes of travel declined, vehicle occupancy declined, and national investment in highway infrastructure fell far behind the demand for road use (Pisarski, 1987). The suburbanization trend continued in the 1990s, with more than 50% of commuters in the 2000 census being suburban residents, and suburb-to-suburb commuting grew 64% from 1990 to 2000 (Pisarski, 2006). Quite strikingly, carpooling declined each decade from 1980 to 2010 (Pisarski and Polzin, 2013).

Similar trends have occurred in Europe. Since 1970, 87% of the growth in the transport of people has occurred in the road passenger transport sector (European Commission, 2006). From 1970 to 2003, the automobile share of passenger travel increased from 74 to 79%, while the rail transport share declined from 10 to 6%, and the bus/coach share dropped from 13 to 8%. The information society has not reduced the need for mobility – a European citizen traveled 17 km a day in 1970 and that rose to 35 km three decades later, heightening concern about traffic congestion (European Commission, 2006). In Southeast Asia, mega-urban regions have immense public transport problems, with travel times to work of 2 h being normal, as is the associated lost productivity and poor quality of life (Jones, 2002).

European cities, with their treasured historic centers having limited road space and parking facilities, became concerned about the detrimental effects of traffic congestion and pollution and were keen to impose access restrictions and 'traffic calming' to reduce automobile use and safeguard pedestrians. These trends spurred interest in road pricing (Jones and Hervik, 1992) and individual travel behavior. Computer analytic methodology spurred travel demand modeling, which blended consumer theory in economics with choice theory in psychology, which assume that people use some utility or value function in assessing choice alternatives. Psychological variables continue to facilitate understanding of the use of priced managed lanes (e.g., Devarasetty et al., 2014) and switching to public transit (e.g., Chen and Chao, 2011).

Attention to the transportation needs of special populations, such as working mothers, the elderly, and those affected by physical disabilities, also came into focus during the 1970s and 1980s. An important component of the auto-commuting boom was a large increase in the number of women in the labor force (Grieco et al., 1989; Pisarski, 1987). Single women with children who lived in suburbs were very dependent on the automobile (Rosenbloom, 1992). In the United States, most trips made by working mothers and their children cannot be made by transit, and the impact of travel constraints on the quality of family life is an important topic. In the 1990s, telecommuting emerged, partly as a way to reduce strain on family life. The convergence of technological, economic, social, and psychological factors made work less

place dependent, and telecommuting was heralded as a viable alternative to auto commuting (Mokhtarian, 1990). From 1999 to 2012 in the United States, access to flexible workplace benefits (requiring a formal telework program) increased from 3% of all workers to 6% (Pisarski and Polzin, 2013).

The information age had many effects on transportation systems. The development of intelligent transport systems, such as in-vehicle navigation (route guidance and congestion warnings) and increased sophistication of driver panel displays, as well as mobile phone use and the intensification of information-based work demands, have raised the importance of driver control, visual search efficiency, and mental load on driving proficiency. Both legal sanctions and further technological developments are being guided by understanding psychological factors. On the horizon of advanced technologies are Intelligent Vehicle Highway Systems that automatically control speed, steering, and spacing. Psychological issues of personal control and freedom (long-standing themes intrinsic to car ownership and driving) and matters of privacy come to the fore with regard to having a vehicle under the control of a computerized system, GPS tracking of vehicle usage, or on-board video monitoring of driving behavior.

Human Factors and Driver Performance

Psychological variables affect the operation of vehicles, as can be seen in studies of driver error and performance decrements (cf Rothengatter and Vaya, 1997). Fatigue has been the most researched topic in this regard. Ordinarily understood as physical exhaustion, it can also involve boredom, task aversion, uncontrollable sleepiness, hypnotic trance-like states, or loss of alertness. Driving performance decrements occur before the driver is aware of the fatigue. Fatigued drivers become complacent and underestimate the effort needed to maintain driving performance. Hours of service limitations are implemented for professional drivers and pilots to guard against fatigue-related safety errors. Police officers must be trained to recognize fatigue. Armstrong et al. (2013) found that 19% of 1600 Australian drivers reported having nearly crashed while fatigued, and 2.4% did crash. Physiological correlates of fatigue have been studied in both driving simulator studies and open-road tests. Heart rate decreases with prolonged driving time, but results across studies have not been uniform. Focal fixation, detected by eye-tracking, is associated with fatigue, but that also occurs in conjunction with hazard perception.

Driver alertness is clearly related to collision avoidance. The driver, as an information processor, must be proficient in making time-to-collision judgments – for example, time to collision is underestimated for smaller vehicles. Generally, the more enriched the visual environment, the lower the underestimate of time-to-collision. Vehicle design modifications have been made progressively in rear brake signals and daytime running lights to reduce collisions, and road design features involving reflectors, warning signs to reduce speed on curves, and edge lines are intended to enhance the input of safety information. Major advances have been made in forward collision and lane departure warning systems, some of which take action autonomously.

Driver expectations and interpretations, however, interact with visual information. Expectancies developed from driving experience may have adverse effects. The person may omit perceiving relevant cues or fail to give them appropriate weight in estimating a hazardous condition. 'Driving experience' does not guarantee expertise, as an experienced driver may have fixed expectations about how a driving situation will progress, which may interfere with accurate perceptions and interpretations of relevant cues. Driver behavior is a function of what drivers believe and feel they can do, as well as what they see and are told.

The topical salience of driver distraction has been intensified by the information age as the use of portable devices for information, communication, and entertainment have infiltrated the psychological space of vehicle operations and have been attributed to be the cause of crashes of cars, trucks, busses, and trains. Text messaging is more impairing than is cell phone usage, and it is a prevalent problem among teenage drivers. Media-based campaigns (as occurred regarding seat belt use) aiming to inform, persuade, and motivate the public to change attitudes and behaviors to improve transport safety are an important domain for the application of social psychological knowledge and behavioral modeling.

Driver Risk Taking, Attitudes, Motivation, and Emotion

The study of driver performance lends itself to experimental lab analyses of perception, decision-making, and reaction time. In contrast, the research on driver risk taking has been conducted largely by questionnaires. Driver attitudes and personality are studied as indicators of the propensity to commit driving errors and traffic code violations, thereby raising accident liability. Thus, rather than focus exclusively on the driver's psychomotor or perceptual skills, attention is given to higher order cognitive processes and motivation (cf Rothengatter and Vaya, 1997). However, there is an issue of whether social desirability affects what is self-reported about driving behavior and self-reported about accidents (Wahlberg et al., 2010).

Self-reported violations and risky driving behavior, as assessed by driver behavior questionnaire measures, have been found to predict involvement in accidents, independent of the amount of yearly driving. Because self-evaluations of driving habits are longitudinally predictive of driving violations and accidents, it then becomes feasible to identify problematic drivers and augment their safety motives. Since these relationships hold strongly for young men, they are commonly a target group for intervention.

The most consistent finding about driver risk taking is its association with a sensation-seeking personality, which is thought to have a biological basis. High sensation seekers are more likely to drive drunk, exceed speed limits, take risks in passing and lane changing, not use seat belts, receive citations for traffic violations, and have crashes. These findings have been obtained in studies conducted in several European countries, Canada, the United States, Australia, and New Zealand. There is substantial research literature on the psychosocial factors influencing the risky driving propensity among young drivers, including speeding, drink driving, not wearing a seat belt, and driving while fatigued, as well as proclivity toward distraction.

Human behavior is high adaptive, and drivers adjust to changes in the intrinsic safety of the driving environment. In hazardous weather, drivers slow down and sharpen alertness. Alternatively, as safety in the driving environment is enhanced through improvements in roadway conditions and vehicle design, drivers may be inclined to drive faster, shorten the distance in following, overtake more readily, or be less attentive to hazards. Therefore, improved vehicle safety features may lead some drivers to overcompensate for perceived risk reduction, driving less safely because they feel more insulated from collision or injury.

Aggressive driving and its alias 'road rage' are prominent topics in the traffic safety field in many countries (Brewer, 2000; Galovski and Blanchard, 2004; Shinar, 1998). Aggression has been part of driving since chariots careened around the Circus Maximus in ancient Rome. Indeed, the symbolization of the automobile has been infused with themes of aggression, which is evident in vehicle names and designs. Uncontrolled anger and aggression on roadways do, of course, constitute a significant driver behavior issue, but neither the magnitude of the problem nor its societal trend is clear. While human aggression springs from many causes, aggression on roadways is a product of various disinhibitory influences (Novaco, 1991), such as anonymity, opportunity to escape, increasingly fortified vehicles, and strong elevations in physiological arousal produced by driving situations. The personalized territoriality of the vehicle and diffusion of roadway aggression scripts in the media also facilitate antagonism in driving when the behavior of other drivers is viewed as provoking.

Driver Selection, Training, and Rehabilitation

Psychologists have developed aptitude tests to insure public safety on road, rail, and air travel. Driving safety has been classically viewed as involving the three 'Es': engineering, enforcement, and education. Accident liability is primarily due to underestimation of risks, negligence or carelessness, personality, attitudinal disposition, or substance misuse, rather than to roadway or vehicle engineering. Hence, reduction in risky driving hinges on detection of problematic drivers and on driver training and rehabilitation programs. There is wide variability across countries in how systematically this is approached.

Reconviction for traffic offenses and for serious traffic offenses has been found in New Zealand national data, gathered prospectively, to be a function of age at first conviction, being male, number of prior convictions, and having a traffic conviction as the most serious offense; moreover, statistical models have achieved high predictive power in identifying traffic offenders at high risk for reconviction for an alcohol-related driving offense. In Germany, psychologists developed models to predict reoffending drivers, and their rehabilitation courses reduced drunk-driving recidivism. Specifically targeted driver education programs have been successful, but it has generally been found that secondary school driver training programs do not reduce accidents, as young people learn to drive from relatives and friends, not just schools. The propensity for risky driving, however, may be ascendant in a willingness to commit violations that can be detected in youth as predrivers (Rowe et al., 2013).

Older driver safety is a significant public health issue that is growing in magnitude (Wang and Carr, 2004). The number of people above the age of 70 who drive is progressively increasing, and that segment of the population is at greater risk for motor vehicle accidents due to cognitive and psychomotor functional impairments, the use of pharmaceuticals, and medical conditions. They are also more likely to die in a crash. Older drivers typically self-regulate by modifying their driving patterns (e.g., not driving at night; not using left-turn lanes). Advances in technology, including GPS navigation, collision avoidance systems, improved headlamps, and improved insulation from exterior noise serve to attenuate accident risk, but psychologists will continue to have important roles in developing driver screening systems and driver rehabilitation programs. How older drivers construe the risk of their driving (Siren and Kjaer, 2011) is an important psychological issue that bears on enhancing their safety and that of others.

Travel Demand and Alternative Modes of Travel

Transport systems planners use mathematical theories about choice in modeling travel demand, and psychological variables add to what is otherwise offered by economic models. Because choices are often grounded in perceptions and subjective valuations, psychological research methods help to ascertain the relationship between travel time and the desirability of travel mode options or goodness of fit with personal needs. Travel time, cost of alternative modes, availability of mass transit, activity plans, work demands, and personal preferences are weighed by the traveler in making transportation choices.

Psychological theory has been applied to change travel behavior, such as using reinforcement theory to boost transit ridership. Transit use has been promoted by offering cash and travel prizes through lottery-type drawings, as well as value coupons and tokens (for food, goods, and services) being distributed with bulk purchases of transit passes or dispensed contingent on boarding a bus. Similarly, inducements for ridesharing have been given by providing special lanes on congested freeways for multiple occupancy vehicles. Such schemes are more effective when the special lane has been added to the road, as opposed to when an existing lane has been converted. Other efforts to reduce solo driving have involved the provision of priority parking facilities and/or reduced parking rates for carpools and vanpools. Despite many schemes to boost carpooling and transit use, the 'drive-alone' mode share in the United States increased each decade from 1980 to 2010 (Pisarski and Polzin, 2013), which is unlikely to reverse as auto manufacturers continue to enhance private vehicle comfort and amenities.

Embedded in the issue of transit use is its social image, particularly with regard to buses. In many countries, persons of high social status commonly do intercity travel by rail for work commutes and discretionary trips, but bus travel typically has the image of being less than prestigious. As well, people may be disinclined to use the bus because of a desire to avoid crime victimization on the bus and at bus stops. Psychological knowledge can be utilized in modifying unfavorable images of transit use and to reduce unwarranted fears about safety risk.

Commuting Stress

Exposure to traffic congestion is stressful by virtue of impedance – the blocking or thwarting of movement and goal attainment. Travel impedance has both physical and perceptual dimensions (Novaco et al., 1990). 'Physical impedance' can be objectively measured by the distance and time of the journey, along with the number of road exchanges – each representing a node of congestion. Time of day is relevant, as congestion is greater in the evening, and commuting stress is compounded by strains from the workday. 'Subjective impedance' is the person's perception of commuting constraints, assessed by questionnaire items or structured interview. Because the commute is sandwiched between the home and the job, the demands and affordances of each of these domains, as well as personality factors, bear on the commuting experience.

Commuting stress has been found to be associated with elevated blood pressure, negative mood on arrival at work, negative mood at home in the evening, lowered frustration tolerance, cognitive performance impairments, illness occasions, work absences, job instability, lowered residential satisfaction, and decrements in overall life satisfaction (for reviews, Koslowsky et al., 1995; Novaco and Gonzalez, 2009). Female solo drivers in high-impedance commutes are most negatively affected, independent of job involvement and household income. Because women tend to have more logistical responsibilities for home and family, they are perhaps more sensitized to stress on high-impedance routes (Novaco et al., 1991).

A useful concept in understanding commuting stress is 'inter-domain transfer effects,' whereby the psychological consequences of environmental conditions in one life domain (home, commuting, work, or recreation) transfer to another, either positively or negatively (Novaco et al., 1990, 1991). An example of a positive transfer effect is that a high level of residential choice buffers commuting stress. Examples of negative transfer effects are that exposure to commuting stress affects negative mood at home in the evening (controlling for a great many other relevant variables) and that commuting stress prospectively leads to job location change, independent of job satisfaction. Automobile commuters have higher stress than train commuters having the same potential commute, and effort and unpredictability seem to be key mediating factors (Wener and Evans, 2011).

Dissatisfaction with the commute is a cost the employee absorbs, often with delayed effects on health and family life, but commuting stress also affects work organizations. People will change jobs to cope with traffic congestion, and commuting stress is more often manifested in illness-related absences from work and with impaired work performance. Working from home has possibilities for stress reduction, but telework has the tradeoff of constraining the restorative effects of being at home (Hartig et al., 2007).

Important Future Topics for Transportation Psychology

Transportation mobility restrictions present important psychological issues for a number of special populations, such as single working mothers with children, people who are physically or developmentally disabled, young drivers, and the elderly. Perhaps the most important problem area will be the needs and functioning of older adults, because of the growth in that sector of the population. In the United States in 1900, people aged 65 and more accounted for 4% of the population. By 2010, this percentage had grown to 13.0%, and it is estimated by the US Census Bureau that by 2030 it will be 19.3%.

Transportation problems for the elderly are exacerbated by suburbanization and by the vast majority of older adults being accustomed to driving automobiles. Decrements in visual, auditory, cognitive, and psychomotor skills that progressively occur with aging present societal challenges for scrutinizing the driving proficiency of older adults – for whom the loss of the driving license constitutes a major stressful life event. Without transit alternatives, older people become a transportation-deprived group. Many older adults progressively adjust to the decline in driving skills by ceasing to drive at night or avoiding turns across oncoming traffic; but such adaptations do not suffice for long, leaving many elderly unable to be licensed to drive. Physicians must remain aware of older driver safety issues in their clinical practice.

Projected technologies, such as automated roadways, are expected to ease transportation difficulties for the general population, but if automotive vehicles become significantly more cognitively complex to operate, this may be disadvantageous to the elderly. It can be expected that there will be substantial societal pressure to develop innovations in transportation to meet the needs of older adults who cannot drive and who do not have access to normal transit services. Wang and Carr (2004) called for efforts to optimize the driver, the driving environment, and the vehicle. The development of flexible and perhaps individualized transportation services for the elderly to facilitate grocery shopping, receiving health care, and going to religious, recreation, and cultural events is an important societal agenda that beckons for psychological input.

Advances in telecommunications will continue to affect commuting mode choice, and intelligent transportation systems (including roadway incident management) will affect travel times and travel safety. Governments will strive to get maximum capacity out of existing systems, but associated with government-instituted practices that flow from that motivation are personal privacy considerations. Surveillance, vehicle monitoring, event data recording, and vehicle control infringe on personal freedom, and psychological issues abound in that regard, including legislative matters.

Adaptation: An Over-Arching Concept

A core concept in the psychology of transportation is behavioral adaptation. Drivers and transport users will adjust to conditions in the transportation environment whether that be in roadway or transit conditions, their commuting scheme, their vehicle attributes, or changes in their skill level or state of functioning. People make adaptations to enhance their well-being, but some adaptations go awry, such as when enhanced vehicle design induces one to underestimate collision risk or use of nomadic devices while driving promotes multitasking

that makes for driver distraction. Adaptation occurs at different levels of analysis, ranging from 'high-level' trip and travel mode decisions to 'low-level' vehicle control behaviors that maintain safety margins. When the behavioral adjustments involved in adaptation are unsuccessful in the long term, more substantial behavioral coping can be expected, as when someone changes jobs to reduce commuting strain or stops driving because of age-related driving skill decrements. Innovations in information technology, by decreasing need for vehicle trips, progressively may facilitate positive adaptations to travel constraints imposed by congested roadways and to unwanted reductions in mobility for some populations.

See also: Aggression, Social Psychology of; Forensic Psychology; Industrial–Organizational Psychology: Science and Practice; Personality Assessment; Stress: Psychological Perspectives; Wellbeing and Burnout in the Workplace, Psychology of; Work Motivation.

Bibliography

Altman, I., Wohlwill, J.F., Everett, P.B., 1981. Transportation and Behavior. Plenum Press, New York.

Armstrong, K., Filtness, A.J., Watling, C.N., Barraclough, P., 2013. Efficacy of proxy definitions for identification of fatigue/sleep-related crashes: an Australian evaluation. Transportation Research Part F 21, 242–252.

Barjonet, P., 1997. Transport psychology in Europe: a general overview. In: Rothengatter, T., Vaya, E.C. (Eds.), Traffic and Transport Psychology: Theory and Application. Pergamon, New York, pp. 21–30.

Brewer, A., 2000. Road rage: what, who, when, where, and how? Transport Reviews 20, 49–64.

Chen, C.-F., Chao, W.-H., 2011. Habitual or reasoned? Using the theory of planned behavior, technology acceptance model, and habit to examine switching intentions toward public transit. Transportation Research Part F 14, 128–137.

Deffenbacher, J.L., Lynch, R.S., Oetting, E.R., Swaim, R.C., 2002. The driving anger expression inventory: a measure of how people express their anger on the road. Behaviour Research and Therapy 40, 717–737.

Delfino, R.J., Tjoa, T., Gillen, D.L., Staimer, N., Polidori, A., Arhami, M., Jamner, L., Sioutas, C., Longhurst, J., 2010. Traffic-related air pollution and blood pressure in elderly subjects with coronary artery disease. Epidemiology 21, 396–404.

Devarasetty, P.C., Burris, M., Arthur Jr., W., McDonald, J., Munoz, G.J., 2014. Can psychological variable help predict the use of priced managed lanes? Transportation Research Part F 22, 25–38.

European Commission, 2006. Energy and Transport: Report 2000–2004. Directorate General for Energy and Transport.

Frumkin, H., 2002. Urban sprawl and public health. Public Health Reports 17, 201–217.

Galovski, T.E., Blanchard, E.B., 2004. Road rage: a domain for psychological intervention? Aggression and Violent Behavior 9, 105–127.

Gossling, S., 2013. Advancing a clinical transport psychology. Transportation Research Part F 19, 11–21.

Grieco, M., Pickup, L., Whipp, R., 1989. Gender, Transport, and Employment: The Impact of Travel Constraints. Avebury, Aldershot.

Groeger, J.A., Rothengatter, J.A., 1998. Traffic psychology and behaviour. Transportation Research Part F 1, 1–9.

Hartig, T., Kylin, C., Johansson, G., 2007. The telework tradeoff: stress mitigation vs. constrained restoration. Applied Psychology: An International Review 56, 231–253.

Jones, G.W., 2002. Southeast Asian urbanization and the growth of mega-urban regions. Journal of Population Research 19, 119–136.

Jones, P., Hervik, A., 1992. Restraining car traffic in European cities: an emerging role for road pricing. Transportation Research-A 26A, 133–145.

Koslowsky, M., Kluger, A.V., Reich, M., 1995. Commuting Stress: Causes, Effects, and Methods of Coping. Plenum Press, New York.

Lukas, J.L., Heady, R.B., 2002. Flextime commuters and their driver stress, feelings of time urgency, and commute satisfaction. Journal of Business and Psychology 16, 565–571.

Mayou, R.A., Ehlers, A., Bryant, B., 2002. Posttraumatic stress disorder after motor vehicle accidents: 3-year follow-up of a prospective longitudinal study. Behaviour Research and Therapy 40, 665–675.

Mokhtarian, P.L., 1990. A typology of relationships between telecommunications and transportation. Transportation Research-a 24A, 231–242.

Novaco, R.W., 1991. Aggression on roadways. In: Baenninger, R. (Ed.), Targets of Violence and Aggression. Elsevier, North-Holland.

Novaco, R.W., Gonzalez, O., 2009. Commuting and well-being. In: Amichai-Hamburger, Y. (Ed.), Technology and Well-being. Cambridge University Press, pp. 174–205.

Novaco, R.W., Kliewer, W., Broquet, A., 1991. Home environment consequences of commute travel impedance. American Journal of Community Psychology 19, 881–909.

Novaco, R.W., Stokols, D., Milanesi, L., 1990. Objective and subjective dimensions of commuting stress. American Journal of Community Psychology 18, 231–257.

Organization for Economic Co-Operation and Development, 2013. How's Life? 2013: Measuring Well-being. OECD Publishing. http://dx.doi.org/10.1787/9789264201392-en.

Pisarski, A.E., 1987. Commuting in America: National Report on Commuting Patterns and Trends. ENO Foundation for Transportation, Westport, CT.

Pisarski, A.E., 2006. Commuting in America III: The Third National Report on Commuting Patterns and Trends. Transportation Research Board, Washington, DC.

Pisarski, A.E., Polzin, S.E., 2013. Commuting in America: The National Report on Commuting Patterns and Trends. American Association of State Highway and Transportation Officials.

Rowe, R., Andrews, E., Harris, P., 2013. Measuring risky-driving propensity in pre-drivers: the violation-willingness scale. Transportation Research Part F 19, 1–10.

Rosenbloom, S., 1992. Why working families need a car. In: Wachs, M., Crawford, M. (Eds.), The Car and the City: The Automobile, the Built Environment, and Daily Urban Life. University of Michigan Press, Ann Arbor, MI.

Rothengatter, T., Vaya, E.C., 1997. Traffic and Transport Psychology: Theory and Application. Pergamon, New York.

Shinar, D., 1998. Aggressive driving: the contribution of the drivers and the situation. Transportation Research Part F 1, 137–160.

Shinar, D., Compton, R., 2004. Aggressive driving: an observational study of driver, vehicle, and situational variables. Accident Analysis and Prevention 36, 429–437.

Siren, A., Kjaer, M.R., 2011. How is the older road users' perception of risk constructed? Transportation Research Part F 14, 222–228.

Stokols, D., Novaco, R.W., 1981. Transportation and well-being. In: Altman, I., Wohlwill, J.F., Everett, P.B. (Eds.), Transportation and Behavior. Plenum, New York, pp. 85–130.

Sullman, M.J.M., Stephens, A.N., 2013. A comparison of the driving anger scale and the propensity for angry driving scale. Accident Analysis and Prevention 58, 88–96.

United States Census Bureau, 2011. Commuting in the United States: 2009. American Community Survey Reports, U.S. Department of Commerce.

Wachs, M., Crawford, M., 1992. The Car and the City: The Automobile, the Built Environment, and Daily Urban Life. University of Michigan Press, Ann Arbor, MI.

Wahlberg, A.E., Dorn, L., Kline, T., 2010. The effect of social desirability on self-reported and recorded road traffic accidents. Transportation Research Part F 13, 106–114.

Wang, C.C., Carr, D.B., 2004. Older driver safety: a report from the older drivers project. Journal of the American Geriatrics Society 52, 143–149.

Wener, R.E., Evans, G.W., 2011. Comparing stress of car and train commuters. Transportation Research Part F 14, 111–116.

Wiesenthal, N., Vingilis, E., 2013. The impact of motor vehicle injury on distress: moderators and trajectories over time. Transportation Research Part F 21, 1–3.

Transportation, Supply, and Congestion

Robin Lindsey, Sauder School of Business, University of British Columbia, Vancouver, BC, Canada
André de Palma, Ecole Normale Supérieure de Cachan, Cachan, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Transportation facilities are prone to congestion because of peaks in travel demand, inflexibility of supply, and inability to store transport services. Urban road traffic congestion and congestion in air travel impose significant economic and social costs. Traditional policies to alleviate congestion such as building new roads have had limited success at best. Congestion pricing has the greatest theoretical potential because it can affect all dimensions of travel behavior. However, congestion pricing of roads and airport runway slots has been resisted by policy makers.

Introduction

Technological advances in transportation since the industrial revolution have spurred the growth and spread of cities, enabled people to live far from where they work, and made long-distance business and recreational travel commonplace. Yet growth in travel has been accompanied by increasing adverse side effects: noise, air pollution, accidents, and – most importantly – congestion: the subject of this article. Congestion is certainly not a recent phenomenon. Two thousand years ago chariot riding was banned in Rome during peak hours because of traffic jams (Gibbs, 1997). But the automobile and other motorized forms of transportation have increased the prevalence of congestion in surface travel, and congestion in air travel has become a serious problem.

The most widespread and costly form of congestion is travel delay. Delay on the roads is manifest as reduced speeds, and queues at intersections and entrances to freeways. Air travelers suffer delay in checking in, passing through security, boarding an aircraft, waiting to take off and land, clearing customs and immigration, and retrieving baggage at the destination. These examples feature delay systems in which users experience delays but eventually complete their trips. Delay systems differ from loss systems, which have a finite capacity to accommodate users and reject further demand once their capacity is reached so that service becomes unavailable. For example, a flight cannot be booked if all seats have been reserved. A given transportation facility can sometimes be a delay system and sometimes a loss system. A bus is a delay system when it stops for passengers and waits for them to board and alight, and a loss system at full occupancy. And an airport is a delay system if arriving aircraft have to queue prior to landing, and a loss system if aircraft get diverted to another airport.

Crowding is a third form of congestion that is commonly experienced on buses, subways, and waiting platforms, in using luggage compartments on trains and aircraft, etc. Delay, unavailability of service, and crowding are problems not only in transportation but also in facilities associated with travel including hotels, restaurants, sports stadiums, recreational areas, and telecommunications. These facilities will, however, not be treated explicitly here.

Why Does Congestion Occur?

Transportation facilities are prone to congestion because of three characteristics of travel demand and supply: demand varies over time, supply is relatively fixed over long time periods, and output is not storable.

Temporal Variation in Demand: Travel demand has pronounced temporal fluctuations. Systematic daily and weekly variations are driven by work and school schedules, and by the operating hours of businesses, shops, and entertainment establishments. Seasonal variations in travel are induced by the timing and length of school holidays and religious festivals, by sporting event schedules, and by the seasonal nature of outdoor recreational activities. Travel demand also fluctuates irregularly due to sporting meets, fairs, and other "special events."

Rigidity of Transportation Supply: Transportation supply consists of infrastructure and vehicles. Infrastructure is generally long-lived and very costly and time-consuming to alter. Vehicles come in smaller and cheaper units than infrastructure, and have shorter lifetimes, but still last much longer than the time scale of demand fluctuations. Differences between regions and traveler groups in the timing of peak demands provide some scope to adjust supply by redeploying vehicles between markets although this is limited by constraints on the mobility of vehicles themselves (little or none for subways and trams).

Output Is Not Storable: A traveler or freight shipment must be present to be transported. Thus, like other services such as haircuts and concerts, but unlike commodities such as grain and steel, production cannot be stored to smooth imbalances between supply and demand.

As a consequence of demand fluctuations, rigid supply, and the impossibility of storage, the utilization rate of transportation facilities varies over time. Capacity is also variable due to accidents, strikes, bad weather, and so on. Because land is scarce and capacity is expensive, it would be prohibitively expensive to build enough capacity to prevent congestion at all times and in all circumstances.

Congestion is exacerbated by two further considerations. First, travelers usually lack incentive to account for the congestion costs they impose on other users. Usage fees are not charged on most roads, and the costs of registration, licensing, and insurance do not vary with distance traveled – let alone traffic

volumes. Access fees are levied at facilities such as waterways and airports, but the fees are usually either fixed or vary with usage crudely at best. Thus, congestion is a negative externality because its adverse side effects are not priced in markets. Second, transportation supply tends to lag growth in population and travel. Air travel volume in passenger-kilometers grew at an average annual rate of 5.7% from 1975 to 2005 (ICAO, 2007). Freight traffic measured in ton-kilometers grew at a rate of 6.9% over the same period. These high growth rates were spurred by airline deregulation in the United States in the late 1970s and subsequently in Canada, Australia, New Zealand, and Europe. Airport planning authorities, meanwhile, failed to invest adequately in runway capacity and air traffic control, or to price usage of airports efficiently.

The Costs of Congestion

Congestion imposes a range of costs on travelers: increased fuel consumption and vehicle wear and tear; extra travel time; the inconvenience of using an alternative travel mode or rescheduling, combining, or canceling trips; disutility from crowding; planning and psychological costs arising from uncertainty about arrival times; and (in the longer run) the costs of relocating jobs and residences. Delays are also costly for freight shippers. And unreliable deliveries can upset just-in-time manufacturing and increase the need for inventories.

Because travel delays are a dominant component of the total cost of congestion it is important to measure accurately the cost of travel time – known as the value of time (VOT). For passenger transportation VOT depends on the mode of travel, trip purpose, time of day, household time pressures, and other individual and trip-specific factors. Small and Verhoef (2007: p. 53) conclude from an extensive literature survey that the VOT is typically between 20 and 90% of the before-tax wage rate with an average about 50%. Estimates of VOT for air travel are generally higher than for driving, in part because many air travelers are on business trips with tight schedules and in part because individuals with high VOT prefer to take fast modes of transport. For freight transportation by truck the VOT depends on many factors: the type of commodity being transported, characteristics of the vehicle such as size and fuel consumption, drivers' wage rates, the importance of punctual delivery, whether the vehicle is operated in-house or for-hire, and so on. Average VOTs for truck trips are much higher than for car trips, and higher still when account is taken of the value of reliability (that is, the cost of unexpected delay).

Various estimates of the economy-wide costs of congestion have been compiled. For traffic congestion in the United States in 2011, Schrank et al. (2012) report a value of $121 billion, which is roughly 0.5% of gross domestic product (GDP). This figure includes time delay for passenger vehicles (valued at $16.79 per person hour) and truck operations (valued at $88.81 per vehicle hour), and extra fuel consumption. Comparable, recent figures for Europe are unavailable. The European UNITE project (Nash et al., 2003) estimated the costs of traffic congestion in the UK, France, and Germany to be 1.5, 1.3, and 0.9 of GDP, respectively. Aggregate statistics on the cost of air traffic congestion are patchy. Morrison and Winston (2008) report estimates for the United States that flight delays in 2005 cost passengers $9.4 billion in lost time and airlines $5.9 billion in additional operating costs. Ball et al. (2010) obtain larger estimates for 2007 of $16.7 billion for passengers, $8.3 billion for airlines, $2.2 billion for people who avoid air travel due to delays, and indirect costs of $4 billion in reduced GDP.

For several reasons aggregate statistics such as these should be interpreted with caution. First, they usually include only the more readily measured costs of congestion. Second, valuing time delay with a single number for VOT ignores differences across travelers as well as possible dependency on trip duration and the severity of the congestion encountered. And third, because eliminating congestion is not a practical option, measuring the total costs of congestion under actual travel conditions is not indicative of what might be gained from policies aimed at congestion relief. (On this point, see also Section Ways to Alleviate Congestion.)

These caveats notwithstanding, it is instructive to compare the costs of congestion with the costs of other transportation externalities. Small and Verhoef (2007: Table 3.3, p. 98) report estimates for the United States. Congestion costs account for about 64% of the total external costs of car transport, followed by accidents (24%) and environmental costs (6%). Delucchi and McCubbin (2011) present a wide range of estimates for all major modes of transport. For roads they find that accidents may impose the highest costs, with congestion costs ranking second. For air transport noise costs are highest, followed by congestion. Friedrich and Quinet (2011) compare these estimates with estimates for Europe and conclude that they are similar in both magnitude and ranking. Overall, the costs of congestion thus appear to be appreciable both in absolute terms and relative to other externalities.

Modeling Congestion in Transportation Facilities

Transportation congestion modeling has been taken furthest for road traffic. Various approaches are described in Daganzo (1997), Small and Verhoef (2007), and Lindsey and Verhoef (2008). Modeling the behavior of individual drivers is a daunting task not only because of their sheer number but also because drivers – being self-interested, fallible, and heterogeneous – do not behave like atoms according to universal and immutable physical laws. Traffic microsimulation models have nevertheless been developed that can simulate the behavior of large numbers of individual vehicles within a predefined road network. These models are well-suited to simulating queuing conditions on urban roads, and analyzing signal timing and other traffic control measures as well as policies such as lane-narrowing. Useful insights into the economics of traffic congestion can nevertheless be obtained by using conventional supply–demand diagrams in which the time dependence of traffic flow is ignored and vehicles are treated as a homogeneous continuum.

Consider **Figure 1**, but ignore all its details for now – which will be covered later. The average cost (AC) curve passing through points *F*, *G*, *J*, and *M* depicts the cost borne by each driver to take a particular trip as a function of the traffic volume; i.e., the number of vehicles that make the trip per hour. Average cost includes vehicle operating cost plus travel

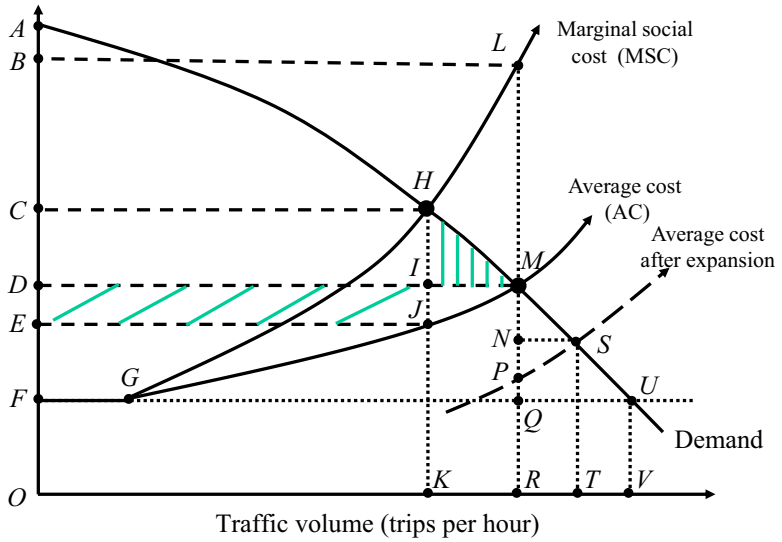

Figure 1 Trip costs, equilibrium, and optimal usage in a delay system.

time, which is assumed to be valued equally by all drivers. At low volumes, vehicles are far enough apart that they can proceed under free-flow conditions, so that the trip cost curve is horizontal between *F* and *G*. At higher volumes, drivers are forced to slow down to maintain a safe following distance. Travel time rises and AC slopes upward. The shape of AC depends on road design, speed limits, vehicle size and handling characteristics, weather conditions, and driver reaction times and preferences for risk-taking. Due to changes in these factors over time, driving speeds have increased and following distances have decreased, thereby boosting effective road capacity and shifting the AC curve down.

The marginal social cost (MSC) curve in **Figure 1**, passing through *F*, *G*, *H*, and *L*, is derived by adding to AC the cost of the congestion that each driver imposes on others. The MSC curve coincides with AC over the interval from *F* to *G*, and lies above AC elsewhere. To determine how many trips are taken, a demand curve can be added as shown by the downward-sloping curve passing through *A*, *H*, *M*, *S*, and *U*. In the absence of traffic control or other means of regulation, equilibrium is established at point *M* where the marginal willingness to pay for a trip as measured by the demand curve matches the cost, AC. The unregulated equilibrium number of trips is *OR*. By contrast, the socially optimal number of trips, *OK*, is found at point *H* where the demand curve intersects MSC. Fewer trips per hour are taken than in the unregulated equilibrium, although travel is still congested because trip cost, *JK*, exceeds free-flow cost, *OF*. One way to support the optimum is to impose a toll equal to *HJ*: the difference between MSC and AC with traffic volume at *OK*. This is the basis behind congestion pricing, to be discussed in Section Congestion Pricing.

At the unregulated equilibrium, the congestion cost imposed by each driver is *LM*. Some studies have estimated this quantity to be several times as big as *MR* during peak periods in large cities. To get a figure for aggregate congestion costs it is tempting simply to multiply *LM* by the number of trips, *OR*, to get area *LMDB*. But this is inappropriate since the number of excessive trips taken in equilibrium relative to the social optimum is only *KR*. Similarly, it does not make sense to evaluate the congestion cost at the optimum, *HJ*, and multiply it by *OK* to get area *HJEC*.

A more practically useful measure of congestion costs is the difference between the net social benefits of travel at the optimum (area *AHJE*) and at the unregulated equilibrium (area *AMD*). This difference can be decomposed into the reduction in costs on the *OK* trips that continue to be taken at the optimum (diagonally shaded area *IJED*) minus the loss of benefits on the *KR* trips that are no longer taken (vertically shaded area *HMI*). The difference in net social benefits can also be determined by computing the net social cost of the *KR* excessive trips taken in equilibrium. This quantity is equal to area *HLM*, which must therefore equal *IJED-HMI*.

Note that the cost of travel under free-flow conditions plays no role in the calculations. One could of course calculate the difference in cost between equilibrium and free-flow conditions, *MQ*, and multiply it by *OR* to get area *MQFD*. But this would be relevant only if it were somehow possible to maintain free-flow conditions at any traffic volume. Traffic volume would then increase from *OR* to *OV*, and a better measure of the gains would be the increase in net social benefits: area *MUFD*.

Supply–demand diagrams such as **Figure 1** are applicable not only to automobile traffic but also to other delay systems such as airport runways, bus service, waterways, and pedestrian walkways. In the case of loss systems, the AC curve is approximately horizontal up to capacity flows and it is therefore optimal to use the facility to capacity while avoiding queuing or other wasteful forms of rationing access. In practice, demand cannot be predicted perfectly and a (probabilistic) tradeoff exists between unused capacity and congestion. Airlines deal with this tradeoff by using sophisticated yield management software that set fares according to the average fraction of seats that are occupied on a given flight.

Though useful, the simple supply–demand approach does abstract from various real-world features of transportation

congestion. First, roads, airports, and other facilities are components of transportation networks. Costs at one facility can depend on usage at other facilities because of conflicts between traffic flows, spillovers of queues, and so on. Such interactions may cause throughput to decline with traffic density and even deteriorate to a state of gridlock. Second, users differ in the congestion they create. Traffic engineers make allowance for differences between road vehicles in size, speed, and maneuverability by using passenger car equivalents that depend on grade and other road characteristics. And third, the static supply–demand approach ignores fluctuations in travel demand over time and does not address trip timing decisions. Vickrey (1969) developed a tractable model of trip-timing decisions which is now widely used. His and other dynamic congestion models are reviewed in de Palma and Fosgerau (2011).

Ways to Alleviate Congestion

Many policies have been adopted to relieve congestion. Some – such as land-use controls – have additional objectives, whereas policies such as road pricing are targeted at congestion. Discussion will be limited here to three policy instruments: capacity investment, congestion pricing, and information technology.

Investment

The traditional approach to congestion relief is to expand capacity. Investment shifts the AC curve in **Figure 1** down, as shown by the dashed line passing through points *P* and *S*. A new unregulated equilibrium is established at *S*. Two benefits result: a reduction in travel costs for the *OR* preexisting trips, and the benefit to travelers from the *RT* new trips which can arise from changes in mode, route, departure time, and other choices. But because the new trips contribute to congestion, travel costs decline by only *MN* rather than by *MP*, and this partly undoes the cost savings to previous users. The generation of new trips is a manifestation of so-called latent (hidden) demand that was not apparent prior to the investment (Duranton and Turner, 2011). The strength of latent demand depends on the geographical scale of the capacity investment, the time period considered, the quality of public transit alternatives, and other factors. Though the increase in usage can be appreciable if the demand curve is relatively flat, latent demand has frequently been overlooked in engineering and cost–benefit studies of investment. Even worse than latent demand, capacity expansion on one link of a transport network can induce shifts of demand to more congestion-prone routes or travel modes that actually raise overall travel costs (Arnott and Small, 1994).

Unlike the unregulated equilibrium, investment always yields a positive gross benefit for the social optimum. The effects can be assessed by tracing the rightward shift of the MSC curve (not shown in **Figure 1**). Depending on the strength of latent demand, the benefits may be greater or smaller than for the unregulated equilibrium. More generally, the benefits of investment, and hence optimal capacity too, depend on how traffic is regulated and priced.

Congestion Pricing

Congestion pricing has been promoted by economists since the 1950s. It makes use of the price mechanism, with all its advantages of decentralized decision-making. And, unlike other policies, congestion pricing restrains latent demand and generates revenue that can be used to fund infrastructure investment and maintenance. The biggest potential gains from congestion pricing in transportation appear to be for road traffic and airport runways.

Road Pricing

Road pricing technology has come a long way since drivers were obliged to stop and pay cash at toll booths (de Palma and Lindsey, 2011). Thanks to advances in automatic vehicle identification and electronic toll collection, tolls can be levied anonymously using prepaid smartcards while vehicles maintain highway speeds. Several European countries now use satellite technology for tolling trucks on intercity motorways. The biggest virtue of road pricing is that it can influence all dimensions of travelers' choice. It is also consistent with the user pay principle.

The optimal toll in the static supply–demand framework of **Figure 1** is *HJ*. The imposition of this toll raises travel cost by *HI* relative to the no-toll unregulated equilibrium. That road pricing tends to make drivers worse off is one of the reasons for longstanding opposition to it. Another is that road pricing is often seen as 'just another tax.' There is growing recognition that to generate political support for road pricing, toll revenues must be used to benefit travelers without undoing incentives for them to curb their travel. Possibilities include improving public transit, reducing gasoline taxes, and investing in roads themselves. New self-financed private toll roads are showing promise as a market-oriented approach to alleviating congestion (Small, 2010).

The simple supply–demand approach of course abstracts from various real-world complications in the implementation of congestion pricing, including setting tolls on a network, varying tolls over time, charging for parking, and avoiding harm to lower income drivers. These obstacles are largely responsible for the slow implementation of congestion pricing. The main successes are electronic road pricing in Singapore, the London congestion charge, the Stockholm cordon charge, and high occupancy toll (HOT) lanes in the United States. None of these schemes adheres rigorously to theoretical prescriptions for congestion pricing. Nevertheless, all except London employ time-varying tolls, and on some HOT lane facilities tolls are adjusted every few minutes to maintain high speeds on the toll lanes. Road pricing practice is reviewed in Small and Verhoef (2007: Section 4.3), Parry (2009), and Santos and Verhoef (2011).

Pricing Airport Runways

Traditional means of allocating runway space are a far cry from congestion pricing. Access to runways has been determined by scheduling committees that may be dominated by representatives of major airlines operating at the airport. And landing fees have been based on aircraft weight, which reflects congestion cost only crudely. Congestion pricing of runways would be administratively easier to implement than road pricing because

takeoff and landings are recorded at airports anyway, and since the transactions costs for a flight are spread over many travelers.

The virtues of congestion pricing of runways were recognized in the late 1960s. Pricing encourages airlines to schedule flights off peak, to use lower cost airports, and to deploy larger aircraft which create less delay per passenger. One complication is that some airlines operate a large fraction of the flights at an airport and have an incentive to internalize the congestion costs their flights impose on each other. Brueckner (2002) shows that the theoretically optimal congestion charge varies inversely with an airline's market power so that optimal tolling would be discriminatory with large carriers charged less than small ones. A number of theoretical and empirical studies have assessed the practical significance of this 'self-internalization' hypothesis with no consensus yet reached. Morrison and Winston (2007) estimate that optimal discriminatory tolling at US commercial airports would yield an annual benefit of $2.7 billion. Nondiscriminatory charges yield only slightly lower benefits of $2.5 billion which suggests that discrimination might not be worthwhile given the higher administration costs and likely political opposition. An alternative to charging for every use of runways is to auction tradable take-off and landing slots. Slot trading was introduced in 1986 at Washington National, Chicago O'Hare, and New York's La Guardia and Kennedy airports, and has worked well. Airport congestion pricing and slot trading are described in Zhang and Czerny (2012).

Information Technology

Development of Intelligent Transportation Systems (ITS) technology began in the 1970s. These systems are now widely used to monitor traffic and to respond quickly to accidents by dispatching emergency services. ITS technology in the form of Advanced Traveler Information Systems is also used to inform travelers about travel conditions. Information is made available by traditional sources such as television, radio, and changeable message signs, as well as via traffic web sites, and it is accessible using cellular phones, smart phones, other handheld devices, and in-vehicle computers and global positioning systems receivers. Travelers can receive traffic alerts by short message service, automated voice call, and e-mail. They can use the information to adjust their travel mode, departure time, route, and other decisions, and they can use it as a navigational aid for way-finding and locating services.

Most studies conclude that ITS technology helps to alleviate traffic congestion and reduce accidents, pollution and carbon-dioxide emissions. Still, the potential of ITS is limited by the fact that drivers can disregard advice if they believe it is not in their self interest. And, similar to the effects of capacity expansion, better information can induce people to travel more because it reduces the costs of travel. This shortcoming can in principle be overcome by implementing road pricing in tandem with ITS.

General Policy Conclusions

From a theoretical perspective congestion pricing is the most effective means of controlling congestion because it influences all travel choices and also contains latent demand. But for various reasons – including the cost of pricing all links on travel networks, equity concerns, and the need to maintain safety – nonprice constraints on travel such as air traffic control and regulations on the movement of heavy road vehicles will persist.

Looking Ahead

Despite the hazards of forecasting it is tempting to speculate about how congestion in transportation will evolve in the years to come. Vehicle ownership and usage is likely to continue growing in China and other developing countries for many years, and cities such as Beijing, Shanghai, and Bangkok are now among the most congested in the world. Until 2006, aggregate congestion indexes suggested that traffic congestion was also getting worse in the United States (Schrank et al., 2012). But congestion has since eased in the United States and most other major developed countries. Evidence is growing that personal traveled has reached a plateau and even started to decline (Millard-Ball and Schipper, 2011). Among the possible reasons are economic stagnation, rising fuel prices, aging of the population, continuing urbanization, and mounting health and environmental concerns. Young adults are postponing purchase of vehicles as well as homes that encourage use of personal vehicles. People can adapt to congestion by adjusting departure times, destinations, residential location, and other choices. Congestion costs are also held in check by the trend toward flexible work times, late night, and weekend shopping, and extended hours of other services. And, most fundamentally, the costs of congestion are borne largely by travelers themselves. Congestion is a self-limiting phenomenon in the sense that the private cost of travel can never exceed the private benefit derived from it.

The effects of current and possible future technological developments on congestion are ambiguous. Innovations such as air conditioning, car stereo systems, books on tape, and traveler information systems mentioned above have made driving more pleasant and allow people to make productive use of time in the car. Similarly, laptops, cellular phones, and executive lounges at airports make delays in air travel less onerous. By reducing the costs of travel these developments might be expected to encourage travel and increase congestion delays.

Technologies such as telecommuting, teleconferencing, teleshopping, Fax, electronic mail, the Internet, and other telecommunications devices offer alternatives to personal and business travel as well as postal mail and courier services. But telecommunications also complement travel by helping people establish new contacts, and learn about new products and places to visit (Mokhtarian, 2002).

And the option of telecommuting may encourage people to live in more remote locations, so that when they do commute they have further to go. Part of the time freed up from commuting may be used to make other trips, and any reductions in travel by telecommuters may be partially offset by latent demand from the rest of the population.

Rapid advances are being made in vehicle design. Electric vehicles promise to deliver major environmental benefits, but their effects on traffic congestion are likely to be more modest.

In the longer term, automated highways, self-driving vehicles, and vehicle-to-vehicle communications could provide huge benefits in congestion relief, fuel savings, and safety. They may also reduce the amount of space required for roads and parking (The Economist, 2012). Nevertheless, these technologies will take many years to phase in, and they pose major technological and legal challenges. In the meantime, the battle against traffic snarls as well as airport delays and congestion on other transport modes will continue to be waged on several fronts.

See also: Automobility; Communication and Transportation, History of; Cost–Benefit Analysis; Infrastructure Investment; Mobility and Travel Activity Patterns; Public Choice Economics; Sustainable Transportation; Transportation Geography; Transportation Planning; Urban Geography; Urban Sprawl; Urban System in Geography.

Bibliography

Arnott, R., Small, K.A., September–October 1994. The economics of traffic congestion. American Scientist 82, 446–455.

Ball, M., Barnhart, C., Dresner, M., Neels, K., Odoni, A., Perterson, E., Sherry, L., Trani, A., Zou, B., 2010. Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States. NEXTOR. http://www.isr.umd.edu.

Brueckner, J.K., 2002. Airport congestion when carriers have market power. American Economic Review 92 (5), 1357–1375.

Daganzo, C.F., 1997. Fundamentals of Transportation and Traffic Operations. Elsevier Science, New York.

Delucchi, M., McCubbin, D., 2011. External costs of transport in the U.S. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK, pp. 341–368.

Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: evidence from US cities. American Economic Review 101 (6), 2616–2654.

Friedrich, R., Quinet, E., 2011. External costs of transport in Europe. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK, pp. 369–395.

Gibbs, W.W., October 1997. Transportation's perennial problems. Scientific American, 54–57.

ICAO, 2007. Outlook for Air Transport to the Year 2025. Cir 313, AT/134, September. http://www.icao.int.

Lindsey, R., Verhoef, E.T., 2008. Congestion modelling. In: Hensher, D.A., Button, K.J. (Eds.), Handbook of Transport Modelling, vol. 2. Elsevier Science, Oxford, pp. 417–441.

Millard-Ball, A., Schipper, L., 2011. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. Transport Reviews 31 (3), 357–378.

Mokhtarian, P.L., 2002. Telecommunications and travel: the case for complementarity. Journal of Industrial Ecology 6 (2), 43–57.

Morrison, S., Winston, C., 2007. Another look at airport congestion pricing. American Economic Review 97 (5), 1970–1977.

Morrison, S., Winston, C., 2008. Delayed! U.S. aviation infrastructure policy at a crossroads. In: Winston, C., de Rus, G. (Eds.), Aviation Infrastructure Performance: A Study in Comparative Political Economy. Brookings Institution Press, Washington, pp. 7–35. http://www.brookings.edu.

Nash, C., et al., 2003. Project UNITE (UNIfication of Accounts and Marginal Costs for Transport Efficiency). Final Technical Report. Fifth Framework Competitive and Sustainable Growth (Growth) Programme, Commissioned by European Commission. DG TREN. http://www.its.leeds.ac.uk.

Parry, I.W.H., 2009. Pricing urban congestion. Annual Review of Resource Economics 1 (1), 461–484.

de Palma, A., Fosgerau, M., 2011. Dynamic traffic modeling. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK, pp. 188–212.

de Palma, A., Lindsey, R., 2011. Traffic congestion pricing methodologies and technologies. Transportation Research Part C 19 (6), 1377–1399.

Santos, G., Verhoef, E.T., 2011. Road congestion pricing. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK, pp. 561–585.

Schrank, D., Eisele, B., Lomax, T., September 2012. The 2012 Urban Mobility Report Powered by INRIX Traffic Data. Texas Transportation Institute, Texas A&M University, College Station, Texas. http://tti.tamu.edu.

Small, K.A., 2010. Private provision of highways: economic issues. Transport Reviews 30 (1), 11–31.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation, second ed. Routledge, London and New York.

The Economist, 1 September 2012. Look, no hands. The Economist Technology Quarterly Q3. http://www.economist.com/node/21560989.

Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review 59 (2), 251–260 (Papers and Proceedings).

Zhang, A., Czerny, A.I., 2012. Airports and airlines economics and policy: an interpretive review of recent research. Economics of Transportation 1, 15–34.

Relevant Websites

http://www.icao.int – International Civil Aviation Organization.
http://www.iteaweb.org/ – International Transportation Economics Association.
http://www.internationaltransportforum.org/ – International Transportation Forum.
http://www.dot.gov – US Department of Transportation.
http://www.trb.org – US Transportation Research Board.
http://www.vtpi.org – Victoria Transport Policy Institute.

Trials

Malcolm M Feeley, School of Law, University of California, Berkeley, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article identifies distinctive features of various forms of trial in different legal systems, and then explores the social functions and the operational details of trials in Western societies. In doing so, it raises perhaps the most persistent question social scientists ask about trials: Which is the best form for resolving disputes?

Introduction

It is tempting to adopt the definition of trials offered by Lon Fuller, the leading authority on the concept adjudication, who defines the trial as a process, which permits disputants "an opportunity for the presentation of proofs and reasoned arguments" in order to give "formal and institutional expression to the influence of reasoned argument in human affairs." As such, trials "assume a burden of rationality not borne by any other form of social ordering" (Fuller, 1981, pp. 93–94). Although this description probably captures cultural assumptions about the contemporary Western trial, it fails to encompass the idea of a trial more generally. It treats human rationality as a given rather than a variable, and it assumes that the primary focus of trials is to resolve disputes, rather than to clarify policy or to assert authority.

I suggest an alternative: a trial is a process by which proof is authoritatively established. Rationality can be but is not necessarily involved in the process, parties affected may or may not have an opportunity to offer proofs or to make reasoned arguments, and there may or may not be a 'dispute.' This more general definition is useful in accounting for a greater variety of processes that are generally understood to be trials, although in widely different settings, including so-called simple or primitive societies, intensely religious societies, and modern Western societies. It also includes processes in which there are no 'disputes,' and where disputes are not 'resolved' so much as they are processed (*see* Courts and Adjudication).

A good place to start in thinking about this more generic definition of trial is Max Weber's analysis of rational and irrational forms of legal decision-making (Weber, 1954; Trubek, 1972). For Weber, forms of legal decision-making vary along two dimensions: the nature of their rationality and their degree of formality. By rationality, Weber meant the presence or absence of 'general rules' that are to be invoked in decision-making. A rational legal system (and decision), according to Weber, involves use of abstract, calculable, and general principles that are invoked when deciding a particular case. In contrast, an 'irrational' system does not rely on the application of general rules to particular facts or cases, but is guided by a holistic process that may not be determinable in advance of the particular decision. Similarly, Weber pointed out, legal decision-making varies in its degree of formality. At one end of the continuum is a formal legal system that is autonomous; it has a distinct set of 'legal' norms and procedures to guide legal decision-making, and perhaps also a distinct set of offices and officials to make decisions. At the other end of the continuum are not casual, informal norms but rather an array of substantive norms, a mixture of legal, religious, moral, and practical precepts that are intermixed and undifferentiated, and so do not constitute a separate and distinct set of (formal) legal norms. In this way, Weber distinguished formal justice from substantive justice. In proceedings dispensing substantive justice, religious, traditional, societal, ethnic, and family norms may blend with distinctively legal norms; here, strictly legal decision-making and decision-makers are not always distinguishable from other realms of social behavior and social actors (*see* Weber, Max (1864–1920)).

Legal decision-making can vary independently along both dimensions, so Weber developed a set of ideal types that has often been used to characterize distinct forms of legal decision-making across time and across cultures. They are presented in Table 1.

Table 1 generates four identifiable types of 'trials': (1) informal irrational (substantive irrationality), (2) informal rational (substantive rationality), (3) formal irrational, and (4) formal rational. The two extremes on the diagonal (cells 1 and 4) are familiar to many students of law and the social sciences. Readers of Weber are familiar with the example for cell 1 – the idea of decision by a *khadi* (an Islamic clerical judge), who after hearing a case hands down a ruling without explicit reference to any general rules. Weber believed that the outcome is understood in Islamic culture and elsewhere as an expression of divine inspiration transmitted through the *khadi*, and that the ruling applies only to the particular case in hand. By contrast, cell 4 represents the trial in the modern, rule-of-law

Table 1 Weberian typology of legal decision-making

Degree of formality	*Degree of rationality (presence or absence of general rules)*	
	Irrational	*Rational*
	I	II
Informality (substantive justice)	Substantive irrationality Khadi justice	Substantive rationality Cost–benefit analysis
	III	IV
Formality (formal justice)	Formal irrationality Trial by ordeal	Formal rationality modern Western trial

Adapted from Inverarity, J., Lauderdale, P., Feld, B., 1983. Law and Society: Sociological Perspectives on Criminal Law. Little Brown, Boston.

state of the sort Fuller (1981) had in mind. Facts are introduced and testimony is taken according to a clear and precise set of procedures, and this evidence is judged in light of predetermined rules. It is this type of decision-making that Fuller had in mind in his discussion of the forms and limits of adjudication. Turning to Weber's two other types of decision-making, cell 2 represents substantive rationality, a form of decision-making that does not entail a distinct and separate legal sphere but melds law with norms in other social realms (or, more precisely, separate realms are not differentiated). Examples include legal decisions based upon cost–benefit analysis or religious precepts. Such decisions are 'substantive' not 'formal,' since they do not derive legitimacy from a distinct and separate legal sphere. Yet they are 'rational' in that they specify clear norms, rules, and procedures. Cell 3 describes a process that is formal in that it depends upon specialized and distinct legal procedures. But because decisions flow from an ineffable process rather than from general rules, the process is irrational. A familiar example is the establishment of proof through appeals to divine intervention – ordeal by fire, ordeal by water, or other attempts to invoke supernatural signs. Whatever its defects, this last set of procedures for establishing proof had one virtue, if only for true believers: certainty of outcome (*see* Law and Development).

The discussion above suggests that what Fuller had in mind as the essence of the trial is in fact only one of several forms of legal decision-making. However, he is not so far from the mark. Weber's purpose in developing the typology outlined above was not only to emphasize structural variability in legal decision-making, but also to set the stage for his theory that modernization requires the adoption of formal legal rationality. He saw formal rational adjudication, like bureaucratization, as one of the essential features and defining characteristics of modernity. It is therefore important to understand what is lost by abandoning the other three forms of decision-making in his typology, and what is gained by adoption of the fourth form. According to Weber, modern law undermines competing forms of authority. To the extent that modern law triumphs – and law is notoriously good at triumphing – legal precedent, legal decision-making, and hence the modern trial overshadow tradition, custom, the family, and religion as sources of authority. In this sense it simplifies and rationalizes autonomous legal authority. The modern idea of the rule of law also aspires to precision; it identifies legal entities and separates them from other forms of control. Thus, it offers both a sense of freedom and predictability at the same time. In a modern system the ruling at a trial trumps other commitments, relationships, and rule systems. That is, the modern trial privileges the state: as officials of the state, modern judges are part of the state-made legal system, and must apply only state-made law. Modern trials undermine other legal systems and challenge the idea of legal pluralism. However, there is a continuing debate about how autonomous modern law is or must be (Nonet and Selznick, 1978).

The Multiple Functions of the Trial

Privileging state law and trials, as formal rational forms of decision-making do, does not mean that there are no differences in the nature and function of trials in modern states. Forms and functions of trials vary widely in modern societies. Damaska (1986) argues that states with traditions of strong government rely on trials not only as means to handle disputes between the parties, but also as opportunities to reassert and to implement state policies. In contrast, weak states tend to view trials in conflict-resolution terms. European countries with traditions of strong central government and civil law systems are examples of the former, while Great Britain and the United States, less-centralized states and common law systems, are examples of the latter. Predictably, settlements in lieu of trial are formally discouraged in the former set of countries and encouraged in the latter. This doubtless helps to explain why common law countries have a much higher proportion of cases commenced in courts are eventually settled short of trial, relative to most countries with civil law systems (Kritzer, 2004). Political scientist Martin Shapiro, who has explored some of the same themes developed by Damaska (Shapiro, 1981), maintains that always and everywhere, courts – including trial courts – in addition to being institutions for conflict resolution, are important agents of social control. He means that in all societies courts are used by the dominant political regime to help promote its agenda. Trials, therefore, always serve two simultaneous and symbiotic functions: they resolve conflicts among the immediate parties, and they provide an opportunity for a type of micromanagement, in the pursuit of regime interests. When trial courts strive to assert independence and do not or will not serve the latter function, Shapiro argues, their jurisdiction will be curtailed. For example in the mid-twentieth century, when English courts failed to rule in favor of labor unions frequently enough at trial, their jurisdiction was curtailed and special labor relations courts, more favorable to labor, were established.

In a more generalized analysis, Durkheim's ideas of social solidarity rest in part upon his understanding of the social functions served by the legal process, including trials (Durkheim, 1964). Although his writings focus on punishment and the criminal process, his overall theory emphasizes the solidifying function of a variety of legal institutions, including the trial. Much of the dramaturgy associated with the process of punishment is also found in the rituals of the trial. Trials, and public reactions to them, represent an important and visible expression and reassertion of Durkheim's 'conscience collective.' Labeling theorists in the social sciences have seized upon this symbolic function, and have explored its relevance for the criminal trial. In a classic article, Garfinkel described the criminal trial as possessing all the necessary 'conditions of a successful degradation ceremony' (Garfinkel, 1956). Similarly, Stuart Scheingold (1977) and others have argued that trials, and particularly high-profile trials, function to foster a 'myth of rights,' which serves to undermine disruptive political struggles. Symbolic victories at trial give the appearance – but not necessarily the reality – of change, and as a consequence they foster political quiescence (Scheingold, 1977). Others have explored the more overtly political functions of some trials (*see* Judicial Review; Political Trials; Political Lawyering).

Clearly one of the basic functions of trials is to resolve conflicts. Yet, some social scientists question even this seemingly obvious notion. Unlike mediation, where the parties

participate by mutual consent, a trial is a process in which at least one party is an unwilling participant. Consider, for example, the criminal defendant whose appearance must usually be coerced. Moreover, modern law does not necessarily express the deeply valued norms of any tightly knit community or even widely held norms. As a consequence, both the institutional framework and the norms applied may be regarded with suspicion, if not hostility, by one or more of the participants at trial (Shapiro, 1981). Various scholars of legal pluralism emphasize just how problematic these challenges are, and often criticize governments and other scholars for giving pride of place to the formal legal trial as the preferred institution for dealing with conflicts (*see* Legal Pluralism). In light of these and related concerns, it is conceivable that a verdict by a court may not always be an effective way to resolve conflicts, especially in the eyes of the losing party. Consequently, many people who lose at trial continue to press their claims in other forums (Abel, 1982, 1983). Others do not even bother to seek redress in courts (Felstiner et al., 1980–91) or seek out formal alternatives to adjudication (*see* Disputes, Social Construction and Transformation of; Law, Mobilization of).

Trial Complexity and Its Consequences

One distinctive feature of modern trials is the replacement of reliance on divine intervention with modern empirically based decision-making. Factual evidence subjected to scrutiny has replaced the evidentiary value of supernatural signs. Eyewitness testimony has superseded oath-taking by the accused. And gathering, sifting, and pondering evidence has been transformed from an informal process into a disciplined arrangement governed by detailed procedures and formal rules. Consider that the earliest-known juries convened in criminal cases in Anglo-Saxon England consisted of men from the local community who had firsthand knowledge about the case and/or the character of the accused. By the nineteenth century, a jury was a group legally barred from any firsthand knowledge of the facts of the case before it, and could receive only evidence screened by highly complex rules and introduced according to complicated procedures. In most ordinary cases, the medieval trial was a relatively informal and brief affair handled by the parties involved in the dispute. But the modern trial is dominated by lawyers – highly trained and paid professional advocates for the parties to the dispute – who must contend with the proliferating subtleties of modern substantive law, and who must practice according to increasingly complex rules of evidence and procedure.

Differences in the structure of the trial process produce a number of important differences in effects. In systems with jury trials, proceedings are likely to rely heavily on oral evidence. But in European and other legal systems that do not have or do not rely so heavily on juries, the trial often proceeds by written submissions to the court. The court confronts a dossier containing statements of facts and conclusions and sworn testimony (depositions), not a series of witnesses telling their stories in open court. Additionally, in systems with juries, the trial once begun is likely to run continuously until each side has rested its case. By contrast, a trial by dossier is likely to be intermittent and extend over a much longer period of time. These differences also affect the pace of true closure to the litigation. Cases in courts with active juries are almost always concluded within a few days of commencement, and verdicts are typically handed down very shortly thereafter. But in systems without juries, with trials by dossier, evidence-gathering is far more protracted, and final judgments are not necessarily issued shortly after submission of the final bit of evidence. Thus, ironically, although the jury-trial system is often criticized as inefficient, it may actually foster expeditious trial proceedings (*see* Juries).

The nature and form of different trial systems have other consequences as well. The conventional understanding is that Anglo-American legal systems have developed elaborate rules of evidence and procedure in order to restrict the power of jurors, who are untrained lay decision-makers. Since civil law systems have had no such need, they have simpler rules. Although these claims have been vigorously disputed, there is no doubt that rules of evidence and procedure in civil law systems with no juries or weak juries are far simpler than those in common law systems with predominant juries. The 'quest for perfect justice' invoked to justify the complexity of common law trial proceedings may be counterproductive (Pizzi, 1999). Trials are swifter, cheaper, and perhaps fairer in many civil law countries than they are in the United States and England. Feeley and Lester (1994) have traced increases in length and complexity of trials in England in the eighteenth and nineteenth centuries, and attribute the changes to the growing complexity of rules of evidence and procedure as well as the greater use of representation by lawyers. Examining the comparative effects of European and American trials in the late-twentieth century, Kagan (2001a,b) has shown that the expansion of 'adversarial legalism' in the United States dramatically increased the cost and length of trials without offering any corresponding increase in benefits (*see* Litigation; Legalization: The Expansion of Law's Domain).

This tendency to increase trial complexity and adversariness in the quest for judicial perfection led legal historian John Langbein to liken the modern adversarial criminal trial to Medieval European proceedings (Langbein, 1976, 1978). Both systems, he argued, aspired to 'perfect justice' but in fact precipitated disastrous results. Impossibly high standards of proof in medieval law led jurists to transform torture into a legal expediency in order to fulfill their need for certainty (which confession could provide). Contemporary jurists have developed a functional equivalent, the coercive practice of plea bargaining, to facilitate a similar objective (*see* Plea Bargaining: A Comparative Analysis). Certainty of outcome in both cases, Langbein argued, becomes a mere legal nicety, since both arrangements depend upon coercion. Both torture and plea bargaining are little more than loopholes to circumvent unrealistically high standards of proof. Simpler proceedings and more realistic expectations, Langbein and many others have maintained, would result in fairer and more workable trials (*see* Legal Formalism; Civil Law; Torture: Legal Perspectives).

Finally, in a very real sense the term 'trial' is misleading and is perhaps an outmoded concept. It implies a single coming together of the parties in a forum where facts are presented and issues resolved. Although a great many cases are handled in

a single session, most cases involve a series of hearings and/or decisions that may extend over a significant period of time. Consider a putatively 'simple' criminal case; a suspect is arrested and taken to court for an arraignment, at which time she/he is asked to enter a plea and bail is set. If pretrial release is not effected, there may be subsequent bail hearings. There may be separate hearings on such issues as a request for representation by a legal aid attorney, probable cause, a motion to suppress evidence, a motion to dismiss, or separate trials for codefendants; there may be a hearing to determine guilt, a sentencing hearing, and appeals on any of the issues just mentioned. Each one of these separate proceedings can be considered a trial in that it can result in an adjudicated determination of an issue. Contrast to this simple criminal case with complex civil cases, which can involve dozens of such issues and hearings. Under such conditions, it does not make much sense to talk about either a 'case' or a 'trial,' since contested issues have been authoritatively resolved at any number of different hearings; issues may have moved from trial courts to appellate courts and back several times, and indeed the name of the 'case' may have changed (Feeley and Rubin, 1998). European legal scholars are more likely to understand the term 'trial' in this elaborated sense than are Anglo-Americans, who are more familiar with the idea of concentrated, continuous oral proceedings. Yet, the ambiguity of the word is likely to persist in the context of most modern, complex systems of litigation. In light of this, it may be more useful to use the term 'dispute processing' (Abel, 1982, 1983).

Whither the Trial?

Trial rates in most industrialized countries have been declining steadily since the early nineteenth century, as negotiated settlements have replaced them. Judges often encourage such settlements, which they often see as rendering greater fairness and effectiveness than would be provided by a trial, and of course settlements are much more efficient (the judge may only have to formally acknowledge the outcome). But since the late-twentieth century, an increasing number of scholars and judges have noted the steady decline of the trial rate (Heyderband and Seron, 1982; Galanter, 2004), and suggested the low level is disturbing. Some argue it represents the triumph of administrative rationalization over justice (Heyderband and Seron, 1982), and others worry that it means that litigants are effectively barred from access to justice, and still others that few trials mean even fewer appeals, which can clarify law and establish precedents (Mulcahy, 2013). Certainly, the arc of the twentieth century saw a marked decline in trials in both civil law and common law countries, but it is difficult to know what to make of this trend. Simultaneously more parties have been represented by lawyers (Feeley and Lester, 1994) and by more and better lawyers, and during this same period the trial process has experienced elaboration. There are more pretrial motions to consider, such as motions to suppress, probable cause hearings, to grant discover, for summary judgment, and the like. Outcomes in such 'preliminary' matters can often be a fairly clear signal as to what the likely outcome of a case may be, and these adjudicated rulings serve as an encouragement to settle. Still, as trial rates approach to zero in the United States and other jurisdictions, and private systems of mediation and arbitration are imposed on parties through contract, the meaningfulness of the idea of one's day in court and a 'fair hearing' can be questioned (Merry and Silbey, 1984).

See also: Dispute Resolutions in Economics; Disputes, Social Construction and Transformation of; Justice and Law; Justice, Conflicts, and the Justice of Conflict Resolution; Justice: Philosophical Aspects; Justice: Social Psychological Perspectives; Legal Process and Social Science: United States; Norms; Rule of Law (and Rechtsstaat); Triangulation: Methodology.

Bibliography

Abel, R., 1982–83. The Politics of Informal Justice: The American Experience, vols. 1 and 2. Academic Press, New York.
Damaska, M., 1986. The Faces of Justice and State Authority. Yale University Press, New Haven, CT.
Durkheim, E., 1964. The Division of Labor in Society. Free Press, New York.
Feeley, M., Lester, C., 1994. Legal complexity and the transformation of the criminal process. In: Gouron, A., Mayali, L., Schioppa, A.P., Simon, D. (Eds.), Subjektivierung des justiziellen Beweisverfahrens. Klostermann, Frankfurt am Main, Germany.
Feeley, M., Rubin, E., 1998. Judicial Policy Making and the Modern State: How the Courts Reformed American Prisons. Cambridge University Press, New York.
Felsteiner, W., Abel, R., Sarat, A., 1980–91. The emergence and transformation of disputes: naming, blaming, and claiming. Law and Society Review 15, 63153.
Fuller, L., 1981. Forms and limits of adjudication. In: Winston, K. (Ed.), The Principles of Order: Selected Essays of Lon Fuller. Duke University Press, Durham, NC.
Galanter, M., 2004. The vanishing trial: an examination of trial and related matters in federal and state courts. Journal of Empirical Legal Studies 1, 459–570.
Garfinkel, H., 1956. Conditions of successful status degradation ceremonies. American Journal of Sociology 61, 4204.
Heyderband, W., Seron, C., 1982. Rationalizing Justice. SUNY Press, Albany.
Inverarity, J., Lauderdale, P., Feld, B., 1983. Law and Society: Sociological Perspectives on Criminal Law. Little Brown, Boston.
Jacob, H., Blanckenburg, E., Kritzer, H., Provine, D., Sanders, J., 1996. Courts, Law, and Politics in Comparative Perspective. Yale University Press, University Press, New Haven, CT.
Kagan, R., 2001a. Adversarial Legalism: The American Way of Law. Cambridge University Press, Cambridge.
Kritzer, H., 2004. Disappearing trials? A comparative perspective. Journal of Empirical Legal Studies 1, 735–749.
Kagan, R., 2001b. Adversarial Legalism. Harvard University Press, Cambridge, MA.
Kirchheimer, O., 1961. Political Justice: The Use Of Legal Procedure for Political Ends. Princeton University Press, Princeton, NJ.
Langbein, J., 1976. Torture and the Law of Proof: Europe and England in the Ancient Regime. University of Chicago Press, Chicago.
Langbein, J., 1978. Torture and Plea-bargaining. University of Chicago Law Review 46, 3–24.
Merry, S., Silbey, S., 1984. What do plaintiffs want? Justice System Journal 9, 148–167.
Mulcahy, L., 2013. The collective interest in private dispute-resolution. Oxford Journal of Legal Studies 33, 59–80.
Nonet, P., Selznick, P., 1978. Law and Society in Transition: Towards More Responsive Law. Harper Row, New York.
Pizzi, W., 1999. Trials without Truth. New York University Press, New York.
Scheingold, S., 1977. The Politics of Rights. Yale University Press, New Haven, CT.
Shapiro, M., 1981. Courts: A Comparative and Political Analysis. University of Chicago Press, Chicago.
Trubek, D., 1972. Max Weber on law and the rise of capitalism. Wisconsin Law Review 3, 720–753.
Weber, M., 1954. In: Rheinstein, M. (Ed.), On Law in Economy and Society. Harvard University Press, Cambridge, MA.

Triangulation: Methodology

Linda Heath, Loyola University Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

At its simplest level, triangulation refers to the use of multiple measures to capture a construct. The triangulation strategy, however, can be applied to multiple operationalizations of treatments and manipulations and to the use of multiple theories, analyses, analysts, methodologies, and research designs and approaches. The history, logic, extensions, limitations, and recent applications and developments of triangulation will be examined in this article.

Introduction

Triangulation has been called "a veritable 'magical' word in mixed methods research" (Tashakkori and Teddlie, 2003: p. 674) and a "near-talismanic method" to combine qualitative and quantitative research (Miles and Huberman, 1994: p. 266). On the other hand, triangulation has been said to be "over-used to the point where it means nothing" (Tashakkori and Teddlie, 2003: p. 674) and to "have too much meaning," therefore having "no meaning at all" (Sandelowski, 2003: p. 328). So what is this magical, near-talismanic, perhaps meaningless thing? At its simplest level, triangulation refers to the use of multiple measures to capture a construct. The triangulation strategy, however, also can be applied to multiple operationalizations of treatments and manipulations and to the use of multiple theories, analyses, analysts, methodologies, and research designs, to name but a few applications. The use of triangulation logic in the social sciences emerged at the time that the quantitative purist or positivist approach was challenged by a qualitative purist or constructivist approach (Johnson and Onwuegbuzie, 2004). At a basic level, the triangulation strategy requires an acknowledgment that no one measure in the social sciences is a perfect measurement of the construct under consideration. With only one measure of the construct, the error and biases inherent in the measure are inextricably confounded with the construct of interest. Although this triangulation logic initially was explicated in the social sciences in terms of measures and methods with Campbell and Fiske's (1959) development of the multitrait-multimethod design, the same logic can be extended to treatments, settings, populations, and many other aspects of the research enterprise, and, many research programs used triangulation logic before the term was first imported into the social sciences. Although simple in concept, triangulation "carries with it a host of implications regarding the proper conduct of social research, the effects of imperfection or unreliability of measurement operations on the development of theory, and the manner in which our field might gradually attain the status of the more developed sciences" (Crano, 1981: p. 320).

History of Triangulation Concept

More than 40 years ago, social scientists called for triangulation of research measures as a means of more adequately capturing the essence of the construct under consideration (Campbell and Fiske, 1959; Webb et al., 1966). The term 'triangulation' was probably borrowed from surveying, where multiple measures from different vantage points are used to locate an object in space by creating a triangle of three points when only two points can be measured directly. Although Erzberger and Kelle (2003) argued that "this notion loses much of its clarity when transferred to the realm of social science methodology (because) it changes from a distinct mathematical concept with a clear meaning to a metaphor with a broad semantic field" (p. 459), explanations based on physical analogs can be helpful (see, for example, Crano's, 1981 example of locating a radio signal by using multiple directional antennae). Social scientists suggested that social and psychological constructs, as well as physical objects, could be brought into clearer focus by using multiple measures from different vantage points with unique sources of error. The uses of triangulation quickly expanded to include triangulation of methods, sources, and theories as well as measures (e.g., Denzin, 1978, 1989).

Some researchers adopted the triangulation strategy wholeheartedly and launched large-scale, multi-investigator, multimethod, multisite, multimeasure studies of social problems (e.g., Lewis, 1981). The need to contend with conflicting results that arose from different methods and measures in the same study led to extensive discussions of the use of qualitative and quantitative research methods and data, across and within the same project (e.g., Reichardt and Cook, 1979; Tashakkori and Teddlie, 2003). Recently, the focus of discussion has shifted to triangulation among multiple paradigms (Greene and Caracelli, 1997; Greene and McClintock, 1985), multiple stakeholders, multiple analysts, multiple theoretical and value frameworks, and multiple targets within a study, a concept Cook (1985) calls "multiplism." Developments in analysis techniques also have promoted the use of triangulation, either at the measurement level, as in casual modeling (Joreskog and Sorbom, 1984), or at the method level, as in meta-analysis (Glass et al., 1981; Rosenthal, 1991).

Logic of Triangulation

Triangulation strategies can be contrasted with what Cook and Campbell (1979) termed "mono-operation bias." Whenever a construct is operationalized in only one way in a research study, the error attached to the method of operationalization, as well as any random error in the operationalization, is

inextricably linked with the true essence of the construct under consideration. This relationship is captured by the simple equation $X = T + e$, meaning that every data point (X) is composed of part True Score (T) and part error (e) (see, for example, Crano, 1981 for further discussion). With only one measure of a construct, the error caused by method bias, response sets, or investigator bias, to name but a few, cannot be disentangled from the true measure of the construct (see Shadish et al., 2002; Webb, 1978 for discussions of myriad types of error). By triangulating, or operationalizing the construct in several ways, the sources of influence on the data point can be disentangled, leading to a clearer understanding of the construct under consideration. Campbell and Fiske's (1959) classic discussion of the multitrait-multimethod matrix lays out a strategy for such specification.

The acknowledgment of the necessity of triangulation represented a maturing of the social sciences beyond the paradigm envy that previously had gripped the field (Crano, 1981). Early social scientists had attempted to gain acceptance and legitimacy by adopting much of the research trappings of the physical or 'hard' sciences. Operating from the positivist view of science (Cook, 1985), social scientists attempted to achieve parsimonious explanations of events and circumstances that would hold across times and settings, much as physical scientists were claiming to do. They sought the precise, well-calibrated measures, the elegant, critical experiments, and the straightforward analyses of the hard sciences. In some areas of research, this approach could be fairly successful. For example, a social scientist could operationalize 'response time' as 'the amount of time that is between stimulus and response,' with stimulus and response both being defined in terms of their operationalizations (e.g., 'the onset of the light' and 'pushing the button,' respectively). But as soon as the social scientists ventured beyond the laboratory and basic research, they discovered that mono-operationalization was inadequate to tap underlying relationships accurately and that simple analysis strategies were inadequate to describe the complicated interactive patterns in the data.

For example, social scientists quickly discovered that they could not define 'family size' as 'the number of people listed on the interview schedule under the heading Family' and be confident that they had captured the true essence of the construct 'family size.' Beyond the problems of deciding whom to list as 'family' (such as deciding what to do with people who are not biologically related but may be committed to the unit and people who are biologically related but rarely present), social scientists realized that some people are not going to report accurately, either because they misunderstand the question, because they are not aware of the true situation, or because they choose to report incorrectly.

To compensate for these problems with error and bias, researchers began to use multiple operationalizations of key constructs, and many claims of triangulation in research design are based solely on the use of multiple items to tap a construct. But, as Webb (1978) and others pointed out, using multiple instances of measures that share critical components of error does not improve the research much beyond identifying some random error. Finding that different measures or methods give the same results does not mean they are giving the right results; they might be giving results that are just similarly wrong (Erzberger and Kelle, 2003). To be most useful, multiple measures must have differing sources of error, including method error, and errors due to response bias and self-presentation, among many others. Campbell termed this important qualification the "heterogeneity of irrelevancies" (Brewer and Collins, 1981).

So, for example, having the same interviewer ask the same respondent 'how many people live in this house?' in a dozen different ways during the same face-to-face interview does not advance the construct validity of 'family size' very much. A respondent who really does not know that her daughter's boyfriend has been living upstairs will report incorrectly every time. A respondent who does not want to reveal that a cousin who is in the country illegally lives there consistently will not report that person as living in the house. An interviewer who consistently reverses threes and fives will do that every time. So repeated questioning only reduces the error associated with not understanding a particular question or with random errors made by the respondent or the interviewer. This is not to imply that reducing such error is unimportant, but rather that triangulation can be much more powerful when used across types of measures instead of within one method of measurement.

Partly in response to the shared error inherent in verbal responses, methodologists suggested the use of unobtrusive measures as a means of tapping constructs without the error inherent in verbal responses (Webb et al., 1966). Unobtrusive measures are not without error, but their usefulness comes from the fact that many of their error sources are different from the error sources that plague verbal responses (e.g., social desirability, response set). In one famous example, researchers measured the popularity of different museum exhibits by measuring the wear and tear on floor tiles surrounding different displays (Webb et al., 1966). Another researcher measured the level of activity among nursing home residents by measuring how dirty the adhesive tape attached to their wheelchair wheels became. Mosteller (1955, cited in Webb, 1978) measured the usage of various sections of the *International Encyclopedia of the Social Sciences* by noting how worn and dirty various sections of the book were. (Note to reader: Feel free to smudge, bend, and dog-ear this page.)

Extensions of Triangulation

The logic of triangulation was quickly extended to methodological orientation (quantitative or qualitative), research designs, treatments, theories, settings, and other aspects of the research enterprise. Denzin (1978) compared the pictures of reality obtained from different researchers using different methodologies at different times to the changing images within a kaleidoscope, sharing some common elements but greatly influenced by the circumstances of viewing. This is not to imply that there is no objective reality, but rather that it is filtered through the researcher's assumptions, research choices, and the factors unique to that point in time. In much the same way that Heisenberg pointed out to psychical scientists that the very act of observing or measuring a physical object can alter the reality of that object, social scientists were coming to appreciate the extent to which their own assumptions, choices, historical time

frame, and even mere presence influenced the data they collected.

Denzin (1978) outlined four basic types of triangulation: data, investigator, theory, and methodological. He further divided triangulation into the three subtypes of time, space, and person. Data triangulation across time refers to both time of the day as well as historical setting, which has been referred to as population changes associated with time (Campbell and Stanley, 1963). Some time effects are profound and obvious. For example, attitudes toward day care among women who engage in telephone interviews during the day might be very different from attitudes of women who respond in the evening (and presumably might have been at work during the day). Similarly, space could have obvious or nonobvious effects. Bosses and employees might have different views of the workplace space, but so too might workers near and far from the elevator.

A strength of triangulation is that the researcher does not need to understand completely the variation inherent in the data collection methods. In the same way that random assignment distributes error randomly across conditions in a true experiment even if the researcher is unaware of the importance of the error, triangulation captures and makes somewhat tractable irrelevancies of the research situation, even if the researcher is unaware of the importance of those irrelevancies. Of course, in the same way that true experiments can be strengthened if the researcher has identified problematic outside variables and blocked on those variables (i.e., randomized within levels of those variables), research designs can be strengthened if researchers can specify which outside factors might be problematic and can triangulate in such a way as to capture those irrelevancies. For example, if a researcher suspects that the age of respondent will relate to the variable of interest, the researcher should specifically select for a variety of ages among respondents and analyze the data specifically to examine age effects, instead of just hoping that triangulating over samples will capture those irrelevancies.

Beyond data triangulation, Denzin (1978) also advocates investigator triangulation, whereby multiple investigators bring their unique perspectives and biases to the research project. This is an acknowledgment that people, no matter how well trained, are not robotic recorders of social events. Again, if the researcher has a priori reason to suspect that some characteristic of the investigator, such as gender, might introduce error into the observations, care can be taken to select equal numbers of male and female investigators and to analyze for differences between the observations of men and women. But the use of multiple investigators should reduce effects of irrelevant observer factors, even in the absence of the researcher's awareness that such factors could be important. Of course, only irrelevant factors that vary among the investigators will be ameliorated.

In a similar vein, Denzin (1978) proposed that researchers approach projects from the perspectives of multiple theories, taking care to extend the scope beyond the theoretical orientation most comfortable for the researcher. He further suggested that rather than pit only two theories against each other in research, projects should be specifically designed to capture as many theoretical orientations as possible. Hypotheses derived from all of these theories could then be tested, and conclusions could be drawn about the relative strengths and weaknesses of the theories. The logic of triangulation is the same as with the previous examples: by using multiple theoretical views, the researcher is better able to bring the research question into focus.

Finally, Denzin (1978) described both within-method and between-method triangulation. Within a single method, such as a telephone survey, multiple items that are intended to tap the same construct can provide much clearer measures of the construct under consideration. Multiple items can be combined via a wide assortment of scaling techniques, ranging from simple additive scales to elaborate factor analytic scales. Again, the triangulation strategy does not result in items measured without error, but, rather, it allows the error that is not shared among the items to be identified and removed. The resulting scale should reflect only the 'true' component of the construct measured plus any error shared by the various items. For example, items that all cause the respondent to respond in a socially desirable manner will not allow the researcher to separate out social desirability components from components that actually tap the construct under consideration. But if some of the items have a response set bias and others do not, the use of multiple measures will allow the response set bias to be identified (within limits).

Because within-method triangulation can leave varying amounts of error inextricably linked to the key construct, researchers are encouraged to triangulate between methods (Brewer and Hunter, 1989; Campbell and Fiske, 1959; Cook, 1985; Hunter and Brewer, 2003; Reichardt and Cook, 1979; Webb et al., 1966). By combining methodologies as different as participant observation and survey research, for example, researchers can study processes under vastly differing assumptions, biases, and errors. For example, the *Reactions to Crime* Project (Lewis, 1981) involved participant observation, content analysis, telephone surveying, and in-person interviewing, all within one large framework. The approach used by the *Reactions to Crime* Project was sequentially multimethod, with the participant observation being completed first, followed by the telephone survey, and finishing with the in-person interviewing. With sequential multimethod work, researchers can use results from previous stages to inform design at later stages. Although this sequential strategy has strengths, Greene and McClintock (1985) pointed out that concurrent multimethod work, in which all methods are implemented at the same time, avoids contamination of one method with results from another and produces a purer triangulation.

Probably the most extensive elaboration of triangulation logic was provided by Cook (1985) who described research that involves multiple stakeholders, multiple data analyses, multiple definitions, multiple methods, multivariate casual modeling, multiple hypotheses, multiple populations, multiple settings, and multiple times. He stated that such multiplism allows the researcher to approximate the "ultimately unknowable truth through the use of processes that critically triangulate from a variety of perspectives on what is worth knowing and what is known. Triangulation requires generating multiple instances of some entity and then testing whether the instances converge on the same theoretical meaning despite the irrelevant sources of uniqueness that any one instance may contain" (1985: p. 57). Like previous writers, however, Cook was quick

to point out that "multiplism is no panacea, for we can multiply replicate the same mistake or the same parochial set of assumptions" (1985: p. 58). Although triangulation can be vastly expanded in scope, its strengths and weaknesses remain the same.

Limitations of Triangulation

Flick (1992) and Fielding and Fielding (1986) have suggested that proponents of triangulation mistakenly believe triangulation solves all construct validity problems. Miles and Huberman (1994) suggested that some researchers think it has near-talismanic properties. So is triangulation a magic wand for solving research problems? Or, at least, did (or do) some researchers believe so? To the contrary, early proponents of triangulation were sounding a warning about the inherent weakness of all measures, and the proposal to use multiple measures with nonoverlapping irrelevancies was an acknowledgment of the profound problems inherent in measuring. In a discussion of the distance between desired and achieved measures of social science constructs and hence of the hypotheses the operationalized measures test, Campbell concluded that "direct knowing, nonpresumptive knowing, is not to be our lot" (1969; p. 52). Therefore, one important limitation of triangulation is that it does not produce perfect measures.

If triangulation does not ensure a perfect reflection of reality, why is it touted as such a boon to research? Researchers have suggested various metaphors to explain the gains from triangulation. Erzberger and Kelle (2003) used a puzzle metaphor to illustrate the gains from triangulation. Each piece of research with a different approach, method, or measurement provides another piece to the puzzle that represents reality. The pieces all together give us a more complete picture of reality than any one piece by itself. They suggested triangulation results not so much in validation as in extension of our understanding. Similarly, Flick (1992) suggested triangulation produces alternatives, rather than validation of the original study, while Olsen (2004) suggested triangulation is not aimed at providing validation but at increasing understanding. Research based on different assumptions and methods can produce results that conflict without either being invalid. Finally, Cresswell et al. (2003) offered the metaphor of maps created by results that stand for the underlying process but that are not to be conflated with the actual process, in the same way that a map is not conflated with the landscape it describes.

Another limitation inherent in the triangulation strategy is that time, finances, and respondent goodwill conspire to prevent a researcher from triangulating on all the constructs in the research program. One can easily imagine how attempting to triangulate on theoretical orientations, research designs, and measurement strategies could lead quickly to a program of research that could never produce a single result, because it never could be completed. The successful researcher must therefore choose where to invest his or her triangulation resources, and the soundness and usefulness of the research outcomes depend to a large extent on the wisdom of those choices during the research design phase.

The amount of effort a researcher chooses to devote to triangulating on the construct should reflect both the importance of that construct for the research question and the level of difficulty in achieving a good measurement. For example, if age is a crucial variable for a research question, investigators might invest extra resources in accessing birth records as well as asking for age and birth date in a survey. If age is less crucial to the question at hand, the researchers might decide to rely on the verbal replies only. Similarly, a true experiment performed in a laboratory might prove adequate for testing a hypothesis about visual processing, but researchers investigating hypotheses about reactions to stress might need to invest in multiple methods to achieve clarity about the processes under investigation.

The researchers' choices about constructs for triangulation resources can have significant effects on the results of the research. For example, if a researcher believes Family Income to be a more important construct in explaining school achievement than Parental Education Level and consequently devotes more effort to triangulating on the measure of Family Income than Parental Education Level, analyses might well reveal a stronger relationship between School Achievement and Family Income than School Achievement and Parental Education, because Family Income was more reliably measured. If the researcher had made the opposite assumption and had devoted more resources to measuring Parental Education than to Family Income, the results might well have shown the opposite effect. The utility of triangulation, therefore, is limited by the soundness of the researchers' initial decisions about resource allocation.

Advances in Triangulation

The logic of triangulation is inherent in both structural equations modeling (Joreskog and Sorbom, 1984) and meta-analysis (Rosenthal, 1991). Structural equations modeling (SEM) is an analysis technique that relies on multiple measures (manifest variables) of some underlying construct (latent variable). Researchers specify which manifest variables relate to which latent variables, and the analysis strategy pulls out common variance among the manifest variables that feed into a latent variable. Variance in the manifest variable that is not shared with the other manifest variables for that latent variable is identified as 'error.' Error terms can be allowed to correlate with other error terms. This is a sophisticated extension of the logic used by Campbell and Fiske (1959) in their explication of the multitrait-multimethod matrix. They pointed out that only part of what researchers are measuring is the key construct they are intending to measure. The rest of the measure includes method variance, random error, and possibly other types of error and bias. SEM requires researchers to identify error explicitly in their models, which makes very clear the frailty of any one measure in isolation. In addition, SEMs show clearly how error terms from similar measures correlate, increasing researchers' awareness of such measurement effects. Recently, this technique has been widely employed and has been used to study spirituality (Sandage and Williamson, 2010), traumatic stress and avoidance (Briere et al., 2010), and early childhood anxiety (Mian et al., 2011), among many other topics.

In a similar vein, meta-analysis exploits multiple studies, with multiple operationalizations, multiple investigators, multiple methods, multiple sites, multiple populations, and many other differences to arrive at a summary statistic that reports the strength of underlying relationships across studies. Instead of requiring one researcher or research team to approach a research question from multiple perspectives, theories, and methodologies, meta-analysis capitalizes on natural variations among research programs to achieve triangulation. A main benefit of meta-analysis is that methodological variation can be built explicitly into the analysis. For example, in a review of the research on environmental and genetic effects on aggression, Miles and Carey (1997) reported that observational measures in laboratory studies showed no genetic effects and strong environmental effects, while self-report and parental rating measures showed both genes and environment to have strong relationships with aggression. Because variance associated with particular methods and measures can be explicitly analyzed in meta-analysis, the increased understanding afforded by triangulation strategies becomes apparent. And the conflicting results from different methods, far from being a nuisance or drawback to triangulation approaches, are rich clues to contingencies that need to be brought to light if social research is ever to reveal fully how these processes operate (Cook, 1985). This method has gained acceptance across a wide range of content areas, including studies of birth order effects in major league baseball (Sulloway and Zweigenhaft, 2010) and religiosity and self-enhancement (Sedikides and Gebauer, 2010).

The importance of triangulation also has been revealed by the prevalence of statistical interactions in social research. Evaluation researchers and others who study social problems rarely find that simple, main effects capture the richness (some would say messiness) of the world beyond the research laboratory (Cook, 1985). Evaluators no longer ask, "Does this work?" but, rather, "For whom does this work, in what settings, with what constraints?" The prevalence of interactions pushes researchers to branch out to different populations, different time frames, and different methods, in other words, to triangulate across many factors in the research setting.

Social research has come a long way from the days when researchers sought pure measures, elegant, simple designs, and straightforward analysis strategies. The norm for social research now is multiple measures, elaborate multimethod design, and analysis strategies that clearly indicate the multiple causation and interactive patterns apparent in the social world. In no small part, this march toward increasing complexity of the research enterprise started with early explications of the logic of and necessity for triangulation.

See also: Content Validity; Generalizability Theory; Indicator: Methodology.

Bibliography

Briere, J., Hodges, M., Godbout, N., 2010. Traumatic stress, affect dysregulation, and dysfunctional avoidance: a structural equation model. Journal of Traumatic Stress 23, 767–774.

Brewer, J., Hunter, A., 1989. Multimethod Research: A Synthesis of Styles. Sage, Newbury Park, CA.

Brewer, M., Collins, B.E., 1981. Perspectives on knowing: six themes from Donald T. Campbell. In: Brewer, M., Collins, B.E. (Eds.), Scientific Inquiry and the Social Sciences. Jossey-Bass Publishers, San Francisco, CA.

Campbell, D.T., 1969. A phenomenology of the other one: corrigible, hypothetical, and critical. In: Mischel, T. (Ed.), Human Action: Conceptual and Empirical Issues. Academic Press, New York.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81–105.

Campbell, D.T., Stanley, J.C., 1963. Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago.

Cook, T.D., 1985. Postpositivist critical multiplism. In: Shotland, L., Mark, M. (Eds.), Social Sciences and Social Policy. Sage, Beverly Hills, CA.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally College Publishing, Chicago.

Crano, W., 1981. Triangulation and cross-cultural research. In: Brewer, M., Collins, B.E. (Eds.), Scientific Inquiry and the Social Sciences. Jossey-Bass Publishers, San Francisco, CA.

Cresswell, J., Clark, V., Gutmann, M., Hanson, W., 2003. Advanced mixed methods research designs. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social & Behavioral Research. Sage, Thousand Oaks, CA.

Denzin, N.K., 1978. Sociological Methods: A Sourcebook. McGraw Hill, New York.

Denzin, N.K., 1989. The Research Act: A Theoretical Introduction to Sociological Methods, third ed. McGraw Hill, New York.

Erzberger, C., Kelle, U., 2003. Making inferences in mixed methods: the rules of integration. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social & Behavioral Research. Sage, Thousand Oaks, CA.

Fielding, N.G., Fielding, J.L., 1986. Linking Data. Sage, Beverly Hills, CA.

Flick, U., 1992. Triangulation revisited: strategy of validation or alternative? Journal of the Theory of Social Behavior 22, 175–197.

Glass, G.V., McGaw, B., Smith, M., 1981. Meta-Analysis in Social Research. Sage, Beverly Hills, CA.

Greene, J., Caracelli, V.J., 1997. Defining and describing the paradigm issue in mixed-method evaluation. In: Greene, J., Caracelli, V.J. (Eds.), Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms. Jossey-Bass Publishers, San Francisco, CA.

Greene, J., McClintock, C., 1985. Triangulation in evaluation: design and analysis issues. Evaluation Review 9, 523–545.

Hunter, A., Brewer, J., 2003. Multimethod research in sociology. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social & Behavioral Research. Sage, Thousand Oaks, CA.

Johnson, R.B., Onwuegbuzie, A., 2004. Mixed methods research: a research paradigm whose time has come. Educational Researcher 33, 14–26.

Joreskog, K.G., Sorbom, D., 1984. LISREL VI, Analysis of Linear Structural Equation Systems by the Method of Maximum Likelihood: User's Guide. International Education Services, Chicago.

Lewis, D., 1981. Reactions to Crime. Sage, Beverly Hills, CA.

Mian, N., Briggs-Gowan, M., Carter, A., 2011. An ecological risk model for early childhood anxiety: the importance of early child symptoms and temperament. Journal of Abnormal Child Psychology 38, 501–512.

Miles, D.R., Carey, G., 1997. Genetic and environmental architecture of human aggression. Journal of Personality and Social Psychology 72, 207–217.

Miles, M., Huberman, M., 1994. Qualitative Data Analysis: An Expanded Sourcebook, second ed. Sage, Thousand Oaks, CA.

Mosteller, F., 1955. Use as evidenced by an examination of wear and tear on selected sets of ESS. In: Daris, K., et al. (Eds.), A Study of the Need for a New Encyclopedia Treatment of the Social Sciences, Unpublished manuscript.

Olsen, W., 2004. Triangulation in social research: qualitative and quantitative methods can really be mixed. In: Holborn, M. (Ed.), Developments in Sociology. Causeway Press, Ormskirk.

Reichardt, C.S., Cook, T.D., 1979. Beyond qualitative versus quantitative methods. In: Cook, T.D., Reichardt, C.S. (Eds.), Qualitative and Quantitative Methods in Evaluation Research. Sage, Beverly Hills, CA.

Rosenthal, R., 1991. Meta-Analytic Procedures for Social Research. Sage, Beverly Hills, CA.

Sandage, S., Williamson, I., 2010. Relational spirituality and dispositional forgiveness: a structural equations model. Journal of Psychology and Theology 38, 255.

Sandelowski, M., 2003. Tables or tableaux? The challenges of writing and reading mixed methods studies. In: Tashakkori, A., Teddlie, C. (Eds.), Handbook of Mixed Methods in Social & Behavioral Research. Sage, Thousand Oaks, CA.

Sedikides, C., Gebauer, J., 2010. Religiosity as self-enhancement: a meta-analysis of the relation between socially desirable responding and religiosity. Personality and Social Psychology Review 14, 17–36.

Shadish, W., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, New York.

Sulloway, F., Zweigenhaft, R., 2010. Birth order and risk taking in athletics: a meta-analysis and study of major league baseball. Personality and Social Psychology Review 14, 402–416.

Tashakkori, A., Teddlie, C. (Eds.), 2003. Handbook of Mixed Methods in Social & Behavioral Research. Sage, Thousand Oaks, CA.

Webb, E.J., 1978. Unconventionality, triangulation, and inference. In: Denzin, N.K. (Ed.), Sociological Methods: A Sourcebook. McGraw Hill, New York.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Non Reactive Research in the Social Sciences. Rand McNally, Chicago.

Tribe

Andre Gingrich, Institute for Social Anthropology, Austrian Academy of Sciences, Vienna, Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Today, a largely defunct comparative concept from the colonial era, the term tribe once designated allegedly 'primitive' local medium-sized social entities with some territorial rights. From its contexts of emergence, the term often attained largely pejorative connotations including nonhistorical qualities of secluded stability. The term still is in current usage for legal and political reasons in many parts of the world, although sometimes being confused with ethnic groups and other more appropriate notions. In ethnoarchaeology, history, and anthropology, it may continue to serve as a scientific concept for certain descriptive purposes if liberated from its ideological connotations.

The notion of tribe emerged in the seventeenth and eighteenth centuries in the course of colonial expansion. Abandoned increasingly by the majority of anthropologists from the middle of the twentieth century onward, the term nevertheless lived on in certain anthropological constructions, until waves of criticism dealt final blows to a concept that is now largely defunct as a general comparative category. Those anthropologists and historians who retain the contested concept of tribe use it mainly as an ethnographic tool and as equivalent to certain vernacular or national legal terms referring to local conceptions of collective sociopolitical identity, while agreeing that a precise general definition is almost impossible. This results in loose working definitions which see tribes as medium-sized, centralized, or acephalous entities displaying a combination of basic characteristics:

1. They are usually associated with a territory, homeland, or tribal area, while using nonterritorial criteria to distinguish between members and nonmembers. The degree of control of the tribal territorial base, and its extension and location tend to change as a result of struggles with states or other opponents. To the outside world, access to territorial usage appears to be confined to members of the tribe, although neighboring groups and superior authorities may also enjoy rights of usage, of tribute, or of taxes. Internally, access to territorial usage may be differentiated according to gender, age, status, wealth, and other aspects of power.
2. Ideologically and socially, tribal members usually share some dominant idiom of common origin, such as descent from a single ancestor, emphasizing group cohesion over outside interests and internal differentiation. This may also serve to legitimize a preference for 'marrying within the tribe,' which in practice, however, is observed far less often than it is upheld in theory. On the contrary, some amount of intertribal and interethnic marriage is indispensable to ensure all kinds of (political, economic, social) alliances with other groups.
3. Although part of the tribal population live more or less permanently in the tribal territory, a considerable portion may live in the world outside. Ritual and kinship ties and obligations are enacted to ensure and strengthen intimate networks among 'home' groups and 'outside' groups, serving as one among several inventories of group adherence within a wider world. In the course of these interactions, tribal group membership may change in many ways: individuals or groups may enter or leave a specific tribe and whole tribes may also collapse or be created.

Tribes and Colonialism

The concept of tribe dates from the earliest phases of anthropology's history, carrying assumptions that are now being criticized and questioned inside and outside academia. Most colonial governments used the term for specific practical interests in administration and dominance while often constructing 'tribes' themselves to that purpose. Beyond these practical interests, colonial settler regimes also valued the term ideologically, since it tended to portray local groups as 'backward' and 'primitive,' which helped to justify a colonial 'mission' of development and civilization. For similar reason, many postcolonial governments and national elites reject the term as being counterproductive to their own claims toward national unity, progress, and modernity.

During decolonization, the term tribe therefore attained a largely negative and pejorative meaning in dominant opinions in many parts of the world. In public discourses concerning these areas, the terms 'ethnic groups' or 'indigenous peoples' are usually preferred to 'tribes,' although the latter usually designates smaller groups. Legal terminology, nevertheless, may continue to apply the term, and in many regions even native speakers may use it with dignity.

In parts of central and western Asia, however, as well as in North Africa, particularly in those areas influenced by Islam, the term tribe and its local equivalents have never had a predominantly pejorative meaning. On the contrary, the term here is usually associated with notions of pride and autonomy, of honor, and of some independence from colonial and local states. Also, notions of 'indigenous' or 'ethnic' groups do not apply well to many of these specific ethnographic contexts. Both of these factors are therefore responsible for the continuing, stronger importance of the term tribe in academic and public discourses inside and about these regions.

Tribes in Anthropology

Before anthropology's emergence, travelers, missionaries, and scholars had used the concept since the seventeenth century in

describing those indigenous peoples whom European powers were about to colonize. Ancient Middle Eastern and Graeco-Roman literature served as sources of reference to depict the newly colonized peoples as representing some kind of forerunners to the 'advanced nations,' equivalent to the *tribus* of the ancient Mediterranean world. The construction of this 'forerunner' image could follow either more sympathetic lines, by emphasizing the enlightened aspect of a noble savage who basically was 'like us,' or it could enhance the more discriminating image of backwardness.

Both ideas presupposed a search for origins, taking European culture as the climax of humanity. In its basic conceptualization, a tripartite world had thus already been constructed, which early anthropology inherited and further elaborated. 'Primitive' cultures consisted of foragers, and pastoral and agrarian 'tribes,' whereas 'oriental' cultures were associated with agrarian states, cities, and with writing. Anthropological evolutionism (e.g., Morgan, 1877; Smith, 1903) elaborated and specified this conceptual inventory, in which, by and large, tribes constituted the decisive transitory stage between early hunters and gatherers and the rise of agrarian states. In a long-term archeological perspective, this may still serve as a valid distinction for certain regions and periods, if liberated from its earlier ideological underpinnings.

Classical anthropological evolutionism collapsed, however, because the increasing body of empirical data showed the failure of any general laws of development, including those regarding tribes. By contrast, the new school of 'historical particularism' or cultural relativism that came to dominate US anthropology around 1900 emphasized the situated, fluid contextualization of each local and tribal group. While this prepared the ground for an empirical orientation that is still valid at the time of writing, the inherent relativism rejected almost any comparative meaning for the concept of tribe. The comparative aspect, by contrast, was elaborated explicitly by anthropological functionalism, largely of British and Northern European provenance. In an earlier variant, Evans-Pritchard (1940) outlined the 'structural relativity' of many tribal group relations around a basic social organization of 'nested' group segmentation (Barth et al., 2005).

In fact, such segmentary relations do represent a dominant model among many local (tribal) groups in Africa and Asia, often being expressed in the prevailing idiom of unilineal genealogy. In these terms, the 'descendants' of one ancestor are contradistinguished, under some conditions, against those from another ancestor, while under different conditions the same group of descendants may split up and oppose each other. Mutual threat, equality of group status, and reproduction of social equilibrium are inherent to this indigenous model, taken up and reconstructed as an anthropological model. Major critiques, however, pointed to implications of timeless stability and secluded cohesion, which to some degree corresponded well with the ideologies of late colonialism.

Gellner (1981, 1990), who was most influential in transferring the segmentary model from sub-Saharan Africa to North Africa and the Middle East, reacted to some of that criticism with an increasingly more refined variant that to an extent absorbed Ibn Khaldun's reasoning on the topic. According to Gellner, Middle Eastern tribal societies interact constantly out of the peripheries with each other and with the states in their centers, sometimes being absorbed or reshaped in the course of these interactions, and sometimes themselves destroying and recreating central states. These processes always promote tribal centralization, inequalities, and tendencies to fusion and fission, which are simultaneously more or less countered by applying the local segmentary model. Jamous (1981) showed in more detail than Gellner how these tendencies to feuding and conflict may be embedded in idioms of 'honor and shame.' His account illustrates how these tensions, which favor the emergence of ephemeral 'big men,' are countered through mediation by third parties, who in many cases are of superior, holy status.

While Marxist, late functionalist, and neoevolutionist critics (for example, Fried, 1975; Godelier, 1973; Southall, 1970) were the first to radically criticize current ideas about tribes in the 1960s and early 1970s, a few anthropologists held on to a general, neoevolutionist conception of tribe. Sahlins (1968) advanced a distinction between acephalous 'segmentary lineage' tribes, more centralized 'chiefdoms,' and 'conical clan' intertribal hierarchies. On the other hand, in Middle Eastern studies the concept gained new respect in the Gellner/Jamous variant, by its consideration of formerly neglected aspects, such as historical dynamism or certain ideological, political, and social factors. A second wave of academic criticism throughout the 1980s and early 1990s came from feminist, postmodern, and postcolonial writers, who deconstructed profoundly the colonial legacy in the concept, and exposed the complete negligence of economic factors and of female lives in the 'tribal' model (Abu-Lughod, 1986; Gilsenan, 1996).

Current Readings

In present discussions, the debate is therefore oscillating between complete rejection of the term 'tribe' and its more or less restricted usage for specific times and regions. Most scholars, however, would agree that the concept is obsolete as a general comparative category outside particular areas. In addition, historical and archeological research might continue, for parts of northern Africa and Asia at least, to consider some conception of tribe as useful (Dresch and Heykel, 1995). Such a restricted, critical elaboration of the concept would, of course, have to take into account seriously previous critiques.

Some of the criticism was creatively synthesized by Bonte and Conte (Bonte et al., 1991). Their approaches basically privilege historical change over repetition, interaction over system, status rivalry over group equality, and marriage and other forms of alliance over descent and ancestry. This helps to outline internal hierarchies among local tribal groups, including the constraints and fields of agency encountered by women, while demonstrating simultaneously how tribal organization is instrumentalized to defend states, to support struggles for central power, or to engage in civil war for various interests.

The postcolonial condition seems in fact to encourage such a restricted, critical elaboration of the concept, rather than discarding it totally. Recent attempts to substitute the term by other European alternatives such as 'principality' (Sneath,

2007) or 'community' (Blumi, 2010) offer few analytical advantages while in fact empirically ignoring what they claim to rephrase. Under the conditions of advanced globalization after the Cold War, and of increased competition within as well as among weakened local states (ranging from Mauritania and Somalia to Afghanistan and Bangladesh), ideologies of ethnicity, but also of tribalism, are gaining significance in many local arenas. It would amount to a strange paradox if critical researchers limited themselves defensively to deconstructing completely the concept of tribe, while many parts of the world in fact go through retribalization of some sort. An active reformulation where empirically appropriate might be the more productive alternative. To that purpose, ideological movements of 'tribalism' will have to be distinguished more clearly from 'tribes,' parallel to the concepts of nationalism and nations. Tribalism and its activists strive more or less successfully to create imagined communities of tribes, with ideological pretexts of 'origins' and of 'cohesion' that serve competition over power and resources. In this perspective, anthropologists are as critical and skeptical about tribalist ideology as they are with regard to nationalism. All the same, they would not deny from the outset that nation states and tribal groups may exist; rather, they would insist on exploring the interrelations between such local entities, their ideological activities, and the world around them.

See also: Chiefdoms, Archaeology of; Clan; Ethnic Identity and Ethnicity in Archaeology; Ethnicity: Anthropological Aspects; Ethnoarchaeology.

Bibliography

Abu-Lughod, L., 1986. Veiled Sentiments: Honor and Poetry in a Bedouin Society. University of California Press, Berkeley, CA.

Barth, F., Gingrich, A., Parkin, R., Silverman, S., 2005. One Discipline, Four Ways. University of Chicago Press, Chicago.

Blumi, I., 2010. Chaos in Yemen: Societal Collapse and the New Authoritarianism. Routledge, London and New York.

Bonte, P., Conte, E., Hames, C., Ould Cheikh, A.W., 1991. Al-Ansāb: la quète des origines. Anthropologie historique de la société tribale arabe. Éditions de la MSH, Paris.

Cooke, M., 2013. Tribal Modern: Branding New Nations in the Arab Gulf. University of California Press, Los Angeles, CA.

Dresch, P., Haykel, B., 1995. Stereotypes and political styles: Islamists and tribesfolk in Yemen. International Journal of Middle East Studies 27 (4), 405–431.

Evans-Pritchard, E.E., 1940. The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People. Oxford University Press, London.

Fried, M.H., 1975. The Notion of Tribe. Cummings, Menlo Park, CA.

Gellner, E., 1981. Muslim Society. Cambridge University Press, Cambridge, UK.

Gellner, E., 1990. Tribalism and the state in the Middle East. In: Khoury, P.S., Kostiner, J. (Eds.), Tribes and State Formation in the Middle East. University of California Press, Berkeley, CA.

Gilsenan, M., 1996. Lords of the Lebanese Marches: Violence and Narrative in an Arab Society. Tauris, London.

Godelier, M., 1973. Le concept de tribu. Crise d'un concept ou crise des fondements empiriques de l'anthropologie? In: Godelier, M. (Ed.), Horizon, Trajets Marxistes en Anthropologie. Maspero, Paris.

Jamous, R., 1981. Honneur et Baraka: Les Structures Sociales Traditionelles dans le Rif. Cambridge University Press, Cambridge, UK. Editions de la MSH, Paris.

Morgan, L.H., 1877. Ancient Society. Holt, New York.

Sahlins, M.D., 1968. Tribesmen. Prentice-Hall, Englewood Cliffs, NJ.

Smith, W.R., 1903. Kinship and Marriage in Early Arabia. Black, London.

Sneath, D., 2007. The Headless State: Aristocratic Orders, Kinship Society, and Misrepresentations of Nomadic Inner Asia. Columbia University Press, New York.

Southall, A., 1970. The illusion of tribe. Journal of Asian and African Studies 5 (1–2), 28–50.

Trickster Ethnography

Jesse W Shipley, Haverford College, Haverford, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The trickster figure in anthropology, folklore, and religious studies has been used to examine how a society understands itself through the study of transgressive characters and local notions of creativity. Tricksters are fundamentally dual raising ontological and epistemological questions about rationality, morality, temporality, and meaning. Tricksters have been important signs of irreverence and inversion and shift between tellers of mythic tales, world makers, culture heroes, selfish hustlers, and agents of change who playfully navigate violent worlds. They are linguistic and ritual mediators, translators standing at the crossroads of mythic and historic time, spiritual and social space, animal and human subjectivities. As tricksters have been identified across various societies, the category of trickster at times tells us more about a process of reductionist inquiry and the anthropological desire for analytic categories than about the lived worlds in which these diverse figures live.

The trickster figure is a mediator who stands on the border between realms, showing how a society understands itself by embodying its transgressions. In anthropology, folklore, and religious studies, identification of the category of the trickster has been a way to fit various apparently transgressive characters and social phenomena into a model of social cohesion. This speaks both to the history of anthropology as well as to fundamental ontological and epistemological questions regarding rationality, duality, temporality, and meaning raised in ethnographic inquiry. Tricksters are fundamentally dual. They are important signs of irreverence and inversion and shift between tellers of mythic tales, world makers, culture heroes, selfish hustlers, and agents of change. In a sense, tricksters have provided a way to theorize the social relevance of creativity. They are linguistic and ritual mediators, translators standing at the crossroads of mythic and historic time, spiritual and social space, and animal and human socialities. The trickster literature speaks to recently renewed debates in anthropology about the relationship between nature and culture as well as difference and translatability.

Coyote, Raven, and Hare in various Native American mythic traditions have been characterized as tricksters igniting scholarly debates about how one character can be both "culture hero" and "selfish-buffoon." Eshu, the Yoruba trickster and Legba in Dahomey are trickster deities that oversee markets and crossroads and mediate spiritual divination and all sorts of risk. The Tortoise in Yoruba (Owomoyela, 1997) and the Akan Spider trickster Ananse are central to mythic storytelling and mediators of social, linguistic, and spiritual transformations. Loki from Norse mythology is a shape-shifting god of mischief. In Greek mythology, Prometheus is a trickster who fooled Zeus to give mankind fire. The Greek Hermes, emerging from Sanskrit and Aryan traditions and later reincarnated as the Roman Mercury, is a trickster as messenger of the gods, inventor of the arts, and patron of thieves. Odysseus is a trickster in his wonderings and wanderings. European festivals of Rabelaisian inversion have been seen as invocations of the trickster (Davis, 2006). B'rer Rabbit in African-American folklore is a trickster who gets in and out of trouble through nimble thought. The Malay "mouse deer is the stereotypical 'trickster' figure: a small and weak but agile creature who survives and triumphs over far more powerful beasts by his wits, his deceit and his cunning" (Scott, 1987: p. 300).

Considering the disparate types who have been called tricksters and their varying social relevances raises how this figuration demarcates a coherent analytic category by uniting a menagerie of superficially similar renegades, messengers, and hustlers. Trickster facilitates a sometimes sloppy, sometimes productive slippage between social analysis and the reflexivity of lived experience. Fundamentally, trickster's various permutations create an analytic category that reveals various social signs of self-consciousness. In doing so, it unites a disparate set of characters and tales showing how ethnography, textual analysis, and social theory make sense of sociocultural reflexivity.

As tricksters have been identified across various societies, the category of trickster at times tells us more about a process of reductionist inquiry and the anthropological desire for analytic categories than about the lived worlds in which these diverse figures reside (Beidelman, 1980). Early theories of tricksters see them as amoral figures who are nevertheless fundamental to the production of social norms in their roles as extraordinary creatures who unwittingly make the world habitable for humans. Trickster has been part of the development of central strands of anthropological theory in relation to comparative religion, folklore, ritual theory, Boasian notions of culture, structuralist analyses of myth, British structural-functionalism, and theories of power and resistance. Analysts often blur an understanding of trickster as a locally identified character in mythic texts and folktales read as socially sanctioned representations of self and as a universal archetype defined by certain types of ambivalent destructive – creative behavior, selfishness and artful hustling. He – and it is certainly a highly gendered, masculine character – provides us a way to examine the changing ways researchers have theorized coherence and rationality and the relationships of representations to culture.

Anthropologists, folklorists, and comparative religion scholars have identified trickster as a figuration for organizing a set of characters and stories from disparate times and places. The category coalesces in the late nineteenth and early twentieth century through interest in Native American folklore and studying

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12227-5

so-called primitive religions as part of comparative religions projects that led to mid-twentieth century interest in identifying the archetypes fundamental to human existence through the folkloric. In the late twentieth century, analysts – most of them relying upon reworked data of older folklorists and anthropologists – see the trickster as a figure of subversion and innovation arising particularly among underclass groups and within historical conditions of inequality. Tricksters and trickster-like behavior gave those on the margins ways to subvert and invert social norms and address power from within through parody and double-voicing. Tricksters in this guise focus on selfish desire, which becomes a form of agency. Rather than seeing them as mythological producers of social order through the inverted representation of social values, they become actors in ongoing, historically specific resistance struggles that address structural inequality. The oscillation between ideas of the trickster as a mythological cultural mediator and as renegade hustler shows how the same evidence can be used to make contradictory arguments while appearing on the surface to be reinforcing the sociohistorical relevance of a general category.

Early twentieth-century ethnographic and folkloric studies take tricksters as representations of specific cultural and mythic systems. The trickster proves a fruitful link between anthropology, religious studies, and psychoanalysis as an archetypical character that links a cultural logic through comparative analyses, to more profound, universal aspects of the humanity. African tricksters tend to be seen as linguistic and ritualistic mediators of human relations. As in Native American examples, they are significant in how they create social relations and values through rituals and signs of inversion. In both areas, analysts struggle to make sense of the relationship between what they see as the contradictions of tricksters: as selfish amoral hustlers and creative heroes. Since the actions that analysts see as contradictory do not seem to concern the societies under consideration, moral coherence and desire for rational cosmology reveal more about the analysts than the people under consideration.

While early studies of Native American and African societies imagined tricksters as figures of traditional cosmology, later scholars presented tricksters as transgressive, historically situated characters that speak to and about power, representing the political strategies of underclass peoples. Zora Neale Hurston (1935), of course, talks of tricksters in a variety of ways, taking the Boasian interest in myth and folklore into understandings of America's life. Her work implicitly grounds the trickster as a figure of collective memory and power struggles, rather than sign of cultural order. Tricksters abound in her tales of African-American and Caribbean politics, folklore, and life. Men and women trick each other. God and white people trick black folks into accepting their historical burdens of labor. Black people in America hustle and charm white society in order to survive. Hurston the teller of tales is a trickster slipping between categories, finding and elaborating different stories for different audiences. For Hurston, trickery is a culturally-historically specific source of pleasure, power, and survival. However, in the hands of others, the trickster becomes a general metaphor applicable to any social type who exhibited irreverent hustler-like behavior. As it is generally applied, the category removes characters from the nuances of social context for the sake of analytic explanation. Tracing how the trickster story has been told by scholars provides insight into ethnography's changing relationship to comparative techniques, mythic logics, ideas of creativity and transgression, and the entanglements of history with culture.

Native American Mythology from Culture to Structuralism

The trickster emerges from late nineteenth and early twentieth century collecting of Native American folkloric-mythological tales. The study of Native American religion and cosmology through mythic tales provides missionaries, anthropologists, and amateur ethnographers' ways to understand the so-called primitive cultures as they face erasure through the encroachment of white civilization and religious conversion. In documenting disappearing native cultures, myths are seen as relatively pure representations of archaic cultures giving insight into the nature of humanity. The trickster figure is central to many of the tales collected. Native American tricksters tend to be the Coyote in California and the Plains, the Raven in the Pacific Northwest, the Hare and Rabbit among Algonquian speaking groups, and others in the Northeast (cf Carroll, 1981). Tricksters tend to have both human and animal characteristics. They live in both mythic and human time.

Paul Radin's *The Trickster: A Study in American Indian Mythology* (1956) is the most influential anthropological study of the trickster, shaping the subsequent popularity of this term. In 1908, Radin enlisted a key informant to find an old man who knew what were understood as authentic versions of the Winnebago trickster cycle, uninfluenced by rising proto-Christian Peyote cults. The tales were purportedly recited in the proper ritual context written down and later transcribed by Radin. The trickster cycle of the Winnebago tells of Wakdjunkaga's exploits in a self-contained narrative focused on the making of the social world. Radin provides analysis and interpretation in a separate essay. The 1956 edition includes summaries of the trickster cycle of Sitconski, the Assiniboine trickster recorded by Robert Lowie (1909) of Raven, the Tlingit trickster recorded by J.R. Swanton (1909), and the Winnebago Hare cycle. It also includes two essays, one by Karl Kerenyi comparing the trickster to Greek mythology and one by C.G. Jung exploring the trickster as archetype.

In Radin's account, the trickster is fundamental to Native North American societies. Wakdjunkaga means "the tricky one" in Winnebago (1956: p. 132). In the cycle of tales, Wakdjunkaga is first removed to a primitive state, as an "amorphous, instinctual, and unintegrated being" without ethical values, social awareness, or self-consciousness. In contrast, the world of nature is represented as "conscious" aware of Wakdjunkaga and his actions (1956: p. 133). Wakdjunkaga's body, emotions, and actions are disordered. The process of the Winnebago trickster cycle, then, orders and develops the trickster out of isolation, creating his body through a social logic. His body is literally shaped with penis and intestines and left and right hands being arranged appropriately. Wakdjunkaga seems at times to strive for "full psychical and social–ethical maturity" though does not seem to achieve it (1956: p. 134). He does not belong "in the world of men but to a much older world" (1956: p. 148). In Radin's analysis, the experience of emotions like

grief and fear lead to self-awareness. The tales describe the trickster breaking myriad taboos and shifting between social categories. For Radin, the representation of the breaking of taboos is shocking for participants, and provides a way of educating, shaping the trickster; joking, craziness, and inversions of social norms lead Wakdjunkaga to be "suddenly represented as a good citizen, as a thoroughly socialized individual" (1956: p. 139).

Many of his activities revolve around inverted and intense relations between sex and violence. Lowie records that the Assiniboine trickster dresses as a woman to marry a young man, and then fools him into thinking he is pregnant by putting a fox under his dress. The trickster breaks taboos by having sex before going on the warpath. The character shifts between being a man and a woman and seducing and having sex with men and women. The Winnebago trickster has a vulva at one point and gives birth to a child by a chief's son. Radin describes how the Winnebago cycle tells how the trickster's huge penis is separable from his body and slung over his shoulder. Subsequent commentators are struck by the ways that the trickster's penis is cut into pieces and made into human form (Hyde, 1998: pp. 28–31). Trickster chases the chipmunk and probes a hollow tree with his giant penis to get him. Chipmunk chews his mobile oversized detachable organ into a human sized one in the proper place on his body with the pieces of his penis cast onto the land and water to "make objects for human beings to use" like potatoes, turnips, and rice (Radin, 1956: p. 39). The Winnebago cycle ends as Wakdjunkaga makes the Mississippi River safe for people, killing and eating "beings that were molesting the people" (Radin, 1956: p. 52). He tells the waterfall to move stating "I am telling you that the earth was made for man to live on and you will annoy him if you stay here. I came to this earth to rearrange it" (Radin, 1956). After wandering the world and getting in various altercations with various people and animals – polecat, mink, woodpecker, coyote – he eats a final meal near the juncture of the Missouri and Mississippi rivers, his buttocks and testicles made impressions in the rock, and his kettle and bowl also remain. He then leaves this world. "Under the world where Earthmaker lives, there is another world just like it and of this world he, Trickster, is in charge" (Radin, 1956). Radin's analysis revolves around questions of morality and motivation probing the human psyche and the social world for the significance of the trickster as a representation of an archaic culture that gives insight into broader issues of humanity. Trickster makes the world safe and habitable for humans but often not out of altruistic interest in humanity but rather through capricious, unthinking action (cf Basso, 1987: pp. 4–8).

The analytic category of the trickster emerges out of the nineteenth century debates about the comparative study of Native American religions as they evolve through a concern with the moral implications and abstract symbolism of myths in so-called archaic societies. Daniel Brinton focused on "Hero Myths" arguing the goal of comparative mythology was to discover "the influence of the myth on the individual and national mind, on the progress and destiny of those who believed it, in other words, its true religious import" (Brinton, 1882: p. xi). Countering prevailing racist notions that "savage races" did not have religion he argued "what appears cruelty, puerility, or obscenity assumes a very different aspect when regarded from the correct, the native point of view, with a full knowledge of the surroundings and the intentions of the myth-makers themselves" (Brinton, 1890: p. 103). Others had argued that hero-gods were actually distorted tales of "chiefs of tribes at some undetermined epoch" and that "tribes" worshiped "external phenomena" and nature (Brinton, 1882: pp. vii & xii). Instead Brinton argues that the Native American hero-god "human in nature, benignant in character, loved rather than feared, and that his worship carried with it the germs of the development of benevolent emotions and sound ethical principles" (Brinton, 1882: p. xii). In the assumed evolutionary hierarchy of societies, studying "primitive peoples" focused on myths as precursors to more advanced theological thinking, which Brinton saw in animal myths, representations of the inanimate as animate and the irrational as rational (Brinton, 1882: p. 21). He compares mythic hero-gods from Native American religions like the Mexican Quetzalcoatl with "Old World mythologies," like those of the Greek Hermes, inventor of the alphabet, music, and the arts, messenger of the gods, and patron of thieves (Brinton, 1882: p. 31). He examines tales of the Algonkin Michabo, The Giant Rabbit, and the Iroquois Ioskeha. For Brinton, hero myths stem from the personification of natural phenomenon and while he argues Native American peoples are not as advanced as Western societies, they value "individual virtue" (Brinton, 1882: p. 235) and the "moral evolution of the nation proceeded steadily from its mythology" (Brinton, 1882: p. 239). Brinton uses Native American hero myths to argue the influence of Christianity and Europeans corrupted and destroyed rather than fostered Native American thought, morality, and virtue, which had its own particular evolutionary track. For Brinton, "the assumption of the psychic unity of mankind underlies all his work in the field of comparative religion and impelled him to argue for the spontaneous origin of religious parallelisms" (Weeks, 2002: p. 5).

Debates about the links between religion and morality and cultural evolution remained central to the study of mythic heroes. Brinton, countering established evolutionary racist theories which describe non-Western people as incapable of abstract thought, argues the selfishness of culture heroes is probably the result of "a gradual deterioration of a purer more primitive form of the myth, and that the more vulgar tales are later additions" (Boas in Teit: p. 4); that is, ancient altruistic sacred tales of creation are layered with more recent profane elements of selfish egotistical behavior.

In the introductory essay to James Teit's 1898 *Traditions of the Thompson River Indians of British Columbia*, Franz Boas mulls over two aspects of "transformer myths" namely the seeming contradiction between the altruism of culture heroes and the selfishness of tricksters that in many tales are, in fact, manifest in the same character. Often, these supernatural human-animals only aid humanity by accident. Humans, in fact, appear incidental to the protagonists of the stories. To folklorists and analysts of comparative religions this is a problem as it contradicts assumptions about moral being.

The lack of concern for humanity in hero myths is a central issue; interpretation revolves around the seeming moral contradictions between a culture hero and the selfish-buffoon. Franz Boas and his students argued that the culture hero is common in Native American myths, describing a "prehistoric era" when "there was no clear distinction between man and

animals" (Boas in Teit: p. 4). The culture hero "gave the world its present shape, who killed monsters that infested the land, and gave man the arts that make life worth living" (Boas in Teit: p. 4). But these figures are not altruistic or allies of humanity. Boas takes an opposing view to Brinton, arguing that the altruistic side of transformer myths develops "with the progress of society. …The higher the civilization of the tribe, the more sharply, it seems, is the line drawn between the culture hero and the trickster" (Boas in Teit, 1898: p. 10). The more developed the altruistic sensibility is, the more there is a contradiction between tricksters and culture heroes splitting the transformer in two. This contradicts Brinton's idealized view that civilization and contact corrupt primitive society and argues for a historical understanding of myth that acknowledges the interplay between the mutually constituted development of ritual and myth and the natural environment and the adoption of various "foreign" tales.

For Boas, detailed historical analyses of myths give insight into outside influences, showing how "foreign and indigenous material is interwoven and worked into a somewhat homogeneous fabric" (Boas in Teit: p. 18). Here, understanding the origins of transformer tales is important because these stories reflect how a culture refracts environment, geography, social organization, and kinship. For Boas, discrepancies between rituals and myths can reveal "survivals," imported images and stories from other cultures and older traditions, giving insight into changing social conditions. Examining historical change reflected in myth can yield an accurate understanding of cultures. As myths are reflections of archaic thought and the ways that people make sense of nature and the changing social conditions in which they find themselves. Boas' interest in myth is a Hegelian lament about something that appears only in its erasure: "It would seem that mythological worlds have been built up, only to be shattered again, and that new worlds were built from the fragments" (Boas in Teit: p. 18). Franz Boas and his students argue for the study of material and symbolic aspects of life in order to form rich ethnographic descriptions of a culture in context. This is a critique of older forms of ethnographic inquiry that sought to compare elements of societies based on surface similarities and assumptions about racial evolution. Instead Boas contends meaning of signs and objects can only be understood through situating specific elements within a culture and its internal logics. Only through the examination in detail of particular cultures can the comparative relevance of social analysis be realized and turned into more abstract generalizations. Studying Native American cultures was a way to record more "primitive," disappearing cultures that, because of their evolutionary proximity to "our" origins, offer comparative insight into the nature of humanity. At the root of the Boasian interest in the culture hero, transformer, and trickster was a concern for the specific epistemological relations between myth and culture.

For Radin, as for his teacher Boas, the trickster has a "twofold function of benefactor and buffoon," which is the "outstanding characteristic of the overwhelming majority of trickster heroes wherever they are encountered in aboriginal America" (1956: p. 124). Wakdjunkaga is significant in preparing the world for humanity and in mediating social being, though in the process can never fully inhabit this world as he straddles the supernatural, the mythic, the archaic, and the lived world. The development of social awareness and self-consciousness is central. While on one level the selfishness and violence of the trickster is amoral and unaware, it, in fact, produces the world through the creation of awareness; his actions constitute self-sacrifice as a form of world-making. Wakdjunkaga transgresses norms and tempts and tricks others, in the process creating social boundaries and categories. Radin points out that the trickster entails multiple interpretations among Winnebago, and its tales reflect social and historical change and the influence of multiple outside sources. Though ultimately the tales are interpreted as about creating and reproducing a normative idea of social life and of good citizenship.

Ultimately for Radin, the trickster is the symbol of an "Archaic and primordial past, where there as yet existed no clear-cut differentiation between the divine and the nondivine … His hunger, his sex, his wandering, these appertain neither to the gods nor to man. They belong to another realm, materially and spiritually, and that is why neither the gods nor man known [sic] precisely what to do with them" (Radin, 1956: p. 168). The trickster in this guise is beyond good and evil, does not fit properly into any time, any type of being, god, human, or animal, or any social category or place. He is inside and outside of each person, responsible for making the lived world but not able to fully inhabit it (Radin, 1956: p. 169).

Radin's account of the Winnebago trickster shapes many scholarly understandings of the character. For Jung the trickster, as Radin outlines him, is a mythical figure that represents and comes out of our inner psychological order. "In his clearest manifestations he is a faithful copy of an absolutely undifferentiated human consciousness, corresponding to a psyche that has hardly left the animal level" (Jung in Radin: p. 200). Jung maps psychic development of humanity onto the evolutionary development of cultures seeing the Winnebago trickster as a window into the primitive mind, which is the root of all human psychology but more clearly accessed in the so-called primitive peoples. He links this trickster to other characters and rituals of inversion in medieval Europe such at the fools' feast. The trickster is manifest in the shape-shifter, poltergeist, and shaman. For Jung, the trickster is "an archetypal psychic structure of extreme antiquity … a faithful copy of an absolutely undifferentiated human consciousness, corresponding to a psyche that has hardly a primitive or barbarous consciousness … a picture of itself on a much earlier level of development" (Jung in Radin: p. 200). But, for Jung, the presence of the trickster in numerous figures in contemporary life is not simply a holdover; it has a function. The trickster, as with all archetypes and psychologically relevant figures, is a way the psyche is speaking about itself. It provides a form for reflection. The trickster in some ways recalls the savior figure. As people move forth from the archaic origins of the trickster myths, the trickster provokes self-consciousness. He is like a shadow of both the collective and the individual, a negative figure, a reflection of consciousness though one that is contradictorily highly unconscious because of his unstable proximity to the self. For Jung, ultimately the trickster plays a role in the liberation of the individual and collective from unconsciousness. The shadow represents the potential of its opposite, that is the development of self-consciousness (Jung in Radin: p. 211). In the trickster myth, we find an assessment of "everything that a man can

never get the better of and never finishes coping with" (Jung in Radin: p. 210).

Native American trickster tales as told by Boasians appear as abstracted texts, readable as literature or myth. They are meant to reflect and go beyond the histories and contexts in which they are collected and interpreted. They contain fragments pieced together of more archaic tales and characters and cultures, and provide clues to the nature of the human mind. The Boasians critique older understandings of Native Americans as either devilish or morally pure by insisting on understanding material, narrative, and symbolic forms in cultural and historical context. But they also make assumptions about the boundedness of cultural groups and make attempts to correct for historical influences to find "true" cultural accounts that will give insight into fundamental human traits by accessing more primitive aspects of life. While the Boasians critique evolutionary ideas of culture, their examinations of trickster tales are shaped by assumptions about an evolutionary hierarchy of cultures that place native peoples closer to archaic humanity.

While Radin and Jung consider the trickster an inverted marker of the structuring principles of psychic and social order, Ellen Basso examines trickster and creation myths of the Kalapalo Indians in Brazil to question the notion of unity of the psyche, the personality, and social order itself. She explores the significance of storytelling as performance in the interpretation of these tales. She focuses on language use in performance to show how Kalapalo ideas of deception are rooted in the contingencies of the lived rather than a cultural unity of myth. Critiquing Radin she explains that the trickster's morality is not what is most significant but rather his immanence, that is "the consequences of the ability of a trickster (or a person) to act inconsistently and experience inconsistencies, in other words, to actively participate in dissolving or scattering the self in order to achieve a new personal understanding, a new identity" (Basso, 1987: p. 7). Taugi is the main male Kalapalo trickster who adapts the environment to make it habitable by people. This creation myth is built upon "a narrative accumulation of the various feelings and powers of different mythic personages who make up the human pedigree. These feelings and powers, passed down from one generation of Dawn People to the next, culminate in the blossoming forth of humanity in all its peculiar, passionate complexity" (Basso, 1987: p. 9). The less common female trickster is Nafigi. The male and female tricksters are according to Basso make a complementary yet opposing pair (Basso, 1987: p. 225). Taugi mocks and resists "signs of relative social competence and power" with which men struggle (Basso, 1987: p. 225). These include concern with wealth, the use of language and sarcasm, the ability to seduce women, insatiable female lust, the achievement of physical perfection, and the scorn of friend and family (Basso, 1987). The female trickster is a "sexually compulsive vampire in her relations with human beings" (Basso, 1987: p. 225). She refracts female concerns with being seen as gossiping and deceptive and as seductive of men. Female tricksters are not usually playful or deceptive but instead reflect the opposite of how women are seen. Most significantly, Basso argues that trickster tales must be understood not in static texts or formal types but rather through language and performance, in "the speech-centered nature of a trickster's activities" (Basso, 1987: p. 229). The trickster is significant as an embodiment of performance and contingency that must be reflected in methods of analysis, whereas conversely, for Boas and Radin, the relevance of the trickster character is contained within myths as detachable, transportable texts. But for both Radin and Basso, the trickster is a form of self-consciousness, a contingent, ambiguous way that a society reflects collective and individual selfhood back at itself.

The trickster plays a small but significant role in Claude Levi-Strauss's development of the structuralist analysis of myth in that the notion of trickster as social mediator becomes a way to argue that myth should be understood through tensions between polar oppositions in signs and categories. Levi-Strauss uses material on North American tricksters as one example in his groundbreaking paper "Structural Study of Myth" (1955). For Levi-Strauss myths hide and distort their true meaning. His structuralism aims to bring order and reveal "some basic logical processes which are at the root of mythical thought" (Levi-Strauss, 1963: p. 224). Whereas Radin examined the narrative structure, character traits, and ordering of myths to find their cultural logic and sense, Levi-Strauss disassembles them, ignoring their surface logics and moralities in search of hidden, deeper realities. Myths are texts that must be broken into constituent parts and reorganized to uncover the more profound grammatical categories not tied to narrative order or proximity in the telling or performance of the myth that will point to their meaning. For him myths are the opposite of poetry. If poetry relies upon language use for meaning and aesthetic sense, myths must be abstracted from language – from any particular enactment – to break them into their significant units and rebundle these parts to get at myths' hidden meanings. Levi-Strauss's concerns with myth, order, and the human mind reimagine the significance of the trickster. The trickster is a mediator caught between uncontrolled consumption and the production of the social world (Carroll, 1981: p. 307). The structuralist logic of the trickster comes as "his mediating function occupies a position halfway between two polar terms, he must retain something of that duality – namely an ambiguous and equivocal character" (Levi-Strauss, 1963: p. 226). Levi-Strauss uses the trickster to demonstrate a fundamental way that signs work in a structuralist system. For Levi-Strauss, the trickster is a mediator between nested pairs of polar oppositions, life and death, agriculture and war, herbivorous and carnivorous animals that become triads. Tricksters are carrion eaters, Ravens and Coyotes, that mediate between herbivores and carnivores. Mediators of primary semiotic oppositions generate oppositions on the next order. Tensions between opposing pairs are played out in mediating signs, which contain that tension and direct it to the next level (Levi-Strauss, 1963: pp. 224–225). Tricksters embody this dialectic. Levi-Strauss's reuse of trickster myths in developing structuralism's logic shows the enduring analytic significance of this category.

African Tricksters, Transgression, and Social Order

As in Native American studies, researchers in Africa in the first half of the twentieth century collected trickster tales as folklore and religious representation. Various African mythic characters have been fit in to the category of trickster – those that blur

social categories, inhabit spaces of moral ambiguity, mediators living at the crossroads of humans and deities. Early Africanist anthropologists seem to have had trouble fitting tricksters into their more systematic research, often collecting stories in separate volumes from their major works of social analysis. Evans-Pritchard's *Azande Trickster* is a collection of stories about Ture the spider, the Azande trickster that follows on his famous *Witchcraft, Oracles, and Magic Among the Azande*. Melville Herskovits collects folklore as part of his more well-known research in Dahomey. R.S. Rattray publishes his collection of Ananse spider trickster tales after his influential studies of Ashanti (Asante) society and religion. Robert Pelton's reworking of classic ethnographic and folkloric materials of Herskovits, Rattray, Marcel Griaule, and various scholars of Yoruba life in *The Trickster in West Africa* has been a major influence on how the trickster figure has been understood. In close readings of trickster tales in relation to West African cultural cosmologies, he argues that the trickster brings social relations into discourse through humor and linguistic and spiritual mediation. Pelton places previously collected tales from a variety of contexts into dialogue to paint a picture of the trickster character. In turn, his textual readings are used by later scholars seeking the trickster's importance. Pelton portrays African cultures as unchanging timeless units and uses the trickster to define what he sees as a shared West African cosmological sensibility.

Whereas Native American analyses focus more on making sense of the moral and cultural logic of tricksters, the study of African tricksters emphasizes their roles in the making and unmaking of social structure through a Durkheimian–Turnerian idea of ritual progression through a liminal phase to symbolic reaggregation and affirmation. They have been taken as embodiments of social mediation, particularly in their control of language, communication, and forms of exchange. Their doubleness and lived contradictions have also been seen as markers of the creative logic of African societies portrayed not only in collected stories but in their contextually specific performances. Scholar of religion, Mircea Eliade, and anthropologist of ritual, Victor Turner, have been influential in shaping understandings of African tricksters as liminal figures, whose symbolic inversions and excesses provide ritual signs of antistructure than in the final instance point to and reproduce normative social values (Douglas, 1970; Babcock-Abrahams, 1975). Ethnographers of African tricksters have paid attention to the poetics of their expression finding meaning in the disjunctures between action and explicit textual intent.

African tricksters also raise questions for Western scholars about morality, who seem perplexed by how characters doing immoral things can be mythic, beloved foundational cultural icons. Rather than resolving contradictions between morality and immorality, tricksters in the process of provoking conflict, lying, being sexually aggressive, stealing, and scheming, in fact, create social relations. African tricksters are manifestations of self-consciousness and social reflexivity. They build the world through producing debt and desire (Vecsey, 1981). Tricksters link "disruptive and recreative possibilities" (Pelton, 1980: p. 128). Conflict and contradiction create sociality, contrary to scholars who worry that the trickster's desires and selfishness are antisocial.

In relation to language, Beidelman points out that for the Kaguru of Tanzania, "Trickery is often expressed through deviant speech patterns. For example, a narrator will portray a hyena's speech as slurred and bumbling, an old woman's as hoarse and grotesque, and a hare's or youngest sibling's as high and lilting. In discerning relations between words and meanings, there are profound ambiguities and difficulties involving misunderstandings and deceit and varying degrees of eloquence and guile" (Beidelman, 1980: p. 31).

Unlike the Winnebago trickster tales, which constitute a cycle, *Anansesem* are an indeterminate set of tales "whose inner principle of order is somehow Ananse himself" (Pelton, 1980: p. 30). Ananse is the spider trickster of the Akan people of Ghana. R.S. Rattray, the first official anthropologist in the Gold Coast colonial service and the preeminent ethnographer of the Asante collected *Anansesem* to accompany his other ethnographic research on religion, politics, and society. Even tales that do not feature Ananse explicitly are included in this genre of folktales. Ananse has human and animal characteristics and the stories he both tells and inhabits shift between mythic time, mundane social settings, and relations with Nyame the Supreme God. He is an eminently social being; his stories revolve around his attempts to dupe those around him into relinquishing wealth, food, sex, and value. This usually involves word play and the complex formulation of tricky plots. While at times he succeeds, usually Ananse is foiled by his own desires and cleverness.

One well-known Ananse story demonstrates his relationship to language, transactions, and social mediation is the tale of Ananse and the pot of wisdom. Its multiple variations that point to the flexibility of the storytelling structure. In one version Ananse aspires to be chief. He has clandestinely collected all of the world's wisdom and carries it in a pot on his head. He looks for a place to hide it so that he alone will have wisdom to outsmart anyone. He ties the pot to his stomach in order to climb a tree and hide it. Because of the awkwardness of the pot, he is unable to climb. If he drops the pot to pull himself up, he will loose all the wisdom but if he does not then he cannot move. Ananse's son finds him stuck in this dilemma and suggests that he place the pot behind him and then climb up to hide it. Having received this advice from his own son, Ananse realizes he does not have all of the wisdom in the world. Ananse, humiliated by being outwitted by his own son, drops the pot, which breaks, spilling the wisdom everywhere. Forevermore, he lives in shame in the tree, which is why spiders live in trees.

In another version that varies Ananse's intents, emotions, and motivations, Ananse begs or tricks God to give him wisdom to share with the world. Ananse is given the wisdom to share with everyone but decides to keep it himself and hide it so no one can get it. As he is climbing the tree with the pot tied around his waist, he gets stuck and his child advises him to remove the pot and climb up. In anger at being embarrassed and outwitted by a young child he smashes the pot and wisdom is scattered everywhere.

Tricksters are highly mobile figures of cultural self-representation and signs of history. By this I mean that rather than remain characters in mythological folkloric tales they are appropriated by various ongoing national, economic, and artistic projects. Indeed in urban cosmopolitan African worlds defined by business people, politicians, and artists, actions and intents are often explicitly debated and discussed as tricksters (Shipley, 2009; Hannerz, 1987). In Kinshasa, as White shows,

the trickster manifest in performances and discourses of popular urban music, shapes a particularly African modernity (White, 1999: p. 158). In Ghana, Ananse and trickster storytelling become the basis of first an eclectic modern variety theater known as concert party theatre and later the National Theatre Movement. Around Ghana's independence from British colonial rule in 1957, Prime Minister Kwame Nkrumah and various playwrights, artists, and scholars identified *Anansesem* as a strong, indigenous cultural tradition that could ground the new nation's attempts to develop modern artistic forms. Efua Sutherland one of the founders of Ghana's National Theatre Movement in the 1950s and 1960s coined a term for a new genre Anansegro, meaning Ananse theater, to adapt the essence of Ananse's cultural wisdom and formal logic for the modern stage (Sutherland, 1975). Her best-known play, "Marriage of Anansewa" tells how Ananse devises a scheme to become wealthy by promising his daughter in marriage to three prominent chiefs. When they all plan to marry her the same day he convinces her to feign her death. This was widely seen as a metaphor for the way that Kwame Nkrumah tried to balance political allegiance and court favor between Cold War powers in the 1950s and 1960s. This play demonstrates how Ananse trickster stories are appropriated by national cultural projects that maintain the importance of this mediating figure and performative forms of indirection and double-voicing long significant in Akan and West African communication and language use (Kwesi Yankah, 1995). The recent boom in film industries in Ghana and Nigeria also appropriate morally ambiguous trickster figures as a key narrative feature. Hip-hop artists and popular musicians have also adopted Ananse tales in their lyrics and poetic forms, with one artist even taking the stage name of Kweku Ananse. Ananse stories while describing particular aspects of Akan sociality have been explicitly appropriated to define a Ghanaian cultural nationhood and transposed to address aspects of modern urban life in Ghana (Shipley, 2009, 2013, 2015).

Ananse is a mythical human-animal-spiritual being, the teller and main character in stories, a humorous hustler; but other West African tricksters like Legba of the Fon and Eshu of the Yoruba are more firmly tied to divinity. Legba, like Ananse, is a master of language, a mediator of social relations. Shrines to him are prominent at the entrance to houses and villages in Dahomey. He is the youngest child of Mawu-Lisa, the supreme being of the Fon of Benin. Legba mediates between their mother Mawu, the female aspect of the divinity, and the various realms ruled by his older siblings. Pelton quotes Herskovits: "Legba knows all the languages known to his brothers, and he knows the language Mawu speaks, too. Legba is Mawu's linguist. If one of the brothers wishes to speak, he must give the message to Legba, for none knows any longer how to address himself to Mawu-Lisa. That is why Legba is everywhere" (Pelton, 1980: p. 73). He is also the patron of divination. Fa is his master though Legba must speak for Fa. Legba is key to "Fon self-understanding" through initiation rituals and through "the special language that is their divination" (Pelton, 1980: p. 112). Legba lives among people, though his transgressions and access to power make him dangerous to bring into the house. Legba is a sign of male potency, and Legba masquerades dance as part of initiation rituals. "Legba's sexuality is the great symbol of his transforming power, not because…of legitimizing promiscuity, but because it forces responses that reveal possibilities hidden with the existent world" (Pelton, 1980: p. 88). But he is a creator and facilitator not a producer, a mediator who has access to power rather than producing power (Pelton, 1980). Legba is present when there is social risk in the market, in relation to raising children, in royal and political administration where he both causes and neutralizes problems; he is there "to rechannel far more than to neutralize, the various forces present in such situations" (Pelton, 1980: p. 83).

Legba, like the other West African tricksters, Pelton examines, "lives where separate worlds meet and can move back and forth between them" (Pelton, 1980: p. 88). Legba is closely related to the Yoruba Eshu. In the Yoruba pantheon of orisha deities, Eshu (Esu, Elegba, Elegbara) is the trickster, a linguist and a mediator between the human world and the gods. Eshu is also present in markets and at crossroads, places of transition, translation, and mediation. Eshu, like Ananse and Legba, is a troublemaker the "embodiment of paradoxes" (Pelton, 1980: p. 129). "When men are quarrelling Elegba is said to be present; when a man has done wrong he is asked if Elegba prompted him" (Wescott, 1962: p. 337). He is an "agent provocateur," "a messenger of the gods," and "a homeless wandering spirit," who inhabits "the market-place, the crossroads, and thresholds of houses" (Wescott, 1962: p. 337). Eshu is also key to Yoruba divination; he "receives a portion of every sacrifice commanded through divination" and "makes his 'presence and power' known each time a diviner holds a session" (Pemberton in Pelton, 1980: p. 129). Eshu is actively present in Yoruba life through sacrifice and divination.

Ananse is a playful schemer, Legba is forgiving and mischievous, and Eshu is vengeful and bold (Pelton, 1980: p. 130). All have huge appetites for sex and valuable things and activities. They desire to be at the center of social action, provoking conflict and producing excitement. While Ananse is motivated by his self-interest or sometimes the pleasure of trickery, Eshu and Legba at times seem to go out of their way to cause trouble for the sake of it. One well-known story with many iterations tells of how Eshu creates conflict between two friends who own farms next door to each other. Eshu wears a multicolored cap and holds his pipe in front of him at his neck and his staff behind him and walks the boundary between the farms. The friends fight about what color his hat was and which way he was going (Frobenius, 1968; Bascom, 1969; Pelton, 1980). Here, Eshu's power comes from the ability to fracture the idea of a single truth into multiple perspectives. Scholarship on these West African tricksters has focused on them as powerful, transgressive, and ambivalent figures, who represent cultural, cosmological logics.

Trickster as Response to Inequality

Studies of tricksters as mythological figures in so-called traditional societies and debates about structuralism, functionalism, archetypes, and neo-Durkheimian religious affirmations ebbed in the early 1980s. Though over the next decade, a new trickster popped up in a number of studies in which older research is reworked to make arguments about power and the global, comparative relevance of the character.

Some of these studies use the trickster to authorize the return to older comparative religions and folkloric projects from the early twentieth century. Others reimagined the trickster in a fundamentally different way than older studies. Rather than providing insight into a culture's cosmology or the psychic unity of the human mind, the trickster became a historical figure, a sign of agency for oppressed social groups and subcultures. Here the trickster is a multivocal character who navigates hostile terrains, speaking back to power, opening up subtle modes of communication and agency for the powerless. Scholars interested in this version of the trickster have focused on the trickster's ability to mediate between realms, to hustle, to aspire, to survive, and to address complex social situations through humor. In the 1980s and early 1990s, trickster scholarship delineated, in effect, a theory of resistance. Reexamining Native American and African tricksters and their contemporary manifestations in Native American literature and culture, African postcolonial worlds, and African diasporic narratives and religion in North America, Haiti, Brazil, Cuba, and beyond scholars have identified a fundamental duality and double-voicing that provides a form of cultural resistance and expression for peoples living under conditions of inequality.

In his 1988 book *Signifying Monkey*: *A Theory of African-American Literary Criticism* literary critic Henry Louis Gates argues Signifyin(g) is the preeminent trope of African-American language use and can be a metaphor for organizing black literary criticism. He focuses on Esu [Eshu] as a key figure of social destruction/production and linguistic mediation in Yoruba cosmology and divination, a trickster who comes with African peoples to the New World through the slave trade and is present in African disaporic religions and ways of thinking. Gates focuses on the "uncertainty or indeterminacy" (Gates, 1988: p. 32), duality, "double-voiced" aspects (Gates, 1988: p. 7), and "meta-linguistic" principles (Gates, 1988: p. 35) of West African trickster discourse manifest in the figure of Esu. As divination is a form of sacred interpretation, translation, and transformation, Esu transports and translates these practices as Pan-African modes of thinking, speaking, and being (Gates, 1988: p. 9). Gates argues that Esu is the patron of literary criticism. "His passage from Africa to the New World and his continued simultaneous existence in the Yoruba-derived cultures of Nigeria, Benin, Cuba, Haiti, Brazil, and the United States allows us to draw upon this compelling black figure as the trope of critical activity as a whole. Whereas Ogun, god of war and the artisan, stands as the writer's muse, it is Esu who stands as the muse of the critic. He can serve us as such because he is both the divine linguist and the divine interpreter, the controlling principle of its representation and its interpretation" (Gates, 1988: p. 35). Gates is interested in "Esu's double-voiced representation in art" (Gates, 1988: p. 51) and the specific historical connections between West Africa and Yoruba practice and black diasporic expression. He compares Esu to African-American forms of Signifyin(g) and double-voicing. Esu becomes a literal and metaphoric way for Gates to talk about black linguistic practices and modes of critical thought. Tricksters and those living under conditions of oppression speak in double-voiced ways such that they say two things at once; the explicit meaning of a statement covers and refracts an indirect, deeper meaning. Pelton and Gates both point to the importance of African tricksters' humor and double voicing as modes of engaging profound philosophical issues for African and African diasporic peoples. Whereas for scholars of African expression, indirection represents a fundamental aspect of West African cosmology, for Gates and scholars of resistance, this is a practice that is developed as African diasporic peoples and other oppressed groups live with intense inequality and violence.

James C. Scott's work which since the 1980s has shaped understandings of resistance in relation to culture and politics, is built upon a metaphoric trickster. His argument in the influential *Domination and the Arts of Resistance* points to African-American trickster tales as strategies of resistance for people living in an oppressive society that determines modes of narrative and spatial relations. He argues "there is a third realm of subordinate group politics" strategically situated between the self-image of elites and the "hidden transcripts" of the oppressed that are voiced in safe private spaces where they can speak about anger or revenge (Scott, 1990: pp. 18–19). The arts of resistance, however, demarcate "a politics of disguise and anonymity that take place in public view but is designed to have a double meaning or to shield the identity of the actors. Rumor, gossip, folktales, jokes, songs, rituals, codes, and euphemisms – a good part of the folk culture of subordinate groups – fit this description. As a case in point, consider the Brer Rabbit stories of slaves, and trickster stories more generally. At one level these are nothing but innocent stories about animals; at another level they appear to celebrate the cunning wiles and vengeful spirit of the weak as they triumph over the strong" (Scott, 1990: pp. 18–19). Here folklore and folk characters give insight into politics and power rather than standing as timeless cultural myths.

While early ethnographers tried to morally reconcile the trickery, humor, and selfishness or tricksters with their heroic and productive capacities, for Scott the trickster's two-sidedness reflects cultural reactions to social inequality. For Scott culture makes sense primarily in relation to struggle. "Nothing illustrates the veiled cultural resistance of subordinate groups better than what have been terms trickster tales … Typically the trickster makes his successful way through a treacherous environment of enemies out to defeat him – or eat him – not by his strength but by his wit and cunning." Trickster escapes by "knowing the habits of his enemies, by deceiving them … the structural position of the trickster hero and the *strategemx* he deploys bear a marked resemblance to the existential dilemma of subordinate groups" (Scott, 1990: pp. 162–163).

Gates and Scott point to the trickster in culture and society, as a figure navigating internal power dynamics shaped by historically structured, unequal relations among cultural groups. Here trickster myths do not stand for fundamental cosmologies or cultural worldviews but reflect struggle and how people without direct access to power use indirect means to claim agency in the midst of violent conditions that continue to hold them down.

In recent years, scholars and popular writers have built on James, Gates, and others to emphasize the potential of the trickster as a general metaphor and social type, in some ways returning to older ideas of archetype. Popular and folkloric writers such as Hyde and Scheub take the trickster around the

globe as a way to think about universal humanity and narratives of creativity. Studies such as these bring together a diverse set of European, African, Native American, and Asian tricksters from older ethnographic and folkloric texts, fiction, politics, and the arts. Scheub takes the trickster character found in oral literature and mythology around the world. He focuses on the temporality of narrative, arguing the trickster stands for a moment of transformation. In this comparative argument heroes are always a bit of the trickster. He provides an impressive, several page list of tricksters from all over the world. For Scheub "Tricksters are the timeless energy, the eternally liminal, the ordering and the chaotic … They are primordial, now sublime and now debased…. We see ourselves in the trickster…. Whenever there is chaos and order, the trickster is there. And for a time, the hero is a trickster, the tale character is a trickster, God plays the role of trickster" (Scheub, 2012: p. 12). Here moments of creative possibility and of heroic (salvation) are always tied to their potential inversions. He argues the trickster is the "fundamental building block of storytelling" that contains with it the "seed of transformation" key to all stories (Scheub, 2012: p. 24). For Hyde, the trickster is also a global figure that points to creativity in the face of uncertainty. Tricksters are travelers and inventors. Structurally, his argument relies on trickster as universal archetype, if in a more functional one than the Jungian version. He finds tricksters everywhere, as writers, painters, political figures, across time and space; at times the category is literal at other times metaphoric. For Hyde the trickster is a romanticized marker of hope, representing a mode of expression and a form of mobility and inversion defined by the potentials of the individual in the face of power. He is not concerned with whether or not the figures he identifies as tricksters have been seen as such in the contexts of their relevance but in arguing for universal idea of human creativity.

Reappropriating Tricksters

While Paul Radin's work on the Winnebago trickster focused on rich explication of a specific trickster cycle and making sense of him in cultural context, his work was influential on a later generations of scholars searching for irreverence and power. The opening paragraph of the preface to the 1956 edition of *The Trickster* anticipates the ways that scholars will elaborate upon this character as a universal archetype:

> Few myths have so wide a distribution as the one, known by the name of *The Trickster*, which we are presenting here. For few can we so confidently assert that they belong to the oldest expressions of mankind. Few other myths have persisted with their fundamental content unchanged. The Trickster myth is found in clearly recognizable form among the simplest aboriginal tribes and among the complex. We encounter it among the ancient Greeks, the Chinese, the Japanese, and in the Semitic world. Many of the Trickster's traits were perpetuated in the figure of the mediaeval jester, and have survived right up to the present day in the Punch-and-Judy plays and in the clown. Although repeatedly combined with other myths and frequently drastically reorganized and reinterpreted, its basic plot seems always to have succeeded in reasserting itself.
>
> Radin, 1972: p. xxiii

Considering his mentor Boas's careful focus on detail and reticence to make generalizations, and his criticism of Radin for rushing to make assertions, it makes sense that this statement comes after Boas's death. Radin seems freed to highlight the social power of transgression, arguing that trickster's inversions can function as satires to voice protest "against the many, often onerous, obligations connected with the Winnebago social order and their religion and ritual" (Radin, 1956: p. 152). In adopting the trickster as a mediator of culture and power that bridges lived worlds and analysis, Scott and scholars of resistance and Hyde and scholars of creativity have lost the specificity of social and historical conditions of inequality, reducing culture, difference, and complex forms of subjectivity to a seductive, preexisting analytic category.

Trickster characters have been good to think with and reveal various scholarly trends in how writing about them has changed over the course of the last century. They show an ongoing scholarly fascination with the seeming contradictions between uncontrolled desire/transgression and social totality, deviance and psychic unity. Scholars seem anxious to reconcile appetites for sex, food, and conquest with logic, rational thought, social good, and totality.

Scholarly accounts of tricksters abound and seem to get more romanticized over time. Analysts of myth and literature find in trickster stories ways to explain profound philosophical and psychological conundrums around human creativity, ambivalence, altruism, chaos, and desire. When the category dissolves into a discomforting mess of particularity, its elements resurface elsewhere. But there are certainly a set of tales and characters from around the world that show an uncanny resemblance to each other when taken from abstracted texts. Trickster tales are sometimes origin stories, tales of the creation of the social order. They can resemble tales of saviors. They can also be tales of chaos and subversion. Tricksters are fundamentally dual in a number of ways and embody and play out various social contradictions. They live in ambiguous temporal-spatial planes and shift between worlds rapidly in mid story. Their tales do not always go in chronological order or follow simple narrative progressions. Their stories often move between mythic time and lived worlds. They inhabit hybrid and changing bodies. They are animals, human-animals, and human-animal-supernatural beings. They are greedy, arrogant, liars, selfish, destructive, devious. They are productive, improvisational, unpredictable creators of culture, heroes and antiheroes.

The trickster's omnipresence and numerous reinventions seem indicative of the continuing need in ethnographic, religious, and literary inquiry to find coherence and commonality especially in the face of contradictions. The category of trickster has been a way that anthropologists have theorized the relationships between individual morality and agency and social organization and power, in the process rethinking how ideas of culture relate to historical and mythic frameworks.

Bibliography

Abrahams, Roger D., 1966. Some varieties of heroes in America. Journal of the Folklore Institute 3, 342.

Abrahams, Roger D. (Ed.), 1983. African Folktales: Traditional Stories of the Black World. Pantheon, New York.

Arnold, A. James (Ed.), 1996. Monsters, Tricksters, and Sacred Cows: Animal Tales and American Identities. University of Virginia Press, Charlottesville.
Babcock-Abrahams, Barbara, 1975. "A talented margin of mess": the trickster and his tales reconsidered. Journal of American Folklore 2, 147–186.
Bakhtin, Mikhail, 1984. Rabelais and His World. Indiana University Press, Bloomington.
Ballinger, Franchot, 2004. Living Sideways: Tricksters in American Indian Oral Traditions. University of Oklahoma Press, Norman.
Bascom, William, 1969. Ifa Divination: Communication between Gods and Men in West Africa. Indiana University Press, Bloomington.
Basso, Ellen B., 1987. In Favor of Deceit: A Study of Tricksters in an Amazonian Society. University of Arizona Press, Tucson.
Beidelman, T.O., 1980. The moral imagination of the kaguru: some thoughts on tricksters, translation, and comparative analysis. American Ethnologist 7, 1.
Boas, Franz, 1898. Introduction. Traditions of the Thompson River Indians of British Columbia. By James Teit. Houghton, Mifflin, and Company, Boston.
Brinton, Daniel G., 1882. American Hero-Myths: A Study in the Native Religions of the Western Continent. A. C. Watts & Co., Philadelphia.
Brinton, Daniel G., 1890. Essays of an Americanist. Porter & Coates, Philadelphia.
Carroll, Michael P., 1981. Levi-Strauss, Freud, and the trickster: a new perspective upon an old problem. American Ethnologist 8 (2), 301–313.
Carroll, Michael P., 1984. The trickster as selfish-buffoon and culture-hero. Ethos 12 (2), 105–131.
Davis, Natalie Z., 2006. Trickster Travels: A Sixteenth-century Muslim between Worlds. Hill and Wang, New York.
Douglas, Mary, 1970. Purity and Danger: An Analysis of Concepts of Pollution and Taboo. Penguin Books, London.
Eliade, Mircea, 1961. Images and Symbols. Sheed and Ward, New York.
Erdoes, Richard, Ortiz, Alfonso, 1998. American Indian Trickster Tales. Viking, New York.
Evans-Pritchard, E.E., 1967. The Zande Trickster. Oxford University Press, London.
Frobenius, Leo, 1968[1913]. The Voice of Africa. Benjamin Blom, New York.
Gates, Henry L., 1988. The Signifying Monkey: A Theory of African-American Literary Criticism. Oxford University Press, New York.
Grottanelli, Cristiano, 1983. Tricksters, scape-goats, champions, saviors. History of Religions 23 (2), 117–213.
Hannerz, Ulf, 1987. The world in creolisation. Africa: Journal of the International African Institute 57 (4), 546–559. Sierra Leone, 1787–1987.
Hurston, Zora N., 1935. Mules and Men. Harper Perennial, New York.
Hyde, Lewis, 1998. Trickster Makes This World: Mischief, Myth, and Art. Farrar, Straus, and Giroux, New York.
La Barre, Weston, 1970. The Ghost Dance: Origins of Religions. Doubleday & Company, Inc., Garden City, New York.
Levi-Strauss, Claude, 1955. The structural study of myth. Journal of American Folklore 68 (270), 428–444.
Levi-Strauss, Claude, 1963. Structural Anthropology. Basic Books, New York.
Lindenbaum, Shirley, 2002. Fore narratives through time: how a Bush spirit became a robber, was sent to jail, emerged as the symbol of Eastern Highlands province, and never left home. Current Anthropology 43 (Suppl.), S63–S73.
Lowie, Robert H., 1909. The hero-trickster discussion. Journal of American Folklore 22 (86), 431–433.
Lowie, Robert H., 1910. Issues 1–2. The Assiniboine. Anthropological Papers, vol. 4. Order of the Trustees, American Museum of Natural History, New York. Esp. pp. 99–134.
M'Baye, Babacar, 2009. The Trickster Comes West: Pan-African Influence in Early Black Diasporan Narratives. University Press of Mississippi, Jackson.
Owomoyela, Oyekan, 1997. Yoruba Trickster Tales. University of Nebraska Press, Lincoln.
Pelton, Robert D., 1980. The Trickster in West Africa: A Study of Mythic Irony and Sacred Delight. University of California Press, Berkeley.
Radin, Paul, 1926. Winnebago Myth Cycles: Their Significance and Function. Part I: The trickster cycle.
Radin, Paul, 1948. Winnebago Hero Cycles: A Study in Aboriginal Literature. Issues 1–5. Waverly Press, Baltimore.
Radin, Paul, 1972[1956]. The Trickster: A Study in American Indian Mythology. Schocken Books, New York.
Rattray, Robert S., 1927. Religion and Art in Ashanti. Clarendon Press, Oxford.
Rattray, Robert S., 1930. Akan-Ashanti Folktales. Clarendon Press, Oxford.
Reesman, Jeanne C. (Ed.), 2001. Trickster Lives: Culture and Myth in American Fiction. University of Georgia Press, Athens.
Ricketts, Mac L., 1966. The North American Indian trickster. History of Religions 5 (2), 327–350.
Scheub, Harold, 2012. Trickster and Hero: Two Characters in the Oral and Written Traditions of the World. University of Wisconsin Press, Madison.
Scott, James C., 1987. Weapons of the Weak: Everyday Forms of Peasant Resistance. Yale University Press, New Haven.
Scott, James C., 1990. Domination and the Arts of Resistance: Hidden Transcripts. Yale University Press, New Haven.
Sekoni, Ropo, 1993. Folk Poetics: A Sociosemiotic Study of Yoruba Trickster Tales. Greenwood, Westport.
Shipley, Jesse W., 2009. Comedians, pastors, and the miraculous agency of charisma in Ghana. Cultural Anthropology 24, 3.
Shipley, Jesse W., 2013. Television tricksters: parody and contradiction in Ghanaian melodrama. Social Text 31 (4 117), 49–79.
Shipley, Jesse W., 2015. Trickster Citizenship: Ghana's National Theatre and a Poetics of Freedom. Indiana University Press, Bloomington.
Swinton, John R., 1909. Tlingit Myths and Texts. Bureau of American Ethnology Bulletin, No. 39. Smithsonian Institution, Washington, DC.
Sutherland, Efua, 1975. The Marriage of Anansewa. Longman, London.
Thompson, Stith, 1978. The Folktale. University of California Press, Berkeley.
Vecsey, Christopher, 1981. The exception who proves the rule: Ananse the Akan trickster. Journal of Religion in Africa 12 (3), 161–177.
Weeks, John M., 2002. The Library of Daniel Garrison Brinton. University of Pennsylvania Museum of Archaeology and Anthropology, Philadelphia.
Wescott, Joan, 1962. The sculpture and myths of Eshu-Elegba, the Yoruba trickster. Definition and interpretation in Yoruba iconography. Africa: Journal of the International African Institute 32 (4), 336–354.
White, Bob W., 1999. Modernity's trickster: "dipping" and "throwing" in Congolese popular dance music. Research in African Literatures 30 (4), 156–175.
Yankah, Kwesi, 1995. Speaking for the Chief: Okyeame and the Politics of Akan Royal Oratory. Indiana University Press, Bloomington.

Trust, Political

Eric M Uslaner, University of Maryland, College Park, MD, USA; Center for American Law and Political Science, Southwest University of Political Science and Law, Chongqing, China; and Aarhus University, Aarhus, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Much of the Western world is experiencing a 'a crisis in public confidence in government.' Levels of political trust – are declining. In the mid-1970s, social, political, and economic turbulence throughout the developed world led to worries that governments had lost their legitimacy. The Organization for Economic Cooperation and Development (OECD), representing 33 largely developed countries, warned, "Citizens across the OECD have lost their confidence in the ability of policy makers to solve economic problems and respond to their needs and demands." What is trust in government? How do we measure it? What are the roots, dimensions, and consequences of this malaise? This article surveys the literature on the meaning and measurement of political trust and then inquires into the causes and consequences of this declining confidence in our institutions for the health of democratic societies.

In much of the Western world, we hear of a crisis in public confidence in government. In the mid-1970s the social and economic turbulence throughout the developed world led to worries that governments had lost their legitimacy and that there was a crisis of confidence in government (Crozier et al., 1975). Since the 1980s, we have heard of a 'confidence gap' in public and private institutions, 'disaffected democracies,' and 'democratic deficits' (Lipset and Schneider, 1983; Pharr and Putnam, 2000; Norris, 2011). The Organization for Economic Cooperation and Development (OECD), representing 33 largely developed countries, warned, "Citizens across the OECD have lost their confidence in the ability of policy makers to solve economic problems and respond to their needs and demands" (OECD, 2013).

What is trust in government? How do we measure it and what are the roots of this malaise? I survey the literature on both measurement and causes and then ask what the consequences of this declining confidence in our institutions are.

There are controversies surrounding each of these issues, including (perhaps starting with) measurement. I consider each in turn.

What Is Trust in Government and How Do We Measure It?

Trust in government is the expectation that the state will provide basic services and keep the country secure and prosperous, an evaluation of government performance. Hetherington (2005: 9) defines political trust "as the degree to which people perceive that government is producing outcomes consistent with their expectations…."

Hetherington implies that trust in government varies over time and that the primary factor shaping political trust is the state of the economy. Hetherington (2005: 9, 10) cautions against a one-to-one correspondence of economic conditions with political trust. Perceptions, not economic statistics, matter for trust in government. People are more likely to evaluate government negatively during periods of financial distress. The 'confidence gap' argument was developed in response to the economic turbulence of the 1970s. The contemporary decline in trust across developed nations reflects "citizens' growing expectations [that] have been hard to address with limited government resources…. The financial and economic crisis that started in 2008 led to a significant loss of trust in government," (OECD, 2013:10, 20).

How do we measure political trust? There are two standard survey questions in the literature. First is the question with the longest history:

> How much of the time do trust the government in [the national capital] to do what is right?

This question has been asked in the United States – in the American National Election Studies – since 1958.

An alternative question, used in the World Values Survey (1981–2008), the United States General Social Survey (1972–2012), and the Gallup World Poll (2011–2012) is (using the Gallup question wording, see OECD, 2013: 25):

> In this country, do you have confidence in each of the following, or not? How about national government?

The confidence and trust measures are close, if not identical. The 1987 General Social Survey in the United States included questions on trusting the federal government and confidence in the executive branch (the Presidency), the legislative branch, the Supreme Court, and the army. I created an index by factor analysis, which was correlated with trust in the federal government at 0.47 – hardly perfect but fairly strong for survey data.

Third – a variation on the trust in government question – is trust in specific institutions. Hetherington (2005: 11) cautions against confounding trust in specific institutions – or the people holding them – with trust in government. Attitudes toward governmental and nongovernmental institutions cluster together, so one can construct a more general index of trust in government. Uslaner (2014a) finds a single dimension

 http://dx.doi.org/10.1016/B978-0-08-097086-8.93151-9

for trust in the Prime Minister, the parliament, the courts, the police, and the military using data from the Afrobarometer, the Arab Barometer, and the Global Barometer for Asia and Latin America.

Finally, for many years the American National Election Study (ANES) included an index of trust in government. In addition to the question "How much of the time do you trust the government in Washington to do what is right?," the trust scale included: "Do you think that quite a few of the people running the government are crooked, not very many are, or do you think hardly any of them are crooked?" The index also included the question: "Would you say the government is pretty much run by a few big interests looking out for themselves or that it is run for the benefit of all the people?"

Most discussions of trust in government in the United States (cf Miller, 1974; Citrin, 1974) use the simple trust question rather than the index. It is unclear whether the scale is measuring a common notion of faith in institutions or a causal chain from being crooked or looking out for yourself on mistrust. In less than half of the years all measures were included, the three measures formed an acceptable scale (alpha > 0.6 in 10 of 21 years). Only once since 1988 did the alpha exceed 0.6; the 'officials crooked' variable was weakly related to the other two. Question wording is a central issue. But so is measurement. First, what is the scale employed? The standard question based upon the ANES has a fourfold response scale ('almost always,' 'most of the time,' 'some of the time,' 'hardly ever'). In the late 1990s, the Swiss Household Panel and the Citizenship, Involvement, Democracy (CID) cross-national surveys in Europe (For the Swiss Household Panel, see http://www.swisspanel.ch/. For the history of the CID project, see http://www.mzes.uni-mannheim.de/projekte/cid/and http://www.mzes.uni-mannheim.de/projekte/cid/background.htm. See Van Deth et al. (2007) for essays based upon the public surveys.) shifted to an 11-point scale. The European Social Survey (ESS) followed suit. (See http://www.europeansocialsurvey.org/.)

The 11-point scale was seen as a methodological advance by the Swiss Household Panel and the CID and ESS surveys. Scherpenzeel (n.d.: 1) argued:

> Scales with relatively few response alternatives force respondents to categorize their reaction towards an attitude object instead of directly mapping it onto the response continuum, thus causing information loss…respondents differentiate more between objects when offered response scales with greater numbers of categories…

Scherpenzeel (n.d.) argues that the 11-point scale demonstrated the greatest reliability and validity of any measure of satisfaction in a survey across 10 nations. Zmerli and Newton (2008) argue that the expanded scale is a better measure. Uslaner (2013a) challenges this claim, arguing that there is a widespread tendency of respondents to choose 'middle' categories (4, 5, 6) even when they give more extreme answers to the measures with fewer categories. These 'middle' responses, which Uslaner calls 'clumping,' do not discriminate well among attitudes – a score of 4 is not significantly different across political institutions than is one of 6.

The Determinants of Trust in Government

What factors shape trust in government?

Satisfaction with the Direction of the Country and the Economy

Hetherington's definition points to the role of 'national conditions' – especially the state of the economy – to be the prime mover of trust. Confidence in political institutions, Citrin argued (1974), depends upon the perceptions of leaders' competence. When they get the job done, they are rewarded with the favor of the public. Getting the job done means keeping the economy strong (Listhaug and Wiberg, 1995; Pew Research Center for The People and The Press, 2010: 19).

In 1980, Republican Presidential nominee Ronald Reagan asked voters: "Are you better off now than you were four years ago?" His argument struck a chord – but it was the wrong question. People don't blame political leaders for their personal misfortunes. What matters is how well they have managed the national economy (Kinder and Kiewiet, 1979).

There is a strong relationship between trust in government and satisfaction with the state of the nation in data from the United States (see **Figure 1**). Cross-nationally the results seem weaker. The OECD report (2013: 25) finds less support for an economic link: "…trust in government may not respond to long-term economic developments or absolute standards of living as much as it does to cultural factors, evolving expectations and political developments." Trust in government is lower where people appear to have greater expectations from government, even if they are not the richest. The economic crisis of 2008 seems to have been responsible for lower trust in government across many countries – from the United States (Uslaner, 2014b) to Greece, Ireland, Latvia, Lithuania, Romania, and Spain (Charron et al., 2013: 8).

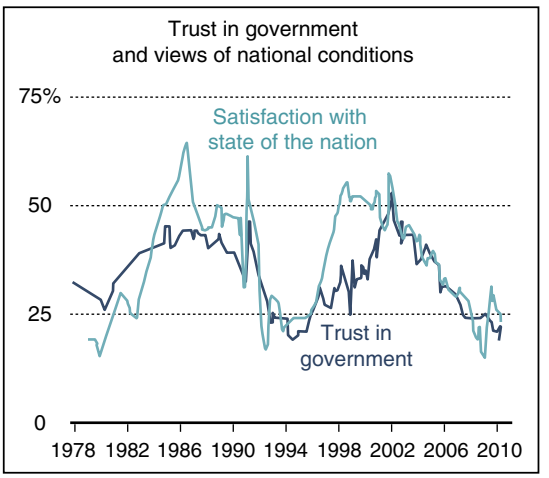

Figure 1 Trends in trust in government: Pew Research Center for the people and the press. Source: http://people-press.org/2010/04/18/distrust-discontent-anger-and-partisan-rancor/.

Data from the Gallup World Poll for 2012 (see Figure 2) support, though not unequivocally, the link between economic optimism and trust in government. The greatest declines in confidence occur among countries that have been most negatively affected by the 2008 economic crisis (notably Ireland, Greece, Slovenia, Portugal, and Spain). Where the crisis was least felt – in China, Switzerland, and Israel – faith in the state moved upward. These data suggest that the economy is a major factor in shaping confidence in government. But it is likely that perceptions of the economy – and especially views about the future – matter more than objective economic statistics.

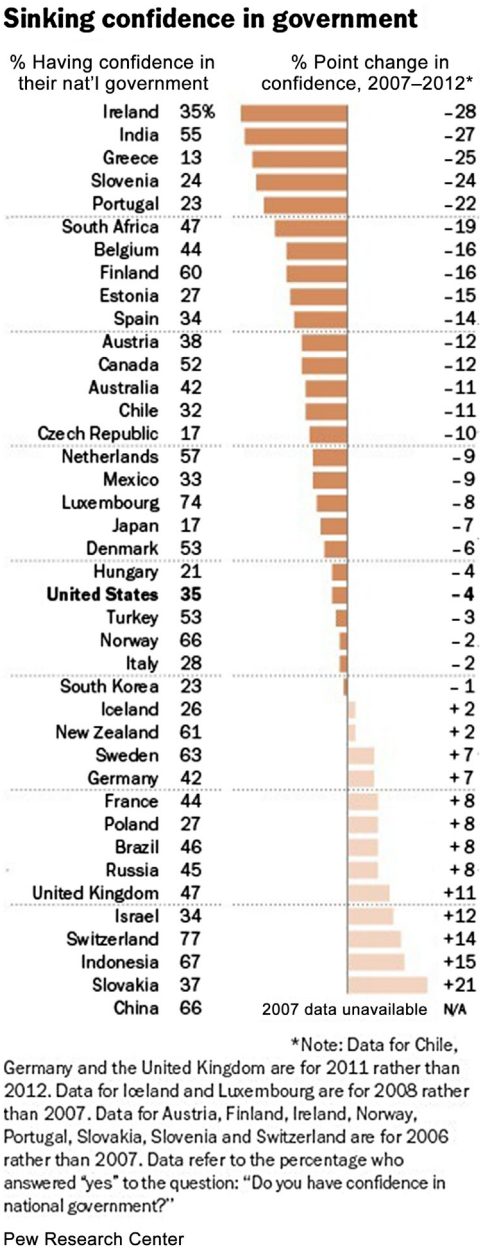

Figure 2 Changes in trust in government (Gallup). Source: Gallup World Poll and OECD. http://www.pewresearch.org/fact-tank/2013/11/21/confidence-in-government-falls-in-much-of-the-developed-world/ (accessed 01.12.13.).

What Government Does?

Government performance is more than just economic outcomes. It is also how governments provide basic services, especially when are dealing with natural disasters.

Outside the developed world, governments may not always provide basic services – water, electricity, telephone – efficiently. When service delivery is intermittent, people blame the government and have less faith in it (Uslaner, 2011a, 2014a). The quality of public service in developed countries also shapes trust in government: Better treatment from public employees leads to more trust in government (Christensen and Laegreid, 2002).

Public confidence in government rises and falls with perceptions of how well the state deals with major disasters. In the United States, the perceived poor response to Hurricane Katrina in Louisiana and Mississippi in 2005 led to less trust in government, particularly among people living in the affected areas (Nicholls and Picou, 2013; Uslaner, 2011a). The perceived inadequate response of both the government and the nuclear energy company Tepco to the Japanese tsunami in 2011 led to sharp declines in trust in both government and private institutions (Foster, 2011; Uslaner, 2011b). When people are most in need of government assistance – and believe that they have not received it – they lose faith in the state. This is especially the case, as in the Japanese tsunami, if people believe that public officials have conspired with the private sector to hide information that could have spared lives.

Corruption

People have less confidence in governments they believe to be corrupt. This seems to be almost tautological, since the ANES index of trust in government included the question of how many officials are crooked. While there is support for this thesis, it is not universal. Uslaner (2011a, 2014a) reports strong support for this argument in transition countries and most of the developing world. Corruption leads to less trust in both government and other people when it is perceived to be linked to inequality. When government officials get rich at the expense of ordinary people, citizens see the political system as unfair. The institutions that are supposed to be the *least* biased in a well-functioning state, the courts and the police, will be perceived as *the most corrupt* (Rothstein, 2000; Uslaner, 2008: Chapters 5 and 8). Petty corruption – small payments to public officials, police officers, teachers, doctors, among others – do not enrich the recipients. So petty corruption does not lead to less trust either in government or other people (Uslaner, 2008: Chapters 5 and 8).

People object to corruption because they see officials (and business people) getting rich at the expense of the state. But as political leaders get rich, corruption is a tax on the poor. There is less money in the public purse for social expenditures such as building schools and hospitals. And often people have to pay bribes to get these services. Corruption makes leaders rich and leads to less effective service delivery and lower economic growth, again taxing the poor (Mauro, 1997).

Does good economic performance mitigate the effects of perceived corruption on trust? Chang and Chu (2006) find no support for this argument in samples from five East Asian

nations. Perceptions of higher corruption lead to less trust in government even when people see strong economic performance, are satisfied with democracy, and expect the future to be bright.

There may be some indirect evidence that strong economic performance might matter when corruption is high and democratic governance is weak. Klasnja and Tucker (2013) report experimental results comparing voters in Sweden and Moldova. When confronted with arguments revealing dishonest politicians in the country with low corruption (Sweden), respondents said that they would vote against these political leaders *regardless of their perception of economic conditions*. In high-corruption Moldova, with less of a history of democratic governance, respondents indicate opposition to dishonest incumbents only if the economy was perceived to be weak. Voting is not the same as trust in government, but this result might well be generalizable.

Winners and Losers

People are more likely to trust the government if they have supported it in the past. If you vote for the party in government, you are more likely to trust it. Anderson et al. (2005: Chapter 5) provide mixed support for this hypothesis across countries. However, the effects are particularly powerful for strong party identifiers for the more ideological extreme voters.

In the United States there have been only modest correlations between partisanship and trust in government. As the parties have become more polarized, the link has strengthened (Hetherington, 2008: 21; King, 1997).

McCarty et al. (2006) show strong linkages between growing income inequality in the United States and partisan polarization. Uslaner (2013b: 1, 7) shows that rising inequality accounts for approximately a third of the long-term decline in trust in government in the United States. Fifty eight percent of Americans trusted their government in 1958 compared to 22% in 2010 (Uslaner, 2013b: 15). In contemporary American politics, the rise of the Tea Party as a movement largely within the Republican party, shows the impact of polarization. The Tea Party, founded as a protest against big government, is founded on mistrust of the state.

In the 2010 ANES Evaluations of Government and Society Study (wave 1), 95% of respondents who said that they support the Tea Party did not trust government–compared to 61% who oppose the Tea Party (and 78% who neither supported nor opposed the movement). Almost 60% of strong Republican identifiers support the Tea Party compared to just 4% of Democrats. Trust in government is becoming increasingly partisan.

An explanation based upon polarization has limited applicability. Polarization is primarily an American development. Many of the mainstream parties in Europe have converged *toward* each other (on Britain, see Adams et al., 2012). Yet in most European countries, right-wing parties based upon opposition to immigration have gathered support.

Rydgren (2007: 253, 254) argues that the "convergence [of mainstream parties] may result in a feeling that the established parties 'are all the same.' This, in turn, may fuel popular distrust and discontent in politicians and political parties and create an audience receptive to parties ready to mobilize protest votes." The rise of right-wing populist parties is not the same as the rise of the Tea Party in the United States – but both the supporters of each fear that immigrants (and minorities more generally) constitute a threat to the dominant (white) culture (Jacobson, 2011; Lucassen and Lubbers, 2012), In both Europe and the United States, the rise of right-wing groups is based upon a loss of faith in the mainstream parties and in the government more generally.

Incivility and Gridlock

Otto von Bismark, Chancellor of Germany in the late nineteenth century, said, "Laws are like sausages, it is better not to see them being made." (See http://www.brainyquote.com/quotes/quotes/o/ottovonbis161318.html (accessed 31.12.13.).) Hibbing and Theiss-Morse (1995) find that Americans don't like Congress because they find the bickering and incivility of the legislative process to be unseemly: The legislative process is not pretty. This leads to less trust in the legislative branch and almost certainly to government in general (although they do not test this).

More recently the decline in trust in government has been linked to gridlock – the inability of Congress to agree to legislation. (The term 'gridlock' comes from a description of traffic patterns in which cars are stalled and intersections are blocked. Political gridlock is the 'logjam' of legislation.) Since 2011 the Republicans have controlled the House of Representatives and the Democrats, the Senate, and the two chambers have been unable to reach compromises on legislation. The ideological divide between the two parties is so wide that there is literally no common ground between them (Resnick et al., 2013). This division led to a shutdown of the federal government for more than 2 weeks in 2013 and to the fewest bills passed in the history of Congress (Sullivan, 2013). Trust in government fell sharply and pollster Andrew Kohut attributed this decline to legislative gridlock. (See http://www.pbs.org/newshour/bb/politics/jan-june13/pew_01-31.html (accessed 31.12.13.).)

Gridlock and the open conflicts (and name calling) within Congress (cf Uslaner, 1993) are certainly responsible for declining trust in government in the United States. Elsewhere polarization is not the norm. Jeering and insults are regular parts of legislative life in many parliamentary systems, although the members do not take their distaste for each other outside the chamber as they do in the United States. And gridlock is a decidedly an American phenomenon. Without rules permitting obstruction, as we find especially in the United States Senate, and without divided party control of the legislative and executive branches, minorities are powerless to stop the flow of legislation. So while this explanation undoubtedly explains part of (maybe even much of) the decline in trust in government in the United States, it cannot account for low confidence in the state elsewhere.

Social Trust

Political trust, some say, is part of a larger syndrome of an upbeat world view. Lane (1959: 164) argues, "Trust in elected officials is seen to be only a more specific instance of trust in

mankind." Advocates of a link between the two types of trust maintain that strong government performance makes people feel better about the government and ultimately more willing to cooperate with each other (Brehm and Rahn, 1997: 1008; Misztal, 1996: 198). Citizens rest easier when dealing with strangers if they know that government will enforce contracts. Rahn et al. (1997: 24) argue that when people trust their government, they are more likely to believe that they can influence it. This growing sense of efficacy makes people more likely to trust each other.

Zmerli and Newton (2008: 707) find strong relationships between the two types of trust, arguing that "…according to social capital theory social and political trust are closely associated and mutually supportive….The theory has prima facie plausibility and, as a basic proposition of political sociology that tries to link the social with the political, a great deal of potential theoretical power." Yet, they admit that the "the claim…has poor empirical support."

They are correct. At the aggregate level, the two forms of trust are highly correlated: from 1964 to 2010, the r^2 between the two trust measures is 0.56 in the United States. Yet, Uslaner (2002: Chapter 5) shows that the foundations of political and social trust are distinct. Evaluations of the state of the economy and of specific political leaders shape only political trust, not social trust. Longer-term measures of optimism only determine social trust.

The individual-level correlations are small for both trust in government and confidence (Uslaner, 2002: 150–153). As Hetherington (2005: 9) also argues, this argument "…conflates two very different concepts….In American life, one can easily identify people who trust other people but do not trust government, and with a little more difficulty, one can also identify the reverse."

It is not quite so problematic as Hetherington claims. Trust in government goes up and down with the state of the economy: It rose sharply when the economy recovered from recessions in the transitions from Jimmy Carter to Ronald Reagan and from George H.W. Bush to Bill Clinton. It rose again when the country faced a common threat after the 9/11 attacks. But social trust remained low (except for a brief period right after the terrorist attacks). Trust in government responds to the economics and politics of the moment, trust in people to the long-term optimism rather than to today's economic report (Uslaner, 2002: 153–158).

Why, then, do Zmerli and Newton, among others, find stronger relationships between the two types of trust than do others? Is it context? Uslaner primarily examines data from the United States, but this is not the answer, because Mishler and Rose (2005) use data from Russia and Kaase (1999: 13) focuses on Europe and concludes that the correlation between the two types of trust 'is small indeed.' The principal difference between the analyses of Uslaner, on the one hand, and Zmerli and Newton, on the other, is the measurement of trust. They use the ESS and the United States Citizenship Involvement Democracy survey. (See http://www.uscidsurvey.org/ (accessed 31.12.13.).) Both use the 11-point trust scales while previous studies employed the dichotomous social trust measure and the 4-point trust in government measure. They claim that these scales are "more reliable measures of attitudes and values" (Zmerli and Newton, 2008: 714). The correlations using the 11-point measures are uniformly stronger across European nations and the United States than are those for dichotomized measures (constructed from the 11-point scales) and Zmerli and Newton (2008: 715–717) take this as evidence in support of the new measures.

It is unclear that the higher correlations for measures based upon 11-point scales are superior to those based on dichotomies or four-point scales. The 11-point scales are problematic (Uslaner, 2013a). Romanians were given both the dichotomous and 11-point social trust questions in the CID: The r^2 between them is a mere 0.164. Seventy-six percent of Romanians scoring 4 on the 11-point measure are mistrusters on the dichotomous measure, as are 69% directly in the middle at 5, and even 58% who scored a 6 on the expanded measure.

Uslaner (2002: 68–74) showed that American respondents to a 'think aloud' experiment about trust understood the question and also that responses are quite stable over time. More complex measures have not fared so well (Uslaner, 2014c) over time or as predictors of outcomes that he demonstrated in earlier work (Uslaner, 2002: Chapter 7). The 11-point scale may be artificially leading us to expect that social and political trust stem from common roots, when plenty of evidence suggests otherwise.

Why Care About Trust in Government?

Do we want trust in government? Both the OECD report and Hetherington suggest that trust is a tool to adopt progressive social policies. Hetherington (2005) argues that trust is necessary to gather support for the government to enact policies (at least in the United States) that are redistributive – welfare, foreign aid, assistance to blacks. People who trust government are not always liberal, but they do have faith that government officials will handle their money responsibly. There is thus a link from trust in government to both liberalism and to outcomes that are generally seen as liberal (Hetherington, 2005: Chapter 3). Even people who are not liberal would be more likely to support health care reform if they trusted government – especially to spend their money responsibly (Hetherington, 2005: Chapter 3). Conservatives distrust government, all government and not just Washington (Uslaner, 2001). But when they trust government, they are more likely to support federal programs that provide aid to those who are less fortunate.

The OECD (2013: 22) report makes similar arguments: "people are prepared to sacrifice some immediate benefits if they have positive expectations of the longer-term outcome of public policies, either at a personal level (pensions) or by contributing to the common good (redistribution of income through taxation)."

Trust in government matters when we need the state the most – when the private sector has neither the means nor the will to tackle major social problems. In the second decade of the twenty-first century, trust in government seems in short supply in the West as new issues come to the fore of the political landscape even as – or perhaps especially as satisfaction with economic performance – the key driver of faith in the state has fallen.

See also: Corruption: Political and Public Aspects; Income Inequality; Legitimacy, Sociology of; Trust, Sociology of.

Bibliography

Adams, James, Green, Jane, Milazzo, Caitlin, 2012. Has the British public depolarized along with political elites? An American perspective on British public opinion. Comparative Political Studies 45, 507–530.

Anderson, Christopher, Blais, Andre, Bowler, Sean, Donovan, Todd, Listhaug, Ola, 2005. Losers Consent: Elections and Democratic Legitimacy. Oxford University Press, New York.

Brehm, John, Rahn, Wendy, 1997. Individual level evidence for the causes and consequences of social capital. American Journal of Political Science 41, 888–1023.

Chang, Eric C.C., Chu, Yun-han, 2006. Corruption and trust: excpetionalism in Asian democracies? Journal of Politics 68, 259–271.

Charron, Nicholas, Lapuente, Victor, Rothstein, Bo, 2013. Quality of Government and Corruption from a European Perspective. Edward Elgar, Cheltenham, UK.

Christensen, Tom, Laegreid, Per, November 2002. Rust in Government – the Relative Importance of Servicer Satisfaction, Political Factors and Demography. Bergen University Research Foundation, Stein Rokkan Centre for Social Studies.

Citrin, Jack, 1974. Comment: the political relevance of trust in government. American Political Science Review 68, 973–988.

Crozier, M., Huntington, Samuel P., Watanuki, Joji, 1975. The Crisis of Democracy. New York University Press, New York.

Foster, Malcolm, 2011. Japan Natural Disasters Shake Public Trust in Government. http://www.huffingtonpost.com/2011/09/01/japan-natural-disaster-government_n_94463 2.html (accessed 30.12.13.).

Hetherington, Marc J., 2005. Why Trust Matters: Declining Political Trust and the Demise of American Liberalism. Princeton University Press, Princeton.

Hetherington, Marc J., 2008. Turned off or turned on? How polarization affects political engagement. In: Nivola, Pietro S., Brady, David W. (Eds.), Red and Blue Nation?, vol. 2. Brookings Institution Press, Washington.

Hibbing, John, Theiss-Morse, Elizabeth, 1995. Congress as Public Enemy. Cambridge University Press, New York.

Jacobson, Gary C., 2011. The President, the Tea Party, and Voting Behavior in 2010. Presented at the 2011 Annual Meeting of the American Political Science Association. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1907251 (accessed 05.09.11.).

Kaase, Max, 1999. Interpersonal Trust, Political Trust, and Non-Institutionalised Political Participation in Western Europe. West European Politics 22, 1–21.

Kinder, Donald R., Kiewiet, D. Roderick, 1979. Economic discontent and political behavior: the role of personal grievances and collective economic judgments in congressional voting. American Journal of Political Science 23, 495–527.

King, David C., 1997. The polarization of American political parties and mistrust of government. In: Nye, Joseph S., Zelikow, Philip, King, David C. (Eds.), Why People Don't Trust Government. Harvard University Press, Cambridge.

Klasnja, Marko, Tucker, Joshua A., 2013. The economy, corruption, and the vote: evidence from experiments in Sweden and Moldova. Electoral Studies 32, 536–543.

Lipset, Seymour Martin, Schneider, William, 1983. The Confidence Gap: Business, Labor, and Government in the Public Mind. Free Press, New York.

Listhaug, Ola, Wiberg, Matti, 1995. Confidence in political and private institutions. In: Klingemann, H.-D., Fuchs, D. (Eds.), Citizens and the State. Oxford University Press, Oxford, pp. 298–322.

Lucassen, Geertje, Lubbers, Marcel, 2012. Who fears what? Comparative Political Studies 45, 547–574.

Mauro, Paolo, 1997. Why Worry about Corruption? International Monetary Fund, Washington.

McCarty, Nolan, Poole, Keith, Rosenthal, Howard, 2006. Polarized America: The Dance of Ideology and Unequal Riches. MIT Press, Cambridge.

Miller, Arthur H., 1974. Political issues and trust in government: 1964–1970. American Political Science Review 68, 951–972.

Mishler, William, Rose, Richard, 2005. What are the consequences of political trust: a test of cultural and institutional theories in Russia. Comparative Political Studies 38, 1050–1078.

Misztal, Barbara A., 1996. Trust in Modern Societies. Polity Press, Cambridge, UK.

Nicholls, Keith, Picou, J. Steven, 2013. The impact of Hurricane Katrina on trust in government. Social Science Quarterly 94, 344–361.

Norris, Pippa (Ed.), 2011. Democratic Deficit: Critical Citizens Revisited. Cambridge University Press, New York.

OECD (Organization for Economic Cooperation and Development), 2013. Government at a Glance. OECD Publishing, Paris. http://dx.doi.org/10.1787/gov_glance-2013-en (accessed 05.11.13.).

Pew Research Center for The People & The Press, April 18, 2010. The People and Their Government: Distrust, Discontent, Anger, and Partisan Rancor. http://www.people-press.org/2010/04/18/distrust-discontent-anger-and-partisan-rancor/ (accessed 05.05.10.).

Pharr, Susan, Putnam, Robert D. (Eds.), 2000. Disaffected Democracies: What's Troubling the Trilateral Countries?. Princeton University Press, Princeton.

Rahn, Wendy M., Brehm, John, Carlson, Neil, 1997. National Elections as Institutions for Generating Social Capital. Presented at the Annual Meeting of the American Political Science Association. August–September, Washington, DC.

Resnick, Brian, McGill, Brian, Krivitchenko, Ella, 2013. This Is What Congressional Gridlock Looks Like in 1 Chart. http://www.nationaljournal.com/congress/this-is-what-congressional-gridlock-looks-like-in-1-chart-20131113 (accessed 31.12.13.).

Rothstein, Bo, 2000. Trust, social dilemmas, and collective memories: on the rise and decline of the Swedish Model. Journal of Theoretical Politics 12, 477–499.

Rydgren, Jens, 2007. The sociology of the radical right. Annual Review of Sociology 33, 241–261.

Scherpenzeel, A., n.d. Why Use 11-Point Scales? http://www.swisspanel.ch/file/doc/faq/11pointscales.pdf (accessed 15.06.12.).

Sullivan, Sean, 2013. The Diminishing Productivity of Congress, in One Chart. http://www.washingtonpost.com/blogs/the-fix/wp/2013/12/02/the-diminishing-productivity-of-congress-in-one-chart/ (accessed 31.12.13.).

Uslaner, Eric M., 1993. The Decline of Comity in Congress. University of Michigan Press, Ann Arbor, MI.

Uslaner, Eric M., 2001. Is Washington really the problem? In: Hibbing, John, Thiess-Morse, Elizabeth (Eds.), What Is It about Government That Americans Dislike? Cambridge University Press, New York.

Uslaner, Eric M., 2002. The Moral Foundations of Trust. Cambridge University Press, New York.

Uslaner, Eric M., 2008. Corruption, Inequality, and the Rule of Law. Cambridge University Press, New York.

Uslaner, Eric M., 2011a. Corruption, the inequality trap, and trust in government. In: Zmerli, Sonja, Hooghe, Marc (Eds.), Political Trust: Why Context Matters. ECPR Press, Colchester, UK.

Uslaner, Eric M., 2011b. Trust in Government and Disasters. Invited talks for the U.S. Embassy, Tokyo, at the Fukuoaka American Center (13 June), Nagoya American Center (June 14), Tokyo American Center (15 June), Iwate Prefecture Cultural Exchange Center (Morioka, Japan, 16 June), and the Kansai American Center (17 June).

Uslaner, Eric M., 2013a. Is Eleven Really a Lucky Number? Measuring Trust and the Problem of Clumping (Unpublished Manuscript). University of Maryland–College Park.

Uslaner, Eric M., 2013b. Income Inequality in the United States Fuels Pessimism and Threatens Social Cohesion (Working Paper). Center for American Progress. http://www.americanprogress.org/wp-content/uploads/2012/12/Uslaner.pdf (accessed 30.12.13.).

Uslaner, Eric M., 2014a. Corruption, inequality, and trust in government. In: Zmerli, Sonja, van der Meer, Thomas (Eds.), Handbook on Political Trust. Edward Elgar, Cheltenham, UK.

Uslaner, Eric M., 2014b. The economic crisis of 2008, trust in government, and generalized trust. In: Harris, Jared D., Moriarity, Brian, Wicks, Andrew C. (Eds.), Public Trust in Business. Cambridge University Press, New York.

Uslaner, Eric M., 2014c. Measuring generalised trust: in defense of the standard question. In: Lyon, Fergus, Möllering, Guido, Sanders, Mark, Hatzakis, Tally (Eds.), Handbook of Research Methods on Trust, second ed. Edward Elgar, London.

Van Deth, Jan W., Montero, José Ramón, Westholm, Anders (Eds.), 2007. Citizenship and Involvement in European Democracies. A Comparative Analysis. Routledge, London, New York.

Zmerli, Sonja, Newton, Kenneth, 2008. Social trust and attitudes towards democracy. Public Opinion Quarterly 72, 706–724.

Trust, Sociology of

Margaret Levi, University of Washington, Seattle, WA, USA; and United States Studies Centre at University of Sydney, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Trust is at issue when someone makes oneself vulnerable to another who can harm if the trust is misplaced. The recipient of trust is either trustworthy or not, and much of the literature revolves around the evaluation of the trustworthiness of the trusted by the trustor. Trust can exist among those who know each other intimately (personal trust) and among strangers (interpersonal, social, or generalized trust). Trust can have as its object other people or institutions and organizations. In conceptualizing trust to undertake empirical research, two crucial distinctions exist: cognitive vs noncognitive trust and personalized vs generalized trust. The more instrumental and cognitive theorists tend to treat trust as an estimate of the trustworthiness of those with whom one has relationships as individuals or within social networks. In contrast are those who claim trust is dispositional or moralistic. While each tradition treats the work trust does as an empirical question, their conceptualizations are sufficiently distinct so that both their presumptions of the role that trust plays in society and their findings are often incompatible. This leads to debate about the sources of trust and whether trust is essential for good government, economic growth, and harmony.

Trust and trustworthiness are distinct but so implicated with each other that it is generally necessary to consider the second in contemplating the first. Indeed, much of the literature revolves around how to establish and assess the trustworthiness of persons, organizations, and institutions.

The Term 'Trust'

There are several distinct literature on trust and several ways to categorize them. One typology differentiates trust among those who know each other intimately (personal trust) from trust among strangers or relative strangers (interpersonal, social, or generalized trust), which in turn is distinct from trust of government, firms, and other institutions and organizations. Another involves the distinction between trust as a relationship and trust as a disposition or value. There are approaches to trust that rely on rational choice, others that call upon structural functionalism, and yet others that evoke social identity theory. There are surveys of trust, experiments on trust, and formal models of trust.

The concept of trust has a long intellectual history. In political philosophy, Thomas Hobbes argued that trust makes cooperation possible, but that betrayals of trust make it hard to rely upon. John Locke emphasized the importance of mutual trust between the governors and the governed within legitimate polities (Dunn, 1990). In sociology, it is a concept used by scholars as various as Talcott Parsons, Peter Blau, Harold Garfinkel, Niklas Luhmann, and Bernard Barber in their efforts to understand the construction of social order, cooperation, institutions, and organizations, and even everyday interactions. To the economist Arrow (1974), 'Trust is an important lubricant of a social system' because it can enhance efficiency. Confidence, reliance, and assurance are related concepts that some scholars conflate and others distinguish albeit in a variety of ways.

What ties various perspectives together is the recognition of trust as a means to cope with uncertainty, but disagreement remains about how trust comes about, when it should be relied upon, and what kinds of interactions it enhances (see, e.g., Gambetta, 1988; Levi and Stoker, 2000; Cook et al., 2005; Tyler, 1998).

For most scholars, trust is relational, involving at least one individual making oneself vulnerable to another individual, set of individuals, or an institution that possesses the potential to do harm or betray. Trust implies a judgment of risk in conditions of uncertainty. Moreover, trust is seldom unconditional; it is given over specific domains. A client may entrust a banker with money but never consider leaving a child in the banker's care; citizens may entrust their lives to their government during wartime but not new taxes to legislators who promise improved public services.

Trustworthiness refers to certain attributes of a person or an institution and can exist even when trust is not at issue. Trustworthiness has two dimensions. The first involves the commitment to act in the interests of the trustor due to moral values about promise keeping, caring about the trustor, or incentive compatibility, or some combination of all three. The second dimension involves competence in the domain over which trust is being given. The trustworthy will not betray the trust as a consequence of either bad faith or ineptitude.

Research on Trust and Trustworthiness

The recent burgeoning of the social science literature on trust was initially a response to the perception that political and social trust is in decline and in part to an argument that trust is essential to a good society. The fact that there are now decades of responses to the same survey questions and that they indicate a reduction in both political and social trust in the advanced industrial democracies has generated an effort to explain and correct this trend. The emphasis on the importance of social capital and the claim that it, too, has decreased over time is another factor producing a focus on trust. Both Coleman (1990) and Putnam et al. (1993) included trust as a critical element in their original accounts of social capital.

Economists and organizational theorists have come to share Arrow's and others' view that more than formal incentives or

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32162-6

credible commitments may be necessary to account for productive internal relationships and innovative change within firms and organizations (Kreps, 1990; Miller, 1992; Kramer and Tyler, 1996). Sociologists, anthropologists, political scientists, and social psychologists are reconsidering the role of trust and trustworthiness in attempting to explain variation in cooperation (Barr, 2004; Gambetta and Hamill, 2005) and better governance (Fukuyama, 1995; Braithwaite and Levi, 1998; Tyler and Huo, 2002; Rothstein, 2011; Uslaner, 2002) are also reconsidering the roles trust and trustworthiness play.

Motivational Assumptions about Trust

Why does someone trust? There are at least four possible kinds of motivations: dispositional, moral, social, and instrumental.

In one widespread perspective, trust is a disposition, that is, some individuals are more inclined to trust or be trustworthy than others due to factors of personality or character. A dispositional view of trust leads to investigations of variation among the population in the propensity to trust and be trustworthy. There is little agreement in the trust literature, however, on the sources of such a disposition. It could be a result of innate character, childhood experience, learning and updating, instrumental manipulation, knowledge, social networks, social bonds, or social norms. The assumption of trust as a disposition fails to capture important elements of trust: trust always has both an object, the person, persons, or institution being trusted and a domain, what the person, persons, or institution is being trusted to do.

By some accounts, trust is motivated by morality. Individuals are trusting because they believe trust is the act of a good person. Trust as a virtue is a very unsatisfying account of motivation, however. To trust a conman or to trust a corrupt politician is hardly virtuous by any common language sense of virtue. Trust may reflect a normative ordering that demands deference and obedience to certain kinds of authorities, but it is important to maintain the distinction between the normative features of trust and the claim that trusting behavior is either morally motivated or a moral good. What may well have moral motivation is trustworthiness, which generally implies commitments, promise keeping, and obligations. We are more likely to consider someone trustworthy, ceteris paribus, when we believe they have ethical values that make it unlikely for them to betray a trust.

A third possible motivation is social in the sense that individuals derive their identity and self-esteem from their relations to others and thus trustor are trustworthy in keeping with their sense of self or to enhance social status, or both. Although the social motivation is intentional, the end sought has more to do with identity than with the achievement of some material benefit. This leads many of the adherents of social motivation to claim that it is neither interest based nor calculative.

A final kind of motivation is, in fact, instrumental and calculative. Trust is given when it is a means to a particular end and when there is reason to believe the trust will be rewarded. There is a positive expectation that the payoffs from taking a risk are worth the possible costs. It is easiest to trust when the risks are sufficiently low, and it may well be worth the risk when the returns are sufficiently high. There are some circumstances, however, when the cost of misplaced trust is so high that trust is rarely given or given only when there is considerable insurance. Trusting one's child to the care of another or investing one's life savings are two such cases.

The development of a reputation for trustworthiness can also be instrumental. The trustworthiness of an individual may enlarge one's possibilities for valued exchange and cooperation, and the trustworthiness of a government or organization may enhance the voluntary compliance and consent of those subject to its rules. Trustworthiness is also likely to increase the range of discretion given to the trusted, which is an issue of particular importance to agents, be they legislators or union officials or managers, acting on behalf of principals who lack the same intimate knowledge of the situation and the alternatives.

Conceptualizing Trust

In conceptualizing trust in order to engage in empirical research, two crucial distinctions exist: cognitive vs noncognitive trust; dyadic vs embedded trust. There is no empty cell in this matrix; a scholar can be concerned with either of the first and either of the second.

Assumptions about motivations are closely linked to whether the conceptualization of trust that informs empirical research is noncognitive or cognitive. Those who argue for motivations based on morality or social bonds tend to have a noncognitive approach, and those who argue motivations are instrumental generally have a cognitive view. The dispositional motivation can fit with either, depending on one's view of disposition formation.

The noncognitive conception of trust draws on the work of psychologists concerned with social identity, particularly Henri Tajfel and John C. Turner. It is advocated by contemporary social psychologists, such as Robyn Dawes, Tom Tyler, Rod Kramer, and Marilyn Brewer, and by social theorists, such as Allan Silver, who perceive trust as having a moral basis that contrasts with, although often complementing, the world of contracting and impersonal transactions. Individuals trust because of their attributions of motives to the person or institutions that must be trusted. If they attribute trustworthiness, they are more likely to trust, and they are most likely to attribute trustworthiness when there are social bonds or group identification between the trustor and the trusted. This perspective is rooted more in the emotions and morals of an individual than in the capacity to reason. It reflects personal relations and cannot be easily regulated by the impersonal relationships embodied in bureaucracy, market, or formal law. The cognitive conception of trust builds on the work of economists, political economists, and social exchange theorists, such as Kenneth Arrow, James S. Coleman Russell Hardin, and Diego Gambetta. The act of trust, of making oneself vulnerable, requires a thoughtful assessment of risk, informed by a rational expectation or a rational belief that the trusted will be trustworthy. Such an expectation or belief is generally grounded in knowledge of the other's past behavior or present motives and constraints. It can be affected by the existence of institutions, sanctions, and incentives that affect the probabilities that the trusted will act in the interest of the trustor.

The second distinction is between dyadic and embedded trust. Trust is relational in either case, but embedded trust

permits conceptualization of generalized trust and institutional trust, and not just trust at the individual level. Dyadic trust depends upon expectations about or social bonds with a particular other. Embedded trust rests in social networks and institutional arrangements that permit expectations about or social bonds based on the social roles, categorization, and rules affecting others.

Measuring and Observing Trust

The most common methods for studying trust include surveys, experiments, and case studies, sometimes but not always informed by formal models (see Levi and Stoker, 2000; Nannestad, 2008, for reviews). The surveys attempt to measure either political or social trust, but there is considerable controversy about how well they measure either. This is particularly true for the questions in the National Election Studies, which seem to be more about confidence in or positive assessment of elected officials than of political trust, and for those in the general social surveys, which seem to be tapping issues of social anxiety rather than social trust. The World Value Survey asks about both institutional and interpersonal trust, but its questions have similar limitations as the American versions. These major surveys do have the advantage of asking the same questions over the course of several decades, but there remains the issue of whether the time range covered is adequate to capture real changes in trust. Moreover, it is possible that people asked at one point in time understand a question differently than when they are asked at a later point in time. Surveys designed explicitly to capture the relationships of trust are even less likely than the major surveys to enhance understanding of changes in trust relations over time, and their questions still tend to be open to multiple interpretations.

Some survey work does offer insight into trust relations, nonetheless. For example, Toshio Yamagishi uses survey evidence, often in combination with experimental work, to provide some interesting observations on variations in trust within and across populations and countries. Comparing the United States and Japan, Yamagishi and Yamagishi (1994) conclude that the embeddedness of the Japanese, at least until recently, in such tight social networks meant that they did not need to develop the capacity for discerning who and what to trust that has long been characteristic of Americans. Others (e.g., Herreros, 2004; Levi et al., 2009; Levi and Sacks, 2009; Rothstein, 2011) use specific questions about trust in particular institutions to understand interactions between governments and citizens.

Experimental work on trust has different strengths and weaknesses. At their best, experiments can control for factors that might influence or be confused with trust and trustworthiness, they can investigate choice behavior and not just the attitudes captured by surveys, and they can be designed so as to clarify the conditions under which individuals are likely to trust. There have been interesting experiments linking trust to issues of exchange and collective action, and there has been some work attempting to get at the effect of dispositional variations within a population. Most of the experiments, however, focus on dyadic relationships, although there are experiments that investigate the operation of social networks. Moreover, much of the experimental work to date has conflated cooperation and trust. For these experiments to be more compelling, they need to clarify the roles that trust plays in producing cooperation rather than assuming that where there is cooperation, there is trust.

Several other approaches to the study of trust also deserve mention. The trust game and other formal models have proved useful to organizational theorists and economic historians attempting to understand the role trust plays in the relationships between principals and agents and in long-distance trade. There have been statistical analyses of indicators of trust in order to determine variation over time, and there have been case studies that attempt to capture the attributes of trustworthy government or trust that affects economic interactions and productivity in and across firms, industries, groups, regions, and countries. However, this research, although multiplying rapidly, is still in fairly early stages of development.

One of the most important advances in the research on trust is in the investigation of how differences in power relationships affect trust and trustworthiness (see, esp. Ensminger and Knight, 1997; Farrell, 2009; also see discussion in Cook et al., 2005, passim).

Is Trust a Useful Concept?

There is some question of whether a concept like trust is even necessary if the purpose is to explain phenomena that rely on impersonal, multiperson, and anonymous interactions. Williamson (1993) makes a case that it is not; to him trust exists only in personal relations with incentives, interests, and institutions regulating political, market, and large-scale social interactions. Silver (1985), coming from a totally different perspective, offers a similar argument. Hardin (1998) claims that trust, as he defines it, can exist only in the rather limited set of cases where the trustor has sufficient knowledge of the trusted to believe that their incentives are compatible; he is wary of such concepts as trust in the government. Levi (2000) claims that trust is not always the solution to distrust. There are, after all, numerous functional alternatives to trust that rely more on assurance or coercion than on trust, such as, third-party insurers, contracts and other legal enforcement devices, professional ethics, fictive kinship, and a variety of social networks and institutional arrangements.

The argument of the social identity theorists and those who come out of the sociological tradition of Durkheim or the liberal political tradition of Locke is that trust most certainly matters and is, in fact, essential to a well-ordered society. It is constitutive, in this view, of the moral and personal glue that enables modern societies to function (Seligman, 1997). These scholars tend to decry the loss of trust and seek to halt its further decline by advocating the maintenance and creation of community, republicanism, and more personalized interactions, even in large-scale organizations. On the other hand, those relying on instrumental and cognitive theories treat the work trust does as an empirical question. Relationships of trust are but one of many means to achieve given ends, and whether they are the best means is still open to considerable research and debate.

See also: Belief, Anthropology of; Control, Social; Emotion: History of the Concept; Emotions, Sociology of; Interpersonal Trust across the Lifespan; Organizations, Sociology of; Psychoanalysis in Sociology; Rational Choice Explanation: Philosophical Aspects; Social System; Trust: Philosophical Aspects; Value Pluralism; Values, Sociology of.

Bibliography

Arrow, K., 1974. The Limits of Organization. Norton, New York.

Barr, A., 2004. Kinship, familiarity, and trust: an experimental investigation. In: Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality. Oxford University Press, New York, pp. 305–331.

Braithwaite, V., Levi, M. (Eds.), 1998. Trust and Governance. Russell Sage Foundation, New York.

Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge, MA.

Cook, K.S., Hardin, R., Levi, M., 2005. Cooperation without Trust? Russell Sage Foundation, New York.

Dunn, J., 1990. Trust and political agency. In: Dunn, J. (Ed.), Interpreting Political Responsibility. Princeton University Press, Princeton, NJ, pp. 26–44.

Ensminger, J., Knight, J., 1997. Changing social norms: common property, bridewealth, and clan exogamy. Current Anthropology 38, 1–24.

Farrell, H., 2009. The Political Economy of Trust. Cambridge University Press, New York.

Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. Free Press, New York.

Gambetta, D. (Ed.), 1988. Trust: Making and Breaking Cooperative Relations. Blackwell, New York.

Gambetta, D., Hamill, H., 2005. Streetwise: How Taxi Drivers Establish Their Customers' Trustworthiness. Russell Sage, New York.

Hardin, R., 1998. Trust in government. In: Braithwaite, V., Levi, M. (Eds.), Trust and Governance. Russell Sage Foundation, New York, pp. 9–27.

Herreros, F., 2004. The Problem of Forming Social Capital. Palgrave Macmillan, New York.

Kramer, R.M., Tyler, T.R. (Eds.), 1996. Trust in Organization: Frontiers of Theory and Research. Sage Publications, Thousand Oaks, CA.

Kreps, D.M., 1990. Corporate culture and economic theory. In: Alt, J.E., Shepsle, K.A. (Eds.), Perspectives on Positive Political Economy. Cambridge University Press, Cambridge, pp. 90–143.

Levi, M., 2000. When good defenses make good neighbors. In: Menard, C. (Ed.), Institutions, Contracts, and Organizations: Perspectives from New Institutional Economics. Edward Elgar, Northampton, MA, pp. 137–157.

Levi, M., Sacks, A., December 2009. Legitimating beliefs: concepts and measures. Regulation & Governance 3, 311–333.

Levi, M., Sacks, A., Tyler, T.R., 2009. Conceptualizing legitimacy, measuring legitimating beliefs. American Behavioral Scientist 53 (3), 354–375.

Levi, M., Stoker, L., 2000. Political trust and trustworthiness. Annual Review of Political Science 3, 475–507.

Miller, G.J., 1992. Managerial Dilemmas. Cambridge University Press, Cambridge.

Nannestad, P., 2008. What have we learned about generalized trust, if anything? Annual Review of Political Science 11 (1), 413–436.

Putnam, R., Leonardi, R., Nanetti, R.Y., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton University Press, Princeton, NJ.

Rothstein, B., 2011. The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective. University of Chicago Press, Chicago, London.

Seligman, A., 1997. The Problem of Trust. Princeton University Press, Princeton, NJ.

Silver, A., 1985. Trust in social and political theory. In: Suttles, G.D., Zald, M.N. (Eds.), The Challenge of Social Control: Citizenship and Institution Building in Modern Society: Essays in Honor of Morris Janowitz. Ablex, Norwood, NJ, pp. 52–67.

Tyler, T.R., 1998. Trust and democratic government. In: Braithwaite, V., Levi, M. (Eds.), Trust and Governance. Russell Sage Foundation, New York, pp. 269–294.

Tyler, T.R., Huo, Y.J., 2002. Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell Sage Foundation, New York.

Uslaner, E., 2002. The Moral Foundations of Trust. Cambridge University Press, New York.

Williamson, O.E., 1993. Calculativeness, trust, and economic organization. Journal of Law and Economics 36, 453–486.

Yamagishi, T., Yamagishi, M., 1994. Trust and commitment in the United States and Japan. Motivation and Emotion 18 (2), 129–166.

Trust: Philosophical Aspects

Karen Jones, The University of Melbourne, Parkville, VIC, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines what trust is, its role, and value in human lives, and when it is justified. It surveys accounts of trust, ranging from the relatively open, which equate trust with reliance, through to more restrictive accounts, which insist that reliance be accompanied by confidence, to highly restrictive accounts which demand that the reliance involved in trust be grounded in specific assumptions about the motivation of the one trusted, such as that they have goodwill toward the truster. It examines whether we need reason to trust, and if we do what kind of reasons can make trust rational.

Three central questions in the philosophy of trust are What is the role and value of trust in human lives? How are trust and trustworthiness to be analyzed? When is trust justified? This article addresses each question in turn; however, the answers to them are interconnected. Any plausible account of what trust is must be capable of explaining its role and value. Any account of what trust is sets constraints on when and how trust can be justified and, since trust and trustworthiness are paired concepts that must be analyzed in tandem, also sets constraints on the analysis of trustworthiness. In the philosophical literature on trust, there is general agreement that trust has both instrumental value for securing the goods of cooperation and intrinsic value as a constitutive part of relationships, such as friendship, that are valuable in themselves. There is also general agreement that trust is a three-place relation: A trusts B to do Z. Disagreement emerges, however, regarding how trust is best analyzed. The least restrictive analysis of trust equates trust with reliance; a more restrictive account insists that reliance be accompanied by confidence; yet more restrictive accounts of trust demand that the reliance involved in trust be grounded in specific assumptions about the motivation of the one trusted, such as that they have goodwill toward the truster, care about maintaining a relationship with the truster, or recognize that they are liable to be held accountable for their failure to act in the way counted on.

The Role and Value of Trust

Trust makes it possible for us to enter, willingly and with confidence, into relationships of dependency. As finite and social creatures, we are dependent on one another in multiple ways. The food we eat typically comes to us having passed through many hands in the stages from production, to distribution, to purchase. Our health rests on the knowledge and professional competence of a host of individuals and institutions, our happiness on the goodwill of friends and strangers, our projects on the contributions of others, our knowledge on testimony and deference to experts, our economies on the competences of government institutions, and so on, through the myriad of dependencies that make our lives possible. In youth and in old age, dependency is even more pronounced.

Dependency brings with it risk: one may be left worse off than one would have been had one not depended. Trust makes such risks seem worth the taking. Trust is required for the proper functioning of divisions of economic and epistemic labor, divisions that can greatly increase the speed with which goods, including knowledge, are produced and, thus, the complexity and scope of the projects that can be undertaken. In these ways, trust, when met with trustworthiness, is shown to have instrumental value insofar as it contributes to bringing about the goods of cooperation.

Social scientists argue that trust is an important form of social capital, the absence of which hampers economic development (Putnam, 1995; Fukuyama, 1995; but for an opposing view see Hardin, 2002). Societies with high levels of trust present more economic opportunities than societies with low levels of trust. Moreover, trust enables transaction costs to be lowered by reducing the need for expensive substitutes for trust such as legal remedies and close monitoring of performance. Monitoring and legal remedy presuppose a relatively determinate understanding of what counts as an acceptable discharging of the responsibilities of a dependency relationship. Trust allows the truster to give trustees discretionary powers with respect to how they fulfill that trust (Baier, 1986: p. 237). Discretionary powers greatly enhance the usefulness of dependency relationships: often the truster cannot precisely specify in advance what it is that they are counting on the other to do either because the truster does not have the ability to predict the contingencies that must be planned for (think of the discretion given to babysitters), or because they do not have the knowledge to specify what would count as an appropriate response to such contingencies. The limits of the trustee's discretionary powers are only vaguely specified and this vagueness opens up the possibility for abuse – an untrustworthy trustee might conceal subtle breaches of trust as appropriate uses of her discretionary powers. Assigning discretionary powers thus brings with it additional risks as well as benefits.

Well-placed trust not only has objective benefits assessable in terms of the results of successful dependencies, it also has subjective benefits. Trust lets us inhabit a simpler and less threatening world in which we need not plan for every contingency (Luhmann, 1973). Trust enables us to forgo anxious monitoring of performance and to feel secure in our interactions with others.

Trust's value is not exhausted by the objective and subjective benefits it brings. Trust also has intrinsic value insofar as it is

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63088-X

constitutive of relationships that are valuable in themselves, such as friendship. Likewise, preemptive distrust, especially when incurred in virtue of one's membership in a negatively stereotyped group, not only has instrumental disvalue insofar as those unfairly distrusted are cut off from the benefits of joint projects that presuppose trust, it also communicates disrespect. The value we assign to being trusted and the disvalue we assign to being distrusted is thus not exhausted by the consequences of others holding those attitudes toward us.

This brief description of the role and value of trust in our lives points to the significance of developing an account of what trust is and also sets constraints on that account: any account of trust must be able to explain how it can fulfill the roles and have the value that it has, or, failing that, must explain how we could be confused about trust's role and value.

Accounts of Trust and Trustworthiness

While trust might appear to be a dyadic relation of the form 'A trusts B' (e.g., the citizens trust their government), trust is best analyzed as a three-place relationship: A trusts B to do Z (Baier, 1986; Holton, 1994). Even among close friends, trust seldom extends to all domains of interaction and there is typically some background assumption in place about that with respect to which the other is trusted. The triadic nature of trust shows that trust need not be an all or nothing affair – one can trust with respect to the performance of some actions, but not others. This opens up the possibility of incrementally increasing trust by gradually extending the range of actions that the other is counted on to do. Paradigmatically, the 'A' and 'B' of the schema 'A trusts B to do Z' stand for persons. However, nonnatural agents, such as corporations and governments can also fill either role, as can perhaps nonhuman animals. Furthermore, on some accounts of trust, but not others, things such as machinery and ropes can occupy the role of B.

Accounts of trust can be taxonomized along two independent axes. The first axis concerns the attitude or stance that trust is taken to be. Is trust actionlike, belieflike, affective attitude-like, some combination of all three, or something different again? The second axis concerns the motivation that trust tacitly ascribes to the one trusted. This axis marks disagreement over the nature of trustworthiness. Accounts can be placed on a graded scale with respect to how open or restrictive they are about the motivational structure that trust tacitly ascribes to the one trusted. A thin account places no restrictions at all on the motivations of the trustworthy; highly restrictive accounts demand that the trustworthy act from goodwill toward the truster (Baier, 1986; Jones, 1996) or from moral integrity (McLeod, 2002). A variety of positions lie in between these extremes.

The simplest account of trust, trust-as-reliance, is open with respect to motivation and emphasizes the respects in which trust is actionlike and seemingly something we can will. On this model, agents can trust machinery and ropes as well as other agents, natural and artificial. However, agents will count as trusting even when they do not expect a favorable outcome. For example, someone may decide to rely on a frayed rope that they do not believe is strong enough to hold them – perhaps it is better to risk probable death or injury from a fall than certain death from the oncoming fire. A trust-as-reliance model conforms to ordinary linguistic usage, for it would not be incorrect to say that such a person decided to trust the rope rather than to take their chances with the fire. Nevertheless, the trust-as-reliance model identifies too heterogeneous a class of dependencies to explain trust's role and value in human lives. For example, the claim that trust contributes to social capital and reduces transaction costs is false if trust is understood as reliance with or without confidence. Someone who relies while expecting the worst will be on the lookout for alternatives to such reliance and will cease to rely once another option emerges. Such wary and reluctant dependence cannot be thought of as social capital. If we are to understand the value of trust, we need a richer conception of it.

According to Thomas Hobbes, "[t]rust is a passion proceeding from the belief of him whom we expect or hope for good, so free from doubt that upon the same we pursue no other way" (1994[1650]: IX 9). Call this the *trust-as-confidence* model. Hobbes emphasizes the affective nature of trust, whereas contemporary trust-as-confidence models emphasize cognitive-based risk assessment. Thus Diego Gambetta writes:

> Trust (or, symmetrically, distrust) is a certain level of subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action.
>
> Gambetta, 1988: p. 217, original emphasis

Confidence can be grounded in all sorts of properties of the one trusted. An agent might trust someone to do Z because they are motivated by self-interest, goodwill, fear, or just because they are in the habit of doing Z. Trust-as-confidence remains a comparatively thin conception of trust. It is able to explain the instrumental value of trust in enabling cooperation, but is unable to explain the ways in which we take trust to have intrinsic value or to explain why we should be distressed by preemptory distrust. The expressive regard shown by trust or distrust can be explained neither by the motivation trust attributes to the one trusted, since any motivation will do, nor by the attitude that trust is, whether this attitude is understood as cognitive or affective.

Richard Holton's account of trust fares better at explaining the intrinsic value of trust while nevertheless remaining open ended about the motivation that trust tacitly ascribes to the one trusted. Drawing on Peter Strawson's (1974) notion of reactive attitudes – those attitudes that we take toward people when we view them as fellow agents who are to be held accountable for what they do – Holton argues that trust is reliance invested with reactive attitudes: "When you trust someone to do something you rely on them to do it and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld" (Holton, 1994: p. 67). When you trust, you do not treat the other as a mere regularity to be worked with or around, but treat them as a fellow agent, subject to normative expectations (or expectations *of*) and not merely predictive expectations (or expectations *that*). (For the distinction between normative and predictive expectations, see Hollis, 1998.) However, it fares less

well than trust-as-confidence accounts in explaining the instrumental value of trust, since, like simple trust-as-reliance accounts, it groups together both confident and reluctant reliances. It is possible to have the dispositions to reactive attitudes that Holton claims are constitutive of trust toward reliances that have been undertaken unwillingly and only because no safer choice presented itself. Merely having dispositions to reactive attitudes does not facilitate cooperation. Confidence seems characteristic of trust, suggesting that prospects for an account of trust fully able to explain its role in our lives might lie in refining and further restricting a trust-as-confidence view.

Russell Hardin's encapsulated interest model of trust proposes to distinguish trust from mere reliance in the following way: "it is trust I have, and not merely inductive expectations *if* I think that you have a reason in some way grounded in me for fulfilling my trust. For example, I might well expect you to rise every morning at 7:00 because I know your habits very well, but I do not trust you to do so if I have no reason to suppose that you do so somehow on my behalf" (1998: p. 12). The reasons for the one trusted performing as they are being counted on to do must make reference to the truster. The reference to 'on my behalf', should not, however, be taken to indicate that Hardin thinks trust requires attributing to the one trusted a direct concern for the truster's welfare. Hardin thinks that typically the one trusted will have self-interested reasons for not letting the truster down. This interest frequently derives from a desire to maintain an ongoing relationship with the truster.

Other, yet more restrictive, accounts see reliance on self-interest as incompatible with trust. Goodwill-based accounts claim that trust tacitly imputes goodwill to the one trusted. According to Annette Baier's influential account, trust is "accepted vulnerability to the other's possible but not expected ill will (or lack of good will) towards one" (Baier, 1986: p. 234). The distinction between trust and mere reliance is the difference between "reliance on their goodwill towards one, as distinct from their dependable habits, or only on their dependably exhibited fear, anger, or other motives compatible with ill will towards one, or on motives not directed on one at all" (Baier, 1986: p. 234). A somewhat different account that shares the intuition that goodwill is central but that stresses the importance of affect or emotion is offered by Jones (1996), who defines trust as an affective attitude of optimism that the goodwill and competence of the other will extend to cover the domain of your interaction, together with the confident expectation that the other will be directly and favorably moved by the thought that you are counting on them. Other theorists (McCleod, 2002; Mullin, 2005) argue that trust tacitly assumes a commitment to shared norms, or that trust tacitly ascribes obligations to the one trusted (Nickel, 2007) or that trust presupposes that the one trusted will act as they are being counted on in part out of recognition that they will be held accountable should they fail so to act (Walker, 2006: p. 80).

The richer the conception of trust, the more likely it is to be able to explain why we take trust to have intrinsic value and the more satisfying the picture it gives of the role of trust in interpersonal relationships such as friendship. Richer conceptions can also explain why trust should be instrumentally valuable, for clearly joint projects will go better when participants have goodwill or shared values. These advantages come at a cost, however, since the richer the conception of trust, the less likely trust is to be in good supply once we step outside the realm of close interpersonal relationships. Highly restrictive accounts of trust have their natural home in theorizing about face-to-face relationships where knowledge about the likely motivations and evaluative outlooks of those with whom we interact is more easily come by. Less restrictive accounts seem better equipped to address trust in more impersonal contexts, including institutional contexts, where it seems we willingly and confidently rely without having much information about the reasons others have to act or about the internal workings and structures of institutions. There are thus advantages and disadvantages to each of the available philosophical accounts of trust and no single account is able to offer a wholly satisfactory explanation of the role and value that trust has in our lives. Perhaps the explanation for this is that we do not have a single pair of trust and trustworthiness concepts, but instead use different concepts in different contexts for difference purposes. If this is so, then we may have to give up on the project of coming up with a single account of trust, and focus instead on analyzing trust in specific contexts.

Justified Trusting

If trusting is to bring about the goods of cooperation, it must be met with trustworthiness. Trust in the untrustworthy breeds exploitation and abuse. Moreover, the costs of misplaced trust are frequently borne by others more vulnerable than the truster – think, for example, of child abuse. Call trust that succeeds in being directed toward the trustworthy 'apt' or 'fitting' trust. Just as true belief must be distinguished from justified belief, apt trust must be distinguished from justified trust. Sometimes, despite our trust being well grounded in evidence available to us, we nonetheless have the misfortune to trust someone who is not trustworthy; other times, we escape getting burned despite trusting unwisely. The problem of being justified in one's trust raises a cluster of questions, some theoretical and some practical: What kinds of reasons can provide grounds for trust? How much evidence of trustworthiness, if any, do we need before our trust counts as justified? How can trustworthiness be promoted and better communicated so that would-be trusters will know where to place their trust and where not? I address each of these questions in turn.

How a theorist analyses trust sets constraints on what they can say about when trust is justified and on the kind of reasons that can justify it. If a theorist holds that trust is fully or partly constituted by a belief that the other is trustworthy, then they must also hold that trust is justified by evidence that would support a belief in trustworthiness. Trust cannot be justified by instrumental or moral reasons, since these are the wrong kind of reasons to support belief in trustworthiness (Hieronymi, 2008).

The problem of what reasons are capable of justifying trust can be bought into focus by considering therapeutic trust, or trust undertaken in the hope of eliciting trustworthiness from someone whose past track record would not support a judgment that they are trustworthy (Horsburgh, 1960). Theorists who take trust to be partly or fully constituted by a belief that the other is trustworthy (Hardin, 2002; Hieronymi, 2008) must

deny that therapeutic trust is genuine trust. However, consider the case of a parent who decides to let their teenaged daughter stay at home overnight without supervision, even while recognizing that she may again show herself irresponsible. It seems that the parent trusts in order to bring about the very trustworthiness that would make their trust apt. Such strategies work, when they do, by eliciting trust responsiveness on the part of the one trusted. Pettit defines trust responsiveness is the "disposition to prove reliable under the trust of others" (Pettit, 1995: p. 203) and argues that it can be grounded in a variety of elements in an agent's motivational set, including positively valued traits such as loyalty, virtue, and prudence, as well as traits that are not so well regarded, such as desire for the good regard of others. Another psychological spring for trust responsiveness is to be found in the agential scaffolding provided by the optimistic truster's image of ourselves as trustworthy, an image we then try to live to up to (McGeer, 2008).

Trust responsiveness has a negative corollary in distrust responsiveness. If others become aware of our distrust, either because we refuse dependency, or because we engage in intrusive surveillance, we risk undermining the positive effect of trust responsiveness and, further, we risk alienating the one being monitored and thus activating untrustworthiness. Surveillance generates resentment, undermines internal motivations toward trustworthiness, and encourages even the honest to attempt to beat the surveillance system (Kramer, 1999: p. 591). It is widely observed (e.g., Baier, 1986; Baker, 1987; Jones, 1996) that trust and distrust are self-confirming. Trusting is often met with perceived trustworthiness and distrust with perceived untrustworthiness. In part, this is to be explained by perception – someone who trusts will interpret the one trusted's actions through the lens of that trust. An action that might have been seen as indicative of untrustworthiness will be explained away as compatible with trustworthiness (Jones, 1996). Equally, someone who distrusts will tend to interpret actions that could be given a favorable interpretation as nevertheless revealing motives indicative of untrustworthiness ('what does she want from me that she is being so nice?'). In part, it is to be explained by opportunities to prove oneself trustworthy: where possible, one steers clear of those one does not trust and so they are never given the chance to demonstrate their trustworthiness. But, in large measure, the self-fulfilling nature of trust and distrust is to be explained by trust responsiveness and its corollary distrust responsiveness.

Once we settle the question of what kinds of reasons – only evidential reasons, or also goal directed reasons or moral reasons – can justify trust, the question remains how strong those reasons must be. How much, if any, evidence do we need of the other's trustworthiness before our trust is justified? How confident, if at all, must we be that goal directed trust will be met with trust responsiveness before instrumental trust, if indeed it is a genuine form of trust, can be justified? Some claim that, on moral grounds, we should begin from a default stance of trust and move from this default only if we have positive evidence of untrustworthiness (Lagerspetz, 1998). Others argue that this question lacks a general answer. How much, and what sort of information agents need before it is rational for them to trust depends on the climate of trustworthiness they inhabit, on what it is that the other is being trusted to do, and on the cost of misplaced trust (Jones, 1996). When the climate is sufficiently bad it might be reasonable to begin from a default stance of distrust and move toward trust only when there is considerable evidence of the trustworthiness of the other. Some actions, such as refraining from harming others as they go about their business, require only minimal decency and no particular competence and thus, unless the climate of trust has broken down completely, it can be assumed that trust is warranted. Other actions will be such that their performance demands competencies or motivational resources that cannot be assumed to be widely distributed – in such cases, the truster will require specific evidence about the trustworthiness of the other. Such evidence may take the form of marks that indicate that the trustee belongs to some subset of the population in which trustworthiness with respect to the required action will be the norm (e.g., board-certified surgeons), or it may take the form of evidence about motivations and competencies specific to the trustee. The consequences of misplaced trust also affect how much evidence is required for rational trust – holding constant climate and kind of action to be performed, it will take more information before I am justified in trusting when I risk a grave loss should the other prove untrustworthy than when the loss is less.

The problem of justified trust has a practical face as well as theoretical one: how can we create societies in which our trust is more likely to be met with trustworthiness and how can would-be trusters work out where to direct their trust appropriately? Intelligently designed systems of accountability can address both problems (O'Neill, 2002, 2005). Trust should not be blind and is not necessarily in tension with accountability. A well-designed system of accountability addresses three concerns: it can help create trustworthiness in the domain it regulates by creating competence (e.g., through systems of training and credentialing) and by creating a confluence of motives toward acting in the ways expected (e.g., through watch dog and oversight structures); it can increase awareness, on the part of those we would trust, of the extent and limits of their competences, and so of the domains in which they are able to respond to trust with trustworthiness; and it can communicate to would-be trusters who is likely to be a fit target for their trust. However, to play these roles, a system of accountability must be well designed. Poorly designed accountability systems (and there are many) undercut internal motivations toward trustworthiness by replacing them with external incentives and penalties, risk activating distrust responsiveness, and substitute proxy measures, which can be checked off by people without field-specific knowledge, for real markers of performance, thereby creating perverse incentives (O'Neill, 2005).

Trustworthiness, like trust, has three-place structure: A is trustworthy with respect to B, regarding action (or action type) Z. One kind of thing with respect to which we can be trustworthy is communicating to would-be trusters those areas in which others may place their trust in us (Jones, 2012). Agents willing to do this help would-be trusters place their trust wisely by signaling which dependencies they will welcome and which they will not. They thus take on some of the burden of helping would-be trusters be justified in their trust. It is not always easy to communicate effectively that one can be trusted. Often trustworthiness cannot be conveyed

directly: "'Trust me!' is for most of us an invitation which we cannot accept at will – either we do already trust the one who says it … or it is properly responded to with, 'Why should I and how can I until I have cause to?'" (Baier, 1986: p. 244). Trusted third parties can thus be important brokers for new trust relationships.

Though some ways of communicating trustworthiness are formal and institutionally backed (e.g., board certification for surgeons), many are informal and rest on self-presentation, body language, and the workings of tacit assumptions regarding who can be trusted by whom for what. Sometimes, attempts to communicate one's trustworthiness in a domain fail because uptake is blocked by stereotype and prejudice (Fricker, 2007). In a related but reversed phenomenon, members of positively viewed groups can be the recipients of unearned trust simply in virtue of their social identities. Prejudice and stereotype are thus major obstacles to justified trusting.

Conclusion

Perhaps the central question in the philosophy of trust is whether and how to separate trust from mere reliance. All plausible philosophical accounts of trust attempt some form of separation between these two phenomena, but there is nonetheless significant disagreement regarding: (1) the attitude they take trust to be and whether this is best analyzed on the model of action, belief, emotion, some combination, or something else again; and (2) the motivation they take trust to tacitly impute to the one trusted. Analyses of trust set constraints on what can be said about trustworthiness and also on when, and by what sorts of reasons, trust can be justified. Answers to these theoretical questions have practical implications for devising strategies to increase the likelihood that trust will be met with trustworthiness.

See also: Social Capital; Virtue Ethics.

Bibliography

Baier, A., 1986. Trust and anti-trust. Ethics 96, 231–260.
Baker, J., 1987. Trust and rationality. Pacific Philosophical Quarterly 68, 1–13.
Fricker, M., 2007. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford.
Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. The Free Press, New York.
Gambetta, D., 1988. Can we trust trust? In: Gambetta, D. (Ed.), Trust: Making and Breaking Cooperative Relations. Basil Blackwell, New York.
Hardin, R., 1998. Trust in government. In: Braithwaite, V., Levi, M. (Eds.), Trust and Governance. Russell Sage Foundation, New York.
Hardin, R., 2002. Trust and Trustworthiness. Russell Sage Foundation, New York, NY.
Hieronymi, P., 2008. The reasons of trust. Australasian Journal of Philosophy 86, 213–236.
Hobbes, T., 1994[1650]. Human Nature. The Elements of Law, Natural and Politic. Oxford University Press, Oxford, UK pt. 1.
Hollis, M., 1998. Trust Within Reason. Cambridge University Press, Cambridge, UK.
Holton, R., 1994. Deciding to trust, coming to believe. Australasian Journal of Philosophy 72 (1), 63–76.
Horsburgh, H.J.N., 1960. The ethics of trust. Philosophical Quarterly 10, 343–354.
Jones, K., 1996. Trust as an affective attitude. Ethics 107, 4–25.
Jones, K., 2012. Trustworthiness. Ethics 123, 61–85.
Kramer, R.M., 1999. Trust and distrust in organizations: emerging perspectives, enduring questions. Annual Review of Psychology 50, 569–598.
Lagerspetz, O., 1998. Trust: The Tacit Demand. Kluwer, Dordrecht.
Luhmann, N., 1973. Vertrauen [Trust: a mechanism for the reduction of social complexity]. In: Trust and Power. Wiley, Chichester, UK, pp. 4–103.
McCleod, C., 2002. Self-Trust and Reproductive Autonomy. MIT Press, Cambridge, MA.
McGeer, V., 2008. Trust, hope, and empowerment. Australasian Journal of Philosophy 86, 237–254.
McLeod, C., 2002. Self-trust and Reproductive Autonomy. MIT Press, Cambridge, MA.
Mullin, A., 2005. Trust, social norms, and motherhood. Journal of Social Philosophy 36, 316–330.
Nickel, P., 2007. Trust and obligation-ascription. Ethical Theory and Moral Practice 10, 309–319.
O'Neill, O., 2002. Autonomy and Trust in Bioethics. Cambridge University Press.
O'Neill, O., 2005. Justice, trust and accountability. Published as Gerechtigkeit, Vertrauen und Zurechenbarkeit. In: Neumaier, Otto, Sedmak, Clemens, Zichy, Michael (Eds.), Gerechtigkeit: Auf der Suche nach eimen Gleichgewicht. Ontos Verlag, Frankfurt, pp. 33–55.
Pettit, P., 1995. The cunning of trust. Philosophy and Public Affairs 24, 202–225.
Putnam, R., 1995. Bowling alone: America's declining social capital. Journal of Democracy 6, 65–78.
Strawson, P., 1974. Freedom and Resentment. In his Freedom and Resentment. Methuen, London.
Walker, M.U., 2006. Moral Repair: Reconstructing Moral Relations after Wrongdoing. Cambridge University Press, Cambridge, UK.

Truth and Credibility in Science

Steven Shapin, University of California-San Diego, La Jolla, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15926–15932, © 2001, Elsevier Ltd.

Abstract

The natural sciences occupy a special place in modern culture. Scientific knowledge substantially defines our culture's notion of truth, and the overwhelming credibility of that knowledge in our society is widely taken to be sufficiently explained by its truthfulness. The social sciences have traditionally concurred in these assessments and have paid the natural sciences the compliment of emulation. Yet that emulation has largely proceeded by way of textbook idealizations of natural scientific knowledge and how it is produced, while the idealizations themselves – produced by both academic epistemologists and reflective scientists – are much more variable than is generally supposed. It has not been until fairly recent times that social scientists have attempted wholly naturalistic inquiries into how scientific knowledge is produced and how it comes to be believed. The sociology of scientific knowledge that emerged from the early 1970s purposely set aside epistemological evaluation and took as its primary task understanding scientific credibility. How is it that scientific claims succeed, or fail, in being believed? This article surveys a number of considerations that recent social studies of science have found to be widely relevant to the interpretation of scientific credibility.

The natural sciences occupy a special place in modern culture. Scientific knowledge substantially defines our culture's notion of truth, and the overwhelming credibility of that knowledge in our society is widely taken to be sufficiently explained by its truthfulness. The social sciences traditionally have concurred in these assessments and have paid the natural sciences the compliment of emulation. Yet that emulation largely has proceeded by way of textbook idealizations of natural scientific knowledge and how it is produced, while the idealizations themselves – produced by both academic epistemologists and reflective scientists – are much more variable than generally is supposed. It has not been until fairly recent times that social scientists have attempted wholly naturalistic inquiries into how scientific knowledge is produced and how it comes to be believed. The sociology of scientific knowledge that emerged from the early 1970s purposely set aside epistemological evaluation and took as its primary task understanding scientific credibility. How is it that scientific claims succeed, or fail, in being believed? This article surveys a number of considerations that recent social studies of science have found to be widely relevant to the interpretation of scientific credibility.

Scientific Truth: As Evaluation and Resource

'Truth' appears only occasionally as a vernacular term, nor does the special notion of intellectual quality the word connotes in academic philosophy figure much in everyday usages. Professional epistemologists worry how to reliably sort genuine knowledge from belief, opinion, or mere knowledge-claims, and this is the major context in which academic theorizing about truth takes place. Many theorists are agreed that knowledge (properly so-called) is 'true belief,' but there is no philosophical consensus about the definition of truth, about the domains in which we can expect to have genuine knowledge, about the grounds for establishing truth, or, indeed, about what it is that can be said to be true (propositions, facts, theories, thoughts, concepts, beliefs, etc.). The major battle pitches correspondence theories against coherence theories of truth, while skirmishes are fought along the margins between conventionalist, pragmatist, performative, redundancy, disquotational, and miscellaneous other theories (or antitheories) of truth.

In ordinary life, the word truth often crops up as a special-purpose term whose employment indicates that some out-of-the-ordinary guarantees are being sought for promises, professions, or claims to knowledge, or that there is some specially occasioned doubt about their genuine status, as in legal proceedings when witnesses are ceremonially sworn 'to tell the truth, the whole truth, and nothing but the truth' (Lynch and Bogen, 1996), or in intimate social relationships when one has misgivings about one's lover's sincerity and looks for explicit warrants of affection: 'When my love swears that she is made of truth,/I do believe her, though I know she lies.' Perhaps even more commonly, everyday actors, and even philosophers in many circumstances, use the words true or truth just as a way of giving added emphasis or value to assertions of what is the case, the lexical equivalent of thumping the table.

The category of everyday actors may be taken, for the most part, to include professional natural scientists engaged in their ordinary knowledge-producing practices. While their expert employments generate knowledge whose content, methods of production, and value are widely said to differ markedly from lay beliefs about the world and lay modes of cognition and evaluation, epistemological reflection does not occupy much of scientists' time, nor is it customary for them to engage in metascientific inquiries into the nature and status of their knowledge. The result is a situation in which the bodies of expert knowledge that are widely taken as the touchstone of truth in our culture, and the object of vast amounts of epistemo- logical reflection by philosophers, tend to be produced by practitioners who rarely concern themselves reflectively or systematically with the nature of their knowledge, or its standing vis-à-vis nonexpert knowledge, or with the methods that produce their so-highly-valued knowledge. The paucity of practitioners' reflection on the nature and quality of their knowledge is an important feature of the modern natural sciences, and the social sciences and humanities are quite different in this respect.

Practicing natural scientists seem to take scarcely any notice of epistemologists' debates or dicta. Indeed, in the natural sciences, epistemological and metascientific musings, as well as sociological reflections, are often taken as a sign of emeritus (or, unkindly, 'past it') standing. They are sometimes explicitly condemned as irrelevant, unprofessional, or even pathological: it may be what philosophers or sociologists do, but (it is said) it is not what practicing scientists ought to do. In the nineteenth century the experimental physiologist Bernard (1865/1957, p. 49) vigorously disengaged the scientific production of truth from scientists' possible reflections on truth or on how truth was produced: 'To find scientific truth, we … have little need to know how our mind reasons; it is enough to let it reason naturally.' Bridgman (1955, p. 157) was one of many modern physicists taking a jaundiced view of philosophers' attempts to characterize both scientific truth and scientific method: philosophers' 'abstractions' tended to attract a 'veneration' that was wholly undeserved. C. P. Snow – a one-time chemist and latterly a spokesman for science – bumptiously admitted that 'By *truth*, I don't intend anything complicated .I am using the word as a scientist uses it. We all know that the philosophical examination of the concept of empirical truth gets us into some curious complexities, but most scientists really don't care' (Snow, 1961, p. 247). Albert Einstein famously celebrated, rather than condemned, what he took to be the fact 'that the man of science is a poor philosopher.' If, Einstein (1936/1954, p. 296, 318) said, 'you want to find out anything from the theoretical physicists about the methods they use, I advise you to stick closely to one principle: don't listen to their words, fix your attention on their deeds.' Moreover, the small number of empirical studies we now have of natural scientists' grasp of the principles of formal logical reasoning strongly suggests that the power and reliability of scientific knowledge do not spring from the superior reasoning capacities of individual scientists (Mahoney, 1979; Mahoney and DeMonbreun, 1977).

When one considers the content of such scientists' musings about truth as are occasionally produced, further problems are posed for any social science wishing to model its epistemic goals on those of the natural sciences. Many academic, and some lay, views of the natural sciences point to scientific facts, laws, and theories as enjoying the status of accredited truth, and to the proper goals of scientific inquiry as the production of such truth. Yet important strands of practitioners' metascientific stipulations identify the goal of producing truth as illegitimate, and such special status for scientific representations as a lay misunderstanding of what genuine science is about. Some scientists do indeed say that science aims at, or even arrives at, one universal truth. However, others say that the truths of sciences are plural, or that science is just 'what works' and that truth – or even correspondence with the world – is none of their concern – just 'what stands the test of experience,' or 'what seems to be the case to the best of our current efforts and beliefs.' Oppenheimer (1954, p. 4) followed the nineteenth-century Scientific Naturalists in warning science to shun 'metaphysical' notions like 'real' and 'ultimate' in talking about its concepts and theories: if we say that physical theories are true, we must say so in a carefully qualified non-metaphysical sense. Einstein (1929/1954, pp. 261–2) found it 'difficult even to attach a precise meaning to the term "scientific truth." Thus, the meaning of the word "truth" varies according to whether we deal with a fact of experience, a mathematical proposition, or a scientific theory.' Bridgman (1955, p. 158) noted that 'In physics we have been forced to recognize … that there are different kinds of truth, depending on the method used for establishing "truth."' Bridgman's student Oppenheimer (1954, p. 95) vigorously agreed: 'Perhaps only a malignant end can follow the systematic belief that all communities are one community; that all truth is one truth.' Just as epistemologists espouse widely varying conceptions of truth, so too do those natural scientists occasionally drawn to philosophizing, while most practicing scientists seem well content to leave philosophy to the philosophers.

From Truth to Credibility: Science as a Topic

For these reasons alone, the posture of the social sciences vis-à-vis the natural sciences, and, especially, with respect to emulating their truth-criteria, is problematic. Should social scientists really wish to pattern themselves on the natural sciences in order to solve problems of intellectual identity, authority, and quality, they would find it no easy thing to do, such is the diversity of natural scientific practices and sensibilities, and of professional metascientific stipulations about the natural sciences (Mills, 1959, pp. 119–31). This diversity is seldom adequately reflected in textbook accounts of scientific method and scientific knowledge, nor is it apparent when one particular natural science – typically some version of modern physics – is made by fiat to stand for all the rest (Shapin, 1999a). Yet once this variability is properly appreciated, it becomes clear that the social sciences will have to make their own claims to knowledge and support them as best they can, using whatever resources and warrants are available and are deemed legitimate.

However, misguided emulation is not the only possible mode of social scientific engagement with the natural sciences. Social scientists can study the cultures of the natural sciences, seeking to understand how a range of natural scientists collectively produce their knowledge, how they make it persuasive, and how they render their intellectual evaluations. This has been the task taken up by the sociology of scientific knowledge over the past 30 years or so. Indeed, the founding gesture of modern sociology of scientific knowledge was the rejection of epistemological evaluation as the proper basis for sociological accounting. Traditional sociology of knowledge – descended from Marx and Mannheim – had reckoned, almost as a matter of course, that true and false beliefs should be explained differently. False beliefs might customarily be referred to distorting social and psychological factors, while the truth of true beliefs – their adequate evidential base, or their logical coherence, or their production and assessment using transparent and impersonal rational methods – could count as its own sufficient explanation (Mannheim, 1936, pp. 239, 261; Phillips, 1974). It was not considered that much of sociological interest could be said about true beliefs, and therefore about the body of currently accepted scientific knowledge. That is why traditional sociology of knowledge explicitly exempted accepted scientific beliefs from its scope.

Bloor's (1976/1991) *Strong Programme* in the sociology of scientific knowledge enjoined sociologists professionally to set aside external judgments of truth and falsity and to explain the acceptance of both reputed true and reputed false beliefs using the same resources. It was not that sociologists were to account all beliefs of equal value – that would itself be an act of epistemological assessment – rather it was that sociologists, in doing sociology, should be impartial with respect to the reputed truth or falsity of beliefs, as true and false seemed to the individual sociologist or as the beliefs were routinely sorted in the culture of which the sociologist was an everyday member. Bloor here identified generally acknowledged academic virtues that would flow from adhering to this maxim of method and academic vices that had traditionally stemmed from rejecting it. On the one hand, it was not the proper business of the social scientist to engage in epistemological assessment; on the other hand, inscribing external judgments of truth and falsity in social scientific accounts illegitimately blocked the free inquiry and rendered explanation circular. In Bloor's view, the sociologist could and should be as curious about why some cultures – in this case, our own – classify the cassowary as a bird as they had traditionally been about such cultures as the New Guinea Karam that classify what we call cassowaries and what we call birds in different taxa (Bulmer, 1973; Barnes, 1973; Barnes and Bloor, 1982).

The *Strong Programme* did not mean to ban consideration of truth and falsity but, rather, to encourage sociologists to take a fuller naturalistic interest in the problem of how actors themselves sorted beliefs as between true and false. Actors' truth- judgments, or judgments of what is the case about the world, were not to be just an asymmetrical resource but a central topic of sociological inquiry. Or, put another way, this newer form of the sociology of knowledge placed the credibility of beliefs at the center of its intellectual agenda. Suppose, taking up the toy taxonomical example, one asked a member of the Karam why it was that they classified cassowaries in a different taxon than the rest of what we call birds. After translating the question into their own vocabulary, the Karam might reply in the prevalent folk- epistemological idiom, rolling their eyes, or saying that the matter is self-evident, or citing all sorts of obvious and important empirical differences between the big, ground-dwelling cassowary and the small, aerial bird-of-paradise. Pressed further, the Karam might point to the views of legitimately respected elders and experts, as sanctioned by custom and use. And so it might plausibly strike the Western social scientist that the grounds of credibility of Karam taxonomy had a lot to do with authority, custom, socialization, and with the vehicles that routinely transmit and maintain culture, including classificatory culture.

Patterns of routine socialization are not the only considerations in explaining structures of credibility – indeed, the anthropologist who most closely studied the particular case in question laid greatest stress upon the supposed functions of Karam taxonomy in sustaining social order (Bulmer, 1973) – but socialization is everywhere pertinent to giving accounts of what people find credible, whether the socializing agencies are those of Karam agricultural society or departments of zoology in modern American universities (Barnes, 1972: p. 271). So, if the sociologist were to put a parallel classificatory question to an academic practitioner of modern Linnaean taxonomy, it is arguable that the responses received would structurally resemble those supplied by the Karam, and, again, patterns of socialization would be major aspects of any explanation the sociologist might offer of the grounds of credibility of taxonomic beliefs in our own expert culture. Note that the empirical adequacy of Linnaean classification is not thereby disparaged – though there is substantial controversy about its theoretical and evidential adequacy within expert taxonomic culture – but the sociologist of scientific knowledge holds that standards of empirical adequacy, and the particular goals towards which classificatory schemes are oriented, are themselves objects of socialization (Dean, 1979). All institutionalized beliefs about nature are causally connected to reality, and 'all beliefs are on a par with one another with respect to the causes of their credibility. It is not that all beliefs are equally true or equally false, but that regardless of truth or falsity the fact of their credibility is to be seen as equally problematic' (Barnes and Bloor, 1982, p. 23; Barnes, 1992). The 'paradigms' of what Kuhn called 'normal science' are exemplary items of technical knowledge, and they are also learned bits of culture. Since this culture is transmitted in authoritative institutions, the sociologist's whole range of resources for dealing with socialization and its consequences is, therefore, pertinent to explaining the credibility of technical scientific knowledge among expert practitioners (Polanyi, 1958; Kuhn, 1962; Barnes, 1982).

Accounting for Credibility: Mundaneness and the Moral Order

So the sociologist of scientific knowledge studies credibility. What actors find credible, and how they account it credible, may include specially occasioned judgments of truth or they may not. The distinction is not crucial, since the credibility of any such reflective truth-judgments (by way of correspondence or coherence or any other criterion) may also be a matter for inquiry. The sociologist, qua sociologist, here differs from the philosopher (and maybe from the layperson): 'Instead of defining (knowledge) as true belief, knowledge for the sociologist is whatever people take to be knowledge,' what they find credible (Bloor, 1976/1991, p. 5). And patterns of socialization are everywhere pertinent to giving a sociological account of the credibility of beliefs. These are, so to speak, ground-conditions for the sociological study of science, and, as such, they are not themselves answers but necessary conditions for arriving at particular answers to particular questions about particular bodies of knowledge – answers that are sociologically proper, constructive, and interesting. For sociologists to say that they are interested in credibility is not to say that they are concerned with a specific aspect or 'factor' bearing on knowledge. Credibility is not a factor in the sense in which, for traditional sociology of knowledge, 'class-allegiance' or 'religious affiliation' counted as 'social factors' contingently bearing upon knowledge – sometimes pertinent, sometimes not (e.g., Barber, 1952, p. 60). Credibility is the state achieved when knowledge-claims become knowledge, and the means for achieving credibility are the assemblage of means by which knowledge-claims become knowledge. From a pertinent point of view, credibility is the only topic for the sociologist of knowledge. As the philosophers might say, credibility has a transcendental

character for sociologists of knowledge: no credibility, no knowledge, no object of sociological study (Shapin, 1995a, pp. 257–8).

It might also be said, no credibility, no moral order. To give an account of what a community finds credible, and how it comes to find it credible, is tantamount to giving an account both of what that community is and how it experiences the world (Douglas, 1975, p. 238). So to say that sociologists of science should shift attention from truth to credibility is a way of orientating practitioners to the proper domain of inquiry, a way of discouraging them from arbitrarily limiting inquiry, and, finally, a way of reminding them of the potential pertinence of a range of social scientific and everyday resources that are available to give accounts of what people find credible.

No matter how highly valued or how technical the form of belief, the sociologist can usefully draw attention to the mundane means by which belief is made credible. In the ethnomethodological tradition, this was a maxim of method derived from the phenomenologists (especially Alfred Schutz) and from the later writings of Wittgenstein. Attention was to be directed to the 'behavior of seeking truth … the institutional and public character of truth, in contrast to the usual psychological and semantic descriptions that depict private disembodiments of that behavior.' For the sociologist, there is no other proper way of engaging with truth save through the study of what people collectively do: 'Truth resides in the rule-guided institutional procedures for conceding it.' In doing sociology, one should accept 'that there are no adequate grounds for establishing criteria of truth except the grounds that are employed to grant or concede it' (McHugh, 1970, pp. 320–1, 333–5). This sensibility strongly influenced the ethnographic study of science that began to emerge in the 1970s, especially with the work of Collins and co-workers (Collins, 1975, 1983, 1985/1992), but it also left a mark on historical studies of science concerned with how past communities established facticity, proof, and legitimate inference (e.g., Shapin, 1994; Dear, 1995; Porter, 1995; Poovey, 1998; Burke, 2000).

If, indeed, the practices for making and maintaining credible knowledge are co-extensive with the moral life of a community, then there are very few aspects of that life which do not potentially bear upon the credibility of belief. It follows that there can be no simple and coherent master-recipe for how credibility is to be achieved or model for understanding it. Much will depend upon the items of knowledge in question, the precise scenes in which the knowledge is assessed, and the state of transmitted background belief. Truth – like God and the devil – is in the details, and the sociologist wanting to understand why it is that a specific individual, or group of individuals, finds the special creation of species, or Darwinian evolution, credible may find it pertinent to consider a long list of mundane, and even some apparently trivial, features of cognition and social interaction (Shapin, 1999b, pp. 7–8, Lynch, 1993, pp. 278–88), while much work in the sociology of scientific knowledge during the 1970s and 1980s pointed to differences in collective goal-orientations (or social interests) as pertinent considerations in judgments of the credibility of scientific claims (e.g., Shapin, 1982, 1995b).

Accounting for Credibility: Economies of Trust

A large body of recent historical and sociological work has studied the means by which the veracity of scientific claims or the legitimacy of inference has been made persuasive. Because most natural know- ledge – that possessed by experts as well as by the laity – is held on the basis of trust, deferring to others whose knowledge is understood to be more direct or competent, even belief in matters of fact has to be actively secured: how persuasively to communicate the experience to be credited? how linguistically (and gesturally) to display one's entitlements to trustworthy knowledge? Studies of both historical and present-day episodes have drawn attention to the rhetorical constitution of authorial transparency, virtue, and intellectual entitlements; to the rhetorical invocation of allies leagued with the author and of the consequences of giving or withholding assent; and to the range of rhetorical means by which authors in general seek to assist the translation of individual claims into collectively credible knowledge (Latour, 1987; Shapin, 1984; 1994; Dear, 1995; Golinski, 1998, Chap. 4).

Recognizing the ineliminable role of trust in the credibility of knowledge is consequential for two diametrically opposed reasons. First, many situations present themselves – both within science and between scientific experts and the laity – in which the relative trustworthiness of sources is in doubt or is contested. The role of scientific expert testimony in courts of law, for example, is one important setting in which lay jurors are set the task of finding the signs of credibility in conflicting scientific accounts, and some studies emphasize the role of common-sense procedures in that task – monitoring testimony for manner and its sources for biasing interest (Brannigan and Lynch, 1987; Lynch and Jasanoff, 1998). Wynne's (1995) work on the public understandings of science, and specifically on the credibility of environmental radioactivity risk-assessments, stresses the complex array of considerations that bear upon lay attitudes towards which scientists, and which scientific pronouncements, are trustworthy. Epstein's (1996) study of the relations between AIDS researchers and AIDS activists is a richly detailed investigation of the grounds of trust in a situation where the very identity of laypersons and experts has been made problematic. Porter (1995) has asserted the crucial role of 'trust in numbers' in modern, complex, and highly differentiated societies, in which, as he argues, personal and familiar warrants for credibility are lacking or deemed inadequate. Jasanoff has examined how trust in expertise is secured in modern environmental science where conflicts over credibility are endemic (Jasanoff, 1992), and Hilgartner's (2000) work on a series of highly contested nutritional reports by the US National Academy of Sciences has drawn attention to the performative aspects of trust in modern expertise.

In the passages of modern 'extraordinary science' studied by Collins (1975), where the existence of specific physical phenomena and the competence of detecting instruments were simultaneously contested, members of a scientific 'core-set' reached out for a diverse array of contingent considerations to locate trustworthiness, and, therefore, to assess the credibility of technical claims: including the personal characteristics of claimants, the reputation of the institutions in which they worked, their allies, their track-records, etc. Historical studies of

scientific controversy (Shapin, 1994, esp. Chap. 6) have drawn attention to the imputed trustworthiness of the early modern gentleman in managing scientific controversy, in deciding the credibility of diverging claims, and even in deciding which controversies ought to be left undecided in favor of maintaining the interaction order.

In such lines of research, reposing trust in scientific claims and claimants appears as an actor's practical problem. Looking to see how and why actors reckon trustworthiness is a way of understanding how they deal with problematic credibility when experts disagree or when it is unclear who indeed is to count as an expert. There are plenty of such occasions in modern science, and especially in those sectors of science in which public participation in decision-making is mandated or pertinent. Is global warming a fact, and, if so, what is to be done about it? Who is to be believed about the risks of genetically modified foods and what policies should be framed for producing and labeling such foods?

Yet the second reason why trust is important in connection with scientific credibility is that the scientific community very often does speak – if not with one voice – then in a comparatively concerted way. Scientists in modern society are regarded as enormously trustworthy, and the credibility of what they concertedly deliver about the natural world is very great. Scientists are masters at trusting each other and they are masters at generating public trust in themselves. That is just another way of saying that consensual scientific knowledge is regarded in our society as the touchstone of truth. Scientific credibility may be an occasional problem, and a vitally important one, but it is also interesting to note how often in our culture the credibility of science does not manifest itself as problematic. Sociologists do not yet adequately understand how scientists achieve and maintain this trustworthiness and the resultant powers of professional credibility (Barnes, 1985, Chaps. 2–3), but the consequences of this trustworthiness for the social and intellectual order of modern society cannot easily be exaggerated (Bauman, 1992, p. 71; Shapin, 1995a). To understand the grounds of scientific credibility is to understand quite a lot about modern society. Scientific credibility is so blended into modern society's institutional life – its schools, universities, government, medicine, the military, and the media – that one cannot readily imagine credible science without the institutions which maintain and transmit it, nor credible institutions without the knowledge which gives them much of their legitimacy.

See also: Experimental Design: Overview; Experimentation in Psychology, History of; Kuhn, Thomas S. (1922–96); Science, History of; Scientific Knowledge, Sociology of; Social Science, The Idea of.

Bibliography

Barber, B., 1952/1962. Science and the Social Order. Free Press, Glencoe, IL.
Barnes, B., 1972. On the reception of scientific beliefs. In: Barnes, B. (Ed.), Sociology of Science: Selected Readings. Penguin, Harmondsworth, UK, pp. 269–291.
Barnes, B., 1973. The comparison of belief systems: Anomaly versus falsehood. In: Horton, R., Finnegan, R. (Eds.), Modes of Thought: Essays on Thinking in Western and Non-Western Societies. Faber & Faber, London, pp. 182–198.
Barnes, B., 1982. T. S. Kuhn and Social Science. Macmillan, London.
Barnes, B., 1985. About Science. Blackwell, Oxford, UK.
Barnes, B., 1992. Realism, relativism and finitism. In: Raven, D., van Vucht Tijssen, L., de Wolf, J. (Eds.), Cognitive Relativism and Social Science. Transaction, New Brunswick, NJ, pp. 131–147.
Barnes, B., Bloor, D., 1982. Relativism, rationalism and the sociology of knowledge. In: Hollis, M., Lukes, S. (Eds.), Rationality and Relativism. Blackwell, Oxford, UK, pp. 21–47.
Bauman, Z., 1992. Intimations of Postmodernity. Routledge, London.
Bernard, C., 1865/1957. Introduction to the Study of Experimental Medicine (H.C. Greene, trans.). Dover, New, York.
Bloor, D., 1976/1991. Knowledge and Social Imagery, second ed. University of Chicago Press, Chicago.
Brannigan, A., Lynch, M., 1987. On bearing false witness: Credibility as an interactional accomplishment. Journal of Contemporary Ethnography 16, 115–146.
Bridgman, P.W., 1955. Reflections of a Physicist, second ed. Philosophical Library, New York.
Bulmer, R., 1973. Why the cassowary is not a bird. In: Douglas, M. (Ed.), Rules and Meanings: The Anthropology of Everyday Knowledge. Penguin, Harmondsworth, UK, pp. 167–193.
Burke, P., 2000. A Social History of Knowledge: From Gutenberg to Diderot. Polity, Oxford, UK.
Collins, H.M., 1975. The seven sexes: a study in the sociology of a phenomenon, or the replication of an experiment in physics. Sociology 9, 205–224.
Collins, H.M., 1983. The sociology of scientific knowledge: studies of contemporary science. Annual Review of Sociology 9, 265–285.
Collins, H.M., 1985/1992. Changing Order: Replication and Induction in Scientific Practice, second ed. Sage Publications, Beverly Hills, CA.
Dean, J., 1979. Controversy over classification: a case study from the history of botany. In: Barnes, B., Shapin, S. (Eds.), Natural Order: Historical Studies of Scientific Culture. Sage, Beverly Hills, CA, pp. 211–230.
Dear, P., 1995. Discipline and Experience: The Mathematical Way in the Scientific Revolution. University of Chicago Press, Chicago.
Douglas, M., 1975. Implicit Meanings: Essays in Anthropology. Routledge & Kegan Paul, London.
Einstein, A., 1929/1954. Scientific truth. In: Einstein, A. (Ed.), Ideas and Opinions. Crown Publishers, New York, pp. 261–262.
Einstein, A., 1936/1954. Physics and reality. In: Einstein, A. (Ed.), Ideas and Opinions. Crown Publishers, New York, pp. 318–356.
Epstein, S., 1996. Impure Science: AIDS, Activism, and the Politics of Knowledge. University of California Press, Berkeley, CA.
Golinski, J.V., 1998. Making Natural Knowledge: Constructivism and the History of Science. Cambridge University Press, Cambridge, UK.
Hilgartner, S., 2000. Science on Stage: Expert Advice as Public Drama. Stanford University Press, Stanford, CA.
Jasanoff, S., 1992. Science, politics, and the renegotiation of expertise at EPA. Osiris 7, 195–217.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers through Society. Harvard University Press, Cambridge, MA.
Lynch, M., 1993. Scientific Practice and Ordinary Action: Ethno-Methodological and Social Studies of Science. Cambridge University Press, Cambridge, UK.
Lynch, M., Bogen, D., 1996. The Spectacle of History: Speech, Text, and Memory at the Iran-Contra Hearings. Duke University Press, Durham, NC.
Lynch, M., Jasanoff, S., 1998. Contested identities: science, law and forensic practice. Social Studies of Science 28, 675–686.
McHugh, P., 1970. On the failure of positivism. In: Douglas, J.D. (Ed.), Understanding Everyday Life: Toward the Reconstruction of Sociological Knowledge. Aldine, Chicago, pp. 320–335.
Mahoney, M.J., 1979. Psychology of the scientist: an evaluative review. Social Studies of Science 9, 349–375.
Mahoney, M.J., DeMonbreun, B.G., 1977. Psychology of the scientist: an analysis of problem-solving bias. Cognitive Therapy and Research 1, 229–238.
Mannheim, K., 1929/1936. Ideology and Utopia: An Introduction to the Sociology of Knowledge (L. Wirth, E. Shils, Trans.). Kegan Paul, Trench, Trubner & Co., London.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Oppenheimer, J.R., 1954. Science and the Common Understanding: The BBC Reith Lectures 1953. Simon and Schuster, New York.
Phillips, D.L., 1974. Epistemology and the sociology of knowledge: the contributions of Mannheim, Mills, and Merton. Theory and Society 1, 59–88.
Polanyi, M., 1958. Personal Knowledge: Towards a Post-critical Philosophy. University of Chicago Press, Chicago.
Poovey, M., 1998. A History of the Modern Fact: Problems of Knowledge in the Sciences of Wealth and Society. University of Chicago Press, Chicago.

Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton, NJ.

Shapin, S., 1982. History of science and its sociological reconstructions. History of Science 20, 157–211.

Shapin, S., 1984. Pump and circumstance: Robert Boyle's literary technology. Social Studies of Science 14, 481–520.

Shapin, S., 1994. A Social History of Truth: Civility and Science in Seventeenth-century England. University of Chicago Press, Chicago.

Shapin, S., 1995a. Cordelia's love: Credibility and the social studies of science. Perspectives on Science 3, 255–275.

Shapin, S., 1995b. Here and everywhere: Sociology of scientific knowledge. Annual Review of Sociology 21, 289–321.

Shapin, S., 1999a. Être ou ne pas être antiscientifique. La Recherche 319 (April), 72–79.

Shapin, S., 1999b. Rarely pure and never simple: Talking about truth. Configurations 7, 1–14.

Snow, C.P., 1961. Address by Charles P. Snow (to Annual Meeting of American Association for the Advancement of Science). Science 133, 256–259.

Wynne, B., 1995. Public understanding of science. In: Jasanoff, et al., (Eds.), Handbook of Science and Technology Studies. Sage, Thousand Oaks, CA, pp. 361–388.

Tversky, Amos (1937–96)

Baruch Fischhoff, Carnegie Mellon University, Pittsburgh, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Amos Tversky was a cognitive and mathematical psychologist who made fundamental contributions to research on judgment, decision making, and measurement. With his long-time collaborator, Daniel Kahneman, he advanced the heuristics-and-biases approach to judgment under certainty and the prospect theory approach to choice. For their joint work, Kahneman received the Nobel Prize in Economic Sciences (2002). A member of the US National Academy of Sciences, Tversky was also an author of the three-volume *Foundations of Measurement* and an influential theory of similarity judgments.

Amos Tversky was a cognitive and mathematical psychologist, whose research examined the foundations of human judgment and decision making. Much of that work described the strengths and weaknesses of the intuitive inferences that guide people as they struggle to cope with life's often diverse, complex, and subtle choices. Tversky's research typically characterized these inferences by comparing observed behavior with a formal model. Sometimes, the model was *descriptive*, predicting how people would behave if they followed basic psychological principles identified as relevant to the task. Sometimes, the model was *normative*, prescribing how ideal individuals would approach the task. Creating such models requires both mathematical and empirical expertise. Together with Daniel Kahneman, Tversky exercised that expertise in two singular accomplishments: the heuristics-and-biases approach to judgment under uncertainty and Prospect Theory for choices involving risk. The former assumes that people base their judgments on general rules of thumb that, although often useful, can produce predictable biases (e.g., exaggerating risks that the news media disproportionately report, making them unduly 'available' in memory). The latter proposes an alternative to utility theory, the normative standard for how decisions should be made, by elegantly integrating established psychological principles. With other collaborators, from diverse disciplines, Tversky developed theories of multi-attribute choice, similarity judgments, and the evaluation of evidence. His research methods ranged from pure mathematics to laboratory experiments to field studies in areas as diverse as medicine, law, and basketball. He coauthored the three-volume *Foundations of Measurement*.

Amos Tversky

Amos Tversky was born in Haifa, Israel, on 16 March 1937. His father was a veterinarian, and his mother a member of Israel's Parliament, the Knesset. He fought in Israel's 1956, 1967, and 1973 wars, and received its highest honor for bravery. He received his BA in psychology at the Hebrew University of Jerusalem in 1961. In 1965, he received his PhD from the University of Michigan's mathematical psychology program. He worked there with Clyde Coombs (his doctoral advisor), Ward Edwards, and David Krantz, among others. He returned to Jerusalem as a faculty member in 1967, moving to Stanford in 1978. A year at the Center for Advanced Study in the Behavioral Sciences provided Tversky and Kahneman with concentrated time to develop their approach to judgment under uncertainty, as did time spent at the Oregon Research Institute (later Decision Research).

Tversky died of melanoma, on 6 June 1996, in Stanford, California. At his death, he was Professor of Psychology at Stanford University, where he was also a Principal of its Center on Conflict and Negotiation. He held positions at Tel Aviv University, as Visiting Professor of Economics and Psychology, and as Permanent Fellow of the Sackler Institute of Advanced Studies. His wife, Barbara, is Professor Emerita of Psychology at Stanford and Professor of Psychology and Education at Teachers College, Columbia University, having studied spatial reasoning and communication. His three children are Oren, Tal, and Dona. He was known for his great energy, joy of life, and sense of humor.

Tversky made fundamental contributions to understanding human thought processes and to the mathematical foundations of the behavioral sciences. His work was distinguished by elegant formal models, tested in simple illustrative experiments. During much of his career, he collaborated with Daniel Kahneman, on studies of judgment and decision making under conditions of risk and uncertainty. He also collaborated with other scientists and scholars from psychology, economics, political science, law, statistics, philosophy, and mathematics. These projects, and the personal interactions surrounding them, helped to integrate these disciplines, despite pressures for specialization. Among his many honors were the Distinguished Scientific Contribution Award of the American Psychological Association, a MacArthur Fellowship, and honorary doctorates from Yale, Chicago, Göteborg (Sweden), and New York at Buffalo. He was a member of the US National Academy of Sciences, which served as one venue for pursuing his life-long commitment to encouraging peace and understanding. For their joint work, Kahneman received the 2002 Nobel Prize in Economic Sciences.

Axiomatic Theories of Choice

An early paper (Tversky, 1969) typifies his approach. Normative theories of choice provide rules that individuals should follow, in order to make effective decisions. The most prominent of these is utility theory, founded on a set of intuitively appealing axioms. It describes options in terms of attributes, or

features, that an individual might like or dislike. For example, when buying a car, the attributes might be price, size, styling, reliability, and comfort. According to utility theory, an individual should consider all relevant attributes, then rank the options according to their attractiveness, or *utility*. If the outcomes are uncertain, then they should be weighted by the probability of receiving them, and decision makers should choose the option with the greatest *expected utility*.

This ordering of the options should have the property of *transitivity*. A person who prefers Car A to Car B and Car B to Car C should also prefer Car A to Car C. Transitivity is an axiom or norm of *rational* decision making (hence the term *normative* analysis). When it describes decision makers' behavior, then they are better off because they are making better choices. So are scientists and policy makers. The former find it relatively easy to predict behavior, knowing that there is a stable overall ranking. The latter can create policy options (e.g., parks, housing, insurance schemes, transportation systems), knowing that the public will respond consistently. The appealing simplicity of the choice theory axioms has emboldened economists to make bold predictions about human behavior and strong prescriptions for public policies; if people are rational, then they can be expected to take full advantage of every choice offered to them.

Tversky, however, demonstrated systematic violations of transitivity, in studies using experimental tasks capturing key features of choices encountered in everyday life. Moreover, these studies did not merely show that utility theory described behavior imperfectly, making it possible to dismiss the violations as curiosities. Rather, Tversky showed that intransitivity arose from essential psychological processes, which he captured in a simple, but insightful model. Those processes reflected well-known limits to human information processing capacity, which make it impossible for people to discriminate fully among all possible options. As a result, unless they have the opportunity for trial-and-error learning, people can be, at best, *boundedly rational*, making orderly inferences about a subset of the full decision problem (Simon, 1957).

In the specific process that Tversky proposed, people simplify complex choices by initially comparing options on the most important attribute (e.g., car price), and then checking whether differences on the second most important attribute (e.g., reliability) are large enough to change their choice. Like many simplifying rules, or *heuristics*, this way of bounding a problem may be helpful in some situations, while producing problems in others. For example, imagine a customer making sequential comparisons between overlapping pairs of cars (A with B, B with C, C with D, and so on), arranged so that each new car is both cheaper and less reliable than the previous one. Looking just at the price would make each new option more attractive, while the reductions in reliability slowly mount up. At some point, though, if the first car were compared with the last car, a buyer might decide that reliability has gotten so low that the price reduction is too small to compensate for it. In that case, the first car would suddenly be preferred, despite having been initially rejected. Such intransitivity could, in principle, turn people into 'money pumps,' willing to pay something to move from one option to the next, but eventually going back to the first option and starting over again. As Tversky noted, car salespeople may try to exploit this cognitive limit by adding options (and costs) to a favored car. Unlike the experimenter, though, they hope to stop just short of the point where customers balk at the increased price and revert to the stripped-down model. Hanson and Kysar (1999), Posner (2010), and Thaler and Sunstein (2007) provide lively discussions of the implications of how to protect people from having their cognitive weaknesses exploited without denying them freedom of choice.

This early research captures a theme that recurs in Tversky's work, namely, that descriptively valid formal theories of human behavior should be assembled from psychologically plausible assumptions about what people actually do, rather than be derived from normatively postulated principles about what they should do (e.g., Tversky, 1967, 1972). Those assumptions gain plausibility if they are grounded in psychological research – which benefits, in turn, from being given formal expression and tested in decision making research. Thus, Tversky's demonstration of intransitivity builds on psychology's long tradition of studying 'context effects,' showing how sensitive choices, especially novel ones, can be to how options are presented. Choosing the relevant behavioral principles, from among the vast array identified by psychology, requires articulating the formal structure of choice situations. In some cases, Tversky did that by adapting existing mathematical approaches. Other cases prompted new mathematics, much of it appearing in three landmark volumes (Krantz et al., 1971/1989/1990). The study of alternative choice rules for multiattribute options has become a cornerstone of *conjoint measurement*, a family of procedures used widely in applied research (e.g., marketing multiattribute consumer products).

Prospect Theory

These strands of formal and empirical research came together in Kahneman and Tversky's (1979) 'prospect theory.' Published in a leading economics journal, it offered an alternative to utility theory, the core of that discipline. It depicts individuals as thinking about their choices in orderly, sensible ways. However, the rules that they follow differ from those prescribed by the axioms of utility theory, in ways that produce systematic violations. It holds that people are both sensitive to features of choices with no representation in utility theory and insensitive to ones in the theory. Prospect theory promises greater descriptive validity than utility, but at the price of abandoning economists' normative aspirations. Markets cannot be trusted to operate efficiently, if people may not make decisions in their own best interests.

Superficially, prospect theory resembles utility theory. 'Prospects' (i.e., options) are represented as sets of possible outcomes. The *value* (rather than the 'utility') of a prospect is the sum of the values assigned to each outcome, weighted by its chances of happening. However, the components of this evaluation differ from their utility theory analogs in ways that made them more realistic psychologically (Tversky and Kahneman, 1991):

1. Outcomes are evaluated relative to a *reference point*, rather than with respect to an individual's *net asset position* (or overall worth). The status quo is a natural reference point;

people focused on it ask how much better or worse off a choice will make them. However, the reference point could also be another salient value, as when people evaluate a pay raise relative to what they expected to get (or what others got), rather than relative to their current salary. This feature of prospect theory captures the psychological principle that people are sensitive to changes (Helson, 1964), unlike economic analyses, which assume that people always count their blessing, looking at any outcome in terms of all that they have in the world.

2. People care less about any given absolute difference (e.g., X dollars, Y minutes) as the overall change increases, for both gains and losses. Utility theory also assumes that people care less about a fixed gain as they have greater wealth. However, it makes the opposite assumption for losses, namely, that a given loss is increasingly painful as total losses mount up. For example, utility theory attributes insurance buying to people being so averse to large possible losses that they willingly incur small certain ones (premiums). This feature of prospect theory captures the psychological principle that it takes a larger difference to be noticed when the overall change is larger – whatever the direction of that change.
3. Before being combined with value judgments, the probabilities of the outcomes are multiplied by a *probability weighting function*. It holds that people pay particular attention to sure things, while being relatively indifferent to probability differences in the middle range (e.g., the difference between 0.35 and 0.45 matters less than that between 0.90 and 1.00). This feature of prospect theory reflects the observation that tangible outcomes mean more than intangible probabilities.

As combined in the theory, these principles (along with some others) predict (among other things) a form of context effect called *framing*, achieved by shifting the reference point. Continuing an earlier example, the same raise can seem like a gain when compared with current wages, but a loss when compared with the expectation of a larger one. In one widely used demonstration, an emergency medical intervention that could save 200 of 600 possible lives is more attractive when described in terms of the 200 lives saved than the 400 lives lost. Such changes can reverse choices, in ways that have no reflection in utility theory (which has no such reference point), hence are irrational (in its terms).

Prospect theory's challenge to utility theory evoked vigorous responses. Some suggested radical methodological skepticism, with partisans attacking any feature of the research that could, conceivably, have artifactually contributed to evidence supporting the theory. Amidst these ad hoc claims, though, were ones that raised fundamental questions, prompted by the power of prospect theory's results. One concerns the role of financial incentives in research, where economics and psychology have conflicting norms. Economists typically believe that people pay little attention to tasks without incentives, while calculating the angles with tasks that have them (e.g., lying about their beliefs if these is some advantage to doing so). In contrast, psychologists typically view all behavior as 'real,' with researchers facing the challenge of deciding (or controlling) how individuals construe their tasks. In one summary, Camerer and Hogarth (1999) find varied impacts of monetary rewards on decision making performance, partially depending on whether people have the skills needed to respond to incentives.

At the other extreme lie economists who accept the validity of prospect theory's results and use them to reflect on their science. Some have examined its implications for the axiomatic basis of utility theory, either refining those axioms to fit the data better or replacing them with different normative principles. Some have examined the practical implications of nonrational behavior, spurring new fields like behavioral finance, which questions efficient market theories, and approaches like libertarian paternalism (Thaler and Sunstein, 2007), which advocates structuring choices so that people are more likely to do what experts hold to be best for them (e.g., saving more for retirement).

Another research direction that prospect theory has propelled focuses on understanding the sort of inconsistent preferences revealed in framing effects. One general approach holds that preferences for specific prospects are often inferred (or *constructed* or *articulated*) from more general, basic values. Contextual cues, like reference points, shape that process, producing inconsistent preferences. Predicting those preferences requires knowing those cues. Informing those preferences requires bringing potentially relevant perspectives to decision makers' attention, so that they can come to know what they want (Fischhoff and Kadvany, 2011; Lichtenstein and Slovic, 2006).

Judgment under Uncertainty: Heuristics and Biases

For economists and psychologists concerned with decision making, values are a matter of personal preference. As a result, showing them to be suboptimal means demonstrating inconsistencies. With beliefs, suboptimality can be established by comparing judgments to one another (looking for inconsistency) or to an accepted external standard (looking for error). In their first published collaboration, Tversky and Kahneman (1971) adopted the latter strategy, finding that mathematically sophisticated scientists systematically misjudged the statistical power of research designs. Although these scientists knew how to calculate power, when forced to rely on intuition, they exaggerated the chances of observing anticipated results, acting as though the law of large numbers applied to small samples as well.

Tversky and Kahneman (1974) proposed that these judgments arose from relying on the *representativeness* heuristic, whereby an event appears likely to the extent that it captures (or 'represents') the salient features of the situation that might produce it. As a result, people who rely on representativeness will neglect factors without salient roles, however relevant they should be. Sample size is one natural candidate for such neglect. If a situation can produce an event, why should it matter how large a sample is observed? A second normatively relevant factor with no role in representativeness is the relative frequency of past observations. People should rely on such *base-rate* (or a priori) information unless there is strongly diagnostic *individuating* information, showing why a specific case is different from the norm. (Medical students are

sometimes told, 'When you hear hoofbeats, think horses, not zebras' – unless in Africa or at the zoo.) However, studies find that even flimsy case-specific evidence can override knowledge of what is generally true. It is natural to think about how well someone fits the image of an engineer (or how much a lump looks like a tumor). It is not so natural (or easy) to combine that vision or theory with a statistical summary of the frequency of engineers (or tumors) in the population (Gilovich et al., 2002; Kahneman et al., 1982).

Analogous reasoning leads to predicting that people who rely on representativeness will underestimate regression to the mean (because error is not represented in the process producing an observation) and exaggerate the predictability of random sequences (because their properties should be represented even in small samples). The latter was demonstrated in studies showing an illusory belief in being able to identify streaks in basketball free-throw shooting. Both collegiate and professional players (and fans) interpreted patches of random sequences as reflecting transient 'hot hands' (Gilovich and Tversky, 1985).

A second heuristic for judgment under uncertainty is *availability*, whereby an event seems probable to the extent that examples easily come to mind. Like other heuristics, it has some validity, when unfamiliar events are unlikely, and draws on natural ways of thinking (people appear to keep mental records of how often they have seen events without even realizing it). However, availability can also predictably lead people astray, as when the news media report some risks much more often than others (e.g., homicide vs asthma) and people cannot correct for the extent to which appearances are deceiving (Slovic, 1987). A third heuristic, also with historical roots in psychology (Poulton, 1994), is *anchoring and adjustment*. Users estimate uncertain quantities (e.g., temperatures, stock prices, building heights) by selecting a salient value (the 'anchor'), then adjusting from there, based on whatever other concerns come to mind. Typically, that adjustment is insufficient, leaving people 'anchored' in the initial value. If the anchor is far from the correct one, then the resulting estimate will be, too.

Just as it drew on basic psychological research, the heuristics-and-biases program also stimulated new directions in it. Tversky and Koehler (1994) subsequently formalized the relationship between evaluations of evidence and probability judgments in 'support theory,' organized around how people weigh the reasons for and against a hypothesis. Like its predecessors, support theory predicts the conditions under which simple, intuitive thought processes lead to sound judgments and poor ones. Tversky's (1977) theory of similarity judgments uses measure-theoretic concepts to predict judgments of representativeness (among things). The theory explains, for example, why people are more likely to say that a son looks just like his father, rather than vice versa (unless they know the son much better). Shafer and Tversky (1985) produced a general account for when judgmental biases raise questions regarding the validity of the normative standard against which judgments are compared.

Amos Tversky's work continues to be profoundly influential, long after his death, as scientists examine their implications and the scientific community hopes for other individuals with his rare combination of intellectual skills, insight, and insistence on 'getting it right.'

See also: Bayesian Theory, History, Applications, and Contemporary Directions; Behavioral Economics; Cognitive Psychology: History; Decision Making, Psychology of; Decision and Choice: Bounded Rationality; Decision and Choice: Economic Psychology; Decision and Choice: Heuristics.

Bibliography

Camerer, C.F., Hogarth, R.M., 1999. The effects of financial incentives in experiment. Journal of Risk and Uncertainty 19, 7–42.
Fischhoff, B., Kadvany, J., 2011. Risk: A Very Short Introduction. Oxford University Press, Oxford.
Gilovich, T., Griffin, D., Kahneman, D. (Eds.), 2002. The Psychology of Judgment: Heuristics and Biases. Cambridge University Press, New York.
Gilovich, T., Tversky, A., 1985. The hot hand in basketball. Cognitive Psychology 17, 295–314.
Hanson, J.D., Kysar, D.A., 1999. Taking behavioralism seriously: some evidence of market manipulation. Harvard Law Review 112, 1420–1572.
Helson, H., 1964. Adaptation-Level Theory. Harper & Row, New York.
Lichtenstein, S., Slovic, P. (Eds.), Construction of Preferences. Cambridge University Press, New York, pp. 653–667.
Kahneman, D., Slovic, P., Tversky, A. (Eds.), 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, New York.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–281.
Krantz, D., Luce, R.D., Suppes, P., Tversky, A., 1971/1989/1990. Foundations of Measurement, vols. 1–3. Academic Press, New York.
Posner, R.A., 2010. Catastrophe. Oxford University Press, New York.
Poulton, E.C., 1994. Behavioral Decision Theory. Cambridge University Press, Cambridge.
Shafer, G., Tversky, A., 1985. Languages and designs for probability judgment. Cognitive Science 9, 309–339.
Simon, H.A., 1957. Models of Man: Social and Rational. Wiley, New York.
Slovic, P., 1987. Perception of risk. Science 236, 280–285.
Thaler, R.A., Sunstein, C., 2007. Nudge. Penguin, New York.
Tversky, A., 1967. A general theory of polynomial conjoint measurement. Journal of Mathematical Psychology 4, 1–20.
Tversky, A., 1969. The intransitivity of preferences. Psychological Review 76.
Tversky, A., 1972. Elimination by aspects: a theory of choice. Psychological Review 79, 281–299.
Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.
Tversky, A., Kahneman, D., 1971. Belief in law of small numbers. Psychological Bulletin 76, 105–110.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131.
Tversky, A., Kahneman, D., 1991. Advances in prospect theory. Journal of Risk and Uncertainty 5, 297–323.
Tversky, A., Koehler, D.J., 1994. Support theory: a nonextensional representation of subjective probability. Psychological Review 101, 547–567.

Twin Studies and Epigenetics

Tiphaine C Martin, Jordana T Bell, and Timothy D Spector, St Thomas' Hospital, Kings College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Epigenetics is the study of chemical and cellular modifications to DNA and chromatin, which alters the expression of genes without altering the underlying DNA sequence. The state of epigenetic elements can change over time and can have a heritable component, but can also be modified by internal and external environmental factors. Therefore, epigenetic phenotype analyses should take into account the dynamic nature of epigenetic mechanisms, and control for potential confounders, such as age and genetic factors. Twin studies provide valuable information about the genetic as well as the epigenetic epidemiology of complex traits and diseases. Such unique studies can explore the relationship between genetic, environmental, and random factors and their interactions with, for example, sex, age, and lifestyle factors underlying a phenotype of interest. Using twins provides a powerful model to help us unravel the multifactorial components of complex diseases and traits.

Epigenetic Mechanisms

The term 'Epigenetics' combines the word 'epi' (Greek: επί – over, above, outer) and 'genetics,' and was first proposed by Waddington (1942) to describe the epigenetic landscape of gene regulation during development. More recently, epigenetics has been defined as the study of chemical and cellular events, such as the modulation of gene expression or cellular phenotypes, caused by mechanisms other than changes produced by the underlying genetic architecture (Holliday, 1994). External and internal environments can stimulate these changes, and the changes could be either persistent or transient. Both persistent and transient epigenetic changes can have lasting physiological impact.

Various studies (cf **Table 1**) have mapped epigenetic elements and annotated the human genome in a way that facilitates identification of potential functional gene sequences. Such general studies underline how epigenetic mechanisms in mammals play an important role in gene expression, regulation, and development and therefore in disease (Murr, 2010). One of the main findings is that epigenetic variants across different epigenetic mechanisms tend to co-occur in specific genomic regions. These epigenetic elements could be either so-called *cis*-regulatory elements, which work through intramolecular interactions (the epigenetic element is in the same chromosomal region as the regulated gene), *trans*-regulatory elements, which work through intermolecular interaction. **Table 2** and **Figure 1** present a general description of epigenetic elements.

DNA methylation (Guibert and Weber, 2013) in mammals is the addition of a methyl group to the cytosine base of the DNA structure, and is the most abundant, stable, easiest to measure, and well-studied epigenetic mechanism in humans. Consequently, the following sections focus mainly on the current knowledge about this epigenetic marker. DNA methylation plays a role in gene regulation, chromosomal stability and X-chromosome inactivation, and parental imprinting. Genetic imprinting is a DNA methylation process by which alleles can be expressed in a parent-of-origin-specific manner, while X-chromosome inactivation (also called lyonization) is a random phenomenon in each cell that inactivates one of the two copies of the X-chromosome present in female mammals. The two latter processes are the first DNA methylation phenomena studied (cf Table 4 (Supplementary Data)). In mammals, DNA methylation occurs on the 5-carbon position of cytosine predominantly in the context of CpGs dinucleotides. CpG-rich regions, called 'CpG islands' (CGIs) which are

Table 1 Large-scale national and international epigenomic consortia[a,b]

Project name	*Start date*	*Affiliations*	*Access data*
Human Epigenome Project (HEP)	1999	The Wellcome Trust Sanger Institute Epigenomics AG The Centre National de Génotypage	http://www.epigenome.org
Encyclopedia of DNA Elements (ENCODE)	2003	NIH EBI The Wellcome Trust Sanger Institute	http://encodeproject.org/ENCODE
The Cancer Genome Atlas (TCGA)	2006	NIH	http://cancergenome.nih.gov
Roadmap Epigenomics Project	2008	NIH	http://www.epigenomebrowser.org/
International Cancer Genome Consortium (ICGC)	2008	15 countries, includes TCGA	http://dcc.icgc.org/web
International Human Epigenome Consortium (IHEC)	2010	7 countries, includes BLUEPRINT, roadmap	http://ihec-epigenomes.org/

[a]Beck et al. (1999).
[b]Rivera and Ren (2013).

Table 2 Epigenomic maps. The table describes the different epigenetic mechanisms occurring at different levels in the cell: at the DNA sequence level – such as DNA methylation, at the level of conformation of DNA sequence – such as chromatin structure or attachment to the nuclear matrix, at the proteomic level – such as histone modification, or at the level of interaction with other epigenetic factors, such as microRNA and transcription-factor binding

Epigenetic modification	*Epigenetic markers*	*Actions*	*Major remarks*	*References*
At the level of genetic sequence				
DNA methylation	CpGs dinucleotides non-CpG sites	Gene regulation Chromosomal stability X-chromosome inactivation Parental imprinting Reprogramming (cell lineage and tissue differentiation)	DNA methyltransferases (DNMTs) catalyse the transfer of methyl groups to cytosine, predominantly at CpG dinucleotides CpG islands are generally hypomethyled and non-CpG site methylation has been reported mainly in ESCs oocytes and adult brain	Guibert and Weber (2013)
DNA hydroxymethylation	CpGs dinucleotides	Not completely defined: may serve as an intermediate in the reaction of DNA demethylation or act as a signal for chromatin modifications	Catalyzed by TET enzymes and abundant in certain cell types	Guibert and Weber (2013)
At the level of conformation of genetic sequence				
Chromatin structure	Nucleosome positioning	Regulation of accessibility of the genetic information (transcription, replication, repair)	Euchromatin (open chromatin) is strongly correlated with gene-dense regions and actively transcribed genes, unlike heterochromatin (closed chromatin) regions	Ozsolak et al. (2007) Fadloun et al. (2013)
Chromatin domain interactions	Interaction between chromatin domains	Regulation of accessibility of the genetic information	Higher order chromatin interactions genome-wide	Fadloun et al. (2013)
At the level of interaction with other epigenetic markers				
Noncoding RNA	rRNA tRNA microRNA Transposable elements Others	Translation RNA splicing Gene regulation genome defense Chromosome structure	Highly abundant and not translated into protein Several groups of ncRNA involved in many cellular processes and diseases such as autoimmune diseases or cancer and senescence	Esteller (2011) Ostankovitch and Pyle (2013)
Transcription factor binding	Transcription factors	Regulation of gene expression, cell identity, and cell fate	Depends on chemical signals such as hormones and neurotransmitters which can be influenced by internal and external environmental factors	Sindhu et al. (2012)
At the proteomic level				
Histone modification	Over 130 different histone modifications identified	Regulation of accessibility of the genetic information (transcribed/repressed genes, regulatory elements)	Combinations of chromatin marks define a limited number of chromatin states	Zentner and Henikoff (2013)
Protein modification	Prions	Misfolded protein state acting like an infectious agent	Pathogenic variant cause Creutzfeldt-Jakob disease (CJD), Normal prions protect against apoptose and act the neuronal development Mechanism of inheritance or conformation change between prions variants not based on change in DNA, reversible in function to environmental stress	Halfmann and Lindquist (2010)

Guibert and Weber, 2013; Ozsolak et al., 2007; Fadloun et al., 2013; Esteller, 2011; Ostankovitch and Pyle, 2013; Sindhu et al., 2012; Zentner and Henikoff, 2013; Halfmann and Lindquist, 2010.

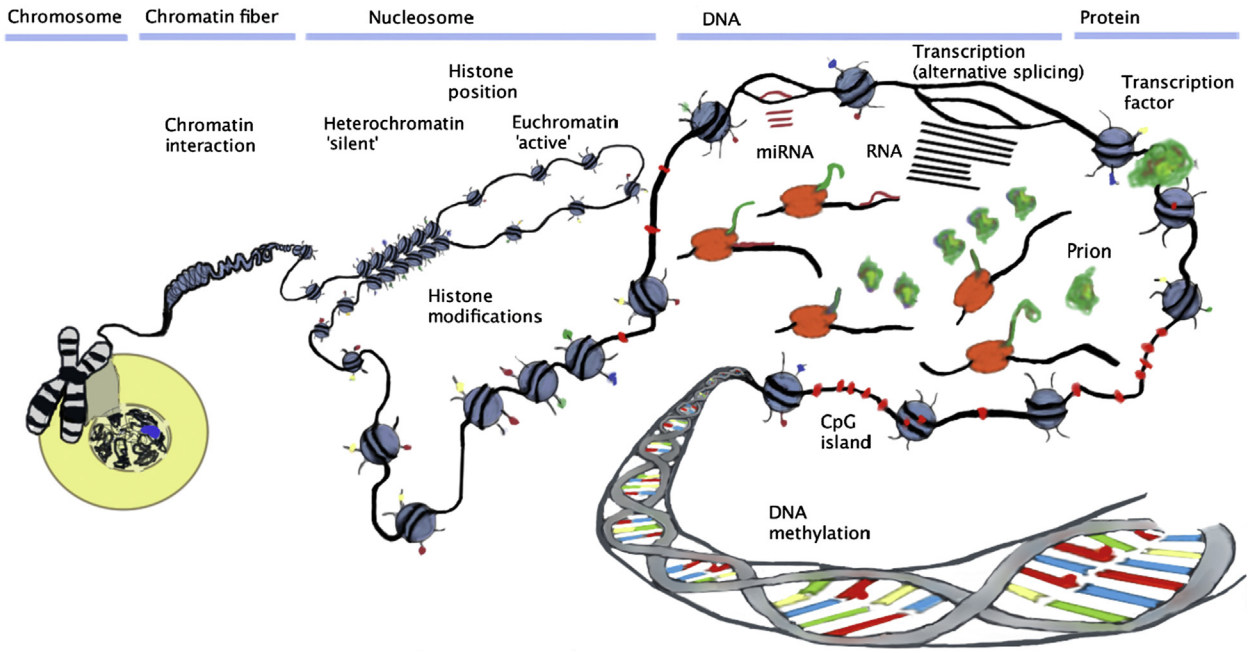

Figure 1 Epigenome maps: the figure represents the different epigenetic mechanisms occurring at different levels in the cell and working together (left to right): at the level of conformation of DNA sequence – such as chromatin structure or attachment to the nuclear matrix; at the chromatin level – such as histone modification; at the DNA sequence level – such as Gene expression; at the level of interaction with other epigenetic factors – such as microRNA and transcription-factor binding; or at the DNA sequence level – such as DNA methylation.

predominantly hypomethylated. CpG islands tend to colocalize with gene promoters, gene-bodies, and regulatory regions. *De novo* methylation in a promoter region usually switches the gene from active to inactive (Klose and Bird, 2006). At other CpGs sites outside promoters, the role played by methylations is less clear as is methylation in non-CpGs sites.

Twins

Twins represent a particular type of full-sibling. They can share part or all of their genomes, their intra-utero environments, and early lifetime events and experiences.

There are different types of twins depending on the degree of genomic identity, called zygosity. The two types of twins are: monozygotic ('identical') (MZ), who share almost 100% of their genome and share most of their *in utero* and early life environment, and dizygotic ('fraternal') (DZ), who share most of their *in utero* and early life environment but only around 50% of the segregation of their genome, and who can be either same-sex or opposite sex pairs. DZ twins come from two different eggs independently fertilized by two sperms at the same time, while MZ twins share the same egg (the same proportion of maternal and paternal genomes). Male–female dizygotic twins are the most common type of dizygotic twins, whereas it is slightly more common to have female–female monozygotic twins than male–male monozygotic twins, due to mortality rates.

Nature generates more diversity in twins than just zygosity (cf **Figure 2(a)**). For instance, the degree and delay of separation between MZ twins *in utero* defines different types of identical twins; or the adoption of twins in different families plays a role on the proportion of shared environment and individual environment. Moreover, random genetic mutations may occur at any time during the development of the embryo and continue to arise during lifetime, increasing the genetic difference between individuals with a frequency of roughly 0.1–100 per genome per sexual generation but currently indistinguishable from 1/300 per cell division per effective genome (Drake et al., 1998).

Monozygotic twins may not be exactly genetically identical (100%) and can have some minor mutational differences (single nucleotide polymorphism – SNP – and small insertion/deletion region – InDels) (Li et al., 2014; Weber-Lehmann et al., 2014). Recently, Li et al. (2014) estimated that somatic mutation (mutation happens in any cell other than a gamete, germ cell, gametocyte, or undifferentiated stem cell) in monozygotic twin pairs has a frequency of 1.2×10^{-7} per base pair. MZ twins may also differ in copy number variation but the extent of these differences remains unclear (Bruder et al., 2008). Mutations that distinguish MZ twins must occur after the morula splits (about four to five days after fertilization), or just beforehand in a confined area of embryo. These mutations can be present in germ cells but not in other tissues and vice versa or in both germline and somatic cells. Recently, researchers (Weber-Lehmann et al., 2014) showed that they could distinguish identical twins based on single nucleotide differences in their genomes in a forensic case that involved monozygotic twins.

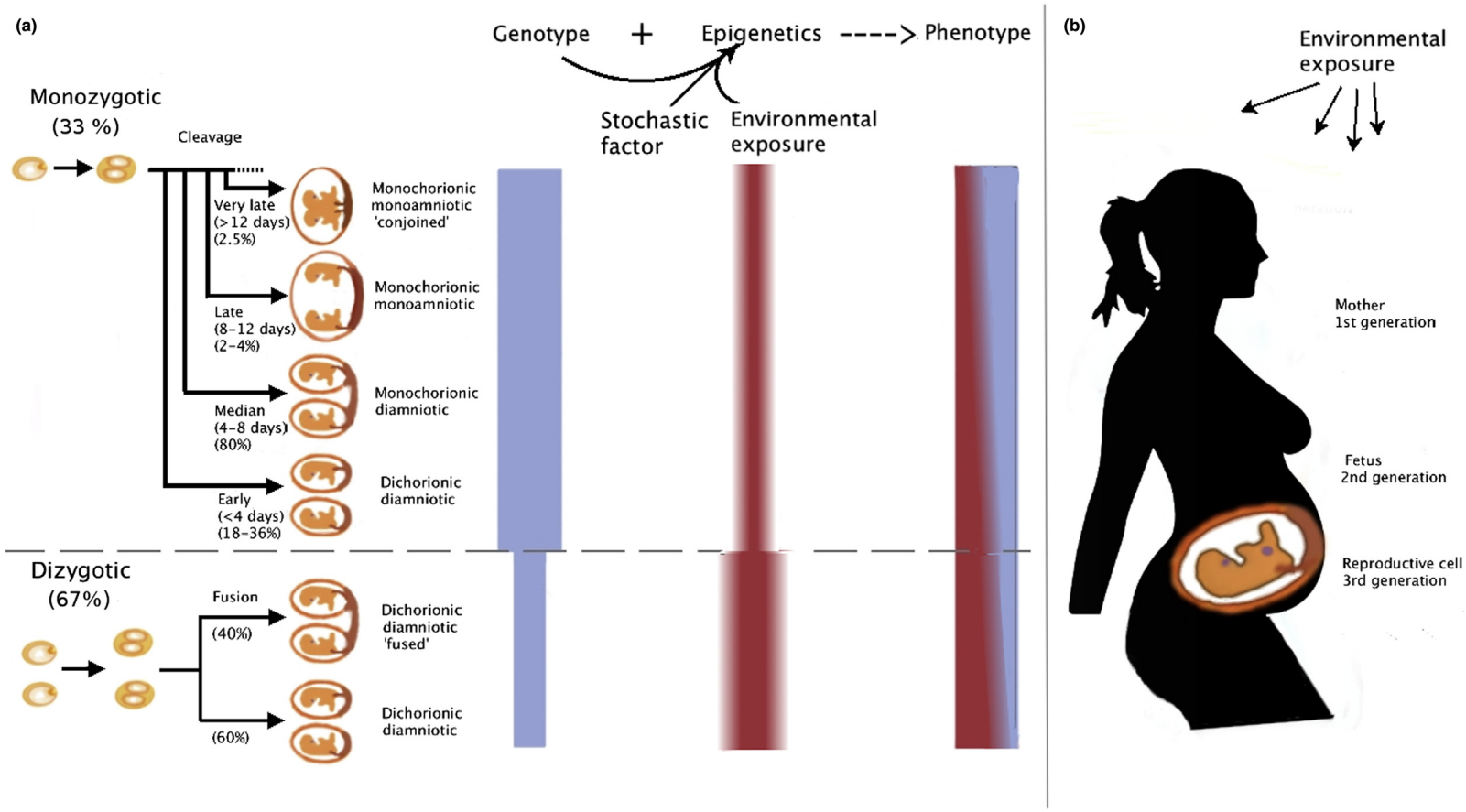

Figure 2 Diversity of twins caused by genetic and environmental factors. (a) Different types of twins and their phenotypic difference in early life are determined by the timing of the egg split. The latter plays on the number of amniotic and chorionic (placenta) sacs, consequently on the intrauterine shared environment. Almost one out of three twins is monozygotic twin, who is partitioned into four different groups. The timing of the split and the percentage of each group were obtained from Hogenson, T.L., 2013. Epigenetics as the underlying mechanism for monozygotic twin discordance. Medical Epigenetics 1. The first rectangle on the right hand side gives the percentage of shared genomic segregation, namely almost 100% of genomic segregation for MZ twins and almost 50% for DZ twins. The next double rectangles with blurred edge indicate the raising influence of environmental factors during embryonic development (from top to bottom). There may be a slight decrease influence of environmental factors between dichorionic-diamniotic MZ twins and dichorionic-diamniotic 'fused' DZ twins, because there is fusion between the blastocysts of DZ twin pairs in the early stage. Furthermore, the proportion of environmental factors between monochorionic MZ twins and dichorionic MZ twins, may be influenced by twin-to-twin transfusion syndrome (TTTS) or other phenomena (Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H., Wong, A.H., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., McRae, A.F., Visscher, P.M., Montgomery, G.W., Gottesman, I.I., Martin, N.G., Petronis, A., 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics 41(2) 240–245; Gordon, L., Joo, J.E., Powell, J.E., Ollikainen, M., Novakovic, B., Li, X., Andronikos, R., Cruickshank, M.N., Conneely, K.N., Smith, A.K., Alisch, R.S., Morley, R., Visscher, P.M., Craig, J.M., Saffery, R., 2012. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Research 22 (8) 1395–1406). The combination and the interaction of genetic effects and environmental effects define the phenotypic variation between twins (described at the top of the figure). The last rectangle shows a potential distribution between genetic effect (blue part) and environmental and stochastic effect (red part) on phenotype at time T and cell type given just after birth. This proportion of different variants (genetic, environment, stochastic) playing on phenotype changes over time and site. (b) The draw shows how external environmental factors affecting the mother plays, also and many more, a role on the epigenetic modifications of the fetus and its given reproductive cells during the development stage. The panel shows how external environmental factors play a role on the epigenetic modifications of the mother, but also potentially on the epigenetic modifications of the fetus and its reproductive cells during the development stage.

Epigenetic States Change over Time and Site

From a general perspective, a multicellular eukaryotic organism, including human organisms, starts from one unique cell, which splits into multiple differentiated cells. A set of epigenetic markers, imprinted genes, and in females' X-chromosome inactivation, is found in all cells of a given organism. These are typically the outcome of epigenetic markers defined in embryogenesis. Only around 100 imprinted genes out of 23 000 are currently known to exist, and this may be because of a reprogramming step that erases most epigenetic signals from parental gametes and also from the zygote. This step ensures the totipotency of early embryonic cells (the ability of a single cell to divide and produce all of the differentiated cells in an organism). Although all cells have almost the same genetic information, they can differ significantly due to epigenetic mechanisms. Moreover, contrary to genetic sequence, epigenetic elements change over time at much higher frequency. Indeed, epigenetic modification plays an important role in regulating gene expression, developmental stages and in tissue differentiation (Colaneri et al., 2013), and therefore in cellular and metabolic memory.

Locus-Specific Effects

Epigenetic variation occurs and develops at different rates across different genomic regions, called also loci. These events depend on three factors, which can act independently or in combination: stochastic drift of epigenetic states, genetic variation, and environmental factors. Martino et al. (2013) observed that in singletons as well as twins, 30% of their assayed CpGs undergo dynamic methylation during the first 18 months of life and some of these methylation events are independent of constrained developmental programs. Moreover, the pattern and the variation of DNA methylation seem to differ according to the annotated function of genetic regions (CpG islands, CpG islands shores, and coding sequence) (Davies et al., 2012). An increase in the variation of DNA methylation over time is observed in annotated genomic regions. By contrast, a significant decrease of the variation of DNA methylation is shown in intragenic regions, enhancers, and low-density CpG promoters (Martino et al., 2013). Epigenetic mechanisms can also give rise to comethylation of several contiguous CpGs in the same direction, and in some cases striking interindividual variability, forming variably methylated regions (VMRs) (Feinberg and Irizarry, 2010).

Cell-Specific and Tissue-Specific Effects

Epigenetics play a key role in the cell cycle, as well as, in cellular and tissue-level differentiation (Khavari et al., 2010). In order to ensure the progression of the cell cycle, epigenetic mechanisms vary across cell cycle phases in governing genome accessibility, and consequently, the expression of individual genes. These variations are shortlived and are not inherited. However, during cellular division, the epigenetic patterns of cell and tissue type are transmitted to daughter cells to maintain tissue integrity. During the mitotic phase and after cellular differentiation, cells reprogram themselves to produce new required functional cell-type.

Data from several sources have identified differences in epigenetic modifications between cells, cell types, and tissues (Davies et al., 2012; Wu et al., 2011; Zhang et al., 2013). Genome-wide DNA methylation patterns are highly tissue-specific, between the tissues of brain and blood for instance, or between the tissues of functionally distinct areas of brain (Davies et al., 2012). Additionally, Wu et al. (2011) observed that DNA methylation differs between different blood cell types. Interestingly, a greater genome-wide epigenetic variation was observed between different tissues of the same individual than between the same cell type of different, unrelated individuals (Eckhardt et al., 2006). This demonstrates clearly a significant tissue-specific effect (Eckhardt et al., 2006). In the epigenetic study of phenotypes, the epigenetic state associated with the complex trait may be a tissue-shared or tissue-specific effect, and therefore the selection of studied tissue or cell type may have an impact on the findings. Thus, the tissue or cell type selected for epigenetic phenotype analysis should be the most appropriate to the complex trait or studied disease.

Unfortunately the most appropriate cell type or tissue for omic studies (for example, epigenomic or transcriptomic studies) of specific complex traits or diseases is not always easily accessible, abundant, or even known. There are clear DNA methylation differences across tissues at cell-type and tissue-specific sites. However, a number of recent studies (Davies et al., 2012; Shenker et al., 2013) suggest that whole blood could be a useful surrogate for less readily available tissues (e.g., brain or lung, in pre-mortem individuals) for molecular biomarker studies for epigenetic modifications. The same cannot be said for gene expression (Witt et al., 2013). Studies comparing the methylation alterations of brain and blood concluded that whole blood might also be a proxy for brain tissue (Davies et al., 2012; Shenker et al., 2013). Within an individual, Davies et al. (2012) have recently shown that a significant correlation ($R^2 > 0.66$) was observed between DNA methylations in whole blood and in tissue derived from brain. The epigenetic patterns observed in blood may therefore be a relevant proxy, depending on the particular genomic region for defining biomarkers of environmental factors that are difficult to quantify in epidemiological studies.

Sex and Gender Effects

DNA methylations and histone modifications differ between sperm and oocyte (Duffié and Bourc'his, 2013). The differences are found in imprinted genes. The majority of imprinted genes come from maternal imprinting control regions (ICRs), which correlate with CpG-rich promoters, while paternal ICR occur more frequently in CpG-poor, intergenetic regions. Apart from X-chromosomal inactivation, which occurs in cells from female individuals after fertilization, the effect of sex and gender on DNA methylation is controversial. A gender-specific effect was detected for some genes in autosomal chromosomes in twins and healthy controls (Boks et al., 2009) in whole blood, whereas some disease studies demonstrated no clear 'gender effect' (Bell et al., 2011) in lymphoblastoid cell lines. Consequently, until we know more, when studying complex traits and diseases affected by gender, it is preferable to use unrelated individuals or twin pairs of the same gender.

Aging Effects

From epigenetic studies of twins and unrelated individuals (for example, Martino et al., 2013; Bell et al., 2012; Rakyan et al., 2010; Teschendorff et al., 2010; Fraga et al., 2005), it was found that aging strongly influences epigenetic markers. A significant discordance is observed in MZ twins in their early life and this discordance and variation in methylation increases with age. Fraga et al. (2005) observed an increase in epigenetic differences between MZ twins with age. In selected CpGs sites they observed a higher DNA hypomethylation, and a higher histone H3 and H4 hyperacetylation (usually associated with transcriptional activation) in the older sibling. Martino et al. (2013) reported that epigenetic differences already exist in MZ twins at birth and increase over their lifetimes in response to stochastic and environmental factors. Moreover, certain DNA methylation sites have been identified as biomarkers of aging (for example, Martino et al., 2013; Bell et al., 2012; Rakyan et al., 2010; Teschendorff et al., 2010; Fraga et al., 2005); and as expected very few of these are related to genetic variations using methylation quantitative trait loci (meQTL) (Bell et al., 2012, Tsai et al., 2012). However, it is possible that the aging methylation effects that have been reported in whole blood are an artifact of increasing blood cell type variability with age (Jaffe and Irizarry, 2014), although several of the aging methylation studies either controlled for whole blood cell type variability (Bell et al., 2012) or the studied specific blood cell subtypes (Rakyan et al., 2010). High levels of confounding between age-related variability and cellular composition as detected by DNA methylation as specific CpG sites were detected in some of the data of previous papers reporting age-related changes of DNA methylation profiles (Jaffe and Irizarry, 2014). Thus, future epigenetic studies of aging should take into account cell composition. Horvath (2013) found that 353 CpG sites together act as an 'ageing clock' in terms of chromatin states, cell states, and tissue variance, and can accurately predict the age of an individual, independent of many tissue types. In addition to DNA methylation markers of aging, other epigenetic marks can also associate with age, such as microRNAs. Multiple miRNAs have been reported to be deregulated during cellular senescence, aging, and disease (Wolfson et al., 2008).

Epigenetic: Genetic, Environmental, and Stochastic Factors

Epigenetic elements differ among related and unrelated individuals and among species; they also change over time, depending on locus. We examine the epigenetic markers that are heritable and the mechanisms that maintain this inheritance.

The Heritability of Epigenetic Modifications

In general, the heritability of a phenotypic trait is defined as the proportion of the total phenotypic variation of this trait that is explained by genetic variation alone. The phenotypic variation of a trait (P) can be partitioned into contributions from latent genetic (G) and environmental (E) factors, and gene–environment interactions (G×E). To a lesser extent, there is the stochastic influence inside each cell (S), such that $V_P = V_G + V_E + V_S + V_{GxE}$. The genetic variation can be due to additive genetic variance (V_A) as well as nonadditive genetic variance that can be related to interactions between alleles at the same locus (dominance, V_D) and/or at different loci (epistasis, V_I): $V_G = V_A + V_D + V_I$. The environmental influence can be split into the environment shared by the study group, for example shared by family members ('the common environment,' V_C), the environment specific to each individual, but common to repeated measures ('the unique environment,' V_U), and the environmental variance that is not explained by other parameters (for example, error variance inside repeat measures, $V\varepsilon$): $V_E = V_C + V_U + V\varepsilon$. Thus, phenotypic variation can be expressed as $V_P = V_A + V_D + V_I + V_S + V_C + V_U + V_{G\times E} + V\varepsilon$. Two estimators of heritability are commonly defined. First, broad-sense heritability (H^2) takes into account total genetic variation (V_G); thus $H^2 = V_G/V_P$. Second, narrow-sense heritability (h^2) takes into account only the additive genetic variation (V_A); thus $h^2 = V_A/V_P$.

Inheritance of epigenetic states may be studied during both mitosis and meiosis. The latter could potentially lead to trans-generational epigenetic inheritance, which is currently a highly controversial topic, due to confusion of the term with inter-generational exposure (Daxinger and Whitelaw, 2010). For example, if environmental exposures occur during pregnancy or during spermatogenesis (Schmidt, 2013), up to three generations (two for paternal environmental exposures) can be affected by the same environmental factors without being dependent on meiotic inheritance mechanisms (cf **Figure 2(b)**). It is therefore important to distinguish between experiments that explicitly test for epigenetic meiotic heritability, and cases of multigenerational exposure. Only epigenetic markers transmitted across four or more generations that exclude other potential confounders such as maternal effects, may be reported as heritable. Rare reports of trans-generational epigenetic inheritance exist (Stern et al., 2012), but the precise mechanisms underlying these effects remain unknown and need to be reconciled with extensive epigenetic reprogramming during embryogenesis.

Heritability of epigenetic states during cellular division is studied where the epigenetic state at each position is the phenotype. Heijmans et al. (2007) performed one of the first studies of epigenetic heritability using twins, in the *IGF2/H19* locus with 196 adolescent and 176 middle-aged twins. They estimated heritability to be 20–70% for the *H19* DMR twins and 57–97% for the *IGF2* DMR twins; and showed that a proportion of the heritabilities is explained by SNPs in cis (the single nucleotide polymorphism is in the same chromosomal region as the epigenetic element studied). They concluded that the IGF2/H19 locus is predominantly driven by heritable factors. Kaminsky et al. (2009) explored three different tissue types: white blood cells (WBC), buccal epithelial cells, and gut biopsies of 114 MZ and WBC and buccal epithelial cells of 80 DZ, using a microarray assaying 12K CpG promoter regions. They observed that intraclass correlation coefficient (ICC) profiles were more similar across tissues of the same individual compared to unrelated individuals and that both CGIs and promoters were less epigenetically variable in WBC and gut cells. The heritability of DNA methylation in

buccal epithelial cells was slightly higher than that in WBC, (mean $ICC_{MZ} - ICC_{DZ} = 0.15 \pm 0.0039$, $p = 1.2 \times 10^{-294}$ compared to mean $ICC_{MZ} - ICC_{DZ} = 0.0073 \pm 0.0034$, $p = 0.044$). Surprisingly, Kaminsky et al. (2009) and recently Gordon et al. (2012) observed that DNA methylation profiles in different tissues (buccal cells for Kaminsky et al. (2009) and cord blood mononuclear cells and human umbilical vascular endothelial cells for Gordon et al. (2012)) are more discordant within monochorionic MZ twins than within dichorionic MZ twins. The explanation of this observation remains unclear and this outcome needs to be validated in a larger sample size. Recently, Grundberg et al. (2013) performed a much larger study in adipose tissue from 648 twins with the Illumina 450K platform (a microarray platform that measures methylation levels at 485 000 CpG sites genome-wide) located predominantly in gene-based regions, but also in some intergenic regions – estimating heritability, h^2, of methylation sites genome-wide at 0.19. But this estimation increased to 0.34 in selecting only the top 10% of variable sites (TVS), enriched in gene bodies and intergenic regions and depleted of promoter regions and CGIs. In concordance with TVS, highly heritable sites ($h^2 > 0.5$) were enriched in hypermethylation, enriched for gene bodies and intergenic regions, and depleted of promoter regions and CGIs, which tended to be hypomethylated.

As discussed previously, aging effects act on epigenetic modifications (Bell et al., 2012; Tsai et al., 2012) or cell composition (Jaffe and Irizarry, 2014); heritabilities can also change over time and site. Based on an original idea by Slatkin (2009), epigenetic events could partly explain the missing causality of complex diseases or traits, but not the missing heritability of complex diseases. Currently, many questions about the extent and mechanisms of such inheritance remain unanswered.

Genetic Effects

Many studies on the heritability of DNA methylation, such as Heijmans et al. (2007), Kaminsky et al. (2009), and Grundberg et al. (2013), explored whether the variation of DNA methylation is related directly to underlying genetic variation. Heijmans et al. (2007) showed that a genetic mutation directly at a CpG site itself changes the interindividual normal distribution of methylation variation to that of a binominal distribution. Using quantitative trait-mapping methods, CpG meQTLs (Bell et al., 2011) were identified in lymphoblastoid cell lines from 77 HapMap Yoruba individuals. The majority of meQTLs were observed in cis and SNPs that associated with gene expression levels overlapped meQTLs, providing evidence for shared mechanisms in a set of genes. Recently, Grundberg et al. (2013) replicated this finding in adipose tissue on larger set of data. Additionally, they detected that 28.5% of CpG sites are associated with a common sequence variant among filtered probes (probes having no known SNPs, InDels, structural variations (SVs), or CNVs both within the probe and at the CpG site). These CpG sites associated with at least one common sequence variant have higher h^2 (0.44); and, on average in TVSs, 15% of common variants explained more than 50% of the methylation heritability per site. Studying such epigenetic modifications under the control of genetic variations may highlight a genetic sequence, environment correlation, or interaction. Consequently, certain epigenetic states may be inherited at the same time as genetic variations.

Fraser et al. (2012) identified and compared meQTLs between the HapMap YRI and CEU samples. They found that 30% of the studied CpG sites differed between populations, but with a small magnitude in population-level methylation differences (approximately 4% of CpG sites having only in average over 10% of population-level methylation difference). Additionally, a wide range of within-population variability in the methylation of individual CpG sites was observed. Population stratification investigating the effect of a genetic factor on disease in epidemiologic studies (Wacholder et al., 2000) could be a manifestation of confounding in epigenetic epidemiologic studies. However, one-third of population-specific DNA methylation variations were not explained directly by genetic variations defined by meQTL methodology (Heyn et al., 2013a). Thus, the methylation divergence between and within populations may be due to a combination of differences in allele frequencies and complex epistasis or gene × environment interactions. Heyn et al. (2013a) concluded that epigenetic mechanisms might help to tune the human genome in response to environmental factors and to explain natural human phenotypic variation.

Liu et al. (2014) defined GeMes as noncontiguous sets of correlated CpGs under the genetic control of one or more methylation quantitative trait loci. They observed that CpGs belonging to same differentially methylated regions (DMRs) cluster together in their variation and are associated to environmental factors such as smoking. Simultaneously, some of these noncontiguous methylation clusters are correlated with each other and can be genetically driven. This pattern is similar to that of linkage disequilibrium (LD) correlation in genetic SNP variation, but for over a much shorter distance, less than 500 bp compared to roughly 3 kb. The mechanisms by which epigenetic variations and genetic variations are related to population variations remain unclear.

Environmental Effects

Epigenetic states can be influenced by the environment, either short or long-lived; and some are transmitted to daughter cells (Cortessis et al., 2012). Environmental factors can be either exogenous, such as diet, medication, lifestyle, and behaviors or endogenous, such as inflammation and hormonal change. Interestingly, Grundberg et al. (2013) observed that 8638 CpG sites of adipose tissue influenced by 30% shared environmental factors are significantly associated with metabolic functions. Overall shared environmental factors contribute to 0.2% of methylation variation of all sites and 2% of TVS, while non-shared environmental and stochastic factors account for 80.8% and 44% respectively (Grundberg et al., 2013).

Tobacco smoke provides the best reproducible example of environmental epigenetics. The first study, using the sparse Illumina 27K platform (Breitling et al., 2011) – a microarray platform that measures methylation levels at about 27 000 CpG sites, which are located mainly in the promoters of RefSeq genes involved in cancer – found a relationship between tobacco smoke and altered DNA methylation at a single locus, the coagulation factor II (thrombin) receptor-like 3 (F2RL3) gene.

Wan et al. (2012) validated the reported loci associated with cigarette smoking and identified a novel locus, G-protein-coupled receptor 15 (GPR15). Using the Illumina 450k array, two studies, one in peripheral blood and lung tissue (Shenker et al., 2013) and the other in whole blood (Zeilinger et al., 2013) replicated the findings and defined a new set of biomarkers and concluded that upon cessation of smoking the DNA methylation state generally reverts back to an intermediate level compared to nonsmokers. Looking at fetal exposure of maternal smoking during pregnancy (PTS), a reduction in DNA methylation of repetitive elements like satellite repeats (e.g., Sat2) or an intra-genetic region of the aryl hydrocarbon receptor repressor gene (AHRR) persisted into adulthood (Joubert et al., 2014; Flom et al., 2011). Similarly, studies on epigenetic changes in rats receiving cigarette smoke concluded that this decrease of DNA methylation could be found also in F2 offspring of F1 offspring exposed during pregnancy.

These studies highlight four main types of relationship between epigenetic and environmental factors: (1) even after stopping exposure of an environmental factor, the persistence of the epigenetic modification can be detected in adulthood, (2) epigenetic states of offspring can be impacted by parental behaviors, and (3) modification or reversal of epigenetic states can take place during adulthood; (4) effects can be seen across tissues.

Stochastic Effects

The variability of epigenetic modifications that is not attributed to environment, measurement error, or genetic factors may be attributed to stochastic ('random') effects. For example, in early life MZ twins already have some differences at the level of DNA methylation, which can potentially be explained by differences their intrauterine environment (Martino et al., 2013). However, when Ushijima et al. (2003) worked on the transmission of DNA methylation between cells during mitosis, they showed that the fidelity in this transmission is not perfect and that there is a larger error rate for epigenetic states than for DNA sequences. The phenomenon of stochastic epigenetic instability, or stochastic drift, can accumulate over time and across cells. Consequently, a phenotype is generated by a combination of stochastic effects, environmental factors, and genetic impacts. The majority of studies found no distinction between stochastic effect and nonshared environment (confounding), and twin studies cannot easily distinguish these two. The proportion of stochastic effects in a complex trait or disease is problematic to estimate. Feinberg and Irizarry (2010) proposed a new non-Lamarckian theory for the role of epigenetics in evolution suggesting that even if a genetic variant does not change the mean phenotype, it might play a role in phenotypic variability mediated by stochastic epigenetic variations.

Twin Studies and Epigenetics

As discussed above, phenotypic variation at the individual/organism and cell level stems from a combination of genetic variation and environmental influence. Compared to other types, twin study designs are a powerful epidemiological tool (Boomsma et al., 2002) that can be used to disentangle and quantify the contributions of the genes; shared and/or individual environments; and the interactions between genetic and environmental factors (Van Dongen et al., 2012; Bell et al., 2012). Indeed, study designs that use unrelated subjects are not able to detect the relative proportions of environmental factors and their interactions with the genome in specific phenotypes, due to the large discrepancies between unrelated subjects at a genomic, age, and environmental level. Utilizing family-based designs or adoptive study designs reduces the impact of genetic variation and theoretically allows us to better define the role of environmental factors. An advantage of twin studies is that several twin registers exist worldwide and the data from these studies are freely available. Twin registers offer a wider diversity of twins thereby facilitating a diversity of studies such as environmental effect, genotype-by-environment-interaction, and the contribution of specific polymorphisms to the total genetic variance in specific complex traits and diseases.

The Classical Twin Study

The classical twin study has been performed since the 1920s and compares phenotypic resemblances between monozygotic and dizygotic twins for one trait or disease. This study design is able to separate genetic and environmental effects (as well as shared and unique environment) and, so, to define heritability of a trait or disease (Boomsma et al., 2002; Rahmioğlu and Ahmadi, 2010). The correlation of a common phenotypic trait between twins (denoted as r_{MZ} for MZ twins or r_{DZ} for DZ twins) depends on their shared genetic and shared environmental variation. This correlation can change with time. If the phenotype under study has a high heritability, then there is a larger concordance in MZ twins compared to DZ twins. Discordance in MZ twins ($d_{MZ} = 1 - r_{MZ}$) for a common phenotypic trait is the noncorrelated phenotypic variability and this depends upon variation at both the genetic and environmental levels. Also the degree of such discordance can change over time.

The estimation of heritability is essential in identifying the major factors influencing complex traits and diseases. More concretely, the term heritability of a phenotypic trait is used to describe the proportion of the total phenotypic variation of this trait that is explained by genetic variation in a given population. The phenotypic variation in a trait (P) can be partitioned into the contributions from latent genetic (G) and environmental (E) factors, and gene–environment interaction (G×E). To a lesser extent, there is a stochastic influence inside each cell (S). This latter is considered as an unique internal environmental factor, such that $V_P = V_G + V_E + V_{G\times E}$. Estimating heritability helps to define the discordance between twins and therefore the role of environment. The latter plays an important role in epigenetic modifications. The exclusion of known and fixed environmental factors such as population, sex, age, cohort, and tissue are usually done in order to study only unobserved genetic and environmental factors and remove some confounders. Theses exclusions assume that these environmental factors do not interact with the genetic variations and other environmental factors. Consequently, the interpretation of heritability should be done cautiously to avoid misconceptions (Visscher et al., 2008).

Three main twin methodologies are used to quantify heritability. Whatever the method used, the computation of heritability is easier and importantly more accurate than those performed on unrelated subjects. A first method is based on MZ twins separated in infancy and reared apart (MZA twins). Assuming that V_E is negligible for MZA twins because of their early separation, interclass correlation between MZA twins (r_{MZA}) becomes an estimate of broad-sense heritability (H^2). In many cases, however, V_E is larger than zero and so, r_{MZA} data tend to yield overestimates of heritability. Despite the power of the MZA design for estimating heritability, the rarity of twins reared apart nowadays generates small samples that are difficult or impossible to replicate.

Two other methods are based on classical twin studies with two further assumptions: first, that MZ twins have identical genomes and DZ twins share 50% of DNA variants and second, that both MZ and DZ twin pairs share the same common environment contributing to total phenotypic variance. Falconer's estimate (Falconer, D. and Mackay, T., 1996) provides a simple first impression of the magnitude of genetic influence in using MZ and DZ intraclass correlation (r). From a mathematical perspective, this comparison allows the assessment of the narrow-sense heritability (h^2) of the trait or disease (or broad-sense heritability if nonadditive genetic variations are incorporated in the formula). This heritability is computed as being twice the difference between monozygotic (r_{MZ}) and the same sex dizygotic (r_{DZ}) concordance rate, that is $h^2 = 2 \times (r_{MZ} - r_{DZ})$, while the shared environment factor is $|r_{MZ} - h^2|$ from MZ twins or $|r_{DZ} - h^2/2|$ from DZ twins. The twin correlation is computed as the addition of additive genetics ($A = h^2$) and common environment (C) factors and the covariance between them, that is, $r_{MZ} = A + C + cov(A,C)$ and $r_{DZ} = ½A + C + cov(½A,C)$, and is close to Pearson's correlation coefficient (variation between −1 and 1).

The second method is the model-fitting analysis of twin data using Restricted Maximum Likelihood (ReML) (Rijsdijk, F.V. and Sham, P.C., 2002), also known as the structural equation modeling (SEM) approach. It allows a more extensive separation and quantification of the observed phenotypic variance into its genetic and environmental components: additive genetic variance (A), nonadditive genetic variance [D (dominant) + I (epistatisis)], shared (or common) environmental variance (C), gene–environment interaction (G × E), and specific (or unique) environmental variance, which also contains measurement error (U (unique) + S (stochastic) + ε (errors in measurement)). The classic model of phenotypic variation is expressed as $V_P = V_A + V_C + V_E$. Two estimators of heritability are commonly defined. First, broad-sense heritability (H^2) takes into account total genetic variation (V_G) when we can define the proportion of all genetic variation (additive and nonadditive); thus $V_P = V_A + V_{D+I} + V_C + V_E$ and $H^2 = V_G/V_P$. Second, narrow-sense heritability (h^2) takes into account only the additive variation of total genetic variation (V_A); thus $V_P = V_A + V_C + V_E$ and $h^2 = V_A/V_P$. However, sometimes ReML can fail to converge when A is very small and also, no negative heritability values are allowed to be computed.

Because of the basic assumptions of the two latter methods, both formulae may lead to underestimates of heritability. Indeed, human DZ twins may share more than 50% of their genetic alleles because of parental homozygosity and assortative mating. Both effects inflate r_{DZ} but not r_{MZ}. Falconer's method can overestimate heritability by underestimating r_{DZ} and overestimating r_{MZ}. Indeed, r_{DZ} is underestimated by the computation of G × E interaction and potentially a sex difference in DZ twins, while r_{MZ} is overestimated by the calculation that identical twins and fraternal twins experience the same degree of environmental similarity. This latter assumption is likely to be violated because of greater similarity in the treatment of MZ twins by other people and the more similar intrauterine environment of MZ twins. On the other hand, Structural Equation Models estimate the proportion of different variables of models based on real data. A comparison between these two methodologies from a simulation of real fMRI data in a twin study showed that both classical ReML estimation and a variant of ReML with Bayesian priors on the variance parameters are more accurate than Falconer's method in both bias and standard deviation comparisons (Ang et al., 2012).

Overall, these studies highlight the power of twin studies to estimate heritability and thereby discordance, even if they may sometimes slightly overestimate heritability and underestimate the contribution of environmental factors.

The Trait/Disease-Discordant Monozygotic Twin Study

The discordant monozygotic twin study, also called the cotwin control method, is a powerful type of case–control study, eliminating genetic effects and subtracting shared environmental factors (Hogenson, 2013). It also has the advantage of completely adjusting for age and cohort effects and minimizing the effects of family and early life influences.

The studies of disease discordance between MZ twins have shown that even for heritable traits there is considerable discordance (Boomsma et al., 2002; Van Dongen et al., 2012) such as 61% of discordance for type I diabetes, 58% for type II diabetes (Condon et al., 2008) and 70–100% of discordance for different cancer types (Lichtenstein et al., 2000). This study design has allowed for the observation of the effect of specific genetic variations such as trisomy 21 (Bonilla et al., 2013; Letourneau et al. (2014)) or Mendelian inherited disorders and chromosomal abnormalities (Zwijnenburg et al., 2010). It also can explore environmental or stochastic factors, such as moderation alcohol consumption, which appears to act beneficially on bone health in female twins without apparent mediation by bone metabolism (Williams et al., 2005), or hemophilia related to a difference in randomization of X-inactivation, of which 30% comes from spontaneous mutation (Bennett et al., 2008).

The discordant MZ model also allows the exploration of causal relationships between two common traits, that is, whether they arise from shared genetic or environmental factors. For example, although autism spectrum disorder (ASD) has a high heritability estimate, Wong et al. (2014) showed that nongenetic factors play a key role in MZ twin discordance. They found that DMRs are site-specific and some are ASD case and family specific. The disease/trait-discordant MZ twin study is the most important and frequently used study to understand the environmental effect on the tuning of phenotype via the proxy that is epigenetic modification (Czyz et al., 2012).

The Multivariate Twin Study

Extension of models to several phenotypes that are potentially influenced by the same and/or different genetic and environmental factors, called multivariate twin models, helps us to unravel causality. Kendler et al. (1987) integrated both phenotype and endophenotypes in his model, to determine their common genetic etiology. Surprisingly, in this study, comorbidity between major depression (MD) and generalized anxiety disorder (GAD) corroborated that the two diseases share genetic factors, but their environmental effects discriminated them. Lervolino et al. (2011) explored the major symptom dimensions of Obsessive-compulsive disorder (OCD). They found that although heritability estimates ranged from 38% to 47%, independent A and E factors specific to each dimension of this disorder explain the remainder of the variance and C was negligible. Thus, the multivariate twin models allow us to explore the causes of comorbidity of two or more traits and test for G×E interactions. It can easily be extended to explore common and unique epigenetic modifications in common traits and diseases.

The Longitudinal Twin Study

Another extension is the longitudinal twin study, which follows the same cohort during a long period. It highlights the changes of a trait over a lifetime, related to environmental and G×E factors. It allows for the monitoring of genetic and epigenetic modifications depending on environmental factors. One example showed that DNA methylation differs in early childhood between young twins and that it was not stable over time (Martino et al. 2013). Using incident disease samples one can pinpoint potentially causal changes (Ng et al., 2012). For example, Heyn et al. (2013b) highlighted the docking protein DOK7 as a candidate for blood-based cancer diagnosis in a cross-section study from a cohort of 15 discordant MZ twins, in which blood samples in a subsample were taken several years before or after cancer diagnosis. This outcome was validated in 21 other discordant MZ twins by pyrosequencing.

Adding other omics data (Bell et al., 2014) or twins' relatives (Weber-Lehmann et al., 2014), the timing of the *de novo* variant or the stability of associations between epigenetic modifications and other omic or phenotypic data, may be defined. For example, Bell et al. (2014) observed both longitudinally stable and variable regions among the 100 top-ranked DMR effects on pain-sensitivity, but the top nine MAP-DMRs were predominantly stable with three-year data and showed altered gene expression in skin tissue.

The Twin Study Combining Different Omic Data

Epigenomic studies may include genotyping or sequencing and other omic data such as RNA-seq, Chip-seq (Van Dongen et al., 2012), RIBO-seq, or metabolomic data (Petersen et al., 2014). In MZ discordant pairs, if the difference between MZ twins cannot be explained at the DNA sequence level (Bruder et al., 2008; Weber-Lehmann et al., 2014), the combination of epigenetic data with other omic data will help to determine nongenetic factors responsible for the change in the trait or disease or to define potential biochemical mechanisms. Grundberg et al. (2013) combined DNA methylation variation (meQTL and meSNP) with gene expression (eQTL) and chromatin mark data (ChiP-seq from ENCODE) in a twin study. They showed a slight enrichment for negative correlation between DNA methylation and gene expression levels in promoter, gene-body regions, and 3′ UTR; but a positive correlation between gene-body methylation and gene expression level. This latter finding might improve transcription efficiency of actively transcribed genes. Additionally, they observed that a significant enrichment of expression meQTLs found in proximal regulatory elements (promoters) compared to all meQTLs for adipose tissue as well as other tissues. The expression meQTLs are also significantly enriched in distal regulatory elements (enhancer) in adipose tissue and metabolic disease loci. This latter outcome confirms that enhancer elements are involved in tissue-specific gene regulation.

Twin Studies Exploring Epigenetic Modifications in Trait or Diseases

Power to Detect Epigenetic Changes in Disease and Complex Trait

Epigenome-wide association studies (EWASs) explore the association between variation in epigenetic sites and a particular phenotype, without an a priori hypothesis. They are similar in concept to genome-wide association studies (GWAS). Statistical power to detect a phenotype-associated epigenetic variant should be a major criterion in EWAS study design. The power of an EWAS is the probability that an EWAS will find a statistically significant association between a trait/disease and epigenetic modifications, given that such an association exists. Power to detect disease-related epigenetic modification is affected by many factors. These include, but are not limited to the choice of epigenetic modifications; study design; "the effect size of [..]" the epigenetic modification "[..] on disease, the concordance of [..]" epigenetic modification "[..] levels between MZ twins at the locus of interest; the sample size; as well as the sensitivity, coverage, and reproducibility of [..]" the epigenetic modification "[..] assay" (Bell and Saffery, 2012). In addition, potential confounding and mediator factors, the choice of tissue and its cell composition, and the stability of the DNA methylation variants over time are also important (Jaffe and Irizarry, 2014).

First, the detection of epigenetic states depends on the accuracy and coverage of the technology selected. For example, methylation arrays such as the Illumina 27k and Illumina 450k arrays do not allow investigation outside of the predesigned probe set and simultaneous interrogation of both genetic mutation and epigenetic modification. A set of recent articles (Chen et al., 2013; Harper et al., 2013; Price et al., 2013) highlights additional technical problems with the probe design and analysis methods used by the Illumina 450k array. Some of them are relevant to other technologies such as polymorphic CpGs and multiple alignments of probes or reads to genome. An important step relevant to analysis of most methylation detection technologies is mapping back the probes or reads to the genome, which is classically done using a reference genome. Consequently, during the study of DNA methylation with Illumina 27k and 450K platform, probes containing

common SNPs or overlapping CNVs anywhere in the sequence are excluded, but the a priori knowledge of each individual's genotype will be required to make an appropriated interpretation and remove only probes affected by SNPs of samples.

Another important step is the definition of epigenetic regions. Some consider single-CpG sites, called differentially methylated positions (DMPs), while others take into account entire continuous genomic regions, which may contain multiple CpG sites. There are currently three types of epigenetic regions considered in epigenetic studies according to variables compared between groups: differentially methylated regions (DMRs) (Jaffe et al., 2012a) having a difference of mean between groups, variably methylated regions (VMRs) (Jaffe et al., 2012b) having a difference of variance between groups, and differentially variably methylated regions (differentially VMRs) (Ahn and Wang, 2013) having a difference of both mean and variance between groups.

Batch effects as well as other confounding factors such as age, population, smoking, and cancer status should be considered early in the study design and in the analysis (Harper et al., 2013; Jaffe and Irizarry, 2014). Other criteria playing a role in the interpretation of epigenetic change are the choice of biological annotation (Price et al., 2013) and network software (Harper et al., 2013), and the management of 'co-methylation' of probes (Liu et al., 2014).

Although MZ twins studies reduce genetic and shared environmental noise (age, early life environment), the composition of cell types and the timing of cell cycle phase across cells in each sample may differ between MZ twins and relate to the phenotype studied. Although cell counts within twin pairs are more similar than between unrelated individuals, as blood composition is highly heritable with a heritability of 61–96% (Evans et al., 1999), there is still a need to perform statistical analysis to control for blood cell composition (Jaffe and Irizarry, 2014). Houseman et al. (2012) designed a method which adjusts for the cell mixture distribution of white blood cells based on the Illumina 27k array, and extended this approach to Illumina 450k array data, with a further extension including reference-free correction. More recently, Zou et al. (2014) also proposed a reference-free method, FaST-LMM-EWASher, to correct for cell type composition. Other methods such as Jaffe and Irizarry (2014) have also been recently developed. These are currently being adapted for application to related samples, such as in twin studies.

The distribution of the studied epigenetic element often has a nonnormal distribution in a sample of individuals. It is possible to transform the values, for example, using a natural log transformation, but the interpretation becomes more difficult. Moreover, variance across the spectrum of methylation is not equal (Relton et al., 2012). Consequently, if the distribution does not conform to normality, the test statistic used should be nonparametric. It is possible to use linear regression but as with all statistical models, there are some assumptions that must be checked, such as constant variation and normality of error/residuals. The robustness of the model should be tested with, for example, resampling methods (permutation, bootstrapping). As the model is applied on a large number of epigenetic sites, multiple testing corrections such as False Discovery Rate (FDR) or Bonferroni adjustment should be applied with stringent threshold. Eventually, replication and validation studies should be performed to validate the outcomes.

Kaminsky et al. (2008) estimated the power to detect differential methylation effects using a high-throughput epigenomics microarray described by Schumacher et al. (2006) in a candidate gene analysis (CpG islands proximal to the homeobox DLX1 gene). They showed that a fold change of 1.2 was sufficiently greater than the null experimental threshold (1.15 null fold change threshold for technical variance) to detect differential methylation with 80% power and type I error of 0.001 after FDR adjustment. To detect a fold change of 1.2 with 80% power, the sample size should be between 15 and 25 of phenotypically discordant twin pairs, depending on number of technical replicates (1 replicate = 25 twins pair versus 4 replicates = 15 twins). The more the number of technical replicates increased, the more power increased, and the more technical variance decreased. Wang (2011) also estimated the power of EWAS but for the case-control design under different parameter settings with a sample size of 250 cases and 250 controls. Other approaches to quantifying methylation effects and power estimations have also been proposed, and using the methOR or methylation odds ratio to describe epigenetic effects (Rakyan et al., 2011b), using generalized regression methods to detect DMRs (Jaffe et al., 2012a) or differential VMR (Jaffe et al., 2012b), and using the first and second moments of methylation distribution (Ahn and Wang, 2013). These methods can also be extended to the epigenetic study of discordant MZ twins.

Epigenome-Wide Association Scans for Complex Traits and Diseases

Epigenome-Wide Association Studies explore variation in epigenetic sites associated with phenotype or environmental factors without an a priori hypothesis (Rakyan et al., 2011b). They are analogous to genome-wide association studies (GWAS), but epigenotype can be considered as an exposure or an outcome or also a mediator. Initially reported EWAS (**Table 3**) to date used the Illumina 27K assay (for example Bell et al., 2012; Gervin et al., 2012), and more recently the Illumina 450K platform (for example, Grundberg et al., 2013; Heyn et al., 2013b; Joubert et al., 2014), as well as high-throughput next-generation sequencing technologies, for example, RRBS (Baranzini et al., 2010) or MeDIP-seq (Bell et al., 2014; Rakyan et al., 2011b; Michels et al., 2013) have reviewed in further detail EWAS design and some examples.

EWAS' may be performed using many study designs e.g., case–control, family or discordant MZ twins. Two important steps in any study are the size of the replication sample and the strategy for validation by other methods. A meta-analysis EWAS combining a discovery discordant MZ twin study and a follow-up unrelated individual study using MeDIP sequencing was used to explore epigenetic effects in pain sensitivity (Bell et al., 2014). This meta-analysis EWAS allowed the definition of nine pain-related phenotype differentially methylated regions (pain-DMRs) without an a priori hypothesis in blood samples. The analysis was performed in two distinct samples, 50 MZ twins identifying about 100 pain DMRs and 50 unrelated individuals identifying 22 pain DMRs respectively. A pain sensitive meta-analysis (MA) highlighted

Table 3 Epigenome-wide epigenetic studies in twins. The list is obtained from a Pubmed search with search terms 'Epigenetic,' 'twin,' 'DNA methylation,' 'histone,' 'microRNA' listing all published studies up to 30 April 2014, without thresholding on sample size

Phenotypes/disease	*Sample Tissue/cell source*[a]	*Epigenetic modification*	*Epigenetic assay*[b]	*Analysis replication*[c]	*Major findings*	*References*
Aging	80 twins; lymphocyte cells; skin cells; muscle and abdominal tissue biopsy	DNA methylation; histone modification	High-performance capillary electrophoresis; AIMS	Anova; replication; validation	Epigenetic drift varies over time and between tissues, but the same overall manner	Fraga et al. (2005)
Risk-taking behavior	1 discordant Female MZ pair; peripheral blood cells	DNA methylation	Microarray array	Paired t-test; replication; validation	Association of risk-taking behavior with 1.2 fold change in DNA methylation of locus DLX1	Kaminsky et al. (2008)
Age and gender, genotype	23 MZ twin-pairs and 23 DZ twin-pairs + 96 healthy controls; whole blood	DNA methylation	Illumina GoldenGate	LMER	Strong association of genetic variant with DMR in cis	Boks et al. (2009)
Heritability of epigenetic change + Twinning process	114 MZ twins + 80 DZ twins; WBC, buccal cells and gut biopsies	DNA methylation	12K microarray array	Methylation difference, ICC; validation in animal studies	Concordance level of methylation: diochorionic MZ > monochorionic MZ > DZ	Kaminsky et al. (2009)
Multiple sclerosis (MS)	3 MZ pairs discordant for MS; $CD4^+$ T cells	DNA methylation; gene expression	Illumina RRBS	N.A.	No association of MS with genetic, epigenetic or transcriptome differences	Baranzini et al. (2010), Baranzini, (2011)
Systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and dermatomyositis (DM)	5 MZ twin pairs discordant for each SLE, RA, and DM; 30 unrelated controls; WBC	DNA methylation; gene expression	Illumina GoldenGate	Student's t-test; validation	49 significant DMRs in SLE discordant MZ twins	Javierre et al. (2010)
Autism spectrum disorders (ASD)	3 discordant Male MZ pairs; lymphoblastoid cell lines	DNA methylation; gene expression	CpG island microarrays	Fisher's exact test; validation	73 significant DMRs with ASD;	Nguyen et al. (2010)
Aging	93 healthy Female: 14 MZ and 17 DZ + 31 singleton; PBMCs	DNA methylation	Illumina 27K	Spearman's rank correlation; replication; validation	213 hyper-Methylated DMRs and 147 hypo-methylated DMRs wit aging	Rakyan et al. (2010)
Autism spectrum disorders (ASD)	3 discordant MZ twins + 1 normal sibling for 2 of the twin pairs + 2 autism concordant pairs and unaffected siblings, and 1 pair of normal MZ twins; LCL derived from peripheral lymphocytes	MicroRNA; gene expresion	Custom-printed miRNA microarrays	Hierarchical cluster analysis and PCA; replication; validation	Association of microRNA with ASD and other co-morbid disorders, including muscle and gastrointestinal diseases, as well as with biological functions implicated in ASD, such as memory and synaptic plasticity.	Sarachana et al. (2010)
Systemic lupus erythematosus (SLE)	1 SLE-discordant MZ in case-control study (26 SLE affected and 26 unaffected controls)	MicroRNA	miRXplore microarray product	miRNA ratio calculation: Hy5/Hy3 ratios validation	Association of SLE with 31 miRNAs differentially expressed	Te et al. (2010)

Aging	34 male MZ pair twin (21-55); saliva	DNA methylation	Illumina 27K	Regression model; validation; replication	Association of 88 DMRs with ageing; 3 genes validated in other cohorts	Bocklandt et al. (2011)
Schizophrenia (SZ) and bipolar disorder (BD)	11 SZ-discordant pairs and 11 BD-discordant pairs; WBC	DNA methylation	Illumina 27K	Paired t-test; validation	Association of ST6GALNAC1 gene with major psychosis.	Dempster et al. (2011)
Childhood-onset type 1 Diabetes (T1D)	15 MZ pairs discordant for T1D; $CD14^{+}$ cells	DNA methylation	Illumina 27K	Wilcoxon signed-rank test; validation; replication	Association of 132 CpG sites with T1D	Rakyan et al. (2011a)
Multiple sclerosis (MS)	8 MS patients, 10 healthy donors, 2 female MZ twins discordant for MS; PBMCs	microRNA	GeneChip miRNA array (Affymetrix).	Anova	Association of microRNA differential expressed with MS	Umeton et al. (2011)
Age and chronological	172 female MZ twins (32-80); whole blood	DNA methylation	Illumina 27K	LMER; replication	Association of age with 16 age-related phenotypes DMRs and 490 chronological age DMRs	Bell et al. (2012)
Schizophrenia (SZ)	38 psychiatric control pairs, 15 SZ discordant twins, 27 SZ concordant twins; whole blood	DNA methylation	Global methylation by nonradioactive elongation-assay and promoter specific methylation	Anova	Association of SZ with DNA methylation differential expressed	Bönsch et al. (2012)
Psoriasis	27 MZ pairs discordant for psoriasis; $CD4^{+}$ and $CD8^{+}$ cells	DNA methylation; gene expression	Illumina 27K; illumina HumanHT-12 v3	Paired t-test; replications	No association of psoriasis with DNA methylation and gene expression	Gervin et al. (2012)
Birth weight, aging	22 MZ and 12 DZ twin pairs; 3 neonatal tissues (CBMC, HUVEC, placenta), blood	DNA methylation	Illumina 27K; EpiTYPING	LMER	Association of 7 genes in DZ CBMCs and 1 in MZ HUVECs with birth weight	Gordon et al. (2012)
Type 2 diabetes + exercise intervention	15 men with (FH+) and 13 men without (FH−) a first-degree FH of T2D + 9 T2D discordant MZ pairs; muscle biopsies	DNA methylation; gene expression; mitochondrial density	MeDIP-Chip The human 2.1 promoter DeLuxe tiling array; affymetrix custom-array NuGO-Hs1a520180-GeneChip; HumanOmniExpress arrays	Two-sample Mann–Whitney; validation	Association of DNA methylation differential expressed with T2D and also with exercise intervention	Nitert et al. (2012)
Type 2 diabetes (T2D)	16 discordant MZ; skeletal muscle and adipose tissue biopsies	DNA methylation	Illumina 27K	Anova; validation	Association of T2D with IL8 and PPARGC1A gene in muscle and with MODY and HNF4A gene in adipose tissue	Ribel-Madsen et al. (2012)
Scleroderma	7 discordant MZ and 1 concordant MZ; PBMC	DNA methylation; X-inactivation	MeDIP; NimbleGen array hybridizatio 4-plex array hybridization kit	Methylation peaks with *p*-score> 1.31	Association of scleroderma with 18 hypermethylated and 25 hypomehtylated sites	Selmi et al. (2012)
Intelligence levels (IQ)	17 MZ pairs; PBMC	DNA methylation; gene expression	Array-based genome-wide DNA methylation and gene expression analyses	Anova; validation	No association of IQ with DNA methylation differentially expressed	Yu et al. (2012)
Down's syndrome (Trisomy 21)	1 MZ twin discordant; fibroblasts	Binding-site of DNA-associated proteins; DNA methylation	ChIP seq for H3K4me3	N.A	Association of trisomie 21 with histone modification;	Bonilla et al. (2013)

(*Continued*)

Table 3 Epigenome-wide epigenetic studies in twins. The list is obtained from a Pubmed search with search terms 'Epigenetic,' 'twin,' 'DNA methylation,' 'histone,' 'microRNA' listing all published studies up to 30 April 2014, without thresholding on sample size—cont'd

Phenotypes/disease	*Sample Tissue/cell source*[a]	*Epigenetic modification*	*Epigenetic assay*[b]	*Analysis replication*[c]	*Major findings*	*References*
Major depressive disorder (MDD)	12 MZ twin pairs discordant for MDD and 12 MZ twin pairs concordant for no MDD and low neuroticism; WBC	DNA methylation	Illumina 450K	Wilcoxon signed-rank test; LMER; replication	No association of MDD with DNA methylation;	Byrne et al. (2013)
Type 1 diabetes (T1D)	Quadruplets (2 sisters with T1D-1 and T1D-2, 1sister with Pre-T1D and 1 healthy control sister); $CD14^+$ monocytes and $CD4^+$ T cells	DNA methylation	Illumina 450K	Cluster analysis; replications	Association of T1D with 164 DMPs	Disanto et al. (2013)
Systemic lupus erythematosus (SLE)	1 MZ twins discordant; PBMC	DNA methylation; gene expression	Genotype using Illumina human 610-Quad BeadChips; Illumina 27K; Agilent 4X44k	N.A.	Association of SLE with CpG methylation and gene expression levels different for 10 genes	Furukawa et al. (2013)
Metabolic traits	648 twins Used (97 MZ pairs, 162 DZ pairs, and 130 singletons); adipose tissue	DNA methylation; gene expession	Illumina 450K; IlluminaHT12 array; combination of Illumina arrays (HumanHap300, human-Hap610Q, 1M-Duo, and 1.2MDuo 1M).	LMER; replication; validation	Association of metQTL with metabolic-trait or disease loci in enhancer; association meQTL with adipose tissue-specific	Grundberg et al. (2013)
Breast cancer	3 discordant MZ twin pairs; whole blood	DNA methylation	Illumina 450K	Wilcoxon rank sum paired test; validation	*DOK7* as a candidate for blood-based cancer diagnosis given a hypermethylation before tumor diagnosis	Heyn et al. (2013b)
Schizophrenia (SZ)	24 medication-free patients with SCZ and 23 non-psychiatric controls + 3 MZ twin discordant for SCZ; peripheral leukocytes	DNA methylation	Illumina 450K	Paired t-test	Association of SZ with 234 DMPs	Kinoshita et al. (2013)
Twin–twin transfusion syndrome (TTTS)	14 individuals (4 twin pairs who survived TTTS; following corrective surgery 6 sole survivors); Peripheral blood; saliva	DNA methylation	Bisulfite pyrosequencing; illumina 27K	LMER	Association of TTTS with hypomethylation of the LINE1 repetitive element	Marsit et al. (2013)
Aging	10 MZ and 5 DZ twin pairs; buccal	DNA methylation	Illumina 450K	Paired t-test; replication; validation	Association of aging with 1/3 of the CpGs sites assayed	Martino et al. (2013)
Rett syndrome	1 MZ for the severity of discordant (milder and more severe); PBMC and skin fibroblasts	X-inactivation; gene expression; DNA methylation	HUMARA; qRT-PCR; Illumina 450K; ABI Prism 310 sequencer; Whole-genome sequencing using HiSeq2000 sequencers	Value of DiffScore of Illumina software; validation	Association of Rett syndrome with 3 DMRs	Miyake et al. (2013)

Autism spectrum disorder (ASD)	50 MZ twin pairs (discordant and concordant for ASD, ASD-associated traits and no autistic phenotype.); Whole blood	DNA methylation	Illumina 27K; Illumina HumanOmniExpress BeadChip	Replication	No association of ASD or ASD-related trait with differential global in DNA methylation, but site-specific such as NFYC promoter;	Wong et al. (2013)
Heat pain sensitivity	50 MZ twins discordant + 50 unrelated individuals longitudinal data for 33 individuals in 50 discordant MZ twins at sampled 2–3 years apart; whole blood	DNA methylation; gene expression	MeDIP-seq	LMER on 2 independent EWASs + meta-analysis; validation	Association of strongest signal in the pain gene TRPA1	Bell et al. (2014)
Major depressive disorder (MDD)	27 UK discordant MZ twin pairs + 23 Australian MZ twin pairs; whole blood cell	DNA methylation; gene expression	MeDIP-seq	LMER; replication	Association of MDD with hypermethylation within the coding region of ZBTB20 and 17 DMRs of genome-wide significance; No association total gene expression level of ZBTB20 with MDD;	Davies et al. (2014)
Congenital renal agenesis	1 discordant MZ twin pair; whole blood cell	DNA methylation	Illumina RRBS; exome- seq	N.A.	Association of Congenital renal agenesis with 514 DMRs in 25 genes	Jin et al. (2014)
Down's syndrome (Trisomy 21)	1 trisomy 21 discordant MZ (T1DS, T2N) + 16 other cohorts of trisomy 21 and normal unrelated individuals; forearm primary fetal skin fibroblasts (T1DS + T2N); primary fibroblasts from umbilical cord tissue	Gene expression + binding-site of DNA-associated proteins + DNA methylation	RNA-seq in Illumina HiSeq 2000 and on the Illumina Hiseq 2500; H3K4me3 ChIP-seq on a HiSeq2500; DNase I hypersensitivity mapping; Illumina RRBS	Paired-t test technical replication: 4 replication: mouse	Definition of gene expression dysregulation domains (GEDDs); associated T21 with regions associated with nuclear lamina proteins or DNA-synthesis phase of the cell–division cycle;	Letourneau et al. (2014)
Type 1 diabetes (T1D)	3 MZ twin pairs discordant for T1D and 6 T1D-concordant MZ twin pairs; lymphocyte cell lines	DNA methylation	Illumina 27K	Paired LIMMA test FDR correction; validation	Association of T1D with 88 CpG sites	Stefan et al. (2014)

[a]*Tissue*: WBC = white blood cells; PBMCs = peripheral blood mononuclear cells.
[b]*Epigenetic assay*: AIMS = amplification of intermethylated sites; Illumina GoldenGate = Illumina GoldenGate bead array technology; 12K = 12K CpG island microarrays; Illumina 27K = Illumina Infinium® Human Methylation27 BeadChip; EpiTYPING = Sequenom MassARRAY EpiTYPING; Illumina 450K = Illumina Infinium® Human Methylation450 BeadChip; HUMARA = human androgen receptor gene locus assay using methylation-specific PCR; Illumina RRBS = DNA methylation: reduced representation bisulfite sequencing; *Gene expression*: Illumina HumanHT-12 v3 = Illumina HumanHT-12 v3 Expression BeadChip; Agilent 4X44k = Whole human genome oligo DNA microarrays 4X44k (Agilent Technologies, Santa Clara, CA); qRT-PCR = quantitative reverse transcription PCR assay; exome-seq = high-throughput exome-capture sequencing.
[c]*Analysis*: LMER = linear mixed-effects regression; DMR = differentially methylated region; N.A. = not described in the article.
Baranzini et al., 2010; Baranzini, 2011; Bell et al., 2012; Bell et al., 2014; Bocklandt et al., 2011; Boks et al., 2009; Bonsch et al., 2012; Bonilla et al., 2013; Byrne et al., 2013; Davies et al., 2014; Dempster et al., 2011; Disanto et al., 2013; Fraga et al., 2005; Furukawa et al., 2013; Gervin et al., 2012; Gordon et al., 2012; Grundberg et al., 2013; Heyn et al., 2013; Javierre et al., 2010; Jin et al., 2014; Kaminsky et al., 2008; Kaminsky et al., 2009; Kinoshita et al., 2013; Letourneau et al., 2014; Martino et al., 2013; Marsit et al., 2013; Miyake et al., 2013; Nguyen et al., 2010; Nitert et al., 2012; Rakyan et al., 2010; Rakyan et al., 2011a; Ribel-Madsen et al., 2012; Sarachana et al., 2010; Selmi et al., 2012; Stefan et al., 2014; Te et al., 2010; Umeton et al., 2011; Wong et al., 2013; Yu et al., 2012.

nine MAP-DMRs of which one was in the promoter of TRPA1, an established pain-sensitivity gene. The 100 top-ranked MAP-DMRs had significant positive correlation between levels of methylation in blood and brain from two postmortem donors and a negative correlation with skin gene expression levels, as expected, which suggests that blood may be a relevant surrogate tissue in this example.

Twin Studies, Epigenetics and Environmental Influences on Diseases

Several epidemiological studies have worked on the association between epigenetic modifications and complex traits, as well as the relationship between epigenetic modifications and these environmental factors (Cortessis et al., 2012). **Tables 3** and 4 (Supplementary Data) reports examples of twin studies exploring the relationship between epigenetics and complex traits or diseases and highlights that most of these twin studies are disease-discordant MZ twins. Few examples are described below.

Smoking is considered to be one of the most important risk factors for many diseases, such as heart disease, chronic obstructive pulmonary disease, stroke, acute respiratory diseases, and cancer. Epigenetics allows for the opportunity to study how linking environmental factor, methylation, and gene expression could give insights into the biological mechanisms underlying multiple diseases. One study has revealed that active smoking as well as secondhand tobacco smoking may trigger breast cancer (Luo et al. 2011). Moreover, levels of DNA methylation in breast cancer such as on tumor suppressor genes change, according to both smoking exposure and diet (Luo et al., 2011). On the other hand, a discordant MZ twin study working only on sporadic breast cancer facilitated detection of 403 DMRs including known and novel potential breast cancer genes (Heyn et al., 2013b). Moreover, they showed that the docking protein DOK7 seems to be a potential candidate as a blood-based cancer diagnostic.

Despite increasing associations between epigenetic modifications, environmental factors, and disease, it is still difficult to gauge if epigenetic biomarkers are the causes or the consequences of disease. Causal modeling has been suggested as an approach to answer this question but may not be definitive (Haslberger et al., 2006; Baedke, 2012; Bellavia et al., 2013). Bellavia et al. (2013) showed with elegance the effect of coarse and fine particulate-matter on blood pressure via DNA hypomethylation by using a mediation analysis. Longitudinal studies, if started early enough, will hopefully allow us to distinguish between cause and consequence and to define potential diagnostic biomarkers.

Future Perspectives and Limitations

Limitations

The study of certain complex traits and diseases from twin registers may be difficult given that the number of discordant MZ twin pairs (sample size) is not large enough to detect significantly differentially modified regions, particularly if we study rare diseases. Moreover, following twins for long periods is a recurrent major challenge for twin registers.

Future Prospects

The detailed study of epigenetic markers, other than DNA methylation, using discordant twin studies or in single cell would improve our knowledge of epigenetics immensely. This will depend on advances in technology making high-throughput methods feasible, such as for example, the LC-MS/MS method that was being recently developed to measure global histone modifications (Zhang et al., 2010), for ChIP-seq (Gilfillan et al., 2012), or for single cell (Lorthongpanich et al., 2013; Nawy, 2013).

Combining epigenetics with other omics data will allow us to better understand the relationship between epigenetic factors, omics data, and environmental factors for the trait/disease of interest, such as for example (Petersen et al., 2014), who analyzed the association between metabolites and DNA methylation. Recently, Letourneau et al. (2014) studied the combination of data from DNA methylation, gene expression, and histone modification to unravel the origin of the perturbation of gene expression in Down's syndrome. They discovered that differential gene expression regions are organized into genome-wide domains, not only on chromosome 21. These regions, denoted gene expression dysregulation domains (GEDDs), could be either upregulated or downregulated and were not correlated with differential DNA methylation, but with H3K4me3 profiles. This discovery opens a new door to unravel complex traits and diseases and it will allow us to understand the impact of environmental factors on phenotype via epigenetic modifications. If similar multi-omic data studies are performed at the population level, the stratification of exposure effect or of treatment efficacy may be identified.

Further work in mediation analysis should lead to the better estimation of the effects of environmental exposure on a phenotype mediated by epigenetic postexposure changes (Bellavia et al., 2013). This would require significant relationships between exposure-outcome, exposure-mediator, and mediator-outcome. Such studies will allow us to understand disease mechanisms and the mediator role of epigenetic changes. Furthermore, they will help to identify biomarkers that may predict disease or inform us about prognosis.

There are at least 10 times as many bacteria as human cells in the body (Eckburg et al., 2005). Epigenetics, as well as the microbiome can constitute proxies of environmental exposure to tune gene expression and, at high level, phenotype. The study of the relationship between genome, phenotype, and endophenotypes (epigenetics, microbiome) are crucial to explore their interactions and symbiotic and pathogenic actions on the host. Using twins will hopefully allow us in the future to unravel the relationship between microbiome and epigenome on the development of disease (Steward et al., 2013).

Each phenotype of each individual is a subtle, personal, and dynamic combination between genetic, environmental, and stochastic factors. Thus, individuals via their behaviors, diets, and lifestyles can act directly or indirectly on their epigenetic markers and consequently, on their phenotypes (Crews, 2008; Powledge, 2011). In medical perspective, complex diseases could be treated by personalized medication being able to act on or by epigenetic markers, called epigenetic readers, erasers, or writers, according to their epigenetic mechanisms

(Szyf, 2009; Cohen et al., 2011; Lu et al., 2012). Additionally, twin studies could also help in modern pharmacogenomics studies (Rahmioğlu and Ahmadi, 2010).

Acknowledgments

The authors would like to thank Idil Erte, Craig Glastonbury, Alan D. Jackson, Laetitia J. Martin, and Leonie Roos for their comments. TDS is an NIHR Senior Investigator and a European Research Council Senior Research Investigator (ERC 250157). The study also received support from the TwinsUK resource, funded by the Wellcome Trust, with support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and the Biomedical Research Center based at Guy's, and the St Thomas' NHS Foundation Trust in partnership with King's College London.

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/B978-0-08-097086-8.82051-6.

See also: Bioethics in the Post-genomic Era; Epigenetic Inheritance; Epigenetics of Brain Plasticity and Behavior; Epigenetics.

Bibliography

Ahn, S., Wang, T., 2013. A powerful statistical method for identifying differentially methylated markers in complex diseases. Pacific Symposium on Biocomputing 69–79.

Ang, S.T., Chen, X., Nichols, T., 2012. A comparison of heritability estimators in brain imaging. Reinvention: A Journal of Undergraduate Research 5, 1.

Baedke, J., 2012. Causal explanation beyond the gene: manipulation and causality in epigenetics. Theoria 27, 2.

Baranzini, S.E., Mudge, J., van Velkinburgh, J.C., Khankhanian, P., Khrebtukova, I., Miller, N.A., Zhang, L., Farmer, A.D., Bell, C.J., Kim, R.W., May, G.D., Woodward, J.E., Caillier, S.J., McElroy, J.P., Gomez, R., Pando, M.J., Clendenen, L.E., Ganusova, E.E., Schilkey, F.D., Ramaraj, T., Khan, O.A., Huntley, J.J., Luo, S., Kwok, P.Y., Wu, T.D., Schroth, G.P., Oksenberg, J.R., Hauser, S.L., Kingsmore, S.F., 2010. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464 (7293), 1351–1356.

Baranzini, S.E., 2011. Revealing the genetic basis of multiple sclerosis: are we there yet? Current Opinion in Genetics and Development 21 (3), 317–324.

Bell, J.T., Saffery, R., 2012. The value of twins in epigenetic epidemiology. International Journal of Epidemiology 41 (1), 140–150.

Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., Pritchard, J.K., 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology 12, R10.

Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., MuTHER Consortium, Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D., Deloukas, P., 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLOS Genetics 8 (4), e1002629.

Bell, J.T., Loomis, A.K., Butcher, L.M., Gao, F., Zhang, B., Hyde, C.L., Sun, J., Wu, H., Ward, K., Harris, J., Scollen, S., Davies, M.N., Schalkwyk, L.C., Mill, J., MuTHER Consortium, Williams, F.M., Li, N., Deloukas, P., Beck, S., McMahon, S.B., Wang, J., John, S.L., Spector, T.D., 2014. Differential methylation of the TRPA1 promoter in pain sensitivity. Nature Communications 2014 (5), 2978.

Bellavia, A., Urch, B., Speck, M., Brook, R.D., Scott, J.A., Albetti, B., Behbod, B., North, M., Valeri, L., Bertazzi, P.A., Silverman, F., Gold, D., Baccarelli, A.A., 2013. DNA hypomethylation, ambient particulate matter, and increased blood pressure: findings from controlled human exposure experiments. Journal of the American Heart Association 2 (3), e000212.

Bennett, C.M., Boye, E., Neufeld, E.J., 2008. Female monozygotic twins discordant for hemophilia A due to nonrandom X-chromosome inactivation. American Journal of Hematology 83 (10), 778–780.

Bocklandt, S., Lin, W., Sehl, M.E., Sánchez, F.J., Sinsheimer, J.S., Horvath, S., Vilain, E., 2011. Epigenetic predictor of age. PLoS ONE 6 (6), e14821.

Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S., Ophoff, R.A., 2009. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PLoS ONE 4 (8), e6767.

Bonilla, X., Letourneau, A., Santoni, F.A., Sailani, M.R., Guipponi, M., Gehrig, C., Antonarakis, S.E., 2013. Domains of histone marks in monozygotic twins discordant for trisomy 21. American Society of Human Genetics, Boston, USA.

Bonsch, D., Wunschel, M., Lenz, B., Janssen, G., Weisbrod, M., Sauer, H., 2012. Methylation matters? Decreased methylation status of genomic DNA in the blood of schizophrenic twins. Psychiatry Research 198 (3), 533–537.

Boomsma, D., Busjahn, A., Peltonen, L., 2002. Classical twin studies and beyond. Nature Reviews Genetics 3 (11), 872–882.

Breitling, L.P., Yang, R., Korn, B., Burwinkel, B., Brenner, H., 2011. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. American Journal of Human Genetics 88 (4), 450–457.

Brothers, J.F., Hijazi, K., Mascaux, C., El-Zein, R.A., Spitz, M.R., Spira, A., 2013. Bridging the clinical gaps: genetic, epigenetic and transcriptomic biomarkers for the early detection of lung cancer in the post-National Lung Screening Trial Era. BMC Medicine 11, 168.

Bruder, C.E., Piotrowski, A., Gijsbers, A.A., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E.C., Tiwari, H., Allison, D.B., Komorowski, J., van Ommen, G.J., Boomsma, D.I., Pedersen, N.L., den Dunnen, J.T., Wirdefeldt, K., Dumanski, J.P., 2008. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Journal of Molecular Biology 82 (3), 763–771.

Byrne, E.M., Carrillo-Roa, T., Henders, A.K., Bowdler, L., McRae, A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Krause, L., Wray, N.R., 2013. Monozygotic twins affected with major depressive disorder have greater variance in methylation than their unaffected cotwin. Translational Psychiatry 3, e269.

Beck, S., Olek, A., Walter, J., 1999. From genomics to epigenomics: a loftier view of life. Nature Biotechnology 17 (12), 1144.

Chen, R.J., Chang, L.W., Lin, P., Wang, Y.J., 2011. Epigenetic effects and molecular mechanisms of tumorigenesis induced by cigarette smoke: an overview. Journal of Oncology 2011, 654931.

Chen, Y.A., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R., 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8 (2), 203–209.

Cohen, I., Poręba, E., Kamieniarz, K., Schneider, R., 2011. Histone modifiers in cancer: friends or foes? Genes and Cancer 2 (6), 631–647.

Colaneri, A., Wang, T., Pagadala, V., Kittur, J., Staffa Jr., N.G., Peddada, S.D., Isganaitis, E., Patti, M.E., Birnbaumer, L., 2013. A minimal set of tissue-specific hypomethylated CpGs constitute epigenetic signatures of developmental programming. PLoS ONE 8 (9), e72670.

Condon, J., Shaw, J.E., Luciano, M., Kyvik, K.O., Martin, N.G., Duffy, D.L., 2008. A study of diabetes mellitus within a large sample of Australian twins. Twin Research and Human Genetics 11 (1), 28–40.

Cortessis, V.C., Thomas, D.C., Levine, A.J., Breton, C.V., Mack, T.M., Siegmund, K.D., Haile, R.W., Laird, P.W., 2012. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. Human Genetics 131, 1565–1589.

Crews, D., 2008. Epigenetics and its implications for behavioral neuroendocrinology. Frontiers in Neuroendocrinology 29 (3), 344–357.

Czyz, W., Morahan, J.M., Ebers, G.C., Ramagopalan, S.V., 2012. Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. BMC Medicine 10, 93.

Dallapiccola, B., Stomeo, C., Ferranti, G., Di Lecce, A., Purpura, M., 1985. Discordant sex in one of three monozygotic triplets. Journal of Medical Genetics 22, 6–11.

Davies, M.N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., Coarfa, C., Harris, R.A., Milosavljevic, A., Troakes, C., Al-Sarraj, S., Dobson, R., Schalkwyk, L.C., Mill, J., 2012. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. Genome Biology 13, R43.

Davies, M.N., Krause, L., Bell, J.T., Gao, F., Ward, K.J., Wu, H., Lu, H., Liu, Y., Tsai, P.C., Collier, D.A., Murphy, T., Dempster, E., Mill, J., Battle, A., Mostafavi, S., Zhu, X., Henders, A., Byrne, E., Wray, N.R., Martin, N.G., Spector, T.D., Wang, J., 2014. Hypermethylation in the ZBTB20 gene is associated with major depressive disorder. Genome Biology 15 (4), R56.

Daxinger, L., Whitelaw, E., 2010. Transgenerational epigenetic inheritance: more questions than answers. Genome Research 20 (12), 1623–1628.

Dempster, E.L., Pidsley, R., Schalkwyk, L.C., Owens, S., Georgiades, A., Kane, F., Kalidindi, S., Picchioni, M., Kravariti, E., Toulopoulou, T., Murray, R.M., Mill, J., 2011. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. Human Molecular Genetics 20 (24), 4786–4796.

Disanto, G., Vcelakova, J., Pakpoor, J., Elangovan, R.I., Sumnik, Z., Ulmannova, T., Ebers, G.C., Ramagopalan, S.V., Stechová, K., 2013. DNA methylation in monozygotic quadruplets affected by type 1 diabetes. Diabetologia 56 (9), 2093–2095.

Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F., 1998. Rates of spontaneous mutation. Genetics 148 (4), 1667–1686.

Duffié, R., Bourc'his, D., 2013. Chapter nine – parental epigenetic asymmetry in mammals. Current Topics In Developmental Biology 104, 293–328. Epigenetics and Development.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., Relman, D.A., 2005. Diversity of the human intestinal microbial flora. Science 308 (5728), 1635–1638.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., Beck, S., 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genetics 38 (12), 1378–1385.

Esteller, M., 2011. Non-coding RNAs in human disease. Nature Reviews Genetics 12, 861–874.

Evans, D.M., Frazer, I.H., Martin, N.G., 1999. Genetic and environmental causes of variation in basal levels of blood cells. Twin Research 2 (4), 250–257.

Fadloun, A., Eid, A., Torres-Padilla, M.E., 2013. Chapter one – mechanisms and dynamics of heterochromatin formation during mammalian development: closed paths and open questions. Current Topics in Developmental Biology 104, 1–45.

Falconer, D.S., Mackay, T., 1996. Introduction to Quantitative Genetics. Longmans Green, London.

Feinberg, A.P., Irizarry, R.A., 2010. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences of the United States of America 107 (Suppl 1), 1757–1764.

Flom, J.D., Ferris, J.S., Liao, Y., Tehranifar, P., Richards, C.B., Cho, Y.H., Gonzalez, K., Santella, R.M., Terry, M.B., 2011. Prenatal smoke exposure and genomic DNA methylation in a multiethnic birth cohort. Cancer Epidemiology Biomarkers and Prevention 20 (12), 2518–2523.

Fraga, M., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suñer, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C., Esteller, M., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. Proceedings of the National Academy of Sciences of the United States of America 102 (30), 10604–10609.

Fraser, H.B., Lam, L.L., Neumann, S.M., Kobor, M.S., 2012. Population-specificity of human DNA methylation. Genome Biology 13, R8.

Fujita, T., Hodaka Fujii, H., 2013. Locus-specific biochemical epigenetics/chromatin biochemistry by insertional chromatin immunoprecipitation (iChIP). Epigenetics Chromatin 6 (Suppl. 1), 20.

Furukawa, H., Oka, S., Matsui, T., Hashimoto, A., Arinuma, Y., Komiya, A., Fukui, N., Tsuchiya, N., Tohma, S., 2013. Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus. Human Immunology 74 (2), 170–175.

Galton, F., 1876. The history of twins, as a criterion of the relative powers of nature and nurture. The Journal of the Anthropological Institute of Great Britain and Ireland 5, 391–406.

Gervin, K., Vigeland, M.D., Mattingsdal, M., Hammerø, M., Nygård, H., Olsen, A.O., Brandt, I., Harris, J.R., Undlien, D.E., Lyle, R., 2012. DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. PLOS Genetics 8 (1), e1002454.

Gilfillan, G.D., Hughes, T., Sheng, Y., Hjorthaug, H.S., Straub, T., Gervin, K., Harris, J.R., Undlien, D.E., Lyle, R., 2012. Limitations and possibilities of low cell number ChIP-seq. BMC Genomics 13, 645.

Gordon, L., Joo, J.E., Powell, J.E., Ollikainen, M., Novakovic, B., Li, X., Andronikos, R., Cruickshank, M.N., Conneely, K.N., Smith, A.K., Alisch, R.S., Morley, R., Visscher, P.M., Craig, J.M., Saffery, R., 2012. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Research 22 (8), 1395–1406.

Grundberg, E., Meduri, E., Sandling, J.K., Hedman, A.K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K.S., Ge, B., Caron, M., Shin, S.Y., Multiple Tissue Human Expression Resource Consortium, Lathrop, M., Dermitzakis, E.T., McCarthy, M.I., Spector, T.D., Bell, J.T., Deloukas, P., 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. Current Opinion in Genetics and Development 93 (5), 876–890.

Guibert, S., Weber, M., 2013. Chapter two – functions of DNA methylation and hydroxymethylation in mammalian development. Current Topics in Developmental Biology 104, 47–83.

Halfmann, R., Lindquist, S., 2010. Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science 330 (6004), 629–632.

Harper, K.N., Peters, B.A., Gamble, M.V., 2013. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. Cancer Epidemiology Biomarkers and Prevention 22 (6), 1052–1060.

Haslberger, A., Varga, F., Karlic, H., 2006. Recursive causality in evolution: a model for epigenetic mechanisms in cancer development. Medical Hypotheses 67 (6), 1448–1454.

Hecht, S.S., 1999. Tobacco smoke Carcinogens and lung Cancer. Journal of the National Cancer Institute 91 (14), 1194–1210.

Heijmans, B.T., Kremer, D., Tobi, E.W., Boomsma, D.I., Slagboom, P.E., 2007. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. Human Molecular Genetics 16 (5), 547–554.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., Esteller, M., 2013a. DNA methylation contributes to natural human variation. Genome Research 23 (9), 1363–1372.

Heyn, H., Carmona, F.J., Gomez, A., Ferreira, H.J., Bell, J.T., Sayols, S., Ward, K., Stefansson, O.A., Moran, S., Sandoval, J., Eyfjord, J.E., Spector, T.D., Esteller, M., 2013b. DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. Carcinogenesis 34 (1), 102–108. http://dx.doi.org/10.1093/carcin/bgs321.

Hogenson, T.L., 2013. Epigenetics as the underlying mechanism for monozygotic twin discordance. Medical Epigenetics 1, 3–18.

Holliday, R., 1994. Epigenetics: an overview. Developmental Genetics 15 (6), 453–457.

Horvath, S., 2013. DNA methylation age of human tissues and cell types. Genome Biology 14 (10), R115.

Houseman, E.A., Accomando, W.A., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T., 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. MC Bioinformatics 13, 86.

Jaffe, A.E., Irizarry, R.A., 2014. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biology 15 (2), R31.

Jaffe, A.E., Feinberg, A.P., Irizarry, R.A., Leek, J.T., 2012a. Significance analysis and statistical dissection of variably methylated regions. Biostatistics 13 (1), 166–178.

Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P., Irizarry, R.A., 2012b. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology 41 (1), 200–209.

Javierre, B.M., Fernandez, A.F., Richter, J., Al-Shahrour, F., Martin-Subero, J.I., Rodriguez-Ubreva, J., Berdasco, M., Fraga, M.F., O'Hanlon, T.P., Rider, L.G., Jacinto, F.V., Lopez-Longo, F.J., Dopazo, J., Forn, M., Peinado, M.A., Carreño, L., Sawalha, A.H., Harley, J.B., Siebert, R., Esteller, M., Miller, F.W., Ballestar, E., 2010. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Research 20 (2), 170–179.

Jin, M., Zhu, S., Hu, P., Liu, D., Li, Q., Li, Z., Zhang, X., Xie, Y., Chen, X., 2014. Genomic and Epigenomic analyses of monozygotic twins discordant for Congenital Renal Agenesis. American Journal of Kidney Diseases pii:S0272–6386 (14) 00498–3.

Joubert, B.R., Håberg, S.E., Bell, D.A., Nilsen, R.M., Vollset, S.E., Midttun, O., Ueland, P.M., Wu, M.C., Nystad, W., Peddada, S.D., London, S.J., 2014. Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance? Cancer Epidemiology Biomarkers and Prevention. http://dx.doi.org/10.1158/1055-9965.EPI-13-1256.

Kaminsky, Z., Petronis, A., Wang, S.C., Levine, B., Ghaffar, O., Floden, D., Feinstein, A., 2008. Epigenetics of personality traits: an illustrative study of identical twins discordant for risk-taking behavior. Twin Research and Human Genetics: the Official Journal of the International Society for Twin Studies 11 (1), 1–11.

Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H., Wong, A.H., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C., McRae, A.F., Visscher, P.M., Montgomery, G.W., Gottesman, I.I., Martin, N.G., Petronis, A., 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genetics 41 (2), 240–245.

Keller, M.C., Medland, S.E., Duncan, L.E., Hatemi, P.K., Neale, M.C., Maes, H.H., Eaves, L.J., 2009. Modeling extended twin family data I: description of the Cascade model. Twin Research and Human Genetics 12 (1), 8–18.

Kendler, K.S., Heath, A.C., Martin, N.G., Eaves, L.J., 1987. Symptoms of anxiety and symptoms of depression. Same genes, different environments? Archives of General Psychiatry 44 (5), 451–457.

Khavari, D.A., Sen, G.L., Rinn, J.L., October 1, 2010. DNA methylation and epigenetic control of cellular differentiation. Cell Cycle 9 (19), 3880–3883.

Kinoshita, M., Numata, S., Tajima, A., Shimodera, S., Ono, S., Imamura, A., Iga, J., Watanabe, S., Kikuchi, K., Kubo, H., Nakataki, M., Sumitani, S., Imoto, I., Okazaki, Y., Ohmori, T., 2013. DNA methylation signatures of peripheral leukocytes in schizophrenia. Neuromolecular Medicine 15 (1), 95–101.

Klose, R.J., Bird, A.P., 2006. Genomic DNA methylation: the mark and its mediators. Trends in Biochemical Sciences 31 (2), 89–97.

Lervolino, A.C., Rijsdijk, F.V., Cherkas, L., Fullana, M.A., Mataix-Cols, D., 2011. A multivariate twin study of obsessive-compulsive symptom dimensions. Archives of General Psychiatry 68 (6), 637–644.

Letourneau, A., Santoni, F.A., Bonilla, X., Sailani, M.R., Gonzalez, D., Kind, J., Chevalier, C., Thurman, R., Sandstrom, R.S., Hibaoui, Y., Garieri, M., Popadin, K., Falconnet, E., Gagnebin, M., Gehrig, C., Vannier, A., Guipponi, M., Farinelli, L., Robyr, D., Migliavacca, E., Borel, C., Deutsch, S., Feki, A., Stamatoyannopoulos, J.A., Herault, Y., van Steensel, B., Guigo, R., Antonarakis, S.E., 2014. Domains of genome-wide gene expression dysregulation in Down's syndrome. Nature 508 (7496), 345–350.

Li, R., Montpetit, A., Rousseau, M., Wu, S.Y., Greenwood, C.M., Spector, T.D., Pollak, M., Polychronakos, C., Richards, J.B., 2014. Somatic point mutations occurring early in development: a monozygotic twin study. Journal of Medical Genetics 51 (1), 28–34.

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K., 2000. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. New England Journal of Medicine 343 (2), 78–85.

Liu, Y., Li, X., Aryee, M.J., Ekström, T.J., Padyukov, L., Klareskog, L., Vandiver, A., Moore, A.Z., Tanaka, T., Ferrucci, L., Fallin, M.D., Feinberg, A.P., 2014. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. American Journal of Human Genetics 94 (4), 485–495. http://dx.doi.org/10.1016/j.ajhg.2014.02.011.

Lorthongpanich, C., Cheow, L.F., Balu, S., Quake, S.R., Knowles, B.B., Burkholder, W.F., Solter, D., Messerschmidt, D.M., 2013. Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. Science 341 (6150), 1110–1112.

Lu, Q., Quinn, A.M., Patel, M.P., Semus, S.F., Graves, A.P., Bandyopadhyay, D., Pope, A.J., Thrall, S.H., 2012. Perspectives on the discovery of small-molecule modulators for epigenetic processes. Journal of Biomolecular Screening 17 (5), 555–571.

Luo, J., Margolis, K.L., Wactawski-Wende, J., Horn, K., Messina, C., Stefanick, M.L., Tindle, H.A., Tong, E., Rohan, T.E., 2011. Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study. BMJ 342, d1016.

Marsit, C.J., Koestler, D.C., Watson-Smith, D., Boney, C.M., Padbury, J.F., Luks, F., 2013. Developmental genes targeted for epigenetic variation between twin–twin transfusion syndrome children. Clinical Epigenetics 5 (1), 18.

Martino, D., Loke, Y.J., Gordon, L., Ollikainen, M., Cruickshank, M.N., Saffery, R., Craig, J.M., 2013. Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. Genome Biology 14 (5), R42.

Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., Milosavljevic, A., Siegmund, K.D., Bock, C., Irizarry, R.A., 2013. Recommendations for the design and analysis of epigenome-wide association studies. Nature Methods 10 (10), 949–955.

Miyake, K., Yang, C., Minakuchi, Y., Ohori, K., Soutome, M., Hirasawa, T., Kazuki, Y., Adachi, N., Suzuki, S., Itoh, M., Goto, Y.I., Andoh, T., Kurosawa, H., Oshimura, M., Sasaki, M., Toyoda, A., Kubota, T., 2013. Comparison of genomic and epigenomic expression in monozygotic twins discordant for Rett syndrome. PLoS One 8 (6), e66729.

Murr, R., 2010. Interplay between different epigenetic modifications and mechanisms. Advances in Genetics 70, 101–141.

Nawy, T., November 2013. Single-cell epigenetics. Nature Methods 10 (11), 1060.

Ng, J.W.W.Y., Barrett, L.M., Wong, A., Kuh, D., Davey Smith, G., Relton, C.L., 2012. The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. Genome Biology 13, 246.

Nguyen, A., Rauch, T.A., Pfeifer, G.P., Hu, V.W., 2010. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology 24 (8), 3036–3051.

Nitert, M.D., Dayeh, T., Volkov, P., Elgzyri, T., Hall, E., Nilsson, E., Yang, B.T., Lang, S., Parikh, H., Wessman, Y., Weishaupt, H., Attema, J., Abels, M., Wierup, N., Almgren, P., Jansson, P.A., Rönn, T., Hansson, O., Eriksson, K.F., Groop, L., Ling, C., 2012. Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. Diabetes 61 (12), 3322–3332.

Ostankovitch, M., Pyle, A.M., 2013. Noncoding RNAs: a Story of networks and long-distance relationships. Journal of Molecular Biology 425 (19), 3577–3581.

Ozsolak, F., Song, J.S., Liu, X.S., Fisher, D.E., Ozsolak, F., Song, J.S., Liu, X.S., Fisher, D.E., 2007. High-throughput mapping of the chromatin structure of human promoters. Nature Biotechnology 25 (2), 244–248.

Petersen, A.K., Zeilinger, S., Kastenmüller, G., Römisch-Margl, W., Brugger, M., Peters, A., Meisinger, C., Strauch, K., Hengstenberg, C., Pagel, P., Huber, F., Mohney, R.P., Grallert, H., Illig, T., Adamski, J., Waldenberger, M., Gieger, C., Suhre, K., 2014. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Human Molecular Genetics 23 (2), 534–545.

Powledge, T.M., 2011. Behavioral epigenetics: how nurture shapes nature. Bioscience 61 (8), 588–592.

Price, M.E., Cotton, A.M., Lam, L.L., Farré, P., Emberly, E., Brown, C.J., Robinson, W.P., Kobor, M.S., 2013. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin 6 (1), 4.

Rahmioğlu, N., Ahmadi, K.R., 2010. Classical twin design in modern pharmacogenomics studies. Pharmacogenomics 11 (2), 215–226.

Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., Leslie, R.D., Deloukas, P., Spector, T.D., 2010. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Research 20, 434–439.

Rakyan, V.K., Beyan, H., Down, T.A., Hawa, M.I., Maslau, S., Aden, D., Daunay, A., Busato, F., Mein, C.A., Manfras, B., Dias, K.R., Bell, C.G., Tost, J., Boehm, B.O., Beck, S., Leslie, R.D., 2011a. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLOS Genetics 7 (9), e1002300.

Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S., 2011b. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics 12 (8), 529–541.

Relton, C.L., Groom, A., St Pourcain, B., Sayers, A.E., Swan, D.C., Embleton, N.D., Pearce, M.S., Ring, S.M., Northstone, K., Tobias, J.H., Trakalo, J., Ness, A.R., Shaheen, S.O., Davey Smith, G., 2012. DNA methylation patterns in cord blood DNA and body size in childhood. PLoS ONE 7 (3), e31821.

Ribel-Madsen, R., Fraga, M.F., Jacobsen, S., Bork-Jensen, J., Lara, E., Calvanese, V., Fernandez, A.F., Friedrichsen, M., Vind, B.F., Højlund, K., Beck-Nielsen, H., Esteller, M., Vaag, A., Poulsen, P., 2012. Genome-wide analysis of DNA methylation differences in muscle and fat from monozygotic twins discordant for type 2 diabetes. PLoS One 7 (12), e51302.

Rijsdijk, F.V., Sham, P.C., 2002. Analytic approaches to twin data using structural equation models. Briefings in Bioinformatics 3 (2), 119–133.

Rönn, T., Volkov, P., Davegårdh, C., Dayeh, T., Hall, E., Olsson, A.H., Nilsson, E., Tornberg, A., Dekker Nitert, M., Eriksson, K.F., Jones, H.A., Groop, L., Ling, C., 2013. A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. PLOS Genetics 9 (6), e1003572.

Rivera, C., Ren, B., 2013. Mapping human epigenomes. Cell 155 (1), 39–55.

Sarachana, T., Zhou, R., Chen, G., Manji, H.K., Hu, V.W., 2010. Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. Genome Medicine 2 (4), 23.

Scesnaite, A., Jarmalaite, S., Mutanen, P., Anttila, S., Nyberg, F., Benhamou, S., Boffetta, P., Husgafvel-Pursiainen, K., 2012. Similar DNA methylation pattern in lung tumours from smokers and never-smokers with second-hand tobacco smoke exposure. Mutagenesis 27 (4), 423–429.

Schmidt, C.W., 2013. Uncertain inheritance transgenerational effects of environmental exposure. Environmental Health Perspectives 121 (10).

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., Ecker, J.R., 2013. Patterns of population epigenomic diversity. Nature 495 (7440), 193–198.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., Petronis, A., 2006. Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Research 34 (2), 528–542.

Selmi, C., Feghali-Bostwick, C.A., Lleo, A., Lombardi, S.A., De Santis, M., Cavaciocchi, F., Zammataro, L., Mitchell, M.M., Lasalle, J.M., Medsger Jr, T., Gershwin, M.E., 2012. X chromosome gene methylation in peripheral lymphocytes from monozygotic twins discordant for scleroderma. Clinical and Experimental Immunology 169 (3), 253–262.

Shenker, N.S., Ueland, P.M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J.M., Vineis, P., 2013. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology 24 (5), 712–716.

Sindhu, C., Samavarchi-Tehrani, P., Meissner, A., 2012. Transcription factor-mediated epigenetic reprogrammin. Journal of Biological Chemistry 287 (37), 30922–30931.

Slatkin, M., 2009. Epigenetic inheritance and the missing heritability problem. Genetics 182 (3), 845–850.

Stefan, M., Zhang, W., Concepcion, E., Yi, Z., Tomer, Y., 2014. DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. Journal of Autoimmunity 50, 33–37.

Stern, S., Fridmann-Sirkis, Y., Braun, E., Soen, Y., 2012. Epigenetically heritable alteration of fly development in response to toxic challenge. Cell Reports 1 (5), 528–542.

Stewart, C.J., Marrs, E.C.L., Nelson, A., Lanyon, C., Perry, J.D., Embleton, N.D., Cummings, S.P., Berrington, J.E., 2013. Development of the preterm gut microbiome in twins at risk of necrotising enterocolitis and sepsis. PLoS ONE 8 (8), e73465.

Sun, C., Burgner, D.P., Ponsonby, A.L., Saffery, R., Huang, R.C., Vuillermin, P.J., Cheung, M., Craig, J.M., 2013. Effects of early-life environment and epigenetics on cardiovascular disease risk in children: highlighting the role of twin studies. Pediatric Research 73 (4 Pt 2), 523–530.

Szyf, M., 2009. Epigenetics, DNA methylation, and chromatin modifying drugs. Annual Review of Pharmacology and Toxicology 49, 243–263.

Te, J.L., Dozmorov, I.M., Guthridge, J.M., Nguyen, K.L., Cavett, J.W., Kelly, J.A., Bruner, G.R., Harley, J.B., Ojwang, J.O., 2010. Identification of unique microRNA signature associated with lupus nephritis. PLoS ONE 5 (5), e10344.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., Savage, D.A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S.A., Jones, A., Beck, S., Wagner, W., Laird, P.W., Jacobs, I.J., Widschwendter, M., 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Research 20 (4), 440–446.

Tsai, P.C., Spector, T.D., Bell, J.T., 2012. Using epigenome-wide association scans of DNA methylation in age-related complex human traits. Epigenomics 4 (5), 511–526.

Umeton, R., Annibali, V., Vittori, D., Mechelli, R., Romano, S., Policano, C., Coarelli, G., Mattei, C., Ristori, G., Salvetti, M., 2011. MicroRNA expression profile of B lymphocytes in multiple sclerosis. In: Poster at 21st Conference of the Italian Neuroimmunology Society (AINI).

Ushijima, T., Watanabe, N., Okochi, E., Kaneda, A., Sugimura, T., Miyamoto, K., 2003. Fidelity of the methylation pattern and its variation in the genome. Genome Research 13 (5), 868–874.

Van Dongen, J., Slagboom, P.E., Draisma, H.H.M., Martin, N.G., Boomsma, D.I., 2012. The continuing value of twin studies in the omics era. Nature Reviews Genetics 13 (9), 640–653.

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era–concepts and misconceptions. Nature Reviews Genetics 9 (4), 255–266.

Wacholder, S., Rothman, N., Caporaso, N., 2000. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. Journal of the National Cancer Institute 92 (14), 1151–1158.

Waddington, C.H., 2012. The epigenotype. 1942. International Journal of Epidemiology 41 (1), 10–13.

Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A., DeMeo, D.L., 2012. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Human Molecular Genetics 21 (13), 3073–3082.

Wang, S., 2011. Method to detect differentially methylated loci with case-control designs using illumina arrays. Genetic Epidemiology 35 (7), 686–694.

Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D.C., Wiehler, J., Rolf, B., 2014. Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Science International: Genetics 9, 42–46.

Williams, F.M.K., Cherkas, L.F., Spector, T.D., MacGregor, A.J., 2005. The effect of moderate alcohol consumption on bone mineral density: a study of female twins. Annals of the Rheumatic Diseases 64, 309–310.

Witt, S.H., Sommer, W.H., Hansson, A.C., Sticht, C., Rietschel, M., Witt, C.C., 2013. Comparison of gene expression profiles in the blood, hippocampus, and prefrontal cortex of rats. Silico Pharmacology 1, 15.

Wolfson, M., Tacutu, R., Budovsky, A., Aizenberg, N., Fraifeld, V.E., 2008. MicroRNAs: relevance to aging and age-related disease. Open Longevity Science 2, 66–75.

Wong, C.C., Meaburn, E.L., Ronald, A., Price, T.S., Jeffries, A.R., Schalkwyk, L.C., Plomin, R., Mill, J., 2013. Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. Molecular Psychiatry 19 (4), 495–503.

Wong, C.C., Meaburn, E.L., Ronald, A., Price, T.S., Jeffries, A.R., Schalkwyk, L.C., Plomin, R., Mill, J., 2014. Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. Molecular Psychiatry 19 (4), 495–503.

Wu, H.C., Delgado-Cruzata, L., Flom, J.D., Kappil, M., Ferris, J.S., Liao, Y., Santella, R.M., Terry, M.B., 2011. Global methylation profiles in DNA from different blood cell types. Epigenetics 6 (1), 76–85.

Yu, C.C., Furukawa, M., Kobayashi, K., Shikishima, C., Cha, P.C., Sese, J., Sugawara, H., Iwamoto, K., Kato, T., Ando, J., Toda, T., 2012. Genome-wide DNA methylation and gene expression analyses of monozygotic twins discordant for intelligence levels. PLoS One 7 (10), e47081, 11.

Zaine, S., Lund, G., 2012. Integration genomic and epigenomic information: a promising strategy for identifying functional DNA variants of human disease. Clinical Genetics 81 (4), 334–340.

Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., Strauch, K., Waldenberger, M., Illig, T., 2013. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One 8 (5), e63812.

Zentner, G.E., Henikoff, S., 2013. Regulation of nucleosome dynamics by histone modifications. Nature Structural and Molecular Biology 20, 259–266.

Zhang, K., Schrag, M., Darmanto, A., Semotiuk, A., Xu, G., Kirsch, W., 2010. Simultaneous quantification of histone acetylation, methylation and ubiquitination at multiple sites by LC/MS/MS-MRM. FASEB Journal 24, 905.5.

Zhang, B., Zhou, Y., Lin, N., Lowdonn, R.F., Hongn, C., Nagarajann, R.P., Chengn, J.B., Lin, D., Stevens, M., Lee, H.J., Xing, X., Zhou, J., Sundaram, V., Elliott, G., Gu, J., Shi, T., Gascard, P., Sigaroudinia, M., Tlsty, T.D., Kadlecek, T., Weiss, A., O'Geen, H., Farnham, P.J., Maire, C.L., Ligon, K.L., Madden, P.A., Tam, A., Moore, R., Hirst, M., Marra, M.A., Zhang, B., Costello, J.F., Wang, T., 2013. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Research 23 (9), 1522–1540.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., Listgarten, J., 2014. Epigenome-wide association studies without the need for cell-type composition. Nature Methods 11, 309–311.

Zwijnenburg, P.J., Meijers-Heijboer, H., Boomsma, D.I., 2010. Identical but not the same: the value of discordant monozygotic twins in genetic research. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 153B (6), 1134–1149.

Relevant Websites

http://encodeproject.org/ENCODE/ – Encode.
http://www.ensembl.org/info/genome/funcgen/index.html – ENSEMBL regulation.
http://www.geneimprint.com – Gene imprinting.
http://igc.otago.ac.nz/home.html – Gene imprinting.
http://regulome.stanford.edu – RegulomeDB.
http://www.twinsuk.ac.uk – TwinsUK.
http://genome.ucsc.edu/cgi-bin/hgTracks?org=human – UCSC genome browser.

Twin Studies in Demography

Jere R Behrman, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Twins studies are used to describe certain characteristics of populations, estimate determinants of population quantity and quality, and identify impacts of critical components of population quantity and quality on each other and on other outcomes. They share in common use of special properties of twins such as shared genes at conception for monozygotic (MZ) twins or very short birth spacing, but use various approaches with different assumptions for different purposes, in order of their frequency of application: (1) ACE model genetic-environmental variance decompositions; (2) MZ twins fixed-effect estimates with controls for endowments; and (3) investigating the quality–quantity fertility model.

Introduction

Demography traditionally has focused on three factors that affect the population quantity: fertility, mortality, and migration. A broader view of human demography also includes population quality – health status, nutritional status, and education – as well as other behavioral outcomes such as marital status, labor force status, and occupational status. Twins studies have been used to attempt to inform the impacts and determinants of population quantity and quality.

Twins studies have been used to describe certain dimensions of populations, to establish the determinants of population quantity and quality, and to identify impacts of critical components of population quantity, and quality on each other and on other outcomes, such as the labor and marriage market returns to health or education. Twins studies share in common the use of special properties of twins, but use a range of approaches with different assumptions for different purposes, around which this article is organized, in order of their frequency of application: (1) ACE model estimates and variance decompositons (additive genetic effects (A), common environment shared by both twins (C) and unique environments that affect each twin differently (E)); (2) monozygotic (MZ) twins fixed-effects estimates of causal effects; and (3) investigating the quality–quantity fertility model.

The second of these approaches exploits identical (monozygotic) twins to control for identical endowments (genetics, environment) at conception. MZ twins are genetically identical at conception, emerging from a single sperm and egg, from which two separate eggs later emerge. MZ twinning occurs at a relatively constant rate among contexts (Kiely and Kiely, 2001). Irrespective of context, MZ twins are rarer than fraternal (dizygotic, DZ) twins. In most prefertility drug populations, about 1 in 85 births are twins (Plomin et al., 2005), of which about a third are MZ, a third same-sex DZ, and a third opposite-sex DZ (Keith et al., 1995).

The other two approaches noted above use both MZ and DZ twins. DZ twins are ordinary siblings in that they are the product of two different eggs and two different sperms. DZs have individual–specific genetics different from family average genetics to the same degree as do other siblings but tend to experience more similar home and broader environments than do other siblings because they are born into the same life-cycle stage of their parents and same larger macroenvironment. The DZ twinning rate is affected by several factors, including maternal age and fertility drugs, and is therefore subject to change over time, across women, and among populations. Some women tend to produce more than one egg at each menstrual period and, therefore, are more likely to have DZ twins, a tendency that may run across generations. Women over 35 years of age are more likely to produce two eggs. Women who have three or more children are also more likely to have DZs. Artificial induction of ovulation and *in vitro* fertilization-embryo replacement can also give rise to DZ and MZ twins. The first twins study approach noted above exploits differences between MZ and DZ twins to identify the basic constructs of interest. The third approach utilizes primarily only the fact of multiple births, not whether they are MZ or DZ.

There are a few datasets of twins raised apart (e.g., Minnesota Study of Twins Raised Apart (MISTRA), Swedish Adoption/Twin Study on Aging (SATSA)). But most twins data include twins that were raised together. Important US twins datasets, for example, include the National Longitudinal Study of Adolescent Health (Add Health) Twin Data, the Midlife Development in the United States (MIDUS) Study, the Minnesota Twins Registry data, the University of Washington Twins Registry data, the Vietnam Veterans Twin Registry Data, and the National Academy of Science–National Research Council (NAS–NRC) Twin Registry of World War II Veterans. There are extensive register-based twins data existing in Australia, Chile, Denmark, Norway, and Sweden, and a few survey-based twins data sets for developing countries (e.g., China). Because twins raised together share both genetic factors – with the genetic overlap at conception being 50% or more (depending on the extent of assortative mating) for DZ and 100% for MZ twins – and important social and economic contexts during childhood and adolescence, they provide an opportunity to better understand how genetic and social endowments affect a variety of behaviors and outcomes that are of key interest to demographers and other social scientists.

Twins are not a random sample of the population. They differ in their developmental environment in the womb and have inferior endowments at birth – lower (Appearance, Pulse, Grimace, Activity, Respiration) APGAR scores; increased incidence of breathing difficulties; and lower birth weight, which has been shown to significantly affect not only infant survival but also adult outcomes (Behrman and Rosenzweig, 2004; Almond et al., 2005; Black et al., 2005). Rosenzweig and Zhang

(2009) summarize the distributions of birth weights from the US white population for twins and nontwins, based on the Minnesota Twins Registry and the US National Longitudinal Survey 1979, and the birth weight distributions for twins and nontwins born in China based on their survey data. In both countries, the average birth weight of singletons is almost 30% higher than that of twins, approximately a 0.7 kg difference. However with respect to a number of observed characteristics, twins differ very little from singletons. Studies on the personality and intelligence of twins, for example, suggest that they have traits very similar to those of nontwins (e.g., Deary et al., 2006). The second twins method above controls for any unobserved characteristics shared by twins, including being raised with siblings born at the same time. So for this approach arguably the aspects in which twins are different from singletons are controlled.

Recent developments have made possible the incorporation of genetic markers in sociodemographic data sets, which clearly enhance the possibilities of incorporating genetics into demographic analyses. The special issue of the *American Journal of Public Health* edited by Kaplan et al. (2013) includes a number of papers that discuss the implications and the limitations of these developments for demography and the social sciences more broadly. These developments may permit identification of the roles of particular genetic markers in various outcomes and otherwise enhance some dimensions of analyses using twins, probably with particular emphasis on the implications of epigenetics for gene–environment interactions in the first, and perhaps the second twins method. But these developments do not eliminate the value of twins studies because, for example, they permit easy characterization of specific genetic markers but not the entire set of genes shared by MZ twins as used in the first twins method, or the control for average family genetics and other shared dimensions of childhood family and neighborhood background experienced at the same life-cycle stage as in the second method or the use of twining as an increase in child quantity that is the essence of the third method.

ACE Model Estimates and Variance Decomposition

The first approach, to estimate heritability and related variance decompositions with ACE or closely related behavioral genetics models, is the most common use of twins in demography and the other social sciences. To illustrate, the Wikipedia article on 'Twin Study' (http://en.wikipedia.org/wiki/Twin_study; accessed 20.09.13.) only includes this approach among the three that are covered in this article.

These analyses generally have been concerned with decomposing the variance of an observed outcome Y ('phenotype') such as fertility or health into the contributions of additive genetic effects (A), common environment shared by both twins (C), and unique environments that affect each twin differently (E) – hence the acronym 'ACE'. The simplest univariate behavioral genetics approach assumes additive genetic models with no assortative mating and with equal environmental influences across MZs and DZs:

$$Y_{ij} = aA_{ij} + cC_{ij} + eE_{ij} \quad [1]$$

where subscripts refer to the *ith* child in the *jth* family in general and to the *ith* twin in the *jth* twinship in twins studies. a measures the extent to which Y_{ij} is affected by genetic endowments A_{ij}, c measures the extent to which Y_{ij} is affected by common environments C_{ij}, and e measures the extent to which Y_{ij} is affected by unique environments E_{ij}. MZ twins are assumed to share 100% of their genes (A) and common environment (C) so that any differences between members of a MZ twinship in this model are due to unique environments (E). The correlation between outcomes for MZ twins (r_{MZ}) is an estimate of $A + C$. DZ twins, like ordinary siblings, are assumed also to share the same C and on average 50% of their genes if there is no assortative mating. In that case the correlation between outcomes for DZ twins (r_{DZ}) is $0.5A + C$. Therefore the additive genetic effect A is twice the difference between the MZ and DZ correlations $A = 2(r_{MZ} - r_{DZ})$ and heritability (h^2) in the simplest univariate ACE model is defined to be the variance in aA relative to the variance in Y:

$$h^2 = a^2\frac{\mathrm{Var}(A)}{\mathrm{Var}(Y)} \quad [2]$$

More recent multivariate behavioral genetics analyses investigate the contributions of genes and the environment to the correlations among multiple outcomes (Plomin et al., 2005). For example, the simplest bivariate ACE model is an extension to two outcomes, say schooling (S) and fertility (F), and allows for cross effects so that, for instance, the determinants of schooling (A^S, C^S, E^S) also affect fertility in addition to the factors that only affect fertility (A^F, C^F, E^F):

$$S_{ij} = a_{SS}A^S_{ij} + c_{SS}C^S_{ij} + e_{SS}E^S_{ij} \quad [3]$$

$$F_{ij} = a_{SF}A^S_{ij} + c_{SF}C^S_{ij} + e_{SF}E^S_{ij} + a_{FF}A^F_{ij} + c_{FF}C^F_{ij} + e_{FF}E^F_{ij} \quad [4]$$

The first three terms in eqn [4] indicate contributions of the three determinants of S (A^S, C^S, E^S) to F and the last three terms indicate contributions of the three factors (A^F, C^F, E^F) that only affect F but not S. Estimation of this bivariate ACE model permits estimation of heritabilities of S and F and the extent to which ACE factors determining S in relation in eqn [3] also determine F in relation in eqn [4]. Estimation of this model, however, does not identify whether the ACE factors determining S determine F indirectly through S or directly. Nevertheless some limited inferences about the possible impact of S on F might be made. In the extreme case in which $a_{SF} = c_{SF} = e_{SF} = 0$, there could not be any impact of S on F. On the other hand, if all the impacts of (A^S, C^S, E^S) on F are through S, $a_{SF}a_{ss} = c_{SF}c_{ss} = e_{SF}e_{ss}$, which can be tested and whether or not satisfied used to establish the maximum possible impact of S on F.

There are many estimates of heritabilities of demographic outcomes, particularly pertaining to the health component of quality. For example related to fertility, Kohler and Rodgers (2003) analyze data on Danish twins born between 1953 and 1970 and conclude that a bivariate behavioral genetics analysis confirms that fertility in this low-fertility setting is subject to important genetic influences and (perhaps somewhat unexpected) genetic variance in fertility is not necessarily shared with genetic variance in completed schooling attainment. Instead, their results show that most genetic variance in fertility is

residual variance that affects fertility but not schooling attainment. Overlapping influences mainly exist for shared environmental factors analyses of females, for whom all shared environmental factors affecting fertility also affect schooling.

A number of questions have been raised or limitations noted about the simple ACE models and, in some cases, extensions developed or empirical evidence presented to attempt to speak to these questions or limitations. Some examples include:

- The assumption of equal common environments for MZ and DZ twins has been questioned and if environments are not common (e.g., $C^{DZ} < C^{MZ}$), then $A = 2(r_{MZ} - r_{DZ})$ above becomes $A + 2(C^{MZ} - C^{DZ}) = 2(r_{MZ} - r_{DZ})$ so if MZ's share more common environments than DZs, in the standard calculation the differences in shared environments are attributed to genetic differences.
- The assumption of no assortative mating seems violated by many studies of assortative mating particularly on schooling and assuming no assortative mating when there is positive assortative mating overstates the genetic contribution (i.e., $A = (s^{-1})(r_{MZ} - r_{DZ})$ where s is the proportion of shared genes among DZs so if s is assumed to be 0.5 when s actually is greater than 0.5 due to assortative mating, A is overestimated); if DZ twins share more than 50% of their genes owing to assortative mating, however, the correlation between the A components across DZs can be increased accordingly by using the actual value of s.
- Epigenetics implies that the expression of genes may differ even for MZ twins; this implies that typical estimates of heritability are lower-bound estimates for heritability at conception because in the simplest ACE model the unique environment E proxies for the part of A altered by epigenetics.
- There may be nonadditive genetics effects (often denoted D for dominance); these can be examined in an ADE model.
- The standard ACE twin design does not permit considering both C and nonadditive genetic effects simultaneously; this limitation can be addressed by expanding the design to include additional siblings.
- Gene–environment correlations are not detectable as distinct effects; incorporating adoption models, or children-of-twins designs may assess family influences uncorrelated with shared genetic effects.
- The implications of standard models of investments in children that endowments of all children in families affect investments in any particular child are ignored (see the second twins study approach below).

As noted, some of these limitations can be dealt with by extending the ACE model or bringing additional information to the analysis. However even if these limitations are dealt with, a fundamental question is why is heritability, a major concern of the ACE literature, of much interest. To illustrate this point, consider **Figure 1** (also see Behrman et al., 1980; Feldman and Lewontin, 1975; Layzer, 1974). In this figure, reaction functions for some outcome (phenotype, such as fertility F) are plotted against the environment (N, combining both C and E). Suppose that genes come only in two forms (1, 2) and reactions of F to changes in N differ for the two gene types as in the figure. Let the dashed line be the average reaction function for the population. The variance of F and the proportion of this variance due to genetics are greater for given environmental changes, the greater the vertical distance between the reaction functions for genetic types 1 and 2 over the range of the change in N. The relative contribution of a small change around N_0, for example, is much smaller than the contribution of an equal change around N_1. Therefore heritability estimates are much different around N_0 than around N_1 even though the reaction functions have the same slopes (and thus indicate the same impacts of changing N on F) around N_0 as around N_1. Within this framework, therefore, estimates of heritability are not a good guide to what happens to F if N is changed. To explore what happens to F if N is changed, the reaction function slopes are needed, not the variance decomposition for F. Obtaining such estimates for changes in such factors as schooling attainment is the focus of the second twins method, discussed next. Also see Kohler et al. (2011) for a recent effort to integrate such concerns with the bivariate ACE framework.

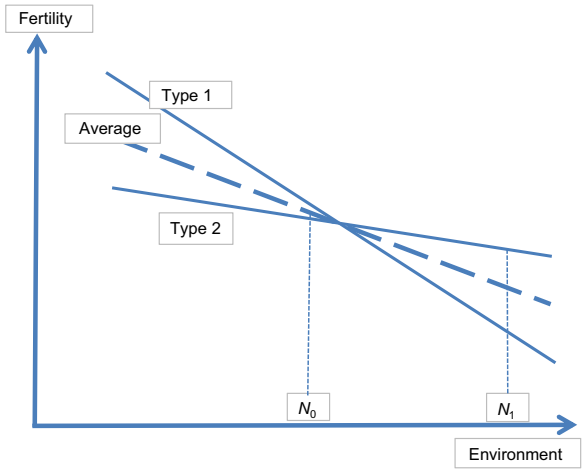

Figure 1 Fertility reaction functions for two genetic types, 1 and 2, and average reaction function all versus the environment.

MZ Fixed-Effects Estimates of Impacts of Demographic Variables or of Other Variables on Demographic Outcomes

The second approach, using MZ twins to control for genetic and other background unobserved confounding factors, is the second most common use of twins for population studies. Social scientists long have used sibling comparisons for this purpose, reasoning that if brothers/sisters are similar with respect to family background and other characteristics, using differences between them in, for example, schooling controls many relevant confounding factors (Griliches and Mason, 1972 is an early study). MZ and DZ twins are more attractive than other siblings in at least one important respect because, as noted above, they share birth dates and, therefore, differences between them are not confounded by family life-cycle differences. Moreover, within-MZ twins estimates also control for genes at conception (for early studies, see Behrman and Taubman, 1976; Behrman et al., 1980).

Twins fixed-effect studies estimate causal effects of one (or more) variable (e.g., own schooling, birth weight, parental schooling, fertility), which may be partly determined directly by unobserved endowments, on other variables (e.g., fertility, mortality, schooling, marital status, spouse characteristics, health-related behaviors and outcomes, social interactions, wages) that are themselves partly determined directly by endowments. Consider the following linear representation of a reduced-form equation relating adult mortality M_{ij} for the *ith* member of paternal family j to his or her schooling S_{ij} and to three sets of unobserved variables that for simplicity are presented as scalars (but which may be vectors) representing (1) endowments f_j that are common among all children of the paternal family j (e.g., exogenous features of paternal family environments in childhood, including family income, parents' human capital, average genetic endowments among siblings, local schooling, and health-related options), (2) the component of endowments that is specific to child i in family j, represented by a_{ij} (e.g., individual-specific deviations from average family genetic endowments), and (3) a random mortality shock that is specific to i in j, inclusive of measurement errors in mortality, represented by v_{ij}:

$$M_{ij} = \beta_S S_{ij} + f_j + a_{ij} + v_{ij} \qquad [5]$$

where β_S is the effect of schooling and the three unobserved variables (f_j, a_{ij}, v_{ij}) are defined in units so that their coefficients are all one. S_{ij} is itself a function of variables that pertain to the paternal family and to the individual children in the paternal family:

$$S_{ij} = \alpha_f f_j + \alpha_a a_{ij} + \alpha_S a_{kj} + u_{ij} \qquad [6]$$

where α_f is the effect of the family-specific endowment f_j on schooling investment in child i, α_a is the effect of the individual-specific endowment a_{ij} of child i on schooling investment in that child, α_s is the effect of the individual-specific endowment a_{kj} of sib k (which is a vector if there are multiple siblings) on schooling investment in child i, and u_{ij} is a disturbance that affects S_{ij} but not M_{ij} except indirectly through S_{ij} (this is a critical assumption discussed below). The stochastic term, for example, might reflect chance events of one twin encountering more inspiring role models that affect schooling attainment but not mortality except through schooling attainment than the other twin.

Note that these relations are related to, but not identical to, the expressions above for the ACE model in the first approach. First, relation in eqn [5] differs from the ACE relations because another outcome S is included among the right-side variables, which indicates a direct effect of one outcome on another that is of central interest in the MZ twins fixed-effects approach. Second, both relations in eqns [5] and [6] differ from the ACE relations in that (1) the family endowments f_j include both the average family (or twinship) component of A and C because there is no need to distinguish between them in this approach and (2) the individual-specific endowment a_{ij}, the individual-specific component of A, is distinguished from the average family component of A. Third, the relation in eqn [6] differs from the ACE relations because the individual-specific component of the other twin a_{kj} (or, more generally, of all other siblings) is included because within family models of investment in children the endowments of all siblings in the family affect the investments in any particular child and parents may allocate resources to their children so as to reinforce or compensate for endowment differentials (Becker and Tomes, 1976; Behrman, et al., 1982, 1995). (Also see discussion of some results below in this section.) Note that relations in eqns [5] and [6] are specified such that the other twin's endowment a_{kj} affects directly S_{ij} but not M_{ij} except indirectly through S_{ij}, a simplifying assumption that can be dropped with no loss in generality.

As is well known, the parameter of central interest β_S is not identified in eqn [5] if α_a or α_f is not zero. β_S is estimated with bias if eqn [5] is estimated across individuals with different values of f_j and a_{ij}. The regression coefficient for an ordinary least squares (OLS) estimate of relation in eqn [5] is $cov(M_{ij}, S_{ij})/\sigma^2(S_{ij}) = [\beta_S(\sigma^2(S_{ij}) + cov(S_{ij}, f_j) + cov(S_{ij}, a_{ij}))]/\sigma^2(S_{ij})$, which is a biased estimate of β_S unless $cov(S_{ij}, f_j) + cov(S_{ij}, a_{ij}) = 0$. Thus, generally, cross-sectional estimates of associations between S and M are biased estimates of the causal impacts of S on M because S is partially proxying for genetic, family background, childhood neighborhood and other endowments.

With no further assumptions, it is clear that β_S, is not identified even if sibling data, including DZ twins, are used to control in the estimation of β_S for common components of endowments f_j because of the specific components of endowments a_{ij}. As long as families or individuals respond to individual-specific differences in endowments, and such differences are important, then sibling estimators do not provide unbiased estimates. Therefore researchers have used MZ twins, between whom there are as minimal as possible endowment differences at conception, to identify β_S in estimates of models such as relations in eqns [5] and [6], which can be rewritten for MZ twins as:

$$M_{ij}^M = \beta_S S_{ij}^M + f_j^M + v_{ij}^M \qquad [7]$$

$$S_{ij}^M = \alpha_f f_j^M + u_{ij}^M \qquad [8]$$

where superscripts M refer to MZ twins. Relations parallel to eqns [7] and [8] can be written for the *kth* MZ twin in the *jth* family. Fixed-effects MZ twin estimators control for all right-side variables in these relations that are common to both members of MZ twinships (including all characteristics of nontwin siblings and whatever is distinctive about being MZ twins). What is and is not controlled in such relations can be seen in equivalent within-MZ twin relations in which all the factors common to the twins are eliminated. Within-MZ-twin estimators are obtained by subtracting relations in eqns [7] and [8] for the *kth* MZ twin in the *jth* family from relations in eqns [7] and [8] for the *ith* MZ twin in the same family. With a within-MZ-twin estimator (or MZ fixed effects), all of the unobserved endowment components in eqns [7] and [8] are controlled so that consistent estimates of β_S are obtained under the maintained assumption that $cov(u, v)$ is zero:

$$M_{ij}^M - M_{kj}^M = \beta_S\left(S_{ij}^M - S_{kj}^M\right) + v_{ij}^M - v_{ij}^M \qquad [9]$$

Thus, MZ fixed-effects estimators identify true reduced-form impacts of S on M under assumptions noted above. Comparisons can be made with estimates of relation in eqn [5] for the same M outcome to learn to what extent estimates of β_S, the impact of S on M, are biased in cross-sectional estimates that fail to control for unobserved endowments f_j and a_{ij}. Comparisons also can be made between MZ

fixed-effects estimates for females and males, among racial and ethnic groups, across birth cohorts, across levels of socioeconomic status (SES), and across countries. Comparisons can also be made between MZ fixed-effects and DZ or other sibling fixed-effects estimators to see if unobserved individual-specific endowments a_{ij} are important so that within-sibling estimates that control only for common family endowments f_j are misleading. Finally, comparisons can be made between DZ fixed-effects and ordinary sibling fixed-effects estimators controlling for birth spacing to investigate the impact of changes in the timing of births and birth order in the estimated impacts.

A major strength of MZ twins fixed-effects estimates, in comparison with ACE estimates (also see Kohler et al., 2011 for an effort to integrate the two approaches), is that under their maintained assumptions they permit estimation of the causal impact of one behavioral variable such as schooling attainment (an indicator of population quality) on other behavioral variables related to population quantity or quality, such as health, mortality, health-related behaviors, fertility, migration, and investment in children. Such estimates often are of considerable interest in terms of prediction or policy assessments. Moreover MZ twins fixed-effects estimates do not require assumptions about MZ and DZ twins having equally shared environments C or the absence of genetic–environmental interactions. It should be noted that instrumental variables (IV) or two-stage least squares (2SLS) are alternative approaches for eliminating bias in estimated impacts of schooling or other variables, but the indentifying instruments used to date for such estimates such as changes in minimum schooling requirements or ages for leaving schooling have the limitation of being local average treatment estimates (LATE) relevant only for those affected by such changes and not for those over the whole distribution of schooling. Differences in schooling for MZ twins, in contrast, tend to occur for many different schooling levels, not just at the minimum required schooling levels or ages, and thus tend to be more like average treatment effects (ATE) estimators (Amin and Behrman, 2014; Behrman et al., 2011; Lundborg, 2013).

MZ twins fixed-effects estimates have been utilized to investigate the impact of schooling on many demographic outcomes – fertility, mortality, own health/disease conditions, own hospitalization, children's schooling and health and nutritional status, spouse's schooling and health conditions – as well as returns to human capital investments in schooling and health in labor and marriage markets (examples include Amin and Behrman, 2014; Amin et al., 2013; Ashenfelter and Krueger, 1994; Ashenfelter and Rouse, 1998; Behrman et al., 1980, 1994; Behrman and Rosenzweig, 1999, 2002, 2004; Behrman and Taubman, 1976; Conley et al., 2006; Li et al., 2008; Lundborg, 2013; Rosenzweig and Zhang, 2009). A common, but not universal, result is that MZ twins fixed-effects estimated impacts of schooling are considerably smaller and often insignificant than OLS estimates, consistent with relations in eqns [5] and [6] and with upward biases in OLS estimates because in OLS estimates S proxies in part for unobserved endowments in relation in eqn [5]. An interesting example is that the coefficient estimate for maternal schooling attainment in relations with the dependent variable being child schooling is significantly positive in OLS estimates but significantly negative in MZ fixed-effects estimates using the Minnesota Twins Registry data for twins born between 1936 and 1955, consistent with the possibility that increasing S and holding endowments constant increases mothers' labor force participation and reduces their time with their children, which in turn reduces the children's schooling attainment (Behrman and Rosenzweig, 2002).

Though the majority of this literature focuses on estimating schooling impacts on various demographic outcomes, there are some studies that use MZ fixed-effects estimates to investigate other types of relations. One example is the impacts of birth weight over the life cycle, in which case MZ fixed-effects estimates of causal effects are larger for schooling and earnings but smaller for adult height and BMI and birth weight for the next generation than OLS estimates, consistent with there being multiple dimensions of unobserved endowments that are controlled in the MZ estimates and negative correlations between underlying intellectual endowments and physical endowments (Behrman and Rosenzweig, 2004). A second example is the estimate of the impacts of fertility and partnership on happiness, in which case controlling for unobserved endowments substantially affects a number of the estimated effects, suggesting that standard analyses of contributions of partnerships and children to well-being are subject to biases, often substantial, both upward and downward (Kohler et al., 2005).

MZ twins fixed-effects approaches also have been used together with relations in eqns [5] and [6] above for DZ twins to explore whether families reinforce or compensate for child endowment differences. Behrman et al. (1994) show how comparisons between the within-twin correlations of human capital outcomes across MZ and DZ twins can be used to identify the individual-specific endowments and the responsiveness of schooling to individual-specific endowments in the family and in the marriage market. The basic idea is to obtain the 'correct' estimates of the impact of schooling on earnings outcomes using relation in eqn [8] with earnings as the outcome and then, conditional on that estimate of β_S, use relations in eqns [5] and [6] for DZ twins to estimate individual-specific endowments a_{ij} and a_{kj} and how schooling investments respond to those endowments. Estimates from two US twins samples indicate that 27 (42)% of the variance in log earnings (obesity) is due to variability in individual-specific endowments (which suggests that they should not be ignored as in DZ or sibling fixed-effects analyses), allocations of schooling reinforce rather than compensate for specific endowments so the family acts to increase inequality among its children (also reported in an earlier twins study by Behrman et al., 1982), and individual-specific earnings endowments of men and their wives' schooling are negatively associated.

Major criticisms of the MZ twins fixed-effects approach and brief reactions to these criticisms include the following four points (see Kohler et al., 2011, for more extensive and more formal discussions of the last three of these):

1. Twins are not representative of the larger population: As noted in the introduction, the MZ twins fixed-effects approach, by controlling for unobserved fixed effects common to the twins, arguably controls for respects in which such twins differ from the overall population.

2. The stochastic terms in relations in eqns [5] and [6] are likely to be correlated: If between-twin differences in S reflect unobserved factors that also directly determine M (or whatever the dependent variable is in relation in eqn [5]) so that $cov(u_{ij}, v_{ij}) \neq 0$, the estimated schooling-mortality impact is still biased in the MZ twins fixed-effect estimator (perhaps the most visible critique along these lines is Bound and Solon, 1999). The sign of this bias is determined by the sign of the correlation. For example, if more favorable *in utero* environments due to proximity to the placenta increase both schooling and health beyond any effect through schooling as suggested by the results in Behrman and Rosenzweig (2004), then the estimate of β_S from eqn [6] is an overestimate of the true value of β_S. The same holds if an accident or illness limits schooling and has persistent negative effects on later outcomes such as increasing age-specific mortality rates. If $cov(u_{ij}, v_{ij}) > 0$ as often seems to be assumed when the independence between the stochastic terms in eqns [5] and [6] is questioned, thus, the MZ twins fixed-effect estimator, all else equal, gives an upper-bound estimate of β_S in which case that the MZ twins fixed-effect estimates of schooling impacts tend often to be smaller in absolute magnitude than the OLS estimates as noted above would seem to be more compeling evidence that the OLS estimates are biased upward in absolute magnitudes. On the other hand, If $cov(u_{ij}, v_{ij}) < 0$ (e.g., the persistent effect of an illness during school ages also, beyond any effects through S, increases life expectancy through directly reducing participating in risky activities), then the MZ twins fixed-effect estimator, all else equal, gives lower-bound estimates of β_S. One possible empirical approach is to include additional variables that might have persistent effects on both schooling and the outcome of interest, such as measures of cognitive ability (Behrman et al., 1980) or birth weight. In these two cases, the estimates of interest are not changed much by including these additional controls, but other applications could reveal different results.
3. Twins react to each other: If there is imitation with respect to the right-side determinant, such as S in relation in eqn [5], there is no bias in either direction for the MZ fixed-effects twin estimator. But if there is imitation with regard to the dependent variable, such as M in relation in eqn [5], the estimated β_S is a lower bound if there is positive imitation and upper bound if there is negative imitation. Of course for some population outcomes such as fertility imitation of outcomes would seem more likely than for others such as mortality. If there is positive imitation on the outcome, the maximum downward bias is 50%, but the actual bias is likely to be considerably less because the maximum is for the unlikely situation in which the other twin's outcome (e.g., fertility) is weighted as much as the own direct determinants of one's own outcome.
4. Measurement error: Fixed-effects estimates in general, including for MZ twins, filter out much of the true signal of S (or whatever the right-side variable is in relation in eqn [5]) without also reducing measurement error, so there is a tendency for more measurement error bias toward zero than in simple cross-sectional estimates. Twin studies that have reports from other respondents (i.e., the other member of a twin pair, the twins' adult children), so that they can estimate measurement error models, report estimated noise-to-signal ratios of 0.04–0.12 (Ashenfelter and Krueger, 1994; Behrman et al., 1994; Ashenfelter and Rouse, 1998; Amin et al., 2013). Therefore a noise-to-signal ratio of about 0.08 is suggestive for the noise-to-signal estimates from these studies, which implies attenuation biases of about 8% for individual estimates and 26% for within-MZ estimates, a fairly substantial effect in the latter. Behrman et al. (1980) observed that estimates of noise-to-signal ratios from other studies could account for up to half of the difference between their fixed-effects estimates and OLS estimates. Ashenfelter and Krueger (1994) and Behrman et al. (1994) introduced use of another report on the twin's schooling to instrument schooling and therefore eliminate the measurement error bias under the assumption that the measurement error in the other report is independent of the measurement error of one's own report. Both studies find that this method for controlling for measurement error increases the estimated returns to schooling in comparison with estimates that do not correct for measurement error. Many, but not all, subsequent studies (including many of those summarized above) use reports from other respondents to control for measurement error.

Quantitiy–Quality Fertility Model

The third approach – but probably less common than the first two approaches – to using twins for population analysis is to use unanticipated multiple births (usually twins) as exogenous increases in child quantity to explore the quantity–quality (Q–Q) fertility model (Becker and Lewis, 1973; Willis, 1973). Rosenzweig and Wolpin (1980a,b) first proposed and implemented this approach, which basically is to estimate effects of an extra birth at parity P due to twins on child quality at previous parities for all women who had at least P births. The incidence of twins at P is used as an instrument to predict the total number of children born in a IV or 2SLS estimation. Early studies found evidence that sibhip size and children's schooling are negatively related (Rosenzweig and Wolpin, 1980b; Hanushek, 1992). But some recent studies (Caceres, 2004; Black et al., 2005; Angrist et al., 2010; Qian, 2009; Li et al., 2008) using twin births to assess the impact of an exogenous increase in child quantity on child quality of nontwins in the same families have found little or no evidence of a Q–Q trade-off with the exception of rural (but not urban) areas of China in the last of these.

Rosenzweig and Zhang (2009) note that these studies do not really examine impacts of total sibship size on children of parity N (or higher) because impacts on the quality of the twins themselves are ignored, perhaps because twins are not comparable to multiple singleton births for two reasons noted in the introduction: (1) twins are more closely spaced and (2) twins have inferior birth endowments. They argue that a key issue for assessing whether existing evidence from twin studies rejects Q–Q trade-offs is thus whether these two features affect investments in nontwins within families. If closer spacing of twins affects relative costs of investing in them vs singleton children and endowments of twins are lower than those of singleton births, then twinning not only may lower average resources per child

due to increases in sibship size, but also may lead to reallocation of resources differentially across children of different birth orders and thus affect inferences about Q–Q trade-offs based on estimates of the effects of twinning on nontwins siblings.

Rosenzweig and Zhang (2009) show that estimates of higher-parity twinning on older nontwin siblings may underestimate Q–Q trade-offs and thus the weak findings in the recent literature may be consistent with the Q–Q model. Given reinforcing behavior, estimates of the effects of twinning on twins and their older nontwin siblings bound the Q–Q trade-offs as long as any economies of scale from twinning are not too large, with estimated effects of twinning on older nontwins providing lower bounds. They also explore additional considerations not incorporated in simple Q–Q models such as public goods and birth spacing. They use the modified Q–Q model with survey data on families with twins and with singletons in China, where the 'one-child' policy that was strictly enforced in urban areas meant that twinning on the first birth resulted in one exogenous extra birth (and likewise for parity two in rural areas where two children were permitted). They find that twin quality is lower than for singleton siblings at N, and that declines in quality for parity-P births are significant so that the basic prediction of the Q–Q model that child quantity increases reduce average quality is not rejected.

See also: Aging and Health in Old Age; Assortative Mating in the Marriage Market; Biomarkers in Demographic Research; Childhood and Adolescence: Developmental Assets; Children, Value of; Early Life Influences on Health and Mortality in Adulthood; Families and Households, Formal Demography of; Family Theory: Competing Perspectives in Social Demography; Family Theory: Economics of Childbearing; Family Theory: Economics of Intergenerational Relations; Family Theory: Economics of Marriage and Divorce; Fertility Theory; Gene–Environment Interplay; Human Development, Theories of; Mortality Differentials: Selection and Causation; Mortality, Biodemography of; Sociobiology and Sociology: A New Synthesis.

Bibliography

Almond, D., Chay, K.Y., Lee, D.S., 2005. The costs of low birth weight. Quarterly Journal of Economics 120, 1031–1083.

Amin, V., Behrman, J.R., Spector, T.D., 2013. Does more schooling improve health outcomes and health related behaviors? Evidence from UK twins. Economics of Education Review 35, 134–148.

Amin, V., Behrman, J.R., 2014. Do more schooled women have fewer children and delay childbearing? Evidence from a sample of U.S. twins. Journal of Population Economics 27, 1–31.

Angrist, J., Lavy, V., Schlosser, A., 2010. Multiple experiments for the causal link between the quantity and quality of children. Journal of Labor Economics 28, 773–824.

Ashenfelter, O., Krueger, A., 1994. Estimates of the economic return to schooling from a new sample of twins. American Economic Review 84, 1157–1173.

Ashenfelter, O., Rouse, C., 1998. Income, schooling and ability: evidence from a new sample of identical twins. Quarterly Journal of Economics 113, 153–284.

Becker, G.S., Lewis, H.B., 1973. On the interaction between the quantity and quality of children. Journal of Political Economy 81, S279–S288.

Becker, G.S., Tomes, N., 1976. Child endowments and the quantity and quality of children. Journal of Political Economy 84, S143–S162.

Behrman, J.R., Hrubec, Z., Taubman, P., Wales, T.J., 1980. Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment, and Schooling. North-Holland Publishing Company, Amsterdam.

Behrman, J.R., Kohler, H.-P., Jensen, V., Pedersen, D., Petersen, I., Bingley, P., Christensen, K., 2011. Does more schooling reduce hospitalization and delay mortality? New evidence based on Danish twins. Demography 48, 1347–1375.

Behrman, J.R., Pollak, R.A., Taubman, P., 1995. The wealth model: efficiency in education and distribution in the family. In: Behrman, J.R., Pollak, R.A., Taubman, P. (Eds.), From Parent to Child: Intrahousehold Allocations and Intergenerational Relations in the United States. University of Chicago Press, Chicago, pp. 113–138.

Behrman, J.R., Pollak, R.A., Taubman, P., 1982. Parental preferences and provision for progeny. Journal of Political Economy 90, 52–73.

Behrman, J.R., Rosenzweig, M.R., 1999. "Ability" biases in schooling returns and twins: a test and new estimates. Economics of Education Review 18, 159–167.

Behrman, J.R., Rosenzweig, M.R., 2002. Does increasing women's schooling raise the schooling of the next generation? American Economic Review 92, 323–334.

Behrman, J.R., Rosenzweig, M.R., 2004. Returns to birthweight. Review of Economic Statistics 86, 586–601.

Behrman, J.R., Rosenzweig, M.R., Taubman, P., 1994. Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment. Journal of Political Economy 102, 1131–1173.

Behrman, J.R., Taubman, P., 1976. Intergenerational transmission of income and wealth. American Economic Review 66, 436–440.

Black, S.E., Devereux, P.J., Salvanes, K.G., 2005. The more the merrier? the effect of family size and birth order on children's education. Quarterly Journal of Economics 120, 669–700.

Bound, J., Solon, G., 1999. Double trouble: on the value of twins-based estimation of the return to schooling. Economics of Education Review 18, 169–182.

Caceres, J., 2004. Impact of Family Size on Investment in Child Quality: Multiple Births as a Natural Experiment. Working Paper. University of Maryland.

Conley, D., Strully, K., Bennett, N.G., 2006. Twin differences in birth weight: the effects of genotype and parental environment on neonatal and post-neonatal mortality. Economics and Human Biology 4, 151–183.

Deary, I.J., Spinath, F.M., Bates, T.C., 2006. Genetics of intelligence. European Journal of Human Genetics 14, 690–700. http://dx.doi.org/10.1038/Sj.Ejhg.5201588. PMID 16721405.

Feldman, M.W., Lewontin, R.C., 1975. The heritability hang-up. Science 190, 1163–1168.

Griliches, Z., Mason, W.M., 1972. Education, income, and ability. Journal of Political Economy 80, 74–103.

Hanushek, E.A., 1992. The trade-off between child quantity and quality. Journal of Political Economy 100, 84–117.

Kaplan, R.M., Spittel, M.L., Spotts, E.L., 2013. Advancing scientific inquiry by blurring research boundaries. American Journal of Public Health 103, S4.

Keith, L.G., Papiernik, E., Keith, D.M., Luke, B. (Eds.), 1995. Multiple Pregnancy: Epidemiology, Gestation, and Perinatal Outcome. Parthenon, New York.

Kiely, J.L., Kiely, M., 2001. Epidemiological trends in multiple births in the United States, 1971–1998. Twin Research 4, 131–133.

Kohler, H.P., Behrman, J.R., Schnittker, J., 2011. Social science methods for twins data: integrating causality, endowments, and heritability. Biodemography and Social Biology 57, 88–141.

Kohler, H.P., Behrman, J.R., Skytthe, A., 2005. Partner + children = happiness? The effects of partnerships and fertility on well-being. Population and Development Review 31, 407–445.

Kohler, H.P., Rodgers, J.L., 2003. Education, fertility and heritability: explaining a paradox. In: Wachter, K.W., Bulatao, R.A. (Eds.), Offspring: Human Fertility Behavior in Biodemographic Perspective. The National Academies Press, Washington, DC, pp. 46–90. http://books.google.com/books?id=poXXt7ta73cC&pg=PA46.

Layzer, D., 1974. Heritability analyses of IQ scores: science or numerology? Science 183, 1259–1266.

Li, H., Zhang, J., Zhu, Y., 2008. The quantity–quality trade-off of children in a developing country: identification using Chinese twins. Demography 45, 223–243.

Lundborg, P., 2013. The health returns to schooling – what can we learn from twins? Journal of Population Economics 26, 673–701.

Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2005. Behavioral Genetics, fourth ed. Worth Publishers, New York.

Qian, N., 2009. Quantity–Quality and the One Child Policy: The Only-child Disadvantage in School Enrollment in Rural China. National Bureau of Economic Research, Cambridge.

Rosenzweig, M.R., Wolpin, K.I., 1980a. Life-cycle labor supply and fertility: causal inferences from household models. Journal of Population Economics 88, 328–348.

Rosenzweig, M.R., Wolpin, K.I., 1980b. Testing the quantity–quality fertility model: the use of twins as a natural experiment. Econometrica 48, 227–240.

Rosenzweig, M.R., Zhang, J., 2009. Do population control policies induce more human capital investment? Twins, birth weight and China's one-child policy. The Review of Economic Studies 76, 1149–1174.

Willis, R.J., 1973. A new approach to the economic theory of fertility behavior. Journal of Population Economics 81, S14–S64.

Tyranny

Stephen D Reicher, School of Psychology and Neuroscience, University of St Andrews, St Andrews, UK
S Alexander Haslam, School of Psychology, The University of Queensland, St Lucia, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews the traditional social psychology of tyranny. We show how this is rooted in the assumption that group psychology per se, and more specifically, immersion in cohesive groups, is at the root of both submission to authority within groups (as exemplified by Janis's work on groupthink) and domination over members of other groups (as exemplified in Zimbardo's analysis of his Stanford Prison Experiment). We also show that the evidence does not support this assumption. Cohesive groups do not exclude intragroup criticism and dissent. Moreover, group cohesion is necessary as much to challenge as to impose intergroup domination. We then provide evidence from the BBC Prison Study to provide an alternative analysis of the relationship between group process and tyranny. On the one hand, shared social identity and group coordination empower group members to express group norms and to contest the imposition of out-group norms upon them. On the other hand, the failure of groups leads people to accept the imposition of tyrannical leadership and extreme forms of action. Rather than urging people to avoid groups and the exercise of group power, we conclude that people need to embrace groups and exercise group power in defense of democratic and humane values.

Introduction

Tyranny is the absent presence that haunts and frames social psychology – and indeed the whole of social science – and has done since the end of the World War II (see Waller, 2010). On the one hand, if you search for the term 'tyranny' in the index of a social psychology textbook, it is almost certain you will not find it (of the 24 such textbooks on our shelves, it appears in none). On the other hand, the topics which we study are shaped by our experience of the Nazi dictatorship and, above all, by the Nazi Holocaust. In the words of Gershon Sholem, we are still reacting to "the great historical and social question … how could this happen" (cited in the frontispiece to Baumann, 1981). This is clear in the areas that dominate social psychology – notably work on prejudice, stereotyping, discrimination, and intergroup conflict. It is equally clear in the topics that we have neglected. Before the War, for instance, leadership was widely studied based on the hope that strong and charismatic leadership would provide a solution to the social problems of an industrial age. But after the war, the experience of such leadership led to a wholesale retreat from research into the topic and strong leaders came to be seen more as psychopaths than saviors (Haslam et al., 2010).

Nonetheless, even if social psychology does not address the issue of tyranny explicitly by name, there is certainly a literature which bears on phenomena that could be included under the rubric of tyranny: the ability of people to question authority or express dissent, their willingness to suppress and oppress those who do dissent or who are simply 'other,' and their willingness to condone oppressive authorities. The most obvious examples of this lie in the huge and ever-growing literature on authoritarianism (Adorno et al., 1950) and the more recent literature on social dominance (Sidanius and Pratto, 1999). These explain support for hierarchical and punitive authorities in terms of stable differences between individuals and, as such, one of the most telling criticisms of them derives from observing the realities of their historical point of origin in Nazi Germany. That is, on the one hand, support for tyranny was a collective phenomenon and on the other, it changed dramatically after the War as the collective context changed (Billig, 1976).

Accordingly, our focus in this article will be on the group psychology of tyranny. We will consider first work which bears on intragroup processes – that is, the ability of people to challenge authority within a group. We will then look at intergroup processes – that is, the willingness of people to oppress (or else support the oppression of) members of other groups.

Despite major differences between these two areas, we will see that they share in common the view that the problems lie in the very nature of group psychology and that therefore the answer to tyranny is for people to shun groups and act as individuals. We shall argue that this neglects the fact that it is only as group members that people can challenge inequality and oppression (Haslam and Reicher, 2012a; Reicher, 2004; Tajfel and Turner, 1979) and therefore such an argument ultimately serves to buttress rather than to challenge tyranny. We shall argue instead that the response should be less to do with whether we embrace groups or not than to question the nature and norms of the groups with which we identify.

Tyranny and Intragroup Processes: Janis's Groupthink Research

For nearly half a century, small-group research has been dominated by the concept of 'groupthink' (see Haslam et al., 2006). What is more, the idea has migrated from academia into the public domain and is used to explain everything from the financial crash to the latest fashions. As Fuller and Aldag observed a while back, "Groupthink is presented as accepted wisdom in sources ranging from *Educational Gerontology* to the *Utne Reader*, from *The New Criterion* to *Vogue*" (1998: p. 166).

In his original text, Janis (1972) compared four of the most calamitous decisions in recent U.S. history – neglecting the defense of Pearl Harbor in 1941, invading North Korea in 1950, invading of the Bay of Pigs in 1960, escalating the war in Vietnam between 1964 and 1967 – and contrasted them to two rather more successful decisions – developing the Marshall plan after World War I and the handling of the Cuban missile crisis in 1962. He argued that the former (unlike the latter) were characterized by an overestimation of in-group wisdom, a denigration of the out-group, and (most strikingly) overwhelming pressures toward conformity even if this goes against one's private beliefs. Indeed self-appointed 'mind-guards' would censor and punish any indications of dissent. In short, everyone agrees with the group 'line' and no one is prepared to mention its flaws, however obvious these might be. In this way, illusion and delusion come to dominate the decision-making process.

These symptoms of groupthink, Janis argued, derived principally from high group cohesion along with poor group structures and high levels of external stress. As Janis himself wrote "the more amiability and esprit de corps among the members of a policy-making group, the greater is the danger that independent critical thinking will be replaced by groupthink, which is likely to result in irrational and dehumanizing actions against outgroups" (1972: p. 13).

It follows from this analysis that the best way to avoid groupthink – to achieve the successful outcome of the Cuban missile crisis rather than the failure of the Bay of Pigs – is to dismantle the group, or at least to weaken the ties between group members. Janis advocates various measures such as bringing in external advisors, creating divisions within the group, and challenging group norms, all of which are designed to counteract 'the we-feeling of solidarity' (1982: p. 7).

There are, however, a series of problems with this analysis. Indeed, some critics mischievously suggest that such is the discrepancy between the continuing academic popularity of the Janis analysis and the actual evidence for it (for reviews, see Aldag and Fuller, 1993; Park, 1990), that just about the only argument for groupthink is the behavior of the research community that has studied this phenomenon (Haslam et al., 2006). Indeed, Fuller and Aldag go so far as to state that "in fact, in no study have all results been consistent with the [groupthink] model" (1990: p. 167).

First of all, then, there is no indication that the various components of groupthink (in-group boosting, out-group denigration, and enforced consensus) form a reliable cluster (Tetlock et al., 1992). Second, there is inconsistent evidence as to whether cohesiveness is an antecedent of groupthink phenomena, especially uncritical conformity (e.g., Flowers, 1977; Leana, 1985). Third, groupthink processes can lead to good organizational decisions as much as bad ones (Peterson et al., 1998). Moreover, as Haslam et al. (2006) argue, lack of shared identity and group cohesion makes it more difficult for group members to maintain the commitment necessary to drive *any* decision to success – especially in the face of opposition by out-groups.

In combination, these various findings point to two important conclusions. The first is that the ability to dissent, the willingness to consider discrepant information, the ways in which out-groups are treated and the nature (as well as the quality) of group decisions may have less to do with group psychology per se than with the specific norms and values associated with particular group identities. That is, whereas some groups may demand total deference to authority, others may be more accepting of dissent (Jetten and Hornsey, 2011; Packer, 2008). This is reflected in research that highlights how nationalism need not be a matter of 'my country right or wrong' but may take the form of critical attachment in which loyalty manifests itself as a willingness to identify and criticize national deficiencies (Roccas et al., 2006; Staub, 1997). Moreover, as studies in Israel, the United Kingdom, and the former Yugoslavia have shown, it is not identification with the group in and of itself which predicts denigration and support for the oppression of out-groups, but rather identification combined with a specific ideology of national glorification (Elcheroth and Reicher, 2014; Roccas et al., 2006). To cite Haslam et al. (2006), then: "although previous research has tended to see poor group decision making as a reflection of *psychological* deficiencies, there are strong grounds for questioning this assumption and arguing instead that problems of over-commitment are more *political* in nature" (p. 623, emphasis in the original).

We can also cite Haslam et al. (2006) in order to introduce the other conclusion. They ask: "how revolutionary movements (e.g. in science, politics and industry) would ever get off the ground if groups disavowed their own groupyness and were forever attentive to individual voices of doubt and dissent within their ranks" (p. 623). That is, to abandon groups is to leave the status quo inviolate. It is to leave existing tyrannies in place more than to create new ones. Such advice therefore serves less as part of the solution than as part of the problem of tyranny.

How, then, should we respond? The analysis provided thus far begins to suggest an alternative approach, one which is based on embracing rather than rejecting cohesive groups but which focuses on the content of group norms and hence on whether group power is used to create or challenge oppression in society. This analysis and these conclusions become clearer when we move from the consideration tyranny in intragroup processes to that of tyranny in intergroup processes.

Tyranny and Intergroup Process: Zimbardo's Stanford Prison Experiment

At just the same time as Irving Janis was developing his analysis of tyranny and oppression within groups, Philip Zimbardo was creating, in microcosm, his own system of tyranny and oppression between groups. What is more, his creation – the Stanford Prison Experiment (Haney et al., 1973; Zimbardo, 2007 – see also Zimbardo's Web site on the study: www.prisonexp.org) – has become, if anything, even better known than the work on groupthink both inside and beyond the academic world. Indeed, it is so well known that it only needs the briefest of outlines.

The study, which was scheduled to last for 2 weeks, began on 14 August 1971. Young men were randomly divided into the role of prisoner or guard. The prisoners were arrested in mock police raids and brought to the basement of the Stanford Psychology department which, for the purposes of the study, had been transformed into a simulated prison environment.

They had their heads shaved, were stripped naked, and made to change into short smocks without underwear, sprayed with what they were told was a delousing agent and then locked into small cells. They were overseen by guards – dressed in khaki shirts and trousers, carrying a whistle, a nightstick, and wearing reflective sunglasses – and under the authority of the 'prison superintendent,' Zimbardo himself.

What is best remembered about the study is the way in which the guards became increasingly abusive and dictatorial – taking pleasure in asserting their power, in degrading and humiliating the prisoners. As Zimbardo reported to Congressional Hearings just 2 months later:

> At the end of only six days we had to close down our mock prison because what we saw was frightening … human values were suspended, self-concepts were challenged, and the ugliest, most base, pathological side of human nature surfaced. We were horrified because we saw some boys ('guards') treat other boys as if they were despicable animals, taking pleasure in cruelty, while other boys ('prisoners') became servile, dehumanized robots.
>
> Zimbardo, 1971: p. 154

For Zimbardo, these findings testify to the power of the situation over the individual and show how "many people, perhaps the majority, can be made to do almost anything when put in psychologically compelling situations" (1971: p. 155). More specifically, Zimbardo argues that participants acted in terms of the role to which they had been assigned and that this happened without any intervention by the researchers:

> Participants had no prior training in how to play the randomly assigned roles. Each subject's prior societal learning of the meaning of prisons and the behavioural scripts associated with the oppositional roles of prisoner and guard was the sole source of guidance.
>
> Zimbardo, 2004: p. 39

Zimbardo's conclusion, then, is that human beings are helpless to resist the requirements and expectations placed upon them in groups. Thus, for instance, the tyrannical actions of his guards were: "emitted simply as a 'natural' consequence of being in the uniform of a 'guard' and asserting the power inherent in that role" (Haney et al., 1973: p. 12). In short, and as with Janis, the message is that group psychology is itself the problem. In line with his work on deindividuation processes (Zimbardo, 1969), Zimbardo contends that immersion among others subverts one's intellectual and moral sensibilities. It leaves one open to accepting influences and to acting in ways that one would normally shun. And where those influences are toxic, decent people will act with toxicity. Groups and power provide the 'compelling situations' that transform ordinary individuals into tyrants.

Zimbardo has applied this analysis to explain many instances of tyrannical behavior, such as the notorious abuses at the Abu Ghraib detention facility in Iraq. Indeed, he acted as an expert witness for the defense, arguing that the abusers were helpless to resist the requirements placed upon them. Staff Sergeant 'Chip' Frederick, accused of torturing detainees was not a 'bad apple,' Zimbardo maintained. Rather, he was a 'chip off the best block' who had been perverted by the 'bad barrel' in which he had been placed (Zimbardo, 2004: p. 344). However, Zimbardo's arguments were rejected in a letter from 49 psychologists to *The Observer* (official magazine of the *US Association for Psychological Science*) who claimed that he had misrepresented the scientific evidence in making his claims (Donnellan et al., 2007). They were also rejected by the judge in the case.

When it comes to looking for evidence that casts doubts on Zimbardo's role account, there is no need to look far. It can be found within the Stanford Prison Experiment itself. Thus, far from participants slipping helplessly into role, there are clear indications that participants resisted the position and the requirements placed upon them throughout the study (for detailed analyses, see Reicher and Haslam, 2006a; Haslam and Reicher, 2012b). For example, the prisoners rebelled from the start, and, if the extent of their rebellion declined over time, it was because Zimbardo actively intervened to achieve this end. He pulled one prisoner out of his cell, sought to recruit him as an informer, and led him to believe that he would not be able to leave the study even if he wanted to. But it is unclear if this impacted on resistance by getting the prisoners to act in role or, as Zimbardo suggests elsewhere, through inducing role confusion: that is, people no longer knew if they were participants in a study or truly imprisoned (Zimbardo, 1989). What is more, even after Zimbardo's intervention, prisoners continued to resist. What changed was the extent to which they supported each other in so doing – and once resistance was individualized, it was easier to contain.

Turning to Zimbardo's guards, the extent of resistance to role is even more striking. Again, Zimbardo himself acknowledges that many guards sought to be firm but fair and others actively sided with the prisoners. Only "about a third became tyrannical in their arbitrary use of power" (1971: p. 154). What is more, of this minority, the fame of the study largely hinges on the behavior of one guard, dubbed 'John Wayne' for his swagger and his aggression.

This raises the question of why this minority – and why this one individual – held sway over the other guards; why, that is, was he allowed to humiliate the prisoners without being questioned, let alone stopped? One answer is that he was seen to represent the values and the wishes of the prison authority, Philip Zimbardo. Certainly, Zimbardo did nothing to intervene or to suggest that abuse was inappropriate. Moreover, there is a growing body of testimony to suggest that Zimbardo and his fellow experimenters actively encouraged the guards to assert their power through use of the tools (e.g., chains, bags to put over the head) with which they had been provided. To cite Carlo Prescott, Zimbardo's chief advisor during the study, the guards were "merely doing what Zimbardo and others, myself included, encouraged them to do or frankly established as ground rules" (2005: p. 8). These ground rules were established from the very start. In particular, in Zimbardo's initial briefing to the guards (which can be seen in his own video of the study), he told them:

> You can create in the Prisoners feelings of boredom, a sense of fear to some degree, you can create a notion of arbitrariness, that their life is totally controlled by us, by the system, you, me … They can do nothing or say nothing that we don't permit. We're going to take

> away their individuality in various ways. In general what all this leads to is a sense of powerlessness.
>
> Zimbardo, 1989

All this is at odds with the claim that the guards acted autonomously and without any training. To borrow Banyard's elegant phrase: "it is not, as Zimbardo suggests, the Guards who wrote their own scripts on the blank canvas of the SPE, but Zimbardo who created the script of terror" (2007: p. 494).

Putting all these various criticisms together, it seems that conformity to role requirements was far from assured and that, to the extent that it did occur, it was as much a consequence of active leadership as of passive followership. On both grounds, the role account of tyranny between groups does not stand up. More generally, there is no support for the notion that there is something inherent in group psychology that makes us slaves to our roles within the group and prone to abuse those in other groups. Indeed, if we take seriously Zimbardo's suggestion that the reason why prisoner resistance was subdued has to do with the way he created role *confusion*, then we again are led to the idea that a coherent sense of group membership is necessary in order to challenge tyranny. This leads us to conclude that there is a need to rethink the relationship between group process, tyranny, and resistance.

Rethinking Groups and Tyranny: The BBC Prison Experiment

We have already seen how 1971 was an important year for social psychology. Janis was developing his groupthink analysis, Zimbardo was conducting his prison experiment, and, to complete the triad, Henri Tajfel and colleagues published their 'minimal group studies' (Tajfel et al., 1971). These studies, which addressed the minimal conditions of in-group bias, were the starting point for the development of social identity theory (Tajfel and Turner, 1979) and, subsequently, of self-categorization theory (Turner et al., 1987), which together form the social identity approach (for a review, see Reicher et al., 2010).

Over the ensuing decades, this tradition of work has transformed our understanding of group processes. Far from suggesting that groups make people less discriminating and liable to follow just any suggestion, the starting point of a social identity analysis is to argue that, when people categorize themselves in terms of their membership of any given social group (a social identification), they act in terms of the norms and values that characterize the group. Moreover, the fact that those who share the same social identification (e.g., 'we are all women') act in terms of the same shared norms/values, helps them to reach consensus, coordinate their actions, and thereby empowers them to achieve what they value (Haslam et al., 1998; Turner, 2005).

The question of when people act in terms of shared identity therefore becomes critical to understand when groups are empowered to reach their goals, and more particularly, when members of subordinate groups are able to challenge their oppressors (cf Haslam and Reicher, 2012a). Tajfel and Turner (1979) spell out some of those conditions. Notably, when a hierarchical social system restricts the upward mobility of subordinate group members (it is impermeable) and when that system is seen as both illegitimate and vulnerable (it is insecure), then they will act together to secure social change.

In 2001, 30 years after the Stanford Prison Experiment, the BBC Prison Experiment was run in order to investigate these ideas, again using a context in which ordinary people were randomly divided into guards and prisoners (see Reicher and Haslam, 2006a; for a full account of the study). While the study clearly drew on Zimbardo's paradigm, it differed in two crucial respects. First, the experimenters were careful not to position themselves alongside one group or the other, or to instruct the guards on how to behave (beyond telling them it was their job to make the prison run smoothly). Second, a number of theoretically based interventions were planned, based on the tenets of social identity theory. Most importantly, the prison system changed from being permeable at the start (prisoners had the prospect of 'promotion' to guard if they exhibited appropriate qualities) to being impermeable after a few days (prisoners were to stay prisoners however they behaved).

The results can be divided into two phases. The early phase provided strong support for a social identity approach to domination and resistance. First, as long as the system was permeable, the prisoners were divided on how to behave. They lacked shared social identity. Once the system became impermeable, however, the prisoners rapidly developed a shared sense of social identity. Second, shared identity enabled the prisoners to coordinate their actions, to support each other, and hence to become an effective unit. It empowered them to challenge the guards. By contrast, the guards lacked shared identity, failed to coordinate their actions, and indeed acted against each other thus canceling out each other's efforts. This undermined their effectiveness in meeting the challenge of the prisoners.

The combination of prisoner identity/coordination/empowerment and lack of guard identity/coordination/empowerment ultimately undermined the initial hierarchy. The prisoners broke into the guards' quarters and the prisoner–guard system collapsed. Together, the participants decided to inaugurate a new egalitarian system – a 'commune' in which all would be equal and work together. This ushered in the second phase of the study and provided new lessons, less in how people respond to an imposed system of inequality than in when and why they will create such a system for themselves.

At first, the Commune functioned well, with members working hard and working together. But there were dissenters. What is more, the dedicated 'Communards' were unwilling to use their authority to defend their system. In this context, they quickly became disenchanted and pessimistic about their ability to create a world rooted in their own values. The dissenters were aware of this and proposed an alternative, a new hierarchical system. As emblematic of the regime they proposed, these 'new guards' requested black uniforms, black berets, and black sunglasses – the iconography of fascism.

The study ended before this new system could be imposed. However the evidence suggests that, while the supporters of the Commune did not enthusiastically embrace it, their opposition was weakening and they certainly lacked the will to resist it. This is reflected not only in what they said and did but also in

the measures of authoritarianism that the researchers administered each day and that showed a steady rise as the Commune faltered. Paradoxically, then, the same people who had rejected, resisted, and overturned a relatively mild form of inequality at the study's start were on the brink of recreating a far more draconian system of inequality at the end.

These findings provide a root and branch challenge to traditional accounts of the psychology of tyranny. On the one hand, they suggest that it is not existence of cohesive groups which, in itself, leads to tyranny. Rather, cohesion gives group members the power to enact their norms and values even against the opposition of others. Far from removing their agency and their capacity to shape their own fate, it is as group members that most people achieve agency in the world (Reicher and Haslam, 2006b). But whether this is emancipatory or oppressive, progressive or toxic depends critically on the content of specific group identities and hence on the sorts of things they prescribe and value.

On the other hand, the findings, particularly from the second phase of the study, suggest that we should be more concerned with group failure than with group strength. When group members feel incapable of acting together to mold their own fate, they are far more likely to begin to shun democracy, to embrace strong leaders who will do things for them, and to accept extreme proposals from those leaders (Haslam and Reicher, 2007).

Conclusion

This analysis has centered on the question 'what is the role of group processes in the psychology of tyranny.' We have seen how the traditional answer has been to suggest that groups 'naturally' make people more slavish within the group and more tyrannical outside the group. The clear message is 'avoid groups' and certainly avoid the exercise of group power because groups and power are bad both for you and society. We have looked at the evidence for this widely disseminated viewpoint and found it wanting. We have also presented alternative evidence pointing to a very different analysis. This suggests that groups provide the basis for overcoming existing tyrannies and that the failure of groups is what allows such tyrannies to be established. The message here is that we should embrace groups with democratic and humane values and that we should use group power in order to defend these values. In short, the traditional psychology of tyranny is not simply wrong, it is dangerously wrong. It does not simply provide a misleading analysis of tyranny, whereas it gives advice that makes tyranny more (not less) likely.

There was a telling moment early on in the BBC Prison Study when the guards were reflecting on their uniforms, resources, and power. Without knowing the name of Zimbardo they expressed what has, by now, become a cultural commonplace: put the uniform on the man (or woman) and he (or she) will become a tyrant. So they shunned their uniform, resources, and power. Equally, later on during the Commune, several of the same people, and for precisely the same reasons, avoided using the power of this group against those who wanted to undermine their values. The irony, as we have seen, is that their very desire not to be tyrants created the conditions for a slide into tyranny. It is this that explains the paradox whereby those who rejected mild inequalities came to accept harsh tyranny. And, in addition to the theoretical reasons, this is why we need to reject the traditional psychology of tyranny.

See also: Collective Behavior, Social Psychology of; Decision Making, Psychology of; Deindividuation, Psychology of; Group Processes, Social Psychology of; Self-Categorization Theory; Social Identity in Social Psychology; Social Psychology.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper & Brothers, New York.

Aldag, R.J., Fuller, S.R., 1993. Beyond fiasco: a reappraisal of the groupthink phenomenon and a new model of group decision process. Psychological Bulletin 113, 533–552.

Banyard, P., 2007. Tyranny and the tyrant. The Psychologist 20, 494–495.

Baumann, Z., 1981. Modernity and the Holocaust. Polity Press, Oxford.

Billig, M., 1976. Social Psychology and Intergroup Relations. Academic Press, London.

Donnellan, M.B., Fraley, R.C., Krueger, R.F., 2007. Not so situational. APS Observer 20 (6), 5.

Elcheroth, G., Reicher, S., 2014. 'Not our war, not our country': contents and contexts of Scottish political rhetoric and popular understandings during the invasion of Iraq. British Journal of Social Psychology 53, 112–133.

Flowers, M.L., 1977. A laboratory test of some of the implications of Janis's groupthink hypothesis. Journal of Personality and Social Psychology 33, 888–895.

Fuller, S.R., Aldag, R.J., 1998. Organizational Tonypandy: lessons from a quarter century of the groupthink phenomenon. Organizational Behavior and Human Decision Processes 73, 163–184.

Haney, C., Banks, C., Zimbardo, P., 1973. A study of prisoners and guards in a simulated prison. Naval Research Review 9, 1–17 (Reprinted in Aronson, E. (Ed.), Readings about the Social Animal, third ed. W.H. Freeman, San Francisco, CA, pp. 52–67.).

Haslam, S.A., Reicher, S.D., 2007. Identity entrepreneurship and the consequences of identity failure: the dynamics of leadership in the BBC Prison Study. Social Psychology Quarterly 70, 125–147.

Haslam, S.A., Reicher, S.D., 2012a. When prisoners take over the prison: a social psychology of resistance. Personality and Social Psychology Review 16, 154–179.

Haslam, S.A., Reicher, S.D., 2012b. Tyranny: revisiting Zimbardo's Stanford prison experiment. In: Smith, J.R., Haslam, S.A. (Eds.), Social Psychology: Revisiting the Classic Studies. Sage, London.

Haslam, S.A., Reicher, S.D., Platow, M., 2010. The New Psychology of Leadership. Psychology Press, London.

Haslam, S.A., Ryan, M.K., Postmes, T., Spears, R., Jetten, J., Webley, P., 2006. Sticking to our guns: social identity as a basis for the maintenance of commitment to faltering organizational projects. Journal of Organizational Behaviour 27, 607–628.

Haslam, S.A., Turner, J.C., Oakes, P.J., McGarty, C., Reynolds, K.J., 1998. The group as a basis for emergent stereotype consensus. European Review of Social Psychology 8, 203–239.

Janis, I.L., 1972. Victims of Groupthink. Houghton-Mifflin, Boston.

Jetten, J., Hornsey, M.J. (Eds.), 2011. Rebels in Groups: Dissent, Deviance, Difference, and Defiance. Wiley-Blackwell, New York.

Leana, C.R., 1985. A partial test of Janis' groupthink model: effects of group cohesiveness and leader behavior on defective decision making. Journal of Management 11, 5–17.

Packer, D.J., 2008. On being both with us and against us: a normative conflict model of dissent in social groups. Personality and Social Psychology Review 12, 50–72.

Park, W., 1990. A review of research on groupthink. Journal of Behavioral Decision Making 3, 229–245.

Penic, S., Elcheroth, G., Reicher, S.D. Can there be patriots after a nationalist war? The struggle between recognition and marginalization of dissenting voices. Political Psychology, in press.

Peterson, R.S., Owens, P.D., Tetlock, P.E., Fan, E.T., Martorana, P., 1998. Group dynamics in top management teams: groupthink, vigilance and alternative models of organizational failure and success. Organizational Behaviour and Human Decision Processes 73, 272–305.

Prescott, C., 2005. The Lie of the Stanford Prison Experiment. The Stanford Daily, 28th April, 8.
Reicher, S.D., 2004. The context of social psychology: domination, resistance and change. Political Psychology 25, 40–62.
Reicher, S.D., Haslam, S.A., 2006a. Rethinking the psychology of tyranny: the BBC Prison Study. British Journal of Social Psychology 45, 1–40.
Reicher, S.D., Haslam, S.A., 2006b. On the agency of individuals and groups: lessons from the BBC Prison Study. In: Postmes, T., Jetten, J. (Eds.), Individuality and the Group: Advances in Social Identity. Sage, London, pp. 237–257.
Reicher, S.D., Spears, R., Haslam, S.A., 2010. The social identity approach in social psychology. In: Wetherell, M.S., Mohanty, C.T. (Eds.), Sage Identities Handbook. Sage, London, pp. 45–62.
Roccas, S., Klar, Y., Liviatan, I., 2006. The paradox of group-based guilt: modes of national identification, conflict vehemence, and reactions to the in-group's moral violations. Journal of Personality and Social Psychology 91, 698–711.
Sidanius, J., Pratto, F., 1999. Social Dominance. Cambridge University Press, Cambridge.
Staub, E., 1997. Blind versus constructive patriotism: moving from embeddedness in the group to critical loyalty and action. In: Bar-Tal, D., Staub, E. (Eds.), Patriotism in the Lives of Individuals and Nations. Nelson-Hall, New York, pp. 213–228.
Tajfel, H., Flament, C., Billig, M.G., Bundy, R.F., 1971. Social categorization and intergroup behaviour. European Journal of Social Psychology 1, 149–177.
Tajfel, H., Turner, J., 1979. An integrative theory of intergroup conflict. In: Austin, W.G., Worchel, S. (Eds.), The Social Psychology of Intergroup Relations. Brooks/Cole, Monterey, CA, pp. 33–48.
Tetlock, P.E., Peterson, R.S., McGuire, C., Chang, S., Feld, P., 1992. Assessing political group dynamics: a test of the groupthink model. Journal of Personality and Social Psychology 63, 403–425.
Turner, J.C., 2005. Explaining the nature of power: a three-process theory. European Journal of Social Psychology 35, 1–22.
Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D., Wetherell, M.S., 1987. Rediscovering the Social Group: A Self-categorization Theory. Blackwell, Oxford.
Waller, J.E., 2010. The social sciences. In: Hayes, P., Roth, J.K. (Eds.), The Oxford Handbook of Holocaust Studies. Oxford University Press, Oxford, pp. 667–679.
Zimbardo, P., 1969. The human choice: individuation, reason and order versus deindividuation, impulse and chaos. In: Arnold, W.J., Levine, D. (Eds.), Nebraska Symposium on Motivation, vol. 17. University of Nebraska Press, Lincoln, NB, pp. 237–307.
Zimbardo, P., 1971. The Psychological Power and Pathology of Imprisonment. Hearings before subcommittee no. 3 of the Committee on the Judiciary House of Representatives Ninety-Second Congress, First sessions on corrections - Part II, Prisons, prison reform and prisoners' rights: California (Serial No. 15, 25 October). US Government Printing Office, Wahington, DC.
Zimbardo, P., 1989. Quiet Rage (video). Stanford University, Stanford, CA.
Zimbardo, P., 2004. A situationist perspective on the psychology of evil: understanding how good people are transformed into perpetrators. In: Miller, A. (Ed.), The Social Psychology of Good and Evil. Guilford, New York, pp. 21–50.
Zimbardo, P., 2007. The Lucifer Effect: How Good People Turn Evil. Random House, London.

U

Uncertainty: History of the Concept

Joshua Wakeham, University of Alabama, Tuscaloosa, AL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the various understandings of the concept of uncertainty in the social sciences. First, it examines the idea of uncertainty and related concepts. Then, it explores the use of the concept in economics, psychology, and sociology.

Within the social sciences, the concept of uncertainty has long been of interest, particularly in areas concerned with decision-making and knowledge. In very broad terms, uncertainty occurs at the limits of knowledge. In those social sciences that take an information-processing perspective on the problem of knowledge and cognition, such as economics, organizational studies, and some parts of psychology, some might describe uncertainty as a fundamental problem, even in places where it is not explicitly discussed. Uncertainty lurks in the background of any seemingly rational decision that is of interest to social scientists. A lack of certainty about the outcome of any action colors the way individuals or groups evaluate the world and their choices. Yet this is not the only understanding of the concept of uncertainty. One can find references to the problem of uncertainty – and related concepts – scattered throughout economics, psychology, sociology, anthropology, and related subfields. Both across and within the social science disciplines, there are differing understandings of the concept of uncertainty, as well as differing interpretations of its significance for social behavior and social science.

In order to better illuminate the history of the concept of uncertainty in the social sciences, this article sets out to do two things. First, it provides some conceptual discussion of uncertainty. Second, it provides historical overviews of the concept of uncertainty in each of the social sciences.

Conceptual Considerations

The Dimensions of Uncertainty

Broadly speaking, uncertainty refers to an epistemic state at the limits of knowledge. It concerns what is known or believed without certainty. It also concerns what is not known. If being certain means having clarity with regard to the truth, then being uncertain means having an obscured view of the truth. The problem of uncertainty points not only to the limits of what any individual might know, but also to the ultimate limits of what is knowable.

To put it another way, there are subjective and objective dimensions to uncertainty (Tannert et al., 2007). First, there are the subjective dimensions of uncertainty. An individual may experience uncertainty with regard to his or her own knowledge about some particular topic. Some experience or some piece of evidence might challenge an individual's once firmly held belief, causing them to have doubts about this belief. An individual may experience uncertainty about the possible outcome of some course of action. In important ways, this subjective, phenomenological experience of uncertainty is both cognitive and emotional. As such, it has obvious relevance to social behavior. There are questions of not only how the experience of uncertainty drives or influences behavior, but also how social context shapes the perception and experience of this cognitive and emotional uncertainty.

Second, there are the objective dimensions of uncertainty, which presume a world out there that is knowable only to a degree. Uncertainty, in this sense, is not necessarily something felt, but rather a feature of living in a complex world. It is useful to distinguish between epistemological and ontological uncertainty (Van Asselt and Rotmans, 2002). In the case of epistemological uncertainty, something that is potentially knowable remains unknown. Further research or searching for more information can reduce such uncertainty. In the case of ontological uncertainty, something remains unknown because it is ultimately unknowable. This may be due to the sheer complexity of the problem, the stochastic nature of the situation, or because natural processes obscure access to the relevant knowledge (such as time-obscuring access to historical knowledge). Thinking about uncertainty as an objective feature of the world is related to social scientific work at two levels. At the first level is the thinking about how other people encounter and deal with the uncertainty of the world around them. Problems of epistemological uncertainty, for example, and the practical problem of distinguishing between epistemological and ontological uncertainty are directly relevant to the social production of knowledge in the sciences, the law, medicine, business, and other venues. More broadly, how people manage ontological uncertainty, particularly as it relates to the complex social, political, technological, and economic systems that make up modernity, is of great interest for the social sciences. At the second level is the thinking about

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03175-5

how the social sciences themselves encounter uncertainty as a feature of the world. This has direct bearing on the limits of proposed models of human behavior, particularly on their predictive validity. Uncertainty, in this sense, is a technical problem that social scientists must account for in their models and explanations of human behavior.

Presumably, most of the time, the subjective experience of uncertainty and the reality of an uncertain world line up, so speaking of the uncertainty in the market or surrounding a particular policy decision or about some organizational strategy practically conflates these two analytic dimensions. Nonetheless, within the social sciences, researchers have emphasized different aspects of uncertainty, depending on their field and their research subject. This expanded view of uncertainty as having both subjective and objective dimensions provides a sufficiently broad framework for capturing the various ways the social sciences have used the term. However, it is also important to recognize that these different points of conceptual emphasis result in research projects that do not necessarily speak to one another.

Related Concepts

Within the social sciences, uncertainty is rarely considered on its own. It is often paired with other ideas related to challenges and problems in knowledge. This includes terms like risk, ambiguity, and ignorance. A brief discussion of how uncertainty has been associated with each of these terms is useful for further conceptual clarity.

Risk and Uncertainty

Uncertainty has most commonly been paired with risk in the social sciences. This is particularly true in economics, where it goes back to Knight's early work (1921), but it is also common in the more recent sociological work on risk and organizations, as well as work on the 'risk society' (Douglas and Wildavsky, 1983; Beck, 1992, 2009).

Risks are generally understood to be the *known* potential negative outcomes of a course of action. Risks include things like a loss on an investment, a nuclear plant having a meltdown, or the environmental and human health impact of widespread use of automobiles. A more technical understanding of risk typically includes some measure or guess of the probability of that negative-outcome happening multiplied by some measure of the presumed cost of that negative outcome. As risks are known potential outcomes, individuals or organizations presumably make efforts to mitigate or manage them, and avoid risks with costs that are too severe.

Uncertainty may be understood as the state of mind of someone deciding on a course of action without a clear outcome. In this subjective sense of the term, an uncertain individual consider the benefits and the risks, or potentially negative outcomes, of any course of action. Perceptions of costly risks are more likely to induce feelings of uncertainty about a particular course of action. However, when paired with risk, uncertainty has been more often understood in the more objective sense of the term.

If risks are known potential negative outcomes, uncertainty, in an objective sense, refers to the broader set of unknowns that may affect the course of action and its outcome. Risks, in theory, can be accounted for, but uncertainty cannot. A deeper critique of theories of risk management (in business, in management, or in other fields; or social sciences, such as economics), is that in the real world, the real probabilities of negative outcomes are never truly known, meaning that real uncertainty is more pervasive than any models of risk reveal (Taleb, 2010).

A final distinction between risk and uncertainty worth noting is the evaluative sense in which they are used. Risk bears a distinctly negative connotation. Objective uncertainty, by contrast, is neutral. Unknown factors may have a positive or negative effect on any course of action.

Ambiguity and Uncertainty

An ambiguous situation or problem is one that invites multiple interpretations. There is a lack of clarity about what it really is. Ambiguity has often been lumped together or even conflated with uncertainty. The amount of overlap between the two concepts, of course, depends on what definition one gives each idea.

Those who take information-processing approach to the problems of knowledge and cognition tend to define ambiguity, like uncertainty, more technically. Ambiguity, like uncertainty, has to do with missing information that is relevant to the outcome of an action (Galbraith, 1973; Frisch and Baron, 1988). Some economists, building on Ellsberg's paradox (Ellsberg, 1961), define ambiguity as uncertainty about the probability of possible outcomes rather than uncertainty about the outcome itself (Camerer and Weber, 1992; Dequech, 2000). Others in organizational studies (Schrader et al., 1993) have argued that in an ambiguous situation, the structure of the problem itself – i.e., the relationship between the potentially relevant variables – is unknown or unsettled, while in an uncertain situation, the structure of the problem is settled, but the value of the relevant variables is not. In these kinds of distinctions, one finds attempts to elucidate not just how the experience of ambiguity differs from uncertainty, but how dealing with ambiguity in order to solve a problem requires different strategies than dealing with uncertainty.

Ambiguity, conceived as a subjective phenomenon, has similarities with the experience of uncertainty. Encountering ambiguity is a troubling cognitive and emotional experience that pushes people to seek out clarity, even at an unconscious level. From experiments on perception in cognitive psychology (Hoffman, 1998) to sociological studies of culture (Zerubavel, 1993), social scientists have demonstrated how people, at the individual or collective level, appear motivated to resolve ambiguity, imposing a sense of clarity. Research using this conception of ambiguity tends to focus on the cognitive and social processes through which individuals and groups perceive and define problems and construct knowledge. Interpretive approaches to social science, particularly in anthropology and sociology, tend to focus on this kind of ambiguity, often conflating it with uncertainty. Arguably, the focus on social and cognitive processes and the production of knowledge rather than the outcomes of actions, where uncertainty has greater practical salience, may account for this emphasis on ambiguity and the underdeveloped sense of uncertainty as a distinctive notion.

Ignorance and Uncertainty

Some social scientists have also been interested in the larger problem of ignorance. Ignorance refers to the broad set of unknowns. Ignorance is of interest not only because it points to the limits of knowledge, but also because of its dynamic role in human affairs, as both a productive and destructive force (Gross, 2007). To that end, some have offered typologies that make distinctions between different kinds of ignorance. Merton (1987), for example, distinguishes between unrecognized and specified ignorance within scientific communities. Unrecognized ignorance refers to 'unknown unknowns' (Kerwin, 1993) or 'meta-ignorance' (Smithson, 1989) – ignorance of what one does not know. Specified ignorance refers to the acknowledged limitations of knowledge, which often drive the search for knowledge (in science and in other fields). Several related terms like *Nichtwissen*, nonknowledge, negative knowledge (Knorr-Cetina, 1999), and nescience (Simmel, 1906) are used in similar ways.

Smithson's taxonomy of ignorance (1989) makes multiple levels of distinction. The first level of distinction is between irrelevance and error. Things that are irrelevant are those that are deemed unworthy of further consideration. Error includes both distortions of information (confusion and inaccuracy) and incompleteness of information. Incompleteness is broken down into absence (e.g., missing piece of information) and uncertainty. Uncertainty includes vagueness, probability, and ambiguity. In this typology, uncertainty is thus a subtype of ignorance related to states where one has partial information.

Recent social scientific interest in the problem of ignorance tends to frame the social production of ignorance as process that is complementary to the production of knowledge (Proctor and Schiebinger, 2008). Cultural norms and social structures play a key role in placing and enforcing limits on what people know about the world around them. Within this framework, uncertainty is often treated as a political tool. For example, politically motivated propagandists use the scientific community's provisional sense of the truth to instill a false sense of uncertainty about the status of official scientific claims regarding things such as cigarette smoking or environmental degradation (Oreskes and Conway, 2011).

Uncertainty in the Social Sciences

Uncertainty in Economics

Of all the social sciences, economics has given the most attention to the concept of uncertainty. Given economics' interest in successfully modeling rational action, uncertainty poses a real problem. Uncertainty, as either an objective or subjective phenomenon, challenges many of the basic assumptions of such modeling efforts. The history of the concept of uncertainty in economics is very much one of attempting to 'tame and domesticate uncertainty' (Quiggin, 2009, p. 195).

For uncertainty in economics, 1921 was an important year. Both British economist John Maynard Keynes and American economist Frank M. Knight published books that dealt, in part, with the problem of uncertainty. In his *Treatise on Probability*, Keynes (1921) lays out an argument about demonstrating logical relationships between two propositions based on both a subjective probabilistic assessment of certainty in one's belief and the accuracy of that belief. While some relationships between two propositions can be established probabilistically, there are some instances where no numerical relationship can be established. Keynes considers these instances to reflect a kind of fundamental uncertainty about the world (Lawson, 1985; Dequech, 2000). Knight's *Risk, Uncertainty, and Profit* (1921) offers a basic definitional distinction between risk and uncertainty. Risk is measurable. The probabilities of the various outcomes are known. Uncertainty refers to those unknowns that cannot be quantified or measured. For both Keynes and Knight, the looming uncertainty of the world also explains some seemingly irrational social behavior (for Keynes, reliance on conventional thinking in the face of long-term uncertainty; for Knight, development of large, hierarchical organizations), as such behaviors have formed to help deal with or manage it. Keynes and Knight's similar understandings of uncertainty had influence on the post-Keynesian and Austrian schools of economics, respectively, but really failed to take hold in mainstream economics despite being cited often (Kelsey and Quiggin, 1992).

Uncertainty as an inescapable, objective feature of the world cannot effectively accounted for in traditional economic models, so most economists ignored or abandoned these particular definitions of uncertainty, treating them as heterodox. Instead, alternative understandings of uncertainty as a subjective phenomenon became more prominent. Being able to incorporate subjective estimates of probability – expressed or revealed – without regard to their accuracy (that is, essentially ignoring whether these estimates reflected objective features of the world) into economic models was a key. This is where development of expected utility theory comes in. Von Neumann and Morgenstern (1944) demonstrated that people evaluate outcomes by their expected utility – and that expected utility was subject to variation. This allowed economists to explain why people were often risk averse under conditions of uncertainty. This model influenced several lines of work, including game theory and several variations on the expected utility model. This includes Savage's (1954) work deriving subjective probabilities and expected utility from people's preferences, as well as Arrow and Debreau's (1954) demonstration of 'sufficient conditions' – namely, market completeness, or a market in which all future outcomes are accounted for by assuming a 'frictionless' market for all current goods – "for the existence of general equilibrium … under conditions of uncertainty" (Quiggin, 2009, p. 196). In what would become the prevailing microeconomic model, this meant that uncertainty, as it was conceived, could be effectively accounted for. Much attention turned to developing ways to better account for subjective probabilities in the models, such as Bayesian approaches (Hirshleifer and Riley, 1992), or variations on expected utility theory like rank-dependent expected utility theory (Kelsey and Quiggin, 1992).

Since the 1970s, a growing body of empirical findings has shown that people consistently violated the predictions of the expected utility theory. In particular, the work of psychologists Daniel Kahneman and Amos Tversky and their collaborators (Kahneman and Tversky, 1979; Kahneman et al., 1982; Tversky and Kahneman, 1992) demonstrated that under conditions of risk and uncertainty, people violated the predictions of the

expected utility model, but in consistent ways. In place of expected utility theory, they propose prospect theory, which takes into account the psychological dimensions of decision making. Prospect theory breaks choices down into two phases: editing/framing and evaluation. In the editing/framing phrase, people simplify available options and outcomes through a variety of heuristics, or cognitive shortcuts, which are determined by the context or formulation of the problem. Such editing or framing makes the evaluation phase easier. During this phase, people evaluate the outcomes or prospects of decision, following several psychological rules, such as using one's starting position to judge gains or losses, overweighing small probabilities or underweighing large probabilities depending on the risks, and loss aversion. The key lesson from all of this is that under conditions of uncertainty, people's behavior is not consistent with the rational actor model, but they do behave in relatively predictable ways. Essentially, people's minds draw on even the flimsiest pieces of evidence from the context in order to impose a subjective sense of certainty in a situation with objective uncertainty. At times, this leads people to overestimate the potential costs and underestimate the benefits (or vice versa) of such risky decisions. This work, of course, contributed directly to the development of the new fields, behavioral economics and behavioral finance.

These developments in the field of economics, along with the Great Recession of 2008, have led some (e.g., Taleb, 2010) to question the wisdom of economics effectively bracketing off the problem of objective uncertainty. However, the problem of effectively modeling or theorizing about unforeseen events and other exogenous shocks to economic markets still remains.

Uncertainty in Psychology

Uncertainty, usually as a subjective experience, often lingers in the background of various theories in psychology. Psychologists have connected uncertainty to the concepts of personality, emotion, and cognition, and the problem of decision making. Although work with these concepts often interacts, they map roughly onto the three distinct normative approaches to the problem of uncertainty that Smithson (2009) identifies.

The first approach comes out of the politicized psychoanalytic tradition of the mid-twentieth century. Tolerance of ambiguity and uncertainty are taken as signs of a positive, well-adjusted personality type, whereas intolerance of such epistemic conditions is often taken as signs of an authoritarian personality (Adorno et al., 1950). While the politics of such an approach have largely been toned down, interest in the concept of personality and relationship to the psychological experience of uncertainty have persisted. At the heart of this research is the idea that there are measurable differences in how individuals respond to uncertainty (Sorrentino et al., 1984; Sorrentino and Short, 1986; Sorrentino and Roney, 2000). Freeston et al. (1994) even developed an Intolerance for Uncertainty scale to measure such individual variation. Differences in tolerance for uncertainty are in turn measured against other psychometric data, including now widely accepted 'big five' measures of personality – openness, conscientiousness, extraversion, agreeableness, and neuroticism (Berenbaum et al., 2008). Even within this more recent work, a lack of tolerance for uncertainty – as an inevitable fact of life – is seen as potentially problematic from the individual psychological point of view.

The second approach to uncertainty focuses the cognitive, emotional, and physiological effects of 'uncertainty, unpredictability, and uncontrollability' (Smithson, 2009, p. 206). Uncertainty, in this sense, is a generally experienced problem. As such, it may even be a variable induced in laboratory settings. In such research, psychologists typically assume that uncertainty induces psychological distress or strain, including anxiety and motivation to seek certainty. For example, Berger and Calabrese (1975) put forth an 'uncertainty reduction theory' of social interaction and communication, arguing that in the initial stages of successful social interaction depends on reducing uncertainty among participants. Social psychologists have been interested in not only how uncertainty affects social interaction and communication (Gao and Gudykunst, 1990), but also social judgment and evaluation (van de Boss 2001, 2009). Recent work in neuroscience and cognitive psychology suggests that the drive for certainty works even at an unconscious level (Burton, 2008). The general lesson is that the experience of uncertainty often drives people into behaviors and beliefs that are potentially problematic because of an underlying cognitive and/or emotional need for certainty. Wilson et al. (2005) have challenged this largely negative assessment of uncertainty, demonstrating that the drive for certainty sometimes lessens the positive emotional experience of a positive event and that uncertainty about a positive event can lengthen the experience of positive emotions.

The third approach to uncertainty that Smithson identifies comes from research focused on decision making. This approach takes an information-processing perspective on cognition (Smithson, 1989). It is often found in cognitive psychology, but it is also related to other approaches to decision-making outside of psychology, including organizational studies, economics, and behavioral economics. Kanheman and Tversky's (1982) work is probably the best example of this approach. Researchers taking this approach tend to treat uncertainty like economists do, often conflating it with notions of probability or risk that can be accounted for in models of decision-making. The central concern is understanding and identifying cognitive biases and heuristics that people rely on when making decisions under conditions of uncertainty.

Uncertainty in Sociology

Considerations of uncertainty figure less prominently within the field of sociology. There are some areas, however, where the problem of uncertainty garners some explicit attention. In particular, in the field of organizational studies and in the social theoretical work on risk culture and society.

Organizational Studies

Organizational sociologists, and others under the broader umbrella of organizational studies, have long been interested in how uncertainty drives organizational behavior. The problem is, like in the other social sciences, the term uncertainty has often been used to refer to slightly different things. For example, of general interest is the uncertainty an organization faces in its task environment, yet environmental uncertainty has been conceived of as a state or an objective feature of

the world, as largely a matter of the perception of organizational actors, or having to do with the unknown effects of an action, or the unknown responses of the environment (Milliken, 1987). Because of such differences in understanding the concept, scholars have used the problem of uncertainty to account for an array of organizational behaviors.

The Carnegie School's early work on the cognitive-behavioral approach to organizational decision making, for example, focuses on the ways in which practical limits constrain or bound rational behavior (March and Simon [1958]1993). In the face of uncertainty, organizational actors often rely on standardization of behavior and responses (Cyert and March [1964]1992). Uncertainty thus contributes to the kind of suboptimal, bounded rationality observed in organizational decisions.

Uncertainty has also played a central role in the work on organizational design (Thompson, 1967; Lawrence and Lorsch 1967; Galbraith, 1973). Organizations cannot have full knowledge of their environment, so they must anticipate the unforeseen. One of the goals in organizational design is not necessarily to eliminate uncertainty (this is impossible), but to cope with it. This means not only paying attention to various sources of information (through the division of labor or specialization), but also designing an organization that is responsive to unanticipated information (through the successful integration or coordination of the parts). Attempting to cope with such uncertainty has been invoked to explain macrolevel organizational behaviors from a wide array of theoretical perspectives, including population ecology (Hannan and Freeman, 1977), resource dependency theory (Pfeffer and Salancik, 1978), and neoinstitutionalism (DiMaggio and Powell, 1983). Uncertainty, as either an objective feature of the environment or as a subjective perception of key organizational actors, can thus possibly explain differences in organizations across environment (more or less certain), as well as seemingly irrational or inefficient organizational behaviors and forms.

Beyond theories of organizations, the problem of uncertainty also plays a central role in the practical considerations of the literature on organizations operating in high-risk environments. The complexity of both technologies and large organizations make accurate assessment of the situations, actions, or consequences more difficult, if not impossible. With the threat of serious human and environmental costs for being wrong, managing risks and uncertainty is paramount. In his normal accident theory, Perrow ([1984]1999) argues that major disasters are an inevitable by-product of working under such epistemic conditions, so some high technological endeavors should be abandoned. In contrast, drawing on cases of successful organizational management of high-risk endeavors, the high-reliability organization theory out of Berkeley (Weick, 1987; Roberts, 1989) proposes that uncertainties and risks can be effectively managed with the right kind of management, organizational culture, and approach to teamwork and organizational learning and knowledge. Arguably, what separates these two different approaches to risk management is an assessment of the nature of uncertainty and our ability or inability to anticipate problems. In the high-reliability approach and in the broader literature on organizational risk management, one finds an understandable focus on preventing foreseeable mistakes and disasters. This places more focus on epistemological uncertainty. Perrow's more cautious approach places greater emphasis on the idea of ontological uncertainty – that is, there are things that are going to happen that could not be predicted.

Risk Culture and Society

The other area within sociology where the concept of uncertainty receives some explicit attention is in the social theoretical work concerned with the role of risk in modernity. Within this area of research, there has been emphasis on both the cultural and structural elements related to problems of risk. Douglas and Wildavsky's (1983) work on the culture of risk explores how perceptions of risk are tied to cultural ideals about what constitutes a good society. Fears and concerns about certain dangers do not necessarily reflect real risks, but rather culturally framed anxieties embedded within a particular social organization. More recently, Cerulo (2006) argues that culture plays a role in preventing people from anticipating the worst possible outcomes, skewing their perceptions to be overly optimistic. While this work emphasizes the problem of risk perception, the problem of uncertainty looms in the background. Culture shapes what areas of uncertainty receive more attention within a given social context, and thus, what areas of uncertainty people may fail to see coming.

Beck (1992, 2009) argues that one of the hallmarks of the postindustrial, globalized world is the unequal distribution of risks and hazards. Scientific and technological advances have brought with them the potential for untold human and environmental costs; yet those costs are not born out equally, especially in a globalized economy. This political dynamic, in effect, means that the powerful are able to minimize negative uncertainties in their lives, while those on the bottom have lives exposed to even more uncertainty.

Conclusion

The concept of uncertainty has had a varied, if somewhat inconsistent, history in the social sciences. At the heart of the matter is a conceptual confusion. Is uncertainty a characteristic of the knowledge itself or a characteristic of the knower? For better or worse, the social sciences have treated it as both. Uncertainty refers to an inevitable feature of the chaotic, unpredictable world, yet it also refers to a psychological state or phenomenological experience. How uncertainty has been conceived of in the social sciences depends on the agenda, methods, and theoretical assumptions of the researcher. Economists, for example, turned away from the idea of uncertainty as a kind of fundamental unpredictability because it essentially stands outside of the models and thus, the basic assumptions of economists. Uncertainty as a subjective experience holds greater sway in psychology and sociology because it is seen as a fundamental driver for an array of seemingly nonrational individual and collective behaviors. Yet this drive for certainty in an uncertain world does not eliminate the intrusion of unpredictable events in human affairs.

The lack of conceptual clarity and consistency does not undermine the importance of this idea to the social sciences. Rather it points to the challenges and the potential intellectual richness of working at the limits of human knowledge.

Bibliography

Adorno, T.W., Frenkel-Brunswik, E., Levinson, D.J., Sanford, R.N., 1950. The Authoritarian Personality. Harper, New York.

Arrow, K.J., Debreu, G., 1954. Existence of an Equilibrium for a Competitive Economy. Econometrica 22 (3).

Beck, U., 1992. Risk Society: Toward a New Modernity. Sage Publications, Thousand Oaks, CA.

Beck, U., 2009. World at Risk. Polity Press, Malden, MA.

Berenbaum, H., Bredemeier, K., Thompson, R.J., 2008. Intolerance of uncertainty: exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. Journal of Anxiety Disorders 22, 117–125.

Berger, C.R., Calabrese, R.J., 1975. Some explorations in initial interaction and beyond: toward a developmental theory of interpersonal communication. Human Communication Research 1, 99–112.

Burton, R.A., 2008. On Being Certain: Believing You Are Right Even When You're Not. St. Martin's Griffin, New York.

Camerer, C., Weber, M., 1992. Recent developments in modelling preference: uncertainty and ambiguity. Journal of Risk and Uncertainty 5, 325–376.

Cerulo, K.A., 2006. Never Saw It Coming: Cultural Challenges to Envisioning the Worst. University of Chicago Press, Chicago, IL.

Cyert, Richard M., March, James G., [1964]1992. A Behavioral Theory of the Firm, second ed. Blackwell Business Publishing, Cambridge, MA.

Dequech, David, 2000. Fundamental uncertainty and ambiguity. Eastern Economic Journal 26, 41–60.

DiMaggio, Paul J., Powell, Walter W., 1983. The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. American Sociological Review 48 (2), 147–160.

Douglas, M., Wildavsky, A., 1983. Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers. University of California Press, Berkeley, CA.

Ellsberg, D., 1961. Risk, ambiguity, and the savage axioms. Quarterly Journal of Economics 75, 643–669.

Freeston, M.H., Rheaume, J., Letart, H., Dugas, M.J., Ladouceur, R., 1994. Why do people worry? Personality and Individual Differences 17, 791–802.

Frisch, D., Baron, J., 1988. Ambiguity and rationality. Journal of Behavioral Decision Making 1, 149–157.

Galbraith, J., 1973. Designing Complex Organizations. Addison-Wesley, Reading, MA.

Gao, G., Gudykunst, W.B., 1990. Uncertainty, anxiety, and adaptation. International Journal of Intercultural Relations 14, 301–317.

Gross, M., 2007. The unknown in process: dynamic connections of ignorance, non-knowledge and related concepts. Current Sociology 55, 742–759.

Hannan, M.T., Freeman, J., 1977. The population ecology of organizations. American Journal of Sociology 82 (5), 929–964.

Hirshleifer, J., Riley, J.G., 1992. The Analytics of Uncertainty and Information. Cambridge University Press, New York.

Hoffman, D.D., 1998. Visual Intelligence: How We Create What We See. W.W. Norton & Co., New York.

Kahneman, D., Tversky, A., 1982. Judgment under Uncertainty: Heuristics and Biases. Cambridge Unversity Press, New York.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47 (2), 263–292.

Kelsey, D., Quiggin, J., 1992. Theories of choice under uncertainty and ignorance. Journal of Economic Surveys 6 (2), 133–153.

Kerwin, A., 1993. None to solid: medical ignorance. Science Communication 15, 166–185.

Knorr-Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.

Lawrence, P.R., Lorsch, J.W., 1967. Organization and Environment: Managing Differentiation and Integration. Harvard Business School Press, Boston.

Lawson, T., 1985. Uncertainty in economic analysis. The Economic Journal 95, 909–927.

March, J.G., Simon, H., [1958]1993. Organizations, second ed. Blackwell Publishers, Cambridge, MA.

Merton, R.K., 1987. Three fragments from a sociologist's notebook: establishing the phenomenon, specified ignorance, and strategic research materials. Annual Review of Sociology 13, 1–28.

Milliken, F.J., 1987. Three types of perceived uncertainty about the environment: state, effect, and response uncertainty. Academy of Management Review 12 (1), 133–143.

Oreskes, N., Conway, E.M., 2011. Merchants of Doubt: How a Handful of Scientists Obscured the Truth from Tobacco Smoke to Global Warming. Bloomsbury Press, New York.

Proctor, R.N., Schiebinger, L. (Eds.), 2008. Agnotology: The Making and Unmaking of Ignorance. Stanford University Press, Stanford, CA.

Perrow, C., [1984]1999. Normal accidents: Living with High-Risk Technologies. Princeton University Press, Princeton, NJ.

Pfeffer, J., Salancik, G.R., 1978. The External Control of Organizations: A Resource Dependence Perspective. Harper & Row, New York.

Quiggin, J., 2009. Economists and uncertainty. In: Bammer, G., Smithson, M. (Eds.), Uncertainty and Risk: Multidisciplinary Perspectives. Earthscan, Sterling, VA.

Roberts, K.H., 1989. New challenges in organizational research: high reliability organizations. Organization and Environment 3 (2), 111–125.

Savage, L.J., 1954. The Foundations of Statistics. Wiley, New York.

Schrader, S., Riggs, W., Smith, R.P., 1993. Choice over uncertainty and ambiguity in technical problem solving. Journal of Engineering and Technology Management 10, 73–99.

Simmel, G., 1906. The sociology of secrecy and secret societies. American Journal of Sociology 11, 441–498.

Smithson, M., 1989. Ignorance and Uncertainty: Emerging Pardigms. Springer Press, New York.

Smithson, M., 2009. Psychology's ambivalent view of uncertainty. In: Bammer, G., Smithson, M. (Eds.), Uncertainty and Risk: Multidisciplinary Perspectives. Earthscan, Sterling, VA.

Sorrentino, R., Roney, C.J., 2000. The Uncertain Mind: Individual Differences in Facing the Unknown. Taylor & Francis, Philadelphia, PA.

Sorrentino, R.M., Short, J.C., 1986. Uncertainty orientation, motivation and cognition. In: Sorrentino, R.M., Higgins, E.T. (Eds.), The Handbook of Motivation and Cognition: Foundations of Social Behavior, vol. 1. Guilford Press, New York, pp. 379–403.

Sorrentino, R.M., Short, J.C., Raynor, J.O., 1984. Uncertainty orientation: implications for affective and cognitive views of achievement behavior. Journal of Personality and Social Psychology 46, 189–206.

Tannert, C., Elvers, H., Jandrig, B., 2007. The ethics of uncertainty. In the light of possible dangers, research becomes a moral duty. EMBO Reports 8 (10), 892–896.

Taleb, N.N., 2010. The Black Swan: The Impact of the Highly Improbable. Random House, New York.

Thompson, James D., 1967. Organizations in Action: Social Science Bases of Administrative Theory. Transaction Publishers, New Brunswick, CT.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation under uncertainty. Journal of Risk and Uncertainty 5, 297–323.

Van Asselt, M., Rotmans, J., 2002. Uncertainty in integrated assessment modeling: from positivism to pluralism. Climatic Change 54, 75–105.

Van de Boss, K., 2001. Uncertainty Management: The Influence of Uncertainty Salience on Reactions to Perceived Procedural Fairness. Journal of Personality and Social Psychology 80 (6).

Van de Boss, K., 2009. Making Sense of Life: The existential self trying to deal with personal uncertainty. Psychological Inquiry.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Weick, K., 1987. Organizational culture as a source of high reliability. California Management Review 29 (2), 112–127.

Wilson, T.D., Centerbar, D.B., Kermer, D.A., Gilbert, D.T., 2005. The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate. Journal of Personal and Social Psychology 88 (1), 5–21.

Zerubavel, E., 1993. The Fine Line: Making Distinctions in Everyday Life. University of Chicago Press, Chicago, IL.

Unconscious: History of the Concept

Donald Levy, City University of New York, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The history of the concept of the unconscious in western thought will be discussed here in three phases of its development: 1. from as early as the sixth century BCE to the seventeenth century, when the concept was not usually distinguished from other, wholly metaphysical concepts in which it was embedded; 2. the beginning of the attempt, under the influence of Descartes, to distinguish the concept of the unconscious from its metaphysical moorings and to submit it to scientific examination; 3. the modern period, in which the concept is very commonly assumed by researchers in the social sciences despite doubts by some about whether the concept's separation from its metaphysical origins is, or ever can be, complete.

Metaphysical Beginnings

The idea that people are sometimes unaware of their own mental processes while these are going on, and may even be incapable of later retrieving them introspectively, has had a long life of its own, independent of and prior to its uses in modern experimental psychology, psychoanalysis, cognitive science, and linguistics. However, the idea's early occurrences do not appear tagged by a single term translatable as 'the Unconscious'; rather, the idea commonly occurs as a corollary of (1) the rationalist doctrine of innate ideas, and more broadly, of (2) the even more ancient idea of the microcosm, the traditional world view according to which each mind contains the universe, in miniature, within itself (Levy, 1967). Since the microcosm concept can be thought of, at least in part, as the view that all ideas are innate, the rationalist doctrine, for example, of Descartes, for whom only some ideas, e.g., of God, or of thought, are innate, is implied by, but does not imply the microcosm view.

Since, presumably, human subjects are never consciously aware of having all ideas within their minds, the concept of the unconscious is implicit in the idea of the microcosm; and to the extent that one might never consciously reflect on (what Descartes meant by) the idea of God or of thought, the concept of the unconscious is implicit in the doctrine of innate ideas, as well. However, rationalists often accepted the microcosm model, too, as for example, Leibniz, who writes of souls that they are 'perpetual living mirrors', each representative of the entire universe; and Plato, who used the example of a young slave untrained in geometry, led only by questioning to the discovery of a mathematical truth, to show the boy in the process of recovering the knowledge from within himself, where the possibility of this recovery of knowledge (*Meno*, 85D, 4–8), which Plato called recollection (*anamnesis*), rests on the claim that "…as all nature is akin, and the soul has learned all things, there is no reason why we should not, by remembering but one single thing…discover everything else" (*Meno*, 81C, 10-D, 4). Since neither Plato nor Leibniz meant to say that human subjects could consciously experience themselves 'mirroring' the universe or consciously recall having known the recovered mathematical truth at some earlier time, the prototype of the concept of the unconscious implicitly employed by them is a thoroughly metaphysical one.

Descartes and the Beginning of Scientific Scrutiny

Descartes's views require separate comment, since, unlike Plato and Leibniz, he had no explicit attachment to the microcosm concept; yet although he had no difficulty with the existence of unconscious mental states insofar as his doctrine of innate ideas implied them (Cottingham, 1998), he has nevertheless gained the reputation of having defined them out of existence as a consequence of his mind–body dualism and his supposed identification of conscious awareness as the essential attribute of mind. L.L. Whyte goes so far as to write, "Descartes … by his definition of mind as awareness, may be said to have provoked as reaction, the European discovery of the unconscious mind … Descartes … had no clear conception of the unconscious mind, for to him all that is unconscious is physiological" (Whyte, 1962: p. 80).

But it is thought, not awareness that Descartes defines as essential to mind (*Meditation II*), and Descartes takes seriously the possibility, concerning his ideas of the external world, that there is some "faculty not yet fully known to me, which produces these ideas without any assistance from external things; this is, after all, just how I have always thought ideas are produced in me when I am dreaming" (*Meditation III*). Such a faculty in him, by implication unconsciously active, apparently would need an intellectual substance to inhere in (*Meditation VI*), so it cannot, presumably, be purely physiological in nature. What finally persuades Descartes that such a deceptive and unconsciously acting mental faculty does not exist is not the thought that the idea of it is incoherent in itself (as would be natural if his view were as Whyte supposed), but rather that once God's existence has been proved, the existence of that faculty is incompatible with God's perfect goodness (*Meditation V*).

What it was in Descartes's philosophy capable of provoking the somewhat desperate efforts to defend the as yet unnamed (Whyte, 1962: pp. 25, 60) concept of the unconscious against him can more properly be found in his rejection of the comforts of the microcosm model, and his insistence on accepting as true only things that are clearly and distinctly perceived, things for which there are 'sure signs' (*Meditation I*). For when such a standard is applied to the concept of the unconscious, its defenders as well as its detractors are at least temporarily at a loss, since no-one then knew, nor is there complete consensus even now, as to where the defining line

separating conscious from other mental states is to be drawn. How exactly it can be proved that an unconscious part of the mind exists in reality, i.e., that it is more than a mere conceptual possibility (if even that), and what sure signs ought to be required of specific contents in it, remain subjects of debate. The contrast on these points with the conscious mind and its states appears quite stark. It was Descartes who first made necessary the project of attempting to separate the concept of the unconscious from the metaphysical moorings that had until that time shrouded it from scrutiny. The concept of the unconscious was itself about to enter consciousness. The search was on for its manifestations in empirical phenomena that could be examined in a scientific way.

The Modern Period

By the end of the nineteenth century, a wide range of psychological phenomena had been identified as seeming to need unconscious mental activity for their explanation: ordinary perception and habit formation, hypnosis and posthypnotic suggestion, dreaming and neurotic symptom formation, to name a few. According to William James (1890/1983), however, all such 'proofs' of the existence of the unconscious must fail, since the very idea of the unconscious mind is an 'unintelligible and fantastical' notion, and more parsimonious explanations of the phenomena in question are in all cases available, e.g., explanations in terms merely of (1) "brain-processes to which no ideation seems attached," (2) "perceptions and volitions…performed consciously, only so quickly and inattentively that no *memory* of them remains," or (3) 'split-off' consciousness, without any reference to unconscious mental activity having to be invoked.

By the end of the twentieth century, the range of psychological phenomena offered in support of the concept of the unconscious had expanded to include language acquisition and 'subliminal perception', 'blindsight', and cognitive dissonance and self-perception attribution processes (Nisbett and Wilson, 1977; Reber, 1993). However, skeptics have continued to doubt the existence of unconscious mental activity entirely, for reasons roughly similar to those laid down by James, though with more sophisticated arguments (Searle, 1992; Strawson, 1994).

There is here an evident lack of prior agreement about what conditions a phenomenon must meet for it to be a genuine instance of unconscious mental activity; one side of the debate assumes that more (and more novel) examples are needed to make their case, while the other (minority) side is already committed to the view that no possible example can succeed. At least it can be said of those who use the sorts of cases mentioned to support the existence of unconscious mental activity that they usually agree about what criteria determine the contents of the unconscious in any given case, and of what those contents are; there is, e.g., no dispute among them about what stimulus is unconsciously perceived in any given subliminal perception experiment, or what forgotten order the subject is obeying in posthypnotic suggestion, even if there are many questions about the way these stimuli are unconsciously processed.

By contrast, the psychoanalytic evidence raises different problems on this point, since there is no agreement concerning what Freud thought the criterion of truth in psychoanalysis is, and of what the consequences of adopting one criterion or another would be. It is widely supposed (e.g., by Grünbaum, 1984; Rey, 1998) that Freud thought therapeutic success is the only (or best) evidence of the correctness of psychoanalytic interpretations. The acceptance of such a criterion of correctness seems to leave Freud vulnerable to the well-known criticism that therapeutic success could as easily be due to suggestion or placebo effects as to interpretive truth. By contrast, Wittgenstein (1966; Levy, 1996) attributes to Freud the view that assent by the subject of an interpretation is ultimately the sole, or main criterion of its correctness. Which criterion Freud accepted, and which he ought to have accepted, are questions impossible to settle here; however, in favor of the assent criterion as the answer to both questions, it can be said first that the assent criterion seems more in accord with Freud's own most extensive discussion of the matter (Freud, 1937). Indeed, he even considers there the possibility that in some cases the mark of a true 'construction' (i.e., interpretation) will be that the patient's symptoms and general condition are not improved but rather are aggravated (Freud, 1937: p. 265). In addition, that criterion also allows interpretations to be true outside the therapeutic situation, where the danger of suggestion and placebo effects can more easily be avoided. Of course, in cross-cultural anthropological testing of non-individual psychoanalytic hypotheses (e.g., on the Oedipus complex, Whiting et al., 1958; Stephens, 1962; Spiro, 1982), neither the assent nor therapeutic success criterion is directly applicable. Finally, for Freud, very generally, it is wishes and ideas, not desires an beliefs, that inhabit the psychoanalytic (i.e., repressed) unconscious (Freud, 1893: p. 251, 1895: p. 10, n1, 1915a: p. 152, 1915b: p. 178). However, ideas and wishes, whether conscious or unconscious, lack the tight connection with behavior to be found in desire and belief; in many cases at least, it is plausible to say that a subject's ideas and wishes are ascertainable only by way of the assent and self-ascription the subject provides, whereas the comparable claim about desires and beliefs (i.e., when these are combined), would have very little plausibility. This provides further evidence that Freud accepted the assent criterion for individual interpretations, after all. Regardless of which criterion for individual interpretations is accepted, the question remains whether unconscious ideas and wishes as referred to psychoanalytically provide genuine exceptions to James's requirements. Without trying to resolve this matter here, it should be noted that the presence or absence of a brain process cannot hinge on the assent of the person whose brain it is; so if the assent criterion were assumed, explanations of James's type (1) would be foreclosed. How, if at all, unconscious contents as psychoanalytically interpreted, might also elude his explanation types (2) and (3) remains to be seen. More broadly, whether psychoanalytic interpretations are capable of experimental testing, and how (if at all) to resolve the difficulties that arise when such testing is attempted are also unanswered questions (Edelson, 1984; Erdelyi, 1985; Erwin, 1996; Levy, 2004).

However, at no point were these or any other epistemological difficulties able to deter philosophical, artistic, and even scientific activity that rested on ideas of the unconscious mind (Ellenberger, 1970; Kandel, 2012). Beginning long before Freud's time, and continuing in the Romantic culture of

nineteenth century Europe right up to the present, such ideas have figured prominently in the philosophies of Hegel, Schopenhauer, and Nietzsche (Mills, 2002; Gardiner, 1967; Lehrer, 1994), in the musical, literary, and visual works and critical theories of many artists (Wintle, 1992; Abrams, 1953; Hopkins, 2004), and continue to be influential in modern critical practice (Tennenhouse, 1976; Fuller, 1988; Gabbard, 2001). Nevertheless, it is also true that from the end of the nineteenth century and for much of the century that followed, in the world of science and the academy, skepticism about the reality of unconscious mental activity and the intellectual legitimacy of even discussing it was widespread outside of psychoanalytic circles. Thus, such diverse thinkers as William James (1890/1983), Jean-Paul Sartre (1943), John Searle (1992), and Galen Strawson (1994) share a view that goes far beyond mere uncertainty about the matter, although both James and Sartre later altered their hostile stance (James, 1902: pp. 214–215; Sartre, 1959). Yet by the end of the twentieth century, the question was rarely raised, so general was the shift away from denying or doubting to conceding the existence of the unconscious, even if what was individually believed varied significantly, beyond the mere agreement concerning its existence. What explains this shift from rejecting to embracing the assumption of the unconscious is not obvious. Conceivably, ideas about the unconscious in the arts and in popular culture may have played a part. It seems unlikely that any direct influence from psychoanalysis alone is the whole explanation. Certainly, the decline of Behaviorism (Graham, 2010) and growing doubts concerning the presuppositions about verification on which it rested (Levy, 1996: pp. 83–127) contributed to the shift. But it was the rise of cognitive science and neuropsychology in the second half of the twentieth century that most clearly turned the tide. However, to say this is simply to raise the question in a different way; for given the background hostility in place before they appeared, why did these new hybrid disciplines with their openness to conceptions of unconscious mental activity arise in the first place? Perhaps at present the most that can be said with certainty is that on all sides non-psychoanalytic psychologists in their own work unwittingly began to echo James Strachey's observation about Freud that he "found that without making that assumption he was unable to explain or even to describe a large variety of phenomena which he came across. By making it, on the other hand, he found the way open to an immensely fertile region of fresh knowledge" (Freud, 1915b: Editor's Note 162).

This is not to say that Freud's very special conception of the unconscious had achieved canonical status beyond the sphere of psychoanalysis. According to Freud, the most basic distinction concerning the unconscious separates those unconscious mental processes that have been repressed from those that have not (Freud, 1915b: pp. 166, 172), and although he thought both types of unconscious processes are capable of acting on consciousness (Freud, 1915b: p. 168), it is only the repressed type that concerned him very much. Since researchers in cognitive science and neuropsychology rarely focus on repression if they even recognize its existence (but for exceptions see Ramachandran and Blakeslee, 1998; Ramachandran, 2011), for most purposes it is reasonable to treat Freud's dichotomy as one that is nearly universally subscribed to. The essence of repression, Freud remarks "…lies simply in turning something away, and keeping it at a distance, from the conscious" (Freud, 1915b: p. 147), but this claim by itself is not adequate to distinguish repression from, for example, the fully conscious refusal to attend to a thing simply because it is a distraction – which is not what he means. Repression was preceded in Freud's early work on hysteria by the idea of defense, and the relation between repression and defense remained tangled, perhaps fruitfully so, since unlike mere avoidance of distraction, defense is something we attempt only against perceived threats and attacks, which is much closer to Freud's idea, but still makes repression sound like something that occurs completely on the conscious level. The term 'repression' in a psychological (i.e., non-political) sense can be found before Freud and some authors have suggested possible sources for Freud's use of the term (on Herbart, e.g., see Jones, 1953: vol. I, p. 309; Andersson, 1962: pp. 116–117); nevertheless, "it was as a clinical *datum* that repression imposed itself from Freud's earliest treatment of hysterics onwards. Freud found that his patients did not have certain memories at their disposition, although these were perfectly vivid once they had been recalled" (Laplanche and Pontalis, 1973: p. 392). It might appear as if Freud was struggling to formulate this *datum* satisfactorily, but that does not seem to be the case. His theory of repression sees it as occurring in two phases, the first of which takes place on the border between the conscious (or 'preconscious') and unconscious (Freud, 1915b: p. 180), as his first use of the term 'repression' in print suggests "…it was a question of things which the patient wished to forget, and therefore intentionally repressed from his conscious thought and inhibited and suppressed" (Breuer and Freud, 1895: p. 10). If repression is a partly conscious mental act at its start, it does not stay that way, he thought. 'Repression proper', as he called the second phase of its development, involved the fully unconscious repressed idea, now out of conscious control, 'putting out derivatives' (Freud, 1915a: p. 149), forming associations with other ideas, taking on extreme, distorted forms of expression (Freud, 1915a), and exerting continuous pressure in the direction of the conscious by way of its derivatives against the force of repression (Freud, 1915a: p. 151). Freud could lay down no general rule governing the degree of distortion or remoteness from its source that an unconscious derivative needed if it is to successfully affect the 'delicate balancing' by which the resistance it faces in its pursuit of becoming conscious is controlled (Freud, 1915a: p. 150). The repression those derivatives are subjected to may not entirely succeed (Freud, 1915a: p. 153), Freud held. Such failure of repression is possible, he thought, because the idea and the emotions that are attached to it can undergo separate transformation when subjected to repression. If those emotions are transformed into conscious pain or anxiety even as the idea remains unconscious, the repression has failed. One such failure of repression, Freud thought, can be seen in neurotic symptoms, which are derivatives that mark, in action, a 'return of the repressed' to consciousness (Freud, 1915a: p. 154). Another failure of repression can be found in the verbal 'free associations' human subjects produce, which are derivatives of repressed ideas sufficiently distorted or far removed from their source to enable them to gain access to consciousness. These are 'substitute formations' (Freud, 1915a) by means of which the repressed idea can be consciously reconstituted, Freud thought (Freud, 1915a: p. 150). His obscure reference to the 'delicate

balancing' governing the success or failure of repression that unconscious derivatives encounter may allude to an unresolved problem in his conception of repression and symptom formation as well, for in his view neurotic symptoms "are the outcome of a conflict…The two forces which have fallen out meet once again in the symptom and are reconciled, as it were, by the compromise of the symptom that has been constructed. It is for that reason, too, that the symptom is so resistant: it is supported from both sides" (Freud, 1916–1917: pp. 358–359). Thus, for Freud, every symptom is a 'compromise-formation', a claim whose universality has been questioned (by Laplanche and Pontalis, 1973: p. 76) on the grounds that symptomatic phenomena are sometimes encountered clinically in which no compromise between defense and wish is apparent – instead, one or the other seems to dominate. However, according to Morris Eagle, in an assertion perhaps consistent with Freud's compromise-formation view of symptoms, we need to recognize that repression itself has both adaptive and maladaptive functions; "repression is both normal and pathogenic" (Eagle, 2002: p. 484).

The fundamental procedure for investigating the repressed unconscious, according to Freud, free association, requires the patient to 'abandon conscious purposive ideas' (Freud, 1900–1901: p. 531), that is, all goal-directed bodily and mental activity, while inhibiting as well attention to the social and practical constraints we impose on our actions when we consciously pursue goals of any kind. The requirement that all activity be suspended has the effect of releasing the mind from everything that normally works to distract it from itself. Then, the method of free association requires that voice must be given to all thoughts entering the mind, which means that considerations of politeness and social acceptability, for example, are suspended, as well as considerations of what is necessary or rational or appropriate in the solution of practical problems, whether physical or mental in nature. Thus, ideally, all inferring, planning, calculating, summarizing, organizing of one's thoughts is consciously inhibited. The ideas that then occur and which the patient gives voice to are 'free' insofar as they are presumably not governed by those social and practical constraints that we normally rely upon in regulating our conscious thinking and acting. However, since the ideas produced in that state have entered consciousness, but are also uncoupled from conscious activities, Freud viewed the ideas we then give voice to as derivatives ('substitute formations') of the repressed ideas which he identified as their source. Since such free associations have partially eluded the force of repression just by entering consciousness, they are not the repressed ideas themselves. But since those free associations are presumed to be detached from the patient's conscious mental or physical activity, their very freedom from connection to such activity allows them to be used as evidence of the repressed ideas from which they came, Freud inferred.

Is it realistic to suppose that anyone can totally suspend conscious purposive ideas as is, required for free association to lead to the recovery of repressed ideas? There is, perhaps, an element of benign deception practiced on the patient here; Freud writes "When I instruct a patient to abandon reflection of any kind and to tell me whatever comes into his head, I am relying firmly on the presumption that he will not be able to abandon the purposive ideas inherent in the treatment and I feel justified in inferring that what seem to be the most innocent and arbitrary things which he tells me are in fact related to his illness. There is another purposive idea of which the patient has no suspicion – one relating to myself" (Freud, 1900–1901: pp. 531–532). The 'deception' here is benign because it is easily practiced on oneself, as Freud himself had done, for example, in using free association to interpret his dreams to his own satisfaction; a shift in the focus of attention is needed here, where there is no treatment in question. So free association, in Freud's view, is a way of generating valuable new evidence that enables the gaps in our understanding of otherwise obscure mental occurrences to be filled.

But is it perhaps more than that? The idea that there is some kind of internal relationship, between the psychoanalytic project and the practice of free association is hard to avoid. It is fair to say that without the technique of free association and the new kind of evidence it summons up, Freud's belief that he had discovered a new method for *analyzing* dreams, jokes, slips, and neurotic symptoms would be very much harder to support. (Freud later added the choice of a profession, of a love-object and of a subject-matter by the creative artist to the list of things suitable for analytic interpretation.) For free association does seem to offer a way of breaking down any individual instance of each of these mental formations into parts and inferring – or ideally, recollecting – the unconscious origins of each, by applying free association to the results; the chain of associations may be long, but there is no reason to suppose it is interminable. The method seems to produce impressive results especially when applied to the interpretation of dreams, which is where Freud began to apply it almost as soon as it was suggested by one of his female patients in regard to her own symptoms. Of course, dreams, unlike symptoms are not normally involved in inhibiting or generating behavior in the way that symptoms often are, which means that inhibiting attention to the 'social and practical constraints' is easier with dreams than with symptoms.

By means of free association, we can temporarily and partially elude the distractions of normal life as well as the force of repression by which we disguise from ourselves the real nature of many of our conscious thoughts and actions. The task of the analyst would be unthinkably greater, if free association were impossible, as is usual with long-dead historical figures, although Freud thought Leonardo da Vinci was an exception about whom he could draw some interpretive inferences, precisely because Leonardo's private notebooks contain numerous drawings and jottings, including slips, which Freud thought could be treated as if they were free associations and analyzed as such (Freud, 1910). To see how valuable in the task of analysis is the evidence from living human beings provided by free association, let us imagine that some people responded to the request for them to free associate with puzzlement. That is, suppose that to them, having ideas while engaged in no conscious, purposive *activity*, whether physical or mental, were entirely empty – they never had such ideas to report, and were puzzled by our expectation that anyone ever did (dreams would therefore be beyond their experience). So although they told jokes, made slips, and exhibited neurotic symptoms, the only ideas they have ever had were ideas bearing on actions, e.g., what they could or should do when they became active; they seemed unable to even conceive of disengaging the having of conscious ideas from physical or mental purposive activity.

It is hard to see how such people (who would appear to us to be woefully deficient in imagination) could engage in free association, or how they could be psychoanalyzed; the best that could be managed in their case would apparently be to draw rough analogies between their actions and our own. On this view, then, free association would be integral to psychoanalysis since it would have to be seen as providing much more than merely collateral evidence for the truth of psychoanalytic interpretations; it would not be simply one heuristic device among others for exploring the repressed unconscious.

When free association is displayed in this way as in effect indispensable to the practice of, psychoanalysis, it then becomes possible to place psychoanalysis within a larger historical framework, one that embraces the psychological contributions of many diverse cultures, especially those of the 'Axial Age' as the period around 700 to 200 BCE in Europe, Asia, and the Middle East has come to be called. Developments within the Asian branch of that pivotal epoch, especially in the cultures of India and China, seem most relevant, given the striking similarities between psychoanalysis and the spiritual disciplines of that time and place. Prominent in all those disciplines was meditation in its various forms, and it is the similarities between free association and meditation that concern us, for in each case, the meditative technique was seen to be integral to the discipline, just as free association is integral to psychoanalysis. Furthermore, all such techniques share a common starting point, i.e., the intentional inhibition of goal-directed activity, just as in free association, and the focusing of one's attention on the nature of that attention itself becomes the source of very different prescriptions concerning its use. These analogies and many more began to interest psychoanalysts in the second half of the twentieth century, and that interest has continued to grow, with no sign of stopping.

Any historical account of such techniques and their respective disciplines would have to begin with Yoga, which shares inhibited action, both mental and physical, as its starting point with free association in psychoanalysis. In this respect, it is like the Alexander Technique, another modern discipline that focuses on correction and improvement of the use of the body (Alexander, 1988). The role of inhibited action in these disciplines as a heuristic device for sensitizing the practitioner to the habitual distortions of posture and movement unwittingly acquired over a lifetime strongly resembles the role of such inhibition in free association, which can also be viewed as a practice whose purpose is sensitization. But Yoga in its original form was very different, and the relevance of that version of it to psychoanalysis is limited, since to start with Yoga seems to have consisted in concentrating attention in order to achieve a trance state of total oblivion, which was held to be the necessary condition for escaping from karma with its cycle of rebirths, and attaining oneness with Brahman, the universal Absolute (Wynne, 2007: p. 82). On that view, there would be little room for very much of an analogy with free association, although inhibiting attention to the constraints governing activity, which must be a part of the path to oblivion, is exactly what free association requires as well. Possibly ancient Yoga had hidden dimensions that would make a connection with psychoanalysis more evident than it is at present.

The inventor, or discoverer, of meditation in a form that does have direct connections to free association, was the Buddha, who self-consciously sought to replace the Yoga of his time with a very different practice, one no longer focused on attaining oblivion. Instead, self-observation and analysis is to be cultivated as a means to achieving Nirvana, the 'snuffing out' or cessation of 'grasping desire,' the kind of desire manifested, most prominently, in addictive behavior and the quest for revenge (Gombrich, 2009). Besides being a pioneer in psychology, the Buddha was a social reformer with a revolutionary moral agenda; thus, he defined Nirvana, which included release from the compulsion to repeat past habits, as requiring non-observance of all caste restrictions (Buddha, 1986, 2000). Of the many Asian Axial Age thinkers, the Buddha is certainly the one who has attracted the most attention from psychoanalysts (see, e.g., Epstein, 1995); as a consequence, it is impossible to deal here with even a small selection of them (but see Safran, 2003 for a representative sampling).

However, key ideas of other Axial Age thinkers, both Asian (e.g., Confucius and Lao Tzu) and non-Asian, have also been found relevant to psychoanalysis (Fingarette, 1963; Bion, 1994; Eigen, 2011). One of the main outstanding questions about the relation between psychoanalysis and all of these Axial Age thinkers is whether psychoanalysis is simply a revival of practices and ideas from that earlier creative period, or is instead a truly new development, and if it is something new, in what way? For example, the apparent absence of a social and political agenda in much contemporary psychoanalysis seems to mark a contrast with its Axial Age predecessors. Whether that absence, if it is real, is a source of strength or is instead a defect remains to be seen. Another dangling problem concerns Freud's espousal of determinism, which threatens to subvert the entire therapeutic project, according to A. MacIntyre, because if it is accepted, there will be no essential difference between "the suffering neurotic and the successfully analysed patient", since both are subject to the "omnipresent causation exerted upon conscious life by the unconscious" (MacIntyre, 2004: pp. 112–113). There is at present no generally accepted response to this problem, either within the psychoanalytic community or outside it.

See also: Causal Analysis, History of; Mind–Body Dualism; Psychoanalysis, History of; Psychology, History of (Twentieth Century); Psychotherapy, History of: Psychiatric Aspects.

Bibliography

Abrams, M.H., 1953. The Mirror and the Lamp: Romantic Theory and the Critical Tradition. Oxford University Press, Oxford.

Alexander, F.M., 1988. Man's Supreme Inheritance: Conscious Guidance and Control in Relation to Human Evolution in Civilization. Centerline Press, Long Beach, CA.

Andersson, O., 1962. Studies in the Prehistory of Psycho-Analysis. Humanities Press, Atlantic Highlands, New Jersey.

Bion, W., 1994. In: Bion, F. (Ed.), Cogitations. Karnac, London.

Buddha, 1986. A Path of Righteousness: Dhammapada. Translation and Commentary by D.J. Kalupahana. University Press of America, Lanham, MD.

Buddha, 2000. The Dhammapada. Translation with Introduction and Notes, J.R. Carter and M. Palihawadana. Oxford University Press, Oxford.

Chomsky, N., 1986. Knowledge of Language. Praeger, New York.

Cottingham, J., 1998. Philosophy and the Good Life: Reason and the Passions in Greek, Cartesian and Psychoanalytic Ethics. Cambridge University Press, Cambridge.

Eagle, M., 2002. Repression. In: Irwin, E. (Ed.), The Freud Encyclopedia: Theory, Therapy and Culture. Routledge, New York, pp. 478–492.

Edelson, M., 1984. Hypothesis and Evidence in Psychoanalysis. University of Chicago Press, Chicago.
Eigen, M., 2011. Contact with the Depths. Karnac Books, London.
Ellenberger, H.F., 1970. The Discovery of the Unconscious: The History and Evolution of Dynamic Psychiatry. Basic Books, New York.
Epstein, M., 1995. Thoughts without a Thinker; Psychotherapy from a Buddhist Perspective. Basic Books, New York.
Erdelyi, M., 1985. Psychoanalysis: Freud's Cognitive Psychology. W.H. Freeman, New York.
Erwin, E., 1996. A Final Accounting: Philosophical and Empirical Issues in Freudian Psychology. MIT Press, Cambridge, MA.
Fingarette, H., 1963. The Self in Transformation: Psychoanalysis, Philosophy and the Life of the Spirit. Basic Books, New York.
Freud, S., 1893. The Complete Letters of Sigmund Freud to Wilhelm Fliess 1887–1904. Harvard University Press, Cambridge, MA.
Freud, S., 1895/1958. Studies on hysteria. In: Breuer, J., Freud, S. (Eds.), The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 2 (J. Strachey et al. Ed. and Trans.). Hogarth Press, London.
Freud, S., 1900–1901. The interpretation of dreams. In: J. Strachey, Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 5. W.W. Norton, New York.
Freud, S., 1910. Leonardo Da Vinci and a Memory of His Childhood. In: Standard Edition, vol. 11.
Freud, S., 1915a. Repression. In: Standard Edition, vol. 14.
Freud, S., 1915b. The Unconscious. In: Standard Edition, vol. 14.
Freud, S., 1916–1917. Introductory Lectures on Psycho-Analysis. In: Standard Edition, vol. 16.
Freud, S., 1937. Constructions in Analysis. In: Standard Edition, vol. 23.
Fuller, P., 1988. Art and Psychoanalysis. Hogarth Press, London.
Gabbard, G.O. (Ed.), 2001. Psychoanalysis and Film. Karnac Books, London.
Gardiner, P., 1967. Schopenhauer. Penguin Books, London.
Gombrich, R., 2009. What the Buddha Thought. Equinox, London.
Graham, G., 2010. Behaviorism. In: Zalta, Edward N. (Ed.), The Stanford Encyclopedia of Philosophy (Fall edition). http://plato.stanford.edu/archives/fall2010/entries/behaviorism/.
Grünbaum, A., 1984. The Foundations of Psychoanalysis: A Philosophical Critique. University of California Press, Berkeley.
Hopkins, D., 2004. Dada and Surrealism: A Very Short Introduction. Oxford University Press, Oxford.
James, W., 1890/1983. The Principles of Psychology. Harvard University Press, Cambridge, MA.
James, W., 1902. The Varieties of Religious Experience. Vintage Books, New York.
Jones, E., 1953. The Life and Work of Sigmund Freud, vol. I. Basic Books, New York.
Kandel, E.R., 2012. The Age of Insight: The Quest to Understand the Unconscious in Art, Mind and Brain [From Vienna 1900 to the Present]. Random House, New York.
Laplanche, J., Pontalis, J.-B., 1973. The Language of Psychoanalysis (D. Nicholson-Smith, Trans.). W.W. Norton, New York.
Lehrer, K., 1994. Nietzsche's Presence in Freud's Life and Thought: On the Origins of a Psychology of Dynamic Unconscious. State University of New York Press, Albany, NY.
Levy, D., 1967. Macrocosm and microcosm. In: Edwards, P. (Editor-in-chief), The Encyclopedia of Philosophy, vol. 5. Macmillan and the Free Press, New York, pp. 121–125.
Levy, D., 1996. Freud among the Philosophers: The Psychoanalytic Unconscious and Its Philosophical Critics. Yale University Press, New Haven.
Levy, D., 2004. Incompleteness and experimental untestability in psychoanalysis. In: Mills, J. (Ed.), Psychoanalysis at the Limit: Epistemology, Mind, and the Question of Science. State University of New York Press, Albany, NY, pp. 177–196.
MacIntyre, A., 2004. The Unconscious: A Conceptual Analysis, revised edition. Routledge, New York.
Mills, J., 2002. The Unconscious Abyss: Hegel's Anticipation of Psychoanalysis. State University of New York Press, Albany, NY.
Nisbett, R.E., Wilson, T.D., 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 84 (3), 231–259.
Ramachandran, V.S., Blakeslee, S., 1998. Phantoms in the Brain. William Morrow, New York.
Ramachandran, V.S., Blakeslee, S., 2011. The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human. W.W. Norton, New York.
Reber, A., 1993. Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious. Oxford University Press, New York.
Rey, G., 1998. Unconscious mental states. In: Craig, E. (Ed.), Routledge Encyclopedia of Philosophy, vol. 9. Routledge, London, pp. 522–527.
Safran, J.D. (Ed.), 2003. Psychoanalysis and Buddhism: An Unfolding Dialogue. Wisdom Publications, Somerville, MA.
Sartre, J.P., 1943. L'Être et le néant. Translated by H.E. Barnes, Being and Nothingness, 1956. Philosophical Library, New York.
Sartre, J.P., 1959. Le Scenario Freud. Translated by Q. Hoare, The Freud Scenario, 1985. Verso, London.
Searle, J., 1992. The Rediscovery of the Mind. MIT Press, Cambridge, MA.
Spiro, M.E., 1982. Oedipus in the Trobriands. University of Chicago Press, Chicago, IL.
Stephens, W.N., 1962. The Oedipus Complex Hypothesis: Cross-Cultural Evidence. Free Press of Glencoe, New York.
Strawson, G., 1994. Mental Reality. MIT Press, Cambridge, MA.
Tennenhouse, L. (Ed.), 1976. The Practice of Psychoanalytic Criticism. Wayne State University Press, Detroit.
Weiskrantz, L., 1986. Blindsight: A Case Study and Implications. Oxford University Press, Oxford.
Whiting, J.W.M., Kluckhohn, R., Anthony, A., 1958. The function of male initiation ceremonies at puberty. In: Maccoby, E.E., Newcomb, T., Hartley, E.A. (Eds.), Readings in Social Psychology. Holt, New York.
Whyte, L.L., 1962. The Unconscious before Freud. Doubleday Anchor Books, Garden City, NY.
Wintle, C., 1992. Analysis and psychoanalysis: Wagner's musical metaphors. In: Paynter, J., et al. (Eds.), Companion to Contemporary Musical Thought, vol. 2. Routledge, London.
Wittgenstein, L., 1966. In: Barrett, C. (Ed.), Lectures and Conversations on Aesthetics, Psychology and Religious Belief. University of California Press, Berkeley.
Wynne, A., 2007. The Origin of Buddhist Meditation. Routledge, New York.

Underclass

Paul A Jargowsky and Lorraine C Minnite, Rutgers-Camden, Camden, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by W.J. Wilson, J.M. Quane and B.H. Rankin, volume 23, pp. 15945–15948, © 2001, Elsevier Ltd.

Abstract

The term 'underclass' came into wide use by academics, policymakers, and the press during the latter quarter of the twentieth century to refer a subset of the urban poor marked by their isolation in decaying ghettos of increasingly concentrated poverty, and their self-defeating behaviors in response to these conditions. Controversy followed, with critics challenging the term's racial and cultural implications, conceptual definition, and empirical measurement. Historical antecedents are found in the work of major political thinkers and propagandists such as Malthus, Marx, Bakunin, and Myrdal. The contentious debates of the last 30 years over the exact meaning and implications of the term diverted attention from the underlying phenomena of urban poverty, joblessness, and dependency. To the extent that the term is used today, it refers to groups who are structurally cut off from the pathways to social and economic mobility.

The controversial term 'underclass' has been used by both academics and the popular press to describe a subset of the poor. The controversy stems from a lack of agreement on the particular characteristics of the subset and the normative content of inclusion in the group. Moreover, the empirical strategies proposed to measure the size of the underclass were problematic and did not gain wide acceptance. In this article, we discuss the origins of the term, its conceptual definitions, attempts to measure the size of the underclass, and current usage of the term.

Historical Origins of the 'Underclass' Concept

As societies grapple with the problems of poverty, there is a long-standing practice of dividing the poor into those who are deserving of assistance and those who are not (Katz, 1989). The deserving poor are said to be destitute for reasons beyond their control, such as economic depressions or the death or disability of a breadwinner. In contrast, the undeserving poor are said to bring poverty upon themselves, by engaging in self-destructive behaviors and failing to embrace mainstream values. In colonial America, this was reflected in the distinction between poverty, 'an unavoidable evil,' in the words of Reverend Charles Burroughs preaching in 1834, and pauperism, "the consequence of willful error, of shameful indolence, of vicious habits" (Katz, 1989: p. 13).

Thomas Malthus, writing at the turn of the eighteenth century, expressed concern about what he saw as a 'redundant population.' "If we multiply too fast," he wrote in *An Essay on the Principle of Population* (1806), "we die miserably of poverty and contagious diseases" (Malthus, 1806; reprinted 1989: p. 88, cited in Morris, 1994: p. 10). Overpopulation, argued Malthus, was the result of early marriage, too many children, and too many children relative to life expectancy. It was the poor who suffered the most from overpopulation, and it was the poor who were least likely to understand the causes of their own poverty. "When the wages of labor are hardly sufficient to maintain two children, a man marries and has five or six. He finds himself miserably distressed," wrote Malthus, but, "the last person he would think of accusing is himself" (Malthus, 1806; reprinted 1989: p. 106, cited in Morris, 1994: p. 11). For Malthus, the problem was personal weakness in the face of natural passions and desires that require regulation, especially sexual behavior. The notion of moral restraint as a check on population growth and, therefore, the spread of poverty through a subculture of vice, is thus deeply implanted in attitudes toward the poor that locate the causes of immiseration and exclusion in the behavior of the immiserated rather than in the organization of economies or societies.

Ideas about social marginality and class exclusion as threats to social order expanded in the nineteenth century with the rise of industrialization, mass employment, and later, modern social science. In the *Eighteenth Brumaire of Louis Napoleon* (1852), Marx identified a 'dangerous class,' the *lumpen-proletariart*, to distinguish between revolutionary and reactionary strata or 'fractions' within the industrial working class. The *lumpen* were a rag-tag group of ruined refugees of the bourgeoisie, vagabonds, pimps, beggars, swindlers, pickpockets, and others of 'dubious means of subsistence and of dubious origin,' incapable of developing a class consciousness (Marx, 1963: p. 75). In Britain at this time, a related notion of a social 'residuum' emerged in debates over the extension of the suffrage. Welshman, for example, notes that in 1867, the radical Birmingham Minister of Parliament John Bright, in debating the merits of the Second Reform Act, argued for the enfranchisement of the respectable, skilled male working class, but also for the exclusion of a substrata of the unskilled and the 'feckless and criminal' poor "because they have no independence whatsoever," he said, and thus, were not worthy of the vote (Welshman, 2006: p. 7). Exclusion was achieved by legislation through the imposition of a modest householder income qualification.

It is a hallmark of the underclass idea that form its various early incarnations, it has been freighted with contentious political, social, and moral meaning. Contrary to Marx, for example, the influential nineteenth-century anarchist theorist Mikhail Bakunin located revolutionary potential among those who had been spared the harsh and degrading discipline of industrial work. Ideas about an 'industrious unemployed,' who

lost work as a result of upheaval and instability in the trade cycle, began to take hold as progress toward working-class enfranchisement in England unfolded. This can be seen in the emergence of scientific philanthropy in the form of the Charity Organization Society (COS) founded in London by Octavia Hill and Helen Bosanquet in 1869, and the associated movement of private charity societies that subsequently spread across Europe and the United States. Hill and other reformers in this movement believed that mere alms giving was not enough. What was needed was investigation and analysis of the deeper causes of poverty if dependency among the poor was to be averted. The COS introduced methods of modern records management, objective empirical investigative techniques that formed the early basis of social casework, and professional training of its 'visitors' to teach the deviant poor values of self-reliance, thrift, and hard work necessary to overcome the difficult hand they had been dealt. The COS did distinguish between the deserving and undeserving poor, but, importantly, a link had been made between economic cycles and the devastating consequences of unemployment, ideas that would reemerge among social scientists in the mid-twentieth century.

Gunner Myrdal (1963), often credited with being the first modern scholar to employ the underclass term, found that a growing population was unable to access employment or the skills and training necessary to participate in the regular labor market. Myrdal's concern, however, was principally with the effects of technological change on the workforce and chronic unemployment rather than the behavior of the poor. Myrdal believed that Americans were familiar with the causes of unemployment, but that the cultural emphasis on individualism and an optimistic belief in class mobility meant they had little understanding of how unemployment was trapping people at the bottom. Other scholars in the 1960s – such as Liebow (1967), Rainwater (1970), and Hannerz (1969) – wrote about a separate lower class culture, but they refrained from using the term 'underclass,' and poverty debates in the United States during this time were mostly consumed with Oscar Lewis's idea of a 'culture of poverty' (1959, 1969). Echoing Myrdal in one of the first scholarly treatments of the underclass in the modern period, Douglas G. Glasgow's *The Black Underclass: Poverty, Unemployment and the Entrapment of Ghetto Youth* likewise argued that "the underclass is distinguished from the lower class principally by its lack of mobility" (1980: p. 8). Whereas lower class status means you are on the bottom rung of the economic ladder, underclass status implies that the ladder is effectively unavailable to you. Entrapment in the underclass is driven by prolonged unemployment, inadequate education, and a "lack of connections to standardized institutions that act as feeder systems to the primary labor market" (Glasgow 1980: p. 7). The underclass is outside the regular channels of the class structure and, therefore, highly unlikely to be able to improve their position over time.

There was little more discussion of an underclass in academic circles until later in the 1970s, as the first troubling signs of the economic dislocation to come began to manifest in cities and communities facing plant closings, white flight, and government disinvestment. In the popular media, especially in the 1970s and 1980s, the term 'underclass' was used to describe the inner-city poor persons, often blacks, who exhibited any deviant or self-destructive behavior. The devastation wrought by the crack epidemic, the rapid rise in out-of-wedlock childbearing, and the surge in violent crime troubled academics, politicians, and the public. Vivid journalistic accounts of the harsh conditions of life in inner-city black ghettos (Auletta, 1982; Dash, 1989; Lemann, 1986a, 1986b; Wilkerson, 1987) emphasized crime, violence, drug use, dropping out of high-school, out-of-wedlock childbearing, and lack of connection to the formal labor market. Lemann emphasized increasing social disorganization in the black ghetto, beginning in the late 1960s, with the out-migration of a better-off black working and middle classes from the inner cities to the suburbs. The underclass, again, were trapped by unemployment in decaying cities left behind, their behavioral deviance an adaptation to their circumstances. This behavioral description of an urban underclass leads to assigning moral culpability for their poverty to those who participate in the ghetto lifestyle. "Underclass," argued Jacqueline Jones (1992: p. 9), "was but one more term used to label and objectify people believed to be morally dissolute and criminally inclined."

Popular and journalistic interest in a new underclass stimulated social science inquiry, as it had done in England a century before. William Julius Wilson's influential book, *The Truly Disadvantaged*, defined the underclass as those "who are outside the mainstream of the American occupational system" (Wilson, 1987: p. 8). This includes those who lack the training and skills to access available jobs and, therefore, experience long-term spells of unemployment or drop out of the labor force altogether. Wilson's work tied the emergence of the underclass to several distinct trends that came together in the inner-city to transform black ghettos. The inner-city black poor, who had always been segregated from whites, became increasingly segregated from the black middle-class due to expanded housing opportunities for economically successful blacks, depriving black ghettos of social capital, and increasing the social isolation of the ghetto poor. At the same time, deindustrialization and the suburbanization of job growth increased joblessness and reliance on alternative means of support, from public welfare to criminality. The result is a 'disproportionate concentration of the most disadvantaged segments of the urban black population,' which created a 'social *milieu*' that fostered low labor force attachment, out-of-wedlock childbearing, drug and alcohol use, and other social pathologies (Wilson, 1987: p. 58). Wilson's analysis addressed the growing divide between the behavior and lifestyle of the inner-city poor and the American mainstream. His analysis acknowledged the emergence of ghetto-specific patterns of behavior that were at variance with mainstream values, but he linked them to macrostructural transformations in the economy and the growing social isolation of the inner-city poor.

Critiques of the Underclass Concept

Wilson's work refocused liberal social science on the problem of urban poverty in the 1980s. It emphasized the devastating impact of broader changes in the American economy and unemployment on the black urban poor, but it also walked a fine line between cultural and structural explanations of poverty. Wilson did not explicitly reject the idea of a culture of poverty (Lewis, 1959, 1969), arguing that weight must be given

to culture, observed as behavior, because cultural practices reflect adaptation to structural conditions. Despite Wilson's efforts to keep the focus on macroeconomic change, his work sometimes was interpreted as an endorsement of the idea of the underclass as isolated from the mainstream by their values and behavior in ways that are damaging to their prospects of escaping the trap of 'ghetto culture' with its 'tangle of pathologies.' In response, the most vigorous criticism of the dominant concept of the underclass that emerged in the 1980s charged that framing the black urban poor as an underclass was supported empirically and marked a dangerous return to discredited notions of an 'undeserving poor' and a 'culture of poverty' as an explanation for their material deprivation.

One of the earliest critics of Wilson was Norman Fainstein (1987) who rejected Wilson's adoption of a language of class as an empirically unsupported means of falsely diminishing the importance of racial discrimination in explaining urban poverty. Fainstein also challenged the empirical validity and significance of the spatial mismatch hypothesis, an analysis relied on by underclass researchers to support their arguments that the inner-city poor were isolated socially (Fainstein, 1987, 1993). Adolph Reed joined Fainstein in criticizing the underclass concept as scientifically questionable and beyond that, as veiling racism in the sociology of the policy research community (Reed, 1990; see also Rosenthal, 1999; Bonilla-Silva, 2001; Feagin, 2001). Other scholars avoided analyzing the underlying attitudes or motivations of social scientists, but similarly found that Wilson ignored the salience of on-going racial discrimination as an explanation for increasingly concentrated poverty (Massey and Denton, 1993; Massey et al., 1994). Some further questioned whether the underclass met the definition of a class or represented a homogenous social grouping (Heisler, 1991; Herpin, 1993; Massey, 1993; Zelley, 1995; Wright, 1994; Martiniello, 1996; Marshall et al., 1996; Wacquant, 2008), and challenged the degree to which the urban poor, and specifically, inner-city blacks, were socially isolated (Oliver, 1988; Newman, 1992; Phillips, 2001) or collective adherents to morally repugnant values (Anderson, 1999).

More than any other scholar, Herbert Gans presented the most sustained critique of the underclass concept. In essays (1990, 1994) and a book, *The War Against the Poor* (1995), Gans argued that 'underclass' was simply a 'buzzword' for the undeserving poor and that 'the term has taken on so many connotations of undeservingness and blameworthiness that it has become hopelessly polluted in meaning, ideological overtone and implications, and should be dropped' (Gans, 1990: p. 272).

The debate, however, is not so much one of terminology, as it is Gans's deep concern with the effects on public opinion of the people so labeled and the policies that follow what for Gans is a pejorative misidentification of a social group. Gans argued that two novel elements in the Wilsonian conception of the underclass mark a departure from its more benign use by Myrdal and heighten the danger of separating groups from the rest of society by pointing selectively to their behavior. One is the term's racialized context, as the underclass is seen as almost entirely black and Hispanic, while racial minorities continue to face various forms of pernicious discrimination. The other is the elaboration of associated pathologies and antisocial behaviors that has the effect of inflaming public opinion and legitimizing a moral indictment of the poor. Moreover, observed behavioral patterns are assumed by most underclass researchers to be caused by norm violations rather than the conditions under which very poor people live. Methodological problems with measuring the underclass (discussed in the following section) are important to Gans's critique, but he goes further to warn of the consequences for civil liberties and democracy itself of a license to stigmatize and therefore scapegoat the disadvantaged.

Measurement Strategies

Ricketts and Sawhill (1988), working with a group of researchers at the Urban Institute, operationalized the underclass concept in a way that could be applied to census data so that national estimates of the phenomenon could be produced. Many of the elements of the conceptual definition of the underclass are not available within the census data, so it is inevitable that the operational definition for quantitative empirical work will be cruder than any particular theoretical conception of the underclass. They sought to identify "a subgroup … that engages in behaviors at variance with those of the mainstream" (Ricketts and Sawhill, 1988: p. 318). Their approach also had a geographic component, in that they identified underclass neighborhoods in which a high proportion of the residents engage in behaviors at variance with mainstream expectations. The neighborhood focus relates to the belief that when dysfunctional behaviors are concentrated, they become normative for the adults, and especially the children, growing up in these neighborhoods. Neighborhoods, proxied by census tracts in their analysis, were designated as underclass if they were simultaneously one standard deviation above the national mean on four characteristics: (1) the percentage of families with children headed by single females; (2) percentage of adult males whole worked fewer than 26 weeks or were not in the labor force; (3) percentage of households receiving public assistance; and (4) percentage of 16- to 19-year-olds not in school and not high-school graduates.

The Urban Institute underclass measure can be criticized on many grounds. For one thing, not all individuals who engage in underclass behaviors live in underclass neighborhoods and not all residents of underclass neighborhoods engage in underclass behaviors. Furthermore, Wilson (1988) criticized the measure on the ground that the second condition combined those who were not in the labor force (a choice) with those who had long spells of unemployment (an economic outcome). Even labor force participation is partly an outcome in the sense that discouraged workers withdraw from the labor market because they see no realistic prospect for employment. Hughes (1989) argued that the use of an underclass neighborhood threshold based on national norms rather than metropolitan norms was problematic, given regional differences in welfare benefit levels and macroeconomic conditions. He argued that the use of metropolitan norms, rather than national norms, would "focus on the concentration of deprivation within certain neighborhoods" rather than on differences in the four characteristics among metropolitan areas (Hughes, 1989: p. 192).

On the basis of 1980 means and standard deviations for the four indicators, Ricketts and Sawhill found that about 2.5 million (or about 1% of the US population) lived in the underclass areas. Virtually all the underclass tracts were in urban areas, disproportionately located in the older industrial cities, such as Newark, New York, and Baltimore. Residents of underclass areas were 59% black and 10% Hispanic. In comparison, 5.6 million persons lived in neighborhoods with poverty rates of 40% or more, about 3% of the US population. Thus, it is a mistake to assume that underclass areas are identical to or a subset of high-poverty areas – for example, the poorest of the poor neighborhoods. Nearly 4 in 10 of all underclass tracts in 1980 were not high-poverty areas, and 72% of these extreme poverty areas were not underclass tracts. Ricketts and Sawhill concluded that the two concepts, while highly correlated, are qualitatively distinct phenomena.

Using 1970 and 1980 Census Data, Ricketts and Mincy (1990) documented a dramatic growth in the underclass during the 1970s. They used 1980 means and standard deviations on the four indicators and applied these retrospectively to the 1970 census data. The number of underclass census tracts grew 331%, from 204% in 1970 to 880% in 1980. The population more than tripled, rising from 752 000 to 2.5 million (Ricketts and Mincy, 1990: p. 140). This rapid growth confirmed the general belief that social conditions of the inner cities, where almost all underclass tracts were located, were deteriorating rapidly during this period. Mincy and Wiener (1993) used the Ricketts–Sawhill definition to estimate the growth of the underclass population during the 1980s, by applying the 1980 thresholds to 1990 Census data. They report that 2.7 million people, 1.1% of the total US population, lived in such areas. These figures represent increases of less than 10% over the 1980 levels for both the number of tracts and their population (Mincy and Wiener, 1993: Table 1). Compared with the dramatic growth in the underclass during the 1970s, the population living in underclass areas increased by only 8% in the 1980s (Mincy and Wiener, 1993: Table 1), while the population living in concentrated poverty areas rose by more than 54% (Jargowsky, 1997: p. 38). Once again, the evidence suggests that the underclass phenomenon differs from concentration of poverty and must be considered a distinct dimension of the urban poverty problem.

The most recent analysis suggests that, by the Urban Institute measure at least, the underclass declined dramatically in the 1990s. Applying 1990 means and standard deviations to the 1990 and 2000 census tract data, Jargowsky and Yang (2006) found a substantial decline in the number of underclass census tracts (−32.5%) and the population living in them (−36.4%). The decrease was driven largely by an increase in high-school graduation rates and a decrease in the receipt of public assistance, the latter being driven by the 1996 welfare reform enacted under President Clinton that substantially reduced welfare receipt. These changes reduced the correlation between the four underclass measures, therefore reducing the probability that any one census tract would be one standard deviation above the mean on all four measures. Although the sensitivity of the underclass measure to correlations among the four measures is problematic, the drop in the underclass phenomenon is consistent with other measures of conditions in inner-city neighborhoods, such as declines in violent crime rates in most major cities during this period (Jargowsky and Sawhill, 2006).

Current Status of the Underclass

The concept of the underclass has largely fallen from use since its heyday among social scientists in the 1980s and 1990s. Despite the controversy over its freighted moral meanings, empirical imprecision, and association with discredited notions of cultural causes of poverty, the term retains a latent power for some. As the cause of alleviating inner-city poverty has declined in the United States, and politics have turned against the poor, yet a new set of conditions reflecting deeply rooted racial and class inequality has arisen, and the underclass idea appears to be making a mild return. The crisis of incarceration and the militarization of the US border have expanded the pool of the poor now legally confined, restricted, and excluded from economic and education opportunity. Under the structural conditions of legal regimes that confer second-class citizenship (rather than simply a function of capitalist exploitation), some – including former critics of Wilson's treatment of the underclass concept – have returned to the notion of a racialized subset of the poor permanently trapped in a subordinate class that threatens the well-being and status of coethnics and even the rest of society. For example, Massey and Pren (2012) argued that current US immigration policies are leading to the emergence of a new Latino underclass. Approximately half of all Mexicans in the United States as of 2012 are undocumented, and therefore they are cut off from mainstream employment opportunities, social services, and public higher education. "The burden of illegality," Massey and Pren wrote, "constitutes the single largest and most potent barrier to Hispanic socioeconomic mobility and integration in the United States" (2012: p. 15).

In the final analysis, the term 'underclass' is seldom used in the twenty-first century as an analytic concept. The contentious debates over the exact meaning and normative implications of the term diverted attention from the underlying phenomena of urban poverty, joblessness, and dependency. To the extent that the term is still used, it refers to groups who are cut off structurally from the pathways to social and economic mobility. At least two groups satisfy this condition – undocumented immigrants, who cannot access the mainstream labor market or many social services, and the ex-offender population, which likewise is cut off from employment opportunities and other vestiges of true citizenship, such as voting. Public policy will need to address these and other structural impediments to economic mobility, or the nation will be at risk of creating a permanent underclass of people with no hope of escaping from poverty.

See also: Culture of Poverty: Critique; Poverty Law: United States; Poverty Policy; Poverty in History; Poverty, Culture of; Poverty, Geography of; Urban Poverty in Neighborhoods.

Bibliography

Anderson, Elijah, 1999. The Code of the Streets: Decency, Violence, and the Moral Life of the Inner City. W.W. Norton & Co, New York.

Auletta, Ken, 1982. The Underclass. Random House, New York.
Bonilla-Silva, Eduardo, 2001. White Supremacy and Racism in the Post-civil Rights Era. Lynne Reinner, Boulder, CO.
Dash, Leon, 1989. When Children Want Children. William Morrow, New York.
Fainstein, Norman, 1987. The underclass/mismatch hypothesis as an explanation for black economic deprivation. Politics and Society 15 (4), 403–451.
Fainstein, Norman, 1993. Race, class and segregation: discourses about African Americans. International Journal of Urban and Regional Research 17 (3), 384–403.
Feagin, Joe, 2001. Racist America: Roots, Current Realities, and Future Reparations. Routledge, New York.
Gans, Herbert J., 1990. Deconstructing the underclass: the term's dangers as a planning concept. Journal of the American Planning Association 56 (3), 271–277.
Gans, Herbert J., 1994. Positive functions of the undeserving poor: uses of the underclass in America. Politics and Society 22 (3), 269–283.
Gans, Herbert J., 1995. The War against the Poor: The Underclass and Anti-poverty Policy. Basic Books, New York.
Glasgow, Douglas C., 1980. The Black Underclass: Poverty, Unemployment, and Entrapment of Ghetto Youth. Vintage Books, New York.
Hannerz, Ulf, 1969. Soulside: Inquiries into Ghetto Culture and Community. Columbia University Press, New York.
Heisler, Barbara S., 1991. A comparative perspective on the underclass: questions of urban poverty, race and citizenship. Theory and Society 20 (4), 455–483.
Herpin, Nicolas, 1993. The urban underclass according to American sociologists: social exclusion and poverty. Revue francaise de sociologie 34 (3), 421–439.
Hughes, Mark A., 1989. Misspeaking truth to power: a geographical perspective on the 'underclass' fallacy. Economic Geography 65, 187–207.
Jargowsky, Paul A., 1997. Poverty and Place: Ghettos, Barrios, and the American City. Russell Sage Foundation, New York.
Jargowsky, Paul A., Sawhill, Isabelle, 2006. The Decline of the Underclass. Policy Brief #36, Center on Children and Families. The Brookings Institution, Washington DC.
Jargowsky, Paul A., Yang, Rebecca, 2006. The 'underclass' revisited: a social problem in decline. Journal of Urban Affairs 28, 55–70.
Jones, Jacqueline, 1992. The Dispossessed: America's Underclasses from the Civil War to the Present. Basic Books, New York.
Katz, Michael, 1989. Undeserving Poor. Random House Digital, Inc.
Lemann, Nicholas, 1986a. The origins of the underclass, part 1. Atlantic Monthly, 31–55.
Lemann, Nicholas, 1986b. The origins of the underclass, part 2. Atlantic Monthly, 54–68.
Lewis, Oscar, 1959. Five Families: Mexican Case Studies in the Culture of Poverty. Basic Books, New York.
Lewis, Oscar, 1969. The culture of poverty. In: Moynihan, Daniel P. (Eds.), On Understanding Poverty. Basic Books, New York.
Liebow, Elliot, 1967. Tally's Corner: A Study of Negro Streetcorner Men. Little, Brown & Company, Boston.
Marshall, Gordon, Roberts, Stephen, Burgoyne, Carole, 1996. Social class and underclass in Britain and the USA. The British Journal of Sociology 47 (1), 22–44.
Malthus, Thomas R., 1989. In: James., P. (Ed.), An Essay on the Principle of Population, vol. 2. Cambridge University Press, Cambridge.
Martiniello, Marco, 1996. Ethnicity, 'race' and social class: the debate on the subject of the underclass in the United States and Europe. The Tocqueville Review/La Revue Tocqueville 17 (2), 117–133.
Massey, Douglas S., 1993. Latinos, poverty, and the underclass: a new agenda for research. Hispanic Journal of Behavioral Sciences 15 (4), 449–475.
Massey, Douglas S., Denton, Nancy, 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge.
Massey, Douglas S., Eggers, Mitchell L., Denton, Nancy, 1994. Disentangling the causes of concentrated urban poverty. International Journal of Group Tensions 24 (3), 267–316.
Massey, Douglas S., Pren, Karen A., 2012. Origins of the new Latino underclass. Race and Social Problems 4 (1), 5–17.
Mincy, Ronald B., Wiener, Susan J., 1993. The Under Class in the 1980s: Changing Concepts, Constant Reality. The Urban Institute, Washington, D.C.
Morris, Lydia, 1994. Dangerous Classes: The Underclass and Social Citizenship. Routledge, London.
Myrdal, Gunnar, 1963. Challenge to Affluence. Pantheon Books, New York.
Newman, Katherine S., 1992. Culture and structure in *The Truly Disadvantaged*. City and Society 6 (1), 3–25.
Oliver, Melvin L., 1988. The urban black community as network: toward a social network perspective. The Sociological Quarterly 29 (X), 623–645.
Phillips, Melanie, 2001. What about the overclass? The Public Interest 145, 38–43.
Rainwater, Lee, 1970. Behind Ghetto Walls: Black Families in a Federal Slum. Aldine Publishing Company, Chicago.
Reed, Adolph, 1990. The underclass as myth and symbol: the poverty of discourse about poverty. Radical America 24 (1), 21–40.
Ricketts, Erol R., Mincy, Ronald, 1990. Growth of the underclass, 1970–1980. Journal of Human Resources 25 (1), 137–145.
Ricketts, Erol R., Sawhill, Isabel V., 1988. Defining and measuring the underclass. Journal of Policy Analysis and Management 7, 316–325.
Rosenthal, Steven J., 1999. How liberal ideology assists the growth of fascism: a critique of the sociology of William Julius Wilson. Journal of Poverty 3 (2), 67–87.
Wacquant, Loic. J.D., 2008. Urban Outcasts: A Comparative Sociology of Advance Marginality. Polity Press, Malden, Mass.
Welshman, John, 2006. Underclass: A History of the Excluded, 1880–2000. Hambledon Continuum, London.
Wilkerson, Isabel, 1987. New studies zeroing in on poorest of the poor. New York Times. December 20 and 26.
Wilson, William J., 1987. The Truly Disadvantaged: The Inner-City, the Underclass and Public Policy. University of Chicago Press, Chicago.
Wilson, William J., 1988. The American Underclass: Inner City Ghettos and the Norms of Citizenship. The Godkin Lecture, John F. Kennedy School of Government, Harvard University.
Wright, Erik O., 1994. Interrogating Inequality: Essays on Class Analysis, Socialism and Marxism. Verso, New York.
Zelley Jr., Edmund W., 1995. Is the 'underclass' really a class? Journal of Sociology and Social Welfare 22 (1), 75–85.

Unemployment and Mental Health

David Fryer, University of Queensland, Brisbane, QLD, Australia; University of South Africa, South Africa; and University of Stirling, Stirling, UK
Rose Stambe, University of Queensland, Brisbane, QLD, Australia

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by D. Fryer, volume 23, pp. 15948–15951, © 2001, Elsevier Ltd.

Abstract

In this article we first review what is generally believed to be the case regarding the relationship between unemployment and mental health. Since we believe the treatment in the dominant literature of both 'unemployment' and 'mental health' to be critically problematic, we then subject each domain to brief critical examination. We close by suggesting ways in which we need to progressively move forward in theory and practice.

Classic Psychological Research on 'Unemployment' and 'Mental Health'

Researchers in the 1930s (e.g., Österreichischen Wirtschaftspsychologischen Forschungsstelle, 1933; Pilgrim Trust, 1938: see Fryer, 1986a; Jahoda, 1982 for a more detailed account) carried out unemployment research in the context of mass unemployment and hunger marches to determine whether mass unemployment would lead to revolution or apathy: the answer given was the latter. Researchers in the 1970s and 1980s were more concerned with the question of whether the association between 'unemployment' and 'indicators of poor mental health' was to be explained by processes through which those with poorer mental health were more likely to become and remain unemployed (individual drift) or processes through which healthy people who became unemployed became less mentally healthy (social causation). While still trying to explicate a causal relationship, research in the 1990s and 2000s has focused on the different effects of unemployment between individuals, e.g., Leana and Feldman (1995), and how identifying these 'moderators' can be used to fine tune intervention technologies, e.g., Paul and Moser (2009). The shift from a more social and political understanding of the problem of unemployment to a problem with the individual reflects a wider and increasing influence of neoliberalism in the Global North.

The accumulated research about the 'psychological costs' of 'unemployment' is now vast. Maynard and Feldman (2011), for example, reported that their search of relevant databases (PsycINFO, SocIndex, etc.) had revealed 31 839 peer-reviewed works with 'unemployment' in the abstracts published in the previous 50 years. However, large bodies of literature had assembled long before this, e.g., in 1938 Eisenberg and Lazarsfeld were able to review well over a 100 studies relating to 'the psychological effects of unemployment' (Eisenberg and Lazarsfeld, 1938) and even in 1909 Taylor was able to assemble a compendious 'Bibliography of Unemployment and the Unemployed' (Taylor, 1909).

A huge and diverse body of unemployment research has been carried out: in a wide variety of geographical settings; across a wide variety of historical periods; from a wide variety of diverse funding bases; at a variety of 'levels' (individual unemployed people, unemployed families, cohorts of school-leavers, whole redundant workforces, populations of states or whole countries, etc.); and has been characterized by researchers' use of a wide variety of methods and research designs (psychiatric assessment, qualitative interviewing, cross-sectional and longitudinal surveys using validated reliable measures, epidemiology, action research) (Fryer, 1986a but see Fryer, 2014). For many researchers, well-designed longitudinal studies in the 1980s using standardized measures and tracking large, carefully matched, samples of people in and out of employment-related transitions were the key studies. Some of the – for many – most persuasive longitudinal quantitative studies were done with young people using standardized measures, deemed reliable and valid, of anxiety, depression, positive and negative self-esteem, a variety of measures of affect and 'psychological well-being.' Typically, these studies measured the 'mental health' of large groups of young people in school and then followed them through labor market transitions, periodically measuring the 'mental health' of those who got jobs and those who did not, and comparing group mean scores cross-sectionally and longitudinally. Groups of unemployed youngsters have been repeatedly demonstrated to have poorer 'mental health' than their employed peers but statistically significant differences are seldom found between the scores of the same groups when at school (Warr, 1987). Many such studies have been done, but three longitudinal research programs have been particularly influential. The first was the 1980s research program based at the Social and Applied Psychology Unit in Sheffield, England – (see Warr, 1987 for an overview). The second program was based at The Flinders University of South Australia – (see Feather, 1992; Feather and O'Brien, 1986 for overviews). The third was based at the University of Adelaide – (see Winefield et al., 1993 for an overview). In brief, such studies provided powerful evidence within a mainstream frame of reference that unemployment causes, rather than merely results from, poor psychological health.

Statistically sophisticated meta-reviews have pooled data from a variety of studies (e.g., McKee-Ryan et al., 2005;

Paul and Moser, 2009) and there has been widespread agreement that 'unemployment' is not only associated with but 'causes' individual misery and mental health problems including anxiety, depression, negative self-esteem, dissatisfaction with life, social dislocation, community dysfunction, and population morbidity – (see Classen and Dunn, 2012; Jefferies et al., 2011; Keily and Butterworth, 2013; Kim et al., 2012). While caveats, criticisms, and controversies have been raised in relation to this huge body of research, they are mostly relatively minor and overwhelmingly raised *within* the frame of reference of such research, rather than of it. It is now difficult to contest, within the dominant modernist research frame of reference, the claim that 'unemployment' has a detrimental impact upon the 'psychological well-being' of 'unemployed people.'

The Scale of 'Unemployment'

In September 2013, unemployment in the UK as indicated by those claiming financial support stood at 1 401 800 (the 'claimant count'). However, according to the International Labour Organisation (ILO) way of counting them, 2 487 000 people in the UK were unemployed. Moreover, 1 447 000 people were *involuntarily part-time employed* being unable to find full-time employment and a further 8 961 000 people were deemed '*economically inactive*' – this included those who want to be employed but had stopped active job search because depressed or pessimistic about the chances of success (all the above figures accessed at http://www.cesi.org.uk/statistics/labour/september-2013 on 15 October 2013).

Unemployment in Europe, and in some European countries in particular, at the time of writing this article, is extremely high and getting worse. Even according to ILO figures, which understate the scale of unemployment, according to Eurostat, in April 2013 *19 375 000* men and women were unemployed in the 17 countries of the euro area, a rate of *12.2%*, and *26 588 000* men and women were unemployed in the Euro-27 countries (a rate of *13%*). Unemployment increased by 1 673 000 in the EU-27 and by 1 644 000 in the euro area since April 2012. Greece, Spain, and Portugal had unemployment rates of *27*, *26.8*, and *17.8%*, respectively. Youth unemployment tends to be proportionately higher than adult unemployment. The youth unemployment rate was close to 25% in the euro area. The ILO-calculated global unemployment was *197 000 000* in 2012, increasing to *202 000 000* in 2013 and *205 000 000* in 2014 (http://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/—publ/documents/publication/wcms_202326.pdf accessed 28.05.13.).

If one accepts that 'unemployment' causes 'mental health' problems and accepts that the scale of unemployment is colossal globally, unemployment is a public mental health catastrophe.

We do not doubt that 'unemployment' (but also under- and overemployment) as currently neoliberally constituted in Organisation for Economic Co-operation and Development (OECD) countries are intimately interconnected with individual, family, community, and societal misery, morbidity, mortality but we do regard the dominant ways in which this is documented, explained, and intervened into to be profoundly problematic. We elaborate on this next.

The Terms 'Unemployment' and 'Mental Health'

How unemployment is conceptualized, defined, and operationalized (translated into a measurement procedure) determines who and how many is/are included (and excluded) in official 'counts' of 'the unemployed,' productively constrains who are selected as participants in empirical unemployment research and has wider ideological implications, which are explored below: by 'ideological implications' we mean implications regarding the interests served by different systems of knowledge claims.

Unemployment used to be widely operationalized in 'developed' countries as a count of those who were 'claimants' of unemployment-specific financial support. Changes in eligibility for financial support could – and often did – reduce the number of claimants without reducing the number of unemployed people. Currently, most countries use the ILO criteria for positioning people as unemployed, and therefore, also to determine the number of unemployed people. These ILO criteria position people as unemployed if they have not undertaken paid employment in the survey reference week, have looked for a job within the previous 4 weeks and are available to start one within 2 weeks, irrespective of whether they are or not in receipt of unemployment-related benefits. People without employment who want paid work but cease to actively look for employment are referred to as the 'discouraged' and are not included in the 'unemployment' figures. The unemployment rate in the UK in September 2013 was 7.3% but the Economics Policy Institute estimated that if unemployed people who had ceased active job search were included that figure would jump to 9.5% (Shierholz, 2013). If decades of psychological research regarded by most unemployment researchers as having demonstrated beyond reasonable doubt that people who become 'unemployed' are disproportionately likely to be depressed, anxious, demoralized, low in self-esteem, socially isolated, etc., (Fryer, 2012; Wanberg, 2012) i.e., discouraged as a result of being unemployed are taken seriously, in countries which use the ILO operationalization of 'unemployment' becoming 'unemployed' can lead to no longer 'counting' as 'unemployed' and thus be effectively 'disappeared' socially, economically, and politically (Fryer, 2014).

Reviews of the field comparing studies drawing upon rather differently constituted unemployed 'samples' should also, therefore, be treated with caution. Reviewers often seem to assume that 'unemployed people' can be lumped together for research purposes whether 'unemployed' in the 1930s, 1980s, or 2010s, although the sampling and the socially constituted nature of 'unemployment' were quite different at different times in different places.

From our standpoint, unemployment is socially constituted in different ways over time and this constitution involves a complex set of dynamic interconnections between dominant discourses, government policies, bureaucratic processes and procedures, administrative technologies, psychological knowledges, forms of subjectification, etc. (see also Baxandall, 2002; Walters, 2000; McDonald and Marston, 2005). The way unemployment is socially constituted has important research, policy, subjective, and ideological implications. The problem is not so much that the ILO operationalization is not broad enough to cover diverse manifestations of unemployment but

rather that it constructs a very particular and ideologically problematic way of understanding 'unemployment.' It is essential not to naturalize the dominant constituted version of 'unemployment' or take it for granted as 'real' but rather to emphasize that it is socially manufactured, historically and culturally contingent, has a genealogy that can be traced and functions to serve some interests rather than others.

If unemployment is a complex social construction, which has changed over time in certain respects, then 'mental health' and psychiatric diagnostic practices are even more evidently, widely problematized, social constructions. So-called 'users of psychiatric services,' survivors of psychiatry, emancipatory disability researchers, antipsychiatry activists, and many others have critiqued the psychopathologization of diverse ways of being human and recently there has been a cacophony of criticism of DSM-5, the text which sets the criteria for 'mental disorder' diagnosis, and the ways of engaging and some would say constituting mental ill-health it implies (e.g., Timimi, 2013). Without denying the distress experienced, the authors of this article position what acritical mainstream scientists refer to as 'major forms of depression' as currently 'existing' only in the sense that they are currently repeatedly socially constituted as 'real' through deployment of dominant discourses, procedures, technologies, etc. Some mainstream authors criticize psychiatric diagnostic practices but position what they consider to be 'appropriately defined forms of major depression' as existing in a realist sense, in themselves but also via elevated risks of comorbidity (Prince et al., 2007) and of suicide (Classen and Dunn, 2012). The notion of cause and effect has long been under fire in relation to human actions and subjectivity. For all these reasons the claim that there is a causal relationship between the changing social constructions, unemployment, and 'mental health,' – or should in contemporary times – be a controversial one.

Explaining the Effects of 'Unemployment'

This brings us to the question of why the lives of so many unemployed people are plagued by misery, morbidity, and, according to some research, mortality. The most influential such explanation is that of Marie Jahoda (1982) who argued that although the intended function of employment was to earn a living, employment also had unintended functions (an imposed time structure, engagement in regular social contact, participation in a collective purpose, receipt of a social identity and required regular activity) the deprivation of which – by unemployment – was responsible for the psychological consequences of unemployment. What is psychologically bad about unemployment, according to this account, is deprivation of benevolent psychologically structuring features of employment. However, many privileged, secure, wealthy, and privileged nonemployed people in the hugely unequal societies of the OECD manage 'psychologically,' relatively at least, very well indeed without having the 'psychological benefits' of employment imposed upon them and of course there is another literature documenting that the employment of many people is psychologically and physically destructive, so this cannot be the whole story. We have long ago offered a wider critique of Jahoda's account and offered an alternative meta-approach in terms of the restriction of agency by unemployment by poverty and powerlessness, especially over the future (e.g., Fryer, 1986b), but these days we are exploring the ways people, when unemployed, are subjectively and materially (re)constituted so that they come to embody, and perform themselves to reproduce a socially and historically produced and psychologically destructive, unemployed identity.

Some variation from person to person in the experience of unemployment is reported, widely explained by researchers in terms of the moderation of the relationship between unemployment and mental health by variables: operationalizing age, length of unemployment, employment commitment, poverty, gender, etc. Strikingly, *however*, the broad psychological impact of unemployment is, according to researchers, with minor caveats, not only psychologically destructive but also, phenomenologically speaking, extraordinarily consistent. *The consistency* in the nature and detail of that destruction across time space and culture *is found* despite huge variation in the nature of employment, the nature of work, the labor market, the nature and degree of support for people, distribution of wealth, dominant social values, etc., from the 1930s to the twenty-first century and from country to country. We regard the consistency of the results of subjectification of unemployed people across time and space as telling us more about the functions served by unemployment, and in particular the unemployed subjectivity, which is repeatedly reconstituted, for the economic and political neoliberal status quo than it does about replication of effects of unemployment (Fryer, 2014; see also McDonald and Marston, 2005; Schram, 2000). Our view is that a network of interconnected constructed and maintained social elements, including discourses of unemployment and mental health (and implicated psy-complex constructions like psychological well-being and misery) whose primary function is to control inflation, reduce wage costs, discipline, etc., also simultaneously constructs and 'makes real' a category of 'the unemployed' necessary to make the neoliberal labor market work in the interests of employers and shareholders, i.e., capital. *The* category *of 'the unemployed'* is composed of different people on the basis of varying criteria at different times and different places and is only meaningful, i.e., only exists, by reference to that network. The subjectivity of 'the unemployed' *is constituted* in such ways as to (re)produce the compliant human means of production required by the employers, shareholders, and government within the contemporary version of the neoliberal labor market. *The* subjectivity of unemployed people *is reconstituted* as phenomenologically aversive, socially stigmatized (for example, see Gibson, 2009), and psychopathological.

Future Directions of Theory and Research

Foucault (1979/2008) lectured specifically about neoliberalism in *The Birth of Biopolitics*. Rejecting the view that neoliberalism is laissez faire in all respects Foucault stated: "neoliberal governmental intervention is no less dense, frequent, active, and continuous than in any other system. But what is important is to see what the point of application of these governmental interventions is now… Government must not intervene on effects of the market. Nor must neo-liberalism, or neo-liberal government, correct the

destructive effects of the market on society… Government must not form a counterpoint or a screen, as it were, between society and economic processes. It has to intervene on society as such, in its fabric and depth" (p. 145).

We can see this very clearly in the logic of 'workfare,' which demands that unemployed persons not only prove they are actively seeking work but are also actively working on themselves through "self-examination, counselling, self-help groups…" (Dean, 1995: p. 575). Psychologically oriented active-labor-market-promoting intervention research projects have included recommendations that unemployed people receive cognitive behavior therapy (Proudfoot et al., 1997; see also Wanberg et al., 2011; Karren and Sherman, 2012) or other employment-specific programs like the JOBS Project (http://www.isr.umich.edu/src/seh/mprc/jobsupdt.html), which "involves the design and evaluation of a preventive intervention aimed at providing job-seeking skills to promote reemployment and to combat feelings of anxiety, helplessness, and depression among the unemployed." The JOBS project resonates, of course, with the 'Improving Access to Psychological Therapies Programme' (http://www.iapt.nhs.uk/). While these may well reduce unemployment in the target intervention groups, in a situation where the number of unemployed people vastly outnumbers the number of employment vacancies, they cannot of course do anything more than reorder the queue of unemployed people looking for employment (Fryer, 1999). At a time (September 2013), 2 487 000 people were unemployed in the UK but there were only 531 000 vacancies for paid work (http://www.cesi.org.uk/statistics/labour/september-2013 accessed on 15.10.13). Such approaches make no sense at a public mental health level since they result in the same number of people being constituted as unemployed with all that implies. These active labor market programs shift the responsibility onto the individual and simultaneously blame the individual for failing to reenter the work force and for their own suffering. Psychology, in providing the psy-complex resources for such (re)subjectification (Rose, 1999), contributes to the normalization of this shift of responsibility onto the individual, which then prohibits or limits possible self-intelligibilities; ones which may lie outside a neoliberal subjectivity (Brady, 2011) including potentialities for resistance.

The reconstitution of subjectivity includes the constitution of subjectivities positioned as pathological in terms of dominant ways of thinking, speaking, and acting in relation to 'mental illness.' Getting to grips with this constitution is a matter of getting to grips with "how statements about mental illness are given the status of being 'true'; how collections of interconnected statements about mental illness are given the status of being 'knowledge'; how forms of surveillance are given the status of 'care'; how forms of assault are given the status of 'treatment'; how the speech and action of some is positioned as 'expertise' with the consequent authority to decide whether others are to be compulsorily detained, have 'treatment' forced upon them etc." (See Stambe et al., 2012: p. 9).

In this article, we have explored a range of positions taken up by unemployment researchers regarding whether the apparent relationship between 'unemployment' and 'mental health' is to do with 'natural' and inevitable psychobiological consequences of depriving an unemployed person of employment-related, psychologically necessary, structures, to do with the frustration of the agentic potential of the individual unemployed person or to do with psychological states that are objectively identifiable as healthy or pathological. These positions have been subjected to a critical examination and we have mapped out our own position, that the apparent relationship between 'unemployment and mental health' is a set of connected manifestations of material, social, and subjective violence, which function to make the neoliberal labor market work in the interests of employers and shareholders. In the view of the current authors, future research on 'unemployment' needs to move away from the psy-complex's individual-blaming manifestations, which work for an oppressive neoliberal status quo, develop progressive emancipatory counterknowledges, and create discursive spaces in which it is possible to think about ourselves without colluding with neoliberal subjection.

See also: Co-morbidity of Mental and Physical Conditions; Depression, Pessimism, and Health; Globalization and Health; Governmentality; Neoliberalism; Social Epidemiology; Social Exclusion, Social Deprivation and Health.

Bibliography

Baxandall, P., 2002. Explaining differences in the political meaning of unemployment across time and space. The Journal of Socio-economics 31, 469–502.

Brady, M., 2011. Researching governmentalities through ethnography: the case of Australian welfare reforms and programs for single parents. Critical Policy Studies 5, 265–283.

Classen, T.J., Dunn, R.A., 2012. The effect of job loss and unemployment duration on suicide risk in the United States: a new look using mass layoffs and unemployment insurance claims. Health Economics 21, 338–350.

Dean, M., 1995. Governing the unemployed self in an active society. Economics and Society 24, 559–583. http://dx.doi.org/10.1080/03085149500000025.

Eisenberg, P., Lazarsfeld, P.F., 1938. The psychological effects of unemployment. Psychological Bulletin 35, 258–390.

Feather, N.T., 1992. The Psychological Impact of Unemployment. Springer Verlag, New York.

Feather, N.T., O'Brien, G.E., 1986. A longitudinal study of the effects of employment and unemployment on school-leavers. Journal of Occupational Psychology 59, 121–144.

Foucault, M., 1979/2008. Lecture 14 Feb 1979. The Birth of Biopolitics. Lectures at the College de France 1978–1979. Palgrave Macmillan, New York.

Fryer, D., 1986a. Being unemployed: a review of the literature on the psychological experience of unemployment. In: Cooper, C.L., Robertson, I. (Eds.), International Review of Industrial and Organisational Psychology 1986. Wiley, Chichester.

Fryer, D., 1986b. Employment deprivation and personal agency during unemployment: a critical discussion of Jahoda's explanation of the psychological effects of unemployment. Social Behaviour 1, 3–23.

Fryer, D., 1999. For better and for worse: interventions and mental health consequences of unemployment. International Archives of Occupational and Environmental Health 72, 34–37.

Fryer, D., 2012. Critical differences: the development of a community critical psychological perspective on the psychological costs of unemployment. In: Kieselbach, T., Mannila, S. (Eds.), Unemployment, Precarious Work and Health. Research and Policy Issues. VS – Verlag für Sozialwissenschaften, Wiesbaden, pp. 473–489.

Fryer, D., 2014. Unemployment. In: Teo, T. (Ed.), Encyclopedia of Critical Psychology. Springer, New York, pp. 2031–2036. Also published online as Unemployment in In: Teo, T. (Ed.), 2013. Encyclopedia of Critical Psychology: Springer Reference (www.springerreference.com). doi: 10.1007/SpringerReference_304973 2013-01-26 18:30:43 UTC. Springer-Verlag Berlin Heidelberg.

Gibson, S., 2009. The Effortful Citizen: discursive social psychology and welfare Reform. Journal of Community & Applied Social Psychology 19, 393–410.
Jahoda, M., 1982. Employment and Unemployment: A Social-psychological Analysis. Cambridge University Press, Cambridge.
Jefferies, B.J., Nazareth, I., Marston, L., Moreno-Kustner, B., Sellón, J.Á., Svab, I., King, M., 2011. Associations between unemployment and major depressive disorder: evidence from an international, prospective study (the predict cohort). Social Science & Medicine 73, 1627–1634.
Karren, R., Sherman, K., 2012. Layoffs and unemployment discrimination: a new stigma. Journal of Managerial Psychology 27, 848–863.
Keily, K.M., Butterworth, P., 2013. Social disadvantage and individual vulnerability: a longitudinal investigation of welfare receipt and mental health in Australia. Australian & New Zealand Journal of Psychiatry 47 (7), 654–666.
Kim, I.H., Muntaner, C., Vahid, S.F., Vives, A., Vanroelen, C., Benach, J., 2012. Welfare states, flexible employment and health: a critical review. Health Policy 104, 99–127.
Leana, C.R., Feldman, D.C., 1995. Finding new jobs after a plant closing: antecedents and outcomes of the occurrence and quality of reemployment. Human Relations 48, 1381–1401.
Maynard, D.C., Feldman, D.C. (Eds.), 2011. Underemployment: Psychological, Economic, and Social Challenges. Springer, New York.
McDonald, C., Marston, G., 2005. Workfare as welfare: governing unemployment in the advanced liberal state. Critical Social Policy 25, 374–401.
McKee-Ryan, F., Song, Z., Wanberg, C.R., Kinicki, A.J., 2005. Psychological and physical well-being during un-employment: a meta-analytic study. Journal of Applied Psychology 901, 53–70.
Österreichischen Wirtschaftspsychologischen Forschungsstelle, 1933. Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langdauernder Arbeitslosigkeit. Mit einem Anhang: Zur Geschichte der Soziographie. Verlag von S. Hirzel, Leipzig.
Paul, K.I., Moser, K., 2009. Unemployment impairs mental health: meta-analyses. Journal of Vocational Behaviour 74, 264–282.
Pilgrim Trust, 1938. Men without Work. Cambridge University Press, Cambridge.
Prince, M., Patel, Saxena, S., Maj, M., Maselko, J., Phillips, M.R., Rahman, A., 2007. No health without mental health. Lancet 370, 859–877.
Proudfoot, J., Guest, D., Carson, J., Dunn, G., Gray, J., 1997. Effect of cognitive-behavioural training on job-finding among long-term unemployed people. Lancet 350, 96–100.
Rose, N., 1999. Powers of Freedom: Reframing Political Thought. Cambridge University Press, Cambridge.
Schram, S.F., 2000. In the clinic: the medicalization of welfare. Social Text 18, 81–107.
Shierholz, H., 2013. 6 September 2013. The unemployment rate is less relevant than ever. Economic Policy Institute: Research and Ideas for Shared Prosperity. http://www.epi.org/publication/unemployment-rate-relevant/ (accessed 16.10.13.).
Stambe, R., Fryer, D., Dauncey, S., Hicks, S., 2012. Is 'mental illness' a barrier to getting involved? Asylum: an international magazine for democratic psychiatry. Psychology and Community, Development 19, 9–11.
Taylor, F.I., 1909. A Bibliography of Unemployment and the Unemployed. P. S. King and Son, London.
Timimi, S., 2013. No more psychiatric labels: campaign to abolish psychiatric diagnostic systems such as ICD and DSM (CAPSID). Self and Society 40, 6–14.
Walters, W., 2000. Unemployment and Government: Genealogies of the Social. Cambridge University Press, Cambridge.
Wanberg, C.R., 2012. The individual experience of unemployment. Annual Review of Psychology 63, 369–396.
Wanberg, C.R., Zhu, J., Kanfer, R., Zhang, Z., 2011. After the pink slip: applying dynamic motivation frameworks to the job search experience. Academy of Management Journal 55, 261–284.
Warr, P.B., 1987. Work, Unemployment and Mental Health. Oxford University Press, New York.
Winefield, A.H., Tiggemann, M., Winefield, H.R., Goldney, R.D., 1993. Growing up with Unemployment: a, Longitudinal Study of its Psychological Impact. Routledge, London.

Relevant Websites

http://www.iapt.nhs.uk/ – Improving Access to Psychological Therapies Programme.
http://www.isr.umich.edu/src/seh/mprc/jobsupdt.html – JOBS Project.

Unemployment: Structural

Assar Lindbeck, Stockholm University, Stockholm, Sweden

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15951–15958, © 2001, Elsevier Ltd.

Abstract

Structural unemployment differs from cyclical unemployment by not disappearing in cyclical booms. In economic theory, structural unemployment is usually analyzed in terms of the concept of equilibrium unemployment (the 'natural unemployment rate' in Friedman's terminology). Two elaborate concepts of equilibrium unemployment – the nonaccelerating inflation rate of unemployment and the unemployment rate that induces firms and workers to accept the same real wage (the PS–WS model) are used as analytical framework for the article. These models are also related to search models for the labor market, in which unemployment equilibrium is defined as a situation where the number of individuals finding jobs equals the number of individuals who are separated from jobs. But it is argued in this article that it is reasonable to widen the concept of structural unemployment to also include unemployment persistence, that is, a tendency for unemployment to continue to be high for quite a while also after temporary unemployment-creating shocks. This article pinpoints various factors that influence the level of structural unemployment, which differs in time and place. It ends with a discussion of how structural unemployment is influenced by policy actions.

Definitions

It is an old tradition in labor economics to distinguish between structural, frictional, and cyclical unemployment. Structural unemployment is then envisioned as a result of the institutional set up of the economy, including private and government organizations, types of market arrangements, demography, laws, and regulations. In the literature, the importance of these institutional features for structural unemployment is tied particularly to their implications for demand for and supply of labor, price and wage formation, and the efficacy of search and matching processes in the labor market. Frictional unemployment may be regarded as a subset of structural unemployment, mainly reflecting temporary unemployment spells as the result of job search and matching difficulties in the connection with quits, new entries to the labor market, and job separation because of the employers' dissatisfaction with individual workers.

Cyclical unemployment differs from structural and frictional unemployment by basically being tied to short-term economic fluctuations. An empirical illustration of the importance of structural unemployment as compared to cyclical is that variations in actually measured unemployment rates have turned out to be much larger between cycles than within cycles, presumably reflecting differences in structural unemployment (Layard et al., 1991). For example, the average unemployment rate over the business cycle in Western Europe has moved from about 3% in the mid-1960s to 6% in the mid-1970s and to 10% from the mid-1980s to the late 1990s. But as pointed out below, various unemployment-persistence mechanisms blur the distinction between cyclical and structural unemployment.

In economic theory, structural and cyclical unemployment usually are regarded as disequilibrium phenomena in the sense that they reflect excess labor supply at existing wages and hence that the labor market does not clear. Then, individual employers informally ration jobs. Nevertheless, technically (analytically) structural unemployment often is analyzed in terms of the concept of *equilibrium unemployment*. This means that the aggregate-unemployment level is in a 'state of rest': existing excess labor supply is assumed to last as long as certain characteristics (parameters) of the economy are unchanged. If this equilibrium is dynamically stable, unemployment equilibrium may also be described as an unemployment level toward which the economy moves as long as new disturbances do not emerge. (The term 'equilibrium' is used in several different ways in economics. It refers sometimes to the equality between demand and supply in a market, i.e., to traditional 'market- clearing,' in other cases to a state that tends to continue over time regardless of whether the market clears or not. The concept of 'equilibrium unemployment' is an example of the latter.)

It is useful to distinguish between two main analytical approaches to equilibrium unemployment: *stock approaches* and *flow approaches*. Stock approaches focus on the difference, at a given point in time, between the workforce desired by firms (aggregate *stock demand* for labor) and the number of workers willing to work (aggregate *stock supply* of labor). Flow approaches deal with the difference between the flows in and out of the unemployment pool during a certain period. Flow-demand for labor is represented then by the supply of 'job slots,' and flow-supply by offers of workers to fill such slots. These two approaches are related because stock demand and stock supply may also be derived in flow approaches. Let us start with stock approaches.

Stock Equilibrium

Milton Friedman's celebrated 'natural unemployment rate' is a stock-equilibrium concept of structural unemployment: 'the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demand and supplies, the costs of gathering information about job vacancies and labor availabilities, the

costs of mobility and so on' (Friedman, 1968). Apparently, this concept of equilibrium unemployment covers structural and frictional unemployment as described earlier.

The nonaccelerating inflation rate of unemployment (NAIRU) is close to Friedman's natural rate, though it emphasizes (more than Friedman does) the nonclearing character of the labor market. The NAIRU is defined explicitly as the unemployment rate at which inflation is constant. It is derived from an expectations-augmented Phillips curve, i.e., a function that assumes that the price-inflation rate is a decreasing function of the unemployment rate and an increasing function of the expected rate of future inflation (Phelps, 1967). In the simplest possible terms (with linear relations):

$$\pi = -au + b + c\pi^{e} \qquad [1]$$

where π is actual inflation, u is the unemployment rate, and π^{e} is expected inflation. For given values of the (positive) parameters a, b, and c, the equation may be depicted as a set of downward sloping short-term Phillips curves (SR_i), one curve for each set and for each value of π^{e} (see **Figure 1**).

In the context of this model, usually (and realistically) it is assumed that economic agents ask for full compensation for expected inflation in the *long run*; thus $c = 1$ in that time perspectives. It is usually also assumed that economic agents do not make *systematic* expectational errors in the long run; thus $\pi = \pi^{e}$. The only unemployment rate consistent with these two assumptions is, from eqn [1], $u^* = b/a$, where * denotes the equilibrium unemployment rate. It is often also assumed that inflation in this situation is constant over time (and not just that actual and expected inflation coincide), because agents have no reason to revise their price and wage setting behavior when inflation is exactly what they had expected when they formed their decisions. u^* is then the NAIRU (the unemployment rate at which inflation is constant), geometrically represented as a vertical long-run Phillips curve, LR in **Figure 1**. Inflation increases to the left of that level and falls to the right. Because inflation in this framework in the long run is independent of unemployment, there is no long-run trade-off between these variables. But it has been argued by Lucas (1972) that there is not even a *systematic* short-term trade-off if economic agents have 'rational expectations' (thus, if they never make systematic expectational errors), and if various agents are free to revise their prices and wages in response to changes in expectations. Such a trade-off will then only arise due to random (nonsystematic) expectational errors, e.g., due to random shocks or randomized government policy.

An alternative exposition of stock-equilibrium in the labor market, and hence structural unemployment, gained ground in the 1980s and 1990s; it emphasizes the requirement of consistency of price- and wage-setting behavior (Shapiro and Stiglitz, 1984; Layard and Nickell, 1986). **Figure 2** gives a simple diagrammatic illustration, with aggregate employment (N) on the horizontal axis and the aggregate real-wage level (w) on the vertical. The LS curve is the traditional aggregate labor supply curve, i.e., the sum of labor supplied by households. The PS curve is the price-setting curve of firms: for alternative unemployment levels and given nominal wages, the curve defines the price level desired by profit-maximizing firms that operate in imperfectly competitive product markets. Or the PS curve can be described as a labor-demand relation, because it also expresses the combination of employment and real wages desired by profit-maximizing firms. Unemployment is then the horizontal difference between the PS and LS curves.

The WS curve defines wage-setting behavior. It depicts the influence of the (un)employment situation on the real-wage rate that the wage-setting process generates. The intuition is that firms feel compelled to offer higher real wages and that workers and unions demand higher real wages when the labor market is tight (low unemployment) than when it is slack. So by contrast to models with perfectly competitive labor markets, a distinction is made between labor supply (by households) and wage setting (by firms, unions, or more realistically a bargaining process).

The equilibrium real-wage rate is now w_0, equilibrium employment N_0 and equilibrium unemployment U_0, the latter being the only unemployment level at which there are no incentives for agents to change prices relative wages and vice versa (corresponding to the equilibrium unemployment rate

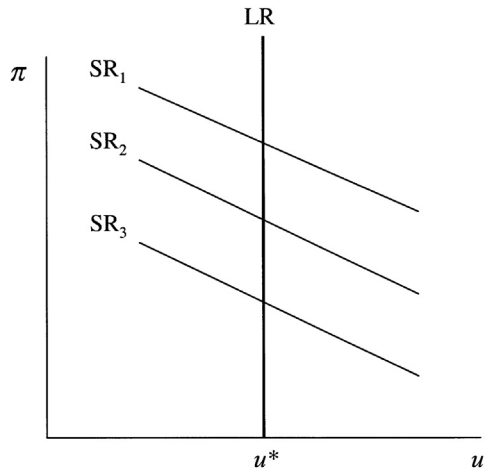

Figure 1 Phillips curves and the NAIRU.

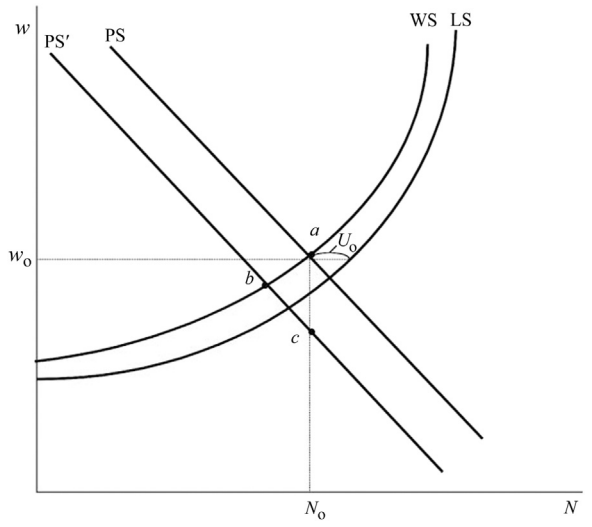

Figure 2 The PS–WS model of equilibrium unemployment.

u^* in Figure 1). This model will be called the PS–WS model of equilibrium (structural) unemployment.

As an illustration, suppose that the real-wage rate in Figure 2 is initially w_0 but that aggregate employment happens to be above N_0, e.g., as a result of positive demand shocks to which nominal magnitudes have not yet adjusted. Firms and workers are then dissatisfied with the existing real wage w_0 Workers and their unions try to raise nominal wages (at existing nominal prices), as can be read off from the WS curve, and firms try to raise prices (at existing nominal wages), as can be read off from the PS curve. The result is a wage-price spiral, as described by the expectation-augmented Phillips-curves model. Firms also start cutting their work force because their desired level is lower than the actual one. Aggregate employment continues to fall until it reaches the unique equilibrium level N_0. The same reasoning, *mutatis mutandi*, holds if aggregate employment is initially lower than the equilibrium level.

Flow Equilibrium

Flow approaches to structural unemployment instead emphasize the flow of workers in and out of unemployment (Phelps, 1970; Hall, 1979; Diamond, 1982). These approaches explicitly grant the heterogeneity of jobs and workers, which means that agents in the labor market must devote time and costs to search to match up idiosyncratic preferences, skills, and skill requirements. In the process, unemployment and vacancies coexist because the latter are not eliminated immediately by the hiring of workers in connection with decentralized negotiations of wages between firms and work applicants.

For simplicity, let us regard the size of the labor force (L) as constant. Further, let s denote the job separation rate, i.e., the fraction of employed workers (N) who lose their jobs in each time period. Let f denote the rate of job finding, i.e., the fraction of unemployed workers (U) who find a job in each time period. We assume that both fractions are constant. The equilibrium unemployment rate (u^0) is then defined as a situation when the number of individuals finding jobs equals the number of individuals who are separated from jobs; hence $fU = sN$. Noting that $N = L - U$, i.e., that the number of unemployed workers equals the labor force minus the number of employed, we get $fU = s(L - U)$ Rearranging terms and dividing both sides by L, we obtain the equilibrium, or structural, unemployment rate $u^0 = U/L = s/(s^{-1} + f)$. As expected, the higher the rate of job separation (s), the higher the equilibrium unemployment rate. And the higher the rate of job finding (f) – due to greater intensity and efficiency of job search and hence better job matching – the lower the equilibrium–unemployment rate.

Though the model basically is designed to highlight frictional unemployment, as described earlier, unemployment in the connection with job search and matching may also be due to continuing changes in the composition of demand and supply of different types of workers. This means that flow models of search and job matching also cover elements of what usually is regarded as structural unemployment, which illustrates the previous assertion that frictional unemployment may be regarded as a subset of structural unemployment.

It is natural and feasible to integrate flow and stock approaches to (un)employment determination. For instance, the efficacy of search and matching processes influences the positions of the PS and WS curves in Figure 2 (and the positions of the Phillips curves in Figure 1). The longer it takes firms to find the right workers for their vacancies, and the longer it takes work applicants to acquire suitable job offers, the more costly it is to hire workers, and the lower will be the labor-demand curve – the PS curve in Figure 2. A poorly functioning search and matching process will force recruiting firms to offer higher wages than otherwise; the result is also a higher WS curve. In this way, flow models are related to previously discussed stock models. Search and matching processes may also be seen as part of the short-term *dynamics* by which stock equilibrium is reached, with positive gross flows in and out of unemployment, netting to zero when there is flow and stock equilibrium.

Microeconomic Foundations

All these macroeconomic models of structural unemployment assume that unemployed workers are not able or willing to get jobs by underbidding the prevailing wages of incumbent workers. The most obvious microeconomic explanation of the absence of wage underbidding is perhaps *minimum wage laws*. But there seems to be rather general agreement among labor market economists that minimum wages have not been high enough in developed countries to explain much of aggregate structural unemployment in that part of the world. But in some countries, minimum wage laws are binding for specific types of workers, such as inexperienced, unskilled, and physically or mentally handicapped people. (In the special case when a firm has a monopsonistic position in a local labor market minimum wages may, up to a certain wage level, increase rather than reduce employment. In this case, the minimum wage simply prevents the firm from keeping down wage costs by holding back its demand for workers.)

There may also be informal *social norms* against the underbidding of the wages of incumbent workers. In other words, wage underbidding may not be socially acceptable behavior among workers, and the acceptance of such offers may not be socially acceptable behavior among firms. Legislation may support such norms.

But more elaborate microeconomic foundations of structural unemployment have been developed since the 1980s. In the *union monopoly model*, unions are assumed to strive for a trade-off between real wages and unemployment. Often, this trade-off is derived formally from a model where unions maximize the expected utility of union members – including those who are employed and those who wind up unemployed due to the union's wage policy. So labor unions are assumed conscientiously to choose a real-wage rate that generates some structural unemployment.

In *union-bargaining models*, the real wage is instead set in a bilateral bargaining process between a union and a firm or an association of firms. Such models qualitatively yield the same result as the union-monopoly model, though the real wage will normally be lower and hence the employment level higher, because the firm is assumed to strive for a lower real wage rate than the union. But neither type of union model answers the questions why there is no underbidding of existing wages by

nonorganized workers and why unemployed union members do not leave the union.

There are also theoretical reasons and empirical evidence why the organization of bargaining, such as the degree of coordination and centralization of bargaining, would influence the outcome of the bargaining process and hence also the unemployment level. The most generally accepted hypothesis is probably that both highly centralized and highly decentralized bargaining are more conducive to employment than wage bargaining on levels in between, such as industry-level bargaining (Calmfors and Driffill, 1988).

The *insider–outsider* theory of unemployment emphasizes the asymmetric market powers of incumbent workers, *insiders*, and disenfranchised workers without jobs in the formal labor market, *outsiders* (Lindbeck and Snower, 1988). Because hiring and firing costs protect incumbent workers, they can push up their wages above the reservation wage of outsiders without losing their jobs. Another way for insiders to prevent wage underbidding is by refusing to cooperate with underbidding outsiders in the production process and by threatening to harass them if underbidding workers are actually hired. With the help of such noncooperation and harassment threats, insiders (possibly with the support of unions) may also be able to create and sustain social norms in their own interest against wage underbidding.

In *efficiency wage* models, by contrast, the microeconomic foundations of structural unemployment are based on the incentive structure of firms (Yellen, 1984). Firms, which are assumed to have all market powers in wage formation, set wages high enough to get a good selection of work applicants and to discourage workers from shirking and quitting. More specifically, firms set wages to balance the cost of higher wages against the productivity increase of having a more efficient work force. Only by accident would the wage level that is optimal for firms clear the labor market.

All these microeconomic theories are rather silent about the employment situation in *informal* labor markets, where minimum wages are not always respected, unions may not exist, collective bargaining agreements are not binding, efficiency wages are not important, and labor turnover costs are very low and hence insider–outsider mechanisms not important. Extreme examples are sellers of flowers or newspapers on street corners, shoeshine boys, and workers in the underground economy. One conceivable explanation why some workers who are excluded from the *formal* labor market remain unemployed rather than choose to work in informal labor markets may be an unwillingness to take jobs that are regarded as highly insecure and demeaning. Generous unemployment benefits help them to make this choice. It is significant that informal labor markets are particularly important in countries without unemployment insurance.

In the popular discussion of structural unemployment, it is often believed that labor-saving technological progress and related increases in labor productivity will necessarily result in higher unemployment. Workers are asserted to become redundant because machines replace workers – a vision of *technological unemployment* as a special variation on structural unemployment. A usual conclusion is that some kind of work sharing is necessary to avoid ever-increasing unemployment. An obvious flaw of this argument is that a gradual rise in productivity will normally be accompanied by higher real wages and profits in the long run, and hence of real income. In this sense, productivity growth creates the foundation for higher demand for output, though deficiencies in aggregate demand may occasionally occur and contribute to recessions. In fact, no long-term trend of the unemployment rate can be detected during the twentieth century.

So work sharing – in the form of a lower general pension age, more generous rules for early retirement, or fewer working hours – is not necessary to prevent gradually rising labor productivity from resulting in rising unemployment. Indeed, the literature in the field clarifies that work sharing is not even likely to contribute to reduced unemployment except in the very short run; one reason is that a reduction of the labor force tends to push up wages. But there is nevertheless a potential problem with continuing productivity growth. Some national economies may fail to adjust to changes in the *composition* of the demand for labor in the connection with uneven rates of productivity growth among sectors and types of workers (sector shifts or skill shifts in the labor market). This is a respectable explanation as to why unemployment rates among low-skilled workers have increased so much (in percentage points) in Western Europe, where relative wages are quite rigid.

Unemployment Persistence

One serious limitation of the concept of unemployment equilibrium as an expression of structural unemployment is that this concept is difficult to distinguish empirically from unemployment persistence, i.e., inertia of unemployment after temporary shocks that have shifted the unemployment rate. Estimated equilibrium rates of unemployment tend to shadow the actual rate, which is an indication of unemployment persistence. Important examples of unemployment persistence are the persistently high unemployment rates after the deflationary policy pursued in several developed countries in the early 1920s and early 1930s; after the oil-price hikes in mid- and late 1970s; and after the restrictive monetary (and in some cases also fiscal) policy in the late 1980s and early 1990s. These experiences raise the suspicion that attempted statistical measurements of equilibrium unemployment also reflect the consequences of various persistence mechanisms. It is hard to find good reasons why the equilibrium unemployment rate in Western Europe should have increased as dramatically as the actual unemployment rate in the 1980s and 1990s.

There are also reasonable microeconomic explanations of unemployment persistence. While some explanations emphasize the behavior of outsiders, others point instead to the behavior of insiders. Examples of persistence mechanisms of the former type are the loss of skill among individuals who have been unemployed for a long time; endogenous changes in preferences in favor of leisure or household work; and the breakdown of social norms in favor of work and hence the emergence of unemployment cultures. We would expect reduced job search in all these cases. Long spells of unemployment may also function as a negative signal to prospective employers about the quality of individual workers. An example of a persistence mechanism that operates via the behavior of insiders is that they (possibly supported by unions) may exploit

subsequent business upswings to push up their wages, hence discouraging firms from expanding their workforce. All this suggests that unemployment persistence is an important factor behind what is usually regarded as structural unemployment. A basic conclusion is that endogenous changes of private behavior after a temporary unemployment-creating shock make the concept of *equilibrium unemployment* too narrow as a description of structural unemployment. For instance, it is realistic to regard prolonged high unemployment in Western Europe after the mid-1970s as the combined result of macroeconomic shocks and institutional rigidities that contribute to unemployment persistence. Hence, cyclical unemployment is asserted to have developed into structural unemployment. The result is strong 'history dependence' of the unemployment rate. So it is reasonable to widen the concept of structural unemployment to also include unemployment persistence.

Policies that Influence Structural Unemployment

Leaving the issue of legislated work sharing aside, what are the influences of government policy on structural unemployment? In the schematic PS–WS model presented earlier, the positions of the PS, WS, and LS curves reflect this influence. The discussion below is confined to taxes, transfer programs, regulations of job-protection, union powers, education/training, and other types of active labor market policy.

Taxes on labor insert a wedge between the labor costs paid by firms, the product wage $(1 + t_w)w$, and the return to workers, the consumption wage, $(1 - t_i)w$. t_w and t_i then denote the wage tax (payroll tax), and the income tax. *The tax wedge*, expressed as the ratio between the product wage and the consumption wedge, is then

$$(1 + t_w)/(1 - t_i) \qquad [2]$$

A consumption tax would enter the tax wedge in the same way as the wage tax. For example, in terms of **Figure 2**, an increase in t_w shifts the labor-demand relation (the PS curve) down, e.g., to PS: at every real-wage rate (w) the demand for labor will be lower than before because the firm must now pay $(1 + t_w)w$ for its labor. As long as the WS curve is unchanged, the equilibrium position shifts from point a to point b in the figure, and the equilibrium employment level (N) and the real-wage rate (w) fall. These effects are relevant in the short run. The long-run effects are less obvious, and they depend on the extent to which wage setting adjusts to the tax wedge, which in turn depends on what happens to alternative incomes of workers, such as unemployment benefits and the return on household work. But for low-skilled workers, wages are often tightly regulated, which means that the wage costs for firms are likely to rise and labor demand will fall for this specific group of workers.

The results of empirical studies on these matters are somewhat ambiguous. But on balance, it seems that aggregate employment is negatively affected by wider tax wedges not only in a short- and medium-term perspective but also in the long run (Nickell and Layard, 1999).

More generous *unemployment benefits* enable workers to be more selective about offered jobs and hence search for longer time periods – as highlighted by the earlier discussed search approach. Frictional unemployment (or *wait unemployment*) will then increase. Because the income loss in connection with unemployment becomes smaller, firms must offer higher wages to prevent shirking and to keep down quits, as highlighted by efficiency wage models. Moreover, workers and unions become less hesitant to ask for high wages, as emphasized by union and insider–outsider models. This tendency is accentuated by the fact that unemployment benefits are, as a rule, financed by taxes rather than by insurance fees directly tied to specific workers. In terms of **Figure 2**, the net effect of all this is that the PS curve is shifted down and the WS curve is shifted up, with rising equilibrium unemployment as a result. Unemployment persistence would also be predicted to increase.

Casual observation and systematic empirical studies, including cross-country regressions, suggest that generous unemployment benefits that may be kept for long periods (in some cases for several years or indefinitely), in fact, have effects of this type (Layard et al., 1991). Other evidence is that the exit rate for individuals from unemployment increases substantially in the US as soon as the benefit period expires after 26 weeks (Katz and Meyer, 1990).

Job-security legislation is often singled out as another factor that tends to raise structural unemployment. Basically, such legislation raises the costs of hiring and of firing workers, the former because firms must consider the possibility of future firing costs when hiring workers. Firms will be less anxious to fire workers in recessions but also less interested in hiring workers in business upswings. A predicted effect is that employment tends to be stabilized at the actually existing level whatever this happens to be. In other words, (un)employment persistence would be expected to increase.

Since such legislation increases the market powers of insiders, this helps them to push up their wages. The predicted effect is an increase in the average unemployment rate over the cycle. Moreover, the market power of insiders and unions depends also on other legislation. Important examples are rules that extend the coverage of bargaining agreements to nonunion firms and give large freedom for unions to use conflict weapons, such as the rights to sympathy strikes, secondary actions and blockades, and the right to take out small 'key groups' in conflicts.

Most government policies discussed earlier tend to raise structural unemployment (including unemployment persistence). Of course, the other side of the coin is that a reversal of these policies is likely to reduce structural unemployment. But there are also well known examples of government policies directly aiming to lower structural unemployment. Competition policy, which accentuates competition in product markets, induces firms to increase production and employment at given real wages; in terms of **Figure 2**, the PS curve shifts to the right. Increased competition in product markets also makes product demand more sensitive to changes in wage costs. This is likely to restrain the wage demands by insiders; the WS curve shifts down. The net effect would be lower equilibrium unemployment.

Active labor market policy is designed to keep workers out of open unemployment in the short term and to subsequently help them to get jobs. In particular, labor-exchange systems and retraining may improve matching between work applicants and vacancies. In terms of **Figure 2**, the PS curve shifts up and the WS curve down with falling structural unemployment as a result. Moreover, subsidized education and training that

raises labor productivity for (previously) low-skilled workers are likely to raise demand for them and reduce the upward wage pressure for skilled workers. This tends to generate similar shifts in the PS and WS curves as in the case of improved labor market exchanges. Unemployment persistence would also be expected to recede.

On balance, empirical studies suggest that labor-market exchange and properly conduced training programs, in fact, are useful for reducing structural unemployment (Layard et al., 1991); in the short term, also public works programs have such effects. But it is natural that the marginal return falls, possibly to zero, when the programs reach a large size (such as several percent of the labor force), and in particular when the programs are used largely to keep down statistically recorded figures on open unemployment.

See also: Labor Supply; Market Structure and Performance.

Bibliography

Calmfors, L., Driffill, J., 1988. Centralized wage bargaining and macroeconomic performance. Economic Policy 6, 13–61.

Diamond, P., 1982. Aggregate demand management in search equilibrium. Journal of Political Economy 90, 881–894.

Friedman, M., 1968. The role of monetary policy. American Economic Review 58, 1–17.

Hall, R.E., 1979. Theory of the natural unemployment rate and the duration of employment. Journal of Monetary Economy 5, 153–169.

Katz, L.F., Meyer, B.D., 1990. Unemployment insurance, recall expectations and unemployment outcomes. Quarterly Journal of Economy 105, 973–1002.

Lindbeck, A., Snower, D.J., 1988. The Insider-Outsider Theory of Employment and Unemployment. MIT Press, Cambridge, MA.

Layard, R., Nickell, S., 1986. Unemployment in Britain. Economica 53, 121–169.

Layard, R., Nickel, S., Jackman, R., 1991. Unemployment: Macroeconomic Performance and the Labour Market. Oxford University Press, Oxford, UK.

Lucas, R., 1972. Expectations and the neutrality of money. Journal of Economic Theory 4, 103–124.

Nickell, S., Layard, R., 1999. The labour market consequences of technical and structural change. In: Ashenfelter, O.C., Card, D. (Eds.), Handbook of Labour Economics. North Holland Press, Amsterdam.

Phelps, E.S., 1967. Phillips curves, expectations of inflation and optimal unemployment over time. Economica 34, 254–281.

Phelps, E.S. (Ed.), 1970. Microeconomic Foundations of Employment and Inflation Theory. Norton, New York.

Shapiro, C., Stiglitz, J.E., 1984. Equilibrium unemployment as a workers disciplinary device. American Economic Review 74, 431–444.

Yellen, J., 1984. Efficiency wage models of unemployment. American Economic Review 74, 200–205.

Unintended Consequences: History of the Concept

Adriana Mica, Institute of Applied Social Sciences, University of Warsaw, Warsaw, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Unintended consequences might be linked to three problems in social sciences: purposive social action, human action (but not human design), and purposive social intervention. The treatments of these evolved rather autonomously within sociology, economics, and public policy analysis, respectively. Sociology established an interest in the unintended as deviation from rational-action accounts of social behavior. Economics investigated the spontaneous social order and the processes of aggregation leading to unintended configurations. Public policy analysis pointed to paradoxes and perverse effects of initiated institutional change. The proliferating analyses of unintended consequences as side effects, of perverse effects, and of paradoxes of modernization projects recently presented the empirical opportunity for the three perspectives to be compared and to further develop their conceptual backgrounds.

The problem of unintended consequences has claimed a central tenet in both sociology and economics. As a matter of fact, it is asserted, over and over again, to be the main concern of social sciences. The fascination with the unintended stems from the fact that processes occurring in a linear fashion do not constitute a problem and challenge for social sciences while unexpected occurrences do (see Popper, [1945]2013: 307; Portes, 2000). The term 'unintended consequences' is used less in relation to overtly stated intentions, but more so to depict outcomes of social action which contravene to the interests of the actors (Popper, 1957: 157). In addition, it pertains to indirect effects and the change of goals during the implementation of certain intentions that are depicted in a paradoxical manner.

Unintended Consequences of Purposive Social Action versus Unintended Consequences of Human Action, But Not of Human Design

The treatment of unintended consequences in sociology and economics was inevitably influenced by the insights of political thought, moral philosophy, and history (Vernon, 1979). This common heritage notwithstanding, an essential difference in approach is apparent. Sociology, from a functionalist background and guided by Robert K. Merton's (1936) analysis of unanticipated consequences of purposive social action, dealt predominantly with the reasons why actions produce unexpected outcomes and the subjective processes conditioning their anticipation. Further, it gave priority to unintended consequences that pertain to negative or undesirable effects. While, economics, building upon Adam Smith's ([1759]1976, [1776]1975), Adam Ferguson's ([1767]1966), and David Hume's (see discussion in Barry, 1982; Hamowy, 1987, 2008) ideas of the invisible hand and spontaneously created social institutions, favored a treatment of unintended that focused on structured patterns resulting from separate actions which did not intend the overall resulting order.

In sociological fashion, unintended consequences were predominantly traced to social action. They were "a particular effect of purposive action which is different from what was wanted at the moment of carrying out the act, and the want of which was a reason for carrying it out" (Baert, 1991: 2). On the other hand, in economic thought, they were defined in relation to social interaction as "establishments, which are indeed the result of human action, but not the execution of any human design" (Ferguson, [1767]1966).

Albert O. Hirschman (1991) identified the three unintended narratives within the 'rhetoric of reaction' against programs of social intervention – i.e., the perversity thesis, the futility thesis, and the jeopardy thesis. Following this logic, the two outlined treatments in the sociology and economics of the unintended, though not reactionary per se, might be termed as the unanticipated-perverse effect perspective (UPE) and the invisible hand-spontaneous order interpretation (ISO).

Intellectual Backgrounds

Within sociology, UPE states that because of limitations of foresight of consequences of social action, the actions of individuals have consequences that are unintended and sometimes even contrary to what was pursued and anticipated. In terms of reactionary rhetoric, Hirschman (1991) traced the formulation of the perversity thesis to Edmund Burke's prediction, and Joseph de Maistre's theological view of the French Revolution, as well as to treatments of the unexpected and invisible hand in Giambattista Vico, Bernard Mandeville, and the Scottish Enlightenment.

Taken in its broadest, UPE covers two out of five of the alternative action sequences to the linear purposive action that were identified by Alejandro Portes (2000, 2010). These two notions are the shift of goals during a given activity (e.g., Max Weber's thesis on the effect of rational ascetism on economic behavior, Robert Michels' iron law of oligarchy) and the unexpected outcome (e.g., Georg Simmel's analysis of cumulative outcomes, Merton's unanticipated consequences of social action). The mid-course shift, after initially being termed as 'the essential paradox of social action' (Merton, 1936), suffered a somehow arrested development on the unintended agenda until it was rediscovered, a few decades later, by public policy analysts (Sieber, 1981; Hirschman, 1977) and sociologists (see Schneider, 1975; Cherkaoui, 2007 – see also discussion in Lal, 1998; Symonds and Pudsey, 2008). In contrast, the unexpected outcome exercised a prevailing

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03176-7

influence in modern sociology from the very beginning. Portes (2000: 10) defined it as involving "end-states that are qualitatively distinct, and sometimes the opposite of, those originally intended."

The pervasive influence of Simmel notwithstanding, it is Merton's (1936) treatment in the seminal article *The Unanticipated Consequences of Purposive Social Action*, which is broadly considered to be the first initiative to establish sociology as the 'analysis of the unexpected' (Portes, 2000; Mica et al., 2011). In this article, Merton (1936: 904) addressed 'social prediction, control, and planning' and differentiated five 'types of social action' by indicating limitations of 'correct anticipation of consequences of action'. As indicated by Portes (2010: 19), "the essay opened the door to a number of modern concepts, all highlighting the paradoxical nature of social life," Merton himself is being one of the main contributors to this list.

The notions advanced by Merton were quite distinct in terms of structural range and type of interest in the relationship between the consequences and the initial intention. Despite this broad compass, they focused sociology on the subjective processes determining the foresight of consequences and, contrary to Merton's intention, reinterpreted the unintended consequences as unexpected and undesirable. As pointed out by Patrick Baert (1991), this path dependence was linked to the broader functionalist framework hosting the unintended debate, and to the fact that sociologists drew their empirical cases from policy-relevant areas, in which the initial intentions became frustrated. On the other hand, the sense of superiority and fascination with paradoxical and counterintuitive outcomes enjoyed by sociologists, who were skeptical of systemic theories grounded on purposive social action, cannot escape notice.

In comparison to sociology, economics, at least in the ISO perspective, seems to have escaped such an 'anti-unintended bias'. ISO depicted social institutions as the unintended consequences of interdependent individual actions, yet not of human design. Such an interpretation highlighted the resulting institutions, or the processes leading to such consequences, and not the deviation from the linear purposive action. An astute exposition of the distinction between sociology and economics of the unintended in terms of research strategy was offered by James S. Coleman (1996: 387), one of the most skilled theorists in the UPE genre from the rational choice theory. According to Coleman, economics added a formal apparatus and demonstrated that equilibrium is possible (at least given the frame of reference).

This approach, which asserts that institutions are the unintended consequences of human actions, is highly embedded in political philosophy, history, and general economics. The classical illustrations pertain to Bernard Mandeville's, *Fable of the Bees: or, Private Vices, Public Benefits* (Schneider, 1990), Hume's theory of morality and justice as spontaneously generated orders, Ferguson's analysis of the evolution of social institutions, Smith's invisible hand, and Alexis de Tocqueville's progress of democracy.

The introduction of ISO is regarded as one of the greatest sociologically relevant contributions of the eighteenth century Scottish Enlightenment. Ramifications of its premises can be followed up in Georg Wilhelm Friedrich Hegel's 'cunning of reason,' as well as in Karl Marx's and Friedrich Engels' treatments of unintended consequences which are revealed as fatal contradictions. Particularly, the invisible-hand metaphor is commonly traced to Mandeville's fable. As it occurred rather sporadically in Smith ([1759]1976, [1776]1975, [1795]1982), its original meaning and background continue to be an issue. Important however is the fact that it evolved in symbiosis with the spontaneous order perspective and contributed to the development of the broader ISO.

In addition to the integration of the intellectual luggage of the invisible-hand notion, the spontaneous order theory found its cues in the works of ancient Chinese philosopher, Chuang Tzu, and the school of Salamanca's writings on economics in the late medieval period. This was followed by the strong contributions from the Scottish school, and from the nineteenth century onward, the compact developments of modern economists and political philosophers, most vividly from the Austrian School of Economics. Carl Menger's explanation of the spontaneous emergence of 'organic phenomena' served as the start of the latter contribution. Further, the twentieth century witnessed a proliferation of ISO explanations as well as an abundance of their critical accounts. In addition to Menger's explanation of the emergence of money, the most prominently referred to processes are Thomas Schelling's model of racial residential segregation, Ludwig von Mises's explanation of the formation of the medium of exchange, Friedrich A. Hayek's 'catallaxy' (i.e., spontaneous market order – see Barry, 1982), and Robert Nozick's account of the rise of the ultraminimal state.

With regards to the theoretical essence of ISO, this was comprehensively depicted in Nozick's (1974, 1994) presentation of the 'invisible-hand' and 'hidden-hand explanations,' and Hayek's ([1976]2003) development of the theory of spontaneous order. According to Nozick, in the invisible-hand explanation, a designed institution is shown to be the outcome of the interaction of actors having no such overall goal in mind. Further, in the hidden-hand explanation, an apparently accidental institutional structure is shown to originate from intended design. In the case of Hayek's treatment, the theory of spontaneous order was used as an argument in the programmatic opposition to the market socialism and central planning models.

As matters currently stand, the relation between the two concepts (the invisible hand and spontaneous order) is unclear, particularly as the former is subject to reinterpretations. Seemingly, the employment of one term over the other seems to be a matter of intellectual taste and background rather than of a straightforward delimitation of separate characteristics.

Current Refinements and Changes in Focus over Time

Although a comparative account of the contributions by Merton to the sociology of the unintended and the legacy of Smith and the Scottish Enlightenment in economics of the unintended are yet to be written, some elements could be pointed out still.

Merton is credited with the introduction and quite successful initial treatment of several concepts, such as unanticipated consequences of action, purposive social action, manifest and latent functions, self-fulfilling and self-defeating prophecy, boomerang effect, Matthew effect, and serendipity

pattern (Sztompka, 1986, 2011). The application area of these terms ranges from unanticipated consequences of purposive social action to cumulative and aggregate effects.

In comparison, Smith is basically associated with the invisible-hand metaphor – a term of appealing designation that was integrated as element of the Scottish Enlightenment's ISO perspective, and further evolved with the Austrian School of Economics. Moreover, the development of this thesis blurred the distinction between economics of the unintended and general economics, not to mention the fact that it pertained to programs outside these two fields (such as philosophy, public policy, biology, and more recently, the naturalistic philosophy of science). Such a strong connection between the UPE problem and particular schools of thought is lacking in the sociology of the unintended. Here, it is rather witnessed the strong consequential appeal of certain subdisciplines, sociological theories, and authors linked with their formulation.

As far as current sociology is concerned, the further prevalence of UPE can be witnessed in the more or less critical extension of Mertonian notions. The perspective is also common to rational choice sociology, analytical Marxism, and economic sociology. The Mertonian conceptual luggage notwithstanding these recent treatments should not be perceived as sole footnotes to Merton's legacy. They belong to consequential authors in their own right, who also benefited from other intellectual influences. Their works advanced the articulation of other dimensions which should be taken into account, in addition to the one of intentionality, when analyzing the unintended consequences of action – see Jon Elster's (1990) discussion of the issue of recognition and functionality, and Baert's (1991) dimensions and modes of unintended consequences. They sophisticated the classification of consequences to the extent that Merton's original treatment of unintended effects became just one of the variants being mentioned – see Portes's (2000, 2010) linear purposive action and alternative behavioral patterns. In addition to rendering and articulating the intellectual heritage of the sociology of the unintended, the sociologists also became more careful and candid in indicating that the notion of unintended does not imply strictly 'undesirable' or 'unexpected' outcomes (Boudon, 1982).

Despite this element of zooming out from the unexpected and undesirable, sociology further reproduces the interest in unexpected consequences of purposive action – whether individual or collective – and in the paradoxical and perverse character of the outcome – see Portes's (2000) 'sociology as analysis of the unexpected,' as well as the 'stronger definition' of Raymond Boudon's 'perverse effect' (Van Parijs, 1982). This preoccupation with the counterintuitive and paradoxical is also visible in accounts dealing with aggregate and cumulative effects – see Elster's (1978) 'counterfinality' and Coleman's (1994) 'social anomalies' brought about by seemingly rational behavior. In addition, it also transpires in the recent theoretical recovery of the paradox of consequences in Weber's work (Cherkaoui, 2007).

When it comes to the state of the art in ISO, two lines of development stand out. The first one pertains to the reevaluation of intellectual roots and the initial meaning of invisible-hand mechanisms in Smith's work (see for instance, Viner, 1972; Rothschild, 2002), as well as to critical perspectives regarding the use of this concept in economics (Barry, 1985; Aydinonat, 2008; Samuels et al., 2011). The reassessment endeavor is impressive and there is no real precedent of such criticism toward related concepts in sociology.

The second notable path is the further refinement of ISO. The general trend comes from interpretations which focus on end-states to rather abstract models, explicating the processes involved in the transition to unintended outcomes – for a review of models of institutions as unintended consequences of human actions in game-theoretic literature and agent-based computational economics, see N. Emrah Aydinonat (2008). Another angle for ordering the literature is through the lenses of Edna Ullmann-Margalit (1978a,b) who framed the evolution of the invisible-hand explanations from nonevolutionary models (the classical invisible hand) to evolutionary ones that became ideologically used against liberals and social planners. She distinguished between aggregate and evolutionary models – the former regarding the occurrence of social institutions, the latter concerning their maintenance (see also Hamowy, 2008).

As stated, according to Nozick (1974: 19; 1994), the invisible-hand explanations depict accounts of "what looks to be the product of someone's intentional design, as not being brought about by anyone's intentions." Although launched within political philosophy, this term has great theoretical potential in both economics and sociology. First, it integrates various treatments of the unintended in these disciplines. Second, it resonates theoretically with the concept of mechanisms, which itself emerges as one of the most relevant intersections between contemporary sociology and economics of the unintended (Cowen, 1998). Hence, there are indications that some of the recently launched notions might provide a common framing across academic disciplines which, in spite of a discursive consciousness claiming a common heritage, studied the phenomenon of the unintended rather in isolation.

Alternative and Emerging Directions in Theory and Research

Still the influence of UPE within sociology and that of ISO within economics should not be taken as thoroughly dominating the treatment of the unintended in these two fields.

When considering sociology, we see that manifestations of ISO enjoy more than an occasional occurrence – see the treatment of the emergence of social order from individual actions in rational choice theory and methodological individualism for instance. The peculiarity of the sociology of the unintended is not the absence of the social order problem (it builds after all on the tradition of the invisible-hand idea), but the fact that it inherited the tendency to qualify the unintended as unexpected and singular, sometimes even as unwanted, and focused on social action and the frustrations of actors' interests.

On the other hand, there are also sociological treatments which tried to dispel the approach that frames the unintended as unexpected. Arguably, the first sociologist to have normalized and discussed unintended consequences as a universal component of social life is Norbert Elias (see Mennell, 1977; van Krieken, 1998: 50). Following, Giddens's (1984) theory of

structuration offered a related demystification of the phenomenon. Inspired by his teacher Elias, Giddens depicted the unintended as recursive side effects which are further constitutive of the circumstances of social action.

In terms of contemporary approaches to the unintended via the side effects, the most vivid manifestation is the 'risk society' angle, within the broader reflexive modernization theory. This framed the 'unintended (second order; institutional) side effects' as consequences of modern social institutions which result in new conditions that call these institutions into question in the context of the second modernity (Beck et al., 2003: 14–15). Here, the "'meta-change' of modern society results from a critical mass of unintended side-effects" (Beck et al., 2003: 2).

The impact of the risk-society paradigm on the sociology of the unintended is manifold. Worthy to note is the fact that it made the concern with unintended consequences of social intervention (modern developments and projects in particular), and the unanticipatable outcomes of collective processes running blindly, visible within mainstream theoretical sociology. Furthermore, it promoted the topic of 'side-effects of side-effects' – e.g., the 'politicization' of side effects and the manner in which the unintended effects result in 'institutional turbulence' and legitimacy debates in global environmental politics (Beck et al., 2003: 14–15).

In Beck's 'first modern society,' he depicted the side effects in a similar manner to externalities within economics. There are two nuances in economists' definition of these. The first one presents the situation when "the activity of one entity (a person or a firm) directly affects the welfare of another in a way that is outside the market mechanism" (Rosen, 2002). Commonly, it is distinguished between positive (e.g., knowledge) and negative externalities (industrial pollution). The second nuance denotes "all the connections, relations, and effects which agents do not take into account in their calculations when entering into a market transaction" (Callon, 1998: 16). These definitions do not quite exclude each other. Importantly however, they presuppose distinct research focuses and conceptualizations in the analysis of the unintended.

In the former interpretation, the spotlight is on the initial intention. This is broadly understood as overall interests of the third party (who is not involved in the original transaction) and the perception of the desirability of the effects for the actor who was not accounted for (positive/negative/neutral). Depicted in these terms, the externalities resemble the side effects from the contemporary sociological discourse on unintended consequences. For economics, these are linked with market failure and pose a theoretical and ideological challenge to the ISO idea.

When it comes to the second framing, this concentrates on attempts to anticipate and include the effects of transactions in actors' calculations and negotiations. The point of interest is less the intention of the action and more the framing of the transaction and the contract. Such activities have the potential to become visible-hand processes, while the invisible-hand concept is linked to market failure. In reference to Ronald H. Coase's (1960) theory that externalities are the results of transaction costs, such understanding differentiates two categories of economic events: 'outcomes generated by encompassing contracts and externalities not covered by contract' (Cowen, 1998: 133). Tyler Cowen considered that this classification would overmatch the one of intended and unintended consequences altogether.

This type of perspective on externalities catches this double theoretical move toward depicting the unintended and negative consequences of collective phenomena, and toward researching the issue of their recording and inclusion in the calculation of costs and negotiations between actors. In other words, this is a research path preoccupied with advancing various strategies (from public to private) of internalizing the negative externalities (Callon et al., 2009) – a common preoccupation for economists and sociologists of the unintended alike.

The Unintended Consequences of Purposive Social Intervention

The recent trend in the study of unintended consequences in sociology and economics introduced the research of the unexpected via the side effects of modernizing programs. In somewhat more radical approaches, these side effects were shown to lead to surprises, paradoxes, and perverse effects of public policy, regulation, and high technology (Sieber, 1981; Brooks, 1986; Thompson et al., 1990; Grabosky, 1995; Tenner, 1997; Hood and Peters, 2004; Roots, 2004; Margetts et al., 2010; Winczorek, 2011). The case studies and conceptual tools advanced by this genre consider the unintended consequences of purposeful intervention. This means that they analyze consequences of attempts to effect social change. According to Sieber (1981: 9), "the term *social intervention* denotes any sort of deliberate effort to alter a human situation in some desired direction, such as a welfare program, a military expedition, an organizational structure, or a law." Although the study of unintended consequences of social intervention has a predilection for the paradoxical and perverse outcome, there are also analyses of phenomena which do not run as explicitly counter to the intentions of the institutional change initiators (Wildavsky, 1980; Pawlak, 2011).

Generally, the study of unintended consequences of purposeful intervention is an investigation field for public policy analysis, planned organizational change, evaluation research, social problems, and the analysis of the impact of high technology on society. In addition, it also surfaces in particular areas of sociology, such as applied sociology, and urban and environmental sociology. From a theoretical and conceptual point of view, it is evidently closer to sociology than to economics of the unintended. The theoretical apparatus builds on sociological framings of the unintended, yet it also advances potent original formulations and classifications. Overall, the observation that the discourses about unintended consequences and risks are socially constructed and are culturally embedded stands out. In addition to the issue of emergence, the studies on outcomes of social intervention also focus on the management of the unintended by various organizational and regulatory regimes.

When incorporating sociological insights, the research on unintended consequences of social intervention mainly builds on Herbert Spencer's application of the idea of regressive effects to governmental programs, on Weber's thesis of the indirect

and unintended effects of rational ascetism, and on Merton's classification of limitations to the anticipation of consequences of social action. The most illustrative author who incorporated and critically approached the sociological knowledge of unexpected outcomes is Sieber (1981). He advanced seven conversion mechanisms which can lead to regressive consequences of purposeful intervention. In addition to such rather traditional references, the analysis of social intervention also rejuvenated the interest in Simmel's and Edward A. Ross' sociologies of the unexpected, and showed their relevance for studying unintended consequences in 'high modernity' and 'knowledge society' (Gross, 2003, 2010).

When it comes to conceptual framings which are divergent from the sociological undertaking, the analysis of risks and unintended consequences by employing cultural theory stands out (see Douglas and Wildavsky, 1984; Thompson et al., 1990; Douglas, 1992). Though this perspective is still pretty much under construction, its prospects of development are quite promising, especially in light of the recent reactivation of the theory by Christopher Hood (2000; Hood and Peters, 2004). The grid-group analysis of unintended consequences has potential because it reveals the rhetorical structure of perspectives on risks and unexpected effects, and the fact that such discourses are culturally embedded.

Concerning the rhetoric of unintended consequences, the most powerful theorist of social intervention by far is Hirschman (1991). This author's input is manifold – see, for example, the perverse effect as a particular thesis of the conservative-reactionary response to purposeful intervention. Further, Hirschman (1975) also pointed out that the rhetorical sensitivity of organizations to particular types of unintended consequences is determined by the overall self-perception of their effectiveness (see discussion of 'fracasomania' – the failure complex), meaning that the social scientists have to be aware of possible exaggerations regarding the pervasiveness of unintended consequences and the implications of such discourses.

The research of unintended consequences of purposeful intervention also points to the management of unintended consequences and risks (Wildavsky, 1988). Notably in this regard is Karl Weick and Kathleen Sutcliffe's (2007) study of high-reliability organizations' (HROs) response to critical situations, and Diane Vaughan's (1996) analysis of the tendency to normalize the unexpected in the 1986 NASA Challenger disaster – see also Charles Perrow's (1984) normal accident. Comparatively to the policy analysis of surprises and paradoxes, the impact of high technology discusses the unexpected and accidents occurring in society.

Observably, the unintended effects of modernization programs and high technology is the social site where the three main types of investigation of unintended consequences – i.e., pertaining to purposive social action; to human action, but not human design; and to purposive social intervention – can most probably bridge a common conceptual framework. It is also noticeable that in spite of common intellectual luggage (traceable to classical authors), and unsystematic attempts to reach a more integral perspective, the three fields on the unintended evolved rather autonomously. The materialization of theoretical affinities that is recently discernable is based on the openness of sociology to influences from economics of the unintended, and also on the refinement of some sociological and philosophical concepts by the policy analysis of the unintended. This theoretical turn is somehow unexpected in light of the fact that sociology has a conceptual vocabulary similar to one of the policy analysis, but anticipatable given the attractiveness of the confident analyses put forward by economics.

See also: Action, Theories of Social; Alternative Schools of Economic Thought; Economic Sociology; Institutions; Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures; Methodological Individualism in Sociology; Modernity; Modernization, Sociological Theories of; Norms; Policy Analysis: Evidence Based Policy-Making; Rational Choice Theory in Sociology; Risk, Sociology and Politics of; Technology and Social Control.

Bibliography

Aydinonat, N.E., 2008. The Invisible Hand in Economics: How Economists Explain Unintended Social Consequences. Routledge, New York.

Baert, P., 1991. Unintended consequences: a typology and examples. International Sociology 6, 201–210.

Barry, N.P., 1982. The tradition of spontaneous order. In: Literature of Liberty, vol. V. Institute for Humane Studies, Arlington, VA, pp. 7–58.

Barry, N.P., 1985. In defense of the invisible hand. Cato Journal 5, 133–148.

Beck, U., Bonss, W., Lau, C., 2003. The theory of reflexive modernization: problematic, hypotheses and research programme. Theory, Culture & Society 20, 1–33.

Boudon, R., 1982. The Unintended Consequences of Social Action. Macmillan, London.

Brooks, H., 1986. The typology of surprises in technology, institutions and development. In: Clark, W.C., Munn, R.E. (Eds.), Sustainable Development of the Biosphere. International Institute for Applied Systems Analysis. Cambridge University Press, Cambridge, pp. 325–350.

Callon, M., 1998. An essay on framing and overflowing: economic externalities revisited by sociology. In: Callon, M. (Ed.), The Laws of the Markets. Blackwell Publishers/The Sociological Review, Oxford, pp. 244–269.

Callon, M., Lascoumes, P., Barthes, Y., 2009. The democratization of democracy. In: Callon, M., Lascoumes, P., Barthes, Y. (Eds.), Acting in an Uncertain World. An Essay on Technical Democracy. The MIT Press, Cambridge, MA and London, pp. 225–254.

Cherkaoui, M., 2007. Good Intentions: Max Weber and the Paradox of Unintended Consequences. Bardwell Press, Oxford.

Coase, R.H., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Coleman, J.S., 1994. A rational choice perspective on economic sociology. In: Smelser, N.J., Swedberg, R. (Eds.), Handbook of Economic Sociology. Princeton University Press, Princeton, pp. 166–180.

Coleman, J.S., 1996. A Vision for Sociology. In: Clark, J., Coleman, James S. (Eds.), Rotledge-Falmer, London and New York, pp. 383–389.

Cowen, T., 1998. Do economists use social mechanisms to explain? In: Hedström, P., Swedberg, R. (Eds.), Social Mechanisms: An Analytical Approach to Social Theory. Cambridge University Press, Cambridge, pp. 125–146.

Douglas, M., 1992. Risk and Blame: Essays in Cultural Theory. Routledge, London and New York.

Douglas, M., Wildavsky, A., 1984. Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers. University of California Press, Berkley, CA.

Elster, J., 1978. Logic and Society. Contradictions and Possible Worlds. John Wiley & Sons, Chichester and London.

Elster, J., 1990. Merton's functionalism and the unintended consequences of action. In: Clark, J., Modgil, C., Modgil, S. (Eds.), Robert Merton: Consensus and Controversy. Falmer Press, London, New York, Philadelphia, PA, pp. 129–135.

Ferguson, A., [1767]1966. An Essay on Civil Society. Edinburgh University Press, Edinburgh.

Giddens, A., 1984. The Constitution of Society: Outline of the Theory of Structuration. Polity Press, Cambridge.
Grabosky, P., 1995. Counterproductive regulation. International Journal of the Sociology of Law 23, 347–369.
Gross, M., 2003. Sociologists of the unexpected. Edward A. Ross and Georg Simmel on the unintended consequences of modernity. The American Sociologist 34, 40–58.
Gross, M., 2010. Ignorance and Surprise. Science, Society, and Ecological Design. Massachusetts Institute of Technology, Cambridge, MA.
Hamowy, R. (Ed.), 1987. The Encyclopedia of Libertarianism. Sage Publications, London, New Delhi and Singapore.
Hamowy, R., 2008. The Scottish enlightenment and the theory of spontaneous order. University of Southern Illinois Press, Carbondale and Edwardsville.
Hayek, F.A., [1976]2003. Cosmos and taxis [Ch. 2 from Law, Legislation, and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy]. In: Hechter, M., Horne, C. (Eds.), Theories of Social Order: A Reader. Stanford Social Sciences, Palo Alto, pp. 221–236.
Hirschman, A.O., 1975. Policymaking and policy analysis in Latin America: a return journey. Policy Sciences 6, 385–402.
Hirschman, A.O., 1977. The Passions and the Interests. Political Arguments for Capitalism before Its Triumph. Princeton University Press, Princeton, NJ.
Hirschman, A.O., 1991. The Rhetoric of Reaction: Perversity, Futility, Jeopardy. The Belknap Press of Harvard University Press, Cambridge, MA and London.
Hood, C., 2000. The Art of the State: Culture, Rhetoric and Public Management. Paperback ed. Clarendon, Oxford.
Hood, C., Peters, G., 2004. The middle aging of new public management: into the age of paradox? Journal of Public Administration Research 14, 267–282.
van Krieken, R., 1998. Norbert Elias. Routledge, London and New York.
Lal, D., 1998. Unintended Consequences: The Impact of Factor Endowments, Culture, and Politics on Long-run Economic Performance. MIT Press, Cambridge, MA.
Margetts, H., Perri, 6, Hood, C. (Eds.), 2010. Paradoxes of Modernization. Unintended Consequences of Public Policy Reform. Oxford University Press, New York.
Mennell, S., 1977. "Individual" action and its "social" consequences in the work of Norbert Elias. In: Human Figurations. Essays for Norbert Elias. Amsterdams Sociologisch Tijdschrift, pp. 99–109.
Merton, R., 1936. The unanticipated consequences of purposive social action. American Sociological Review 1, 894–904.
Mica, A., Peisert, A., Winczorek, J. (Eds.), 2011. Sociology and the Unintended. Robert Merton Revisited. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Warszawa, Wien.
Nozick, R., 1974. Anarchy, State, and Utopia. Basic Books, New York.
Nozick, R., 1994. Invisible-hand explanations. The American Economic Review 84, 314–318. Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association.
Pawlak, M., 2011. Unintended consequences of institutional work. In: Mica, A., Peisert, A., Winczorek, J. (Eds.), Sociology and the Unintended. Robert Merton Revisited. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Warszawa, Wien, pp. 355–370.
Perrow, C., 1984. Normal Accidents: Living with High-Risk Technologies. Basic Books, New York.
Portes, A., 2000. The hidden abode: sociology as analysis of the unexpected: 1999 presidential address. American Sociological Review 65, 1–18.
Portes, A., 2010. Economic Sociology: A Systematic Inquiry. Princeton University Press, Princeton and Oxford.
Popper, K., 1957. The Poverty of Historicism. The Beacon Press, Boston.
Popper, K., [1945]2013. The Open Society and Its Enemies. Princeton University Press, Princeton.
Roots, R.I., 2004. When laws backfire: unintended consequences of public policy. American Behavioral Scientist 47, 1376–1394.
Rosen, H.S., 2002. Public Finance. McGraw-Hill/Irwin, Boston.
Rothschild, E., 2002. Economic Sentiments. Adam Smith, Condorcet, and the Enlightenment. Harvard University Press, Cambridge, MA and London.
Samuels, W.J., Johnson, M.F., Perry, W.H., 2011. Erasing the Invisible Hand: Essays on an Elusive and Misused Concept in Economics. Cambridge University Press, New York.
Schneider, L., 1975. The Sociological Way of Looking at the World. McGraw-Hill Book Company, New York, St. Louis, San Francisco.
Schneider, L., 1990. Paradox and Society: The Work of Bernard Mandeville. Aldine de Gruyter, New York.
Sieber, S., 1981. Fatal Remedies. The Ironies of Social Intervention. Plenum, New York.
Smith, A., [1759]1976. The Theory of Moral Sentiments. Clarendon Press, Oxford.
Smith, A., [1776]1975. An Inquiry into the Nature and Causes of the Wealth of Nations. Clarendon Press, Oxford.
Smith, A., [1795]1982. The history of astronomy, edition used: essays on philosophical subjects. In: Wightman, W.P.D., Bryce, J.C. (Eds.), vol. III of the Glasgow Edition of the Works and Correspondence of Adam Smith. Liberty Fund, Indianapolis.
Sztompka, P., 1986. Robert K. Merton. An Intellectual Profile. Macmillan, Houndmills, Basingstoke, Hampshire and London.
Sztompka, P., 2011. Existential uncertainty and its remedies. On the shoulders of Robert K. Merton. In: Mica, A., Peisert, A., Winczorek, J. (Eds.), Sociology and the Unintended. Robert Merton Revisited. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Warszawa and Wien, pp. 75–87.
Symonds, M., Pudsey, J., 2008. The concept of "paradox" in the work of Max Weber. Sociology 42, 223–241.
Tenner, E., 1997. Why Things Bite Back: Technology and the Revenge of Unintended Consequences. Alfred a Knopf, New York.
Thompson, M., Ellis, R., Wildavsky, A., 1990. Cultural Theory. Westview, Boulder, CO.
Ullmann-Margalit, E., 1978a. The invisible hand and the cunning of reason. Social Research 64, 181–198.
Ullmann-Margalit, E., 1978b. Invisible-hand explanations. Synthese 39, 263–291.
Van Parijs, P., 1982. Perverse effects and social contradictions: analytical vindication of dialectics? The British Journal of Sociology 33, 589–603.
Vaughan, S., 1996. The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press, Chicago.
Vernon, R., 1979. Unintended consequences. Political Theory 7, 57–73.
Viner, J., 1972. The Role of Providence in the Social Order. American Philosophical Society, Philadephia.
Weick, K.E., Sutcliffe, K.M., 2007. Managing the Unexpected. Resilient Performance in an Age of Uncertainty, second ed. John Wiley & Sons, Inc., San Francisco.
Wildavsky, A., 1980. The Art and Craft of Policy Analysis. The Macmillan Press Ltd, London and Basingstoke.
Wildavsky, A., 1988. Searching for Safety. Transaction, New Brunswick, NJ.
Winczorek, J., 2011. Why do procedures have unexpected outcomes? In: Mica, A., Peisert, A., Winczorek, J. (Eds.), Sociology and the Unintended. Robert Merton Revisited. Peter Lang, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Warszawa, Wien, pp. 307–334.

United Nations and its Agencies

Janice W Wetzel, International Association of Schools of Social Work, United Nations
Nancy E Wallace, World Federation for Mental Health, United Nations

© 2015 Elsevier Ltd. All rights reserved.

Abstract

As the social work profession increases its international presence in the twenty-first century, involvement with selected United Nations (UN) bodies is vital. The interface between social work and UN agencies that are most relevant to the concerns of both will be addressed. Important information regarding the UN will be presented to enable knowledgeable participation. Examples of social work advocacy over the years, likely to remain relevant for decades to come will be described, alongside recommendations for social work's future involvement.

The United Nations' Social Agenda

The social agenda of the United Nations (UN) and that of the social work profession are closely aligned. Both are concerned with a broad range of psychosocial issues that span social justice and human rights, promotion of peace, economic and social development, eradication of poverty, disaster relief, the advancement of women, and support for other marginalized populations including children, older people, immigrants, and disabled people (Dominelli, 2010). The promotion of health and psychosocial well-being are also shared concerns. While each UN agency may focus their efforts on particular areas of involvement, one should be aware that they are all committed to human rights regardless of specialization (Healy and Link, 2012).

The UN system is comprised of the General Assembly, the Security Council, the Economic and Social Council, and its specialized agencies, funds, and programs (Wilson, 2012). The UN Charter calls for formal relationships between non-governmental organizations (NGOs) and the UN like those representing international social work and social welfare. The work of NGOs includes consulting with Member States, providing technical expertise, advocacy, participating in joint operational projects, disseminating information, and raising awareness between the UN and civil society. International social work organizations are among these NGOs.

International Social Work Organizations at the UN

The International Association of Schools of Social Work (IASSW), the International Federation of Social Workers (IFSW), and the International Council on Social Welfare (ICSW) were all founded in Paris in 1928 at the First International Conference on Social Work. These global social work organizations, the largest in the world, hold consultative status with the Economic and Social Council (ECOSOC) at the UN and with the Department of Public Information (DPI), among others. All three organizations participate as NGOs in principal UN locations including New York, Geneva, and Vienna and have regional offices globally.

IASSW promotes the development of social work education throughout the world and develops standards to enhance the quality of social work education and research in 3000 schools of social work at tertiary level in 74 countries. They are concerned with social justice and social development through human rights-oriented policy and advocacy activities. IASSW was the first social work organization to acquire Special Consultative Status in 1947 and was present at deliberations on the Universal Declaration of Human Rights (Healy, 2012).

IFSW provides a global voice for the profession with over 750 000 members in 90 countries. They strive for social justice, human rights, and social development through the promotion of social work and best practices. IFSW is affiliated with the United Nations Children's Fund (UNICEF), and like its sister organizations, also works with the World Health Organization (WHO), the United Nations High Commissioner for Refugees (UNHCR), and the Office of the High Commissioner for Human Rights (OHCHR) (Hall, 2012).

ICSW promotes social development, social welfare, and social justice undertaken by tens of thousands of independent community organizations in 70 countries that work with people in poverty, hardship, or distress. Its membership is made up of multiple professions, including social workers. They involve a network of major international organizations, as well as people who are themselves experiencing hardship, working in their own communities. ICSW serves a range of vulnerable, marginalized people across the life span in numerous cultures. Their work with the UN has long been collaborative across governments, missions, and NGOs (Correll, 2012).

Together, these three organizations launched a Global Agenda for Social Work and Social Development that was presented at the UN at the 29th Annual Social Work Day in March 2012. They agreed on an action agenda to prioritize collaborative activities between 2012 and 2016 around four themes: promoting social and economic equalities, advancing the dignity and worth of all people, working toward environmental sustainability, and strengthening recognition of the importance of human relationships. Action steps will be reported annually through the Global Agenda Observatory (Zelenev, 2013). A European Observatory for Social Work and Social Development was launched in 2013. There will be a special supplement on the Global Agenda presented in 2014 that will be used to provide data to influence the UN.

As a first subsequent effort, IASSW, ICSW, and IFSW made a joint Statement for the High Level Segment at the Annual

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28098-7

Session of ECOSOC held in July 2012 in New York, addressing the Annual Ministerial Review theme, "Promoting productive capacity, employment and decent work to eradicate poverty in the context of inclusive, sustainable and equitable economic growth at all levels for achieving the MDGs."

Social Work Day at the UN was founded under the aegis of IFSW in 1983. It became a joint enterprise with IASSW in the early 1990s. Over the years, the event has attracted thousands of students and professionals. It is designed to educate social workers and students about the UN and the UN about social work. The themes of Social Work Day have emanated from the shared concerns of the UN and the international social work community. Topics have ranged from practice and policy issues regarding human rights, women, children, families, indigenous populations, mental and physical health, social development and disasters – both natural and artificial (Dominelli, 2012). In recent years, a Global Social Work Student Day, organized by students, has accompanied this event.

Education of the community, UN advocacy, and social action are important social work roles. Examples of social workers' involvement include providing training, workshops, and curriculum development for the profession on humanitarian, social, and economic development issues; and utilizing international human rights conventions to support local advocacy efforts. They also participate in UN World Summits on social development, sustainable development, women, HIV/AIDS, xenophobia, climate change, and aging. They write policy statements and make presentations to governments and the UN through the NGO forums associated with these events.

Social workers engage in activities under the aegis of the UN Conference of NGOs (CoNGO), the umbrella organization that facilitates the participation of NGOs in the UN system. Affiliation allows social workers and other NGOs to address themes and issues through participation in pertinent NGO committees. Social workers, for example, were instrumental in founding and providing leadership for the UN NGO Committee on Mental Health. Its creation has fostered collaboration among NGOs that has strengthened efforts to bring mental health issues to the global agenda at the UN. They introduced World Mental Health Day to the UN, which formally incorporated the event in its annual calendar of celebrations. A social worker delivered the first address concerning women and mental health at World Mental Health Day in 1996. The committee has been an important vehicle for advocating social work concerns.

Social work educators have established schools of social work where none existed before, e.g., Argentina, the Cameroons, and Vietnam; undertaken study trips in countries to address social issues such as trafficking in women and children, aging populations, organized crime, health and mental health, economic crises, natural disasters, and other issues that require urgent attention and benefit from social work perspectives.

UN World Conferences and Summits

Social work NGOs have been involved with the UN for decades. The launching of world conferences and on-going events since 1990 has fostered increased opportunities for advocacy and decision-making. Social workers provided leadership at these conferences, serving on numerous planning committees and drafting policies, conducting global meetings and panels, and coordinating the conferences themselves in some instances. If the profession is to be a viable participant in the twenty-first century, social work leadership and visible involvement will be required at all appropriate UN ventures.

There are nine World Conferences that have influenced governments and NGOs. These conferences continue to have periodic follow-up meetings to evaluate progress on the action plans and integrate emerging issues in order to plan for the future. In-depth information regarding each conference is available on the Internet. World Conference action plans indelibly changed how development has been conceptualized. While economic and social policies previously excluded the reality of people's lives, all development programs are now people-centered and focused on eliminating poverty and fulfilling basic needs, alongside realizing human rights, investing in health and well-being, education and training, and improving the status of women. Training and technical consultations have been replaced by capacity building concepts and sustainability. Social work is well-suited to join in such a participative model. See http://www.un.org/geninfo/bp/intro.html for a comprehensive discussion of these new approaches.

Millennium Development Goals

Any discussion of the UN and its priorities of relevance to social work must cover the Millennium Development Goals (MDGs). Building on the work of earlier World Conferences, the eight MDGs were launched in 2000 by consensus of member states for the purpose of reducing extreme poverty by 2015 and aimed to:

1. Eradicate extreme poverty and hunger;
2. Achieve universal primary education;
3. Promote gender equality and empower women;
4. Reduce child mortality;
5. Improve maternal health;
6. Combat HIV/AIDS, malaria, and other diseases;
7. Ensure environmental sustainability; and
8. Develop a global partnership for development.

A 2014 MDG report indicates that progress has been made on some fronts, but the MDGs will not be fully realized by the target date. Consequently, governments in cooperation with civil society are developing a post-2015 agenda to meet today's challenges. This bears in mind, lessons learned from the MDG framework, while taking into account current realities that did not exist at the start of the millennium. While the importance of the MDGs is not disputed, there are many who regret the absence of other key concerns such as the escalating global issues of aging, mental health, violence, and disasters. NGOs are advocating for their inclusion. Social workers have important roles to play in the post-2015 process to ensure the goals are met and social work interests are included in it.

Sustainable Development Goals

In July, 2014, the UN Open Working Group (OWG) of the General Assembly concluded 18 months of consultations and

negotiations to produce an Outcome Document which contains 17 proposed goals and 169 related targets creating the recommended framework for the Sustainable Development Goals (SDGs) for post-2015. Health is highlighted in Goal 3: "Ensure healthy lives and promote wellbeing for all at all ages." This includes a target to "reduce by one-third premature mortality from non-communicable diseases through prevention and treatment, and promote mental health and wellbeing" (SDGs, 2014). This should be of particular importance for social work practice, advocacy, and education.

Non-Communicable Diseases

On the threshold of the post-2015, SDGs, one of the greatest challenges, is in the area of global health. Of particular interest to social workers are Non-Communicable Diseases (NCDs) which include mental and neurological disorders. In 2011, the General Assembly mandated a comprehensive review and assessment of the progress achieved in the prevention and control of NCDs. In July 2014, UN member states, the UN system, and civil society met for the UN NCD Review to assess the progress and gaps. Emphasis in the final outcome document focused on developing a set of commitments and actions for global and national implementation, prevention and treatment across the life span, mobilization of resources and recognition of NCDs as a priority in the post-2015 development agenda (NCDs, 2014). This critical emerging issue has important implications for social work and opportunities for collaboration with the UN and civil society in supporting and advocating for implementation, accountability, and civil society capacity building. A UN publication, The Global Conversation Begins: Emerging Views for a New Development Agenda, has been launched to assess progress (United Nations, 2013).

Post-2015

The year 2015 marks the 70th anniversary of the United Nations. It is both a target year and the 20th anniversary for a number of historic internationally agreed upon documents (UN Agreed Upon Documents, 2014). The close of the MDG process galvanized a post-2015 agenda to the year 2030. This time, the focus is on SDGs, a more inclusive, intergenerational perspective that also aligns with Rio + 20, the world's Sustainable Development Conference held in June of 2012.

The 20th Anniversary of the Copenhagen Declaration and Summit on Social Development is also a 2015 event in which social workers are engaged. Concerns about inequalities and advocacy for a social protection floor will continue to be crucial undertakings. The end of poverty continues to be a worldwide post-2015 objective.

The year 2015 is also the 20th anniversary of the Beijing Platform for Action, passed in 1995 at the 10-year follow-up to the end of the Decade of Women, also called the Fourth World Conference on Women. ECOSOC will review the 12 essential areas of concern that are designed to achieve the goals of gender equality, development, and peace. It has become clear over the last two decades that there are forces that are devoted to rolling back the gains of the international women's movement. Therefore, it was decided not to hold a global women's conference as had been done previously, but to launch a review of every government's progress and challenges to making the Beijing Platform for Action a reality. UN Women is committed to ensuring that gender equality is a fact across the world, locally, nationally, and internationally.

Convention on the Elimination of All Forms of Discrimination against Women

A core principle leading to Beijing was based upon CEDAW, the Convention on the Elimination of All Forms of Discrimination Against Women, created at the middle of the Decade of Women in Copenhagen. CEDAW is essentially a Bill of Rights for Women. To date, 187 out of 194 countries have ratified and signed the agreement. Those that have not are the United States, Iran, Sudan, South Sudan, Somalia, and two small Pacific Island nations (Palau and Tonga). There is a US movement designed to provide pressure for implementation at the national level that include Cities for CEDAW and CEDAW 2014. Attention will also be given to the many governments who have signed on but are found wanting in its execution, a major concern of the United Nations and the women of the world (CEDAW, 2014).

UN Agencies Relevant to Social Work

Economic and Social Council

ECOSOC, the body most relevant to social work at the UN, is concerned with the world's economic, social, and environmental challenges. Policies are recommended following discussion of pertinent issues by member states. Meetings are held throughout the year with prominent academic and business leaders, including 3200 registered NGOs in consultative status with ECOSOC.

NGO affiliation is assured by the Charter of the United Nations, Article 71. Eligibility requirements, rights, and obligations are governed by ECOSOC resolution 1996/31. Consultative status is granted by ECOSOC upon recommendation of the ECOSOC Committee on NGOs. Every year, at meetings convened by ECOSOC, NGO representatives including social workers are invited to provide technical expertise, interventions, and position statements that address the topics being discussed. One example is the "Intervention on Participation in Poverty Reduction," written by the IASSW representative to the UN Commission on Social Development.

Department of Public Information

DPI was established to promote UN global awareness and understanding using media including publications, radio, TV, the Internet, and video conferencing. DPI partners with diverse audiences throughout the world, including NGOs at the UN in New York where it maintains an NGO Resource Center. During weekly briefing panels, topics being highlighted at the UN are discussed. NGOs are given the opportunity to affiliate with DPI and access the UN just as they do through ECOSOC. Both

IASSW and IFSW are affiliated with the DPI. DPI also coordinates an annual, well-attended international conference in various locations around the world. These meetings are planned by a large group of NGO volunteers, including social workers, in cooperation with UN staff.

Department of Economic and Social Affairs

Department of Economic and Social Affairs (DESA), the think tank for the UN, works closely with ECOSOC and other divisions. The Social Integration Branch, Division for Social Policy and Development, is particularly relevant to social work. This provides substantial intergovernmental support at the UN, while facilitating international cooperation on social and economic development issues. Key concerns are hunger and malnutrition, sustainable development, disability, indigenous people, employment issues, and social protection. Their work includes aging, youth, family, and inclusive policy issues that are submitted to the UN Commission for Social Development, the ECOSOC, and the General Assembly.

DESA's mandate calls for increasing capacities in developing countries so that they might better function and contribute nationally and internationally. Capacity development is the process by which people, organizations, and societies become self-sustaining, capable of achieving their own goals. Building capacities increases the likelihood that the MDGs and post-2015 development agendas will be realized. Social workers, educated to facilitate capacity building, are natural partners.

United Nations Development Program

United Nations Development Program (UNDP) is the UN's global development network. UNDP operates on multiple levels – local, national, regional, and global – in 177 countries and territories, and within the UN itself. Like social workers, UNDP advocates for change and helps people to connect to resources at local, regional, and country levels. Partnership is at the center of all aspects of their work, which focuses on solutions to development challenges as voiced by people themselves. The protection of human rights and empowerment of women are embedded prominently in their mission.

UNDP is also concerned with capacity development as an evolving approach based on research. They have identified four drivers of change that yield significant capacity and are assessable: institutional arrangements, leadership, knowledge, and accountability. UNDP was instrumental in setting the standard for gender mainstreaming at the UN. Thus, all policies, practices, and programs are assessed through a gender-specific lens. UNDP and social work are kindred spirits. International social work organizations could engage further with the UNDP Resident Representative in each country to strengthen their voice in UN development activities.

UN Human Rights Council

The mandate of the UN Human Rights Council (HRC) is to promote and protect human rights worldwide, identifying and developing responses to challenges and violations. The main office of the HRC, located in Geneva, Switzerland, is supported by the OHCHR. OHCHR provides the UN with research, education, and advocacy. Assistance is also given to governments. OHCHR helps to implement social justice, legislative reform, and electoral processes to facilitate human rights standards. Consistent with the creation of the UN Charter, the OHCHR is charged with mainstreaming human rights by incorporating a human rights perspective in all UN programs. To that end, it partners with governments, civil society, and all UN bodies and international organizations, including NGOs.

Social workers have been active with HRC and OHCHR, advocating extensively on human rights concerns, including writing manuals on human rights and providing policy statements and consultations on human rights violations. IFSW's and IASSW's policies on human rights are utilized as the basis for advocacy with these UN entities. For example, IFSW's policy on refugees addresses their physical needs for food, shelter, clothing, and health, alongside psychosocial needs essential to human well-being. Other social work interventions have included human rights coalition building with other Consultative NGOs. IASSW and IFSW collaborated with the UN in 1994 on a Human Rights Manual for educating social workers (UN, 1994). Social workers have long encouraged the inclusion of international human rights in social work curricula (Ife, 2012; Wronka, 2007, 1995; Reichert, 2007; Wetzel, 2000). A number of the world's schools are becoming totally human rights oriented.

Office for the Coordination of Humanitarian Affairs

Office for the Coordination of Humanitarian Affairs's (OCHA) mandate includes coordination of interagency decision-making in response to complex emergencies. Participants include NGOs and the International Federation of Red Cross and Red Crescent Societies, as well as the UN agencies. Without their expertise, responses may be fragmented and inappropriate. Relief is needed in the short term, providing necessary resources and meeting basic human needs. Disaster response must also engage in sustainable long-term efforts and be grounded in a human rights orientation and research to be effective.

Social workers should be key players in the dissemination of information and coordination of responsible disaster policies and services on the ground and have an important role to play in humanitarian aid work (Tan, 2009). Social work has a comprehensive perspective regarding psychosocial responses to natural and artificial disasters ranging from hurricanes and tsunamis to conflict-related emergencies and the infrastructures to ensure professional accountability and deliver training (Dominelli, 2013a). Like OCHA, social workers advocate for the rights of people in need and can be partners in fostering sustainable programs. They are well positioned to ensure support to meet clients' needs, strengthen leadership coordination and evidence-based humanitarian decision-making, planning, and resource allocation.

Social workers have been at the forefront of climate change awareness and disasters for many years, including issuing policy statements on the subject at the United Nations (Dominelli, 2012). Social workers at the UN in collaboration with other NGOs have addressed the impact of climate change on mental health and psychosocial well-being (Weissbecker et al., 2007).

Social work practitioners and academics have been involved in disaster response and preparedness, providing education and training for the UN, governments, universities, municipalities, first responders, and health care workers (Desai, 2007). The Inter-Agency Standing Committee (IASC) under the auspices of OCHA is mandated to improve delivery of humanitarian assistance to populations affected by natural disasters, conflict-related emergencies, global food crises, and pandemics. Social workers have assisted IASC in producing a series of guidelines for responding to humanitarian crises. These have included gender-based violence, disasters, social development, and psychosocial interventions. The IASC Guidelines for Mental Health and Psychosocial Support in Emergency Settings are widely used and have been lauded worldwide as exemplary practical psychosocial interventions that eliminate stigma, inappropriate practices, and policies and integrate a human rights approach.

UN Women

The UN Charter (United Nations, 1945) and the Universal Declaration of Human Rights (United Nations, 1948) reflect the United Nations' long-standing commitment to the plight of women. In 2010, in response to the UN reform agenda, the United Nations General Assembly created UN Women, the UN Entity for Gender Equality and the Empowerment of Women. In doing so, it provided an umbrella organization equipped to merge resources and mandates for greater impact. The offices involved in the merger were: the Division for the Advancement of Women (DAW), International Research and Training Institute for the Advancement of Women (INSTRAW), Office of the Special Adviser on Gender Issues and Advancement of Women (OSAGI), and United Nations Development Fund for Women (UNIFEM). Information is available on the UN Women website.

Despite concerted efforts and progress made on behalf of women over the past decades, UN Women notes that gender inequalities are entrenched throughout the world, denying women basic human rights (United Nations, 2009, 2006). Not only is this disparity oppressive to women, but it also affects their families and communities and has vast socioeconomic ramifications. UN Women is not alone in recognizing that empowering women fuels thriving social and economic development. Social workers have long called for the advancement of women and their human rights (Reichert, 2007; Wetzel, 1993). Curricula in many countries entail awareness of the situation of women and girls, and the importance of supporting and facilitating changing attitudes toward them (Dominelli, 2013b). Social work and UN Women are natural partners.

United Nations Children's Fund

UNICEF, operating in 190 countries, is dedicated to the rights of all children, having the global authority to influence decision-makers at the highest levels, and establishes relationships with grassroots partners, including young people, who have innovative ideas that need assistance to turn them into reality. The organization is the standard bearer for the Convention on the Rights of the Child, essentially the bill of rights for the world's children. UNICEF is committed to removing the challenges that block women and, in turn, the development of children, whether concerned with poverty, malnutrition, violence, preventable disease, or discrimination. Recognizing that girls are at risk throughout the world, UNICEF works to promote their education as fundamental to their advancement and well-being.

Social workers representing many NGOs have been involved with UNICEF for years. In 2012, UNICEF commissioned a Child Protection Systems Strengthening paper that consolidates current thinking and lessons learned, including the need for a strong child protection workforce (McCaffery et al., 2012). Participation of social workers at UNICEF-sponsored conferences has led to other key initiatives by international social work organizations. The journal, *International Social Work*, for example, published special issues on Globalization/Indigenization (Lyons, 2010) and on Children's Rights in Africa (Sewpaul and Matthias, 2013).

Working Group on Girls

The Working Group on Girls (WGGs) is dedicated to advancing their rights and status. Their efforts focus on the implementation of the commitments to girls' rights that have been established by international agreements. Advocacy in the form of policy statements, program development, and resource allocations includes the participation of girls as agents of change in their own families and communities.

World Health Organization

WHO is the coordinating authority for health within the United Nations system. It provides leadership on global health and mental health issues, shapes the research agenda, sets norms and standards, and provides best practices based upon evidence-based policy research.

The WHO's Constitution states that "health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (WHO, 1946, 2006). They recognize that depression affects more than 350 million people of all ages, in all communities, and is a significant contributor to the global burden of disease. Women worldwide are twice as likely to experience depression as are men, regardless of economic status, racial, or ethnic background. By the year 2020, WHO predicts that it will be the second leading cause of disability, with women most affected (WHO, 2008).

In response to the escalating global mental health risk, WHO launched its Mental Health Gap Action Programme in Geneva. Although aware of needs in developed countries, its aim is to increase services for mental health and substance abuse in low- and middle-income countries (WHO MhGAP, 2008). At the core of mhGAP is the establishment of productive partnerships that address psychosocial interventions, including family- and community-based treatment and rehabilitation for depression, alcohol, and drug dependence. More social workers are needed in many developing countries and rural areas if their objectives are to be realized. Regardless of cultural differences, early intervention, basic education, and training of professional and lay people are essential (Wetzel, 2012).

WHO's most recent initiative is the development of the 2013–2020 Global Mental Health Action Plan, which is a transformative agenda on the global burden of mental disorders and the need for a comprehensive, coordinated response from health and social sectors at the country level. This provides an important opportunity for social workers to become more involved in psychosocial policy development and advocacy.

United Nations Educational, Scientific, and Cultural Organization

UNESCO is a specialized agency of the UN system, based in Paris. It is devoted to fostering peace throughout the world. Universal respect for social justice and the rule of law are advanced by promoting collaboration among nations through education, science, culture, and shared communication. Human rights and fundamental freedoms are core affirmations regarding race, sex, religion, and language.

UNESCO's celebrations during the 60th Anniversary of the Universal Declaration of Human Rights at its headquarters in Paris in 2008 provided a forum for diverse participants from governments, media, academia, and the private sector. Its central objective was to identify and acknowledge the most effective ways in which civil society was making a positive impact on respect for human rights. Under the aegis of the DPI, programs included a panel of experts on "International Community Mental Health Education: Human Rights-Based Grassroots and Professional Models" that involved IASSW as organizer and moderator. Panelists' shared messages included the empowerment of indigenous people as leaders; the creation of schools of social work in developing communities; and the unique relevance of social work in such communities.

UNESCO has objectives well suited to partnering with international social work education. Relying primarily on research, training, and teaching activities, UNESCO strives for the advancement, transfer, and sharing of knowledge. Technical cooperation encompasses the preparation and adoption of international instruments and legal recommendations that are consistent with social work goals.

Social Work Skills, Best Practices, and the UN

The following best practices are suggested for international social workers. Most are familiar to the profession, but the provisos may be thought provoking.

Psychosocial Perspective

The strengths perspective concerned with the resourcefulness and resilience of people, long embraced by social workers, is an appropriate global orientation, particularly when applied within a psychosocial framework.

Self-Determination and Cultural Relativism

Human rights in a global context should respect and embrace alternative cultural practices. Provided that the customs are not abusive, they can be very effective. While cultural sensitivity and cultural competence are essential social work values to be honored, cultural relativism must not overshadow human rights. When a local practice is abusive, even if it is accepted by the community, abuse in the name of self-determination is still abuse (NASW, 1997).

Human Rights-Based Social Work and Gender Mainstreaming

By adopting human rights as the foundation principle upon which all of social work theory and applied knowledge rest, social workers involved with individuals, families, small groups, large communities, or institutions will have a value-based tool in which to apply their knowledge and skills. Gender mainstreaming is the UN human rights method designed to ensure the appropriate assessment of practices, policies, and programs in terms of their disparate impact on girls and women and boys and men across the life span. If they are not appropriate for everyone, it is not appropriate that we engage in them (OSAGI, 2008).

Human Rights and Social Justice

Social justice is a significant but insufficient concept in that it is inherently associated with cultural mores that may limit one's rights. Human rights encompass social justice and transcend civil, economic, and political customs, in consideration of the basic life sustaining needs of all human beings, without distinction (NASW, 1997).

Capacity Building, Personal and Social Development

Many definitions of capacity building have emerged since the latter part of the twentieth century when the UNDP promoted capacity building to alleviate poverty as an international policy directive. Increasing the capacities of individuals through enhancing their potential ability and skills, so that they can become fulfilled members of society, is a shared objective with social work. Whether working with groups or communities, increasing culturally-appropriate self-awareness is a universal means of increasing self-esteem, so essential to personal development (UNDP, 2013).

Train the Trainers Model

Train the Trainers is a model that has evolved from the international women's movement in the developing world since the mid-twentieth century. It is designed to address the shortage of professionals, while empowering women to take charge of their own lives and futures. The model is also suitable for men. A participatory, non-hierarchical group-dynamic approach is taken whereby experts train leaders who in turn train others in an organization or community. They in turn train others, and so on. The approach has been used in comprehensive health programs, with factory workers, and micro-credit programs throughout the world (Femnet, 2013; Wetzel, 1993).

Participative Social Action Research and Programming

Participatory action research is a cyclical process that begins with diagnosing problem situations, planning action steps, and

implementing and evaluating the outcomes that in turn lead back to reassessment based upon data collected. The findings are applied to effective policy and programs relevant to the real world of participants. By including the vital input of stakeholders, the likelihood that studies will be implemented is increased. In the process, individuals, organizations, and communities each develop capacities that enhance success and development. The empowering participatory process of social action research is as effective an intervention as the end result (Powers and Allaman, 2012; Participatory Action Research Toolkit, 2010).

Social Work and Global Economics

It is recommended that social work education consider including more information on economics, particularly from a global perspective (Ortiz and Cummins, 2012; Staub-Bernasconi, 2007). Formerly grounded in a colonialization perspective, international development policy has progressed over the years to focus on the empowerment and inclusion of recipients. It is important that social workers understand the relationship between economics and social conditions, and the impact of economic policy on development policy. For example, HIV/AIDS donor nations have had strings attached that hamper recipients on the local level. Despite good intentions, Western governments, the World Bank, International Monetary Fund, and United States Agency for International Development have had to consider debt forgiveness in the face of growing poverty due to structural adjustment requirements (Lewis, 2005).

Social Work and the Social Compact

In developing countries where entitlements are often nonexistent, human rights are championed by NGOs (Gibney, 2003). Throughout the world, in recent years, the social compact that has provided a safety net for people in need has been waning. Social work can act as a catalyst to help to reactivate each nation's commitment on behalf of disenfranchised populations by encouraging social workers to join grassroots community coalitions. Everyone can learn from one another all over the world, appropriately transferring, adapting, and applying their hard-earned knowledge and skills to the problems in their own countries (Jones, 2007; Hokenstad and Midgley, 2004). Social workers and social work educators also must continue to teach students how to develop resources, build coalitions, engage in advocacy, and provide support and missing expertise where needed (Hokenstad, 2008). By doing so, the United Nations will find social workers to be invaluable partners.

See also: International Association of Schools of Social Work; International Council on Social Welfare.

Bibliography

Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), 2014. http://www.cedaw2014.org/ (accessed 30.07.14.).

Correll, D., 2012. International council on social welfare (ICSW). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. Oxford, New York, pp. 286–291.

Department of Economic and Social Affairs (DESA), 2014. http://www.un.org/en/development/desa/index.html (accessed 30.07.14.).

Dominelli, L., 2010. Social Work in a Globalizing World. Polity Press, Cambridge.

Dominelli, L., 2012. Green Social Work. Polity Press, Cambridge.

Dominelli, L., March 13, 2013a. Learning from the 2004 Tsunami: Guidelines for Empowering Disaster Interventions and the Development of Resilient Communities. Humanitarian Aid Seminar. Strand Hotel, Copenhagen.

Dominelli, L., March 2013b. Environmental justice at the heart of social work practice: greening the profession. International Journal of Social Welfare. http://dx.doi.org/10.1111/ijsw.12024.

Femnet, 2013. Train the Trainers Curriculum for Gender Mainstreaming. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CB8QFjAA&url=http%3A%2F%2Fwww4.worldbank.org%2Fafr%2Fssatp%2Fresources%2Fhtml%2Fgender-rg%2FSource%2520%2520documents%255CTraining%2520materials%2FTRGEN1%2520Femnet%2520Gender%2520Mainstreaming%2520TOT.pdf&ei=QGfZU6K6O86VyASblYHYDw&usg=AFQjCNGsD4_8NY495wYXzLTshg3-kO344Q&sig2=gBN2jZW7zVsPypBlEeN9WQ&bvm=bv.72185853,d.aWw (accessed 30.07.14.).

First International Conference of Social Work, 1929, July 8–13, 1928. Conference Proceedings. Author, Paris.

Gibney, M.J. (Ed.), 2003. Globalizing Rights. Oxford Amnesty Lectures 1999. Oxford Press, UK.

Hall, N., 2012. International federation of social workers. In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. Oxford, New York, pp. 275–280.

Healy, L.M., 2012. International association of schools of social work (IASSW). In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. Oxford, New York, pp. 281–285.

Healy, L.M., Link, R.J. (Eds.), 2012. Handbook of International Social Work. Oxford, New York.

Hokenstad Jr., M.C., 2008. International social work: education. In: Mizrahi, T., Davis, L.E. (Eds.), Encyclopedia of Social Work, twentieth ed. Oxford University Press and NASW Press, New York, pp. 488–493.

Hokenstad Jr., M.C., Midgley, J., 2004. Lessons from Abroad. NASW, Washington, DC.

Ife, J., 2012. Human Rights and Social Work: Towards Rights-based Practice, third ed. Cambridge University Press, Cambridge, UK.

International Federation of Social Workers (IFSW) and International Association of Schools of Social Work (IASSW), 2009. Human Rights Manual, second ed. http://ifsw.org/resources/publications/human-rights/ (accessed 02.08.13.).

Jones, D.N., November 2007. News and views…from IFSW: renewed global interest in the profession. In: Borowski, A. (Ed.), On Human Dignity and Social Work, International Social Work, vol. 50, no. 6, pp. 859–861. http://isw.sagepub.com/content/50/6/859.full.pdf+html (accessed 08.08.13.).

Lewis, S., 2005. Race against Time. Anansi Press, Toronto.

Lyons, September 2010. In: Globalization/Indigenization Special Issue. International Social Work, vol. 53, pp. 595–598. http://dx.doi.org/10.1177/0020872810378796 (accessed 08.08.13.).

McCaffery, J., Davis, R., Conticini, A., 2012. Strengthening Child Protection Systems in Sub Saharan Africa: A Working Paper. UNICEF and the Inter Agency Child Protection Working Group. http://www.unicef.org/wcaro/english/strengthening_child_protection_systems_in_sub-Saharan_Africa_-_August_2012_.pdf (accessed 08.08.13.).

NASW (National Association of Social Workers), 1997. International policy on human rights. In: Social Work Speaks: NASW Policy Statements, fourth ed. NASW Press, Washington, DC, pp. 188–191. Author.

Non-Communicable Diseases (NCDs), Outcome Document. http://www.un.org/en/ga/president/68/pdf/letters/782014Non-communicable%20diseases_Outcome%20Doc-8%20July%202014.pdf (accessed 30.07.14.).

OSAGI, 2008. Gender Mainstreaming. Office of the Special Advisor on Gender Issues and Advancement of Women. DESA. www.un.org/womenwatch/osagi/gendermainstreaming.htm (accessed 02.08.13.).

Ortiz, I., Cummins, M. (Eds.), 2012. A Recovery for All: Rethinking Socio-economic Policies for Children and Poor Households. UNICEF. http://www.unicef.org/socialpolicy/index_62107.html (accessed 30.07.14.).

Participatory Action Research Toolkit, 2010. INCITE: Women of Color against Violence. http://www.incite-national.org/media/docs/5614_toolkitrev-par.pdf (accessed 04.08.13.).

Powers, C.B., Allaman, E., 2012. How participatory action research can promote social change and help youth development. In: Boyd, D., Palfrey, J., Sacco, D. (Eds.), The Kinder & Braver World Project: Research Series. Born This Way Foundation & the Berkman Center for Internet & Society at Harvard University (Supported by the John D. and Catherine T. MacArthur Foundation).

Reichert, E. (Ed.), 2007. Challenges in Human Rights: A Social Work Perspective. Columbia University Press, New York.

Sewpaul, V., Matthias, C. (Eds.), January 2013. Special Issue on Child Rights in Africa. International Social Work, vol. 56, no. 1. http://isw.sagepub.com/site/Special_Issues/Special_Issues_Index.xhtml (accessed 08.08.13.).
Staub-Bernasconi, S., 2007. Economic and social rights: the neglected human rights. In: Reichert, E. (Ed.), Challenges in Human Rights: A Social Work Perspective. Columbia University Press, New York, pp. 138–161.
Sustainable Development Goals (SDGs), 2014. http://sustainabledevelopment.un.org/index.php?menu=1549 (accessed 30.07.14.).
Tan, N.T., June 2009. Disaster management: strengths and community perspectives. Journal of Global Social Work Practice 2 (1).
United Nations, 2013. The Global Conversation Begins: Emerging Views for a New Government Agenda. UN, New York.
United Nations, 2014. Non-Communicable Diseases Outcome Document. http://www.un.org/en/ga/president/68/pdf/letters/782014Non-communicable%20diseases_Outcome%20Doc-8%20July%202014.pdf (accessed 17.07.14.).
United Nations, 2009. Violence against Women. http://www.un.org/en/women/endviolence/pdf/factsheets/unite_the_situation.pdf (accessed 27.12.12.).
United Nations, 1994. Human Rights and Social Work: A Manual for Schools of Social Work and the Social Work Profession (Professional Training Series No.4). UN, New York and Geneva.
United Nations, 1945. Charter of the United Nations and Statute of the International Court of Justice. http://www.un.org/en/documents/charter/index.shtml (accessed 15.07.14.).
United Nations, 1948. Universal Declaration of Human Rights. UN, Geneva.
United Nations Agreed Upon Documents, 2014.
United Nations Secretary-General, 2006. UN Women. In-depth Study on All Forms of Violence against Women. http://www.un.org/womenwatch/daw/vaw/v-sg-study.htm (accessed 26.12.12.).
Weissbecker, I., Wallace, N.E., et al., 2007. Working Group on the Impact of Climate Change on Mental Health and Psychosocial Well-being: Guidelines for Action. Report to the Secretary General. 60th Annual NGO/DPI Conference on Climate Change. United Nations, New York. http://excelser.net/climatecaucus/mental.htm (accessed 04.08.13.).
Wetzel, J.W., 2012. Global mental health. In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. Oxford University Press, New York, pp. 214–219.
Wetzel, J.W., 2007. Human rights and women: a work in progress. In: Reichert, E. (Ed.), Challenges in Human Rights: A Social Work Perspective. Columbia University Press, New York, pp. 162–187.
http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CB8QFjAA&url=http%3A%2F%2Fwww.unicef.org%2Fsocialpolicy%2Ffiles%2FA_Recovery_for_All_FINAL_Web.pdf&;ei=amXZU8HKJ42pyATE5IKACQ&usg=AFQjCNGbi7DVD2byZ3dsEtx2ja8iAIWhhA&sig2=IJHp3YnpnAqgfFCmey7D0w&bvm=bv.72185853,d.aWw (accessed 30.07.14.).
Wetzel, J.W., April 2000. Women and mental health: a global perspective. International Social Work 43 (2), 205–215.
Wetzel, J.W., 1993. The World of Women: In Pursuit of Human Rights. MacMillan Press Ltd, London.
Wilson, S., 2012. The united nations agencies. In: Healy, L.M., Link, R.J. (Eds.), Handbook of International Social Work. Oxford University Press, New York, pp. 305–310.
WHO MhGap, 2008. Mental Health Gap Report. Geneva. http://www.who.int/mental_health/mhgap/en/.
WHO, 2008. World Report on Disability and Rehabilitation. http://www.who.int/disabilities/actionplan/en/ (accessed 30.07.14.).
WHO, 1946, 2006. Constitution of the World Health Organization. www.who.int/governance/eb/constitution/en/ (accessed 05.08.13.).
Working Group on Girls. http://www.girlsrights.org (accessed 20.07.14.).
Wronka, J., 2007. Human Rights and Social Justice. Sage, London.
Wronka, J., 1995. Human rights. In: Edwards, R.L. (Ed.), Encyclopedia of Social Work, nineteenth ed., vol. 1. NASW Press, Washington, DC, pp. 1405–1418.
Zelenev, S., 2013. IASSW, ICSW and IFSW establish global agenda observatory. International Social Work 56, 420. http://isw.sagepub.com/content/56/3/420.full.pdf+html (accessed 04.08.13.).

Relevant Websites

www.ngocongo.org/ – Conference of Non-Governmental Organizations.
http://www.un.org/en/development/desa/index.html – Department of Economic and Social Affairs.
www.un.org/en/hq/dpi/ – Department of Public Information.
www.iassw-aiets.org – International Association of Schools of Social Work.
www.icsw.org – International Council of Social Welfare.
www.ifsw.org – International Federation of Social Workers.
http://www.un.org/millenniumgoals/ – Millennium Development Goals.
www.undp.org – United Nations Development Fund.
http://www.unwomen.org/ – UN Women.
www.unesco.org/ – UNESCO.
www.who.org – World Health Organization.
http://www.who.int/mental_health/mhgap/en/ – WHO Mental Health Gap Action Programme (mhGAP).

Universities and Science and Technology: Europe

Barbara Sporn, Vienna University of Economics and Business Administration, Vienna, Austria

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 15974–15978, © 2001, Elsevier Ltd.

Abstract

Science and technology at European universities have changed substantially. With the emergence of a global knowledge economy, science (as basic research) and technology (as a product of applied research) are more tightly coupled to form the basis for national wealth creation. Universities are expected to provide well-trained scientists and engineers, new knowledge, and applied science. European systems of higher education have been challenged by these new economic conditions. State budgets are increasingly constrained. Ministries have been deregulating authority to the institutions. Information technology has influenced knowledge production and delivery. A more diverse student body has been seeking a university education as part of their life-long learning need. Globalization and international competition have pushed European universities to revise and reorganize their institutions. Consequently, academic science and technology are in a process of restructuring. New knowledge transfer strategies are emerging that aim at increased alliances between companies and universities. Research has become more interdisciplinary and problem oriented. Market-based funding develops as a system supporting industry-related as opposed to curiosity-driven research. More management capacity within universities is needed to deal with these changes. Overall, universities must refocus their mission in order to integrate knowledge production (research), knowledge dissemination (teaching), and knowledge application (outreach/service).

The role of science and technology at European universities has been undergoing tremendous changes (Etzkowitz and Leydesdorff, 1997; Zinberg, 1991). Given the increased importance of information technology and the emerging global knowledge-based economy, scientific research and development, as well as academic training, form the basis of national policies and institutional strategies to promote economic growth. As a consequence, universities are expected to respond to the need for well-trained scientists and engineers, new knowledge, and applied science (Zinberg, 1991). The industry and technology-oriented academic fields like engineering and business management find it easier to combine science (as basic research) with technology (as a product of applied research) in order to meet the new requirements of the economy and to raise additional funds. Other disciplines in the social sciences or humanities may have greater difficulty in maintaining a key position within universities and obtaining necessary resources (Slaughter and Leslie, 1997).

In order to understand the current trends in academic science and technology in Europe, it is important to describe the institutional varieties of universities. Trends in different European countries serve as a basis for analyzing new challenges for higher education systems. Consequences for the production and management of knowledge show the new imperative for the future, which emphasizes a redefinition of the relationship between universities, industry, and the state in Europe (Etzkowitz and Leydesdorff, 1997; Neave and Van Vught, 1991).

Academic Science and Technology in Europe

European systems of higher education consist of a variety of universities and colleges (Gellert, 1993; Teichler, 1996). On a very general level, two types of universities prevail: one of a comprehensive character and the other specializing in one field or area. The comprehensive university represents a conglomerate of different faculties, academic programs, and consequently a larger number of students. Specialized universities are focused on a particular academic discipline or field, such as business administration or engineering, with only one faculty and fewer students. Specialized universities tend to focus on applied research and professional training, thereby responding more easily to external needs of the labor market and private industry. Comprehensive universities are more likely to be dedicated to basic research and training (Teichler, 1996).

The modern European university which combines research and teaching emerged in the early nineteenth century and went through a lively period of transitions up to the present. The proliferation of fields (e.g., biotechnology, telecommunications, electronic commerce) and new institutional types (e.g., business schools, vocational colleges) have been based on the increasingly diverse research and education needs of national economies. In retrospect, this differentiation of functions appears to have been a consequence of the new requirements of the labor market and changes in the knowledge infrastructure (Etzkowitz and Leydesdorff, 1997). At the same time, the role of higher education shifted emphasis from social mobility to wealth creation (Nowotny, 1995). In other words, universities had to find new forms to translate science into technology, moving away from earlier concerns with liberal education and curiosity-driven research (Slaughter and Leslie, 1997). Many European governments and ministries have played a pivotal role in this shifting of emphasis in higher education systems by supporting universities through research money and other financial incentives (Etzkowitz and Leydesdorff, 1997).

These new trends in the social and economic context of universities have also changed the role of knowledge (Gibbons et al., 1994). Traditionally, institutions of higher education across the world have performed core missions and functions: knowledge production (research), knowledge dissemination (teaching), and knowledge application (service–outreach). For

universities operating in a social context of application, research has to accumulate not only disciplinary-based knowledge but also useful results for society. The reason for the increasing application–orientation of higher education systems, especially in Europe, has been mass higher education leading to new forms of knowledge production focusing more on a link between science and technology. As a consequence, the university has moved from an ideal model (i.e., social change) to a pragmatic model (i.e., wealth creation) (Gibbons et al., 1994; Trow, 1973).

In Europe, mass higher education not only entails a sharp increase in students, faculty, and institutions. Generally, it also implies increasing institutional diversity, organizational complexity, and academic heterogeneity. Also, public pressure has challenged the ideal of autonomous universities as 'ivory towers' and has called for more accountable institutions (Dill and Sporn, 1995). In the following section these challenges are analyzed more closely.

New Challenges for European Systems of Higher Education

Colleges and universities in Europe have become part of mass higher education systems because of expansion and diversification. Expansion refers to the large increases in student numbers in most European countries caused by public policies to provide education to a larger portion of the population. Diversification means mostly a binary system of higher and vocational education translating into different types of institutions for different target groups (Sporn, 1999). In this process, knowledge production and dissemination have become very important. "The driving force behind accelerated supply and demand of marketable knowledge lies in the intensification of international competition in business and industry. In many cases in-house research is no longer sufficient to meet competitive demands. In order to commercialize knowledge, firms have to look for new types of links with universities, government laboratories as well as with other firms" (Gibbons et al., 1994, p. 46).

An additional factor aggravating the socioeconomic challenges for higher education has to do with the trend to 'scientification' of employment and work (Teichler, 1996). The service and production industry has become increasingly dependent on the results of research. Additionally, a growing number of positions are labeled as 'knowledge work' and call for high skills and solid academic training. Management positions, especially in large companies with a wide reach, require a university degree (Teichler, 1996).

As many observers of higher education have outlined, major challenges within European universities have occasioned the following changes (Gibbons et al., 1994; Sporn, 1999): diversification of functions, different social profiles of student population, more professional education, diverse funding of higher education, shifting power between academic and administrative roles, focus on efficiency and effectiveness, and new forms of knowledge production. In this realm, science and technology have gained importance by being more tightly linked with each other and with external constituencies. Scientific research has been translated into application-oriented results, thereby raising public credibility and legitimacy. For this amalgamation of science and technology, higher education scholars employ to the widely used term 'technoscience', implying a combination of "basic and applied research, discovery and innovation, science and product, knowledge and commodity" (Slaughter and Leslie, 1997, p. 38).

As outlined earlier, these changes in university science and technology have been based on new economic conditions in many European countries caused by budgetary constraints and new patterns of public spending. Additionally, restructured job markets call for constant relearning of the workforce. These different areas display a general cross-national dynamic and include five major trends: scarce financial resources, changing role of the state, new technologies, continuing education, and increasing globalization (Sporn, 1999).

State budgets for higher education have been restructured in many European countries. Although public funds for higher education are stable or even increasing in some European countries, they cannot respond to the rising number of students. As a consequence, new sources of funding are needed. Public spending is very often based on output measures like the number of graduates or publications. As a consequence, universities have acquired a more diversified public and private funding base and have concentrated their strategies on enhancing efficiency and effectiveness of teaching and research (Clark, 1998; Williams, 1995).

Closely linked to the problem of limited public funds are new mechanisms of coordination and control in higher education systems, especially regarding the role of the state. Many European systems formerly characterized by 'state control' have moved to a state 'supervising' model, with increased institutional self-regulation and autonomy while the state steers the system from a distance. In many cases like Sweden or The Netherlands, leadership structures have been strengthened and quality control is handled through external audits and evaluation (Neave and Van Vught, 1991). As national economies in Europe have become more volatile and dynamic, the life-cycle of knowledge has shortened and employees need to refresh their expertise more frequently. Hence, continuing education and lifelong learning have become key elements of national policies for enhancing economic productivity. Universities responded by expanding their programs, offering adult education, interdisciplinary programs, and part-time studies with a more vocational orientation. This has expanded the number of continuing, postsecondary, and professional education programs. In Europe, this trend has been labeled as the 'professionalization' of general education (Gellert, 1993; Teichler, 1996).

The new knowledge economy is mainly based on the widespread use of technology. Information technology has also changed university teaching, research, and management. The expansion of the Internet and affordable hardware and software have led to computer literacy of large parts of the student population. The impacts on how students learn, professors teach and conduct research, and administrators manage the institution are complex. Recent trends like virtual universities and distance education programs, international research projects, and administrative information systems are a few examples of changes influenced by information technology (Sporn, 1999). Globalization and international competition

have put pressure on colleges and universities. The importance of research and development, the creation of common markets, rising mobility of students and staff, and the free movement of capital have increased the need for universities to contribute to national wealth by making the economy more competitive (Etzkowitz and Leydesdorff, 1997). Consequently, market-oriented university practices and national policies have been pushing for university–industry research and development programs which aim at stimulating regional economic development (Slaughter and Leslie, 1997).

These trends in European higher education systems have caused a process of restructuring of academic science and technology. Developments include new technology transfer strategies, shifting relations between applied and basic research, new funding patterns, and rising need for management capacity within universities.

Consequences for Academic Science and Technology

Knowledge Transfer

Government and universities have been designing strategies which should create and sustain competitive advantage in industry. The critical role of knowledge transfer either through research or through teaching has triggered new alliances between industry, government, and academic institutions. Trends include partnerships between universities and businesses, technology transfer units, incentives for a combination of basic and applied research, and for professors contributing to spin-off or start-up companies (Etzkowitz and Leydesdorff, 1997).

Knowledge transfer is often enhanced through funding for research that supports economic development. At the same time, industry relationships help universities to raise additional funds and support companies in their research and development. In this process, academic science and technology are increasingly linked to form the basis for this relationship. For example in The Netherlands, the University of Twente moved from a third-class technology-oriented institution to a successful university through the creation of new entrepreneurial practices focusing on transfer, research, and development (known as TRD activities) (Clark, 1998).

International organizations like the European Union (EU) have a high interest in knowledge transfer as well. EU projects like EUREKA reflect new state interests by establishing large supranational networks of scientists and practitioners to stimulate technological innovations. EUREKA promotes 'market-driven' collaborative research and development through the use of advanced technologies thereby strengthening European competitiveness.

Problem Orientation

As higher education systems all over Europe have expanded, research and teaching have shifted their emphasis as well. Given the complex requirements of the job market, of knowledge-based industries, and of society at large, research and teaching have to be oriented more to problems than to disciplines. This orientation helps universities to better meet the expectations and needs of external constituencies and offer relevant programs to their students. Increasingly, universities define fields of competence (e.g., environmental economics, social medicine) which combine different disciplines and areas to provide solutions to problems like pollution or AIDS (Gibbons et al., 1994; Dill and Sporn, 1995).

As a consequence, new forms of knowledge production have become necessary. Gibbons et al. (1994) coined the term 'Mode 2' knowledge production, ascribing five basic attributes to it: application context, transdisciplinarity, heterogeneity and organizational diversity, social accountability and reflexivity, and quality control. 'Mode 1' refers to the traditional discipline-oriented research of tenured faculty members. 'Mode 2' knowledge production includes a wider, more temporary, and heterogeneous set of researchers and practitioners collaborating on a problem defined in a specific and localized context (Gibbons et al., 1994).

Market-Based Funding

In Europe, scarce resources for higher education have given rise to new claims for public accountability and performance-based funding of universities. These developments have increased the importance of market mechanisms for financing higher education (Williams, 1995). Departments or groups which work in a field that is highly relevant to industry and society have more opportunities for raising additional funds and for gains in the process of resource allocation. They are also better able to contribute to the university budget through their share of overheads. These additional funds can then be used to cross-subsidize other less 'marketable' areas or programs (Clark, 1998). Mechanisms of market-based funding have also created a more entrepreneurial culture within many European universities. Faculty members responsible for their own research institute or center at universities can form a periphery which obtains additional income and also serves the public need for applied research, teaching, and service (Clark, 1998; Sporn, 1999). Again, fields which can more easily combine science and technology, like telecommunications, are especially inclined to get involved in these activities. Conflicts arise if less successful areas which belong to the tradition of university research and teaching get scrutinized for downsizing or closure (Slaughter and Leslie, 1997).

Managerial Capacity

The new relationship between universities, government, and industry requires institutional management capacity. Given the need to translate external demands into internal responses and strategies, faculty members in Europe – who have been steering universities for decades – have become overburdened. A new type of professional university manager is emerging, who is responsible for institutional efficiency and effectiveness (Gumport and Sporn, 1999). Reforms in many countries like Austria, The Netherlands, Sweden, or Spain were aimed at strengthening the leadership structure of universities.

European universities historically were dominated by strong state control and little decision-making power. But with ongoing reforms in many European countries, professional administrators and university managers have gained importance and governments have delegated many tasks to the institutional level. Consequently, in countries like Norway or

Sweden (Gornitzka et al., 1998) an increase in the number of university administrators can be observed while the size of the faculty stagnates. The rising interest in knowledge transfer based on science and technology is adding to the need for management support for new activities.

Tensions arise from these trends as to the new status of employees and the relationship between basic and applied research. While in many continental European countries university professors have been state employees, new trends call for more flexible and decentralized forms of employment. Also, at the institutional level, basic and applied research will have to be redefined. Although the basic mission of the university includes basic research a more market-oriented system requires a larger amount of applied and problem-oriented research projects.

To summarize, the link between science and technology has played a critical role within universities in Europe. National policies have converged around increasing national economic competitiveness through product and process innovations, channeling students and resources into well-funded curricula that meet the needs of a global marketplace, preparing more students for the workplace at lower costs, and managing institutions more effectively and efficiently (Slaughter and Leslie, 1997).

The expansion of European higher education systems will continue to aim at high participation rates. European economies will need more qualified specialists as well as valid research results in applied fields in order to stay abreast of major technological and scientific developments. Governments will move from a control to a supervising role, thereby trying to enhance new partnerships between higher education and industry to boost the economy. Universities have to emphasize their core commitment to a service function in order to meet these high expectations and needs. Science and technology will continue to play a pivotal role in this development.

See also: Higher Education Management; Higher Education Market; Higher Education Research; Policy Knowledge: Universities; Research and Development in Organizations; Social Science and Universities; Universities and Science and Technology: United States; von Humboldt, Wilhelm (1767–1835).

Bibliography

Clark, B.R., 1998. Creating Entrepreneurial Universities: Organizational Pathways of Transformation. Pergamon, Oxford, UK.

Dill, D.D., Sporn, B., 1995. The implications of a postindustrial environment for the university: an introduction. In: Dill, D.D., Sporn, B. (Eds.), Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly. Pergamon, Oxford, UK, pp. 1–19.

Etzkowitz, H., Leydesdorff, L., 1997. Introduction: universities in the global knowledge economy. In: Etzkowitz, H., Leydesdorff, L. (Eds.), Universities and the Global Knowledge Economy. Pinter, London, pp. 1–8.

Gellert, C., 1993. Structures and functional differentiation – remarks on changing paradigms of tertiary education in Europe. In: Gellert, C. (Ed.), Higher Education in Europe. Jessica Kingsley, London, pp. 234–246.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzmann, S., Scott, P., Trow, M., 1994. The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. Sage, London.

Gornitzka, A., Kyvik, S., Larson, I.M., 1998. The bureaucratisation of universities. Minerva 36, 21–47.

Gumport, P.J., Sporn, B., 1999. Institutional adaptation: demands for management reform and university administration. In: Smart, J. (Ed.), Higher Education: Handbook of Theory and Research. Agathon, New York, pp. 103–145.

Neave, G., Van Vught, F.A., 1991. Prometheus Bound: The Changing Relationship between Government and Higher Education in Western Europe. Pergamon, Oxford, UK.

Nowotny, H., 1995. Mass higher education and social mobility: a tenuous link. In: Dill, D.D., Sporn, B. (Eds.), Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly. Pergamon, Oxford, UK, pp. 72–89.

Slaughter, S., Leslie, L., 1997. Academic Capitalism: Politics, Policies, and the Entrepreneurial University. The Johns Hopkins University Press, Baltimore, MD.

Sporn, B., 1999. Adaptive University Structures: An Analysis of Adaptation to Socio-economic Environments of US and European Universities. Jessica Kingsley, London.

Teichler, U., 1996. Higher education and new socio–economic challenges in Europe. In: Burgen, A. (Ed.), Goals and Purposes of Higher Education in the 21st Century. Jessica Kingsley, London, pp. 96–111.

Trow, M., 1973. Problems in the transition from elite to mass higher education. In: OECD (Ed.), Policies for Higher Education. OECD, Paris, pp. 51–101.

Williams, G.L., 1995. The 'marketization' of higher education: reforms and potential reforms in higher education finance. In: Dill, D.D., Sporn, B. (Eds.), Emerging Patterns of Social Demand and University Reform: Through a Glass Darkly. Pergamon, Oxford, UK, pp. 170–193.

Zinberg, D.S., 1991. The Changing University: How Increased Demand for Scientists and Technology is Transforming Academic Institutions Internationally. Pinter, London.

Universities and Science and Technology: United States

Michael A Dennis, Georgetown University, Washington, DC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The American research university's history is inseparable from the growing importance of science and technology to the state in the late nineteenth and twentieth centuries. Successive world wars and ongoing global conflict have firmly connected academic science and technology to government patronage. Even as government support declines, federal policies shape the means through which the research university and its products enter the commercial marketplace.

Introduction

The American research university's history is inseparable from the growing importance of science and technology to the state in the late nineteenth and twentieth centuries. From the passage of the 1862 Morrill Land Grant Act to the establishment of the National Science Foundation in 1950, government and university have established deep and long-lasting connections, although the nature and character of those connections has changed profoundly over time. Once reliant upon private contributions and philanthropic foundations, the archetypical American university, as well as American science and technology, became wards of the state during the twentieth century. By the beginning of the twenty-first century, the American research university was busy recasting its social relations and its position in the nation's institutional and intellectual ecology.

In many respects, this history repeats itself. At the beginning of the twentieth century, a new kind of institution – the private research organization – emerged as a potential threat to the university as the primary site for the production of scientific and technical knowledge. Although such institutions survived, they were incapable of mounting a successful challenge to the university's intellectual hegemony. Research and teaching were powerful resources that other institutions could not match whatever their potential virtues to researchers. Today, universities and government-sponsored research are challenged by private corporations intent on capitalizing upon new discoveries, especially in the biomedical disciplines such as genomics. Ironically, the intellectual property issues that ruled out significant pre-World War II (WWII) government patronage of academic science and technology have returned in a similar guise.

Equally important, the university's internal structure has altered profoundly over the past century, a change perhaps most visible in the shrinking role of university presidents. The presidents of the late nineteenth and early twentieth centuries were dynamic individuals willing to risk reputation and even wealth for their beliefs as they sought to shape and mold institutions. Charles Eliot at Harvard and Daniel Coit Gilman at Johns Hopkins invented the research university in the United States, bringing German ideals to bear in a distinctly American fashion. Later, Harvard University president James Bryant Conant played a critical role in building the atomic bomb and developing postwar national security policy. Conant also led a vigorous public debate over the nature and character of the postwar American state. Presidents in the year 2012 have become largely fund-raisers, anonymous even within their own institutions, let alone – with some exceptions – on the wider stages of national policy.

The American College

The antebellum (i.e., before the American Civil War) American college was an institution devoted to the production of civic leaders, ministers, and school teachers. At heart, it was a character development institution complete with a psychological model of the mind legitimizing its pedagogical routines (Guralnick, 1975; Veysey, 1965). 'Faculty psychology,' the belief that distinct physical areas of the brains were the locations of different emotional and intellectual functions, was the early theory of cerebral localization undergirding the college's character-building program. Whether mandatory Latin or Greek memorization or mathematical reasoning, specific kinds of intellectual work were thought to strengthen the faculties to produce better, more capable citizens. The extent to which science and technology played a role in the antebellum college rested upon their ability to lay claim to this process of disciplining and shaping young minds. Social or economic utility was not yet an acceptable reason for the addition of science and technology to the college curriculum.

Passage of the 1862 Morrill Act granted each state federal land for the establishment of a college dedicated to the mechanic and agricultural arts. These new institutions, including Cornell University, the University of Wisconsin, the University of Illinois, and others became sites for the development of new academic roles and new kinds of knowledge. Because graduate education in the sciences was largely unheard of in the United States, the majority of instructors had studied abroad, mainly in Germany, and to a lesser extent in England and France. The land grant and its equally important sister institution, the agricultural experiment station, were hothouses for the production of a distinctive vision of American science. Utility had been the driving force in the creation of the new institutions, but many of the researchers hired at these new sites desired to transform them into institutions resembling the laboratories of their German teachers, such as the famous chemist Justus Liebig. As Charles Rosenberg (1997 [1976]) argued, the agricultural experiment station yielded the first American hybrid, the scientist entrepreneur. Obtaining resources for research demanded that academics develop techniques and practices capable of

satisfying multiple constituencies, ranging from the local farmer who wanted to know why crops would not grow in area soil, to the dean who wanted researchers to solve problems of interest to potential donors, and the politician who wanted results to justify reelection. Land grant colleges and agricultural experiment stations provided reasons for the support of scientific research so long as such research delivered useful results (Veysey, 1965; Geiger, 1986).

Research

The American research university was the product of two distinct trends – the increasing number of Americans traveling abroad for graduate education and the problems posed by the rapidly industrializing US polity following the Civil War. First, American researchers who studied in Germany came to believe that the German university was the model of a research institution. This was an illusion fostered by the Americans' peculiar status in the German university system – they were as much sophisticated tourists as students and few, if any, pursued the full course of study. American students in German laboratories worked with outstanding teachers, such as Carl Ludwig, but they failed to observe that the German university system trained more professionals – physicians, lawyers, and bureaucrats – than pure academic researchers. When American students returned to the United States to take up academic posts, they brought only part of the German university system back with them (Geiger, 1986; Owens, 1985).

The founding of the Johns Hopkins University in 1876 is often taken as the origin of the American research university. Although President Daniel Coit Gilman promoted the university as a center of *Wissenschaft* in Baltimore, Hopkins represented less Göttingen on the Chesapeake than another form of hybridization: the meshing of research with the antebellum college's character-building function. Gilman benefited from the fact that Johns Hopkins had provided in his will for the foundation of a new university and medical school. However, the meaning of Gilman's innovations – laboratory- and seminar-based instruction – is often lost in the promotion of research. One should see Gilman and the staff he handpicked – including the physicist, Henry Rowland; the chemist, Ira Remsen; the mathematician, J.J. Slyvester; and the biologist, Henry Newell Martin – as men who brought new resources and practices to bear upon old problems. Gilman and other contemporaries such as Harvard University President Charles Eliot believed that post-Civil War America was a nation in crisis, suffering from effects of unprecedented technological change and the breakdown of local social orders in the building of a national culture. Education, that most universal remedy for social ills, was once again called upon to protect and reinvent social order. Research, manifested in the labors of the laboratory and the seminar room, was the perceived antidote for a hyperkinetic market culture. It was in the laboratory, whether working on the cutting of diffraction gratings, as in Rowland's laboratory, or in the seminar room engaged in the philological study of classical texts, that the student would encounter a new form of discipline in which steady and methodical intellectual work would prepare the young mind for the task of living and working in a new social world. More importantly, such students would themselves work as teachers in the nation's colleges and secondary schools, transferring this new form of discipline and understanding to a wider populace (Kargon, 1986; Hannaway, 1976).

In the hands of educational reformers, the laboratory and the seminar room became new technologies for the (re) production of social order. They were also sites of intellectual innovation whose products might have value independent of their pedagogical origins. The *fin de siècle* saw the emergence of two new types of research institutions – the private, philanthropic research institution, such as the Carnegie Institution of Washington (CIW) (now the Carnegie Institution for Science), and the Rockefeller Institution for medical research (now the Rockefeller University), and the industrial research facility, such as those established at General Electric and AT&T. Both institutions sought to break the university's stranglehold on the production of knowledge and each was modestly successful in this ambition. In the new organizations, research had a value for its own sake unknown in the university. Part of the problem was that many American academics were unable to understand why one would want to separate research and teaching. Andrew Carnegie charged his private research facility, the CIW, with the mission of seeking out the 'exceptional man' and providing him with funds and resources to complete his work free of distractions. While the CIW consulted with academics on how to put Andrew Carnegie's generous gift to work, they were hard pressed to keep academics from applying for funds to relieve them of their teaching burden or to pay for graduate research assistants. Neither was an acceptable use of CIW funds. Robert S. Woodward, the CIW's second president, established a variety of research divisions to insulate 'exceptional' men from both the university and other demands on their time (Reingold, 1979).

Industrial research facilities were rather different, and represented a wholesale transfer of the laboratory as a knowledge production site to the factory. Instead of students and theses, the new laboratories produced patents and processes necessary for corporate survival. The high-tech industries of the early twentieth century – electricity and communications – led the way, integrating their corporate structures into the very manufacture of knowledge that might have industrial implications. Despite early press that these laboratories were universities in exile, later studies recognized that industrial research was always basic to corporate needs and goals. Successful laboratories integrated themselves into the firm's business; unsuccessful laboratories fought for independence that quickly evaporated when the firm suffered in the marketplace (Beer, 1958; Reich, 1985; Wise, 1985; Noble, 1977; Dennis, 1987).

The Visible Hand

World War I (WWI) brought about a fundamental reorganization of the university with respect to the production of technical knowledge. Although American academics made no great contributions to the development of weapons technology (save for the peripheral development of intelligence tests Carson, 1993, 2006), they organized themselves in new and powerful ways that allowed for the development of new forms of patronage. In particular, the National Research Council established by Caltech president and astrophysicist George Ellery

Hale, to resuscitate the National Academy of Sciences, proved an invaluable mechanism for the support of university-based research. The early twentieth century had seen the emergence of incredibly wealthy private philanthropic foundations, including the gargantuan Rockefeller Foundation (RF), the Carnegie Corporation of New York, and the Guggenheim Foundation (Kohler, 1991; Kay, 1993). The RF easily dwarfed all its competitors and would forever alter the topography of American higher education. In the wake of WWI, the Foundation supported a postdoctoral fellowship program administered by the National Research Council (NRC). As the historian Robert Kohler (1987) argued, the NRC was a 'trade association' for American science, providing philanthropies with a means of dispensing funds without having to support mundane university needs. That is, the foundations were not interested in supporting undergraduate education, or in building facilities; instead, the RF emphasized a policy of 'making the peaks higher.' Those who had excellent resources and professors became sites where NRC postdoctoral students clustered. In turn, those fellows left for jobs and attempted to build institutions capable of attracting such fellows.

One should note that the RF did not distribute its monies across the disciplines. The biomedical and social sciences were its primary targets of patronage; even during the Great Depression, the RF maintained its funding of these sciences, especially through efforts to import the techniques and instrumentation of the harder sciences into the biomedical and social sciences. It is to an RF officer, Warren Weaver, that one owes the term 'molecular biology,' and the RF's interest in genetics, human biology, and social engineering was to have lasting repercussions in another *fin de siècle*'s most public work, the Human Genome Project (Kohler, 1976; Kay, 1993).

Private support of the sciences was not matched by government support before WWII. Indeed, efforts to have the federal government pay for university-based research ran up against long-standing beliefs that government money invariably brought with it government control. Curiously, American academics did not seem to fear RF funds for analogous reasons. Various efforts to organize federal support of the sciences during the Great Depression foundered on this belief (Kargon and Hodes, 1985; Genuth, 1987; Hart, 1998). Equally important was a fundamental inability of university administrators and government bureaucrats to determine who would own the fruits of government sponsored research (Owens, 1990). Such disagreements made federal support of university-based research nearly impossible, save for research lacking any potential commercial value. Nor was industry a powerful patron of interwar academic science and technology. In large part, corporations sought such tight control over sponsored academic research that researchers almost welcomed the war and the possibility of relatively unconstrained government support (Galison et al., 1992). In postwar America, the state emerged for science as a patron of unprecedented wealth and power.

Wards of the State

WWII's effect upon the American university and American science and technology was profound. American researchers, organized by their leader, Vannevar Bush of Massachusetts Institute of Technology (MIT), and the CIW, instituted a virtual revolution in warfare, bringing about a permanent revolution in munitions, whether it was the atomic bomb or attempts at the mathematization of strategic thinking through operations research. Scholars, especially David Noble (1984), Stuart W. Leslie (1993), and Rebecca Lowen (1997), have begun the task of examining how the war and postwar military patronage transformed specific universities. Others, such as Paul Forman (1987, 1992, 1995), have led an effort to understand how the very content of technical knowledge changed to meet the needs of new patrons. The rise of the military as the dominant patron of American university-based science had far-reaching effects upon the university, not the least of which was a strengthening of the faculty with respect to administrators. It was the faculty with whom military officers established relationships and it was the faculty that produced both the new technologies and the students that the military wanted. In turn, administrators attempted to reassert their authority through the development of new means of accounting, most notably the invention of indirect overhead as a means of capturing revenue from grants and contracts made to specific individuals (Dennis, 1994; Galison, 1997).

Postwar military patronage focused upon four major research areas: guided missiles, continental defense, the managerial sciences, and intelligence related research in areas as diverse as area studies and cryptography. These four domains capture the range of military interest and share a common tool, the digital computer (Edwards, 1996; Ceruzzi, 2003). The development of the ballistic missile brought with it the development of miniature electronic components, including the integrated circuit and the ubiquitous microprocessor. Continental defense, the effort to develop a shield against a possible Soviet aerial attack, produced a range of technologies, including the modem and magnetic core memory. The US continental defense system, semiautomatic ground environment (SAGE), provided corporations like IBM with the technological experience to engage in massive commercial software development, including SABRE, the airline reservation system (Redmond and Smith, 2000). The managerial sciences encompassed a multitude of fields, ranging from the quantitative social sciences, such as operations research and systems analysis, to the emerging field of management science. In turn, interest in the managerial sciences, work that ranged from Simon and Ackoff's views on information processing and management to decision support systems, flourished in the reformed business school of the postwar era (Schlossman et al., 1994; Khurana, 2007). That management might be taught as a science, in the most positivistic sense, was one reason for the growth and development of graduate business education, among the most notable stories of Cold War education. Finally, the effect of the US intelligence services – the CIA and the NSA – remains among the most secret effects of the Cold War on the American university (Simpson, 1998). While it is clear that Area Studies and various departments of communications research were adjuncts to the cold war effort, what remains unclear is the extent to which research on computers, artificial intelligence, and mathematics was driven by the ever increasing needs of national security. The digital computer and its associated academic departments – applied

mathematics, linguistics, computer science, electrical and communications engineering, information theory, etc. – owe much to the nation's cryptographic needs for their development. Given that such work is still considered among the most sensitive, one may never know the depth of the state's penetration into the university.

Markets, Universities, and the State

Even before the Cold War's end, relations between universities and the state began to change. Although national security remained a significant cause for federal research funding, academic science and technology increasingly became viewed as engines of economic growth at both the regional and national levels. Hence, there were an ambitious set of efforts to 'clone' Silicon Valleys or Route 128s across the nation. If Americans were once all Keynesians, they had now become Schumpeterians, believing that the creative destruction provoked by advances in science and technology was the cure for economic stagnation (Berman, 2012; McCraw, 2010; Geiger, 2004). Market logics quickly became the grammar of academic success, especially in the sciences and engineering disciplines. Universities increasingly viewed themselves as incubators for start-up companies in fields as diverse as information technology and biotechnology. Genentech's 1980 initial public offering produced the amount of money that universities had heretofore seen only in federal grants. Blockbusters, long Hollywood's domain, became the next academic frontier as researchers sought to patent and publish, even when the federal government subsidized research. Ironically, few seemed to notice that only a handful of start-ups in any field ever became successful over time.

Even before the 1980 passage of the Bayh-Dole Act allowed universities to patent federally sponsored research, academic institutions had worked with the patent system to raise revenues. During the Great Depression, MIT and other schools had created patent policies to share monies that faculty patents might provide (Metlay, 2006). Although these policies yielded little actual revenues, by 1980 other changes made patents an increasingly viable source of funding. In addition to Bayh-Dole, which removed any questions about university patenting, several other government policies made a big difference. First, the 1978 Revenue Act cut the capital gains tax rate from 49 to 28% and the Department of Labor issued a regulatory clarification allowing pension funds to invest some of their funds in venture capital. Within a few years, there were hundreds of biotech start-ups, as biomedical researchers and venture capitalists looked for a new El Dorado. Second, the 1980 Supreme Court decision *Diamond v. Chakrabarty*, which allowed for the patenting of genetically engineered organisms, greatly expanded what the new start-ups could patent as they attempted to create new markets. Finally, the 1982 creation of a new Court of Appeals for the Federal Circuit, specializing in patent matters, made it clear that the State would work to preserve intellectual property rights. Government policy played a key role in making university-based science and technology marketable (Berman, 2012).

Commercialization had its critics, but the last decade of the twentieth century and the twenty-first century's first decade saw the triumph of a set of beliefs – neoliberalism – arguing that markets and the private sector were the real sources of economic power. Although this view found itself challenged in the wake of the financial collapse of 2008–09 when the Federal government forestalled another Great Depression, the privatization of academic American science and technology continues (Mirowski, 2011). The increasing focus upon STEM – science, technology, engineering, and mathematics – education bodes well for one dimension of the American research university, but an increasing segment of those earning degrees in those fields are foreign nationals who often return to their home country after receiving their degrees. In 2009, the last year for which there are data, 27% of the Masters degrees in the STEM fields went to non-US citizens. Just as US companies have increasingly shifted their physical manufacturing overseas, so have US universities shifted their manufacture of new technical knowledge-producers to foreigners (NSB, 2012). What the future holds remains unclear, but university presidents and faculties will have to negotiate a future in which their work is increasingly viewed as economically vital, but unworthy of massive federal support.

See also: Biocitizenship; Chemical Science: History and Sociology; Patenting in the Biological Sciences; Science and The state; Universities and Science and Technology: Europe.

Bibliography

Beer, J.J., 1958. Coal tar dye manufacture and the origins of the modern industrial research laboratory. Isis 49, 123–131.
Berman, E.P., 2012. Creating the Market University: How Academic Science Became and Economic Engine. Princeton University Press, Princeton, NJ.
Carson, J.S., 1993. Army alpha, army brass, and the search for army intelligence. Isis 84, 278–309.
Carson, J.S., 2006. The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750–1940. Princeton University Press, Princeton, NJ.
Ceruzzi, P., 2003. A History of Modern Computing, second ed. The MIT Press, Cambridge, MA.
Dennis, M.A., 1987. Accounting for research: new histories of corporate laboratories and the social history of American science. Social Studies of Science 17, 479–518.
Dennis, M.A., 1994. "Our first line of defense": two university laboratories in the postwar American State. Isis 85 (3), 427–455.
Edwards, P., 1996. The Closed World: Computers and the Politics of Discourse in Cold War America. The MIT Press, Cambridge, MA.
Forman, P., 1987. Behind quantum electronics: national security as basis for physical research in the United States, 1940–1960. Historical Studies in the Physical Sciences 18, 149–229.
Forman, P., 1992. Inventing the maser in postwar America. Osiris 7 (second series), 105–134.
Forman, P., 1995. "Swords into ploughshares": breaking new ground with radar hardware and technique in physical research after World War II. Reviews of Modern Physics 67, 397–455.
Galison, P., 1997. Image and Logic: A Material Culture of Microphysics. University of Chicago Press, Chicago.
Galison, P., Hevly, B., et al., 1992. Controlling the monster: stanford and the growth of physics research, 1935–1962. In: Galison, P., Hevly, B. (Eds.), Big Science: The Growth of Large Scale Research. Stanford University Press, Stanford, CA, pp. 46–77.
Geiger, R.L., 1986. To Advance Knowledge: The Growth of American Research Universities. Oxford University Press, New York.
Geiger, R.L., 2004. Research Universities and the Paradox of the Marketplace. Stanford University Press, Stanford, CA.
Genuth, J., 1987. Groping towards science policy in the 1930s. Minerva 25, 238–268.
Guralnick, S.M., 1975. Science and the Ante-bellum American College. American Philosophical Society, Philadelphia.

Hannaway, O., 1976. The German model of chemical education in America: Ira Remsen at Johns Hopkins (1876–1913). Ambix 23, 145–164.
Hart, D.M., 1998. Forged Consensus: Science, Technology and Economic Policy in the United States, 1921–1953. Princeton University Press, Princeton, NJ.
Kargon, R.H., 1986. Henry Rowland and the physics discipline in America. Vistas in Astronomy 29, 119.
Kargon, R.H., Hodes, E., 1985. Karl Compton, Isaiah Bowman, and the politics of science in the great depression. Isis 76, 301–318.
Kay, L.E., 1993. The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology. Oxford University Press, New York.
Khurana, R., 2007. From Higher Aims to Hired Hands: The Social Transformation of American Business Schools and the Unfulfilled Promise of Management as a Profession. Princeton University Press, Princeton.
Kohler, R.E., 1976. The management of science: the experience of Warren Weaver and the Rockefeller Foundation programme in molecular biology. Minerva 14, 279–306.
Kohler, R.E., 1987. Science, foundations, and American universities in the 1920s. Osiris 3, 135–164.
Kohler, R.E., 1991. Partners in Science: Foundations and Natural Scientists, 1900–1945. University of Chicago Press, Chicago.
Leslie, S.W., 1993. The Cold War and American Science: The Military-Industrial-Academic Complex at MIT and Stanford. Columbia University Press, New York.
Lowen, R.S., 1997. Creating the Cold War University: The Transformation of Stanford. University of California Press, Berkeley.
McCraw, T.K., 2010. Prophet of Innovation: Joseph Schumpeter and Creative Destruction. Harvard University Press, Cambridge, MA.
Metlay, G., 2006. Reconsidering renormalization: stability and change in 20th century views on university patents. Social Studies of Science 36 (4), 565–598.
Mirowski, P., 2011. Science-Mart: Privatizing American Science. Harvard University Press, Cambridge, MA.
Noble, D., 1977. America by Design: Science, Technology, and the Rise of Corporate Capitalism. Knopf, New York.
Noble, D., 1984. Forces of Production: A Social History of Industrial Automation. Knopf, New York.
NSB, 2012. Science and Engineering Indicators 2012. National Science Board, Washington D.C. http://www.nsf.gov/statistics/seind12/start.htm (accessed 10.07.13.).
Owens, L., 1985. Pure and sound government: laboratories, playing fields, and gymnasia in the 19th-century search for order. Isis 76, 182–194.
Owens, L., 1990. MIT and the federal "angel": academic R&D and federal-private cooperation before WWII. Isis 81, 188–213.
Redmond, K.C., Smith, T.M., 2000. From Whirlwind to MITRE: The R&D Story of the SAGE Air Defense Computer. MIT Press, Cambridge, MA.
Reich, L.S., 1985. The Making of American Industrial Research: Science and Business at GE and Bell, 1876–1926. Cambridge University Press, New York.
Reingold, N., 1979. National science policy in a private foundation: the Carnegie Institution of Washington. In: Oleson, A., Voss, J. (Eds.), The Organization of Knowledge in Modern America, 1860-1920. Johns Hopkins University Press, Baltimore, pp. 313–341.
Rosenberg, C.E., 1997 [1976]. No Other Gods: On Science and American Social Thought, Revised and Expanded Edition. Johns Hopkins University Press, Baltimore.
Schlossman, S., Gleeson, R.E., et al., 1994. The Beginnings of Graduate Management Education in the United States. Graduate Management Admission Council, Santa Monica, CA.
Simpson, C. (Ed.), 1998. Universities and Empire: Money and Politics in the Social Sciences during the Cold War. The New Press, New York.
Veysey, L.R., 1965. The Emergence of the American University. University of Chicago Press, Chicago.
Wise, G., 1985. Willis R. Whitney, General Electric, and the Origins of U.S. Industrial Research. Columbia University Press, New York.

Unobtrusive Methods, History of

Raymond M Lee, Royal Holloway University of London, Egham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In a classic book, *Unobtrusive Measures*, first published in 1966, Eugene Webb, Donald Campbell, Richard Schwartz, and Lee Sechrest made a case for the greater use of data sources *not* dependent on the elicitation of data directly from respondents or informants. They further argued for the importance of combining data obtained using different methods in ways that allowed the methodological weaknesses inherent in particular methods to compensate for one another. The various sources of potential data Webb et al. identified and the justifications they advanced for the use of unobtrusive measures are explored, as is the wider context that surrounded the book's original production. Problems associated with unobtrusive measures, including the ethical issues involved and the difficulty of generating measures relevant to particular research problems are discussed. The relationship of the unobtrusive measures tradition to the wider 'cultural turn' in the social sciences and to new technologies for data capture and collection is outlined.

Introduction

The recognition by social scientists at the beginning of the twentieth century that it was both feasible and desirable to elicit information directly from people about their circumstances, behavior, and subjective states had profound consequences. It opened the way for widespread adoption of survey methods, a variety of interviewing techniques, and the use of questionnaires and self-completion instruments. The difficulty is, of course, that obtaining information from respondents or informants is not unaffected by the interaction between researcher and researched. The act of asking, in other words, shapes the character of the answers given. In a book displaying an irreverent skepticism toward the methodological practices dominant at the time, Eugene Webb, Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest proposed that the "interrogation of experience, attitude and belief" (Kellehear, 1993, 1) might better be accomplished through the use of what they termed 'unobtrusive measures' (Webb et al., 1966); data gathered without a direct interchange between the researcher and the researched. In their view, the great benefit of unobtrusive measures lies in their 'nonreactive' character, thus avoiding threats to data quality caused by the researcher's presence. Accordingly, Webb et al. suggest that social researchers should pay more attention to sources of data such as the physical traces associated with people's passage through their physical environment, observing situations without participating in them, and the use of documentary sources of various kinds. Instead of using interviews and questionnaires, researchers might just as easily answer research questions by watching what people do, by looking at physical evidence of various kinds, or by turning to the words or images inscribed in various media.

Webb et al.'s (1966) book *Unobtrusive Measures* encouraged researchers to seek out novel and innovative ways of using unobtrusive data in research rather than falling back on what had become by the time they were writing standard and predictable ways of collecting data. Arguing that researchers should seek to be imaginative and creative in their thinking about data relevant to particular research problems, *Unobtrusive Measures* offered its readers such intentionally off-beat examples as the use of wear on the floor tiles surrounding a museum exhibit to measure visitor flows; the changing size of suits of armor as a way of measuring changes in human stature over time; and the relationship in the 1960s, at a time when long hair tended to signal political radicalism, between psychologists' hair length and their methodological preferences. However, Webb et al. went beyond simply advocating that researchers should add unobtrusive measures to the methodological repertoire. Their principal objection to traditional data sources is "that they are used alone" (1966, 1). In other words, data collection methods used singly are inferior to the use of multiple methods. Webb et al.'s argument is that any research finding is potentially vulnerable to an assertion that the results found are the product of fallibilities inherent in the data collection process. These fallibilities, however, are not the same across methods. In particular, problems of reactivity do not affect data collected unobtrusively. As a result, configuring different methods, each of which is fallible in a different way, is preferable to overreliance on a single method.

Origins and Contexts

In 1962, the US Office of Education offered Donald Campbell a contract for work on research methods. As he records, part of the final report on the contract was entitled: "Other Measures: A Survey of Unconventional and Cooperation-Free Measures for Social Science" (Campbell, 1988, 19). This report was to provide the basis for Webb et al.'s later book. In its breadth and its emphasis on imagination, creativity, and the use of unusual methods, *Unobtrusive Measures* reflected the circumstances of its production. Individually and collectively, the authors of *Unobtrusive Measures*, all faculty members at Northwestern University, appear to have been both genial and congenial, but their collective production of the book reflected more than conviviality. It was the product of their collective experience and expertise, the breadth of their interests, and the intellectual milieu that surrounded them. The authors of *Unobtrusive Measures*, three psychologists and a sociologist, were scarcely narrowly focused as social scientists. This is particularly so in Campbell's case. Even in his undergraduate days, Campbell

seems to have made a point of being open to a wide and diverse range of intellectual opportunities. His career is notable for the breadth of its engagement with many different scholarly strands within and beyond psychology, intellectual changes of direction, and a penchant for collaboration that included a variety of productive associations with anthropologists and sociologists, including Louis Wirth at Chicago (Campbell, 1988, 10; Shadish and Luellen, 2004). A psychologist by training, Webb's professional background was in journalism (see Webb and Salancik, 1966). Schwartz, the sociologist in the group, worked at the interface between law and sociology (Garth and Sterling, 1998). Both Webb and Schwartz, in other words, worked in fields that relied heavily but skeptically on interviewing, and in which the cross-checking and weighing of sources against one another is important to the assessment of evidence. Since early in his career, Sechrest has also worked across a range of substantive and methodological domains, making contributions to a number of different fields and subspecialisms within psychology (Bootzin and McKnight, 2006).

The book itself grew out of a weekly 'Sack Lunch' at which it became a game to come up with novel methods (Campbell, 1981, 481). (For non-US readers, a 'sack lunch' or 'brown bag lunch' is an informal lunchtime meeting to which one brings one's sandwiches in the eponymous paper sack or bag.) The studied informality of this setting was apparently no accident. Campbell had been at the University of Chicago before coming to Northwestern University. He had found his time at Chicago less than satisfying (Campbell, 1988). Tenure pressures there meant that levels of competition between junior faculty operated in ways, which inhibited the free flow of ideas. In consequence, Shadish and Luellen (2004, 85–86) suggest, Campbell tried to create an intellectual climate at Northwestern University that encouraged open discussion and intellectual diversity, and in which the outlandishness of an idea did not constitute a barrier to its consideration.

One wider, and usually unremarked, context for Webb et al.'s work was the Cold War. The funding of Campbell's original methodological research was made available under the provisions of the National Defense Education Act, which had been enacted as a response to criticisms of the American education system that followed the launch by the Soviet Union of the first artificial satellite, *Sputnik*, in 1957. Some of Campbell's earlier work and further work on the development of nonreactive measures by Sechrest and Phillips (1979) were funded by the Office of Naval Research (ONR). As Solovey (2012) points out, such funding did not necessarily imply that the research had direct utility for the military; ONR, in particular, had something of a reputation for having a hands-off relationship with the researchers it funded. One of the four authors of *Unobtrusive Measures*, Eugene Webb did, however, have a close relationship to the US defense establishment.

From 1967 to 1968, Webb was on leave serving as a staff member at the Institute for Defense Analyses (IDA) (Hastorf et al., 1995), a body mostly funded by contracts from the US Department of Defense that by the mid-1960s had developed an interest in counterinsurgency research (Wakin, 1992). It is not clear how Webb came to take the position with IDA, or how, if at all, his involvement with the organization was related to his authorship of *Unobtrusive Measures*, and the attention the book had attracted on publication. However, one insider account (Deitchman, 1976) records that unobtrusive measures began to be of interest as a component of counterinsurgency studies in Thailand in response to perceived problems with interview-based data collection. Webb went on to undertake a methodological review of social science research in Vietnam critical of how existing studies had been conducted (Rohde, 2011), and to sit on a number of boards and committees looking at utilization of social science research by the US Department of Defense (Klare, 1972; National Research Council Advisory Committee, 1971).

The other authors of *Unobtrusive Measures* do not seem to have taken public positions for or against the Vietnam War. The use during the war of body counts as a supposed measure of combat effectiveness did, though, give rise to what has become known as Campbell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979). (The same idea is also expressed in what is popularly known as Goodhart's law, "When a measure becomes a target, it ceases to be a good measure.")

Sources of Unobtrusive Measures

Unobtrusive Measures is organized around a loose typology of data sources: traces, archives, and observation. Traces are the physical residues that social activities leave on the physical environment. Webb et al. distinguish between 'erosion measures' produced through the wearing away of the environment, and 'accretion measures' that reflect the addition of material to the environment typically through abandonment or adornment. Wear on library books as measured by smudges, finger marks, turned-down pages, and the like is a classic erosion measure of a book's use by readers. Accretion measures, Webb et al. suggest, include such things as the amount of litter or graffiti to be found in some setting. 'Garbology,' the application of archaeological techniques to the material discarded in contemporary societies, provides accretion measures important to the understanding of consumption patterns (see, e.g., Rathje and Murphy, 2001). Archives are written records of various kinds. Webb et al. make a somewhat arbitrary distinction between what they call 'running records' and 'episodic records.' Running records are the "ongoing continuing records of a society" (1966, 53), vital registration data, for example, or recurring items in newspapers. Episodic records have a more discontinuous form. In consequence, they do not permit longitudinal research in ways that are possible with running records. When they refer to 'observation' what Webb et al. have in mind are situations in which activity in some setting is recorded without direct interaction between researcher and researched, and, in many cases, without the latter being aware that they are being studied. Observations can be conducted directly by researchers but they can also involve the use of data capture technologies to record activities or interactions as they occur.

Webb et al. pay careful attention to the problematic aspects of using such sources. Thus, traces are ubiquitous, easily

quantifiable, and available at low cost but it can be difficult to make inferences from them. Over time wear on a surface becomes an unreliable measure of use, for example, because the worn area eventually wears faster than it did before. In addition, baseline information that would enable a rate to be calculated from trace data for some measure is often unavailable. Running records reflect the organizational processes by which they are produced (Garfinkel, 1967). The extent to which documents have differentially been retained or survived affects the availability of episodic records, while personal and autobiographical documents might not be free of pressures toward positive self-presentation. Observation can also be problematic. What one would like to observe is not always available to be observed. Nor is the meaning of some observed behavior necessarily obvious without access to other more reactive sources of data. Careful attention needs to be paid to sampling and data recording issues, and to the role and possible impact of the observer on the setting. Serious ethical issues arise in situations where people do not realize they are being observed.

None of the data types described by Webb et al. can be said to be unknown to social scientists. If outside archaeology the analysis of trace data remained relatively uncommon in the mainstream social sciences, the examples of Freud and, from fiction, Sherlock Holmes provide a popular cultural mandate for linking environmental traces to behavioral activity (Ginzburg, 1979). Nonparticipant observation has roots in a tradition going back to the late nineteenth century in which members of the middle class, for reasons of curiosity, adventure, investigation, or reform, went 'undercover' to gain access to and understanding of the lives of the poor or the working class (Pittenger, 2012). In the 1920s and 1930s, the perceived failings of research based on self-report instruments encouraged direct and systematic observation of children at play and in classrooms (Cairns and Cairns, 2006), much of it conducted by female researchers such as the Austrian psychologist Charlotte Bühler, the sociologist Dorothy Swaine Thomas, Mildred Parten, Ruth Addington, and others. Eclipsed, on the one hand, by experimental methods and, on the other, by the rise of self-report methods, the use of documents, personal and otherwise, formed an important strand within a number of social science disciplines in the first half of the twentieth century (Allport, 1942; Gottschalk et al., 1945). Webb et al.'s achievement was to take a residual category, the 'other methods' of Campbell's original study, and animate it with a plausible conceptual frame and a detailed methodological rationale, which they then populated with compelling instantiations.

Triangulation

Although the term does not occur very frequently in *Unobtrusive Measures*, the concept of 'triangulation' is central to Webb et al.'s work. As they put it, "The most persuasive evidence comes through a triangulation of measurement processes. If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it" (1966, 3). This stance derived largely from Donald Campbell's work. Early on in his career, Campbell had become concerned with difficulties involved in measuring attitudes and establishing the validity of data derived from laboratory experiments. Working with Donald Fiske, he eventually proposed an approach, dubbed 'multiple operationalism,' based around the intertwined notions of 'convergent and discriminant validation' (Campbell and Fiske, 1959). Where different measures for the same trait, construct, or concept converged in the sense of being highly correlated, reliability could be assumed. Where correlations between measures of different traits, constructs, or concepts were low, thus discriminating between them, validity could be assumed.

The term 'triangulation' has its origins in geodetic measurement (Blaikie, 1991; Hewitt, 2010). (It might be relevant here that Donald Campbell had been a naval officer during the war and can be presumed to have been thoroughly familiar with navigational uses of the term.) Campbell and Fiske's work on the multitrait–multimethod matrix had taken a relatively narrow approach that depended on the results of different methods being available in the form of test results, rating scores, and the like. Assessing the effect of 'a series of imperfect measures' therefore involved examining simultaneously and systematically the interrelationships within a (triangular) array of correlation coefficients. This approach has remained strong in psychology. Fifty years on, Campbell and Fiske's work remains widely cited, and a sizable literature now exists devoted to the development, refinement, and extension of their original work (Eid and Nussbeck, 2009).

Popularized by the sociologist Norman Denzin (1970), among others, the notion of triangulation soon took on a wider aspect. In particular, it began to refer in a more general way to the combination of quantitative and qualitative methods. A substantial debate ensued about whether it was possible, desirable, or feasible to combine such methods (for a review, see Bryman, 2006). One position, strongly held, was that particular methods are embedded within wider epistemological and ontological frameworks, which preclude the very possibility that they might be combined. In another view, perhaps more prevalent among those in applied fields and in evaluation research, the case for combining methods was a largely pragmatic one in which epistemological issues, while relevant, were far from determining. In all of this, the notion of 'triangulation' took on a variety of meanings, sometimes referring to the loose configuration of different methods, sometimes to their concomitant use, and sometime to their deployment in an integrated way (Moran-Ellis et al., 2006).

One irony of this debate was that it tended to lose sight of Webb et al.'s specific focus on reactivity. While they are likely to have thought that using survey interviewing and depth interviewing within the same study was better than using either one on its own, Webb et al. would have seen both methods as reactive, even if that reactivity manifested itself in different ways. While the notion of triangulation had opened up the debate about the use of multiple methods, it also paradoxically reproduced a situation of the kind Webb et al. had originally criticized: a reliance on a set of conventional methods, the specific methodological weaknesses of which remained relatively unexamined. Lost, too, was their insistence of being open to the using unexpected or off-beat data sources. From a different perspective, however, what these developments underlined was one of the major criticisms of unobtrusive

measures: the absence of a distinct methodology for generating specific measures relevant to a particular research problem.

Impact, Use, and the Generative Problem

Unobtrusive Measures can with some justification be called a methodological classic. By 1979 the book had sold in excess of 125 000 copies and had gone into its fourteenth printing (Sechrest and Phillips, 1979). It was subsequently published in a revised and expanded edition (Webb et al., 1981) with a new title, *Nonreactive Measures in the Social Sciences*. The original edition was republished in 2000. A number of textbooks aimed, not uncritically, at encouraging students to understand the value of unobtrusive approaches have also appeared (see, e.g., Kellehear, 1993; Lee, 2000; Emmison et al., 2012).

Even quite early on, however, proponents of the approach became concerned about the apparent disparity between the popularity of the book and the extent to which unobtrusive measures were actually used. Webb and Weick in an article on the usefulness of unobtrusive measures in the study of organizations – somewhat pointedly subtitled 'A Reminder' – speak of the 'ceremonial citation' of Webb et al.'s book (1979, 650). Describing the slow take up of unobtrusive measures as "somewhat disappointing and a bit puzzling," Sechrest lamented that while "the idea or concept of unobtrusive measures has clearly caught on, the practice of unobtrusive measurement has not" (1975, 103). Sechrest pointed to a number of reasons for the relatively slow adoption of the innovations proposed in *Unobtrusive Measures*. One was the book itself which he candidly admitted "was little more than a catalog or compendium of disparate techniques of varying degrees of worth" (1975, 103). In addition, the authors of *Unobtrusive Measures* had not attempted to develop further the methodological aspects of the approach in a systematic way. As a result, there was little information about the reliability and validity of possible measures. Perhaps the fundamental problem, though, was the absence of explicit guidance for those seeking to use unobtrusive methods in relation to a particular research problem. Beyond commending particular authors for being 'imaginative' and briefly and jokily invoking Sherlock Holmes as a role model (1966, 35), *Unobtrusive Measures* said little about the topic and provided little explicit guidance to anyone seeking to apply the lessons of the book to their own research.

Eventually, two broad strategies were to emerge to deal with the issue of generating measures relevant to a particular research problem (Lee, 2000). Webb and Weick (1979) detailed a list of strategies by which researchers could orient themselves to problems, methods, or data sources. These include seeking data in unlikely places, using multiple sources of data in multiple ways, being playful and creative, and being alert to patterns and anomalies in particular measures. Sechrest and Phillips proposed a second approach, a generative taxonomy for the identification of unobtrusive measures (1979; see also Webb et al., 1981: Chapter 9). The taxonomy took the form of a matrix with a list of purposes to which measures might be put on one dimension and the characteristics of responses on the other. For example, one might want to assess the degree of interest or involvement in some activity by reference to the frequency with which that interest manifests itself in the activity, or assess affective states through examining how far the choices people make indicate positive or negative evaluations of what has been chosen.

It can be argued that neither of these approaches is entirely satisfactory. The 'orientational' approach adopted by Webb and Weick offers some guidance to researchers but still seems to be too general and inexact to be directly useful. The generative taxonomy proposed by Sechrest and Phillips has the benefit of systematicity. However, it can be seen as being unwieldy (as Sechrest and Phillips themselves admit) and too narrowly focused on particular domains to have much independent generative power. Subsequently, Lee (2000) suggested abandoning both the orientational and taxonomic approaches. He argued instead for developing a variety of heuristics for generating unobtrusive measures. These heuristics were to be based on relevant modes of data acquisition – 'finding,' 'capturing,' and 'retrieving' – understood in an active way rather than on Webb et al.'s passive typology of potential data sources. However, the utility of this approach has yet to be demonstrated.

Ethics

The original edition of *Unobtrusive Measures* makes practically no mention of ethical issues. This was an intentional omission. In the preface to the book, Webb et al. indicated that they disagreed among themselves about the issues involved (1966, v–vi). They also felt that the topic was important enough and complex enough to warrant a separate, extended treatment. A number of reviewers homed in on the book's failure to address itself to ethical concerns. Richard Hill reviewing the book in the *American Sociological Review* notes Webb et al.'s disclaimer but judged it 'not enough' given the increased attention to ethics that was beginning to be seen in the social sciences at the time (1966, 565). In the *Annals of the American Academy of Political and Social Science*, Roy G. Francis complained that, "Except for a disclaimer in the Preface, there is not one damned word about invading privacy, not one hint about the civil rights problems, and no real concern about the fact that social scientists are bent upon studying human beings" (1966, 230).

Webb et al.'s insouciant stance toward ethical issues seems particularly egregious because they devoted an entire chapter of their book to 'contrived observation' and the use of hidden hardware for tracking and monitoring behavior. Contrived observation refers to situations where behavior is observed under induced conditions, such as when a confederate creates a set of circumstances designed to see how those in some particular setting react. Used to study topics such the willingness of bystanders to intervene in emergencies or usually hidden discriminatory practices, such methods rely directly on deception. Standing at the beginning of an era of transistorization, miniaturization, and later digitization that widely enhanced the usefulness and usability of recording devices (Lee, 2004), Webb et al. were clearly intrigued by the possibilities then emerging for logging data from natural settings using hardware of various kinds. The advantage in doing so is that it becomes possible to minimize observer effects produced by inattention, misperception, or fatigue. Deploying hidden cameras and microphones can come to seem, however, uncomfortably like the use of surveillance techniques.

It is perhaps not surprising, then, that the second, revised (1981) edition of the book contained a separate chapter devoted entirely to the ethical implications of using unobtrusive measures. Perhaps raising more questions than they answered, Webb et al. (1981) addressed issues relating to privacy and informed consent. They note that the gathering of trace data and the use of observation often occur in relatively public contexts where expectations of privacy might be attenuated, and point, too, to the practical difficulties of obtaining informed consent in such settings. They also emphasize the responsibility placed on researchers to ensure that research procedures meet ethical standards.

Webb et al. point to the wider role that had been adopted by the U.S. Federal government in relation to the ethical regulation of scientific research, noting that at the time the extension of such regulation to the social sciences was still controversial. Were they writing today it is likely that Webb et al. would note that while the issues they raise have scarcely been resolved in any definitive way, many countries now implement systems of prior ethical review of social science research. The ethical dilemmas surrounding use of unobtrusive measures are no longer, therefore, left to the discretion of individual researchers. Surveying the situation in Canada, especially in relation to unobtrusive studies of discriminatory behavior, Page (2000) has argued that research is increasingly constrained by what he sees as an overrigid implementation of informed consent procedures.

Unobtrusive Measures and the Cultural Turn

Webb et al.'s work predated the 'cultural turn' in the social sciences, the shift beginning in the 1970s that emphasized the understanding and interpretation of culture as a central component of social inquiry. Crucial to the cultural turn was a rejection of measurement, whether unobtrusive or not, as a focal preoccupation of the social sciences. Drawing on a wide range of theoretical and methodological approaches, including postmodernism, semiotics, phenomenology, discourse analysis, and the like, researchers in many disciplines turned instead to interpretive or critical understandings of the social world. Work of this kind would seem to have little in common with traditional conceptions of unobtrusive measures. What these approaches share with Webb et al.'s work, however, is "a rather sideways glance at traditional sources of data in social research" (Lee, 2000, 7). Treating the world as irredeemably cultural opens to study, aspects of social life that in other contexts might be regarded as peripheral, insignificant or taken-for-granted. Much work in cultural studies has been shaped by approaches from literary studies and the humanities that traditionally made little use of elicited data but utilized instead sources historically deprecated or marginalized within mainstream social science. The importance Webb et al. placed on a playful, irreverent stance toward potential sources of data has a degree of resonance with certain kinds of postmodern sensibility.

One starting point for much writing in cultural studies is that the shift toward the use of interviewing and the growth in quantitative techniques that began in the social sciences in the 1920s fostered a sensibility toward data in which the spoken word of the interview was elevated over the textual and the visual. In consequence, the use of personal documents, such as letters, diaries, memoirs, and so on declined (Plummer, 2001), as did the use of photography as a mainstream research method (Stasz, 1979). In recent years, increasing interest in discursive and textual practices, the interpretation of lived experience, and a commitment to overcome the traditional invisibility of disadvantaged, exploited, and colonized groups have all encouraged the revalorization of such sources. Indeed, the new-found popularity of documentary and visual methods and work focused on material culture has produced what is now a very disparate range of work covering a wide variety of substantive topics and methodological approaches.

In the field of visual research, a number of writers have developed integrated frameworks in an attempt to bring some unity to the diversity of approaches (see, e.g., Pauwels, 2010; Emmison, 2010). One benefit of these frameworks is that they allow particular perspectives, including the unobtrusive measurement approach, to be located relative to one another in terms of their assumptions, procedures, products, and the like. Pauwels, for example, has developed a graphical representation that provides a synoptic view of the various pathways potentially linking visual materials of various kinds with relevant research design issues and with the formats and purposes associated with outputs from the research. Within this framework he includes the use of found materials and observational methods of various kinds. The Australian researcher Michael Emmison has developed a framework that begins from the assumption that researchers have tended to conceive of the visual in a relatively narrow way that focuses primarily on objects such as photographs, advertisements, or cartoons. Instead, he suggests a much broader understanding of the visual as that which is "seen and observable" (2010, 236). Emmison in a sense recasts Webb et al.'s preoccupation with the unobtrusive and the ubiquitous by emphasizing how in social life it is often that which is pervasive in the sense of being before our eyes which is routinely invisible. Acknowledging "obvious methodological affinities with an older and these days somewhat neglected and unfashionable – branch of social research, the use of unobtrusive or non-reactive measures" (2010, 243), Emmison goes on to explore ways of understanding what can be seen, whether this takes a two-dimensional form, as in, for example, visual images, a three-dimensional form, as with material objects, or is available to be observed.

Prior (2008) has demarcated a variety of approaches to the study of documents. He points to emerging strands of work concerned with the use and function of documents within particular social settings as opposed to the content of the documents themselves. For the most part, however, Prior contends, documentary research has been preoccupied with content; the dominant approach has been to treat documents as sources of evidence. Even when researchers have obeyed the ethnomethodological injunction to treat documents as 'topic' rather than 'resource' (Zimmerman and Pollner, 1970) they have done so, according to Prior, by examining how content is socially produced. Webb et al.'s discussion of documentary sources stands well within the dominant tradition of documentary research, with its focus firmly on using content as a resource and on traditional concerns with the authenticity, credibility, and representativeness of documents themselves

(Scott, 1990). Having said this, their discussion of the potential for reactivity in documents depends on an implicit, if astute, understanding of how documents are socially produced, and the extent to which the use and function of documents within a setting might index some of its relevant features. (See, for example, Webb and Weick's (1979) discussion of the relationship between organizational constraint and the thickness of procedure manuals.)

Unobtrusive Measures and Technological Change

One area of considerable growth in the period since the emergence of *Unobtrusive Measures* has been telemetrics, the use of wireless information systems for the remote capture and transmission of data relating to physical, spatial, cognitive, and even emotional states. The increasing availability of portable and handheld devices and the convergence of various digital technologies has had a transformative effect on many areas of research concerned with the ongoing flow of activity and behavior (Goodwin et al., 2008).

At the time they were writing Webb et al. could hardly have envisaged the Internet. Its impact on the field of unobtrusive measures has been substantial and significant. Large repositories of records are now readily available online. Social media sites, blogs, wikis, the ratings, rankings, and comments that appear on many Web sites, and online communities devoted to particular topics, all offer boundless research opportunities. It is now often relatively straightforward to combine unobtrusive measures with other sources of online data. For example, one might relate the response to a question on an online survey to the time taken for the question to be answered. One can also note that an ironic consequence of the emergence of online methods is that in the future are they likely to make social scientists less reliant on self-report methods (see, e.g., Savage and Burrows, 2007).

Conclusion

Although their work has its iconoclastic elements, *Unobtrusive Measures* is a gentle book. Webb et al. prefer to make their case through humor, wit, and irreverence rather than through the kind of stridently programmatic writing that was common in many methodological texts in the 1960s and 1970s. Behind its 'oddball' or 'off-beat' facade the book is written in a quietly cajoling style, continually drawing the reader back to the deficiencies of research based on single measures or reactive methods. If Webb et al.'s original rather relaxed take on ethical issues is no longer acceptable, the message of *Unobtrusive Measures* is that everything we do, everything we see, and everything we read is grist to the analytic mill, and that cultivating our methodological sensibilities to recognize the implications of that is no bad thing. Webb et al.'s work also resonates in a period when technological change and the methodological transformations associated with it are increasingly challenging the dominance of self-report methods. The methodological repertoire of the mid-twenty-first century will be far from that challenged by Webb et al. in the mid-twentieth, but its unlikely that their message will have been completely obliterated.

See also: Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures.

Bibliography

Allport, G.W., 1942. The Use of Personal Documents in Psychological Science. Social Science Research Council, New York.
Blaikie, N.W.H., 1991. A critique of the use of triangulation in social research. Quality & Quantity 25, 115–136.
Bootzin, R.R., McKnight, P.E., 2006. Introduction. In: Bootzin, R.R., McKnight, P.E. (Eds.), Strengthening Research Methodology: Psychological Measurement and Evaluation. American Psychological Association, Washington, DC, pp. 3–8.
Bryman, A., 2006. Editor's introduction: mixed methods research. In: Bryman, A. (Ed.), Mixed Methods, vol. I. Sage Publications Ltd, London, pp. xxv–xxxix.
Cairns, R.B., Cairns, B.D., 2006. The making of developmental psychology. In: Damon, W., Lerner, R.M. (Eds.), Handbook of Child Psychology, Theoretical Models of Human Development, vol. 1. John Wiley & Sons, New York, pp. 89–165.
Campbell, D.T., 1979. Assessing the impact of planned social change. Evaluation and Program Planning 2, 67–90.
Campbell, D.T., 1981. Comment: another perspective on a scholarly career. In: Brewer, M.B., Collins, B.E. (Eds.), Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell. Jossey-Bass, San Francisco, pp. 454–486.
Campbell, D.T., 1988. Perspective on a scholarly career. In: Campbell, D.T., Overman, E.S. (Eds.), Methodology and Epistemology for Social Sciences: Selected Papers. Chicago University Press, Chicago, IL, pp. 1–26.
Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81–105.
Deitchman, S.J., 1976. The Best-laid Schemes: A Tale of Social Research and Bureaucracy. MIT Press, Cambridge, MA.
Denzin, N.K., 1970. The Research Act in Sociology: A Theoretical Introduction to Sociological Methods. Butterworth Ltd, London.
Eid, M., Nussbeck, F.W., 2009. The multitrait-multimethod matrix at 50! Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 5, 71.
Emmison, M.J., 2010. Conceptualizing visual data. In: Silverman, D. (Ed.), Qualitative Research. Sage Publications Ltd, London, pp. 233–249.
Emmison, M.J., Smith, P.D., Mayall, M., 2012. Researching the Visual: Images, Objects, Contexts and Interactions in Social and Cultural Inquiry, second ed. Sage Publications Ltd, London.
Francis, R.G., 1966. Review of: unobtrusive measures: nonreactive research in the social sciences. The Annals of the American Academy of Political and Social Science 368, 229–230.
Garfinkel, H., 1967. Studies in Ethnomethodology. Prentice-Hall, Englewood Cliffs, NJ.
Garth, B., Sterling, J., 1998. From legal realism to law and society: reshaping law for the last stages of the social activist state. Law and Society Review 32, 409–472.
Ginzburg, C., 1979. Clues. Theory and Society 7, 273–288.
Goodwin, M.S., Velicer, W.F., Intille, S.S., 2008. Telemetric monitoring in the behavior sciences. Behavior Research Methods 40, 328–341.
Gottschalk, L.R., Kluckhohn, C., Angell, R.C., 1945. The Use of Personal Documents in History, Anthropology, and Sociology. Social Science Research Council, New York.
Hastorf, A.H., Leavitt, H.J., Mann, J.K., Holloway, C.A., 1995. Memorial Resolution: Eugene J. Webb, 1933–1995 [WWW Document]. http://facultysenate.stanford.edu/archive/1995_1996/reports/105164.html.
Hewitt, R., 2010. Map of a Nation: A Biography of the Ordnance Survey. Granta Books, London.
Hill, R.J., 1966. Review of: unobtrusive measures: nonreactive research in the social sciences. American Sociological Review 31, 565.
Kellehear, A., 1993. The Unobtrusive Researcher: A Guide to Methods. Allen Unwin Pty Ltd, St. Leonards, NSW.
Klare, M.T., 1972. War without End: American Planning for the Next Vietnams. Alfred A. Knopf, New York.
Lee, R.M., 2000. Unobtrusive Methods in Social Research, Understanding Social Research. Open University Press, Buckingham.
Lee, R.M., 2004. Recording technologies and the interview in sociology, 1920–2000. Sociology: Journal of the British Sociological Association 38, 869–889.

Moran-Ellis, J., Alexander, V.D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., Thomas, H., 2006. Triangulation and integration: processes, claims and implications. Qualitative Research 6, 45–59.

National Research Council Advisory Committee on the Management of Behavioral Science Research in the Department of Defense, 1971. Behavioral and Social Science Research in the Department of Defense: A Framework for Management. National Academy of Sciences, Washington, DC.

Page, S., 2000. Community research: the lost art of unobtrusive methods. Journal of Applied Social Psychology 30, 2126–2136.

Pauwels, L., 2010. Visual sociology reframed: an analytical synthesis and discussion of visual methods in social and cultural research. Sociological Methods and Research 38, 545–581.

Pittenger, M., 2012. Class Unknown: Undercover Investigations of American Work and Poverty from the Progressive Era to the Present. NYU Press, New York.

Plummer, K., 2001. Documents of Life 2: An Invitation to a Critical Humanism. Sage Publications Ltd, London.

Prior, L., 2008. Repositioning documents in social research. Sociology: Journal of the British Sociological Association 42, 821–836.

Rathje, W.L., Murphy, C., 2001. Rubbish!: The Archaeology of Garbage. University of Arizona Press, Tuscon, AZ.

Rohde, J., 2011. The last stand of the psychocultural cold warriors: military contract research in Vietnam. Journal of the History of the Behavioral Sciences 47, 232–250.

Savage, M., Burrows, R., 2007. The coming crisis of empirical sociology. Sociology: Journal of the British Sociological Association 41, 885–899.

Scott, J., 1990. A Matter of Record: Documentary Sources in Social Research. Polity Press, Cambridge.

Sechrest, L., Phillips, M., 1979. Unobtrusive measures: an overview. In: Sechrest, L. (Ed.), Unobtrusive Measurement Today. Jossey-Bass, San Francisco, pp. 1–18.

Shadish, W.R., Luellen, J.K., 2004. Donald Campbell: the accidental evaluator. In: Alkin, M.C. (Ed.), Evaluation Roots: A Wider Perspective of Theorists: Views and Influences. Sage Publications Ltd, London, pp. 80–87.

Solovey, M., 2012. Cold war social science: spectre, reality, or useful concept? In: Solovey, M., Cravens, H. (Eds.), Cold War Social Science: Knowledge Production, Liberal Democracy, and Human Nature. Palgrave Macmillan, Basingstoke, pp. 1–22.

Stasz, C., 1979. The early history of visual sociology. In: Wagner, J. (Ed.), Images of Information: Still Photography in the Social Sciences. Sage, Beverly Hills, CA, pp. 119–136.

Wakin, E., 1992. Anthropology Goes to War: Professional Ethics and Counterinsurgency in Thailand. University of Wisconsin Center for Southeast Asian Studies, Madison, WI.

Webb, E., Weick, K.E., 1979. Unobtrusive measures in organizational theory: a reminder. Administrative Science Quarterly 24, 650–659.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., 1966. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago, IL.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., Grove, J.B., 1981. Nonreactive Measures in the Social Sciences, second ed. Houghton Mifflin, Boston.

Webb, E.J., Salancik, J.R., 1966. The interview or the only wheel in town. Journalism Monographs 2, 1–49.

Zimmerman, D.H., Pollner, M., 1970. The everyday world as a phenomenon. In: Pepinsky, H.B. (Ed.), People and Information. Praeger, New York, pp. 33–65.

Urban Anthropology

Peter JM Nas, Leiden University, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S.M. Low, volume 24, pp. 15999–16004, © 2001, Elsevier Ltd.

Abstract

Urban anthropology consolidated in the 1970s–80s and became an important section of anthropology and urban studies. It embraces four main fields: (a) urban ethnography to highlight urban conditions and lifestyles, (b) the comparative study of cities as wholes, (c) anthropological studies of urbanization as a process of regional, national, and global social change related to all sorts of problems, especially poverty and slums, urban inequality and ethnicity, and urban environmental and conservation issues, and (d) fieldwork under specific urban conditions.

Introduction

Urban anthropology is a section of general anthropology and urban studies. It embraces urban ethnography and is diversified by regional differences. In short, the subject is quite complex, as it examines various sources of influence and stimuli. The formational phase of urban anthropology took place in the 1960s, a reflection of the contemporary rapid world urbanization especially in Latin America and Asia that occurred after the Second World War. Europe and the United States already had reached high levels of urbanization.

Urban studies demonstrate that the city is one of the most important inventions of mankind. The city is basically an unfolding societal process of increasing interdependency and cooperation of high-density populations in their quest to realize increasingly complex activities by attaining high levels of heterogeneity and specialization (Nas, 1990: p. 35; Karsten in Toelichting, 1938). Cities have paved the way for a cumulative increase in living standards.

Urban anthropology is developing in response to the escalation of almost complete world urbanization, which will be reached in the second half of the twenty-first century when most probably 80% of the world population will be located in urban centers. At this stage, urban studies and urban anthropology will dominate almost the whole society and paradoxically this will mean their termination as a specific subfield in anthropology and other social sciences.

The Formational Stage

After the Second World War, anthropology was dominated by functionalism and structuralism. The attention paid to evolutionist and diffusionist theories had weakened. The study of family and kinship, religion, and hierarchical relationships, based on empirical research by means of fieldwork, was prominent. One obvious indication of this is the book *Tristes Tropiques* by Lévi-Strauss (1955), the founder of structuralism, who relates his travels to some Indian groups in the interior of Brazil. The aims of anthropology were listed by Evans-Pritchard (1963) as the description of important features of societies and cultures, the characterization of the underlying structures and the comparison of cultures all over the world, but outside the West. Participant observation during long fieldwork periods was considered the specific anthropological method, and the principal aim, known as the holistic approach, was to reach an understanding of society and culture as a whole. The well-known introduction to cultural anthropology by Keesing (1958) covers such issues as material culture, social, economic and political patterns, religion, art, and language.

An influential example of a functionalist study on the city in this period is Sjoberg's *The Preindustrial City* (1960). These cities, Sjoberg argues, are rather small with 5000 to 10 000 inhabitants, rarely more than 100 000. They maintain political, administrative, military, economic, religious, and educational functions generally mixed in the city area. Commercial, religious, and administrative activities are situated in the center and professional and ethnic characteristics are bound to specific neighborhoods. The lower class is concentrated in the urban peripheries and the differences between higher, lower, and outcast population groups are indicated by such expressions of cultural differences as language and clothing. Living, working, and religious practices are combined in the same ward. Sjoberg shows that family life was dominated by the extended family, whose functions were many, including procreation, socialization, and social security. Furthermore, these families also played an economic, political, and religious role. The marriages of the children were arranged so that class boundaries were maintained. The roles assigned in marriage were fairly formal, with the women, who were responsible for the upbringing of the children, staying out of the public sphere. The major religious functions that exerted a binding influence were performed by the elite. The political functions were also dominated by the elite who exercise traditional leadership based principally on personal relations and less on strict administrative rules. Sjoberg describes how economic life was dominated by the guilds, which among their other tasks, maintained monopolies and were responsible for the selection and training of members. They also regulated capital and ceremonies.

Although this study by Sjoberg highlights the similarities between preindustrial cities, it disguises the cultural differences between cities in diverse regions. By the 1960s, for example, the specific characteristics of Yoruba, Maya, and medieval cities in Europe had already been highlighted. Krapf-Askari (1969) described the layout of the Yoruba cities (present-day Nigeria) characterized by a palace, plaza, and holy wood

within the city walls. She shows how the different types of family systems, either patrilineal or bilineal, determined the layout of the traditional city and the social and ceremonial integration with the villages, respectively, in a cake wedge and chess board system, as no distinction between city and countryside was made by the inhabitants. This integration of town and countryside also applied to the traditional Maya cities that had a highly ceremonial character (Hardoy, 1973). In contrast, the predominating feature of the medieval city in Europe was a pronounced gap between town and countryside (Pirenne, 1952).

Besides the study of the city as a whole and the comparison of cities embedded in different cultures, the accelerating process of urbanization and its related problems attracted attention. This led to a heavy emphasis on research into two issues, poverty and slums, and to the related matter of rural–urban migration.

In the 1950s and 1960s, Lewis (1959, 1968) published a series of books about his research in the slums of Mexico, Puerto Rico, and New York. He developed the concept of the 'culture of poverty,' implying that a core group in slums, probably all over the world, is not just poor, but has culturally adapted to their poverty, transmitting this subculture from generation to generation. He argues that the specific traits of culture of poverty have been generated by long-term adaptation to a life both marginal and poor, generating distrust toward the institutions in society at large, dependency, and even specific personality traits. His family histories, on which the concept is based, and the descriptions of daily life in slums are an extensive and intricate example of urban ethnography. By this time, urban anthropology was no longer limited to societies outside the West as the Puerto Ricans in New York were one of the groups researched. The study of a black ghetto in Washington, DC, by Hannerz (1969) is another example of the change in focus toward the urban ethnography of poor groups in Western societies. In France the *bidonvilles* of Paris, including that of Noisy-le-Grand mainly populated by French families, were an early field of functionalist slum study focusing on deviant behavior and *exclusion sociale* (Labbens, 1965; Klanfer, 1965).

Migration was the second important subject at the end of the formational phase. By then, the rural exodus and the migration of people from all over the world to Western cities involved ever-growing numbers and could no longer be ignored. The migration model of Lee (1967) distinguished the characteristics of the emigration territory, of the immigration territory, the intervening obstacles such as distance, and the personal characteristics of the migrants such as gender and education. This 'bipole' and push–pull factors model integrated easily into the migration 'laws' that were developed in response to urbanization in Britain in the nineteenth century (Nas, 1974). Anthropologists followed the migrants to the cities in their own countries to discover their motives, their ways of adaptation and subcultures. 'Peasants in cities' became one of the influential titles that betrayed this evolving interest (Mangin, 1970). Later, migration studies also incorporated the growing number of international labor migrants to the Western and Middle Eastern cities in their research. As migrants often, although certainly not always, ended up as poor people in ghettos and slums, migration and slum culture were interrelated to some extent. As main elements in the rapid urbanization process, together they formed the roots of new forms of urbanism.

The formational phase ended in two approaches that are still dominant now and are exemplified by two types of 'introduction to urban anthropology,' one by Fox and the other by Gutkind. Fox (1977) singles out holism, participant observation, and cultural comparison as the main features of anthropology and he shows the different pathways urban anthropology has taken to study urbanism (lifestyle of city people in general), poverty (especially the poor based on life histories), and urbanization (alterations in urban social, associational, ethnic, and other structures in combination with urban problems). His holistic approach has led to the distinguishing of five types of cities: regal-ritual, administrative, mercantile, colonial, and industrial. These categories are discussed in their broader state and administrative settings and elaborately illustrated by case studies. He predicts that the dualism between holistic and microstudies in urban anthropology will be bridged by a focus on the relationships of the different social groups in the city with the larger regional and national social and administrative contexts. Gutkind (1974) has concentrated on African cities, demonstrating their relational, organizational, and institutional complexities in the context of the process of urbanization and the style of life of the inhabitants. He advocates the anthropological study of cities by exploring such concepts as tribalism, ethnicity, pluralism, and modernity, and dedicates considerable thought to rural–urban relations as well as political, work, family, and religious relations and ways of life. He discusses the specific methodological problems of field research in urban areas at length and, acknowledging the need for the application of knowledge, he warns of the false dichotomy between pure and applied studies.

Early Roots

In the 1960s and the 1970s, in attempts to develop urban studies and urban anthropology, urbanists began to search for earlier urban research results and approaches. Sennett's (1969) *Classic Essays on the Culture of Cities* with papers from the German School including Weber, Simmel, and Spengler and the Chicago School with Park, Wirth, and Redfield proved very helpful in deepening theoretical and methodological insights. The contributions of these scholars were widely read by students of urbanism.

Marx, Weber, Redfield, and the works of the Chicago School became especially influential. In 1921, Weber (1958) published his essay on the city in which he compared the medieval city in the Low Countries, the classical ancient cities, the early Renaissance city in Italy, and the Far Eastern cities. He applied his ideal-type method to his analysis and concluded that only the medieval cities in Europe had all the characteristics of the fully developed city, namely, a fortification, market, own juridical system, local political structure, and high level of autonomy. These medieval cities reached the highest level of heterogeneity that Weber claimed was characteristic of the basic role and function of the city. In

their linear evolutionary approach, Marx and Engels (1972, original 1846) were much more negative about the city as they interpreted the gap between town and countryside as an expression of the separation of cerebral and manual labor and one of the dominating forms of alienation. In their contrasting negative and positive views on the role of the city in culture and societal development, the views of Marx and Weber reveal the basic controversy that still characterizes Western culture. The positive attitude toward the countryside is also rooted in the outlook of the traditional European rural nobility, which romanticized country life.

The works of the Chicago School have been interesting not so much for their theoretical approach that uses an ecological analogy but for their urban ethnographic richness. Particularly influential are Wirth's view on urbanism as a way of life (1938) and Burgess' 1925 concentric zone model of Chicago, showing a number of concentric circles with different social and housing characteristics, varying in lifestyle, wealth, and ethnicity. This concentric zone model has been replicated in other cities all over the world, generating a lot of inspiration and new alternative modeling.

The difference between rural and urban was intensively researched by Redfield (1941) who chose a range of communities from traditional folk society to peasant village and city to analyze their characteristics, including worldview, family relations, and cohesion. His work was part of a tradition of community studies initiated by Robert and Helen Lynd in Middletown in the United States (1929) and restudied in 1937. In other countries such as the United Kingdom (Frankenberg, 1966), the community studies approach was espoused. The research on kinship and extended families in cities led Bott (1957) to distinguish close-knit and loose-knit family relations, thereby laying a basis for the network approach. The study of the differences between town and countryside, including those in Western countries, and applying quantitative approaches has attracted many scientists of diverse origins. It led to typologies of cities in various countries and to stereotyping the differences of rural and urban areas and later it has shown that many cities are home to traditional types of communities and some small rural communities can display a rather urban type of outlook.

The Chicago School combined the overview of different settlements in Mexico with elaborate research on areas, communities, and situations in Chicago. They covered a wide variety ranging from the ghetto and slum to the Gold Coast and from the hobo bohemians and gangs to the dancing immigrants in the taxi-dance hall. These studies have continued to captivate and stimulate urban anthropologists and have laid foundation for an abundance of studies of urban components, such as wards, communities, professions, meeting places, and events all over the world.

Two other sources of inspiration have been the Copperbelt and Java studies. Hannerz (1980) presents an overview of the early research in the Copperbelt, Zambia (Africa), by the Rhodes-Livingstone Institute with authors such as Wilson, Gluckman, Southall, Epstein, and Mitchell. Their work analyzed adaptations to life in the towns and raised the matter of whether an African townsman with his intensive tribal and rural relations remained tribesman or not. The conclusion was that a townsman was a townsman and that tribalism in towns was an entirely urban phenomenon. The migrant was detribalized in the city and deurbanized when he returned to the village. Particularly interesting are the situational analyses and the extended case study methods as developed in the Copperbelt studies. Gluckman (1940) takes the opening of a bridge as point of departure to obtain insights into the various persons and groups that take part in the ceremony and describes their relationships to explain the colonial conditions, *la situation coloniale* (Balandier, 1951). Mitchell (1956) also took a specific event, namely, the Kalela dance and the roles performed by the medical doctor, the nurse, and so on to shed light on tribal and colonial relations.

In the colonial period, the Java studies of cities were written in Dutch, often by scholars who had been trained as colonial administrators or were educated in Dutch universities in language and law of the Netherlands East Indies or architecture. They were extremely interested in the architecture, cultures, and languages of the Indonesian Archipelago and wrote extensively about what they encountered in their places of residence. Only in 1958 under the title *The Indonesian Town* (Wertheim et al., 1958) was part of their work published in translation. Town development in the Netherlands Indies, including traditional town formation, the analysis of the urbanization process and its accompanying problems, the layout of a regency town, the living conditions of the coolies and the administrative alterations in a small town were issues discussed at length. Many works from this colonial period have never been translated, including those of Tillema and Karsten, not to mention the masterpiece of de Haan (1922) on Old Batavia that was amply used in her PhD dissertation by Milone (1966). To abstract the specific features of colonial society in the Netherlands Indies composed of a Western and native section, the concepts of bazaar economy, dual economy, and plural society were defined. Later, these were elaborated by other concepts like the informal sector. All the studies on the Javanese and Copperbelt towns have greatly broadened the scope of urban anthropology and strengthened its consolidation in the 1970s and 1980s.

Consolidation

In the 1970s and 1980s, urban anthropology was still young, but its consolidation on the basis of the growing number of research works undertaken in cities, the increasing migration streams to Western cities leading to integration problems, and the rapid urbanization in cities outside Europe and the United States strikingly resulting in overwhelming slum areas and conspicuous poverty was imminent. Its three main pillars were (a) urban ethnography to highlight urban lifestyles and urbanism as city culture, (b) the comparative study of cities as wholes in various societies, and (c) anthropological studies of urbanization as a process of regional and global social change related to all sorts of problems, especially poverty and environmental issues.

At the end of the 1960s and the beginning of the 1970s, the general feeling in Western society and social science was fundamentally changed by the revolutionary quest for new societal forms distinct from the 'conservative' post-Second

World War reconstruction mentality. In urban anthropology, center–periphery and dependency theories were used to explain dependent urbanization in Latin America (Frank, 1967), Africa, and Asia. Galtung's (1971) model emphasizing a center and a periphery in both central and peripheral countries was intricate and shed considerable light on the dependency relations between cities in the world and the differences between Western and Southern urbanization processes. It was used for the interpretation of city size distributions in the world and the explanation of internal and international migration patterns (Nas, 1974).

La question urbaine by Castells (1972), presenting an elaborate Marxist approach to the city, was particularly revealing. Castells rejected the culturalist and harmonic approach to the city that explained urban culture as the consequence of simple demographic density and social complexity factors intensified by rural–urban migration and rapid urbanization. He was convinced that city culture should be explained by the social structure of the society as a whole and, particularly, the historical formation of the capitalist class structure and imperialism. Consequently, the cities in the First, Second, and Third World are completely different in character as they have been produced by capitalist industrialization in the West; imperialist, industrial, and financial domination in the Third World; and the socialist mode of production in the socialist countries. The Third World city consists of a concentration in large agglomerations tending to lack integration into a hierarchical urban network, with a great distance between large settlements and the countryside, an ecological juxtaposition of the autochthonous and Western city wards, a legacy of the colonial period, and a production system set up to serve the dominant Western societies. Hence, slums are not marginal or disintegrated as they often display an important degree of social cohesion. What characterizes inhabitants of slums is that they have interests different from those of the dominant class in society. In contrast to Third World urbanization, socialist urbanization is characterized by the primacy of politics, while First World urbanization displays large metropoles and megalopoles with a wide distribution of activities, functions, and groups, plus the disappearance of the difference between town and countryside. At the end of the 1980s, after the fall of the Berlin Wall and new economic policies in China, the Marxist approach was more or less abandoned. What did remain, however, was the view of the city as an arena in which various interest groups try to realize their ideas or their 'urban meanings' about the development of the city. A focus on urban protest movements in the urban arena was related to this (Castells, 1983). When one or other interest group has formed an elaborate vision for the future development of a city and is backed by the resources to realize its plans, it is called a strategic group (Evers and Schiel, 1988). Urban arena and urban meaning are still pivotal concepts in urban anthropology.

Later, the urbanization branch of urban anthropology was continued, for instance, in the work of Southall (1998), who focuses on the crisis in the capitalist mode of production where it is being affected by all sorts of environmental and social problems such as climate change, heightened risk of flooding, pollution, and other ecological disasters. Besides giving an analysis of the pristine, ancient, and feudal city, Southall discusses the Asian, Third World, and Western cities. His volume, which has apocalyptic overtones, is fairly comprehensive.

Meanwhile, besides their slums, both the informal sector of the cities of the Third World was widely investigated in an attempt to determine possible chances of development as well as urban ethnic minority problems arising from the influx of migrants from Third World countries to Western cities where these groups formed what became known as the Fourth World. These studies were of an applied nature and their purpose was to solve concrete problems. This drive to obtain research results in the expectation of defining and 'solving' problems provided a strong impetus to the consolidation of urban anthropology.

In 1979, the Society for Urban Anthropology was founded as a subdivision of the American Anthropological Association and in 1983 the International Union of Ethnological and Anthropological Sciences installed the Commission on Urban Anthropology (Nas and de Groot, 2009).

By now, urban anthropology was generally accepted as an important part of general anthropology and this implied that anthropology had turned its attention to both applied studies and to Western urban societies. These applied studies concentrated principally on slums, low-cost housing and *kampong* improvement, the informal sector, and urban ethnic minority studies tended to serve as a backup for urban policy and planning. Urban anthropology proliferated with an abundance of research on cities and urbanization in various countries on different continents, often published in collections of articles. It was also influenced by scholars from other disciplines as among them Lynch (1960), Turner (1976), and Kostov (1991) from architecture and Mumford (1961) and Braudel (1987) from history, and by a multitude of sociological studies (Hatt and Reiss, 1951; Ward, 1976), geography (Hauser and Schnore, 1965; Santos, 1971), and urban planning (Epstein, 1973; Friedmann, 1973).

Urban ethnography also steadily continued to broaden its boundaries by including all sorts of lifestyles and professions present in the urban arena, investigating specific ethnic and age groups, scavengers, construction workers, *becak* or pedicab drivers, prostitutes, and small middlemen and traders.

This proliferation of urbanization and ethnographic studies eclipsed the comparison of cities as wholes. Earlier Redfield and Singer (1954) had distinguished between orthogenetic and heterogenetic cities, the former being characterized by the continuation of traditional high culture and the latter by the integration of influences from other cultures. Hannerz (1980) differentiated Coke, Commerce, and Court towns and proposed an analytical model based on roles and role inventories to compare cities as wholes. In geography and sociology, the comparison of cities developed empirically and quantitatively, using large numbers of variables analyzed by factor analytical techniques.

Urban anthropology also made progress in theoretical approaches to urban societies, for example, by applying network theory to survey the city as a network of networks (Hannerz, 1980) and the dramaturgical perspective of Goffman (1959), who approaches institutions in urban society as if they had a front and a back stage on which persons present themselves in different ways.

Examples of important urban anthropological collections of articles that have broadened the horizon to different continents and subjects are the first publications of the aforementioned Commission on Urban Anthropology under the title *Town-Talk: The Dynamics of Urban Anthropology* (Ansari and Nas, 1983), with examples from the Gulf States, West Africa, Japan, and Venezuela, and *City and Society* (Southall et al., 1985) with Bulgarian, African, Kuwaiti, and Indian examples and studies on lifestyles, ethnic groups, middlemen in the grain trade, and workers in the supermarket industry. Other contributions with regional scope deal with Mediterranean (Kenny and Kertzer, 1983), African (Peil and Sada, 1984), and Indonesian cities (Nas, 1986) to mention just a few. Many publications broadening the scope of urban anthropology have followed, often with the words 'Third World cities' or 'cities in developing countries' in the title (Breese, 1972).

As mentioned earlier, in the 1970s and 1980s, slums and low-cost housing, the informal sector, and ethnicity were the hot topics. The thinking about slums in urban anthropology was heavily influenced by the architect John Turner (1978). He took exception to the idea that slums are cancers in the urban constellation and must be eradicated by destruction, relocation, and the introduction of massive low-cost housing projects by the government. Turner realized that new, low-cost housing was often too expensive for the inhabitants of the slums and argued for a decentralist approach in which planning and construction decisions were made by the users. He also stressed that slums represented enormous investments to their inhabitants, value that should not be ignored but appreciated. Turner reversed the negative attitude toward slums, turning it in a basically positive direction and pointed out that the needs of slum people are diverse, varying from location near work and low costs in the first phase of settlement in the city to housing security, only moving on to services in a later phase. These needs were related to the migration pattern in the city. *Le droit à la ville*, the right to the city, of the citizens was also stressed by Lefebvre (1968).

In the 1970s, the concept of 'informal sector' was very influential in urban anthropology. The idea was that, in developing countries especially, there were major informal activities that were not regulated, not counted in the statistics and hence remained outside the official administrative system. Among other factors, these activities dealt with small amounts of goods, services, and capital, thereby involving the application of traditional techniques, and they were often family based. In the formal sector, such activities would have required large amounts of goods and capital, and they would have been officially registered and based on contractual relationships. This distinction generated a flood of anthropological research in attempts to describe the informal sector and its development potentials. In the course of time, it became clear that the distinction was not as clear cut as had been expected. Members of one family could work in the formal and informal sectors and persons could even be involved in both sectors simultaneously. Some activities were shifted from one sector to another as it could be profitable not to be registered in the formal sector for tax avoidance purposes, even though this might imply less security. The conclusion was that perhaps the concepts of informalization and formalization would be better research tools. In explorations of development, expectations of the informal sector were raised high and many projects initiated. Hernando de Soto (1989) who gave examples from housing and transport generated particular optimism in the development of Peru by proposing to eliminate all sorts of official bureaucratic obstacles to entrepreneurship in the informal sector. The results turned out to be less than convincing, although the importance of entrepreneurship in development is widely recognized nowadays. The informal sector approach generated a wealth of research on all sorts of urban activities, including street-sweepers, scavenger communities, street children, small itinerant traders, construction workers, and persons operating traditional means of transport.

Alongside slums and the informal sector, urban ethnicity emerged at this phase of consolidation. It concerns people who share a system of norms and values and have to interact with other population groups in 'complex societies,' as cities tend to be increasingly labeled (Cohen, 1974). Hence, urban ethnicity is a very general concept, rightly ignoring the almost racial connotations it might have. Cohen (1974) shows the possible interrelationship of urban ethnicity with social inequality and political power. When class and ethnicity coincide, they could reinforce each other, and political power could be acquired on the basis of ethnic connections. Some of the examples of ethnic groups in Cohen's book are the Hausa in Nigeria, the Protestant and Roman Catholic groups in Northern Ireland, and the economic elite dominating the city of London. However, many other countries also provide examples of this type of group, for instance, the Pakistanis in the United Kingdom, the Chinese in Indonesia, and the Moroccans in the Netherlands. When migrants settle in a society, as time passes, their adaptation process can lead to integration or assimilation, but this often produces problems in the larger society. Importantly, all sorts of other modes of connection of ethnic groups with the receiving society are possible and are found in history. As seen earlier, in Africa urban ethnicity was discussed in terms of tribalism and in Latin America regional clubs have been important (Jongkind, 1974). Initially, these clubs were considered as a vehicle of security and adaptation; later it emerged that the leaders and members of regional clubs were already successful in the city and did not need them for upward social mobility.

To round off this phase on the basis of two introductions in urban anthropology (Eames and Goode, 1977; Press and Smith, 1980), one can list the following main issues: (a) types of cities and their roles, (b) urbanism and the process of urban change, (c) urban differentiation (gender, social class and ethnic groups, urban sites and places, and professions in cities), (c) urban integration (families, associations, networks, situations, and meetings), and (e) methods of research.

Current Developments

The last two decades have been shaped by various fundamental developments. These have directed the urbanization branch of urban anthropology away from its focus on complex societies to the even more complex global network society. The fall of the Berlin Wall has created new political conditions that unite the world in many aspects and the distribution of new

telephone and Internet facilities has transformed the whole world providing easier direct long-distance connections that have transformed the world financial system, international business, market relations, and local and regional identities. Generally speaking, this has become an intense process in both globalization and localization. In a three-volume masterpiece, Castells (1996–98) has described and analyzed these current processes and their impact on the city. He distinguishes the 'space of places' and the 'space of flows' in megacities all over the world, referring, respectively, to a local and historical community part of a city and a global, ahistorical part linked to the global flow of ideas, money, and so on. Both have their own ways of life, even architecture, featuring the local, more traditional constructions on the one hand and the global, starkly, and modern architecture on the other hand.

Different regional urban experiences were combined by Gugler (1996) into one volume listing Monsoon Asia, China, India, Indonesia, the Arab world, Africa South of the Sahara, and Latin America in an attempt to describe convergences and divergences in urbanization. In this branch of urban anthropology, the different sciences have been almost completely integrated. Hence, environmental aspects have been added to these studies by dint of focusing on certain relevant areas in the city and the purpose is often applied research. Other examples of regional specification deal with Japanese cities (Karan and Stapleton, 1997) and Caribbean cities (Jaffe, 2008).

In the ethnographic branch of urban anthropology, more and more parts of cities are being covered, including particular buildings, the mall, the square, plaza or *alun-alun*, the street, the neighborhood, urban greenery, water problems, graveyards, and urban agricultural and architectural conservation areas.

Data are also being acquired about different ethnic and social groups in cities as well as about social and economic activities. The employment of and changes in the use of mobile phones and the Internet and participation in Internet communities is especially interesting. There is an increasing tendency to frame ethnicity as multiculturalism. Guliana Prato (2009) raises the question to what degree multiculturalism is compatible with citizenship. She points out that the receiving society is not static and does not consist of one single culture, while the immigrant cultures often are also composed of many cultures and are being altered because of their connections with the larger societal settings (Harris, 2011). The number of studies on women in urban areas, investigating their entrepreneurship, is growing, which means that gender studies are increasingly integrated into urban anthropology (Jarvis et al., 2009).

The four volumes of the *Encyclopedia of Urban Cultures: Cities and Cultures around the World* (Ember and Ember, 2002) form an inclusive inventory of data on a great number of cities in the world, providing general urbanization and ethnographic information.

In the branch of urban anthropology directed toward comparing cities as a whole, a new approach has evolved under the label 'urban symbolism' (Nas, 1993; Nas and Samuels, 2006). In this approach, the means of existence is not the point of departure in the characterization and comparison of cities, but has to cede precedence to the symbolic and ritual dimensions that reflects other aspects such as the work relationships. A multitude of cities have been studied in this framework and this has led to new concepts for the study of cities as a whole inspired by the Chicago School spatial models, semiotic signification ideas, and social cohesion studies. Cities seem to abound in symbols (Nas, 2011) that can be used to characterize them and to describe their soul. Urban symbolism studies can be used as a tool in city branding and marketing. This urban symbolism approach corresponds to the insights of Lefebvre (1974) into the production of space, that distinguishes natural space, the perception of space, and the representation of space; Harvey (1990) on time and space; and the work of Nora (1984–92) and his realms of memory. In the framework of urban symbolism, Dürr (2005) has published a case study of Albuquerque in the United States and Till (2005) of Berlin in Germany. Shahshahani (1998) has analyzed Esfahan in Iran and Colombijn (1998) Canberra in Australia to mention a few examples.

One can conclude that in current urban anthropology studies, many old themes are being continued in new ways. Migration has been transformed into international mobility by a network approach and takes place in transnational spaces (Smart, 1999), urban ethnicity as a further phase beyond multiculturalism (Prato, 2009), and the urban way of life with solitude alongside solidarity and loneliness (Coleman, 2009). Interesting new developments cover many topics such as gated communities (Low, 2003), the influence of mobile phones and the Internet (Postill, 2008), wounded cities (Schneider and Susser, 2003), urban heritage conservation (Brumann, 2009), traditional and colonial architecture (Schefold et al., 2003–08; Nas, 2006), and a host of books and essays about urban agriculture and farming, low-density agrarian-based urbanism (labeled *kotadesasasi*), urban soundscapes, the city at night, environmental activism, besides all sorts of urban localities and communities, among them neighborhoods, an urban skateboarding community, and a boxing gym.

Many of these themes refer to worldwide processes. In 1998, the prevailing view on globalization prompted the Society for Urban Anthropology in the United States to change its name to the Society for Urban, National, and Transnational Anthropology. In 1983, the cooperation between urban anthropologists and representatives of other sciences led to the setting up of an Urban Laboratory (as part of National Center for Scientific Research (CNRS)) in Paris. Jacques Gutwirth and Colette Pétonnet (1987) played an important role in the foundation of this *Laboratoire d'Urbanisme*.

Urban anthropology has passed through a great number of theoretical-, concept-, or reality-driven approaches. These have covered the urban ecology analogy, functionalism, network analysis, the dramaturgical perspective, migration models, culture of poverty research, community studies, situational analysis, the extended case study approach, ideal-type comparison, Marxist approaches, structural imperialism, informal sector research, social inequality research, urban symbolism, transnational research, and so on. The anthropological study of the city has proved flexible and able to absorb the developments in general anthropology and other related sciences as well as the outside urban realities that shape our world.

Fieldwork

Field research in cities and megacities differs from the traditional fieldwork in rural and tribal areas as the researcher is generally part of a multimillion population and often known to a limited number of people in the city. The globalization of society has created new possibilities for frequent short visits to the research area. Once long fieldwork visits of one or several years dominated and were preferred, whereas nowadays a series of shorter visits to the same area are possible, even necessary because of the publish-or-perish pressures in the academic world. It has also become possible to study multiple locations, for example, in the case of migration both the area of departure and the receiving area. Marcus (1995), Hannerz (2003), and Falzon (2009) have elaborated on this relatively new fieldwork tendency called multisited or multilocal ethnography, and contrasted it to the 'fly-on-the-wall perspective' (Katz, 2010: p. 26). Under the motto of "follow people, connections, associations, and relationships across space (because they are substantially continuous but spatially non-continuous)" (Falzon, 2009: pp. 1–2), now it is even possible to investigate certain research objects in different locations and describe their fate in various communities. An interesting example of 'follow the thing' is Greek whisky (Bampilis, 2010). In urban anthropology, besides one-location fieldwork in a specific neighborhood or on a specific square, multisited fieldwork is quite common, as when symbols and rituals are 'followed.' In a nutshell, under the influence of globalization, present-day fieldwork tendencies cover a much wider sequence of locations rather than being concentrated in the area of one city or agglomeration.

To illustrate the specific character of urban fieldwork, the concept of the *flâneur* might be instructive. In contrast to research in small urban communities where everybody is soon aware of the stranger in the community, in the city as a whole, particularly in megacities, the anthropologist is immersed in a crowd of people and a multitude of places. His first task is to familiarize himself with different places: their uses, functions, and the people who frequent them. This exercise transforms the anthropologist into a *flâneur*: a stroller, a wanderer, who has no specific relationship with the places he or she is interested in and is eager to make contact with everyone who happen to cross his path in his quest to discover what the inhabitants do and think. The *flâneur* is an observer and a participant collecting general knowledge about the city. This task is the starting point for developing a view founded on his personal background, including the relevant literature, but can later blossom into an intimate knowledge of the views expressed by certain categories and groups of inhabitants, especially those of the key informants who play a prominent role in the research. Besides being an observer and participant, the *flâneur* is also a detective because his ultimate goal is to combine all he has learned into one narrative that will unite the fragmented views held by the inhabitants in one, single model and thereby expose the soul of the city, its structure, and dynamics.

Charles Baudelaire, who introduced the concept of *flâneur* in 1863, wrote the following: "For the perfect *flâneur*, for the passionate spectator, it is an immense joy to set up house in the heart of the multitude, amid the ebb and flow of movement, in the midst of the fugitive and the infinite. To be away from home and yet to feel oneself everywhere at home; to see the world, to be at the centre of the world, and yet to remain hidden from the world …" (Baudelaire, 1995: p. 9). The flâneur is "the painter of the passing moment and of all the suggestions of eternity that it contains" (Baudelaire, 1995: p. 5). "He marvels at the eternal beauty and the amazing harmony of life in the capital cities … He gazes upon the landscapes of the great city …" (Baudelaire, 1995: pp. 11). The message conveyed by Baudelaire is that the *flâneur* is a member of the crowd and a connoisseur of urban life; he is an active, intellectual observer driven by curiosity and he combines the casual eye of the stroller with the purposeful gaze of the detective.

The true home of the *flâneur* was nineteenth-century Paris whose arcades provided his principal domain (Hazel Hahn, 2006); the city was his interior, his house. That is why sometimes, in the guise of a public type, the *flâneur* is restricted to nineteenth-century Paris in the scientific literature. He is considered to have disappeared when the shopping arcades were superseded by large department stores exuding a full-fledged consumer ideology, and when the interior of those shops became the exterior of the city and the domain of the consumer public.

Walter Benjamin (1991: p. 524) elaborates the contrast between the *flâneur* and the traveler or tourist. The tourist restricts his attention to the monumental and places of special interest, 'the exotic and picturesque' as it were (Goebel, 1998: p. 379), and is not actually involved in either urban history or city life. The *flâneur* tries to incorporate the native view, as he is interested in the background and history of the urban landscape and society. *Flânerie* emerges from the confrontation between the city and the biographical background of the *flâneur*, or the 'dialectical movement of familiarity and strangeness' (Goebel, 1998: p. 379). In this dialectical movement, the *flâneur* is also a detective. It suits him to hold himself apparently inert among the urban masses, but in reality he is a passionate observer and participant (Benjamin, 1991: p. 554), eager to understand the city and, like a journalist, takes notes on urban life. *Flânerie* is 'an activity of ethnographic exploration' (Goebel, 1998: p. 389).

In the course of time, the theory on the *flâneur* and *flânerie* has given rise to specific methods of enjoying and exploring the city. Schulte Nordholt (2008) discusses these in her examination of Georges Perec and his sociology of daily life. The immobile *flâneur* might choose to take up his post in a particular café or some other location and sit there to observe what is happening in the street. The mobile *flâneur* can follow one person during the day and write down what the person observed has been doing, perhaps visiting shops, talking with friends, and so forth. He can also take up his position three times a day at different spots in a particular square and describe what happens from different angles. Alternatively, he can observe the same place for one day over a period of several years. An even more common method is just to roam the city at random happening upon the unknown places and vistas, becoming acquainted with the architecture and plazas, relishing accidental encounters, and festivities. Nowadays, urbanists and architects still opt for this way of exploring the city.

Another possible step could be to include an imaginary exploration of the city by drawing mental maps or even by

narrative mapping without setting a map down on paper. In present-day research on urban symbolism, mental and narrative mapping play an essential role as they form a rich source of insight into the city as it has been expressed by informants.

As *flâneur,* the aim of the urban anthropologist is to achieve a synthetic approach to the order and dynamics of a city, repeatedly using observation, participation, interviewing, and reflection. This is a creative and iterative process repeated until an optimal insight has been achieved. The *flâneur* is the symbol of the anthropologist mastering the city. The *flâneur* is an urban anthropologist *avant la lettre* and the urban anthropologist is a modern *flâneur*.

See also: Brokers and Brokerage, Anthropology of; Center–Periphery Relationships; Cities: Capital, Global, and World; Community Studies: Anthropological; Cross-Cultural Research Methods in Sociology; Development and Urbanization; Ethnography in Applied Social Research; Ethnography; Ghetto; Global Migration; Informal Sector; Inner City; Multiculturalism, Anthropology of; Multisited Field Studies; Poverty, Culture of; Symbolism in Anthropology; Urban Geography; Urban History; Urban Sociology; Urban Studies: Overview.

Bibliography

Ansari, Ghaus, Nas, Peter J.M. (Eds.), 1983. Town-Talk: The Dynamics of Urban Anthropology. E.J. Brill, Leiden.
Balandier, George, 1951. La situation coloniale: approche théorique. Cahiers Internationaux de Sociologie 11, 44–79.
Bampilis, Tryfon, 2010. Greek whisky: The localization of a global commodity, PhD Thesis. Leiden.
Baudelaire, Charles, 1995. The Painter of Modern Life and Other Essays. Phaidon Press, London (Original 1863).
Benjamin, Walter, 1991. Das Passagen-werk, Gesammelte Schriften, Band V-1. Suhrkampf, Frankfurt am Main (Written between 1927–1940).
Bott, Elizabeth, 1957. Family and Social Network: Roles, Norms and External Relationships in Ordinary Urban Families. Tavistock, London.
Braudel, Fernand, 1987. Beschaving, Economie en Kapitalisme (15e-18e Eeuw). Deel 1. De Structuur van het Dagelijks Leven. Contact, Amsterdam.
Breese, Gerald (Ed.), 1972. The City in Newly Developing Countries: Readings on Urbanism and Urbanization. Prentice Hall International, London.
Brumann, Christoph, 2009. Outside the glass case: the social life of urban heritage in Kyoto. American Ethnologist 36, 276–299.
Burgess, Ernest W., 1925. The growth of the city: an introduction to a research project. In: Park, Robert E., Burgess, Ernest W., McKenzie, Roderick D. (Eds.), The City: Suggestions for the Investigation of Human Behavior in the Urban Environment. University of Chicago Press, Chicago, pp. 47–62.
Castells, Manuel, 1972. La Question Urbaine. François Maspero, Paris.
Castells, Manuel, 1983. The City and the Grassroots: A Cross-Cultural Theory of Urban Social Movements. University of California Press, Berkeley.
Castells, Manuel, 1996–1998. The Information Age: Economy, Society and Culture, vols. 3. Blackwell, Cambridge.
Cohen, Abner (Ed.), 1974. Urban Ethnicity. Tavistock, London.
Coleman, Leo, 2009. Being alone together: from solidarity to solitude in urban anthropology. Anthropological Quarterly 82 (3), 755–777.
Colombijn, Freek, 1998. Canberra: a sheep in wolf's clothing. International Journal of Urban and Regional Research 22 (4), 565–581.
Dürr, Eveline, 2005. Identitäten und Sinnbezüge in der Stadt: Hispanics im Südwesten der USA. Lit, Münster.
Eames, Edwin, Goode, Judith Granich, 1977. Anthropology of the City: An Introduction to Urban Anthropology. Prentice-Hall, Englewood Cliffs.
Ember, Melvin, Ember, Carol R., 2002. Encyclopedia of Urban Cultures: Cities and Cultures around the World, vols. 4. Scholastic Company, Danbury.
Epstein, David G., 1973. Brasília, Plan and Reality: A Study of Planned and Spontaneous Urban Development. University of California Press, Berkeley.
Evans-Pritchard, Edward E., 1963. Essays in Social Anthropology. The Free Press of Glencoe, New York (Original 1962).
Evers, Hans-Dieter, Schiel, Tilman, 1988. Strategische Gruppen: Vergleichende Studien zu Staat, Bürokratie und Klassenbildung in der Dritte Welt. Dietrich Reimer Verlag, Berlin.
Falzon, Mark-Anthony (Ed.), 2009. Multi-sited Ethnography: Theory, Praxis and Locality in Contemporary Research. Ashgate, Burlington.
Fox, Richard G., 1977. Urban Anthropology: Cities in their Cultural Settings. Prentice-Hall, Englewood Cliffs.
Frank, André Gunder, 1967. Capitalism and Underdevelopment in Latin America. Monthly Review Press, London.
Frankenberg, Ronald, 1966. Communities in Britain: Social Life in Town and Country. Penguin, Harmondsworth.
Friedmann, John, 1973. Urbanization, Planning and National Development. Sage, Beverly Hills.
Galtung, Johan, 1971. A structural theory of imperialism. Journal of Peace Research 2, 81–117.
Gluckman, Max, 1940. Analysis of a social situation in modern Zululand. Bantu Studies 14, 1–30, 147–174.
Goebel, R.J., 1998. Benjamin's flâneur in Japan: urban modernity and conceptual relocation. The German Quarterly 71 (4), 377–391.
Goffman, Erving, 1959. The Presentation of Self in Everyday Life. Doubleday, Garden City.
Gugler, Josef (Ed.), 1996. The Urban Transformation of the Developing World. Oxford University Press, Oxford.
Gutkind, Peter C.W., 1974. Urban Anthropology: Perspectives on 'Third World' Urbanization and Urbanism. Van Gorcum, Assen.
Gutwirth, Jacques, Pétonnet, Colette (Eds.), 1987. Chemins de la ville: Enquêtes ethnologiques. Editions du Comité des Travaux Historiques et Scientifiques (C.T.H.S.), Paris.
de Haan, Frederik, 1922. Oud Batavia, vols. 2. Kolff, Batavia.
Hahn, Hazel, Haejeong, 2006. Du flâneur au consommateur: spectacle et consommation sur les Grands Boulevards, 1840–1914. Romantisme 134 (4), 67–78.
Hannerz, Ulf, 1969. Soulside: Inquiries into Ghetto Culture and Community. Columbia University Press, New York/London.
Hannerz, Ulf, 1980. Exploring the City: Inquiries toward an Urban Anthropology. Columbia University Press, New York.
Hannerz, Ulf, 2003. Being there … and there … and there! Reflections on multi-site ethnography. Ethnography 4 (2), 201–216.
Hardoy, Jorge E., 1973. Pre-Columbian Cities. Allen and Unwin, London.
Harris, Rosemary, 2011. Multiculturalism: theoretical challenges from anthropology. Urbanities 1 (1), 70–75.
Harvey, David, 1990. The Condition of Postmodernity. Blackwell, Cambridge.
Hatt, Paul K., Reiss Jr., Albert J. (Eds.), 1951. Cities and Society: The Revised Reader in Urban Sociology. The Free Press of Glencoe, New York.
Hauser, Philip M., Schnore, Leo F. (Eds.), 1965. The Study of Urbanization. John Wiley, New York.
Jaffe, Rivke (Ed.), 2008. The Caribbean City. Ian Randle, Kingston.
Jarvis, Helen, Kantor, Paula, Cloke, Jonathan, 2009. Cities and Gender. Routledge, Abingdon.
Jongkind, Coenraad F., 1974. Regionale clubs in Lima, Peru, PhD thesis. Amsterdam.
Karan, P.P., Stapleton, Kristin (Eds.), 1997. The Japanese City. University Press of Kentucky, Lexington.
Katz, Jack, 2010. Time for new urban ethnographies. Ethnography 11 (1), 25–44.
Keesing, Felix M., 1958. Cultural Anthropology. Rinehart, New York.
Kenny, Michael, Kertzer, David I. (Eds.), 1983. Urban Life in Mediterranean Europe: Anthropological Perspectives. University of Illinois Press, Urbana.
Klanfer, Jules, 1965. L'exclusion sociale: Etude de la marginalité dans les sociétés occidentales. Cahiers Science et Service, vol. 2. Bureau de Recherches Sociales, Paris.
Kostov, Spiro, 1991. The City Shaped: Urban Patterns and Meanings through History. Thames and Hudson, London.
Krapf-Askari, Eva, 1969. Yoruba Towns and Cities: An Enquiry into the Nature of Urban Social Phenomena. Clarendon Press, Oxford.
Labbens, Jean, 1965. La condition sous-proletarienne: L'héritage du passé. Cahiers Science et Service, vol. 1. Bureau de Recherches Sociales, Paris.
Lee, E.S., 1967. A theory of migration. Ekistics 23 (137), 211–216.
Lefebvre, Henri, 1968. Le droit à la ville suivi de Espace et politique. Anthropos, Paris.
Lefebvre, Henri, 1974. La production de l'espace. Anthropos, Paris.
Lévi-Strauss, Claude, 1955. Tristes Tropiques. Librarie Plon, Paris.
Lewis, Oscar, 1959. Five Families: Mexican Case Studies in the Culture of Poverty. Basic Books, New York.

Lewis, Oscar, 1968. La Vida: A Puerto Rican Family in the Culture of Poverty: San Juan and New York. Vintage, New York.
Low, Setha, 2003. Behind the Gates: Life, Security, and the Pursuit of Happiness in Fortress America. Taylor and Francis, New York.
Lynch, Kevin, 1960. The Image of the City. M.I.T. Press, Cambridge.
Lynd, Robert S., Lynd, Helen M., 1929. Middletown: A Study in Modern American Culture. Constable, London.
Lynd, Robert S., Lynd, Helen M., 1937. Middletown in Transition: A Study in Cultural Conflicts. Harcourt Brace Jovanovitch, New York.
Mangin, William (Ed.), 1970. Peasants in Cities: Readings in the Anthropology of Urbanization. Houghton Mifflin, Boston.
Marcus, George E., 1995. Ethnography in/of the world system: the emergence of multi-sited ethnography. Annual Review of Anthropology 24, 95–117.
Marx, Karl, Engels, Friedrich, 1972. De Duitse Ideologie, Deel 1, Feuerbach. SUN, Nijmegen (Original 1846).
Milone, Pauline, 1966. Queen City of the East: The Metamorphosis of a Colonial Capital. University Microfilms, Ann Arbor.
Mitchell, J. Clyde, 1956. The Kalela Dance. Manchester University Press, Manchester.
Mumford, Lewis, 1961. The City in History: Its Origins, Its Transformations, and Its Prospects. Harcourt, Brace and World, New York.
Nas, Peter J.M., 1974. Imperialism, city size distributions and migration. Sociologia Neerlandica 10 (3), 219–232 (Original in Dutch 1973).
Nas, Peter J.M. (Ed.), 1986. The Indonesian City: Studies in Urban Development and Planning. Foris Publications, Dordrecht and Cinnaminson.
Nas, Peter J.M., 1990. De stad in de Derde Wereld. Coutinho, Muiderberg.
Nas, Peter J.M. (Ed.), 1993. Urban Symbolism. E.J. Brill, Leiden.
Nas, Peter J.M. (Ed.), 2006. The Past in the Present: Architecture in Indonesia. NAi Publishers, Rotterdam.
Nas, Peter J.M. (Ed.), 2011. Cities Full of Symbols: A Theory of Urban Space and Culture. Leiden University Press, Leiden.
Nas, Peter J.M., de Groot, Marlies, 2009. The IUAES from past to present. In: Nas, Peter J.M., Jijiao, Zhang (Eds.), Anthropology Now. Intellectual Property Publishing House, China, pp. 417–461.
Nas, Peter J.M., Samuels, Annemarie (Eds.), 2006. Hypercity: The Symbolic Side of Urbanism. Kegan Paul, London.
Nora, Pierre, 1984–1992. Les lieux de mémoire. Gallimard, Paris.
Nordholt, Schulte, Annelies E., 2008. Georges Perec: topographies parissiennes du flaneur. Relief 2 (1), 66–86.
Peil, Margaret, Sada, Pius O., 1984. African Urban Society. John Wiley, Chichester.
Pirenne, Henri, 1952. Medieval Cities: Their Origins and the Revival of Trade. Garden City, Doubleday.
Postill, John, 2008. Localizing the internet beyond communities and networks. New Media and Society 10 (3), 413–431.
Prato, Giuliana B. (Ed.), 2009. Beyond Multiculturalism: Views from Anthropology. Ashgate, Farnham.
Press, Irwin, Smith, M. Estellie (Eds.), 1980. Urban Place and Process: Readings in the Anthropology of Cities. Macmillan, New York.
Redfield, Robert, 1941. The Folk Culture of Yucatan. University of Chicago Press, Chicago.
Redfield, Robert, Singer, Milton, 1954. The cultural role of cities. Economic Development and Cultural Change 3 (1), 53–73.
Santos, Milton, 1971. Les villes du Tier Monde. Editions M.-Th. Génin, Paris.
Schefold, Reimar, Nas, Peter J.M., Domenig, Gaudenz, 2003–2008. Indonesian Houses, vols. 2. KITLV Press, Leiden.
Schneider, Jane, Susser, Ida (Eds.), 2003. Wounded Cities: Destruction and Reconstruction in a Globalized World. Berg, Oxford.
Sennett, Richard (Ed.), 1969. Classic Essays on the Culture of Cities. Meredith, New York.
Shahshahani, Soheila, 1998. Esfahan's gardens and mosques: on the instrumentality of symbols as a means of retaining urban identity. International Journal of Urban and Regional Research 22 (2), 602–613.
Sjoberg, Gideon, 1965. The Preindustrial City: Past and Present. The Free Press, New York (Original 1960).
Smart, Alan, 1999. Participating in the global: transnational social networks and urban anthropology. City and Society 11, 1–2, 59–77.
de Soto, Hernando, 1989. The Other Path: The Invisible Revolution in the Third World. Harper and Row, New York.
Southall, Aidan, 1998. The City in Time and Space. Cambridge University Press, Cambridge.
Southall, Aidan, Nas, Peter J.M., Ansari, Ghaus (Eds.), 1985. City and Society: Studies in Urban Ethnicity, Life-Style and Class. Institute of Cultural and Social Studies, University of Leiden, Leiden.
Till, Karen E., 2005. The New Berlin: Memory, Politics, Place. University of Minnesota Press, Minneapolis.
Toelichting, 1938. Toelichting op de 'Stadsvormingsordonnantie Stadsgemeenten Java'. Landsdrukkerij, Batavia.
Turner, John F.C., 1976. Housing by People: Towards Autonomy in Building Environments. Marion Boyars, London.
Turner, John F.C., 1978. Menselijk Wonen – Anders Wonen: Over Zelfbeschikking in de Woonomgeving. Het Wereldvenster, Baarn.
Ward, Barbara, 1976. The Home of Man. Penguin, Harmondsworth.
Weber, Max, 1958. The City. The Free Press, New York (Original 1921).
Wertheim, Willem F., et al., 1958. The Indonesian Town. Van Hoeve, The Hague and Bandung.
Wirth, Louis, 1938. Urbanism as a way of life. The American Journal of Sociology 44 (1), 1–24.

Urban Conflicts

Liam O'Dowd, Queen's University, Belfast, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Three sources of urban conflicts are identified: (1) changing state–city relationships; (2) the relationship between the dynamics of capitalist development and cities; and (3) the specific dynamics of urban life and the urban environment where the city itself is seen as a causal variable. Two sets of questions crosscut all three strands. The first addresses how violent conflicts can be regulated, transformed, and rendered into more constructive nonviolent conflicts through the processes of urban civil society. The second concerns how, why, and where urban conflicts turn violent and with what consequences. In summary, cities now rival states as arenas and stakes in political conflict and urban conflicts have an increasing transnational and transcultural salience, which underlines the necessity for more sustained comparative analysis.

Three Research Frameworks: An Introduction

Urban conflicts have been framed and constituted historically by the making and unmaking of states, by changing patterns of economic development and by the specific effects of densely settled populations living in close proximity. As objects of research, they vary enormously across time and space, in terms of their causes, outcomes and the forms they assume. In order to make sense of a highly diverse and unwieldy research field, it is useful to identify three underlying frameworks of analysis, each of which sees urban conflicts as shaped by a different process or dynamic. These include (1) changing state–city relationships; (2) the dynamics of global capitalism; and (3) the specific dynamics of urban life and the urban environment. While these approaches may be distinguished in terms of what they choose to emphasize, in practice, i.e., in any one study, these processes may be frequently conceived as overlapping, intertwining or interpenetrating.

Violent and Nonviolent Urban Conflicts

Implicitly at least, social scientific debates on all three processes distinguish between violent and nonviolent urban conflicts. In practice, however, the research and policy literature has been increasingly drawn to questions of urban violence, its causes, consequences, multiple forms, and how it might be defined, regulated, or transformed. Research on urban violence needs to be contextualized within a much wider social science tradition on the multiple origins, definitions, manifestations, and consequences of violence generally and the various ways in which it may be studied empirically (see Heitmeyer and Hagan, 2003). Understanding the scale, diverse manifestations, and variability of urban violence has become the touchstone of much research on urban conflicts. Muggah (2012), for example, provides a detailed overview of the global literature on urban violence and includes an annex, which lists the most prominent research networks and Internet resources on the topic (see also, World Bank, 2010; Agostini et al., 2010). At times, the problematizing of urban violence tends to obscure the fact that conflict per se is the *sine qua non* of urban life and that much of this conflict remains nonviolent, constructive, and agonistic rather than violent, destructive, and antagonistic.

Cities generally have a long history of managing nonviolent conflict and in creating institutions that regulate agonistic conflict. Cities' association with the development of civil society is a recurring theme linking them to a capacity to 'politicize' or moderate conflicts among key violence-prone groups (Varshney, 2002; Keane, 2006). In his History of Violence, for example, Muchembled (2012) accords European towns of the Late Middle Ages, a precocious status in the 'civilizing process'. He suggests that, through a form of 'municipalization of violence', they regulated and pacified those prone to disorder and violence such as young males, prostitutes, and criminals. Young males, however, continue to be seen as a problematic category, i.e., as perpetrators and victims of urban violence, both transhistorically and transculturally (Muggah, 2012: 48). Nevertheless, there is huge variability in rates of urban homicide, with dramatically declining rates in Europe over the past four centuries Muchembled (2012) and currently high rates in some Latin American and sub-Saharan cities (Muggah, 2012: 47–50).

There is no unambiguous connection between cities and forms of conflict. On the one hand, cities can represent protection, safety, resilience, dense networks of reciprocity, economic productivity, and functioning welfare systems (formal and informal). On the other hand, they may be sites of extreme forms of violence, war, insecurity, inequality, poverty, and fear (Muggah, 2012: 2).

In sum, much contemporary research on urban conflicts, across all three frameworks identified above, revolves around two sets of questions. The first addresses how violent urban conflicts can be regulated, transformed, and rendered into more constructive nonviolent conflicts through the processes of civil society. The second concerns how, why, and where conflicts turn violent and with what consequences? Agostini et al. (http://www.eisf.eu/resources/item/?d=41760), for example, seek to develop an analytical framework for understanding the processes of economic, political, and social violence by focusing on the interaction between risk factors in three cities: Nairobi, Kinshasa, and Bogota. The Urban Tipping Point research project led by Moser and Rodgers, studies four cities, Nairobi, Dili, Patna, and Santiago/Chile, in an attempt to understand how the conflicts in inherent in urban life (associated with poverty, youth, political exclusion, and gender-based insecurity) can tip

over into generalized urban violence (http://urbantippingpoint.org). Violent urban conflicts are highly heterogeneous, multicausal, spatially uneven, and typically episodic. Moreover, they also generate long-term urban effects enhancing spatial segregation, discrimination, strengthening ethnic, class, gender, racial, and intergenerational divisions (Savage and Muggah, 2012: 3–4).

Contemporary social research across the three frameworks identified above suggests that it utopian, however, to pursue some kind of urban or societal consensus aimed at ending urban conflict *tout court*. Programs of social homogenization and assimilation driven by projects of state and nation-building and capitalist development, almost by definition, are as likely to generate as to eliminate urban conflicts. Similarly, it is impractical to assume that strategies of urban planning, control, surveillance, or the use of lethal force have the capacity to ensure long-term urban stability. Coercive forms of domination inevitably produce collective resistance. The more constructive or agonistic urban conflicts are those that encourage negotiation, bargaining, and mutual respect among diverse groups as a means of combatting poverty, marginalization, and inequality. It remains difficult in practice, however, to disentangle urban conflicts from the often violent process of state formation and distintegration.

Changing City–State Relationships as a Source of Urban Conflicts

As Tilly (1992) has shown, cities have long been integral to European state formation and the wars associated with it. In the era of city-states, urban conflicts were self-evidently state conflicts. With the emergence of the great overseas and land empires from the sixteenth century onward, and the spread of national states from 1945 onward, cities were increasingly contained and governed by overarching states. States, not cities, became the preeminent stakes in political conflict, the locus of military, political administrative, and ideological power and, additionally, the key regulators of economic power. Consequently, one of the major strands in urban research addresses how state formation and governance have become key determinants of conflict in cities. With accelerating urbanization globally, urban space is increasingly used to make political claims and urban conflicts are constituted by wider intra- and interstate conflicts (Savitch, 2008; Coward, 2009; Graham, 2010; Davis and Libertun de Duren, 2011; http://www.conflictincities.org).

The proliferation of states (from 51 in 1945 to 193 in 2012) has generated much conflict (violent and otherwise) although the scale and intensity of this conflict has fluctuated considerably over space and time. In general, intrastate wars have predominated over interstate wars – the former accounting for an estimated 16.2 million deaths, the latter for 3.3 million deaths in the 60 years after 1945 (Fearon and Laitin, 2003). With the increasing urbanization of the world's population, and the pivotal economic, administrative, and symbolic roles that cities play in national states and in the global, geopolitical system, the changing relationship between cities and organized political violence has come in for increased scrutiny (Graham, 2010; Gregory, 2010; Coward, 2009). Cities have become key sites for the assertion, erosion, and contestation of state sovereignty (Beall and Fox 2009; Davis and Libertun de Duren 2011). According to Graham (2010), the geopolitics of interstate war has generated a new military urbanism, where the city is a 'battle space' provoking a new emphasis on securitization. Here the historical capacity of the most powerful states to externalize organized political violence is undermined, as their own cities become key frontiers of terrorism, counterinsurgency, and crime prevention.

Cities, here, are largely, if not exclusively, seen as the pawns or the victims of the various protagonists of interstate and intrastate wars. While the number of UN recognized states have multiplied, they vary hugely in terms of their infrastructural power (economic, political, ideological, and military). This impacts on their capacity to control, respond to, or exacerbate urban conflicts on home territory or in other, weaker states – for example, in the Middle East, the Balkans, or Africa (Graham, 2010; Beall et al., 2013). As Mann (2012) argues, the core states most deeply embedded in the global capitalist economy in the West, wield greatest infrastructural power. More recently established, poorer states in the Global South or in the historic frontier zones or peripheries of historic empires (Anderson, 2008; O'Dowd, 2012) are less capable of regulating urban conflicts. Weak 'infrastructural power' can be a recipe for generating fear, insecurity, and uncertainty in cities which, in turn, become factories of irregular forms of pro- and antistate violence ranging from criminal activity to organized conflict rooted in ethnic, racial, national, or religious antagonisms (www.conflictincities.org; Bollens, 2007). Such cities risk becoming 'decivilized' by new and repressive forms of communitarianism, and routinized informal violence in everyday life.

An alternative understanding of the state–city interface finds in the very vulnerability and complexity of large cities, antidotes or capacities to combat or decompose organized political violence over the state. Here the totalizing and polarizing thrust of much state-related violence finds an antidote in the differentiating capacities of cities, their shared institutions of civil society, innovative forms of economic development, creativity, and mutual interdependence, and their shared rules for governing relationships between culturally diverse populations (Varshney, 2002; Sassen, 2010). Cities in this view have the potential to be sites of regulated public contest between multiple groups struggling for power, recognition, inclusion, and self-affirmation. 'Open cities', lodged in complex global interurban interconnections (economic, political, and cultural) also have the potential to help create a transnational civil society, which promises to moderate, if not eliminate, wars over bordered national states.

Capitalist Dynamics and Urban Conflicts

A second strand of research emphasizes the rise of neoliberal capitalism globally and the way in which it has diminished the role of states in shaping urban conflicts since the 1980s. The state–city interface is now seen to be no longer so important as commodification, privatization, and deregulation reduce the capacity of states to provide urban social services or mitigate enhanced socioeconomic inequalities. These trends are most apparent in the weaker, more fragile states of the Global South,

faced by unprecedented urban growth. As cities are reordered according to the logic of the neoliberal market, states are less likely to be a focus for individual or collective demands for citizenship rights. Urban conflicts become more fragmented and differentiated in everyday struggles over housing, clean water, schools, policing, physical safety, and the right to use the streets to make a livelihood (Bayat and Biekart, 2009). State accountability becomes an ever more remote prospect in urban contexts formed by a kaleidoscope of corporate interests, corrupt states, pro- and antistate paramilitaries, rival gangs, and informal urban settlements sustained by subsistence economies.

At one level, then, cities are conceptualized as key drivers of economic growth and urbanization, i.e., as central to the operations of corporate capitalism, the rise of megacities and the restructuring, segmentation, or hollowing out the power of national states (Taylor, 2004; Harvey, 2012; Sassen, 2012 – for a critique of this perspective, see Therborn, 2011). At another level, these same cities are the sites of hugely diverse, and often violent, urban conflicts rooted in ever-widening social disparities. Seen through this lens, urban conflicts are fundamentally capitalist conflicts and, directly or indirectly, manifestations of class struggle.

Harvey (2012: 5) argues that contemporary cities have arisen through the geographical and social concentration of a surplus product and that capitalism needs urbanization to absorb the surplus product that it continually produces. The control over the disbursement of surplus, including the ownership and control of cities, typically lies in few hands (Harvey, 2008: 24). Capitalist crisis polarizes wealth further and produces cities that are kept under constant surveillance in gated communities and privatized public spaces (2012: 15). Drawing on Lefebvre, Harvey (2012: 25) recognizes that there are a multitude of diverse struggles over 'the right to city' but he goes beyond the theorists of agonistic pluralism to hold out the prospect of an emergent unity that would pit the 'People' against the 'Party of Wall Street' (2012: 164). For Harvey, gaining control of "the inner connection between urbanization and surplus production and use" (2012: 25), can contribute to the overthrow of contemporary finance capitalism.

From this perspective, the significant urban conflicts are viewed through the lens of class struggle over property, human rights, and other urban resources. This allows for a trans-historical and transcultural approach, which can applied, for example, to the evolution of the Western capitalist city, the colonial city, and the rapidly developing cities of the Global South in the twenty-first century.

There is a substantial comparative and historical literature on urban riots that explores the complex ways in which class divisions interact with other social divisions – how economic inequality fuses with other factors ethnic, racial, gender, and political divisions to fuel periodic, if recurring, violent conflict (e.g., Gilje, 1996; Arrom and Ortoli, 1996; Brass, 1996). Some of this comparative research also addresses state–city relations, notably the issue of policing (Waddington et al., 2009) and the existence of urban-level 'riot systems' (Brass, 1996) while at the same time, recognizing the way in which these issues are framed by a wider economic system that generates social inequality and exclusion.

Writers, such as Wacquant (2008), concentrating on Western Europe and North America, study the scope, spread, and intensity of marginality or socioeconomic seclusion in the city. Here too capitalism shapes the state, but the latter is a more coherent force than its counterpart in the Global South. Wacquant (2008) contrasts the experience of the American city, where ethnicity is amplified by the state and trumps class, with the Western European city where 'class' (and the class-structured state) in the key determinant of social segregation and marginalization as in the multinational French banlieue riots of 2005 (see also, SSRC, http://riotsfrance.ssrc.org). US literature on urban riots, for example, tends to concentrate on the nexus in the inner city between race, ethnicity, and socio-economic inequality and the responses of the police and legal system. In the French suburban riots of 2005, Wacquant (2008) portrays the state's role as more direct as it has been shaped by historical forms of institutionalized forms of class conflict that have forced it to mitigate urban inequality through universalist social provision or targeted interventions. Wacquant here adopts a Bourdieusian or Weberian approach to the city that sees it as site where various forms of capital – economic, cultural, social, and symbolic – are accumulated in ways that are shaped by the state. For him, the struggle to claim or protect these forms of urban capital is at the root of urban conflict.

There is a huge and varied tradition of empirical research on urban conflicts that does not link them directly to class struggle in the strictly Marxist sense. For these studies, capitalist dynamics are mediated through a complex of specific local, regional, and historical structures of power, domination, and resistance. They conceptualize the city more as a dependent rather than an independent variable in its own right (see framework three below). The focus is on how imperatives of commodification working through urban planning, tourism, property development, gentrification, inward investment, and big sports projects dramatically displace and relocate disadvantaged populations (including migrants), the homeless, the poor, and the disabled from desirable urban sites (e.g., Marcuse and Van Kempen, 2002).

Infrastructure may be conceived as a material manifestation of urban conflicts typically resolved in favor of the powerful (Rodgers and O'Neill, 2012) or resisted (sometimes temporarily) by the disadvantaged and the poor. Thus, the gated communities, the wealthy suburbs, the banal, and homogenized urban public spaces are one side of a process driven by a desire on the part of elites and middle classes to distance themselves from direct social contact with the 'dangerous' classes because of the latter's proclivities to criminality, disease and violence. The deep and multiple social inequalities that flow from the capitalist system may be experienced differently in working class Parisian suburbs than in the favelas and shantytowns of the new megacities, however. Local and regional differences remain significant in shaping the form of urban conflicts. For example, there are no counterparts in Middle Eastern cities to the lethal drug gang wars in Mexican and other Latin American cities (Rodgers, 2009; Molloy, 2013). Similarly, the conflicting urban fiefdoms of Beirut or the struggle between fundamentalists and secularists in Cairo have their own regional specificities. Nevertheless, the long history and contemporary prevalence of urban riots, protests, racial, ethnic, and sectarian conflicts, gang warfare and criminality may be interpreted as struggles over unequal economic, political, and cultural resources.

Two of the major theorists of urban conflicts, David Harvey (2012) and Manuel Castells (2012) provide somewhat different perspectives on these conflicts although both recognize the ubiquitous influence of neoliberal capitalism. The former emphasizes the primacy of 'class struggle', however, where the latter stresses power inequalities. Harvey continues to emphasize the direct link between urban protests and the contemporary crisis of finance capitalism – assimilating them under the banner of the collective 'right to the city'. Castells' (1983, 2012) work, on the other hand, has moved away from his early structuralist analysis of the capitalist system and the specificity of the 'urban' (Castells, 1977), to the comparative study of urban social movements. He conceives these as grassroots phenomena, that are spontaneous, often temporary or transient, and prone to co-option or disappearance. He argues that they continue to challenge the multiple structures and forms of power and domination to be found in cities across the globe. For Castells, however, these urban conflicts are less coherent and more unpredictable, than in Harvey's analysis that links them rather more directly to the nexus between urbanization and capital accumulation.

The 'City' as a Source of Urban Conflicts

The third research strand draws on theories of the 'urban' per se, where the city itself is treated as a causal variable in generating and shaping distinctive patterns of human behavior and culture (Wirth, 1938; Jacobs, 1961). This approach has been typically associated with periods where cities develop highly visible characteristics that sharply distinguish them from their immediate hinterlands (or the rest of state territory) – for example, Chicago in the early twentieth century, the colonial city at the zenith of Western imperialism and colonialism, and more recently, the rapidly expanding cities in 'developing' countries. Within this framework, urban conflicts are seen to arise directly from the sociospatial and physical environment. Alternatively, anthropologists ascribe to cities a mythology, charisma, or soul, which help to constitute their particular urban conflicts (e.g., Hansen and Verkaaik, 2009).

While these perspectives accord cities an integrity and distinctiveness of their own that is not reducible to political economy, they also acknowledge the impact of the state and economy on urban conflicts. The association of cities with conflict, violence, disorder, poverty, and a range of other social problems is rooted in the Western intellectual response to the growth of the industrial city. Moral outrage over the ills and evils of urban life have typically shaded from ambivalence into outright antiurbanism (White and White, 1962; Williams, 1973) in the context the portrayal of social change as framed by rural–urban, country–city, and gemeinschaft–gesellschaft distinctions. Angotti's (2006) critique of Davis (2006) has trenchantly argued that this tradition continues to survive in contemporary Western views of the megacities of the Global South. These analyses tend to see cities as 'things' that produce all manner of urban ills (including mass poverty, violence, and disorder) rather than as complex sets of social relationships, which provide multiple opportunities for progressive, grassroots urban movements to oppose oppression and create democratic institutions.

The most systematic and empirically grounded example of the tradition that sees the city as an independent variable in its own right remains the work of the Chicago School of Sociology (1890–1940). This culminated in Wirth's (1938) famous theoretical synthesis. Wirth proposed a universalist definition of the city as a "relatively large, dense, and permanent settlement(s) of social heterogenous individuals" (1938: 8) that produced varied forms of social disorganization, conflict and 'deviant behavior'. Although much criticized for their reliance on biological/ecological metaphors and its conservative and ethnocentric biases (e.g., Rodgers, 2010), the influence of the Chicago survives in many social scientific and planning approaches to the city.

The 'Chicago' approach also pioneered the practice of according 'exemplar' or 'paradigmatic' status to key cities (Beauregard, 2003), because they are deemed to be stereotypical or generic, archetypal, or prototypical – a trend setter (Brenner, 2003: 208). This approach is an attempt to impose a degree of analytical manageability and coherence on the huge population of highly diverse historical and contemporary cities. Typologies (either explicit or implicit) are thereby generated while a small number of key exemplars of each type get intensive research scrutiny. Contemporary examples include the 'global city', the capitalist city, the industrial city, the '(post)modern city', the 'fundamentalist city', the 'postsecular city', 'contested cities', and 'cities of the Global South/North'. These conceptualizations influence the interpretation of urban conflicts deemed to be characteristic of each type.

Anthropologists typically rely on 'cultural' analyses to provide more holistic interpretations of cities and their complex conflicts. For example, Hansen and Verkaaik (2009) employ the concept of urban charisma as embedded in people, things, places, and situations (2009: 8). They suggest that the urban must be understood as "a way of being in the world, as a dense and complex cultural repertoire of imagination, fear and desire" (2009: 5). As such urbanism is defined differently than in the Chicago School – cities are not reducible to their functions or the dynamics of their social networks rather they should be understood at several levels as (1) sensory regimes; (2) specific forms of knowledge and intelligibility; and (3) specific forms of power, connectivity, and possibility. The key charismatic urban individuals or urban specialists – big men, hustlers, brokers, even criminals who have extralocal links to centers of power and who link informal urban life to its formal institutions. They register, facilitate and mediate the constant conflicts over urban space between rich and poor, planners and pavement dwellers (2009: 11). In the process, they give cities a form of unity in the face of their multiple conflicts and fissiparous tendencies.

Summary

The research traditions identified above often overlap in practice yet they reveal three important analytical options in the search for the source of urban conflicts. Whether the source is located in state–city relationships, the global capitalist economy, or the city as a distinctive material and social phenomenon, has far-reaching implications for understanding the substance and

potential significance of urban politics. The examples discussed under each heading are merely illustrative of a vast and complex field. While much of this literature has its provenance in the 'West', there is an overall tendency across the three approaches to encompass the study of urban conflicts within an overarching global framework. Nevertheless, the three approaches identified above do not embrace all the relevant literature on urban conflicts notably that which accords more significance to individuals. Gupte (2012), for example, argues that collectivist terms such as ethnic and communal violence might be replaced by the term 'civil violence' which accommodates a more microperspective that better accommodates the agency of individuals. Karl-Dieter Opp (2009) also advances a more capacious framework that synthesizes a range of disciplinary perspectives on political protest and social movements while advancing a structural-cognitive model of his own. This framework seeks to make explicit links between macrostructural perspectives and microperspectives focused on why people participate in different forms of social protest. However, Opp does not link protests and social movements explicitly and systematically to cities or the urban environment.

Urban conflicts have an increasing transnational and transcultural resonance. Despite their vast differences, cities in the Internet age are now embedded in global networks, which make the performance of some urban conflicts iconic and instantly communicable. These conflicts also highlight shifting relationships between violent and nonviolent strategies. Wars are increasingly urban. Cities now rival states as arenas and stakes in political conflict. Struggles between rich and poor, Al-Qaeda bombing campaigns, the 'Arab Spring' protests, the 'occupy' movements, drugs wars in Latin America or sectarian violence in Asia and Africa all demand specific historical analysis. However, they are also reshaping cities in recognizable patterns that demand more sustained comparative study.

See also: Agonism; Conflict and Consensus; Ethnic Conflicts; Nation-State and War; Territorial Conflicts; Violence, History of; Violence: Public.

Bibliography

Agostini, G., Chianese, F., French, W., Sandhu, A., 2010. Understanding the Processes of Urban Violence: An Analytical Framework. http://www.eisf.eu/resources/item/?d=41760 (accessed 24.10.13.).

Anderson, J., 2008. From Empires to Ethno-national Conflicts: A Framework for Studying 'Divided Cities' in 'Contested States'. Divided Cities/Contested States Working Paper No.1. www.conflictincities.org (accessed 17.10.13.).

Angotti, T., 2006. Review essay. Apolcalyptic anti-urbanism: Mike Davis and his planet of slums. International Journal of Urban and Regional Research 30 (4), 961–967.

Arrom, S.M., Ortoli, S. (Eds.), 1996. Riots in Cities: Popular Politics and the Urban Poor in Latin America 1765–1910. Scholarly Resources, Wilmington, DE.

Bayat, A., Biekart, K., 2009. Cities of extremes. Development and Change 40 (5), 815–825.

Beall, J., Fox, S., 2009. Cities and Development. Routledge, London.

Beall, J., Goodfellow, T., Rodgers, D., 2013. Cities and conflict in fragile states in the developing world. Urban Studies, 1–19.

Beauregard, R., 2003. City of superlatives. City and Community 2 (3), 183–199.

Bollens, S., 2007. Cities, Nationalism and Democratization. Routledge, London.

Brass, P.R. (Ed.), 1996. Riots and Pogroms. New York University Press, New York.

Brenner, N., 2003. Stereotypes, archetypes and prototypes: three uses of superlatives in contemporary urban studies. City and Community 2 (3), 205–216.

Calame, J., Charlesworth, E., 2009. Divided Cities: Belfast, Beirut, Jerusalem, Mostar and Nicosia. University of Pennsylvania Press, Philadelphia, PA.

Castells, M., 1977. The Urban Question: A Marxist Approach. MIT Press, Cambridge.

Castells, M., 1983. The City and the Grassroots: A Cross-cultural Theory of Urban Social Movements. Edward Arnold, London.

Castells, M., 2012. Networks of Outrage and Hope: Social Movements in an Internet Age. Polity Press, Cambridge.

Coward, M., 2009. Urbicide: The Politics of Urban Destruction. Routledge, Oxon and New York.

Davis, M., 2006. Planet of the Slums. Verso, London.

Davis, D., Libertun de Duren, N. (Eds.), 2011. Cities and Sovereignty. Indiana University Press, Bloomington.

Fearon, J.D., Laitin, D.D., 2003. Ethnicity, insurgency and civil war. American Political Science Review 97 (1), 75–90.

Gilje, P.A., 1996. Rioting in America. Indiana University Press, Bloomington.

Graham, S., 2010. Cities under Siege: The New Military Urbanism. Verso, London.

Gregory, D., 2010. Seeing red: Bagdad and the event-ful city. Political Geography 29 (5), 266–279.

Gupte, J., 2012. What's Civil about Inter-group Violence? Five Inadequacies of Communal and Ethnic Constructs of Urban Riots. MICROCON Research Working Paper 62. MICROCON, Brighton.

Hansen, T.B., Verkaaik, O., 2009. Introduction: urban charisma: on everyday mythologies in the city. Critique of Anthrop 29 (5), 5–26.

Harvey, D., 2008. The right to the city. New Left Review 53, 23–40.

Harvey, D., 2012. Rebel Cities: From the Right to the City to the Urban Revolution. Verso, London.

Heitmeyer, W., Hagan, J. (Eds.), 2003. International Handbook of Violence Research. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Hepburn, A.C., 2004. Contested Cities in the Modern West. Palgrave, Basingstoke.

Jacobs, J., 1961. The Death and Life of Great American Cities: The Failure of Town Planning. Penguin, Harmondsworth.

Keane, J. (Ed.), 2006. Civil Society: Berlin Perspectives. Berghahn Books, Oxford.

Mann, M., 2012. The Sources of Social Power, vols. III and IV. Cambridge University Press, Cambridge.

Marcuse, P., Van Kempen, R. (Eds.), 2002. Of States and Cities: The Partitioning of Urban Space. Oxford University Press, Oxford.

Molloy, M., 21 August 2013. The Mexican undead: toward a new history of the "drugs war" killing fields. Small Wars 9 (8). http://smallwarsjournal.com (accessed 06.04.14.).

Moser, C., Rodgers, D., (Eds.). Understanding the Tipping Point of Urban Conflict: Violence, Cities and Poverty Reduction in the Developing World. Working Paper Series, University of Manchester. http://www.urbantippingpoint.org (accessed 21.10.13.).

Muchembled, R., 2012. A History of Violence (J.Birrell, Trans.). Polity Press, Cambridge.

Muggah, R., 2012. Researching the Urban Dilemma: Urbanization, Poverty and Violence. International Development Research Centre, Ottawa (Chapter 4), 106 pp.

Opp, K.-D., 2009. Theories of Political Protest and Social Movements: A Multidisciplinary: Introduction, Critique and Synthesis. Routledge, New York.

O'Dowd, L., 2012. Contested states, frontiers and cities. In: Wilson, T.M., Donnan, H. (Eds.), A Companion to Border Studies. Wiley-Blackwell, pp. 158–176.

Rodgers, D., 2009. Slum wars of the 21st century: gangs, mano duro, and the new geography of conflict in Central America. Development and Change 40 (5), 949–976.

Rodgers, D., 2010. Urban Violence Is Not (Necessarily) a Way of Life. Working Paper No. 2010/20, UNU-WIDER. World Institute for Development Economics Research, Helsinki, FI.

Rodgers, D., O'Neill, B., 2012. Infrastructural violence: introduction to special issue. Ethnography 13 (4), 401–412.

Sassen, S., 2010. When the city itself becomes a technology of war. Theory, Culture and Society 27 (6), 33–50.

Sassen, S., 2012. Cities in a World Economy, fourth ed. Sage, London.

Savage, K., Muggah, R., 19 Jan 2012. Urban violence and humanitarian action: engaging the fragile city. Journal of Humanitarian Assistance. http://sites.tufts.edu/jha/archives/1524.

Savitch, H., 2008. Cities in a Time of Terror: Space, Territory and Local Resilience. Cleveland University Press, Cleveland, OH.

SSRC (Social Science Research Council), November 2005. Riots in France: civil unrest in the French suburbs. http://riotsfrance.ssrc.org (accessed 03.12.13.).

Taylor, P.J., 2004. World City Network: A Global Urban Analysis. Routledge, London.

Therborn, G., 2011. End of a paradigm: the current crisis and the idea of stateless cities. Environment and Planning A 43, 272–285.

Tilly, C., 1992. Coercion, Capital and European States AD 990–1992. Blackwell, Oxford.

Varshney, A., 2002. Ethnic Conflict and Civic Life: Hindus and Muslims in India. Yale University Press, New Haven, CT.

Wacquant, L., 2008. Urban Outcasts: A Comparative Sociology of Advanced Marginality. Polity, Cambridge.
Waddington, D., Jobard, F., King, M. (Eds.), 2009. Rioting in the UK and France: A Comparative Analysis. Willan Publishing, Cullompton, Devon.
White, M., White, L., 1962. The Intellectual versus the City: From Thomas Jefferson to Frank Lloyd Wright. Harvard University Press and MIT Press, Cambridge, MA.
Williams, R., 1973. The Country and the City. Oxford University Press, Oxford.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44 (1), 1–24.
World Bank, 2010. Violence and the City: Understanding and Supporting Community Responses to Urban Violence. World Bank Social Development Department, Conflict, Crime and Violence Team, Washington, DC.

Urban Design and Public Space

Ali Madanipour, Newcastle University, Newcastle upon Tyne, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public space is a primary ingredient in urban environment; places of encounter where public life unfolds. Following structural changes in economy and society since the 1970s, it became a focus of attention in a shifting division of labor in spatial production, raising anxiety about the role of the state, decline of public services, and privatization of urban environment. It is acknowledged as a vehicle of urban regeneration and sustainable development, an essential part of building competitive and compact cities, based on services and experience economy. However, concerns about social and environmental consequences of globalization, consumerism, urban spectacle, and gentrification remain.

Public space has increasingly become a subject of inquiry for different disciplines and professions (Carmona et al., 2008; De Souza et al., 2012; Hou, 2010; Low and Smith, 2006; Orum and Zachary, 2010; Parkinson, 2012; Sadeh, 2010; Watson, 2006). How can this rising interest be explained and why is there such a new emphasis on a subject as old as the city itself? Public space is one of the primary ingredients that make up a city, ingredients that in one form or another have continually been part of the vocabulary of urbanism. A city is an agglomeration of people, where developing a common infrastructure of institutions and spaces is an essential prerequisite for making collective life possible. As the remains of the earliest cities show, human settlements have always been formed by an accumulation of household territories linked together through common spaces such as streets, markets, and temples (Benevolo, 1980; Morris, 1994). By now, therefore, after centuries of making and remaking cities, we should have worked out how to develop and maintain our common infrastructure; so why should there be intensified attention paid to this subject? In this article, I will examine the theme of public space, and urban design as the process of shaping urban space, in the context of the major social and economic changes that have transformed the now predominantly urban world. Investigating the organization of space in ancient ruins gives the archaeologists some understanding of how society was organized in these lost cities. If we study our contemporary urban spaces, what can we learn about our society today? If public space is a subject of renewed interest, should it be seen as a necessary ingredient in facilitating and supporting urban change? Or might it indicate an anxiety about change, rendering us no longer able to travel on a well-trodden path? What are the possible causes and consequences of these changes for the urban society?

Anxiety for Democratic Urbanism

The concern about public space is partly rooted in the structural changes in society that endangered the provision and maintenance of public services (Madanipour, 2011). The 1970s marked the end of a model of economic development based on a tight regulation of the market by the state, and of the direct involvement of the state in the production of the built environment. During the nineteenth and twentieth centuries, urbanization went hand in hand with industrialization, bringing people into cities at ever larger numbers and faster rates, resulting in a continually unstable combination of high productive capacities and poor living conditions. The global economic crisis of the 1930s, social upheavals in Europe, and the Second World War had encouraged western governments to take a more active role in the economy, stimulating demand through a better distribution of resources and improving the conditions of life in cities. Public investment in essential public services such as health, education, transport, and housing increased through direct involvement by the state.

The planning and design ideas that had been developed in an earlier period, from Ebenezer Howard's garden cities (Howard, 1965) to Le Corbusier's city of future (Le Corbusier, 1987), were now employed to construct houses, roads, schools, hospitals, and new towns. Modernism and garden cities movement became the twin design ideas that shaped the twentieth century cities (Jacobs and Appleyard, 1987). Le Corbusier had called for 'architecture or revolution' and Howard's book was subtitled, 'the real path to reform.' Industrial technology and functionalist design were used in the mass production of the built environment, justified by social concerns. These ideas, in turn, were routed in a long history of social reform and utopian thinking that had originated from the Renaissance ideal cities and continued up to the urban improvements and the model towns of the eighteenth and nineteenth centuries. While these ideas had been implemented in small-scale experiments, they were now implemented systematically through comprehensive planning. The twentieth century was a century of organizations (Aglietta, 2000), as partly reflected in the complexity and expansion of the state and the emergence of a mass society, the manifestation of which could be visibly seen in the development and transformation of the built environment. Provision of public open space was one of the key demands of these changes, as indicated by the Charter of Athens, demanding for new breathing spaces in the overcrowded and badly built cities (Sert, 1944).

This model of development, however, ran out of steam by the 1970s, which became a turning point; the following decade witnessed the start of a new phase, in which the Keynesian accord between the state and the market was broken. The role of the state in urban development changed in accordance with

epochal shifts in the economy. Technological innovation in information, communication, and transport transformed working practices, manufacturing industries were relocated to lower cost regions of the world, and the global map of economic production was redrawn. Structural change eroded the state's ability or willingness to invest in, and the urban populations' strength to demand for, public services. The size and scope of the state was reduced, so as to open up new spaces for the market to rejuvenate. State institutions now saw their role as enablers rather than producers, paving the way for the market rather than directly engaging in urban development. The midcentury urban renewal projects had not only failed to solve the social problems, they had now become a new social problem themselves, demanding urgent attention. The accelerated process of deindustrialization, casualization of the labor force, international immigration, and dwindling public budgets threatened the provision and maintenance of public services. Urban development projects were no longer planned and implemented by the public sector, which was now seen to be bureaucratic, clumsy, and ineffective. The end of the social democratic phase led to a period of neglect for public spaces and public institutions.

The new laissez-faire phase that started in the 1980s transformed the organization of spatial production. The rundown industrial sites, and later city centers and public housing schemes, were opened up to a new model of development, in which private companies were supported by public subsidies to take on the role of urban development and regeneration. Private firms, however, had a limited remit, obliged to provide a return on the investment of their shareholders, rather than delivering services for the general public. If public authorities were not making or maintaining public spaces, how could private firms be expected to do so? This question opened up a crisis of confidence and a rising sense of anxiety about how, to what extent, and by whom, should public services be provided. For their urban development projects to function, private developers needed some form of intermediate space, providing access to properties, as well as representing the development as a whole to the possible customers. But developers desired to control these spaces, especially if they were on private land, so as to ensure the comfort and protection of the users and to safeguard their own investments. The regeneration of cities after industrial decline often meant investing in parcels of land surrounded by a sea of decline, where new upmarket areas could be transplanted in the middle of a deprived district, a disconnection that could find social manifestation, such as resentment, anxiety, and fear, and spatial forms such as highly guarded, walled, and gated areas. In an increasingly unequal society, the copresence of the two ends of the socioeconomic spectrum could be interpreted as undesirable and even dangerous. Public space was increasingly associated with homelessness and crime, to which the private reaction was fearful withdrawal, and the policy response was forceful conquest, both with negative social consequences. Privatization of public space could shield the higher income users of these spaces and keep the lower income groups and undesirable activities out, while gentrification and heavy-handed policing would displace them. This was a crisis for public space, and by extension a crisis for the city as a whole.

The campaign against gentrification and the privatization of public space, therefore, had a wider anxiety about the loss of the idea of the city, and the disabling impact of social inequality and exclusion on democratic practices. The western liberal democratic tradition had long encouraged the development of an active public sphere, in which citizens were able to participate, communicate freely, and develop opinions about the affairs of their society, which would enable them to make informed political decisions in democratic governance. The heritage of the enlightenment was based on the idea of generating and sharing information and knowledge, as exemplified by the British coffee shops, French salons, print newspapers and pamphlets, collective efforts in writing dictionaries and encyclopedias, all part of the emerging democratic nation states. Major theorists of public sphere, such as Habermas (1989) and Arendt (1998), attempted to show how this public sphere was an essential ingredient of a well-functioning democratic life. If the public sphere was controlled by the private corporate world of the property developer and the media tycoon, how could such democratic traditions be maintained and improved? Anxiety about public space was an anxiety for the urban society and for democracy. The loss of access to public space meant a decline of democracy, as the primary condition for a place or activity to be public is access (Madanipour, 2010).

The Challenge of Diversity

With its roots in the Latin word for people, the word 'public' indicates people as a whole, which is often defined in contrast to the private sphere, which is the sphere of individuals and households. The liberal political theory has argued for maintaining a clear divide between the public and the private spheres, so as to avoid corruption (Nolan, 1995; Wacks, 1993). The interplay of concealment and exposure, i.e., some control over what is revealed to others and what is kept hidden from them, remains an important dynamic in social and political life (Nagel, 1998). Encroachment of the public sphere into the private sphere would result in the loss of privacy and individual freedom, while the encroachment of the private sphere into the public sphere may create individual gain and collective loss. There are, however, difficulties with thinking about people as a uniform whole, as any society is made up of a range of different types of people, and ultimately of a multitude of individuals each with a different life story and trajectory. As the campaigners for different minority or disadvantaged groups have argued, the notion of the public in law and politics has historically been based on a narrow basis, ignoring the diversity of the population to whom the law applies. The public sphere is made up of many social and spatial ingredients, which include the public spaces of towns and cities. Many cities have historically emerged at the meeting points of major roads, and public spaces of cities, which took shape around these intersections, may be broadly defined as crossroads, as places of encounter where the life of the city unfolds. It is here that diverse populations may meet, their trajectories collide or coincide, and where the different spheres of political action, sociability, economic exchange, and cultural performance may play out through instrumental and expressive means (Madanipour, 2003).

Anxiety over public space is not only about sociability, equality, and democracy; it is also about the loss of community, which is a long-term problem of modernity. The populations of the small agrarian towns had developed tight cohesive interrelationships, which were shattered as these populations were uprooted from their towns and villages and brought by force or encouragement to large cities as paid workers. The early students of urban societies (Engels, 1993; Tönnies, 1957; Simmel, 1950) were horrified about the conditions of these newcomers, interested in the way metropolitan crowds were alienated from one another, anonymous and atomized, withdrawing into a small mental space that would allow them a degree of psychological safety. Many romantics, revolutionaries, and reformers who observed the life of large cities recommended a return to the countryside, even dismantling large cities, by force if needed, or at least the reconstruction of smaller communities in which a sense of cohesive relationships could be developed. Others advocated the creation of new religious or secular bonds inside cities to help develop new communities.

In these communitarian ideals, public space played a significant role, as a place in which people of a small community could be linked to one another (Taylor, 1995; Etzioni, 1995). The possibility of recreating such communities, of course, is in serious doubt, as the forces of mobility, diversity, individualization, and metropolitan growth are not showing any slowdown. The design ideas that were used in this laissez-faire stage were all a rejection of the modernist functionalism, advocating playful juxtapositions or a return to the premodern city. Communitarian ideas have been a strong influence on these design directions. A return to the idea of garden cities, as shown in new urbanism and urban villages, has often been based on the centrality of public space as a cohesive force in spatial and social organization. In practice, however, it can only help create a semblance of a community, an appearance that may or may not in the long run turn into a cohesive social phenomenon. In a mobile society, the possibility of developing long term social bonds is reduced, and the image rather than the practice of cohesive urban life may be what might be possible.

Public space, in its physical form, no longer plays the crucial role that it had once had in urban life. In the smaller cities of agrarian societies, the public space was simultaneously a place of sociability, trade, religion, and politics. The central public spaces of the city were places of high theater of power by the religious and political elite of the day, placing their highest towers, most magnificent buildings, most elaborate works of art, their most impressive statues and fountains, so as to produce a social order through symbolic means. However, the growth of the large industrial city, the transformation of the cultural norms, the breakup of the social bonds and the atomization of urban populations, the mediated and mass forms of communication, and the complexity of political and economic processes all splintered the public events into many different locations. The symbolic and the functional values of the public space were taken over by the high rise corporate offices in the new cities, indicating the cathedrals of the newly powerful. The symbols that were inherited from the past were transferred to an aesthetic sphere, but the public space kept some of its earlier values, as evidenced by political demonstrations and campaigns as well as a place of everyday sociability and intermingling. What constituted public space was the possibility of copresence and face-to-face encounter, which remains the most potent form of social interaction. Much hope has been placed in the communicative possibilities of the information and communication technologies, which have created the possibility of forming new social networks and new spaces of possibility across the world. These technologies, however, are essentially libertarian, allowing individuals to keep a distance from others and cut their links as easily as setting up those links. The strength and potential durability of social connections that face-to-face encounters could bring about were exposed to further pressures of contingency and ephemerality.

A Shifting Division of Labor

The campaign for public space, which started to grow from the 1980s onward, was a central theme of urban design. In its broadest sense, urban design is the process of shaping cities. It plays a vital part in the production of the built environment, with significant political, economic, and cultural dimensions, going beyond appearances, limited scope and scale, and mere physical transformation. In the common use of the term, design is sometimes limited to form and appearance, visual expression and aesthetic experience, and so restricted to beautiful objects and luxury lifestyles, branding and superficial treatment. Design, however, goes beyond appearances; it also addresses the function and use of objects, as well as providing instructions for making them. It reorganizes space and signifies cultural meaning, political performance, and economic value. Design in its modern sense was seen as a sign of applying order and rationality, in contrast to the accidental and piecemeal development of medieval cities (Descartes, 1968). Design, therefore, is a plan, purpose or intention, articulated by visual or other means, envisaging a finished outcome and instructing how it should be made. It involves "the playful creation and strict evaluation of the possible forms of something, including how it is to be made" (Lynch, 1981: p. 290). It deploys creativity, as 'the generation of new ideas,' toward innovation as 'a specific end' (Cox, 2005: p. 2), as well as offering a code with which the contemporary society can be understood (Sudjic, 2009: p. 49). When it is applied in shaping cities, therefore, the role of design is wide ranging in scale and scope. It goes beyond the narrow scope of townscape beautification or the limited scale of the spaces between buildings, to engage in transforming significant parts of urban areas or even the design of entire new towns and cities. This purposeful process of shaping cities is not limited to physical transformation, but also includes a wide range of economic, political, and cultural considerations.

The roots of urban design go far back to the dawn of urban living 10 millennia ago, as living in towns and villages has always required thinking about how to build and transform them in such a way as to support life of a large number of households living together in a limited area. The remains of the earliest cities are the evidence of this purposeful thinking: relating one building to another, moving from one place to another, providing places of significance, and relating the town to its surroundings, all reflecting and consolidating social structures and habits in a human agglomeration. Urban

design in its current form, meanwhile, emerged as a response to the changing division of labor in spatial production. As the state withdrew from urban development, urban planning lost much of its remit and ability to envisage the future shape of the city and to shape the urban environment through forward planning. Envisaging the future of the city in an uncertain world seemed like a futile enterprise, and the previous generation who had engaged in this visioning was condemned as utopian.

As planning moved toward policy development and implementation, it severed its ties with architecture and abandoned the development of design skills. Rooted in the modernist tradition, these skills were thought to be merely involved in aesthetic considerations, rather than rational calculation, which increasingly seemed to be possible by computers. Land-use planning, therefore, became limited in responding to the initiatives of the market, rather than initiating change on behalf of the local authorities. Architects, on the other hand, lost their sense of confidence in redesigning the city, as the postwar experiments were seen to have failed and as the public and private sectors were renegotiating their relative roles and powers. Architects were now primarily part of the private developer's team, limited to the confines of the site and the interests of the client, rather than the wider problems of the city. The social rhetoric of modernism had already been abandoned once its ideas were taken over by the large private corporations, but it had survived in the public urban development schemes, which were now lost as a result of structural change. Urban design emerged to fill a gap, providing the necessary skills and knowledge to design and transform parts of cities. The new pattern of urban development involved a number of stakeholders from the public and private sectors involved in potentially fragmented schemes, which now needed direction and coordination, delivered through master plans for large-scale projects and emphasis on public space as an integrative force, linking various stakeholders together and mobilizing them around a shared vision. After several phases of change, planning eventually embraced a strategic spatial form, which would allow it to focus on a particular area of the city, as a pragmatic use of the limited resources focusing on the most troubled parts of the city with the highest potential for economic revival. It was through urban design master plans and strategic spatial plans that urban design and urban planning came closer to one another once again. Large-scale urban projects became the vehicle of mobilizing resources for urban regeneration (Lecroart, 2007).

While the technical division of labor has been shifting to find a proper deployment of the necessary knowledge and skills, there has been a campaign for intensifying participatory practices, which would compensate for the democratic deficit of this technocratic effort. Public participation, in a sense, is the process through which the production of space can become inclusive, as a counterweight to the exclusionary social pressures and the narrow focus of the economic agendas (Madanipour, 2010). If genuinely open, it has the potential to offer a better chance of delivering inclusive public spaces. The question, however, often revolves around the extent of participation: is it led and framed by the elite to legitimize their decisions or is there a genuine space for coproduction in the process?

Economic and Environmental Rationale

The initial stages of economic and political restructuring coincided with rising anxieties. These stages were followed by a new wave of optimism, as public space was eventually placed on the agenda of the urban development. The early shocks of deindustrialization and radical liberalizations were partly moderated by the idea of sustainable development, with its blend of social, economic, and environmental concerns. From international organizations such as the European Union down to national governments and local authorities seemed to accept the necessity of investing in public space, and many municipalities around the world have developed programmes of public space development and improvement. The division of labor between the public and private sectors have been continually shifting, always leading to controversial outcomes when substantial parts of the public sector roles and duties were taken over by the private sector. Now that housing and intervention in the private property was no longer a public sector remit, its shift of focus on public space could help it find legitimacy and visibility, recovering a spatial platform in the urban development process. This new attention is often associated with economic, social, and environmental justifications.

In the globalized economies, cities are the engines of economic activity, as evidenced by the restless growth of urban populations around the world, changing the demographic balance in favor of cities for the first time in history. As the two State of European Cities reports show, larger European cities show a trend of growth in population and in the accumulation of wealth, while the middle sized cities have stagnated and peripheral regions decline (ECOTEC, 2007; RWI et al., 2010), an uneven pattern of development which has been exacerbated by the global financial crisis. The competitiveness of cities, therefore, has become a key consideration for policy makers, hoping to ensure this capacity through innovation and attractiveness, both with significant roles for public space.

The European Union's 10-year strategy, EU 2020, advocated 'smart growth,' which was "an economy based on knowledge and innovation" (EC, 2010a: p. 5). The transition out of labor-intensive manufacturing industries was to lead into knowledge-intensive services and higher levels of added value. Innovation, therefore, lay at the heart of the new economic activities, in science and technology as well as in cultural and creative industries, which were involved in utilizing the information and communication technologies and high skilled workers in developing new products, encouraging consumption through the 'experience economy' and setting and amplifying social and cultural trends that could stimulate consumer demand (EC, 2010b: p. 17). Innovation drew on the meeting of minds, the social encounters that would facilitate collaboration and exploration, partly made possible through a vibrant urban environment.

Attractiveness, meanwhile, was another way that urban public space was envisaged to contribute to economic competitiveness. Cities found a decisive importance in the EU Lisbon Strategy, which advocated making Europe "a more attractive place to work and invest," and as most Europeans lived in urban areas, the "attractiveness of European cities" was

thought to "enhance their potential for growth and job creation" (EC, 2006: p. 1). In this context, the role of public space becomes recognized by the Leipzig Charter, signed by European ministers of development in 2007, who saw public spaces as "soft locational factors," which are "important for attracting knowledge industry businesses, a qualified and creative workforce and for tourism" (EC, 2007: p. 3). Economic competitiveness, therefore, was thought to be achieved through innovation, which in turn was aided by an attractive and vibrant urban environment, a crucible in which new ideas and practices could be developed through intensified communication and exchange. The idea of the death of the city, which was once imagined as inevitable (Martindale, 1966), has become only a fading memory. Rather than disappearing under the pressures of mobility and communication technologies, cities have grown larger, reflecting and intensifying the need for copresence as an economic force. In addition to economic value, public spaces were also acknowledged to have environmental significance.

Cities have major responsibilities in both creating and finding ways of addressing the problem of climate change. As the Aalborg Charter, signed at a meeting of European towns and cities in 1994, admits, "We understand that our present urban lifestyle … make us essentially responsible for many environmental problems humankind is facing" (EC, 1994: p. 1). The lack of open space, the Charter maintained, was one of the core environmental problems. Search for individual comfort and quality of domestic space had led to urban sprawl, which had adverse effects on society and environment, consuming higher rates of energy and land, generating higher levels of traffic, air, and noise pollution (EEA, 2009). Compact cities, with their mixed land use and high density and reliance on public transport, have been identified as a way of reducing energy consumption and protecting the environment from encroachment by the city. Living at high-density compact urban environments, however, needs to be facilitated and supported by the provision of high quality public open space. Public open space, therefore, becomes a central theme in the policy documents that advocate sustainable development (EC, 2007; EC, 2010c). The vision of a sustainable city envisages "safe areas, green, and other public spaces, as well as … short distances to facilities and services," which would make them "sufficiently attractive to counter urban sprawl" (EEA, 2009: p. 40). As the European Environment Agency puts it, "Urban areas need to provide for their citizens the foundations for choices leading toward more sustainable lifestyles, such as affordable housing in more compact urban areas that provide high quality public spaces and a healthy environment" (EEA, 2009: p. 102).

Anxiety over the privatization of public space and the delivery of public services triggered the initial phase of campaign for public space. Now that many national and international organizations have acknowledged the significance of public space in urban life and many municipalities have active programmes of public space improvement, should the old anxieties be put to rest? If the public space agenda is now so strongly supported by the state, have the problems gone away or at least a path is established for dealing with these problems?

From Old to New Anxieties

Some signs indicate that there are grounds for new anxieties, as public space improvement may serve as a pathway to gentrification and consumerism and the capacity for its improvement may be jeopardized by the decline of public finances. The economic phase that started in the 1980s has in turn come to a halt with a new global economic crisis in 2008, which has demanded massive intervention by the state, but this time saving the banks as the infrastructure of the market, rather than a return to direct intervention in the economy, indicating the possibility of deepening the laissez-faire directions, rather than a return to the social democratic phase.

Economic austerity and dwindling public finances may limit the future projects of public space improvement, triggering a new, intensified phase of concern over the provision and maintenance of public services. As the opening words of the European Union's 10-year strategy, EU 2020, put it, "Europe faces a moment of transformation. The crisis has wiped out years of economic and social progress and exposed structural weaknesses in Europe's economy" (EC, 2010a: p. 5). The economic crisis has been so deep that the ability of the governments to maintain their commitments to public service delivery has been stretched to the limit. In this context, the future of public space as an inclusive and accessible place may come under pressure. Austerity, however, is not a universal problem and many cities around the world continue to grow in size and prosperity, and carry on investing in their public spaces. The problem in these cities may lie in the way public services and public space are conceptualized and delivered.

Some have called the new phase in economic production as spatial and immaterial, hinting at the nature of digital streams that can turn into profitable commodities. The globalized economy, however, is based on an intensified production and consumption of material products (Madanipour, 2011); people's lifestyles in wealthy cities have never been more entwined with a larger number of material possessions (Sudjic, 2009). A new global division of labor has enabled the relocation of manufacturing from some cities to others, opening up new productive capacities, which require an expanded capacity for consumption. What has been called a consumerist society is the foundation of the global economic relationships, as without intensified consumption, the global economic relations of intensified production and exchange would collapse. The impact of consumerism on the urban space has been enormous, as economic regeneration in many cities has been made in the form of new shops, restaurants, and spaces of leisure, encouraged through an 'experience economy,' in which the urban environment is envisaged as a spectacle. Public spaces play a central role in the development of spaces of experience and in encouraging consumption, to the extent that they might be even seen in some instances as the preconditions for the development of new consumerist activities in an area. So the question becomes, is this the reason why public space has become a key commitment by the municipal authorities, who think they need to follow the market's lead and see consumption-based activities as the only way of economic revival and political legitimacy?

Globalization and state entrepreneurialism have also encouraged cities to behave as firms, thinking of citizens as customers, and public authorities as managing directors of corporations in competition with other corporations from around the world, attempting to attract investment. The basic assumption is that a city needs to attract the floating resources, which can be captured through a variety of means from mass tourism to new companies. Investment in the quality of public environment is one of the key elements of this strategy, which helps branding places as commodities for sale. It is with the same logic that cities compete for having the tallest buildings, recruiting the most famous architects, and developing the most expressive architectural forms. The question, however, is whether the local investment in public infrastructure is steered toward the needs of these outside investors or it also includes the needs of the local citizens. If one is ignored at the cost of the other, the result would be differential treatment, fragile spaces, uneven development, and an unequal society.

A new anxiety emerges about the impact of public space development on the vulnerable communities in deprived and rundown parts of cities. Is the emerging interest in public space improvement coinciding with an interest in the fate of deprived neighborhoods, or would investment in public spaces be a stepping stone to gentrification of urban areas, pricing out existing communities and replacing them with people who have stronger financial abilities? The positive contribution of investment in public spaces to the improvement in the quality of urban life has to be acknowledged. The question is whether such improvement is beneficial for everyone, or that it serves some at the cost of others. Concern over the lack of public space, which was a symptom of a particular model of development, may therefore turn into a concern over the impact of public space provision on public life. An uneven pattern of development, both within cities and between cities, has been noted even in Europe, where the tradition of the welfare support is still strongest in the world (ECOTEC, 2007; RWI et al., 2010). Added to inequality is the challenge of social diversity, which is fueled by the plurality of lifestyles and the intensification of mobility and migration. Investment in public spaces of the deprived neighborhoods, therefore, may be thought to contribute to social inclusion of the marginalized groups (Madanipour, 2010). Similarly, the environmental imperative requires attention and investment in urban public spaces. The question is whether in the context of new economic conditions the willingness and the ability to invest in these areas will not suffer (Madanipour et al., 2014).

See also: Afrocentric Approaches to Social Work; Eco-cities and Sustainable Urbanism; Gentrification; Globalization and World Culture; Globalization: Geographical Aspects; New Urbanism; Planning Theory: The Good City and Its Governance; Postmodern Urbanism; Urban Internal Spatial Structure; Urban Neoliberalism; Urban Planning (Western), History of.

Bibliography

Aglietta, M., 2000. A Theory of Capitalist Regulation: The US Experience, new ed. Verso, London.
Arendt, H., 1998. The Human Condition, second ed. The University of Chicago Press, Chicago.
Benevolo, L., 1980. The History of the City. Scolar Press, London.
Carmona, M., de Magalhães, C., Hammond, L., 2008. Public Space: The Management Dimension. Routledge, London.
Cox, G., 2005. Cox Review of Creativity in Business: Building on the UK's Strengths. HM Treasury/HMSO, London.
De Souza, E., Silva, A., Frith, J., 2012. Mobile Interfaces in Public Spaces: Locational Privacy, Control, and Urban Sociability. Routledge, London.
Descartes, R., 1968. Discourse on Method and the Meditations. Penguin, London.
EC, 1994. The Aalborg Charter. European Commission, Brussels.
EC, 2006. Thematic Strategy on the Urban Environment. European Commission, Brussels.
EC, 2007. The Leipzig Charter on Sustainable European Cities. European Commission, Brussels.
EC, 2010a. Europe 2020: A Strategy for Smart, Sustainable and Inclusive Growth. European Commission, Brussels.
EC, 2010b. Green Paper: Unlocking the Potential of Cultural and Creative Industries. European Commission, Brussels.
EC, 2010c. Making Our Cities Attractive and Sustainable. European Commission, Brussels.
ECOTEC, 2007. State of European Cities Report. European Commission, Brussels.
EEA, 2009. Ensuring Quality of Life in Europe's Cities and Towns. European Environment Agency, Copenhagen.
Engels, F., 1993. The Condition of the Working Class in England. Oxford University Press, Oxford.
Etzioni, A., 1995. The Spirit of Community: Rights, Responsibilities and the Communitarian Agenda. Fontana Press, London.
Habermas, J., 1989. The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. Polity Press, Cambridge.
Hou, J., 2010. Insurgent Public Space: Guerrilla Urbanism and the Remaking of Contemporary Cities. Routledge, London.
Howard, E., 1965. Garden Cities of Tomorrow. Faber and Faber, London.
Jacobs, A., Appleyard, D., 1987. Towards an urban design manifesto. Journal of American Planners Association, 112–120 (Winter).
Le Corbusier, 1987. The City of Tomorrow and Its Planning. The Architectural Press, London.
Lecroart, P., 2007. From the Plan to the City: Managing Large-Scale Urban Development Projects in Europe, Les Cahiers de l'Institut d'Aménagement et d'Urbanisme de la Région d'Île de France, No. 146, June 2007, pp.110–118.
Low, S., Smith, N., 2006. Politics of Public Space. Routledge, London.
Lynch, K., 1981. Good City Form. MIT Press, Cambridge, MA.
Madanipour, A., 2003. Public and Private Spaces of the City. Routledge, London.
Madanipour, A. (Ed.), 2010. Whose Public Space? International Case Studies in Urban Design and Development. Routledge, London.
Madanipour, A., 2011. Knowledge Economy and the City: Spaces of Knowledge. Routledge, London.
Madanipour, A., Knierbein, S., Degros, A., 2014. Public Space and the Challenges of Urban Transformation in Europe. Routledge, London.
Martindale, D., 1966. Prefatory remarks: the theory of the city. In: Weber, M. (Ed.), The City. The Free Press, New York.
Morris, A.E.J., 1994. History of Urban Form: Before the Industrial Revolution, third ed. Longman, Harlow.
Nagel, T., 1998. Concealment and exposure. In: Philosophy and Public Affairs, vol. 27. (1), pp. 3–30 (Winter 1998).
Nolan, L., 1995. Standards in Public Life, vol. 1. HMSO, London. Report.
Orum, A.M., Zachary, N., 2010. Common Ground? Readings and Reflections on Public Space. Routledge, London.
Parkinson, J., 2012. Democracy and Public Space: The Physical Sites of Democratic Performance. Oxford University Press, Oxford.
RWI, DIFU, NEA and PRAC, 2010. Second State of European Cities Report. European Commission, Brussels.
Sadeh, E., 2010. Politics of Public Space: A Survey. Routledge, London.
Sert, J.L., 1944. Can Our Cities Survive? An ABC of Urban Problems, Their Analysis, Their Solution. Harvard University Press, Cambridge, MA.
Simmel, G., 1950. The metropolis and mental life. In: Wolff, K. (Ed.), The Sociology of Georg Simmel. The Free Press, New York, pp. 409–424.
Sudjic, D., 2009. The Language of Things. Penguin, London.
Taylor, C., 1995. Liberal politics and the public sphere. In: Etzioni, A. (Ed.), New Communitarian Thinking. University Press of Virginia, Charlottesville, pp. 183–217.
Tönnies, F., 1957. Community and Society (Gemeinschaft und Gesellschaft). Harper and Row, New York.
Wacks, R. (Ed.), 1993. Privacy. The International Library of Essays in Law and Legal Theory, vol. 1. Aldershot, Dartmouth.
Watson, S., 2006. City Publics: The (Dis) Enchantments of Urban Encounters. Routledge, London.

Urban Entrepreneurialism

Andrew Wood and Timothy Brock, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urban entrepreneurialism is a concept developed by the geographer David Harvey to characterize a profound change in the period after 1970 in the way capitalist cities are governed. The article defines urban entrepreneurialism, its distinct characteristics, and the range of urban strategies that are associated with the entrepreneurial city. The article examines how the concept has been deployed since its introduction in 1989.

The term 'urban entrepreneurialism' is used to characterize a broad historical shift in the way in which the cities of the capitalist world are governed. The term dates to an annual address by the prominent geographer David Harvey to mark his receipt of the Anders Retzius Medal by the Swedish Society for Anthropology and Geography and was subsequently published in *Geografiska Annaler* (Harvey, 1989). David Harvey has a long and distinguished record of contributions to the discipline of geography and his influence extends well beyond the field (see Castree and Gregory, 2006 for a review of his work). In what follows, the authors first provide a brief summary of Harvey's argument. While the term dates to 1989, and like all concepts it is rooted in the particular circumstances of the time, the ideas encompassed in Harvey's argument have exerted significant influence on subsequent work. The second section of the article provides a review of the various ways in which scholars have sought to illustrate and examine the different aspects of urban entrepreneurialism set out in Harvey's account. The third section of the article examines the position of the concept in relation to subsequent developments in how one might best theorize and investigate the changing political and economic dynamics of the capitalist city. While the concept jostles for position with more recent contributions, the arguments that underpin urban entrepreneurialism continue to have lasting resonance.

Urban Entrepreneurialism Defined

The concept of urban entrepreneurialism is rooted in David Harvey's general interest in the historical relationship between the process of capitalist development and its manifestation in the urban landscape or the 'built environment' (Harvey, 1978, 1985, 2003). Harvey's work has traversed a variety of geographical contexts and scales, although the concept of urban entrepreneurialism is very much rooted in the study of the contemporary capitalist city. His book *Social Justice and City* (Harvey, 1973) marked Harvey's turn to the theoretical framework developed by Marx and by subsequent Marxist-inspired scholars. The analytical strength of urban entrepreneurialism lies in its ability to bridge the powerful insights of a Marxian account of capitalist development with the particular ways in which the place and management of cities have changed over the past 50 years.

Harvey argues that the 1970s and 1980s had been marked by a series of changes in the way in which economic and political interests sought to govern the city. The dominant approach to that point was one of 'urban managerialism' in which the state and its various agencies served as the key locus for administrating or running the city. Using ideas similarly inspired by Marx, Manuel Castells had developed an influential argument that urban politics was largely centered on conflict and division around the reproduction of labor power (Castells, 1977). His work on the politics of state housing provision in Paris proved particularly influential for subsequent scholarship on the politics of urban social movements (Castells, 1983). In short, urban managerialism is an approach to governing the city that principally concerns the allocation of resources and services to meet the political demands of different constituencies. Accordingly, urban political conflict is very much centered on struggles over the equitable distribution of state services and resources.

Harvey argued that while urban managerialism was the typical mode of governing the capitalist city in the 1960s, the 1970s and 1980s had seen a general reorientation of the concerns and issues deemed significant to how cities should be run. The period had been marked by the steady emergence of a more 'entrepreneurial' form of urban governance in which local ruling elites had begun to adopt a much more proactive role in seeking to determine the future trajectory of the city. Managerialism has centered on questions of social reproduction, whereas the urban agenda was becoming ever more shaped by the economic fortunes and prospects of the city. Entrepreneurialism reflected the development of what Cox has termed a 'new urban politics' as local economic development took center stage (Cox, 1993).

There are three dimensions or distinguishing characteristics of urban entrepreneurialism. The first concerns the actors engaged in governing the city and here Harvey's use of the term 'urban governance' is particularly significant. Governance reflects the view that the ordering and management of the city encompasses a range of actors beyond those of city government itself. Indeed, urban entrepreneurialism's first distinguishing characteristic is that it is centered on 'the notion of a public–private partnership' (Harvey, 1989: p. 7). Subsequent work in the urban entrepreneurial tradition has commonly sought to describe how 'partnerships' and partnership vehicles in particular cities have emerged.

While formal public–private partnerships are of relatively recent vintage, Harvey recognizes that local business interests have long sought to promote or boost their particular territories of interest. Along these lines the argument clearly resonates with other scholarly work seeking to theorize the politics of urban development. The concept of the urban growth machine developed by Harvey Molotch and John Logan (Molotch, 1976; Logan and Molotch, 1987) and that of the urban regime developed by Clarence Stone (Stone, 1989) are particularly germane. For Harvey, the novelty of the current condition is the way in which the traditional business interest concern with promoting the city "is integrated with the use of local government powers to try and attract external sources of funding, new direct investments, or new employment sources" (Harvey, 1989: p. 7). Key to Harvey's argument is not that local business interests have only recently developed a concern with the management and future trajectory of their city but rather that these interests have increasingly sought to pursue their concerns through the incorporation and use of the powers of the local state. Harvey's argument starts with economic interests and seeks to examine how those interests are realized through the mobilization of state powers. What ultimately drives the activities of those involved in governing the city is the attempt to secure the flows of capital investment and labor that will serve to realize the interests of those vested in the city. The argument certainly lends itself to research examining the history of business–state relationships at the urban scale, although work deploying the concept has very largely focused on contemporary circumstances.

The second feature of urban entrepreneurialism is the speculative nature of the activities undertaken by the partnership and, in particular, the distribution of the risks associated with these activities. Again Harvey recognizes that speculation is a necessary feature of capitalist investment and that investors will always seek to weigh risks against potential gains. What is key to urban entrepreneurialism is that the risks are now commonly borne by local public agencies. Again, Harvey is keen to distinguish how the assumption and distribution of risk marks a break from earlier phases of entrepreneurial activity. This second aspect of urban entrepreneurialism provides the focus for a significant amount of recent research. A number of studies have sought to describe and examine the different vehicles and schemes through which public agencies have encumbered the risks of speculative investment. These include the myriad ways in which local states and state agencies abate or defer the collection of taxes and revenues. This work, in turn, dovetails with a broad research tradition in urban studies examining the fiscal health of local and municipal government. The novelty of urban entrepreneurialism rests not in the presence of such risks – New York City in 1975 and Cleveland in 1976 serve as two high-profile cases of municipal bankruptcy – but in the proactive way in which local governments have assumed the risks associated with speculating on the future of the city.

Work on the fiscal side of entrepreneurial activities includes studies of the increasing use of property and other tax abatements that seek to encourage capital investment. The use of such incentives has a long history in the United States but the period since 1970 has witnessed their proliferation at the urban scale. While the onset of recessionary conditions in 2008 might be seen to harbor greater caution in the use of such programs, many cities and local governments are now faced with the fiscal consequences of speculating on future growth as developments are abandoned or delayed as a consequence of recessionary conditions. Property tax abatements were an early measure designed to encourage urban investment, although the complexity of financial measures has increased significantly over the past four decades. One particularly widespread vehicle currently employed in the United States is tax-increment financing in which projected increased taxes from urban development are used to fund capital improvements and other expenses that enable and accompany the planned development (Johnson and Man, 2001). The reliance on the projection of estimated future revenues brings into sharp relief the speculative nature of such schemes.

Harvey's third characteristic of urban entrepreneurialism – and perhaps the least well specified of the three – reflects his long-standing concern with the geographies of the activities that underpin the changing form of the capitalist city. Harvey argues that the effects of entrepreneurial activities can no longer be readily conflated with the territories in which those initiatives and projects are located. Baltimore's Inner Harbor development – in many ways a classic flagship urban entrepreneurial project – had economic, political, and cultural significance that extended well beyond the residents immediately impacted by its development. In this aspect of Harvey's argument, one can perhaps begin to trace the influence of a now burgeoning set of authors and studies highlighting the fragmented nature of the contemporary city (Graham and Marvin, 2001).

Urban Entrepreneurial Strategy

The three distinguishing features of urban entrepreneurialism provide a valuable means of unpacking changes in how cities are governed. While each provides a window on the contemporary city, it is the historical confluence of the three that marks a distinctive change in how one might best theorize and study the city. One of urban entrepreneurialism's particular strengths is its ability to encompass a wide range of different activities undertaken by place-based coalitions seeking to promote the city and their particular stakes within it. Mediating between the broad concept and the details of particular cities, Harvey identifies four options or strategies that provide a template for examining how particular coalitions of interest seek to place a city within broader economic and political structures. In this way, the concept has considerable flexibility and cleverly avoids conflating urban entrepreneurialism with any one particular strategy or policy. What counts as urban entrepreneurialism depends on the interests engaged in urban governance, the nature of the relationship among them, and the general character of the activities pursued. In these terms, the claims made by Harvey can be examined across a diversity of different contexts.

The first option, and arguably the broadest of the four, involves positioning the city to exploit particular advantages in the global division of labor for the production of goods and services. This opens up a wealth of different competitive

possibilities from a strategy of commodity production based on low-cost labor to the mobilization of key labor skills and talents attached to particular niche sectors. The possibilities here are multiple and the framework encourages us to recognize the politics involved in assembling local economic development strategies. In the US context, emphasizing the low cost of industrial labor has commonly been allied to a series of initiatives designed to inhibit the presence of unions and ensure an adequate and reliable supply of labor. Yet competition on the basis of labor costs is only one of a number of different possible strategies and, in certain instances, cities have sought to attract industry and economic activity that rests on professional, skilled, and relatively well-paid labor. The competition between cities to host creative industries and the employment associated with them is a popular example of competition on the basis of labor skills rather than labor costs. Similarly, the rash of initiatives to construct science, technology, and research parks commonly plays to the theme of attracting and retaining skilled labor.

The second pathway is to seek improvement of the city's place within the spatial division of consumption. The strategy is similarly geared to the attraction of investment and revenue into the locality. Activities that fall into this category include the promotion of cultural events and facilities to attract tourists and tourist spending. This is often accompanied by the construction or redevelopment of public facilities including art galleries, museums, concert halls, and aquariums. The overall strategy commonly involves a cluster of facilities in the form of a cultural district or neighborhood designed to attract visitors to the city. Significant capital investment whether in the form of sporting venues or convention centers can also form part of the strategy designed to position the city as a destination for visitors and visitor spending. Indeed, urban entrepreneurialism usually entails far more than the simple reimaging or rebranding of the city through the use of promotional materials. The consumption-based model covers a wide variety of different activities and pathways. In the European context, Barcelona tends to be heralded as a benchmark of urban development centered on cultural activities (Blanco, 2009). Yet the shift to entrepreneurialism is a widespread and general feature of the contemporary period and the competition surrounding the award of major cultural and sporting events provides further evidence of the growing intensity of consumption-based interurban competition.

A third strategy involves enhancing the position of the city within the spatial division of command and control functions. Here, Harvey links urban entrepreneurialism to the then burgeoning interest in global cities and their place in marshalling the flows of commodities, capital, and information that make up the global economy (Sassen, 1991). For cities such as New York, London, and Chicago, their key role as critical nodes within global financial markets and networks is a position deemed to require a constant stream of initiatives and improvements to the social and physical infrastructures that sustain their place within the wider flows of the global economy. Whether in the form of the new office space for financial services – such as those housed in the London docklands – or the role of airport hubs in encouraging the relocation of head offices – such as Boeing's move from Seattle to Chicago – the interests at the core of urban entrepreneurialism seek to protect and enhance the city's position in orchestrating the flows that connect the global economy.

While the first three strategies are largely concerned with the city's economic position, the fourth is very much centered on investment managed by the state. Harvey argues that urban entrepreneurialism can assume a concern with the redistribution of surpluses as cities seek to draw upon the resources and revenues collected and spent by higher tiers of government. The competition among cities and local governments for state facilities – whether in the form of university campuses, government offices, or even penal facilities – heralds the fourth strategy through which cities seek to promote their position within wider networks. In the United States, a concern with the geographic distribution of military and military-industrial facilities provides a particularly acute example of the competition for national state spending, albeit one that commonly encompasses actors and interests beyond those coincident with the city.

The menu of options from which local coalitions can draw points to the range of activities and initiatives compatible with the shift to urban entrepreneurialism. Furthermore, those invested in the city might seek to pursue more than one particular strategy. Indeed, the proactive role of urban-based interests in determining the strategies and policies pursued by local agencies provides grist to the mill for the study of urban politics and a sizable body of work reflects how this has played out in a variety of urban contexts (Hall and Hubbard, 1998). Through the link between territorial interests and the strategies they pursue, there are clear connections between the issues raised in Harvey's account and the broad and influential literature on urban growth coalitions and urban regimes. For Logan and Molotch (1987), the glue that holds together the otherwise potentially disparate interests animating the local coalition is the pursuit of growth. Yet how growth is manifest for Logan and Molotch is very largely reduced to the suburbanization of the city and the transformation of green space into urban form. In this formulation, property developers and land speculators lie at the heart of the growth coalition. Urban entrepreneurialism, by contrast, accommodates a much wider set of avenues through which ruling interests might seek to realize their stake in the development of the city. Accordingly, the concept has wide resonance.

Urban Entrepreneurialism in Context

Work adopting regime theory and the growth coalition shares urban entrepreneurialism's concern with the proactive role of local interests in seeking to manage the agenda of urban governance. Where urban entrepreneurialism differs from these concepts is in the emphasis placed on the constraints that shape both the practices of local elites and the extent to which entrepreneurial activities are successfully realized. In Harvey's formulation, activities within any one city are conditioned and shaped by the activities taking place elsewhere. The argument works at two levels. First, urban entrepreneurialism represents a general shift in how cities are managed and the new phase is marked by the replication of similar strategies and initiatives

across multiple urban contexts (Wood, 1996). The promotion of an arts and entertainment district becomes a more or less universal feature of the urban landscape. The recent policy mobilities literature is explicitly concerned with how urban initiatives and policies proliferate, emphasizing how they develop in one particular context and are then mobilized and deployed elsewhere (McCann and Ward, 2011). Such work provides significant insight into how the practices encompassed by urban entrepreneurialism – from the development of cultural districts to the pursuit of city-wide visions – seem to diffuse so quickly. For Harvey, the "serial reproduction of similar forms of urban redevelopment" (1989: p. 10) is a hallmark of the turn to entrepreneurialism.

At a second level, the constraints inherent in urban entrepreneurialism arise from the manner in which interurban competition for investment is structured or, in short, the rules that determine which strategies are successful and which are not. For Harvey, the suite of entrepreneurial activities is subject to the discipline of capitalist competition such that policies other than those serving the general interests of capital are always liable to be undermined. Harvey is less clear on how the discipline of competition is achieved in practice and the rather pessimistic view of the prospects of alternative strategies is moderated in his more recent work. In examining the discipline imposed on entrepreneurial activities, the work of Jason Hackworth on the neoliberal city is particularly helpful (Hackworth, 2007). For Hackworth, critical mediators in the interurban competition for investment are the bonding agencies that evaluate the extent to which local public authorities are capable of repaying their long-term debt. As Hackworth puts it, "the connection that cities have with capital markets is determined more than ever by the neoliberal standards enforced by rating agencies" (Hackworth, 2007: p. 33). For Hackworth, it is no coincidence that the turn to entrepreneurialism coincides with the growing dependence of cities on municipal debt to cover their spending on the provision of social services and capital infrastructure.

While the growth coalition and the urban regime frameworks tend to bracket out the broader economic and political forces that shape urban politics, urban entrepreneurialism encourages us to examine how the initiatives registered at the urban scale articulate with the broader changes characterizing the capitalist space economy. Rather than isolating the urban as a discrete political arena or reducing urban change to 'the effects' of broader economic forces, the initiatives and pressures at the core of entrepreneurialism are seen to have an important role in shaping the contours of capitalist development. In preparing the city for external investment, urban entrepreneurialism serves to encourage further rounds of place-based competition as entrepreneurial activities lubricate the flows of mobile investment that produce change in the global political economic landscape. Consequently, the urban is coupled to changes taking place across a range of geographical scales and the transition to urban entrepreneurialism can be read as integral to increasing place-based competition rather than the residual effect of the increasing mobility of capital. Indeed, central to Harvey's account of urban politics is not the mobility the capital favored by certain schools of thought but rather the endemic tension between the value locked in the immobile or fixed material and social infrastructures that make up the city and the flows of investment required to unlock that value and realize profit. The tension between fixity and mobility is similarly at the core of Cox and Mair's argument concerning the local dependence or geographic fixity of certain agents and their efforts to attract the flows of investment necessary to secure the economic health of the locality (Cox and Mair, 1988). Furthermore, the mobilization of place-based identity engendered by interurban competition provides a means of uniting social interests and groups and thus displacing the social and spatial divisions that Harvey and others see as endemic to the capitalist city.

While Harvey's argument has been widely cited, work deploying the concept of urban entrepreneurialism has proceeded unevenly. A significant body of recent work centers on the production of the facilities, landscapes, and images associated with the promotion of the entrepreneurial city. Short et al. provide an intriguing case study of how political and economic interests in Syracuse, New York, sought to rebrand the city from one marked by its association with an industrial past to a new image playing to the city's role as a destination for conference goers and retail consumers (Short et al., 1993). This is not to suggest that place marketing should be conflated with the turn to entrepreneurialism as the work of Gold and Ward (1994) on the history of place promotion clearly indicates (see also Kearns and Philo, 1993). Harvey's own work on the Baltimore Inner Harbor has also provided a model for subsequent work examining how 'flagship' projects – such as convention centers, museums, or cultural districts – mark the development of increasingly entrepreneurial strategies (Harvey, 2001). The shift to entrepreneurialism has also provided the context for a range of studies that have sought to detail how entrepreneurial strategies and the interests that promote them have displaced traditional arrangements for managing the city (Boyle and Hughes, 1994).

Unpacking the business interests that animate the drive to entrepreneurialism is a second area that has garnered attention. In urban studies, the term 'business community' is commonly taken as sufficient to capture the range of private interests and agents that have a stake in the development of the city. Harvey's account encourages a more nuanced view and scholars in the United States and Britain have sought to examine how different fractions of the business class have engaged with and encouraged the shift to entrepreneurial strategies (North et al., 2001; Valler, 1996; Rogerson and Boyle, 1998). How the practice of creating and mobilizing local business interests works in different national contexts raises the more general issue of the portability of the concept of urban entrepreneurialism and the arguments that underpin it. Scholars working in Great Britain have suggested that the framework tends to overplay the capacity of local business interests to bring about change. Peck and Tickell's work on Manchester provides a careful and thorough investigation of the weakness of the Manchester business community, reflecting both the development of national and increasingly transnational capital on the one hand and the autonomy of local government from business interests on the other (Peck and Tickell, 1995). The development of radical local authorities in Liverpool, Sheffield, London, and elsewhere

during the 1980s is witness to the strength of cross-national differences in the nature and role of the local state and, in particular, its relation to local business interests.

However, there are a number of areas in which work on urban entrepreneurialism remains underdeveloped. Rarely have studies examined how local residents have experienced the impact of entrepreneurial initiatives and how the distribution of costs and benefits has played out in particular cities. MacLeod's study of Glasgow provides a rare exception and examines how the repositioning of the industrial city has been accompanied by the marginalization of its least powerful residents (MacLeod, 2002). Yet how residents come to identify with the projects, ideas, and policies that underpin entrepreneurial strategies remains a question for further work (Hall and Hubbard, 1998).

While the concept of urban entrepreneurialism served to capture a number of different trends in the nature and purpose of governing the capitalist city, there has commonly been a significant gap between the concept – pitched at a relatively high level of abstraction – and the concrete changes examined in the study of the city. Urban entrepreneurialism has tended to provide a context for study rather than a set of propositions that serve as research questions or themes that lend themselves to empirical investigation. Studies have commonly sought to examine how urban entrepreneurialism plays out in a particular case rather than using a particular case or set of cases to examine the conceptual limits of the arguments. Furthermore, much of the work to this point has centered on cases drawn from North America and Western Europe, which raises a host of interesting questions about the insights that might (or might not) be gained from deploying the concept in a range of quite different settings.

Since its original formulation in 1989, additional concepts have come to frame the study of the capitalist city. Neoliberalism and the neoliberal city are, arguably, the most prominent and it is this concept that has displaced urban entrepreneurialism as the favored theme for radical urban studies. Harvey has made a significant contribution to understanding neoliberalism, although these arguments are commonly pitched at the global scale rather than focused specifically on its implications for urban form and how cities are governed (Harvey, 2005). Many of the themes raised in relation to neoliberalism clearly have a bearing on cities and the freedom of trade, the liberalization and deregulation of markets and the veneration of the market relative to the state are each consistent with the turn to urban entrepreneurialism and entrepreneurial governance. Moreover, the continued resonance of urban entrepreneurialism might be seen to rest in its capacity to link the latest phase of capitalist development with the profound set of changes that have characterized how cities are governed.

See also: City Marketing; Gentrification; Local Economic Development; Urban History; Urban Megaprojects; Urban Neoliberalism; Urban Policy in Europe; Urban Policy in North America; Urban Regeneration; Urban Sociology; Urbanization.

Bibliography

Blanco, I., 2009. Does a 'Barcelona Model' really exist? Periods, territories and actors in the process of urban transformation. Local Government Studies 35, 355–369.
Boyle, M., Hughes, G., 1994. The politics of urban entrepreneurialism in Glasgow. Geoforum 25, 453–470.
Castells, M., 1977. The Urban Question: A Marxist Approach. The MIT Press, Cambridge MA.
Castells, M., 1983. The City and the Grassroots. University of California Press, Berkeley.
Castree, N., Gregory, D. (Eds.), 2006. David Harvey: A Critical Reader. Blackwell, Oxford.
Cox, K., 1993. The local and the global in the new urban politics: a critical review. Environment and Planning D, Society and Space 11, 433–448.
Cox, K., Mair, A., 1988. Locality and community in the politics of local economic development. Annals Association of American Geographers 78, 307–325.
Gold, J., Ward, S., 1994. Place Promotion: The Use of Publicity and Marketing to Sell Towns and Regions. John Wiley, Chichester.
Graham, S., Marvin, S., 2001. Splintering Urbanism. Routledge, London.
Hackworth, J., 2007. The Neoliberal City: Governance, Ideology and Development in American Urbanism. Cornell University Press, Ithaca.
Hall, T., Hubbard, P. (Eds.), 1998. The Entrepreneurial City: Geographies of Politics, Regime and Representation. Wiley, Chichester.
Harvey, D., 1973. Social Justice and the City. Edward Arnold, London.
Harvey, D., 1978. The urban process under capitalism: a framework for analysis. International Journal of Urban and Regional Research 2, 101–131.
Harvey, D., 1985. The Urbanization of Capital. Blackwell, Oxford.
Harvey, D., 1989. From managerialism to entrepreneurialism: the transformation of urban governance in late capitalism. Geografiska Annaler 71B, 3–17.
Harvey, D., 2001. Spaces of Capital. Routledge, New York.
Harvey, D., 2003. Paris, Capital of Modernity. Routledge, New York.
Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press, Oxford.
Johnson, C., Man, J. (Eds.), 2001. Tax Increment Financing and Economic Development, Uses, Structures and Impact. State University of New York Press, Albany, NY.
Kearns, G., Philo, C., 1993. Selling Places: The City as Cultural Capital, Past and Present. Pergamon, Oxford.
Logan, J., Molotch, H., 1987. Urban Fortunes: The Political Economy of Place. University of California Press, Berkeley.
MacLeod, G., 2002. From urban managerialism to a 'revanchist city'? On the spatial injustices of Glasgow's renaissance. Antipode 34, 602–624.
McCann, E., Ward, K. (Eds.), 2011. Mobile Urbanism: Cities and Policymaking in the Global Age. University of Minnesota Press, Minneapolis.
Molotch, H., 1976. The city as a growth machine: towards a political economy of place. American Journal of Sociology 82 (2), 309–355.
North, P., Valler, D., Wood, A., 2001. Talking business: an actor-centred analysis of business agendas for local economic development. International Journal of Urban and Regional Research 25, 830–846.
Peck, J., Tickell, A., 1995. Business goes local: dissecting the 'business agenda' in Manchester. International Journal of Urban and Regional Research 19, 55–78.
Rogerson, R., Boyle, B., 1998. Reluctant mavericks: towards an explanation of the limited interest of capitalist firms in shaping Glasgow's inward investment strategy. Scottish Geographical Magazine 114, 109–119.
Sassen, S., 1991. The Global City: New York, London, Tokyo. Princeton University Press, Princeton NJ.
Short, J., Benton, L., Luce, W., Walton, J., 1993. Reconstructing the image of an industrial city. Annals of the Association of American Geographers 83, 207–224.
Stone, C., 1989. Regime Politics: Governing Atlanta, 1946–1988. University of Kansas Press, Lawrence KS.
Valler, D., 1996. Locality, local economic strategy and private sector involvement: case studies in Norwich and Cardiff. Political Geography 15, 383–403.
Wood, A., 1996. Analysing the politics of local economic development: making sense of cross-national convergence. Urban Studies 33, 1281–1295.

Urban Geography

Rita Schneider-Sliwa, University of Basle, Basle, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urban development processes are greatly affected by structural transformation of societies. Over the past decade, societies have witnessed changes that were unprecedented in recent history: the collapse of political and economic systems, new integration policies, globalization of the economy and the decline of the nation-state, mass migration, governmental retrenchment, and social restructuring. Structural changes in societies manifest themselves distinctly in urban areas and every generation needs to engage in the discourse on a country's built politics, the sustainability of urban form and design, and the ecological footprint of major cities and suburban areas. Urban geography deals with the analysis of the complex dimensions of urban social, economic, cultural, and political processes, patterns, and structures and urban planning processes to build up or retain local comparative advantages, while retaining cities as livable urban places and securing their sound social and environmental development for future generations. This article describes the conceptual and methodological approaches and the long-term prospects of urban geography.

Introduction

For the first time in human history, the majority of the world's population lives in cities. The megacities of low-income countries are subject to rapid growth and as a result are increasingly difficult to govern. For megacities in developing countries problems arise from their ever-growing populations and the fragmentation of cities with parts of the urban economy partially integrated into the world economy and parts being left out even of national development. Their challenge is to accelerate economic and balanced, sustainable development in order to reduce the potential of violent conflict and to support peace processes. Urban/metropolitan areas in highly industrialized nations are likewise facing challenges, including increased intraurban competition for investments and tax payers', and to increase their capacity to protect and strengthen collective security.

Theoretically informed social science research is necessary to monitor trends of urban development, to provide basic information for the optimization of local strengths and to meet local needs, and for urban planning that respects collective decisions, increased competition, modern urban structures, and the individual 'feeling' of a city. Social science research defines and analyzes the generators of urban development and brings to light basic information about urban structures and processes, perceptions, for example, of inner city problem areas, and the attractiveness of cities. This is where urban geography has an important role to play.

Urban Geography: Tasks and Approaches to a Social Science Discipline

The Scope of Urban Geography

Urban geography deals with the analysis, explanation, and prognosis of urban forms, urban social fabric, and economic structures and functions. Questions encompass economic, political, social, and ecosystem geography in their urban contexts at various scales, from the macroscale of the urban system to the mesoscale of metropolitan functional urban areas to the microscale of urban neighborhoods, streets, and even individual households. The scientific results of urban research increasingly serve as a basis for decisions on public investment, allocation of resources, and urban development planning. Urban research can be broken down into interurban (system of cities) and intraurban analyses that address the processes going on in cities (*see* Cities: Internal Structure). Conceptual approaches pertain to urban form and urban morphology, the urban social fabric and economic structures, and urban development and urban policy.

Urban Form and Urban Morphology

This approach deals with analysis of the built environment. Changes in urban morphology, including historic/heritage preservation, the creative reuse or restoration of landmark buildings, and urban revitalization, have gained importance in recent years for at least the following two reasons. (1) The built heritage of cities displays layers of significance and narratives that bridge the past with the present. As heritage properties have symbolic, esthetic, and emotive values, measures to protect heritage planning can improve a city's attractiveness for residents and investors and hence, the city's fiscal base. (2) Urban morphology has been linked to national defense strategies as was the case during World War II and the Cold War when dispersed urban development became an integral part of national policy in the United States (Augur, 1948).

Urban form and urban morphology have also been taken up in Neo-Marxian/Neo-Marxist urban geography. The regulationist school, for example, links power structures to the functioning of urban regimes. According to this school of thought, investment in modern urban centers is a means of accumulating capital at the cost of other urban subareas, a process accompanied by policies of neglect and disinvestment. Social and welfare disparities as are manifest in differing morphologies across urban space have thus become another major area of urban research.

Urban Function and Urban Social Structure

Analysis of Urban Land Use Development, Urban Functions, and Specializations

Specialized economic activities seek locations with the greatest competitive advantage. This approach helps to understand urban land use development and to determine those land use patterns that most competitively provide basic support for the city. The internal structures of specialized urban functions (wholesale, retail, service activity, etc.) within the city are points of focus. On the microscale of the city, the analysis of central business districts and neighborhood business have received priority (Weaver and Duerksen, 1977; Alperovich, 1982; Mallet, 1994; Briffault, 1999; Ward, 2007; Duranton and Puga, 2005). On the meso- and macroscales of urban systems analysis, the degree of specialization in and between cities is considered the base for identifying a city's market potentials, strengths, and niches within a regional or national hierarchy of cities. Results from urban systems research on urban specializations have found their way into policy analysis of regions that are in need of economic, social, and territorial cohesion. Here, public funds fostering place-based policies are channeled into places best able to contribute to the sustainable and balanced development, as identified by urban systems research. This, for example, was the case in Germany after reunification (Spangenberg, 2008; Bundesinstitut für Bau-, Stadt- und Raumforschung, 2012).

Empirical Examination of Factors Organizing Space and Megalopolitan Structure

This approach deals with the metropolis as a changing configuration and the processes by which metropolitan areas grow and expand beyond their rigid corporate limits. Studies of metropolitan patterning examine the functional differentiation of suburbs, the interrelation of the city and its surrounding area through economic linkages and commuting, and the way metropolitan communities reorganize themselves into supra-metropolitan functional urban areas (Antikainen, 2005; Perlik et al., 2001; Anthrop, 2004). The development of megacities, the increasing functional specialization of metropolitan centers and communities, and their growing interdependency are key research issues, as is the general theme of urban governance of such larger functional urban entities that extend beyond administrative borders (Coombes et al., 1982; Van den Berg and Braun, 1999; Meijers 2005).

Urban Social Geography and the Factorial Ecology of Cities, Social Status Differentiation, and Segregation

Urban areas are highly differentiated complex systems (see also Cities: Internal Structure in this encyclopedia). These approaches deal with urban subareas, urban subpopulations, social and economic characteristics of urban neighborhoods, the social patterning of local residents by race, ethnicity and class, and the behavior of subpopulations as a function of their social group, race, and class, or as mediated by the characteristics of the neighborhood, urban subarea, and the urban system. This approach has recently become more important given the tendencies for growing social exclusion of individuals, household, groups, and communities with respect to education, equal employment opportunities, cultural and gender norms, incomes, occupational status, and social networks based on race, gender, class, origin, or religion, and given new policies against social exclusion. Combatting such tendencies, for example, has been integrated into the European Union (EU) Treaty of Maastricht and related European economic development programs, and it is by means of geographic social science research that one can track progress or stagnation in this regard.

Urban subarea characteristics and urban social and spatial differentiation are commonly analyzed using factorial ecology. This term refers to various statistical approaches using factor analytic methods. Small area analyses with a limited set of variables grounded in social theory is common. Factor analysis, on the other hand, depends on a large number of variables that are then reduced in an exploratory way to essential properties of a particular phenomenon of urban subareas or urban space. It determines urban subareas according to common social characteristics or in terms of households or individuals with common characteristics.

The urban ecologist or social geographer does not focus on spatial differentiation per se. Rather, the urban geographer studies such patterns with a view of uncovering the social, political, economic, or cultural processes that may be responsible for these patterns. Residential segregation of different social status groups, for example, occurs in many different cultural settings. In increasingly multicultural societies, patterns of differentiation among social groups and neighborhoods and their particular spatial geometry are extremely complex. Factorial ecological investigations help to monitor processes of social distance as reflected in the degree of physical distance and residential separation. Of note in this context is a major EU research program called ESPON – European Observation Network, Territorial Development and Cohesion. This applied research program provides pan-European geographic information systems (GIS)-based and social science-based evidence and knowledge about European territorial structures, trends, perspectives, and policy impacts. The research enables comparisons among regions and cities with the overall goal to tailor result-oriented policy toward promoting economic prosperity, competitiveness, territorial cooperation, social cohesion, and more sustainable development throughout the EU.

Perception of the Urban Environment

This approach sees age- and social-status-related perceptions of the urban environment, personal activity radius, and spatial behavior as related variables. Individual perceptions of reality affect spatial behavior, be it shopping trips or intraurban migration. Different groups such as the elderly, the youth, and the poor have distinct ranges of activities and different patterns of spatial behavior. Market research utilizes the results of studies on the perception of the urban environment and spatial behavior. At the microscale of individual urban geography, studies concentrate on activities that a person does regularly and places that are visited regularly, i.e., daily or weekly. These may be analyzed using the methods of time geography and indicate an individual's capabilities to overcome distance by means of mobility and communication. At the macroscale, the collective spatial behavior of social groups and the spatial

geometries of social group behavior are the focus. Decisions regarding distances between shopping centers and other central urban functions, for example, make use of empirical findings of collective spatial behaviors.

Perceptual geographic urban analysis upholds the theory that spatial behavior, like intraurban migration or shopping and recreational activity, is often limited to a closed field of perception and reference, and it is affected by personal spatial evaluation. The overall image of a city from the point of view of commercial interests is affected by 'soft' locational factors (e.g., amenities, attractiveness of the urban environment, social and demographic processes in the city, or business climate). Image analysis can contribute significantly toward finding reasons for the exodus of population, companies and enterprises, or disinvestment. Moreover, measures for the improvement of a negative city image or for instigating desired development can be recommended. Image analysis, therefore, has become an important instrument and base for (re)defining strategies for the marketing and branding of cities and urban regions.

Urban Development

The Analysis of Urban Development and Urban Morphology as Related to Historical, Political, and Geographic Peculiarities

This approach that dates to the 1930s and was rooted in the analysis of European cities deals in particular with the development dynamics of the city and the determinants of that development. Urban development analysis takes the positive consequences (e.g., concentration of social and economic performance, importance as center of innovation/innovation hub) and negative consequences of development (e.g., environmental problems or urban blight) into consideration. Today, this approach offers important insights for the political, economic, and social components of urban development planning.

Urban Cultural Geography, Intercultural Comparison of Cities, and Urbanism

This approach emphasizes the historical and cultural context of cities. Processes of urbanization and the current physical urban structures of cities, including their inner differentiation, are assumed to be products of a series of sociopolitical systems. In cultural genetic urban geography (see Section Culture Type of Cities below), persistent historical urban structures are identified according to the appropriate historical social system and cultural sphere of influence. Because society changes faster than do physical urban structures, linking historical patterns in the urban fabric with historical social systems is a major focus of research in cultural genetic urban geography.

Culture Type of Cities

This research focuses on culture-specific differences in urbanization and urban structure that were created in the course of historical development. The physical structure of the city and the differentiation of urban quarters are seen as related to the value system and demands of a dominant culture. Commonly, such cultural imprints are evident in the central parts of many modern cities. Proponents of this approach hold, for example, that cultures and historic contingencies have shaped cities differently according to the predominant cultural values of the respective society and epoch. Cities in all cultures have reflected the visions, fundamental beliefs, and ideological orientation of their founders and of successive generations.

In today's competitive globalized and technically oriented world, historical built heritage as well as cultural and spiritual norms provide an important record of human achievement. As pointed out by Gospodini (2004), the constructed landscape narrative serves as a place identity generator. Being able to read historical urban landscapes may make unfamiliar environments familiar and overcome xenophobia.

Postmodern Urbanism

Postmodern urbanism aimed to provide a consolidated body of theory challenging the Chicago School: The so-called Los Angeles School utilized Marxist and poststructuralist theories to interpret sociospatial differences in late-capitalist, globalizing cities (Dear, 2000, 2003, 2005; Dear and Dahmann, 2008). Los Angeles, deemed to show exceptions from established urban theory, became the laboratory for postmodern urban studies. The *gestalt* of the late-twentieth-century city no longer possessed concentric order or homogenous sectors according to postmodernists. Rather, postmodernists saw a mosaic of fragmentation, with poverty landscapes and spectacular centers of symbolic and commercial significance, representing the activity of global capital. Such centers could be in the center or peripheral nodes. These new urban developments were seen as dramatic and polarizing. Spectacular urban design elements were interpreted as manifestations of fierce global competition and necessary 'unique selling points' that could distinguish one city from another in the competition for global capital. To capture both the driving process and results of such urbanization, many concepts have been put forth, including *world city*, *dual city*, *hybrid city*, *cyber city*, *fragmented urbanism*, *edge cities*, and *fortress urbanism*.

Despite its claim to novelty, postmodern urbanism may be understood as a variation of the theme of culture-specific urban development: Whereas in former times the combination of mercantile feudal systems and ecclesiastic power and wealth shaped cities, creating monumental worldly and church architecture but leaving entire city quarters neglected, in contemporary society, the combination of laissez-faire capitalism, big corporations, and government devolution shapes cities. Now, as then, hegemony and capital are driving forces of polarized urban development. Postmodern urbanists drew attention to problems of fragmentation and to the need to explore them more intensively and make them a political issue.

Urban Policy Analysis

Globalization and Urban Policy

Cities are social systems in space that react to the following local and supralocal developments:

- Increasing competition among metropolitan areas as a result of globalizing economies and cultures.
- Deregulation and separation of social politics from trade and industrial politics.

- Economic and political integration as well as political and environmental crises that initiate migration flows.
- Change of societal values and pluralism of lifestyles.

Global economic restructuring, increased competition among cities, and the decline of the welfare state force cities to pursue entrepreneurial urban development strategies. These may include economic incentives, the promotion of investments or megaprojects to increase urban attractiveness, and a focus on growth rather than redistributive urban development. Also, public–private partnerships or urban regimes that emphasize private sector planning have emerged.

Urban regimes follow different approaches to development; some implicitly exclude or include social policies as part of their overall strategy, with differing impacts on long-term development processes. For an individual city, analyzing the interrelationships among urban policy, urban development, and the changing geometries of segregation helps to understand how and why planning may unintentionally influence social and spatial polarization. Given the adoption of Western urban policy and planning models in other countries, it is important to understand social, spatial, and economic implications of each type of policy for urban development in different cultural settings.

Government Devolution: The State and Local Response

In a national urban system, the relationship between national/federal urban policy and local policy and local urban development is of interest to urban geographers: what happens in cities when federal policies change? Recently, financial responsibility has shifted away from national governments toward state and local governments. Noteworthy, however, is the co-occurring tendency since the early twenty-first century for central governments to assume extensive responsibility for funding issues deemed vital to national interests, with impacts on every household locally.

The trend toward government devolution with respect to issues considered local has been observed particularly in the US and Western European federal systems; it has prompted concern that social problems may worsen if the federal/national role of counterbalancing unfavorable social and economic developments is restricted. While federal/national policies may at first glance adversely affect areas of concentrated poverty, one must look at local policies and responses to Federal/national policy to achieve balanced view of long-term adjustments and trends in societal development.

Policy-oriented urban geography studies the problems and potentials of local planning responses to current problems of urbanism and federal and state policies, such as social/economic disparities, caused by policies at various levels. Research on restoring inner city neighborhoods in the United States reveals examples of innovative community concern expressed by public and private sectors with the potential to counterbalance public sector devolvement to some extent. European nations, among others, have developed interesting local policy mechanisms to deal with the ills and imbalances of society manifest in urban areas (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2012, 2013).

Methods in Urban Geography

Conceptual approaches focus mainly on two perspectives.

- Those related to urban space and urban subarea characteristics, the basic dimensions of city patterning and the determinants and processes behind changing structural characteristics of urban space over time.
- Those related to people in urban space, e.g., individual and collective spatial behavior (for example, shopping trips, commuting, daily activity ranges), people's perceptions of urban space, and planning processes or projects.

Concomitantly, two major, complementary methodological approaches characterize urban research: (1) empirical regional/urban subarea analysis and (2) empirical social science, particularly survey methods.

Methods of Empirical Regional/Urban Subarea Analysis – Urban Social Monitoring

Societal developments manifest in urban space and are reflected in structural characteristics of urban subareas. Urban geographers analyze the aerial structure of urban communities in terms of aerial attributes. The methods correspond to those found in regional geography in so far as they help to delineate and observe structural changes of agglomerations, city centers/cores, urban expansion, and suburban areas. Within the city, research units are districts and neighborhoods as well as other 'official' spatial units, whether defined for planning, political, or statistical purposes. Urban subareas may be of any scale; census tracts and smaller census units are commonly used as statistical reference areas to characterize the increasing differentiation of urban social milieus.

Urban subarea analysis enables *urban social monitoring*: documenting and analyzing detailed socioeconomic structural patterns and processes of change. Complex spatial processes are broken down into individual components. The following three statistical approaches are important. (1) For *descriptive purposes*, methods include computer-assisted cartography and the refined cartographic and analytic methods enabled by GIS. These produce unidimensional or multidimensional maps of social, demographic, or other phenomena as differentiated in urban space. GIS, of course, is also suited for the establishment of a long-term statistical–cartographic database, which can be periodically updated. Such a database simplifies thematic longitudinal onsite analysis of the target urban region with regard to social, economic, and demographic processes and forecasts. (2) In a more exploratory sense, *factorial ecological investigations* use a number of multivariate descriptive statistical techniques (the methods of factor analysis) to identify the essential characteristics that differentiate one urban subarea from another. *Cluster analyses* subsequently performed on factor-analyzed urban subareas can help identify groups of subareas with common patterns of variability. (3) In order to understand the determinants of such patterns, one may combine descriptive and analytical statistical techniques. Factor scores from factorial analyses may, for example, be used as input data in *multiple regression*

analyses that relate these aggregate characteristics to explanatory variables.

Social monitoring of urban subarea characteristics over time enables evaluation of current structural changes: urban geography relies on statistical data collected by public and private institutions or public welfare organizations. As the data reflects institutional norms and goals, analysts have no influence on the survey method or the aggregation and systematization of the indicators. Consequently, theoretically informed urban research is limited by the quality of these (secondary) data sources, which is continually improving.

Survey Methods – Empirical Social Science Research

Survey research methods are needed to analyze activity systems/patterns of individuals or households, as well as preferences, felt needs, perceived quality of the urban environment, and residential behavior in specific urban contexts. Survey research methods in social geography concentrate on the *interrelationship/correlation between spaces and social group behavioral patterns*, designation and appraisal, as well as group perception and evaluation of spaces. An *advantage* of incorporating surveys early in the research process is the ability to identify important issues that are not addressed with official statistics.

The Contribution of Empirical Urban Research

In sum, research questions in urban geography address different sets of issues:

- the physical attributes of spaces or indicators of their functional interconnection
- immaterial aspects (spatial perception or space-related activity)
- both of these or
- aspects of spatial change that result from the interaction between immaterial and physical elements (e.g., the physical results of political visions for the urban area and the consequent measures, or sociospatial disparities caused by power relations in society)

Studies of objective reality (e.g., sociogeographic structural analysis) use data that reflect numerically definable spatial characteristics that can be utilized in factorial analysis (as opposed to subjectively perceived spatial attributes). This data is mostly found in official statistical material, published and unpublished, the basis of which are extensive censuses (e.g., of population) or surveys (e.g., of enterprises). Personal, small-scale, and tailored surveys are necessary if the time lapse between official censuses is too great or if certain key data elements are missing (e.g., building conditions in different neighborhoods). However, if the research focus concerns the activity patterns of individuals and groups, interviews are necessary if spatially related behavior, socioeconomic background, and possible consequences of behavioral patterns are to be dealt with. Between factorial ecology and survey methods, only the data sources and the scale of the evaluation differ. Even though the goals may vary, the *basic principle* of *preparing and evaluating* official or specifically collected social data from survey research in a descriptive and/or quantitative analytical manner stays the same.

Key Features of Urban Geography

The standing of urban geography within the social sciences draws on three elements that other disciplines dealing with urbanism do not share to the same extent.

Combination of Conceptual and Methodological Capabilities

- Empirically and theoretically informed research from an integrated perspective.
- Relevance to society.
- Diversity of urban research topics in social, economic, and environmental fields, among others. The range of *urban research topics* includes, for example:
  - Changing patterns of residential segregation, the persistence of urban poverty, the rise of gated communities, housing and urban labor market developments.
  - Urban environmental quality, questions of environmental racism.
  - Functional interrelatedness of cities and their suburban areas.
  - The territorial embeddedness of global impacts.
  - Local policy responses to supralocal trends.
  - Entrepreneurial urban policy and microscale urban social development.
  - Cross-cultural urban development.
  - Urban systems development.

Specific Methodological Competencies and Diversity of Analytical Responses

- Empirical research: primary survey methods, such as monitoring, mapping, or field surveys.
- Statistics: drawing conclusions about research questions/hypotheses via processing spatial primary data and official statistical secondary data.
- Cartography: tool for working with and presenting spatially differentiated phenomena.
- Remote sensing: interpretation of aerial photos and satellite photos as an important source of information.
- Using GIS to present and analyze data and developments in space.

Decoding Highly Complex Encrypted Urban Social Landscapes

Urban geographic thought ranges from notions of culturally shaped cities to cities formed by modern societies and cities having lost their character as a result of postmodern societal development or the culture of globalism. Concomitantly, urban analysis has proliferated to be able to track each facet of urban development as it was conceptualized and postulated to exist. The cumulative analytical expertise is shown in Table 1.

Conclusions

Global economic restructuring, political change, government devolution, and social restructuring alter urban and regional

Table 1 Urban geographic analysis: spatial reference levels and research emphases

	Systems, spatial reference levels	*Research emphasis (examples)*
Interurban systems	**Global/international levels**	● Worldwide or international urbanization ● Worldwide distribution of megacities, global cities, etc. ● Types of cities/culture type of cities (comparative perspective) ● Cities in industrial and developing countries
	National, regional, and interregional levels	● Regional types of cities
	Intraregional level including functional urban regions and large metropolitan areas	● Development of conurbations ● Relocation of businesses/corporate headquarters, administrative authorities ● Mobility, including commuters, retail activities, leisure related ● Central place systems, urban networks, and urban systems
Intraurban systems	**Entire city or municipality**	● Urban development and structure ● Inner city centers or system of centers ● Shopping centers at the urban fringe vs inner city retail centers ● Suburbanization processes (population, commerce, and industry) ● Sociospatial segregation/fragmentation ● Inner city mobility/leisure and shopping behavior ● Urban image and city marketing ● Urban traffic systems (public transportation)
	Mesoscale urban areas (city quarters, neighborhoods)	● Living conditions in neighborhoods ● Housing market and residential location
	Microlevel urban areas (retail locations, street sections)	● Social structure, social problems in inner cities and neighborhoods ● City functions, central business districts and commercial streets ● Inner city structure, commercial developments ● Urban form, built environment ● Urban renewal, neighborhood rehabilitation and planning, pre-analysis and accompanying studies ● Environmental pollution (noise, emissions, fine dust, etc.)

Based on Heineberg, H., 2014. Stadtgeographie. Grundriss Allgemeine Geographie, fourth ed. Schöningh, UTB, Paderborn, p. 23.

fortunes and give rise to new sociospatial patterns of development. Fragmentation, polarization, and new local spatial identities form new urban realities. Formerly prosperous cities may lose their comparative advantage. To retain competitiveness and build up local (endogenous) potentials or to govern uncontrollable cities, new models and modes of planning are being developed. In cities of the Western world, public–private partnerships – coalitions of major local economic and political players – have been outlining/developing and implementing grand visions of development. Many megacities in low-income countries have difficulty coping with the basic infrastructural needs of their populations.

Thus increasingly, in cities worldwide, urban planning and the prevailing urban regime and its orientation and complex interaction with the regional, national, or supranational economic and political systems have a significant effect on the development of a city. In order to deal with contemporary urbanism, decision makers will need to rethink urban processes, structures, and policies. Understanding the city from a systems perspective is required because urban development, form, and social and economic structure are intertwined with the larger society, economy, and politics nationally and internationally. A systems perspective also understands the city as an organism/entity with enormous local capacities and strengths that might successfully counteract the local or regional effects of globalization or other supralocal forces.

Urban geography is a systems-oriented social science discipline with great relevance to interdisciplinary solutions of problems of urbanism and urban and urbanizing areas. The systems perspective, the methodical skills required, and the variety of theories and research questions in urban geography make this discipline well suited to understand and deal with problems of contemporary urbanism.

See also: Cities: Internal Structure; Cultural Geography; Residential Segregation: Recent Trends; Social Geography; Urban System in Geography; Urbanization.

Bibliography

Alperovich, G., August 1982. Density gradients and the identification of the central business district. Urban Studies 19, 313–320.

Anthrop, M., 2004. Landscape change and the urbanization process in Europe. Landscape and Urban Planning 67 (1–4), 9–26.

Antikainen, J., 2005. The concept of functional urban area. Findings of the ESPON project 1.1. Informationen zur Raumentwicklung 7, 447–452.

Augur, T.B., 1948. The dispersal of cities as a defense measure. Journal of the American Institute of Planners 14 (3), 29–35.

Briffault, Richard, 1999. A government for our time? Business improvement districts and urban governance. Columbia Law Review 99, 365–477.

Bundesinstitut für Bau-, Stadt- und Raumforschung im Bundesamt für Bauwesen und Raumordnung, 2012. Zentrale Orte. http://www.bbsr.bund.de/BBSR/DE/Raumentwicklung/RaumentwicklungDeutschland/Daseinsvorsorge/Projekte/ZentraleOrte/ZentraleOrte.html.

Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2012. Nationale Stadtentwicklungspolitik Eine Gemeinschaftsinitiative von Bund, Ländern und Gemeinden. www.nationale-stadtentwicklungspolitik.de.

Bundesministerium für Verkehr, Bau und Stadtentwicklung, Bundestransferstelle Soziale Stadt, 2013. Soziale Stadt – Investitionen im Quartier. Gute Beisppiele aus der Praxis vor Ort. Deutsches Institut für Urbanistik und Bundesinstitut für Vau-, Stadt- und Raumforschung, Belrin.

Coombes, M.G., Dixon, J.S., Goddard, J.B., Openshaw, S., Taylor, P.J., 1982. Functional Regions for the Population Census of Great Britain. See more at: http://www.popline.org/node/401503#sthash.TNlwf0tm.dpuf. http://www.popline.org/node/401503.

Dear, M., 2005. Comparative urbanism. Urban Geography 26 (3), 247–251.

Dear, M., 2000. The Postmodern Urban Condition. Blackwell Pub., Oxford.

Dear, M., 2003. The Los Angeles school of urbanism: an intellectual history. Urban Geography 24 (6), 493–509.

Dear, M., Dahmann, N., 2008. Urban politics and the Los Angeles school of urbanism. Urban Affairs Review 44 (2), 266–279.

Duranton, G., Puga, D., 2005. From sectoral to functional urban specialization. Journal of Urban Economics 57 (2), 343–370.

Gospodini, A., 2004. Urban morphology and place identity in European cities: built heritage and innovative design. Journal of Urban Design 9 (2), 225–248. http://dx.doi.org/10.1080/1357480042000227834.

Heineberg, H., 2014. Stadtgeographie. Grundriss Allgemeine Geographie, fourth ed. Schöningh. UTB, Paderborn, p. 23.

Mallet, William J., 1994. Managing the Post-industrial city: business improvement districts in the United States. Area 26 (3), 276–287.

Meijers, E., April 2005. Polycentric urban regions and the quest for synergy: is a network of cities more than the sum of the parts? Urban Studies 42, 765–781.

Perlik, M., Messerli, P., Bätzing, W., 2001. Towns in the Alps. Urbanization processes, economic structure, and demarcation of European functional urban areas in the Alps. Mountain Research and Development 21 (3), 243–252. http://dx.doi.org/10.1659/0276-4741(2001)021[0243:TITA]2.0.CO;2.

Spangenberg (BBR Bonn), M., 2008. Das System der Zentralen Orte in der aktuellen Landes- und Regional planung der Bundesländer, insbesondere vor dem Hintergrund des demografischen Wandels. www.stgb-brandenburg.de/fileadmin/user_upload/stgb-brandenburg.de/dokumente/themengebiete/StGB_ZOS_Spangenberg.ppt.pdf.

Van den Berg, L., Braun, E., 1999. Urban competitiveness, marketing and the need for organizing capacity. Urban Studies 36, 5–66, 987–999.

Ward, K., 2007. Business improvement districts: policy origins, mobile policies and urban liveability. Geography Compass 1 (3), 657–672.

Weaver, C.L., Duerksen, C.J., 1977. Central business district planning and the control of outlying shopping centers. Urban Law Annual 14, 57–79.

Urban Growth Models

Britten Harris, University of Pennsylvania, Philadelphia, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16018–16021, © 2001, Elsevier Ltd.

Abstract

Models can perform the dual functions of organizing knowledge of how urban systems work and grow, and providing support for analyzing policy and planning proposals. Modern urban modeling has progressed substantially in the last 50 years, as a result of extensive development supported by massive data collection and growing computing power. Much of the development has been impelled by the needs of transportation analysis, but land use or locational modeling has advanced well beyond initially fragmented and descriptive theories. The major difficulty arises out the underlying importance of resource allocation, which is influenced by complex behavioral patterns and requires an interdisciplinary approach. These behaviors are defined too narrowly in purely economic models, and too descriptively in purely observational models of discrete choice. This dichotomy is a major focus of modeling, and is being gradually overcome.

An urban model is a statement about the interrelationships between parts of an urban system, which not only describes these interactions, but makes possible the prediction of the effects of some kinds of changes in the system or in its environment. This is the essence of any scientific theory.

Viewed as a part of a scientific discipline, urban models provide an organizing framework for bringing together many aspects of knowledge about urban life, for testing subsidiary theories, and for identifying gaps in knowledge and guiding future research. Viewed as an aid to professional practice in urban planning or engineering, a model which is operational at a relevant level of detail can provide insights into the future results of new policies, and guidance in their formulation and improvement. To serve these two divergent purposes, a model must be understandable and acceptable to both the scientist and the professional.

Prior to the mid-1950s, many disciplines and professions engaged in disjointed efforts to analyze urban affairs. In Geography and Sociology, there were efforts to achieve an overall view in largely descriptive terms. Examples include Burgess (1927), Hoyt (1939), and Harris and Ullman (1945). Starting in a period of optimism about scientific management, the last 50 years have seen substantial growth in urban modeling, but there are still many imperfections in current models, and obstacles to their filling their dual role. We will try to expose the bare bones of both the progress and the difficulties as they have developed over five decades.

Basic Considerations

An urban metropolitan area comprises a set of organized activities and residential spaces, occupying land and buildings, and using infrastructure. These activities and their containers are at higher densities than their nonurban surroundings, and are more closely connected with each other than with their ambient environment. Their existence and development require the proper functioning of markets or allocation methods for land, for housing, offices, and factories, for labor, and for public and private services, including those of the transport and communications systems which knit these other functions together.

Because of the paramount importance of resource allocation, and the role of real or artificial markets, these interrelationships are largely governed by economic considerations – but with many differences from received economic dogma. There is a division of decision-making between the private and public sectors, necessitated in part by economies of scale and indivisibilities. In the ultimate political sense, economic efficiency is only one of several means to a bundle of humanistic ends. In analyzing behavior, especially that of households and their members, economic forces are greatly influenced and modified by social and psychological ones. Economic analysis depending on equilibrium tendencies ignores or belittles the lags in personal and organizational adjustment to change, and the durability of investment in infrastructure, buildings, and human capital.

There is another and fundamental schism between economic descriptions of 'rational' behavior and the realities of urban life. This divergence has led to two trends in urban modeling which are only now being resolved. In economics, similar actors in similar conditions make similar decisions – people shop at the nearest center, traffic follows the shortest route, and employees seek out the nearest jobs. In reality, some actors make long trips, and roundabout ones (but with diminishing frequency) for the same purposes which might be pursued using shorter and more direct trips. The 'rational' behavior is defined in economic models of optimizing choices, while diverse behavior is described, but not explained, in a number of ways by 'gravity models,' 'entropy maximization,' and 'discrete choice' – three very similar approaches.

By far the most important issue in model development has been this antinomy between an economic view that posits uniform behavior and a humanistic view which posits diversity. These two views underlie different lines of model development, and we will follow their reconciliation, though a recognition of ignorance – a reconciliation which clears the way for the resolution of other but less important difficulties.

Inventing a Paradigmatic Model

In the early 1950s, a number of factors favoring the discovery of a new paradigm in urban modeling came together: on the

ground, through the suburban redirection of American urbanism, and planning and building urban expressways; in technique, through the optimism of instrumental rationalism, seen in Operations Research and Management Science and buttressed by the commercial availability of computers; and in planning, through the widespread influence of the University of Chicago Program. In the middle of the decade, urban land planning began a slow movement in the direction of modeling (see Voorhees, 1959), and the Chicago Area Transportation Study (CATS, 1959) developed an innovative approach, drawing on many sources, which in retrospect must be considered a new paradigm for urban modeling.

A new paradigm in scientific or professional matters involves a framework for studying phenomena and defining their relationships. The importance of such innovation does not imply perfection; implementing and extending the paradigm and overcoming its imperfections may require decades. The essential novelty of this paradigm for analyzing the critical aspects of urban settlement systems extends far beyond transportation itself, and embodies the following elements:

1. Examine the system in all relevant detail, with massive data collection and thorough analysis of user behavior and system responses. (This was done in transportation studies by large-scale sampling of travel behavior and traffic conditions.)
2. Simulate user behavior over the entire system or subsystem, together with the system response, and with feedback in which the system response influences the user behavior. (The simulations in CATS were inventive and completely novel.)
3. Draw heavily on other disciplines that can contribute to the accuracy and realism of the simulations; use exact formulations wherever possible, and implement data management and simulation through computers where this is necessary and relevant. (CATS employed sociologists, planners, economists, engineers, and computer experts – and used computers for its most advanced innovations.)

Similar simulations could be imagined, and indeed have been undertaken, for the housing market, the labor market, the location of economic activity and public services, and the environment. Clearly, these subsystems are mutually interdependent, and in the final analysis their simulations would have to interact.

Despite its seminal influence, the CATS study has often been denied its paradigmatic role because of its shortcomings, most obviously inadequate attention to transit and failure to take into account the reaction of land uses to transportation. It shares many other problems with land-use modeling, as we will see in exploring them. It also embodied the schism between economics and behavior: in connecting residences and workplaces by trips it used a gravity model of dispersed behavior, but in assigning trips to the highway system it used the deterministic idea of taking only shortest routes. As a paradigm for transport analysis, it has been and is being reworked; in principle its methods can be traced even in divergent but related urban models.

Residential Models: Two Directions and Their Linkages

The major steps in land-use modeling are easily summarized. A series of theoretical models, initially of residential location, were generated by Wingo (1961), Alonso (1964), and Mills (1972) – with other contributions. These models all assumed a monocentric city (having a single employment center), which produced rings of settlement, each occupied by a single residential income class. Density declined and income rose with distance from the center. Residents were trading off commuting costs against space, under the constraints of their income. Mills's model dealt with zones in a linear programming framework, while the other models used continuous space.

The rigor and simplicity of these models inspired efforts to make them operational, hindered by their restrictive assumptions. Herbert and Stevens (1960) developed very early the idea of bid-rents (i.e., specifications of willingness to pay for various types of housing in various locations), and postulated a multicentered city with competition for land in a linear programming format. They did not operationalize the measures of preferences that would underlie a bidding process, or the connection between place of residence and place of work. The National Bureau of Economic Research (NBER) (Ingram et al., 1972) extended this work in a larger operational form, with actual employment centers, a diversity of choice of spatially distributed housing submarkets, and a final determination of residency by linear programming based on the cost of work-to-home travel. Building, remodeling, and redeveloping housing was in a separate supply model.

The alternative route of replicating consumer choice in housing and shopping through the use of a gravity model (that specified a declining probability of choice with increasing distance) was more realistic and more appealing to planning practitioners, but was initiated at the cost of disregarding some economic realities. The Lowry Model (Lowry, 1964) thus gave the appearance of modeling the behavior of all locators (including residents and their choices), and gave room for diversity through a gravity model. The same approach was widely adopted; for a review, see Goldner (1971). It was popularized by Putman (1982) in the USA, and in structure planning and by Echenique in the UK (reviewed in Echenique and Owers, 1994). Most of these efforts went beyond modeling residential choice to consider other locational trends, and other improvements were introduced to make the Lowry Model still more realistic. This model was thus a major contribution to the paradigm shift, by its introduction of a partially unified system of activity location.

The stage was set for a new type of advance that tended to bring these two trends much closer together. The necessity for multicentered models in applied analysis was firmly established. The NBER had introduced a certain amount of diversity through a probabilistic choice of housing markets. Echenique had realized the necessity for constraints on residential location, and began to introduce rents as an incentive for supply and a brake on choice in the face of income limitations. Three important events in the 1970s solidified these advances. In Geography, Wilson (1970) explained the gravity model by analogy with entropy (or constrained randomness) in physics. In Economics, McFadden (1973) formalized the idea of

'discrete choice,' which made diverse behavior acceptable to economists; the gravity model became a special case of this accepted theory. Independently, Anas (1975) showed in a seminal paper that interaction under a gravity model could be constrained by ability to pay in a competitive housing market, leading to an economic equilibrium solution, with market clearing.

All three models – gravity, discrete choice, and maximum entropy – can produce mathematically identical results. There is a common recognition that there are aspects of behavior which are not represented in models, and which contribute to a kind of statistical error in any deterministic explanation. The degree of error varies, and may or may not be susceptible to resolution by further analysis.

Complete Models

While residential choice models are the most important and perhaps the most interesting, a complete model as described at the outset would be a model suite containing numerous submodels. First, land development and its uses are determined by way of locational decisions that in turn are strongly influenced by transport considerations. Putman and Echenique have taken the lead over the last three decades in driving this lesson home – at least to the transport planning community. We will assume that any complete land-use model will embody or interface with an appropriate transportation model and use a forecast of transportation system changes. Second, even the most advanced residential models suffer from some of the general difficulties discussed below, and in particular, models of residential housing supply are weaker than those of demand. Third, models are needed for all land uses, most particularly manufacturing industry and trade and services. With the increasing shift in the overall economy toward business and consumer services, these sectors and their representation in models need thorough review.

There is a modest number of complete model suites publicly available worldwide (see Wegener, 1994, for a review). Putman's ITLUP and EMPAL models have been extended to METROPOLIS, which is widely used, and maintains the Lowry tradition of non-economic constraints. Its employment location is somewhat descriptive in style. Echenique has developed an elaborate input–output framework to generate goods traffic as well as personal trips, with extensive economic constraints. Anas (1987) is continuing to develop a range of economically rigorous and ambitious models, most recently for the New York Metropolitan Transportation Commission (in progress). New work is being undertaken which we cannot elaborate here, but special mention must be made of de la Barra, who follows Echenique in the Americas, and Waddell (2000), who is approaching the residential location from a microeconomic viewpoint in direct collaboration with state planning and transportation studies. Most model sets are provided on a made-to-order basis, which raises their cost and severely limits their availability.

In our judgment, the characteristic nonresidential models have various weaknesses in almost all model sets. The manufacturing sector is undergoing changes that are not yet fully understood and which may upset conventional wisdom. Business services have barely begun to be studied in the locational literature, let alone embodied in models. Retail trade is the best understood of these three fields, but inadequate use is being made of existing knowledge.

Comments on the State of the Art

It is apparent that considerable advances have been made in urban models, but that results are still short of what might be desired. There are still no established scholarly standards in the field, nor is there a clear link between research and practice. Limited tests of models on the same or similar data sets show little unanimity in their conclusions.

There are many social and equity issues being newly faced in planning of all sorts. Some of these have appeared in the USA in the implementation of the clean air act and its amendments, and related issues in transport planning have been litigated. It appears that there is inadequate disaggregation in most models, so that the behavior of minority groups and their reactions to policy inputs could be studied, while some aspects of disaggregation affect the evaluation of transport and planning policy. These needs conflict with the needs for simplicity, transparency, and speedy computation which arise in public discourse.

With increased computer capabilities, microsimulation is appearing as a possible approach to the disaggregation problem, but its relation to the speed of computation must be studied. Indications in the transportation planning field are that it will extend greatly the time required for analysis – and even today transportation models are the biggest bottleneck in operational land-use models.

The growth of urban modeling is linked closely with the expansion of computational power through both mainframe and personal computers. The first steps in modeling were conceptual, and computationally limited in their demands. Growing computer power has made it possible to develop and implement more accurate and more elaborate models. Computing for planning has been decentralized to the desks of planners and administrators. Computer users have readier access to large data sets, now organized in Geographic Information Systems. Research workers of all types have greatly increased power to investigate localized and sectoral phenomena, having better access to computation, data, and modeling methods. All of these investigations contribute to a web of knowledge that expands understanding of urban phenomena, and undergirds improved modeling. Examples include the work of Anas and Waddell discussed above, as well as of specialists like Landis, Cevero et al. (1999), and Shen (2000).

Bibliography

Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.

Anas, A., 1975. The empirical calibration and testing of a simulation model of residential location. Environment and Planning A 7, 899–920.

Anas, A., 1987. Modeling in Urban and Regional Economics. Harwood, Chur, Switzerland.

Burgess, E.W., 1927. The determinants of gradients in the growth of the city. Publications, American Sociological Society 21, 178–184.

CATS, 1959. Final Report, vol. 1. Chicago Area Transportation Study, Chicago.
Cevero, R., Rood, T., Appleyard, B., 1999. Tracking accessibility: housing opportunities in the San Francisco Bay area. Environment and Planning A 31 (7), 1259–1278.
Echenique, M., Owers, J. (Eds.), 1994. Research into Practice: The Work of the Martin Center in Urban and Regional Modeling. Environment and Planning B 21 (5), 513–514.
Goldner, W., 1971. The Lowry model heritage. Journal of the American Institute of Planners 37, 100–110.
Harris, C.D., Ullman, E.L., 1945. The nature of cities. Annals of the American Academy of Political and Social Sciences 242, 7–17.
Herbert, J.S., Stevens, B.H.A., 1960. Model of the distribution of residential activity in urban areas. Journal of Regional Science 2, 21–36.
Hoyt, H., 1939. The Structure and Growth of Residential Neighborhoods in American Cities. Federal Housing Administration, Washington, DC.
Ingram, G.K., Kain, J.F., Ginn, J.R., 1972. The Detroit Prototype of the NBER Urban Simulation Model. National Bureau of Economic Research, New York.
Lowry, I.S., 1964. A Model of a Metropolis, RM-4035-RC. The Rand Corporation, Santa Monica, CA.
McFadden, D., 1973. Conditional logit analysis and quantitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.
Mills, E.S., 1972. Urban Economics. Scott Foresman, Glenview, IL.
Putman, S.H., 1982. Integrated Urban Models: Policy Analysis of Transportation and Land Use. Pion, London.
Shen, Q., 2000. Spatial and social dimensions of commuting. Journal of the American Planning Association 66, 68–82.
Voorhees, A. (Ed.), 1959. Journal of the American Institute of Planners 25 (1).
Waddell, P., 2000. A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. Planning and Design, (Environment and Planning B) 27, 247–264.
Wegener, M., 1994. Operational urban models: the state of the art. Journal of the American Planning Association 60, 17–29.
Wilson, A.G., 1970. Entropy in Urban and Regional Modeling. Pion, London.
Wingo Jr., L., 1961. Transportation and Urban Land Use. Johns Hopkins University Press, Baltimore, MD.

Urban History

Rosemary Sweet, University of Leicester, Leicester, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous article by K. Ugawa, volume 24, pp. 16021–16026, © 2001, Elsevier Ltd.

Abstract

This article discusses the emergence of urban history as a subdiscipline in Western Europe and North America from its eighteenth-century origins and considers some of the methodological and theoretical approaches that have informed urban historical studies, ranging from the medieval to the contemporary periods. It considers the emergence of new paradigms of urban history in the rapidly urbanizing societies of South and East Asia and the implications for traditional urban historical scholarship. Finally, it provides an overview of the institutional and associational presence of urban history as a discipline.

Introduction

Urban histories have been written of towns and cities since antiquity: a sense of the past is a crucial component in the construction of a sense of place and local histories documenting a town's antiquity and its achievements continue to inform the wider historical study of urban places. The study of urban history, however, that is the study of towns as particular types of settlement or society, as agents of change, or as the locus of particular social, cultural, or political formations, is of more recent origin, and the emergence of a specific subgenre of academic study is unequivocally a twentieth-century phenomenon. In the twenty-first century, urban history has a clear status as a field of knowledge, although the extent to which it truly forms a distinct academic discipline, with its own methodologies or theories of knowledge, is more debatable. In the words of one urban historian, it is rather a 'focus for a variety of forms of knowledge' about the historical development of towns and cities (Dyos, 1982).

Origins

In the eighteenth century, historians and political economists, such as William Robertson and Adam Smith, recognized the importance of towns as historical agents in bringing about the rise of the modern European state and the system of global commerce that underpinned it, the collapse of feudalism, and the emergence of more widespread participatory politics (Robertson, 1769; Smith, 1776). Other historians drew attention to the role of towns in undermining feudalism by securing liberties and charters from monarchs who looked for commercial benefits and a counterweight against the influence of overmighty barons. They pointed also to the critical role urban society had played in facilitating the exchange of ideas, fostering creativity in the arts and encouraging the growth of civility. Thus, although the emergence of urban history as a field of research was still some distance away, the transformative role played by urban society in the broader narrative of historical development had been clearly identified. Political developments in the nineteenth century directed new attention to the historical urban experience: in Britain, the combination of political and municipal reform in the 1830s focused interest on the historical evolution of urban governance giving rise to some of the first synthetic studies of urban history, as opposed to the purely local study of a single town (Merewether and Stephens, 1835). The rise of democratic movements in Italy and France directed attention backward toward the communes of the medieval period, as historians sought models of more egalitarian and participatory politics in the annals of the medieval city republics (Sismondi, 1809–18). Building on the eighteenth-century recognition of the role of urban society in what would now be called the civilizing process, cities were also recognized as centers of cultural exchange and the arts, even as contemporaries denigrated the cultural philistinism of modern industrial and commercial urban centers. Thus, the Italian renaissance has always been discussed as an urban phenomenon seen through the histories of individual city states such as Florence or Siena (Burckhardt, 1860). But it was the unprecedented growth of the modern industrial city, the achievements of engineering and technology that lay behind it, and the extremes of wealth and poverty that it generated, which directed attention toward towns as the incarnation of modern capitalist modes of production and the nurseries of the modern age.

It was these changes that provoked the German sociologist, Max Weber, to write one of the most influential analyses of urban history of the twentieth century. In *The City* (published posthumously in 1921 and first translated into English in 1958), Weber set out a typological model for understanding the city and its historical development from the consumer/producer cities of antiquity, to the largely autonomous city states of the medieval period, to the age of 'great cities' in the modern era. According to Weber, the western city emerged in the medieval period, defined by its market functions as a center of exchange, but also by its administrative autonomy in feudal society, its system of corporate governance and modes of association that were independent of ties of kinship or personal fealty, and its status as a defensive organization given physical form in the city walls. Weber's conception of the city was an 'ideal type' and one that was shaped by his own understanding of the rise of democracy and participatory politics as the natural end of civil society. He excluded from his definition of the city both the settlements of Eastern Europe which were subject to feudal overlords and

'oriental' cities, such as Edo or Peking, despite their size, because they failed to display the requisite degree of autonomy and corporate governance. The model can be recognized today as Euro-centric, gender blind, and of limited use in an age of global cities and urban conglomerations or urban regions, but nonetheless it continues to carry value as a heuristic tool and remains an important reference point in historical discussions of urbanity.

At the same time, the rapid pace of urban growth and industrialization in Western Europe and the USA focused the attention of reformers, economists, and philanthropists upon the city as a source of social problems and unexpected failures, as well as the engines of prodigious wealth generation and expansion. The city became an object of study and an entity to be described, planned, and controlled. This line of enquiry is perhaps more accurately seen as laying the foundations for urban sociology where attention was focused upon finding explanations and solutions to the problems of modern urban society. This was a field of research in which the Chicago School of Sociology led the way in the 1920s and 1930s. Of particular significance to the development of urban historical studies would be the work of Robert Park and Ernest Burgess, who applied a theory of human ecology to urban development, arguing that competition for land and resources inevitably led to differentiated zones of deprivation and affluence (the concentric zones theory) and Louis Wirth who concentrated on the way in which human behavior adapts to both the urban environment and the social structure of cities, a theme most famously articulated in his essay 'Urbanism as a way of life' (Park et al., 1925; Wirth, 1938).

Emergence of a Subdiscipline

Building on these developments, historians on both sides of the Atlantic in the postwar era made a deliberate attempt to move away from the 'urban biography,' the study of a particular town. Rather, they set about identifying urban systems and typologies based on scale (capital cities, metropolitan cities, megalopolis, administrative units), function (ports, railway centers, textiles, shipbuilding and heavy engineering, iron and steel, spas, tourist centers), categories (industrial cities, new towns, Mediterranean cities, imperial cities, sun belt/rust belt cities), or ideology (capitalism, communism, utopian planning): analytical tools which continue to be a staple of urban historical analysis. They also focused upon common processes: the development of capitalism, urbanization, and the emergence of a distinctive urban consumer economy or the interaction between property, power, and social relations, using thematic studies based on transport, demography, property ownership, income, and occupational patterns. From this perspective, an understanding of the particular evolution of an urban settlement might be significant in the local context, but for urban historians it was more important to locate such information within a broader understanding of systems of power or social relationships. The urban historian, in other words, sought to generalize and rise above the unique particularities of local studies that so often sustained the civic pride of antiquarians and local historians. This pursuit of the general rule, the sense of what was typical, was very much in tune with the positivism of social science–based research in the 1950s and 1960s, particularly in the Anglo-phone world. This was the period when economic history was riding high and the advent of new technologies, such as main frame computing opened up new possibilities for the quantitative analysis of evidence and the creation of statistical data from which generalities could be drawn and probabilities asserted. Social science-based historical research exercised a profound influence on urban history, where the town represented the locus within which economic and social change could be studied, but also constituted a peculiar nexus of social, economic, and cultural factors, often described as the 'urban variable.'

Within this urban history agenda, medievalists and early modernists were often uncomfortable with theory and a form of quantitative empiricism which often depended upon evidence of a kind which simply did not survive from earlier periods and, crucially, an assumption that urban history is essentially the story of urbanization, and the parallel narratives of industrialization, democratization, and modernization. Their concerns were various: the medieval and early modern town was a very different phenomenon to that of the nineteenth or twentieth century. There was even considerable debate as to what constituted urban status in a period when few towns in Europe exceeded 10 000 people and many were far smaller: a size that no modern historian would recognize as urban. All towns must be understood in relation to their hinterlands, but in an essentially agrarian economy, medieval and early modern towns need to be studied as part of much broader systems and processes in which a distinctive urban experience is less easily isolated and defined than in the modern period. Quantification, which was axiomatic for so many of the questions posed by historians of the modern city, was often problematic: even establishing basic population figures for medieval and early modern urban settlements has been a challenge. The value systems of medievalists and early modernists also sat uncomfortably with those analyzed by urban historians of the modern period. Questions of legal status, systems of authority, or the organizing role of religion have loomed large. Concepts of class or systems of governance, by contrast, were far less important. Rather than urban growth, it was frequently decline and decay that attracted historians' attention. At times, it seemed that there was little in common between the towns of medieval Europe, which numbered a few thousand or less, and the great cities of the nineteenth and twentieth centuries of Europe and North America. This temporal divide hardened, with social scientists dominating urban historical writing in the period after 1750 and historians colonizing the earlier period – crudely a division either side of the main thrust of industrialization and modernization in Britain. This in itself is a (social science based) generalization, but the emergence of a separate Pre-Modern Towns group in 1987 in the UK was an expression of this intellectual and philosophical divide in Britain, and the disciplinary distinction between social scientists and historians in their respective approaches to towns and cities should be acknowledged.

More recent developments in historical writing have encouraged something of a rapprochement between historians across this chronological divide in the UK and Western Europe, at least. (In the USA, the study of colonial urban history, while vigorous, continues to be heavily overshadowed by studies of

the modern American city; beyond Europe and North America, a different temporal framework applies, see later text.) Since the 1980s, the rise of cultural history, the linguistic turn, post-modernism, and the influence of the later Annales School (as reflected in interest in the construction of historical mentalités, for example) cast doubt on the empiricism and the theoretical assumptions of the social science–based approaches. Quantitative methods, which were implicit in much early social science–based urban history, could be criticized for having reified towns as depersonalized, abstract entities which simply grow or decline, which experience problems or resolve them. Such an approach distorted the nature of historical inquiry by focusing on issues and questions for which quantifiable sources exist, and by marginalizing aspects of urban life which did not lend themselves so easily to quantification, such as women's work or the experience of the marginal poor. Urban history, as Griffiths and Jenner observed, was rendered as 'a series of graphs and tables, or a succession of maps' (Griffiths and Jenner, 2000). More recently, however, the opportunities offered by personal computing and the digitization of historical sources have altered the balance of historical enquiry yet again. The quantitative and qualitative can be combined ever more effectively and with greater sophistication due to the rapid steps taken in digital software and the emergence of a new field of digital humanities: the potential is evident in projects such as 'London Lives' which allow the researcher to combine the qualitative, microhistorical approach of researching the personal history of an individual with the ability to retrieve data on a macroscale through 15 different datasets created from 8 archives and comprising over 3.35 million names.

Interdisciplinarity

Urban history has defied definition, embracing as it does all aspects of the history of urban settlements whether as physical structures, economic entities, or communities of people. Its catholicity has also been reflected in its multi- and interdisciplinary approach to the past. The early gatherings of urban historians in North America and in Europe, respectively, included economic historians, sociologists, geographers, and architectural and planning historians. While geographers and planners are principally concerned with questions concerning the production, use, and meaning of space, historians share these interests but with a greater emphasis on the temporal dimension, extending their analysis across time. Influenced by theoretical works such as Henri Lefebvre's La production de l'espace (1974; translated into English, 1991) historians, and urban historians in particular, have sought to 'spatialize' understandings of identity, social relations, and human activity. The urban environment and the material fabric of streets, houses, and public buildings can no longer be seen as passive actors in the historical process: rather urban space both molded and was molded by the behavior and actions of urban inhabitants. The advent of more sophisticated mapping technologies through the development of GIS has allowed urban historians to open up new agendas in both the analysis of the spatial dimensions of urban society over time and also the presentation of research results. The potential to visualize data through mapping techniques is evident in projects such as the hypercities project at UCLA and Visualizing Urban Geographies, a collaboration between the National Library of Scotland and the University of Edinburgh.

Other historiographical developments, such as the 'linguistic turn,' and the school of cultural history which grew from it, directed urban historians toward the analysis of language and the way in which it shaped perceptions of identity and experience, particularly in terms of social status, class, and gender. This has been an approach that has lent itself to sources from antiquity to the twentieth century. The concentration of people within towns and cities and the rich potential for different forms of communication in urban society multiplied the opportunities to establish and create meaning or express identity. The emphasis on 'experience and identity' as a category of analysis has also opened up new avenues for urban historians to explore which were previously ignored or regarded as incidental, for example, the sensory dimensions and experience of the city – its sight, smell, touch, and sound – and the emotions that the city provokes.

Historians of medieval and ancient cities have always cooperated closely with archaeologists being heavily dependent upon the evidence of material remains for understanding of both the physical structure and social life of the town. Maurice Beresford notably drew upon archaeological evidence for his path-breaking study of medieval planned towns (Beresford, 1967): a key text that refuted earlier assumptions that urban planning was an attribute only of modernity. This kind of collaboration characterizes urban history wherever excavations have taken place, whether the British Iron Age town of Silchester (where archaeological evidence has transformed our understanding of British urban settlements prior to the Roman conquest (Fulford, 2012)) or the Japanese medieval castle town of Ichijo-dani, constructed in 1472 and destroyed in 1573. Historians and archaeologists studying classical and medieval urban settlements, influenced by cultural and spatial theory like their counterparts in later periods, have recently attempted to move beyond the description of topographical and architectural developments as ends in themselves, recognizing that the physical structure of towns was an embedded part of daily life, political institutions, and religious belief. Equally, archaeologists have become more interested in the activities within the physical structures as well as nonarchitectural or 'activity spaces' and the history of the everyday bringing the agenda of urban historians and archaeologists ever closer (Grig, 2013).

The convergence also reflects the 'material turn' in historical studies: in recent years, historians have been moving away from a text-centered approach to the past and have become increasingly appreciative of the insights to be derived from studies of material culture for later periods, as well as those for which little in the form of textual evidence remains. This greater openness to using material culture has also influenced urban history on a conceptual level by decentering the importance of human agency (drawing also on Actor Network Theory) and focusing attention on the agency of inanimate objects, buildings, spaces in shaping the behavior of individuals and society at large. Histories of the automobile and of roads, for example, demonstrate how these inanimate objects and structures have a powerful determinative impact on human agency in the

urban context, while historians of the nineteenth-century city have analyzed how the modern liberal society was brought into being through techniques of governance that depended on an infrastructure of water, sewerage, gas, and electricity (Gunn, 2011; Joyce, 2003; Otter, 2008).

As the temperature of the global village rises, a historically informed understanding of the interaction between cities and the environment has become increasingly pressing. In the latter half of the twentieth century, the field of urban environmental history has emerged as historians seek to understand urban society as the product of the relationship between human agency, technology, and nature and to recognize the fundamental and irreversible changes that modern urban society has wrought upon the globe (Chakrabarty, 2009). Although the urban form is often seen as 'unnatural,' that is, man-made in contrast to the 'natural' rural or agrarian landscape, environmental historians break down that distinction, demonstrating how towns and cities have exercised a profound influence over the surrounding countryside, reshaping 'nature': an approach that is exemplified in William Cronon's study of Chicago (Cronon, 1991). Another highly topical area of urban environmental research lies in the histories of pollution (water, air, noise) which almost always have an urban focus, as towns are the primary sources of pollutant activity. This approach to urban environmental history is often concerned with questions of social justice and inequality, and the interplay of race, gender, and class in urban environmental politics.

As the comments above might suggest, urban history always stands in danger of losing its coherence as a particular approach, due to its inherent interdisciplinarity. At the other extreme, there is also a tendency to assume that anything, by virtue of taking place in a town or city, can be said to qualify as urban history. For most urban historians, however, the urban context is a necessary though not a sufficient condition, because the urban dimension itself may be considered an independent actor in the process of change and development in towns and cities. As the British historian H.J. Dyos explained:

> the authentic measure of urban history is the degree to which it is concerned directly and generically with cities themselves and not with the historical events and tendencies that have been purely incidental to them … it is the study of the characteristically symbiotic relationships of their (cities') different characteristics, of the ways in which their components fitted together or impinged on other things that distinguishes urban historians from those who may be said merely to be passing through their territory. (Dyos, 1974)

Questions of identity and social experience, meaning and representation, environment and material culture equally transcend the chronological and the disciplinary boundaries. If the distinctiveness of urban history as a discipline is no longer so clearly demarcated in the twenty-first century, this is due to changes in the nature of historical discipline itself rather than to any crisis of confidence in the validity of the town or city as the object of historical research. The 'cultural' turn, for example, as noted above, has helped to dissolve the boundaries between many different subgenres of history, not just urban history. The cogency of urban history as a framework for historical inquiry, however, remains powerful and it continues to evolve in new directions.

Globalization

Urban history as a field of study emerged in Western Europe and North America largely in response to the specific historical circumstances of rapid urbanization, industrialization, and, in Western Europe, postwar reconstruction, which focused historians' and sociologists' attention upon urban society as a field of paramount importance for academic research. In the postcolonial era, there has been a global shift in attention as South America, Asia, and to a lesser extent Africa have undergone their own process of rapid urbanization. Just as in Europe and North America, much urban history in Asia, Africa, and South America was originally constituted of local studies and urban biographies, but increasingly urban historical scholarship in these areas has taken a more conceptual and comparative approach. The experience of urbanization in China and India, for example, challenges many of the paradigms by which Western urban history has been understood: the scale of the problems (and opportunities) faced by Asian cities dwarfs those of the Western experience of urbanization, and there is a telescoping of changes and processes for which there is equally no Western parallel: demographic expansion, industrialization, the emergence of democracy, the communications revolution – all of which in Western history followed sequentially – are taking place simultaneously and introducing an entirely new paradigm of urbanism. The models by which we have traditionally understood modernization and urbanization are being systematically upstaged.

Developments in the history of Chinese cities reflect those in the wider urban historiography. In the mid-twentieth century, a Eurocentric view of urban history saw Imperial China as static and autocratic, only spurred into modernization by the arrival of the West in the nineteenth century. It was argued that the form of Chinese cities reflected a long-established Confucian cosmology and that culturally they were indistinct from the countryside (Mote, 1971). Latterly, historians began to consider how internal events were equally, if not more, important than external influences in bringing change to China. This led to a reassessment of the history of cities, and two trends have dominated scholarship. The first views cities in Imperial China as dynamic, with complex commercial, cultural, social, and political identities, while the second explains the emergence of urban modernity as a result of both indigenous and external factors. Thus, for example, the former capital Kaifeng had a population of at least 1 million, and was characterized by lively markets, painting, poetry, and guidebooks, as well as sites of worship. Later, despite the disruption of dynastic overthrow, cities retained and developed their own distinct identities. Merchants of all classes formed large-scale trading networks that reached across the empire and abroad, while proto-industries in cotton, silk, and other sectors developed.

By the collapse of the Qing dynasty in 1911, treaty ports, the outposts of colonialism in China, had already been transformed by modern innovations such as new roads, sewage systems, department stores, and a vibrant print culture.

While it is true that Shanghai emerged as the center of Chinese urban modernity, historians now recognize that other cities were not simply copies of this metropolis. Histories of cities as far afield as Chengdu in Western China and Harbin in the north reveal how economic and political elites relied on local elements as much as global notions of urban modernity to create hybrid identities. Despite the disruption of the Anti-Japanese War of Resistance and the Communist revolution of 1949, we are now beginning to understand that even as urbanization stretches across whole regions, such distinctive identities continue to shape the development of cities in the twenty-first century.

In India and South Asia, the village and not the city has traditionally dominated both national self-perception and the outsider's view of the country. Weber dismissed Indian towns because they lacked the defining sense of unitary community that he identified in the West; rather they were simply loose federations of extended families, castes, or occupational groups (Gupta, 1981). For Indian nationalists, the city was associated with colonial rule and was regarded as a Western import, while the movement for independence was rooted in rural society, where Gandhi located the true India. Under imperial rule, it was assumed, a handful of presidency cities such as Madras, Bombay, Delhi, and Calcutta, and a clutch of railway towns had burgeoned; the others suffered deindustrialization and deurbanization throughout the nineteenth century. Postcolonial and subaltern studies reinforced this antiurban tendency, with a literature where rural society and peasant farmers, rather than the urban proletariat, were foremost. Serious interest in the city did not begin until the 1950s when sociologists and economists began to study the processes of urbanization in response to the influx of post-partition migrants to Indian cities after 1947. The field was dominated by sociologists and anthropologists, however, and a historical dimension was missing until the 1960s when historians began to recognize the importance of the city as a backdrop – if not as an agent or a 'variable' – in the emergence of modern politics and nationalism and in the processes of industrialization and modernization. By the 1970s, Indian historians, influenced by some of the same historiographical trends as in the West, began to focus more upon the city as the theater for the display of colonial power through architecture and the built environment, on the social history of the urban working classes, and on the public culture of colonial Indian cities. The foundation of the Urban History Association of India in 1978 bears witness to the emergent profile of the field. But it is the astonishingly rapid urbanization of the last 30 years – Indian cities now account for more than 10% of the world's urban population – that has directed attention toward the city and its colonial as well as the postcolonial past. Economic liberalization and globalization have created a newly wealthy and influential urban middle class: the same forces have also reshaped Indian cities and deepened the inequities of wealth and political opportunity and as such the city is the object of renewed historical study, shaped by both its colonial and postcolonial past. Even more so than in China, the urban history of India has had to engage with the historiography of the West, but what the more recent literature has demonstrated is the hybridity of Indian urbanism: colonial models of urban governance were imported, but they were also modified, resisted, and adapted in the colonial context, giving rise to a very different version of indigenous modernity and urbanism.

As South Asia, China, and parts of Africa undergo unprecedented urbanization and represent an ever-increasing proportion of the world's urban population, the study of urban history is losing its Eurocentric focus and taking on an increasingly global perspective: urbanity and modernity are being reconfigured and traditionally held assumptions regarding periodization and the origins of urban modernity in the West or the dominance of Western forms of urbanism have had to be refined. Questions of urban historical process such as the relationship between state power and urban governance or the growth of civil society and the rise of the public sphere have to be seen now in global perspective (Kidambi, 2007). The emergence of 'world cities' such as Bombay and Delhi, Shanghai, and Singapore, as well as London and Los Angeles as the crucial nodal points of global networks makes them key players in the emergent field of transnational history which looks at interactions and developments beyond the nation state and across continental boundaries. The field of transnational and global urban history is one which is rapidly expanding and opening up a rich seam of new questions and new directions.

Institutionalization

The institutionalization of urban history can be traced through the establishment of dedicated research chairs, research centers, formal associations of scholars, publications, and the presence of urban history within university curricula. The formal institutionalization of urban history goes back to the early twentieth century when the Stockholm Institute for Urban History was founded by the confederation of Swedish Towns in 1919, a development which led later to the establishment of the world's first chair in urban history at the University of Stockholm in 1953. The inspiration for the Swedish Center for Urban History, however, was rooted in the traditions of urban biography and localism (its remit was the ordering and publication of documents relating to the history of Swedish towns). A different version of urban history began to be formalized, however, at around the same time: across the Atlantic in 1953, a loose association of American academics agreed to meet informally to discuss urban history, deliberately distancing themselves from both the traditions of antiquarianism in the European tradition and the urban boosterism that was characteristic of North American cities. A regular series of conferences and publications followed in a movement that was identified as the 'new urban history' (Stave, 1977). Nine years after the North American Urban History group held its first meeting, H.J. Dyos of the University of Leicester and other colleagues belonging to the Economic History Society formally met as the Urban History Group 1962 in the United Kingdom. During the late 1950s, 1960s, and 1970s, associations of urban historians were established across Europe and further afield, such as the Arbeitskreis für landschaftliche deutsche Städteforschung (1959) and the Arbeitskreis für südwestdeutsche Stadtgeschichtsforschung (1960) and the Dutch Urban History Group in 1974. Beyond Europe, the Tokyo Study Group in Comparative Urban

History was formed with six charter members in 1971, for example, and the Urban History Association of India, as noted above, was established in 1978.

While most of these national associations were also interested in urban history beyond the national boundaries, there was also a strong international and collaborative impulse to urban historical research as countries across Europe had to come to terms with the devastation that World War II had wrought upon towns and cities and sought to put the conflict of the first half of the twentieth century behind them. In this spirit, the International Commission for the History of Towns (ICHT) was established in 1955: 'pour objet d'établir des contacts et de faciliter la coordination des recherches entre les spécialistes de toutes les branches de l'histoire des villes' (to establish contacts and to promote the coordination of research between specialists in all areas of the history of towns and cities). In a similar vein, the Institute für vergleichende Städtegeschichte was established in Munster in 1969. The most notable achievement of the ICHT to date is the support given to the Historic Towns Atlas project, the aim of which is to coordinate the production of atlases of towns plans, to common scales, accompanied by commentaries and subsidiary maps, to facilitate the comparative study of the history of European towns.

Comparative European, and increasingly comparative global urban historical research, is also fostered through the biennial conferences of the European Association of Urban Historians (EAUH), established in 1989. The lists of sessions of recent conferences give a fair indication of current trends in urban history with themes such as environment, heritage, deindustrialization, consumption, and construction of community all areas of particularly topical concern, obviously reflecting contemporary priorities in wider society. Recent conferences have also seen a marked increase in scholars from the former Eastern Bloc, who were formerly isolated from the traditions of Western European and American historical writing, for whom topics such as the legacy of industrialization and deindustrialization for modern cities and the problems of urban sustainability are particularly pertinent. These conferences also serve to showcase methodological advances, and in particular the progress that has been made with IT-assisted mapping and GIS as well as the construction of other digital resources. In the USA, the Urban History Association and Society for American City and Regional Planning History hold conferences in alternate years, and while their focus is chiefly upon the urban and planning history of North and South America, their meetings are similarly taking on an increasing global and transnational perspective.

Once scholars had identified a common interest in urban history and formed an association, publications soon followed, initially in the form of newsletters to disseminate information, but invariably leading to the emergence of journals dedicated to urban historical research. Thus the *Journal of Urban History* which first appeared in 1974 has its origins in the Urban History newsletter of the 1950s and 1960s. In the United Kingdom, H.J. Dyos was similarly inspired to issue his British newsletter until 1974 when its publication was formalized as the *Urban History Yearbook*. This, in turn, evolved into the *Urban History Journal* published by Cambridge University Press in 1991. While *Urban History* and *Journal of Urban History* share a comparative approach and global coverage, many other urban historical journals have emerged in the last 50 years with a stronger national or regional approach, including the German journal *Städteforschung* (first issued in 1970), *Stadsgeschiedenis* (covering Belgium and The Netherlands), *Città e Storia* (primarily Italy), and *Histoire Urbaine* (predominantly France) while in Japan, *The Comparative Urban History Review* focusing on Japanese urban history has been published twice a year since 1982.

The establishment of research centers dedicated to urban history at various universities has helped to ensure that the subject maintains an institutional profile and a presence in the curricula of undergraduate and postgraduate courses (Rodger and Menjot, 2006). H.J. Dyos became the first professor of Urban History in the United Kingdom; following his death in 1978, the tradition of urban history at Leicester continued, with the foundation of the Center for Urban History in 1985. Its current priorities reflect some of the key developments in urban history of recent years: an emphasis on the importance of transnational approaches and recognition of the importance of Asia (notably China and India) in modern urban history and the nature of urban heritage and the management of the past for cities in postindustrial society. Other specialist research centers for the study of urban history include the Center for Metropolitan History at the Institute of Historical Research, University of London, the Center for Urban History, University of Antwerp, the Danish Center of Urban History at the University of Aarhus, and the Center for Urban History of East Central Europe at Lviv.

See also: Bourgeoisie and Middle Classes, History of; Citizenship, Historical Development of; History and the Social Sciences; Modernization and Modernity in History; Social History; Weber, Max (1864–1920).

Bibliography

Bairoch, P., 1991. Cities in Economic Development: From the Dawn of History to the Present. Chicago University Press, Chicago, Illinois.

Beresford, M., 1967. New Towns of the Middle Ages: Town Plantation in England, Wales and Gascony. Praeger, New York and Washington.

Burckhardt, J., 1860. Die Cultur der Renaissance in Italien. Ein Versuch, Basel.

Burke, P., 1975. Some reflections on the pre-industrial city. Urban History Yearbook. Leicester University Press, Leicester, UK, pp. 13–21.

Chakrabarty, D., 2009. The climate of history: four theses. Critical Inquiry 35 (2), 197–222.

Cowan, A., Steward, J. (Eds.), 2007. The City and the Senses. Urban Culture since 1500. Ashgate, Aldershot.

Cronon, W., 1991. Nature's Metropolis: Chicago and the Great West. W.W. Norton, New York.

Dyos, H.J., 1974. Editorial. Urban History Yearbook 1, 3–10.

Dyos, H.J., 1982. Urbanity and suburbanity. In: Cannadine, D., Reeder, D. (Eds.), Exploring the Urban Past: Essays in Urban History by H.J. Dyos. Cambridge University Press, Cambridge, UK.

Fraser, D., Sutcliffe, A. (Eds.), 1983. The Pursuit of Urban History. Edward Arnold, London.

Fulford, M., 2012. Silchester and the study of Romano-British urbanism. Journal of Roman Archaeology (Suppl. 90).

Griffiths, P., Jenner, M. (Eds.), 2000. Londinopolis. Essays in the Cultural and Social History of Early Modern London. Manchester University Press, Manchester, UK.

Grig, L., 2013. Cities in the long late antiquity 2000–2012: a survey essay. Urban History 40 (3), 554–566.

Gunn, S., 2011. The Buchanan report and the problem of traffic in 1960s Britain. Twentieth-Century British History 22 (4), 521–554.
Gupta, N., 1981. Twelve years on: urban history in India. Urban History 8 (1), 76–79.
Hohenberg, P.M., Lees, L.H., 1995. The Making of Urban Europe 1000–1994, second ed. Harvard University Press, Cambridge, Massachusetts.
Joyce, P., 2003. The Rule of Freedom. Liberalism and the Modern City. Verso, London.
Kidambi, P., 2007. The Making of an Indian Metropolis. Colonial Governance and Public Culture in Bombay, 1890–1920. Ashgate, Aldershot.
Massard-Guilbaud, G., Rodger, R. (Eds.), 2011. Environmental Social Justice in the City. Historical Perspectives. White Horse Press, Cambridge.
Merewether, H.A., Stephens, A.J., 1835. The History of the Boroughs and Municipal Corporations of the United Kingdom, from the Earliest to the Present Time, vols. 3. London.
Mote, F.W., 1971. Intellectual Foundations of China. Knopf, New York.
Nead, L., 2000. Victorian Babylon: People, Streets and Images in Nineteenth-Century London. Yale University Press, New Haven and London.
Otter, C., 2008. The Victorian Eye. A Political History of Light and Vision in Britain, 1800–1910. University of Chicago Press, Chicago, Illinois.
Park, R., McKenzie, R.D., Burgess, E., 1925. The City: Suggestions for the Study of Human Nature in the Urban Environment. University of Chicago Press, Chicago, Illinois.
Robertson, W., 1769. The History of the Reign of the Emperor Charles V. with a View of the Progress of Society in Europe, from the Subversion of the Roman Empire to the Beginning of the Sixteenth Century, vols. 3. W. Strahan, London.
Rodger, R., Menjot, D., 2006. Teaching Urban History in Europe. L'enseignement de l'histoire urbaine en Europe. Centre for Urban HIstory, Leicester, UK.
de Sismondi, J.C.L., 1809–18. Histoire des républiques Italiannes du moyenne age, vol. 16. Paris.
Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell, London.
Stave, B., 1977. The Making of Urban History: Historiography through Oral History. Sage Publications, Beverley Hills and London.
Weber, M., 1958. In: Martindale, D., Neuwirth, G. (Eds.), The City. Free Press, Glencoe, Illinois.
Wirth, L., 1938. Urbanism as a way of life. American Journal of Sociology 44 (1), 1–44.

Relevant Websites

http://hypercities.ats.ucla.edu/ – Hypercities project.
http://www.londonlives.org/ – London Lives.
http://www.eauh.eu/ – The European Association of Urban Historians.
http://www.historictownsatlas.org.uk/hta-programme – The Historic Towns Atlas project.
http://www.historiaurbium.org/ – The International Commission for the History of Towns.
http://www.sacrph.org/ – The Society for American City and Regional Planning History.
http://uha.udayton.edu/ – The Urban History Association.
http://geo.nls.uk/urbhist/contact.html – Visualizing Urban Geographies.

Urban Informality: The Production and Regulation of Space

Ananya Roy, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urban informality is a polyvalent concept with roots in the debates of the 1970s about capitalism, informal work, and marginality. Most recently, urban informality has come to signify either an apocalyptic vision of a global epidemic of slums or a utopian vision of entrepreneurial poverty and improvised urbanism. Eschewing such conceptualizations, in this article, I present urban informality as a mode of the production and regulation of space. Urban informality is thus a heuristic device for understanding the territorial practices of state power and the processes of capital accumulation.

Urban informality has recently emerged as a prevalent concept in urban theory and research. However, the term is increasingly deployed in multiple and often divergent ways, rendering the concept somewhat chaotic. In addition, the distinctive genealogy of the term is often overlooked, thereby obscuring the historically specific political economy that the concept is meant to illuminate. In this article, I seek to explain the concept of urban informality in the following ways. First, I trace its genealogy to the debates of the 1970s about the structural logic of capitalism. Second, I show how today the idea of urban informality has been revived in the crucible of debates about urban citizenship. Third, I put forward the provocation that to truly consider the 'urban' character of informality requires us to think about the production and regulation of space. In particular, I argue that urban informality is best understood as a heuristic device that makes possible an analysis of the territorial practices of state power. Finally, I conclude by sketching emergent debates about informality, especially those that return us to the relationship between informality and the future of capitalism.

The Informal Sector

The term *informal* can be traced to the work of Keith Hart and the International Labor Organization (ILO) in the early 1970s. While modernization theory struggled to explain how a 'marginal mass' was not absorbed by industrialization, Hart showed that such forms of marginality and informality were a structural feature of urban economies. Writing in the context of Accra, Hart (1973: 61, 68) identified a "world of economic activities outside the organised labor force" carried out by an "urban sub-proletariat." Hart designated these activities as informal. At more or less the same time, the ILO (1972) in Kanbur, (2009) defined the informal sector as the activities of "petty traders, street hawkers, shoeshine boys and other groups 'underemployed' on the streets of big towns, and includes a range of wage-earners and self-employed persons, male as well as female." Hart's conceptualization of informality has been commonly interpreted as "the relationship of economic activity to intervention or regulation by the state" (Kanbur, 2009: 5). This is accurate. In a recent reflection on his 1973 paper, Hart (2006: 25, in Kanbur, 2009: 2–3) notes that, following Weber, he had argued "that the ability to stabilise economic activity within a bureaucratic form made returns more calculable and regular for the workers as well as their bosses. That stability was in turn guaranteed by the state's laws, which only extended so far into the depths of Ghana's economy." Informal work lay outside this realm of calculable and stable transactions and was thus erratic, with low returns.

In many ways, Hart's argument is a companion to the conceptualization proposed by dependency theorists: that the state maintains the informal economy in unregulated form in order to subsidize a system of global capital accumulation (Portes et al., 1989). Indeed, a whole generation of social science research, much of it produced in Latin America, was to explode the 'myth of marginality' (Perlman, 1977) and show how informal work and habitat was an integral part of the capitalist city, especially the 'dependent city' at the periphery of the world-system (Castells, 1983). But Hart went further. Unlike later interpretations of the informal sector, he "did not identify the informal economy with a place or a class or even whole persons" (Hart, 2006: 25). Instead, he argued that many of Accra's residents sought to forge a multiplicity of livelihoods and income opportunities, in other words that informality was a generalized condition, a way of life if you will. I will return to this idea of informality as a way of life later in this article.

Hart's conceptualization of the informal sector closely resonates with more contemporary interpretations of informality, especially scholarship that traces the formation of a "new urban marginality"–not only in the Global South but also in Europe and the Americas. Wacquant's work (1996, 1999, 2007), for example, documents the emergence of an "advanced marginality" that is linked to the "territorial stigmatization" faced by residents of marginalized spaces: the ghetto, the banlieue, the favela. Similarly, Auyero (2000) charts the emergence of the hyper-shantytown. Such research is united in its emphasis on the connections between such advanced marginality and the hollowing out of economies and welfare states through neoliberal capitalism. The hyper-shantytown is thus produced by 'hyper-unemployment' (Auyero, 1999), a systematic process of deproletarianization and labor informalization. While the theorists of advanced marginality acknowledge the poverty-targeting efforts of the state (for example, in Brazil, the upgrading of favelas through the provision of services), they insist that such programs are minor palliatives in the face of a massive structural crisis. In particular, they argue that the

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74051-7

communities of the urban poor are now overwhelmed by violence: the violence of state repression, the symbolic violence of stigma and discrimination, and the material violence of poverty and unemployment (Perlman, 2003). In short, yesterday's 'slums of hope' are today's 'slums of despair' (Eckstein, 1990: 165).

Perhaps the most prominent interpretation of informality as the structural logic of capitalism is Mike Davis's (2004) apocalyptic account of a planet of slums. For Davis (2004: 14–15), informal urbanization is a stark manifestation of 'overurbanization' or "urbanization without growth," which in turn is "the legacy of a global political conjuncture – the worldwide debt crisis of the late 1970s and the subsequent IMF-led restructuring of Third World economies in the 1980s." Davis (2004: 23) thus designates this world-system as a "planet of slums," a warehousing of the "surplus humanity" released by deproletarianization and agricultural deregulation in hazardous and miserable forms of urban settlement. He concludes that "informal survivalism" is "the new primary mode of livelihood in a majority of Third World cities" (Davis, 2004: 24).

Urban Citizenship

Mike Davis's account of a planet of slums is accompanied by considerable anxiety about the urban politics of informality. Davis (2004: 28) asks: "To what extent does an informal proletariat possess that most potent of Marxist talismans: 'historical agency'?" He concludes that "uprooted rural migrants and informal workers have been largely dispossessed of fungible labour-power, or reduced to domestic service in the houses of the rich" and that thus "they have little access to the culture of collective labor or large-scale class struggle." Davis's lament rehearses a much older argument presented by dependency theorists. For example, in the seminal text, *The City and the Grassroots*, Castells (1983) presents an ambitious theory of urban social movements. Studying both formal political organizations and the mobilizations of informal squatter communities, Castells acknowledges the central role of politics in the transformation of the capitalist city but also analyzes the limits of such politics. In particular, he designates the populist politics of squatter communities as a symptom of the "dependent city," a "city without citizens." He argues that while squatters mobilize to secure access to land, services, jobs, and at times even tenure, they are simultaneously co-opted into systems of political patronage. Thus, they are clients rather than citizens, disciplined subjects of urban populism rather than active agents of structural change.

Such understandings of urban politics are directly contradicted by a new genre of scholarship that positions informality as the material basis of formations of urban citizenship. For example, in his work on South Asia, Chatterjee makes a distinction between a bourgeois civil society and the political society of the urban poor. Civil society is bourgeois society and in the Indian context, an arena of institutions and practices inhabited by a relatively small section of people able to make claims as fully enfranchised citizens. By contrast, political society is the constellation of claims made by those who are only tenuously and ambiguously right-bearing citizens. The "paralegal" practices and negotiations of political society is for Chatterjee the politics of much of the people in most of the world: "The paralegal then, despite its ambiguous and supplementary status in relation to the legal, is not some pathological condition of retarded modernity, but rather part of the very process of the historical constitution of modernity in most of the world" (Chatterjee, 2004: 75). His work echoes that of Appadurai, who sees in the political actions of Mumbai's slum-dwellers a form of deep democracy, a counter-governmentality that challenges the top-down strictures of planning and development. Similarly (Benjamin, 2008: 721), makes the case for occupancy urbanism, the appropriation of urban space by poor groups. Most recently, Holston (2007: 4) has argued that the auto-constructed peripheries of Brazilian cities are the site of new imaginations and practices of urban citizenship, what he calls insurgent citizenship or where citizenship is produced "not primarily through the struggles of labor but through those of the city."

What is at stake in such work is the conceptualization of informality as an alternative urban order, a different way of organizing space and negotiating citizenship. AlSayyad (2004) thus presents the idea of "urban informality as a way of life." His title refers to the classic 1938 essay by Wirth, "Urbanism as a Way of Life." AlSayyad argues that today urban informality is a generalized urban condition. In a similar vein, Bayat (2007) argues that informality is the "habitus of the dispossessed." In earlier work, Bayat (2000), working in the context of Middle Eastern cities, outlined the repertoire of tactics through which urban 'informals' appropriate and claim space. According to Bayat, this "quiet encroachment of the ordinary" by subaltern groups creates a street politics that shapes the city in fundamental ways. Bayat's analysis is similar to Michel de Certeau's (1984) conceptualization of the "practice of everyday life" as a set of tactics that can undo the oppressive grid of power and discipline. In similar fashion, Simone (2006a, b) presents the African city as pirate towns, where urban residents develop forms of everyday practice that allow them to operate resourcefully in underresourced cities. This is a context of crisis, where "production possibilities" are severely limited; but this is also a context where "existent materials of all kind are to be appropriated" (Simone, 2006a,b: 358). While Simone does not use the term *informality*, his analysis suggests that he is describing practices that can be designated as such:

> African cities are characterized by incessantly flexible, mobile, and provisional intersections of residents that operate without clearly delineated notions of how the city is to be inhabited and used. … These conjunctions become an infrastructure—a platform providing for and reproducing life in the city. (Simone, 2004: 407–408)

Such conceptualizations of urban informality are also shaping the professions of architecture and urban planning. Indeed, informality is the inspiration for the imagining of urban futures. It is thus that star architect Rem Koolhaas interprets the urbanism of Lagos as a 'culture of make-do' (Enwezor, 2003: 116). In his encounter with Lagos, part of Harvard's Project on the City, Koolhaas is taken with the inventiveness of its residents as they survive the travails of the megacity. He sees such experimental responses as generating "ingenious, critical alternative systems," a type of 'self-organization' creating "intense emancipatory zones" (Godlweski, 2010). Koolhaas draws the following conclusion: "Lagos is not catching up with us. Rather, we may

be catching up with Lagos." In similar fashion, in 2011, the Center for Architecture in New York hosted an exhibition titled "Jugaad Urbanism: Resourceful Strategies for Indian Cities." Curated by Kanu Agrawal, the collection of models, photographs, and video installations sought to highlight the 'inspired, duct-taped ingenuity' of communities of poverty. From tin-can canopies to frugal latrines, the exhibition was meant to capture how the poor 'make-do' and how such forms of making-do are a new idiom of urban ingenuity and entrepreneurialism:

> Set in the radically uneven urban landscapes of Delhi, Mumbai, Ahmedabad and Pune, India, Jugaad Urbanism explores how the energy of citizens "making-do" is translated by architects, urban planners, and governmental and nongovernmental entities into efficient and inventive strategies for sustainable urban growth. From energy-generating spinning wheels to the extensive skywalks of Mumbai—the exhibition highlights how 'jugaad' interventions (a term in Hindi used to describe an innovative, resourceful approach) are challenging traditional spatial hierarchies and mechanistic planning principles.
> (http://cfa.aiany.org/index.php?section=upcoming&expid=136)

The Jugaad Urbanism exhibition is an example of a new global imaginary about poverty, which views the economies of the poor as economies of entrepreneurship and dynamic informality. I will discuss the implications of such an approach in the concluding section of this article.

The Production and Regulation of Space

In his initial analysis of the informal sector, Keith Hart put forward a cautionary note that is as valid today as it was in 1973:

> Socialists may argue that foreign capitalist dominance of these economies determines the scope for informal (and formal) development, and condemns the majority of the urban population to deprivation and exploitation. More optimistic liberals may see in informal activities, as described above, the possibility of a dramatic bootstrap operation, lifting the underdeveloped economies through their own indigenous enterprise. Before either view—or a middle course stressing both external constraint and autonomous effort—may be espoused, much more empirical research is required. (Hart, 1973: 88–89)

Taking up Hart's caution against structuralist and populist views of the informal sector, I now turn to a different conceptualization of urban informality. With particular attention to the "urban" character of informality, I suggest that informality must be understood as a mode of producing and regulating space. In particular, urban informality is best understood as a heuristic device that makes possible an analysis of the territorial practices of state power. Such a definition has three implications. First, it means that informality is not a separate and bounded sector of unregulated work, enterprise, and settlement but instead a pervasive socio-spatial logic. Indeed, as (AlSayyad, 2004: 25) has noted, what requires explanation is not so much informality as a way of life as it is formality, as a 'new mode' of urbanism, one that "was introduced to organize urban society only in the nineteenth century." Second, the entanglement of formality and informality suggests that no legal or regulatory boundaries can be drawn between these domains. The formal and the legal are perhaps better understood as fictions, as moments of fixture in otherwise volatile, ambiguous, and uncertain systems of spatial regulation. Indeed, in my work on Calcutta (Roy, 2003), I have shown that informality exists at the very heart of the state and is an integral part of the territorial practices of state power. This builds on the previous work of (Meagher, 1995: 259, 279) who calls for a conceptual shift "from informality represented as a marginalized sector to 'informalization' conceived as a wider economic response to crisis," and notes that such a process of "informalization" does not happen "outside the state" but rather is a "socioeconomic restructuring instigated by the state." Finally, to conceive of informality as a mode of producing and regulating space implies that informality cannot be seen to be synonymous with poverty. The splintering of urbanism does not take place at the fissure between formality and informality but rather, in fractal fashion, within the informalized production of space. A closer look at the metropolitan regions of much of the world indicates that informal urbanization is as much the purview of wealthy urbanites and suburbanites as it is that of squatters and slum-dwellers. The urgent question then becomes how and why elite informalities are valorized and even regularized while subaltern informalities are criminalized.

Such issues have been taken up in considerable detail in recent scholarship on urban informality. Gidwani (2006: 12) describes how unauthorized farmhouses, built illegally on agricultural land, dot the periphery of Delhi. "The owners of these farmhouses, notables on Delhi's social circuit, are often the very same people who profess an urban environmentalism and advocate the 'cleaning up' of Delhi." The "cleaning up" of Delhi has usually entailed the dispossession and the displacement of the poor through slum evictions and demolitions, what Baviskar (2003) provocatively titles, "bourgeois environmentalism." Similarly, in the case of Mumbai, Weinstein (2008: 23) highlights the construction of new shopping centers, many of them built illegally by the city's "largest and most notorious mafia organization, on land belonging to the state government's public works department." Indeed, as she notes, this mafia organization was running "a highly profitable property development business since the mid-1990s, constructing retail and residential buildings in other parts of Mumbai, as well as in Dubai and Karachi." It is thus that Ghertner (2008) notes that almost all of Delhi violates some planning or building law, such that much of the construction in the city can be viewed as "unauthorized." He poses the vital question of why some of these areas are designated as illegal and worthy of demolition, while others are protected and formalized. How and why is it that the law has come to designate slums as 'nuisance' and the residents of slums as a 'secondary category of citizens,' those that are distinguished from 'normal,' private property-owning citizens? It is through such designations that differentiated spatial value is produced and that space comes to be produced and regulated. Similarly, Holston (2007: 228) notes that Brazilian cities are marked by an "unstable relationship between the legal and illegal." While it may seem obvious and apparent that the urban poor are engaged in an informal and illegal occupation of land, much of the city itself is occupied through the 'misrule of law': "Thus in both the wealthiest and the poorest of Brazilian families we find legal landholdings that are at base legalized usurpations" (Holston, 2007: 207). Who then is authorized to (mis)use the law in such ways to declare

property ownership, zones of exception, and enclaves of value? (Yiftachel, 2009: 249–250) characterizes such processes as "the new politics of *gray spacing*" whereby subordinated groups are relegated to the "pseudo-permanent margins of today's urban regions, which exist partially outside the gaze of state authorities and city plans" while elite encroachers are allowed to colonize urban space.

What is at stake in such a genre of research is a relational understanding of urban informality. Yiftachel (2009: 250) correctly notes that gray spacing is "a ceaseless process of producing social relations." Such relationalities are starkly manifested in the territorial practices of the state. In fact, to describe urban informality as the production and regulation of space requires careful attention to the territorialized flexibility of the state. It is of course the state that determines what is informal and what is not, thereby generating an uneven geography of spatial value. It is the state that considers the informal subdivisions of corporate economies outside of Mexico City or Bangalore legal and formal while rendering illegal shantytowns in the same peripheries. It is the state that allows sprawling, elite 'farm houses' on the edges of Delhi to function legally as appendages of the agrarian land laws while criminalizing and violently demolishing slums and squatter settlements throughout the city. Such forms of state power often operate under the sign of 'public interest.' As Bhan (2009) demonstrates, judicial rulings seeking to assert the public interest have recently created a slate of slum clearances in Delhi. These rulings have hinged on judicial interpretations of encroachment that have likened the urban poor to illegal trespassers and thus as unworthy of legal and constitutional protection (Bhan, 2009: 139). Such forms of desubjectivation are not only expressions of punitive power but also the practice of distinction. Following Bourdieu (1979), distinction is the "judgment of taste" and thus the social practice of class power. It regulates the "economy of cultural goods" and produces an "aristocracy of culture" (Bourdieu, 1979: 1, 11). This social practice of class power is amply evident in the differentiated character of urban informality, or what Ghertner (2011) calls "rule by aesthetics." In his analysis of the production of space in Delhi, he shows how the judgment of taste determines which urban developments are granted amnesty, and even "heralded as monuments of modernity" (Ghertner, 2008: 66) while others are labeled "nuisance" and thus criminalized and demolished. Ghertner (2010: 185) thus argues that it is necessary to "expand our understanding of the epistemology of government to include attention to a more diverse array of governmental technologies, some more aesthetic than strictly calculative." This type of 'aesthetic normativity,' a phrase I borrow from Ghertner (2010: 199) is of course the mirror image of the aesthetic vision that marks 'jugaad urbanism' and its celebrations of the 'duct-taped ingenuity' of the informal city.

Frontiers: Informality and Capitalism's Future

In a profound and poetic intervention in the hegemony of capitalism, Gibson-Graham (2008) puts forward the call to "perform new economic worlds." Seeking to recover "diverse economies," they foreground "projects of economic autonomy and experimentation" (Gibson-Graham, 2008: 614). These projects include "squatter, slum-dweller, landless, and cohousing movements, the global ecovillage movement, fair trade, economic self-determination, the relocalization movement, community-based resource management, and others" (Gibson-Graham, 2008: 617). Such an "ontology of economic difference" (Gibson-Graham, 2008: 614) is potentially radical. But it also demands critical scrutiny. As Samers (2005: 876) notes, many of the diverse economies celebrated by Gibson-Graham are also informal economies. And these informal economies are in fact fully capitalized domains and often sites of intense exploitation. In fact, there is an uncanny resemblance between Gibson-Graham's postcapitalist utopias and the recent rediscovery of the informal sector by right-wing libertarians.

Perhaps, the most influential articulation of such a paradigm comes in the work of Hernando de Soto who presents informality as an economy of 'heroic entrepreneurs.' For de Soto, the Third World slum is rich in assets, albeit in the defective form of dead capital. The 'mystery of capital' is how such assets can be transformed into liquid capital, thereby unleashing new frontiers of capital accumulation. It is important to note that de Soto presents capital not as a relationship of production but rather as a 'representational process.' For him the poor are not workers but rather informal entrepreneurs who own assets, albeit those rendered impotent by faulty representation. 'Marx,' De Soto writes, "would probably be shocked to find how in developing countries much of the teeming mass does not consist of oppressed legal proletarians but of oppressed extralegal small entrepreneurs with a sizeable amount of assets." In de Soto's mediations, the diverse and alternative economies imagined by Gibson-Graham become new frontiers of capital accumulation.

In my recent work (Roy, 2010), I have designated these efforts to convert economies of poverty into frontiers of capital accumulation as poverty capitalism. In particular, I have drawn upon the case of microfinance to illustrate this process. Microfinance, the provision of financial services to the poor, is a highly popular poverty alleviation tool. Premised on the idea that the poor are inherently entrepreneurial, microfinance is an example of what I call "bottom billion capitalism," a set of dispersed but coherent efforts to construct, and make productive, a global economy where the world's bottom billion–the billion or so people living under conditions of poverty–are integrated into circuits of capital accumulation. This is what (Gates, 2008: 23) has called 'creative capitalism': "an attempt to stretch the reach of market forces … to bring far more people into the system – capitalism – that has done so much good in the world." Gates's vision is striking and seductive, for he is able to position poverty as not only the primitive other of the modern economy but also as its future. "There are markets all over the world that businesses have missed," he asserts (Gates, 2008: 27), and the poor constitute a particularly important and lucrative market, a 'billion bootstraps.' Put another way, informality, once the realm of deproletarianized survival, or imagined as an ontology of economic difference, is now the economy of bootstrap entrepreneurship, ready to be monetized and capitalized.

But such capitalization remains fragile. If nothing else, it remains subject to appropriations by the very informality celebrated by de Soto. These are diverse economies being

produced in the interstices of global capitalism, what Iquierdo (2012) calls "Capitalismo Amarillo" or "yellow capitalism." Documenting the activity of 'pirate traders' in the Tepito street market in Mexico City, Iquierdo traces circulations of mass-produced goods and 'original fakes' between Mexico City and the Special Economic Zone of Guangzhou, China. This is a counterfeit economy, one that belies the naïve populism of bottom billion capitalism as well as the romanticism of postcapitalist utopias. This is a thoroughly commodified and differentiated economy with a multiplicity of livelihoods and spatial strategies that would have seemed familiar to Keith Hart. Yet, it provides glimpses of tactics, appropriations, and disruptions, allowing subaltern informality to find a place in the world. This seems more convincing as an image of the future than the apocalypse of a planet of slums or a world of self-organized entrepreneurs.

Bibliography

AlSayyad, N., 2004. Urban informality as a way of life. In: Roy, A., AlSayyad, N. (Eds.), Urban Informality: Transnational Perspectives from the Middle East, Latin America, and South Asia. Lexington Books, Lanham.

Auyero, J., 1999. 'This is a lot like the Bronx, isn't it?': lived experiences of marginality in an Argentine slum. International Journal of Urban and Regional Research 23, 45–69.

Auyero, J., 2000. The hyper-shantytown: Neoliberal violence(s) in the Argentine slum. Ethnography 1, 93–116.

Baviskar, A., 2003. Between violence and desire: space, power, and identity in the making of metropolitan Delhi. International Social Science Journal 55 (175), 89–98.

Bayat, A., 2000. From 'dangerous classes' to 'quiet rebels': the politics of the urban subaltern in the Global South. International Sociology 15 (3), 533–557.

Bayat, A., 2007. Radical religion and the habitus of the dispossessed: does islamic militancy have an urban ecology? International Journal of Urban and Regional Research 31 (3), 579–590.

Benjamin, S., 2008. Occupancy urbanism: Tadicalizing politics and economy beyond policy and programs. International Journal of Urban and Regional Research 32 (3), 719–729.

Bhan, G., 2009. This is no longer the city I once knew. Evictions, the urban poor and the right to the city in millennial Delhi. Environment and Urbanization 21 (1), 127–142.

Bourdieu, P., 1979. Distinction: A Social Critique of the Judgment of Taste (translated by R. Nice) (1984 edition). Harvard University Press, Cambridge, MA.

Castells, M., 1983. The City and the Grassroots. University of California Press, Berkeley, CA.

Chatterjee, P., 2004. The Politics of the Governed: Reflections on Popular Politics in Most of the World. Columbia University Press, New York.

Davis, M., 2004. Planet of slums: urban involution and the informal proletariat. New Left Review 26, 5–34.

De Certeau, M., 1984. The Practice of Everyday Life. University of California Press, Berkeley, CA (S. Rendall, translator).

Eckstein, S., 1990. Urbanization revisited. Inner-city slum of hope and squatter settlement of despair. World Development 18, 165–181.

Enwezor, O., 2003. Terminal modernity: Rem Koolhaas' discourse on entropy. In: Patteeuw, V. (Ed.), What Is OMA: Considering Rem Koolhaas and the Office for Metropolitan Architecture. NAi Publishers, Rotterdam, the Netherlands.

Gates, B., July 2008. How to fix capitalism. Time 31, 23–29.

Ghertner, A., 2010. Calculating without numbers: aesthetic governmentality in Delhi's slums. Economy and Society 39 (2), 185–217.

Ghertner, A., 2011. Rule by aesthetics: world-class city making in Delhi. In: Roy, A., Ong, A. (Eds.), Worlding Cities: Asian Experiments and the Art of Being Global. Wiley-Blackwell (Studies in Social and Urban Change Series), Chichester, UK.

Ghertner, A., May 2008. Analysis of new legal discourse behind Delhi's slum demolitions. Economic and Political Weekly 17, 57–66.

Gibson-Graham, J.K., 2008. Diverse economies: performative practices for other worlds. Progress in Human Geography 32 (5), 613–632.

Gidwani, V., 2006. Subaltern cosmopolitanism as politics. Antipode 38 (1), 7–21.

Godlweski, J., 2010. Alien and distant: Rem Koolhaas on film in Lagos, Nigeria. Traditional Dwellings and Settlements Review XXI (II), 7–20.

Hart, K., 1973. Informal income opportunities and urban employment in Ghana. Journal of Modern African Studies 11 (1), 61–89.

Hart, K., 2006. Bureaucratic form and the informal economy. In: Guha-Khasnobis, B., Kanbur, R., Ostrom, E. (Eds.), Linking the Formal and Informal Economy: Concepts and Policies. Oxford University Press, New York, pp. 21–35.

Holston, J., 2007. Insurgent Citizenship: Disjunctions of Democracy and Modernity in Brazil. Princeton University Press, Princeton, NJ.

ILO, 1972. Incomes, Employment and Equality in Kenya. ILO, Geneva.

Iquierdo, J., 2012. "Capitalismo Amarillo: Special Economic Zone" Exhibition at Manifesta 9, Genk, Belgium. http://capitalismoamarillo.net/exposiciones/.

Kanbur, R., 2009. Conceptualising Informality: Regulation and Enforcement. Department of Economics and Applied Management, Cornell University, WP 2009–11.

Meagher, K., 1995. Crisis, informalization and the urban informal sector in Sub-Saharan Africa. Development and Change 26 (2), 259–284.

Perlman, J., 1977. The Myth of Marginality. University of California Press, Berkeley, CA.

Perlman, J., 2003. Marginality: from myth to reality in the favelas of Rio de Janeiro. In: Roy, A., AlSayyad, N. (Eds.), Urban Informality: Transnational Perspectives from the Middle East, Latin America, and South Asia. Lexington Books, Lanham, MD.

Portes, A., Castells, M., Benton, L., 1989. The Informal Economy. Johns Hopkins University Press, Baltimore, MD.

Roy, A., 2003. City Requiem, Calcutta: Gender and the Politics of Poverty. University of Minnesota Press, Minneapolis, MN.

Roy, A., 2010. Poverty Capital: Microfinance and the Making of Development. Routledge, New York.

Samers, M., 2005. The myopia of 'diverse economies', or a critique of the 'informal economy'. Antipode 37 (5), 875–886.

Simone, A., 2004. People as infrastructure: intersecting fragments in Johannesburg. Public Culture 16 (3), 407–429.

Simone, A., 2006a. Pirate towns: reworking social and symbolic infrastructures in Johannesburg and Douala. Urban Studies 43 (2), 357–370.

Simone, A., 2006b. Urban circulation and the everyday politics of African urban youth: the case of Douala, Cameroon. International Journal of Urban and Regional Research 29 (3), 516–532.

Wacquant, L., 1996. The rise of advanced marginality: notes on its nature and implications. Acta Sociologica 39, 121–139.

Wacquant, L., 1999. Urban marginality in the coming millennium. Urban Studies 36, 1639–1647.

Wacquant, L., 2007. Ghetto, banlieue, favela, etc.: Tools for rethinking urban marginality. In: Urban Outcasts: A Comparative Sociology of Advanced Marginality. Polity, Cambridge, United Kingdom.

Weinstein, L., 2008. Mumbai's development Mafias: globalization, organized crime and land development. International Journal of Urban and Regional Research 32 (1), 22–39.

Yiftachel, O., 2009. Theoretical notes on 'gray cities': the coming of urban apartheid? Planning Theory 8 (1), 88–100.

Urban Internal Spatial Structure

Yan Song, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines the term 'urban internal spatial structure' and summarizes a schematic view of research areas related to urban internal spatial structure. Specifically, discussions are included of the observation of urban internal spatial structure, identification of driving forces of urban spatial structure, and investigation of the connection between urban spatial structure and a variety of outcomes, such as land and housing market performances, environmental safety and quality, and individual well-being.

Introduction

The term 'urban internal spatial structure' loosely encompasses dimensions such as urban development patterns, urban form, location of land uses, and location of firms and households. The way that urban spaces are arranged affects many aspects of how cities function and has implications for economic efficiency, accessibility, environmental sustainability, safety, social equity, and cultural creativity (Jacobs, 1961). For this reason, the subject has caught abundant attention by a variety of researchers. In recent decades, increasing public concern over the set of negative impacts of urban sprawl – the loss of open land resources, longer commutes, increased greenhouse gas emissions, and lower levels of physical activity – has led to policy efforts that seek to change the course of urban growth patterns. Notably, a range of urban and city planning efforts, such as the smart growth movement that emerged in the 1980s across the world, openly sought to avoid urban sprawl and improve public health through reshaped city forms, and to create good urban structures to help policymakers and the public choose desirable patterns of urban development. These movements still popularly appear in the dialog between research scholars, urban planners, policymakers, and the general public.

This article aims to summarize a schematic view of research areas related to urban internal spatial structure, including (1) observation of urban internal spatial structure; (2) identification of driving forces of urban spatial structure; and (3) investigation of the connection between urban spatial structure and a variety of outcomes, such as land and housing market performances, environment safety and quality, and individual well-being (Figure 1). The following sections synthesize main research topics and findings in each of these categories.

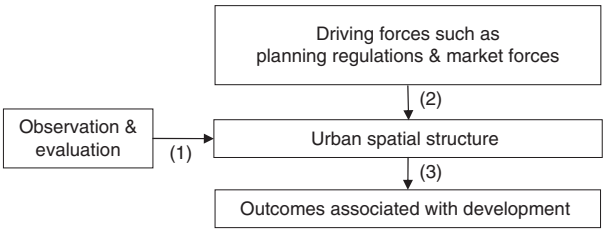

Figure 1 Research areas related to urban internal spatial structure.

Observation of Urban Internal Spatial Structure

The continuing efforts in creating efficient urban development patterns have brought increasing attention to the need for understanding urban spatial structures in order to describe patterns of development. Researchers and practitioners have developed numerous methods and focused on the development of detailed and policy-relevant measures to quantify urban spatial structures. The different types of measures of urban spatial structure are described in the following sections.

Monocentricity and Polycentricity of Cities

The most classical measures are monocentricity and polycentricity of cities or urban areas. As cities expand with prevalent usage of automobiles, there has been a higher level of urban decentralization, which is referred to the movement of population and businesses away from the established city center toward the urban fringe.

To measure decentralization, population and employment density gradients can be used. One example of a reliable estimation procedure of density gradient is the two-point estimation procedure. The estimates of density gradients can be made using available data on two areas: central city and total urbanized area. More specifically, the mathematical notation of the assumption that population density in an urban area declines as a negative exponential function of distance from the city center is $D(x) = be^{-ax}$, where $D(x)$ represents population density at distance x from the city center, b is the intercept (or the central density), and a is the density gradient of the function. A lower value in the density gradient indicates that an urban population or employment is more dispersed or spread out over the urban land area. For example, a value of 1.5 for the density gradient indicates a highly concentrated urban population or employment, whereas a value of 0.01 for the density gradient indicates a dispersed urban population or employment with equal density over the urban area.

Urban decentralization depicts two waves of urban expansion in the United States: decentralization of both population and employment. In the first wave of decentralization, which started in the late nineteenth century, the population moved to the suburbs. As late as in 1960, however, most jobs (63%) were still in the central city while 51% of the population lived in suburbs. In the subsequent second wave of decentralization,

jobs were also decentralized. By the year 2000, people both lived and worked in the suburbs in the United States. Across regions, the share of employment within three miles from the city center is rarely more than 19%. Consequently, employment subcenters emerged, establishing polycentricity of urban areas. Many studies have defined employment subcenters using a variety of methods, such as density maps, employment and population size, and minimum-density thresholds (Giuliano and Small, 1991). For examples, Giuliano and Small (1991) defined 32 employment subcenters in the five-county Los Angeles region by identifying a contiguous set of zones, each with 10 employees per acre, and with a combined minimum of 10 000 employees. Cho et al. (2008) identified subcenters for Mecklenburg County, part of the Charlotte metropolitan area in North Carolina. Because their study area is a much smaller metropolitan area than Los Angles (LA) in terms of population (its population is 6.58% of LA's), size (1.55% of LA's), and employment (10.35% of LA's), they lowered the minimum thresholds to four employees per acre for the employment density – that is, the mean employment density in the study area. As a result, a total of 10 subcenters, including the Central Business District (CBD), were identified in Mecklenburg County.

Similar decentralization patterns have been found in developing countries in recent decades. For example, Song (2013) investigates how road infrastructure in Chinese cities has shaped the degree of decentralization of Chinese urban regions in the last 20 years. There has been strong evidence that population density has been decreasing in the central city areas in Chinese cities. The improved mobility because of investments in road infrastructure, along with other factors such as increased urban income, has led Chinese cities to decentralize and evolve from monocentric to polycentric cities. It is evident that identification of population and employment subcenters can help understand the urban spatial structure of cities.

Metrics of Urban Internal Spatial Structure

There is a great amount of interest in extending the discussion of urban spatial structure by including measures of urban form. Understanding how land uses are organized within cities, for example, is essential for scholars and policymakers to tackle urban issues such as sprawl.

Galster et al. (2001) conducted a very detailed analysis of urban form with eight dimensions: density, continuity, concentration, clustering, centrality, nuclearity, mixed uses, and proximity. Each dimension reflects spatial relationships among units of analysis. These units are defined by assigning housing units from 1990 census blocks into 1-mile or 1/2-mile grids. Then, using Geographic Information System (GIS) tools, they computed each measure for 13 metropolitan areas. Finally, they normalized all the measures and, based on equal weights for each normalized measure, offered new and interesting information about urban development patterns in the US metropolitan areas. Consistent with conventional wisdom, Atlanta, Detroit, Miami, and Denver ranked high on the sprawl index while New York, Philadelphia, Boston, and Chicago ranked low. These measures developed by Galster et al. (2001) provide considerable information about forms of urban spatial structure beyond density and suburban population growth. Specifically, they reveal differences in spatial relationships among 1-mile grids within metropolitan areas and thus reveal quantitative insights about urban form. Yet their measures offer little of use to urban residents or policymakers. Should public officials in Houston, for example, be concerned or pleased that it ranked high in clustering and low in nuclearity?

Detailed and policy-relevant measures of urban form have since then been developed by many researchers. For example, Allen's measures, part of a planning support system called INDEX, include over 30 measures computed at a variety of geographic scales (Allen, 2001), offering a number of advantages over previous measures. Like all good indicators, they are well defined, relatively easy to compute when GIS data are available, easily interpreted, and policy relevant. Whereas gross measures of density, nuclearity, and centrality provide interesting information about metropolitan form, measures of transportation options, residential proximity to retail and industrial uses, and accessibility of parks, shops, and transit are of direct concern to citizens and policymakers. Furthermore, these measures serve well to evaluate impacts of market and planning on urban development patterns. Song and Knaap (2004) developed a variety of measures to illustrate urban development patterns in a range of urban areas in the United States. Their results suggest that some characteristics of development patterns differ significantly within and across study areas and over time; this raises doubt about the utility of one single index of urban spatial structure for an entire metropolitan area. Their results also suggest that by investigating characteristics of urban spatial structure, the good news is that density is getting higher, cities are becoming more compact, and neighborhoods are becoming more internally accessible; the bad news, however, is more extensive: residential neighborhoods are becoming more isolated, land uses remain separated, and pedestrian accessibility to various destinations is falling.

Despite the above-mentioned efforts in developing urban form metrics to quantify urban development patterns, there are several issues with current practice in computing measures. First, the set of measures included in one study sometimes do not sufficiently capture the complexity of urban form. Description of urban form relying on one-dimensional measures, like development density or land use, has failed to adequately describe urban form in a comprehensive manner. Thus there is a need for a set of metrics covering different dimensions of urban form. Second, some commonly used indicators are possibly ill-defined. As Talen (2003) writes, "too much discussion about cities is devoid of measurement … Examples are words such as suburb, public realm, mixed use, diversity and access. These concepts are vital to the discussion, but have been difficult to pin down." It is necessary to have better defined and measurable metrics, which can be either quantitative or qualitative. Third, more recently, the proliferation of spatial data sources and GIS tools for spatial data analysis has made a variety of new spatial metrics available. However, many of these computed spatial metrics are correlated and redundant measures. For example, when quantifying street network connectivity for cities, should we calculate the *Alpha* or the *Beta* Connectivity Index, or both?

In summary, there is a need for a set of better defined urban form metrics to depict urban spatial structure. Recent research on urban form calls for multidisciplinary approaches to measure urban spatial structure (Clifton et al., 2008). Four conclusions are drawn in this area of research. First, over the last two decades, substantial progress has been made in the ability to measure and analyze spatial patterns that help characterize urban form. Second, at multiple scales and for a variety of reasons, there are advantages to development that is mixed and compact. Third, normative principles and policies for addressing urban form need to be crafted at multiple scales and carefully designed to address the disparate issues that arise at each scale. Fourth, with so many disparate measures now used to operationalize the same constructs, it would advance urban form research to have some standardization in operational definitions and measurement protocols.

Built Environment Types within Urban Areas

More recently, researchers have developed quantitative methods of classifying and locating different types of urban developments within urban areas. In the past few years, urban designers have developed neighborhood typologies for the improvement of urban policy; the *transect* is a recent example. The central notion of the *transect* is a geographical cross-section of a region including a gradient of area types, ranging from rural to urban (Talen, 2002). Components of the built environment such as buildings, lots, land uses, and streets can then be organized into each area type. As shown in Figure 2, the rural-to-urban continuum can be divided into six discrete area types: natural, rural, suburban, general neighborhood, neighborhood center, and urban core. As one moves along the gradient of these area types, differences in design, ecology, and social structure are apparent. Policies used to manage urban development elements such as housing types, street design, and setback can be specified, using the *transect*, by each area type.

To apply the above described concept and to examine how different built environment types are allocated in the urban area in the rural-to-urban continuum in the US urban areas, Song and Knaap (2007) developed a method with the following steps: compute a range of measures of urban form, distill these into a small set of design factors using factor analysis, and develop statistically defined neighborhood classifications using cluster analysis. Variables associated with the physical form of the built neighborhoods have been established in the literature, as mentioned earlier. Street design, mixed land uses, accessibility, density, and alternative transportation modes are common variables used to identify neighborhood types (Song and Knaap, 2004). For examples, *street design measures* include number of intersections, number of cul-de-sacs, length of street miles, size of the street blocks; *lot design and density measures* include lot setback, size of the parcel, floor space of the single-family house, and number of single-family lots; *mixed land use measures* include acres of commercial, industrial, public, and multifamily land use; *accessibility measures* include distance to commercial stores or to bus stops; *natural environment measures* include acres of open space and amount of areas with tree canopy. Some of the above measures are highly correlated. The distribution of cul-de-sacs, for example, is highly correlated with the distribution of large blocks. Therefore, in the second step, it is useful to condense these variables into a smaller set of variables that removes redundancy and correlation in the data. Factor analysis, a technique for data reduction, can be used to understand the dimensional structure of the group of variables and improve the overall efficiency of the analysis. In the third step, cluster analysis, a method of combining observations into groups based on their similarity within a set of predetermined characteristics, can be performed to facilitate the identification of neighborhood types. K-means cluster analysis can be used to classify objects into different neighborhood types on the basis of similarities and dissimilarities in the values of the factors scores in such a way that each neighborhood type is internally as similar as possible but externally dissimilar to other

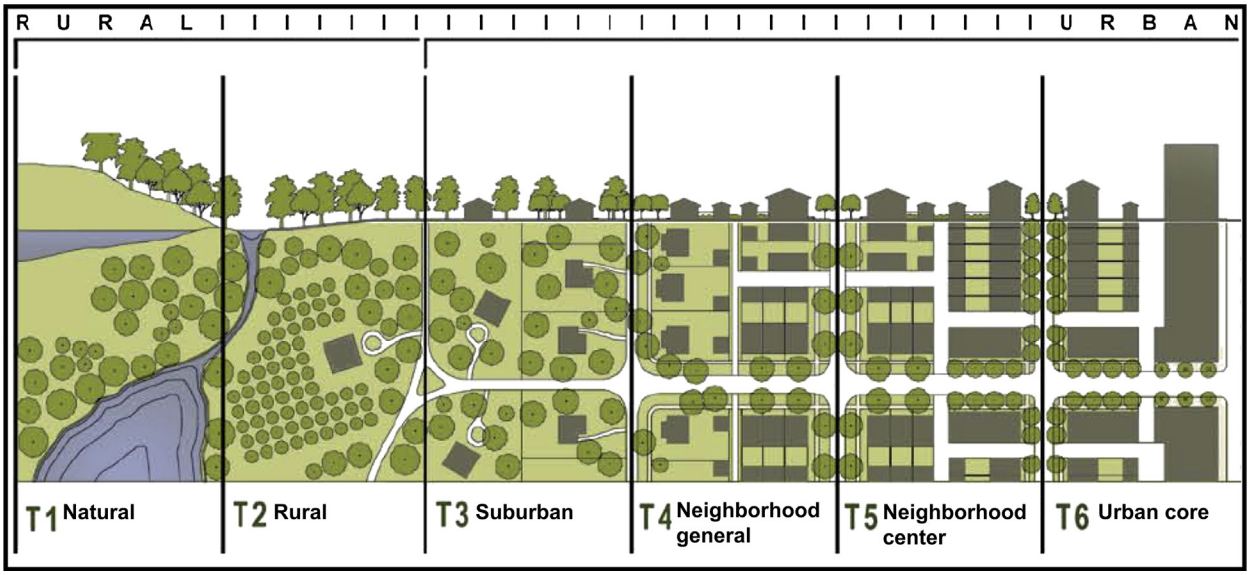

Figure 2 Urban transect. Source: http://www.transect.org/transect.html.

neighborhood types. The method has been applied to a range of metropolitan areas across the world to examine how urban developments within an urban area differ by their characteristics and locations. In what follows, there are two examples on identification of built environment types within urban areas in the United States and China, respectively.

In the rural-to-urban continuum in a typical US metropolitan area, there are in general, the following types of developments ranging from the outskirt to the central area.

Neighborhood Type 1: Sporadic Rural Developments

This type of development has a set of typically rural features: very sparse and unconnected transportation networks with poor access to transit, large lots and houses with low density, dominant rural land use with abundant open spaces but little mixed use such as public, industrial, or multifamily residential uses.

Neighborhood Type 2: Bundled Rural Developments

This type of development is poorly connected, dense with small lots, proximate to public land uses such as schools but not close to other land uses such as commercial and multifamily residential use.

Neighborhood Type 3: Outer Ring Suburbs

The outer ring developments are characterized by disconnected street networks, numerous cul-de-sacs, relatively large lots, predominantly detached single-family land uses with little nonresidential land use, and moderate amount of open space. These neighborhoods, developed in the decades after World War II, as developers began to add curvilinear street forms and new street block and lot layouts to the landscape, are the very image of mid- to late twentieth century automobile suburbs. In some cities, the outer ring suburbs, accompanied with concentrated employment centers, are identified as 'edge cities.'

Neighborhood Type 4: Composite Greenfields

Type 4 neighborhoods are located in recent greenfields and are characterized by modified grid and connective street networks, high density, moderate accessibility to bus stops, ample multifamily residential use but little commercial land use in the immediate neighborhoods.

Neighborhood Type 5: Downtown, Inner and Middle Ring Suburban Redevelopments/Infill

Type 5 developments, which are located in downtown areas or at the inner and middle ring of cities, are characterized by connective street networks and small street blocks, small lots and high density, accessible bus services, mixed land use such as commercial and multifamily residential use, but little open space. These neighborhoods correspond to the mid-nineteenth and the early twentieth century developments. The homes in downtown areas were built in the late eighteenth and early to mid-nineteenth centuries and had characteristics such as grid street networks and small, square, and 200-foot blocks. These small blocks were developed by the prewar land developers to maximize the rent received from higher-priced corner lots – the smaller the blocks, the more corners to rent per land claim. In response to the advent of streetcars, neighborhoods developed in the 1870s and 1880s had larger rectangular blocks and shops along streetcar corridors. From the 1930s, when the widespread use of automobiles was added to the settlement pattern of earlier years, the inner and middle suburban neighborhoods began to show a modified rectilinear grid with blocks of varying sizes.

Similar analysis has been done to identify built environment types in developing Chinese cities (Zhang and Song, 2013), demonstrating a variety of five built environment types based on variations in the built environment measures. These five neighborhood types also depict a continuum from urban and suburban areas to rural areas. Three main themes are identified for developing Chinese cities: the renewed urban area, the decentralized suburbs, and the urbanizing rural area. More specifically, type 1 neighborhoods are the urbanizing rural neighborhoods, featuring low density, poor transport connections and commercial or occupational linkages, and reflecting a new transformational force in many rural neighborhoods, where urbanization is motivated by the conversion of farmland into medium- to small-sized industrial factories. This transformation can increase landowners' income and absorb surpluses in the local labor force. Types 2 and 3 neighborhoods, the new urban fringe and the expanded urban neighborhoods, respectively, are suburbanized area types that were once greenfields. Mostly thanks to market forces (such as increased income and demand for new and bigger housing units) and planning efforts, the new urban fringe neighborhoods have now developed into lively urban places with moderate population density, high accessibility, and high land use diversity. Driven by employment and population decentralization, these suburbanized districts are now acquiring an urban feel as streets are paved, public spaces are beautified, and housing increasingly takes the form of high-density, multistory apartment buildings. Types 4 and 5 neighborhoods, the renewed urban core and the urban district neighborhoods, respectively, are clearly gentrifying. This process is likely attributed to the urban renewal movement motivated both by the marketplace and by the government policies, which are enforced to redevelop the deteriorated old, public, and welfare housing into high-end or moderate commercialized housing and core business uses; and to upgrade street networks and transit service.

Formation of different built environment types within urban areas is driven by the joint forces of government policies and the marketplace. In the case of Chinese cities, the city core is quite vibrant with housing and commercial rehabilitation in the core area, due to cultural preferences, market forces, and government policies. The role of the state in China has been more direct and powerful in setting the suburbanization process and urbanization process of rural areas in motion. Among the forces driving suburbanization were marketization of urban land, the shift of industrial land to tertiary use, state and local funding of transportation improvement, the availability of foreign and domestic capital, and new housing construction in the suburbs. The government policies have also been active in promoting the construction of the 'New Countryside' in rural areas, aiming to decrease the discrepancies between urban and rural living conditions.

From these two examples, it is evident that identifying different built environment types within urban areas facilitates our understanding of the internal structure of urban areas, which further assists with drafting relevant and varied urban policies for different types of neighborhoods within urban areas.

Types and Quantity of Land Conversions

Identification of urban internal spatial structure also involves the detection of land use conversions; that is, land use changes resulting from converting a vacant plot or changing the current use of a given land parcel. Some typical land use classifications include (Wu and Yeh, 1999) agricultural, industrial, commercial (including large department stores, offices, hotels, markets, and recreation facilities), residential development, government, institute, and community facilities.

To identify land conversions, one method is to conduct a supervised classification of Landsat satellite imagery and detecting change in urban areas. Landsat is a commonly used data source for analyzing landscape change. It has global coverage and captures data in seven bands, at a resolution of 30 m, in scenes that are approximately 180 km^2. Landsat five and Landsat seven detect blue, green, and red light in the visible spectrum as well as near-infrared, mid-infrared, and thermal-infrared radiation that human eyes cannot perceive. Landsat records this information digitally and it is downlinked to ground stations, processed, and stored in a data archive. By classifying the patterns of reflectance in each grid cell, researchers can interpret the land cover located there and once satellite images have been classified into land cover categories, researchers can study how those categories change over time. Typical steps involved for detecting land use changes include the following: (1) creating and preparing a training dataset to support the satellite image classification; (2) deriving a signatures file containing spectral characteristics of land cover classes of interest; (3) performing a supervised classification of Landsat satellite imagery; and (4) detecting and quantifing the observed changes in urban areas over time.

Numerous studies have been carried out to detect land use changes within urban areas. By detecting the amount, type, and location of land use conversion patterns, researchers are able to evaluate how the structure of urban areas changed over time. For example, by examining the changing urban landscape patterns between 2001 and 2006 for the United States using the land cover data from the US Geological Survey, Kaza (2013) characterizes the landscape metrics for each county in the continental United States. He found uneven and heterogeneous patterns of growth across counties. More importantly, his study shows that although urban counties are becoming less fragmented, some rural counties in the Southern and Western United States are experiencing significant leapfrog development. Land conversion studies can also be carried out for single urban area or metropolitan regions. For example, Song (2013) tracks land use changes from nonurban to urban use by each cell of 30 by 30 m for the City of Shenzhen, China. The land use change analysis yielded a total of 8.2% of the cells that changed between 1994 and 2005. The spatial distribution of changes is shown in Figure 3. Most changes occurred along the coastline and highway or arterial corridors.

A second method is to analyze land parcel changes by comparing parcel shape files in successive periods to determine whether they had split or changed functions (Wilson and Song, 2011). GIS shape files over time are necessary for such an analysis. ArcGIS software and the Python scripting language can be used to automate and streamline spatial data processing and manipulation tasks.

In general, recent land conversion studies have demonstrated that urban sprawl has been occurring, thus calling for public policies designed to reverse the trend and correct for negative externalities brought about by urban sprawl.

Identification of Driving Forces of Urban Spatial Structure

Many studies have focused on identify driving forces (including both market and regulatory factors) shaping urban development patterns, such as urban scale, and rate and location of land use conversions. For example, researchers have used standard monocentric city models (both closed city and open city models), and extended monocentric city models to investigate the determinants of urban spatial scale of urbanized areas across the world (Ke et al., 2009; Song and Zenou, 2006). Theoretical models define that four basic factors – population, income, commuting cost, and price of rural land – explain variations in urban spatial scales. Empirical analyses of cities across the world confirm the theoretical derivations with the following results. For the US cities, the population and income variables have positive and significant effects on the spatial size of urbanized areas; the expenditure on transportation, a proxy for commuting costs, has positive effect on urban spatial scale and thus confirms the predicted negative relationship between urban size and commuting costs; and the higher agricultural land rents surrounding urban areas can result in smaller cities in the United States (Song and Zenou, 2006). Other studies (Ke et al., 2009) have shown strong evidence that urban economics theory has common value in explaining urban spatial scale in different countries, provided that market forces are not suppressed.

Another set of studies confirmed that the decision to convert undeveloped land or to change (usually to intensify) the existing use is influenced by a variety of economic, social, political, and personal considerations. Sheer population growth increases the absolute demand for housing and other developed land uses, and demographic changes (e.g., an influx of young professionals) can impact the types of residential development that are favored. The type and magnitude of economic activity can, therefore, exert a great deal of influence on the number and characteristics of the local population. The political climate also impacts the nature of regulation and extent to which growth management policies are implemented to affect land conversion processes. In addition to leadership at the regional and state levels, a robust economy and an educated population are key correlates of planning and development management influencing land conversions (Wilson and Song, 2011).

A variety of techniques, including hazard models and discrete choice models, have been employed to study the conversion of land uses. For example, Bockstael (1996) brought together concepts from several disciplines to model the probability of land parcel conversion within the Patuxent watershed in Maryland. The key contribution of this study was its representation of the landowner's decision to convert a given parcel from one state (use) to another as a comparison of the present value of expected returns under the current state less the costs of converting the parcel now, and the expected returns under alternate states minus conversion costs. Zax and Skidmore (1994) used a hazard model to investigate the effect of tax rate changes on the probability of parcel conversion and found

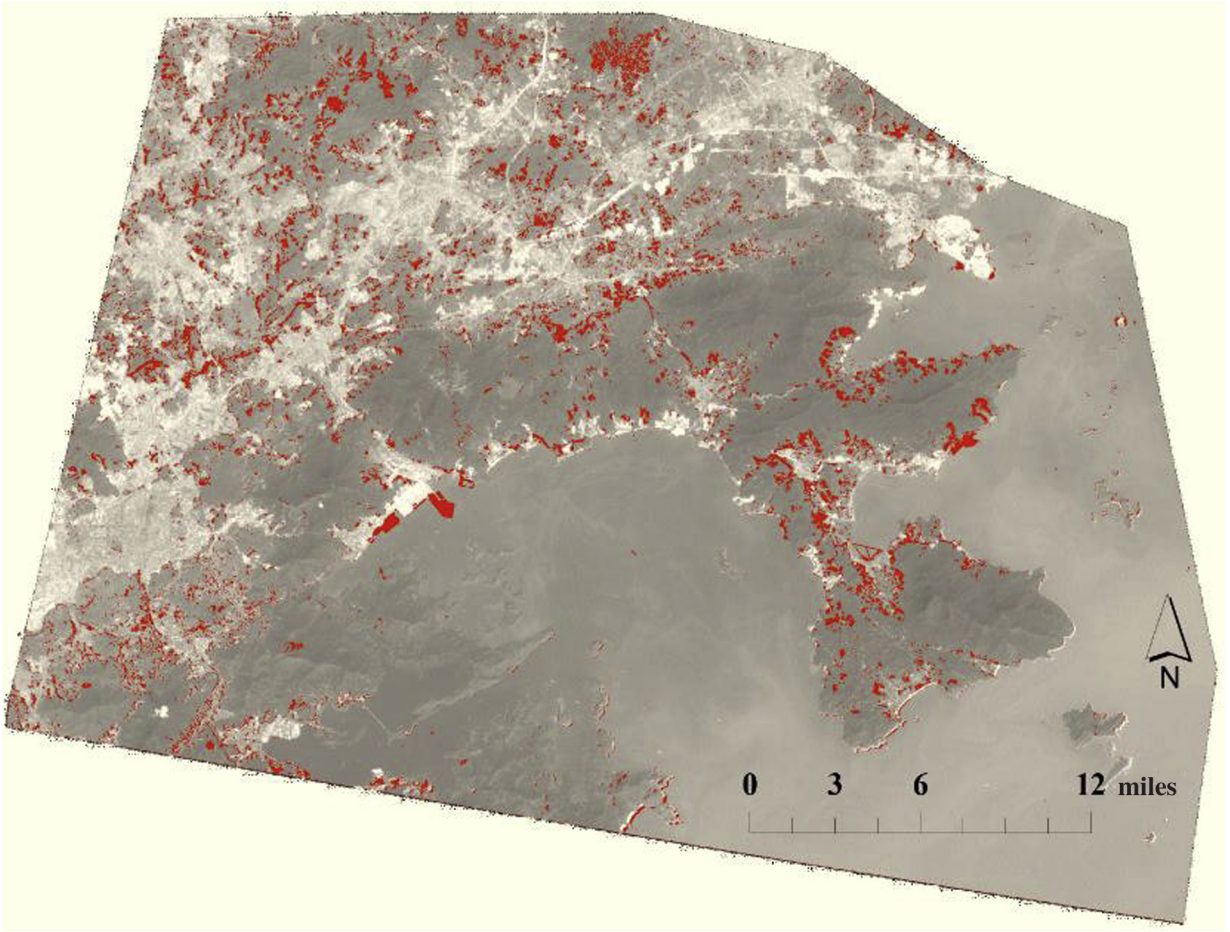

Figure 3 Land use changes between 1994 and 2005 in Shenzhen. Source: Author.

evidence that when increases in the tax rate are known in advance, the effect is an increase in the short-term probability of development. Wilson and Song (2011) analyzed parcel changes and then conducted a spatially explicit hazard analysis of land use change in Mecklenburg County in North Carolina in the United States, to identify a set of factors that predict which vacant land parcels are most likely to be subdivided during a given one-year time period. They found evidence of a positive association between large residential subdivision events and an increased likelihood of subsequent subdivision activity. This corroborates that major residential subdivisions drive land conversion.

As the examples illustrate, by understanding the factors that explain land conversions, planners, elected officials, and researchers can be in a stronger position to manage growth, craft public policy, and refine our understanding of how urban systems function.

Investigation of Urban Internal Spatial Structure and Associated Outcomes

Investigation of the connection between urban spatial structure and outcomes encompasses a range of topics. For example, there are a variety of studies on how alternative urban forms (monocentric versus polycentric, or compact versus sprawling) are valued in a land or housing market, whether there is any relationship between alternative urban forms and individual well-being (such as commuting time and level of satisfaction), and whether alternative urban development patterns in the long term can affect water and air quality.

To explore how alternative urban forms (compact versus sprawling) are valued in a housing market, quantitative price valuation of alternative urban forms can be performed. Spatial hedonic housing models are used to evaluate how certain features of urban development patterns, such as density, connectivity, land use mixture, and accessibility are valued in a housing market (Song and Knaap, 2003). The results have shown in general some price premium associated with more compact developments with a range of amenities, such as access to public transit, neighborhood stores, and parks.

Many studies explore the relationship between urban spatial structure and individual well-being. For example, Cho et al. (2008) explored how polycentric urban spatial structure affects residential location choice and they find that access to certain employment subcenters, measured in terms of generalized cost, is an important determinant of household residential location decisions and thus brings a higher level of utility to the households. Other research questions include, for example, would monocentric or polycentric spatial structures result in different commuting behaviors? How would different spatial structures result in different

outcomes in resident satisfaction, physical activity, or stress levels? More research is needed in these areas.

Scholars have also explored whether alternative urban forms (compact versus sprawling) can affect air quality in the long run. Integrated land use–transportation–emissions models have been developed to test, for example, whether polycentric and transit-oriented developments can reduce the vehicular emissions contributing to tropospheric ozone and to fine particulate matter. One example of this area of research developed a modeling system assessing long-term development scenarios for Charlotte and Mecklenburg County, North Carolina's largest metropolitan area and part of an ozone nonattainment area. The results demonstrate that more efficient urban spatial structure can result in reduced emissions and subsequently improved air quality (Morton et al., 2007). Nevertheless, research in this area is still scarce, more studies using different methods, such as simulation modeling or regression analysis using empirical data, are needed to explore what form of urban spatial structure can make cities more green and healthy.

In conclusion, the term 'urban internal spatial structure' encompasses many related dimensions, including a range of concepts concerned with the organization, configuration, and form of urban land use and activities. The way that land use and activities are arranged affects many aspects of how cities function and has implications for economic efficiency, accessibility and connectivity, social equity, and environmental sustainability. When most cities around the world are urbanizing or suburbanizing, a shift to a sustainable urban spatial structure for urban areas provides a more desirable growth path.

See also: Eco-cities and Sustainable Urbanism; Suburbanization and Suburbanism; Urban Sprawl.

Bibliography

Allen, E., 2001. INDEX: software for community indicators. In: Brail, R.K., Klosterman, R.E. (Eds.), Planning Supporting Systems: Integrating Geographic Information Systems, Models, and Visualization Tools. ESRI Press, Redlands, CA, pp. 229–261.

Bockstael, N.E., 1996. Modeling economics and ecology: the importance of a spatial perspective. American Journal of Agricultural Economics 78 (5), 1168–1180.

Cho, E.J., Rodriguez, D., Song, Y., 2008. The role of employment subcenters in residential location decisions. Journal of Transport and Land Use 1 (2), 121–151.

Clifton, K., Knaap, G.J., Ewing, R., Song, Y., 2008. Quantitative analysis of urban form: a multidisciplinary review. Journal of Urbanism: International Research on Placemaking and Urban Sustainability 1 (1), 1–31.

Galster, G., Hanson, R., Ratcliffe, M.R., Wolman, H., Coleman, S., Freihage, J., 2001. Wrestling sprawl to the ground: defining and measuring an elusive concept. Housing Policy Debate 12 (4), 681–717.

Giuliano, G., Small, K.A., 1991. Subcenters in the Los Angeles region. Regional Science and Urban Economics 21, 163–182.

Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York.

Kaza, N., 2013. The changing urban landscape of the continental United States. Landscape and Urban Planning 110 (2), 74–86.

Ke, S., Song, Y., He, M., 2009. Determinants of urban spatial scale—Chinese cities in transition. Urban Studies 46 (3), 1–19.

Morton, B.J., Rodriguez, D.A., Song, Y., Cho, E.J., 2007. Using TRANUS to construct an integrated land use-transportation-emissions model of Charlotte, North Carolina. In: Transportation Land Use, Planning and Air Quality Conference Proceeding.

Song, Y., 2013. Chapter 1, "Infrastructure and urban development: evidence from Chinese cities". In: Ingram, G., Brant, K. (Eds.), Infrastructure and Land Policies. Lincoln Institute of Land Policy.

Song, Y., Knaap, G.J., 2003. New urbanism and housing values: a disaggregated assessment. Journal of Urban Economics 54, 218–238.

Song, Y., Knaap, G.J., 2004. Measuring urban form: is Portland winning the war on sprawl? Journal of American Planning Association 70 (2), 210–225. (This article was evaluated as "noteworthy and classic" article by JAPA).

Song, Y., Knaap, G.J., 2007. Quantitative classification of neighborhoods: the neighborhoods of new single-family homes in the Portland Metropolitan Area. Journal of Urban Design 12 (1), 1–24.

Song, Y., Zenou, Y., 2006. Property taxation and urban sprawl. Journal of Urban Economics 60, 519–534.

Talen, E., 2002. Help for urban planning: the transect strategy. Journal of Urban Design 7 (3), 293–312.

Talen, E., 2003. Measuring urbanism: issues in smart growth research. Journal of Urban Design 8 (3), 195–215.

Wilson, B., Song, Y., 2011. Do large residential subdivisions induce further development? A spatially explicit hazard analysis of land use change in Charlotte. Journal of American Planning Association, 5–22. Winter 2011.

Wu, F., Yeh, A.G.-O., 1999. Urban spatial structure in a transitional economy. Journal of the American Planning Association 65 (4), 377–394.

Zax, J.S., Skidmore, M., 1994. Property tax rate changes and rates of development. Journal of Urban Economics 36 (3), 314–332.

Zhang, Y., Song, Y., 2013. A typology of neighborhoods in Chinese cities: the case of Zhongshan Metropolitan Area. Working paper.

Urban Life and Health

Roderick J Lawrence, Institute of Environmental Sciences, Faculty of Economic and Social Sciences, University of Geneva, Geneva, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A.T. Geronimus, volume 24, pp. 16026–16029, © 2001, Elsevier Ltd.

Abstract

Urban environments are complex with many constituents. Likewise, health is multidimensional. Hence, both these subjects are not structured within traditional disciplines and professional sectors. Therefore, urban life and health ought to be considered in terms of the multiple factors that influence living conditions and health status, as well as the interrelations between them. An ecological perspective recognizes that behavioral, biological, cultural, economic, social, physical, and political factors need to be considered if a comprehensive understanding of urban health is to be developed and applied. In order to integrate all these dimensions, it is necessary to go beyond interpretations that rely solely on the biomedical model of health. Interdisciplinary and collaborative research contributions are necessary.

Introduction

One hundred years ago, about 80% of the world's population lived in rural areas, whereas in the year 2008 a half of the global population lived in cities. The twentieth century was characterized by a growth in the number, population size, and total surface area of cities on a scale previously unknown, and this trend is expected to continue during the twenty-first century. Some have argued that the last century corresponded to a period of 'the urban revolution' that has transformed the physical, psychological, and social dimensions of daily life including housing, transport, and other characteristics of metropolitan areas. Urban life has important health benefits including easy access to job markets, education, health care, medical services, as well as diverse cultural and leisure activities (Lawrence, 2000). However, at the beginning of the twenty-first century, urban health can be characterized by relatively high levels of tuberculosis, respiratory and cardiovascular diseases, cancers, adult obesity, and malnutrition, tobacco smoking, mental ill health, alcohol consumption and drug abuse, sexually transmitted diseases (including AIDS), as well as fear of crime, homicides, violence, and accidental injury and deaths. It is noteworthy that in the 1990s, mental ill health was integrated into the etiology of urban health, and that the promotion of physical and mental health was accepted as a complementary goal for national and local policymakers and professionals.

What Is Urban Health?

The World Health Organization (WHO) states that health is "not merely the absence of disease and infirmity but a state of optimal physical, mental, and social well-being" (World Health Organization, 1946). This definition is idealistic, but it has the merit of not focusing on illness and disease, which have often been considered as either temporary or permanent impairment to health, or the malfunctioning of a single or several constituents of the human body. Given that the WHO's definition of health includes social well-being, the most common interpretations of health ought to be enlarged. The WHO also states that the enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political, economic, or social condition.

Health can be defined as a condition or state of human beings resulting from the interrelations between humans and their biological, chemical, physical, and social environment. All these components of residential environments should be compatible with their basic needs and their full functional activity including biological reproduction over a long period (Lawrence, 2001). Health is the result of both the direct pathological effects of chemicals, some biological agents, and radiation, and the influence of physical, psychological, and social dimensions of daily life including housing, transport, and other characteristics of metropolitan areas. For example, improved access to medical services is a common characteristic of urban neighborhoods that is rare in rural areas. In the field of health promotion, health is not considered as an abstract condition, but as the ability of an individual to achieve her or his potential and to respond positively to the challenges of daily life. From this perspective, health is an asset or a resource for everyday life, rather than a standard or goal that ought to be achieved. This interpretation is pertinent because the environmental and social conditions in specific residential environments do impact on human relations, induce stress, and can have positive or negative impacts on the health status of groups and individuals (Hartig and Lawrence, 2003). It also implies that the capacity of the health sector to deal with the health and well-being of populations is limited and that collaboration with other sectors would be beneficial.

A discussion of urban health should begin by clarifying what is meant by urban and what interpretation of health is applied. Although there is no consensus about the definition of 'urban,' which varies from country to country, the United Nations uses national definitions that are commonly based on population size. Many countries use 10 000 resident persons as a benchmark to define urban, whereas other countries including India use 5000 persons as the referent. Other definitions of urban are based on the administrative or political

authority of municipalities, especially the degree of autonomy in relation to the national or state administration. Other definitions include the socioeconomic status of the resident population, especially their livelihood (e.g., the share of all employed persons having nonagricultural occupations). A combination of these characteristics could be used to interpret rural and urban areas, but this has been rare, especially in recent published research on the health of urban populations.

Urban health is a vast and complex subject. In order to understand the multidimensional nature of urban health, all the constituents of urban ecosystems and the interrelations between them should be considered over time. These constituents include four interrelated sets of hazards:

Environmental hazards including ambient air quality, ambient noise levels, soil and water contamination, and solid waste disposal.

Social hazards including criminality, violence, and community discord, as well as the lack of education and training especially for immigrants, women, and children.

Economic hazards such as no affordable housing, permanent unemployment, or inequalities of access to diverse kinds of resources including primary health care.

Technological hazards including traffic accidents, industrial and chemical disasters on industrial sites, and contamination from mass-produced foods and synthetic products.

These four main sets of risk factors are variable within and between urban areas and neighborhoods and even the specific characteristics of buildings impact on health. In addition, exposure to hazards can vary over short and relatively long periods of time. For example, the internal conditions of a room in a housing unit may be related to the physical condition of the residential building and its site, and the urban neighborhood in which it is located. In turn, the neighborhood, the city, the region, and the country are all directly or indirectly interrelated. Hence, global challenges, including climate change, can have different impacts on cities and urban neighborhoods within cities, depending on their geographical location, local climate, infrastructure, and building construction (McMichael, 2000). In essence, local and international levels are linked across cultural, geographical, political, and temporal boundaries.

The exposure of different population groups (e.g., children, ethnic groups, the elderly, and the unemployed) to these four sets of risk factors should also be taken into account (Galea and Vlahov, 2005). It cannot be assumed that an individual, household, or a particular social group exposed to one or more of the above sets of hazards will necessarily suffer negative impacts to their health. In principle, biologically inherent mechanisms are mediated by the social and environmental circumstances of residential neighborhoods (Hartig and Lawrence, 2003). Hence, someone exposed to a hazard can have a different risk or negative impact compared with another person also exposed to the same hazard. Therefore, it is necessary to interpret the health of urban populations in terms of both individual and social differences by explicitly accounting for age, gender, socioeconomic class, occupational status, and the geographical and temporal distribution of the population in terms of the four sets of risk factors presented above.

An ecological interpretation of urban life and health considers four main sets of interrelated factors: the individual, who has a specific genetic code with a susceptibility and immunity to illness and disease, as well as lifestyle traits; the agent or vector of illness and disease, including not only biogeophysical components of the environment but also the social and psychological dimensions of human settings; the physical and social environment of the individual that affects the susceptibility of the host, the virulence of biophysical agents, and the exposure, quantity, and nature of the contact between host and vector; and the available resources used by the individuals and households including housing, nutrition, money, information, and access to health and medical services, which ought to be affordable for all groups of the population. This perspective implies that an analysis of the interrelations between multiple components of any human ecosystem is necessary. Systemic interpretations of human illness, health, and local environments have a long history. They can be traced back at least as far as the Hippocratic treatise 'On Airs, Waters, and Places' published initially about 2600 years ago (Hippocrates, 1849).

The distinction between biomedical models and ecological interpretations of urban health is fundamental. The germ theory, for example, is an incomplete explanation of human illness and disease because it ignores the contribution of numerous physical and social dimensions of the environment that can impact on health. Ecological interpretations maintain that the presence of a germ is a necessary but not a sufficient condition for an individual to become ill. They accept that some individuals become more susceptible to certain illnesses because of their differential exposure to numerous environmental, economic, and social factors that can promote or be harmful to health and well-being. This interpretation does not ignore the influence of genetics, individual behavior, or primary health care. However, it maintains that, alone, these do not address possible relations between social problems and illness (e.g., inequalities) or positive social dimensions and health promotion (e.g., public education). The distinction between potential and actual health status can be the foundation for a new interpretation of health that includes the way environmental, economic, social, and technological risk factors transgress traditional disciplinary boundaries while remaining locality and temporally specific. Cues for this interpretation were already provided by human ecologists in the 1920s.

The early studies in human ecology at the University of Chicago plotted the geographical distribution of some characteristics of the resident population of Chicago including their ethnic origin, socioeconomic status, birth and mortality rates, delinquency, mental, and other illnesses. These cartographic studies enabled the authors to overlay the maps of the spatial distribution of these characteristics in order to identify those that occurred in the same urban area. This approach established many correlations (Park et al., 1925). For example, the distribution of cases of tuberculosis corresponded with the highest incidence of delinquency. This finding led the authors to suggest that cities comprise 'natural' areas that are defined geographical, economic, social, and cultural dimensions. The term 'natural' was used because these areas were not planned but constituted as the city developed. Each 'natural' area is characterized not only by its geographical location but also by

the market price of land and property, and sets of customs, norms, and nonmonetary values that are part of the lifestyle of the residents.

The first generation of studies in human ecology from the 1920s marked a growing interest in collaboration among researchers from disciplines who worked to deal with complex multidimensional questions including those of urban health. This kind of interdisciplinary collaboration has been extremely fruitful since the 1970s.

Research on the impacts of the activities of human societies on biophysical components of the biosphere and on humans themselves has been completed (Boyden, 1992). This research examines the adaptation processes used in response to changes in human ecosystems. The term ecodeviation is used to refer to a general biological principle that describes life conditions of human and other species that are different from those in the natural habitat of the species. When these differences become large, sudden perhaps irreversible behavioral and physiological maladjustments may occur (Laughlin and Brady, 1978). Some physiological maladjustments during the last 10 000 years include diseases such as scurvy, typhoid, cholera, smallpox, and influenza.

Another approach used by Boyden and his colleagues analyzes the advent and acceptance of new technologies including those that are not essential for basic human needs and sustenance. These inventions become an integral part of human societies by a process termed as technoaddiction. The spatial and social organization of societies can become dependent on them as shown by the reliance on combustion engines using fossil fuels. For example, this has led to transportation policies in many countries based heavily on road traffic using private cars in preference to public transport by rail. Some consequences of this kind of technoaddiction include the extensive use of arable land for urban and regional development, the incidence of air pollution and noise, as well as nonactive lifestyles.

An in-depth case study of Hong Kong illustrates the principles of ecodeviation and technoaddiction. This case study was completed as part of the UNESCO Man and Biosphere Program dealing with ecological studies of human settlements (Boyden et al., 1981). The case study of Hong Kong by Boyden and his colleagues identifies and measures the changing ecological characteristics of the human settlement and its hinterlands prior to the formal founding of the British Colony in 1842 until the 1970s. This study notes that with the advent of colonialism traditional rules and customs were challenged by foreign ones leading to the introduction of novel forms of land tenure, new resources including materials, machinery, and techniques, and the proliferation of cash and market economies. The specialization of tasks in Hong Kong increased. The economic consequences of these kinds of changes are important, because the total cost of energy consumption, materials, and labor increased.

This case study examines patterns of energy production and consumption flows, the production and disposal of wastes including pollutants, the densification and dispersal of the built environment, and impacts on the health and well-being of the population. The case study also includes a comparison of traditional Chinese and contemporary imported methods of food production, processing, and packaging. The outcomes of the change from one set of process to the other include impacts on land uses, agricultural production methods, the import and export of diverse kinds of materials, and the diet and health of the population. The authors note that, at the beginning of the twentieth century, inadequate nutrition was common in Hong Kong owing to a diet almost exclusively limited to polished rice. By the late twentieth century, however, the local environment provided a range of food including fresh fruit, fish, meat, and vegetables. Although nutritional diseases, such as beriberi, have been eradicated, in recent decades there have been changes in the diet of the population especially the increasing consumption of refined carbohydrates (e.g., white flour and refined sugar), which imply reduced dietary fiber. These changes can lead to increased dental caries, overnutrition, and obesity, as well as an increase in diabetes. The authors note that the death rate from diabetes was 4.3 per 100 000 in 1974, which is a threefold increase since 1961 and a fourfold increase since 1949.

Boyden and his colleagues raise many questions about the costs and benefits of industrialized food production, the toxicity of fertilizers, the ecology of soil and water catchment areas, and deforestation. In essence, this kind of case study can identify both the intended and unintended consequences of human products and processes over the long term. It also shows that it is important to distinguish between the tacit know-how of the local population prior to colonization (in which knowledge and practice are indistinguishable) and the explicit know-how of the colonial administration (in which theory and practice are interrelated but clearly distinguishable). The gradual suppression of implicit instruments of regulation by a growing number of explicit ones has occurred with urbanization in Hong Kong and elsewhere. This example illustrates that there is a need to reconsider both implicit and explicit regulatory means and measures, which are fundamental constituents of people–environment relations.

What Is a 'Healthy City'?

According to Hancock and Duhl (1988), "a healthy city is one that is continually creating and improving those physical and social environments and expanding those community resources which enable people to support each other in performing all the functions of life and in developing themselves to their maximum potential." This definition of a healthy city implies that health is determined by both short- and long-term processes. A healthy city is not only a quantified outcome such as the measurement of health status of the population. The long-term goals of the WHO Healthy Cities project are to integrate health in the agenda of all policy decision makers in cities, to create a strong partnership for health promotion between groups in the public and private sectors, and to apply a local, participatory approach when implementing projects (Goldstein, 2000; World Health Organization, 2000).

The health status of populations in specific urban areas is the result of not only many material and nonphysical constituents listed in **Table 1** but also the interrelations between them. Therefore, it is inappropriate to isolate one factor from the contextual conditions in which it occurs. Instead, systemic ecological approaches ought to be applied to understand both

Table 1 Qualities of a healthy city

1. The meeting of basic needs (for food, water shelter, income, safety, and work) for all citizens
2. A clean, safe physical environment of high quality, including housing quality
3. An ecosystem that is stable now and sustainable in the long term
4. A diverse, vital, and innovative economy
5. A strong, mutually supportive, and nonexploitive community
6. A high degree of participation and control by the public over the decisions affecting their lives, health, and well-being
7. The encouragement of connectedness with the past, with the cultural and biological heritage of city dwellers, and with other groups and individuals
8. Access to a wide variety of experiences and resources with the chance for a wide variety of contact, interaction, and communications
9. A built form that is compatible with and enhances the preceding characteristics
10. An optimum level of appropriate public health- and sick-care services accessible to all
11. High health status (high levels of positive health and low levels of disease)

the factors and the interrelations between them. It is recommended that although social scientists, psychologists, and medical or public health researchers have not often collaborated, there are good reasons for them to share experience and apply interdisciplinary approaches.

The 11 qualities of a healthy city listed in **Table 1** need to be understood and applied in specific localities using innovative empirical research and professional practice. This stems from the fact that many contributions that are meant to address health promotion and prevention have not been wholly successful even though many urban planners, public health officers, and medical practitioners are convinced that they have the 'right answers.' There is an urgent need for innovative approaches to address the continuing failure of the wealthiest countries of the world to provide all citizens with secure employment, affordable housing, and appropriate health care that meet at least minimal requirements. The failure of so-called 'model' housing estates and urban planning projects constructed in the 1960s and 1970s in numerous cities around the world clearly shows that new ideas, working methods, objectives, and criteria are needed (Burridge and Ormandy, 1993; Halpern, 1995).

Our incapacity to deal with the above-mentioned problems is related to the complexity of dealing with urban health, the compartmentalization of scientific and professional knowledge about urban ecosystems, the bureaucratic division of responsibilities in cities, and the increasing diversity of living conditions between various cities and within specific cities. In addition, the lack of effective collaboration between scientists, professionals, and policy decision-makers has led to the 'applicability gap' in sectors that deal with urban planning, public health, and many other sectors concerned with the construction and maintenance of cities. These shortcomings of mainstream scientific research and professional practice are not necessarily the result of the lack of political commitment, financial resources, or viable propositions. They are, above all, the logical outcome of the narrow vision of so-called experts who do not address fundamental issues but only topics isolated from their urban context.

In order to deal with these limitations, at least four sets of obstacles need to be overcome. First, those conceptual frameworks that do not recognize the pertinence of an ecological interpretation of urban health and living conditions need to be challenged. Second, those methodological contributions that value rational, quantified interpretations of illness and disease at the expense of qualitative interpretations of health and well-being need to be replaced by more holistic research. Third, the segmentation and bureaucratization of professional knowledge and expertise often at the expense of the experience of laypeople should be challenged. Finally, the philosophy and conceptual framework for the education and training of professionals and scientists is too shortsighted. Some of these shortcomings are being addressed by the WHO Healthy Cities project in order to promote health and well-being of citizens.

The WHO Healthy Cities Project

The Healthy Cities project was founded in 1987 by 11 European cities and the WHO Regional Office for Europe. Today, there are more than 30 national and regional networks in Europe involving about 600 municipalities, now complemented by many hundreds more in each of the regions of the world (Goldstein, 2000). The Health for All strategy provides the strategic framework for this project. The Healthy Cities project in the WHO European region includes four main components. First, the designated cities are committed to a comprehensive approach to achieving the goals of the project. Second, national and subnational networks together with Euronet facilitate cooperation between partners. Third, multi-city action plans are implemented by networks of cities collaborating on specific issues of common interest. Finally, special (model) projects are being implemented in Central and Eastern Europe.

The Healthy Cities project involves collaboration among sectors by formulating a 'City Health Plan' that identifies the interrelations between living conditions in urban areas and the health of the residents (Green et al., 2003). It is argued that health can be improved by addressing the physical environment, and the social and economic determinants of health in all situations (such as the home, the school, the workplace). This broad interpretation has meant that equity and social inequalities are identified as key factors in cities that need to be addressed. In particular, the plight of vulnerable social groups (including the handicapped, homeless, unemployed, single mothers, and street children) is ranked as a high priority for interventions. This approach is meant to focus not only on specific groups but also on particular neighborhoods where there are concentrations of vulnerable people with relatively high health risks.

Supportive Environments

The concept of supportive environment has been used in the Healthy Cities project to emphasize that policy definition and implementation should focus on all the determinants of

health, not just those within the health sector (Bistrup, 1991). Therefore, it includes the role of physical environmental factors that influence health and not just the lifestyle of individuals and groups in specific localities. In addition, it is not limited to the physical characteristics of the environment because it accounts for the cultural, social, economic, and political dimensions. When these dimensions are explicitly addressed then it is necessary to deal with equality and equity in societies and how these impact on health and well-being in precise residential environments.

From this perspective, the layout, design, and maintenance of residential environments should meet the requirements of all groups of the population including the increasing number of people with special needs, in particular, the most vulnerable in society: for example, the homeless, a group that comprises an increasing number of adolescents and young adults in industrialized countries; the elderly who need domiciliary care; people with disabilities who require easy access to and within housing units; single-parent households that may need access to special child care services; and refugees and immigrants that have specific cultural customs in and outside their housing unit that should be accommodated. This suggests that an integrated, interdisciplinary understanding of how the health of populations in precise localities can be improved is essential if a range of interventions are to be applied effectively again at the beginning of the twenty-first century.

An interesting example of the application of this approach is public transport in the region of Zurich, the largest city in Switzerland. This region accommodates 1 million residents. The public administration has implemented government policies to promote public transport at a lower economic and ecological cost than reliance on private motorcars. Public transport is more energy efficient and less polluting than private motorcars, especially when electric motors are used for trains and trams. Public transport uses less land surface area per passenger than the roads and car parking required for private motorcars, thus conserving natural landscapes and sites for allotments and leisure activities. Since the implementation of this policy in the 1980s an electric train network has been coordinated with tram and trolley bus routes so that a person has access to all neighborhoods in the region without the need to use a motorcar. The implementation of this coordinated transport policy in the Zurich region has led to a substantial increase in the use of public transport from 363 journeys per person per year in the 1980s to 532 journeys per person per year in the 1990s.

During the same period, the city administration implemented plans for diverting road traffic from residential streets to main roads and converting some of these streets into pedestrian zones with landscaping and children's playgrounds. In addition, new residential neighborhoods have been consciously planned with medium- and high-density residential buildings and other functions that are accessible for pedestrians in the region who can use public transport. The environmental, economic, health, and social impacts of this vast coordinated policy should not be underestimated. The coordinated transport project in Zurich shows the wide range of benefits of collaboration spanning across several sectors and explicitly dealing with health in tandem with the qualities of the local residential environment. It stems from a critique of piecemeal approaches to road transport, car parking, and traffic safety. This coordinated approach across several sectors reinterprets accessibility and mobility in and between urban areas in terms of not only public and private modes of transport but also air and noise pollution, consumption of nonrenewable resources, monetary costs and public investments, active and sedentary lifestyles, as well as health and well-being. This ecological approach not only raises questions about the high priority attributed to private motorcars during the twentieth century but also shows that direct investments in efficient public transport systems and pedestrian precincts can also be considered as indirect investments to promote environmental quality, reduce energy consumption and air pollution while promoting health-promoting behavior (physical activity by walking and cycling) and human well-being.

Conclusion

Today, policymakers in most countries still have great difficulty in measuring, describing, and explaining constancy, change, and differences in health, housing, and environmental conditions in cities. Part of the difficulty has been the lack of systematic data collection (Lawrence, 1996). A dynamic set covering several sectors is required across a range of administrative levels and geographical scales. Alone, official statistics based on national census returns do not provide comprehensive accounts of the quantity and quality of the housing stock, urban infrastructure, and services in rapidly developing urban areas and they ignore illegal buildings in informal settlements. Today, there are several kinds of innovative techniques and tools that can be used to monitor and analyze the spatial distribution, dynamics, and interrelated nature of environmental, housing, demographic, and health profiles in urban areas. There have been significant developments in the collection and interpretation of data, indicators, and information during the 1990s that warrant further systematic applications. For example, geographical information systems can represent data from diverse sources (including remote sensing) in order to identify relationships between the represented components of urban areas. This relatively new approach has not been widely applied in the health sector in many countries, but its use to monitor ambient environmental conditions, the built environment, and the incidence of ill health and disease warrants a commitment from public administrations at both national and local levels.

Evidence on the determinants of health of urban populations has increased since the 1990s, especially in relation to the socioeconomic inequalities of people living in urban areas. Nonetheless, too little attention has been given to assessing the effectiveness of interventions to change these determinants in specific localities and monitor the outcomes on health. If these evaluations are not available, then policy decision-making is handicapped. More research is required to evaluate different kinds of interventions that improve the built environment in urban areas in order to identify impacts on health and well-being by the collection of time series data. Environmental impact assessment, health impact assessment, and cost–benefit analysis can be used to study interventions that are meant to promote health. Today, there are too few studies that evaluate

different approaches for the communication of information to inform and educate laypeople, even though, studies show that the health impact of water supply and sewage disposal is influenced by the way the population uses (or misuses) new infrastructure and services. In addition, this research can identify economic, cultural, and political barriers to change. An understanding of the political feasibility of policy options should include an account of these barriers, as well as the legal framework and the behavioral norms and values of the population.

See also: Ecology and Health; Environmental Stress and Health; Globalization and Health; Injuries and Accidents: Psychosocial Aspects; Migration and Health; Social Exclusion, Social Deprivation and Health; Socioeconomic Status and Health; Violence as a Public Health Problem.

Bibliography

Bistrup, M.-L., 1991. Housing and Community Environments: How They Support Health. National Board of Health, Copenhagen.

Boyden, S., 1992. Biohistory: The Interplay between Human Society and the Biosphere. Past and Present. UNESCO, Paris.

Boyden, S., Millar, S., Newcombe, K., O'Neill, B., 1981. The Ecology of a City and It People: The Case of Hong Kong. Australian National University Press, Canberra.

Burridge, R., Ormandy, D. (Eds.), 1993. Unhealthy Housing: Research, Remedy and Reform. E & FN Spon, London.

Galea, S., Vlahov, D. (Eds.), 2005. Handbook of Urban Health: Populations, Methods and Practice. Springer, New York.

Goldstein, G., 2000. Healthy cities: overview of a WHO international program. Reviews on Environmental Health 15 (1–2), 207–214.

Green, G., Acres, F., Price, C., 2003. City health development planning. In: Tsouros, A., Farrington, J. (Eds.), WHO Healthy Cities in Europe: A Compilation of Papers on Progress and Achievements. World Health Organization European Office for Europe, Copenhagen, pp. 103–133.

Hancock, T., Duhl, L., 1988. Promoting Health in the Urban Context. FADL Publishers, Copenhagen.

Hippocrates, 1849. On air, waters, and places. In: The Genuine Works of Hippocrates. Translated with a Commentary by Francis Adams. The Sydenham Society, London.

Halpern, D., 1995. Mental Health and the Built Environment. Taylor and Francis, London.

Hartig, T., Lawrence, R.J. (Eds.), 2003. The residential context of health. Journal of Social Issues 59, 455–676 (special issue).

Laughlin, C., Brady, I. (Eds.), 1978. Extinction and Survival in Human Populations. Columbia University Press, New York.

Lawrence, R.J., 1996. Urban environment, health and the economy: cues for conceptual clarification and more effective policy implementation. In: Price, C., Tsouros, A. (Eds.), Our Cities, Our Future: Policies and Action Plans for Health and Sustainable Development. World Health Organization European Office for Europe, Copenhagen, pp. 38–64.

Lawrence, R.J., 2000. Urban health: a new research agenda? Reviews on Environmental Health 15, 1–11.

Lawrence, R.J., 2001. Human ecology. In: Tolba, M.K. (Ed.), Our Fragile World: Challenges and Opportunities for Sustainable Development, vol. 1. Eolss Publishers, Oxford, pp. 675–693.

McMichael, A., 2000. The urban environment and health in a world of increasing globalization: issues for developing countries. Bulletin of World Health Organization 78 (9), 1117–1126.

Park, A., Burgess, E., McKenzie, R., 1925. The City: Suggestions for the Study of Human Nature in the Urban Environment. University of Chicago Press, Chicago.

World Health Organization, 1946. Constitution. World Health Organization, Geneva.

World Health Organization, 2000. Healthy Cities in Action: 5 Case-studies from Africa, Asia, Middle East and Latin America. World Health Organization, Geneva. Document WHO/SDE/PHE/00.02.

Relevant Website

www.euro.who.int/healthy-cities – World Health Organization.

Urban Life: Future Developments

MA Hughes, University of Pennsylvania, Philadelphia, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16029–16031, © 2001, Elsevier Ltd.

Abstract

The future of urban settlements, as was their past, will likely be driven by three fundamental mechanisms. First, the changing technology of the digital economy will reshape the possibilities of urban life over the next generation. Second, the limits of post-1945 development patterns, especially the low density suburban form prevalent in the United States for 50 years, will be tested. Third, the attractive power of centers will wax or wane; given the diminished power and autonomy of the large industrial urban center over the preceding century, this third mechanism is likely to be closely related to changes in the first two. The multiple contingencies of these and other mechanisms will frame the important issues facing decision makers over the coming decades. Three are already apparent. First, if largeness is a consequence of a passing industrial era, then how will the more fundamentally urban quality of density be manifest in the coming decades? Second, if the well-known mismatch between formal and informal urban organization (local–state–national versus neighborhood–regional–global) is accurate, then how resistant or malleable will formal structures be in the service of informal imperatives? Third, urban life will confront the access issue of personal automobiles versus every other form of mobility. While epitomized in US settlements, this tension is rapidly becoming global in both its prevalence and its consequences.

The future of urban settlements and the lives lived therein will depend on a mix of predictable trends and human contingencies. History is the best guide to those future trends, and three are discussed here: technology, the limits to growth at urban peripheries, and the attractive power of urban centers. Caution is perhaps the best guide to discussing future contingencies, but three major issues are so apparent that they will likely confront urban decision makers in the coming generation: large versus dense settlement structures, local versus regional decision-making processes, and automobiles versus other means of mobility and access to urban resources.

Three Determinants of Urban Futures

Since the beginning of the urban era, technologies in various forms and the relative 'gravitational' pull of urban centers and rural hinterlands have been powerful determinants of urban life. These fundamental mechanisms are surely the product of human action and are never completely beyond a human capacity to intervene (Mumford, 1961). Yet, the power of such mechanisms (for example, the technology of production and the related attractiveness of center versus peripheral locations) often exerts a defining influence on urban settlements in ways that can overwhelm individual or collective action in a given time or place. Introduced below are three such large trends that may be expected to influence the future of urban life.

Waves of innovation in communications and transportation technologies occurred throughout the twentieth century. At present, these roiling changes are understood and discussed in a summary way as the emergence of an Economy (Mitchell, 1995). A new digital geography will reshape the possibilities of urban life over the next generation (Kotkin and Siegel, 2000). In particular, the Economy changes many of the incentives (e.g., economies of scale and various costs that increase with distance) that led to concentration in urban centers in the nineteenth and twentieth centuries. By weakening and sometimes inverting such incentives, the New Economy is characterized by much more footloose (or non-location-specific) production and consumption. It is now both possible and increasingly common to participate in employment, education and many other daily activities well beyond the commuting radius of a daily urban system. Under such conditions, the nineteenth and twentieth century comparative advantages of specific locations (e.g., entrepot for commodities, large physical agglomerations of workers and consumers) are being eclipsed by locational amenities such as weather and natural beauty as well as cumulative advantages that accrue to particular locations that happened to innovate first in some industry or activity.

Second, the limits of post-1945 development patterns, especially the low density suburban form prevalent in the United States for 50 years, will be tested and some expect those limits to be reached. Technological changes (with, in the United States, the full support and encouragement of several national policies) have maximized the opportunity for low-density choices and this has led to a half-century of suburbanization in urban settlements, in which the increase in built-up land area has far exceeded the increase in population or employment. The resulting development patterns have been associated with a long list of social costs, including increased traffic congestion and lengthening commutes, loss of both open space and agricultural land uses, a rise in social and ethnic segregation as well as in aggregate housing costs, and the decline of communal participation and social capital (Downs, 1995; Kunstler, 1996; Duany et al., 2000). These critiques have sought to portray as unsustainable the current low-density development regime, often labeled sprawl. Policy recommendations range from government controls on new peripheral growth to the modification of cultural and aesthetic values related to suburbanization.

Third, the attractive power of centers will wax or wane; given the diminished power and autonomy of large industrial urban centers over the preceding century, this third mechanism is likely to be closely related to changes in the first two. Paradoxically, the decentralizing trends of both the digital economy and a low-density development regime have created a powerful role for centralized command and control exploited by a select group of 'global cities' (Sassen, 1991). This centralizing function is also present on a smaller scale within regional systems, somewhat reviving a role played for smaller urban centers. In addition to this emerging economic function, urban centers may also gain a new importance derived from the limits on further growth at the urban periphery (Rusk, 1999), whether those limits are imposed by technology or by governments.

Three Issues Facing Future Urban Decision Makers

The multiple contingencies of these and other mechanisms will likely frame the important issues facing decision makers over the coming decades. Without knowing how the possibilities will turn out (a placeless city of bits or not? More growth on the undeveloped urban edge or less? A sustained re-centering of urban settlements or their permanent de-centering?), it is imprudent to list specific issues to be grappled with. But three critical issues are already apparent in the foreseeable (or more seeable) future.

First, if largeness is a consequence of a passing industrial era, then how will the more fundamentally urban quality of density be manifest in the coming decades? While the emergence of a New Economy is neither fully understood nor even universally accepted, the technological framework within which urban settlements are constructed and maintained seems to have changed profoundly. While industrialism's appetite for physical agglomeration has weakened if not collapsed, there appears to be a powerful revival of demand for density on a very large scale in a select set of global cities and on a smaller scale in centers throughout the urban system. This demand for density appears to be related less to production and distribution, unlike during the industrial era, and more toward lifestyle amenities. Breaking the link between density and largeness in the minds of urban dwellers and decision makers, which seems fundamental to the suburban reform agenda discussed above, will be a major challenge in coming decades. In many ways, of course, this leads back to the future and to a revival of Ebenezer Howard and Garden City and New Town ideas.

Second, if the well known mismatch between formal and informal organization (local–state–national versus neighborhood–regional–global, respectively) is accurate, then how resistant or malleable will formal structures be in the service of informal imperatives? Many of the changes demanded in the preceding paragraph require collective action and governmental capacity. That the ideas being debated are neither new nor complex suggests the difficulties of actually implementing a desirable future. Whether or not the public apparatus of decision–making related to urban life can be adapted to match the new and emerging scales at which urban life is being lived through new technologies and relationships is perhaps the single most important issue framed by observable trends.

Third, urban life will continue to confront the issue of automobiles versus every other form of mobility. While epitomized by US settlements, this tension is rapidly becoming global in both its prevalence and its consequences. Regardless of the scale and governance issues outlined above, physical access to some resources (perhaps a different set than those of today) will continue to be a fundamental rationale for urban settlements. The access problem is solved through mobility of some form. The worldwide trend toward individual automobile mobility (whether that trend moves from the large base of car use in the United States or the smaller but rapidly growing base found in the developing world) will likely join if not eclipse other specific critical priorities facing urban decision makers. Ironically, the negative environmental impact of increasing automobile use is as global an urban issue as the positive relationships of the New Economy. The intersection of combustion engines, urban settlement patterns, and cross-national environmental degradation will challenge urban life in powerful and perhaps decisive ways.

See also: Mumford, Lewis (1895–1990); Urban Geography; Urban History; Urban Sociology; Urban Sprawl; Urban Studies: Overview.

Bibliography

Downs, A., 1995. New Visions for Metropolitan America. Brookings Institution Press, Washington, DC.

Duany, A., Plater-Zyberk, E., Speck, J., 2000. Suburban Nation: The Rise of Sprawl and the Decline of the American Dream. North Point Press: Farrar, Strauss and Giroux, New York.

Hall, P., 1988. Cities of Tomorrow. Blackwell, London.

Kotkin, J., Siegel, F., 2000. Digital Geography: The Remaking of City and Countryside in the New Economy. The Hudson Institute, Indianapolis, IN.

Kunstler, J.H., 1996. Home from Nowhere: Remaking Our Everyday World for the 21st Century. Touchstone: Simon & Schuster, New York.

Lynch, K., 1981. Good City Form. MIT Press, Cambridge, MA.

Mitchell, W.J., 1995. City of Bits: Space, Place, and the Infobahn. MIT Press, Cambridge, MA.

Mumford, L., 1961. The City in History: Its Origins, Its Transformations, and Its Prospects. Harcourt, Brace & World, New York.

Rusk, D., 1999. Inside Game, Outside Game: Winning Strategies for Saving Urban America. Brookings Institution Press, Washington, DC.

Rybczynski, W., 1995. City Life. Touchstone: Simon & Schuster, New York.

Sassen, S., 1991. The Global City. Princeton University Press, Princeton, NJ.

Urban Megaprojects

Fernando D Orueta, Universidad de La Rioja, Logrono, La Rioja, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the context of the intercity competition favored by urban entrepreneurialism, urban megaprojects play a significant role not only for their ability to attract investment and generate high returns, but also as iconic symbols of the new economy. In addition to their scale, they are characterized by the mixed nature of their components and land uses, the use of various forms of public–private partnership and their preferred location in areas of former residual use. Their expansion to a large number of cities around the world has increased their diversity and with it the need to expand comparative analytical studies.

Introduction

A historical reading of urban processes can identify over time the development of large urban projects of various types in any of the world's most significant cities. However, when the term megaproject is used in contemporary urban studies it does not refer generically to these developments, but rather to specific types of intervention in the urban space.

According to Altshuler and Luberoff (2003), during the 1950s and 1960s there was a strong flow of investment in American cities. A highly significant part of these investments went into promoting megaprojects, largely linked to new transport infrastructures, e.g., freeways and airports, but also aimed at urban renewal. Just as occurred in much of Western Europe, these investments must be linked to the development of postwar Fordist state (Diaz Orueta and Fainstein, 2008).

Europe also had to rebuild its war-torn cities. In fact, in European cities the emphasis was on the construction of housing, both in districts destroyed by war and in previously nonurbanized areas on the urban periphery (Fainstein, 2008). By contrast, interventions in the USA primarily focused on inhabited areas, in many cases involving the demolition of slums. While public provision of housing generally played a crucial role in Europe, the United States relied primarily on the private sector and nonprofit organizations to invest in sites freed up by slum clearance.

Megaprojects played an important role in shaping a new urban environment characterized, among other features, by proliferating highways used by growing numbers of people, who progressively agreed to the use of private vehicles and the construction of large buildings, which were said to ensure a more efficient functioning of the city, while improving housing shortages and facilitating the attraction of capital (Diaz Orueta and Fainstein, 2008).

Since the late 1960s and very clearly in the 1970s, such developments began to meet problems, for several reasons. One was growing opposition by movements to protect the environment and national heritage (Altshuler and Luberoff, 2003). A significant proportion of new buildings erected on cleared land represented a break with the collective memory of the city. In addition, various working class movements also put up strong arguments based on the social consequences of many megaprojects, which often involved the displacement of hundreds of thousands of people, especially in the USA, with the consequent destruction of communities, and also exposing the way potential benefits were unevenly distributed socially. As Salet (2008) suggests, from the late 1960s progrowth coalitions were faced with the debate about social polarization, articulated around issues, such as social inequality, racism, education, and unemployment. Already in the 1970s, and especially following the oil crisis, environmental concerns were also forcefully entering urban agendas. On the other hand, increasing fiscal constraints and decentralization policies on the part of the state acted as additional constraints (Diaz Orueta and Fainstein, 2008). It is clear that by the mid-1970s developers had already entered what some writers have termed the 'do no harm' phase, as projects were transformed to avoid emerging resistance (Altshuler and Luberoff, 2003).

Within a few years a new generation of urban megaprojects began to appear. From the late 1970s thinking centered on the idea that the fundamental problem cities had to face was economic crisis and decline. There emerged what some authors have called the new urban policy, aimed at ensuring economic growth of cities, promoting competitiveness and implementing new forms of local government to respond to the demands of economic globalization (Swyngedouw et al., 2004). By the early 1990s, these trends had become consolidated, giving rise to a new generation of urban megaprojects, the evolution of which has been inescapably linked to the neoliberal urban restructuring experienced by North American and European cities (Brenner and Theodore, 2002). Megaprojects came used as a way to reposition cities in the global economy (Lehrer and Laidley, 2008).

A New Generation of Urban Megaprojects

In this new era, development of megaprojects constituted a central tool in strategies of political and economic elites to improve urban competitiveness (Del Cerro, 2013). As Fainstein (2008) argues in her analysis of New York, London, and Amsterdam, these actions have been shown to be extremely useful in attracting international investment of many kinds. In these three cities, as in others, there was a major transformation of their economic bases, leaving behind industrial specialization to focus increasingly on offices, leisure, and tourism, while markedly increasing new spaces for luxury homes. A significant number of the urban megaprojects have been developed in old abandoned industrial areas that have been completely

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74052-9

transformed, giving way to new spaces that symbolically represent the progress of the new economy. Furthermore, although the megaprojects are placed in specific locations, their analysis requires consideration of different territorial scales. Urban megaprojects are important for lobbying and decision-making networks at interregional, transnational, and international levels. Therefore, the analysis of local policy making must be carried out from a multilevel perspective (Salet, 2008).

New megaprojects have not led to the demise of other developments focused on building large transport related infrastructures, the energy sector or supply of basic necessities, such as water (Diaz Orueta and Fainstein, 2008). In fact, the analysis of these types of megaprojects remains a very noteworthy subject of attention, especially with regard to major transport infrastructure (Flyvbjerg et al., 2003; Priemus and van Wee, 2013).

However, contemporary urban research focuses primarily on other types of megaprojects. These are large developments that feature different content: mixed residential use, tertiary spaces, utilities of various kinds, new transport infrastructure, etc. Megaprojects thus affect large surface areas where a change in land use and new construction or total rehabilitation of old buildings occurs. Their symbolic spatial component is crucial; hence the centrality given to the designers of these spaces, converted into real superstar architects who stamp their distinctive signatures on iconic reference points of contemporary cities.

Moreover, the scale of the projects poses a growing challenge to traditional financial instruments and urban management (Lungo, 2002). Directly linked to the change of scale, one of the clearest peculiarities of new megaprojects as already indicated is their mixed character. For example, when analyzing the waterfront development in the city of Toronto, Lehrer and Laidley (2008) emphasize the mixed use including housing, different types regarding tenure and size, built with spaces for offices and small trade, and surrounded by parking areas, repositories, and natural amenities.

To finish this outline of an introductory approach to the new generation of urban megaprojects, it is necessary to add a third dimension to complete the above references to its mixed character and scale. This is the location: what urban spaces are chosen to situate the megaprojects? In general, the preferred location is in old industrial areas, ports in decline or other terrains with a residual urban use, usually but not necessarily outside the city center. It is common that these spaces are not well integrated with the rest of the city, reinforcing the tendency toward specialization and spatial monofunctionalization (Salet, 2008). In the US and European cities, these major projects, characterized by repetitive architectural designs that contribute to the commodification of the city, are generally located in areas devoid of interaction between old and new spaces, large and small, which gives centers its special and distinctive character (Fainstein, 2008) (**Figure 1**).

On the other hand, it is often noted that in contrast to what happened with the old megaprojects, location on urban fringes has reduced social conflicts, as the population it less likely to be directly affected by their construction. However, this latter view must be qualified since the increasingly abundant literature on urban megaprojects shows several case studies in moderately developed countries in which the choice of location has impacted directly and negatively on important population groups. This is the case in Buenos Aires, where the development of Puerto Madero megaproject, on the waterfront, threatens the survival of the Villa 31, a shantytown inhabited by more than 20 000 people and located alongside some of the most sought after neighborhoods in the city. Another example is the attempt to build a new airport for Mexico City, the location of which would impact directly on the lives of the communities in that

Figure 1 Buenos Aires: Puerto Madero Waterfront. Photo by Fernando Diaz Orueta.

area. In fact, this megaproject led to a social conflict of huge proportions that eventually brought about its closure in 2002 (Flores Dewey and Davis, 2013). In another context, Beijing's hosting of the Olympic Games and the major urban developments it deriving therefrom generated the transformation of large areas of the city causing, e.g., the gentrification of different historic neighborhoods (Ren, 2009).

Urban Entrepreneurialism and Megaprojects

Already in the late 1980s, some authors were calling attention to the transition from a model of urban governance based on managerialism to another based on entrepreneurialism (Harvey, 1989). In this new model, urban policies favor competition among cities, in the hope that economic growth will generate prosperity, welfare, and employment for the majority of the population. Although initially developed in the United States, urban entrepreneurialism also reached the cities of Western Europe. Urban megaprojects are a strategic element of urban entrepreneurialism, transmitting an image of status and power and thus attracting investment (Dyble, 2013). Therefore, they are primarily oriented toward the market, competitiveness and profit, leaving social commitments, such as employment and affordable housing in the background. Ultimately, it comes down to getting significant returns, and that is basically achieved through the construction of luxury residences and hotels, large-footprint office towers, and shopping malls (Fainstein, 2008).

Compared to other developments, the process of definition, approval, and implementation of urban megaprojects makes their management enormously complex. In fact, a large part of the literature on these behemoths analyzes not only the emergence of new forms of management and planning, but also funding and taxation mechanisms. All this is in a context of growing multilevel interrelationships in which the number and diversity of those involved in the decision-making process multiplies. Given the complexity of these projects, their implementation spreads over time, subject to the vagaries of economic cycles. Therefore, study of the coalitions supporting megaprojects and their multilevel relationships is essential in order to understand not only their gestation and administration, but also their sustainability over time and even, when it does occur, their failure.

Such coalitions work hard to obtain social support in favor of the development of megaprojects. This support is achieved through institutional and politicoeconomic structures that redirect public attention away from controversy and, therefore, generate or at least try to generate the depoliticization of megaprojects (Lehrer and Laidley, 2008). For example, in the case of the waterfront development in Toronto, a city studied by these authors, public participation was not initiated at the time of considering new development choices, but only once the decision as to which option was preferable had been taken. In fact, the majority of megaprojects erected in many cities of the world have not had any participative processes, either before or after approval.

Such has been the case with most of the megaprojects in Latin American cities. In fact, in these cities it is common, once the decision to go ahead with such a project has been made, for a process to operate that openly bypasses both the urban planning offices of and even political channels, regarded as overly reliant on democratic public debate (Lungo, 2002). As a result, these projects serve the interests of a small minority while producing displacement of the less advantaged from their neighborhoods, further deepening the processes of urban segregation and social exclusion. This tends to reinforce the existence of islands of modernity surrounded by large areas of poverty.

Promoters of megaprojects linked to major transport infrastructure are also perceived to evade and even violate the practices of good governance, transparency, and participation in decision-making processes (Flyvbjerg et al., 2003). Such practices are perceived as hindrances hampering development of the projects. This approach can produce profound and lasting environmental damage, given that by not taking into account the need for impact assessments or doing them in a very lax way, infrastructures that should never have been erected or built without the necessary precautions are constructed.

In addition, this markedly business-friendly orientation exerts substantial pressure on local governments, especially in a climate of severe financial constraint. For example, Dyble (2013), analyzing the case of Skyway in Los Angeles, highlights the way its management evolved so that, in the context of globalization, the emphasis on a transport policy that prioritizes long-distance connections rather than local or regional integration led to this major infrastructure becoming an important part of urban governance now run largely according to financial criteria. In fact, given public finance constraints, governments have been more inclined to negotiate the entry of private funding. But there is no doubt that the growing presence of private funding means that the public sector loses an important part of the ability to control projects.

Since the contexts where megaprojects arise are varied, both socioeconomically and from the political and cultural perspectives, it is possible to detect significant differences and peculiarities between different experiences. In this regard, comparative studies are very revealing. For example, Doucet (2013) discusses two projects, one local initiative in the city of Rotterdam and one in Glasgow, promoted by the private sector. Comparative analysis identifies important differences between European cities, highlighting the way that when municipal authorities lead and coordinate the project, business objectives are more easily achieved than when the public sector is limited to facilitating the activities of private developers. Doucet, without abandoning the perspective of urban entrepreneurialism, also stresses the need to analyze the role of individual actors and leaders involved in the process.

In an equally comparative key, Fainstein (2008), analyzing different megaprojects in the cities of New York, London, and Amsterdam, highlights how in the latter two cities, and particularly in Amsterdam, a true social democratic ethos had survived. This means that, in contrast to what happened in New York, the two European cities at least until 2008 held developers to certain social and environmental requirements. Above all, even with difficulties the welfare state seems to present greater resistance in countries like the Netherlands, which would imply a model of urban governance with greater government leadership and a certain level of commitment to

egalitarian goals. In London, recent research shows how some of the large urban operations associated with hosting the 2012 Olympics have generated powerful gentrification processes (Watt, 2013). In other European cities, such as Barcelona, the urban model is increasingly oriented toward competitiveness, displacing other considerations to the background (Degen and Garcia, 2012).

In the coming years, with the consolidation of the neoliberal vision of the European Union, it is foreseeable that the reality will increasingly resemble that of the USA, with a clear deepening of social inequality, a simultaneous increase in poverty indicators and the concentration of wealth in the hands of very small sectors of the population.

Public–Private Partnership

Although the relationship between the public and private sectors is far from new in the history of urban development, it has become more intense with the expansion of urban entrepreneurialism adopting new and varied forms of collaboration. This is understood as essential for competition between cities and has reached such extremes that it is sometimes not easy to distinguish the specific role played by the public and private sectors and the meaning of the performance of each. Therefore, results of the collaboration between both sectors vary from case to case, although insofar as there is a very strong tendency to prioritize economic benefits, the extent of their commitment to egalitarian goals is generally very limited.

Cooperation between the public and private sectors has resulted in the creation, in numerous cities and under various formulations, of institutionalized collaborative spaces. For example, in 1991 Bilbao created the Bilbao Metropoli-30 association, formed by different public and private entities, which was instrumental in establishing the new image of a city trying to overcome a deindustrialization crisis by projecting an innovative character (Rodriguez and Martinez, 2003). A year later, the creation of the public company Bilbao Ria 2000 brought it well within the governance by business interests. This new company was responsible for coordinating projects and developments subsequently implemented in the fields of urban planning, transport and the environment, including the Guggenheim Museum in the Abandoibarra district, the flagship project in reshaping Bilbao (Gonzalez, 2003; Vicario and Martinez Monje, 2003) (**Figure 2**).

Analysis of these entities helps identify the composition of the coalitions that promote megaprojects and the specific weight and capacity for determination of urban agendas of each of their different public and private members. While in some cities leadership can be held by local public authorities trying to control and direct growth, in others it is business sectors that set priorities and determine ways forward. This happened, for example, in most Spanish cities during the economic boom before the outbreak of the crisis in 2007/2008, when corporate power and political will completely overlapped to favor the growth of a massive speculative bubble that was reinforced inter alia, by the development of a number of megaprojects of diverse nature, new museums and cultural centers, major housing estates with tens of thousands of new homes, oversized airports, theme parks, large office complexes, industrial parks, shopping malls, etc.

To understand the real scope of the public–private partnership in the development of megaprojects, consideration of the socioeconomic and political context is crucial. For example, in many countries these large urban operations have coincided in time with profound economic and political changes that have greatly affected the role of the state. Specifically in Latin

Figure 2 Bilbao: Guggenheim Museum. Photo by Fernando Diaz Orueta.

American cities, in the 1990s as a result of decentralization and privatization, there was a redistribution of regulatory powers between different levels of government (local, regional, national), as well as the roles played by private institutions and local community organizations (Lungo, 2002). This led to the emergence of new urban management structures reflecting the confluence, sometimes not without conflict, of the public and private sectors, and the current multilevel nature of urban processes.

Some writers believe that, given the complexity in terms of networks of entities involved, the heterogeneous and not infrequently divergent variety of interests and contradictions inherent in megaprojects, it is to be highly recommended that management should be led by the public sector and, in particular, by local governments. Lungo (2002), taking as reference the Latin American urban reality but reaching conclusions that can be extrapolated to other territories, states that municipalities are better able to ensure that other levels of government in a reduction of the negative external factors of megaprojects, ensuring the involvement of locals and weaker members and promoting a more balanced distribution of benefits.

Public–Private Partnership and Urban Planning

The relationship between urban megaprojects and urban planning processes is a relevant concern. In many cities, the development of these projects collides with the procedures and regulations characteristic of the traditional tools of urban planning. In fact, new ideas have been boosted since the 1980s, such as strategic planning, one of the main objectives of which has been to expedite the implementation of these major developments, favoring public–private partnership (PPP) strategies. Also, as Salet (2008) recalled, it should not be forgotten that today urban planning is not a question of urban or regional planners or even local coalitions, but that an approach based on interconnectivity in the network society.

Also with reference to planning it is necessary to stress the need to differentiate urban realities, given that the meaning of planning in different cities is very distinct. For example, when Bezmez (2008) analyzes Istanbul she points out that, unlike most European or American cities, in that city planning has never been instrumental in the development of megaprojects. Conversely, Fainstein (2008) in her comparative analysis of New York, London, and Amsterdam highlights the importance of the planning process in these three cases, as well as the peculiarities she found. So while in London and New York megaprojects are part of a more comprehensive planning process than seen in earlier times, in Amsterdam, a city where overall planning has historically dominated, a rethinking of spatial strategies beyond a polycentric model can be observed. In contrast to the above, and referring to the city of Manila, Shatkin (2008) points out the need to carry out a proper analysis that goes beyond the idea that globalization implies an inexorable process of 'Westernization'. Thus, in relation to the planning and development of megaprojects, Shatkin concludes that Manila is undergoing a massive privatization of urban and regional planning. That means that the major developers have become responsible for urban development planning, assisted by government authorities.

Results of Megaprojects

Many of the scientific studies that examine the new generation of urban megaprojects maintain a critical perspective on their results. The supposedly favorable effects on urban economies and regional development are difficult to measure in quantitative terms and, on the contrary, there are many indicators that point to significant negative impacts from the social and environmental point of view. However, as already pointed out above, this does not mean that the promoters of megaprojects have not made a notable effort to present them publicly as key tools in moving toward a social and environmentally balanced model city. The image projected of megaprojects through urban marketing refers to a sustainable and equitable city where new and compelling urban attractions appeal to visitors and new residents interested in leisure and culture. As such, a city in which economic growth would support citizen participation, deepening democracy, social equality, and respect and even environmental recovery.

However, analytical studies of megaprojects portray a very different panorama. These studies mostly reflect the growing power of some economic and political elites that border the democratic mechanisms of decision making to boost urban entrepreneurialism. As a result, in the last two decades megaprojects have been one of the elements that have contributed to increased urban segregation and particularly gentrification.

Another aspect often queried in relation to megaprojects is that of additional costs. The already high planned budgets are almost systematically exceeded. In fact, it is common to establish demand forecasts well above actual expectations, underestimate the environmental impact and exaggerate the expected economic benefits of the megaproject. Ultimately, the low estimate of costs does not appear to be due so much to inadvertent error as to deliberate deception (Flyvbjerg et al., 2003). Thus, both in terms of cost overruns and treatment of financial risks, similar trends are detected in different types of megaprojects, including major transport infrastructures (Altshuler and Luberoff, 2003). These writers suggest the need to improve accountability measures in order to detect and resolve the systematic underestimation of costs. This is not a trivial matter, since in some cities the problems created are extremely burdensome. For example, dragging the huge debt the city of Madrid (7500 million Euros at year-end 2012) is explained in part by the significant overspending that occurred on a megaproject that was in itself enormously costly: the burial of the M30 urban motorway ring. But although the most extreme, this is not the only case of a Spanish city in which the development of one or more urban megaprojects, coupled with the usual extras, has helped lead to choke the public purse. The same thing happened with, e.g., Valencia's City of Arts and Sciences or in Zaragoza's large urban projects boosted by the International Water Expo 2008 (**Figure 3**).

Yet despite this negative balance, the use of megaprojects as instruments for construction and transformation of cities has become increasingly common, leading to what other writers have called the paradox of megaproject development (Flyvbjerg et al., 2003; Altshuler and Luberoff, 2003; Del Cerro, 2013).

Weak social opposition, partly explained by the efforts made to try to present megaprojects as virtuous developments that should therefore fall outside the scope of urban disputes,

Figure 3 Valencia: City of Arts and Sciences. Photo by Fernando Diaz Orueta.

have facilitated this apparent paradox. As already noted, in most cities in the developed world the number and intensity of social movements has been less than at previous historical moments. The strength of legitimizing discourse and sophisticated marketing strategies, together with the mixed nature of the uses of megaprojects and their peripheral location, have contributed to the maintenance of consensus, or at least of social disinterest. In these countries, the weakness of social resistance to the new generation of megaprojects, although these have sometimes occurred (Diaz Orueta, 2007), contrasts with the strength of the protests that arose over previous megaprojects (Lehrer and Laidley, 2008). But in any case, when the perspective is extended to cover the cities of the global South and other countries of intermediate development, the situation changes dramatically with important and sometimes successful social mobilizations taking place (Flores Dewey and Davis, 2013).

Problems and Methodological Challenges in the Study of Urban Megaprojects

As mentioned above, in the last 10 years there has been a sharp increase in research focused on the new generation of urban megaprojects. The remarkable growth in the number of case studies has been accompanied by an increasingly intense effort to achieve generalizable results, especially by conducting research of a comparative nature involving very different cities. However, today there are still a number of important features for which only partial answers are so far available and that, therefore, continue to pose challenges for research on aspects of urban megaprojects.

First of all, it is necessary to go further in establishing the differences and similarities between the different experiences, developing a comprehensive and systematic comparative analysis. To begin with, the comparison of urban megaprojects in the cities and countries of the global North should be extended. But besides this perspective at least two other types of investigation are required: those that compare the experience of the global North to that of the global South and those focusing exclusively on cities of the global South. In this regard, it is important not to turn the conclusions of research conducted in countries of the global North into general categories, extrapolatable to other urban areas. Of course, this does not mean that there are not more or less intense connections between some cities. For example, the experience of urban transformation undergone by Barcelona with the hosting of the 1992 Olympic Games had a major influence on the development of urban megaprojects in several Latin American cities (Lungo, 2002). However, this in no way meant slavish imitation of what happened in Barcelona. The political, economic, and social contexts, and the role played by urban planning and the public sector, are so diverse, that results necessarily differ profoundly.

In that sense, one of the main problems faced by these studies is the difficulty, even higher than in other types of research, of standardizing information to enable rigorous comparisons. This question arises not only in regard to the construction of comparable quantitative indicators, but also the great diversity in regulatory frameworks, planning traditions, functioning of political systems, economic context, etc.

A second aspect, partially addressed to date and in need of more thorough and systematic consideration, is the impact produced by megaprojects on urban social structures. According to research to date, most of these megaprojects primarily reflect the specific interests of certain political and economic elites and thus favor the development expected by these groups to the detriment of others. Therefore, as suggested by, e.g., Lehrer and Laidley (2008), it is necessary to undertake a class-based analysis aimed at understanding to what extent and how

the proliferation of megaprojects could be influencing the evolution of social inequality in cities.

Third, research to date has interpreted urban megaprojects as a given factor of the new urban reality. A characteristic of contemporary cities that can be analyzed, sometimes very critically, but for which no alternatives are discussed (Diaz Orueta and Fainstein, 2008). However, with the level of maturity that has been reached by now, megaproject research should urgently address this issue. Consequently, it is necessary to question their inevitability, consider whether it is worth keeping these megaprojects and, if so, under what conditions and for what purposes.

A fourth aspect, indubitably linked to the foregoing, is that the global economic crisis and the interpretation thereof should occupy a central place on the agenda of megaproject research. The crisis affects the whole planet, albeit unevenly, and its implications are clearly visible in the cities. This crisis has called into question the bases of the justification of urban policies of the past two decades, including urban megaprojects. In any case, given the complexity of the moment, it is premature to be sure that neoliberalism will soon be outdated.

There can be no doubt that the global crisis affects in an especially intense way some essential conditions that have allowed the development of megaprojects. Thus, for example, the difficulties of the private sector to access credit have increased, the public sector faces a situation of deep cuts and severe indebtedness and, on the other hand, the demand for new office space, housing, etc. has decreased markedly, even disappearing in some countries (Diaz Orueta and Fainstein, 2008). Furthermore, given the lengthy time required for their development, some have been paralyzed, at least temporarily, while others remain although undergoing major reformulation.

It should also be remembered that the crisis does not affect all cities with the same intensity. Consider, for example, the scenario of growth in China compared to the structural decline since 2008 has been driven economy of the countries of Southern Europe. Furthermore, while some cities are raised through the crisis trying to shape a different model, others rely on returning in a few years previous to the path of growth. In the latter case, it is not intended to change the model so as to take a series of measures to promote economic recovery. It is in this last key that should be understood otherwise incomprehensible bets, such as how, for example, that by the leisure resort Eurovegas would occupy an area of over 750 ha on the periphery of metropolitan Madrid that, if it goes ahead, would become one of the largest facilities of this type in the world. This action is not only hard to justify from the social and environmental point of view, but also in economic terms taking into account the situation of suffocation and bankruptcy that territory is currently undergoing.

A further consequence of the global crisis is a revival of citizens' rejection of megaprojects. In a context of strong social cuts, with a sharp increase in poverty, unemployment, and the number of homeless people, it is not feasible to maintain previous levels of consensus about megaprojects. The conditions that, for a long time, helped many cities to place them outside the realm of public debate have disappeared and now, on the contrary, the demand for accountability for the disastrous management of some of the megaprojects developed in previous years continues to grow. Possibly more than a few western countries have now entered a phase of social mobilization focused on urban issues, in which the debate on megaprojects occupies an important place.

Briefly, the challenges to be faced by urban research on megaprojects in years to come are varied and complex. However, the growth of studies in this field of research, the diversity of approaches and the wide geographical coverage achieved, guarantee an excellent position from which to continue to deepen our knowledge.

See also: Cities: Capital, Global, and World; Gentrification; Just City; Urban Entrepreneurialism; Urban Neoliberalism; Urban Regeneration.

Bibliography

Altshuler, A., Luberoff, D., 2003. Mega-projects: The Changing Politics of Urban Public Investment. Brookings Institution Press, Washington, DC.

Bezmez, D., 2008. The politics of urban waterfront regeneration: the case of Haliç (The Golden Horn), Istanbul. International Journal of Urban and Regional Research 32.4, 815–840.

Brenner, N., Theodore, N. (Eds.), 2002. Spaces of Neoliberalism. Urban Restructuring in North America and Western Europe. Blackwell, Oxford.

Degen, M., Garcia, M., 2012. The transformation of the 'Barcelona model': an analysis of culture, urban regeneration and governance. International Journal of Urban and Regional Research 36 (5), 1022–1038.

Del Cerro, G. (Ed.), 2013. Urban Megaprojects. A Worldwide View. Emerald Group Publishing, Bingley.

Diaz Orueta, F., 2007. Madrid: urban regeneration projects and social mobilization. Cities 24 (3), 183–193.

Diaz Orueta, F., Fainstein, S., 2008. The new mega-projects: genesis and impacts. International Journal of Urban and Regional Research 32 (4), 759–767.

Doucet, B., 2013. Variations of the entrepreneurial city: goals, roles and visions in Rotterdam's Kop van Zuid and the Glasgow harbour megaprojects. International Journal of Urban and Regional Research 37 (6), 2035–2051.

Dyble, L., 2013. Chicago and its skyway: lessons from an urban megaproject. In: Ingram, G., Brandt, K. (Eds.), Infrastructure and Land Policies. Lincoln Institute of Land Policy, Cambridge, MA.

Fainstein, S., 2008. Mega-projects in New York, London and Amsterdam. International Journal of Urban and Regional Research 32 (4), 768–785.

Flores Dewey, O., Davis, D., 2013. Planning, politics and urban mega-projects in developmental context: lessons from Mexico city's airport controversy. Journal of Urban Affairs 0 (0), 1–21.

Flyvbjerg, B., Bruzelius, N., Rothengatter, W., 2003. Megaprojects and Risk: An Anatomy of Ambition. Cambridge University Press, Cambridge.

Gonzalez, S., 2003. The role of the Guggenheim Museum in the development of urban entrepreneurial practices in Bilbao. International Journal of Iberian Studies 16 (3), 177–186.

Harvey, D., 1989. From managerialism to entrepreneurialism: the transformation in urban governance in late capitalism. Geografiska Annaler Series B 71 (1), 3–17.

Lehrer, U., Laidley, J., 2008. Old mega-projects newly repackaged? Waterfront redevelopment in Toronto. International Journal of Urban and Regional Research 32 (4), 786–803.

Lungo, M., 2002. Large urban projects. A challenge for Latin American cities. Land Lines 14 (4). http://www.lincolninst.edu/pubs/544_Large-Urban-Projects.

Priemus, H., van Wee, B. (Eds.), 2013. International Handbook on Megaprojects. Edward Elgar Publishing, Cheltenham.

Ren, X., 2009. Olympic Beijing: reflections on urban space and global connectivity. International Journal of the History of Sports 26 (8), 1011–1039.

Rodriguez, A., Martinez, E., 2003. Restructuring cities: miracles and mirages in urban revitalization in Bilbao. In: Moulaert, F., Rodriguez, A., Swyngedouw, E. (Eds.), The Globalized City. Economic Restructuring and Social Polarization in European Cities. Oxford University Press, Oxford.

Salet, W., 2008. Rethinking urban projects: experiences in Europe. Urban Studies 45 (11), 2343–2363.

Shatkin, G., 2008. The city and the bottom line: urban mega-projects and the privatization of planning in Southeast Asia. Environment and Planning A 40 (2), 383–401.

Swyngedouw, E., Moulaert, F., Rodriguez, A., 2004. Neoliberal urbanization in Europe: large scale urban development projects and the new urban policy. In: Brenner, N., Theodore, N. (Eds.), Spaces of Neoliberalism. Urban Restructuring in North America and Western Europe. Blackwell, Oxford.

Vicario, L., Martinez Monje, M., 2003. Guggenheim effect? The generation of a potentially gentrifiable neighbourhood in Bilbao. Urban Studies 40 (12), 2383–2400.

Watt, P., 2013. 'It's not for us'. Regeneration, the 2012 Olympics and the gentrification of East London. CITY 17 (1), 99–118.

Relevant Websites

http://www.bm30/Welcome_uk.html – Bilbao Metropoli 30.
http://www.inura.org/whatis.html – International Network for Urban Research and Action.
http://www.ltgdc.org.uk – London Thames Gateway.
http://www.puertomadero.com – Puerto Madero. Buenos Aires.
http://issuu.com/gsapponline/docs/urban-v9-is2 – URBAN. The Return of Megaprojects.

Urban Neoliberalism

Ugo Rossi and Alberto Vanolo, Università di Torino, Turin, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The rise of neoliberalism as the hegemonic art of governing contemporary capitalist cities traces its origins back to the 1980s, being associated with the ascent of conservative governments in the Unites States and the United Kingdom pursuing economic development strategies based on the conventional free market ideas. Since then, urban neoliberalism has spread across the globalizing world through an increasing emphasis being laid on the entrepreneurialization of local government, the privatization of public services, and the commodification of urban space. In this context, urban neoliberalism has taken the form of a highly mobile government technology, giving rise to a tremendous variety of politico-economic regimes across the globe through processes of hybridization and variegation. This article describes the variegated geographies of neoliberalism, ending with an analysis of the recent global recession which is understood as a crisis of (urban) neoliberalism.

Introduction: The Neoliberal Art of Governing Cities

Over the last two decades, the notion of neoliberalism has become key to contemporary debates and empirical inquiries within the social sciences concerned with the investigation of urban and regional development processes. Since the late 1970s onward, the economic policy that subsequently came to be known as neoliberalism supported the development of allegedly free (or 'anarchical': Jessop, 2002) market relations led by an ideology of competitive individualism, combined with policies aimed at socializing the costs of deficits in the private sector through generous capital injections to firms and banks during recessions. The success of neoliberal policies, being dogmatically imposed in the post-cold war era as the only viable option (a 'new religion': Peet, 2003) after the decline of the Keynesian mode of regulation, has been concomitant with the globalization of the international economy, impinging upon economic development processes at a variety of geographical scales, including the urban and regional levels.

Prior to the rise of neoliberalism, there is the end of an era of intense and apparently unlimited economic growth commonly recognized as the 'golden age of capitalism' (Marglin and Schor, 1991), beginning after the World War II and ending with the economic and geopolitical turbulences of the 1970s. In the United States, the 1980s witnessed the embrace of the so-called Reaganomics, the supply-side economic policy pursued by the Federal Administration during Reagan's presidency. According to the supply-side approach, in order to revitalize markets and attain higher rates of economic growth, it is necessary to reduce the fiscal pressure on businesses and capital holders so as to stimulate private savings and investments. The conservative and entrepreneurialist turn changed in depth the way in which economies and societies were governed in the United States as well as in the United Kingdom, which at the time was undergoing a parallel neoliberal '(counter)revolution' led by the Prime Minister Margaret Thatcher: from the privatization of state corporations to the controversial Poll Tax, which introduced a fixed tax per adult resident, leading to protests all over the country and in the city of London most intensely. While the controversy that developed around the so-called 'Poll Tax in Britain' presaged the end of Margaret Thatcher's premiership, as the 'iron lady' resigned few months after the approval of the tax, in the United States the adoption in 1978 of the Proposition 13 in the State of California – a constitutional amendment deliberating that the maximum amount of taxation on real property should not exceed the 1% of the property value – paved the way for Ronald Reagan's ascent from the governorship of California to the Federal Presidency. Indeed, the adoption of the amendment inspired an antifiscal campaign known as 'taxpayers' revolt' throughout the country, which prepared the ground for Ronald Reagan's successful presidential election in 1980.

The neoliberal (counter)revolution of the Reagan–Thatcher era had strong repercussions at the urban scale, opening the way for an unprecedented deregulation of housing markets and generally for an irreversible shift to the entrepreneurialization of local government and the privatization and commodification of urban public space (Smith, 1996; Mitchell, 1997; MacLeod, 2002). The translation of the neoliberal ideology into an instrumental and geographically mobile 'intellectual technology' relied on a globalizing system of policy exchanges and transfers, mobilizing a variety of institutional intermediaries and local importers of 'ideas from elsewhere' (Ong, 2007; Prince, 2012). In the Unites States, in order to build the reputation of neoliberalism as the emerging conventional wisdom within the wider public, politico-economic elites have drawn on the contribution of existing or newly established consultancy agencies funded by public and private donors (Peck, 2010). The neoliberal 'revolution' has thus changed the way in which at one and the same time the economic base and the politico-institutional realm of cities are conceived and concretely organized. As 'post-cold war' neoliberalism can be generally understood in a Foucauldian sense as the 'art of government' regulating and dictating the expansion of the economic form of capitalism into the social sphere (Lemke, 2001), through processes of individualization and entrepreneurialization of the self (Burchell, 1993; Lazzarato, 2009), urban neoliberalism more specifically is to be viewed as the translation of the logic of free market capitalism into the urban domain of sociospatial relations. As the next section of this article shows, in the late 1980s, David Harvey was the first social scientist to clearly

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74020-7

identify the rise of a neoliberal art of governing cities, through his conceptualization of the entrepreneuralization of urban governance in US cities over the 1980s, even though at that time he did not explicitly use the term neoliberalism. Harvey's understanding of urban governance in incipient neoliberal times serves as a point of departure here, preparing the ground for the following two sections of this article which are devoted to the analysis of the geographical expansion of urban neoliberalism over the last three decades. Finally, the concluding section focuses on the supposed crisis of urban neoliberalism as a consequence of the ongoing global recession, whose urban roots are customarily associated with the 'credit crunch' of 2007–08 that affected real estate markets in the US economy and elsewhere.

The Entrepreneurialization of Urban Governance

In a path-breaking article, David Harvey (1989) accurately illustrated the causes and the effects of the process of change in the sphere of local government, which he analyzed in terms of a shift from the public-managerial approach prevailing until the urban crisis of the 1970s to the entrepreneurialization of urban governance associated with the advent of a pro-growth mode of regulation. In his article, Harvey distinguished between the 'government' and the 'governance' of cities and regions, anticipating a conceptualization that became widely accepted within the field of urban studies and the wider social sciences in the following years. In his account, the former (government) identifies a hierarchical and 'managerial' governing style, founded on the primacy of the public sector and aimed at redistributing city revenues through the provision of services to firms and households. On the other hand, the latter (governance) highlights the decentralization of governing procedures in a context of public sector's reduced institutional capacity and of rising importance of public–private partnerships and related processes of negotiated decision making.

To urban and regional scholars, further expanding the conceptual framework elaborated by Harvey, entrepreneurial cities appeared to be crucial sites where politico-economic elites are in a position to experiment with innovative capital accumulation strategies, tending toward the immaterialization of the economic process (Hall and Hubbard, 1998). While Harvey verified these dynamics in light of the case study of Baltimore and its regenerating waterfront, related processes of socio-spatial restructuring were observed in a number of cities in the West and elsewhere. In Britain, the deindustrializing cities in the North of England (Newcastle upon Tyne, Leeds, Manchester, Liverpool) dealt with policies and plans of urban regeneration aimed at stimulating private businesses, especially small- and medium-sized enterprises, and at revitalizing inner-city areas which were affected by economic and demographic decline (Ward, 1997). Related dynamics of urban regeneration and changing economic governance were also noticed in other large cities and smaller towns in North America and Western Europe (Judd and Parkinson, 1990) and subsequently in Asia, as it is discussed below in this article (Jessop and Sum, 2000; Wu, 2002).

Since the 1990s, the discursive and material forces of globalization exerted a growing influence on the evolution of urban entrepreneurialism and the trajectories of economic development and socio-spatial restructuring. As Harvey had explained, politico-economic elites across the world were urged to connect cities and regional spaces to economic, political, and cultural relations of global reach. The growing interplay of global and local spatialities shapes in depth urban governance processes, which are driven by the so-called 'local politics of globalization' (Swyngedouw, 1992; Cochrane et al., 1996). The 'new urban politics' becomes thereby a crucial terrain of investigation with reference primarily to Anglo-American cities and metropolitan areas but also to cities in other contexts of advanced and emerging capitalism.

Variegated Geographies of Urban Neoliberalization

As Aiwha Ong (2007) has pointed out, the neoliberal regime of institutional governance and capitalist accumulation should be viewed as a mobile governmental technology, constantly traveling through a variety of geographical contexts and related sociopolitical cultures. For instance, the Chinese experience, with its unique combination of authoritarian state and unregulated market, reads as an example diverging from the Anglo-American ideal type of neoliberalism (He and Wu, 2009). By approaching neoliberalism as a political-economic practice, it will be evident that the neoliberal mode of regulation relies on existing sociocultural institutions, producing a wide range of hybrid formations (Brenner et al., 2010).

At the same time, the hybridization of political and institutional forms should not lead to believe that the transition to urban neoliberalism is linear and consistent with local values and sociocultural diversities. In the vast majority of cities in the South of the world, the effects of neoliberal policies have been pervasive, eradicating existing cultural, economic, and sociopolitical institutions. In the 1980s and the early 1990s, the Washington-based financial institutions (essentially the International Monetary Fund and the World Bank) recommended a set of economic and social policies (the so-called 'Structural Adjustment Programs'), which were approved and implemented without concerting decisions with the local citizens, with strong repercussions on the most deprived urban areas (Peet, 2003). Adjustment programs were forcefully adopted by those countries receiving financial aid and rescheduling debt payments (on the basis of the so-called 'conditionality principle'). These plans were deeply informed by neoliberalism as an ideology and policy rhetoric, under the banner of the so-called 'Washington Consensus Decalogue', which prescribed allegedly universal recipes for the recovery of national and urban–regional economies (Escobar, 1995; Power, 2003).

Macroeconomic and sector-specific policies, mainly elaborated at the national level, have had deep implications for cities in the South of the world. The joint effect induced by the privatization of important public services, the closure of a number of state-owned corporations, and the bankruptcy of manufacturing industries that previously benefited from public subsidies, has had serious consequences on local societies, leading to rising unemployment rates and generating novel forms of exclusion from the labor market (Riddel, 1997). These processes have also led to the expansion of the informal

underground economy. Job losses, the shrinkage of the state sector, and the fall in living standards have forced the city dwellers to rely on a variety of sources of experimentation and commonality; shared relational capital, unused land converted into community gardens, informal trading and street vending are customary responses to poverty and deprivation in the cities of the Global South (Simone, 2004).

In this context, existing welfare states have been restructured and widely dismantled: educational and health care services have been drastically reduced, the already scarce supply of public housing has been halved on average, the price of imported food has considerably increased as a consequence of the devaluation of local currencies, while large portions of fertile land have been reconverted to export-oriented productions (Harvey, 2005). These processes of socioeconomic restructuring have led to the expansion not just of the informal economy but also of the illegal, criminal economies: newly formed street gangs and urban mafias have proliferated, engaging in large-scale organized theft and drug trade through increasingly transnational networks (Wacquant, 2008). In a context of increased poverty and hardship, large cities in the South indeed tend to become less safe places to live in. In the face of perceived or actual threats, the so-called 'transnational capitalists' (Sklair, 2002), that is those involved in the most profitable sectors of the economy, such as the oil industry for instance, respond to this situation by tightening security systems and militarizing their residential spaces in a context of neoliberalization of violence where exceptional violence is transformed into exemplary violence regulating the relationship with 'the other' (Springer, 2012). In general, social disadvantage has intensified in large cities and metropolitan areas, particularly affecting migrants and the ethnic minorities. The ambivalent effects of neoliberal urbanism are highly visible in the most dynamic city regions (such as the industrial coastal cities of South China, the fast-developing cities in the Middle East such as Dubai, the most famous as well as established global cities in the West like London and Los Angeles), where migrants are employed on casual bases in the industrial and the service sectors under conditions of unbearable workloads, low wages, uncertain contractual status, and precarious housing (Buckley, 2012).

The transnational governance of development processes in the South have also evolved over the time. The normative and straightforward approach to economic restructuring characterizing the Structural Adjustment Programs in the 1980s has become softer in many respects. The basic logic underlying development programs, however, appears to be immutable. For instance, despite the asserted good intentions, the financial aids provided in order to attain the Millennium Development Goals (MDGs) identified by the United Nations (from poverty alleviation to the reduction of infantile mortality rates, from the pursuit of an environmentalist agenda to the promotion of interstate and supranational agreements for international development) are persistently subject to IMF conditionality, which dictates the transition toward modes of social and economic organization and development based on the neoliberal agenda (Amin, 2006). At the same time, recent development programs like the MDGs are sustained by increasingly more sophisticated governmental apparatuses mobilizing calculative practices of information profiling and knowledge exchange, aimed at enhancing the neoliberal values of responsibilization and entrepreneurialization of the poor in the Global South (Ilcan and Phillips, 2010).

Whereas economic development policies adopted in poorer and emerging countries in the South have gradually abandoned the shocking therapies of first-generation neoliberalism, the 'original' forms of neoliberalism (those elaborated and commonly practised in the West) have changed over the years, and they will certainly change in the future, without losing their distinguishing features, relating to the emphasis laid on goals of economic growth to the detriment of those of social justice and wealth redistribution. After the rise and the triumph of ultra-liberal approaches during the Reaganian–Thatcherian decade, the 1990s and the first half of the 2000s have seen the reformulation and the reframing on novel discursive bases of neoliberalism as a politico-economic discourse and as a policy practice. In the attempt to conceptualize the recent evolution of neoliberalism, geographers Jamie Peck and Adam Tickell (2002) have provided an influential interpretation of the shift from an unconditionally deregulatory policy rationale, defined as 'roll-back neoliberalism,' which characterized the Reagan–Thatcher era, aiming to extinguish the social-democratic legacies created by the Keynesian regulation of capitalism, to a renewed neoliberal approach, more positive and 'constructive,' aimed at rebuilding the capitalist state on entrepreneurial bases and at reshaping in innovative ways the relationships between business interests and the public sector. The latter approach, defined as 'roll-out' stage of neoliberalism in Peck and Tickell's terminology, has been advocated during the 1990s by the New Democrats and the New Labor Party led by Bill Clinton and Tony Blair in the United States and Great Britain, respectively.

Neo-progressive parties took the lead, therefore, in a second stage of the neoliberal 'revolution,' replacing the conservative parties dominating the first stage, in the United States as well as in their most influential allied countries, notably those belonging to the Atlantic Alliance that has remained unchallenged after the collapse of the Soviet Union. Having acquired this role, the progressive parties have engaged in a vast endeavor of self-legitimization in the eyes of business actors, by precipitously converting to the basic tenets of the neoliberal orthodoxy: the obsession for budget rigor leading in Europe to the adoption of a restrictive fiscal and monetary policy as the distinguishing trait of the newly founded European Union in 1992, the reduction of public spending in social assistance and, finally, most importantly in the context of post-Fordist capitalism, the flexibilization of the workforce. Unlike the conservatives, refusing the idea itself of society as a collectively organized entity and embracing the purest free market ideology, the center-left governments have sought to reconcile objectives which were considered mutually antithetic until a recent past. Combining the enhancement of social cohesion with the stimulation of economic competitiveness and turning the public sector into a more entrepreneurial and accountable organization have been distinguishing features of the so-called 'new liberal formulation' (Harloe, 2001). The 'Third Way' publicly announced by the British Prime Minister Tony Blair and theorized by sociologist Anthony Giddens (1998) came to represent the most consistent example of neoliberal formulation being proposed on the progressive side of the political spectrum as an

alternative to the ultraliberalism advocated by the conservatives. The emphasis placed on 'global governance' and transnational civil society, in contrast to the neoliberal aversion to 'big government' and social justice, alongside the attention paid to issues of environmental sustainability and cultural pluralism, in lieu of the conservative predilection for the integrity of the family and the nation, have been the essential principles of Giddens–Blair's Third Way (Peck and Theodore, 2001).

Within the sphere of urban government, apparently indissoluble beliefs and taboos received from the past shattered in the face of local and nationwide campaigns against petty and organized crime and the variety of phenomena of social deviance, including homelessness, drug abuse, and even 'art crimes' such as graffiti and writing. While urban administrations led by conservative mayors gave rise to highly mediatized repression campaigns against urban crime, such as most famously the 'zero tolerance' initiative adopted in the 1990s by the Mayor of New York Rudolph Giuliani (Smith, 1998), the center-left governments sought to restore their credibility in security policies; a field in which progressive parties were still conventionally associated with tolerant attitudes. In this vein, Prime Minister Tony Blair announced, in the political Manifesto of the New Labor in 1997, a new course of action against crime under the banner of the slogan: "tough on crime and tough on the causes of crime," which explicitly attempted at reconciling policies addressing issues of social exclusion and urban deprivation with measures to combat crime and deviance.

Broadly speaking, the second stage of neoliberalism in the 1990s testifies to an important redefinition of the ways in which capitalist cities are governed and administered. In the unified European space, these changes have taken form through institutional dynamics of multilevel governance, in which cities and regions play a crucial role, after the hollowing out of the state and of the institutions gravitating within its sphere (Jessop, 2004). In the public service sector, on the one hand, local and regional governments assume key roles in the governance process; on the other hand, however, subnational administrations externalize their prerogatives to other subjects of private or not-for-profit status, in line with the general trend of contractualization of social services, by which services are delivered on the basis of a contract signed between the providers (either within or outside the formal public sector) and welfare recipients. Job centers subcontracted to private companies or quasi-public agencies replace the old state-led employment bureaus, while similar processes of outsourcing occur in the health care services and other social assistance sectors organized at the local level with a view to achieve a modernized local welfare state based on dynamic local–central and public–private relationships (Cochrane, 2004). In times, in which economic and financial flows appear to be increasingly out of juridical and political control and in which the strategies of transnational firms and other 'foreign' investors are caught in a hardly decipherable chaos, large cities, and metropolitan areas – in their capacity to foster internationally competitive economies of agglomeration and diversity – are assumed to be crucial institutional arenas in the regulation (i.e., the deregulation and reregulation) of capitalist economies and societies (Brenner and Theodore, 2002). At the same time, owing to the reduced government spending and the fierce competition in obtaining public and private funds for economic regeneration policies and particularly for costly infrastructural projects, urban governments are compelled to devise a wide range of local economic development strategies: place marketing and promotion campaigns, mega-events, company subsidies, and the creation of free economic zones.

Neoliberalizing Urban Economies

In the described context of hegemonic and at the same time spatiotemporally variegated neoliberalism, two local economic regeneration schemes are particularly illustrative of the dominant approach to urban development: the 'Urban Enterprise Zones' policy, which was first carried out at the time of the Thatcher premiership and the Business Improvement Districts initiatives. The former, based on fiscal incentives and ambitious infrastructure projects, is designed to trigger economic development dynamics in declining and peripheral urban areas. In neocapitalist China, the creation of 'economic special zones' in the 1980s made a decisive contribution to the booming of its export-led economy, particularly in the southern coastal regions which have become the drivers of China's path of economic development: the industrial cities of Shenzhen, Dongguan, and Canton in the rampant Pearl River Delta region. The Business Improvement Districts policy, on the other hand, which has been embraced by local governments in a number of countries including Britain, Germany, New Zealand, and South Africa since the 1970s onward, focuses on the physical renewal and the aesthetic amelioration of attractive and potentially profitable urban districts (usually shopping and entertainment precincts located in downtown areas), using funds obtained from the self-taxation of local private firms, whose financial efforts are rewarded with the assignment of autonomous responsibilities in the management of the regenerated spaces (Ward, 2006; Cook, 2008).

Contemporary processes of urban and regional governance highlight the rise of relatively autonomous scales of economic and spatial regulation at the local level, but also the persistent influence of the national state in this realm, particularly in the East Asian context, but in different ways also in Western liberal democracies. In the unified European space, the last two decades have indeed seen the national state repositioning itself as a key actor in the coordination and the material structuring of the complex multilevel governance structure regulating the processes of urban and regional development (Brenner, 2004). This institutional context demands attention being paid to the ways in which neoliberal imperatives dictated by international and national politico-economic elites become incorporated into economic development strategies at the urban and regional levels, particularly in the form of public–private partnerships aimed at revitalizing local economies through the creation of technological districts, the abatement of local taxes, the demolition of obsolete housing, the redesign of residential spaces, the megaprojects of downtown and waterfront regeneration, the construction of iconic buildings, the invention of luxury and mass-consumption spaces, and the adoption of surveillance devices guaranteeing the safety of inhabitants and city users (Asian, European, and North American case studies are available in Olds, 1995; Moulaert et al., 2005; Orueta and Fainstein, 2008).

Outside the Western world, powerful think tanks and supranational institutions have made a decisive contribution to the diffusion of the approach to urban policy that has been described above. At the global scale, a crucial role has been played by the World Bank and the Organization for Economic Cooperation and Development (OECD). The World Bank's "corporate goals of reducing poverty, promoting market-based growth, building durable institutions, and protecting the environment" (World Bank, 2000, p. 40; *see also* World Bank, 1991) stem from a conventionally neoliberal understanding of economic development issues. From the standpoint of the Washington-based organization, cities are not to be considered as containers of social problems, let alone places where socio-spatial inequalities have to be addressed, but as 'engines' of economic growth, more precisely as centerpieces for processes of economic-relational networking reconnecting local economies to the global flows of capitals, commodities, information, and policy discourses (Theodore and Peck, 2012). In the emerging economies of the Global South, as a result, local elites are pressed with demands from international organizations for turning the major urban agglomerations into globalizing cities competing with their Western counterparts. Even in cities located in what was previously labeled as 'Third World,' the 'urban question' is no longer concentrated on housing issues and the allocation of adequate services for the local population, but on the optimization of local economic productivity. The currently common wisdom on development issues holds that even antipoverty initiatives – which used to have priority in the urban policy agenda until a recent past – have to be combined with strategies of economic growth.

The link between antipoverty and economic growth policies is justified not only along discursive lines (for instance, as a way to avoid parasitic social assistance), but also by reference to structural changes within urban development dynamics. In the cities and metropolitan areas in the Global South, economic growth processes and the regeneration of large urban areas have been accompanied by the proliferation of informal settlements, variously defined slums, favelas, bidonvilles, shantytowns, depending on the geographical–linguistic context. This 'planet of slums' (Davis, 2006) has taken shape also as an indirect consequence of selective and elitist strategies of urban development, concentrated on a limited set of 'competitive' urban spaces. In the discursive politics of neoliberalism, even slums and favelas are portrayed as 'opportunities' for economic development, waiting to exploit the potential of their entrepreneurial talents (Mukhija, 2003).

Reproducing Neoliberalism

A result of the dynamics that have been shortly described and analyzed in the previous section has been the global expansion and circulation of neoliberalism, which has adapted to the varying (geographically and historically) configurations of post-Fordist urban capitalism. Many authors have approached the urbanization of neoliberalism through the lens of the 'new urban politics' originally identified by Kevin Cox and other critical urban geographers (Cox, 1993; Cochrane, 2007). The characteristic, and the ambiguity, of urban neoliberalism rests on its ambivalent relationship with the public sector and the state in particular. On the one hand, in line with the tenets of classic liberalism and its more recent formulations in the last quarter of the twentieth century, supporters of urban neoliberalism postulate the retreat of the state from the socioeconomic realm, advocating an involvement of private actors within the institutions of urban and regional governance; on the other hand, supporters of neoliberal policies do not hesitate to exert pressure on national governments when it comes to obtain financial support for megaprojects and international events capable of fostering the so-called 'urban boosterism' (Harvey, 1989).

Apart from internal contradictions as well as the relentless attempts to mediate between conflicting interests, the strength of urban neoliberalism rests on its capacity to incorporate all sorts of urban problems and issues (social, economic, and environmental) into the seductive domain of spatial competitiveness and economic growth, proposing policies flexibly adapting to a variety of geographical contexts. What can be defined the 'politics of the mime' – namely, the alternatively passive or creative imitation of successful models of urban growth – has expanded across the globe, albeit this process has occurred in varying forms and with contrasting outcomes (Rossi and Vanolo, 2012). As soon as the depressive effects of the geopolitical and monetary crises of the 1970s vanished and the Japanese challenge to the Western hegemony on technology and business management unraveled, 'ideas from America' became again predominant. As a result, since the 1980s strategies of urban economic development and restructuring in the United States have acquired a renewed leading role, offering success stories of innovation and entrepreneurship in a number of advanced sectors such as finance, business administration, information, and communication technologies. In an increasingly competitive and interconnected environment such as that of the globalizing capitalist economy, the rapidity and intensity of economic and spatial changes being observed in the North American context have forced those governing cities and regions in other contexts to accelerate the implementation of regeneration policies and broader projects of urban development. As Peck and Tickell (2002) have underlined, the length of time covering the manifestation of an urban issue, the selection of objectives and priorities, the mobilization of institutions and financial sources, the organization of projects and urban planning initiatives and, finally, their implementation has increasingly shortened, inducing local elites to rely on ready-made policy recipes ('off-the-shelf solutions' as the two authors put it), more easily implementable and less likely to be contested by the local population in comparison to strategies formulated at the local level.

For the reasons explained above, neoliberalism has succeeded in reproducing and 'territorializing' itself in apparently efficient and even creative ways, rapidly migrating from one place to another across the globe. In doing so, it has drawn on the contribution of a variety of powerful agents and intermediaries: technocratic elites, think tanks, global consultants, political networks, and development gurus (such as Richard Florida and Michael Porter, the theorists of the urban 'creative class' and the 'competitive advantages' of the inner city, respectively). The planetary circulation of presumably winning ideas and policy solutions is conventionally labeled as 'policy transfer' within the mainstream political sciences, while it is alternatively understood in terms of 'policy mobility' by critical

urbanists pointing to the effect of reinvention associated with the process of circulation (McCann, 2010; Peck, 2011). The phenomenon of policy mobility and transfer is related not only to strategies of growth-led development and governance (such as in the cases of the Special Economic Zones and the Business Improvement Districts policies), but also to issues and problems firmly anchored in local and institutional specificities, such as poverty, deviance, social marginality, and urban criminality (Wacquant, 1999). Using a metaphor drawn from evolutionary sociobiology, it can be observed what happens when an intrinsically complex system (such as the city or the world economy) reduces its own 'political biodiversity' (i.e., the variety of conceptual and operational tools normally used in order to identify policy issues and formulate appropriate strategies to tackle them), becoming more vulnerable and exposed to the risks associated with the reduction of viable options for action (the selection mechanisms in the evolutionary terminology). An illuminating example of the dangers associated with the reduction of complexity generated by the global expansion of neoliberal urbanism is that of the financial crisis that hit the world economy in 2008 and 2009. Owing to the role played by the housing finance sector in triggering an economic slowdown in the United States that has rapidly traveled toward other advanced capitalist economies, today's cities appear to be crucial sites in which the limits and the contradictions of neoliberal capitalism and globalization can be observed and critically scrutinized.

The Global Recession: A Crisis of (Urban) Neoliberalism?

The intimate relationship linking capitalism, the housing sector, and processes of urban development has come to the fore in an extraordinarily visible fashion over the last three decades or so. First, the so-called 'financialization' of the global economy (the steady increase in financial transactions at a world scale and their autonomization from the 'real' economy) has played a key role in the ascent of established or emerging global cities, such as London and New York in the West, and Shanghai, Hong Kong, Singapore, Mumbai and Dubai in Asia; in addition, before the crisis, the housing boom had been the engine of economic growth after the end of the short period of contraction that followed the bursting of the New Economy bubble (the economy based on 'dotcoms' firms which expanded between 1995 and 1999). The noticeable but also virtual profits linked to the financialization of the economy, alongside the vertiginous rise of property values, have generated – during the relatively short golden age of post-cold war neoliberalism, the two decades between the mid-1980s and the first half of the 2000s – widespread illusions and expectations of unlimited wealth accumulation until the 2008–09 financial crisis.

The financialization of the economy, which has been brought about by neoliberal globalization and the dismissal of Keynesian economic policies, has sparked a trajectory that has eventually led to the economic slump of 2008–09 and the subsequent recession. The process of financialization has produced an unprecedented availability of capitals, which have been partly reinvested in the mortgage market, resulting in the growing expansion of the real estate sector. In the early 2000s, the United States, and particularly the southern and western states (the so-called 'Sun Belt' which had been at the forefront of the post-Fordist transition), such as California, Texas, Florida, witnessed a steady rise in housing prices, as a consequence of the expansion of homeownership due to the illusory increase in mortgage opportunities. Related phenomena of rising demand in the housing market have been observed in those European countries that have laid out permissive regulatory frameworks on financial services and consumer credit, such as Spain, Ireland, United Kingdom, and Iceland. These countries, along with the American states that were in the forefront of the housing bubble, have been those that most have suffered the consequences of the 'credit crunch' that hit the world economy in 2007–2008. A variety of social groups have been affected by the economic downturn, especially those that had the illusion of attaining a higher standard of life thanks to easier access to homeownership and to a number of goods (not always strictly necessary) purchased through consumer credit.

The push for homeownership has been actively encouraged by the deregulation of mortgage markets in the United States and in many European countries as well. Since the early 1990s, the acceptance policies of banks have become increasingly lenient, and credit limits (i.e., the maximum amount that can be borrowed through a mortgage) have expanded (Aalbers, 2008). The illusion of a relentlessly expanding homeownership is, therefore, at the origin of the subprime mortgage crisis in the United States. Subprime mortgages are riskier loans, in many cases, refinances, offered at higher (predatory) interest rates to economically vulnerable and potentially insolvent clients: low-income individuals belonging to ethnic minorities and other disadvantaged groups, frequently employed on a casual basis and with a background of credit default (Wyly et al., 2007). High-risk 'predatory' loans were introduced in the 1990s, preparing the ground for the housing bubble of the subsequent decade. In 2006, the apparently unlimited expansion of the mortgage market started slowing down, giving rise to a set of interrelated negative phenomena, eventually leading to a downward spiral: growing insolvencies by mortgage borrowers, foreclosure proceedings, increase in housing abandonment at the neighborhood level, crisis, and in some cases bankruptcy of credit institutions, fall of housing prices and depreciation of property values. In September 2008, the bankruptcy of Lehman Brothers, one of the largest and oldest financial firms in the United States, triggered by the subprime mortgage crisis, created a banking panic which was followed by the global economic recession.

The economies of cities and metropolitan regions that in previous years had taken the lead in the process of globalization and neoliberalization have been seriously affected by the credit crunch, the crisis of the housing market and of the wider economy: from established global cities (New York, Los Angeles, London) to globalizing cities (Dubai, Shanghai, Singapore) where spectacular megaprojects were underway when the crisis began, to the new centers of urban Europe such as Dublin, Barcelona, Manchester, and even the quiet Reykjavik have faced a deep depression of the economy, with the closure of several firms and commercial activities, the generation of new reserves of unemployed and working poor, while a number of urban

megaprojects have been suspended. In the United States, which has been the epicenter of the global economic shock – after having been at the core of the organizational and technological 'revolution' of post-Fordism and of the process of financialization – it is not only the private sector that has suffered the consequences of the downturn. With increasingly passive local budgets owing to the risky investment operations undertaken in the years of the financial boom and as a result of austerity measures being adopted at the national level, the administrators of several cities in the United States and across the advanced capitalist world have been forced to drastically reduce the supply of important services (transportation, health care, cultural events) so as to manage to repay the debts and to meet fiscal consolidation targets.

Whereas orthodox economists are blamed for not having predicted the economic and financial crisis, but also for having actively contributed to foster the belief of self-regulating markets, the expansion and the subsequent cataclysm of urban capitalism in the neoliberal era were prophesied by David Harvey and those sociologists and geographers who, drawing on his pioneering studies, have analyzed in detail and at the same time have denounced the destructive link between finance, capitalism, and urban development (Aalbers, 2012). More generally, the present global recession is commonly perceived as the 'crisis of neoliberalism,' namely as the outcome of the contradictions inherent in the neoliberal strategy of capitalist development and related macroeconomic policies (see, for instance, Duménil and Lévy, 2011; Overbeek and van Apeldoorn, 2012), even though critical commentators lament the fact that powerful economic actors have managed to retain and even increase their influence and wealth thanks to state bailouts and the unchallenged dominance of free market ideas among mainstream policy makers and politicians (Crouch, 2011). The coming years will show, therefore, whether this structural crisis will stimulate a shift in the dominant paradigm of economic-political regulation or it will just reassert the power of neoliberalism on novel material and discursive bases.

See also: Cities and the Financial Crisis; Cities: Capital, Global, and World; City Marketing; Globalization: Geographical Aspects; Housing and Cities; Neoliberalism; Real Estate Development; Urban Entrepreneurialism; Urban Geography; Urban Informality: The Production and Regulation of Space; Urban Megaprojects; Urban Planning in China; Urban Policies in Asia; Urban Policy in Europe; Urban Policy in North America; Urban Violence and Militarization.

Bibliography

Aalbers, M.B., 2008. The financialization of home and the mortgage market crisis. Competition and Change 12 (2), 148–166.
Aalbers, M.B. (Ed.), 2012. Subprime Cities. The Political Economy of Mortgage Markets. John Wiley and Sons, Chichester (UK).
Amin, S., 2006. The millennium development goals. A critique from the south. Monthly Review 57 (10), 1–15.
Brenner, N., 2004. New State Spaces: Urban Governance and the Rescaling of Statehood. Blackwell, Oxford.
Brenner, N., Peck, J., Theodore, N., 2010. Variegated neoliberalization: geographies, modalities, pathways. Global Networks 10 (2), 182–222.
Brenner, N., Theodore, N., 2002. Cities and the geographies of 'actually existing neoliberalism'. Antipode 34 (3), 341–347.
Buckley, M., 2012. From Kerala to Dubai and back again: construction migrants and the global economic crisis. Geoforum 43 (2), 250–259.
Burchell, G., 1993. Liberal government and techniques of the self. Economy and Society 22 (3), 267–282.
Cochrane, A., 2004. Modernisation, managerialism and the culture wars: the reshaping of the local welfare state in England. Local Government Studies 30 (4), 481–496.
Cochrane, A., 2007. Understanding Urban Policy. A Critical Approach. Blackwell, Oxford.
Cochrane, A., Peck, J., Tickell, A., 1996. Manchester plays games: exploring the local politics of globalization. Urban Studies 33 (8), 1319–1336.
Cook, I.R., 2008. Mobilising urban policies: the policy transfer of US business improvement districts to England and Wales. Urban Studies 45 (4), 773–795.
Cox, K.R., 1993. The local and the global in the new urban politics: a critical view. Environment and Planning D: Society and Space 11 (4), 433–448.
Crouch, C., 2011. The Strange Non-Death of Neoliberalism. Polity Press, Cambridge, UK.
Davis, M., 2006. Planet of Slums. Verso, London.
Duménil, G., Lévy, D., 2011. The Crisis of Neoliberalism. Harvard University Press, Cambridge, MA.
Escobar, A., 1995. Encountering Development. The Making and Unmaking of the Third World. Princeton University Press, Princeton, NJ.
Giddens, A., 1998. Third Way. The Renewal of Social Democracy. Polity Press, Cambridge.
Hall, T., Hubbard, P. (Eds.), 1998. The Entrepreneurial City. Geographies of Politics, Regime and Representation. John Wiley, Chichester, UK.
Harloe, M., 2001. Social justice and the city: the new 'liberal formulation'. International Journal of Urban and Regional Research 25 (4), 889–897.
Harvey, D., 1989. From managerialism to entrepreneurialism: the transformation in urban governance in late capitalism. Geografiska Annaler B: Human Geography 71 (1), 3–17.
Harvey, D., 2005. A Brief History of Neoliberalism. Oxford University Press, New York.
He, S., Wu, F., 2009. China's emerging neo-liberal urbanism: perspectives from urban redevelopment. Antipode 41 (2), 282–304.
Ilcan, S., Phillips, L., 2010. Developmentalities and calculative practices: the Millennium development goals. Antipode 42 (4), 844–874.
Jessop, B., 2002. Liberalism, neoliberalism, and urban governance: a state-theoretical perspective. Antipode 34 (3), 452–472.
Jessop, B., 2004. Hollowing out the nation-state and multilevel governance. In: Kennett, P. (Ed.), A Handbook of Comparative Social Policy. Edward Elgar, Cheltenham, pp. 11–25.
Jessop, B, Sum, N.-L., 2000. An entrepreneurial city in action: Hong Kong's emerging strategies in and for (inter)urban competition. Urban Studies 37 (12), 2287–2313.
Judd, D.R., Parkinson, M. (Eds.), 1990. Leadership and Urban Regeneration. Cities in North America and Europe. Sage, Newbury Park, CA.
Lazzarato, M., 2009. Neoliberalism in action: inequality, insecurity and the reconstitution of the social. Theory, Culture and Society 26 (6), 109–133.
Lemke, T., 2001. 'The birth of bio-politics: Michel Foucault's lecture at the Collège de France on neoliberal governmentality. Economy and Society 30 (2), 190–207.
MacLeod, G., 2002. From urban entrepreneurialism to a 'revanchist city' on the spatial injustices of Glasgow's renaissance. Antipode 34 (3), 602–624.
McCann, E., 2010. Urban policy mobilities and global circuits of knowledge: toward a research agenda. Annals of the Association of American Geographers 101 (1), 107–130.
Marglin, S., Schor, J.B. (Eds.), 1991. The Golden Age of Capitalism. Reinterpreting the Postwar Experience. Oxford University Press, New York.
Mitchell, D., 1997. The annihilation of space by law: the roots and implications of anti-homeless laws in the United States. Antipode 29 (3), 65–89.
Moulaert, F., Rodriguez, A., Swyngedouw, E. (Eds.), 2005. The Globalized City: Economic Restructuring and Social Polarization in European Cities. Oxford University Press, Oxford.
Mukhija, V., 2003. Squatters as Developers? Slum Demolition and Redevelopment in Mumbai, India. Ashgate, Aldershot.
Olds, K., 1995. Globalization and the production of new urban spaces: Pacific Rim megaprojects in the late 20th century. Environment and Planning A 27 (11), 1713–1743.
Ong, A., 2007. Neoliberalism as a mobile technology. Transactions of the Institute of British Geographers 32 (1), 3–8.
Orueta, F.D., Fainstein, S., 2008. The new mega-projects: genesis and impacts. International Journal of Urban and Regional Research 32 (4), 759–767.

Overbeek, H., van Apeldoorn, B. (Eds.), 2012. Neoliberalism in Crisis. Palgrave Macmillan, Basingstoke, UK.
Peck, J., 2010. Constructions of Neoliberal Reason. Oxford University Press, Oxford.
Peck, J., 2011. Geographies of policy: from transfer-diffusion to mobility-mutation. Progress in Human Geography 35 (6), 773–797.
Peck, J., Theodore, N., 2001. Exporting workfare/importing welfare-to-work: exploring the politics of Third Way policy transfer. Political Geography 20 (4), 427–460.
Peck, J., Tickell, A., 2002. Neoliberalizing space. Antipode 34 (3), 380–404.
Peet, R., 2003. Unholy Trinity: The IMF, World Bank and WTO. Zed Books, London.
Power, M., 2003. Rethinking Development Geographies. Routledge, London.
Prince, R., 2012. Policy transfer, consultants and the geographies of governance. Progress in Human Geography 36 (2), 188–203.
Riddell, B., 1997. Structural adjustment programmes and the city in tropical Africa. Urban Studies 34 (8), 1297–1307.
Rossi, U., Vanolo, A., 2012. Urban Political Geographies. A Global Perspective. Sage, London.
Simone, A., 2004. For the City Yet to Come. Changing African Life in Four Cities. Duke University Press, Durham, DC.
Sklair, L., 2002. Globalization: Capitalism and Its Alternatives. Oxford University Press, Oxford.
Smith, N., 1996. The New Urban Frontier. Gentrification and the Revanchist City. Routledge, New York.
Smith, N., 1998. Giuliani time. The revanchist 1990s. Social Text 16 (4), 1–20.
Springer, S., 2012. Neoliberalising violence: of the exceptional and the exemplary in coalescing moments. Area 44 (2), 136–143.
Swyngedouw, E., 1992. The Mammon quest. 'Globalization', interspatial competition and the monetary order: the construction of new scales. In: Dunford, M., Kafkalas, J. (Eds.), Cities and Regions in the New Europe. Belhaven Press, London, pp. 255–274.
Theodore, N., Peck, J., 2012. Framing urban neoliberalism: translating 'common-sense' urban policy across the OECD. European Urban and Regional Studies 19 (1), 20–41.
Wacquant, L., 1999. Les prisons de la misère. Raisons d'agir, Paris.
Wacquant, L., 2008. The militarization of urban marginality: lessons from the Brazilian metropolis. International Political Sociology 2 (1), 56–74.
Ward, K., 1997. Coalitions in urban regeneration: a regime approach. Environment and Planning A 29 (8), 1493–1506.
Ward, K., 2006. 'Policies in motion, urban management and state restructuring: the trans-local expansion of business improvement districts. International Journal of Urban and Regional Research 30 (1), 54–75.
World Bank, 2000. Cities in Transition: World Bank Urban and Local Government Strategy. World Bank, Washington, DC.
Wu, F., 2002. China's changing urban governance in the transition towards a more market-oriented economy. Urban Studies 39 (7), 1071–1093.
Wyly, E.K., Atia, M., Lee, E., Mendez, P., 2007. Race, gender, and statistical representation: predatory mortgage lending and the US community reinvestment movement. Environment and Planning A 39 (9), 2139–2166.

Urban Places: Statistical Definitions

Richard L Forstall, Alexandria, VA, USA
Kam Wing Chan, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by S. Moses, volume 24, pp. 16035–16040, © 2001, Elsevier Ltd.

Abstract

Virtually all countries report population and other data as urban or rural, but specific definitions vary greatly as to the geographic units used and how populous an urban place must be. Many countries recognize urban as including overflow and suburbs outside municipal limits, giving rise to additional definitions such as urbanized areas and metropolitan areas. This article discusses the US definition of urban and its evolution, summarizes the US metropolitan definition, and compares the urban definitions of the 40 countries with the largest urban populations, with special attention to China.

Introduction

Studies of cities and urbanization rely on the availability of relevant statistics, especially censuses and other data from national, state, and municipal governments. Most countries publish a range of statistics for their cities, and many do so for aggregates identified as urban and rural. However, what is termed a city varies greatly across countries and sometimes even within countries, and there is little international consistency in what is defined as urban. The United Nations Population Division publishes estimates and projections of urban population for the world and individual countries, but in doing so it accepts each country's own particular urban definition and has not made any recent effort to establish an international standard.

This wide variation in official census definitions of urban is not due primarily to differences in the concept of what constitutes an urban locality. In nearly all countries, the adopted definition of urban seeks to identify a high-density concentration of nonagricultural population, typically with some kind of municipal government. However, the minimum population size required for recognition varies widely by country.

Countries also differ in how they delimit the individual urban place, which may be an administrative unit, such as a city or commune, or may be defined with a nonadministrative boundary. Each alternative has some advantages; choice of an administrative unit generally will maximize the amount of statistical data available for the place, from censuses, other government-produced statistics, and private sources. On the other hand, places defined with nonadministrative boundaries may more precisely represent the place as a dense concentration of nonagricultural population, but be less convenient as to data availability and public recognition (Gibbs, 1961: Part I; Goldstein and Sly, 1974; Arriaga, 1975).

Besides recognizing urban places as a category, many countries have defined larger entities such as urban areas, urbanized areas (UAs), and metropolitan areas (MAs). These concepts developed early in the twentieth century, as the population and geographical extent of urban development overflowed the boundaries of most large and many small cities, leaving the city administrative unit underbounded and less representative of the actual urban concentration. But just as with urban places themselves, definitions of these larger urban entities have varied greatly across countries.

The following sections of this article discuss the definition of urban used in the United States, the US metropolitan definition, and the urban and metropolitan definitions currently used by some larger countries. The concluding section, by the second author, presents the urban definition used in China.

United States: Definitions of Urban

Official concern with determining urban and rural populations in the United States can be traced back to the 1850 decennial census. A statistical atlas of 1874 and the 1880 census recognized municipalities of at least 8000 population as urban and also presented data for an alternative threshold of 4000. A supplementary analysis of the 1900 census was the first that officially designated places of 2500 or more as urban (Walker, 1874; Truesdell, 1949). That size threshold for urban status has been retained in every subsequent US decennial census, although no extended justification for it has been offered.

However, besides the changes in threshold there have been significant changes in the choice of geographic unit to which the population threshold is applied. Moreover, the US experience illustrates how the nature of a country's administrative divisions can affect its definition of urban territory.

US Incorporated Municipalities

Under the federal structure of the United States, major differences in the characteristics of municipalities appear across states, because each state creates and oversees its own administrative subdivisions. All states have counties or county equivalents, and subcounty units of various degrees of local self-government, collectively referred to as minor civil divisions (MCDs). Incorporated municipalities exist in all but one state (Hawaii); they are variously termed cities, towns, boroughs, or villages. In most states, they are associated with locally recognized nonagricultural population concentrations and contain relatively little open-country population.

These 'incorporated places' (the Census Bureau term) comprise only a small share of the total US area, while the

remainder is sometimes referred to as 'unincorporated.' In 2010, 19 533 incorporated places contained 192.0 million people or 62.2% of the US population, but only 122 236 m^2 or just 3.5% of the land area. Many incorporated places are very small, with roughly half having fewer than 1000 residents (U.S. Census Bureau, 2012: Tables 8, 9, and 38).

The six New England States (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont) diverge from this pattern. Here, the basic subdivisions are incorporated municipalities termed cities or towns, which collectively comprise all or most of the inhabited area of the state. Unlike the typical incorporated place in the other states, in New England most towns and some cities include considerable open-country territory along with at least one high-density concentration of population.

This New England pattern in which local municipal-type governments include rural as well as urban territory somewhat resembles the commune-type pattern found widely in Europe, although the New England units average much larger in extent than typical communes. It dates from the first period of English settlement in the early 1600s. While some other Northern states have subcounty municipal-type governments, commonly termed townships, they also have many incorporated municipalities like those of the remaining states – in some cases within and subordinated to the townships. Among these states are New York and Wisconsin, where the entities commonly called townships are officially termed towns.

Since 1900, the US Census Bureau has regarded as 'incorporated places' only municipalities incorporated as cities, boroughs, or villages, and towns except in New England, New York, and Wisconsin. While not questioning that the towns in the latter states are incorporated municipalities, the census does not regard them as places, because they typically include much open country as well as population concentrations.

Township governments are also especially strong in New Jersey, Pennsylvania, Michigan, and Minnesota. In all 12 strong-township states, recent censuses have provided much the same range of data for towns and townships as for incorporated places of comparable size. Most of the other Midwestern states also have subdivisions termed townships, as do Arkansas and North Carolina, but their municipal activities are few or nonexistent today. The Census Bureau treats them like the MCDs of other states (U.S. Bureau of the Census, 1994: Chapter 8; U.S. Census Bureau, 2012: Table 39).

New England Towns: Urban or Rural?

The existence of these town-type municipalities in New England and other states not surprisingly has created problems of comparability for an urban category defined in terms of municipalities of a specified size. Under the varying rules of the censuses prior to 1930, all or most New England towns that met the population threshold were included as wholly urban.

In 1930, an adjustment was made under which only towns in Massachusetts, New Hampshire, and Rhode Island could qualify, provided they contained "a village or other thickly settled area having more than 2500 inhabitants and comprising, either by itself or when combined with other villages in the same town, more than 50 percent of the total population of the town" (Truesdell, 1949: p. 9). However, in these cases, it was still the whole town that was treated as urban, not just the thickly settled portion.

Also in 1930 and 1940, a few political subdivisions in other states, regarded as analogous to the New England towns, were treated as urban under a special rule if they had at least 10 000 residents, a population density of at least 1000 per square mile, and included no subordinate incorporated place. Most of these special-rule areas were suburban in character and located in states with strong town or township governments. However, at that time the urban category still excluded extensive suburban development located in 'unincorporated' areas outside municipal boundaries.

Before the 1940 census, totals of urban population using the 2500 threshold were determined all the way back to the first US census in 1790. Towns in New Hampshire, Massachusetts, and Rhode Island were included for census years in which they appeared to meet the criteria adopted in 1930. The special-rule areas in other states were included only for 1940 and 1930.

The remainder of the population was defined as rural. With minor revisions and the addition of data for 1950 determined on the 1940 basis, these urban and rural totals for 1790 to 1950 based on places of at least 2500 have been published in each subsequent census (U.S. Census Bureau, 2012: Table 7).

Urbanized Areas

As early as 1790, urban development overflowed across some municipal boundaries, notably around Philadelphia. If the overflow was either annexed by the parent city or separately incorporated in places meeting the population threshold, it qualified to be included in the official urban category. However, by about 1900, annexation activity by large cities had slowed and a significant share of suburban development was taking place in unincorporated territory (Weber, 1899/1967).

The census responded to such changes from 1910 through 1940 by defining 'metropolitan districts,' which represented a functional approach not limited to the continuously built-up environs of cities. The next section of this article discusses them further.

The 1950 census introduced the concept of the UA, with the stated aim of providing a better separation of urban and rural populations in the vicinity of larger cities. A UA was defined to comprise at least one city of at least 50 000 and its more or less contiguous built-up suburbs, incorporated or unincorporated. Incorporated places were never split by a UA boundary, but in New England towns, suburban townships, and unincorporated territory generally only the built-up parts were included.

The 1950 UAs were defined using precensus fieldwork and a density criterion of at least 500 housing units per square mile. Industrial areas and other zones of distinctly urban land use also were included even if they did not meet the housing density requirement, and the UAs have always allowed gaps of up to 1.5 miles in built-up continuity, under specified conditions. In 1960, the density limit was lowered slightly to 1000 persons per square mile applied to small enumeration areas. Starting in 1970, the UA definitions began to exclude sparsely settled portions of certain incorporated places.

Census Designated Places

The 1950 census also identified and defined many population concentrations outside UAs that were not separately incorporated, including many in New England. These unincorporated places, which were renamed 'census designated places' (CDPs) in 1980, were defined for concentrations as small as 1000; those of at least 2500 were treated as urban. CDPs located within UAs were not separately identified in 1950, but many have been recognized in later censuses.

Together, the recognition of UAs and CDPs added a net 6.7 million to the 1950 urban population, increasing it from 59.6 to 64.0% of the US total (U.S. Census Bureau, 2012: Table 7). Collectively, the 1950 changes altered the US urban definition from one based almost entirely on municipal status and boundaries to one that embraced densely populated unincorporated territory near larger cities, and also recognized qualified population concentrations even if they were not separately incorporated. These modifications, however, still did not include most unincorporated suburban development around smaller cities and towns.

After 1970, some UAs were defined as of that year around certain cities of less than 50 000, and starting in 1980 a UA was recognized for each area whose total population was at least 50 000. Supplementing the official series, Pickard (1967a,b) used the 1970 census criteria to define UAs of at least 50 000 for 1920–40, and also UAs not already defined by the census for 1950 and 1960, mainly because lacking any city as large as 50 000.

By 1990, technical developments permitted replacing manual determination of UA boundaries from detailed maps with interactive analysis of population density at the block level, in turn allowing UA definitions to systematically exclude sparsely settled portions of incorporated places (U.S. Bureau of the Census, 1994: Chapter 12).

Urban Clusters

In 2000, UA delineation was essentially fully automated, and the criteria were loosened slightly to permit inclusion of some development down to a population density of 500 per square mile. But the major development of 2000 was the recognition of urban clusters (UCs), identified by applying the delineation rules for UAs to identify all concentrations of at least 2500 population.

For purposes of defining the urban category, the clusters superseded smaller incorporated places or CDPs, either including them in a cluster, or eliminating them by determining that they did not have as many as 2500 residents at the required density level. Thus, as of 2000, urban territory in principle was defined completely independently of administrative boundaries.

To document the effect of the 2000 changes on the size of the urban category, the 2000 criteria were also applied to 1990 data. The changes would have added a net of 7.0 million to the 1990 urban population and raised the urban share from 75.2 to 78.0% (U.S. Census Bureau, 2012: Table 7).

Changes in the defining procedure for the 2010 census were relatively few, mainly involving qualification of specified types of nonresidential areas as urban if adjacent to a UA or UC.

In 2010, altogether UAs and UCs included 249.3 million people and a land area of 106 386 m^2, amounting to 80.7% of the national total population of 308.7 million, but only 3.0% of the land area. As with most earlier definitions, 'rural' comprised the rest of the national territory.

Useful as the UAs are, they are far less well known than the MAs discussed in the next section. Indeed, the public's chief awareness of UA boundaries may stem from the reduced speed limits that generally apply when entering them (Plane, 2004).

For users who consider 2500 too low a threshold for the urban category, data are available for urban territory by cluster population size. For example, if the threshold for cluster recognition were 25 000 instead of 2500, the 2010 urban share would comprise 73.9% of the total population and 2.6% of the area (U.S. Census Bureau, 2012: Table 6).

United States: MAs

The general term 'metropolitan area' means a large city including its suburbs. The term has developed mostly independently of the related term 'metropolis,' a city of major importance in its country or region; many officially recognized MAs in the United States are not metropolises in that sense. In contrast to a UA, which is a high-density concentration of continuous urban development and thus in a sense a single place, an MA is conceived as a functional entity, embracing both a UA as its core and surrounding territory closely related to that core.

The MAs used in official US statistics since 1950 are 'metropolitan statistical areas' (MSAs) defined by the US Office of Management and Budget (OMB) as a geographic standard for all federal statistical agencies, ensuring that statistics from different agencies will refer to consistent boundaries. OMB defines an MSA as a large population nucleus or core plus adjacent communities having a high degree of economic and social integration with that core. Integration is measured using census data on commuting to work. The Census Bureau conducts the technical work needed to arrive at the individual MSA delineations.

In general, entire counties form the MSA building blocks, so that besides a city and its UA, MSAs typically include more distant satellite communities and much open country. In the 2010 census, MSAs were reported with a total of 258.3 million residents, of which 218.0 million, or 84.4%, were in UAs, while the MSA total land area was 912 992 m^2 of which only 85 177, or 9.3%, was in UAs. UCs within MSAs included another 10.0 million residents and 6520 m^2, leaving 30.3 million residents and 821 295 m^2 defined as both rural and metropolitan (U.S. Census Bureau, 2012: Tables 8 and 9).

Recognizing and Delineating MSAs

The first US official recognition of functional-type MAs came in the 1905 Census of Manufactures, which presented data on 13 'industrial districts' defined in terms of subcounty MCDs. In the 1910 decennial census, 'metropolitan districts' were officially defined, generally in terms of MCDs, which were included chiefly on the basis of a population density of at least 150

per square mile. The list of metropolitan districts was updated and boundaries revised after 1920, 1930, and 1940 (U.S. Bureau of the Census, 1913: pp. 73–77; Thompson, 1948). These were probably the earliest national metropolitan definitions that stated specific criteria (Forstall et al., 2009).

However, definition in terms of subcounty units greatly limited the usefulness of metropolitan districts for the compilation of statistical data other than from the decennial census, since most such data were not available except by county. As a result, in 1949, the US Bureau of the Budget (later renamed the OMB), in consultation with the various statistical agencies, established criteria to define 'standard metropolitan areas' (SMAs) for presentation in the 1950 census.

This called for decisions on four fundamental aspects: (1) the choice of geographic building blocks, (2) the minimum size for qualification as an SMA, (3) how to define the core area of the SMA, and (4) how to determine which if any additional units to include. To maximize their usefulness, generally the SMAs were defined in terms of counties. They included every city of at least 50 000, with the city's county constituting the central core. However, just as with the urban definition, the unusual character of New England county subdivisions resulted in special treatment; the SMAs and cores in those states were defined in terms of cities and towns instead of counties.

Commuting data were chosen for determining whether any counties beyond the core should be included in the SMA. There were no national commuting data at the county level, but transportation and other agencies had conducted many local surveys. The final criteria required 15% of workers commuting to the core for an outlying county to qualify. Counties also had to meet specified requirements of 'metropolitan character,' such as having less than one-third of their labor force engaged in agriculture and at least half their residents living in contiguous MCDs with 150 persons or more per square mile. There were 169 SMAs defined for 1950, consisting of 265 counties and 208 New England cities and towns.

The SMAs were rechristened 'standard metropolitan statistical areas' (SMSAs) in 1959 to emphasize that they were not necessarily suitable as general-purpose MAs, and various other changes were made in the criteria after 1950 (Forstall, 2012).

As the availability of data for them made the SMSAs better known, they could be said to suffer from their own popularity. Some local areas wanted to achieve SMSA status in spite of not qualifying for it under the official rules. For example, in 1959, portions of the New York and Chicago SMSAs were split off to create several separate SMSAs, and the New York and Chicago areas as previously defined were recognized as 'standard consolidated areas' (SCAs), thereby establishing a two-level hierarchy of SMSAs within SCAs. Finally, in 1972, the Nassau-Suffolk SMSA on Long Island was created out of the New York area, although it did not even contain a qualifying central city.

Changes for 1980

The census-defined UAs, first established in 1950, were finally accepted as the basis for defining the SMSA central core, consisting of counties with at least half their population in the UA. To deal with the inconsistencies presented by the Nassau-Suffolk, New Jersey, and Chicago definitions, the criteria expanded the two-level hierarchy of areas. Within SMSAs of 1 million or more, defined strictly according to the published rules, component 'primary metropolitan statistical areas' (PMSAs) were recognized based mainly on specified commuting criteria. Any area in which PMSAs qualified was termed a 'consolidated metropolitan statistical area' (CMSA) while all other areas remained MSAs, thus supplanting the terms SCA and SMSA. Recognition of a PMSA required support from local opinion, which OMB ascertained through congressional offices. The need to recognize both CMSAs and PMSAs emphasized a fifth fundamental question in defining MAs: how to decide when two hitherto separate areas have become a single multicentered area.

After the updating based on the 1980 census, there were 351 areas in June 1983, including 253 MSAs and 22 SCAs comprising 76 PMSAs.

Developments 1990–2010

The 1990 review of the MA standards made only minor changes in the criteria, but many users declared that they should be less complicated, even while recognizing that metropolitan structure had become more complex. During the 1990s, OMB and the Census Bureau undertook an extensive review of the standards, including commissioning proposals by outside scholars (Dahmann and Fitzsimmons, 1995). Revised standards in December 2000 included converting to county-unit definitions of MSAs in New England, although ancillary definitions in terms of cities and towns were retained as 'New England city and town areas.' The commuting level required to include an outlying county was raised from 15 to 25%, and the earlier requirements of metropolitan character (density, recent growth, etc.) were eliminated. PMSAs were replaced by 'metropolitan divisions' but only within MSAs of at least 2.5 million.

The major change for 2000 was the recognition of 'micropolitan areas' around the newly recognized UCs, provided the UC population was at least 10 000. The term 'core based statistical areas' (CBSAs) was adopted for the collective metropolitan and micropolitan areas. Provision was also made for recognizing 'combined statistical areas' (CSAs) consisting of adjacent CBSAs (metro or micro) if their employment interchange was at least 15%, with local support required unless interchange exceeded 25%; the components of CSAs retained their status as CBSAs.

Only a few changes were made in the standards for 2010. The term 'definition' was replaced by 'delineation,' and the delineation of CSAs was revised to qualify inclusion of any adjacent CBSA if employment interchange was at least 15%, without reference to local opinion.

In February 2013, the MSA delineations were updated using estimates of commuting from the 2006–10 five-year American Community Survey (U.S. Office of Management and Budget, 2013). The 2013 updating resulted in 381 MSAs, totaling 262.5 million people or 85.0% of the 2010 US total, compared with 84.9 million and 56.1% in the SMAs in 1950. There were also 536 micropolitan areas, accounting for another 8.8% of the 2010 total and leaving only 6.2% of the population outside CBSAs. Together the CBSAs accounted for 47.4% of the US land area (U.S. Census Bureau, 2013).

The official US MAs have come into wide use by both government agencies and private industries. They also have served as a model for defining MAs abroad (International Urban Research, 1959). Changes in the criteria over time have reflected such factors as the availability of new data and techniques of analysis, more extensive use of the delineations outside of government, and increased interest in urban, suburban, and exurban trends.

Urban Definitions in Other Countries

Definitions Used in 40 Large Countries

The most significant difference in how countries define urban relates to how they delimit the individual urban place. In the first edition of its *Demographic Yearbook*, the United Nations Statistical Office (1948: Table 9) recognized three main types of locality definitions in what is still a useful classification:

(a) Agglomerations without regard to official boundaries or forms of government.
(b) Areas with official boundaries under the jurisdiction of local or "urban" forms of government. These are special areas and do not comprise the whole country.
(c) Relatively small (or smallest) areas or administrative subdivisions with fixed boundaries which, in sum, comprise the whole country. These may or may not be under local forms of government.

The United Nations' (UN's) current biennial projections of world population include 2011 estimates of each country's urban population and a note specifying the country's definition of its urban category (United Nations Population Division, 2012). Forty countries are listed with at least 15 million urban population; together they account for 3057 million or 82% of the estimated world urban total of 3632 million. Based primarily on the information in that publication, the following discussion groups these 40 large countries into nine categories, broadly based on the UN categories of 1948. Following the UN's usage, 'agglomeration' denotes a concentration of continuous urban-type development.

A – Agglomerations, not based on administrative boundaries
A1 – Agglomerations defined primarily by population density: the United States, United Kingdom, Canada, and Australia, with a total urban population of 356 million, use density for small enumeration units to arrive at agglomeration boundaries. Their population thresholds for urban qualification range from 1000 in Canada and Australia to 2500 in the United States and 10 000 in the United Kingdom.
A2 – Agglomerations defined based on urban continuity, which may extend across administrative boundaries: France and Algeria, with a total urban population of 80 million. France's definition requires 2000 residents in the main agglomeration of a commune, defined as consisting of houses and other urban structures separated by less than 200 m. However, it is the whole commune that is treated as urban, as are other communes with built-up continuity with a qualifying commune, even though their population is below the 2000 limit. Thus, France follows an agglomeration approach, but defines the agglomeration in terms of entire communes.
A3 – Agglomerations described as localities, population centers, or places: Mexico, Argentina, Colombia, Vietnam, Venezuela, and Chile, with total urban population of 233 million. Urban thresholds range from 2000 (Argentina) to 4000 (Vietnam). Exactly how localities are defined is not always specified.
A4 – Agglomerations provided they are seats of local administrative units: Brazil and Peru, with a total urban population of 189 million. Available evidence suggests that this criterion includes some very small localities, although their effect on the country's urban total may be trivial.

B – Urban units are municipalities, distinguished from nonmunicipal territory
B1 – Urban units are incorporated municipalities: Japan, the Russian Federation, Nigeria, Pakistan, Iran, Ukraine, Thailand, Poland, Saudi Arabia, and Iraq, with a total urban population of 540 million. Most countries in this group do not cite an urban threshold, but Nigeria uses 20 000 and Saudi Arabia 5000. How closely this urban definition compares with the other categories depends on whether municipalities are tightly or generously defined. For example, Japan's cities, which comprise its urban population, included 90.7% of its population and 57.2% of its area in 2010 (Japan Statistics Bureau, 2012: Table1).
B2 – Most urban units are municipalities, but some are nonmunicipal areas with specified urban characteristics: India, Bangladesh, Malaysia, Morocco, and Myanmar (Burma), with a total urban population of 486 million. India and Bangladesh generally use a threshold of 5000 and Malaysia of 10 000. Myanmar's criteria were not available to the UN in 2012; it is placed here assuming that (like India and Bangladesh) it still follows the model established decades ago by the Census of India.
B3 – Urban units limited to capitals of provinces and districts: Turkey and Egypt, with a total urban population of 89 million. Especially in Egypt (2011 urban proportion 43.5%), this definition probably omits considerable urban-type development.
B4 – Urban units identified largely in terms of cities or municipalities, not necessarily including their entire areas, and sometimes including noncity areas: China, Indonesia, Philippines, Republic of Korea, South Africa, and the Democratic Republic of Congo, with a total urban population of 945 million. The urban criteria for these countries typically are lengthy and detailed and some 'cities' are overbounded. This is especially true of China, whose urban definition is presented in the final section of this article.

C – Urban units are small administrative units (communes). Germany, Italy, and Spain, with a total urban population of 138 million comprise this category. Italy and Spain use a threshold of 10 000.

Changes in Urban Definitions

Some countries have changed therir urban definitions in recent decades. In Western Europe, definitions in many countries used to be based on municipalities, whose territory was largely urban in character. Subsequent amalgamations of local-government units resulted in expanding most municipalities

to include much rural territory (Council of European Municipalities and Regions, 2009: pp. 4–7). This encouraged adoption of urban definitions based on individual agglomerations, for example, in Sweden, Norway, Denmark, and the United Kingdom. Such changes not only improve the preciseness of the urban definition, but also add to the difficulty of making valid comparisons of urban population across countries and over time.

Official Urbanized and Metropolitan Definitions in Other Countries

Besides the countries listed in categories A1 and A2 above, only a few countries publish officially defined urban agglomerations. India's urban agglomerations are defined in terms of contiguous municipalities and nonmunicipal urban outgrowths. Japan has published densely inhabited districts (DIDs) since 1960, based on small enumeration areas, density of at least 4000 persons per km^2 and 5000 total population. Individual DIDs do not cross municipal boundaries, but detailed maps permit identifying continuous agglomerations. Formerly, the DIDs were the basis for defining Japan's urban population.

Many countries define functionally conceived MAs under various labels, generally in terms of minor administrative divisions. Besides the United States, these include Canada, Mexico, Colombia, Brazil, Belgium, Netherlands, Switzerland, Austria, Hungary, the Scandinavian countries, Bangladesh, Japan, Australia, and New Zealand. The defining criteria vary widely and not all countries specify them in detail, but most appear to use commuting data to determine metropolitan boundaries.

Besides these national definitions, some international bodies have presented analogous definitions on their Web sites, such as functional urban areas (European Observation Network, or ESPON) and 'larger urban areas' (Eurostat), supplementing rather than superseding the nationally developed definitions.

The UN's list of largest so-described urban agglomerations (United Nations Population Division, 2012), although widely used, unfortunately is an internally inconsistent mixture of the MA, UA, or city-proper definitions of individual countries (Forstall et al., 2009). Systematic efforts to define MAs or UAs internationally include those by International Urban Research (1959), Hall and Hay (1980), and Moriconi-Ebrard (1993, continued by the Geopolis Web site). Champion and Hugo (2004), Hall and Pain (2006), and Champion (2011) have pointed to evolutionary trends that raise new questions about urban and metropolitan definitions. The unofficial Web sites of Geographia and the Lincoln Institute of Land Policy present extensive data on large world urban areas (Angel, 2012; Angel et al., 2012), and various other Web sites offer population data on cities and towns by country and in some cases on MAs or UAs.

Urban Definitions in China

The People's Republic of China has what may be the world's most complicated and confusing way of defining its urban population. The sources of confusion are multiple and multilayered. The complexity is first compounded by the Chinese household registration (*hukou*) system, which does not consider most migrants to cities as permanent residents and excludes many of them from the city's annual *de jure* population counts. In contrast, the decennial census reports population on a *de facto* basis, producing counts often quite different from the annual count. Moreover, the coexistence of multiple urban and city definitions also complicates the task of defining a city and city population. Finally, rapid development over the past three decades in the country and its major cities, including numerous changes in the urban definition and local administrative boundaries, has also added complexity.

A Chinese city or municipality (*shi*) is an administrative unit, but the 'city' label can be misleading. Large cities refer to those in the categories of prefecture-level and provincial-level cities. There are also many smaller cities (county-level cities) and several thousand towns. The principles used in defining 'urban' within them are similar to those for large cities. Most large Chinese cities today encompass an extensive area, which contains an urbanized core (high-density built-up area), surrounded by numerous scattered towns and large stretches of rural territory, usually with dense farming populations. These 'cities' are so large in area that they are more aptly called 'regions.'

In almost all large cities, the urban core, together with some close-in built-up and rural areas, is divided into 'city districts' (*shiqu*), while the surrounding rural areas (including many towns) are divided into counties (*xian*). The urban administrative area (the collective city districts) generally is larger than what can be called the continuously built-up UA around the core.

Besides referring to the entire region-sized administrative area, the word *shi* is also often used to refer to a smaller area – the aggregate of city districts only. Even more confusing, the term 'city population' (*shi renkou*) can mean the population of any of the following: the city region (including its subordinate counties), the aggregate of city districts, the aggregate urban areas (see below) within the city region, or the aggregate urban areas within the city districts only.

Up until about the mid-1990s, the aggregate of city districts did roughly correspond to the UA for most large cities. However, since then, many largely rural counties have been reclassified administratively as city districts in the pursuit of rapid 'urbanization' (Yew, 2011). In other words, many city districts now include large stretches of rural territory. As a result, in the 2000 census, China's National Bureau of Statistics (NBS) defined two urban categories, city urban and town urban. Under this definition, city urban includes, first, all city districts with a population density of at least 1500 persons per km^2. Second, for other city districts, urban areas are (1) all subdistrict units (*jiedao* (streets), *zhen* (towns), and *xiang* (townships)) that include the city or district government headquarters, or street government offices; (2) subdistrict units with built-up areas contiguous to the units specified in (1); and (3) all other *jiedao* units. Town urban comprises village-level units, which are residents' committees (*jumin weiyuanhui*) or include the town government headquarters, and other units whose built-up area is con-

tiguous to the preceding. Based on that urban definition, China's *de facto* urban population accounted for 36% of the total population in 2000. Many specialists considered the results acceptable and generally comparable to urban counts used in other countries (see, e.g., Chan, 2007).

China's urban definition underwent further changes in 2006 and 2008. The latter definition was used in the 2010 Census and is in use currently. In cities and towns, urban areas consist of residents' committee units where local-government headquarters (city, district, county, or township) are located, and units with built-up structures contiguous to those units. Additional units that are not contiguous but have a population of at least 3000 (such as development zones and mining areas) are also considered urban areas. Based on this definition, the 2010 Census counted 666 million urban population, or 50% of the nation's population.

While the city population statistics from the census and other NBS sources are for the *de facto* population, the *de jure* set published by the public security authorities using local household registers (*hukou*) is widely used within various levels of governments. These *hukou* population data count only the population with legal permanent registration in the area, and do not reflect the actual population in residence. Under China's *hukou* system, each citizen is required to register in one and only one place of (permanent) residence. An individual's *hukou* status defines his or her rights and eligibility for social welfare and various services, including public education and housing, within a specific administrative unit. Since there are often significant disparities in services, especially between rural and urban units, change of one's *hukou* registration is ordinarily not granted even though one has migrated to a new place (Chan, 2010). Some minor changes to the *hukou* system have been implemented locally in recent years and bolder changes are currently being discussed and considered.

For almost all large cities, the *hukou* population is smaller than the *de facto* population because of in-migration of people whose *hukou* remains back in their original place of registration. In some extreme cases, the differences between the registered (*hukou*) population and the actual *de facto* population can be huge. In 2010, in Shenzhen adjacent to Hong Kong, local *hukou* population was only 2.6 million, whereas the *de facto* census population was 10.4 million.

Suburbanization began to emerge around China's large cities in the 1990s and has accelerated in the last decade. Their expanding high-density cores are increasingly encircled by somewhat less dense suburban development. For a variety of analytical and policy purposes, it is very useful to be able to define a metropolitan boundary based on a daily labor-shed concept, as is done in many other countries.

However, there are no official or standard criteria for defining MAs in China. While the MA concept has frequently been mentioned by researchers, often they do not mean by it the daily labor shed. Since commuting data are still not systematically collected or available publicly, some researchers have proposed alternative methods to delimit MAs (e.g., Chan and Forstall, 2012).

See also: Urban Planning: Methods and Technologies.

Bibliography

Angel, S., 2012. Planet of Cities. Lincoln Institute of Land Policy, Cambridge, MA.

Angel, S., Parent, J., Civco, D.L., Blei, A.M., 2012. Atlas of Urban Expansion. Lincoln Institute of Land Policy, Cambridge, MA.

Arriaga, E., 1975. Selected measures of urbanization. In: Goldstein, S., Sly, D.F. (Eds.), The Measurement of Urbanization and Projection of Urban Population. IUSSP Committee on Urbanization and Population Distribution, Ordina, Dolhain (Belgium), pp. 19–87. Working Paper 2.

Champion, A., 2011. The changing nature of urban and rural areas in the United Kingdom and other European countries. In: Zlotnik, H. (Ed.), Population Distribution, Urbanization, Internal Migration and Development: An International Perspective. United Nations, New York, pp. 144–160.

Champion, T., Hugo, G., 2004. New Forms of Urbanization. Ashgate, Aldershot.

Chan, K.W., 2007. Misconceptions and complexities in the study of China's cities: definitions, statistics, and implications. Eurasian Geography and Economics 48 (4), 383–412.

Chan, K.W., 2010. The Chinese household registration system and migrant labor in China: notes on, a debate. Population and Development Review 36 (2), 357–364.

Chan, K.W., Forstall, R.L., 2012. Delimiting Beijing's Metropolitan Area. Paper Presented at Annual Meeting of Association of American Geographers February 25, New York.

Council of European Municipalities and Regions, 2009. Changes in Local and Regional Structures in Europe. www.ccre.org/docs/changes_in_local_and_regional_structures_web_EN.pdf (accessed 13.06.26.).

Dahmann, D.C., Fitzsimmons, J.D., 1995. Metropolitan and Nonmetropolitan Areas: New Approaches to Geographical Definition. Working Paper Series No. 12. U.S. Bureau of the Census, Washington, DC.

Forstall, R.L., 2012. Metropolitan areas. In: Anderson, M.J., Smith, R.Z., Salvo, J.J. (Eds.), Encyclopedia of the U.S. Census, second ed. CQ Press, Thousand Oaks, CA, pp. 311–314.

Forstall, R.L., Greene, R.P., Pick, J.B., 2009. Which are the largest? Why lists of major urban areas vary so greatly. Tijdschrift voor Economische en Sociale Geografie 100 (3), 277–297.

Gibbs, J.P. (Ed.), 1961. Urban Research Methods. Van Nostrand, Princeton, NJ.

Goldstein, S., Sly, D.F. (Eds.), 1974. Basic Data Needed for the Study of Urbanization. IUSSP Committee on Urbanization and Population Distribution, Ordina, Dolhain (Belgium). Working Paper 1.

Hall, P., Hay, D., 1980. Growth Centres in the European Urban System. University of California Press, Berkeley.

Hall, P., Pain, K., 2006. The Polycentric Metropolis. Earthscan, Sterling, VA.

International Urban Research, 1959. The World's Metropolitan Areas. University of California Press, Berkeley.

Japan Statistics Bureau, 2012. Densely Inhabited Districts. Somusho Tokeikyoku, Tokyo.

Moriconi-Ebrard, F., 1993. L'Urbanisation du Monde depuis 1950. Anthropos, Paris.

Pickard, J.P., 1967a. Dimensions of Metropolitanism. In: Research Monograph, vol. 14. Urban Land Institute, Washington, DC.

Pickard, J.P., 1967b. Appendixes to Dimensions of Metropolitanism. In: Research Monograph, vol.14A. Urban Land Institute, Washington, DC.

Plane, D.A., 2004. Population distribution—geographic areas. In: Siegel, J.S., Swanson, D.A. (Eds.), The Methods and Materials of Demography. Elsevier Academic Press, San Diego, pp. 81–104.

Thompson, W.S., 1948. The Growth of Metropolitan Districts in the United States: 1900–1940. U.S. Government Printing Office, Washington, DC.

Truesdell, L.E., 1949. The development of the urban-rural classification in the United States: 1874–1949. U.S. Bureau of the Census, Current Population Reports, Series P-23, No. 1. U.S. Government Printing Office, Washington, DC.

United Nations Population Division, 2012. World Urbanization Prospects: The 2011 Revision. United Nations, New York.

United Nations Statistical Office, 1948. Demographic Yearbook 1948. United Nations, Lake Success, NY.

U.S. Bureau of the Census, 1913. Thirteenth Census of the United States Taken in the Year 1910. In: Population, 1910, vol. I. U.S. Government Printing Office, Washington, DC.

U.S. Bureau of the Census, 1994. Geographic Areas Reference Manual. Bureau of the Census, Washington, DC.

U.S. Census Bureau, 2012. 2010 Census of Population and Housing, Population and Housing Unit Counts, CPH-2-1. United States Summary, Washington, DC.

U.S. Census Bureau, 2013. Population Change for Metropolitan and Micropolitan Statistical Areas in the United States and Puerto Rico (February 2013 Delineations): 2000 to 2010 http://www.census.gov/population/www/cen2010/cph-t/CPH-T-5.pdf.

U.S. Office of Management and Budget, February 28, 2013. Revised Definitions of Metropolitan Statistical Areas, Micropolitan Statistical Areas, and Combined Statistical Areas, and Guidance on Uses of the Delineations of These Areas. OMB Bulletin No. 13-01.

Walker, F.A., 1874. Statistical Atlas of the United States Based on the Results of the Ninth Census 1870: With Contributions from Many Eminent Men of Science and Several Departments of the Government. J. Bien, New York.

Weber, A.F., 1899/1967. The Growth of Cities in the Nineteenth Century. Cornell University Press, Ithaca.

Yew, C.P., 2011. Pseudo-urbanization? Competitive government behavior and urban sprawl in China. Journal of Contemporary China 20 (72), 281–298.

Relevant Websites

http://www.citypopulation.de – City Population (accessed 13.07.05.).
www.demographia.com – Demographia (accessed 13.07.05.).
http://www.espon.eu – ESPON (accessed 13.07.05.).
http://epp.eurostat.ec.europa.eu – Eurostat (accessed 13.07.05.).
http://www.geohive.com – GeoHive (accessed 13.07.05.).
www.e-geopolis.eu – Geopolis (accessed 13.09.21.).
http://www.lincolninst.edu – Lincoln Institute of Land Policy. (accessed 13.07.05.).
http://www.statoids.com – Statoids (accessed 13.07.05).
un.org/en/development/desa/population/ – United Nations Population Division. Topics: Biennial Projections of National Totals, National Urban, Large Urban Agglomerations (accessed 13.07.05.).
unstats.un.org – United Nations Statistics Division. Topics: Demographic Yearbook, National Census Dates (accessed 13.07.05.).
www.census.gov – U.S. Census Bureau. Topics: Publications, Geography, Urban, Metropolitan; Web sites of National Statistical Agencies (accessed 13.07.05.).

Urban Planning (Western), History of

Robert Freestone, University of New South Wales, Sydney, NSW, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Planning ranks as a significant social movement of the twentieth century. Its modern history is an ongoing story of achievement, resilience, failure, and adaptation to place, culture, and circumstance. The scope of planning has expanded dramatically from its nineteenth-century origins addressing the challenges of industrial urbanization. Early planning goals crystallized around the themes of health, convenience, and beauty. By mid-century, the guiding ideal was the modernist functional city. After World War II, planning became ensconced in the welfare state. Toward the end of the century, planning came under attack on many fronts, fragmenting its traditional land-use focus into many specialties. In recent decades, the planning agenda has become more diffuse in response to intractable challenges in the urban environment. Sustainability has provided a new focus. In the early twenty-first century, planning operates at many different scales in a variety of forms but land-use allocation, development control, and stewardship of the public realm constitute recurrent themes. The balance sheet of more than a century of modern urban planning in Western cities shows a positive contribution to their productivity, sustainability, and quality of life.

The evolution of urban planning is inextricably linked with the processes of modernization and the nature of modernity. The emergence of planning in Western cities reflected the response of both private interests and the state to problems of accelerated urbanization from the end of the nineteenth century. The agenda and scope of urban planning have subsequently expanded dramatically. Its forms have ebbed and flowed in concert with myriad economic crises, social problems, environmental and demographic pressures, political opportunism, and technological changes. At the start of the twenty-first century, urban planning is a multifaceted set of policy and administrative processes that resist easy categorization.

While an early focus on reforming the physical framework for urban life has braided in various social, economic, and legalistic directions, some common threads have characterized the mainstream history of planning theory and practice. These include a pervasive ideology of holism and the benefit of the 'bird's eye view' in directing urban development; direct involvement in the allocation of land uses; a practical commitment to reconciling the diverse interests of stakeholder groups; an emphasis on order, scientific analysis, and rationality in public policy decision making; belief in the power of environmental improvement to deliver community benefits; and an ongoing faith in the potency of enlightened state intervention to secure the 'public interest' in the face of market failure.

Antecedents and Beginnings

For such a quintessential twentieth-century activity, urban planning has a remarkably long history. But this premodern planning invariably meant rudimentary forms of spatial layout and architectural set-pieces disengaged from broader policy concerns. Indeed, planning was all too frequently co-opted as a mere technical device for the imposition of authoritarian will. Planning as civic design stretches back to antiquity – to ancient Egypt, Greece, and Rome – with the organization of urban spaces manipulated to reflect the dominant political and social order. This thread continues through the grand public squares and piazzas of the renaissance city and into the planning templates for imperial expansion, notably the Vitruvian-inspired Spanish Laws of the Indies of the sixteenth century.

Planned towns were variously seats of government, ostentatious displays of autocratic power, military footholds, and service centers for land settlement schemes. Rectilinear, gridiron, or chessboard street plans were the most time-honored morphologies, but by the start of the nineteenth century these models were being studded with open spaces and framed by belts of parkland. Since antiquity, ideal cities assumed a variety of alternative geometries. Radial plans would decisively influence the first generation of early modern planners. Dating from the beginning of the eighteenth century, Karlsruhe (1715) in Germany became one exemplar for a new form of street pattern that was both aesthetic and functional in fanning out from a central edifice.

A more modern urban planning began to emerge in the late nineteenth century to address challenges unleashed by industrial urbanization. Existing narrowly focused administrative procedures and primitive building regulations were ill-equipped to address an interdependent set of new challenges: unprecedented rates of urban population growth; housing shortages and overcrowding; new forms of economic enterprise and infrastructure needs; and a rising tide of environmental pollution.

Diverse modes of thought and practical innovations were tapped to help shape the new discourse of planning. At the theoretical end of the spectrum, for example, a nexus was forged between a longer tradition of social and economic utopian thought and the newer notion of better cities through works like James Silk Buckingham's *National Evils and Practical Remedies* (1849) and Edward Bellamy's *Looking Backward* (1887). Other thinkers, like the influential anarchist Peter Kropotkin, looked toward a more decentralized society aligning village life with new modes of production. The most substantial experiments in industrial decentralization were more socially conservative. These were the planned industrial villages expressing an emergent form of welfare capitalism geared to educating and disciplining a productive labor force. What Walter Creese in *The Search for Environment* (1966) called

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74015-3

the Bradford–Halifax School of model village builders in the English Midlands developed into full-blown community building in places like Bournville and Port Sunlight. The same sort of settlements arose in continental Europe and the United States, for example, George Pullman's eponymous industrial community near Chicago.

Utilitarian issues preoccupied reformers within the cities proper, as basic infrastructure (transport, water supply, sewerage) struggled under new conditions of rapid growth. Poor housing conditions for the growing working class in the form of high-density slum and tenement dwellings became a major concern. Exposes by reformers like New York's Jacob Riis in *How the Other Half Lives* (1890) shocked middle-class sensibilities. Philanthropic model lodging houses and stricter regulations over space standards for rooms, dwellings, allotments, and, in turn, streets ensued. Provision of parks and open spaces was another major concern. Utilizing the organic metaphors that were so prevalent in early planning propaganda, parks were the 'lungs of the city' and needed to be reserved in advance of development. Citizens improvement associations, the precursor to civic art and 'tidy towns' movements, provided a bottom-up chorus of complaint. A major theme through all of this was the need to secure higher standards of public health and hygiene, a crusade that extended to cognate missions of social, moral, and cultural cleansing. Those who could afford to fled the city for the suburbs. Elite low-density communities such as Riverside (1869) near Chicago helped advance planning technique.

Synthesizing ideas in the air, the early basic precepts of planning were expressed through a rich palette of modernist thought. An American contribution was the park system – the concept of interlinked networks of open spaces mixing formal and informal elements and parkways developed by Frederick Law Olmsted Sr., that gave cities like Boston, Denver, and Kansas City a green legacy enjoyed by later generations. The enormously influential garden city idea was similarly associated with an individual, British shorthand reporter Ebenezer Howard, and advanced in his book *To-morrow: A Peaceful Path to Real Reform* (1898), reprinted as *Garden Cities of Tomorrow* (1902). At its most visionary, the manifesto called for clusters of new, self-contained 'slumless, smokeless cities' of limited size on community-owned land as an alternative to the sprawling urban-industrial city. Alternative new urban structures were mooted. In 1892 Spanish engineer Arturo Soria y Maya invented the idea of the linear city (*la ciudad lineal*), which resurfaced in twentieth-century forms such as Edgar Chambless's Roadtown (1910), the ideas of the Soviet de-urbanists in the 1930s, and corridor city planning in the 1960s.

A string of seminal contributions signposting ways ahead emerged in Europe in the late nineteenth century. Town extension planning and land-use zoning were already accepted municipal practices in Germany before 1900. Important texts were Reinhard Baumeister's *Stadterweiterungen in technischer, baupolizeilicher und wirtschaftlicher Beziehung* (Town Expansions Considered with Respect to Technology Building Code and Economy, 1876) and Joseph Stubben's *Der Städtebau* (1890). The *Lex Adickes* regulations for Frankfurt in 1891 prefigured land-use zonation based on sets of standardized regulations. Baron Haussmann's extraordinary transformation of Paris under the patronage of Napoleon III from a largely medieval city into a modern masterpiece of boulevards, star-points, and strategically located public buildings had begun in the 1870s. The ambitious 1859 extension plan of Ildefons Cerdà in Barcelona was an even earlier initiative. A very different aesthetic to these large-scale interventions was that of Viennese architect Camillo Sitte whose *Der Städtebau* (1889) helped codify more artistic, intimate, and historically sensitive techniques for the modern 'art of designing cities and suburbs.'

International exhibitions provided unique opportunities to showcase new urban planning ideas as planned cities in microcosm. The most sensational synthesis was the 1893 World's Columbian Exposition in Chicago. This collaborative integration of grand architecture, landscape design, public art, and early amusement park was overseen by local architect Daniel Burnham, who was to emerge as the leading American architect-planner of the early 1900s.

The crucial intellectual breakthroughs made simultaneously around the Western world in the 1880s and 1890s were recognition of the limitations of piecemeal improvement, the interconnectedness of urban life, and the necessity for more coordinated intervention to avoid the undermining of the social order through outbreaks of disease, community dissent, and loss of economic competitiveness. A new approach was being invented around the world adapted to particular institutional circumstances: town planning in Great Britain, city planning in North America, *städtebau* in Germany, *planificación urbana* in Spain, and *urbanisme* in France.

The First Half of the Twentieth Century

At the turn of the last century, modern planning goals crystallized around recurrent themes of convenience, efficiency, healthfulness, beauty, and order. The actual means to these ends were quite diverse, ranging from totally planned landscapes through selective interventions in the urban fabric to new regulatory frameworks for private development. Certain kinds of planning assumed a cultural distinctiveness. House-and-garden were at the center of British concerns, France was known for its monumentalism, while planning advocates looked to Germany for its innovative land management and to the United States for the business-minded aestheticism of city beautiful thought. The planning movement had become genuinely international. Even a relatively closed Asian country like Japan, in acknowledging the importance of modernization in economic and political advancement, had discovered Western planning ideas through study tours and private visits. Political alliances were crucial for the export of knowledge, so the British impact was strongest in Commonwealth nations and French ideas in various francophone settings.

Major conferences were important agents in the diffusion of knowledge. A spin-off from Howard's Garden City Association (1899; renamed the Town and Country Planning Association in the early 1940s) was the International Garden Cities Association (1913), which would evolve into the International Federation for Housing and Town Planning. Information sharing would be damaged by World War I, but planning was well established as a global movement with astonishing commonalities of purpose and method in diverse settings. With that came moves to institutionalize and inevitably

bureaucratize the activity as a genuine profession occupying the interstices between the traditional built environment disciplines of architecture, surveying, and engineering. Urban planning was also developing its own methodology, with the accent on scientific civic surveys, an extension of settlement worker investigations. Scottish polymath Patrick Geddes supplied the requisite mantra of 'survey-analysis-plan.'

In the English-speaking world, several interconnected events in Britain in 1909 would have major reverberations, especially around the anglophone world: the first formal town planning course started with establishment of the William Lever-endowed Department of Civic Design at the University of Liverpool; Raymond Unwin's classic *Town Planning in Practice* appeared; and the first planning legislation was passed in the form of the *Town Planning, Housing etc* Act. The British Parliament was told of the government's intention to realize 'the home healthy, the house beautiful, the town pleasant, the city dignified, and the suburb salubrious.' The following year, the Royal Institute of British Architects organized an international conference and exhibition that attracted the leading names in world planning and showcased the state of the art.

Early incursions into planning controls were cautious, with governments mindful of private property rights. The British law was initially restricted to zoning for new extensions only but was later extended to existing development. Land-use zoning established itself as the standard form of microregulation in the United States and elsewhere. A popular early mode of zoning was the residential district, a concept that was grafted readily onto existing local government regulations and widely invoked as a tool of social exclusion. A 1916 New York City zoning ordinance introduced novel controls on building height and shape that led to the Manhattan 'wedding cake' skyscraper form. It was not until the historic 1926 US Supreme Court case *Euclid, Ohio vs. Ambler Realty Co*. that the legal validity of planning was finally confirmed in America. British and continental models interconnected planning and zoning, but in the United States city planning commissions and zoning boards could pursue independent agendas.

The early twentieth century is distinguished by some landmark plans that codified modernist planning ideas. The 1901 plan for Washington, D.C., by the Senate Park Commission headed by Daniel Burnham reconfigured Pierre L'Enfant's eighteenth-century plan into a monumental triumph for city beautiful ideas. Many plans for American cities would follow, but the most impressive was the privately sponsored *Plan of Chicago* (1909) by Burnham and Edward Bennett. Artistic in presentation and regional in scope, it set a new standard for comprehensive planning. Edward Lutyens's New Delhi (1912) transplanted grand manner planning to an imperial context. The 1911–12 international competition for the new federal capital of Australia produced a unique survey of early modern planning thought. The winning scheme by Walter Burley and Marion Mahony Griffin, both former associates of Frank Lloyd Wright, was a brilliant landscape synthesis of garden city, city beautiful, and deeper democratic ideals. The most idealistic plan of its day was the footloose *World City for Peace* (1912) by American sculptor Hendrik Anderson and French architect Ernest Hébrard. The multivolume Regional Plan of New York masterminded by Thomas Adams in the late 1920s was the archetypal pragmatic, professional, growth-oriented plan of the interwar period. A celebrated critique by urbanist Lewis Mumford channeling the decentralist ideas of the rival Regional Planning Association of America branded the scheme as dinosauric in entrenching rather than curbing trends toward megalopolitan development. Mumford had further clashes with Robert Moses, a powerful New York official known as the 'master builder' with a penchant for large-scale and often ruthless public works.

Mumford was deeply influenced by the garden city movement and advocated a more organic, environmentally sensitive variant of regional planning. While stopping short of Howard's dream of a 'better, brighter civilization,' garden city advocates campaigned tirelessly for new towns. Their successes were modest but resulted in impressive showpieces of planning in practice. Two British garden cities were built: Letchworth (1904) designed by Raymond Unwin and Barry Parker and Welwyn Garden City (1919) by Louis de Soissons. Welwyn was conceived as an overspill satellite community for London, an idea readily taken up elsewhere. An early German response was the fringe *grossiedlung* with apartment blocks of social housing. In the United States, echoes of the garden city notion can be read into the 1920s speculative sunbelt 'new towns' such as those designed by John Nolen in Florida; the New Deal towns of the late 1930s such as Greenbelt, Maryland sponsored by Rexford Tugwell's Resettlement Administration; and Clarence Stein and Henry Wright's Radburn, New Jersey, with its cul-de-sacs and 'superblocks' representing an experimental 'garden city for the motor age.'

The garden city label and concept was translated to many countries: cite-jardin, gartenstadt, ciudad jardin. The most ubiquitous application was to suburban rather than new town settings, becoming more popular and prolific through the prosperous 1920s in concert with rapid suburbanization generally and with privately master-planned communities often leading the way. These ventures refined further an enduring amalgam of site planning ideas revolving around house-and-garden living, hierarchical street systems, provision of open space and public facilities, and attention to landscaping and streetscape qualities. Clarence Perry synthesized these ideas into the widely influential notion of cellular 'neighbourhood units' but the low-density suburban ideal reached its zenith in Frank Lloyd Wright's *Broadacre City* dream of the early 1930s.

Other expressions of modernism were more vertiginous. New York City's Rockefeller Center in the 1930s foreshadowed the penchant for large-scale urban redevelopment. The most startling theoretical schemes were those of European theorists like Auguste Perret, Ludwig Hilbersheimer, and most famously the messianic Swiss architect Le Corbusier. Through the 1920s, Le Corbusier worked his way through a succession of urban utopias – La Ville Contemporaine, City of Three Million Inhabitants, and the La Ville Radieuse – to perfect a design vocabulary of towers-in-parkland, multilevel transportation systems, and functional land-use segregation. His *Plan Voisin* applied these ideas to a portion of central Paris, sweeping away narrow streets and 'diseased quarters' for a forest of cruciform high-rise tower blocks. The legacy of these ideas can be seen in large-scale state housing programs and planned capital cities such as Chandigarh and Brasilia. The guiding spirit here was architectural modernism, and emerging out of the

deliberations of CIAM (*Congrès internationaux d'architecture moderne*) was the Athens Charter – the definitive town planning manifesto of the 1930s for 'the functional city' distinguished by specialized transportation and freeway systems, high-rise buildings, extensive open spaces, and planned dormitory communities.

Postwar Master Planning

World War II delayed implementation of this blueprint but unleashed a universally strong planning impulse under the banner of postwar reconstruction. In many European cities, the necessary reconstruction was as much physical as ideological. But in most Western nations cross-political agreement on the relevance of state involvement in urban development processes moved planning decisively beyond the rhetoric, experimentation, and broad theoretical acceptance attained by the 1930s into the realm of implementation. Through the 1950s, the scale of planning activity grew significantly as planning settled comfortably into the policy machinery of the welfare state. The bureaucratization of planning was underpinned by a steady professionalization and growth in tertiary education to train the larger numbers of experts required.

The dominant discourse was that of urban planning as an unambiguously beneficent activity of government: comprehensive, technocratic, scientific, and socially progressive. More was better if cities were to remain healthy, efficient, prosperous, and beautiful. The major focus was on land-use planning. The elements of modernist planning had surfaced before the war and were now synthesized into a total ensemble embracing master plans; comprehensive redevelopment of the inner city; monofunctional land-use zoning; high-rise social housing towers; low-density middle-class suburbs; planning and replanning of residential areas as neighborhood units; freeways, ring roads, and pedestrianized town centers; green webs of open space; industrial decentralization; and new towns. From mid-century, these planning ideas, variously interpreted and combined, but broadly accepted by architects, developers, and government, made decisive impacts on urban landscapes worldwide.

All-encompassing metropolitan blueprints assumed a central status. Three early master plans from the mid-1940s influencing many that followed were those for London, Stockholm, and Copenhagen. Patrick Abercrombie's definitive plans for the city and county of London, 1943–44, were built around a cellular community structure, an encircling greenbelt to contain suburban sprawl, and a ring of new towns. Sven Markelius's general plan for Stockholm 1945–46 skillfully integrated transportation, land use, and environmental planning, with new subcenters developed around railway stations. And the first Copenhagen finger plan of 1947 also anticipated the transportation corridor and green wedge planning that would gain popularity from the 1960s when encircling greenbelts seemed doomed by escalating suburbanization.

Postwar city strategies targeted both obsolete inner areas and planning opportunities on new greenfields sites. Urban renewal schemes involving high-rise housing, upgraded transportation infrastructure, private- and public-sector offices, and other new uses sought to modernize the late industrial city, frequently doing so at the expense of existing community structure and historic building fabric. Efficient urban mobility remained a major concern, and an influential prospectus was Colin Buchanan's *Traffic in Towns* (1963), which legitimized planning ideas to become commonplace in Britain and internationally such as clearways, flyovers, pedestrianized precincts, and multistorey car parks. The garden city dream was kept alive in new towns. The most heroic program was in Britain, with the *New Towns Act* 1946 facilitating construction of over 30 new towns by public corporations between 1947 (Stevenage) and 1970 (Milton Keynes). Similarly, around Paris a ring of *villes nouvelles* was conceived within an ambitious 1965 master plan (SDAURP). Private-sector new towns (such as Reston and Columbia on the east coast) were more common in the United States. Everywhere, however, and even if it did not look like it, considerable planning energy was devoted to the processes of facilitating suburban development. The high point of technocratic master planning allied to 'scientific' methodologies and modernist architectural solutions was the late 1960s.

Postconsensus Urban Planning

The new world order that emerged in the 1970s constituted a markedly different context. The notion of the centralized, benevolent state crumbled. Economic policy was forced to look beyond the industrial fordism that had been the touchstone of social progress in the first half of the twentieth century. Dissent over unpopular wars and civil rights spilled onto the streets. More frequent calls were heard for greater community participation and self-help. The sterility of modern design trends was challenged. And the environmental movement awakened. In the face of these challenges, even if planning seemed to be realizing prewar dreams in physical terms, it was not always delivering attractive, equitable, and environmentally sound outcomes.

As the world changed, urban planning came under attack on many fronts: for standardized formulaic solutions, freeway proposals slicing through established neighborhoods, destruction of historic buildings, minimal attention to social issues, and facilitation of suburban sprawl. A famous critique of mainstream planning ideology with its fixation on the steamroller 'radiant garden city beautiful' was Jane Jacobs's *The Death and Life of Great American Cities* (1961). Robert Goodman portrayed planners as the 'soft cops' in *After The Planners* (1972). If the demolition of the award-winning planned housing project Pruitt–Igoe (1972) in St. Louis was symbolic of the end of architectural modernism, it also represented the loss of faith in the 'brave new world' international-style planned landscapes of the postwar era. Other countries had their version of Pruitt–Igoe, such as the partial collapse of the high-rise Ronan Point development in London in 1968. Many unpopular grand redevelopment plans for Western cities were scuttled by citizen action during these years. But there was also a higher-level backlash against planning evident in the 1980s that fundamentally questioned the central role of government planning as an agent for the wider public interest. The political consensus that had been the platform for the rise of urban planning after World War II was severely strained.

New directions consistent with a search for postmodern solutions were sought as negative impacts from the decline of traditional industries took hold. There were moves to reinvent central cities hit hard by suburbanization and industrial abandonment. The American developer James Rouse promoted the festival marketplace concept to rejuvenate the retail appeal of struggling downtowns like Boston and Baltimore. Observers like William Whyte in his *The Social Life of Small Urban Spaces* (1980) began to turn their attention to the design and use of the public realm in more human-centered ways. Alternatives to both wasteful low-density and bleak high-rise housing were more explored. A more environmental approach to design also developed, a particular concern for more liveable and safer spaces captured by C. Ray Jeffery's *Crime Prevention through Environmental Design* (1971). Planning responded to its own crises of legitimacy by fragmenting into more popular smaller targets such as advocacy planning, participatory planning, cultural planning, economic development planning, transportation planning, historic preservation, and urban design. In recent decades, the planning agenda has become more diffuse as cities have become ever more difficult to manage.

New Millennium Planning

In the early twenty-first century, cities face a raft of pressures; some enduring if updated from the early days of the town planning movement, others new and just as troublesome. Financing constraints, calls for greater participation and accountability, transport congestion, environmental degradation, continuing sprawl, social polarization, peak oil, global warming, and risks from extreme events and disasters are some of the problems now being confronted. Balancing productivity and livability is an ongoing challenge. Not all places have to contend with growth issues; a 'shrinking cities' syndrome is also evident with long-term structural economic decline. Globalization has introduced intense competitive pressures. The rise of global cities, captured in Saskia Sassen's *The Global City* (1991), is associated with new growth vectors: the information and service economies in particular. New York, London, and Tokyo are the three alpha cities dividing up a 24-h world. Below them is a fiercely competitive battle for prominence in particular time zones across the eastern and western hemispheres. Michael Porter's *The Competitive Advantage of Nations* (1990) was influential in shaping public policy agendas and planning responses.

Unlike an earlier era when the state would assume fiscal leadership, the new era of neoliberalism instead sees governments of all political persuasions privileging market forces. Reaganism and Thatcherism in the 1980s were the templates, although the Global Financial Crisis of 2007–10 and its lingering effects prompted a rethink and a return in some sectors to more command and control interventionism. Nevertheless, the early twenty-first-century city is one where ideologies of privatization, commercialization, and entrepreneurialism hold sway. The manifestations within planning systems have been diverse and include reforms to cut the red tape of planning regulation, the preference for public–private partnerships to share risks and responsibilities in a climate of declining public funding, elevating city branding strategies to public policy, and designating hot spots for accelerated development with special administrative relationships outside everyday arrangements. The enterprise zone concept is a related outcome of the latter trend, intended to stimulate free enterprise through financial incentives with a minimum of bureaucratic rules. Energetic city-states like Hong Kong and Singapore served as a model.

Urban regeneration has become a key policy concern. Three sets of circumstances are common: offering opportunities to redevelop often vast tracts of redundant industrial and port lands; ensuring the vitality of central areas in decline because of loss of commercial activity to rival suburban centers as captured in Joel Garreau's *Edge City* (1991); and reinvigorating the decline of former suburban showpieces like shopping malls and planned social housing estates. In the shift from greenfields to brownfields (industrial) and grayfields (underdeveloped suburban lands), a common approach is precinct development based on purpose-created agencies created to fast-track development with major private-sector involvement. Recurring land-use types include office towers and business parks, higher density private housing, waterfront upgrades, adaptive reuse of historic districts, convention centers, sporting stadia, trade centers, and casinos.

The London Docklands redevelopment centered on Canary Wharf, which began in the 1970s, first showcased showpiece regeneration possibilities on an expansive scale. Exemplars worldwide have followed, including Battery Park City in Manhattan, the Tokyo Waterfront, Barcelona, Marseilles, Bilbao, Liverpool, and Vancouver. Progressive civic leadership is a recurring feature of many successful transformations. In Germany, International Building Exhibitions from the early 1980s featured actual physical demonstration projects. Emscher Park (from 1999) in the former powerhouse of the Ruhr comprises over 90 separate employment and cultural projects, many drawing on the region's industrial heritage. Major urban regeneration strategies employ eye-catching iconic projects. The Grands Projets of former French presidents Mitterrand and Sarkozy in Paris reflect the highest-level political priority to quality design outcomes. The staging of hallmark events like the Olympic Games is now linked to transformative postgame urban strategies. The London Games (2012) were described as the 'regeneration games' for improving the quality of life in the inner eastern suburbs of the Lea Valley. Megaprojects such as large infrastructure and city gateway schemes in various cities are all geared ultimately to increasing accessibility and connectivity and enhancing competitiveness in what Manuel Castells in *The Rise of the Network Society* (1996) evocatively called 'the space of flows,' but all carry inherent financial and political risks.

The integration of land use, transport planning, and design has become a new priority. Peter Newman and Jeff Kenworthy's *Cities and Automobile Dependence* (1989) was an influential sourcebook that documented the costs of car reliance in low-density cities, but the real breakthroughs have been in practical schemes. European-style light rail systems have proven popular worldwide. They were adopted progressively in many American cities from the 1980s, even in notoriously car-dependent cities like Los Angeles (1990) and Houston (2004). Promoting active transport usage has become an integral aspect of urban public policy. For example, citywide bicycle rental programs, which

began appearing in the 1970s and received a fillip from the Parisian (Velib) scheme launched in 2007, numbered over 400 worldwide by 2012. Superior urban design standards are now factored into expansion and retrofit projects. The Danish urbanist Jan Gehl, author of *Cities for People* (2010), has been an influential promoter of walkability, cyclability, and liveliness in human-scaled central cities.

Culture has assumed a growing importance in urban revitalization, its possibilities sketched in influential tracts such as Charles Landry's *The Creative City* (2000) and Richard Florida's *The Rise of the Creative Class* (2002) where place-based creativity and mobile artistic professionals are seen as fundamental drivers of economic growth. Many cities have opted for cultural planning strategies. The most successful transnational program is the annual European City of Culture launched in 1985 and targeted at city marketing, physical regeneration, and tourism. Culture has been employed at various levels in urban development processes, including major investment in new flagship performance spaces, museums, and art galleries; development of cultural precincts and 'quarters'; support for public art and more grass-roots ventures; and the staging of seasonal activities such as fairs, music festivals, and cinema in the park.

The dominant planning paradigm in organizing responses to the major challenges presented by the twenty-first century city is sustainability. The standard definition of sustainable development is 'development which meets the needs of the present without compromising the ability of future generations to meet their own needs.' This concept was popularized in *Our Common Future*, a report published by the United Nations' World Commission on Environment and Development in 1987, also known as the Brundtland Report. It is generally accepted that sustainable development calls for a convergence between economic development, social equity, and environmental protection.

Endorsed and elaborated at the highest level of the United Nations beginning with a landmark Summit in Rio de Janeiro in 1992, these ideas have inexorably devolved to guide state, provincial, regional, and local planning strategies worldwide.

Sustainability has helped to reinvigorate the mission of planning and has proven adaptable to different interpretations, although concrete policy implementation has proven difficult. It guides many different types of strategic initiatives and regulatory control, but environmental goals involving protection, conservation and replication of natural systems tend to be paramount. Progress has been made on indicators or metrics to measure the impact of sustainable development policies. Green building rating systems for water and energy efficiency, materials recycling, indoor environmental quality and other criteria have been institutionalized in many jurisdictions, for example, LEED (USA), CASBEE (Japan), NABERS (Australia), and BREEAM (United Kingdom). Sustainability aims are now routinely incorporated into environmental assessments of new development. Sustainability has also reinvigorated attention to the important links between human health and the built environment. This was an early theme of the planning movement nurtured through the postwar period and by later contributions such as Leonard Duhl's *The Urban Condition* (1963) and then the World Health Organization's Healthy Cities Program (1987), but only since the 1990s has there been a renewed interest in urban form. Another emerging issue has been planning to secure sustainable food supplies; Carolyn Steel's *Hungry City* (2008) highlights the need for urban food planning.

New approaches to urban and suburban development channel broader imperatives toward sustainable design. These share similar objectives for a sustainable urbanism but with different emphases. What American planners call Smart Growth strategies encompass anti-sprawl measures, higher development densities, preservation of rural land and greenspace, and designation of urban growth boundaries. Portland, Oregon is considered an exemplar of such policies. Transit Oriented Development (TOD) promotes the creation of compact communities centered around high-quality public transit systems to lessen dependence on cars. One of the earliest and most successful examples of TOD is Curitiba in Brazil, based on an extensive bus system network. The New Urbanism is associated with American architect-planners like Andrés Duany, Elizabeth Plater-Zyberk, and Peter Calthorpe. It emphasizes walkable communities featuring mixed-use development with a range of housing types and commitment to superior design qualities enforced by architectural-based codes. The seasonal vacation community Seaside (1984) on the Florida Panhandle coast embodies the connection to neo-traditional design inspired by vernacular styles. Poundbury (1993), designed by Luxembourg designer Leon Krier, captures a British version drawing on English vernacular design traditions. The impact of the New Urbanism can be traced into master-planned communities, gated communities, infill projects, and 'grayfield' redevelopments.

Climate change has assumed a prominent place in sustainable development thinking. It was cemented in the public imagination by the documentary film *An Inconvenient Truth* (2006) narrated by former U.S. Vice President Al Gore. Setting the global agenda is the United Nations Intergovernmental Panel on Climate Change established in 1988 to provide comprehensive scientific assessments of changing climatic parameters, risks, consequences, and responses to the world's continuing dependence on carbon-based energy. Major assessment reports are released every few years. The role of spatial planning in developing both mitigation measures to reduce emissions and adaptation measures to ameliorate the effects of climate change has grown. The mitigation measures encompass actions around reducing car travel, increasing energy efficiency, and encouraging alternative energy sources; the adaptation measures pertain more to controlling development in high-risk areas.

Cities are now expanding and even merging to create urban settlements on an unprecedented scale. Mega city-regions and corridors are emerging in various parts of the world to become the new engines of both global and regional economies. Central government now has much greater awareness of the role of cities in economic development. There has also been a resurgence of metro-regional strategies and strategic visions to address the management of powerful centrifugal forces for dispersion, coordinate provision of infrastructure, facilitate joint ecological projects, secure balanced economic development, and guarantee spatial equity in terms of access to opportunities. Planning systems are struggling to keep up with the challenges and pressures. Moves toward addressing problems of political fragmentation through constituting new

unitary city-regional authorities are evident but struggle because of entrenched political opposition to radical reform. Incremental steps can be taken through voluntary regional agreements evincing the moves toward the popular theory of collaborative planning. International treaties can encourage the positive convergence of city responses to key issues, such as climate change. This has been evident in 'borderless' Europe. The European Spatial Development Perspective enacted in 1999 constitutes a policy framework aimed at balanced and sustainable spatial development. A Territorial Agenda agreed to by EU members in 2007 builds on the earlier compact. There is a strong theme of orderly management of polycentric city-regions. Randstad Holland was an early model, popularized by Peter Hall in *The World Cities* (1966).

The Achievements and Legacy of Urban Planning

Urban planning ranks as a significant social movement that came of age in the twentieth century. Its modern history is a complex and ongoing story of achievement, failure, resilience, and redemption. Over time, chameleon-like, it has taken a role that has been constantly adapted to place, culture, and circumstance. It is a history of both continuities (the quests for health, justice, efficiency, environment, and amenity) and discontinuities (in response to global technological, economic, political, and climate shifts).

Modern urban planning is now over a century old. The scholarly study of urban planning history burgeoned in the last quarter of the twentieth century, with an International Planning History Society formed in England in 1974. Planning over the decades since then has grown further in complexity at different scales, driven by a mix of motivations, and often striving to reconcile top-down and bottom-up aspirations. It has numerous public and private stakeholders enjoying different levels of power and influence. It is defined in different ways as community building, urban management, development control, and strategic planning. And it is rare, at least at the supra-local level, to find a singular planning department as a self-contained or even peak coordinating agency. Instead, planning is usually distributed across many agencies dealing with transport, economic development, communities, and environmental protection. Modern urban planning in Western countries is far from monolithic and even at the most general level assumes different cultural inflections – for example, British town and country planning focused on managing land use, European urbanism on design, and American planning with its social and economic policy orientation. The core foci remain regulation of development, design of the built environment, and valuation of the public realm.

Historically, planning thought has experienced its greatest periods of innovation in response to periodic social crises – explosive population growth, war, global economic shifts. Planning's aspirations have always soared above the practicality of their comprehensive realization, being buffeted and reshaped continually by broader social constraints. At times, planning has been a deeply flawed practice, with outcomes often the reverse of those intended (destruction rather than enhancement of community, promotion of segregation rather than social cohesion, dreary townscapes rather than inspired urban design). This is planning's 'noir' side. Always more political than merely technical, sometimes it has been exploited as a tool of social control to pursue policies of exclusion, segregation, and marginalization.

At the same time, there is a raft of value-adding achievements in areas such as open space provision, affordable housing, mass transit, public infrastructure, development of liveable, new communities, injection of sanity into myriad land-use determinations, and protection of natural and cultural resources. There is general community acceptance that constraints on individual freedoms and property rights may be warranted for collective benefits. On balance, urban planning initiatives have meant that a significant proportion of humanity on every continent finds itself living in better circumstances at the beginning of the twenty-first century than 100 years previously. World Town Planning Day, an initiative launched in the late 1940s, is now celebrated worldwide every November.

See also: Cities: Capital, Global, and World; Community, Social Context of: The US Case; Modernization and Modernity in History; Neighborhood Revitalization and Community Development; Planning Ethics; Planning Issues and Sustainable Development; Urban Geography; Urban Sociology; Urban Sprawl; Urban Studies: Overview.

Further Reading

Bosma, K., Hellinga, H., 1997. Mastering the City 1: North-European City Planning 1900–2000. NAI Publishers/EFL Publications, Rotterdam.

Freestone, R. (Ed.), 2000. Urban Planning in a Changing World: The Twentieth Century Experience. Routledge, London.

Hall, P., 2013. Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century, fourth ed. Blackwell, Cambridge.

Hall, P., Tewdwr-Jones, M., 2011. Urban and Regional Planning, fifth ed. Routledge, London.

LeGates, R., Stout, F. (Eds.), 1998, Early Urban Planning: Selected Essays, vol. 9. Routledge, London.

Sanyal, B. (Ed.), 2005. Comparative Planning Cultures. Routledge, New York.

Sies, M.C., Silver, C., 1994. Planning the Twentieth-Century American City. Johns Hopkins University Press, Baltimore.

Ward, S.V., 2002. Planning the Twentieth-Century City: The Advanced Capitalist World. John Wiley and Sons, Chichester.

Ward, S.V., Freestone, R., Silver, C., 2011. The new planning history: reflections, issues and directions. Town Planning Review 82, 231–261.

Urban Planning in China

Tingwei Zhang, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article summarizes the historical and contemporary urban planning practices and policies in China, and discusses China's rapid urbanization following its economic reform since the 1980s and the role of urban planning in promoting urbanization and stimulating economic growth.

Urban planning could be understood as the professional work of guiding urban development through making physical plans and development regulations, or more broadly as social, economic, environmental, and constructive efforts to make an urban area a better, healthier, and more just place to live, work, and get around. If urban planning means physical development, then China is one of the civilizations with a long history and rich experiences of urban planning activities. Planned towns were built before the Zhou Dynasty of 1000 BC when town planning principles and rules were officially recorded in *Zhou-Li*: *Kao-gong-ji* (Figure 1). If urban planning includes physical as well as economic and social development, then the urban planning practice in China has shown an approach and a consequence different from that in Western countries.

Both traditional Chinese culture and preferences of China's rulers over 4000 years have impacts on China's urban planning practice. The former determines the location, space order, and form of towns and buildings based on social norms and cultural beliefs, such as *Feng-Shui*, and it has more consistent influence; the latter defines and regulates planning activities following the ideology of the government at that time and personal biases of the ruler in power.

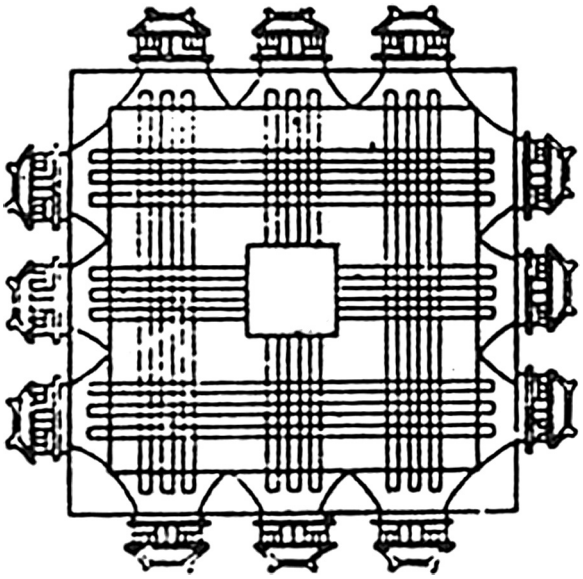

Figure 1 The ideal city model defined in *Zhou-Li*: *Kao-gong-ji*, the Zhou Dynasty, 1000 BC. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

Urban planning activities in China may be divided into three periods:

1. Recorded urban planning activities and planned cities in ancient and medieval China from the Shang Dynasty, 1700 BC, to Middle Qing Dynasty of 1840 (the year of the Opium War, which ended with the China–Britain Treaty of 1842, the symbol of China's entering into the quasicolonial age of modern times in Chinese history);
2. Urban planning in China's modern times from Middle Qing Dynasty to the Republic of China in mainland in 1949 with planning activities in China's early industrial and colonial cities;
3. Contemporary urban planning in the People's Republic of China after 1949 to present. This is the period with a fluid trajectory of urban policy and planning practice. The present urban planning practice accompanies large-scale urbanization and rapid growth of urban economy and urbanized areas. It is in this period that the function of urban planning shifted significantly from a central government's tool to implement central decisions under the planned economy to a local strategy to stimulate economic growth in the post-reform era.

Taiwan, where the government of the Republic of China has operated since 1949, and Hong Kong and Macau, China's two Special Administration Districts that used to be ruled by foreign administrations, follow regulations and principles different from the planning practices on the mainland. However, with the adoption of the open-to-outside reform policies since 1978, the differences between urban planning practices in the mainland and in Taiwan, Hong Kong, and Macau have become much smaller. While cities in various ethnic Chinese societies still show some influences of traditional Chinese culture, their appearances are becoming more similar in the age of globalization. Neoliberalism policy and 'selling the place at the global market' promotion are popular in all Chinese cities whether they are located on the mainland or across the Taiwan Strait.

Urban Planning Activity in Ancient and Medieval China (from Shang Dynasty, 1700 BC, to Middle Qing Dynasty, 1840)

Ancient Human Settlement in China and Feng-Shui

The oldest human settlement of about 10 000 years ago in China is recorded in *I Ching* or *Yi Jing*, one of the oldest Chinese classic texts in 1000 BC. Archaeologists found remains and sites

of ancient human settlements in several areas in China, including Lon-shan village and Ban-po village; most of these settlements were built 5000–7000 years ago. The plan and form of these settlements do not show recognizable planned patterns, but they do seem to follow some sort of order based on certain cultural and/or religious beliefs, as found in many ancient settlements of other ancient civilizations (A. Rapoport has done excellent work on the topic.) The cultural beliefs and social norms were developed gradually, and they eventually became important guidelines in planning and building towns and houses in the ancient Chinese society. The most famous guidelines are Feng-Shui (meaning the Wind and the Water).

Feng-Shui is a summary of experience learned in building houses, making tombs, and planning towns, plus many cultural beliefs passed on for generations among Chinese people. Feng-Shui lists a group of regulations for location selection and orientation. According to Feng-Shui, a town should be located in an area backed by mountains at the north, facing a river/lake at the south, with roads and agricultural fields on the east and west (Figure 2). Given China's geographic location in the northern hemisphere, the ideal location and orientation seem based on practical considerations – orienting to the south for direct sunlight and to block wind from the north. Other Feng-Shui regulations also provide advices for locating a house and making a layout of a room.

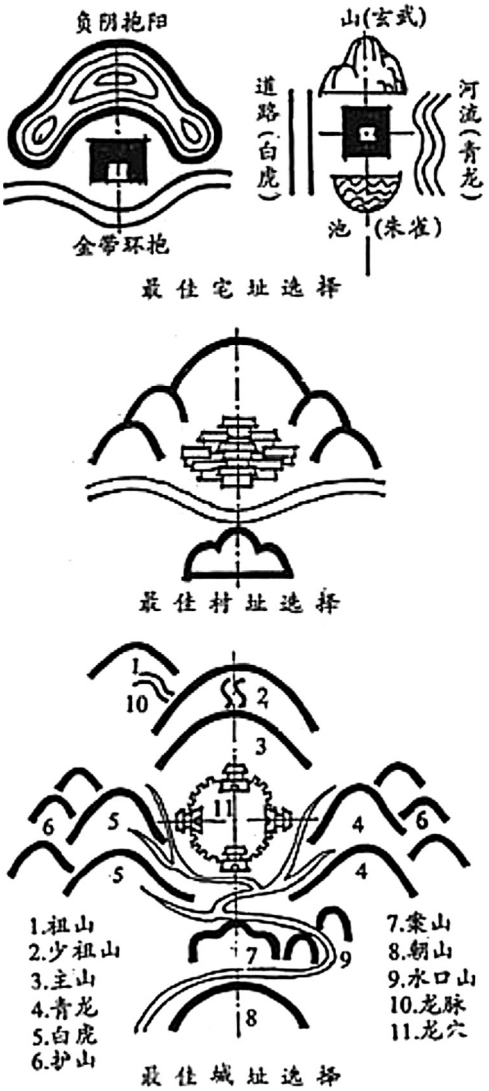

Figure 2 *Feng-shui* principles for selecting an ideal location of a town. Reproduced from Peng, G., 2001. History of Chinese Architecture. China Architecture and Building Press.

Recorded Planned Cities in Ancient and Medieval China

Guided by scholars' research findings, archaeologists have found sites and remains of several important ancient cities. There are two types of ancient cities: formal cities as capitals of states, kingdoms, and the nation and informal ones as main trading and transportation hubs. Both show certain planned patterns.

Chang-an city, the capital of the Tang Empire (618–907, Xi'an today), located in the Shan'xi Province in northern China, was the largest city in the world in the seventh to the tenth century (Figure 3). It was rebuilt on the old Chang-an city of the Qing Dynasty (221–206 BC), which was severely damaged during the war. As the capital of the world's most powerful nation at that time, Chang-an city had a population of over one million and a walled area of over 8700 ha. The city's plan reflected the ideal city model defined in *Zhou-Li*: *Kao-gong-ji* in the Zhou Dynasty (see Figure 1): it had a well-organized grid street system with three gates on each side of the city walls, and it was divided into many functional zones for the royal family (palaces in the middle north), administration agencies and senior officials (south of the palaces), and residential and commercial uses with a hierarchical order. This model was duplicated in other capital cities in various China's dynasties, even copied in Japan while building Heiankyo City (Kyoto today), the capital of Japan then.

While Chang-an is a typical planned formal city, Ping-jiang city (Suzhou today) of the Song Dynasty (960–1279) is recognized as a typical planned informal city with a more free style. Ping-jiang was the economic hub of the Yangtze River Delta, China's richest area. The city had a walled area of 3.8 km × 4.2 km, and five city gates instead of the ideal model of 12 gates. On comparing Ping-jiang to Chang-an, one can find similarities as well as differences: the plan of Ping-jiang located the regional administration core (the governor's mansion as well) in the central area, the Temple of Confucius (the most important temple in the Chinese culture) to the south of the administration core, and mansions of rich families in the surrounding areas. The load network also had a grid pattern (Figure 4). However, the layout respected the topographic setting of the city, so it was cut by many canals and had only a total of five gates rather than the ideal model of three gates on each side of the city walls.

China's most important capital city is Beijing. The first walled city in Beijing was Ji, the capital of the State of Ji from the eleventh to seventh century BC, and later the capital of the State of Yan, one of the powers of China's Warring States period (473–221 BC). Beijing as the national capital as seen today was founded during the Yuan Dynasty (1279–1368), and functioned as China's capital through the Ming Dynasty (1368–1644), the Qing Dynasty (1644–1911), and the

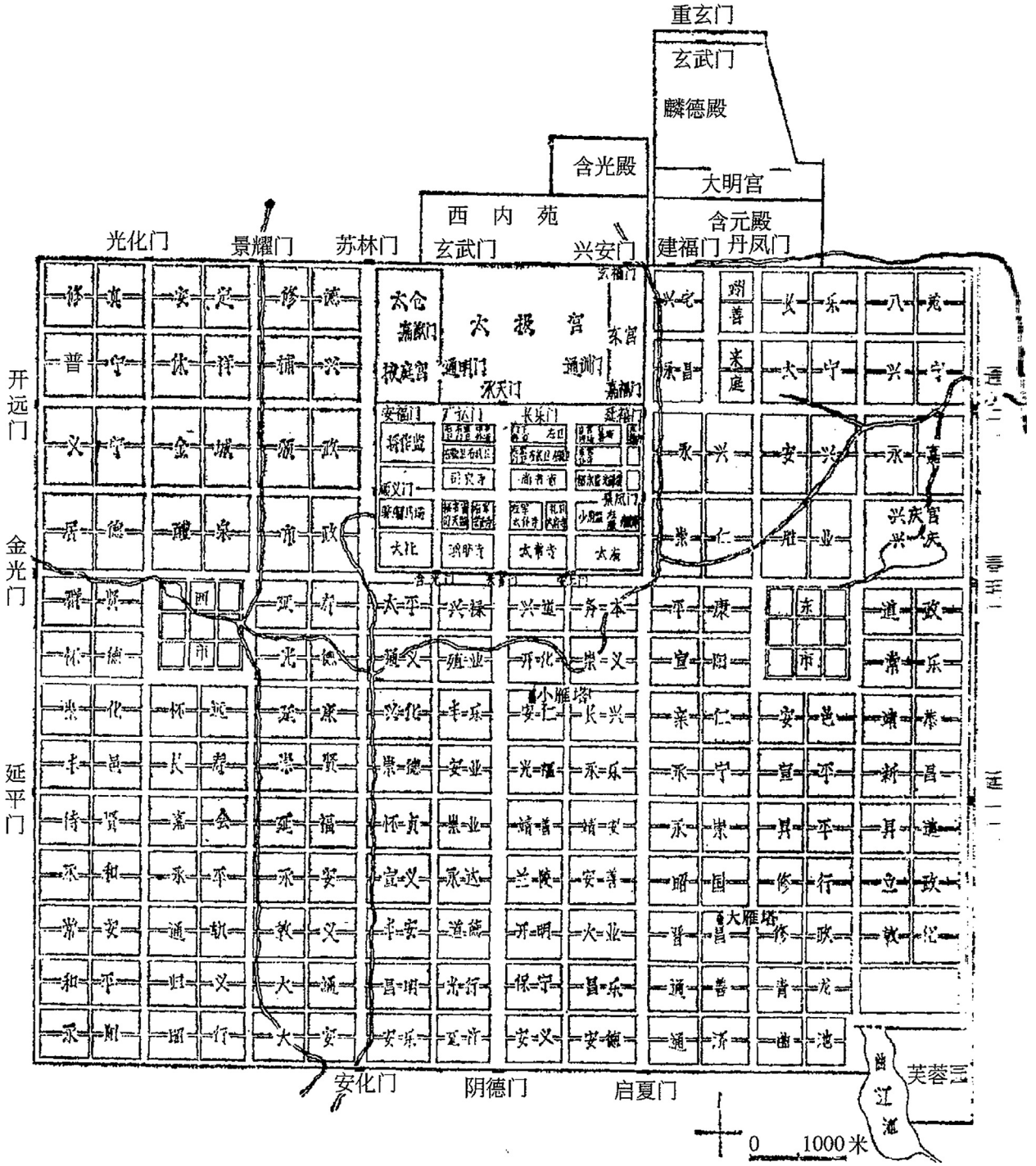

Figure 3 Chang-an city, the planned capital of the Tang Dynasty (618–907). Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

Republic of China on the mainland. It is still the capital of the People's Republic of China. The plan of Beijing represents all characters of an ideal Chinese capital: a hierarchical space order, strictly defined functional zones, and a grid street network (Figures 5 and 6). Many features displayed in Chang-an city of the Tang Dynasty could be found in Beijing, such as the location of the Forbidden City (the Emperor's palace) in the middle north, and a well-defined hierarchical order of space and buildings (Figure 7).

Kevin Lynch identified the five 'elements of urban form' in Beijing: paths, edges, nodes, districts, and landmarks. His analysis indicates that Chinese cities in ancient and medieval periods were developed based on a sophisticated tradition with full planning skills (Figure 8).

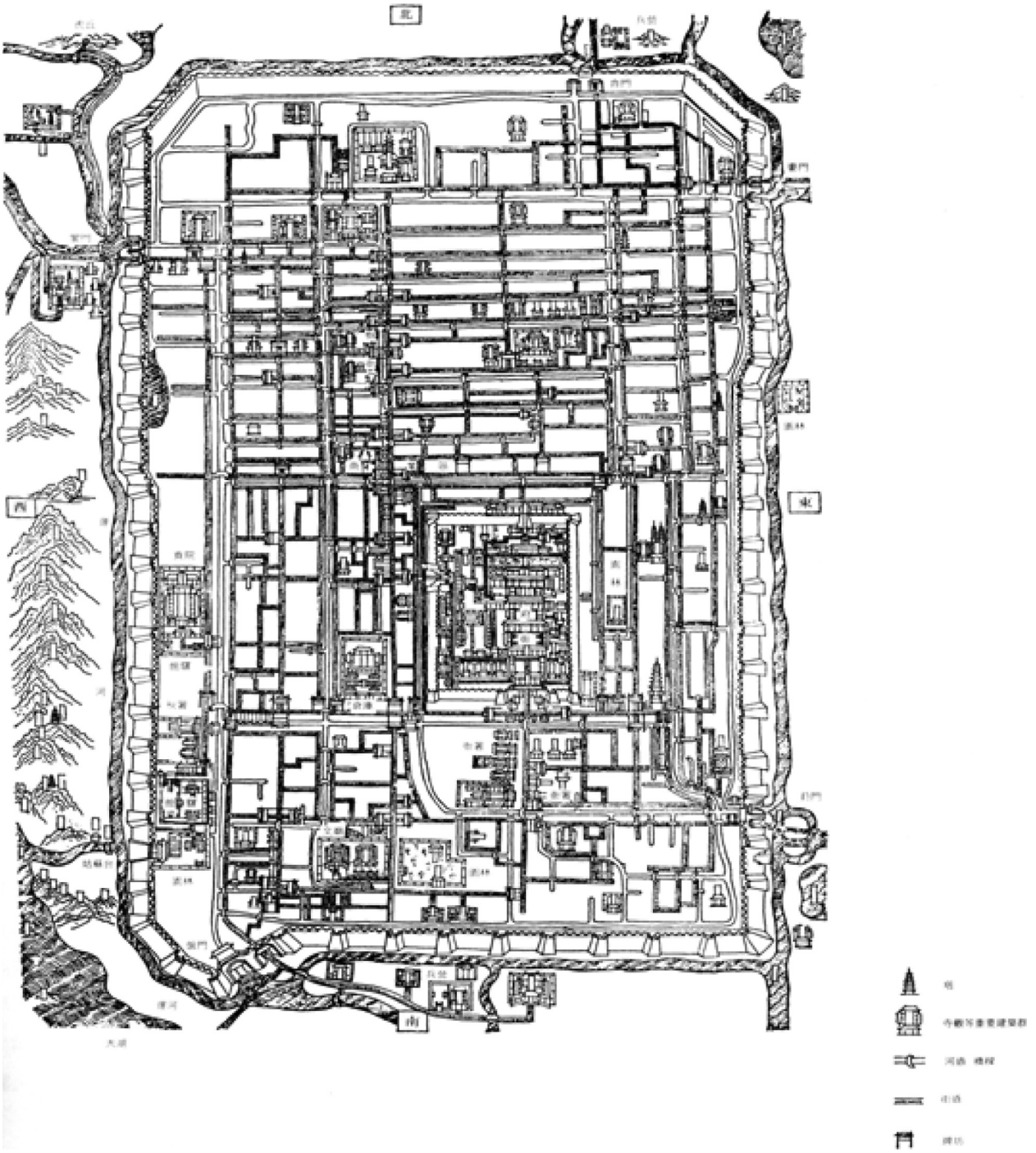

Figure 4 A planned informal city: Ping-jiang (Suzhou) is a regional commercial hub in the Song Dynasty, AD 960. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

Urban Planning in China's Modern Times (Middle Qing Dynasty, 1840, to Republic of China in the Mainland, 1949)

As China was forced to open to the Western countries after the 1842 China–Britain Treaty, urban planning and city construction had two types of practices: in traditional Chinese cities and in cities with foreign concessions.

Traditional Chinese Cities

In cities where there were no colonial zones or foreign concessions, traditional planning principles were still honored. Most of those cities kept the traditional space order, and their townscape remained unchanged for hundreds of years due to lack of investment. In a few cities with new projects, planning activities were limited to infrastructure improvement and new

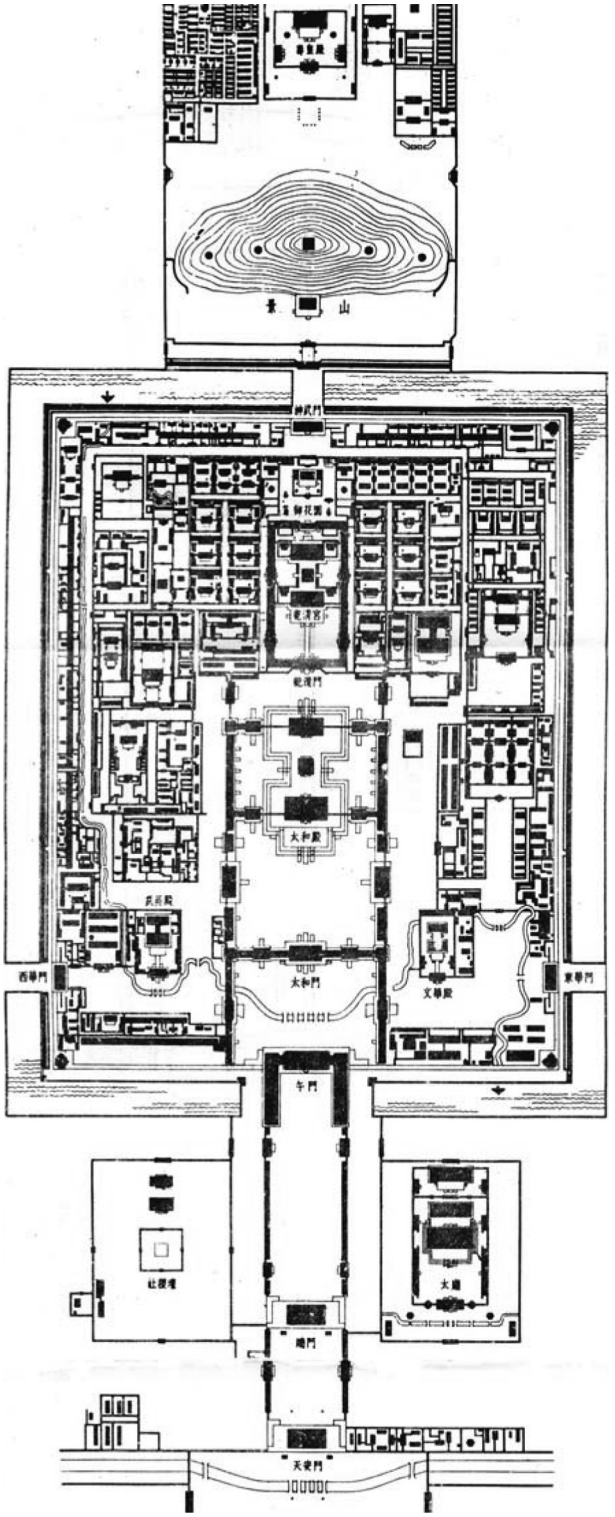

Figure 5 Central Beijing, Ming Dynasty, AD 1368, a well-defined space order and many functional zones. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

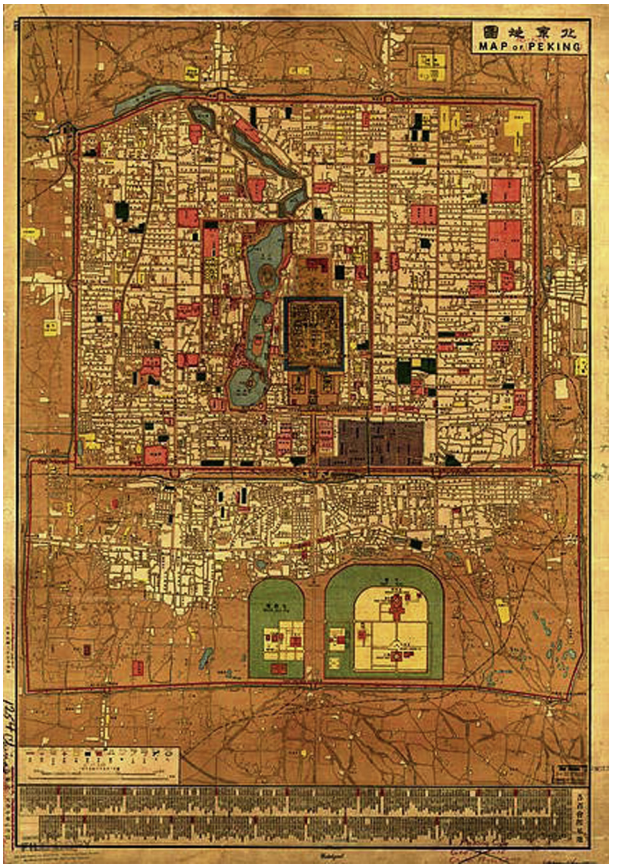

Figure 6 Map of Beijing, 1914, still maintained the original spatial structure. Reproduced from *Wikipedia,* originally by the Deutsch Ostasiatisches Expeditions Korps.

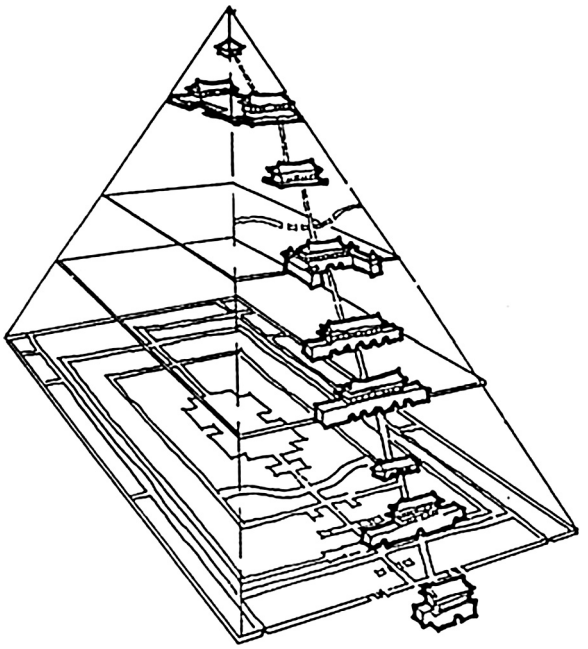

Figure 7 A strictly defined hierarchical space order in the forbidden City, central Beijing. Reproduced from Peng, G., 2001. History of Chinese Architecture. China Architecture and Building Press.

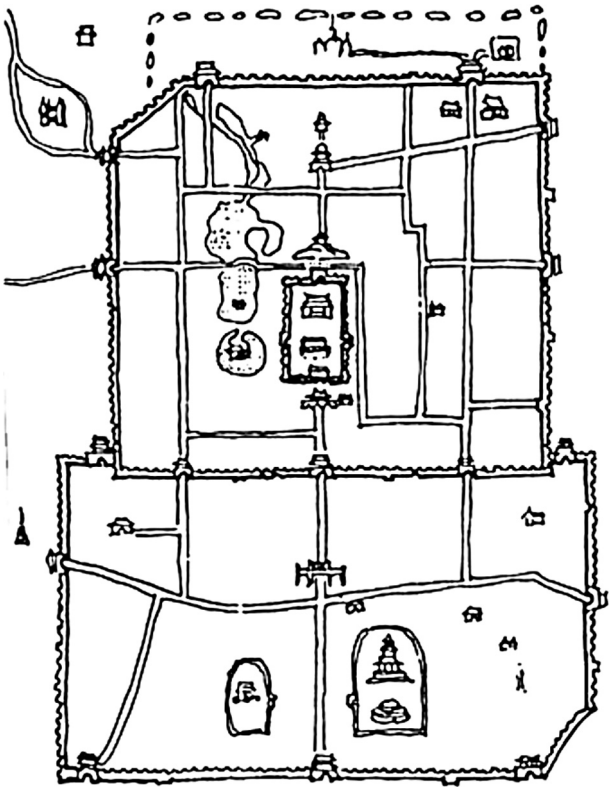

Figure 8 Paths, edges, districts, nodes, and landmarks in Beijing: The five elements of urban form described in Kevin Lynch, 1963. *The Image of the City*, Cambridge: M.I.T. press.

administration zones. Nanjing, the capital of the Republic of China in the late 1920s, was planned following the traditional Chinese space order and building form, even though the plan was made by planners and architects educated in the West (Figure 9). In general, not much change had occurred to the cityscape of traditional Chinese cities in the period from the 1840s to the 1940s, 100 years of chaos with endless wars.

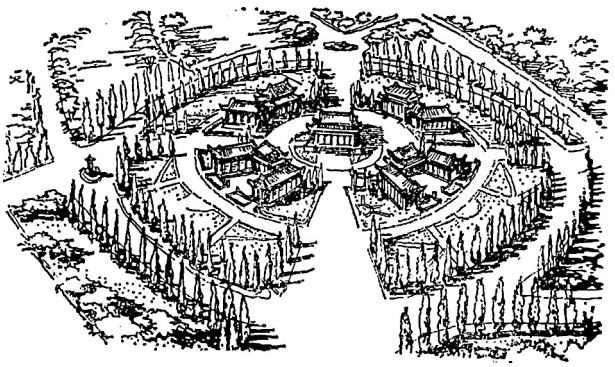

Figure 9 A plan for the National Administration Center in Nanjing, 1929. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

Cities with Foreign-Occupied Concessions

After the Opium War of 1840, China was forced to open to the Western nations; concessions and colonial zones occupied and managed by western forces emerged in a number of cities along the east coast regions. The cityscape in these cities showed a coexistence feature of Western-style districts and traditional Chinese neighborhoods, similar to other colonial cities. Shanghai, the city opened to the West in the 1840s, had two urban areas: a Chinatown under the control of the Chinese government and several concessions managed by representatives of Western countries (Figure 10). The walled Chinatown kept a traditional Chinese form with an administration core in the center; outside the Chinatown, there were the British, French, and American concessions. These concessions were planned by Western civil engineers, surveyors, and architects with a Western space pattern different from the Chinese tradition. Through the planning of concessions, modern planning practice was introduced to China, especially planning regulations and the construction, maintenance, and management of infrastructure systems.

Chinese planners who graduated from West planning, engineering, and architecture programs participated in making Shanghai's master plans. From 1946 to 1949, three sets of master plans were completed. European planning theories and practice such as satellite towns and green belt system were applied (Figure 11).

Several Chinese cities were completely occupied by foreign forces as defined in treaties, such as Dalian in Liaoning Province and Qingdao in Shandong Province in the early 1900s. These cities were planned by architects and planners of foreign countries, so the plans reflected the culture and planning practice of the occupiers. The plan of Dalian occupied by Russia in 1900s, for instance, displayed a boulevard and plaza system common in European nations then (Figure 12). This pattern is quite different to the Chinese tradition; however, some Chinese rulers viewed the Western cityscape a symbol of 'modernization.'

Urban Planning in the People's Republic of China (1949 to Present)

The Soviet Model and the Anti-City Movement (1949–1978)

As China became a socialist country in 1949, a close relationship to the Soviet Union was established. Since then, planning in the People's Republic of China had a new meaning – not only in the physical development realm but also in economic and social development terms. The National Planning Committee and its local branches were in charge of all economic and physical development plans and projects through the Five Year Plan system. Every five years, a new Five Year Plan was issued by the National Planning Committee in which infrastructural, industrial, agricultural, and housing plans and projects from the national and regional to the local levels for the five years were described and arranged.

The Soviet-style planning system functioned for over 30 years until the 1978 reform era. Under the planned economy, the urban planning bureau in a municipality became an institutional tool to ensure the implementation of the national Five Year Plan and the municipality's local plans.

Figure 10 Map of Shanghai, 1915, showing areas under different administrations with a walled Chinatown at the south. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

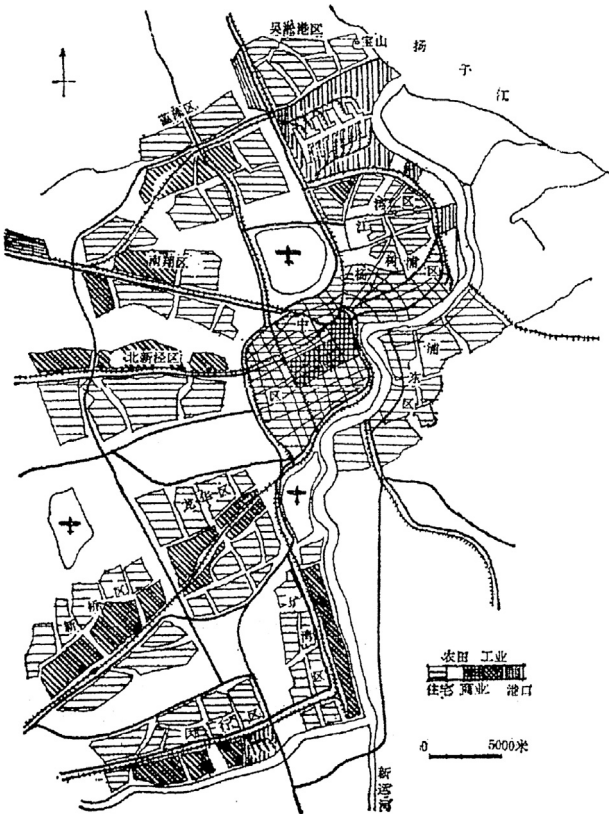

Figure 11 Shanghai Master Plan, the first edition, 1946. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

There was little space left for local initiatives for economic and urban development. The planning system contributed to China's industrialization and urbanization with both positive and negative consequences. It helped the establishment of China's industrial base at the cost of low efficiency and a low urbanization level.

From 1959 to 1962 when a recession occurred, an antiurban policy was adopted, and a movement of forced urban–rural relocation to reduce pressure on urban services was promoted.

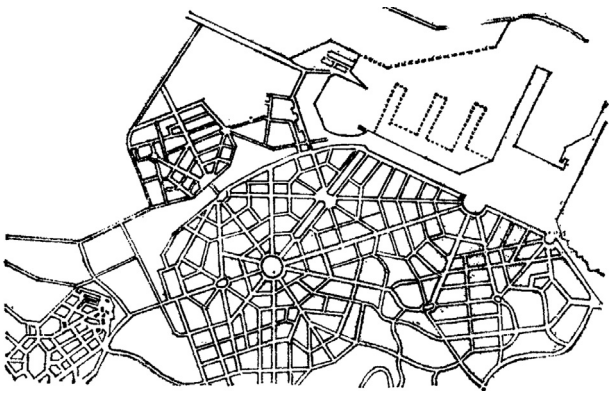

Figure 12 Plan for Dalian, the city center, 1900, by the Russian administration. Reproduced from Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

The antiurban policy had its ideological roots since the Chinese communist revolution was essentially a peasant movement, and many leaders of the revolution had bias against cities.

Another controversial policy in the prereform age was the resident registration (Houkou) system that requires all urban residents to register at urban police stations, put in place in the 1950s aiming to control rural–urban migration and keep China's rural population in farm fields. The antiurban policy was at its peak during the political movement of the Culture Revolution from 1966 to 1976. Over 16 million urban residents, mostly young students, were relocated to the countryside, which caused a considerable reduction in the urban population in the period. In 1965, China's urbanization level (the percentage of registered urban residents in total population) reached 18%. However, 2 years later by the end of 1966, the first year of the Culture Revolution, the urbanization level dropped to 17.3%. Urban planning was viewed a useless or even harmful action, so all planning departments in cities were removed and all planning programs in universities were closed.

Since the 1978 reform, promoting urbanization by allowing rural–urban migration has become a national strategy aiming to stimulate economic growth. The 2010 census reports that 49.68% of China's total population was living in urban areas in 2010. The rapid increase of urban population means a huge demand for urban planning service in the postreform era.

Reform and Rapid Urbanization since the 1980s

The reform era beginning in 1978 opened a new page in China's history. The reforms constituted a fundamental paradigm shift from an ideology-led to an economy-led society and government. In general, the reform consists of two stages: reforming the agriculture sector in rural areas from 1978 to early 1980s followed by reforms in manufacturing and trading sectors in cities in the 1980–90s. With the implementation of the urban reforms, three main policies were formulated: introducing a market mechanism to replace the planned system in most economic realms, decentralizing decision-making power regarding urban development issues from the central to local government, and establishing urban land and housing markets. China's constitution has been revised several times to reflect the new foundation of the nation – a free market economically with an authoritarian government politically. For example, the 1988 constitution separates urban land ownership (state owned) from land-use right (a commodity), which lays the foundation for the property law in which a partial property right is created and protected.

The new policies ended much of the central planning authority in urban development issues with carrying out the decentralization oriented reform. Urban planning has become a local stake in negotiating with the central government in sharing revenue and making development decisions. According to China's Urban Plan Act of 1990, the master plan of a city, which is mandated by the Act, has to be approved by the provincial planning committee, and master plans of key cities such as Beijing, Shanghai, and Guangzhou have to be approved by the State Council. The central and provincial governments thus maintain their impacts over important development decisions via reviewing and approving master plans. At the city

level, plans not required for approval by the Act such as strategic plans, are made on behalf of local interest and used to guide local growth. The new institutional function of urban planning has brought a 'golden age' to the planning profession in China.

By introducing market forces – domestic and foreign investments as well as entrepreneur ideas – the reform created China's economic miracle and urban boom since the 1980s. The 1978 'Open door to outside world' reform was implemented at the early stage of the globalization era. It was historically a perfect timing. As Western investors and cross-national companies were looking for growth opportunities globally, an open China provided the ideal place due to its vast market potential, cheap labor, inexpensive land and energy, fewer regulations on environmental and labor protection, and, more importantly, a market-friendly government providing almost free infrastructure and favorable tax regulations. As the world's second largest economy, mainland China's gross domestic product reached $6717.2 billion in 2010 and its foreign reserve was $3.18 trillion by the end of 2011. China's economy had grown continuously at an average annual rate of nearly 10% from 1990 to 2010. Economic growth stimulates China's urbanization, and urbanization fosters economic growth. The strong economic performance lays the foundation for city construction and urban development. Urban planning has made its contribution to the economic miracle in fields of land acquisition and preparation, infrastructure development, and place making and selling at the global market.

China's urban population (831 million in 2008) is 2.7 times the total population of the United States (304 million in 2008) (Table 1). In 2007, China had two cities (Shanghai and Beijing) among the 20 largest cities in the world; the number will increase to four (the two already mentioned plus Guangzhou and Shenzhen) in 2025, according to the United Nations-Habitat predictions.

The urbanization process over the past 60 years has been especially fluid, although the main trend is a continuous urban population growth (Figure 13 and Table 2). Urbanization rates increased rapidly especially since the 1990s when China was becoming the 'world's manufacturing plant,' a by-product of the globalization process. Urban population increased by 18.2 million annually from 1995 to the 2000s, and the urbanization level reached 51.3% in 2011, the highest in China's history. The growth of the urban population has created a huge demand for new urban services on the urbanized land, which in turn has created an enormous demand for urban planners.

Table 1 Basic information about China's urbanization (2008)

Urban population			
Urban residents	*Migrants*	*Total population in urban area*	*Number of cities*
606 million (45.7% of total population)	225 million	831 million	655[a]

[a]Four state-administrated cities (Beijing, Shanghai, Tianjin, and Chongqing; a status similar to Washington DC in the United States) plus 651 other cities.
Reproduced from CPOPIN (China Population Information Network), 2009. Reports on the Web of National Bureau of Statistic of China, www.stats.gov.cn.

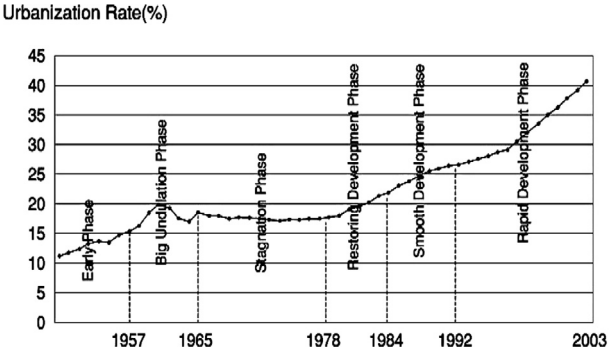

Figure 13 Trajectory of China's urbanization. Based on the National Statistics Bureau of China, 2004 (urban population refers to 'officially registered urban population' excluding migrants).

Given the increasing demand for urban housing, services, and jobs, urbanization has been officially employed as a means to promote economic growth. In a national debate in the 1990s about 'the importance of urbanization to China's economy,' supporters argued that urbanization had fallen behind China's industrialization and economic development so it should be encouraged. At that time, the central government adopted the policy of 'speeding up urbanization to stimulate economic growth' by promoting urbanization at a rate of 1.5–2% annually. In practice, expanding existing cities by creating new urban districts to absorb new factories and employees (migrants as majority) and reducing the amount of peasants by merging rural townships and villages ('che xiang bing cun') were the main strategies. From 1995 to 2002, the average urbanization level increased by 1.44% annually. Some Chinese scholars warn that given China's huge population baseline, the urbanization growth rate is too high to be sustainable. They suggest that an annual growth rate of 0.8%, or at most 1% will be more appropriate and rational.

The postreform urbanization has been facilitated not only by the government but also by the marketplace. Foreign and domestic investment in real estate and manufacturing stimulated the demand of urban land and urban labor. Most urban expansion or the so-called 'Chinese version of sprawl' has been a result of the establishment of suburban 'Economic Development Zones' (EDZs) where local government attracts foreign direct investment by providing low-priced land and free infrastructure. Researchers reported that 34–46% of urban land expansion was caused by the government-led land acquisition in EDZ. The immature land and housing market often fueled irrational land acquisition via large-scale land-use conversion from agriculture to urban uses. Urban planners also played a role in implementing the growth-oriented policy.

Stimulating Economic Growth in the Globalization Era through 'Place Making and Place Selling': The Task for China's Urban Planning Today

With the establishment of China's leading position in the world economy, both China's central government and several

Table 2 China's urbanization level[a]

Year	*1951*	*1965*	*1975*	*1978*	*1995*	*2000*	*2002*	*2008*	*2011*
Urbanization	11.8	18.0	17.3	17.9	29.0	36.2	39.1	45.7	51.3

[a]In percentage, in selected years.
Reproduced from www.stats.gov.cn (urban population refers to 'officially registered urban population and migrants lived in urban areas for over 6 months').

city governments seek for the next development goal: stimulating economic growth and promoting Beijing, Shanghai, and several other well-developed Chinese cities to be international or global cities. Urban planning is expected to carry out the new task. The main approach of making an international city is through 'big international events' such as the 2008 Olympic Games in Beijing, the 2010 World Expo in Shanghai, and the 2011 Asian Games in Guangzhou.

In preparing for the 2008 Olympic Games, the Beijing municipal government under the leadership and financial support of the central government made and completed a bunch of plans including infrastructure development projects (building new and upgrading existing airports, subways, and expressways), constructing Olympic Games facilities (total of 31 sport fields and halls) and public buildings (the National Performing Art Center the China Central Television (CCTV) compound, and new museums and libraries), renovating historic buildings as well as preserved historic blocks, and building new commercial and office campus and housing projects (Figure 14). In Shanghai, similar strategies were adopted in preparing for the 2010 World Expo. Many innovative planning concepts were generated and implemented, including the application of ecofriendly technology (water and solid waste recycle and use of solar energy), reuse of old factory buildings for the Expo, integrating Expo facilities into city life after the Expo, and infrastructure improvement projects (Figures 15 and 16). Guangzhou followed the same development model in welcoming the 2011 Asian Games. In all cases, planners made significant contributions to these successful events and to promoting cities to be international hubs (Figure 17).

China's Urban Planning Profession and Planning Education Today (in the 2000s)

The Urban Planning Profession

From the 1930s to the 1940s, China had only a few planners; most of them were educated in the West, and the majority graduated from architecture or engineering schools. The old-generation planners laid the foundation for China's planning profession and planning education. Some of them participated in making master plans for Shanghai, Nanjing, and other key Chinese cities.

In the 1950s, before the divorce of China and the Soviet Union in 1960, a number of young planners got their planning training in Soviet universities. These planners played a critical role in China's early industrialization and urbanization under the planned economy. Their impacts on planning theory and practice, a mix of the classic European physical planning

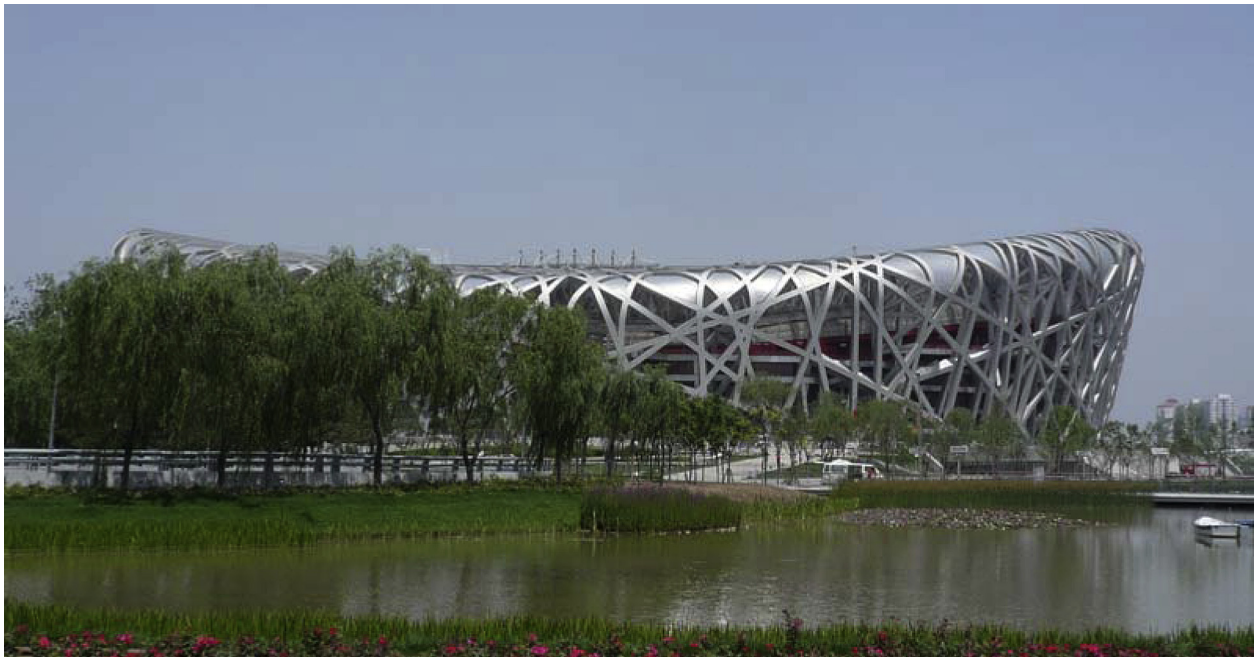

Figure 14 The National stadium for the 2008 Olympic Games in Beijing. Photo by the author.

Figure 15 Master plan of the 2010 World Expo, Shanghai. Reproduced from Zhou, Z., 2010. Illustrated Architectural Guide to Expo 2010 Shanghai China, Time + Architecture, 2010–3.

principle and a centralized decision-making system, can still be found.

Today, over 110 000 planners of various levels are serving on city planning committees, working in planning bureaus and planning institutes. China has more urban planners than any other country. However, the number of planners per 1000 residents in China (0.008 planners for every 1000 residents) is slightly less than that of the United States (0.11 planners for every 1000 residents, total of about 33 000 American planners in 2010).

China established its certified planner system in 2000 with the help of the American Institute of Certified Planners (AICP), Royal Town Planning Institute, and Hong Kong Institute of Planning. Up to 2010, a total of 97 831 planners have taken the certification examination, and 14 928 have been awarded the 'National Certified Planner' status (CSUS and UPSC, 2010, 2011). The rate of passing the examination is much lower than that of the AICP examination.

New-generation Chinese planners are much better armed with planning knowledge and skills than ever before. Most of them are good at physical planning, urban design, geographical information systems, computer-assisted design, computer-generated modeling, and other software applications, but their training in the social science and public policy aspects of planning is still limited. In part this situation can be attributed to the demands on Chinese planners. Urban planning is still viewed as a profession handling infrastructure, city beautification, housing, etc. Social issues are not fully explored in the domain for planners. The situation is also due to China's developmental stage: to make a bigger pie is more important than to share the pie fairly. To a certain degree, underemphasis on social developments also is deeply rooted in Chinese history.

China's Urban Planning Education

Several Chinese universities offered urban planning courses in the 1930s, but none had independent planning programs; rather, most of the courses were offered in architecture or civil engineering departments, such as Tongji University and

Figure 16 Shanghai Pudong District. Photo by the author.

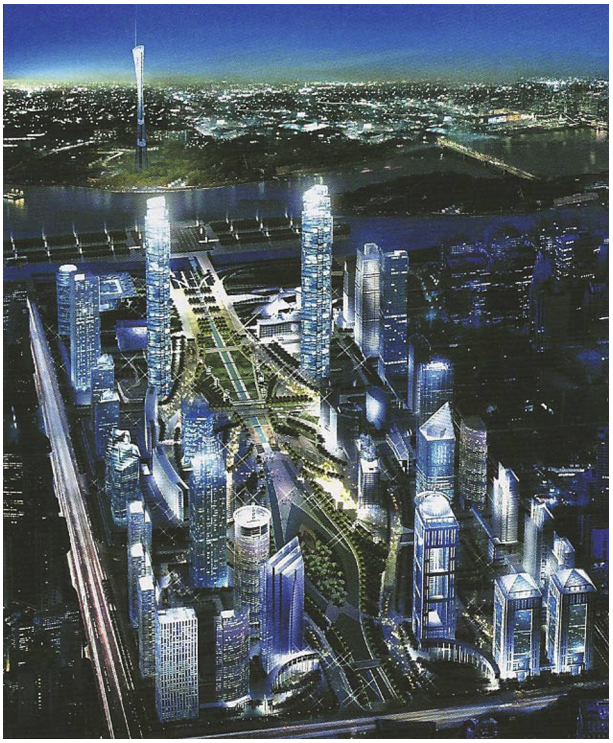

Figure 17 Bird's eye view of Guangzhou Zhujiang Newtown, 2010. Reproduced from Guangzhou City Planning Bureau, 2011. Urban Design in Guangzhou, China Architecture and Building Press.

Jiaotong University in Shanghai and Zhongyang University in Nanjing.

As China entered its industrialization period in the early 1950s, two universities, Tsinghua University in Beijing and Tongji University in Shanghai, established independent planning concentrations in their architecture departments. Five other universities followed this model soon after in the late 1950s. The seven urban planning programs educated Chinese planners for over 20 years in the prereform era. All of them were physical planning oriented, with limited coverage of social and economic development courses. In the 1980s, the first geography-based urban planning program was founded in Hangzhou University. Urban studies and regional economic development planning are the main topics for students studying planning in geography departments.

As more rural residents migrate to urban areas and the number of cities increases rapidly, the demand for urban planners is increasing considerably. The huge demand stimulates the growth of planning education programs. Since the 1990s, planning education programs have been established in almost all big Chinese cities. The number of planning programs increased from 28 in 1995 to 269 in 2011, with a total of 37 000 planning students enrolled, and 20 000 undergraduates, 4000 masters, and 300 PhDs graduating annually in recent years (CSUS and UPSC, 2011). In North America, the United States and Canada together had 99 planning educational programs in 2010. Thus, China becomes the country with the largest planning education capacity in the world (Figure 18).

China's urban planning programs have diverse backgrounds. Although most city planning departments are still within architecture or engineering colleges, about 35–40% of planning programs are housed in departments of geography,

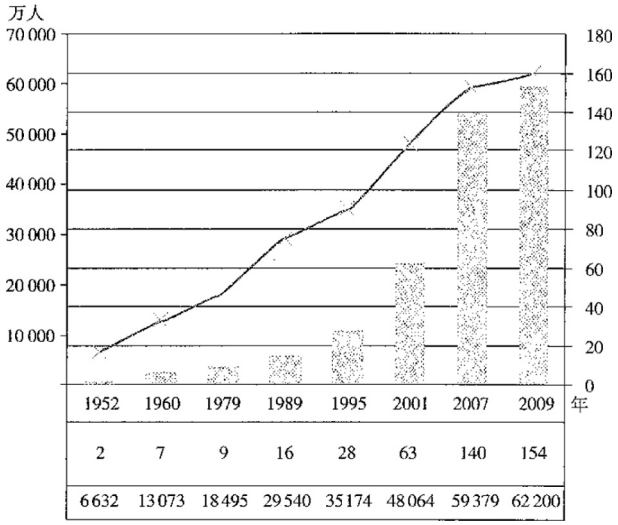

Figure 18 Numbers of urban planning programs in Chinese universities vs. numbers of China's urban population (in 10 000) from 1952 to 2009. Reproduced from Chinese Society for Urban Studies (CSUS) and Urban Planning Society of China (UPSC), 2010 and 2011. China's Urban Planning, Annual Report 2010, 2011, China Architecture and Building Press.

and one in the college of public administration (in Renmin University in Beijing). Because most demand is in physical development given China's rapid urbanization and huge urban population baseline, students with physical planning and urban design training are the majority and enjoy better job opportunities. Students of planning programs with urban studies background in geography departments do more work in research institutions and government planning agencies, but their job opportunities are less bright than that of design-oriented graduates. This fact reflects the nature of China's urban planning today: urban planning is more involved in physical development projects, and social development does not yet receive that much attention, although community planning is a growing area where many young planners are engaged with enthusiasm.

While the quantity of planning programs has been growing significantly, the quality of some programs is problematic. The Planning Accreditation Committee of China has only assessed and accredited 10% of all planning programs so far. It warned many newly established planning programs that they should take the quality issue more seriously since many of them are simply profit driven (CSUS and UPSC, 2010, 2011).

A positive development in terms of educational quality is that Chinese planning programs have been actively searching for partners in Western planning educational departments in order to exchange teaching experience and improve quality with the help of senior professors from well-known Western universities. Top Western and ethnic Chinese planning scholars from American and European universities are frequently invited to lecture in Chinese planning schools. This could be viewed as an important benefit from increasing international exchange in the globalization era.

Bibliography

CPOPIN (China Population Information Network) 2009, reports on the Web of National Bureau of Statistic of China, www.stats.gov.cn.

Chinese Society for Urban Studies (CSUS), Urban Planning Society of China (UPSC), 2010. China's Urban Planning, Annual Report 2010. China Architecture and Building Press.

Chinese Society for Urban Studies (CSUS), Urban Planning Society of China (UPSC), 2011. China's Urban Planning, Annual Report 2011. China Architecture and Building Press.

Dong, J., 2004. History of City Building in China. China Architecture and Building Press.

Guangzhou City Planning Bureau, 2011. Urban Design in Guangzhou. China Architecture and Building Press.

Lynch, K., 1963. The Image of the City. M.I.T. Press, Cambridge.

National Bureau of Statistic of China, various years, www.stats.gov.cn.

Peng, G., 2001. History of Chinese Architecture. China Architecture and Building Press.

Zhou, Z., 2010. Illustrated Architectural Guide to Expo 2010 Shanghai China. Time + Architecture, 2010-3. Tongji University Press.

Urban Planning, Competitions and Exhibitions

Carola Hein, History of Architecture and Urban Planning, TU Delft, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Public and private institutions or stakeholders use exhibitions and competitions as means of urban planning to stimulate the production and exchange of ideas and as tools to reshape urban form as well as governance. *Competitions* aim to elicit ideas from various people – mostly professionals, but occasionally also the general public – on a defined topic, including urban plans. *Exhibitions*, both educational and exemplary, provide spaces respectively for the comparison of architectural and urban representations or actual construction. Both methods have long histories of changing the goals and themes that have evolved with the shifting definition of the profession.

Urban planning – defined as a complex and multifaceted attempt to adapt the form and function of a city to the multiple needs of its various stakeholders and as a tool to effect political, social, and physical change – deploys a broad range of methods to reshape urban form, including *competitions* and *exhibitions*. Both tools permit the comparison of projects and built designs from a variety of contributors, who often come from other cultures and different places. They aim to enhance good design and to prompt public discussion and awareness, even consensus. Urban planning or urban design competitions have a long history that is rooted notably in the French architectural education and German design tradition. Since the turn of the nineteenth century, urban competitions and (educational as well as exemplary) exhibitions helped define the emerging profession. Competitions and exhibitions also give visibility to a specific city and help brand a location; public and private organizers around the world thus consider competitions and exhibitions to be important tools in positioning cities and regions in the global marketplace. Both originally emerged in Europe and have spread throughout the world, where various parties have tested them in different cultural contexts. Their focus has also evolved, from being primarily focused on aesthetics and top-down approaches to addressing functional concerns or issues such as environmental sustainability, and they have come to include the bottom-up involvement of citizens (for general literature, see Speiregen, 1979; Omi, 1986; Haan et al., 1988; Becker, 1992; Jong and Mattie, 1994; Aymone, 2007).

Competitions

A planning competition is only one of several options available to an organizer when choosing an urban plan and a planner. Others include directly commissioning an individual to come up with a plan, holding planning workshops or seminars, and appointing a planning committee or a consultant to make the decision. If the topic and its constraints are clearly defined so that the participants can make meaningful statements within a limited period of time, and if the political and administrative contexts and the economic situation permit consequent implementation or adaptation of the results, competitions are good ways to find solutions. Competitions allow the sponsor to select from a variety of proposals, even if the organizer decides not to implement the proposal as such or to use only some ideas, as the competition may help to stimulate further insights into the planning problem and ideas for responding to it.

Such competitions can take several forms. They can be open to the general public, restricted to members of the relevant professional institute, or restricted even further to a limited number of participants, whether appointed by the organizer or chosen after an application process. (These procedures may also follow upon each other, and restricted competitions can include payment to participants.) Restricted competitions have been criticized for being limited to famous names or big companies. In contrast, critics note, anonymous and open competitions allow unknown and young planners to participate and present their ideas, and may result in new concepts. This can, however, mean large numbers of participants, or lead to the choice of an author who lacks the means and experience to realize the proposal. As only prizewinners are paid, much unpaid work is done, particularly in the case of complex urban competitions, which, therefore, are sometimes considered to be unproductive. Nonetheless, even unbuilt contributions are part of the architectural culture and history of a city and a period. In some cases, unrealized competition projects have had a stronger impact on planning history and theory than built ones. Peter and Alison Smithson's entry in the 1958 urban planning idea competition 'Capital Berlin,' featuring a platform over the city, serves as an example: never built, it nonetheless inspired projects and debates internationally.

Cities and nations around the world see competitions as a particularly useful way to include citizens in the decision-making process, and also as a way to justify their planning choices and expenses. In turn, the democratic character of a competition depends primarily on whether its preparation and ensuing realization are part of a public and open discussion. The process of a competition per se is apolitical, and any political system can use one as a purely technical device to identify the most appropriate response, as the Soviet government did in holding an international competition for plans for Greater Moscow in 1931–32. More generally, the outcome of a competition depends on the participants and is strongly influenced by the aims and quality of the competition brief and the organizers. The institutional character of the organizers – whether they are public institutions, private sponsors,

professional organizations, architectural schools, or magazines – and the level of their concrete power over the competition site determine whether the competition leads to realization or is limited to creating ideas, and whether it is supposed to adapt to the existing context or provide a long-term vision for the future.

Preparatory groups – mostly professionals and functionaries – make major decisions on the overall distribution of functions in a city, and on the use, size, and form of an urban plan. They prepare a set of competition documents that can range from primarily descriptive to very precise and technically detailed. These documents might include concept plans, maps, and data on the socioeconomic structure of the population; requirements for educational, housing, leisure, or infrastructure equipment; and general guidelines on urban function, structure, and form. The preparatory documents might end up guiding future development, whether or not the competition produces a winning project that the city implements.

History of Major Large-Scale and City Center Competitions

Architectural competitions have been held for many centuries, particularly for the design of religious and public buildings. Competitions for entire cities or urban areas also have a long history. An early example is the one held after the London fire of 1666, as well as the international idea competitions for urban embellishment opened by the Russian empress Catherine II for the city of Saint Petersburg (1763) or the contests for the Place Louis XV (today's Place de la Concorde) in Paris (1748).

Urban planning was usually seen as a part of architecture, and the competitions held by professional and educational architectural institutes included urban design projects, such as for a 'Seaside Resort for a Sovereign on the Mediterranean Coast,' a topic at the Paris Ecole des Beaux Arts in 1872. Educational institutions held student competitions for grading and awards such as the 'Rome prize,' which included travel and a several-year stay in Rome. These projects were not intended to be built, but nevertheless reflected the design spirit of their time. The 'Rome prize' winners were assured an honored part in public building projects on their return. With their long history of academic competitions, French planners won late early twentieth-century international competitions for the enlargement of Barcelona (1903), Bucharest (1906), Guayaquil (Ecuador) (1907), Antwerp (1910), Yass-Canberra (1911) (Australia), and other cities in North and South America as well as in France itself (Risler, 1915).

The history of urban competitions reflects the development of the planning profession: following industrialization and rapid urban growth, early competitions tackled large areas on the outskirts of a city or even a whole city. With the growing scale and complexity of cities, their parts and compounds became the main topic of later competitions. In spite of the large number of competitions held over the last century, some of the earliest major transformations and extensions of cities, including the work of Eugène Haussmann in Paris (1850s–70s) and the extension of Berlin under James Hobrecht (1862), were based on direct commissions or procedures other than competitions.

A comprehensive and international analysis of urban planning competitions is still lacking, and the number of competitions is too large to be analyzed here: analysis of the late 1920s, for example, shows between 15 and 30 urban planning competitions annually in Germany alone (Becker, 1992). But a brief genealogy of major examples, presented here in chronological order, shows overall trends.

One of the most interesting examples of nineteenth-century urban planning, the transformation of the Vienna fortification zone, did result from an international urban competition in 1857–58 (Albers, 1997; Hall, 1997); it requested ideas for traffic planning, plot division, and urban beauty for a zone both sides of which were already built. None of the prize-winning or recommended designs was realized, but the city used them as references for a basic urban plan. In 1892, Vienna was once again the site of a major urban competition; this one, like the competition for Munich in the same year, invited ideas for a master plan of the whole city. These two competitions can be seen as milestones marking the early development of urban planning – intimately connected to the works of major urban planning theoreticians of that time, such as Reinhard Baumeister, Camillo Sitte, and Josef Stübben, whose theoretical work inspired the implementation of the later master plan for Vienna (Herscher, 2003).

Some famous urban designs included competitions as one part of their complex history. The mid-nineteenth-century extension plan for Barcelona resulted from conflict between the liberal central government and the conservative communal government. Initially, several plans were on the table, including a large grid expansion by Ildefons Cerdà. But the commune supported a small-scale project, and sponsored a competition in 1859. Antoni Rovira i Trias won with a proposal for a star-shaped extension of the city's historic core. The national government rejected this proposal and reverted to Cerdà's plan.

A milestone in urban planning history was the Berlin competition of 1908–10. Initiated by architects' associations and organized by several independent municipalities, it concerned a large urbanized area of more than 5 million inhabitants. Although the two winning proposals were not realized as such, the process helped clarify major problems and led to several competitions for smaller parts of the city. Through exhibitions, publications in magazines, and private brochures, the city disseminated the competition formula and its results. They became important references for other cities and planners, stimulating other large-scale competitions in major and minor European cities in the late nineteenth and early twentieth centuries, including Gothenburg (1861), Strasbourg (1877), Krakow (1910), Reval (1913), Dublin (1917), Paris (1919), Budapest (1936), and Zagreb (1929) (Albers, 1997). Some planners, including the Frenchman Léon Jaussely, participated in several of them (Albers, 1997).

In Europe, early competitions concentrated on remodeling and enlarging existing cities. Meanwhile, new nations around the world sought cultural representation, and they planned and constructed entirely new capital cities. Competitions held for Canberra in 1911–12 (winner: Walter Burley Griffin) and for Brasilia in 1956–57 allowed Australia and Brazil, respectively, to obtain a state-of-the-art urban plan ready to impress the nation and foreign countries. The winning project for Brasilia's master plan, by Lucio Costa, was successful in not only laying out a comprehensive plan but also establishing a foundation myth for the new capital (Holston, 1989). Following another international competition in 1977, an

American team, International Planning Associates, won the prize of developing the master plan for the new federal capital of Nigeria, Abuja. Such competitions gave government authorities the opportunity to consult with leading planners and, as in the case of Canberra and Abuja, eventually appoint foreigners to realize the plan.

Even in densely built Europe, governments held competitions for new towns. Italy, under Benito Mussolini, reclaimed marshland and then sponsored a series of competitions to build new cities on it, including Sabaudia; the regime thus demonstrated its capacity to provide housing for its growing urban population. Great Britain and Germany held competitions for new towns, designed according to modernist planning principles in the post–World War II period to create new population centers. Examples include Aarne Ervi's winning proposal for the center of the new town of Tapiola in Espoo City, Finland, announced in 1953; a competition for the central area of the new town of Ashdod, Israel, in 1965; and Richard Llewelyn-Davies's prize-winning entry to the 1967 competition for Milton Keynes, a new town in northern England, with a decentralized, car-friendly, low-density, and flexible design.

European institutions, regions, cities, and architectural schools sponsored city competitions and projects for new cities, all in a quest to build a capital for newly unifying Europe; but these competitions were, generally, flawed by the sponsors' lack of political and financial power, the dominance of nationalism, as well as the unclear future of the new European communities. Furthermore, such competitions demonstrated the difficulty of defining Europe. In the competition sponsored in 1954 by the Saarland – a region that considered a European future to be a welcome escape from both integration into France and the impossibility of a return to Germany – the jury selected three winning projects, so different from each other that none could represent a common Europe (Hein, 2004).

Reconstruction after World War II stimulated urban competitions of various scales in several countries, leading to intense and often controversial discussions. The German city of Nürnberg made a great effort to involve citizens, opening one idea competition to all citizens (yielding some 1192 proposals) and another one to planning professionals. The latter competition proved so successful that it also sparked the German Building Exhibition of 1949, which attracted some 300 000 visitors and helped rebrand the city of the Nazi Party gatherings and of the Nürnberg trials as a modern and democratic center (Diefenbacher and Henkel, 2009). Meanwhile, the most political of the large-scale competitions was probably the international urban idea competition for 'Capital Berlin' of 1957–58 (Hein, 1991), organized by the West German and West Berlin governments, which targeted an area in the city center that was largely under East German administration – a provocation to which the East German government reacted with a competition of its own.

In contrast, the competition for rebuilding the atom-bombed center of Hiroshima was a unique case among the more than 200 cities that had been bombed in the war. Japan had no previous experience with warfare on its land, with postwar rebuilding, or even with competitions. A confluence of several elements in 1948 led to the unusual building of Hiroshima Peace Park: international interest, a special law providing supplementary funding (the Peace City Law), a comprehensive project chosen by competition, and a sensitive architect. The government had dispatched Kenzô Tange in 1947 to draw the land use plan for Hiroshima; his prizewinning project for the Peace Park established connections with the immediate surroundings and even projected a network of related facilities for other areas of the city. Reaching beyond the Nakajima neighborhood, it visually integrated the A-Bomb Dome with the remainder of the Hiroshima Industry Promotion Hall (a brick-and-steel building located close to the center of the explosion), while also integrating urban design, architecture, and landscaping (Hein, 2002).

Other large-scale competitions held in the wake of natural disasters highlight the global approach of urban planning in the postwar era. After an earthquake destroyed the Macedonian city-capital Skopje, the Yugoslav government established a group of international consultants who chose an outline plan by Doxiadis Associates and the Polish architect Adolf Ciborowski (who had supervised the rebuilding of Warsaw). In 1965, the country held an international competition for a design to rebuild the city center: the consultants invited four Yugoslav firms, as well Van den Broek en Bakema (Netherlands), Maurice Rotival (United States), Luigi Piccinato (Italy), and the first-prize winner, Kenzô Tange (Japan).

One engine for the global spread of competitions was colonialism: European countries brought the practice of holding competitions to their colonies. These continued the practice even after independence as a sign of modernization and emancipation. One example is the international competition held in 1961 for the piercing of the citadel, the casbah, of the capital city of Tunis. The consequent period of increased globalization has contributed to the growing number of international competitions since the 1970s, as can be seen in the listing of those approved by the Union Internationale des Architectes (UIA). Standouts among these in the early years were student competitions in the field of housing, sponsored by the Institut National du Logement Belge (INL) or by the UIA. By the 1980s, the number of competitions proposed by non-European institutions started to increase, including one in the United Arab Emirates for plans for the University of El Ain (won by SWECO-AS (Sweden)) in 1983 and another in Kuwait for the proposal for a city on the sea, the Pearls of Kuwait, won by P. Cox, Richardson & Taylor (Australia) and Jerde Partnership (United States) in 1989 (www.uia-architectes.org).

By the 1990s, competitions were a common tool around the world for a city or other body to gain international advice and attention. In 1996, Bucharest sponsored an international competition for proposals to transform its center both to host the offices of the capital city and to celebrate the year 2000. Out of 235 projects from 33 countries, the jury invited 15 authors to continue their research in a second phase; it finally awarded the first prize to Meinhard von Gerkan and Joachim Zeiss (Germany). The number of competitions has grown over the years, with the locations sponsoring them spreading out even further around the world and bringing together international juries and international participants, such as the 2003 competition for the Freedom Park Trust in South Africa, where three finalists from various parts of the world were honored (US firm Obra Architects, Tai Fai of Hong Kong, and Vladimir Djurovic Landscape Architecture in collaboration with the architect Imad Gemayel of Lebanon); or the 2004 idea

competition for the Place des Martyrs and the Grand Axis of Beirut, won by a group from Greece.

The 1990s also saw planning competitions on regional scales, often with a focus on infrastructure. In May 1990, the cities of Strasbourg, France, and Kehl, Germany, opened a competition for ideas for a border-crossing district along a proposed new highway. Connecting Strasbourg to the German high-speed road network, the highway would help revive the neglected 300 hectare Fronts de Neudorf area. With urban planning departments from two different nations, the project was European rather than national. The departments tried to coordinate their local planning laws and to create a concrete example of collaboration among regions of different European Union (EU) member states and between former enemies. Their success spurred further cross-border competitions.

As planning doctrines and urban needs shift, so do the foci of competitions. In recent years, environmental sustainability has become a prominent guideline. In 2007, the French president announced a competition on the future of Greater Paris in the next several decades. The theme was that the reduction of greenhouse gases, agreed upon in the Kyoto Treaty, and the larger societal changes needed to preserve energy, should all start in cities. The president invited architects to rethink Paris' urban spaces to develop a more ecological way of life and to integrate the center and the periphery; he imagined a new global development plan that would resolve the tension between rapid and uncontrollable development, on the one hand, and the need to preserve energy and create an ecological equilibrium, on the other. The competition itself and the underlying presumptions have since raised numerous questions and provoked a flurry of responses in the media and among academics. However, in terms of realization, very little has ensued.

Competitions for Limited Areas and Special Functions

Planning an entire city area brings with it numerous difficulties and demands extensive knowledge that cannot be expected from competition participants. Simultaneously with the long and rich history of competitions for large-scale designs for new cities, which was mapped in Section History of Major Large-scale and City Center Competitions, public and private sponsors also opened competitions for limited areas and clearly defined tasks such as campuses, government centers, commercial or recreational centers, and urban parks. Examples can be found worldwide: a privately sponsored competition for the University of California (1897–99), city-held competitions for educational districts in Tokyo (1940s), design competitions for civic centers in the United States, and the urban design competition for the government districts in the Spreebogen and Spreeinsel areas in Berlin (1993–94).

It is likely that the largest number of competitions has been organized for residential layouts. Examples can be found in various countries; examples from Germany are given here for illustration. The country has been a major organizer of competitions over the last century, and their topics and results reflect the main concerns of housing over this period. Early competitions focused on small-city and company housing, such as for the Krupp colony Altenhof in Essen (1892). Metropolitan housing, economic land use, and the creation of a rational block structure became central topics in the competitions of 1911 for Tempelhofer Feld and Schöneberg South that followed the Greater Berlin competition of 1910. The search for an economical solution for residential areas climaxed in competitions for the experimental district Berlin-Haselhorst (1928) and the Dammerstock-Siedlung in Karlsruhe (1929).

The French tradition of student competitions continues to influence contemporary events. The French Programme d'Architecture Nouvelle (PAN) invited students to improve the quality of subsidized housing between 1972 and 1987, awarding the prize winner the opportunity to realize his or her project. Building on this, Europan is a student competition in which municipalities propose some 50–70 strategic sites in each session, asking students to solve spatial problems in urban contexts and to formulate new urban paradigms and sustainable strategies (http://www.europan-europe.eu).

Federal and municipal governments in the United States have also sponsored some notable targeted design competitions. The World Trade Center site in New York saw a design competition in 2002, and the redesign of the Highline in New York is the outcome of a 2004 competition launched by the Friends of the Highline and the City of New York. Recent examples include a competition for a new mixed-income housing model opened by the Chicago Housing Authority (with the City of Chicago and the US National Endowment for the Arts), awarded to Brian Healy Architects of Boston; the Congress for New Urbanism's AuthentiCITY open design competition for West Palm Beach, Florida, in 2012; and the ONE Prize 2012, a competition that asks entrants to explore the regenerative possibilities of urban transformation in the struggling US economy. Several of these events exemplify the opportunity that competitions have to inspire societal change (http://www.dexigner.com/news/25074). Nonetheless, the US sponsors far fewer competitions than other countries.

More recently, China has emerged as a major player in the practice of holding competitions for targeted urban design. In conjunction with the creation of the new Yangshan deep-water port, the Shanghai municipal government held an urban planning competition in 2002 and selected a Hamburg-based design firm, Von Gerkan, Marg und Partner Architects, to design New City for 800 000 people as one of 11 new satellite towns of Shanghai (Gerkan, 2008). Other recent examples include a 2002 competition for an art, cultural, and entertainment district in West Kowloon (awarded to Foster and Partners); the Quanzhou Luoyang River Urban Design Competition (2011); the Pingdongshan Baigui Lake Urban Design Competition (2011) and the Hangzhou Bay New Town Competition (2010). These competitions attract participants from around the world.

Exhibitions

Exhibitions are an important way for cities and other bodies to present urban planning theories, concepts, and realities to specialists and the general public. They can follow on competitions or accompany conferences but can also be events on their own. Their purpose and impact are multifold and have shifted over time. They can serve as a tool to experiment with

and spread new concepts to an international as well as local, professional, and lay public. They may serve as a focus for decision makers and citizens and can also attract funding or help to focus it. Media attention lavished on these events can contribute to a city's brand and promote its position in the urban hierarchy.

Among the multiple forms of exhibitions, two forms can be broadly distinguished: educational, such as the public presentation of historic or contemporary projects, plans, and models for concrete or visionary proposals (often organized in conjunction with specialist conferences) in one locale or as a traveling exhibition; or exemplary, such as the building or rebuilding of an urban area with temporary or permanent buildings, as a specific urban design project or as part of another event such as a world's fair. Both are part of the planning process of many countries that aim to promote democratic participation, and they can go hand in hand. Both can also be primarily cultural events, displaying the history, present, or future of planning and helping to promote an area economically. And both have increased over the last decades together with the proliferation of visionary and unbuildable projects and the creation of architectural institutions. The following proposes a chronological overview of select examples.

Major educational exhibitions focusing specifically on urban planning have been held since the turn of the nineteenth century, reflecting ongoing debates in the emerging profession. In 1903, Dresden was the site of the first German Municipal Exhibition (Erste Deutsche Städte-Ausstellung), which showcased the achievements in engineering and urban planning of several German cities. In 1911, an exhibition on hygiene was held in the same city, and included a large display on urban planning. The theme of overcrowding was at the center of the tenement exhibitions organized by Lawrence Veiller in 1899 and 1900 in New York. City planning became a focus of the work of Benjamin Marsh; his 1908 exhibition 'Congestion of Population in New York' led to his 1909 publication *An Introduction to City Planning*, as well as municipal activity (Commission on Congestion of Population (1911); Veiller, 1900–1901). Other exhibitions were held to promote new urban plans, such as the McMillan Plan for Washington in 1902 or Daniel Burnham's plan for Chicago sponsored by the Commercial Club of Chicago in 1909. Yet others occurred in parallel with congresses of the International Garden City and Town Planning Association or with international architectural congresses, such as that of London in 1906, which displayed works of major planners, including Raymond Unwin, Josef Stübben, Charles Buls, and Eugène Hénard (Meller, 1995).

Exhibitions held in Berlin and London in 1910 were key events in urban planning history. Featuring projects from numerous countries, they aimed to convey political, cultural, and aesthetic statements. The London exhibition was held in conjunction with an international conference on town planning, which had an impact as far away as Japan and included the first elements of Patrick Geddes's 'Cities and Town Planning Exhibition.' Major publications by Werner Hegemann based on the exhibitions of Berlin (1910), Düsseldorf (1912), and Gothenburg (1923) (Hegemann, 1913, 1925) expanded the range of those exhibitions' impact. Enthusiasm for urban planning was on the rise. Three major urban planning exhibitions were held in 1913 alone: in Nancy, Ghent, and New York. Other exhibitions followed even in national realms, including in Australia, the United Kingdom, and the United States (Freestone and Amati, 2014). Town planner and urban theorist Patrick Geddes expressed a common sentiment when he praised exhibitions for modifying the civic atmosphere:

> "In each city the Town Planning Exhibition has effected, more or less, this education of public opinion, and with practical results. Sometimes this impulse is a diffused one, as in Edinburgh, with results manifold, but not easily separately traced. Sometimes there are immediate definite results to which we can point, as in Dublin: e.g. (a) formation of a "Cities and Town Planning Department" of the National Museum, (b) formation of a "Town Planning Association of Ireland," with (c) immediate initiatives of practical Improvements in Dublin itself." (Geddes, 1913)

Geddes's exhibition on 'City and Town Planning' was sent to India and started a major trend of traveling shows – a means to bring the concept of planning to a broad group of people.

Other planning exhibitions followed, and for some institutions, notably the New York Museum of Modern Art (MoMA), traveling exhibitions became a major instrument in spreading planning ideas within the United States and beyond. Starting with the 1932 exhibition 'Modern Architecture: International Exhibition,' which included a section on *Siedlungen*-type community housing, select members and associated advisors of the institution used exhibitions to translate European modernist planning into American practice. From the 1930s onward, exhibitions were often inspired by modernist planning ideas, local problems, and the desire for public involvement: the 'Coventry of Tomorrow' exhibition in 1939; the Rebuilding Britain Exhibition of 1943; or New York MoMA's circulating 1944 exhibition 'Look at Your Neighborhood,' which was designed by Rudolf Mock (with support from Clarence Stein) to educate the American public on the need for comprehensive postwar urban planning (Mock, 1945).

The need for adequate housing for workers during the war, the wartime experience with planning as a helpful tool for organizing large spaces, and the looming car-based sprawl of postwar cities spurred calls for comprehensive planning and for intense reflection on the peacetime future of cities, whether war destroyed or not. The reconstruction exhibition 'Berlin Planterster Bericht' of 1946 highlights attempts to redefine Berlin's urban area even beyond the war-ravaged areas. Many other rebuilding exhibitions followed, notably in the war-damaged German cities (Petz, 2010). Many municipalities around the world held planning exhibitions to introduce their population to new urban concepts. Edmund Bacon, who headed the Philadelphia City Planning Commission from 1949 to 1970, organized the 1947 Better Philadelphia Exhibition that laid the groundwork for the postwar transformation of the city. The exhibition attracted some 385 000 people; and the pamphlet accompanying the exhibition tried to convince the general public of the need for comprehensive urban transformation.

Exhibitions continued to be places to imagine the future as the approach to urban planning changed in the 1960s. MoMA's 1967 exhibition 'The New City: Architecture and Urban

Renewal' promoted rational land development, integrated transportation, citizen participation, neighborhoods with public facilities, and housing with the goal of economic development in search of a new 'wholesome urban environment,' in the words of Sidney J. Frigand, the former deputy executive director of the New York City Planning Commission (Museum of Modern Art, 1967: p. 7). Many of the projects in the show were just as heavy-handed, but his words showed a changing approach to urban form and life. The exhibition was not limited to documenting the past or proposing abstract solutions; it intervened in the future of New York with very specific design proposals by several teams, including one from Cornell led by Colin Rowe and Thomas Schumacher.

In recent decades, architectural museums and archives have sponsored a variety of exhibitions on planning in different periods, making them instruments for the promotion of particular concepts or styles. At times, these institutions intervene in planning, as was the case with the 'Berlin Morgen' exhibition in Frankfurt am Main in 1991, where proposals later reappeared as competition entries (Lampugnani and Mönninger, 1991). Other exhibitions concentrate on historic plans: 'La Ville. Art et Architecture en Europe 1870–1993' (1994), held at the Center Pompidou in Paris and later staged in Tokyo, provided both the general and the professional public with detailed insights into cross-cultural planning. Traveling exhibitions of printed boards and models can reach people around the globe, on themes such as the rebuilding of the Tohoku Region in Japan after the 2011 tsunami and the Fukushima reactor disaster.

Exemplary exhibitions also take built form, whether provisional or permanent. Such displays are a means to try out visions, promote urban concepts, or display recent achievements. World's fairs have concretely featured urban forms. The American world's fairs in Chicago (1893) and St. Louis (1904) were temporary manifestations of City Beautiful concepts and helped to popularize the new style. The New York World's Fairs in 1939 and 1964 helped promote the concepts of the motorized city of the future. More recent events include those in New Orleans in 1984, Tsukuba in 1985, Vancouver in 1986, Brisbane in 1988, Seville in 1992, Lisbon in 1998, Hanover in 2000, Zaragoza in 2008, Shanghai in 2010, and Yeosu, South Korea, in 2012, with the most recent ones focusing on issues of urban sustainability.

World's fairs have also been the locus for the display of urban transformations as embodied in large-scale urban models. Models of the Liverpool docks and Hamburg's harbor, for example, were displayed respectively at the world's fair in London in 1851 and Paris in 1900 to showcase and promote these locations. Le Corbusier displayed his vision of a modern Paris in 1925 in the Pavillon de l'Esprit Nouveau at the Exposition international des Arts Décoratifs et modernes. The 1939 New York World's Fair displayed a giant model of the city of the future, sponsored by General Motors. The Shanghai Urban Planning Exhibition Center with its huge model of current and approved buildings continues this tradition.

Exemplary exhibitions have created a range of permanent urban elements, transport infrastructure, riverfront redevelopment, parks, or housing. Housing has been a focus for several international building exhibitions, the most famous of them in Germany, where the form and function of residential design in each epoch were represented through its own exhibition. Examples included the artists' colony at Mathildenhöhe near Darmstadt (1901), the 1925 Stuttgart exhibition entitled 'Die Wohnung,' modernist projects such as the Dammerstock mentioned in this article, or the 'traditionalist' Stuttgart-Kochenhof Siedlung (1933) (Cramer and Gutschow, 1984). Many building exhibitions showcasing urban design and housing accompanied rebuilding after World War II in numerous German cities, including Nürnberg, München, and Hamburg. Among the most influential ones are the Constructa (1951) in Hanover, which aimed to lay out principles for reconstruction, and the Interbau Berlin (1957) that featured the reconstruction of an inner-city district as a high-rise housing area, the Hansaviertel.

Germany has continued the practice of building exhibitions in recent decades with several innovative events, including the International Building Exhibition (IBA) in Berlin of the 1980s, which aimed at a critical reconstruction of the city and inspired the Kumamoto Artpolis of the 1990s in southern Japan (a building exhibition transformed into a traveling exhibition that toured the world); and the International Building Exhibition Emscher Park in the Ruhr region (1989–99), which aimed at setting transformational impulses for an old industrial area (Shaw, 2002). The Hannover 2000 Expo became the flagship for a combination of world's fairs, model planning, and housing exhibition. There, an overall city regeneration plan included 2000 housing units built to respond to environmental and social sustainability concerns. The IBA Hamburg (2013) and other concurrent events continue this tradition into the present (e.g., the Netzwerk 'IBA Meets IBA') (Hamburg, 2010).

The German concept of building exhibitions has found international interest. Numerous cities elsewhere have undertaken housing regeneration expos such as Bo97 in Staffanstop (Sweden, 1997), and Glasgow's Homes for the Future Housing Expo (1999), Malmö (2001), and Stockholm (2002) continued this tradition. The trinational exhibition, the IBA Basel 2020, may be an indication of further internationalization of the concept. Exemplary exhibitions are extremely costly and time-consuming, and they require broad participation; they do, however, help to create larger interest in urban development issues – bringing together socioeconomic and physical development issues, providing an innovative instrument to flexibly address large-scale problems within a broad framework, allowing the mobilization and focusing of resources, developing paradigmatic examples, and possibly becoming models for collaboration among various urban stakeholders (Hamburg, 2010). The insertion of viewing platforms and information pavilions as part of large-scale urban transformations such as that of the HafenCity in Hamburg, or institutions such as the Shanghai Urban Planning Exhibition Center, build on the success of exemplary exhibitions and make entire areas of cities effectively an exhibition site.

A Cultural Phenomenon

Over the last century, competitions and exhibitions have been important tools in the spread of urban theories, concepts, ideas, and planning practices. They have not only echoed the development of modern planning but also furthered it through

the promotion of exchange, particularly between European countries and North America, since the beginning of the profession at the turn of the nineteenth century when there were no examples to follow. They were helpful tools in cities' and planners' search for the best functional and aesthetic solutions. They have been a tool in cultural transfer and have contributed to the development of the discipline and of building culture.

There are, however, cultural differences in the historical and contemporary use of these instruments. Germany has a long tradition of public competitions and exhibitions. France's competitions and exhibitions have been sustained not only by the public but also by architectural schools. Both countries continue to promote these tools. Other countries, such as Japan, with different planning traditions and different aims have held only a limited number of competitions, and those have had limited impact. Instead, these countries developed other approaches for the selection of urban design projects. Japan, for example, historically preferred to invite foreign experts, send planners on study trips, and introduce foreign projects and ideas in magazines or book translations. Direct commissioning continues to be highly valued.

Over the last century, competitions and exhibitions have become an important tool for urban transformation. The increased desire for urban marketing and branding emphasizes their importance for the major stakeholders in the city. Most surprising among the current promoters may be China, with its current flurry of competitions and exhibitions bringing together practitioners from around the world.

See also: Critical Planning Theory; Planning Ethics; Strategic Planning for Public and Nonprofit Organizations; Urban Places: Statistical Definitions; Urban Planning (Western), History of; Urban Sprawl.

Bibliography

Albers, Gerd, 1997. Zur Entwicklung der Stadtplanung in Europa. Vieweg, Braunschweig.

Aymone, Nicolas, 2007. L'apogée des concours internationaux d'architecture. L'action de l'UIA 1948–1975. Picard, Paris.

Becker, Heidede, 1992. Geschichte der Architektur und Städtebauwettbewerbe. Verlag W. Kohlhammer, Stuttgart.

Commission on Congestion of Population, February 28, 1911. Report of the New York City Commission on Congestion of Population: Transmitted to the Mayor and the Board of Aldermen. Lecouver Press Company, New York.

Cramer, Johannes, Gutschow, Niels, 1984. Bauausstellungen. Eine Architekturgeschichte des 20. Jahrhunderts. W. Kohlhammer, Stuttgart, Berlin.

Diefenbacher, Michael, Henkel, Matthias, Stadtarchiv, Nürnberg, 2009. Wiederaufbau in Nürnberg. Schmid, Neustadt an der Aisch.

Freestone, Robert, Amati, Marco. Exhibitions and the Development of Modern Planning Culture February 2014.

Geddes, Patrick, 1913. Two steps in civics: "Cities and Town Planning Exhibition" and the "International Congress of Cities": Ghent International Exhibition, 1913. The Town Planning Review 4, 78–94.

Gerkan, Meinhard Von, 2008. Projects in China. Architects von Gerkan, Marg and Partners. Prestel, Munich.

Haan, Hilde De, Haagsma, Ids, Sharp, Dennis, Frampton, Kenneth, 1988. International Architectural Competitions of the Last 200 Years. Thames and Hudson, London.

Hall, Thomas, 1997. Planning Europe's Capital Cities: Aspects of Nineteenth-Century Urban Development. E&FN Spon, London.

Hamburg, I.B.A., 2010. IBA Meets IBA: eine Ausstellung zur 100jährigen Geschichte der Internationalen Bauausstellungen. Jovis, Berlin.

Hegemann, Werner, 1913. Der Städtebau nach den Ergebnissen der Städtebau-Ausstellungen in Berlin und Düsseldorf 1910–1912. E. Wasmuth, Berlin.

Hegemann, Werner, 1925. Der Städtebau nach den Ergebnissen der internationalen Städtebau Ausstellung Gothenburg. E. Wasmuth, Berlin.

Hein, Carola, 1991. Berlin Städtebaulicher Ideenwettbewerb 1957–58. Gebr-Mann Verlag, Berlin.

Hein, Carola, 2002. Hiroshima: the atomic bomb and Kenzo Tange's Hiroshima Peace Center. In: Ockman, Joan (Ed.), Out of Ground Zero. Case Studies in Urban Reinvention. Temple Hoyne Buell Center for the Study of American Architecture, Columbia University, Prestel, New York, München.

Hein, Carola, 2004. The Capital of Europe. Architecture and Urban Planning for the European Union. Greenwood/Praeger, Westport, CT.

Herscher, Andrew, 2003. Stadtebau as imperial culture: Camillo Sitte's urban plan for Ljubljana. The Journal of the Society of Architectural Historians 62, 212–227.

Holston, James, 1989. The Modernist City: An Anthropological Critique of Brasilia. The University of Chicago Press, Chicago and London.

Jong, Cees De, Mattie, Eric, 1994. Architectural Competitions 1792–1949/Architektur Wettbewerbe 1792–1948/Concours d'Architecture 1792–1949. Benedikt Taschen/V+K, Cologne/Naarden.

Lampugnani, Vittorio Magnano, Mönninger, Michael (Eds.), 1991. Berlin Morgen. Ideen für des Herz einer Großstadt.

Meller, Helen, 1995. Philanthropy and public enterprise: international exhibitions and the modern town planning movement, 1889–1913. Planning Perspectives 10 (3), 295–310.

Mock, Elizabeth (Ed.), 1945. Built in USA since 1932. The Museum of Modern Art, New York.

Omi, Sakae, 1986. Kenchiku sekkei kyôgi. A History of Architectural Competitions in Japan. Kajima Shuppankai, Tokyo.

Risler, George, 1915. Préface. In: Agache, Donnat-Alfred, Auburtin, Jean-Marcel, Redont, Edouard (Eds.), Comment reconstruire nos cités détruites, notions d'urbanisme s'appliquant aux villes, bourgs et villages. A. Colin.

Shaw, Robert, 2002. The International Building Exhibition (IBA) Emscher park, Germany: a model for sustainable restructuring? European Planning Studies 10 (1), 77–97.

Speiregen, Paul, 1979. Design Competitions. McGraw Hill, New York.

Veiller, Lawrence, 1900–1901. The tenement-house exhibition of 1899. Charities Review 10, 19–25.

Petz, Ursula von, 2010. City planning exhibitions in Germany, 1910–2010. Planning Perspectives 25 (3), 375–382.

Urban Planning: Methods and Technologies

Richard E Klosterman, What if?, Inc., Hudson, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The collection, analysis, and dissemination of the information needed to improve public and private decision making is one of the major justifications for urban planning. As a result, planning methods and technologies have played a central role in planning practice and education since the profession's inception over 100 years ago. This article reviews the evolution of planning methods and techniques, considers the most important technologies used in contemporary planning practice, and speculates on the future of planning methods and technologies.

Evolution of Planning Methods and Technologies

Three relatively distinct eras can be traced in the evolution of planning practice, methods, and techniques. The first era, 'planning as design, characterized planning from its inception until the 1960s. The second, 'planning as applied science, characterized the profession in the 1960s and 1970s. The third, 'planning as communication, has characterized the profession since the 1980s. The following sections examine each of these eras and their corresponding methods and technologies in turn.

Planning as Design

The planning profession in the United States emerged at the beginning of the twentieth century in response to the unregulated urban growth and inadequate public facilities of the new industrial city. Guided by a rather naive form of environmental determinism, the early planners assumed a professional responsibility for improving society through changes in the physical environment and protecting an overarching public interest from the presumably self-interested and uninformed actions of elected officials and private citizens. Planners' methods were largely intuitive methods of manual design adapted from the profession's institutional homes in architecture and landscape architecture, taught primarily by planning practitioners in the ubiquitous studio workshop.

By the 1950s, these largely intuitive methods had evolved into a standard land use planning approach codified in F. Stuart Chapin's *Urban Land Use Planning* (Chapin, 1957) and T.J. Kent's *The Urban General Plan* (Kent, 1964). This model defined planning as the preparation of a formal document laying out a long-term, comprehensive, and general vision for a community's future physical development, including public and private uses of land and related public facilities. These plans included a summary of existing and emerging conditions and needs, a statement of planning goals, a 20- or 30-year development plan expressed in map form, and policies for implementing the plan. This conception of planning continues to dominate planning education and practice, as reflected in the fifth edition of Chapin's text (Berke et al., 2006).

Planning as an Applied Science

The planning profession's attention turned in the 1960s from the design of the physical city to new concerns with improving transportation access, reducing urban blight, and improving social and economic conditions at the urban and regional scales. The dominant perception of the field changed from an image of 'planning as design' to 'planning as an applied science' as new students, faculties, and departments joined the profession from outside the traditional schools of design. An increased emphasis on quantitative methods and social science theory led to the abandonment of the once ubiquitous design studio for new courses in research design, statistics, and quantitative techniques utilizing texts such as Isard's monumental *Methods of Regional Analysis* (Isard et al., 1960) and Krueckeberg and Silvers' *Urban Planning Analysis* (Krueckeberg and Silvers, 1974). Planners' professional role was now seen as delivering unbiased, 'objective, and politically neutral advice to elected officials and the public and collecting and disseminating more and better information that could inform and improve the policy-making process.

Computers first began to be used in planning in the 1960s, as part of a general faith in science and technology and the emergence of new academic fields such as operations research, urban economics, and regional science. The most notable early attempts to use computers in planning were the ambitious and highly visible federally funded efforts to build large-scale metropolitan land use – transportation models and integrated municipal information systems. These efforts were seen as providing the foundations for a new 'science' of planning by helping planners understand and guide the urban development process and improve the information base for public and private decision making (Harris, 1965; Webber, 1965).

Planners' optimistic faith in computer technology was severely tested in the 1970s. The pioneering efforts to develop large-scale urban models and municipal information systems failed spectacularly, largely as a result of their overly ambitious goals, the limitations of available computer equipment, and the lack of required data. Sophisticated analytic tools such as mathematical programing which were assumed to provide the foundations for 'scientific planning' were found to be largely inappropriate for planners' 'wicked problems' which are hard to define or solve, have outcomes that are difficult to predict, and involve considerable uncertainty (Lee, 1973; Rittel and Webber, 1973).

The practical failure of the early computer-based urban models reflected a more fundamental rejection of the 'rational planning' model that was assumed to underlie all attempts to use computers in planning. Planning methods and technologies

that were once assumed to be 'objective' and politically neutral were found inevitably to increase the power of administrators and technical experts at the expense of those who lacked the expertise to use them effectively. Information that appeared to be more accurate, credible, and objective simply because it had been generated by a computer was found to reflect politically important choices in the selection of data, the application of computational procedures, and the distribution of results (Wachs, 1982). By the decade's end, most planners had no access to – or perceived need for – the expensive, fragile, and hard-to-use mainframe computers of the era and continued to rely on manual computational techniques that had changed little since the 1950s.

Planning as Communication

Planners' faith in computing technology was reborn in the 1980s with the arrival of microcomputers which made computer-based methods and techniques readily available to planners the world over. By the century's end, planners routinely used computers for general-purpose office functions such as processing documents, monitoring budgets, and maintaining records. By 2010, planners around the world routinely used geographic information systems (GIS; to be discussed below) to process permits, maintain land-related information, and prepare thematic maps. Equally important, the increased availability of information from municipal information systems and the Internet allowed planners to quickly and easily process increasingly large quantities of spatially-related information.

Underlying these developments were new images of 'planning as communication' (Innes, 1995). Planners increasingly realized that planning involves not only the collection and distribution of numerical data and formal documents but, more importantly, face-to-face communication with colleagues and clients involving less traditional types of 'information': stories, advice, and personal experiences. As a result, planners began to recognize that the *ways* in which they transmit information are often more important than *what* they say. For example, planning forecasts could be described either as the result of highly sophisticated computer-based procedures that exclude nonexperts or as implementing easily understood models and assumptions that invite public comment.

Today's Planning Technologies

Before the advent of computers, planning data were available only in voluminous printed volumes and paper maps that were incomplete, inconsistent, and could only be updated manually. Planning analysis and communication required manually copying data from printed documents, laborious and error-prone hand computations, and preparing text, tables, maps and figures by hand.

Today all of this has changed. A wealth of spatially-related data is readily available in digital form from public and private data providers. Planners' computer-based tools allow them to analyze these data quickly and use them to produce attractive documents, charts, and maps that they can easily distribute to clients and other professionals. As a result, it is almost impossible to conceive of planning *without* computers and related information and communication technologies (ICTs). The following section briefly reviews the technologies that are most widely used by today's planners.

Geographic Information Systems

GIS are computer systems for capturing, storing, querying, analyzing, and displaying information about natural and built entities and activities that can be displayed on a map. GIS are particularly powerful analysis tools because they consider cartographic, attribute, and topological data about spatially related entities and activities. The cartographic data describe *where things are*, that is, their location relative to a reference system, and provide the information needed to prepare a map. For example, the cartographic data for the lines on a map include their location (beginning and ending points), shape (straight or curved), type (solid, dotted, or dashed), and color. Attribute data describe *what things are*, such as, their magnitude, classification, and ownership. For example, the attribute data for a tax assessor's map would contain information on the street address, use, ownership, and assessed value for land parcels. Topological data describe spatial relationships, that is, *where things are located relative to other things*. Topological relationships include: (1) adjacency, for example, the fact that France and Poland are both next to Germany but are not next to each other; and (2) containment, for example, that Paris is located in France, not Germany.

GIS have proven extremely useful for a wide range of planning applications such as analyzing and displaying spatially related information, maintaining land-based inventories, and administering permitting processes. Unfortunately today's GIS only consider the present and past and cannot support planners' particular concerns with projecting future conditions, evaluating the long-term implications of alternative public choices, and considering public policy goals, benefits, and costs.

Planning Support Systems

In response to the limitations of the current generation of GIS, the last decade has witnessed a rebirth in the development of planning models that use the tremendous analysis and display power of GIS and the increased availability of spatially-related data to develop customized planning support systems (PSS) that meet the needs of planners (see, e.g., Brail and Klosterman, 2001; Brail, 2008). The following approaches have dominated efforts in this area: (1) large-scale urban models, (2) cellular automata (CA) models, (3) agent-based models, and (4) rule-based models. (Other modeling approaches are reviewed in Trianstakonstantis and Mountrakis (2012)).

Large-Scale Urban Models

Large-scale urban models for studying the interaction between land uses and spatial interaction first attracted planners' attention to computer-based models in the 1960s (see, e.g., Harris, 1965). These models project future land uses, housing and employment for zones within a metropolitan area, and the travel by transportation mode between these zones (Miller et al., 1999; Hunt, 2005). More sophisticated models attempt to represent the complex interactions between land uses and

transportation demand, the market for land and space, and the location of households and firms over time and space. Models of this type that have been widely implemented in practice include DRAM/EMPAL (Putman, 1983), MEPLAN (Bertuglia et al., 1990), TRANUS (de la Barra, 2001), and UrbanSim (Waddell, 2002).

CA Models

CA models of urban growth have received a tremendous amount of academic attention in recent years (Trianstakonstantis and Mountrakis, 2012). CA models represent an urban area with a lattice of cells, each of which exists in one of a finite set of states; for example, 'developed' and 'undeveloped. The progression of time is modeled as a series of discrete steps with future patterns determined by transition rules that specify the behavior of cells over time (e.g., whether a cell switches from undeveloped to developed), as a function of conditions at each cell and its neighboring cells at each time step. A widely used CA model, SLEUTH, which stands for the input requirements of the models – slope, land cover, exclusion, urbanization, transportation, and hill shade – uses information on land uses at two points of time to calibrate a statistical model that projects future land use changes (Clarke, 2008).

Agent-Based Models

As their name suggests, agent-based systems attempt to model the behavior of complex systems by representing the behavior of the various agents that make up the system. Thus, for example, an agent-based urban model would represent the actions of a region's households, firms, and land developers who interact to buy and sell land via a simulated market. The behavior of each of these agents is modeled, along with their interactions with other agents and with the environment, which together determine future urban development patterns (Heppenstall et al., 2012).

Rule-Based Models

Rule-based models first became popular in planning with the publication of John Landis's landmark California Urban Futures (CUF) model, the first GIS-based urban development model (Landis, 1994). CUF and other rule-based models do not attempt to duplicate the complex spatial interaction and market clearing processes that underlie the large-scale urban models. Instead, they incorporate explicit decision rules that allow model users to specify how the model will behave. Thus, for example, rule-based urban growth models such as CUF and What if? (Klosterman, 1999) incorporate user-specified rules for determining the relative suitability of different locations, projecting future land use demands, and allocating projected land use demands to suitable sites. Similarly, rule-based impact assessment models such as CommunityViz (Walker and Daniels, 2011) and INDEX (Allen, 2008) specify rules that determine the impacts that proposed actions will have on future development patterns, public infrastructure, and the like.

The Internet

The Internet and the World Wide Web (or 'Web') provide a 'network of networks' that connect millions of computer users around the world into a seamless web of electronic communication and information dissemination. The Internet has been extremely useful in providing planners and the public with anytime, anyplace access to large databases such as the entire contents of the national census and interactive audio and video displays. It also allows plans and proposals to be distributed quickly and easily, allowing people with different skills and viewpoints to participate in the public policy-making process, independent of time and place. The recent development of powerful Web services promises to make a new generation of online models and methods readily available to planners around the world.

Today's Planning Methods

Continued advances in the power and capabilities of modern GIS and the availability of high-quality spatial data have stimulated a revolution in the area of spatial analysis. Important issues such as spatial autocorrelation and the modifiable area problem which had been largely ignored in the past are now being addressed thoroughly, and exciting new tools and techniques for analyzing and describing spatially related phenomena are being developed and refined (Fotheringham and Rogerson, 2009). Unfortunately, these developments have had little impact on planning practice and education. A host of planning methods texts were published in the 1970s, but only two (Klosterman, 1990; Wang and Hofe, 2007) have been published since then. A new textbook demonstrating how current GIS technology and traditional planning can be used to address the realities of today's planning is being prepared.

The Future of Planning Methods and Technologies

Planning methods and technologies have changed dramatically over the last 50 years, due largely to the incredible advances in ICT that have occurred since the birth of computer-aided planning in the 1960s. Fragile and expensive room-sized computers and paper 'punch cards' have been replaced by low cost and more powerful machines that fit into our hands. Today's GIS provide tools for high-quality cartography, three-dimensional visualization, and spatial analysis that could not be imagined a generation ago. A wealth of spatially related data is currently available from government and commercial data sources, increasingly via the Web. Free and easy-to-use online mapping and route-finding tools have made spatially referenced information and services a routine part of people's lives. The rapid development of more powerful Web services promises to make digital data and advanced computing tools readily available to planners around the world. Perhaps most importantly, the grand challenges of climate change, resource depletion, and environmental degradation increasingly require computer-based analysis, forecasting, and evaluation. Together these trends are encouraging academic institutions, innovative government agencies, and entrepreneurial organizations around the world to develop and apply a new generation of Web-based models and methods.

Nevertheless, experience has demonstrated that progress in the field of computer-aided planning will not be quick or easy.

A vast body of academic literature on computer-aided models and methods has accumulated since computers entered planning in the 1960s (see e.g., Harris, 1985; Klosterman, 1992; Trianstakonstantis and Mountrakis, 2012). Complex and expensive integrated land use/transportation models are used increasingly in practice, but only by well-funded academic institutions and regional transportation planning agencies. An overwhelming array of GIS-based PSS models is being developed around the world but most of them are academic prototypes or 'one off' professional applications that have not been adopted elsewhere. A number of commercial off-the-shelf PSS have been developed but these systems have been adopted only by a small number of agencies and are rarely used on a routine basis.

Important obstacles can be identified that have limited the adoption of computer-based planning models and methods (Vonk et al., 2005; Vonk and Geertman, 2008). Academic institutions are excellent venues for developing models but cannot provide the long-term support required for their continued use in professional practice. Efforts by commercial firms to develop planning models are hindered by a vicious circle: practitioners are understandably reluctant to adopt an immature and poorly known technology; the limited market for planning models restrains the ability of developers to improve and market their tools; this completes the circle by limiting planners' awareness of, and willingness to adopt, computer-based models and methods.

More fundamentally, the adoption of computer models and methods is limited by the somewhat paradoxical fact that today's planners often do little planning. Planners are so occupied by routine administrative tasks such as processing permits, enforcing building codes, and administering regulations that they all too often fail to help the communities they serve engage an uncertain and politically contentious future. As a result, the current effort to develop GIS-based planning models is part of a larger effort to return the planning profession to its traditional concern with using information and analysis to help the public more effectively engage its collective future (Isserman, 1985; Myers, 2001; Hopkins and Zapata, 2007).

The current effort to apply advanced ICTs to planning is also part of a fundamental change in our image of planning and the role that technology can play in planning research and practice. The first generation of computer models and methods was assumed to provide the foundations for value-neutral 'planning for the public' that would allow planners, as technical experts on the city, to understand the urban development process, accurately forecast the future, and identify the best policy option, with little need to involve the public. Today's planning models generally pursue new ideals of citizen-based 'planning with the public' which limit planners' role to structuring and informing a citizen-led decision process and attempt to involve stakeholders and private citizens in all stages of the planning process – understanding the past and present, identifying problems, evaluating alternative solutions, and attempting to make decisions they can support (Innes, 1996). ICT can play an important role in this effort but it will require the development of new models and methods that allow nonexperts to understand present and past conditions, clearly identify alternative public policy options, and easily evaluate their implications.

We live in an era of exciting technical possibilities, an increased appreciation for spatially referenced information, widespread concern with issues such as climate change and resource scarcity, and a growing demand for more meaningful public participation. These demands will only increase with the growing interdependence, complexity and uncertainty of a rapidly urbanizing world. Dramatic advances in ICTs may provide the technological foundations a rejuvenated profession that will once again claim its traditional role as a source of inspiration about what the future might – and should – be and promote more meaningful citizen involvement. Only time and professional practice will tell whether this potential is achieved.

See also: Finance, Geography of; Geographic Information Systems and Remote Sensing.

Bibliography

Allen, Eliot, 2008. Clicking toward better outcomes: experience with INDEX, 1994 to 2006. In: Brail, R.K. (Ed.), Planning Support Systems for Cities and Regions. Lincoln Institute of Land Policy, Cambridge, MA.

Berke, Phillip R., Godschalk, David R., Kaiser, Edward J., Rodriguez, Daniel A., 2006. Urban Land Use Planning, fifth ed. University of Illinois Press, Urbana, IL.

Bertuglia, C.S., Leonardi, G., Wilson, A.G. (Eds.), 1990. Urban Dynamics: Building an Integrated Model. Routledge, London.

Brail, Richard K. (Ed.), 2008. Planning Support Systems for Cities and Regions. Lincoln Institute of Land Policy, Cambridge, MA.

Brail, Richard K., Klosterman, Richard E. (Eds.), 2001. Planning Support Systems: Integrating Geographic Information Systems, Models and Visualization Tools. ESRI Press, Redlands, CA.

Chapin, F. Stuart, 1957. Urban Land Use Planning. Harper, New York.

Clarke, Keith C., 2008. A decade of cellular urban modeling with SLEUTH: unresolved issues and problems. In: Brail, R.K. (Ed.), Planning Support Systems for Cities and Regions. Lincoln Institute of Land Policy, Cambridge, MA.

de la Barra, Tomas, 2001. Integrated land use and transport modeling: the tranus experience. In: Brail, R.K., Klosterman, R.E. (Eds.), Planning Support Systems: Integrating Geographic Information Systems, Models, and Visualization Tools. ESRI Press, Redlands, CA.

Fotheringham, A. Stewart, Rogerson, Peter A. (Eds.), 2009. The SAGE Handbook of Spatial Analysis. SAGE Publications, London.

Harris, Britton, 1965. Urban development models: new tools for planning (Guest Editor). Journal of the American Institute of Planners 31, 90–183.

Harris, Britton, 1985. Synthetic geography: the nature of our understanding of cities. Environment and Planning, A 17, 443–464.

Heppenstall, Alison J., Crooks, Andrew T., Linda, M. See, Batty, Michael (Eds.), 2012. Agent-Based Models of Geographical Systems. Springer, Dordrecht.

Hopkins, Lewis D., Zapata, Marisa A. (Eds.), 2007. Engaging the Future: Forecasts, Scenarios, Plans, and Projects. Lincoln Institute of Land Policy, Cambridge, MA.

Hunt, J.D., 2005. Integrated Land-Use and Transport Models: An Introduction. Transportation Research Board, Washington, DC (Workshop 162).

Innes, Judith E., 1995. Planning theory's emerging paradigm: communicative action and interactive practice. Journal of Planning Education and Research 14, 183–189.

Innes, Judith E., 1996. Planning through consensus building: a new view of the comprehensive planning ideal. Journal of the American Planning Association 62 (4), 460–472.

Isard, Walter, et al., 1960. Methods of Regional Analysis: An Introduction to Regional Science. MIT Press, Cambridge, MA.

Isserman, Andrew M., 1985. Dare to plan: as essay on the role of the future in planning practice and education. Town Planning Review 36 (4), 483–491.

Kent, T.J., 1964. The Urban General Plan. Chandler, San Francisco.

Klosterman, Richard E., 1990. Community Analysis and Planning Techniques. Rowman and Littlefield, Savage, MD.

Klosterman, Richard E., 1992. Evolving views of computer-aided planning. Journal of Planning Literature 6 (3), 249–259.

Klosterman, Richard E., 1999. The what if? collaborative planning support system. Environment and Planning, B: Planning and Design 26, 393–408.

Krueckeberg, Donald A., Silvers, Arthur L., 1974. Urban Planning Analysis: Methods and Models. Wiley, New York.

Landis, John D., 1994. The California urban futures model: a new generation of metropolitan simulation models. Environment and Planning, B 21, 399–420.
Lee Jr., Douglas B., 1973. Requiem for large-scale models. Journal of the American Institute of Planners 39 (3), 163–178.
Miller, Eric J., Kriger, David S., Hunt, John David, 1999. Integrated Urban Models for Simulation of Transit and Land Use Policies: Guidelines for Implementation and Use. Transit Cooperative Research Program Report 48. Transportation Research Board, Washington, DC.
Myers, Dowell, 2001. Symposium: putting the future in planning. Journal of the American Planning Association 67 (4), 365–401.
Putman, Steven H., 1983. Integrated Urban Models: Policy Analysis of Transportation and Land Use. Pion, London.
Rittel, Horst W.J., Webber, Melvin M., 1973. Dilemmas in a general theory of planning. Policy Sciences 4, 155–169.
Trianstakonstantis, Dimitrios, Mountrakis, Giorgos, 2012. Urban growth prediction: a review of computational models and human perceptions. Journal of Geographic Information Systems 4, 555–587.
Vonk, Guido, Geertman, Stan, 2008. Improving the adoption and use of planning support systems in practice. Applied Spatial Analysis 1, 153–173.
Vonk, Guido, Geertman, Stan, Schot, Paul, 2005. Bottlenecks blocking widespread usage of planning support systems. Environment and Planning, A 37, 909–924.
Wachs, Martin, 1982. Ethical dilemmas in forecasting for public policy. Public Administration Review 42 (6), 562–567.
Waddell, Paul, 2002. UrbanSim: modeling urban development for land use, transportation and environmental planning. Journal of the American Planning Association 68 (3), 297–314.
Walker, Doug, Daniels, Tom, 2011. Planners Guide to CommunityViz: The Essential Tool for a New Generation of Planning. APA Planners Press, Chicago, IL.
Wang, Xinhao, Hofe, Ranier vom, 2007. Research Methods in Urban and Regional Planning. Tsinghua University Press, Beijing.
Webber, Melvin M., 1965. The roles of intelligence systems in urban systems planning. Journal of the American Institute of Planners 31, 289–296.

Urban Policies in Asia

Mee Kam Ng, The Chinese University of Hong Kong, Shatin, Hong Kong, China

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Asia is a colorful mix of countries with varying degree of urbanization and can roughly be classified geographically into East, Southeast, and South Asia. While the state is central to the formulation of urban policies in Asia, the more developed parts put more emphasis on the green agenda, whereas the less developed parts continue to be snared by the brown agenda. Asian countries are keen to join the globalized worlds but various shades of civil societies have also challenged the impacts of the trend of globalization on local development as well as environmental and social sustainability.

Diversities

By 2030, more than half of the world's urban population will be in Asia (ADB, 2011; Clark et al., 2009). Asia is not a unified whole and geographically can further be divided into East Asia, Southeast Asia, and South Asia with distinctive internal differences. East Asia consists of Japan, poor and dictatorial North Korea, wealthy and more democratic South Korea, China (with her Special Administrative Regions, Hong Kong, and Macau), Taiwan, and Mongolia. In Southeast Asia, fully urbanized Singapore leads in terms of general economic development, followed by Malaysia, Thailand, Indonesia, the Philippines, Brunei, and the poor countries in the Indo-Chinese region including Vietnam, Cambodia, Laos, and Myanmar where only one-fifth to one-third of their population is urban. South Asia is a region with diversified cultures, people, religions, and languages. It is led by India in terms of land area, population, and economic development, followed by the rivalry Muslim countries of Pakistan and Bangladesh and the small island countries of Sri Lanka and Maldives, and landlocked mountainous countries of Nepal and Bhutan. From a political economic perspective, one end of the spectrum consists of state-centered market economies at different levels of development and the other end of the spectrum is made up of transitional socialist market economies including China and the countries in former French Indo-China.

It is impossible to talk about Asian urban policies without understanding the meaning of 'urban' in the historical traditions of diversified Asia. In the bygone past, the urban realm arose as a result of agricultural surpluses and as such cities in China used to be just administrative and market centers for their rural counterparts. The evolving urban policies in Asia are undeniably outcomes of mega trends such as millennia of indigenous histories; the age of exploration, colonialism, and imperialism in the last two centuries, the Cold War and the recent renewed round of accelerated globalization; entrenched and historically evolved microsocioeconomic and political processes; and power play among different levels of states, the business sector, and the civil society. Whether developed or developing, the nature of urban challenges faced by Asian countries is formidable: how to develop a pathway that manifests their time-honored traditions (such as emphasis on the roles of state and family relationships), conserves their fragile environments and social networks, and, at the same time, allows them to 'converge' with the rapidly globalizing world dominated by Western standards and values. In terms of the actual contents of urban challenges, these countries are manifestly very different.

The more developed urban realms in East and Southeast Asia are often seen as products of progrowth developmental and bureaucratic states that have developed close and productive relationships with the private sectors (Amsden, 1992; Wade, 1990; Waley, 2012; Woo-Cummings, 1999) in economic growth. However, increasingly this economic development strategy has given ways to concerns about environmental and social sustainability. Hence, naturally their urban policy agenda revolve around economic restructuring, environmental, social, and cultural issues in face of heightened globalization. Urban development in the less developed countries is often fueled by accelerating rural–urban migration, leading to the growth of informal economies, shanty towns, environmental degradation, pollution problems, health issues, conspicuous poverty, urban crimes, and extreme social polarization.

Whether a country is highly or hardly urbanized, the urban realm is critical for its development (UN Habitat, 2011) and cities contribute over 80% of gross domestic product (GDP) in many Pacific Asian countries (ADB, 2012). The quality of urban policies is often determined by the level of decentralization, including the dedication of resources to build hard and soft infrastructure. It is because according to Clark et al. (2009, 6), "the central government controls more than 70% of the revenues." As argued by Skocpol (1982, 6), "a state's means of raising and deploying financial resources tell us more than could any other single factor about its existing capacities to create or strengthen state organizations, to employ personnel, to co-opt political support, to subsidize economic enterprises, and to fund social programs." At the local level, the mode and quality of local governance have to rely on the availability of resources (including human resource such as the quality of the bureaucracy), and the scale of urban challenges in terms of economic opportunities, urban poverty, levels of air and water pollution, solid waste production, and so on. The trajectory of actual development on the ground very often is determined by the philosophy and power held by the stakeholders involved.

While it would be impossible to review urban policies in all the countries in Asia, the following, nevertheless, attempts to highlight key urban policies adopted in individual cities in

countries within the three subregions in order to give readers a flavor of the array of urban policies in diversified Asia.

East Asia

Although Japan has not really turned around since her economic crisis in the last decade of the twentieth century, her capital city, Tokyo, has always been at the forefront of promoting environmental protection and sustainable development. The city has faced socioeconomic challenges such as declining birth rates, aging population, and keen city competition and the Tokyo Metropolitan Government formulated the "Tokyo Vision 2020: Driving change in Japan/Showing our best to the world" in 2011. The key urban policies are related to Japan's vulnerability to natural hazards, including the construction of a disaster-resilient city; the need for efficient, locally produced, consumed and independent, distributed power production; and boosting global competitiveness through development as an international hub (op cit., 2011). To produce a resilient city, priorities are given to develop earthquake-resistant, affordable, and sustainable housing and communities fit for different age groups and will last for generations (BUD, 2011, 4). The city has also promoted public–private partnership, land readjustment means, and urban design to regenerate the Capital. Flagship projects include Roppongi Hills and the 31-ha Shiodome railway yard next to Shinbashi Station. They were undertaken by the private sector within rules and regulations set by the government and their eventual developments were also shaped, to certain extent, by voices from local communities (Waley, 2007). The 10-year project for Green Tokyo and the 10-year project for a Carbon-Minus Tokyo (Tokyo Climate Change Strategy) have been formulated to realize the Tokyo Vision 2020. The New Strategic Guidelines for Greenery formulated in 2006 aims to involve the public and private sectors and the civil society in greening the city. And various plans and laws are in place to use systematic and integrated measures to recycle resources. About 1300 facilities with a total annual consumption of fuels, heating, and electricity of at least 1500 kl crude oil equivalent are required to formulate, report, assess, and announce on measures against climate change (BOE, 2009). The city also uses its cutting edge technology to launch the Tokyo Ubiquitous Technology Project such as IC tags to allow access of information by "anyone, anytime, anywhere" in the capital city.

Similar to other Asian countries, South Korea is facing an aging population situation, demanding more elderly housing, elderly friendly environment, and transportation policy (OECD, 2012: 77). South Korea could have done a better job in developing a coordinated urban development strategy and help promote the internationalization and competitiveness of Seoul and other cities (OECD, 2012) but she has continued to practice an urban containment policy. As in the case of Japan, South Korea has nurtured big business interests, chaebols, headquarters of which can be found in the capital city, Seoul. In her capital region, the city, Seoul, has suffered from high-density development, escalating property prices, traffic congestion, intensive air pollution, and intense competition with and outside Korea. Subcenters have been developed to decentralized development. Popularized by the pop song *Gangnam Style*, Gangnam, a sub-center developed through land readjustment in the 1970s has become "the most expensive residential area, the most dynamic business area, and the most active entertainment area" (Kim and Han, 2012, 147).

South Korea has also experienced a 'green turn' of their urban development strategy. The revitalization of the once concrete-covered Cheonggye stream in 2005 has become a symbol of successful city-center regeneration. In 2008, the government launched the "Green New Deal" and in 2009, "the National Strategy for Green Growth up to 2050," both aim to "mitigate climate change, create new engines for economic growth and improve quality of life" based on green technology and low-carbon production fueled by renewable energy (op cit., 2012, 88). A five-year plan is developed to specify policies, objectives, and targets, as well as resources allocation. In 2008, Korea published her *Urban Vision for 2020* aiming to nurture eight growth engines (business, information technology (IT), tourism, green industry, biomedical, design, finance, and contents), improve urban living conditions, establish urban identity, and restore the natural environment (op cit., 2012, 102). In order to build a "polycentric mega-city," planning authority is decentralized to the local level to encourage redevelopment, densification, and mixed and transit-oriented development. Given the increasing number of immigrants to cities such as Seoul, the urban policies need to be adjusted to the cultural needs and practices of these immigrants.

Development of Taiwan took off in the 1960s when the Nationalist Government pursued the Ten Great Infrastructure Projects across the island state. Today, key urban challenges include an aging population, environmental risks, economic and spatial restructuring, sustainable development, and international competition. And the goal is to develop Taipei as a forward thinking, charming, livable, cultural, eco, cybernetic, and safe city (UDD, 2005). In 2003, the Taipei City Sustainable Development Council was established to build Taipei into a sustainable world city that pursues smart growth, recycling, and progress sharing (Taipei City SD Council, 2004). Similar to Tokyo, Ubiquitous Intelligent Taipei is a major policy for the Taipei City Government, providing residents access to various kinds of services such as transport, parking, telecare, etc. (TMG, 2010). The city has been proactive to reduce carbon emission and an interdepartmental "Greenhouse Gases Reduction Promotion Team" has been formed to execute the "Energy Saving and Carbon Reduction Plan" that involves industrial and commercial energy conservation, subsidization of environmentally friendly vehicles, waste recycling and reuse, etc. (Secretariat of TCG, 2009). And the active civil society has played an instrumental role in shaping the environmental and social policies (Lyons, 2009). However, the politically isolated island state has faced keen international competition especially with the rise of Chinese cities and has in recent years adopted more neoliberal measures in marketing the city such as tax cut for the rich, relaxing rules and regulations related to the protection of labor rights, and encouraging redevelopment through the provision of plot ratio and gross floor areas bonuses.

The largest country in East Asia is Socialist China, once an underurbanized country with a *hukou* (household registration)

system that strictly controls internal migration between rural and urban areas. However, since late 1978, the giant nation started to embark on economic reforms and open door policies, leading to decentralization of administrative functions and rapid urbanization processes especially along the coastal regions, first through the development of the Special Economic Zones (SEZs), then the 14 open coastal cities, the open economic zones, free trade zones, and development zones. The transition from a centrally planned socialist economy into one with market mechanisms has significant impacts on old urban areas in the reforming cities. The heavily used compact inner city areas that carry the historic heritage and collective memories of many urbanites were increasingly seen as problematic sites that need to be 'fixed' and redeveloped into land uses with higher exchange value such as commercial and retailing uses. Furthermore, following the practice of capitalist Hong Kong, the leasing of land use rights has become an important financial resource to fuel infrastructure development. According to Article 18 of China's Administration Law on Real Estate (1994), land premium paid by developers for the right to use land was collected by the State Treasury and when this law was revised in 2004, 30% of the land revenue goes to the Ministry of Finance and 70% to the relevant local government (Choe et al., 2008, 38). And many municipal governments have used the financial resources to provide infrastructure in order to attract outside investments to propel economic growth. And no place in China has a more dramatic urban story than the legendary 'instant city,' Shenzhen SEZ north of Hong Kong.

The country's first SEZ has grown from a city of merely 0.3 million in 1980 to an informally quoted figure of 17 million today. In three decades' time, the city's GDP had grown more than 4184 times, from RMB196 million in 1979 to RMB820 billion in 2009 (Shenzhen Statistics Bureau, 2010: 30). The first master plan of SEZ was formulated in 1980 and had a planned area of only 50 km^2 and by the third master plan, the SEZ had incorporated its two neighboring rural counties as districts and the whole city was expanded to 2000 km^2 (Office of Shenzhen History, 1997). Certain slogans summarize Shenzhen's philosophy in developing her urban realm: "Time is money, efficiency is life" and "Hong Kong's efficiency, Singapore's environment." At the turn of the twenty-first century, the dramatic growth of the city has led to constraints of the carrying capacity of the natural environment, shortage of land supply, rising costs of production, and pressing need to restructuring the economy in face of keen competition within and beyond China. To cope with these challenges, the city repositions herself as "an experimental site for integrated reforms to develop a creative and 'circular' economy," to collaborate with Hong Kong to develop newly reclaimed Qianhai as "an international financial, trade and logistics centre," to serve as a base for national high-tech industries and cultural heritage and an integrated transport hub as well as an internationally renowned tourism city by the sea (Ng, 2011: 640). The city also began to pay more attention to sustainable urban development and the quality of environment. To cope with excessive land occupation, the city adopts a new land policy to divide land uses into 'no-go' areas (44%), 'restricted development' areas within the 'ecological control line' (5.8%), 'developed' areas (32.5%), and 'suitable development' areas (17.7%).

The Shenzhen SEZ was built north to Hong Kong, a British colony between 1842 and 1997. Hong Kong's return to Chinese rule coincided with the outbreak of the Asian financial crisis and the city plummeted into a prolonged economic depression that threatened the tacit social contract between the growth-oriented government and the once politically apathetic civil society. Hong Kongers, especially the younger generation, have become much more vocal after 1997 and have seized every opportunity to exert their right to participate in deciding on the planning and development of the city. As a self-crowned 'Asia's World City,' the nondemocratic administrative Special Administrative Region government has to maintain the city's competitiveness amidst the dramatic growth of aspiring 'world cities' in China and beyond while also addressing domestic issues. Key challenges in the urban realm include escalating property prices, air pollution, solid waste management, undesirable urban living environment, as well as social conflicts arising from the intensified integration between Hong Kong and the China mainland. The current government has stepped up measures to search for more housing sites to check the mounting property prices. Following China's carbon intensity reduction target, the government has also published the city's *Climate Change Strategy and Action Agenda* in 2011. The city's latest Urban Renewal Strategy 2012 was reformulated to encourage community-initiated redevelopment efforts, that is, owners of buildings in poor conditions within identified redevelopment sites could seek the service of the Urban Renewal Authority to facilitate their redevelopment efforts. Environmental policies have also been developed to improve air quality, minimize waste, and promote recycling. New building regulations have been developed to ensure more sustainable urban form. Hong Kongers have been very proud of their core values including rule of law, anticorruption institutions, and privileging markets over government intervention. The government has to be extremely vigilant in responding to the community's concerns over the erosion of the city's core value as a result of inevitable and intensifying integration with the mainland at all levels.

Southeast Asia

Multicultural Singapore has been known for its urban affordable housing, transit-oriented transport policies, and heritage conservation efforts. In recent years, its "Active, Beautiful and Clean" (ABC) Waters program can be seen as one of the most creative and sustainable urban development policies in face of climate change and global warming. Active waters aims at providing new community spaces and bringing people to water for recreational purposes, Beautiful waters refers to integrating aesthetically reservoirs and waterways with neighborhood parks and landscape, and Clean waters concerns enhancing water quality through holistic management of the city's surface runoff through natural cleansing elements (National Waters Agency, 2011). The ABC Water Master Plan includes 100 project sites and efforts have been made to promote the ABC Waters Design concepts such as green building design elements (Building and Construction Authority's Green Mark Scheme has been in place since 2005), artificial water bodies creation, and flood control of storm water runoff at source into the

projects undertaken by public agencies and the private sector (op cit., 2011: 4). The Prefabricated Extensive Green Roof Tray system is also developed to provide cost-effective 'instant' green roofs (BRI, 2011). Indeed, all Singaporeans are served by modern sanitation and the government manages the water loop in an integrated manner, with two-thirds of the city functioning as local catchment areas collecting and storing rainwater. And a 3P (people, public, private) Partnership Approach to engage the community to take ownership of the water bodies or to encourage the private sector and grassroots organizations to organize or participate in activities launched at the ABC water sites (op cit., 2011: 6). Unlike her more developed Asian neighbors, a vocal and resistant civil society is absent. Nevertheless, the planning authority has encouraged public participation and the formation of focus groups to solicit information, the typical state-centered Singaporean style.

Another rapidly growing Southeast Asian country is Malaysia. Like Singapore, Malaysia is a multicultural and multiracial society, although its 'bumiputras' ('sons of the soil') policy favors the Malays over the Chinese and Indians and has indeed been increasingly challenged. Under the leadership of the then Prime Minister Mahathir bin Mohamad, Malaysia's economy has moved toward high technology-based industries. To facilitate this restructuring process, the Multimedia Super Corridor project between Kuala Lumpur City Centre and the International Airport with the Cyberjaya intelligent city, the country's 'Silicon Valley', at its center was constructed. Within the self-contained Cyberjaya are central flagship enterprise, commercial and residential zones with ecofriendly and low-density developments that aim to preserve nature and maintain an excellent quality of life (Yuen et al., 2006: 238). The target is to locate 500 IT companies in the city by 2020 (op cit., 2006: 240). And this initiative demands interrelated changes in the urban realm: providing urban services that are on a par with international standards, redesigning curriculum and education system, and issues of urban poverty as high-tech industries typically employ fewer workers.

The capital city of Thailand, Bangkok, has suffered from traffic congestion, air pollution, water pollution with untreated wastewater discharged into the public sewers, land subsidence, and growth of slums. The city develops the Metropolitan Master Catalogue for 50 districts to guide physical development through the formulation of a plan based on local strategic agenda, regulations, and database (Cities Alliance, 2005: 58). Through the Bangkok Comprehensive Plan, air pollution is tackled through the development of mass transit, establishment of vehicular emissions standards, promotion of nonmotorized transport, and the use of alternative fuels (op cit., 2006: 59), as well as mobilizing volunteers to spot 'black smoke vehicles' and to 'protect (monitor) air quality' in factories and construction sites. To arrest land subsidence and damage to infrastructure, policies and regulations on groundwater extraction have been developed to control effluent standards and wastewater treatment plants to protect water quality (op cit., 2006: 59). Community and volunteer efforts have also been mobilized to identify solutions to garbage dumping, wastewater discharge, and to develop environmental projects.

Indonesia experienced a dramatic decentralization push in the late 1990s, a result of the Asian financial crisis and the widespread demands from the civil society. With the implementation of Law 25/99 (Fiscal Balance between Central and Local Governments), except national defense and security and foreign affairs that remain under the purview of the central government, local governments are responsible for the judicial system, fiscal and monetary matters, and macroeconomic planning. Various functions, financial resources, as well as 2.3 million staff from 4000 government departments and 20 000 public facilities were transferred to local governments within 1 year (Turner et al., 2003). At the same time, democratization was initiated allowing for the direct election of national, provincial, and district governments and both trends have led to the need in revising various rules and regulations and a renewed understanding of the roles of the private sector in the development process (Sarosa, 2006). Despite the good will of the central government in decentralization, the transferred responsibilities have become challenges to the local governments that are often short of administrative skills and revenue sources. Inspired by Singapore, the mayor of Tarakan, a city in East Kalimantan, devised 'three pillars' of development: human resource development, rule of law, and provision of people's welfare as economic development (op cit, 2006: 174). The mayor mobilized the community to manage and conserve the environment and turned down an investment to exploit coal in the city (op cit., 2006: 174). In the capital city Jakarta, high urbanization has resulted in traffic jam, flooding issues, waste management, health services, and education and employment opportunities, especially for the poor. To tackle the key issues, the municipal government has endeavored to adopt the transit-oriented development concept and develop mass transit while reducing the road ratio to 10%. It has also adopted an integrated flood control system to restore the riverside environment, enhancing the drainage systems and optimizing the functions of urban rivers.

China's reawakening has stirred other transitional economies in Asia such as Vietnam and *doi moi* (renovation) was initiated in 1986. And *doi moi* has unleashed economic reforms, urban planning efforts, urbanization, and calls for decentralization. Similar to other early urbanized countries, providing infrastructure to sustain economic growth and a basic quality of life in the urban realm has been a key challenge, especially when the central government continues to control the budget and there is a general lack of land use planning and development enforcement capacity. A high percentage of motor cycles (4 million in Hanoi alone) contributes to heavy traffic congestion, and haphazard urban and industrial developments present pollution problems and difficulties in rationalizing land uses. Not all urban areas are connected with sewage, and water and electricity supply is also an issue. The lack of relevant human resources to tackle these issues through policies, legal means, and administrative measures in a holistic manner has worsened the situation. As a result, foreign official development assistance has been instrumental in building the country's institutional capacity in the urban realm. For instance, the project of Nhieu Loc-Thi basin canal of Ho Chi Minh City, funded by the World Bank and Ho Chi Minh People's Committee, involving the separation of sewage and storm water drainage, lessen the threat of flooding, promoting economic growth and enhancing public management of

private entities. Through extensive public engagement, the project not only helps improve the quality of life, but also has boosted the local capacity to plan, construct, and manage this type of urban projects (Lang, 2006).

South Asia

In South Asia, the key challenge is related to the mode of governance and the lack of resources at the municipal level to provide, operate, and maintain various urban services. The central government plays an important role in South Asian countries. In India, the central government decentralized more powers to urban and rural local government bodies through the 73rd and 74th constitutional amendments in 1992 (Choe et al., 2008: 4). The 28 state governments are responsible for urban development, management, and service delivery but the central government initiates policies and programs of national importance such as poverty reduction, housing for low-income urban households, and infrastructure. State governments reserve the rights to allocate responsibilities and functions to local governments through tax sharing or expenditure transfer (Mathur, 2006: 137). In 2005, the Indian government launched the Jawaharlal Nehru National Urban Renewal Mission in 63 cities to improve and expand the economic and social infrastructure of cities, to ensure access to basic services including housing and land for the urban poor, to initiate wide-ranging urban sector reforms, and to strengthen municipal government functions (Choe et al., 2008: 6). In the capital city, the Delhi Municipal Corporation started the Bhagidari initiative in 2000 to ensure that utilities providers cater to the needs of the residents through interactive meetings with citizens' groups such as welfare associations, private sectors, industrial organizations, and school welfare committees. While the central and state governments provide grants and finance to aid infrastructure development at the local level, the key challenge, nevertheless, is to boost financial capacity of local municipal governments to maintain service delivery functions such as encouraging public–private partnership, sale of land use rights, and application of cost-recovery principle to construct urban infrastructure (Choe et al., 2008: 141).

A case in point is the Ahmedabad Municipal Corporation (AMC), established in 1950 under the Bombay Provincial Municipal Corporation Act, 1949. Key functions of AMC are to provide water and urban amenities and to run hospitals, medical colleges, and local transport; its main sources of revenue are taxes on properties and goods that enter the city. To improve the collection of the property tax, the AMC disconnected water supply and drainage lines and other 'name and shame' measures on nonpayment of property tax, and a valuation manual and database are developed to allow professionals to check and collect taxes on goods that enter the city (op cit., 2008: 145). AMC even pioneered in the issuing of municipal bonds to raise the necessary funds for financing the much needed infrastructure. Increased cash flow has helped improve the delivery of services, which in turn restores a more trusting relationship between the government and citizens. Securing financial resources enables the city to launch environmental improvement and sustainable transport initiatives. Ahmedabad has the 'only successful running Bus Rapid Transit' in India but research shows that the urban poor do not use formal public transport. In fact, 72% of women and 51% of men in the low-income households 'traveled less than 3 km' and rely on nonmotorized transport, with 60% poor women walking to work (Mahadevia et al., 2012: 3–4). Making public transport affordable is an urban challenge in these situations.

Bangladesh is a secular democracy but her mode of urban governance has a lot to be desired because of inadequate decentralization, a general lack of local political leadership, and numerous interrelated urban problems such as weak economic growth, social polarization, insufficient utilities, transport, education, health and leisure facilities, slums and squatters and the accompanying social distress in law and order, or the conservation of heritage and culture (Islam, 2006: 54–55). The City Corporation Ordinance or the Pourashava Ordinance constrains the power of local city corporations, rendering them little authority to mobilize resources or develop programs of commercial interest (op cit., 2006: 58). Hence, international organizations have been instrumental in building her capacity in urban governance. For instance, the United Nations Development Programme and the UN-Habitat had assisted in the country's urban policy formulation and drafting of master plans for cities. The World Bank has supported water supply and sewage treatment and transport projects in Dhaka and the Asian Development Bank has provided assistance in service deliveries as well as enhancing urban governance. In the country's second largest city, Chittagong, besides her Nobel laureate Professor Muhammad Yunus who first developed the concepts of microcredit and microfinance and later developed the Grameen Bank, the city corporation has been instrumental in establishing sustainable social, economic, and environmental projects. With the able leadership of the then three consecutive time mayor Chowdhury, the city developed self-sustaining programs on infrastructure development, garbage disposal, parks and playgrounds maintenance, water supply to poor neighborhoods, power plant installation, solid waste recycling plant, shopping malls, hospitals and health care centers, and various education facilities from kindergartens to colleges (op cit., 2006: 58–60). Mayor Chowdhury's downfall in 2007 as Bangladesh changed to an army-backed government highlighted the problem of the country's centralized mode of governance. It also highlights the challenge of political stabilities in the developing countries.

Conclusion

Perhaps one defining characteristic of the urban realm in Asia is the important role of the state, especially the central government, and it was coined as "the developmental state" (Wade, 1990; Woo-Cummings, 1999). Urban development at the local level is dependent on the degree of decentralization set by the central government. Deployment of fiscal and human resources by growth-oriented states to develop and tailor-made appropriate and integrated policies is undoubtedly the key to success in the developed cities. Traditionally, the Asian states and the private sectors have developed strong relationships such as the roles of zaibatsu in Japan, chaebols in Korea, the state-led market system in Singapore, not to mention the unique socialist market economic system

in transitional China. The strong relationships between the state and the private sector have produced unique political economies over time, which are different from a typical free market economy that emphasizes on the existence of a level-playing field for all economic actors. However, accelerated globalization and the need for more foreign investments have led to the launching of 'neoliberal' economic policies. Yet, one may argue that the nature of Asian 'neoliberalism' is different from her Western counterpart. Instead of rolling back the state welfare system and rolling out market-driven reforms, Asian 'neoliberalism' is more about 'weakening' certain state–private sector relationships in order to nurture the growth of a more inclusive market.

Obviously, developed or developing, the green and the brown agenda continue to be relevant as these cities struggle to address domestic issues including social polarization, poverty, and degrading environment and at the same time, the need to compete for overseas investment and international trading links and status. The green urban policies are often launched to boost global competitiveness and these include measures to ensure disaster resilience, development of renewable energy, ICT infrastructure, and green urbanism to counteract climate change and global warming. For the developing Asian countries, urban policies are related to enhancing the capacity of local governments to tackle pollution issues and to provide adequate infrastructure to satisfy basic needs. Indeed, evolving social contexts have complicated the launching of these various policies. Domestic and international migrants into cities have created challenges to the local governments. While these migrants usually make great contributions to urban development, accommodating their multifaceted needs and harmonizing their relationships with the local communities and building a truly multicultural society could be daunting tasks. The developing and maturing civil societies in Asia are slowly (re)shaping the various urban policies, challenging the state–private sector nexus in the more developed countries, and substituting inefficient government or weak private sector to satisfy their socioeconomic needs in the less developed ones. The Asian urban drama will most likely continue to grip the world's attention as the twenty-first Pacific Century unfolds.

See also: Cities Under Postsocialism; Citizen Participation; Eco-cities and Sustainable Urbanism; Economic Transformation: From Central Planning to Market Economy; Geography: Overview; Globalization: Geographical Aspects; Just City; Planning Issues and Sustainable Development; Planning Theory: The Good City and Its Governance; Urbanization in China.

Bibliography

Amsden, A.H., 1992. Asia's Next Giant: South Korea and Late Industrialization. Oxford University Press.

Asian Development Bank (ADB), 2011. Competitive Cities in the 21st Century: Cluster-Based Local Economic Development.

Asian Development Bank (ADB), 2012. Urban Development. http://www.adb.org/themes/urban-development/main (accessed April 2013).

Building Research Institute, 2011. Peg Tray. http://www.bri.sg/product/peg-tray/ (accessed April 2013).

Bureau of Environment (BoE), Tokyo Metropolitan Government, 2009. Best Practices of ANMC2. Available at: http://www.anmc21.org/english/bestpractice/Tokyo9.html (accessed April 2013).

Bureau of Urban Development (BUD), Tokyo Metropolitan Government, 2011. Urban Development in Tokyo. BUD, TMG, Tokyo.

Choe, K., Laquian, A., Kim, H., 2008. Urban Development Experience and Visions: India and the People's Republic of China. Asian Development Bank.

Cities Alliance, 2005. Liveable Cities: The Benefits of Urban Environmental Planning. Cities Alliance, ICLEI and UNEP.

Clark, A.L., Fujiki, M.D., Davidson, M. (Eds.), 2009. The Urban Transformation in Asia: Policy Implications of Decentralization. East-West Center.

City of Taipei: The Urban Development Context, 2005. Urban Development Department (UDD).

Islam, N., 2006. Bangladesh. In: Roberts, Kanaley (Eds.), Urbanization and Sustainability in Asia. ADB and Cities Alliance, Manila, Philippines, pp. 43–69.

Kim, H.M., Han, S.S., 2012. Seoul. Cities 29, 142–154.

Lang, N.T., 2006. Viet Nam. In: Roberts, Kanaley (Eds.), Urbanization and Sustainability in Asia. ADB and Cities Alliance, pp. 369–401.

Lyons, D., 2009. The two-headed dragon: environmental policy and progress under rising democracy in Taiwan. East Asia 26, 57–76.

Mahadevia, D., Joshi, R., Datey, A., 2012. Sustainability and Social Accessibility of Bus Rapid Transit in India. UNEP Riso Centre, Roskilde.

Mathur, O.P., 2006. India. In: Roberts, Kanaley (Eds.), Urbanization and Sustainability in Asia. ADB and Cities Alliance, pp. 135–154.

National Waters Agency, Singapore, 2011. ABC Waters Design Guidelines.

Ng, M.K., 2011. Strategic planning of China's first special economic zone: Shenzhen city master plan (2010–2020). Planning Theory and Practice 12 (4), 638–642.

OECD, 2012. Urban Policy Reviews. Korea.

Office of Shenzhen History, 1997. Establishment and Development of Special Economic Zones in China: Shenzhen (in Chinese). Chinese Communist Party History Press, Beijing.

Secretariat of Taipei City Government, 2009. Taipei to Join the CDP Cities Programme. Available at: http://www.sec.taipei.gov.tw/ct.asp?xItem=185897&ctNode=31198&mp=101001 (accessed 22.02.11.).

Shenzhen Statistics Bureau, 2010. Shenzhen Statistics Yearbook 2009. China Statistics Bureau.

Skocpol, T., 1982. Bringing the state back in. Items 36 (1–2).

Taipei City SD Council, 2004. Taipei City's Vision, Declaration and Commitment for Sustainable Development. Available at: http://tsd.utrust.com.tw/hope/index.htm. in Chinese, (accessed 16.04.04.).

Tokyo Metropolitan Government, 2010. UI Taipei. Available at: http://english.taipei.gov.tw/ct.asp?xItem=1553185&ctNode=8516&mp=100002 (accessed 17.01.11.).

Tokyo Metropolitan Government, 2011. Principal Policies of the Tokyo Metropolitan Government: Tokyo Vision 2020. Available at: http://www.metro.tokyo.jp/ENGLISH/PROFILE/policy01.htm (accessed 01.05.13.).

Turner, M., Podger, O., Sumardjono, M., Tirthayasa, W., 2003. Decentralisation in Indonesia: Redesigning the State. Asia Pacific Press, Canberra.

UN Habitat, 2011. State of the World's Cities 2010–2011—Cities for All: Bridging the Urban Divide.

Wade, R., 1990. Governing the Market: Economic Theory and the Role of Government in East Asia. Princeton University Press.

Waley, P., 2007. Tokyo-as-world-city: reassessing the role of capital and the state in urban restructuring. Urban Studies 44 (8), 1465–1490.

Waley, P., 2012. Japanese cities in Chinese perspective: towards a contextual, regional approach to comparative urbanism. Urban Geography 33 (6), 816–828.

Wang, Y., Wang, Y., Wu, J., 2009. Urbanization and informal development in China: urban villages in Shenzhen. International Journal of Urban Regional Research 33 (4), 957–973.

Woo-Cummings, M., 1999. The Developmentalist State. Cornell University Press, New York.

Yuen, B., Ahmad, S., Ho, C.S., 2006. Malaysia. In: Roberts, Kanaley (Eds.), Urbanization and Sustainability in Asia. ADB and Cities Alliance, Manila, Philippines, pp. 223–243.

Relevant Websites

http://www.adb.org/ – Asian Development Bank.
http://www.asean.org/ – Association of Southeast Asian Nations.
https://www.asian-studies.org/ – The Association for Asian Studies.
http://www.unhabitat.org/categories.asp?catid=9 – UN-Habitat.

Urban Policy in Europe

Patrick Le Galès, Sciences Po, CNRS, Centre d'études européennes, Paris, France.

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article argues that urban policies have a long inheritance in Europe as they were invented in medieval trade cities all over the continent. The long-term territorialization of European societies and the importance of the state explain why urban policy is particularly concerned with places in Europe. But urban policy are not only explained by urban factors, they are one type of public policies among others hence the emphasis on public policy analysis and capitalism to understand the transformation of urban policy.

Urban policies are public policies and Europe is the land of public policy. Historically, urban policies in Europe started with the institutionalization of power in Greek cities and the Roman Empire. Taxes were raised; some specialized experts and administration emerged; ideas were put forward; protection, control, and war were at stake; securing food supply or water was central; and elites invested resources to build, plan, develop cities, and control their population. In Europe, urban policies developed with the coming age of the medieval communes organizing those flourishing cities which, along the Baltic and the North Sea on the one hand (St Petersburg, Hamburg, or Amsterdam, for instance), the Méditerranée on the other hand (i.e., Venice, Florence, or Barcelona), and those in between (Bruge, Cologne, Prague, or Lyon), became the center of the first capitalism, that of merchants and traders. Within those fast developing cities, new ways to govern and to implement urban policies emerged: statistics, bureaucracies, planning, public private partnership to develop new neighborhoods, and forms of policing. Then, the conquest, organization, and control of space became a classic task states develop in the process of state making and state institutionalization.

Like any public policy, urban policies are always historically and geographically located in a given society and at the same time influenced or structured by large-scale processes or paradigms. Politics is part of urban policies, hence the importance of power relations, forms of capitalism, and state. Contradictions are essential to understand urban policy: not only the contradictions of the capitalist order where public policies aimed at dealing with uneven developments and inequalities while supporting capitalist accumulation and competition between cities but also the contradictions between organizations, laws, and agencies. Urban policy must be understood as a policy with the classic question raised by Harold Laswell in the US in 1936: *Who gets what, when, and how*? An urban policy is always an attempt to change existing situations, a set of power relations more or less institutionalized, and is also an experiment as suggested by Dewey (1927) and the pragmatists, hence the political dimension. However, by contrast to the pragmatist urban scholars have emphasized on structural political dimensions: conflict, political order, class, coercion, resistance, and power relations. As any public policy, an urban policy combines: (1) some principles, ideas, representations, values, and goals to be achieved; (2) actors, individual or collective with unequal resources (money, expertise, legitimacy, and networks), more or less organized in different ways; (3) institutions, budgets, rules, and instruments; and (4) results. Urban policies are in part inherited, but can also be constructed and the label may reveal strategic actors' activities rather than substance. Implementation is central. Urban scholars too often consider urban policies as specific: they fall into the sector centered trap as do environment, social policy, gender, or economic policy specialists. Urban policies, including planning, are public policies which are for the most part similar to the rest of public policies.

Analyzing urban policy is fraught with difficulties for different reasons: Firstly, defining the urban is no small problem. Long gone are the days when a city was surrounded by walls with different social structures and rights for citizen. Beyond classic definition of cities in terms of concentration of people, buildings, exchanges, and land, i.e., forms of density of individuals, networks, or the built environment, many now see a gigantic urban world where no one nowhere is immune from the urban, a global urban world, i.e., a planetary uneven urbanization (Brenner, 2013: p. 109). If the urban is everywhere, it is nowhere and cities then do not matter in that perspective. The question of boundaries is never solved. It follows that critical geographers may analyze urban policy as either irrelevant, or in constant flux, as an unstable assemblage of mobile policies. This has proved useful to mock fads in urban policy such as the 'creative city' discourse promoted by policy entrepreneur, selling his 'best practices' to cities all over the globe and in Europe alike (McCann and Ward, 2011). But neither urban policies nor cities are so fluid and the assemblage metaphor is often not so relevant for urban policies.

Alternatively, urban policy is explained by the transformation and the requirements of capitalism. Here the argument can be developed in more or less functionalist way. Neoliberalism as a global paradigm has been used in all sorts of more or less rigorous ways to explain urban policy change (Brenner, 2013). Others focus on urbanization processes and convergence. If there is a general theory of urbanization, for instance a Marxist-inspired theory (Harvey, 1985; Brenner, 2004 among others), urban policy is understood as either the attempt by the state to accompany or limit the problems associated with the urbanization process, or, as Brenner put it urban policy is defined as "encompassing all state activities orienting towards the regulation of capitalist accumulation" (p. 2). Urban policy in London, Helsinki, or Prague may be

 http://dx.doi.org/10.1016/B978-0-08-097086-8.74058-X

seen as an attempt to deal with the spatial contradictions of capitalism.

There is another problem linked to the definition of the urban: while it was classically understood in relation to fixity, as a place in space, a booming research current has shown the importance of physical and virtual networks, circulations, and links. Milan is Milan but is also structured by Senegalese neighborhoods of Dakar through regular in and out migration, to London and New York through financial circulations, and so on. Images, ideas, representations of cities travel, and the virtual cities of science fiction writers, movies, or games are also part of the urban world.

The tension between state and cities is a powerful reminder of the dynamics of rescaling. The scale question was brought back on the agenda by critical geographers, most notably Neil Brenner (2004), and political scientists (Keating, 2013); so, for instance, in urban policy terms, one cannot just describe policy initiatives developed by central governments. Urban policies are initiated and implemented at different levels closely articulated, enmeshed, or in competition, and with many contradictions: neighborhoods, cities, metropolitan governments, regional or federal public authorities, the state, the EU or influenced by international organizations (Le Galès, 2002). This does not occur at random; there is a politics of rescaling where interest groups mobilize to locate resources and policies. Local welfare policies are a typical example of this general dynamic (Kazepov, 2010).

Urban Policy: How Does Europe Matter?

Urban policy in Europe makes sense in relation to some particular features of Europe. Firstly, European societies are deeply territorialized, cities have a long history, and Europeans have low rate of transnational mobility.

Urban policies in Europe mainly deal with places: European societies are deeply territorialized, levels of transnational mobility are lower than in other parts of the world and only increase slowly. With the exception of Britain, many urban Europeans have dense social networks in cities and when they move, they often come back where they were before, in particular in the south (Andreotti et al., 2013). By contrast, American urban policies, for instance, make sense if one remembers the origins of colonial towns, the industrial development of the North East, the Western Frontier, the race issue, the federal system, and a particular understanding of freedom which makes public policy often illegitimate and by contrast, mobility a pillar of individual freedom (Imbroscio, 2012). But if Americans are born to run, Europeans are rather born to stay, or to move slowly.

Historically, Europe has been created and shaped by cities (Pinol, 2012; Le Galès and Therborn, 2010), and the long history of urban Europe made of medium-sized regional capitals has been well accounted for (Hohenberg and Lees, 1985). Urban Europe came into full development during medieval times, thanks to the development of the first capitalism of trade but in relations to existing settlements inherited from the Catholic Church or the Roman Empire. Cities gave rise to modern Europe and states came of age in relation to them. The European urban structure has been relatively stable ever since, with two important caveats: the institutionalization of the modern state (the making of national and/or imperial capital and the role of various wars), and the massive disruption of the industrial revolution in Britain, in particular, and also in Belgium, Eastern France, Germany, and Scandinavia.

European societies are deeply territorialized societies because of the tradition of stable peasants on their land and the long history of medium-size cities (Kaelble, 1984; Bagnasco and Le Galès, 2000). Britain, where industrialization shaped the urban structure more than anywhere else, is an exception. The political order of modern Europe is the result of the interaction between this longer standing structure of cities and the latter emerging as nation-state territorial units (Tilly, 1990). In due course, 'voracious states' consolidated with or against 'obstructing cities' (Tilly and Blockmans, 1994). The making of the modern state and the coming age of the second (industrial capitalism) marked the end of the golden age and autonomy of occidental cities. Urban policies gained prominence in those medieval prosperous cities, which were the centers of trade. The importance of the urban government was stressed as the city developed upon the movement of medieval 'communalization' (i.e., communes being formed through acquiring a *charter*). Cities became institutionalized associations, autonomous and autocephalous, active territorial corporations characterized by autonomy and capacity for action toward the outside (the lord, the prince, the abbey, the bishop, the state, the emperor, and rival cities), and led by urban officials. Medieval urban governments, controlled by the urban bourgeoisie, developed democratic institutions, and commerce left their mark through the edification of monuments symbolizing their power: squares, town halls, beffrois, or campanile. They invested resources in city centers. Urban policy all over Europe first inherits a conception of power, the centralization of public buildings, an understanding of politics (including protest), a rather encompassing view of the city as a political beast with long-term legitimacy resources associated to the elected representative, for instance, the mayor in some countries.

The long history of state making is replete with stories of national elites allying or fighting those burghers and their administrators. From the early seventeenth century, as Foucault explained (2007), the modern idea of police and government emerged. Urban policies were conceived in relation to the ministries of police and the framing of society by state dispositifs (Miller and Rose, 2008; Dean, 2010). The rise of nations and modern states, more or less operational in Western Europe by 1870, paved the way for powerful processes of nationalization of European societies and differentiation among them through urbanization, politics, culture, labor market mobility, industrialization, production, consumption, the making of a more class-structured society, wars, and the one hundred years development of public policies: infrastructure, education, welfare state, labor market, or urban policy. In industrial cities, for instance, in the UK and also in Germany, France, and Scandinavia, the scope of social problems led elites in urban government to pioneer policy programmes in housing, planning, basic elements of welfare, and education (De Swaan, 1988), and hygienist concerns led to the 'haussmanization' movement of city rebuilding together with the emergence of local public goods or to the rise of modern planning. Urban governments played a key role in providing

basic utilities and services such as water, sewage, street lighting, later gas and electricity, firemen and transport, not to mention slaughterhouse. This development was diverse, fragmented, contested between a conservative petty bourgeoisie and the municipal socialism movement, and more consistent in the North of Europe than in the South. Most local government in Europe gained legal recognition in the second part of the nineteenth century. Gradually, a professional local bureaucracy emerged to deal with those developments.

Later, in most of the post-1945 period, within the social democratic compromise of most European states, the role of urban government was structured by center–periphery relations (Goldsmith and Page, 1987, 2010), i.e., in national terms and as part of local government. Urban governments differed in Europe because each country had a different constitutional setting, different rules, different public finances systems, different political systems and traditions, different organizations to provide services, and different police forces. Urban policy was mostly understood in terms of infrastructure provision. European cities have been structured by large amount of public expenditure, translated in terms of welfare services, education, buildings, roads, housing, etc. Urban policy was in many ways part of the larger development of the welfare state and the modernization for industrialization. Urban policy was on the other hand related to modernist planning promoted by rational heroic planners against narrow local interests from the path. The reconstruction of cities after the war in Cologne, Dresden, Coventry, or Le Havre gave ample evidence of this modernist urban policies, accentuated in the case of Eastern Europe under the socialist regimes as evidenced in Bucarest or Varsaw.

Urban Policy: Inheritance, Institutionalization

Urban policies are for a large part inherited and sometimes, the result of ad hoc combination. This is central in the European case because of the long standing of the autonomous cities and municipal government on the one hand and the particular strength of states and public policies on the other, including in postsocialist cities (Stanilov, 2007). Urban policies deal with the inheritance from the past (city centers, infrastructure, renovation of working class/immigrants estates) with a mix of social, cultural, economic, and policies to redevelop the material side of the cities, buildings, housings, and utilities. Inheritance means institutions, finances, banks, professions, ministerial bureaux, urban forms, material environments, agencies, ways of doing things, knowledge, budgetary rules, clienteles, professional experts, and interest groups. Urban policy in Europe is strongly institutionalized with rules, budgets, interest groups, policy instruments, and political organizations rooted in the urban policy sector. Problems emerging on the political agenda are in part the result of this past: renovating and protecting old city centers (policy for centro storico in Italy), modernizing centers with strong local bourgeoisies, i.e., the major issue in European cities of heritage maintenance, or dealing with the long-term decline of industrial cities or neighborhoods (a few shrinking cities in oriental Europe or Northern England like Sheffield, the suburban large social estates in Britain, France, or in Roma), or renovating the utility infrastructure.

There is a huge literature about urban policies. Unfortunately, as often in the policy world, a good deal is written about specific initiatives, the 'new'policy of the day analyzed in urban policy institutes. A lot of 'new initiatives,' specific program, and much heralded 'urban new project' are no more than the recombination of those different elements, describing government initiatives, sometimes evaluating them under limited parameters. A classic New Public Management recipe consists for political leaders in announcing new policy initiatives all the time in order to create movement, pressure for change, complete lack of transparency, and incapacity to evaluate the impact, i.e., another way to play the blame game (Hood, 2011). At worse, the influence of public choice models lead to rankings, the search for best practice, and normative best recipes which make sense for neoclassic economists but not for social scientists, assuming there is more to policy goals than optimizing economic development.

In most cases, not always, urban policies in Europe nowadays are in part a continuation of what was done before and is deeply institutionalized from one country to the next and from one city to the next, i.e., investment in key infrastructures, transport and utilities (could be privatized in particular in the UK), renovation of the city center including a strong cultural drive and marketing, forms of planning, recreation of public spaces, high profile urban projects and buildings, policing (controlling marginal groups, the poor) and maintaining public order, social policy, economic policy to promote innovation and support to firms and finally policies to renovate neighborhoods in crisis, mostly in outer cities estates but sometimes in the center (Brussels), with a mix of integrated policies and some sort of integrated approach for specific places.

Urban policies do not change all the time in relation to any new idea in the wind but the implementation process shows strong uncertainty and diversity (Crosta, 2009). And the ideas are also relatively stable. Stockholm, Helsinki, or Copenhagen are characterized by long-term economic, cultural, and social urban strategies developed in tension between state and urban elites. Urban policies in London are still characterized by the need to develop the British capital, to account and deal with the interest of the City of London, to develop infrastructure, and to renovate the eastern part of London, including the former docks. By contrast, there are reasons and a particular set of social and political relations to explain why urban policy in Napoli or Marseille has achieved little over the past 30 years. There are increasingly international powerful economic interests playing a role in urban policy, but powerful interest groups within cities, including voters, should not be underestimated either.

Historically, three countries had clearly defined or labeled urban policy: the Netherlands, where cities created the state, Britain, and France. In the 1980s and 1990s, not without connection with the making of European networks supported by the EU, those national policies became more mixed, disseminated all over Europe and became more similar, encompassing all these different dimensions. Kazepov's important edited volume on cities in Europe (2005) clearly

provides examples of the characteristics of urban policies in Europe.

Protests, Conflicts, and Implementation Failures

Urban policy, like any public policy, is a way to manage conflicts within society, to contribute to make and to legitimize a political order. They have to deal with interest groups, protest, resistance, and many implementation failures. The obsession with 'new urban policies' often reveals frustration with lack of results and failures.

This became obvious in the post-1968 period; when the 'urban question' took a different turn, Urban movements participated to class struggles, together with unions but outside factories (Castells, 1977). Urban mobilization concurred to claim for public goods, and enlarge the quest for better working condition coming from the workers movement to issues of quality of life and welfare conflicts. They also massively contested many large prestigious urban projects associating urban political elites and private developers and put forward alternative ways of live, including massive squatting movements in Swiss, Dutch, and German cities. This led to the development of urban policies addressing issues of historical centers renovation, environment, deindustrialization, race, and protest. More diverse urban policies also led to more interest groups, more organizations, more agencies, and more contradictions between different elements of urban policies within a territory.

Empirical researches on European urban conflicts, policies, and mobilizations have increased in the last years (Vitale and Podesta, 2011): neighbors, inter-ethnic, religious, xenophobic, and class conflicts including the most progressive and the most conservative interest groups and mobilization. Classic dynamics include contention for collective goods, reactive conflicts on urban planning, and contentious quest for public recognition. Current forms of urban contention seems independent from neoliberal policies, but affected by classical problems of private property, migration, eviction of poor neighborhood, lifestyles and individualization, density, commuting, the use of public spaces and transports, the 'entre-soi' of upper-middle classes, territorial inequalities and equalization claim-making, and localization of polluting infrastructure.

But neoliberal policies play a role. It is an ambivalent, twofold role (Brenner, 2013). On the one hand, neoliberal concerns deeply influence the means and goals of urban mobilizations. Some of the emerging actors in these mobilizations point to commodification and market virtues (critical consumerism, raves, craftsmen informal markets, third-sector lobbying for quasi-market). On the other hand, neoliberalism is attacked with distopic narratives, and collective action is organized against neoliberal policies with a major role played by political parties, and not only by grassroots organizations. One issue facing social actors is the difficulty to identify responsibility in urban policy, to blame actors: there are many experts, a strong judicialization; difficult translocal coordination; solution shopping, and some combination of policy ideas. Impacts of urban conflicts on neoliberalization relate to three main mechanisms (coalition building, emulation, and brokerage). Conflicts produce the weaving of alliances with some decision makers and some public bodies, changing power relationships, naming the public role of urban utilities and collective goods, thus affecting policy outcomes.

Classic public policy analysis (which is not rational choice) has also emphasized that policy implementation is central and that it is particularly central for urban policy for two reasons: the goals are usually very general (building an environment friendly, socially inclusive, economically dynamic, democratic, and culturally vibrant city region) and many interest groups are in competition. Urban policy has to be understood in the details of the operationalization (policy instruments) and implementation (Pressman and Wildavsky, 1974) where the winners, the losers, and the real goals are defined in the process of implementation. By contrast to large generalization, urban policy implementation often reveals contradiction, power relations, and weak top-down capacity of coercion (Laino, 2012). In most countries, the coherence of urban policy is an illusion created by discourse (and discourses are important). Often several urban policies, in contradiction with each other, are developed at the same time, with different organizations, different professional groups, or different levels of government in charge. Urban scholars too often take for granted national plans, and scenarios and maps to analyze urban policy when looking at organizations, rules, budgets, and interest groups is usually far more precise to explain the policy and its results. For instance, the renovation of historic city centers has been going on for decades in many countries, and the fact that the same policy is now labeled cultural policy, or policy to attract tourists, or social mix policy does not change much about the substance.

The point is not to argue that as far as European urban policies are concerned, *plus ça change plus c'est pareil* but to stress the fact that many policies are not implemented, that societies encompass many collective actors with resources mobilized to prevent change to take place, that urban policies are political activities thus include political actors who, at the same time, defend their interest and their constituency while pursuing some elements of common good. Public policy activities include important cognitive dimensions to label and 'sell' the policies that have to be critically examined that by contrast to strong constructivist views underlining ongoing changes or fragile assemblages always subject to changes; urban policies in Europe are also characterized by an important degree of institutionalization (rules, organizations, interests, budgets, ways of doing things) that are inherited and change slowly, due to lack of consistency, layers of policies added over the years, contradictions, and implementation failures. Beyond the labels, the goals, the substance of the policies are rather stable with a given set of repertoires of action.

Urban Policy Change

What has changed over the last decade is a growing emphasis on the environment, social cohesion and social mix, policing and the international dimension, both economic competition and integration of foreign born populations.

Changes do take place: the rise of environmental concern in urban policy is one example. All over Europe, European urban

policies, national urban policies, and policies development by urban governments include a vast array of programmes to limit carbon emission, and take into account environmental concerns. The political ecology of cities (Swyngedouw et al., 2006; Béal, 2011) has become an institutionalized field of urban policy, supported on the one hand by the rise of green party councilors in many cities and on the one other hand by giant utility firms (water, waste disposal) or information technology (IBM and the smart city) eager to sell massively expensive technological solutions to deal with water pollution, waste disposal, or the optimization of transport. Those policies include investment in recycling, limits to individual car transportation (in favor of bicycles or collective transport), reducing energy consumption of public building and activities, improving school food with local producers, setting environmental norms for building, or, more visibly, creating much heralded and visible 'econeighbourhood' much loved by ecologically concerned middle classes. Depollution of industrial land, brownfields, or the air is also on the card. Groups are enrolled, experts and consultants mushroom, national agencies are created, European networks of cities take the issue on board, political entrepreneurs rise, all sorts of citizen groups, scholars, and NGOs push in this direction.

Increasing spatial and social inequalities within European cities have made the question of poor neighborhoods more acute, a problem made more visible by the presence of immigrants. Various types of policies have been developed, usually with a strong integration element, in order to deal with those neighborhood: area-based policy, neighborhood policy, together with urban renewal have been tried out in particular to prevent riots, and systematic exclusion, or worse case of social segregation (Andersson et al., 2010; Epstein, 2012), in particular, in urban peripheries.

Policies do not change easily. Ideas about new policies and new concept travel change all the time and may have some contribution to changes (Peck, 2010; Theodore and Peck, 2012). In that sense policy mobility does matter and contributes to urban policy change but the more institutionalized the urban policy, the more difficult it is to bring changes based upon a new urban policy idea. Policy transfers are common in public policies but their impact is often limited. The idea of the compact city, developed in the Netherlands, gained prominence in debates all around Europe to limit suburbanization. Social mix policies have been at the core of fierce debate in urban research: do they contribute to the urban fabric or do they legitimize the colonization of middle classes and the domination of the poorest (Bridge et al., 2011)?

Macroprocesses also lead to deep long-term transformation of urban policy. David Harvey's seminal 1989 paper on urban entrepreneurialism is based upon his reinvigorated Marxist framework and analysis of the spatial fix of capitalism (Harvey, 1985, 1989). He explained how the rise of globalized capitalism was radically changing the logic of urban policy: competition between cities was the name of the game and urban policy had to be analyzed as the result of this disciplinary pressure on working and lower middle classes by a bourgeoisie and large economic interest taking advantage of the global competition between spaces. In Europe, urban policy shifted toward more entrepreneurial policies, destroying part of the existing capital (and social relations) in order to favor different conditions for capital accumulation. From a rather different perspective, Bob Jessop conceptualized the change as the end of the *national Keynesian welfare state* and the coming age of the *Schumpeterian globalized workfare state*. Analyzing 'the crisis of the national spatiotemporal fix and the ecological dominance of capitalism' (2000), Jessop's analysis was illuminating: national states were not just undermined by bottom-up political pressure leading to decentralization or metropolitan government, they were active and central in the restructuring of national spaces, to foster urban governance in order to accelerate the process of 'creative destruction' (Schumpeter) characteristic of capitalism. Urban policies were therefore to shift from classic planning, housing provision, infrastructure and welfare policies to supply side policies, development corporations, free tax zones, and competition. This crucial insight proved decisive to understand radical urban policy changes in Thatcherite Britain or in the US. Neil Brenner went further, building on the insights of Harvey and Jessop and also Neil Smith and Henri Lefevre to design his own model precisely linking the rescaling of the state, or statehood as he puts it, with the rise of urban governance. To sum up his argument: "'it is no longer capital that is to be molded into the (territorially integrated) geography of state space, but state space that is to be molded into the (territorially differentiated) geography of capital' and thus 'the goal of national, regional and local state spatial policies is no longer to alleviate uneven geographical development but to intensify it through the deployment of locational policies designed to strengthen the place-specific socio-economic assets of strategic, globally linked regions'" (p. 16).

National urban policy shift neatly reflects the trend. At the urban level, as was seen in the UK in the 1980s, more resources have been used to address the logic of competition to attract media attention, middle classes, capital, firms or public organizations in London or old industrial cities alike. That progressively also applied to cities elsewhere. Competition is not new as such; after all, all the countries saw fierce competition between cities to attract research centers, railway stations (nineteenth century), or airports but it is not regulated by political influence within the national space any more. Similarly and even more tellingly, national spatial policies have shifted: instead of dealing with regional inequalities, they support urban mega project and they shift resources to reinforce the prestige and competitive edge of capital cities. States use their capital cities as national champion for economic growth (Crouch and Le Galès, 2011). The more centralized the country, the more obvious the shift as evidenced in London, Paris, Copenhagen, Budapest, or Lisbon. Urban competition has become one important factor to explain urban policy change and that translated into financial transfers, public investments, and decreasing support for those left behind, in other words a shift "from regional development policy aimed at assisting backward areas to one aimed at improving further the performance of the most successful" (Crouch and Le Galès, 2011). Many in Britain have singled out this new form of capitalism (The phrase is due to urban policy scholar Alan Harding from the University of Liverpool) where the state invests in the capital and uses spectacular events, like the Olympic games, to legitimate large public expenditure in the capital city. Attempts to reorganize 'Great Helsinki' or 'le Grand

Paris' and the infrastructure projects associated to them (IT platform, new airport, new metro, international university) are no different.

The competition logic for urban policy is now promoted as end to itself: many city leaders publicly engage in the competition to attract the next international Expo (Milan, 2015 winning against Izmir, Turkey), a best practice award from the United Nations Programme of Best Practices to improve living conditions in human settlements, The EU green capital award (Vitoria in Spain after Hamburg and Stockholm), the European capital of culture (Marseille and Kosice before Umea and Riga). Honors and labels are like medals, classic instruments to create competition for prestige, orienting the conduct of conduct in the Foucauldian sense and legitimizing the UN or the EU. For urban elites, those labels are seen as a positive political symbol for elections and provide an opportunity to legitimize entrepreneurial policies, and benchmarking is an opportunity to create reputations, trust, and to attract investment or private public partnership.

Brenner has also worked with a group of critical political economy geographers to analyze the rise of neoliberalism, the characteristics of the paradigm, and the way it has translated into urban policies (Peck, Theodore, Brenner 2013) in order to carefully identify existing neo-liberalism and its incompleteness in most places.

There are lots of evidence to show how urban policies have been transformed, for instance, through the movement of privatization of services, utilities, and the introduction of quasi-markets; the financiarization of various urban majors actors (utilities, private developers, and even housing organizations) (Halpern and Pollard, 2011); the competition between cities justifying large investment in prestigious cultural projects, mega events (Olympic games), and emblematic buildings by star architects (Ponzini and Nastasi, 2011); supply side policies to support firms and economic development; more flexible land planning to allow for diverse project and private actors profit; gentrification processes favoring middle classes; new forms of policing including the control, exclusion, or surveillance of poor groups and immigrants; new ways to deliver policies through contracts to private firms (Raco, 2012); benchmarking between cities; massive urban renewal strategies (Epstein, 2013) or policies supporting gentrification. Whatever may be the dynamics between these different processes, globalization, capitalism expansion and financiarization, neoliberalization, and state restructuring matter for urban policy.

Urban policies, as most public policies, are the result of collective action where some of the actors have particular type of resources, political legitimacy, and capacity. The political sociology of urban public policy has emphasized the political role of other actors from social movements to NGOs and also from large firms, a classic theme in US urban studies which was less influential and fruitful in the Europe. Things are different now, thanks to globalization and deregulation process. Some actors in urban policies have accumulated so much wealth that the translation in power resources is becoming a threat for democracy (Crouch, 2011). Europe was for a long time rather behind this movement but things have changed: property markets, banks, and urban developments are now more and more part of international strategies from financial organizations.

In urban policy, giant utility firms have become massively influential actors, as EDF, Centrica, Gazprom, RWE, Eon, GDF Suez, ENEL, Veolia environment, or Iberdrola. They invest in different parts of the world in energy, the built environment, renewable energy in particular but also engineering, transport, or even water. The European market for utilities is estimated at around 800 billion euros per year. Capitalism is also urban capitalism and it takes different forms in Europe (Lorrain, 2005). Those firms have amalgamated to become globalized multiservice utility firms. Eon, the giant German utility group, had in 2012 a market capitalization of 27 billion euros and employed 75 000 people. EDF, the French electricity and utilities group, employs 150 000 people, with markets sales of around 65 billion euros in 2011. Most utilities negotiate some time length monopoly in delivering some services management or delivery, built infrastructure or energy with municipalities. Special funds dedicated to the financing of utilities have become major players of urban policies. Utility funds developed to finance the investment and the firms in utilities, energy, water, transport, waste disposal, and IT infrastructure who work with urban government. The deregulation of utilities, initiated in the UK by Mrs Thatcher's government, paved way for a formidable boom of the financing of utilities investment.

Top investors in property development are huge financial organizations, such as the Norway Sovereign Wealth funds, the leading investor in property development in Europe in 2012, but leading property development investors include the German fund manager Union Investment, the famous US private equity firm 'Blackstone,' or the real estate branch or top insurance companies such as Generali in Italy, Allianz in Germany, or Axa in France.

Urban policies in Europe deal with contemporary European cities. Urban areas all over Europe are dealing with the consequences of the crisis and financial cuts are limiting their otherwise important investment capacity. At best they enjoy slow growth but some are facing deep crisis. Western Europe's classic largest metropolis (i.e., London and Paris, around 10 to 12 million inhabitants) are not in the top 20 in the world any more and aging is a growing concern. By contrast, Istanbul and Moscow are taking over as major development centers on the other side of the continent. Despite the welfare state, inequalities increase, riots regularly take place, in particular in former colonial countries (London, Paris) or not (Stockholm), or against the crisis in Athens, Madrid, or Barcelona. The EU has not delivered in terms of regional and urban policy and the much heralded change toward a more sustainable development and less energy proned growth model is proving slow and difficult. Utilities and private developments are financiarized. Democracy is under pressure from large global business firms, extreme right parties, and general decreasing engagement with politics. European cities and urban areas are becoming more ethnically diverse, thanks to immigration, even if the social models are still robust and cities experiment democratic participation (Kübler, 2012). Since the 1980s, at least until 2010, most Western European cities enjoyed demographic and economic growth; even former industrial cities joined the trend with a few exceptions in Germany, Italy, or the UK. But Eastern European cities went the other way in the aftermath of post-1989 economic deindustrialization even if the capitals

now recover. The EU urban policy has not been included and for one good reason: after a slow emergence a decade of important initiative led scholars to expect the development of an ambitious Urban policy, the dog did not bark. The liberal turn of the EU in the late 1990s marginalized the whole regional policy aiming at reducing inequalities. Enlargement became an excuse to get rid of the urban policy and the 'mainstreaming' of EU urban policy meant residualization (Atkinson and Eckardt, 2004; Halpern and Le Galès, 2011).

European cities are part of the restructuring process of a continent searching for a different model. Urban policies in Europe are part of this work in progress, if progress is the right word. But the economic crisis and what looks like a decade of austerity is deeply undermining European cities and urban policies alike.

See also: Capitalism; Governance; Institutions; Neoliberalism; Public Policy; Urban Policies in Asia; Urban Policy in North America.

Bibliography

Andersen, T.H., Van Kempen, R. (Eds.), 2001. Governing European Cities. Social Fragmentation, Social Exclusion and Urban Governance. Ashgate, Aldershot.

Andersson, R., Bråmå, Å., Holmqvist, E., 2010. Counteracting segregation: Swedish policies and experiences. Housing Studies 25 (2), 237–256.

Andreotti, A., Le Galès, P., Moreno Fuentes, F., 2013. Controlling the urban fabric: the complex game of distance and proximity in European upper middle classes residential strategies. International Journal of Urban and Regional Research 37 (2), 578–597.

Atkinson, R., Eckardt, F., 2004. Urban policies in Europe: the development of a new conventional wisdom. In: Eckardt, F., Kresil, P. (Eds.), City Image and Urban Regeneration. Peter Lang, Franckfurt, pp. 33–65.

Atkinson, R., Rossignolo, C., 2008. The Recreation of the European City, Governance, Territory and Polycentricity. TechnePress, Amsterdam.

Bagnasco, A., Le Galès, P. (Eds.), 2000. Cities in Contemporary Europe. Cambridge UniversityPress, Cambridge.

Béal, V., 2011. Urban governance, sustainability and environmental movements: post-democracy in French and British cities. European Urban and Regional Studies 19 (4), 404–419,.

Béal, V., Gauthier, M., Pinson, G. (Eds.), 2011. Le développement durable changera-t-il la ville? Le regard des sciences sociales. Publications de l'Université de Saint-Etienne, Saint-Etienne.

Bolt, G.S., Kempen, R. van, 2011. Successful mixing. Effects of urban restructuring policies in Dutch neighbourhoods. Tijdschriftvoor Economische en Sociale Geografie 102 (3), 361–368.

Brenner, N., 2004. New State Spaces. Oxford university Press, Oxford.

Brenner, N., 2009. Restructuring, rescaling and the urban question. Critical Planning 16, 60–79.

Brenner, N., 2013. Theses on planetary urbanization. Public Culture 25 (1), 85–114.

Bridge, G., Butler, T., Lee, L. (Eds.), 2011. Mixed Communities Gentrification by Stealth?. Policy Press, Bristol.

Buck, N., Gordon, I., Harding, A., Turok, I. (Eds.), 2005. Changing Cities, Rethinking Urban Competitiveness, Cohesion and Governance. Palgrave, Basingstoke.

Castells, M., 1977. The Urban Question, a Marxist Approach. Edward Arnold, London.

Cochrane, A., 2007. Understanding Urban Policy. A Critical Approach. Blackwell, Oxford.

Cochrane, A., 2011. Making up global urban policies. In: Bridge, Gary, Watson, Sophie (Eds.), The New Blackwell Companion to the City. Wiley-Blackwell, Chichester, UK, pp. 738–746.

Coutard, O., Dupuy, G., Fol, S., 2004. Mobility of the poor in two European metropolises: car dependence versus locality dependence. Built Environment 30 (2), 138–145.

Crosta, P.L. (Ed.), 2009. Casi di politiche urbane. La pratica delle pratiche d'usodelterritorio. Franco Angeli.

Crouch, C., 2011. The Strange Non Death of Neoliberalism. Polity Press, Cambridge.

Crouch, C., Le Galès, P., 2011. Cities as national champions. Journal of European Public Policy 19 (3), 405–419,.

Cunha, A., Knoepfel, P., Leresche, J.P., Nahrath, S., 2005. Enjeux du développement urbain durable. Presses Polytechniques et Universitaires Romandes, Lausanne.

Davis, J., Imbroscio, D. (Eds.), 2009. Theories of Urban Politics, second ed. Sage, London.

Dean, M., 2010. Governmentality, Power and Rule in Modern Society. Sage, London.

Dewey, J., 1927. The Public and Its Problems. Holt, New York.

Epstein, R., 2013. La rénovation urbain. Presses de Sciences Po, Paris.

Flint, J., Raco, M. (Eds.), 2012. The Future of Sustainable Cities. Policy press, London.

Foucault, M., 2007. Security, Territory, Population, Lectures at the College de France 1977–1978. Palgrave, Basingstoke.

Gandy, M., 2005. Cyborg urbanization: complexity and monstrosity in the contemporary city. International Journal of Urban and Regional Research 29 (1), 26–49.

Giersig, N., 2008. Multilevel Governance and the European City. VSverl, Wiesbaden.

Goldsmith, M., Page, E. (Eds.), 2010. Changing Government Relations in Europe: From Localism to Intergovernmentalism. Routldege/ECPR, London.

Goldsmith, M., Page, E. (Eds.), 1987. Central and Local Government Relations: Comparative Analysis of West European Unitary States. Sage, London.

Le Galès, P., 2002. European Cities, Social Conflicts and Governance. Oxford University Press, Oxford.

Le Galès, P., Therborn, G., 2010. Cities. In: Immerfal, Stefan, Therborn, Göran (Eds.), Handbook of European Societies. Springer, pp. 59–89.

Hamedinger, A., Wolffhardt, A. (Eds.), 2012. The Europeanisation of Cities: Policies, Urban Change & Urban Networks. Techne Press.

Halpern, C., Pollard, J., 2011. Les acteurs de marché font-ils la ville?. Working papers du Programme Villes & Territoires, 2011-06 Sciences Po, Paris.

Halpern, C., Le Galès, P., 2011. No autonomous public policy without ad hoc instruments. A comparative and longitudinal analysis of the European Union's environmental and urban policies. Revue Française de Science Politique, (English) 2011/1 61, 43–67.

Harloe, M., 1995. The People's Home. Blackwell, Oxford.

Harvey, D., 1985. The Urbanisation of Capital. Blackwell, Oxford.

Harvey, D., 1989. The Condition of Postmodernity. Blackwell, Oxford.

Healey, P., 2012. Thinking planning and urbanism. International Journal of Urban and Regional Research 36 (4), 871–872.

Heinelt, H., Kubler, D. (Eds.), 2005. Metropolitan Governance: Capacity, Democracy and the Dynamics of Place. Routledge, London.

Hohenberg, P., Lees, Hollen L., 1985. The Making of Urban Europe. Harvard University Press, Cambridge.

Hood, C., 2011. The Blame Game: Spin, Bureaucracy, and Self-preservation in Government. Princeton University Press, Princeton.

Imbroscio, D., 2012. Beyond mobility, the limits of liberal urban policy. Journal of Urban Affairs 34 (1), 1–20.

Imrie, R., Lees, L., Raco, M. (Eds.), 2008. Regenerating London, Sustainability and Community in a Global City. Routledge, London.

Jessop, B., 1997. The entrepreneurial city: re-imagining localities, redesigning economic governance and restructuring capital? In: Jewson, N., Macgregor, S. (Eds.), Realising Cities: New Spatial Divisions and Social Transformation. Routledge, London.

Kaelble, H., 1984. The European Way: European Societies during the Nineteenth and Twentieth Centuries. Berghan Books, New York.

Kazepov, Y. (Ed.), 2005. Cities of Europe. Blackwell, Oxford.

Kazepov, Y. (Ed.), 2010. Rescaling Social Policies towards Multilevel Governance in Europe. Ashgate.

Keating, M., 2013. Rescaling Europe. Oxford University Press, Oxford.

Kübler, D., 2012. Metropolitanization and metropolitan governance. European Political Science 11 (3), 402–408.

Laino, G., 2012. Il fuoconelcuore e il diavolo in corpo. La partecipazione come attivazione sociale. Franco Angeli, Milano.

Laswell, H., 1936. Who Gets What, When, How. Meridian Books, Cleveland.

Lorrain, Dominique, 2005. Urban capitalisms European models in competition. International Journal of Urban and Regional Research 29 (2), 231–267.

McCann, E., Ward, K. (Eds.), 2011. Mobile Urbanism: City Policymaking in the Global Age. University of Minnesota Press, Minnesota.

Miller, D., Rose, N., 2008. Governing the Present. Polity Press, Cambridge.

Moulaert, F., Rodriguez, A., Swyngedouw, E. (Eds.), 2003. The Globalized City. Economic Restructuring and Social Polarization in European Cities. Oxford University Press, Oxford.

Musterd, S., Ostendorf, W. (Eds.), 1998. Urban Segregation and the Welfare State: Inequality and Exclusion in Western Cities. Routledge, London.

Nicholls, W., 2008. The urban question revisited: the importance of cities for social movements. International Journal of Urban and Regional Research 32 (4), 1468–2427.

Palermo, P., Ponzini, D., 2010. Spatial Planning and Urban Development. Springer Verlag.
Peck, J., 2005. Struggling with the creative class. International Journal of Urban and Regional Research 29 (4), 740–770.
Peck, J., 2010. Constructions of Neoliberalreason. Oxford UniversityPress, Oxford.
Peck, J., Theodore, N., Brenner, N., 2013. Neoliberal urbanism redux? International Journal of Urban and Regional Research 37 (3), 1091–1099.
Pinol, J.L. (Ed.), 2012. Histoire de la villeeuropéenne. Le Seuil, Paris.
Pinson, G., 2009. Gouverner la ville par projet. Presses de Sciences Po, Paris.
Phelps, Nicholas A., Wu, Fulongeds, 2011. International Perspectives on Suburbanization. A Post Suburban World? Palgrave Macmillan, pp. 192–210.
Ponzini, D., Nastasi, N., 2011. Star architecture. Scene, Attori et spectacoli, nellecittacontemporanee. Allemandi, Milano.
Porter, L., Shaw, K. (Eds.), 2008. Whose Urban Renaissance? An International Comparison of Urban Regeneration Strategies. Routledge, London.
Pressman, S., Wildavsky, A., 1974. Implementation: How Great Expectations in Washington Are Dashed in Oakland or Why It's Amazing that Federal Programs Work at All: This Being a Saga of the Economic Development Administration. California University Press, Berkeley.
Raco, M., 2012. The new contractualism, the privatization of the welfare state, and the barriers to open source planning. Planning Practice, and Research.
De Swaan, A., 1988. In Care of the State. Cambridge UniversityPress, Cambridge.
Sellers, J., 2001. Governing from Below. Urban Politics and Post-industrial Economy. Cambridge University Press, Cambridge.
Stanilov, K. (Ed.), 2007. The Post Socialist City. Springer.
Storper, M., 2013. Keys to the City. Princeton University Press, Princeton.
Swyngedouw, E., Heynan, N., Kaila, M. (Eds.), 2006. In the Nature of Cities, Urban Political Ecology and the Politics of Urban Metabolism.
Theodore, N., Peck, J., 2012. Framing neoliberal urbanism: translating "commonsense" urban policy across the OECD zone. European Urban and Regional Studies 19 (1), 20–41.
Tilly, C., 1990. Coercion, Capital and European States. Blackwell, Oxford.
Tilly, C., Blockmans, W. (Eds.), 1994. Cities and the Rise of States in Europe. WestviewPress, Boulder.
Vitale, T., Podesta, N., 2011. Dalla proposta alla protesta, e ritorno. Conflittilocali e innovazion e politica [From Proposal to Protest, and Back Again. Local Contention and Political Innovation]. Bruno Mondadori, Milano, p. 256.

Urban Policy in North America

Michael B Teitz, University of California, Berkeley, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urban policy in North America comprises a blend of historically established laws and government policies together with newer efforts to address urban problems. In the former category are key social agreements, for example, that urban development will be primarily carried out by the private sector, constrained only by planning laws that are mostly in the domain of local and state governments. Examples in the latter category include efforts to solve specific urban problems such as inadequate and unaffordable housing for low-income populations, or to respond to infrastructure deficiencies in transportation. All three countries in North America – the United States, Canada, and Mexico – are federal in their governance structure, so policy is necessarily distributed across the levels of government. As a result, central government plays only a limited role in key aspects of urban growth, though in Mexico there is a stronger tradition of centralization of policy. Within this framework, urban policy is diverse, but falls into three broad groupings. First, land use and development underpin urbanization; policy in these areas has existed from the earliest days in all three countries, exemplified by regulation of urban form. In North America, such regulation has, to varying degrees, supported market driven development. Second, urban policy deals with management of key functions, exemplified by transportation, public health, utilities, and environmental quality. Such policy occurs at all levels of government. Third, urban policy addresses social issues that have varied over time from the acculturation of immigrants, to poverty, housing, and race relations.

Urban policy in North America is something of an anomaly. In all three countries – the United States, Canada, and Mexico – most people live in urban areas, yet national governments have few policies explicitly identified as urban, even though most national domestic policies have powerful urban implications. Urban areas contain their own local governments within larger federal and state structures, and those governments pursue the welfare of their inhabitants and other organizations within their boundaries. States and provinces occupy an intermediate role within the federal structure of each nation-state, establishing the ground rules for much that urban governments may do while intervening directly. Although urban policy has come to be identified particularly with some sets of issues, notably those to do with land use, housing, transportation, poverty, and racial segregation, in fact it is much broader than this. Indeed, it may be argued that many of the most important components of urban policy in the countries comprising North America are scarcely visible. Cities are the most dynamic form of human social organization. Urban policies reflect fundamental social agreements about how cities and towns continue to be built and transformed and how their inhabitants relate to each other. Only when those understandings are challenged does the visible portion of urban policy emerge, usually in the context of political and social conflict. Resolution of those conflicts has been a part of urban life throughout history and it will continue to be so.

'Urban policy' may take on two closely related meanings, depending on how the term 'policy' is understood. In a broad sense, 'policy' denotes a general and high-level guide to decision and action, particularly by governmental or other public entities. In this case, urban policy should be seen as comprising general principles that guide public action in urban areas and frame the issues associated with them. For example, urban development through market processes, and support of home ownership have been such policies in the United States for the past century. 'Policy' also specifies a definite course of action selected from among alternatives to realize specific goals. In this sense, 'urban policy' refers to such courses of action in relation to urban areas and issues, especially in the public realm. An example might be the policy of supporting home ownership by making mortgage interest tax deductible. In practice, it is difficult to disentangle the two, especially in advanced societies, where most of the population lives in urban areas. Furthermore, the diversity of urban areas within a national political system, be they metropolitan regions, cities, or smaller urban places, engenders specific forms of complexity that are reflected in public policy. Thus, to understand urban policy, it is necessary to see how public decisions are formulated, as well as to examine the character of urban policies in both their general and specific senses.

Public Action in Relation to Urban Areas

The organization of public action reflects a profound dilemma associated with geographically differentiated places whose populations have political power. Metropolitan areas, cities, and other urban entities are the most dominant such places in our time. Considerations of technical efficiency often suggest that public action should be shaped by functional requirements, rather than in accord with governmental structure (examples are transportation, public utilities, and public safety). Such function-based structures might be seen as a 'vertical' form of organization, in which a particular function is managed coherently to take advantage of the economies of scale. Most national states have ministries that reflect such a functional division, as do lower levels of government. In strongly unitary states, such national ministries extend their reach to virtually all aspects of life, even at the most local or individual level.

Society, however, comprises much more than the nation-state, being a vastly complex structure of social, economic, and

political relationships among individuals and groups. The reality of everyday life engenders strong relationships in a 'horizontal' form, crosscutting functional concerns and focusing on place itself. Such 'place-based' concerns are accentuated where people in subnational communities have the political power to express their collective desires, even though they may run counter to national concerns expressed through vertical forms of organization. Democratic societies in which legislatures are elected on a geographic basis are particularly likely to express such horizontal pressures. However, the fundamental social organizational forces that create subnational communities are present everywhere, with concomitant tensions between the vertical and horizontal, the functional and the communal, at all levels of government.

One solution to this dilemma is a federal system, in which legislative and executive responsibilities are explicitly allocated to different levels of government, usually associated with accompanying geographic disaggregation of the national space. This is so in all three countries in North America. Conventionally, systems of this type exhibit at least three formal levels of government, national, state (also called regional or provincial), and local, with the latter often further subdivided. Urban areas are rarely national in scale or scope, though Singapore provides an important exception as a city that is a unitary state, albeit with local forms of administration within itself. Urban areas at the first level below the national in a federal system are also relatively infrequent, though Mexico City and Washington, DC, are principal examples. Even there, the capital is a unique political entity rather than simply a state within the federal system. Thus, most urban areas in explicitly federal systems fall under the jurisdiction of local governments.

Although urban areas are geographically local in nature, urban policy is not also essentially local. To the contrary, cities are intricately involved in national politics in urbanized societies. Many strands of national policy affect urban areas or are intended to respond to the concerns of their inhabitants. The formation and execution of urban policy occurs at all levels of government. In addition, many policies that deeply affect urban areas and their inhabitants are couched in terms of other policy domains. Major governmental ministries or agencies whose explicit mission is to shape urban policy are rare. Nonetheless, urban places are both the target of many explicit policies, while also the recipients of the consequences of others. It is often difficult to distill the urban policy implications of policies formulated in other fields.

Urban Policy in North America

North America is a continent comprising three separate nation states, each with a distinctive history, its own cultural and demographic makeup, and its own economic, social, and political system. Nonetheless, Canada, Mexico, and the United States do have much that is common in their forms of urbanism, their governance, and the policies they have evolved. All three countries are highly urbanized, although Mexico has a somewhat lower proportion of its population living in urban areas, reflecting its status as a less-developed economy. In each country, market forces drive urbanization, though the extent to which they are regulated varies substantially. Cities in each are shaped by the dominant technologies of our time, especially by the automobile as the primary means of transportation. Demographic changes, especially immigration and aging are affecting urban areas in each country. Nonetheless, there are important themes of urban policy in each. Critical to understanding those themes are the categories of public policy involved and the governments that addressed them.

Although 'public policy' is a widely used term, which now drives political debate, as well as graduate and professional education and research, it is seldom voiced in legislation. Much urban public policy, in fact, may be settled, broadly accepted, and challenged only when conditions change sufficiently to lead to questions being raised about its validity. Urban policy, in this view, is something like an iceberg, much of which is virtually invisible.

Urban Policy as Land Use and Development

Consider the fundamental question of how urban areas in most of North America are designed and built. Current policy debates concerning land use planning, urban sprawl, and transportation infrastructure investment are shaped by long agreed public policy that urban development should be carried out by private investors, regulated by government, and in part supported by public investment. Europeans established this policy in the early years of settlement of North America, even though the first settlements were far from pure capitalist enterprises. As colonies of Britain, the Netherlands, France, and Spain, they were intended to benefit their parent countries, and to varying degrees they were rigorously regulated and their forms prescribed. Some of this heritage carries over, especially in Mexico. Nonetheless, in the United States, especially, there had emerged by the early nineteenth century a distinctive form of settlement that was to define the American form of urbanism, namely the grid-plan.

The urban grid, a rectilinear structure of streets, was the physical expression of a fundamental policy that urbanization would be market driven, privately organized, and minimally regulated. It enabled two key elements necessary for such a process to work. The first was the creation of easily laid out and marketable lots that could be sold speculatively to buyers at a distance. The second was the easy expandability of the urban structure, simply by adding further streets and subdivisions as the market expanded. Thus, a distinct and recognizable form of American city emerged, speculative in its origins, highly adaptable to changes in growth and technology, and based on the conception of land as a marketable commodity. Even when the grid was replaced in the twentieth century by the curvilinear streets of the suburb, development remained a function of the private market. The suburb, with its stronger appeal to the consumer's sense of aesthetic quality and lifestyle choice, necessarily expanded the realm of public or quasi-public regulation, but it was equally the product of private, speculative development. With this shift, however, in the United States came major public policies that have supported this form of urbanization, notably the federal funding of major highways, and the deductibility of mortgage interest, the latter justified by an explicit policy of encouraging home ownership rather than rental tenure. The most recent challenge to suburban forms of urbanism, by neo-traditional or 'New Urbanist' critics, calls

only for a change in the way in which urbanization is treated by regulation and taxation. Only relatively rarely in modern American history has government attempted directly to create urban places, even though national, state, and local laws have fundamentally shaped urban settlement patterns. United States' efforts to build new towns during the Great Depression of the 1930s were unusual in this respect and they failed.

Clearly, this account simplifies a complex, continental scale reality. The legacy of Spain in Mexico was one of a much more centralized, class defined, and bureaucratic society. Mexican urban development reflected those facts in its slower growth and more constrained forms of urbanism. Similarly, development in Canada reflected both the earlier French centralizing influence and the British legacy of unitary parliamentary government, albeit grafted onto a federal system in which the provinces were relatively strong. It is no accident that public control of urban development in both Mexico and Canada remained historically much stronger than in the United States, yet in both countries cities are not built by governments.

The idea that urbanization is market driven has had powerful consequences. Paradoxically, the market itself implies public constraints. No market functions without some prior structure of behavior or public order; some constraints must be established for markets to function at all. For urban policy, this implies public regulation of the urban development process. The nature and locus of that regulation has defined many of the key urban policy issues in North America over the past 100 years. In the United States, control of public policy is ultimately vested either at the level of the states or of the federal government, depending on the Constitution and its interpretation by the Supreme Court. With respect to land use, control is generally agreed to rest legally at the state level, although the basic legal doctrines that support land use regulation rest on constitutional decisions that are national. In fact, states have delegated land use regulation to local governments under a variety of names. California, for example, delegates the power under the doctrine of 'Home Rule.' On this legal basis, aided by a long history of piecemeal settlement and independent formation of urban places, US urban policy toward development and change is played out at the most local level.

This situation contrasts sharply with that in many other advanced countries, notably in Europe and Asia, where state or central governments play more centralized roles in land use planning, regulation, and development. Canada and Mexico also present substantially more centrally regulated versions of urban development. In Canada, cities in the central and western provinces tend to look more like those in the United States. The eastern provinces have a stronger tradition of central control, though always empowered at the provincial level. Toronto's ability to control its suburban development contrasts markedly with any US city, but a change in political complexion at the Ontario provincial level has had substantial implications for the city's policy. In Mexico, until very recently, with a single-party state, local control of development was limited. Although state or central governments did not always have the capacity to enforce their will more than intermittently over economically or politically powerful local groups, the tradition of central government finance and fiat remained powerful. How this has evolved in the first decade of the twenty-first century, as the political monopoly of the Partido Revolutionario Institutional (PRI) has ended will be discussed later.

Urban Policy as Functional Management

Local power over development has major implications for urban development, particularly when viewed from a functional perspective. For decades now US observers and advocates have argued that urban problems of transportation, energy, utilities, and environmental quality require the political ability to act at a regional level. Again and again, such arguments have been turned back as localities have refused to form larger regional governments or to cooperate in accepting land uses that they deem undesirable. That refusal has been encapsulated in the term 'NIMBY' ('Not in my back yard'), which has become both a rallying cry for control by local neighborhoods in the face of larger economic and political forces, and a term of opprobrium conveying selfishness and exclusion. Nonetheless, metropolitan areas cannot be built and cannot operate at a high level of amenity and income without necessary functional efficiency. One widely employed means around the problem has been the creation of single-purpose special districts, for example, to handle water supply. These are typically limited in their range of action and their ability to override local decision making. They are established by state law, but with the clear requirement that local political representatives support their formation. In recent decades, some entities of this type have also emanated from federal governments in all three countries.

Management of air quality is good example of such urban policy, illustrating both its urban character and its placement within another organizational realm. Awareness of the dangers of air pollution has a long history in urban areas, but its modern manifestation in the United States may be dated from the passage of clean air legislation by the city of Pittsburgh in the 1950s and the enactment of state and federal air quality standards and legislation in the 1970s. At first, efforts to improve urban air quality focused on stationary sources of pollution, especially industrial sources such as steel mills and oil refineries. As they were controlled, however, it became apparent that the single greatest source of air pollution was from mobile sources, especially private automobiles. In metropolitan areas and cities such as Los Angeles, dirty air was identified as a contributing cause of serious public health problems and reduced quality of life associated with smog. Clearly, these were regional and in some cases interstate problems. The federal government's response was to treat environmental quality by regulatory legislation and the formation of the US Environmental Protection Agency in 1970. State governments reinforced the effort with similar legislation.

Air quality degradation by automobiles can be countered in many ways, including regulating the emissions permitted, increasing the cost of driving, providing incentives for use of cleaner vehicles, and changing the way in which urban areas are structured so as to reduce the number of miles driven. Any of these policies are conceivable, but the specifics of implementation differ among them. Furthermore, the mode of intervention might have a great deal or very little relationship to urban areas themselves. Regulation of automobile emissions through national specification of permissible technologies is

only indirectly urban in character, and its effectiveness will vary with the standards applied from place to place. On the other hand, consider the problem of regulating development as a means to control emissions in metropolitan Los Angeles. Although both federal and state agencies have regional offices, neither would typically be able to carry out such a policy effectively at the local level. To do so for the metropolitan area required also a regional agency, in this case, the South Coast Air Quality Management District. This agency, established directly by the state, legally has the power to override local development decisions. However, in the face of overwhelming political opposition to such actions, which would violate home rule, it has been reluctant to do so, preferring to employ its regulatory powers in other directions. Alternative air quality policies may have similar goals, but their urban manifestation may be quite dissimilar.

This story of policy directed toward urban problems, defined on a functional basis, and legislated and implemented at multiple levels of government can be found again and again in North America, as it can elsewhere in the urbanized and urbanizing world. Apart from being much worse than in Los Angeles, air pollution in Mexico City presents much the same set of problems. In transportation, public health, housing, water supply, wastewater disposal, garbage removal, and public safety, governments have been attempting to find solutions within their own financial and political situations. It is not a new phenomenon. From the beginning of the industrial revolution, much of what we recognize as urban policy has been driven by attempts to deal with the dysfunctions of the industrial process. This process occurred a little earlier in Britain than in most of Europe and North America, but once on course it has barely slackened. Some urban concerns, of course, are much older. For example, laws to prevent disastrous fires by prohibiting wooden buildings may be found in the very earliest records of urbanization. However, the emergence of systematic urban policies really dates from the nineteenth century.

For early industrial cities in North America, the most urgent issues were public health, public safety, and transportation. The previous section indicated how the use of a planned grid permitted both a functioning land market and lines of transportation. Although the cores of cities were often congested, especially where, as in Boston, their cores were essentially unplanned, they could function with pedestrian and horse-drawn transportation before the advent of the railroads. However, such cities were necessarily densely built, outran their water supplies, and had primitive sewage disposal. Frequently, they endured epidemics, notably of cholera and yellow fever. As epidemiology evolved, the importance of clean water and sewerage became evident and the first great modern age of urban policy began. Although the delivery of pure water was, at first, a privately financed and run form of enterprise, it soon became a municipal monopoly, as did sewerage. Provision of electric power and gas, on the other hand, primarily remained private enterprises in North America, with a few minor exceptions. These utilities were regulated after the turn of the twentieth century in order to limit their monopoly power. Public safety became a key municipal function with the creation of police and fire forces in the nineteenth century and has generally remained so.

Beginning with the development of urban trolley lines and railroads in the nineteenth century, transportation investments have played an especially critical role in urban policy over the past 200 years. Although streets and street maintenance have been primarily managed by local governments, the provision and regulation of transportation has involved both private markets and public entities and governments. In the United States, most public policy in this realm was enacted by state and local governments throughout the nineteenth and early twentieth centuries. The ability of dense but growing cities to expand depended on fast and reliable forms of transportation so long as employment remained centralized. Privately run trolley lines, buses, subways, and commuter railroads, operating under local or state charters, provided a workable, if often uncoordinated solution. Later, many were consolidated under public ownership. Together, they enabled development of the suburbanized city of the early twentieth century, brought to full flower by the automobile. With the advent of automobiles, the focus of public policy shifted from private investment in infrastructure to public investment in highways, funded in large part by taxes on the sale of gasoline. National government engagement in transportation within cities remained minor until the 1950s. By that time, the invention of the freeway had raised the cost of new investment beyond the capacity of both local and state governments.

Large-scale federal entry into the process began with the creation of the Interstate Highway Program in the 1950s. Despite its formal purpose of enhancing national defense by providing high-speed, controlled-entry highways linking the entire country, the Interstate system played a substantial role in urban development through the second half of the twentieth century. By the year 2000, urban freeways ran both through and circumferentially around every major urban area in the United States, fostering suburban decentralization of cities across the continent. Similar freeway development took place in Canada and, more recently, has been occurring in Mexico. Although there has been much debate about the precise intent and impact of the Interstate system on US cities, there can be little doubt that the availability of federal funds significantly enhanced suburban development in the past 40 years. Whether these were conscious policy choices and whether this was the primary cause of the decline of central cities, discussed later, cannot be demonstrated. Given the rapid growth of population, productivity, and income, together with consumers' desires for single-family housing and automobile transportation, it seems likely that economic and political pressures would have had similar effects, whether or not the federal government was involved. Nonetheless, this period saw the entry of federal policy into urban transportation in a major way. It remains there today, although the nexus has shifted. The past two decades has seen a resurgence of political support for urban mass transportation, driven by the perception that the freeway solution is either too costly or too destructive of environmental quality to be sustained. This perspective parallels the critique of the suburb and urban sprawl discussed earlier. As a result, federal transportation funds now include substantial subventions to mass transportation – buses, light rail, and commuter rail in metropolitan areas. While the effectiveness of this shift is

being debated, it now appears to be an entrenched element of national policy toward urban areas.

Urban Policy as Social Policy

Since the 1960s, urban policy in the United States has not been primarily associated in the media and public debate with the issues discussed so far, though the recent resurgence of concern about urban sprawl and climate change has shifted attention back to the effects of urban growth. Rather, for several decades, urban policy has been primarily identified with problems of poverty, racial discrimination, segregation, and the decline of central cities. As is true in so many other instances, such a definition of urban policy is not new. It emerged in tandem with the industrial city in the nineteenth century, when low wages, intense overcrowding, dilapidated housing, crime, social disorganization, and disease combined to generate a new form of urban slum. The public health effects of industrial urbanism first engendered policy responses that were engineering solutions, especially in the form of potable water supplies and waste disposal. Epidemics and infectious diseases, such as tuberculosis, also required medical responses, which took the form of public hospitals and health services. Crime problems engendered police forces and prisons.

However, by the late nineteenth century, the fact that so many of the new urban populations in the United States and Canada were immigrants led to concerns for public safety in a larger sense, namely the stability of the society itself. Public education became a key element of local urban policy, its intent being both to provide a literate and capable labor force for the expanding economy and to ensure that children were able to speak English and become acculturated into society at large. Fears that immigrants would bring with them European socialist ideas and form labor parties were not unjustified, despite the fact that they were not realized. The momentum for these policies was certainly national in scope and linked to the prevailing sense of the United States as an emergent imperial power on a world scale. However, their realization was almost entirely at the state and local levels, where governments and influential voluntary groups were at times almost indistinguishable.

Poverty presented not only a risk to the health and social position of the affluent in American cities, it also affronted their belief in the United States as a place of freedom and opportunity. Efforts to help the 'worthy poor' raise themselves out of poverty took many forms, of which the Settlement House movement is the best known, evolving into forms of assistance associated with the social work profession. This movement also provided an early means for women of higher social status to become active in professional and social realms. Thus, by the early twentieth century, many believed that urban poverty called for effective governmental and voluntary actions.

A key realm of response was housing, and it is here that the largest challenge to the idea of urban development through the market emerged. The slum was perceived in many ways, but the vision of grossly overcrowded, unhealthy, dilapidated housing was intrinsic to all of them. Observers and critics of capitalism had noted the inability of markets to provide adequate urban housing from early in the industrial revolution. Public efforts to regulate construction or management were largely ineffectual, however. By the end of the nineteenth century in the United States, as in Britain and Europe, there had emerged two competing policy responses. Public regulation of housing quality took the form of building and housing codes, enforced by local bureaucracies. Those who saw regulation as insufficient or ineffective called for direct public or non-profit ownership and rental of housing for low-income families. Market advocates opposed both.

The twentieth century has seen the parallel evolution of these approaches to low-income housing policy in all three countries of North America. Mexico, which was undergoing a revolution and installation of a one-party state, probably exemplifies the highest level of commitment to public provision. However, the lack of resources due to underdevelopment, and the mistaken effort to adhere to unrealistic standards, ultimately ensured that public housing would become only a very limited part of its mass urbanization. In common with other developing countries, urbanization for the poor primarily meant informal housing on the margins of major cities. Canada, with a stronger tradition of government intervention, though varying greatly across the provinces, has also adhered to policies that give greater weight to provision of low-income housing. In both the United States and Canada, the tradition of local regulation continues, with rising standards and incorporation of new areas of oversight. In the United States, a determined effort to create a public housing program in the 1930s saw its culmination in the Housing Act of 1949, which led to construction of federally funded rental housing for the poor in most cities. However, the program became discredited, in part because of social problems emanating from its increasing racial segregation and focus on the very poorest populations, and in part because of insufficient funding to maintain and control high-rise housing that was ill-designed for its purpose. In a major shift of federal policy affecting cities, the program was largely replaced after 1974 by a low-income rental subsidy, which remains the basis of federal policy in this area. As a result, direct provision of new housing for the poor has switched toward the nonprofit sector.

Entangled with housing and poverty policy, especially in the United States, is the issue of race. American cities have long been segregated, both by race and income. African-Americans were confined by economics and social pressure to limited residential areas within cities. This phenomenon was exacerbated by major migration from the rural south to northern cities throughout the first half of the twentieth century. While that migration resulted at first in opportunity and rising incomes, by the 1950s there had also emerged ghettoes of African-American and other minority groups in poverty, in which serious social problems persisted. This process was exacerbated by the out-migration from older cities of white middle- and working-class populations seeking a suburban lifestyle and impelled also by race-based fear and prejudice. As a result, cities' financial capacity to address the needs of their populations declined. Education and public service quality diminished, which reinforced the migration. With improvements in civil rights during the 1960s, yet with racial discrimination still widespread, the contrasts between racial groups' aspirations and achievement burst out into urban civil disorder on a massive scale. Large areas of cities including Detroit, Los Angeles, and Washington, DC, were devastated.

The question of appropriately responsive policy is almost paradigmatic of urban policy in general. On one hand, efforts were made to improve incomes and opportunity through broad programs, such as those for education and job training. On the other hand, there were efforts to empower local communities through a variety of supports and incentives for the formation of community organizations. The so-called War on Poverty of the 1970s, exemplified both kinds of policy. It certainly led to the formation of new political networks and organizations among the urban poor, with long-term results on the racial and ethnic makeup of local politics. However, the position of the poorest and most socially disadvantaged populations improved little, if at all, in the face of a widening gap in the income distribution at large and a decline in the demand for unskilled labor. Increasing crime and social disorganization associated with rising drug use also contributed to the problem. With falling national rates of economic growth, budget stringency led to the withering away of urban poverty programs. In the mid-1990s, these issues were once more addressed by national policy, operating this time through the medium of the welfare program, and especially its support for unmarried parents. Insistence on entry into the labor market has reduced the level of dependency substantially, but the results have yet to be tested outside a boom economy.

Urban Policy in the Twenty-First Century

Cities as physical and social artifacts last for centuries, while policies come and go. Nonetheless, new physical and social challenges, and shifting political configurations do force change. After the first decade of the twenty-first century, it seems clear that both are at work. For North American cities, perhaps the greatest challenge is that presented by the prospect of major climate change through the process of global warming. Cities will be affected in a multitude of ways, ranging from rising sea levels to extreme temperatures and catastrophic weather events. The United States as a major emitter of greenhouse gases has made little national effort to address the problem, but states and cities in all three countries have begun to encourage efforts both to reduce carbon emissions and to prepare for the adverse consequences of climate change. California is especially notable in its enactment of legislation, together with administrative actions to reduce carbon dioxide. Many local governments have been pursuing policies toward both prevention and mitigation of the effects of climate change. It is striking that in this instance state and local governments are ahead of their national counterparts, but they are a minority.

This lack of action coincides, not accidentally with a shift to the right in the political environment in all three countries during the first decade of the twenty-first century. In the United States, that shift stalemated any effort to ratify or implement the Kyoto treaty on climate change, even as the federal government cut urban programs, especially in low-income housing, reinforcing the drift toward higher-income inequality. However, overall racial segregation declined as African-Americans and Latinos moved to the older, inner suburbs. At the same time, the tragic events of 11 September 2001 led to a huge increase in efforts to ensure security, with funds being poured into urban police and surveillance. In Canada, a conservative government came to power and similarly scaled-back programs directed toward urban issues after a long period of liberal domination. Income inequality has increased, along with residential segregation and gated communities. Perhaps the most dramatic change came in Mexico, where the PAN defeated the PRI after more than 40 years of one-party rule. Urban policy, particularly toward housing shifted toward the market, with an emphasis on middle- and working-class home ownership. New institutions for housing mortgage finance and land provision resulted in a large increase in overall supply, but at the cost of increased income segregation. Whether this shift to the right will survive the effects of the global financial crisis starting from 2008 remains to be seen, but there can be no doubt that it has constituted a major change in the environment for urban policy in North America.

See also: Cities and the Financial Crisis; Growth Management and Urban Planning in the United States; Housing and Cities; Just City; Planning Theory: The Good City and Its Governance; Planning and Development; Urban Neoliberalism; Urban Planning (Western), History of; Urban Policies in Asia; Urban Policy in Europe.

Bibliography

Babatz, G., 2004. The role of the *Sociedad Hipotecaria Federal* in the development of the Mexican mortgage market. Housing Finance International 18 (3), 42–51.

Bobo, L., O'Connor, A., Tilly, C., 2000. Urban Inequality in the United States: Evidence from Four Cities. Sage Foundation, New York.

Bradford, N., 2007. Whither the Federal Urban Agenda? A New Deal in Transition? Canadian Policy Research Network, Ottawa.

Burchell, R.W., 1998. The Costs of Sprawl—Revisited. National Academy Press, Washington, DC.

Commission Nacional de la Vivienda (CONAVI), 2007. The National Housing Program 2007–2012: Towards Sustainable Residential Development. Commission Nacional de Vivienda, Mexico.

Cullingworth, B., 1997. Planning in the USA: Policies, Issues, and Processes. Routledge, New York.

Duany, A., Plater-Syberk, E., Speck, J., 2000. Suburban Nation: The Rise of Sprawl and the Decline of the American Dream. North Point Press, New York.

Dye, T.R., 1997. Understanding Public Policy, nineth ed. Prentice-Hall, New York.

Fischler, R., 1998. Health, safety, and the general welfare: markets, politics, and social science in early land-use regulation and community design. Journal of Urban History 24 (6), 675–719.

Fowler, E.P., Siegel, D., 2002. Urban Policy Issues: Canadian Perspective. Oxford University Press, Oxford.

Frisken, F. (Ed.), 1994. The Changing Canadian Metropolis: A Public Policy Perspective. Institute of Governmental Studies Press, Toronto.

Garcia Peralta, B., Hoffer, A., 2006. Housing for the working class on the periphery of Mexico city: a new version of gated communities. Social Justice 33 (3), 105–120.

Glaeser, E., 2002. The Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier. The Penguin Press, New York.

Hall, P., 1996. Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century. Blackwell, New York.

Hall, P., 1998. Cities and Civilization. Pantheon, London.

Hayes, R.A., 1995. The Federal Government and Urban Housing Policy; Ideology and Change in Federal Policy, second ed. SUNY Press, New York.

Henderson, J.V., Thisse, J.-F. (Eds.), 2004. Handbook of Urban and Regional Economics. Elsevier, Amsterdam.

Judd, D.R., 1998. City Politics: Private Power and Public Policy, second ed. Longman, New York.

Keating, W.D., 1996. Revitalizing Urban Neighborhoods. University Press of Kansas, Lawrence.

Monkkonen, P., 2009. The Housing Transition in Mexico: Local Impacts of National Policy. Institute of Business and Economic Research, Berkeley, CA. Program on Housing and Urban Policy, Working Paper No. D09-001.

Newman, P., Kenworthy, J., 1999. Sustainability and Cities: Overcoming Automobile Dependence. Island Press, Washington, DC.

O'Flaherty, B., 1996. Making Room: The Economics of Homelessness. Harvard University Press, Cambridge.

Shabas, J., 2011. The impact of legislative reforms to Canadian Federalism on Toronto's ability to reduce poverty. Berkeley Planning Journal 24, 14–25.

Stein, R.M., 1991. Urban Alternatives: Public and Private Markets in the Provision of Local Services. University of Pittsburgh Press, Pittsburgh, US.

US Department of Housing and Urban Development, 1998. The State of the Cities. US Department of Housing and Urban Development, Washington, DC.

Wilson, W.J., 1990. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago.

Urban Poverty in Neighborhoods

JE Grigsby, III, University of California, Los Angeles, CA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 25, pp. 16079–16082, © 2001, Elsevier Ltd.

Abstract

After nearly four decades of research and efforts at policy reform and program intervention, the number of persons in poverty is increasing and persons in poverty are becoming more and more concentrated. This article provides an overview of the history of poverty in America, its causes, and indicates why scholars and policy-makers should be concerned. The article concludes by calling for a more comprehensive approach to the solutions of poverty – one which takes into account the structural sources of poverty, its multigenerational roots, and the link between individual neighborhoods and regions.

Opinions differ regarding what causes poverty in America and why it seems to be concentrated in certain areas. Regardless of these differences, there is a relatively new focus on urban poverty and an agreement that the number of persons in poverty is growing as is the concentration of poverty in urban neighborhoods.

Poverty in America

National interest in poverty was first triggered when John F. Kennedy, while running for president, was taken on a tour of poor areas in the rural south. What he and his entourage saw was staggering: children with bloated bellies, families living in tar paper shacks, high rates of unemployment, and illiteracy. In short, the other America – families living in a state of poverty, the likes of which most Americans had never seen. This experience so affected the young presidential aspirant that he vowed, if elected, his administration would do everything within its power to address issues of poverty in America.

During the 1960s, politics and the publication of Michael Harrington's *The Other America* (1963), helped to focus the nation's attention on the plight of the poor in urban areas. President Johnson, taking office after Kennedy's assassination, launched the 'War on Poverty' in an attempt to have the federal government significantly address problems of poverty. However, due to the Vietnam conflict this effort was short lived.

It wasn't until the publication of James Wilson's *The Truly Disadvantaged* (1987) nearly two decades later, that interest in poverty was rekindled. Wilson's book sparked a flurry of research projects seeking to provide answers to such questions as: how widespread is poverty in America, what causes poverty, where is poverty most severe, what types of individuals and/or households are likely to wind up in poverty, what does being raised in poverty do to people, can people escape poverty, and perhaps most important, what can be done to reduce the incidence of poverty, and who should be responsible for addressing poverty issues? The results of most of these studies have revealed that poverty in America is not limited to rural areas. Every major city in the USA contains poverty neighborhoods and, in spite of growing economic prosperity, more Americans find themselves increasingly concentrated in poverty neighborhoods.

A person in poverty is defined by the United States Bureau of the Census as one whose total family income falls below a federally defined level. This level varies according to family size and is adjusted annually for inflation. In 2000, for example, the poverty threshold for a family of four was $17,463. The poverty rate for a particular neighborhood is determined by dividing the number of poor persons who live there by the total number of poor and nonpoor persons. Census tracts are used as the geographic basis for identifying the location of poverty neighborhoods. Researchers differ on the cut-off level for defining poverty neighborhoods. The range is from 20 to 40%, with the higher level being used most commonly. While different names have often been ascribed to poverty neighborhoods, Jargowsky (1977) refers to African-American high-poverty areas as ghettos, while barrios describe Latino high-poverty neighborhoods, and slums are poverty neighborhoods comprised primarily of whites.

What Causes Poverty and Why Does It Seem to Be Concentrated in Certain Neighborhoods?

A number of different explanations for what causes poverty and why poverty appears to be concentrated have been put forth. Some of these include: discrimination and segregation, immigration of rural farm workers from other countries to cities in America, the poor quality of inner-city schools, crime and the prevalence of drugs, white flight, and the emergence of a global economy.

The Creation of Ghettos and Barrios: The Role of Segregation

During the 1940 and 1950s, the USA underwent a rapid expansion of its industrial and manufacturing sectors. African-Americans seeking to improve their quality of life migrated in large numbers from the south to major urban growth centers throughout the USA. Segregated housing, restrictive covenants, and zoning practices forced many of these African-American migrants to seek shelter in older sections of cities. Segregation and discrimination also limited employment and advancement opportunities, which in turn impacted earning power. As a consequence, African-Americans were disproportionately concentrated in low-income and often poverty neighborhoods

called ghettos (Drake and Cayton, 1945; Clark, 1965; Dubois, 1967). Even after the passage of Title VI of the Civil Rights Act of 1964 and Title VIII of the Civil Rights Act of 1968, real estate agents, mortgage brokers and lenders continued to make it difficult for African-Americans to acquire housing outside of inner city areas. As a consequence, African-Americans, unlike their white counterparts, have much lower home ownership rates, one of the key ways Americans accumulate wealth (Bullard et al., 1994; Oliver and Shapiro, 1995). Massey and Denton are two scholars who maintain that racial segregation and not economic segregation have played a major role in the continued growth of ghetto poverty (Massey and Denton, 1993).

During the 1970, 1980 and 1990s, a large number of immigrants from Latin America, in particular from Mexico, have come to the USA. To a certain extent, these immigrants find themselves relegated to poverty areas or barrios, much like African-Americans did before them. More recently, immigrants from various Asian countries, particularly Vietnam, Cambodia, and Laos have been moving to American cities. In 1990, foreign-born residents constituted 10% of all those living in concentrated poverty areas, up from 3.5% two decades earlier (Mincy and Wiener, 1993, Table 6). As of yet, no single term has emerged to describe Asian poverty neighborhoods. Although references to China Towns, Little Tokyos, and Little Vietnams often imply similar conditions to ghettos and barrios. Like African-Americans, the new immigrants also find that their poverty neighborhoods suffer from inferior schools, high crime rates, inadequate housing, and limited employment opportunities. Language barriers and low educational achievement seem to be key factors contributing to the concentration of these new immigrants in poverty neighborhoods.

While researchers have easily identified poverty neighborhoods with a predominance of racial minorities, that is ghettos and barrios, few predominately white poverty neighborhoods or slums are evident in metropolitan areas. This is not to say that there are no whites in poverty; quite the contrary, the largest number of individuals in poverty are whites. It is just that poor whites are distributed throughout cities and not particularly concentrated in certain neighborhoods.

Social Mobility

Social mobility, or as Wilson (1987) calls it, "neighborhood sorting," is another reason often put forward to help explain the growing concentration of urban poverty. There are two dimensions of this concept. The first focuses on white flight; that is, as more and more minorities move into urban areas, better educated and more affluent whites move to the suburbs. Reasons most often given for this out-migration of whites from central city areas include: a desire to provide better educational opportunities for their children, safer neighborhoods, better and cheaper services, more neighborhood amenities, and the ability to realize greater return on their investment in housing.

These same interests are propelling the move of middle-class African-Americans from inner city areas, leaving poor African-Americans behind. The consequence is, further concentration of poor people in central city poverty neighborhoods (Wilson, 1987).

The Impact of the Global Economy

More recent literature has focused on what role deindustrialization, brought on by a growing global economy, has had on both increasing the number of persons in poverty and the growing spatial concentration of poverty (Bluestone and Harrison, 1982). The growth of new jobs is occurring faster in suburban areas than in inner-city areas. This makes it increasingly difficult for residents of poverty neighborhoods to acquire employment. Kain (1968, 1992) calls this the spatial mismatch phenomena.

Rapid job growth in the service sector has reduced the impact of some losses of industrial employment. However, these jobs tend to pay relatively low wages with few, if any, benefits. This often means that, even though individuals residing in poverty neighborhoods can secure work, the jobs they obtain do not provide adequate income or sufficient benefits to lift them or their families out of poverty. Thus, there are a growing number of working poor residing in poverty neighborhoods (Pastor et al., 2000; Bernstein et al., 2000).

The Significance of Regional Economic Growth and Poverty Neighborhoods

Some of the more recent literature suggests that there is a link between the size and growth of poverty neighborhoods and the robustness of a region's economy. Jargowsky, for example, shows that a rise in overall metropolitan mean income resulted in sharp declines in ghetto poverty concentration among African-Americans (Jargowsky, 1977). Similarly, Pastor et al. found in 74 regions throughout the USA that regions which performed best economically were also ones that either reduced inner-city poverty or reduced the rate of growth of poverty neighborhoods (Pastor et al., 2000).

To date, the most comprehensive work examining the causes of poverty can be found in *Poverty and Place: Ghettos, Barrios, and the American City* by Paul Jargowsky. Unlike previous studies which are locale specific, this study examines national poverty data. Jargowsky finds that structural changes in the economy have in fact adversely affected those with lower educational levels and job skills by reducing their wages. He also finds that deconcentration of employment further reduces the possibility of these individuals finding work and that the flight of higher-income individuals from poverty neighborhoods is an important factor in increasing neighborhood poverty. However, he could not find any evidence that strongly supports the notion that segregation plays a direct role in the recent growth of poverty neighborhoods. Jargowsky concludes by suggesting that neighborhood poverty is the

> predictable result of the economic status of minority communities and the degree to which minorities are residentially segregated from whites and from each other by income.
>
> Jargowsky, 1977: p. 19

Why Is It Important to be Concerned about Urban Poverty?

One of the reasons for the concern with urban poverty is that place and neighborhoods matter. Therefore what happens to

people living in these neighborhoods is of concern to policymakers. Since the majority of residents of poverty neighborhoods are children often in female-headed households, they are of particular concern. What happens to children residing in poverty neighborhoods is the subject of considerable debate. There is ample evidence, for example, that many of these children are not adequately nourished, are in need of more regular medical care, are not properly supervised, and are constantly exposed to drugs and criminal acts. But there is no consensus on what should be done to improve the quality of life for these children, let alone who should do it, and how it should be financed. At the same time, most urban dwellers believe that something needs to be done. Passage of the 1996 Welfare Reform Act may have a significant impact on many of these households as mothers are required to transfer from public assistance to work.

To a lesser extent, some of these concerns apply to the elderly who are often trapped in poverty neighborhoods. Even though municipalities, for the most part, have been conscientious in providing some services to the elderly (most notably transit, housing, and nutrition), environments in which these elderly reside, by most accounts, can be hazardous.

A second reason for concern about poverty neighborhoods revolves around the question of equity, particularly in the provision of public services. Poverty neighborhoods contain the oldest housing stock most in need of repair, the most deteriorated infrastructure, have the highest levels of toxic waste concentrations, and the least amount of open space, particularly in comparison with more affluent neighborhoods. Given these conditions, how should public resources be allocated to address these problems?

A third reason for concern is that poverty neighborhoods are believed to be the cause of a number of undesirable social conditions, high crime rates, family disintegration, incidences of major health problems including high infant mortality, tuberculosis, HIV/AIDS, alcoholism, and other antisocial behavior. These conditions if left unchecked could spread to other neighborhoods.

Finally, the concern with poverty neighborhoods centers around the widely held belief that these neighborhoods adversely impact other neighborhoods. Here the concern is often with crime and violence, the impact on property values, and how much of a tax drain poverty neighborhoods cause for other areas of the city. There is also the concern, as was evidenced in Los Angeles in 1992, that intolerable conditions can lead to civil unrest which has proven to be extremely costly (Orfield, 1997; Pastor, 1993).

Conclusion

After nearly four decades of research and efforts at policy reform and program intervention, one thing is abundantly clear – the number of persons in poverty is increasing and persons in poverty are becoming more and more concentrated. Between 1970 and 1990, the number of poor people in high-poverty neighborhoods nearly doubled, from 1.9 million to 3.7 million. This rate of growth exceeded the overall rate of increase in the total population in poverty neighborhoods during this same time period, in spite of the fact that during the 1980s the nation experienced good economic times. Most perplexing is the growing poverty rate among households where one or both adults are working. A robust economy might have provided jobs but not at wage levels that moved people out of poverty.

What is wrong? One might conclude that problems of poverty cannot be solved. Or one might suggest that those who are responsible for providing policy leadership on this issue simply have not found the right solution yet. More disturbing, however, is the thought that the vast number of Americans not in poverty simply do not care much about those who are.

Clearly, there is no simplistic solution to the question of poverty. Indeed this may be one of the problems. Politicians in their zeal to provide solutions respond to constituents or press accounts which usually focus on symptoms or sensational exposes. Researchers tend to focus on individual policy issues. What is needed, perhaps, is a more comprehensive policy approach which takes into account the structural sources of poverty, its multigenerational roots, and the link between individual neighborhoods and regions.

See also: Community Sociology; Neighborhood Revitalization and Community Development; Neighborhood; Poverty Policy; Poverty in History; Poverty, Culture of; Poverty, Sociology of; Prostitution and Sex Work; Residential Segregation: Geographic Aspects; Underclass; Urban Geography; Urban Sociology.

Bibliography

Banfield, E., 1970. The Un-Heavenly City. Little, Brown and Company, Boston.
Bernstein, J., McNichol, E.C., Lawrence, M., Zahradnik, R., 2000. Pulling Apart a State-by-State Analysis of Income Trends. Center on Budget And Policy Priorities Economic Policy Institute, Washington, DC.
Bluestone, B., Harrison, B., 1982. The Deindustrialization of America: Plant Closing, Community Abandonment, and the Dismantling of Basic Industry. Basic Books, New York.
Boger, J.C., Wegner, J.W. (Eds.), 1996. Race Poverty and American Cities. The University of North Carolina Press, Charlotte, NC.
Boston, T.D., Ross, C.L. (Eds.), 1997. The Inner City Urban Poverty and Economic Development in the Next Century. Transaction Publishers, New Brunswick, NJ.
Bullard, R.D., Grigsby, J.E., Lee, C., 1994. Residential Apartheid: The American Legacy. In: CAAS Urban Policy Series, vol. 2. University of California Press, LA.
Clark, K.B., 1965. Dark Ghetto: Dilemmas of Social Power. Harper & Row, New York.
Denton, N.A., Massey, D.S., 1998. Residential segregation of blacks, Hispanic, and Asians by socioeconomic status and generation. Social Science Quarterly 69, 797–817.
Drake, S.C., Cayton, 1945. Horace R. Black, Metropolis: A Study of Negro Life in a Northern City, vol. 1. Harper Tourch Books, New York.
Dubois, W.E., 1967. The Philadelphia Negro: A Social Study. Schocken Books, New York.
Harrington, M., 1963. The Other America Poverty in the United States. Penguin, New York.
Jencks, C., 1993. Rethinking Social Policy: Race, Poverty, and the Underclass. Harper Perennial, New York Reality. Working Paper. The Urban Institute, Washington, DC. Unpublished manuscript.
Jargowsky, P.A., 1977. Poverty and Place Ghettos, Barrios, and the American City. Russell Sage Foundation, New York.
Kain, J.E., 1968. Housing segregation, Negro employment, and metropolitan decentralization. Quarterly Journal of Economics 82, 175–197.
Kain, J.E., 1992. The spatial mismatch hypothesis: three decades later. Housing Policy Debate 3, 371–460.
Massey, D.S., Denton, N.A., 1993. American Apartheid Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Mincy, R., Wiener, S., 1993. The under Class in the 1980s: Changing Concept. Constant Reality. Working Paper. The Urban Institute, Washington, DC. Unpublished.

Oliver, M.L., Shapiro, T.M., 1995. Black Wealth White Wealth: A New Perspective on Racial Inequality. Routledge, New York.

Ong, P. (Ed.), 1994. The State of Asian Pacific America: Economic Diversity, Issues & Policies. LEAP Asian Pacific American Public Policy Institute and UCLA Asian American Studies Center, Los Angeles.

Orfield, M., 1997. Metropolitics: A Regional Agenda for Community and Stability. Brookings Institute Press, Washington, DC.

Pastor Jr., M., 1993. Latinos and the Los Angeles Uprising: The Economic Context. The Thomas Rivera Center, Claremont, CA.

Pastor Jr., M., Drier, P., Grigsby III, J.E., Garza-Lopez, M., 2000. Regions that Work. How Cities and Suburbs Can Grow Together. University of Minnesota Press, Minneapolis, MN.

Sunquist, J.L. (Ed.), 1969. On Fighting Poverty. Basic Books, Inc., New York.

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. The University of Chicago Press, Chicago.

Urban Regeneration

Claudio de Magalhães, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article discusses the concept of urban regeneration and summarizes the types of policies that make up urban regeneration as a public policy field. As urban regeneration policies are heavily context-dependent, the article uses the example of the United Kingdom to illustrate how urban regeneration policies have emerged and how they have been shaped by urban problems and by the dominant responses to them. The article concludes with a discussion of the current challenges to urban regeneration as both a concept and a policy field.

The Concept of Urban Regeneration

Defining what 'urban regeneration' is and what sort of interventions and policies that term encompasses is not straightforward. At the broadest level, it has been applied to any significant intervention improving rundown urban areas and is roughly synonymous with terms such as 'urban renewal,' 'urban revitalization,' or 'urban renaissance,' although there are subtle but not entirely negligible differences in meaning between them. 'Urban regeneration' has also become the most widely accepted translation in British English for expressions in other European languages designating similar interventions: stedelijke vernieuwing, renouvellement/renovation/regeneration urbaine, stadterneuerung, renovación urbana, etc.

The idea that urban areas need regenerating comes from a particular understanding of the causes of the problems of economic and social decline in industrial cities and of the appropriate policy responses to those problems. That understanding has at its core the proposition that the economic decline of a locality, with its attending social and environmental problems could be more effectively addressed as problems *of* that locality rather than as economic, social, or environmental problems that happened to take place in that locality. This notion began to take hold in the late 1960s in policy programs addressing social deprivation in inner-city areas (McCarthy, 2007) and was later to inform a wide variety of urban policy interventions, especially but not exclusively those concerned with improving the physical fabric of urban areas as a means to achieving wider policy goals. From its emergence as a distinct policy field in the 1970s to today, urban regeneration has come to signify that field of public policy that deals with the regrowth of economic activity, the restoration of social function or social inclusion, and the re-establishment of environmental quality in localities in which those elements have deteriorated (Couch and Fraser, 2003).

Urban regeneration policies seek to address the problems of a locality in all their multiple dimensions. A well-known British definition of urban regeneration emphasizes this interplay of physical, economic, and social dimensions of urban problems:

> "Comprehensive and integrated vision and action which leads to the resolution of urban problems and which seeks to bring about a lasting improvement in the economic, physical, social and environmental condition of an area that has been subject to change." (Roberts and Sykes, 2000: 17)

However, this multidimensionality of the concept of urban regeneration begs the question of what is that interaction, what are the causal nexuses it contains, and how they could be acted upon and in what direction. The notion of urban regeneration is based on the assumption that economic and social problems are not a-spatial, and the character of a location can determine the nature of those problems and compound them. Extensive areas of physical dereliction or inadequate physical infrastructure are likely to deter investment in a locality and make the spontaneous renovation of those areas far more difficult. The spatial concentration of socially disadvantaged groups might compound their problems by weakening their interactions with the rest of society and reinforcing patterns of social exclusion. It should be noted that some of these causal links have been assumed rather than demonstrated (see e.g., Cheshire, 2006).

Another element in the notion of urban regeneration is that the decline (physical, social, or economic) is of a structural nature and will not be reversed spontaneously as part of the normal process of change and adaptation that is part of the life of cities. In this case, change requires purposive action by the state through policy and direct intervention, as market forces are not sufficient to trigger and ensure adaptation or transformation. Another definition of urban regeneration puts it in this way:

> "The large scale process of adapting the built environment, with varying degrees of direction from the state." (Jones and Evans, 2013: 2)

The classical problems addressed by urban regeneration policies are widely acknowledged in the literature as those pertaining to the deterioration of the social, economic, and physical health of cities or parts of cities in the older industrialized economies of Europe and North America. These are in turn the result of structural changes in the last quarter of the twentieth century that saw the demise of traditional industries often leaving behind a weak economic basis, persistent problems of unemployment, and derelict sites; the suburbanization of the middle classes taking away from the city its tax base; the economic, physical, and environmental decline of central areas of cities with a deterioration of living conditions for its dwellers; and increasing and harder to bridge social inequality (Karadimitriou et al., 2013).

The problems addressed by urban regeneration are often described as examples of the 'wicked issues' of the policy literature. They are policy problems that seem intractable, persistent, and not amenable to simple solutions. Moreover, they are multifaceted; their consequences span over the areas of competence of many different government agencies; they are not amenable to solutions coming from one single powerful governmental actor and the involvement of actors beyond government; the timing for solutions to work tends to be much longer than conventional program horizons and electoral cycles.

There is also a scalar dimension to urban regeneration problems. Gripaios (2002) suggests that those problems could be classified roughly in two categories. The first comprises deteriorated areas within otherwise relatively prosperous cities, whose economy and population have persistently failed to benefit from the surrounding prosperity. The second encompasses those cities and areas of cities situated in regions with a depressed economy, which require interventions that go well beyond the locality and the city.

The multidimensionality of the problems addressed by urban regeneration interventions underpins the complexity of the objectives those interventions seek to achieve. These are marked by tensions, not just between different views of what a better city or city quarter might be, but also between different perceptions of the role of those different dimensions in bringing about the desired outcomes. The tensions between physical redevelopment and environmental improvement on the one hand, and social and economic benefits on the other are well documented (Turok, 1992; Swyngedow et al., 2002). Similarly, Butler and Hammet (2009) point out to a longstanding tension between encouraging economic competitiveness, with an emphasis on physical transformation, and encouraging social inclusion.

Consequently, over the nearly four decades in which urban regeneration policies have been pursued, their objectives have varied not only because of the varying nature of the problems associated with urban decline. The variation is also the result of different understandings of how multiple dimensions of regeneration relate to one another and how that relationship could be mobilized toward the desired aims, of different ideological perspectives from successive governments and varying balance of power of the potential beneficiaries and losers of regeneration policies. Across countries, different policy and institutional contexts have also shaped variations in urban regeneration policy objectives.

Current regeneration practice accepts that interventions need to go beyond physical redevelopment, and while this remains an important component, the main objectives tend to refer to stimulation of economic growth, together with the decrease of social inequality and the increase in community cohesion. More recently, environmental sustainability and climate change issues have gained considerable prominence as objectives of regeneration policy (Jones and Evans, 2013).

Those objectives are still quite broad and admit several different conceptions of how to attain them. As the public sector in most countries no longer has the capability or the appetite for large-scale direct intervention in the urban environment, achieving the economic goals of urban regeneration has come to depend largely on the performance of the property market. Stimulation of economic growth has often meant the provision of accommodation for businesses, or even just the employment generation and economic multiplier potential of the development industry (Turok, 1992). In both cases, the success of the intervention relies on altering the dynamics of local property markets. New buildings in renovated surroundings with a buoyant property market can change investor's perception of a locality and lead to more investment in the local economy. There is a well-explored connection between urban regeneration and city marketing, with physical changes in a locality (often through flagship developments) being used to boost the competitiveness of a location in attracting businesses, dwellers, and investment (Evans, 2005). In many cases, the re-integration of a derelict area into the mainstream property market itself has been seen in itself as a regeneration objective. The objective of stimulating the economy has also been translated into investment in urban infrastructure, especially as a way of increasing the competitiveness of urban regions (with e.g., improvement in public transport networks), although much of the multiplier effect of this investment also tends to depend on the stimulation of property markets. However, since the 1990s, the principle that regeneration should not be circumscribed to physical redevelopment has gained wide acceptance in theory, if not always in practice (Imrie et al., 2009).

Social equity objectives were originally restricted to housing and the replacement of inadequate dwellings. They were addressed mainly through the provision of more and/or better housing units, mostly by the state but subsequently with increasing degree of dependence on the workings of the housing market. More recently, this has come to include incentives to the creation of mixed income communities as a way of breaking the spatial concentration of poorer households, offer more opportunities for social mix, and thus lead to higher levels of social inclusion. Those objectives have also been frequently addressed through the provision of health, education, or leisure facilities, or indirectly, with the provision of employment that would result from the interventions described in the previous paragraphs.

The pursuit of economic competitiveness has gone beyond property redevelopment and has encompassed initiatives to foster the skills and competencies of local population through social capital building programs. In parallel, perception of the social objectives of regeneration as alleviating poverty and social deprivation has broadened to those of propagation of a cohesive society and a socially inclusive city (McCarthy, 2007; Imrie et al., 2009). These broader social objectives have been addressed through skills and training programs and social capital building initiatives, including communal asset ownership and community integration programs. The list also includes initiatives to reinforce community identity and change the perception of a locality among both its residents and outsiders removing stigma and injecting pride about the place (leading to stronger citizenship), and initiatives addressing the causes and consequences of crime and fear or crime.

Environmental sustainability objectives have now become an integral part of urban regeneration policy agendas as well. They have been pursued with varying degrees of effectiveness through decisions on location, development layout, density,

and use mix, infrastructure provision and building specifications, and indirectly through making possible and incentivizing more sustainable lifestyles (Flint and Raco, 2012).

However, although urban regeneration is often regarded as a comprehensive, holistic discipline, encompassing all the aspects described earlier, in practice it is rarely, if ever, truly comprehensive. The complexity of the problems it tries to address, together with the nature of the government machinery available to implement the policies and the balance of interests associated with each objective means that most urban regeneration practice is skewed toward either physical/environmental objectives, or economic growth and competitiveness or social inclusion. Tallon (2010) mentions a categorization of urban regeneration into 'people,' 'business,' and 'place'-based regeneration, reflecting that division. He also suggests a similar classification of economic, social/cultural, physical/environmental, plus a fourth category of 'governance,' encompassing those urban regeneration interventions whose main focus is on the engagement of local communities, the institutional framework for that engagement and the skills required for it.

Most noticeable has been the disjuncture between urban regeneration interventions with a physical and economic focus, and those with a neighborhood focus (or place/business and people focus, to use the terminology suggested earlier). In the UK, during the New Labour government (1997–2010), social inequality and community cohesion issues were increasingly separated from urban regeneration, and subsumed under the discourse and practice of *neighborhood renewal* (Jones and Evans, 2013). This term has been used to signify a much stronger focus on social issues than mainstream urban regeneration, with its increasing emphasis on economic development and competitiveness trough changes in the built environment. On the opposite end of the spectrum, in the first half of the 2000s, the term urban renaissance came to be understood as a form of urban regeneration that sought to address economic and social objectives through the promotion of high-quality urban and architectural design in towns and cities, in a spirit similar to US-based New Urbanism.

However, to understand how these different concepts of urban regeneration objectives and their associated policies and practices came about, we need to look at the context in which they emerged. The following section explores the history of urban regeneration in the UK as an example that illustrates the links between policy substance and context.

Urban Regeneration: Historical Evolution in the UK

As mentioned earlier, some of the core policy elements of what we now understand as urban regeneration began to appear in the 1960s as part of efforts to tackle problems of decline of cities in industrialized economies. In the UK, these polices were also part of the effort of *reconstruction*, not only of war-damaged urban areas, but also of the significant heritage of nineteenth century slum dwellings. In practice, these policies were very similar to those carried out at the time in the United States to counteract urban decline under the title of urban renewal. They involved extensive demolition of the inner cities and their replacement with modern buildings, mass housing, and road and transport infrastructure.

However, the demolish-and-rebuild solutions to problems of urban decline became gradually seen as inadequate or downright harmful to the communities they sought to help. Urban riots and unrest in the late 1960s demonstrated the inability of those interventions to deal with the complex social problems caused by de-industrialization, suburbanization, and the ghettoisation of poorer neighborhoods. In the UK, the first policies that would fit the contemporary understanding of urban regeneration appeared in the late 1960s and early 1970s (McCarthy, 2007). These were central government grants to local government for targeting urban areas with concentrated poverty and social needs.

As the decline in traditional urban industries and related unemployment, deterioration in living conditions, and physical dereliction continued throughout the 1970s and early 1980s, the economic dimension of the urban regeneration challenge grew in importance. That was now seen primarily as one of facilitating the restructuring of local economies and ensuring that they remained competitive, and that the significant investment embedded in existing physical and social infrastructure was not wasted. To the pro-market governments of the 1980s, that meant freeing up market forces and allowing them to trigger the regeneration of declining urban quarters. It also meant a reliance on property markets as the main instrument for the delivery the physical infrastructure for a regenerated local economy. Emblematic of this period are the large private-led but state-supported flagship projects such as Canary Wharf in London, which was meant to redefine the economic role of its location and create the wealth that would ultimately address problems of social need and deprivation. Whether or not they succeeded or were even capable of doing that is still a contentious point (see Swyngedow et al., 2002; Imrie et al., 2009).

Over time, some of these interventions did indeed succeed in reshaping the local economy, but their record in addressing the underlying social problems is far more questionable. Turok (1992) pointed out early on the limits of regeneration strategies based on an excessive reliance on property development and the weakness of the claims made in its favor. These strategies were always controversial, but in the end the property market crash of the late 1980s made their limitations clear. British urban regeneration policies in the 1990s went back to a previous model of area-based initiatives, spatially concentrated interventions with a broad set of aims including improvement in the housing stock, health and education facilities, training and employment, public realm, etc. However, differently from their 1960s and 1970s predecessors which had been awarded through need, these new area-based programs were secured through competition and required partnership arrangements between public sector bodies, private sector player, and increasingly voluntary and community sector organizations. In this, as pointed out by Leach and Percy-Smith (2001), urban regeneration reflected the increasing dominance of New Public Management approaches to government, with its emphasis in introducing private sector management practices into the running of the state.

Environmental sustainability objectives have been incorporated in urban regeneration agendas from the late 1990s and

in the following decade as part of concerted international efforts. This is manifest in the effort to reuse and regenerate previously developed land, or 'brownfields,' to increase urban densities and secure compact cities, and to better integrate regeneration projects and public transport networks. This period saw also a renewed emphasis on communities, expressed in efforts to address the use and income mix of neighborhoods, as well as their environmental quality. However, as suggested earlier, the emphasis on community and neighborhood regeneration meant also a tension with the economic development dimension of urban regeneration, which has remained the dominant component. In the UK as in other countries, the reconfiguration of the state toward a more reduced role led to a reliance on large property development projects with a significant component of private capital for the delivery of economic development objectives and the provision of urban infrastructure and all types of social and merit goods (Karadimitriou et al., 2013). Given the increasing role of culture and knowledge-related activities in the urban economies of industrialized countries, 'culture-led' regeneration became a common strategy, bolstered by the success of interventions in Bilbao and Barcelona, and also the Tate Modern in London and other examples elsewhere (Evans, 2005). Most of these interventions relied on the creation of flagship buildings with cultural functions as a way of changing the perception of a location at the national and often international level. The regenerative effects would come with increasing levels of inward investment, tourism (with all its multiplier effects), and a larger resident population, attracted by the new cultural facilities.

The financial crisis and economic downturn of 2008 slowed down the majority of regeneration interventions based on physical redevelopment not only in the UK but also in most of Europe, as property markets slumped. Paradoxically, austerity policies and cuts in public sector budgets have made more difficult for public sector bodies to meet urban regeneration objectives on their own. In the UK, the coalition government that came to power in 2010 has reacted to those challenges by concentrating government funding into economic regeneration, with grants, tax incentives, and aid to interventions that could create employment, boost the economy, and increase its competitiveness (Jones and Evans, 2013). On the other hand, under the banner of localism and 'the big society,' government has signaled its intention to devolve aspects of regeneration, and especially neighborhood regeneration to civil society, with incentives to the proliferation of community trusts, social enterprises, business improvement districts, local enterprise partnerships between businesses and local authorities, and similar mechanisms.

Finally, it is worth noticing that for most of its history hitherto, urban regeneration has been the concern of the industrialized economies of Europe and North America. However, in the last few years policies addressing urban decline and the need for urban redevelopment have been adopted in the recently industrialized economies of East Asia and elsewhere. South Korea has embarked in urban regeneration interventions to cope with the legacy of fast and unplanned urban growth (Shin, 2009), and more recently, China has started to look for solutions for the problems of urban degradation created by the phenomenally fast growth of its cities (He and Wu, 2007). These policies tend to follow the extensive demolition and rebuilding model of urban renewal, but there are also signs that these countries are rapidly becoming aware of the limitations of that model.

This very brief historical overview takes us to the current challenges facing urban regeneration and urban regeneration practices and policies.

The Challenges for Urban Regeneration

From the earlier discussion it becomes clear that urban regeneration is foremost an empirical policy field that tries to grapple with the complex connections between the built environment, the economy, and society at different scales and to intervene in processes and variables that are possibly too complex and often with unclear causation. However, the problems of urban decline are real and can dramatically affect the economic health and the social well-being of neighborhoods and put considerable pressure on governments. Urban regeneration practice has often tried to replicate 'what works,' even if it often cannot explain why it works or whether it would work in a different context. Time, however, has allowed lessons to be learned from the various regeneration interventions with their varying degrees of success. Nonetheless, some issues seem to persist and challenge regeneration practice as a whole. Some of the most important ones are discussed.

The first issue is the unresolved causal nexus between the built environment and the social and economic processes that take place in it. Current regeneration practice has gone a long way from simplistic assumptions of straightforward causality of the postwar years, but the relationship between physical improvement and social and economic outcomes is still poorly understood – if not in theory, certainly in regeneration practices. The existing evidence is inconclusive about whether regeneration programs based on physical restructuring and redevelopment have been the main determinant of long-term structural changes in the economy and social well-being of the areas where they have been applied, and whether change was in the intended direction and the benefits were received by those to whom they were intended (see Cheshire, 2006; Ball and Maginn, 2005; Webber et al., 2010). The persistence of such approaches in e.g., regeneration interventions associated with large sporting events, such as the 1992 Olympic Games in Barcelona Commonwealth Games in Manchester in 2002, the Sydney Olympics in 2000, or the recent 2012 London Olympics suggests a strong belief in the transformative powers of physical redevelopment. Clearly, the production of new buildings in previously derelict areas, given the right conditions, can attract an influx of occupiers and this in itself will change the economic profile of the target area as in Canary Wharf or Kings Cross in London. Similarly, the development of iconic buildings in the right location and circumstances (e.g., the Guggenheim Museum in Bilbao or the Tate Modern in London) can boost the local economy and help achieve broader social goals – although it might also cause displacement and gentrification and therefore direct the benefits from regeneration away from those to whom they had been targeted. However, the evidence is much less conclusive as regards more subtle and complex social objectives, such as the

creation of social mix through the provision of different types of housing or the creation of a sense of place through new and renovated public space (see Madanipour, 2005; Cheshire, 2006 for conflicting perspectives). There is also the question of whether the causal nexus, if and when it is present, is strong enough to counteract other social and economic determinants and produce all the outcomes that are expected in an urban regeneration intervention.

Urban regeneration is by nature place-focused and the second issue is that of the effectiveness of localized interventions in dealing with social and economic problems that have their roots in the structure of the regional or national economy and society, and might not be amenable to solutions at local level. This is the case e.g., with issues of unemployment, levels of skills in local population, and many of the factors that determine social exclusion. Even when these problems can be addressed to a modest extent at the level of a locality, the benefits produced by regeneration interventions are likely to leak out of the target area and community, as there are no spatial barriers preventing it. Similarly, incentives to improve the economic performance of a deprived area by e.g., providing tax incentives of new accommodation to incoming businesses might just lead to displacement of businesses from elsewhere if there are no changes in the regional economy (see Rhodes et al., 2005).

A third issue is the nature of the knowledge required for effective urban regeneration. Because of the complexity and multidimensionality of the problems urban regeneration has to deal with (see the earlier reference to the 'wicked issues'), no one single actor is likely to have all the necessary knowledge or the capacity to solve those problems on their own. The need for partnerships of various kinds between public sector bodies (especially local authorities), the private sector, voluntary sector organizations, and communities has been widely recognized in urban regeneration theory and practice. For reasons of efficiency as well as equity, the need for participation, partnership work, and co-production of policy outcomes has been reaffirmed in nearly all the urban regeneration literature (see Harding, 1998; Ball and Maginn, 2005). However, these different actors have different capacities for expressing their knowledge and integrating it in coherent set of objectives. Moreover, there is no reason to expect those objectives not to be conflictive with those of other actors, since they will derive different kinds of benefits from an urban regeneration intervention.

This takes us to the fourth issue. An enduring challenge in urban regeneration practice is to ensure that the appropriation of the benefits it creates serves the intended policy objectives and that those it was supposed to benefit actually do so. Urban regeneration implies a concerted effort to change the configuration and the social and economic dynamics of a place. This process almost always produces both winners and losers, but determining and controlling who they are is far from simple. Power structures in society, amounts of social capital accessible by different groups, distribution of property rights, mobilization capacity etc., will determine who captures what benefits (Cars et al., 2002). A common critique of urban regeneration through property development projects is that its benefits are more likely to be appropriated by development interests, local government, and incoming residents than by existing residents and businesses, whereas the disbenefits are more likely to fall to the latter (Imrie and Thomas, 1993; Swyngedow et al., 2002). It is a contentions point whether this is because converting regeneration benefits into property values, rents, local taxes, local GDP growth, and other monetizable outputs is more straightforward than converting them into less tangible and more uncertain social benefits for resident communities, as opposed to this being the result of an inherent structural bias in physical, property-based regeneration.

This is a weaker version of the argument that says that the inequalities in the appropriation of the benefits of urban regeneration are structural and intentional, and result from the dominance of neoliberal, business-dominated approaches to city management, and urban policy, which facilitates the appropriation of the city by large finance capital (Swyngedow et al., 2002; MacLaran, 2003). According to this argument, urban regeneration projects are essentially instruments to promote the re-valorization of land and buildings, gentrification, and the re-appropriation of parts of the city by the elites, the main beneficiaries of the economic impacts of such projects. Some of this argument draws on Harvey's (1985) formulation about the relationship between the built environment and capital accumulation, by which capital switches to the built environment when the profit rates in other productive spheres fall (see also Ball, 1986 for the critique). However, the fact that cities all over the world, under governments of different ideological profiles, insist on pursuing urban regeneration interventions with a significant component of land and property redevelopment would suggest that even with structural inequalities of voice and power, there is still the possibility of producing benefits that can reach disadvantaged groups. The key issues are therefore, how to control the appropriation of benefits and, in the unavoidable occurrence of losers, ensure that their concerns are also addressed. Once again, partnerships and other participatory mechanisms from policy definition to the long-term monitoring and management of outputs have been so far the answer.

The earlier discussion raised the question of how the public interest and, more specifically, the interests of deprived communities might be protected in urban regeneration interventions. For much of the history of urban policy in representative democracies, this was seen as the natural role of the state and the public sector. However, changes in the nature of local governance in mature democracies has led to an increase in collaborative forms of governing, with a less clear separation of roles between public and private sector and between government and civil society (Leach and Percy-Smith, 2001; Sullivan and Skelcher, 2002). The expression 'enabling state' has been used to describe a situation in which the state does not seek to deliver policy outcomes by itself, but rather enables others to do it. Urban regeneration is not an exception, and in many countries, regeneration policy objectives are predicated upon interventions in which the state is only one and not necessarily the dominant partner. The fifth issue is therefore how to secure accountability and representativeness in partnerships and other similar governance mechanisms, in a way that ensures that all those with a stake in the regeneration intervention are represented, but also that the conflicts that are unavoidable in that sort of

policy action do not prevent it from taking place (see Osborne, 2010).

A sixth issue relates to the nature of resources for urban regeneration. Traditionally, urban regeneration has relied on a mix of public funds and subsidies leveraging private investment. The more recent global economic downturn has affected the status-quo because of increasing pressure on public sector budgets. As a result, private and community resources have become essential for the delivery of policy outcomes. There has been an increasing reliance on private developer to deliver social goods (in the UK, this includes nonmarket housing) or in community organizations to secure long-term maintenance of amenities and social facilities. One of the main implications of this role of public, private, and community actors is that urban regeneration interventions require governance arrangements that ensure that the different types of uncertainties and risks faced by the actors are considered and reduced as far as possible (Karadimitriou et al., 2013). Urban regeneration is inherently uncertain and risky, and this includes financial and market risks for private partners, financial and political risks for public bodies, and risks to the welfare for the communities involved. Negotiating and arranging how risks and rewards are distributed is becoming therefore an important function of public sector bodies in multiactor regeneration interventions. In recent times, instruments such as land-readjustment, public land ownership, dedicated funding mechanisms, direct public investment, and tax mechanisms have been deployed with this purpose. The discussion around value capture – how the benefits and positive externalities caused by regeneration interventions can be captured and utilized to pay for public goods – is relevant in this regard (see Ingram and Hong, 2012). Mechanisms such as tax increment finance, planning obligations, and levies have been used with different degrees of success to convert what otherwise might be appropriated as private value into sources of funding the public goods urban regeneration should produce.

A final issue refers to the long-term sustainability of urban regeneration interventions, in the fullest sense of environmental, social, and economic sustainability. Under the influence of national and international directives, sustainability has become an essential objective in urban regeneration. The environmental dimension of sustainability has been addressed satisfactorily through considerations about the carbon footprint of renovated urban quarters, the energy efficiency of their buildings, and more broadly through public transport accessibility and the density and compactness of the urban form (Jones and Evans, 2013). However, this is not so with the other two dimensions of sustainability: social and economic sustainability are well defined in the literature, but their practical incorporation into policy is still problematic. A combination of urban design, participatory governance regimes, public transport investment, and strategic municipal planning and land acquisition has demonstrated the possibility of environmental and social sustainability in places such as Freiburg in Germany or Hammarby Sjöstad in Sweden. Nevertheless, generalizing such exceptional experiences remains difficult and the capacity of local level interventions to change structural processes that are regional, national, and often global remains a real issue.

Final Remarks

Urban regeneration has become the major component of urban policy in many countries. However, rather than a discipline firmly anchored in a solid body of theory, it is an empirical field of policy, trying to address the perceived problems of urban decline in all their complexity on the basis of interventions that have worked in similar circumstances and of views of causality that have not (yet) been disproved in the real world. As suggested earlier, its agenda reflects the prevailing understandings of urban problems, their causes, and remedies, themselves contested by the various interests affected by urban regeneration interventions. Nonetheless, those problems are real, and the need to reuse urban resources in a sustainable and socially inclusive way means that urban regeneration interventions will remain important tools for dealing with urban change. How well they will be able to produce the kinds of policy outcomes expected of them might well depend on how challenges such as those listed earlier are addressed and what sort of urban regeneration practice emerges from that.

See also: Gentrification; Housing and Cities; Urban Design and Public Space; Urban Policy in Europe.

Bibliography

Ball, M., Maginn, P., 2005. Urban change and conflict: evaluating the role of partnerships in urban regeneration in the UK. Housing Studies 20 (1), 9–28.

Ball, M., 1986. The built environment and the urban question. Environment and Planning D: Society and Space 4, 447–464.

Butler, T., Hammet, C., 2009. Regenerating a global city. In: Imrie, R., Lees, L., Raco, M. (Eds.), Regenerating London: Governance, Sustainability and Community in a Global City. Routledge, Abingdon, pp. 40–57.

Cars, G., Healey, P., Madanipour, A., De Magalhães, C. (Eds.), 2002. Urban Governance, Institutional Capacity and Social Milieux. Ashgate, Aldershot.

Cheshire, P., 2006. Resurgent cities, urban myths and policy hubris: what we need to know. Urban Studies 43 (8), 1231–1246.

Couch, C., Fraser, C., 2003. Introduction: the European context and theoretical framework. In: Couch, C., Fraser, C., Percy, S. (Eds.), Urban Regeneration in Europe. Blackwell, Oxford, pp. 1–16.

Evans, G., 2005. Measure for measure: evaluating the evidence of culture's contribution to regeneration. Urban Studies 42 (5/6), 959–983.

Flint, J., Raco, M. (Eds.), 2012. The Future of Sustainable Cities: Critical Reflections. Policy Press, Bristol.

Gripaios, P., 2002. The failure of regeneration policy in Britain. Regional Studies 36, 568–577.

Harding, A., 1998. Public-private partnerships in the UK. In: Pierre, J. (Ed.), Partnerships in Urban Governance: European and American Experience. Macmillan, Basingstoke, pp. 71–92.

Harvey, D., 1985. The Urbanisation of Capital. Blackwell, Oxford.

He, S., Wu, F., 2007. Socio-spatial impacts of property-led redevelopment on China's urban neighbourhoods. Cities 24 (3), 194–208.

Imrie, R., Lees, L., Raco, M. (Eds.), 2009. Regenerating London: Governance, Sustainability and Community in a Global City. Routledge, Abingdon.

Imrie, R., Thomas, H., 1993. The limits of property-led regeneration. Environment and Planning C: Government and Policy 11, 87–102.

Ingram, G., Hong, Y.-H. (Eds.), 2012. Value Capture and Land Policies. Lincoln Institute of Land Policy, Cambridge, MA.

Jones, P., Evans, J., 2013. Urban Regeneration in the UK. Sage, London.

Karadimitriou, N., De Magalhães, C., Verhage, R., 2013. Planning, Risk and Property Development: Urban Regeneration in England, France and the Netherlands. Routledge, Abingdon.

Leach, R., Percy-Smith, J., 2001. Local Governance in Britain. Palgrave, Basingstoke.

MacLaran, A. (Ed.), 2003. Making Space: Property Development and Urban Planning. Arnold, London.

Madanipour, A., 2005. Value of place: can physical capital be a driver for urban change? The experience of Castle Vale, Birmingham. In: CABE (Ed.), Physical Capital: How Great Places Boost Public Value. The Commission for Architecture and the Built Environment, London, pp. 48–71.

McCarthy, J., 2007. Partnership, Collaborative Planning and Urban Regeneration. Ashgate, Aldershot.

Osborne, S. (Ed.), 2010. The New Public Governance: Emerging Perspectives on the Theory and Practice of Public Governance. Routledge, Abingdon.

Roberts, P., Sykes, H. (Eds.), 2000. Urban Regeneration: A Handbook. Sage, London.

Rhodes, J., Tyler, P., Brennan, A., 2005. Assessing the effect of area based initiatives on local area outcomes: some thoughts based on the national evaluation of the Single Regeneration Budget in England. Urban Studies 42 (11), 1919–1946.

Shin, H.B., 2009. Property-based redevelopment and gentrification: the case of Seoul, South Korea. Geoforum 40, 906–917.

Sullivan, H., Skelcher, C., 2002. Working across Boundaries: Collaboration in Public Services. Palgrave, Basingstoke.

Swyngedow, E., Moulaert, F., Rodriguez, A., 2002. Neoliberal urbanisation in Europe: large-scale urban development projects and the new urban policy. Antipode 34 (3), 542–577.

Tallon, A., 2010. Urban Regeneration in the UK. Routledge, Abingdon.

Turok, I., 1992. Property-led urban regeneration: panacea or placebo? Environment and Planning A 24 (3), 361–379.

Webber, C., Larkin, K., Tochtermann, L., Varley-Winter, O., Wilcox, Z., 2010. Grand Designs? A New Approach to the Built Environment in England's Cities. Centre for Cities, London.

Urban Sociology

Anne B Shlay and John Balzarini, Temple University, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Scaglia, volume 24, pp. 16082–16087, © 2001, Elsevier Ltd.

Abstract

Space, place, and its connection to social relations make up the study of urban sociology. Urban sociology views cities as the spatial basis for the reproduction of inequality and as an integral part of the capitalist accumulation process. It relies on qualitative and quantitative methods to explain urban development, neighborhoods growth and decline, and their effects on people and their communities. Urban sociologists also rely more heavily on spatial methods to connect social relations to place. Contemporary research topics include global cities, globalizing urbanization, culture and neighborhood change, community and social networks, segregation, neighborhood change, concentrated poverty, and gentrification.

Introduction

Everyone the world over sooner or later becomes part of cities. What were rural places are rapidly merging with the metropolis as urban places continue to grow and conquer the hinterland. People either are moving to cities or finding themselves swept up in tidal forces of metropolitan development. Urbanization is becoming more and more a ubiquitous local, national, and global process. According to the United Nations (2012), sometime in early 2007, world urbanization crossed the 50% threshold, and for the first time in history, a majority of human beings lived in cities. By mid-century, the figure is projected to reach two-thirds. The future of humanity is clearly urban.

Urban sociology is a field designed to understand and make sense of these spatial trends and patterns for the people who live in cities. It is a field that explores city life from the vantage point of economic forces and global cities to a focus on the homeless person hanging out on the corner. It places cities and city life within the broad contours of mega theories, while also looking at how the lives of individuals are affected by the specific characteristics of places. Urban sociology's distinguishing feature is its focus on the role of space in people's lives. It looks at the connection between urban and metropolitan spatial organization and the social and economic situations of different social groups – people who vary by age, race, ethnicity, family type, country of origin, income, and socioeconomic status. Urban sociology embraces the diversity of social life and explores how it is played out across the physical contours of cities.

The focus on space is critical. Urban sociology is not the study of everything socially going on in cities, which ultimately, given the universality of the urban, would dominate the entire field of sociology. Urban sociology is about space – it is about land, development, housing, density, architecture, open space, distributions of land use, urban infrastructure, roads and transportation, neighborhood housing and population composition, and ultimately, location, location, location. Space, place, and its connection to people make up the study of urban sociology. Space literally and figuratively anchors the entire field.

Theory, Perspectives, and Schools of Thought

Urban sociology was born with the Industrial Revolution, the decline of feudalism, and the concomitant shift from a rural to urban society. Following Durkheim (1997 [1893]), these economic and spatial changes, according to Tönnies (1963 [1887]) and Simmel (1950 [1903]), altered fundamental social relationships and brought about society-wide transformations of increasing complexity. Moving from Gemeinshaft (community) to Gesellshaft (society) created an urban social order based on money and work, not family and faith – a condition both socially liberating and disconcerting (Simmel, 1950 [1903]). Economic structural changes and spatial reorganization required new mechanisms to foster social integration on the urban scene – a theoretical beginning that focused on the maintenance of social order and control. This quest for understanding this new urban order gave birth to the Chicago School and the field of urban ecology, which studied the forces underlying human spatial organization – the city.

A paradigm shift emerged with the recognition that cities were not the natural spatial expression of human organization. Instead, they represented the interests of a particular economic system (capitalism), were managed deliberately (through political policy), were used to enhance power and privilege (at the expense of others), and were the basis for the reproduction of inequality. These ideas heightened interest in a critical 'new' urban sociology that along with theory building viewed the field as a source for social change.

The Chicago School

The creation of a specific field called urban sociology took place in Chicago in the early years of the twentieth century, which focused exclusively on urban spatial formations and change and the social behaviors emerging within this new environment. With urbanization becoming a dominant trend, foreign immigrant's as well domestic migrants from rural America flocked to U.S. cities. Chicago was a prime destination. With the rapid and intense growth of the city, Chicago sociology observed the process of urbanization with wonder and awe. The people moving to cities spoke different languages and had

different customs, and world views, making the potential for chaos enormous. Yet order prevailed, both in terms of land use and behavior.

Sociologists turned to the established field of plant and animal ecology for guidance in understanding the prevailing order of human settlement patterns. Sociologists married ecological principles to the study of cities. Within this framework, the Chicago School of Urban Sociology was born. Leading the way was Robert Park, Ernest Burgess, Robert McKenzie, and others who solidified sociology's place as an American discipline focusing on the country's urban experience.

Urban ecology explained spatial dynamics as a complex web of competition between various groups for control over settlement space. Urban spatial organization was viewed as a natural process of environmental adaptation. Concepts including competition, dominance, adaptation, invasion, and succession explained urban spatial order and change.

The concentric zone introduced by Ernest Burgess (1967 [1925]) illustrated how housing and population organized themselves based on these principles. These principles of urban social-spatial organization preserved community by allowing different groups to select themselves into their own urban neighborhoods: Little Italy, Chinatown, and the like. Neither chaotic nor disorderly, the city exhibited a new urban order that the Chicago School sought to describe and explain.

The Chicago School based the creation and recreation of community on discovered processes of how neighborhoods change and conform to principles of segregation. The Chicago School explained segregation like other urban patterns as natural, inevitable, and uncontrollable (Park, 1936, 1967 [1925]; Burgess, 1967 [1925]; McKenzie, 1967 [1925]). As the mechanism through which community was produced and maintained, segregation was also a positive outcome, a process tied to what the Chicago School called the moral order of the city. The professed inevitability of the segregation of both people and land use conformed to prevailing attitudes toward race, class, and ethnicity as well as the economic fundamentals associated with early free market capitalism associated with cities.

Chicago School theorization of the city unilaterally portrayed urban spatial organization as the inevitable product of natural forces beyond the control of any group or individuals. The fundamental belief in the inevitability of urban order as well as the emphatic conviction that urban spatial patterns, particularly segregation, were positive developments eventually was called into question. Recognition that segregation by race represented discriminatory processes, not self-selection, was ultimately rendered illegal by civil rights legislation. Urban inequality, particularly by race, fostered a new look at those theories that appeared to provide sociological justification for an unfair spatial status quo. With the urban race riots of the 1960s, urban sociology's professed natural forces were revealed to be scholarly rationalizations for racism, discrimination, and the ghetto.

The New Urban Sociology, Urban Political Economy, and the Production of Space

The new urban sociology emerged as a deliberate critique of urban ecology. It incorporates the role of power and politics in the guise of the state organizing urban development in the interests of capitalism. The new urban sociology embraces and adapts ideas of Marxist geographers like David Harvey (1973) and Neil Smith (1979). Yet it asserts a role for human agency that often is missing from the more structural orientations of Marxist urbanism. It also retains urban ecology's focus on and belief in methods and research.

With a genesis in community power structure research pioneered by Floyd Hunter (1953), Harvey Molotch (1976) – and later in collaboration with John Logan (Logan and Molotch, 1987) – explicitly introduced the role of power and politics, not ecological forces, in shaping urban development. The city as a growth machine thesis argues that the shape and direction of urban development, particularly of downtowns, is the product of behind-the-scenes land-based activities by locally based coalitions of urban actors who stand to benefit from increased land and property values, land use intensity, and population expansion. These institutional actors include corporations, utilities, newspapers, universities, and unions as well as government who work collectively and concertedly because of shared interests.

With the power to orchestrate local political decision making, growth coalition activities are magnets for public underwriting and subsidy. Although these developments yield significant economic advantages to coalition members, their public payoffs appear minimal or even harmful. The growth machine agenda holds sway because of a well-articulated and disseminated ideology that maintains that growth is in the public interest despite ample evidence to the contrary. Popular acquiescence to growth machine projects is indicative of the hegemonic dominance of growth ideology. Contrary to urban ecology's insistence that cities are organized spatially according to processes that provide mutual benefits to all groups, the city as a growth machine stipulates that cities are a reflection of political, economic, and social manipulation and domination with clear winners and losers.

Marxist geography informed a Marxist urban sociology and it often is difficult to disentangle the disciplinary differences. David Harvey (1973) was the first to advocate for an urban political economy that explicitly links urban development to capitalism. Urban political economy focuses on how cities are involved in the circulation of capital and the reproduction of labor power. As part of capitalism's need to combat the falling rate of profit, cities also serve as the locale for collective consumption (Castells, 1977) – that is, the construction of publicly funded urban infrastructure that supports capitalist enterprises, subsidizes profits, and assists labor force reproduction. Cities are not simply places where capital is accumulated. Cities are an integral part of the capitalist accumulation process.

The physical expansion of cities reflects the use of cities to grow finance capital. Urban development and suburbanization are activities that require investment into the built environment. Urban development represents the shift of capital from manufactured commodities (the primary circuit of capital) to building homes and buildings (the secondary circuit of capital). Capital shifts between primary and secondary circuits reflect cycles of investment and disinvestment between these two circuits of capital (Harvey, 1973).

Urban decline and uneven development in the form of abandonment and neighborhood deterioration is not the absence of activity within these communities, but rather represent the active disinvestment of capital from communities (Smith, 1979). Reinvestment creates values while disinvestment removes it. Disinvestment is not simply a process in which buildings are neglected. Disinvestment devalues property as part of the process for later profitable revalorization. Urban political economy is the study of how capital accumulation actively develops and underdevelops neighborhoods, cities, and communities. The recognition of the crucial link between neighborhood vitality and housing finance led to many urban sociologists partnering with community-based organizations to fight neighborhood disinvestment (Shlay, 1989).

Urban political economy views cities as sites of production and capital accumulation (primary circuit of capital) and as places for consumption (the secondary circuit of capital) and for the reproduction of labor power (the generation of present and future labor market participants). Cities are an integral part of the reproduction of labor power by providing the physical environment in which labor power is recreated daily.

But Manuel Castells (1977) argued that cities are also sites for collective consumption i.e., publicly financed goods and services that are shared by urban residents. Collective consumption subsidizes the wage rate and increases capitalist profits therefore making city services a key factor in capitalist accumulation. In addition, different capacities of labor power are produced through the provision (or lack thereof) of education. Urban political economy implicates city services as mechanisms that create a stratified urban labor force. What is explicitly urban about collective consumption is that this occurs spatially.

Much of the original theoretical work defining urban political economy focuses on the production of space. Under capitalism, this is the initial and subsequent work at turning land into a commodity and using it as a fundamental component of capitalist accumulation and labor force reproduction. Henri Lefebvre (1991, 1996 [1968]), a French urban scholar and political activist, introduced the concept of production of space. The translation of his work was delayed, however, and so his direct contribution to the English-speaking world was mediated through the works of other urban Marxists (Harvey, 1973; Castells, 1977).

According to Lefebvre the ways in which space is produced is an integral part of the economic, political, and cultural dimensions of society. In *The Production of Space*, Lefebvre (1991) argued that capitalist forces produce abstract space, that is, homogenized spaces that are turned into exchange values. The abstraction of space is how urban development is used to create exchange values under capitalism despite the reality that space or spaces are unique locations that never can be abstract. Space cannot be fully abstract because space operates as distinct social places that are constructed by people through their daily activities and lived experiences. These social spaces of use values are areas where the culture of urban life thrives, often in direct conflict with the abstract space idea generated by capitalism. People use – that is, produce – space to live. Capitalism produces space to raise money. These uses are in conflict.

Abstract space, like abstract labor is the means for capitalist appropriation. Just as labor power can be abstracted and sold on the labor market, abstract space permits the exchange of land as a commodity. But the problem for capitalist domination is that labor and space are unique. For Lefebvre this conflict over social and abstract space is the decisive class struggle, leading him to reject the legitimacy of property rights and to argue that urban inhabitants by virtue of their residency have the right to participate in all land use decisions and to appropriate space. This is the 'right to the city' (Lefebvre, 1996 [1968]).

Securitization (the trading of finance raised off of the exchange value of land) has increased the abstraction of space to the point that space itself is no longer featured in financial transactions (Gotham, 2006). This ultimate abstraction of space is its financialization in which land is produced through the sale of financial instruments, that are, at least in theory, attached to the land (Gotham, 2006). Financialization represents land as a series of economic debts, obligations, and returns so that land in rural Idaho is rendered equivalent to land in New York City. The financialization of cities is an essential step for capitalist accumulation.

Urban theory's journey from urban ecology to urban political economy is one that continues to seek to explain urban development, neighborhood growth and decline, and their effects on people and their communities. It has moved its theoretical focus on explaining urban development away from ecological forces of self-selection. Urban political economy looks at the forces of capitalism – the commodification of land, the movement of capital into and out of the city's physical infrastructure, and the conflict for urban residents between property use and exchange values. Urban theory continues to look at large-scale forces that shape the urban scene.

But contemporary urban theory also emphasizes the role of agency – the decision-making power of growth coalitions is about the behind-the-scenes activities of urban actors with shared interests. The power of urban growth coalitions, however, appears to be diminishing with the tidal forces associated with financialization – forces directly tied to decision making within global financial markets.

Global Cities and Globalizing Urbanization

The major contemporary world force is globalization. Globalization involves the development of mobile global capital and a global economic market, the integration of a global division of labor, the rise of transnational corporations, and the diminished importance of nation-states. Theorizing about a global order and cities is the major frontier for urban sociological theory.

The focus of urban theory on globalization emphasizes a shift in how uneven development is addressed. Urban sociology initially addressed the uneven development within individual cities. With globalization, uneven development is a feature of the world. Urban development is expressed globally across an array of increasingly economically integrated cities.

Foreign direct investment in the global core, increased exchange between global financial markets, and the movement of industrial production from wealthier regions to poorer regions have corresponded with increased centralization of various firms and corporations that manage this increasingly complex and ever-growing system. As introduced by Saskia Sassen (1991) and others, global cities represent the 'command and control centers' of the globally integrated economy. Global cities, such as New York, London, and Tokyo, among others, are the locations of the major global stock exchanges, the headquarters of transnational corporations, and world trade and lending organizations. These organizations direct capital flows and investment and command institutional control of global resources and policy decisions, which make global cities the locus of power in an ever-globalizing world.

The concept of globalizing urbanization focuses on how global forces are treated on the ground by local urban institutions and actors. Globalizing urbanization addresses how local urban development reflects international trends associated with heightened immigration, intensifying racial and ethnicity segregation, the spatial polarization of rich and poor, the incorporation of market mechanisms into public planning, widespread gentrification, and the growing specter of fiscal crises at every level of government. Theorizing about globalizing urbanization emphasizes local decision making tied to forces of globalization. This is because globalizing cities are the frontiers for the implementation for global trends, such as neoliberalism, and are sites for local resistance and acquiescence (Brenner and Theodore, 2002).

Culture and Urban Places

Urban sociology's initial focus on culture and its connection to urban development began as a response that was seen as an overly deterministic model represented by urban ecology (Alihan, 1938). These early calls for assessing culture's role in urban spatial patterns argued that attitudes in the form of place-based attachments could counter the strongest economic forces.

Nonetheless, culture was largely ignored within urban sociology as scholars sought to distance themselves from culture-of-poverty arguments that allegedly blamed the victims. Culture as a central variable within urban sociology became more prominent as a feature of the consumption of housing, goods, and services within cities. Culture is now theorized as a critical accompaniment to the dynamics of both local and global forces associated with gentrification. Pioneer urban sociologist Sharon Zukin (1987) illustrated the demise of authenticity with the commodification of culture (and culture's corporate globalization) within gentrifying places. Culture is a market signal that attracts reinvestment and eventually global capital. Culture is a critical part of how urban space is commodified and transformed.

In a very different way, culture is also connected to spatially concentrated poverty (Wilson, 1987). Contemporary concentrated urban poverty is linked to a series of variables, including culture. But how culture is linked to the continuation of urban poverty is argued to be a complex phenomenon based on a combination of structural, spatial, and cultural variables. Those that connect culture to the spatial dynamics of poverty are careful to maintain that culture is not a determining factor of urban poverty but rather is an important feature of concentrated poverty (Wilson, 1996).

Methods in Urban Sociology

Urban sociology was born of a tradition rich in theory as well as method. Urban sociology's founders, the Chicago School, were data fiends embracing both quantitative and qualitative methods, including ethnographic research. The earliest urban sociologists plied their trade by mapping, counting, observing, and surveying. They pioneered participant observation, conducting some of the earliest studies using ethnographic methods.

The early urban sociologists were empirical. They cared about data, how they got it, and how they used it. The emphasis in urban sociology on using a wide range of rigorously applied methods continues. Although urban sociology, like most fields in sociology, periodically engages in pedantic methodological battles, it typically tends to value and support all types of empirical research, emphasizing the quality of the methods used.

This focus on careful empirical work gives urban sociologists standing within applied networks that rely on quality research for policy development, assessment, and application. In the applied research arena, the empirical work of urban sociology can rival the work of urban economics because sociologists, unlike most economists, know to collect primary data. One of the strengths of urban sociology, and typical of sociology generally, is a focus on research design.

An enduring strength of urban ethnographic research is exposing the 'hidden' worlds and clandestine cultures of urban inhabitants. But urban ethnography now studies the unusual as well as the seemingly mundane. Urban ethnography seeks to explore people's lifestyles, behaviors, decisions, and actions as well as the ways people frame their experiences and construct their identities. Urban ethnography captures a piece of urban life and connects these pieces to larger urban structures.

The standards for rigorous quantitative research in urban sociology are high. Urban sociologists rely heavily on primary and secondary data, are skilled in sampling and survey design, and have grown their skills at statistical analysis.

But being rigorous is not the sole domain of quantitative sociologists. Qualitative research in urban sociology also has become much more meticulous about method. The best urban researchers appear to understand the utility of both methods and many now heed the call for using multiple methods as part of a multifaceted research design (Small, 2009).

Urban sociologists also rely more heavily on spatial methods that permit them to connect social relations to place (Logan, 2012). The ideas behind the methods are less new than the ability to manipulate more data in a spatial way. More data sets contain geographic identifiers. Computerized mapping, geographic information system (GIS), and other software permit more sophisticated analysis.

Topics in Urban Sociology

Many make a fundamental mistake of thinking that the field of urban sociology is the generic study of people living in cities. This conflation stems in part from the media, which labels particular types of music as urban and stereotypes of what urban represents (e.g., minorities, criminals, poor). So it often is assumed that the studies of racial inequality, poverty, crime, and so on are *de facto* topics in urban sociology.

But studying people in cities is not urban sociology. Urban sociology is spatial in its orientation. If space is not a salient element in the research, it is not urban sociology. Place and space are the fundamental ingredients of urban sociology, a focus that links spatial location to social relations. This definition, therefore, provides some limits around the definition of topics within the field of urban sociology. Although many sociologists study race and many racial minorities live in cities, studying race is not urban sociology unless space and location are critical analytical variables.

This article highlights a few of most important topics in the field of urban sociology. These are community and social networks, segregation, neighborhood change, concentrated poverty, and gentrification.

Community and Social Networks

By definition, much of the discipline of sociology is about community – the underlying mechanisms that attach people together. Indeed, the construction, maintenance, and source of community were major topics for early urban sociologists who were concerned about the loss of social cohesion with the transition from rural to urban society.

Much of urban sociology is concerned with role of place-based communities in social and economic life. Studies of communities examine issues of power, class, gender, family composition, housing type and tenure, ethnic and racial diversity, crime and safety, and employment. Of particular importance are questions associated with the role of community in social mobility.

A central focus of contemporary community studies is the role of social networks (Wellman, 1999). Social networks, a key aspect of community, represent the basis for acquiring social capital. Social capital consists in part of relationships that may provide access to employment and other resources. A major question within urban sociology is whether the mixing of economic groups at the neighborhood level increases low-income families' access to social capital. For urban sociology, the question is whether mixed-income communities have positive effects on poor household's upward mobility. Linking mixed-income communities to economic mobility have been addressed most recently by sociologists involved in the Moving to Opportunity Experiment who have examined whether moving poor African-American families to middle-income white suburbs increases access to social capital by varying their social networks at the community level (Briggs et al., 2010).

Segregation

Segregation, like community, is a topic of long-standing interest within urban sociology. The Chicago School argued that social segregation was part of the underlying ecological structure that preserved community. Cities in early stages of urbanization were characterized by both racial and ethnic segregation – a period when ethnic differences between groups had great social significance. The city represented the amalgamation of difference. Segregation created village-like environments, thus permitting the social reconstruction of the village at the neighborhood level. Early urban sociologists viewed both racial and ethnic segregation in a positive light because they permitted socially similar people to make community together.

With the change in immigration policy, urban ethnic groups began the process of social homogenization that occurred with the reduced numbers of immigrants. Urban ethnic groups became white while the increased urbanization of African-American residents escalated. The segregation of newly white ethnic groups became less relevant to community construction and segregation by default was less prevalent. Yet the unabated racial segregation of African-Americans continued.

The civil rights movement and the urban riots of the 1960s propelled urban racial segregation into the public domain making the study of African-American segregation one of the most important research areas for urban sociology. Developing appropriate methods and measures for tracking racial segregation is a critical contribution of urban sociologists (Massey and Denton, 1993).

With the continuation of excessively high levels of racial segregation across most cities, urban sociologists also have investigated the underlying and not so mysterious causes of racial segregation. Urban sociologists have been at the forefront of researching housing market dynamics and segregative processes that create and maintain high levels of segregation (e.g., racial steering). In addition to the causes of racial segregation, urban sociologists also have been active in researching the consequences of segregation by examining neighborhood access and racial barriers to important income and wealth-generating resources across neighborhoods, such as housing finance (Squires, 1994; Shlay, 1988). This type of research has expanded to include examining the segregation of Hispanic and Asian households. Recent research suggests that segregation has decreased in some cities although the significance of this increase is debated (Logan and Stults, 2011; Glaeser and Vigdor, 2012).

With the Fair Housing Act, the Community Reinvestment Act, and other federal and local initiatives, urban sociologists play an important role in assessing the effectiveness of policy and the need for further intervention. Sociological research on discrimination in mortgage lending and homeowners insurance combined with effective political organizing and advocacy has helped propel the passage and increased the enforcement capabilities of legislation. Monitoring segregation and the effects of segregation continues to be crucial as communities work to stem the negative effects of race-based subprime and predatory lending and faulty underwriting.

Neighborhood Change

Urban sociology focuses on the dynamics of neighborhood change, specifically the precipitants of change, the process of change, the effects of change on households moving in and out,

and the effects of neighborhood change on the city and region as a whole. The Chicago School viewed neighborhood change as a process initiated by the 'invasion' of one type of different household (e.g., African-American) into a homogenous neighborhood (e.g., white). According to this perspective, this invasion served as a market signal that the neighborhood was transitional, leading eventually and seemingly inevitable from one type of neighborhood to another. In this way, urban ecology argued that invasion and succession were part of a normal set of mechanisms through which neighborhoods change.

Research now views the dynamics of neighborhood change much differently. Rather than viewing the movement of households as causing neighborhood change, urban sociology looks at the institutional mechanisms that precipitate particular types of changes. Neighborhoods are affected by the decision making within housing market institutions (e.g., real estate brokers, lenders, insurance agencies, etc.). What may appear to be a process governed by alleged principles of invasion and succession is actually a set of changes fueled by decisions made by housing market actors.

The focus on how housing market institutions affect neighborhoods is important largely because these decisions can be changed. The dynamics of racial change, abandonment, gentrification, and the like are not driven by housing preferences independent of the housing market context that is shaped by institutional actors such as lenders. Understanding the role of institutions in affecting neighborhood change is vital to promoting fair housing and increasing housing opportunities across the racial, ethnic, and class spectrum.

Concentrated Poverty

Concentrated poverty describes neighborhoods of largely extremely poor people devoid of income diversity. Neighborhoods characterized by concentrated poverty are believed to be socially and economically set apart from others because of their combined social and spatial isolation. Concentrated poverty is believed to have effects beyond the sheer numbers of extremely poor people. Rather, people living in neighborhoods with concentrated poverty maintain destructive and overtly negative lifestyles and behaviors that prolong their impoverishment. The combination of spatial and social variables in the form of concentrated poverty has fostered a spatial orientation to problems of poverty (Wilson, 1987).

Deconcentrating poverty has become part of a set of policy tools to alter the life circumstances of extremely poor families. As part of Hope VI, destroying large and dense public housing developments was also a form of deconcentrating poverty. Similarly Hope VI developments of mixed-income housing have sought to deconcentrate poverty. The Moving to Opportunity federal initiative that moved poor families out of public housing and into white suburban communities is yet another attempt at poverty deconcentration.

Whether deconcentrating poverty is effective in promoting more opportunities and economic advancement for poor families is not entirely clear. Much of what appears to be a form of poverty deconcentration may really be the residential displacement of poor families. Moreover, concentrated poverty continues to be a feature of most American cities, making it an important area for continued research on the spatial aspects of urban poverty.

Gentrification

Gentrification is a type of neighborhood change. A neighborhood undergoing gentrification is typically a community of low and moderate income with old but architecturally significant housing to which middle- and upper-income households are moving. The initiation of gentrification often occurs with the initial move in of artists looking for inexpensive homes and studios. A neighborhood that is fully gentrified contains predominately higher income families who live in housing whose values have soared and that have undergone extensive renovations.

In its initial phase in the early 1980s, gentrification was thought to be driven by a 'back to the city' movement. Later, scholars recognized that gentrification was driven by the reinvestment of capital in these depressed communities (Smith, 1979). The idea was that property values were so low that windfall profits were likely in the wake of gentrification. Gentrification represents property revalorization based on changing flows of capital. Yet these gentrified communities were not widespread and aptly were characterized as 'islands of renewal in seas of decay' (Berry, 1985).

In major cities, gentrification has become more widespread and is now considered a worldwide phenomenon. In many ways, the scope of what may be called gentrification has gone far beyond the speculative flipping of a few old properties. Gentrification has been suggested to be a 'global urban strategy' (Smith, 2002). This newer form of gentrification is believed to be fueled by state subsidies of transportation and other forms of infrastructure that creates large incentives for private sector investment. In this context, gentrification may no longer be a characteristic of how particular neighborhoods change but may represent large change in the city as a whole.

Urban Sociology and Policy

The genesis of urban sociology occurred during the formative years of American urbanization – a time of dramatic change in people's living conditions and a physical crush of social and economic difference amid new rules of behavior. The Chicago School was attracted by the mystery and magic of urbanization. They delved into this new spatial arrangement to gain an understanding of these dramatic changes and to do good. They saw a role for urban sociology to make cities better places. Of course, they embodied the racist, classist, and elitist ideologies of the period in which they lived and created. But within that ideological context, they sought to apply sociology to steer away from bad outcomes and to achieve what they believed was the moral order of the city.

Contemporary urban sociology is less self righteous about city life and less holistic in its approach to urban problems. While theoretical work on cities takes into account massive forces of capitalism and global inequality, the empirical work of urban sociology is more focused and compartmentalized. In the tradition of the Chicago School, many urban sociologists continue with the goal of doing good. But their good works are

focused on very specific problems. Urban sociology purports to do some good and to inform public policy. But the field has no vision of what constitutes a good city and its efforts within the policy field are piece meal.

Nonetheless, urban sociology may claim credit for working on a large number of issues in urban policy. These issues include fair housing, redlining, mortgage discrimination, affordable housing, low-income homeownership, exclusionary zoning, housing abandonment, homelessness, welfare dependency, unemployment, concentrated poverty, gentrification, downtown development, uneven development, neighborhood decay, gender discrimination, and tenants rights.

The Future of Urban Sociology

Although urban policy may not be the stuff of presidential elections and federal political maneuvering, the centrality of cities and place continues. Indeed, urban scholars suggest that matters associated with place will intensify with the heightened metropolitan inequality that is becoming more visible spatially. Urban sociology's future will continue to be concerned with the forces that shape cities and the quality of urban life. These topics will include the following:

- Gentrification and the intensifying capture of central city neighborhoods by elite groups
- Immigration and the flow of unparalleled diversity and difference moving to cities around the globe
- Globalization of housing finance and its effects on housing and neighborhoods
- The role of neighborhoods and the underpinnings of neighborhood effects
- Changing growth politics and power dynamics associated with place
- The role of racial segregation in maintaining intransigent urban poverty
- The impediments to racial and ethnic spatial assimilation
- The decentralization and suburbanization of poverty
- The growth of urban inequality and social movements
- Obstacles and possibilities for environmentally sustainable cities

In the midst of the ongoing disinvestment and reinvestment fueled by global capitalism, urban sociology will tackle thorny issues around the predictable cycles of uneven development. In the meantime, cities will remain places of magic, mystery, and misery.

See also: Capitalism: Global; Community Sociology; Discrimination: Racial; Inner City; Neighborhood Effects; Poverty, Sociology of; Racial Relations; Residential Segregation in the United States; Spatial Mismatch; Underclass; Urban Violence.

Bibliography

Alihan, M., 1938. Social Ecology: A Critical Analysis. Columbia University Press, New York.

Berry, B.J.L., 1985. Islands of renewal in seas of decay. In: Peterson, P.E. (Ed.), The New Urban Reality. Washington DC, Brookings, pp. 69–96.

Brenner, N., Theodore, N., 2002. Cities and the geographies of "actually existing neoliberalism". Antipode 34 (3), 349–379.

Briggs, X. d. S., Popkin, S., Goering, J., 2010. Moving to Opportunity: the Story of an American Experiment to Fight Ghetto Poverty. Oxford University Press, New York.

Burgess, E.W., 1967 [1925]. The growth of the city: an introduction to a research project. In: Park, R.E., Burgess, E.W., McKenzie, R.D. (Eds.), The City. University of Chicago Press, Chicago, pp. 47–62.

Castells, M., 1977. The Urban Question. MIT Press, Cambridge.

Durkheim, E., 1997 [1893]. The Division of Labor in Society. The Free Press, New York.

Glaeser, E., Vigdor, J., January 2012. The End of the Segregated Century: Racial Separation in America's Neighborhoods, 1890–2010. Manhattan Institute. Number 66 http://www.manhattan-institute.org/html/cr_66.htm (accessed 30.05.13).

Gotham, K., 2006. The secondary circuit of capital reconsidered: globalization and the U.S. real estate sector. American Journal of Sociology 112 (1), 231–275.

Harvey, D., 1973. Social Justice and the City. Johns Hopkins University Press, Baltimore.

Hunter, F., 1953. Community Power Structure: A Study of Decision Makers. University of North Carolina Press, Chapel Hill.

Lefebvre, H., 1991. The Production of Space. Blackwell Publishing, Malden.

Lefebvre, H., 1996 [1968]. Writings on Cities. Blackwell Publishing, Malden.

Logan, J.R., 2012. Making a place for space: spatial thinking in social science. Annual Review of Sociology 38, 507–524.

Logan, J.R., Stults, 2011. The Persistence of Segregation in the Metropolis: New Findings from the 2010 Census. Census Brief Prepared for Project 2010 (accessed 30.05.13) http://www.s4.brown.edu/us2010.

Logan, J.R., Molotch, H., 1987. Urban Fortunes: the Political Economy of Place. University of California Press, Berkeley.

Massey, D., Denton, N., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge.

McKenzie, R.D., 1967 [1925]. The ecological approach to the study of the human community. In: Park, R.E., Burgess, E.W., McKenzie, R.D. (Eds.), The City. University of Chicago Press, Chicago, pp. 63–79.

Molotch, H., 1976. The city as a growth machine: toward a political economy of place. American Journal of Sociology 82 (2), 309–332.

Park, R.E., 1936. Human ecology. American Journal of Sociology 42 (1), 1–15.

Park, R.E., 1967 [1925]. The city: suggestions for the investigation of human behavior in the urban environment. In: Park, R.E., Burgess, E.W., McKenzie, R.D. (Eds.), The City. University of Chicago Press, Chicago, pp. 1–46.

Sassen, S., 1991. The Global City. Princeton University Press, Princeton.

Shlay, A.B., 1988. Not in that neighborhood: the effects of population and housing on the distribution of mortgage finance in the Chicago SMSA. Social Science Research 17 (2), 137–163.

Shlay, A.B., 1989. Financing community: methods for assessing residential credit disparities, market barriers, and institutional reinvestment performance in the metropolis. Journal of Urban Affairs 11 (3), 201–223.

Simmel, G., 1950 [1903]. The metropolis and mental life. In: Wolff, K.H. (Ed.), The Sociology of Georg Simmel. Free Press, Glencoe, pp. 409–424.

Small, M., 2009. How many cases do I need? On logic and the science of case selection in field-based research. Ethnography 10 (1), 5–38.

Smith, N., 1979. Toward a theory of gentrification: a back to the city movement by capital, not people. The Journal of the American Planning Association 45 (4), 538–548.

Smith, N., 2002. New globalism, new urbanism: gentrification as global urban strategy. Antipode 34 (3), 427–450.

Squires, G.D., 1994. Capital and Communities in Black and White. SUNY Press, Albany.

Tönnies, F., 1963 [1887]. Community and Society. Harper and Row, New York.

United Nations, 2012. World Urbanization Prospects, the 2011 Revision. United Nations, New York.

Wellman, B., 1999. Networks in the Global Village: Life in Contemporary Communities. Basic Books, New York.

Wilson, W.J., 1987. The Truly Disadvantaged. University of Chicago Press, Chicago.

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Vintage Books, New York.

Zukin, S., 1987. Gentrification: culture and capital in the urban core. Annual Review of Sociology 13, 129–147.

Relevant Websites

The Urban Portal (http://urbanportal.org) is an online gateway to the latest research, events, and resources in urban social science.

Research Committee 21 (RC21) (http://rc21.org/) on the Sociology of Urban and Regional Development of the International Sociological Association promotes theory and research in the sociology of urban and regional development.

The Community and Urban Sociology Section (http://commurb.org) Of the American sociological Association includes those involved in community and urban sociology in their careers and vocations in teaching, scholarship, and work.

City and Community (http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291540-6040) is a journal of the Community and Urban Sociology Section of the American Sociological Association that advances urban sociological theory, promote the highest quality empirical research on communities and urban social life, and encourage sociological perspectives on urban policy.

Urban Sprawl

Robert Bruegmann, University of Illinois at Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The term *urban sprawl* has been used to describe low-density automobile-oriented settlement patterns with little comprehensive public planning. Opponents of this kind of settlement pattern claim that it is economically inefficient, socially inequitable, environmentally damaging, and aesthetically ugly. They have called for more planning, growth restrictions, growth boundaries, greenbelts, open space conservation, environmental laws, historic preservation, and other mechanisms to promote many of the compact, transit-oriented cities. Other observers suggest that the problems with sprawl have been overrated and the efforts to stop it are likely to be ineffectual or to backfire.

Urban sprawl is an imprecise, usually derogatory, term used most commonly in affluent countries of the English-speaking world to describe certain kinds of low-density automobile-oriented urban settlement patterns with little comprehensive public planning. Opponents of this kind of settlement pattern claim that it is economically inefficient, socially inequitable, environmentally damaging, and aesthetically ugly. To halt sprawl, reformers have called for more highly planned, compact, transit-oriented cities, growth restrictions, growth boundaries, greenbelts, open space conservation, and historic preservation. Although the term has been used by some planners and other experts for most of the last century, it became a major public concern in most countries only in the economic boom period of the last years of the twentieth century. At the same time, however, it has become clear that even some of the most basic assumptions about sprawl, its causes and effects, are in dispute.

History of the Term

The kinds of condition described by the term *sprawl* have been visible in urban areas for centuries. Already by the beginning of the modern era, for example, there were complaints about settlements straggling out from large European cities like London and Paris, and there were attempts to curtail this growth. These complaints seem to have coalesced in Britain following World War I when the term sprawl as a noun with something like its current meaning came into general usage (Bruegmann, 1996, 2000, 2005). One of the most conspicuous opponents of sprawl was Thomas Sharp, the planner, who issued vitriolic attacks on 'semi-suburbia,' 'ribbon developments' along roads leading out of town, and suburban villas marring the countryside (Sharp, 1932). His cause was taken up by groups such as the Council for the Preservation of Rural England with its constituency of architects, planners, large landowners, and members of Britain's artistic and intellectual elite. Several bills passed in the 1930s aimed at restricting ribbon development and creating greenbelts around cities (Hall, 1988).

The original criticisms of sprawl were primarily based on aesthetic and symbolic considerations. Critics derided it as ugly and as causing a blurring of the traditional distinctions between the city and the countryside. Complaints about sprawl have continued to accompany every era of major prosperity, but their content has changed considerably depending on the time and place. Immediately after World War II, for example, sprawl became less of an issue in Britain because it, like many continental European countries, had been exhausted economically by the war. Furthermore, in Britain and northern Europe, governments increasingly adopted stringent controls on peripheral development, so it was widely believed that it was no longer a problem. In the United States, on the other hand, the prosperity of the postwar years without any strong growth control planning led to a tremendous suburban boom. Consequently, the most conspicuous negative reactions to sprawl in the postwar years came from the United States. Although aesthetic objections still loomed large, the particular circumstances of American suburbia led to increased objections on social grounds.

The most conspicuous use of the term *sprawl* in the postwar decades came from William H. Whyte, who organized a conference for *Fortune* magazine on the subject and then published a jeremiad against sprawl in *The Exploding Metropolis*, a book that he edited (Whyte, 1958). For Whyte, as for Jane Jacobs and many other American critics of suburbia, Los Angeles became the prime example of all of the evils of sprawl.

Anti-sprawl sentiments subsided somewhat during the economic downturn of the 1970s when development slowed significantly. During the last decades of the twentieth century, however, development boomed once across much of the developed world and even more so in some of the developing world. Concerns about sprawl became widespread, particularly because of important new arguments against it on environmental grounds.

As the various strands of the sprawl argument emerged in recent years, the term has become more common, but it never acquired a clear definition. By the early twenty-first century, it could mean many different things. Almost everyone agreed that it referred to low-density development, but there has been a major debate in academia about how to measure sprawl and to understand its dynamics (Wolman et al., 2005).

For example, sprawl commonly refers to settlement that is scattered, discontinuous, leapfrog, or ribbon, but as many observers have pointed out, with additional development the low-density discontinuous settlement of one era often becomes the compact, and high-density settlement of the next

(Lessinger, 1962). Sprawl also refers to automobile-dependent development or development with poor access from home to work or to public open spaces, but many other observers have pointed out that, as cities have become wealthier, they have almost inevitably seen more automobile usage and, despite the problems they create, automobiles have greatly increased mobility for a larger and larger part of the urban population.

For many years scholars outside the English-speaking world tended to avoid the term. Labels like *città diffusa* in Italian or *ciudad dispersa* in Spanish were coined to talk about entire urban regions in continental Europe. Variations of the term *peri-urbain* or *rurbain* were used to discuss parts of the urban realm that in the English-speaking world were called suburbs and exurbs. In Germany, Thomas Sieverts introduced the concept of *Zwischenstadt* to designate the interstices within and between cities. In Japan, the term *supurôru* was coined to designate what English commentators called sprawl, although the target was more haphazard growth and lack of adequate infrastructure than any specific desire to make development more compact (Hebbert, 1994; Sorenson and Ikata, 2009; Sieverts, 1997; Monclus, 1998; Nivola, 1999; European Union, 1990; Brake et al, 2001; Whitehand and Carr, 2001; Richardson and Bae, 2004; European Environmental Agency, 2006; Phelps, 2006; Crouch et al, 2007; Hall and Pain, 2009).

Despite the confusion over definitions, by the turn of the twenty-first century the term *sprawl* had gained acceptance not just in the affluent West but also in many other places around the world. Important coalitions to fight scattered, low-density, automobile-oriented growth at the periphery of urban areas have sprung up in virtually every place where growth has been significant (Sudjic, 1992; Bogart, 2006). In the United States, the term *smart growth* became important in the 1990s to designate efforts to combat sprawl. In England, Canada, and Australia, the *compact city* tended to be the term used to refer to the opposite of sprawl. Within the last two decades throughout much of the world, sprawl has been seen as contrasting with sustainable development, another term with a large literature and extremely divergent definitions (Jenks et al., 1996).

Objections to Sprawl

Objections to sprawl raised by reformers have shifted significantly over the years according to the times and conditions in various parts of the world. They can, however, be classified under four major headings that range from the apparently quite objective to the obviously very subjective. These categories might be designated economic efficiency, social equity, environmental sustainability, and aesthetic value.

Efficiency Concerns

Although the earliest and most constant objections to sprawl have been aesthetic, the debate about public policy to curb sprawl has primarily been conducted in more objective terms. One of the most extensive segments of this literature has targeted the alleged inefficiencies of sprawl. The most important early summary of these issues in the United States in the postwar decades appeared in the Costs of Sprawl report issued by the Real Estate Research Corporation (1974). Using hypothetical models of various kinds of development, this report concluded that master-planned communities at fairly high densities were more efficient in terms of private development costs and public infrastructure costs than what they called unplanned communities at low densities. This report was attacked immediately on several fronts. Before the report had appeared, several researchers had noted that what appeared to be scattered, low-density development was often only the first stage of urban growth and that the undeveloped land might eventually invite even higher densities than if the settlement had been built all at one time (Lessinger, 1962). Another attack on the conclusions in the report came from reviewers who stated that the apparent savings actually were primarily due to the failure of the report to hold conditions constant in the various models (Altshuler, 1977; Windsor, 1979). Other researchers suggested that master planning in practice, as opposed to the hypothetical models used by the Real Estate Research Corporation, not only did not always yield the savings that should have been expected but often turned out to be more expensive and risky because of high upfront costs (Peiser, 1984). The supposed efficiencies of high density were also disputed (Ladd, 1992). More recently, a new report on the costs of sprawl has tried to respond to these challenges (Burchell et al., 1998). According to the authors of this report, the dire predictions of previous researchers have not materialized but, they opined, the higher costs of low-density settlement, particularly road and other infrastructure costs, would soon become intolerable.

Social Concerns

The second charge against sprawl, that it exacerbates economic and social inequities, has been more common in the United States than elsewhere because for many years the suburbs were substantially wealthier and housed fewer minorities than the central cities. This was particularly true after the unrest in American cities in the 1960s when many affluent families moved from the central city to the far suburbs. According to the view of many anti-sprawl advocates, this move in the population was motivated in large part by racism, and the low-density suburbs, were an attempt by middle-class citizens to distance themselves from social problems in the city. Other observers, however, have noted that sprawl has occurred in much the same way in metropolitan areas with few minorities as in areas with large minority populations, suggesting that race, while important, may not have been a decisive factor. They also note that American suburban areas have always been more diverse than they were often portrayed in the media and that in recent years they have become increasingly so (Harris and Larkham, 1999).

Another strand of thought has involved the treatment of central cities and suburban areas by governmental bodies. Many opponents of sprawl believe that sprawl in the United States was not inevitable or even a freely made choice by most of the population. Instead, they posited, it was a condition that was induced artificially by massive federal subsides in favor of homeownership, notably homeowner insurance guarantees, mortgage interest deductions, and federal funding for the massive postwar highway-building program. This is disputed by those who point out that many of these policies could have

benefited central cities at least as much as suburbs if there had been demand for city living and that massive decentralization has occurred throughout the world in places where these policies were not in place. Whatever the causes, as affluent citizens moved out of the central city, this move unquestionably did tend to maroon many poor and minority families who were unable or unwilling to move. At the same time, it tended to deplete the tax base of the older municipalities as they lost sales taxes and property taxes to outlying municipalities.

Another charge against sprawl in the United States has been a belief that sprawl leads inevitably to impoverished social relationships and less robust communities. Antagonism to sprawl on social grounds has often been bound up with attacks on suburbs generally. A group of famous sociological studies in the postwar years, for example, the work of Daniel Riesman, portrayed American suburbs as alienating, soulless places. This view has, however, been rebutted by other sociologists, for example, Herbert Gans (Riesman, 1950; Whyte, 1958; Gans, 1967).

Another target of criticism aimed primarily at the suburbs has been the privatization of public spaces, for example, the rise of gated communities and the decline of Main Street shopping centers in favor of suburban shopping malls (Sorkin, 1992). For others, however, what is important in social relationships and community has not changed. Only the outward forms of it have changed. In this perspective, the public areas of the mall actually do function much like the streets of city centers, and gated communities are no different in kind than condominium buildings in central cities with their doormen and security systems.

Environmental Concerns

Environmental concerns started to play an increasing role in anti-sprawl polemics following the ecological activism of the late 1960s. Indictments of sprawl on environmental grounds became common in the 1990s following a wave of interest in the idea of sustainability. According to this line of thought, low-density settlement inevitably leads to excessive energy usage, particularly gasoline for automobiles (Newman and Kenworthy, 1989). This in turn leads to the depletion of resources and to increased pollution and greenhouse gas emissions. Low-density settlement also leads to the loss of farmland, most seriously prime farmland near urban areas, to the loss of forests, wetlands, and other open lands, and to the loss of animal habitat. Sustainable cities, according to opponents of sprawl, will be higher density with a greater use of transit and a better balance of housing and jobs and less encroachment on agricultural and open land (Jenks et al., 1996; Brotchie et al., 1995).

Sprawl opponents particularly decry cities in the United States, because they use more energy than their counterparts in Europe and much more energy than cities in poorer parts of the world. For many environmentalists, this constitutes a fundamental human challenge that can only be solved by stringent conservation measures and replanning cities to be more compact and energy-efficient. Other scholars have noted a history of dire warnings about looming energy shortages and ecological disaster, none of which has come to pass, and have suggested that there are much more efficient and inexpensive ways to cut down on energy use and greenhouse gas emission than remaking the urban environment. In any case, they claim, no amount of replanning and conservation will by itself be able to compensate for the enormous growth in energy use that will almost inevitably occur as the poorest nations of the world are able to gain for themselves middle-class lifestyles. Only major advances in technology are likely to solve the problem.

Perhaps the most contentious area involving both energy consumption and environment remains the use of the private automobile. Sprawl reformers are almost unanimous in the belief that private automobile usage must be curtailed and public transportation greatly strengthened to reduce both congestion and pollution. However, the idea that it will be possible to counter the rise in automobile ownership and per capita vehicle miles traveled, which is a feature of virtually every city in the world as it has become more affluent, or to counter the long-term drop in the proportion of citizens riding public transportation, which has also been true in virtually all of these places, is, according to some observers, both unlikely and undesirable (Gordon and Richardson, 1997; Staley, 1999). These observers also point out that automobiles have become so much more energy-efficient that they often use less energy and emit fewer greenhouse gases per vehicle mile traveled than buses.

Another emotional part of the debate about sprawl has concerned the loss of farmland. According to many anti-sprawl reformers, cropland, particularly prime cropland, is an irreplaceable asset that is being fast eroded by urban expansion. Those who take an opposing view counter that saving farmland is not that urgent because, due to new agricultural methods, farmland productivity has soared over the last several decades, even as cropland has been lost to urbanization. In much of the affluent world, they point out, government subsidies have led to a great deal of excess production and surpluses that are often then sold to some of the poorest countries in the world at very low prices, badly undercutting agricultural production in these places.

Aesthetic Concerns

The final complaint about sprawl, that it is ugly, has inspired the most constant and abundant literature throughout the years. The sprawling city has been said to be unplanned, chaotic, monotonous, neither urban nor rural, but a combination of the worst aspects of each. One of the chief problems with this kind of judgment, of course, is that it is inherently subjective. Various individuals and groups differ widely in their aesthetic sensibilities, and even within a single individual or group, responses can change dramatically over the years. It has also been conspicuous, from the interwar complaints of British landowners down to the beginning of the twenty-first century, that there are class-bound aesthetic interests at work in this kind of critique.

Proposed Remedies for Sprawl

The most obvious and dramatic remedy for sprawl would be to stop new development outside existing urban areas altogether. This could be done by outright governmental prohibitions or

by limiting the development of necessary infrastructure, notably roads, water delivery systems, and wastewater treatment facilities. Although this has worked in Russia and China and other places with highly authoritarian regimes and even in places like Singapore and Hong Kong, it has been very difficult in the democratic countries of the affluent West. Instead governments in these countries have tried to channel growth.

Channeling or Managing Growth

One of the earliest and most important attempts to manage growth was the greenbelt system instituted around British cities, notably around London. In this system, which was made possible by a nationalization of all development rights in Britain after World War II, a growth boundary was established around British cities marking the inner edge of a greenbelt in which development was forbidden. Any additional growth that could not be accommodated within the greenbelt was to be confined to existing towns and villages beyond the greenbelt, and to newly established 'new towns (Abercrombie, 1945).' This system, which borrowed many features from Ebenezer Howard and his garden city concepts, did stop growth at the edge of the greenbelt. It was not nearly as successful in containing growth beyond the belt. In fact, growth beyond the greenbelt eventually scattered across much of southeast England. Moreover, by restricting the land available for new development adjacent to the already built-up area, it very likely raised land prices considerably, making it more difficult for people of modest means to afford housing, and perhaps driving them further out than would have been the case without the restrictions (Hall et al., 1973).

However, flawed it might have been in practice, the greenbelt model has been very attractive to planners. A similar kind of greenbelt, together with a set of policies encouraging compaction and public transportation, has been put in place around cities from Toronto to Seoul, South Korea (Troy, 1996). One of the most carefully examined cases has been the one in the state of Oregon since the 1970s. The experience of Portland, which has been America's most conspicuous exemplar of anti-sprawl planning, has been watched closely and debated hotly (Abbott et al., 1994; Mildner et al., 1996).

In other cities, other schemes were devised to channel growth. Immediately after World War II, Copenhagen experimented with a finger plan in which corridors of growth would alternate with green wedges. This kind of system was widely emulated. In other places, planners advocated widely spaced new centers to channel growth. Around Paris, for example, new development was to be channeled into specific growth axes organized around transportation lines and large new towns like St. Quentin en Yvelines and Marne-la-Vallee. One virtue of these finger or corridor proposals was that they mirrored the way development had occurred historically along rail lines. In practice, however, it proved increasingly difficult to enforce these corridors as expanded car ownership freed citizens from the constraints of remaining close to transit stops.

In the United States, in the wake of the postwar boom period, the slow growth and no growth movements became prominent in the 1960s and 1970s as a way to slow change, particularly at the urban edge. In most cases, growth was slowed in a given area, but subsequent research has suggested that the growth often merely migrated to adjacent, less regulated communities as the constraint on supply put sharp upward pressure on prices in the regulated areas (Porter, 1997; Garvin, 1996). By the 1990s, there was an increased push to reign in sprawl by managing infrastructure development. Concurrency rules dictate that adequate roadway, water, and sewer capacity be in place before development is permitted.

Open Space Conservation

Another measure sometimes counted as an anti-sprawl technique involves the creation of permanent open space. This can be done by legislative fiat, by the purchase of land, by the purchase of development rights, or by some combination of these mechanisms. Of course, in itself, these measures don't stop sprawl. In fact, absent other controls, it is possible that they simply push development out further than it would otherwise go. One could draw this conclusion from the London greenbelt or from San Francisco where a great many local initiatives to save open space and stop development have pushed settlement out into California's Central Valley.

A dramatic example of the use of open space to guide development can be seen in Hong Kong where British planning methods were adopted with even more rigor than they could be in Britain. Outside of central Hong Kong, authorities set aside a large percentage of the territory as country parks. All new development had to be contained in new towns built at very high densities above railroad stations. The scheme has been carried out largely as planned. The new towns have densities as high as central Hong Kong, making this the only affluent city in the world with densities equal to those of large nineteenth-century cities. It has meant that Hong Kong has been able to provide extremely efficient public transportation. On the other hand, it appears to have driven up land and house prices very dramatically.

Governmental Structures and Incentives

Other techniques for curbing low-density settlement and promoting high-density growth involve governmental policies and structure. In the United States, many sprawl reformers believe that the American tradition of deciding land-use policies at the local level, combined in many cities with a fragmented system of municipal government, has led to a lack of coherence in planning and created major incentives for municipalities to encourage sprawl (Gottdiener, 1977). The remedy proposed by many reformers was to institute metropolitan-wide governments with strong planning responsibilities, or to push responsibility for planning up to the state or federal level, so that local competition for tax dollars would be less important in planning decisions. Moreover, tax revenues would be shared more evenly, making it difficult to move to avoid taxes and allowing the possibility of a more equitable distribution of income for schools and other public services (Orfield, 1997). Among the most important metropolitan governments in North America are those of Portland, Minneapolis-St. Paul, and Toronto. States that have initiated strong growth management laws include Hawaii, Oregon, Florida, and New Jersey. Whether or not these policies will have any

substantial effect in stopping the outward decentralization of people and jobs in the long run is still unknown, and in recent years a number of these policies have been overturned or scaled back.

Another area targeted by reformers involves tax policies and other governmental fiscal measures. According to these reformers, sprawl in the United States was subsidized by federal mortgage insurance policies and mortgage interest deduction on the income tax. Low-density settlement was also subsidized by heavy federal subsidies for roads and the lack of subsidies at the same scale for public transportation. Many American reformers would like to see the end of the federal mortgage interest deductions, the imposition of heavy taxes on gasoline and vehicles to offset the damage they do to the environment, and much greater funding for transit (Downs, 1994). Other observers question whether any program to increase fees on private transportation would, given the apparently inexorable trend toward low density, actually lead to enough transit use to offset the undoubted increases in highway congestion and pollution (Gordon and Richardson, 1997). The prospect of alternative energy for private vehicles and the possibility of new forms of hybrid transportation systems, such as private rapid transit, very small, and driverless vehicles, might well alter all perspectives.

Design Solutions

A final set of recommendations to deal with sprawl has been made at the level of urban design. Throughout the affluent world there have been advocates for undoing what they consider to be the damage wrought by Modernist architecture and town planning and a corresponding segregating out of land uses into single-purpose districts. An important example in the United States has been the movement loosely called the New Urbanism which involves an attempt to create higher density traditional settlements with a mix of uses to promote community and transit use (Katz, 1994; Duany et al., 2000). A related movement is the push to create transit-oriented design in which high-density mixed-use developments are planned near transit stations. The new developments around some Washington DC Metro stations are considered a good example. It is again too early to judge the success of these efforts. On the one hand, New Urbanists have created a large number of new communities, and their form-based zoning which regulates building envelopes as an alternative to typical American zoning by land use has been adopted by a large number of local jurisdictions. However, it is not clear whether the techniques of Smart Growth have started to change mainstream patterns of urban development in any fundamental way.

The Sprawl Debate

Beyond the question of problems supposedly created by sprawl, or the efficacy of any particular remedy, is a larger set of questions about some of the most basic assumptions about sprawl itself, whether it is recent or has been evident for a long time, whether it is accelerating or is actually declining. There has actually been very little research on many of these questions. However, there does seem to be some evidence that in the United States, where the debate at the end of the twentieth century was hottest, decentralization and sprawl might actually have peaked in the period between the 1920s and the end of the postwar decades, meaning that the ratio of land consumption to population growth was greater during these periods than anytime since. Certainly, residential lot sizes at the suburban edge have been falling, and an increasing number of American urbanized areas are growing in density. Most strikingly, the Los Angeles region, still in much of popular opinion the prime exemplar of sprawl, has gotten so much denser over the last half century that it is currently the densest urbanized region in the United States and denser than a number of European urban areas (Demographia.com; Bruegmann, 2005).

There is also a fundamental question about whether the denser cities of Europe are different from American cities or whether they exhibit all of the same characteristics but are merely evolving at a slower pace because affluence came later, their population growth has been slower and there remains a relatively larger mass of historic fabric at the core (Bruegmann, 2005). A good deal of recent research seems to indicate that the American example is not an aberration, but that when cities become more affluent they virtually always tend to decrease in density (Echenique and Saint, 2001). In fact, virtually every urban area in the world seems to be following the lead of the United States in the way overall densities are dropping and development is pouring out into peripheral areas with highways, shopping centers, and residential subdivisions (Angel et al., 2011 and 2012).

See also: Sustainable Development: An Economic Perspective; Urban Geography; Urban History; Urban Poverty in Neighborhoods; Urban Sociology; Urban System in Geography.

Bibliography

Abbott, C., Howe, D., Adler, S. (Eds.), 1994. Planning the Oregon Way: A Twenty Year Evaluation. Oregon State University Press, Corvallis, OR.

Abercrombie, P., 1945. Greater London Plan 1944. His Majesty's Stationery Office, London.

Altshuler, A., 1977. Review of the costs of sprawl. Journal of the American Institute of Planners 43, 207–209.

Angel, S., Parent, J., Civco, D., Blei, A., 2011. Making Room for a Planet of Cities. Lincoln Land Institute, Cambridge, MA.

Angel, S., Parent, J., Civco, D., Blei, A., 2012. Atlas of Urban Expansion. Lincoln Land Institute, Cambridge, MA.

Bogart, W., 2006. Don't Call it Sprawl: Metropolitan Structure in the Twenty-First Century. Cambridge University Press, Cambridge.

Brake, K., Dangshchat, G., Herfert, G. (Eds.), 2001. Suburbanisierung in Deutschland, aktuelle Tendenzen. Leske + Budrich, Opladen.

Brotchie, J., Battey, M., Blakely, E., Hall, P., Newton, P. (Eds.), 1995. Cities in Competition: Productive and Sustainable Cities for the 21st Century. Longman, Melbourne, Vic, Australia.

Bruegmann, R., 1996. The American city: urban aberration or glimpse of the future? In: Cohen, M., Ruble, B., Tulchin, J., Gaylord, A. (Eds.), Preparing for the Urban Future: Global Forces and Local Pressures. Johns Hopkins University Press, Baltimore, MD.

Bruegmann, R., 2000. The paradoxes of urban sprawl reform. In: Freestone, R. (Ed.), Urban Planning in a Twentieth Century World. Routledge, London.

Bruegmann, R., 2005. Sprawl: A Compact History. University of Chicago Press, Chicago.

Burchell, R., Listokin, C., Shad, N., Phillips, H., Downs, A., Seskin, S., Davis, J., Moore, T., Helton, D., Gall, M., 1998. Costs of Sprawl Revisited, Transportation. Research Board, National Research Council, Washington, DC.

Crouch, C., Leontidou, L., Petschel-Held, G., 2007. Urban Sprawl in Europe, Landscapes, Land-Use Change and Policy. Blackwell, Oxford.
Demographia.com.
Downs, A., 1994. New Visions for Metropolitan America. Brookings Institution, Cambridge, MA. Washington, DC, and Lincoln Institute of Land Policy.
Duany, A., Plater-Zyberk, E., Speck, J., 2000. Suburban Nation: The Rise and Decline of the American Dream. North Point Press, New York.
Echenique, M., Saint, A., 2001. Cities for the New Millennium. Spon Press, London.
European Environmental Agency, 2006. Urban Sprawl in Europe: The Ignored Challenge.
European Union, 1990. Green Paper on the Urban Environment. European Commission, Brussels, Belgium.
Gans, H., 1967. The Levittowners: Ways of Life and Politics in a New Suburban Community. Allen Lane, London.
Garvin, A., 1996. The American City: What Works, What Doesn't. McGraw Hill, New York.
Gordon, H., Richardson, P., 1997. Are compact cities a desirable planning goal? Journal of the American Planning Association 63, 95–106.
Gottdiener, M., 1977. Planned Sprawl: Private and Public Interests in Suburbia. Sage Publications, Beverly Hills, CA.
Hall, P., Gracey, H., Drewett, R., Thomas, R., 1973. The Containment of Urban England, vols. 2. PEP, London.
Hall, P., 1988. Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century. Basil Blackwell, Oxford.
Hall, P., Pain, K., 2009. The Polycentric Metropolis: Learning from Mega-City Regions in Europe. Earthscan, London.
Harris, R., Larkham, P., 1999. Changing Suburbs: Foundation, Form and Function. E & FN Spon, London.
Hebbert, M., 1994. Urban sprawl and urban planning in Japan. In: Shapira, I., Masser, I., Edington, D. (Eds.), Planning for Cities and Regions in Japan. Liverpool University Press, Liverpool, UK.
Jenks, M., Burton, E., Williams, K., 1996. The Compact City: A Sustainable Urban Form? E & FN Spon, London.
Katz, P., 1994. The New Urbanism: Toward an Architecture of Community. McGraw-Hill, New York.
Ladd, H., 1992. Population growth, density and the costs of providing public services. Urban Studies 29 (2), 273–295.
Lessinger, J., 1962. The case for scatteration. Journal of the American Institute of Planners 28 (3), 159–169.
Monclus, F. (Ed.), 1998. La ciudad dispersa, Suburbanizacion y nuevas periferias. Centre de Cultura Contemporaria, Barcelona.
Mildner, G., Dueker, K., Rufolo, A., 1996. Impact of the Urban Growth Boundary on Metropolitan Housing Markets. Center for Urban Studies, Portland, OR.
Newman, P., Kenworthy, J., 1989. Cities and Automobile Dependence: An International Sourcebook. Gower Publishing, Aldershot, UK.
Nivola, P.S., 1999. Laws of the Landscape: How Policies Shape Cities in Europe and America. Brookings Institution, Washington, DC.
Orfield, M., 1997. Metropolitics: A Regional Agenda for Community and Stability. Brookings Institution Press and Lincoln Institute of Land Policy, Washington, DC. Cambridge, MA.
Peiser, R., 1984. Does if pay to plan suburban growth? Journal of the American Planning Association 50, 419–433.
Phelps, N., 2006. Post-Suburban Europe: Planning and Politics at the Margins of Europe's Capital Cities. Palgrave McMillan, Basingstoke.
Porter, D., 1997. Managing Growth in America's Communities. Island Press, Washington, D.C.
Real Estate Research Corporation, 1974. The Costs of Sprawl, vols. 2. United States Government Printing Office, Washington, DC.
Richardson, H., Bae, C.-H., 2004. Urban Sprawl in Western Europe and the United States. Ashgate, Aldershot, Hampshire.
Riesman, D., 1950. The Lonely Crowd: A Study of the Changing American Character. Yale University Press, New Haven, CT.
Sharp, T., 1932. Town and Countryside: Some Rural Aspects of Urban and Rural Development. Oxford University Press, Oxford, UK.
Sieverts, T., 1997. Zwischenstadt Ort und Welt, Raum und Zeit. Vieweg, Braunschweig.
Sorenson, A., Ikata, 2009. Megacities: Urban Form, Governance and Sustainability. Springer, Tokyo.
Sorkin, M. (Ed.), 1992. Variations on a Theme Park. The Noonday Press, New York.
Staley, S., 1999. The Sprawling of America: In Defense of the Dynamic City. Reason Public Policy Institute, Los Angeles.
Sudjic, D., 1992. The 100-Mile City. Andre Deutsch, London.
Troy, P., 1996. The Perils of Consolidation: A Discussion of Australian Housing and Urban Development Policies. The Federation Press, Annandale, NSW, Australia.
Whitehand, J., Carr, C., 2001. Twentieth-century suburbs. A Morphological approach. Routledge, London.
Whyte, W., 1958. Urban Sprawl. In: The Exploding City. Fortune Magazine pp. 115–139.
Windsor, D., 1979. A critique of the costs of sprawl. Journal of the American Planning Association 45 (3), 279–292.
Wolman, H., Galster, G., Hanson, R., Ratcliffe, M., Furdell, K., Sarzynski, A., 2005. The fundamental challenge in considering sprawl: which land should be considered. Professional Geographer 57 (1), 94–105.

Urban Studies: Overview

Ronan Paddison, School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. LeGates, volume 24, pp. 16092–16099, © 2001, Elsevier Ltd.

Abstract

This article reviews the scholarship developed in the subject of urban studies. The boundaries of urban studies are constantly in flux. This article examines the questions through which urban studies has become defined and its relationship with other core social sciences, and its place in university education. Urban research and scholarship is frequently interdisciplinary in its practice. Urban research is becoming increasingly globalized in the practitioners involved in it. The call for urban theory to be more responsive to the diversity of the city has challenged the dominance of earlier perspectives. Fundamentally, there is a lack of consensus around its ontological core, what defines the city, and how should it be conceptualized. The article also examines the implication of interdisciplinary and transdisciplinary for the status of urban studies in relation to the core social science subject.

Unpacking urban studies – the questions through which it has become defined, the story of its development, its relationship with the core social sciences, its place in university education – is revealing in helping to identify what contribution urban scholarship has been able to make to the wider project of intellectual endeavor and the problems that have surrounded its development as an autonomous field of study. As a term 'urban studies' invites immediate identification with specific aspects of human life; furthermore, as it has become commonplace to remark, more than half of the world's population is urban. In other words, studying cities – the endeavor of urban studies – should be central to the social sciences just as it should occupy an important position pedagogically in universities. Paradoxically, however, as a term 'urban studies' has a somewhat restricted use geographically, as well as in university curricula, while its status as an autonomous field of study within the social sciences has not reflected the importance the city has to modernity.

Such an argument has been expressed by others who have sought to understand the nature and role of urban studies as an intellectual endeavor; Robert Beauregard (2010) in a recent overview considered that among the social sciences urban studies occupies a somewhat marginal position. Yet, adopting such a position should not be taken that 'urban studies' lacks a core, nor that it lacks theory. Rather that, by comparison with the core social sciences, sociology, anthropology, political science, economics, and human geography, for example, it appears to be a relatively poor relation in (for example) its representation in university social science faculties. Further, while, as will be argued, there is a recognizable core of questions that help define urban studies, by its nature an appreciation of 'the urban' needs to be able to come to terms with the (multiple and growing) number of intersections with human life. Concepts such as 'urban greening,' considering the contribution of cities to global warming, are rapidly assuming an important place in urban research, yet were absent from the field of urban studies little more than a generation ago. In other words, the boundaries of what constitutes urban studies are constantly in flux, enlarging its reach, developments that may further problematize its definition as a discrete area of enquiry.

As an overview this article begins with an attempt to identify the core of what constitutes 'urban studies,' the issues that have been key to its research agenda. The discussion then turns to the problems linked with the studying of cities. The issues underpinning urban studies and its research agenda have unfolded in ways that reflect the changing understanding of the nature of cities and the epistemologies through which they should be studied. The problems of undertaking urban analysis both reflect the complexities of cities and offer some plausible explanation as to why the study of cities, as significant as it is to understanding contemporary (and past) life, does not enjoy a similar position to the core social sciences.

Identifying a Core

Identifying the core of an academic field as broad as urban studies is likely to be both a fraught as well as a contested exercise. To the extent that intellectual enquiry is influenced by fashion, what is understood as constituting the core may change, just as the development of competing epistemologies will foreground different types of (urban) issue. Defining what is the core of urban studies too differs from what *should* be its primary concerns – what research questions should we be asking about cities as opposed to those that are broached.

One route into the question of identifying the core issues of urban studies is to approach it from how cities are studied. In their successful *City Reader* Richard T. LeGates and Frederic Stout (2003) identify 12 key issues that "have most intrigued scholars who study cities" (p. 13) and present key readings representative of each likely to appeal to a wide audience, undergraduates through to scholars working in the area of urban studies (*sic*) and allied fields. In other words, their selection of topics sought to appeal to the wide variety of disciplinary backgrounds with which urban studies is associated. These issues included the evolution of cities; urban culture; urban society; urban politics and governance; urban economics, urban public finance, and regional science; urban and metropolitan space and city systems; megacities and global city systems; urban planning, design, landscape, and

architecture; race, ethnic, and gender relations in cities; urban problems and policy and urban futurism.

What needs emphasizing is that LeGates and Stout identified the range of issues with which urban studies considered in its broadest sense can be associated. Urban texts devised by scholars working from a more disciplinary perspective (e.g., Hall (2006) or Stevenson (2013)) will be more focused and less comprehensive of the range of topics urban studies entails. The concerns identified by Gates and Stout are revealing for the span covered by urban studies, yet while each was drawn out illustratively through key examples and authors involved in the study of different aspects of cities, the presentation was less transparent on the specific questions that have preoccupied urban analysis and which continue to do so.

The key questions around which urban studies has evolved, and which continue to dominate the field, entail those that have been historically enduring, added to which are other newer entrants that have developed particularly as a response to the recognition of the problems cities create or can be associated with. Among the enduring questions are the following:

- Why do cities grow? Why do some cities grow more than others? What are the implications of the uneven growth of cities? How do we understand why and how cities grow against the context of the multiscalar world in which these processes unfold?
- How do we define the city? How do we define and understand cities within the deepening processes of globalization?
- How are cities – their development and social processes within them – to be theorized? Do megatheories of urban structure and change have continued importance for urban analysis or are they too universalist in their claims and therefore better substituted by middle-range theories? How can and should urban theory respond to the urban growth in the global south?
- What methods should we use to study cities? What types of urban process are able to be meaningfully explored through macroscale perspectives as opposed to those more suited to microscale forms of analysis? What are the implications of using different research methods to explore urban processes?
- What defines the nature of urban life? How is social life in the city organized and conducted?
- How are cities governed? How should they be governed to meet the ideals of democracy? How are they governed to meet the aims of achieving economic growth and the delivery of public services in efficient and equitable ways?
- How are the issues of social diversity responded to in everyday life and through urban policy? What problems arise from different patterns of behavior between those living in the city (e.g., antisocial behavior) and how are these responded to through the governance of the city?
- What is the role of the market in the city in influencing the distribution of land uses and the provision of basic needs and what role should the state have in ensuring that these needs – for housing, transport, education, for example – are met?
- What have been and are the implications of new technologies in fashioning urban structure, urban economics, and social change? How can new technologies be used to address urban problems (e.g., Smart Urbanism) and what are its soci(et)al implications?
- How are cities planned and how should they be planned? What are the implications of the planning process for enhancing the economic growth of the city, social life within it, its governance and for local democracy? What is the role of architecture in the planning of cities?

These questions are representative of the more enduring issues that have preoccupied urban studies but do not make a claim to be comprehensive.

The unfolding development of urban studies has been marked by the augmentation of these questions in the light of the changing contexts within which cities are located. In other words, how and why urban studies has developed has often been as a response to the changing environment in which cities function and are represented by economic and political, including policy, shifts. As it has been frequently commented on, the study of cities as an intellectual endeavor has been responsive to major societal problems and shifts – indeed, urban studies has enjoyed major spurts of growth as a response to wider societal crises in the 1920s, the 1960s, and the 1970s and at the present day. Within recent decades the 'enduring questions' of urban analysis have been added to by new questions often in response to the identification of societal problems, the importance of which has propelled them to achieve core status. Two examples can be cited to show how the core agenda of urban studies is expanding and fluid:

- How can more sustainable forms of urban development contribute to reducing environmental problems including climate change? What are the policy and political implications linked to making cities more sustainable?
- How can and should cities be more socially just? How is the just city to be defined and how can social inequalities in the city, an enduring but reemergent feature of them, be addressed?

Again, these new questions do not exhaust how the core of urban studies has been augmented in recent decades, though the question of social justice implicitly refers to how urban neoliberalism is a dominant assemblage of techniques affecting the city in multiple ways, the play of power relations in the city, and the role of resistance.

What the range of core questions amply demonstrates is that urban studies is a multidisciplinary field aiming to understand (and steer) what is a highly complex 'structure.' By their nature, many of the questions may be suggestive of the primacy of particular disciplines. Indeed, how urban studies has evolved reflects this as the analysis of particular questions has developed with the development in turn of new subbranches and their attendant infrastructure such as specialist academic journals. Simultaneously, however, relatively few questions are the exclusive domain of particular disciplines; urban research and scholarship is frequently interdisciplinary in its practice.

Urban Studies in Flux

Few would deny that as a research field the contemporary study of the city is a vibrant and rapidly developing area, the nature of which has changed dramatically within recent decades. Its multidisciplinary nature together with the obvious ways in which its key concerns have such obvious relevance to societal (including global) problems contribute to its vibrancy and significance as an intellectual endeavor.

The vibrancy of urban research is reflected in its professional development. As an exemplar of its development the increase in academic and professional journals catering to urban researchers is one obvious indicator of the broadening reach of urban studies. At the middle of the twentieth century there were no academic journals specifically focused on publishing urban research. This was dependent on journals belonging essentially to the mainline disciplines that endorsed urban research as a branch of their wider endeavors. During the 1960s two journals were initiated, which were wholly focused on publishing urban research, the *Urban Affairs Quarterly* (later titled *Urban Affairs Review*) and *Urban Studies*. Both were responsive to publishing urban research that was interdisciplinary in its focus with papers covering a wide range of different research. Other journals were to follow, in some cases as in the *International Journal of Urban and Regional Research* focusing on theoretically driven, critically oriented research papers, while being responsive to the breadth of the urban research agenda. More numerous were specialist journals identified around the emergence of particular subbranches including *Urban Geography*, the *Journal of Urban Economics*, and *Urban Anthropology*. Whereas, then, at the mid-twentieth century there were no academic journals in the Anglophone world exclusively committed to publishing urban research, some two generations later the number is nearer to 20. Added to this are a growing number of professional journals published in the planning and other fields related to urban development.

Vibrancy too is reflected in the growing geographical spread of researchers drawing in those increasingly from the global south. For much of the twentieth century urban studies was dominated by researchers from North America and particular European and other northern countries, a domination that is hardly challenged by recent shifts but which, increasingly, is being complemented by urban researchers working from the global south. Thus, within the first decade of the twentieth century paper submissions to one of the leading generalist urban journals from China increased from accounting for less than 1% of all submissions to more than 8%. (For the same period the submissions originating from the USA hovered around 20% and, for the UK, a little lower.) The submissions from newly emergent nations closely follows the BRIC and TIGER groupings, but, if at the margins, is beginning to draw in researchers from other less developed nations. In other words, urban research is becoming increasingly globalized in the practitioners involved in it.

Vibrancy too is reflected in the multiplicity of perspectives, theoretical and methodological, through which urban research is conducted. While in the earlier history of the subject the study of cities had tended to be dominated by universalizing theories and arguments, contemporary research is more acutely aware of the intellectual dangers of meta-theory and the influence paradigmatic cities and perspectives can have in framing research in the very different circumstances represented by the diversity of cities. Whereas, then, for much of the twentieth century understanding the social life of the city was influenced by the tenets of the Chicago School, before the end of the century other perspectives had become dominant drawing out in many cases very different forms of analysis. Remnants of the Chicago School remain and, in some cases, the popularity of ethnographic methods of analysis, for example, is being revisited. Further, the influence of the paradigmatic city remains a seductive force in understanding cities. Nevertheless, the call for urban theory to be more responsive to the diversity of the city, global north and south, has challenged the dominance of earlier perspectives (Robinson, 2006; Roy, 2011).

The multiplicity of perspectives through which urban research is conducted makes for it being a broad church but one that challenges the coherence of the field. Fundamentally, there is a lack of consensus around its ontological core, what defines the city, and how should it be conceptualized. It is an ongoing problem that has confronted urban studies at different stages in its intellectual evolution – what is a city? How is it to be defined? Earlier attempts to define the city tended to focus around what it is not, the rural. Yet, such definitions are increasingly challenged by the blurring of the urban and rural and by the increasing awareness of the problems of defining the city in fixed terms in a world defined by rapidly emergent fluidity.

The reality is that as a broad church the research of urban studies embraces perspectives that are mutually inconsistent in ontological and epistemological terms. Positivist forms of analysis remain, particularly in econom(etr)ic analysis, in which the city is defined in fixed terms and in which rigorous, frequently quantitative, scrutiny seeks to claim the identification of 'true' relationships. Alongside such studies are those that embrace a postmodern rejection of the objectivity and totalizing expectations of positivism in favor of interpretations that are more receptive to diversity, ambiguity, and contingency. Whether such epistemological diversity is a source of weakness is a moot point (at least for some), but one that is likely to be challenged by researchers who have sought to embrace the wider 'turns' in social science research, cultural, linguistic, and spatial. In other words, the strength of urban studies as a field is to be sought more through the diversity of epistemologies it adopts and its willingness to embrace new perspectives able to explore the richness and complexity of the city.

Explaining the Status of Urban Studies

Paradoxically, given the apparent vibrancy of urban research and the broadening array of issues around which urban studies has become defined, its position within the social sciences, as suggested at the outset, is marginal. Many of those involved in urban research would subscribe to its contemporary practice being exciting, particularly for its epistemological diversity, even if this eclecticism challenges the coherence of understanding the city and their ability to talk to one another across what are sometimes disciplinary fences. Yet as much as urban

studies is vibrant, its status among the social sciences and particularly to its core disciplines is more one of a relatively 'poor' and dependent relation rather than an equal. To those social science disciplines that have been key in unraveling the nature and working of cities – economics, sociology, political science, geography, and anthropology in particular – the study of cities is a branch or subbranch; at the same time the field defining itself as urban studies has been unable to break into the ranks of what are considered the core social sciences. Explaining the paradox entails reexamining the nature of urban studies as well as it does the wider political environments in which urban research, particularly in universities, is located.

Problems in defining what is the city, which successive urban analysts have sought to explore, is matched by the inherently problematic nature of the urban itself. As a space or an area it tends to cut across the disciplines around which social science has become defined – that is, because of the nature of the urban it is inescapably of interest to political scientists as it is to geographers and sociologists. To the political scientist, then, acknowledging the role of cities, as Mossberger et al. (2012) argue in their keynote volume, follows from its ability "to conceptualise and understand local/urban political processes, public problems, and forms of decision-making to a particular space or location, rather than just analyze the subareas of voting, policy making and social movements…Urban politics is both a field in its own right, with particular dynamics and constraints that vary from place to place; and an area of research that tells political scientists and other scholars much about politics more generally, whether at the local, national or supranational level" (p. 4). Understanding urban politics is important to understanding political processes more generally while, interestingly, the argument also captures the diversity and uniqueness of local urban politics. Yet, to the wider project of political science understanding the role of urban politics is contributory to, rather than a core element of, the discipline. A similar story is repeated in the relationships the urban has with the other core social sciences. There are variations in how the story has evolved, not least because of the evolution of the separate social sciences themselves and the significance of the urban to the central tenets of the disciplines themselves. These variations, however, do not mask the reality that urban analysis in the core social disciplines has at best been a part of them, often qualifying the central tenets of the core disciplines rather than being necessarily central to their identity.

Another way of viewing this positioning of urban studies within the social sciences is to highlight a claim frequently (and rightly) made of it, its multidisciplinary nature. Two points follow from this – that, as it has been argued, the urban overlaps with the interests of multiple social science disciplines, but that, critically, it raises the specter of urban analysis needing to be interdisciplinary in how researching cities is conducted. The first raises the possibility that urban scholarship becomes 'dependent' on the efforts of those working from the perspective of the key disciplines to which it is allied, the second the problem of interdisciplinarity that has been an enduring issue not just to urban studies but to other fields of intellectual endeavor that transcend established frameworks of knowledge understanding, the 'disciplinarization' of knowledge production.

The dependency of urban analysis on the social science (and other) disciplines has had beneficial as well as more problematic impacts. Were we to identify the most influential analyses linked to the city in the last century, this would most likely include outputs from the Chicago School and texts such as Robert Dahl's (1961) exploration of urban power, *Who Governs?*, and David Harvey's (1973) book, *Social Justice and the City*; the list would also include others from outside the social sciences, notably Lewis Mumford (1937) and Jane Jacobs, particularly in *The Death and Life of Great American Cities*. Their influence in the development of urban studies has been indisputable in influencing in turn much subsequent research. Yet, as much as urban research draws attention to its interdisciplinary nature, its researchers have often failed to throw off their disciplinary legacy.

The problem identifies the power of the core disciplines in furthering (urban) knowledge and the challenges in being able to conduct research in ways that are more faithful to the ambitions of interdisciplinarity. Ramadier (2004) proposes a spectrum of disciplinary thinking in which its boundaries are progressively challenged through interdisciplinary and, ultimately, transdisciplinary analysis, arguing that it is really the latter that is able to address the complexity of urban issues. Yet, transdisciplinary analysis remains an aspiration, reliant as it is on the subordination of disciplinary to responsive, new ways of thinking about the city and its problems. Even its 'predecessor,' interdisciplinarity, poses challenges to the ways in which researchers approach urban research. Its realization may be more likely where urban research is conducted around closely defined issues and problems – possibly ones that have distinct policy implications – by teams of researchers representing different disciplinary backgrounds and working cooperatively.

The close connections between urban analysis, the understanding of the major problems of the city, and its ability to inform public policy have characterized the development of urban studies. Responding to the problems of cities – in the 1920s and the 1960s as well as the present day – urban studies has experienced significant shifts in how cities are understood. Further, precisely because the problems cities often tend to experience – poverty, deprivation, crime, racial conflict – can only be understood from a multidisciplinary perspective, so their analysis and their policy prescription has been interdisciplinary in its scope. Thus, the race riots in London in the early 1980s, together with the accompanying problems of urban social and economic deprivation, were to draw in researchers from a wide variety of social disciplines – sociologists, geographers, political scientists – and from beyond, including theologians (Archbishop of Canterbury: Commission on Urban Priority Areas, 1985).

While the development of urban studies has been closely connected with urban problems and policy, the relationship has not always been unproblematic. Urban policy research, and the presentation of policy solutions to urban problems, has run the gauntlet of criticism from other urban researchers for the shallowness of its research and its failure to understand the complexity of the city or, as in the case of the lively debates surrounding Richard Florida's notion of the creative city, the distributional consequences of policy tropes. Further, to the

extent that urban policy-oriented research falls short in the comprehensiveness and depth of its analysis, it can become 'relegated' in qualitative evaluations of research that have themselves become increasingly widespread in national university systems.

A Brief Postscript

Urban studies is defined around a paradox – its ambivalent relationship with the social sciences and its vibrancy. Entrenched systems of knowledge production reflected through the core social science disciplines, the complexity of the city and the interdisciplinary nature of urban problems, the footprint of the urban cutting across disciplinary boundaries, the challenge of conducting interdisciplinary research that is able to meet its demands – each has influenced how urban studies has developed but in ways that have effectively prevented the field from being able to create a sufficiently powerful identity in its right so that it is able to claim an equality to the core social sciences. Such a conclusion would not necessarily be subscribed to by all urban scholars; furthermore, to the extent that the argument carries weight the problem may be in the shortcomings of those involved in urban analysis in being able to fashion a sufficiently rigorous and intellectually autonomous field.

Simultaneously, urban studies is an increasingly rich field of intellectual endeavor. Its productive adoption of the multiple 'turns' – linguistic, cultural, spatial – has been key here. The centrality of cities has ensured that they have been rediscovered not only by the core social sciences but increasingly by scholars from the humanities. As during previous crises in the last hundred years, urban studies has been enriched, a richness that is defined around its epistemological diversity and its (necessary) tolerance to it.

See also: Urban Anthropology; Urban Geography; Urban History; Urban Sociology.

Bibliography

Archbishop of Canterbury's Commission on Urban Priority Areas, 1985. Faith in the City – a Call for Action by Church and Nation: Report of the Archbishop of Canterbury's Commission on Urban Priority Areas. Published for the General Synod of the Church of England by Church House Publishing, London.

Beauregard, R.A., 2010. Urban studies. In: Hutchison, R. (Ed.), Encyclopaedia of Urban Studies. Sage, London; Thousand Oaks, CA, pp. 931–936.

Dahl, R.A., 1961. Who Governs? Yale University Press, New Haven, CT.

Gates, R.T.L., Stout, F. (Eds.), 2003. The City Reader, third ed. Routledge, London; New York.

Hall, T., 2006. Urban Geography, third ed. Routledge, Milton Park, Abingdon; New York.

Harvey, D., 1973. Social Justice and the City. Edward Arnold, London.

Jacobs, J., 1961. The Death and Life of the Great American Cities. Vintage, New York.

Mossberger, K., Clarke, S.,E., John, P. (Eds.), 2012. The Oxford Handbook of Urban Politics. Oxford University Press, Oxford; New York.

Mumford, L., November 1937. What is a city? Architectural Record LXXXII, 92–96. Reprinted in Gates, R.T.L., Stout, F. (Eds.), 2003. The City Reader.

Mumford, L., 1989/1961. The City in History: Its Origins, Its Transformations and Its Prospects. Harcourt, New York.

Ramadier, T., 2004. Transdisciplinarity and its challenges: the case of urban studies. Futures 36, 423–439.

Robinson, J., 2006. Ordinary Cities: Between Modernity and Development. Routledge, Milton Park, Abingdon, UK.

Roy, A., 2011. The 21st century metropolis: new geographies of theory. In: Neuman, M., Hull, A. (Eds.), The Futures of the City Region. Routledge, Milton Park, Abingdon, UK; New York, pp. 31–42; originally published in Regional Studies 43(6).

Stevenson, D., 2013. The City. Polity Press, Cambridge, UK; Malden, MA.

Urban System in Geography

Juan M Kanai, University of Miami, Miami, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by P. Hall, volume 24, pp. 16099–16104, © 2001, Elsevier Ltd.

Abstract

In the early twenty-first century geographers continue to examine relations between cities, towns, and what may lie beyond their perceived borders. Yet efforts are now underway to redefine the urban system canon in order to better address questions pertaining to the global system of urbanization. The approach still holds explanatory purchase for both systems of cities and cities as systems. Research agendas have been updated, however. The emphasis for interurban relations has shifted from rank to network – or from national hierarchy to transnational connectivity. Research concerning core–periphery relations at the intraurban scale increasingly takes into account the rise of massively expansive yet functionally integrated polycentric city-regions in virtually every world realm. Nonetheless, poststructuralist perspectives such as mobile urbanism pose further challenges to system-based constructs.

From National Urban Systems to Global City Networks

In the decades of modernization and fast-paced economic growth following World War II, policy-makers concerned themselves with planning balanced and synergistic urban systems within the framework of tightly bounded national space-economies. Ideas influential during this period included the rank-size rule and central place theory. The following three propositions summarize the extensive literature on location theory: cities and towns with known population sizes are regularly spaced across a given territory (Lösch, 1954); predictable decreases in population size occur for nodes progressively ranking lower in the urban system – i.e., the population of the nth-ranked town being reduced to just $1/n$ of the country's largest city (Stewart, 1958); and urban nodes acquire heightened economic importance when they are able to serve extensive market areas with their specialized goods and services (Christaller, 1966/1933). Currently, the systems approach to urban centrality remains the most influential within geography's cognate of regional science, particularly among those focused on developing a general systems theory of urbanization (Hall, 2009).

In an earlier assessment of the urban systems approach, Hall (2001) listed its weaknesses. Despite the theory's purchase on territorial development initiatives of the mid-twentieth century, empirical corroboration was questioned early on. Not only did its proponents assume urbanization to be taking place on an abstract isotropic Euclidean plain; they also required territories to be self-contained with negligible interactions with external regions. Moreover, the field's economistic rationale of purported universal applicability notwithstanding, studies were not able to develop convincing explanations of why the expected regularities could be corroborated in certain regions but not others. Departures from the assumed norm were particularly conspicuous in what was then called the developing world. Urban primacy, a larger-than-expected concentration in capital cities or primate cities, came to be seen as a problematic feature, particularly in countries with burdensome colonial inheritances (Berry, 1961). Furthermore, following Lipton (1977) critics of the 'urban bias' argued that city-centric economic initiatives and aid schemes actually constituted a hindrance to development. This thesis, however, is now being critically revisited. Jones and Corbridge (2010) posit that the major advantages and increasing returns to urban agglomeration also accrue beyond the Global North, where they have been most thoroughly studied by the new economic geography.

In the current era of globally attuned urban geography, it is important to notice that Brazilian scholars have made an important contribution to urban systems research particularly through their studies of urbanization in the Amazon Basin (Becker, 1995; Corrêa, 1987). This body of research is noteworthy because it draws from location theory – by problematizing systemic interurban flows and relations of relative centrality, while also rising above the traditional limitations to this approach. Long considered an anomaly in terms of spatial development, Amazonia is both a vast natural space of planetary significance and a thoroughly urbanized territory with an extensive, albeit unevenly developed and thinly connected, network of cities and towns. While clearly lacking in second-rank nodes, which location theory valued as ideal for a region's integration to the national urban system, the network includes city cores with global ambition, such as Metropolitan Manaus, as well as hybrid settlements; these depart decidedly from rural configurations but are also devoid of even basic aspects of urban infrastructure and economic resources.

Becker (1995) points out that the unusual patterns found in the 'urbanized forest' cannot be properly understood without accounting for geopolitics, particularly the multiple interventions that the Brazilian state has undertaken to develop the region since the 1960s. Adopting a longer-term perspective, Corrêa (1987) emphasizes the need to develop a periodization of the urban network's evolution according to the shifting mode of regional integration to the world system, as e.g., a resource-extraction periphery and then a development frontier for industrial capitalism. Thus, rather than an anomaly to a stylized model of urban system configuration, Amazonian urbanization may be understood more productively as part of

and not apart from the unstable historical geographies of capitalism and the unevenness produced by its dominant modes of production through time. Thus, a geo-historical analysis of urbanization is required not only to chart out the spatial constitution of urban networks but also to examine the drivers of their dynamic evolution through time in a context of capitalist uneven development.

Urban system configurations have also been reexamined at the global scale. Sassen (1991) points out that since the later third of the twentieth century, the organization of production through extensive value chains of transnational reach has promoted the revalorization of select urban economies such as e.g., New York, London, and Tokyo but also Shanghai, Johannesburg, and São Paulo. These now function as command and control centers for the global economy with major clusters of specialized producer services. Concerned with the connectivity and economic specialization of such global cities, a vibrant new frontier of research on urban systems has consolidated around the notion of a world city network (Beaverstock et al., 2004). Earlier versions of this work may have represented a simplistic scaling-up of traditional central place theory notions from national systems to a global economy assumed to be smoothly integrated through transnational networks. These early efforts were also overly concerned with controversial rankings and a hierarchical categorization of world cities. Yet Taylor (2009) has called for a theoretical reorientation to provide a more fruitful conceptualization of intercity cooperation and mutual complementarities implicated in the production of economic globalization.

The Urbanization of the World and Rise of City-Regions

Beyond economistic reductionisms, urban geographers posit that globalization cannot be fully understood without theorizing the expansion of urbanism as the dominant form of social organization worldwide. This assertion rests on the recognition that the conditions for global economic integration depend largely on the unprecedented extent and scale of contemporary urbanization, a process that has also been discussed as generalized planetary urbanization (Brenner and Schmid, 2011). The complexity of resulting spatial patterns cannot be easily captured through traditional urbanization concepts let alone through the urban systems approach. Brenner (2013: 87) points out that geographies of urbanization are "assuming new, increasingly large-scale morphologies that perforate, crosscut, and ultimately explode the erstwhile urban/rural divide."

The city-region constitutes an important example of efforts to re-conceptualize functional specializations at the intraurban scale and core–periphery relations within agglomerations in light of this expansive urban reality. The literature explicitly seeks to move beyond notions limited to the scale of twentieth-century metropolitan urbanization (Scott, 2001). Parr (2005) explains that formal administrative boundaries and even the contours of contiguously urbanized metropolitan cores no longer reflect underlying structures of economic and social organization adequately. Relations between cities and surrounding territories have been re-functionalized. Increasingly widespread and complex, the patterns of interaction have also given rise to polycentric urban regions with multiple, often complementary, and specialized, city cores.

Economic geographers have shown particular interest in the city-region because the concept highlights the ongoing relevance of urban agglomeration and territorialized economies in the age of globalization and the purported end of geography (Rodríguez-Pose, 2008). Methodological questions remain on the definition of areal extents and the conceptualization of how city-regional formation fosters economic innovation and competitiveness (Turok, 2009). Yet, city-regions are increasingly seen as a key unit for growth and development in the contemporary world economy, and the city-regional coordination of infrastructures and economic activities is argued to play a crucial role in the establishment and maintenance of advanced global city centers to service extensive transnational markets (Scott, 2001). Empirically, several studies document the emergence, functional articulation, and increasing policy relevance of regional urbanization in the European Union (Hall and Pain, 2006); China and the Asia Pacific region (Douglass, 2000); and North America (Lang and Knox, 2009). The United Nations has begun to recognize the worldwide emergence of extensive urban configurations that, taking the form of mega-regions, urban corridors, functional city-regions, and massive conurbations, are announced to be the new engines of global and regional economies (UN-Habitat, 2010).

Critical perspectives on the emergence of city-regions also warrant attention. Roy (2009) points out that even though the city-region is becoming one of the dominant notions in the field of urban global studies, its theorization is largely predicated on specific experiences of urbanization in North America and Western Europe and therefore does not reflect other multiple forms of rapid urbanization occurring throughout the entire planet. Robinson (2006) calls for renewed efforts in cosmopolitan, globally comparative research in order to better capture such urban heterogeneity beyond imposed categorical dichotomies of North–South, rich–poor cities usually conceptualized as either entrepreneurially geared to ascend in global city hierarchy or locally focused on remedying developmental deficits. A case in point can be found in Johannesburg. **Figure 1** contains a schematic diagram representing Johannesburg's complex spatial transition from an industrial metropolis in the apartheid era of South African development to one of the cores in the thriving yet still highly unequal and fragmented Gauteng city-region with multiple and heterogeneous connections to the African continent and world economy.

Politics also matter in the making of city-regions. Jonas and Ward (2007) call for more explicit analyses of the role that power relations play in the varying construction of city-regions as territorial entities at the service of diverging political agendas and spatial goals. Moreover, Brenner (1999) demonstrates the centrality of city-regional agendas in wider state projects to reconfigure the economic capacities and rescale the regulatory functions of nation-states transitioning to a global economic order in the face of neoliberal hegemony determining market-centric approaches to urban and regional development. Such politics of scale, however, are far from linear. Institutional contestation and social mobilization mediate their outcomes (Jonas and Ward, 2007). In fact,

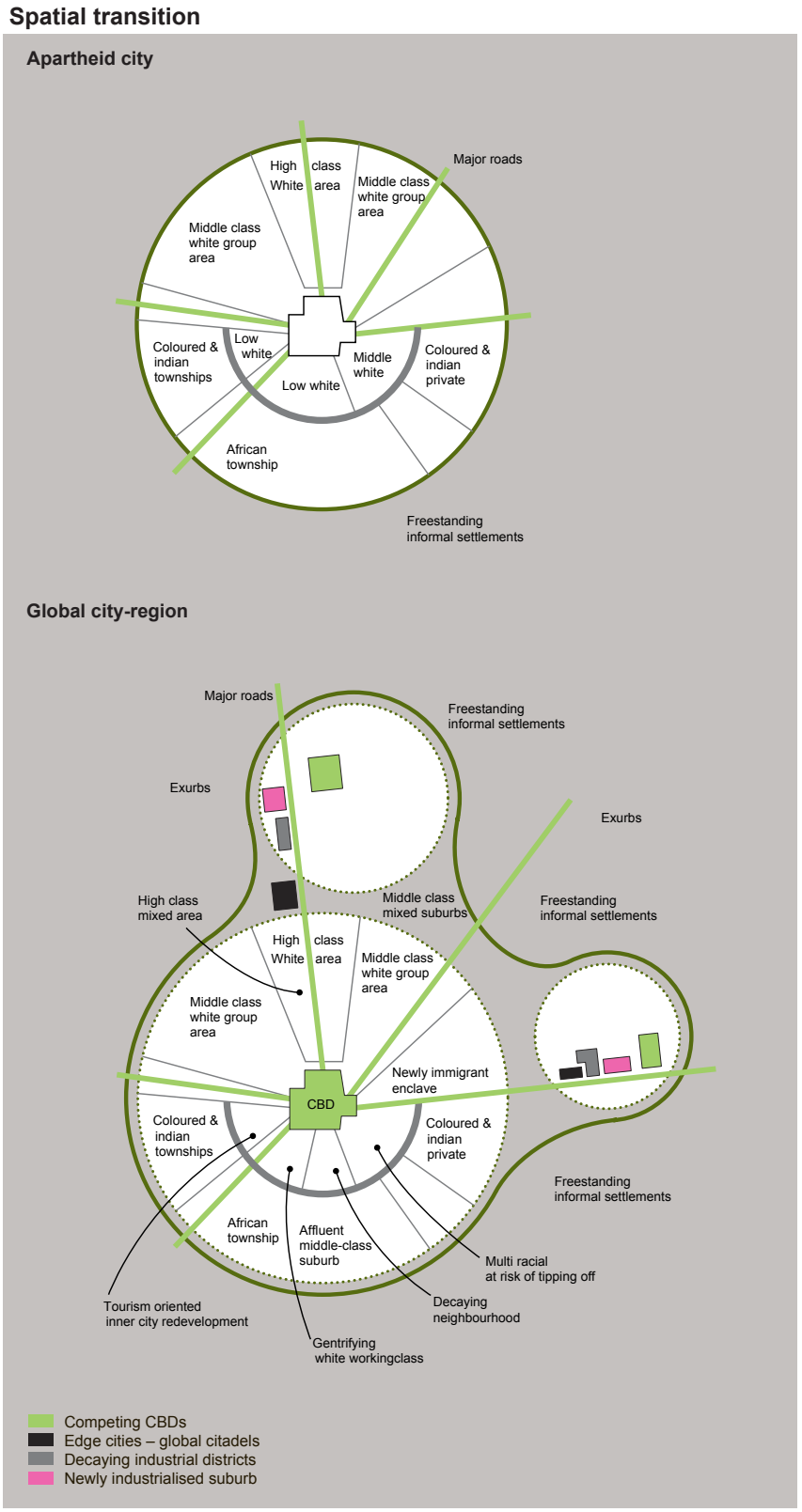

Figure 1 Schematic diagram of Johannesburg's spatial transition. Source: July 2006. Urban Age Conference Newspaper "Johannesburg Challenges of Inclusion?" (n.p.), London School of Economics and Political Science. http://v0.urban-age.net/0_downloads/JohannesburgNewspaper.pdf.

city-regional interdependencies provide opportunities for innovative forms of community organizing and coalition-building aiming for more egalitarian forms of urbanization (Benner and Pastor, 2011).

Poststructuralist Perspectives and Mobile Urbanism

Poststructuralist perspectives have also gained influence among geographers and pose additional challenges to the traditional urban system literature by taking on cognate questions of intercity relations and *trans*-local city-making. These efforts, however, adopt a diverging, less deterministic, approach that emphasizes contingent connections among multiple and heterogeneous intervening actions of both humans and the more-than-human world. Scholarship on questions of urban assemblage, and particularly its application to mobile urbanism, provides salient examples. Inspired by the work of French philosophers Deleuze and Guattari and science historian Latour, studies on urban assemblage call attention to multiple processes of becoming, whereby agents/objects of varied materialities come together, transforming each other in unstable configurations that help to explain the contemporary pace and complexity of contemporary urban-global change (Anderson and McFarlane, 2011). Thus, assemblage urbanism holds the promise of less reductive narratives of urbanization and city-making. Yet its place within the broader fields of critical urban studies and geopolitical economy remains in question.

Mobile urbanism has advanced an alternative view of the urban system as assemblage. In a pioneering study, Graham and Marvin (2001) focused on complex networked infrastructures in and among cities with global ambitions, such as e.g., a round-the-clock international airport that is seamlessly linked to a state-of-the-art business district through a multibillion high-speed transit megaproject that bypasses the disadvantaged neighborhoods that lie along the way. These (dis) connections were explicitly designed to selectively enable certain flows – while confining others – to serve goals of selective economic globalization while managing its local consequences within determined territories.

More recent contributions to mobile urbanism focus on the burgeoning global transfer of urban policies and schemes across highly differentiated territorial conditions, leading to 'best practices' in one locality to be soon emulated by many others (McCann and Ward, 2012). Thus, while more challenging to measure and model than economic networks and material flows, traveling policies constitute an additional layer to understand how cities relate to one another in the global era through *trans*-local multifarious connections. Whereas understanding the hegemony of neoliberal ideas provides a useful entry point to trace linkages, Robinson (2011) points out that, as is the case with urban policies related to water and sanitation, there is not only a spread of market-centric approaches but also an important hybridization of territorial schemes shaped by localized political struggles and conjunctures that should not be disregarded.

See also: Cities: Capital, Global, and World; Information Society, Geography of; Urban Geography; Urbanization in Africa; Urbanization in China; Urbanization.

Bibliography

Anderson, B., McFarlane, C., 2011. Thinking with assemblage. Area 43 (2), 162–164.
Beaverstock, J.V., Smith, R.G., Taylor, P.,J., 2004. World-City network: a new metageography? Annals of the Association of American Geographers 90 (1), 123–134.
Becker, B.K., 1995. Undoing myth: the Amazon, an urbanized forest. In: Clüsener-Godt, M., Sachs, I. (Eds.), Brazilian Perspectives on Sustainable Development of the Amazon Region, vol. 15. The Parthenon Publishing Group, UNESCO, Paris, pp. 53–89.
Benner, C., Pastor, M., 2011. Moving on up? Regions, megaregions, and the changing geography of social equity organizing. Urban Affairs Review 47 (3), 315–348.
Berry, B.J.L., 1961. City size distributions and economic development. Economic Development and Cultural Change 9 (4), 573–588.
Brenner, N., 1999. Globalisation as reterritorialisation: the re-scaling of urban governance in the European Union. Urban Studies 36 (3), 431–451.
Brenner, N., 2013. Theses on urbanization. Public Culture 25 (1), 69, 85–114.
Brenner, N., Schmid, C., 2011. Planetary urbanization. In: Gandy, M. (Ed.), Urban Constellations. Jovis, Berlin.
Corrêa, R.L., 1987. A periodização da rede urbana da Amazônia. Revista Brasileira de Geografia 49 (3), 47–52.
Christaller, W., 1966/1933. Central Places in Southern Germany (C.W. Baskin, Trans.). Prentice-Hall, Englewood Cliffs, NJ.
Douglass, M., 2000. Mega-urban regions and world city formation: globalisation, the economic crisis and urban policy issues in Pacific Asia. Urban Studies 37 (12), 2315–2335.
Graham, S., Marvin, S., 2001. Splintering Urbanism, Networked Infrastructures, Technological Mobilities and the Urban Condition. Routledge, London.
Hall, P., 2009. Looking backward, looking forward: the city region of the mid-21st century. Regional Studies 43 (6), 803–817.
Hall, P., 2001. Urban System in Geography. International Encyclopedia of the Social and Behavioral Sciences. Elsevier, 16099–16104.
Hall, P.G., Pain, K., 2006. The Polycentric Metropolis: Learning from Mega-city Regions in Europe. Earthscan, James & James.
Jonas, A.E.G., Ward, K., 2007. Introduction to a debate on city-regions: new geographies of governance, democracy and social reproduction. International Journal of Urban and Regional Research 31 (1), 169–178.
Jones, G.A., Corbridge, S., 2010. The continuing debate about urban bias the thesis, its critics, its influence and its implications for poverty-reduction strategies. Progress in Development Studies 10 (1), 1–18.
Lang, R., Knox, P.K., 2009. The new metropolis: rethinking megalopolis. Regional Studies 43 (6), 789–802.
Lipton, M., 1977. Why Poor People Stay Poor: A Study of Urban Bias in World Development. Temple Smith, London.
Lösch, A., 1954. The Economics of Location (W.H. Woglom, W.F. Stolper, Trans.). Yale University Press, New Haven, CT.
McCann, E., Ward, K., 2012. Assembling urbanism: following policies and studying through the sites and situations of policy making. Environment and Planning A 41, 42–51.
Parr, J., 2005. Perspectives on the city-region. Regional Studies 39 (5), 555–566.
Robinson, J., 2006. Ordinary Cities: Between Modernity and Development, vol. 4. Routledge, London.
Robinson, J., 2011. 2010 Urban geography plenary lecture – the travels of urban neoliberalism: taking stock of the internationalization of urban theory. Urban Geography 32 (8), 1087–1109.
Rodríguez-Pose, A., 2008. The rise of the "City-Region" concept and its development policy implications. European Planning Studies 16 (8), 1025–1046.
Roy, A., 2009. The 21st-century metropolis: new geographies of theory. Regional Studies 43 (6), 819–830.
Sassen, S., 1991. The Global City: New York, London, Tokyo. Princeton University Press, Princeton, NJ.
Scott, A.J. (Ed.), 2001. Global City-regions: Trends, Theory, Policy. Oxford University Press.
Stewart Jr., C.T., 1958. The size and spacing of cities. Geographical Review 48, 222–245.
Taylor, P.J., 2009. Urban economics in thrall to Christaller: a misguided search for city hierarchies in external urban relations. Environment and Planning a 41 (11), 2550–2555.
Turok, I., 2009. Limits to the mega-city region: conflicting local and regional needs. Regional Studies 43 (6), 845–862.
UN-Habitat, 2010. State of the World's Cities 2010/2011: Bridging the Urban Divide. Earthscan, James & James.

Urban Violence

Douglas S Massey, Office of Population Research, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urban violence may be sporadic or endemic. Sporadic violence refers to periodic eruptions of communal violence between social groups whereas endemic violence is built into the ecological organization of urban areas by structural conditions in society. Although communal violence often appears to be spontaneous, it is often orchestrated by social entrepreneurs for political purposes. The most common form of communal urban violence in the United States is the race riot. Race riots swept urban America during two periods – just after World War I and during the 1960s – and they erupt occasionally to the present day, usually in response to a triggering event, though research suggests the underlying cause is usually competition for scarce resources. Endemic urban violence varies across neighborhoods and is strongly associated with concentrated disadvantage caused by the segregation of a poor group by race and class. In the United States, high levels of violence are disproportionately concentrated in racially isolated poor black neighborhoods and exposure to such urban violence has been shown to have strong negative effects on health, cognition, and well-being.

Although many people assume that rates of violence are invariably higher in urban areas than in rural areas, this is not necessarily true. Indeed, Pinker (2011) has convincingly demonstrated that rates of violence historically were much higher among foraging and rural peoples and that rates of violent death have steadily dropped with the industrialization and urbanization of society during the nineteenth and twentieth centuries. Although in the United States today violent crime rates tend to be higher in urban than in rural areas (Wells and Weisheit, 2004), once accidents and other sources of violence are factored in, violent death rates are actually lower in rural counties (Myers et al., 2013). According to data from the United Nations, the level of urbanization across nations is only weakly correlated with homicide rates with an r-square value in the single digits (personal computation).

Urban violence occurs in two basic forms: sporadic and endemic. Sporadic violence refers to periodic eruptions of communal violence between social groups, usually defined on the basis of religion, race, or ethnicity. Examples include race riots in the United States, Hindu attacks on Muslims in India, and pogroms against specific groups such as Jews in Czarist Russia or Indians in East Africa (Brass, 1996). In contrast, endemic urban violence consists of ongoing violence, often criminal, that varies across neighborhoods but is built into the ecological organization of urban areas by the structural conditions in society, such as poverty and segregation (Sampson, 2012a).

Communal violence within cities has characterized cities throughout history and generally transpires within a larger system of racial, ethnic, or religious subordination and is directed against a despised or envied out-group. Such out-groups are typically framed as unlikable and incompetent or competent but cold (Fiske et al., 2002). Episodes of communal violence are generally triggered by perceived affronts or injuries to the majority population, real or imagined, and often occur during times of social disorder when normative structures have broken down or are temporarily suspended by dramatic events (Cuddy et al., 2008). Classic examples include violence directed against Jews during times of plague in Medieval European cities (Poliakov, 2003) and the violent confrontations between Hindus and Muslims at the end of the British Raj (Saxena, 2013).

Although communal violence between hostile groups may arise spontaneously in response to precipitating events (Ghassem-Fachandi, 2012), they may also be created deliberately by social entrepreneurs who specialize in the production of riots for political purposes, disguising them as spontaneous acts of self-defense against injustice (Brass, 2011). Horowitz (2001) studied 150 communal riots in 50 countries and concluded that despite their seeming spontaneity, riots are often planned and usually patterned with respect to timing, targeting, and location. According to Brass (2011), actors such as the police, criminals, businessmen, and politicians engage in continuous and often clandestine efforts to prepare for communal violence when an opportunity rises, creating 'institutionalized riot systems' that select locations, plan, rehearse, and refine narratives in advance.

In the United States, the most common form of communal violence over the years has been the urban race riot. During the Great Black Migration that began toward the end of the nineteenth century, African-American migrants from the rural south poured into northern and western cities where they were segregated into black ghettos (Massey and Denton, 1993). As ghetto neighborhoods filled up and began to encroach on white residential areas, the new arrivals were met with escalating violence that culminated in a wave of white-on-black race riots that swept urban America after the First World War (McWhirter, 2011), with two of the most famous examples being Chicago in 1919 (Tuttle, 1996) and Tulsa in 1921 (Madigan, 2001).

Race riots erupted in urban American again in the 1960s, but this time violence was not perpetrated by whites but by blacks rising in revolt against oppressive conditions in the nation's ghettos, venting their anger at the police and other symbols of white authority. The principal actions of the rioters were the burning and looting of ghetto businesses, causing Janowitz (1968) to label the violence 'commodity riots' as opposed to communal riots. Although the Kerner Commission

on Civil Disorders famously attributed the violence to segregation itself, announcing that "our nation is moving toward two societies, one black, one white—separate and unequal" (Kerner, 1988), sociological analyses at the time generally invoked a 'deprivation–frustration–aggression model' in which high levels of poverty and joblessness generated frustrations that exploded into violence in response to a 'triggering event,' such as an instance of police brutality (McPhail, 1971). However, Spillerman (1976) later pointed out that most of the variance in rioting was explained by just two variables: region and size of the black population. More recent event history analyses have concluded that the eruption of racial violence in the 1960s is best explained by intergroup competition for scarce resources such as jobs and political power (Myers, 1997).

Although nationwide urban violence has not swept American cities since the 1960s, race riots did not end in 1970 but have erupted sporadically over the years in response to various triggering events, as in Miami during 1980 (Portes and Stepick, 1993), Brooklyn in 1991 (Waddington, 2007), Los Angeles in 1992 (Baldassare, 1994), and Cincinnati in 2001 (Bronson, 2006). Despite the occasional outbreak of race riots from time to time, the most common form of urban violence in the United States today is endemic, expressed on a daily basis in certain neighborhoods in response to structural circumstances prevailing in the larger urban landscape. Across neighborhoods, rates of violence are strongly associated with material deprivation. Whether the urban area is Stockholm or Chicago, violence is predicted by concentrated neighborhood poverty (Sampson and Wikström, 2008). Although average levels of both concentrated disadvantage and concentrated violence are much lower in Stockholm than in Chicago, the relationship between the two variables follows essentially the same curve (Sampson, 2012b).

The reason why violence and disadvantage are less concentrated in Stockholm is that poverty rates are much lower and there is much less segregation by race and class. Massey and Denton (1993) demonstrated that a combination of high poverty within a group and a high level of residential segregation between that group and the rest of society axiomatically produce concentrated poverty, a fundamental process that is exacerbated by elevated levels of segregation on the basis of income (Massey and Fischer, 2000). Although this explanation was critiqued by Jargowsky (1997), Quillian (2012) ultimately demonstrated that, given a class structure dividing people into rich and poor, concentrated poverty results from a combination of racial segregation, class segregation between racial groups, and class segregation among members of the segregated group.

The spatial concentration of poverty also geographically concentrates any attributes that are correlated with poverty, such as crime, delinquency, joblessness, marital instability, welfare dependency, and substance abuse, thereby creating a very unstable and hostile environment that breeds social disorder and a lack of what Sampson (2006) calls collective efficacy – social cohesion between neighbors characterized by a willingness to intervene on behalf of the common good. Within any urban ecology, the various manifestations and sequelae of material deprivation are highly spatially correlated with one another (Sampson, 2012b). Thus, neighborhood violence is but one of many manifestations of spatially concentrated disadvantage that poor, racially segregated groups routinely experience in American cities, making it difficult for researchers to isolate the specific causes and effects of neighborhood violence from the many other maladies that are correlated with concentrated poverty (Sampson, 2011).

Nonetheless, careful research carried out by Sampson and colleagues indicate that high rates of violence are strongly and independently predicted by low levels of collective efficacy and high concentrations of disadvantage, controlling for a variety of ancillary factors (Sampson et al., 1997; Morenoff et al., 2001; Sampson and Morenoff, 2004; Sampson, 2012b). The likelihood that an individual engages in violence is similarly predicted by living in a segregated neighborhood characterized by high levels of crime and moral cynicism – a belief that legal authorities are ineffective and illegitimate (Sampson et al., 2005).

When poor black households move from segregated city neighborhood into white affluent suburbs, their exposure to neighborhood violence is dramatically reduced (Sharkey and Sampson, 2010). Massey (2004) has argued that long-term exposure to high levels of neighborhood violence produced by the combination of high poverty and high segregation have strong negative effects on health, cognition, and well-being. Children and adolescents are especially vulnerable to the negative effects of violence as exposure may alter the course of brain development to produce long-lasting physical, mental, and emotional problems such as attachment deficits, anxiety, depression, and aggression (Margolin and Gordis, 2004; Finkelhor et al., 2009; Duke et al., 2010). According to research done by Massey and Brodmann (2014), poor black adolescents growing up in the United States experience four times the degree of neighborhood disadvantage and twice the level of peer violence as affluent white adolescents. They go on to show that exposure concentrated neighborhood disadvantage increases the risk of obesity, raises the likelihood of failure at school, and increases the odds of committing a violent crime while simultaneously lowering scores on vocabulary tests and reducing the likelihood of eventual employment as adults. They also find that exposure to peer violence increases the risk of ADHD, school failure, violent crime, and criminal justice involvement and that contrasting neighborhood environments go a long way toward explaining socioeconomic differentials by race and class (Massey and Brodmann, 2014).

Exposure to concentrated disadvantage and its accompanying violence has been shown experimentally to have significant negative effects on mental health and happiness (Ludwig et al., 2012) and to increase the risk of diabetes and obesity (Ludwig et al., 2011). Using quasi-experimental methods, Sharkey has shown that exposure to neighborhood violence reduces children's performance on cognitive tests (Sharkey, 2010) and undermines children's attention and impulse control (Sharkey et al., 2012). Also using a quasi-experimental design, Massey et al. (2013) demonstrate that exposure to neighborhood violence also increases the frequency of negative life events among residents and promotes higher levels psychological distress. In addition, by increasing the frequency of negative life events, which often interfere with work responsibilities, exposure to neighborhood disorder and violence also reduces rates of employment and lowers earnings, while exposure to violence in schools reduces grade achievement among children.

Owing to persistent racism, lagging socioeconomic status, and restrictive zoning in American suburbs, levels of black segregation and spatial isolation remain high in most of the nation's large metropolitan areas (Rugh and Massey, 2013). As a result, in the United States, the foregoing negative effects occur very disproportionately to African-Americans (Massey and Denton, 1993). Indeed, the distribution of black and white neighborhoods by per capita income, crime rates, and concentrated disadvantage barely overlap, meaning that the most affluent African-Americans generally experience worse neighborhood conditions than the poorest whites (Peterson and Krivo, 2010; Sampson, 2012b; Massey and Brodmann, 2014). In the past, black poverty was characterized by the inheritance of racial disadvantage (Duncan, 1969), whereas it is now perpetuated by the inheritance of place disadvantage (Sharkey, 2013). The unique exposure of African-Americans to spatially concentrated disadvantage and violence because of poverty and segregation is the principal reason for America's stalled progress toward racial equality.

See also: Community, Social Context of: The US Case; Crime, Immigration, and Ethnicity; Discrimination: Racial; Ethnicity, Sociology of; Income Inequality; Minorities; Prejudice in Society: Psychological Perspectives; Racial Relations; Racism, Sociology of; Residential Segregation in the United States; Social Stratification; Underclass; Urban Sociology.

Bibliography

Baldassare, M., 1994. The Los Angeles Riots: Lessons for the Urban Future. Westview Press, Boulder CO.

Brass, P.R., 1996. Riots and Pogroms. New York University Press, New York.

Brass, P.R., 2011. The Production of Hindu-Muslim Violence in Contemporary India. University of Washington Press, Seattle.

Bronson, P., 2006. Behind the Lines: The Untold Story of the Cincinnati Riots. Chilidog Press, Milford, OH.

Cuddy, A.J.C., Fiske, S.T., Glick, P., 2008. Competence and warmth as universal trait dimensions of interpersonal and intergroup perception: the stereotype content model and the BIAS map. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 40. Academic Press, New York, pp. 61–149.

Duke, N.N., Pettingell, S.L., McMorris, B.J., Borowsky, I.W., 2010. Adolescent violence perpetration: associations with multiple types of adverse childhood experiences. Pediatrics 124, e778–e786.

Duncan, O.D., 1969. Inheritance of poverty or inheritance of race? In: Moynihan, D.P. (Ed.), On Understanding Poverty. Basic Books, New York, pp. 85–110.

Finkelhor, D., Turner, H.A., Ormrod, R., Hamby, S., Kracke, K., 2009. Children's Exposure to Violence: A Comprehensive National Survey. U.S. Department of Justice, Washington, DC.

Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 878–902.

Ghassem-Fachandi, P., 2012. Pogrom in Gujarat: Hindu Nationalism and Anti-Muslim Violence in India. Princeton University Press, Princeton, NJ.

Horowitz, D.L., 2001. The Deadly Ethnic Riot. University of California Press, Berkeley.

Janowitz, M., 1968. Social Control of Escalated Riots. University of Chicago Press, Chicago, IL.

Jargowsky, P.A., 1997. Poverty and Place: Ghettos, Barrios, and the American City. Russell Sage Foundation, New York.

Kerner, Otto, 1988. The Kerner Report: The 1968 Report of the National Advisory Commission on Civil Disorders. Pantheon, New York.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L.F., Kessler, R.C., Kling, J.R., Lindau, S.T., Whitaker, R.C., McDade, T.W., 2011. Neighborhoods, obesity and diabetes: a randomized social experiment. New England Journal of Medicine 365, 1509–1519.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2012. Neighborhood effects on the long-term well-being of low-income adults. Science 337, 505–1510.

Madigan, T., 2001. The Burning: Massacre, Destruction, and the Tulsa Race Riot of 1921. Thomas Dunne Books, New York.

Margolin, G., Gordis, E.B., 2004. Children's exposure to violence in the family and community. Current Directions in Psychological Science 13, 52–155.

Massey, D.S., 2004. Segregation and stratification: a biosocial perspective. The DuBois Review: Social Science Research on Race 1, 1–19.

Massey, D.S., Albright, L., Casciano, R., Derickson, E., Kinsey, D., 2013. Climbing Mount Laurel: The Struggle for Affordable Housing and Social Mobility in an American Suburb. Princeton University Press, Princeton, NJ.

Massey, D.S., Brodmann, S., 2014. Spheres of Influence: The Social Ecology of Racial and Class Inequality. Russell Sage Foundation, New York.

Massey, D.S., Denton, N.A., 1993. American Apartheid: Segregation and the Making of the Underclass. Harvard University Press, Cambridge, MA.

Massey, D.S., Fischer, M.J., 2000. How segregation concentrates poverty. Ethnic and Racial Studies 23, 670–691.

McPhail, C., 1971. Civil disorder participation: a critical examination of recent research. American Sociological Review 36, 1058–1073.

McWhirter, C., 2011. Red Summer: The Summer of 1919 and the Awakening of Black America. Henry Holt and Co, New York.

Morenoff, J.D., Sampson, R.J., Raudenbush, S., 2001. Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology 39, 517–560.

Myers, D.J., 1997. Racial rioting in the 1960s: an event history analysis of local conditions. American Sociological Review 62, 94–112.

Myers, S.R., Branas, C.C., French, B.C., Nance, M.L., Kallan, M.J., Wiebe, D.J., Carr, B.G., 2013. Safety in numbers: are major cities the safest places in the United States? Annals of Emergency Medicine 62, 408–418.

Peterson, R.D., Krivo, L.J., 2010. Divergent Social Worlds: Neighborhood Crime and the Racial-Spatial Divide. Russell Sage, New York.

Pinker, S., 2011. The Better Angels of Our Nature: Why Violence Has Declined. Viking Publications, New York.

Poliakov, L., 2003. The History of Anti-Semitism. In: From the Time of Christ to the Court Jews, vol. 1. University of Pennsylvania Press, Philadelphia.

Portes, A., Stepick, A., 1993. City on the Edge: The Transformation of Miami. University of California Press, Berkeley.

Quillian, L., 2012. Segregation and poverty concentration: the role of three segregations. American Sociological Review 77, 354–379.

Rugh, J.S., Massey, D.S., 2013. Segregation in post-civil rights America: stalled integration or end of the segregated century? DuBois Review 10 (2), 1–25. http://dx.doi.org/10.10170S1742058X13000180.

Sampson, R.J., 2006. Collective efficacy theory: lessons learned and directions for future inquiry. In: Cullen, F.T., Wright, J.P., Blevins, K. (Eds.), Taking Stock: The Status of Criminological Theory. Transaction Publishers, New Brunswick, NJ, pp. 149–167.

Sampson, R.J., 2011. Neighborhood effects, causal Mechanisms, and the social structure of the city. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, New York, pp. 227–250.

Sampson, R.J., 2012a. Neighborhood inequality, violence, and the social infrastructure of the American city. In: Tate, W.F. (Ed.), Research on Schools, Neighborhoods, and Communities: Toward Civic Responsibility. Rowman and Littlefield, Lanham, MD, pp. 11–28.

Sampson, R.J., 2012b. Great American City: Chicago and the Enduring Neighborhood Effect. University of Chicago Press, Chicago, IL.

Sampson, R.J., Morenoff, J.D., 2004. Spatial (dis)advantage and homicide in Chicago neighborhoods. In: Goodchild, M., Janelle, D. (Eds.), Spatially Integrated Social Science. Oxford University Press, New York, pp. 145–170.

Sampson, R.J., Morenoff, J.D., Raudenbush, S.W., 2005. Social anatomy of racial and ethnic disparities in violence. American Journal of Public Health 95, 224–232.

Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277, 918–924.

Sampson, R.J., Wikström, P.O., 2008. The social order of violence in Chicago and Stockholm neighborhoods: a comparative inquiry. In: Shapiro, I., Kalyvas, S., Masoud, T. (Eds.), Order, Conflict, and Violence. Cambridge University Press, New York, pp. 97–119.

Saxena, H., 2013. Communal Politics: Effects on Indian Independence and Its Aftermath: 1906–1947. Manak Publications, New Delhi.

Sharkey, P., 2010. The acute effect of local homicides on children's cognitive performance. Proceedings of the National Academy of Sciences 107, 11733–11738.

Sharkey, P., 2013. Stuck in Place: Urban Neighborhoods and the End of Progress toward Racial Equality. University of Chicago Press, Chicago, IL.

Sharkey, P., Sampson, R.J., 2010. Destination effects: residential mobility and trajectories of adolescent violence in a stratified metropolis. Criminology 48, 639–682.

Sharkey, P., Strayer, N., Papachristos, A., Raver, C., 2012. The effect of local violence on children's attention and impulse control. American Journal of Public Health 102, 2287–2293.

Spillerman, S., 1976. Structural characteristics of cities and the severity of racial disorders. American Sociological Review 41, 771–793.

Tuttle, W.M., 1996. Race Riot: Chicago in the Red Summer of 1919. University of Illinois Press, Urbana.

Waddington, D., 2007. Policing Public Disorder. Willan Publishers, New York.

Wells, L.E., Weisheit, R.A., 2004. Patterns of rural and urban crime: a county-level comparison. Criminal Justice Review 29, 1–22.

Urban Violence and Militarization

Jonathan Mendel, Geography, School of the Environment, University of Dundee, Dundee, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

There is a long history of cities playing key roles in different types of violence and military conflict. Urban violence remains a significant factor in cities today. Urban warfare is increasingly prominent in 'new' wars and there has been a blurring of boundaries between military action 'outside' of states and domestic governance 'inside.'

This article will outline how cities are affected and formed by, and play significant roles in shaping, violence and warfare. Urban warfare will be used as an example of the prominent role of cities in violence, and the article will consider the increasing militarization of 'domestic' urban life. Surveillance will be considered both as an attempt to engage with (fears of) urban violence and as an example of a blurring of military/domestic lines. The article will conclude by considering potential responses to urban violence.

Introduction

This article will begin by outlining a historical context. Urban violence is often presented as purely or predominantly a problem for cities and urban life. However, it will be argued that there is a long-running and intimate relationship between cities and violence, to the extent that they have often been mutually constitutive of one another.

This article will then consider broader questions of urban warfare. This article will look at how 'new' approaches to war have led to what Gregory (2008) describes as a cultural turn in counterinsurgency, incorporating diverse types of action from counterinsurgency as 'armed social work' through to 'urbicide' aimed at eradicating the 'disease' of urban lifestyles. Urban militarization will be shown to impact on policing and control in diverse ways: ranging from increasing use of SWAT-style tactics to military technologies and techniques (SWAT – Special Weapons and Tactics). This article will end by looking at other potential ways out of this militarization and other ways of building urban life and safety.

Urban Histories and Urban Violence: 'The City Is the Result of War'

While violence is often presented simply as a threat to urban life, this belies a far more complex history. Virilio argues powerfully that war and cities have been inextricably linked since the beginning of urban living: for Virilio, "the city is the result of war, at least the preparation for war" and he finds military reasons behind the original concentration of people in cities (Virilio and Lotringer, 2008: 19). Drawing on Virilio's work, Graham (2004a: 1) thus moves beyond a focus just on military action against cities to argue that "[c]ities, warfare and organised political violence have always been mutual constructions." Following such work, this article will argue that rather than cities simply being important targets for violence or places where violence takes place, cities and urban violence are more closely intertwined.

There is a long history of violence against cities. For Hewitt (1983: 259) attempts to destroy urban places go back at least to the ancient Roman assault on Carthage. However, Hewitt (1983: 259) also finds a 'more-than-physical reality' to urban places (which feeds into the sometimes extreme measures that can be taken in the hope of annihilating them). Hewitt (1983: 275) argues that particular aspects of this more-than-physical urban life – for example, high-density ways of living – are specifically targeted in area bombing. Hewitt (2009: 344–345) thus thinks about the bombing of London in World War II as part of a 'preoccupation with demonstration pieces or spectacle': finding links between this and the more recent targeting of cities in 'shock and awe.' Similarly, the 'terrorist' targeting cities speaks both to a desire for publicity (cities are often locations for media organizations, for example) and recognition that these are important places for societies and states.

However, the interplay between cities and violence goes beyond targeting. In ancient warfare, Virilio observes that "defence was not speeding up but slowing down. The preparation for war was the wall, the rampart, the fortress" (Virilio and Lotringer, 2008: 20). Militarization now takes place in different ways, though, and military space is now "a space of time, a space of the rapidity of attack and reaction" (Virilio and Lotringer, 2008: 86). For example, missiles and munitions mean that city walls and fortifications cannot offer adequate defense from destruction by large-scale state violence (even to the limited extent that they ever did). Attempts to protect against the speed of attack – for example, through missile defense systems to defend against attacks with nuclear weapons – have not been realized and there are serious doubts about their feasibility. Even ignoring the limitations of missile defense systems, the openness on which cities depend – for example, substantial amounts of goods (including illegal and gray market goods) move in and out of cities – means that ruling out all potential attacks is not feasible.

The role of the military and militarization is also important here. Inasmuch as it plays a role as a defensive site, the city has always been – at least to an extent – militarized. With everything from city walls and barracks to more opaque military involvement (for example, being used to supply essential goods for conflict and populations to serve in wars) cities have always been, to a certain extent, militarized places. It is, therefore, important to follow Woodward in considering how

military activities "imprint themselves across … spaces, places, environments and landscapes. Furthermore, these military geographies also have a far wider imprint than armed conflict, marking and shaping places and spaces far distant from the points of military engagement – including those that I see from my [Gateshead] window" (Woodward, 2005: 730). This history is visible in many urban landscapes. The location of settlements (for example, on hilltops) often has defensive elements and the structure of cities (for example, inside city walls) also often reflects a defensive role. While some of aspects of urban violence and militarization should be seen as novel, we should start by making clear that it is part of a far longer history.

While this move to modern war may have changed the ways in which cities can work as defensive sites – an acceleration of the possibilities for offense limiting the role of military fortifications – cities certainly have not lost their relevance to contemporary warfare. For Woodward (2005: 730), "if all geographies are military in some way, the actual pursuit of armed conflict is only part of the story." For example, Woodward (2005: 718–719) notes the presence of weapons factories and barracks in urban sites – militaries continue to need access to weapons and personal.

Cities have played key roles in the last century's major wars. They were important in resourcing the violence and then in memorializing the losses (see Edkins, 2003: Chapter 3). While there are complex cultural reasons why urban areas have been the target of political violence (as will be discussed in more depth below), this importance of cities in war has led to the development of multiple ways of targeting cities in such conflict. This is perhaps best-illustrated in Walt Rostow's (1981) strategy for the bombing of Germany during World War II; targets are justified in terms of potential damage to the German war effort and the German state. As will be discussed below, this development of rationalities for the targeting of cities continues in 'new' wars too.

While one would hope that there will not be another world war, cities and violence remain intimately linked today. As Graham (2004a: 5) argues, we see cities '*reemerge*, as *geopolitically* charged spaces.' Instead of offering spaces outside of conflict and militarization, in 'new' wars and newer types of violence cities remain very much caught up in the conflicts of contemporary geopolitics. As will be discussed below, cities have also played and continue to play significant roles in the development of 'new' wars, urban warfare, and terrorism.

Military Operations in Urban Terrain and Vertical Geopolitics

For some time now, it has appeared that the US and its allies have the ability to dominate large-scale, high-technology conflicts in 'flat' open spaces. This was especially clear in the 'Turkey shoot' that ended the 1991 Gulf War, with large numbers of retreating Iraqi troops wiped out from above with relatively low risk to US-led forces (Hiro, 1992: 388–389; Matthews, 1993: 166–167). Echoes of this were seen in the more recent violence in Iraq, where – when Iraqi fighters were grouped together in relatively open spaces – the US was also able to use 'dumb' munitions to kill everyone within a given area at low risk to their forces (Clark, 2003: 44).

However, such dominance is not without consequences. Unsurprisingly, opponents of US-led forces frequently seek not to be wiped out from above. In part, because of the capacity of US-led forces to wipe out opponents when they gather in 'flat' and open spaces, such opponents are increasingly fighting away from this type of space (see Graham, 2004b: 12–13). Instead of massing forces in large open spaces, there has been a move toward 'new' ways of fighting in spaces where targeting is harder. While various nonurban landscapes have been used (for example, Afghan caves have been important), the use of urban areas is clearly most relevant to this article.

There has increasingly been a shift away from the geopolitics of conflict being seen as a 'flat' discourse – pivoting around open spaces – and toward a vertical geopolitics (Graham, 2004b). Faced with the prospect of rapid annihilation should they try to contest US action in 'flat' spaces, opponents of US geopolitical goals are increasingly moving to deploying violence in the vertical spaces of cities. As Elden (2013: 49) argues, "[j]ust as the world does not just exist as a surface, nor should our theorisations of it; security goes up and down; space is volumetric."

'New' wars are thus increasingly urban and, as discussed above, cities are reemerging as prominent and volumetric geopolitical spaces. This has led to the development of strategies aimed at using military violence 'better' in cities and to cities being increasingly theorized as problems facing the deployment of military violence. This article will discuss these two approaches consecutively.

The increasing attention given to cities as spaces and places of military violence is leading to increasing focus on training for and simulation of military action in urban areas; the US military, in particular, has aimed to increase its ability to use violence in urban areas. While this is often linked to military actions as part of the 'war on terror,' one might note an October 2001 publication that was already referring to a "renaissance in interest in military operations in urban terrain (MOUT) within the American military" (Desch, 2001: 2). Two 'conventional' explanations for this are acknowledged: "current world demographic trends suggest that increasingly the population of the world will be urban. Thus, the likelihood of any military operation being conducted in urban areas is also likely to increase. Second, a number of recent catastrophic failures in MOUT – the costly U.S. fight in the Bakara market place in Mogudishu, Somalia, in 1993, and the Russian defeat in the first battle for Grozny Chechnya, in 1994–95 – have refocused attention on the challenges of MOUT operations in general" (Desch, 2001: 2). It is also suggested that "there is another factor that is also prompting the U.S. military to think hard about MOUT. This factor is the dramatically different international and domestic security environment the United States faces now that the Cold War is over [including] the fact of U.S. primacy – there is no single nation or even likely coalition of nations that can challenge the United States in conventional or nuclear combat" (Desch, 2001: 2). Subsequent events – including problems faced by US-led forces fighting in Iraqi cities – have clearly worked to further intensify the attention focused on military operations in urban terrain (MOUT).

This interest in MOUT has led to a range of different approaches, which are having impact in terms of both military and (as will be discussed below) civilian approaches to maintaining order. There has been an emphasis on military 'boots on the ground' (sometimes alongside law enforcement) in attempts to maintain order in cities in the course of operations such as Operation Iraqi Freedom. Large numbers of troops can be needed to do this, to the extent that some have argued that the force requirements for doing so in Iraq exceeded what was available to the US government (Enterline et al., 2009: 250). However, such deployments have also been accompanied by a range of high-technology attempts to engage with cities and urban conflict, with often strikingly different effects. A broad range of technologies – from drones to 'precision' guided weapons to Area Impact Munitions – have been used to surveil urban areas and to deploy violence in these areas (Gregory, 2008). There is a move from the 'flat' spaces of older concepts of geopolitics to violence taking place in vertical cityscapes that necessitate very different tactics – for example, 'boots on the ground' have been needed (bringing an associated increase in casualty levels on the US-led side).

Alongside such deployments of technology, there has been what Gregory (2008) terms a cultural turn in counterinsurgency: a type of 'armed social work' in which local communities and the views of local populations become very much part of the terrain or 'battlespace' in which conflict takes place. While this may seem progressive, it is worth following Gregory by flagging up that this type of armed social work need not be an alternative to large-scale urban military violence; instead, it is intimately linked with violence. For example, a focus on 'precision' weapons and on understanding the urban areas in which conflict takes place can very much exist alongside the use of Area Impact Munitions to destroy all life in significant parts of these areas. It is not the case that a focus on greater cultural understanding will necessarily lead to a decrease in military violence; while this 'armed social work' is clearly aimed at achieving a reduction in insurgent violence, its efficacy at achieving this goal has not been evidenced.

Alongside this move to more 'boots on the ground,' there are (potentially science fiction-inspired) dreams of new ways of doing urban violence. There are hopes that increasing surveillance and precision targeting might render city spaces flat and replace 'our' boots on the ground with machines (Graham, 2008). Again, though, this approach to MOUT has not been shown to work with current technologies.

Killing Disease and Killing Cities

The difficulties caused for militaries by these vertical spaces can mean that these spaces are conceptualized in diverse and often deeply troubling ways. Urban life and the structures of urban areas can be conceptualized as a problem for military intervention; part of the goal of intervention is to 'treat' this problem. As Graham (2003) argues, this has led to particular types of urban life being seen through disease metaphors: for example, to conceptualizing "Palestinian urbanization as a 'cancer' whose spread is undermining the organic constitution of the modern State of Israel." Types of urban violence or disorder can then be viewed as part of the disease, while large-scale military violence may be constructed as somehow cleansing or healing of the disease. This discourse can lead to everything from calls to simply destroy significant urban areas through area bombing or large-scale military operations through to attempts to 'treat' the illness through counterinsurgency operations.

The idea of urbicide is significant here – attempts to kill cities, to make urban life impossible, can be presented as 'surgery,' as the use of military, police, or economic violence to destroy 'sick' urban life so that the 'healthy' can be saved. The term 'urbicide' has been used in a range of ways. Berman (1987) introduced this idea in order to describe how "[t]he physical and social destruction of the [South Bronx] began with the construction of the Cross Bronx Expressway in the late 1950s and early 1960s spreading gradually southward … then in the early 1970s the disintegration began to spread at a spectacular pace, devouring house after house and block after block, displacing thousands of people like some inexorable plague." This meant that, where blocks of housing had once stood, the city "now contained nothing at all. Most of the rubble from their buildings had long since been carted away and they were as open and empty as the desert." For Shaw (2004), though, urbicide is more explicitly violent: "a form of genocide, the fundamentally illegitimate form of modern war in which a civilian population as such is targeted for destruction by armed force." For Coward (2004) this is targeted at destroying heterogeneity – at destroying a plural culture.

'Treatments' for urban life might thus go so far as urbicide. This can range from 'unbuilding' programs of planning decay (see Graham, 2004c) through to violence being deployed in much more literal ways – as in Coward's (2004) discussion of the targeting of heterogeneity. State action that goes so far as to deliberately destroy essential urban infrastructure can be presented as a move to save the body politic by destroying the diseased tissue that is supposedly parasitic upon it. A market for weapons-come-redevelopment-tools such as armored bulldozers – to be used in modifying and demolishing urban areas where violent conflict is taking place – has thus emerged: redevelopment and counterinsurgency offered by a single machine (Gregory, 2008). This militarization and these metaphors feed into particular types of urban violence, which continue the long history of complex interrelationships between urban areas and conflict but also offer some novel features.

Policing

The mutually constitutive nature of cities and war is not just visible in such stark examples as cities contributing to war efforts, being targets of military violence or being places where insurgency takes place. Increasingly, there is a blurring of lines between 'military' violence abroad and 'domestic' law enforcement and control within state boundaries. Military technologies and techniques are increasingly being turned inwards and metaphors of a 'war on terror' have led in some circumstances to an escalation of violence and an escalation of the tactics that are viewed as justifiable within the 'domestic' sphere.

For example, Balko (2013: xvi) documents a move to 'SWAT'-style policing in some US cities on some occasions:

military technology proliferated, and war metaphors (for example, the 'war on drugs' or 'war on terror') were applied to policing. Fisher (2010: 5) sees SWAT-style and militarized approaches to and equipment for policing increasingly becoming part of – and militarizing – routine duties such as crowd control. The same military theories of 'new' and networked wars that have been influential in developments in military doctrine (e.g., Arquilla and Ronfeldt, 1996) do not offer a clear divide between war and 'domestic' violence. Instead, theories of net war have long conceptualized a spectrum of opponents from 'terrorist' groups that aim to kill civilians through to wholly or largely peaceful protest groups.

There has also been a militarization of urban policing through drawing on military surveillance technology and techniques for policing uses. Drones have been marketed – and used – for policing within the US: what Wall and Monahan (2011) call a 'drone stare' has moved from Afghan and Iraqi war zones into US cities. Data-led surveillance has also pulled much of the virtual chatter of urban life into a blurring of military-policing boundaries: one of the interesting aspects of recent revelations regarding National Security Agency and Government Communications Headquarters actions has been moves to allow military agencies notionally targeting those 'outside' the state to increasingly conduct surveillance domestically, and for policing agencies 'inside' to work with military agencies 'outside' in order to access intelligence on citizens.

Likewise, military and militarized technologies and techniques are used in the hope of finding – in these data – ways to detect a 'terrorist' before they have even been radicalized. Consortia of military, arms and security companies are working to monitor the ways in which people live in, move through and travel from cities – in the hope of finding some 'print' in these data from which our future proclivities can be predicted (Hall and Mendel, 2012). In contrast to hopes that cities may provide a space of openness and possibility, the figurative violence here lies in a closing down of the future (which may also feed into the far more literal violence of the detention of citizens the restriction of the ways in which some 'suspect' people can live in, come to and move between cities) (Hall and Mendel, 2013).

Clearly, various types of nonstate violence can be part of urban life. A lot of attention is frequently given to the spectacular violence that one occasionally – but too often – sees in terrorist attacks. These attacks are horrific for those caught up in them, and for those who lose people close to them as a result. Such attacks are, however, also thankfully uncommon. There are reasons why people may be especially concerned about this type of violence as opposed to more 'everyday' risks that come with urban life, but clearly there are many things in the city, which are far more likely to harm one than terrorist attack: ranging from road traffic to more commonplace crimes. Cities also remain attractive places to live. One should, therefore, be cautious about militarized responses to the diversity of urban life, which are framed as responses to spectacular and horrific, but thankfully very rare, terrorist violence: such militarized approaches carry real costs, and risk damaging significant aspects of urban life.

This is not to say that one can find any complete escape from the entangling of war and the city, but I would emphasize the opportunities within and beyond the war machine. While Deleuze and Guattari (1988: 214) are right to warn that "[t]here is fascism when a war machine is installed in each hole, in every niche," I would also follow their emphasis on the multiple lines of flight that remain possible for and from war machines. Thrift (2005: 337) responds to Virilio's theorizing of the city by arguing for "social science research on the city that makes it possible to construct a somewhat more hesitant account, not least … because hesitant accounts are closer to the way the world is." Thinking about and engaging with urban violence and militarization might let us find ways to hesitate and to rethink the long-standing relationship between war and cities in some different ways: finding lines of flight that do not end in war, violence, or urbicide.

Conclusions

As this article has shown, there is a long history of the coevolution of cities and war and military-city links are longstanding and still developing. This is very visible today through urban warfare, MOUT and increasing surveillance, control and militarization. Clearly, violence takes place in and has been deployed against cities in some horrific ways. However, there is also hope for ways of developing urban life that do not simply feed into this militarization and violence. I would argue that there is a real imperative to build and develop such approaches.

Urban violence is a serious issue for many, and impacts on people's lives in diverse ways. However, it is also worth noting than many people continue to choose to live in cities. While cities may be targeted, they can also be remarkably hard to destroy. Urban life can persist, even during and after brutal assaults. While war and violence are often viewed as solely disruptive to urban life, this article has instead detailed the mutually constitutive nature of cities and certain types of urban violence. It is also, though, worth looking at other responses to perceived urban threats and violence – reactions that may build on this militarization to offer lines of flight beyond it, or that may offer alternatives or challenges to it.

Doreen Massey (1993: 66–67) draws on the example of a walk down her local Kilburn High Street in order to illustrate a progressive sense of place in the city: taking real joy in the noise, bustle, and diversity of urban life. Against an excessive emphasis on threats of terrorism and against the increasing militarization and surveillance of urban life, I would highlight the importance of making space for this noise and disorder (see Mendel and Riesch, forthcoming). The 'medicine' often proffered in the name of making urban life safer and offering protection from violence may be far more dangerous than the 'illness.' States can still project significantly larger scale violence than any nonstate terrorist group is likely to in the foreseeable future, and moves toward increasing surveillance and a closing down of future possibilities and uncertainties also threaten the freedoms and opportunities that make urban life so attractive. The complexities of cities' vertical geopolitics offer pleasures and possibilities that cannot be found in (literally or metaphorically) 'flattened' places.

One of the responses to the 7 July 2005 attacks on London was the *we're not afraid* campaign: a refusal to respond to terrorism by bowing to fear. Responses in London included initiatives that built a sense of urban community as part of a robust response to terrorism: for example, supporting those

who felt nervous after the attacks in their use of public transport. The response to urban violence – even quite horrific violence – thus need not always or not only be a crackdown or an increasing curtailing of possibility: creative responses that build on and strengthen urban communities are also real options. Rather than accepting ever-increasing surveillance and militarization in response to the perceived threat of terrorism and other urban violence, I would argue instead for a refusal of fear and for attempts to build a more progressive sense of place and city. Rather than bowing to measures which – although nominally aimed at protecting cities – risk destroying what makes urban life so attractive, a refusal of fear and a refusal of certain control measures may let the noise and bustle of cities continue, even if the price of doing so is accepting some risk of nonstate urban violence. While acknowledging the codevelopment of cities, war, and violence I would hope – albeit hesitantly – that urban life still offers lines of flight in more progressive directions.

Acknowledgments

I am grateful to Nicholas Fyfe for his helpful feedback on this work, and have also benefited from discussions with students as part of Dundee University's *Cities* module. This article also draws on ideas discussed with Hauke Riesch, as part of our work on a different topic.

See also: Cities: Capital, Global, and World; Cities: Internal Structure; Gated Communities; Geopolitics; Networks, Geography of; Place in Geography; Political Geography; Surveillance and Privacy, Geography of; Urban Geography; War and Terrorism, Geography of.

Bibliography

Arquilla, J., Ronfeldt, D.F., 1996. The Advent of Netwar. RAND, Santa Monica.
Balko, R., 2013. Rise of the Warrior Cop: The Militarization of America's Police Forces. PublicAffairs, New York.
Berman, M., 1987. Among the ruins. New Internationalist (178).
Clark, G.W.K., 2003. Winning Modern Wars: Iraq, Terrorism, and the American Empire. PublicAffairs, New York.
Coward, M., 2004. Urbicide in Bosnia. In: Graham, S. (Ed.), Cities, War and Terrorism: Towards an Urban Geopolitics. Blackwell, Oxford.
Deleuze, G., Guattari, F., 1988 (B. Massumi, Trans.). In: A Thousand Plateaus: Capitalism & Schizophrenia. Athlone, London.
Desch, M.C., 2001. Why MOUT now? In: Desch, M.C. (Ed.), Soldiers in Cities: Military Operations in Urban Terrain. Strategic Studies Institute, Carlisle.
Edkins, J., 2003. Trauma and the Memory of Politics. Cambridge University Press, Cambridge.
Elden, S., 2013. Secure the volume: vertical geopolitics and the depth of power. Political Geography 34 (1), 35–51.
Enterline, A.J., Greig, J.M., Gortzak, Y., 2009. Testing Shinseki: speed, mass and insurgency in post-war Iraq. Defense and Security Analysis 25 (3), 235–253.
Fisher, J., 2010. SWAT Madness and the Militarization of America's Police. Praeger, Santa Barbara.
Graham, S., Jan–Feb 2003. Lessons in urbicide. New Left Review 19.
Graham, S., 2004a. Introduction: cities, warfare, and states of emergency. In: Graham, S. (Ed.), Cities, War and Terrorism: Towards an Urban Geopolitics. Blackwell, Oxford, 202–226.
Graham, S., 2004b. Vertical geopolitics: Baghdad and after. Antipode: A Radical Journal of Geography 36 (1), 12–23.
Graham, S., 2004c. Postmortem city. City: Analysis of Urban Trends, Culture, Theory, Policy, Action 8 (2), 165–196.
Graham, S., 2008. Robowar™ Dreams. US Military Technophilia and Global South Urbanisation 12 (1), 25–49.
Gregory, D., 2008. 'The rush to the intimate': counterinsurgency and the cultural turn. Radical Philosophy 150, 8–23.
Hewitt, K., 1983. Place annihilation: area bombing and the fate of urban places. Annals of the Association of American Geographers 73 (2), 257–284.
Hewitt, K., 2009. Proving grounds of urbicide: civil and urban perspectives on the bombing of capital cities. ACME: An International E-journal for Critical Geographies 8 (2), 340–375.
Hall, A., Mendel, J., 2012. Threatprints, threads and triggers: imaginaries of risk in the 'war on terror'. Journal of Cultural Economy 5, 9.
Hall, A., Mendel, J., 2013. Digital traces and the 'print' of threat: targeting populations in the war on terror. In: Harvey, P., Casella, E., Evans, G., Knox, H., McLean, C., Silva, E., Thoburn, N., Woodward, K. (Eds.), Objects and Materials. A Routledge Companion. Routledge, London.
Hiro, D., 1992. Desert Shield to Desert Storm: The Second Gulf War. Routledge, London.
Massey, D., 1993. Power-geometry and a progressive sense of place. In: Bird, J., Curtis, B., Putnam, T., Tickner, L. (Eds.), Mapping the Futures: Local Cultures, Global Change. Routledge, London, pp. 59–69.
Matthews, K., 1993. The Gulf Conflict and International Relations. Routledge, London.
Mendel, J., Riesch, H., forthcoming. Science blogging below-the-line: a progressive sense of place? In: Cupples, J., Lukinbeal, C., Mains, S. (Eds.), Mediated Geographies/Geograpies of Media. Springer, The Netherlands.
Rostow, W.W., 1981. Pre-invasion Bombing Strategy: General Eisenhower's Decision of March 25, 1944. University of Texas Press, Texas.
Shaw, M., 2004. New wars of the city: relationships of 'urbicide' and 'genocide'. In: Graham, S. (Ed.), Cities, War and Terrorism: Towards an Urban Geopolitics. Blackwell, Oxford.
Thrift, N., 2005. Panicsville: Paul virilio and the esthetic of disaster. Cultural Politics 1 (9), 337–348.
Virilio, P., Lotringer, S., 2008. Pure War, new ed. Semiotext(e), Los Angeles.
Wall, T., Monahan, T., 2011. Surveillance and violence from afar: the politics of drones and liminal security-scapes. Theoretical Criminology 15 (3), 239–254.
Woodward, R., 2005. From military geography to militarism's geographies: disciplinary engagements with the geographies of militarism and military activities. Progress in Human Geography 29 (6), 718–740.

Urbanization

Gordon McGranahan, International Institute for Environment and Development, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The world's population growth in coming decades will be predominantly in urban Asia and Africa. With about half of the world's population now urban, the salience of the rural–urban dichotomy is declining: urban densities are decreasing; polycentric regions are on the rise; and the cultural, social and economic distinctions between rural and urban are increasingly blurred. Nevertheless, urbanization provides critical opportunities and challenges for sustainable development. Urbanization and economic growth have combined to create various global environmental pressures, and often bring social and economic inequalities. However, dense urban settlement has environmental advantages, and measures to inhibit urbanization often exacerbate inequalities.

Background

To say a country or region is urbanizing implies that it is becoming more urban. Demographers have interpreted this to mean that a growing share of the population lives in urban settlements, with the level of urbanization being the urban share, and the rate of urbanization being the annual growth rate of this urban share (United Nations Population Division, 2014). Many other urban researchers also define urbanization in this way, but the term can also refer to land use, economic activity, or culture becoming more urban; and sometimes carries with it the presumption that these different dimensions of urbanization occur together. Increasingly, they do not. Indeed, some of the conventional distinctions made between rural and urban settlement patterns, economic activities, and cultures are becoming increasingly blurred. Moreover, the implications of demographic urbanization can be different from when economic activity, cultural norms, or land use becomes more urban. For example, demographic urbanization implies a shift in population from less- to more-dense settlement patterns; whereas the conversion of rural to urban land, sometimes described as urbanization, is increasingly the result of a shift from more- to less-dense urban settlement patterns.

For this article, the demographic definition will be used. The next section reviews estimates of past and future urban populations and urbanization levels, relying heavily on the official United Nations statistics (as indicated in **Box 1**, these statistics need to be interpreted with care). Globally, we are at a point where the rates of population growth, urbanization, and urban population growth have all been declining for decades, but the number of additional people living in urban areas each year is just peaking, and accounts for virtually all of the world's population increase. The following section looks beyond the narrow rural–urban dichotomy to consider changing settlement forms and rural–urban linkages. The final sections examine urbanization and the challenges and opportunities it creates for economic development, social inclusion, and environmental health and sustainability.

Estimates of Past and Future Urban Populations and Urbanization Levels

Contrary to popular claims, contemporary urbanization and urban population growth rates are not unprecedented. Sub-Saharan Africa is not the fastest urbanizing region now, let alone in history. As described below, available estimates indicate that average world urban growth rates have actually been declining since at least the 1950s, and those for Sub-Saharan Africa since the 1970s (United Nations Population Division, 2014). Urbanization rates are highest in Asia and global rates have been declining over the past half century. Both urbanization rates and urban growth rates are predicted to continue their descent.

The absolute increase in urban populations each year *is* unprecedented, however. Despite the declining growth rates, the estimated annual increase in the number of urban people in the world is just peaking at almost 80 million, though it is now set to begin a slow decline. The United Nation's urban population projection for 2010–50 has an additional 940 million urban dwellers in Africa and 1450 million in Asia (United Nations Population Division, 2014). These two population increments combine to form over a third of the estimated urban population of the world in 2050, and as illustrated in **Figure 1** are about the same magnitude as the overall population growth predicted for the period. Accommodating this many more urban dwellers efficiently and equitably will be difficult.

It can be helpful to think of these urban demographics in terms of two stylized transitions, which are overlapping, interrelated, but distinct. The first, known as the demographic transition, posits that when modern socioeconomic development leads to falling mortality rates in previously stable populations, the population size grows until fertility eventually falls to the point where the population again stabilizes, at lower mortality and fertility rates (Dyson, 2010). The second, known as the urban transition, associates modern socioeconomic development with a shift from a predominantly agricultural population living mostly in dispersed rural settlements toward a predominantly urban-based population engaged mostly in industrial and service activities.

The first set of rows in **Table 1** illustrate the changing overall population growth rates for major regions of the world, and conform roughly to the demographic transition. While population growth rates are now low or falling in all regions, European and North American population growth rates have been the lowest, with the European rates now close to zero (despite migration into the region). At the other extreme, the population growth rates in Sub-Saharan Africa are still well above 2%. Growth rates in Asia and Latin America have been declining particularly rapidly, with Asia's decline having started

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72120-9

Box 1 Interpreting and misinterpreting international statistics on urban populations

Using Different Definitions of 'Urban'

There is no international consensus on how to identify or determine the boundaries of urban areas, as evidenced by the diversity of national urban definition summaries in the reports on Urbanization Prospects by the United Nations Population Division (2014). Definitions may prioritize population criteria (e.g., settlement size), functional criteria (e.g., the share of the working population engaged in agricultural activities), or administrative criteria. The United Nations has tried to keep a degree of conformity in the sorts of settlements labeled 'urban.' Nevertheless, among countries with size criteria, the minimum size generally ranges from 1000 to 5000. Outliers include Sweden, where a built up area with at least 200 households with gaps of no more than 200 m between them is defined as urban, and Mali where the censuses up to 1987 used a cutoff of 5000, the 1998 census used a cutoff of 30 000, and the 2009 census used a cutoff of 40 000. Attempts to come up with internationally comparable definitions have relied mainly on size and density criteria (World Bank, 2009). As the resolution and availability of remote sensing data improve, users of new data systems may be able to apply their own urban definitions, based on variables such as density, contiguity, and distance (Montgomery and Balk, 2011). For the moment, however, there is little standardization or flexibility, and for most purposes, country-specific definitions of urban settlements must be accepted.

Confusing Urban Population Growth and Urbanization

Urbanization is often treated as equivalent to urban population growth. Thus the Wikipedia definition of urbanization at the time of writing this (18 January 2013) states that: "Urbanization refers to the increasing number of people that live in urban areas." However, only about half of estimated global urban population growth can be ascribed to the increasing share of the population that is urban (i.e., to urbanization), the other half being the result of natural population growth. Thus, the world's population between 2000 and 2010 is estimated to have grown at 1.2% per annum, and the urbanization level is estimated to have been growing at 1.0% per annum, and the urban population growth rate has been 2.3.

Relying on Old Census Data and Simplified Population Projections

Reliable estimates of rural and urban population have long depended on national censuses, which aspire to complete counts of everyone in the country, in their place of residence, at the time of the census. Modern technology is slowly rendering enumeration-based censuses obsolete, particularly in countries with well-developed registration systems; and in Western Europe, registration-based and mixed censuses are becoming the norm. Most other international estimates of urban and rural population still rely heavily on enumeration censuses undertaken every five or more often ten years. Even censuses are likely to undercount the homeless, illegal, or unregistered migrants and people living in informal settlements. Moreover, censuses can also take several years to process and release, and in low income or politically unstable countries censuses are often delayed or even canceled. Thus current as well as future population estimates rely on projections. Unfortunately, United Nations projections, on which most researchers depend, are in need of improvement to reduce systematic bias (Bocquier, forthcoming).

more recently. These differences are consistent with the contrasting economic situations in the different regions, with Europe and North America the wealthiest regions, Sub-Saharan Africa the poorest, and Asia quite mixed but with many of the most rapidly growing countries.

The second set of rows in Table 1 contains the urbanization rates, which together with Table 2 reflect movement along the urban transition. Asia is currently the fastest urbanizing region, with an average urbanization rate estimated at 1.4% this decade, with China in particular both urbanizing and growing economically very rapidly. Sub-Saharan Africa is also still urbanizing at about 1.3%, and its urbanization rate is declining more slowly than Asia's, with the result that by the 2020s Sub-Saharan Africa is projected to be urbanizing faster than Asia. With the exception of North Africa, all the other regions are urbanizing at well under half a percent a year, which is consistent with all of them having had more than two thirds of their populations living in urban areas in 2010 (see Table 2).

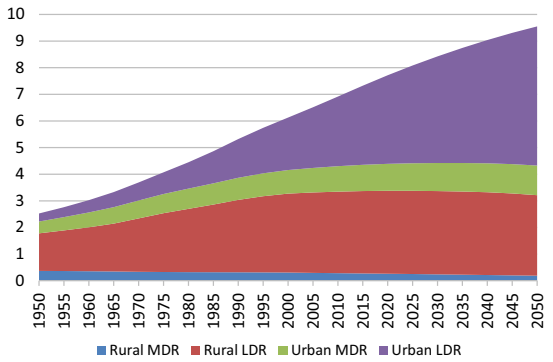

Figure 1 Population estimates of the world 1950–2050 by rural/urban and more/less developed regions. Source of Statistics: United Nations Population Division, 2014. World Urbanization Prospects: The 2014 Revision [Online]. Available: http://esa.un.org/unpd/wup (accessed 21.07.14.).

The third set of rows in Table 1, the urban population growth rates, are (approximately) the sum of the population growth rates and the urbanization rates. Sub-Saharan Africa's 4.0% urban growth rate this decade is the highest of any region, despite being well below the 5.5% registered in the 1950s. More than half of this growth can be accounted for by its high overall population growth rate. Asia, by way of contrast, has a lower (2.3%) urban population growth rate, but more than half of this growth reflects urbanization.

While the demographic and urban transitions can help make sense of the statistics summarized in Tables 1 and 2, the underlying relationships and drivers are still unclear and the generalizations underpinning the transitions hold limited predictive power or policy relevance. Considerable debate remains over the principal shifts in the demographic transition, including why, when, and for whom health improves and mortality rates decline, and whether the primary driver of fertility decline is, for example, changes in education (Lutz and KC, 2011), urbanization (Martine et al., 2013) or better access to contraception (Cleland et al., 2012). The relationships between the two transitions are also poorly understood, with many observers treating them

Table 1 Estimates of population growth rates, urbanization rates, and urban population growth rates (all in % per annum) by region for decades between 1950 and 2050

	1950–1960	*1960–1970*	*1970–1980*	*1980–1990*	*1990–2000*	*2000–2010*	*2010–2020*	*2020–2030*	*2030–2040*	*2040–2050*
Population growth rates										
World	1.8	2.0	1.9	1.8	1.4	1.2	1.1	0.9	0.7	0.6
Sub-Saharan Africa	2.1	2.5	2.8	2.8	2.7	2.7	2.6	2.4	2.2	2.0
Northern Africa	2.7	2.7	2.5	2.6	1.9	1.7	1.6	1.3	1.0	0.8
Asia	2.0	2.3	2.2	2.0	1.5	1.1	1.0	0.6	0.4	0.2
Europe	1.0	0.8	0.6	0.4	0.1	0.2	0.0	−0.1	−0.2	−0.2
Latin America and the Caribbean	2.8	2.7	2.4	2.0	1.7	1.3	1.0	0.8	0.5	0.3
Northern America	1.8	1.3	1.0	1.0	1.1	0.9	0.8	0.7	0.6	0.5
Oceania	2.2	2.2	1.6	1.6	1.5	1.6	1.4	1.2	1.0	0.9
Urbanization rates										
World	1.3	0.8	0.7	0.9	0.8	1.0	0.9	0.7	0.5	0.5
Sub-Saharan Africa	3.3	2.1	2.1	1.9	1.3	1.4	1.3	1.2	1.0	0.9
Northern Africa	2.0	1.6	1.1	1.0	0.6	0.4	0.5	0.5	0.6	0.6
Asia	1.9	1.2	1.4	1.7	1.5	1.8	1.4	1.0	0.7	0.6
Europe	1.0	1.0	0.7	0.4	0.1	0.2	0.3	0.3	0.3	0.3
Latin America and the Caribbean	1.8	1.5	1.2	0.9	0.7	0.4	0.3	0.2	0.2	0.2
Northern America	0.9	0.5	0.0	0.2	0.5	0.2	0.2	0.2	0.2	0.2
Oceania	0.7	0.6	0.0	−0.1	0.0	0.0	0.0	0.1	0.1	0.2
Urban population growth rates										
World	3.2	2.8	2.6	2.7	2.3	2.3	2.0	1.5	1.2	1.0
Sub-Saharan Africa	5.5	4.6	4.9	4.8	4.0	4.1	4.0	3.6	3.2	2.9
Northern Africa	4.7	4.4	3.6	3.6	2.5	2.1	2.1	1.8	1.6	1.4
Asia	3.9	3.5	3.5	3.8	3.0	3.0	2.3	1.6	1.1	0.8
Europe	2.0	1.8	1.2	0.8	0.2	0.4	0.3	0.2	0.2	0.1
Latin America and the Caribbean	4.6	4.2	3.6	3.0	2.4	1.7	1.4	1.1	0.7	0.5
Northern America	2.7	1.8	1.0	1.2	1.6	1.2	1.0	0.9	0.8	0.6
Oceania	3.0	2.9	1.6	1.5	1.5	1.6	1.4	1.2	1.1	1.0

Source of statistics: United Nations Population Division, 2014. World Urbanization Prospects: The 2014 Revision [Online]. Available: http://esa.un.org/unpd/wup (accessed 21.07.14).

Table 2 Urbanization levels (percent urban) 1950–2050 by geographical region

Geographical area	*1950*	*1970*	*1990*	*2010*	*2030*	*2050*
World	30	37	43	52	60	66
Sub-Saharan Africa	11	18	27	35	45	55
Northern Africa	26	37	46	50	56	63
Asia	18	24	32	45	56	64
China	12	17	26	49	69	76
India	17	20	26	31	39	50
Europe	52	63	70	73	77	82
Latin America and the Caribbean	41	57	71	78	83	86
Northern America	64	74	75	81	84	87
Oceania	62	71	71	71	71	74

Source of statistics: United Nations Population Division, 2014. World Urbanization Prospects: The 2014 Revision [Online]. Available: http://esa.un.org/unpd/wup (accessed 21.07.14).

as largely independent but some presenting the urban transition as driven primarily by the demographic transition (Dyson, 2010). Moreover, the relationship between urbanization and socioeconomic development is interpreted in widely divergent ways.

More important, the superficial inexorability of the demographic and urban transition models hide the human struggles involved in urbanization – the conflicts it engenders, the opportunities it creates, and the enormous individual and collective ingenuity required to overcome the challenges it brings. A simplistic perspective on urbanization can be accentuated by presenting the process of urbanization in narrowly statistical terms, as being driven by people migrating from rural to urban locations or living in locations that are being reclassification from rural to urban. This presents urbanization as a relatively passive process, involving one decision to move or perhaps no moves at all. Actually, people often move back and forth between and among rural and urban areas and urbanization occurs as much through decisions not to go back to rural areas as it does through decisions to move to town. Moreover, reclassification is generally linked to population growth in previously rural localities; and migration is likely to be a bigger factor in the growth of peri-urban populations as well as in a large share of the small towns becoming urban, rather than in existing urban areas.

Moreover, conformity to the demographic and urban transition models can also hide the enormous variety of forms that urbanization can take. Urbanization is an open-ended process, locally and globally. At the local level, it is difficult to predict

which towns and cities will succeed and prosper and which will wither, and what the consequences will be for the urban and rural economies and environments. Seemingly small differences in initial conditions and in the choices taken and changes imposed along the way, can make all the difference. It is also difficult to predict the increasingly global economic and environmental consequences of the system of cities that form the core of the world economy and its dominant sources of culture. It is particularly difficult to disentangle the role of urbanization in transforming the world we live in from the economic, cultural, and technological shifts that tend to accompany urbanization.

Moving beyond the Rural–Urban Dichotomy

It is often pointed out that 'rural–urban' should be understood as a continuum rather than as a dichotomy, though there are at least two continua. One extends from the urban center through the ambiguous periphery into the rural hinterland. The second extends across settlement sizes from the urban megacity, through the ambiguous small towns, down to the isolated rural hamlets and homes. But the very treatment of settlement as occurring in distinct clusters is also a crude and increasingly inaccurate simplification.

At least since the 1980s it has been recognized that some large population concentrations are best conceived as polycentric urban regions. Also, in some parts of the world conventionally 'urban' activities and livelihoods are increasingly found in areas of low density and small settlements. This is not just the outcome of the sprawl and suburbanization of settlements in affluent countries, but also of the sort of transformations ongoing in low- and middle-income countries.

With the urban–rural distinction is losing its salience, researchers have suggested that the classification of settlement forms should be reworked, using a multidimensional approach (Champion and Hugo, 2004). Such a reclassification is likely to emerge slowly, along with changes in mapping and information technologies. But even when considering urbanization in the more traditional sense, it is important to take into account the changing distribution of settlement sizes and forms, changing urban densities, and the relationship between urbanization and rural–urban linkages.

Changing Settlement Sizes and Locations

Settlement population statistics give no indication of sizes around which a disproportionate number of urban settlements cluster or where population growth is particularly rapid. As urbanization has progressed, even with a stable distributional form (Henderson, 2010), the number of large cities has increased. The number of cities with populations over a million grew from 2 in 1800 to 380 in 2000 (Satterthwaite, 2013). More recently, the number of megacities has been growing, from 10 in 1990 to 23 in 2010, with their share of the world's urban population growing from 6.7% to 10.3% over this period (United Nations Population Division, 2014). Nevertheless, in 2010, over half of the world's urban population was still estimated to live in settlements of under half a million.

Over the last two centuries, cities have also shifted their center of gravity from heavily populated Asia toward Europe and North America, and then more recently back toward Asia again. Thus while at the start of the nineteenth century 63% of the world's largest cities were in Asia this had fallen to 22% by the start of the twentieth century; but it was back up to 49% by the start of the twenty-first century (Satterthwaite, 2013: p. 239). These shifts are somewhat similar to those that have been documented for the world's economic center of gravity (Quah, 2011).

In interpreting statistics on the cities in different size categories, it should be kept in mind that the forms of urban settlement are changing along with the sizes. This can easily lead to misinterpretations. Peri-urban growth, declining urban densities, and the increasing prevalence of polycentric urban regions have become part of the new urban transformation. When smaller urban centers in polycentric urban regions grow more rapidly than the larger centers, this can be interpreted as reflecting a shift in population growth toward smaller settlements or as a change in the form of large urban agglomerations; but it is not reducible to either.

Changing Urban Densities

Suburbanization, sprawl, and other manifestations of declining urban densities have been observed in affluent countries at least since the early nineteenth century and particularly in North America where the shift to the automobile was vigorously promoted. Urban densities have also been declining, though less dramatically, in all other major regions of the world. A major recent study of urban settlements with populations over 100 000 (Angel et al., 2011) concluded that average built-up area densities declined at about two percent between 1990 and 2000, with urban land cover growing at about twice the rate of urban population.

Unplanned, low density urban expansion, or urban sprawl, is often blamed for creating settlement forms that are environmentally damaging (Wilson and Chakraborty, 2013) and unhealthy (Frumkin et al., 2004). Damaging sprawl can result from not taking these environmental and health impacts into account when infrastructural and residential decisions are being made, and it is sometimes amplified by 'perverse' subsidies. Particularly in affluent countries, sprawl is often presented as something that smart growth and compact cities need to work against. In lower income countries, declining density can push affordable housing far out from city centers; and finding better ways of reconciling density with healthy and convenient housing for lower income residents is often a major challenge.

Urbanization and Rural–Urban Linkages

Urbanization is often treated as something that just happens in urban settlements, though by definition it involves a shift in population from rural to urban locations, transforming rural as well as urban landscapes and livelihoods (Tacoli, 2006).

Urbanization involves changes in demographic, economic, and environmental flows between rural and urban areas, though not always in predictable ways.

Many families in rapidly urbanizing countries adopt livelihood strategies that span the rural–urban divide. They may invest in securing an urban foothold for at least some family members for at least some periods of time. Rural-bound remittances from urban workers often make up a significant share of the incomes of these multilocal 'households.'

Rural–urban linkages have been identified as an important influence on progress toward the Millennium Development Goals (World Bank, 2013a). Indeed, in reading the following section on urbanization and development, it is important to keep in mind that many of the advantages that cities can provide depend on healthy economic and environmental relations with rural areas; and rural livelihoods and environments depend on what is done in cities.

Urbanization and Sustainable Development

Urbanization and Economic Development

Cross-country data show that higher levels of urbanization are associated with higher per capita incomes, with some scatter and some shift in the relationship over time (see Figure 2). Considerable theory and empirical evidence support the view that urbanization is integral to economic growth. Whereas for agricultural and subsistence economies, it is more productive for people to be dispersed, industry and services favor urban clustering. In the language of urban economists, agglomeration facilitates sharing (e.g., of large facilities), matching (e.g., of jobs and people) and learning (e.g., about more productive ways of working) (Duranton and Puga, 2004). The advantages of urban locations are less tangible for postindustrial production but include advantages of face-to-face contact that electronic communications have not managed to replace (Storper, 2013; especially Chapter 11).

Not all aspects of urbanization are economically advantageous. Sharing, matching, and learning, as well as low transport and communication costs, depend on having people and enterprises being located conveniently in relation to each other. People and enterprises tend to seek out economically advantageous locations and given the opportunity, will seize many of the economic benefits of agglomeration. However, they will also create urban crowding and congestion. Reconciling the various conflicting private and collective interests in cities is a major political and governance challenge. Even strong adherents of the market economy acknowledge that people in urban areas are especially dependent on good government, and not just to defend property rights. Cities are shot through with interrelations and impacts that are external to the market; the benefits of urbanization and the people to whom they accrue, depend on how government agencies and other powerful actors and organized groups plan for and respond to these externalities. Notably, many cities have shown themselves capable of contributing to economic growth despite creating serious social and environmental problems.

Given the potential economic benefits of urbanization, and the attraction of most governments to economic growth, one might expect officials in urbanizing countries to favor rapid urbanization. However, the standard survey of world population policies from 2011 found that 82% of developing countries said that they had implemented policies to curb rural–urban migration (United Nations, 2013). Less than 3% claimed to have policies to encourage rural–urban migration, though it is noteworthy that China was among this small group.

Antiurbanization sentiments among officials in Africa has probably been amplified by a tendency to confuse urban growth with urbanization and to exaggerate the speed at which Sub-Saharan Africa is urbanizing (Potts, 2012; Satterthwaite, 2007). Influential research from the World Bank at the start of the 2000s also portrayed Africa experiencing 'urbanization without growth' (Fay and Opal, 2000), reinforcing a widespread view that urbanization should be curbed. However, later

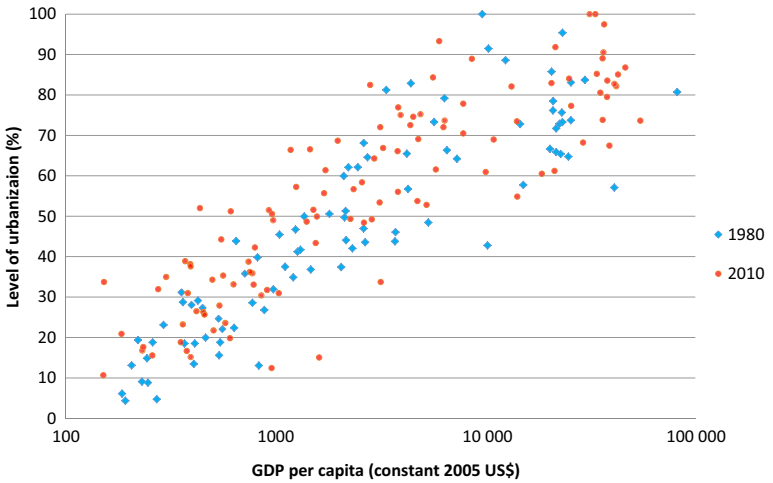

Figure 2 Levels of urbanization plotted against real GDP per capita, 1980 and 2010. Source of statistics: World Bank. 2013b. World Development Indicators [Online]. Available: http://databank.worldbank.org/data/views/variableSelection/selectvariables.aspx?source=world-development-indicators (accessed 27.09.13.).

and more thoroughly researched publications by the World Bank have reversed this interpretation (World Bank, 2009).

A recent review of the evidence that urbanization contributes to economic growth, focusing especially on Asia and Africa, concluded that urbanization has the potential to promote growth and does so when the infrastructure and institutional settings are conducive (Turok and McGranahan, 2013). Even when there is urbanization without economic growth, urbanization can be economically beneficial. If, for example, a country's rural economy is in crisis, pushing people to move to urban areas, the result may be urbanization with economic decline. However, if cities take measures to curb this migration, this is likely to make economic conditions worse, particularly for the aspiring migrants. Of course, the potential economic benefits of urbanization are no justification for leaving rural problems to fester or starving agriculture of productive investment.

Urbanization, Inequality and Social Exclusion

Those who support urbanization as a means of economic advancement generally accept that it also tends to be associated with rising inequalities (Kanbur and Venables, 2005), part of which involves rural–urban differences. Intraurban inequalities are also of growing importance. Even according to conventional income-based measures, a growing share of world poverty is located within urban areas (Ravallion et al., 2007) and most conventional measures miss the depth and scale of urban poverty and inequality (Mitlin and Satterthwaite, 2013).

Somewhat paradoxically, while urbanization is often associated with growing inequality, measures to inhibit urbanization can amplify these inequalities. There are a number of clear examples of this, even among the more successful emerging economies of Africa, Latin America, and Asia (McGranahan and Martine, 2014). South Africa's apartheid system, built around racial controls on urbanization, was not only among the most oppressive systems of urban exclusion when it was in operation but it also left a legacy of spatially and racially fragmented cities. Brazil's more passive resistance to urbanization contributed to the emergence of its favelas, which are also at the center of many of the country's current conflicts and social and economic inequalities. While less clear, China's rural–urban migrant controls, the *hukou* system, could leave a legacy of social and economic inequality and fragmentation.

The exclusion of rural–urban migrants from cities is often bound up with the exclusion of low-income urban populations from urban benefits. UN-Habitat estimates that about 45% of the urban population in developing countries live in 'slum' households, defined as those lacking improved water, improved sanitation, adequate space or solid construction – up from 35% in 1990 (UN-Habitat, 2012: p. 150). A large but 'difficult-to-estimate' share of this population lives in what have come to be termed 'informal settlements,' which represent a 'grey zone of urban exclusion' (Koonings and Kruijt, 2009). Many of these disadvantaged urban residents cannot afford to secure housing in formal markets or through public provisioning. They may not be physically evicted from their cities and towns, but end up living informally, or even illegally, in locations where neither private nor public ownership controls are tightly enforced. Though the net effect is not clear, this intraurban exclusion of disadvantaged groups makes it more difficult for low-income migrants to take advantage of urban benefits and almost certainly pushes some to spend more time in rural areas.

Of course rural exclusion can also exacerbate inequalities and create social problems, and during urbanization the most inequitable outcomes are likely to arise when the same social groups are being excluded from both urban and rural locations. In rural areas, exclusion may come about through policies favoring land consolidation and the increasing dominance of large commercial farms but it can also come about through rural economic decline and population growth. The relatively well-educated and wealthy rural residents may find it relatively easy to find alternative places to work and live in urban areas. Particularly if no efforts are being made in urban areas to accommodate their growing low-income populations, the less educated and poorer rural residents will find it difficult to secure an urban home and livelihood.

Urbanization and Shifting Environmental Burdens

In recent centuries, economic development has shifted the principal locus of urban environmental burdens from local to city-regional and eventually toward global scales (McGranahan et al., 2001). Looking across urban centers in different parts of the world, the poorest urban populations in the poorest countries tend to have the worst environmental health conditions in and around their homes. The ambient environments of heavily industrialized and motorized cities of middle-income countries tend to be most environmentally distressed, with the worst ambient air pollution problems. It is the populations of the wealthiest cities that have the largest global environmental footprints per capita and contribute the most to climate change.

These differently scaled environmental burdens can overlap and create environmental injustices, wherein those who face the most hardship as the result of larger scale burdens, contribute the least to them. In polluted industrial cities, the poorest citizens tend to live downstream and downwind of the major polluters. Globally, those most at risk from climate change include the world's lowest income citizens, rural and urban, who are contributing least to global climate change.

Climate change is creating new environmental risks and exacerbating old ones, including the risks of inadequate water and sanitation provision. Urbanization is putting in place infrastructure that will not only contribute to or protect people from more local environment burdens, but will determine whether cities, and urbanization itself, contribute to climate change mitigation and adaptation. Where urban growth takes place will also influence how many people are living in risk-prone regions and localities.

Urbanization has been part of modern economic growth and is sometimes blamed for contributing to climate change and other global, environmental burdens associated with high consumption levels. It is important, however, to distinguish the effects of urbanization from those of economic growth, and to set both in the context of the prevailing political economy. When urbanization leads to higher productivity, it can be said

to be contributing to the larger ecological footprint that this productivity often facilitates. However, looking forward, this higher productivity is also likely to be a key ingredient in any successful attempt to reduce global environmental burdens without causing undue economic hardship.

Furthermore, excluding people from cities tends to exacerbate the impact of prevailing environmental problems. In low-income countries, addressing severe environmental health problems is generally cheaper in urban than in rural areas; and measures to put off low-income migrants, such as not providing services to rapidly growing low-income areas, is likely to prevent improvements in environmental health conditions for deprived urban residents and also for those forced to stay in rural areas Alternatively, when people in affluent countries move out from cities to low-density suburban or rural settlements, their environmental impact is likely to increase, particularly if they become more dependent on automobile transport.

See also: Cities: Capital, Global, and World; Development Theory in Geography; Development and Urbanization; Eco-cities and Sustainable Urbanism; Growth Management and Urban Planning in the United States; Housing and Cities; Internal Migration: Developing Countries; Just City; Planning and Development; Smart Cities; Spatial Equity; Urban Anthropology; Urban Geography; Urban Life and Health; Urban Sprawl; Urban Studies: Overview; Urban System in Geography; Urbanization in Africa; Urbanization in China.

Bibliography

Angel, S., Parent, J., Civco, D.L., Blei, A.M., 2011. Making Room for a Planet of Cities. Policy Focus Report. Lincoln Institute of Land Policy, Cambridge, MA.

Bocquier, P., The future of urban projections: suggested improvements on the UN Method. Spatial Demography, forthcoming.

Champion, A.G., Hugo, G., 2004. New Forms of Urbanization: Beyond the Urban–Rural Dichotomy. Ashgate, Aldershot.

Cleland, J., Conde-Agudelo, A., Peterson, H., Ross, J., Tsui, A., 2012. Contraception and health. The Lancet 380, 149–156.

Duranton, G., Puga, D., 2004. Microfoundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), Handbook of Regional and Urban Economics, Cities and Geography, vol. 4. Elsevier North Holland, Amsterdam, pp. 2063–2117.

Dyson, T., 2010. Population and Development the Demographic Transition. Zed Books, London.

Fay, M., Opal, C., 2000. Urbanization without Growth: A Not So Uncommon Phenomenon. World Bank, Washington, DC.

Frumkin, H., Frank, L., Jackson, R., 2004. Urban Sprawl and Public Health: Designing, Planning and Building for Healthy Communities. Island Press, Washington, DC.

Henderson, J.V., 2010. Cities and development. Journal of Regional Science 50, 515–540.

Kanbur, R., Venables, A. (Eds.), 2005. Spatial Inequality and Development. Oxford University Press, Oxford.

Koonings, K., Kruijt, D., 2009. The rise of megacities and the urbanisation of informality, exclusion and violence. In: Kruijt, D., Koonings, K. (Eds.), Mega-cities: The Politics of Urban Exclusion and Violence in the Global South. Zed, London, pp. 8–26.

Lutz, W., KC, S., 2011. Global human capital: integrating education and population. Science 333, 587–592.

Martine, G., Alves, J.E., Cavenaghi, S., 2013. Urbanization and Fertility Decline: Cashing in on Structural Change. Discussion Paper. International Institute for Environment and Development (IIED), London.

McGranahan, G., Jacobi, P., Songsore, J., Surjadi, C., Kjellén, M., 2001. The Citizens at Risk: From Urban Sanitation to Sustainable Cities. Earthscan, London.

McGranahan, G., Martine, G. (Eds.), 2014. Urban Growth in Emerging Economies: Lessons from the BRICS. Routledge, Abingdon.

Mitlin, D., Satterthwaite, D., 2013. Urban Poverty in the Global South: Scale and Nature. Routledge, New York.

Montgomery, M., Balk, D., 2011. The urban transition in developing countries: demography meets geography. In: Birch, E.L., Wachter, S.M. (Eds.), Global Urbanization. University of Pennsylvania Press, Philadelphia, PA, pp. 89–106.

Potts, D., 2012. Whatever Happened to Africa's Rapid Urbanisation? Counterpoints. Africa Research Institute, London.

Quah, D., 2011. The global Economy's shifting Centre of gravity. Global Policy 2, 3–9.

Ravallion, M., Chen, S., Sangraula, P., 2007. New Evidence on the Urbanization of Global Poverty. World Bank, Washington.

Satterthwaite, D., 2007. The Transition to a Predominantly Urban World and Its Underpinnings. International Institute for Environment and Development, London.

Satterthwaite, D., 2013. Urbanization in low- and middle-income nations. In: Desai, V., Potter, R.B. (Eds.), The Companion to Development Studies. Routledge.

Storper, M., 2013. Keys to the City: How Economics, Institutions, Social Interactions, and Politics Shape Development. Princeton University Press, Princeton, NJ.

Tacoli, C. (Ed.), 2006. The Earthscan Reader in Rural–Urban Linkages. Earthscan, London and Sterling, VA.

Turok, I., McGranahan, G., 2013. Urbanization and economic growth: the arguments and evidence for Africa and Asia. Environment and Urbanization 25, 465–482.

UN-Habitat, 2012. The State of the World's Cities Report 2012/2013: Prosperity of Cities. United Nations Human Settlements Programme, Nairobi.

United Nations, 2013. World Population Policies 2011. Department of Economic and Social Affairs, New York.

United Nations Population Division, 2014. World Urbanization Prospects: The 2014 Revision [Online]. Available: http://esa.un.org/unpd/wup (accessed 21.07.14).

Wilson, B., Chakraborty, A., 2013. The environmental impacts of sprawl: emergent themes from the past decade of planning research. Sustainability 5, 3,302–3,327.

World Bank, 2009. World Development Report 2009: Reshaping Economic Geography. The World Bank, Washington, DC.

World Bank, 2013a. Global Monitoring Report 2013: Rural–Urban Dynamics and the Millennium Development Goals. The World Bank, Washington, DC.

World Bank, 2013b. World Development Indicators [Online]. Available: http://databank.worldbank.org/data/views/variableSelection/selectvariables.aspx?source=world-development-indicators (accessed 27.09.13.).

Urbanization in Africa

Deborah Potts, King's College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Urbanization in Africa has become somewhat differentiated from the processes experienced elsewhere, particularly in Asian countries. Global market forces caused some deindustrialization in sub-Saharan Africa and urban economies weakened in the last two decades of the twentieth century. Urban incomes declined in real terms, presenting major challenges to policy makers. In a range of countries, the rate of growth in levels of urbanization slowed as net in-migration rates fell as urban economic opportunities became less attractive. GDP growth rates have since strengthened largely due to natural resource-based developments. The impact of these on urbanization is very variable between and within countries.

Africa, south of the Sahara, has been profoundly rural throughout human history. In 2013, at the time of writing, this was no longer true for some countries – including Cameroon, Ghana, and South Africa – which had passed the significant tipping point whereby most people were residing in urban settlements. However, most large countries in the region are still mainly rural. Higher levels of urbanization tend to be found in the smaller nations with small populations. North African countries are different: all are mainly urban except Egypt but their economic and social history, embedded so firmly for millennia within the Mediterranean region, has been quite different from sub-Saharan Africa (see **Figure 1**).

Historical Background

Since it is such a vast continent, it is often misleading to generalize about phenomena in Africa. There is huge variation between the historical influences, which laid the basis for urban settlements, for example. North Africa has some of the world's oldest settlements, while those in central Southern Africa are some of the newest. Their locations reflected their roles in the societies which existed when they developed, and the mode of production of the moment. Consider the differences between cities such as Alexandria in Egypt, established by the Pharaohs, developed by Alexander the Great, and which served as Egypt's capital for a 1000 years until the Muslim conquest in the seventh century, and town like Lusaka, Zambia's capital, a small British colonial railway settlement that only became the capital city in 1935.

Many of the large towns of sub-Saharan Africa, including capital cities, were essentially established either to serve the needs of European colonizing powers in the late nineteenth or early twentieth century, or developed before formal colonies were established as a result of economic changes wrought by the rapid expansion of sea-borne European mercantilism from the end of the fifteenth century. Ports along the West and Central African coastlines often began in this way, from Saint Louis on the border of Senegal and Mauritania to Luanda and Lobito in Angola. In many cases, the primary stimulus for their establishment was the slave trade. By contrast, in West Africa historical, indigenous African urban settlements such as Timbuktou, Sokoto, Maiduguri, and Kumasi were in the interior, serving the commercial and political needs of various former great African empires and states such as the Bornu Empire, the Sokoto Caliphate, and the Asante Kingdom. Indigenous urban settlements also included the Yoruba agro-towns of the Oyo kingdom (Krapf-Askari, 1969). All major towns in these systems were involved to a greater or lesser extent in trans-Saharan trade with North African societies. Those in the Sahel reflect the influence of Islam, as do the cities of northern Africa. The advent of colonial administrative control and the creation of new countries with borders shaped entirely by European strategic interests at the end of the nineteenth century was often a devastating blow for these indigenous towns. Their political and economic roots were frequently transformed or destroyed either by outright conquest or the more insidious forces of economic geography as colonial administrations and companies wrenched the region's productive forces and urban systems away from their northern or inward-facing activities and refocused them on channeling exports of minerals and agricultural products to the ports and Europe. Some declined, others took on new roles in the colonial administration or commercial developments. In these cases, hybrid cities emerged with new, often racially segregated, colonial elements alongside the established indigenous towns (O'Connor, 1983): an example is Kano in northern Nigeria.

In southern and eastern Africa, with the exception of the coast, nearly all towns developed under European colonial or white settler administrations. Racial segregation was 'planned' into these towns, and African workers who migrated to them for work were discouraged from permanent settlement for decades. The further south you went, the more fiercely and effectively were such policies implemented, reflecting the racist sentiments and imperatives of the large numbers of white settlers in countries like South Africa, South-West Africa (today's Namibia), and Southern Rhodesia (now Zimbabwe). They culminated in apartheid policies in South Africa. Their primary aim was to limit the growth of a permanently settled black African urban population because that would inevitably create political forces which would undermine the white settler state. While these policies succeeded in reducing the rate of African urbanization for some decades, they could not stop it altogether. Eventually, in 1994, the inevitable came to pass and South Africa became a democratic state, but only after a bitter struggle against apartheid fought mainly in the African

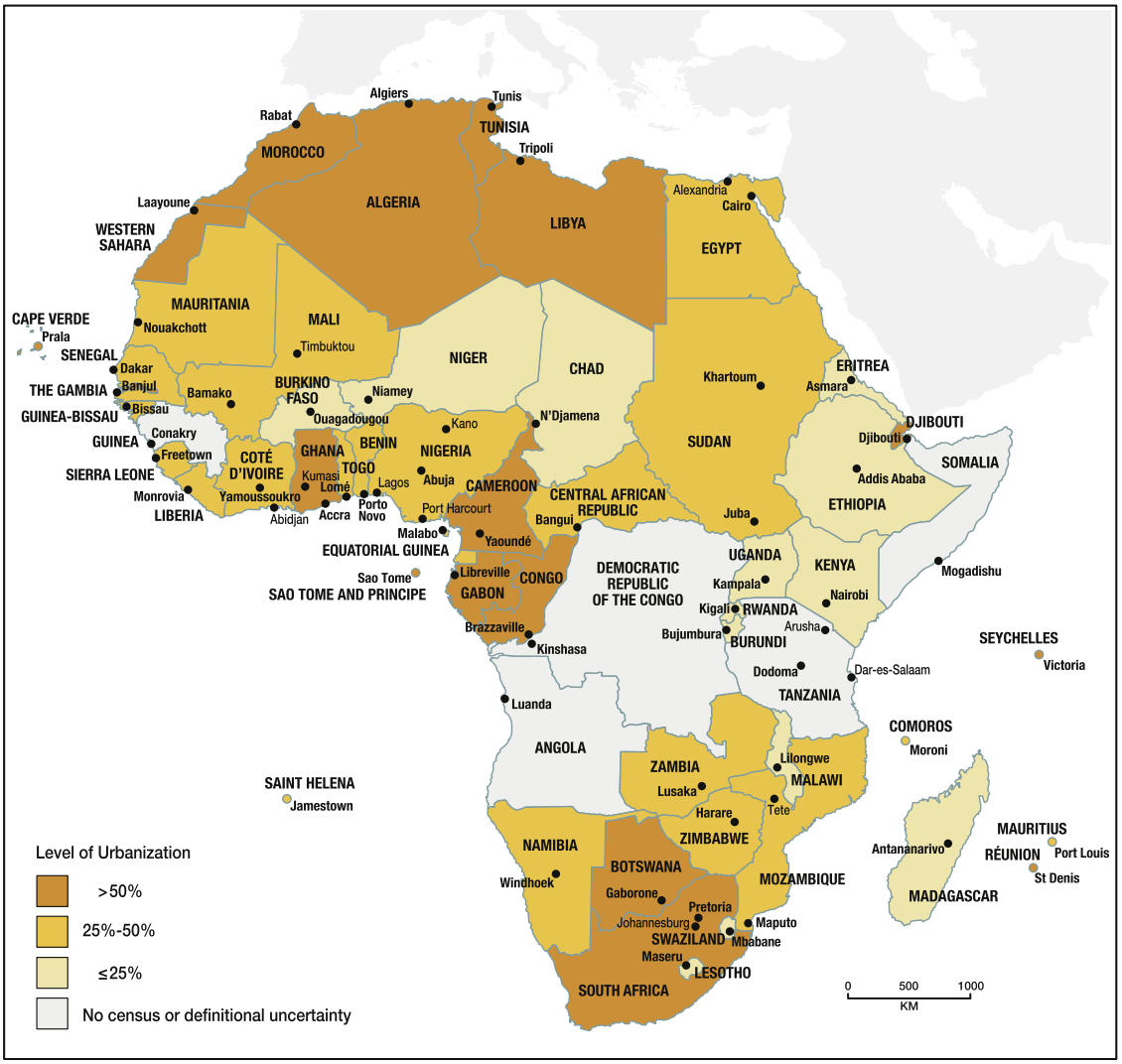

Figure 1 Urbanization levels in Africa. Notes: where possible census data have been used to calculate urbanization levels. The uncertainties about some countries' levels are discussed in the text. UN-Habitat data from the State of African Cities 2010 have been used particularly for those countries where it is known the level is over 50%. Levels for small island states are represented by the color of the dot for the capital city. Sudan is now two states but it has not had separate censuses yet, so the prepartition state is shown.

residential areas (known as townships) of the larger cities. The legacy of this racist, colonial past in the built environment of contemporary cities in this part of Africa is still evident, and problematic for the development of more inclusive towns, although time and subsequent physical expansion in many eastern and central African towns means they are essentially now expressions of postcolonial urbanization. The towns on Africa's east coast, north of South Africa, were often established prior to European contacts and had been strongly influenced in their function and location by contacts with Asian and Arabic trade, including slavery. Where they survived, the Arabic influences are also apparent in their historic centers, as in Zanzibar.

The major exception in urban historical terms in East Africa is Ethiopia. First, it was not colonized for any significant period and, second, its capital city, Addis Ababa, was the last of a series of famously 'wandering capitals' (Horvath, 1969) and not established until 1886. As it developed under a feudal political economy, its morphology is exceptional with an extraordinary mixture of house types and social classes in its densely settled central areas. The contrast between this 'inclusive' city and the divided cities of South Africa is stark.

African Urbanization in a Globalizing World Economy

African towns are now mainly shaped by the economic forces of contemporary global capitalism. From about 1980 to the global financial crash of 2008, sub-Saharan African urbanization, perhaps more so than any other part of the world, was subject to the ideological fashion for neoliberal capitalism, with increasingly unregulated free market forces determining urban economies and the livelihoods of millions of urban residents. This was a significant departure from the

modernizing policies of the post independence era (usually from the 1960s), which in most African countries had sought, via government policies, to diversify national economies away from what was felt to be the dangerous and restrictive over-reliance on commodity exports inherited from European colonialism. Those policies had fostered infant industries and investments in woefully limited social and economic infrastructure, much of which created strong urbanizing tendencies. In this way, for a brief period, postcolonial Africa attempted the type of development strategy which typified the postwar period for developing countries. However, given the lateness of independence in the region, most countries had little time, often a decade or less, before the first oil crisis in the early 1970s changed the global economic context drastically. The second oil crisis in 1979 meant many African countries' debt became intractable at the same time as the profound western ideological shift toward neoliberal economic policies began to take effect.

As a result, to a greater extent than elsewhere, the main economic influences on sub-Saharan Africa's postcolonial urbanization have been increasingly unregulated market forces and a sharply reduced public sector role. Most countries in 1980 still had very small economies and, once heavily indebted, lacked the economic weight or political clout to temper the economic 'medicine' of neoliberal financial institutions, unless they were perceived to have some strategic significance in Cold War terms, such as Zaire (now the Democratic Republic of the Congo (DRC)). Donor conditionality ensured that most countries increasingly conformed to the tenets of neoliberal market reforms, with open economies, liberalized trade, uncontrolled currencies, and cutbacks in government expenditure. The impact on urban economies and the livelihoods of urban residents was more dramatic than elsewhere in the world largely because of three factors: first, most countries were still very poor and urban poverty had already begun to rise in the 1970s as governments became more indebted; second, welfare nets for urban residents scarcely existed; and third, as rules of comparative advantage came to determine what African countries could produce, many urban-located large-scale enterprises proved uncompetitive and national economies were pushed back toward dependence on primary commodity exports. Formal sector employment dwindled relative to the size of the general workforce in African towns and cities, and towns across sub-Saharan Africa informalized to a remarkable degree.

The Links between African Urban Economies and Livelihoods and the Built Environment of the City

The nature of economic activities shapes cities in many ways. It sets the parameters of urban workers' incomes, the city tax base, and the practicability of different policies to provide essential services such as housing, water, sanitation, and transport. A major constraint for African towns is that many struggle to be major forces of production within their countries, instead relying on trade and services for their economic base. Of fundamental importance has been Africa's inability to industrialize. In comparison to Asia, African towns have fared extremely poorly over the past few decades in the creation of steady, wage-paying jobs for unskilled and semiskilled workers in industrial enterprises. Millions of such jobs have been created in Asia, from South Korea to Cambodia, from Vietnam to China. Sub-Saharan African towns have been unable to compete for these jobs for a range of reasons. Their workforces tend to be considerably less well-educated. Urban infrastructure tends to be very poor. Electricity supplies are often irregular. National governments, constrained by donor conditionality and prevailing neoliberal ideology, have found it difficult to intervene and foster industrialization via tariff barriers, or by the myriad of other policies practiced in other countries to encourage manufacturing industry beyond the 'natural' limits of comparative advantage, such as making artificially cheap loans, electricity, or water available. The fact that African countries tend to remain among the world's poorest also constrains these sorts of interventions. Furthermore, as the center of global economic gravity has shifted toward Asia, simple geography enhances the competitiveness of countries in that region for any outsourcing of the cheaper end of supply chain activities. Thus cities in countries like Cambodia have, so far, benefited from industrial investments looking for cheaper labor forces than can be found in, say, China, rather than African cities. There has been some Chinese manufacturing investment in Africa, particularly in countries with large markets like Ethiopia and Egypt, and cheap labor is one factor (Xiaochen, 2012) but the numbers of jobs overall, as yet, remain extremely limited by comparison.

African urban economies have often actually deindustrialized. The structural adjustment programmes (SAPs) enforced by the World Bank in the 1980s and 1990s allowed imports of cheaper products, which led to many industries in African towns closing down or reducing output. In Zimbabwe, the manufacturing industry accounted for about a quarter of GDP at the end of the 1980s. SAPs were introduced in 1991 and by 1998 the share had fallen to 17% (Potts, 2010). The example of Nigeria is detailed in **Box 1**. Light manufacturing exports from sub-Saharan Africa comprised only 0.9% of the world total in 2008, a decline from 1.2% in 1980, and the African Competitiveness Report for 2011 stated that the conditions for heavy manufacturing are simply not met in most low- and middle-income sub-Saharan African countries. It is rather sobering to realize that in 2008 Vietnam exported more light manufacturing products than all of sub-Saharan Africa, including South Africa (World Economic Forum et al., 2011).

The problem for African urban livelihoods is that while manufacturing industries are far from the only source of formal employment and prosperity in cities in developing countries, they create very positive multiplier effects and growth in many other sectors such as formal financial and commercial services is greatly enhanced. The regular wages from such jobs create virtuous circles, stimulating the housing, retail, and leisure sectors. These incomes can also make it possible to provide crucial services and recover their costs. Yet in many African urban settlements, this is a challenge simply because even basic infrastructure and housing, provided by formal, legal channels, is unaffordable to the majority.

Contemporary African urbanization needs to be understood in relation to these underlying economic constraints. It is generally claimed that in most sub-Saharan African towns most people work in informal sector jobs, and that the majority of

Box 1

It is estimated that 80% of Nigeria's textile factories have closed (Green and Macnamara, 2008; citing Alden) and 250 000 urban textile workers have lost their jobs because of foreign imports – equivalent to just under a quarter of the manufacturing workforce in the late 2000s. In the north of the country, textile manufacturing was the only urban-based industrial sector of any significance, but much has now closed, exacerbating already very high levels of urban poverty (Burgis, 2009). The southeast of Nigeria has a far more broadly based and dynamic tradition of domestic manufacturing. This includes the town of Aba, famed for its local small- and medium-scale (and sometimes handmade) shoes and clothes industries. Many of these used machinery imported during the 1970s and 1980s when the Nigerian currency was strong due to the oil boom. However, in recent years these have also been very badly affected by the 'flood of cheap imports that has accompanied China's strengthening ties with the continent' (Mahtani, 2007). It is estimated that by 2007 half or more of the market for Aba's shoes had been captured by East Asian imports. Similar problems led to many industries closing in nearby Nnewi, which was known for its motor parts trade (Financial Times, 2007). Attempts by the Nigerian government to reintroduce protection for domestic textile manufacturing have been widely circumvented by importers who recertificate Chinese textiles as locally made (Burgis, 2009). On the other hand, clothes manufacturers who did depend on certain types of high-quality imported fabrics have suffered as these have become difficult to obtain, except through smuggling (Mahtani, 2007). Nigeria has, like many other African countries, tried to encourage foreign investment in manufacturing by setting up Free Trade Zones. There is one in Calabar, also in Nigeria's southeast. However, domestic manufacturers outside the zone argue that this also causes them to suffer from unfair competition from within their own country (Beattie, 2007).

new jobs are found in this sector too. As reliable and comparable employment data in different towns may be hard to find, it is difficult to be precise. For most African urban residents, however, their livelihoods, the places where they live, and the things that they buy can be characterized as informal. While there is much disagreement about where formality ends and informality begins, it is usually agreed that typical urban informal jobs include self-employment as petty traders, selling everything from vegetables to cheap clothes (now often imported from Asia) to charcoal to drinking water in sachets, as well as small-scale producers (e.g., of furniture or metal products, often recycled), provision of cooked food and local beer production and sale, and a host of services such as hairdressing, motorbike taxis, house construction, car repairs, tutoring, or tailoring. Other typical features are that the activities do not conform to official regulations about, for example, business premises or health and safety conditions, rarely maintain accounts, and pay no income tax. Family members may work together in some enterprises, although some may not get paid. The production and incomes generated by informal sector work are also very imperfectly captured by government statistics, if at all.

Urban informality has its pros and cons. In much of Africa, economic safety nets such as pensions or child benefits are nonexistent or very inadequate – although some countries in southern Africa, and Kenya and Senegal have been gradually introducing these. Therefore, most adults must work if they are to survive and, if the formal sector cannot provide jobs, self-employment is the answer. The informal sector has a great capacity for 'involution' whereby many people produce and/or trade items that could be more 'efficiently' produced and distributed by far fewer in the formal sector (Armstrong and Mcgee, 1968). This ability to absorb the urban workforce in Africa when formal job growth is outpaced by population growth is beneficial. For decades, many reports on urban Africa argued that urban unemployment rates were very high but these did not properly account for the informal sector. It is now recognized by the African Development Bank (2012) that, in contrast to previous assumptions, poorer African countries tend to have *lower* urban unemployment rates. Another advantage is that 'informal' goods are sometimes cheaper than formal sector products because labor is cheap, or the products are lower quality or use recycled materials. Even when informal traders' prices are more expensive than formal sector outlets, if there is a market for the goods evidently the producers and sellers provide for a need – classic examples being the subdivision of luxury items into units that poor people can sometimes afford (half a loaf; one cigarette) or are bringing goods (e.g., groceries brought from wholesale formal outlets) to locations where the formal sector cannot or will not reach, probably because profit rates would be too low or formal outlets are not facilitated by the state.

The informal sector has negative aspects too. It is characterized by very low incomes, which are also often very irregular. This makes people's lives very difficult and vulnerable. The lower unemployment rates in poor cities are associated with a desperate survivalist need to work and earn even tiny incomes in highly insecure social and economic environments. In these circumstances, low unemployment can be an indicator of poverty rather than opportunity. Research for the African Development Bank (2012) found that "the unemployed are less likely to suffer from poverty than many self-employed or underemployed" (p. 102), the reason being that unemployment may only be possible if an individual has access to some other source of income or support. Furthermore, African governments, both at the national and city scale, have been highly ambivalent in their attitudes toward the informal sector. They sway back and forth between regarding it as deeply negative and a hindrance to development, and seeing it as a partial solution to unemployment and a starting point for new entrepreneurs. Since the 1970s, many official policies include at least some element of supporting SMMEs (Small, Medium, and Micro-Enterprises), and this usually embraces the more profitable informal enterprises. Partly, and understandably, the incentive for governments is to bring such economic activities within the official tax net, thus strengthening the city's revenues. However, the intractable obstacle for this worthy approach is that nearly all surveys of African urban informality find that the majority of workers are self-employed, usually in trade. Many are women who may combine their flexible work patterns with household responsibilities, including child care. Their characteristics do not fit the SMME projects.

Official attitudes unfortunately frequently veer off into the punitive, sometimes in drastic ways. City laws (often inherited from colonial governments) often deem the 'informal' 'illegal.' For most of the time informality is tolerated, otherwise African cities would not work (in both senses of the word). And government officials and politicians are often deriving some of their own incomes from the so-called informal activities. For example, the renting of rooms in one of Nairobi's large slums,

Kibera, provides large profits for some people in the government. Nonetheless, urban planners are rarely in favor of the urban informal sector because, almost by definition, it is unplanned and flouts regulations. The sector's 'image' is also problematic. Governments often want their capital cities, in particular, to appear 'modern' and 'planned.' Downtown traders in particular undermine that image and are periodically discouraged or even banned. The police often harass informal traders, sometimes with a view to obtaining bribes, or steal goods under the pretence that fines are being levied for breaking laws. An infamous example of a government campaign against urban informality occurred in Zimbabwe in 2005, when virtually every informal worker and enterprise was banned from operating, and all the unplanned houses were demolished in Operation Murambatsvina ('Drive out the Trash'). All this adds to the vulnerability of informal sector workers, and undermines their already low incomes.

The typically low incomes also mean many African urban residents can neither afford reasonable urban housing nor 'formal' sector provision of essential services like water, sanitation, and electricity. At another level, it also means that there is a significant proportion of Africa's urban residents for whom it is almost impossible to provide these services via the formal sector on a cost recovery basis, let alone through any large-scale profit-making private venture. As a result, extremely high proportions of urban dwellers live in unplanned housing, with very poor services. Sanitation conditions can be very dangerous for health. While everyone does have access to water or they could not survive, it may have to be purchased in small quantities from local vendors, or obtained illegally by tapping into piped supplies, or taken from local wells, which may be contaminated. If electricity is available it is often too expensive for ordinary urban residents: again tapping illegally into nearby power supplies may solve the issue although this is dangerous. Many use charcoal or kerosene.

Highly innovative, often community-based, approaches can make very positive improvements in these areas (Satterthwaite and D'Cruz, 2005). For example, the now global organization, Shack Dwellers International, has worked with many low-income residents in African cities to help them plan how their housing conditions could be improved and then negotiate with city authorities to implement their solutions. City officials may welcome practical solutions to what sometimes seem impossible urban conditions. However, if city politicians are highly authoritarian and antagonistic toward informality, this may not work as the approach does depend on a degree of flexibility about urban planning rules, which must often be relaxed, and a reasonably benign government. Negotiating for some form of tenure security so that residents of unplanned settlements do not exist under threat of eviction is often a starting point for other improvements to occur in urban living conditions. One feature of many African cities, except for those furthest south where major towns were mainly built on land alienated by white settlers from indigenous people, is that large numbers of low-income residential areas occupy land which was obtained via various forms of customary tenure. This may make their residents more secure than those squatting on private or public land held under capitalist property rights. It also tends to be much cheaper. However, unless such settlements are recognized by the local government, they may find it difficult to obtain urban services. What forms of tenure are acceptable and effective for low-income residential areas is one of the most significant debates about African urbanization. UN-Habitat, the global agency, which focuses on human settlements and which is based in Nairobi in Kenya, took this as one of the main themes of its 2010 *State of African Cities* report (UN-Habitat, 2010).

Urban Population Growth and Urbanization Levels

African countries tend to have high population growth rates because birth rates remain high. Even in southern and North African countries, where they are lowest, the rates tend to be higher than in most of South Asia and much higher than anywhere else in the Global South. In the rest of Africa they are much higher still, usually over 30 per 1000 and sometimes more than 40. Natural increase rates also tend to be high in African towns – their rate of growth simply from annual increments to their populations from births minus deaths is very high relative to other urban areas in the world. For these reasons, African urban populations tend to be growing faster than those in Asia or Latin America. This is often taken to mean that Africa is the world's fastest urbanizing region, but this is a mistake. The rate at which a country urbanizes – how fast its population is shifting from being rural to urban – depends not on absolute growth rates but the gap between urban and national growth rates. Current data suggest that Asia has been urbanizing faster in the past 10–20 years.

The reason is that Asian urban economies have been stronger than those in Africa and this has encouraged more long-term in-migration to Asian towns. The rate at which African populations were shifting from being rural to urban slowed in many countries toward the end of the twentieth century due to the weakening of urban economies. In the 1960s and 1970s, urban growth rates had often been phenomenal: many large towns like Nairobi, Lusaka, and Abidjan grew at annual rates of over 10% for about a decade, largely due to rural–urban migration. However, the falls in urban incomes in real terms during SAPs squeezed the gap between rural and urban incomes, which is a key driver of rural–urban migration. This phenomenon was charted by economists for a range of African countries in a book entitled *Africa Misunderstood, or whatever happened to the rural-urban income gap*? (Jamal and Weeks, 1993). As the gap reduced, net in-migration to urban areas in some countries fell. This did not necessarily mean that rural–urban migration rates fell steeply, though there is evidence that this happened in some countries. Many migrants still came to towns for cash wages and in the hope of better incomes. However, urban livelihoods had become very vulnerable, and lack of work and low incomes meant that many people found they could not afford to stay in towns and so urban–rural migration rates rose. In sub-Saharan Africa, rural–urban migrants may still maintain access to land through indigenous tenure systems and this facilitates circular migration between rural and urban areas. The pattern had often been encouraged or, in southern Africa, enforced by colonial regimes. It continued in the postcolonial years although with many adaptations: families came to join migrant household heads in towns and length of stay in town increased

significantly. Sex ratios in towns swiftly normalized in the 1960s and 1970s. Nonetheless, rural–urban linkages were still strong and older migrants 'retiring' back to rural areas remained quite common. When the rural–urban income gap narrowed sharply, these linkages and migration patterns strengthened again, and the annual net addition to urban populations from migration fell.

For Africa as a whole, it has been estimated that the contribution of net in-migration to urban growth was about 40% in the 1960s and 1970s, but fell to only 25% in the 1980s (Chen et al., 1998). In Tanzania in 2001–02, the contribution has been calculated to have been only 17% with most growth coming from natural increase and the reclassification of rural populations. Circular migration was significant – rural–urban and urban–rural migration flows totaled about 389 000 people – but net in-migration to urban centers was only around 44 000, 0.6% of the urban population. In addition, about 190 000 moved between urban centers (Muzzini and Lindeboom, 2008).

The first country where these impacts were recorded in urbanization levels was Zambia. In the 1970s and 1980s, it experienced a particularly negative set of economic and political forces: the oil price rises leading to debt; a massive fall in the prices of its key export, copper; and violent conflict in the region surrounding it, which proved especially costly since it was landlocked. In 1990, Zambia's census showed that it had counterurbanized in the 1980s: the urban share of the population had fallen from 40% to 38%. Net migration rates shifted to being urban–rural and several mainly rural provinces experienced strong in-migration. Since then a few other countries also experienced periods of counterurbanization: Cote d'Ivoire, Mali, the Central African Republic, and Zimbabwe. Zambia continued to counterurbanize in the 1990s but reurbanized from 2000 to 2010, although the urban share in 2010 was still less than it had been 30 years earlier. Mali's urbanization levels have also increased again. More commonly, large sub-Saharan African countries experienced a marked slowing in the rate of urbanization. Many censuses toward the end of the 1990s and in the twenty-first century found that the increase in the urban proportion had been under 2%, and in some cases 1% or less, over intercensal periods, which were often much longer than 10 years. Benin, Ethiopia, Malawi, Mauritania, Mozambique, Niger, Senegal, Sudan, Togo, and Uganda are in this category. Analysis combining the limited available data from Nigeria's 2006 census and from Africapolis – a French research project using aerial photographs to estimate town populations – suggests that Nigeria should also be included (Potts, 2012). Africapolis reports even argue that urbanization levels across much of West Africa have been overestimated by UN agencies and the World Bank (Africapolis Team, 2008). This is largely due to long gaps between censuses, which had led to urban populations being estimated from projections based on earlier, more rapid, growth. Two of sub-Saharan Africa's largest countries, the Democratic Republic of the Congo and Angola, have not had any census for decades so their urban populations can still only be estimated. The combination of this past tendency to overestimate urbanization levels and the recent census data indicating lower levels can cause confusion. The data set on urbanization levels held by UN-Habitat, the main global agency focused on urban settlements, in 2013 suggested that urbanization levels in 17 African countries had declined between 2000 and 2010. At the same time, the same agency strongly promotes the view that Africa is urbanizing rapidly. The contradictions arise because the overgenerous estimates for 2000 have not been corrected. Yet, even the downwardly revised 2010 estimates are too high for some countries according to their recent censuses; for example, Chad, Madagascar, Liberia, Malawi, Mozambique, and Zimbabwe.

In addition to the slowing of urbanization attributable to changes in migration patterns to large towns, for some African countries reported urbanization levels may also overestimate how many people are living in truly urban settlements and give a false impression of the extent to which a country's workforce has shifted away from rural economic activities. Urbanization levels are often used as a proxy for economic development on the assumption that they reflect important sectoral changes in the constitution of a country's GDP and employment. Yet many African countries not only define rather small settlements as urban, but also do this without regard to whether their residents are still working in agriculture. Were the urban settlement definitions used in India, which factor in economic activity patterns, applied in such countries, their urbanization levels would fall sharply.

This definitional issue is of significance for two of East Africa's major countries, Kenya and Tanzania. Both are portrayed by their governments and agencies like UN-Habitat and the World Bank as urbanizing rapidly. The capital cities of Nairobi and Dar es Salaam have grown fast for decades. Nonetheless in both the level of real urbanization is lower than generally claimed. In Tanzania, there is no clear definition of what is urban. The census figures may include not only the main town, but also periurban wards, which are designated 'urban,' and even mixed rural–urban wards where much or part of the population is definitely rural. There is no doubt that this led to a general overestimation of real urban population growth (Holm, 1992). Furthermore, one third of those employed in the urban population of Tanzania (including large towns) was recorded as working in agricultural activities in 2002. In small towns and townships, the vast majority (67%) were working in agriculture (Muzzini and Lindeboom, 2008). For these reasons, it seems certain that a functional urbanization level for the country in 2002 was less than the 23% recorded in the census. Similarly, in Kenya's recent census of the urban population has included both 'core' urban areas and 'periurban' areas. Often the latter include many rural people and for some urban areas, it is evident that the vast majority of the population are farmers. Kenya's 2009 census recorded its urbanization level as 30% but the core urban population was only 23% of the national total.

Africa continues to urbanize but the various factors discussed show that the rate at which this is occurring has slowed and that the level of urbanization is less high than often assumed. The primary reasons are economic. Distinguishing between rural and urban populations is important if urbanization is being used as an indicator of certain types of economic development and change. However, from the perspective of evaluating the livelihoods of most people in sub-Saharan Africa, the distinctions may be less useful. One problematic aspect is that in many urban areas, including some very large cities like

Kinshasa, the capital of the Democratic Republic of the Congo, many residents who work in urban-based activities also derive some or occasionally most of their food from urban agriculture. This was encouraged by the fall in real urban incomes from the 1980s, which led to poor households spending very high proportions of their incomes on food. In addition, it is common for low-income urban households to maintain linkages with rural kin and vice versa. These linkages include social and economic aspects such as remittances of cash and goods, food produced in rural areas being brought to towns, and visits and circular migration between households at both ends of the spectrum. The composition of connected households can shift over time as rural versus urban opportunities and needs change. In this way, the rural–urban divide is blurred. This adds a further layer to discussions about African urbanization.

Africa's Urban Prospects

From about 2003, many sub-Saharan African economies experienced a surge in growth rates of GDP after a period of up to two decades of stagnation and decline. Some countries, like Angola, have been among the world's fastest growing economies. The primary reason has been the growth in demand for a range of natural resources from Asian countries, particularly China, to fuel their industrialization and infrastructural development. Africa is rich in natural resources and there have also been many new mineral and energy finds. Offshore oil and gas finds of global significance are shaping new economic directions in both West and East African countries, which previously were not involved in these sectors, such as Ghana, Tanzania, and Mozambique. High gold, copper, and other metal prices encouraged new large-scale corporate investments in many countries, including these three and Zambia. The proposed development of extraordinary iron ore reserves in and near Guinea – at least 50 billion tonnes – by international investors and the IFC of the World Bank has been promoted as possibly SSA's 'largest industrial investment project ever' (AIOG, 2012). The ending of conflicts in southern African countries as white minority rule and the Cold War ended has also helped economies to reestablish growth.

These economic changes are having urbanizing effects. For example, ports are creating new jobs and attracting new migrants, as with Tema and Sekondi-Takoradi in Ghana, both of which grew much faster from 2000 to 2010 than the capital, Accra. In Nigeria, oil cities like Port Harcourt and Warri in the southeast have similarly experienced higher population growth than most of Nigeria's other towns, including Lagos, the commercial capital. Kumasi, in Ghana's interior, has been growing at about 5.5% annually between 1984 and 2010 and its population now rivals that of Accra, the capital city. It has always had a key marketing role in the local economy and functions as a central place for a hinterland with strong exports of cocoa, gold, and hardwood. Arusha in Tanzania also serves a productive commercial agricultural hinterland and has similarly experienced very high rates of population growth, exceeding those of the capital city of Dar es Salaam in the 1990s. In addition, it has the offices of the regional grouping, the East African Community, as well as being a central node for tourists, often flying into Nairobi from Europe, since it is well placed for access to some of the world's best wildlife parks. Small-scale artisanal gold mining in Tanzania has had vigorous urbanizing impacts in small towns in Tanzania (Bryceson and Jønsson, 2010). Tete, in Mozambique, was hard hit by that country's civil war in the 1980s but became one of the fastest-growing large towns in the country with new foreign investment in its huge coal reserves.

The evidence suggests that, in sub-Saharan Africa, natural resource-based economic stimuli are increasingly the cause of urban population growth where it is significantly in excess of that derived from natural increase or the reclassification of small settlements. This is logical since this is the main source of national economic growth. This is a direct contrast to Asian urbanization, which is being driven by significant industrialization with millions of jobs created within urban settlements. In Africa, many of the new jobs created by recent economic changes are not urban-located, such as those in offshore oil and gas, timber, and agricultural developments. Large-scale mining of commodities like copper and coal, on the other hand, is more often directly associated with the creation of major towns, as in Zambia and the DRC, for example. The largest conurbation in Africa in Gauteng Province in South Africa, centered on Johannesburg, has its origins in gold mining. Nonetheless, mineral developments can be associated with the creation of enclave economies, with relatively few jobs in very concentrated locations, and a lack of forward and backward economic linkages stimulating other urban employment-creating enterprises. Africa has experienced this before during the colonial and early postcolonial periods. The patterns are illustrated by Zambia's experience: it began to reurbanize after 2000, after 20 years of counterurbanization, as copper prices surged and the output of its copper mines in the Copperbelt towns increased. Yet four of the five largest of these towns grew more slowly than the national population from 2000 to 2010; the most vigorous growth was experienced in the capital city, Lusaka, which grew at nearly 5% per year.

The future of African urban economies and internal migration trends will partly depend on the extent to which governments are able and willing to translate increased income from natural resource production into diversified and specialized urban-based formal employment opportunities. The multiplier effects of such employment would foster further sustainable urbanization with the development of more viable urban livelihoods. Yet the history of shocking inequality in incomes and the limited spatial spread of the benefits of GDP generated in oil-rich countries like Angola, Nigeria, Equatorial Guinea, and Congo does not provide a positive precedent. African governments have also struggled to negotiate reasonable deals with the world's extremely powerful transnational mining companies; in some cases, they have been poorly advised by international financial institutions and have reaped far less from mineral investments than countries like Chile (Larmer and Fraser, 2010). If strong demand for natural commodities continues, however, their bargaining power increases and there are signs that African states are becoming more assertive about the nature of mineral contracts. A good example here is Botswana, which has experienced very positive economic growth for decades driven by its valuable diamonds, which are mined in collaboration with De Beers. The government has managed to derive reasonable levels of revenue from

the diamonds, although this is partly because the sheer scale of its resources gave it a strong bargaining hand. Nonetheless, it is now more than 50% urbanized with some of the urbanization occurring as the occupational structure of its large, traditional agro-villages shifted from agriculture-based to a range of more diversified and specialized urban-type work as the national economy grew and developed. The government has made very specific efforts to diversify, in very difficult circumstances. The national population is small, only 1.9 million in 2011, and the regional economic powerhouse of South Africa across the border provides tough competition. Nonetheless, Botswana counts among the most successful exporters of light manufactures in SSA (World Economic Forum et al., 2011). Small island states like Mauritius have been more successful in industrializing and urbanizing, and Lesotho has developed a textile sector. But these countries have tiny populations and are not typical of sub-Saharan African nations; furthermore, their seeming success is often due to special trade treatment by the United States and EU, such as the USA's African Growth and Opportunity Act (AGOA), without which their manufactured exports would be uncompetitive.

The classic linear development models, which assumed that economic development, urbanization, and industrialization progressed together, are challenged by Africa's experience. Urbanization has slowed, stalled, and even occasionally reversed at times and the same is true of industrialization. The natural resource-driven economic growth pattern may not dominate in the long term, but it does impose constraints on urban livelihoods and economies. Where vigorous urbanization is occurring currently, it tends to be associated less with production based in towns and more with trade and services linked to primary commodity sectors. The extent to which this can be translated into better and more dependable incomes for Africa's urban dwellers is still uncertain.

See also: Africa's Environment: A Stressed Biogeographical and Cultural Landscape; Development Geographies; Development and Urbanization; Urban Geography; Urban System in Geography; Urbanization.

Bibliography

African Development Bank, 2012. African Economic Outlook 2012. African Development Bank.

Africapolis Team, 2008. Urbanization Trends 1950–2020: A Geo-statistical Approach. Analytical Report. Africapolis, AFD-SEDET. http://www.afd.fr/jahia/webdav/site/afd/users/administrateur/public/publications/BT/Africapolis_Final-Report_EN.pdf.

AIOG [African Iron Ore Group] 31 July 2012. Iron ore infrastructure to stimulate Africa's growth. Advertisement in Financial Times, London.

Armstrong, W., Mcgee, T., 1968. Revolutionary change and the Third World city: a theory of urban involution. Civilization 18, 353–377.

Beattie, A., 12 July 2007. Free Trade Zones: Resort Is Designed to Keep Business at home. Financial Times, London.

Bryceson, D.F., Jønsson, J., 2010. Gold digging careers in rural Africa: small-scale miners' livelihood choices. World Development 38 (3), 379–392.

Burgis, T., 20 July 2009. The North: Abandoned in an 'Incubator of Poverty'. Financial Times.

Chen, N., Valente, P., Zlotnik, H., 1998. What do we know about recent trends in urbanization? In: Bilsborrow, M. (Ed.), Migration, Urbanization and Development: New Directions and Issues. United Nations Population Fund (UNFPA) and Kluwer Academic Publishers, Massachusetts, pp. 357–390.

Financial Times, 12 July 2007. Dying to Keep Up with Competition from China. Nigeria.

Green, M., Macnamara, W., 24 January 2008. Pros and Cons of Importing. Financial Times, London.

Holm, M., 1992. Survival strategies of migrants to Makambako: an intermediate town in Tanzania. In: Baker, J., Pedersen, P.O. (Eds.), The Rural-Urban Interface in Africa: Expansion and Adaptation. Scandinavian Institute of African Studies, Seminar Proceedings No. 27, pp. 223–238.

Horvath, R.J., 1969. The wandering capitals of Ethiopia. Journal of African History 10 (2), 205–219.

Jamal, V., Weeks, J., 1993. Africa Misunderstood: Or Whatever Happened to the Rural-Urban Gap? Macmillan, Basingstoke.

Krapf-Askari, E., 1969. Yoruba Towns and Cities: An Enquiry into the Nature of Urban Social Phenomena. Clarendon Press, Oxford.

Larmer, M., Fraser, A. (Eds.), 2010. Zambia, Mining, and Neoliberalism: Boom and Bust on the Globalized Copperbelt. Palgrave Macmillan, New York.

Mahtani, D., 29 March 2007. Nigeria's Leaders Fail to Nurture Potential of Region's Industrial Powerhouse. Financial Times, London.

Muzzini, E., Lindeboom, W., 2008. The Urban Transition in Tanzania: Building the Empirical Base for Policy Dialogue. World Bank, Washington, DC.

O'Connor, A., 1983. The African City. Hutchinson, London.

Potts, D., 2010. Circular Migration in Zimbabwe and Contemporary Sub-Saharan Africa. James Currey, Oxford.

Potts, D., 2012. Challenging the myths of urban dynamics in Sub-Saharan Africa: the evidence from Nigeria. World Development 40 (7), 1382–1393.

Satterthwaite, D., D'Cruz, C., 2005. Building Homes, Changing Official Approaches: The Work of Urban Poor Organizations and Their Federations and Their Contributions to Meeting the Millennium Development Goals in Urban Areas. Human Settlements Working Paper. IIED, London.

UN-Habitat, 2010. The State of African Cities 2010: Governance, Inequalities and Urban Land Markets. United Nations Human Settlements Programme (UN-HABITAT), Nairobi.

World Economic Forum, the World Bank, the African Development Bank, 2011. The Africa Competitiveness Report. World Economic Forum.

Xiaochen, F., 2012. Foreign Direct Investment from China to Ethiopia. Masters Dissertation. John F. Kennedy School of Government, Harvard University.

Urbanization in China

Hyun Bang Shin, The London School of Economics and Political Science, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article on China's urbanization examines the main drivers behind China's urban transformation, its impact on the society, and how the impact is unevenly spread across populations and geography. It begins with a brief historical summary of China's urbanization focusing on the post-Liberation period, that is 1949 onward, and moves on to discuss some of the major factors that have contributed to people's migration and urban physical expansion. Three key characteristics are paid particular attention: (1) industrial production and uneven development; (2) land-based accumulation; (3) local state building and state entrepreneurialism. The article further examines some of the important sociospatial outcomes and challenges that China's urbanization poses.

Introduction

Viewed from a demographic perspective, urbanization refers to the growth of the urban share of the national population. This growth can be achieved by a number of ways: the natural growth of the existing urban population, the migration of rural residents to urban areas, the physical expansion of extant urban administrative boundaries, and the emergence of new towns as beachheads of further urban growth in neighboring areas. For China, while the fertility control through the implementation of the One Child Policy since 1979 has been a constraint upon urban population growth to some extent, migration as well as the territorial expansion of urban areas has substantially increased the rate of urbanization in recent decades (Song and Timberlake, 1996; Hsing, 2010). While China's urbanization rate reached just under 20% in the late 1970s at the time of embarking on its economic reforms, China's 2010 Population Census revealed that almost 50% of its national population was now classified as urban. The subsequent year 2011 was the landmark point for China, seeing the majority of the national population marked as urban for the first time in its history. The United Nations Population Division has predicted that China will add another 10% by 2020 (United Nations Population Division, 2012: Table A.2). China's new leadership from 2012 has also expressed explicitly their eagerness to make sure China's urbanization reaches 60% by 2020, which will help boost the country's consumption basis to create domestic demand and therefore sustain economic development in the next decade or so (Shanghai Daily, 2013; see also Chan, 2014). While the reliability of these official statistics may be subject to scrutiny and proper interpretation (Brenner and Schmid, 2014), it is important to understand what the state aspiration of urbanization aims to achieve.

The recent emphasis on 'urban' is not a surprise. Increasingly, an urban way of life has been regarded as 'modern' and 'civilized' while the rural has been associated with backwardness (Anagnost, 2004). The theme of the Shanghai World Expo 2010 written in English was 'Better City, Better Life.' To many spectators and visitors from outside China, the English slogan was reflecting a common sense understanding. Improved urban conditions including infrastructure, services, and environment would result in a better life for citizens. However, the Chinese version of the Shanghai World Expo meant entirely something else. In Chinese, the slogan reads '*Chengshi, Rang Shenghuo Geng Meihao*,' which can be literally translated as 'City makes life happier.' The emphasis was no longer on a 'better city' but on the 'city' itself. Constructing "urbanization as a political and ideological project" (see Shin, 2014a), the Chinese Party State was sending out a clear signal to its domestic population that an urban way of living was the way forward for the nation.

This article on China's urbanization examines the main drivers behind China's urban transformation, its impact on the society, and how the impact is unevenly spread across populations and geography. It begins with a brief historical summary of China's urbanization focusing on the post-Liberation period, that is 1949 onward, and moves on to discuss some of the major factors that have contributed to people's migration and urban physical expansion. Industrial production and uneven development, land-based accumulation, and local state building are the three theses that are given particular attention.

Urbanization Trend

China's rates of urbanization have been very much influenced and shaped by the national policies that alternated between strict control and relaxation (Song and Timberlake, 1996). Since the Liberation in 1949, China has gone through dramatic changes in terms of urbanization and industrialization. The 10-year period between 1950 and 1960 was devoted to raising the industrial output through the implementation of the first Five-Year Plan (1953–57) and the political campaign known as the Great Leap Forward, which resulted in high rates of urban in-migration. The following 5 years (1961–65) were marked by the sharp decline of urban population, which resulted from the return of migrants to their origin places in order to supplement labor shortage in rural areas due to the nationwide famine at the time (Chan, 1994). It was this period that witnessed the strict control of the grain ration distribution in cities and the tightening of population move through the implementation of the household registration system known as *hukou* in Chinese. The 1960s and the 1970s also saw a major sociopolitical

uprooting due to the reign of the Cultural Revolution that lasted almost a decade. Some of the most notorious campaigns and policies exercised in this period were resettlement campaigns, known as '*shangshan xiaxiang*' (up to the mountain and down to the villages) and '*xiafang*' (sending down). These sent hundreds of thousands of urban cadres, youths, and intellectuals to the countryside, and the number of these youths reached about 20 million (Naughton, 1995; Chan, 1994). A famous statement by Mao Zedong may express one of his main concerns during this period: '*Chengshi taidale, buhao*' (It is no good if cities are too big). Based on the apparent restraints of urban growth in the 1960s and 1970s, Barry Naughton states that economic planners in China "consistently made choices about the role and function of cities that tended to undervalue the contribution of cities" (Naughton, 1995: p. 61). On the one hand, the stagnant growth of urban areas during Maoist era has been evaluated by many scholars as a result of strong antiurban bias. However, it must be reminded that rather than ideological antiurban and prorural bias of the Chinese leadership, it was the strong motivation to drive industrialization for maximum accumulation that resulted in the urban retreat during the planned economy period (Kirkby, 1985).

China's economic reform since the death of Mao Zedong in 1976 has brought profound changes to urban life in China. Cities and towns, which had undergone a long period of hibernation during the period of Cultural Revolution that lasted about a decade since 1966, have been given the impetus to thrive again under the reform experiments that steer the nation to follow the path of so-called 'socialist market economy.' After a somewhat hibernating status of 'zero urban growth' during the 1960s and 1970s that saw no real growth of urbanization rate, China's urban share of its national population reached 32% in 1997, reflecting the rapid expansion of urban areas and high levels of urban in-migration lured by the new business and employment opportunities. It is however necessary to highlight the fact that the initial emphasis of economic reforms on the growth of township and village enterprises (TVEs) absorbed a substantial number of rural reserve labor, facilitating the transformation of rural farmers into industrial workers, and in doing so, achieving an economic expansion through what can be termed as 'urbanization from below' (Ma and Fan, 1994). From the mid-1990s, these TVEs gave way to large enterprises and foreign firms, which were increasingly concentrated in cities of the eastern and coastal provinces.

'Urban' as the Unit of Accumulation

From the early 1990s, after a brief spell of stagnation immediately after the Tian'anmen uprising, Deng Xiaoping's Southern Tour reignited China's drive for the economic reform. This time, its focus was more on prioritizing the eastern coastal region and special economic zones for resource concentration including the attraction of foreign direct investments and rural surplus labor for low production costs. These measures were epitomized in the widely publicized policy of 'Get Rich First,' which hoped for the eventual trickle-down of eastern coastal region's economic success to its hinterlands. The core and distinctive nature of China's urbanization during this period could be the rise of the 'urban' as the unit of accumulation. Three key characteristics are examined here: (1) industrial production and uneven development; (2) land-based accumulation; and (3) local state building and state entrepreneurialism.

Rise of Industrial Production and Uneven Development

Through the official endorsement of promoting a 'socialist market economy' by the Chinese Communist Party (CCP) in 1992, a full economic reform thrust was put in place, and major cities in the eastern coastal region received various preferential treatments to lead China to become the 'factory of the world.' While the exogenous factors of foreign direct investments and the formation of the global value chains have all facilitated China's transformation, the core domestic thrust in relation to urbanization could be the production and exploitation of the country's cheap labor force (Buck, 2007). This was facilitated by the household responsibility system in People's Communes in rural areas, which not only increased the agricultural productivity but also freed farmers from being tied to the collective commune system. This accompanied decollectivization, which allowed household members to engage in nonfarming activities as long as they fulfilled their obligation to the village collectives in terms of meeting the grain quota imposed upon each household (Zou, 2003). The 'Scissors' gap' between the agricultural and industrial products was abolished, and the control of household and employment registration had been considerably lessened throughout the 1990s, prompting the influx of migrants into cities. The production of cheap labor force is further aided by the household registration system that prevents migrant workers and their families from accessing full services in cities, resulting in what Kam Wing Chan (2014) refers to as *hukou*-based urbanization. China's rise as the factory of the world was also facilitated by the rapid development of the private sector and the restructuring of the state sector. Industrial outputs increased to an unprecedented level. The scale of foreign direct investment escalated exponentially, helping China to address its earlier resource constraints for financing investment in fixed assets. The FDIs largely focused on eastern coastal provinces such as Guangdong and Zhejiang, which saw a number of special economic zones emerging to accommodate production facilities for goods manufacturing for overseas market.

The economic reform and the prioritization of selected provinces and cities have resulted in uneven development with aggravating regional disparities, which became particularly prominent in the 1990s and early 2000s (Dunford and Li, 2010). Intraregional disparities have shown some signs of improvement over the years especially since the mid-2000s, but interregional disparities are yet to exhibit amelioration (Kanbur and Zhang, 2005). Such aggravation of regional disparities has led to the state implementation of regional policies that aim to redirecting resources to poorer regions. One of such policies includes the 'Go West' policy that aimed at diverting resources for investment in central and western regions (Tian, 2004). The attention of the central state to poorer and less-performing regions and provinces also owes its greater awareness of the need to boost the domestic consumption in order for the economy to sustain its current growth momentum. Some critics have noted China's vast territorial size as a safeguard for

China's interaction with the world economy on its own terms, and for the country's economy to take the advantage of the economy of scale (see for instance Liew, 2005). The expansion toward the western region in this regard has a great significance in terms of sustaining economic development as well as addressing political needs of addressing regional disparities (Shin, 2012). In this regard, cities such as Chongqing have grown to become a major regional hub, serving the domestic market in particular. With the increase in labor costs in eastern coastal provinces such as Guangdong, factories are also migrating further inland, relocating themselves to inland places that offer more preferential conditions in terms of lower wages and incentives for attracting these firms. These efforts by inland cities and provinces may involve the surge of state-led investments in fixed assets to provide the required infrastructure and facilities, reproducing in this process the speculative urbanization that characterized the cities in eastern coastal provinces (Shin, 2014a).

Land-Based Accumulation

Urbanization entails urban population increase but it also gets facilitated by the horizontal expansion of cities. In China, this equates to the territorial expansion that brings more land resources under the control of urban governments. For critics such as You-tien Hsing (2010), urbanization is synonymous with local state's territorial expansion, the process of which is summarized as 'urbanization of the local state':

> Local state leaders seek to legitimize themselves as urban promoters and builders, and urban agendas dominate local development policy, while local politics predominantly revolves around farmland conversion and industrial or commercial development projects. Urban construction has therefore expanded from an accumulation project to a territorial project of local state building. Urban modernity, now captures the political imagination of local state leaders. I call this dynamic the 'urbanization of the local state.'
>
> Hsing, 2010: p. 6

The control of urban land resources has become vital for Chinese cities' development, as the land reserves are the major sources of financing investments. Chinese local governments are known for their pursuit of land-based revenues, a practice that has gained importance throughout the 2000s in particular along with increased land commodification (Lin et al., 2014). With the reform of taxation system and decentralization, local governments are to retain a certain share of budgetary revenues, while extrabudgetary revenues raised locally are kept in the hands of local governments without having to be transferred to the central government. Land-related revenues such as land use premium and administrative fees upon land-use transaction constitute a substantial share of extrabudgetary revenues. Land-related revenues may also include loans borrowed by local government investment arms with the use of land as collateral (Ding, 2005). It is estimated that "These loans plus the revenue generated from conveyance fees accounted for 40 to 50 percent of the Hangzhou municipal government budget in 2002. In turn these revenues were used to fund more than two-thirds of the city's investments in infrastructure and urban services" (Ding, 2005). A recent study on Guangzhou suggests that the municipality's land conveyance income accounted for "about 20–25% of the total fiscal revenue" (Lin et al., 2014: p. 13) in the 2000s and "contributed 40–60% of the total urban fixed capital construction fund" (Lin et al., 2014: p. 14). China's dualist structure of land-ownership places the ownership of urban lands in the hands of the state (hence state-ownership of urban lands), while rural lands are owned by rural collectives. This provides monopoly control of land resources for local governments as *de facto* landlords (Shin, 2009), whose power of land administration has been delegated by the central government.

The utilization of land resources therefore becomes an important avenue of revenue maximization for local governments that strive to address the need of development, especially in the context of China's macroeconomic growth that depends heavily on investment in fixed assets (Shin, 2014a: pp. 505–507). The utilization of land resources may happen in four ways. First, urban governments may strive to gain their control over existing urban space that used to be occupied and controlled by what You-tien Hsing (2010) referred to as 'socialist landmasters,' that is, government institutions, state-owned enterprises, and militaries, which often make claims on the use rights of lands they occupied for years. They would do so in order to generate land revenues for their own benefits. Second, rural lands that used to lie outside the urban government control are subject to reclassification as urban lands through administrative boundary adjustments so that rural counties merge with neighboring municipalities. Third, the designation of development zones in rural counties bring greater opportunities for urban governments to raise more land-related revenues, as the strategic insertion of excluded landscapes boost the exchange value of land use rights (Wu and Phelps, 2011). Fourth and finally, existing urban spaces may be subject to revalorization through redevelopment in order to put them into a higher and better use (Shin, 2009). This process will involve expropriation of various fragmented property rights as well as the sorting of existing residents and businesses.

Local State Building and State Entrepreneurialism

China's urbanization therefore can be understood as the territorial consolidation of the power in the hands of urban governments. However, the pursuit of economic gains extends beyond land-related revenues. The rise of Chinese local states has been accompanied by their greater power to gain control on regulating local development initiatives (e.g., urban planning and land administration). The rising importance of land as commodity in financing local development has also led to the transformation of land governance into an entrepreneurial one. Land release or administering the transfer of land use rights has become a means for local governments to channel investment into the city. Having been granted with the power to use land as resources to bring money to their cities, municipalities have effectively become "the actual manager of state property" (Haila, 1999: p. 585). In this process, the city itself has become an entrepreneurial entity.

In addition to the local state's role as the manager of state land assets, the rise of local states has been facilitated further by the proactive intervention of local states in the process of

economic transition as well as in the formation of the market. The state intervention in the market has been carried out to such an extent that the local state itself has become an active participant in the market rather than remaining only as the embodiment of the regulatory practices. Branches of administrative bodies have often transformed itself into state companies or joint ventures in order to participate in the markets such as real estate. Critics have been referring to these as the rise of 'local state corporatism,' 'local state entrepreneurialism,' and so on (Duckett, 1998; Oi, 1995; Shin, 2009).

Nevertheless, the rise of local states and their entrepreneurial turn are to be treated with caution especially in consideration of their relationship with the central state. As China experiences city-based accumulation, local states have also increased its territorial power. The key question is the extent to which this enhanced local autonomy reconciles with China's ongoing centralization of political power. From the perspective of intercity competition and economic decentralization, it may be possible to argue that local states seek ways in which they can be free of central state intervention. Local political elites may embark on strategies including strategic plan-making and the promotion of special economic zones to facilitate more inward foreign direct investment and the expansion of their accumulation capacities (Wu and Zhang, 2007). However, it is equally possible to envisage that the central state has strived to make sure their presence is felt in local policy making (Chen, 2009; Xu and Yeh, 2009). Alternatively, local states may aspire to their territorial ambition, both domestic and international, under the auspices of the central state (Shin, 2014b). It has been argued by various critics that the rise of key cities such as Shanghai as the world city has been supported by the central state (Wu, 2009). For instance, Xu et al. (2009) examine the process of land commodification and developmental politics, arguing that both local and central states engage in scalar politics to exert greater power in their own rights. In particular, they remind the attempts by the central state "to control the articulation of scales through which a more centrally consolidated power can be achieved" (p. 910). The territorial politics become ever more sensitive and complex as China sees the rise of various regional scales including city regions. Disposition of land as resources is often at the center of this politics, for the reasons as discussed above.

Remaking Chinese Urban Landscape

Spatial restructuring in Chinese cities has become the cause and outcome of China's rapid urbanization. It is a well-known fact that investment (that is, fixed capital formation) has been the most influential contributor to the macroeconomic development. Real estate investment in particular has been influential in the formation of gross domestic capital in cities at all levels. With the commodification of land and housing, major commercial housing projects have been carried out initially near suburban areas where land mobilization was relatively easier, and eventually in inner city areas in the form of urban redevelopment projects. The rise of household income as a result of China's economic development also provided a huge demand for improved housing conditions. The speculative environment in the real estate housing sector further prompted the notion of housing as an asset.

It is often discussed that work-unit compounds characterized the urban landscape of most Chinese cities. These work-unit compounds were the spatial representation of production and consumption spaces, and were the loci of microlevel politics that filtered the exercise of state power through the unique institutional forms of managers–workers relations (Bray, 2005; Walder, 1986). Spatially, work-unit compounds were effectively gated compounds that provided all daily necessities and family needs including nursery, schooling, medical care, and old age care. The quality of these provisions depended largely on the political strength and affluence of work-units. The administrative allocation of urban lands and the subordination of land use to production needs led to the homogenization of urban landscapes, minimizing the chances of spatial and social diversities.

Urbanization during the reform era brought about fundamental changes to the ways in which spatial production and consumption were organized. New spaces have emerged, catering for the needs of diverse urban interests including those of increasingly affluent urban residents, of local elites who often aspire to transforming their cities into 'world class cities,' and of migrants who struggle to settle down in cities. These new spaces may include art districts, central business districts (CBDs), gated commercial estates, heritage neighborhoods, monumental spectacles, and so on (Broudehoux, 2010; Pow, 2009; Ren and Sun, 2012; Shin, 2010, 2012). One of the major spatial manifestation that has captured the attention of urban critics is the presence of what is often referred to as 'urbanized villages,' 'villages-in-the-city,' or 'villages within a city.' These are former rural villages that have become urbanized in times of rapid expansion of neighboring urban areas. Having succumbed to the power of urbanization, these former rural villages come to exhibit a unique spatial expression of China's urbanization and informality particularly in southern China, where a huge influx of migrant population has been witnessed for many decades (Tang and Chung, 2002). Migrants' need of affordable rental housing is seen to have propelled the rise of such settlements, further fueled by the emergence of informal landlordism on the part of former villagers whose income generation opportunities were severely undermined due to the loss of agricultural income after farmland expropriation resulting from urban expansion. Critics often question the absence of urban sprawl in the manner as seen in other developing countries with rapid urbanization. Latin American cities have also seen the prevalence of informal settlements through land invasion, illegal subdivision, and informal extension. A close examination of the urbanized former rural villages and their emergence indicate a unique process of China's urbanization that involves a historic compromise between local states and village collectives, resulting in a less conflictual process of urban expansion.

The process of Chinese cities' remaking in recent decades has come with some huge costs especially in the form of local residents' displacement. Land assembly for the provision of various urban infrastructure and facilities or urban redevelopment projects targeting densely populated inner city districts have led to the displacement of hundreds of thousands of local citizens. For instance, the Geneva-based Centre on Housing

Rights and Eviction estimated on the basis of official records that about 1.5 million Beijing citizens would have been affected by various developmental projects between 2000 and 2008. This estimation does not take migrants into consideration and, therefore, the actual scale of displacement would far outweigh the official estimation (COHRE, 2007). In Shanghai, official data also suggest that about half a million households were affected by redevelopment between 2003 and 2010 (Shanghai Municipal Statistical Bureau, 2011). The scale of relocation and permanent displacement of local residents become particularly pronounced when cities embark on major developmental projects to address the aspiration to become 'world class cities,' a strategy that often accompanies the hosting of megaevents such as the 2008 Beijing Summer Olympic Games and the 2010 Shanghai World Expo (Shin, 2012).

Urbanization, Citizenship, and Rights Awareness

As the natural population growth in cities has been under control especially with the implementation of the One Child Policy since the early 1980s, China's urbanization is primarily driven by people's migration from rural counties to cities. This further raises questions about the usefulness of the household registration system that has been in place for decades since the late 1950s. Household registration was initially designed as a means to devise the ways in which place-based entitlements could be distributed. Rural residents were to have access to welfare entitlements provided by village collectives on the basis of their membership in the village collectives, while urban residents were to benefit from welfare provisions by their employers. Local governments intervened in the case of those few residents who failed to receive any kind of support from employers. Over time, the *hukou* system has become a tenacious measure to dissociate migrants with no permanent local *hukou* from accessing urban-based social services (Solinger, 1999), hence playing the role of invisible barrier to rural-to-urban migration. Migrants in cities were "still officially classified as temporary residents in the city and treated as noncitizens" (Zhang, 2002: pp. 317–318).

With the growing concern of inequalities and of more just distribution of economic benefits in urban areas, various localities are increasingly enacting urban and social policies in the manner that could be described as 'local citizenship'. That is, instead of making a strict distinction between urban and rural residents, the *hukou* system seems to transform itself into strengthening the division between local and nonlocal residents, emphasizing economic securities for local residents while minimizing benefits for nonlocal migrants (Smart and Lin, 2007). The shaping of local citizenship will see geographical unevenness. For those less affluent provinces and cities that aim to attract industries as well as compete for migrant workers, the *hukou*-related barriers are expected to be incrementally removed. More affluent cities may further raise the bar to selectively recruit qualified skilled migrant workers. The extent to which these uneven processes will reproduce dualism with regard to local and nonlocal citizens' access to urban services will require close examination in the future. This will also raise some important questions about the nature of citizenship in mainland China, especially with regard to the ways in which collective membership becomes the precondition for rights entitlement, and also with regard to the relationship between state citizenship, urban citizenship, and collective citizenship in rural areas.

Another important dimension is the issue of rights awareness and how the growing degree of rights claims communicate with the traditional notions of rights awareness in China. In particular, an increasing emphasis has been placed on building the middle class in China (Tomba, 2004; Zhang, 2010). This attracted attention not only from the scholars studying China's economic development and social progress but also from policy-makers and international consultancies. For domestic and global businesses, the expansion of middle class populations means greater buying power, creating opportunities to market high-end products including luxury goods. For policy-makers, their recent emphasis on boosting consumption is designed to supplement investment and sustain economic growth in times to come. For critics on China's societal progress, the middle class expansion raises the question of whether or not the expansion of China's middle class will bring about the rise of civil society and greater awareness and call for social, economic, and political rights. It is however unclear if the middle class populace will exert pressure on the state for more political reform. A recent study suggests that China's new middle class is keen to protect their individual rights but reluctant to exercise collective rights to challenge the state (Chen, 2013). Critics also point out that China's rights claims have been largely on economic subsistence (e.g., pay arrears), and there are signs that the middle class protests on various social and environmental issues are on the rise but rarely to the level of making claims against the state (Perry, 2008; Shin, 2013). Furthermore, the growth of the middle class populace in China is part of a dual process of middle class making and proletarianization of rural farmers in the urban contexts of exacerbating inequalities. For middle classes to be acting as the social agent of changes there is a need for them to find a common ground to form a broader cross-class alliance. This also requires a greater attention to what the position of working classes and farmers is going to be in a country that is arguably witnessing the "loss of a socialist state" (Lin, 2014: 43; see also Shin, 2014a).

Conclusion: Uneven Geography of China's Urbanization

This article has examined the contemporary trend of China's urbanization, the key drivers of this urbanization, and the major sociospatial outcomes and concerns. China's reintegration with the world economy during the post-1978 period has branded China as the 'factory of the world.' This was made possible by the presence of a rapidly growing number of migrant workers, central and local state policies to provide incentives and preferential treatments for inward investment, and China's geographical scale that enabled an economy of scale for domestic production and consumption. From the perspective of local states, the urbanization has increasingly become synonymous with the process of local state building.

As local governments are delegated with the power to administer the transaction of urban land use rights on the basis of state-ownership of urban lands, territorial politics have become prevalent especially with regard to gaining control on land resources, which are nowadays prominent sources of financing investment in fixed assets.

The article has also discussed some of the major sociospatial outcomes of China's urbanization, focusing on the remaking of Chinese cities that accompanied the rise of CBDs to bring the urban landscape to meet the urban aspiration to become global in appearance. While Chinese cities used to share the key characteristics of 'socialist cities' during the era of planned economy, they have trodden the path of diversity and heterogeneity, seeing the emergence of various cityscapes that range from commercial gated communities and heritage neighborhoods to art districts, and the CBDs. Together with heavy investment in urban infrastructure and facilities, these developmental projects nevertheless produce heavy costs for on-site residents. Official aggregate estimations often report millions of local citizens being affected by these projects, and these numbers are largely underestimated as migrant tenants are largely excluded from the counting. Major cities often produce explicit policies to depopulate city centers, and transform the area to meet the demands of domestic and foreign professionals and investors, thus paving the way for the class remake of urban space.

As China's 'urban age' is expected to intensify under the Party State, the remaining question is its long-term stability of current development models. Discussions on this topic propel China observers to think of the value of China studies and its contribution to the general scholarship. For instance, while China's development thus far has focused on major sites of accumulation along the eastern coastal region, more attention is now being paid to its hinterlands, that is, central and western regions. In order to further boost both investment and consumption, cities in central and western regions may hold the key to the future of China's sustainable economic development to similar rates as experienced during the post-1978 reform period. This requires the decentering of urban China studies away from major cities such as Beijing and Shanghai. It is no unknown secret that urban China studies have been concentrated on a selected few number of cities (or on a so-called 'first-tier' cities). China's urbanism is expanding to influence lower-tier cities in China, and the new leadership's drive for further urbanization will require the growth of smaller towns and counties where there are more rooms for expansion. It will be necessary for China scholars to closely examine the conditions of urban accumulation and city-remaking therein. The other major issue will be the extent to which political stability at all geographical scales (i.e., urban, regional, and national) will be maintained. The CCP has been reinventing itself to adopt to the changing social, economic, and political contexts (the best known example being the admittance of entrepreneurs as Party members since 2003), but it has also proven its strict, brutal approach when social protests have made attempts to make cross-class or cross-regional alliances. The rise of middle classes and their increasing demands for better living conditions and the security of their household assets have the potential to be turned against the state itself. The aggravating inequalities and the continued pursuit of proletarianization of rural surplus labor will exert pressure on the state, while growing gaps between the new middle class and the new urban poor also become a major source of concern. To this extent, the long-term prospect of China's urbanization hinges upon the way in which political stability can be maintained by the state, especially in times of cyclical economic difficulties generated domestically as well as globally. This also makes it important for China observers to closely follow the rise of discontents and rights claims, and scrutinize the ways in which Chinese Party State responds to these.

See also: Asia, Sociocultural Aspects: China; Cities Under Postsocialism; Development and Urbanization; Development and the State; Urban Planning in China; Urbanization.

Bibliography

Anagnost, A., 2004. The corporeal politics of quality (Suzhi). Public Culture 16 (2), 189–208.

Bray, D., 2005. Social Space and Governance in Urban China: The Danwei System from Origins to Reform. Stanford University Press, Stanford, CA.

Brenner, N., Schmid, C., 2014. The 'urban age' in question. International Journal of Urban and Regional Research 38 (3), 731–755.

Buck, D., 2007. The subsumption of space and the spatiality of subsumption: primitive accumulation and the transition to capitalism in Shanghai, China. Antipode 39 (4), 757–774.

Broudehoux, A.-M., 2010. Images of power: architectures of the integrated spectacle at the Beijing Olympics. Journal of Architectural Education 63 (2), 52–62.

Chan, K.W., 2014. China's urbanization 2020: a new blueprint and direction. Eurasian Geography and Economics 55 (1), 1–9.

Chan, K.W., 1994. Cities with Invisible Walls: Reinterpreting Urbanization in Post-1949 China. Oxford University Press, Hong Kong.

Chen, J., 2013. A Middle Class Without Democracy: Economic Growth and the Prospects for Democratization in China. Oxford University Press, Oxford.

Chen, X. (Ed.), 2009. Shanghai Rising: State Power and Local Transformations in a Global Megacity. University of Minnesota Press, Minneapolis; London.

COHRE, 2007. Fair Play for Housing Rights: Megaevents, Olympic Games and Housing Rights. Centre on Housing Rights and Evictions, Geneva.

Ding, C., 2005. Property tax development in China. Land Lines 17 (3). http://www.lincolninst.edu/pubs/1041_Property-Tax-Development-in-China.

Duckett, J., 1998. The Entrepreneurial State in China: Real Estate and Commerce Departments in Reform Era Tianjin. Routledge, London.

Dunford, M., Li, L., 2010. Chinese spatial inequalities and spatial policies. Geography Compass 4 (8), 1039–1054.

Haila, A., 1999. Why is Shanghai building a giant speculative property bubble? International Journal of Urban and Regional Research 23 (3), 583–588.

Hsing, Y-t, 2010. The Great Urban Transformation. Oxford University Press.

Kanbur, R., Zhang, X., 2005. Fifty years of regional inequality in China: a journey through central planning, reform, and openness. Review of Development Economics 9 (1), 87–106.

Kirkby, R.J., 1985. Urbanization in China: Town and Country in a Developing Economy 1949–2000 AD. Columbia University Press, New York.

Liew, L.H., 2005. China's engagement with neoliberalism: path dependency, geography and party self-reinvention. The Journal of Development Studies 41 (2), 331–352.

Lin, C., 2014. The language of class in China. In: Panitch, L., Albo, G. (Eds.), Socialist Register 2015: Transforming Classes. The Merlin Press, London, pp. 24–53.

Lin, G.C.S., Li, X., Yang, F.F., Hu, F.Z.Y., 2014. Strategizing urbanism in the era of neoliberalization: state power reshuffling, land development and municipal finance in urbanizing China. Urban Studies [online first]. http://dx.doi.org/10.1177/0042098013513644.

Ma, L., Fan, M., 1994. Urbanization from below: the growth of towns in Jiangsu, China. Urban Studies 31 (10), 1625–1645.

Naughton, B., 1995. Cities in the Chinese economic system: changing roles and conditions for autonomy. In: Davis, Deborah S., et al. (Eds.), Urban Spaces in Contemporary China: The Potential for Autonomy and Community in Post-Mao China. Woodrow Wilson Centre Press, Washington, DC, pp. 61–89.

Oi, J.C., 1995. The role of the local state in China's transitional economy. The China Quarterly 144, 1132–1149.
Perry, E.J., 2008. Chinese conceptions of "rights": from Mencius to Mao – and Now. Perspectives on Politics 6, 37–50.
Pow, C.-P., 2009. Neoliberalism and the aestheticization of new middle-class landscapes. Antipode 41 (2), 371–390.
Ren, X., Sun, M., 2012. Artistic urbanization: creative industries and creative control in Beijing. International Journal of Urban and Regional Research 36 (3), 504–521.
Shanghai Daily, March 07, 2013. Expanding Cities to Drive the Economy. URL: http://www.shanghaidaily.com/article/print.asp?id=525399 (accessed 17.03.13.).
Shanghai Municipal Statistics Bureau, 2011. Shanghai Statistical Yearbook 2011. China Statistics Press.
Shin, H.B., 2014a. Contesting speculative urbanization and strategizing discontents. City 18 (4–5), 509–516.
Shin, H.B., 2014b. Urban spatial restructuring, event-led development and scalar politics. Urban Studies 51 (14), 2961–2978. http://dx.doi.org/10.1177/0042098013515031.
Shin, H.B., 2013. The right to the city and critical reflections on China's property rights activism. Antipode 45 (5), 1167–1189.
Shin, H.B., 2012. Unequal cities of spectacle and megaevents in China. City 16 (6), 728–744.
Shin, H.B., 2010. Urban conservation and revalorization of dilapidated historic quarters: the case of Nanluoguxiang in Beijing. Cities 27 (Suppl. 1), S43–S54.
Shin, H.B., 2009. Residential redevelopment and the entrepreneurial local state: the implications of Beijing's shifting emphasis on urban redevelopment policies. Urban Studies 46 (13), 2815–2839.
Smart, A., Lin, G.C.S., 2007. Local capitalism, local citizenship and translocality: rescaling from below in the Pearl River Delta Region, China. International Journal of Urban and Regional Research 31, 280–302.
Solinger, D.J., 1999. Contesting Citizenship in Urban China: Peasant Migrants, the State, and the Logic of the Market. University of California Press, Berkeley.
Song, F., Timberlake, M., 1996. China's urbanization, state policy and the world economy. Journal of Urban Affairs 18 (3), 285–306.
Tang, W.-S., Chung, H., 2002. Rural-urban transition in China: illegal land use and construction. Asia Pacific Viewpoint 43 (1), 43–62.
Tian, Q., 2004. China develops its west: motivation, strategy and prospect. Journal of Contemporary China 13 (41), 611–636.
Tomba, L., 2004. Creating an urban middle class: social engineering in Beijing. The China Journal 51, 1–26.
United Nations Department of Economic and Social Affairs/Population Division, 2012. World Urbanization Prospects: The 2011 Revision. United Nations, New York.
Walder, A.G., 1986. Communist Neotraditionalism: Work and Authority in Chinese Industry. University of California Press, Berkeley.
Wu, F., 2009. Globalization, the changing state, and local governance in Shanghai. In: Chen, X. (Ed.), Rising Shanghai: State Power and Local Transformations in a Global Megacity. University of Minnesota Press, Minneapolis; London, pp. 125–144.
Wu, F., Phelps, N.A., 2011. (Post)suburban development and state entrepreneurialism in Beijing's outer suburbs. Environment and Planning A 43 (2), 410–430.
Wu, F., Zhang, J., 2007. Planning the competitive city-region. Urban Affairs Review 42 (5), 714–740.
Xu, J., Yeh, A.G.O., 2009. Decoding urban land governance: state reconstruction in contemporary Chinese cities. Urban Studies 46 (3), 559–581.
Xu, J., Yeh, A., Wu, F., 2009. Land commodification: new land development and politics in China since the late 1990s. International Journal of Urban and Regional Research 33 (4), 890–913.
Zhang, L., 2010. In Search of Paradise: Middle-Class Living in a Chinese Metropolis. Cornell University Press, Ithaca.
Zhang, L., 2002. Spatiality and urban citizenship in late socialist China. Public Culture 14 (2), 311–334.
Zou, W., 2003. The Changing Face of Rural Enterprise. China Perspectives. http://chinaperspectives.revues.org/773.

Utilitarian Social Thought, History of

James E Crimmins, The University of Western Ontario, London, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The beginnings of utilitarian social theory can be traced to Greek philosophy. In the late seventeenth and early eighteenth centuries, it began to take on its classical form. Jeremy Bentham is usually identified as the fountainhead of the doctrine, but the contributions of Hutcheson, Hume, Beccaria, Helvétius, Godwin, and the 'religious utilitarians' should also be noted. After Bentham, the theory received its most celebrated treatment by John Stuart Mill. The article describes the essential elements of utilitarian theory, delineates the dissemination and development of utilitarian ideas, explores problems of theory and interpretation, and provides a survey of later developments.

Standard utilitarian theory is underwritten by the following moral claims: the rightness/wrongness of an action is determined by the goodness/badness of its consequences; the only thing that is good in itself is pleasure and the only thing bad in itself is pain; and happiness/misery is constituted of the aggregation of pleasures and pains. Implicit in this ethical hedonism is the premise that the motives to action are rooted in self-interest. Based on these elements, the doctrine is then expressed in the form of the greatest happiness principle – that is, the rightness of an action is determined by its contribution to the happiness of everyone affected by it. In *A Fragment on Government* (1776) the legal philosopher Jeremy Bentham (1748–1832) declared it "a fundamental axiom" that "it is the greatest happiness of the greatest number that is the measure of right and wrong," and that "the obligation to minister to general happiness, was an obligation paramount to and inclusive of every other" (Bentham, 1977: pp. 393, 440–441). In 1781, in recording a dream in which he imagined himself as the founder of a "sect", Bentham coined the name "utilitarian" for this theory (Crimmins, 1990: p. 314).

Origins

Although Bentham is usually cited as the founder of utilitarianism, the antecedents of utilitarian principles have a far older vintage in the philosophy of Plato, Aristotle, and Epicurus, and in early Christian thought (Scarre, 1996: Chapter 2). Other significant dimensions of the theory can be traced to the seventeenth century writings of Thomas Hobbes, John Locke, and Richard Cumberland. However, it was the 'moral sense' philosopher Francis Hutcheson (1694–1746) who first cast the objective of morality in explicitly utilitarian terms when he wrote, in *An Inquiry into the Original of our Ideas of Beauty and Virtue* (1729), "*that Action is* best, which accomplishes the *greatest Happiness* for the *greatest Numbers*; and that, worst, which in *like manner*, occasions *Misery*" (Hutcheson, 2004: p. 125). In addition, Hutcheson speculated about the appropriateness of 'applying mathematical Calculation to moral Subjects', including the possibility of assessing happiness in terms of 'addition' and 'diminution', and hazarded the view that what matters in calculating the virtue of an action is the degree of happiness and "the Number of Persons to whom the Happiness shall extend" (Hutcheson, 2004: pp.123, 125, 128). By indicating that quantities and the extent of happiness matter in making moral claims Hutcheson set the agenda for later utilitarian moralists, though he himself continued to believe that the disinterested motive of benevolence was the foundation of virtue and later expunged all traces of the discussion of mathematical calculation in moral matters from subsequent editions of the *Inquiry*.

Like Hutcheson, William Godwin (1756–1836), in *An Enquiry Concerning Political Justice* (1793), also stipulated that virtue was located in 'universal benevolence', in seeking the pleasures of others (Godwin, 1793: vol. 1, pp. 87, 106). Nor, again like Hutcheson, could he be considered an outright consequentialist, since he believed that a man's 'disposition' was a surer guide to virtue than his actions (Godwin, 1793: p. 122). On the other hand, in a similar vein to Bentham, Godwin depicted society as a simple aggregate of individuals and subscribed to the view that justice, properly understood, was 'coincident with utility' (Godwin, 1793: pp. 90, 121). The gulf between Godwin and Bentham widened considerably, however, in their reflections on the purpose of law. For Godwin, reputedly the first philosophical anarchist, law was only a temporary aid to virtue. In the belief that men are not perfect but are perfectible (Godwin, 1793: p. 118), Godwin argued that moral instruction was the principal way to bridge the conflict between egoism and benevolence. Once individuals properly understood their duty to act in the public interest, then the sanctions of law would become redundant. The optimism embedded in this idea led Godwin to envisage a time when societies might be able to function entirely on the principle of cooperation to produce the greatest happiness without the need for government.

It was in the moral philosophy of the eighteenth-century religious advocates of the utility principle that utilitarian theory began to assume its classic form. Proponents of religious utilitarianism included John Gay (1699–1745), John Brown (1715–66), Edmund Law (1703–87), Abraham Tucker (1705–74), and William Paley (1743–1805). The ethical hedonism developed by these moralists was premised on the existence of a God who would punish and reward in the afterlife according to the degree to which individuals acted to enhance general happiness in the present life; in short, personal happiness – encompassing the prospect of eternal happiness in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.03143-3

the life to come – is achieved by pursuing social happiness in the here and now (Crimmins, 1998). Gay's major contribution to the utilitarian tradition is the theory of 'associationism', which explained actions based on habit and was further developed and given common currency by David Hartley (1705–57). Hartley's theory of association is a constituent assumption in the thought of nearly all the later utilitarians and in particular was a major influence on James Mill.

Among the religious exponents of utility, only Paley offered a comprehensive and systematic social philosophy founded upon the utility principle, encompassing morals, government, law, and political economy, in his popular *Principles of Morals and Political Philosophy* (1785). But Paley's politics, with certain liberal exceptions (religious toleration and the abolition of slavery), was essentially conservative, and his moral theory tended to encourage complacency toward the causes of misery in the present life (Crimmins, 2013a).

Bentham's secular version of the theory was derived from a more progressive strain of enlightenment thought: empirical, analytical, and antagonistic toward religious beliefs and other moral mythologies. Unaware of Hutcheson's formulation of the utility principle, Bentham attributed it to the Italian law reformer Cesare Beccaria (1738–94) and sometimes to the dissenter and natural philosopher Joseph Priestley (1733–1804). Other progenitors to whom he traced his intellectual inheritance included David Hume (1711–76) and Claude Helvétius (1715–71) (Rosen, 2003: Chapters 3 and 5). In Hume's secular moral philosophy, the source of the rules of justice were located in general utility, but he explained the psychology of moral conduct in terms of a theory of 'moral sentiments' rather than motives supplied by pleasures and pains. Bentham's acknowledgment of the importance of Hume was tempered by his general irritation with the subjectivism of the Scottish 'moral sense' school, and by Hume's failure to apply the utility principle as a critical tool of analysis and guide to social improvement (Crimmins, 2011a: Chapter 3). In *De l'esprit* (1758), Helvétius provided a systematic if rudimentary explanation of human motivation in terms of utility perceptions and expectations and some useful suggestions for the application of this knowledge to law and other social questions. But his philosophy lacked the systematic arrangement, conceptual clarity and rigor necessary for a 'science' of morals and legislation. This is what Bentham set out to supply.

Bentham's Science of Morals and Legislation

In *An Introduction to the Principles of Morals and Legislation* (1789), Bentham delineated the component parts of his 'science' of morals and legislation founded upon the utility principle. The general purpose of the science was "to rear the fabric of felicity by the hands of reason and of law" (Bentham, 1970: p. 11). The *Introduction* was then considered by Bentham to be the foundational text of his legal philosophy and his plans for future works were intended to elucidate the practical application of the utility principle in the areas of civil and penal law, judicial procedure, legislative reward, constitutional law, legislative procedure, international law, finance, political economy, and 'universal jurisprudence' (Bentham, 1970: pp. 5–6).

The Principle of Utility

At the base of Bentham's theory was the assumption of "the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do" (Bentham, 1970: p. 11). As a first principle used to establish everything else, the utility principle was not itself subject to proof, but Bentham claimed, there was sufficient testimony to the fact that it had always been generally accepted as the standard of right and wrong (Bentham, 1970: p. 13). This foundational principle referred to a feeling or sentiment that served to approve the utility of an action. As Bentham explained, "The principle here in question may be taken from an act of the mind; a sentiment; a sentiment of approbation; a sentiment which, when applied to an action, approves of its utility, as that quality of it by which the measure of approbation or disapprobation bestowed upon it ought to be governed" (Bentham, 1970: p. 12). Utility also meant public utility, and the utility of the individual was a part of the public utility in which the individual shared. Thus a person approved of an act or law to the degree to which the happiness of all the individuals in 'the community in question' were advanced. For Bentham, the relationship between happiness and pleasure and pain was straightforward and did not require elaboration: pleasure contributes to happiness, while pain detracts from it. In these terms the public interest is constituted of the aggregate of individual interests, construed in terms of the balance of pleasures and pains experienced by each person.

Bentham's intention was not that the agent should only consider his own personal interests when contemplating action, nor that we should *always* bear in mind the happiness of the whole of humanity. Rather, when a person considers the utility of an action he should take into account the impact of the action on the interests or well being of those he knew would be affected by it: where the action concerns only the agent, then that is the only interest to be considered; when the action impacts upon others, then their interests must be counted too; and when lawmakers enact a new law or policy it is the impact on the happiness of the people at large they must bear in mind, up to and including the inhabitants of an entire society, and quite possibly the inhabitants of other societies. At his most universal Bentham, who was responsible for introducing the new word 'international' into both the French and English languages, held that the object of international law is "the greatest happiness of all nations taken together" (Bentham, 1838–43: vol. 2, p. 538).

Motives and Consequences

For Bentham all motives have their source in the anticipation or expectation of pleasure or pain. He delineated four *sanctions* or sources of pain and pleasure, which he may have learned from either Locke or Gay: physical, political, moral, and religious. These sanctions are available to the moralist and to the legislator in guiding and determining an individual's moral conduct, and explain how an essentially self-interested individual can be directed to perform actions, which enhance general happiness. In Bentham's theory of motives, the dictates of utility were neither more nor less than the dictates of the most extensive and enlightened benevolence. Other motives

were categorized either as 'semisocial' motives (such as love of reputation, desire of amity, and those rooted in religion), which may or may not coincide with the principle of utility, and purely 'self-regarding' motives (such as physical desire, pecuniary interest, love of power, and self-preservation), which were the least likely to coincide with utility. All motives, however, including the most extensive benevolence, are rooted in self-interest in one form or another; in these terms, the notion of a *dis*interested motive is implausible.

While the consequences of an act are partly dependent on the motives which give birth to the act (and partly on the empirical circumstances in which the act is performed), in Bentham's theory the utility of an act is not dependent on its originating motive(s). The relative utility of an act is determined solely by its consequences: the advantages and/or disadvantages that result. When deciding whether to act or which act to perform the individual must calculate as best as he or she can the pleasures and pains which may reasonably be expected to accrue to the persons (including him- or herself) affected by the acts under consideration. Bentham's famous *hedonic* or *felicific calculus* (he did not use these terms) by which the value of pleasure and pain may be reckoned was constituted of the following elements: intensity, duration, certainty or uncertainty, propinquity or remoteness, fecundity, purity, and extent. A similar calculation should guide the legislator in formulating policy and law, though Bentham recognized that precision in such matters could only be impressionistic – the closer the calculation came to the 'ideal' the greater the prospect of devising effective measures. Moreover, Bentham recognized that it was not normally feasible nor rational for an individual to engage in repeated calculations prior to acting in everyday life, hence he spoke of habit (based on association) as the general source of action and an understanding of the general tendencies of actions to enhance or diminish happiness as sufficient to guide us in most situations.

Punishment and the Panopticon

The consequences of acts were central to Bentham's theory of punishment, in part derived from Beccaria's path-breaking treatment of the subject in *On Crimes and Punishments* (1764). The utilitarian objective of criminal law is general deterrence and this is served by ensuring that punishments are in proportion to the mischiefs produced by crime. To this end Bentham developed a series of rules to guide the lawmaker in the construction of a utilitarian penal code, including the elements involved in the calculation of the mischief caused by an offense and the form and degree of punishment necessary to effect deterrence (Crimmins, 2011b). Following Beccaria, though on more clear-cut utilitarian grounds, Bentham opposed the death penalty, arguing that deterrence in capital crimes is best served by the example of offenders serving life in prison with hard labor.

Bentham's controversial panopticon, best known as the design for a circular prison in which prisoners lived under the apprehension of constant observation by officials positioned in a central watchtower, has occasioned a good deal of attention over the years since its inception in the late 1780s. Michel Foucault famously took the perspective that *panopticism* defined a 'new physics of power', an experimental 'laboratory of power' in which behavior could be modified, and viewed the panopticon as a symbol of the repressive, disciplinary society, the modern 'society of surveillance' (Foucault, 1977: p. 208). This view of the panopticon has opened some interesting lines of discourse on the encroaching methods of social control in contemporary liberal societies. As a critique of Bentham's proposals, however, it is less impressive.

In *Panopticon; Or the Inspection House* (1791) Bentham gave practical expression to the utilitarian theory of punishment by stipulating a range of subordinate principles to which his circular prison should conform: economy, since the prison should be a private self-sustaining operation based on the productive labor of the inmates and not require financial assistance from the public; severity, because it was necessary for the offender to suffer to serve the ends of general deterrence and personal reformation; and humanity, which prescribed that prisoners should be deprived only of liberty not of health or life. Upon their release, reformed convicts should be able and willing to live on the rewards of their own labor. To these guiding principles Bentham added several devices to produce transparency and accountability, chief among them opening up the operations of the prison and its officials to public scrutiny (Semple, 1993).

Bentham also speculated about the application of the panopticon idea to other institutional arrangements in which a high degree of supervision of the 'inmates' would produce beneficial results, including hospitals, asylums, poor houses, factories, and schools. He developed a myriad of improvements to accompany such buildings, and in particular devoted an extraordinary amount of time to poor law reform (Bentham, 2001, 2010), but did not live to see the construction of a panopticon of any sort. In later life, when he was primarily occupied with developing a complete code of laws for 'all nations professing liberal opinions', he found time to advocate reform of the laws governing anatomical research and willed his own body for public dissection following his death in 1832 (Crimmins, 2002). Two of the greatest social campaigners of the first half of the nineteenth century, the medical and health reformer Thomas Southwood Smith (1788–1861) and the poor law and factory reformer Edwin Chadwick (1800–90), worked closely with Bentham in this period and developed their ideas under his influence.

John Austin

Although sometimes described as an act utilitarian, it is generally acknowledged that for Bentham properly constituted *rules* based on calculations of general utility, ought to be the guides for human action (Baumgardt, 1966). John Austin (1790–1859), a disciple of Bentham's jurisprudence, is usually cited as the first explicit rule utilitarian. However, in *The Province of Jurisprudence Determined* (1832), he made it clear that he learned the significance of rules from Paley, from whom he also borrowed the theological elements of his theory. Austin became acquainted with utilitarianism during his undergraduate years at Cambridge. In 1821 he was hired to tutor J.S. Mill in Roman law, began attending meetings of the Utilitarian Society established by the younger Mill in 1823, and in 1826 was appointed to the Chair of Jurisprudence at the newly founded University of London. Austin's legal philosophy had a major impact on

English and American jurisprudence, described by one commentator as "a canonical statement of modern legal positivism and analytical jurisprudence" (Lieberman, 2013).

James Mill

James Mill (1773–1836) has often been credited with influencing Bentham's conversion to democratic principles in 1808–09; however, Bentham had already developed utilitarian arguments for democratic reform in manuscripts written around the time of the French Revolution. What Mill certainly brought to the political cause of utilitarianism was clarity of expression, which he employed to telling effect in his *Essay on Government* (1819). Based on the premise that government has the purpose of producing the greatest happiness of the greatest number, in this essay Mill made a succinct case for substantive parliamentary reform (universal manhood suffrage, secret ballot, and annual parliaments). Nevertheless, Mill was taken to task by the Irish Benthamite socialist William Thompson for excluding women from the proposed reform of the franchise (*Appeal of One Half of the Human Race Women against the Pretensions of the other Half Men*, 1825). More famously, the Whig T.B. Macaulay mounted a biting criticism of Mill's essay in the *Edinburgh Review* in 1829, largely for its narrow view of human nature and a priori science of politics. Both Bentham and the younger Mill sided with Thompson on votes for women, though for tactical reasons Bentham remained silent on the question when presenting his mature democratic theory in *Plan of Parliamentary Reform* (1817) and other constitutional writings from the 1820s. Nor were Bentham and the younger Mill impressed by James Mill's recommendation that the vote be based on a modest property qualification and limited to men over the age of 40; they opted for universal suffrage at the age of 21, subject only to a literacy test. Mill also considered the advantages of proportional representation and plural voting in *Thoughts on Parliamentary Reform* (1859) and *Considerations on Representative Government* (1861).

Dissemination of Utilitarian Ideas

Bentham's writings – with the notable exceptions of *A Fragment on Government* and *Defence of Usury* (1787) – were often disparaged for their tortuous style, marked by the irksome use of parentheses and a distracting tendency to convey ideas in newly minted terminology. It is doubtful that he would have achieved an extensive reputation as a moralist and legal philosopher had it not been for the Genevan lawyer Étienne Dumont (1759–1829), who assumed the role of Bentham's translator and editor, and fashioned an international audience for his work. In 1802 Dumont published *Traités de législation civile et pénale* (3 vols., 1802), the main parts of which were the treatises on civil and penal law, and followed this with several other publications based on Bentham's writings: *Théorie des peines et des récompenses* (1811), *Tactiques des assemblée législatives, suivi d'un traité des sophismes politiques* (1816), *Traité des preuves judiciare* (1823), and *De l'organisation judiciaire et de la codification* (1828). None of these texts was a straightforward translation of Bentham, but are rather styled redactions or recensions, either edited from Bentham's original papers or combining manuscript material with extracts from published works. In this form utilitarian ideas reached many more readers than Bentham ever could have hoped (Crimmins, 2014).

After its initial circulation in France, the *Traités* was translated into Russian, Spanish, German, Hungarian, Polish, and Portuguese. Reportedly, 50 000 copies of Dumont's various recensions were sold in Europe in the early decades of the century and 40 000 in Spanish translation in Latin America alone, where Bentham's ideas were adopted in the nascent law schools. The impact of utilitarian ideas on legal education in the United States began with David Hoffman (1784–1854) and his celebrated *A Course of Legal Study* (1817, revised and expanded 1836), later complemented by *Legal Outlines* (1829). Soon after, the 'American Utilitarian' Richard Hildreth (1807–65) translated the main parts of Dumont's *Traités* as *Theory of Legislation* (1840) – a work destined to remain at the center of utilitarian studies in the English-speaking world through to the middle of the twentieth century – and followed that with his own utilitarian *Theory of Morals* (1844).

The reaction to the spread of utilitarianism in the United States was mixed: praise for Bentham's legal philosophy, particularly the theory of civil law, was offset by objections to its underlying moral assumptions and disdain for religion. Among the notable American intellectuals inspired by Bentham in this period were Thomas Cooper (1759–1839), a prolific author and controversialist whose writings included political economy, law, science, and philosophy; Edward Livingston (1764–1836), the famous author of codes of law for Louisiana who acknowledged Bentham as the world's leader in this domain; and John O'Sullivan (1813–95), the quixotic editor of the *United States Magazine and Democratic Review* who wrote sympathetic reviews of Bentham, Livingston, and Hildreth, and followed them in advocating utilitarian law reform, particularly the abolition of capital punishment. Bentham's influence continued throughout the century in America, where the *Traités* paved the way for the reception of other editions and versions of his writings, which in turn led to sympathetic responses to the more amenable forms of utilitarian moral and legal theory offered by Austin, Mill, and Henry Sidgwick (Crimmins and Spencer, 2005).

John Stuart Mill

In his *Autobiography* (1873) Mill relates that when he first encountered Bentham's ideas in Dumont's edition of the *Traités* "I now had opinions, a creed, a doctrine, a philosophy; in one among the best senses of the word, a religion; the inculcation and diffusion of which could be made the principal outward purpose of a life" (Mill, 1963–91: vol. 1, pp. 67–8). Later, partly in response to the challenge of Macaulay's criticism of his father and partly the result of his own reflections on his intellectual inheritance, he developed a revised version of utilitarian theory. This revisionism was first announced to the world in the essay 'Bentham' (1838) and received definitive expression in *Utilitarianism* (1861). In the latter, Mill sought to defend the basic hedonism of utilitarian theory while offering refinements designed as responses to critics of the doctrine. The most problematic of these is the

distinction between the *qualitative* differences of types of pleasures, a view first enunciated in rudimentary form in Paley's *Principles*.

While adhering to the basic Benthamic analysis of motives, Mill introduced the concept of higher pleasures, by which he meant the pleasures of the intellect, which he claimed were intrinsically more desirable than other pleasures. However, this qualitative distinction undermined the aggregative dimension of the theory and introduced questions concerning the appropriate judge of which pleasures were preferable. Mill only raised further questions about the integrity of the theory by arguing that the constituent elements of an individual's happiness might encompass objectives, such as the pursuit of truth, beauty, and virtue, without consideration for the utility they may bring. This was a view later adopted by G.E. Moore (1873–1958) in the *Principia Ethica* (1903). Rejecting the hedonism of Bentham and Mill, Moore set forth a 'consequentialist' theory now known as *ideal* utilitarianism, in which the 'good' to be attained is more expansive than the summation of pleasures, and contains within itself certain objective features of the universe, preeminently the beauty of its constituent parts (Scarre, 1996: p. 114).

In other respects, Mill's utilitarianism remained within the Benthamic framework. To the criticism that utilitarianism sacrificed principles for expediency, he replied that it was expedient to observe the *rules* associated with general utility and implied that only saints could be thoroughgoing utilitarians, able to calculate continually the consequences of their actions. Like Bentham too, Mill gave priority to security in explaining the objectives of a system of justice. Finally, the supposed conflict between utilitarianism and Mill's commitment to individualism and individual liberty in the essay *On Liberty* (1859) is resolved when liberty is defended, not for its own sake but on the utilitarian grounds of the benefits, which accrue to the individual and to the society of which the individual is a member.

Individualism and Collectivism

Utilitarianism has enjoyed a sustained if ambiguous presence within the liberal tradition. Bentham's 'liberalism' is largely based on his writings on civil law and political economy, where he introduced subordinate principles to guide the legislator in giving effect to the utility principle: security, subsistence, abundance, and equality. The stress on the *security* of the individual and *security of expectations* gave a priority to individual liberty when crafting law and tended to introduce a conservative element into the legislator's reform agenda, while the egalitarian principle held out the prospect of redistributive policies. Which of the subordinate principles was most in accord with the dictates of utility at any one point, and therefore should be acted upon by the legislator, depended on the particular context and particular issue. To the 'security-providing' principle Bentham later added the 'disappointment-preventing' principle. Security required that legitimate expectations associated with property and place ought not to be disappointed, but where this is not possible (as in a time of crisis, such as famine or military defense) or when manifestly beneficial social improvements may be introduced without causing general concerns about the protection of security, then disappointed expectations (loss of public office as a result of necessary administrative reform or loss of land under a compulsory purchase order in the public interest, for example) must be compensated in the legislative process.

Bentham shared Adam Smith's view that each person is the best judge of their own interests. The route to a nation's prosperity is by individuals pursuing their interests in ways of their own choosing, and government intervention should be limited accordingly. However, Bentham also developed the theory of diminishing marginal utility, which dictated that one could have too much of a good thing; beyond a certain threshold the utility value of pleasures to an individual would begin to diminish. This furnished the legislator with an additional conceptual tool when considering the uneven distribution of social happiness and suggested policies to redistribute wealth and resources. The collectivist conclusions Bentham drew from this were modest in scope, but later economists, like W.S. Jevons (1835–82), developed the theory in the direction of modern welfare economics. F.Y. Edgeworth (1845–1926) and the Fabian socialists Sidney (1859–1947) and Beatrice Webb (1858–1943) were also important in applying utilitarian ideas to maximize and better distribute 'total utility' (see the relevant articles in Crimmins, 2013b).

Commentators who emphasize the importance of the subordinate principles that inform Bentham's theory of civil law and political economy tend to view him as an *individualist* or *liberal* utilitarian, who intended that laws should be modeled to facilitate individuals in the pursuit of happiness in ways they, rather than the legislator, thought appropriate (Kelly, 1990). This interpretation contrasts with the traditional view of Bentham as an aggregative legislative utilitarian. Mill is also interpreted as a *liberal* utilitarian, whose theory of justice stressed the utility value of *security* of the person and property (Riley, 1998). This placed definable limits on the functions of law and implied a gradualist approach to achieving optimal utilitarian outcomes through social and political reforms. Nevertheless, in a society of literate people neither Bentham nor Mill saw any barriers to the introduction of democratic institutions, including the extension of the suffrage to women on equal terms with men. Moreover, in later life Mill inclined toward collectivist solutions to social problems (Riley, 1996).

In *Man versus the State* (1884) and other writings, Herbert Spencer (1820–1903) employed the utility principle to enhance the liberty of the individual, defend the existing social order, and attack the drift toward socialism and 'slavery.' James Fitzjames Stephen (1829–94), Henry Sidgwick (1838–1900), and A.V. Dicey (1835–1922) all advocated versions of utilitarian individualism, although Sidgwick occasionally gave voice to 'socialist' sentiments in developing his *intuitional* utilitarian theory in *The Methods of Ethics* (1874). Reform-minded liberals such as J.A. Hobson (1859–1940) and L.T. Hobhouse (1864–1929) found in the protean nature of utilitarianism a justification for distinctly nonindividualist policies. Thus Hobhouse in *Liberalism* (1911) incorporated utilitarianism into a new liberal discourse that, in addition to laissez-faire economics (and notwithstanding Herbert Samuel's quasi-Benthamic motto 'the greatest liberty of the greatest number') also bridged elements of socialism, social

Darwinianism, and the idealism of T.H. Green (1836–82), and out of which emerged the beginnings of Britain's welfare state.

Criticisms

In the nineteenth century, utilitarianism was assailed from all sides: by sundry religionists, intuitionists like William Whewell (1794–1866), and idealists like Green, F.H. Bradley (1846–1924), Bernard Bosanquet (1848–1923), and D.G. Ritchie (1853–1903). If Marx considered utilitarianism an ephemeral bourgeois ideology, the latter critics combined to attack its atomism, crude materialism, narrowly construed theory of motivation, and lack of appreciation of the spiritual dimension of the human condition (Mill notwithstanding).

One of the most effective modern criticisms of utilitarianism is found in John Rawls' *A Theory of Justice* (1971). Rawls argued that in counting *only* interests and aggregating *across* persons to arrive at the appropriate action or policy the utilitarian fails to 'take seriously' the distinction between persons. In reply it could be argued that at the level of private ethics utilitarian calculations implicitly acknowledge that different and distinct persons are involved in most situations about which individuals have to make moral judgments. This means that a consistent utilitarian will treat the interests of different people affected by his or her actions as of equal importance. However, the problem with this reasoning is that it risks making utilitarianism primarily a theory of equal consideration and effectively relegates the aim of maximizing utility to the status of a byproduct, not the ultimate ground of the theory. This remains an unresolved tension in utilitarian theory. Critics of the theory, for example, frequently note the unpalatable sacrifices entailed by simple calculations of optimal utility to social problems; in an indiscriminate calculation where only total quantity of pleasures count, minorities are especially vulnerable to risk. Although Bentham was aware of the problems associated with the potential exclusion of minority interests under the phrase 'the greatest happiness of the greatest number', the literature on utilitarianism abounds with discussions of the potentially negative consequences of employing unqualified aggregative calculations of social utility. This debate is represented in the exchange between the modern act utilitarian J.J.C. Smart and his critic Bernard Williams in *Utilitarianism: For and Against* (1973). Many other dimensions of the history and theory of utilitarianism and the numerous debates associated with its ideas and their practical application are explored in *The Bloomsbury Encyclopedia of Utilitarianism* (2013).

See also: Bentham, Jeremy (1748–1832); Consequentialism (Including Utilitarianism); Eudaemonism; Field Theory; Hume, David (1711–76); Liberalism: Historical Aspects; Mill, John Stuart (1806–73); Pragmatist Social Thought, History of; Rawls, John (1921–2002); Spencer, Herbert (1820–1903); Weberian Social Thought, History of.

Bibliography

Baumgardt, D., 1966. Bentham and the Ethics of Today. Princeton University Press, Princeton, NJ.

Bentham, J., 1838–43. In: Bowring, John (Ed.), The Works of Jeremy Bentham, 11 vols. William Tait, Edinburgh.

Bentham, J., 1970. In: Burns, J.H., Hart, H.L.A. (Eds.), An Introduction to the Principles of Morals and Legislation. Athlone Press [The Collected Works of Jeremy Bentham], London.

Bentham, J., 1977. In: Burns, J.H., Hart, H.L.A. (Eds.), A Comment on the Commentaries and a Fragment on Government. Athlone Press [The Collected Works of Jeremy Bentham], London.

Bentham, J., 2001, 2010. In: Quinn, M. (Ed.), Writings on the Poor Laws, 2 vols. Clarendon Press [The Collected Works of Jeremy Bentham], Oxford.

Crimmins, J.E., 1990. Secular Utilitarianism: Social Science and the Critique of Religion in the Thought of Jeremy Bentham. Clarendon Press, Oxford.

Crimmins, J.E., 1996. Contending interpretations of Bentham's utilitarianism. Canadian Journal of Political Science 29, 751–777.

Crimmins, J.E. (Ed.), 1998. Utilitarians and Religion. Part I. Religious Utilitarians Part II. Secular Utilitarians. Thoemmes Press, Bristol, UK.

Crimmins, J.E., 2002. Jeremy Bentham's Auto-Icon, and Related Writings. Thoemmes Press, Bristol, UK.

Crimmins, J.E., Spencer, M. (Eds.), 2005. Utilitarians and Their Critics in America, 4 vols. Thoemmes Continuum, Bristol, UK, pp. 1789–1914.

Crimmins, J.E., 2011a. Utilitarian Philosophy and Politics: Bentham's Later Years. Continuum, London & New York.

Crimmins, J.E., 2011b. The principles of utilitarian penal law in Beccaria, Bentham and Mill. In: Koritansky, P.K. (Ed.), The Philosophy of Punishment and the History of Political Thought. The University of Missouri Press, Columbia, MO.

Crimmins, J.E., 2013a. Religion, utility and politics: Bentham versus Paley. In: Crimmins, J.E. (Ed.), Religion, Secularization and Political Thought: Thomas Hobbes to J.S. Mill, second ed. Routledge, London (Chapter 6).

Crimmins (Ed.), 2013b. The Bloomsbury Encyclopedia of Utilitarianism. Bloomsbury, London & New York.

Crimmins, J.E., 2014. Bentham and utilitarianism in the early nineteenth century. In: Eggleston, B., Miller, D. (Eds.), The Cambridge Companion to Utilitarianism. Cambridge University Press, Cambridge (Chapter 2).

Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Pantheon Books, New York.

Godwin, W., 1793. An Enquiry Concerning Political Justice, 2 vols. G.G.J. and J. Robinson, London.

Halévy, E., 1972. The Growth of Philosophic Radicalism [La Formation du Radicalisme Philosophique, 1901–4]. (1928. M. Morris, Trans.) Augustus M. Kelley, London.

Hutcheson, F., 2004. An Inquiry into the Original of our Ideas of Beauty and Virtue in Two Treatises, third ed. (1729, Wolfgang Leidhold, Ed.), Liberty Fund, Indianapolis.

Kelly, P.J., 1990. Utilitarianism and Distributive Justice: Jeremy Bentham and the Civil Law. Oxford University Press, Oxford.

Lieberman, D., 2013. John Austin (1790–1859). In: Crimmins, J.E. (Ed.), The Bloomsbury Encyclopedia of Utilitarianism. Bloomsbury, London & New York.

Mill, J.S., 1963–91. In: Robson, J.M. (Ed.), Collected works of John Stuart Mill, 33 vols. University of Toronto Press, Toronto and London vol. 1: Autobiography (1873); vol. 10: Bentham (1838), Utilitarianism (1861); vol. 18: On Liberty (1859); vol. 19: Thoughts on Parliamentary Reform (1859); Considerations on Representative Government (1861).

Riley, J., 1996. J.S. Mill's liberal utilitarian assessment of capitalism versus socialism. Utilitas 8, 39–70.

Riley, J., 1998. Mill on justice. In: Boucher, D., Kelly, P. (Eds.), Social Justice: From Hume to Walzer. Routledge, London.

Rosen, F., 2003. Classical Utilitarianism from Hume to Mill. Routledge, London & New York.

Scarre, G., 1996. Utilitarianism. Routledge, London.

Semple, J., 1993. Bentham's Prison: A Study of the Panopticon Penitentiary. Clarendon Press, Oxford.

Smart, J.J.C., Williams, W., 1973. Utilitarianism: For and Against. Cambridge University Press, London.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF

JAMES D. WRIGHT

University of Central Florida, Orlando, FL, USA

VOLUME 25

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

EDITOR IN CHIEF BIOGRAPHY

James D. Wright received his bachelor's degree in philosophy from Purdue University in 1969, a master's degree in sociology from the University of Wisconsin in 1970, and his PhD in sociology from Wisconsin in 1973. His first academic appointment was at the University of Massachusetts, Amherst, where he was promoted to Associate Professor in 1976 and to Full Professor in 1979. Most of his 15 years at Massachusetts were spent as Director, Associate Director, or Director of Research in the Social and Demographic Institute.

Wright left Massachusetts for Tulane University in New Orleans in 1988 where he served as the Charles and Leo Favrot Professor of Human Relations in Tulane's Department of Sociology. He also enjoyed a courtesy appointment in the Tulane School of Public Health, Department of Biostatistics and Epidemiology. He left Tulane in 2001 to become the Provost's Distinguished Research Professor of Sociology at the University of Central Florida. In 2003, he was named as the Director of the University's Institute for Social and Behavioral Sciences and in 2013, was designated as a UCF Pegasus Professor, the University's highest faculty honor.

Wright was the Editor in Chief of the scholarly journal *Social Science Research* from 1978 until 2014. His previous editing experience also includes a 20-year stint as editor of the Aldine de Gruyter book series *Social Institutions and Social Change*, two editions of the *Handbook of Survey Research* (Academic Press, 1983; Emerald Publishing, 2010), and service on the editorial boards of numerous journals. He is the author of 21 books and scholarly monographs on topics ranging from homelessness to research methods to NASCAR, and he has published more than 300 journal articles, book chapters, reviews, essays, and polemics. Wright lives with his wife Christine Stewart and their multiple dogs and cats in Winter Springs, Florida, where in his spare time he likes to cook fancy meals for large crowds.

This page intentionally left blank

PREFACE TO THE FIRST EDITION

Large and ambitious works such as the present Encyclopedia depend on countless instances of input, cooperation, and contextual support. Therefore, the editors in chief would like to express their gratitude to several institutions and individuals. Our most general thanks go to our respective home institutions, the Center for Advanced Study in the Behavioral Sciences, Stanford, and the Max Planck Institute for Human Development, Berlin. We are certain that without the effective infrastructures of these institutions and the rich collegial networks and intellectual climate they provide, implementing this Encyclopedia in such a short amount of time would not have been possible. In this context, we also need to mention that due to the administrative budgets provided to the editors in chief by the publisher, the financial strains on our home institutions were minimal. Such a situation may be a rarity in the modern world of scientific publishing where publishers often press scholars and institutions into taking on larger and larger shares of the publication enterprise.

Aside from the more than 4000 authors, our deepest thanks goes to the editors of the 39 sections, to whom we delegated many decisions at various stages, who were primarily responsible for developing the lists of articles and authors, and who stood as the main gatekeepers of scientific quality for the entries in their sections. As a group, the section editors displayed remarkable energy, intelligence, and tolerance for the inevitable frustrations of tending large numbers of authors over a long period of time. Because of the brevity of the acknowledgment of the main editorial coproducers here, we alert the reader to the list published as part of the front matter and the description of their extensive work we offer in the Introduction. We would also like to mention that one of our highly esteemed section editors, Franz Weinert, died unexpectedly during the last phases of his editorial work. For people who had the pleasure of knowing Franz Weinert, it comes as no surprise that he completed his editorial duties without ever complaining about the difficult health conditions he was facing. He was a gentleman and a distinguished scholar.

The International Advisors also gave wise counsel on several designated occasions, and a number of these scholars represented on the International Advisory Board took more initiative — always helpful — than we had originally asked. The International Advisory Board was particularly helpful in the process of choosing section editors. Some of its members assisted us ably as we attempted to make the Encyclopedia as international as possible and also in dealing with special problems that are part of a large project with close timelines. Thus, we remember a few occasions where members of the International Advisory Board helped us with their substantive and social competence to deal with matters of editorial disagreements. Because of the overall quality, collaborative spirit, and professional responsibility of our section editors, these events were rare indeed, but dealing with them required masterful and accelerated input. We appreciate the collegiality that members of the Board displayed when called upon in these special circumstances.

In addition, and as described in the Introduction, there was a host of esteemed colleagues who gave us advice on numerous questions such as author and editor selection, organizational matters, topics for entries, content and quality of selected articles, substantive and methodological niches to be covered, as well as last-minute author replacements. The number of such individuals is large indeed, and most likely, significantly larger than the list presented below. There were too many short, but nevertheless significant collaborative encounters that we likely did not deposit into our long-term memory or written documentation. We apologize for such oversights.

In this spirit, we appreciate the advice given in the three planning meetings in Stanford, Uppsala, and Dölln/Berlin by the following scholars: Peter Behrens, Burton Clark, Gordon Clark, Lorraine Daston, Meinolf Dierkes, Jean-Emile Gombert, Torsten Husén, Gérard Jorland, Ali Kazancigil, Wolfgang Klein, Gardner Lindzey, Renate

Mayntz, Andrew Pettigrew, Denis Phillips, Marc Richelle, Ursula M. Staudinger, Piotr Sztompka, Eskil Wadensjö, Björn Wittrock, and Robert Zajonc.

We would like to extend special thanks to Laura Stoker, who pinch-hit for Nelson Polsby (section editor for Political Science) who was unable to attend the meeting of section editors in April 1998; and to Linda Woodhead, Senior Lecturer in Religious Studies, Lancaster University, who assisted David Martin with editorial work at all phases for the section on Religious Studies.

We also acknowledge the help of scholars who advised Smelser and Baltes at the stage of pulling together the entry lists, proposing authors, and occasionally reviewing the content and quality of articles submitted.

Smelser thanks: Jeffrey Banks, Peter Bickel, Edgar Borgatta, Charles Camic, Jennifer Chatman, Jean Cohen, Michael Dear, Pierre Ducrey, Eckart Ehlers, Sylvie Faucheux, Steven Guteman, Frank Furstenberg, Helga Haftendorn, Peter Katzenstein, David Laibson, Stephan Lauritzen, Douglas McAdam, Eleanor Maccoby, Phyllis Mack, Cora Marrett, Douglas Massey, Donald Melnick, Harvey Molotch, Lincoln Moses, James Peacock, Trond Petersen, Andrew Pettigrew, Alejandro Portes, Matilda Riley, István Rév, John Roberts, Dorothy Ross, Rob Sampson, Fritz Scharpf, Melvin Seeman, James Short, Fritz Stern, Carol Swain, Ann Swidler, Ken'ichi Tominaga, Charles Tilly, Wolfgang van den Daele, Sidney Verba, Margaret Weir, Thomas Weisner, and Jennifer Wolch.

Baltes is very grateful to: Nancy Andreasen, Gerhard Arminger, André-Jean Arnaud, Jens Asendorpf, Jan Assmann, Margret Baltes (deceased), Jürgen Baumert, Peter Behrens, Manfred Bierwisch, Niels Birbaumer, Peter Bloßfeld, Walter Borman, Robert F. Boruch, Mark Bouton, Michael Bratmann, David Buss, Shelley Chaiken, Lorraine Daston, Juan Delius, Marvin Dunnette, Georg Elwert, Helmut Fend, Hans Fischer, Peter Frensch, Alexandra M. Freund, Dieter Frey, Jochen Frowein, Gerd Gigerenzer, Snait Gissis, Peter Gollwitzer, Ian G. Gotlib, David Gyori, Heinz Häfner, Helga Haftendorn, Giyoo Hatano, Adrienne Héritier, Theo Herrmann, Otfried Höffe, Ludger Honnefelder, Paul Hoyningen-Huene, James Huang, Günther Kaiser, Heidi Keller, Martina Kessel, Gábor Klaniczay, Wolf-Hagen Krauth, Achim Leschinsky, Karen Li, Shu-Chen Li, Ulman Lindenberger, Elizabeth Loftus, Gerd Lüer, Ingrid Lunt, Hans J. Markowitsch, Laura Martignon, Randolf Menzel, Dietmar Mieth, Susan Mineka, Setsuo Miyazawa, John R. Nesselroade, Claus Offe, Vimla Patel, Meinrad Perrez, Rosa Lynn Pinkus, Robert Plomin, Neville Postlethwaite, Thomas Rentsch, István Rév, Peter Roeder, Richard Rorty, Hubert Rottleuthner, Peter Schäfer, Heinz Schuler, Norbert Schwarz, Richard J. Shavelson, Joan E. Sieber, Burton Singer, Wolf Singer, Edward Smith, Hans Spada, Günter Spur, Rolf Steyer, Michael Stolleis, José Juan Toharia, LeRoy B. Walters, Elke U. Weber, Peter Weingart, and Reiner Wimmer.

There was one more group that played a special role in the editorial review (see also Introduction). Baltes was assisted by a team of colleagues who provided expert input to the section editors and to him during final review: Gregor Bachmann, Alexandra M. Freund, Judith Glück, Wolfgang Klein, Olaf Köller, Shu-Chen Li, Ulman Lindenberger, Ursula M. Staudinger, and Christine Windbichler. Their expertise in helping us to ensure quality is gratefully acknowledged.

The day-to-day organizational work of the editors in chief occurred at their respective research centers — the Center for Advanced Study in the Behavioral Sciences and the Max Planck Institute for Human Development in Berlin. Smelser singles out Julie Schumacher, his main assistant in the project, for her organization of the meeting of section editors, and her superb coordination of the flow of correspondence and manuscripts over several years. She remained firm when the rest of us were faltering, and displayed the greatest efficiency, intelligence, and good cheer. Smelser also relied on periodic help from Leslie Lindzey, Kathleen Much, Jane Kolmodin, and Anne Carpinetti. Michelle Williams of the University of California, Berkeley, served as his main editorial assistant for nearly 2 years, going over all entry manuscripts from the standpoint of readability, and carrying a great part of the cross-referencing work; her judgment was always the best, and Smelser is forever indebted to her.

At the Max Planck Institute for Human Development, and aside from Julia Delius (see below), Baltes expresses his deeply felt thanks to the main secretarial staff in his office (Helga Kaiser, Romy Schlüter, Conor Toomey) who during the preparation of this Encyclopedia took on larger shares of responsibilities for other projects and occasionally helped out when the editorial office was overloaded. More directly involved in the day-to-day operation of the Encyclopedia was Penny Motley, who efficiently functioned as project secretary during the planning phase. During the main phases of the project, it was Yvonne Bennett who was responsible for secretarial and organizational matters. She deserves much applause for her superb and efficient help in the day-to-day running of the editorial office, making sure that all manuscripts and proofs were processed quickly, and also for assisting with the organization of various meetings. Not least, Baltes thanks the administrative office of the Max Planck Institute, and there especially Nina Körner and Karin Marschlich, for their competent work in managing the project budget, as well as Sabine Norfolk who handled most of the numerous fax

transmissions with ever present smiles and friendliness. Finally, in the Berlin editorial office, thanks are also due to Susannah Goss for her excellent work in translating individual articles into English.

Individual section editors wish to thank the following scholars who gave them substantive advice in their work: Richard Abel, Itty Abraham, John Ambler, Karen Anderson, Mitchell Ash, Alan Baddeley, Boris Baltes, Manfred Bierwisch, Michael Bittman, Sophie Bowlby, John Carson, Shelly Chaiken, Roger Chartier, François Chazel, Stewart Clegg, Carol Colby, Philip Converse, James Curran, Jerry Davis, Natalie Z. Davis, Peter Dear, Juan Delius, Michael Dennis, Josh DeWind, Sherry Diamond, Elsa Dixler, Georg Elwert, Howard Erlanger, Drew Faust, Malcolm Feeley, Nancy Folbre, Michael Frese, Angela Friederici, Lawrence Friedman, Bryant Garth; Victor Ginsburgh; Marcial Godoy, Frances Goldscheider, Calvin Goldscheider, John H. Goldthorpe, Reginald Golledge, William Graziano, Stephen Gudeman, Mauro Guillen, Doug Guthrie, John Hagan, Eric Hershberg, Steve Heydemann, Stephen Hilgartner, Judith Howard, William Howell, Jill Jaeger, Bob Kagan, Joe Karaganis, Roger E. Kasperson, Ronald Kassimir, Robert W. Kates, Wolfgang Klein, Hans-Dieter Klingemann, Bert Kritzer, Kay Levine, Steven Lukes, Michael Lynch, Akin Mabogunje, Stuart Macaulay, Donald MacKenzie, David Magnusson, Michael McCann, Sally Merry, Peter Meusburger, Sieglinde Modell, John Monahan, Kevin Moore, Gerda Neyer, Hiroyuki Ninomiya, Jodi O'Brien, Eva Oesterberg, Mark Osiel, James L. Peacock, Susan Phillips, Trevor Pinch, Wolfgang Prinz, Sheri Ranis, Harry Reis, Estevao C. de Rezende Martins, Marc Richelle, Leila Rupp, Gigi Santow, Austin Sarat, Simon Schaffer, Linda Scott, Tim Shallice, Seteney Shami, Ronen Shamir, Susan Silbey, Wolf Singer, Sheila Slaughter, Gerhard Strube, Ursula M. Staudinger, J. Stengers, Stephen M. Stigler, Michael Stolleis, Mark Suchman, Denis Szabo, Verta Taylor, Ashley Timmer, David Trubek, Leslie Ungerleider, Don Van Arsdol, Jakob Vogel, Rita Vörg, Judy Wajcman, Mary Wegner, Wlodzimierz Wesolowski, Steve Wheatley, Björn Wittrock, and Vincent Wright (deceased).

Section editors also thanked the following persons for their research assistance: Pamela Anderson, Susan Augir, Anja Berkes, Chantale Bousquet, Aida Bilalbegović, John Clark, Susanne Dengler, Kathrine Derbyshire, Annie Devinant, Casey B.K. Dominguez, Barbara Dorney, Elizabeth Dowling, Tracey L. Dowdeswell, Carolyn Dymond, Debbie Fitch, Susannah Goss, Verhan Henderson, Andrew Hostetler, Gudrun Klein, Heike Kubisch, Angie Lam, Ellen Lee, Valerie Lenhart, Kay Levine, Allison Lynn, Helena Maravilla, Carol B. Marley, Marion Maruschak, Michael McClelland, Rhonda Moats, Birgit Möller, Katja Neumann, Linda Peterson, Justin Powell, Paul Price, Xandra Rarden, Chris Reiter, Deborah Sadowski, Heidi Schulze, Heidi Sestrich, Keith Smith, Judith Thompson, Karen Varilla, Stuart Vizard, Danelle Winship.

It is difficult to single out one person as especially helpful. Despite the risks involved, we would like to highlight one person (see also Introduction). As shown in the front matter, beginning in the second year of the planning phase, Julia Delius served as our scientific editorial assistant. For our internal process of editorial review and management, Julia Delius was the mastermind of editorial coordination. Her work was simply outstanding.

Finally, we express our thanks to several individuals at Elsevier publishers whose work was both essential and helpful to us throughout. Barbara Barrett started the ball rolling for the entire Encyclopedia project, and was a master at negotiating out the fundamental arrangements at the beginning. Geraldine Billingham, as executive editor, also played an important role in the development of the project, as well as carrying the heavy responsibility for ensuring that all the financial and operational aspects of the project were realized. She executed this work with directness, skill, and tact. Angela Greenwell coordinated the editorial process with dexterity, calmness, and good judgment. Three persons were responsible for monitoring the progress of entries from Elsevier's side: Michael Lax for a time at the beginning; Jayne Harrison, who carried an overwhelming load of manuscripts month after month, and Helen Collins, who was responsible for the equally demanding job of shepherding the manuscripts through the entire production process. Ruth Glynn skillfully and imaginatively masterminded the development of the electronic version within Science Direct. We know how much the success of the Encyclopedia has depended on these individuals and their staffs, and our appreciation and admiration is here recorded.

Our expressions of gratitude and respect for the authors, section editors, reviewers, counselors, and editorial team, however, are not meant to distract from our own responsibilities as editors in chief. Let us hasten to add, therefore, that those whom we acknowledge deserve most of the credit but little of the blame for whatever shortcomings remain.

Neil J. Smelser
Paul B. Baltes
Palo Alto and Berlin
June 2001

This page intentionally left blank

INTRODUCTION

Alan Sica's article in this encyclopedia observes, "The urge to systematize and rationalize both common and esoteric knowledge has been strongly and continuously in evidence for at least 2400 years in the West" (Sica, 2015). Sica dates the origins of this urge to Plato's nephew Speusippus, who took up the challenge of Plato's call for an *encyclopedia* (from Greek words meaning 'general education'), and traces the urge though Pliny the Elder, Vincent of Beauvois, Francis Bacon, Denis Diderot, and a great many others. Today, an encyclopedia is simply a self-contained reference work "with two aims: to include up-to-date knowledge about a particular discipline or group of disciplines and to make this knowledge conveniently accessible" (Smelser and Baltes, 2001). Such are the aspirations of the present work.

Histories of the concept of an encyclopedia are widely available and need not occupy us here. Those interested in where the *International Encyclopedia of the Social and Behavioral Sciences* (IESBS) fits into the long history of encyclopedias will do well to consult the Sica article quoted above or the Smelser–Baltes 'Introduction' to the first edition of this work.

This is the second edition (2/e) of the IESBS. Elsewhere, I acknowledge my immense debt to the editors of the first edition, Dr Neil Smelser and Dr Paul Baltes. Without their perspective and template to follow in assembling the second edition, this version would have scarcely been possible. The very thought of an encyclopedia covering *all* the social and behavioral sciences, one that drew on scholars from *all* corners of the globe, is itself so audacious that one would hesitate even to begin without a pretty detailed road map to the terrain.

On the occasion of the publication of Robert E.L. Faris's widely acclaimed *Handbook of Modern Sociology*, Joseph Gittler was moved to ask, "With the plethora of new handbooks, reviews, and compendia in sociology, one wonders whether there is a need for yet another one." That question was posed almost 50 years ago; the ensuing half-century has been marked, as Sica says, by "an avalanche of reference materials the likes of which had never been seen in libraries prior to about 2000." So readers and critics are within their rights to ask: Why a second edition? Why now?

The first edition (1/e) was reviewed in *Contemporary Psychology* as "the largest corpus of knowledge about the social and behavioral sciences in existence" (Park, 2004). According to readily available data sources, IESBS 1/e has been cited nearly 6000 times in subsequent scholarly literature and has an h-index ($h = 24$) that would be the envy of many full professors. Citations are only one measure of impact, of course, and probably not the most important measure for a basic reference work intended more for students and a general audience than for practicing professionals. In addition to the thousands of scholarly citations, items from IESBS 1/e were 'hit' almost two *million* times on Science Direct between online publication in November 2002 and August 2014 — two million instances where IESBS 1/e was likely consulted as a basic reference source, to get quickly up to speed in a new area of research or to provide background for a paper or project, or just to satisfy a reader's curiosity. The first edition, in short, emerged soon after publication as a go-to reference work throughout the social and behavioral sciences. Michael Shanahan referred to the first edition as "the atomic bomb of reference works." It was thus an easy conclusion that the work had sufficient value to keep it alive and current.

In reading selected articles from the first edition, however, I was taken aback by how much and how quickly they had aged. The first articles I looked at were those pertinent to my own areas of research and in nearly every case, there was important recent work that I felt needed to be discussed. With the assistance of colleagues in other areas of the social and behavioral sciences, it became obvious that this was a general problem. To be sure, much of the material in 1/e has enduring value; indeed, roughly half the articles in 2/e are updated and revised items from the first edition. Still, it is sobering to realize just how quickly the frontiers of social science

knowledge shift and thus how quickly a basic reference source becomes dated. I am certain there are articles in the current edition that are already in need of some updating.

Smelser and Baltes enumerated six criteria that guided their decisions about what to include and what to leave out. The criteria are as follows:

1. secure knowledge, realizing that security of knowledge is a dynamic and relative term,
2. knowledge with balance and comprehensiveness,
3. knowledge that is integrative rather than fragmented,
4. knowledge that places the evidence into historical and theoretical context,
5. knowledge that highlights connections between topics and fields, and
6. knowledge that combines, where possible, theory and practice.

I would like to say I have followed these same criteria in determining the content of 2/e, but all six criteria have proved problematic in one or another way and thus deserve some comment in light of the present work:

1. What is secure knowledge today is apt to be recognized as wrong tomorrow. Many of our requests for updates to 1/e articles generated some response to the effect, "All of this is completely obsolete. We need to write a new entry from scratch." One would expect this to be true of areas of research such as behavioral genetics, say, or cognitive neuroscience or other fast-moving scientific fields, but it also turns out to be true of archaeology, history, and philosophy. The leading edge of research in *every* discipline threatens received wisdom; those who content themselves with received wisdom ('secure knowledge') will quickly find themselves behind the times. So our charge to authors was to focus less on the 'secure knowledge' in their respective fields and more on the cutting edge — the new perspectives, methods, findings, and issues that will animate social and behavioral science research in the coming decade.
2. Throughout the social sciences, the distinction between theory and ideology is sometimes difficult to discern, and so too the distinction between applied research and community or political activism. We need not invoke Max Weber's conclusions about the impossibility of a truly value-free social science to understand that ideological undertones and political preferences can often be detected in social science research work; sometimes, they are quite explicit. Nor is it much of a surprise to learn that many social scientists struggle, not always successfully, to distinguish between what they want the world to be and what the world in fact is. Such struggles are part of the process by which knowledge develops and are thus in evidence in many of the 2/e articles. It would be unwise to enforce by editorial fiat a consensus on issues that the social sciences themselves have yet to resolve.
3. The fragmentation of knowledge *across* disciplines is far more serious that the fragmentation of knowledge *within* them, although the latter is a problem as well. Many well-informed sociologists, for example, take pride in reading widely and can speak with some authority about what is going on in the field, even in areas of sociology that do not directly concern them. I assume this is also true of the other social and behavioral science disciplines and areas. But rare is the sociologist who is cognizant of work being done, say, in social geography *even when*, as is often the case, the fundamental problems, theoretical issues, and methodological approaches are similar. This is a widespread problem across the disciplines: The evolutionary psychologists do not know as much as they should about what the anthropologists are doing; the economists ignore a lot of pertinent work in sociology and political science; etc. For better or worse, disciplinary differentiation is an inherent element in the process of knowledge development and discovery, so this is a reality to be faced, not a problem to be solved. We dealt with this reality mainly in two ways: (1) Authors were encouraged wherever possible to think across rather than strictly within their disciplines. (2) Calling relevant work done in other disciplines to the reader's attention was a guiding principle for the choice of cross-references for each article.
4. Social-constructionist theories of knowledge would argue that knowledge does not become knowledge until it is placed in historical, cultural, social, political, and theoretical context. As put very simply and succinctly in the opening line of the famous text by Berger and Luckmann, "reality is socially constructed" (1966: 1). This is no less true of social science knowledge than knowledge in any other realm. Unfortunately, a great deal of what we think of as social science knowledge is based on biased samples of WEIRD people — that is, Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). (The same point has, of course, been made by many others.) The WEIRD, thus defined, represent a rather exotic minority group judged by world standards, and the challenge for IESBS 2/e was therefore twofold: (1) to recruit non-WEIRD editors and authors where possible; and (2) to persuade all authors to incorporate comparative, cross-cultural, and

international themes and perspectives wherever it was possible to do so. This effort was more effective in some sections than in others. Moving away from nation-specific literature, themes, and results to a truly *international* (the I in IESBS) account of the social and behavioral sciences was an ever-present *desideratum* in my discussions with editors and authors.

Alas, an emphasis on cross-culturalism often brings to light that everything, or nearly everything, seems to vary by culture, subgroup, ethnicity or nation, and at that point the question arises whether social science generalization is even possible. Are there universals that cross all social groupings? Or does social science dissolve into a mist of group-specific particularities that adduce no general principles? That hardly seems desirable. Yet myopic and ethnocentric fixations on one's own society or culture are scarcely to be encouraged either.

The worst offenders in this regard tend to be social scientists from the USA (and from the West more generally), who, of course, also comprise the largest share of editors and authors for IESBS 2/e. True, as with greenhouse gases, the USA (and the West) also produce the largest share of the world's science, social and otherwise, and so it is inevitable that they also represent the largest share of IESBS 2/e contributors. The Swedish social anthropologist Ulf Hannerz wrote about this very issue in his marvelous essay, "Editing Anthropology" (Hannerz, 2010). Hannerz quotes a Canadian anthropologist who said of the "neighborly presence" of US anthropology that it was like being "in bed with the elephant." The same is true of most or all of the disciplines covered in this work. But then in a passage remarkable for both its insight and its generosity, Hannerz remarks, "No doubt tensions and conflicts may arise out of this situation, but perhaps it is better to understand that the elephant cannot help being an elephant, and that it makes sense for it, in its habitat, to behave in elephantine ways. A well-trained elephant, too, can do a lot of good, such as in carrying heavy loads." So if the elephant that is Western social and behavioral science has been pressed to carry the heavier load in these pages, that is, the price to be paid for being an elephant in the first place.

The larger issue here is that the social and behavioral sciences, which were largely birthed in the West, in Europe and the United States, mainly address a trans-Atlantic scientific dialogue, a dialogue that is carried out almost exclusively in English. Indeed, throughout much of the world, a scientist's stature is indexed in part by his or her success in publishing in English. Publications in the native language are simply *assumed* to be inferior. The main exceptions, tellingly, are French and German social scientists whose native-language publications are valorized. A Dutch or Portuguese or Hungarian social scientist who only published in the native language would be seriously punished in merit and tenure reviews. Elsevier's citation database Scopus *requires* English-language abstracts before a journal will even be considered for inclusion. And English (with American spellings) is likewise the official language of this encyclopedia. No resources were made available to get articles submitted in other languages translated (although several such articles were submitted). Inevitably, this biases the encyclopedia's coverage in favor of social science addressed to the trans-Atlantic dialogue. Consequently, a great deal of excellent social science published in Spanish, or Mandarin, or Farsi, or any of the other world languages, is overlooked.

That said, steps were taken to improve the internationalism of the encyclopedia's coverage throughout the 4 years of its development. To begin, I sought area and section editors from around the globe and urged them whenever possible to seek international (vs strictly Western) authors. Articles were often returned to their authors with explicit instructions to add international coverage. And when all efforts failed and we ended up with articles that were unmistakably restricted to a single nation in their coverage (frequently but not always, this was the US), I indicated the national focus in the article subtitle. Still, a great deal more WEIRD social and behavioral science remains in these pages than would be desirable in an *international* encyclopedia. In a later section, I give the national breakdown of everyone involved in this project.

5. An encyclopedia that "highlights connections between topics and fields" is easier to envision than to accomplish. Criterion (3) above speaks to the disciplinary fragmentation of our social science knowledge. William Easterly, the NYU economist, has written about the 'tyranny of experts' but the tyranny of the disciplines is an equal problem, a tyranny sanctified by the very academic structure of the modern university, where the separate and often warring colleges are each comprised of assorted departments, the latter representing basic academic disciplines. Interdisciplinary programs are almost invariably relegated to various centers, institutes, or specialized research laboratories — in short, to the periphery of the university's power structure. And while everyone *says* they want to encourage multi- and interdisciplinarity, very little is done to make that happen. Instead, departments remain as the fundamental unit of academic organization.

   Hortatory admonitions to editors and authors have made IESBS 2/e somewhat more interdisciplinary than its several predecessors, but in the end, the disciplinary emphasis in many of our articles would be hard

to overlook. Still, there are many sections of the work that were clearly developed with interdisciplinarity in mind and in which one finds interdisciplinary social science at its current best: environmental and ecological sciences; health; life course; sexuality; GLBT studies; public policy; evolutionary sciences; war, peace, violence, and conflict; and genetics, behavior, history, and society are a few sections that come immediately to mind.

I am particularly pleased with the extensive coverage given in 2/e to the interactions between the biological and social sciences. Harry Whitaker once remarked in a private e-mail that "we are all witnesses to the marriage of biology and the social sciences." Many of the articles found here are offspring of that marriage. And those articles leave little doubt that the future offspring of this marriage will be among the most exciting and insightful social science products of the coming decade.

The social sciences once resisted everything biological as rank reductionism; in addition, the social scientists have always feared that a mature and complete biology would eventually drive us out of business. Articles included in IESBS 2/e make it clear that 'nature' versus 'nurture' was the wrong question from the very beginning, that what we need to understand is how biological inheritance and social and cultural forces *interact* to produce individual and collective outcomes. This is where a lot of the interesting action in the social sciences is going to be in the coming years.

6. The coverage of the practical or applied sides of the various disciplines has been increased in this edition, so the applied content of 2/e certainly exceeds that of its predecessor. New to this edition is the entire section on applied social and behavioral sciences; also new is our large section on social work and the separate section on applied, industrial and organizational psychology. But as Sara Strickhouser and I say in our article on the history of applied social research, "The history of 'applied social science' is indeed the history of the social sciences themselves" (Strickhouser and Wright, 2015). All of the social sciences in their modern form were birthed in the post-Enlightenment effort to understand why people, their institutions and society as a whole were as they were — why greed was more visible than generosity, why oppression was more common than freedom, why inequalities of various sorts characterized every social institution. To the founders of the disciplines, the modern distinction between basic and applied scientific concerns would have seemed meaningless, as indeed it was.

The Planning Phase and Editorial Structure

Elsevier's original acquisitions editor for this project, Scott Bentley, discussed the broad outlines of our attack on IESBS 2/e with me in a series of phone calls in the early months of 2010. By May 2010, certain of these outlines had come into sharp relief. There would be a single editor in chief with a tier of 'responsible lieutenants' to assist. These lieutenants started out as advisory editors but soon morphed into the 10 area editors in the final editorial structure. Each area editor would in turn oversee various sections and each section would have its own editor or editors. The original plan was that the area editors would each be responsible for approximately 400 articles and each section editor responsible for about 100 articles, so that no one would have an excessive amount of work compared to everyone else. In the end, however, some sections required more articles and others fewer, so the areas vary from fewer than 300 to more than 600 articles, and the sections vary from fewer than 30 articles to well more than 200.

Prior to my first discussions with Scott Bentley, Elsevier had commissioned about 20 'market surveys,' roughly half from professional librarians and the other half from academics, each focused on 'what was right and what was wrong' with the first edition. The most insistent demand that surfaced was for a fully electronic, computer-searchable work either in addition to or instead of a print version. In the end, both an e-version and a print version are being made available. But there were also consistent concerns expressed about spotty coverage of some areas: no coverage of social work to speak of; inadequate coverage of accounting, marketing, and business; no separate section of criminology; haphazard coverage of applied topics and concerns; and so on. And there were the predictable concerns about undue Western domination of the work.

Working from the 1/e Table of Contents and this stack of 20 surveys, I tinkered for a few months with how best to expand coverage to address the concerns our reviewers had expressed; the result was (more or less) the editorial structure shown elsewhere in the front matter of this work. Thus, the 53 sections of 2/e are basically the 39 sections of the first edition plus sections that were added to cover important emerging areas of research interest (e.g., gay, lesbian, bisexual, and transsexual studies) or to provide deeper coverage of disciplines and subdisciplines that the first edition treated lightly if at all (applied social and behavioral sciences, social work). A

few of the first edition sections were eliminated entirely: specifically, integrative concepts and issues (something of a catch-all category for 1/e articles that were highly interdisciplinary or simply did not fit anywhere else) and gender studies (gender is now so thoroughly integrated into the research and thinking of all the social and behavioral science disciplines and subdisciplines that a separate section was unnecessary and somewhat retrograde). Others were altered or expanded in significant ways (organizational and management studies, for example, became the second edition section on management, organizations, business, marketing and finance; expressive forms evolved into the 2/e section on culture and the arts; linguistics was subdivided into three separate sections; etc.) And some subdisciplinary areas that were part of other sections in the first edition were given their own sections in the second (life course studies; sexuality; criminology; labor studies; and war, peace, violence, and conflict are the most prominent examples), just to ensure adequate coverage.

The Introduction to 1/e (Smelser and Baltes, 2001) discusses at great length the intellectual architecture of the first edition and its evolution; the agonizing decisions about rules for the biographical articles; whether to organize the work by discipline, by concept, by an a priori taxonomy of social science knowledge, or in some other manner; and related structural matters. What resulted in 1/e was something of a hybrid model with some sections clearly representing specific disciplines, others representing 'overarching topics,' still others representing 'intersecting fields,' and so on. My task was far less challenging: I thought the architecture of the first edition served its purposes admirably and so I adopted it wholesale for the second edition, with the small modifications noted above.

While each of the 53 sections of 2/e has some more or less obvious justifying logic or substantive or disciplinary coherence, their arrangement into the 10 areas of the second edition does not. As already indicated, decisions about what sections to include and exclude were made very early in the development of the second edition, based on what was in the first edition, what Elsevier's marketing surveys suggested as significant lacunae, and my own sense of things. The arrangement of the sections into areas came later and was dictated largely by the interests and expertise of those who had been persuaded to sign onto the enterprise. Thus, all the Linguistics sections are in Area Five along with the cognitive and neuroscience sections because Harry Whitaker at Northern Michigan University had been a linguist first, a neuroscientist second, and a historian of psychology third. Likewise, Area Eight contains its seemingly chaotic hodgepodge of evolutionary sciences, genetics, religious studies, science and technology studies, and law because these happen to be the fields of interest of the redoubtable Barbara Prainsack of King's College, London, whose editorial talents are as deep as her interests are broad. Simplifying only slightly, section editors were chosen for their substantive expertise in particular areas of research (as well as their international or cultural 'range'); the plan was that section editors would review each of their articles for completeness of coverage, currency, fairness, and the like. In contrast, area editors were chosen for their breadth of interest and knowledge spanning multiple sections – people who by nature, experience, and training would be good at seeing the big picture. The role of the area editors was to assure that the articles were at least modestly international and interdisciplinary and formed some sort of coherent whole.

As laid out in an early e-mail to one of the area editors, "In my mind, Advisory [Area] Editors are generalists and Section Editors are content specialists. The Section Editors need to bring breadth of knowledge to their responsibilities, of course, but the primary job requirement is expert understanding of the particular field or section. We need section editors who can commission pieces that adequately review past developments in the subject area, but who can also shape their portfolios to identify major unanswered questions and new research developments." This was the general model followed throughout the recruitment process.

The approach just sketched inevitably left a residual area made of up sections that none of the other area editors were willing or able to claim. That proved to be Area Ten. A very large number of scholars from various disciplines and world regions were approached about serving as the Area Ten Editor. All demurred until I approached Martin Bulmer of the University of Surrey, who served the project for about a year but was unable to continue and subsequently left the Board. So for the final year of the project, the editing of Area Ten was done mainly by me.

The online subject classification for IESBS 1/e did not correspond precisely to the sections of the first edition; the subject classifications listed articles by broad general topic regardless of which 1/e section developed them. So for the first year or so, my assistants and I labored to assign every 1/e topic to one of the 53 2/e sections. With 53 sections to populate with articles and about 4000 articles to classify, the process of assigning articles to unique sections occupied huge amounts of time and generated endlessly complicated spreadsheets. (After about a year of this, a document surfaced in the Elsevier archives that showed exactly where each 1/e article had originated.)

Along with populating the 2/e sections with articles, searching for area and section editors occupied a major part of the early going. The original idea was to identify the 10 area editors first and then enlist them in the

search for their section editors, but in fact, these processes were simultaneous and were about half-completed when the two planning meetings were held: the first in San Diego on 19 and 20 May 2011; the second in London on 5–7 June 2011. Although not part of the plan, the first meeting was mostly attended by area editors and the second, by section editors.

The agenda for both meetings was largely identical: achieve clarity on the respective roles of the various editors; solicit advice on filling the unfilled editorial slots; get buy-in on the project *mantra* (internationalism, interdisciplinarity, inclusiveness!); explain the relationship between the first and second editions (e.g., that all 1/e authors would be offered the opportunity to update their articles); explain the process of assigning articles to sections; and the like. In almost all cases, consensus was quickly forthcoming.

The largest substantive change in the second edition is the treatment of psychology. (Whether measured by students, practitioners, professors or programs, psychology is the largest of the social and behavioral sciences.) In the first edition, the whole of psychology was contained within four large sections: clinical and applied psychology; developmental, social, personality, and motivational psychology; cognitive psychology and cognitive science; and behavioral and cognitive neuroscience. Cognitive science and neuroscience are, of course, rapidly growing sectors of the psychological discipline and it seemed appropriate to devote an entire area of the second edition to them. This area contained seven separate sections (three of them devoted to linguistics). And the 'softer' psychological sciences were placed in a separate area comprised of eight sections (including psychiatry and social work). These changes were made to assure that 2/e had adequate coverage of the entire psychological enterprise (a concern that was raised in some of the reviews of 1/e).

The structure of the editorial team was developed to serve two purposes: first, to divide up the labor into manageable parts; and secondly, to assure that every article in the final product would be reviewed by at least three scholars: the respective section editor, area editor, and the editor in chief. Since we were all aware of one another's identities throughout, the review process was not blinded, but every article in this work has been thoroughly peer-reviewed and vetted. In many cases, these articles have been through four, five, six, or even more revisions before being accepted for inclusion in the work.

Recruiting Editors and Authors

IESBS 2/e faced two different recruiting tasks: recruiting editors and recruiting authors. As indicated elsewhere, the original plan was that I would recruit area editors largely from my own personal networks and they in turn would recruit their section editors. In practice, the process was much more convoluted. First off, many of my 'first choice' candidates were hesitant to involve themselves in such a daunting and ambitious project. In some cases, an initial agreement to come aboard was followed a few weeks or months later by withdrawal as the true scope of work came into sharper focus. Some people that I tried to recruit as area editors demurred but agreed to serve as section editors; in other cases, candidates for area editors were nominated by section editors. In a couple of cases, there was a complete team of section editors prior to the successful recruitment of a corresponding area editor. And so on. The process of assembling an editorial team for the project was by no means linear.

Personal networks – mine and those of the people I persuaded to sign on – were the largest single source of ideas for other editors, by far. But personal networks tend to follow disciplinary and national lines and become rather quickly exhausted. In selecting both editors and authors, it was of course essential that all relevant perspectives be covered and that the breadth of fields was adequately represented. So other sources of ideas were tapped: the expertise of the Major Reference Works team at Elsevier; editorial boards of leading journals in the respective fields; Google Scholar citation counts; the knowledge and networks of people who had been involved in preparing the first edition – basically, anywhere names of top scholars in particular areas could be located. I did not keep systematic track of the number of people approached, much less of the outcomes; but in reviewing my e-mail archives from the early period of the project, I probably asked (on average) somewhere between six and eight prospects before I found someone to agree (an average that was highly variable across the areas and sections).

In the initial plan, area editors would recruit their own section editors and the section editors would in turn recruit their own authors, but the reality was again different. Inevitably, my negotiations with prospective editors (both types) would often result in a declination of the editorial role but an offer to write an article on a particular topic (or topics), offers I readily accepted. So in some cases, I had committed authors of particular articles well before I had the corresponding section or area editor in place. And many of the section editors likewise leaned on their area editors and on me for author suggestions, contacts, and outreach. So all the editors

were involved in locating willing authors (and not coincidentally, in making suggestions for articles to be included in the final work).

No one was asked to keep precise statistics on how many authors were sought before an agreeable author was found. My sense is that the 2/e editors had a somewhat tougher time with this than the 1/e editors had (but on the other hand, the 2/e editors had all the 1/e authors to work with, which Smelser and Baltes certainly did not have). The 1/e editors report that about 60% of those invited to write articles agreed to do so on the first 'ask' (with equivalent success rates in each subsequent 'ask'). In 2/e (or so I would guess), the success rate was probably closer to 50% and possibly even lower. They also report that something in excess of 90% of those who agreed to write for 1/e in fact contributed "acceptable entries in time to be included…" In the 2/e case, we found ourselves still soliciting authors within a few weeks of the submission deadline and some dozens of authors pulled out of the project at the last minute. Still, when all the dust settled, there were just about 800 articles planned for both editions that never materialized and both editions ended up with just fewer than 4000 articles. Perhaps these are 'natural limits' for a work of this sort.

Biographies*

In the Introduction to 1/e, Smelser and Baltes provided a short history of and rationale for the decision to include a limited number of biographical articles (150 of them, to be exact). This was evidently a controversial and contested decision since some of those associated with 1/e felt that biographical articles had no place in an encyclopedia focused on substantive knowledge (versus the personalities of those who produced it). As Newton said (paraphrasing a passage from Bernard of Chartres), "If I have seen further, it is by standing on the shoulders of giants" (a passage made famous among social scientists by Robert K. Merton's *On the Shoulders of Giants* (1965)). In other words, discovery is only possible because its discoverer is able to refer to knowledge already gathered and accumulated by his or her predecessors; and since all discoveries – small and large, substantial and minor – contribute to the production and accumulation of knowledge, singling out particular scholars for special attention would be both pointless and invidious.

This position, although understandable, was abandoned in the first edition and remains so in the second, largely because it is not possible to understand the social sciences without knowing the individuals who contributed to it. C. Wright Mills once defined the sociological imagination as the intersection of biography and history (Mills, 1959). To know the social sciences therefore requires biographical and historical knowledge. The social sciences are distinguished from the physical and biological sciences in the way they produce knowledge because individual human beings, including social scientists, invest meaning in their actions. Only through understanding the meaning that individuals give to their actions are social scientists able to provide explanations of human behavior and its outcomes. And what applies to social action in general applies to the production of social science knowledge as well. It thus makes sense to include biographical accounts of social science personages who have been particularly distinguished in making sense of human action and have in turn made major contributions to social science knowledge. To acknowledge individual social scientists with biographical articles is also to acknowledge their special and particular insights.

Once the decision was made to include biographies in 1/e, the next major decision was whom to include and whom to leave out. Two arbitrary but defensible principles governed these decisions: (1) the total number of biographies was not to exceed 150 (chosen because it seemed like a manageable number) and (2) only biographies of the dead would be included. The 1/e section editor for biographies, Karl Ulrich Mayer, consulted handbooks and encyclopedias, asked other section editors to name and rank the most important names in their respective disciplines, ran citation checks, and finally submitted a consolidated list to various experts for review. This resulted in some additions and some deletions and in the end 147 biographies made it into the first edition.

Instructions to authors writing these biographies directed them to include "a brief sketch of the major dates and events in the life of the biographee," to describe "the major contours of the [person's] substantive contribution," and to highlight the "importance and relevance of the biographee's work for the social and behavioral sciences." Largely the same instructions were given to 2/e biographers.

The selection criteria for 2/e were very different. First, on the reasoning that well-done biographies of leading figures is one of the things that draw readers to a work such as this, the total number of biographies was

* My thanks to Andreas Hess for drafting this section of the Introduction.

increased by more than 100 (to just over 250). In a major departure from 1/e, much of the increase is made up of biographies of the living. The 1/e restriction that excluded living persons resulted in anomalies that we thought needed to be addressed. As life expectancy has expanded in the twentieth and twenty-first centuries, so have the life spans of social scientists. Claude Levi-Strauss, to cite one obvious example, lived to the age of 101 years. When 1/e was published, he was in his 90s, was widely cited, was recognized universally as a modern classic, indeed a genius, of social anthropology – and yet, owing simply to his longevity, he was not included in the 1/e biographies. Biographies of, for example, the economist Herbert Simon, the political economist Albert O. Hirschman, the sociologist and philosopher Jürgen Habermas, and the linguist Noam Chomsky were all excluded for the same reason; indeed, Habermas and Chomsky still live as this is being written (Fall, 2014). In all these cases, the major works had been produced decades earlier, their major imprints on the social and behavioral sciences were already evident, their eventual inclusion in works such as this was inevitable. Are readers to be deprived of accounts of their lives, times, and contributions just because they have survived into deep old age?

Once the decision was made to include biographies of the living, however, it was not always easy to decide who among the living merited inclusion. Often the editors had to make decisions on a case-by-case basis, frequently in extended consultation with colleagues or experts in the field. Criteria used to make these decisions were whether the potential biographee had a reputation beyond one culture, nation, or language group; whether there was a record of sustained production over several decades; and whether there was an impact beyond the confines of a single discipline. These criteria reduced our list of suggested biographees considerably. Early lists of candidates numbered as many as 200 new names (in addition to the 147 biographies included in 1/e). In the end, as noted earlier, somewhat more than 100 new names were added.

In a few cases, persons were included that arguably did not satisfy our inclusion criteria. One such is Raul Hilberg, who became famous mainly for having written one very important book, *The Destruction of the European Jews* (1961). Hilberg's book is still regarded as the seminal work in Holocaust studies, a field of research that came to prominence much later; we felt that his signal contributions to Holocaust studies merited a biographical article. Much the same could be said, perhaps, of André Gorz, a well-known social theorist from France, and Albert Memmi, whose work is on colonialism, anticolonialism, and racism. These and other examples could and probably will raise questions about the porousness of our criteria. Whether we were right in giving 'trespassers' their space is open to debate. In any case, whether Hilberg, Gorz, Memmi, and others do or do not merit their biographies is for readers, not us, to decide.

Other important changes in the 2/e biographies section include, first, a conscious effort to attain a better gender balance, including more women both as biographees and as authors; and second, expanding the biographies both geographically and culturally. In the end we made some inroads in both respects but fell short of our internal goals. One of the main reasons for this is that any encyclopedia must, in the first instance, be true to the disciplines it represents. This means being true to the real status of important players, their networks and disciplinary power relations, and their academic capital in their respective fields and beyond. Change in these respects has been noticeable; certainly, the growth of the social sciences in Asia has been remarkable; and in most disciplines, the proportional presence of women has increased. But modest change in the right direction does not a revolution make. Too many nonwhite, non-Western, and nonmale practitioners are still being left out on the disciplinary peripheries and are not fully represented – not in their home disciplines and not in our encyclopedia either.

We state this fact with regret. In many cases, the limiting factor is not excellence but simple visibility. Being part of the trans-Atlantic social science dialogue and engaging this dialog mainly in English translates into more visibility, as discussed earlier. This gets a scholar on the radar screen. Equally excellent work engaging other intellectual dialogues in languages other than English does not. A great many outstanding scholars have thus gone unrecognized in 2/e for precisely this reason. That being said, strenuous efforts were made to make the 2/e biographies more inclusive. Among those included are the Argentinean educationalist Sarmiento, the Brazilian sociologist Cardoso, and the Japanese political scientist Maruyama, to name just a few. This is, however, 'just a few' of the many dozens of prominent social and behavioral scientists 'out there' whom we would have included had we been given more time and more resources (to get non-English works translated, for example). We do feel that we have made some important inroads and beg the reader's understanding of the limits within which we worked. We can only hope that the next edition will strive for more inclusiveness and better representation than we were able to achieve.

Another entirely new feature in the 2/e biographies is the series of 'collective biographies.' In many instances, the intellectual lives of individuals who work in research groups intersect and form a network that makes the

entire group (or paradigm) successful. Examples include articles on the Durkheimians, the Annales School, the Frankfurt School, and several others.

As indicated earlier, instructions to authors were changed little from 1/e. All authors were asked (1) to include the most important biographical facts (country of origin, generation, education, influences, other formative experiences, networks, etc.); (2) to list the most important output (discoveries, publications, etc.) and to briefly discuss the subject matter and content of the most important publications or discoveries; (3) to look at the lasting impact of the person and his or her work (if possible, beyond just one discipline, language group, or culture); and (4) to provide a bibliography of primary and secondary sources related to the biographee and his or her work. Most authors stuck to these guidelines.

Finally, we state for the record that there are a number of important social and behavioral scientists that we would have liked to include but for whom we were unable to find a proper biographer. The search for biographers went on for nearly 4 years. In many cases, our first-choice author agreed at once, but in other cases, we approached four, five, or even more people before an agreeable author was found. And in about two dozen cases, no such author was ever found, with the result that a deserving person's biography had to be dropped. Forgoing a case-by-case discussion of these deletions, let it be simply noted that this is why we have not included biographies of people such as Amartya Sen, the economist and social philosopher; Raymond Boudon, the French sociologist; Ronald Dworkin, the legal theorist; Avishai Margalit, the philosopher; or a collective article for the British Marxist historians (Eric Hobsbawm, E.P. Thompson, Christopher Hill, and Perry Anderson). Perhaps our most regrettable omission is Neil Smelser himself, for whom no biographer could be found despite at least a dozen attempts. (Paul Baltes's biography is included.) Neil is one of the giants on whose shoulders we all stand.

Redundancy

I mentioned earlier that in the beginning of the project, there was no 'map' linking each of the 1/e articles to its respective 1/e Section. Our efforts to create such a map invariably bogged down every time we hit an article that could plausibly belong to any of several sections, which was often. Obviously, it would have been possible to force every article into a 2/e section, but at the time it seemed advisable to involve the editors themselves in these decisions. So on numerous occasions, I distributed lists of 'orphaned' articles (articles for which no home had yet been designated) to the area and section editors, asking for foster parents to volunteer. Eventually, of course, every article found a home, but at a relatively late date in the process, I became aware that a few dozen articles had found more than one home. That is, some of the editors adopted articles without filing the appropriate paperwork with the editor in chief! In such cases, the result was two articles on the 'same' topic. (The 2/e Table of Contents was scoured multiple times looking for precisely these kinds of duplications but in the end, several proved elusive.)

In the large majority of cases, the content of the 'duplicate' articles was sufficiently distinct that both articles were retained, often with some retitling to help readers differentiate between them. For example, readers will find two articles on 'social identity.' One was recruited for the section on social psychology; the other for the sociology section. With rare exception, the two articles discuss different aspects of social identity from entirely different disciplinary angles; indeed, the two articles share only a small handful of common references. So it was an easy decision to retain them both: one appears here as Social Identity in Social Psychology; the other as Social Identity in Sociology. Careful readers will find perhaps a score of similar cases in the encyclopedia.

Alas, in a few regrettable cases, the overlap between duplicate articles was so pronounced that one of the articles had to be deleted. Since by the time these problems surfaced the articles had (often) been written and rewritten several times and had been accepted by both the section and area editors, the last-minute decision to delete a redundant article was painful. My sincere apologies to the authors involved.

Some redundancy in a work of this scope and aspiration is not only inevitable, it is probably also desirable. A reader interested in, say, social inequality will (one hopes) want to know what the sociologists, anthropologists, economists, and geographers all have to say about that topic. And if a few of the many articles on the topic say more or less the same thing, then what harm is done? Editors at all levels were asked to be alert to redundancies across articles and the editorial record shows that many were. That being said, quite a bit of topical redundancy no doubt remains, and that is not (in general) a matter of concern. It is what Smelser and Baltes refer to as the "inevitable residue of overlapping."

The Review Process

Once the table of contents had stabilized, each of the editors was asked to review the 1/e articles assigned to their section or area and come to various decisions. In some cases, 1/e articles could be reprinted more or less 'as is,' with perhaps some updating of references for readers seeking entry points into the current literature. This was especially the case with what were eventually designated as 'legacy articles,' e.g., Gabriel Almond's article on Civic Culture, James Gregor's article on Dictatorship, &c. In other cases, the 1/e articles needed some updating to be suitable for 2/e, either by their original authors when available or by others recruited specifically for the purpose. The general rule was that all 1/e authors would be offered the opportunity to do the updates themselves, and that other authors would be sought in the event of the original authors' declinations. In still other cases, the 1/e article was so obsolete than an entirely new article would be required. And then finally, the review of 1/e articles often revealed areas of current research that were inadequately covered; in these cases too, completely new articles would be sought.

Where appropriate, the relationship between the 1/e and 2/e versions of an article is spelled out in the article tagline. In many cases, there were issues of authorship to resolve — a 1/e article written by *X* and completely revised by *Y* is *Y*'s article although *X* might be identified in the tagline; but what of an article written by *X*, now deceased, and only modestly revised and updated by *Y*? Every decision rule that I formulated to resolve such issues soon confronted an exceptional case where the rule did not seem to apply. So in the end, the only workable rule was that everyone should get proper credit for their intellectual property, whether prime author or reviser, whether living or dead. If exceptions to this rule have crept into the final product (as they almost certainly have), I offer my apologies to those whose work has been slighted or improperly recognized. Deciding on proper authorship for many of our articles was a far more involved and difficult task than expected.

The section editors formed the first line of editorial defense and were the ones to whom article drafts were originally submitted. The section editors read and commented upon the draft articles and returned them to authors for revision. It is a rare article that appears here in its original form: on average, articles were revised two or three times before being sent to the area editors and subsequently on to me; in many cases, it was more like five or six revisions, with subsequent revisions at the request of the area editors and the editor in chief. And in some cases, articles were submitted to outside readers for independent review. As I stated earlier, the reviewing process was not blind, but it was quite thorough, as is reflected in the exceptional quality of the final product.

Internationalism and Inclusivity

A truly *International* Encyclopedia of the Social and Behavioral Sciences would of course require a global network of editors and authors *and* a commitment to internationalize the content of the work wherever possible. With respect to the editorial team for 2/e, this was definitely realized. The section editors and editors in chief who produced the first edition were drawn from only nine countries: The United States, Germany, Italy, Canada, Sweden, Australia, France, Switzerland, and the UK. Two-thirds of those editors were from the USA and only 13% were women. Area and section editors for 2/e represent 21 countries, US scholars make up only 38% of the total, and women comprise 36%. Thus, the proportional representation of women among the editors was nearly tripled in 2/e over 1/e while the predominance of the USA was halved. Nations represented among the editors of 2/e that were not represented in 1/e include Luxembourg, Greece, Finland, Japan, China, Ireland, Singapore, Spain, Russia, Austria, Israel, and the Netherlands. The presence of several Asian nations on this list is particularly gratifying. To be sure, the North Americans and Western Europeans still predominate, but they no longer monopolize the field.

Adding the International Board of Consulting Editors (see the following statement of Acknowledgments) increases the internationalism of the editorial team even further. Of the 45 members of that Board, only 12 (27%) are US social and behavioral scientists. Countries represented on that Board that are not otherwise part of the editorial team include three Central and South American nations (Venezuela, Argentina, Costa Rica), several Asian nations (the Philippines, Malaysia, South Korea, India), a few Western European nations (Norway, Denmark), two representatives from Central and Eastern Europe (Hungary, Croatia), two from the Middle East (Lebanon, Qatar), and one from South Africa. These editions bring to 35 the total number of countries with representation on the IESBS editorial team.

Much the same can be said of the 2/e authors, who are also more international than the 1/e authors had been. Smelser and Baltes noted in their Introduction that while authors from 51 nations were represented in the work,

58% of the authors were from North America and another 35% were from (mainly Western) Europe. "As to gender composition, 21% of the [1/e] authors were women." In 2/e, the proportion of women among the authors has increased slightly to approximately 25% and the percentage of North American authors has declined to fewer than half. Again, North America and Western Europe still contribute the larger share of the 2/e authors but the growing representation of social and behavioral science authors outside the West is both noteworthy and encouraging.

Another important point is that the international scope of the project is much more pronounced than editors' and authors' current affiliations would lead one to conclude. (This would have been equally true in 1/e as well.) Consider, for example, Michiru Nagatsu, author of our article 03053 on the history of behavioral economics. Dr Nagatsu was born in Japan and retains Japanese citizenship but was educated at Exeter and the London School of Economics and Political Science and wrote his IESBS 2/e article while on a postdoctoral fellowship in Finland. Or Guillermina Jasso, area editor for Area Three. Jasso was born in the USA, received her PhD from Johns Hopkins, and is on the faculty at New York University. By any standard, she would count as a US sociologist in our compilations. Yet her parents were Mexican of Basque, Spanish, Portuguese, and Dutch origins; Spanish is her native language; over the course of her distinguished career, she has published with coauthors from Germany, India, Israel, Mexico, the Netherlands, and Sweden; and she holds a permanent research fellowship at IZA, the Institute for the Study of Labor, in Bonn, Germany. Scoring her as 'USA' scarcely does justice to her biography or to the international perspective she brought to the editorial task.

What is true of author Nagatsu and editor Jasso is true in degrees of virtually the entire roster of IESBS participants. If one includes place of birth, citizenship, where educated, sites of research, visiting appointments, sites of fellowships, honorary professorships, invited lectures, and on through the list of highlights in a modern academic career, it is safe to say that every nook and cranny of the globe (save, possibly, Antarctica) is represented somewhere in the biographies of IESBS authors and editors. The true measure of internationalism, in short, is not so much where people come from or where they currently are, but rather the extent to which they bring a sensitivity and commitment to cultural and national differences into their work. Whether this was or was not realized in 2/e is for readers to decide.

Concluding Observations

Smelser and Baltes aspired in the first edition of this work to "assemble the whole range of knowledge – vast and complex as it is – of the social and behavioral sciences in one place." Since what does and does not qualify as 'social and behavioral science' is itself contested, it is not surprising that nearly every one of the market reviews commissioned by Elsevier to guide the development of the current edition noted content omissions in 1/e that reduced its usefulness and that needed to be addressed in a second edition. Critics can and no doubt will say the same about 2/e. I have already mentioned the nearly 800 articles that appeared on the 2/e Table of Contents at one time or another but that will not be found in the final product, either because agreeable authors could not be located or (less often) because agreeable authors never produced an acceptable article. And it is pretty easy to identify whole chunks of present-day social and behavioral science that are given short shrift, or no shrift at all, in this work – not because these chunks were judged to be unimportant but because limits of time, space, resources, and networks meant that some elements of the contemporary social science enterprise never came to our attention.

The social and behavioral sciences begin with what artificial intelligence guru Marvin Minsky once described as a "three pound computer made out of meat" – the human brain, which for all we know is the most complicated three-pound hunk of matter in the universe – and end with a thoroughly globalized culture and social structure that encompasses all aspects of our species' physical, biological, psychological, social, emotional, cultural, and spiritual lives. We are the most successful species ever to have inhabited the planet; we have colonized every inhabitable piece of dirt; our range extends everywhere; all the earth's resources lay at our command. To record all that has come to be known about human origins, evolution, language, culture, and social structure – everything we know about human emotions, identity, love, religion, family, economics, politics, domination, exploitation, and on through a very long list – would require orders of magnitude more volumes than the mere 26 volumes that comprise the present work.

We have come increasingly to the realization that what is social and human about our species is very intimately connected to our biological, genetic, and evolutionary inheritance. At the same time, the brain specifically and biological evolution more generally "cannot be viewed separately from the material, historical and

social circumstances that produce it" (Berntsen and Baker, 2015). As these authors put it, biology and culture are 'mutually constitutive.' Fifteen years ago, the editors of the first edition remarked on "ferment at the boundaries between the biological and the behavioral and social sciences." And while there are still important controversies, much of the ferment has yielded to cooperation as we are just now beginning to explore this endlessly complex web of interconnectedness between our biological and social selves. It is an easy and nearly inevitable conclusion, thus, that the third edition of the IESBS will contain a great deal more of the work being done at this intersection than 2/e was able to accommodate.

To conclude on a purely personal note, the 5 years I have spent working on this project produced many professional and personal gratifications, none quite so delicious as the certain knowledge that I would learn something new every single day. My work on 2/e exposed me to literature I never knew existed and to aspects of the social and behavioral sciences of which I was at best dimly aware. I can only hope that readers who come to these pages will find them as enriching as I have.

Bibliography

Park, Denise C., 2004. Review of the International Encyclopedia of the Social and Behavioral Sciences, first ed. Contemporary Psychology: The APA Review of Books 49 (6).
Berger, P.L., Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Doubleday, New York.
Berntsen, Baker, L.A., 2015. Defiant behavior during adolescence and cultural variations. This volume.
Faris, R.E.L., 1964. Handbook of Modern Sociology. Rand McNally, Chicago.
Gittler, J., 1965. Review of Faris' Handbook of Modern Sociology. American Journal of Sociology 71, 335—336.
Hannerz, U., 2010. Editing anthropology: two experiences in space and time (Chapter Two). In: Hannerz, U. (Ed.), Anthropology's World: Life in a Twenty-First Century Discipline. Pluto Press.
Henrich, J., Heine, S.J., Noranzayan, A., 2010. The Weirdest People in the World? Working Paper Series des Rates für Sozial- und Wirtschaftsdaten, No. 139. Available at: http://hdl.handle.net/10419/43616.
Hilberg, R., 1961. The Destruction of the European Jews. Yale University Press, New Haven, CT.
Merton, R.K., 1965. On the Shoulders of Giants. University of Chicago Press, Chicago.
Mills, C.W., 1959. The Sociological Imagination. Oxford University Press, London.
Sica, A., 2015. Encyclopedias, handbooks and dictionaries. This volume.
Smelser, N.J., Baltes, P., 2001. Introduction. In: Smelser, Baltes (Eds.), The International Encyclopedia of the Social and Behavioral Sciences, first ed. Elsevier, New Tork and Amsterdam, pp. xxxi—xlviii.
Strickhouser, S.M., Wright, J.D., 2015. Applied social research, history of. This volume.

James D. Wright
University of Central Florida, Orlando, FL, USA

ACKNOWLEDGMENTS

A work of this scale is inevitably a communal enterprise involving many hundreds of editors, authors, reviewers, technical support people, colleagues, professional acquaintances, and friends who have provided input, advice, encouragement, and assistance over the past five years. Any listing is bound to be incomplete, so my apologies in advance to those who expected to be mentioned here but are not. Be assured that the slight is not intentional.

My negotiations with Elsevier about this project began early in 2010. By May of that year, I had committed to the project and have been working on it at varying levels of intensity ever since — often on weekends, during holidays, and whenever else I could snatch a useable chunk of time. So my first acknowledgment must be to my wife Chris for suffering through my five year obsession, always with understanding and affection, usually with good cheer. Chris is probably the only person in the world who is happier than I am that the project is now completed.

I must also acknowledge an enormous intellectual debt to the editors in chief of the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, Drs Neil Smelser and Paul Baltes (the latter now deceased). Dr Smelser was gracious enough to serve as Chair Emeritus of the project's International Board of Consulting Editors and to put himself at my disposal as, in his words, "a battle-scarred veteran." The basic intellectual structure of this work is adopted more or less wholesale from the Smelser—Baltes edition, with a few exceptions noted earlier in the *Introduction*. Without the original to serve as the starting point and blueprint, it is hard to imagine even embarking on such a task, much less bringing it to completion. Their edition was hailed as "the largest corpus of knowledge about the social and behavioral sciences in existence," which is to say that it formed an *extremely* solid foundation from which to build the work you now hold in your hands.

The first edition was organized into 39 sections overseen by a team of 53 section editors. The structure of the second edition is slightly different. It is divided up into 10 broad 'Areas,' each with an Area Editor overseeing the work of between four and eight Section Editors. In all, the second edition is thus comprised of the 10 Areas subdivided into 53 Sections; the 53 Sections were developed by a team of 76 Section Editors. So whereas Smelser and Baltes had 53 colleagues to help complete the work, I had 84 (two of the Area Editors also edited their own sections). On pages i-v in the front matter to this work one finds an Editorial Board list showing the Area Editors, Section Editors, Sections, names, and affiliations of the entire editorial team — a team whose dedication and effort in behalf of the project has vastly exceeded the rate of pay. A more helpful, productive, or delightful group of scholarly collaborators is difficult to imagine.

What readers see in the listed Editorial Structure of the work is the *final* editorial team as it existed when the work went into production. As would be expected in a project of this duration, there was considerable turnover within the team as the project developed and as death, competing demands, and general indifference all took their toll. I am grateful, nonetheless, to everyone who agreed, at least initially, to participate in the project, whether or not they were able to stay involved to the very end.

Susan Hanson (Geography) and Ulf Hannerz (Anthropology) are owed a special note of thanks because, of the 53 Section Editors involved in the first edition, these are the only two who agreed to reprise their roles in the second. (Both also wisely requested and received coeditors to carry part of the load.) Persuading people to serve who had no real idea what they were getting into was challenge enough, but these two suffered through 1/e and were willing to do it all over again for 2/e. And for that, I am especially grateful.

Many of the editors benefitted from the assistance of colleagues and graduate students at their various institutions. There was no mechanism built into the editorial process to keep systematic track of this

assistance and so the listing below is certain to be incomplete. But I can at least acknowledge those of whom I am aware who contributed time and intellectual energy to the effort: Liam Morton and Andrea Livingstone (Psychiatry); Jennifer O'Connor and Alana Dorris (Applied, Industrial and Organizational Psychology); Jen Hanis-Martin (Sexuality); David Budde, Annette Stimmer, Nikolas Ott, and Franziska Pfeifer (Institutions and Infrastructure of Social and Behavioral Sciences); Matthew LeDuc and Scott Freeman (Contemporary Cultural Concerns); and Simone Dudziak (Public Policy). To those named, my thanks. To those omitted, my apologies.

A special note of thanks to Dr Donald J. Barrett. Don was originally recruited to develop an entirely new section of the second edition on Gay, Lesbian, Bisexual, and Transsexual Studies, completed the section in short order, then e-mailed me to ask if there was anything further he could do to be helpful. Consequently, he took a major hand in developing the section on Biographies and was quite helpful with many other sections too. He appears in the Editorial Structure both as a Section Editor and as a Senior Editorial Associate to acknowledge his yeoman service to the entire enterprise. Also listed as a Senior Editorial Associate is my colleague Dr Amy M. Donley, who has been with the project since the very beginning and who, more than once, was all that stood between me and madness. I am proud to count them both as colleagues, collaborators, and friends.

I am also pleased to acknowledge Don Barrett, Andreas Hess, Harry Whitaker, John Berry, Barbara Miller, Scott Bentley, and Claire Byrne for their incisive and helpful commentary on previous drafts of the Acknowledgments and Editor's Introduction.

The project's International Board of Consulting Editors, chaired by Smelser and Dr Peter Marsden of Harvard University, was formed late in 2011 as a 'sounding board' to be used by the Area and Section Editors in searching for topics and authors and to help push for greater internationalism in our coverage. As it turned out, several members of this Board also stepped up to write entries in their areas of expertise. The composition of this Board and the affiliations of the consulting editors (as they were when the Board was first formed, not necessarily as they are today) are as follows: Neil J. Smelser, *University of California, USA*; Peter V. Marsden, *Harvard University, USA*; Ronald F. Abler, *International Geographical Union, USA*; John Bacon-Shone, *University of Hong Kong, China*; Roberto Briceño-León, *Universidad Central de Venezuela, Venezuela*; Roel Bosker, *Rijksuniversiteit Groningen, The Netherlands*; Michael Burawoy, *University of California, Berkeley, USA*; Nestor T. Castro, *University of the Philippines, Philippines*; Dalton Conley, *New York University, USA*; Freek Cronjé, *North West University, South Africa*; Scott Davies, *McMaster University, Canada*; Abdul Rahman Haji Embong, *Malaysian Social Science Association, Malaysia*; Joan-Maria Esteban, *Barcelona Graduate School of Economics, Spain*; Robert S. Feldman, *University of Massachusetts, USA*; Sari Hanafi, *American University of Beirut, Lebanon*; Gudmund Hernes, *Norwegian Business School, Norway*; Sheila Jasanoff, *Harvard University, USA*; Tim Jensen, *University of Southern Denmark, Denmark*; Colin Jones, *University of London, UK*; Nadezhda Lebedeva, *Higher School of Economics; and National Research University, Russia*; Miklós Lévay, *Eötvös Lorand University, Hungary*; Dinka Marinovic Jerolimov, *Institute for Social Research, Croatia*; Laurie M. Joyner, *Rollins College, USA*; Ferenc Kiefer, *Hungarian Academy of Sciences, Hungary*; Pan Suk Kim, *Yonsei University, South Korea*; Katsuya Kodama, *International Peace Research Association; and Mie University, Japan*; Allan Lavell, *Facultad Latinoamericana de Ciencias Sociales (FLACSO), Costa Rica*; Jake Lynch, *University of Sydney, Australia*; Cecilia Mabragana, *Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Argentina*; William McBride, *Purdue University, USA*; Ishwar Modi, *University of Rajasthan, India*; Leonardo Morlino, *School of Government, Italy*; Melvin L. Oliver, *University of California, USA*; Angela M. O'Rand, *Duke University, USA*; Kenneth Prewitt, *Columbia University, USA*; Thilo Rehren, *University College London Qatar, Qatar*; Colette Sabatier, *Université Victor Segalen, France*; Philip Schofield, *University College, UK*; Rainer K. Silbereisen, *University of Jena, Germany*; Chamhuri bin Siwar, *National University of Malaysia, Malaysia*; Tom W. Smith, *World Association for Public Opinion Research, USA*; Catalina Smulovitz, *Universidad Torcuato Di Tella Miñones, Argentina*; Philip Spinhoven, *Leiden University, The Netherlands*; Sukhadeo Thorat, *Indian Council of Social Science Research, India*; Nigel Thrift, *University of Warwick, UK*; John R. Townshend, *University of Maryland, USA*; Gisèle Yasmeen, *Social Sciences and Humanities Research Council of Canada, Canada*.

Searching for an editorial team to represent every social science discipline and every region of the globe quickly exposed the deficiencies in my personal network, so I am indebted to a very large number of scholars around the world for calling names of prospects to my attention, allowing me to drop their names in the recruitment effort, or otherwise helping me find people with the expertise and commitment required for the job. In most cases, this is a list of people who were invited onto the editorial team in some capacity,

demurred, but made valuable suggestions about others to contact. In several cases, people are listed because they engaged me in intellectual discussions about the structure, format, and even advisability of this work, or simply replied to a cold-call e-mail in pleasant, kind, helpful, or informative ways. Again ignoring the risk of omission, my sincere personal thanks to those named below for their contributions large and small to the final product:

Doug Anderton, Juri Allik, George Armelagos, Susan Ayers, Amanda Baumle, Janeen Baxter, Jess ben-Habib, Joao Biehl, David Blanchflower, Hans-Peter Blossfeld, Jason Boardman, Dorret Boomsma, Roel Bosker, Henry Brady, Martin Bulmer, Craig Calhoun, Miguel Centeno, William Cockerham, Mark Nathan Cohen, Amelie Constant, Simon Cottle, Connie Citro, Eileen Crimmons, James Curran, Scott Davies, Donatella Della Porta, George Downs, Francis Dodoo, David Dozois, Sonja Drobnick, Jaap Dronkers, John Dryzek, Riley Dunlap, Lauren Edelman, David Elesh, Paula England, Nancy Eisenberg, Geoffrey Evans, Miriam Faust, Robert Feenstra, Marcus Feldman, Stanley Feldman, Ronald Fischer, James Fowler, Jeremy Freese, Angela D. Friederici, Michael Frumpkin, Masahisa Fujita, Nicole M. Gage, Nancy Gallini, Rosemary Gartner, Morton Ann Gernsbacher, Robert Goodin, John Grattan, Lance Gravlee, David Greenberg, Wendy Griswold, Melanie Guldi, Monica Das Gupta, Peter Hatemi, Robert Haveman, Karen Heimer, John Hibbing, Sarah Blaffer Hrdy, Sheila Jasanoff, Kelly Kadera, Cigdem Kagitcibasi, David Kavanaugh, Roger Keil, Byungkook Kim, Rob Kitchin, Jack Knight, Lyle Konigsberg, Candace Kruttschnitt, Helen Ladd, Phillip Lane, Graeme Lang, Janet Lauritsen, Ashley Leeds, Evelyn Lehrer, Richard Lerner, David Lewis, Tim Liao, Tom Linneman, Norman Long, Karen Lyons, Ruth Mace, Gary Marks, Herbert Marsh, Karl-Ulrich Mayer, Lorraine Mazerolle, Rose McDermott, Susan Michie, Harry Minas, Peter Miller, S. Phillip Morgan, Victor Nee, Vin-Kim Nguyen, Paul Nieuwbeerta, Helga Nowotny, Howard Nusbaum, Claus Offe, Thomas Ollendick, Joan Orme, Dennis O'Rourke, Michael Overington, Catherine Panter-Brick, Susan Parnell, Deborah M. Pearsall, Gerard Pfann, Jon Pierre, Mike Ping, Jennifer Platt, Ken Prewitt, Cathy Price, Jill Quadnagno, Charles Ragin, Thilo Rehren, Ben Rosamond, Bo Rothstein, Beth Rubin, Heinrich Sauer, Peer Scheepers, Ralf Schwarzer, Moshe Semyonov, Michael Shanahan, Sara Shostak, Alan Sica, Rainer K. Silbereisen, Beth Simmons, Catalina Smulovitz, Michael Sobel, Diane Stone, Jill Stephenson, Rafe Stolzenberg, Miron Straf, Richard Swedberg, Alan C. Swedlund, Sidney Tarrow, Jacques-Francoise Thisse, Stefan Timmermans, Florencia Torche, Erik Tsang, Judy Treas, Billie Turner, Cees van der Eijk, Paul van Lange, James Vaupel, Karin Wahl-Jorgensen, Mitchell G. Weiss, Neil Whitehead, Pamela Wilcox, Bjorn Wittrock, Cornelia Woll, Gisele Yasmeen, and Barbie Zelizer.

Gratitude is owed as well to the many colleagues at Elsevier who contributed to the production of this volume. My first debt is to Claire Byrne, the project's Content Project Manager and the head of the Major Reference Works (MRW) team that managed the day-to-day *Sturm und Drang* of the project: aggrieved, confused, or delinquent authors, missed deadlines, testy and intransigent editors, and the often-maddening work style of the Editor in Chief. I am sure there were days when Claire wanted to tear her hair out, but to external appearances, she was an island of calm in a sea of turmoil. She was also exceptionally good at her job.

Claire's MRW team comprised Richard Berryman, Will Bowden-Green, Esmond Collins, Mark Harper, Sam Mahfoudh, Mike Nicholls, Gemma Tomalin, Joanne Williams, and Marise Willis. No matter what issue or problem I faced, one of these fine, capable people was always available to step up with a solution, invariably offered in a spirit of cooperation and good cheer. They have been to IESBS 2/e what tech support is to the modern computerized office. Getting this work completed without their help is inconceivable.

Thanks also to the succession of Elsevier's Acquisitions Editors who managed the finances and internal politics of the project: to Scott Bentley, whose gentle suasions pulled me into the project in the beginning; to Karen Maloney and Peter Labella who ran the show in the middle period; and to Lisa Tickner who was in charge at the very end. These are the people on the front line in the endless war between the publisher, who wants things done now and at minimal expense, and the Editor in Chief, who was always asking for more time and more money. The many dozens of amicable compromises necessary to bring the project to completion were all ably negotiated by Scott, Karen, Peter, and Lisa, and for that, I am grateful indeed.

The University of Central Florida and its Department of Sociology have provided a congenial home for the project. My thanks in particular to Dean Michael Johnson; to the current and former chairs of my department, Jana L. Jasinski and Libby Mustaine; to my faculty colleagues Liz Grauerholz, Phillip 'Hutch' Pollock, Bruce Wilson, Melanie Guldi, Stacy Barber, Tosha Dupras, and John Schultz; and to current and former graduate students Thomas Hall, Sara Strickhouser, Betsy Swart, Jordana Navarro, Marie Gualtieri, Shannon Simone, and Shannon Frey.

And most of all, my sincere thanks to the many thousands of authors whose work this is. When comparing the quality of the work contained within these pages to the authors' honoraria, it is obvious that IESBS 2/e is an act of collective charity by our authors to the disciplines they represent. It is my pleasure to thank our authors on behalf of the global social science community.

James D. Wright
Orlando, Florida
January 2015

GUIDE TO USING THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order.

There are four features to help you easily find the topic you are interested in an alphabetical contents list, cross-references to other relevant articles within each article, subject classification index, and a full subject index.

1. Alphabetical Contents List

The alphabetical contents list, which appears at the front of each volume, lists the articles in the order that they appear in that volume of the Encyclopedia. It includes both the volume number and the page number of each article.

2. Cross-references

All of the articles in the Encyclopedia have been cross-referenced. The cross-references, which appear at the end of an article as a See also list, serve three different functions:

a. To draw the reader's attention to related materials in other articles.
b. To indicate material that broadens and extends the scope of the article.
c. To indicate material that covers a topic in more depth.

3. Subject Classification Index

This index appears in Volume 26 and groups articles under subject headings that reflect the broad themes of Social and Behavioral Sciences. This index is useful for making quick connections between articles and locating the relevant article for a topic that is covered in more than one article. Under some section headings, you will find subheadings, if it is appropriate for the articles in a subject area to be grouped into more specific subareas.

4. Contributors

At the start of each volume, there is list of the authors who contributed to that volume.

5. Index

The index appears in Volume 26. The index includes page numbers for quick reference to the information you are looking for. The index articles differentiate between references to whole article, a part of an article, and a table or figure.

This page intentionally left blank

CONTENTS TO VOLUME 25

Editorial Board *v*

Editor in Chief Biography *xi*

Preface to the First Edition *xiii*

Introduction *xvii*

Acknowledgments *xxix*

Guide to Using the Encyclopedia *xxxiii*

Contributors to Volume 25 *xliii*

VOLUME 25

Valency and Argument Structure in Syntax 1
Jae Jung Song

Value: History of the Concept 7
John E King and Michael McLure

Value of Life, Economics of 14
Glenn C Blomquist

Value Pluralism 21
Ruth Chang

Values Across Cultures, Development of 27
Purnima Singh

Values, Anthropology of 33
James L Peacock and Hannah M McFadden

Values, Psychology of 36
Daphna Oyserman

Values, Social Psychology of 41
Jan Cieciuch, Shalom H Schwartz, and Eldad Davidov

Values, Sociology of 47
Helmut Thome

Veblen, Thorstein (1857–1929) 54
Thomas Hall and Andreas Hess

Venture Capital 59
Martin Kenney

Vestibular System 63
Soroush G Sadeghi and Kathleen E Cullen

Vestibulo-Ocular Reflex, Adaptation of the 70
Diana E Mitchell and Kathleen E Cullen

Victimless Crime 75
Steven Block

Video Cultures: Television 80
Denise D Bielby and C Lee Harrington

Vienna Circle: Logical Empiricism 87
Friedrich Stadler

Vietnamese Revolution, The 95
Hy Van Luong

Violence and Media 100
D Charles Whitney and Ellen Wartella

Violence as a Public Health Problem 105
Raymond P Lorion and Michael B Blank

Violence, Criminalization, and the Homeless 111
Amy M Donley, James D Wright, and Jerian Benwell-Lybarger

Violence, History of 116
Heinz-Gerhard Haupt

Violence in Anthropology 122
Kay B Warren

Violence: Public 128
Charles Tilly

Virtual Reality and Spatial Cognition 133
Gregor Hardiess, Hanspeter A Mallot, and Tobias Meilinger

Virtual Reality in Psychotherapy 138
Julia Diemer, Paul Pauli, and Andreas Mühlberger

Virtue Ethics 147
Christine Swanton

Vision, High-Level Theory of 153
Johan Wagemans

Vision, Low-Level Theory of 158
Hugh R Wilson

Visual Imagery 163
Paolo Bartolomeo

Visual Images in the Media 169
John Huxford

Visual Memory, Psychology of 175
Svein Magnussen

Visual Methods in the Social Sciences 181
Douglas Harper

Visual Perception, Neural Basis of 184
Oliver Braddick

Visual Sociology 191
Piotr Sztompka

Visual Space, Geometry of 197
Josef Lukas

Visual Streams: Dorsal and Ventral 201
A David Milner

Visual Turn 207
Francesco Ventrella

Visual Word Recognition, Neurocognitive Psychology of 214
Arthur M Jacobs and Johannes C Ziegler

Visualization in Science and Technology 220
Michael Lynch

Vocabulary Acquisition 226
Fernand Gobet

Vocational Education 232
Alison Fuller

Vocational Interests, Values, and Preferences, Psychology of 239
Barbara Schneider and Walter G Cook

Vocational Training 245
Núria Rodríguez-Planas

Voluntary Associations, Sociology of 252
David Horton Smith

Voluntary Organizations 261
David Horton Smith

Volunteerism, Psychology of 268
Mark Snyder and Alexander Maki

von Hayek, Friedrich A. (1899–1992) 273
Viktor Vanberg

von Humboldt, Alexander (1769–1859) 279
Tobias Kraft

von Humboldt, Wilhelm (1767–1835) 286
Jurgen Trabant

von Ranke, Leopold (1795–1886) 290
Wolfgang Hardtwig

Voting and Direct Democracy 293
Lawrence LeDuc

Voting, Explanations of: Social Class 299
Geoffrey Evans

Voting, Sociology of 302
Nonna Mayer

Voting: Turnout 309
Benjamin Highton

Vygotsky, Lev Semenovic (1896–1934) 312
René van der Veer

Vygotsky's Theory of Human Development and New Approaches to Education 316
Eugene Matusov

Vygotsky's Theory of Cognitive Development 322
Alex Kozulin

Vygotsky's Sociocultural Theory 329
Natalia Gajdamaschko

Wage Differential and Structure 335
Robert Drago

Wallerstein, Immanuel (1930–) 340
Thomas E Reifer

Walzer, Michael (1935–) 346
Gerard Boucher

Wanting vs Needing 351
Mike JF Robinson and Kent C Berridge

War and Democracy 357
Hans Joas

War and Nationalism 363
Daniele Conversi

War and Terrorism, Geography of 371
James A Tyner and Samuel Henkin

War: Anthropological Aspects, Historical Development of 377
Sverker Finnström and Carolyn Nordstrom

War Crimes Tribunals, Crimes against Humanity and Transitional Justice 382
Jamie Rowen

War, Political Violence, and Effects on Children 389
Christine E Merrilees

War Propaganda 395
Stacey Scriver

War, Social Causes and Consequences 401
Edgar Kiser

War, Sociology of 405
Lars B Kaspersen

Warfare in History 414
Michael S Neiberg

Wars among Nation-States: Patterns and Causes 420
Christian Olsson

Water, Health and Social Inequality 428
Linda M Whiteford and Cecilia Vindrola-Padros

Water Resources and Sustainable Water Management 437
Jim L Wescoat, Jr

Water Wars 443
Ashok Swain

Watson, John Broadus (1878–1958) 448
Peter Harzem

Wealth Distribution 450
Edward N Wolff

Weber, Max (1864–1920) 456
Stephen P Turner

Weberian Social Thought, History of 462
Lawrence A Scaff

Weight Regulation 468
Jane Wardle and Fiona Johnson

Welfare 475
Lorraine C Minnite and Joan M Mazelis

Welfare and Education 481
Klaus Hurrelmann, Katharina Rathmann, and Julian Kickbusch

Welfare and Work 491
Robert A Moffitt

Welfare Economics 497
Yew-Kwang Ng

Welfare: Philosophical Aspects 504
Jonathan Riley

Welfare Policy Comparisons 511
Gosta Esping-Andersen

Welfare Reform 515
Torben M Andersen

Welfare Retrenchment 521
Bernhard Ebbinghaus

Welfare State, History of 528
EP Hennock

Wellbeing and Burnout in the Workplace, Psychology of 534
Raymond T Lee

Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences 537
Sabine Sonnentag

Western and Non-Western Views of the Social and Behavioral Sciences:
De-Westernizing the Social and Behavioral Sciences 541
Thomas Brisson

White Collar Crime: Recent Trends and Debates 551
Michael L Benson

White, Gilbert (1911–2006) 558
James L Wescoat, Jr

White-Collar Workers, History of 564
M Prinz

Whiteness and White Supremacy 569
June Ying Yee

Whorf, Benjamin Lee (1897–1941) 575
Joseph L Subbiondo

Williams Syndrome 579
Joanne S Camp and Annette Karmiloff-Smith

Wilson, William Julius (1935–) 584
Peter I Rose

Wisdom, Psychology of 590
Judith Glück

Wisdom: Philosophical Aspects 598
Wolfgang Welsch

Witchcraft 604
Peter Geschiere

Wittgenstein, Ludwig (1889–1951) 611
Eike von Savigny

Women and Criminal Justice 616
Loraine Gelsthorpe

Women and Psychiatry 622
Carol Cooperman Nadelson and Carol Ann Bernstein

Women in Buddhism 628
Lisa J Battaglia

Women's Religiosity 635
Grace Davie, Tony Walter, and Darren E Sherkat

Women's Status and National and International Security 638
Valerie M Hudson

Word Order 644
Jan Rijkhoff

Work and Industry, Sociology of 657
Tony J Watson

Work, Anthropology of 662
Jim Weil

Work, History of 669
J Ehmer

Work, Labor: History of the Concept 675
Bénédicte Zimmermann

Work Motivation 680
James M Diefendorff and Gina A Seaton

Work, Sociology of 687
Frederick C Gamst

Working Classes, History of 694
M Van der Linden

Working Conditions 698
Françoise J Carré

Working Memory, Neural Basis of 703
Michael Petrides

Working Memory, Psychology of 711
Robert H Logie

Workplace Autonomy 718
Linda C Forbes and John M Jermier

Workplace Deviance 722
Rebecca Bennett and Shelly Marasi

Workplace Safety and Health 727
Simo Salminen

Workplace Stress 733
Michael Sliter and Zhenyu Yuan

World Republic 740
Otfried Höffe

World Systems Theory 748
Daniel Chirot

Worldview 753
Rik Pinxten

Wright, Sewall (1889–1988) 758
James F Crow

Writing, Evolution of 761
Denise Schmandt-Besserat

Writing Instruction 767
Steve Graham

Writing Systems 773
Roger D Woodard

Writing Systems, Psychology of 780
David R Olson

Wrongful Convictions 785
Marvin D Free, Jr

Wundt, Wilhelm Maximilian (1832–1920) 792
Lothar Sprung

Xenophobia and Nativism 796
Peter Hervik

Xenophobia: Social Psychological Aspects 802
Margarita Sanchez-Mazas and Laurent Licata

Youth Culture, Anthropology of 808
Vered Amit

Youth Culture, Sociology of 813
Ryan Moore

Youth Gangs 819
Ross Deuchar

Youth, Geography of 824
Tracey Skelton

Youth Joblessness 830
David NF Bell and David G Blanchflower

Youth Movements 837
Ken Roberts

Youth, Music, and Peace Building 842
Mwenda Ntarangwi

Zooarchaeology 849
Jack M Broughton

CONTRIBUTORS TO VOLUME 25

Vered Amit
Concordia University, Montreal, QC, Canada

Torben M. Andersen
Aarhus University, Aarhus V, Denmark

Paolo Bartolomeo
Brain and Spine Institute, Groupe Hospitalier Pitié-Salpêtrière, Paris, France; Fédération de Neurologie, Hôpital Pitié-Salpêtrière, Paris, France; and Catholic University, Milan, Italy

Lisa J. Battaglia
Samford University, Birmingham, AL, USA

David N.F. Bell
Stirling Management School, University of Stirling, IZA and CPC, Stirling, UK

Rebecca Bennett
Louisiana Tech University, Ruston, LA, USA

Michael L. Benson
University of Cincinnati, Cincinnati, OH, USA

Jerian Benwell-Lybarger
University of Central Florida, Orlando, FL, USA

Carol Ann Bernstein
New York University School of Medicine, New York, NY, USA

Kent C. Berridge
University of Michigan, Ann Arbor, MI, USA

Denise D. Bielby
University of California Santa Barbara, Santa Barbara, CA, USA

David G. Blanchflower
Dartmouth College, Hanover, NH, USA; Stirling Management School, University of Stirling, Stirling, UK; and Peterson Institute for International Economics, IZA, CESifo and NBER

Michael B. Blank
Center for Mental Health Policy and Services Research, University of Pennsylvania, Philadelphia, PA, USA

Steven Block
Central Connecticut State University, New Britain, CT, USA

Glenn C. Blomquist
University of Kentucky, Lexington, KY, USA

Gerard Boucher
UCD School of Sociology, University College Dublin, Dublin, Republic of Ireland

Oliver Braddick
University of Oxford, Oxford, UK

Thomas Brisson
University Paris 8, Saint-Denis, France

Jack M. Broughton
University of Utah, Salt Lake City, UT, USA

Joanne S. Camp
Goldsmiths, University of London, London, UK

Françoise J. Carré
Center for Social Policy, J.W. McCormack Graduate School of Policy and Global Studies, University of Massachusetts Boston, Boston, MA, USA

Ruth Chang
Rutgers University, New Brunswick, NJ, USA

Daniel Chirot
University of Washington, Seattle, WA, USA

Jan Cieciuch
University Research Priority Program Social Networks, University of Zurich, Zurich, Switzerland; and Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

Daniele Conversi
IKERBASQUE, Basque Foundation for Science, Bilbao, Spain; and Department of Contemporary History University of the Basque Country UPV/EHU, Leioa, Spain

Walter G. Cook
College of Education, Michigan State University, East Lansing, MI, USA

James F. Crow
University of Wisconsin, Madison, WI, USA

Kathleen E. Cullen
McGill University, Montreal, QC, Canada

A. David Milner
Durham University, Durham, UK

Eldad Davidov
Institute of Sociology, University of Zurich, Zurich, Switzerland

Grace Davie
University of Exeter, Devon, UK

Ross Deuchar
University of the West of Scotland, Hamilton, UK

James M. Diefendorff
The University of Akron, Akron, OH, USA

Julia Diemer
University of Regensburg, Regensburg, Germany

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Robert Drago
Precision Numerics, Springfield, MA, USA

Bernhard Ebbinghaus
School of Social Sciences, University of Mannheim, Mannheim, Germany

J. Ehmer
University of Salzburg, Austria

Gosta Esping-Andersen
Universitat Pompeu Fabra, Barcelona, Spain

Geoffrey Evans
University of Oxford, Oxford, UK

Sverker Finnström
Uppsala University, Uppsala, Sweden

Linda C. Forbes
Ancell School of Business, Western Connecticut State University, Danbury, CT, USA

Marvin D. Free
University of Wisconsin—Whitewater, Whitewater, WI, USA

Alison Fuller
Institute of Education, University of London, London, UK

Natalia Gajdamaschko
Simon Fraser University, Vancouver, BC, Canada

Frederick C. Gamst
University of Massachusetts, Boston, MA, USA

Loraine Gelsthorpe
Institute of Criminology, University of Cambridge, Cambridge, UK

Peter Geschiere
University of Amsterdam, The Netherlands

Judith Glück
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Fernand Gobet
University of Liverpool, Liverpool, UK

Steve Graham
Mary Lou Fulton Teachers College, Arizona State University, Tempe, AZ, USA

Thomas Hall
University of Central Florida, Orlando, FL, USA

Gregor Hardiess
University of Tübingen, Tübingen, Germany

Wolfgang Hardtwig
Humboldt University, Berlin, Germany

Douglas Harper
Duquesne University, Pittsburgh, PA, USA

C. Lee Harrington
Miami University, Oxford, OH, USA

Peter Harzem[†]
Auburn University, Auburn, AL, USA

Heinz-Gerhard Haupt
University of Bielefeld, Bielefeld, Germany

Samuel Henkin
Kent State University, Kent, OH, USA

E.P. Hennock
University of Liverpool, Liverpool, UK

Peter Hervik
Aalborg University, Aalborg, Denmark

Andreas Hess
University College Dublin, Dublin, Ireland

Benjamin Highton
University of California, Davis, CA, USA

Otfried Höffe
Universität Tübingen, Tübingen, Germany

Valerie M. Hudson
Texas A&M University, College Station, TX, USA

[†] Deceased.

Klaus Hurrelmann
Hertie School of Governance, Berlin, Germany

John Huxford
School of Communication, Illinois State University, Normal, IL, USA

Arthur M. Jacobs
Dahlem Institute for Neuroimaging of Emotion (D.I.N.E.), Freie Universität Berlin, Berlin, Germany

John M. Jermier
College of Business & Patel College of Global Sustainability, University of South Florida, Tampa, FL, USA

Hans Joas
Humboldt University, Berlin, Germany; and University of Chicago, Chicago, IL, USA

Fiona Johnson
Cancer Research UK Health Behaviour Research Centre, University College London, London, UK

Annette Karmiloff-Smith
Centre for Brain and Cognitive Development, School of Science, Birkbeck, University of London, London, UK

Lars B. Kaspersen
Copenhagen University, Copenhagen K, Denmark

Martin Kenney
University of California, Davis, CA, USA

Julian Kickbusch
Hertie School of Governance, Berlin, Germany

John E. King
Greensborough, VIC, Australia

Edgar Kiser
University of Washington, Seattle, WA, USA

Alex Kozulin
Feuerstein Institute, Jerusalem, Israel

Tobias Kraft
Institut für Romanistik, Universität Potsdam, Potsdam, Germany

Lawrence LeDuc
University of Toronto, Toronto, ON, Canada

Raymond T. Lee
I.H. Asper School of Business, University of Manitoba, Winnipeg, MB, Canada

Laurent Licata
Université Libre de Bruxelles (ULB), Bruxelles, Belgium

Robert H. Logie
University of Edinburgh, Edinburgh, Scotland, UK

Raymond P. Lorion
College of Education, Towson University, Towson, MD, USA

Josef Lukas
Martin-Luther Universitat, Halle-Wittenberg, Germany

Hy Van Luong
University of Toronto, Toronto, ON, Canada

Michael Lynch
Cornell University, Ithaca, NY, USA

Svein Magnussen
University of Oslo, Oslo, Norway

Alexander Maki
University of Minnesota, Minneapolis, MN, USA

Hanspeter A. Mallot
University of Tübingen, Tübingen, Germany

Shelly Marasi
Eastern New Mexico University, Portales, NM, USA

Eugene Matusov
University of Delaware, Newark, DE, USA

Nonna Mayer
Centre for European Studies, Sciences Po-National Centre for Scientific Research, Paris, France

Joan M. Mazelis
Rutgers, The State University of New Jersey, Camden, NJ, USA

Hannah M. McFadden
University of North Carolina, Chapel Hill, NC, USA

Michael McLure
Business School, University of Western Australia Business School, Crawley, WA, Australia

Tobias Meilinger
Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany

Christine E. Merrilees
State University of New York, Geneseo, NY, USA

Lorraine C. Minnite
Rutgers, The State University of New Jersey, Camden, NJ, USA

Diana E. Mitchell
McGill University, Montreal, QC, Canada

Robert A. Moffitt
Johns Hopkins University, Baltimore, MD, USA

Ryan Moore
CUNY-Queensborough Community College, New York, NY, USA

Andreas Mühlberger
University of Regensburg, Regensburg, Germany

Carol Cooperman Nadelson
Harvard Medical School, Boston, MA, USA

Michael S. Neiberg
United States Army War College, Carlisle, PA, USA

Yew-Kwang Ng
Nanyang Technological University, Singapore

Carolyn Nordstrom
University of Notre Dame, Notre Dame, IN, USA

Mwenda Ntarangwi
Calvin College, Grand Rapids, MI, USA

David R. Olson
Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada

Christian Olsson
Université Libre de Bruxelles (ULB), Bruxelles, Belgium

Daphna Oyserman
University of Michigan, Ann Arbor, MI, USA

Paul Pauli
University of Würzburg, Würzburg, Germany

James L. Peacock
University of North Carolina, Chapel Hill, NC, USA

Michael Petrides
Montreal Neurological Institute, McGill University, Montreal, QC, Canada

Rik Pinxten
Ghent University, Ghent, Belgium

M. Prinz
Emsdetten, Germany

Katharina Rathmann
Hertie School of Governance, Berlin, Germany

Thomas E. Reifer
University of San Diego, San Diego, CA, USA

Jan Rijkhoff
Aarhus University, Aarhus C, Denmark

Jonathan Riley
Tulane University, New Orleans, LA, USA

Ken Roberts
University of Liverpool, Liverpool, England, UK

Mike J.F. Robinson
Wesleyan University, Middletown, CT, USA

Núria Rodríguez-Planas
IZA – Institute for the Study of Labor, Bonn, Germany

Peter I. Rose
Kahn Liberal Arts Institute, Smith College, Northampton, MA, USA

Jamie Rowen
University of Toronto, Toronto, ON, Canada

Soroush G. Sadeghi
State University of New York at Buffalo, Buffalo, NY, USA

Simo Salminen
Finnish Institute of Occupational Health, Promotion of Occupational Safety, Helsinki, Finland

Margarita Sanchez-Mazas
Université de Genève, Genève, Switzerland

Lawrence A. Scaff
Wayne State University, Detroit, MI, USA

Denise Schmandt-Besserat
University of Texas, Austin, TX, USA

Barbara Schneider
College of Education, Michigan State University, East Lansing, MI, USA

Shalom H. Schwartz
The Hebrew University of Jerusalem, Jerusalem, Israel; and Laboratory of Socio-Cultural Research, National Research University – Higher School of Economics, Moscow, Russia

Stacey Scriver
School of Political Science and Sociology, NUI Galway, Galway, Ireland

Gina A. Seaton
The University of Akron, Akron, OH, USA

Darren E. Sherkat
Southern Illinois University, Carbondale, IL, USA

Purnima Singh
Indian Institute of Technology Delhi, New Delhi, India

Tracey Skelton
National University of Singapore, Singapore

Michael Sliter
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

David Horton Smith
Boston College, Retired in Bradenton, FL, USA

Mark Snyder
University of Minnesota, Minneapolis, MN, USA

Jae Jung Song
University of Otago, Dunedin, New Zealand

Sabine Sonnentag
University of Mannheim, Mannheim, Germany

Lothar Sprung
Humboldt University, Berlin, Germany

Friedrich Stadler
Institute Vienna Circle (Institut Wiener Kreis), University of Vienna, Wien (Vienna), Austria

Joseph L. Subbiondo
California Institute of Integral Studies, San Francisco, CA, USA

Ashok Swain
Uppsala University, Uppsala, Sweden

Christine Swanton
University of Auckland, Auckland, New Zealand

Piotr Sztompka
Jagiellonian University, Krakow, Poland

Helmut Thome
Martin-Luther-University Halle-Wittenberg, Halle-Wittenberg, Germany

Charles Tilly
Columbia University, New York, NY, USA

Jurgen Trabant
Free University, Berlin, Germany

Stephen P. Turner
University of South Florida, Tampa, FL, USA

James A. Tyner
Kent State University, Kent, OH, USA

Viktor Vanberg
Walter Eucken Institute, Freiburg, Germany

M. Van der Linden
International Institute of Social History, Amsterdam, The Netherlands

René van der Veer
Center for Child and Family Studies, Leiden University, Leiden, The Netherlands

Francesco Ventrella
University of Sussex, Brighton, UK

Cecilia Vindrola-Padros
London South Bank University, London, UK

Eike von Savigny
Universität Bielefeld, Bielefeld, Germany

Johan Wagemans
University of Leuven (KU Leuven), Leuven, Belgium

Tony Walter
University of Bath, Bath, UK

Jane Wardle
Cancer Research UK Health Behaviour Research Centre, University College London, London, UK

Kay B. Warren
Anthropology Department, Brown University, Providence, RI, USA

Ellen Wartella
University of Texas at Austin, TX, USA

Tony J. Watson
University of Nottingham, Nottingham, UK

Jim Weil
Science Museum of Minnesota, St Paul, MN, USA

Wolfgang Welsch
Friedrich Schiller University, Jena, Germany

James L. Wescoat
Massachusetts Institute of Technology, Cambridge, MA, USA

Jim L. Wescoat
Aga Khan Program for Islamic Architecture, Massachusetts Institute of Technology, Cambridge, MA, USA

Linda M. Whiteford
University of South Florida, Tampa, FL, USA

D. Charles Whitney
University of Texas at Austin, TX, USA

Hugh R. Wilson
York University, Toronto, ON, Canada

Edward N. Wolff
New York University, New York, NY, USA

Roger D. Woodard
The University of Buffalo, Buffalo, NY, USA

James D. Wright
University of Central Florida, Orlando, FL, USA

June Ying Yee
School of Social Work, Ryerson University, Toronto, ON, Canada

Zhenyu Yuan
Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

Johannes C. Ziegler
Brain and Language Research Institute, Aix-Marseille Université and Centre National de la Recherche Scientifique, Marseille, France

Bénédicte Zimmermann
Ecole des Hautes Etudes en Sciences Sociales, Paris, France

V

Valency and Argument Structure in Syntax

Jae Jung Song, University of Otago, Dunedin, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The purpose of this article is to explain the valency and argument structure of the predicate. The article begins with definitions of valency and argument structure. This then leads to illustration of a number of ways in which arguments are removed from, or added to, the argument structure, whereby changes in the valency of the basic predicate may be brought about. The article demonstrates that the argument structure can also be altered or rearranged without involving a change of valency. Also illustrated is how languages may keep the valency of the derived predicate from exceeding that of the basic predicate. The article addresses the question of whether or not identical coding of nominals – especially in the context of valency increasing – entails doubling of grammatical relations. Finally, a brief discussion of how valency-changing operations are handled or represented in current grammatical theories is provided.

The purpose of this article is to describe the valency and argument structure of the verb or predicate. The notion of argument is known to have been adapted from mathematics. In mathematics, a function may take one or more arguments, which will be used to determine the value of that function. In linguistics, argument is used as a technical term for syntactic elements that combine obligatorily with the verb (or the predicate as the verb is usually referred to in relation to arguments). For instance, the English verb *die* must take at least one syntactic element, e.g., *the mayor*, to form sentences, e.g., *The mayor died*. This syntactic element is taken to be the (sole) argument of the predicate *die*. The verb *kill*, on the other hand, requires two arguments to form sentences, one representing the killer (*the man*) and the other the person killed (*the mayor*), as in *The man killed the mayor*. The number of arguments that are required by a predicate is then specified in the argument structure of that predicate. Additional syntactic elements, e.g., *in the office* or *in 2011*, can also potentially combine with the predicate *die*, e.g., *The mayor died in the office*, *The mayor died in 2011*, or *The mayor died in the office in 2011*. But these are regarded only as optional, because their absence has no bearing upon the grammaticality of the 'minimal' sentence, *The mayor died*. Such optional syntactic elements are referred to as adjuncts. The distinction between arguments and adjuncts, albeit not always easy to make, is a useful one (see Comrie, 1993 for a discussion on how arguments and adjuncts should be distinguished).

The argument structure also contains information concerning the relation between the predicate and its argument(s). For instance, in the sentence *The man killed the mayor*, the predicate *kill* and its arguments *the man* and *the mayor* have logical (or grammatical) relationships between them with the effect that *the man* is the actor (or the subject) that carried out the killing, and *the mayor* is the undergoer (or the direct object) that was affected by the actor's killing (see Grammatical Relations; also Bickel, 2011 for an overview of grammatical relations). Adjuncts, e.g., *in the office* in *The mayor died in the office*, also bear logical (or grammatical) relationships, e.g., location (or oblique), with the predicate. Information on adjuncts, however, is not included in the argument structure of the predicate itself.

The term valency (or valence in the USA) is said to have been borrowed from chemistry into syntax. In chemistry, the valency of a chemical element is defined as the capacity of that element for combining with a fixed number of atoms of another element, for instance, two hydrogen atoms combining with an atom of oxygen in the case of water (i.e., H_2O). The notion of valency, when applied to verbs, refers to the number of arguments that are required for a given verb to form sentences. For example, the verb *die* is regarded as having a valency structure of a single argument. Thus, verbs such as *die* are characterized as one-place (or monovalent) verbs. The verb *kill* is a two-place (or divalent) verb, because it requires two arguments. The verb *give* is a three-place (or trivalent) verb, because it requires three arguments.

The rest of the article is concerned principally with illustrating how argument structure can be altered or manipulated with or without a concomitant change of valency. How languages may keep the valency of the derived verb from exceeding that of the basic verb by manipulating argument structure will also be described (see Morphology in Linguistics for derivation; also Comrie, 1985). Finally, how

valency-changing operations are handled in current grammatical theories will be discussed briefly.

Change of Argument Structure with Change of Valency

There are two main mechanisms through which the argument structure of the predicate can be altered, together with a change of valency: argument reduction and argument addition. In argument reduction, one of the arguments of the basic predicate may be demoted to adjunct status. In argument addition, on the other hand, adjuncts may be promoted to argument status, or new arguments may be introduced into the argument structure. There are five important grammatical devices utilized in argument reduction or argument addition: passivization, antipassivization, noun incorporation, advancement, and causativization. The first three belong to argument reduction, whereas the last two fall under argument addition (but cf the section Change of Argument Structure without Change of Valency for examples of noun incorporation and causativization consisting of both argument reduction and addition).

Passivization

In the passive clause, one of the arguments of the basic two-place predicate (i.e., the actor) is demoted to an adjunct with the other argument (i.e., the undergoer) retaining its argument status (but being promoted to the subject of the passive clause). This operation is referred to as passivization (e.g., Abraham and Leisiö, 2006; Siewierska, 1984, 2005). The number of arguments of the basic predicate is thus reduced from two (i.e., the actor and undergoer) to one (i.e., the undergoer) in the corresponding passive clause. This is substantiated by the fact that it is possible to omit the newly created adjunct. This change of valency (or valency decreasing) is signaled on the verb either by a verbal affix or by a periphrastic element in the verb phrase. Korean provides an example of the passive:

1. Korean

cuin-i	totuk-lɨl	cap-əss-ta
owner-NOM	thief-ACC	catch-PST-IND

'The owner caught the thief.'

2. Korean

totuk-i	(cuin-eke)	cap-hi-əss-ta
thief-NOM	(owner-by)	catch-PASS-PST-IND

'The thief was caught (by the owner).'

(Abbreviations used to gloss the data: ABS = Absolutive; ACC = Accusative; APASS = Antipassive; APPL = Applicative; ASP = Aspect; AUX = Auxiliary; CS = Causative Suffix; DAT = Dative; DEM = Demonstrative; DET = Determiner; DO = Direct Object; ERG = Ergative; F = Feminine; FUT = Future; INCMP = Incompletive; IND = Indicative; INTR = Intransitive; IO = Indirect Object; IPFV = Imperfective; LOC = Locative; M = Masculine; NFUT = Non-Future; NOM = Nominative; O = Object; OBL = Oblique; PASS = Passive; PERF = Perfective; PL = Plural; PRS = Present; PST = Past; S = Subject; SG = Singular; TR = Transitive; Q = Question.)

In some languages (e.g., Finnish, Nanai, Ulcha, and Ute), passivization involves only demotion of the actor of the basic predicate, with the undergoer remaining unaltered in its grammatical status (i.e., direct object). This is illustrated in

3. Ulcha

ti	dūse-we	hōn-da	ta-wuri
DEM	tiger-ACC	how-Q	do-PASS

'What's to be done about that tiger?'

Antipassivization

Antipassivization is often regarded as the counterpart of passivization in languages with the ergative–absolutive case system (see Polinsky, 2005a for a cross-linguistic survey of antipassive constructions). But it should not be inferred from this that the passive is never found in languages with the ergative–absolutive case system; both the passive and the antipassive may occur in the same language (e.g., Basque, Eskimo, and Mam). In the antipassive, the actor argument of the basic two-place predicate remains the (sole) argument of the corresponding antipassive (albeit marked by the absolutive case), with the undergoer argument demoted to adjunct status or, optionally, omitted. If and when expressed, however, the undergoer is marked accordingly by a dative or oblique case. This suggests that the antipassive is as much intransitive or detransitivized as is the passive: a decrease in valency from two (i.e., the actor and undergoer) to one (i.e., the actor) argument. The verb in the antipassive also bears an affix as a marker of valency decreasing. Dyirbal, an Australian Aboriginal language, provides an example of the antipassive (i.e., (5), as opposed to the basic sentence in (4)):

4. Dyirbal

balan	dʸugumbil	baŋgul	yaɽa-ŋgu	balga-n
F-ABS	woman-ABS	M-ERG	man-ERG	hit-NFUT

'The man is hitting the woman.'

5. Dyirbal

bayi	yaɽa	bagun dʸugumbil-gu	balgal-ŋa-nʸu
M-ABS	man-ABS	F-DAT woman-DAT	hit-APASS-NFUT

'The man is hitting the woman.'

In some languages, the demoted original undergoer must be completely eliminated from the antipassive clause. Bandjalang, another Australian Aboriginal language, is one such language, as illustrated in (6) and (7).

6. Bandjalang

mali-yu	ɖa:ɖam-bu	mala-Ø	bulan-Ø	ɖa-ila
that-ERG	child-ERG	that-ABS	meat-ABS	eat-PRS

"The child is eating the meat."

7. Bandjalang

mala-Ø	ɖa:ɖam-Ø	ɖa-le-ila
that-ABS	child-ABS	eat-APASS-PRS

'The child is eating.'

Noun Incorporation

Noun incorporation is another type of valency-changing operation found in some languages of the world, whereby one of the arguments of the predicate becomes part of the

predicate, thereby losing its argument status (Mithun, 1984). The resulting predicate may function as a single word. In Chukchi (a Chukotko-Kamchatkan language), for instance, the vowels of incorporated nouns must participate in word-internal vowel harmony in conjunction with host verbs (i.e., *kupre* in (8) vs *kopra* in (9)):

8. Chukchi

tumg-e	na-ntəwat-ən	kupre-n
friends-ERG	set-TR	net-ABS

'The friends set the net.'

9. Chukchi

tumg-ət	kopra-ntəwat-gʔat
friend-NOM	net-set-INTR

'The friends set nets.'

Note that, with the undergoer incorporated into the predicate, the actor argument of the basic two-place predicate in (8) loses its ergative marking and is instead marked by the nominative case in (9). In Chukchi, noun incorporation also motivates the verb to shed its transitive marking (i.e., *-ən*) in favor of the intransitive marking (i.e., *-gʔat*). These changes indicate that noun incorporation involves reduction of two arguments of the basic predicate to one (i.e., valency decreasing).

Advancement of Adjuncts to Arguments

So far, argument reduction has been illustrated. In many languages, it is also possible to do the opposite, that is, to increase the valency of the basic verb by promoting an adjunct (e.g., beneficiary/recipient, comitative, instrumental, locative, etc.) to argument status. In order to signal such advancement or promotion, the verb is typically marked by a so-called applicative affix; the construction as a whole is commonly referred to as the applicative construction (see Peterson, 2007 for a cross-linguistic survey; also Polinsky, 2005b). In Kinyarwanda (a Niger–Congo language), for instance, instrumental adjuncts can be promoted optionally to argument or direct object status. Compare the following two sentences:

10. Kinyarwanda

úmwáalímu	a-ra-andik-a	íbárúwa	n'í-íkárámu
teacher	he-PRS-write-ASP	letter	with-pen

'The teacher is writing a letter with the pen.'

11. Kinyarwanda

úmwáalímu	a-ra-andik-iish-a	íbárúwa	íkárámu
teacher	he-PRS-write-APPL-ASP	letter	pen

'The teacher is writing a letter with the pen.'

The instrumental adjunct in (10) is promoted to direct object in (11), evidence for which can be adduced from the fact that the instrumental nominal, which is marked by the instrumental prefix in (10), appears as a bare nominal (i.e., no case marking) in (11). This advancement must also be registered on the verb by the applicative suffix *-iish*, as in (11). Both the old direct object (a.k.a. the basic object) and the newly created direct object (a.k.a. the applied object) in (11) exhibit grammatical properties of direct objecthood (but see the section 'Doubling of Grammatical Relations'). This suggests that the valency of the derived verb in (11) is greater by one argument than that of the basic verb in (10). In (10), there are two arguments (i.e., *teacher* and *letter*) and one adjunct (i.e., *pen*). In (11), on the other hand, there are three arguments (i.e., *teacher*, *letter*, and *pen*). Moreover, the argument structure of the predicate in (10) specifies the relation between the actor and the undergoer, and the argument structure of the derived predicate in (11) specifies the relation between the actor, the undergoer, and the instrument. To put it differently, (10), without the instrumental phrase *n'í-íkárámu*, will remain grammatical, whereas (11), with *íkárámu* deleted, will be rendered ungrammatical. Incidentally, the most common semantic role that is encoded by the applied object is that of beneficiary/maleficiary and/or recipient (Peterson, 2007, p. 40).

Causativization

In common with advancement of adjuncts, causativization adds an argument to the argument structure of the basic predicate. However, it does so not by converting an existing adjunct into an argument, but by introducing a completely new argument into the argument structure (see Song, 1996, 2005 for a cross-linguistic survey of causative constructions; but cf. Kittilä (2009) for the non-valency-changing function of causative morphology). This newly introduced argument is the causer argument. Thus, if the basic verb is a one-place verb, the corresponding causative verb will be a two-place verb (the newly added causer argument and the sole argument of the original basic verb). If the basic verb is a two-place verb, the corresponding causative verb will be a three-place verb (three arguments including the causer argument). This is illustrated by the following pair of Turkish sentences. As opposed to the basic verb in (12), the causative verb in (13), with the causative suffix *-t*, has the causer argument, *dişçi*.

12. Turkish

müdür	mektub-u	imzala-dɪ
director	letter-DO	sign-PST

'The director signed the letter.'

13. Turkish

dişçi	mektub-u	müdür-e	imzala-t-tɪ
dentist	letter-DO	director-IO	sign–CS–PST

'The dentist made the director sign the letter.'

The valency of the causative verb thus exceeds that of the basic verb by one argument, as can be neatly expressed in

14. NONCAUSATIVE: n vs CAUSATIVE: $n + 1$
[n: number of arguments]

The schema in (14) indicates that causativization is a case of argument addition (i.e., valency increasing), with the status of the argument(s) of the basic predicate remaining intact. This may indeed be true of one-place verbs; causativization only increases the valency of the basic one-place verb from one to two arguments. Causativization of two- or three-place (or trivalent) verbs, in contrast, would give rise to three or four arguments, respectively, more than many languages may be able to cope with – regardless of whether verbs are basic or derived. Indeed, there is much cross-linguistic variation in causativization of two- or three-place verbs, as opposed to one-place verbs. For instance, there are languages that derive causative verbs only from basic one-place verbs (e.g., Comrie,

1989; Dixon, 2000; Dixon and Aikhenvald, 1997; Song, 2001). Moreover, in some languages argument structure is manipulated in such a way that the valency of the derived verb is prevented from exceeding that of the basic verb (Dixon, 2000; Song, 1996), as will be discussed in the section 'Change of Argument Structure without Change of Valency.'

Change of Argument Structure without Change of Valency

There are languages that introduce a new argument into the argument structure of the causative verb, only if one of the arguments of the basic two- or three-place verb is demoted to an adjunct or omitted completely. In other words, causativization of the basic two- or three-place verb in such languages results in no increase in valency, contrary to the schema in (14), but rather the valency of the derived causative verb is equal to that of the basic verb. In Babungo (a Niger–Congo language), for instance, when the causative suffix *-s* is added to two-place verbs, the direct object of the basic verb must be omitted or expressed as an adjunct. Thus, the original direct object in the noncausative sentence in (15), *zɔ̃*, can either appear as a prepositional adjunct, *nə zɔ̃*, or be omitted in the corresponding causative sentence in (16).

15. Babungo

ŋwə́	fɛ̀e	zɔ̃
he	fear-PERF	snake

'He was afraid of a snake.'

16. Babungo

mə̀	fɛ̀sə̀	ŋwə́	(nə̀ zɔ̃)
I	fear–CS–PERF	him	(with snake)

'I frightened him with a snake.'

There also exist some languages that maintain the canonical number of arguments with respect to verbs or predicates by not encoding either the causee or some argument other than the causer when these verbs are morphologically causativized. In Afar, an Eastern Cushitic language, there are two causative suffixes, one exclusively for intransitive verbs and the other for transitive verbs. Many causativized transitive verbs, unlike causativized intransitive ones, do not have the original subject of transitive verbs (i.e., the causee NP) specified, as in

17. Afar

ʼoson	ʼgarca	gey-siis-ee-ʼni
they	thief	find–CS–they/PERF-PL

'They caused the thief to be found/they caused someone to find the thief.'

So normally, in Afar, the number of arguments per causative verb, regardless of the transitivity of the original verb, is no more than two.

Songhai (or Sonrai), a Nilo-Saharan language, has a causative suffix *-ndi*, which can be added to intransitive and transitive verbs. When it is suffixed to ditransitive verbs, either the causee NP or the indirect object NP of the original verb has to be left out. This explains why the following sentence involving a three-place predicate *neere* 'to sell' is ambiguous in two ways.

18. Songhai

garba	neere-ndi	bari	di	musa	se
Garba	sell-CS	horse	the	Musa	IO

'Garba had Musa sell the horse' *or*
'Garba had the horse sold to Musa.'

In other words, there is a strict requirement in this Nilo-Saharan language that there should be no more than three arguments per causative verb.

Moreover, there are languages that avoid exceeding the maximum number of arguments by drawing on detransitivizing strategies already available in the grammar. In Halkomelem, a Salish language, morphological causativization is possible only if the noncausative verb is first detransitivized via the antipassive suffix *-əm*. Thus, (19) is ungrammatical, since the causative suffix *-st* is added to the transitive verb stem, *q'wəl-ət-*. In contrast, (20) is grammatical, since the transitive verb is antipassivized (i.e., detransitivized) before morphological causativization takes place.

19. Halkomelem

*ni	cən	q'ʷəl-ət-stəxʷ	kʷθə	səplíl	ʔə	łə	sɬéniʔ	
AUX	1SG	bake-TR-CS	DET	bread	OBL	DET	woman	

"I had the woman bake the bread."

20. Halkomelem

ni	cən	q'ʷə́l-əm-stəxʷ	θə	sɬéniʔ	ʔə	kʷθə	səplíl
AUX	1SG	bake-APASS-CS	DET	woman	OBL	DET	bread

'I had the woman bake the bread.'

In Blackfoot, an Algonquian language, there are two causative suffixes, *-ippi* and *-atti*, which apply to both intransitive and transitive verbs. However, when these suffixes are added to them, transitive verbs must first be detransitivized via the suffix *-a:ki*. The net effect of this detransitivization is that the direct object of the original verb is changed to adjunct status and thus left out of the argument system of the verb concerned. Note that in this language, the verb internally records arguments. Conversely, nonrecorded elements are by definition adjuncts or optional elements. After detransitivization of the two-place verb, only the subject is registered in the verb, not the direct object. Only after that is the causative suffix added to the now detransitivized verb, whereby the causer NP can also be registered in the verb as an argument. Thus, only the causer and causee NPs are recorded in the verb. In Blackfoot, therefore, the number of arguments per causative verb never exceeds two owing to the detransitivization. In effect, the number of arguments in causatives is exactly the same as in noncausative (transitive) sentences. Similarly, in Bandjalang, an Australian Aboriginal language, the causative suffix *-ma* can be added to intransitive verbs. However, when they undergo the morphological causativization, transitive verbs must first be detransitivized via the antipassive suffix *-li*.

In Southern Tiwa (a Kiowa-Tanoan language), the causative verb registers only two arguments, that is the causer and the causee NPs. When the causative suffix *-ʼam* is added to transitive verbs, the direct object of the original verb is obligatorily incorporated into the verb

complex (see the section 'Noun Incorporation' for noun incorporation).

21. Southern Tiwa

I-'u'u-kur-'am-ban
1SG.S:2SG.O-baby-hold–CS–PST

22. Southern Tiwa

*'U'ude i-kur-'am-ban
baby 1SG.S:2SG.O-hold–CS–PST
'I made you hold the baby.'

The sentence without incorporation of the direct object of the original verb as in (22) is ungrammatical. In Southern Tiwa, arguments should be registered in the verb complex, while optionally they can appear in full forms as well. Thus, *'u'ude* 'baby' in (22) is intended to be an argument. But as the causative verb has already reached its full capacity of registering the arguments (i.e., the causer and causee NPs), thereby being unable to record the full NP *'u'ude*, the sentence is ruled out as ungrammatical. On the other hand, in (21) the same NP is incorporated into the verb complex, thus effectively making the original verb intransitive. Now, the subject NP of the detransitivized verb and the causer NP associated with the causative suffix *-'am* can be registered in the verb without any difficulty.

What the foregoing data suggest is that, regardless of whether the predicate is basic or derived, languages allow only a certain number of arguments per predicate (i.e., two or three) (for a detailed discussion, see Song, 1996, Chapter 6; also Kemmer and Verhagen, 1994 from a cognitive–linguistic perspective). The net effect of the various grammatical strategies pressed into service to maintain the canonical number of arguments is no change of valency in spite of the introduction of the causer argument, although there may be a change of argument structure between the basic predicate and the derived predicate: addition of the causer argument offset by demotion to adjunct status (or omission) of one of the arguments of the basic predicate.

Another instance of a change of argument structure with no change of valency comes from noun incorporation (cf the section 'Noun Incorporation'). When one of the arguments is incorporated into the verb, the position vacated by that argument may then be usurped by an adjunct, e.g., instrument, location, or even possessor (Mithun, 1984, p. 856–859). In Tupinambá (a Tupi-Guaraní language), the possessor can be promoted to the direct object position when the possessed noun is incorporated into the verb. Compare (23) and (24).

23. Tupinambá

s-oβá a-yos-éy
his-face I-it-wash
'I washed his face.'

24. Tupinambá

a-s-oβá-éy
I-him-face-wash
'I face-washed him.'

In (24), the possessor is represented on the verb by the prefix *s-*, whereas in (23), without noun incorporation, the direct object is *s-oβá* 'his face,' and it is thus registered on the verb by the prefix *yos-*. In Yucatec Mayan (a Mayan language), the argument position vacated by the incorporated direct object (i.e., *če'* 'tree') is taken up by the locative phrase (i.e., *in-kool* 'my cornfield'), the argument status of which is indicated by its loss of the locative preposition (i.e., *ičil*).

25. Yucatec Mayan

k-in-čak-Ø-k če' ičil in-kool
INCMP-I-chop-it-IPFV tree in my-cornfield
'I chop the tree in my cornfield.'

26. Yucatec Mayan

k-in-č'ak-cče'-t-ik in-kool
INCMP-I-chop-tree-TR-IPFV my-cornfield
'I clear my cornfield.'

Doubling of Grammatical Relations

One theoretical issue in argument addition concerns the question of whether or not identical coding of nominals entails doubling of grammatical relations. In some languages, two nominals in the causative construction may be coded identically, for instance. In Kinyarwanda causatives, both the actor argument (or subject) and the undergoer argument (or direct object) of the basic predicate are coded identically as direct objects. Comrie (1989) interprets such identical coding to be indicative of doubling of the direct object. This position, however, has been challenged by Polinsky (1994), who demonstrates that the two identically coded nominals in Kinyarwanda causatives behave differently in terms of syntactic properties. In fact, only one of the identically coded 'direct objects' is syntactically a direct object, the other being an indirect object, identical coding notwithstanding. Kozinsky and Polinsky (1993) describe more or less the same situation in Korean and Dutch causatives. Similarly, the basic object and the applied object in the Kinyarwanda applicative sentence (11) are bare nominals (that is, coded identically) (see the section 'Advancement of Adjuncts to Arguments' for applicative constructions). This also led to a controversy as to whether or not (11), as opposed to (10), has two direct objects, i.e., doubling of the direct object. Gary and Keenan (1977) take the view that (11) has two direct objects. Perlmutter and Postal (1983), on the other hand, argue that (11) does not have two direct objects, but rather one direct object and one indirect object. Further evidence against doubling of the direct object in Kinyarwanda applicative constructions is provided by Dryer (1983) and Polinsky (1994). Peterson (2007) provides a cross-linguistic survey of the distribution of direct object properties in applicative constructions. His preliminary generalization is that when the newly created or applied object in applicative constructions encodes semantic roles other than beneficiary/recipient (i.e., instrumental, locative, comitative, etc.), it tends not to display all direct object properties with the old or basic object claiming some direct object properties. If the encoded semantic role is beneficiary/recipient, however, the tendency is that the applied object possesses direct object properties, with the basic object displaying direct object properties only in a few attested instances.

Theoretical Approaches to Valency-Changing Operations

Current grammatical theories have proposed different approaches to representation of valency-changing operations; more than one approach may also be possible within one and the same grammatical theory (Brown and Miller, 1996). These approaches can be characterized broadly as lexical or syntactic. Lexical Functional Grammar regards valency-changing operations as productive lexical processes giving rise to alternative assignments of grammatical functions (or grammatical relations) to arguments. Relational Grammar invokes revaluation rules to handle valency-changing operations. These revaluation rules are regarded as syntactic, because they operate on grammatical relations between initial and noninitial strata (or different syntactic levels) of the relational structure. In Principles and Parameters Theory, valency-changing operations are treated as syntactic. In morphological causativization, for instance, the basic verb undergoes movement to combine with the causative affix at an abstract syntactic level. This movement is then claimed to comply with the same principles that are obeyed by other movements in syntax. Research interest in valency-changing operations, however, has all but dissipated in Minimalism, the latest generative perspective on linguistic theory. This may not come as a surprise because the primary objective of Minimalism is to evaluate theories such as Principles and Parameters Theory in terms of metatheoretical desiderata (e.g., virtual conceptual necessity, economy, symmetry) rather than develop any particular theory.

See also: Grammatical Relations; Linguistics: Incorporation; Syntax–Semantics Interface.

Bibliography

Abraham, W., Leisiö, L.. (Eds.), 2006. Passivization and Typology. John Benjamins, Amsterdam.

Bickel, B., 2011. Grammatical relations typology. In: Song, J.J.. (Ed.), Oxford handbook of Linguistic Typology. Oxford University Press, Oxford, pp. 399–444.

Brown, K., Miller, J. (Eds.), 1996. Concise Encyclopedia of Syntactic Theories. Pergamon, Oxford.

Comrie, B.., 1985. Causative verb formation and other verb-deriving morphology. In: Shopen, T.. (Ed.), Language Typology and Syntactic Description, Grammatical Categories and the Lexicon, vol. 3. Cambridge University Press, Cambridge, pp. 309–348.

Comrie, B.., 1989. Language Universals and Linguistic Typology. Blackwell, Oxford.

Comrie, B.., 1993. Argument structure. In: Jacobs, J., von Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax: An International Handbook of Contemporary Research. Walter de Gruyter, Berlin, pp. 905–914.

Dixon, R.M.W.., 2000. A typology of causatives: form, syntax and meaning. In: Dixon, R.M.W., Aikhenvald, A.Y.. (Eds.), Valency Changing: Case Studies in Transitivity. Cambridge University Press, Cambridge, pp. 35–83.

Dixon, R.M.W., Aikhenvald, A.Y.., 1997. A typology of argument-determined constructions. In: Bybee, J., Haiman, J., Thompson, S.A.. (Eds.), Essays on Language Function and Language Type. John Benjamins, Amsterdam, pp. 71–113.

Dryer, M.S.., 1983. Indirect objects in Kinyarwanda revisited. In: Perlmutter, D.M.. (Ed.), Studies in Relational Grammar 1. University of Chicago Press, Chicago, pp. 129–140.

Gary, J.O., Keenan, E.L.., 1977. On collapsing grammatical relations in universal grammar. In: Cole, P., Sadock, J. (Eds.), Syntax and Semantics, Grammatical Relations, vol. 8. Academic Press, New York, pp. 83–120.

Kemmer, S., Verhagen, A.., 1994. The grammar of causatives and the conceptual structure of events. Cognitive Linguistics 5, 115–156.

Kittilä, S., 2009. Causative morphemes as non-valency increasing devices. Folia Linguistica 43, 67–94.

Kozinsky, I., Polinsky, M.., 1993. Cause and patient in the causative of transitive: coding conflict or doubling of grammatical relations? In: Comrie, B., Polinsky, M. (Eds.), Causatives and Transitivity. John Benjamins, Amsterdam, pp. 177–240.

Mithun, M.., 1984. The evolution of noun incorporation. Language 64, 847–894.

Perlmutter, D.M., Postal, P.M.., 1983. Some proposed laws of basic clause structure. In: Perlmutter, D.M.. (Ed.), Studies in Relational Grammar 1. University of Chicago Press, Chicago, pp. 81–128.

Peterson, D.A.., 2007. Applicative Constructions. Oxford University Press, Oxford.

Polinsky, M.., 1994. Double objects in causatives: towards a study of coding conflict. Studies in Language 19, 129–221.

Polinsky, M.., 2005a. Antipassive constructions. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures. Oxford University Press, Oxford, pp. 438–441. Also available at: http://wals.info/chapter/108.

Polinsky, M.., 2005b. Applicative constructions. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures. Oxford University Press, Oxford, pp. 442–445. Also available at: http://wals.info/chapter/109.

Siewierska, A.., 1984. Passive: A Comparative Linguistic Analysis. Croom Helm, London.

Siewierska, A.., 2005. Passive constructions. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures. Oxford University Press, Oxford, pp. 434–437. Also available at: http://wals.info/chapter/107.

Song, J.J.., 1996. Causatives and Causation: A Universal-Typological Perspective. Addison Wesley Longman, London.

Song, J.J.., 2001. Linguistic Typology: Morphology and Syntax. Pearson (Longman), Harlow.

Song, J.J.., 2005. Nonperiphrastic causative constructions. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures. Oxford University Press, Oxford, pp. 450–453. Also available at: http://wals.info/chapter/111.

Value: History of the Concept

John E King, Greensborough, VIC, Australia
Michael McLure, Business School, University of Western Australia Business School, Crawley, WA, Australia

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We distinguish two broad categories of value theories, objective and subjective, which focus, respectively, on the conditions of production and on the preferences of consumers. We discuss the contributions to the labor theory of value made by David Ricardo and Karl Marx, outline some criticisms of Marxian value theory, and summarize Piero Sraffa's attempt to rehabilitate the classical, objective theory of value. We then outline the development of the marginal utility of value and discuss its relationship to market equilibrium, before assessing the marginal productivity theory of distribution, some extensions to utility analysis, and the connection between utility and welfare.

Introduction

We distinguish two broad categories of value theories, objective and subjective, which focus, respectively, on the conditions of production and on the preferences of consumers. While thinking on the issue can be traced back at least as far as Aristotle (Lowry, 1987), modern work on the objective value theory was most influential in the era of classical political economy, especially in the work of David Ricardo (1772–1823) and Karl Marx (1818–83), and was revived after 1960 by the followers of Piero Sraffa (1898–1983). Subjective value theory became prominent after 1870 under the influence of William Stanley Jevons (1835–82), Léon Walras (1834–1910), and Carl Menger (1840–1921), and remains the approach to the theory of value that is taken by mainstream economists today.

The Objective Classical Theory of Value

The objective classical theory of value dates from the mid- to late seventeenth century, reflecting increased interest in the production of commodities and reduced emphasis on the circumstances of their exchange. It was the cornerstone of classical political economy (Dobb, 1973; O'Brien, 2004). In the two decades after 1670, Sir William Petty (1623–87) "came remarkably close to the idea that the exchange value of a commodity is determined by the quantity of labour needed to produce it" (Meek, 1973, p. 35). Half a century later, Richard Cantillon (1680–1734) distinguished the market price of a commodity from its 'intrinsic value,' the latter being proportional to the land and labor needed for its production.

In *The Wealth of Nations* (1776), Adam Smith (1723–90) took three distinct approaches to the problem of value. The first, following Petty, was a simple 'labor embodied' theory, which Smith believed to apply without qualification to the 'early and rude state of society.' The second, which he thought more appropriate to contemporary capitalism, was an 'adding-up theory' that explained value as the sum of the costs of production, including land and capital in addition to labor. His third theory, in some ways which anticipated the later subjective value theory (but was not fully developed by Smith), saw value as being determined by the 'toil and trouble' experienced by the producers.

David Ricardo and the Labor Theory of Value

A much more systematic treatment of value was provided by David Ricardo. Ricardo's primary interest was in the distribution of the social product, and in particular the determinants of the rate of profit, since he regarded this as by far the most important influence on the growth prospects of the British economy. He needed a theory of value less for its own sake than to provide a reliable way of measuring these distributional variables. Ricardo seems originally to have hoped that he could evade the problem by formulating a 'corn model' of agricultural production in which both inputs and outputs consisted entirely of physical quantities of grain. Soon, however, he replaced corn with labor "as the quantity in terms of which product, wages, and surplus were alike expressed" (Dobb, 1973, p. 74).

Ricardo was critical both of Adam Smith's 'adding-up theory' of value and of Jean-Baptiste Say's subjective approach to the problem. He began his own *Principles of Political Economy and Taxation* with an extended discussion of the labor theory of value, noting that air and water were much more useful than gold, but also much less valuable. "Utility then is not the measure of exchangeable value," he concluded as a general principle, "although it is absolutely essential to it" (Ricardo, 1821, p. 11). For the small minority of useful objects that could not be freely reproduced, such as rare statues and pictures, scarce books and coins, and vintage wines, value is indeed determined by their scarcity relative to the demand for them. But the great majority of commodities can be freely reproduced 'almost without any assignable limit' by the employment of human labor, and the labor theory of value does apply to all such commodities, "on the production of which competition operates without constraint" (Ricardo, 1821, p. 12). Fluctuations in supply and demand cause only "accidental and temporary deviations of the actual or market price of commodities" from their 'primary and natural price,' which is given by "the comparative quantity of labor which is necessary to their production" (Ricardo, 1821, p. 88).

Integrated with his labor theory of value was Ricardo's surplus theory of income distribution. Wages were determined

by the consumption requirements of the laborers, and were the first charge on the net product (the difference between total output and the means of production – seed corn, for example – that were needed to produce it). The residual was the 'surplus produce,' which Ricardo believed to be divided between capitalists and landlords in accordance with his theory of differential rent. His treatment of rent allowed Ricardo (unlike Cantillon) to eliminate land from the analysis of value, since the value of corn was determined by the quantity of labor needed to produce it on the least fertile land, where no rent was paid.

Capital, however, was altogether more troublesome. It posed four difficult problems, all related to the question of time (O'Brien, 2004, p. 100). First, the ratio of capital to labor differed from industry to industry. Second, the durability of fixed capital – which Ricardo distinguished from circulating capital – was unequal between industries. Third and fourth, the period of production and the turnover rate of capital were also different in different industries. These four factors would cause prices to diverge from labor values, even in the long period when rates of profit tended toward equality, so that an increase in wages would lead to changes in natural prices even if the quantities of labor embodied in the various commodities was unaltered. But Ricardo convinced himself that these complications were relatively minor. "The greatest effects which could be produced on the relative prices of those goods from a rise in wages," he claimed, "could not exceed 6 or 7 per cent" (Ricardo, 1821, p. 36), leaving him with what the neoclassical economist George Stigler (1958) famously described as a 93% labor theory of value.

Ricardo continued to wrestle with these problems until the very end of his life, acknowledging just before he died that the labor theory of value was only an approximation to the truth (King, 2013, pp. 79–80). But that did not lead him to embrace a subjective theory of value, or even to withdraw his objections to Adam Smith's adding-up theory. A change in wages would not lead to a change in the prices of all commodities in the same direction, as Smith's principle entailed. 'On the contrary,' Ricardo noted, "if wages fell, those commodities only would fall in value, which had a less proportion of fixed capital employed upon them, than the medium in which price was estimated [that is, gold]; all those which had more, would positively rise in price" as a consequence of the increased rate of profits (Ricardo, 1821, p. 46).

Karl Marx and the Labor Theory of Value

Marx was a strong admirer of Ricardo, but also a severe critic. His most fundamental objection was that Ricardo had failed to distinguish different modes of production, instead regarding capitalism as timeless and eternal. Where the theory of value was concerned, Marx believed, the crucial distinction was between capitalism and the earlier, classless 'simple' or 'petty commodity production.' In simple commodity production, commodities normally sold at prices equal to their labor values, but in capitalism matters were considerably more complicated.

He also criticized Ricardo for failing to make a second vital distinction, between constant and variable capital: the former was the value of the 'dead' or 'stored-up' labor in raw materials and machines, while the latter, the 'living labor,' represented the value of the workers' labor power, and was the sole source of surplus value, and hence of profit. This affected his analysis of the rate of profit, which for Marx (no less than for Ricardo) was the single most important economic variable. In Marx's system, the rate of profit was positively related to the rate of exploitation (the ratio of surplus value to variable capital) and negatively related to the organic composition of capital (the ratio of constant to variable capital) (Marx, 1867: chs. VIII–IX; Marx, 1895: chs. III–IV; Howard and King, 1985: ch. 6).

In a competitive capitalist economy, Marx argued – again like Ricardo – there would be strong forces tending to equalize both the rate of profit and the rate of exploitation in all industries. But there were inescapable technical conditions that gave rise to large differences in the organic composition of capital in different sectors. This meant that commodities would not, in general, sell at prices equal to their labor values. They would instead sell at what Marx termed 'prices of production,' which would exceed labor values in industries with an above-average organic composition and fall below values if the organic composition was below the social average. In aggregate, however, the sum of prices would equal the sum of values; the sum of profits would be equal to the sum of surplus value; and the rate of profit would be the same, whether it was calculated in terms of values or prices of production.

In the numerical example in volume III of *Capital*, published in 1895, 12 years after Marx's death, there are five industries, with an organic composition of capital ranging from 1.5 in industry III to 19.0 in industry V, and also significant differences between industries in the durability of constant capital. After transformation, the ratio of price of production to labor value ranges from 1.85 in industry III to 0.86 in industry V (Marx, 1895: ch. IX). This analysis, to repeat, did not appear in Marx's lifetime, and before its publication Friedrich Engels (1820–95) amused himself by announcing a 'prize essay competition' in which he invited the readers of volume II of *Capital* to supply the solution to the so-called 'transformation problem.' No one succeeded, though several entrants came very close (Howard and King, 1985, 1989: part III, ch. 2).

Criticism of Marx's procedure for transforming labor values into prices of production began almost immediately after it was revealed in 1895. The Austrian economist Eugen von Böhm-Bawerk (1851–1914) rejected the entire theoretical apparatus, including the distinction between constant and variable capital and the theory of exploitation in addition to the labor theory of value itself, while his former student Rudolf Hilferding (1877–1941) came to Marx's defense. Ladislaus von Bortkiewicz (1868–1931), himself sympathetic to the classical tradition in political economy, took issue with the details of Marx's analysis, criticizing him for failing to transform input values and pointing out that the aggregate equalities on which Marx had insisted would hold only under special circumstances (all three essays are reprinted in Sweezy, 1975). Other significant contributors to the pre-1914 debate included Wolfgang Mühlpfort (1872–1928) and V.K. Dmitriev (1868–1913) (Howard and King, 1989: ch. 3).

In the 1920s some Marxists began to ask whether these quantitative aspects of the labor theory of value had not been given too much attention, at the expense of the deeper and more interesting qualitative questions. The best known of these Hegelian Marxists was the Russian Menshevik, Isaak Illich

Rubin (1886–1937), whose thinking has stimulated a large and continuing literature on the 'form of value' in Marx (Rubin, 1973). Some Western Marxists began to ponder the implications of the decline in competition associated with the rise of cartels, trusts, and giant corporations at the end of the nineteenth century. If competition was no longer strong enough to equalize the rate of profit across industries, what relevance did the (quantitative) labor theory of value have in the new stage of 'monopoly' or 'state monopoly capitalism'? Paul Sweezy (1910–2004) borrowed extensively from neoclassical analyses of oligopoly pricing, while continuing to endorse the qualitative labor theory of value (Sweezy, 1942). Discussion of the (quantitative) transformation problem also continued after 1945, with a range of alternative solutions being proposed (Howard and King, 1992, part IV).

Piero Sraffa and the Revival of the Classical Theory of Value

Meanwhile, a young Italian scholar was working in Cambridge on the definitive edition of Ricardo's collected works, and also attempting to reinterpret the entire classical tradition in value theory. The eventual outcome was a slim volume entitled *Production of Commodities by Means of Commodities*, with the modest subtitle 'Prelude to a critique of economic theory.' Piero Sraffa (1898–1983) took as given the outputs of the various commodities, the inputs of those commodities and of labor that were needed to produce them, and one distributive variable. This could be either the real wage (as in Ricardo), or the wage share in net output, or – Sraffa's own preference – the rate of profit, itself possibly determined by the rate of interest. He demonstrated that, on the assumption of an equal rate of profit in all industries, these parameters were sufficient to determine all prices and the other distributive variables, without any reference to demand conditions or to marginal utilities (Sraffa, 1960).

In fact Sraffa shared the classical emphasis on 'physical real cost' (Kurz, 2003, p. 169), and had no time whatever for subjective value theory. His analysis pointed to fundamental weaknesses in the neoclassical theory of capital, since the capital stock could not be defined independently of the rate of profit, still less inserted into an aggregate production function that could then be differentiated with respect to 'capital' in order to determine the rate of profit. Sraffa also exposed a 'monotonic fallacy' in neoclassical theory: a rise in the profit rate, he demonstrated, might reduce the value of capital (the phenomenon of 'capital reversal') and might also induce capitalists to adopt productive techniques with a higher capital–labor ratio ('reswitching'). Sraffa's work encouraged the search for alternatives to the marginal productivity theory of distribution and gave rise to the celebrated 'Cambridge controversies' in the theory of capital (Harcourt, 1973).

There were also important implications for the Marxian labor theory of value. Sraffa's intention seems to have been to rehabilitate Ricardo's value theory ('seems,' because he published almost nothing after 1960 and never responded to the substantial critical literature that his book soon generated). But Meek (1973, pp. xxviii–xliv) soon recognized that Sraffa's analysis could also be used to solve the transformation problem. Sraffa had devised a 'Standard Commodity' in which the ratio of outputs to inputs was identical for all commodities. This Standard Commodity, Meek suggested, could be used as the *numéraire* in transforming values into prices of production, overcoming the difficulties that Bortkiewicz and others had identified with Marx's solution, so that (as in the one-commodity Ricardian corn model) the maximum rate of profit depended solely on the technical conditions of production and did not vary with prices or with the actual rate of profit.

This procedure might be sufficient to rescue the labor theory of value. Whether it was also necessary was another matter. Neoclassical critics like Paul Samuelson (1915–2009) maintained that the entire process of transformation was an 'unnecessary detour,' since prices of production could be derived directly from knowledge of conditions of production and income distribution, without any need to calculate labor values. Samuelson was supported by the former Marxist Ian Steedman (b. 1941), who also drew on Sraffa to identify serious additional problems with the labor theory. The existence of joint production and alternative techniques of production could produce paradoxical results, including negative labor values and negative surplus value, even while prices and the rate of profit remained positive (Steedman, 1977). It was soon recognized that these perverse outcomes posed a significant challenge, not just to Marxian value theory but also to any Sraffian rehabilitation of classical theory (Howard, 1987).

As already noted, Sraffa's own intentions were never made clear. The fundamental question that he left unanswered was whether *Production of Commodities* should be regarded as a brick or a broom, that is to say, as the basis for a modernized version of classical value theory or only as a critique – certainly, a profound and damaging critique – of the subjective, neoclassical theory of value and distribution. Thus his work remains open to a variety of different, if not entirely inconsistent, interpretations (Roncaglia, 2009).

One final question concerns the ethical dimension of objective value theory, and in particular the use of labor values as instruments for the purposes of socialist planning. It is unclear what Marx's own position was on this issue (compare Meek, 1973, pp. 256–263 and Robinson, 1942, pp. 23–24). Several serious problems have been identified, including the environmental damage caused by neglecting the costs of land, water, and other natural resources; the excessive capital-intensity induced by ignoring differences in the durability of capital goods; and the failure to relate prices to use values when there is no genuine market for consumer goods (Nove, 1983). Arguably, efficient planning requires a synthesis of objective and subjective value theory, as proposed (unsuccessfully) in the former Communist countries by would-be reformers like Viktor Novozhilov (1892–1970) and Ota Šik (1919–2004).

The Subjective Neoclassical Theory of Value

In the years following the publication of William Stanley Jevons's *Theory of Political Economy* (1871), Carl Menger's *Grundsätze der Volkswirtschaftslehre* (1871), and Léon Walras's *Élements d'Économie Politique Pure* (1874), the concept of 'marginal utility' gravitated to the very center of economic

theory. While there were many important classical precursors to significant elements of 'neoclassical' theory (Ekelund and Hébert, 2002), the 1870s marked a turning point because the entire body of theory started to be systematically reformulated around the proposition that the value of economic goods is fundamentally a consequence of subjective assessment by individuals.

Marginal Utility and Equilibrium

The utility that an individual derives from the consumption of a quantity of a particular good is determined by his or her subjective assessment of the pleasure, or satisfaction, derived from consumption. 'Marginal utility' refers to the satisfaction that an individual gains from the consumption of an additional unit of a particular good in a given context. For example, the context for the marginal utility derived from the consumption of the 10th unit of a good is given by the individual having already consumed 9 units of that good, and so on. While H.H. Gossen (1810–58) and Jules Dupuit (1804–66) had already posited that marginal utility diminishes with additional consumption, the early neoclassical economists reflected on the implications of that proposition for the relationships between demand and price, and supply and price.

Of the three founders of the marginal revolution, Menger's contention that "value does not exist outside the consciousness of men" (Menger, 1871, p. 121) was the strongest endorsement of a strictly subjective character of value theory. His consequential treatment focused on causal sequences and inspired the development of an 'Austrian school' that was distinct from the mainstream equilibrium analysis associated with Jevons and Walras, where, to varying degrees, objectively determined supply considerations equilibrate with the subjectively determined demand within a competitive environment on the assumption that individuals choose to maximize their utility. The central pillar of their theory is the proposition that a maximizing individual will attain equilibrium through exchange when the relative prices for goods possessed are equal to his or her subjective assessment of their relative marginal utilities. If the relative price of a good is less than an individual's assessment of its marginal utility relative to other goods, a maximizing individual will wish to acquire more of that good in exchange for other goods he or she possesses with relative prices that are higher than his or her assessment of their relative marginal utility.

From this, three key principles emerged: quantity demanded for a particular good decreases with respect to the price of that good; quantity supplied for a particular good increases with respect to the price of that good; and, when one equilibrium price prevails for each particular good (the 'law of one price') in a competitive environment, quantities of goods are exchanged up to the point where each individual's assessment of relative marginal utilities reflects relative equilibrium prices, at which point the value of all goods supplied is equal to the value of all goods demanded for both individuals and markets. Some early neoclassical economists used these relationships to reinterpret Adam Smith's discussion of the distinction between use value and exchange value to suggest that water has a low exchange value because its relative abundance suggests low marginal utility, whereas diamonds have a high exchange value because their relative scarcity suggests high marginal utility. Conversely, the high total utility of water is equated with high use value, whereas the low total utility of diamonds is equated with low use value.

The general conception of equilibrium was formulated by Léon Walras by treating utility for each individual as 'separable,' so that the utility from consuming a good depends on the quantity of that good alone, and 'additive' in the sense that total utility from consuming a number of goods is derived by summing of utility for each good. If a, b, … , m indicate quantities of economic goods A, B, …, M held by individual i following voluntary exchange, individual i's total utility U is

$$U_i = f(a_i) + g(b_i) + \ldots + r(m_i) \qquad [1]$$

The marginal utilities of each good are given by the first derivative of the respective separable elements of relation [1], so that relative prices under competitive equilibrium reflect the various ratios of these first derivatives. However, in general equilibrium the quantity demanded of any one good is not just related to the price of that good; rather, it is related to the equilibrium price for each and every consumer good, with each equilibrium price expressed as the rate at which it exchanges for a common *numéraire* (good A). But the relationship between the price of each particular good and the quantity demanded of each good remains negative; it is just that the demand relationship is conditioned by the relative prices of other consumer goods.

Walrasian equilibration extends beyond exchange to encompass production and capital formulation. Walras also argued that a 'groping' process of price adjustments, which he called *tâtonnement*, would lead to a system of equilibrium prices, to which quantities would adjust until markets fully clear. From the perspective of value theory, the key point for Walras was that marginal utility, or *rareté* as he called it, was the cause of value due to the general interdependence deriving from the utility-maximizing behavior of each individual entrepreneur, capitalist, landlord, and worker. An entrepreneur making decisions about the formation of a particular capital good has regard to the demand for the productive service provided by that good, as well as the costs of accessing capitalists' savings, just as an entrepreneur purchasing productive services has regard to the demand for consumer goods derived from the theory of exchange, as well as the price of productive services. The direction of causation for all values originates from the relative prices of consumer goods, which reflect relative marginal utilities.

In 1889 Francis Ysidro Edgeworth (1845–1926) was the first to point out a consistency problem with Walras's equilibrating notion of *tâtonnement*, as a static theory does not demonstrate a path from disequilibrium to equilibrium, which is significant when dynamic adjustments to disequilibrium prices take time to complete. For example, disequilibrium decisions concerning the formation of heterogeneous capital goods have flow-on effects for the supply of productive services and the availability of consumer goods. Late in life, Walras removed disequilibrium trading from his theory, with an auctioneer crying out prices but transactions themselves not settled until equilibrium prices have been attained. In

contrast, Vilfredo Pareto (1848–1923), Walras's successor in Lausanne, distinguished between complete and incomplete equilibrium, and provided for the owners of particular capital goods, among others, to acquire positive or negative rents under incomplete equilibrium.

Alfred Marshall (1842–1924), a second-generation neoclassical economist, created an alternative approach to Walras's that has become known as partial equilibrium. By invoking an 'other things being equal' assumption, Marshall developed a theory of the relationship between price, quantity demanded, and quantity supplied, for a particular good in isolation from other goods. For the individual, and the market for a particular good in aggregate, a rise in price leads to a decrease in the quantity demanded; and equilibrium within a market is given at the point where the demand and supply intersect.

In Marshall's analysis, the 'marginal demand price' is a point on the demand curve and represents the amount of money that an individual forgoes to acquire one additional unit of the good. Consequently, the marginal demand price is a monetary representation of marginal utility, provided that the marginal utility of money is constant. Pareto objected that this is an unnecessary assumption, and derived the equation for a negatively sloped demand curve when permitting the marginal utility of the money (Walras's *numéraire*) to vary (Pareto 1893–94). In 1915, Eugen Slutsky (1880–1948) then used Pareto's demand equation to isolate the income effect of a price change (the change in the purchasing power of a consumer's budget resulting solely from a change in the price of a good) from the substitution effect (the substitution of one good for another attributable solely to the change in the relative price of the good). In 1934 J.R. Hicks (1904–89) and R.G.D. Allen (1906–83) independently developed an alternative formulation of the income and substitution effects, with Hicks, in *Value and Capital* (1939), subsequently pointing out that the 'Marshallian' demand curve shows the combined influence of the income and substitution effects, whereas a study of the pure effect of changes in relative prices would derive the demand curve based only on the substitution effect by compensating the consumer for any income effect associated with a change in the price of a good. In doing so, Hicks provided a succinct mechanism to reconcile the fundamental difference in the demand theories of Walras–Pareto and Marshall.

Marginal Productivity and the Distribution of the Product

The implication of the theories of production and capital formation for the distribution of the product were not fully articulated in the 1870s. For example, Walras's theory of production was initially predicated on 'fixed' coefficients of production. It was not until the technique of production was subject to change that questions related to the substitution of different services of production, and to the relationship between such substitution and variations in wage rates and rentals on capital and land, that analysis of the distribution of product, in the form of income, took center stage in the 1890s. Major studies in that regard include the article by J.B. Clark (1847–1938) on 'Distribution as determined by a law of rent' (1891); the influential book by Phillip Wicksteed (1844–1927) *An Essay on the Co-ordination of the Laws of Distribution* (1894); and the 1894 review of that book by A.W. Flux (1867–1942), which simplified and crystallized the meaning of Wicksteed's work by famously invoking Euler's theorem. J.B. Clark's (1899) book, the *Distribution of Wealth*, is also important for bringing the complete story together.

To maximize profit, entrepreneurs reflect on the productivity of each productive service when offering a price for that service, just as the owners of each productive service have regard to the best price they can obtain for providing that service. In the same way that consumers maximize utility by focusing on comparative marginal utilities under competitive conditions, so too would entrepreneurs maximize profits, and the owners of productive services maximize incomes, when the price of productive services reflects their comparative 'marginal product' (the increase in output from adding one final unit of that particular productive service while holding the quantity of all other services unchanged), provided of course that those services are used to produce goods that are the most highly valued by consumers. In such circumstances, Euler's theorem confirms that total product is completely allocated to the owners of productive services in direct proportion to the product of their productive service's marginal product and quantity.

But the limits within which the 'marginal productivity theory of distribution' is valid have been the subject of much discussion, even among early neoclassical economics (Pullen, 2010). Pareto used first-order homogeneous production functions in 1894 to consider the substitution of productive services by changing the coefficients of production, but when he subsequently treated the issue of marginal productivity at length he raised concerns about the limits of substitutability (as some factors are necessarily fixed) and noted that productive services are not independent of output (when there are economies or diseconomies of scale). He also convinced Barone and Wicksteed of the importance of these limitations, with Wicksteed subsequently abandoning much of the marginal theory of distribution. Walras, in contrast, quickly adopted the marginal productivity theory of distribution.

Extended Dimensions to Utility Analysis

The utility relation [1] referred to earlier is only additive because it is separable and expressed as a cardinal quantity. Second-generation neoclassical theorists, however, soon questioned both these requirements, with separability in particular seen as an unnecessary restriction that constrained the scope of utility analysis. In 1889, Rudolf Auspitz (1837–1906) and Richard Lieben (1842–1919) presented the utility of one commodity as depending on the quantities of all commodities, so that the separable and additive utility for individual i given by relation [1] is effectively replaced by a general nonseparable and nonadditive relation, such as relation [2] below:

$$U_i = f(a_i, bi, \ldots m_i) \qquad [2]$$

This allowed Auspitz and Lieben to undertake pioneering theoretical work on the complementary or substitutable relationship between different goods, which was later adopted and extended by Edgeworth, Pareto, and others.

But the extension of the analysis to encompass demand derived from multi-good utility functions also posed serious questions regarding the stability of the utility relation. When utility at a particular point in the utility system alters, depending on the path taken to attain that point, demand relationships will fluctuate. Utility is no longer completely exogenous – some elements of an individual's subjective valuation of utility, and the associated demand relationships, are influenced by the economic process itself. Stable demand equations depend on each individual's assessment of utility being 'path independent,' so that utility at a particular point is the same irrespective of the path to that point. In mathematical terms, path independence is contingent upon establishing the 'integrability' of the utility function, sometimes referred to as the 'integrability problem.' The formal conditions that establish integrability for economic analysis were first devised in Italian by G.B. Antonelli (1858–1944) in 1886, but their importance was not appreciated at the time. Nicolas Georgescu-Roegen (1906–94), Herman Wold (1908–92), and Paul Samuelson subsequently provided formally correct solutions.

The second significant change to utility theory was the move away from specifying utility as a quantity (cardinal utility) toward preferences over particular combinations of goods (ordinal utility). After Irving Fisher (1867–1947) conjectured, in 1892, that the measurement of utility may not be strictly required for equilibrium theory, Pareto (1900, 2006–09) demonstrated analytically that the shape of Edgeworth's indifference curves is unchanged for an individual irrespective of whether utility is considered cardinally or ordinally, and that the point of economic equilibrium is identical in both cases. He then suggested that experimental data on choice between pairs of goods could be collected in two ways: by direct observation of actual market decisions to derive what Hicks and Allen (1934) would call the marginal rates of substitution (which play the same role as relative marginal utilities in early neoclassical theory), or by conducting a series of binary choice experiments to notionally derive indifference curves after establishing whether respondents reject or accept proposed transactions. The subsequent and definitive specification of choice theory replaced the indifference curve framework with axiomatic formalism when Paul Samuelson (1938) and Hendrik Houthakker (1950) respectively introduced the 'weak' and 'strong' axioms of 'revealed preference.' One feature of the 'strong axiom' is that it establishes conditions for preference revelation that are consistent with integrability having been established for demand relationships and utility.

Utility and Welfare

While Marshall and Edgeworth took important tentative steps in the direction of analyzing the relationship between individual utility and collective welfare, it was Pareto who argued that any suggestion that individual maximization necessarily implies collective economic maximization is a proposition that needs to be demonstrated analytically. In demonstrating that relationship with the aid of his famous welfare criterion (that an unequivocal gain in collective economic welfare requires at least one person to gain and no one to be made worse off), Pareto (1906–09) also underlined his view that interpersonal comparability of utility must be excluded from economic theory generally, and from the analysis of collective economic welfare in particular. However, Pareto (1913) also introduced a formal analysis of social welfare by including arguments pertaining to the welfare of other members of the collective within each individual's utility function. This necessarily presumed that each individual person in society makes interpersonal comparison of utility and, as such, opened the possibility for analysis of redistribution in particular and social welfare more generally. But in extending the scope of the utility function in such a way, Pareto himself abandoned both his own welfare criterion and the use of ordered preferences (relying instead on quantities of utility). Ever since, however, the economic approach to social welfare has been subject to repeated attempts to reinsert both the Pareto criterion and the Paretian preference ordering back into the analysis of social welfare. The works of Abram Bergson (1938) and Kenneth Arrow (1950) are two of the more prominent examples.

Conclusion: Value in a Digital Age

One final problem is specific to the digital age. How does objective value theory apply in cyberspace? What, exactly, does Twitter produce? Which of its employees are productive laborers, which are unproductive, and why? Given that the marginal quantity of labor embodied in the 'product,' whatever that might be, is very close to zero, while the average quantity of labor is very much larger, what is the relevant quantity for the purposes of defining value? Given the importance of intellectual property and hence of 'intangible' or 'fictitious capital' (Marx's phrase) in the digital economy, is there any tendency for the rate of profit to be equalized across the economy? How would Marx – or for that matter Ricardo – have reacted to a world of temporary monopolies, protected by intellectual property laws but subject to constant erosion by new technologies of production and consumption? The difficulties posed by these questions seem rather less damaging to the subjective than to the objective theory of value.

See also: Equilibrium: History of the Concept; Field Theory; Microeconomics, History of; Political Economy, History of; Profits; Utilitarian Social Thought, History of.

Bibliography

Antonelli, G.B., 1886. On the mathematical theory of political economy. In: Chipman, J.S., Hurwicz, L., Richter, M.K., Sonnenschein, H. (Eds.), Preferences, Utility, and Demand. Harcourt Brace Jovanovich, New York, 1971.

Arrow, K., 1950. A difficulty in the concept of social welfare. Journal of Political Economy 58, 238–246.

Auspitz, R., Lieben, R., 1889. Untersuchungen über die Theorie des Preises. Verlag von Dunker und Humblot, Leipzig.

Bergson, A., 1938. A reformulation of certain aspects of welfare economics. Quarterly Journal of Economics 52, 310–334.

Clark, J.B., 1891. Distribution as determined by a law of rent. Quarterly Journal of Economics 5, 289–318.

Clark, J.B., 1899. The Distribution of Wealth: A Theory of Wages, Interest and Profits. Macmillan, New York, 1927.
Dobb, M.H., 1973. Theories of Value and Distribution since Adam Smith: Ideology and Economic Theory. Cambridge University Press, Cambridge.
Edgeworth, F.Y., 5 September 1889. The mathematical theory of political economy. Nature 40, 434–436.
Edgeworth, F.Y., 1897. The pure theory of monopoly. In: Bridel, P. (Ed.), The Foundations of Price Theory, vol. 3. Pickering and Chatto, London, 2001.
Ekeland, R.B., Hébert, R.F., 2002. The origins of neoclassical microeconomics. Journal of Economic Perspectives 16, 1997–2215.
Flux, A.W., 1894. Review of Wicksteed's essay. Economic Journal 4, 305–308.
Fisher, I., 1892. Mathematical Investigations in the Theory of Value and Prices. Yale University Press, New Haven, CT, 1925.
Harcourt, G.C., 1973. Some Cambridge Controversies in the Theory of Capital. Cambridge University Press, Cambridge.
Hicks, J.R., 1939. Value and Capital. Clarendon Press, Oxford.
Hicks, J.R., Allen, R.G.D., 1934. A reconsideration of the theory of value: parts 1 and 2. Economica 1 (52–76), 196–219.
Houthakker, H.S., 1950. Revealed preference and the utility function. Economica 17, 159–174.
Howard, M.C., 1987. Economics on a Sraffian foundation: a critical analysis of neo-Ricardian theory. Economy and Society 16, 317–340.
Howard, M.C., King, J.E., 1985. The Political Economy of Marx, second ed. Longman, Harlow.
Howard, M.C., King, J.E., 1989, 1883–1929. A History of Marxian Economics, vol. I. Macmillan, Basingstoke and Princeton University Press, Princeton, NJ.
Howard, M.C., King, J.E., 1992, 1929–1990. A History of Marxian Economics, vol. II. Macmillan, Basingstoke and Princeton University Press, Princeton, NJ.
Jevons, W.S., 1871. Theory of Political Economy. Macmillan, London.
King, J.E., 2013. David Ricardo. Palgrave Macmillan, Basingstoke.
Kurz, H.D., 2003. The surplus interpretation of the classical economists. In: Biddle, J., Davis, J.B., Samuels, W.J. (Eds.), The Blackwell Companion to the History of Economic Thought. Blackwell, Oxford, pp. 167–183.
Lowry, S. Todd (Ed.), 1987. Pre-Classical Economic Thought: From the Greeks to Adam Smith. Kluwer-Nijhoff, Boston, MA.
Marx, K., 1867. Capital, vol. I. Foreign Languages Publishing House, Moscow, 1961.
Marx, K., 1895. Capital, vol. III. Foreign Languages Publishing House, Moscow, 1962.
Meek, R.L., 1973. Studies in the Labour Theory of Value, second ed. Lawrence & Wishart, London.
Menger, C., 1871. Principles of Economics (Grundsätze der Volkswirtschaftslehre). New York University Press, New York, 1981.
Nove, A., 1983. The economics of feasible socialism. Allen & Unwin, London.
O'Brien, D.P., 2004. The Classical Economists Revisited. Oxford University Press, Oxford.
Pareto, V., 1893-94–2007. In: Marchionatti, R., Mornati, F. (Eds.), Considerations on the Fundamental Principles of Pure Political Economy. Routledge, London.
Pareto, V., 1900. Summary of some chapters of a new treatise on pure economics by professor Pareto. Giornale degli Economisti e Annali di Economia 67 (2008), 453–504.
Pareto, V., 1906-1909. In: Montesano, A., Zanni, A., Bruni, L., Chipman, J.S., McLure, M. (Eds.), Manual of Political Economy: A Critical and Variorum Edition. Oxford University Press, Oxford, 2014.
Pareto, V., 1913-1992. The community's utility maximum in sociology. In: Pasinetti, L. (Ed.), Italian Economic Papers, vol. 1. Oxford University Press, Oxford.
Pullen, J., 2010. The Marginal Productivity Theory of Distribution: A Critical History. Routledge, London.
Ricardo, D., 1821. On the Principles of Political Economy and Taxation, third ed. Cambridge University Press, Cambridge. 1951.
Robinson, J., 1942. An Essay on Marxian Economics. Macmillan, London.
Roncaglia, A., 2009. Piero Sraffa. Palgrave Macmillan, Basingstoke.
Rubin, I.I., 1973. Essays on Marx's Theory of Value. Black Rose Books, Montreal.
Samuelson, P.A., 1938. A note on the pure theory of consumers' behaviour. Economica 5, 61–71.
Slutsky, E., 1915. On the theory of the budget of the consumer. In: Boulding, K., Stigler, G. (Eds.), Readings in Price Theory. Irwin, Homewood, IL, pp. 27–56, 1952.
Sraffa, P., 1960. Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory. Cambridge University Press, Cambridge.
Steedman, I., 1977. Marx after Sraffa. New Left Books, London.
Stigler, G., 1950. The development of utility theory, I and II. Journal of Political Economy 58, 307–327, 373–396.
Stigler, G., 1958. Ricardo and the 93% labour theory of value. American Economic Review 48, 357–367.
Sweezy, P.M., 1942. The Theory of Capitalist Development. Monthly Review Press, New York, 1970.
Sweezy, P.M. (Ed.), 1975. Karl Marx and the Close of His System. Merlin, London.
Walras, L., 1874. Elements of Pure Economics (Élements d'Économie Politique Pure, ou, Theorie de la richesse sociale). Irwin, Homewood, IL, 1954.
Wicksteed, P., 1894. Essay on the Co-Ordination of the Laws of Distribution. London School of Economics, London, 1932.

Value of Life, Economics of

Glenn C Blomquist, University of Kentucky, Lexington, KY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

For practical purposes placing an economic value on life is straightforward. People make tradeoffs between small changes in their probability of survival and finite amounts of valuable time and money. Estimates of these values of changes in mortality risks, or values of statistical lives, come from analysis of jobs, consumption, and direct questioning. Evidence indicates typical values for adults are from $2 million to $12 million (2011 US dollors). Heterogeneity is expected rather than a single, universal value. Evidence that values are lower for seniors is mixed, but for young children values are 1.5–2.0 times higher. Altruism can be relevant and is greatest for children in the household. These theoretically preferred values have mostly displaced cost-of-illness. Gradual displacement of quality-adjusted life years for valuing changes in length of life has facilitated moving beyond cost effectiveness analysis. Deontological and estimation issues aside, economic values of life based on tradeoffs are likely to continue to be useful in policy decisions.

Thinking about the Economics of Value of Life

The economics of the value of life is about what individuals and societies are willing to sacrifice to get longer expected lifetimes. Value of life is about private choices that individuals make implicitly and explicitly about their own health and safety. Value of life is also about collective, public choices that societies make concerning expenditure, tax, and especially regulatory programs that affect mortality risks. Inferences from individual choices inform public decisions about such things as traffic safety and air quality regulations and are a vital input into benefit–cost analysis. Although ethical and estimation concerns exist, these values are useful in policy decisions about health, safety, and the environment.

The Term 'Value of Life'

Value of life typically refers to monetary values of things that individuals are willing to trade for small changes in probabilities of their own survival. These situations are typically what individuals face in life and what decision makers face in making public policy. Value of life is about tradeoffs involving small changes in risk; it is not about willingness to pay to avoid certain death.

Value of a Statistical Life

The typical situation is illustrated by thinking about an individual who is one of a group of 10 000 people. Everyone in the group is identical and knows that during the next year nine people in the group will die. Everyone also knows that the number of people who are going to die could be reduced to eight if sufficient funds can be raised. If somehow it is known that each individual is willing to pay $600 for the reduction of one death in the group, then the value of life is $600 per person times 10 000 people, or $6 million. Sometimes this amount is referred to as the 'value of a statistical life' (VSL) because the identity of the individual who would have been the ninth death, yet lives, is unknown at the beginning of the period when the decision is made.

Value of a Change in Mortality Risk

The numerical example illustrates another way of interpreting value of life. The reduction in the number of deaths implies that the probability of death faced by each individual decreases from 0.0009 to 0.0008. The $600 that the individual is willing to pay reduces the probability of his or her death by 0.0001. Because information is available from similar situations with small changes in the risk of death, but the changes are not exactly the same, for convenience, the value is standardized to a unit change (1–0). The standardized 'value of a change in mortality risk' is $600 times the 10 000 needed to standardize 0.0001 to 1, or $6 million. Both the 'value of a statistical life' and the standardized 'value of a change in mortality risk' represent what is typically meant by value of life. It is the rate at which individuals are willing to trade small amounts of money and risk.

Characteristics of Values of Life, Values of Changes in Mortality Risks

The nature of value of life as defined above can be described further by considering a simple equation of individual expected utility. Let

$$E(U) = PU(C) \qquad [1]$$

where $E(U)$ is expected lifetime utility for an individual. U is a well-behaved utility function with marginal utility $U'>0$ and diminishing marginal utility $U''<0$. C is consumption and P is the probability of survival. Differentiating eqn [1] keeping the individual at the same level of $E(U)$, and solving for $\mathrm{d}C/\mathrm{d}P$ yields the rate at which the individual is willing to trade off current consumption for a small change in the probability of survival:

$$-\mathrm{d}C/\mathrm{d}P = U(C)/PU' \qquad [2]$$

Equation [2] says that the (negative of the) rate at which an individual is willing to trade consumption for a small change in mortality risk equals the utility of consumption divided by the expected marginal utility of consumption. The amount is positive because the tradeoff itself is negative, i.e., less C for more P. If consumption is expressed in monetary terms, U' is the marginal utility of income, and PU' would be expected marginal utility of income. This would mean that the value of a change in mortality risk, $-dC/dP$, would be the expected monetary value of the utility of consumption; it is willingness to pay for risk reductions. For small changes eqn [2] suggests that it makes little difference whether the individual is paying for small reductions in mortality risk or accepting compensation for small additions to mortality risk. For example, for small, equal (and opposite) changes in mortality risk, the amount the individual is willing to pay for a slightly safer job is approximately the same as the amount the individual is willing to accept for a job which is slightly riskier.

While the simple equation above captures the essence of the risk-dollar tradeoff, formal models of consumer and worker behavior include behavior which affects the probability of survival, budget constraints, insurance, workers' compensation, bequests, multiple periods, multiple risks, and risk perception. See Jones-Lee (1976), Rosen (1981), Cropper and Freeman (1991), and Freeman (2003).

Labor Earnings

Within the context of intense interest in economic growth and national income accounting and from an emphasis on manufacturing and machinery following World War II grew an appreciation for the potentially high returns from investment in human beings. In a special conference issue of the *Journal of Political Economy*, Mushkin (1962) articulated the case for investment in human beings through promoting better health. She argued that the return would be through future increases in labor earnings. If one more person were to live because of an investment in health, then national income would be greater by the amount that person would earn in the labor market. Expected future labor earnings became a standard measure of the value of life. Notwithstanding the fundamental contribution made by establishing human health as an important factor in economic growth, eqn [2] yields no clear, precise connection between the value of life and labor earnings. P may be close to 1, and C may be positively related to labor earnings, but the primary difference is that the value of $U(C)$ is not likely to equal C. Assuming that $U(C)$ equals C means that no allowance is made for either consumption which occurs outside the formal market or for the value of living. Instead, in eqn [2] what matters is the tradeoff between consumption and mortality risk to the individual who is affected. Schelling (1968) argued this point persuasively at a Brookings conference more than 40 years ago. The value to the individual influences behavior and should be the starting value for public decisions which affect risks to the individual's life. This point is a reminder that gross national product, or national income measured with traditional accounting, is a highly imperfect indicator of well being.

Costs of Illness

In valuing investments in health, Mushkin (1962) added to earnings the saving of future health expenditures. The sum of these health expenditures saved (direct costs) and the labor earnings not forgone (indirect costs) became known as the 'cost of illness' avoided. Cost of illness was used extensively because of the wide acceptance of the importance of investment in human beings and because estimation was manageable. The shortcoming that homemakers would be valued at zero was overcome by estimating the value of the services performed or the amount of earnings the person would have had if the individual had worked in the labor market. Low values for infants whose future earnings were heavily discounted and zero values for retirees who are not in the market remain awkward implications. The most troublesome aspect of cost of illness, however, is the lack of a theoretical basis which should include the individual's own value of living. Tolley et al. (1994) provide a comparison of the cost of illness and the value of life based on an individual's willingness to trade money and risk.

Ex Ante and Unidentifiable

Equation [2] displays another characteristic which is inherent in the relevant situations. The value of a statistical life with one fewer person of 10 000 people dying and the value of reduction in mortality risk by 0.0001 share an *ex ante* perspective. The situation is confronted, the life lottery must be played, but the outcome is not yet known. The identity of the individual who would have been the ninth death among the 10 000 is not known. Alternatively, whether the individual facing a lower probability of death survives or dies is not known. Decisions which must be made by individuals and societies about risks to their health, safety, and environment are often made in a similar manner.

Situations in which the individual is identifiable are inherently different. If the individual is thrust into a life-threatening situation by an unforeseen natural event, such as a tornado which causes entrapment in a collapsed building, the individual and society will often spend whatever is available or do whatever is possible to save the life. The situation involves a potentially large change in survival for a known individual whose probability of survival has fallen greatly. Saving the individual's life from almost certain death is usually valued highly. If the trapped person is saved, all may end well. If the person cannot be found in the collapsed building after some time, then the probability of saving the person becomes small and the rescue effort decision is transformed into one involving the more typical value of a change in mortality risk. Medical situations involving identifiable patients with known diagnoses have some similarities.

Court cases involving wrongful death of a specific individual are also inherently different because they too are *ex post*. Nothing can change the fact that the individual is dead. Nothing can be done to compensate the deceased for death. The value relevant to forensic decisions is the amount that can compensate the estate for the individual's death. The value depends on the deceased individual's contribution to the well being of the other members of the family, an amount that is related to labor market earnings. From an efficiency

perspective, the value relevant for forensic situations should consider generating sufficient incentive to influence future decision makers who deal with similar risks. Viscusi (2007) compares the different concepts used in public policy and in court cases.

Criticisms of Value of Life

A value of life based on individual preferences as defined above is not without criticism. Criticism ranges from the ethical to the technical.

Deontology

One saying goes that economists know the price of everything and the value of nothing. In this spirit, one objection to the value of life as defined above is that it is vulgar and ethically wrong to place a monetary value on human life. Deontologists such as Plato or Kant espouse moral theory based on obligate moral rules independent of the desirability of the consequences of acting on the rules. If it is morally wrong that people are exposed to risks of death, then it is wrong to construct a value of life to be used to guide decisions involving risks of death.

In contrast, teleologists such as Bentham or Mill whose moral theory is based on the goodness of the consequences are highly interested in a value of life which might guide decisions so as to generate the greatest good for the greatest number. Teleology is the ethical basis for benefit cost analysis including analyses that use values of life (values of small changes in risk) in evaluating policy alternatives. Absolutist deontology can play a useful role in ruling out objectionable policies which many deem morally wrong. However, categorical criticism of applying individuals' values of changes in their own risks in analysis tends to reduce transparency by driving unavoidable tradeoffs from consideration rather than explicitly including them in the analysis. See Brandt-Rauf and Brandt-Rauf (1980), who analyze the ethical conflict over occupational safety and health.

Tension and misunderstanding associated with the value of life concept might well be reduced with terminology that emphasizes small tradeoffs for risks instead of dollars for life. Cameron (2010) makes the case that a term such as microrisk that describes the willingness to swap the value of alternative goods and services for small changes in risks would be a cosmetic change well worth making. Such change is probably unnecessary in countries that accept benefit cost analysis as routine. However, careful wording might calm policy discussion in others, such as the US, and avoid deontological mischaracterizations of changes as devaluation of life, see Viscusi (2009).

No Single Value

Equation [2] is useful for demonstrating that there is no single value of life which applies to all individuals in all circumstances. The amount an individual is willing to trade for a small change in his or her probability of survival depends on P, the base level of the probability. Because P is in the denominator, the value of life can be expected to increase as P decreases. For example, adult aging reduces the probability of survival and can be expected to increase the value of life, holding constant consumption. Automobile drivers who must travel alone at night on winding roads covered with snow and ice (facing lower P) can be expected to have higher values of life than at other times.

The value of life depends on idiosyncratic factors that are reflected in the utility function. $U(C)$ can vary among individuals with the same measured consumption, C, owing to differences in family, environment, or genes. Individuals enjoying close, supportive family relationships and living in clean, ecologically rich natural environments and societies with high-quality civil institutions and public services can be expected to have higher values of life.

Lastly, from eqn [2] the value of life depends upon C. Since C can be expected to increase with income, $U(C)$, and the value of life will be higher with higher income and consumption. Adult aging can be expected to reduce future consumption and, in turn, value of life. The negative effect of aging due to less future consumption (C) with a shorter time horizon and the positive effect due to lower probability of survival (P) make the net effect of aging on value of life theoretically ambiguous and something to be determined empirically. Given the various factors that determine tradeoffs for risks, there is no reason to expect to find a single value even with this simple model. More complex models consider the value of changes in expected health profiles over remaining life where profiles describe the timing, sizes, and sources of mortality risks change, types of morbidity, and recovery. In such models, there are additional reasons to expect differences in values changes in mortality risks, see Cameron and DeShazo (2012).

Quality-Adjusted Life Years

Early evaluation of medical outcomes was in terms of life years added. Desire to recognize quality of life and deontological objection to money values lead to adjusting a life year added by a quality weight to yield a quality-adjusted life year (QALY). The quality weight for a health state associated with an illness (or injury) can be determined by a direct questioning method. The questioning may be as straightforward as picking a point representing the health state on a straight-line scale between death and full health. It might be deciding what probability of full health in a standard gamble with full health and death as the possible outcomes is equivalent to the health state. QALYs place a value on the expected number of years of life.

QALYs emerged to be used in cost utility, or cost effectiveness in evaluating medical interventions for several reasons. One reason was the deontological repugnance for health professionals of putting a monetary value on changes in health and, especially, life. Another was the relative paucity and suspected unreliability of estimates of individual willingness to pay in the 1980s when QALYs were being developed. Given the current abundance of good information about willingness to pay for changes in mortality risks and the tenuous theoretical grounding for QALYs, use of QALYs for benefit–cost analysis involving mortality risks is questionable. For morbidity risks (nonfatal illness or injury) the continued refinement of contingent valuation and the sizable stock of QALY estimates suggest future practice is likely to blend willingness to pay,

QALYs, and related measures such as disability-adjusted life years (DALYs). See Gold et al. (1996), Johannesson (1996), Drummond et al. (1997), and Kenkel (2006). All are consequentialist in that they focus on the values of outcomes.

Estimates of Value of Life

Value of life is about private choices that individuals make implicitly and explicitly about their own health and safety. Bland, healthy diets are traded for tasty food and drink which increase blood pressure and cholesterol levels. Comfortable, safe dwellings are left for skiing and rock climbing. Jaywalking is done to save valuable time. Time and sweat are spent exercising to improve physical condition. Travel is modified to cope with treacherous conditions and increase the chances of a safe trip. Jobs are given up to pursue a safer, healthier lifestyle. For such tradeoffs, if the changes in risk and the values of whatever is traded are known by the individuals and can be known by researchers, then values of life can be estimated by analysis of observable behavior. Even with the insights and complexities that behavioral economics adds, these values can be useful for policy analysis, see Robinson and Hammitt (2011).

In addition, if realistic risk tradeoff scenarios can be constructed, then values of life can be elicited directly. Values of life have been estimated using these approaches at least since the early 1970s. The models upon which the estimation is based typically use a utility function which in some way resembles the one shown in eqn [1], make P endogenous so that individuals can change the probability of survival through their behavior, and finally introduce a resource constraint to reflect limited money and time. An implication of these models is that individuals will engage in risk-changing activity to the extent that the marginal value of the activity equals the marginal cost. Letting V be the value of life as defined in eqn [2], P' the change in the probability of survival, and K the cost of changing P, the condition of interest is

$$VP' = K \qquad [3]$$

If P', the change in risk, and K, the cost of changing P, can be determined, then V, the value of life implied by individual behavior, can be estimated. See Cropper and Freeman (1991), Freeman (2003), Johansson (1995), and Johannesson (1996).

Risk-Compensating Wages

The labor market has proven to be a fertile area for producing estimates of individuals' willingness to trade off safety and money. Faced with an array of jobs with different characteristics, workers will choose jobs that suit them best. If two identical workers are confronted with two jobs which are known to be identical except for risk of death associated with the job and pay, then equilibrium implies that the worker who chooses the riskier job must be compensated by a higher wage. This risk-compensating wage difference implies a value of life.

For example, if the estimated tradeoff is \$0.25 per hour in wage for an additional 0.0001 in annual mortality risk and the individual works 2000 h per year, the wage premium is \$500 per year. Multiplying by 10 000 to standardize to a unit risk change (for convenience) means the implied value of life is \$5 million.

Metaanalyses of the many wage-risk studies done for the US and countries around the world yield a range of estimates from \$2 million to \$12 million with an average nearer to the upper end, perhaps \$7 million. (These values are reported in 2011 US dollars as are all estimates reported below. Adjustments are made for inflation using the Consumer Price Index, but none is made for changes in real income over time.) Several concerns exist about the reliability of these estimates. The fatality rates used to measure risk may not reflect actual risks on the job. Worker information about the risks may be lacking or worker perception of the risks may not reflect actual risks. Risks of nonfatal injury risks may be correlated with fatality risks and may not be measured well or at all. Unmeasured and omitted job disamenities may be correlated with fatality risks. Estimates of values of life may not be representative because of self-selection. Recent studies that use risk data that more clearly match workers' jobs and panel data that control for unobserved worker risk ability and productivity characteristics address a number of these concerns. For a review of the *metaanalyses* and advances made in recent studies, see Cropper et al. (2011).

Risk Tradeoffs in Consumption

Markets other than the labor market and activities other than work have also produced estimates of individuals' willingness to trade off safety and other things they value. By estimating the effect of environmental quality on housing values and isolating the effect of environmental quality on mortality risks, the value of life implicit in the housing market can be estimated. By estimating the effect of the occupant fatality rate on the prices of automobiles, the value of life implicit in the automobile market can be estimated. These situations match well with eqn [3] in that numerous options are available from which the consumer can choose. Other consumption decisions are less continuous, but still imply values of life. Wearing a seat belt while driving in a motor vehicle involves a tradeoff between the gain in safety versus the time and discomfort associated with belt use. Installing smoke detectors involves a tradeoff between a reduction in risk of a fatal fire and installation costs. Blomquist's (2004) review suggests that value of life estimates from consumption studies tend to be less than the estimates from the labor market and fall in the range of \$2 million to \$9 million with a best estimate around \$5 million.

The concerns about the reliability of these estimates are similar to the estimates from the labor market. Some concerns can be readily addressed. For example, if risks are misperceived according to known relationships, the implied values of life can be adjusted. Two other concerns may not be addressed as easily. One is about the ability to separate the risk of fatality from other characteristics of the product. The other is the ability to estimate the value of disutility associated with the consumption.

Stated Preferences in Contingent Markets

Early in the emergence of value of life based on individual willingness to pay, Schelling (1968: pp. 143–144) and Mishan

(1971: p. 174) encouraged economists to employ a questionnaire method to estimate values of life. Since the early 1970s, economists working primarily in environmental, safety, and health economics have developed the stated preference methods. These methods elicit tradeoff choices through direct questioning of individuals who state their preferences about the situation confronting them. The hypothetical setting or market is described, the choice is presented, and information about the respondent is collected. The survey might be in person, by mail, by phone, or some combination. Individuals have been presented with more rapid-response ambulance service for higher taxes, changes in risk on the job for compensating changes in wages, safer highway travel for more money, and safer medications for more out-of-pocket expense. Risk–dollar, risk–risk, and hybrid risk–risk and risk–dollar tradeoffs have been designed to elicit values of life. Contingent valuation, choice experiments, and other stated preference methods have been used in laboratory and field settings for valuing changes in mortality risks due to pollution, travel, and medical treatment. *Metaanalyses* of the many stated preference estimates tend to fall in the range of $2 million to $8 million, somewhat less than the estimates from the labor market and closer to the estimates from tradeoffs in consumption.

A great advantage of contingent valuation is the ability to provide information and describe the tradeoff precisely. Concern exists about reliability. One is related to the challenge of successfully communicating small changes in risk. Another more worrisome concern is potential hypothetical bias due to not actually having to pay. Skepticism remains, but substantial progress has been made through internal validity checks and innovations such as cheap talk pleas for individuals to say what they would really do, convincing individuals that their responses matter and are consequential for policy, and adjusting responses by how certain they are. Future research should increase confidence in these estimates. For a review of the *metaanalyses* and issues, see Cropper et al. (2011).

Heterogeneity

Values for Various Types of Individuals

Benefit–cost analysis of environmental, health, and safety proposals value anticipated changes in mortality risks using the average of values of life for all members of society. In this way the values which individuals place on their own lives inform public policy. For public policies such as improvements in traffic safety which have broad effects across many types of individuals in society, average values of life are appropriate. However, some policies primarily affect specific types of individuals who may have values of life which differ from the typical individual. Whether it is socially acceptable to value changes in mortality risks differently is a sensitive issue. Using different values for differences in gender, race, and lifestyle is usually not considered. A consensus seems to be emerging in the US that it is unacceptable to use different values for groups which differ by income; an average value of life across all of society is normally used. However, adjustments to average values based on income changes over time and across countries are made. Robinson (2007) discusses income and age adjustments in the USA. While no consensus exists yet concerning what is politically acceptable regarding age, research has been done estimating values of life for individuals of different ages. An entire issue of the *Journal of Risk and Uncertainty* (Viscusi, 2010) is devoted to heterogeneity of the value of statistical life with respect to factors such as income and responsibility as well as age.

Values for Children, Working-Age Adults, and Seniors

Values for seniors (age over 65) and children (age under 18) are of special interest because they are vulnerable groups. Life-cycle models of value of life often imply that value of life will vary with age, in part, because of discounting. Discounting future risks can be relevant to people of all ages, but it can matter greatly for the very young, for whom the distant future is discounted greatly. For senior adults there are fewer future periods. A possible offsetting factor for seniors is that they often have accumulated more economic resources. Early contingent valuation studies in the USA and Sweden which address age and discounting find lower values of life for the elderly and that future changes in mortality risks are discounted substantially. See Cropper et al. (1994) and Johannesson et al. (1997). For adults, the issue evolved into a question about discounting the value of life years (VSLYs) and whether VSLY remains constant over the life cycle, see Hammitt (2007).

The vast literature on wage-risk tradeoffs provides estimates of how values of life vary with age for workers. The most recent labor market studies use age-specific fatality rates, estimate by age cohorts, and account for life cycle consumption patterns. The review by Aldy and Viscusi (2007) finds that VSL increases with age peaks at midworklife and declines with the decline flatter than the increase. Stated preference studies have the advantage of including individuals of all ages and those not working in the labor market. Characterizing the effect of age depends on how quality is judged and how much weight is put on different studies, but the review by Krupnick (2007) finds the value of mortality risks appears to be fairly constant during adult life until declining moderately for seniors. The labor market studies point to a value of mortality risks for seniors that is less than the value for midcareer workers, and the stated preference studies give a hint that seniors older than 70 have a value that is only slightly less. A notable recent study by Cameron and DeShazo (2012) that estimates values of mortality risk reductions implicit in values of changes in complex life health profiles finds that values for age 75 are less than half of values for age 65. Currently, no senior discount is made in the US or Europe.

The literature for valuation of mortality risks for children is not as developed as it is for adults, but there has been recent activity. US EPA (2003) developed a special handbook on valuation of children's health risks, but did not suggest different values due to the paucity of estimates. Early estimates based on parents' provision of travel safety for their children hinted at values that were at least equal to parents' value of their own mortality (and morbidity) risks and probably greater, see Blomquist (2004). The Organization for Economic Cooperation and Development (OECD) devoted an entire book (Scapecchi, 2006) to economic evaluation of health risks to children. In that book Dickie and Gerking (2006) model ways in which distributions of resources within households can

influence parents' values of children's risk. They estimate values of reduction in cancer risks for children and find that values for children are markedly greater than parents' values of changes in their own risks. Two other recent, stated preference studies elicit parents' values of reducing fatal risks from consuming pesticide residue in food and from adverse reactions to asthma medication. Both Hammitt and Haninger (2010) and Blomquist et al. (2011) find parents' values of mortality risk reductions are roughly 80% greater than their values of risk reductions for themselves. A calibrated model of valuation of changes in life expectancy in which parents are altruistic and fertility is endogenous produces a similar result; see Birchenall and Soares (2009). OECD (2012) is now willing to recommend no age adjustment for adults, but values of life for children that are 1.5–2.0 times the average adult value. In the US valuing reductions in mortality risks for children more than for adults for policy might be better received than the senior discount. More research on value heterogeneity with respect to age will be valuable. For example, parents values of children's health risks may well depend on resource allocation in the household, characteristics of the parent or child, or opportunities for reducing health risks for each child. Much is known about values for typical adults, but much less about values for the very young and very old.

Values of Others' Lives – Altruism

Values which individuals place on changes in their own probabilities of survival may be less than the value to all of society because the values that others place on the individual's life are not counted. The strongest case for including others' values is within the individual's immediate family. Since people sort themselves into households based on mutual caring it makes sense that interdependencies are greatest within the household. Bonds within the household are thought to be strong enough that decision making is sometimes modeled at the household level and transfers are made within the household. Models of parents' values of children's risks are based on altruism within the household. Beyond the household it is not clear that the values others have for an individual's change in the probability of survival should be included in benefit–cost analysis. In a general model of a society constrained by limited resources, it can be demonstrated that the value of a statistical life is the same with universal pure altruism with caring about the overall well being of others and universal pure self-interest with no caring at all for others. In other words, the individual's value of his or her change in risk need not be supplemented for benefit–cost analysis. This result follows from the recognition that adding others' values causes the individual to consume more health and safety and less of other desirable things than the individual would choose to consume. The individual's value is sufficient as long as there is no substantial concern with the distribution of income. With impure, safety-focused, paternalistic altruism, however, it can be appropriate to augment the individual's value by others' willingness to pay. How much, if at all, altruism outside the household should augment individuals' values of their own mortality risk is a fascinating question that remains a topic for future research. For further readings, see Jones-Lee (1992), Johansson (1995), and Bergstrom (2006).

Public Expenditures and Social Burden

Another reason for considering adjusting the individual value of life is the impact that an individual's death has on others through public tax and expenditure programs. Bailey (1980) argued for the use of individual value of life for guiding policy decisions. One adjustment he made was for the individual's contribution to the public tax and insurance system. The amount of adjustment is potentially greater the more extensive is the welfare state in the society to which the individual belongs. More recently, the observation that lifestyle impacts government expenditures on health has led to something which might be called social burden analysis. The monetary impact on others of individuals who smoke cigarettes, drink heavily, or have sedentary lifestyles has been estimated. Such social burden calculations are not estimates of value of life in that they do not represent values of small changes in mortality risks. They are not individual's values, but are an external effect. Interestingly, if there is a net expenditure burden and it were considered as an adjustment to individual values of life for benefit–cost analysis, adjusted values would be lower than average values of life; benefits of an environmental program would be less than if it affected groups with healthier lifestyles. However, the social burden notion is rarely used this way. Also interesting is the inconsistency with which impacts on government revenues and expenditures are included in the burden analysis. For example, a RAND Corp. study by Manning et al. (1991) estimates that smokers, roughly speaking, pay their own way because of the excise taxes they pay on cigarettes and the amounts they bequeath to survivors through net contributions to pension plans because smokers die younger. Others who do burden analysis, including those who determine what is admissible in court, are more selective in what they are willing to accept as relevant. They will count increases in some medical costs financed by others such as those associated with hospitalization but exclude other medical costs which would be saved, such as costs of long-term care. Social burden analysis awaits a theoretical grounding to put it on equal footing with benefit–cost analysis. Social burden aside, changes in government surplus (revenue-expenditure) can be theoretically appropriate to include along with changes in consumer and producer surpluses in benefit-cost analysis of policies that change mortality risks. Such a benefit–cost analysis would not arbitrarily exclude relevant changes in expenditures and revenues; see Boardman et al. (2011).

What Is Known

What may seem on the surface to be impossible, placing an economic value on life, for practical purposes is straightforward. People, individually and collectively, make choices all the time in which they implicitly make tradeoffs between small changes in their probability of survival and finite amounts of valuable time and money. For more than 40 years economists have recognized these values as conceptually preferred. Estimates of these values of statistical lives, or, alternatively, values of changes in mortality risks, come from analysis of jobs with different wages and risks, consumption decisions involving

changes in risk and time and money, and from direct questioning involving risk-money tradeoffs in constructed markets. The evidence from a large number and wide variety of studies suggests the typical value of life for adults falls in a range from $2 million to $12 million (2011 US dollars) with estimates from consumption and contingent valuation tending to be less than estimates from the labor market. This range may seem wide, but it has proven useful in many policy analyses which often turn on other factors. Theoretically no single value of life is expected. Age is expected to be a relevant characteristic, but possibly in a complex way. Evidence that values are lower for seniors is mixed, but the evidence for children suggests values for young children are 1.5–2.0 times the values for typical adults. Altruism is thought to be the greatest for children in the household where ties are close. The fact that these theoretically preferred, willingness-to-pay values have displaced the early measure, cost-of-illness, matters because they are typically greater although the relationship is not exact. The displacement of QALYs for valuing changes in length of life has allowed moving beyond cost effectiveness analysis among alternative health interventions. Deontological and estimation issues aside, the economic values of life based on tradeoffs for small changes in mortality risks is likely to continue to be useful in policy decisions about health, safety, and the environment.

See also: Children, Value of; Cost–Benefit Analysis; Health Economics; Safety, Economics of; Wage Differential and Structure.

Bibliography

Aldy, Joseph E., Viscusi, Kip W., 2007. Age differences in the value of statistical life: revealed preference evidence. Review of Environmental Economics and Policy 1, 241–260.

Bailey, Martin J., 1980. Reducing Risks to Life: Measurement of the Benefits. American Enterprise Institute for Public Policy Research, Washington, DC.

Bergstrom, Theodore C., 2006. Benefit-cost in a benevolent society. American Economic Review 96, 339–351.

Birchenall, Javier A., Soares, Rodrigo R., 2009. Altruism, fertility, and the value of children: health policy evaluation and intergenerational welfare. Journal of Public Economics 93, 280–295.

Blomquist, Glenn C., 2004. Self-protection and averting behavior, values of statistical lives, and benefit-cost analysis of environmental policy. Review of Economics of the Household 2, 89–110.

Blomquist, Glenn C., Dickie, Mark, O'Conor, Richard M., 2011. Willingness to pay for improving fatality risks and asthma symptoms: values for children and adults of all ages. Resource and Energy Economics 33, 410–435.

Boardman, Anthony E., Greenberg, David H., Vining, Aidan R., Weiner, David L., 2011. Cost-Benefit Analysis: Concepts and Practice. Pearson, Prentice-Hall, Boston.

Brandt-Rauf, Sherry I., Brandt-Rauf, Paul W., 1980. Occupational health ethics: OSHA and the courts. Journal of Health, Politics, Policy and Law 5, 523–534.

Cameron, Trudy A., 2010. Euthanizing the value of statistical life. Review of Environmental Economics and Policy 4, 161–178.

Cameron, Trudy A., DeShazo, J.R., 2012. Demand for health risk reductions. Journal of Environmental Economics and Management. http://dx.doi.org/10.1016/j.jeem.2012.05.008.

Cropper, Maureen L., Aydede, Sema K., Portney, Paul R., 1994. Preferences for life saving programs: how the public discounts time and age. Journal of Risk and Uncertainty 8, 243–265.

Cropper, Maureen L., Freeman, A. Myrick, 1991. Environmental health effects. In: Braden, John B., Kolstad, Charles D. (Eds.), Measuring the Demand for Environmental Quality. North-Holland, New York, pp. 165–221.

Cropper, Maureen L., Hammitt, James K., Robinson, Lisa A., 2011. Valuing mortality risk reductions: progress and challenges. Annual Review of Resource Economics 3, 313–336.

Dickie, Mark, Gerking, Shelby, 2006. Valuing children's health: parental perspectives. In: Scapecchi, Pascale (Ed.), Economic Valuation of Environmental Health Risks to Children. OECD Publishing, Paris.

Drummond, M.F., O'Brien, B., Stoddart, G.L., Torrance, G.W., 1997. Methods for the Economic Evaluation of Health Care Programmes. Oxford University Press, New York.

Freeman, Myrick A., 2003. The Measurement of Environmental and Resource Values: Theory and Methods. Resources for the Future, Washington, DC.

Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, New York.

Hammitt, James K., 2007. Valuing changes in mortality risks: lives saved versus life years saved. Review of Environmental Economics and Policy 1, 228–240.

Hammitt, James K., Haninger, Kevin, 2010. Valuing fatal risks to children and adults: effects of disease, latency, and risk aversion. Journal of Risk and Uncertainty 40, 57–83.

Johannesson, Magnus, 1996. Theory and Methods of Economic Evaluation of Health Care. Kluwer, Boston, MA.

Johannesson, Magnus, Johansson, Per-Olov, Löfgren, Karl-Gustaf, 1997. On the value of changes in life expectancy: blips versus parametric changes. Journal of Risk and Uncertainty 15, 201–219.

Johansson, Per-Olov, 1995. Evaluating Health Risks: An Economic Approach. Cambridge University Press, Cambridge.

Jones-Lee, Michael W., 1976. The Value of Life: An Economic Analysis. University of Chicago Press, Chicago.

Jones-Lee, Michael W., 1992. Paternalistic altruism and the value of statistical life. Economic Journal 102, 80–90.

Kenkel, Don, 2006. WTP- and QALY-based approaches to valuing health for policy: common ground and disputed territory. Environmental and Resource Economics 34, 419–437.

Krupnick, Alan, 2007. Mortality-risk valuation and age: stated preference evidence. Review of Environmental Economics and Policy 1, 261–282.

Manning, Willard G., Keeler, Emmett B., Newhouse, Joseph P., Sloss, Elizabeth M., Wasserman, Jeffrey, 1991. The Costs of Poor Health Habits. Harvard University Press, Cambridge, MA.

Mishan, E.J., 1971. Cost–Benefit Analysis. An Introduction Praeger, New York.

Mushkin, Selma J., 1962. Health as an investment. Journal of Political Economy 70, 129–157.

OECD, 2012. Mortality Risk Valuation in Environment, Health and Transport Policies. OECD Publishing.

Robinson, Lisa A., 2007. How US government agencies value mortality risk reductions. Review of Environmental Economics and Policy 1, 283–299.

Robinson, Lisa A., Hammitt, James K., 2011. Behavioral economics and regulatory analysis. Risk Analysis 31, 217–234.

Rosen, Sherwin, 1981. Valuing health risk. American Economic Review 71, 241–245.

Scapecchi, Pascale (Ed.), 2006. Economic Valuation of Environmental Health Risks to Children. OECD Publishing, Paris.

Schelling, Thomas C., 1968. The life you save may be your own. In: Chase, S.B. (Ed.), Problems in Public Expenditure Analysis. Harvard University Press, Cambridge, MA, pp. 127–176. Reprinted in his book Choice and Consequence. Brookings Institution, Washington, DC, (2006).

Tolley, George S., Kenkel, Don, Fabian, Robert (Eds.), 1994. Valuing Health for Policy: An Economic Approach. University of Chicago Press, Chicago.

U.S. Environmental Protection Agency, 2003. Children's Health Valuation Handbook. EPA 100-R-03-003. http://yosemite.epa.gov/ee/epa/eed.nsf/cbd494e04061784d85256a2b006c1945/6ed3736d44c87a4a85256dc1004da4ac/$FILE/handbook1030.pdf.

Viscusi, Kip W., 2007. The flawed hedonic damages measure of compensation for wrongful death and personal injury. Journal of Forensic Economics 20, 113–135.

Viscusi, Kip W., 2009. The devaluation of life. Regulation and Governance 3, 103–127.

Viscusi, Kip W., 2010. The heterogeneity of the value of statistical life: introduction and review. Journal of Risk and Uncertainty 40, 1–13.

Value Pluralism

Ruth Chang, Rutgers University, New Brunswick, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

'Value pluralism' as traditionally understood is the metaphysical thesis that there are many values that cannot be 'reduced' to a single 'supervalue.' While it is widely assumed that value pluralism is true, the case for value pluralism depends on resolution of a neglected question in value theory: How are values properly individuated? Value pluralism has been thought to be important in two main ways. If values are plural, any theory that relies on value monism, for example, hedonistic utilitarianism, is mistaken. The plurality of values is also thought to raise problems for rational choice. If two irreducibly distinct values conflict, it seems that there is no common ground that justifies choosing one over the other. The metaphysical plurality of values does not, however, have the implications for rational choice that many have supposed. A charitable interpretation of value pluralist writings suggests a 'nonreductive' form of value pluralism. Nonreductive value pluralism maintains that in the context of practical choice, there are differences between values – whether or not those values reduce to a single supervalue – that have important implications for rational choice. This article examines the main arguments for metaphysical value pluralism, argues that metaphysical value pluralism does not have certain implications that it is widely thought to have, and outlines three forms of nonreductive value pluralism.

'Value pluralism' as traditionally understood is the metaphysical thesis that there are many values that cannot be 'reduced' to a single 'supervalue.' Although value monism has an impressive pedigree of proponents (e.g., Bentham, Mill, and, arguably, Aristotle and Nietzsche among others), it is now widely assumed to be false. The arguments for value pluralism, however, are surprisingly inconclusive. Indeed, as we will see, they critically depend on the resolution of a neglected question in value theory: How are values properly individuated?

The metaphysical plurality of values has been thought to be important in two main ways. First, if values are plural, any theory that relies on value monism is mistaken. So, for example, hedonistic utilitarianism and most forms of preference-utilitarianism must be rejected. Second, the plurality of values is thought to raise problems for rational choice. If justice, for instance, is irreducibly distinct from mercy, how can there be rational choice between them? Without reduction, it seems there is no common ground that justifies choosing one value over another.

On closer inspection, however, it turns out that metaphysical or reductive value pluralism does not have the implications for rational choice that many have supposed. A charitable interpretation of value pluralist writings suggests a second, nonreductive, form of value pluralism. Nonreductive pluralism is neutral on the metaphysical question of plurality but insists that in the context of choice, there are differences between values – whether or not those values reduce to a single supervalue – that have important implications for rational choice. Since the differences claimed to hold between values vary from author to author, nonreductive value pluralism is not itself a particular view about values but merely a convenient rubric under which a loose collection of different views about values may be grouped.

This article examines the main arguments for reductive value pluralism, argues that reductive value pluralism does not have certain implications it is widely thought to have, and outlines three forms of nonreductive value pluralism.

Reductive Value Pluralism

Value pluralists maintain that whatever values are, there are ultimately many of them: They do not all reduce to a single ultimate value (Berlin, 1950, 1959; Nagel 1979; Richardson 1994; Taylor 1982). Exactly how this metaphysical thesis is to be understood depends on how the notion of reduction is to be understood.

Reduction in the context of values is best understood as an explanatory relation: If one value reduces to another, what it is to bear the one value is fully explained by what it is to bear, promote, or respect the other value. This reduction is neutral on the general ontological question of whether there 'really' are any values at all. The pluralist maintains only that there are many values, whether or not they are to be regarded as entities in their own right.

Two paradigmatic relations of explanatory reduction are 'is merely instrumental to' and 'is wholly constituted by.' (Others include 'is merely symbolic of,' 'is merely contributory to,' 'is merely a part of'). If one value is merely instrumental to another, there is nothing more to having the one value than promoting the value it is a means to. For example, if beauty is merely instrumental to pleasure, what it is to be beautiful is fully explained by the pleasure it brings. If one value is wholly constituted by another, there is nothing more to having the one value than being a way in which the other value is borne. For example, if the evening's pleasure is wholly constituted by the thrill at the gaming tables, what it is to have the evening's pleasure is fully explained by the thrill it is constituted by.

If all values reduce to a single value, that value is the only ultimate value, and value monism is correct. If, however, there is something more to two or more values than the values they are instrumental to, constituted by, etc., then those values are irreducibly distinct, and value pluralism is correct.

Three Arguments for Reductive Value Pluralism

Although there has been no agreement on which is the one value to which all others reduce, the view that there is such a value has

two main attractions. If all values reduce to one, then values can be neatly systematized as instruments to, constituted by, etc., a single supervalue. Value monism appears to ensure a simple and elegant axiology, with a supervalue at the trunk of a structure that branches out to the other values that each derive from the supervalue in some way.

Moreover, if value monism holds, it seems that all conflicts between values are only apparent. For if there is ultimately only one value, then options for choice can be tidily arrayed according to how much of the supervalue they bear, promote, or respect. Any evaluative choice would ultimately be a choice between two amounts of the supervalue. Choosing between values would always be like choosing between two lumps of coal or three. Thus, value monism seems to dissolve the threats to practical rationality posed by tragic choices, moral dilemmas, and 'incommensurable' options.

Despite these attractions, most contemporary value theorists assume that value pluralism is true. There are three main arguments for pluralism. First is the intuitive implausibility of value monism: Given the apparent diversity of values, how could there be a single value 'common to' all valuable items? Second is the thought that *akrasia* (actions against one's better judgment) can be explained only if values are plural: How can it make sense for someone to choose something that she believes is worse overall unless there is something attractive about the worse option that is not 'included' in the better one? A closely related third argument maintains that some choice situations involve unavoidable loss: No matter which alternative one chooses, something valuable will be forgone, and thus, there must be plural values at stake in the choice.

Ordinary Intuition

Value monism seems to run afoul of common sense. Two intuitions suggest that any monistic account will be vulnerable to counterexample. First, how could there be just one value that runs through valuable items as diverse as, for example, achieving philosophical insight and eating a slice of delicious cheesecake? While perhaps the value of both options is instrumental to or constituted by some value (e.g., pleasure), it is hard to believe that bearing, promoting, or respecting that value is ultimately all there is to their being valuable. Second, if there were a single value that exhausted the value of all valuable things, the evaluative difference between things could always be given by some *amount* of the supervalue. But how could the difference in value between achieving philosophical insight and eating delicious cheesecake be a matter of quantity of some one thing? It is incredible to think that evaluative differences among diverse goods are just a matter of more or less of a single value.

To these charges, monists offer two replies. The first is to suggest that the troublesome intuitions rely on an uncharitable view of monism, one according to which what it is to bear the supervalue is to bear or bring about a kind of feeling or experience like pleasure or happiness. It is, indeed, hard to believe that all valuable things involve having or producing a pleasurable or happy experience. But monism need not be so crude. Some monists propose instead that the ultimate value is given by the satisfaction of one's 'fully informed' or 'rational' desires or preferences. The one property that runs through all valuable items, then, is the property of satisfying constrained desires, and it is the value of that property that is the supervalue. Evaluative differences between items are a matter of the strength or number of desires they satisfy.

Whether desire-satisfaction accounts of the supervalue are plausible depends on what it is for a desire to be fully informed or rational. Full-information is usually understood as having all the relevant facts and being free from logical error. Rational desires might be procedurally rational (passing certain evaluatively neutral tests) or substantively rational (passing certain evaluative tests). Insofar as full-information or rational is a value-neutral constraint, there is no guarantee that such constrained desires or their strengths will track what is intuitively valuable or more valuable: It is perfectly possible that everyone have fully informed or procedurally rational desires to perform evil and malicious deeds, and that the desires for those deeds be stronger than desires for angelic ones. If, in an effort to secure this tracking, one places evaluative constraints on desires, as one might do by insisting that desires be substantively rational, then one must give up monism. For the values that operate as constraints on desires must be distinct from the value they constrain, and so there are ultimately many values, not one.

Monists have a second, more promising, reply. This reply was introduced by Mill in an effort to defend utilitarianism against counterintuitive consequences of Bentham's quantitative hedonism. There are, Mill insisted, not only different *quantities* of pleasure, but different *qualities* of pleasure. The pleasure of philosophical insight is a 'higher' pleasure than the pleasure of eating delicious cheesecake. Thus, while there is ultimately only one value, there are different qualities of it that explain the seeming diversity of values. Moreover, the evaluative differences between valuable items are not simply given by some *amount* of the supervalue; achieving philosophical insight and eating cheesecake may differ in the *quality* of the supervalue they bear or instantiate – one pleasure is 'higher.'

This reply gets to the heart of the dispute between monists and pluralists. The monist insists that the commonsense belief in multiple values is in fact a belief about different qualities or aspects of a single value. Whether this position is plausible depends on how values are to be individuated. What makes two considerations two distinct values as opposed to two qualities or aspects of a single value? Unfortunately, this question has received almost no philosophical attention. Until the question of how values are to be individuated is settled, ordinary intuitions cannot provide good reasons for thinking value pluralism is true.

Akrasia

Some philosophers (Wiggins, 1978; Nussbaum, 1986) have argued that for *akrasia* to be explained as a coherent phenomenon, plural values are required, and since *akrasia* is a coherent phenomenon, it follows that there are plural values. This argument is often presented as establishing the 'incommensurability' of values, but incommensurability is used here, misleadingly, as synonymous with or as entailing plurality.

If one judges that one has most reason to choose one alternative but instead chooses the other, one is weak of will. Suppose one must choose between going to a party and staying home to work. Although one believes that staying home to work has the greater value – and therefore that one has most reason to

stay home – one chooses to go to the party. How can such a choice be explained? If monism is correct, it seems that weakness of will must be an incoherent phenomenon. For suppose that there is a single ultimate value, say, pleasure. It seems, then, that there is nothing about the lesser option that could possibly attract one to it. For everything the lesser option has going for it – a certain amount of pleasure *X* – is included in what the greater option has going for it – *X* plus more pleasure besides. One cannot choose to stay home for a reason, no matter how bad, for there is no reason for choosing the one option that is not already a reason for choosing the other.

The argument, however, does not succeed in establishing value pluralism. For akratic choice can be explained simply by attributing to the akrates a *belief* that there are plural values at stake, there need not actually *be* plural values. One might mistakenly believe that the value of going to the party is irreducibly distinct from the value of staying home and thus be attracted by the allure of the former. One's reason for going to the party, while based on a false belief, could nevertheless be a reason that makes one's weakness of will coherent.

But perhaps for *akrasia* to be coherent in an 'objective' sense, that is, coherent for agents with, inter alia, relevant beliefs that are true, plural values *are* required. This version, however, also fails. One possible difficulty is that what grounds the akrates' attraction to the lesser option may be not the plurality of the values at stake but rather some contingent feature of the circumstances in which the value of the lesser option is instantiated or realized. So, for example, although there might ultimately be only pleasure at stake, the fact that the lesser pleasure of going to the party occurs in a seedy part of town has a special allure for the akrates. There are arguably no plural values here, only circumstances extrinsic to the supervalue whose special appeal to the agent might provide a reason for choosing the lesser option.

A more significant problem is the fact that attraction to the lesser option can be explained by appeal to different qualities or aspects of the supervalue. Take, for example, a choice between two glasses of wine, one laced with pepper and one not (Stocker, 1990). Although one judges that the unadulterated wine will provide the greater pleasure, one is charmed by the particular pleasure of the wine laced with pepper and so chooses the lesser pleasure for the reason given by its particular charm. Similarly, one might choose to go to the party while judging that staying home is better for the reason that the party option provides a particular quality of the supervalue that staying at home lacks.

The success of this reply depends on an account of the individuation of values. Is what attracts about the lesser option a value distinct from the value of the greater option or some distinctive quality of an ultimate supervalue?

Unavoidable Loss

In some choice situations, it seems that no matter which alternative one chooses, something valuable will be lost. If, for example, one must choose between attending a lecture on Kant's ethics and having cheesecake with friends, one is bound to lose out on value no matter which alternative one chooses – and not simply because one cannot have both. If one chooses to attend the lecture, one will forgo the gustatory pleasure of eating cheesecake; if one chooses to go out for cheesecake, one will forgo the philosophical insight one would have gained at the lecture. The fact that any choice in such situations entails value loss shows that the choice involves a conflict between distinct values. Therefore, there are plural values. (Essentially the same argument is sometimes put in terms of the possibility of rational regret over having foregone a lesser good.)

Monism, it seems, must deny that such cases exist. For if monism were true, then choice would always be a matter of choosing more or less of a single value. If one chooses the option with more value, there is no value loss. If one chooses the option with the lesser value, then there is value loss, but it is not unavoidable.

Both the argument from *akrasia* in its 'objective' form and the argument from unavoidable loss rely on the conviction that values are sometimes *incompatible* with one another: Pursuit of (respect for, instantiation of, etc.) one excludes pursuit of (etc.) the other. It is the incompatibility of values that grounds the akrates' attraction to the lesser option and that ensures that no matter which value one chooses, there will be value loss. This incompatibility of values might be conceptual, as Isaiah Berlin believed was true of value concepts like *justice* and *mercy*, or an intrinsic feature of values given the circumstances in which they arise. If values are incompatible, it seems that they cannot be reduced to a common value.

The monist reply here is already familiar from the previous two arguments. Sophisticated monists of the Millian variety might insist that there can be incompatibilities among different *qualities* or aspects of a single value. In cases of unavoidable loss, there need not be more than one ultimate value. The languorous pleasure of basking in the sun might be incompatible with the piquant pleasure of hearing unexpected good news: Having the one pleasure now rules out having the other now, though they are nevertheless instances of a single value, pleasure (cf. Stocker, 1990). Similarly, the qualitative value of achieving philosophical insight may be incompatible with the qualitative value of eating delicious cheesecake. Choosing between the lecture and the cheesecake involves a loss in value no matter how one chooses, but there is only one ultimate value.

All three arguments for value pluralism crucially depend on the answer to the question, how are values properly individuated? The fundamental idea behind the common monist response to pluralist arguments is that there is a *nonultimate* sense of value according to which it can be true that there are distinct values and yet those values cannot be values in the robust sense implied by the claim that there are two distinct *ultimate* values.

Exactly how this is to be worked out remains to be seen. Hurka (1996) has suggested that one pleasure is not a different ultimate value from another pleasure if it does not differ in any of its intrinsic features. In explaining what it is for a feature to be intrinsic to a value, Hurka urges that the location of a value is extrinsic to it. Thus, John's pleasure and Joanna's pleasure are two instances of the same value, pleasure. Stocker (1990), however, has suggested that if pleasure is located in different people, there are different ultimate values: There is the pleasure-for-John and the pleasure-for-Joanna. Stocker (1990, 1997) maintains that if it is rational to care about any difference between two evaluative considerations, it follows that those considerations are

different ultimate values. Thus, there is sharp disagreement about how values are properly individuated.

Monism, if it is to be plausible, must hold that the correct account of value individuation leaves room for a tenable notion of a value quality or aspect that serves to undermine intuitions that values are plural, accounts for weakness of will, and explains away choices that seem to involve unavoidable loss. If there is such a notion, then value monism may yet win the day. But any such victory for monism may be hollow. For if monism is to be cluttered with complex relations among value qualities or aspects, the two chief attractions of monism – in short, the tidiness and simplicity it seemed to offer – fall by the wayside. In the end, then, the debate between value pluralists and value monists may be much ado about nothing.

Implications of Reductive Value Pluralism

Assuming that there are ultimately many values, what, if anything, of interest follows? There are three main implications that are thought to be important.

First, if reductive value pluralism is true, then reductive value monism and any normative theory that relies on it must be rejected. So, for example, Benthamite utilitarianism has to go. Perhaps more significant is the result that preference utilitarianism, or, more generally, desire-satisfaction theories of value must go. (This is not to say that desire-satisfaction theories of what it is for something *to be* a value must be rejected. One might hold that every value must satisfy some or other desire to be a value while denying that all values reduce to the value of desire satisfaction.) Many economists and some philosophers favor desire-satisfaction accounts of value, and their accounts at the very least require modification if value pluralism is true. Economists who model rational choice on satisfaction of preferences, for example, must relinquish any claim that *goodness* is a matter of satisfaction of one's preferences and instead maintain only that the rationality of choice depends on one's preferences conforming to certain axioms (cf. Broome, 1991). Philosophers who think that individual well-being is the only value there is and that what makes one's life go best is satisfaction of one's desires must give up either the claim that well-being is the only value there is or the claim that well-being is simply a matter of desire-satisfaction.

Some have thought that value pluralism shows that any form of utilitarianism is mistaken. But what is essential to utilitarianism is compatible with value pluralism (Sen, 1981). Others have thought that Kantian ethical theory must be monistic, but neo-Kantians have argued that Kantian ethics is compatible with a plurality of fundamental principles (Hill, 1992). Pluralists have urged that value pluralism is the foundation of political liberalism: If values are metaphysically plural, then liberalism is the correct theory of justice (Galston, 2002). But the metaphysical plurality of values is plausibly neutral between liberalism and nonliberalism. The fact that there are ultimately many values does not entail that a state should not compel its citizens to pursue one value over others, nor does monism entail that a state should not protect its citizens' choices to pursue nonultimate values in ways that do not best promote or respect the supervalue. As some political theorists have pointed out, liberalism itself might be understood as monistic about value: The ultimate value is the value of permitting people to pursue different nonultimate values. Whether there is, in the end, some indirect way in which value pluralism supports liberalism remains an open question.

The classical pluralist theory of morality is W.D. Ross's intuitionism, which posited a multitude of plural foundational, self-evident principles of prima facie duty (Ross, 1930). Many contemporary philosophers have followed in his wake, arguing that there are multiple fundamental values (Nagel, 1979; Finnis, 1980; Griffin, 1986).

A second implication of value pluralism is that it allows for indeterminacy in what ought to be done and thus underwrites the legitimacy of normative disagreement. If values in conflict are irreducibly plural, there may be no truth of the matter as to how they are to be weighed against one another. If you disagree with me about what ought to be done, value pluralism allows that we may both be right. This indeterminacy in what ought to be done in turn entails a kind of relativism – not the sort according to which what is right for you is different from what is right for me, but rather a relativism within each individual as to what ought to be done. For each of us, there may be no determinate right answer as to what we ought to do (Wolf, 1992).

Value pluralism has been thought to have a third implication. If value pluralism is correct, then difficulties supposedly follow for rational choice and deliberation (for general discussion, see Railton, 1992). The difficulties alleged are various, but the reason for thinking that none follows from value pluralism is the same. Many have supposed, for example, that alternatives for choice bearing plural values are *incomparable*. How can the option of respecting someone's right to free speech be compared with an option that doubles everyone's pleasure? Insofar as the value of a right to free speech is irreducibly distinct from the value of pleasure, it seems that there is no common basis on which to make a comparison between their bearers. The argument goes, if options for choice bearing plural values are incomparable, rational choice between them is precluded.

There is good reason, however, to think that the plurality of values does not entail the incomparability of options that bear them. Take any two putatively ultimate values such as the value of the right to free speech and the value of pleasure. One can always imagine some option that bears the one value in a notable way that can be compared with another option that bears the other value in a nominal way. So, for instance, an option that involves violating everyone's right to free speech is worse than an option that involves reducing one person's pleasure by a small amount. Given any two putatively irreducibly distinct values, there will always be some comparison between a notable bearer of the one value with a nominal bearer of the other value. The existence of 'nominal-notable' comparisons demonstrates that if alternatives are incomparable, it is not the plurality of values per se that entails their incomparability (Chang, 1997). Assuming that rational choice depends on the comparability of alternatives, the plurality of values, then, does not preclude rational choice in this way.

Although value pluralism does not entail incomparability, perhaps the issue of value pluralism is important because if value monism is true, complete comparability follows. If there is ultimately only one value, evaluative differences between items must always reduce to differences in amount or quality of the

supervalue, and quantities or qualities of the same thing can always be compared. On some views, if alternatives are comparable, the existence of a rational choice is assured. Thus, monism seems to guarantee that there is a rational choice in every choice situation, no matter how intractable the conflict may seem.

The inference from monism to comparability, however, is mistaken. It is not clear that different qualities of a single value can be compared. How can one compare the luxurious, wallowing pleasure of lying in the sun and the intense, sharp pleasure of quenching a fierce thirst? Making the case for comparability is arguably just as hard under the assumption of monism as it is under the assumption of pluralism. Moreover, even assuming that the supervalue does not admit of different qualities, it is a mistake to assume that all *quantities* of a single supervalue are comparable. Pleasure can be greater in amount without the value of that pleasure being thereby greater. The five-course feast that ends with petit fours involves a larger quantity of pleasure than the five-course feast without, for the petit fours add a quantity of pleasure to the pleasure already had. But the additional quantity of pleasure may make the meal less valuable with respect to pleasure – the petit fours induce pleasure 'overkill.' Since a larger quantity of pleasure is not always better with respect to pleasure, it seems possible in principle that different quantities of a single value are incomparable. Assuming that there are such genuine cases of monistic conflict, the issue of value pluralism cuts across the question of whether bearers of value are comparable or incomparable.

Nonreductive Value Pluralism

In general, the question of how many ultimate values there are is irrelevant to the question of how conflicts between them are to be resolved. What matters in conflict resolution is how the values relate to one another in the context of choice, not how they stand to one another metaphysically. As already suggested, a monist might happily allow conflicts between value qualities, while insisting that, as a metaphysical matter, there is ultimately only one value.

Given that reductive value pluralism does not have the implications for practical reason that many value pluralists seem to think it has, it is natural to think that there is some other notion of value pluralism besides the metaphysical one. There is, however, no single specific idea that would account for the concerns of all value pluralists who think the 'plurality of values' has implications for rational choice. At best, the cluster of views can be generally characterized as holding that there is some or other difference among values – something apart from mere metaphysical numerical difference – that has important implications for rational choice. The loose collection of views about the ways in which differences between values can have practical implications might be called nonreductive value pluralism.

Nonreductive value pluralism comes in many different varieties, but there are three leading kinds (for general discussion see Griffin, 1997). First is the view that some values are *incompatible* that they cannot be pursued (instantiated, etc.) together. The incompatibility of values raises the possibility of unavoidable loss of value and, more alarmingly, the specter of moral dilemmas in which the unavoidable loss of value is expressed by a moral requirement.

Second is the view that some values are so different that they – and therefore their bearers – are *incomparable*. Suppose you must choose between a career as a clarinetist and one as a lawyer (Raz, 1986). If the values of those careers are incomparable, and if their incomparability precludes rational choice between them, it seems that whichever career one chooses, the choice cannot be rational. Thus, this kind of nonreductive value pluralism has important implications for practical reasons. It suggests that there is perhaps a wide swath of choices that are beyond the scope of practical reason.

A third view is that some values are so different that their bearers cannot be measured by the same cardinal scale of value (cf. Sunstein, 1994; Anderson, 1993, 1997). Suppose you must choose between guns and butter. Since there is no cardinal scale along which the value of guns and butter can be measured, the alternatives are *incommensurable*. Many pluralists have thought that rational choice between incommensurable options is impossible. But it is clear that the lack of cardinal comparability does not preclude rational choice. One alternative might be better than another though there is no cardinal scale according to which it is better. Lexically superior options, for instance, cannot be measured on the same cardinal scale as the options they are lexically superior to, but they are nevertheless better. Moreover, the failure of precise cardinal comparability need not entail incomparability. Items may be 'imprecisely' comparable (Parfit, 1984; Ms; Griffin, 1986) or 'indeterminately ranked' (Seung and Bonovac, 1992) or 'on a par' (Chang, 2002). Although incommensurable options do not preclude rational choice, the fact that they are incommensurable entails that any justification for choosing one over another cannot appeal to the idea that there is some amount of one that provides the equivalent *measure* of some amount of the other.

See also: Consequentialism (Including Utilitarianism); Economics and Ethics; Relativism, Cognitive.

Bibliography

Anderson, E., 1993. Value in Ethics and Economics. Harvard University Press, Cambridge, MA.
Anderson, E., 1997. Practical reason and incommensurate goods. In: Chang, R. (Ed.), Incommensurability, Incompatibility and Practical Reason. Harvard University Press, Cambridge, MA.
Berlin, I., 1950. Concepts and Categories. Oxford University Press, Oxford, UK.
Berlin, I., 1959. The Crooked Timber of Humanity. Random House, New York.
Broome, J., 1991. Weighing Goods: Equality, Uncertainty and Time. B. Blackwell, Cambridge, MA.
Chang, R., 1997. Introduction. In: Chang, R. (Ed.), Incommensurability, Incomparability, and Practical Reason. Harvard University Press, Cambridge, MA.
Chang, R., 2002. The possibility of parity. Ethics 112, 659–688.
Finnis, J., 1980. Natural Law and Natural Rights. Clarenden Law Series, Oxford.
Galston, W., 2002. Liberal Pluralism: The Implications of Value Pluralism for Political Theory and Practice. Cambridge University Press, Cambridge.
Griffin, J., 1986. Well-Being: Its Meaning, Measurement and Moral Importance. Clarendon Press, Oxford, UK.

Griffin, J., 1997. Incommensurability: what's the problem? In: Chang, R. (Ed.), Incommensurability, Incomparability, and Practical Reason. Harvard University Press, Cambridge, MA.
Hill, Thomas, 1992. Kantian pluralism. Ethics 102, 743–762.
Hurka, T., 1996. Monism, pluralism, and rational regret. Ethics 106, 555–575.
Kekes, J., 1993. The Morality of Pluralism. Princeton University Press, Princeton, NJ.
Nagel, T., 1979. Mortal Questions. Cambridge University Press, Cambridge, UK.
Nussbaum, M.C., 1986. The Fragility of Goodness. Cambridge University Press, Cambridge, UK.
Parfit, D., 1984. Reasons and Persons. Oxford University Press, Oxford, UK.
Parfit, D. (Ms). How to Avoid the Repugnant Conclusion.
Railton, Peter, 1992. Pluralism, determinacy, and dilemma. Ethics 102, 720–742.
Raz, J., 1986. The Morality of Freedom. Oxford University Press, Oxford, UK.
Richardson, H.S., 1994. Practical Reasoning about Final Ends. Cambridge University Press, Cambridge, UK.
Ross, W.D., 1930. The Right and the Good. Oxford University Press, Oxford, UK.
Sen, A., 1981. Plural utility. Proceedings of the Aristotelian Society 81, 193–215.
Seung, T.K., Bonovac, D., 1992. Plural values and indeterminate rankings. Ethics 102, 799–813.
Stocker, M., 1990. Plural and Conflicting Values. Oxford University Press, Oxford, UK.
Stocker, M., 1997. Abstract and concrete value: plurality, conflict and maximization. In: Chang, R. (Ed.), Incommensurability, Incompatibility and Practical Reason. Harvard University Press, Cambridge, MA.
Sunstein, C., 1994. Incommensurability and valuation in law. Michigan Law Review 92, 779–851.
Taylor, C., 1982. The diversity of goods. In: Sen, A., Williams, B. (Eds.), Utilitarianism and Beyond. Cambridge University Press, Cambridge, UK.
Wiggins, D., 1978. Weakness of will, commensurability and the objects of deliberation and desire. Proceedings of the Aristotelian Society 79, 251–277.
Wolf, S., 1992. Two levels of pluralism. Ethics 102, 785–798.

Values Across Cultures, Development of

Purnima Singh, Indian Institute of Technology Delhi, New Delhi, India

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The present article examines values and their development across cultures. Since values are central to any understanding of the person as well as the society, they are said to represent desirable goals and guiding principles in people's lives. The article also examines some of the social correlates of values. Values of social collectives representing the cultural level shared values are explored. Examining the research related to shared cultural values, the article suggests the need to look at within cultural variations in values, more specifically examining the differences in values among social and cultural groups so as to have a more nuanced understanding of values in different cultural contexts.

Values and Social Life

The proposition that individuals develop in cultures and that culture frames their values is well accepted in the literature. In other words, how we behave, how we perceive our own and others behavior and how we address social issues and take positions in our social, cultural, and political lives are all influenced by our values. Today, the inextricable and complex linkage of values and culture is even more intense as it has been found that at both personal and cultural levels, values have an impact on the social interactions, judgments, and explanations. Any analysis of values will be incomplete without examining both personal and cultural values because both impact an individual's behavior and social relations. Personal values are the broad socially desirable goals and vary in the importance assigned to them from person to person. They have been posited to serve as guiding principles in people's lives. Divergences in different cultures often induce large variations in personal values. For example, some cultures give more importance to certain values while others give importance to some other values. Asian cultures may give more importance to the traditions than the Western cultures. Even among Asians, individuals may give different degrees of importance to traditions. Thus, we see that individuals belonging to a particular culture may share some values and at the same time differ in other values with their fellow beings. In today's scenario, the composite nature of different cultures may perhaps minimize these differences across cultures but at the same time within a culture, differences in values may get accentuated. It is also important to mention here that culture is not a fixed set of well-defined parameters. Asian and Western cultures have large variations, not only in terms of degree of importance or acceptability of different values but also in terms of intrinsic nature of these values. Even different regions, nations, and states within the same cultural context may have variations in values preferences. But despite saying this, we do find some common strands in values across cultures.

As has been discussed above, it is true that both personal and cultural values, influence the perceptions of legitimacy, acceptance, and effectiveness of behaviors of individuals, groups, and institutions as well as the rejection of certain behaviors. We imbibe and inculcate values as we see and meet others, gain knowledge about acceptable and unacceptable social norms and standards, and adapt to customs and practices held important in one's culture. Values have also been seen to have social functions such as to foster development and progress, facilitate cooperation or competition between groups, and create conflicts as well as to promote harmony, happiness, and peace. In order to understand the numerous social functions of values, it is important to understand the role of various social entities who judge or categorize certain behaviors as socially appropriate or inappropriate, disapprove of others, and even justify their own behavior and demands, based on their values. Before we start delving deep into the role of values and their importance across cultures, it is important to understand what exactly values are. Most scholars see values as representing desirable goals that function as principles in the life of individuals, groups, and societies. They are seen to reflect conceptions of what is preferred and considered appropriate in a given social context. Another important feature of values is that all values are organized hierarchically both in the case of an individual, the society, as well as the culture. Some values may be considered as more important by an individual and some other values by other individuals. This variedness is seen both at the societal level and the cultural level. In other words, not only individuals have a unique value hierarchy, but societies and cultures may also have their own unique hierarchy of values. One set of theories concerning values have postulated that by and large, although people differ in terms of their value priorities, the structure of the human value system is universal (e.g., Schwartz, 1992). People may differ in terms of the relative importance they place on a set of universally important value types.

Values of social and cultural groups allow us to understand shared ideas that a particular social and cultural group considers right and desirable, encourage these in their members, and may consider those who do not espouse these values as outsiders. The maintenance of boundaries between persons, groups, and societies is facilitated and shaped by values. Today, when we are all influenced by the forces of globalization and the fast changes taking place both in the societies and in different cultures worldwide, values seem to be the most affected by these changes. This significance of values to individual and social life, however, has been recognized since long by psychologists, sociologists, anthropologists, philosophers, and other social scientists (Rokeach, 1973;

Schwartz, 1992) .Time and again people have been discussing and debating about values, lamenting the decline in values, and reasserting their impact on behavior, emotions, decisions, and social interactions. Kluckhohn (1951), an anthropologist, was one of the earliest scholars to examine the concept of values. His writings paved the way for later work by providing a definition of values "as a conception, explicit or implicit, distinctive of an individual or characteristics of the group, of the desirable which influences the selection from available modes, mean and ends of action" (p. 395).

One of the more comprehensive research on values was conducted by Rokeach (1973) wherein he defines "a value is an enduring belief that specific mode of conduct or end-state of existence" (p. 5). He posited two set of values – terminal values and instrumental values. By terminal values he referred to the desirable, end states of existence. These consist of goals that a person would like to achieve during his/her lifetime. Different groups of people in various cultures may differ in terms of their terminal values. Instrumental values on the other hand are the modes of behavior that are preferred, which are means of achieving the terminal values.

Schwartz (1992) in his widely acknowledged work, proposed that values pertain to the desirable end states or behaviors. Summarizing the main strands that run through various conceptualizations of values, he highlighted the following attributes of values: values are motivational constructs that transcend specific actions and situations and guide the selection or evaluation of actions, policies, people, and events. Schwartz and Sagiv (1995) further clarified the concept in their articulation of values as "transitional goals (terminal or instrumental)" that reflect both individual and collective goals that are guiding principles in a person's life.

Schwartz (1992) proposed 10 types of values that are distinct but recognized by people in most societies. These include the following. Conformity, which emphasizes going along with others. As a result, those actions that are likely to harm others or will not be liked by others are generally not given importance. Tradition focuses on respect and acceptance. A strong commitment to customs and traditions are endorsed. Values espoused by different religions also come under tradition. Benevolence refers to enhancing the welfare of others or being nurturant. Another value called universalism emphasizes understanding, appreciation, tolerance, and protection of people and nature. The value of self-direction considers safety, harmony, and stability of society, of relationships, and of self as imperative. Excitement, novelty, and accepting challenges in life are emphasized by stimulation, which is another important value in the list. Hedonism emphasizes on pleasure and gratification for oneself. Personal success and competence is referred to as achievement, which was considered by Schwartz as another core value. The last two values in the list consist of power, which is understood in terms of social status and prestige, control, or dominance in the society, and security, which stresses the safety and stability of society, relationships, and the self. Schwartz (1992) further added a dimension of congruence and incongruence to this set of values. Certain values like (1) power and achievement, (2) achievement and hedonism, (3) hedonism and stimulation, (4) stimulation and self-direction, (5) self-direction and universalism, (6) universalism and benevolence, (7) tradition and conformity, (8) conformity and security, and (9) security and power were delineated as congruent values as they go together and at times even support each other. The incongruent values are (1) self-direction and stimulation versus conformity, tradition, and security; (2) universalism and benevolence versus achievement and power; (3) hedonism versus conformity and tradition. For Schwartz, the individual actions taken in the process of exploration of these values have both psychological and social consequences for individuals.

Several popular but cognate concepts in psychology need to be examined to further clarify the meaning of the concept of values. Attitude is one such concept that is generally defined as a relatively enduring organization of beliefs, feelings, and behavioral tendencies toward objects, groups, events, and symbols. Attitudes enable us to evaluate people, behaviors, events, issues, or any object. According to Schwartz (2012) "values underlie our attitudes; they are the basis for our evaluations. We evaluate people, behaviors, events, etc. positively if they promote or protect attainment of the goals we value. We evaluate them negatively if they hinder or threaten attainment of these valued goals. Beliefs are ideas about how true it is that things are related in particular ways. Examples of beliefs are 'war never solves problems,' 'Africa is larger than Europe,' and 'psychologists are wise.' Beliefs vary in how certain we are about their truth and their validity. General beliefs that people hold about how the world functions are called social axioms (Leung and Bond, 2004). Unlike values, beliefs refer to the subjective probability that a relationship it true, not to the importance of goals as a guiding principles in life" (p. 16).

Having clarified the concept of values, another issue related to values is their measurement. Measurement of values has been a challenge for researchers and several attempts have been made in the past, starting from Spranger's (1928) '*Types of men*' wherein he identified major human values, namely, theoretical, economic, aesthetic, social, political, and religious. Allport et al. (1931) operationalized Spranger's list of values. Several other attempts have been made to measure values but the most popular ones are Rokeach's (1973) Values Survey, and Schwartz's (1992) Value Scale. Schwartz is credited with developing a survey that can be used to identify a set of values that can differentiate an individual from another or that can also differentiate a society from another. He asked his respondents to rate the extent to which each of 56 values was a guiding principle in his or her life. The values were presented as a single word or a phrase with a brief clarification of what the author meant by each of them.

The Sociocultural Correlates of Values

The social psychological research has delved into variables in the sociocultural context that are potentially important in shaping value preferences. Age, life stage, educational level, occupational status, and gender, are some of the variables that have received considerable research focus. Some of this research is discussed here to illustrate the trends. Studies have reported that there are differences in values due to age.

Schwartz (1992) reported that young people attributed greater importance to openness (stimulation and hedonism) and less importance to conservation (tradition, conformity, and security). It has been found that younger people expressed more individualistic and less collectivistic values. Also, people living in urban settings in comparison to those living in rural settings typically show more individualistic and less collectivistic orientation and view 'openness to change' and 'universalism' as more important. In the case of gender, it has been argued that men would emphasize agentic-instrumental values like power, whereas women emphasize expressive-communal values like benevolence. In general, women have been found to be more collective, interdependent, relational, and allocentric in comparison to men. Differences in values due to different occupational statuses are also seen and have been demonstrated by studies comparing managers and workers. Most of these studies have reported that managers and workers have distinct value systems. Some studies conducted in India have reported that the managers perceive greater differences between themselves and workers. Education has been another variable that has received research attention. Kohn and Schooler (1983) argued that education has the tendency to promote flexibility, openness, and self-direction values. Results related to education and values are somewhat mixed. Few studies have shown that educational experiences promote intellectual openness, flexibility, and breadth of perspective necessary for self-direction. Education has been found to be related to self-direction and stimulation rather than conformity and tradition. However, some other studies have not shown much differences due to educational status. Despite such mixed results, one would like to acknowledge that schools have been known to encourage the dominant social ideals in most cultural contexts, at times even religious ideals and values. Schools are considered the 'institutions of socialization' and certainly have an impact on values. Therefore, a lot of importance is given to value education in schools in almost all cultures.

Development of Values across Cultures: Differences and Similarities in Value Preferences

Since values are tied to sociocultural context and are shaped by the context, it would not be unfair to say that they are socially and culturally constructed and maintained. The study of values and analysis of cultures is closely linked. Culture provides principles to people to sample information in their environment (Triandis and Suh, 2002). Researchers have looked at culture in different ways. Some have explored culture in terms of structures, some have defined its functions, while others have examined the processes. Whatever be the lens through which culture has been seen, all these conceptions have helped in understanding the impact of culture on human interaction. When trying to make sense of others' behavior, their own behavior, or even taking a stand on issues, people from Eastern cultures may focus more on family reactions (like what will my family/community say, or what will be the repercussions of my action on my family, etc.), which can be referred to as elements of the collective self. These cultures emphasize interdependence among individuals more while people in Western cultures may focus more on their personal self as these societies are more individual oriented. Also, Western culture seems to hold clear-cut dichotomies between the self and the other, man and nature. Thus, we see that there are cultural variations in the way people relate and interact with each other and these are shaped by values. Relationality seems to be very important in Asian cultures but not in Western culture. Cultural values reflect what a social collective would think is right and desirable. How relationality works in Asian cultures and under what conditions it does not work needs exploration.

Another question of interest to theorists working on culture is that can a nation be used as a synonym for culture or also do nations have a culture. Rohner (1984) attempted to answer this question. He proposed that the essence of 'culture' lies in the shared way in which individuals interpret what goes on around them. For example, women in rural parts of northern India generally cover their heads before the elderly. If many people think that covering the head by a woman is a sign of respect for elderly, one can say that there is a shared culture of showing respect to elderly by covering the heads. Rohner further distinguishes between a culture and a social system. In case of a social system, an example of which could be a family or even a nation, the emphasis is on the patterns of shared behavior and interactions. A very important question raised by Smith et al., which ought to be addressed by theorists working in this field is – "are culturally shared meanings achieved because there is a system of shared values and beliefs, or is there some set of normative reference points in a culture that leads individuals to interpret what goes on around them in similar ways even though their values and beliefs differ from those around them?" (p. 6). How does an Indian (originally from an Indian village with his baggage of traditional values) now an immigrant in the United States derive shared meanings? Or to rephrase it differently, how Indian traditional values and individualistic American cultural orientation together influence values of Indian immigrants? People may, however, sometimes behave inconsistently with their personal values. One may value certain things in his or her personal life but behave completely different when interacting with others in a specific cultural or even in a specific situational context. Rohan (2000) was of the view that one's attitudes and behaviors interact with the personal and cultural value systems. Recently, Boer and Fischer (2013) provided an interesting framework that specifies the motivational basis of social attitudes concerning fairness, care, purity, in-group, and authority.

Are there cultural differences within nations? And how theorists handle this is a pertinent question as most cross-cultural psychology research makes national comparisons. It is also true that some nations have more cultural diversity than others. As an Indian, I am certainly aware that within my country from north to south and east to west, there are vast cultural differences. Hence my skepticism for any such research claiming to compare Indian values and people's behavior with other nations, as there are variations in values among Indians from different regions, social groups, etc. As Smith et al. (2013) articulated "nations, as well as the ethnic, regional and socio-economic groups that may be present within them, can be defined as cultures to the extent that there is evidence that their

members interpret the events around them in relatively similar ways. We need next to look at the ways in which cross-cultural psychologists can test for such homogeneity of outlook within any particular social system" (p. 7).

Major Cross-Cultural Value Surveys

A seminal work examining values across cultures was undertaken by Hofstede (1980, 2001). He was associated with a worldwide survey of IBM employees in more than 70 nations, which was later published in his book *Culture's Consequences*. He describes the effects of a society's culture on the values of its members, and further how these values relate to the behavior. In his initial work, he identified four major dimensions along which cultural values could be analyzed: individualism–collectivism; uncertainty avoidance; power distance, and masculinity–femininity. Individualism–collectivism refers to the importance attributed to the self or the collective/society. In collectivistic societies, individuals are more embedded in their groups. Individualism gives more importance to personal distinct choices and individual's achievements. Power distance is another dimension identified by Hofstede, which refers to the degree to which people accept power of institutions and practices or whether people endorse more egalitarian values. The dimension of masculinity–femininity considers the importance given to attributes that are considered masculine. In contrast, femininity refers to nurturance, care, and benevolence. Uncertainty avoidance refers to the stability in things happening. Later, Hofstede added a fifth dimension, i.e., long-term orientation. This refers to how much a society values long-standing – as opposed to short-term – traditions and values. Cultures differ in their emphasis on short-term values in social relations to long-term values. In the 2010 edition of *Cultures and Organizations: Software of the Mind*, Hofstede added a sixth dimension also, which he called indulgence versus self-restraint. This refers to the extent to which members in a society attempt to control their desires and impulses. Indulgent societies provide for gratification of basic human desires while restrained societies are more restrictive and try to regulate basic human desires.

Of all the dimensions proposed by Hofstede, over the years the dimension of individualism and collectivism generated a lot of enthusiasm among researchers and resulted in a lot of research. One of the major criticisms of Hofstede's work is that within a nation, differences based on regions, social class, ethnicity, gender, etc. were not addressed. More recently, Minkov and Hofstede (2012) tested whether there was much variation in values between regions within nations, compared to the variation between nations. Their results show that regional differences in values within nations tend to be somewhat less than differences between nations. This result may be attributed to several methodological issues related to the study but this certainly needs to be further explored. It has also been reported that people's values, specifically people's collectivistic values, are target specific. This brings in other factors in the understanding of people's values. The implication of this assumption is that the target of interaction is the determining factor. When the target is close to the person, more collectivistic behaviors will be shown than when the target is not so close. In other words, for our family, friends, and even at times in-group members, we are guided by collectivistic values; however, for others we may be guided by individualistic values.

Schwartz (1992) (whose work has already been discussed) provided an alternative approach to the study of values across cultures. His view was that any comparison of national level values would not be meaningful unless one has understood the structure of values that are held by individuals. When we know the structure of values within the minds of individuals in various cultures, we may also know whether the psychological meaning given to a value is shared or not. He identified seven dimensions that differentiate societies. These include embeddedness, intellectual autonomy, affective autonomy, hierarchy, egalitarianism, mastery, and harmony.

Another significant work in this area that yields nation-level means for values is based on the World Values Survey (Inglehart, 1997). Beginning in Europe in the 1980s today this survey covers more than 80 nations. Inglehart has analyzed the changes in these values over time and has identified well-being versus on survival and preference for secular-rational versus traditional authority as major dimensions affecting people's political preferences. The famous globe project (House et al., 2004) identified the following societal cultural values: uncertainty avoidance, institutional collectivism, in-group collectivism, power distance, gender egalitarianism, assertiveness, and performance orientation from a survey on business managers in 62 nations. In this survey, managers rated their society 'as it is' and 'as it should be' on nine dimensions of national culture. These societal cultural values were found to be related to various organizational variables.

To sum up, these major value surveys show that cultures develop values that are shared by their members. How do cultural values develop? Suppose in a particular culture, people give preference to the needy in resource allocation. This preference may be due to the fact that resources are scarce in that culture and therefore protecting the needy is an important concern and this develops into a value. Values are generally received through cultural means, especially transmission from parents, teachers, and other socializing agents (in today's time media has a major role) to children. Fischer (2012) suggested that at the national level, one may expect the success or failure of different types of social systems to predict which values will go together. The Euro-American culture emphasizes the individualistic, personal, or independent orientation. In contrast, in many African and Asian cultures, the major focus is on collective, social, and interdependent orientation (Smith and Schwartz, 1997). Religion also specifies the desirable values, end states, or behavior. Although the essence of all religions is the same and the values emphasized by major religions are complementary to each other, differences may be seen. Suar (2013) posits "that values propounded by major religions can be broadly put under three categories: (1) self-purifying values (e.g., truthfulness, honesty, compassion, fasting, right speech, avoidance of murder, adultery and stealing), (2) others' welfare- and service-promoting values (e.g., nonviolence, forgiveness, charity, right understanding, refraining from

bearing false witness against neighbors), and (3) faith-inculcating values in a supreme power/God (daily prayer, pilgrimage, right concentration and submission to God). Values such as avoidance of murder, adultery and stealing are self-purifying and in the context of others are welfare-promoting. Essentially, faith in supreme power, self-control, and others' welfare are the mainsprings of values in all religions" (p. 971). Thus, although religion has impact on personal values and values across cultures, it is also true that dominant religious groups in most cultures define value priorities for most in a particular cultural context but less dominant groups have their own priorities. In order to have a harmonious social order, it is important that value priorities of most groups be recognized.

In different nations, which we may consider as different cultural contexts (although one has to understand that there will be several subcultures in a nation), diverse values get precedence over others. In this section, an examination of values in different national (cultural) contexts will be discussed to examine values across cultures. Needless to say, a lot of this work has been done in work contexts but has ramifications for other contexts also. This is just a sample of some studies that have been done and is by no means exhaustive. Studies from non-Western contexts have been sampled and discussed as compared to Western contexts they are less represented. The concept of 'bao' is a traditional Chinese concept but is important for understanding social relations as well as values (Liu, 1999). Six subconcepts of bao have been identified, some having a positive direction (repayment of gratitude, repayment of favor, and good retribution) and some having a negative direction (retaliation, revenge, and evil retribution). The social significance of bao is that it shapes the way people maintain balance in social relationships. Confucious values emphasize repayment of a grievance with rectitude and returning virtue with virtue. Such reciprocity, which is different from the notion of reciprocity in Western culture, governs social relationships in China. In India, several researchers have focused on Indian work values. These studies show that Indians continue to be hierarchical and are relationship oriented, where family plays a major role in people's lives. Relationship orientation, harmony and tolerance, norms and duty, and preference for hierarchy are quite important in the contemporary Indian society. In general, the Indian society has been considered as collectivist. The main features of collectivists include interdependent self-construal, compatibility with in-group goals, roles of norms, duties, and obligations in determining social behaviors, and priority to relationships. People try to remain embedded in the group. However, as D. Sinha and Tripathi have noted, India is a case of coexistence of opposites where both individualistic and collectivistic tendencies are present.

The relationship between values and social identity has been a concern for many researchers in different parts of the world, particularly in the region considered to be high on individualism. For instance, some studies conducted in Australia have observed that there was close association between endorsing value types of achievement, hedonism, security, and stimulation as associated with assuming Australian identity. In a study conducted in South Africa, it was seen that in case of Afrikaners international harmony and equality were best predicted by being progressive and religious. National strength and order were predicted by being an African speaker for white Afrikaners, while among black South Africans the best predictor was being South African, a global citizen, and progressive. These studies have started looking at the finer variations in values within groups in the same cultural context. This trend needs to continue to facilitate a more in-depth understanding of the development of values across cultures. A multilevel analysis of values examining how individual values interact with values embraced and shared by groups in a specific culture and how such interaction helps to explain values across different cultures needs to be done. Issues regarding stability and change of values in the wake of fast paced changes in most cultures seem to be an important future direction of research.

See also: Conformity: Sociological Aspects; Egalitarianism; Individualism versus Collectivism: Philosophical Aspects; Socialization in Infancy and Childhood; Tradition, Anthropology of; Values, Anthropology of; Values, Psychology of; Values, Social Psychology of; Values, Sociology of.

Bibliography

Allport, G.W., Vernon, P.E., Lindzey, G., 1931. A Study of Values. Houghton Miffin, Boston.

Boer, D., Fischer, R., 2013. How and when do personal values guide our attitudes and sociality? Explaining cross-cultural variability in attitude-value linkages. Psychological Bulletin 139, 1113–1147.

Fischer, R., 2012. Value isomorphism in the European Social Survey: exploration of meaning shifts in values across levels. Journal of Cross-cultural Psychology 43, 883–898.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work Related Values. Sage, Beverly Hills, CA.

Hofstede, G., 2001. Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Cultures, second ed. Sage, Thousand Oaks, CA.

Hofstede, G., Hofstede, G.J., Minkov, M., 2010. Cultures and Organizations: Software of the Mind, third ed. McGraw-Hill, USA.

House, R.J., Henges, P.J., Javidan, M., Dorfman, P.W., Gupta, V., 2004. Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies. Sage, Thousand Oaks, CA.

Inglehart, R., 1997. Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Societies. Princeton University Press, Princeton, NJ.

Kluckhohn, C.K.M., 1951. Values and value orientations in the theory of action. In: Parsons, T., Sils, E. (Eds.), Toward a General Theory of Action. Harvard University Press, Cambridge, MA, pp. 388–433.

Kohn, M., Schooler, C., 1983. Work and Personality. Ablex, Norwood, NJ.

Leung, K., Bond, M.H., 2004. Social Axioms: A Model for Social Beliefs in Multicultural Perspective. In: Advances in Experimental Social Psychology, vol. 36. Academic Press, San Diego, CA, USA, pp. 119–197.

Liu, C., 1999. The concept of bao and its significance in organizational research. In: Kao, H.S.R., Sinha, D., Wilpert, B. (Eds.), Management and Cultural Values. Sage, New Delhi, pp. 152–168.

Minkov, M., Hofstede, G., 2012. Hofstede's fifth dimension: new evidence from the world values survey. Journal of Cross-cultural Psychology 43 (1), 3–14.

Rohan, M., 2000. A rose by any name? The value construct. Personality and Social Psychology Review 4, 255–277.

Rohner, R.P., 1984. Toward a conception of culture for cross-cultural psychology. Journal of Cross-cultural Psychology 15, 111–138.

Rokeach, M., 1973. The Nature of Human Values. Free Press, New York.

Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology 25, 1–65.

Schwartz, S.H., 2012. An overview of the Schwartz theory of basic values. Online Readings in Psychology and Culture 2 (1).

Schwartz, S.H., Sagiv, L., 1995. Identifying culture-specifics in the content and structure of values. Journal of Cross-cultural Psychology 26, 92–116.

Smith, P.B., Schwartz, S.H., 1997. Values. In: Berry, J.W., Segall, M.H., Kagitcibasi, C. (Eds.), Handbook of Cross-cultural Psychology, second ed., vol. 3. Allyn & Bacon, Boston, MA, pp. 77–118.

Smith, P.B., Fischer, R., Vignoles, V., Bond, M.H., 2013. Understanding Social Psychology across Cultures. Engaging with Others in a Changing World. SAGE, London.

Spranger, E., 1928. Types of Men (P.J.W. Pigors, Trans.), fifth ed. Niemeyer, Halle.

Suar, D., 2013. Valuing the values. In: Misra, G. (Ed.), Psychology and Psychoanalysis. History of Science, Philosophy and Culture in Indian Civilization. Centre for Studies in Civilizations, New Delhi, pp. 957–988.

Triandis, H.C., Suh, E.M., 2002. Cultural influences on personality. Annual Review of Psychology 53, 133–160.

Values, Anthropology of

James L Peacock and Hannah M McFadden, University of North Carolina, Chapel Hill, NC, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Peacock, volume 24, pp. 16145–16148, © 2001, Elsevier Ltd.

Abstract

From the anthropological view, values are considered important sociocultural indicators, deriving from and diversified by human thought and interaction. Values are an inextricable part of the anthropologist's work and are commonly identified as points of interest in contemporary topics spanning a variety of disciplines. As part and parcel of the human experience, the worth of value studies continues to be emphasized.

Contemporary anthropology views values through a holistic lens, acknowledging both their diversity and their subjectivity. Values are seen as both interconnected and varied within and between individuals and contexts of culture, space, and time. Anthropologically, values are seen as human (i.e., psychosocial–cultural) constructions, yet, once constructed, values define the good and the bad, and hence guide and even control behavior. As a discipline centered on the investigation of humans and their constructions, anthropology is appropriately situated to analyze values, their implications, and their importance (Peacock, 1986).

The importance of values in anthropology was recognized by the mid-twentieth century by the designation of two chapters on this topic in Alfred Kroeber's *Anthropology Today: An Encyclopedic Inventory* (1952). Both chapters were written by philosophers, F.S.C. Northrop and David Bidney, rather than practicing anthropologists, which perhaps suggests a sense that the topic should be grounded in a humanistic tradition. Both authors trace philosophical underpinnings while also discussing issues still current such as the debate between cultural relativism and human rights. While this article is much briefer than the two noted, they serve as a baseline against which we can note new developments.

Clyde Kluckhohn, a foremost influential author of one anthropological theory of value, notes that values, in the sense of 'conceptions of the desirable,' tend to be justified morally by aesthetic judgment, reasoning, or some combination of these (Kluckhohn, 1951). Religious practice is one sphere where we find examples of Kluckhohn's morally justified conceptions of the desirable. Religious doctrine and spiritually inspired stories set forth basic principles for living – the belief that one should treat others as one would like to be treated, or the 'golden rule,' exemplifies an *in vivo* ethic. Institutions and individuals alike interpret and disseminate such an ethic, but it is subjects who morally justify, consume, adapt, and integrate this rule for living as a part of their personhood, and it thereby becomes their value.

Values enter anthropology both as a subject for study and as an aspect of the anthropologist's own experience. As a subject for study, values are critical, contemporary topics for investigation in a variety of anthropological subfields. Ecological anthropology, for instance, considers how values affect the way humans interact with and conceptualize their environment and perhaps destroy it. Elucidated by the Society for Applied Anthropology (SfAA), business anthropology considers how ethics are employed in the workplace and studies consumer behavior as it is affected by shared values. In an assessment of 'sense of place' values in Alaska, Brown et al. (2002) further reflect the nuances of the term, indicating that people have and share 'values' and also express 'value' for things – the complex study of valuing economic value is one focus of economic anthropology.

Treating values as an object of study, anthropologists have taken either a cultural or a sociological approach, either emphasizing the patterning of values in themselves and the power of values as cultural conceptions that shape experience, or, *mutatis mutandis*, emphasizing the interrelationship between values and the sociological context.

Examples of the first approach come largely from cultural anthropology in the American tradition (which has philosophical roots in German idealism). Early examples are based on configurationalism: that is, holistic analysis of values that define broad themes of a particular culture. *Patterns of Culture* is the classic example where Ruth Benedict (1934) distinguishes Apollonian and Dionysian values among Plains compared with Pueblo Indians. Later, during and following World War II, Benedict, Mead, and others applied the same approach to whole nations, depicting the dominant values and psychological patterns of Japan, Germany, and Russia; thus was created the national-character school in anthropology (Benedict, 1946; Mead, 1943). Kluckhohn, Vogt, and Albert (1966), and Kluckhohn and Strodbeck (Kluckhohn, 1951; Kluckhohn and Strodbeck, 1961), later refined the somewhat undisciplined holism of these studies by explicitly centering on values. They defined dimensions for comparison and made systematic comparisons of particular ethnic or regional groups within the United States. Clifford Geertz adopted a Parsonian definition of culture (values, symbols, and ideas) that further refined the older configurationalistic notion, and then pursued a Weberian project in his *Religion of Java* (1960), where he distinguished the major divisions within a complex culture according to distinctive values. *Habits of the Heart* (Bellah et al., 1985) applied this approach to the United States, producing an influential critique of American individualism and calling for a return to earlier values that recognized societal responsibility. Early configurationalists along with German influences remained important in these later, Weberian-inspired works; the recurrent theme is that

values, as a manifestation of culture, have meaning and significance irreducible to any societal context.

The second, sociological approach places more importance on social relations, which are viewed as a crucial source and ground of values. Examples of the second approach come classically from the British social anthropological tradition, which had philosophical roots in France, especially in the works of Émile Durkheim. Building from the foundation that thought derives from the relations, experiences, and behaviors we incur as actors in our local social scenes, Durkheim, in *Elementary Forms of Religious Life* (1915), argues that values derive from thought, generally. An excellent application of this thesis is Campbell's *Honor, Family, and Patronage* (1964). Campbell demonstrates cogently how bilateral family patterns explained ideas of honor expressed by the Sarakatsani of Greece. Honor is an ultimate value, for which the Sarakatsani man will sacrifice his life. Yet in this society, honor is rooted in kinship; one's honor depends on one's kin. The values derive from and sustain the family. Campbell's analysis is one of the best in showing the intricate functional integration of value and social organization. Building on this particular analysis, Pitt-Rivers (1983), Peristiany (1992), and others also showed how the concept of honor can be seen as a Middle Eastern and Southern European value, somewhat transcending the uniqueness of social organization but also grounded in a regional cultural and social pattern they termed 'Mediterranean.' Unni Wikan's *In Honor of Fadime* (2008) traces such values to the horrifying case of a young Kurdish woman living in Sweden who was murdered by her father, who felt justified by the dishonor she brought upon her family in having relations with a man outside of her community. In this case, the power of values is explicit: honor is justified by a society, despite undermining basic human rights.

An effort to reduce the first approach to the second was Douglas' *Cultural Bias* (1978). Douglas distinguishes two dimensions of social life, group and grid; the grid is the networks that bind and classify members internally, and the group is the identity of the group contrasted to those outside its boundaries. Utilizing her social variables 'grid' and 'group,' Douglas argues that many cultural values – conceptions of time, of health and disease, and of nature – depend on which particular social configuration is dominant. Thus, by knowing the social pattern, one can predict which cultural value will dominate; the societal configuration explains cultural values.

Since the peak of the cultural and social anthropological structural-functionalist approaches to value, anthropology has developed in a myriad of directions. Comaroff and Comaroff's *Ethnicity, Inc.* (2009) illustrates indigenous groups transmuting their local culture to mimic that of corporations. In the struggle to maintain identity, culture and the values therein are sold as commodities, which are consumed by tourists and companies. Drawing from cases across the globe, the Comaroffs illustrate the globalized, capitalistic forces blurring the distinction between values and value. Further exemplifying the sociological emphasis but with a Marxist twist, one anthropological approach treats political economy as the driving force of culture. This approach tends to diminish the role of values, which are treated as false consciousness, as ideological constructions motivated by class interests, or as a drive to power and profit. Studies of colonialism and postcolonialism illustrate this approach, for example Comaroff's *Body of Power, Spirit of Resistance* (1985), where religious movements and their accompanying values are contextualized within a colonial and postcolonial framework in South Africa.

While anthropologists who emphasize political economy moved values from center stage, the topic has migrated to other disciplines, especially those concerned with globalism and international relations. Huntington's *The Clash of Civilizations* (1996), for example, reifies certain value configurations, notably those of Islam and the West, then foresees the clash between them, replacing the ideological conflict of the Cold War. Scott's *Seeing like a State* (1998) also focuses on the power of values, notably rationalization, when coupled with the state, to shape human life and thought.

Turning to the place of values in anthropology and for the anthropologist, the basic debate is between engagement and objectivity. While Weber's 'value-free social science' concept converges with positivism to express at least an ideal of objectivity, his *Verstehen* converges with a Gadamerian post-modernism to acknowledge the role of values (what Gadamer terms 'prejudice') as projected by the analyst into the process of analysis. These Weberian nuances can still inform the anthropological debate (see, e.g., the debate between D'Andrade and Scheper-Hughes, 1995). Like Weber, anthropologists agree in acknowledging the role of values in their research. Methodologically, this translates as reflexive analysis, the turning of the anthropologist's focus onto him or herself and the values and biases that may influence research or fieldwork. Still, there is debate regarding whether the analyst should attempt to control the influence of values or embrace their influence (i.e., become engaged). Engagement means that the analyst should express rather than suppress values, whether those of one's informants, oneself, or others (e.g., of some ideology or vision of the good). Objectivity means that the analyst should control or suppress the expression of values (other than those of scientific or accurate description and analysis).

A major concern of anthropologists is with the values and ethics governing fieldwork. This concern is illustrated by both history and recent issues treated by the American Anthropological Association (AAA). The overriding principle of the Ethics Code of AAA is "Do no harm to those with whom one does field research." This principle abides in multiple revisions of the code, including the current code (revised in 1998) and pending codes. Means of preventing harm include informed consent (explaining one's research to those involved and insuring that they consent to involvement) and protection against risk, for example by concealing the identity of those involved.

Two examples of controversies expressing this principle are the Darkness at El Dorado affair and the Human Terrain Program (HTP), both of which provoked debate by numerous anthropologists and assessments by committees appointed by AAA. *Darkness in El Dorado* by Patrick Tierney is an expose by a journalist of research by a geneticist, an anthropologist, and others among the Yanomamo of Brazil and Venezuela. The researchers were accused of practices such as taking blood samples inappropriately and manipulating the

Yanomamo inappropriately. Committees appointed by AAA extensively researched these allegations, rejecting some and accepting others. Some claim the controversy threatened the reputations of both individuals and the discipline.

The HTP was instituted by the US military in Afghanistan. Several hundred scholars were recruited to work with the military in combat areas. Such HTP recruits did research on local situations and sometimes established community programs, such as health clinics. A number of anthropologists criticized this effort on the ground that it violated the principle of 'do no harm' in that local people involved in the research or programs could be placed in danger by being brought to the attention of the Taliban. Further, HTP could endanger or jeopardize anthropological fieldwork by allying it with military and intelligence work, hence rebutting claims of objectivity and neutrality. In response to such concerns, an AAA committee has carried out extensive analysis of military practices in relation to anthropology and anthropologists, supporting, to some degree, critiques of the ethics.

Such examples illustrate a general trend for anthropologists and anthropology to link scholarly work and societal issues. One term that captures this trend is 'public anthropology' (Peacock, 1997; Borofsky, 2011). Ronald Niezen's *A World beyond Difference* (Neizen, 2004) offers a searching analysis of a spectrum of values related to practice, with a Human Rights Declaration at one extreme, postulating universal values, and postmodernist refusal to affirm any moral value in terms other than relativistic ones at the other.

In the early twenty-first century, efforts toward defining the place of values in the discipline of anthropology display three major streams. The first affirms positivism and objectivity. The second affirms subjectivity, an interpretive, postmodernist stance leading toward reflexivity – explicit acknowledgment of one's values as part of one's analysis. The third also affirms the place of one's values but leads not so much inward as outward, minimally toward ethical behavior in relation to those among whom one does research, and maximally toward engagement and practice – a mission of reform and change of the world.

The debate about the proper place of values in the experience of the anthropologist is important within the discipline. Arguably of greater impact, however, is the first topic of this article: the anthropological study of value as expressed by humankind. Values will continue to be studied as phenomena deriving from the intricate relations between all worldly things. Anthropology continues to illuminate the importance and worth of these studies, as values are critical elements in the construction and understanding of life by peoples past and present.

See also: Anthropological Research, Ethics of; Cultural Psychology; Culture: Contemporary Views; Globalization, Anthropology of; Honor and Shame; National Character; Organizations and Culture; Religion, Anthropology of; Values, Psychology of; Values, Sociology of.

Bibliography

Bellah, N., Madsen, R., Sullivan, W., Swidler, A., 1985. Habits of the Heart: Individualism and Commitment in American Life. University of California Press, Berkeley, CA.

Benedict, Ruth, 1934. Patterns of Culture. Houghton Mifflin, Boston.

Benedict, Ruth, 1946. The Chrysanthemum and the Sword: Patterns of Japanese Culture. Houghton Mifflin, Boston.

Bidney, D., 1952. The concept of value in modern anthropology. In: Kroeber, A.L. (Ed.), Anthropology Today. University of Chicago Press, Chicago, pp. 682–699.

Borofsky, Rob, 2011. Center for a Public Anthropology. http://www.publicanthropology.org/public-anthropology/.

Brown, G.G., Reed, P., Harris, C.C., 2002. Testing a place-based theory for environmental evaluation: an Alaska case study. Applied Geography 22 (1), 49–76.

Campbell, J.K., 1964. Honour, Family and Patronage: A Study of Institutions and Moral Values in a Green Mountain Community. Clarendon Press, Oxford.

Comaroff, Jean, 1985. Body of Power, Spirit of Resistance: The Culture and History of South African People. Univeristy of Chicago Press, Chicago.

Comaroff, J.L., Comaroff, Jean, 2009. Ethnicity, Inc. University of Chicago Press, Chicago.

D'Andrade, R., Scheper-Hughes, N., 1995. Objectivity and militancy: a debate. Current Anthropology 36, 399–420.

Douglas, Mary, 1978. Cultural Bias. Royal Anthropological Institute, London.

Durkheim, Emile, 1965. Elementary Forms of Religious Life. Free Press, New York.

Geertz, Clifford, 1960. The Religion of Java. Free Press, Glencoe, IL.

Huntington, S., 1996. The Clash of Civilizations and the Remaking of World Order. Simon and Schuster, New York.

Kluckhohn, Clyde, 1951. Values and value-orientation in the theory of actions: an exploration in definition and classification. In: Parsons, T. (Ed.), Toward a General Theory of Action. Harvard University Press, Cambridge, MA, pp. 338–433.

Kluckhohn, F., Strodbeck, F., 1961. Variations in Value Orientations. Row & Peterson, Evanston, IL.

Kroeber, A.L. (Ed.), 1952. Anthropology Today. University of Chicago Press, Chicago.

Mead, Margaret, 1943. And Keep Your Powder Dry: An Anthropologist Looks at America. William Morrow, New York.

Neizen, Ronald, 2004. A World beyond Difference: Cultural Identity in the Age of Globalization. Blackwell Publishing, MA.

Northrop, F.S.C., 1952. Cultural values. In: Kroeber, A.L. (Ed.), Anthropology Today. University of Chicago Press, Chicago, pp. 668–681.

Peacock, James, 1986. The Anthropological Lens: Harsh Light, Soft Focus. Cambridge University Press, Cambridge.

Peacock, James, 1997. The future of anthropology. American Anthropologist 99, 9–29.

Scott, James C. (Ed.), 1998. Seeing like a State – How Certain Schemes to Improve the Human Condition Have Failed. The Yale Agrarian Studies Series. Vail-Ballou Press, Binghamton, NY.

Tierney, Patrick, 2002. Darkness in El Dorado: How Scientists and Journalists Devastated the Amazon. W. W. Norton & Company Inc., New York, NY.

Wikan, Unni, 2003. In Honor of Fadime: Murder & Shame (Anna Paterson, Trans.). University of Chicago Press, Chicago.

Values, Psychology of

Daphna Oyserman, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Values are internalized cognitive structures that guide choices by evoking a sense of basic principles of right and wrong, a sense of priorities, and a willingness to make meaning and see patterns. Like other cognitive constructs, values can be studied at the individual level or at the group level. That is societies, cultures, and other social groups have value-based norms, priorities, and guidelines, which describe what people ought to do if they are to do the 'right,' 'moral,' 'valued' thing. The study of values currently focuses more explicitly on the circumstances in which values predict action.

Values: Psychological Perspectives

Values are internalized cognitive structures that guide choices by evoking a sense of basic principles of right and wrong (e.g., moral values), a sense of priorities (e.g., personal achievement vs group good), and create a willingness to make meaning and see patterns (e.g., trust vs distrust). Like other cognitive constructs, values can be studied at the individual level. However, while it is unusual to discuss attitudes as a group-level phenomenon, values are amenable to analysis at the societal or group level. That is societies, cultures, and other social groups have value-based norms, priorities, and guidelines, which describe what people ought to do if they are to do the 'right,' 'moral,' 'valued' thing. A society, political party, or region can be described as having conservative or liberal values, traditional or progressive values, religious or secular values, and so on.

Values are supposed to influence behavior and people are suspicious of others if they suspect that they hold different values in part because of the presumed link between values and behavior. Knowing a person's value system provides a sense that one also knows what he or she will do in a particular situation or across situations. In spite of the strong belief that values predict behavior, the effect of values on behavior is subject to situational constraints and affordances. That is, if mobilized or made salient, individual values are linked with behavior and choices but not necessarily otherwise.

However, if values do not come to mind, they are unlikely to influence judgment and behavior. Cultures may be said to provide concrete and social embodiments of values. Thus, if cultural values are mobilized or made salient, they also predict behavior and choices of groups and to some extent, of individuals within these groups. Whether cultural cues (e.g., the national flag) serve to cue relevant values may depend on whether their appearance in context seems unremarkable or an obvious attempt to cue a particular value response. Some kinds of values, those that form the basis of moral judgments, appear to have a more visceral effect on judgment, influencing how people feel about a topic, situation, or choice separately from their cognitive evaluation of the topic, situation, or choice. Thus values can cue affective (feeling) as well as cognitive (thinking) responses. The effect of values on judgments and behavior can be seen whether the value is explicitly part of the decision-making or not. That is, a value can be associated with a choice without the association being made explicit.

Overview

This article addresses the following questions. First, what are values, how do psychologists study values, and how are values distinguishable from related concepts – motives, goals, and attitudes? Second, are values best considered at the individual level as located within individuals, or at the context level, as if located in social structures, or are both levels of analyses needed to understand values? Third, what do values imply for behavior? With regard to the first question, values are often studied by asking people to rate how important certain statements are to them or to people in their group. These methods rely on explicit judgments which may be subject to the same variety of constraints as all self-report measures (including, for example, whether people can make explicit judgments about what they value, and social desirability – the desire to say things that the questioner will approve of). Reliance on self-report makes values difficult to study and persistent questions arise as to whether they are 'real,' that is, whether they actually can be shown to have causal influence on behavior.

In spite of this difficulty, values have intuitive appeal as an explanatory construct. Much of everyday life is cast in terms of values, value judgments, or value preferences. Think of ethics, law, religion, politics, education, art, lifestyle, child rearing, and more. Choices in these domains are often explained in terms of values and the value trade-offs one is willing or unwilling to make. Abstract value judgments are embodied in seeming gut reactions that something is right, moral, or natural verses wrong, immoral, or unnatural. Another way to 'see' values in action is to contrast cultures or subcultures in what seems right, natural, or moral.

Empirically, an emerging literature uses priming techniques to show that values matter when brought to mind, not otherwise and that what a value means depends on the context in which it is considered. Values related to loyalty (e.g., patriotism) and to acceptance of authority (e.g., traditionalism), for example, are associated with more trust in government. This effect occurs only in contexts in which government functions brought to mind are relevant to these (e.g., trusting the government to make the right choices for defense spending)

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24030-0

not otherwise (e.g., trusting the government to make the right choices for health care). A similar argument has been made for the role of values in acceptance of social policies, polices that appear to come from one's own political party are preferred in part because in-group loyalty is valued.

Another way in which values are studied is by comparison. Comparisons in choices, behavior, social structures have all been used to infer differences in values. One of the great contributions of cultural and cross-cultural research is the way that it brings Western cultural values into sharp relief by making these values stand out in comparison to alternative frames of core values. These and other differences are also found in comparisons of political parties, social classes, religions, and even professions. Consider American cultural values, Americans are said to value life, liberty, and the pursuit of happiness. But what does 'value' mean? Implicitly or explicitly we evaluate or assign value to everything – regarding things as good or bad, a truth or falsity, a virtue or a vice. How do we know? One important means is through values. Values can be thought of as priorities, internal compasses, or springboards for action – moral imperatives. In this way, values or mores are implicit or explicit guides for action; they are general scripts framing what is sought after and what is to be avoided. Values can be thought of as scaffolded on universal human needs for security, autonomy, and connection, which would imply that since all cultures face similar problems they are likely to overlap in values (though not necessarily in their ranking of the importance of each value). To sustain a group, for example, people have to believe that their place within the group is fairly assigned, that authority is justly applied and not used in a harmful way.

Each of these is a value judgment, while the need for these core values may emerge from an evolutionary basis, the relative importance of each value may emerge from socialization within the family. Thus, saying that Americans value life, liberty, and the pursuit of happiness does not necessarily imply that people from other societies value these things less. Rather it implies that Americans are likely to consider security, autonomy, and connection as needing to fit together in ways that maximize autonomy while not undermining security and connection. Moreover, Americans may differ in the relative weight that they put on these core values depending on their social class, political affiliation, and profession, because each of these provides a particular meaning to what connection, security, and autonomy mean. Connection can focus on loyalty to in-group and security can mean purity and sanctity to differing degrees among different groups. This has been studied both by cultural and cross-cultural psychologists interested in cultures of honor and by psychologists interested in morality both generally and as part of the study of other domains, including religion, politics, and organizational processes.

Definitions

Modern theories of values are grounded in the work of Kohn (class and values), Rokeach (general value systems), Merton (social structure and values), and Kluckhohn (group-level values). For example, Rokeach defines values as shared proscriptive or prescriptive beliefs about ideal modes of behavior and end states (outcomes) – what is worth striving for. Social groups couple these ideals about modes of behavior and outcomes with regulations of allowable ways to attain these outcomes. Thus values can focus on both the outcomes and the procedures or process (behaviors) used to attain the outcomes. Moral foundations theory (Haidt and Joseph) builds on this work focusing on moral principles – values one should live by, and bridges from psychological and sociological accounts to anthropological and evolutionary accounts of morality to understand core values, which they term virtues. Actualizing one's talents, enjoying oneself, leading a purposeful life, caring for the next generation are all examples of operationalization of values into behaviors and outcomes.

Values can be conceptualized on the individual and group level. At the individual level, values are internalized social representations or moral beliefs that people appeal to as the ultimate rationale for their actions. Though individuals in a society are likely to differ in the relative importance assigned to a particular value, values are an internalization of sociocultural goals that provide a means of self-regulation of impulses that would otherwise bring individuals in conflict with the needs of the groups and structures within which they live. Thus, discussion of values is intimately tied with social life. At the group level, values are scripts or cultural ideals held in common by members of a group; the group's 'social mind.' Differences in these cultural ideals, especially those with a moral component, determine and distinguish different social systems. In this sense Weber's Protestant 'ethic' and 'spirit' of capitalism describe value systems.

Values, to which individuals feel they owe an allegiance as members of a particular group or society, are seen as the glue that makes social life possible within groups. Yet, they also set the stage for frictions and lack of consensual harmony in intergroup interactions. Values are thus at the heart of the human enterprise; embedded in social systems, they are what makes social order both possible and resistant to change. Values are not simply individual traits, they are social agreements about what is right, good, to be cherished.

What is common to all value phenomena? At the individual level, values contain cognitive and affective elements and have a selective or directional quality; they are internalized. Preference, judgment, and action are commonly explained in terms of values. Values may be grounded in evolutionary press to prefer people, contexts, and situations which do not spread disease or harm, resulting in values of purity, fairness, caring, loyalty to in-group, and acceptance of authority that are differentially emphasized across time, place, and situation. Individuals take on values as part of socialization into a family, group, and society. Once taken on, values are assumed relatively fixed over time (*see* Values Across Cultures, Development of). Indeed, values that are individually endorsed and highly accessible to the individual do predict that individual's behavior. Conversely, even personally endorsed values would not influence action when they are not made salient to the individual at the time of action. Moreover, in any given situation more than one personally endorsed value may apply, and the behavioral choice appropriate for one value may conflict with the behavioral choice appropriate to another value.

Values are codes or general principles guiding action; they are not the actions themselves nor are specific checklists of what to do and when to do it. Thus, two societies can both value achievement but differ tremendously in their norms as to what to achieve, how to achieve, and when pursuing achievement is appropriate. Values underlie the sanctions for some behavioral choices and the rewards for others. A value system presents what is expected and what is hoped for, what is required and what is forbidden. It is not a report of actual behavior but a system of criteria by which behavior is judged and sanctions applied. Values scaffold likes and dislikes, what feels pleasant and unpleasant, and what is deemed a success or failure. Values and value systems are often evoked as rationales for action; for example, values of freedom and equality were evoked to elicit American support for the Civil Rights movements. Values differ from goals in that values provide a general rationale for more specific goals and motivate attainment of goals through particular methods.

History and Current Developments

Initially viewed with suspicion by Western social scientists as too subjective for scientific study, the concept of values found increasing use beginning with *The Polish Peasant in Europe and America* (Thomas and Znaniecki, 1921). Impetus for the study of cultural values comes from the work of Alfred Kroeber, Clyde Kluckhohn, Talcott Parsons, Charles Morris, Robert Redfield, Ralph Linton, Raymond Firth, A.I. Hallowell, and more currently from Michael Bond, Jonathan Haidt, Geerdt Hofstede, Ronald Inglehart, Milton Rokeach, and Shalom Schwartz.

Kluckhohn and Strodtbeck (1961) proposed that cultural value systems are variations of a set of basic value orientations that flow from answers to five basic questions about being. These are first, what is human nature – evil, neutral, mixed, or good? Second, how do we relate to nature or supernatural – subjugation, harmony, or mastery? Third, what is the nature of time – past, present, or future? Fourth, what is the nature of human activity – being, being-in-becoming, or doing? And fifth, what is the nature of our relationship to others – are we joined vertically, horizontally or are we simply separate individuals? Klukhohn and Strodtbeck also organized a system for comparing values in terms of their level of generalization and function in discourse and conduct, proposing that values fit into a pyramid of ascending generalization. For each society, a few central or focal values were proposed to constitute a mutually interdependent set of what makes for the 'good life.' These include the unquestioned, self-justifying premises of the society's value system and definitions of basic and general value terms; for example, happiness, virtue, beauty, and morality. The basic questions raised by Klukhohn and Strodtbeck continue to be studied, for example, by asking if societies differ in the propensity to accept or value vertical relations in which power differential is clear.

Since American researchers dominated values research, much early work focused on documenting American values. As early as 1944, need for achievement was described as an important or even a defining American value which was in decline (Spates, 1983). The idea that American values are in decline continues to be a hot topic of political debate. Some relevant research focuses on change over time in narcissism, the extent that individuals (over) value themselves, and their own needs and desires compared to others.

Early values studies documented the influence of education, age, type of employment, and socioeconomic status on value preferences of Americans, adding to Weber's thesis of the influence of religion (Protestantism vs Catholicism) on achievement and work values in Europe. Kohn (1977) was responsible for a number of important values surveys documenting that in various European countries and the United States, parents of higher socioeconomic status value self-direction in their children more than parents of lower educational and occupational levels. These findings have been verified cross-nationally in 122 societies. The influence of education on values continues to be hotly debated, the concern being that education carries with it reduced acceptance of traditional values (e.g., submission to authority, obedience). However, within the United States, for example, there is no evidence that political party affiliation or religiosity is causally influenced by obtaining a college degree.

Extending the documentation of American values, Rokeach (1973) validated empirically 36 values related to preferred end states and preferred ways of behaving. Using Rokeach's scale, value differences tied to class, age, race, subculture, and level of differences were documented in many countries. Building on Rokeach, Schwartz (1992) delineates values as ways of articulating universal requirements of human existence – to survive physically, have social interchange, and provide group continuity. For Schwartz, values represent operationalizations of these needs as goals that fit together in meaningful clusters (achievement, self-direction, stimulation, hedonism, universalism, benevolence, tradition, conformity, security, and power). Some clusters are compatible (e.g., stimulation and hedonism) and others compete (e.g., self-direction and conformity).

Using mostly data from teachers and college students in 20 primarily Western countries, Schwartz shows that, with the exception of China, specific values mostly do 'cluster' and 'compete' as expected. Thus, 'honest,' 'forgiving,' and 'helpful' cluster together as 'benevolence,' and 'self-direction,' and 'stimulation' cluster far from 'conformity,' 'tradition,' and 'security.' These data suggest important universality to how values are organized cross-culturally and that societies differ in which clusters of values predominate public life.

More recent analyses, using as its base, all literature published between 1980 and 2000 comparing Americans either with one another or with people from other countries on the extent that they endorse values of individualism and collectivism suggest that Americans are neither the most individualistic nor the least collectivistic of people (Oyserman et al., 2002). There was no difference between America and other English-speaking countries and only relatively small differences between Europe (Western and Central) and America, with Americans being higher in individualism. The majority of available studies (40%) focused on comparisons between East Asians and Americans, making them the most common comparison in the cross-national individualism literature. Combined effect sizes for comparisons with East Asia were moderate in size, as were combined effect sizes for Africa and

the Middle East. Whereas these findings corroborate conventional expectations of cultural theorists, surprisingly the combined effect size for countries in the "other Asia" region including India and Pakistan was small. Moreover, no difference in individualism was found between the United States and South America, with Americans in the United States being less individualistic than Latin Americans in South America in more than half of the country samples tested by Oyserman and her colleagues. With regard to collectivism, Americans were found to value collectivism less than were others from all regions of the world, with the exception of English-speaking countries. Surprisingly, Americans were *lower* in collectivism than Europeans were, befitting the idea of a uniquely American way of being (high individualism and low collectivism) but challenging the notion of a single 'Western' cultural set of values. In addition, effect sizes for Asian regions were similar to those for European regions, with large differences in valuing collectivism only when US and African samples were compared. Almost half of all studies focused on comparisons between East-Asian regions and America but effects were heterogeneous implying that simple between group comparisons would not necessarily yield value differences.

Priming people to consider their cultural values changes the extent that they feel obligated to act to help others rather than do their own thing (for a review, Oyserman and Lee, 2008). Values frameworks have been broadened to include Confucian values such as valuing relationships, valuing the family as the model of relationships, valuing a basic benevolence in engaging with others, and valuing hard work, skill development, and thrift (education, industriousness) (Hofstede and Bond, 1988). Cultural researchers have also shown that changes in societies produce rapid changes in values as well as in which personality and social interaction styles are valued. Emotional expressivity and outgoingness are valued more as societies shift toward individualistic values.

Controversies

Key tensions in the values literature focus on the conditions under which they may influence behavior, and the appropriate level of analyses for seeing values in action. Interest in values as a research focus has ebbed in the past as each paradigm for studying values has been criticized for lack of specificity of findings as due to values as distinct from other constructs, including social norms, attitudes, or situational constraints. Current cultural psychology focuses attention on social structures as the repository of values such as personal freedom, group harmony, personal happiness, and duty or filial piety.

How do we know that values exist? A number of options are available: (1) Individual testimony – people say what values they hold. Yet, self-reports of values are subject to pronounced context effects (*see* Attitudes and Behavior). (2) Behavioral choices – either in naturalistic or laboratory settings, value differences may be imputed from behavior. Yet, behavior is influenced by many variables other than values. At the individual level, values themselves are assumed to link to behaviors via their influence on norms and attitudes, but people may infer their values from their behavior, reversing the causal relationship. (3) Cultural and social structures – expenditure of resources, time, energy, and structuring of the natural environment; cultural products can be seen as concrete residues of value-based choices (see Cross-Cultural Psychology; Cultural Psychology). (4) Social interchange – observation of behavior in situations of conflict, and more generally observation of what is rewarded or punished, praised or vilified provides data for identifying what is socially valued. Here, too, the question of appropriate evidence arises. To what extent is it appropriate to assume that differences in social structures and societies are evidence of value differences? Political and economic influences and simple inertia may set the stage for behaviors, without a causal influence of necessarily values.

Future Directions

Cross-cultural perspectives are currently becoming increasingly central to values discussion. For example Inglehart (1990) documented values and value change in a large multinational study, and thus a large number of two-nation comparison studies has emerged. Another important topic of research is the connection between values of individuals, values of subcultural groups, and values of larger cultural systems and methods for identifying and studying each of these. Perhaps in addition to identifying value vocabularies at each level, it is time to begin to ask whether values appropriately are studied as fixed traits of individuals or as embodied in groups and to what extent values research is synonymous with cultural and cross-cultural research.

Given that any particular behavior importantly is influenced by context effects that make certain information salient at the moment of action, it is not surprising that the effects of individual value endorsements on behavior have a 'sometimes you see it, sometimes you don't' quality about them. But focusing on individual endorsement of values may miss much of the power of value systems to influence everyday life. That is, individuals may not need to personally endorse or have salient particular values in order for their influence to be felt. The most profound influence of values may be through the ways that they influence rules, norms, procedures within a society, and in this way structure the everyday life choices for individuals within a society.

Thus, whereas previous researchers have documented values using survey techniques in which individuals rated the extent to which various values were important to them, future assessment of values may need to consider more indirect approaches such as what services a society provides its members, what behaviors are rewarded or sanctioned and so on.

See also: Attitudes and Behavior; Collectivism and Individualism: Cultural and Psychological Concerns; Critical Psychology; Cross-Cultural Psychology; Levels of Analysis in Social Psychology; Social Cognition; Social Psychology; Values Across Cultures, Development of; Values, Anthropology of; Values, Sociology of; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Darley, J., Shultz, T., 1990. Moral rules: their content and acquisition. Annual Review of Psychology 41, 525–556.

Grusec, J.E., 2011. Socialization processes in the family: social and emotional development. Annual Review of Psychology 62, 243–269.

Haidt, J., Joseph, C., Fall, 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Dædalus 55–66.

Hofstede, G., Bond, M.H., 1988. The confucius connection: from cultural roots to economic growth. Organizational Dynamics 16, 5–21.

Inglehart, R., 1990. Culture Shift in Advanced Industrial Society. Princeton University Press, Princeton, NJ.

Kluckhohn, F., Strodtbeck, F., 1961. Variations in Value Orientations. Row, Peterson, Evanston, IL.

Kohn, M., 1977. Class and Conformity. University of Chicago Press, Chicago.

Merton, R.K., 1957. Social Theory and Social Structure. Free Press, Glencoe, IL.

Oyserman, D., Coon, H.M., Kemmelmeier, M., 2002. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. Psychological Bulletin 128, 3–73.

Oyserman, D., Lee, S.W.S., 2008. Does culture influence what and how we think? Effects of priming individualism and collectivism. Psychological Bulletin 134, 311–342.

Rokeach, M., 1973. The Nature of Human Values. Free Press, New York.

Schooler, C., 1996. Cultural and social-structural explanations of cross-national psychological differences. Annual Review of Sociology 22, 323–349.

Schwartz, S., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology 25, 1–66.

Spates, J., 1983. The sociology of values. Annual Review of Sociology 9, 27–49.

Sullivan, J., Transue, J., 1999. The psychological underpinnings of democracy: a selective review of research on political tolerance, interpersonal trust, and social capital. Annual Review of Psychology 50, 625–650.

Thomas, W., Znaniecki, K., 1921. The Polish Peasant in Europe and America. Chicago University Press, Chicago.

Values, Social Psychology of

Jan Cieciuch, University Research Priority Program Social Networks, University of Zurich, Zurich, Switzerland; and Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland
Shalom H Schwartz, The Hebrew University of Jerusalem, Jerusalem, Israel; and Laboratory of Socio-Cultural Research, National Research University – Higher School of Economics, Moscow, Russia
Eldad Davidov, Institute of Sociology, University of Zurich, Zurich, Switzerland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Psychologists view values as an aspect of personality that underlies and motivates attitudes and behavior. We present overviews of Allport's motivational theory and Rokeach's cognitive theory, and elaborate on Schwartz's integration of earlier approaches. Schwartz defines values as transsituational goals, varying in importance, that serve as guiding principles in the life of a person or group. Values form a circular motivational continuum that reflects their conflicts and congruence and may be universal. We discuss the location of values in personality structure and relations of values to behavior.

The contemporary psychology of values grew out of two sources: personality psychology, inspired by philosophy, and social psychology, closely linked to sociology. Allport, author of the first psychological theory of values (Allport and Vernon, 1931), pioneered the former approach, drawing upon Dilthey's philosophy and Spranger's personality typology (*Lebensformen*). Durkheim (1897/1964) and Weber (1905/1964) introduced values into sociological discourse as a tool for explaining social structure and social change.

The psychology of personality treats value preferences as individual differences that permit a holistic description of the human being. Sociological approaches treat values as an aspect of culture and study their influences on the individual. Social and cross-cultural psychologists often combine these approaches. They use individual differences in value preferences to understand attitudes and behavior and cultural differences in value emphases to characterize and compare societies or other collectivities.

Of the many theories that have emerged in the history of the psychology of values, three have had the most influence on empirical research. Allport's theory of values (Allport and Vernon, 1931), which was also intended as a psychology of personality, came first. Later, Rokeach (1973) proposed a theory of values as a central aspect of personality useful for understanding individual and group differences in attitudes and behavior. Most recently, Schwartz presented a theory of value content and structure (1992, 1994) that integrated earlier approaches and facilitated values research in many branches of psychology, including personality, social, political, cross-cultural, developmental, and positive psychology as well as in the related disciplines of education, sociology, law, and economics. Below we briefly describe Allport's and Rokeach's theories and elaborate on the Schwartz theory that is currently predominant in psychology. We do not discuss the Inglehart value theory (e.g., Inglehart and Baker, 2000) that is prominent in political science and sociology, because it is not a psychological theory.

Roots of the Contemporary Psychology of Values

Allport's Theory of Values

In formulating his theory, Allport was inspired by Dilthey's psychological and philosophical writings. Dilthey treated psychology as a branch of the humanities (*Geisteswissenschaft*), whose primary goal was to understand rather than to explain, the latter being the domain of the natural sciences (*Naturwissenschaften*). Spranger (1914) followed Dilthey by proposing a personality typology aimed at a holistic description of human beings that would facilitate understanding. Allport and Vernon (1931) viewed this typology as the best a priori theoretical foundation for developing a list of values and constructing a tool for their measurement. According to Spranger, the following six personality types, each characterized by its most prominent values, account for the diversity of human personality. The personality of each individual is a synthesis of the different types, with one being dominant and the others subordinate.

1. Theoretical: truth, intellect, criticism, and individualism
2. Economic: usefulness, frugality, practicality
3. Aesthetic: harmony, beauty, form
4. Religious: a sense of unity with the world, a quest for unity through spiritual development
5. Social: love, respect, help for people
6. Political: power, control over people (not necessarily in the political sphere)

Allport and Vernon (1931) offered a four-point justification for assigning values a privileged role in the study of personality. First, they asserted that measuring isolated traits, capacities, or other specific constructs yields a misleading understanding of personality because such information provides only fragments but not a holistic view of a person. They averred that personality is not the sum of isolated elements but the overarching structure that determines those elements. Second, they argued that the personality of each person is a unique system, not strictly

comparable with any other, and that this assumption is central to a scientific account of personality. Following from the above was their third point: An appropriate construct for investigating personality must be able both to describe each individual as a coherent system rather than a sum of isolated aspects and to provide a precise account of the variety of human personalities. Finally, Allport and Vernon concluded that the value construct best fulfills the requirements in their third point.

Allport and collaborators (Allport and Vernon, 1931) proposed a value instrument to measure personality based on the Spranger typology, entitled *A Study of Values* (SoV). Together with Lindzey, Allport published a modified version of the SoV 20 years later (Allport et al., 1951).

The concept of value that Allport and colleagues adopted combined two psychological meanings: (1) value as an interest and (2) value as an evaluative attitude. Both the original and modified versions of the SoV questionnaire used the terms *values* and *interests* interchangeably. The interpretation of values as interests facilitated the practical application of the SoV and helped to popularize the theory. Applied psychology soon picked up this psychological adaptation of Spranger's typology and employed it in such fields as vocational guidance, management, and education.

When discussing the theoretical foundations of their questionnaire, Allport and Vernon (1931) also used the terms *evaluative attitudes* (or *forms of valuation*) interchangeably with values. Upon perceiving any object, people engage in the everyday activity of evaluation based on their values. Allport and Vernon proposed that people might make six possible types of evaluations, corresponding to the six personality types defined above. In explaining this conception of values as evaluations, they cited Spranger's example of evaluations of a ring. Depending on their dominant values, people may regard a ring as a glittering object (an aesthetic evaluation), an object made of a certain metal (a theoretical evaluation), an object having a certain market value (an economic evaluation), a symbol of love and loyalty (a social evaluation), a symbol of certain rights and duties (a political evaluation), or an object with some sacred or mystical value (a religious evaluation).

As bases of evaluation, values are likely to influence one's perceptions. This view of values created opportunities for empirically verifying their psychological reality in non-questionnaire surveys. In an early review paper, Allport and Cantril (1933) described the first studies of this type (e.g., effects of values on selective perception of the news). The concept of value proposed by Allport was used in the New Look approach to perception. According to this approach, perception is an active and constructive process influenced by many factors and values are one of the most important (Postman et al., 1948).

Rokeach's Theory of Values

Rokeach (1973) also conceptualized values within the framework of personality theory. In contrast to Allport, Rokeach adopted a cognitive approach. He viewed personality as composed of concentric subsystems of beliefs organized from the most central to the person to the most peripheral. The core subsystem of personality is the set of beliefs about oneself, with the next most central subsystems concerning their values.

Rokeach started a new tradition in psychological research on values by adopting Kluckhohn's (1951) view of values as explicit or implicit conceptions of the 'desirable.' These conceptions of the desirable may characterize groups as well as individuals. This reconnected the psychological study of values with its sociological roots and drew attention to the influence of social experience and location on value priorities.

Rokeach (1973: p. 5) defined values formally as "enduring beliefs that a specific mode of conduct or end state of existence is personally or socially preferable to an opposite or converse mode of conduct or end state of existence." By referring to values as beliefs about more and less preferred modes of conduct or end states, Rokeach conceptualized values as forming a hierarchical system. When referring to them as 'enduring,' he explained that values are acquired in an absolute, all-or-none manner that resists change, thereby insuring the continuity of personality. He held, however, that values do change through experience and processes of maturation in which absolute beliefs about the desirable come into conflict. Thus, his values approach makes it possible to consider a human being in terms of both permanence and changeability. Rokeach (1973) saw this changeability as an advantage of values over traits for describing personality, because traits presumably do not change. In sum, for Rokeach, beliefs are the main category for describing values and values are the basic unit for describing personality.

Rokeach criticized Allport because his list of values was too general; profiles on his six personality types profile could not distinguish adequately among the personalities of different people. Rokeach (1982) exemplified this problem by noting that two people for whom political values were dominant in their personality profile might be very dissimilar or even radically different if one of them embraced leftist values and the other rightist ones. This is an important criticism because it demonstrates that Allport's approach could not achieve its main objective, to describe individuals' personalities comprehensively and enable clear and informative interindividual comparison.

Rokeach (1973) sought to formulate a catalog of values that would be as complete as possible. Based on intuitive, theoretical, and empirical grounds, he estimated that the catalog of values held by people is not large. Rather than each person having a unique combination of values, people differ in the relative importance they attribute to the same limited set of values, that is, in their hierarchy of values. In other words, the same values guide all human beings, but their preference ranks differ. Rokeach proposed a list of 36 values, composed of an equal number of terminal or end-state values (e.g., wisdom, justice) and of instrumental or means/modes of conduct values (e.g., honest, ambitious). He used these categories in developing a measurement instrument entitled the *Rokeach Value Survey* (RVS) which consists of two lists, one of terminal values and the other of instrumental values. Respondents rank the 18 values in each list from most to least important as guiding principles in their own lives. Although the RVS presents the values as abstract terms with short definitions, the ranking of values is sufficiently specific to reveal subtle

differences in value hierarchies between individuals. This overcomes the weakness that Rokeach identified value surveys formulated in the Allport tradition. Hundreds of studies have employed the RVS to characterize and compare individuals and groups in an informative manner.

Schwartz's Theory of Values

Common Characteristics of Values in Different Theories

Schwartz (1992) defines values as transsituational goals, varying in importance, that serve as guiding principles in the life of a person or group. What differentiates among values is their motivational content. This definition integrates Rokeach's cognitive understanding of values as goals and Allport's motivational content of values derived from his view of values as both interests and evaluative attitudes.

Schwartz (1992, 2006) suggested that five characteristics of values are common to most previous psychological approaches to values: values (1) are beliefs linked to emotions; (2) refer to desirable goals that motivate action; (3) transcend specific actions and situations; (4) serve as standards for evaluating actions, policies, people, and events; and (5) form a relatively enduring hierarchical system ordered by importance. To this he added that (6) the impact of values on everyday decisions is rarely conscious, and (7) any action or attitude is guided by the relative importance of multiple, competing values, that is, by trade-offs among relevant values. Compared with values, attitudes are less abstract and less central to the self, relate to specific objects, derive their valence from the values that guide them, and do not serve as standards.

Circular Continuum of Values

Schwartz (1992) distinguished between the value hierarchies of individuals and the structure of relations among values themselves. Each individual has his or her unique hierarchy of values that determines the trade-offs that guide his or her behavior and attitudes. Schwartz's major innovation is his assumption that relations among different values have an inherent structure that is probably universal. He postulates that this structure takes the form of a circular motivational continuum (see Figure 1). Circularity expresses three features of the structure: adjacent values are motivationally compatible (e.g., security, tradition, and conformity), the similarity between values decreases with the distance between them around the circle, and motivationally opposed values are located on opposite sides of the circle (e.g., tradition opposes stimulation, self-direction, and hedonism).

Schwartz proposed three principles that explain the universal order of value relations around the circle. First, one can simultaneously pursue adjacent values in a single action (e.g., tradition and conformity by participating in religious rituals); adjacent values are compatible motivationally. Conversely, one cannot simultaneously pursue values located on opposite sides of the circle in the same action (e.g., climbing Everest as an expression of stimulation undermines security); opposite values are incompatible motivationally.

Viewing values as organized along two bipolar dimensions summarizes the oppositions between competing values. As the

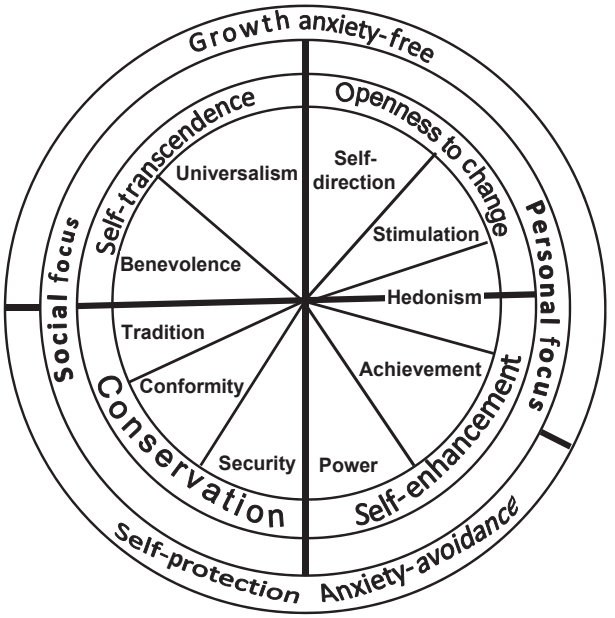

Figure 1 The motivational continuum of values with the principles underlying their circular order. Modified version of the figure published in Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna, M. (Ed.), Advances in Experimental Social Psychology, vol. 25. Academic Press, London, pp. 1–65.

second circle from the middle of Figure 1 shows, one dimension contrasts *openness to change* and *conservation* values. This dimension captures the conflict between values that emphasize independence of thought, action, and feelings and readiness for change (self-direction, stimulation) and values that emphasize order, self-restriction, preservation of the past, and resistance to change (security, conformity, tradition). The second dimension contrasts *self-enhancement* and *self-transcendence* values. This dimension captures the conflict between values that emphasize concern for the welfare and interests of others (universalism, benevolence) and values that emphasize pursuit of one's own interests and relative success and dominance over others (power, achievement). The poles of these dimensions are called higher order values.

A second principle that organizes the value circle is the interests that value attainment serves. Values on the left of Figure 1 primarily regulate how one expresses personal interests and characteristics. Values on the right primarily regulate how one relates socially to others and affects their interests. Relations of values to anxiety are a third organizing principle. Values in the bottom part of the circle are based in the need to avoid or control anxiety and threat and to protect the self (Schwartz, 2006). Conservation values (bottom left) emphasize avoiding conflict, unpredictability, and change by submission and passive acceptance of the status quo. Power values (bottom right) emphasize overcoming sources of anxiety by actively controlling threat. In contrast, values in the top part of the circle are relatively anxiety free, expressing growth, and self-expansion. Self-transcendence values (top left) emphasize promoting the

welfare of others. Openness to change values (top right) emphasize autonomous, self-expressive experience. Achievement values do both: Successfully meeting social standards may control anxiety and it may affirm one's sense of competence.

One can segment a circular continuum in many ways (cf the color circle). Schwartz (1992) discriminated the 10 basic values and four higher order values in **Figure 1** arbitrarily, for scientific convenience. Most subsequent research has used these distinctions. **Table 1** lists the 10 values in their order around the circle, with a short definition of each.

Depending on one's objectives and the desired precision of one's tools, however, one can divide the continuum differently. A recent refinement of the Schwartz model partitions the same continuum into 19 more narrowly defined values whose effects are distinguishable (Schwartz et al., 2012). It distinguishes two types of self-direction (thought and action), two types of power (dominance and resources), two types of security (societal and personal), two types of conformity (rules and interpersonal), two types of benevolence (caring and dependability), and three types of universalism (nature, concern, and tolerance), plus two new values (face and humility).

Any catalog of values could be criticized as incomplete regardless of its theoretical bases. Researchers are bound to find new value concepts to add to the catalog of basic values as they study additional cultural groups. Thinking of values metaphorically as arrayed on a straight line, even if it is a continuum, one could always add new values at the ends of the line. The circular model is different because the space is closed and finite. Any supposedly new value must necessarily fit somewhere on the existing circle; it must represent some finer discrimination in the circle of motivations or a combination of recognized values. Adopting a conception of values as forming a circular motivational continuum has important, far-reaching implications. Four implications are particularly significant.

1. All the values that are recognized across human cultures are located on the motivational continuum of values.
2. The motivational continuum may be divided into as many or as few specific values as fit a researcher's objectives, but the relations among the values a researcher chooses are predetermined by the order of the circular continuum.
3. The structure of values consists of more than a list of values or groups of values. The circular arrangement reflects the similarities of adjacent elements of the circle and dissimilarities (or conflicts) between elements on opposite sides of the circle.
4. The structure of values provides a matrix that can reveal the motivational meaning of any particular value by examining its location relative to the other values in the circle.

Table 1 Ten basic individual values in the original theory, defined in terms of their core motivational goals

Basic value	*Core motivational goal*
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact.
Tradition	Respect, commitment, and acceptance of the customs and ideas that one's culture or religion imposes on the individual.
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
Security	Safety, harmony and stability of society, of relationships, and of self.
Power	Social status and prestige, control or dominance over people and resources.
Achievement	Personal success through demonstrating competence according to social standards.
Hedonism	Pleasure and sensuous gratification for oneself.
Stimulation	Excitement, novelty, and challenge in life.
Self-direction	Independent thought and action choosing, creating, exploring.

Adapted from: Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna, M. (Ed.), Advances in Experimental Social Psychology, vol. 25. Academic Press, London, pp. 1–65.

Empirical Verification of the Circular Model of Values

Evidence supporting the circular model of values comes from cross-cultural studies with questionnaires designed to measure the 10 (or 19) values and from experimental studies and studies of value change. The questionnaires include the Schwartz Value Survey (Schwartz, 1992) that asks respondents to rate the importance of 57 abstract values as guiding principles in their life, three versions of the Portrait Values Questionnaire that ask respondents how similar they are to each of 21, 40, or 57 individuals who are described in terms of their important values (Schwartz, 2003, 2006; Schwartz et al., 2012), the Picture-Based Value Survey with 20 pictures that depict a child performing a value-relevant act which children rank for importance as they would like to be themselves using a five category Q-sort (Döring et al., 2010), and two types of forced-choice questionnaires in which respondents indicate which value is more and which less important to them as a guiding principle (Bilsky et al., 2012; Lee et al., 2008). Data using these different methods in over 75 countries, from hundreds of convenience samples and from some 40 national representative samples, support the circular value structure, although it is not always possible to discriminate among the adjacent values in the circle (Schwartz, 2011).

The circular structure implies that if a value becomes more important, values adjacent to it in the circle should also increase in importance whereas those opposed to it should lose importance. This pattern of structural change has been found both in laboratory experiments that employed value-priming interventions (Pakizeh et al., 2007; Maio et al., 2009) and in longitudinal studies in real life (Bardi et al., 2009). Maio (2010) summarizes additional experimental evidence to show that people represent values mentally in terms of the motivational conflicts and compatibilities the value circle depicts.

Values and Personality

McAdams and Pals' (2006) proposal to differentiate among three levels of personality provides a convenient framework to locate values in the personality structure. Their first level is consists of dispositional traits (usually Big Five personality traits), the second corresponds to cognitive and motivational constructs ('characteristic adaptations'), and the third concerns life narratives. Values might be located at the first or second level of this personality structure. At the first level, values can be viewed as a dynamic, motivational, aspect of personality. One could then draw a parallel between traits and values. Personality traits are biologically based formal characteristics of behavior, whereas values can be viewed as biologically based motivational tendencies (Cieciuch, 2012; McCrae, 2009). The biological basis of each may explain why studies across so many cultures support the Big Five traits and the value circle.

However, Schwartz (1992) follows Rokeach (1973) in arguing that values are cognitive/emotional transformations of needs or drives into goals. He grounded the values he identified in three universal requirements of human existence: needs of individuals as biological organisms, requisites of coordinated social interaction, and survival and welfare requirements of groups. Every value represents a goal that expresses at least one or more of these universal requirements (Schwartz, 1994). Values serve as socially acceptable terms that people use to communicate with others about their goals and to coordinate in pursuing them. Every person must cope with the universal requirements that underlie values. However, each person's unique biological heritage and social experiences combine to influence the relative importance assigned to each value, the person's own value hierarchy. Thus, values are characteristic adaptations shaped by the interplay of basic personality traits and the stimuli of the social and natural world.

Values and Behavior

Although many factors typically influence any specific behavior, values are often an important contributing factor. For values to be linked to behavior, they must be experienced as relevant. Because basic values are abstract, their relevance in particular situations is often not obvious. As Maio (2010) noted, people may see equality, for example, as relevant in situations of ethnic or gender diversity, but may fail to see it as relevant to discrimination based on weight. He proposed that people's abstract values relate more consistently to their behavior when they have 'concrete instantiations' of those values, that is, when they have thought of specific ways in which a value is or can be expressed in real-life situations. Maio (2010) summarizes series of experiments which demonstrate that values affect behavior more when people have thought about tangible applications of those values and when those applications are typical (gender equality) rather than atypical (handedness equality). Typicality makes people perceptually ready to 'see' a value as relevant in subsequent situations.

Several mechanisms link values to behavior (Bardi and Schwartz, 2003; Schwartz, 2009). First, values influence choices because people act in order to attain, affirm, or preserve the goals that their important values express. For example, people vote for parties whose platforms facilitate attaining the values they cherish and against parties whose platforms threaten attainment or preservation of these values (Caprara et al., 2006). Put differently, people's values determine the valence that people assign to the consequences of any action. Actions are more attractive, more subjectively positive, to the extent that they promote attainment of valued goals. The value-based assessments of potential choices typically occur outside of conscious awareness. Values enter awareness when the actions, one is considering, have conflicting implications for different values one cherishes.

Second, values influence behavior through their effects on planning. The higher the priority given to a value, the more likely people will be to form action plans that can lead to its expression. Planning focuses people on the pros of desired actions rather than the cons. It increases people's belief in their ability to reach the valued goal, their persistence in the face of obstacles and distractions, and their belief in the desirability of the goal. By promoting planning, value importance increases value-consistent behavior.

Third, values influence behavior through their effect on attention, perception, and interpretation in situations. People attend more to those aspects of situations that threaten or offer opportunities to attain their cherished values. One person may perceive a job offer as an opportunity for greater self-direction, another as a threat to her security. One may interpret the offer as a chance to help his family, another as a chance to gain power and influence. Each value-based perception or interpretation suggests that a different line of action is desirable.

Values on the Boundary of Psychology

The approaches discussed above treat values as characteristics of individuals, thereby placing value theories in mainstream psychology. Value priorities are also used to describe and compare societies and cultures. Research on societal or cultural values usually aggregates the values of group members to obtain mean priorities (e.g., Inglehart and Baker, 2000). Theories differ in whether they view cultural values as located in the minds of individuals (e.g., Hofstede, 2000) or as hypothetical, latent variables that support and justify the functioning of societal institutions (Schwartz, 2011). This research primarily investigates effects of cultural values on individual attitudes, beliefs, and behavior, and relations of cultural values to other characteristics of groups and societies (affluence, social policies, socioeconomic change, etc.). Thus, the study of values extends out from psychology to all of the social sciences.

See also: Attitudes and Behavior; Big Five Factor Model, Theory and Structure; Cross-Cultural Psychology; Personality Assessment; Personality, Trait Models of; Values Across Cultures, Development of; Values, Psychology of; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Allport, G.W., Cantril, H., 1933. Recent applications of the study of values. The Journal of Abnormal and Social Psychology 28 (3), 259–273.

Allport, G.W., Vernon, P., 1931. A test for personal values. The Journal of Abnormal and Social Psychology 26 (3), 231–248.

Allport, G.W., Vernon, P., Lindzey, G., 1951. Study of Values: Measuring the Dominant Interests in Personality. Manual of Directions. Houghton Mifflin Company, Cambridge.

Bardi, A., Schwartz, S.H., 2003. Values and behavior: strength and structure of relations. Personality and Social Psychology Bulletin 29, 1207–1220.

Bardi, A., Lee, J.A., Hofmann-Towfigh, N., Soutar, G., 2009. The structure of intra-individual value change. Journal of Personality and Social Psychology 97 (5), 913–929.

Bilsky, W., Janik, M., Schwartz, S.H., 2011. The structural organization of human values-evidence from three rounds of the European Social Survey (ESS). Journal of Cross-Cultural Psychology 42 (5), 759–776.

Bilsky, W., Gollan, T., Grad, H., Mendes Teixeira, M.L., Rodriguez Monter, M., Schweiger-Gallo, I., 2012. Computerized Paired Comparison of Values (CPCV): Cross-Cultural Evidence for the Validity of an Alternative Form of Measuring Value. Paper presented at the International Association for Cross-Cultural Psychology 21st Congress, Stellenbosch (18-21.07.12).

Caprara, G.V., Schwartz, S.H., Capanna, C., Vecchione, M., Barbaranelli, C., 2006. Personality and politics: values, traits, and political choice. Political Psychology 27, 1–28.

Cieciuch, J., 2012. Big Five and Big Ten – between Aristotelian and Galilean physics of personality. Theory and Psychology 22 (5), 689–696.

Döring, A.K., Blauensteiner, A., Aryus, K., Drögekamp, L., Bilsky, W., 2010. Assessing values at an early age: the picture-based value survey for children. Journal of Personality Assessment 92, 439–448.

Durkheim, E., 1897/1964. Suicide. Free Press, Glencoe, Il.

Hofstede, G., 2000. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations, second ed. Sage, Beverly Hills, CA.

Inglehart, R., Baker, W.E., 2000. Modernization, cultural change, and the persistence of traditional values. American Sociological Review 65, 19–51.

Kluckhohn, C.K.M., 1951. Values and value orientations in the theory of action. In: Parsons, T., Sils, E. (Eds.), Toward a General Theory of Action. Harvard University Press, Cambridge, MA, pp. 388–433.

Lee, J.A., Soutar, G., Louviere, J., 2008. The best-worst scaling approach: an alternative to Schwartz's Values Survey. Journal of Personality Assessment 90, 335–347.

Maio, G.R., 2010. Mental representations of social values. In: Zanna, Mark P. (Ed.), Advances in experimental social psychology, vol. 42. Academic Press, Burlington, pp. 1–43.

Maio, G.R., Pakizeh, A., Cheung, W.-Y., Rees, K., 2009. Changing, priming, and acting on values: effects via motivational relations in a circular model. Journal of Personality and Social Psychology 97, 699–715.

McAdams, D., Pals, J., 2006. A new Big Five: fundamental principles for an integrative science of personality. American Psychologist 61, 204–217.

McCrae, R., 2009. The physics and chemistry of personality. Theory and Psychology 19, 670–687.

Pakizeh, A., Gebauer, J.E., Maio, G.R., 2007. Basic human values: inter-value structure in memory. Journal of Experimental Social Psychology 43 (3), 458–465.

Postman, L., Bruner, J., McGinnies, E., 1948. Personal values as selective factors in perception. The Journal of Abnormal and Social Psychology 43 (2), 142–154.

Rokeach, M., 1973. The Nature of Human Values. The Free Press, New York.

Rokeach, M., 1982. On the validity of Spranger-based measures of value similarity. Journal of Personality and Social Psychology 42 (1), 88–89.

Schwartz, S.H., 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In: Zanna, M. (Ed.), Advances in Experimental Social Psychology, vol. 25. Academic Press, London, pp. 1–65.

Schwartz, S.H., 1994. Are there universal aspects in the content and structure of values? Journal of Social Issues 50, 19–45.

Schwartz, S.H., 2006. Basic human values: theory measurement and applications. Revue francaise de sociologie 47 (4), 929–968.

Schwartz, S.H., 2009. Basic values: how they motivate and inhibit prosocial behavior. In: Mikulincer, M., Shaver, I.P. (Eds.), Herzliya Symposium on Personality and Social Psychology, vol. 1. American Pychological Association Press, Washington, pp. 221–241.

Schwartz, S.H., 2011. Values: individual and cultural. In: van de Vijver, F.J.R., Chasiotis, A., Breugelmans, S.M. (Eds.), Fundamental Questions in Cross-Cultural Psychology. Cambridge University Press, Cambridge, pp. 463–493.

Schwartz, S.H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., Dirilen-Gumus, O., Konty, M., 2012. Refining the theory of basic individual values. Journal of Personality and Social Psychology 103, 663–688.

Schwartz, S.H., 2003. Value Orientations. European Social Survey Core Questionnaire Development (Chapter 7). At: http://www.europeansocialsurvey.org/index.php?option=com_docman&task=cat_view&gid=83&Itemid=80.

Spranger, E., 1914. Lebensformen. Geisteswissenschaftliche Psychologie und Ethik der Persönlichkeit. Niemeyer, Halle.

Weber, M., 1905/1958. The Protestant Ethic and the Spirit of Capitalism. Scribner's, New York.

Values, Sociology of

Helmut Thome, Martin-Luther-University Halle-Wittenberg, Halle-Wittenberg, Germany

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R. Rezsohazy, volume 24, pp. 16153–16158, © 2001, Elsevier Ltd.

Abstract

The first section of this article presents core elements in the object- or concept-focused definition of values, including some of the functions ascribed to them: guiding and coordinating actions, marking personal and collective identities. The next section introduces some additional analytical dimensions: the substantive, temporal, and social generalization of values; their position in means–ends relationships; conceptual distinctions with regard to needs, attitudes, and norms. The third section addresses some issues regarding the development of values on the personal and macrosocial level, concentrating on the cognitive developmental approach. The final section presents a brief outline of two prominent approaches (initiated by Inglehart and Schwartz) in studying the development and spread of values on the basis of data collected in a long-term series of cross-national surveys conducted in an increasing number of countries worldwide.

Values are generally considered to play a crucial role in motivating and guiding human action and to be constitutive elements in the construction of personal and collective identities; they serve as a potent source of social integration but also of conflict and division. Values are conceptualized quite differently in various sociological theories. This article focuses on those analytical distinctions and fundamental assumptions that are widely shared among sociologists and other social scientists, but some of the issues on which theoretical perspectives diverge will also be mentioned.

What Are 'Values,' How to Define Them?

There are numerous definitions – or rather explications of meaning – available in the sociological literature. Most of them can be classified into two groups: (1) 'values' as something directly attached or ascribed to preferred objects and (2) 'values' as (enduring) beliefs or conceptions that construe something as preferable or desirable. The first approach is exemplified by Thomas and Znaniecki (1958: p. 21) who write: "By a social value we understand any datum having an empirical content accessible to the members of some social group and a meaning with regard to which it is or may be an object of activity. Thus, a foodstuff, an instrument, a coin, a piece of poetry, a university, a myth, a scientific theory, are social values." In a similar vain, R.H. Turner (1968: p. 100) states: "A value is a category of objects…." The second approach, which has become the dominant one within sociology (in contrast to economic theory), is represented by C. Kluckhohn (1951: p. 395) who, drawing upon Dewey, writes: "A value is a conception, explicit or implicit, distinctive of an individual or characteristic of a group, of the desirable which influences the selection from available modes, means, and ends of action." And he emphasizes, "that affective ('desirable'), cognitive ('conception'), and conative ('selection') elements are all essential to this notion of *value*" (ibid.).

The crucial point here is the distinction between something simply desired (matter-of-factly) and something conceived to be desirable, based on some reasoned judgment. An object may be desired directly as such for whatever qualities it innately has (or is assumed to have). But in order to 'value' an object – according to Kluckhohn's conception – it must fulfill certain criteria that transcend it. This transfer of meaning, however, may advance into something like a transformation of the object: it may take on a symbolic quality, which makes it appear and then be treated as a value in itself, while its connection to some 'conception' of desirability might recede in the background; patriots, for example, may endow the national flag with such a quality. Thus, the object-focused and the concept-focused definitions are commensurate to some extent: a value may be instantiated in or symbolically represented by some object or behavior, which in turn becomes indicative of the value construct.

There is an element of circularity in Kluckhohn's definition: how can desirability be stated without referring to preconceived values? This circularity can be mitigated by interpreting values as 'second-order' desires (H.G. Frankfurt) and explicating the experiences and processes by which values are generated or arrived at. A first-order desire can be conceived as some want (drive, urge, impulse), directed toward some object that is given independently from any judgment regarding its appropriateness. We may act upon this want spontaneously. But since we experience different sorts of such first-order desires, some of them competing with each other and with those of other people, we are led into reflections seeking or constructing criteria that enable us to evaluate and rank order our desires, abstracting and transforming them into second-order desires and, in an additional step, into a second-order volition. Of course, there is always the possibility of getting overwhelmed by our spontaneous urges and desires.

The evaluative standards applied in such processes comprise different elements; Parsons, for example, distinguishes cognitive, appreciative, and moral standards (Parsons and Shils, 1951: pp. 162ff.), and Kluckhohn (1951: p. 397) emphasizes that "the consistent connection between values and notions of approval and disapproval implies the potentiality for rational justification"; he also sees a "union of reason and feeling inherent in the word *value* (ibid., 400)." The meaning of such

evaluative standards is revealed in binary (or bipolar) codes in terms of which judgments are being made – like 'good' vs 'evil/bad' (in the case of moral standards), or 'true' vs 'false' (cognitive standards), 'beautiful' vs 'ugly' (aesthetic standards). Some of these codes may be applied reflexively. For example, one may judge it as good or bad (right or wrong) to deliver or not deliver a moral judgment (in terms of good or bad) in a specific situation; moralizing, however, might be considered inappropriate with regard to certain issues in certain social contexts. We thus move from second- to higher-order desires. The 'moral matrix,' i.e., the set of issues or dimensions requiring or allowing for moral judgment varies between cultures and among social groups (Haidt, 2012).

Any culture offers to the person a repertoire of evaluative standards, *social* values; they are conveyed and inculcated, over various stages of development, via socialization processes and supported by social practices and interaction rituals (Durkheim, Collins), but they also need to be interpreted individually and variedly depending on the specifics of the situational context in which they are employed. A single act normally touches upon various (and often competing) values (like honesty and politeness); on the other hand, a single value may be related to a larger set of actions. Values thus come into play not singularly but as a plurality, which means they must somehow be harmonized, arranged into a value 'system' or comprehensive sets of value orientations in which they are hierarchically ordered, to some extent at least. Such an arrangement, however, must also allow for a considerable degree of flexibility, it must remain open for context-specific interpretations and modifications over time. In this sense, values are enduring but not static.

As cited above, Kluckhohn defines values as "distinctive of an individual or characteristic of a group"; as such they may serve as markers of personal and collective identity. On whatever level of social generalization values are analytically located, for most sociologists the concept contains an element of normativeness. Desirability implies a notion of shared beliefs and commitments, including a commitment to one's true self, which incorporates the perspective of a 'generalized other' (Mead). The striving to realize some value is thus connected with a sense of oughtness, something requested from oneself and from others, a request that, if not met, may elicit guilt and shame feelings (imposed upon oneself) or indignation and despise (at someone else), or, if met, pride and respect. The reach and intensity of the ought component varies; there are values considered to be (more or less) 'private,' at one's personal disposal, others (particularly moral values) are more or less obligatory, and placed under public control.

Kluckhohn includes, as we have seen, a functional description in his definition of values by stating that "[a] value influences the selection from available modes, means, and ends of action." Some additional functions can be added here (for a longer list see Rokeach, 1973: pp. 12–16): (1) Values not only guide actions (though they are, by far, not the sole determinant), they also influence perceptions and feeling states. (2) They provide a 'vocabulary of motives' (Gerth and Mills, 1953), which helps to make sense of and to interpret, explain, justify, or criticize one's own as well as other people's behavior, prospectively and retrospectively. (3) Shared (social) values provide a base for commonalities and trustful communication; they stabilize expectations needed to coordinate interaction over the long run. They perform these functions to the extent that they are uncontested. "They can be taken for granted and this taken-for-grantedness can also be taken for granted" (Luhmann, 1996: p. 31). In other situations, however, values can also be disputed among persons and groups and thus instigate conflicts.

Some Further Analytical Dimensions and Distinctions

Dimensions of Generalization

As concepts, values are symbolically (re-)presented by words, behavior (e.g., gestures and rituals), and visible objects (e.g., badges or monuments). As symbols, they "involve 'interpersonalizing' the kind of 'abstraction' or 'generalization' which characterizes all stable systems of orientation" (Parsons and Shils, 1951: p. 161). Luhmann (1998: pp. 340f.) conceives of values as 'symbolically generalized media of communication,' and he distinguishes three interdependent dimensions of generalization: the substantive, the temporal, and the social.

The first dimension refers to the delimitation of semantic meaning, the substantive content, and thematic unity of the concept, which is specified by some contrast or difference: what pertains to it and what is excluded, e.g., equality vs inequality, respect vs disrespect, granted to whom and whom not. The semantic meaning of a value concept can be abstracted to lower or higher levels and specified across many or not so many situations and institutional contexts.

The temporal generalization is brought about by various forms of custom building and habit forming or by explicit norm setting. Particularly, the norms connected with or inherent in values lend certainty and stability to expectations derived from them: In case of disappointment it is not the expectation that needs to be corrected, rather the perpetrator is in the wrong (on the distinction between norms and values see below). Value surveys often fail to adequately capture the element of normativity tight up with values.

Social generalization seems to be an obvious ingredient of values. They are shared among different persons – but shared to what extent? On the lower end of the scale we have 'personal' or 'privatized' values, shared by only a few people; or they may be held in common by many people but unknowingly so and/or without any normative request. 'Social' generalization is thus intimately linked with temporal or normative generalization, and there are at least two subdimensions that need to be distinguished here. One is the factual range of agreement (consensus) or disagreement (dissent), the other is the assumed or perceived range of consensus or dissent. Perceived and factual agreement may fall apart in both directions. People not only act on their own attitudes and personal values, but also on what they perceive to be other people's beliefs and values. Unfortunately, empirical value surveys usually limit themselves to registering only the degree of factual consensus (by aggregating individual data, calculating means and variances); they regularly fail to ascertain at least the first level of reciprocity, i.e., their respondents' perceptions of what other people think and value (more on this and normativity in Thome, 2003).

Instrumental and Terminal Values, Empirical and Transcendental Ends

In his famous book on *The Nature of Human Values*, Rokeach (1973: pp. 7f.) distinguishes between 'instrumental' and 'terminal' values; in the same sense he also speaks of 'means and ends values.' The former refers to 'modes of conduct' (e.g., being 'ambitious,' 'self-controlled,' 'independent'), which are thought to be instrumental in attaining the latter, also referred to as 'end-states of existence,' like 'happiness,' 'freedom,' 'salvation,' or 'self-respect' (ibid., 12). When Rokeach considers all 'moral' values (like 'honesty') to form a subgroup within the category of instrumental values, his conception runs counter to what many other scholars construe as moral values. He acknowledges the possibility "that one terminal value, so defined, is instrumental to another terminal value or that an instrumental value is instrumental to another instrumental value" (ibid.), but his focus lies on the assumption that there is a certain type, a fixed set of values, which are all instrumental to a second type of values, the 'terminal' ones, which in turn cannot (by definition) be employed as means to attain an 'instrumental' value.

Parsons' conception of 'means–ends chains' is more flexible on this: "concrete action does not take place in terms of isolated ends with only one set of means, but rather in terms of very long and complex 'chains' of means-end relationships, so constituted that the 'end' of one sector of the chain is a means to some further end (or more than one)" (Parsons, 1990: p. 322). The ultimate ends (Parsons also speaks of 'ultimate ends or values,' p. 323) envisioned or gradually emerging (in the course of action) can be 'empirical' or 'transcendental.' Whether or not an empirical end is reached (by causally adequate means) can be checked by the actor and by an outside observer. Whether or not a transcendental end has been attained is not arguable by an external observer in terms of scientific criteria. In this case, the means–ends *relationship* becomes 'symbolic,' e.g., when people perform religious rituals. The 'symbolic appropriateness' attributable to this relationship can be interpreted by an external observer, it does not need to be explicitly conscious to the actor himself (Parsons, 1968: p. 258).

Means–ends relationships are often looked at in only one direction: what has been attained as an end can, in the next step, be employed as a means to reach another end. But Simmel already pointed out that originally trivial means can be turned into (ultimate) ends, referring to money as a striking example.

Related Concepts: Needs, Attitudes, and Norms

Needs

Needs are generally conceived as more or less urgent wants or desires, which are registered or make themselves felt as such by the actor herself/himself or by an observer. They can be given independently from any evaluation (i.e., they are 'first-order desires'), though the way they are shaped and perceived (e.g., as necessities that assure survival in a natural and social environment) is likely to be influenced by some evaluative reflections or other external considerations. Often they are seen as rooted in the organic system (physiological processes) or in 'human nature' (including its social embeddedness), independent from one's volition. There are different conceptions concerning the interconnectedness of needs and values. They can be conceived as the input, the raw material from which values are abstracted. Ball-Rokeach and Loges (1992: pp. 2222f.) contend that "(v)alues are cognitive representations of human needs … [they] unite individual's needs with the broader needs of society." Kluckhohn (1951: pp. 428f.) states: "Values *both* rise from and create needs", but he also notes that "some choices do not involve value elements but solely need elements." The (relative) urgency of needs (depending on their respective level of satisfaction) changes with changing natural environments, technologies, and social conditions; needs may thus serve as a mediating factor in the development of values.

Attitudes

Thomas and Znaniecki (1958: p. 22) define *attitude* as "individual consciousness which determines real or possible activity of the individual…. The attitude is thus the individual counterpart of the social value." Kluckhohn (1951: p. 423), too, emphasizes the 'exclusive referability to the individual' as a core definitional element in the concept of 'attitude,' but he adds a second 'principal difference from *value*' namely "absence of imputation of the 'desirable.'" Though 'attitude' implies a positive, negative, or indifferent stance taken toward some object, for whatever reason, the concept as such implies no element of oughtness. Attitudes may engage values, but values may also become the object of an attitude.

Norms

Norms are generally defined as (formally or informally given) rules about what persons – in most cases a specified set of persons – ought to do or not to do in given circumstances, enforced by the threat of sanctions in the case of transgression. In contrast, values inspire and guide behavior, but they are quite independent of specific situations, and they do not directly prescribe or proscribe a specific behavior. Norms are restrictive in their sheer oughtness; values instead are enhancing, they inspire a striving for something 'greater than oneself' (see the concluding paragraph in the third section below).

Values are regularly employed to justify and legitimate norms, but it can be misleading to say of norms (making it a part of their definition) that they 'specify' values, or that they are 'derived' from them. Norms are needed to coordinate action and maintain social order particularly in circumstances where the actors involved hold different or conflicting values. To perform this function (especially in highly differentiated societies), there must be available a repertoire of (mainly legal) norms that are neutral with regard to any set of potentially conflictive particularistic values (see the long-standing discussions about the nature of 'positive' law and the relationship between legitimacy and legality). The construal of more highly abstracted values that transcend the apparent discrepancies between more concretely specified values may also help to prevent or mitigate conflict. Individuals who are in disagreement with norms that are imposed upon them might still decide to obey them, not only for fear of sanctions but also because they have made their compliance to legitimately established common or legal norms, a higher-order norm (a principle) that they are willing to adhere to in order to serve the (universally) valued goal of peaceful coordination of

action, whatever conflicts might arise on the level of lower-order values. When norms are highly abstracted and formulated as principles, the analytical difference to values might get blurred. These constellations have also triggered much theoretical debate among philosophers and social scientists seeking to clarify the priority relationship between the 'right' and the 'good.'

Developmental Perspectives: Adopting, Generating, and Changing Values

The adoption and reconstruction of values on the personal level and in different stages of life have mainly been conceptualized and studied as a major element in processes of socialization shaped by the given and further evolving (sub-) cultural systems and by their respective location within the differentiated structure of social systems. These processes have been theorized in widely differing ways by psychologists and social scientists. In this section, I will concentrate on core concepts of the cognitive developmental approach, which focuses on moral norms and values and has its major roots in the work of Kohlberg. This approach is particularly interesting because it offers hypotheses and empirical insights not only on personal development, but also relates these processes to certain features of large-scale societal structures that have been evolving in recent history. At the end of this section, I will briefly point out a rather different approach, which conceives of identity-constitutive values as arising out of emotion-laden 'experiences of self-transcendence.'

Kohlberg (1981, 1984) describes individual moral development as a sequence of three major (and further subdivided) stages characterized by rising levels of (cognitive) understanding and different types of conformity motives. At the 'preconventional level,' norms are taken to be valid because they are (re-)presented by authorities; they are followed in order to avoid punishment or receive rewards. At the 'conventional level,' moral values and the norms associated with them are deemed valid because they are propagated in one's group or society; they are primarily accepted and obeyed in order to gain social acceptance and to avert guilt feelings. At the 'postconventional level,' reached by a smaller portion of people, individuals ascribe superior validity to universal moral principles (centered around the concepts of justice and fairness and equal basic worth of all human beings), which they respect and voluntarily submit to on the basis of reasoned insight in their legitimacy. The major factor propelling forward this movement from one stage to the next, Kohlberg (taking the lead from Piaget) saw in the developmental 'logic' inherent in the ontogenetic advancement of cognitive capacities from early childhood into adolescence. He also assumed a gradually increasing consistency between the cognitive and the motivational dimension of moral development, between judgments (concerning the rightness of moral norms and the validity of moral values) and the willingness (motivation) to act in accordance with them.

More recent research has led to several modifications and additional analytic distinctions. A major correction achieved herein was to differentiate and disentangle the cognitive and the motivational dimension – moral understanding and moral motivation are now seen as independent dimensions (Colby and Damon, 1995; Nunner-Winkler, 2004: pp. 306f.). Each level of cognitive moral understanding can be associated with different types and varied strength of moral motivation, though some (correlational) affinity between these dimensions seems to persist. A high level of moral understanding might be tied up with a high level of moral motivation: acting morally may become a central part of a person's self-identity; but it can also be associated with a low level of motivation or with an inclination to make use of 'neutralization techniques' and cognitive distortions, which pave the way to moral disengagement (Bandura et al., 1996). There is also some debate on whether or not or in which way such a 'minimalized' set of highly abstracted and universally valid moral principles (most essentially: not to inflict harm on other people) can be or in fact are supplemented with more concretely prescribed moral rules associated with broader sets of values concerning the 'good life' as perceived and construed within particular communities of varied size and social location (see, e.g., Joas, 2000).

Nunner-Winkler (2004) registers a large-scale historical trend in societal development that has strengthened the advancement toward higher levels of moral reasoning reached by a growing number of individual persons. Her starting point is Elias' claim that there is a coevolution of societal and personality structure (Elias, 1994). In these processes, chains of interdependence among increasing numbers of interacting persons, institutions, and administrative units were elongated, thus calling for ever more individual insight and considerateness, i.e., self-control. Consequently, 'outside constraints' (level 1 in Kohlberg's developmental theory) were more and more replaced by internalized 'self-constraints' (level 2). Many other theoretical models for analyzing long-term societal developments could be introduced at this point, e.g., Durkheim's ideas on the erosion of collectivism; Weber's propositions concerning the increasing and multifaceted import of processes of rationalization; Riesman's theses on the shifting weights of tradition-, inner-, and outer-directed orientations as characteristic features of evolving cultures and personality structures; more recent work by Wouters (1986) has emphasized the increasing importance of processes of 'informalization.' Most theorists are agreed upon assuming that societies in the long run have become less hierarchically organized and more differentiated into interdependent subsystems performing specialized functions more or less autonomously and reaching out beyond national boundaries. These processes – including the dismantling of hierarchical structures within the family – have made individuals become increasingly more autonomous: free, but also forced, to make their own choices, free to enter cooperative relationships but also pushed to develop their competitive capacities. According to Nunner-Winkler (and others), the sociohistorical changes in societal structure and in socialization styles have, on the level of individual personalities, promoted the transition from the second to the third level of Kohlberg's scheme of moral development: "a change from super-ego controls and habitualized conformity dispositions towards more ego-syntonic types of conformity motivations" (Nunner-Winkler, 2004: pp. 309f.). But she also warns that "there might be a rising risk that the increasing availability of ego-syntonic

conformity motivation might go along with the dismantling of previously reliable barriers against amorality" (Nunner-Winkler, 2000: p. 239). The structurally induced release from moral bind has been a central topic in theories of anomie, with a broad range of empirical evidence particularly produced in studies of crime development (see, e.g., Messner et al., 2008).

As indicated above, cognitive developmental theories assume that on the highest level of personal moral development high-conformity motives, i.e., the willingness to act in accordance with moral values and norms considered universally valid, emanates directly from one's reasoned insight into what is 'right.' Other scholars have argued that rational reflections as such cannot produce motivation, that there is no motivation without the press of emotions. Joas (2000), drawing on Ch. Taylor and others, expounds the idea that motivating conceptions of the desirable are anchored in and ultimately emerge from emotion-laden experiences of self-transcendence, in which a person is captivated by something that carries her beyond her personal self, without losing it. Such (elevating, but in other cases dejecting) experiences are evoked when smoothly proceeding routine activities of everyday life get interrupted or turned into a new and unexpected direction, thus calling into question previous paths of behavior and hitherto held beliefs and interpretations. Examples are religious revelations, falling in love, loosing one's job, getting intensely praised or humiliated, having a friend dying in an accident, being suddenly confronted with the misery of other people. The elevating or disturbing feelings elicited in such situations call for articulation and narration in order to find new interpretations that clarify, partially revise and partially confirm, one's own standing on what is good or bad (values) and how this is related to what is right or wrong (in terms of norms). In this perspective, the adequacy of norms can be judged rationally (with regard to their efficiency in coordinating social action), the validity of values, however, is comprehended in the interplay of reflections and emotions arising from elevating or disturbing experiences that press for articulation and interpretation (for a challenging view on the role of emotions in grounding moral values see De Sousa, 2008).

Empirical Research on Values and Value Change

There is a long tradition of empirical studies on values and value change, in which a large array of strategies and techniques has been used to collect information and interpret data from all sorts of written text, from participant and nonparticipant observation, from face-to-face interviews or questionnaire responses. A brief overview of empirical value studies of various types produced till the early 1980s is offered by Spates (1983). To study values in national and cross-national perspective, large-scale survey studies based on representative samples of whole populations (or of specific subpopulations like highly qualified employees of large corporations, school teachers, or university students) and repeated overtime have become the prime research instrument. Since the 1970s, sociological discussions on value change have been increasingly dominated by the work of Inglehart (e.g., 1977, 1997) and the *European* and *World Values Surveys* he has initiated. Since the mid-1990s, another approach on cross-cultural value research developed by S.H. Schwartz, has received increasing attention as well. What follows here is a brief outline of the major concepts and empirical findings advanced in both of these research groups (with respect to Inglehart, I draw upon a more comprehensive review in Thome, 2008). I also want at least to mention the cross-cultural studies by Hofstede (1980, 2001) and by Triandis (1995) as well as the values research based on experimental gaming strategies applied in a laboratory setting (see, e.g., Wilke et al., 1986), which I cannot review here due to space limitations (for an overview on these approaches see Karp, 2000).

Inglehart is best known for his interpretation of value change (since the late 1960s) as a shift in preferences away from 'materialist' to 'postmaterialist' values. His initial hypotheses were derived from Maslow's theses on need hierarchies (Maslow, 1954) stating that "people act to fulfill a number of different needs, which are pursued in hierarchical order, according to their relative urgency for survival" (Inglehart, 1997: p. 22). On this view, the physical need for sustenance and freedom from bodily harm takes precedence over all else. Only to the extent that these are satisfied do we, according to these assumptions, turn our attention to the social need to belong and to be respected, and finally to the personal need for self-actualization. With respect to political objectives (Inglehart's primary value indicators), basic needs find expression, for example, in the pursuit of economic stability and continuous growth; the need for security requires public order and protection from criminality and (seemingly) hostile countries; the need to belong and to be given respect motivates calls for participation in the workplace and in politics; self-actualization requires legally enshrined freedoms and opportunities to arrange our lives as we see fit. Inglehart's idea was to ask people how urgent they considered political goals he presented to them via questionnaire items classified in this way as indicating a stronger or weaker leaning toward the materialist or postmaterialist pole of such value orientations. The item batteries that Inglehart had designed to measure value orientations via a rank ordering of political goals were included in the *Eurobarometer Surveys* (EBS) conducted on behalf of the European Commission since 1973 twice a year (the data are available as an online trend file, see references below, which also includes data from three earlier European surveys as well). In addition to the EBS, Inglehart et al. initiated more comprehensive *European Values Studies* and *World Values Surveys* starting in 1981, now in its sixth wave scheduled to be concluded in March 2014. These surveys have by now covered about hundred countries in all continents.

The European data till (roughly) 1990 confirmed, by and large, Inglehart's prognoses concerning the shift toward postmaterialism in central European countries. Since then the trends diverged cross nationally, and despite further economic growth postmaterialism declined in several countries. There was (and still is) considerable debate not only on Inglehart's way of theorizing, but also on the adequacy of his instruments of measurement (see, e.g., Clarke, 2000; Flanagan, 1987; Vinken, 2006), which have become increasingly questionable as the surveys were expanded

to include non-European countries as well. Inglehart himself has extended his system for categorizing relevant value orientations. The dimension of materialism vs postmaterialism has been embedded in the more broadly conceived dimension of 'survival vs self-expression,' which includes additional measurements of subjective well-being, trust in other people, political engagement, and tolerance for nonconformist lifestyles. The ensuing research has made increasingly clear that there has not been a departure from materialist values but rather an increasing emphasis on combining economic prosperity with libertarian, antiauthoritarian value orientations (Welzel, 2009). Inglehart has also constructed a second analytical dimension, which contrasts 'traditional' and 'secular-rational' orientations (Inglehart and Baker, 2000; Inglehart and Welzel, 2010). Key points here are belief in God and religious commitments, respect for authorities, regard for one's own nation and the family. Inglehart acknowledges in principle that these two dimensions may be combined in different ways in different countries. While a continuous increase in economic prosperity promotes both secularization and self-expression values, this development is 'path dependent,' considerably determined by cultural traditions. Nevertheless, for the long run he sees both dimensions converge into a two-stage process of 'modernization': Traditional values are replaced (or dominated) by secular-rational values in the process of industrialization; self-expression values become predominant during the transition from industrial to postindustrial (postmodern) societies. More recently, this approach has been further expanded and modified into a 'human development model' (Welzel and Inglehart, 2010).

Inglehart's ordering of countries from all continents along these two analytical dimensions is, in its overall configuration, pretty much what one might expect. European countries with large portions of Protestants load high on both analytical dimensions, those with large portions of Catholics are located around the center, whereas South-Asian and Islamic countries rank low on both dimensions. However, it appears highly problematic (in some cases at least, e.g., when inquiring religious beliefs) to use identical questionnaire items in very different cultural regions. More generally, it is highly problematic to use questionnaire items (as Inglehart does), which address different categories of values on rather different levels of analytical abstraction and specificity. Also, the rankings of a set of value items (particularly if specified as political goals) might be strongly influenced not just by the degree of desirability ascribed to them, but also by the degree to which respondents consider them to be at risk or rather fulfilled under given circumstances. These assessments should therefore also be explicitly measured in these surveys and statistically controlled for when evaluating priority rankings of values (cf Thome, 2003).

A theoretically and methodologically more elaborated system of value categories and survey instruments applied in cross-cultural studies has been worked out by the social psychologist S. H. Schwartz (drawing on Hofstede, 1980). He has proposed a system of 10 basic value types, each specified by individual-level motivational emphasis and translated into a set of more than 50 personal goals (Schwartz, 1994): power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, security. According to Schwartz, these values represent three 'universal requirements of human existence,' which he does not rank in advance: needs of individuals as biological organisms, coordinated social interaction, survival, and welfare needs of groups. Surveys conducted in a set of 66 nations proved 45 of the 56 (occasionally 57) value items to have similar meanings across nations (Schwartz, 2004). The 10 basic values are grouped along two higher-order bipolar dimensions: the first contrasts 'openness to change' (seeking stimulation and self-direction) with 'conservation' values (seeking security and conformity, upholding tradition), the second opposes 'self-transcendence' (orientation toward universalism and benevolence) with 'self-enhancement' (seeking achievement and power); the tenth value, hedonism, cuts across the borderlines between self-enhancement and openeness to change. Thus, the 10 basic values can be ordered to be more congruent or more opposing to each other, rather close or rather remote in their meaning. Schwartz (1994: p. 97) sees the cultural and the individual level of values conceptually related, but he also points to specific cultural differences in dealing with 'critical issues that confront all societies.' He classifies these into three bipolar dimensions referring to the relative weight given to (a): 'embeddedness' vs affective or intellectual 'autonomy' (of individuals in relation to social groups), (b) 'hierarchy' vs 'egalitarianism' (with respect to social status and power distribution), and (c) 'harmony' vs 'mastery' (striving for peace and for unity with nature, or for self-assertion in attaining group or personal goals) (Schwartz, 2004: pp. 45f.). This scheme thus comprises seven categories of opposed or supplementary cultural orientations. Schwartz sees the validity of this scheme confirmed by empirical analyses based on data collected on those 45 value items that were found to have similar meanings across nations and cultures. For 67 nations and cultural groups (represented by teacher and student samples), the mean scores for each of these items were registered and pairwise correlated. Multidimensional scaling analyses were then performed to map the cultural differences between nations. In this way Schwartz identified seven transnational cultural groupings, which show 'striking parallels' with the clusters found by Inglehart (based on different analytical dimensions). So, for example, West European countries load high on autonomy, egalitarianism, and harmony, whereas Confucian countries lean toward the opposite pole on all of these dimensions: embeddedness, hierarchy, and mastery.

Schwartz et al. (2001) also developed a second measurement instrument, the 40-item 'Portrait Values Questionnaire' (PVQ), which demands lower levels of abstract thinking so that it can be used in large population surveys not limited (like the original instrument) to samples of highly educated people. This instrument, too, has been validated across many cultures. To be included in the series of *European Social Surveys* (ESS) starting (with 2 years interval) in 2002/03, a shortened, 21-items version of the PVQ has also been set up. In the meantime it has been shown, however, that this instrument

does not represent the full range of the 10 value types identified before (Meuleman et al., 2012). A thoroughly conducted comparative assessment of the validity of both the Inglehart and the Schwartz value theories is offered by Datler et al. (2013).

See also: Anomie; Collective Beliefs, Sociological Explanation of; Culture, Sociology of; Modernization, Sociological Theories of; Norms; Public Opinion: Social Attitudes.

Bibliography

Ball-Rokeach, S.J., Loges, W.E., 1992. Value theory and research. In: Borgatta, E.F., Borgatta, M.L. (Eds.), Encyclopedia of Sociology, vol. 4. Macmillan, New York, pp. 2222–2228.

Bandura, A., Barbaranelli, C., Caprara, G.V., Patorelle, C., 1996. Mechanisms of moral disengagement in the exercise of moral agency. Journal of Personality and Social Psychology 71, 364–374.

Clarke, H.D., 2000. Refutations affirmed: conversations concerning the Eurobarometer values battery. Political Research Quarterly 53, 477–494.

Colby, A., Damon, W., 1995. The development of extraordinary moral commitment. In: Killen, M., Hart, D. (Eds.), Morality in Everyday Life. Developmental Perspectives. Cambridge University Press, New York, pp. 342–370.

Datler, G., Jagodzinski, W., Schmidt, P., 2013. Two theories on the test bench: internal and external validity of the theories of Ronald Inglehart and Shalom Schwartz. Social Science Research 42, 906–925.

De Sousa, R., 2008. Really, what else is there? Emotion, values, and morality. Critical Quarterly 50, 12–23.

Elias, N., 1994. The Civilizing Process. Blackwell, London (German orig. 1939).

Flanagan, S.C., 1987. Values in industrial societies. American Political Science Review 81, 1303–1319.

Gerth, H., Mills, C.W., 1953. Character and Social Structure. The Psychology of Social Institutions. Harcourt, Brace & World, Inc., New York.

Haidt, J., 2012. The Righteous Mind. Why Good People are Divided by Politics and Religion. Penguin, London.

Hofstede, G., 1980. Culture's Consequences: International Differences in Work-Related Values. Sage, Beverly Hills.

Hofstede, G., 2001. Culture's Consequences. Comparing Values, Behaviors, Institutions, and Organizations across Nations. Sage, Beverly Hills.

Inglehart, R., 1977. The Silent Revolution. Changing Values and Political Styles in Advance Industrial Society. Princeton University Press, Princeton, NJ.

Inglehart, R., 1997. Modernization and Postmodernization. Cultural, Economic, and Political Chance in 43 Societies. Princeton University Press, Princeton, NJ.

Inglehart, R., Baker, W., 2000. Modernization, cultural change, and the persistence of traditional values. American Sociological Review 65, 19–51.

Inglehart, R., Welzel, C., 2010. Changing mass priorities: the link between modernization and democracy. Perspectives on Politics 8, 551–567.

Joas, H., 2000. The Genesis of Values. University of Chicago Press, Chicago, IL (German original 1997).

Karp, D.R., 2000. Values theory and research. In: Borgatta, E.F., Montgomery, R.J.V. (Eds.), Encyclopedia of Sociology, vol. 5. Macmillan Reference USA, New York, pp. 3212–3227.

Kluckhohn, C., 1951. Values and value-orientations in the theory of action. An exploration in definition and classification. In: Parsons, T., Shils, E.A. (Eds.), Toward a General Theory of Action, second ed. Harvard University Press, Harvard, pp. 388–464.

Kohlberg, L., 1981. Essays on Moral Development. In: The Philosophy of Moral Development. Moral Stages and the Idea of Justice, vol. 1. Harper & Row, San Francisco.

Kohlberg, L., 1984. Essays on Moral Development. In: The Psychology of Moral Development. The Nature and Validity of Moral Stages, vol. 2. Harper & Row, San Francisco.

Luhmann, N., 1996. The sociology of the moral and ethics. International Sociology 11, 27–36.

Luhmann, N., 1998. Die Gesellschaft der Gesellschaft. Suhrkamp, Frankfurt am.

Maslow, A.H., 1954. Motivation and Personality. Harper and Row, New York.

Messner, S.F., Thome, H., Rosenfeld, R., 2008. Institutions, anomie, and violent crime. Clarifying and elaborating institutional-anomie theory. International Journal of Conflict and Violence 2, 163–181.

Meuleman, B., Davidov, D., Schmidt, P., Billiet, J., 2012. Social location and value priorities. A European-wide comparison of the relation between social-structural variables and human values. In: Gabriel, O., Keil, S.I. (Eds.), Society and Democracy in Europe. Routledge, New York, pp. 45–67.

Nunner-Winkler, G., 2000. Von Selbstzwängen zur Selbstbindung (und Nutzenkalkülen). In: Endreß, M., Roughley, N. (Eds.), Anthropologie und Moral. Philosophische und soziologische Perspektiven. Königshausen & Neumann, Würzburg, pp. 211–243.

Nunner-Winkler, G., 2004. Sociohistoric changes in the structure of moral motivation. In: Lapsley, D.K., Narvaez, D. (Eds.), Moral Development, Self, and Identity. Lawrence Erlbaum Ass, Mahwah, NJ, pp. 299–334.

Parsons, T., 1968. The Structure of Social Action. Free Press, New York (Orig. publ. 1937).

Parsons, T., 1990. Prolegomena to a theory of social institutions. American Sociological Review 55, 319–333.

Parsons, T., Shils, E.A., 1951. Values, motives, and systems of action. In: Parsons, T., Shils, E.A. (Eds.), Toward a General Theory of Action. Harvard University Press, Harvard, pp. 47–243.

Rokeach, M., 1973. The Nature of Human Values. The Free Press, New York.

Schwartz, S.H., 1994. Beyond individualism/collectivism. New cultural dimensions of values. In: Kim, U., Triandis, H.C., Kâğitçibaşi, Ç., Choi, S.-C., Yoon, G. (Eds.), Individualism and Collectivism. Theory, Method, and Applications. Sage, Thousand Oaks, pp. 85–119.

Schwartz, S.H., Melech, G., Lehmann, A., Burgess, S., Harris, M., Owens, V., 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. Journal of Cross-Cultural Psychology 32, 519–542.

Schwartz, S.H., 2004. Mapping and interpreting cultural differences around the world. In: Vinken, H., Seoters, J., Ester, P. (Eds.), Comparing Cultures. Dimensions of Culture in a Comparative Perspective. Koninklijke Brill, Leiden, Netherlands, pp. 43–73.

Spates, J.L., 1983. The sociology of values. Annual Review of Sociology 9, 27–49.

Thomas, W.I., Znaniecki, F., 1958. The Polish Peasant in Europe and America, second ed., vol. 1. Dover Publ, New York (originally published in 5 Volumes 1918–1920).

Thome, H., 2003. Soziologische Wertforschung. Ein von Niklas Luhmann inspirierter Vorschlag für die engere Verknüpfung von Theorie und Empirie. Zeitschrift für Soziologie 32, 4–28.

Thome, H., 2008. Value change in Europe from the perspective of empirical social research. In: Joas, H., Wiegandt, K. (Eds.), The Cultural Values of Europe. Liverpool University Press, Liverpool, pp. 277–319.

Trandis, H.C., 1995. Individualism and Collectivism. Westview, Boulder, CO.

Turner, R.H., 1968. The self-conception in social interaction. In: Gordon, C., Gergen, K.J. (Eds.), The Self in Social Interaction. Wiley, New York.

Vinken, H., 2006. East Asian Values Surveys. Making a Case for East Asian-Origin Values Survey Concepts. Working paper no. 5 in 2006: ZUMA-Arbeitsbericht Nr. 2006/05. Mannheim.

Welzel, C., 2009. Werte- und Wertewandelforschung. In: Kaina, V., Römmele, A. (Eds.), Politische Soziologie. VS Verlag für Sozialwissenschaften, Wiesbaden, pp. 109–139.

Welzel, C., Inglehart, R., 2010. Agency, values, and well-being: a human development model. Social Indicators Research 97, 43–63.

Wilke, H.A.M., Messick, D.M., Rutte, C.G. (Eds.), 1986. Experimental Social Dilemmas. Peter Lang, Frankfurt.

Wouters, C., 1986. Formalization and informalization: changing tension balances in civilizing process, theory. Culture & Society 3, 1–18.

Relevant Websites

http://ec.europa.eu/public-opinion – Eurobarometer.
http://www.europeansocialsurvey.org – European Social Survey.
http://europeanvaluesstudy.eu – European Values Study.
http://worldvaluessurvey.org – World Values Study.
(They all refer to the last section: Empirical Research on Values and Value Change).

Veblen, Thorstein (1857–1929)

Thomas Hall, University of Central Florida, Orlando, FL, USA
Andreas Hess, University College Dublin, Dublin, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Thorstein Veblen witnessed significant changes in the American political, economic, and cultural landscape. Industrialization replaced agriculture as the primary source of income for the American economy. Urban population centers flourished, and across the country, politicians debated social justice initiatives. The writings of the 'muckrakers' informed and mobilized the general public to action regarding unfair business practices and corrupt political elites. Veblen also witnessed the rise of the corporation and spread of unprecedented economic disparities.

The Theory of the Leisure Class was Thorstein Veblen's most popular work. He was a prolific writer in both economics and sociology. To better understand his writings, Veblen must be located in history and the traditions of his community. In his lifetime, the United States fought three wars: the Civil War, the Spanish–American War, and World War I. Veblen lived during a time when Charles Darwin's theory of the origin of species was groundbreaking. Herbert Spencer's application of natural selection or 'survival of the fittest' to explain social stratification supported the viewpoints of neoclassical economists. These apologists of business interests advised against any regulation or interference with the 'hidden hand of the market.' In addition, Spencer's schema justified a lack of humanity or compassion for those who did not possess wealth. The shift in America's economic engine from agriculture to manufacturing paved the way for European immigration required to meet manufacturing demands. Urbanization weakened individuals' ties to their faith, family, and community. A new social order was emerging that would have a profound effect on several generations that followed.

Veblen's contemporaries include the 'fathers' of American pragmatism – Charles Peirce, William James, and John Dewey. Their ideas considerably influenced Veblen's convictions regarding social justice and humanistic philosophy. William Sumner also had an impact on Veblen's theory of the evolution of social systems. Though Veblen never supported Marxism, because of his radical stances, he is often referred to as the 'American' Marx (Mitchell, 2001).

Family Background and Early Life

Thorstein Veblen was born on 30 July 1857 in Manitowoc County, Wisconsin. His parents, Thomas and Kari Veblen, immigrated to America from Norway in 1847. They eventually saved enough money to purchase a 290-acre farm in Minnesota that became the family's permanent homestead. Thorstein Veblen took after his father, a respected leader of a small farming community. He respected his father's views on social order and injustice, but unlike his father, Veblen was precocious; he tested the limits of prescribed behavior in his austere Norwegian community. When Veblen had reached a mature age, his father arranged for him to attend Carleton College. Thorstein was never consulted on the matter, and perhaps this is one of the reasons why he struggled with accepting the doctrines and Christian values of the College.

After graduating, Veblen briefly attended Johns Hopkins University. There he met Charles Peirce, who first introduced him to American pragmatist ideas. After a brief stint at Johns Hopkins, Veblen moved to Yale to study under the tutelage of William Sumner, a scientist who embraced Darwinism and the views of the British sociologist Herbert Spencer. Through Sumner, Herbert Spencer's ideas took on a form that was to be important and influential in Veblen's work (Hess, 2001).

Veblen received a PhD degree in philosophy from Yale in 1884. However, he was subsequently unable to find a faculty position, perhaps because of his Norwegian descent or his unwillingness to adhere to the teachings of the Christian faith. When he returned to his parents' community in Minnesota, Veblen reported that he was ill and needed time to recuperate. During his recuperation, he devoted his time to reading and studying botany. He also found time to court Ellen Rolfe, whom he married in 1888.

In 1891, Veblen enrolled as a graduate student in economics at Cornell. He received a graduate fellowship to study with Laurence Laughlin, and in 1892, he followed Laughlin to the University of Chicago. At Chicago, he was introduced to the works of William James and John Dewey; James' theory on 'habit' and 'instinct' would later come to influence many of Veblen's published works.

After rumors of marital infidelity, in 1906, Veblen decided to move to Stanford University in California, where he had been offered an appointment as associate professor in economics. However, due to his unpopularity as a teacher and dissatisfaction with the institution, he resigned the post in 1910. In 1911, he accepted an appointment at the University of Missouri–Columbia but soon resigned his position there to move to the east coast to work for the Food Administration in Washington, DC. Veblen found working in a government bureaucracy vexing and left the agency in 1918. In the fall of the same year, he joined the editorial staff of *The Dial*, a radical journal of literary and political opinion located in New York City. He also started to collaborate with Charles Beard, James Harvey Robinson, and John Dewey to establish the New School for Social Research. After his retirement in 1926, Veblen moved

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61309-0

back to California, where he lived in a cabin near Stanford University. He died there in 1929.

The Gilded Age

The term Gilded Age was coined by Mark Twain and Charles Dudley Warner in their book, *The Gilded Age: A Tale of Today* (1873). The book satirized the rising inequality between economic classes and the apparent lack of interest among government officials concerning the consequences of their laissez-faire policies. By the turn of the nineteenth century, America's industrial production outpaced England, Germany, and France combined. The result of this growth was a bifurcation of income and social class. A few large corporations dominated the steel, oil, sugar, meat, and farm machinery industries. Because the number of corporations was small, owners colluded with one another to create a supply chain that controlled prices and limited competition. These men became known as 'robber barons.' Their power ultimately led to widespread corruption and eventually loss of public trust in a representative democracy.

The loss of public faith in American representative democracy developed during hard economic times. The first post-Civil War depression occurred in the 1870s. During this time, a grasshopper plague devastated farm income in the mid-Western states. Poverty was widespread and communities struggled to survive. Despite the dire economic times, the attitude of railway and eastern promoters toward farmers was 'charge all the traffic can bear' and 'the public be dammed.' In 1893, another economic depression hit, which further contributed to undermining public faith in elected representatives. This time the depression would last 4 years, and unlike the grasshopper plague of the 1870s, most businesses, banks, agricultural interests, and factories were affected. Millions became unemployed and were unable to provide for their families. Criticism of the ruthless practices of the robber barons reached a fever pitch. However, the corporations redoubled their efforts to resist change. Corporate 'thugs' were employed to intimidate and even murder striking workers. Many federal and state governments became 'puppets' of these absentee owners. No wonder that public sentiment turned against absentee owners and elected officials.

Wealth Is Waste: the Rise of Class Consciousness in America

In 1899, Veblen published *The Theory of the Leisure Class*. The study was a critique of economic stratification in postindustrial America. It focused on the outrageous waste of the so-called 'pecuniary culture.' Veblen had first hoped that society would evolve in a peaceable direction; it seemed that the 'instinct of workmanship' guided social development away from social atomism toward the collective good. However, the rise and the new power of business interests of the late nineteenth and early twentieth centuries soon rendered such hopes utopian: social stratification intensified. A new leisure class began to set the standards for desirable goods. Their norms regarding consumption turned out to be the norms to which all other classes aspired. Inexpensive yet functional goods were labeled as 'cheap' and 'unsuitable.' Instead, the consumption of useless goods became a litmus test for social class and status. For Veblen, "so thoroughly has the habit of approving the expensive and disapproving the inexpensive been ingrained into our thinking that we instinctively insist upon at least some measure of wasteful expensiveness in all our consumption, even in the case of goods which are consumed in strict privacy and without the slightest thought of display" (Veblen, 2007: 103f).

Veblen suggested that the privilege of the leisure class had evolved from "primitive savagery to barbarism… from a peaceable to a consistently warlike habit of life" (Veblen, 2007: 11). Class distinctions arose when individuals with desirable traits were rewarded and those with nondesirable traits became subservient to others. Those men rewarded for their warlike traits were held in high esteem; they were imbued with status, power, and leisure time. In modern times, the leisure class included the industrial upper class, athletes, politicians, military officers, and the clergy. For this new class, "time is consumed non-productively from a sense of the unworthiness of productive work, and as an evidence of the pecuniary ability to afford a life of idleness" (Veblen, 2007: 32).

The privileges of the leisure class were similar to those afforded to warriors of primitive cultures. Norms for men in the leisure class were granted permission to engage in behaviors that were taboo for other classes. Men celebrated public drunkenness, the use of stimulants, and promiscuity. Women, whom these men regarded as their property, had to behave along prescribed norms for 'proper ladies.' These men had little care for those women within their social strata, much less for those 'undesirables' on the outside. Women and children as well as other possessions were chattel used to promote the good 'character' of the genteel, well-bred man. Veblen spoke avidly against the crass public demonstrations of waste celebrated by these men and warned against the new conspicuous consumption, which he saw as a return to primitive, barbaric practices.

Critique of Neoclassical Economic Theory

Herbert Spencer's social Darwinism had a major impact on neoclassical economic theory in the Gilded Era. The collusion of a relatively small group of men shaped most decisions regarding governmental regulations on economic activities. As a result, it was not big business that carried the burden of taxation but farmers and other subaltern classes. This, in turn, paved the way for the creation of monopolies that limited competition and distribution of goods and services. Neoclassical economics did not support institutional or governmental interference with market economies because intervention was believed to inhibit market exchanges. According to the theory, it is the individual actor who determines the degree to which wealth is obtained. Those who become wealthy do so by hard work and frugality, while those who are poor are unmotivated or unwilling to work and save.

Veblen was critical of neoclassical economics theory. Veblen viewed economics as an evolutionary science. He offered an alternative hypothesis based on social evolution, accelerated by the spirit of workmanship, and led by peaceable men. His

schema of evolutionary economic growth was influenced by the belief that instincts are not conscious, but are conditioned by and activated through habits. In line with pragmatist reasoning Veblen argued that the nexus of change lies between the individual and macro-social institutions: "every change so made is necessarily made by individuals immersed in the community and exposed to the discipline of group life as it runs in the community, since all life is necessarily group life" (Veblen, 2007: 103f). Accordingly, "social wealth is the stock of human norms, mores and structures of truthfulness" (Veblen, 2007: 173).

Time and again, Veblen critically referred to emulation and predation as a means of creating a real or imagined status based on an outward demonstration of wealth and strength. In contrast to emulation and predation, the instinct of workmanship is an innate drive that seeks to increase productivity and efficiency both for the individual and for the betterment of the community. For example, Veblen used the phrase 'parental bent' to explain altruism, or the tendency to extend one's self to provide for future generations. He also used the term 'idle curiosity' to describe creativity or taking time to think about innovation, as opposed to accepting well-worn 'solutions.' In his day, the colloquialism 'idle hands are the work of the devil' was juxtaposed to his description of 'idle curiosity.'

The Machine Process: Social Rationality in Postindustrial Economies

The corporation, the joint-stock company, and the absentee owner destabilized the relationship between the captain of industry and the community. The working class bore the consequence of the shift. For absentee owners and corporations, there were not any economic incentives for craftsmen to oversee the production of serviceable goods at a reasonable cost. The pecuniary culture supported the production of goods at the lowest cost. Unskilled workers replaced craftsmen, and labor was a commodity bought and sold without concern for societal costs.

In *The Engineers and the Price System*, Veblen argues that the instinct of workmanship will neutralize the predatory instinct of pecuniary culture. Veblen envisioned that workers would eventually evolve to become a force that mediated the influence of business interests (Veblen, 1975). He believed that industrialization would lead to the development of rational worker actors motivated by creative instincts rather than to mindless machine tenders. Industry, as it then existed, was not capable of generating creativity. It was a stifling matter-of-fact condition that reduced workers to parts and pieces of an impersonal assembly line. The machine industry led to 'trained incapacity' or what Dewey described as 'occupational psychosis.' But industrialization could also be a progressive force that would result in an alternative to repetitive work that required little thought or imagination. Innovation would lead to increased efficiency and replace the predatory instincts of business with a community-oriented, creative instinct of workmanship. If inexpensive goods were available to the population, the waste created from conspicuous consumption would lose its appeal and the profitability of the business interests would return to communities as opposed to absentee owners.

Despite such hopes, toward the end of his life Veblen became more and more pessimistic about the promise of evolutionary social change. His hope for the emergence of rational actors, increased innovation, and efficiency gave way to less idealistic observations. As he aged, he was increasingly concerned about the instinct of predatory, coercive barbarism increasing in strength. The engineers who once held promise in Veblen's blueprint for production over profit had now become allies of the class-conscious business interests and distanced themselves from unskilled workers. A hierarchical order of middle-class 'white collar' and lower-class 'blue collar' workers came to be accepted.

The background to Veblen's skepticism lay in World War I, which unified the country through the zeal of patriotism. The post-War years were a celebration of economic prosperity and conspicuous consumption. This was the era of the 'flappers,' a time when new norms for women were recognized. Politicians were bribed and absentee owners were concealed in their feudal kingdoms. At the same time, the upper class became oblivious of the communities it plundered. An economy built on deregulation and nonintervention was breaking under the weight of speculation and unsecured debt. The system was designed for economic growth, not stagnation. For decades, Veblen had warned of the risk of confidence men, who fixed prices and manipulated stock markets, yet few people were listening. Coincidentally, the prophet Veblen died just a few months before his warnings were realized. However, Veblen would have been surprised, first, by the way the Great Depression supplanted conspicuous consumption but also, secondly, by the amount of government regulation of industry that occurred following the Great Depression.

Veblen and "the Academy"

Thorstein Veblen's conflict with the bureaucracy within institutions of higher education may have been established by his early experiences as a student who chaffed against the moral philosophy prescribed by Carleton College or perhaps his status as an outsider in the highly esteemed institutions of his day. Veblen's scholarly works often put him at odds with the 'captains of industry,' who through philanthropy funded many of the academic institutions in which he was employed. Early in his academic career, while still serving the University of Chicago, Veblen had criticized the power of corporations and the impact they had on knowledge production and scholarship. However, his main manifesto, *The Higher Learning in America: A Memorandum on the Conduct of Universities by Business Men*, was not published until 1918.

In it, Veblen criticized the academy for neglecting its moral obligation to the community. He believed that university faculty had a duty to cultivate scientific inquiry. He contrasted the role of the university man with that of the schoolmaster: "The schoolmaster and his work may be equally, or more, valuable to the community at large – presumably more rather than less – but in so far as his chief interest is of the pedagogical sort his place is not in the university" (Veblen, 1965a: 18). Veblen objected to business and political interests giving

vocational colleges the rank of universities. He simply did not believe that the university was a place for schools of law, engineering, teaching, or home economics. He accused universities of adding fields of study that were vocations to "comfort and edify the unlearned with lyceum lectures, to dispense erudition by mail order, and to maintain some putative contact with the amateurs scholars and dilettanti beyond the pale" (Veblen, 1965a: 192).

Business interests supported the development of what Veblen described as 'quasi-universities.' He criticized the appointment of businessmen on university boards. Veblen noted that boards do not "understand or appreciate scholarship or the pursuit of knowledge" (Veblen, 1965a: 142). He criticized those unlearned businessmen who decreased the quality of academic rigor, programs, and personnel. Furthermore, Veblen warned of an academia being compromised by businessmen whose judgments were centered on convenience, as opposed to scientific curiosity. Boards of trustees led by business interests had a negative impact on academia, its programs, and personnel, with scholarship morphing from an 'avocation to that of a vocation.' Veblen warned that "modern learning is of a similarly matter-of-fact, mechanistic complexion, and that it similarly leans on statistically dispassionate tests and formulations…It is the 'dry light of science' that is always in request, and great pain is taken to exclude all color of sentimentality" (Veblen, 1965a: 7). Veblen cautioned that to "the laity a 'university' has come to mean, in the first place and indispensably, an aggregation of buildings and other improved real-estate …. while the pursuit of learning is a relatively obscure matter … the academic work is, after all, unseen" (Veblen, 1965a: 139). He asserted that massive undergraduate training, sports, and amusements, as well as credit for correspondence classes undermined the pursuit of scholarship.

The 'new' universities relaxed academic qualifications and in place of rigor 'student activities' encouraged school spirit. Athletics events were given more attention than scholarly pursuits or the provision of excellent libraries (Veblen, 1965a). In short, Veblen foresaw a time when attendance at universities would become commonplace and vocational training would become the norm, mainly because "men of affairs have taken over the direction of the pursuit of knowledge" (Veblen, 1965a: 77).

Contributions to Modern Sociology and Economics

In his lifetime, Veblen was a highly regarded, albeit misunderstood theorist in sociology and economics. Pragmatism and Darwinism influenced his economic and sociological contributions to both fields of study. His synthesis of pragmatism, Darwinism, and conflict theory continues to influence social scientists today. Veblen's conceptualization of social evolution was unlike any of his predecessors. He considered the effects of economic, social, and cultural changes on how different actors think and act. Veblen wrote that change is "driven by the human instincts of emulation, predation, workmanship, parental bent, and idle curiosity" (Veblen, 1965b).

Veblen viewed social evolution as an illogical process; social evolution was dynamic and unpredictable. However, unforeseen shifts in the social composition of a class, race, or community could create the momentum to change. A new level of functioning could alter how parts related to one another. While Veblen viewed the individual actor as a product of the social system, some powerful individuals could still exert disproportional power on the system. In the late nineteenth and early twentieth centuries, Veblen witnessed these effects and their detrimental consequences, ranging from robber barons and steel magnates to the captains of higher learning.

Veblen's insights about conspicuous consumption, market speculation, and dehumanizing conditions still ring true today. His work serves as a reminder of the risks associated with consumerism, emulation, and predation. Individuals still continue to seek status through the purchase of objects that are, after a short period of time, first discarded and then immediately replaced by new fashionable objects. This cycle is repeated with every new product. Veblen was obviously not fond of marketing and advertisement. For him "falsehood permeates business….As you look out a car window, the rocks and trees are placarded all over with lies … Most advertisements, other than mere announcements for public information, are 'intentional deceptions'" (quoted in Dorfman, 1966: 210).

Contemporary consumerism drives economic growth – and not just in America. Much of the credit crunch of the twenty-first century resembles the economic and political circumstances of Veblen's day. No wonder then that modern economists and sociologists are rediscovering Veblen. The resurgence of interest in Veblen's heterodox economic theories obviously represents an interest in hearing a different point of view, a view that dissents somewhat from standard economics. The fact that Veblen's ideas are still considered worthy of discussion proves that some of his ideas have outlasted those of the more fashionable economic theories of recent times that treated Veblen's ideas and conceptualizations as merely those of a rather eccentric character.

See also: Conspicuous consumption and distinction, history of; Consumption, Geography of; Consumption, History of; Consumption, Sociology of; Darwinism: Social; Darwinism; Leisure and Cultural Consumption: US Perspective; Leisure, Sociology of; Social Inequality in Cultural Consumption Patterns.

Bibliography

Brantlinger, Patrick, Higgins, Richard, 2008. Waste and value: Thorstein Veblen and H.G. Wells. Criticism 48 (4), 453–475.

Camic, Charles, Hodgson, Geoffrey Martin, 2011. The Essential Writings of Thorstein Veblen. Routledge, Milton Park, Abingdon, Oxon.

Cassano, Graham, 2009. Choosing our ancestors: Thorstein Veblen, radical institutionalism and sociology. Critical Sociology 35 (3), 367–377.

Coser, Lewis A., 1977[c1977]. Masters of Sociological Thought: Ideas in Historical and Social Context/ Lewis A. Coser; Under the General Editorship of Robert K. Merton. Harcourt Brace Jovanovich, New York.

Dorfman, Joseph, 1966[c1961]. Thorstein Veblen and His America, with New Appendices. Kelley, New York.

Dowd, Douglas, 2000[c2000]. Thorstein Veblen. Transaction Publishers, New Brunswick, NJ.

Dyer, Alan W., 1986. Veblen on scientific creativity: the influence of Charles Peirce. Journal of Economic Issues 20, 21–41.

Hedoin, Cyril, 2009. Weber and Veblen on the rationalization process. Journal of Economic Issues 43 (1), 167–188.

Hess, Andreas, 2001. Concepts of Social Stratification: European and American models. Oxford Palgrave, Houndmills, Basingstoke, Hampshire.

James, Spencer, 2009. The invidious Veblen and the virulent victors: the salience of the leisure class today. Michigan Sociological Review 23, 59–80.

Jo, Tae-Hee, 2013. Saving private business enterprises. American Journal of Economics and Sociology 72 (2), 447–467.

Mills, C., 2000[c2000]. The Sociological Imagination/C. Wright Mills. With a New Afterword by Todd Gitlin. Oxford University Press, Oxford [England], New York.

Mills, C., Horowitz, I., 1964[c1964]. Sociology and Pragmatism: The Higher Learning in America [By] C. Wright Mills. Edited with an Introduction By Irving Louis Horowitz. Paine-Whitman Publishers, New York.

Mitchell, Ross E., 2001. Thorstein Veblen: pioneer in environmental sociology. Organization & Environment 14 (4), 389–408.

Oncu, A., 2009. Wither business ideology: revisiting Veblen's theory of engineers as revolutionary actors. Review of Radical Political Economics 41 (2), 196–215.

Penner, Devin, 2011. Monopoly and crisis in the era of the "Giant corporation": Neo-Marxist versus radical institutionalist approaches. Science & Society 75 (2), 180–205.

Preparata, Guido G., 2013. Industrious rebels and captains of deterrence: defiance interpreted through a Veblenian reformulation of strain theory. Crime Law and Social Change 1 (60), 25–38.

Schwartz, J.M., 1990. Tracking-down the nordic spirit in Thorstein Veblen's sociology. Acta Sociologica 33 (2), 115–124.

Seybold, Peter, 2008. The struggle against corporate takeover of the university. Socialism and Democracy 22 (1), 115–125.

Tilman, Rick, 2004. Karl Mannheim, Max Weber, and the problem of social rationality in Thorstein Veblen. Journal of Economic Issues XXXVIII (1), 155–172.

Tilman, Rick, 2007. Thorstein Veblen and the Enrichment of Evolutionary Naturalism. University of Missouri Press, Columbia, MO.

Tilman, Rick, 2002. Emile Durkheim and Thorstein Veblen on epistemology, cultural lag and social order. History of the Human Sciences 15 (4), 51–70.

Twain, Mark, Warner, Charles Dudley, Budd, Louis J., 2001. The Gilded Age: A Tale of Today. Penguin Books, New York.

Veblen, Thorstein, 1930[c1919]. The Place of Science in Modern Civilisation and Other Essays/By Thorstein Veblen. Viking Press, New York.

Veblen, Thorstein, 1945[c1945]. An Inquiry into the Nature of Peace and the Terms of its Perpetuation, By Thorstein Veblen. Viking Press, New York.

Veblen, Thorstein, Mills, C. Wright, 1992. The Theory of the Leisure Class. Transaction Publishers, New Brunswick, USA.

Veblen, Thorstein, 1964a. The Vested Interests and the Common Man: The Modern Point of View and the New Order. A.M. Kelley, New York.

Veblen, Thorstein, 1964b[c1914]. The Instinct of Workmanship and the State of the Industrial Arts. Norton, New York.

Veblen, Thorstein, 1964c. Imperial Germany and the Industrial Revolution. Kelley, New York.

Veblen, Thorstein, 1964d. An Inquiry into the Nature of Peace and the Terms of its Perpetuation. A.M. Kelley, bookseller, New York.

Veblen, Thorstein, 1965a[c1918]. The Higher Learning in America: A Memorandum on the Conduct of Universities by Business Men. A.M. Kelley, New York.

Veblen, Thorstein, 1965b[c1899]. The Engineers and the Price System. A.M. Kelley, bookseller, New York.

Veblen, Thorstein, 1966[c1966]. Imperial Germany and the Industrial Revolution. University of Michigan Press, Ann Arbor.

Veblen, Thorstein, 1969[c1919]. The Vested Interests and The Common Man. (The Modern Point of View and The New Order.) Capricorn Books, New York.

Veblen, Thorstein, 1975[c1904]. The Theory of Business Enterprise/by Thorstein Veblen; 1904; With a Prefatory Note by Joseph Dorfman and A Review by James Hayden Tufts. A.M. Kelley, Clifton, NJ.

Veblen, Thorstein, 1997. Absentee Ownership: Business Enterprise in Recent Times: The Case of America. Transaction Publishers, New Brunswick.

Veblen, Thorstein, 2007[c1899]. The Theory of The Leisure Class; Edited with an Introduction and Notes by Martha Banta. Oxford University Press, New York.

Venture Capital

Martin Kenney, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Venture capitalists invest in relatively young, privately owned firms judged to have the potential for fast growth and resultant opportunities to be sold to either the public or existing firms. The first formal venture capital organizations were formed in the United States in the aftermath of World War II. Typically, venture capitalists are actively involved in advising and monitoring their portfolio firms. The dominant organizational form is the venture capital limited partnership, although there are corporations, government entities, and individual investors that also provide venture capital. Today, venture capital as an investment activity is global in scope with approximately 45 nations having some venture capital activity.

Venture capital is a specialized form of investment in relatively young, privately owned firms, judged to have the potential for fast growth. The overriding goal of venture capitalists is to reap outsize capital gains on their original investment within less than 7 years after the investment. This is achieved by selling their stake either through a trade sale or a listing on public stock markets. Venture capitalists are active investors and invariably demand and receive seats on the firm's board of directors from where they monitor their investment and assist management by providing advice, recruiting executives, and participating in important strategic decisions. In the past 50 years, a number of fastest growing US technology firms received funding from venture capitalists. While venture capital as a formal activity first emerged in the United States, in 2012, approximately 45 nations had some venture capital activity (see, for example, Kenney et al., 2008; Guler and Guillen, 2010).

Introduction

The term 'venture capital' was first used in 1939 by Lammot du Pont in Congressional Hearings (Kenney, 2011). Since then, venture capital has evolved into a specialized form of finance supporting small privately owned companies judged to have the potential for fast growth. With few exceptions, the overriding goal of venture capitalists is to reap capital gains in excess of 10 times their original investment in less than 7 years. A most strict definition of venture capital excludes buyouts, loans, and other forms of financial transactions. Venture capitalists invest in equity in the form of common stock, preferred stock, convertible debentures, and/or other financial instruments convertible into common stock when their portfolio company is sold to an acquirer or in a public equity offering. At or shortly after a liquidity event, venture capitalists realize their return as capital gains.

The Venture Capital Industry

Since the 1950s, a number of fast growing US firms, particularly in new technologies, but not confined to them, received venture capital. Examples include Amazon.com, America Online, Amgen, Apple Computer, Cisco Systems, Compaq, DEC, Facebook, Federal Express, Genentech, Google, Intel, Lotus, Netscape, Oracle, Seagate, Sun Microsystems, 3Com, Yahoo!, and countless others. In Europe and Asia, venture capitalists have funded far fewer of these universally known successes; however, there have been some important firms funded including ARM Holdings (UK), Research in Motion (Canada), and Skype (Luxembourg). These firms were central to the establishment of entirely new industries such as biotechnology, semiconductors, software, the Internet, magnetic data storage, a wide variety of computer market segments, and data communications equipment. Today, venture capital is an important institution in the US national system of innovation (Kenney, 2011). In the United States, it has long been known that venture capital investing is and has been regionally concentrated in order of importance in Silicon Valley, Boston, New York, and Southern California (Florida and Kenney, 1988; Sorenson and Stuart, 2001; Chen et al., 2010). Globally, while the United States is by far the largest market for venture capital, significant venture capital markets exist in Canada, China, Israel, Japan, Taiwan, and the United Kingdom (Jeng and Wells, 2000; Megginson, 2004; Wright et al., 2005).

With the formation of the first formal venture capital organization immediately after World War II, there was significant institutional experimentation seeking the best format to support venture investing. Since the late 1950s, there has been a secular increase in the amount of venture capital available, and the number of nations with a venture industry has increased. In 1965, venture capitalists invested less than \$100 million in new ventures, and in 1999, investment peaked at $\$51 \times 10^9$. In 2011, total investment in the United States was $\$28.4 \times 10^9$ (National Venture Capital Association, 2012).

History

For centuries, wealthy individuals have invested in risky ventures in the hopes of securing an outsize return. Such investments can be traced back at least as far as the involvement of merchants in organizing consortia to finance long-distance trading ventures. The financing for firms participating in the Industrial Revolution came from merchants and other wealthy individuals willing to back entrepreneurs building manufacturing facilities or exploiting a new technology. These informal investors are now known

as 'angels,' i.e., individuals who are not professional investors in start-ups (one of the earliest academic articles was Wetzel, 1983).

During the 1930s, in both the United States and the United Kingdom, there were discussions of lack of investment capital for small businesses. In the United Kingdom, the 1931 Macmillan Report identified a shortage of private sector funds for supporting small businesses. The report contributed to a contemporaneous debate in the United States about how to provide funds for the small businesses severely affected by the Great Depression. There was little agreement about what types of small businesses should receive funding and the proper vehicles for providing financing. Prior to World War II, there were a few wealthy families, such as the Mellons, Phipps, Rockefellers, and Whitneys, investing in young ventures. Shortly after World War II, in New York City, these families established offices and hired professional managers to discover, evaluate, and invest in small companies with superior growth prospects. These were the first family venture capital funds. Although each of them changed many of their organizational characteristics, the Rockefeller (Venrock), Whitney (J. H. Whitney), and Phipps (Bessemer Ventures) funds continued to be active in 2000 (Kenney, 2011). The other significant private sector initiative was the establishment of American Research and Development (ARD) in Boston in 1946. ARD was important, not only because of the firms it funded but also because of the Harvard Business School (HBS) faculty member, General Georges Doriot, who established it. Doriot influenced a number of HBS students who later became venture capitalists. As important, Doriot through his writing, speaking, and example defined many of the practices of venture investing. Finally, ARD's $70 000 investment in Digital Equipment Corporation in 1958 was worth $350 million in 1971 – ample proof that significant return was possible (Hsu and Kenney, 2005).

The role of the US government in the development of venture capital has been limited, with a few important exceptions. In 1958, the US government passed legislation creating the Small Business Investment Corporations (SBICs) to encourage investment in small businesses. Despite being plagued by abuse during its first decade, the SBIC program made two important contributions to the development of the industry. The first contribution was its loan guarantees that permitted participants to leverage their private funds. This made low-cost capital available to investors, a number of whom soon left the SBIC program and established important independent venture capital firms. The SBIC Act also loosened Depression-era laws forbidding banks from owning more than 5% equity in any company. By having an SBIC, the banks were permitted to own large equity stakes in their venture-type investments. Particularly in the 1960s and 1970s, banks were a significant source of venture capital. During the 1980s, the limited partnerships displaced them (Kenney, 2011).

There was another federal legislation that had a significant impact on venture capital. The most important was a series of changes in employee pension fund laws during the 1970s. The first change was dramatic tightening by the Employment Retirement Investment Security Act of 1974, which operated to nearly cut off institutional investment in venture capital. The legislation was gradually loosened from 1977 through 1982, and new pension fund monies flooded into the independent firms. The lowering of capital gains tax rates may also have encouraged venture investing (Bygrave and Timmons, 1992; Poterba, 1989; Gompers, 1994). With these important exceptions, the US government has only had a limited role in creating the industry. In other nations, governments have been far more proactive (for example, for East Asia, see Kenney et al. 2002 and Lockett and Wright, 2002; for India, see Dossani and Kenney, 2002; for China, see Ahlstrom et al., 2007; and for Israel, see Avnimelech and Teubal, 2010).

The Venture Capital Limited Partnership

Not surprisingly, the organizational forms for venture capital investing have varied through time. Today, although the independent private venture capital limited partnership is the dominant organizational form globally, it is not the only one. Financial institutions, particularly banks, continue to be active. At times, industrial and other financial corporations have had significant venture capital operations. In the United States and many other nations, subnational government entities such as states or provinces have launched venture capital operations. In some nations, there are also national government-sponsored venture capital organizations. Other organizational formats have included publicly traded closed-end funds, other publicly traded vehicles, and family funds.

By the end of the 1970s in the United States and later in many other nations, the private limited partnership became the dominant format for venture investing. In the limited partnership format, the venture capitalists raise and manage individual partnership funds, each of which is denoted by a Roman numeral in increasing order. The venture capitalists are the general partners, and investors are the limited partners. Investors include pension funds, foundations, endowments, corporations, wealthy individuals, and various other entities. The duration of a particular fund was typically between 7 and 10 years. In the initial 1–3 years the fund actively invests, then in the remaining years it harvests the investments. Successful firms have two or more funds operating simultaneously, each of which is in a different phase of a fund's life cycle.

Compensation for the general partners has two components. The first component is an annual management fee; this can range from 2 to 3% of the limited partner's investment. The fees pay the costs of operating the firm and the salaries of all the employees including the venture capitalists. The other component is the carried interest, which refers to the percentage of the capital gains the venture capitalists retain from the investments after reimbursing the initial investment of the limited partners. The carried interest varies by fund from 20 to 30% of the total returns. After initial investments are returned, the limited partners and general partners receive their prorated share of the capital gains. Returns vary by fund, with certain successful funds having experienced returns in excess of 40% compounded per annum for the life of a specific fund (Metrick and Yasuda, 2010).

Investment Process

Venture capitalists are not passive investors. Because of their equity stake, they have an interest in the firm's success, which is

defined in terms of a rapid increase in the firm's value. Invariably, one or more venture capitalists 'lead investors' join the firm's board of directors. From this position they actively monitor the firm's progress. However, they do more than monitor. They often assist and counsel management. This involvement comes in many forms and is situation dependent, but includes counseling on major strategic decisions, assisting in recruiting key personnel, and providing contacts in the legal, investment banking, and other business service communities. Finally, the investment by a prestigious venture capital firm creates the perception that the small firm will be a success, an important asset for convincing customers and suppliers that the start-up is viable and worthy of patronage.

Entrepreneurship depends on the structure of investment opportunities; in the typical venture capital investment scenario, an entrepreneur or entrepreneurial team approaches a venture capitalist with a business plan describing the firm, the management team, potential competitors, and the market. Since venture capitalists receive hundreds of business plans per annum, those recommended by individuals they know receive the greatest attention. If one or more of the venture capitalists are interested, then the entrepreneurial team is invited to present its business plan in person. Should the venture capitalist's interest continue, a due diligence process is initiated during which the entrepreneurs' references and others who might be able to speak of the candidate's qualifications, personal and technical, are contacted. If there are technological unknowns, the venture capitalist gathers information on the project's feasibility from technical experts. Prospective customers are interviewed to gauge their need or desire for the product. In other words, they gather as much information as possible about the investment and its growth potential. If these hurdles are surmounted, then the team is invited to present the plan to a meeting of all the partners. Should the partners agree, a contract detailing the terms of an offer is extended to the entrepreneurs. Upon acceptance the venture capitalist invests in the firm. With this investment the relationship with the entrepreneur changes, as the venture capitalist now is committed to the success of the firm. Venture capital investments are staged, and, if the firm meets preset milestones, then it receives new investments at a higher valuation. Should the firm not meet its targets or the market changes, at the next stage the firm's valuation can decrease. In some cases, the venture capitalists, if they own more than 50% of the firm and are dissatisfied with the management's performance, will replace the management including, if necessary, the firm's founder(s).

Usually, venture capitalists syndicate their investments with other venture capitalists, thereby spreading risk, providing multiple contacts and sources of advice for the start-up firm, and cross-checking on the investment decision (Gompers and Lerner, 1999). Firms attractive to venture capitalists have the following characteristics. First, the firm is entering a market that is potentially large and promises to grow rapidly. Second, the market will allow rapid scaling, with revenues growing faster than further capital investment. Third, the firm should be capable of increasing rapidly in value allowing a quick payout in terms of capital gains. Some combination of these three characteristics is necessary to ensure that venture capitalists can secure sufficient capital gains on their successes to offset losses on the unsuccessful firms.

There are three outcomes for venture investments: sale of the stock through an initial public offering, sale of the firm to another firm, or bankruptcy. On any of these liquidity events, the proceeds are returned to the limited partners as either cash or marketable securities. As a rule of thumb, three or four investments are expected to fail at a near total loss. Another three or four neither fail nor are easily liquidated. These are termed the 'living dead' and are judged failures by the venture capitalists. It is the final few that have enormous returns that determine a fund's success, compensate for the failures, and complete the venture capital cycle, enabling the venture capitalists to raise a new fund.

The Globalization of Venture Capital

Venture capital as a formal institution was a US innovation. In other countries and other historical periods, merchant banks, governmental bodies, and even family networks performed some similar functions of supporting new entrepreneurs. Beginning in the late 1970s, a few of the US funds looked abroad for funding opportunities and foreign firms began venture investing in the United States; also, in other countries, governments, financial institutions, and individuals wishing to reproduce the US experience established funds targeting local opportunities. Initially, most of these non-US efforts floundered because of a dearth of entrepreneurs; inadequate institutional, financial, and legal infrastructures to nurture and support entrepreneurship; and a lack of experienced venture capitalists. In the late 1990s, due to the opportunities for new firm formation to exploit the Internet, there was a proliferation of venture capital firms, not only in Europe but also in Latin America and Asia. Furthermore, US firms became increasingly active in Asia, most recently in China and India, Europe, and Israel. With the collapse of the Internet bubble many of these failed, but in 2012, most of the Organisation for Economic Co-operation and Development nations have a venture investment community. The most vibrant of these are located in Israel, China, and, to a lesser degree, India (Ernst and Young, 2010). In 2010, China was the second largest market in the world for venture capital investing.

Conclusion

There are continuing debates regarding the efficacy of venture capital-based innovation systems. Today, venture capital is a significant component of the US national innovation system and, while less important in most other nations, it now exists as a source of capital in at least 45 nations. Undoubtedly, governments around the world see venture capital as a socially desirable institution, and nearly all economically significant nations have adopted policies encouraging venture capital formation, although, in most nations, government-established or government-funded venture capitalists have had very limited success. More recently, due to a variety of factors including a lack of high-quality exits, the emergence of extremely large funds, and changing technologies, some industry observers claim that classical venture investing is in decline or being replaced by angel investors. A recent report released by the Kauffman Foundation

has questioned whether the current limited partnership model is the best structure for the investors in venture capital limited partnerships (Mulcahy et al., 2012). As long as there are a significant number of opportunities to create firms capable of extremely rapid growth in value, there will be opportunities for venture investing.

See also: Capitalism: Global; Capitalism; Information Society; Stockholders' Ownership and Control.

Bibliography

Ahlstrom, D., Bruton, G., Yeh, K., 2007. Venture capital in China: past, present, and future. Asia Pacific Journal of Management 24 (3), 247–268.

Avnimelech, G., Teubal, M., 2010. Venture capital start-up co-evolution and the emergence and development of Israel's new high tech cluster. Economics of Innovation and New Technology 13 (1), 33–60.

Bygrave, W.D., Timmons, J.A., 1992. Venture Capital at the Crossroads. Harvard Business School Press, Boston.

Chen, H., Gompers, P., Kovner, A., Lerner, J., 2010. Buy local? The geography of venture capital. Journal of Urban Economics 67 (1), 90–102.

Dossani, R., Kenney, M., 2002. Creating an environment for venture capital in India. World Development 30 (2), 227–253.

Ernst, Young, 2010. Back to Basics: Global Venture Capital Insights and Trends Report 2010, http://www.ey.com/GL/en/Services/Strategic-Growth-Markets/Global-venture-capital-insights-and-trends-report.

Florida, R., Kenney, M., 1988. Venture capital, high technology and regional development. Regional Studies 22 (1), 33–48.

Gompers, P.A., 1994. The rise and fall of venture capital. Business and Economic History 23, 1–24.

Gompers, P.A., Lerner, J., 1999. The Venture-Capital Cycle. MIT Press, Cambridge, MA.

Guler, I., Guillen, M.F., 2010. Institutions and the internationalization of US venture capital firms. Journal of International Business Studies 41, 185–205.

Hsu, D., Kenney, M., 2005. Organizing venture capital: the rise and demise of American Research & Development Corporation, 1946–1973. Industrial and Corporate Change 14, 579–616.

Jeng, L.A., Wells, P.C., 2000. The determinants of venture capital funding: evidence across countries. Journal of Corporate Finance 6, 241–289.

Kenney, M., 2011. How venture capital became a component of the U.S. NSI. Industrial and Corporate Change 20 (6), 1677–1723.

Kenney, M., Haemmig, M., Goe, W.R., 2008. The globalization of the venture capital industry. In: Macher, J.T., Mowery, D.C. (Eds.), Innovation in Global Industries: US Firms Competing in a New World. National Academies Press, Washington, DC, pp. 313–340.

Kenney, M., Han, K., Tanaka, S., 2002. Scattering Geese: The Venture Capital Industries of East Asia: A Report to the World Bank. Berkeley Roundtable on the International Economy WP146, Berkeley, CA.

Lockett, A., Wright, M., 2002. Venture capital in Asia and the Pacific Rim. Venture Capital 4 (3), 183–196.

Megginson, W.L., 2004. Toward a global model of venture capital? Journal of Applied Corporate Finance 16 (1), 89–107.

Metrick, A., Yasuda, A., 2010. Venture Capital and the Finance of Innovation. Wiley, New York, NY.

Mulcahy, D., Weeks, B., Bradley, H.S., 2012. We Have Met the Enemy… and He is Us: Lessons from Twenty Years of the Kauffman Foundation's Investments in Venture Capital Funds and The Triumph of Hope over Experience. Ewing Marion Kauffman Foundation. http://www.kauffman.org/uploadedFiles/vc-enemy-is-us-report.pdf (accessed 21.05.12).

National Venture Capital Association, 2012. Annual venture investment dollars increase 22% over prior year, according to the Moneytree Report. January 20, 2012.

Poterba, J.M., 1989. Capital gains tax policy toward entrepreneurship. National Tax Journal 42, 375–389.

Sorenson, O., Stuart, T., 2001. Syndication networks and the spatial distribution of venture capital investments. American Journal of Sociology 106, 1546–1588.

Wetzel, W., 1983. Angels and informal risk capital. Sloan Management Review 24, 23–34.

Wright, M., Pruthi, S., Lockett, A., 2005. International venture capital research: from cross-country comparisons to crossing borders. International Journal of Management Reviews 7 (3), 135–165.

Vestibular System

Soroush G Sadeghi, State University of New York at Buffalo, Buffalo, NY, USA
Kathleen E Cullen, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A.H. Clarke, H. Scherer, volume 24, pp. 16170–16176, © 2001, Elsevier Ltd.

Abstract

The vestibular system is phylogenetically one of the oldest sensory systems. Its sensory organs detect angular and linear motion of the head relative to space, as well as gravitational forces. The vestibular system generates reflexes in response to head movements to both stabilize the visual axis (gaze) and maintain head and body posture, and is also necessary for our sense of self-motion and orientation in space. In contrast to importance of vision or audition in our daily lives, which can easily be understood by simply shutting our eyes or plugging our ears, the significance of vestibular function is more difficult to appreciate. When the system is functioning normally, we are usually unaware of a distinct sensation arising from vestibular inputs. This is because vestibular signals are integrated with visual, proprioceptive and other sensory information to provide a sense of motion. The critical role played by the vestibular system in everyday life becomes obvious once it has been compromised. For this reason, clinical studies have provided important insights into function of the vestibular system.

Introduction

The vestibular sensory organs, located on the inner ear, are comprised of the (1) two otolithic organs (the saccule and utricle) that sense linear acceleration (i.e., gravity and translational movements) and (2) three semicircular canals that sense angular acceleration in three perpendicular planes. The receptor cells of the otoliths and semicircular canals send signals through the vestibular nerve fibers to the neural structures that control eye movements, posture, and balance, and provide our sense of self-motion (Figure 1). Integration of vestibular information with proprioceptive, motor, visual, and other sensory signals occurs at the earliest stages of sensory processing in the vestibular nuclei and vestibular cerebellum.

When the vestibular system is impaired, normal activities become difficult if not impossible due to gaze instability and the loss of postural control. Fortunately, the brain is capable of substantial recovery after vestibular loss. However, while many symptoms resolve over time as the result of compensatory changes in the vestibular pathway and sensory substitution for the lost vestibular inputs (Sadeghi et al., 2012), recovery is incomplete. In particular, patients often show deficiencies in gaze stability and postural control during activities that include more challenging (i.e., high frequency and velocity) movements.

In recent years, the findings of clinical, behavioral, and neurophysiological studies have led to a better understanding of how the vestibular system functions in normal conditions and after compensation. Below we review some of these findings.

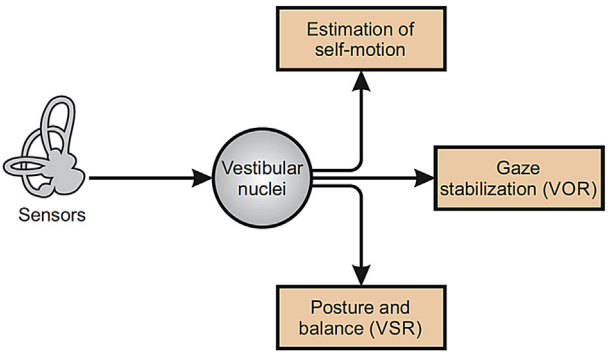

Figure 1 Vestibular signals from the labyrinth of the inner ear are transferred to the vestibular nuclei (VN) via the vestibular afferents of the VIII nerve. In turn, the VN projects to other brain areas to (i) stabilize the visual axis of gaze via the vestibuloocular reflex, (ii) control posture and balance, and (iii) produce estimates of self-motion. VOR, vestibuloocular reflex; VSR, vestibulospinal reflex.

Anatomy of Vestibular End Organs

Movements in three-dimensional space are a combination of rotation and linear movements as well as the omnipresent linear force of gravity. The vestibular end organs, (Figure 2), comprise three semicircular canals (horizontal, posterior vertical, and anterior vertical) that are mutually perpendicular and detect angular motion in all three dimensions of rotation. In addition, each end organ also contains two otolithic organs that detect linear motion: (1) the utricle, which is oriented in the horizontal plane and (2) the saccule, which is oriented in the vertical plane. During everyday activities, our head motion is sensed by the relative movement of the endolymph and the receptor cells within each of these end organs.

The otoliths contain a sensory epithelium called the macula, comprised of hair cells and associated supporting cells (Figure 3). A gelatinous membrane, called the otolithic membrane, overlies the hair cells and their hair bundles also called cilia. Calcium carbonate crystals called otoconia (i.e., Greek for 'ear stones') are embedded in this membrane, and serve to make the membrane heavier than the surrounding

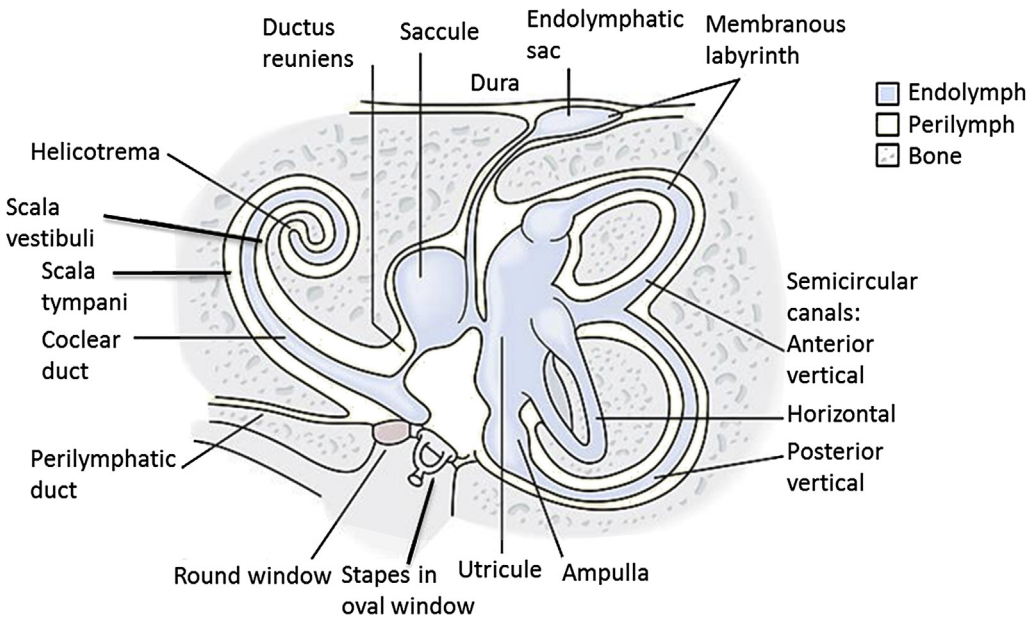

Figure 2 Schematic drawing of the inner ear of man with three semicircular canals, the two otolithic organs and the cochlea.

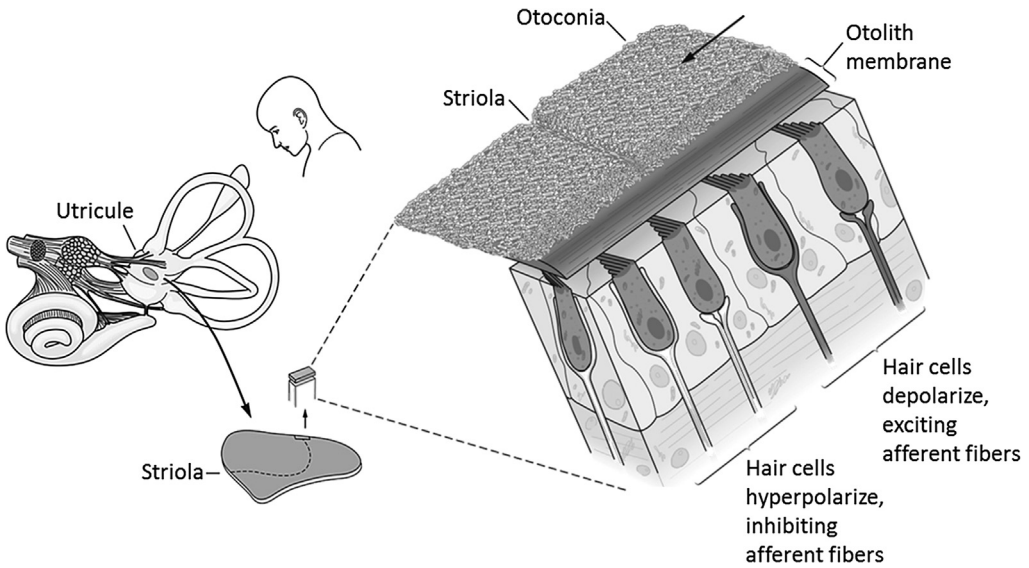

Figure 3 Schematic of the macula and otolithic membrane in an otolithic organ.

fluid. As a result during linear motion, a relative shearing motion between the otolithic membrane and the macula displaces the hair cell bundles, in turn generating a receptor potential in the hair cells. In particular, the relatively larger mass of the otolithic membrane causes it to lag behind the hair cells of the macula, leading to transient shearing of the hair cell bundle.

The sensory epithelium within each of the three semicircular canals is called the crista and is located in an expansion called the ampulla (**Figure 4**). The cilia of the hair cells of the semicircular canals extend out of the crista into a gelatinous membrane, called the cupula. Importantly, each cupula spans the entire cross section of its ampulla, effectively forming a water-tight barrier through which endolymph cannot flow. Consequently, when the head turns in the plane of one of the semicircular canals, the inertia of the endolymph fluid causes it to lag relative to the hair cells of the crista. Notably, because of the small diameter of the canals (~0.3 mm in humans) and the viscosity of the endolymph, the canals perform a mathematical integration; while the motion stimulus required to activate the canals is rotational head acceleration, the hair cells of the semicircular canals encode rotational head velocity (Chapter 4; Goldberg et al., 2012).

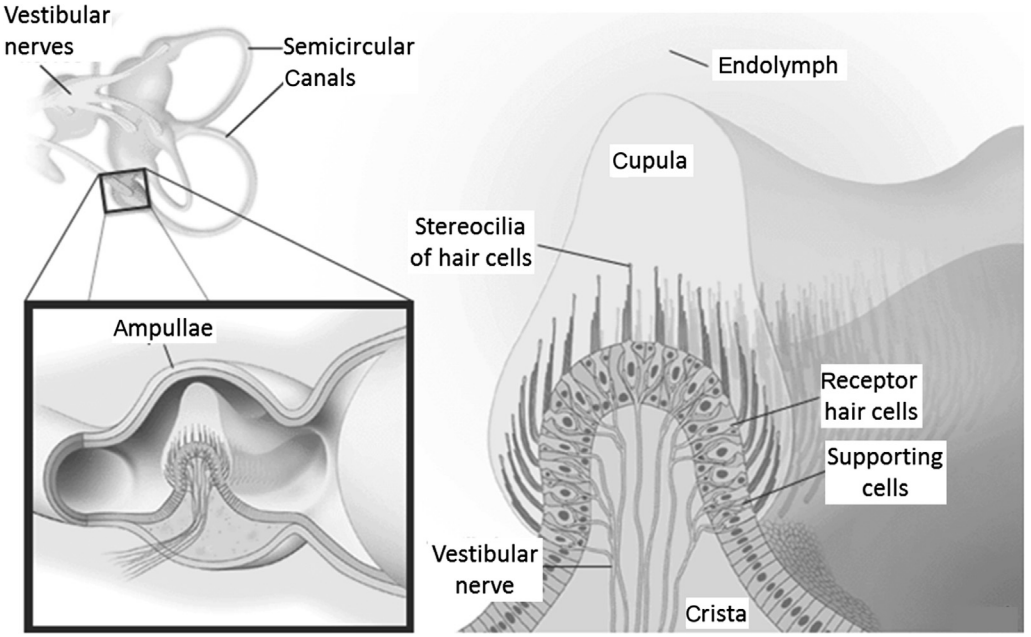

Figure 4 Schematic of the ampulla and cupula of a semicircular duct.

Physiology of the Vestibular System

Vestibular Hair Cells and Afferents

Bending of the cilia of the hair cells opens mechanically gated channels, leading to the inflow of positive ions, and hair cell depolarization. This in turn leads to opening of calcium channels and the calcium-dependent release of neurotransmitters from hair cells onto the afferent nerve terminals. There are two types of vestibular hair cells: type II hair cells which are cylindrical and present in all animals, and flask-shaped type I hair cells that are found only in amniotes (i.e., mammals, reptiles, and birds).

Each type II hair cell is innervated by about 20 bouton afferent terminals that receive inputs from single ribbons. In contrast, each type I hair cell receives a calyx terminal that completely covers its basolateral walls and receives inputs from ~20 ribbons (Lysakowski and Goldberg, 2008). Recent studies have shown that these calyx terminals can receive either synaptic or nonsynaptic inputs from hair cells (Songer and Eatock, 2013) that result in accumulation of potassium ions in the synaptic cleft during stimulation (Contini et al., 2012; Lim et al., 2011). Functionally, afferents can be grouped based on the regularity of their resting discharge, using a normalized coefficient of variation (CV*) (reviewed Goldberg, 2000). Regular units show low variability in interspike intervals (ISI) of resting discharge ($CV^* < 0.1$), while irregular units have higher variability in resting discharge ISI ($CV^* > 0.1$). Afferents with pure bouton terminals that contact type II hair cells have regular discharge properties and mostly innervate the periphery of the vestibular neuroepithelium. In contrast, afferents with pure calyx terminals that contact type I hair cells have irregular discharge and mostly innervate the central area of the neuroepithelium. Most afferents, however, have both bouton and calyx terminals (i.e., dimorphic), and thus contact both type I and II hair cells. These dimorphic afferents can have regular or irregular resting discharge, with regulars and irregulars innervating the peripheral and central areas of the neuroepithelium, respectively. To date, the contribution of boutons versus calyces to the output of dimorphic afferents is not understood.

The responses of semicircular canal afferents with less resting discharge variability (i.e., regular units) encode more information than those with more irregular discharges (Sadeghi et al., 2007). In contrast, more irregular afferents have higher response gains than regular afferents, a difference which increases with increasing frequency (reviewed in Goldberg, 2000). Thus, information from the end organs is sent to the brain via two pathways or information channels: (1) regular afferents which provide more detailed information about the specific trajectory of head motion, and (2) irregular afferents which act as high frequency motion event detectors. Current research is aimed at understanding the functional implications of these differences in coding strategies at higher levels of signal processing (reviewed in Cullen, 2012).

Vestibular Nuclei

The vestibular afferents project to the vestibular nuclear complex which is comprised of four main nuclei: the medial vestibular nucleus (receives inputs mainly from horizontal canals), the superior vestibular nucleus (inputs mainly from vertical canals), the lateral vestibular nucleus (LVN, inputs mainly from saccule), and the inferior (or descending) vestibular nucleus (inputs mainly from utricle and saccule). In addition, vestibular nuclei also receive inputs from other regions of the brainstem, the vestibular cerebellum and higher brain areas (e.g., cortex and cerebellum), relaying extravestibular inputs such as somatosensory, visual, eye movement-related and premotor head movement commands. The

functional implications of this convergence of input at the level of the vestibular nuclei are discussed below in relation to the stabilization and control of gaze (i.e., the vestibuloocular reflex (VOR); Section Stabilization of Gaze by Vestibuloocular Pathways) and posture (i.e., vestibulospinal reflexes; Section Stabilization of Posture by Vestibulospinal Pathway).

Stabilization of Gaze by Vestibuloocular Pathways

As we move through the world, the VOR produces compensatory eye movements equal in magnitude but opposite in direction to head movements to help stabilize the visual world on the retina. These eye movements serve to effectively stabilize gaze and ensure clear vision during our everyday activities. The efficacy of VOR can be easily quantified by measuring eye movements (typically by electrooculography, and more recently by video-based systems) and has thus played a major role in the diagnosis of vestibular disorders. Arguably, the best known clinical case is that of physician who called himself JC. In 1952, JC completely lost all vestibular functions after he was treated with an ototoxic antibiotic (streptomycin) for leg infection. One of his main symptoms was the lack of gaze stability. For instance, he was unable to read street signs or recognize people's faces while walking. However once he stopped moving these images became clear. Notably, for JC as well as other patients with substantial vestibular loss, normal activities that we take for granted become difficult since even small head movements are accompanied by gaze instability since the VOR is no longer effective.

The VOR produces compensatory eye movements in response to both rotational and translational head movements. Specifically, the rotational VOR (r-VOR) is evoked in response to angular head motion that stimulates the semicircular canals (e.g., while shaking head back and forth to indicate the answer 'no'). In contrast, the translational vestibuloocular reflex (t-VOR) is evoked in response to linear head motion that stimulates the otolithic organs (e.g., while riding on a subway or train). In everyday life, rotational and translational head movements generally occur together, and thus the r-VOR and t-VOR are simultaneously evoked.

The neural pathways that mediate the VOR have been well characterized. Additionally, as detailed in the following Chapter entitled *The Vestibulo-ocular Reflex: Adaptation and Compensation* the VOR pathway shows impressive adaptation in response to behavioral requirements. For example, when wearing corrective lens for vision, the VOR adaptively increase or decrease as needed in order to maintain stable gaze during self-motion. In particular, the relative simplicity of the neuronal pathway that mediates the VOR and its calibration has made this reflex an excellent model system for bridging the gap between the cells, neuronal circuits, and behavior. The most direct VOR pathway consists of projections from vestibular afferents to neurons in the vestibular nuclei, which in turn project to extraocular motoneurons, providing eye movements 5–6 m after head movements (Huterer and Cullen, 2002). A schematic showing the pathways mediating the r-VOR evoked by horizontal head movements (i.e., yaw rotations that activate the horizontal semicircular canal) is shown in **Figure 5**. In response to a rightward head motion, the afferents from the right horizontal canals are excited. In turn, these afferents activate central neurons in the vestibular nuclei, which in turn project to the eye muscle motoneurons that drive the movement of both eyes to the left (i.e., the lateral rectus muscle of the left eye and the medial rectus muscle of the right eye). Since the mechanical range of eye rotation is limited anatomically by the extraocular muscle and optical nerve structures, the eye can only move so far to compensate for sustained head motion in any direction. Accordingly, compensatory VOR eye movements are reset by means of a quick phase in the anti-compensatory direction. The inset in **Figure 5** illustrates the eye movements evoked by sustained rightward head rotation. Note the eye movement has a sawtooth pattern, termed 'vestibular nystagmus,' characterized by compensatory 'slow phase' eye movements to the left and rapid resetting 'fast phase' eye movements to the right. The analysis of such vestibular nystagmus is often used clinically to aid in the diagnosis of vestibular disorders.

A similar pattern, known as optokinetic nystagmus, can also be elicited by rotation or translation of the visual field while the person remains stationary. Notably, the optokinetic system uses visual rather than vestibular inputs to stabilize gaze (*see* Visual Streams: Dorsal and Ventral). These eye movements are evoked for example, when a person looks out at the world through a train's window or the side window of a car. In this case the motion of the visual scene causes a compensatory eye movement in the direction opposite to the train's motion with a fast phase in the direction of the train's motion. Importantly, the optokinetic system has response dynamics complementary to those of the VOR. As reviewed above, the VOR has a short latency and is most effective for transient head movements (i.e., >0.1 Hz). In contrast, while the optokinetic system has

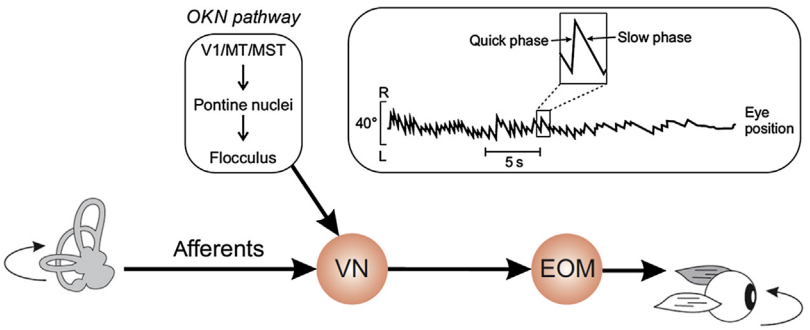

Figure 5 Diagram of the reflex pathway generating the vestibuloocular reflex (VOR), showing the response to rightward head motion.

a longer latency, it responds more robustly to low frequency, more sustained head movement where the VOR is less effective. Neurophysiological studies have shown that optokinetic eye movements are largely mediated by subcortical pathways, which project to the second stage of the 3 neuron VOR pathway (see OKN Box in **Figure 5**).

In real world situations, where vestibular and optokinetic stimulation occur simultaneously (as would be the case when walking down the street during the day), commands from the two pathways are integrated within single neurons in the vestibular nuclei to produce sustained responses, that in turn generate robust stabilizing eye movements. In addition, when behaviorally advantageous, it is possible to suppress the VOR. For instance, if we are reading a newspaper when riding as a passenger in a car or bus, our goal is no longer to stabilize the world on the retina. Instead, our goal is to stabilize the words that we are reading – on a page of the newspaper that is moving together with us. In this case, both visual and nonvisual (behaviorally contingent) mechanisms work together to suppress the VOR so that our eyes remain stable relative to the words that are moving with us (rather than relative to the stable world). This mismatch between the vestibular and visual inputs can cause motion sickness if it becomes persistent. In addition, the VOR is suppressed when we make combined eye-head movements to look at targets of interest. Such suppression is useful since the reflexive VOR would drive the eye in the direction opposite to the intended voluntary eye movements made to shift the axis of gaze (*see* Eye Movement, Control of: Oculomotor Control).

As detailed in Section Clinical Manifestations of Vestibular Dysfunction below, spontaneous vestibular nystagmus in the absence of vestibular or visual stimulation is the hallmark of an imbalance of input from the vestibular end organs from the two sides. It is important to eliminate visual fixation in order to fully observe spontaneous nystagmus and in turn diagnose vestibular imbalance. This can be done by having patients wear thick lenses (Frenzel) that prevent them from fixating on visual targets in the local environment.

Stabilization of Posture by Vestibulospinal Pathways

The vestibular system also makes important contributions to the maintenance of posture and control of balance during everyday life. Neurons in the vestibular nuclei that receive direct inputs from the vestibular nerve in turn send direct projections to the spinal tract to control posture and balance. This streamlined circuitry allows the rapid postural corrections in response to unexpected motion vital to prevent falling in normal subjects.

The pathways that control balance have been relatively well characterized. The vestibulospinal tract emerges from the LVN in the brainstem and runs unilaterally to the spinal motoneurons of the extensor muscles (**Figure 6**). When a person experiences rightward motion, the vestibulospinal tract activates the extensor muscles on the right, to counteract the motion that has been experienced. Below we specifically consider the vestibulocollic (VCR) reflex – a vestibulospinal reflex that stabilizes the head in space by activating the neck musculature in response to vestibular stimulation.

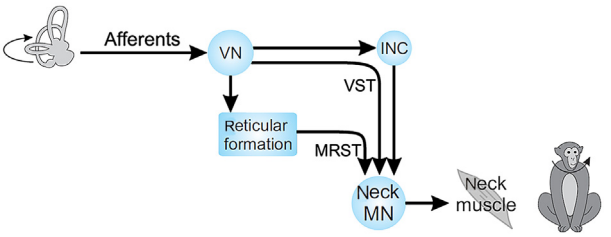

Figure 6 Diagram of the reflex pathway generating the vestibulocollic reflex (VCR), showing the response to rightward head motion.

The VCR produces a command to move the head in the direction opposite to that of the current head-in-space motion. Accordingly, it functions to stabilize the head relative to inertial space. Consistent with its role in head stabilization, experiments have shown that the head becomes unstable when the ducts of semicircular canals are blocked to prevent activation of the hair cells during motion. As was the case for the VOR (see Section Stabilization of Gaze by Vestibuloocular Pathways above) the most direct pathway mediating the VCR is comprised of a three-neuron arc. However, in contrast to the VOR, more complicated multisynaptic circuitry makes the dominant contribution to the reflex pathways.

As was the case for the VOR, vestibulospinal reflexes can be counterproductive during certain behaviors. For example, when we make active neck movements the stabilization responses produced by the VCR would produce head-movement commands in the direction opposite to those of the intended self-generated motion. Recent studies have shown that vestibulospinal pathways are suppressed, on average, by about ~70% during the active head movements (reviewed in Cullen, 2012). Interestingly, these same neurons faithfully encode information about passive head rotations, which occur during the execution of voluntary movements. This ability to selectively respond to passive vestibular stimuli during concurrent active and passive motion is likely to underlie our ability to reflexively respond to changes in vestibular input that the brain does not expect. For example, in order to recover from tripping over an obstacle while walking or running, it is crucial that vestibulospinal reflexes generate robust responses to the unexpected component of head motion.

The accurate control of posture and estimation of self-motion further depends on the integration of vestibular and proprioceptive information, which is profoundly disrupted in cerebellar patients. The neuronal computation performed in the cerebellum combines vestibular and proprioceptive information to provide an explicit representation of unexpected motion that ensures accurate spatial orientation and postural control (Brooks and Cullen, 2013). In response to unexpected motion, the cerebellum computes unexpected motion within milliseconds to then send an appropriate signal to the spinal cord to rapidly adjust balance (**Figure 7**). When we move voluntarily, our brain normally construct an estimate of the sensory inflow that it expects from our sensory systems (i.e., vestibular, proprioception, and somatosensory). The cerebellum then computes the mismatch between this expectation and the actual sensory input.

It is thought that athletes can compute this mismatch for very complex motions; consider for example a gymnast doing

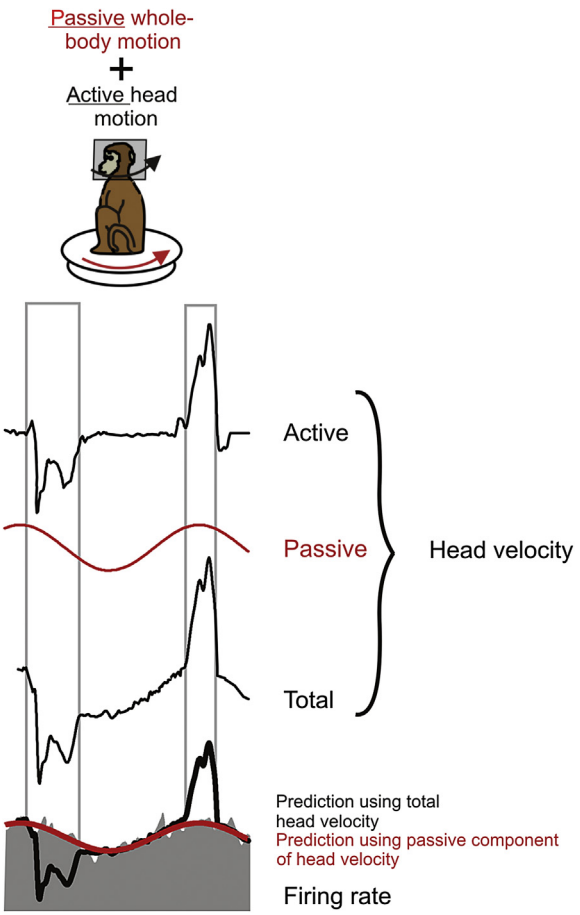

Figure 7 Cerebellar nuclei neurons selectively respond to passive head motion when experienced concurrently with active motion.

a back flip on a balance beam. In contrast, cerebellar patients cannot do this, even for a simple movement like placing a foot on a step. In addition, as people age, their sense of balance worsens because they lose some of the sensory input – most notably vestibular input that was present in younger years. As detailed in Section Peripheral Vestibular Disorders below, this makes balance more challenging because there is less reliable information to estimate movements. As a result, older adults often rely more on vision, which is a relatively slow sense compared with vestibular and proprioceptive systems. Rehabilitation programs often encourage patients to retrain their sensory organization and use faster sensory inputs such as vestibular and proprioception to improve their balance.

Interestingly, the sensation of height vertigo has also been explained within the context of sensory mismatch. This sensation occurs when a person stands on the top of a high tower or the edge of a cliff. While standing, most people display some degree of body sway that is regulated by the multisensory convergence in the vestibular system. In this particular context, vision plays a major role. Notably there is a discrepancy in visual perception of motion as a result of the difference in distance between the background – located far from the foreground. The lack of consistent information impairs postural control and thus increases body sway potentially operating in a 'positive feedback' loop. Instability can be counteracted having the subject focus their view on the near foreground, or by providing other stabilizing sensory cues – for example, by allowing him or her to hold on to a hand rail or touch the ground.

Multisensory Integration during Self-Motion in Everyday Life

As reviewed above, information from the vestibular sensors is combined with extravestibular cues, such as proprioception, visual inputs and motor efference signals, at the earliest stages of central vestibular processing. As a result, the multimodal representations of self-motion are used to provide: (1) Behaviorally appropriate gating of reflexes including the VOR and VCR during active shifts of gaze and active head movements, respectively; (2) Estimates of the relative motion of neighboring parts of the body (e.g., body vs head motion) required to ensure stable posture and perception; and (3) Accurate representations of unexpected self motion required for perceptional stability. This integration is also evident at higher levels of vestibular processing. Notably, the convergence of visual, proprioceptive, and vestibular inputs is observed in the main vestibular cortical area, the parietoinsular vestibular cortex (PIVC) as well as in head direction neurons (found in the lateral mammillary nuclei, retrosplenial cortex, lateral dorsal thalamus, striatum, and entorhinal cortex) that discharge in relation to the animal's directional heading with respect to the environment.

Clinical Manifestations of Vestibular Dysfunction

Disorders of the vestibular system are more difficult to identify than those of other sensory systems (i.e., hearing loss or blindness in the case of the auditory and visual systems). Nevertheless, the symptoms of vestibular dysfunction namely, dizziness, imbalance, and disequilibrium are extremely common. The probability of vestibular dysfunction increases with age. For example, a recent epidemiological study (National Health and Nutrition Examination Surveys (NHANES)) suggests that 35% of adults in the United States aged 40 and older have vestibular dysfunction, corresponding to nearly 70 million Americans.

The most common peripheral vestibular disorders include: (1) Benign paroxysmal positional vertigo (BPPV) in which patients have the false sense of self-rotation when adopting certain static head positions, (2) Ménière's disease in which patients experience episodes of vertigo, fluctuating hearing loss, and tinnitus (illusionary ringing in the ears), (3) sudden unilateral peripheral vestibular loss, and (4) superior semicircular canal dehiscence syndrome, which leads to symptoms and signs of vestibular and auditory dysfunction. Each of these disorders can cause gaze instability and postural imbalance and in addition often have complex psychological manifestations (including anxiety and phobias). Finally, it is important to mention that the loss of peripheral vestibular input with aging plays a prominent role in the cause of falls in the elderly. Below, we consider the physiological cause of each disorder as well as current treatments used to ameliorate deficits. Comprehensive

descriptions of central disorders (e.g., migraine-associated dizziness, vertebrobasilar ischemic stroke, and vertebrobasilar insufficiency) can be found elsewhere (Dieterich, 2007).

Peripheral Vestibular Disorders

Benign paroxysmal positional vertigo (BPPN) is the most common cause of vertigo, experienced by more than 50% of people over the age of 70. BPPN occurs when dislodged calcium carbonate crystals from the otoconial membrane of the utricle move into a semicircular canal duct. Typically, due to the geometry of the labyrinth, dislodged crystals accumulate in the posterior semicircular canal. Accordingly, the resultant nystagmus is most often comprised of vertical and torsional eye movement. The dynamics of the crystals motion through the labyrinth endolymph account for the long latency of the pathological nystagmus (typically 10–15 s after the head movement). To treat BPPV, the patient's head is moved to position the plane of the affected posterior canal vertically. In turn, this causes the crystals to move from the canal duct into the crus commune and then into the vestibule as a result of gravity, from where they are subsequently absorbed.

Ménière's disease is characterized by periods of vertigo, hearing loss, and tinnitus and results in a substantial reduction in quality of life. Episodes of vertigo, especially their unpredictability, often prevent Ménière's disease-afflicted patients from conducting daily activities and can create anxiety. Historically, the underlying cause was thought to be related to an excess of fluid in the inner ear called 'hydrops.' However, the disease probably has multiple etiologies including viral and autoimmune responses, hereditary predisposition, and head injuries.

Sudden unilateral loss of peripheral function is most commonly caused by vestibular neuritis which is thought to be of either a viral or vascular origin. The resulting inflammation and possible ischemic injury to hair cells and/or afferent nerve fibers leads to a loss of peripheral function in the affected side. In addition, surgery for acoustic neuromas produces unilateral loss of vestibular function. More rapid onset of the loss of vestibular function produces more severe symptoms of gaze instability and postural instability. Patients also commonly experience a sense of severe rotatory vertigo which can be accompanied by nausea and vomiting.

Immediately, following vestibular loss, patients will have a spontaneous nystagmus, that includes horizontal and torsional components (with slow phases of the nystagmus directed toward the affected side). Furthermore, the unilateral loss of otolithic function produces an ocular tilt reaction characterized by a counter-rolling of the eyes toward the side of vestibular hypofunction when the head is tilted in the roll axis. Finally, and most important clinically, a patient's horizontal VOR becomes asymmetric, such that head rotations that would have excited the horizontal canal on the damaged side (and inhibit canals on the intact side) generate a very small angular VOR.

In most patients, many of these symptoms resolve within a week, however gaze stability remains subpar for the more challenging head rotations. The functional improvement observed is the result of compensatory processes in central vestibular pathways, as well as the substitution of information from the proprioceptive, visual and motor systems (Sadeghi et al., 2012). Interestingly, the degree to which compensation and substitution strategies are helpful varies considerably across patients. Common rehabilitation procedures such as Cawthorne-Cooksey exercises effectively incorporate multimodal (i.e., combined vestibular and proprioceptive) stimulation into rehabilitative training to promote compensation.

Superior semicircular canal dehiscence syndrome is a relatively rare disorder. It is caused by a thinning or opening in the bone overlying the superior semicircular canal. Symptoms include autophony (the sensation that one's voice is too loud in the affected ear), sound-induced loss of balance (Tullio phenomenon), and low frequency hearing loss. The cause is thought to be either congenital or the result of physical trauma to the head. Treatment is usually surgical resurfacing/replacing of the affected bone.

Vestibular Function and Aging

Numerous studies provide evidence for age-related changes in peripheral vestibular sensory organs. In particular, aging is accompanied by progressive hair cell degeneration, otoconial degeneration in the otolithic organs, and decreasing number of Scarpa's ganglion neurons. Type II hair cell counts decline at the same rate for all 5 sensory epithelia. Type I hair cell counts decrease in all three semicircular canals at a significantly faster rate than in the otoliths. There is also a decrease in primary vestibular neurons. As a result, the efficacy of the VOR decreases with age and older adults also become more susceptible to falls and dizziness.

See also: Cerebellum: Cognitive Functions; Eye Movement, Control of: Oculomotor Control; Visual Streams: Dorsal and Ventral.

Bibliography

Brooks, J.X., Cullen, K.E., 2013. The primate cerebellum selectively encodes unexpected self-motion. Current Biology 23 (11), 947–955.

Contini, D., Zampini, V., Tavazzani, E., Magistretti, J., Russo, G., Prigioni, I., Masetto, S., 2012. Intercellular K^+ accumulation depolarizes type I vestibular hair cells and their associated afferent nerve calyx. Neuroscience 227, 232–246.

Cullen, K.E., 2012. The vestibular system: multimodal integration and encoding of self-motion for motor control. Trends in Neuroscience 35 (3), 185–196.

Dieterich, M., 2007. Central vestibular disorders. Journal of Neurology 254 (5), 559–568.

Goldberg, J.M., 2000. Afferent diversity and the organization of central vestibular pathways. Experimental Brain Research 130 (3), 277–297.

Goldberg, J.M., et al., 2012. The Vestibular System, a Sixth Sense. Oxford University Press.

Huterer, M., Cullen, K.E., 2002. Vestibuloocular reflex dynamics during high-frequency and high-acceleration rotations of the head on body in rhesus monkey. Journal of Neurophysiology 88 (1), 13–28.

Lim, R., Kindig, A.E., Donne, S.W., Callister, R.J., Brichta, A.M., 2011. Potassium accumulation between type I hair cells and calyx terminals in mouse crista. Experimental Brain Research 210 (3–4), 607–621.

Lysakowski, A., Goldberg, J.M., 2008. Ultrastructural analysis of the cristae ampullares in the squirrel monkey (Saimiri sciureus). Journal of Comparative Neurology 511 (1), 47–64.

Sadeghi, S.G., Chacron, M.J., Taylor, M.C., Cullen, K.E., 2007. Neural variability, detection thresholds, and information transmission in the vestibular system. Journal of Neuroscience 27 (4), 771–781.

Sadeghi, S.G., Minor, L.B., Cullen, K.E., 2012. Neural correlates of sensory substitution in vestibular pathways following complete vestibular loss. Journal of Neuroscience 32 (42), 14685–14689.

Songer, J.E., Eatock, R.A., 2013. Tuning and timing in mammalian type I hair cells and calyceal synapses. Journal of Neuroscience 33 (8), 3706–3724.

Vestibulo-Ocular Reflex, Adaptation of the

Diana E Mitchell and Kathleen E Cullen, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Ito, volume 24, pp. 16176–16179, © 2001, Elsevier Ltd.

Abstract

Motor learning not only allows us to develop the skills necessary to perform fine-tuned movements, such as a tennis serve, but also ensures the accuracy of simpler reflex behaviors. Notably, the vestibulo-ocular reflex (VOR), which functions to stabilize images on the retina, shows impressive adaptation in response to changes in environment requirements, such as those imposed by a new pair of eyeglasses. The pathways that mediate the VOR and its adaptation are relatively simple, making this reflex an excellent model system for bridging the gap between brain and behavior. This article reviews behavioral, electrophysiological, and lesion studies of VOR adaptation.

Introduction to Vestibulo-Ocular Reflex

As reviewed in the previous article entitled *The Vestibular System*, the vestibulo-ocular reflex (VOR) produces eye movements that are equal in magnitude but opposite in direction to the head motion in order to stabilize gaze over a wide range of head movements. For example, when a subject experiences rightward head rotation in darkness, the VOR will produce an eye motion to the left. The gain of the rotational VOR, defined as compensatory eye velocity divided by head velocity, is ~1.0. The three-neuron arc that is responsible for mediating the direct VOR pathway consists of projections from vestibular afferents to interneurons in the vestibular nuclei, which in turn project to extraocular motoneurons (**Figure 1(a)**). The simplicity of this pathway is reflected in its fast response time (**Figure 1(b)**): compensatory eye movements lag head movements by only 5–6 ms in the primate (reviewed in Cullen, 2012). Because, the VOR initiates eye movements at much shorter latencies than visually mediated eye movements and is thus fast enough to stabilize the gaze during common activities such as locomotion.

Within the vestibular nuclei, the primary interneurons that mediate the direct VOR pathway are called position-vestibular-pause (PVP) neurons (**Figure 1(a)**; reviewed in Cullen, 2012), named so because in addition to responding to vestibular stimulation, they encode eye position signals and pause during saccades. A second class of interneurons, termed floccular target neurons (FTNs), also contribute to the VOR pathway (**Figure 1(a)**; reviewed in Cullen, 2012). Notably, FTNs receive input from the flocculus of the cerebellum as well as from the vestibular nerve. The responses of FTNs complement those of PVP neurons during our daily activities, and play a vital role in calibrating the VOR (reviewed in the following section).

Adaptive Capability of the VOR

A mismatch between visual and vestibular stimuli drives VOR adaptation. A typical training paradigm used to induce changes in the VOR gain is illustrated in **Figure 2**. Before any training, the VOR gain is approximately unity, so that, the eyes move with the same amplitude, but in the opposite direction, as the head. A decrease in VOR gain is induced by training with visual motion in the same direction as the head (gain-down stimulus). Following this type of training, the magnitude of the eye movement is less than that of the head and the VOR gain is reduced (<1). An increase in VOR gain occurs when visual motion is in the direction opposite to the head (gain-up

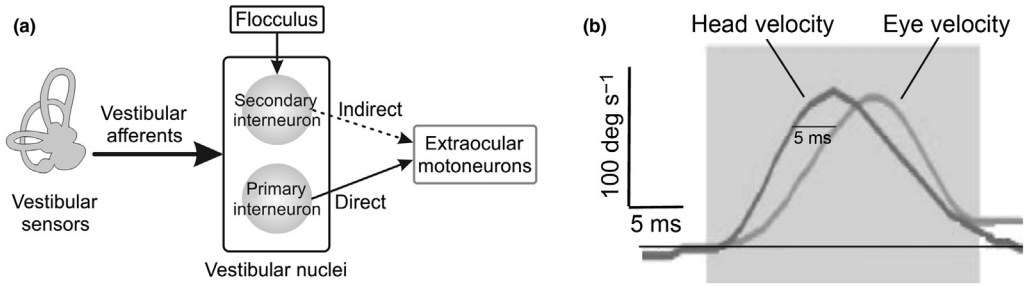

Figure 1 (a) Circuit diagram describing pathways that mediate the VOR. The direct pathway mediating the VOR consists of vestibular afferents, which project to the primary VOR interneurons in the vestibular nuclei, which in turn project to ocular motoneurons. Secondary VOR interneurons receive input from the vestibular afferents and the cerebellar flocculus, and also project to ocular motoneurons. These cells are part of the pathway that is involved in the calibration of the VOR. (b) Eye response following a transient head perturbation. Following the onset of head rotation, the eye begins to counter-rotate after a latency of 5–6 ms. The gain of the reflex is eye velocity divided by head velocity. Note: eye velocity has been inverted to facilitate comparison. Modified from Huterer and Cullen (2002).

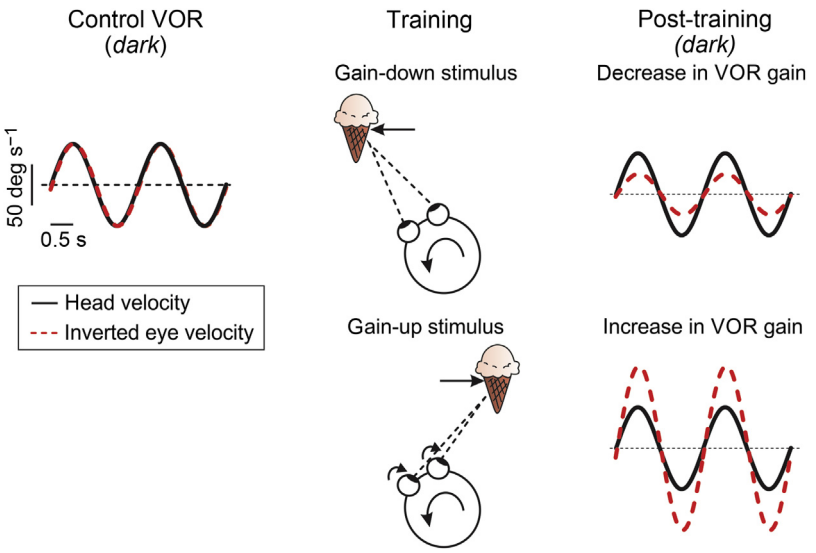

Figure 2 A typical training paradigm used to induce changes in the VOR gain. Before any training, the VOR gain is ~1, that is the eyes move with the same amplitude, but in the opposite direction from the head. A decrease in VOR gain is induced by training with image motion in the same direction as the head (gain-down stimulus). Following this type of training the magnitude of the eye movement is less than that of the head and the VOR gain is reduced (<1). An increase in VOR gain occurs when image motion is in the direction opposite that of the head (gain-up stimulus). Subsequently, the magnitude of the eye movement is more than that of the head and the VOR gain is increased (>1).

stimulus). Subsequently, the magnitude of the eye movement is more than that of the head and the VOR gain is increased (>1).

One of the most impressive examples of VOR adaptation was shown by Gonshor and Jones (1976). In this study, participants wore optical lenses that induced vision-reversal such that left was right and vice versa. The gain of the VOR immediately dropped within 2 days of wearing lenses. After 3–4 weeks the response of the VOR actually reversed: VOR eye movements moved in the same direction of the head. Therefore, VOR was able to adapt to these dramatic new conditions resulting from the change in optics so that gaze stability could be achieved. While in everyday life the demands on the VOR are less challenging, VOR adaption is nevertheless essential to ensure gaze stability. In humans, during the first few years of life, the VOR is continuously adjusted to account for large changes in the size of the head (head circumference increases by 30% in the first year). Later in life, the VOR adapts to account for the magnification of corrective lenses when prescribed, such as for myopia. For example, individuals suffering from myopia are typically prescribed convex corrective lenses resulting in their VOR gains being lower than people who do not wear glasses. During aging, the reflex is also able to adapt to the incremental loss of the receptor cells. Finally, the VOR shows remarkable recovery following the loss of unilateral labyrinthine input as a result of injury.

Cerebellar Flocculus

The floccular complex is a structure within the cerebellum and is required for VOR motor learning; the gain of the VOR cannot be adapted after this structure has been removed. In primates, the flocculus proper is defined as the four most medial lobules caudal to the posterolateral fissure. Adjacent to the fissure is a lobule that constitutes a transition zone, termed the medial extension, which is then followed by five to six more rostral lobules that make up the ventral paraflocculus. Like all of the cerebellum, the flocculus receives two main types of inputs: (1) mossy fibers and (2) climbing fibers (Figure 3(a)). Mossy fiber inputs provide head movement-related information (i.e., vestibular) via projections from the vestibular nerve and vestibular nuclei as well as eye movement-related information from the vestibular, reticular, raphe, and perihypoglossal nuclei. The mossy fibers project to granule cells, whose axons form the parallel fibers. These parallel fibers make synapses onto the dense dendrites of Purkinje cells, which, in turn, project to FTNs in the vestibular nuclei. The climbing fiber input originates in the inferior olive and makes strong synaptic connections onto Purkinje cells. Although typically only one climbing fiber will project to a single Purkinje cell, the strength of the synapse is powerful enough to evoke repetitive discharge in that Purkinje cell. This inhibitory side loop through the cerebellar flocculus (Figure 3(a)) plays an important role in mediating VOR adaptation (reviewed in the following section).

The mossy fibers and climbing fibers influence the discharge pattern of Purkinje cells differently. Mossy fiber input will give rise to simple spikes (Figure 3(b)) in Purkinje cell firing, which can reach discharge rates as large as 300 sp/s. On other hand, climbing fiber input causes complex spikes in Purkinje cells (Figure 3(b)), which occur less often than simple spikes (at a rate of ~1 sp/s) and are longer in duration than simple spikes, lasting for as long as 2–5 ms.

Two Influential Models of Learning in the VOR

In order to change the gain of the VOR response, synaptic weights within the pathway that mediates this reflex must be

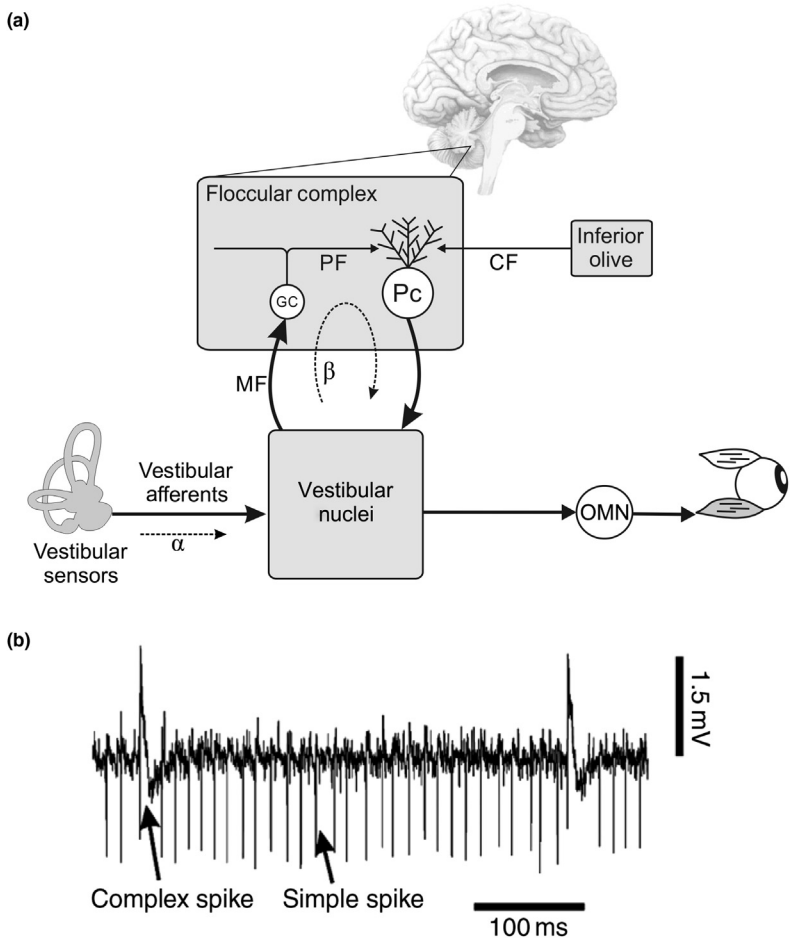

Figure 3 (a) Two pathways control the gain of the VOR. The main reflex path with a gain α is paralleled by an inhibitory side branch through the floccular complex of the cerebellum with a gain β, so that total gain is proportional to (α–β). *Abbreviations:* CF, climbing fibers; GC, granule cells; MF, mossy fibers; OMN, oculomotor nuclei; Pc, Purkinje cells. (b) Recordings from an example Purkinje cell showing its high-frequency simple spike responses as well as a complex response modulated by a climbing fiber. Climbing fiber excitation results in an EPSP followed by a pause in firing rate. Courtesy of Simpson, J., and Maruta, J.

altered. Where and how these adaptive processes occur has been debated, and two hypotheses of motor learning in the VOR have dominated the literature (reviewed in Broussard and Kassardjian, 2004). The principal difference between these two models is the location of the modified synapses that underlie VOR adaptation.

In 1972, Ito proposed a theory of VOR adaptation based on the Marr–Albus model of cerebellar motor learning. In Ito's framework, the climbing fibers from the inferior olive send retinal slip or an error signal to the floccular Purkinje cells. This generates complex spikes, which triggers a change in the weight of the synapse between the parallel fibers and the Purkinje cells. This error signal modifies the head velocity-related modulation of Purkinje cells in order to change the gain of this cerebellar loop of the VOR circuitry so that the VOR is compensatory (Figure 4(a); red star).

On the other hand, Miles and Lisberger proposed that the projection from floccular Purkinje cells to neurons in the vestibular nuclei induced learning, causing plasticity exclusively within the brainstem (Figure 4(b); green star). This hypothesis proposes that during normal VOR, the eye and head movement inputs effectively cancel out and floccular Purkinje cells are not modulated. When there is a visual–vestibular mismatch, however, retinal slip would drive smooth pursuit eye movements in order to maintain stable gaze. This eye movement signal would get sent to Purkinje cells, whose output would encode the error between the eye and head movement. Finally, the error signal guides plastic changes at the level of neurons in the vestibular nuclei, so that the modified VOR pathways will produce stable gaze.

Evidence for a Multisite Hypothesis of VOR Motor Learning

More recent evidence from single unit recording experiments indicates that the site of VOR motor learning is not exclusively restricted to the floccular complex or the vestibular nuclei (reviewed in Broussard and Kassardjian, 2004). During

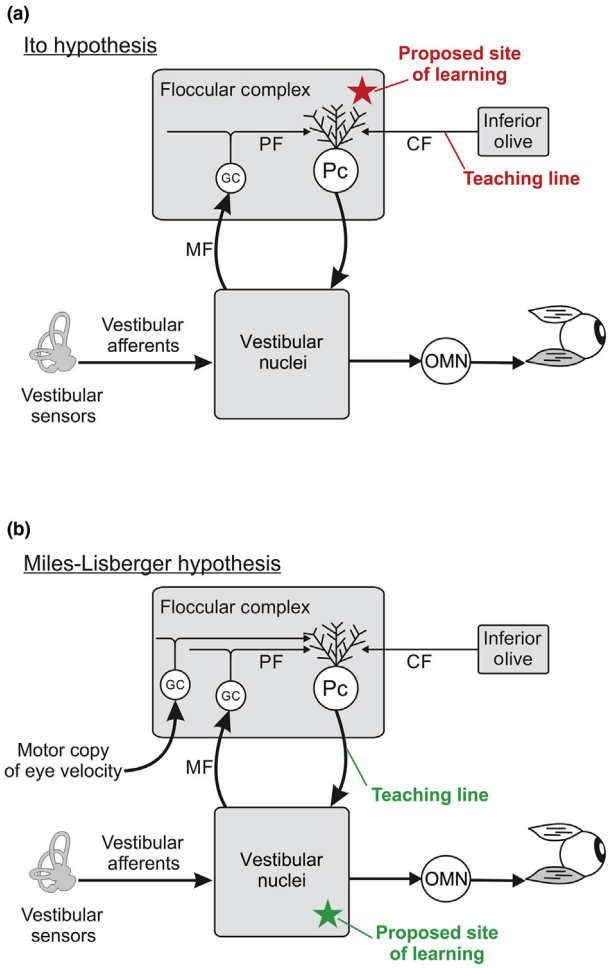

Figure 4 Two classic models of VOR motor leaning. (a) In Ito's model, climbing fibers encode a visual error signal which induces learning at the synapses of parallel fibers onto the Purkinje cells. (b) In the Miles–Lisberger model, floccular Purkinje cells encode an error signal which induces the modification of synapses onto target neurons in the vestibular nuclei. *Abbreviations:* CF, climbing fibers; GC, granule cells; MF, mossy fibers; OMN, oculomotor nuclei; Pc, Purkinje cells.

conditions that induce VOR motor learning, changes in the simple and complex spike activity of Purkinje cells are appropriate to drive these adaptive changes providing evidence for Ito's proposal of VOR motor learning. However, once a new VOR gain state has been established, the responses of FTNs in the vestibular nuclei are also modified.

The results of lesion studies have furthered the proposal that the site of VOR motor learning is not restricted to either the floccular complex or the vestibular nuclei. Specifically, changes in VOR gain are completely abolished by inactivation of the floccular complex immediately following learning, while inactivation of the floccular complex does not completely abolish long-term changes in VOR gain. Taken together, recording and lesion studies provide strong evidence that the gain changes required for VOR adaptation are initially stored in the floccular complex, and drive the formation of long-term synaptic modifications at the level of the vestibular nuclei such that long-term memory is chiefly consolidated in the brainstem rather than the cerebellum. This process has been termed the transfer or consolidation hypothesis (reviewed in Broussard and Kassardjian, 2004).

Cellular Mechanisms for VOR Motor Learning

As described above, there is general agreement that the synaptic plasticity required for VOR adaptation first occurs in the floccular complex, and the Purkinje cell response in turn drives the formation of long-term synaptic changes at the level of the vestibular nuclei. Much work has been devoted to investigating the cellular mechanisms of these synaptic changes in the floccular complex and vestibular nuclei. As is reviewed in articles "Long-term potentiation; Long-term depression," the two forms of synaptic plasticity are long-term depression (LTD) and long-term potentiation (LTP), which lead to a decrease and increase, respectively, in synaptic efficacy.

It has long been known that LTD occurs at the parallel fiber to Purkinje cell synapse when the climbing fibers and parallel fibers are simultaneously stimulated. More recent work has furthered this providing insight into the role of LTD in VOR adaptation (reviewed in Cullen, 2008). For example, blocking cerebellar LTD pharmacologically impairs VOR adaptation. In addition, transgenic mice expressing an inhibitor of protein kinase C, a molecular required for cerebellar LTD, exhibit impairments of VOR adaptation. There have also been reports on the importance of cerebellar LTP in VOR adaptation. Blocking the T-type channel, required for cerebellar LTP, in mice impaired VOR adaptation (Ly et al., 2013). Taken together, these studies suggest that both cerebellar LTD and LTP play a role during VOR motor learning.

Recently, *in vitro* studies have provided insight into the cellular mechanisms that mediate synaptic plasticity at the level of the vestibular nuclei. It has been shown that repetitive stimulation of the vestibular nerve can evoke LTD at the afferent-vestibular nuclei neuron synapse when paired with postsynaptic depolarization or LTP when paired with postsynaptic hyperpolarization (McElvain et al., 2010). LTP appears to depend on AMPA receptors whereas LTD requires NMDA receptors. Another complementary study has shown that either LTD or LTP can be evoked at the afferent-vestibular nuclei neuron synapse using repetitive stimulation of vestibular afferents depending on the pattern of pulse trains that are being delivered (Scarduzio et al., 2012). Short durations between pulse trains lead to LTP whereas long durations results in LTD. These results show that synaptic plasticity can occur within the vestibular nuclei and suggest that these changes, alongside synaptic modifications within the cerebellum reviewed above, likely play a role in guiding VOR motor learning.

VOR Adaptation and Treatment Options for Vestibular Disorders

Under normal conditions the VOR is effective (i.e., gain of unity) over the full range of head movements, which can easily reach accelerations and velocities of up to 25 000 deg s^{-2} and 500 deg s^{-1}, respectively. Immediately following unilateral lesion, however, VOR responses are

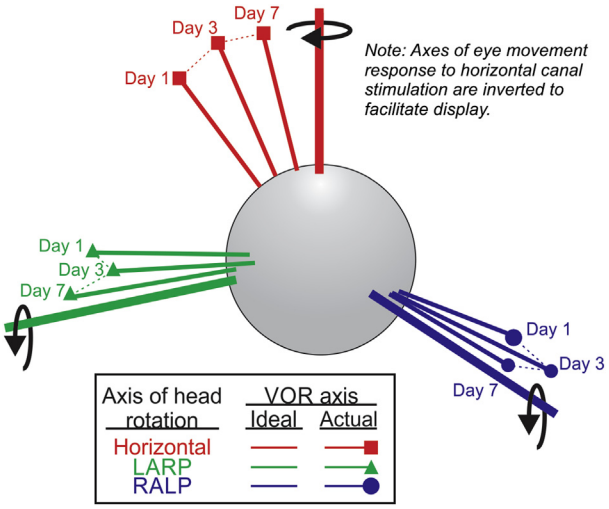

Figure 5 Changes in the axis of VOR eye movements induced with the vestibular prosthesis for a monkey over the course of 7 days of wearing the device. Each data vector depicts the axis and magnitude of VOR responses during sinusoidal rotations about the horizontal canal axis, and rotations in axes that stimulate the right anterior canal with the left posterior canal (RALP), and the left anterior canal with the right posterior canal (LARP) in push–pull pairs. Vector length indicates the peak response velocity and the direction of the vector indicates the axis of the VOR eye movements; for comparison, thick axes without markers depict ideal responses (i.e., VOR gain of 1). Modified from Dai, C., Fridman, G.Y., Chiang, B., Rahman, M.A., Ahn, J.H., Davidovics, N.S., Della Santina, C.C., 2013 Directional plasticity rapidly improves 3D vestibulo-ocular reflex alignment in monkeys using a multichannel vestibular prosthesis. Journal of the Association for Research in Otolaryngology 14, 863–877.

decreased, and patients experience gaze instability and accompanying vertigo. Fortunately, most patients exhibit remarkable recovery within 1 week for sinusoidal rotations with low frequencies and velocities (reviewed in Cullen et al., 2009). Yet, in response to more challenging stimuli (e.g., rapid rotations of up to 500 deg s^{-1}), compensation is not complete and VOR responses are asymmetric (VOR gains are ~0.7 and 0.3 for contralesional and ipsilesional-directed rotations, respectively).

As mentioned above, patients suffering from vestibular deficiency typically exhibit a unilateral vestibular hypofunction and exhibit asymmetric VOR responses. Therefore, training that would lead to increases in VOR responses during both ipsilesional and contralesional directed rotations would not be beneficial for these patients. One recent study has shown that the VOR is capable of unilateral adaptation (Migliaccio and Schubert, 2013). Following a unilateral VOR adaptation paradigm, researchers found that VOR gains toward the adapted side increased when tested during passively applied head movements, whereas VOR gains toward the non-adapted side did not. Clinicians can take advantage of these results to further the development of vestibular rehabilitation exercises.

Vestibular patients with bilateral loss of function suffer from debilitating symptoms for which there remain no adequate treatment options. Ongoing research is focused on developing a chronically implanted vestibular prosthesis as a treatment option for these patients. This device consists of a gyroscope, which senses the movement of the head and transmits this signal to electrode arrays implanted in each semicircular canal, which stimulate the vestibular nerve and sends this information to the rest of the brain. The vestibular prosthetic stimulation, like natural stimulation, can produce both reflexive eye and head movements (Mitchell et al., 2013). Furthermore, animals rendered bilaterally vestibular deficient and implanted with a prosthesis achieve partial restoration of the VOR responses, which improve when the device is worn for extended periods of time (Lewis et al., 2010). Finally, the misalignment of the VOR eye movement relative to the axis being stimulated decreases over the course of 1 week (**Figure 5**; Dai et al., 2013). The ability of the VOR neural circuitry to adapt to stimulation delivered by a vestibular prosthesis is promising for patients in need of treatment to relieve symptoms associated with bilateral vestibular deficiency.

See also: Eye Movement, Control of: Oculomotor Control; Long-Term Potentiation and Long-Term Depression; Vestibular System.

Bibliography

Broussard, D.M., Kassardjian, C.D., 2004. Learning in a simple motor system. Learning and Memory 11, 127–136.

Cullen, K.E., 2008. Procedural learning: VOR. In: Byrne, J.H. (Ed.), Learning and Memory: A Comprehensive Reference. Academic Press/Elsevier, Oxford, UK, pp. 383–402.

Cullen, K.E., 2012. The vestibular system: multimodal integration and encoding of self-motion for motor control. Trends in Neurosciences 35, 185–196.

Cullen, K.E., Minor, L.B., Beraneck, M., Sadeghi, S.G., 2009. Neural substrates underlying vestibular compensation: contribution of peripheral versus central processing. Journal of Vestibular Research 19 (5-6), 171–182.

Dai, C., Fridman, G.Y., Chiang, B., Rahman, M.A., Ahn, J.H., Davidovics, N.S., Della Santina, C.C., 2013. Directional plasticity rapidly improves 3D vestibulo-ocular reflex alignment in monkeys using a multichannel vestibular prosthesis. Journal of the Association for Research in Otolaryngology 14, 863–877.

Gonshor, A., Jones, G.M., 1976. Extreme vestibulo-ocular adaptation induced by prolonged optical reversal of vision. Journal of Physiology 256, 381–414.

Huterer, M., Cullen, K.E., 2002. Vestibuloocular reflex dynamics during high-frequency and high-acceleration rotations of the head on body in rhesus monkey. Journal of Neurophysiology 99, 13–28.

Lewis, R.F., Haburcakova, C., Gong, W., Makary, C., Merfeld, D.M., 2010. Vestibuloocular reflex adaptation investigated with chronic motion-modulated electrical stimulation of semicircular canal afferents. Journal of Neurophysiology 103, 1066–1079.

Ly, R., Bouvier, G., Schonewille, M., Arabo, A., Rondi-Reig, L., Léna, C., Casado, M., De Zeeuw, C.I., Feltz, A., 2013. T-type channel blockade impairs long-term potentiation at the parallel fiber-Purkinje cell synapse and cerebellar learning. Proceedings of the National Academy of Sciences of the United States of America 110, 20302–20307.

McElvain, L.E., Bagnall, M.W., Sakatos, A., du Lac, S., 2010. Bidirectional plasticity gated by hyperpolarization controls the gain of postsynaptic firing responses at central vestibular nerve synapses. Neuron 68, 763–775.

Migliaccio, A.A., Schubert, M.C., 2013. Unilateral adaptation of the human angular vestibulo-ocular reflex. Journal of the Association for Research in Otolaryngology 14, 29–36.

Mitchell, D.E., Dai, C., Rahman, M.A., Ahn, J.H., Della Santina, C.C., Cullen, K.E., 2013. Head movements evoked in alert rhesus monkey by vestibular prosthesis stimulation: implications for postural and gaze stabilization. PLoS One 8, e78767.

Scarduzio, M., Panichi, R., Pettorossi, V.E., Grassi, S., 2012. The repetition timing of high frequency afferent stimulation drives the bidirectional plasticity at central synapses in the rat medial vestibular nuclei. Neuroscience 223, 1–11.

Victimless Crime

Steven Block, Central Connecticut State University, New Britain, CT, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Victimless crimes are difficult to define and categorize, but they unarguably account for considerable attention and resources in legal discussion and criminal justice system processing. Crimes commonly labeled victimless include drug use, gambling, prostitution, pornography, adultery, and sodomy. These behaviors are normally legislated on one or more grounds: harm to others, offense to others, harm to self, or a violation of morals. Yet, critics suggest that society would benefit from decriminalization or legalization.

Most criminal acts involve identifiable victims who directly suffer as a result of the crime. In contrast, victimless crimes have no direct victims and are often the subject of intense public policy discussion. Drug use, gambling, prostitution, and pornography are among actions that are traditionally classified in this category (Veneziano and Veneziano, 1993). Whether other behaviors, such as driving an automobile under the influence of alcohol, committing tax fraud, and illegally possessing a firearm, are victimless crimes is debated. Although the word 'victimless' is rarely used in legal settings, the term is more commonly utilized to illuminate the debate about the role of government in this set of activities.

Definitions of victimless crime have been formulated since the 1960s. Packer (1968) described victimless crimes as "offenses that do not result in anyone's feeling that he has been injured so as to impel him to bring the offense to the attention of the authorities" (p.151). Similarly, Morris and Hawkins (1970) state that "these crimes lack victims in the sense of complainants asking for the protection of the criminal law" (p. 6). Earlier, Schur (1965) wrote that the term "refers essentially to the willing exchange, among adults, of strongly demanded, but legally proscribed goods or services" (p. 169). Other words are also used to describe similar, if not identical, groups of behaviors including vice (Skolnick, 1998), consensual (McWilliams, 1996), public order, and fringe crimes (Clement and Barbrey, 2008).

While there are few themes that connect all victimless crimes, many concepts separate the majority of victimless crimes from other groups of criminal behavior, such as personal violent and property crimes. Foremost, as indicated in the proposed definitions, victimless crimes are actions that do not have clearly identifiable victims. Violent, property, and white-collar crimes normally have victims who are likely to feel as their direct interests have been harmed, but victimless crimes contain only secondary victims or no victims at all.

Analyses of drug use, gambling, prostitution, and pornography involve questions about both harmfulness and morality. Each behavior has the potential to harm the actor and indirectly harm other people. Harms may be physical, financial, or psychological. Further, several victimless crimes also violate standards of morality for certain populations. Although morality is most often linked to sexual behavior, ethical and religious concerns are also raised about personal drug use and gambling. Morality is rarely discussed for other victimless crimes that relate to strictly personal safety, such as motor vehicle violations or jaywalking.

Another unique characteristic of victimless crime that separates this category from other offenses is the controversial legal status. Nearly all organized societies prohibit *mala in se* crimes such as sexual assault, robbery, and theft. Conversely, there is great division over the legal statuses of *mala prohibita* victimless crimes. This subjectivity is well represented by considering a historical account of laws against victimless crimes. Most countries have experienced time periods when laws have shifted for drug use, gambling, and prostitution.

A last theme of victimless crimes is that they are generally viewed by the public as being less serious than other forms of crime. Multiple studies have constructed rankings of crime seriousness ranging back to criminological research in the 1960s and 1970s. Consistently, various types of violence and property offenses are ranked as more serious than victimless crimes by respondents using different sample populations (Cullen et al., 1982; Rossi et al., 1974).

Justifications for Legislation

The philosophical debate about the role of law in private behavior began with John Stuart Mill's 'On Liberty' written in 1859. Mill assisted in creating the fundamental position of liberalism when he posited that the only justification for infringing on a citizen's individual liberties is when the interests of others are directly harmed. Mill's core arguments were eventually challenged by James Fitzjames Stephen who believed that society maintains the right to enforce laws, apart from harming others, under particular circumstances. The Mill–Stephen conflict laid the framework for two diverging viewpoints. The release of the United Kingdom's Wolfenden Report on prostitution and homosexuality in the 1950s reinvigorated the academic discussion about the role of harm and morality in law nearly a century after Stephen's response to Mill. Consequently, Lord Patrick Devlin and H.L.A. Hart engaged in debate about the proper scope of criminal law. Hart extended and clarified Mill's points in creating the idea of legal positivism, which challenged that any relationship should exist between law and morality. Devlin, echoing the words of Stephen, claimed that society's future relied upon a shared set of values. Further, he questioned whether a lack of morality would result in any less impact than acts violating loyalty toward a nation. Years later, Joel Feinberg (1984, 1985, 1986, 1988) produced

an authoritative four-volume series that meticulously outlined the four main justifications for criminal law. Although this philosophical debate persists today, legal scholars claim that the importance of morality has diminished with arguments from both liberal and conservative perspectives now considering whether specific acts involve enough harm to justify legislation (Harcourt, 1999). Research has shown that justifications for legal intervention toward victimless crimes, related to harm and morality, are also reflected in citizen opinions (Stylianou, 2010).

Harm to Others

Mill's sole justification for creating limitations on personal freedom was the harm principle. There is virtual consensus that if an act harms other people directly, this may be a justifiable reason for creating laws. However, several complex challenges emerge in the application of the harm principle. First, there are several forms of harm that go beyond physical ramifications. Other examples of harm include financial and emotional impact. A compulsive gambler might commit acts of violence to fund a habit, cause family bankruptcy with reckless wagering, and create tense relationships within the household. While all of these results are considered harmful, there is no agreement upon which forms of harm justify legal intervention. Second, it is uncertain whether harm must be a definite result caused by an action to be considered within this category of 'harm to others.' When an intoxicated driver violates drunk-driving laws, he or she increases the risk of harm for other individuals, but the actor does not always cause actual harm to other individuals. In most cases, no one suffers harm from this infraction, yet it is agreed that driving while intoxicated increases the likelihood that harm will occur.

Offense to Others

Some actions offend others in society without causing direct harm to their interests. Joel Feinberg (1985) expressed the idea that people other than the actor may be affected by an action without directly harming them. Victimless crime laws targeting public nudity and obscenity are more closely related to offensiveness than harmfulness. Several other offenses are viewed as a public nuisance, which, in many cases, serves as grounds for legal action. Yet, offensiveness is subjective, which leads to difficulties in determining which activities are offensive enough to warrant legal intervention. The behaviors that offend one individual do not necessarily correlate with the offenses that will offend other members of society.

Legal Paternalism

Beyond affecting other people and their interests, victimless crimes often impact the offender themselves. Although there may be indirect harm toward other people, laws requiring the use of motorcycle helmets and motor vehicle seatbelts chiefly protect individuals from harming themselves. Legal paternalism assumes that the government possesses the ability to make better decisions than individuals who may be swayed by emotion, excitement, and poor analytical skills. Feinberg (1986) divides paternalism into two categories: soft paternalism and hard paternalism. Soft paternalists, like Feinberg, argue that it is justifiable to make paternalistic laws when an individual makes a decision that is not entirely voluntary. Hard paternalists suggest that the protection of an individual alone, voluntary or not, is a sufficient reason to create a law. Critics of enforcing laws based on paternalistic grounds argue that individuals should have the right to engage in risky behaviors as long as it only harms their own interests. Further, several self-harming behaviors such as eating unhealthy foods, sun tanning, and climbing mountains are rarely considered for criminal legislation. One important complication to analyzing the paternalism doctrine is the perspective that acts that harm an actor do indirectly impact the lives of others in some manner. For instance, a drug habit influences family members emotionally, as well as the overall productivity of society, therefore harming all persons in that society indirectly.

Legal Moralism

Apart from harm, the legal moralism doctrine supposes that acts may be outlawed to enforce a consensus morality that exists within a society. This rationale for prohibiting victimless acts is perhaps the most controversial and was the primary concentration of the Mill–Stephen and Hart–Devlin debates. Questions of morality commonly focus on sexual acts such as sodomy, fornication, and adultery. Critics of legal moralism doubt whether it is just to use law to evaluate the moral systems of individuals. Further, in many instances no moral consensus emerges. In practice, the enforcement of morality often reflects the moral standards of persons in positions of power rather than all citizens within a society. The moral beliefs of minority populations are likely to be diminished and persecuted through law under legal moralism.

Perspectives on Victimless Crimes

Several unique political and social perspectives provide insight about the variety of approaches for handling victimless crimes. Beyond the fundamental liberal and conservative viewpoints, ideological perspectives on victimless acts include the libertarian movement, harm reduction, the 'broken windows' approach, and feminism. Libertarians support less governmental intrusion in private lives. Accordingly, the libertarian perspective on victimless crimes pushes for a minimization of criminal laws. In addition, libertarians do not support governments taking control of markets, but rather they support open markets that allow for individuals to exercise their own freedoms.

Another perspective on victimless crime is the harm reduction movement. Harm reductionists accept that people will engage in potentially risky activity and rejects 'zero tolerance' attitudes. Harm reductionists support systems that apply a public health approach that allows for treatment rather than criminal justice approaches that rely mainly on punishment. Additionally, education and awareness focuses on producing accurate information about safe engagement in these activities. Harm reduction supporters claim that zero tolerance approaches depend upon exaggeration and intimidation to deter participation.

Based on the assumption that community disorder leads to more serious criminality, the 'broken windows' perspective promotes strict enforcement of laws against quality of life offenses that occur in public and will portray a community in a negative manner (Kelling and Wilson, 1982). Several victimless crimes contribute to community disorder including street prostitution, obscenity, panhandling, graffiti, and jaywalking. Some cities in the United States, including New York City, used the 'broken windows' approach to change policing tactics during the 1990s. Critics of the 'broken windows' logic and implementation question whether there is an overuse of criminal justice resources. Others suggest that problematic activities are simply displaced to other settings when enforced in particular areas.

Many victimless crimes involve women or the portrayal of women. The feminist movement has been particularly vocal about prostitution and pornography. One branch of feminism supports tougher, stricter laws toward prostitution and pornography with a focus on eliminating them from society. Proponents of this viewpoint typically argue that gender roles are reinforced and that there is rarely a truly voluntary act that takes place. Rather, these feminists believe that coercion and abuse are rampant within any system of prostitution and pornography. Oppositely, another segment of feminism views sex work within prostitution and pornography as empowering. This side of the argument argues that when women act as the supply within the sex trade, they perform legitimate work that is in demand with a great amount of control that is rarely available in other industries for women.

Arguments for and against Prohibition

Proponents of legal restrictions for victimless crimes rely upon several separate arguments. Most centrally, an important logical assumption is that unwanted behaviors can be reduced or eliminated through prohibition. For some people, this reduction in the behavior is caused by a fear of legal repercussions. Other people may abstain from the behavior due to a respect for the role of law. Examples of declining cocaine and opium use following the Harrison Narcotics Act of 1914 and lower levels of alcohol use during and after alcohol prohibition in the 1920s and 1930s are examples of this success in the United States. The prohibition of a disapproved act also demonstrates total governmental condemnation of a particular behavior irrespective of whether the prevalence of the act declines.

The alleged connection between victimless crimes and other forms of predatory crime provides further justification for legal intervention for prohibitionists. Some victimless crimes possess a clear relation with secondary forms of crime. Prohibition supporters commonly rely on a causal link between the act in question and violence or property crimes. Suggested relationships include cocaine, amphetamines, and alcohol with violence (Parker and Auerhahn, 1998), sex work with physical and sexual assaults (Farley, 2004), gambling with income-generating crimes (Stitt et al., 2003), and pornography with forcible rape (Malamuth et al., 2000).

Even in cases where there is no apparent link with secondary harms, most victimless crimes are considered to be unproductive, because the behavior is recreational. Several drugs that lack a clear connection with serious crime, such as marijuana, are condemned for creating more apathetic users. Similarly, gambling is often critiqued as an exchange of money with no tangible, productive outcome for society.

A separate argument proposed by prohibitionists relates to accessibility for youth. Both prohibitionists and legalization supporters typically agree that children should not have access to participating in victimless acts. When policies are developed short of full prohibition, completely reducing access to children is unlikely. Most countries set a particular age at which young adults are permitted to use alcohol. Yet, these age-graded restrictions do not stop youth from accessing alcohol. Efforts also are enacted to reduce access to pornography, gambling establishments, and registered firearms, although full prohibition is argued to be the most aggressive approach for limiting access to youth.

The movement to remove criminal sanctions for victimless crimes also produces several multidimensional arguments. Some of these perspectives are based on the philosophical role of law in society, while others are formed around practical concerns. Foremost, it is suggested that creating legal restrictions for acts that do not directly harm other individuals is an undue infringement upon individual freedoms. These sentiments are rooted in the writings of John Stuart Mill and other liberal thinkers. In some countries, arguments are made that restrictions on personal freedoms violate constitutional protections, such as the 9th and 14th Amendments in the United States.

Labeling theorists suggest that interactions with the criminal justice system create an unnecessary stigma associated with being labeled a criminal or deviant. These labels have both a formal and informal impact on the offender and their ability to function within society. Because criminals face these difficulties, there is an increased likelihood that they will commit more serious crimes. Persons supporting decriminalization and legalization often believe that the removal of criminal sanctions for minor offenses will reduce further criminality by eliminating these labels.

Another argument in support of legalization is the difficulty associated with preventing citizens from engaging in private behaviors. In the case of drugs, like alcohol and tobacco, millions of users worldwide are addicted to these substances and prohibition alone does not address their addiction. Legislation that changes a behavior's status from being legal to illegal must also consider the ramifications of addiction and compulsion.

A related issue is the cost and difficulty associated with enforcing criminal laws. If victimless crime laws are enforced, offenders will be arrested, processed, and incarcerated which costs money to taxpayers. Based on the assumption that there is no complaining victim, law enforcement must resort to aggressive techniques including using paid informants, listening to wiretaps, searching suspect vehicles, and conducting potentially risky sting operations. These approaches are often time-consuming and dangerous. Proponents of legalization believe that these valuable police, court, and corrections system resources should be shifted toward more serious crimes with identifiable victims. In a recent year, over 500 000 prisoners were incarcerated in the United States alone for drug and public order offenses, according to the

Bureau of Justice Statistics (Carson and Sabol, 2011). The maintenance of these prisoners is expensive and most do not contribute productively to the workforce while incarcerated.

When crimes without victims are enforced, enticing opportunities for corruption emerge for criminal justice system practitioners. Critics suggest that a major downfall of alcohol prohibition in the United States was the corruption of persons responsible for enforcement. Without the lack of a complaining victim, police or other individuals in the justice system are presented with opportunities for bribery that do not exist with other offenses. Another related problem is selective policing and prosecution. Victimless crime enforcement allows for targeting of populations based on class, race, gender, age, or personal vendetta.

Legalization supporters suggest that illegal markets will develop when a prohibited commodity or good is in demand. Because participants in illegal markets do not have protection from the justice system and the ability to sue when there are disagreements or misunderstandings, violence ensues. In the country of Bhutan, lawmakers banned the sale of tobacco in 2004, but an illegal market reportedly emerged, as did violence during alcohol prohibition in the United States in the 1920s. In the case of drug violence, critics of victimless crime laws claim that violence associated with increasing income to purchase prohibited goods (economic compulsive) and violence related to illegal markets (systemic) could both be greatly reduced, if not eliminated, with legalization. In addition to promoting violence, illegal markets allow for individuals and organized crime to gain financial proceeds from illicit activity, while restricting governmental tax revenue.

The removal of criminal sanctions provides governments with the ability to control several aspects of a particular act. Where applicable, governments can tax, regulate use, and monitor use and distribution in ways that are not possible under a prohibition system. In the Netherlands, the creation of red light districts permits authorities to restrict the sale of sexual services to a specific area within cities. In the US state of Nevada, prostitution is often zoned to secluded areas where sex workers operate under strict regulations. The control of prostitution also results in lower levels of violence (Brents and Hausbeck, 2005). Additionally, the use of excise taxes, another form of control, has been shown to limit alcohol and drug consumption (Cook, 2008). Under a criminalization scheme, these forms of governmental control disappear.

Several scholars have weighed the arguments of both sides of the legalization debate and concluded that a more complex legal solution is needed beyond simple prohibition or legalization. One sociologist proposes creating a dualistic prostitution legal system that legalizes and controls indoor prostitution while outlawing street prostitution and enforcing such laws (Weitzer, 1999). Meanwhile, others have suggested that drug policies should rely on licensing and exclusion as a more moderate approach to prohibition (Leitzel, 2013).

Approaches to Reducing Victimless Crime

Strategies to reduce victimless crime normally focus on either the supply or demand associated with the illegal markets. Supply reduction techniques target the supply of drugs, prostitution, gambling rings, and pornography. Critics of supply reduction claim that such policies tend to be expensive and intrusive. Further, supply markets for drugs and prostitution rebuild and reconfigure even following significant interventions. In the case of drug trafficking organizations, arrests of drug carriers create opportunities for others to become involved in the trade. Likewise, consumers may switch drug allegiances or shift from street prostitution to indoor prostitution when supply markets are changed. Common controversial techniques targeting illicit markets include the eradication of drug-producing crops, civil forfeiture of assets, the creation of increased penalty zones, and various covert operations.

Other approaches focus on directly reducing the demand within illegal markets. These techniques include forms of prevention, punishment, and treatment. Both educational and awareness programs are commonly instituted to minimize drug use, prostitution, gambling, and other similar acts. Educational programs often involve schools or community programs to deliver messages, while awareness programs utilize several outlets including the media. Punishments toward the demand side, include, but are not limited to, criminal sanctions (e.g., probation, incarceration, fines) and formal restrictions on access to public services and voting rights. Under the category of treatment, rehabilitative techniques, such as peer-based 12-step programs and individual professional counseling, attempt to change established behavior.

Other policies reflect the harm reduction movement. Harm reduction focuses on limiting the actual harm that results from acts of drug use, gambling, and prostitution rather than solely attempting to eradicate the problem. The harm reduction movement is criticized for accepting and promoting negative behaviors, yet proponents cite evidence of lower sexually transmitted disease rates in prostitution and overdose levels for serious drug use when harm reduction policies are applied. In Switzerland, voters in 2008 approved the continuation of a pilot program to prescribe heroin to perpetually addicted users. According to one study, the measure positively impacted illicit heroin markets and reduced criminal activity associated with the illegal drug trade in the country (Killias and Aebi, 2000). Most innovative harm reduction initiatives are in Western European nations or Canada and public support for the movement appears strongest in these countries, as well.

The Future of Victimless Crime

Controversy about the role of government in private behavior has lasted for centuries. The first strong, organized movement to outlaw several victimless acts took place in the late nineteenth century. In the United States, state and federal laws were passed restricting personal drug use, prostitution and sex trafficking, and forms of gambling such as the lottery. Through the twentieth century, many laws against drugs, prostitution, and gambling were maintained or strengthened. By the 1960s and 1970s, a decriminalization movement occurred in the US grounded in a changing population and belief that prohibition caused more problems than it solved. Several drugs were decriminalized in individual states. Accordingly, state lotteries emerged, the presence of large casinos spread, and

pornography became more regulated in the US and across the world. Many other countries have similarly experienced a liberalization of several victimless crimes throughout the twentieth century.

Recent legal changes indicate that the shift toward a liberal philosophy continues around the world. Portugal became one of the first developed countries to decriminalize drug possession in 2001, instituting a system in which users were referred to treatment, opposed to traditional criminal prosecution. After a decade, these changes appeared to have positive effects on overall usage rates and accessibility to treatment (Greenwald, 2009). Other countries have followed the Portuguese model. The Argentinian Supreme Court ruled that laws prohibiting personal consumption of marijuana to be unconstitutional, because the state lacked the power to enforce morals. During elections in November of 2012, voters in Colorado and Washington in the US approved legalization measures for marijuana. These states became the first to legalize recreational marijuana use, although many others have approved marijuana use for medicinal purposes.

Prostitution approaches have also become more progressive in several countries. In many Western European nations, including Germany, the Netherlands, and Hungary, prostitution is now legalized and controlled. In Hungary, for instance, sex workers pay taxes and maintain appropriate documentation for their work. In Sweden, a 1999 law removed criminal laws targeting sex workers while maintaining the ability to arrest and prosecute consumers. Within the next decade, Norway and Iceland followed the Swedish model to decriminalize the sale of sex.

Drastic changes in victimless crime policy have also impacted gambling. Since the 1980s, casinos in the United States have expanded from two locations in Las Vegas, Nevada and Atlantic City, New Jersey to more than 35 states today. Government-sanctioned lotteries exist in many industrialized countries, including the United States and United Kingdom. Further, the progression of Internet-based gambling has increased the challenges associated with enforcing regulations and prohibitions.

Liberalization has also impacted laws that are rarely prosecuted. In a 2003 landmark US Supreme Court decision, *Lawrence v. Texas*, the Court majority ruled that laws prohibiting consensual sodomy were unconstitutional. Lower courts have applied the *Lawrence* finding to cases involving other consensual sex acts. Other countries have repealed laws against sodomy or had challenges to laws prohibiting nonreproductive sex. Early in the twenty-first century, many jurisdictions legalized assisted suicide, as well.

Despite this apparent shift in favor of liberalism, other changes to victimless crime laws have become more restrictive. In the US state of Rhode Island, lawmakers recriminalized indoor prostitution in 2009. Drug laws targeting several newer substances vary by state in the US with many states moving to criminalize drugs that are viewed as emerging problems. Further, international treaties bind nations together, which presents a challenge for countries considering legalizing particular drugs. Based on historical considerations and contemporary viewpoints, debates about the proper level of intervention for victimless crimes are likely to continue in the future as they have for more than 150 years across the world.

See also: Addictions: General Considerations; Crime: Organized; Drug Control: International Aspects; Drugs, Decriminalization of: Cultural Concerns; Legalization: The Expansion of Law's Domain; Mill, John Stuart (1806–73); Online Pornography; Pathological Gambling and Gambling Disorder; Prostitution and Sex Work.

Bibliography

Brents, B.G., Hausbeck, K., 2005. Violence and legalized brothel prostitution in Nevada: examining safety, risk, and prostitution policy. Journal of Interpersonal Violence 20 (3), 270–295.

Carson, E.A., Sabol, W.J., 2011. Prisoners in 2011. U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC.

Clement, K.E., Barbrey, J.W., 2008. Criminal laws on the fringe: an analysis of legislated punishments for morality crimes in the 50 states. Critical Criminology 16 (2), 105–121.

Cook, P., 2008. A free lunch. Journal of Drug Policy Analysis 1 (1), 1–5.

Cullen, F., Link, B.G., Polanzi, C.W., 1982. The seriousness of crime revisited. Criminology 20 (1), 83–102.

Farley, M., 2004. "Bad for the body, bad for the heart": prostitution harms women even if legalized or decriminalized. Violence Against Women 10 (10), 1087–1125.

Feinberg, J., 1984. Harm to Others. In: The Moral Limits of the Criminal Law, vol. 1. Oxford University Press, Oxford.

Feinberg, J., 1985. Offense to Others. In: The Moral Limits of the Criminal Law, vol. 2. Oxford University Press, Oxford.

Feinberg, J., 1986. Harm to Self. In: The Moral Limits of the Criminal Law, vol. 3. Oxford University Press, Oxford.

Feinberg, J., 1988. Harmless Wrongdoing. In: The Moral Limits of the Criminal Law, vol. 4. Oxford University Press, Oxford.

Greenwald, G., 2009. Drug Decriminalization in Portugal: Lessons for Creating Fair and Successful Drug Policies. The Cato Institute, Washington, DC.

Harcourt, B., 1999. The collapse of the harm principle. Journal of Criminal Law and Criminology 90 (1), 109–194.

Kelling, G., Wilson, J.Q., 1982. Broken windows. The policy and neighborhood safety. Atlantic Monthly 249 (3), 29–38.

Killias, M., Aebi, M.F., 2000. The impact of heroin prescription on heroin markets in Switzerland. In: Natarajan, M., Hough, M. (Eds.), From Illegal Drug Markets: From Research to Prevention Policy, Crime Prevention Studies, vol. 11. Criminal Justice Press, Monsey, NY, pp. 83–99.

Leitzel, J., 2013. Toward drug control: exclusion and buyer licensing. Criminal Law and Philosophy 7 (1), 99–119.

Malamuth, N.M., Addison, T., Koss, M., 2000. Pornography and sexual aggression: are there reliable effects and can we understand them? Annual Review of Sex Research 11 (1), 26–91.

McWilliams, P., 1996. Ain't Nobody's Business if You Do: The Absurdity of Consensual Crimes in Our Free Country. Prelude Press, Los Angeles, CA.

Morris, N., Hawkins, G.J., 1970. The Honest Politician's Guide to Crime Control. University of Chicago Press, Chicago, IL.

Packer, H., 1968. The Limits of the Criminal Sanction. Stanford University Press, Palo Alto, CA.

Parker, R.N., Auerhahn, K., 1998. Drugs, alcohol, and violence. Annual Review of Sociology 24, 291–311.

Rossi, P.H., Waite, E., Bose, C.E., 1974. The seriousness of crimes: normative structure and individual differences. American Sociological Review 39 (2), 224–237.

Schur, E.M., 1965. Crimes without Victims: Deviant Behavior and Public Policy-Abortion, Homosexuality, Drug Addiction. Prentice-Hall, Englewood Cliffs, NJ.

Skolnick, J., 1998. The social transformation of vice. Law and Contemporary Problems 51 (1), 9–29.

Stitt, B.G., Nichols, M., Giacopassi, D., 2003. Does the presence of casinos increase crime? An examination of casino and control communities. Crime & Delinquency 49 (2), 253–284.

Stylianou, S., 2010. Victimless deviance: toward a classification of opposition justifications. Western Criminology Review 11 (2), 43–56.

Veneziano, L., Veneziano, C., 1993. Are victimless crimes actually harmful? Journal of Contemporary Criminal Justice 9 (1), 1–14.

Weitzer, R., 1999. Prostitution control in America: rethinking public policy. Crime, Law and Social Change 32 (1), 83–102.

Video Cultures: Television

Denise D Bielby, University of California Santa Barbara, Santa Barbara, CA, USA
C Lee Harrington, Miami University, Oxford, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociological study of television has not kept pace with the new media environments that have developed from emerging digital-interactive technologies. As a result, sociology's theoretical approaches to understanding the cultures of television production, distribution, and consumption lag in their ability to account for the transformations that have taken place. To address this gap we discuss foundational work and recent developments on topics that are central to the study of television – media technologies and social life, audiences, production, and globalization. For each, we draw upon research from the fields of communication and media that are relevant to sociological concerns alongside scholarship by sociologists to understand the video cultures of contemporary television.

Introduction

Television was introduced as an experimental technology in the 1920s and 1930s in Europe, Asia, the former Soviet Union, and the Americas, and following the Second World War it rapidly became a widely accepted form of mass communication around the globe under a variety of national regulatory arrangements. Among early postwar adopters were France, who kept the production, distribution and broadcasting of television programs as a state monopoly in 1945; England, as a subsidized public service in 1946; the United States, as a commercial medium in 1948; Canada, as national system in 1952; Japan, in 1953; as dual public and commercial networks; and Australia, as a free and governmentally regulated broadcast service in 1956. Although television's long-term business cycle – its innovation and diffusion as a novel technology, establishment and system growth as a communications industry, maturation and popularity, and specialization and diversification – took decades to unfold (Cunningham, 2000), its widespread adoption as an important source of news, information, and entertainment quickly led to a multifaceted approach to its study. Initially, scholarly analysis focused on television in several definitive ways – as a cultural form, an industry, a technology, and as a medium shaped by national policy – and this eventually expanded to encompass its study as a global influence and its audience(s). Early attention by social scientists, especially among those in the United states, brought focus to this comprehensive vision by paying particular attention to television's social impact (see, for example, Bogart, 1956). (Bogart (1956) covered television's history, appeal, content, and consumption, effect on other media, relationship with advertisers, and influence on politics and youth.)

However, as Grindstaff and Turow (2006) have observed, since the 1950s, sociology's dominant conceptual approaches to the study of television have fallen short in two key ways. One is their inability to account for video cultures – the new media environments that develop from the "continually emerging sets of multifaceted digital-interactive technologies" (p. 103) – that have grown around contemporary television's production, distribution, and consumption. The other is the inadequacy of sociology's theoretical approaches to the study of television – the political economy perspective that largely attends to issues of power at a macrolevel along with how ownership and control of television and the organization of television production practices shape and influence content, and the cultural perspective that focuses more on the expressive and symbolic dimensions of television programming and reception – to keep up with the industry's transformation. As Grindstaff and Turow (2006) observe, "contemporary changes in the medium threaten to make past research on television appear quaint and anachronistic" (p. 103), but this theoretical lag should challenge scholars to find ways "to draw from enduring perspectives from the older work and assess how they apply to the new-media environment" (p. 103).

In this article we consider the issues raised by Grindstaff and Turow's (2006) thoughtful assessment of television sociology. We agree that there is a need to update existing theoretical approaches for capturing television's newer formulations, especially television's more extensive global presence on the one hand and its deeper penetration of everyday social life on the other. Over two decades ago, media scholar George Gerbner (1993) observed that "Mass media, particularly television, form the common mainstream of contemporary culture. They present a steady, repetitive, and compelling system of images. For the first time in human history, most of the stories are told to most of the children not by their parents, their schools, or their church but by a group of distant corporations with something to sell" (p. 207). This was echoed nearly a decade later when leading media sociologist Todd Gitlin stated that "the obvious but hard-to-grasp truth is that living with the media is today one of the main things Americans and many other human beings do" (2001: p. 5). More recently, this fact was underscored when leading communications scholar Sonia Livingstone challenged colleagues, in her 2008 presidential address to the International Communication Association, to more thoughtfully probe how "the media mediate everything, entering into and shaping the mundane yet significant relations among individuals and between individuals and society" (2009: p. xi). (In posing this challenge, Livingstone drew the distinction between mediation, "in which the mass media comprise one of many influential but independent institutions whose relations with the media can be usefully analyzed" and mediatization, which implies that "that all influential institutions in

society have themselves been transformed, reconstituted by contemporary processes of mediation" (2009: p. 2). While mediatization – the subordination of important social institutions to the media – remains an open question, mediation does not.) In short, with media now so unquestionably ubiquitous and so widely recognized as permeating every facet of social life, sociology is at a crucial juncture if it is to remain relevant as a contributor to the study of media in general and to the field of television, in particular.

To this end, we discuss foundational work and recent developments on topics that are central to the study of television – media technologies and social life, audiences, production, and globalization. For each, we draw upon research from the fields of communication and media that are relevant to sociological concerns alongside scholarship by sociologists to understand the video cultures of today's television.

Media Technologies and Social Life

Television's initial cultural setting was the household. Although there is scattered early ethnographic work about its presence in extensions of the domestic sphere such as taverns, hotels, and hospitals (e.g., May, 1999), television's 'ambience' was more thoroughly conceptualized when Anna McCarthy (2001) examined the personal uses of television in public spaces that included airports, train stations, retail outlets, and the like, outside the home. McCarthy's research enabled scholars to focus comparatively on the ways in which the medium interacted with its setting, and how that interplay shaped its use by viewers. McCarthy's conceptualization built upon the important scholarship by Lynn Spigel (1992) on television's contribution to the evolution of the post–Second World War (WWII) domestic sphere. Spigel's (1992) work demonstrated how the public's adoption of television as a domestic medium was influenced not just by the quest for audiences by large corporations, but also and equally importantly by the subjective responses to the technology by housewives of the 1950s whose then-prescribed social responsibility as homemakers made them gatekeepers of that domain. It was the response of housewives who ultimately shaped not only television's acceptability as a presence in the domestic realm, but equally importantly its cultural form, because in order for family gatekeepers to find the medium acceptable, its content had to be seen as relatable. (The marketing of the Princess telephone by Bell Telephone (Lupton, 1993) and of the cell phone by Nokia (Djelic and Ainamo, 2005) are other examples of how technological innovation, in order to become widely adopted, had to be subordinated to the cultural expectations of potential users before they could become instruments of sociable interaction.)

A consequence of the research on television's early adoption spawned the important line of scholarship on audiences, which we discuss below. However, it is first necessary to address the significance of how recent new technologies of distribution that include the secondary marketing of series on digital video disk (DVD) (and in the older physical format of videotape), TV subscription outlets like video-on-demand (VOD) and pay-per-view, and streaming over the Internet through other so-called over the top electronic services such as Netflix and Hulu have transformed access to the medium and are creating new forms of consumption. These forms of *a la carte* access were presciently anticipated by W. Russell Neuman in *The Future of the Mass Audience* (1991), in which he wrote of the future fragmentation of the mass audience and its fallout, niche programming. This (r)evolution in access is highly socially significant because it has implications for how we organize as a society around access, including as voluntary communities of like-minded individuals who share interests and tastes; how we conduct ourselves in public spaces, such as the workplace, that integrate the leisure/entertainment/information content that television offers; how consumption of the traditional serialized narrative structure of television has evolved into 'binge' or 'marathon' viewing wherein the watching of a series becomes like consuming a novel and allows for greater scrutiny of aesthetic quality (Chmielewski, 2013; McNamara, 2012); and the prevalence of so-called social TV that consists of real-time conversations happening online while viewing (Jurgensen, 2013). According to Jenkins (2006), these developments have so profoundly transformed the processes of media consumption that they have given rise to a "convergence culture" between media users and producers that has enhanced the extent to which audiences eagerly anticipate and avidly participate in developments in media narratives (Leaver, 2013; Milner, 2009). At the crux of this cultural shift is the development of what Jenkins calls a "transmedia impulse" for storytelling (2006: p. 133), a narrative "universe" that arises when audiences trace and combine plots that unfold across multiple media distribution formats (books, movies, comic books, games, and fan-generated knowledge and content). (Fans have always relied upon secondary and tertiary texts to supplement their reading of primary texts (see Fiske, 1987), and they have created a robust industry out of fan fiction (Alter, 2012), but Jenkins (2006: p. 98) argues that convergence culture relies upon collective intelligence on the part of the consumer and the audience to read an entertainment franchise or brand (i.e., a narrative universe) horizontally across formats and platforms. While Fiske's textual layers and Jenkins's convergence culture are clear adjuncts to the primary text that are partly audience-generated, Gray's (2008) concept of paratext regards promos, trailers, advertising, and the like as overt attempts by the industry to direct audience expectations of the primary text.)

Transformations in media consumption like these will continue to evolve, and they are readily sociologically contextualized. It is well known that marketers target consumers on age-based demographic characteristics, and while 18- to 49-year-olds are the single most important consumer market post–WWII (Turow, 1997), the child/youth ('tween') market has become equally important (because broadcast networks and advertisers seek to establish an early relationship with consumers), especially as an indicator of the adoption of new media technologies. Although the transformations in television consumption described above and the emergence of other online use patterns are presently more prevalent among online teens, GenY, GenX, and Younger Boomers relative to older age groups (Turow, 2011: p. 494), as the range of distribution technologies becomes more accessible, simplified, and understandable, consumption practices will eventually change for everyone. That said, recent statistics suggest that traditional

broadcast television will be around for some time to come. (Due to space limitations we are unable to address the growing 'posttelevision' scholarly literature which examines the future of traditional broadcast television in a world of convergence culture; see, for example, Spigel and Olsson (2004) and Lotz (2013).) Recent Nielsen data indicate that "teens consume the vast, vast majority of their video content via traditional television" (Juenger, quoted in Flint, 2013: p. B1), despite their ready adoption of new media formats and technologies. Finally, the introduction and adoption of new distribution technologies and transformation of consumption practices indicate there is still much for sociologists to learn about the impact of the social uses of these technologies on society and in everyday life. Cultural differences reflected by chronological age are often presumed (see Freese et al., 2006, for example), but recent analysis suggests chronological age is less relevant to the ways in which viewers relate to television content (Harrington and Bielby, 2010; Harrington et al., 2011). Finally, large questions still remain about the societal implications of the Internet, which is the genesis of this revolution. These include the relationship of the digital divide to inequality, the extent to which the Internet furthers social isolation or social capital and community formation, its beneficial or detrimental impact on politics, its impact on bureaucratic structure and worker control, and its potentially diversifying, hyper-segmenting, or massifying effect on societal culture (political elections and terrorism being equally compelling if diametrically opposed uses) (DiMaggio et al., 2001).

Audiences

The audience is crucial to the existence of television, not only as a market that creates demand for the shows advertisers subsidize through commercials, but also as the locus of interpretative activity that contributes to making the content of television programs culturally significant. As an industry, television is in the business of creating and selling programs, but from the industry's standpoint its real business is the creating and selling of audiences – as consumers who will buy what they see advertised, and as commodities to be marketed to advertisers who sponsor the shows. Interpretative activity is shaped by the audience's social activities and cultural identities as it seeks out and consumes particular programs. These two views of the relevance of the audience to television have created separate lines of scholarship from different disciplinary perspectives and only recently, through the study of fandom, are they in potential dialog with one other.

The measurement of the television audience that the industry relies upon to understand its market grew directly out of measurement techniques used in radio; these included the systematic measurement of audience size that yielded information about a potential market, and the measurement of program ratings that indexed a show's popularity (see, for example, Lasswell, 1948; Lazarsfeld, 1940). Because the concept of reaching the largest possible audience of potential consumers is pivotal to the success of commercial television, early research by sociologists on audiences focused on the concept of the mass audience, the broad reach that advertisers desired and that networks needed to establish themselves. Although then-influential scholarly work of the Frankfurt School aligned the notion of the mass audience with a vulnerability to persuasion by powerful social institutions, others challenged the presumptive notion of the audience as a mass of indistinguishable individuals, groups, or categories (Lang, 1957; Rosenberg and White, 1957) (Subsequent research has revealed how audience segmentation (e.g., social status, gender, age, employment, religion, and political affiliation) is associated with media use (Roe, 2000; van Eijck and van Rees, 2000; van Rees and van Eijck, 2003).), and still others who directly studied media effects challenged the presumption of persuasiveness, arguing that "media content was filtered by interpersonal networks, by individual need, and by selectivity in exposure and perception" (Katz, 1977: p. 22).

Issues of measurement precision have always been a matter of ongoing discussion between the industry and ratings companies, the leading one being A.C. Nielsen, because advertisers rely upon the accuracy of ratings data to determine advertising rates and where best to place their sponsorship. Recent innovations by Nielsen to the measurement of TV viewership to better capture the impact of new digital technology upon audience engagement include joining with Twitter, Inc., which is the leader in real-time television engagement, to create a new, separate industry metric "that will gauge penetration based entirely on data from the popular microblogging website and social network" (O'Connell, 2012). Another modification is Nielsen's decision to expand beyond the traditional TV definition of television by introducing a comprehensive plan to capture all video viewing including broadband, Xbox, and iPads (Block, 2013). A still unresolved debate between the networks and Nielsen is the networks' push for advertisers to pay for live ratings plus online and digital video recorder (DVR) viewing within 7 days, as opposed to the live ratings plus 3 days they pay for now (Bond, 2013). (DVRs, which are present in nearly half of all homes with a television, are primarily used to record scripted fare, not sports programming or awards shows. The television industry would like to see DVRs phased out and video-on-demand (VOD) become the predominant alternative viewing platform because VOD does not allow commercials to be skipped (Flint, 2012).) In short, viewers' use of digital technologies has changed the very definition of the practical ways in which audiences construct their viewership, and this has, in turn, transformed not only the commodification of the audience but also traditional dynamics between the industries of advertising, marketing, and television.

The study of fandom has opened up intriguing avenues of study for television scholars that align the characteristics of audiences with aspects of their engagement. Two decades ago, the study of *Star Trek* fans by Henry Jenkins (1992) and Camille Bacon-Smith (1992), and of soap opera fans by Harrington and Bielby (1995a), launched the first wave of fandom studies, the study of fans' interpretative communities. These foundational works, along with Joli Jensen's (1992) penetrating insights into cultural assumptions about fans in general, established the importance of studying fan cultures in order to understand the social legitimacy, meaning, and significance of intensive viewing practices. The wealth of research produced in the second wave of fandom studies documented how fandom is integrated with modern life (see Gray et al., 2007), and yet a great deal of work

remains to be done on fans' individual motivations, enjoyment, and pleasures, which comprise the third wave of scholarship on fandom. This entails, in particular, "furthering our understanding of how we form emotional bonds with ourselves and others in a modern, mediated world" through fan communities (2007: p. 10). (For example, new media technologies have transformed how fans record encounters with a celebrity. Fans' insistence on taking a photo of oneself with a famous person has all but replaced the standard artifact of the autograph (Villarreal, 2013). In another example, as ever more people spend more of their lives online, social media sites are now, also, becoming a public space for the conduct of traditional cultural rituals such as mourning the death and preserving the memory of beloved cultural figures (Khatchatourian, 2013).) Very recent scholarship on fandom's contribution to the social bonds that comprise everyday life, how individuals utilize fandom as a resource for navigating the boundaries of ascribed sociodemographic categories such as age, gender, race, and social origins, and even fandom's role in rewriting of the institutional structure, organization, and dynamics of the life course offer rich insight into the ever-deepening embeddedness of media institutions like television in the lives of individuals in the twenty-first century (Gatson and Reid, 2011; Harrington and Bielby, 2010; Harrington et al., 2011; Harrington and Brothers, 2010). However, as some have already noted, fandoms, whether they are virtual or face-to-face, are still stratified communities with cliques, hierarchies, and conflict (Pearson, 2010) and associated power dynamics (Harrington and Bielby, 1995b), as is the case in all societies.

Production

Guided by the production of culture perspective (Peterson and Anand, 2004), the sociological study of television as a system of industrial organization has yielded considerable insight into how television, as a culture industry, fabricates expressive-symbolic elements. In conjunction with the concepts of the art world (Becker, 1982), the culture world (Crane, 1992), and the fiction complex (Griswold, 2000), the production of culture perspective captures the organizational complexity of the television industry and its myriad institutional participants, such as policy makers who determine availability of content or products from other sources, professional critics who appraise product quality, and audiences whose taste preferences influence producers and shape product trends. Early work influenced by the production of culture approach focused primarily on elaborating the structure and dynamics of television's industrial systems per se (see, for example, Muriel Cantor's 1971 study of the television producer and Intintoli's 1984 ethnography of the production of soap operas). However, over time, scholars incorporated analysis of other pivotal aspects of the industry such as industry boundary spanners and gatekeepers who serve as intermediaries that link the industry's loosely coupled components, target innovation, and move products along (Hirsch, 1972), and industry executives' use of symbolic or ritualistic practices to manage the extreme uncertainty, risk, and ambiguity in audience preferences (Bielby and Bielby, 1994). As a whole, this body of work reveals that television strives to be a highly rationalized industry, but does not necessarily succeed at this goal.

The disconnect between the industry's rationalized aspects and its extreme business uncertainty opened the door to research on how the distinctive features of creative industries like television shape the subjective work experience of its employees. Important work by Hesmondhalgh (2007) (Hesmondhalgh and Baker, 2011) reveals that there is a chasm between scholars who celebrate creative jobs in the technology-driven economy, where there is flexibility, safety, autonomy, intrinsic satisfaction, name recognition, and the integration of conception and execution, and those who recognize that knowledge economy jobs are also marked by high levels of insecurity, casualization, and long working hours that result in self-exploitation. Research on localized production cultures among industry workers has specifically targeted the individual-level sense-making that creative workers undertake as they labor (Caldwell, 2008). Questions about sense-making become particularly relevant when studying workers who use information and communication technologies that allow the labor of production to be, on the one hand, dispersed across geographic locations, and, on the other, 'virtually' close transnational collaborations. How is collaboration enacted? How are artistic conventions understood? A recent line of scholarly work focuses on the analysis of change in television formats and production in the 'postnetwork' (i.e., mass/broadcast) era of multichannel, multiplatform television (Lotz, 2007), and its effect on traditional television formats such as the talk show (unchanged due to its segmented format that satisfies Internet-affected taste for 'video snacks') and the nightly news (in decline because of competition from on-demand availability of news around the clock). It remains to be seen how the study of localized production cultures in particular and the production of culture more generally will be affected as distinct lines of inquiry and as scholarly paradigms by the developments of the postnetwork era, in particular, as the demand for new job skills arises and established ones die out, whether a postbroadcast era will continue to translate into a wider range of desirable program options (pay cable networks have already vastly enhanced the content, artiness, and quality of serialized dramas(McNamara, 2013)), what the textual attributes and features of newer program options might be, and how this broadening of program options will affect the industry's labor market, and access to it, for talent behind the camera as well as in front (Ross, 2011).

Television's Globalization

Why has sociology lagged as an equal contributor to the analysis of these industry developments? In this section we discuss one last development in the television industry, its globalization. (Use of the term globalization here is in the sociological sense of the increasing interconnectedness of different parts of the world. Our use here refers specifically to cultural globalization, which Crane (2002: p. 1) defines as "the transmission or diffusion across national borders of various forms of media and the arts." See also Giddens (1990). We recognize the debates among globalization scholars about the adequacy of their own conceptualizations of culture and

globalization (see Bielby, 2010).) With few exceptions (Cantor and Cantor, 1986; Havens, 2006; Bielby and Harrington, 2008), not only has this topic received limited attention by sociologists, it reveals in precise ways where concerted theoretical work is needed in order for sociology to keep abreast with the industry's transformation.

The study of the global market for television was launched by the now-classic United Nations Educational, Scientific and Cultural Organization (UNESCO) study by Nordenstreng and Varis (1974), which provided one of the earliest attempts to systematically document the export of television programming. Their work uncovered a 'one-way flow' of television content from the United States, Britain, France, and the Federal Republic of Germany to the rest of the world, with the United States leading this group. Although their statistics revealed that a large number of hours of airtime were exported, it is now well understood that importing was deployed for a limited time by nations in the process of building their own emerging national television systems, and that as a strategy this proved to be far more cost effective in the short run than underwriting local productions to fill schedules and attract audiences. Subsequent findings by Varis (1984) provided equally interesting evidence that other countries were also major producers of programming for international distribution within regional markets; these included Mexico, which distributed throughout Latin America, and Lebanon and the United Arab Republic, who were major producers for the Middle East. Still others, like India and Brazil, had robust internal industries, which subsequently developed into vigorous export markets. Since the pioneering work by Nordenstreng and Varis, vital regional geolinguistic markets – such as Japan, which is a major supplier throughout Asia and South Asia and has a global reach that includes the West, Egypt, which supplies to the Arab world, India to the Indian populations in Africa and Asia, and Hong Kong and Taiwan to the Chinese-speaking population – have developed based on common cultural, linguistic, historical, and geographic similarities (Sinclair et al., 1996).

Statistics like those from UNESCO about television markets abroad were the basis for early and compelling assertions of US cultural imperialism by communication scholars (Schiller, 1976). A more recent (1990s) version of these concerns is the work by media scholars who study the internationalization of media conglomerates as purveyors of US cultural hegemony, global cultural homogenization, and cultural synchronization (Herman and McChesney, 1997; McChesney, 2004; Miller et al., 2001). Even more recently, there is renewed interest in programming content, which, it is argued, foregrounds neoliberal values such as consumption (Couldry, 2010); this is yet another iteration of a macro-oriented, top-down approach.

Arguments like these are often provocative and compelling, and they can be a helpful starting point for trying to understand key questions about the power of national media culture around the world and the landscape of diverse media markets. However, too often discussion of global media ends rather than starts at questions of imperialism (Tunstall, 1994). Their generally macrooriented, top-down analyses note the increasing power and scope of a small number of media corporations around the world; however, this is a problem because broad statistics about the distribution of shows or the ownership of conglomerates, although interesting and obvious, also locks us into assumptions and narrows our explanations and understanding of the global market. Important questions are missed when one focuses only on the broad picture of television trade flows or industry consolidation, especially in such an unpredictable market (Cantor and Cantor, 1986). For instance, although national industries have developed many business strategies that include coproductions and the selling of formats (TV show concepts that are repackaged to suit particular (national) markets and tastes (Freedman, 2003)) in an attempt to systematize revenue from abroad, even these arrangements are subject to the vagaries of potential buyers who must be responsive to cultural proscriptions and policies, as well as to audience preferences (Esser, 2010). Also left undetermined is why specific programs are marketed, adopted, or rejected by distributors. Even larger questions loom about how television programs are transformed at the local level in diverse international contexts of reception and interpretation.

This theoretical gap directs us to so-called middle-range approaches that bridge micro- and macroconcerns and target mesolevel conceptualization, evidence, and analysis (Sinclair et al., 1996). Mesolevel analysis attends to the concrete operational and cultural logics of global television markets: the features and operations of the marketplace rather than the market per se; the cyclicality of the forms and practices of the business of the industry rather than a presumed teleological development; industry participants who are located at all levels of organizational settings rather than just industry leaders or other highly visible decision makers; and the possibility that television series are malleable cultural products rather than immutable texts with inherent meanings (Bielby and Moloney, 2008). In short, it is necessary to study not only the production of television/video cultures, but also their distribution, and the contexts and mechanisms of their distribution, at many concrete levels (Bielby and Harrington, 2008). In sum, if a small number of huge media corporations now dominate the majority of media industry markets, this is a useful starting point for investigation of global television ownership and control and its consequences. But research must move beyond this important starting point. A useful research focus would be seeking understanding of the institutional logics of the organizations that make up the global marketplace for television and its video cultures, and targeting its mechanisms for managing market uncertainty in its business transactions, and, how local cultures affect these arrangements/solutions. This conceptual vantage point invites attention to organizations, participants, market features, discourses, locations, and properties of the cultural products produced (which pose a challenging management issue because they are so culturally/symbolically laden and their success or failure is so difficult to predict). More importantly, it would call for concerted attention to the locally specific ways in which such mechanisms are culturally understood, practiced, and engaged, (see Chung and Hamilton, 2001; Wherry, 2012). Finally, it would pay close attention to the consistency of findings that indicate reception by audiences is a crucial middle-range factor in this market (Gray, 2007; Kraidy, 2009; McMillin and Fisherkeller, 2009; Sanson, 2011).

Conclusion

The research reviewed here provides several challenges for future sociological research on television's video cultures. The theoretical gap noted by Grindstaff and Turow (2006) that consists of the distance between the now dominant approaches to the study of television – the political economy perspective and the cultural perspective – calls for concerted attention by scholars of culture industries. Such efforts call for a bolder approach to reading across disciplinary boundaries and attending to scholarly developments in cognate fields, while at the same time not losing sight of what is inherently sociological in that quest. For those still interested in macrolevel debates, a useful starting point might be the nuanced, encyclopedic synthesis and review of the vast body of evidence that is used to invoke notions of cultural imperialism and cultural globalization (Hesmondhalgh, 2007; Mirrlees, 2013). For those interested in the fertile opportunities for expanding conceptualization of the middle range of television's video cultures related to globalization and cultural production, there is useful work by cultural geographers who study regional/urban agglomerations of cultural production and distribution (see Scott, 2005, for example). There is also useful work by media studies scholars who examine how the institutional logics of politics and markets of other countries thwart television distribution (Curtin, 2005). For those interested in closing the theoretical gap in other ways lies the opportunity to explore the vastly overlooked topic of the properties of television as a cultural form (Bielby and Harrington, 2004; Mittell, 2004), the embeddedness of its form in everyday life (Lembo, 2000), and its critical appraisal and analysis (Bielby et al., 2005). Together, these topics made up an important part of early scholarship on television (see Boddy, 1990), which has, regrettably, been all but abandoned and yet remains pivotal to understanding television's ubiquity, popularity, societal penetration, and global presence. In conclusion, regardless of how or where one enters the gap, what is essential is building out from more established lines of thought and traditions and finding new challenges and opportunities for empirically grounded insights.

See also: Artists, Competition and Markets; Audience Measurement; Audiences, Media; Culture and Institutional Logics; Culture, Production of: Prospects for the Twenty-First Century; Globalization and World Culture; Internet and Culture; Narrative Networks; Networks and Cultural Consumption; Organizations and Culture; Television: General.

Bibliography

Alter, A., June 14, 2012. The weird world of fan fiction. Wall Street Journal. Available at: http://online.wsj.com/article/SB10001424052702303734204577464411825970488.html.

Bacon-Smith, C., 1992. Enterprising Women. University of Pennsylvania Press, Philadelphia, PA.

Becker, H., 1982. Art Worlds. University of California Press, Berkeley, CA.

Bielby, D., 2010. Globalization and cultural production. In: Hall, J., Grindstaff, L., Lo, M-C. (Eds.), Handbook of Cultural Sociology. Routledge, New York, NY, pp. 588–597.

Bielby, D., Harrington, C.L., 2004. Managing culture matters: genre, aesthetic elements, and the international market for exported television. Poetics 32 (1), 73–98.

Bielby, D., Harrington, C.L., 2008. Global TV: Exporting Television and Culture in the World Market. New York University Press, New York, NY.

Bielby, D., Moloney, M., 2008. Considering global media: sociological contributions. In: Rice, R. (Ed.), Media Ownership: Research and Regulation. Hampton Press, Cresskill, NJ, pp. 269–287.

Bielby, D., Moloney, M., Ngo, B., 2005. Aesthetics of television criticism: mapping critics' reviews in an era of industry transformation. In: Jones, C., Thornton, P. (Eds.), Research in the Sociology of Organizations: Transformations in Cultural Industries. JAI Press, Greenwich, CT, pp. 1–43.

Bielby, W., Bielby, D., 1994. 'All hits are flukes': institutionalized decision-making and the rhetoric of network prime-time program development. American Journal of Sociology 99, 1287–1313.

Block, A.B., February 20, 2013. Nielsen agrees to expand definition of TV viewing. Hollywood Reporter. Available at: http://www.hollywoodreporter.com/news/nielsen-agrees-expand-definition-tv-422795.

Boddy, W., 1990. Fifties Television: The Industry and Its Critics. University of Illinois Press, Urbana, IL.

Bogart, L., 1956. The Age of Television. Frederick Unger Publishing Company, New York, NY.

Bond, P., March 4, 2013. Les Moonves Declares DVR Viewing Should Be Counted 'Equally' in Ratings Measurements. Available at: http://www.hollywoodreporter.com/news/les-moonves-declares-dvr-viewing-425980.

Caldwell, J., 2008. Production Culture. Duke University Press, Durham, NC.

Cantor, M., 1971. The Hollywood TV Producer: His Work and His Audience. Basic Books, New York, NY.

Cantor, M., Cantor, J., 1986. American television in the international marketplace. Communication Research 13 (3), 509–520.

Chmielewski, D., 2013. Binge-viewing transforms TV. Los Angeles Times. A1, A10.

Chung, W-K., Hamilton, G., 2001. Social logic as business logic: guanxi, trustworthiness, and the embeddedness of Chinese business practices. In: Appelbaum, R., Felstiner, W., Gessner, V. (Eds.), Rules and Networks: The Legal Culture of Global Business Transactions. Hart, Oxford, England, pp. 325–346.

Couldry, N., 2010. Why Voice Matters: Culture and Politics after Neoliberalism. Sage, London, England.

Crane, D., 1992. The Production of Culture: Media and the Urban Arts. Sage, Newbury Park, CA.

Crane, D., 2002. Culture and globalization: theoretical models and emerging trends. In: Crane, D., Kawashima, N., Kawasaki, K. (Eds.), Global Cultures. Routledge, New York, NY, pp. 1–25.

Cunningham, S., 2000. History, context, politics, policy. In: Turner, G., Cunningham, S. (Eds.), The Australian TV Book. Allen & Unwin, St. Leonards, NSW, Australia, pp. 13–32.

Curtin, M., 2005. Murdoch's dilemma, or "What's the price of TV in China?". Media, Culture, and Society 27 (2), 155–175.

DiMaggio, P., Hargittai, E., Neuman, W.R., Robinson, J., 2001. Social implications of the internet. In: Annual Review of Sociology, vol. 27. Annual Reviews Press, Palo Alto, CA, pp. 307–336.

Djelic, M-L., Ainamo, A., 2005. The telecom industry as cultural industry? The transposition of fashion: logics into the field of mobile telephony. In: Jones, C., Thornton, P. (Eds.), Transformation in Cultural Industries. Elsevier, Ltd, Oxford, England, pp. 45–80.

Esser, A., 2010. Television formats: primetime staple, global market. Popular Communication 8 (4), 273–292.

Fiske, J., 1987. Television Culture. Methuen, London, England.

Flint, J., October 29, 2012. Higher DVR usage becomes a mixed blessing for TV industry. Los Angeles Times. A1, A9.

Flint, J., March 9, 2013. Teens still tune in on TV sets. Los Angeles Times. B1, B3.

Freedman, D., 2003. Who wants to be a millionaire? The politics of television exports. Information, Communication, and Society 6 (1), 24–41.

Freese, J., Rivas, S., Hargittai, E., 2006. Cognitive ability and internet use among older adults. Poetics 34, 236–249.

Gatson, S., Reid, R., 2011. Race and ethnicity in fandom. Transformative Works and Cultures 8. Accessed online at: http://journal.transformativeworks.org/index.php/twc.

Gerbner, G., 1993. Learning productive aging as a social role: the lessons of television. In: Bass, S., Caro, F., Chen, Y. (Eds.), Achieving a Productive Aging Society. Auburn House. Westport, CT, pp. 207–219.

Giddens, A., 1990. The Consequences of Modernity. Stanford University Press, Stanford, CA.

Gitlin, T., 2001. Media Unlimited. Metropolitan Books, New York, NY.

Gray, J., 2008. Television pre-views and the meaning of hype. International Journal of Cultural Studies 11 (1), 33–49.
Gray, J., 2007. Imagining America: the Simpsons go global. Popular Communication 5 (2), 129–148.
Gray, J., Sandvoss, C., Harrington, C.L., 2007. Fandom: Identities and Communities in a Mediated World. New York University Press, New York, NY.
Grindstaff, L., Turow, J., 2006. Video cultures: television sociology in the "new TV" age. Annual Review of Sociology, vol. 32 Annual Reviews Press, Palo Alto, CA, pp. 103–125.
Griswold, W., 2000. Bearing Witness: Readers, Writers, and the Novel in Nigeria. Princeton University Press, Princeton, NJ.
Harrington, C.L., Bielby, D., 1995a. Soap Fans. Temple University Press, Philadelphia, PA.
Harrington, C.L., Bielby, D., 1995b. "Where did you hear that?" Technology and the social organization of gossip. The Sociological Quarterly 36 (3), 607–628.
Harrington, C.L., Bielby, D., 2010. A life course perspective on fandom. International Journal of Cultural Studies 13 (5), 429–450.
Harrington, C.L., Bielby, D., Bardo, A., 2011. Life course transitions and the future of fandom. International Journal of Cultural Studies 14 (6), 567–590.
Harrington, C.L., Brothers, D., 2010. Constructing the older audience: age and aging on soaps. In: De Kosnik, A., Ford, S., Harrington, C.L. (Eds.), The Survival of Soap Operas: Strategies for a New Media Era. University Press of Mississippi, Jackson, MS, pp. 300–314.
Havens, T., 2006. Global Television Marketplace. BFI, London, England.
Herman, E., McChesney, R., 1997. The Global Media: The New Missionaries of Corporate Capitalism. Cassell, London, England.
Hesmondhalgh, D., 2007. The Cultural Industries, second ed. Sage, Los Angeles, CA.
Hesmondhalgh, D., Baker, S., 2011. Creative Labor: Media Work in Three Cultural Industries. Routledge, London, England.
Hirsch, P., 1972. Processing fads and fashions: an organization set analysis of culture industry systems. American Journal of Sociology 77 (4), 639–659.
Intintoli, J., 1984. Taking Soaps Seriously: The World of Guiding Light. Praeger, New York, NY.
Jenkins, H., 1992. Textual Poachers. Routledge, New York, NY.
Jenkins, H., 2006. Convergence Culture: Where Old and New Media Collide. New York University Press, New York, NY.
Jensen, J., 1992. Fandom as pathology: the consequence of characterization. In: Lewis, L. (Ed.), The Adoring Audience. Routledge, New York, NY, pp. 9–29.
Jurgensen, J., April 26, 2013. Can we please stop talking about TV? Wall Street Journal, D1–D2.
Katz, E., 1977. Social Research on Broadcasting: Proposals for Further Development. BBC, London, England.
Khatchatourian, M., July 31, 2013. The memory of Cory Monteith lives on (and on) via social media. Variety. Available at: http://variety.com/2013/digital/news/the-memory-of-cory-monteith-lives-on-and-on-on-social-media-1200569547/.
Kraidy, M., 2009. Reality television, gender, and authenticity in Saudi Arabia. Journal of Communication 59, 345–366.
Lang, K., 1957. Mass appeal and minority tastes. In: Rosenberg, B., White, D.M. (Eds.), Mass Culture: The Popular Arts in America. Free Press, Glencoe, IL, pp. 379–387.
Lasswell, H., 1948. The structure and function of communication in society. In: Bryson, L. (Ed.), The Communication of Ideas. Institute for Religious and Social Studies, New York, NY, pp. 37–51.
Lazarsfeld, P., 1940. Radio and the Printed Page. Duell, Sloan and Pearce, New York, NY.
Leaver, T., 2013. Joss Whedon, Dr. Horrible, and the future of web media. Popular Communication 11 (2), 160–173.
Lembo, R., 2000. Thinking through Television. Cambridge University Press, Cambridge, MA.
Livingstone, S., 2009. On the mediation of everything: ICA presidential address 2008. Journal of Communication 59 (1), 1–18.
Lotz, A., 2007. The Television Will Be Revolutionized. New York University Press, New York, NY.
Lotz, A., 2013. Television 2013. Cinema Journal 52 (3), 190–197.
Lupton, E., 1993. Mechanical Brides: Women and Machines from Home to Office. Cooper-Hewitt National Museum of Design. Smithsonian Institution, and Princeton Architectural Press, New York, New York.
May, R.A.B., 1999. Tavern culture and television viewing: the influence of local viewing culture on patrons' reception of television programs. Journal of Contemporary Ethnography 28, 69–99.
McCarthy, A., 2001. Ambient Television: Visual Culture and Public Space. Duke University Press, Durham, NC.
McChesney, R., 2004. The Problem of the Media: U.S. Communication Politics in the 21st Century. Monthly Review Press, New York.
McMillin, D., Fisherkeller, J., 2009. Local identities in globalized regions: teens, everyday life, and television. Popular Communication 7 (4), 237–251.
McNamara, M., January 15, 2012. The pros and cons of binge TV viewing. Los Angeles Times. D1, D27.
McNamara, M., June 27, 2013. Brett Martin and Alan Sepinwall look at the TV revolution. Los Angeles Times. E1, E10.
Miller, T., Govil, N., McMurria, J., Maxwell, R., 2001. Global Hollywood 2. British Film Institute, London, England.
Milner, R., 2009. Working for the text: fan labor and the new organization. International Journal of Cultural Studies 12 (5), 491–508.
Mirrlees, T., 2013. Global Entertainment Media: Between Cultural Imperialism and Cultural Globalization. Routledge, New York, NY.
Mittell, J., 2004. Genre and Television: From Cop Shows to Cartoons in American Culture. Routledge, New York: NY.
Neuman, W.R., 1991. The Future of the Mass Audience. Cambridge University Press, Cambridge, England.
Nordenstreng, K., Varis, T., 1974. Television Traffic – a One Way Street? Reports and Papers on Mass Communication No. 70. UNESCO, Paris.
O'Connell, M., December 17, 2012. Nielsen and Twitter partner for social media TV ratings. Hollywood Reporter. Available at: http://www.hollywoodreporter.com/live-feed/nielsen-twitter-partner-social-media-403360.
Pearson, R., 2010. Fandom in the digital era. Popular Communication 8 (1), 84–95.
Peterson, R., Anand, N., 2004. The production of culture perspective. Annual Review of Sociology. Annual Reviews, Palo Alto, CA.
Roe, K., 2000. Socioeconomic status and children's television use. Communications: The European Journal of Communication Research 25 (2), 3–18.
Rosenberg, B., White, D., 1957. Mass Culture: The Popular Arts in America. The Free Press, New York, NY.
Ross, S., 2011. Introduction: In focus: writing and producing TV in the post-network era. Cinema Journal 50 (2), 128–131.
Sanson, K., 2011. We don't want your must-see TV: transatlantic television and the failed "Coupling" format. Popular Communication 9 (1), 39–54.
Schiller, H., 1976. Communication and Cultural Domination. International Arts and Sciences Press, White Plains, NY.
Scott, A.J., 2005. On Hollywood. Princeton University Press, Princeton, NJ.
Sinclair, J., Jacka, E., Cunningham, S., 1996. New Patterns in Global Television: Peripheral Vision. Oxford University Press, Oxford, England.
Spigel, L., 1992. Make Room for TV: Television and the Family Ideal in Postwar America. University of Chicago Press, Chicago, IL.
Spigel, L., Olsson, J., 2004. Television after TV: Essays on a Medium in Transition. Duke University Press, Chapel Hill, NC.
Tunstall, J., 1994. The Media Are American, second ed. Constable, London, England.
Turow, J., 2011. Media Today: An Introduction to Mass Media, fourth ed. Routledge, New York, NY.
Turow, J., 1997. Breaking Up America: Advertisers and the New Media World. University of Chicago Press, Chicago, IL.
van Eijck, K., van Rees, K., 2000. Media orientation and media use: television viewing behavior of specific reader types from 1975 to 1995. Communication Research 27 (5), 574–616.
van Rees, K., van Eijck, K., 2003. Media repertoires of selective audiences. Poetics 31 (5), 465–490.
Varis, T., 1984. The international flow of television programs. Journal of Communication 34 (1), 143–152.
Villarreal, Y., January 30, 2013. Look, I really did meet…. Los Angeles Times. A1, A11.
Wherry, F., 2012. The Culture of Markets. Polity Press, Cambridge, England.

Vienna Circle: Logical Empiricism

Friedrich Stadler, Institute Vienna Circle (Institut Wiener Kreis), University of Vienna, Wien (Vienna), Austria

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Vienna Circle of Logical Empiricism or Logical Positivism, sometimes better known as 'Schlick-Circle,' is described as a philosophical movement between the two World Wars in Vienna with its main proponents and its theoretical program based on the manifesto of 1929 covering 'scientific philosophy' and 'scientific conception of the world.' On the one side, the international networking and acknowledgment is characterized, and at the same time the marginalization in Austria and Central Europe caused by the rise of Fascism and National Socialism, on the other, which resulted in the forced migration and placement in the Anglo-Saxon world. As one typical member of this movement Otto Neurath is characterized, who was also dealing with sociology, the social sciences including picture language, although the Vienna Circle is mainly known by its focus on logic, mathematics, and the natural sciences.

The Vienna Circle

The Vienna Circle (as the so-called Schlick Circle) consisted of a group of about 36 thinkers coming from the natural and social sciences, logic, and mathematics, which met regularly in Vienna between the two World Wars to discuss philosophy and its relation to the sciences. The work of this group constitutes one of the most important and most influential philosophical contributions of the twentieth century, in particular in the development of analytic philosophy and history and philosophy of science (Stadler, 2001, 2003a; Richardson and Uebel, 2007).

Prior to World War I, the predecessor of the later Vienna Circle (the 'First Vienna Circle') had begun to take shape both as an organization and as a philosophy (Uebel, 2000). Within a discussion circle (inter alia, with Frank, Hahn, and Neurath) at a coffeehouse, traditional 'academic philosophy' grew more scientific. This so-called 'First Vienna Circle' met regularly as of 1907 to discuss the synthesis of empiricism and symbolic logic as modeled after Mach, Boltzmann, and the French conventionalists (Pierre Duhem and Henri Poincaré). This was also seen as an indirect answer to Lenin's polemical remarks against Mach in his book, *Materialism and Empirio-Criticism* (1909), which remained very influential in Eastern Europe up to the Velvet Revolution of 1989–90.

This early phase in the development of logical empiricism can also be interpreted as an anti-Cartesian turn in epistemology and philosophy of science, which undermined both the synthetic a priori and the secure foundations of knowledge. In the middle of the permanent crisis of philosophy between reform and revolution in society and science, the further development of this 'scientific philosophy' had, in any case, been initiated.

The Vienna Circle was first publicly announced in 1929 with the publication of what came to be called its manifesto, *Wissenschaftliche Weltauffassung. Der Wiener Kreis* (*The Scientific Conception of the World: The Vienna Circle*), edited by the Verein Ernst Mach (Ernst Mach Society) and authored by Rudolf Carnap, Hans Hahn, and Otto Neurath (Carnap et al., 1929). This group was essentially a modernist movement, at the center of which was the so-called 'Schlick Circle,' a discussion group organized in 1924 by Moritz Schlick. Rudolf Carnap, Herbert Feigl, Philipp Frank, Kurt Gödel, Hans Hahn, Otto Neurath, Felix Kaufmann, Viktor Kraft, Karl Menger, Friedrich Waismann, and Edgar Zilsel belonged to its inner circle. Their meetings were also attended by Olga Taussky-Todd, Olga Hahn-Neurath, Rose Rand, Gustav Bergmann, and Richard von Mises, and on several occasions by visitors such as Hans Reichenbach, Alfred J. Ayer, Ernest Nagel, Willard Van Orman Quine, and Alfred Tarski. Members of the periphery, most of them as participants and occasional guests, were Egon Brunswik, Karl Bühler, Josef Frank, Else Frenkel-Brunswik, Heinrich Gomperz, Carl Gustav Hempel, Eino Kaila, Hans Kelsen, Charles Morris, Arne Naess, Karl Popper, W.V.O. Quine, Frank P. Ramsey, Kurt Reidemeister, and the alleged 'genius,' Ludwig Wittgenstein, who was not attending the Schlick Circle but only regularly with Schlick and Waismann and had a special influence on some members of the group. In addition, the mathematician Karl Menger organized in the years 1926–36 an international 'Mathematical Colloquium,' which was attended by Kurt Gödel, John von Neumann, and Alfred Tarski among many others (Menger, 1994).

This international and interdisciplinary discussion circle was pluralistic and committed to the ideals of the Enlightenment. It was unified by the aim of making philosophy scientific with the help of modern logic on the basis of experimental and everyday experience. The general aims of the movement were expressed in its publications such as the two book series, *Schriften zur Wissenschaftlichen Weltauffassung* (*Publications on the Scientific Conception of the World*) 1929–37 with 11 volumes, *Einheitswissenschaft* (*Unified Science*) 1933–38 with seven volumes, the journal *Erkenntnis* 1930–40 (the 1939 volume was called *Journal for Unified Science*), and the *International Encyclopedia of Unified Science* 1938–70 (Neurath et al., 1971).

Given this story of scholarly success, the fate of the Vienna Circle was tragic. The Verein Ernst Mach was suspended in 1934 by Austro-Fascism for political reasons; Schlick was murdered in 1936 by a fanatic student, and around this time, many members of the Circle were forced to leave Austria for racial and political reasons. Thus, soon after Schlick's death, the Circle disintegrated. As a result of the emigration of so many of its members and adherents, however, the Circle's ideas became more and more widely known, especially in

Northern Europe, Britain, and North America, where they contributed hugely to the emergence of modern philosophy of science (Timms and Hughes, 2003; Hardcastle and Richardson, 2003; Manninen and Stadler, 2010). In Germany and Austria, however, the break that was caused by the forced emigration of the Vienna Circle's members was felt in the philosophical and mathematical scene for a long time (Heidelberger and Stadler, 2003; Stadler, 2010a,b).

The name 'Vienna Circle' was suggested by Neurath and was supposed to have a pleasant connotation similar to the 'Vienna Woods' or the 'Viennese Waltz.' At the same time the term was to indicate the origin of this philosophical movement and its collective orientation (Frank, 1949). In the programmatic essay of 1929 the position of the antimetaphysical 'radical' left wing around Carnap, Neurath, Hahn, Frank, and others, was especially prominent. This group supported the idea of a physicalist unity of science, most commonly referred to as 'logical empiricism' as found later in the program of the *International Encyclopedia of Unified Science* (Neurath, 1945/1946). By contrast, the more moderate wing of the Vienna Circle around Schlick, Waismann, Feigl, and others, emphasized their adherence to a dualism of science and philosophy with different names like 'consistent empiricism' (Schlick, 1950) or 'logical positivism' (Kraft, 1950).

The widely used term 'logical positivism' actually originated in Blumberg and Feigl's (1931) article with that title which was published in the *Journal of Philosophy*. Here both authors gave a concise description of the new anti-Kantian synthesis of logical and empirical factors proclaiming the impossibility of synthetic a priori truths. They went on to describe the philosophical transformation from the old to the new positivism with the adoption of symbolic logic, epistemology, and research into the foundations of science. Finally, they explained, following Wittgenstein, "the purpose of philosophy as the clarification of the meaning of propositions and the elimination of … meaningless pseudo-propositions" (Blumberg and Feigl, 1931: 269).

Proponents of 'scientific philosophy' thought of philosophy not as an autonomous discipline existing prior to science but as a critical discipline dependent upon the results of the natural and social sciences, logic, and mathematics. Turning around Kant's dictum, they claimed that philosophy without science is empty, and science without philosophy is blind. Adoption of this scientific conception of philosophy does not, however, determine what epistemology, methodology, and ontology one is committed to. Nonetheless, all adherents of scientific philosophy demanded exact methods, a critical attitude, and a more or less empirical orientation. They opposed irrational and theological systems of philosophy (*Systemphilosophie*) and viewed science in general in a positive way.

Historically, Mach's philosophy provided the foundation for the development of the positions adopted within the Vienna Circle. The term 'logic of science' (Carnap's 'Wissenschaftslogik'; Carnap, 1934), known since the mid-1930s as 'philosophy of science,' was later used to describe these positions. This implied a general scientific conception of philosophy as well as an attempt to provide a philosophy for all sciences (including the humanities). In addition, within the Vienna Circle, philosophy was regarded both as a form of linguistic analysis and as a discipline drawing on the foundations of the natural and social sciences.

At the same time there were divergences of philosophical approaches within the Vienna Circle. Those such as Schlick defended a methodological dualism of philosophy and science and those such as Carnap and Neurath sought to integrate philosophy altogether within a scientific conception of the world or encyclopedic unity of the sciences. In Schlick's view, the classical philosophical positions of empiricism and rationalism were integrated with the help of modern logic and mathematics, but a distinction between philosophy and science still remained. Neurath's more radical physicalism or 'encyclopedism' of logical empiricism aimed at overcoming philosophy itself within his collective project of an *International Encyclopedia of the Unity of Science* (Neurath, 1945/1946). This divergence in philosophical approaches left room for debates within the Circle on such topics as the merits of phenomenalist and physicalist languages, coherence and correspondence theories of truth, logical syntax and semantics, verification and confirmation, and ideal and natural languages. At the same time there was a certain consensus on the merits of logical analysis of language, a fallibilist epistemology, a scientific attitude to the world, and the unity of scientific explanation and knowledge in general.

The rivalry between Schlick's 'consistent empiricism' and Neurath's physicalist (later empiricist) unified science is a complex matter. Certain views were shared by both, such as the view of philosophy as a critique of language in accordance with Wittgenstein's philosophy of the *Tractatus* of 1922. However, while the principle of verification, logical atomism, and the picture theory of language, are constitutive features of the entire movement, by themselves they do not characterize the Vienna Circle. Theoretical elements like logicism, verifiability, methodological phenomenalism and physicalism, a fallibilist theory of knowledge, conventionalism, and realism, together with an empiricist encyclopedism were cornerstones of the internal pluralistic development of logical empiricism from the 1930s onwards. This development also reflected the influence of Neurath's historico-pragmatic point of view within the Circle. In particular, the objection toward any dualism of 'language' and 'world' (as '*Wirklichkeitsphilosophie*') with the attendant denial of any absolute 'foundation of knowledge' (Schlick, 1934) is representative of this nonreductive naturalism and methodological holism in the spirit of Pierre Duhem's and Henri Poincaré's philosophy of science. This form of relativism and naturalism already anticipated the historical turn after World War II in the philosophy of science (with Kuhn, 1962) and which contributed to overcoming the linguistic turn and the so-called 'received view' of philosophy of science.

Nevertheless, the rejection of synthetic a priori judgments remained an important element of the Vienna Circle. According to Russell and Whitehead in the *Principia Mathematica*, symbolic logic and mathematics were regarded as purely analytic and a priori (independent of any experience). Analytic truths of these kinds were contrasted with contingent statements of the natural sciences and ordinary everyday experience, as synthetic *a posteriori* judgments. But there was no further class of synthetic a priori judgments: instead there was thought to be an important class of 'meaningless' sentences, without any cognitive content. The elements of this class were seen as 'metaphysical' in the sense that they are not part of knowledge at all

even though they may express some realm of common sense experience (as a moderate summary see von Mises, 1951/1990).

This position of the classical Vienna Circle is most prominently represented by Carnap's *Elimination of Metaphysics through Logical Analysis of Language* (1931), which developed a program for a unified rational reconstruction of science. But the question as to whether an empirical basis could serve as the foundation for all knowledge received strongly divergent answers from coherence theorists about truth influenced by Neurath and correspondence theorists influenced by Schlick (Hempel, 1993). Also, the apparently strict distinction between analytic and synthetic sentences was questioned (Menger, 1979: 1–60). The ideal of one language of science, logic, and mathematics was radically weakened within the Vienna Circle itself with Menger's and Carnap's principle of tolerance long before Quine (1953) put forward his critique of the *Two Dogmas of Empiricism*. Thus, contrary to popular belief, a heterogeneous pluralism of views was in fact characteristic of the Vienna Circle: for example, regarding ethics (Schlick, Menger, Kraft), the alternatives of realism *versus* positivism (Schlick, Carnap, Feigl, Kraft, Kaufmann), verificationism *versus* falsificationism (both positions criticized by Neurath, especially against Popper's *Logic of Scientific Discovery*), and last, but not least, in matters of ideological and political preference, for example, conservative liberalism versus leftist socialism. In the later period of the Vienna Circle the contested verification principle was gradually abandoned and replaced by some form of a probabilistic confirmation methodology based on the principle of 'connectibility' (von Mises, 1951).

The unity of science movement with its six 'International Congresses for the Unity of Science' (1935–41) and the ambitious publication project 'International Encyclopedia of Unified Science' (1938–70) had a broader cultural meaning and goal, most notably the attempt to improve the human condition and to promote social reform and the intellectual struggle against irrationalism and totalitarian *Weltanschauungen*. It was a manifestation of a late Enlightenment conception of science with a socially inspired antimetaphysics. Between the two World Wars metaphysics was seen as a correlative feature of German idealism as well as of (Austro-) fascist 'universalism' as represented by the economist Othmar Spann.

The practical impulse behind this therapeutic destruction of metaphysical systems, then, was the desire for a scientific attitude based on human experience, directed against the *Zeitgeist* of totalitarism and cultural pessimism (as criticized already by Neurath in 1921a and 1931). Therefore, traditional philosophy, first of all, had to be reduced to a critical analysis of language, because most proponents of logical empiricism thought that an exact and sober usage of the scientific language is a precondition for all problem-oriented philosophizing – and moreover a sort of moral obligation.

Social criticism and collective work in philosophy of science formed a programmatic unity striving for a sweeping improvement of the human condition. Whereas in the natural sciences considerable progress had already been made, the situation in the social and cultural sciences was not so transparent, and influenced by the ongoing *Methodenstreit* since the turn of the century (Kaufmann, 1936). Although some members of the Vienna Circle like Kaufmann, Neurath, and Zilsel contributed essentially to this neglected field, their publications have been largely ignored in the historiography on the Circle for a long time. In this respect it is worth mentioning that, after the disintegration of the Vienna Circle, its former members still occasionally made reference to the 'scientific conception of the world' when speaking about general ideological questions. For example, Carnap spoke about 'scientific humanism' as a view shared by the majority of the logical empiricists (Carnap, 1963: 81ff.).

After the dissolution of the Vienna Circle, the forced migration of most of its members and the dispersion of the logical empiricist movement from its centers in Central Europe, the twin aims of a transformation of philosophy and the establishment of philosophy of science could only be envisaged once the ties to their previous cultural context and audience had been severed. But even in these difficult times the proponents of the exiled Vienna Circle organized six well-attended prestigious international conferences, the 'International Congresses for the Unity of Science': Paris (1935 and 1937), Copenhagen (1936), Cambridge, UK (1938), Cambridge, Massachusetts (1939), and Chicago (1941). One can thus say that the demise of the Vienna Circle in the German-speaking world was accompanied by the transformation of Viennese '*Wissenschaftslogik*' into philosophy of science in the Anglo-Saxon scientific community converging with North American currents like (neo)pragmatism.

Looking at the current historiography on the Vienna Circle, one can quickly recognize the difficulty of providing a representative description of the Circle and of logical empiricism in its entirety. Even the autobiographical accounts of Vienna Circle members since the classical period of the Schlick Circle show a remarkable variance – depending on focus and underlying motivations.

What these texts have in common is the refutation of metaphysics as well as of philosophy as a discipline in its own right. As an alternative, one finds a tendency toward a (physicalist) unified science that later culminated in an empiricist encyclopedia project and that includes the principle of tolerance as applied to logic and scientific languages. Already here, the contours of epistemological options emerge. If one also takes into account that the manifesto represents only one variant of the Vienna Circle at the end of the 1920s, then it becomes amply clear that there existed only a limited consensus.

In addition, it is obvious that neither the autobiographical accounts of contemporaries nor the historical accounts originating shortly after 1945 were able to provide an adequate picture of the Vienna Circle. Moreover, there exists only a partial, albeit broad, overlap of the concept of Vienna Circle with that of logical empiricism in general when one takes into account the related movements of the Berlin Circle around Hans Reichenbach or the Warsaw Group around Alfred Tarski (Danneberg et al., 1994).

Despite this pluralism is it still possible to find a sort of basic agreement here – one that unites the members of the Vienna Circle – both the central figures and those on the periphery? First of all, it is a way of philosophizing based on linguistic analysis and a great amount of problem-oriented, open-ended discussion. This was experienced personally by Arne Naess, who focused several times on the Vienna Circle's 'thought style' which, in (not only) his opinion, leads to an

inherent "pluralism of tenable worldviews" (Naess, 2003). Secondly, the use of an unambiguous language, together with exact methods, is certainly a main legacy of the Circle and those associated with it. It is only when this exact formal approach is adopted that the content and positions can be constructively criticized and refuted – a characteristic which most current modern and postmodern philosophies lack.

The explicit and hidden history of the Vienna Circle from '*Wissenschaftslogik*' to recent philosophies of science documents the wide range, pluralism, and diversity of the its heritage and message. Be it called 'scientific philosophy' (as initiated by Schlick), 'scientific humanism' (according to Carnap), or a modern encyclopedism as a 'republic of scholars' (following Neurath), it is a guide to an intellectual journey, which continues through the present day and probably on into the future.

Otto Neurath – Encyclopedia and Utopia

The Vienna Circle is mainly known by its focus on logic, mathematics, and the natural sciences, though its effects on the social sciences are noteworthy. To illustrate the effects of the Vienna Circle on the social sciences, included here is a short biography of Otto Neurath and the impact of the Vienna Circle, through him, on Sociology.

Born 10 December 1882, in Vienna to the social reformer and political economist Wilhelm Neurath, Otto Neurath attended school in Vienna and later in Berlin (at the recommendation of Ferdinand Tönnies), where he studied mathematics, political economy, and history. In 1906, he completed his military service and from 1907 to 1914 he taught political economy at the Neue Wiener Handelsakademie (New Vienna Trade School). In 1907, Neurath married the social scientist and feminist Anna Schapire (1877–1911), who died from complications related to the birth of their son Paul. In 1912, he married the blind mathematician Olga Hahn and they jointly published mathematical articles before World War I. This period also encompassed the early philosophical and scientific activities in the 'First Vienna Circle' together with Hans Hahn and Philipp Frank, with whom he maintained private and scholarly contacts until the interwar years. Prior to 1914 he also published sociological studies, inter alia, on the so-called 'theory of war economy' and on the economical situation in the Balkans, to which he took a number of study trips with the support of a grant awarded by the Carnegie Foundation for International Peace.

Following the outbreak of World War I, Neurath served on the east front and in Vienna. He was appointed director of the German war economy museum in Leipzig and at same time he worked in the war economy section of the war ministry in Vienna. In this period Neurath developed his theory and practice of graphic representation of socioeconomic relations and his natural science models. He also completed his Habilitation and became private lecturer in political economy at the University of Heidelberg with Max Weber in 1917 – a position which, due to his war service and later his involvement in the Munich 'Räterepublik,' he was never to actively exercise.

In the revolutionary phase of the postwar period Neurath became head of the Central Planning Office in Bavaria, where as 'social engineer' he had futilely tried to carry out his vision of full socialization on the basis of an economy in kind in 1919. After the movement was suppressed, Neurath was sentenced to one and a half years imprisonment for 'aiding and abetting high treason.' After 6 weeks of detention – during which he wrote his *Anti-Spengler* book – he was able to travel to Austria on Otto Bauer's intervention, but he was no longer permitted to enter Germany (until 1926) and thus also lost his lectureship in Heidelberg.

After his return to Vienna, Neurath devoted his efforts to social economy and the settlement movement in Viennese communal politics. In 1920, he became secretary general of the organization for settlement and allotment movement. In 1923, he founded the Museum for Settlement and City Planning and, from this he developed the Social and Economic Museum in Vienna (1924–34). Here he, together with an interdisciplinary team, developed his 'Viennese method of pictorial statistics' as a highly effective tool for adult education. In addition to experts from various fields, he also collaborated with artists and constructivist graphic designers (most notably Gerd Anrtz). As an architect he was able to secure his old friend Josef Frank, and as collaborator Marie Reidermeister, the sister of the mathematician Kurt Rademeister. During the same period, he had contacts with the Bauhaus in Dessau. Still in Vienna, Neurath attempted, in view of the emerging fascism, to promote the internationalization of pictorial statistics by founding branches and institutions abroad. After he emigrated in the wake of the events of the Austrian civil war of 12 February 1934, the 'Foundation for Visual Education' in The Hague was to become the platform for the further activities of the encyclopedia movement and pictorial pedagogics. There, under the most difficult conditions, he continued his work on the project known as International System of Typographic Picture Education (Isotype). His wife Olga Hahn died in 1937 in Holland from complications after a kidney operation.

At the beginning of the 1920s, still during his Vienna years, Neurath resumed his philosophical work from before World War I and was actively involved in the Vienna Circle and further circles in Neurath's private apartment and at the Social and Economic Museum. Important highlights of these activities were his initiation of and collaboration on the manifesto *Wissenschaftliche Weltauffassung: Der Wiener Kreis* (Scientific World View: The Vienna Circle) in 1929, which he wrote with Hans Hahn and Rudolf Carnap, and the founding and organization of the Ernst Mach Society/Verein Ernst Mach (1928–34), the organ for the popularization of the Vienna Circle in connection with the Vienna school reform and adult education movement. Neurath was also active as a teacher at Vienna's Workers' University, in the labor union movement, and at diverse adult education institutes, in particular at the Ottakring 'Volksheim.' Finally, as the main representative of the so-called 'left wing' of the Vienna Circle, Neurath was the leading figure for public relations and – increasingly after 1934 – for the internationalization of logical empiricism (1935–41: with six International Congresses for the Unity of Science in Paris, Copenhagen, Cambridge, Harvard, Chicago). In this context, he founded the Institute for the Unity of Science in The Hague.

In 1940, after German troops invaded the country, Neurath undertook a second adventurous flight to England, where he was interned on the Isle of Man. He subsequently established

another Isotype Institute together with Marie Reidemeister, whom he married in 1941. From 1941 to 1945 he worked as a lecturer at the University of Oxford. He served as a consultant of the city of Bilston in a city renewal project and edited the book series *Unity of Science* and *Encyclopedia of Unified Science* from 1938 on.

Otto Neurath died of heart failure on 22 December 1945, in Oxford.

Otto Neurath was one of the last polymaths of the now lost old Austrian scientific culture, is slowly but continuously being rediscovered (Sandner, 2014). This process of rediscovery is taking place on an international and interdisciplinary level although it is still a selective process and happens at varied pace intervals. It seems not to be due merely to chance that this exercise in intellectual archaeology takes place in our era between Modernism and Postmodernism.

Neurath would, on principle, have rejected the title 'Universal Genius' which William Johnston awarded him (with the best of intentions) in his book *The Austrian Mind* (Johnston, 1972).

For Neurath knowledge was always a collective enterprise embedded in an all-encompassing social context. Throughout his life he never jettisoned his enlightened view of a social and cognitive totality. For him the discovery and explication of seemingly disparate connections between phenomena, both in terms of text and image, and the development of lines of argument based thereon was an essential precondition for all intellectual work. To view science as an end in itself was totally alien to him. He believed that knowledge – even as part of a hardly predictable historic process – should serve life in all the areas between the poles of everyday life and specialist research activity. This concept of reflective science will be surprising only to those who have been taken in by the clichéd images of the so-called *positivism debate* or that about the *one-dimensional society* (Marcuse, 1964; Dahms, 1994). Both these distortions absolutized individual elements of Logical Empiricism and used them to form a hostile portrait by dialectic magic. On the other hand, traditional academic metaphysics, obsessed with the subtleties of meaning and oblivious to the consequences of material being feels equally challenged by the radical and down-to-earth 'scientific world conception' clothed in the garb of the neo-Enlightenment, which Neurath promoted.

It is not easy to approach the complex and highly diverse life-work of Otto Neurath without getting lost in details, on the one hand, or without, on the other, resorting to inadequate generalizing descriptions such as 'positivist,' 'physicalist,' 'Austro-Marxist,' etc. How can one obtain an overview of a man, both of theory and practice, who covered the entire spectrum of knowledge including architecture, economic history, mathematics and logic, philosophy, history and theory of knowledge, visual education, and museology in about 300 publications (including around 30 monographs and books), and who worked in at least four countries (Austria, Germany, The Netherlands, England) as a social scientist and teacher at commercial colleges, workers' colleges, adult education schools, and at universities, a man who pursued a demanding profession (as museum director), yet also figured centrally in the Vienna Circle?

It seems that the answer to this question lies in Neurath's principle of thinking and acting: to proceed from incomplete and unstable elements temporarily united in theory and its related practice, and to employ, under the constraints of uncertainty and mere probability, the criteria of empiricism without ever losing sight of the fragmentary starting point or the contemplation of a historical process. This approach in itself constitutes a criticism of the idea of a fundamental and hierarchical *system* and at the same time an appeal to think in ways differing from those which prevail in society and the sciences, an appeal to risk a concrete *utopia* and to contribute to the promotion of the common good by means of cooperative *planning* for freedom and happiness. After World War I, Neurath envisaged an international 'republic of scholars,' much in the tradition of the French encyclopedists. Yet despite his optimism, he realized that the growth of Fascism and National Socialism threatened to end all these young democracies and he was not naive enough to believe that the proper proportion of rationalism and empiricism alone could reverse the 'demise of reason.' In his unpublished response to Max Horkheimer's *Der neueste Angriff auf die Metaphysik* (1937) (*The Most Recent Attack on Metaphysics*). Neurath, anticipating the main arguments against the dialectic of Enlightenment, referred to the limitations of social criticism which result from the fragility of knowledge including sociology and its history (Neurath, unpublished manuscript: 2): "What is not at all noticed at one time attracts the attention of other times and then can become the center of important considerations. […] for basically nothing is secure – everything is in flux." Moreover "we know no other authority beyond science which sits in judgement over science and studies its foundations," that is, there also is not available some form of transcendental reason. A central affinity to the standpoint of Wittgenstein's middle and late periods is unmistakable: philosophy is viewed as a language game, a game from which we cannot escape but about which we can certainly speak intelligently.

The basis for the remarkable rediscovery of Otto Neurath was laid some decades ago: in the English-speaking world the research began (motivated not least by Willard Van Orman Quine's repeated reference to the ship metaphor as an image of knowledge) with the edition of two selected volumes of Neurath's writings (1973 and 1983), whereas in the German-speaking world, besides one volume of selected works and a monograph, with the volumes of an edition of his writings; first those on philosophy and methodology (1981), then those on visual education (1991) and now those on the life and work and on economics (1996f).

In Austria an exhibition on his life and work together with a centenary symposium on Neurath (in conjunction with that on Moritz Schlick), held in 1982, augmented the new interest by the public and scholars. Topics which attracted this rediscovery ranged from his educational work to his economic theories, and his links to the architecture of the *Bauhaus* and the Viennese *Werkbundsiedlung* and to constructivist art (Haller, 1982; Stadler, 1982). A second main focus of interest moved toward the critical exploration of Neurath's theoretical potential and led to a reconstruction of the 'forgotten' Vienna Circle (Uebel, 1991, 1992, 2007). As a result of all this activities, we can now state that the depiction by others and even self-descriptions of 'logical positivism' after 1945 – especially of Neurath's role – were derived from appalling inadequacies and prejudices. For

many innovations of current history and philosophy of science were, in fact, anticipated in Neurath's oeuvre. The rediscovery of Neurath is therefore not merely a phenomenon of academic nostalgia, but itself constitutes research into the conditions and possibilities of changing a paradigm in the philosophy of science (see recently Cat, 2010).

The process of reassessment did not occur only within the narrow confines of philosophy and the theory of knowledge, but the most abstract structural context of Neurath's various subjects which, on first glance, seem all too disparate and range from the *Encyclopedia of Unified Science* to the picture language and his museology and visual education. From the start, Neurath himself had underlined the essential unity of his patterns of thought and of action. Not only are holism in philosophy of science (Duhem-Neurath-Quine principle) and naturalism extending to the "orchestration of the science by the encyclopedism of logical empiricism" (Neurath, 1945/1946) and their precise formulation pursued as systematically as his interest in visual education and the Viennese method of visual statistics and the *Isotype*. Loose but deliberate analogies can be found in Neurath's original writings, as can attempts to apply his theoretical concepts to education, social reform, and the politics of knowledge (Heinrich et al., 2011; Kraeutler, 2008; Neurath, 2013).

The following areas must be seen from this point of view: economy in kind, war economics, planning theory as well as his involvement in the Bavarian Revolution, the Viennese housing movement and town planning projects in England, not to mention his visual education projects and the founding of the 'Social and Economic Museum in Vienna' and of the 'International Foundation for Visual Education' in Holland (The Hague). Thus Neurath also emphasized several times, in a reference to Leibniz and Comenius, that the future encyclopedia did not aim merely at a standardization of the *language of science* but also at a standardization of the *visual representation* (he projected 26 volumes of text and a 10 volume 'visual thesaurus'). This combination of text with visual language had already been realized in part in his publications *Fernunterricht* (1931–33) and his successful book *Modern Man in the Making* (1939). Toward the end of his life Neurath referred to the 'mosaic of the sciences.' In the spirit of this formulation we can arrive at an understanding of his life's work by means of a kind of collage, employing the regulative idea of the unity of science and society – a term that is most appropriately used as the subtitle to the latest volume of selections from his work (P. Neurath and Nemeth, 1994). Neurath's phrase of a 'mosaic of sciences' also contains (long before C.P. Snow's well-known critique) an implied criticism of distinct (second and third) scientific cultures.

The biography of Otto Neurath is marked by the historical caesuras of 1918/19, 1934, and 1938/40, which brought major interruptions to his work. The fact that he nevertheless carried on between destruction and reconstruction under the most trying circumstances can offer encouragement and indicate that, even in a hostile environment the humanization of knowledge can make a modest contribution to the democratization of society (Nemeth and Stadler, 1996).

First, there is Neurath's vision of a joint reconstruction of the enterprise of sciences, which, poetically formulated, reappears in different variations throughout his work (Neurath, 1921b/1973: 199).

> We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.

Then there is the ethos with which Neurath pursued this reconstruction. In the middle of the revolutionary postwar phase, Neurath delivered an analysis of the *Zeitgeist* which, particularly in our present times of change, gives us reason to reflect (Neurath, 1919/1981: 137f.).

> How often has the word "utopian" been said with a tone of soft disdain or commiserating recognition! However stood on the ground of what happens to be the present and only knows it, sat in judgement over he who tried to glance behind the curtain of the stage of fate, who went to pains to prepare the game of the future. Now the curtain is being raised, clumsily the actors emerge from behind the scenes. Unknown figures meet unknown figures and unknown things. Some try to feign ease, others look at the prompt boxes and listen gratefully to the utterances coming from there. Rathenau, Atlanticus, Popper-Lynkeus and others are suddenly no longer simply 'utopians', but social engineers who were ahead of their times. From all corners we now hear catchwords and demands which we find in Fourier, Cabet, Bellamy, even in Thomas Morus or Plato. The utopia today are the only attempts at total constructions of social engineering at our disposal. Whoever does not want to meet completely unarmed the innumerable stimuli and endeavours breaking over us today, should read the utopia which until now has been dealt with only marginally by economics.

Finally, there is the consistency and continuity of his thought. In his intellectual testament, where Neurath (1946: 526f.) states

> I always promoted monism as a means of empiricist communication and I promoted pluralism as an attitude in making hypotheses. I have contended with thinkers of all kinds who tried to declare one system as marked out before others and therefore I tried to convey the insight that we need a kind of 'decision' wherever we have to make a "choice," even when we are trying a scientific theory [...]. One may evolve more than one theory of light starting from the same basis, as one may plan more than one holiday tour from the same starting point.

Note: This article is an abridged and updated version of my article on the Vienna Circle in: James Robert Brown, 2012; The general background for this entry in Stadler (2001) and in the publications of the Institute Vienna Circle: www.univie.ac.at/ivc.

See also: Feyerabend, Paul (1924–94); Hempel, Carl Gustav (1905–97); Lazarsfeld, Paul Felix (1901–76); Logical Positivism and Logical Empiricism; Popper, Karl Raimund (1902–94); Von Neumann, John (1903–57); Wittgenstein, Ludwig (1889–1951).

Bibliography

Blumberg, Albert, Feigl, Herbert, 1931. Logical positivism. A new movement in European philosophy. Journal of Philosophy 28, 281–296.

Brown, James Robert (Ed.), 2012. The Philosophy of Science: The Key Thinkers. Continuum, London and New York.

Carnap, Rudolf, 1931. Überwindung der Metaphysik durch logische Analyse der Sprache. Erkenntnis 2, 219–241.

Carnap, Rudolf, 1934. Logische Syntax der Sprache. Springer, Wien. English Translation: The Logical Syntax of Language. Routledge, London, 1937.

Carnap, Rudolf, 1963. Intellectual autobiography. In: Schilpp, Paul A. (Ed.), The Philosophy of Rudolf Carnap. Open Court, La Salle, IL, pp. 3–84.

Carnap, Rudolf, 1967. The Logical Structure of the World. University of California Press, Berkeley and Los Angeles.

Carnap, Rudolf, Hahn, Hans, Neurath, Otto, 1929. In: Ernst Mach, Verein (Ed.), Wissenschaftliche Weltauffassung. Der Wiener Kreis. Artur Wolf Verlag, Wien. Abridged English Translation as: "The Scientific Conception of the World: The Vienna Circle." In: Neurath, 299–318 (1973).

Cartwright, Nancy, Cat, Jordi, Uebel, Thomas (Eds.), 1996. Otto Neurath: Philosophy between Science and Politics. Cambridge University Press, Cambridge.

Cat, Jordi, 2010. "Neurath, Otto", Stanford Encyclopedia of Philosophy. http://plato.stanford.edu.

Dahms, Hans-Joachim, 1994. Positivismusstreit. Die Auseinandersetzungen der Frankfurter Schule mit dem logischen Positivismus, dem amerikanischen Pragmatismus und dem kritischen Rationalismus. Suhrkamp, Frankfurt/M.

Danneberg, Lutz, Kamlah, Andreas, Schäfer, Lothar (Eds.), 1994. Reichenbach und die Berliner Gruppe. Braunschweig-Wiesbaden.

Frank, Philipp, 1949. Modern Science and Its Philosophy. Cambridge University Press, Cambridge, MA.

Haller, Rudolf (Ed.), 1982. Schlick und Neurath – Ein Symposion. Rodopi (= Grazer Philosophische Studien 16/17), Amsterdam.

Hardcastle, Gary, Richardson, Alan W. (Eds.), 2003. Logical Empiricism in North America. University of Minnesota Press, Minneapolis.

Heidelberger, Michael, Stadler, Friedrich (Eds.), 2003. Wissenschaftsphilosophie und Politik/Philosophy of Science and Politics. Springer, Wien-New York.

Heinrich, Richard, Nemeth, Elisabeth, Pichler, Wolfram, Wagner, David (Eds.), 2011. Image and Imaging in Philosophy, Science and the Arts, vol. 2. Ontos, Frankfurt/M.

Hempel, Carl G., 1993. In: Stadler (Ed.), Empiricism in the Vienna Circle and in the Berlin Society for Scientific Philosophy. Recollections and Reflections, pp. 1–10.

Hofmann-Grüneberg, Frank, 1988. Radikal-empiristische Wahrheitstheorie. Eine Studie über Otto Neurath, den Wiener Kreis und das Wahrheitsproblem. Hölder-Pichler-Tempsky, Vienna.

Horkheimer, Max, 1937. Der neueste Angriff auf die Metaphysik. Zeitschrift für Sozialforschung VI (1).

James Robert Brown (Ed.), 2012. Philosophy of Science. The Key Thinkers. Continuum International Publishing Group, London.

Johnston, William, 1972. The Austrian Mind. An Intellectual and Social History 1848–1938. University of California Press, CA, Chapter 12.

Kaufmann, Felix, 1936. Methodenlehre der Sozialwissenschaften. Julius Springer, Wien. Reprint: Springer, Wien-New York, 1999.

Kraeutler, Hadwig, 2008. Otto Neurath. Museum and Exhibition Work. Peter Lang, Frankfurt/M.

Kraft, Viktor, 1950. Der Wiener Kreis. Der Ursprung des Neopositivismus. Springer, Wien-New York.

Kuhn, Thomas S., 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago and London. (= Foundations of the Unity of Science. Toward an International Encyclopedia of Unified Science. Otto Neurath, Rudolf Carnap, Charles Morris, (Eds.). second ed., enlarged II/2, 1970).

Kuipers, Theo (Ed.), 2007. Handbook of the Philosophy of Science: General Philosophy of Science – Focal Issues. Elsevier, Amsterdam etc.

Manninen, Juha, Stadler, Friedrich (Eds.), 2010. The Vienna Circle in the Nordic Countries. Networks and Transformations of Logical Empiricism. Springer, Wien-New York.

Marcuse, Herbert, 1964. The One-Dimensional Man. Studies in the Ideology of Advanced Industrial Society. Beacon Press, Boston, MA.

Menger, Karl, 1979. Selected Papers in Logic and Foundations, Didactics, Economics. Reidel, Dordrecht-Boston-London.

Menger, Karl, 1994. In: Galland, Louise, McGuinness, Brian, Sklar, Abe (Eds.), Reminiscences of the Vienna Circle and the Mathematical Colloquium. Kluwer, Dordrecht-Boston-London.

Mohn, Erich, 1977. Der logische Positivismus: Theorien und politische Praxis seiner Vertreter. Campus Verlag, Frankfurt-New York.

Müller, Karl H., 1991. Symbole, Statistik, Computerdesign. Otto Neuraths Bildpädagogik im Computerzeitalter. Hölder-Pichler Tempsky, Vienna.

Naess, Arne, 2003. Pluralism of Tenable Worldviews. In: Stadler (Ed.), The Vienna Circle and Logical Empiricism: Reevaluation and Future Perspectives. Kluwer, Dordrecht, Netherlands.

Nemeth, Elisabeth, 1981. Otto Neurath und der Wiener Kreis. Revolutionäre Wissenschaftlichkeit als politischer Anspruch. Campus, Frankfurt/New York.

Nemeth, Elisabeth, Stadler, Friedrich (Eds.), 1996. Encyclopedia and Utopia. The Life and Work of Otto Neurath (1882–1945). Kluwer, Dordrecht-Boston-London.

Nemeth, Elisabeth, Schmitz, Stefan W., Uebel, Thomas (Eds.), 2007. Otto Neurath's Economics in Context. Springer, Dordrecht.

Neurath, Otto, 1919/1981. Utopien. Dresden. In: Wirtschaft und Lebensordnung. Reprint: Neurath, Gesammelte philosophische und methodologische Schriften, 137f.

Neurath, Otto, 1921a. Anti-Spengler. Callwey, München. English in Neurath (1973), 158–213.

Neurath, Otto, 1921b. Anti-Spengler. Georg Callwey, Munich. Reprint in: Neurath (1973) and in Neurath (1981), 184 and 199.

Neurath, Otto, 1931. Empirische Soziologie. Der wissenschaftliche Gehalt der Geschichte und Nationalökonomie. Springer, Wien. English in Neurath (1973), 319–421.

Neurath, Otto. Einheitswissenschaft und Logischer Empirismus. Eine Erwiderung. Vienna Circle Foundation, Rijksarchief Noord Haarlem (NL), p. 2, unpublished manuscript.

Neurath, Otto, 1939. Modern Man in the Making. Alfred Knopf, New York.

Neurath, Otto, 1946/1946. The orchestration of the sciences by the encyclopedia of logical empiricism. Philosophy and Phenomenological Research 6 (4), 496–508.

Neurath, Otto, 1946. For the discussion: just annotations, not a reply. Philosophy and Phenomenological Research VI, 526f.

Neurath, Otto, 1973. In: Neurath, Marie, Cohen, Robert S. (Eds.), Empiricism and Sociology. Reidel, Dordrecht-Boston.

Neurath, Otto, 1979. In: Hegselmann, Rainer (Ed.), Wissenschaftliche Weltauffassung, Sozialismus und Logischer Empirismus. Suhrkamp, Frankfurt/M.

Neurath, Otto, 1981. In: Haller, Rudolf, Kinross, Robin (Eds.), Gesammelte philosophische und methodologische Schriften. (GPMS), 2 vols. Hölder-Pichler-Tempsky, Vienna.

Neurath, Otto, 1983. In: Cohen, Robert S., Neurath, Marie (Eds.), Philosophical Papers 1913–1946. Reidel, Dordrecht-Boston-Lancaster.

Neurath, Otto, 1991. In: Haller, Rudolf, Kinross, Robin (Eds.), Gesammelte bildpädagogische Schriften (GBS). Hölder-Pichler-Tempsky, Vienna.

Neurath, Otto, 2004. In: Uebel, Thomas E., Cohen, Robert S. (Eds.), Economic Writings. Selections 1904–1945. Kluwer, Dordrecht-Boston-London.

Neurath, Otto, 2013. In: Eve, Matthew, Burke, Christopher (Eds.), From Hieroglyphics to Isotype: A Visual Autobiography. Hyphen Press, London.

Neurath, Paul, Nemeth, Elisabeth (Eds.), 1994. Otto Neurath oder die Einheit von Wissenschaft und Gesellschaft. Böhlau, Vienna-Cologne-Weimar.

Neurath, Otto, Carnap, Rudolf, Morris, Charles (Eds.), 1938ff./1971. Foundations of the Unity of Science. Toward an International Encyclopedia of Unified Science, 2 vols. The University of Chicago Press, Chicago-London.

Neurath, Otto, Carnap, Rudolf, Morris, Charles (Eds.), 1971. Foundations of the unity of science. Toward an International Encyclopedia of Unified Science, volume I, nos. 1–10, volume II, nos. 1–9. Chicago University Press, Chicago-London.

Quine, Willard Van Orman, 1953. From a Logical Point of View. 9 Logico-Philosophical Essays. Harvard University Press, Cambridge, MA.

Richardson, Alan, Uebel, Thomas (Eds.), 2007. The Cambridge Companion to Logical Empiricism. Cambridge University Press.

Sandner, Günther, 2014. Otto Neurath. Eine Politische Biographie. Zsolnay, Wien.

Schlick, Moritz, 1918/1925. General Theory of Knowledge (Albert E. Blumberg, Trans.). Springer, Wien-New York, 1974.

Schlick, Moritz, 1934. Über das Fundament der Erkenntnis. Erkenntnis 4, 79–99.

Schlick, Moritz, 1950. Moritz Schlick. In: Ziegenfuss, Werner, Jung, Gertrud (Eds.), Philosophen-Lexikon. de Gruyter, Berlin.

Stadler, Friedrich. History of the Philosophy of Science. From Wissenschaftslogik (Logic of Science) to Philosophy of Science: Europe and America, 1930–1960. In: Kuipers (Ed.), pp. 577–658.

Stadler, Friedrich (Ed.), 1982. Arbeiterbildung in der Zwischenkriegszeit. Otto Neurath – Gerd Arntz. Löcker, Vienna-Munich.

Stadler, Friedrich (Ed.), 1993. Scientific Philosophy: Origins and Developments. Kluwer, Dordrecht-Boston-London.

Stadler, Friedrich, 2001. The Vienna Circle. Studies in the Origins, Development, and Influence of Logical Empiricism. With the first publication of the protocols (1930–1) of the Vienna Circle and an interview with Sir Karl Popper (1991). Springer, Wien-New York.

Stadler, Friedrich, 2002. The emigration and exile of Austrian intellectuals. In: Steininger, Rolf, Bischof, Günter, Gehler, Michael (Eds.), Austria in the Twentieth Century. Transaction Publishers, New Brunswick and London, pp. 116–136.

Stadler, Friedrich (Ed.), 2003a. The Vienna Circle and Logical Empiricism. Reevaluation and Future Perspectives. Kluwer, Dordrecht-Boston-London.

Stadler, Friedrich, 2003b. The 'Wiener Kreis'. In: Timms, Hughes (Eds.), Great Britain: Emigration and Interaction in the Philosophy of Science, pp. 155–180.

Stadler, Friedrich, 2003c. In: Hardcastle, Richardson (Eds.), Transfer and Transformation of Logical Empiricism: Quantitative and Qualitative Aspects, pp. 216–233.

Stadler, Friedrich (Ed.), 2010a. The Present Situation in the Philosophy of Science. Springer, Dordrecht-Heidelberg-London-New York.

Stadler, Friedrich (Ed.), 2010b. Vertreibung, Transformation und Rückkehr der Wissenschaftstheorie. Am Beispel von Rudolf Carnap und Wolfgang Stegmüller. LIT Verlag, Wien-Berlin.

Stadler, Friedrich, Weibel, Peter (Eds.), 1995. The Cultural Exodus from Austria. Springer, Wien-New York.

Timms, Edward, Hughes, Jon (Eds.), 2003. Intellectual Migration and Cultural Transformation. Refugees from National Socialism in the English-speaking World. Springer, Wien-New York.

Uebel, Thomas (Ed.), 1991. Rediscovering the Forgotten Vienna Circle. Austrian Studies on Otto Neurath and the Vienna Circle. Kluwer, Dordrecht-Boston-London.

Uebel, Thomas, 1992. Overcoming Logical Positivism from within. The Emergence of Neurath's Naturalism in the Vienna Circle's Protocol Sentence Debate. Rodopi, Amsterdam-Atlanta.

Uebel, Thomas E., 2000. Vernunftkritik und Wissenschaft. Otto Neurath und der Erste Wiener Kreis. Springer, Wien-New York.

Uebel, Thomas, 2007. Empiricism at the Crossroads. The Vienna Circle's Protocol-Sentence Debate. Open Court, Chicago and La Salle, IL.

von Mises, Richard, 1951. Positivism. A Study in Human Understanding. Dover, New York.

Vietnamese Revolution, The

Hy Van Luong, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The Vietnamese revolution combined a nationalism-inspired struggle against Western powers and a collectivism-oriented socioeconomic transformation within Vietnam. It reached its peak from 1945 to 1986. This article provides a historical background on the Vietnamese revolution, the processes of revolutionary transformation in Vietnam since 1945, and the theoretical debate in Western social sciences on the Vietnamese revolution.

In one and a quarter century since the French conquest of southern Vietnam in the 1860s, Vietnam underwent a radical transformation from a Confucian monarchy to a Marxist–Leninist state. The Vietnamese revolution involved a violent clash with two major Western powers, France and the United States, as well as a fundamental transformation of the Vietnamese socioeconomic system.

Historical Background

Citing the Vietnamese monarchy's persecution of Catholics and refusal of a trade request, France launched a series of attacks on Vietnam from 1858 onward. By 1867, France had seized as a direct colony the frontier southern third of Vietnam (Cochinchina or *Nam bộ*) from the Vietnamese monarchy. By 1884, northern and central Vietnam (respectively Tonkin or *Bắc bộ*, and Annam or *Trung bộ*) had fallen under French control, although the Nguyen monarchy nominally retained its direct administration of the latter.

In the aftermath of the French conquest, the resistance to French colonialism, mainly in the name of an 'Aid-the-King' movement, was led primarily by Confucian literati. It lasted the longest in central and northern Vietnam. By the time the Union of French Indochina was established in 1897, with Laos and Cambodia as the remaining two of the five components, the armed anticolonial resistance led by Confucian literati had virtually collapsed (see Marr, 1971).

Many members of the Vietnamese elite collaborated with the French rulers, while a number of others searched for a more effective path of resistance, strongly influenced by Chinese reformist and revolutionary thoughts and inspired by the rise of Japan as an Asian power. At one end of the collaboration–resistance spectrum in the first quarter of the twentieth century was the Confucian scholar Phan Bội Châu, a native of northern Central Vietnam, who organized a movement of students to Japan and inspired isolated armed resistance against French colonialism (Marr, 1971). In the context of strong French repression, Thailand and southern China served as important bases for exiled resistance leaders (Goscha, 1999). At the other end were many collaborators, such as the northern neotraditionalist Phạm Quỳnh, as well as many southerners who were rewarded with large tracts of land sold at low prices as the cultivated acreage in Cochinchina increased five times from 1870 to 1920 thanks to water control efforts by the French. Many of these southerners acquired French citizenship, experimented with the more open politics of the direct French colony of Cochinchina, or even played a role in the southern millenarian movements of Cao Đài and Hòa Hảo (Tai, 1983, 1992), coopting many poor peasants in the sharply stratified Cochinchina society. With the last Confucian examination being held in Vietnam in 1919, a new Western-educated elite began playing a salient role in the anticolonial resistance from 1925 onward. The major uprisings took place in northern Vietnam in 1930 under the leadership of the Vietnamese Nationalist Party (VNP), and in central and northern Vietnam in 1930–31 under the Indochinese Communist Party (ICP). The ICP-led movement in 1930–31 represented the most serious challenge to French colonialism since the Aid-the-King movement of the nineteenth century, as it involved the participation of half a million Vietnamese and threatened the colonial order in 25 provinces. At the peak of this movement, the colonial administrative apparatus virtually collapsed in many villages in the provinces of Nghệ-An (birthplace of Hồ Chí Minh) and Hà Tĩnh in northern Central Vietnam. The ICP local leadership spontaneously seized village administrations, abolished taxes, reduced rents, and redistributed up to 8500 acres of communal land in these two provinces alone.

In the short run, both the VNP- and ICP-led movements in 1930–31 collapsed under strong repressive measures by the French colonial regime with its superior military forces. Many revolutionary leaders were imprisoned, while Thailand and southern China again became important bases for exiled political activists, including Hồ Chí Minh. It was not until the release of a number of political prisoners and the introduction of a more liberal political atmosphere in colonial Indochina under the Popular Front regime in France (1936–39) that revolutionary activism gained significant momentum. In the context of the Japanese military's advance into Indochina in September 1940 and the French colonial government's subsequent collaboration with Japanese military authorities, the Vietminh (an abbreviation of *Việt Nam Độc Lập Đồng Minh Hội*, or Vietnam Independence Alliance League) under Hồ Chí Minh began building a base of operation in northern mountainous areas, working with many members of ethnic minorities, against both the Japanese and the Axis-aligned French colonial administration (Marr, 1995, Chapter 3).

Wars and Vietnamese Revolutionary Momentum

The major turning point was 1945, when French colonial power was humiliated by the Japanese-supported coup of 9 March 1945, and when in the context of a power vacuum with the Japanese surrender to Allied forces in August 1945, the Communist-led Vietminh organized the seizure of power throughout the country. The last Vietnamese monarch abdicated under revolutionary pressure. Hồ Chí Minh's Declaration of Independence from France in September 1945 evoked a powerful nationalist sentiment from all social strata, and garnered widespread support for the first independent government of Vietnam (Marr, 1995, 537–9). In the Vietnamese Marxist orthodox discourse, this seizure of power has been referred to as the August Revolution (*cách mạng tháng tám*). It was a revolution only in the narrowest sense of the term: a seizure of administrative power and an official marking of the end of the monarchical era. However, even the maintenance of this administrative power in a territorially unified Vietnam came under serious challenges by France's desire to regain its influence and control of Vietnam, the complications of post-World War II international politics, and the postwar rise of U.S. power and the Cold War. France's attempt to regain control of Vietnam and its underestimation of the support for Vietnamese independence led to the Franco-Vietnamese war from 1946 to 1954. The United States' greater concern with Europe and Soviet influence after World War II, in combination with Chinese Communists taking control of mainland China in 1949, shifted U.S. policy toward a support for France and away from U.S. President Franklin Roosevelt's support for the independence of colonized peoples in Asia and Africa. The violent Franco-Vietnamese war ended in 1954, after the Vietminh defeated the French at Điện Biên Phủ in the northern highlands. Vietnam was temporarily divided into two parts at the 17th parallel, with the north under Communist control and the south under American influence.

Hồ Chí Minh's vision of a unified Vietnam was not achieved until 1975, after a long and violent war that, in the context of the Cold War, involved a massive U.S. ground involvement in South Vietnam and U.S. bombing of the north and parts of Laos and Cambodia. The Vietnamese Communist Party began lending support to the armed resistance against the U.S.-allied government in Saigon in 1959, helped to establish the National Liberation Front of South Vietnam in 1960, and increased the flow of personnel and arms from the north to the south from 1960 onward (Lượng et al., 1995, pp. 285ff.). The United States responded by increasing military advisers in South Vietnam to 15 000 by 1963, bombing North Vietnam from 1964 onward, and introducing ground combat troops to South Vietnam in 1965. At its peak, the United States had half a million combat troops in South Vietnam. In the context of the Cold War, North Vietnam in turn received significant military and economic aid from the Soviet Union, China, Eastern Europe, and Cuba. The intense warfare led to the migration of millions of peasants to urban areas in South Vietnam, seriously damaged the South Vietnamese economy, and increased the dependence of the U.S.-supported South Vietnamese government on U.S. aid.

In the United States, the draft system and the protracted warfare in Vietnam, exposed daily in the mass media, intensified the antiwar movement, especially among the young. This contributed to the U.S. policy of handing the war to South Vietnamese troops, and to the withdrawal of U.S. ground troops and the end of bombing North Vietnam by 1973 in the aftermath of the Paris Peace Agreement. South Vietnam enjoyed no peace, as the Paris Agreement was violated by both the Saigon government's troops, on the one hand, and the North Vietnamese troops and their southern allies, on the other. The U.S.-supported government in South Vietnam eventually collapsed in April 1975, under the attack of North Vietnamese troops and their southern allies.

Vietnamese Marxists' consolidation of political control (in the north by 1954 and in the entire country by 1975) strengthened the foundation for the enactment of their socioeconomic vision. In North Vietnam, the socioeconomic revolution accelerated throughout the 1950s under a Communist vision of a collectivized economy where the state played the dominant role in industrial and agricultural production, as well as in the distribution of raw materials and finished products. By 1965, the northern economy was based mostly on the state and cooperative sectors, and sustained by aid from the socialist bloc, as North Vietnam became directly involved in the war against the United States and the American-supported government in Saigon. The Communist party's economic vision was imposed on the south after 1975. The economic crisis in the late 1970s, fueled by the resistance of peasants in the southern Mekong Delta to collectivization and the disenchantment of northern cooperative farmers, led to a gradual shift toward a market economy. This process of economic reforms (*đổi mới*) started in agriculture in 1981. The economic reform was reaffirmed at the 1986 Communist Party Congress, and implemented in the rest of the economy by 1989. Cultivators in the north consequently have had long-term leases on agricultural land since 1994, and they have regained control over production and distribution processes. In the southern third of the country, land has been under *de facto* private ownership again since 1988, although the principle of collective ownership is reaffirmed in Vietnam's 1992 constitution (Article 17) and various land laws. Private industrial and commercial enterprises as well as large private plantations have been allowed to re-emerge.

Foreign direct investments since 1988 have helped to fuel generally strong economic growth, which averaged 7.3% annually in the 1990–2010 period. Vietnam has moved from being one of the poorest countries in the world in 1990 to joining the lower level of middle-income countries by 2008. Poverty has been greatly reduced. Although Vietnam is still a one-party state officially committed to socialism, socioeconomic inequality has sharply arisen.

The Vietnamese Revolution: Theoretical Debates

Many an important theoretical model on revolution has been constructed partly on the basis of scholarly knowledge of the Vietnamese revolution, a revolution that combined a nationalism-inspired struggle against Western powers with a collectivism-oriented socioeconomic transformation within Vietnam, and that reached its peak from 1945 to 1986. These models are embedded in the major traditions of contemporary Western social theory (John Stuart Mill, Karl Marx, and Emile

Durkheim). In the literature on revolution and agrarian unrest, the political scientist S. Popkin's rational choice model offers a good example of a theoretical model within the Millean tradition of inquiry (Popkin, 1979). It considers the behavior of both the peasantry and the elite as the goal-directed acts of self-interested individuals. Popkin argues that modern revolutionary movements such as the Vietminh succeed because, as political entrepreneurs, their leaders can offer peasants concrete welfare improvements and effective organization against the marauding lords and notables of the old order (cf. Race, 2010: chapters 4 and 7; see also Montesano, 2011). He emphatically rejects the view that modern agrarian movements involve defensive reactions to the violation of precapitalist normative structures by the colonial and capitalist order. In Popkin's argument, the incorporation of indigenous agriculture into the capitalistic world market, the establishment of the colonial regime, and the expansion of state power are not necessarily deleterious to peasant welfare. According to Popkin, colonialism brings vital stability and an improved communication system that keeps prices from fluctuating widely and thus keeps peasants alive in times of local famine.

Intriguing as it is, Popkin's model encounters its share of empirical anomalies in the Vietnamese case. For example, on which matrix of material costs and benefits did a large number of Vietnamese from different social strata base their decision to volunteer for the Vietminh armed forces in 1946–50? Recall that this was when Vietminh troops faced an enemy of overwhelmingly technological superiority, and it was before poor recruits could foresee any substantial socioeconomic reforms in their favor.

In contrast to the Millean tradition of inquiry, the Marxist tradition seeks to situate economic transformation and revolutionary movements within the structure of conflict-ridden class relations. Both Western and Vietnamese Marxist analysts consider Western capitalist expansion as the root of the Vietnamese revolution. In this perspective, French colonialism in Vietnam as the representation of Western capitalist interests sought the maximal return on capital at the expense of the native working classes. The sociologist Martin Murray argues that capitalist exploitative measures included the concession of indigenous land to colonial settlers for the development of major cash crops; the introduction of direct and indirect taxes, and the conversion of tax payments from kind to cash to facilitate capitalist growth; and the oppressive appropriation of labor through the corvee system, the maintenance of rural labor reserves in Tonkin and Annam, and repressive labor laws to hold down labor costs for capitalist agro-mineral ventures. Murray suggests that the impact of world capitalism was far from uniform in Vietnam. It was most striking in Cochinchina, which emerged as one of the top three rice-exporting areas in the world, where approximately two-thirds of the rural population were separated from the means of agricultural production, where land and labor became strictly commodified, and where classes became increasingly differentiated. In contrast, in the north and center (Tonkin and Annam), although the rural population was increasingly drawn into market relations, the majority still tenaciously held onto their minuscule land holdings, buffered in many communities by their small shares of communal land. In North and Central Vietnam, the proportions of inalienable communal land averaged, respectively, 21% and 25% of the cultivated acreage. Murray suggests that the persistence of communal land in North and Central Vietnam contributed to the formation of noncapitalist labor reserves for the plantations, mines, and factories within French Indochina and beyond, reserves where workers could rely on kinship and communal ties for survival in times of recession and during youth and old age. It has been suggested that the cost of capitalist production was consequently held down because capitalist firms had to pay only the cost of short-term labor reproduction (the cost to sustain workers physically during the period of employment), and not the cost of long-term labor renewal (the cost of replacing a physically debilitated workforce with a new generation of laborers and of maintaining a temporarily idle workforce during economic downturns). In the long run, Murray suggests, "These 'blockages' to the homogeneous spread and deepening of the extended reproduction of capital coincided with economic stagnation, declining productivity, and the vicious cycle of rural poverty associated with the 'underdeveloped' regions of the globe" (Murray, 1980, p. x).

In the context of the capitalist world system, Marxist analysts consider the proletariat to be either located largely in the periphery of world capitalism, or composed of 'ethnicities' originating in the periphery; the bourgeoisie is heavily concentrated in, although by no means restricted to, the Western core. Within the Marxist world system framework, class conflict takes on an international dimension, and revolution is seen as originating in the periphery, where proletarian class interests emerge most clearly.

Within the same Marxist framework, the sociologist Jeffrey Paige suggests, "The strength of colonial and imperial political controls long prevented the political expression of these conflicts, but with the decline of colonial power in the postwar era, the commercial export sectors of the underdeveloped world have become centers of revolutionary social movements" (Paige, 1975, p. 3). Assuming inherent conflicts between agricultural producers and the dominant class, Paige attempts to construct an overall model of the forms that class conflict takes (Paige, 1975, 1983). Paige deduces these actors' behavior patterns and the form of agrarian unrest from their income sources. More specifically, he proposes that, on the one hand, the economic base of the landed dominant class is narrower, more static, and less efficient than that of its commercial and industrial counterparts. The former's economic weakness gives rise to zero-sum class conflict and its tyrannical political control of the means of production (both land and labor). On the other hand, the broader, more dynamic, and more efficient capital base of the commercial or industrial upper class can expand nonproducers' resources. This factor facilitates the adoption of more compromising solutions by the dominant class. In this context, class conflict involves the economic distribution of goods and products rather than the political control of labor and land.

Adopting Marx's theoretical arguments in his analysis of the French peasantry (Marx, 1964), Paige further hypothesizes that cultivators' dependence on land as their main source of income gives rise to a conservative, competitive, and structurally dependent peasantry. In contrast, their dependence on wages would result in a more radical, solidary, and structurally interdependent cultivator class. In the Vietnamese case, Paige's

model predicts that agrarian unrest will take the form of rebellion when both cultivators and noncultivators derive their incomes from land, as is the case in North and Central Vietnam. Furthermore, in Cochinchina, "A combination of noncultivators dependent on income from land and cultivators dependent on income from wages leads to revolution" (Paige, 1975, p. 71; Paige, 1983, p. 706). The Marxist framework has its share of empirical anomalies. First of all, the active resistance to French colonialism did gain support among numerous members of the native elite whose organizational efforts provided a crucial link among the rural communities. It is reflected in the necessity of the special category 'resistance-supporting landlord' in the land reform campaign in the north after independence. The active support of these members of the indigenous elite for anticolonial movements cannot be mechanically explained in terms of the struggle by the working classes of the periphery against the capitalists of the core. It is their nationalist aspirations that propelled them toward a path of political activism (see also Emerson, 1960; Woodside, 1976, p. 303).

Vietnam's regional variations in revolutionary strength and socioeconomic conditions present the Marxist models with a second major anomaly. In colonial Vietnam, the peasantry was most heavily dispossessed of the means of production in Cochinchina (South Vietnam). However, from the mid-nineteenth century to 1975, the armed resistance to French colonialism and capitalism was generally more limited in the south than in the north and the center. With the sole exception of the 1940 armed insurrection in certain areas of Cochinchina (cf. Paige, 1975, pp. 323–4; Toàn et al. 1985, pp. 318–9), at no point did the anticolonial armed movement in the south exceed or even approach the intensity of the unrest in the center and the north. In the pre–World War II period, it was in Annam (Central Vietnam) that the 1930–31 unrest posed the greatest threat to the colonial order. During the First Indochina War (1946–54), the French encountered considerably greater resistance in Annam and Tonkin. During the period of direct and massive American intervention (1964–75), it was not in the former Cochinchina but in the southern part of Central Vietnam that the National Liberation Front developed its strongest roots (Mitchell, 1968; Paige, 1975, pp. 326–32; Luong 1992, pp. 227–8). However, it should be noted that among Vietnam specialists writing on the Vietnamese revolution, David Hunt has a different perspective on this issue, and focuses primarily on the transformation of class relations within Vietnam. He seems to agree with Paige in arguing, "[t]he Mekong Delta [in Cochinchina] was the most revolutionary section of the country…[during the 1946–54 war] Viet Minh cadres disregarded [VCP] united front guidelines and chased away rural elites. [In the 1960s] activists in the South … showed little interest in cross-class alliances and resumed their campaign to topple 'the entire socioeconomic elite which constituted the foundation of the Saigon regime.'" (Hunt, 2008: p. 7). In other words, Hunt considers Mekong Delta cultivators revolutionary because they were class conscious, were engaged in class struggle, and did not necessarily lose their solidarity after getting land and becoming middle-class peasants (Hunt, 2003: pp. 603, 615). He considers the northern part of the country less revolutionary than the Mekong Delta because many members of the village elites in the north were allowed to join the Vietnamese Communist Party (Hunt, 2003: p. 603, 615). However, without the success of the struggle (with the support of many elite members) against Western powers, it would not have been easy for the collectivist socioeconomic revolutionary vision to be implemented. It was also in the Mekong Delta that the collectivist socioeconomic vision of the Vietnamese Communist Party encountered the greatest resistance from cultivators (see also Long, 1988; Luong, 1994).

It is a historical fact that North and Central Vietnam, despite their lesser degree of class polarization, constituted the areas with both stronger resistance to Western intervention and lesser resistance to a collectivist revolutionary vision.

While influenced to varying degrees by the Marxist approach, anthropological analyses of agrarian revolutions, and of historical events in general, have paid closer attention to the dynamic interplay between capitalism and indigenous noncapitalist social formations as sociocultural systems (Wolf, 1969).

In James Scott's moral economy model, agrarian unrest is examined in terms of the reactive response of the peasantry in a precapitalist social formation to capitalism and colonialism (Scott, 1976). In the context of the world market and the colonial state, the precapitalist order's subsistence ethic, by which lords and the state are supposed to respect peasants' right to subsistence in exchange for legitimacy, is more often violated. The expansion of Western capitalism and the establishment of European colonies led to the violation of the subsistence ethic through the erosion of patron–client ties and the existing welfare mechanisms in the relation between cultivators and noncultivators. Scott considers peasant movements quintessentially defensive reactions of a moral nature: defensive reactions against the violation of the subsistence ethic and the increasingly serious threat to peasants' survival during subsistence crises. Under this threat and violation, any significant natural or human-induced disaster potentially sparks off agrarian unrest (cf. Luong, 1985). The prominent role of Third World peasants in twentieth-century political movements, Scott argues, involves this dynamic interplay between a precapitalist sociocultural order, on the one hand, and colonialism and capitalism, on the other. Unrest is thus defined, within the Durkheimian tradition of social thought, in terms of a breakdown of the normative framework during a period of transition.

On the basis of his study of the Vietnamese revolution in a Mekong Delta province, David Elliott suggests that the dynamics of this revolution in the Mekong Delta in its earlier period were underlain by a moral economy framework, despite the fact that "the collective sense of moral community… [was] less powerful [in the Mekong Delta] than in other areas of Vietnam" (Elliott, 2003, p. 15). Over time, "the vast social dislocation of the war which atomized the rural communities [in the Mekong Delta] also strengthened the 'rational peasant' individualism of many people struggling to survive in an environment in which familiar community social mechanisms had broken down. ….[I]ndividualism cut into the base of voluntary support for the revolution, in part because the revolution had little to offer a 'rational peasant' by 1971" (Elliott, 2003, 1235; cf. Hunt, 2003, pp. 611ff.)

In the context of the strong socialist revolutionary potential in North and Central Vietnam, Luong hypothesizes

that the precapitalist tradition of the north and the center provided stronger ideological and organizational resources for active resistance to French colonialism in particular, and to capitalist imperialism in general. Despite considerable internal tension, this tradition nurtures values conducive to the growth of nationalism, to collectivism in a relatively mild form, and to a hierarchical organizational framework (Luong, 2010).

On one level, the indigenous precapitalist framework was characterized by both class cleavages, and a male-oriented and kinship-centered hierarchy. Luong suggests that the hierarchical structure of the local tradition, maintained partly through the socialization process, accounts for numerous features in Vietnamese anticolonialism and the Vietnamese revolution: the male-centered elitism in the VNP recruitment and organizational methods, the Vietminh's reliance on male members of the intelligentsia for local leadership in the early period of anti-French resistance, and the male dominance in both the domestic and public arenas in postcolonial Marxist Vietnam.

On another level, the growth of nationalism and a collectivistic vision of political economy are facilitated by important communal institutions and corresponding conceptual categories in the local traditions of the north and the center. Northern and central villages maintained sharp boundaries through villagers' participation in the collective worship of tutelary deities, an extraordinary degree of village endogamy, the institution of communal land, and a nucleated settlement pattern. In contrast, group boundaries were considerably less rigid in southern Vietnamese communities, where settlements were nonnucleated and where there existed significant geographical mobility due to the frontier environment. Luong suggests that through a sharp distinction between insiders and outsiders in the native conceptualization of kinship, communality, ethnicity, and nationality, the northern and central tradition intensified the negative reaction of many Vietnamese to any foreign intrusion on the indigenous landscape and to French domination in the colonial racial diarchy. Luong also suggests that in northern and central Vietnam, the relatively strong institution of communal land in particular, and the subsistence ethic of the precolonial period in general, nurtured a collectivist vision of political economy even among many VNP members in the 1920s, well before their exposure to Marxism. The existence of communal land and the subsistence ethic also facilitated the process of land collectivization in northern and central villages, in comparison to its very slow progress in the south from 1975 to 1985. Finally, within the corporate rural community, the extensive kinship and communal ties facilitated the mobilization of resources by local revolutionary leaders, especially the generally well-regarded members of the native intelligentsia. Luong suggests that the indigenous noncapitalist tradition thus cannot be ignored in any analysis of the Vietnamese encounter with French colonialism in particular and the capitalist world system in general (see Marr, 1971; Woodside, 1976).

See also: Colonization and Colonialism, History of; Eastern European Studies: Society; Southeast Asia: Sociocultural Aspects; Wars among Nation-States: Patterns and Causes.

Bibliography

Emerson, R., 1960. From Empire to Nation. Harvard University Press, Cambridge, MA.

Elliott, D.W.P., 2003. The Vietnamese War: Revolution and Social Change in the Mekong Delta 1930–1975, 2 volumes. M E Sharpe, Armonk, NY.

Emerson, R., 1960. From Empire to Nation. Harvard University Press, Cambridge, MA.

Goscha, C.E., 1999. Thailand and the Southeast Asian Networks of the Vietnamese Revolution, 1885–1954. Curzon Press, London.

Hunt, D., 2003. Review essay: revolution in the delta. Critical Asian Studies 35, 599–620.

Hunt, D., 2008. Vietnam's Southern Revolution: From Peasant Insurrection to Total War. University of Massachusetts Press, Amherst, MA.

Long, N.V., 1988. Some aspects of cooperativization in the Mekong Delta. In: Marr, D., White, C. (Eds.), Postwar Vietnam: Dilemma in Socialist Development. Cornell University Southeast Asia Program, Ithaca, NY, pp. 163–173.

Lượng, C.V., Tạo, V., Cường, T.Đ, Cúc, ĐT., Đính, T.H., Nhật, N.V., 1995. Lịch Sử Việt Nam, 1954–1965 [Vietnamese History, 1954–1965]. Nhà Xuất bản Khoa học xã hội, Hanoi, Vietnam.

Luong, H.V., 1985. Agrarian unrest from an anthropological perspective: the case of Vietnam. Comparative Politics 17, 153–174.

Luong, H.V., 1994. The Marxist state and the dialogic restructuration of culture in rural Vietnam. In: Elliott, D., et al. (Eds.), Indochina: Social and Cultural Change. Claremont-McKenna College, Claremont, CA, pp. 79–117.

Luong, H.V., 2010. Tradition, Revolution, and Market Economy in a North Vietnamese Village, 1925–2006. University of Hawaii Press, Honolulu, HI.

Marr, D.G., 1971. Vietnamese Anticolonialism, 1885–1925. University of California Press, Berkeley, CA.

Marr, D.G., 1995. Vietnam 1945. University of California Press, Berkeley, CA.

Marx, K., 1964. The Eighteenth Brumaire of Louis Bonaparte. International Publishers, New York.

Mitchell, E.J., 1968. Inequality and insurgency: a statistical study of South Vietnam. World Politics 20, 421–438.

Montesano, M.J., 2011. War comes to Long An: the classic we hardly know? Journal of Vietnamese Studies 6, 87–122.

Murray, M.J., 1980. The Development of Capitalism in Colonial Indochina. University of California Press, Berkeley.

Paige, J.M., 1975. Agrarian Revolution: Social Movements and Export Agriculture in the Underdeveloped World. Free Press, New York.

Paige, J.M., 1983. Social theory and peasant revolution in Vietnam and Guatemala. Theory and Society 12, 699–737.

Popkin, S., 1979. The Rational Peasant: The Political Economy of Rural Society in Vietnam. University of California Press, Berkeley, CA.

Race, J., 2010. War Comes to Long An: Revolutionary Conflict in a Vietnamese Province (updated and expanded). University of California Press, Berkeley, CA.

Scott, J.C., 1976. The Moral Economy of the Peasant. Yale University Press, New Haven, CT.

Tai, H.T.H., 1983. Millenarianism and Peasant Politics in Vietnam. Harvard University Press, Cambridge, MA.

Tai, H.T.H., 1992. Radicalism and the Origin of the Vietnamese Revolution. Harvard University Press, Cambridge, MA.

Toàn, N.K., Bình, N.C., Tạo, V., Nam, P.X., Thanh, B.Đ, 1985. Lịch Sử Việt Nam (History of Vietnam), vol. 2. Nhà xuất bản Khoa Học xã hội, Hanoi, Viet Nam.

Wolf, E.R., 1969. Peasant Wars of the Twentieth Century. Harper and Row, New York.

Woodside, A.B., 1976. Community and Revolution in Modern Vietnam. Houghton Mifflin, Boston.

Violence and Media

D Charles Whitney and Ellen Wartella, University of Texas at Austin, TX, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16187–16192, © 2001, Elsevier Ltd.

Abstract

Summarized is research, primarily from the United States, arguing that violent content in mass media has small but significant impact on real-world violence. Three forms of effects, learning of aggressive attitudes and behaviors, desensitization, and cultivation of fear, are discussed, as are three theoretical models – Bandura's social learning theory, Berkowitz's priming effects theory, and Heusmann's social developmental model – for such effects. Contextual factors in effects, and the special nature of child audiences are discussed. Recent research on television content is presented, and public policy alternatives are suggested.

Public and academic concern about media's contribution to real -world violence are about as old as the mass media and the social sciences themselves (Wartella and Reeves, 1985). Despite frequent framing of the matter as 'controversial,' extensive research – an estimated 3000 (Donnerstein et al., 1994) to 3500 (Wartella et al., 1998) studies in the United States alone – have examined the impact of media violence, and a number of recent major reviews (Huston et al., 1992; Murray, 1994; see also Potter, 1999; Paik and Comstock, 1994; Comstock and Paik, 1991) have concluded that media violence plays a measurable role in real-world violence. A variety of US agencies, including the Centers for Disease Control of the US Public Health Service (1991), and medical and public interest organizations, including the American Medical Association, the American Academy of Pediatrics, and the National Conference of Parent-Teacher Associations, have identified media violence as a public health problem. The review below focuses most heavily on US research and US media, most notably American television, primarily because a large majority of the published social science research on media and violence is US research on American audiovisual media. Potter (1999, pp. 44–45), for example, reports 42 published content analyses of US television since 1954, and just 19 from the rest of the world. Moreover, American media are among the world's most violent – and most exported – and real-world violence is a recurring public policy issue. Further, much of this literature concerns impacts of media violence on children and adolescents, for the interrelated reasons that young audiences are considered the most impressionable and most vulnerable. Adults are generally viewed to be more resistant to the deleterious influences of violence, and, as some would argue (cf. Huesmann, 1997), violent behaviors in adulthood may be traced to media use during childhood.

Theories of Effect

Three models have been proposed to describe the process by which such learning and imitation of media violence occurs: social learning theory, priming effects theory, and a social developmental model of learning (Wartella et al., 1998).

First proposed by Albert Bandura in the 1960s, social learning theory is the best known theoretical account of violence effects. Bandura asserts that through observing television models, viewers come to learn behaviors which are appropriate, that is, which behaviors will be rewarded and which punished. In this way, viewers seek to attain rewards and therefore imitate these media models. When both children and adults are shown an aggressive model who is either rewarded or punished for their aggressive behavior, models who are positively reinforced influence imitation among the viewers. Even research in the field has demonstrated that aggression is learned at a young age and becomes more impervious to change as the child grows older. In a longitudinal study to examine the long-term effects of television violence on aggression and criminal behavior, Huesmann et al. (1984) studied a group of youth across 22 years, at ages 8, 18, and 30. For boys (and to a lesser, though still significant extent for girls), early television violence viewing correlated with self-reported aggression at age 30 and added significantly to the prediction of serious criminal arrests accumulated by age 30. These researchers find a longitudinal relationship between habitual childhood exposure to television violence and adult crime and suggest that approximately 10% of the variability in later criminal behavior can be attributed to television violence.

Priming effects theory serves to augment the more traditional social learning theory account of television violence effects. In the work of Leonard Berkowitz and his colleagues, this theoretical account asserts that many media effects are immediate, transitory, and short-term (Berkowitz, 1984). Berkowitz suggests that when people watch television violence, it activates or 'primes' other semantically related thoughts which may influence how the person responds to the violence on television. Viewers who identify with the actors on television may imagine themselves like that character carrying out the aggressive actions of the character on television, and research evidence suggests that exposure to media aggression does indeed 'prime' other aggressive thoughts, evaluations, and even behaviors such that violence viewers report a greater willingness to use violence in interpersonal situations.

Only Rowell Huesmann's (1986; see also Huesmann, 1997) theoretical formulation of the social developmental model of violence effects offers a true reciprocal theoretical account of how viewers' interest in media violence, attention to such violence, and individual viewer characteristics may interact in a theory of media violence effects. Using ideas from social cognition theory he develops an elaborate cognitive mapping or script model. He argues that social behavior is controlled by 'programs' for behavior which are established during childhood. These 'programs' or 'scripts' are stored in memory and

are used as guides to social behavior and problem solving. Huesmann and Miller (1994, p. 161) submit that "a script suggests what events are to happen in the environment, how the person should behave in response to these events, and what the likely outcome to those behaviors would be." Violence from television is 'encoded' in the cognitive map of viewers, and subsequent viewing of television violence helps to maintain these aggressive thoughts, ideas, and behaviors. Over time such continuing attention to television violence can thus influence people's attitudes toward violence and their maintenance and elaboration of aggressive scripts.

This theory suggests that while viewing violence may not cause aggressive behavior, it certainly has an impact on the formation of cognitive scripts for mapping how to behave in response to a violent event and what the outcome is most likely to be. Television portrayals, then, are among the media and personal sources that provide the text for the script which is maintained and expanded upon by continued exposure to scripts of violence.

Huesmann has demonstrated that there are key factors which are particularly important in maintaining the television viewing–aggression relationship for children: the child's intellectual achievement level, social popularity, identification with television characters, belief in the realism of the TV violence, and the amount of fantasizing about aggression. According to Huesmann, a heavy diet of television violence sets into motion a sequence of processes, based on these personal and interpersonal factors, that results in many viewers becoming not only more aggressive but also developing increased interest in seeing more television violence.

It must be emphasized that all serious scholars of the impacts of media on violence are careful to note that media are not the only, nor perhaps among the most important, contributors to real-world violence. Violent behavior is a complex, multivariable problem, formed of many influences. Racism, poverty, drug abuse, child abuse, alcoholism, illiteracy, gangs, guns, mental illness, a decline in family cohesion, a lack of deterrents, the failure of positive role models, among others, all interact to affect antisocial behavior. As Huesmann has argued, aggression is a syndrome, an enduring pattern of behavior that can persist through childhood into adulthood. The impact of media violence appears strongest as a predictor of real-world antisocial behavior as one facet of a 'culture of violence.'

Types of Effects

As Potter (1999, Chap. 9) notes, media-violence effects fall into five categories – physiological, emotional, cognitive, attitudinal, and behavioral – and both immediate or short- and long-term effects have been studied. While some attention has focused on direct, short-term imitative or modeling effects (cf. Phillips, 1980, 1982, but see Hessler and Stipp, 1985), more attention and public policy concern has focused on the long-term impact of repeated exposure to violence. More generally, three overarching categories of effect receive most attention: learning of aggression, desensitization to real-world violence, and the cultivation of fear in repeated exposure to media violence (Wilson et al., 1997).

Clearly, not all violent depictions should be treated equally, nor all viewers. The (US) National Television Violence Study (Wilson et al., 1997) identified several contextual factors within a representation that may influence audience reactions to media violence which include the following.

The Nature of the Perpetrator

Where individuals perceive perpetrators of violence as attractive, as heroes, and/or as similar to themselves, the likelihood of stimulating attention (Bandura, 1986) and aggression (Paik and Comstock, 1994) increases.

The Nature of the Victim

While the commission of violence on an attractive character with which an audience member identifies might serve to inhibit aggressive behavior, its principal impact would seem to be in arousing fear among the audience members.

The Reason for the Violence

Wilson et al. (1997, p. 24) note that violence viewed as justified likely heightens aggression, while violence viewed as unjustified arouses fear. The impact of justification has been documented with fictional as well as realistic programming (Meyer, 1972), and with adult as well as child viewers (Liss et al., 1983). In fact, a recent meta-analysis of 217 media studies documents that a justified portrayal of violence can enhance aggressive behavior among viewers (Paik and Comstock, 1994).

The Presence of Weapons

A number of studies, including a meta-analysis of 56 published experiments (Carlson et al., 1990) have demonstrated that the presence of weapons, either pictorally or in the natural environment, can enhance aggression among subjects. While, for ethical reasons, the large majority of such research involves adult subjects, in at least one study (Frodi, 1975) the presence of weapons enhanced aggression among adolescents. 'Conventional' weapons such as guns and knives are more likely than unconventional means for priming the effect, social learning theory would suggest, because their use as a means of aggression are stored in memory (Berkowitz, 1990; Leyens and Parke, 1975).

The Extent and Graphicness of the Violence

A review (Wilson et al., 1997) for the National Television Violence Study suggested that more research is needed, but several tentative conclusions about extent and graphicness could be reached: (1) extensiveness of violence *within* media presentations should be associated with increased desensitization to violence, at least in the short- to medium-term; (2) graphicness of violence should be associated with increased cultivation of fear; (3) longitudinal studies clearly suggest that extensiveness of *viewing*violent media presentations heightens the likelihood of engaging in aggressive behavior.

The Degree of Realism of the Violence

In brief, realistic violence has been found to induce aggressive behavior, and to induce fear, more than violence believed to be

less realistic or more fantastic. An extremely important qualification deals with younger children, who may be unable to distinguish realistic from fantastic characters, behaviors, and situations. In one study, however, where *perceived realism* was manipulated for older children (9 – 11 in Feshbach, 1972; 10 – 13 in Atkin, 1983), those subjects who were led to believe that footage was realistic news were more likely subsequently to behave aggressively than those led to believe it was taken from an entertainment program.

Whether Violence is Rewarded or Punished

Rewarded violence is more likely to be imitated than violence which is punished. Significantly, and particularly for children (since, as we will show below, television programming most frequently presents violent actions that are neither rewarded nor punished), the absence of punishment may enhance imitation, even in the absence of explicit reward (Bandura, 1965; Walters and Parke, 1964). Paik and Comstock's meta-analysis (1994) suggests that rewarded violence stimulates aggression among both child and adult audiences. One study suggests that punishment of criminal violence decreases fear (Bryant et al., 1981).

Consequences of Violence

In general (and exceptions are noted in Wilson et al., 1997, p. 30), mediated depictions of violence which show either pain cues or other short- or longer-term negative effects or consequences of violence are likely to depress the learning of aggression. There is little research on the effects of pain cues or violence consequences on desensitization and the cultivation of fear.

Presence of Humor

As the National TV Violence Study review also noted, further research is needed here as well, but the present state of knowledge suggests, other things being equal, that violence coupled with humor is more likely to heighten aggression, and to increase desensitization, than violence without the presence of humor:

> Several mechanisms can be used to explain such a facilitative effect of humor on aggression. Humor might elevate a viewer's arousal level over that attained by violence alone, and increased arousal has been shown to facilitate aggression. … Humor could serve as a reinforcement or reward for violence, especially if the perpetrator is funny or admired or his or her wit. And humor may diminish the seriousness of the violence and therefore undermine the inhibiting effects of harm and pain cues in a scene. … However, we should underscore that our conclusion about the facilitative effect of humor on aggression is tentative until more systematic research … is undertaken (Wilson et al. 1997, p. 32).

Young Viewers

As noted, research indicates that certain factors may be processed differently by young viewers. First, children below about age 8 have more difficulty distinguishing reality from fantasy and often imitate superheros with magical powers such as the *Power Rangers* (Boyatzis et al., 1995). Second, young children may have difficulty connecting scenes and drawing inferences from the plot. Timing of punishments and rewards becomes important in this instance. In many programs, the crime or violent behavior may go unpunished until the end of the program. Young children may have difficulty connecting the ending punishment with the initial violent act and may, therefore, believe that the violence went unpunished (Wilson et al., 1997). Thus, learning of aggressive attitudes and behaviors from television varies by both the nature of the portrayals and the nature of the viewers. The presence of contextual factors in the portrayals which may inhibit young children's social learning of aggression decreases the negative consequences of such portrayals and should be encouraged. Not all violent portrayals are the same and the context of violence is clearly quite important. Similarly, young children, those under the age of 7 or 8, may be particularly susceptible to learning from exposure to television violence because of differences in how they make sense of television compared to adults.

The Media Environment

Television's role as the central mass medium in much of the world for the past half-century, and its ubiquity and ability to enter almost every home, often without parental supervision, has meant that more public and scholarly concern has focused on its contents than on any other medium's, and this concern has accompanied its diffusion into every corner of the earth.

Unfortunately, cumulative and comparative research on television's violent content is hampered by a lack of consistency in defining violence and especially defining the population and sampling frame in studies of television.

The most extensive single content study of US television was the 1994–97 National Television Violence Study (National Television Violence Study, 1997; Center for Communication and Social Policy, 1997, 1998). Examined were the 6 a.m.–11 p.m. contents of a multistage probability sample constructed sample week of programming on 23 network-station, independent-station, and basic-cable and premium-cable channels; thus about 8000 programs were analyzed over the 1994–95, 1995–96, and 1996–97 television 'seasons.' Certain programs, including 'hard news,' religious shows, sporting events, quiz shows, and educational shows, were sampled but not analyzed. About three-fifths of the remaining programs contained some visual violence, a figure that like most summary statistics remained stable over the 3 years of the study. In descending order, premium cable, basic cable, independent-station, broadcast network station, and public broadcast stations' programming were likely to contain violence. By content genre, in decreasing order, movies, dramas, children's shows, music videos, and reality-based and comedy programs were likely to contain violence. Violence was far more prevalent during prime-time than during daytime hours.

Of signal concern to the NTVS researchers was the *context* of televised violence; it was often glamorized (more than a quarter of all violence was perpetrated by 'good' or attractive characters,

and some 40% by characters with at least some good qualities); sanitized (about 7/8 of violent scenes show no blood and gore; almost half show no harm to victims of violence, although more than half of violent interactions show infliction of harm that would be lethal in 'real life,' and about half depict no pain cues in victims of violence); and unsanctioned (in a majority of scenes, violence perpetrators were neither rewarded nor punished ['punishment' was considered any noticeable sanction, including a perpetrator's oral expression of remorse], and among other scenes, rewards and punishments were about equally likely; in three-quarters of cases, characters perpetrating violence were either never punished, or were punished only at the program's conclusion). Moreover, only 3% of programs with violence had any antiviolence theme (Center for Communication and Social Policy, 1998, Chaps. 3–4). Potter (1999) has an extensive discussion of definitions of media violence and the results of content analyses from a variety of studies.

Media Violence and Public Policy

As noted, the consensus in the social scientific community regarding media violence is that it serves as a contributor to aggression in the real world, and virtually all public opinion surveys confirm that wider publics believe this as well. Nonetheless, such findings are 'controversial' in the media industries, and among a minority of academics (see, e.g. Fowles, 1999 and sources cited therein). In the United States in particular, since the 1950s the media industries – television in particular, but also recorded music, motion pictures, pictorial comic books, and video games – have responded to public and governmental outcries against violent content by promises to reform under self-regulation. Motion pictures, video-games, and recorded music all list 'ratings' for their products which suggest age-groups for which the industry self-regulatory groups think the content for these products is appropriate. The US Telecommunications Act of 1996 mandated that the television networks create a 'voluntary' ratings system or face creation of one by the federal government. The system the broadcast industry has created suggests appropriate age categories and levels of sex, language, and violence to allow consumers and parents to make program choices. The same act mandated a 'v-chip' in newly manufactured television sets to allow parents to screen or filter out violent programs. There is to date insufficient research to indicate whether and what sorts of parents and other viewers are using either the ratings or the v-chip technology to screen violence. However, experimental research by Cantor and her colleagues (Cantor and Nathanson, 1998, Cantor et al., 1997) suggests that young children may use aged-based ratings systems to shield themselves from violent content while for older children and adolescents, there may be a boomerang or 'forbidden fruit' effect whereby ratings *attract* them to more 'adult' violent or sexually explicit material. In the United States Congress, a moratorium on discussion of television violence is in effect, pending further information on the effects of ratings and the v-chip, but the legislature is focusing its attention on violent video-games, the target of significant public criticism in the wake of a number of firearms murders in public schools.

Conclusion

It is clear that where children and television violence are concerned, the question that remains is not whether media violence has an effect, but rather how important that effect is in comparison with other factors in bringing about the current level of crime in the United States and other industrialized nations. Future research should also aim to establish who precisely is most susceptible to media violence, and, most importantly, what sorts of intervention might help diminish its influence. At the same time, any interventions that help establish policies and practices to reduce the socially inappropriate ways of portraying violence and increase the socially responsible ways (such as using violence to assert antiviolence messages) should be encouraged as well. Long-term solutions to problems caused by violence in the real world, however, will require attention to a much wider variety of causal agents.

See also: Media Effects on Children; Media Effects.

Bibliography

American Psychological Association, 1993. Violence and Youth: Psychology's Response. Volume 1: Summary Report of the American Psychological Association Commission on Violence and Youth. American Psychological Association, Washington, DC.

Atkin, C., 1983. Effects of realistic TV violence vs. fictional violence on aggression. Journalism Quarterly 60, 615–621.

Bandura, A., 1965. Influences of models' reinforcement contingencies on acquisition of imitative responses. Journal of Personality and Social Psychology 1, 589–595.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall, Englewood Cliffs, NJ.

Berkowitz, L., 1984. Some effects of thoughts on anti-social and pro-social influences of media events: a cognitive-neoassociationistic analysis. Psychological Bulletin 95 (3), 410–427.

Berkowitz, L., 1990. On the formation and regulation of anger and aggression: a cognitive-neoassociationistic analysis. American Psychologist 45, 494–503.

Berkowitz, L., LePage, A., 1967. Weapons as aggression-eliciting stimuli. Journal of Personality and Social Psychology 7, 202–207.

Boyatzis, C., Matillo, G., Nesbitt, K., 1995. Effects of 'The Mighty Morphin Power Rangers' on children's aggression with peers. Child Study Journal 25 (1), 45–55.

Bryant, J., Carveth, R.A., Brown, D., 1981. Television viewing and anxiety: an experimental examination. Journal of Communication 31 (1), 106–119.

Cantor, J., Harrison, K., Nathanson, A., 1997. Ratings and advisories for television programming, Part III. In: Center for Communication and Social Policy, University of California. In: Barbara, Santa (Ed.), National Television Violence Study, vol. 2. Sage, Thousand Oaks, CA, pp. 267–322.

Cantor, J., Nathanson, A., 1998. Ratings and advisories for television programming, Part III. In: Center for Communication and Social Policy, University of California. In: Barbara, Santa (Ed.), National Television Violence Study, vol. 3. Sage, Thousand Oaks, CA, pp. 285–321.

Carlson, M., Marcus-Newhall, A., Miller, N., 1990. Effects of situational aggression cues: a quantitative review. Journal of Personality and Social Psychology 58, 622–633.

Center for Communication and Social Policy, University of California, Santa Barbara, 1997b. National Television Violence Study, vol. 2. Sage, Thousand Oaks, CA.

Center for Communication and Social Policy, University of California, Santa Barbara, 1998. National Television Violence Study, vol. 3. Sage, Thousand Oaks, CA.

Centers for Disease Control, 1991. Position Papers from the Third National Injury Conference. US Department of Health and Human Services, Atlanta, GA.

Comstock, G., Paik, H., 1991. Television and the American Child. Academic Press, San Diego, CA.

Donnerstein, E., Slaby, R., Eron, L., 1994. The mass media and youth violence. In: Murray, J., Rubinstein, E., Comstock, G. (Eds.), Violence and Youth: Psychology's Response, vol. 2. American Psychological Association, Washington, DC, pp. 219–250.

Feshbach, S., 1972. Reality and fantasy in filmed violence. In: Murray, J.P., Rubinstein, E.A., Comstock, G. (Eds.), Television and Social Behavior: Television and Social Learning, vol. 1. US Government Printing Office, Washington, DC, pp. 318–345.

Fowles, J., 1999. The Case for Television Violence. Sage, Thousand Oaks, CA.

Frodi, A., 1975. The effects of exposure to weapons on behavior from a cross-cultural perspective. International Journal of Psychology 10 (4), 283–292.

Gerbner, G., Gross, L., Morgan, M., Signorielli, N., 1980. Living with television: the dynamics of the cultivation process. In: Bryant, J., Zillmann, D. (Eds.), Media Effects. Lawrence Erlbaum, Hillsdale, NJ, pp. 17–41.

Hessler, R.C., Stipp, H., 1985. The impact of fictional television suicide stories on United States fatalities: a replication. American Journal of Sociology 90, 151–167.

Huesmann, L.R., 1986. Psychological processes promoting the relation between exposure to media violence and aggressive behavior by the viewer. Journal of Social Issues 42 (3), 125–139.

Huesmann, L.R., 1997. Screen Violence and Real Violence: Understanding the Link. Huesmann, Ann Arbor, MI.

Huesmann, L.R., Lefkowitz, M.M., Walder, L.O., 1984. The stability of aggression over time and generations. Developmental Psychology 20 (6), 1120–1134.

Huesmann, L.R., Miller, L., 1994. Long-term effects of repeated exposure to media violence in childhood. In: Huesmann, L.R. (Ed.), Aggressive Behavior: Current Perspectives. Plenum, New York, pp. 153–183.

Huston, A.C., Donnerstein, E., Fairchild, H., Feshbach, N., Katz, P., Murray, J., Rubinstein, E., Wilcox, B., Zuckerman, D., 1992. Big World, Small Screen: The Role of Television in American Society. University of Nebraska Press, Lincoln, NE.

Leyens, J.P., Parke, R.D., 1975. Aggressive slides can induce a weapons effect. European Journal of Social Psychology 5, 229–236.

Liss, M., Reichardt, L., Fredriksen, S., 1983. TV heroes: the impact of rhretoric and deeds. Journal of Applied Developmental Psychology 4, 175–187.

Meyer, T.P., 1972. Effects of viewing jusitified and unjustified real film violence on aggressive behavior. Journal of Personality and Social Psychology 23, 21–29.

Murray, J.P., 1994. Impact of televised violence. Hofstra LawReview 22, 809–825.

National Television Violence Study, 1997. National Television Violence Study, vol. 1. Sage, Thousand Oaks, CA.

Page, D., O'Neal, E., 1977. 'Weapons effect' without Demand characteristics. Psychological Reports 41, 29–30.

Paik, H., Comstock, G., 1994. The effects of television violence on antisocial behavior: a meta-analysis. Communication Research 21, 516–546.

Phillips, D.P., 1980. Airplane accidents, murder and the massmedia. Social Forces 58, 1001–1024.

Phillips, D.P., 1982. The impact of fictional television stories on United States adult fatalities. American Journal of Sociology 87, 1340–1359.

Potter, W.J., 1999. On Media Violence. Sage, Thousand Oaks, CA.

Turner, C.W., Layton, J.F., Simons, L.S., 1975. Naturalistic studies of aggressive behavior: aggressive stimuli, victim visibility, and horn honking. Journal of Personality and Social Psychology 31, 1098–1107.

Walters, R.H., Parke, R., 1964. Influence of response consequences to a social model on resistance to deviation. Journal of Experimental Child Psychology 1, 269–280.

Wartella, E., 1995. Media and problem behaviours in young people. In: Rutter, M., Smith, D.J. (Eds.), Psychosocial Disorders in Young People: Time Trends and Their Causes. J. Wiley, Chichester, UK, pp. 296–323.

Wartella, E., Olivarez, A., Jennings, N., 1998. Children and television violence in the United States. In: Carlson, U., von Feilitzen, C. (Eds.), Children and Media Violence: Yearbook from the UNESCO Clearinghouse on Children and Violence on the Screen. UNESCO International Clearinghouse on Children and Violence on the Screen. Göteborg, Sweden, pp. 55–62.

Wartella, E., Reeves, B., 1985. Historical trends in research on children and the media: 1900–1960. Journal of Communication 35 (2), 118–133.

Whitney, C., Wartella, E., Lasorsa, D., Danielson, W., Olivarez, A., Lopez, R., Klijn, M., 1996. Part II: Television violence in 'reality' programming. In: University of Texas. In: Austin (Ed.), National Television Violence Study, vol. 1. Sage, Thousand Oaks, CA, pp. 269–360.

Wilson, B., Kunkel, D., Linz, D., Potter, W.J., Donnerstein, E., Smith, S., Blumenthal, E., Gray, T., 1997. Part I: Violence in television programming overall. In: Center for Communication and Social Policy, University of California. In: Barbara, Santa (Ed.), National Television Violence Study, vol. 1. Sage, Thousand Oaks, CA, pp. 1–268.

Violence as a Public Health Problem

Raymond P Lorion, College of Education, Towson University, Towson, MD, USA
Michael B Blank, Center for Mental Health Policy and Services Research, University of Pennsylvania, Philadelphia, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Rather than approaching pervasive community violence (PCV) as a criminal justice problem requiring responses from police, judiciaries, and corrections, this article describes violence as a public health problem requiring responses from a variety of health and public policy perspectives. Here, we posit that community violence represents a toxin to the social ecology of communities that has radiating negative effects analogous to other environmental toxins. Drawing from local and global examples, we argue that assessing exposure to the toxic effects of PCV is a necessary first step in planning for interventions that mitigate negative impacts on individuals and communities. Notably, we argue that the burden of disease is highly varied and unpredictable, ranging from the most extreme internalizing outcomes (e.g., suicide) to extreme externalizing disorders (e.g., homicide). We propose a few approaches for measurement of the various fear circuitry disorders that may result as individual responses to PCV, as well as the use of community-based participatory research methods to develop evidence-based interventions to address PCV from a social justice and public policy perspective. The article concludes with a brief discussion of vulnerability and adaptation in communities exposed to pervasive violence, as well as factors related to resiliency and health-related quality of life.

Introduction

Just as the writing of this article neared completion, the following headline appeared in a local newspaper: *Short life marred by a milieu of violence* above an article reporting the fatal shooting of a 16-month-old boy by a gunman targeting the infant's father. The article described the infant's brief life and the *milieu* in which he was raised. This milieu was marked by nearly continuous and inescapable exposure to violence. His mother sought to escape from elements of this milieu by distancing herself from his abusive father; by reluctantly and infrequently taking her son outside; and by shielding him from angry and threatening neighbors. Despite his young age, she wondered if he would reach adulthood safe from harm.

As described below, such stories could be written any day in any week in many of the largest cities, both in the United States and elsewhere around the globe. Living under such conditions we describe as exposure to pervasive community violence (PCV) and liken its effects to other environmental toxins that can be fatal with direct contact and indirectly corrode physical, psychosocial, and emotional well-being of such ecologies. Not unlike the conditions within insurgencies and civil wars, PCV can have malignant effects on interpersonal trust within families and across neighborhoods; can isolate young and old alike in what they experience as the uncertain safety of their homes; and can undermine assumption about the future for oneself and especially, members of one's family that are male and minority.

Because of PCV, violence and the threat of violence are increasingly recognized as serious public health problems in the United States and globally. Violence and its aftermath leave no demographic entity unaffected. In the United States in 2009, 16 800 people were victims of homicide; an additional 37 000 took their life (http://www.cdc.gov/violenceprevention/). The incidence of violent deaths in the United States is alarming; its incidence across the globe, especially in war-torn areas is staggering!

These lost lives, however, represent merely a fraction of the impact of PCV exposure! We argue herein that PCV exposure cumulative impact is the creation of a sustained psychosocial environment as toxic to human health as many other forms of ecological toxins. These effects include not only physical injury and death following direct contact with violence but also deep-seated psychological and psychosocial 'bruises' which may interfere with health-promoting interpersonal connections; limit access and openness to intimacy; and through an increase in anticipatory and reactive violence, actually increase the ambient PCV level. PCV exposure may result in permanent physical and emotional scars. PCV settings also require the diversion of substantial public resources to provide health care, public safety (i.e., through police and other forms of security), and lost productivity. Violence also has the potential to erode communities by decreasing property values; disrupting social services; and infecting relationships among neighbors with suspicion, caution, and trepidation.

In 2002, the Centers for Disease Control and Prevention and the World Health Organization produced a startling joint report of the impact of violence worldwide (Krug et al., 2002). Globally, interpersonal violence accounts for more than 1.6 million deaths annually and is one of the leading causes of death for people between the ages of 15 and 44. Although more than 90% of violent deaths occur in low- and middle-income countries, affluent nations are not insulated from this toxin. In the United States, for example, the December 2012 mass shooting of educators and primary grade children at Sandy Hook Elementary School; of moviegoers in Aurora Colorado; and the bombings during the 2013 Boston marathon occurred in affluent communities. Each of these tragic events resulted in public discussions of gun control, mental illness, and acknowledgment of the US's culture of violence. Each also expanded the number of people who no longer feel secure in their homes, in their schools, and in public venues.

Despite such high-profile cases generating cries for change (but not necessarily actual change!) and increasing awareness

that PCV is present across the United States and the world, its impact on those exposed remains poorly understood. Thus, we argue in this article that PCV exposure has become a characteristic of the environments in which many children, adolescents, and adults (teachers, parents, and citizens) live, work, study, and play across the world. During the 1990s, measures to assess the nature and levels of PCV exposure were developed for youth ranging in age from preschool through adolescence. Findings from the resulting surveys across multiple settings (urban, suburban, and rural) confirmed that PCV exposure is a widespread and damaging phenomenon. As noted above, documented physical, emotional, and behavioral consequences support consideration of PCV as an environmental toxin with measurable negative effects, including increase in its ambient level. Thus, PCV has the potential to increase its pervasiveness within the very environments in which it is encountered and thus to gradually expand beyond its original borders to contaminate adjoining areas and beyond.

The need to assess PCV exposure in adults is argued as a necessary antecedent to the design and implementation of interventions to protect against exposure and mitigate its initial effects.

Many of the epidemiological studies reported herein were conducted in the United States; it is imperative that similar epidemiological work be undertaken globally. Far too much of the world's population currently lives in war-ravaged or threatened areas or have escaped from such settings only to find themselves in dangerous refugee settings. Far too many women and children experience the impacts of domestic, family, and relational violence, and far too many others witness both these forms of violence, and their physical and emotional consequences.

As work began to understand and measure PCV during the 1980s, many policy makers and social scientists viewed PCV as a problem limited to low-income neighborhoods trapped in struggles among gangs competing to sell and distribute illegal drugs, especially cocaine and crack cocaine. From this perspective, the challenge for communities was to locate and restrain perpetrators and respond to their victims. Less public were efforts to contain such turbulence within these neighborhoods and thereby avoid or limit contamination of outlying areas. Gradually, there was recognition that community violence may affect all segments of the population. Those fearing victimization feel a need to protect themselves; often with tragic consequences. As this article was being prepared, a middle-class teenager died because he mistakenly entered a home close in location and in appearance to his own. Confused by the combination of alcohol and fatigue, the youth entered the wrong home, encountered an anxious middle-class homeowner who fearing for his family's safety killed the intruder before ascertaining whether the perceived threat was confirmed as valid.

We hypothesize but have yet to confirm empirically that PCV impairs judgment in such instances resulting, as noted, in unnecessary and unintended acts of violence. Increasingly, across the world seemingly clear distinctions between victim and perpetrator seem to fade. Heightened exposure to violence directly or through media reports of its occurrence appears to increase acts of aggression in anticipation of violence from another. Effectively, the timing rather than the nature of an act may differentiate victim from victimizer. Consider the aforementioned accidental death of the teenage victim who entered the wrong home. Had the homeowner paused long enough to be sure of the threat, he and his family may have been the victims and the teenager, the perpetrator. Acting too swiftly may be as dangerous as acting too slowly in a PCV environment!

In such an environment, the sense of unease may seem permanent, the need for vigilance continuous, and suspicion of others a required attribute. That state may, in turn, disrupt sleep, upset vegetative functions, impair interpersonal relationships, and disable effective concentration on work and studies. Together, this constellation of 'fear circuitry disorders' (Andrews et al., 2009) representing a specific group of stress-induced and fear-based disorders are now identified in DSM-5 as a distinct classification as Stress-Induced and Fear Circuitry Disorders. This classification brings together diverse fields as neuroscience and community psychology and may lead to more refined treatments for these conditions based on a better understanding of the biological and environmental factors that contribute to their development. These disorders range from posttraumatic stress disorder, panic disorder and agoraphobia, social phobia, and specific phobias, and recent work suggests that they may be closely related in terms of brain pathophysiology. PCV, therefore, may negatively alter the essence of community life, broadly disrupt its functioning, and presumably thereby impair residents' general physical and emotional health.

The potential for such broad impact must be appreciated and respected *before* one can enlist communities to investigate PCV's nature and consequences. Investigators who do not appreciate that quality of the study depended on this aspect may be surprised by the reluctance of community leaders (e.g., school administrators, county health officer, and representatives of the mayor's office) to participate in its study. The leaders may deny the problem's extent or hesitate to make public the negative quality of their neighborhood or community. For them genuine partnerships require time-tested, abiding historical involvement and future commitment. Historical involvement enables disclosure and trust. Intimate aspects of a community are rarely shared with strangers. Assurance that the investigator will be there in the future (even after external funds are gone) is important, given community leaders' appreciation of the intransigence of social problems and the fickleness of social scientists.

Thus, from the outset it must be appreciated that to study and alter the nature and consequence of PCV requires a genuine and long-term partnership between communities and social scientists that is exemplified by Community-Based Participatory Research (CBPR). Roberts (2013) suggests that CBPR activities flow from the foundational principles of respect for persons, beneficence, and justice that were outlined in the Belmont Report (1979). Roberts (2013) insists that CBPR activities must always demonstrate a fundamental and deep regard for the dignity and autonomy of individuals who constitute a community by virtue of residing in a certain geographic location, having a shared racial, ethnic, or cultural identity: or living with a particular illness. Advances in biomedical ethics have invariably occurred very slowly, even glacially, and historically only after

public outrage at abuses such as were revealed at Nuremberg, Tuskegee, and other infamous projects and experiments.

What differentiates CBPR from traditional approaches to conducting research is the sharing of power and participation with communities from the very beginning of any project. CBPR is not conducted 'on' participants with particular characteristic or illnesses, or 'in' communities where they live, but rather it conducts research 'with' communities and stakeholders. That difference is essential! Shared decision-making and embracing an empowerment process (Rappaport, 1981) are the cornerstones of CBPR. When Rappaport argued for a social policy of empowerment rather than prevention, he described it as paradoxical because he recognized that only by recognizing and valuing the strengths in individuals, groups, and communities could programs be successfully implemented to address deficits, inequities, or illness. There is much to be learned from CBPR efforts that failed, and the behind-the-scenes narratives of what did not work and speculations as to why they failed, since those failure experiences are often as illustrative as are successes, and those types of lessons are often far more plentiful that than are the stories of success (Primavera and Brodsky, 2004).

Definition of Terms

In reviewing the growing PCV literature, the need for defining relevant terms becomes apparent. *Violence* refers to both acts and consequences of the *intentional* use or threat of use of force to cause injury, harm, or death to another. The qualifier 'intentional' excludes unintended or accidental acts and consequences. Determining intent represents an as yet unresolved methodological challenge. For PCV, the *consequences of violent acts* refer to the rippling health and mental health effects experienced by: (1) individuals directly victimized by an act or threat of violence; (2) individuals indirectly victimized as a bystander, witness, acquaintance, or loved one of a direct victim; and (3) individuals indirectly victimized by their awareness of, and anxiety about, the occurrence of acts or threats of violence. *Community violence* locates the acts and their consequences in a geographic setting (e.g., a neighborhood or a school), an identifiable social grouping (e.g., a grade cohort, a gang, a nuclear or extended family), or both. *PCV* refers to the chronic and widespread occurrence of acts or threats of violence within such settings or social groups.

Thus, PCV exists when *substantial* acts and consequences of violence are encountered across *multiple* settings and circumstances over *extended* time. To investigate PCV systematically, the qualifiers 'substantial,' 'multiple,' and 'extended' must be operationally defined and rigorously measured. Information must be gathered about a community's rate of violent acts and about rates of individual encounters with violent acts. For methodological and political reasons, PCV measures vary in their consideration of violence occurring within or outside of the home. The majority of such measures limit their focus to outside encounters. Yet, by definition, PCV implies cross-setting exposure. At its most extreme, PCV represents the absence of any sanctuary or setting in which relief and a sense of safety is possible. Thus, the most serious impacts on physical and psychological health are likely to be observed in those exposed at home, at work or school, and en route between the two.

In a series of papers and articles, Lorion and colleagues proposed consideration of *PCV as an environmental contaminant* with measurable effects on health and mental health (Cooley-Quille and Lorion, 1999; Lorion, 1998, 1999, 2001a,b; Lorion et al., 1998; Lorion and Saltzman, 1993) From that perspective, PCV contaminates settings and thereby the actions and interactions of at least some inhabitants. *Contamination* refers to the corrosive or pathogenic effect of PCV exposure on community climate. Slowly investigators are documenting PCV's negative effects on: (1) individual assumptions about the threatening quality of others, (2) the need to limit one's contact with others, and (3) the urgency of responding aggressively to actual or anticipated attacks. This ecological contamination appears to vary in its pace as well as intensity across individuals, situations, and settings. Understanding the determinants and mechanisms of such patterns and their malleability within and across individuals should inform the design of interventions to limit PCV exposure and reduce its consequences.

Measuring PCV

As reflected in this section of the article, attention has shifted in the past decade from continued examination of PCV as a significant ecological influence on emotional and physical behavior in spite of its continued persistence and, we believe, expanding toxicity. Ideally, reviewing past work will have a heuristic influence on renewing interest in understanding and preventively controlling the PCV's repercussions. Investigation of PCV's environmental influence requires the availability of measures to estimate its presence and to confirm its impacts. Applying a common measurement approach across diverse populations and settings, Richters and associates (Lorion and Saltzman, 1993; Martinez and Richters, 1993; Osofsky et al., 1993; Richters and Martinez, 1993) scientifically documented the extent of children's exposure to violence. They also provided early evidence of the relationship between PCV exposure and children's emotional and behavioral functioning. Subsequent investigations of such relationships (e.g., Hill and Madhere, 1996; Singer et al., 1995) confirmed the seriousness of such links.

Measures of exposure to community violence share many similarities with the *Survey of Children's Exposure to Community Violence* (Richters and Saltzman, 1990). Instructed explicitly *not* to include media exposure (i.e., what was seen or heard about on radio, television, or in movies), respondents indicate along a five-point Likert scale (from 0 = *never* to 4 = *a lot of the time*) how often they have been victimized by, heard about, or witnessed violent events or related activities (e.g., shooting, stabbing, mugging, seeing a dead body). For each instance of reported exposure, respondents provide information about: (1) where the exposure occurred (i.e., in or near school, in or near home, in the neighborhood); (2) who perpetrated the violence (e.g., a stranger, a family member); (3) who, if not the respondent, was victimized (e.g., a stranger, a family member); and (4) when the incident occurred (e.g., 1 week ago, 5 years ago).

Variations of this measure were applied by Martinez and Richters (1993) and Osofsky et al. (1993) in studies of mothers of elementary graders and by Richters and Martinez (1993) and Saltzman (1992) in their studies of fifth and sixth graders. A cartoon-based version of this measure was designed by Richters and Martinez (i.e., Things I Have Seen and Heard, 1990) to assess PCV exposure in primary graders. Test–retest reliability was the most acceptable $r = 0.81$. A second cartoon-based measure, Levonn, assessed distress symptoms associated with exposure (Richters et al., 1990). Subsequently, Fox and Leavitt (1995) developed the Violence Exposure Scale for Children (VEX-R) as a cartoon-based measure of PCV exposure in children from preschool through the elementary grades. Using this measure, Shahinfar et al. (2000) documented PCV's impact on preschoolers. The negative behavioral consequences of such exposure were confirmed in a sample of Israeli elementary school children by Raviv et al. (2001).

A further variation of the Richters and Saltzman (1990) measure was applied by Saltzman (1995) to examine PCV exposure's link to adolescent involvement in antisocial and delinquent behaviors. Building on these variations, Singer et al. (1995) designed the Recent Exposure to Physical Violence and the Past Exposure to Violence surveys to obtain adolescent reports of exposure during the prior 12 months or 'while growing up, not including the past year,' respectively. The acts of violence surveyed by Singer et al. included threats, slapping, hitting, punching, beatings, knife attacks, and shootings. Separate reports for each (other than knife attacks and shootings) were obtained for home, school, and neighborhood. Respondents were also asked how often they had witnessed someone else victimized at each of the aforementioned sites. Two additional items inquired about respondents' having been 'made to do a sexual act' or having witnessed someone else 'being made to do a sexual act.' Overall, measurement efforts to date support the conclusion that psychometrically acceptable reports of PCV exposure can be obtained from youth ranging in age from preschoolers to adolescents.

In addition to the above work, Jenkins and associates (Bell et al., 1988; Jenkins and Thompson, 1986) designed and conducted qualitative studies of African-American children's encounters with violence. Results from four studies are summarized in Bell and Jenkins (1993). Using related measures, Jenkins and colleagues obtained information about youth's exposure in terms of: (1) knowing someone who was a victim of violence; (2) having witnessed violent events *in real life*; (3) having been a victim of a violent incident; and (4) having perpetrated a violent act. Comparable to the work mentioned above, information about such forms of exposure was obtained from students enrolled in grades 2 through 12.

Although not a universal experience, victimization occurs at levels approaching 50% for threats and 40% for forced entry into their home. Indirect victimization (e.g., being a witness to a violent act) levels were consistently higher for incidents ranging from drug use to shootings, stabbings, and muggings. Reports of seeing a dead body in the neighborhood ranged from 24 to 40%. Using an extension of the Richters and Saltzman measure, Singer and colleagues (Singer et al., 1995; Song et al., 1998) validated these reports in their surveys of high-school students ($N = 3735$). As expected, considerable variability is present across settings. Nevertheless, levels of victimization and witnessing (both at home and in school) justify recognition of PCV as a major environmental contaminant. Apfel and Simon's (1996) description of communal violence as referring to settings in 'which every child has witnessed or expects to witness violence and has been or expects to be violated' (pp. 4–5), seems to apply to many of the nation's neighborhoods and schools.

Increasingly, evidence suggests that there are important associations between perceptions of neighborhood problems and health-related quality of life (HRQL). The perceptions of the same neighborhood are often quite divergent between different people living in the same place (Dahl et al., 2010). Perceptions of neighborhood problems are often found to be negatively predictive of psychological, social, and environmental HRQL. This suggests that the various aspects of well-being may be influenced as much by perceptions of neighborhood problems as actual neighborhood characteristics (Erin et al., 2012).

Assessing PCV's Health Impact

Older studies also report on PCV's toxicity. Saltzman (1992), for example, observed significantly higher levels of generalized emotional distress in PCV-exposed fifth and sixth graders. This pattern is also reported by Richters and Martinez (1993), Martinez and Richters (1993), Osofsky et al. (1993), and Bell and Jenkins (1993). Across these studies of preadolescents, PCV exposure relates positively to affective and vegetative signs of depression and indices of stress-related disorders. With both preadolescents and adolescents, Singer et al. (1994) linked PCV exposure and increased aggression, conduct disorder, running away, and lowering of self-esteem and social competence. Support for the link between PCV exposure and involvement in violent acts is reported by Saltzman (1995). In Saltzman's study, PCV exposure significantly related to adolescent involvement in antisocial and violent acts independent of peer, familial, and other recognized predictors of aggression during that developmental period. Rubinetti (1996) linked PCV exposure and violent activity most strongly in youth with decreased empathy, hopelessness, and lowered self-esteem. Provocatively, Song et al.'s (1998) findings suggest that indirect victims are more likely than direct victims to aggress *in anticipation of* violent acts. Although further study is needed, PCV may: (1) create interpersonal distance among the residents of an affected setting; (2) desensitize them to the consequences of violence on others; and (3) heighten the perceived need to defend oneself against assault. Overall, there appears to be substantial evidence that heightened exposure to PCV increases the likelihood that one will engage in violence. In turn, the resulting acts of violence increase levels of PCV exposure, providing a mechanism for a recursive, continuing cycle.

Two other studies (Cooley-Quille and Lorion, 1999; Scarpa et al., 2000) reinforce the physical impacts of PCV exposure. Each study found physiological concomitants of PCV exposure, including increased heart rate, respiration, and

cortisol levels, long-established indices of heightened stress, and arousal. Cooley-Quille and Lorion (1999) also reported preliminary evidence of a link between PCV and sleep disorders. Anecdotal reports from youth living in PCV-heightened neighborhoods report difficulty falling asleep, repeatedly waking up in response to gunshots, loud noises, and police activity and, generally, only rarely sleeping throughout the night. If these reports are confirmed along with further information about the vegetative effects of PCV exposure, new understandings may emerge which change existing assumptions about the lack of impulse control and academic failure, often associated with urban poverty, from attributions about personal failure to questions about environmental toxicity. Thus, although examination of the physiological consequences of living or working within settings characterized by high PCV levels is just beginning, preliminary results suggest that these impacts are real, extensive, and potential triggers for further pathogenesis across multiple parameters of functioning.

Such findings support the hypothesized toxicity of PCV exposure, which appears to impair cognitive, emotional, behavioral, and biological functioning simultaneously, sequentially, or synergistically. As yet unexamined systematically, however, is the breadth of these 'PCV-consequences' (assuming they are causally linked) within settings and communities. In essence, we have yet to study how the toxin spreads and the transmission routes by which the occurrence of violent events (presumably their nature, frequency, and intensity would influence the spread of effects) contaminates the environment. Longitudinal information is necessary to causally link exposure and its observed concomitants. It is also necessary to determine at what point interventions can be introduced to protect against or, if necessary, to treat such effects.

It should be recognized that what is presently known about PCV relates almost entirely to school-aged youth. We know little about the effect of maternal PCV exposure on neonates, of exposure on toddlers and preschoolers, and even less about the nature, extent, and consequences of exposure on adults including on senior adults for whom PCV may lead them to limit their movements outside of their homes and thus this contact with others. What of the parents, grandparents, and others who live in such neighborhoods? Within schools, what of the staff and teachers who may live in such neighborhoods, travel through such neighborhoods, or hear the tales of their students?

Work reported to date pays little attention to these members of exposed communities, except perhaps to suggest that they could buffer youth against the impacts of PCV. Whether we can assume the capacity of exposed adults to do so merits careful examination!

We must therefore examine exposure to PCV across the entirety of a community's population and settings. Anecdotal reports, informal focus groups, and clinical encounters suggest that exposure reduces the quality of teachers' instructional effectiveness, capacity to apply classroom management strategies, and availability to students before or after the school day. Reasonably, one might expect that PCV similarly disrupts parent–child, parent–teacher, and teacher–teacher interactions. As reported above, Song et al. (1998) found a significant link between indirect exposure and anticipatory aggression. Ambiguous interpersonal cues may be interpreted by youth (and perhaps adults?) as threatening, leading to what is assumed by the perceiver as a defensive act. Presumably (although we do not really know), when self-defense seems unlikely to succeed, ambiguity leads to retreat, avoidance, or perhaps acquiescence with stated or unstated demands.

If this explanation of Singer's finding is valid, would it not extend to adult interactions within and across generations? Might it not explain, at least in part, what from the outside appears as hypersensitivity to insult or disrespect, the immediacy with which one asserts one's right, or the reluctance to give way, lest courtesy be mistaken for weakness or fear? Might it not explain urban parents' use of physical discipline or its reverse, that is, seeming inability or unwillingness to control their children's behavior? What of neighborhoods in which the elderly limit, and the young have limited, time outside of the home because of concerns about danger and vulnerability? Would not isolated streets facilitate violence with immunity? Would not the biological concomitants related to stress, sleep deprivation, and concentration impair adults' (be they teachers, storekeepers, bus drivers, or elderly neighbors) capacity to interact comfortably with each other and to neutralize or at least dilute PCV's toxicity and thereby protect youth? To ignore the status of adults in such settings implies that they are unaffected by PCV or other community toxins.

Albeit at levels of concern, PCV is linked to increased but not universal pathology. Not only does a majority of those so exposed not manifest pathology, but most residents do not report exposure. PCV exposure appears to relate not only to the occurrences of violent acts, but also to awareness of and sensitivity to those occurrences. Somehow, many exposed are seemingly unaware of that and, potentially, for them there is some bliss in ignorance. The individual and ecological factors that determine sensitivity remain to be investigated. That such differences exist, however, argues for attending to issues of adaptation, coping, and wellness. As noted recently:

> By implication it seems reasonable to view wellness as representing the outcome of normative developmental processes, in programming terms, as the default option. In the absence of that relatively rare combination of individual and situational circumstances, individual development appears to proceed within socially expected and appropriate boundaries reflecting the general capacity to adapt to situational demands and cope with stressful events. Just as canalization accounts for the limits within which genetic processes emerge within species (Gottlieb, 1992), an analogous process may account for the human capacity to adapt to a multitude of settings and circumstances, and to cope with seemingly debilitating and pathogenic stressors.
>
> (Lorion, 2000: pp. 13–14)

If wellness, adaptation, and adjustment represent the default option in human development, and individuals, families, neighborhoods, schools, and communities in general cope with PCV, economic (dis)advantage, familial instability, critical life events, inadequate school budgets, and so on (the length of one's list merely underscores the point), understanding those mechanisms may inform the design of interventions to protect those who cannot cope. Answers to questions about how to optimize development and coping

must be out there, in the homes, the schools, the classrooms, and the neighborhoods. To find those answers, we in the social sciences must enter the communities to observe, inquire, and learn. I propose we go not as experts bringing esoteric knowledge, but as students and collaborators, hopefully bringing tools and perspectives whose utility will be determined relative to what already exists, and what can be sustained over time and across levels of resources.

See also: Ecology and Health; Environmental Health and Safety: Social Aspects; Globalization and Health; Health Interventions, Community-based; Occupational Health; Urban Violence; Violence in Anthropology; Violence: Public; War, Political Violence, and Effects on Children.

Bibliography

Andrews, G., Charney, D., Sirovatka, P., Regier, D., 2009. Stress Induced and Fear Circuitry Disorders: Refining the Research Agenda for DSM-V. American Psychiatric Publishing, Arlington, VA.

Apfel, R.J., Simon, B. (Eds.), 1996. Minefields in Their Hearts. Yale University Press, New Haven, CT.

Bell, C.C., Jenkins, E.J., 1993. Community violence and children on Chicago's southside. Psychiatry: Interpersonal and Biological Processes 56, 46–54.

Bell, C.C., Hildreth, C.J., Jenkins, E.J., Levi, D., Carter, C., 1988. The need for victimization screening in a poor, outpatient medical population. Journal of the National Medical Association 80, 853–860.

Belmont Report, 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Retrieved 13 December, 2011, from: hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Cooley-Quille, M.C., Lorion, R.P., 1999. Adolescents' exposure to community violence: sleep and psychophysiological functioning. Journal of Community Psychology 27, 367–376.

Dahl, T., Ceballo, R., Huerta, M., 2010. In the eye of the beholder: mothers' perceptions of poor neighborhoods as places to raise children. Journal of Community Psychology 38, 419–434. http://dx.doi.org/10.1002/jcop.20372.

Erin, M.H., Shepherd, D., Welch, D., Dirks, K.N., McBride, D., 2012. Perceptions of neighborhood problems and health-related quality of life. Journal of Community Psychology 40, 814–827. http://dx.doi.org/10.1002/jcop.21490.

Fox, N.A., Leavitt, L.A., 1995. The Violence Exposure Scale for Children (VEX). University of Maryland, College Park, MD.

Gottlieb, G., 1992. Individual Development and Evolutions: The Genesis of Novel Behavior. Oxford University Press, New York.

Hill, H.M., Madhere, S., 1996. Exposure to community violence and African American children: a multidimensional model of risks and resources. Journal of Community Psychology 24, 26–43.

Jenkins, E.J., Thompson, B., 1986. Children Talk about Violence: Preliminary Findings from a Survey of Black Elementary School Children. Paper presented at the 19th Annual Convention of the Association of Black Psychologists, Oakland, CA.

Krug, E.G., et al. (Eds.), 2002. World Report on Violence and Health. World Health Organization, Geneva.

Lorion, R.P., 1998. Exposure to urban violence: contamination of the school environment. In: Eliott, D.S., Williams, K., Hamburg, B. (Eds.), Violence in American Schools. Cambridge University Press, New York, pp. 293–311.

Lorion, R.P., 1999. Community prevention and wellness. In: Herson, M., Ammerman, T. (Eds.), Advanced Abnormal Child Psychology. Erlbaum, Hillsdale, NJ, pp. 251–266.

Lorion, R.P., 2000. Theoretical and evaluation issues in the promotion of wellness and the protection of 'well enough'. In: Cichetti, D., Rappaport, J., Sandler, I., Weissberg, R. (Eds.), The Promotion of Wellness in Children and Adolescents. Sage, Thousand Oaks, CA, pp. 1–28.

Lorion, R.P., 2001a. Violence as a problem of health: emotional, behavioral and physical concomitants of exposure to pervasive community violence. In: Smelser, N., Baltes, P. (Eds.), Encyclopedia on the Social and Behavioral Sciences. Elsevier Publishers, London.

Lorion, R.P., 2001b. Exposure to urban violence: shifting from an individual to an ecological perspective. In: Schneiderman, N., Tomes, H., Gentry, J., Speers, M., Silva, J. (Eds.), Integrating Behavioral and Social Sciences with Public Health. American Psychological Association, Washington, DC.

Lorion, R.P., Saltzman, W., 1993. Children's exposure to community violence: following a path from concern to research to action. Psychiatry: Interpersonal and Biological Processes 56, 55–65.

Lorion, R.P., Brodsky, A.E., Cooley-Quille, M., 1998. Exposure to pervasive community violence: resisting the contaminating effects of risky settings. In: Biegel, D.E., Blum, A. (Eds.), Innovations in Practice and Service Delivery across the Life Span. Oxford University Press, New York, pp. 124–144.

Martinez, P., Richters, J.E., 1993. The NIMH Community Violence Project: II. Children's distress symptoms associated with violence exposure. Psychiatry: Interpersonal and Biological Processes 56, 22–35.

Osofsky, J.D., Wewers, S., Hann, D.M., Fick, A.C., 1993. Chronic community violence: what is happening to our children? Psychiatry: Interpersonal and Biological Processes 56, 36–45.

Primavera, J., Brodsky, A.E., 2004. Introduction to the special issue on the process of community research and action. American Journal of Community Psychology 33 (3–4), 177–179.

Rappaport, J., 1981. In praise of paradox: a social policy of empowerment over prevention. American Journal of Community Psychology 9, 1–25.

Raviv, A., Erel, O., Fox, N.A., Leavitt, L.A., Raviv, A., Dar, I., Shahinfar, A., Greenbaum, C.W., 2001. Individual measurement of exposure to everyday violence among elementary school children across various settings. Journal of Community Psychology 29 (2), 117–140.

Richters, J.E., Martinez, P., 1990. Checklist of Child Distress Symptoms: Self-Report Version. National Institute of Mental Health, Rockville, MD.

Richters, J.E., Martinez, P., 1993. The NIMH Community Violence Project: I. Children as victims of and witnesses to violence. Psychiatry: Interpersonal and Biological Processes 56, 7–21.

Richters, J.E., Saltzman, W., 1990. Survey of Exposure to Community Violence – Parent Report Version. Unpublished measure. Child and Adolescent Disorders Research, National Institute of Mental Health.

Richters, J.E., Martinez, P., Valla, J.P., 1990. Levonn: A Cartoon-Based Structured Interview for Assessing Young Children's Distress Symptoms. National Institute of Mental Health, Washington, DC.

Roberts, L.W., 2013. Community-Based Participatory Research for Improved Mental Healthcare: A Manual for Clinicians and Researchers. Springer, New York, NY.

Rubinetti, F., 1996. Empathy, Self-Esteem, Hopelessness and Belief in the Legitimacy of Aggression in Adolescents Exposed to Pervasive Community Violence (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Saltzman, W., 1992. The Effect of Children's Exposure to Violence (Unpublished master's thesis). University of Maryland, College Park, MD.

Saltzman, W., 1995. Exposure to Community Violence and the Prediction of Violent Antisocial Behavior in a Multi-Ethnic Sample of Adolescents (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Scarpa, A., Fikretoglu, D., Luscher, K., 2000. Community violence exposure in a young adult sample: II. Psychophysiology and aggressive behavior. Journal of Community Psychology 28, 417–426.

Shahinfar, A., Fox, N.A., Leavitt, L.A., 2000. Preschool children's exposure to violence: relation of behavior problems to parent and child reports. American Journal of Orthopsychiatry 70, 115–125.

Singer, M.I., Anglin, T.M., Song, L., Lunghofer, L., 1994. The Mental Health Consequences of Adolescents' Exposure to Violence. Case Western Reserve University, Cleveland, OH.

Singer, M.I., Anglin, T.M., Song, L., Lunghofer, L., 1995. Adolescents' exposure to violence and associated symptoms of psychological trauma. Journal of the American Medical Association 273, 477–482.

Song, L., Singer, M., Anglin, T.M., 1998. Violence exposure and emotional trauma as a contributor to adolescents' violent behavior. Archives of Pediatric Adolescent Medicine 152, 531–536.

Violence, Criminalization, and the Homeless

Amy M Donley, James D Wright, and Jerian Benwell-Lybarger, University of Central Florida, Orlando, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Homeless people are sometimes the perpetrators and quite often the victims of crime and violence. This article reviews the available research on this topic, focusing mainly on studies conducted in the United States. Homeless teens, women, the mentally ill, and the substance-abusive are shown to be especially high-risk subpopulations. Random acts of violence committed against homeless people are evidently on the rise, both in the United States and around the world. Around the world, many communities have responded to the rise in homelessness by criminalizing homeless behaviors. Unfortunately these measures do not assist people in exiting homelessness.

In popular thinking, homelessness and crime are tightly interwoven (Amster, 2008). Homeless shelters and other locales frequented by homeless people are routinely avoided because they are perceived as crime hot spots, and homeless people themselves are routinely avoided because of their real or imagined criminal potential (Snow et al., 1989). Homeless individuals are often seen as irresponsible and deviant, leading to a public perception that equates homelessness and crime (Barak, 2002). While the media and many policies targeting homeless people focus on the perpetration of crime by homeless people, scholarly research has tended to focus on the victimization of homeless people (Crawford et al., 2011).

The Homeless as Perpetrators of Violence

Most of what we know about criminal perpetration by homeless people comes from studies examining previous arrest and prison experience. In a study of 767 homeless people in Austin Texas, Snow et al. (1989) found that the typical arrest of a homeless person was for a minor offense such as vagrancy or shoplifting. Out of their total sample, 32% had been arrested at least once, but the vast majority of those arrests were for nonviolent crimes. While homeless men are no more likely than the general male population to commit violent crimes, they are arrested more frequently for nonviolent crimes (Snow et al., 1989). A national survey of homeless providers and homeless clients in the United States found that 49% had been incarcerated for five or more days in jail, but only 18% had spent time in a federal prison (Burt et al., 1999). The authors note that, "some of these jail experiences may have been a direct result of their homelessness (i.e., the charges might be for behaviors that are difficult to avoid if one is homeless, such as loitering)" (Burt et al., 1999: 25). Kushel et al. (2005) found that among their sample of ~1500 homeless people, nearly a quarter had spent time in prison. Drug use (specifically cocaine, methamphetamine, and heroin) as well as mental illness were positively correlated with past prison experience. In the United Kingdom, Fitzpatrick et al. (2013) conducted a statistically generalizable survey of homeless people and found that 46% had been incarcerated at some point in their lives. However, detailed information on the nature of their incarceration was not available and, thus, the nature of the crimes committed is not known.

Both mental illness and substance abuse among homeless people increase the likelihood that they will be arrested and incarcerated. In a study of homeless adults, 20% of those with substance abuse issues had been arrested in the past year compared to 10% of those that did not have a substance abuse problem (O'Toole et al., 2004). McGuire and Rosenheck (2004) found in their study of 7000 homeless people with serious mental illness that two-thirds had histories of incarceration. However, of these about half had spent a total of six months or less of their lives incarcerated. Among those with more extensive incarceration histories, substance abuse was a positive correlate, illustrating the relationship between mental illness and substance abuse.

In the United States, there are well-established links between mental illness, incarceration, and homelessness (DeLisi, 2000; Ditton, 1999; Snow et al., 1989). Deinstitutionalization of the mentally ill in the United States began in 1955; thus, by 1980 or so (usually taken as the onset of the large upsurge of homelessness that happened in American and cities abroad), most persons destined to be deinstitutionalized already had been. The proper concern for the past two decades has, therefore, been not with deinstitutionalized but rather with the 'never institutionalized,' namely, mentally ill homeless people who had never been brought into a system of care. A related and even more pressing concern has been with the 'trans-institutionalized'; that is, those who might have been in mental hospitals in years past but are now in prisons and jails instead (Raphael and Stoll, 2013). Many homeless mentally ill people spend a great deal of time incarcerated; indeed, Torrey et al. (2010) conclude that there are more mentally ill people in jails and prisons than in psychiatric hospitals. Similarly, Markowitz (2006) found that in the United States, increased mental health and psychiatric hospital capacity was correlated with lower levels of homelessness. Most homeless people with mental illness problems in the United States find that there is nowhere to go for help. These individuals are often on the streets, increasing the likelihood of arrest and incarceration.

Among homeless youth, less is known about criminal perpetration. However, in the research that has been done, victimization and perpetration are both salient parts of a typical homeless youth's life (Crawford et al., 2011). One of the first studies examining criminal perpetration among homeless youths found that nearly 1 in 6 of the Los Angeles area youths in their sample had attacked someone with

a weapon while being on the streets, although this percentage was much lower than the rates of victimization these youths experienced (Kipke et al., 1997). In the Crawford et al. study, homeless youth who were female, nonminority, and heterosexual were less likely to engage in violence than their opposites; alcohol and drug-abusers were more likely to engage in violence, a common finding (Chen et al., 2006; Kipke et al., 1997). Moreover, increased time on the streets is positively correlated to rates of criminal perpetration (Whitbeck and Hoyt, 1999).

Most studies of violence perpetration among the homeless depict it as relatively minor and benign and the crimes they commit largely as survival behaviors, and not predatory criminality. While it is clear that the homeless are responsible for some crime and violence, it is yet to be established that the crime and violence they commit is any more common or more serious than that perpetrated by equally impoverished but housed individuals.

The Homeless as Victims of Violence

Most research examining homelessness and violence focuses on the victimization of homeless people and finds an association between homelessness and criminal victimization (Lee and Schreck, 2005; Thrane et al., 2006; Whitbeck and Simons, 1990). Even when compared to similar housed counterparts, homeless people report higher levels of victimization (Lee and Schreck, 2005), often by wide margins. It is nearly axiomatic that life on the streets or in the shelters exposes homeless people to high crime risks, and while the reward for victimizing a homeless person may be small, they are exceedingly easy targets, especially if they are also substance abusive, mentally ill, or both.

Violence as a Precursor to Homelessness

Victimization is both a cause and consequence of homelessness. Many studies have found an association between adult experiences of homelessness and childhood experiences of abuse and neglect (Mayock and Sheridan, 2012; McIntyre and Widom, 2011; Paradise and Cauce, 2002; Thrane et al., 2006) and repeated bullying (Bouffard and Koeppel, 2012). In a study of homeless adults in the United Kingdom, 22% reported being victims of physical abuse while 16% reported being victims of sexual abuse during their childhood (Fitzpatrick et al., 2013). Not only does the experience of childhood abuse and neglect increase the likelihood that people will become homeless, it also increases the likelihood that adult homeless people will experience victimization (Lee and Schreck, 2005; McIntyre and Widom, 2011; Jasinski et al., 2010).

Homeless Youth

Like homeless adults, homeless youths report higher rates of child abuse and neglect as compared to nonhomeless youths (Stoltz et al., 2007). Violence at home is, for many youths, the reason they become homeless in the first place. Oftentimes, they leave of their own volition (sometimes fleeing domestic situations so abusive that life on the streets becomes the preferred alternative) while other times they are kicked out. In a survey of homeless youth aged 13–17 years in Australia, 45% had experienced violence at home prior to becoming homeless (Australian Institute of Health and Welfare, 2008).

Homeless Women

Studies of victimization of homeless people have consistently found that homeless women are more likely to be victims of violence, particularly sexual assault and rape, as compared to homeless men (Evans and Forsyth, 2004; Kim et al., 2010; Kushel et al., 2003; Lee and Schreck, 2005). While homeless, the lack of suitable and stable shelter increases the likelihood that a woman will be a victim of violence, and here too, this is particularly true if substance abuse or mental illness is also part of the picture.

A great deal of research concerns the role of violence in the lives of homeless women. While violence is unfortunately a salient part of homelessness for many people, this is particularly true among women. For many homeless women, victimization begins at home in childhood and continues as an adult experiencing homelessness. Across numerous studies, homeless women who report having been abused as a child also report higher rates of violent victimization as an adult. Research examining why this correlation exists has not been conclusive. Some of the cited relevant factors include a lack of skill in developing healthy and safe relationships, a lack of knowledge of what constitutes a healthy relationship, the inability to recognize warning signs of an abusive relationship, general low self-esteem, and depression (see Jasinski et al., 2010; for an overview of relevant literature).

Domestic violence is a main cause of homelessness among women. In the US data reported by Jasinski et al. (2010), approximately three-quarters of homeless women reported histories of domestic violence and about one in four cited intimate partner abuse as a or the main reason for their current homelessness. Other studies have suggested that as much as half of all homelessness among women is precipitated by domestic violence and abuse.

While the majority of research in this area comes from the United States, the connection between violence and homeless women is not unique to the United States. In 2007, half of the homeless women in Australia seeking homeless services cited violence as the cause of their homeless status (Homelessness Taskforce, 2008). In a study of 60 homeless women in Ireland, 72% reported being victimized as children and two-thirds reported experiencing intimate partner violence as an adult (Mayock and Sheridan, 2012).

Victimization While Homeless

Homeless people have been victims of physical attacks and even murder around the globe including in the United States (National Coalition for the Homeless, 2012) and Japan (Fujii and Tamaki, 2003). Among the homeless adult population, those sleeping rough or unsheltered and those that are mentally ill (Dennis and Steadman, 1991) are particularly vulnerable to violent attacks. A pertinent CNN story of 20

February 2007 referred to this as 'sport killing' and reported that "Teen 'sport killings' of homeless [are] on the rise."

Since 1999, the National Coalition for the Homeless (NCH) has issued an annual report on attacks against homeless people in the United States. Through the end of 2011, the NCH has documented 1289 attacks of violence against homeless persons perpetrated by housed individuals in the United States, including 339 murders (National Coalition for the Homeless, 2012). Among these documented attacks, the most common perpetrator is a young man (under the age of 30 years) while the typical victim is an older man (aged 40–60 years). The NCH focuses on attacks on homeless people perpetrated by housed people. Information about violence perpetrated against homeless people by other homeless people is extremely scarce.

In regard to homeless youths, one of the most commonly found salient factors is the history of child abuse. Among homeless youth this has been associated with greater levels of victimization as adults in many studies (Melander and Tyler, 2010; Tyler et al., 2001; Tyler and Melander, 2013; Whitbeck et al., 2001). Among the homeless youth population, those that identify as gay, lesbian, or bisexual seem to be at disproportionate risk of victimization as compared to their heterosexual counterparts (Tyler, 2008).

In the United States, several states have responded to attacks against homeless people by adding housing status to their state hate-crime statutes. Despite several attempts, housing status is not included in the United States federal hate-crime definition. Likewise, while there have been widely publicized attacks against homeless persons in the United Kingdom, to date homeless people are not protected under the UK hate-crime law (Garland, 2012). While hate crimes of any sort are relatively rare occurrences, many argue that homeless people are particularly vulnerable and that these occurrences serve to further delineate between the value and rights of the housed versus those of the unhoused (Wachholz, 2009).

Criminalization of Homelessness as a Policy Response

Across the United States, Europe, and elsewhere, the response to the issue of homelessness has often been to criminalize behaviors associated with homeless people (for the US case, see Donley and Wright, 2008; for a case study in Europe, see Bence and Udvarhelyi, 2013; for the Asian case, see e.g., Wardhaugh, 2012). These criminalization measures are not designed to address the causes of homelessness but rather serve as measures to remove homeless people from public places. Criminalization tactics come in many forms and vary across jurisdictions. Some of the more common measures include the enactment of law targeting homeless-related behaviors and 'street sweeps,' where homeless people are forcibly removed from certain areas within cities.

Law enforcement has attempted to protect citizens, business owners, and public spaces by criminalizing unavoidable, yet undesirable actions, but this approach may be ineffective and counterproductive. "Police, attorneys, judges, and other actors in the criminal justice system … see panhandling as a minor offense the prosecution of which consumes valuable time, leaves root causes intact, and fails to deter reoccurrence" (Lee and Farrell, 2003: 318). Deterrence cannot have an effect when the person sometimes has no other choice but to commit a crime for basic subsistence (Davis, 1992). The NLCHP (2011) concurs that the criminalization of homelessness only has temporary effects on homeless criminal activity, jailing them for a short period of time then releasing them with no choice but to commit the same 'subsistence crimes' again.

Los Angeles is an example of a major American city that has used aggressive policing as a response to homelessness. This effort is also one of the few that was subsequently examined in terms of cost and effectiveness. The Safer Cities Initiative carried out in the notorious LA Skid Row area issued citations at 50 to 70 times the average rate and only 21 of the 1346 nondrug arrests in the area were for violent crimes (Blasi, 2007). Many individuals were cited for small crimes such as sitting on the sidewalk or throwing cigarettes on the ground. Homeless individuals often do not have the ability to pay the associated fines, would consequently be arrested, and released back onto the street with no other options. In just under a year of the Safer City Initiative, 24 individuals were arrested 201 times at an estimated cost of 3.6 million US dollars. The police officers used to carry out this initiative cost the city of Los Angeles an additional 6 million dollars. In the end, there was a decrease in the number of homeless individuals in Skid Row but an increase in the number of homeless individuals in surrounding areas. Thus, these measures did not decrease the number of homeless people but rather resulted in nothing more than some relocation of the homeless to other areas.

Budapest is an example of a European city that has attempted to end homelessness by outlawing it. Among the measures adopted by the Hungarians is a federal law allowing local municipalities to ban the 'inadequate use' of public spaces; in Budapest, this became a prohibition on the use of public space for 'habitual residence' or storage of one's belongings (Bence and Udvarhelyi, 2013). In the same vein, various districts in Budapest have been declared 'homeless-free zones,' and there has been a recent Constitutional amendment guaranteeing the right of local governments to criminalize 'habitual residence' in places not intended for human habitation.

Of all the behaviors associated with homelessness in the public mind, panhandling is evidently one of the most offensive (Lee and Farrell, 2003). As a result, this behavior has been a main target of effort to 'clean up the streets.' However, "contrary to common belief, panhandlers and homeless people are not necessarily one and the same. Many studies have found that only a small percentage of homeless people panhandle, and only a small percentage of panhandlers are homeless" (Scott, 2002). Although exact numbers are unavailable, estimates of the percentage of homeless people that engage in panhandling range from 5 to 40% (Lee and Farrell, 2003). Moreover, many panhandlers are not homeless. Goldstein (1993) found in his research in New Haven, Connecticut that only one-third of panhandlers were actually homeless. Nevertheless, panhandling and homelessness are intimately linked in public thinking.

The criminalization of homelessness is an international concern as research has pointed to an increase in privatization of public spaces, anti-homeless legislation, and practices of forced displacement in many countries including Canada,

United Kingdom, Germany, Hungary, Australia, New Zealand, the Netherlands, and Japan, as well as the United States.

China too has undergone major changes in the way it responds to homeless people over the past 15 years. Centers that used to be places where homeless people were incarcerated for activities such as begging have been converted into homeless aid centers. While many hail this change as a major improvement, criminalization persists. Anti-begging laws are common and beggars are portrayed in the media as criminals (Kennett and Mizuuchi, 2010).

Conclusion and Future Research

The criminalization of homelessness in the United States has been the subject of numerous lawsuits that argue certain ordinances infringe on citizens' freedom of speech and movement as well as freedom from unreasonable search and seizure, cruel and unusual punishment, discrimination, and forced evictions. The Supreme Court has ruled against the vagueness of some anti-homeless legislation (*Papichristou vs City of Jacksonville*). It has also been ruled that certain anti-homeless legislation is unconstitutional if it prohibits basic 'life-sustaining activities' and if alternatives do not exist. In the absence of these conditions, people cannot be prosecuted for things such as sleeping in the park or relieving themselves in public (*Pottinger vs City of Miami*, *Jones vs City of Los Angeles*). The United States Congress passed the Helping Families Save their Homes Act of 2009 requiring the development of alternative solutions to criminalizing homelessness. Anti-panhandling initiatives have been more successful. Although general panhandling is protected under the First Amendment in the United States, aggressive panhandling has been legally banned in communities across the country (Wright and Donley, 2011).

While this is not the place to review all the root causes of homelessness, it can be said that homelessness results when there are more low-income people than there are low-income housing units. At its base, in other words, homelessness results from inadequate incomes and inadequate supplies of affordable housing. Being homeless places people at high risk for criminal victimization, and this often puts these same people in positions where petty criminality is nearly required for daily survival. Criminalizing homelessness by criminalizing the behaviors associated with homeless does nothing to address root causes; this attacks the symptoms of the problem rather than the problem itself. No city in the world has eliminated homelessness via criminalization, although many have tried.

Future studies should examine the potential impact of programs designed to address root causes of homelessness as opposed to criminalizing homelessness-related behaviors. In addition, it is important to better understand levels of criminality among homeless people. While some studies have been done to date, typically these must rely on arrest data, self-reports among sheltered homeless individuals, or analysis of individual cases. A great deal of research today examines reintegration of individuals exiting criminal justice institutions and reentering communities. The reintegration of homeless individuals presents unique challenges and, thus, research that focuses on these issues would be particularly important.

See also: Domestic Violence: Sociological Perspectives; Homelessness in the United States; Intimate Partner Abuse, Applied Research on; Poverty, Sociology of; Street Children: Cultural Concerns; Urban Poverty in Neighborhoods; Violence: Public.

Bibliography

Ammann, J.J., 2000. Addressing quality of life crimes: criminalization, community courts and community compassion. Saint Louis University Law Journal 44, 811–820.

Amster, R., 2008. Lost in Space. LFB Scholarly Publishing LLC, New York.

Australian Institute of Health and Welfare, 2008. Making Progress: The Health, Development and Wellbeing of Australia's Children and Young People (Cat. no PHE 104). Author, Canberra.

Barak, G., 2002. Integrative theories. In: Levinson, D. (Ed.), Encyclopedia of Crime and Punishment. Sage, Thousand Oaks.

Bence, Rita, Udvarhelyi, Eva Tessza, 2013. The growing criminalization of homelessness in Hungary – a brief overview. European Journal of Homelessness 7 (2), 133–144.

Blasi, G., The UCLA School of Law Fact Investigation Clinic, 2007. Policing Our Way Out of Homelessness?: The First Year of the Safer Cities Initiative on Skid Row.

Binder, S., 2002. The Homeless Court Program: Taking the Court to the Streets. Available at: http://www.americanbar.org/content/dam/aba/events/homelessness_poverty/lwteh_article.authcheckdam.pdf.

Burt, M.R., Aron, L.Y., Douglas, T., Valente, J., Lee, E., Iwen, B., 1999. Homelessness: Programs and the People They Serve. Urban Institute, Washington, DC.

Bouffard, L.A., Koeppel, M.D.H., 2012. Understanding the potential long-term physical and mental health consequences of early experiences of victimization. Justice Quarterly. http://dx.doi.org/10.1080/07418825.2012.734843.

Chen, X., Thrane, L., Whitbeck, L., Johnson, K., 2006. Mental disorders, comorbidity, and postrunaway arrests among homeless and runaway adolescents. Journal of Research on Adolescence 16 (3), 379–402.

Crawford, D.M., Whitbeck, L.B., Hoyt, D.R., 2011. Propensity for violence among homeless and runaway adolescents: an event history analysis. Crime & Delinquency 57, 950–968.

Daily, China, June 2003a. Tragedy Spurs End to 21 Year-old Rule on Vagrants. http://www.chinadaily.com.cn/en/doc/2003-6/20/content_239912.htm.

Daily, China, August 2003b. Vagrants Get Aid as New System Begins. http://www.chinadaily.com.cn/en/doc/2003-08/01/content_250903.html.

Daily, China, March 2004. Downtown Beggars Can Be Choosers. http://www.chinadaily.com.cn/english/doc/2004-03/18/content_315678.htm.

Davis, M., 1992. Fortress Los Angeles: the militarization of urban space. In: Sorkin (Ed.), Variations on a Theme Park, pp. 154–180.

DeLisi, M., 2000. Who is more dangerous? Comparing the criminality of adult homeless and domiciled jail inmates: a research note. International Journal of Offender Therapy and Comparative Criminology 44, 59–69.

Dennis, D.L., Steadman, H.J., 1991. The Criminal Justice System and Severely Mentally Ill Persons: An Overview. U.S. Department of Health and Human Services, Washington, DC.

Ditton, P.M., 1999. Mental Health and Treatment of Inmates and Probationers. NCJ 174463. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC.

Donley, A.M., Wright, J.D., 2008. Cleaning up the streets: community efforts to combat homelessness by criminalizing homeless behaviors. In: McNamara, R.H. (Ed.), Homelessness in America. Praeger Publishers, Westport, CT, pp. 75–91.

Evans, R.D., Forsyth, C.J., 2004. Risk factors, endurance of victimization, and survival strategies: the impact of the structural location of men and women on their experiences within the homeless milieus. Sociological Spectrum 24, 479–505.

Fujii, K., Tamaki, M., 2003. Henken Kara Kyousei (From the Prejudice to the Association). Fubaisha, Nagoya, Japan.

Fitzpatrick, S., Bramely, G., Johnsen, S., 2013. Pathways into multiple exclusion homelessness in seven UK cities. Urban Studies 50 (1), 148–168.

Gaetz, S., 2004. Safe streets for whom? Homeless youth, social exclusion, and criminal victimization. Canadian Journal of Criminology and Criminal Justice 46 (4), 423–455.

Garland, J., 2012. Difficulties in defining hate crime victimization. International Review of Victimology 18, 25–37.

Goldstein, B.J., 1993. Panhandlers at Yale: a case study in the limits of law. Indiana Law Review 27, 295–359.

de Graaf, W., Van Doorn, L., Kloppenburg, R., Akkermans, C., 2011. Homeless families in the Netherlands: intervention policies and practices. Journal of Social Research and Policy 2 (1), 5–18, 9 (3), 1281–1311.

Hansel, K., 2011. Constitutional othering: citizenship and the insufficiency of negative rights-based challenges to anti-homeless systems. Northwestern Journal of Law and Social Policy 6 (2), 445–472.

Homelessness Taskforce, 2008. The Road Home: A National Approach to Reducing Homelessness. Commonwealth of Australia, Canberra.

Jasinski, J., Wesely, J., Mustaine, E., Wright, J.D., 2010. Hard Lives, Mean Streets: Violence in the Lives of Homeless Women. University Press of New England, Boston, MA.

Kelling, G.L., Wilson, James Q., 1982. Broken Windows. The Atlantic.

Kennett, P., Mizuuchi, T., 2010. Homelessness, housing insecurity and social exclusion in China, Hong Kong and Japan. City, Culture and Society 1, 111–118.

Kim, M.M., Ford, J.D., Howard, D.L., Bradford, D.W., 2010. Assessing trauma, substance abuse, and mental health in a sample of homeless men. Health and Social Work 35, 39–48.

Kipke, M.D., Simon, T.R., Montgomery, S.B., Unger, J.B., Iversen, E.F., 1997. Homeless youth and their exposure to and involvement in violence while living on the streets. Journal of Adolescent Health 20, 360–367.

Kushel, M.B., Hahn, J.A., Evans, J.L., Bangsberg, D.R., Moss, A.R., 2005. Revolving doors: imprisonment among the homeless and marginally housed population. American Journal of Public Health 95 (10), 1747–1752.

Kushel, M.B., Evans, J.L., Perry, S., Robertson, M.J., Moss, A.R., 2003. No door to lock: victimization among homeless and marginally housed persons. Archives of Internal Medicine 163, 2492–2499.

Lee, B., Farrell, C., 2003. Buddy, can you spare a dime? Homelessness, panhandling, and the public. Urban Affairs Review 38, 299–324.

Lee, B.A., Schreck, C.J., 2005. Danger on the streets: marginality and victimization among homeless people. American Behavioral Scientist 48, 1055–1081.

Mahs, Jurgen von, 2004. The sociospatial exclusion of single homeless people in Berlin and Los Angeles. The American Behavioral Scientist 48 (8), 928–960.

Mayock, P., Sheridan, S., 2012. Women's 'Journeys' to Homelessness: Key Findings from a Biographical Study of Homeless Women in Ireland. Women and Homelessness in Ireland. Research Paper 1. School of Social Work and Social Policy and Children's Research Centre, Trinity College Dublin, Dublin.

Markowitz, F.E., 2006. Psychiatric hospital capacity, homelessness, and crime and arrest rates. Criminology 44 (1), 45–72.

McGuire, J., Rosenheck, R.A., 2004. Criminal history as a prognostic indicator in the treatment of homeless people with severe mental illness. Psychiatric Services 55, 42–48.

McIntyre, J.K., Widom, C.S., 2011. Childhood victimization and crime victimization. Journal of Interpersonal Violence 26 (4), 640–663.

Melander, L.A., Tyler, K.A., 2010. The effect of early maltreatment, victimization, and partner violence on HIV risk behavior among homeless young adults. Journal of Adolescent Health 47, 575–581.

Merscham, C., Van Leeuwen, J.M., McGuire, M., 2009. Mental health and substance abuse indicators among homeless youth in Denver, Colorado. Child Welfare 88, 93–110.

National Coalition for the Homeless, 2012. Hate Crimes against the Homeless: The Brutality of Violence Unveiled. Available at: http://www.nationalhomeless.org/publications/hatecrimes/hatecrimes2011.pdf.

National Law Center for Homelessness and Poverty (NLCHP), 2009. Homes Not Handcuffs: The Criminalization of Homelessness in U.S. Cities. A Report by The National Law Center on Homelessness & Poverty and The National Coalition for the Homeless. Washington, DC.

National Law Center for Homelessness and Poverty (NLCHP), 2011. Criminalizing Crisis: The Criminalization of Homelessness in U.S. Cities. A Report by the National Law Center on Homelessness and Poverty. Washington, DC.

Newburn, T., Jones, T., 2007b. A very special relationship? Criminal Justice Matters 67 (1), 12–47.

O'Toole, T.P., Conde-Martel, A., Gibbon, J.L., Hanusa, B.H., Fine, M.J., 2003. Health care of homeless veterans: why are some individuals falling through the safety net? Journal of General Internal Medicine 18, 929–933.

O'Toole, T.P., Gibbon, J.L., Hanusa, B.H., Freyder, P.J., Conde, A.M., Fine, M.J., 2004. Self-reported changes in drug and alcohol use after becoming homeless. American Journal of Public Health 94 (5), 830–835.

Paradise, M., Cauce, A.M., 2002. Home street home: the interpersonal dimensions of adolescent homelessness. Analyses of Social Issues Public Policy 2, 223–238.

Raphael, Steven, Stoll, Michael A., 2013. Accessing the contribution of the deinstitutionalization of the mentally ill to growth in the US incarceration rate. Journal of Legal Studies 41 (1), 187–222.

Rasnow, T.L., 2011. Traveling justice: providing court based pro se assistance to limited access communities. Fordham Urban Law Journal 2.

Scott, M.S., 2002. Panhandling. Community Policing Services, Washington, D.C.

Snow, D., Susan, G. Baker, Anderson, Leon, 1989. Criminality and homeless men: an empirical assessment. Social Problems 36 (5), 532–549.

Stoltz, J.M., Shannon, K., Kerr, T., Zhang, R., Montaner, J.S., Wood, E., 2007. Associations between childhood maltreatment and sex work in a cohort of drug using youth. Social Science & Medicine 65, 1214–1221.

Sylvestre, M.E., 2010b. Policing the homeless in Montreal: is this really what the population wants? Policing and Society 20 (4), 432–458.

Thrane, L., Chen, X., Johnson, K., Whitbeck, L., 2008. Predictors of post-runaway contact with police among homeless adolescents. Youth Violence and Juvenile Justice 6 (3), 227–239.

Thrane, L.E., Hoyt, D.R., Whitbeck, L.B., Yoder, K.A., 2006. Impact of family abuse on running away, deviance, and street victimization among homeless rural and urban youth. Child Abuse & Neglect 30, 1117–1128.

Torrey, E. Fuller, Kennard, Aaron D., Eslinger, Don, Lamb, Richard, Pavle, James, 2010. More Mentally Ill Persons Are in Jails and Prisons than Hospitals: A Survey of the States. Treatment Advocacy Center, Arlington, VA.

Tyler, K.A., 2008. A comparison of risk factors for sexual victimization among gay, lesbian, bisexual, and heterosexual homeless young adults. Violence & Victims 23 (5), 586–602.

Tyler, K.A., Hoyt, D.R., Whitbeck, L.B., Cauce, A.M., 2001. The impact of childhood sexual abuse on later sexual victimization among runaway youth. Journal of Research on Adolescence 11, 151–176.

Tyler, K.A., Melander, L., 2013. Child abuse, street victimization, and substance use among homeless young adults. Youth & Society. http://dx.doi.org/10.1177/0044118X12471354.

Wachholz, S., 2009. Pathways through hate: exploring the victimisation of the homeless. In: Perry, B. (Ed.), Hate Crimes, The Victims of Hate Crime, vol. 3. Praeger, Westport, CT, pp. 199–222.

Wardhaugh, J., 2012. Beyond the workhouse: regulating vagrancy in Goa, India. Asian Criminology 7 (3), 205–223.

Whitbeck, L.B., Hoyt, D.R., Yoder, K.A., Cauce, A.M., Paradise, M., 2001. Deviant behavior and victimization among homeless and runaway adolescents. Journal of Interpersonal Violence 16, 1175–1204.

Whitbeck, L., Hoyt, D., 1999. Nowhere to Grow: Homeless and Runaway Adolescents and Their Families. Aldine de Gruyter, New York.

Whitbeck, L., Simons, R., 1990. Life on the streets: the victimization of runaway and homeless adolescents. Youth and Society 22, 108–125.

Wright, J., Donley, A., 2011. The Poor and Homeless in the Sunshine State: Down and Out in Theme Park Nation. Transaction Publishers, New York.

Violence, History of

Heinz-Gerhard Haupt, University of Bielefeld, Bielefeld, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

History of violence is not a field of its own in historiography, but it participates in broader debates. These debates are dealing with long-term development and questions whether violent means in public and private life are regressing or expanding. They are part of the history of social movements, criminality, terrorism, genocides, and ethnic cleansing. Political, social, and cultural history approaches are mobilized in these studies.

Violence in its various forms has been an aspect of historical writing since Thucydides, but to this day there is no independent specialty in historical research on violence. The history of violence is thus written as part of military and war history, of research on revolution and protest, and of the history of gender and the body. In Europe and elsewhere, it cuts across the institutional and disciplinary logic of the study of history. This lack of independence has two consequences for historical research on violence. First, it is always interdisciplinary and includes approaches proper to political science and social history, as well as issues of cultural history. Second, it cannot look back on a unified development, but on disparate traditions in the individual disciplines. While revolutionary violence, for example, is discussed primarily in the context of the French Revolution, research on social protest has developed under the influence of 1950s Anglo-American social history and of the reactions to the 1960s urban riots in the US. The historiography of research on violence is thus many stranded and international.

Definitions of Violence

The definitions of violence are as varied as its manifestations. The term itself takes on disparate coloring in the context of each respective national language. English, French, and Italian reflect the Latin and emphasize the injuries caused by violence and its violation of corporal integrity. By comparison, the German term *Gewalt* is less negatively charged, also being used as a synonym for 'power' or *pouvoir*. Only its extension *Gewalttatigkeit* captures the injurious character of the Anglo-American term. In English, the term *force* is designing legitime violence. In German, the legitimation of violence means must use the generic term *Gewalt*. A comparative study of the historical development and significance of terms is a desideratum for research. The spectrum of scholarly definitions of violence is broad. The distinctions between direct and indirect, collective and individual, legitimate and illegitimate, concrete and structural, physical and psychological, and manifest and symbolic violence reflect varying accents in the discussion. Depending on the definition, the focus can be on the manifestations, the justifications, or the effects of violence. The term has been specified further in accordance with its intended use for qualitative or quantitative studies. Modern research on protest for instance sees collective violence when at least 20 persons take part in acts in which persons or objects are damaged.

In the current discussion among historians, violence is widely understood as injury to people's physical integrity, caused by various historical actors in various contexts. Here, violence is not seen as an anthropological constant, nor as a universal historical trait held in common, but tied to the actions of specific groups and conditions that are subject to change in various national societies and epochs. The German sociologist Heinrich Popitz has advanced this view (Popitz, 1986). He sees violence as "an act of power that leads to the intentional bodily injury of others." In this view, corporality distinguishes violence from other means of domination, such as orders, though it contributes to their effect. Since violence injures bodily integrity, it possesses a massive potential for threat and evokes fears. These can contribute to the avoidance of acts of violence, but can also amplify them. The pair of terms 'physical' and 'psychological' violence touch on this connection. Not only does the experience of physical violence, such as torture or rape, have psychological consequences; psychological violence, such as brainwashing, can also have physical effects. Popitz' definition also points to the concept of power. It thus designates the content of violent conflicts as an often unequal structural situation of victims and perpetrators. This underscores that the history of violence is also always a part of societal processes in which the distribution and resources of power are up for discussion.

Beyond its physical core, the term violence has taken on various meanings in the course of the modern age.

1. In the process of state building, legitimate violence is distinguished from illegitimate private violence, independent of whether the latter is wielded in revolutions or in defense of the status quo. The state has defended and extended its monopoly on violence as a constituting component of being a state since early modern times. This monopoly prevailed through the disarmament of the feudal rulers. With the army and (since the eighteenth and especially the nineteenth centuries) the police, state organs exercised the monopoly on legitimate violence. This monopoly required justification in its form and degree, but not in its existence. Lindenberger and Ludtke saw the hallmark of modernity in the network of relationships between physically suffered violence, on the one hand, and the state monopoly on violence, on the other (Lindenberger and

Ludtke, 1995). Thus they not only claim an increase in the state's practice of control and repression, but also touch upon the everyday experience of violence. In the modern age, citizens can come into conflict with state regulations and sanctions which in turn provoke violent actions. In modern protest research, Charles Tilly, among others, has held state reactions to unrest responsible for the extent of existing violence (Tilly, 1975).

2. The experiences of dependency and inequality between the centers of power, colonies, and the so-called 'Third World' led the Norwegian social scientist Johan Galtung to develop the concept of 'structural violence' (Galtung, 1975). From the perspective of peace politics, he drew attention to a discrepancy. He saw a chasm between the existing potential for development based on technology and available societal wealth, on the one hand, and the respective levels of development of disadvantaged groups and countries, on the other. His concept emphasizes not individual perpetrators, but social structures, and not primarily physically experienced violence, but poverty and exploitation. He was not so much concerned with legitimating counterviolence, as vehemently defended by the Algerian sociologist Frantz Fanon, but with social peace, which he thought would be secured for all through equality and opportunities for development (Fanon, 1961). The term 'structural violence' flourished in the 1970s, but also met sharp criticism. It was accused not only of failing to define 'structure,' but also for lacking specificity in its analysis of violence. "By tendentiously deciphering all relations as relations of violence, these relations are leveled," wrote Wolf-Dieter Narr (1973); i.e., the analyses of violence and society are indistinguishable. Galtung was able to demonstrate that the term 'violence' can be applied meaningfully in the analysis of acts in the context of the forcible relations effective in acts of exchange, in inculcating literacy, and in migrations.
3. The concept of symbolic violence is beset with similar problems of distinction. Pierre Bourdieu and Claude Passeron used it in their theoretical analysis of the French school system and thus designated "that form of violence" that "is exercised against a social actor with the complicity of this 'actor'" (Bourdieu and Passeron, 1970). The experience is violent in that the actors are subjected to a code of language and behavior alien to them; it is symbolic through the verbal and semiotic form of its mechanisms. These shape the school's demands and enable or prevent scholastic careers. According to Bourdieu, by accepting this code, the actors contribute to the lasting legitimation of the scholastic system of selection. This term rightly touches on the linguistic and semiotic dimension of violence, but expands it to the point that it no longer has clear contours.
4. In gender history during the 1990s, the concept of violence has been extended to private forms of violence within the family. Feminist literature has used the term 'sexual violence.' This refers to actions that appear in a form tying societies to sexuality. They are carried out against the will of persons who do not have the same power resources as the perpetrators. Sexual violence is thus analyzed as part of the inequality effective in the relationship between the sexes. It shifts from an exceptional phenomenon to one in terms of which the paradigmatically prevailing ideas of power relations and sexuality, and the experiences of women and girls can be grasped. This extension of the concept of violence not only opens new fields to scientific research, but, since the 1980s, has also been taken up in the penal codes of individual societies.

Historical Research on Violence

In the past, experience with violence was subsumed again and again in theoretical sketches. Among them, the theories of Thomas Hobbes, Max Webber, and Hannah Arendt were doubtless the most influential, though they arose under disparate conditions (Hobbes, 1914; Weber, 1919; Arendt, 1970). But, especially since the 1950s, their influence on historical research on violence has been superseded by that of Norbert Elias, Michel Foucault, and most recently Clifford Geertz (Elias, 1976; Foucault, 1975; Geertz, 1973).

Thomas Hobbes was the primary inspirer of historical research on violence (Hobbes, 1914). In seventeenth-century English society, which was no longer integrated by religion, he demonstrated in terms of contract theory the necessity of the state's use of violence. Hobbes postulated that, to pacify an English society composed of private persons hostile to each other, these persons relinquished their rights to the Leviathan state, i.e., in the English case, to the absolute monarch. In return, they received protection from hostile others and the guarantee of societal peace. The double function of the state's monopoly on violence – on the one hand, to protect the interests of individuals, and on the other, to limit their possibilities of self-realization and their rights – has become the common possession of most analyses of violence since Hobbes. The individual's relationship to state violence, thus often develops from his specific experiences either with the state's protection or with its limitation of his freedom. Norbert Elias has provided important stimuli for historical research on violence (Elias, 1976). Elias described European history since the Middle Ages as a history of civilization. This hypothesis is based on two central assumptions. First, Elias says that, in the course of the expansion of the trades, commerce, transportation, and the activity of the state, the interaction among members of society became more frequent and dense. To the degree that they competed for influence and shares of power, agreements on behavior and behavioral norms were required. In this process, civilized manners prevailed. On the other hand, the state monopoly exercised an increased external compulsion over the behavior of the members of a society. This found its most complete form in courtly society. In the realm regulated by courtly etiquette, the individual's success depended on new civilized modes of behavior. To the degree that peoples' definition of themselves conformed to this framework, forms of self-compulsion and control of feelings resulted. Violent structures of interpersonal societal relations receded as the model of civilized and inner-directed acting spread from the upper strata to the rest of society. Elias' historical periodization has been criticized by Hans-Peter Duerr, but his secular-based hypothesis has nonetheless been very fruitful for research (Duerr, 1988–96).

Michel Foucault emphasized the shift in form and content in state practices of punishment, which shaped both societal reality and the individual's life practice (Foucault, 1975). In his view, the practices of domination targeting the body have shifted since 1800 to strategies targeting the mind. The disciplinary measures carried out on the human body in clinics, psychiatric facilities, and prisons took on a more total character in the modern age. As in Jeremy Bentham's 'Panoptikum,' the individual is now subjected to control by institutions and the effects of techniques of power. In Foucault's model, the direct violence of the exercise of power made way for the violence of discourses – at least after the threshold between the epochs, which he locates around 1800. Historical studies have relativized this hypothesis and made it more specific, as they did with Elias' theories, but it too has retained its importance as a challenge to detailed studies.

While Elias and Foucault locate the changing use of bodily violence in epochal contexts and societal macrosituations, cultural anthropological studies have gained importance over the years, and they are no longer devoted to the causes, but to the forms of violence. Attention to the 'how' of the practice of violence, a maximum of 'dense description' of constellations of violence, and even an 'ethics of precision' (von Trotha, 1997) is demanded for the description and explanation of acts of violence and their results. In referring to Clifford Geertz, recourse is taken to anthropological approaches or variants of symbolic interactionism are tested (Geertz, 1973).

Emphases and New Fields in Historical Research on Violence

An International Comparison of Violence

Historical research on violence has so far privileged the national context and has seldom conducted international comparisons. This concentration on the nation-state suggested itself, since the state monopoly on violence stood in the center of research on violence. But there are still some comparative works, based partly on national research or inspired by the latter's questions. Among these is the long-term comparison of the period from the thirteenth to the nineteenth centuries, in which P. Spierenberg describes behavior toward victims, the weak, and animals to demonstrate a general increase of empathy in Europe (Spierenburg, 1991). Charles Tilly formulated a typology of collective violence in Europe, in which he names the parochial conflicts – the feuds between geographical communities (Tilly, 1975). He uses the term 'reactive violence' to designate resistance against state interference in the citizens' way of life. According to Tilly, proactive violence is exercised within those organizations that acquire, and seek to defend, their political power.

All these investigations face the problem of having to work with unequal and incomplete national data. Since forms of protest and criminality take different cultural shapes in different societies, views of reality are often grasped as reality itself. In comparative research on revolution, violence has often not been accorded central importance. The only exception is Arno Mayer's study that contrasts the violence of the French and the Russian Revolutions (Mayer, 2000). Finally, research on totalitarianism has pinpointed state terror as a central instrument of totalitarian systems. Dietrich Beyrau has presented a critical, detailed analysis contrasting and comparing violent state measures (Beyrau, 2000). It elaborates the similarities as well as the differences between the violence of the National Socialist and that of the Stalinist systems of rule. These comparisons had been continued by Michael Geyer and Timothy Snyder. Comparative studies permit discussion of the charged question of modern research on violence: whether private and public violence dwindles in or is compatible with the functioning of modern society.

Collective Violence as Social Protest

Research on protest is one of the most flourishing branches of research on violence. The extent and significance of acts of violence are examined in it. It has taken on new contours since the 1950s, under the influence of English, often Marxist-inspired social history and of the 1960s American research on urban riots. Insight into the inherent logic of popular action has supplanted the condemnation of rage-blinded, aggressive mob, and crowd activity. Their violent nature has been interpreted less as a defining characteristic than as an aspect of social protests and less as irrational than as a rational strategy. Georges Rude and Edward P. Thompson in particular have carried out this revision (Rude, 1964; Thompson, 1971). Thompson has also drawn attention to the 'moral economy.' With this term, he designates the model of violence as a fitting behavior by peasants, artisans, and workers in response to the markets in times of economic crisis or famine. Acts of popular violence in Great Britain during the transition from the eighteenth to the nineteenth centuries are set in relation to this model and contrasted as a traditional alternative to the developing market society. Thompson's continuing hypotheses that these protests were directed against the gentry have meanwhile been revised. Nevertheless, building on Thompson, attention has shifted to the significance of violent protests in a strategy based primarily on negotiating and compromising with the authorities. In this context, John Bohstedt speaks of a 'protocol of riot' that the protesters strictly adhered to, in order to initiate negotiations, rather than destructive activities (Bohstedt, 1988). When violence was nonetheless employed, its form was often borrowed from traditional rituals or it was applied symbolically against objects. Between 1770 and 1850 in Great Britain, France, and Germany, the majority of protests against hunger, tax revolts, and resistance against conscription were usually conducted nonviolently. Following the American urban unrest, Graham and Gurr later questioned this hypothesis of violence's positive significance for negotiating processes (Graham and Gurr, 1969). They underscored that violence is, as a rule, unproductive, while in Charles Tilly's interpretation of violent means as part of power struggles, they can indeed be successfully employed (Tilly, 1975). New research is now asking whether traditional scripts of food riots are taken up by protest against high prices and food shortage during and after the two world wars.

Since the mid-1980s, this still influential view of protest has been accused of attributing too much rationality to it and of denuding it of its emotionality. Alain Corbin in particular has pointed out the degree to which fears, hopes, and rumors have been responsible for acts of violence (Corbin, 1990). George

Lefebvre already traced plundering and acts of violence against the landed nobility in France in 1789 to the effects of alarming news and fears in the rural population (Lefebvre, 1970). For his part, Corbin describes how, on a market day in August 1870, a crowd of 300–800 people tormented a young noble and then burned him alive. He had been accused of being a Prussian and of calling out 'Vive la Republique.' These acts, like the Grande Peur in the French Revolution, can be traced to the spread of rumors, but that does not explain their cruelty. For historical research on protest, this finding suggests that violent activity should be understood as part of a history of the emotions. A growing actual interest of historians in emotions, their origin and effect in different surroundings, will surely develop this field (Bourke).

The investigation of the history of violence has not been limited to the extent and significance of acts of violence but has also asked about its instigators and causes. The mass psychologist Gustav Le Bon has attributed mass violence primarily to women (Le Bon, 1908). Historical research on protest has shown that women have indeed taken part in acts of violence. But women remained in the minority and were more intensely active when the well-being of the family was at stake for instance in food riots than, for example, in conflicts with the military. But they were also victims of police repression when they attacked the property or public rules. In order to determine the actors of violence, the social profile of people arrested for manifestations of violence has been worked out. They usually do not come from the 'Lumpenproletariat,' but are rural and urban wage earners, sometimes from tradesmen milieus. The people, more than the mob, carried out these actions. Bohstedt, in particular, has pointed to disturbed societal relations as one of the causes of violence in Britain, especially where social contact and dependencies between the upper classes and the masses were attenuated (Bohstedt, 1988). In his study of extreme international violence, Christian Gerlach is also stressing the importance of crisis. But it remains open whether this hypothesis can be confirmed for other protests. Whether acts of violence always followed social disintegration, and whether rituals of violence do not in turn presuppose relatively unified communities is a question for additional research. In the past, historical research on protest has concerned itself most with the phase in which civic society was formed. But now it is also examining the period between the world wars as well as the second half of the twentieth century and is including the problem of political violence to a greater degree. Its approaches and methods must now prove themselves in this field.

Historical Research on Violence as Crime

Violent crime in its long-term historical development, societal distribution, and meaning has been investigated primarily in terms of Europe in the Middle Ages and early Modern Age, and in terms of the twentieth-century US. Modern historical studies of crime that also address nineteenth- and twentieth-century Europe are rare. Quantitative international comparative studies of violent crime often exhibit methodological deficiencies, because the database in the individual countries and through time is uneven, and above all because the prevalence of recorded acts of violence determined also depends on the strength and severity of prosecuting authorities. Additionally, Eastern and Central Europe have been almost completely omitted from almost all long-term studies due to the lack of research there or to language barriers. Nevertheless, trend analyses from a bird's-eye perspective can provide hypotheses that can then be worked out in detail or modified by historical microstudies. Spierenberg hypothesized a drop in the rate of homicide in Europe before 1850 (Spierenburg, 1991); in contrast, studies of Western Europe and the US from 1800 to the beginning of the twentieth century by Gurr and Lane find a rising rate of crime and homicide in the US (Gurr et al., 1977; Lane, 1997). These rates allegedly fell in the century from 1840 to 1940, only to rise again in the middle or end of the 1950s, but individual studies of crime in discrete nation-states or cities do not confirm this chronology. American research has confirmed the hypothesis that the US was, and remains, the most violent nation among the Western democracies. But it is still debated whether this can be traced primarily to 'frontier violence' (Hollon, 1974) and its consequences. In Europe, during the last 50 years homicide rates are rising again as a consequence – as Manuel Eisner is suggesting – of economic polarization and cultural individualism.

Modern research on crime focuses on the relationship between delinquents, the penal system, and prosecution. Social history concerns itself with the social origin and situation of criminals; legal history and sociology investigate penal 'labeling,' which criminalizes specific behaviors; and institutional and political history is devoted to the construction of prosecuting bodies. Nicole Castan and Gerd Schwerhoff, among others, have conducted historical studies of cities or regions, examining premodern societies. The following general finding resulted. Until the early modern age, even homicide was an infraction that was not punished automatically by death, but it could be atoned for and expiated through rituals and payments of money. In sixteenth-century Germany, laws called for more severe punishments, but court practice was orientated more toward requiring expiatory measures benefiting the surviving relatives than toward retaliation. This assessment of violence was not surprising in a time that still accepted it as a definite part of societal practices and rituals in the premodern period. Among youth, villagers, and journeymen and masters in trades, not only were conflicts resolved by violence, but initiation rites were also organized around violence. The criminalization and limitation of violence is a process that began in the early modern age but did not fully prevail until the full-blown modern age. For France, and based on the discourse of elites, Muchembled reports on rapid advance of the process of civilization, locating the criminalization of homicide in the eighteenth century (Muchembled, 1988). But even for the eighteenth century, doubts have been raised whether the draconian penalties prescribed can be regarded as an indicator that more severe norms prevailed, or if these penalties were intended to compensate for the weakness of prosecuting authorities.

For it has been pointed out that, precisely in premodern times, legal jurisdiction and police control had not thoroughly penetrated into rural communities in a number of European societies. For the nineteenth century, Peter Gay has advocated the hypothesis that, despite the criminalization of violence, it was indeed tolerated in specific societal areas, to the degree that it remained limited to and ritualized in them (Gay, 1993). He

adduces the examples of the duel and students' fencing to gain scars, corporal punishment in schools and the army, and violence within marriage. The question of when which practices were prosecuted judicially and considered illegitimate violence has not yet been investigated adequately in international comparison.

Recent Approaches of Historical Research on Violence

Historical research on violence has taken on new dimensions in the newer style of the history of war and the military. The latter is concerned with military perpetrators of violence that exceeds normal military action: participation in mass executions, acts of annihilation, torture, or rape. Browning and Bartov have been exemplary in pursuing this approach in studying National Socialist police troops and the *Wehrmacht* on the Eastern front, respectively (Browning, 1992; Bartov, 1985). Concentrating on military perpetrators opens up the analysis of wars that accompany decolonization, the Vietnam War, and warlike conflicts accompanying the collapse of the bipolar world order. Beyond the discovery of formerly unknown or repressed atrocities committed by soldiers and officers, attention has focused on analyzing the causes of such inhumane behavior. Current discussion addresses the question of whether specific images of community and or group cohesion motivate the perpetrators. Beyond soldiers' and officers' motivation, interaction, and acts, research is also concerned with how they deal with experiences with violence. To the degree that interest in the history of memory and experience increased, the effects of violence on postwar societies has also been researched. World War I has been more thoroughly considered than World War II. Even today, it is becoming clear that societal discourse on wars and on the legitimacy of the violence exercised in them has as large a role as direct experience of violence in dealing with the past. It turns out that there is a need to correct the one-dimensional assumptions that experiences on the front are transformed directly into a brutalization of postwar behavior.

The construction of the state monopoly of legitime physical violence – as Max Weber defined it – has been part of studies on state building. But this interest has overstressed the implementation of this monopoly which was equal in time and space. State monopoly for instance was more hegemonic in France than in Italy, where the Mafia organizations used frequently violent means to defend their interests. In several countries, zones of limited statehood continue to exist in which warlords used military violence in order to assure their dominance. This approach has been used in order to explain the barbarous violence in African societies during the last 5 years as well as the atrocities of civil wars. But nation-states themselves are actors of uncivilized and uncontrolled violence either in the context of colonization and decolonization. Or by acts of 'ethnic cleansing', which is considered by Marc Mazower to be more widespread than genocides during the twentieth century. This approach is situating – as the recent studies on genocide are doing – violence in a broader historical context.

Historical studies on terrorism were more frequent after 9/11 than before, but even then the scientific outcome is rather limited. The wave of anarchist violence at the end of the nineteenth century and its ideological protagonists (Kropotkin, Bakunin, etc.) attracted a lot of attention as well as the terrorist organizations, their actors, and the state reaction to them in Western Germany, Italy, and France. Those studies are focusing often on national cases and comparative views are the exception. Terrorism is often situated in the relationship of state and terrorist groups and explained by an aggression–reaction schema. It was also perceived as a communicative act by which terrorist groups tried to provoke state power by their demands. Those studies which shared this interpretation insist on the media coverage and stress performative acts.

Research on violence has profited in the late twentieth century from interest in questions of cultural history. In this context, the taboo against violence is no longer seen as the dominant characteristic of the modern age, whose ambivalence Baumann has underscored (Baumann, 1991). The media's conveying and staging, which depicts and overemphasizes acts of violence, have contributed to this ambivalence of the modern age. These connections are often mentioned in current diagnoses, but there is a lack of comparative historical research devoted to the respective cultures of violence in various milieus and at various times. For most epochs, not even the discursive boundaries between violence and nonviolence are well enough known. Special attention should be paid to the degree to which state authorities and the upper strata of societies have intervened in the justification of forms of violence and in its condemnation (Haupt). For the respective concepts of violence, all have a mediating link between the state monopoly on violence and everyday experience.

See also: Arendt, Hannah (1906–75); Elias, Norbert (1897–1990); Foucault, Michel (1926–84); Rape and Sexual Coercion; Social Movements, History of: General; Terrorism; Torture and its Consequences, Psychology of; Violence: Public; War, Sociology of; Warfare in History; Weber, Max (1864–1920).

Bibliography

Arendt, H., 1970. On Violence. Harcourt Brace, New York.
Bartov, O., 1985. The Eastern Front, 1941–45. German Troops and the Barbarisation of Warfare. Macmillan, London.
Baumann, Z., 1991. Modernity and Ambivalence. Cornell University Press, Ithaca, NY.
Bentham, J., 1787–1962. Panopticon or the inspection-house. In: Bowring, J. (Ed.), The Works of Jeremy Bentham. Russell & Russell Inc., New York, pp. 37–172.
Beyrau, D., 2000. Schlachtfeld der Diktaturen Osteuropa im Schatten _on Hitler und Stalin. Vandenhoeck & Ruprecht, Gottingen, Germany.
Bloxham, D., Moses, A.D. (Eds.), 2010. The Oxford Handbook of Genocide Studies. Oxford.
Bohstedt, J., 1983. Riots and Community Politics in England and Wales, 1790–1810. Harvard University Press, Cambridge, MA.
Bohstedt, J., 1988. Gender, household and community politics: women in English riots 1790–1810. Past and Present 120, 88–122.
Bourdieu, P., Passeron, J.P., 1970. La Reproduction. Paris.
Brockhaus, U., Kolshorn, M., 1993. Sexuelle Gewalt gegen Madchen und Jungen, Mythen, Fakten, Theorien. Campus, Frankfurt, Germany.
Browning, C., 1992. Ordinary Men. Reserve Police Battalion 101 and the Final Solution in Poland. Harper Collins, New York.
Castan, N., 1980. Justice et repression en Languedoc a l'epoque des lumieres. Flammarion, Paris.
Corbin, A., 1990. Le village des cannibales. Flammarion, Paris.
Duerr, H.-P., 1988–96. Der Mythos _vom Zivilisationsprozeß, 4 vols. Frankfurt, Germany.

Elias, N., 1976. U ber den Prozeß der Zivilisation, Soziogenetische und psychogenetische Untersuchungen, 2 vols. Suhrkamp, Frankfurt, Germany.
Emsley, C., Knafla, L.A. (Eds.), 1996. Crime History and Histories of Crime. Studies in the Historiography of Crime and Criminal Justice in Modern History. Greenwood Press, Westport, CT.
Eisner, M., 2002. Langfristige Gewaltentwicklung; Empirische Befunde und theoretische Erklärungsansätze. In: Heitmyer, W., Hagan, J. (Eds.), Internationales Handbuch der Gewaltforschung, Opladen, pp. 58–80.
Fanon, F., 1961. Les Damnes de la Terre. Maspero, Paris.
Foucault, M., 1975. Surveiller et punir. Naissance de la Prison. Paris.
Fitzpatrick, S., Geyer, M. (Eds.), 2009. Beyond Totalitarianism. Stalinism and Nazism compared, Cambridge.
Gailus, M., 1990. Straße und Brot. Sozialer Protest in den deutschen Staaten unter besonderer Beru_ cksichtigung Preußens, 1847–1849. Vandenhoeck & Ruprecht, Gottingen, Germany.
Galtung, J., 1975. Strukturelle Gewalt. Beitrage zur Friedens- und Konfliktforschung. Rowohlt, Reinbek, Germany.
Gay, P., 1993. The Cultivation of Hatred. New York.
Geertz, C., 1973. The Interpretation of Culture. Basic Books, New York.
Gerlach, C., 2011. Extrem gewalttätige Gesellschaften. Massengewalt im 20.Jahrhundert. München.
Graham, H.D., Gurr, T.R. (Eds.), 1969. The History of Violence in America. Praeger, New York.
Gurr, T.R., et al., 1977. The Politics of Crime and Conflict. Sage, Beverly Hills, CA.
Haupt, H.G., Weinhauer, K., 2011. Terrorism and state. In: Bloxham, D., Gerwarth, R. (Eds.), Political violence in twentieth century Europe, pp. 176–2009. Cambridge.
Haupt, H.G., 2012. Gewalt und Politik im Europa des 19.und 20.Jahrhunderts. Göttingen.
Hobbes, T., 1914. Leviathan. J.M. Dent & Sons, London.
Hollon, W.E., 1974. Frontier Violence. Another Look. Oxford University Press, New York.
Kocka, J., Jesson, R., 1990. Die abnehmende Gewaltsamkeit sozialer Proteste vom 18. zum 20. Jahrhundert. In: Albrecht, P.A., Backes, O. (Eds.), Verdeckte Gewalt. Suhrkamp, Frankfurt, Germany, pp. 33–57.
Lane, R., 1997. Murder in America; a History. Ohio State University Press, Columbus, OH.
Le Bon, G., 1908. Psychologie des Foules, thirteenth ed. Alcan, Paris.
Lefebvre, G., 1970. La Grande peur de 1789. Colin, Paris.
Lindenberger, T., Ludtke, A., 1995. Physische Gewalt—eine Kontinuita _ t der Moderne. In: Lindenberger, T., Ludtke, A. (Eds.), Physische Gewalt. Suhrkamp, Frankfurt, Germany.
Mayer, A.J., 2000. The Furies. Violence and Terror in the French and Russian Revolution. Princeton University Press, Princeton, NJ.
Mazower, M., 2002. Violence and the state in the twentieth century. American Historical Review 1147–1167.
Muchembled, R., 1988. L'invention de l'homme moderne. Culture et sensibilites en France du Xe au XVIIIe siecle. Fayard, Paris.
Narr, W.-D., 1973. Gewalt und Legitimitat. Leviathan 1, 7–41.
Popitz, H., 1986. Phanomene der Macht. Autoritat—Herrschaft— Gewalt—Technik. Mohr, Tu bingen, Germany.
Rude, G., 1964. The Crowd in History, France and England. Wiley, New York.
Schwerhoff, G., 1991. Ko In im Kreuzverhor. Kriminaiitat, Herrschaft und Gesellschaft in einer fruhneuzeitlichen Stadt. Bouvier, Bonn, Germany.
Spierenburg, P., 1991. The Broken Spell. Macmillan, London.
Ther, P., 2011. Die dunkle Seite der Nationalstaaten "Ethnische Säuberungen" im modernen Europa. Göttingen.
Thompson, E.P., 1971. The moral economy of the English crowd in the eighteenth century. Past and Present 50, 76–136.
Tilly, C., 1975. The Rebellious Century. Harvard University Press, Cambridge, MA.
von Trotha, T., 1997. Zur Soziologie der Gewalt. In: von Trotha, T. (Ed.), Soziologie der Gewalt (KZfSS, 37). Kolner Zeitschrift fur Soziologie und Sozialpsychologie. Opladen_Wiesbaden, Germany, pp. 9–56.
Weber, M., 1919. Politik als Beruf. Duncker & Humblot, Munich, Germany.

Violence in Anthropology

Kay B Warren, Anthropology Department, Brown University, Providence, RI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Anthropologists have moved away from static categories and hierarchies of violence, commonly used by other disciplines, and toward the dynamics of continually interacting and transforming formations of violence. This article illustrates contrasting dynamics of violence for the Cold War and post-Cold War periods. It shows how anthropological perspectives complicate assumptions made about transitions from 'war' to 'peace.' Anthropologists offer new perspectives on war and peace, truth and reconciliation commissions, illicit migration and human trafficking, health care for impoverished urban sectors, and humanitarianism in action.

Introduction

The study of violence in the field of anthropology has been reshaped in response to a sea change of global transformations and regional conflicts. Anthropologists have moved away from the study of discrete categories and hierarchies of violence, commonly used in other disciplines, toward the dynamics of continually interacting and transforming formations of violence as they are mediated by institutions and ongoing social relations. In the process, anthropologists no longer see 'warfare' and 'peace' as discrete stages or states of being. The 'violence paradigm' – that distinguishes predator, victim, and audience – is no longer convincing.

The roles of civilians in warfare and the impact of war on civilian life immediately complicate these relations. In Latin America's counterinsurgency wars of the 1980s – when state armies fought revolutionary forces – armies on both sides were composed of civilians with limited training and unsophisticated arms. State authorities forcefully recruited boys and young men from schools, family farms, and their workplaces. Rebels found foot soldiers among leftist volunteers. They were joined by families that fled their communities when army patrols used scorched earth tactics in the countryside. After seeing dead bodies along the road, young community leaders in schools, religious groups, and local committees grew anxious that their names might appear on the local assassination list. Should they stay and keep a low profile where they were or leave and join the other side which, in fact, was not strong enough to guarantee their safety?

Anthropologists illustrate how war lives on through its after-effects and the violent fragmentation of people's lives. In Guatemala City, the gangs that poor joined as youth groups and alternative families in the mid-1980s became increasingly brutal in the 1990s as they affiliated with organized crime. The extreme violence they came to practice had its roots in the hardest moments of the counterinsurgency war.

Dispossession occurred after these wars as well. State-mandated relocation programs sent families away from 'dangerous' borders, leaving them without basic resources. Agriculturalists lost their lands and social networks in the process. Some turned to the anonymity of city life. Others joined the surge of internal refugees. Many fled for their lives to other countries, such as Mexico and the United States. There is no universal story of the struggle to survive these wars. It is clear that the erasure of civilians from scholarly accounts of war needs to be rectified.

Cold War Agendas for Anthropology

Across the Cold War and the collapse and transformation of state socialism (1946–89), research on violence followed great power tensions, clashing ideologies, and violent authoritarian states. Land, labor, class conflict, state's regulation of national markets, and other aspects of single-party governance were key aspects of socialism in practice. Analysts have documented socialist regimes and counterinsurgency armies in many parts of the world. Other formations of violence – including ethnic nationalism, sectarianism, racism, communal violence, genocide, and organized crime – complicated and/or served the interests of these political systems in different parts of the world.

During the Cold War, social scientists on the left in the United States, Europe, and major zones of conflict circulated important critiques of capitalist imperialism and colonialism. They demonstrated how the dynamics of structural violence in these economies reproduced inequality for urban workers and rural peasants. Precarious families faced physical violence, small unstable incomes, life-threatening poverty and debt, disease, and high rates of child mortality. Activist researchers supported anti-imperialist agendas and participated in solidarity movements, speaking out at major forums. Engaged scholars collaborated with their regional counterparts, especially in Latin America, on agrarian histories, studies of repressive regimes, and indigenous issues. The 400-year history of elite racism as a form of coordinated governance through the church, military, and government was another line of path-breaking work. Others were involved in long-term innovative community and regional studies. Histories of rebellions, revolutions, urban violence, and activism that challenged the status quo have been important products of this research.

These scholars and public intellectuals generated critiques of violence and capitalist governance for international audiences. Many stressed the regional dynamism of struggles and agendas for change. The special brutality of counterinsurgency warfare was revealed in the dual goals of crushing revolutionary

movements and preemptively punishing civilians who might at any point become politicized and join rebel forces or oppositional organizations. Counterinsurgency wars in Latin America involved genocidal violence, mass killings, widespread 'disappearances,' and decentralized military bases that detained and tortured suspects who came to their attention in the routine surveillance of nearby towns and cities. Spies were everywhere; public buses were routinely stopped and searched for subversives by armed soldiers. No matter what the age of the bus rider, a regional knit bag from an activist area or a small tape recorder was seen as a signal that someone was politically dangerous and likely to be involved with the rebels.

National organizations on the popular left challenged state violence with independent research and public protest. Activist groups came together in public to personally testify about the human cost of state violence, holding up jarring photos of the 'disappeared' in their families. Critiques of human rights abuse, labor exploitation, and state violence were discussed and circulated by activist research groups and featured at major events. In some instances, this was a hotly contested process. Human rights, particularly labor rights, were denounced by conservative critics as dangerous revolutionary claims rather than as internationally accepted liberal norms. Peace activists were denounced as communists.

In spite of continuing repression, research was done in support of the peace process. The international community supported respected researchers who produced accountings of civilian and armed forces deaths on both sides for official truth and reconciliation committees. Locally, townspeople were able to pursue something that had been too dangerous to undertake during the war. They disinterred corpses of loved ones from secret cemeteries hastily created by the troops after massacres in order to do their own personal accounting and mourning of the dead. Years after the massacres, these were still risky ventures; in some areas local activists were aided by professional forensic anthropologists who used their scientific training to identify remains and establish how individuals died.

This historical period produced new kinds of political media. To draw international attention to and humanize shocking mass violence in faraway places, which had received limited international media coverage during counterinsurgency wars, the Cuban government gave awards for outstanding books based on the testimonies of indigenous Marxists, mine workers, and women activists for their struggles against capitalist exploitation and repression in Latin America. Feminist anthropologists in the 1980s documented labor abuse and exploitation in newly emerging workplaces, most especially at maquiladoras – huge 'offshore' factories in special economic zones with high-volume assembly lines operating under the deafening din of modern machinery. These young Third World women were under constant pressure to keep up with work quotas that wore out their bodies as they assembled clothing for mass markets in the United States.

Postsocialist Transition versus Armed Forces Demobilization

Research agendas changed dramatically after the end of the Cold War. State-managed socialist economies in China, Vietnam, Russia, Eastern Europe, and socialist countries in Latin America and Africa took different paths in their transitions to market economies. Research turned to economic globalization, neoliberalism, the privatization of government functions, and new forms of warfare. The opening of Cold War borders took center stage as did transnational research on governance and conflict within and across world regions.

Post-Cold War formations of violence took new turns as rebels lost momentum and the UN oversaw negotiations between key parties in a peace process that sometimes took years to negotiate. Reforms called for cuts in the size of the army and the demobilization and disarmament of rebels. Well-known rebels and leftist activists turned to party politics and the electoral system. As former soldiers looked for employment, gangs formed that engaged in extortion, murder for hire, robberies, and illicit commerce. Crime rates increased. The principals of progressive rural schools supported by international nonprofit organizations were objects of extortions and shake downs, streets continued to be unsafe at night, and daylight robberies were common. Violence reflected the corruption of authorities, the failure of the criminal justice system, the power of organized crime, and transnational scope of neoliberal markets. Lynchings, that is, spontaneous collective violence against individuals judged by the local community to be criminals, was practiced in the cities and countryside. Some attacks on foreign women tourists were the result of fast-moving local gossip that the tourist's real motive for visiting the countryside was to steal babies or children and take them home to the United States. This is part of a wider anxiety about the adoption market, brokered by elites, between countries like Guatemala and the United States.

Anthropologists and related scholars have written powerful post-Cold War works on globalization, neoliberalism, political violence, security, counterinsurgency warfare transitions, and transnationalism. They have researched different world regions and traced the dramatic growth of informal economies, organized crime, and large-scale illicit businesses, networks that serve as forms of governance in weak states in Africa and Latin America. One can hear the echoes of these abstract issues in the day-to-day life and violence of urban neighborhoods and local communities.

The Anthropology of Care, Public Health, and Humanitarianism in Neoliberal Times

The scope of the anthropological project has grown as researchers questioned the conventional emphasis on violence as the state's monopoly over the use of legitimate force by the armed forces, police, and judicial systems. Cultural anthropologists have chosen not to uniquely privilege formal institutions or physical harm in their studies. Rather they include informally instituted structures of violence generated in situations of civil war, state terrorism, fragmented states, globalized economies, and the marginalization and dispossession of individuals and communities.

In the process, the 'anthropology of care' has taken on seemingly intractable social problems and shed new light on the paradoxes and possibilities of current systems and practices of care for people at the margins. These ethnographic case

studies interrogate neoliberal market expansion into impoverished urban neighborhoods, as well as the extension of state services dealing with addiction into areas outside larger cities. At issue is whether new models of health delivery can gain effective access to the poor and their needs, and what happens when larger systems fail to effectively connect. How do people at the extreme margins take care of others when the wider system fails them?

Anthropologists pursue the many faces and paradoxes of humanitarianism and the relation of humanitarianism to governance. The transnational recruitment of cheap labor, the targeted closing of borders to dangerous or undeserving others, and deportation regimes are some of the compelling issues in this line of research. Making ones' way across borders is a powerful challenge. The technologies of border surveillance pursue access to everyday social life, individuals' backgrounds, and personal intentions (Miriam Ticktin, 2011). Interviews by state authorities generate unbearable anxieties. The interviewer demands answers to key questions. To be convincing, answers need to fall in the domain of the right answers and convey the correct effect. In practice, however, ideal answers continually change as do the officials. Those trying to gain special visas for access to health care on humanitarian grounds face these dilemmas. They had to impress officials in these interrogations to stand a chance for a visa. These processes echo some aspects of the arbitrariness, unfairness, and inefficiencies that Akhil Gupta (2012) addresses in his discussions of bureaucracy, structural violence, and urban poverty in India.

Topical Examples

1. Transitions: war and the ironies of peace
2. Truth and reconciliation processes
3. A crime with many names: human trafficking, modern day slavery, and the transportation of undocumented migrants by organized crime
4. The anthropology of care, public health, borders, and humanitarianism.

Transitions: War and the Ironies of Peace

Diane Nelson generates innovative framings and analytical momentum from post-structuralist and feminist theory in her book "*Reckoning: The Ends of War in Guatemala*" (2013). This work exposes problematic forms of structuralism in social scientific and cultural research framings and media analysis. Her analysis stresses transitions in a very different and more critical analysis than the mainstream's focus on peace, democracy, and business expansion. She responds to current discussions of the shift of socialist to postsocialist governance with new directions for research on transitions. Nelson critically engages the political and moral dilemmas that accompanied the 'capitalist victories' of the 'peace process' with its focus on 'democratic' and 'market-oriented governance.' She argues that 'reckoning' is a very revealing political process that benefits from her critical positioning with its unconventional and innovative mix of critical theory and ethnographic engagement. Her most striking move is to posit virtually universal participation and complicity in the morphing patterns of violence across transitions. Grassroots Marxist rebels leave their weapons and Cold War grammars of conflict to be absorbed by the market economy, illicit economies, and maquiladora production for international markets. The greater contact between Mayans and the United States than ever before starts with rural refugees fleeing genocidal violence who become transnational labor migrants, providing new sources of income for their families, and gaining membership in violent gangs anchored in LA and Guatemala.

A hidden wound appears in her work, one that addresses an unspeakable question for anthropologists and activists during the counterinsurgency war: 'Was the insurgency a mistake or not?' The question is whispered by Nelson and others in fleeting terms.

Her coedited book with 'Carlota McAllister *War by Other Means: Aftermath in Post-Genocide Guatemala*' (2013) is a very successful collaborative engagement with the continuing after effects of 'la violencia,' as it is called in Guatemala. The article in this book explore new formations of local and transnational violence. They pursue the internalization of violence in rural and urban populations, especially in gangs, lynchings, and attacks on foreigners.

Truth and Reconciliation Processes

Anthropologist Kimberly Theidon (2013) has contributed to innovative critical studies on truth and reconciliation commission (TRC) processes. Her perspective cites the power of postwar human rights activism to create technologies that establish the truth of particular genocidal events in order to create paths for reconciliation through TRC processes in different parts of the world. In tension with this view, she pursues the project of documenting community responses to situations of extreme violence in the 1980s, especially the legacies of rebels seeking to overthrow government and counterinsurgency soldiers responding with widespread violence designed to halt radicalization.

This work transcends fragmented accounts that TRC interviewers classically produce with their templates of issues and lists of questions. By contrast, Theidon pursues carefully contextualized ethnographic work on the interpersonal contours of violence, how local communities dealt with and continue to deal with fellow residents who were Maoist Shining Path sympathizers and rebels during and after the war. Theidon offers a gender-inflected account of the impact of peasant soldiers who occupy rural communities and their sexual violence against women. Anti-Shining Path community organizing was also violent, another legacy for widows, single mothers, and their children.

This ethnographic approach generates original findings on masculinity, class, and violence; on women's battles to limit antagonists' violence against their children, at a substantial personal cost. Theidon's ethnography highlights the contrast between the goals of the TRC processes, which offers the psychological therapeutic recovery of memories of violence versus the local Quechua community attempts to live together with their own construction of reconciliation, 'by letting go of hateful memories.' Theidon's commitment to the legitimacy of indigenous representations of reconciliation versus her own preference for the establishment of TRC reparations for women

who were raped during the conflict and militarization of these communities is fascinating.

Illegal Migration: A Crime of Many Names and Competing Forms of Governance

There have been a variety of constructions of illegal migration: 'human trafficking,' 'modern day slavery,' and the 'transportation of illegal migrants across borders.' Anthropology and related fields have examined the implications of these rhetorics for criminalizing migration and concentrating on deportation as a dominant form of state governance (De Genova and Peutz, 2010; Hansen and Stepputat, 2005; Ticktin, 2011; Warren, 2012).

Activists, governments, and scholars use the term 'human trafficking' to identify labor recruiters and criminal organizations that treat job seekers as commodities that can be bought and sold. Criminal organizations look for migrants who can be convinced to take what are often demeaning jobs at a great distance from their homes and social networks. While these jobs may be idealized by recruiters/traffickers, they are often similar to jobs at home but with the promise of better pay. Human trafficking is often assumed to be sex trafficking, that is prostitution. But many other kinds of work are offered to people seeking employment: jobs that involve physical coercion, exploitation, and irregular pay. For instance, labor migrants to the United States from Central America are recruited to harvest fruits, vegetables, and tobacco; work in slaughterhouses; take on graveyard shifts cleaning buildings or restocking shelves; and as live-in domestics. The legal issue that determines if people are victims of the crime of human trafficking is whether individuals have given 'their consent' to work in these jobs or whether they are forcefully coerced by recruiters and kept by supervisors in situations without being paid or are faced with violent threats if they do not meet production quotas. These jobs commonly require labor that is physically dangerous and subject workers to demeaning treatment as objects in labor markets. In some cases, people are in effect bought and sold as they are parceled out to work sites. They are controlled by inflated debts they must repay to cover recruitment fees, transportation to jobs sites, and forged travel documents to cross-national borders. Additional fabricated expenses such as paying off fines for breaking 'rules' invented by their handlers are common practices. Extra fees for housing, food, clothing, and medical care are deducted from workers' pay, with or without a formal accounting. Workers generally pay off these illegal debts to regain freedom of movement, just as they would in their home countries if they worked on Central American plantations or urban businesses. These forms of regulation are exploitative and abusive; they violate recognized but often unenforced human rights. An alternative term for these situations is 'modern day slavery' (Bales, 2008). This term includes many different kinds of work that force people into slave-like conditions in different parts of the world. For instance, these practices include the exploitation of youths who are controlled through debts inherited across generations, which means they are bound to menial work to pay debts that they may not fully understand. Youths work very long shifts and often receive token compensation, beyond food and improvised housing, for jobs in brick making, agriculture, and domestic service. Or they may be forcefully recruited into armies – national and rebel forces – in conflict zones. They become indentured workers who cannot leave their positions. International law criminalizes the recruitment of youths and children under 18 years of age for these and other kinds of work because they are too young to give their own legal consent to work. A third approach focuses on the illegal transportation across borders by organized crime and the work of border governance to criminalize and stigmatize migration and organize deportations. These forms of state regulation often stand in tension with the government's active promotion of migration for its citizens. In Latin America, for example, governments with international development assistance from the United States and the UN have developed nonformal education initiatives to educate citizens in safer work choices for migrants (Warren, 2012). In many countries, much of the national population sees their future in terms of tactical migration. There are too few job opportunities, little continuing education, and chronic violence for the urban poor at home. Many families dream of escaping to better job markets or to joining members of their families already working in the United States. There are ironic forms of governance at play in this situation. Each construction of migration, as 'human trafficking,' 'modern day slavery,' or 'competing forms of governance' has its own complexity. Undoubtedly more emphasis should be put on the role of organized crime in migration.

The 2014 crisis of unaccompanied youths and children crossing illegally into the United States and surrendering to border authorities is the result of an aggressive and cynical scam. Criminal organizations – in a move to diversify their operations beyond serving as brokers for the drug trade which caters to the lucrative demand from the US and European consumers – are selling Central American families made-up stories of easy visas and successful family reunification if they have a parent or other relatives living in the United States. They have extorted thousands of dollars from each impoverished extended family to take their youths and children to the Promised Land. These recruiters failed to disclose the truth that this wave of migrants will most likely not be allowed to stay in the United States without visas. These youths and children risked death and sexual violence along the way. In seeking a better future for their children and an escape from chronic urban violence, corruption, and unemployment in their home communities, their families have assumed extraordinary debts. However lucrative the scam has been for organized crime, it has been a very cruel one for everyone else.

The Anthropology of Care, Public Health, Borders, and Humanitarianism

Cultural and medical anthropologists have chosen to study the daily practice of formal institutions and their capacity for physical harm rather than healing through acts of omission. They also seek to understand the social fragmentation and suffering generated by informally instituted structures of violence (Das, 2007). Governments often picture inner cities as zones of lawlessness; in contrast, researchers discover that marginal neighborhoods are products of urban policies that condemn these neighborhoods as hopeless and dangerous. There are few legal job opportunities for adults and young

people in underclass urban zones. Governments provide few public services, given prioritized decisions that limit funding. As a result the schools are inadequate. Marginalized urban life in much of Latin America is dominated by gang violence, corrupt policing, organized crime, and lucrative illicit economies. It is a nexus of both precarity and care.

The 'anthropology of care' has become a major site of social critique. Anthropologists have successfully used ethnographic methods to pursue case studies of chronic poverty in the United States and other world regions. They are studying drug addition, inadequate health services, criminal gangs, and state violence, particularly but not uniquely in urban settings, as understood in their wider regional and institutional contexts.

With neoliberal economic reforms, more of the burden of care falls to families and kinship networks in the United States. Angela Garcia (2010) worked on state programs in New Mexico that deal with chronic drug addiction and found many barriers to successful treatment built into these programs. The larger program, funded by the state, was broken. Garcia found patterns of love and care between mothers and daughters at home to calm the suffering of withdrawal. In effect, securing drugs for their adult children when the clinic refused to help them has become part of the cycle of care. She wants readers to understand the dynamics of this situation, tied to the state funding and addiction treatment norms and to the corrosive poverty of these communities with few options for work. The reality is that there is no place in these settings to be outside heroin; the drugs are everywhere. In a fascinating argument, she ties this existential dilemma to the longer history of these communities' dispossession from the lands that their families had worked and owned in the historical past.

Clara Han, in long-term research in Chile, learned how the urban poor had to cope with a very violent and fragmented environment – drugs, gangs, violent policing, and economic pressures that require the continual balancing of social obligations with limited financial reserves since so little stable work was available. In order be involved in the care of others, these urbanites have found ways to create their own social networks and forms of care by making use of shared credit cards, something they are able to do as the financial industry has recently looked for ways to expand their client base for credit cards and debt (Han, 2012).

It is in these dilemmas that anthropologists come to understand the deep desire to help others and the limits that corrosive poverty and violence places on the possible. There is inevitably a larger range of constraints – which these researchers identify for each particular situation – that limit people's options and shape their moral dilemmas. The challenge for anthropologists and others is to see the humanity in these situations.

Public health on a much larger scale is another arena for critical anthropology. At issue is the potential and limits of market economies in progressive governments to create new systems of public health delivery. In Brazil, the economics and politics of public health have been the center of a series of innovative policies dealing with international markets, which has widened the access of the poor to public health resources in crises such as HIV/AIDS treatment and allows people to live full lives with HIV/AIDS as a chronic illness. João Biehl (2007) produced a path-breaking multisited study of Brazilian governance and public heath bureaucracies in action. Biehl produced a wide-angle view of the multinational production of drugs in India and Brazil's negotiations with international drug companies to lower the cost of medications, to create the possibility of state health-care delivery that could widely radically increase the poor's access to drugs for HIV/AIDS.

João Biehl and Adriana Petryna have joined with other investigators to produce an international volume of case studies dealing with critical public health studies (2013). These studies examine the biopolitics of health and disease and the neoliberal state's role in regulating bodies and human life in the context of the circulation of new health technologies and the operation of treatment regimes in very different local social and political contexts with striking inequalities of access.

See also: Cold War, The; Global Migration; Global Violence Against Women; Globalization and Human Trafficking; Peace Movements; Re-entry and Reintegration After Incarceration; Refugees: Geographical Aspects; Revolutions, History of; Surveillance Studies; Terrorism; Torture and its Consequences, Psychology of; Transnational Organized Crime; War Crimes Tribunals, Crimes against Humanity and Transitional Justice; Warfare in History.

Bibliography

Bales, K., 2008. Ending Slavery: How We Free Today's Slaves. University of California Press, Berkeley, CA.

Berlant, L., 2011. Cruel Optimism. Duke University Press, Durham, NC and London, UK.

Biehl, J., 2007. Will to Live: AIDS Therapies and the Politics of Survival. Princeton University Press, Princeton, NJ.

Biehl, J., Petryna, A. (Eds.), 2013. When People Come First: Critical Studies in Global Health. Princeton University Press, Princeton, NJ.

Butler, J., Athanasiou, A., 2013. Dispossession: The Performative in the Political. Polity Press, Cambridge, UK and Malden, MA, US.

Das, V., 2007. Life and Words: Violence and the Descent into the Ordinary. University of California, Berkeley.

De Genova, N., Peutz, N. (Eds.), 2010. The Deportation Regime: Sovereignty, Space, and the Freedom of Movement. Duke University Press, Durham, NC and London, UK.

Fassin, D., 2012. Humanitarian Reason: A Moral History of the Present. University of California Press, Berkeley.

Fassin, D., 2010. Contemporary States of Emergency: The Politics of Military and Humanitarian Interventions. Zone Books, New York.

Feldman, I., Ticktin, M. (Eds.), 2010. In the Name of Humanity: The Government of Threat and Care. Duke University Press, Durham, NC and London, UK.

Garcia, A., 2010. The Pastoral Clinic: Addiction and Dispossession along the Rio Grande. University of California Press, Berkeley, CA.

Gupta, A., 2012. Red Tape: Bureaucracy, Structural Violence, and Poverty in India. Duke University Press, Durham, NC.

Han, C., 2012. Life in Debt: Times of Care and Violence in Neoliberal Chile. University of California, Berkeley, CA.

Hansen, T., Stepputat, F. (Eds.), 2005. Sovereign Bodies: Citizens, Migrants, and States in the Postcolonial World. Princeton University Press, Princeton, NJ.

Levenson, D., 2013. What happened to the revolution? Guatemala City's Maras from life to death. In: McAllister, C., Nelson, D.M (Eds.), War by Other Means; Aftermath in Post-Genocide Guatemala. Duke University Press, Durham, NC, pp. 195–217.

McAlister, C., Nelson, D.M. (Eds.), 2013. War by Other Means: Aftermath in Post-Genocide Guatemala. Duke University Press, Durham, NC.

Montejo, V., 1999. Voices from Exile: Violence and Survival in Modern Mayan History. University of Oklahoma Press, Oklahoma City.
Nordstrom, C., 2007. Global Outlaws: Crime, Money, and Power in the Contemporary World. University of California Press, Berkeley.
O'Neill, K.L., Thomas, K. (Eds.), 2011. Securing the City: Neoliberalism, Space, and Insecurity in Postwar Guatemala. Duke University Press, Durham, NC.
Penglase, R.B., 2014. Living with Insecurity in a Brazilian Favela: Urban Violence and Daily Life. Rutgers University Press, New Brunswick, NJ.
Rosas, G., 2012. Barrio Libre: Criminalizing States and Delinquent Refusals of the New Frontier. Duke University Press, Durham, NC.
Rothenberg, D. (Ed.), 2012. Memory of Silence: The Guatemalan Truth Commission Report. Martin's Press, New York.
Stoler, A.L. (Ed.), 2013. Imperial Debris: On Ruins and Ruination. Duke University Press, Durham, NC and London, UK.
Theidon, K., 2013. Intimate Enemies: Violence and Reconciliation in Peru. University of Pennsylvania Press, Philadelphia, PA.
Ticktin, M., 2011. Casualties of Care: Immigration and the Politics of Humanitarianism in France. University of California Press, Berkeley.
Warren, K., 2012. Indigenous Movements and Their Critics: Pan-Mayan Activism in Guatemala. Princeton University Press, Princeton, NJ.
Warren, K., Winter 2012. Troubling the victim/victimizer dichotomy in efforts to combat human trafficking: the unintended consequences of moralizing labor migration. Indiana Journal of Global Legal Studies 19 (1), 105–120.

Violence: Public

Charles Tilly[†], Columbia University, New York, NY, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16206–16211, © 2001, Elsevier Ltd.

Abstract

Public violence is episodic social interaction that immediately inflicts physical damage on persons and/or objects, results at least in part from coordination among persons who perform the damaging acts, and occurs at least partly in publicly accessible spaces. Forms of public violence differ with respect to (1) extent of coordination and (2) salience of damage in the social interaction. Cross-classification of (1) and (2) yields five rough types: organizational outcomes, coordinated destruction, opportunism, fragmented resistance, and violent games. All five types of public violence depend heavily on adjacent nonviolent political processes.

Visibly violent social processes simultaneously attract and repel analysts of human behavior. War, rape, genocide, assassination, assault, football, looting, official execution, collective suicide, mutual flagellation, automobile collisions, airline crashes, gang fights, and struggles pitting police against demonstrators all qualify by some definitions as public violence. Most of them regularly attract headlines when they occur – the more extensive, public and grisly their consequences the bigger the headlines. Violence makes the news.

In these circumstances, reporters and readers commonly turn to very simple explanations: some humans are violent in nature, others peaceable, so violent events occur when violent people congregate and enjoy free rein. Unfortunately for that sort of explanation, the boundary between violent and peaceable people actually blurs badly. Even individuals who participate energetically in violent events generally spend most of their time and energy on nonviolent activities. Meanwhile persons who live mostly peaceful lives occasionally get involved in organizing or perpetrating violent activities, for example, by doing military service, serving on juries, or engaging in contact sports. Other people who flee violent interactions in their daily routines nevertheless support one variety of public violence or another in the proper circumstances, for example, by voting for candidates who favor the death penalty, by cheering one group of nationalists in their competition with rivals or by cooperating in what they call a just war. Indeed, most citizens of most states distinguish sharply between coercive acts committed lawfully by duly constituted agents of their own states (legitimate force) and similar acts committed by other people (illegitimate violence). Nonviolence is a beguiling principle, but one honored more in the breach than in the observance.

In general discussions of the subject, three competing views of violence prevail: as propensity-driven behavior, as instrumental interaction, and as cultural form. Treatments of violence as propensity-driven behavior locate its causes within the actor, calling attention to genetic, emotional, or cognitive peculiarities that incline a given individual, group, or category of persons, to damaging behavior more than others. Portrayals of violence as instrumental interaction characterize everything from petty assaults to all-out warfare as means (however inefficient and self-defeating) to power, wealth, prestige, or other ends. To call violence a cultural form – as in the claim that because of the frontier, slavery, or capitalist competition the USA has an exceptionally violent culture – argues that the ready availability of certain ideas, practices, models, and beliefs itself promotes violent action.

The three views suggest very different approaches to inhibiting or promoting violence. Propensity-driven violence calls for the alteration or inhibition of motives, instrumental violence for shifts of incentives and outcomes, cultural violence for teaching new ideas, practices, models, and beliefs. Alas, none of these three broad approaches to violence has yet yielded systematic explanations of the genesis, trajectory, outcome, or sheer variety of public violence as a whole. Perhaps we await the social scientific Isaac Newton who will detect underlying regularities in all the diverse phenomena people call violent. More likely the notion of violence resembles such other widespread but diffuse terms as disorder and goodness: morally powerful categories that do not refer to any causally coherent domain.

In any case, the word 'violence' almost always arrives with baggage of disapproval. The distinction of violent from nonviolent social interactions usually depends on a moral boundary, or at least activates one. Precisely because of that moral charge, advocates of one cause or another frequently extend the term 'violence' to varieties of harm far outside the range of direct, short-term infliction of physical damage on one person by another. They label hate speech, pornography, poverty, unemployment, malnutrition, and other ills as forms of violence. To call a phenomenon virtual violence is to condemn it, to claim that it damages something valuable even if it produces no physical destruction in the short run. As a consequence, every social scientific attempt to delimit, describe, classify, and explain public violence generates controversy. This attempt will be no exception.

We can nevertheless discipline the inquiry by adopting a restrictive definition of the phenomenon under analysis. Public violence, for present purposes, refers to episodic social interaction that:

- immediately inflicts physical damage on persons and/or objects

[†] In memoriam: Charles 'Chuck' Tilly (1929–2008)

- results at least in part from coordination among persons who perform the damaging acts
- occurs at least partly in publicly accessible spaces.

(The ensuing discussion will, however, extend 'physical damage' to forcible seizure of persons or objects over restraint or resistance.) At one edge, such a definition excludes strictly individual, private, impulsive, and/or accidental damage to persons or objects. At the other, it excludes long-term, incremental damage such as communication of infectious disease, cumulative wear and tear, exposure to toxic substances, and death hastened by neglect or social pressure. Nevertheless, it includes an immense range of social interactions, from minor scuffles among rival groups of voters to wholesale civil war. Despite a century or more of effort, social scientists have not so far assembled any reliable, important and systematic body of knowledge that applies to the whole range thus defined.

We therefore have no choice but to differentiate, reviewing analyses of narrower phenomena that bear some family resemblance to each other, probably share some causal processes, and have attracted sustained scholarly attention. Assuming episodic interaction in publicly accessible space, at least a little communication among actors, and some minimum of physical damage to persons or objects, Figure 1 combines the extent of coordination among damage-doers and the salience of damage in the overall pattern of interaction to produce a crude typology of public violence. ('Low' coordination means individualized action with little collective planning and signaling, while 'High' coordination means activation of differentiated and bounded organizational structure with extensive planning and signaling. 'Low' salience of violence means that most of the social interaction involved produces no short-term physical damage, while 'High' salience means interaction strongly concentrated on damage.)

Other distinctions come immediately to mind. Separation of symmetrical from asymmetrical interaction, for example, usefully distinguishes relatively even contests from situations in which a gang or an oppressor visits violence on a victim having no means of retaliation. Whether agents of a constituted government participate in public violence, as we shall see, significantly affects its character; thus we might also classify public violence by the presence or absence of governmental agents. Again, public violence probably incorporates scale effects, so that damaging interaction involving just two or three people differs qualitatively from otherwise similar interaction involving thousands of people at a time. The present classification gives priority to coordination and salience on the hunch that those two dimensions capture a significant part of the systematic variation about which social scientists have produced useful ideas and findings.

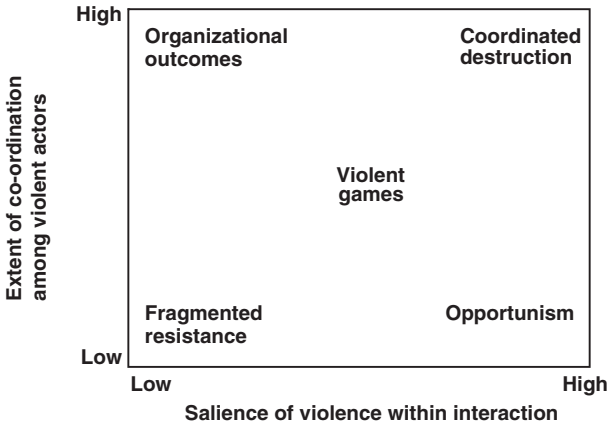

Figure 1 Crude typology of public violence.

Within the coordination-salience space we can conveniently distinguish five locations:

1. organizational outcomes: various forms of collective action generate resistance or rivalry to which one or more parties respond by actions that damage persons and/or objects; examples include demonstrations, protection rackets, governmental repression, and military coups, all of which frequently occur with no more than threats of violence, but sometimes produce physical damage;
2. coordinated destruction: persons or organizations specialized in the deployment of coercive means undertake a program of damage to persons and/or objects; examples include war, genocide, collective self-immolation, and some kinds of terrorism;
3. opportunism: as a consequence of shielding from routine surveillance and repression, individuals or clusters of individuals use immediately damaging means to pursue generally forbidden ends; examples include looting, gang rape, revenge killing, and some sorts of military pillage;
4. fragmented resistance: in the course of widespread small-scale and generally nonviolent interaction, a number of participants respond to obstacles, challenges, or restraints by means of damaging acts; examples include sabotage, clandestine attacks on symbolic objects or places, assaults of governmental agents, and the more immediately damaging forms of what James Scott (1985) calls 'weapons of the weak';
5. violent games: two or more relatively well-defined and coordinated groups follow a known interaction script entailing the infliction of damage on others as they compete for priority within a recognized arena; examples include gang rivalries, contact sports, some election battles, and some struggles among supporters of teams or stars.

Of course these crude types overlap, for example, in the activities of the South Asian thugs who sometimes act as enforcers for landlords in disputes with squatters (organizational outcomes) but at other times follow their patrons in large-scale raids on their caste, religious, or political enemies (coordinated destruction). Similarly, the various actions known as terrorism show up in all four corners of Figure 1 – as clandestine resistance, as a by-product of peaceful claim-making in highly repressive regimes, as part of genocide, and as opportunistic elimination of old enemies. It is nevertheless useful to retain the five-part division, because within each area thus delineated social scientists have accumulated at least a modicum of systematic knowledge. Let us therefore review them one by one. The following discussion gives extensive attention to organizational outcomes and coordinated destruction, then uses analogies and comparisons with the first two as it treats opportunism, fragmented resistance, and violent games much more summarily.

Organizational outcomes. A significant share of public violence occurs in the course of organized social processes that are not in themselves intrinsically violent. That is notably the case in collective political struggle. Political regimes differ dramatically in the scope they allow for nonviolent collective making of claims, for example, by petitioning, shaming, marching, voting, boycotting, striking, forming special-interest associations and issuing public messages. On the whole, democratic regimes tolerate such claim-making more readily than do their undemocratic neighbors; that is one way we recognize a regime as democratic. Even in democratic regimes, nevertheless, such forms of collective claim-making occasionally generate open violence. That occurs for three main reasons.

First, every regime empowers agents – police, troops, headmen, posses, sheriffs, and others – to monitor, contain and, on occasion, repress collective claim-making. Those agents always have some means of collective coercion at their disposal, and always enjoy some discretion in the use of those means. In one common sequence, claimants challenge repressive agents, occupy forbidden premises, attack symbolically significant objects, or seize property, then agents reply with force. Because variants on that sequence frequently occur, when repressive agents are at hand they actually perform the great bulk of the killing and wounding that occurs in public violence.

Second, collective claim-making often concerns issues that sharply divide claimants from regimes, from powerful groups allied with regimes, or from rival groups; examples are campaigns to stop current wars, outlaw abortion, or expel immigrants. In these circumstances, offended parties often respond with counterclaims backed by force, whether governmental or nongovernmental.

Third, in relatively democratic regimes an important share of collective action centers not on specific programs but on identity claims: the public assertion that a group or a constituency it represents is worthy, united, numerous, and committed (WUNC). Assertions of WUNC include marches, demonstrations, mass meetings, occupations of plants or public buildings, vigils, and hunger strikes. Even when the means they adopt are currently legal, all such assertions entail implicit threats to direct WUNC energy toward disruptive action, implicit claims to recognition as valid political actors, and implicit devaluation of other political actors within the same issue area. These features sometimes stimulate counter-action by rivals, objects of claims, or authorities, with public violence the outcome.

Organizational outcomes also include some encounters that do not begin with concerted making of claims. Border guards, tax collectors, military recruiters, census-takers, and other governmental agents, for example, sometimes generate intense resistance on the part of whole communities as they attempt to impose an unpopular measure. Similarly, audiences at theatrical performances, public ceremonies, or executions occasionally respond collectively to actions of the central figures by attacking those figures, unpopular persons who happen to be present or symbolically charged objects. By and large, organizational outcomes connect with issues over which groups are also currently contending in nonviolent ways.

One subclass of organizational outcome, however, displays a rather different pattern. Some organizations specialize in controlling coercive means, threatening to use those means if necessary, but seeking compliance without violence if possible. Examples include not only established agents of repression but also *mafiosi*, racketeers, extortionists, paramilitary forces, and perpetrators of military coups. When such specialists in coercive means encounter or anticipate resistance, they commonly mount ostentatious but selective displays of violence. Their strategy resembles that of many Old Regime European rulers, who lacked the capacity for continuous surveillance and control of their subject populations, but often responded to popular rebellion with exemplary punishment – rounding up a few supposed ringleaders, subjecting them to hideous public executions and thus warning other potential rebels of what might befall them. The strategy is most successful, ironically, when specialists in coercion never actually have to deploy their violent means.

Coordinated destruction differs in precisely that regard. It refers to those cases where persons or organizations specialized in the deployment of coercive means undertake a program of actions that damage persons and/or objects. Coordinated destruction overlaps with organizational outcomes because specialists in coercion participate in both and because threats of force sometimes escalate into struggles between coercive organizations. Here, the organizations' strategies center, however temporarily, on the production of damage. Examples include war, genocide, collective self-destruction, public penance, and government-backed terror. The major distinctions separate (1) lethal contests, (2) campaigns of annihilation, and (3) ritual harm.

In lethal contests, at least two organized groups of specialists in coercion confront each other, each one using harm to reduce or contain the others' capacity to inflict harm. War is the most general label for this class of coordinated destruction, but different variants go by the names of civil war, guerrilla, low-intensity conflict, and conquest. Although lethal contests of various sorts stretch back as far as humanity's historical record runs, the standard image of two or more disciplined national armies engaged in destroying each other within generally accepted rules of combat only applies to a very small historical segment: roughly 1650 to 1950 for Europe, a few much earlier periods for China, even rarer intervals elsewhere in the world. Outside of those exceptional moments, autonomous raiding parties, temporary feudal levies, mercenary assemblages, bandits, pirates, nomads doubling as cavalry, mobilized villages, and similar conglomerate or part-time forces have fought most historical wars.

Lethal contests shade over into campaigns of annihilation when one contestant wields overwhelming force or the object of attack is not an organization specialized in the deployment of coercive means. Analysts have employed the term genocide for those campaigns in which attackers define their victims in terms of shared heritage and politicide for those in which victims belong to a common political category; so far no commonly accepted term has emerged for similar campaigns aimed at members of religious or regional categories. The usual stakes in campaigns of annihilation are collective survival, on one side, and recognition as the sole party with the right to territorial control, on the other. Because of those stakes, such

struggles tend to generate vast mobilizations of support extending far beyond the specialists in coercion who initiate them.

Ritual harm plays a much smaller part in contemporary public violence than do lethal contests and campaigns of annihilation. At times, however, they have assumed considerable importance, for example, in the potlatch (aggressive, competitive destruction of the performer's own symbolically charged property) that once marked struggles for precedence along America's northwest coast. Other well known instances include the processions of flagellants once organized regularly by Renaissance Italy's penitent confraternities. Today self-immolation, collective suicide, and hunger strikes occasionally occur as gestures of protest, serving chiefly to dramatize the positions of groups that lack other political resources.

Within the broad category of coordinated destruction, a remarkable change occurred during the half century following World War II. Where interstate wars among well-identified national armies had predominated for a century or more, shifts toward a much wider variety of coercive forces and toward campaigns of annihilation elevated civil war, broadly defined, to the chief setting of coordinated destruction. Decolonization, expansion of world trade in arms and drugs, reappearance of mercenary forces, and the weakening of central state capacity in many world regions all contributed to that change.

Opportunism combines low levels of coordination with high salience of violence within the total interaction. Opportunism benefits from some of the same processes that have favored recent moves away from interstate war toward violent struggle within state boundaries. Looting, gang rape, revenge killing, and some sorts of military pillage all occur more frequently in interstices, peripheries, and lapses of central state control. In fact, opportunistic public violence often occurs as an outgrowth of coordinated destruction. Disbanded but still armed troops, for example, frequently use their arms to prey upon civilians. The destruction wrought by mercenaries during truces, marches, and aftermaths of wars made them the bane of European civilian life during the sixteenth and seventeenth century heyday of mercenary armies. Similarly, when major protests or disasters divert the attention of repressive forces, looters and killers often take advantage of relaxed control to pursue their own ends.

Fragmented resistance involves segmented or weakly coordinated damage to persons or objects in the course of interactions that are not predominantly violent. When governmental agents, landlords, employers, religious authorities, or other powerful people impose innovations, exactions, or constraints on subject populations, members of those populations often resist indirectly and on the small scale. They engage in sabotage, ambushes, and clandestine attacks on symbolic objects or places, and temporary acts of resistance that either inflict damage on authorities or incite violent reprisals from authorities. Fragmented resistance differs from organizational outcomes chiefly in the balance of forces among parties and the scale of organization among the weaker parties. Otherwise, the same regularities that apply to organizational outcomes apply to fragmented resistance.

In its most stylized forms, such as the medieval tournament, coordinated destruction comes to resemble violent games. Likewise, such violent games as ritual fights between youths of neighboring villages edge over into coordinated destruction. Still, we can usefully distinguish a class of public violence in which two or more relatively well defined and coordinated groups follow a known interaction script entailing infliction of damage on others as they compete for priority within a recognized arena. Age grades of local youths, organized gangs, sports teams, supporters of powerful or popular figures, electoral factions and fraternal orders inducting new members all sometimes engage in violent games. Frequently the outcome of the contest produces confirmation or alteration of a rank order that in turn entitles winners and their followers to prestige or privilege. In that sense, violent games stand somewhere between (1) WUNC performances (whether violent or not) and (2) the public displays of wealth, power, and patronage by which warlords, crime bosses, political magnates, heads of influential families, *caciques*, and entertainment stars maintain their credibility.

In all its forms, public violence interacts closely with nonviolent politics. Violence as organizational outcome or coordinated destruction clearly entails struggles for power and takes shape from the character of power struggles spanning violent and nonviolent action. Violence as opportunism and fragmented resistance often occurs as a by-product of grand political processes: state formation, revolution, war, military conscription, taxation, and patron–client realignment. Violent games follow their own rules, but frequently intersect with or mutate into other forms of public violence. At first glance, violence may seem the antithesis of orderly public politics. At second glance, most public violence connects closely with politics in general.

See also: Conflict and Conflict Resolution, Social Psychology of; Conflict and War: Anthropological Aspects; Ethnic Conflict, Geography of; Genocide, Historical Aspects of; Military History; Military Sociology; Police: A Sociology of Knowledge Approach; Revolutions, Sociology of; Sport, Sociology of; Terrorism; Violence and Media; Violence in Anthropology; Violence, History of; War, Sociology of; Warfare in History.

Bibliography

Beissinger, M.R., 1998. Nationalist violence and the state. Political authority and contentious repertoires in the former USSR. Comparative Politics 30, 401–433.

Brass, P.R. (Ed.), 1996. Riots and Pogroms. New York University Press, New York.

Brubaker, R., Laitin, D.D., 1998. Ethnic and nationalist violence. Annual Review of Sociology 24, 423–52.

Dudley, L.M., 1991. The Word and the Sword. How Techniques of Information and Violence Have Shaped the World. Blackwell, Oxford, UK.

Fogarty, B.E., 2000. War, Peace, and the Social Order. Westview Press, Boulder, CO.

Gambetta, D., 1993. The Sicilian Mafia. The Business of Private Protection. Harvard University Press, Cambridge, MA.

Gould, R.V., 1999. Collective violence and group solidarity: evidence from a feuding society. American Sociological Review 64, 356–380.

Grimshaw, A.D., 1999. Genocide and democide. In: Encyclopedia of Violence, Peace, and Conflict, vol. 2. Academic Press, San Diego, CA, pp. 53–74.

Gurr, T.R. (Ed.), 1989, Violence in America, vols. 2. Sage, Newbury Park, CA.

Gurr, T.R., Harff, B., 1994. Ethnic Conflict in World Politics. Westview Press, Boulder, CO.

Kaldor, M., 1999. New and Old Wars. Organized Violence in a Global Era. Polity Press, Cambridge, UK.

Licklider, R. (Ed.), 1993. Stopping the Killing. How Civil Wars End. New York University Press, New York.

della Porta, D., 1995. Social Movements, Political Violence, and the State. A Comparative Analysis of Italy and Germany. Cambridge University Press, Cambridge, UK.

Reiss Jr., A.J., Roth, J.A. (Eds.), 1993. Understanding and Preventing Violence.. National Academy Press, Washington, DC.

Reyna, S.P., Downs, R.E. (Eds.), 1999. Deadly Developments. Capitalism, States and War. Gordon and Breach, Amsterdam.

Rule, J.B., 1988. Theories of Civil Violence. University of California Press, Berkeley, CA.

Scott, J.C., 1985. Weapons of the Weak. Everyday Forms of Peasant Resistance. Yale University Press, New Haven, CT.

Scott, J.C., 1990. Domination and the Arts of Resistance. Hidden Transcripts. Yale University Press, New Haven, CT.

Scott, J.C., 1998. Seeing Like a State. How Certain Schemes to Improve the Human Condition Have Failed. Yale University Press, New Haven, CT.

Stanley, W., 1996. The Protection Racket State. Elite Politics, Military Extortion, and Civil War in El Salvador. Temple University Press, Philadelphia.

Thomson, J.E., 1994. Mercenaries, Pirates, and Sovereigns. State-Building and Extraterritorial Violence in Early Modern Europe. Princeton University Press, Princeton, NJ.

de Waal, A., 1997. Famine Crimes. Politics & the Disaster Relief Industry in Africa. Indiana University Press, Bloomington, IN.

Zolberg, A.R., Suhrke, A., Aguayo, S., 1989. Escape from Violence. Conflict and the Refugee Crisis in the Developing World. Oxford University Press, New York.

Virtual Reality and Spatial Cognition

Gregor Hardiess and Hanspeter A Mallot, University of Tübingen, Tübingen, Germany
Tobias Meilinger, Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the significance of virtual reality within the field of spatial cognition is outlined. The role of virtual reality is grouped in three sections addressing (1) the current and latest technology of virtual reality regarding the two main functions within virtual reality, that is, technology to interact with virtuality (input devices used to record observer actions and output devices used to simulate sensory stimuli) and technology for presenting the virtual environments to the user, (2) the usage of this technology for the purpose of research in the field of spatial cognition regarding behavioral and neuronal processes (discussing advantages and disadvantages of virtual reality), and (3) virtual reality experiments and their results that are relevant in current research of spatial cognition covering place memory, wayfinding in large-scale spaces, and the neural representations of spatial features.

The term 'virtual reality' was first used and introduced by Jaron Lanier in 1989. By definition, virtual reality is a collection of technologies for generating a human–computer interface that allows people to interact efficiently with, become immersed in, and to feel the presence in computerized 3D environments, while using their natural senses and motor skills in real time. Here, immersion refers to the technical capability of the system to deliver a surrounding and convincing environment with which the user can interact. Presence is the sense of being (embodied) in a virtual environment rather than in the physical space where the real user's body is actually located.

During the last years, computer technology has improved substantially and with it the factors supporting immersion and presence. Improvements concern the complexity and richness of the details of the virtual environments displayed or the diversity and fidelity of sensory and motor interactions involved in the virtuality. The following sections will address these advancements and discuss their application within the field of spatial cognition.

The Technology of Virtual Reality

Virtual reality is largely based on computer graphics. The quality of computer graphics increased substantially within the last years affecting features such as the size of virtual environments, their level of detail, the realism of surface texture, the modeling of illumination, shading and shadowing, and the animation of objects and avatars. At the same time, the fidelity of sound and the speed of rendering (i.e., the process of computing a 2D image from a 3D database) have improved. Through these advancements, virtual environments are now able to closely approximate real world conditions. As a result of a development mostly driven by the computer game industry, graphics engines are now frequently available as open-source software (e.g., Unreal Engine 3, CryENGINE 3, OpenSceneGraph, and Ogre3D) and run on cheap, custom hardware with satisfactory results, that is, engines enable rendering of and allow interactions with virtual environments in real time and in a highly realistic manner.

In terms of interaction with virtual environments, users can rely on classical devices like the computer mouse, joystick, and gamepad (Figure 1(a)) to act in the virtual space. In addition, a large number of more sophisticated devices have been developed for special purposes. The flystick is an extended joystick where the position is tracked in six degrees of freedom (Figure 1(b)). The Phantom® device combines position sensing of a single finger with force feedback to that finger to create a closed loop of haptic interaction. Motion capture with devices such as the Kinect controller (Microsoft®) is now common even in the consumer market. High-precision motion capturing of the head, the hand (e.g., data glove; Figure 1(c)), or the whole body allows the use of real body movements to alter the viewing direction, to manipulate objects, to change the body position, or to do all these actions together in the virtual reality. Furthermore, body capturing can be used to animate an avatar (a virtual agent), usually a person who is visible within a virtual environment and can interact with it. Movement devices such as the VirtuSphere (Figure 1(d)) or the omnidirectional 'Cyberwalk' treadmill (Figure 1(e)) allow users to walk for unlimited distances within a virtual environment thus expanding the physical limitations of a restricted tracking space in laboratory rooms. Furthermore, such movement devices facilitate the sense of real walking and provide movement tracking at the same time, which is a combination hard to realize in reality.

Concerning visual stimulation, today a huge number of devices are available that allow the presentation of virtual sceneries to the user. In most cases stimuli are presented on flat desktop screens, covering just a restricted part of the human field of view. Usually, users have to rely on a mouse or a keyboard to interact. A more direct interaction becomes possible when using head tracking together with a head mounted display (Figure 1(a)). In a head mounted display, stimuli are presented by the use of two small LCD screens placed in front of the eyes in order to interlink real head movements with rotations and translations in the virtual world. Such head mounted displays are also applied when walking on treadmills or in walking spheres (see above). Stimulating the entire field of view of

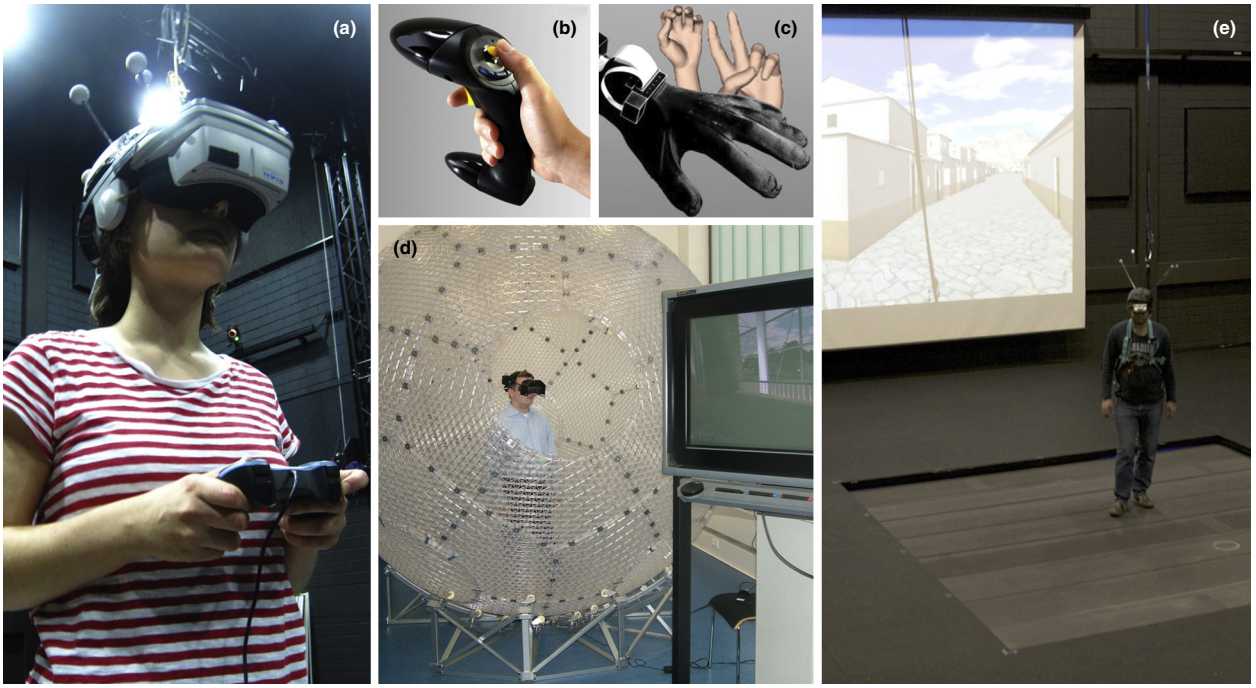

Figure 1 Devices for interacting with and presenting of virtual environments. (a) Subject is using a gamepad to interact within a virtual environment presented via head mounted display (from Max Planck Institute for Biological Cybernetics, Tübingen). (b) Flystick to interact in all possible degrees of freedom. (c) Data glove to control the movement of a virtual hand. (d) Subject is walking in the VirtuSphere (from Cognitive Neuroinformatics, University of Bremen). (e) Subject walking on the omnidirectional 'cyberwalk' treadmill (from Max Planck Institute for Biological Cybernetics, Tübingen; http://www.cyberwalk-project.org/).

a human subject requires large projection screens (curved or flat; Figure 2) or the combination of such multiple screens in a so-called Cave. Furthermore, stereoscopic stimulation can increase the feeling of immersion by providing realistic binocular depth cues. Stereoscopic 3D effects are produced by means of head mounted displays, anaglyph glasses, shutter glasses, polarization glasses, or by special 3D displays.

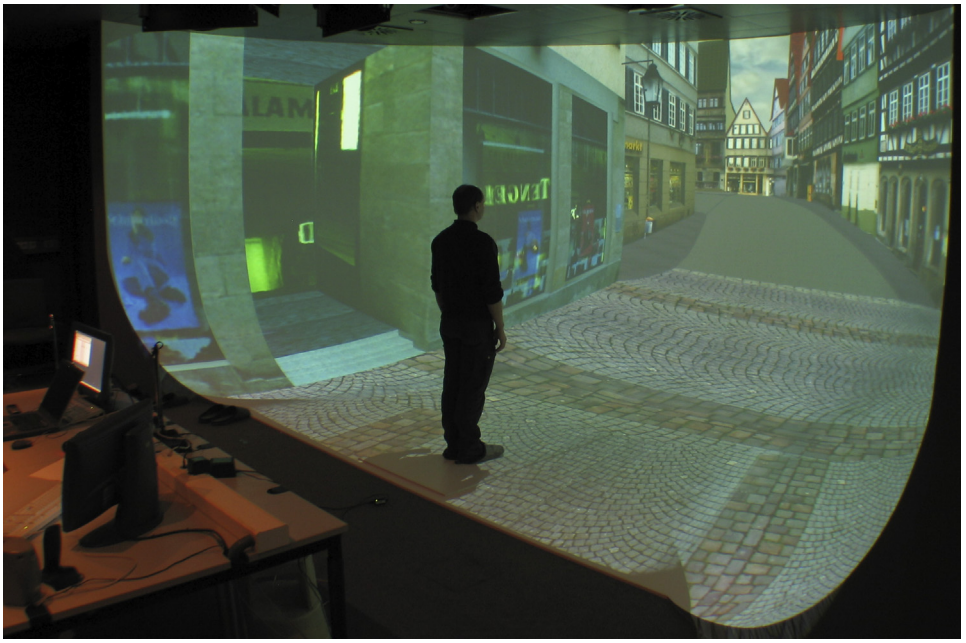

Figure 2 Virtual Tübingen displayed in the PanoLab, a wide-area high realistic projection system for interactive presentations of virtual environments (http://virtual.tuebingen.mpg.de/PanoLab.html).

The Application of Virtual Reality in Spatial Cognition

In the field of spatial cognition, researchers investigate behavioral and neuronal processes associated with the perception, representation, revision, and utilization of knowledge about spatial environments or conditions and the behaviors bearing on such knowledge. Spatial knowledge is apparently associated with locations of something in relation to a spatial reference. Research on behavior attainments focuses on orienting in spatial layouts, homing (i.e., the ability to find the way back to a certain home place), pointing toward goals in an environment, extracting visuospatial information by head and eye movements, navigating in large-scale (urban or cluttered) environments or mazes, and the communication about spatial facts. Concerning neuronal processes, questions about brain areas involved in spatial cognition, the representational format as well as the metrics of spatial codes, the involvement of different senses (i.e., multisensory processing), and the characteristics of spatial features crucial for spatial behavior are examined.

Spatial cognition research is often closely related to virtual reality technology. In fact, for the majority of research questions, the simulation of spatial conditions is a requirement that at least substantially simplifies the experimental procedures. Imagine a navigational experiment where subjects have to find certain routes within a town or another large-scale terrain. Here, it would be hardly possible to ensure repeatability in a highly controlled manner when performing such a task in a real urban environment. Furthermore, manipulations of the spatial layout or metrics, objects within the scenery, or the sensorimotor processing cannot be realized. Consequently, spatial cognition profits greatly from the advancements in virtual reality technology:

1. Virtual reality is highly controllable and makes procedures repeatable with respect to the design of the environment and the way of interacting with it.
2. Body position, pointing direction, and exploration movements can be measured with high precision in real time.
3. Manipulations of the environment, of viewpoints or metrics of the virtual space (e.g., physical inconsistencies or non-Euclidean metrics), and of interactions with the virtual reality are possible in real time.
4. Different sensory modalities related to spatial perception can be tested selectively and can be brought into competition with one another in order to estimate their relative contributions for a given task.
5. Virtual reality enables the measurement of spatial behavior in large-scale environments also under real walking conditions (e.g., treadmill or walking sphere).
6. Virtual reality enables recordings of neuronal activities during spatial performance (e.g., by use of EEG or fMRI). Single or multicell recordings are done in rats or mice using an adapted version of the walking sphere (Hölscher et al., 2005; Figure 4).
7. Experiments in the real space may be inappropriate or impossible because of inaccessibility, cost, excessive danger, security requirements, etc. These limitations can be bypassed when using virtual reality.

On the other hand, virtual reality entails some disadvantages or attentions the user must take into consideration. Cyber- or simulator sickness is one of them (Stanney et al., 1998). It can be caused by temporal interrupts or delays in the interaction and stimulation due to hard- or software processes or by cross-sensory conflicts between vestibular and visual stimuli. For example, the perception of self-movement (vection) can be induced purely visually without stimulating the vestibular or the proprioceptive system, yielding a conflict that might result in simulator sickness. The symptoms of this sickness include disorientation, disequilibrium, drowsiness, salivation, sweating, and vomiting. Another aspect when using virtual reality is the conflict between the virtual and the physical reality. Both environments might be represented partially at the same time and generate conflict between each other, for example, when a virtual and a real room differ in orientation (May, 2004). Just like with simulator sickness, the amount of disturbance will depend on the used setup and the task at hand.

In virtual reality, the user is always confronted with virtual stimuli, generated by a certain technology to simulate the reality. The degree of realism in virtual reality as well as the way of presentation and interaction has to be chosen with respect to the scientific question or the kind of task to investigate. Hence, experimenters need the knowledge about the range of possibilities when simulating real conditions. Additionally, when applying virtual environments, the user needs a certain amount of training to get familiar with the simulated realism. Setups for virtual reality can also be very expensive and need a high amount of engineering power for successful running (e.g., Cave or treadmill setup).

Spatial Cognition: Experiments in Virtual Reality

Virtual reality is applied in various research areas including spatial cognition. The following experiments in human and animal navigation will show how virtual reality can contribute to examining such a field. These examples address (1) the contribution of different information sources, (2) the selection of environmental features, and (3) the neuronal implementation in the brain.

Virtual reality can help to identifying whether humans integrate and memorize their paths walked and/or rather rely on visual information such as landmarks (i.e., environmental, salient objects that are easy to recognize) encountered. Foo et al. (2005) had participants physically walk through a hall while displaying a virtual environment to them via a head mounted display. When putting spatial information derived from paths walked before and visual landmarks into conflict, participants heavily relied on visual landmarks and tended to ignore conflicting body information. So in the case of conflict, humans identify locations visually rather than use on body-based information. However, body cues do contribute in addition to visual information as well. Ruddle et al. (2011) compared walking through a large-scale space on an omnidirectional treadmill with navigating the same space only visually. Participant's direction and distance estimates to remote locations profited from physical walking distances within the environment, but not from physical rotations alone as compared to navigating only visually. Although humans do rely mainly on visual information,

body-based cues especially traversing distances contribute to our memory as well.

If vision is important, which features of the visual information are used by navigators? Steck and Mallot (2000) displayed local and global landmarks (i.e., houses and mountains) on a screen covering 180° of the horizontal field of view. Participants memorized both types of landmarks using them to retrace a route when the one or other type was absent. When putting local and global landmarks into conflict (by displacing the global ones) participants showed quite individual preferences. Stankiewicz and Kalia (2007) had participants navigate through a virtual maze with certain objects and intersections appearing rarely or more often. Subjects relied more strongly on landmarks that were more salient or specific. Structural landmarks such as walls were valued higher than object landmarks. Humans seem flexible with respect to which visual information to use. In general, they focus on informative landmarks and on the surrounding geometry. Gillner et al. (2008) used virtual reality to generate an environment where smooth color distributions provided the sole cue to position while features and objects that could have provided landmark information were absent. Still, subjects were able to home and recognize places, indicating that a simple, 'snapshot-matching' mechanism of place recognition is available to humans.

Virtual reality can not only be used to examine newly learned spaces. Frankenstein et al. (2012) asked inhabitants of Tübingen to point to locations within a photorealistic model of their hometown (virtual Tübingen; Figures 2 and 3). Participants performed the more accurately the more closely they were facing north within virtual Tübingen and their pointing performance did not differ at all for close-by and far away pointing targets. This indicates that spatial memory within their hometown was acquired from maps and combined with a different source of information, namely navigation-based information in order to point from within an environment.

Virtual reality is also highly suited to examining the neural basis of location memory and navigation. This is, for example, done by measuring brain activity during moving through a virtual environment. In an early study, Maguire et al. (1998) showed that visual navigation as compared to simply following an indicated path is associated with activity in parietal areas and the hippocampus. The more directly navigators approached a familiar route the higher was the hippocampal activity. Wolbers and Büchel (2005) showed that the integration of locally experienced views within a virtual environment is paralleled by increased hippocampal activity. Hippocampus is involved in relating multiple views; other regions are involved in different aspects of navigation. Janzen and Weststeijn (2007) showed that the relevance of landmarks (i.e., whether they were located at decision points or at nondecision points) is processed in the parahippocampus, whereas the processing of route order is related to various other structures, for example, parietal areas.

Virtual reality is not limited to humans. Dombeck et al. (2010) as well as Hölscher et al. (2005) had rats walking a movable sphere while displaying their progression visually (Figure 4). Then they measured electric currents in implanted electrodes within the hippocampus and identified so-called place cells that code for a certain location within the space (O'Keefe and Dostrovsky, 1971). This shows that even animals react toward a virtual reality in a way they behave within real environments.

The examples show how virtual reality can be fruitfully used within spatial cognition research. Despite some disadvantages such as the occurrence of simulator sickness, virtual reality allows for highly realistic stimulation while at the same time keeping full experimental control. With the technology further improving and especially becoming cheaper and more easy to use, virtual realities will also become more widespread not just within spatial cognition research.

Figure 3 Virtual Tübingen. View on the city hall from the virtual model of Tübingen, Germany (http://virtual.tuebingen.mpg.de/). Virtual Tübingen was used by Frankenstein et al. (2012).

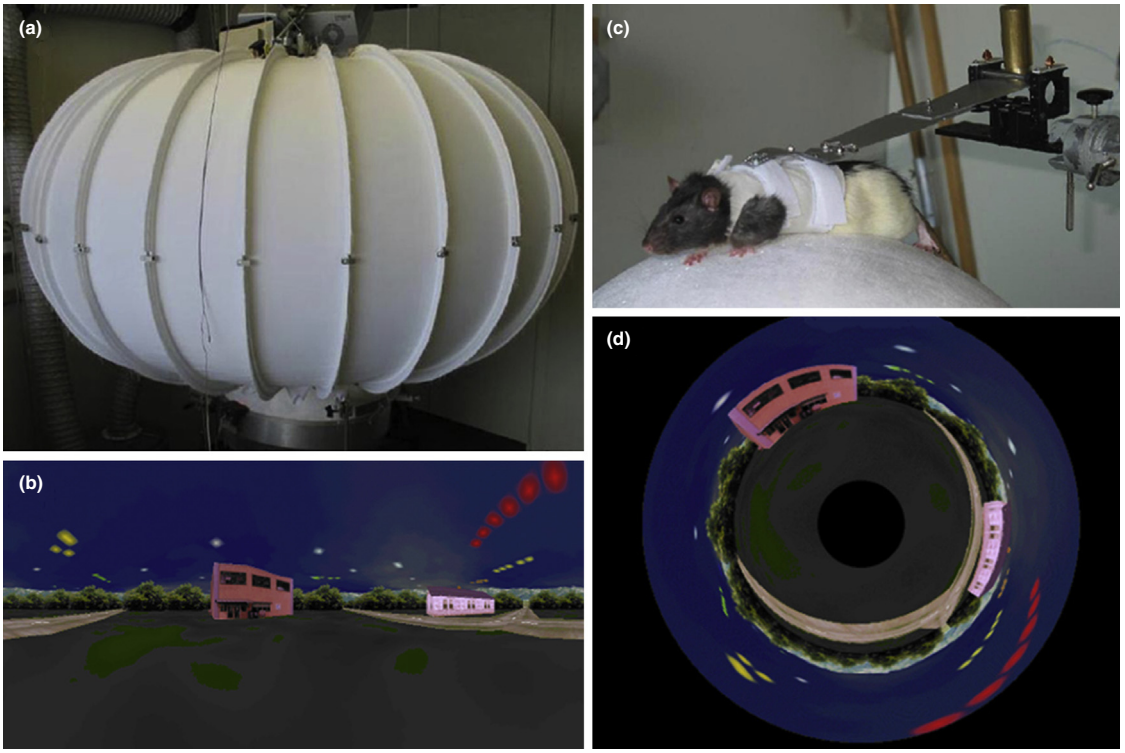

Figure 4 Virtual reality projection setup for rats running on top of an air-cushioned polystyrene sphere as used by Hölscher et al. (2005). (a) A 360° projection screen. (b) Image generated by the world cameras with a 360° field of view in horizontal direction. (c) Rat fixed on top of the sphere is located inside of the screen. (d) Distorted image projected via one video projector onto the 360° projection screen.

See also: Human Spatial Orientation, Neural Basis of; Navigation in Virtual Space: Psychological and Neural Aspects; Perception of Action Space: Using Multiple Frames of Reference; Workplace Deviance.

Bibliography

Dombeck, D.A., Harvey, C.D., Tian, L., Looger, L.L., Tank, D.W., 2010. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. Nature Neuroscience 13, 1433–1440.

Foo, P., Warren, W.H., Duchon, A., Tarr, M., 2005. Do humans integrate routes into a cognitive map? Map- versus landmark-based navigation of novel shortcuts. Journal of Experimental Psychology: Learning, Memory, and Cognition 31, 195–215.

Frankenstein, J., Mohler, B.J., Bülthoff, H.H., Meilinger, T., 2012. Is the map in our head Oriented North? Psychological Science 23, 120–125.

Gillner, S., Weiss, A.M., Mallot, H.A., 2008. Visual homing in the absence of feature-based landmark information. Cognition 109, 105–122.

Hölscher, C., Schnee, A., Dahmen, H., Setia, L., Mallot, H.A., 2005. Rats are able to navigate in virtual environments. The Journal of Experimental Biology 208, 561–569.

Janzen, G., Weststeijn, C.G., 2007. Neural representation of object location and route direction: an event-related fMRI study. Brain Research 1165, 116–125.

Maguire, E.A., Burgess, N., Donnett, J.G., Frackowiak, R.S.J., Frith, C.D., O'Keefe, J., 1998. Knowing where and getting there: a human navigation network. Science 280, 921–924.

May, M., 2004. Imaginal perspective switches in remembered environments: transformation versus interference accounts. Cognitive Psychology 48, 163–206.

O'Keefe, J., Dostrovsky, J., 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Research 34, 171–175.

Ruddle, R.A., Volkova, E., Bülthoff, H.H., 2011. Walking improves your cognitive map in environments that are large-scale and large in extent. ACM Transactions on Computer–Human Interaction 18, 1–22.

Stankiewicz, B.J., Kalia, A., 2007. Acquisition and retention of structural versus object landmark knowledge when navigating through a large-scale space. Journal of Experimental Psychology: Human Perception & Performance 33, 378–390.

Stanney, K.M., Mourant, R.R., Kennedy, R.S., 1998. Human factors issues in virtual environments: a review of the literature. Presence 7, 327–351.

Steck, S.D., Mallot, H.A., 2000. The role of global and local landmarks in virtual environment navigation. Presence: Teleoperators and Virtual Environments 9, 69–83.

Wolbers, T., Büchel, C., 2005. Dissociable retrosplenial and hippocampal contributions to successful formation of survey representations. Journal of Neuroscience 25, 3333–3340.

Virtual Reality in Psychotherapy

Julia Diemer, University of Regensburg, Regensburg, Germany
Paul Pauli, University of Würzburg, Würzburg, Germany
Andreas Mühlberger, University of Regensburg, Regensburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Virtual reality (VR) provides a new medium for the presentation of emotionally relevant, complex stimuli in an ecologically valid and at the same time highly controlled manner. Therefore, VR applications have attracted the attention of psychotherapists and researchers. In the past years, viable VR systems for varied applications in psychotherapy have been developed, most notably for exposure therapy in anxiety disorders. This article resumes the major applications of VR in psychotherapy, gives an overview of the current literature on its efficacy, and provides an outlook on future perspectives for psychotherapy and clinical research.

Virtual reality (VR) is a computer-generated presentation that differs from common video or audio technology by its interactive nature. VR aims for the greatest possible immersion in order to create in the user a sense of presence in the virtual environment. To achieve this, VR systems provide input to the user's sensory channels and replace sensory input from the real world. The user views the VR environment via a head-mounted display (HMD; **Figure 1**) or large projection screens (CAVE system; **Figure 2**). Besides, most VR systems also provide auditory input, and input of other modalities like olfactory and haptic stimulation is also possible. Importantly, the computer system that controls the VR presentation receives user information through sensors and/or input devices, e.g., head tracking systems, joystick, buttons, or a personal computer mouse. Adapting the VR environment in real time, the computer enables the user to experience the illusion of being in the virtual world. Both user and VR system receive mutual feedback, making the system truly interactive.

VR systems were first used for simulator training in aviation, spaceflight, or for military purposes. In these applications, VR enables users to learn how to operate a complex technical system in a risk-free simulation, before applying their skills to the real, often dangerous machinery. In clinical psychology, VR has attracted attention for its potential to create an emotional experience and to instigate processes of emotional learning, which lie at the heart of psychotherapy. Today, VR technology is used in a fast-growing number of applications in clinical research and therapy, particularly as a form of exposure therapy for anxiety disorders called virtual reality exposure therapy (VRET). The history of VR therapy is summarized in **Box 1**.

Exposure Treatment: A Standard Technique in Cognitive Behavioral Therapy for Anxiety Disorders

State-of-the-art cognitive-behavioral therapy for anxiety disorders includes *in vivo* or in *sensu* exposure treatment. After cognitive preparation (including psychoeducation), patients are encouraged to overcome their avoidance behavior and to confront the feared objects or situations in real life or in their imagination until a reduction of subjective and physiological symptoms of anxiety occurs. *In vivo* exposure is especially effective: Response rates of up to 80–90% have been reported in randomized-controlled trials (Choy et al., 2007).

Probably the most influential theory about the processes underlying the efficacy of exposure therapy was provided by Foa and Kozak (1986). Their emotional processing theory postulates that irrational fear may be conceptualized as a 'fear network' that links the anxiety-provoking stimuli with cognitive (e.g., evaluation as threatening), emotional (subjective experience of fear and anxiety), physiological (e.g., increase in heart rate), and behavioral (e.g., avoidance) responses. According to Foa and Kozak (1986), exposure therapy works by altering this fear structure. Information elements that are incompatible with the fear structure, e.g., the information that the feared consequences do not occur, are incorporated into the fear structure. Emotional processing theory also considers psychophysiological arousal a necessary first step in exposure therapy, as arousal indicates that the fear structure has been activated and is thus accessible to change.

Virtual Reality Exposure Therapy

While the high efficacy of *in vivo* exposure therapy is well documented, in practice many obstacles and problems with this form of therapy remain. Major drawbacks of *in vivo* exposure are its time-consuming nature and the difficulty of creating relevant exposure situations for many anxiety disorders (e.g., social anxiety disorder, agoraphobia, and posttraumatic stress disorder (PTSD)) (Mühlberger and Pauli, 2011). VR offers an alternative medium for conducting exposure therapy. As an emerging alternative, VRET might help to overcome these obstacles. VRET has been tested for most anxiety disorders, including specific phobia, panic disorder and agoraphobia, PTSD, and social anxiety disorder (Bouchard et al., 2007). For generalized anxiety disorder and obsessive–compulsive disorder, few studies exist so far.

Advantages and Limitations of VRET

VRET offers a range of advantages that make this technique attractive for mental health professionals. The 'gold standard'

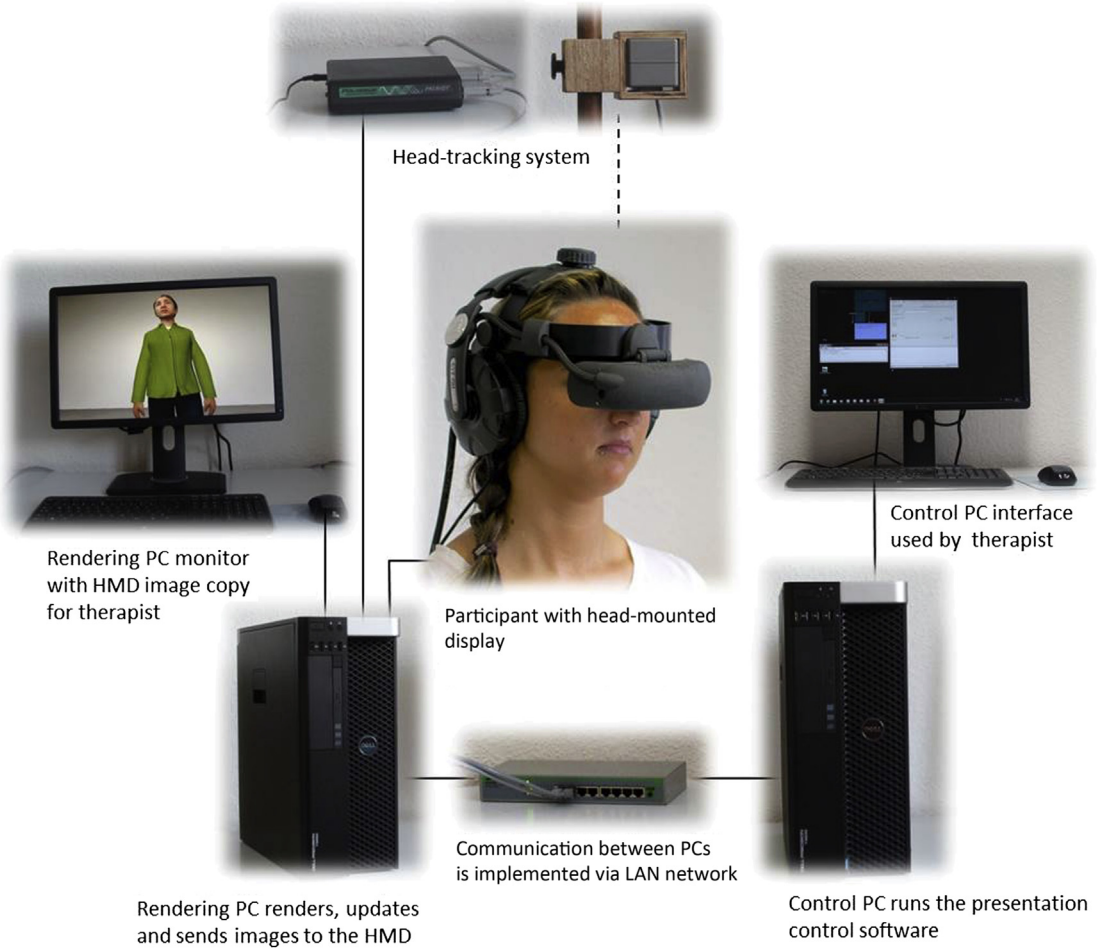

Figure 1 Basic components of a VR system as used at the University of Regensburg, Germany.

Figure 2 Driving environment in the PsyCAVE at the University of Würzburg, Germany. Reproduced from Mühlberger, A., 2014. Virtuelle Realität in der Klinischen Emotions- und Psychotherapieforschung. In: Jeschke, S., Kobbelt, L., Dröege, A. (Eds.), Exploring Virtuality, Springer, Berlin/Heidelberg.

in exposure-based treatment, *in vivo* exposure, is often time-consuming, costly (e.g., in aviophobia), and requires a great deal of preparation on the side of the therapist in order to provide the necessary stimuli or create the situation that is needed. These factors limit the availability of exposure-based treatment. As a result, many patients who would benefit from this form of therapy do not receive exposure therapy at all. VRET offers an alternative that can take place in the therapist's office. Relevant exposure elements can be repeated effortlessly as many times as needed. For some anxiety disorders, especially PTSD, *in vivo* exposure to the relevant stimuli may be dangerous, or impossible. Therapists often rely on their patients' imagination to recreate the relevant experience from their memory, a task that is inherently difficult for most patients suffering from PTSD. Here, VRET not only provides a medium to present relevant environments to recall and confront traumatic memories, but also helps the therapist to pace the exposure process and to keep control of its gradual progression.

In general, due to high levels of apprehension and avoidance, many patients with anxiety disorders do not seek treatment, or drop out of treatment. There are indications that patients might accept VRET more easily than standard *in vivo* exposure. García-Palacios et al. (2007) found that of a mixed

Box 1 History of VR therapy

Already in the 1960s, foreseeing the potential of computer technology, Sutherland conducted first experiments with a still cumbersome HMD that was hung from the ceiling (Sutherland, 1968). This display could project simple three-dimensional structures. It was already equipped with a sensor that recorded head movements of the user, and the display was updated using this information. Before VR technology was ready for clinical applications, it was used mainly for training purposes, especially in aviation, spaceflight, and for military purposes. The first publications on VR as an exposure technique for specific phobia date back to the 1990s. Since then, the developments of interactive computer graphics in the field of PC games have led to a fast-developing range of applications. Apart from clinical psychology and psychotherapy, VR scenarios are also increasingly used as aids to motor rehabilitation for patients suffering from neurological disorders.

patient group (64.7% specific phobia, 35.3% social phobia; $n = 150$), 76% preferred VRET over *in vivo* exposure. About 27% indicated they would refuse *in vivo* exposure therapy, while just 3% reported they would reject VRET. Further, the highly controllable setting makes the processes that occur during VRET accessible to researchers; self-reported, behavioral, and physiological levels of anxiety can be assessed online. Thus, VR technology is a promising research tool, too. **Box 2** summarizes the advantages of VRET.

Limitations of VRET include technological and clinical barriers. VR systems may provoke simulator sickness, characterized by symptoms like nausea, visual strain, headache, and dizziness. Simulator sickness resembles motion sickness, and is related to a mismatch between sensory, proprioceptive, and vestibular information (e.g., a user is moving visually through a VR environment, but is physically standing still). However, severe simulator sickness is rare. In fact, fewer than 5% of the patients discontinue VRET due to simulator sickness. Another obstacle is that some patients do not experience considerable anxiety during VRET; however, whether this affects treatment outcome has not been investigated yet. Because of limited empirical data, VR researchers suggest that patients with difficulties to differentiate between real events and fiction, e.g., patients with symptoms of acute psychosis, should not receive VR-based treatment, and psychosis is commonly defined as an exclusion criterion in studies (e.g., Gorini et al., 2010). Interestingly, recent preliminary studies hint at a positive effect of VR skills training for patients with psychosis (Kurtz et al., 2007; Park et al., 2011; see below). Thus, further research might lead to a revision of this limitation and individually tailored VR treatments might be useful even in those patients after careful consideration of the individual case and symptoms. An important limitation for the clinician is the lack of commercially available software packages for conducting VRET. So far, researchers have relied heavily on shareware solutions that are restricted to a small selection of fears or require varying degrees of programming skills. We expect, however, that appropriate software solutions will appear on the market soon, making VRET available to clinicians on a large scale. Additionally, new hardware developments (e.g., cost-effective HMDs with large field of view) will enhance presence, reduce side effects, and promote the application of VRET.

VRET for Specific Phobia

By its nature, VRET is most easily applicable to anxiety disorders with distinctive phobic stimuli, which can be readily implemented in a VR environment, as is the case for specific phobia. VRET has been applied most often to aviophobia, acrophobia, and arachnophobia. Other, more complex stimuli, like virtual humans (agents) for VRET of social phobia are more difficult to render in VR. It is therefore not surprising that specific phobia has been the first and most intensely studied disorder in VRET research. For an example of VR therapy for tunnel phobia, see **Box 3**.

VRET for Aviophobia

In vivo exposure for aviophobia commonly includes at least one commercial flight, making this treatment rather costly for the patient. Here, VRET offers some specific advantages over *in vivo*

Box 2 Advantages of VRET

- Cost-effectiveness (especially for aviophobia and driving phobia).
- Less effort in organizing exposure sessions (especially for agoraphobia and social phobia).
- Confidentiality and safety.
- Exposure may be repeated easily.
- Specific stimuli (e.g., turbulence) may be implemented and presented repeatedly as needed.
- Type, intensity, and duration of exposure are entirely controlled by the therapist.
- Patients' reactions may be observed and recorded.
- Patients may be motivated more easily to try out new behavior.
- High patient acceptance and low threshold for seeking treatment.
- Patients lacking in imagination may be treated more effectively than with imaginal exposure.
- The laboratory setting makes VRET easily accessible for research.

(Adapted from Mühlberger, A., Pauli, P., 2011. Virtuelle Realität in der Psychotherapie. (Virtual reality in psychotherapy). PiD - Psychotherapie im Dialog 12 (2), 143–147 with permission.)

Box 3 Example case: VRET for tunnel phobia

Before therapy, Ms. W. suffered from intense fear of tunnels. The treatment consisted of four sessions over the course of 6 weeks. After cognitive preparation, Ms. W. was exposed to repeated tunnel drives in VR. During the first session, Ms. W. reported that she experienced strong anxiety and found herself unable to look at anything except the road straight ahead of her. Her breathing was shallow, her palms were sweaty, and her heart rate was accelerated. In the second session, the therapist started to slow down the drive through the tunnel, and introduced breaks in the tunnel. Ms. W. was encouraged to try behaviors she had long avoided, like looking around in the tunnel. At first, this elicited intense fear, but this anxiety subsided gradually. By the fourth session, the drive through the tunnel no longer triggered anxiety, breathing had normalized, and Ms. W. was able to look around in all directions without experiencing fear. After therapy, Ms. W. was able to drive through real tunnels. She also found herself to be much more relaxed handling her own anxiety in different situations. No return of fear while driving through road tunnel was reported even 2 years after treatment.
(Adapted from Mühlberger, A., Pauli, P., 2011. Virtuelle Realität in der Psychotherapie. (Virtual reality in psychotherapy). PiD - Psychotherapie im Dialog 12 (2), 143–147 with permission.)

exposure. First, VRET allows exposing patients to as many flights as necessary at little cost. Further, the most feared aspects of flying for aviophobic patients are usually takeoff and turbulence. In VRET, these specific flight phases can be manipulated and repeated many times. For example, a motion platform to effectively simulate turbulence is used, which is found to have a marked effect on self-reported fear and physiological indicators of arousal (Mühlberger et al., 2005)(see Figure 3). A simpler and cost-effective technological solution that nonetheless has been proven to be effective is the use of subwoofers to simulate airplane movements and provide haptic stimulation. Overall, numerous studies have investigated the efficacy of VRET for aviophobia, with very good results (for a review, see da Costa et al., 2008). Wiederhold and Wiederhold (2003) reported on stable effects of VRET for aviophobia at follow-up after 3 years.

Figure 3 Flight simulator as used at the University of Würzburg, Germany. Reproduced from Mühlberger, A., 2014. Virtuelle Realität in der Klinischen Emotions- und Psychotherapieforschung. In: Jeschke, S., Kobbelt, L., Dröege, A. (Eds.), Exploring Virtuality, Springer, Berlin/Heidelberg.

VRET for Acrophobia

One of the earliest and most impressive studies on the efficacy of VRET was conducted by Emmelkamp et al. (2002) for acrophobia. Emmelkamp et al. (2002) directly compared *in vivo* exposure with VRET and therefore reproduced the actual *in vivo* exposure settings in the VR environment (e.g., looking down from a gallery in a shopping mall). The randomized-controlled study revealed VRET to be equally effective as *in vivo* exposure. Importantly, comparable treatment effects were found not only for ratings of fear, but also in a posttreatment behavior avoidance test conducted *in vivo*. Thus, VRET effects were successfully transferred to the real world. Treatment effects were further found to be stable after 6 months in both the groups. Since then, many studies have confirmed the efficacy of VRET for acrophobia (Meyerbröker and Emmelkamp, 2010). Unfortunately, direct comparisons between VRET and *in vivo* exposure are still rare. Figure 4 depicts VR scenarios used for the treatment of specific phobias.

VRET for Phobia of Small Animals

A number of studies have investigated VRET for arachnophobia, with positive results (e.g., Côté and Bouchard, 2005). Figure 5 shows the reduction of fear within and between sessions of VRET for arachnophobia (Shiban et al., 2013). An interesting development of VRET that has been applied to the phobia of small animals is augmented reality. An augmented reality system makes use of a transparent HMD to project images of feared objects onto the real-life environment. This way, patients may encounter and may be exposed to virtual spiders, cockroaches, or other small animals superimposed on the real environment, e.g., the therapy room, while they can see their own body in this environment. Thus, this technology also allows seeing the feared animals crawling across the own body (Juan and Calatrava, 2011). Arachnophobia is also frequently used as

Figure 4 VR scenarios for the treatment of phobias. Reproduced from Mühlberger, A., 2014. Virtuelle Realität in der Klinischen Emotions- und Psychotherapieforschung. In: Jeschke, S., Kobbelt, L., Dröege, A. (Eds.), Exploring Virtuality, Springer, Berlin/Heidelberg.

a paradigmatic anxiety disorder in VR research aimed at dismantling processes of anxiety and anxiety reduction (e.g., Diemer et al., 2013; Shiban et al., 2013; see below).

VRET for Other Specific Phobias

Other specific phobias that have been treated with VRET include fear of driving and fear of tunnels. The specific advantages of VRET for these phobias include safety and convenience. As is the case in aviophobia, the phobic situation may be prolonged and repeated as many times as needed. **Figures 2** and 4 present VR environments used in VRET for driving and tunnel phobia.

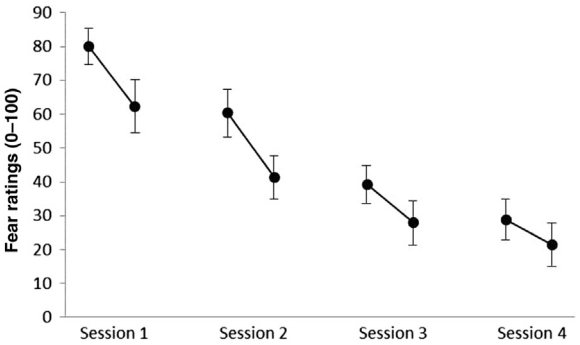

Figure 5 Fear reduction during VRET for arachnophobia. Subjective fear ratings at the start and end of each of four sessions are shown.

VRET for Other Anxiety Disorders

VRET for Posttraumatic Stress Disorder

Another important application of VRET is the treatment of PTSD. Several studies have been conducted, especially for combat-related PTSD and PTSD in survivors of terrorist attacks. Military personnel are at an increased risk for developing PTSD and related disorders. One of the first studies on VRET for combat-related PTSD was provided by Rothbaum et al. (2001), with positive results. In recent years, the US military has begun a series of investigations in an effort to establish and provide appropriate treatment. This effort comprises studies on the efficacy of VRET for combat-related PTSD (Rizzo et al., 2011). Virtual scenarios replicating Iraqi and Afghan environments have been developed and tested. These scenarios include a range of visual, auditory, haptic (vibrations), and olfactory stimuli that can be employed in an individualized treatment protocol (see Cukor et al., 2009). In 2010, McLay et al. reported the first use of VRET in a military camp in the actual conflict area where soldiers were engaged in combat. Future perspectives in this area include studies on the usefulness of VR as a diagnostic tool that might help to identify individuals at an increased risk for PTSD, in order to improve assessment of fitness for duty, i.e., resilience (Rizzo et al., 2011). In the civilian field, VRET has been used to treat survivors of terrorist attacks (Difede et al., 2007) and motor vehicle accidents (Beck et al., 2007).

VRET for Social Anxiety Disorder

Social anxiety disorder is characterized by fear and apprehension of situations involving social interaction and/or social performance. Since social interactions are still difficult to create in a meaningful and believable way in VR, most VRET studies on social anxiety disorder so far have focused on the fear of public speaking, which is very common in patients suffering from social anxiety disorder. First results of VRET for fear of public speaking revealed positive outcomes (see Pull, 2012). Researchers have also found that patients respond differentially to virtual audiences, depending on the behavior of the agents in the audience (friendly vs hostile) (Pertaub et al., 2001). The development of more realistic computer-generated simulations of human

behavior (artificial intelligence) promises sophisticated future applications. Already, investigators have successfully simulated complex social interactions like small talk (Pan et al., 2012). The application of simulated humans in different forms of social interactions for exposing and training patients suffering from social anxiety disorder might be the next development in this field.

VRET for Panic Disorder and Agoraphobia

Another anxiety disorder that is increasingly treated with VRET is panic disorder with agoraphobia. Like individuals suffering from social anxiety disorder, agoraphobia patients fear various and many rather complex situations. Nonetheless, some common agoraphobic situations, like public transport, shopping centers, and tunnels have been realized in VR to treat agoraphobia patients successfully (Botella et al., 2007). With regard to panic symptoms, first results indicate that symptom simulation in VR (e.g., blurred vision) might be a promising addition to VRET for panic disorder with or without agoraphobia (Botella et al., 2007).

VRET for Obsessive–Compulsive Disorder and Generalized Anxiety Disorder

Little research exists on VRET for generalized anxiety disorder and obsessive–compulsive disorder, most likely because the feared stimuli in these disorders are very individual and because cognitive processes might play a greater role than in other anxiety disorders. Nonetheless, VR applications for these disorders have been tested (Gorini et al., 2010; Kim et al., 2009) and shown positive results. However, more research in these areas will show up soon.

Efficacy of VRET

Until 2012, four meta-analyses on the efficacy of VRET for the anxiety disorders had been published (Gregg and Tarrier, 2007; Powers and Emmelkamp, 2008; Parsons and Rizzo, 2008; Opris et al., 2012). In general, the results are promising, with large effect sizes. For example, Opris et al. (2012) report an average weighted effect size (across the anxiety disorders) of $D = 1.12$ for the comparison with wait-list control groups, while no significant differences were found between cognitive-behavioral therapy with *in vivo* exposure and VRET. However, these results are still based on a limited number of studies, with just $n = 21$ VRET studies reviewed by Opris et al. (2012). Of these, only 11 studies carried out a direct comparison of VRET with *in vivo* exposure (Opris et al., 2012). Looking at specific anxiety disorders, most research is available on aviophobia and acrophobia. For these disorders, VRET may already be considered an empirically justified treatment (Meyerbröker and Emmelkamp, 2010). More efficacy studies, especially randomized-controlled trials that include *in vivo* or imaginal exposure as control treatments, are necessary to test the efficacy of VRET for other anxiety disorders and in comparison to other interventions.

VR Applications for Other Mental Disorders

Apart from exposure treatment, VR scenarios have been developed and examined for a range of applications. One is the use of VR worlds because of their distracting potential for pain management during painful medical procedures. Hoffman et al. (2011) report pain reductions of 35–50% when patients treated for burn injuries engaged during treatment in an interactive VR game (medication + VR) compared to a medication only condition. Comparable effects were observed in patients experiencing severe pain. VR scenarios are also tested as a tool to enhance (social) learning in autistic children. First results are positive (for a review, see Beelani et al., 2011), but more research is needed. Other applications in psychotherapy are the use of VR in eating disorders, substance-related disorders, and schizophrenia.

Therapeutic Uses of VR for Eating Disorders

Body dissatisfaction, in combination with a distorted body image, is a central feature of eating disorders. VR scenarios have been created that allow patients to adapt a three-dimensional virtual body to their body image, and to compare this to their actual and desired body shapes, helping patients to realize and change the distorted nature of their own body image (Ferrer-García and Gutiérrez-Maldonado, 2012). First studies have used these VR body image scenarios in conjunction with cognitive therapy for the treatment of eating disorders, with positive results especially on measures of body satisfaction (Ferrer-García and Gutiérrez-Maldonado, 2012). Other VR scenarios for eating disorders including emotionally significant environments related to food and body image are being developed; however, little research has evaluated this approach so far.

VR Applications in Substance-Related Disorders

The main VR application for substance-related disorders is the use of VR scenarios for cue exposure. VR environments equipped with substance-related cues have been found to elicit craving in alcohol-dependent patients (Lee et al., 2009), and in users of cocaine (Saladin et al., 2006). Repeated cue exposure in VR has been reported to reduce craving in alcohol-dependent patients (Lee et al., 2009). In an ingenious, randomized-controlled study ($n = 91$), Girardi et al. (2009) found that crushing cigarettes in VR reduced nicotine addiction as assessed with a questionnaire and improved rates of abstinence and adherence to a treatment program that was delivered alongside the study. These results indicate a possible therapeutic application of VR for habit reversal.

VR-Assisted Therapy for Schizophrenia

Researchers have highlighted a number of different therapeutic VR applications for patients suffering from schizophrenia. Deficits in social interaction are central to schizophrenia and require special attention from researchers and clinicians, as the social withdrawal associated with negative symptoms of schizophrenia accounts for much of the social and occupational impairments experienced by patients. In 2011, Park et al.

published promising results of a randomized-controlled trial ($n = 91$) comparing VR-assisted vs traditional social skills training. The authors found VR role plays to be associated with greater motivation and better outcome in terms of assertiveness and conversational skills, while the *in vivo* role playing group showed better outcome on one measure only (nonverbal communication). Surely these results warrant replication and further study. Another area that has been explored with VR is medication management (Kurtz et al., 2007). Freeman et al. (2008) suggest further applications of VR as a part of a cognitive treatment program including psychoeducation, behavioral experiments, and coping with psychotic symptoms.

VR as a Research Tool

VR enables researchers to put patients into complex, yet standardized scenarios. Consequently, VR has numerous and varied applications not only as a research tool in clinical psychology, especially in therapy research, but also in assessment and basic research on the processes underlying different psychopathologies. In most of these areas, research has just begun, but we expect this area to expand considerably in the near future.

VR for Research on Therapy Processes

VR brings a form of exposure therapy into the laboratory that is far more suitable than the traditional forms of exposure for the assessment of the processes that make exposure-based psychotherapy work. For example, several studies have investigated psychophysiological parameters of treatment processes during VRET. Diemer et al. (submitted for publication) recently reviewed this field of research and found clear indications that VR exposure is accompanied by marked psychophysiological arousal, while psychophysiological habituation processes during VRET are still poorly understood. This seems to be related to the great heterogeneity of study designs used in VRET research, and the still low numbers of studies on therapy processes during VRET. In a study exploring context effects of exposure therapy, VR scenarios were used to demonstrate that exposure in multiple contexts improves the generalization of VRET treatment effects to the real word, presumably by reducing the renewal of fear (Shiban et al., 2013).

An important aspect of VR applications in psychotherapy is the concept of presence. Presence is usually defined following Slater and Wilbur (1997) as "the sense of being in the virtual environment" (see also Schubert et al., 2001). Presence is supposed to be the mediator that allows an artificial VR environment to engage real emotions (Parsons and Rizzo, 2008). As Schubert et al. (2001) have shown, presence is in fact a multifaceted construct that can be divided into three aspects: spatial presence, i.e., 'the sense that we are located in and act from within the virtual environment'; involvement, i.e., 'the attention dedicated to the virtual environment'; and realness, i.e., 'the perceived verisimilitude of the virtual environment.' Research on the significance of presence and its facets for VR therapy is underway. As Price et al. (2012) could show in a VRET study of social phobia, different components of presence may be related to different aspects of treatment.

VR for Assessment

As VR allows a direct, continuous, and unobtrusive observation of symptoms and problematic behavior (e.g., phobic avoidance), VR scenarios provide a tool for improving diagnostic accuracy. For example, Mühlberger et al. (2007) have shown that physiological responses in a VR tunnel scenario reliably differentiate tunnel-fearful from nonfearful participants. Similarly, Sorkin et al. (2006) demonstrated that performance in a VR maze differentiated patients with schizophrenia from healthy controls. Freeman et al. (2008) review a number of other VR applications for schizophrenia, including the assessment of paranoid thinking in a social situation. Other investigators have used VR scenarios to assess emotion recognition deficits in schizophrenia (Dyck et al., 2010). For attention-deficit hyperactivity disorder (ADHD), the work group of Rizzo developed a VR classroom. An attention test delivered in this VR scenario, which incorporates typical distractors, was shown to differentiate better between children affected by ADHD and controls than standard measures (Adams et al., 2009). One step further is the use of VR to monitor and gauge treatment effects. This was done by Pollack et al. (2010), who measured the effects of methylphenidate on the performance of children with ADHD in a continuous performance task in a VR classroom simulation.

VR Research on Psychopathological Processes

Adding a level of complexity to laboratory research that was previously impossible, VR is increasingly used to investigate psychopathological processes in different disorders. The research program outlined by Freeman et al. (2008) on the use of VR scenarios to identify correlates and predictors of psychotic symptoms is one approach that falls into this category. Already, VR has been used in conjunction with electroencephalography (EEG) (Lee et al., 2009) to discover neurophysiological correlates of craving in alcohol dependence. In basic research on fear and anxiety, a VR paradigm that is increasingly used combines fear conditioning with complex VR scenarios, allowing the assessment of context-conditioning and related variables. For example, Glotzbach-Schoon et al. (2013) assessing context-conditioning in a VR paradigm comparing the effects of predictable and unpredictable aversive events in participants high vs low in trait anxiety. While both groups showed contextual fear conditioning, high trait anxious subjects were found to associate the context with the aversive events especially rapidly. This research may help shed light on the processes underlying the acquisition of dysfunctional anxiety.

Further, anxiety-provoking VR scenarios can be used as a nonpharmacological challenge procedure to elicit and investigate fear in the laboratory in a controlled manner. Recently, this approach was used to test the anxiolytic potential of the atypical antipsychotic quetiapine in a randomized, placebo-controlled trial in a sample of $n = 58$ patients suffering from arachnophobia (Diemer et al., 2013). The laboratory

setting allowed the assessment of fear and anxiety in a multimodal manner, and to detect domain-specific drug effects.

One of the growing research areas in clinical psychology and psychiatry is the identification and investigation of endophenotypes. Endophenotypes are conceptualized as genetically determined traits that confer a risk of psychopathology. They are supposed to bridge the gap between genetics and diagnoses, representing underlying pathomechanisms that may be shared by various diagnoses. One such endophenotype is the fear-potentiated startle response, which has been linked to increased vulnerability for anxiety disorders. Startle paradigms can be easily combined with VR scenarios, allowing the assessment of endophenotypes in an ecologically valid environment (cf Glotzbach-Schoon et al., 2013). Many other applications of complex VR environments to investigate possible endophenotypes – which may be cognitive, psychological, or psychophysiological in nature – of different disorders are possible.

Future Perspectives in VR Therapy and Research

The last decades were associated with a continuous expansion of VR applications and research. Computer technology and software development are advancing at a fast pace, making a widespread use of VR technology in mental health areas likely. As for VRET, a particular increase of VR applications for the treatment of aviophobia and PTSD is expected, given the obvious advantages of using VR scenarios for these disorders. Particularly, augmented reality has the potential to further boost VR applications in research and health care.

In general, the development of VRET research in the last decade shows two trends: the expansion of the different applications, and the dissemination of VRET, i.e., the increase of its availability in the health care system. In terms of diagnoses, more research into VRET applications for the anxiety disorders, especially social phobia, PTSD, and panic disorder and agoraphobia is to be expected. Further, research into VR applications beyond anxiety is likely. In terms of availability, feasibility studies on the introduction of VRET into routine health care will be the next step and are bound to be conducted in this decade.

As for VR research, the possibilities provided by VR to investigate psychotherapeutic processes are only beginning to be used. Possible variables of interest during exposure-based treatments include changes in psychophysiological reactivity during exposure, homework assignments, and effects of additional treatment elements (e.g., attention focusing). The emerging applications of VR in basic research are numerous (see above). They are supplemented by technological possibilities that include the combination of VR scenarios with different assessment methods, among them physiological recordings, eye tracking, EEG, and neuroimaging techniques.

See also: Behavior Therapy: Background, Basic Principles, and Early History; Diagnostic Processes in Clinical Psychology and Psychotherapy; Evidence-Based Psychotherapy; Experiential Psychotherapy; Panic Disorder and Agoraphobia Across The lifespan; Phobias Across the Lifespan; Psychological Treatment, Effectiveness of; Psychological Treatments Delivered Nontraditionally; Psychological Treatments, Empirically Supported; Psychological Treatments: Randomized Controlled Trials; Psychotherapy Integration; Psychotherapy Process Research; Social Phobia across the Lifespan.

Bibliography

Adams, R., Finn, P., Moes, E., Flannery, K., Rizzo, A.S., 2009. Distractibility in attention/deficit/hyperactivity disorder (ADHD): the virtual reality classroom. Child Neuropsychology 15 (2), 120–135.

Beck, J.G., Palyo, S.A., Winer, E.H., Schwagler, B.E., Ang, E.J., 2007. Virtual reality exposure therapy for PTSD symptoms after a road accident: an uncontrolled case series. Behavior Therapy 38, 39–48.

Beelani, M., Fornasari, L., Chittaro, L., Brambilla, P., 2011. Virtual reality in autism: state of the art. Epidemiology and Psychiatric Sciences 20, 235–238.

Botella, C., García-Palacios, A., Villa, H., Baños, R.M., Quero, S., Alcañiz, M., Riva, G., 2007. Virtual reality exposure in the treatment of panic disorder and agoraphobia: a controlled study. Clinical Psychology & Psychotherapy 14 (3), 164–175.

Bouchard, S., Côté, S., Richard, D.C.S., 2007. Virtual reality applications for exposure. In: Richard, D.C.S., Lauterbach, D. (Eds.), Handbook of Exposure Therapies. Elsevier Academic Press, Amsterdam, pp. 347–388.

Choy, Y., Fyer, A.J., Lipsitz, J.D., 2007. Treatment of specific phobia in adults. Clinical Psychology Review 27 (3), 266–286.

Côté, S., Bouchard, S., 2005. Documenting the efficacy of virtual reality exposure with psychophysiological and information processing measures. Applied Psychophysiology and Biofeedback 30 (3), 217–323.

Cukor, J., Spitalnick, J., Difede, J., Rizzo, A., Rothbaum, B.O., 2009. Emerging treatments for PTSD. Clinical Psychology Review 29, 715–726.

da Costa, R.T., Sardinha, A., Nardi, A.E., 2008. Virtual reality exposure in the treatment of fear of flying. Aviation, Space, and Environmental Medicine 79 (9), 899–907.

Diemer, J., Domschke, K., Mühlberger, A., Winter, B., Zavorotnyy, M., Notzon, S., Silling, K., Arolt, V., Zwanzger, P., 2013. Acute anxiolytic effects of quetiapine during virtual reality exposure – a double-blind placebo-controlled trial in patients with specific phobia. European Neuropsychopharmacology 23 (11), 1551–1560.

Diemer, J., Mühlberger, A., Pauli, P., Zwanzger, P. Virtual reality exposure in anxiety disorders: impact on psychophysiological reactivity, submitted for publication.

Difede, J., Malta, L.S., Best, S., Henn-Haase, C., Metzler, T., Bryant, R., Marmar, C., 2007. A randomized controlled clinical treatment trial for World Trade Center attack-related PTSD in disaster workers. The Journal of Nervous and Mental Disease 195 (10), 861–865.

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Mathiak, K., 2010. Virtual faces as a tool to study emotion recognition deficits in schizophrenia. Psychiatry Research 179, 247–252.

Emmelkamp, P.M.G., Krijn, M., Hulsbosch, A.M., de Vries, S., Schuemie, M.J., van der Mast, C.A.P.G., 2002. Virtual reality treatment versus exposure in vivo: a comparative evaluation in acrophobia. Behaviour Research and Therapy 40 (5), 509–516.

Ferrer-García, M., Gutiérrez-Maldonado, J., 2012. The use of virtual reality in the study, assessment, and treatment of body image in eating disorders and nonclinical samples: a review of the literature. Body Image 9, 1–11.

Foa, E.B., Kozak, M.J., 1986. Emotional processing of fear: exposure to corrective information. Psychological Bulletin 99 (1), 20–35.

Freeman, D., Gittins, M., Pugh, K., Antley, A., Slater, M., Dunn, G., 2008. What makes one person paranoid and another person anxious? The differential prediction of social anxiety and persecutory ideation in an experimental situation. Psychological Medicine 38, 1121–1132.

García-Palacios, A., Botella, C., Hoffman, H., Fabregat, S., 2007. Comparing acceptance and refusal rates of virtual reality exposure vs. in vivo exposure by patients with specific phobias. Cyberpsychology & Behavior 10 (5), 722–724.

Girard, B., Turcotte, V., Bouchard, S., Girard, B., 2009. Crushing virtual cigarettes reduces tobacco addiction and treatment discontinuation. Cyberpsychology and Behavior 12 (5), 477–483.

Glotzbach-Schoon, E., Tadda, R., Andreatta, M., Tröger, C., Ewald, H., Grillon, C., Pauli, P., Mühlberger, A., 2013. Enhanced discrimination between threatening and safe contexts in high-anxious individuals. Biological Psychology 93 (1), 159–166.

Gorini, A., Pallavicini, F., Algeri, D., Repetto, C., Gaggioli, A., Riva, G., 2010. Virtual reality in the treatment of generalized anxiety disorders. Annual Review of Cybertherapy and Telemedicine 8, 31–35.
Gregg, L., Tarrier, N., 2007. Virtual reality in mental health: a review of the literature. Social Psychiatry and Psychiatric Epidemiology 42 (5), 343–354.
Hoffman, H.G., Chambers, G.T., Meyer, W.J., Arceneaux, L.L., Russell, W.J., Seibel, E.J., Richards, T.L., Sharar, S.M., Patterson, D.R., 2011. Virtual reality as an adjunctive nonpharmacologic analgesic for acute burn pain during medical procedures. Annals of Behavioral Medicine 41, 183–191.
Juan, M.C., Calatrava, J., 2011. An augmented reality system for the treatment of phobia to small animals viewed via an optical see-through HMD: comparison with a similar system viewed via a video see-through HMD. International Journal of Human-Computer Interaction 27 (5), 436–449.
Kim, K., Kim, C.H., Kim, S.Y., Roh, D., Kim, S.I., 2009. Virtual reality for obsessive-compulsive disorder: past and the future. Psychiatry Investigation 6 (3), 115–121.
Kurtz, M.M., Baker, E., Pearlson, G.D., Astur, R.S., 2007. A virtual reality apartment as a measure of medication management skills in patients with schizophrenia: a pilot study. Schizophrenia Bulletin 33 (5), 1162–1170.
Lee, S.H., Han, D.H., Oh, S., Lyoo, K., Lee, Y.S., Renshaw, P.F., Lukas, S.E., 2009. Quantitative electroencephalographic (qEEG) correlates of craving during virtual reality therapy in alcohol-dependent patients. Pharmacology, Biochemistry and Behavior 91, 393–397.
McLay, R.N., McBrien, C., Wiederhold, M.D., Wiederhold, B.K., 2010. Exposure therapy with and without virtual reality to treat PTSD while in the combat theater: a parallel case series. Cyberpsychology, Behavior, and Social Networking 13 (1), 37–42.
Meyerbröker, K., Emmelkamp, P.M., 2010. Virtual reality exposure therapy in anxiety disorders: a systematic review of process-and-outcome studies. Depression and Anxiety 27 (10), 933–944.
Mühlberger, A., Bülthoff, H.H., Wiedemann, G., Pauli, P., 2007. Virtual reality for the psychophysiological assessment of phobic fear: responses during virtual tunnel driving. Psychological Assessment 19 (3), 340–346.
Mühlberger, A., Pauli, P., 2011. Virtuelle Realität in der Psychotherapie. (Virtual reality in psychotherapy). PiD - Psychotherapie im Dialog 12 (2), 143–147.
Mühlberger, A., Wiedemann, G., Pauli, P., 2005. Subjective and physiologic reactions of flight phobics during VR exposure and treatment outcome: what adds motion simulation? Annual Review of CyberTherapy and Telemedicine 3, 185–192.
Opris, D., Pintea, S., Garcia-Palacios, A., Botella, C., Szamoskozi, S., David, D., 2012. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. Depression and Anxiety 29 (2), 85–93.
Pan, X., Gillies, M., Barker, C., Clark, D., Slater, M., 2012. Socially anxious men interact with a forward virtual woman: an experimental study. PLoS One 7 (4), e32931.
Park, K.-M., Ku, J., Choi, S.-H., Jang, H.-J., Park, J.-Y., Kim, S.I., Kim, J.-J., 2011. A virtual reality application in role-plays of social skills training for schizophrenia: a randomized, controlled trial. Psychiatry Research 189, 166–172.
Parsons, T.D., Rizzo, A.A., 2008. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: a meta-analysis. Journal of Behavior Therapy and Experimental Psychiatry 39 (3), 250–261.
Pertaub, D.P., Slater, M., Barker, C., 2001. An experiment on fear of public speaking in virtual reality. Studies in Health Technology and Informatics 81, 372–378.
Pollack, Y., Shomaly, H.B., Weiss, P.L., Rizzo, A.A., Gross-Tsur, V., 2010. Methylphenidate effect in children with ADHD can be measured by an ecologically valid continuous performance test embedded in virtual reality. CNS Spectrum 15 (2), 125–130.
Powers, M.B., Emmelkamp, P.M., 2008. Virtual reality exposure therapy for anxiety disorders: a meta-analysis. Journal of Anxiety Disorders 22 (3), 561–569.
Price, M., Mehta, N., Tone, E.B., Anderson, P.L., 2012. Does engagement with exposure yield better outcomes? Components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia. Journal of Anxiety Disorders 25 (6), 763–770.
Pull, C.B., 2012. Current status of knowledge on public-speaking anxiety. Current Opinion in Psychiatry 25 (1), 32–38.
Rothbaum, B.O., Hodges, L.F., Ready, D., Graap, K., Alarcon, R.D., 2001. Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. The Journal of Clinical Psychiatry 62 (8), 617–622.
Rizzo, A., Parsons, T.D., Lange, B., Kenny, P., Buckwalter, J.G., Rothbaum, B., Difede, J., Frazier, J., Newman, B., Williams, J., Reger, G., 2011. Virtual reality goes to war: a brief review of the future of military behavioral healthcare. Journal of Clinical Psychology in Medical Settings 18, 176–187.
Saladin, M.E., Brady, K.T., Graap, K., Rothbaum, B.O., 2006. A preliminary report on the use of virtual reality technology to elicit craving and cue reactivity in cocaine dependent individuals. Addictive Behaviors 31 (10), 1881–1894.
Schubert, T., Friedmann, F., Regenbrecht, H., 2001. The experience of presence: factor analytic insights. Presence: Teleoperators and Virtual Environments 10 (3), 266–281.
Shiban, Y., Pauli, P., Mühlberger, A., 2013. Effect of multiple context exposure on renewal in spider phobia. Behaviour Research and Therapy 51 (2), 68–74.
Slater, M., Wilbur, S., 1997. A framework for immersive virtual environments (FIVE). Presence: Teleoperators and Virtual Environments 6 (6), 603.
Sorkin, A., Weinshall, D., Modai, I., Peled, A., 2006. Improving the accuracy of the diagnosis schizophrenia by means of virtual reality. American Journal of Psychiatry 163, 512–520.
Sutherland, I.E., 1968. A head-mounted three-dimensional display. Proceedings of the American Federation of Information Processing Societies 68, 757–764.
Wiederhold, B.K., Wiederhold, M.D., 2003. Three-year follow-up for virtual reality exposure for fear of flying. CyberPsychology & Behavior 6 (4), 441–445.

Relevant Websites

www.cybersession.info – Virtual Reality Therapy Inferface "CyberSession".
www.cybertherapy.info – The Virtual Reality and Ambient Intelligence Reference Site for Medicine, Psychology and Neuros.
www.hitl.washington.edu/home – Human Interface Technology Laboratory. University of Washington.
ict.usc.edu – USC Institute for Creative Technologies.
www.interactivemediainstitute.com – Interactive Media Institute.
www.labpsitec.uji.es/eng – Laboratory of Psychology and Technology at Universitat Jaume.
www.psychiatry.emory.edu/programs/tarp – EMORY University School of Medicine, Trauma and Anxiety Recovery Program.
www.recveb.ucsb.edu – Research Center for Virtual Environments & Behavior.
www.virtuallybetter.com – Virtually Better, Inc.
www.vtplus.eu – Virtual Therapy + Research Systems.
w3.uqo.ca/cyberpsy – Laboratoire de Cyberpsychologies de l'UQO.

Virtue Ethics

Christine Swanton, University of Auckland, Auckland, New Zealand

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Virtue ethics is a type of normative ethical theory generally seen as opposed to both Kantian ethics and consequentialism. On orthodox conceptions, virtue ethics has been regarded as Aristotelian, but virtue ethics is now with greater frequency seen as belonging to a broader church including Confucian ethics, Hume, and even Nietzsche. Common to all virtue ethics is the centrality of character (understood in various ways) to ethics, although in applied ethics, the application of virtue concepts to action is a rich area of debate.

Virtue ethics is a type of normative ethical theory generally seen as opposed to both Kantian ethics and consequentialism. In its modern incarnations, it has been largely inspired by the eudaemonistic tradition of the ancient Greeks, and in particular Aristotle. However, we can expect further developments in virtue ethics that are inspired by other sources, such as Nietzsche and Hume. Virtue ethics is characterized by the centrality but not necessarily the primacy of character in normative ethical theory, where norms of excellence of character provide norms for both the good life in general and acting well and rightly. In virtue ethical traditions, norms of excellence of character are rich, and involve being well disposed in regard to reason, feelings, desires, motives, emotions, as well as action; hence the importance of the expression of fine inner states in virtue ethics. Problems for virtue ethics include the foundations of virtue, the universality of virtue, the relation between character and issues in applied ethics such as role obligations and right action generally, demands for codifiability and determinacy, and 'character skepticism' suggested by situationist psychology.

Definition

It has often been said that the development of virtue ethics (in modern dress) is in its infancy. For this reason, although virtue ethics is a type of substantive normative ethical theory opposed to consequentialism and Kantian ethics, its definition is contested (Swanton, 2013). The problem of definition is compounded by the fact that although much contemporary virtue ethics is neo-Aristotelian, it need not be, and has not always been. Increasingly, David Hume is being read in a way more sympathetic to virtue ethics, Nietzschean and neo-Confucian virtue ethics are developing, and Plato and Martineau have inspired 'agent-based' virtue ethics (Slote, 1996, 2001). Virtue ethicists, then, need not, and do not agree on such central questions as the definition of right action and what makes a trait of character a virtue. Common to virtue ethicists, however, is the view that virtue notions pertaining to character and action rather than duty, rules, or consequences is central to ethics. This is not to say that rules, consequences, and the motive of duty are not important, but virtue notions will produce virtue-centered conceptions of moral rightness, objectivity in ethics, impartiality, demandingness or stringency of ethics, and practice, which reconfigure the moral universe.

The major source of inspiration for the development of modern virtue ethics is Aristotle (1955), particularly his *Nicomachean Ethics*. Two central features of his view have influenced definitions of modern virtue ethics. The first is his idea that the judgment of a virtuous agent is the rule and the measure of the right. This feature has spawned the view that virtue ethics is opposed to moral rules and principles and that these must be entirely replaced by the judgment of a qualified judge (a virtuous agent). However, there are 'virtue rules' such as requirements to be kind, considerate, fair, courageous, and so on: rules that we use to educate our children into the moral life, and that we may use in morally charged situations where we might ask, for example, 'What would be the kind or tactful thing to do here?' Virtue ethicists, however, follow Anscombe (1958) in believing that moral rules should be couched in the 'thick' virtue concepts.

The second feature sometimes thought to be defining of virtue ethics is Aristotle's view that virtues are essentially connected with the flourishing of their possessor. David Hume, however, thought that there are four potential grounds for status as a virtue: agreeableness and usefulness to self and agreeableness and usefulness to others. What constitutes right action and what gives status as a virtue are points of disagreement among virtue ethicists, so particular conceptions of right action and virtue status should not be employed to define virtue ethics as such.

Content of Virtue Ethics

Discussion of the content of virtue ethics can center usefully on three issues: (1) virtue ethical conceptions of moral rightness of action, (2) the nature of virtue in virtue ethics, and (3) different conceptions of what makes a trait of character a virtue, within virtue ethics.

Virtue Ethical Conceptions of the Right

It is often thought that the virtue ethical conception of rightness, indeed one that is defining of virtue ethics is the following. (1) An act is right if and only if it is what a virtuous

agent would characteristically (i.e., acting in character) do in the circumstances (Hursthouse, 1996, 1999). However, other possibilities within virtue ethics exist, and have been canvassed. According to 'agent-based' virtue ethics, rightness is understood thus. (2) An act is right if and only if it exhibits good motivation (Slote, 1996, 2001). Statman (1997) alludes to another possibility. (3) An act is right if and only if it is expressive of virtuous character.

Finally, we might say (4) an act is right if and only if it hits the targets of relevant and contextually salient virtues in relation to action (Swanton, 2003, 2013). Concept (4) differs from (1) in that it is possible for a virtuous agent acting in character to do in the circumstances something that is wrong according to (4). For example, according to (4), an act expressive of a benevolent disposition, and benevolently motivated, may fail to be a benevolent act. For a virtuous agent is not omniscient, although she is (on orthodox views) endowed with practical wisdom and the act may turn out to be extremely harmful. For the same reason (4) differs from (2) and (3). Furthermore, virtuous acts, as opposed to actions from or expressive of a state of virtue, need not possess all the features required for action from a virtuous state (as Aristotle recognized). In certain contexts, they may not express fine character or even good motives. In short, a virtue ethicist may share the common conception of right action as one that is successful in hitting its target. Where virtue ethicists who share this view differ from, for example, consequentialists, is their conception of 'hitting the target.' The target is set not by consequentialist criteria of promoting the most good, but by the ends of the various virtues. These ends will of course differ from virtue to virtue, so unless one subscribes to a strong doctrine of the unity of the virtues, virtue ethicists will possess a pluralistic view of rightness of action.

Where virtue ethicists do not always require fine inner states for rightness (as opposed to goodness) of actions, they may distinguish between rightness and excellence of acts, where the criteria for the latter are more stringent than those for the former.

The Nature of Virtue in Virtue Ethics

A virtue is an excellence of character. It is a disposition of acknowledging or responding to items in the field of the virtue in an excellent way. According to virtue ethical conceptions of virtue, the standards of excellence point to states inside the agent, as well as to the world outside the agent. In short, for a virtue ethicist, the realization of the end of a virtue must express fine inner states, if that realization is to display virtue.

Virtues have a specialized focus, that is, they have a field. The field of a virtue may be human beings in general; one's children, property, friends, honors, and pleasure; the natural world, cultural and artistic objects, money, abstract objects such as knowledge, and states of affairs such as dangerous situations, interpersonal conflict; and so on. There are a wide variety of virtues that focus on different fields or domains – virtues such as friendship, courage, justice, environmental virtues, appropriate assertiveness, honesty, politeness and tact, liberality, generosity, as well as thrift. Some virtues are universal such as courage, temperance, and justice, while others are relative to roles, such as professional virtues and virtues of parenthood (call these role-differentiated virtues (Swanton, 2007a)). Yet others are relative to situation, for example, according to Aristotle, virtues whose field concerns (the expenditure of) money vary according to how much money one has, and range from thrift to 'magnificence.'

Insofar as 'virtue' is a success word, the form of acknowledgment of items in the field of a virtue required by a virtue must be appropriate or successful. What constitutes success or appropriateness will vary according to two types of things: first, according to the different kinds of morally significant feature in those items which it is the end of the virtue to recognize, sustain, further, or bring about (such as their value or status), and second, according to the differing forms or modes of moral acknowledgment of items required by the various virtues (such as promoting (value or good) or respecting individuals in ways befitting their status). Virtue ethicists will differ on what are the basic morally significant features in the items acknowledged and what are the basic forms of moral acknowledgment inherent in virtue. The former may include value, status, bonds, and good; the latter promotion, honoring (Pettit, 1997), respect, love, appreciation, and creativity. The important point is that the criteria of success in meeting the various ends of the different virtues will vary, depending on which forms of moral acknowledgment are involved in their exercise. We may ask, for example, whether such virtues as justice, benevolence, and friendship are centrally concerned with all, or some of, the following: the promotion of good, the promotion of value, the honoring of value, respect for individuals in ways befitting their status, or love for individuals in virtue of bonds between those individuals and the agent.

According to virtue ethicists, an exercise or manifestation of virtue requires the successful expression of fine inner states, that is, in displaying virtue, the agent must both have fine inner states and successfully express them in her behavior. For example, virtuous creativity does not merely involve meeting the standards for success in the external manifestation of creativity that confers artistic merit in the created product. The creative response must also be expressive of fine inner states, such as the expression of artistic passion, a desire to create something of value to society or of high artistic quality, practical wisdom in the production of value, originality, and a desire to create something novel, interesting, or original. Similarly, love as a virtuous response or form of moral acknowledgment is not merely the promotion of the good of an individual, or the inculcation of a belief in the beloved that she is loved. Virtuous love is expressive of genuinely caring motivation: motivation that is not pathological, such as passively or morbidly dependent. Again, respect for status as a virtuous response is not merely the formal trappings of respect according to societal convention.

Although virtue ethicists share the view that virtue requires excellence of inner states, what kinds of inner states need to be excellent for virtue to be displayed is a matter of some controversy. Neo-Aristotelians emphasize the necessity of practical wisdom, but this view has been challenged by work on the so-called virtues of ignorance (Driver, 1989), by those who emphasize the sufficiency of admirable motivation for virtue (Slote, 1996, 2001), and by those who emphasize the pivotal role of passion in the life of virtue (Solomon, 1998).

Say, for example, that benevolence as a virtue is concerned centrally with promoting the good of individuals, we still need to know answers to the following questions. Is it enough for benevolence to be displayed that the agents display good motives in their attempt to promote good? Is it necessary that she display practical wisdom? Is it sufficient that she display practical wisdom or is she also required to display knowledge in cases where there is large-scale ignorance concerning good? To what extent does virtue require rising above the general ignorance of a population or culture?

Some progress on the issue of what constitutes excellence in inner states appropriate to human beings may be secured if Nietzsche's emphasis on excellence in depth psychological states is followed up. For Nietzsche, virtuous action is action expressive of fine depth states, and this view enables him to draw a distinction between virtuous and nonvirtuous altruism, virtuous and nonvirtuous asceticism, and so on. Forms of beneficence, sublimation, and self-discipline, which are expressive of self-contempt and resentment, lack virtue for Nietzsche. Post-Nietzschean analytic-oriented psychologists also distinguish virtue and vice on the basis of depth states expressed in habitual behavior. For example, genuine and passive dependent or 'morbid' dependent love is distinguished on that basis by, for example, Horney (1951). Fromm (1977) too, distinguishes between healthy (benign) and unhealthy (malignant) forms of aggression. Rather than aggressiveness being simply characterized as a vice, Fromm's account distinguishes between malignant and benign forms based on the human's characterological orientation to the satisfaction of basic needs such as the needs for excitation or stimulation, and relatedness.

What Makes a Trait of Character a Virtue

On the issue of what makes a trait of character a virtue, virtue ethics can be divided into eudaemonistic and non-eudaemonistic forms.

Eudaemonism

According to eudaemonistic virtue ethics, which characterizes Greek and neo-Aristotelian virtue ethics in its modern form, virtue is at least partially constitutive of the flourishing of the possessor of the virtues. Common to all versions of eudaemonistic virtue ethics is the view that what constitutes the flourishing of the agent cannot be understood entirely independently from the idea of virtue itself. That is, in determining what the virtues are, we cannot simply ask, first, what is good for a person, and then define the virtues in terms of traits that promote that good. For example, according to Aristotle, fineness of character is the greatest good for a person. Pleasure without temperance, wealth without justice, and the possession of friends without the virtue of friendship is not good for an individual, that is, does not benefit an individual. However, unlike the Stoics, Aristotle did not believe that virtue is sufficient for eudaemonia or flourishing: a modicum of good fortune and money, and children who do not disgrace one, are also necessary.

Due to the fact that eudaemonistic virtue ethics has been the dominant form of virtue ethics, it has been the main target of opponents of virtue ethics. The most important and intractable criticisms involve the relation between being good for a human being and being good *qua* human being and the nature of the naturalism that ensues. Other criticisms include the following. Because eudaemonism does not understand flourishing to be characterizable wholly independently of virtue, it has been accused of circularity. However, a coherentist epistemology, which is indeed adopted by Aristotle and many neo-Aristotelians, may remove the sting from this criticism. It does so provided the network of concepts and background theories of human nature in which conceptions of virtue and flourishing are embedded is sufficiently rich, informed by the social sciences, broad ranging, and does not rely on a problematically narrow, prejudiced, and uninformed set of intuitions. Another criticism is the charge of egoism occasioned by the belief that, according to eudaemonism, the rationale for the virtues is their benefiting the possessor of the virtues. This criticism is met by the claim that virtuous action need not be, and is not standardly, motivated by the self-interest of the possessor of the virtue. This reply in turn begets the criticism of indirection, namely, the criticism that a paradoxical and unmotivated dissonance opens up between the rationale for status of a trait as a virtue and the motivation and rationale of virtuous action.

Noneudaemonistic Virtue Ethics

Criticisms such as these have motivated, at least in part, the development of noneudaemonistic virtue ethics, which does not require that status of a trait as a virtue is grounded in the flourishing of the possessor of the virtues. At least two non-eudaemonistic forms have been explored.

1. Agent-based virtue ethics. According to 'agent-based' views, developed by Slote and inspired by Plato and Martineau (1891), a virtue is a trait that is admirable, and its admirability need not be based on its contribution to further value, such as the welfare of human agents or the environment generally. Nor need a virtue be even partially constitutive of the flourishing of a human agent. Such a view relies for its development on aretaic intuitions about admirability, rather than the more standardly appealed to axiological intuitions about what is valuable.
2. Response views. Although all virtues are dispositions of appropriate responsiveness to or acknowledgment of items within their fields, according to response views this feature is what makes a trait of character a virtue. It is not also necessary that a virtue characteristically constitute (in part) the flourishing of the agent. Response views may be more or less pluralistic about both the bases of moral acknowledgment (such as bond, value, status, and the good for) and the forms of moral acknowledgment (such as promotion, honoring, love, respect, appreciation, and creativity). Response views that are genuinely virtue ethical insist that important goods and states (such as pleasure, money, and friends) cannot be understood as goods entirely independently of virtue (Swanton, 2003, 2013).

Intellectual Context of Modern Virtue Ethics

Three main concerns have motivated the rise of virtue ethics in recent times:

1. Increasing emphasis on the importance of good character and character education for a healthy society.
2. Dissatisfaction with rival moral theories – particularly consequentialism and Kantianism.
3. Increasing recognition of the centrality of virtue in philosophers such as Hume, Nietzsche, Confucius, and neo-Confucians.

The emphasis on character and the character training of children is increasingly thought to be important, indeed central, for an ethically sound society. Moral rules, where important, need to be made, interpreted, and applied in complex situations, and in ways that are not distorted by such vices as greed, impatience, cowardice, lack of perseverance, racism and other forms of hostility, callousness, oversensitivity, lack of concern for or overconcern with, fairness, and so on. A difficulty for a reorientation of ethical thought toward virtue ethics, in line with the above concern, is the apparent linkage of virtue ethics with social conservatism. This view, however, is based on several misunderstandings. First, virtue ethics as such is open as to how we delineate the virtues. Specific versions of virtue ethics will differ on this question. Second, although it would be agreed by all that it is imperative to bring up children to possess character traits that are not personally and socially destructive, virtue ethics as such is open as to whether a correct version of virtue ethics provides a monistic or pluralistic answer to the question 'What is a good life for a human being?' Third, it is sometimes thought that virtue ethics is too 'individualistic' and fails to appreciate the limitations of a virtuous agent, and her embeddedness in society. However, a virtue ethicist could agree that even a virtuous human agent necessarily has limited experience, possesses limited perspectives, and has limited expertise and knowledge. Accordingly, it is open to a virtue ethicist to adopt a dialogical view of virtues of practice, and moral epistemology generally (Swanton, 2003).

The philosophical/theoretic push to virtue ethics has been driven by several factors, including the following. One factor is dissatisfaction with the lack of emphasis on, or a denigration of, the 'inner' – feelings, emotions, passions, and practical wisdom – in the life of the moral agent. Virtue ethicists emphasize the importance of outer action as expressive of inner states: benevolence, for example, is not just a matter of promoting value or good. Fine emotions, for example, are important for the situational appreciation required by practical wisdom. Emphasis on the emotions also brings into view forms of moral acknowledgment not sufficiently emphasized in analytic philosophical ethics, such as (virtuous) caring and love. It also highlights the centrality in ethics of certain partialistic virtues such as parental love and friendship.

A second factor in the intellectual context of modern virtue ethics is its objection to decontextualized understandings of moral theory. For example, for a virtue ethicist, ethics is neither essentially impartial nor essentially partial. Rather there are both virtues and vices of impartiality, and virtues and vices of partiality relative to different fields. An understanding of partialistic vices, such as nepotism and favoritism (toward one of one's children, for example), and impartialistic vices (such as dispositions to seriously neglect near and dear in the interests of 'the common good'), requires a highly nuanced and contextualized understanding of the boundaries between these vices and allied virtues.

Related to the above feature of virtue ethics is a third and much-emphasized factor: skepticism about the codifiability of ethics in terms of decisive universal principles. Virtue ethicists favor a conception of ethical expertise that emphasizes experience and practical wisdom. Because of their emphasis on detail, particularity, and complexity, they recognize that moral expertise, which relies on knowledge of, for example, medicine, business, educational, and other practices, does not generalize to all contexts.

A fourth factor is a desire to make the concept of living well central to moral evaluation. Commonly for virtue ethicists, the supposed gap between the good life and the moral life is reduced via the desideratum of living a virtuous life. Living a virtuous life reduces this gap primarily through a reconfiguration of the contrast between altruism and self-interest. A virtuous agent, characteristically and in normal circumstances, enjoys the moral life as Aristotle points out, because her capacities for strength, love, concern, respect, and lack of resentment enable her to see her own concerns as not sharply in contrast with the interests of others. Increasing interest in understanding Nietzsche, Hume, and Confucius through virtue ethics has made this clear.

Virtue Ethics and Applied Ethics

The application of virtue ethics to social practice raises the question of an epistemology suitable for virtue ethics. Such an epistemology may be monological or dialogical. Commonly, it is thought that virtue ethics is inherently monological because the arbiter of rightness of action is thought to be the (or a) virtuous agent, that is, an agent of good moral character, practical wisdom, fine sensibilities, and the ability to appreciate a situation and context in fine detail. Nonvirtuous agents simply emulate virtuous ones. However, even virtuous agents are necessarily limited because of their human condition (e.g., a virtuous man cannot have the experiences of women who are contemplating an abortion). Arguably, a suitable virtue ethical epistemology for the world of practice, for example, education, medicine, business, aid work, family relations, must be properly socially grounded. Participants in a practice, recognizing their epistemic and other limitations, exercise the moral/epistemic virtues of practice, which include the dialogical virtues. These of course are legion, including imaginative deliberation, insight, perseverance, respect in various forms, open-mindedness, facilitativeness, and an absence of those vices that distort practical wisdom and communication, for example, cowardice, rigidity, hostility, callousness, excessive love of power, squeamishness, and lack of concern for or excessive concern for fairness. For a virtue ethicist, determining whether a participant is, for example, excessively concerned with fairness or insufficiently concerned with fairness is often a matter of fine judgment applied in contexts where many values are at stake, and knowledge and

appreciation of relevant facts, including the feelings, preferences, and so forth of affected parties, commonly is gleaned only through well-conducted dialogue.

Serious application of virtue ethics to ethics of roles in the analytic tradition is relatively speaking still in its infancy but see Oakley and Cocking (2001) and Swanton (2007b).

Skepticism about Character

One important criticism of virtue ethics derives from the social sciences, specifically 'situationist' psychology (Ross and Nisbett, 1991). According to some philosophers who make use of this psychology, virtue ethics is inadequate because it is based on a characterological psychology that is empirically inadequate. According to situationist psychology:

> Behavioral variation across a population owes more to situational differences than dispositional differences among persons. Individual dispositional differences are not as strongly behaviorally individuating as we might have supposed… (Doris, 1998: p. 507)

The claim that experiments such as Milgram's (1963) demonstrate the empirical inadequacy of virtue ethics because there is a lack of cross-situational consistency in behavior betrays a lack of understanding of the virtue ethicist's concept of virtue. For a virtue ethicist, a virtue is not simply a disposition to perform acts of a certain type (e.g., beneficent acts). The possession of virtue also requires the possession of fine inner states. Admittedly, Milgram's experiments show that a remarkable number of subjects administer electric shocks of considerable severity, in experimental situations of a certain type. The tendency to perform beneficent acts is arguably not as robust as one might hope. However, there was considerable variation in the mental states of those prepared to administer such shocks, and these differences may point to character traits of, for example, compassion, benevolence, respect for authority, and commitments, which manifest themselves in various ways in dilemmatic situations. According to Milgram, the subjects differed markedly in their emotional reactions to increasingly pressuring requests to administer shocks. Milgram claims that an unanticipated effect "was the extraordinary tension generated by the procedures." He says: "One might suppose that a subject would simply break off or continue as his conscience dictated. Yet this is far from what happened. There were striking reactions of tension and emotional strain." While some subjects reported and manifested intense strain and anguish, "some subjects (who had administered maximum shocks) had remained calm throughout the experiment, and displayed only minimal signs of tension from beginning to end" (Milgram, 1963: p. 376).

For the virtue ethicist, emotional reactions are crucial for the attribution of virtue and vice, and the mark of the virtuous is the propensity to violate a 'virtue rule' such as 'be benevolent' only with reluctance, anguish, and distress. The degree to which these emotions are present will vary according to the circumstances. Furthermore, it is clear from Milgram's account that the anguished saw themselves as operating under a rather severe virtue dilemma. A benevolent agent may also possess the virtues of respect for authority, trust, honoring commitments, and appreciation of (the value of) knowledge. All these virtues were in play during the experiment (1963). Subjects rightfully assessed the experimental situation as warranting trust, the honoring of a commitment, and as promoting a worthy objective. Now where virtues of, for example, trust and fidelity appeared in acute dilemmatic relation to another – nonmaleficence – most subjects arguably gave insufficient weight to the latter. But virtue ethicists are not committed to the view that many actual agents are capable of resolving such dilemmas correctly when under severe stress. The degree of virtue in agents is, however, able to be determined by their emotional reactions to their admittedly faulty resolutions of dilemma (e.g., reactions of anguish or indifference). And those differences were evident. It is open to the virtue ethicist to argue that while trust was not misplaced, and although obedience to authority is a virtue, in this case most subjects displayed this virtue to excess. Arguably, this failure hindered the exercise of the virtue of nonmaleficence present in those who suffered anguish.

An excellent book-length reply to the situationist critique is Snow (2010). In general, according to virtue ethicists, good character is a centrally important component of the good life for an individual in society, and virtue ethics is a philosophically rich and fruitful approach to moral theory. The further development of this theory requires cooperation with the social sciences, if a richer understanding of virtue, as an excellence of character in human beings, is to be obtained.

See also: Kantian Ethics and Politics; Moral Reasoning in Psychology; Religion: Morality and Social Control; Social Work Ethics; Values, Sociology of; Wisdom, Psychology of.

Bibliography

Angle, S.C., 2009. Sagehood: The Contemporary Significance of Neo-confucian Philosophy. Oxford University Press, New York.
Aristotle, E., 1955. Nicomachean Ethics (J.A.K. Thomson, Trans.). Penguin, London.
Anscombe, 1958. Modern moral philosophy. Philosophy 33, 1–19.
Annas, J., 1993. The Morality of Happiness. Oxford University Press, New York and Oxford.
Annas, J., 2011. Intelligent Virtue. Oxford University Press, Oxford and New York.
Baron, M.W., Pettit, P., Slote, M., 1997. Three Methods of Ethics. Blackwell, Oxford, UK.
Crisp, R., Slote, M. (Eds.), 1997. Virtue Ethics. Oxford University Press, Oxford and New York.
Doris, J.M., 1998. Persons, situations, and virtue ethics. Noûs 32, 504–530.
Driver, J., 1989. The virtues of ignorance. The Journal of Philosophy LXXXVI, 373–384.
Driver, J., 1996. The virtues and human nature. In: Crisp, R. (Ed.), How Should One Live? Essays on the Virtues. Oxford University Press, Oxford and New York.
Foot, P., 1978. Virtues and Vices and Other Essays in Moral Philosophy. Blackwell, Oxford, UK.
Fromm, E., 1977. The Anatomy of Human Destructiveness. Penguin, London.
Horney, K., 1951. Neurosis and Human Growth: The Struggle for Self Realization. Routledge and Kegan Paul, London.
Hume, D., 1978. Enquiries Concerning Human Understanding and Concerning the Principles of Morals, third ed. Clarendon Press, Oxford, UK.

Hurka, T., 1992. Virtue as loving the good. Social Philosophy and Policy 9, 149–168.
Hurka, T., 1999. Three faces of flourishing. Social Philosophy and Policy 16, 44–71.
Hursthouse, R., 1991. Virtue theory and abortion. Philosophy and Public Affairs 20, 223–246.
Hursthouse, R., 1995. Applying virtue ethics. In: Hursthouse, R., Lawrence, G., Quinn, W. (Eds.), Virtues and Reasons: Philippa Foot and Moral Theory. Clarendon Press, Oxford, UK, pp. 57–75.
Hursthouse, R., 1996. Normative virtue ethics. In: Crisp, R. (Ed.), How Should One Live? Essays in the Philosophy of Virtue. Clarendon Press, Oxford, UK.
Hursthouse, R., 1997. Virtue ethics and the emotions. In: Statman, D. (Ed.), Virtue Ethics: A Critical Reader. Edinburgh University Press, Edinburgh, UK.
Hursthouse, R., 1999. On Virtue Ethics. Oxford University Press, Oxford.
Kant, I., 1996. The doctrine of virtue. In: MacGregor, M. (Ed.), The Metaphysics of Morals. Cambridge University Press, Cambridge, UK.
Louden, R., 1984. On some vices of virtue ethics. American Philosophical Quarterly 21, 227–236.
MacIntyre, A., 1984. After Virtue. University of Notre Dame Press, Notre Dame, IN.
Martineau, J., 1891. Types of Ethical Theory, vol. 2. Oxford University Press, Oxford, UK.
McDowell, J., 1979. Virtue and reason. Monist 62, 331–350.
Milgram, S., 1963. Behavioral study of obedience. Journal of Abnormal and Social Psychology 67, 371–378.
Milgram, S., 1974. Obedience to Authority. Harper and Row, New York.
Nietzsche, F., 1973. Beyond Good and Evil (R.J. Hollingdale, Trans.). Penguin, London and New York.
Nietzsche, F., 1996. The Genealogy of Morals (D. Smith, Trans.). Oxford University Press, Oxford, UK.
Oakley, J., Cocking, D., 2001. Virtue Ethics and Professional Roles. Cambridge University Press, Cambridge.
Pettit, P., 1997. The consequentialist perspective. In: Baron, M.W., Pettit, P., Slote, M. (Eds.), Three Methods of Ethics. Blackwell, Oxford, UK.
Ross, L., Nisbett, R.E., 1991. The Person and the Situation. Temple University Press, Philadelphia, PA.
Slote, M., 1996. Agent based virtue ethics. In: French, P.A., Uehling Jr., H.K., Wettstein, T.E. (Eds.), Midwest Studies in Philosophy, Moral Concepts, vol. XX. University of Notre Dame Press, Notre Dame, IN.
Slote, M., 2001. Morals from Motives. Oxford University Press, Oxford.
Solomon, D., 1988. Internal objections to virtue ethics. In: French, P.A., Uehling Jr., H.K., Wettstein, T.E. (Eds.), Midwest Studies in Philosophy, Ethical Theory: Character and Virtue, vol. XIII. University of Notre Dame Press, Notre Dame, IN, pp. 428–441.
Solomon, R.C., 1998. The virtues of a passionate life: erotic love and 'the will to power'. Social Philosophy and Policy 15, 91–118.
Statman, D. (Ed.), 1997. Virtue Ethics: A Critical Reader. Edinburgh University Press, Edinburgh, UK.
Snow, N.E., 2010. Virtue as Social Intelligence: An Empirically Grounded Theory. Routledge, New York.
Swanton, C., 2003. Virtue Ethics: A Pluralistic View. Oxford University Press, Oxford.
Swanton, C., 2005. Nietzschean virtue ethics. In: Gardiner, S.M. (Ed.), Virtue Ethics, Old and New. Cornell University Press, Ithaca, New York, pp. 179–192.
Swanton, C., 2007a. Can Hume be read as a virtue ethicist? Hume Studies 33, 91–113.
Swanton, C., 2007b. Virtue ethics, role ethics, and business ethics. In: Walker, Rebecca L., Ivanhoe, Philip J. (Eds.), Working Virtue: Virtue Ethics and Contemporary Moral Problems. Clarendon Press, Oxford, pp. 207–224.
Swanton, C., 2011. Nietzsche and the virtues of mature egoism. In: May, S. (Ed.), Critical Guide to Nietzsche's on the Genealogy of Morality. Cambridge University Press, Cambridge, New York.
Swanton, C., 2013. The definition of virtue ethics. In: Russell, D.C. (Ed.), Cambridge Companion to Virtue Ethics. Cambridge University Press, New York, pp. 315–337.
Trianosky, G., 1990. What is virtue ethics all about? American Philosophical Quarterly 27, 335–344.
Watson, G., 1990. On the primacy of character. In: Flanagan, O., Rorty, A. (Eds.), Identity Character and Morality: Essays in Moral Psychology. MIT Press, Cambridge, MA, pp. 445–469.

Vision, High-Level Theory of

Johan Wagemans, University of Leuven (KU Leuven), Leuven, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In a hierarchical view on vision, low-level vision concerns the extraction of simple, local features from retinal images, while high-level vision concerns the interpretation of their meaning. Classic work on object recognition is reviewed, followed by more recent work on scene interpretation, action and event understanding, and appreciation of visual art. In these areas, the evidence supports an intricate interplay between low-, mid, and high-level visions, which blurs the traditional distinctions. Alternative frameworks, such as reverse hierarchy theory and hierarchical predictive coding, are proposed to explain better what is currently known about high-level vision.

Introduction

The concept of 'high-level vision' is used to denote aspects of vision that are influenced by information stored in memory such as an object's identity, its name, or the context in which it belongs. It is usually distinguished from low-level visual processing, which is primarily driven by the visual input. High-level vision involves functions such as object recognition, face perception, picture naming, visual categorization, scene interpretation, action and event understanding, and even appreciation of visual art.

Background

In hierarchical models of vision (e.g., Marr, 1982; Serre et al., 2007), higher levels of visual processing operate on the building blocks delivered by more primitive visual mechanisms. In Marr's approach to perception, each stage has its own algorithms and its own format of representing processing output. The primal sketch represents the basic features and their spatial relations (low-level vision) and the 2.5-dimensional (2.5-D) sketch represents surfaces and shapes from the observer's viewpoint (midlevel vision). Recognition consists of establishing a match between the descriptions derived from the visual input and those stored in visual long-term memory. Marr was convinced that this required a representation of the full three-dimensional (3-D) structure of an object and its components in a so-called 3-D object model (high-level vision).

Marr's approach is more than 30 years old now, but it is still highly influential as a framework to understanding vision in terms of representations and computations. However, his assumption of strictly bottom–up processing in sequential stages has gradually lost attraction. One reason for this theoretical evolution has been the increased popularity of models in which different levels of processing interact with each other. Partial outputs from lower level processes initiate higher level processes and the outputs of higher level processes feed back into the lower level processes. When different levels are distinguished in more recent theories, great care is taken to avoid the assumption of strict sequential processing.

However, the issue of where early vision ends, where high-level vision starts, and how to demarcate the two has remained important. Pylyshyn (1999) has proposed the notion of cognitive impenetrability as the major criterion for separating early (or low-level) vision from later (or higher level) vision. Early visual processes run autonomously – they cannot be influenced by what is known about the world or expected in a specific situation. Most aspects of visual processing that seem cognitive (e.g., recovering unique 3-D percepts from ambiguous two-dimensional inputs) are attributed to embodied natural constraints derived from principles of optics and projective geometry, internal to the visual mechanisms themselves. Only effects of object-specific beliefs on earlier visual processing would count as cognitive penetration. A large number of so-called top–down effects of higher level processes on early vision are then explained as effects of attention (preceding early vision) or as postperceptual effects on the decision stage based on the perceptual output.

Implicit in Pylyshyn's (1999) effort to delineate early vision from cognitive influences is his belief – shared by many vision scientists – that real scientific progress is possible only in early vision. Arguably, however, in the last decades considerable progress has been made in the area of high-level vision as well, initially with regard to object recognition, but more recently also with regard to the interpretation of scenes, actions, events, and even art.

Classic Work on Object Recognition

Biederman (1987)

Object recognition concerns the identification of an object as a specific entity (i.e., semantic recognition) or the ability to tell that one has seen the object before (i.e., episodic recognition). Building on Marr's ideas about matching perceptually derived 3-D volumetric descriptions to those stored in memory, Biederman (1987) has formulated a theory of object recognition that has incorporated some crucial aspects of midlevel vision as well.

Biederman proposed that objects are identified at their entry level on the basis of a structural description of its components and their spatial relations. All object parts are modeled by a limited number of simple primitives called 'geons,' a specific

subset of generalized cones. A generalized cone is the volume swept out by a cross section moving along an axis. Geons are defined by the categorical attributes on the following dimensions: the axis (straight or curved), the shape of the cross section (straight or curved, symmetric or asymmetric), and the size of the cross section (constant, expanded, or expanded and contracted).

One major claim in Biederman's recognition-by-components (RBC) theory is that geons themselves are viewpoint invariant because the underlying dimensions can be distinguished on the basis of so-called nonaccidental properties (NAPs). These are special features in the image, such as collinearity, curvilinearity, cotermination, parallelism, and symmetry, which are reliable in the sense that they are most likely caused by similar characteristics in 3-D space, at least under the assumption of a general viewpoint. Biederman argued that Gestalt principles of grouping support the extraction of these NAPs and that they are thus detected fast and reliably enough to support bottom–up identification of geons.

Allowing for different combinations of different geons, with different spatial relations between them, yields enormous combinatorial possibilities. Based on computational grounds, Biederman argued that the recovery of a small number of geons (3 or 4) is enough to quickly recognize complex objects even when they are occluded, rotated in depth, degraded, or lacking color or texture. This claim has also been supported empirically by experiments in which parts of objects (in fact, line drawings of objects) were deleted until only two or three parts were left. Other supporting evidence came from recognition of contour-deleted line drawings (which is possible as long as the geons are recoverable), priming with contour-deleted pictures (which is situated at the level of the parts, not the image fragments and not the whole picture), and viewpoint invariance of 3-D object recognition (under certain limited conditions). In more recent years, additional support for this theory of object recognition was obtained in infants, non-Western cultures, and at neural levels.

Edelman (1999)

Another important theoretical step forward in high-level vision was Edelman's (1999) book on *Representation and Recognition in Vision*. Edelman argued that the attempt to base recognition on geometrically reconstructed representations of distal objects (like Marr's 2.5-D sketch) suffers from serious theoretical and practical problems. First, it would lead to the homunculus problem and, more pragmatically, it appears, at least up till now, not to work, not even when using the most sophisticated computer vision techniques. The major reason for this failure might be that previous approaches have tried to achieve veridicality of object representation by aiming for isomorphisms between objects in the world and their corresponding internal representations (i.e., first-order isomorphism or representation by similarity). Instead, Edelman proposed to aim for the formally just as good isomorphism between the relations among several external objects and the relations among their corresponding internal representations (i.e., second-order isomorphism or representation of similarity).

To this end, Edelman introduced the notion of a shape space, a formalism borrowed from mathematical statistics. A shape space is a metric space in which each point corresponds to a particular shape and in which geometric similarity between shapes can be defined rigorously as proximity, a quantity inversely related to distance. A novel shape (or a novel view of a previously seen shape) can be localized in such a shape space based on its similarity to known objects (or stored views of known objects). A tuned unit (or module of units) that responds optimally to some shape (i.e., the landmark) and progressively less to progressively less similar shapes, is enough to implement this mechanism. Edelman used a radial basis function approximation network that can be trained and then generalizes its response from a series of given views of an object to other views of that object. As a by-product of this learning, the network also responds progressively less to progressively less similar objects, precisely what is needed for it to work as an active landmark in internal shape space. This scheme works not only for known objects but also for novel objects. The trick is that a new object must be represented by a vector of proximities to several reference shapes or landmarks. For example, even if one has never seen a giraffe before, it is possible to compute the proximities to known animals such as a camel, a goat, a pig, or even a leopard.

Edelman has called this system a 'chorus of prototypes' to emphasize that the prototypical landmarks act together in representing shape. Initial simulation results with this system have been encouraging, especially in light of the wide variety of tasks: identification of novel views of familiar objects, categorization of novel object views, discrimination among views of novel objects, local viewpoint invariance for novel objects, recovery of a standard view and of pose for novel objects, prediction of a novel view for a novel object, and so on. None of its theoretical competitors (e.g., Biederman's RBC) could handle such a variety of tasks.

More Recent Work on Scene Interpretation, Action and Event Understanding, and Art Appreciation

While object recognition was at the center of high-level vision in the 1980s and 1990s, the last couple of decades have seen a significant surge of interest in other aspects of high-level vision, which appear even harder because of the complexity of the stimuli to be interpreted and the variety of possible responses (interpretations and even social and emotional reactions).

Scene Interpretation

Scene interpretation has always constituted a great challenge, because scenes are more complex than isolated objects; they consist of several different objects in several mutual spatial relationships, usually with more clutter and stronger effects of lighting and occlusion than single objects. In a fruitful collaboration between psychology and computer vision, Oliva and Torralba (2006) showed that it is possible to extract a number of interesting scene characteristics from relatively simple image features. They analyzed global image characteristics instead of first segmenting the image into regions, extracting isolated object shapes, and computing spatial relations between them. For instance, the roughness of

the luminance profile across the whole image happens to be smooth for open natural scenes like a beach and rough for closed urban scenes, the spatial envelope can be low or high on openness and expansion, and so on.

When lots of images of different scenes are plotted in a multidimensional graph of such global image characteristics, they appear to cluster together by semantic category (e.g., street images versus highway images), which means that the gist of a scene can be recognized on the basis of a combination of global image characteristics. In total, seven global scene properties could be identified that can easily be derived from natural images and that appear important for scene interpretation: openness, expansion, mean depth, temperature, transience, concealment, and navigation. Greene and Oliva (2009) then showed experimentally that human perceivers are also capable of extracting these global scene properties from natural images flashed as briefly as 20–70 ms. Similar critical exposure durations were also obtained for scene categorization of the images (e.g., desert, forest, mountain, and ocean). The theoretically important message from this line of work is that semantically high-level interpretation is sometimes relatively easy for the visual system to derive from surprisingly simple low-level (but global) image features.

Action and Event Understanding

Humans have a strong tendency to interpret sequences of movements as meaningful events and actions, even with simple geometric shapes moving dynamically across a screen. Building on early work by Albert Michotte and Fritz Heider, Brian Scholl and colleagues have recently studied such action and event understanding in more detail. Specifically, they wanted to show that the perception of animacy is really a perceptual phenomenon and not the result of a higher order cognitive judgment. In order to do this, they systematically manipulated a number of subtle stimulus parameters and measured their effect on task performance, in addition to their effect on the perceptual impressions as such.

In one series of experiments (Gao et al., 2009), simple geometric shapes moving dynamically across a screen were interpreted as the chasing of one shape (a 'sheep') by another shape (a 'wolf'), amidst a large number of other shapes. Observers could control the sheep's movements by moving the computer mouse and were instructed to avoid being touched ('caught') by the wolf. While the sheep was highlighted by being drawn in a different color, the wolf looked identical to the distracters and could therefore be identified only on the basis of its spatiotemporal behavior (i.e., whether it followed the sheep's movement or not). Performance in this chasing detection task was strongly dependent on the chasing subtlety, which was manipulated by the maximal angular deviation of the wolf's heading direction relative to perfectly following the sheep (for a demonstration, see http://www.yale.edu/perception/chasing/). When the chasing subtlety was minimal, the chasing was obvious and could easily be avoided. When the chasing subtlety was maximal, performance was high again but mainly because the wolf was actually not chasing competently. For a large range of intermediate levels of chasing subtlety, the wolf was chasing competently but the chasing was difficult to detect. Such a nice dissociation between actual chasing and perceived chasing as a function of a subtle stimulus manipulation supports the idea that perceived animacy is really a perceptual process (with nicely controllable psychophysical characteristics) and not the consequence of a higher level cognitive judgment. Indeed, subjects were often unaware of the differences between the different stimulus conditions, which had a clear effect on their task performance.

In another series of experiments (Gao et al., 2010), a green disc moved quasirandomly across the screen, along with a number of 'dart' shapes that moved completely independently from one another and from the green disc. In one condition, the darts' orientations were continuously adjusted during their motion trajectories so that they were always pointing directly at the disc. Observers spontaneously misinterpreted this in terms of animate behavior such as chasing, presumably on the basis of the darts' facing their target, a phenomenon that the authors have called the 'wolfpack effect' (for a demonstration, see http://www.yale.edu/perception/wolfpack/). In a perpendicular control condition, the facing direction was updated continuously to be pointing 90° away from the green disc, which completely eliminated the wolfpack effect. Hence, although these two conditions were well-matched in terms of all lower level properties, one evoked a rich percept of animacy and intentionality, while the other did not.

In one of the experiments with these stimuli, the authors combined this manipulation with the explicit task of detecting a wolf, this time clearly distinguished from both the sheep and the darts by being colored differently. When the wolfpack was facing the sheep, performance was reduced by more than 20%, compared to the perpendicular control condition, as if the wolfpack (which consisted of independently moving darts) distracted attention away from the real wolf (moving in relation to the sheep). When the wolfpack was facing the wolf, the effect was attenuated. Hence, although subjects were fully informed about the irrelevance of the darts' orientations, they could not refrain from interpreting them as an animate and intentional wolfpack. This provides strong evidence that this is an automatic perceptual process rather than a higher level cognitive judgment.

Art Appreciation

Art is still more complex, obviously, and one's reactions are even richer, from pure perceptual stimulation to aesthetic pleasure and the experience of thrills and being deeply moved. In fact, trying to understand people's reactions to art requires the integration of all the processes psychology usually distinguishes (attention, perception, emotion, etc.) and of all the factors known to influence these processes (in addition to stimulus factors, also the person's background, interest, knowledge, expertise, etc.). In this way, experimental psychoaesthetics – the science of understanding art appreciation – is a fascinating microcosmos and test bed for psychology as a whole (Wagemans, 2011). In spite of these complexities, here as well recent research has made good progress in relating these very high-level processes to known midlevel visual processes.

A good example is a study by Koenderink et al. (2012) with line drawings of human figures made by Picasso in the 1940s. These drawings provide an interesting case study of high-level vision because they contain a lot of white space, which appears to be filled in one's mind when interpreting them as bodily surfaces in particular attitudes (i.e., orientations in space) and because Picasso intentionally combined multiple perspectives in a single drawing, which means that one can never have experienced such an image before. As a rendering of a 3-D object, the line drawing is blatantly incorrect and yet one has no problem in seeing a human figure with clearly defined body parts that are visually coherent and meaningful.

Koenderink et al. asked their observers to adjust an elliptical gauge figure (which looks a bit like a pushpin with its needle sticking out) so as to appear as a circle painted on the surface, at a large number of predetermined locations within the surface area bounded by a simple outline consisting of a few well-placed pen strokes. Based on these local measurements, a reconstruction of the global shape is possible, which can then be analyzed in great detail, for instance, to compare the perceived shape between different observers, to relate the perceived shape to local image cues, and so on. In the context of this article, only a few highlights can be summarized. First, the purely planar structure of the drawings is sufficient to account for much, although by no means all of the data. The observed reliefs of the body surfaces have structures that are twisted in depth such that their outlines are nonplanar. When looking at these images, the perceptual processes giving rise to awareness apparently act in depth (e.g., they may make an arm move forward or backward in pictorial space). Depth cues such as contour curvature, contour terminations, and T-junctions (which also play an important role in Biederman's RBC theory) have a noticeable influence on the observed reliefs, but they are insufficient to account for all of the structure in the data. In these Picasso drawings, most of the paper is left blank and perceived structure must therefore be due to effects from the neighborhood of that area or it has to be supplied by the perceiver. Examples of both were obtained in this study. A clear instance of the former case is a convexly curved arc whose direction changes continuously over 60° or more, perceived as an ovoid shape (e.g., breasts or buttocks). A clear instance of the latter case involves the relief of the human back, which was perceived as singly convex by one observer and including a centrally elongated groove of the spine by another observer. These observers clearly applied a different type of template of the human back to fill in the empty space in the drawing.

It is tempting to speculate that the interplay between these image-based and knowledge-based influences, and the inter-individual variations in this interplay, may well contribute to the aesthetic pleasure experienced by the perceiver and to the artistic value of the artwork. Perhaps such a piece of visual art, which provides just enough cues to induce particular percepts, while still leaving enough to the perceiver's imagination, is appreciated so much because it imposes the optimal degree of constraints on our perceptual awareness (Van de Cruys and Wagemans, 2011). Such a view on the rewarding aspects of art is compatible with one of the new theoretical frameworks replacing the traditional view on the hierarchy of low-, mid-, and high-level visions.

New Theoretical Frameworks Replacing the Traditional Hierarchical View

The current mainstream view of human visual cortex emphasizes functional specialization and hierarchical organization (e.g., Grill-Spector and Malach, 2004). Functional specialization means that different parts of the brain are assumed to carry out different functions, for simple stimulus attributes like form, motion, and color, as well as for more complex semantic categories such as faces, bodies, and places (e.g., Kanwisher, 2010). This also involves the distinction between ventral stream processing for visual awareness of object identity and dorsal stream processing for visually guided action, also labeled as 'what' versus 'where' stream processing (e.g., Goodale and Milner, 1992). The hierarchical organization entails a number of anatomical and functional characteristics of the different cortical areas along the ventral stream (e.g., V1, V2, V3a, V4, LOC), such as increasing receptive field size, decreasing retinotopy, increasing invariance, and increasing complexity of the features that are encoded, from simple stimulus features like orientation and spatial frequency, to junctions, parts, and whole objects (e.g., Serre et al., 2007).

In line with this anatomical and functional characterization of human visual cortex, the current mainstream view of visual processing (often implicitly) assumes that high-level percepts are constructed from combinations of low-level attributes, much like whole objects are assembled from parts (cf. Biederman's RBC), parts arising from spatial relationships between line segments and junctions, line segments being the outputs from edge detectors, and so on. In such a view, processing is mainly feedforward, starting from local features encoded in low-level areas and ending with global interpretations in high-level areas. In the section 'More Recent Work on Scene Interpretation, Action and Event Understanding, and Art Appreciation,' several lines of research have been reviewed, showing that such a simple story needs revision. Scenes can be interpreted quickly based on global but low-level image characteristics. Dynamic displays can give rise to automatic, spontaneous high-level interpretations of animacy and intentions, which appear strictly tied to simple perceptual attributes like relative orientations and trajectories. High-level percepts of human bodies from simple line drawings consist of a mixture of low-level and midlevel processing of image features as well as knowledge-based shape templates, and aesthetic pleasure may be derived from the mixture of image-based constraints and subjective filling in of empty spaces in 'the eye of the beholder.' In the last decade, a number of alternative views on the relationships between the levels of cortical processing have emerged that might do better justice to these findings.

Reverse Hierarchy Theory

First, Hochstein and Ahissar (2002) have proposed the 'reverse hierarchy theory,' in which they argue that a fast feedforward sweep quickly activates global percepts (e.g., the overall gist of a scene) in high-level areas with large receptive fields, whereas feedback from these higher areas to lower areas and recurrent processing in the low-level areas with small receptive fields is necessary for fine-grained processing of local details (e.g., a difficult visual search task). Hence, this theory

distinguishes the anatomical and structural aspects of the hierarchy of the visual system (low- versus high-level representations) from the temporal and functional aspects of it (early versus late stages of processing). Reverse hierarchy theory can explain a lot of the findings on fast object and scene categorization, visual search, perceptual learning, and so on.

Second, Moshe Bar and colleagues (e.g., Bar, 2003; Bar et al., 2006) have proposed a general theory of visual object recognition, in which two streams of processing initially occur in parallel, one in which the high spatial frequencies of input images are processed in a relatively slow, feedforward sequence according to the standard hierarchy of stages (V1, V2, V4, …), and another in which the low spatial frequencies are quickly transmitted to areas in the prefrontal cortex to generate hypotheses about possible identities of the object as well as their most likely scene context, after which these two streams are integrated and refined in an iterative, reentrant way. Again, this view can explain a lot of the findings on object recognition, scene context effects, priming, and so on.

Both these views are similar in spirit to traditional ideas from Gestalt psychology (for recent reviews, see Wagemans et al., 2012a,b), because they both postulate that global processing comes first, both are highly interactive and dynamic views on cortical processing, and they are not even limited to purely visual areas. Indeed, Moshe Bar has more recently extended his view to include cognitive interpretation and emotional significance (e.g., Bar, 2007; Barrett and Bar, 2009), which would make it more suitable to explain the findings on animacy, intentionality, and aesthetic pleasure reviewed above. This view also bears family resemblance to another recent theoretical framework, called 'hierarchical predictive coding.'

Hierarchical Predictive Coding

In the mainstream view summarized before, the brain was largely conceived to passively encode and decode the available sensory information. In this new framework of understanding cortical processing, the brain is instead conceived to actively predict the forthcoming input. According to this so-called predictive coding account (for reviews, see Clark, 2013; Huang and Rao, 2011), expectations about the state of the world are stored or generated in higher level brain areas, which are then sent to lower level areas, where they are evaluated against the upcoming information. Discrepancies between prior expectations and the actual input are translated into prediction error (or 'surprise') signals, which are sent to higher level areas, where they are used to update the predictions. This process is reiterated across the complete cortical hierarchy until the prediction error is sufficiently minimized at all levels, which is why this view is labeled as 'hierarchical predictive coding.'

Friston (2010) has argued that this view of the brain as a hierarchical network, in which predictions are generated and evaluated, has the potential of unifying perception, action, and learning into a single theoretical framework. There is currently a blossoming literature in which many psychophysical and neural findings are reconceptualized in a predictive coding framework. In addition, new studies are carried out to try to support this view with direct empirical evidence, mainly as enhanced neural responses to violations of expected input. These findings are then taken to confirm the idea that higher level expectations are used to 'explain away' the lower level representations that are compatible with these expectations (Clark, 2013). It remains to be seen how productive this new framework will be in making unique, falsifiable predictions that cannot be made by other theories, and that are then also supported empirically.

Bibliography

Bar, M., 2003. A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of Cognitive Neuroscience 15, 600–609.

Bar, M., 2007. The proactive brain: using analogies and associations to generate predictions. Trends Cognitive Sciences 11, 280–289.

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E., 2006. Top-down facilitation of visual recognition. Proceedings of the National Academy of Science of the USA 103, 449–454.

Barrett, L.F., Bar, M., 2009. See it with feeling: affective predictions in the human brain. Philosophical Transactions of the Royal Society, Series B 364, 1325–1334.

Biederman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychological Review 94, 115–147.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36 (3), 181–204.

Edelman, S., 1999. Representation and Recognition in Vision. MIT Press, Cambridge, MA.

Friston, K.J., 2010. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11, 127–138.

Gao, T., McCarthy, G., Scholl, B.J., 2010. The wolfpack effect: perception of animacy irresistibly influences interactive behavior. Psychological Science 21, 1845–1853.

Gao, T., Newman, G.E., Scholl, B.J., 2009. The psychophysics of chasing: a case study in the perception of animacy. Cognitive Psychology 59, 154–179.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends in Neurosciences 15, 20–25.

Greene, M.R., Oliva, A., 2009. The briefest of glances: the time course of natural scene understanding. Psychological Science 20, 464–472.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annual Review of Neuroscience 27, 649–677.

Hochstein, S., Ahissar, M., 2002. View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36, 791–804.

Huang, Y., Rao, R.P.N., 2011. Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science 2, 580–593.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the USA 107, 11163–11170.

Koenderink, J.J., van Doorn, A.J., Wagemans, J., 2012. Picasso in the mind's eye of the beholder: three-dimensional filling-in of ambiguous line drawings. Cognition 125, 394–412.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman, San Francisco, CA.

Oliva, A., Torralba, A., 2006. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Research 155, 23–36.

Pylyshyn, Z.W., 1999. Is vision continuous with cognition? the case for cognitive impenetrability of visual perception. Behavioral and Brain Sciences 22, 341–423.

Serre, T., Oliva, A., Poggio, T., 2007. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Science of the USA 104, 6424–6429.

Van de Cruys, S., Wagemans, J., 2011. Putting reward in art: a tentative prediction error account of visual art. i-Perception 2, 1035–1062. http://dx.doi.org/10.1068/i0466aap (special issue on Art & Perception).

Wagemans, J., 2011. Towards a new kind of experimental psycho-aesthetics? Reflections on the Parallellepipeda project. i-Perception 2, 648–678. http://dx.doi.org/10.1068/i0464aap (special issue on Art & Perception).

Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., von der Heydt, R., 2012a. A century of Gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. Psychological Bulletin 138, 1172–1217.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J.R., van der Helm, P., van Leeuwen, C., 2012b. A century of Gestalt psychology in visual perception: II. conceptual and theoretical foundations. Psychological Bulletin 138, 1218–1252.

Vision, Low-Level Theory of

Hugh R Wilson, York University, Toronto, ON, Canada

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16232–16237, © 2001, Elsevier Inc.

Abstract

Models of visual function may be construed as low level theories in two related senses. First, theories may be regarded as low level when they deal explicitly with the lower levels of the retino-cortical visual hierarchy. For the purposes of this article this will be assumed to comprise the retina and cortical areas from primary visual cortex through area V4 in the ventral or form vision pathway. The second sense in which models may be construed as low level is with regard to the level of description. Thus, while high level vision models generally deal with abstractions that are at present but tenuously related to underlying neural network activity, low level vision theories are concerned with the explicit development of neural network models which are consistent with single unit neurophysiology on the one hand and are capable of predicting perceptual data on the other. The models discussed in this article are all low level in this network sense as well. Thus, this article will focus on contemporary neural models of retinal, striate, and extrastriate cortical processing in the form vision system. Key topics include 'feedback,' 'gain controls,' 'configural pooling,' and 'cooperative–competitive interactions.'

Introduction

The primate visual system comprises over 30 cortical areas in addition to the retina and other subcortical structures. To model the lower levels of form vision or motion perception, it therefore becomes necessary to develop neural network theories of retinal, striate, and extrastriate cortical function that are consistent with known anatomy and physiology and that are capable of predicting relevant psychophysical data. The mathematical techniques necessary for this enterprise include convolution and 'nonlinear dynamical systems' theory (see also Self-Organizing Dynamical Systems). This article will review current theories of retinal and striate cortical function followed by a discussion of neural models related to areas 'V2' and 'V4' in the form vision hierarchy.

Theories of Retinal Function

Retinal neurons transform the optical image in order to extract biologically relevant visual information concerning light intensity changes in space (i.e., contours and edges), changes in time, and chromaticity. Furthermore, the retina must shift its operating point so as to signal useful information over more than a brightness range, a process termed light adaptation. These transformations are carried out by networks comprising at least five different types of neurons: photoreceptors, bipolar cells, horizontal cells, ama-crine cells, and ganglion cells (see Wässle and Boycott, 1991). As the extraction of color information is covered elsewhere let us focus here on light adaptation and the extraction of contours.

Light intensities in natural scenes cover more than a 10^6 range, yet the ganglion cells providing the retinal output to the brain can only respond over about a 300:1 range. Thus, the retina must compute the mean level of illumination at any given time and adjust the circuitry to respond to about a 300:1 range around that operating point. Key experiments on retinal light adaptation have demonstrated that there are two major processes involved: one subtractive and one divisive (see Hood, 1998). The direct path through the retina involves photoreceptors activating bipolar cells, which in turn activate ganglion cells. In addition the retina incorporates two major feedback loops: (a) from photoreceptors to horizontal cells then back to photoreceptors and (b) from bipolars to amacrine cells back to bipolars. A recent retinal model has tentatively identified each of these loops with one of the light adaptation processes (Wilson, 1997). In particular, loop (a) was conjectured to provide sub-tractive feedback, while loop (b) was suggested to be the site of divisive feedback. In a subtractive loop, the feedback signal is subtracted from the loop input, while in a divisive loop, the feedback signal divides the loop input signal thus reducing its gain. (More likely, each feedback loop incorporates both subtraction and division, suggesting a somewhat more complex formulation.) It is significant that both horizontal cells and the amacrine cells, the key units in these two feedback loops, have a relatively broad horizontal spread across the retinal surface and thus provide feedback that is averaged over an extended area. Subtractive feedback is well understood in the context of linear systems theory, but divisive feedback moves into the realm of 'nonlinear dynamics'. Designating the bipolar cell response as B, the amacrine feedback response as A, and the light-derived signal received from the photoreceptors as L, the simplest divisive feedback equations are of the form (Wilson, 1997):

$$\begin{aligned} \frac{dB}{dt} &= -B + \frac{L}{1+A} \\ \frac{dA}{dt} &= -A + B^N \end{aligned} \qquad [1]$$

These equations may be analyzed using standard techniques (Wilson, 1999) of nonlinear dynamical systems theory. In particular, one can solve for the steady state or equilibrium value at which $dB/dt = 0$ and $dA/dt = 0$. The results for very small and very large values of L are approximately:

$$\begin{aligned} B &\approx L \text{ for } L \leq 1 \\ B &\approx L^{1/(N+1)} \text{ for } L \gg 1 \end{aligned} \qquad [2]$$

Thus, at very low light intensity levels the bipolar response B is approximately equal to its input, so there is no attenuation of weak signals. At high light intensity levels, however, B is compressed to approximately the $(N+1)$ root of L. For $N = 2$, a physiologically plausible figure, the divisive feedback loop in eqn [1] will compress an $L = 10^6$ input into a $B = 100$ output, easily compatible with the response range of the ganglion cells. A further property of eqn [1] is that the temporal response becomes more transient as L increases (Wilson, 1999), and this change is well documented in the literature on flicker sensitivity and light adaptation (Wilson, 1997).

As is well known, retinal receptive fields have an excitatory center and inhibitory surround (or vice versa) spatial profile (see also Visual Perception, Neural Basis of). Direct bipolar input to the ganglion cells produces the receptive field center, while the lateral spread of both horizontal and amacrine cells creates the antagonistic surround. Thus, the neurons engaged in light adaptation via feedback also produce the spatial surround of ganglion cell receptive fields. As is evident from the discussion above, divisive feedback has almost no effect at low light levels (eqn [2]), so one would expect receptive field surrounds to be very weak at low luminances but much stronger at higher luminances, and this has been demonstrated experimentally. For some simulation purposes the spatial characteristics of retinal receptive fields may be approximated as a circularly symmetric difference of Gaussians:

$$G(R) = A\exp\left(-R^2/\sigma_c^2\right) - B(L)\exp\left(-R^2/\sigma_s^2\right) \qquad [3]$$

where R is radius, and σ_c and σ_s are the space constants of the center and surround, respectively. The inhibitory gain $B(L)$ would vary from $B(0) \approx 0$ to $B(L) \approx A\,\sigma_c/\sigma_s$ for $L \gg 0$, which makes the integral of the receptive field $\int G\,dR = 0$. For further discussion of retinal models see Wilson (1997) and Hood (1998).

Theories of Striate Cortical Function

Hubel and Wiesel (1968) categorized striate (V1) cortical cells as simple or complex (see also Visual Perception, Neural Basis of). Both of these cell types show a preference for a particular orientation and also for a range of spatial frequencies. The now classical model for a simple cell receptive field (RF) is just an oriented linear filter with parallel excitatory and inhibitory zones defined by horizontal (x) and vertical (y) coordinates:

$$\begin{aligned} RF(x,y) &= A\left\{\exp\left(-x^2/\sigma_c^2\right) - \frac{\sigma_c}{\sigma_s}\exp\left(-x^2/\sigma_s^2\right)\right\} \\ &\times \exp\left(-y^2/\sigma_y^2\right) \end{aligned} \qquad [4]$$

Equation [4] describes a vertically oriented receptive field, but other orientations (varying in about 15° increments between 0° and 180°) can be produced by coordinate rotation. Based on physiological data, $\sigma_y \approx 3.2\sigma_c$ produces appropriate orientation band-widths. The activation $R(x, y)$ of a cortical array of simple cell receptive fields is produced by convolution:

$$R(x,y) = \iint_{-\infty}^{\infty} RF(x-x', y-y')I(x', y')\mathrm{d}x'\mathrm{d}y' \qquad [5]$$

where $I(x', y')$ represents the distribution of light in the image. Simulated neural responses are then produced by passing the results of eqn [5] through a nonlinear function exhibiting both a threshold and saturation such as the Naka–Rushton function

$$S(R) = \left\{\begin{array}{ll} \frac{MR^N}{\beta^N+R^N} & \text{for } R \geq 0 \\ 0 & \text{for } R < 0 \end{array}\right\} \qquad [6]$$

where M is the maximum response level and N ≈ 2–3 for cortical neurons. As $S(R) = M/2$ when $R = \beta$, β is referred to as the semisaturation constant. Dynamics may now be added to this formulation as follows (Wilson, 1999):

$$\tau\frac{\mathrm{d}E}{\mathrm{d}t} = -E + S(R) \qquad [7]$$

where E is the excitatory firing rate and τ is the time constant for the rate of exponential approach to the asymptotic firing rate.

The simple cell model in eqns [4–7] can be used to generate complex networks of interconnected model neurons for simulations of cortical function (Wilson, 1999). However, contemporary research indicates that orientation selectivity is more complex than this. In particular, the linear filtering in eqn [4] suggests that orientation tuning originates entirely from a linear superposition of nonoriented lateral geniculate (LGN) filters similar to that in eqn [3]. While this was Hubel and Wiesel's (1968) original suggestion for orientation selectivity, the preponderance of evidence now supports additional factors in orientation tuning. First, convergence of as few as two nonoriented LGN filters conveys a modest orientation bias on cortical neurons. This very broad orientation bias is then sharpened considerably by two operations occurring in neural networks at the cortical level: contrast gain control and collinear facilitation (Sompolinski and Shapley, 1997).

Contrast gain controls were first modeled by Heeger (1991) as divisive feedback operations. He suggested that the summed activity of many different oriented units provided a divisive feedback signal to each of the units in turn. The divisive feedback loop in eqn [1] can be altered to function as a cortical gain control as follows:

$$\begin{aligned} \tau\frac{\mathrm{d}E_\theta}{\mathrm{d}t} &= -E_\theta + S\left[\frac{R_\theta}{1+A}\right] = \frac{MR_\theta^N}{\beta_\theta^N(1+A)^N + R_\theta^N} \\ \frac{\mathrm{d}A}{\mathrm{d}t} &= -A + \sum_\theta E_\theta \end{aligned} \qquad [8]$$

where θ designates the preferred orientation of each unit. Substitution of the Naka–Rushton function S from eqn [6] and rearrangement shows that this cortical divisive feedback circuit has the effect of scaling β by $(1+A)$. This network has been implemented and shown to sharpen cortical orientation tuning (Wilson, 1993). Due to the dynamics of the network, this orientation sharpening evolves over time as the divisive feedback takes effect, and this has recently been shown to occur in most cortical neurons (Sompolinsky and Shapley, 1997).

The final contribution to cortical orientation selectivity is provided by collinear facilitation (Sompolinsky and Shapley, 1997). Recent evidence indicates that neurons with the same preferred orientation are interconnected if their receptive fields are situated on a common line defined by their preferred orientation. Such reciprocally linked receptive fields are typically separated by 1–2 receptive field lengths so that their

mutual facilitation will only be triggered by stimulus edges of sufficient length. The excitatory connections are probably mediated by NMDA receptors so that both cells in a collinear pair must be independently activated by the stimulus for facilitation to occur. Obviously, collinear facilitation will sharpen orientation tuning when measured with extended lines or gratings. The function of collinear facilitation is to enhance the salience of extended contours, thus making them prominent candidates for further processing.

Global Processes in Extrastriate Form Vision

The complex cells of Hubel and Wiesel can be readily modeled by spatial pooling of simple cells with the same preferred orientation. More interesting are end-stopped complex cells, which are sensitive to both contour length and curvature. As illustrated in Figure 1(a), an end-stopped complex cell can be modeled by full-wave rectification of simple cell responses followed by second-stage pooling using a 2–3 times larger version of the oriented filter defined by eqn [4]. Full-wave rectification is mathematically equivalent to summing responses from both on-center and off-center simple cells that have the same spatial location and preferred orientation. The first-and second-stage filters in this filter–rectify–filter sequence usually have different preferred orientations, with the orthogonal case being illustrated in Figure 1(a). The optimal line stimulus for this end-stopped unit will obviously be a horizontal line with length equal to the width of the second filter's excitatory zone as shown superimposed on the diagram. In addition to lines of fixed length, end-stopped cells will respond well to curves that are tangent to the preferred orientation of the first-stage simple cells, and they respond well to the orientation of texture boundaries (Wilson, 1993). Although Hubel and Wiesel (1968) reported the presence of end-stopped neurons in V1, subsequent work by von der Heydt and Peterhans (1989) indicates that they predominate in the second visual area, V2.

Following V1 and V2 the cortical form vision pathway projects to V4, and finally to inferior temporal cortex (IT). Mean receptive field dimensions increase by a factor of approximately 2.5–3 from area to area within this pathway, reaching diameters of 10°–30° in IT. It was noted above that the receptive fields of end-stopped neurons must be approximately 2–3 times larger in each dimension than the simple cell receptive fields that provide their inputs in order to function optimally. Thus, the increasing receptive field size throughout the form vision pathway indicates that pooling operations are combining information over increasingly large areas. The nature of pooling in V4 was first investigated by Gallant et al. (1993). They studied macaque V4 responses to concentric and radial gratings as well as to conventional sine wave gratings. One subclass of neurons responded best to concentric gratings, while a second subclass responded best to radial gratings. Although many neurons

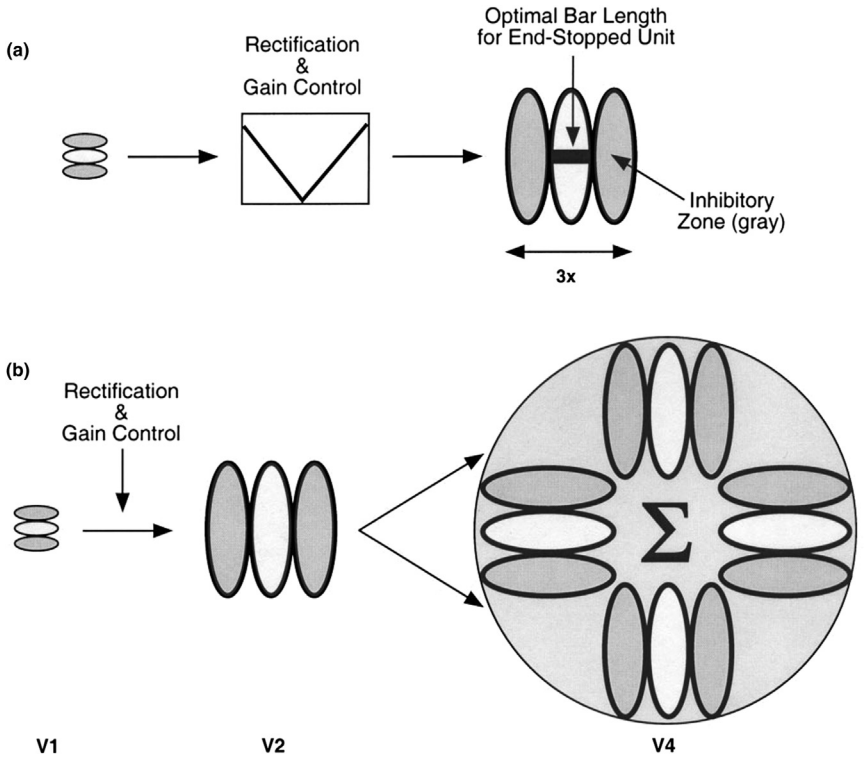

Figure 1 Models of (a) an end-stopped complex cell and (b) a V4 concentric unit. The end-stopped cell (a) is produced by oriented simple cell filtering followed by rectification and subsequent filtering at a different orientation. The optimal bar length and orientation for stimulation of this unit is shown. The V4 concentric model in (b) builds upon the hierarchical theme in (a) by adding a final stage of configural pooling in which responses of a concentric array of end-stopped units are summed. The three stages of the V4 model are hypothesized to occur in V1, V2, and V4.

responded to conventional sine wave gratings, few were optimized for these patterns.

Humans are extremely good at detecting concentric structure in glass patterns, and this results from concentric summation of orientation information over considerable regions (Wilson et al., 1997). A glass pattern is produced by randomly positioning pairs of signal dots such that each pair is oriented tangentially to the desired concentric structure. This produces a random dot pattern appearing to contain concentric swirls of dots. Furthermore, fMRI studies have shown that human V4 is more sensitive to concentric or radial gratings than to conventional gratings (Wilkinson et al., 2000). A quantitative model for V4 concentric receptive fields is depicted in **Figure 1**(b) (Wilson et al., 1997). The initial two stages of the model are oriented simple cells followed by rectification and orthogonal second stage filtering. As mentioned above, these model end-stopped complex cells map onto V1 and V2 physiology, and they respond to curved contours that are locally tangential to the orientation of the simple cell filter. The final stage of the V4 model consists of linear summation (Σ) of concentrically arranged curvature information from V2. This represents an example of configural pooling, as each curvature response is pooled only if it arises in a spatial subregion of the receptive field consistent with the concentric global pattern. The final V4 receptive field size is approximately three times the diameter of its V2 subunits, thus conforming to the observed receptive field size increase in the form vision pathway. Similar V4 configural pooling models may be constructed to extract radial and other structures from the stimulus, and one challenge to future research will be an enumeration of configural pooling mechanisms.

What might be the function of V4 concentric units? Several observations suggest that they may be involved in the perception of faces and other ellipsoidal shapes. First, computer simulations have shown that model V4 units can detect faces in complex scenes (Wilson et al., 2000). Second, fMRI studies have shown that the face-selective area in the fusiform gyrus responds not only to faces but also to circular or ellipsoidal contours (Wilkinson et al., 2000), thus supporting the hypothesis that analysis of head shape is a major ingredient in face perception.

Perceptual Oscillations

There are many instances of instabilities resulting in perceptual oscillations, two of the most common being the Necker cube and binocular rivalry. A third example is the Marroquin illusion in which a static, periodic dot pattern triggers the percept of flashing illusory circles (see Wilson et al., 2000). A Marroquin pattern is constructed by superimposing three square grids of dots, each rotated $\pm$ 60° with respect to one another. It is natural to conjecture that competitive interactions among V4 concentric units might be the cause of the illusory circles triggered by this stimulus. Indeed, there is physiological evidence that competition among V4 neurons forms a basis for spatial and object based selective attention (Reynolds et al., 1999).

A neural network model incorporating spatially regional competitive interactions among V4 concentric units can explain the illusory flashing Marroquin circles (Wilson et al., 2000). In this simulation V4 neural responses were represented by a variant of eqn [7]:

$$\tau_E \frac{\mathrm{d}E_n}{\mathrm{d}t} = -E_n + \frac{100\,P_+^2}{(\beta + H_n)^2 + P_+^2}$$

where

$$P = S_{\text{Marroquin}} - k \sum_{n \neq k} I_k \exp\left(-R_{nk}^5/\sigma^5\right) \qquad [9]$$

$$\tau_I \frac{\mathrm{d}I_n}{\mathrm{d}t} = -I_n + E_n$$

$$\tau_H \frac{\mathrm{d}H_n}{\mathrm{d}t} = -H_n + gE_n$$

In these equations the input P to each neuron is the response of the V4 model in **Figure 1**(b) to the Marroquin pattern, $S_{\text{Marroguin}}$ minus a spatially weighted sum of inhibitory activity I_k generated by neighboring V4 neurons, the inhibitory neurons being driven by the E cells as described by the second equation. Parameters k and σ are respectively the gain and space constant of the inhibitory region. Mathematical analysis of equations such as these shows that one further ingredient is necessary for the network to produce limit cycle oscillations: the firing rate of active excitatory neurons must decline over time (Wilson, 1999). Excitatory cortical neurons are known to contain slow hyper-polarizing currents that cause the firing rate to drop about threefold over the course of several hundred milliseconds, and this is incorporated into the mathematical model via the H or hyperpolarizing variable (Wilson, 1999). The effect of this variable, governed by g, is to increase the semisaturation constant of any active E neuron, thereby reducing the firing rate. Simulations of this network provide a good quantitative explanation of the illusory flashing circles generated by Marroquin patterns (Wilson et al., 2000). Extensions of this work suggest that models analogous to eqn [9] can also explain oscillations and traveling wave phenomena in binocular rivalry.

Conclusions and Future Directions

The low-level vision theories developed here are based on nonlinear differential equations describing firing rates (rather than individual spikes) of retinal and cortical neurons. Large-scale neural networks incorporating these principles can now be used to simulate visual processing from the retina up to V4. Although this approach certainly represents but a first-order approximation to the underlying neural activity, it nevertheless provides a solid foundation for further progress.

Many issues, of course, remain to be addressed. Key among these is the ubiquitous cortical feedback: IT back to V4, V4 back to V2, etc. One conjecture is that these connections mediate a rapid dynamical optimization of the processing circuitry based upon initial feed-forward estimates of the stimulus being analyzed: a sophisticated example of a self-organizing system (see also Self-Organizing Dynamical Systems). A second likelihood is that this feedback mediates both top-down influences on selective attention and figure-ground segmentation, which appears to involve extra-striate feedback to V1 (Lamme and Roelfsema, 2000). The future promises a cooperative interaction between theory and experiment in elucidating these and many other exciting issues.

See also: Vision, High-Level Theory of; Visual Perception, Neural Basis of.

Bibliography

Gallant, J.L., Braun, J., VanEssen, D.C., 1993. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science 259, 100–103.

Heeger, D.J., 1991. Nonlinear model of neural responses in cat visual cortex. In: Landy, M., Movshon, J.A. (Eds.), Computational Models of Visual Processing. MIT Press, Cambridge, MA, pp. 119–133.

von der Heydt, R., Peterhans, E., 1989. Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. Journal of Neuroscience 9, 1731–1748.

Hood, D.C., 1998. Lower level visual processing and models of light adaptation. Annual Reviews in Psychology 49, 503–535.

Hubel, D.H., Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology 195, 215–243.

Lamme, V.A.F., Roelfsema, P.R., 2000. The distinct modes of vision offered by feedforward and recurrent processing. TINS 23, 571–577.

Reynolds, J.H., Chelazzi, L., Desimone, R., 1999. Competitive mechanisms subserve attention in macaque areas V2 and V4. Journal of Neuroscience 19, 1736–1753.

Sompolinsky, H., Shapley, R., 1997. New perspectives on the mechanisms for orientation selectivity. Current Biology 7, 514–522.

Wässle, H., Boycott, B.B., 1991. Functional architecture of the mammalian retina. Physiology Reviews 71, 447–480.

Wilkinson, F., James, T.W., Wilson, H.R., Gati, J.S., Menon, R.S., Goodale, M.A., 2000. An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings. Current Biology 10, 1455–1458.

Wilson, H.R., 1993. Nonlinear processes in visual pattern discrimination. Proceedings of the National Academy of Sciences USA 90, 9785–9790.

Wilson, H.R., 1997. A neural model of foveal light adaptation and after image formation. Vision Neuroscience 14, 403–423.

Wilson, H.R., 1999. Spikes, Decisions, and Actions: Dynamical Foundations of Neuroscience. Oxford University Press, Oxford, UK.

Wilson, H.R., Krupa, B., Wilkinson, F., 2000. Dynamics of perceptual oscillations in form vision. Nature Neuroscience 3, 170–176.

Wilson, H.R., Wilkinson, F., Asaad, W., 1997. Concentric orientation summation in human form vision. Vision Research 37, 2325–2330.

Visual Imagery

Paolo Bartolomeo, Brain and Spine Institute, Groupe Hospitalier Pitié-Salpêtrière, Paris, France; Fédération de Neurologie, Hôpital Pitié-Salpêtrière, Paris, France; and Catholic University, Milan, Italy

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Visual mental imagery is the faculty whereby we can 'visualize' an object from memory. This article summarizes how damage to various brain areas can impair or not different visual perceptual or imaginal abilities. Whereas damage to occipital visual areas (striate and extrastriate) can of course produce impairments on several visual perceptual abilities, it does not seem either necessary or sufficient to impair visual mental imagery. On the other hand, damage to more anterior areas in the temporal lobe can impair visual imagery for objects, color, and orthographic material. More dorsal loci of brain damage, perhaps coupled with disconnection between the right and the left hemisphere, can induce neglect for the left part of visual mental images. Thus, there seems to be a partial overlapping between the neural correlates of high-level (but not low-level) visual perception and visual mental imagery abilities. This partial overlapping of neural resources may reflect essential differences between perception and imagery, which might be related to Sartre's distinction between more immediate and more reflective forms of consciousness.

Visual Mental Imagery and Brain Damage

Try to imagine a leopard and establish whether it has round or pointed ears. How can we revisualize an object, a face, or a color from memory? There is currently much debate on the psychological and neural aspects supporting these abilities, but we know that brain damage can disrupt them and produce particular patterns of dissociations between perceptual and imagery abilities. The first reported case of a relatively isolated deficit of mental imagery came under Charcot's observation in the Salpêtrière Hospital in Paris (Charcot and Bernard, 1883). M. X used to experience particularly vivid visual mental images for object forms and colors, which he described as being precise and glowing as reality itself. He used his ability as a mnemotist, until one day he found himself to be completely unable to recall forms and colors from memory. For example, he lost his ability to produce detailed drawings from memory, or to precisely recall poems and letters by 'reading' them in his 'mind's eye.' This deficit was relatively selective, being accompanied by only minor problems in reading and a slight decrease in color perception. There is no record concerning the potential presence of a brain lesion in M. X, but his reading problems in the absence of more elementary visual deficit might suggest left temporal damage with sparing of the occipital cortex (Bartolomeo, 2008). However, more recently Zago et al. (2011) have suggested that loss of imagery in the case of M. X may in fact have been of psychiatric origin.

Other neuropsychological studies found a frequent association of perceptual and imagery deficits in brain-damaged patients (Farah, 1988). Together with results of neuroimaging experiments in normal subjects, this evidence prompted the conclusion that visual mental imagery, much as visual perception, depended on the functioning of retinotopic areas in the occipital lobe, including V1 (Kosslyn et al., 2006). However, since the 1990s other patient studies showing double dissociations between perception and imagery abilities (reviewed by Bartolomeo, 2002) have produced a strong challenge to the perception-imagery equivalence hypothesis.

Visual Mental Imagery and the Striate Cortex

Hypotheses postulating a large anatomo-functional overlapping between visual perception and visual mental imagery assign a central role to V1 in imagery abilities (Kosslyn et al., 2006). If this is true, then patients with cortical blindness resulting from V1 damage should lose the ability to revisualize items from memory. However, neuropsychological studies have provided abundant evidence challenging this hypothesis. For example, Chatterjee and Southwood (1995) studied three cortically blind patients. Two of these patients had bilateral medial occipital damage but showed nevertheless an intact capacity for imagining object forms. In contrast, the third patient had damage extending well forward the left temporal lobe and was unable to imagine object forms. More recently, another patient was described with complete and permanent cortical blindness after bilateral damage to V1, who had perfectly preserved visual mental imagery for objects, colors, and spatial relationships (Zago et al., 2010). In contrast, another cortically blind patient, who had signs of brain hypometabolism not only in the occipital cortex, but also extending to the inferior and mesial temporal cortex, found himself unable to perform any tasks based on the visual appearance of places, symbols, objects, and animals (Policardi et al., 1996). Recently, Bridge et al. (2011) described a patient with a lesion restricted to the gray matter of the calcarine sulcus (V1). Despite near-complete cortical blindness, this patient had vivid visual mental imagery. This patient also underwent functional neuroimaging (fMRI), which demonstrated a pattern of cortical activation during visual mental imagery in occipitotemporal extrastriate areas, and in parietal and frontal regions, which was indistinguishable from normal participants. Altogether, this evidence is compelling in showing that lesions restricted to V1 do not typically produce deficits of visual mental imagery.

Another prediction resulting from the perception-imagery equivalence hypothesis is that *unilateral* occipital lesions producing lateralized visual field defects such as lack of vision in one side of the visual fields (homonymous hemianopia)

should determine corresponding lateralized deficits in visual mental imagery; i.e., impaired imagery on the blind side of the visual field. Butter et al. (1997) reported results consistent with this prediction in eight patients with unilateral field defects, whose recall of the position of previously presented dots was worse on the blind side than on the normal side. The authors concluded that visual imagery involves topographically organized visual areas of the occipital lobe. However, there are several problems with this conclusion (Bartolomeo, 2002), including the presence of lesions extending beyond the occipital cortex in at least two of the patients. A recent study on 12 patients with homonymous hemianopia (six with right-sided and six with left-sided hemianopia) (Dulin et al., 2011) challenged more directly the conclusions of the Butter et al.'s (1997) study. Using a similar test paradigm as the one employed by Butter et al. (1997), Dulin et al. (2011) found evidence for normal lateralized imagery in all the tested patients.

Visual Mental Imagery and Extrastriate Visual Areas

Even if visual mental imagery does not require the integrity of low-level visual areas such as V1, it could well rely on visual areas at a higher level of processing in the ventral cortical visual stream (Milner and Goodale, 1995; Mishkin et al., 1983). However, the precise level of visual processing putatively important for visual mental imagery remains uncertain. Bartolomeo et al. (1998a) described a striking dissociation between impaired perception for high-level visual attributes and preserved imagery in a patient (Madame D) with bilateral extrastriate lesions, at the borders between the occipital and the temporal lobes. After the occurrence of her lesions, Madame D had severely impaired visual perception for object form (visual agnosia), words and letters (alexia without agraphia), colors (cerebral achromatopsia), and faces (prosopagnosia). However, her visual mental imagery for the same visual entities was perfectly preserved. She produced plausible drawings from memory, while being unable to identify her drawings on subsequent testing. Madame D had thus intact mental imagery for the very same visual entities that she could not perceive. More recently, Moro and collaborators (2008) made a detailed report of the complementary dissociation in two patients with traumatic brain injury, who had impaired visual mental imagery with reasonably preserved perception. In both cases, there was an involvement of the left temporal lobe in the absence of visible damage to the occipital cortex.

These and many other case reports (Bartolomeo, 2002) challenge the hypothesis that visual occipital areas are necessary to visual mental imagery, and suggests instead that extensive damage to the left temporal lobe is crucial to observe deficits of visual mental imagery, at least for object shape and color and for orthographic material. Since this conclusion was proposed (Bartolomeo, 2002, 2008), it has received further support from several more recent studies of patient with occipital damage and cerebral blindness (Bridge et al., 2011) or homonymous hemianopia (Dulin et al., 2011). An interim conclusion at this point could thus be that regions in the occipital cortex are not necessary to visual mental imagery abilities; however, the possibility remains open that areas in more anterior locations along the ventral visual processing streams, such as the temporal lobe, might be important for these abilities.

Domains of Visual Mental Imagery: Color, Words, and Space

Visual mental imagery is not a unitary, undifferentiated faculty. Neuropsychological evidence suggests the possibility to distinguish at least five kinds of visual entities whose imagery can be independently affected by brain damage: shapes of objects, colors of objects, faces, orthographic material, and spatial relationships (Goldenberg, 1993). We will present here a selection of evidence concerning the neuropsychology of mental imagery for three of these domains: color, orthographic material, and spatial arrays.

Visual Mental Imagery of Color

Color processing is a privileged domain for the study of visual imagery abilities, because it has more chances of requiring genuinely visual processes than the other domains, whose spatial characteristics may authorize the use of nonvisual strategies (Farah, 1988). After several early reports of patients with an association of impaired color perception resulting from brain lesions (cerebral achromatopsia) and loss of color imagery (Damasio, 1994), further case descriptions pointed out the possibility of normal color imagery in achromatopsic patients, and of normal color vision with loss of color imagery (Bartolomeo, 2002). For example, soon after her strokes Madame D (Bartolomeo et al., 1997) complained to see the world in shades of gray or, later on, in reddish-brown nuances, but could form vivid mental images of colors and even perform subtle hue discriminations between imagined colors. Miceli and coworkers (2001) reported a detailed description of patients PCO and IOC, with normal color perception but impaired color knowledge. These patients could not attribute the correct color to objects, or select the correct crayon to color them. Of note, IOC, but not PCO, had apparently normal knowledge of the form and function of objects. Although the issue of mental imagery was not explicitly addressed in this study, the pattern of performance of these two patients strongly suggests a selective deficit of visual mental imagery for colors for IOC, and an impairment of color and form imagery for PCO. PCO had suffered from encephalitis resulting in bilateral lesions (more extensive on the left than on the right) of the temporal pole and the inferolateral temporal regions (middle and inferior temporal gyri), prefrontal convexity, and orbital gyri. Damage was less extensive in the parietal lobe and the superior-rostral insula. Both hippocampi were atrophic. IOC had lesions in the calcarine cortex, the lingual gyrus, the isthmus and the adjacent cingulate gyrus, the parahippocampal gyrus and the rostro-medial part of the fusiform gyrus, including part of the ventral portion of the temporal pole, with left hippocampal atrophy, as a consequence of a stroke in the territory of the left posterior cerebral artery. Interestingly, a control patient reported in the same study, SLA, who had more posterior lesions in the left

hemisphere involving the posterior part of the lingual gyrus, the cuneus, and the calcarine cortex, with largely spared temporal cortex, had right homonymous hemianopia but no cognitive impairment. In particular, SLA had normal knowledge of colors. Thus, an extensive damage of the left temporal lobe seems both necessary and sufficient to produce deficits of color imagery.

Visual Mental Imagery of Orthographic Material

Literacy induces peculiar modifications of the visual system (Dehaene et al., 2010). In literate people, left posterior brain damage can disrupt the processing of visually presented letters and words in the relative absence of other deficits (and notably, with preserved writing), a disorder known as alexia without agraphia or pure alexia (Bachoud-Lévi and Bartolomeo, 2003; Dejerine, 1892). We have already mentioned the case of Madame D, who, despite being profoundly alexic and incapable to identify even single letters without errors (Bartolomeo et al., 1998b), could perform at ceiling difficult tests concerning the visual appearance of letters and words (Bartolomeo et al., 1998a). For example, she could easily answer questions such as: imagine the letter M. Cut it into half vertically. Drop off the right side and turn the remaining part upside down (the answer is V). A similar dissociation between impaired perception and preserved imagery also held for numbers in this patient. For example, Madame D was unable to recognize visually the number 6, but described it as 'a half circle open toward the right, making a loop at its bottom.' Complementary evidence concerning the neural bases of visual knowledge of orthographic material came from the description of another pure alexic patient, VSB. In addition to his reading deficit, VSB was found to suffer from a selective deficit for visual mental imagery of orthographic material, or loss of letter shape knowledge (Bartolomeo et al., 2002). In this patient, an extensive left temporal–parietal lesion severely impaired reading, whereas writing was relatively preserved. At variance with other pure alexic patients described in the literature, VSB claimed to have become unable to mentally visualize letters and words. Indeed, his performance on a series of tests tapping visual mental imagery for orthographic material was severely impaired. In the present framework, the extensive left temporal damage suffered by VSB may account for his mental imagery deficit. However, allowing VSB to trace each item with his finger dramatically ameliorated performance on the same tests. Visual mental imagery for nonorthographic items was comparatively spared. The pattern of dissociation shown by VSB between impaired visual mental imagery and relatively preserved motor-based knowledge for orthographic material lends support to the view that separate codes, respectively, based on visual appearance and on motor engrams, may be used to access knowledge of the visual form of letters and words.

Damage to domain-specific regions (Kanwisher, 2000) or circuits (Epelbaum et al., 2008) in the ventral temporal lobe might thus contribute to the occurring of selective deficits of distinct domains of visually based knowledge. However, it remains uncertain whether there is a strict overlapping of perceptual and imagery functions within the same high-level visual regions in the temporal lobe (Rich, et al., 2006), or whether nearby but distinct regions are important for these different abilities.

Imaginal Neglect

A particular class of deficits of spatial imagery concerns the impaired ability to conjure up or produce a complete and symmetrical description of an imagined scene. Patients with damage to the right hemisphere may behave as if the left part of the world does not exist anymore, thus showing signs of left visuospatial neglect (Bartolomeo, 2007; Halligan and Bartolomeo, 2012). About a third of these patients may also show neglect for the left part of their mental images (Bartolomeo et al., 1994). These patients may omit to mention the left side of the mental space when describing places from memory, thus demonstrating 'imaginal' neglect (Bisiach and Luzzatti, 1978). Not all patients with visual neglect show imaginal neglect. A possibility to explain this dissociation is that an important aspect of neglect is that patients' attention is 'captured' by right-sided objects and cannot easily disengage from these objects to explore the remaining space; however, imagined details may have less attention-capturing power than real ones, thus decreasing the possibility to observe imaginal neglect (Bartolomeo et al., 1994). Nevertheless, imaginal neglect can also occur in the absence of signs of perceptual neglect, either at the onset or, perhaps more commonly, as a result of selective compensation for the perceptual aspects of the syndrome. Patients often learn with time (and possibly the help of people around them) to explore more thoroughly their visual environment. However, compensation may be more difficult to obtain in the more abstract imaginal domain, which is rarely the object of rehabilitation or of more informal reminders to 'look to your left' (Bartolomeo and Chokron, 2001). Thus, similar to other domains of visual mental imagery, several studies have reported the existence of double dissociations between imaginal and perceptual neglect (Anderson, 1993; Beschin et al., 1997; Coslett, 1997; Guariglia et al., 1993; and Ortigue et al., 2001).

However, the study of imaginal neglect raises peculiar methodological problems. Often, very different task are used to evaluate spatial perception and spatial imagery. In particular, in several studies paper-and-pencil tests were used to test perceptual neglect and description from memory was used to test imaginal neglect (Bartolomeo et al., 1994; Rode and Perenin, 1994; Rode et al., 2007). Moreover, description from memory might rely more on verbal semantic memory than on visual imagery, and thus produce symmetrical descriptions even in the presence of imaginal neglect (Rode et al., 2004). To encourage the use of a visual mental strategy, a response time 'geographical' test was devised (Bartolomeo et al., 2005), with strictly comparable perceptual and imaginal conditions (Bourlon et al., 2008, 2011). In different tasks, participants either saw towns/regions on a map of France or heard their names, and pressed one of two keys according to the stimulus location (left or right of Paris). The results obtained with these tasks confirmed the rarity of imaginal neglect with respect to perceptual neglect.

In a recent case report of imaginal neglect (Rode et al., 2010), structural and diffusion MRI demonstrated damage to

several white matter tracts in the right hemisphere and to the splenium of corpus callosum (the principal connection between the two brain hemispheres). The same study reported on a second right-brain damaged patient, who showed signs of perceptual but not imaginal neglect, and had damage to the same intrahemispheric tracts; the callosal connections, however, were spared. Imaginal neglect might thus result from the association of frontoparietal dysfunction, which impairs orienting toward left-sided items (Bartolomeo et al., 2007) and additional posterior callosal disconnection, which prevents the symmetrical processing of spatial information from long-term memory.

In clinical settings, drawing from memory is often used to assess imaginal abilities and then directly compared to drawing copying. However, the visual feedback in drawing test may influence the patient by inducing an attentional capture of the right-sided details the patient had just drawn (Chokron et al., 2004). To address this issue, drawing without visual feedback was proposed, e.g., while blindfolded (Chokron et al., 2004) or using a pen that leaves no visible traces on the sheet (Cristinzio et al., 2009). While in general patients demonstrate more neglect with visual feedback than without visual feedback (Chokron et al., 2004), thus confirming the attention-capturing effect of right-sided visual items (Bartolomeo et al., 1994), one recent case report (Cristinzio et al., 2009) showed the opposite effect: the patient had more neglect without than with visual feedback, perhaps as a consequence of concomitant impairment of working memory, which prevented the patient from remembering what he had already drawn. To summarize, one possibility to account for the rarity of imaginal neglect is that this form of neglect might depend on deficits of top-down processes, such as endogenous attention or active rehearsal of spatial knowledge, that are less impaired than exogenous attention in patients with perceptual neglect (Bourlon et al., 2011).

Assessment of Visual Mental Imagery

How to test for visual mental imagery abilities in brain-damaged patients? The usual strategy, borrowed from experimental psychology studies in normal individuals, is to ask patients to perform tasks that most likely cannot be resolved with simple access to (nonvisual) semantic information. For example, when exploring visual knowledge of colors, one can ask questions about the typical color of objects, with the precaution of avoiding items whose color can be retrieved by using nonvisual associations (e.g., red–blood, white–snow) (Bartolomeo et al., 1997; Beauvois and Saillant, 1985). A recent case report (Zeman et al., 2010), however, raises the disquieting possibility that at least some of these tasks can be correctly performed in the absence of the corresponding 'quasi-visual' phenomenological experience, which should be typical of visual mental imagery. These authors described a patient who performed normally on mental imagery tasks while negating to experience visual mental images. During attempted imagery, this patient, who had no visible brain lesion, showed reduced fMRI activation in a network of posterior regions, while activity in frontal regions was increased compared to controls. A possible conclusion could be that V1 activity is crucial to the phenomenological qualities of mental imagery. Against this conclusion, however, the recently described patient with V1 damage mentioned before (Bridge et al., 2011) did report experiencing vivid visual images of faces and houses, while demonstrating extrastriate and frontal activations on fMRI.

"The Batterie Imagination et Perception" (BIP)

A standardized paper-and-pencil battery of tests has recently been developed in order to assess perceptual and imaginal deficits in patients with brain damage (BIP, Bourlon, et al., 2009). The BIP, which currently exists in French (Bourlon et al., 2009) and Italian language (Antonietti et al., 2008) versions, examines distinct domains of visual mental imagery and visual perception abilities such as object form and color, animals, orthographic material, numbers, faces, and spatial relationships. It comprises two parts: one composed of 14 tests assessing mental imagery, another composed of eight tests assessing various visual perceptual abilities. The BIP was designed to investigate associations and dissociations between perceptual and imaginal deficits in patients and thus provide evidence relevant to the debate concerning the neural bases of visual mental imagery. Here follows a short summary of the tests used.

The imagery section of the BIP consists of tests of animal comparison (participants mentally compare the size of an animal to that of a goat) (Kosslyn et al., 1995); mental representation of physical details (establish whether the ears of a determined animal are round or sharp) (Sheridan and Humphreys, 1993); object comparison (determine from memory which of the two items named by the examiner is bigger); morphological discrimination (compare from memory pairs of objects, animals, or fruits that are morphologically similar, and mention the difference in their visual appearance); mental representation of objects (recall from memory the typical shape and the orientation of objects); recall of object color (recall the typical color of objects mentioned by the examiner) (Bartolomeo et al., 1997); color comparison (determine which of the two objects mentioned by the examiner is lighter in color) (De Vreese, 1991); mental representation of numbers (decide whether numbers mentioned by the examiner have closed loops (e.g., 8) or not (e.g., 2)) (Coltheart et al., 1975); mental construction of letters (identify upper-case letters after having mentally built them following oral instructions) (Behrmann et al., 1994); mental representation of letters (in the first part of this test, participants have to imagine lower case letters named by the examiner and decide whether they appear in the space line, exceed above space line or beneath space line (Weber and Castleman, 1970); in the second part, participants have to state from memory whether upper-case letters named by the examiner contain any curved lines or not) (Coltheart et al., 1975); mental representation of words (decide if words named by the examiner contain one or more letters with an ascender (t, l, d), a descender (j, p, y), or neither (m, e)) (Weber and Castleman, 1970); mental representation of faces (answer questions about the shape of the nose (thin–large), and the

shape of the lips (thin–fleshy) of famous faces) (Bartolomeo et al., 1998a); mental representation of spatial relationships (evaluate the position of imagined dots) (Butter et al., 1997); and modified o'clock test (establish the position of the hands of an imagined clock for a verbally given hour) (Grossi et al., 1993). The perception section of the BIP consists of analogous tests, visually presented stimuli.

Conclusion

Mental imagery abilities draw on particular sorts of conscious processes, whose phenomenological quality can be distinguished from visual perception (Sartre, 1936), and is closer to visual memories. For example, you can be surprised by a visual perception, but not by the mental image that you conjured up (Sartre, 1940). Another crucial difference between mental images and their 'real' counterparts is that mental images are based on active processes of reflective consciousness, as opposed to the more immediate character of perception. As a consequence, imagery abilities make us ontologically free from the constraints of reality (Sartre, 1940), as well as from its influence on elementary sensory processing (Mesulam, 1998). In this sense, mental imagery is a central feature of our being human. Lesion neuropsychology provides key causal evidence to establish its neural underpinnings.

See also: Attention, Neural Basis of; Hemispatial Neglect, Neural Basis of; Visual Memory, Psychology of.

Bibliography

Anderson, B., 1993. Spared awareness for the left side of internal visual images in patients with left-sided extrapersonal neglect. Neurology 43, 213–216.

Antonietti, A., Bartolomeo, P., Colombi, A., Incorpora, C., Oliveri, S., 2008. Rapporti tra capacità percettive e immaginative nell'anziano. Ricerche di Psicologia 1-2, 105–120.

Bachoud-Lévi, A.C., Bartolomeo, P., 2003. Mechanisms of pure alexia: spatially-based impairment, letter identification deficit, or both? Neurocase 9 (2), 164–176.

Bartolomeo, P., 2002. The relationship between visual perception and visual mental imagery: a reappraisal of the neuropsychological evidence [Review]. Cortex 38 (3), 357–378.

Bartolomeo, P., 2007. Visual neglect. [Review]. Current Opinion in Neurology 20 (4), 381–386.

Bartolomeo, P., 2008. The neural correlates of visual mental imagery: an ongoing debate. [Review]. Cortex 44 (2), 107–108.

Bartolomeo, P., Chokron, S., 2001. Levels of impairment in unilateral neglect. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, second ed., vol. 4. Elsevier Science Publishers, Amsterdam, pp. 67–98.

Bartolomeo, P., D'Erme, P., Gainotti, G., 1994. The relationship between visuospatial and representational neglect. Neurology 44, 1710–1714.

Bartolomeo, P., Bachoud-Lévi, A.C., Denes, G., 1997. Preserved imagery for colours in a patient with cerebral achromatopsia. Cortex 33 (2), 369–378.

Bartolomeo, P., Bachoud-Lévi, A.C., de Gelder, B., Denes, G., Dalla Barba, G., Brugières, P., et al., 1998a. Multiple-domain dissociation between impaired visual perception and preserved mental imagery in a patient with bilateral extrastriate lesions. Neuropsychologia 36 (3), 239–249.

Bartolomeo, P., Bachoud-Lévi, A.C., Degos, J.D., Boller, F., 1998b. Disruption of residual reading capacity in a pure alexic patient after a mirror-image right hemispheric lesion. Neurology 50 (1), 286–288.

Bartolomeo, P., Bachoud-Lévi, A.C., Chokron, S., Degos, J.D., 2002. Visually- and motor-based knowledge of letters: evidence from a pure alexic patient. Neuropsychologia 40 (8), 1363–1371.

Bartolomeo, P., Bachoud-Levi, A.C., Azouvi, P., Chokron, S., 2005. Time to imagine space: a chronometric exploration of representational neglect. Neuropsychologia 43 (9), 1249–1257.

Bartolomeo, P., Thiebaut de Schotten, M., Doricchi, F., 2007. Left unilateral neglect as a disconnection syndrome. Cerebral Cortex 45 (14), 3127–3148.

Beauvois, M.F., Saillant, B., 1985. Optic aphasia for colours and colour agnosia: a distinction between visual and visuo-verbal impairments in the processing of colours. Cognitive Neuropsychology 2 (1), 1–48.

Behrmann, M., Moscovitch, M., Winocur, G., 1994. Intact visual imagery and impaired visual perception in a patient with visual agnosia. Journal of Experimental Psychology: Human Perception and Performance 20 (5), 1068–1087.

Beschin, N., Cocchini, G., Della Sala, S., Logie, R., 1997. What the eyes perceive, the brain ignores: a case of pure unilateral representational neglect. Cortex 33, 3–26.

Bisiach, E., Luzzatti, C., 1978. Unilateral neglect of representational space. Cortex 14, 129–133.

Bourlon, C., Pradat-Diehl, P., Duret, C., Azouvi, P., Bartolomeo, P., 2008. Seeing and imagining the "same" objects in unilateral neglect. Neuropsychologia 46 (10), 2602–2606.

Bourlon, C., Chokron, S., Bachoud-Levi, A.C., Coubard, O., Bergeras, I., Moulignier, A., et al., 2009. [Presentation of an assessment battery for visual mental imagery and visual perception]. Revue Neurologique (Paris) 165 (12), 1045–1054.

Bourlon, C., Duret, C., Pradat-Diehl, P., Azouvi, P., Loeper-Jeny, C., Merat-Blanchard, M., et al., 2011. Vocal response times to real and imagined stimuli in spatial neglect: a group study and single-case report. Cortex 47 (5), 536–546.

Bridge, H., Harrold, S., Holmes, E.A., Stokes, M., Kennard, C., 2011. Vivid visual mental imagery in the absence of the primary visual cortex. Journal of Neurology.

Butter, C.M., Kosslyn, S.M., Mijovic-Prelec, D., Riffle, A., 1997. Field-specific deficits in visual imagery following hemianopia due to unilateral occipital infarcts. Brain 120, 217–228.

Charcot, J.M., Bernard, D., 1883. Un cas de suppression brusque et isolée de la vision mentale des signes et des objets (formes et couleurs). Le Progrès Médical 11, 568–571.

Chatterjee, A., Southwood, M.H., 1995. Cortical blindness and visual imagery. Neurology 45, 2189–2195.

Chokron, S., Colliot, P., Bartolomeo, P., 2004. The role of vision in spatial representation. Cortex 40 (2), 281–290.

Coltheart, M., Hull, E., Slater, D., 1975. Sex differences in imagery and reading. Nature 253, 434–440.

Coslett, H.B., 1997. Neglect in vision and visual imagery: a double dissociation. Brain 120, 1163–1171.

Cristinzio, C., Bourlon, C., Pradat-Diehl, P., Trojano, L., Grossi, D., Chokron, S., et al., 2009. Representational neglect in "invisible" drawing from memory. Cortex 45 (3), 313–317.

Damasio, A.R., 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Publishing, New York.

De Vreese, L.P., 1991. Two systems for colour-naming defects: verbal disconnection vs colour imagery disorder. Neuropsychologia 29 (1), 1–18.

Dehaene, S., Pegado, F., Braga, L.W., Ventura, P., Nunes Filho, G., Jobert, A., et al., 2010. How learning to read changes the cortical networks for vision and language. Science 330 (6009), 1359–1364.

Dejerine, M.J., 1892. Contribution à l'étude anatomo-pathologique et clinique des différentes variétés de cécité verbale (Contribution to the pathological and clinical investigation of different varieties of word blindness). Mémoires de la Société de Biologie 4, 61–90.

Dulin, D., Cavezian, C., Serrière, C., Bachoud-Levi, A.-C., Bartolomeo, P., Chokron, S., 2011. Colour, face and visuo-spatial imagery abilities in low-vision individuals with visual field deficits. The Quarterly Journal of Experimental Psychology 64, 1955–1970.

Epelbaum, S., Pinel, P., Gaillard, R., Delmaire, C., Perrin, M., Dupont, S., et al., 2008. Pure alexia as a disconnection syndrome: new diffusion imaging evidence for an old concept. Cortex 44 (8), 962–974.

Farah, M.J., 1988. Is visual imagery really visual? Overlooked evidence from neuropsychology. Psychological Review 95 (3), 307–317.

Goldenberg, G., 1993. The neural basis of mental imagery. Baillière's Clinical Neurology 2 (2), 265–286.

Grossi, D., Angelini, R., Pecchinenda, A., Pizzamiglio, L., 1993. Left imaginal neglect in heminattention: experimental study with the O'clock test. Behavioural Neurology 6, 155–158.

Guariglia, C., Padovani, A., Pantano, P., Pizzamiglio, L., 1993. Unilateral neglect restricted to visual imagery. Nature 364, 235–237.

Halligan, P.W., Bartolomeo, P., 2012. Visual neglect. In: Ramachandran, V. (Ed.), Encyclopedia of Human Behavior, second ed., vol. 3. Academic Press, pp. 652–664.

Kanwisher, N., 2000. Domain specificity in face perception. Nature Neuroscience 3 (8), 759–763.

Kosslyn, S.M., Thompson, W.L., Kim, I.J., Alpert, N.M., 1995. Topographical representations of mental images in primary visual cortex. Nature 378, 496–498.

Kosslyn, S.M., Thompson, W.L., Ganis, G., 2006. The Case for Mental Imagery. Oxford University Press, New York.

Mesulam, M.M., 1998. From sensation to cognition. Brain 121 (Pt 6), 1013–1052.

Miceli, G., Fouch, E., Capasso, R., Shelton, J.R., Tomaiuolo, F., Caramazza, A., 2001. The dissociation of color from form and function knowledge. Nature Neuroscience 4 (6), 662–667.

Milner, A.D., Goodale, M.A., 1995. The Visual Brain in Action. Oxford University Press, Oxford.

Mishkin, M., Ungerleider, L.G., Macko, K.A., 1983. Object vision and spatial vision: two cortical pathways. Trends in Neurosciences 6, 414–417.

Moro, V., Berlucchi, G., Lerch, J., Tomaiuolo, F., Aglioti, S.M., 2008. Selective deficit of mental visual imagery with intact primary visual cortex and visual perception. Cortex 44 (2), 109–118.

Ortigue, S., Viaud-Delmon, I., Annoni, J.M., Landis, T., Michel, C., Blanke, O., et al., 2001. Pure representational neglect after right thalamic lesion. Annals of Neurology 50 (3), 401–404.

Policardi, E., Perani, D., Zago, S., Grassi, F., Fazio, F., Ladavas, E., 1996. Failure to evoke visual images in a case of long-lasting cortical blindness. Neurocase 2 (5), 371–394.

Rich, A.N., Williams, M.A., Puce, A., Syngeniotis, A., Howard, M.A., McGlone, F., et al., 2006. Neural correlates of imagined and synaesthetic colours. Neuropsychologia 44 (14), 2918–2925.

Rode, G., Perenin, M.T., 1994. Temporary remission of representational hemineglect through vestibular stimulation. NeuroReport 5, 869–872.

Rode, G., Rossetti, Y., Perenin, M.T., Boisson, D., 2004. Geographic information has to be spatialized to be neglected: a representational neglect case. Cortex 40 (2), 391–397.

Rode, G., Revol, P., Rossetti, Y., Boisson, D., Bartolomeo, P., 2007. Looking while imagining: the influence of visual input on representational neglect. Neurology 68 (6), 432–437.

Rode, G., Cotton, F., Revol, P., Jacquin-Courtois, S., Rossetti, Y., Bartolomeo, P., 2010. Representation and disconnection in imaginal neglect. Neuropsychologia 48 (10), 2903–2911.

Sartre, J.-P., 1936. L'imagination. Presses Universitaires de France, Paris.

Sartre, J.-P., 1940. L'imaginaire. Gallimard, Paris.

Sheridan, J., Humphreys, G.W., 1993. A verbal-semantic category-specific recognition impairment. Cognitive Neuropsychology 10 (2), 143–184.

Weber, R.J., Castleman, J., 1970. The time it takes to imagine. Perception and Psychophysics 8, 165–168.

Zago, S., Corti, S., Bersano, A., Baron, P., Conti, G., Ballabio, E., et al., 2010. A cortically blind patient with preserved visual imagery. Cognitive and Behavioral Neurology 23 (1), 44–48.

Zago, S., Allegri, N., Cristoffanini, M., Ferrucci, R., Porta, M., Priori, A., 2011. Is the Charcot and Bernard case (1883) of loss of visual imagery really based on neurological impairment? Cognitive Neuropsychiatry 16 (6), 481–504.

Zeman, A.Z.J., Della Sala, S., Torrens, L.A., Gountouna, V.-E., McGonigle, D.J., Logie, R.H., 2010. Loss of imagery phenomenology with intact visuo-spatial task performance: a case of 'blind imagination'. Neuropsychologia 48 (1), 145–155.

Visual Images in the Media

John Huxford, School of Communication, Illinois State University, Normal, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article offers a review and assessment of the polemics surrounding the nature of the visual image (i.e., a reproduced sight) in the media (mass-produced communication), and of its social implications. The author explores the intellectual history of the debate, which is most typically advanced as an opposition between the qualities and effects of the word, or text, and of the image. The discussion is tracked across three key areas: child literacy and education, commercial and political persuasion, and issues surrounding the news. It is also suggested that in an era when the image is seen to have gained ascendency over text, digital tools and the power of the Internet bring potent new influences on the way we create, consume, and comprehend visual images.

The world is filled with visual images. Images appear on television and movie screens, billboards, posters, magazines, newspapers, and computer monitors. While, as a specific field of academic study and teaching, visual communication is little more than 50 years old, the properties and impact of the visual image have been the subject of research for centuries across a range of academic disciplines. In recent years, the emphasis has moved away from theories of art toward a psychosocial perspective that draws on research from semiotics, branches of psychology, sociology and communications, and cultural theory.

In this article, the discussion will focus on three key areas: child literacy and education, commercial and political persuasion, and issues surrounding news, as well as on the repercussions of digital imagery and the way that the power of cyberspace is changing the social role of visual communication.

The Age of the Image

Increasingly, it is argued that the word, once the dominant medium, is being supplanted by the image. Civilization has taken, in W.J.T Michell's phrase, a 'pictorial turn' (Mitchell, 1994: pp. 2–3). As Kellner suggests: "We are living in one of the most artificial visual and image-saturated cultures in human history, which makes understanding the complex construction and multiple social functions of visual imagery more important than ever before" (2002: p. 81).

In earlier centuries, the balance between verbal and visual representation strongly favored the word. Text dominated, while images in books typically served in a subsidiary role as illustrations. However, this status quo was challenged by the emergence of the photographic process in the nineteenth century, with photographs then spreading into newspapers and advertisements. Film accelerated the process with the development of a series of audiovisual technologies, including movies and television, tipping the balance between word and image in favor of the latter.

Today digital imagery and streaming video in cyberspace have added new dimensions to our ever-developing and increasingly complex visual culture. Text and images interact in new ways in the multimedia terrain of the Internet, reigniting the long-standing debate over the properties and functions of the visual versus those of the word.

Visual Images versus the Word

There has long been debate over the social implications of the image. Often this has centered on controversies that surround certain visual media; most prominently, racial and gender stereotyping, sexually explicit material, and violence as depicted on television, film, the Web, and in video games. However, this debate has also been advanced as an opposition between the qualities and social effects of the word or text, and those of the image, with the former typically being privileged over the latter.

History of the Debate

Whichever medium may be the most ubiquitous, the premise on which this debate turns has always been the supposed superiority of text over image, a bias informed by an intellectual tradition that stretches back centuries. (For a detailed discussion, see Stephens, 1998.) The Bible, in associating the Word with creation in Genesis while forbidding the production of the 'graven image or any likeness,' has long been used to support this presumption. Even as the image has established itself at the center of Western culture, writers and philosophers from Plato to contemporary thinkers have remained critical of the qualities and effects of the image compared to those of the word.

This intellectual antipathy is prompted by a series of overlapping concerns. Partly, it is founded on a long-standing suspicion of *verisimilitude*. In one view, such a capturing of likeness evokes a high degree of 'truth.' It is this perspective that typically reflects a traditional, although now contested, history of art as the progress of visual reproduction toward an ever more accurate reflection of the natural world, an ideal established by the Romans, and one that finds its most perfect expression in the photographic process.

However, verisimilitude has also driven a more negative perspective; that what the visual image offers is not truth but seductive illusion. A painter, Plato wrote, can deceive 'children

and fools' with the 'imitation of appearance,' instead of bringing 'truth' or 'real things' (1956: p. 463). Substantively, this argument remains the root of one strain of criticism laid against television and movies, that their illusionary realism may offer an ideal vehicle for propaganda.

A second charge leveled against the visual image is that, with the growth of technical reproduction, images devalue the objects they represent. The uniqueness of the original is lost in the deluge of visual copies. This argument is perhaps most acute when what is being reproduced are images of cherished buildings, statues and landmarks, and most especially, works of art. When the Mona Lisa becomes a common sight in the everyday world beyond the louver – for example, on T-shirts and advertising hoardings – the 'quality of its presence,' in the words of Walter Benjamin, is "always depreciated" (1969: p. 221).

A third charge involves the long-held belief that the image speaks more directly to the emotions than to the logical mind. Aquinas argued that images "excite the emotions more effectively aroused by things seen than by things heard" (cited in Freedberg, 1989: p. 162). This rousing of emotions is seen as making the problems surrounding pornography and the visual depiction of violence all the more acute.

Finally, the image's accessibility may distract the populace from the more cerebral merits of the word, be it the religious tract, improving essay, or newspaper text.

Visual Images and Child Literacy

If there is unease that the image is replacing the word for society in general, this concern is heightened in debates over child literacy. At different times, anxiety has centered on the presumed ill effects on children's reading abilities and habits of a series of visual media, including movies, comic books, and video games. However, most commonly criticized is television. Researchers have offered a number of hypotheses concerning the effects of television on child literacy and weighed the empirical evidence for each.

A *stimulation* hypothesis argues that there is no conflict between images on television and words in print because watching television stimulates or improves reading in the young. This theory, although not widely held, rests largely on studies concerned with the reading of a book based on a television show after watching the program, or more generally, on children being prompted by television to research an interest. However, in a study of the juvenile book market, only a small percentage of children reported reading a book after seeing or hearing of it through television. Similarly, the findings from other studies imply that any interests that are aroused by television tend to be short-lived.

More systematically studied are variations of a *reduction* hypothesis, where the consumption of television is predicted to have a negative impact on reading. There are five common variations of this hypothesis. First, a *displacement* hypothesis argues that television viewing takes time away from reading. However, studies have shown no conclusive evidence of this, indicating instead that time given to reading, although always scant, has remained remarkably stable.

Second, the *passivity* hypothesis argues that television viewing causes children to become mentally indolent and so less inclined to read. Although studies have shown that television does require less mental effort than reading, it is argued that synthesizing visual and audio information engages children in a process that is active, not passive. Third, the *retardation* hypothesis, an associated theory that argues television viewing deteriorates the brain, is seen as implausible for the same reason.

A fourth variety of reduction hypothesis is *concentration deterioration*, which suggests that a child's ability to concentrate is weakened by television. This too, although popular among critics, has found little scientific support.

Finally, *anti-school* hypotheses argue that children come to expect school to be as entertaining and immediately gratifying as television programs, and when it fails to meet these expectations, pupils lose motivation. Once more, the evidence for this is contradictory and inconclusive.

Visual Images in Education

In contrast to the attribution of negative effects to television viewing, some educationalists have called for the visual image to play a larger role in school curriculums which, traditionally, have been dominated by the word. Despite the criticism of television in general, some educational shows – such as 'Sesame Street' – have achieved strong reputations.

More recently – and despite the negative issues surrounding some games – there has been increasing interest in the potential of video games as educational tools. Some scholars have argued that such games may teach not only specific educational content but also foster the ability to learn by stimulating cognitive development. Proponents point to the focus on problem solving that is central to many games, and it has been argued that video games may even encourage children to develop a basic understanding of scientific methodology, in which hypotheses are formed and tested, as they seek to understand and manipulate the game world. (For a detailed examination of these issues, see Brown, 2008.)

In broader terms, this move toward the educational value of images may represent a belief in the benefits of acquiring what has been termed 'visual literacy.' The view is that irrespective of content, engagement with visual media may lead to the development of mental skills applicable beyond the understanding of the medium itself. To some extent, this argument is predicated on the widely influential work of Howard Gardner who, in developing a theory of 'multiple intelligences,' argued that different modes of communication are beneficial to the development of different types of intelligence.

However, the theory that visual imagery may encourage both specific cognitive development and a distinctive worldview is more generally a reflection of similar beliefs regarding textual and verbal communication, as expressed, most notably, by linguists Edward Sapir and Benjamin Lee Whorf. The perspective may also be traced to theories associated with Marshall McLuhan, who links cognition to the technical characteristics of varying media.

From this perspective, it is suggested that television-based teaching may have a positive effect on a number of areas of cognition, including the formation of a distinctive worldview,

analytic reasoning, and spatial intelligence: the ability to envision mentally, for example, what a shape might look like from various angles. It is also suggested that exposure to certain visual media, such as advertisements that employ metaphors, may encourage and develop abstractive and analogical thinking.

However, it is argued that because images primarily function as *analogs* of reality and are so less inclined to divide the perceived world into particular categories in the manner characteristic of language, the development of a specific worldview through exposure to visual imagery appears unlikely. (This is not to say that the *uses* to which the visual image may be put – e.g., to depict violence on television – may not have significant impact on one's worldview.)

Similarly, it is judged improbable that engagement with visual forms will enhance a viewer's skills in analytic thinking, since images lack many of the propositional elements of language (for example, it is difficult to make statements in the negative through images, e.g., 'there is no more ice cream'). While a stronger case may be made for the connection between visual imagery and spatial intelligence, evidence for these benefits is also weak. Finally, it is argued that aspects of abstractive or analogical thinking may indeed profit from viewing certain visual material, although the argument remains controversial (see Messaris, 1994 for a full discussion).

More generally, there is a growing body of opinion that, rather than teaching new cognitive skills, the representational conventions of images are founded on informational cues that people *routinely* learn to interpret through their visual experiences in real life. This calls into the question the very existence of a distinctive 'visual literacy' based on cognitive abilities acquired through exposure to mediated imagery.

Properties of Visual Images

Overall, there is scant evidence then, that visual images – most typically in the form of television – are either detrimental to child literacy, or significantly contribute to distinctive cognitive development.

However, it is clear that the image does possess distinctive qualities. Peirce's classification of signs into iconic, indexical, and symbolic (Pierce, 1991) is seen as an important key to these properties. Images that bear a resemblance to some aspect of reality are examples of *iconic* signs. At the same time, the photograph may be defined as an *index*; a sign that has some physical connection to the object or event to which it refers, (such as fingerprints or footprints) since it is produced by the physical effect of reflected light.

These features clearly set the image apart from the word. Iconic signs have little role in language and indexicality is entirely absent. Conversely, the visual image would normally not rely on the type of arbitrary conventions that link signifier (representation) with signified (what is represented) in the manner that constitutes the symbolic mode most characteristic of language.

Another crucial distinction can be made between words and visual images. Whereas verbal communication possesses detailed conventions governing word order (or syntax) that facilitate the linking of concepts, there is no such formal convention for combining images into larger meanings, although those who work with images, such as artists, may develop their own. Images exhibit what is termed 'syntactic indeterminacy' that increases their ambiguity as carriers of meaning (Messaris, 1997).

These distinct features have important consequences for the roles that the visual image plays within society. While it is difficult for images to convey precise *statements* about the objects they represent (e.g., 'the man no longer enjoys sarcasm'), still the image's characteristic iconicity, together with less pronounced but flexible symbolic properties, have allowed the medium to convey certain, descriptive information with a precision unmatched by words. For example, a range of scientific fields have been advanced through the use of detailed depictions and diagrams.

Equally, the visual image has been a boon in the development of user-friendly computers, where iconic representations, such as scissors and folders, have been used to replace lines of text. At the same time, photographic indexicality has become a vital component in the creation of historical records.

However, these properties may also have troubling consequences.

Visual Images and Persuasion

Iconicity, indexicality, and syntactic indeterminacy bring distinctive properties to visual persuasion, be it set to commercial or political purposes. In an extensive study of these issues, Messaris (1997) offers a broad range of devices through which the advertising image may engage attention and mold attitudes. Iconic representations of physical appearance and interpersonal behavior may reproduce real-world visual cues associated with a series of emotional responses; including sexual interest and excitement. This prompting of emotion is central to advertising strategies that seek to associate positive feelings with the possession or consumption of products (see, for example, Williamson, 1978). Further, iconic devices that manipulate the viewer's point of view (such as close-ups or high and low angles) can incite a series of other responses, including curiosity and trust.

The ability of iconicity to communicate information quickly and with relatively little mental effort may contribute to another advantage the image may have in persuasion. Some cognitive psychologists believe that visual images are typically processed peripherally – rather than centrally – in ways that may decrease the likelihood of viewers engaging their full critical and analytical faculties, especially when the speed of the message leaves little time for reflection.

The 'truthfulness' implied by indexicality is also a critical ingredient in visual persuasion whenever a photographic image is offered as proof of an advertiser's claims. However, visual images may readily be made to lie through the application of a wide range of techniques, from the staging of photographs to airbrushing to improve appearances. Moreover, digital technology has become a powerful aid to those who would practice such visual deception.

Finally, the syntactic indeterminacy of the image may also have important consequences that make it more useful in persuasion than the explicit meanings offered by words. For example, advertisers may take advantage of this property to

imply associations or statements that they would be unwise to state verbally (such as the association of cigarette smoking with healthy activity).

The implications of these properties have led to a heightened attention to the exploration of image production and the messages imbedded within them, most commonly through the perspective of critical analysis. Critical cultural analysis views society as a 'contested terrain,' a battleground for "domination and hegemony between competing gender, race, sex, ethnic, class, and other forces" (Kellner, 2002: p. 85). From this perspective, "codes of power are embedded in all visual images to varying degrees" (Sperling, 2011: p. 28).

These concerns focus on a range of issues and forms of media, for example, the sexual objectify of women in music videos, celebrity gossip magazines, and video games.

Two of the most debated arenas, however, remain the persuasive effects of imagery in political and commercial communication, and in the news.

Visual Images and Politics

While political advertising has existed, in its modern form, since at least the 1820s, television ads have now become a primary source of political information. Television has been accused of having a number of detrimental consequences for the political process. These include, in a political equivalent to the displacement theory associated with child literacy (see above), distracting the populace from civic involvement. However, the area of most concern with regard to the visual image is its effects within political advertising and televisual campaigning.

The potency of the visual image in political persuasion is thought to have been most graphically displayed in a debate between Richard Nixon and John F. Kennedy. It is reported that the majority of those who saw the 1960 debate on television judged Kennedy the victor, while most who listened on radio thought Nixon had won. Typically, this result has been taken to indicate that the more attractive image that Kennedy presented was the deciding factor, even though other influences have been offered – including the possibility that it was radio that may have favored Nixon's patterns of delivery, while exaggerating Kennedy's Boston accent in a manner that repelled listeners (Schudson, 1995).

The use of visual imagery for persuasive purposes becomes most intense in political advertising. Although the public is becoming more skilled at 'reading' visual cues, it is argued that viewers usually do not take the time to consider the way these cues may be attempting to manipulate their attitudes, making the visual element of political ads the most crucial.

Political advertisements use many of the same visual strategies as their commercial counterparts, although often exhibiting a strong reliance on iconic visual effects. Patriotic appeals are common in the political advertisement or campaign, and the use of certain visual symbols – including the national flag – has been found to be especially prevalent. Equally, the use of family imagery, such as images of spouses and of children, remains common.

The iconicity of visual imagery has been held responsible for other political strategies that accentuate appearance over content. Many make extensive use of devices that appear to be modeled on people's real-world experiences of interpersonal space. For example, politicians interviewed on television will typically turn toward the camera, and therefore toward the viewer, in order to imitate the real-world appearance of a direct and candid approach. Similarly, the use of color to create emotional impact is widespread in political advertisements. Bright colors are used to connote positive images, while gray or black and white are frequently employed in attack ads to help prompt a negative response.

Visual Images and the News

Equally controversial has been the influence of the visual image in news. The entry of the image into an arena that had traditionally been the province of the word was originally viewed with consternation, both by journalists and by many outside of the profession, as the use of visual material developed from artwork through woodcut engraving to early photographs. Visual images continued to be regarded as a medium that bypassed the reasoning mind and, as such, appeared to threaten journalism's Enlightenment heritage. While the photograph was held in higher esteem in the picture magazine, this had little impact on its critics.

Yet at the same time, the photographic record, as metaphor, as well as in practice, came to underpin the journalistic enterprise. By the mid-nineteenth century, there was widespread typification of the newspaper as a 'daguerreotype' (an early camera) of the social and natural world, with the indexicality of the process being used to exemplify, and give credence to, claims of journalistic objectivity (Zelizer, 1998: p. 9).

However, Photoshop and other digital techniques have made it increasingly easy to falsify photographs, with such practices finding their way into the news. More broadly, critics have argued that the apparent immediacy of the news photograph works to render invisible the human agency involved in its production. In practice, every image is subject to selection, interpretation, and, often, alteration as it works its way through the news process.

The news photograph embodies the perspective of the photographer who selects a particular camera angle; the page designer or editor who frequently selects one shot from a large number taken of an event and who may then crop the image; and the headline or caption writer where the photograph will be combined with, and interpreted through, text, while simultaneously being bounded within the newspaper's policy on how photographs may be deployed. At each stage the image is made to conform to a certain narrative structure – a news angle or theme – with its meaning inevitably being colored and manipulated by that association. As Stuart Hall observes: "News photos operate under a hidden sign marked 'this really happened, see for yourself.' Of course, the choice is this moment of an event as against that, of this person rather than that, of this angle rather than any other. Indeed the selection of this photographed incident to represent a whole complex chain of events and meaning, is a highly ideological procedure" (Hall, 1973: p. 241).

Similarly, in response to claims that the photography offers a transparent 'codeless' window on the world of newsworthy events, scholars have pointed to the use of photographs, in all news media, as being a practice based on cultural codes of understanding (such as an eagle representing freedom in America), as well to the connotative properties inherent in the practice itself. Roland Barthes (1977), among others, has argued that the photograph fulfills not only a denotative (or referential) but also a connotative (or interpretive) function, which undermines its claim to be unmediated truth.

Visual symbolism is frequently used in news to carry messages that stretch far beyond the photograph's referential qualities, with such techniques used to provide photographic validation of "abstract issues and qualities that cannot, in the normal sense, be validated visually" (Huxford, 2001: p. 45). Huxford offers three classifications of photographic symbolism routinely employed in the press on both sides of the Atlantic; *temporal*, in which time frames are evoked and utilized within the newspage; *metaphorical*, where analogical associations are employed; and *synthetic* which entails gross distortions of reality.

In all these ways, the news photograph, be it on the page, computer screen, or in broadcast film, may be far from either 'neutral' or 'objective.'

More recently, the focus of research has shifted away from issues surrounding the veracity of the news photograph, and toward the social and cultural functions that news photography may fulfill. This development highlights the ritual work of journalism in social life, specifically its role in creating and sustaining both communal and individual identities. The unifying work of ritual and of ritualistic journalism becomes particularly crucial when death or destruction threatens the social order. Research following this line of inquiry has included Dayan and Katz's exploration of the live coverage of events of significant social meaning; what is described as the 'contests,' 'conquests,' and 'coronations' that relate to one or more of the core values of society (1992: p. 199). Similarly it is argued that news images may be translated into 'lived images' during community ritual, being appropriated and reworked in ways that may empower and heal the community. For example, the striking and dramatic news imagery that was a central focus of the attack on the United States in 9/11 later reemerged in new ways in local carnivals and parades, with floats featuring pictures and models of the Twin Towers restored to their former glory (Coonfield and Huxford, 2009).

Digital Images

The ascendency of the image and the controversy over its effects in such areas as education, politics, and news have been accelerated and intensified by the rise of digital media.

While the Web has, arguably, given fresh impetus to text and allowed the word and image to interact in new ways, Web designers frequently give more space and visual weight to imagery. Moreover, streaming video is gaining increasing importance on the Web as technology and bandwidth improve.

A significant aspect of this new digital frontier is the manner in which it has fostered expertise in creating and manipulating digital imagery, turning users from passive viewers of images into active participants. The development of digital photography has prompted a substantial leap in the number of pictures being produced. By 2004, 28 billion digital photos then were being taken annually in the U.S., 6 billion more photos were shot on film, even though twice as many people owned film-roll cameras than digital (Murray, 2008: p. 152). The downloading and placing of such photographs on home pages, social networking sites, and blogs is now ubiquitous, while many individuals are also producing their own video clips, segments of animation, and even short movies. At the same time, a host of online communities that hone and spread this digital expertise have emerged in recent years. As Sperling argues, "Today's students also form individual identities and act online; they are at home with the Interweb's visual architectures and intellectual platforms; and they actively embrace the informality of its connections, its unstable transactions, the fluidity of its borders, and its shifting and disorienting discursive spaces" (2011: p. 28).

While the Web has been described as a 'digital utopia' (Rosen, 2001: p. 318), many of the issues surrounding the deceptive and potentially damaging effects of visual representation have followed images into cyberspace. Digital media has provided those who would perpetrate visual fakery with an arsenal of increasingly subtle and effective tools. At the same time, the range of visual networks through which today's online users move and the speed of those operations raise the likelihood that many images will register subconsciously, as unexamined beliefs and assumptions are prompted and absorbed.

However, the apparent vulnerability of the digital audience to visual trickery has been questioned. Some scholars argue that the key feature of today's 'post-photographic age,' where the digital image has replaced traditional emulsion-based photography, is that all pictures are seen as artificial representations of reality and are thus treated with increasing skepticism (e.g., Farid, 2009: pp. 42–44). Others are less sanguine. Messaris (2012), while acknowledging that the Web's digital networks have made it easier for the detection of visual manipulation to be publicized, also points out that awareness of the likelihood of digital manipulation does not necessarily translate into an ability to detect its use in specific instances.

Examples of digital manipulation of photographs that go undetected can be found in two key areas discussed in this article: political communication and news coverage. For instance, one experiment staged at Stanford University at the time of the 2004 Presidential election has implications for political persuasion. In this, photographs of the facial features of test subjects were subtly combined with the faces of either Republican George Bush or Democrat John Kerry. The subjects were then asked to make a preference between the photographs of the two politicians.

Researcher Jeremy Bailenson and his colleagues found that while the choices of highly partisan voters were unaffected by this manipulation, among those with less certain political preference the facial 'morphs' had the effect of swinging people's choices in the direction of the morph that included their own facial features. This held true even though the subjects demonstrated no conscious awareness of the manipulation (Bailenson et al., 2008).

Similarly, there are instances in news, as Messaris points out, when even journalists have been fooled by digitally doctored photographs. For example, when a fake photograph of the dead Osama bin Laden was posted online by unknown sources, the image was used on a number of well-regarded news sites before being exposed as a hoax.

Summarizing this debate, Messaris suggests, "In addressing this issue a researcher would have to go beyond the methodology of traditional surveys. The evidence would have to come from the examination of people's spontaneous reactions to images. Without a systematic look at such data, all claims about the effects of digital media on the public's attitudes towards images will necessarily remain highly conjectural" (2012: p. 10).

See also: Advertising and Advertisements; Advertising: Effects; Advertising: General; British Cultural Studies; Computer Games and Academic Achievement; Educational Media: Potentials for Learning; Film and Video Industry; Journalism and Journalists; Journalism; Mass Media, Representations in; Media Effects on Children; Media Effects; Media Events; Media and Child Development; New Media, News Production and Consumption; News: General; Photography as a Medium; Political Advertising; Political Communication; Semiotics; Social Media; Television: General; Violence and Media.

Bibliography

Bailenson, J.M., Iyengar, S., Yee, N., Collins, N., 2008. Facial similarity between voters and candidates causes influence. Public Opinion Quarterly 72 (5), 935–961.

Barthes, R., 1977. Image-Music-Text. Fontana, London.

Benjamin, W., 1969. The Work of Art in the Age of Mechanical Reproduction. Illuminations (English translation). Schocken Books, New York.

Brown, H.J., 2008. Videogames and Education (History, Humanities, and New Technology Series). M.E. Sharpe, Armonk, NY.

Coonfield, G., Huxford, J., 2009. News images as lived images: media ritual, cultural performance, and public trauma. Critical Studies in Media Communication 26 (5), 457–479.

Dayan, D., Katz, E., 1992. Media Events. Harvard University Press, Cambridge.

Farid, H., 2009. Seeing is not believing. IEEE Spectrum 46 (8), 44–48.

Freedberg, D., 1989. The Power of Images: Studies in the History and Theory of Response. University of Chicago Press, Chicago, IL.

Hall, S., 1973. The determinations of news photographs. In: Cohen, S., Young, J. (Eds.), The Manufacture of News: A Reader. Sage, Beverley Hills, CA, pp. 176–190.

Huxford, J., 2001. Beyond the referential: uses of photographic symbolism in the press. Journalism: Theory, Practice and Criticism 2 (1), 45–72.

Kellner, D., 2002. Critical perspectives on visual imagery in media and cyberculture. Journal of Visual Literacy 22 (1), 81–90.

Messaris, P., 1994. Visual 'literacy': Image, Mind, and Reality. Westview Press, Boulder, CO.

Messaris, P., 1997. Visual Persuasion. Sage, Thousand Oaks, CA.

Messaris, P., 2012. Visual 'literacy' in the digital age. Review of Communication 12 (2), 1–12.

Mitchell, W.J.T., 1994. Picture Theory: Essays on Verbal and Visual Representation. University of Chicago Press, Chicago, IL.

Murray, S., 2008. Digital images, photo-sharing, and our shifting notions of everyday aesthetics. Journal of Visual Culture 7 (2), 147–163.

Pierce, C.S., 1991. In: Hoopes, J. (Ed.), Peirce on Signs: Writings on Semiotics by Charles Sanders Peirce. University of North Carolina Press, Chapel Hill.

Plato, 1956. Great Dialogues of Plato (W.H.D. Rouse, Trans.). New American Library, New York.

Price, C.J., 2011. Using visual theories to analyze advertising. Visual Communication Quarterly 18 (1), 18–30.

Rosen, P., 2001. Change Mummified: Cinema Historicity, Theory. University of Minnesota Press, Minneapolis.

Schudson, M., 1995. The Power of News. Havard University Press, Cambridge, MA.

Sperling, J., 2011. Reframing the study of American visual culture: from national studies to transnational digital networks. The Journal of American Culture 34 (1), 26–35.

Stephens, M., 1998. The Rise of the Image, the Fall of the Word. Oxford University Press, New York.

Williamson, J., 1978. Decoding Advertisements: Ideology and Meaning in Advertising. Boyars,, London.

Zelizer, B., 1998. Remembering to Forget: Holocaust Memory through the Camera's Eye. University of Chicago Press, Chicago, IL.

Visual Memory, Psychology of

Svein Magnussen, University of Oslo, Oslo, Norway

© 2015 Elsevier Ltd. All rights reserved.

Abstract

When you open your eyes, your visual world immediately makes sense. This is achieved by some process by which online visual signals are matched to stored visual representations, and attest to the extraordinary capacity of visual memory. Information that enters the brain through the eyes is stored and remembered in terms of visual representations – in the form of images and sequences of images – or recoded and stored by the brain in terms of verbal or categorical representations. This multiple-coding principle of visual memory is part of the explanation why long-term visual memory generally outperforms other forms of memory. Visual memory recruits many of the brain areas that are involved in online perception, suggesting that the processes of visual perception and visual memory are closely interwoven.

Human memory is not a single mental faculty or cognitive system; rather the cognitive processes which we collectively call memory are composed of a number of independent and specialized cognitive systems that encode and store information in different formats. The various taxonomies of human memory agree on two main distinctions: First, memory may be explicit, or declarative, in the sense that personal knowledge and previous experiences are consciously recollected or recognized, or memory may be implicit, or nondeclarative, in the sense of being expressed indirectly in behavior without accompanying conscious recollection of previous learning episodes (Schacter et al., 2000). Obviously, in most situations, including those of the experimental laboratory, memory systems are activated in parallel, and the resulting performance is supported by several memory systems (Tulving, 2002). It is a challenging task for memory researchers to delineate the contribution of the various memory systems.

Second, a distinction is drawn between short-term memory, called immediate memory by early memory researchers, which is the seat of consciousness and active processing and is able to store limited quantities of information for limited periods of time, and long-term memory, which stores unlimited amounts of information for unlimited periods of time. This distinction, associated with the so-called 'modal model' of memory, is generally, though not universally, accepted among memory researchers (Jonides et al., 2008), however, it is also accepted that the border between short- and long-term memories is diffuse, and that short-term memory performance may be influenced by long-term memory representations.

The concept of *visual memory* has two related meanings: In the most general textbook usage, visual memory refers to the storage of information transmitted through the visual system. According to this definition it is a form of memory, which may be compared with memory for information transmitted through other sensory systems such as auditory, tactile, or olfactory memory. The type of information that is processed (colors, music, smells) and which cannot be translated between modalities captures important characteristics of modality-specific memories. However, this 'channel of communication' definition of visual memory misses the point that vision is a gate to many memory systems, and conversely that information in different sensory channels may be processed by the same amodal memory system. In a more restricted usage, visual memory refers to a code by which visual perceptions are inscribed and retrieved from memory in the form of images or sequences of images. According to this definition, visual memory is not a separate memory system but a form of representation of information that is processed by different memory systems. This definition also acknowledges that the channel of communication does not specify the memory format; what is viewed by the eye may be stored amodally, e.g., in memory systems that rely on verbal codes.

With these distinctions in mind, the article first discusses the relationship between visual perception and visual memory; this section is followed by two sections devoted to visual short-term and long-term memory, respectively, with the contributions of explicit and implicit memory discussed within each section, and the article ends with a discussion of the empirical evidence for the existence of visual memory codes.

Visual Perception and Memory

Perhaps the most direct expression of visual memory is the perceptual process itself, in the recognition of colors, textures, shapes, objects, and visual scenes. This is achieved by some process by which online information from the visual system is matched to stored descriptions. Perceptual recognition is immediate, representing a fast, automatic and implicit (nonconscious) process, and it precedes semantic and episodic memory retrieval, which deals with the identification of particular, familiar objects and visual scenes. The vast collection of stored representations that allow perceptual recognition is at least partly associated with a separate memory system, called the *perceptual representation system* (PRS). It is believed to consist of several subsystems, one of which is a structural description system, which stores descriptions of coherent or meaningful objects. Object recognition is further influenced by the process of *priming*, which is a largely nonconscious or implicit memory system belonging to the PRS (Schacter et al., 2000). In priming, recognition of objects is speeded up by previous presentation of related information and slowed down by previous presentation of conflicting information. These influences may cover extended time periods, well into the long-term memory domain.

The close link between visual perception and visual memory is further illustrated by the observation that certain forms of perceptual deficits due to local brain lesions appear to affect memory as well. For example, persons who lose their ability to see colors owing to bilateral lesions of a region of the occipital lobe known as V4 usually also lose their ability to remember or imagine colors (Sacks, 1995).

Visual Short-Term Memory

Most current treatments of visual short-term memory are based on the theory of *working memory* proposed 40 years ago by Baddeley and Hitch (1974), later developed in a very influential book by Baddeley (1986) and in several subsequent papers (Baddeley, 2000, 2012). According to this theory short-term memory is not a simple mechanism for passive storage of information, rather it is a coordinated set of mechanisms that combine incoming information from sensory systems with information retrieved from long-term memory, and consists of a central executive operating with the assistance of two domain specific support systems or buffers, dealing with visual–spatial and verbal information, respectively, and one multimodal system, the episodic buffer. Broadly speaking, the concept of *visual working memory* refers to the short-term memory system, which stores information that enters the brain through the eye. But as already noted, the definition is too broad since visual information such as written words or pictures of naturalistic scenes is easily recoded and stored in verbal memory systems. A more restricted definition of visual working memory is *the maintenance and manipulation of information represented as visual codes*. Since, with complex, meaningful visual stimuli memory performance may by supported by visual and verbal working memory systems operating in concert, much of the research in visual working memory has aimed at isolating the visual component of memory.

The Visual–Spatial Sketchpad

The support system referred to as the visual–spatial sketchpad is not a purely visual processing mechanism, but thought to combine information from visual, tactile, and haptic sensory channels. To a first approximation the visual–spatial sketchpad can be compared with the phenomenon of visual imagery. Visual imagery is the ability to produce 'inner' images of previously seen persons, objects, and geographical scenes, and to produce images of scenes that are simply imagined. Visual mental imagery tasks require subjects to retrieve information from long-term memory or to maintain visual information that has recently been viewed, and to perform cognitive operations on, or make judgments about, this information. The purely visual nature of visual imagery is supported by experiments showing that cognitive operations on visual mental images are governed by the same laws as similar operations performed on online visual images, and that visual imagery recruits many of the brain regions involved in online perception, even perhaps including some of the earliest brain regions in the visual process. This process is quite similar to the task of the visual–spatial sketchpad, and there is evidence that visual memory and imagery activate a number of common brain regions (Slotnick et al., 2012). However, whereas visual imagery is conceived as a conscious process of image generation, maintenance in visual working memory is not conceived as a conscious process in the same sense. Differences between visual imagery and the visual–spatial sketchpad are also suggested by the finding that visual imagery is disrupted by task-irrelevant visual noise whereas maintenance in visual working memory is not disrupted by visual noise (Repovs and Baddeley, 2006). Thus, despite similarities, visual imagery and the visual–spatial sketchpad are not identical.

Capacity Limitation of Visual Working Memory

The capacity of visual working memory depends upon several factors, such as the type of information tested, the test procedure employed, the time allowed for memory encoding, and the discriminability of the items in the visual display. There are individual differences in working memory capacity and there are age effects. Visual working memory performance in the adult population starts to decline around 60–65 years of age, and decreases steadily toward older age, a trend it shares with most memory tasks involving explicit but not implicit memory (Nilsson, 2003).

Until quite recently, most research on visual working memory has been concerned with the spatial aspect of visual displays and comparatively little was known about memory for the other properties of visual objects and visual scenes, such as their color, texture, and orientation. Experiments investigating the object property aspect of visual working memory suggest that visual object working memory is a unitary mechanism, with a capacity limitation that is fairly independent upon object properties and complexity. To a first approximation, visual working memory may be conceived as storing integrated objects, thus when remembering objects and their properties, it does not matter if they have different colors and orientations or if they share these features. The capacity limitation is four objects; when the number of objects exceeds this limit, memory suffers (Awh et al., 2007). This does not mean that the visual working memory capacity is completely unaffected by stimulus complexity. While the number *four* pops up in most experiments, it is clear that several stimulus and encoding parameters determine the capacity of the visual working memory. For example, if object complexity is high and encoding time is low, the estimated visual working memory capacity decreases, but if object complexity is high and there is unlimited encoding time, capacity increases but appears to not exceed the magical number four.

Fractionation of Visual Working Memory

Recent experiments indicate that visual working memory may be composed of additional subsystems. It is now realized that there is one cognitive subsystem that is partly specialized for processing objects and object properties such as color, orientation, and shape, and a second cognitive system that is specialized for processing the spatial characteristics of visual displays (Repovs and Baddeley, 2006).

A powerful experimental strategy in studying the independence or interaction between cognitive processing systems, are the so-called *dual-task experiments* where participants are

required to carry out two tasks simultaneously. For example, participants may have to remember a previously presented visual pattern while counting backwards or making a spatial judgment. The results of such experimental manipulations provided strong evidence for separate verbal and visual memory systems, showing that the performance on visual working memory tasks is not impaired by a parallel verbal task but it is impaired by a parallel visual task, with complementary pattern found for verbal memory tasks. Concurrent visual processing tasks compete for processing resources whereas verbal and visual tasks are processed independently. A similar pattern of interference is found between visual and spatial tasks, both handled by the visual–spatial sketchpad: Visual short-term memory tasks are more strongly disrupted by visual than spatial interference, and spatial memory tasks are simultaneously more strongly disrupted by spatial than visual interference. This double dissociation strongly supports a fractionation of visual short-term memory into separate, parallel, limited-capacity, visual and spatial components (Klauer and Zaho, 2004).

Implicit Visual Short-Term Memory

While the concept of visual working memory is largely based on studies of explicit memory where participants are requested to remember specific stimulus items, implicit influences on visual working memory through perceptual priming is recognized (Baddeley, 2010). Further experimental evidence for an implicit component of visual short-term memory comes from a parallel line of research on so-called *perceptual memory* (Magnussen, 2000). In classical visual working memory tasks, participants are required to consciously remember the stimuli to be compared and the difference between the patterns is easily detected in simultaneous displays. In a typical experiment one object in the display may change in color, orientation, or shape, and across experimental conditions the number of objects in the display and/or the number of object features is varied. This approach is governed by what some cognitive psychologists call a *storehouse* metaphor – how much is retained in memory? Perceptual memory research, on the other hand, is governed by a *correspondence* metaphor, investigating the fidelity of memory. How well are the details of the original display retained, and what distortions of color, texture, and orientation are observed? In these experiments the memory of elementary attributes of visual images – size, orientation, movement – is measured using psychophysical discrimination tasks where participants are requested to decide, or guess, if a test pattern has a higher or lower value than a previously shown reference pattern on the stimulus dimension is question, and a discrimination threshold is determined, usually representing the value of 75% correct guessing. The results from these experiments show, somewhat surprisingly, that information about such elementary attributes is not degraded or distorted, but maintained with a fidelity of the sensory image during intervals up to 2 or 3 min. The storage is impaired by a concurrent processing task along the same dimension, but not by a concurrent task along another dimension, and there is evidence that at least three stimulus dimensions may be stored in parallel. Thus the pattern of interference mimics the interference pattern observed in the more traditional visual working memory tasks. However, the performance on this task does not require conscious recollection or recognition, the differences that are discriminated in memory are so small that they escape any verbal labeling. The high-fidelity memory performance depends upon the forced choice technique in which the participants often feel that they are just guessing, but also that they are guessing correctly. From such observations it was concluded that the memory performance in such delayed discrimination experiments was based on an implicit type of memory, rather than on explicit memory. This low-level memory system may be a part of the visual working memory system or form a parallel implicit memory system associated with the PRS. Its role in visual memory is not known but it might store information about visual features that helps building permanent object representations (Magnussen et al., 2010).

Visual Long-Term Memory

Visual Representation in Episodic and Semantic Memory

The picture in the morning paper is not just a sidewalk cafe, it is the Deux Magots in Paris, and the person in the front row is not just a female figure, it is the movie star, *la grande dame* of French movies, Catherine Deneuve. The conscious recognition of particulars is a visual expression of semantic memory – a memory system that stores information about general facts. And if you are a Parisian or happened to be a lucky tourist, you might also remember seeing Catherine Deneuve walking by as you were sitting outside the very same cafe, and you might remember her looks at that particular day. This production of a memory image to the 'inner eye' is a visual representation of episodic memory, a memory system that stores information about personal facts and episodes. Semantic and episodic memory represent declarative or explicit forms of memory. Both personal observation and experimental research testify to the human capacity for visual episodic memory.

The experimental investigation of visual long-term memory is closely associated with the study of pictorial memory. In a series of classical studies in cognitive psychology (Standing, 1973) participants viewed large numbers of black-and-white and color slides of natural scenes, objects, persons, interiors, and landscapes, and memory was later probed by presenting pairs of slides containing one previously seen and one novel picture and asking the participants to pick from each pair the picture they saw in the study phase. Human participants are amazingly good at this task. With a learning set of 2500 pictures the score was on average 90% correct after a 36 h memory interval, and with a learning set of 10 000 pictures the score was 84% correct in the test trials. This easily outperforms the memory for words tested with similar methods. The excellent recollection of pictures and visual scenes has been explained in terms of the dual coding theory, which distinguishes between visual and verbal memory codes (Paivio, 1995). Picture memory has an advantage because pictures are effectively coded in both memory systems whereas words are coded primarily, but not exclusively, in verbal systems. The effectiveness of dual coding is illustrated by memory enhancing techniques, e.g., in the method of loci where items to be remembered are attached to places in a visualized familiar

environment and later retrieved by mentally traveling the same path in memory. The distinction between visual and verbal coding systems is also a central part of theory of working memory.

Implicit Visual Long-Term Memory?

What is remembered in the experiments of Standing (1973) and subsequent studies of pictorial long-term memory? Is it really possible to remember that many pictures in detail or is the memory performance based on 'gist' memory? Gist memory is the general meaning of the scene that is independent of explicit knowledge of what the scene contain. In addition, the method that is used in most of these studies, the two-alternative forced choice technique, is open to guessing and implicit memory influences. Thus, the impressive memory performance in the classical experiments may be result of gist memory and/or heavily influenced by implicit memory. Rensik (2000) proposed a theory of visual scene perception in terms of three systems: A low-level vision system that analyzes the basic features of visual stimuli, a second system that of focused attention that provides information about particular attended objects, and a third system that takes care of scene schema, providing information about layout and gist. So, in harmony with the view of Tulving (2002) pictorial long-term memory is supported by both explicit and implicit memory systems.

There is now strong evidence for the contribution of implicit memory to long-term pictorial memory. In a study by Vogt and Magnussen (2007) participants studied 400 color photographs of front doors, each picture presented for 5 s, and were later tested for the memory of the pictures across a period of nine days. There were two conditions, original pictures and doctored pictures where minor details were removed, details that were largely undetected by the participants (see **Figure 1**). Eliminating the influence of gist – all pictures were pictures of doors – the results of the experiment, in conjunction with the results of the classic Standing (1973) study, demonstrated the importance of both visual details and of gist memory: A comparison of the results for the original pictures with the memory performance of the participants in the comparable condition of the original Standing (1973) experiment, showed that by eliminating pictorial diversity memory, long-term memory performance dropped by 10–15%, and an additional drop in memory performance by 15% was observed in the 'doors' experiment when the unnoticed details were removed in the 'doors' experiments (see **Figure 1**). The results of a series of studies with large number of naturalistic pictures (Brady et al., 2008) further argue that details are important in long-term visual memory. Even after viewing thousands of pictures in different categories, and with many exemplars in each category, the subsequent memory performance was very high. In general, the results of long-term visual memory studies tend to support the multiple-system theory of Rensik (2000), which acknowledges contributions from both explicit and implicit memory systems to long-term visual memory.

The Evidence for Visual Memory Codes

The necessity for postulating visual codes has been questioned by cognitive psychologists who favor a unitary memory system based on propositional, languagelike codes. As an illustration, studies of memory expertise have shown that the memory of chess masters is vastly superior to that of ordinary players for real game positions, but not for random positions of chess pieces on the board (De Beni et al., 2007). The expertise memory is explained by the fact that masters have a long-term storage of a large number of visual 'chunks', where a chunk is a constellation of actual game positions involving a few pieces, and the estimated number chunks is about 50 000, which is comparable to the number of words commanded by

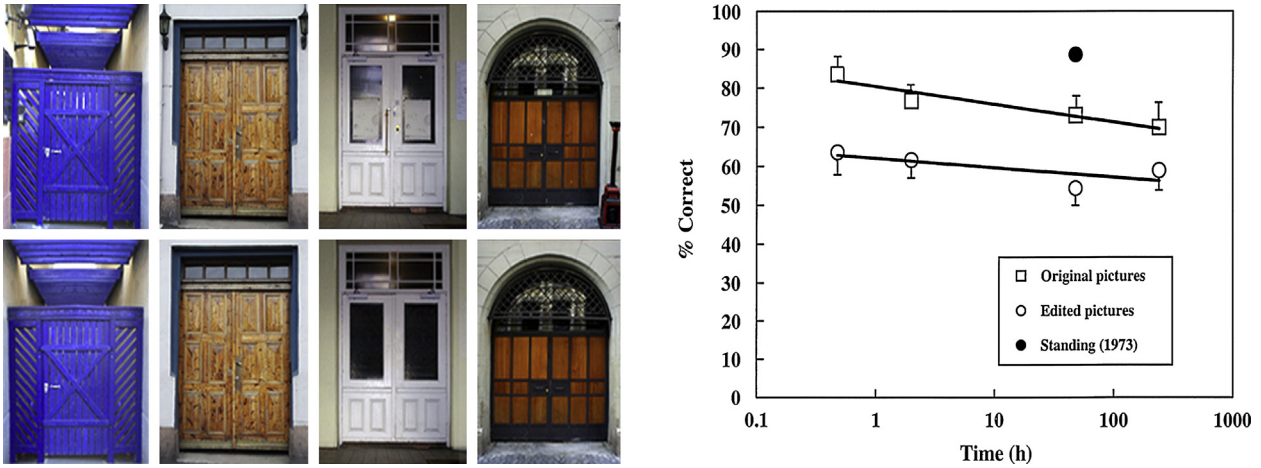

Figure 1 Visual memory for naturalistic pictures when the gist memory is eliminated. Participants viewed 400 color photographs of front doors and were later tested for the memory of the doors, at each session 100 studied doors were paired with 100 novel doors and the participant decided which of the doors had been previously shown. Results (right-hand panel) are shown for original pictures (left-hand panel, upper row) and pictures where minor details were removed (left-hand panel, lower row). Single point plots the result of Standing (1973), who had participants study a comparable number of random naturalistic pictures. Eliminating 'gist' memory indicates a drop in memory performance of about 15%, and eliminating minor details in the picture leads to an additional 15% drop in memory performance. The results strongly support a multiple-component theory of visual memory. Reprinted from Vogt, S.E., Magnussen, S., 2007. Long-term memory for 400 pictures on a common theme. Experimental Psychology 54, 298–303.

persons with a college education. Obviously, such chunks may be effectively represented either in terms of image codes or in terms of propositions or in both.

However, it is now widely recognized that visual memory, at least partly, depends upon visual memory codes, and some of the research cited above provides strong evidence that the brain employs such codes: The implicit memory representations that are responsible for the immediate perceptual recognition of the visual world cannot possibly be understood in terms of verbal or categorical codes. Also, the studies of perceptual memory provide strong evidence for pure visual memory codes, as the memory performance on these tasks escapes any verbal or categorical representation – the only way to perform these tasks is to rely on purely visual representations. The results of dual-task experiments provide additional evidence for separate visual and verbal representations, at least in short-term memory range. In long-term memory, the extraordinary retention of naturalistic pictures and the results of the experiments on the importance of visual details are difficult to interpret in terms of verbal or categorical representation. Additional arguments may be summarized under four headings.

Nonhuman Primates Have Detailed Visual Memories

A large number of experiments with monkeys have shown that these animals can learn to discriminate between large numbers of complex abstract fractal or snow-flake type visual patterns, store these in long-term memory, and utilize the information in single-trial short-term memory tasks. Few, if any scientists have claimed elaborate propositional representation in monkeys.

Perceptual Learning and High-Precision Storage of Sensory Information

In one famous experiment, an eidetic subject was able to identify the three dimensional figure hidden in random-dot stereograms when the left and right-eye patterns were viewed on different days (Stromeyer and Psotka, 1970). Such high-fidelity images cannot reasonably be stored in terms of propositions. Long-term memory changes in the form of perceptual learning are also well established. Subjects get better at performing fine discriminations between orientations or textures across large number of training sessions. The learning is specific for the training conditions and does not transfer to other conditions; for example, when subjects are trained in discriminating visual textures in one part of the visual field, the training effect does not transfer to other parts of the visual field (De Beni et al., 2007). Thus, at least some forms of visual memory must be due to local neural changes closely associated with the perceptual process itself.

Dissociation between Visual Memory and Other Forms of Memory

Lesions of the brain, which disturb semantic and episodic memory may leave the perceptual recognition process intact. The patient's perception of the world is normal but he or she does not remember seeing these objects or persons before. Recognition memory may, however, be demonstrated though perceptual priming: If the patient is shown a drawing of an object and at a later time is shown the same drawing again, there is a gain in the time required to identify and name drawings that were previously seen compared with novel drawings, and the gain is comparable for amnesic and normal subjects. If, on the other hand, the link between visual memory representations and online perception is disrupted, amodal (verbal) episodic and semantic memory may be intact but the person can no longer recognize what he or she sees. In the neuropsychological syndrome of *visual agnosia* the patient sees but the visual world is meaningless. There may be nothing wrong with the perceptual process per se, and the person navigates safely in the geographic environment and may even copy diagrams and drawings. However, there is no visual recognition and no visual memory, although other forms of working memory (and long-term memory) may be completely intact, but the information, which comes from vision is meaningless (Farah, 1990).

Visual Processing Areas of the Brain Are Recruited in Memory Tasks

Modern brain imaging techniques such as functional magnetic resonance imaging that allow fairly local and precise brain activity patterns to be mapped during the execution of cognitive tasks, would appear to offer a unique possibility of isolating the visual versus verbal components of working memory. But isolation of specialized memory systems is not a simple and straightforward procedure, both because cognitive tasks activate many brain systems concerned with perception, attention, and memory, and because most neural networks in the brain probably perform many processing tasks in parallel. Even if the performance on a specific visual memory task is not assisted by verbal memory, the simple fact that we identify a recognizable pattern that can be verbally classified and later remembered implies activation of verbal memory processes. Imaging studies have demonstrated distinct areas localized to the occipital and temporal lobes of the brain, specialized for processing particular classes of stimuli such as objects, scenes, and human faces (Haxby et al., 2004). These regions are activated in visual memory tasks with these classes of stimuli but not in verbal memory tasks, suggesting that at least certain kinds of information is represented as visual codes. Otherwise, imaging of memory processes in the human brain has shown that there are no regions specialized for storing or maintaining memories (D'Esposito, 2007). Memory involves widely distributed networks and all major regions of the cerebral cortex have been implicated in the various forms of memory. Since visual memory is a form of representation, a code rather than a system, one would expect it to share processing space with other cognitive memory systems, in addition to activation of the regions that are specific for the processing of visual information. Memory researchers generally agree that memories may involve many of the neural ensembles that were involved in the processing of the information in the first place, what is debated is how far down the processing hierarchy recruitment extends. Some brain imaging studies claim that the primary visual cortex (area V1), i.e., the very earliest level of sensory analysis receiving direct input from the eye, is activated in visual memory tasks (Harrison and Tong, 2009). The role of these early regions in short-term and long-term visual memory is a challenge for future research. They may be activated in visual recognition and delayed

discrimination tasks, but it is less likely that they are activated in memory tasks involving visual imagery. Since memory images are normally much less vivid and detailed than perceptual images, and memory is rarely confused with perception, they must activate at least partly different neural circuits (Slotnick et al., 2012).

Looking Ahead

One of the major advances of modern cognitive science is the idea that the human mind is functionally organized in terms of multiple independent processing systems, each system devoted to a particular cognitive task. In vision research, the idea of parallel, specialized processing systems was quickly accepted following the pioneering studies of the Nobel laureates David Hubel and Torstein Wiesel (1959), which showed that there were specialized neurons – in cognitive research termed channels – for processing features of the visual stimulus, such as color, orientation, size, and motion. The idea of modular organization of memory has had a slower progress. The current concept of visual memory identifies several memory systems as detailed above, and the distinctions between implicit and explicit memory systems, and the multiple-component working memory model, have been very successful in organizing the vast cognitive research literature. However, current neuroscience does not provide evidence for such a modular organization of memory, at least not if modules are defined in terms of localized brain processes. In the case of visual memory, an alternative view would be a model of memory that parallels the visual perceptual processing strategy of the brain, with the low- to higher-order specialization of the processing of visual features, objects, and scenes, and where different sensory areas are recruited depending upon the specific memory task.

Modern research confirms that visual memory is superior to other forms of memory, being able to store vast amounts of information in terms of pictures or natural scenes. Visual memory, in the general definition of memory for information transmitted through the visual system, partly depends on information stored in terms of visual representations, and partly on information stored in terms of verbal or categorical representations. The extremely good retention of visual information of pictures and geographical scenes is likely to be the result of the multiple-systems operation of visual memory. It is the task of future research to specify the relative contributions of the different memory systems to the short-term and long-term visual memory performance, and to decide if the concept of memory systems is indeed a fruitful one.

See also: Declarative Memory, Neural Basis of; Episodic Memory; Implicit Memory; Visual Imagery; Working Memory, Neural Basis of.

Bibliography

Awh, E., Burton, B., Vogel, E.K., 2007. Visual working memory represents a fixed number of items regardless of complexity. Psychological Science 18, 622–628.

Baddeley, A., 1986. Working Memory. Oxford University Press, New York.

Baddeley, A., 2000. The episodic buffer: a new component of working memory? Trends in Cognitive Sciences 4, 417–423.

Baddeley, A., 2010. Long-term and working memory: how do they interact? In: Bäckman, L., Nyberg, L. (Eds.), Memory, Aging and the Brain. Psychology Press, Hove, UK, pp. 7–23.

Baddeley, A., 2012. Working memory: theories, models and controversies. Annual Review of Psychology 63, 1–29.

Baddeley, A., Hitch, G.J., 1974. Working memory. In: Bower, G. (Ed.), 1974. Recent Advances in Learning and Motivation, vol. 8. Academic Press, New York, pp. 47–89.

Brady, T.F., Konkle, T., Alvarez, G.A., Oliva, A., 2008. Visual long-term memory has a massive storage capacity of object details. Proceeding of the National Academy of Science, USA 105, 14325–14329.

De Beni, R., Cornoldi, C., Larsson, M., Magnussen, S., Rönnberg, J., 2007. Memory experts: visual learning, wine tasting, orienteering and speech-reading. In: Magnussen, S., Helstrup, T. (Eds.), Everyday Memory. Psychology Press, Hove, UK, pp. 201–227.

D'Esposito, M., 2007. From cognitive to neural models of working memory. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 362, 761–772.

Farah, M., 1990. Visual Agnosia. MIT Press, Cambridge, MA.

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635.

Haxby, J.W., Gobbini, M.I., Montgomery, K., 2004. Spatial and temporal distribution of face and object representation in the human brain. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 889–890.

Hubel, D.H., Wiesel, T.N., 1959. Receptive fields of single neurones in the cat's striate cortex. Journal of Physiology 148, 574–591.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, B., Moore, K.L., 2008. The mind and brain of short-term memory. Annual Review of Psychology 59, 193–224.

Klauer, K.C., Zaho, Z., 2004. Double dissociations in visual and spatial short-term memory. Journal of Experimental Psychology: General 133, 355–381.

Magnussen, S., 2000. Low-level memory processes in vision. Trends in Neurosciences 23, 247–251.

Magnussen, S., Greenlee, M.W., Baumann, O., Endestad, T., 2010. Visual perceptual memory – *anno* 2008. In: Bäckman, L., Nyberg, L. (Eds.), Memory, Aging and the Brain. Psychology Press, Hove, UK, pp. 53–75.

Nilsson, L.G., 2003. Memory function in normal ageing. Acta Neurologica Scandinavia 107 (Suppl. 179), 7–13.

Paivio, A., 1995. Imagery and memory. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 977–986.

Rensik, R.A., 2000. The dynamic representation of visual scenes. Visual Cognition 7, 17–42.

Repovs, G., Baddeley, A., 2006. The multi-component model of working memory: explorations in experimental cognitive psychology. Neuroscience 139, 5–21.

Schacter, D.L., Wagner, A.D., Buckner, R.L., 2000. Memory systems of 1999. In: Tulving, E., Craik, F.I.M. (Eds.), The Oxford Handbook of Memory. Oxford University Press, Oxford, pp. 627–648.

Sacks, O., 1995. An Anthropologist on Mars. Alfred A. Knopf, Inc., New York.

Slotnick, S.D., Thompson, W.L., Kosslyn, S.M., 2012. Visual memory and visual mental imagery recruit common control and sensory regions of the brain. Cognitive Neuroscience 3, 14–20.

Standing, L., 1973. Learning 10,000 pictures. Quarterly Journal of Experimental Psychology 25, 207–222.

Stromeyer, C.F., Psotka, J., 1970. The detailed texture of eidetic images. Nature 225, 346–349.

Tulving, E., 2002. Episodic memory: from mind to brain. Annual Reviews of Psychology 53, 1–25.

Vogt, S.E., Magnussen, S., 2007. Long-term memory for 400 pictures on a common theme. Experimental Psychology 54, 298–303.

Visual Methods in the Social Sciences

Douglas Harper, Duquesne University, Pittsburgh, PA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16266–16269, © 2001, Elsevier Ltd.

Abstract

This article is an overview of visual methods in the social sciences, with an emphasis on sociology. The article notes that, relative to other methods, visual methods are underdeveloped in the social sciences. It details forms of visual methods, from empirical to semiotic. Particular attention is paid to the use of photographs in the research process. The article also discusses problems in visual methods, including ethics and matters of technology and representation. Finally, it discusses issues specific to the publication and other dissemation of visual research in the social sciences.

Visual methods, relative to other methods, are underdeveloped in the social sciences. The probable explanation is that social science reasoning generally relies on the analysis of aggregate data, and photography and other visual methods locate their attention on the particular. Thus, it requires a reversal of typical social science reasoning to generalize from the particular, especially when the particular is a visual image.

The nature of visual evidence is also problematical for social science. Photographic (film or video) images appear to represent the world 'objectively,' that is, without interpretation. At the same time, it is recognized that photographs are socially and technically constructed, thus embodying points of view and interpretive influences. The tension between the competing qualities of objectivity and subjectivity influence the use of visual images in the social sciences.

Forms of Visual Methods

Empirical Evidence

Photographs (or photos linked together in film or video) are the most common form of visual sociology. The most common use of visual images is to catalog aspects of material culture or social interaction (see Collier, 1967), which was a common aspect of articles published in the first issues of the *American Journal of Sociology* (see Stassz, 1979). In the first 20 volumes of the *American Journal of Sociology*, until 1916, sociologists used photographs to make arguments about the character of rural and urban life; the working conditions in offices and factories, and living conditions in industrializing cities. As sociology sought a more scientific posture, the visual research method (which was not well developed in these early examples) was largely abandoned.

Anthropologists, working in the first decades of the twentieth century, also used photography extensively (Edwards, 1992). Anthropology's primary mission during its early eras was to document and classify human groups. Photography's ability to gather a large amount of information in a single document made it an important data-gathering method. As anthropology abandoned paradigms based on the taxonomic organization of racial classifications, photography became less important as a research method.

The full development of photography as part of ethnographic analysis is generally attributed to Gregory Bateson and Margaret Mead's (1942) study of Balinese culture. In this project, the researchers used 759 photographs (selected from more than 25 000 made during their fieldwork) to support and develop ethnographic observations and analysis. Bateson and Mead used sequences of photographs to show how the Balinese performed social rituals or engaged in routine behavior. These images typically record the unfolding of events, and thus are similar to short movie clips. The researchers also used images to survey material culture such as dwellings and agricultural techniques. This promising expansion and redefinition of visual anthropology has been followed by contemporary examples such as Danforth and Tsiaras' study of death rituals of Greek villagers (1982), and several other examples in sociology and anthropology.

Howard Becker was the first modern sociologist to suggest a link between photography and sociology (1974). Much of Backer's argument referred to the ability of photographs to record the world as observed. For example, Becker noted that photography and sociology, since their inception, had both been concerned with the critical examination of society. Becker noted that the early twentieth century photographers, such as Lewis Hine and Jacob Riis, were concerned with topics such as urbanization and industrialization, the exploitation of labor, and the conditions of immigrants, all of central concern to sociologists. Becker suggests that a modern visual sociology should emerge from documentary practice engaged with a sociological theory, and the sociological use of visual data developed in the context of common concerns with validity, reliability, and sampling. Becker's article, reprinted in several anthologies, is credited with the rebirth of visual sociology in the US.

In the meantime, sociologists have used photographs as evidence in the study of social change (Rieger, 1996), in sociological classification of farmsteads (Harper, 1997), and other topics that reflect conventional sociological research. Visual ethnography, however, remains a relatively small movement within the field methods tradition in sociology and anthropology.

Photos as Cultural Narratives

Anthropologists have used film and video (and, to a lesser extent, the 'photo essay') to communicate typical events of a cultural scene. This tradition is largely attributed to Robert

Flaherty's seminal ethnographic film, *Nanook of the North*, completed during the 1920s (see Rotha, 1980). The resulting film tradition in anthropology has both embraced and challenged the capacity of visual narrative format to communicate cultural information. Nevertheless, the use of film (and photography) to tell 'cultural stories' remains a viable part of visual anthropology, and constitutes the single most influential demonstration of visual methods in social research. In anthropology, ethnographic films are presented at professional meetings, scholarly film festivals, and as a part of normal pedagogy.

Photographs in the Research Process

Collier (1967) first described how researchers may employ photographs to elicit cultural information. The process, now referred to as 'photo-elicitation,' uses images to stimulate culturally relevant reflections in interviews. Researchers may take the photos prior to the interview, or the photographs may be from a personal collection or professional archive. Photo-elicitation has become a common method in visual sociology, used in a wide range of studies, including public responses to social problems (Faccioli and Zuccheri, 1998), cultural definitions of an urban neighborhood (van der Does et al., 1992), the 'working knowledge' of a rural artisan (Harper, 1987), and ethnic recognition among Asian immigrants (Gold, 1991). The promise of photo-elicitation lies in the understanding of how photographs communicate different meanings to the researcher and the subjects; thus the images become a basis of interviewing that explores the taken-for-granted assumptions of both researcher and subject.

Social science film-makers have integrated imagery into research in a different manner. Beginning in the early 1970s, anthropologists have taught cultural informants to use film and video cameras in order to make films that tell cultural stories from the vantage point of the informant. The first example of this method was Sol Worth and John Adair's research among the Navajo, which resulted in seven films made by native informants (described in Worth and Adair, 1972). With the advent of easy to use video cameras, native-produced visual ethnography has become more common. While controversy continues to surround the redefinition of theoretical authority that native-produced films challenge, most visually oriented social scientists find the native-produced film or video a viable source of ethnographic information.

Problems in Visual Methods

Ethics

There are at least three aspects to the question of ethics and visual methods in research. The first concerns the matter of anonymity; the second concerns the matter of informed consent; and the third represents the burden of 'telling the visual truth.' These three issues are inter-related.

Because social science is generally concerned with the patterns of social behavior, individuals are represented by words, or more commonly, by responses they have made to a questionnaire. Thus the identity of individuals are protected. The protection of the privacy of subjects is, indeed, codified in the ethical canons of the social science disciplines. In the case of visual research, however, it is often impossible to protect the privacy of subjects. All discussion of ethics and visual methods rests on this fundamental fact; the photograph produces a likeness which is often identifiable.

Visual researchers thus face a more pronounced problem of informed consent. Subjects must be made aware of implications of publishing their likeness in social science research, and agreement to participate should be made in writing and prior to the research.

The matter of ethics in visual research is unique among research methods for one final reason. Because it is often possible to identify subjects in research images, visual researchers must be absolutely certain that these images 'tell the truth.' Visually oriented social scientists recognize that it is unethical to alter images to change their meaning, or to present an image that connotes a reality that was not consistent with the understood reality of a situation. Most visual researchers feel that the ethical considerations of visual research most separates it from that of their peers.

Technology

Like all social science methods, visual research depends upon technology. Like all technologies, visual technology makes it possible to record some types of information, but not others. Steiger (1995) detailed how specific cameras, lenses, film types, and camera settings create different kinds of filmic statements, which, in turn, make possible different types of sociological thinking. The specific connection between 'technologies of seeing' (whether relating to photography, film, video, or virtual reality) and social ideas remains an area of enormous potential and scant attention. Steiger's article raises the important point that to be a skilled researcher using visual methods is to master cameras, video and film cameras, and possibly photo printing. Now that many of these technologies have an electronic dimension, skilled visual researchers also need to become adept at software that catalogs visual images, that allows for the improved presentation of images, and that facilitates the production and distribution of images.

New methodologies, such as computer-based multimedia, the World Wide Web, CD publication, and virtual reality, all offer possibilities for the development of innovative visual methods. Each offers an approach for recording, analyzing, and participating in the visual dimensions of society. The future of visual methods in social research will depend, to a large extent, on how well these new technologies are integrated into social science methods and thinking.

Publication: Organizational Basis

In the past two decades, visual research in anthropology and sociology has been institutionalized in intellectual organizations such as the Society for Visual Anthropology, a part of the American Anthropological Association, and the International Visual Sociology Association, an independent scholarly organization with ties to several academic disciplines. Publication of visual methods requires special consideration, and the

journals such as *Visual Anthropology* and *Visual Sociology* feature the presentation of images. The full development of visual methods in social research will require that the mainstream journals and books accept the publication of images as a normal part of social science research, no different than the publication of tables or graphs.

Bibliography

Bateson, G., Mead, M., 1942. Balinese Character: A Photographic Analysis. New York Academy of Sciences, New York.

Becker, H.S., 1974. Photography and sociology. Studies in the Anthropology of Visual Communication 1 (1), 3–26.

Collier Jr, J., 1967. Visual Anthropology: Photography as a Research Method. Holt, Rinehart and Winston, New York.

Danforth, L.M., Tsiaras, A., 1982. Death Rituals of Rural Greece. Princeton University Press, Princeton, NJ.

van der Does, S., Edelaar, S., Gooskens, I., Liefting, M., van Mierlo, M., 1992. Reading images: A study of a Dutch neighborhood. Visual Sociology 7 (1), 4–68.

Edwards, E., 1992. Anthropology and Photography 1860–1920. Yale University Press, New Haven, CT.

Faccioli, P., Zuccheri, N., 1998. The double vision of alcohol. Visual Sociology 13 (2), 75–90.

Gold, S., 1991. Ethnic boundaries and ethnic entrepreneurship: a photo-elicitation study. Visual Sociology 6 (2), 9–22.

Harper, D., 1987. Working Knowledge: Skill and Community in a Small Shop. University of Chicago Press, Chicago.

Harper, D., 1997. Visualizing structure: reading surfaces of social life. Qualitative Sociology 20 (1), 57–77.

Rieger, J., 1996. Photographing social change. Visual Sociology 11 (1), 5–49.

Rotha, P., 1980. Nanook and the north. Studies in Visual Communication 6 (2), 33–60.

Stassz, C., 1979. The early history of visual sociology. In: Wagner, J. (Ed.), Images of Information: Still Photography in the Social Sciences. Sage Press, Beverly Hills, CA.

Steiger, R., 1995. First children and family dynamics. Visual Sociology 10 (1–2), 28–49.

Worth, S., Adair, J., 1972. Through Navajo Eyes. Indiana University Press, Bloomington, IN.

Visual Perception, Neural Basis of

Oliver Braddick, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Following active selection of the input by eye movements, visual perception requires sequential transformations of neural signals, in the retina, area V1, and extrastriate cortex. These transformations create patterns of activity that represent the information needed for recognition, location, and guidance of actions. These representations rest on population coding, in which single neurons carry univariant information, and each neuron acts as a labeled line. Each cortical neuron has a receptive field structure that makes it a tuned filter for properties such as contour orientation. Grouping processes create representations of structure, which serve as the basis, via the ventral cortical stream, for recognition of objects, faces, and scenes in specialized temporal lobe areas. In parallel, the dorsal stream to the parietal lobe creates spatial representations in various frames of reference appropriate for the control of actions. The binding problem – how different aspects of visual information are related to a common object – is important but not yet well understood. Psychophysical thresholds can be related to signals in noise, in specific populations of cortical neurons. Visual cortical activity is not determined solely by bottom-up processing but is subject to central modulation by attention and task demands. The visual system is also plastic and its connectivity can be modified by experience. Possible relations between visual awareness and neural activity are discussed.

Structure of the Visual Pathway

Visual perception is an interactive process in which eye movements are actively used to select the patterns of light entering the eyes. These patterns of light energy are encoded into neural signals in the retina of the eye, and these signals are then transformed as they pass through the neural network of the retina and are transmitted through a series of interconnected brain areas. These transformations create patterns of neural activity that represent the visual information needed for recognizing objects, people, and events; for locating ourselves in the environment; and for guiding actions. We concentrate here on the relatively early stages of this process, often referred to as 'visual perception' but it should be recognized that there are no clear cut boundaries between the brain processes underlying perception and those, sometimes called 'cognitive,' which work with visual information to direct attention, make decisions, retain and access visual memories, and control our actions.

The neural encoding of the visual scene begins in the photoreceptor cells (rods and cones) in the retina, which signal the local light intensity falling on each receptor by their levels of electrical polarization. A network of horizontal, bipolar, and amacrine cells connect these receptors to the retinal ganglion cells. The photoreceptors and ganglion cells are not uniformly distributed across the retina, but are much more densely packed in the central region (the fovea). The closely packed cone receptors in the fovea provide spatially fine-grained information about the pattern of light and dark, giving the high acuity needed for detail vision in daylight vision, and the number of ganglion cells (equivalent to one ganglion cell per cone receptor in the central fovea), ensures that this fine spatial grain is preserved in the information sent to the brain. Thus our sense of a richly detailed visual world is in fact provided by a relatively small high-resolution patch in the fovea, which continually moves around through eye movements to flesh out the coarse layout provided by the low-resolution vision of the peripheral retina.

At low light levels, the more sensitive rod photoreceptors take over from the cones: while rods are 20 times as numerous as the cones, they feed a much smaller number of ganglion cells. These cells pool the rod signals to enable detection of faint lights, at the expense of providing only a coarse picture of the visual scene.

The axons of the ganglion cells form the fibers of the optic nerve, and so the signals that they carry constitute the information that the eye sends to the brain. The nature of this information is discussed in the Section Representation of Spatial Pattern and Shape. At the optic chiasm, these fibers are routed so that signals from the two eyes can be transmitted to each side of the brain, making it possible to integrate binocular information at a higher level.

The lateral geniculate nucleus (LGN) of the thalamus provides a relay station where visual signals can be modulated, but the first major reorganization of the signals from the retina occurs when they arrive in the occipital lobe of the cerebral cortex (*see* Cerebral Cortex and Occipital Lobe (Visual Cortex): Functional Aspects). The primary receiving area, at the posterior pole of the brain, is known as striate cortex, or cortical area V1. It is surrounded by a series of extrastriate visual areas. These areas can be distinguished both by specializations of function (see Section The Hierarchy of Multiple Visual Areas) and by the fact that many of them form distinct, ordered maps of the visual field ('retinotopic' maps). Area V2 receives much of the input from V1, and sends information on to areas V3, V4, and V5. Pathways through these and other visual areas send information to the temporal and parietal lobes of the brain, and are considered to provide a hierarchy of visual processes of increasing complexity.

Within each cortical area, most connections between neurons are quite local. However, information is passed from one cortical area to another in fiber tracts, in an orderly way. Specific layers of the cortex receive input from areas at a lower hierarchical level, while other layers either transmit to higher levels, or receive feedback from higher levels (Bullier, 2004).

 http://dx.doi.org/10.1016/B978-0-08-097086-8.56018-8

Thus, cortical areas are commonly thought of as a set of modules, each performing particular visual computations and passing the results of these computations to other modules.

A minority of fibers in the optic nerve transmit information not to the cerebral cortex, but to midbrain structures, in particular the superior colliculus, which also incorporates a topographic map of the visual field. In lower vertebrates, the corresponding midbrain structures provide the main central visual pathway, but in mammals, certainly in primates, this pathway mainly serves the function of orienting to stimuli by eye movements rather than any complex analysis of the visual image, and probably does not contribute directly to conscious visual perception (see Section The Neural Basis of Visual Awareness). However, there are connections running both ways between this subcortical visual pathway and the much more elaborate cortical network for visual information processing.

Key Concepts of Neural Representation and Their Evolution

Specific Nerve Energies and Place Coding

In developing modern concepts of the neural basis of vision, an early key idea was that messages from the eye to the brain could convey distinct information according to which element of the nervous system carried them. Thomas Young, around 1800, proposed that different colors were represented according to the degree that three forms of 'sensitive filament' in the eye were stimulated, a view that in its essentials, with cone photoreceptors as the sensitive elements, is accepted today. Developments of this idea in the nineteenth century were Lotze's concept that each separate fiber in the optic nerve is associated with a 'local sign' – that is, we can know which point in space stimulated our eyes from which particular fiber is active – and Müller's 'doctrine of specific nerve energies' that the quality of stimulation (e.g., sound versus light) is conveyed by which fibers convey a signal, and not by any distinct quality in the signals themselves. This principle (sometimes called 'place coding' or 'labeled line') is now recognized as fundamental for understanding how visual information is represented within the brain, as well as in the sensory pathways leading to it.

Univariant Coding by Impulse Frequency

The work of Adrian in the 1920s established that sensory information is carried by a series of nerve impulses (action potentials). As the impulses are identical all-or-none events, quantitative variations, for example, in the contrast at edges between light and dark, have to be encoded by the rate at which impulses occur. In principle, more complex statistical properties of the train of impulses could carry rich information, but visual neurophysiology has not yet yielded convincing examples of this. Thus, the activity of a single neuron is univariant – it can encode only a single variable in the visual image. The classic example is in color vision: a particular receptor may respond relatively more strongly to green light than to red light. However, because its response is univariant, its responses to a dim green light and a bright red light are indistinguishable. Thus, to represent even a single variable, within a multidimensional domain such as color, requires information to be combined or compared from two or more neural signals. The modern neuroscience of color vision identifies such comparison in the opponent processing of different cone signals in retinal ganglion cells and higher levels of the visual pathway.

Receptive Fields and Tuned Filters

These principles of neural coding have been particularly applied in considering how the visual system represents spatial information. A concept which developed early was that each visual neuron has a receptive field – a region of the field of view within which light can influence the neuron's activity. Further work showed that receptive fields have internal structure, such that a particular spatial pattern of light and dark within this region is optimal for activating a given neuron. This led to the idea that neurons could be 'feature detectors,' signaling the presence of a specific pattern element, such as a small black moving spot that might be a fly in the visual field of a frog. However, in the mammalian brain at least, it currently seems unlikely that any single neuron on its own can provide such a signal; in all cases that have been investigated, the distribution of activity across some population of neurons would be needed to determine reliably the presence of a particular feature or object (Churchland and Sejnowski, 1992). This idea of population coding should not obscure the fact that, in visual brain areas, coding is sparse. That is, for any particular visual stimulus, only a small fraction of the available neurons are active, and so it is appropriate to think of particular subsets of neurons as being involved in representing particular visual attributes.

Neurophysiological studies in the 1960s and 1970s, pioneered by Hubel and Wiesel, concentrated on how populations of neurons in V1 encoded particular visual properties, such as the orientation of an edge. Each neuron was found to be a 'tuned filter,' which responded to a limited range of orientations. The perceived orientation of a contour can then be derived from the peak of activation across a distribution of differently tuned filters. This idea, originating in recordings of single neuron activity, was supported by psychophysical experiments, which showed that the visual system could be selectively adapted (pioneered by Blakemore, Campbell, and others; the method is summarized by Braddick (2010)). For example, intense stimulation by stripes angled at 5° from vertical reduced the ability to detect low contrast stripes at that orientation, and made vertical stripes appear tilted in the opposite direction. These results are what would be expected if the adaptation reduced the sensitivity of a group of neurons (sometimes called an orientation 'channel') tuned to orientations close to the adapting stimulus and so shifted the distribution of activity across the population of tuned neurons. Cortical neurons also act as arrays of tuned filters for other properties, such as direction of motion, and analogous effects of selective adaptation can also be found for these properties.

The arrays of tuned neurons, as well as allowing properties such as contour orientation to be registered, also make it possible to segregate visual information according to these local properties. Common orientation, color, or motion of texture elements makes it possible to segment the scene into distinct objects and surfaces, and to group features which belong to the same object. Such segmentation plays a key role in the organized perception of a visual scene.

The Hierarchy of Multiple Visual Areas

The finding, that neurons in area V1 were tuned to a range of stimulus properties, led to exploration of higher levels of cortical processing. Experiments with macaques, from the mid-1970s onward, showed that there is a multiplicity of distinct visual areas beyond V1 (*see* Occipital Lobe (Visual Cortex): Functional Aspects). As well as progressive transformations through a hierarchy of visual areas, there was found to be parallel processing by areas such as V4 and V5, which receive information from V1 in parallel via separate routes. These areas are specialized, so that, for example, V5 contains a high proportion of neurons tuned to specific directions of motion but few that are wavelength-tuned, while V4 shows the opposite balance (Zeki, 1993). This has led to a picture of functional separation, within the cortex, of representations which emphasize different aspects of visual information and serve distinct purposes.

This picture of functional segregation was consistent with findings in human patients who had localized brain damage; for instance, such damage can abolish color discrimination with little effect on other aspects of visual perception. Since the 1990s, new methods of functional brain imaging have made it possible to explore localization of function in the normal human brain (*see* Functional Brain Imaging of Language processes). These have confirmed that visual stimuli activate multiple brain areas, each responding to particular stimulus properties. The activation of a particular area also depends on the visual task the subject is set and whether they are attending to the stimulus, showing that vision cannot be thought of as a purely 'bottom-up' or 'feed-forward' neural process, in which information originating in the receptors is progressively transformed. Rather, the results of high-level processes can modulate how low-level processes operate. An anatomical basis for this is provided by 'descending' connections from higher to lower visual areas, which are at least as rich and complex as those transmitting information from lower to higher processing (Bullier, 2004).

These concepts of neural information processing inform current research on all aspects of vision. In this article they are illustrated primarily with respect to the representation of pattern, shape, and spatial layout. They also underlie current understanding of other perceptual functions such as motion, color, and depth perception.

Representation of Spatial Pattern and Shape

Receptive Field Structure and Specificity

The richness of vision as a source of information comes largely through the detailed pattern of light and dark that arises from the spatial structure of the environment. Encoding of this pattern begins in the retina. The receptive fields of retinal bipolar and ganglion cells have a center–surround opponent organization. In an 'on-center' cell, light in the receptive-field center activates the cell, but this is countered by inhibition from light in the surround. Off-center cells show the opposite relationship. In either case, the activity of the cell signals the presence of spatial contrast between center and surround, with little or no response to a uniform field of light covering the whole receptive field.

The signals in the optic nerve therefore highlight the field locations where spatial structure is present, but carry little explicit information about the nature of this structure. Ganglion cell receptive fields are circular, and so respond equally to the contrast in an edge crossing the field whether this edge is horizontal, vertical, or oblique. Neurons in area V1 also respond to local contrast, but mostly have receptive fields with a particular preferred orientation, as discussed in the Section Receptive Fields and Tuned Filters. Neurons with different preferred orientations ensure that contours at any angle around the clock are represented by activity within V1. Orientation seems to be a key organizing principle for V1, demonstrated by the way neurons are laid out in a semiregular array of 'orientation stripes' across the cortical surface; presumably this provides the initial basis for the neural representation of shapes and textures.

A second important organizing principle is that of multiscale representation. Some cortical neurons respond to very fine-scale spatial detail; others, covering the same region of space with larger receptive fields, respond to similar features on a much coarser scale (De Valois and De Valois, 1988). Mathematically, this variation of scale corresponds to different neural channels responding to different spatial frequencies within a region of the image. Computational studies show that this local analysis in terms of component spatial frequencies is an efficient way to capture the significant structure of scenes, as outlined in the next section.

From Measurements to Spatial Primitives

The multiscale array of orientation-selective neurons in V1 can be considered as making a comprehensive set of measurements of the image. To develop a useful representation of scenes and objects, these measurements need to define primitives – elements of an 'alphabet' of spatial structure through which more complex entities can be represented. Thus, a measurement of contrast in a particular orientation and scale does not itself unambiguously define an edge segment; but in conjunction with nearby measurements at a range of scales, it can provide the basis for an assertion of a local edge-segment primitive (Frisby and Stone, 2010: ch. 5). While such primitives are important in the computational description of what the visual system is doing, it is unclear whether neurons in visual cortex explicitly represent primitives, or rather whether these are implicit in the pattern of connections amongst V1 cells and from V1 to higher levels. In any case, the separation of information into different spatial frequency bands (scales) appears to be a characteristic of the neural organization at the level of V1 but is not a feature of the higher levels of the processing hierarchy.

Grouping Processes and Figure/Ground

To yield descriptions of object shapes and bounded surfaces, local edge primitives must be grouped together to form extended contours and areas with a common texture. The importance of grouping phenomena was first recognized by the Gestalt psychologists in the early twentieth century, who

argued that perception could not be understood in terms simply of the response to local elements but that 'the whole is greater than the sum of its parts.' A plausible neural basis for some aspects of grouping exists in the horizontal connections between V1 cells tuned to a similar orientation. Contours, in turn, define the boundaries of surfaces and objects. In particular, such boundaries establish the key perceptual distinction between the 'figure' to which the contour belongs and defines its shape, and the 'ground' or background; the contour marks the point where the ground is concealed by the object, but indicates nothing about the properties of the ground itself. If there are markings on the background, these are interrupted by the occluding contour. This interruption by occlusion can give the perception of an edge, even if there is no difference in overall luminance level of the figure and the ground – an effect sometimes called an 'illusory contour.' Peterhans and Von der Heydt (1991) have found neurons in V2 which respond to the characteristic configuration of contour elements where one surface occludes a background surface. Such neurons respond to an illusory contour or a contrast edge at the same preferred orientation, and so represent the boundary of an object in an invariant way, whether it is indicated by a difference in light intensity or by the interruption of background markings. The activity of neurons in V1 and V2 also reflects the figure/ground distinction; the response of a neuron that prefers an edge, which is dark on the left and which may be modulated by whether the left-hand region belongs to an occluding figure or the occluded background. The timing of this modulation suggests that it is the result of feedback from higher visual areas that extract figure-ground relations from the image – an example of the role played by recurrent loops in computing information about the world from information in the image.

Recognition

One important end-point of the processing of spatial pattern is the visual recognition of objects, faces, and scenes. Damage to the temporal lobe of the brain can specifically impair object recognition (agnosia), sometimes limited to a particular class of objects, notably a lost ability to recognize faces (*see* Prosopagnosia). Functional imaging studies show a network of areas that is selectively activated during the processing of faces (in particular, the fusiform face area on the lower surface of the temporal lobe) (*see* Face Recognition, Psychological and Neural Aspects), and another that is active in visual recognition of locations (the parahippocampal place area). Recordings from individual neurons are not generally possible in humans, but in the temporal lobe of the monkey, Gross, Rolls, and others found single neurons that respond specifically to faces. The pattern of activation over a small number of such cells can convey sufficient information to distinguish one individual from another. The inferotemporal region of the monkey brain contains cells that respond to features that can define different types of object. In both face- and object-selective neurons, the neuron's response shows position generalization; that is, it is selective for a particular class of stimulus, irrespective of its position within a wide receptive field.

Thus, the principles of a sparse population code, found in the representation of local pattern in V1, appear to apply also to the higher-level representation of visual identity (reviewed by Thorpe in Kriegeskorte and Kreiman, 2011; see also Barlow, 1995)). There must be a series of neural transformations that assemble information about local pattern features, to identify the characteristics that define a particular object category. Investigations of neurons along the ventral processing stream (see below) are beginning to reveal some of the two- and three-dimensional elements which are used to define perceived shape – for example, neurons which encode the location of sharply curved points around an object's contour (reviewed by Kourtzi and Connor, 2011). However, the hierarchy of neural transformations that connects the representation of shape elements with the identification of objects is still not well understood.

Spatial Location and Layout

As well as allowing the identification of objects and scenes, vision defines the space within which objects are located and are the targets of actions. Disturbances of spatial abilities are found with damage to the parietal lobe in both human and monkey brains (Ungerleider and Pasternak, 2004). These disturbances are quite distinct from the impairment of object recognition seen from temporal lobe damage, leading to the proposal of a broad division between a *ventral stream* of cortical visual processing, from V1 to temporal lobe and concerned with identifying what we are looking at, versus a *dorsal stream* from V1 to parietal lobe which conveys the information needed to know where it is and how we can act upon it. Different specialized visual areas contribute to the two streams; in particular, the motion area V5 (MT) sends information primarily via the dorsal stream (*see* Visual Streams: Dorsal and Ventral).

Spatial information requires a frame of reference. Area V1, and other early visual cortical areas, appear to be organized within a retinotopic frame – that is, the receptive fields, and their topographic arrangement across the cortex, are defined in terms of image locations on the retina. As the eyes are constantly moving in the head, a stable representation of space requires more than retinotopic information. Eye direction must also be factored in, to establish a head-centered frame. Such combination of retinal and eye-direction information occurs in a number of parietal areas. However, the properties of cells in different parts of parietal cortex suggest that they are preparing visual information in an appropriate form for controlling different action systems – for example, eye movements, reaching and grasping by the hands, and locomotion. Parietal neurons appear to be quite flexible in the use of eye-, head-, and body-centered frames, although it has been suggested that the retinotopic frame may serve to communicate between different action systems. An alternative frame of reference is allocentric, that is, referred to the layout of the environment rather than any part of the body of the observer. It has been suggested that this is a feature of the ventral stream, which has to identify objects independently of any particular viewpoint.

Parietal areas transmit information to motor-related areas of the frontal lobe, both directly and via the cerebellum. Comparison of cells in the parietal lobe with those in the

premotor frontal areas to which they project emphasizes how difficult it is to define any boundary between visual and motor processing. For instance, activity in some parietal neurons is associated both with viewing a particular three-dimensional shape (e.g., cylinder) and with the grasping action required by that particular shape. 'Mirror neurons' in premotor cortex respond both in association with the hand action of the monkey, and when the monkey is passively viewing another hand executing the same action. The linked parietal and frontal systems are best thought of as a set of visuomotor circuits which embody the spatial information needed to control and define a particular type of action, converted between reference frames as necessary.

The Binding Problem

Research on visual neuroscience has emphasized the fragmentation of brain processing. Any visual stimulus will activate many different neurons. Distinct brain areas and subsets of neurons process different attributes such as shape, color, and motion, and there are multiple spatial representations in different visuomotor circuits. In contrast, our experience of the visual world appears unified, and the different attributes of a visual object are bound together – if a white cup and a green book are together on the table, we do not confuse which color belongs to which object. How this perceptual unity arises from fragmented neural representations is known as the binding problem (Treisman, 1996).

Two, not necessarily incompatible, answers have been widely discussed. One is based on Anne Treisman's feature integration theory of attention. This theory, supported by evidence from visual search, proposes that a 'spotlight' of focal attention modulates processing within distinct 'feature maps' for different attributes (perhaps embodied in the multiple retinotopic areas of extrastriate cortex), and so multiple feature attributes are accessed only for the object that is the current focus of attention. The second idea is that cortical neurons show oscillations in their activity at frequencies around 40 Hz (the 'gamma-band'), and that temporal coherence between these oscillations acts as a label for different signals contributing to the representation to a common object. Such coding by temporal synchrony has attractions; it could provide a basis for understanding how features at different locations within a single map can be linked or segregated in object-based representations, as well as how binding occurs across different maps. However, the evidence for it remains limited and controversial. Gamma-band activity in electroencephalographic (EEG) recordings of large cell ensembles shows intriguing variations with perceptual processing, but its functional role is not yet understood.

Neural Basis of Visual Sensitivity and Thresholds

Much detailed knowledge of vision comes from the analysis of psychophysical thresholds – measures of how sensitive the visual system is in making fine discriminations (reviewed by Braddick, 2010). If these data are to be integrated with knowledge from neurophysiology, the relation between psychophysical threshold and patterns of neural activity needs to be understood.

'Threshold' conveys the suggestion of a stimulus intensity below which no neural activity is elicited. However, in practice, sensory neurons are rarely completely silent, and their activity, even with a fixed stimulus, shows variability or 'noise.' Thus, the ability to detect weak stimuli is determined by the statistical problem of distinguishing a stimulus-driven change in activity from those due to random noise.

Several studies have determined the 'thresholds' of single neurons in monkey visual cortex, in terms of the stimulus level which achieves a statistically defined level of reliable response. These can be compared with behavioral performance in detection of a trained animal; in some cases behavioral and neural measures can be made simultaneously. The information carried by single neurons proves to be remarkably close to that implied by the behavioral responses. Remarkable, because the animal has the opportunity to pool the information from thousands of neurons, which would be expected to yield much more reliable information. The data can be reconciled by realizing that although large numbers of neurons may be involved, the random variations in their activity are correlated so the number of genuinely independent neural signals is much smaller (Newsome et al., 1995).

Whether a particular judgment depends on pooling activity in tens, hundreds, or thousands of neurons, there must be many more neurons in the same brain area whose activity is irrelevant because they do not respond selectively to the stimulus concerned. If activity of such neurons was included in the pooling process, it would introduce noise and degrade sensitivity. Thus sensitive discrimination, either in the laboratory or in real-life situations, must depend on selecting a small minority of neurons whose output is informative for the task in hand – likely one of the skills that is acquired in perceptual learning.

Modulation of Neural Activity by Attention and Task Demands

It could be conceived that neural activity, relevant to a particular task, was channeled to appropriate decision and response mechanisms through a purely 'bottom-up' process. However, the Section Representation of Spatial Pattern and Shape has mentioned the ample evidence for 'top-down' control of visual processing, modulating activity at many levels according to the central decisions about what is relevant. For instance: (1) neural activity, as early in the pathway as the LGN, is strongly modulated according to the sleeping–waking state of the organism; (2) the response of single neurons in areas such as V4 and V5 in monkeys can be changed radically according to whether the preferred stimulus of the neuron is behaviorally relevant or irrelevant on a particular occasion; (3) functional brain imaging shows that the overall activation of an area such as V5 in humans is enhanced by attention to the property that it analyses (in this case, motion), even if the stimulus is unchanged; and (4) within a topographically organized area such as V1, activation of a particular region is enhanced when attention is directed to the corresponding region of visual space (*see*

Attention, Neural Basis of). Thus the neural mechanisms of visual perception work in a flexible way, directed by the processes of attention and executive function to be adaptive for the current goals of behavior.

Development and Plasticity of Visual Neural Organization

The organization described in the preceding sections is the product of a developmental process. The vision of the newborn baby has low acuity, in part because the foveal cone receptors are very immature and loosely packed, but also because most of the connections between neurons in the brain are not yet established. Over the first few months of life, V1 becomes functional as the characteristic tuning of the V1 neurons starts to develop, first for orientation followed by motion and then stereoscopic depth (Atkinson, 2000; Braddick and Atkinson, 2011) This is followed by the development of integrative processing in extrastriate cortical areas (Braddick and Atkinson, 2011). The dorsal and ventral cortical streams do not develop in step; global integration of motion in the dorsal stream emerges before corresponding processes for static form in the ventral stream, but this is followed by radical reorganization of brain networks, in which the dorsal stream appears to be particularly vulnerable in a range of developmental disorders, ranging from Williams Syndrome and autism to very premature birth and congenital cataract.

This vulnerability is an example of the plasticity of the visual system in development, which includes long-term adaptive changes, in response to the visual inputs received by the system. For instance, if the input to visual cortex has an anomalous distribution across the retina (e.g., as a result of a congenital absence of cones in the central fovea) the usual retinotopic map in V1 is distorted to match this distribution. When the balance of input from the two eyes is disturbed (e.g., when the input of one eye is reduced in a critical period during early development) the cortex adjusts to receive less input from the less active eye. The plasticity seen in these abnormal cases must also play a role in normal development: development of the normal neural organization of visual brain areas must depend on the regularities and correlations in the visual input that these areas receive.

Perceptual Learning

The Sections Modulation of Neural Activity by Attention and Task Demands and Development and Plasticity of Visual Neural Organization indicate ways in which the function of the visual system can be dynamically modulated, both from moment to moment and more permanently. This is also shown in the strong learning effects found for visual abilities; almost any visual discrimination becomes strikingly more sensitive with practice. This learning can be highly specific, for example, showing improvement for detecting stimuli only at the orientation that has been practiced, suggesting that changes occur at relatively low levels of visual processing such as V1. The discussion in the Section Neural Basis of Visual Sensitivity and Thresholds suggests that such learning may be based on establishing highly specific connections to decision mechanisms from the neurons most sensitive to the stimulus concerned. However, while changes with learning have been demonstrated in the responses of visual cortical neurons, the relation of these changes to modified neuronal connections is still speculative.

The Neural Basis of Visual Awareness

The detailed understanding of how visual information is represented in the nervous system does not answer the question: what aspect of neural activity corresponds to the conscious awareness of the observer? Many aspects of vision are not accessible to consciousness, even though the relevant information exists within the nervous system. For example, we cannot consciously tell whether a stimulus was delivered to the left or right eye, even though the two eyes activate distinct neurons in V1 and the distinction is critical for stereoscopic depth perception. This raises the question of whether there is any subset of visual neurons whose activity contributes specifically to awareness, constituting the so-called Neural Correlate of Consciousness (*see* Consciousness, Neural Basis of).

Individuals with damage to V1 report that they are blind in the region of the visual field corresponding to the damage. However, they can sometimes make reliable responses to stimuli they do not report seeing, a phenomenon named 'blindsight' by Weiskrantz and Warrington who discovered it in the 1970s. Thus there must be routes for visual information which can control behavior but whose activity has no conscious correlate – a pathway is known from the superior colliculus in the midbrain, bypassing V1 en route to the motion-sensitive area V5, which could provide such a possible route. However, it is not known whether V1 activity is itself a necessary correlate of conscious visual experience, or whether V1 simply serves as the only access route to higher processes that have such a correlate; there are difficulties with either alternative.

In binocular rivalry, discrepant patterns are presented to the two eyes, and alternate in conscious perception. Neuronal activity can be monitored in monkeys who are trained to indicate which pattern is perceptually dominant. For some cortical neurons, activity fluctuates in correspondence with perceptual dominance, but for others there is no correlation. The proportion in the former group increases in going from lower to higher cortical areas (Logothetis, 1998). These results suggest that awareness is not uniquely associated with particular cortical areas, but may emerge cumulatively as information is processed through the system. The true relation between brain and consciousness will remain speculative for many years, but visual processing is likely to provide the key test cases for this central problem of the human sciences (*see* Consciousness, Neural Basis of).

See also: Attention, Neural Basis of; Cerebral Cortex; Computational Neuroscience; Consciousness, Neural Basis of; Eye Movement, Control of: Oculomotor Control; Face Recognition, Psychological and Neural Aspects; Functional Brain Imaging of Language Processes; Neural Representations

of Objects; Neural Synchrony as a Binding Mechanism; Occipital Lobe (Visual Cortex): Functional Aspects; Perception of Action Space: Using Multiple Frames of Reference; Topographic Maps; Visual Streams: Dorsal and Ventral.

Bibliography

Atkinson, J., 2000. The Developing Visual Brain. Oxford University Press, Oxford.

Barlow, H.B., 1995. The neuron doctrine in perception. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 415–435.

Braddick, O., 2010. The psychophysical approach. In: Goldstein, E.B. (Ed.), Encyclopedia of Perception. Sage Publications, Thousand Oaks, CA.

Braddick, O., Atkinson, J., 2011. Development of human visual function. Vision Research 51, 1588–1609.

Bullier, J., 2004. Communication between cortical areas of the visual system. In: Chalupa, L.M., Werner, J. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 522–540.

Churchland, P.S., Sejnowski, T.J., 1992. The Computational Brain. MIT Press, Cambridge, MA.

De Valois, R.L., De Valois, K.K., 1988. Spatial Vision. Oxford University Press, New York.

Frisby, J.P., Stone, J.B., 2010. Seeing. MIT Press, Cambridge, MA.

Kourtzi, Z., Connor, C.E., 2011. Neural representations for object perception: structure, category, and adaptive coding. Annual Review of Neuroscience 34, 45–67.

Kriegeskorte, N., Kreiman, G. (Eds.), 2011. Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging. MIT Press, Cambridge, MA.

Logothetis, N.K., 1998. Object vision and visual awareness. Current Opinion in Neurobiology 8, 536–544.

Newsome, W.T., Shadlen, M.N., Zohary, E., Britten, K.H., Movshon, J.A., 1995. Visual motion: linking neuronal activity to psychophysical performance. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA, pp. 401–440.

Peterhans, E., von der Heydt, R., 1991. Subjective contours – bridging the gap between psychophysics and physiology. Trends in Neurosciences 14, 112–119.

Treisman, A., 1996. The binding problem. Current Opinion in Neurobiology 6, 171–178.

Ungerleider, L.G., Pasternak, T., 2004. Ventral and dorsal cortical processing streams. In: Chalupa, L.M., Werner, J. (Eds.), The Visual Neurosciences. MIT Press, Cambridge, MA, pp. 541–562.

Zeki, S., 1993. A Vision of the Brain. Blackwell Scientific, Oxford.

Visual Sociology

Piotr Sztompka, Jagiellonian University, Krakow, Poland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Visual sociology is a new and rapidly expanding field of sociological research and theory. It is considered as either (1) a new subdiscipline of sociology concerned with the iconosphere and the visual sociosphere of contemporary society or (2) as a methodological orientation supplementing standard methods of qualitative research, applicable to all substantive areas of sociology. The emergence of visual sociology reflects an important shift in the culture of modern society, from written texts to images. Both the purposefully created pictures and the visible, outer layer of social behavior are rich and dense. Interpersonal communication is also pervaded with images. We live in the exhibitionist and voyeuristic society, and this is manifested in six ways: as the society of the icons, of the spectacle, of the self-presentation, of the design, of the surveillance, and of gazing. Visual sociology explores these new features and uses the cognitive opportunities they provide. As a young direction of inquiry, it grapples with some philosophical dilemmas: realism versus constructivism, and universalism versus relativism. Irrespective of these controversies, the main message of visual sociology is to look around and to develop visual imagination, that is, the skill of translating abstract concepts and theories of sociology into their observable manifestations in everyday life, and to interpret the iconosphere and visual sociosphere in terms of sociological concepts and theories.

The Visual in the Focus of Current Debates

The visual aspects of social life – that is, all that is accessible to sight as well as all social phenomena and processes related to seeing and being seen – have become one of the central concerns of the contemporary social sciences. Some scholars believe that this signals a new turn, similar to the linguistic turn in the philosophy of the 1970s: "visuality is emerging as a central topic of discussions in the human sciences in the way that language did …. The focus shifts from words to pictures, the sayable to the seeable" (Mitchell, 2006: 9, 10).

In 1994 two scholars, working independently, gave the emerging trend its labels. The Swiss art historian Gottfried Boehm coined the term 'iconic turn,' and the American literary critic W.J.T. Mitchell introduced the idea of the 'pictorial turn.' Boehm focused mainly on the changes at the intellectual level: a new approach to the visual domain departing from the traditions of the history of art. The main point is to go beyond the sphere of art and extend research to all sorts of visual representations, establishing the more comprehensive 'image science' (Boehm, 1994). This is justified by the fact that "pictures privileged by the descriptive label 'art' constitute only a small portion of the sea of imagery in which the world cultures are adrift" (Moxley, 2008: 136).

Mitchell, on the other hand, calls for the departure from the traditional perspective that treats pictures only as vehicles of meaning, as representations of other objects, toward recognizing that they have some independent status within culture, 'lives' and 'wishes' of their own, and that they 'want' something and 'say' something to us. As such they possess some agency, exerting influence on people and their behavior. Therefore they should be treated as valid objects of inquiry per se, and not only as indicators of some other objects such as meanings or structures (Mitchell, 1994). This idea was taken up later by the proponents of the so-called performative approach, Jeffrey Alexander and others (Alexander et al., 2006, 2012).

Both Boehm and Mitchell, but even more their followers, recognize that such changes at the intellectual level are not autonomous but reflect changes in the society of our time, in the period usually called late modernity, high modernity, or postmodernity (Giddens, 1990; Jay and Brennan, 1996). They join a number of scholars who perceive the shift in popular culture and everyday life toward pictures and away from texts. Some, more sociologically inclined, observe the proliferation of pictures as forms of communication and representation. Others, closer to psychology, claim that images and not only words become the means of human thinking. The emergence of the iconic language is supposed to signify the end of the 'Guttenberg Galaxy,' the rule of the printed word. As a famous photographer, Andreas Feininger, puts it: "In contrast to the printed page, a photograph can be 'read' at a glance – and people seem to get more and more restless, impatiently waiting to go on. The language of words – reading – is slow and generally losing its importance in favor of the language of the image – TV, pictorial advertising and photography" (Feininger, 2004: 4). Already at the end of the 1970s and the beginning of the 1980s, the work of Jean Baudrillard (1994), Roland Barthes (1983), and Susan Sontag (1978) emphasized this striking ontological feature of contemporary society, namely its saturation with pictures, images, the expanding iconosphere, as well as tremendous enrichment of the directly observable outer layers of social life, of the visible sociosphere. Modern, but particularly late modern capitalist, consumer society is particularly rich in visibility, saturated with pictures and icons, bristling with images, diversified and colorful visual impressions. And people are paying more attention to external, visible layers of social reality. We live in the exhibitionist and voyeuristic society. Sociologists could not remain insensitive to these changes. As Douglas Harper puts it, "making sociology visual was a parallel to making society visible" (Harper, 2012: 2, 3).

The Exhibitionist and Voyeuristic Society of Our Time

The growing importance of the visual domain is manifested at six levels of everyday life. First, we live in the *society of the icons*. We are surrounded with lots of intentionally created pictures: billboards, commercials, advertisements, banners, flags, coats of arms, color magazines, TV programs, the Internet, and multimedia presentations. Not only in tabloids, but in serious journals or even daily newspapers, photographs appear together with the written texts. Interestingly, the photo editors are also influenced by the Windows style, and instead of the full-page photos in the old days of *Life*, *Look*, *Paris Match*, and the *Illustrated London News*, we have large number of small images resembling another webpage on the Internet. The icons in all their variety are representations of something, but their main function is not to describe but to communicate some meanings, convey a message, 'make a statement' of some sort, and they have been created for that purpose. Their connotation is more important than their denotation.

Second, we live in the *society of the spectacle*. Many of the collective situations in which people find themselves are orchestrated according to some scenarios and occur in carefully designed settings. The rich visuality of social situations and social scenes is intended to impress, to convey directly visible, obvious meanings. In the domain of leisure, just think of the rock concerts where music no longer dominates and the lights, costumes, hairdos, choreography, fireworks, mists, and smokes are central elements of the performance. Or consider carnivals, celebrations of the New Year. In the area of sport, competitions turn to spectacles watched by millions of fans, as in the opening ceremonies of the Olympics, the World Cup in soccer, or the Rose Parade at Pasadena. In the area of the economy and consumption, just remind yourself of the shopping galleries or malls designed like baroque churches or royal palaces, shop windows arranged like theater stages or opera settings, Christmas markets, bank halls resembling museums, and executive offices with huge mahogany tables and carefully arranged family portraits. In the domain of politics, just look at the party conventions, presidential inaugurations, parliamentary rituals, patriotic parades, mass rallies, protest gatherings, and street riots. Even the domain of science has not escaped the tendency. Perhaps the most obvious case is that of Nobel Prize ceremonies, but the opening of each major international congress or academic conference is now accompanied by all kinds of ritual or even art performances. Science has also developed some characteristic formats of spectacles in which the debates are conducted or results presented: panel discussions, roundtables, and presidential addresses with large audiences. The educational, university context is also rich in spectacles. The commencements, parades of professors and students through town dressed in all these strange gowns and hats, and the honorary degree ceremonies, are rich in visuality and carried out according to strict scenarios. And it is hard today to attend a university lecture that does not use visual illustrations with the help of PowerPoint. Even the domain of death illustrates the trend. Funerals, particularly of public heroes or celebrities, turn into spectacles. No wonder that in most of such collective occasions we see hordes of press photographers. It is easier to grasp the meaning of these situations in pictures than in words.

Third, we live in the *society of the display*. The presentation of the individual, autonomous self, the universal human tendency so aptly demonstrated by Erving Goffman (1959), has found much more extensive and enriched manifestations. The syndrome of Thorstein Veblen's 'conspicuous consumption' (1994 [1889]), which in the old days was a way of life of the elites, *nouveau riches*, or celebrities, has become universal. The external, observable requisites – the cars we drive, houses we live in, dresses and shoes we wear, watches on our hands, jewelry on our necks, pens we write with, leather suitcases in which we carry office papers – are the symbols of status and building blocks with which we construct our identity, to be admired and envied by others. The celebrities whose lifestyles have become more accessible than ever through tabloids and snobbish magazines are serving as role models to be imitated. Famous brands have become democratized through franchising; they are much cheaper and more accessible than before. And for those who still cannot afford Rolex, Omega, Louis Vuitton, or Dolce y Gabana there are cheap fakes from Asia easily available in the street markets. In extreme cases, social identity is constructed exclusively on such grounds: there are communities of taste, of style, of trends and fashions, fans of sport cars, motorcycles, or boats, devotees of certain brands, demonstrating strong forms of solidarity formerly reserved to the believers in the same God or in the same Nation.

Fourth, we live in the *society of the design*. The attention devoted to the packaging, shape, and color of various products (just think of cars, boats, mobile phones, computers, skis, dresses, and ties); furniture of apartments; architecture of houses; and urban planning of streets, parks, stadiums, and airports is incomparable to any earlier era. The experts in design are everywhere and in great demand; it is their time. Everyday life of consumer society is saturated with tremendous diversity and variety of intentionally created visual impressions: of colors, forms, styles, and fashions. Consumers respond accordingly. We very often buy various products attracted by their design, and we often change still functionally efficient gadgets, TV sets, cars, computers, and cameras because of their new design, in order to catch up with novelty. This tendency is driven by the logic of the commercial competitive market, the need for novelty, originality, standing out, drawing attention, and even shocking the consumer.

Fifth, we live in *the society of surveillance*. The penetration of state coercive, regulatory, or administrative power to many corners of our lives occurs often by means of video surveillance: the hidden cameras in factories, stores, banks, hotels, offices, gas stations, streets and squares of the cities; and photo-controls of speeding on the roads. Criminal investigators use hidden video recordings as proofs at courts, arrange provocations in corruption cases, and so forth. We are amazed looking at Google Maps at how detailed the observations are that are provided by all those satellites flying invisibly over our heads. We are seen and observed by others without our awareness much more often than ever before. We are also made aware of social control and sanctioning of deviant behaviors by visible signs demonstrating binding rules, demanding certain behaviors or prohibiting others: traffic signs, and iconic signs at the airports, in railway stations, and on the streets.

Sixth, we live in the *society of gazing*. We avail ourselves of these new, immense opportunities of visual experiences, by intensifying our practices of observing, looking at others and the surroundings. The custom of Paris street cafes of leisurely looking at passersby has extended to many cities in the world. In the streets of any city – Rome and Paris, Amalfi and Florence, New York and Shanghai – what the people do most often is look at shop windows and take snapshots of everything. Taking the little digital cameras to their eyes has become the most frequent and common gesture of the human species. And returning home, we sit and look at TV, look at pictures on the Internet, glance through illustrated magazines, play computer games, and so on. A particular variety of this is gazing at movie stars, celebrities, or the 'rich and famous.' Crowds do it directly at film festivals, Oscar ceremonies, and the like; and readers of tabloids do it indirectly by looking at pictures of 'high society' with a mixture of awe and envy.

These new phenomena provide particular opportunities for sociological inquiry. All that we see around us, and not only what is to be found in pictures, may become the subject of study. Such inquiry may serve two purposes. First, it extends the field of research, the category of objects to be studied. We begin to treat pictures and observed images of social life as phenomena to be studied as intrinsically interesting parts of culture, the visual culture. They are the ultimate targets of research. But the iconosphere and visual sociosphere may also serve as means of research, as resources helpful in digging into the structures, mechanisms, and regularities of social life, a new research opportunity opening for social scientists due to the enhanced visuality of social life. Here the visual turn means adding a new approach (Prosser, 1998), a new orientation to those in the standard repertoire of the social sciences. In the discipline of sociology, the concern with visuality is marked by the birth and rapid development of the new area of research called visual sociology.

Emergence and Career of Visual Sociology

Sociology's concern with visuality is relatively new, and dates only to the last decades of the twentieth century. Among the classical authors, only Georg Simmel wrote about the observability of social interaction and its implications for sociological method in an essay published in 1908. He was particularly concerned with everyday life in the city, in the urban context, by its nature full of visual richness and diversity. He is also remembered for a striking remark that the eye is for a sociologist the most important of the senses (Simmel, 1921 [1908]: 358). But for the whole twentieth century the domination of techniques and procedures based on words and texts – surveys, questionnaires, interviews, content analysis, and more and more sophisticated statistical methods of interpretation – pushed visual data to the margin. It was only in the 1970s, when the so-called subjectivist turn toward symbolic interactionism, phenomenology, ethnomethodology, and so-called dramaturgical theory raised interest in the subjective data and qualitative methods, that the first large-scale photographic projects appeared in sociology. Their authors – Dough Harper, Bill Aron, Phyllis Ewen, Bruce Jackson, Mark Rosenberg, Edward Hall, and Erwin Zube – document the social life of various communities, with particular attention given to the more 'exotic' or deviant subcultures such as homeless people, vagabonds, hippies, or religious sects (Wagner, 1979; Collier and Collier, 1986; Harper, 2012). The year 1972 witnessed the birth of reflexive photography, known also as a collaborative visual method: Sol Worth and John Adair give simple cameras to some of the Navaho Indians and elicited spontaneous visual evidence of their way of life, systems of values and rules, and cherished objects as seen by the insiders (Worth and Adair, 1972; Collier and Collier, 1986). This approach has been used later on many other occasions.

The truly classical forerunner of visual sociology is Erving Goffman. He was the first to submit rich photographic material to a deep interpretational analysis aimed at unraveling hidden cultural premises and symbolic meanings. In his 1979 book, *Gender Advertisements*, he collected 508 photographs from the media, commercials, and private archives representing men and women separately and in various interactions in many contexts of everyday life. The expressive behavior captured in the pictures is, he believed, a good indicator of deeply hidden, often subconscious, culturally accepted standardized definitions of gender (1979: 7). According to these culturally imposed and internalized definitions, people present their selves to others on social stage. Being a man or a woman is expressed by self-presentations congruent with cultural images of masculinity or feminity.

But the founding event of visual sociology as a new, institutionalized discipline was the exhibition titled "Exploring Society Photographically" organized by a pioneer of the field, the well-known sociologist Howard Becker at Northwestern University in Evanston, Illinois in 1981 (Becker, 1981). Among 12 photographic projects included in the exhibition, most were of a clearly sociological character: the study of life in an Arkansas penitentiary, of migrant fruit pickers, of a community of elderly Jews; a look at the roller-skating culture of Venice, California, at patients in the hospital, at various work experiences, and the assimilation of American Indians. The exhibition catalog provided sociological interpretations of photographs as a method of unraveling hidden cultural and social patterns. In this way it set the standards for later similar efforts. Another milestone was a conference at Amsterdam in 1989 and the accompanying exhibition titled "Eyes Across the Water" (Flaes, 1989). Since that time, visual sociology has developed quite rapidly and acquired institutional and organizational forms. In the mid-1980s the International Visual Sociology Association was established. Specialized journals started to appear: in 1991 *Visual Sociology*, later renamed *Visual Studies*; *Journal of Visual Culture*; and others. A series of conferences followed, classes and seminars were opened at various universities, sections were created at some national sociological associations, and quite recently (2013) the Working Group on Visual Sociology was established within the International Sociological Association (ISA), well on the way to becoming a full-fledged ISA research committee.

There are several ways in which visual sociology is practiced. One tendency is treating visual sociology as one of the specialized subdisciplines with its own objects of study and research methods. There is a new 'tribe' of sociologists whose

self-defined identity is as visual sociologists (Nathanson and Zuev, 2013; Harper, 2012). But there is also the different approach in which visual sociology is considered as a general methodological orientation applicable in all specialized fields of sociological research. There is some parallel with other, rapidly developing areas of sociological interest: historical sociology or the sociology of knowledge (as distinct from the sociology of science). In spite of the misleading names, rather than narrow specializations, both are in fact wide theoretical and methodological orientations that can be applied in a number of – if not all – disciplines of sociology. The same is true of visual sociology. Treated in this way, it is only distinguished at the methodological level and not at the substantive level (Rose, 2001). It is relevant for many substantive areas, where the iconosphere, visual sociosphere, or visual practices are rich enough – perhaps most obviously for the sociology of everyday life and of interpersonal behavior; but also for urban sociology, sociology of work, sociology of religion, sociology of leisure, sociology of sports, sociology of social movements, and many other subdisciplines. In most of them the visual orientation comes down mainly to interpretational procedures (Sztompka, 2006/2012, 2012).

Those who consider visual sociology as a separate discipline still differ as to its proper objects of study. Some limit its target to the iconosphere, the purposefully created pictures communicating some meaning or carrying a certain message encountered in the environment of everyday life (Sztompka, 2008). Among them the central place is usually reserved for photography, but other types of visual representations are also included: video recordings, movies, TV pictures, computer presentations, and even murals, graffiti, tattoos, photo-blogs on the Internet, and so forth. The wider approach also includes the visual sociosphere, that is, not only purposefully produced pictures but the whole visual, observable layer of social life, reflected as images in the minds of the people. As Emmison and Smith declare: "Visual inquiry is no longer just the study of the image, but rather the study of the seen and observable" (2000: ix). The study of the iconosphere and visual sociosphere requires interpretation, the search for meaning, either what the pictures and observable manifestations of social life represent (denotation) or what they imply, suggest, and symbolize (connotation).

A somewhat exotic focus has become fashionable recently. It is no longer concerned with the iconosphere or visual sociosphere and its content, or meaning, but rather on social practices through which pictures, icons, or images are produced or perceived (imagined), distributed or shared, preserved or remembered. The proposed methods of research are congruent with this approach. The study of the visual practices of depicting and perceiving treats visual data as any other cultural component, and approaches visual data not from the internal perspective of meaning but from the external perspective of the cultural and social contexts or 'discourses' in which they appear, including even the material aspects of certain tangible objects (e.g., snapshots) that people handle and use in specific ways. It proposes looking 'behind the pictures' rather than 'at the pictures' or 'through the pictures' (Wright, 2008).

Current Debates

More deep and fundamental divisions within visual sociology have to do with philosophical assumptions, whether consciously adopted or only implied by the research modes. There are two frontiers of the battle: realism versus constructivism and universalism versus relativism. The early forerunners of visual sociology – Roland Barthes, Susan Sontag, Margaret Mead, Erving Goffman, as well as classical photographers Edward Weston, Laszlo Homoly-Nagy, August Sander, and Andreas Feininger – accepted a realist perspective: photographs represented something that really exists in front of the lens, the objective facts, phenomena, and events (Cartier-Bresson, 1952). For Sontag, photographs were the "miniatures of reality" (1978: 4). The photographic techniques (at least of the traditional analog form) assure that the visual, observable features of reality are reflected, imprinted on the light-sensitive paper. Photographs therefore constitute a proof, the strong evidence of something that happened. In the same vein, the sociological photographs capture and preserve the looks and behavior of the people, individually or collectively; the social events and occasions that actually occurred; social situations and civilizational environments as they really existed: "photography is the somehow natural witness of 'what has been'" (Wright, 2008: 17).

At the opposite pole to the realist perspective are various versions of constructivism (Berger, 1972). In the neo-Kantian mood they claim that the world does not exist per se, but is to a great extent the construct of the people who act in the world, perceive the world, or describe it or theorize about it. In a similar way to which linguistic champions considered reality as shaped or at least filtered by the language used by the people, some proponents of the visual perspective and some representatives of visual sociology treat pictures, icons, and images as shaped or at least distorted by the agents involved in visual practices. 'Cameras don't take pictures, people do' is the main argument. And people are subjective and culturally shaped creatures; they cannot escape their minds or their cultural milieu, and both are reflected in the pictures they take. Photographs, they say, are not representations of something existing independently, but presentations, emanations, embodiments of at least five circumstances:

1. The subjective reasons, intentions, and emotions – in brief the whole mind-set of the photographer.
2. The social status, role, and identity of the photographer – for example, as professional or amateur.
3. The cultural conventions, rules, and values constraining or facilitating certain photographic themes, aesthetic styles, or perspectives.
4. The wider context in which the picture is taken, such as travel, sightseeing, or family ritual.
5. The concretely situated interaction of the photographer and the subject – for example, the voyeurism of a hidden camera versus an open partnership.

The middle of the road position, closest it seems to common sense, may be called critical realism. It claims that there is a real world out there, and also the social world consisting of living people, their actions and interactions, the

material effects of their activity including the civilizational and technical environment, and also the organizations and institutions they build. But photographs are not simple mirror reflections of this world; they involve considerable selectivity and biases; they are constructed reflections depending upon subjective as well as cultural contexts. The first choice refers to the topic or theme of a picture. This is of course a personal decision linked with intentions, purposes, and interests. The theme of the picture may also reflect cultural conventions and rules: what is the right topic, what observations or experiences should be preserved, and so on. Then the second choice is the focus on an element that is considered central, that should be sharp and put in the middle or in another strategic location. The third decision has to do with framing. This involves some idea of priorities, because by framing we include what we consider important, worth recording, and cut off everything that lacks significance. Then the fourth and even more subtle decision concerns the depth of the field. This reflects further priorities of significance in the area already selected by framing. What is the foreground, and what can be left less sharp and clear as the background? Recent photographic technology, assuming that mass amateur photography is usually capturing people (family members or friends), makes this decision automatic by the tricks of so-called face recognition and moving autofocus tracing the actions of the people present. But switching on or off such features is the photographer's decision. The fifth choice refers to the lens we use. Are we interested in the context, environment, surroundings of the event or person captured, and hence we use a wide-angle lens; or are we only concerned with the object or person in its intrinsic details, detached from context. When some authors claim that for sociological photography the wide-angle lens is most useful, behind this choice there is a theoretical assertion that social events are always situated in complex contexts and cannot be properly understood ('seen') in isolation. The sixth choice is the shutter speed. It is related to our assumptions concerning the various speeds of social events in various contexts of social life. As the very term 'social life' suggests, when shooting society we are shooting movement, change. But the movement is of diverse tempo, rhythm, and duration. To grasp it adequately, the proper shutter speed is crucial. Finally, seventh, we have to make the most important decision, when to press the shutter. Social life is a fluid process, an incessant sequence or sometimes a chaos of events. To take a still photograph is to freeze that movement. As a famous press photographer, Henri Cartier-Bresson, emphasized, the mark of a great photograph is the right choice of a fleeting moment, which will never repeat itself but which is most loaded with meaning, most indicative of the problem we want to represent visually. Usually there is only one 'right moment.' Thus a middle-of-of the road perspective of critical realism, not falling into the trap of solipsism or pure constructivism, asserts that the photograph is of something real, which occurred in front of the lens, but admits at the same time that it is a selective reconstruction based on at least six choices taken by the photographer.

Another line of controversy of the fundamental, philosophical sort is the debate between universalism and cultural relativism. Universalism claims that pictures provide the language that can be understood across cultural or civilizational divisions. In other words, pictures are interculturally translatable. Such was the typical view of outstanding photographers. Andreas Feininger claims that "photography is picture-language, the only language that can be understood anywhere in the world" (2004: 3). And Martin Jay argues that there is "the capacity of the visual to break free from linguistic and cultural constraints." In his view, this is demonstrated in "the relative ease with which images cross the line," which "undermines the assumption that they are entirely relative to specific cultures, which are incommensurable with their neighbors" (Jay, 2002: 274). This intrinsic quality of pictures is enhanced by the process of globalization. It implies, among other consequences, the homogenization of cultures, or at least their hybridization or 'creolization' (Hannerz, 1996; Tomlinson, 1999). It means that no culture is now a self-contained, isolated unit, but rather a kind of a cultural melting pot: each consists of various components originating in many cultures. They provide bridges with the cultures of origin and facilitate intercultural translation. Pictures are the first to use this "transcultural mechanism of visual experience" (Jay, 2002), because they are not mediated by natural language, which is usually the most resistant core of autonomy within a culture. The visual culture is most sensitive to globalization (Mirzoeff, 1999; Mitchell, 2002).

The opponents to this view rallying under the banner of cultural relativism claim that the meaning of pictures, and also stylistic conventions, aesthetic preferences, perception rules, and reception regimes, are all shaped by culture, and cultures are unique wholes resisting intercultural mixing or translation. "It should be remembered that photography is a cultural product, despite its 'realistic' nature" (Wright, 2008: 50). Each picture may be understood only within the confines of its own culture, with the help of local cultural codes. There are various visual worlds separated from each other by impenetrable borders, a mosaic of visual monads.

In the debate between universalism and cultural relativism, the middle of the road position may be called incomplete universalism. The pictures, as more direct and unmediated vehicles of meaning, are more accessible for representatives of various cultures. Due to modernization and globalization, the ways pictures are taken become more uniform, as for example in the case of photography and a photo camera as a truly universal gadget. The content of photographs is also becoming similar, with the progressing homogenization of cultures and civilizations. All this makes 'reading,' interpreting, and understanding photographs more independent of cultural filters. But of course there remain local peculiarities, deep cultural stereotypes and prejudices, stylistic conventions, value priorities, and hierarchies of social issues – which become reflected both in the ways societies are depicted by photographs and in the manner in which photographs are contemplated and deciphered. In this area of remaining (or perhaps eternal) diversity, the claims of cultural relativism are relevant.

A Glimpse at the Future

The development of visual sociology is facilitated by the wider tendencies of its mother discipline. From a concern with abstract social systems and structures, sociology seems to move toward growing interest in everyday life and particularly the field of interpersonal relations. Simmel replaces Marx and Weber, and Goffman dethrones Parsons as the gurus of sociology for the twenty-first century. This new substantive focus is accompanied by the growing importance of qualitative methods instead of the purely quantitative and statistical style of earlier research.

For visual sociology this probably implies two trends. First, growing emphasis on visual methods as supplementary to more traditional qualitative research: recording observation by visual techniques, photo elicitation in interviews, the content analysis of the iconosphere, and studying private and family collections of photographs as a new type of personal documents (Prosser, 1998). Second, treating visual imagination as the necessary skill of all sociologists, independent of their field of research. And the visual imagination is a capacity of double translation: of pictures and visual impressions into sociological conceptual and theoretical frameworks; and the reverse, finding representations of sociological concepts and theories in images and observable features of social life. After all, there is not much more to society than what people do, individually and collectively. And most of what they do is visible. Such ontological commitments, typical it seems for recent sociology, provide a strong rationale for the future career of visual sociology.

See also: Documentary and Ethnographic Film; Photography as a Medium; Semiotics; Visual Images in the Media; Visual Methods in the Social Sciences.

Bibliography

Alexander, Jeffrey, Giesen, Bernhard, Mast, Jason (Eds.), 2006. Social Performance: Symbolic Action, Cultural Pragmatics, and Ritual. Cambridge University Press, Cambridge.
Alexander, Jeffrey, Bartmanski, Dominik, Giesen, Bernhard (Eds.), 2012. Iconic Power: Materiality and Meaning in Social Life. Palgrave Macmillan, New York.
Barthes, Roland, 1983. The photographic message. In: Selected Writings. Fontana, New York.
Baudrillard, Jean, 1994. Simulacra and Simulation. University of Michigan Press, Ann Arbor.
Becker, Howard (Ed.), 1981. Exploring Society Photographically. Northwestern University Press, Evanston.
Berger, John, 1972. Ways of Seeing. Penguin, London.
Boehm, Gottfried, 1994. Was ist ein Bild? Fink, Munich.
Cartier-Bresson, Henri, 1952. The Decisive Moment. Simon and Schuster, New York.
Collier, John, Collier, Malcolm, 1986. Visual Anthropology. Photography as a Research Method. The University of New Mexico Press, Albuquerque.
Emmison, Michael, Smith, Philip, 2000. Researching the Visual. Sage, London.
Feininger, A., 2004. A philosophy of photography. In: That's Photography. Katje Cants Verlag, Ostfilder-Kemnat, pp. 1–4.
Flaes, Robert Bonzajer (Ed.), 1989. Eyes across the Water: The Amsterdam Conference on Visual Anthropology and Sociology. Het Spinhuis, Amsterdam.
Giddens, Anthony, 1990. The Consequences of Modernity. Polity Press, Cambridge.
Goffman, Erving, 1959. Presentation of Self in Everyday Life. Doubleday, New York.
Goffman, Erving, 1979. Gender Advertisements. Macmillan, London.
Hannerz, Ulf, 1996. Transnational Connections. Routledge, London.
Harper, Douglas, 2012. Visual Sociology. Routledge, London.
Jay, Martin, 2002. Cultural relativism and the cultural turn. Journal of Visual Culture (3), 267–278.
Jay, Martin, Brennan, Teresa, 1996. Vision in Context: Historical and Contemporary Prespectives on Sight. Routledge, London.
Margolis, Eric, Pauwels, Luc, 2011. The Sage Handbook of Visual Research Methods. Sage, London.
Mitchell, Claudia, 2011. Doing Visual Research. Sage, London.
Mirzoeff, Nicholas, 1999. An Introduction to Visual Culture. Routledge, London.
Mitchell, W.J.T., 1994. Picture Theory: Essays on Verbal and Visual Representation. The Unioversity of Chicago Press, Chicago.
Mitchell, W.J.T., 2002. Showing seeing: a critique of visual culture. Journal of Visual Culture (1), 165–181.
Mitchell, W.J.T., 2006. An interview with W.J.T. Mitchell. Image and Narrative. Online Magazine of the Visual Narrative (15), 1–17.
Moxley, Keith, 2008. Visual studies and the iconic turn. Journal of Visual Culture 7 (2), 131–146.
Nathanson, Regev, Zuev, Dennis (Eds.), 2013. Sociology of the Visual. Routledge, London.
Prosser, Jon (Ed.), 1998. Image-Based Research. Routledge, London.
Rose, Gillian, 2001. Visual Methodologies. Sage, London.
Rose, Gillian, 2011. Visual Methodologies. An Introduction to Research with Visual Materials. Sage, London.
Simmel, Georg, 1921 [1908]. Sociology of the senses: visual interaction. In: Park, Robert, Burgess, Ernest (Eds.), Introduction to the Science of Sociology. University of Chicago Press, Chicago.
Sontag, Susan, 1978. On Photography. Farrar, Strauss and Giroux, New York.
Spencer, Stephen, 2011. Visual Research Methods in the Social Sciences. Routledge, London.
Spencer, Stephen, 2012. Visual Research Methods in the Social Sciences: Awakening Visions. Routledge, London.
Sturken, Marita, Lisa, Cartwright, 2001. Practices of Looking: An Introduction to Visual Culture. Oxford University Press, Oxford.
Sztompka, Piotr, 2006/2012. Socjologia Wizualna: Fotografia Jako Metoda Badawcza (Visual Sociology: Photography as a Research Method), second ed. Wydawnictwo PWN, Warszawa (In Polish, also published in Russian and Czech).
Sztompka, Piotr, 2008. The focus on everyday life: a new turn in sociology. European Review 16 (1), 23–38.
Sztompka, Piotr, 2012. Visible meanings. In: Alexander, Jeffrey, et al. (Eds.), Iconic Power. Palgrave Macmillan, New York, pp. 233–245.
Tomlinson, John, 1999. Globalization and Culture. Polity Press, Cambridge.
Veblen, Thorstein, 1994 [1889]. The Theory of the Leisure Class. Penguin, London.
Wagner, Jon (Ed.), 1979. Images of Information: Still Photography in the Social Sciences. Sage, Beverly Hills.
Worth, Sol, Adair, John, 1972. Through Navajo Eyes. Indiana University Press, Bloomington.
Wright, Terence, 2008. Visual Impact. Berg, Oxford.

Visual Space, Geometry of

Josef Lukas, Martin-Luther Universitat, Halle-Wittenberg, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16274–16278, © 2001, Elsevier Ltd.

Abstract

Human orientation in space is based primarily on vision: we are able to judge the size of an object even from a distance, its form and shape, and its egocentric distance. The set of all objects in the visual field together with the perceived spatial relationships between them is called visual space. To some extent, our perceptions are veridical. There are, however, marked and stable deviations from the physical relationships. This observation raises two questions: (1) are the regularities of perceived spatial relationships characterized by some particular geometry? and (2) if so, what are the psychophysical functions relating this geometry to the well-known Euclidean geometry of physical space?

Human orientation in space is based primarily on vision: we are able to judge the size of an object even from a distance, its form and shape, and its egocentric distance. The set of all objects in the visual field together with the 'perceived' spatial relationships between them is called 'visual space'. To some extent, our perceptions are veridical. There are, however, marked and stable deviations from the physical relationships. This observation raises two questions: are the regularities of perceived spatial relationships characterized by some particular geometry? and if so, what are the psychological functions relating this geometry to the well-known Euclidean geometry of physical space? This article is only concerned with the first question. It has four main parts. After some introductory and historical remarks on first observations and perceptual phenomena leading to the notion of non-Euclidean geometry of visual space in the early 1900s (Section 'Is Visual Space Non-Euclidean?') a formal characterization of what is meant by a geometry is given in the Section 'Geometry as a Mathematical Language for Spatial Structure'. The Section 'Theories on the Geometry of Visual Space' presents a discussion of theoretical developments in the field of geometries for visual space, in particular the theory of Luneburg (1947). The Section 'Empirical Results on Geometrical Structures of Visual Space' reviews related empirical investigations and results.

Is Visual Space Non-Euclidean?

The question of whether visual space is Euclidean was not formulated explicitly until the beginning of the twentieth century. Answers to this question, however, can be traced back as far as the ancient Greek philosophers. Suppes (1977) presents a list of classical contributions to this problem, starting with Euclid's theory of perspective (rated as a 'Yes' answer). His list contains a majority of 'No' statements, some undecided cases rated as 'not sure', and only a few 'Yes' conclusions, mainly from authors using philosophical arguments rather than empirical results. Strong evidence against Euclidean geometry as an appropriate description of the structure of visual space arose from observations in so-called alley experiments. Investigating the phenomenon of size constancy, Hillebrand (1902, cited in Lukas, 1996) asked subjects to arrange small light points in the horizontal plane along two lines such that they should form an alley-like configuration, running from two fixed points toward the observer. He used two similar tasks but with different instructions. The first was to match the 'horizontal distances' of points to make sure that, perceptually, the width of the alley should be perceived as equal at any egocentric distance (*D*-alley).

The second was to pay attention to the alley's right and left border and make sure that they are perceived as 'parallel straight lines' (*P*-alley). To Hillebrand's surprise, the two instructions led to different configurations. Both alley types were curved, with *D*-alleys consistently lying 'outside' the corresponding *P*-alleys (see **Figure 1**). Blumenfeld (1913, cited in Lukas, 1996) confirmed this finding and was the first to mention a far-reaching consequence: since in Euclidean geometry lines with constant distance everywhere are called parallel (and vice versa), visual space has to be non-Euclidean. Blumenfeld's geometric reasoning elicited a vivid discussion, but for more than 30 years thereafter, it was not worked out mathematically in any substantive way. Luneburg (1947) was the first to present both a complete geometrical framework for visual space, and explicit psychophysical functions that related physical and visual coordinates in order to explain the striking results of the alley experiments and other empirical facts in binocular space perception known at that time.

Geometry as a Mathematical Language for Spatial Structure

Within mathematics there are at least two different approaches to geometry, the 'synthetic' and the 'analytic'. In the spirit of Euclid's axiomatic approach, synthetic geometry is defined over a set of points and a set of straight lines as primitives by specifying a number of relations considered to be true. Over many centuries Euclid's axioms have been considered to be the only reasonable description for spatial relationships. They include incidence axioms (for example, any two distinct lines have at most one common point), order axioms, axioms concerning motion and congruity, metric axioms, and, finally, the famous parallel axiom stating that for any point P and any line g not containing P there is one and only one line g' through P parallel to (not intersecting) g. At the beginning of the nineteenth century, however, Gauss, Bolyai, and Lobachevski

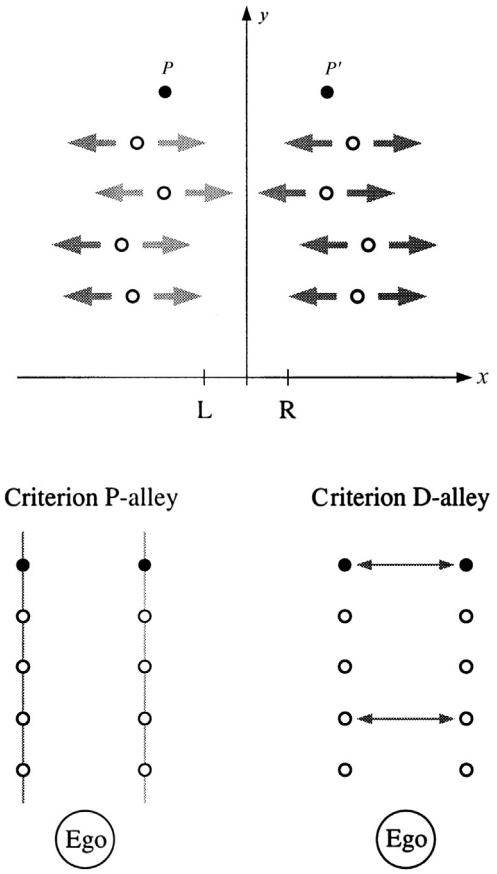

Figure 1 Classical setting for alley experiments. P and P' are fixed points at eye level, L and R denote the observer's left and right eye, respectively. The subject adjusts movable points according to one of the criteria shown below (left: parallel alley, right: distance alley).

discovered independently that the parallel axiom can be replaced by the postulate that there exist at least two parallels, yielding a different type of geometry (later called 'hyperbolic'). Moreover it has been shown that even a third version of the parallel axiom, stating that any two lines have at least one point in common, is compatible with the remaining Euclidean axioms, and leads to an 'elliptic' geometry.

The properties of both types of non-Euclidean geometry are demonstrated easily by models that fulfill the respective set of axioms. A well-known model for an elliptic plane is the surface of a sphere in Euclidean space $I\mathbb{R}^3$, where the set of elliptic lines is defined as the set of all (Euclidean) great circles. Similarly, models for hyperbolic geometry can be constructed on surfaces with constant negative curvature in $I\mathbb{R}^3$ (for example, saddle-shaped surfaces). Hyperbolic geometry is therefore said to have constant negative curvature, whereas elliptic geometry is characterized by its constant positive curvature.

Theories on the Geometry of Visual Space

Let V be the set of all points in the three-dimensional physical space which a fixed observer can see without moving the head. If x, y, z denote the Euclidean coordinates in physical space, then V is a subset of the positive half space $y > 0$ right in front of the observer. The origin of the coordinate system is interpreted as the egocenter of the observer and is assumed to lie in the closure of V, the x-axis denotes the direction from left to right, and the z-axis corresponds to the vertical axis. We restrict the following discussion to the horizontal plane $z = 0$ at eye level. The well-known Euclidean metric d_{phys} represents the physical distances in V.

In his famous attempt to represent the perceptual structure of binocular visual space geometrically, Luneburg (1947) starts out with the basic assumption that in addition to d_{phys} there exists some metric d_{vis} on V that represents the 'perceived' spatial structure. If γ and ϕ denote binocular parallax and bipolar latitude, respectively, his theory can be summarized by the following basic hypotheses:

LH1 There exists a Riemannian metric d_{vis} on the visual field V that represents perceived distances. The geodesics of d_{vis} are perceived as straight lines.

LH2 The visual space (V, d_{vis}) has constant Gaussian curvature K.

LH3 The perceived distance d_{vis} $(0, P)$ of point P to the observer is a function of the bipolar parallax only.

LH4 The ϕ-translations (γ, ϕ)–$(\gamma, \phi + \text{const})$ are visual isometries (leave all perceived distances unchanged).

LH5 The γ-translations (γ, ϕ)–$(\gamma + \text{const}, \phi)$ are angle-preserving.

Assumption LH1 provides the basis for the remaining assumptions: In a Riemannian metric space a coordinate system can be introduced that allows numerical representations for any geometrical objects such as lines, angles, triangles, or distances. LH2 restricts the discussion to geometries that are constantly curved everywhere. The remaining assumptions LH3 through LH5 define the psychophysical relationship between (physical) stimulus space and (perceptual) visual space. LH3 implies, for example, that points lying on a Vieth–Müller circle (VMC) (a circle containing the nodal points of the observers eyes) are perceived as equidistant to the observer because they have the same bipolar parallax γ. These latter assumptions are explained and discussed extensively in Binocular Space Perception Models.

Eschenburg (1980) gave a complete proof of Luneburg's conjecture that LH1 through LH5 uniquely determine the metric d_{vis} by two parameters K and σ up to equivalence. More specifically, the line element of the metric is represented in terms of the bipolar coordinates ϕ and γ by

$$\mathrm{d}s^2 = \frac{2\sigma^2(\mathrm{d}\gamma^2 + \mathrm{d}\phi^2)}{e^{\sigma\gamma} + Ke^{-\sigma\gamma}}$$

(Eschenburg, 1980: p. 7). K denotes the Gaussian curvature of visual space, and σ is an individual scaling factor.

Empirical Results on Geometrical Structures of Visual Space

Whereas LH3 through LH5 are closely connected to the psychophysical relationship between physical and visual coordinates, LH1 and LH2 define the inner geometry of visual space. Empirical investigations have concentrated on two

questions: (1) is K constant? (2) if so, is $K = 0$ indicating Euclidean geometry, or $K < 0$ (hyperbolic geometry) or $K > 0$ (elliptic space)? The early observations of distance alleys lying consistently outside the corresponding parallel alleys are only compatible with hyperbolic geometries (Indow and Watanabe, 1984). Direct numerical estimates for K can be derived from alley experiments or from so called 3- and 4-point experiments using double VMC suggested by Luneburg (1947) (see **Figure 2**). Using these methods K has mostly been found to be negative (Hardy, Blank, Boeder, Rand, and Rittler, 1953, cited in Lukas, 1996). However, numerical estimates of K cannot always reliably be reproduced in repeated measurements, estimates resulting from 3- and 4-point experiments are not valid predictors for data obtained from alley experiments, and experiments over large egocentric distance ranges coincide qualitatively but not quantitatively (Indow, 1991).

Moreover, for a test of LH2, numerical estimates of K are only of limited utility, because they presuppose the psychophysical mapping functions. Several methods have been suggested to test LH1 and LH2 independently from the remaining assumptions. In the triangle method used by Blank (1961) subjects had to bisect two sides of a given triangle *ABC*. Six of a total of seven subjects gave judgments indicating a negative curvature. However, the results seem to depend on the size of the triangle. Using the same method, Higashiyama (1981) found strong variations in curvature depending on lateral positions of the triangles and their distance from the observer. Battro et al. (1976) used stakes to produce experimental alleys of variable size in large open fields in natural environments. The curvature of visual space did not show a negative constant value. In addition, the results of experiments with equidistant lines and visual triangles under these viewing conditions also invalidated the constant curvature assumption.

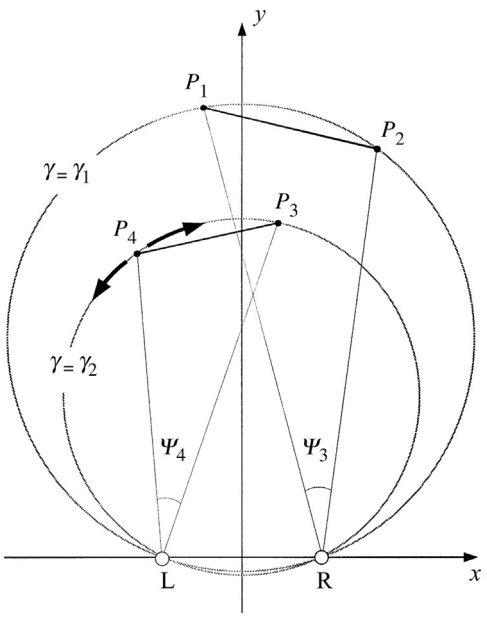

Figure 2 Experimental procedure in 4-point experiments. P_1, P_2, and P_3 are fixed on the Vieth–Müller circles $\gamma = \gamma_1$ and $\gamma = \gamma_2$, respectively. P_4 is adjusted by an observer such that P_4P_3 matches P_1P_2 perceptually.

Luneburg (1947) used quite a different argument and referred to the fact that Riemannian spaces with constant curvature are often characterized geometrically by a property called free mobility (any configuration can be moved to any position in space without changing the shape). Eschenburg (1980) showed that, contrary to Luneburg's argumentation, the existence of visual isometries alone does not guarantee free mobility in a strict sense. In Blank's (1958, cited in Blank, 1961) axiomatic version of Luneburg's theory, free mobility is guaranteed by an axiom that implies both the so-called Desarguesian property, and constant curvature in Riemannian spaces. Empirical consequences of this axiom have been tested successfully by Foley (1964, cited in Foley, 1972). However, a much stronger test in subsequent experiments failed (Foley, 1972). Lukas (1996) gave detailed arguments as to why the first test might have been too weak and the latter one unreasonably strong. From an empirical point of view Eschenburg (1980) provides the most powerful and sensitive test of the constant curvature hypothesis. He shows that an appropriate combination of the double VMC-experiments produces necessary and sufficient conditions for LH2. Experimental tests of these conditions reported in Lukas (1996) reflect a situation that seems to characterize most results in this field: the data are nicely fitted qualitatively by the predictions using constant curvature, yet there are numerically small, but statistically significant deviations.

Nevertheless, there are few models in the literature that try to generalize Luneburg's geometrical assumptions (contrary to his psychophysical assumptions, which have been relaxed in several attempts). Yamazaki (1987) developed a concrete non-Riemannian geometry and suggested its application to visual space. Because this model does not use the concept of length, its utility for empirical research seems limited.

A different approach is reported in Lukas (1996). Starting out with an axiomatic analysis of size constancy, he defines 'apparent distance' in terms of modern measurement (*see* Measurement, Representational Theory of) and shows that a very weak invariance condition resembling Luneburg's isei-conic transformations (defined in LH4 and LH5) forces the geometric structure of visual space to be Euclidean.

Conclusion and Future Directions

Critical reviews of empirical results seem to indicate that there is little hope of characterizing visual space globally by one particular geometry. Suppes (1995) concludes that visual space is context-dependent and no reasonable simple set of axioms in the spirit of synthetic geometry can be given for the structure of visual space. This view seems to be supported by recent results from Zimmer (1998). In a series of experiments, subjects failed in a test of very basic requirement of 'betweenness' indicating that not even simple order structures on a straight line can be assumed to hold a priori. Nowadays, geometric models in visual perception are actively discussed almost exclusively in the context of mathematical psychology and foundational aspects of science. The main stream of research in vision has obviously shifted to computational models (*see* Vision, Low Level Theory of) and the analysis of cognitive processes underlying human performance in spatial perception.

See also: Human Spatial Orientation, Neural Basis of; Perception of Action Space: Using Multiple Frames of Reference; Vision, High-Level Theory of; Vision, Low-Level Theory of.

Bibliography

Battro, A.M., di Pierro Netto, S., Rozestraten, R.J., 1976. Riemannian geometries of variable curvature in visual space: visual alleys, horopters, and triangles in big open fields. Perception 5, 9–23.

Blank, A.A., 1961. Curvature of binocular visual space. Journal of the Optical Society of America 51, 335–339.

Eschenburg, J.H., 1980. Is binocular visual space constantly curved? Journal of Mathematical Biology 9, 3–22.

Foley, J.M., 1972. The size-distance relation and intrinsic geometry of visual space: implications for processing. Vision Research 12, 323–332.

Higashiyama, A., 1981. Variation of curvature in binocular visual space estimated by the triangle method. Vision Research 21, 925–933.

Indow, T., 1991. A critical review of Luneburg's model with regard to global structure of visual space. Psychological Review 98, 430–453.

Indow, T., Watanabe, T., 1984. Parallel- and distance alleys with moving points in the horizontal plane. Perception and Psychophysics 35, 144–154.

Lukas, J., 1996. Psychophysik der Raumwahrnehmung. Psychologie Verlags Union, Weinheim, Germany.

Luneburg, R.K., 1947. Mathematical Analysis of Binocular Vision. Technical Report, Princeton, NJ.

Suppes, P., 1977. Is visual space Euclidean? Synthese 35, 397–421.

Suppes, P., 1995. Some foundational problems in the theory of visual space. In: Luce, R.D., D'Zmura, M., Hoffman, D.D., Iverson, G.J., Romney, A.K. (Eds.), Geometric Representations of Perceptual Phenomena: Papers in Honor of Tarow Indow on His 70th Birthday. Lawrence Erlbaum Associates, Mahwah, NJ.

Wagner, Mark, 2006. The Geometries of Visual Space. Lawrence Erlbaum, Mahwah, NI.

Yamazaki, T., 1987. Non-Riemannian approach to geometry of visual space: an application of affinely connected geometry to visual alleys and horopter. Journal of Mathematical Psychology 31, 270–298.

Zimmer, K., 1998. Experimentelle Untersuchungen zur Geo-metrischen Struktur des Binokularen Sehraums. Shaker, Aachen, Germany.

Visual Streams: Dorsal and Ventral

A David Milner, Durham University, Durham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The human brain contains two quasi-distinct visual systems operating in parallel, each originating in the primary visual cortex. The ventral (occipitotemporal) stream, codes information as a semantic representation suitable for storage and for deploying off-line in cognitive processes such as imagining, recognizing, and planning. It provides the visual contents of our perceptual experience. The dorsal (occipitoparietal) stream serves the more immediate function of guiding our movements visually from moment to moment, and so needs to code transitory 'here and now' information in an egocentric format. Its contents appear not to inform cognitive processing, and to be inaccessible to conscious monitoring.

Our Visual Inheritance

Evolution has provided primates with a patchwork of visual areas occupying almost 50% of the posterior half of the cerebral cortex. Despite the undoubted complexity of the interconnections among these different areas, two broad 'streams' of processing have been identified in the macaque monkey brain, each originating from the primary visual area (V1): a ventral stream projecting eventually to the inferior temporal cortex (ITC), and a dorsal stream projecting to the posterior parietal cortex (PPC). The dorsal stream also receives visual inputs via a number of subcortical visual structures, notably the superior colliculus in the midbrain and the pulvinar nucleus of the thalamus (Figure 1).

Functional Evidence: Primate Neurobehavioral Studies

In their seminal 1982 paper, which first explicitly identified the two streams of visual processing, Leslie Ungerleider and Mortimer Mishkin (1982) proposed that the two systems play complementary roles in the perception of incoming visual information. According to their account, the ventral stream governs the recognition of objects, whereas the dorsal stream mediates the localization of those same objects. Apparent support for this idea came from work with monkeys. Lesions of ITC produced deficits in the animal's ability to discriminate visually between objects, but had less effect on its performance on a task requiring it to register the location of a 'landmark' located between two alternative response sites. Conversely, lesions of the PPC produced deficits in performance on the landmark task, but did not affect object discrimination learning. Although the evidence available at the time was largely consistent with Ungerleider and Mishkin's (1982) 'what versus where' proposal, the results of subsequent research fit better with a distinction not between subdomains of perception, but between perception on the one hand and the guidance of action on the other (Jeannerod, 1997; Milner and Goodale, 2006). For example, subsequent research showed that monkeys with PPC lesions fail to register not only the *location* of a 'landmark' object but also its visual features as well, and show a loss of overt orienting movements toward the landmark. In other words PPC lesions cause a fundamental deficit of visual orienting, thereby limiting what items in the visual environment are attended to and processed. The deficit on the landmark task is primarily one of visually guided orienting movements, not one of perception – spatial or otherwise.

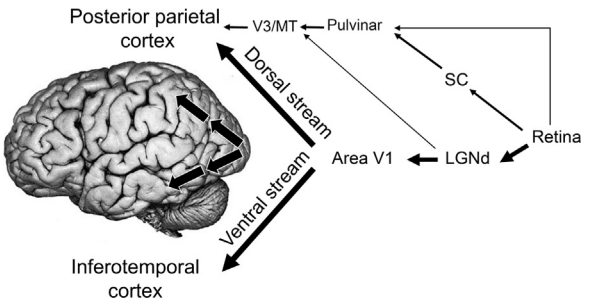

Figure 1 A schematic diagram of the two streams of visual processing in primate cerebral cortex. The ventral stream receives nearly all of its visual input from the primary visual cortex (V1), which in turn receives its input from the lateral geniculate nucleus (LGNd) of the thalamus. The dorsal stream also receives input from V1, but in addition gets substantial inputs from the superior colliculus (SC) via the pulvinar (Pulv), another nucleus in the thalamus. This pathway and others shown on the diagram remain intact in patients who have damage to V1, enabling certain unconscious visual abilities to survive such damage. The arrows on the inset photograph of the human brain show the approximate route of the two streams within the cerebral hemispheres.

Functional Evidence: Visual Neurons in the Ventral and Dorsal Streams

Neurons in ventral-stream areas culminating in ITC are tuned to the features of objects, and many of them show remarkable categorical specificity; for example, for faces or for bodily movements. Many of these cells maintain their selectivity irrespective of viewpoint, retinal image size, and lighting conditions. These ventral-stream neurons are not affected directly by the monkey's concurrent behavior, though many are modulated by how often the visual stimulus has been encountered, and by its association with reward. Thus the evidence strongly suggests that the ventral stream is more

concerned with the enduring characteristics and significance of objects than with moment-to-moment changes in the visual array.

Neurons in the dorsal stream show quite different properties from those in the ventral stream. Neurons in the PPC are activated by visual stimuli according to the response the monkey makes to those stimuli. For example, some cells respond when the stimulus is the target of an arm reach; others when it is the object of a grasp response; others when it is the target of a saccadic eye movement; others when the stimulus is moving and elicits a slow pursuit eye movement; and still others when the stimulus is stationary and becomes the object of fixed gaze (Jeannerod, 1997; Milner and Goodale, 2006). Other neurons, located in the same region as those associated with saccadic eye movements, respond in association with shifts of visual attention to fresh locations across the visual field.

These findings can readily explain why posterior parietal lesions impair landmark task performance: the monkey has lost the neural machinery for making head or eye movements toward the landmark, thereby failing to attend to it and learn its significance. They also explain the most immediately striking and long-known effects of PPC lesions – an inability to reach accurately for a moving or stationary food morsel, and a failure to shape and orient the hand and fingers appropriately to grasp or pick up the item. The evidence now shows clearly that visual networks in the dorsal stream compute more than just spatial location – they process the orientation and intrinsic geometry of visual stimuli as well, all in the service of efficient and well-coordinated action.

It should be noted that there are some visual areas that do not fit neatly into either stream. Although by some definitions the area known as the 'middle temporal' (MT) or 'visual area 5' (V5) lies in the dorsal stream, it behaves very much as if it lies in neither stream, or perhaps in both. Anatomically it is monosynaptically and reciprocally connected with the relatively unspecialized early visual areas V1, V2, and V3, and it sends projections directly to several areas in both dorsal and ventral streams. At first sight area MT looks like a ventral-stream area in being highly specialized for processing one particular visual property, namely motion, but research has shown that it provides this information both to dorsal stream areas mediating one's ocular and manual tracking of moving stimuli, and to ventral-stream areas concerned with putting the '4th dimension' into our conscious perception. This latter convergence in certain higher reaches of the ventral stream, notably within depths and banks of the superior temporal sulcus, allows static information (e.g., about objects, faces, and bodies) to be integrated with their dynamic properties, enabling us to recognize facial expressions and bodily actions.

Evidence from Neuroimaging Research

Functional neuroimaging of the visual brain in humans has advanced dramatically in recent years. Functional MRI (fMRI) in particular has provided extensive supporting evidence for the division of labor between the two visual streams that is described here. First, numerous studies beginning with Rafi Malach and his colleagues in 1995, have established a close association between ventral-stream activation and different aspects of visual perception (Grill-Spector and Malach, 2004). One area, for example, known as the lateral occipital (LO) area, is activated whenever the observer is looking at pictures of objects, and is localized by subtracting the brain activation pattern that results from viewing scrambled versions of the same objects. The net activations remaining after the subtraction indicate the regions concerned with shape perception in healthy individuals. Separate areas within the ventral stream are selectively activated by other visual categories (faces, hands, bodies, scenes), or by particular visual features (color, texture). Nancy Kanwisher and her colleagues, for example, discovered in the 1990s that faces preferentially activate an area in the ventral occipitotemporal region that has become known as the 'fusiform face area' (FFA). In other work, they found that places and buildings preferentially activate an area lying more medially, frequently named the 'parahippocampal place area' (PPA).

No area in the dorsal stream has been found to share this kind of close association with a dimension or category of visual perception. Instead, recent work has shown that areas within the dorsal stream are selectively activated when people actively respond to their visual environment in particular ways, whether by reaching out to targets and grasping objects or simply making rapid eye movements toward targets. Thus in a series of studies during the 2000s, Jody Culham and her colleagues have established that an area known as 'the human homologue of the monkey's anterior intraparietal area' (hAIP) is closely associated with the visual guidance of object grasping. At the same time they have also established that the 'superior parieto-occipital area' (SPOC) is linked with the reaching movements of the arm toward a visual target (Culham et al., 2006).

Visuomotor Deficits following Brain Damage in Humans

The first systematic description of a patient with bilateral PPC damage was published by Rudolph Bálint in 1909. Bálint's patient was unable to reach accurately toward objects with his right hand. In many respects this disorder, which Bálint called optic ataxia, resembles the visuomotor problems of the PPC-lesioned monkey. Humans with PPC damage not only fail to reach in the right direction, but also have difficulty orienting their wrist and forming their manual grasp appropriately to pick up particular individual objects (Jeannerod, 1997). This inability to use visual information to preshape the hand while reaching out means the patient often has to grope for an object. For example, they make errors in hand rotation as they try to reach toward and through a large oriented slot. Also, they often fail to show the normal unconscious opening of the hand in proportion to the size of different objects when reaching out to grasp them. Yet despite such visuomotor failures, these patients often have little difficulty in giving perceptual reports of the orientation and location of those very objects they fail to grasp. Parallel problems have been documented in controlling reaching and grasping movements so as to avoid potential obstacles lying close to the target object. Again this occurs without any concomitant

problem in judging where the potential obstacles are located. It thus seems that dorsal stream damage in humans, as in animals, selectively affects visually guided action, across both spatial and nonspatial domains.

Perceptual Deficits following Brain Damage in Humans

On the other side of the equation, damage to the ventral stream typically impairs aspects of visual perception. These range from difficulties in color or texture discrimination, to problems in visual recognition, in other words to different varieties of visual agnosia. In this context the condition known as visual form agnosia has been particularly well studied. These patients not only fail to recognize faces or objects, they fail even to distinguish simple geometric shapes. One such patient, DF, has been studied extensively over more than 20 years (Milner and Goodale, 2006; Goodale and Milner, 2013). The scientific value of these studies hinges on DF's uniquely fortuitous pattern of brain damage, which gives rise to severe visual form agnosia as an isolated symptom, in the absence of any motor impairment or notable cognitive loss. Most other similar patients in the literature either have a less severe impairment of shape discrimination, or suffer from cognitive or motor impairments.

What DF can do is as important as what she cannot. For example, her attempts to make a perceptual report of the orientation of an oriented slot show little relationship to its actual orientation, whether her reports are made verbally or by manual means. However, when she is asked to insert her hand or a hand-held card into the slot, she effortlessly moves her hand or the card toward the slot in the correct orientation, and inserts it smoothly. Indeed her hand begins to rotate in the appropriate direction as soon as it leaves the start position. In short, although she cannot report the orientation of the slot, she can act upon it with considerable skill.

We observed similar dissociations between perceptual report and visuomotor control when DF was asked to deal with properties of objects such as their size and shape (Goodale et al., 1991). Thus she showed excellent visually controlled anticipatory hand posture when reaching out to pick up blocks of different sizes that she could not distinguish perceptually. Just like healthy subjects, DF adjusted her finger–thumb separation well in advance of contacting the object, doing so in the normal linear fashion in relation to the target width. Yet when she was asked to use her finger and thumb to give a perceptual indication of the object's width, her responses were unrelated to the actual stimulus dimensions, and showed high variation from trial to trial.

Various studies of DF show that she is able to carry out many actions that are governed by visual information of which she has no perceptual awareness. But we have found that this is only true of actions that are directly targeted at the visual stimulus. She cannot successfully use the same visual information to guide a response that is superficially similar but is displaced to another location. Presumably the difference is that this kind of displaced response is arbitrary or symbolic in nature. DF seems to depend on a dedicated visuomotor processing system that will normally only come into play when she carries out natural goal-directed actions. She has no explicit visual awareness of the shapes and sizes that guide her responses when she is using this preserved visuomotor system.

There are temporal as well as spatial limits on DF's ability to drive her motor behavior visually. In one study, DF was shown a rectangular block, which was then removed while her eyes were closed, after which she reached out and pretended to grasp it. The preparatory grip size of normal subjects still correlates well with object width even after a 30 s delay. In DF, however, all evidence of grip scaling during her reaches had evaporated even after a delay of 2 s. The problem seems to be that this kind of delayed reach is again not a natural goal-directed movement, instead requiring subjects to recall the object they were shown earlier and then to 'pantomime' what it would be like to grasp it. DF simply does not have the wherewithal to implement such a strategy: to put it starkly, she cannot recall the appearance of something that she has not perceived in the first place. Apparently her preserved visuomotor brain can only respond to present or imminent states of the visual world, and routinely disregards past states, which typically will no longer be relevant to the task in hand.

Parenthetically, a complementary improvement in both reaching accuracy and grip scaling has been recorded in two patients with optic ataxia when making pantomimed responses after a 5 s delay, as opposed to acting immediately (Milner and Goodale, 2006). Presumably the delay forced these patients, who had sustained bilateral damage to the PPC, to use their (relatively) undamaged ventral system, effectively overriding the imperfect visuomotor control that would be provided by their damaged dorsal stream.

We have carried out several experiments using fMRI with DF herself in recent years. First we discovered using high-resolution structural MRI that the densest areas of damage in her brain are concentrated bilaterally in a region that corresponds well to the location of area LO in the normal brain (James et al., 2003). As mentioned earlier, area LO is activated in association with the perception of object shape in healthy individuals. Clearly therefore the loss of area LO would readily explain DF's visual form agnosia – effectively she would have lost the cerebral machinery for constructing percepts of whole objects from their elements. This interpretation was upheld when we carried out a fMRI experiment with DF herself, making the same contrast between viewing intact and scrambled objects that defines area LO in healthy subjects. The absence of any net activation in this experiment (James et al., 2003) established that her brain damage has effectively deprived her of the neural circuitry required to form percepts of intact objects: her brain shows no sign of being able to distinguish intact objects from scrambled ones. The data also allow us to frame an important converse argument. Normally, fMRI only allows us to make *correlations* between a particular pattern of brain activation and some particular psychological process. But if a patient who has evidently lost area LO bilaterally has also lost the ability to perceive object shape, then we can make a *causal* inference. We already knew that area LO is active when we perceive objects: now we can argue that LO forms an *essential* element in the healthy brain's circuitry for perceiving the shapes of objects. Its activity is not merely an epiphenomenon of shape perception.

We also tested whether DF's dorsal stream was activated when she reached out and grasped objects. We confirmed that area hAIP is indeed active when DF grasps objects under visual guidance, as is her area SPOC when she simply reaches out to touch objects (James et al., 2003). These further data provide strong confirmatory evidence for our hypothesis that the calibration of DF's many visually guided skills depends on the intactness of visuomotor control mechanisms within her dorsal stream.

Operation of the Two Streams in Normal Vision

It is most likely that healthy observers too, like DF, control their actions using visual information that is not present in their awareness. In healthy people of course, this visuomotor processing will be operating in parallel with intact perceptual processing within the ventral stream – with minimal interaction between the two. In fact the dorsal stream seems to act in relative isolation not only from the perceptual system, but also from the brain's acquired database about the visual world.

The implicit assumptions that underlie the brain's ability to confer stability and constancy on our perceptual world are based on such stored real-world knowledge. These assumptions sometimes cause us to misjudge what we see, as when a shift in a large frame makes a stationary spot inside it seem to shift in the opposite direction. In contrast, the dorsal system, by and large, is not deceived by such spatial illusions. Instead, it directs our saccadic eye movements and hand movements to where a target really is, instead of where our perceptual system tells us it is. Similarly, under appropriate circumstances geometric illusions can be seen to affect visually guided reaching and grasping far less – if at all – than they affect perceptual judgments (Goodale, 2008). In other words, you may perceive an object as bigger than it really is, but you open your finger–thumb grip correctly when reaching for it. In one study by Tzvi Ganel and his colleagues, the stimuli were adjusted such that one looked larger than the other, although actually physically smaller due to the influence of the Ebbinghaus illusion (**Figure 2**). The results were dramatic and thoroughly convincing: perceptual judgments went one way, while grasp calibration went the opposite way. No other model of visual function predicts, or can plausibly explain, such observations.

In another study, Richard Dyde and I (2003) compared two illusions of visual orientation (**Figure 3**). One, the simultaneous tilt illusion (STI), was used because it is believed to originate in early visual cortex, as a result of lateral interactions between columns of neurons sensitive to slightly different stimulus orientations. We argued that this illusion should therefore be 'passed on' to both visual streams, and so deceive action as much as perception. The other, the rod-and-frame illusion (RFI), was chosen because it is believed to be generated in the higher reaches of the perceptual system, and so, ex-hypothesi, anatomically well within the ventral stream. We reasoned that the orientation of responses directed at the central stimulus in the RFI should be immune to the illusion, despite there being a strong effect on perception. The results bore out these predictions. We found a similar strength of *perceptual* illusion in both STI and RFI, but as predicted, only those movements made toward the STI were deceived by the illusion, and indeed they were deceived to an extent very similar to the subjective STI illusion.

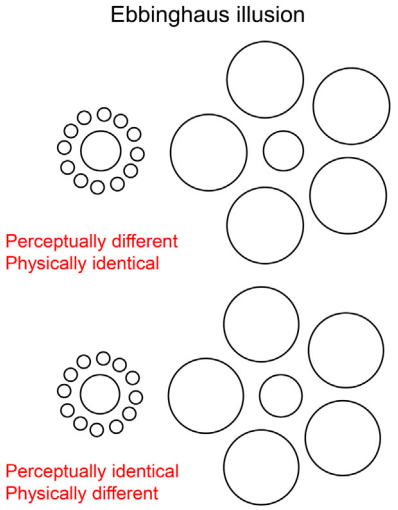

Figure 2 In the Ebbinghaus illusion, shown here, the two middle circles in the top two arrays appear to be different in size even though they are actually physically identical. The two middle circles in the bottom display appear to be identical but their real size is actually different. (To convince yourself of this, view each display through a piece of paper with two holes cut in it that reveal only the two central circles.)

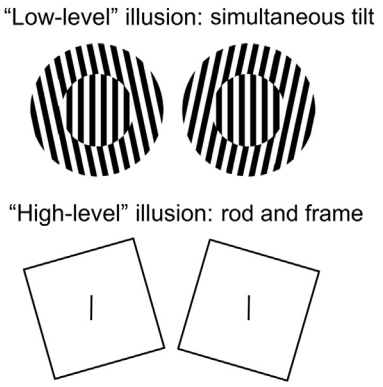

Figure 3 The 'simultaneous tilt illusion' (top) and the 'rod-and-frame illusion' (bottom). In both illusions, we see the stripes or central line to be tilted in opposite directions according the tilt of striped background (top) or the tilt of the frame (bottom). The reasons for this, however, are different for the two illusions. The simultaneous tilt illusion is probably the result of local inhibitory effects within the primary visual cortex, and therefore would be passed on to both visual streams. As a consequence this illusion affects both perception and action. The rod-and-frame illusion, however, depends on the same kinds of perceptual mechanisms as the pictorial illusions already mentioned, and like those, affects perception but not action.

It seems that the processing accomplished by the ventral stream both generates, and is in turn informed by, stored abstract visual knowledge about objects and their spatial relationships. It may be that these 'top-down' influences, and indeed the continual two-way traffic within the ventral stream, are what render our perceptual representations accessible to awareness (Lamme, 2006). This would fit neatly with the idea that coded

descriptions of enduring object properties ('invariants'), rather than transitory egocentric views, are needed for mental manipulations like the planning of action sequences and the mental rehearsal of alternative courses of action.

Perception and Action

The distinction between the visual roles of the dorsal and ventral streams has forced a careful consideration of what constitutes a simple visually guided act. As mentioned earlier, the dorsal stream is associated with immediate goal-directed movements, but is unable to guide movements based on visual information separated either spatially or temporally from the act itself. In addition, the dorsal stream's role seems to be limited to the guidance of movements that are well practiced. The use of visual illusions illustrates this well. Claudia Gonzalez and her colleagues (2008) showed that unfamiliar grasps made by the thumb and ring finger were subject to visual illusions like the Ebbinghaus, whereas natural grasps between forefinger and thumb were not (Goodale, 2008). She also found that right-handers using the left hand are also subject to visual illusions to which their preferred hand is immune. Recent evidence from Thomas Schenk suggests further that an action carried out 'in thin air', even when guided by a virtual image of a real object apparently located at the same place as the action, may require mediation by the ventral stream. At any rate, DF fails to show visual calibration of her grip size under these circumstances, when there is nothing tangible to grasp (despite appearances). Evidently the dorsal stream is not so easily fooled into treating a virtual image as real. It is deeply grounded in hard physical reality. In contrast of course, the ventral stream has no problem with virtual reality – as testified by our insatiable thirst for movies and television.

Integration of Visual Function

Although the dorsal and ventral streams have separate functions, there are multiple connections between them, and indeed adaptive goal-directed behavior in humans and other primates must depend on a successful integration of their complementary contributions. Thus the execution of a goal-directed action might depend on dedicated control systems in the dorsal stream, but the selection of appropriate goal objects and the action to be performed depends on the perceptual machinery of the ventral stream. Recent research suggests that direct connections from areas in the ventral stream to area hAIP may permit the dorsal stream mechanisms controlling grasping not only to calibrate grip size, but also to calibrate the grip force that needs to be applied by the fingers to grasp objects composed of different materials. Only the ventral stream seems to analyze the surface color and texture of objects, thereby enabling the brain to infer the 'stuff' that a given object is made of. A grip suitable for picking up a hardball is not going to work well with a ripe tomato or for lifting a ball of polystyrene.

Of course, the fact that ventral stream information may be useful to modulate dorsal stream processing by bringing visual 'semantics' into play, does not imply that it also brings conscious visual content with it. The evidence we have points to a rather strict segregation of conscious access, with only the ventral stream able to bring its products into visual phenomenology. fMRI has helped to confirm this segregation.

As mentioned earlier, the use of fMRI has delineated several specialized ventral-stream areas, two of which are the FFA, and the PPA, respectively activated by pictures of faces, and by pictures of buildings or scenes. Researchers have used this difference between the two areas to study the changes in visual consciousness during 'binocular rivalry.' They presented pictures of a face and a house simultaneously, one picture to each eye (**Figure 4**). Sometimes subjects would see the house and sometimes the face, but almost never both at once. They were asked to press a key whenever one percept was replaced by the other. Whenever they reported a switch to seeing the face, there was increased activity in their FFA – but when they reported seeing the house instead, there was more activity in their PPA. In other words, the activity in the FFA and the PPA reflected what they consciously perceived, not what was on their retina (which never changed).

Research like this demonstrates that areas in the ventral stream are associated with conscious visual processing, but in most cases the dorsal stream has not been tested. Yet pictures of objects can activate areas in the dorsal stream too – including area hAIP, where visual information about objects drives the formation of our grasping movements. Only recently have researchers asked the question whether levels of activation in these dorsal stream areas are correlated with awareness and unawareness just like activations in the ventral stream. Fang Fang and Sheng He (2005), instead of pitting two different patterns against each other, one shown to the left eye and the other to the right, created 'interocular suppression' by presenting high-contrast dynamic noise to just one eye. This visual noise prevented the observers from consciously seeing pictures of objects that were presented to the other eye. Yet these pictures still elicited substantial fMRI activation in the dorsal stream – and more importantly this activation was just as great as when there was no interocular suppression (i.e., the pictures were consciously perceived). As would be expected from previous work, however, Fang and He did find large differences in activation in the ventral-stream area LO between these 'aware' and 'unaware' testing conditions. So it seems that object-related areas in the dorsal stream respond to visual objects to the same extent, quite independently of whether the visual information is perceived consciously or not. That perceptual distinction seems to be entirely the province of the ventral stream, acting in parallel.

On the other hand, as was noted earlier, the lateral intraparietal area is not only concerned with guiding visually elicited saccadic eye movements, it also guides switches of our visual attention to new locations, even when our eyes remain stationary. Since this area is located in the dorsal stream, it has sometimes been argued that the dorsal stream must play a role in determining the contents of our visual consciousness. Indeed in experiments using binocular rivalry, fMRI activations in both the dorsal stream and associated areas in the frontal lobes are seen whenever there is a shift of perception from one of the two stimulus patterns to the other. But of course this activation does not correlate with the *direction* of the shift – it is just the same whether there is shift from percept A to percept B or from B to A. Dorsal stream activity determines our perceptual experience only indirectly – by switching activation

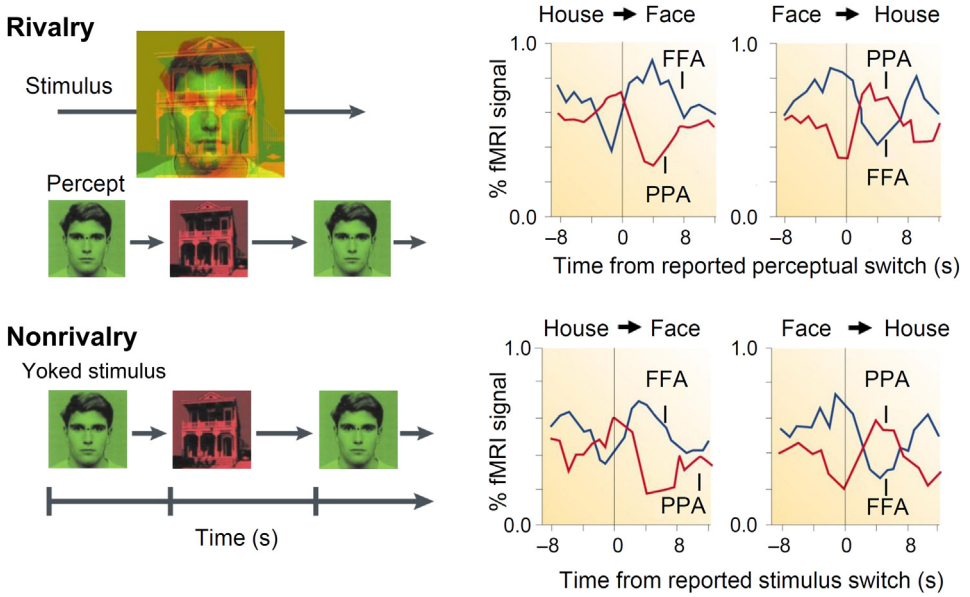

Figure 4 This figure shows the results of a fMRI study using binocular rivalry. Brain activation was measured in the FFA (selective for faces) and the PPA (selective for buildings) when observers viewed rivalrous images of a face and of a house (top left). The level of activation switched between the two brain areas in step with what the observers reported seeing (top right). The changes in activation in the two areas were just the same as those seen when the two pictures were presented successively (rather than simultaneously) in a manner mimicking the alternations of rivalry (bottom left and right). From Tong, F., Nakayama, K., Vaughan, J.T., Kanwisher, N., 1998. Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21, 753–759 (Figure 1).

between different perceptual representations within the ventral stream. It is uncorrelated with the current contents of our perception.

In summary, only the perceptual system in the ventral stream can provide the raw materials suitable for thought processes to act upon, and suitable for storage in memory. In contrast, the dorsal stream provides the raw materials for guiding action and switches of attention. It is designed to guide actions purely in the 'here and now,' and its products are discarded after use since they have to be constantly superseded by new ones.

See also: Occipital Lobe (Visual Cortex): Functional Aspects; Visual Perception, Neural Basis of.

Bibliography

Culham, J.C., Cavina-Pratesi, C., Singhal, A., 2006. The role of parietal cortex in visuomotor control: what have we learned from neuroimaging? Neuropsychologia 44, 2668–2684.

Fang, F., He, S., 2005. Cortical responses to invisible objects in the human dorsal and ventral pathways. Nature Neuroscience 8, 1380–1385.

Goodale, M.A., 2008. Action without perception in human vision. Cognitive Neuropsychology 25, 891–919.

Goodale, M.A., Milner, A.D., 2013. Sight Unseen: An Exploration of Conscious and Unconscious Vision, second edition. Oxford University Press, Oxford.

Goodale, M.A., Milner, A.D., Jakobson, L.S., Carey, D.P., 1991. A neurological dissociation between perceiving objects and grasping them. Nature 349, 154–156.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annual Review of Neuroscience 27, 649–677.

James, T.W., Culham, J., Humphrey, G.K., Milner, A.D., Goodale, M.A., 2003. Ventral occipital lesions impair object recognition but not object-directed grasping: a fMRI study. Brain 126, 2463–2475.

Jeannerod, M., 1997. The Cognitive Neuroscience of Action. Blackwell, Oxford.

Lamme, V.A.F., 2006. Towards a true neural stance on consciousness. Trends in Cognitive Sciences 10, 494–501.

Milner, A.D., Goodale, M.A., 2006. The Visual Brain in Action, second ed. Oxford University Press, Oxford.

Tong, F., Nakayama, K., Vaughan, J.T., Kanwisher, N., 1998. Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21, 753–759.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.

Visual Turn

Francesco Ventrella, University of Sussex, Brighton, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The visual turn encompasses a number of critical concerns in the humanities and social sciences, which explore visuality as both socially constructed experience, and as what makes social experience possible. The visual does not simply imply what can be seen by the eye as part of the phenomenal world, but it also refers to the social organization, partition, and administration of looking. As a challenge to the alleged transparency of images, the visual turn represents a critique of realism across media and disciplines.

Turning the Visual Turn

The visual turn encompasses a number of critical concerns in the humanities and social sciences, which explore visuality as both socially constructed experience, and as what makes social experience possible. The term visuality does not simply refer to a 'social practice of vision' but to a specific technique of power that can be visualized in history. How did visual culture convey representations of homosexuality before there was even a name for it? Why is the visual discourse of ethnicity often associated with attributes bearing a reference to authenticity, naturalness, and documentary realism? Such ways of interrogating the visual aim at reorganizing the relationship between what enters the field of vision and our abilities to interpret it. Seeing is not necessarily a pivoted and perspectival mode of apprehension, but an extended and expanded activity that turns images into signifying practices. Images present themselves as intermedial events that do not originate within, although they might not transcend from genres, traditions, canons, and conventions. As an independent and abstract term, the visual does not simply imply what can be seen by the eye as part of the phenomenal world, but it also refers to the social organization, partition, and administration of looking. Thus, the visual is also the medium that makes social experience possible by making visible social shifters such as gender, sex, class, ethnicity, race, age, ability, etc.

Within the field of vision, both material and immaterial representations can be approached in relation to regimes of visuality (what can be seen and how, under which circumstances), and in conjunction with the most disparate media of portrayal, depiction, symbolization, and simulation that make things visible (painting, speech, photography, clothing, cinema, dream narratives, writing, the Internet, etc.). The intellectual history of visuality is embedded in a number of disciplinary practices traditionally ranging from art history, philosophy, psychology, aesthetics, anthropology, and rhetorics. Its displaced genealogy is reflected in the effects that the visual turn has had across the disciplines. The lack of a common disciplinary object of visual culture has triggered a series of animated discussions among scholars. Yet, as W.J.T. Mitchell seems to suggest, the departure of a common object could be fortunate if it accompanies the emergence of a common objective to overcome "the veil of familiarity and self-evidence that surrounds the experience of seeing" (2002). The visual turn challenges the idea that images are transparent and obvious; the visual is turned every time seeing is displayed to make it available to analysis.

Genealogies of the Visual Turn

During the 1970s, an awareness about the intermediality of the visual initially triggered a shift away from research methodologies traditionally organized in monolithic academic encampments. Svetlana Alpers employed the term 'visual culture' with the aim to interrupt the seemingly natural link between works of visual art and their historical antecedents (1996). By this move, Alpers wanted to shift the priority from a history of art told by the works alone, on to a history of art which would intersect modes of visualization and apperception, and the way in which these too are inherently and structurally cultural phenomena. Visual texts should thus be interpreted in relation not only to the way in which one culture makes itself visible, but also to the way in which culture conceives of representation and representing as ways of knowing.

Such changes would be quickly registered in the emerging fields of film studies and cultural studies. The social history of art in Britain was particularly receptive to the role of ideology in the production and reception of artifacts, and started to trouble the distinction between high art and popular culture, thus hindering the conservative aesthetic canons of appreciation through a critical analysis of the politics of representation. Journals like *Screen* and *Block* record a number of pivotal contributions to the analysis of visual culture, often written with a pedagogical purpose to encourage critical teaching. Particularly important for these discussions was the theoretical input, deriving from "a combination of work on representation as a structure and process of ideology producing subject positions, and the social disciplining of technologies and regimes of power having real effects upon our lives and experiences" (Bird et al., 1996). The politicized voice of these journals, and the involvement with psychoanalysis, Marxist theory and semiotics was partly grown from a dialogue with other journals in continental Europe, notably France. The interdisciplinary journal *Tel Quel*, featuring contributions from theorists like Julia Kristeva and Hubert Damisch, will prove extremely influential for the development of the visual turn, on both sides of the Atlantic (Ffrench and Lack, 1998). Today, visual culture

is still committed to analyzing the commodification of the networks of production and exchange of visual culture, and attempts to avoid the reproduction of "value for transnational capital" (Mirzoeff, 2003).

The visual turn, and cognate phrases such as the 'pictorial turn' and the 'iconic turn', is not circumscribed to a specific academic discipline, but it envelops a wider array of interests that can be located in relatively recent transdisciplinary fields often referred to as visual studies, visual theory, and visual culture studies (Mitchell, 1994; Holly, 1996; Jay, 2002; Elkins, 2003; Moxey, 2008; Smith, 2008). There seems to be a general agreement that the visual turn has been somehow triggered by the ubiquitous presence of images, arguing that since "Western culture has come to be dominated by visual rather than oral or textual media" (Cartwright and Sturken, 2001), it has become possible to claim that "modern life takes place onscreen" (Mirzoeff, 1999). However, the spectacularization of life is not solely a product of contemporary visuality, but intersects with the historical layers of power relations that structure human imagination, and even fantasies.

Whither Historical Objectives?

Other scholars, particularly in North America, prefer to postpone the moment of the visual turn to the early 1990s (Mitchell, 1994; Jay, 2002), and identify a twofold emergence of visual studies in both the crisis of traditional disciplines and the spreading of new visual objects. Nicholas Mirzoeff (1999), in particular, has urged academics to acknowledge the need for the humanities to respond to a visual crisis arising from the excess of visual artifacts that were being made available by digital technologies. James Elkins (1999) too supports this idea by stressing not only the effects on 'visual literacy' deriving from the profusion of images, but also the changing quality of images we create and consume. The question must be asked whether the visual turn is a response to historical conditions that have changed visual experience, or it is a methodological revision of old issues. In fact, some have objected that the inundation of images in contemporary Western culture does not justify the need for a new methodology. Indeed, it is still important to take into account a genealogical approach to visual culture, one which challenges the presentism effected by the endless cycles of regeneration and renewal prompted by academic audit processes. Scholars who are skeptical of the historical abilities of visual studies have dismissed it as a mere disciplinary innovation that deals with popular illustrations, new media, artifacts, and design.

However, visual culture is not oblivious to the historical problems about vision. The very term 'visual turn' should be taken to task, for it invites us to explore the multiple directionality of turning around and turning to what calls for our attention in order to overturn what we perceive to be a problem in visual culture. Feminist theory has a crucial role in asking historical questions, which turn upside down the patriarchal separation between sexuality and regimes of visibility, particularly in the colonial/postcolonial arena. Mirzoeff acknowledged "the continuingly dynamic force of feminism (taken in the broad sense to incorporate gender and sexuality studies) to challenge disciplinarity of all kinds" (2003). The role of feminist visual theory is particularly significant to challenging the way in which sex, gender, race, ethnicity, and class are exchanged in culture as differences, which are historically constituted and institutionalized, especially in relation to a critique of the White privilege of the Western academic discourse (Pollock, 2003). In her account of the historical construction of oppositional Black spectatorship, bell hooks notoriously confronted the fact that many feminist film theorists were unable to see how race informed their looking relations. Hence, she discussed how the Black female spectator has to develop a position which is neither that of 'woman as image', nor that of 'man as bearer of the look' in order to create another space of identification, that is oppositional. "Looking at films with an oppositional gaze, black women were able to critically assess the cinema's construction of white womanhood as object of the phallocentric gaze and choose not to identify with either the victim or the perpetrator" (Hooks, 1992).

Irit Rogoff has argued that the political project of visual culture is deeply historical, for it revisions conventional narratives by showing that the difference between whom we see and whom we do not see is linked to the difference between 'aspects of the historical past', which have been given circulation, and some which have not (2003). Why do some images become intelligible, and why have others remained unintelligible for a long time? These questions do not enter the field of the humanities and social sciences with the visual turn but are already embedded in the agenda of critical theory. Walter Benjamin had already reflected on the readability of images in relation to historical materialism. He suggested that the readability of history is inscribed in its concrete, immanent, and material visibility, a form of 'graphicness' (*Anschaulichkeit*) that allows us to 'break with vulgar historical naturalism' (Benjamin, 1999). Benjamin's position lays bare a distinction between seeing and knowing that ought to be constantly interrogated. Within history writing, this entails a consideration of visual documents not simply as an illustration of a fact, but as a historiographical tool to mobilize the story told (or untold) by archival sources.

Interdisciplinarity, Transdisciplinarity, Indiscipline

The emergence of visual studies also responds to a cross-disciplinary definition of 'theory', not as a prescription of methods of interpretation but as a reflexive analysis of the discourses through which texts and meanings are interlinked (Culler, 1988). However, as the etymology seems to suggest (from the Greek *theōria*, meaning 'a sight', 'a spectacle', 'a form of viewing'), theory too is to be turned, for it is part and parcel of the structures by which seeing is regulated. What the English-speaking academia calls 'theory' in the globalized arena is not only intrinsically bound up with the supremacy of visuality in the Western epistemological tradition, as extensively discussed by Michel Foucault, but also inextricably intertwined with power relations, fantasies of mastery, surveillance, and containment that inform the very possibility to represent difference.

Thus, the visual turn presents itself as a metatheoretical question, and too often it has been dismissed as an exquisitely academic issue. Regular debates around the alleged usefulness of certain methodologies have tinted the visual turn with a polemical tone that responds to the perceived overlapping of

scholarly interests, particularly with regard to traditional academic disciplines such as art history, anthropology, literature, and sociology. Many scholars identify the visual turn with postdisciplinary academia, and this has triggered the dissemination of a number of different, if not contrasting definitions. According to Mieke Bal, the concept of transdisciplinarity presupposes an "immutable rigidity, a travelling without changing", while multidisciplinarity would be to "subject the fields of the two disciplines to a common analytic tool". She does not consider either of these options as viable ones and thus wishes for a 'travelling' theorization of concepts across the humanities and social sciences, one which is open to continuous negotiations, transformations, and reassessments (2002). Mitchell has put forward the idea that visual culture may even be defined as an 'indiscipline' (2002), for it develops a 'parasitical approach' that questions and critiques the material ways in which singular disciplinary subjects employ images and objects to create, contain, or enable different viewing positions.

Such suggestions have triggered some defensiveness on why and how academic disciplines might need new terms. In confronting directly these disciplinary anxieties voiced in the October questionnaire of 1996, Svetlana Alpers reminded that "Disciplinary boundaries, like differences between artistic mediums, are a subject of investigation, not of denial" (Alpers, 1996). This still seems a relevant advice to overcome the fear of theoretical catachresis, and it is also an invitation to foster a critical examination of the cultural layers that enfold the practice of looking within and outside academia. The need for a new theoretical posture proposed by the visual turn is not justified by a fancy for innovation, but derives from the institutional and historical awareness that "we cannot ask the new questions in the old language" (Rogoff, 2003).

Beyond the Word/Image Opposition

In an iconographical compendium of the seventeenth century, made for the use of artists and scholars, Cesare Ripa says that images are made to signify something different from what can be seen. This early modern claim can be taken to signal a semiotic attitude that, in part, may still be valid for those who analyze pictures like units of language comprising of signs, symbols, icons, and attributes whose connection produces meaning. The import of linguistic models to analyze visual objects is not a recent one. The genealogy of visual studies can be traced back to the rhetorical dispute about the supremacy of words over pictures. The ancient topic of ekphrastic literature, for instance, or the poetics of paragone have informed the very inception of iconology as a study of the image in modern times, and still represent a spur for scholarship in visual semiotics and intersemiotic translation (Mitchell, 1986; Bal and Bryson, 1991). Unlike linguistics concerned with the quantifying of the intersemiotic translation between words and images studied within a circumscribed context of signification (Groupe μ, 1992), visual studies pay particular attention also to the way in which visual signs may exceed systems of representation, moving across different sensoria, affects, materialities, and media (Schefer, 1995). Unlike visual hermeneutics or traditional iconology in which historical agents are represented as responding to specific cultural contexts, visual studies reintroduces the subjectivity of the analyst that the iconologist had evacuated, in order to show how meaning is rhetorically produced and reproduced through history, rather than frozen in a single moment of the past (Holly, 1996).

Visuality and Textuality

Following Jacques Derrida's dictum that "there is nothing outside the text", visual studies turn down the separation between objects of analysis (text/images) and contexts (social or historical background, biography, institutional spaces) sustained by historicist and psychologizing methods of interpretation that champion the concept of intentionality. Moreover, it has taken from semiotics the antirealist tenet that visual culture is made of signs that stand for something other than themselves and are always in relation to each other. This approach erodes the illustrative function of the image, thus liberating tasks and potentials that are charged with political significance. Consequentially, some have attacked visual studies for having undermined the historical function of images. Yet, it is precisely in the critique of the naturalized match between what can be said of what we see that the historical project of visual studies must be found.

Visual studies reflect on the problematic permutability between verbal and pictorial utterances. The interest in analyzing modes of presentation and representation is notably subsumed from structuralist linguistics. Early visual theorist such as Meyer Schapiro, Hubert Damisch, and Louis Marin have in fact looked at semiotics and linguistics to examine pictures as significant units structured like a language with a specific visual grammar and stressed the ability of images to perform acts of enunciation. Mieke Bal and Norman Bryson have argued that certain dematerialized linguistic categories, such as the pronoun 'I' and its coterminous 'you', but also adverbs like 'there' or 'today' have no referential value: they do not stand for a real person, object, place, or time external to the system of representation. Instead, their value is relational and internal to discourse itself (1991). These are 'deictic' terms that define the subject of discourse visually, while marking the possible not correspondence between speaker and subject. The insistence on the contiguity between 'I' and the 'eye' in a picture, for instance, is based only on the assumption that the subject may always have the power to be the one who looks. Here there is a question about the relationship between knowing and seeing that informs a number of poststructuralist methodologies, most notably feminist, women's, and gender studies, which critique the disparity between look and sight, i.e., looking and being looked at. For instance, in Édouard Manet's famous painting of *Olympia*, first exhibited in 1865, her look is actively fixed outside the painting, and this challenges the objectification of looking: the viewer is no longer the 'I' who can take possession of the woman, for Olympia's look positions him as a 'you' who is accountable for his act of looking. The notions of gaze, look, voyeurism, scopophilia, and fetishization devised by feminist visual theories in the 1970s have instigated a revision of the narrative strategies and technological supports that keep certain mystifications of gender in place within the field of art history, media, and film studies (Pollock, 2003). The possibility of mobilizing subject positions within a picture is precisely the signal that the codes

of visuality are not naturally fixed, but inherently constructed through gender.

Envisioning the Linguistic Turn

The intersection between semiotics, linguistics, and visual studies goes beyond a mere borrowing of theoretical jargon, for it is embedded in the very conceptualization of the turn. In fact, the phrase 'visual turn' has gained popularity by its assonance to other postmodern turns within the humanities and social sciences, most notably the 'linguistic turn'. The linguistic turn, which in the humanities and social sciences is often associated with a shift from structuralist to poststructuralist semiotics, bears the proposition that language constitutes reality and that the linguistic model ought to be the model for all communications. While metaphysical thinking takes language as a label that is tagged to concepts and objects, thinkers like Jacques Derrida and Michel Foucault borrowed from Ferdinand de Saussure's structuralist linguistics the idea that the relationship between signified and signifier is arbitrary. Thus, the linguistic turn is particularly interested in the metaanalysis of the linguistic tropes and narratives that convey truth and believability as such.

There are two key issues that arguably made the linguistic turn attractive for what here is being termed as the visual turn. The first has to do with the fact that, by shifting our attention from an abstract conception of text to a material source of textualities, the linguistic turn emphasized the unacknowledged role of writing as discourses in which the subject is inscribed. The emphasis on discourse is grounded on the idea that the speaker is not the subject of language, but the speaker exists as long as he or she is subjected to language. Thus, language and symbolic systems in general are not an expression of subjectivity, but rather they represent the agency that produces subjectivity by projecting around the speaker the space of the subject (Benveniste, 1971[1958]). Michel Foucault, who referred to this process as *assujettissement*, defined discourse as a body of knowledge that both defines and delimits what can be said about something. Discourses obviously change over time and are specific to certain social and historical contexts. As discourses regulate the administration of power and knowledge between the sayable and the seeable, there is no other way to visualize the subject but in intersection with discursive formations of identity and identification. Hence, visuality could be considered as the multifarious and fragmented discourse about vision, power, and knowledge.

Therefore, the second attractive aspect it follows is that the linguistic turn represents a break with the humanist conception of the subject, and challenged the idea that meaning should be traced back to the intentions of an agent or maker. Once it has been acknowledged that the meaning of a text cannot be tied up to a univocal source, but it has a life of its own, so to speak, it becomes easier to understand that the same might apply to images and visual culture that have in language their model of manifestation. For Derrida, the interpretation of a text or an image is enfolded within an irreducible and generative multiplicity of contexts and readings that he compared to 'dissemination'. Derrida proposes that meaning is a function of context, and that there are no limits or barriers to an endless sequence of recontextualizations. Hence, it follows that what dissemination displaces is the concept of an 'exhaustively determinable' context of meaning, that is a context originating in conscious intention (Derrida, 1982).

Frames and Contexts

The problematic of the context is a crucial one to establish a provisional distinction between the linguistic and the visual turn – at least, within the terms that have characterized certain academic debates. In fact, whereas scholars concerned with literary works have to join text and text, the visual turn has posed a new problem dealing with the translation between picture and commentary. Norman Bryson argues that within the historical disciplines, these elements traditionally stand to one another in a mimetic fashion, rather than in an analytical relation. The difference between text and context is conventionally established as "a separation that is at the same time an evident hierarchy, for the expectations that context will control the text" (Bryson, 1992). However, this separation also entails that, unlike the text, which still requires to be read textually, the context is often placed outside the text as an unquestionable propeller of meaning. Visual studies scholars have been extremely receptive to the claim, advanced within deconstruction philosophy and the linguistic turn, that the context does not partake with the purity of origins, nor is it a natural given, but it is something we make. If the context is produced by interpretative strategies, this means that contexts too are just as much in need of a critical analysis as texts are (Culler, 1988; Bal and Bryson, 1991; Bryson, 1992).

The context can be conceived visually, for it operates like a frame which, both and at the same time, encloses and displays its object. The frame is neither inside nor outside: it institutes the work. It is the frame, Derrida says, "gives raise to the work" (Derrida, 1987[1978]: 9). The frame oscillates between the two grounds that it constitutes – the work and the setting – and yet, with respect to each of these, it always dissolves into the other. The frame is all show and yet evades visibility; it is disavowed, yet already at work in fixing the look. One may think of a display within a museum. On the one hand, different framings encourage variable readings of the object, which are not necessarily related to its original function, and may even mystify its original meaning. On the other hand, a certain exhibition of the object can also expose ideas that were completely concealed in other forms of display or framing. This iterability of the sign illuminates the operations of dissemination in the field of cultural production.

Indeed, this is an issue that exceeds the museum walls, for the whole world is potentially contained into an exhibitionary paradigm under the powers of visual colonialism. Timothy Mitchell has convincingly examined the way in which anthropological exhibitions, a popular feature in world's fairs and international exhibitions, elicit the production of what he calls 'orientalist reality'. According to this principle, the Orient was defined as the product of unchanging racial and cultural essences, whose lack of progress was framed and exposed by the very exhibitionary paradigm that meticulously ordered races, costumes, and artifacts like specimens to put under scientific observation. However, the pretense objectivity flaunted by these systematic displays in fact betrayed a flawed image of the other as backward and primitive, thus providing

a visible and verifiable justification to the civilizing mission of Europe. As Timothy Mitchell puts it: "objectiveness was a matter not just of visual arrangement around a curious spectator, but of representation. What reduced the world to a system of objects was the way their careful organization enabled them to evoke some larger meaning, such as History or Empire or Progress" (1992).

The Biopolitics of the Visual

James Herbert stressed the tendency of visual studies to dematerialize what we see, in that "they shift the focus of analysis away from things seen toward the process of seeing" (Herbert, 2003). However, the alleged dematerialization in the methods of discourse analysis does not necessarily lead to the endorsement of a disembodied vision per se. In fact, the visual turn has also created a space to reflect on the effects of visual regimes upon the body. Take the representation of gender in general, and of femininity in particular. Griselda Pollock explains that, within the field of vision, women are looked at not because they are women: "I learn femininity – she says – as that repeated experience of being subjected to being looked at, appraised, accosted, commented upon and so on that will agonize my teenage years and cause older women to want their faces to be cut up in order to maintain the beautiful appearance" (2003). This first person account dramatizes what Teresa de Lauretis defines as "the power of discourses to 'do violence' to people, a violence which is material and physical, although produced by abstract and scientific discourses as well as discourses of the mass media" (quoted in Pollock, 2003). Culture inscribes on the body fixed meanings whose physical and material effects should be continually undermined if we want to establish new relations. Many feminist theorists turned to semiotics and psychoanalytic theory in order to question the association between the signifier 'woman' and concepts such as beauty, nature, and salability. Semiotics does not take the relationship between signifier and signified for granted, but takes that relationship precisely as an object of study.

The events of 11 September 2001 have represented a turning point for the discussion about the impact of institutional discourses upon corporeality. Mirzoeff argues that war is not only the subject of September 11, "but it is also a means of creating subjects" (2003). These events have reshaped the way in which we perceive the relationship between surveillance and safety, intimacy and publicness to a degree that has overstepped the mere domain of the visual and has invaded the whole sensory experience. In Foucault's terms, all bodies are shaped through biopower, that is institutional discourses which measure, define, identify, and categorize the body through political practices such as hygiene, birth control, health prevention, and beyond, including the broader fields of safety and security.

Our culture has become extremely saturated with images and devices that perpetuate the representation of violence and terror which trigger fantasies of previous horrors that have been traumatically interlocked with collective memory. Griselda Pollock and Max Silverman have borrowed from French poet and Buchenwald survivor David Rousset the phrase 'concentrationary universe' to define a system of total domination that annihilates the human in the concentration camp. Although condemnable, this system has somehow seeped into contemporary cultural imagination, particularly through forms of entertainment. Pollock and Silverman argue that the concentrationary is being replicated into novel and sometimes unexpected sites of visual experience: "via our spectacular identification as consumers of films, games and other cultural practices". These cultural objects often rely on processes of identification and misrecognition that "lure participant spectators into fantasies of mass destruction" (2013). The aestheticization of fascism in leather porn or video games like Manhunt 2 (2006), and the genetic paranoia conveyed in films like Children of Men (2006) represent only an unfinished archive of aesthetically different objects, whose politics of total domination continue to trigger a constant erosion of the human and its definitions.

Technological innovations constantly rewrite established visual theories, as in the case of the Foucauldian panopticon that has been reconceptualized across the intersection of sex and surveillance informing visualities on the Web. In reality shows, the live endears a compression of time and place that is reminiscent of classic scopic regimes like the peep show or the panopticon. However, Olivier Asselin has noted that the webcam instead subverts the parable of total surveillance epitomized by Foucault's theory of the panopticon, for this is not about one individual surveilling a group, but the other way around. Explicit representations of sex are rarely represented on live webcam shows; the webcam mainly broadcasts deserted interior spaces. Yet, sexuality is always there, as it is omnipresent, forever hoped for, forever suggested by the very architecturality of the medium (2008). Thus, the webcam can be considered as a pornographic medium not for what is represented, but for how representations operate by it.

Another important question deals with the way in which old visual technologies, like the cinema, deal with the representation of hitherto little seen sexual identities. In the past decade queer theory has been particularly receptive to the questions posed by the visual turn in order to overturn the relationship between identity and identification of non-heteronormative sexualities. Judith Halberstam has looked at the materiality of the cinematic medium to explain how the effacement of representations of the transgender body in the blockbuster movie *Boys Don't Cry* (1998) may be sustained by certain temporal and spatial axes embedded within the cinematic medium (what can be seen at what point in the plot of the movie). In this movie, the character of Brandon Teena passes as a boy, as long as no reference is shown to his past. Once transgenderism is exposed through the crude representation of a criminal episode of rural homophobia with lethal effects, Teena can only come back in the movie as a revenant that features in the guise of a ghostly voice from a police interrogation previously recorded. Halberstam writes: "When we 'see' the transgender character, then, we are actually seeing cinematic time's sleight hand. Visibility, under these circumstances, may be equated with jeopardy, danger and exposure, and it often becomes necessary for the transgender character to disappear in order to remain visible" (Halberstam, 2005).

This issue raises a question of how visual studies approach invisibility as a problem lying within, and not outside the field of representation. In recent phenomenological debates often

referred to as 'the affective turn', the body has become a pivotal standpoint from which to ground political and social questions. In this sense, the representation of corporeality and sensorial experience is not lost in discourse or language, but it is continually rematerialized in a number of contexts that demonstrate how much the invisible matters to the social production of fantasies, fears, and anxieties that have the body as the first object and objective. One should think of the cultural perceptions of HIV/AIDS which created a number of cosmetic visualizations within popular visual culture as a means to cope with the spread of fears about touch, contagion, and the skin; or also the spectacularization of sonograms in the public arena, which has material effects upon women's health in the context of abortion and self-determination (Cartwright and Sturken, 2001).

French philosopher Jacques Rancière has recently rebooted the political agenda of the visual turn by putting an emphasis on the way in which regimes of visuality intervene within what he calls the 'distribution of the sensible' (*partage du sensible*). Distribution here refers to both forms of inclusion and exclusion of what can be apprehended by the senses: what can be said of what we see, what can be heard of what is said, what is made visible by what we say, etc. Rancière comes to this proposition from an elaboration on the aesthetic regime of art and critiques the way in which a certain distribution of the sensible might endorse forms of consensus about, e.g., what can be said of what we see, rather than fostering creative forms of 'dissensus' that make democracy possible (2009). It is perhaps in this possibility for disagreement that the pedagogical position of visual studies may ultimately lie, in between self-reflexivity and self-critique, if they really are about enabling emancipation, rather than simply delivering another form of training.

Re-Turning the Visual Turn

Nicholas Mirzoeff has defined visual culture as a "tactic with which to study the genealogy and functions of today's life from the point of view of the consumer" (1999). Such statement does not necessarily bind visual culture to the analysis of contemporary design, video games, or advertisement, since also the past should be studied as a commodified projection, which comes in the shape of coffee table books, blockbuster exhibitions, and cruises to heritage sites. Yet, his statement could also be turned against the visual turn itself, for that too is imbricated in the profitable circulation of knowledge promoted within globalized academia. A possible way to interrogate this issue could emerge from a reflection on the generations and temporalities of the visual turn, that is an analysis of the rhetorical strategies and chronological narratives through which visual studies have been discussed and indeed presented as a turn.

A survey of the many voices that have partaken in such debates has shown that the history of the visual turn is often told through a narrative of progress, success, and overcoming. The old guardians of aesthetic value, experts, and connoisseurs, have been supplanted by analysts whose methodological and theoretical concerns are presented as more sophisticated, discursive, democratic, inclusive, etc. Yet, beside this account of emancipation lies a modernist fantasy of identity that eschews a confrontation with the effects of the postdisciplinary commodification of 'theory'. Shifts or turns not only highlight an epistemological direction, but also a temporal one. Clare Hemmings has argued that the narrative of progress is a textual technique describing a trajectory from sameness to difference, singularity to multiplicity, simplicity to complexity. Yet, progress narratives are also intertwined with narratives of loss: they claim that the past is over, and in one single move they want to make the new more attractive and, therefore, more salable. Of course, this is not specific to the visual turn (in fact, Hemmings here is talking about feminist theory), but this may initially help us to disentangle and break down its 'political grammar' (2011).

Take the impact of the visual turn on art history. It has been said that visual studies is not 'a discipline' that has stretched the art historians' interests in order to accommodate new media (photography, video, the press); neither is it a mere methodological updating. Visual studies rightly stress the fact that vision itself is already a form of critique that enables us to rewrite culture. This position has been welcomed among art historians interested in upsetting conservative aesthetic values by mingling disparate objects in order to reshuffle historiographical periodizations, which maintained low and high culture apart. Irit Rogoff has fostered the impact of visual culture upon art history for bringing forward a cultural awareness about vision that has made it possible to shift from the 'good eye' to 'the curious eye' so that issues like desire, corporeality, sexuality can now be treated historically (2003). This is a vital shift, and should not be undermined. Still, we owe to ask what happened to the old values, which used to be defined by the 'good eye', and whether they have been lost, or rather they have been transposed onto other spheres of public life. It is equally vital to ask whether the more attractive term 'free choice' may have become the new word for the outmoded normative concept of 'taste'.

Martin Jay urged that an indiscriminate leveling of visual culture could make us lose the "ability to make distinctions among different kinds of images and experiences". He also suggested, "traditional art history with its canonical restrictions still has a lot to teach us about the ways in which optical virtuosi, those with the gift of training to explore and extend the limits of the visual experience, transcend the conventions of their visual environment and open up new worlds for our eyes" (2002). In a similar vein, Thomas Crow already pointed out, "to surrender a history of art to a history of images will indeed mean a de-skilling of interpretation" for a whole generation of scholars (1996) and, it should be added, the expendable re-skilling of the younger one. With reference to this statement, Mitchell concluded, "the erosion of the forensic skills of connoisseurship and authentication among art historians in favour of a generalised iconological interpretative expertise is a trade-off that ought to trouble us" (Mitchell, 2002: 168).

Perhaps, it is due to that de-skilling that today we can see a whole new generation of scholars whose theoretical plasticity becomes all the more attractive the lesser it is resilient to the fluctuations and demands of the academic and publishing markets. All these considerations should be assessed in their own terms. No one is calling for a return to connoisseurship or elitist aesthetic values. Nonetheless, embracing new skills or

losing old ones appear as mutually exclusive options only until they remain fastened to a narrative of progress, that is also a story of consumption. These considerations, then, pose a question that becomes the more urgent the more we lose our ability to counter the argument that the value of the arts should be calculated only according to profitable terms. Visual culture appears to us only as the reified memorabilia that have often been buried without even entering the realm of political experience. Instead, one should not underestimate the powers of the outmoded in triggering unrealized theoretical imaginaries. Indeed, if we are to embrace a narrative of progress in the visual turn, one must also sustain a narrative of return to old questions. It would be very idle to be contented with fostering theoretical progress by simply disseminating methodological amnesia.

See also: Art History; Body, History of; Continental Philosophy of Language; Critical Theory; Deconstruction; Foucault, Michel (1926–84); Gender History; History: Theories and Methods; Media and History; Multiculturalism; Reference and Representation: Philosophical Aspects; Surveillance Studies.

Bibliography

Alpers, 1996. Visual culture questionnaire. October 77, 26.
Bal, M., 2002. Traveling Concepts in the Humanities: A Rough Guide. University of Toronto Press, Toronto.
Bal, M., Bryson, N., 1991. Semiotics and art history. The Art Bulletin 73 (2), 174–208.
Benjamin, W., 1999. The Arcades Project (H. Eiland, K. McLaughlin, Trans.). Harvard University Press, Cambridge, MA.
Benveniste, E., 1971[1958]. Subjectivity in language. In: Problems in General Linguistics. (M.E. Meek, Trans.). University of Miami Press, Miami.
Bird, J., et al., 1996. The Block Reader in Visual Culture. Routledge, London and New York.
Bryson, N., 1992. Art in context. In: Cohen, R. (Ed.), Studies in Historical Change. University of Virginia Press, Charlotesville, pp. 18–42.
Cartwright, L., Sturken, M., 2001. Practices of Looking: An Introduction to Visual Culture. Oxford University Press, Oxford and New York.
Crow, T., 1996. Visual culture questionnaire. October 77, 34–36.
Culler, J., 1988. Framing the Sign: Criticism and Its Institutions. Oklahoma, Norman.
Derrida, J., 1982. Signature, event, context. In: Margins of Philosophy. (A. Bass, Trans.). The University of Chicago Press, Chicago.
Derrida, J., 1987[1978]. The Truth in Painting (G. Bennington, I. McLeod, Trans.). University of Chicago, Chicago.
Elkins, J., 1999. The Domain of Images. Cornell University Press, Ithaca.
Elkins, J., 2003. Visual Studies: A Skeptical Introduction. Routledge, London and New York.
Ffrench, P., Lack, R.F., 1998. The Tel Quel Reader. Routledge, London and New York.
Groupe μ, 1992. Traité du signe visuel: Pour une rhétorique de l'image. Seuil, Paris.
Halberstam, J., 2005. In a Queer Time and Place: Transgender Bodies. Subcultural Lives. New York University Press, New York and London.
Herbert, J., 2003. Visual culture/visual studies. In: Nelson, R.S., Shiff, R. (Eds.), Critical Terms for Art History. The University of Chicago Press, Chicago, pp. 452–464.
Hooks, b, 1992. Black Looks: Race and Representation. South End Press, Boston.
Holly, M.A., 1996. Past Looking: Historical Imagination and the Rhetoric of the Image. Cornell University Press, Ithaca and London.
Jay, M., 1996. Visual culture and its vicissitudes. October 77, 42–44.
Jay, M., 2002. Cultural relativism and the visual turn. Journal of Visual Culture 1, 267–278.
Mirzoeff, N., 1999. Introduction to Visual Culture. Routledge, London and New York.
Mirzoeff, N., 2003. The Visual Culture Reader, second and third ed. Routledge, London and New York.
Mitchell, T., 1992. Orientalism and the exhibitionary order. In: Dirks, N. (Ed.), Colonialism and Culture. University of Michigan Press, Ann Arbor, pp. 289–318.
Mitchell, W.J.T., 1986. Iconology: Image, Text, Ideology. The University of Chicago Press, Chicago and London.
Mitchell, W.J.T., 1994. Picture Theory: Essays on Verbal and Visual Representation. The University of Chicago Press, Chicago and London.
Mitchell, W.J.T., 2002. Showing seeing: a critique of visual culture. Journal of Visual Culture 1, 167–181.
Moxey, K., 2008. Visual studies and the iconic turn. Journal of Visual Culture 7, 131–146.
Pollock, G., 2003. The visual. In: Eagleton, M. (Ed.), A Concise Companion to Feminist Theory. Blackwell Publishing, Oxford, pp. 173–194.
Pollock, G., Silverman, M., 2013. Concentrationary Memories: Totalitarian Terror and Cultural Resistance. I.B. Tauris, London.
Rancière, J., 2009. Aesthetics and Its Discontents. Polity, Cambridge.
Rogoff, I., 2003. Studying visual culture. In: Mirzoeff, N. (Ed.), The Visual Culture Reader. Routledge, London and New York, pp. 14–26.
Schefer, J.L., 1995. The Enigmatic Body. Cambridge University Press, Cambridge.
Smith, M., 2008. Visual Culture Studies: Interviews with Key Thinkers. Sage Publications, London.

Visual Word Recognition, Neurocognitive Psychology of

Arthur M Jacobs, Dahlem Institute for Neuroimaging of Emotion (D.I.N.E.), Freie Universität Berlin, Berlin, Germany
Johannes C Ziegler, Brain and Language Research Institute, Aix-Marseille Université and Centre National de la Recherche Scientifique, Marseille, France

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.A. Moreno, G.C. Van Orden, volume 24, pp. 16555–16561, © 2001, Elsevier Ltd.

Abstract

Together with eye movement control, visual word recognition is the fundamental process underlying reading performance, but while the former is natural and basically innate, word recognition is highly artificial and usually takes years to learn. Neurocognitive psychologist use a variety of methods and models to investigate how humans (learn to) perform the multiple subprocesses underlying word recognition, for example, discriminating and recognizing the written equivalents of elementary speech sounds (phonemes), or automatically constructing the meaning(s) of a word. Modern neuroimaging research suggests that word recognition is a nonlinear dynamic, embodied process temporally distributed over several hundred milliseconds and neuroanatomically distributed across a range of cortical and subcortical networks, questioning the classical notion of a passive retrieval of word knowledge from a hypothetical mental lexicon. In a world full of illiterates and dyslexics, the biggest challenge for neurocognitive psychology is to provide the applied disciplines with the necessary and sufficient knowledge making it possible to design efficient and reliable methods for diagnosis, teaching, training, and rehabilitation of visual word recognition and reading.

Reading is one of the greatest achievements of both human civilization and the human mind. It is based on two fundamental processes: word recognition and eye movement control. While eye movement control is a natural, basically innate process – we move our eyes naturally about three to four times per second across sections of the environment to bring objects into focus – visual word recognition is not natural, but a highly artificial activity of the mind that comes at a cost: for most people it takes years to learn it. Unlike native speech and auditory word recognition which are acquired more naturally, children have to learn visual word recognition as much as adults who study a new language, and possibly, even a new script.

The neurocognitive psychology of visual word recognition investigates the mental and neuronal processes that allow readers to map printed symbols onto sounds and meanings and those that allow readers to become amazingly skilled in this highly complex activity. It uses a variety of neurocognitive methods (e.g., psychophysics, mental chronometry, brain imaging, and stimulation techniques) and models (e.g., boxological, mathematical, and computational; for a classification see Jacobs and Grainger, 1994) to explain how humans (learn to) perform the fundamental subprocesses underlying word recognition: (1) discriminate and recognize the written equivalents of elementary speech sounds (phonemes), that is, letters/graphemes (in alphabetic scripts) from the environment, for example, a text; (2) transform information about those symbols (e.g., phoneme/letter identities and their relative position) into information about sound and meaning.

Major Issues

Most of the neurocognitive research on visual word recognition studies the question how humans can gain rapid and automatic access to the meanings of single words. Other major issues concern: (1) how beginning readers learn and apply the alphabetic principle in perception and action; that is, the mapping of letters (or letter groups) onto elementary speech sounds, which allows them to 'decode' the thousands of spoken words that are in their mental lexicon prior to reading; (2) the functional units of learning/processing written words (e.g., letters/graphemes, phonemes, syllables, or morphemes) and their neural correlates; (3) the question as to which role orthographic, phonological, and morphological processes play in learning to recognize and process visual words and the temporospatial dynamics of these processes in the brain; (4) the neural and behavioral causes and consequences of individual differences and deficits in the acquisition of literacy; and (5) the development of intervention methods to detect and remedy such deficits.

Learning Visual Words

Learning to read starts later than learning to speak, for most people around the age of 5 or 6 years. During several years thereafter children must acquire the code used by their culture for representing speech (sounds) as a series of visual symbols, that is, mapping letters (orthography) onto phonemes (phonology) and vice versa. In most languages, the relationship between symbol and sound is systematic, whereas the relationship between symbol and meaning is largely arbitrary. According to one influential model (Frith, 1986), learning to read words is not gradual, but progresses in steps. When the letters of the alphabet have been learned, in step 1 the child applies a logographic strategy; that is, instant word recognition on the basis of salient graphic features, such as the oval around the sign ESSO. In step 2, the alphabetic strategy is applied, that is, a strictly sequential blending of sounds to evoke a word, such as cuh-a-tuh which English-speaking children often use to

read CAT. In step 3, finally, the orthographic strategy is applied, that is, instant word recognition on the basis of orthographic, syllabic, or morphemic cues.

Another important theory of reading development assumes that children rely on phonological decoding skills to learn novel words (Share, 1995). According to this theory, every successfully decoded word provides the child with an opportunity to set up direct connections between a given letter string (orthography) and the spoken word, which results in the development of an orthographic lexicon. Phonological decoding thus provides a powerful self-teaching device because the explicit learning of a small set of spelling–sound correspondences allows the child to decode an increasingly large number of words, which bootstraps orthographic and lexical development (for an implementation of this theory, see Ziegler et al., 2014).

Learning to read words correctly is made difficult by a number of learning/processing problems that differ across alphabetic languages (Ziegler and Goswami, 2005). The *availability problem* reflects the fact that not all phonological units are consciously (explicitly) accessible prior to reading. Thus, connecting letters to phonemes that are not yet readily available requires further cognitive development. The *consistency problem* reflects the fact that identical visual units can represent different sounds and vice versa, as illustrated in the famous "Beware of HEARD a dreadful WORD, it looks like BEARD and sounds like BIRD … ." Finally, the *granularity problem* reflects the fact that there are many more orthographic units to learn when access to the phonological system is based on bigger grain sizes as opposed to smaller grain sizes, like in English. That is, there are more words than there are syllables, more syllables than there are rimes, more rimes than there are graphemes, and more graphemes than there are letters. A related problem is that higher-level representations are more accessible than lower-level ones. Thus, English-speaking kindergarteners have better access to the number of syllables in a word than to the number of phonemes in a syllable. They also have better access to rimes than to phonemes. A further problem is the dominance of meaning over form. An example is the observation that common semantic features of two words (e.g., the fact that TIGER and CLOWN are associated with a circus can mask phonological differences between them (e.g., that TIGER and CLOWN have different onsets). An example which should be familiar to all adult readers is that one has better mnestic access to semantic features than to form features (e.g., upper- vs lowercase) of words (cf Jacobs, 2001).

As concerns the problem of gaining access to meaning of written words, the debate revolves around two alternatives: (1) skilled readers access meaning through phonological recoding of written words (i.e., the translation of visual units into sound codes in reading) (Van Orden, 1987) or (2) they access meaning directly from orthography without phonological mediation (Coltheart et al., 1993). Current computational theories agree that both pathways are used during skilled reading (Coltheart et al., 2001; Perry et al., 2007) and the weight given to either of them might depend on factors such as the orthographic transparency of the writing system or the proficiency of the reader (see Section Meaning Construction in Context). A related open question is whether orthographic and phonological processes are interactive or independent, and whether they develop interactively or sequentially. Many studies suggest that written words automatically activate their sound representations and there is increasing evidence for strong bidirectional interactions between orthographic and phonological processes even in skilled adult readers (Ziegler et al., 1997).

Which visual units are functional in learning and processing words? The majority of studies suggest that reading is letter based (e.g., Nazir et al., 1998), but intermediate units, such as phonemes and graphemes (Rey et al., 1998, 2000), vowel–letter clusters (Chetail and Content, 2013), syllables, and bigrams (Conrad et al., 2009, 2010) or morphemes (Rastle et al., 2004) have been found to play a role in going from letter representations to whole words. Proficient readers thus seem to make use of any kind of helpful information that efficiently and reliably leads to the disambiguation of a word in a given context (see Section Processing Visual Words).

How is the recent cultural invention of 'words' (<6000 years) constructed in the human brain? Since evolution hardly had time to develop inbuilt mechanisms for word recognition and reading, learning to read must involve a preemption of parts of the existing visual system and its conversion, by minimal modification, to process the shapes of letters and words. Indeed, in most expert readers, a localized region of the left occipito-temporal sulcus, just lateral to the fusiform gyrus, systematically takes on the function of identifying visual letter strings. This 'visual word form area' is thought to play a pivotal role in informing other temporal, parietal, and frontal areas of the identity of the letter string, for the purpose of both semantic access and phonological retrieval (Dehaene et al., 2005).

Processing Visual Words

For many decades the prevailing theory on word recognition was that word representations are stored in a mental lexicon which is a part of long-term memory and from which the grammar processor retrieves them to understand phrases and sentences. More recent approaches favor an alternative view; according to which, words are objects, that is, sensory stimuli that act directly on mental states, meaning, they are active operators rather than passive operands, and rather than having a fixed meaning, they are 'cues to meaning,' which means, their meaning lies in the context-dependent causal effects they have on mental states (Elman, 2004). The view that words are objects, and thus are coded via their sensory–motor features, that is, primarily the sound–image and articulatory program, goes back to early neurologists/aphasiologists like Freud (1891) and language theoreticians like Bühler (1934), who also anticipated recent 'embodied cognition' theories of word recognition, according to which the meaning of words is grounded in the sensory–motor activities of the objects/actions/events they refer to (see Section Meaning Construction in Context).

Modern neuroimaging techniques, such as magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) provide interesting insights into how the brain processes written words, showing that this process is far from being as instantaneous as the subjective intuition of an adult reader – who can read 250 words per minute without being

aware of the fact that fully recognizing each word might take from one-fourth to half a second – might suggest. Even highly automatized word recognition in skilled readers is thus a process distributed over time with a variable time course depending on many context factors, such as word familiarity, orthographic, phonological, morpho-syntactic or semantic complexity, or reading skill. When a printed or written object (i.e., a word) is projected onto the retina, areas in both occipital lobes (e.g., V1 to V4, occipito-temporal sulcus) start to extract visual features like strokes, angles, curves, or contours from the object and synthesize them into letter shapes. Already after one-tenth of a second, the above-mentioned 'visual word form area' in the fusiform gyrus of the left hemisphere has helped to construct a preconscious word representation from these features and letters. For this neuronal representation to become consciously available, it needs to be coupled with the corresponding sound representation, taking about another one-tenth of a second of processing, mainly in areas specialized for speech in the left temporal lobe. About another one-tenth of a second is then taken up by activity in neural networks of the left frontal cortex related to constructing the context-dependent meaning of the word. All these activities largely develop not only in parallel and in a distributed fashion involving large parts of the left cortex, but also as bilateral subcortical structures of the limbic system, or the cerebellum, as well as some right-hemispheric areas in order to construct the full meaning of a word, including simple affective features, such as 'pleasant' or 'disgusting,' morpho-syntactic features like case role, or semantic features like 'human' (Citron, 2012; Conrad et al., 2011; Hofmann et al., 2009, 2011; Kuchinke et al., 2005; Ponz et al., 2013; Recio et al., 2014; Vosse and Kempen, 2009).

Meaning Construction in Context

Studies on single word recognition, that is, processing words in isolation, can only reveal limited insights on how the ultimate goal of reading is achieved: making meaning out of a series of written symbols. The widely spread idea that readers decode words by decomposing them into semantic 'atoms,' thus leading to conscious access to their meaning independent of context, is not compatible with recent neurocognitive evidence. Many results suggest that the metaphor of a mental lexicon containing fixed literal meanings of words which are passively 'accessed' is a fallacy. Rather, during reading, the brain constructs several possible meanings of a given word (including figurative ones) in a context-sensitive, 'on-line' way – some aspects of meaning often being processed separately from others in different parts of the brain – as the result of nonlinear, dynamical interactions between bottom-up and top-down processes (Bazan, 2011; Elman, 2004; Forgacs et al., 2012; Jacobs, 2011, in press; McClelland et al., 1989; Schrott and Jacobs, 2011). These processes are much more complex than popular models of 'cold' information processing in reading (Just and Carpenter, 1980), word recognition (Coltheart et al., 2001), sentence comprehension (McClelland et al., 1989), or text processing (Kintsch, 1988) have proposed: it involves 'embodied' processing of sensory–motor and 'hot,' that is, affective or aesthetic, aspects and meaning construction processes such as the activation of context-sensitive frames/schemas, conceptual blending of information from different domains of reference or subject variables, such as ideas of reference or odd beliefs (Aryani et al., 2013; Bohrn et al., 2013; Briesemeister et al., 2011, 2014; Coulson, 2006; Jacobs, 2011, 2014; in press; Juhasz et al., 2011; Kutas, 2006; Nazir et al., 2008; Schrott and Jacobs, 2011; Siakaluk et al., 2008; Vo et al., 2006; Zwaan and Taylor, 2006).

Methods and Tasks

A wealth of (neuro-)cognitive methods (e.g., rating scales, mental chronometry, oculo- and pupillometry, electroencephalography (EEG), fMRI, transcranial magnetic stimulation (TMS), and near-infrared spectroscopy (NIRS)) allows to study the multiple processes involved in visual word recognition at different levels of processing (e.g., sensory–motor, mnestic, and affective) using a variety of dependent variables (e.g., subjective–experiential, neuronal, and behavioral) in a multitude of tasks (e.g., perceptual identification, silent reading, naming, lexical decision, and semantic decision). Research using these methods and tasks has identified a multitude of factors that influence visual word recognition, such as word frequency/familiarity, word length, lexicality, neighborhood density and frequency, imageability, concreteness, regularity/consistency, age of acquisition, emotional valence, figurativeness, or semantic cohesion (e.g., Balota et al., 2004). Some of these effects appear to be ubiquitous and independent of the task or the context, such as the word frequency effect; that is, the observation that familiar words are recognized faster than unfamiliar ones; others, such as the effects of emotional valence, appear to be more sensitive to context, task, or subject characteristics (e.g., Veltkamp et al., 2012).

In order to allow a synthesis of all the different effects observed in a multitude of tasks, it is necessary to develop a functional overlap modeling strategy identifying (with the help of computational modeling) the task- and context-independent core processes underlying visual word recognition (Grainger and Jacobs, 1996; Jacobs and Grainger, 1994). This also involves greater efforts at quantifying the relevant variables among the mass of factors mentioned above, for example, by replacing the subjective rating procedures used to 'operationalize' certain word features, such as imageability or emotional valence, by computational models allowing to quantify them in an objective manner (Westbury et al., 2014). This approach is valuable because it solves a long-standing methodological and epistemological problem of studies on word recognition and empirical psychology in general: the theoretically unwarranted use of dependent variables obtained in one study (i.e., ratings of imageability or valence) as independent variables in the factorial or other experimental design of another study. A final ambitious goal is to connect the cognitive and computational models in a meaningful way with both behavioral and neuroimaging data in order to distinguish between processes that are localized to specific brain structures (e.g., sensory–motor processing) and processes where specialization arises in the distributed pattern of activation over many different areas, each participating in

multiple functions (Hofmann and Jacobs, in press; Price, 2012; Taylor et al., 2012).

Models

As a rule of (neuro-)cognitive psychology, a given neurocognitive test should be accompanied by a falsifiable neurocognitive (process or task) model that makes explicit the processes and structures underlying performance in the test and allows to assess its validity, as well as qualitative and quantitative predictions about performance. Contemporary cognitive and computational models of visual word recognition can be distinguished according to about a dozen criteria (Jacobs and Grainger, 1994), such as pre-quantitative versus quantitative or learning versus performance models.

First, there are a number of abstract mathematical models that simulate word recognition as an optimal decision process that accumulates noisy evidence over time. While successfully capturing some aspects of the empirical data obtained in standard word recognition experiments (e.g., frequency effect in a lexical decision task), they are badly underdetermined with regard to the nature of the information driving optimal decisions (but see Dufau et al., 2012). These models thus simply fit the human data, and although they do so very accurately, they 'post-dict' existing results rather than predict novel results.

In contrast, algorithmically concrete models of the connectionist family, such as interactive activation models (e.g., Grainger and Jacobs, 1996; McClelland et al., 1989) are concerned with how specific word features activate specific orthographic, phonological, or semantic units through processes of interaction and competition (Hofmann et al., 2011). These models can not only simulate and predict a variety of standard behavioral word recognition effects, but also be meaningfully connected with neuroimaging data (Braun et al., 2006; Hofmann and Jacobs, in press).

Similarly, dual route models of reading aloud (e.g., Coltheart et al., 2001) also make the hypothetical mental processes that underlie word recognition transparent. They assume two pathways for going from print to sound; a 'direct' orthographic route that maps letters onto whole word units in an orthographic and phonological lexicon using interactive activation, and a sublexical route that maps graphemes onto phonemes. These routes operate at the same time and the results of their computations are integrated at the phoneme output level. While the sublexical route in earlier dual-route models (e.g., Coltheart et al., 2001) was serial, rule-based, and hardwired, recent dual-route models, namely the connectionist dual process model or CDP+, have replaced the rule-based sublexical route with a two-layer associative network that learns the mappings between graphemes and phonemes (Perry et al., 2007). This model has also been extended to process multi-syllabic words (Perry et al., 2010), which requires syllabic segmentation and sublexical stress assignment (see also Conrad et al., 2010) and to simulate reading development and developmental dyslexia (Ziegler et al., 2014). A notable advantage of this modeling approach is that it is already successfully used in guiding applications, in particular the diagnosis and treatment of dyslexia (Engl et al., 2013; Ziegler et al., 2008).

Applications

Basic research on visual word recognition and reading development has been tremendously important for our understanding of how reading should be taught (for review, see Rayner et al., 2001) and how reading disorders, such as acquired or developmental dyslexia, should be diagnosed and treated (e.g., Engl et al., 2013; Heine et al., 2012; Thaler et al., 2009). With regard to the former, it is today well established that (1) mastering the alphabetic principle (that written symbols are associated with phonemes) is essential to becoming a proficient reader in all languages (Ziegler et al., 2010) and (2) methods that teach this principle directly are more effective than those that do not (especially for children who are at risk for having difficulty learning to read). Thus, elementary school teachers who make the alphabetic principle explicit are most effective in helping their students become skilled, independent readers.

With regard to the diagnosis and treatment of developmental dyslexia, theories of reading development and skilled reading have become increasingly important in our understanding of how deficits in various subcomponents of the reading network can cause the kind of reading problems seen in children with dyslexia (Ziegler et al., 2008). It has become clear that the heterogeneity that is seen in the disorder can be directly related to the complexity of the reading task, which requires among others good visual and attentional skills, good phonological skills, and good spoken language skills (e.g., vocabulary). Although deficits in any of these basic components can cause dyslexia, it is important to note that phonological deficits tend to outweigh visual and/or orthographic deficits both in terms of the magnitude of the deficit and the number of affected children (Ziegler et al., 2008).

On the basis of research on skilled word recognition and reading development, a variety of reading interventions have been proposed that typically attempt to improve processing at one or several of the above-described levels. For example, computer programs that explicitly train the mapping between orthographic units and their corresponding sounds have been shown to be of help for training children with dyslexia in decoding skills(Engl et al., 2013; Kyle et al., 2013). Programs that train phonological awareness skills prior to reading or during the initial stages of reading have been shown to significantly improve reading (Ehri et al., 2001). Extra-wide letter spacing of text has been shown to improve reading accuracy and speed in children with dyslexia presumably by reducing the deleterious effects of crowding (Zorzi et al., 2012). A technique based on a rapidly moving window that unmasks the to-be-read text as children read along has also been shown to improve reading speed for children with dyslexia (Breznitz et al., 2013). This list is not exhaustive but shows the impact of fundamental research on word recognition for a better understanding and treatment of dyslexia.

Challenges

In a world full of illiterates and dyslexics, the biggest challenge for neurocognitive psychology surely remains to provide the

applied disciplines with the necessary and sufficient fundamental knowledge making it possible to design efficient and reliable methods for diagnosis, teaching, training, and rehabilitation of visual word recognition and reading. Given the enormous, heterogeneous set of models and methods in the neurocognitive literature, another challenge is theoretical and methodological unification. For example, fundamental research on eye movement control or speech perception still remains relatively disconnected from research on visual word recognition; research on implicit learning and working memory is still widely separated from research on reading. Also, more research on standards for model building and evaluation, as well as for evaluating diagnostic and teaching methods is needed, if we want to fully explain the conditions favoring optimal acquisition and mastering of reading. Finally, so far, too little effort has been spent on developing a more ecologically valid model of *literary* reading that not only takes into account the wealth of results that have been accumulated by studies on single word recognition and sentence comprehension – the vast majority of which used only very simple (i.e., monosyllabic and monomorphemic) words or simple, descriptive, nonliterary texts – but also considers the complex dynamics of emotional, immersive, reflective, or aesthetic processes that accompany reading of literary texts such as poems, short stories, or novels (Iser, 1976). A recent neurocognitive poetics model of literary reading that integrates elements of rhetoric, aesthetics, and cognitive poetics with neurolinguistic and psychonarratological concepts (Jacobs, 2011, 2014) makes first steps in this direction. First empirical support for this descriptive comprehensive model comes from studies on sentence and text processing, as well as poetry reception (Altmann et al., 2012; Bohrn et al., 2012, 2013; Hsu et al., in revision; Lehne et al., in revision; Lüdtke et al., 2014).

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Dyslexias, Acquired and Agraphia; Eye Movements in Reading; Language Acquisition; Language Processing, Functional Magnetic Resonance Imaging of; Neural Basis of Reading; Psycholinguistics: Overview; Spelling, Psychology of.

Bibliography

Altmann, U., Bohrn, I.C., Lubrich, O., Menninghaus, W., Jacobs, A.M., 2012. Fact versus fiction – how paratextual information shapes our reading processes. Social Cognitive and Affective Neuroscience, 09/12.

Aryani, A., Jacobs, A.M., Conrad, M., 2013. Extracting salient sublexical units from written texts: 'Emophon', a corpus-based approach to phonological iconicity. Frontiers in Psychology Language Science 4, 654. http://dx.doi.org/10.3389/fpsyg.2013.00654.

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., Yap, M., 2004. Visual word recognition of single-syllable words. Journal of Experimental Psychology: General 133 (2), 283–316.

Bazan, A., 2011. Phantoms in the voice. A neuropsychoanalytic hypothesis on the structure of the uncon- scious. Neuropsychoanalysis 13, 161–176.

Bohrn, I.C., Altmann, U., Lubrich, O., Menninghaus, W., Jacobs, A.M., 2012. Old proverbs in new skins – an FMRI study on defamiliarization. Frontiers in Psychology 3, 204. http://dx.doi.org/10.3389/fpsyg.2012.00204.

Bohrn, I.C., Altmann, U., Lubrich, O., Menninghaus, W., Jacobs, A.M., 2013. When we like what we know – a parametric fMRI analysis of beauty and familiarity. Brain and Language 124, 1–8. http://dx.doi.org/10.1016/j.bandl.2012.10.003.

Braun, M., Jacobs, A.M., Hahne, H., Ricker, B., Hofmann, M., Hutzler, F., 2006. Model-generated lexical activity predicts graded ERP amplitudes in lexical decision. Brain Research 1073–1074, 431–439.

Breznitz, Z., Shaul, S., Horowitz-Kraus, T., Sela, I., Nevat, M., Karni, A., 2013. Enhanced reading by training with imposed time constraint in typical and dyslexic adults. Nature Communications 4, 1486.

Briesemeister, B.B., Kuchinke, L., Jacobs, A.M., 2011. Discrete emotion effects on lexical decision response times. PLoS ONE 6 (8), e23743. http://dx.doi.org/10.1371/journal.pone.0023743.

Briesemeister, B.B., Kuchinke, L., Jacobs, A.M., 2014. Emotion word recognition: discrete information effects first, continuous later? Brain Research 1564, 62–71.

Bühler, K., 1934. Sprachtheorie (Theory of Language). G. Fischer, Stuttgart.

Chetail, F., Content, A., 2013. What is the difference between OASIS and OPERA? Roughly five pixels: orthographic structure biases the perceived length of letter strings. Psychological Science. http://dx.doi.org/10.1177/0956797613500508.

Citron, F.M.M., 2012. Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. Brain and Language 122, 211–226. http://dx.doi.org/10.1016/j.bandl.2011.12.007.

Coltheart, M., Curtis, B., Atkins, P., Haller, M., 1993. Models of reading aloud: dual-route and parallel-distributed-processing approaches. Psychological Review 100, 589–608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., Ziegler, J.C., 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. Psychological Review 108, 204–256.

Conrad, M., Carreiras, M., Tamm, S., Jacobs, A.M., 2009. Syllables and bigrams: orthograpic redundancy and syllabic units affect visual word recognition at different processing levels. Journal of Experimental Psychology: Human Perception and Performance 35, 461–479.

Conrad, M., Recio, G., Jacobs, A.M., December 2011. The time course of emotion effects in first and second language processing: across cultural ERP study with German–Spanish bilinguals. Frontiers in Psychology – Language Science 2 (351), 1–16.

Conrad, M., Tamm, S., Carreiras, M., Jacobs, A.M., 2010. Simulating syllable frequency effects within an interactive activation framework. European Journal of Cognitive Psychology 22, 861–893.

Coulson, S., 2006. Constructing meaning. Metaphor and Symbol 21, 245–266.

Dehaene, S., Cohen, L., Sigman, M., Vinckier, F., 2005. The neural code for written words: a proposal. Trends in Cognitive Sciences 9, 335–341.

Dufau, S., Grainger, J., Ziegler, J.C., 2012. How to say "no" to a nonword: a leaky competing accumulator model of lexical decision. Journal of Experimental Psychology: Learning, Memory, and Cognition 38 (4), 1117–1128. http://dx.doi.org/10.1037/a0026948.

Ehri, L.C., Nunes, S.R., Willows, D.M., Schuster, B.V., Yaghoub-Zadeh, Z., Shanahan, T., 2001. Phonemic awareness instruction helps children learn to read: evidence from the National Reading Panel's meta-analysis. Reading Research Quarterly 36 (3), 250–287.

Elman, J.L., 2004. An alternative view of the mental lexicon. Trends in Cognitive Sciences 8, 301–306.

Engl, V., Thaler, V., Heine, A., Jacobs, A.M., 2013. Individuelle Förderung von Kindern mit Lese- und/oder Rechtschreibauffälligkeiten: Das Guckomobil-Förderprogramm (Individual Training of dyslexic children: The Guckomobile Training Program). BMBF Band 38, Berlin.

Forgacs, B., Bohrn, I.C., Baudewig, J., Hofmann, M.J., Pleh, C., Jacobs, A.M., 2012. Neural correlates of combinatorial semantic processing of literal and figurative noun-noun compound words. NeuroImage 63, 1432–1442.

Freud, S., 1891. Zur Auffassung der Aphasien (On aphasia). Deuticke, Wien.

Frith, U., 1986. A developmental framework for developmental dyslexia. Annals of Dyslexia 36, 69–81.

Grainger, J., Jacobs, A.M., 1996. Orthographic processing in visual word recognition: a multiple read-out model. Psychological Review 103, 518–565.

Heine, A., Engl, V., Thaler, V., Fussenegger, B., Jacobs, A.M., 2012. Neuropsychologie von Entwicklungsstörungen schulischer Fertigkeiten (Neuropsychology of developmental deficits in skills underlying school achievement). Hogrefe, Göttingen.

Hofmann, M.J., Jacobs, A.M., 2014. Interactive activation and competition models and semantic context: from behavioral to brain data. Neuroscience and Biobehavioral Reviews, in press.

Hofmann, M.J., Kuchinke, L., Tamm, S., Vo, M.L., Jacobs, A.M., 2009. Affective processing within 1/10th of a second: high arousal is necessary for early facilitative processing of negative but not positive words. Cognitive Affective Behavioral Neuroscience 9, 389–397.

Hofmann, M.J., Kuchinke, L., Biemann, C., Tamm, S., Jacobs, A.M., 2011. Remembering words in context as predicted by an associative read-out model. Frontiers in Psychology Language Science 2 (252), 1–11.

Hsu, C.T., Jacobs, A.M., Conrad, M. Can Harry Potter still put a spell on us in a second language? an fMRI study on reading emotion-laden literature in late bilinguals, in revision.

Iser, W., 1976. Der Akt des Lesens: Theorie ästhetischer Wirkung (The Act of Reading: Theory of Aesthetic Impact). Fink Verlag, München.

Jacobs, A.M., 2001. The cognitive psychology of literacy. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behavioral Sciences. Elsevier, Amsterdam, pp. 8971–8975.

Jacobs, A.M., 2011. Neurokognitive Poetik: Elemente eines Modells des literarischen Lesens (Neurocognitive Poetics: Elements of a Model of Literary Reading). In: Schrott, R., Jacobs, A.M. (Eds.), Gehirn und Gedicht: Wie wir unsere Wirklichkeiten konstruieren. Hanser, München, pp. 492–520 (Brain and Poem: How We Construct Our Realities.

Jacobs, A.M., 2014. Towards a neurocognitive poetics model of literary reading. In: Willems, R. (Ed.), Towards a Cognitive Neuroscience of Natural Language Use. Cambridge University Press.

Jacobs, A.M. Metaphern beim Lesen in Gehirn und Geist (Metaphors during Reading in the Mind/Brain). Psychosozial, in press.

Jacobs, A.M., Grainger, J., 1994. Models of visual word recognition: sampling the state of the art. Journal of Experimental Psychology: Human Perception and Performance 20 (6), 1311–1334.

Juhasz, B.J., Yap, M.J., Dicke, J., Taylor, S.C., Gullick, M.M., 2011. Tangible words are recognized faster: the grounding of meaning in sensory and perceptual systems. The Quarterly Journal of Experimental Psychology 64 (9), 1683–1691.

Just, M.A., Carpenter, P.A., 1980. A theory of reading: from eye fixation to comprehension. Psychological Review 87, 329–354.

Kintsch, W., 1988. The use of knowledge in discourse processing: a construction-integration model. Psychological Review 95, 163–182.

Kuchinke, L., Jacobs, A.M., Grubich, C., Võ, M.L.-H., Conrad, M., Herrmann, M., 2005. Incidental effects of emotional valence in single word processing: an fMRI study. NeuroImage 28, 1022–1032.

Kutas, M., 2006. One lesson learned: frame language processing—literal and figurative—as a human brain function. Metaphor and Symbol 21, 285–325.

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., Goswami, U., 2013. Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the United Kingdom: GG Rime and GG Phoneme. Reading Research Quarterly 49, 59–74.

Lehne, M., Jacobs, A.M., Koelsch, S. On First Looking into Hoffmann's "The Sandman": Neural Correlates of Suspense during the First Reading of a Literary Text, in revision.

Lüdtke, J., Meyer-Sickendiek, B., Jacobs, A.M., 2014. Immersing in the stillness of an early morning: testing the mood empathy hypothesis in poems. Psychology of Aesthetics, Creativity, and the Arts 8, 363–377.

McClelland, J.L., St John, M., Taraban, R., 1989. Sentence comprehension: a parallel distributed processing approach. Language and Cognitive Processes 4, 287–336.

Nazir, T.A., Jacobs, A.M., O'Regan, J.K., 1998. Letter legibility and visual word recognition. Memory and Cognition 26, 810–821.

Nazir, T.A., Jeannerod, M., Hauk, O., 2008. The role of sensory-motor systems for language understanding. Foreword. Journal of Physiology Paris 102, 1–3.

Perry, C., Ziegler, J.C., Zorzi, M., 2007. Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. Psychological Review 114, 273–315.

Perry, C., Ziegler, J.C., Zorzi, M., 2010. Beyond single syllables: large-scale modeling of reading aloud with the connectionist dual process (CDP++) model. Cognitive Psychology 61 (2), 106–151. http://dx.doi.org/10.1111/cogs.12030.

Ponz, A., Montant, M., Liegeois-Chauvel, C., Silva, C., Braun, M., Jacobs, A.M., Ziegler, J.C., 2013. Emotion processing in words: a test of the neural re-use hypothesis using surface and intracranial EEG. Social Cognitive and Affective Neuroscience, 03/2013.

Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage 62, 816–847.

Rastle, K., Davis, M.H., New, B., 2004. The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. Psychonomic Bulletin and Review 11, 1090–1098.

Rayner, K., Foorman, B.R., Perfetti, C.A., Pesetsky, D., Seidenberg, M.S., 2001. How psychological science informs the teaching of reading. Psychological Science in the Public Interest 2, 31–74.

Recio, G., Conrad, M., Hansen, L.B., Jacobs, A.M., 2014. On pleasure and thrill: the interplay between arousal and valence during visual word recognition. Brain & Language 134, 34–43.

Rey, A., Jacobs, A.M., Schmidt-Weigand, F., Ziegler, J.C., 1998. A phoneme effect in visual word recognition. Cognition 68, 71–80.

Rey, A., Ziegler, J.C., Jacobs, A.M., 2000. Graphemes are perceptual reading units. Cognition 75, B1–B12.

Schrott, R., Jacobs, A.M., 2011. Gehirn und Gedicht: Wie wir unsere Wirklichkeiten konstruieren (Brain and Poem: How We Construct Our Realities). Hanser, München.

Share, D.L., 1995. Phonological recoding and self-teaching: Sine qua non of reading acquisition. Cognition 55, 151–218.

Siakaluk, P.D., Pexman, P.M., Aguilera, L., Owen, W.J., Sears, C.R., 2008. Evidence for the activation of sensorimotor information during visual word recognition: the body–object interaction effect. Cognition 106, 433–443.

Taylor, J.S.H., Rastle, K., Davis, M.H., 2012. Can cognitive models explain brain activation during word and Pseudoword reading? a meta-analysis of 36 neuroimaging studies. Psychological Bulletin 45, 0–26.

Thaler, V., Urton, K., Heine, Hawelka, S., Engl, V., Jacobs, A.M., 2009. Different behavioral and eye movement patterns of dyslexic readers with and without attentional deficits during single word Reading. Neuropsychologia 47, 2436–2445.

Van Orden, G.C., 1987. A ROWS is a ROSE: spelling, sound, and reading. Memory and Cognition 15, 181–198.

Veltkamp, M., Recio, G., Jacobs, A.M., Conrad, M., 2012. Is personality modulated by language? International Journal of Bilingualism. http://dx.doi.org/10.1177/1367006912438894, 04/2012.

Vo, M.L.-H., Jacobs, A.M., Conrad, M., 2006. Crossvalidating the Berlin affective word list (BAWL). Behavior Research Methods 38, 606–609.

Vosse, T., Kempen, G., 2009. The Unification space implemented as a localist neural net: predictions and error-tolerance in a constraint-based parser. Cognitive Neurodynamics 3, 331–346. http://dx.doi.org/10.1007/s11571-009-9094-0.

Westbury, C., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B.B., Hofmann, M.J., Jacobs, A.M., 2014. Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. Frontiers in Psychology 4, 991.

Ziegler, J.C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., 2010. Orthographic depth and its impact on universal predictors of reading: a cross-language investigation. Psychological Science 21 (4), 551–559. http://dx.doi.org/10.1177/0956797610363406.

Ziegler, J.C., Castel, C., Pech-Georgel, C., George, F., Alario, F.X., Perry, C., 2008. Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. Cognition 107, 151–178. http://dx.doi.org/10.1016/j.cognition.2007.09.004.

Ziegler, J.C., Goswami, U., 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. Psychological Bulletin 131, 3–29.

Ziegler, J.C., Perry, C., Zorzi, M., 2014. Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. Philosophical Transactions of the Royal Society B. http://dx.doi.org/10.1098/rstb.2012.0397.

Ziegler, J., Van Orden, G.C., Jacobs, A.M., 1997. Phonology can help or hurt the perception of print. Journal of Experimental Psychology: Human Perception and Performance 23, 845–860.

Zorzi, M., Barbiero, C., Facoetti, A., Lonciari, I., Carrozzi, M., Montico, M., 2012. Extra-large letter spacing improves reading in dyslexia. Proceedings of the National Academy of Sciences 109 (28), 11455–11459. http://dx.doi.org/10.1073/pnas.1205566109.

Zwaan, R.A., Taylor, L.J., 2006. Seeing, acting, understanding: motor resonance in language comprehension. Journal of Experimental Psychology: General 135, 1–11.

Visualization in Science and Technology

Michael Lynch, Cornell University, Ithaca, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Starting in the 1970s, historians and ethnographers of scientific practice began to treat images, inscriptions, and the technologies for producing them as centrally important aspects of science, rather than as secondary products of theoretical and experimental work. Visualization in science includes arrays of practices with instruments and tools that enable researchers to make phenomena visible and publicly available. Visualization is not limited to vision, nor is it limited to sensory perception, as it includes drawing, graphic recording, and other means of inscription on paper, and more recently projection on to screens. In many cases, phenomena of interest to researchers are rendered into visible and palpable form only through such instrumental intervention and mediation. Different complexes of practice and instrumentation are associated with distinctive standards of objectivity modes of objectification. Images and models not only depict preexisting objects; but they sometimes have a proactive role in articulating configurations that only later are realized by imaging technology. Studies of visualization in science emphasize the links between different forms of data representation, and the labor process that develop and maintain the integrity of those links.

Scientific knowledge is represented in many forms: spoken and printed words, graphs, formulae, tables, drawings, engravings, diagrams, photographs, charts, maps, models, diorama, demonstrations, and many other more specialized renderings. Although linguistic propositions, measurable quantities, and mathematical formulae continue to be treated in philosophy of science as the primary bearers of scientific knowledge, historians, sociologists, anthropologists, and even a few philosophers have grown accustomed to treating images, maps, sounds, and soundscapes as significant and irreducible tools and products in science. Moreover, the explosion of digital image processing in recent decades has drawn scholarly attention to the remarkable ease with which digital and digitized data are rendered into two- and three-dimensional images, animations, and models that combine and recombine multiple forms of graphic, pictorial, sensual, and verbal formats (Carusi et al., 2014; Coopmans et al., 2014).

Some of the most interesting work in science and technology studies (STS) examines complex relations among material, pictorial, and discursive constituents in scientific work. Often a scientific text, or even a single image, incorporates mathematical, verbal, and pictorial expressions in a single frame, so that cross-referential relations are at least as significant as the particular graphic forms and formats. Published tables and figures are just the tip of the iceberg, and illustration is but one of the many functions of visual images.

Three related senses of visual representation are important for current historical, literary, and ethnographic studies of science. A passive or static sense of the concept of representation refers to an image, inscription, or sign that depicts, signifies, or symbolizes something else. An active or dynamic sense of the word refers to the practice of *making* images and the action of *doing* representing; a sense that is closer to Ian Hacking's (1983) emphasis on how observation involves a practical *intervention* with the material things and fields observed. Yet another sense, developed in actor-network theory, plays off of the concept of political representation and applies it to the material domain in which an image, model, or prototype is a delegate that stands proxy for, and acts on behalf of, a nonpresent multitude (Latour, 1993). The term *visualization* often is preferred to representation, as it more specifically draws attention to the material practices and products of drawing, inscribing, engraving, graphing, mapping, amplifying, enhancing, and combining what is seen, heard, touched, or mechanically exposed and recorded. Research in the transdisciplinary field of sound studies (Thompson, 2004; Pinch and Bijsterveld, 2012; Sterne, 2012) has convincingly shown that scientific observation and technological intervention are not limited to visual apprehension and optical instrumentation. Observations, measurements, and demonstrations in many research fields often involve a kind of synesthesia in which sounds, vibrations, spectrographic patterns, nonvisible electromagnetic radiation are pictorially and graphically presented. However, such data almost invariably take visual form on paper (and, increasingly, on screens), even when the phenomena presented are not visual or optical.

Representing and Visualizing

The study of visual representation in science crystallizes lines of work from diverse academic disciplines. In the 1970s and early 1980s historians of science began to pay close attention to nonverbal representations and nonpropositional forms of reference (Rudwick, 1976), and at about the same time art historians took greater interest in the broad array of what Elkins (1999) has called "images that are not art": pictures and models that are crafted as products and by-products of scientific and medical work, but which are rarely considered as art (Edgerton, 1976, 1991; Alpers, 1983). Such scholarly attention to visual images, as material products of interest in their own right, motivated interest in the role of artisan skills, craft techniques, and visual technologies in the scientific revolution, and in the continuing place of visual arts and crafts in day-to-day laboratory work.

Visual Representation

Studies of visual representation in science include analyses of published and unpublished images, and of the historical crafts and technologies responsible for their production. Currently, there are numerous disciplinary and cross-disciplinary approaches to visual representation in the philosophy, history, and social study of science. Social, cultural, and literary studies of visual representation include a bewildering array of historicist, semiotic, feminist, cognitivist, epistemological, pragmatist, and other lines of interpretation. As many cultural and literary studies have emphasized, visual representations do not simply resemble natural objects: they express normative assumptions about the natural and cultural world; they incorporate historically and culturally specific artistic conventions; and they mediate political and cultural relations of patronage, commodity exchange, and subjugation. The active sense of the concept of representation points to a broad array of practices and technologies for making, producing, and shaping images and relations among images.

Visualization

Visualization means more than seeing or depicting, as it refers to the use of research tools and practices for uncovering hidden features, magnifying objects and enhancing analyzability, facilitating measurement and comparison, and making things, properties, relational configurations, and entire fields publically witnessable, intelligible, and accountable. Visualization includes efforts to convince as well as to exhibit. Consistent with a programmatic aim to investigate the content of natural science knowledge (Bloor, 1976), ethnographers, ethnomethodologists, and social historians closely examine the situated practices, inscription devices, and textual forms through which scientific data and pictorial representations are exhibited, analyzed, stabilized, objectified, and disseminated (Latour and Woolgar, 1986; Shapin and Schaffer, 2011[1985]; Lynch and Woolgar, 1990). These studies stress the significance of visualization as a crucial aspect of a continuing scientific revolution. This revolution involves the incessant invention, adaptation, and use of novel instrumental and literary technologies for revealing invisible objects, analyzing and demonstrating abstract relations, and disseminating knowledge.

Renaissance innovations in print and engraving technology, together with novel optical devices and artistic conventions, enabled natural philosophers to inscribe and reproduce faithful representations of observational experience (Ivins, 1969; Eisenstein, 1979). Latour (1990) uses the expression "immutable mobiles" to capture the dual sense of stability and mobility attributed to imprinted or engraved representations. Scientific matters of fact, while still controlled by those who had the means to purchase, build, and operate equipment like the telescope, microscope, and pneumatic air pump, were communicated widely through networks of coinvestigators, patrons, and other "virtual witnesses" (Shapin, 1984). In most cases, virtual witnesses who read, for example, Robert Boyle's experimental reports were in no position to replicate the original experiments, but Boyle's detailed descriptions illustrated by engravings of observed phenomena and laboratory instruments stood proxy for hands-on experience. The literary technology of virtual witnessing enabled Boyle to disseminate experimental knowledge widely, while sparing all but a few well-equipped adepts from the difficulties and uncertainties involved in actually trying to replicate the experiments. The dissemination of printed and engraved representations facilitated, but did not guarantee, the stability (or immutability) of the knowledge in question, since printing, engraving, and publication compounded the contingencies of knowledge-production, enlarging the circle of potential disputants as well as adherents. The stability and mobility of images required a disciplining of the instruments and persons involved in their production and distribution. Instruments like the microscope and persons like engravers, printers, and publishers were subject to variable degrees of trust and distrust, and their activities became subject to legal institutions such as copyright (Johns, 1998). Consequently, whether located in a gradual historical process or in the temporal development of a particular laboratory or museum project (Knorr-Cetina and Amann, 1990; Star and Griesemer, 1989), 'immutable mobiles' depend upon continual efforts to control sources of distortion, disruption, and misinterpretation at every point of production and distribution.

Image, Object, and Objectivity

Numerous S&TS studies over the past four decades have examined the conceptual and interpretive practices associated with the production of visual images, and the communicative and community-building work associated with them. By elucidating the practical handiwork and collective labor process hidden behind the grand polarity of image and object, these studies challenge existing conceptions of the scientific image and its relationship to the natural world.

Studies of visualization address classic philosophical questions about the objectivity of reference and representation, but they give these questions an original twist. Instead of addressing these questions abstractly, S&TS researchers investigate specific cases in which scientists encounter practical dilemmas and devise provisional solutions. An ideal-typical version of such a dilemma is furnished by Hanson (1958: 4):

> Consider two microbiologists. They look at a prepared slide: When asked what they see, they may give different answers. One sees in the cell before him a cluster of foreign matter: it is an artifact, a coagulum resulting from inadequate staining techniques. This clot has no more to do with the cell, *in vivo*, than the scars left on it by the archaeologist's spade have to do with the original shape of some Grecian urn.

In this example, the observational instrument is both a material thing and a transparent medium through which other things appear. Depending upon what the microscopist sees on the slide, the instrument will be deemed a source of subjective deception or a means for revealing an objective cellular structure. One of Hanson's microscopists sees an artifact, but in the particular research context it is less like an artifact uncovered at an archaeological site (a Grecian urn) and

more like the unwanted traces on that artifact left by the spade that uncovered it. This example suggests that the distinction between objective structure and artifact can be contentious and difficult to resolve. In their restricted domain, the microscopists are faced with a standard philosophical question: "How do you know that the image stands for a real object?" Hanson argues that practitioners answer (or in many cases do not even raise) the question by relying upon their theoretical and experiential preconceptions about what can possibly exist in the domain under study. In his terms, they see the image as a theoretically possible entity or artifact. This conception of theory-laden perception denies that observational evidence is independent of interpretation, so that sensory evidence (aided by instruments) cannot unequivocally resolve deep theoretical disputes.

Like Hanson, Ian Hacking (1983) rejects any effort to treat representation in science in terms of passive observation, but, in contrast to Hanson's emphasis on the theoretical and conceptual background, Hacking points to the practical interventions performed with instruments. These interventions include cross-checks, calibrations, and probes that compare and contrast the outputs from different instruments and techniques. Convictions about reality inform, and develop from, a microscopist's competent handling of the laboratory instruments and materials. Hacking remains focused on the philosophical problem of how natural reality is validated, but he places the solution in the hands of the competent microscopist, aided by an arsenal of tools and instruments. The classical confrontation between subject and object is thus mediated (and sometimes obviated) by instruments and techniques. They are part of the means of perception, and like our eyes and our imaginations they sometimes are deemed as untrustworthy sources of subjective illusion. In microscopy and other laboratory sciences, correct judgment involves a mastery of technique as much as sharp eyes and an educated mind. A whole chain of preparatory actions precedes and sets up the dramatic confrontation between microscopist and cell. Cellular preparations are cultured, fixed, stained, counterstained, thin-sectioned, and mounted on a slide before they become visible (Lynch, 1985a). If the microscopist sees a foreign body – an artifact – in the cell, one or another of the preparatory techniques may be held responsible for the illusion. What the microscopist sees (or does not see) is thus bound up with a labor process that includes much more than looking down the tube of the microscope and focusing the eyes on an object.

When considered in light of the focus on instrumental praxis, visual representation (on paper, or on instrument screens) does not stand proxy for seeing with the eyes. Janet Vertesi (2014) suggests that "drawing-as" is a more adequate account of visualization than Hanson's interpretation of the Wittgensteinian theme of "seeing-as" (seeing a figure as something in particular, in line with the classic gestalt figure that can be seen alternatively as a duck or as a rabbit). The material media through which phenomena (including duck-rabbits) are inscribed on paper or projected on to screens are far from incidental features of observation. The philosophical question of how the mind relates to the world is not resolved on its own terms by reference to material graphic media; instead, it is transposed to a different space, where its resolution is provisional and practical.

Historicizing and Situating Objectivity

The concept of objectivity itself has been historicized in terms of artistic media and visualization technologies. Lorraine Daston and Peter Galison (1992, 2007) identify different eras of objectivity in the history of science. Objectivity, in their conception, is both a relationship of image to object and a historically situated moral ideal for apprehending that relationship. They identify these eras through an analysis of conventions used in illustrated atlases in various fields of natural history, science, and medicine starting in the sixteenth century and running through the present. The earliest era, "truth to nature" involved a commitment to the detailed rendering of the essential appearance and structure of natural phenomena. For example, a botanical specimen selected to illustrate its type, would be painted or sketched with great care to show characteristic details. Such careful depiction was attentive to, and productive of, representational standards designed to standardize the framing and perspective, and eliminate idiosyncratic variation. Renderings of this type can still be found in popular nature-study field guides (Law and Lynch, 1990), but Daston and Galison argue that "truth to nature" was succeeded (though not displaced entirely) in the nineteenth century by the moral ideal of "mechanical objectivity," in which mechanical reproduction of phenomena (through photography and other modes of imaging and data recording) was highlighted and artistic handiwork and interpretive intervention were actively discouraged (also see Porter, 1995). However, practitioners eventually realized that (suitably disciplined) judgment and (theoretically guided) imagination were necessary features of objectivity, because mechanical renderings could not by themselves stand as "images uncontaminated by interpretation" (Daston and Galison, 1992: 120), as they also were sources of noise, artifact, and other forms of deceptive appearance. Rather than keeping hands off of mechanical sources of data, contemporary scientists move among various instruments to configure fields of data and enable diverse forms of image production and analysis.

In their final chapter, Daston and Galison (2007: Chapter 7) suggest that a novel era of objectivity may be dawning in the twenty-first century, exemplified by the use of probe microscopes – instruments such as the scanning tunneling microscope (STM) and atomic force microscope (AFM). These instruments have been central to the coalescence of the multifield nanotechnology movement in science and engineering (Mody, 2011), and unlike optical instruments, they form topographic images at the nanoscale (a nanometer is one-billionth of a meter), analogous to the way a blind person sweeps a probe across a surface to feel the texture of a street or sidewalk and to detect any irregular features or obstacles lying in the path. Such instruments, according to Daston and Galison, afford haptic rather than optical apprehension, and the reality they access is 'presentational' rather than 'representational' – meaning that the properties of the object are inseparable from the operations of the instrument (as Prentice (2012) observes,

haptic apprehension has long been important for the practice of surgery).

Daston and Galison's provocative thesis effectively challenges the idea that objectivity is a static and coherent standard. Research on specific historical and contemporary cases of visualization supports much of what they argue, but it also complicates their narrative of a succession of eras characterized by the ascendency of particular moral/epistemological conceptions of objectivity. For example, Emma Frow's (2012) research on contemporary efforts by biology journal editors to police the use of digital image processing ('Photoshopping') indicates an unresolved tension that arises from technological possibilities that the editors hope to resolve with updated conceptions of truth to nature. The objective status of mechanical images also seems highly contextual in science and other fields, such as law. For example, while US courts in the nineteenth and early twentieth centuries were reluctant to treat photography and X-rays as self-sufficient forms of evidence (sufficient without supportive testimony by qualified witnesses), the popular (and, potentially, the legal) reception of brain scan evidence and DNA profile evidence seems much stronger and less critical (Golan, 2004; Dumit, 2004; Lynch et al., 2008). Such research suggests that, while notions of objectivity are historically situated, multiple versions of objectivity can be found at any given time in distinctive contexts of research, and that their relationship to one another is unlikely to fit into an orderly progression of eras.

Images and Models as Harbingers of Materialization

One of the most intriguing suggestions from studies of visualization is that images and models not only depict what scientists see with their instruments, they also have a proactive role in shaping how phenomena are visualized mechanically. Cambrosio et al. (1993) describe a controversy over diagrams that Paul Ehrlich drew at the turn of the twentieth century to illustrate his theory of the microbiological immune reaction. The diagrams included pictorial elements, but did not resemble any object that could be seen with microscopes or other instruments at the time. Instead, they were diagrammatic representations of hypothetical immunological entities (Cambrosio et al., 1993: 673; also see Kaiser, 2005 on Feynman diagrams). The ambiguity of the diagrams became a subject of controversy over whether their pictorial features presented viewers with a misleading "slippage from model to reality" (p. 684); a slippage from speculative chemical formulas to diagrams that included distinctive pictorial aspects reminiscent of cellular membrane configurations. To make a long story short, however, Ehrlich's pictures turned out to have heuristic value for materializing a novel entity – the antibody – and establishing the lock-and-key imagery that persist today in the field of immunology.

The idea that pictures, diagrams, and models not only stand proxy for observations but also are used proactively as materials for speculations, calculations, and tests of hypotheses is particularly clear in the fields of molecular biology and nanotechnology. The hand-held models, and later the animated computer models, used in molecular chemistry provide materials for puzzle-solving. In principle, the models incorporate structural features and combinatorial laws that act to constrain possible solutions, so that working with the models is akin to experimentation and measurement (Francoeur, 1997; de Chadarevian and Hopwood, 2004). Natasha Myers suggests that the image of subcellular "machinery" is more than a metaphor: it "embodies a 'fruitful collapse of metaphor and materiality'" (Myers, 2014; quoting Haraway, 1997: 97). Protein structures not only are depicted as though they fit together in gearbox-like assemblies, but they begin to look that way as imaging technology progressively reveals their structure.

The brief history of the nanotechnology movement also illustrates the idea of imaginary models and images as leading edges for mechanical visualization. K. Eric Drexler and the Foresight Institute were major promoters of nanotechnology in the 1980s. Many of Drexler's futuristic ideas about building functional and self-fashioning robots at the nanoscale were later dismissed as impossible, and the images of 'nanobots' (often depicted as though they were miniaturized submarines or space vehicles on voyages of discovery (Nerlich, 2008)) have been succeeded by more prosaic images of material surfaces and molecular constructions. Nanoscale phenomena are often said to be subject to quantum forces, and to differ profoundly from macroscale or even microscopic entities, and yet visual conventions for depicting clearly outlined, colorful, discretely positioned objects are abundantly used in both popular and professional depictions of nanoscale fields and objects. The translation of electronic current variations measured by a probe microscope into visual images of surface topography allows for a great deal of flexibility, and conventional models and images provide templates for resolving and visualizing such topography. Both serious and playful efforts to design molecules often configure them along the lines of cartoon figures and macroscopic machines and vehicles (humanoid stick-figures, nanocars and other nanovehicles) (Toumey, 2009; Ruivenkamp and Rip, 2010; de Ridder-Vignone and Lynch, 2012). Consequently, familiar things and artifacts provide imaginary templates for visualizing nanoscale entities and designing molecular machines.

Chains of Industry and Credibility

The STS treatment of visualization as temporal, embodied, material labor opens up a domain of investigation that is not covered by mentalistic and perceptual accounts of scientific observation. It ties in with a graphic and material, rather than cognitive or idealist, legacy, because the appearance of the scientific object is bound up with a temporal process, typically involving many hands in a division of labor. Marx famously credited Feuerbach with having stood Hegel on his feet, meaning that he placed Hegel's abstract dialectic on the firm footing of historical materialism. Less famously, Marx suggested a related way to bring history of science down to earth:

> … historiography pays regard to natural science only occasionally, as a factor of enlightenment and utility arising from individual great discoveries. But natural science has invaded and transformed human life all the more *practically* through the medium of industry; and has prepared human emancipation, however directly and much it has

> had to consummate dehumanization. *Industry* is the *actual*, historical relation of nature, and therefore of natural science, to man.
>
> Karl Marx, *Economic & Philosophic Manuscripts of 1844*.

A further inversion from feet to hands can be inspired by studies of visualization in science. For Marx, industry, rather than history of ideas, embodies the modern human relation to natural science. For STS, natural science is not simply the historical basis for modern industry and its relations of production. The laboratory's own labor process is itself a site of collective handiwork aided by instruments, with variable degrees of complexity, automation, and historical development. Marx speaks of the "working up" of raw materials into the form of a product; an artifact that embodies human purposes and provides a resumé of the labor process that produced it: "The product is after all but the summary of the activity of production." Among the products of laboratory work are data – carefully framed and emplotted observations and measurements which stand to nature much in the way that processed food does to raw ingredients. Scientists sometimes speak of raw data collected in the field or from electronic instruments, and they distinguish such data from processed images that result from 'working up' the data into presentable form.

Likening the visualization of a scientific data to a manufacturing process may seem to miss an essential point: that the object of scientific investigation is not an artifact, even though it may be difficult to establish the ultimate difference between natural and artificial features. In the vernacular of laboratory science, the object is found, not made (unless, of course, the scientist aims to file a patent claim), and it is distinguished both from deliberately constructed tools and inadvertently produced scratches, residues, and deformations. Close studies of day-to-day laboratory work have furnished many examples of how research materials are cultivated, handled, and reconfigured, but to conclude, for example, that all cellular structures are artifacts collapses a vernacular distinction that provides laboratory work with a fundamental task and tension. Such a conclusion also invites the sterile polarity and tiring polemics of the realist-constructivist debate. Latour tries to avoid the polemics by devising a neologism that straddles the polarity. According to Marx, the industrial product becomes a fetish – a commodity whose exchange value obscures the work of its production. According to Latour (1999: 272), the laboratory product becomes a 'factish' – a fact whose solidity and epistemic significance hides the day-to-day work of establishing and maintaining that solidity and significance. The 'factish' is never independent of its manufacture, but this does not reduce the fact to an empty screen for projecting arbitrary beliefs. The robust credibility of fact remains in place, not because it exists in an eternal space remote from human activity, but because of the way it is fashioned.

Latour's creative vocabulary is less instructive than his examples. In an ethnography of a Brazilian field project, Latour (1995) closely examined the series of practices through which a field site was made graphically analyzable. The project was concerned with a boundary between rainforest and savannah, and the question the researchers addressed was whether the trees were advancing into the grassland or gradually retreating. As scientific projects go, this one was simple. It involved only a few researchers, a down-to-earth problem, no complex equations, and relatively simple technology. Nevertheless, it involved a complicated series of representational practices: plotting out the field site into a systematic grid; selecting and tagging plant specimens; collecting and analyzing soil samples; plotting data on paper graphs; and endless record-keeping with multiple registers. The field site, the tagged specimens, the binned and color-coded soil samples, and countless other materials used in the project were hybrid objects. None of these things were purely natural or artificial, concrete or abstract, material or symbolic, real or imaginary, but a great deal of care was taken to preserve specimens, keep track of identities, and cross-check measurements and records. The project resulted in a publication in which a graphic figure represented the forest–savannah boundary, and the data collected during the project were used to support conjectures about the movement in question. Again, this was not a big or powerful project, and it resulted in tentative findings, but it served well enough to illustrate an important point about the credibility of the project's evidence. The researchers staked the credibility of their data on a complex network of accountability in which no single representation corresponded to an independent natural object, and yet all of them, in different ways, contained, encoded, traced, or preserved arguably natural properties of the field site. The credibility of the production was more like the strength of a rope, weaving together separate strands, and less like the compelling realism of a photographic image. Perhaps another example will help elucidate this point.

The police in the United States use the expression "chain of custody" to refer to the continuity of evidence as it is collected and transported from one place to another. So, for example, detectives and forensic pathologists collect blood samples at a crime scene, place them in tubes, and transport them to forensic labs, where they are analyzed with DNA profiling methods. Such evidence typically travels a complicated route, and it is transformed and repackaged several times in the process. Courtroom challenges have led the police to use a whole series of administrative procedures for checking, cross-checking, and validating the identity of evidence samples. Standardized forms, tamper-evident bags, stickers with bar codes on them, automated scanners, filing systems, and many other mundane bureaucratic artifacts create an administrative envelope that protects the credibility of evidence against challenge. The 'scientific' results of DNA profile analysis stand only as long as the administrative envelope resists challenge and no gaps are exposed in the chain of custody. None of the forms, bar codes, and protocols that make up the chain guarantee its integrity and continuity. Even an elaborately documented and impressively consistent chain of custody does not guarantee the truth of the evidential result against all possibility of error, mishap, fraud, and conspiracy, but when the numerical indices, signatures, and analytic results tell a consistent story, they are more likely than not to resist adversary challenge (Lynch et al., 2008).

It often seems natural to employ metaphors of vision, perception, and cognition when discussing scientific observation and representation, but the analogy with an industrial

labor process suggests a more material, and less ocular and mental, conception of visualization and its natural products. Consistent with the analogy with the assembly line, a scientific project is organized into a chain of tasks performed by a series of technicians. Specimen materials are collected, selected, prepared, rendered, and reorganized to expose, select, and mark relevant features, to upgrade analyzability, and to enframe data with instrumental and graphic fields (Knorr-Cetina, 1981; Lynch, 1985b; Goodwin, 1994; Latour, 1995).

Visualization thus includes the most fundamental relations between science and nature; it is not limited to pictorial images and illustrations. Studies of visualization take up epistemology's most basic topics – observation, representation, measurement, and objectivity – but in concrete, materially specific, situations of practice.

See also: Cultural Studies of Science; Laboratory Studies: Historical Perspectives; Nanotechnology: Social and Cultural Aspects; Science and Technology Studies, Ethnomethodology of; Science, Sociology of; Scientific Knowledge, Sociology of; Visual Sociology; Visual Turn.

Bibliography

Alpers, S., 1983. The Art of Describing: Dutch Art in the Seventeenth Century. University of Chicago Press, Chicago.

Bloor, D., 1976. Knowledge and Social Imagery. Routledge and Kegan Paul, London.

Cambrosio, A., Jacoby, D., Keating, P., 1993. Ehrlich's "beautiful pictures" and the controversial beginnings of immunological imagery. Isis 84, 662–699.

Carusi, A., Hoel, A.S., Webmoor, T. (Eds.), 2014. Visualization in the Age of Computerization. Routledge, London and New York.

de Chadarevian, S., Hopwood, S. (Eds.), 2004. Models: The Third Dimension in Science. Stanford University Press, Stanford, CA.

Coopmans, C., Vertesi, J., Lynch, M., Woolgar, S. (Eds.), 2014. Representation in Scientific Practice Revisited. MIT Press, Cambridge, MA.

Daston, L., Galison, P., 1992. The image of objectivity. Representations 40, 81–128.

Daston, L., Galison, P., 2007. Objectivity. Zone Books, Brooklyn, NY.

Dumit, J., 2004. Picturing Personhood: Brain Scans and Biomedical Identity. Princeton University Press, Princeton.

Edgerton, S.Y., 1976. The Renaissance Rediscovery of Linear Perspective. Harper & Row, New York.

Edgerton, S.Y., 1991. The Heritage of Giotto's Geometry: Art and Science on the Eve of the Scientific Revolution. Cornell University Press, Ithaca, NY.

Eisenstein, E., 1979. The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early Modern Europe. Cambridge University Press, Cambridge.

Elkins, J., 1999. The Domain of Images. Cornell University, Press, Ithaca, NY.

Francoeur, E., 1997. The forgotten tool: the design and use of molecular models. Social Studies of Science 27 (1), 7–40.

Frow, E., 2012. Drawing a line: setting guidelines for digital image processing in scientific journal articles. Social Studies of Science 42 (3), 369–392.

Golan, T., 2004. Laws of Man and Laws of Nature: A History of Scientific Expert Testimony. Harvard University Press, Cambridge, MA.

Goodwin, C., 1994. Professional vision. American Anthropologist 96, 606–633.

Hacking, I., 1983. Microscopes. In: Hacking, I. (Ed.), Representing and Intervening. Cambridge University Press, Cambridge, pp. 186–209.

Hanson, N.R., 1958. Patterns of Discovery. Cambridge University Press, Cambridge.

Haraway, D., 1997. Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse: Feminism and Technoscience. Routledge, New York.

Ivins, W., 1969. Prints and Visual Communication. Da Capo, New York.

Johns, A., 1998. The Nature of the Book: Print and Knowledge in the Making. University of Chicago Press, Chicago.

Kaiser, D., 2005. Drawing Theories Apart: The Dispersion of Feynman Diagrams in Postwar Physics. University of Chicago Press, Chicago.

Knorr-Cetina, K., 1981. The Manufacture of Knowledge. Pergamon, Oxford.

Knorr-Cetina, K., Amann, K., 1990. Image dissection in natural science inquiry. Science, Technology, & Human Values 15, 259–283.

Latour, B., 1990. Drawing things together. In: Lynch, M., Woolgar, S. (Eds.), Representation in Scientific Practice. MIT Press, Cambridge, MA, pp. 19–68.

Latour, B., 1993. We Have Never Been Modern. Harvard University Press, Cambridge, MA.

Latour, B., 1995. The 'pédofil' of Boa Vista: a photophilosophical montage. Common Knowledge 4, 144–187.

Latour, B., 1999. Pandora's Hope. Harvard University Press, Cambridge, MA.

Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton, NJ.

Law, J., Lynch, M., 1990. Lists, field guides, and the descriptive organization of seeing: Birdwatching as an exemplary observational activity. In: Lynch, M., Woolgar, S. (Eds.), Representation in Scientific Practice. The MIT Press, Cambridge, MA, pp. 267–299.

Lynch, M., 1985a. Art and Artifact in Laboratory Science. Routledge & Kegan Paul.

Lynch, M., 1985b. Discipline and the material form of images: an analysis of scientific visibility. Social Studies of Science 15, 37–66.

Lynch, M., Woolgar, S. (Eds.), 1990. Representation in Scientific Practice. MIT Press, Cambridge, MA.

Lynch, M., Cole, S., McNally, R., Jordan, K., 2008. Truth Machine: The Contentious History of DNA Fingerprinting. University of Chicago Press, Chicago.

Mody, C., 2011. Instrumental Community: Probe Microscopy and the Path to Nanotechnology. MIT Press, Cambridge, MA.

Myers, N., 2014. Rendering machinic life. In: Coopmans, C., Vertesi, J., Lynch, M., Woolgar, S. (Eds.), Representation in Scientific Practice Revisited. MIT Press, Cambridge, MA, pp. 153–175.

Nerlich, B., 2008. Powered by imagination: nanobots at the science photo library. Science as Culture 17 (3), 269–293.

Pinch, T., Bijsterveld, K. (Eds.), 2012. The Oxford Handbook of Sound Studies. Oxford University Press, Oxford.

Prentice, R., 2012. Bodies in Formation: Remaking Anatomy and Surgery Education. Duke University Press, Durham, NC.

Porter, T., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press, Princeton.

de Ridder-Vignone, K., Lynch, M., 2012. Images and imaginations: an exploration of nanoimage galleries. Leonardo 45 (5), 447–454.

Rudwick, M., 1976. The emergence of a visual language for geological science, 1760–1840. History of Science 14, 149–195.

Ruivenkamp, M., Rip, A., 2010. Visualizing the invisible nanoscale study: visualization practices in nanotechnology. Science Studies 23 (1), 3–36.

Shapin, S., 1984. Pump and circumstance: Robert Boyle's literary technology. Social Studies of Science 14 (4), 481–520.

Shapin, S., Schaffer, S., 2011[1985]. Leviathan and the Air Pump, second ed. Princeton University Press, Princeton (first ed., 1985).

Star, S., Griesemer, J., 1989. Institutional ecology, "translations" and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. Social Studies of Science 19, 387–420.

Sterne, J. (Ed.), 2012. The Sound Studies Reader. Routledge, New York.

Thompson, E., 2004. The Soundscapes of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900–1933.

Toumey, C., 2009. Truth and beauty at the nanoscale. Leonardo 42 (2), 151–155.

Vertesi, J., 2014. Drawing as: distinctions and disambiguation in digital images of Mars. In: Coopmans, C., Vertesi, J., Lynch, M., Woolgar, S. (Eds.), Representation in Scientific Practice Revisited. MIT Press, Cambridge, MA, pp. 15–35.

Vocabulary Acquisition

Fernand Gobet, University of Liverpool, Liverpool, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Shatz, volume 24, pp. 16292–16294, © 2001, Elsevier Ltd.

Abstract

Acquiring the words of one's language is, in principle, a challenging problem, but children solve it surprisingly easily. This article covers (1) the skills (speech perception, speech segmentation, and speech production) that children must acquire before producing their first words; (2) the main theories attempting to explain how the meaning of words is acquired; (3) the role of working memory in learning words; and (4) the acquisition of vocabulary beyond the early years. The tension between nativist and constructivist theories is discussed, as well as the challenges faced by future research.

Introduction

First words are typically produced between 10 and 18 months of age. Before reaching this stage, infants must acquire several competencies in phonemic perception, segmentation, and speech production. In spite of considerable empirical research and theoretical debate (mostly between nativist and constructivist researchers), we still do not know how children acquire vocabulary and language more generally. This article will consider, in order: (1) the skills that children must acquire before producing their first words; (2) how the meaning of words is acquired; (3) the role of working memory in learning words; and (4) the acquisition of vocabulary beyond the early years.

Acquiring Pre-Vocabulary Competencies

There is now good evidence that language acquisition starts before birth and that infants' perception of speech is affected by prenatal maternal speech. Three days after birth, infants already display a preference for familiar speech (Decasper and Spence, 1986). But there is still a long way to go before speech perception and production skills are sufficiently developed to enable the production of words. At the very least, before acquiring words, infants must develop three types of competencies: they must be able to perceive, segment, and produce speech.

Speech Perception

A phoneme can be defined as the smallest class of sounds that leads, in a specific language, to differences in meaning. For example, in English, /l/ and /r/ are two phonemes because replacing one with the other produces words with different meanings, as is apparent with '*r*oyal' and '*l*oyal.' (Note that phonemes are usually indicated by two slashes.) By contrast, /l/ and /r/ do not affect meaning in languages such as Japanese and Thai and thus are two instances of a single phoneme.

Infants face two sources of difficulty in finding out which phonemes to use (Rowland, 2014). First, the sound pattern of phonemes lacks acoustic invariance. Phonemes vary considerably in the way they are pronounced (the variations are called 'phones'): people speak with different accents, different speeds, and different pitches of voice (e.g., male vs female voices). In fact, it has been found that even a word as short and common as 'the' is pronounced in more than 50 distinct ways (Waldrop, 1988)! Another source of difficulty is that infants do not receive phonemes as an orderly sequence of sounds that are independent. On the contrary, these sounds overlap and interact – a phenomenon named *co-articulation*. For example, the phoneme /p/ is pronounced differently in 'my pet' and in 'your pet.'

Second, different sets of phonemes are used in different languages, as we have seen with the example of /l/ and /r/ above. It has been estimated that, in the world's languages, there are about 600 different consonants and about 200 different vowels. While most languages use from 20 to 37 phonemes (Maddieson, 1984), the problem faced by infants is to identify which phonemes are used in their language. Although the input obviously provides strong cues, this does not fully solve the problem because some groups of phones constitute a phoneme in some languages but not in others.

Segmentation

The speech signal is continuous and does not come with indications of where the words start and end. In this respect, it is different from many written languages that use spaces between words. Thus, one important and challenging task faced by infants is to segment speech into words – a task which is known as *word segmentation*. It is important to note that pauses in speech do not provide reliable cues to word boundaries, because pauses occur both between and within words.

Research has investigated the kind of cues that children might use to segment speech, and we briefly discuss the main candidates (see Ambridge and Lieven, 2011; Rowland, 2014; for a more detailed discussion). A first cue is provided by prosody, which concerns the rhythm, stress, and intonation of speech. Several languages have a characteristic stress pattern. For example, English has a trochaic pattern, where most words consist of a heavy stress followed by a light stress (e.g., language, infant). However, several languages do not offer a regular stress pattern (e.g., Inuktitut, an Inuit language of Canada, and French; Rowland, 2014), so this cue is not sufficient by itself.

We have mentioned earlier that the same phoneme can be pronounced with different variations (phones) depending on the surrounding phonemes and on the position within a word. Allophonic variation thus provides cues that can be used by infants for word segmentation. Another cue is offered by phonotactic regularities (i.e., the acceptable combinations of phonemes in a given language). Some combinations of sounds are not allowed within a word but could occur between words (or vice versa), and some combinations of sounds are more likely to occur at the beginning rather than the end of words (or vice versa). Thus, phonotactic cues can help segmentation.

Another source of information is provided by words that occur in isolation. Caregivers occasionally pronounce words that are clearly segmented. Using as criterion a gap of at least 300 ms in the speech stream after the preceding word and before the following word, Brent and Siskind (2001) found that 9% of the words met this criterion and were thus pronounced in isolation.

A final cue is provided by the transitional probabilities between words. Different syllables have a different probability of following a given word, and if infants learn even approximations to these probabilities, they can then use them as information to select the speech stream. That infants can do this, at least in an experimental situation, was shown by Saffran et al. (1996), who used a continuous sequence of nonsense syllables (e.g., bidakupadotigolabubidaku …).

Experiments have shown that infants can use the cues that we have reviewed; however, by themselves, these cues are not sufficient for successfully segmenting the speech stream. Ambridge and Lieven (2011) have therefore proposed that infants are likely to use multiple cues. In addition, infants gradually understand that speech conveys meaning and use meaning to isolate words.

Speech Production

While infants are precocious in perceiving speech sounds, production of the sounds of their language progresses only slowly. This is in part due to the immaturity of their vocal organs (e.g., the infant's larynx is located higher than the adult's, making it impossible to pronounce some sounds). It is generally admitted that speech production improves through a series of stages. Here we briefly describe the stages proposed by Vihman (1996) (note that the ages given are approximate).

In the first stage (*reflexive vocalization*, from 0 to 2 months), the sounds produced by infants include sounds expressing discomfort (e.g., crying), vegetative sounds (e.g., burping), and sounds related to physical activity (e.g., grunts). The second stage (*cooing and laughter*, from 2 to 4 months) is characterized by the appearance of comfort sounds, which normally occur when other people smile and talk. The third stage (*vocal play*, from 4 to 7 months) is characterized by vocal play sounds (e.g., squeals), which are made possible by the maturation and better control of articulation. This stage also sees the beginning of primitive babbling, with sounds starting to resemble consonants and vowels. The final stage (from 7 months on) comprises two parallel substages. In the *reduplicated/canonical babbling* substage, starting from 7 months of age, infants repeat speech syllables (e.g., da-da-da). In the *variegated babbling* substage, infants produce combinations of syllables (e.g., ka-da-bu-da). This second substage increasingly dominates the first substage over time.

It is interesting that the first words, which tend to occur at this stage, contain the syllables ma, pa, and na, such as in *papa* and *mama*. Jakobson (1962) has argued that these speech sounds, which contain bilabials and the open vowel /a/, are the simplest to produce. Whether they are really intended to refer to parents, as parents commonly think, is debatable.

Even after the first words have been produced, it takes a long time for infants to learn to pronounce words correctly. Lust (2006) has identified four types of phonological processes that produce errors in early child language: omission of sounds (e.g., 'broke' is pronounced as 'bok'); substitution of sounds (e.g., 'rabbit' is pronounced as 'wabbit'); assimilation of sounds (e.g., 'doggy' is pronounced as 'goggy'); and repetition of sounds (e.g., 'daddy' is pronounced as 'dada'). Several explanations have been proposed to account for these errors. First, infants mispronounce some words because they misperceive them (Lust, 2006). Second, infants need time to learn and optimize the way they use their speech organs to articulate words correctly. Third, the production of correct speech sounds depends on a maturational program, which is universal (Jakobson, 1941/1968). Finally, pronunciation depends on learning and correctly using word patterns or word 'templates.' When these templates are used incorrectly, errors occur (Macken, 1995).

Learning the Meaning of Words

Children's vocabulary learning begins slowly, but rapidly increases – at the age of 16 months children know around 40 words (Bates et al., 1994), yet by school age children learn up to 3000 words each year (or just above 8 words every day) and a 17-year-old learns up to 10 000 words each year (or more than 27 words every day), mainly through reading (Nagy and Herman, 1987).

A classic phenomenon reported in most textbooks on child language is the 'vocabulary spurt': after having learnt around 50 words, children suddenly and dramatically increase the rate at which they learn new words. However, it seems that this spurt is a statistical artifact. When the appropriate statistical analysis is carried out, it appears that, with most children, there is no sudden increase in learning rate (Ganger and Brent, 2004).

At face value linking words to their meaning involves overcoming a number of insurmountable difficulties, of which two have been classically considered as particularly critical. The first is the *reference problem*, discussed at length by the philosopher Quine (1960). When hearing a word (e.g., 'gavagai') in a given situation (e.g., a rabbit scurries by), how does a child (or a linguist) know what the referent of gavagai is? Is it the rabbit, its color, the action of running, some part of the rabbit (e.g., its ear), the time of the day, the color of the grass the rabbit is running on, or an indefinite number of other possible referents? The second problem is the *extension problem*, which is actually common to the acquisition of all concepts, not only those labeled by words. Let us assume that a child learns that 'dog' refers to the family dog ('Snowy'). How does she correctly extend the class of animals and/or objects to which the label

'dog' can be correctly applied? How does she learn that 'dog' not only applies to 'Snowy,' but also to 'Fido' or 'Rintintin?' That it also applies to labradors, greyhounds, and bulldogs, but not to cats and coyotes? That it applies not only to living dogs but also to dogs made of plastic, wood, or metal? And that it does not apply to the parts of a dog, such as its ears or legs?

While these philosophical questions are interesting, they do not seem to bother infants and children – as we have just seen, they learn new words and their meaning fairly rapidly. Thus, an important scientific task is to uncover the mechanisms by which children can avoid the twin problems of reference and extension. A number of theories have been proposed to answer this question, and we briefly review six of them.

Theories Based on Innate Constraints

A first class of theories (Markman, 1990; Landau et al., 1988; Golinkoff et al., 1994) argue that the presence of innate constraints helps infants make the correct inferences when a new word is met. The details differ between the specific theories, and a large number of innate constraints have been proposed, but the following constraints provide a good sample of the kind of innate biases that have been proposed. A first constraint is the *whole object assumption*: when hearing a word, the child assumes that it refers to an entire object (or organism) rather than to its parts. A second constraint, the *mutual exclusivity assumption*, states that objects should have only one name. Thus, for example, if the child knows that the family cat's name is Pepsi, she would assume that words spoken in relation to this cat refer to something other than its name (e.g., they refer to an action or to a part of the cat). A third example is the *noun bias*: infants find nouns easier to learn than other kinds of words. A fourth constraint is the *shape bias*: children tend to label objects that have a similar shape with the same word. As a final example, consider the *principle of extendibility*: children take it for granted that words do not refer to a single exemplar, but rather to a category containing exemplars that share similarity.

Assuming such innate constraints meets with a number of theoretical and empirical difficulties (Rowland, 2014; Ambridge and Lieven, 2011). Theoretically, some constraints would actually make it difficult to learn many words. For instance, in our family cat example, the mutual exclusivity assumption makes it impossible for a child who knows that her cat's name is Pepsi to learn that it is a 'cat.' Another problem is that some of the biases are not empirically valid. Consider the *noun bias* assumption, which implies that nouns are learnt more quickly than verbs. One argument for this (putative) bias is that nouns tend to refer to individuals and concrete objects, while verbs are less directly related to perceptual input (e.g., Gentner, 1982). As it is plausible on biological grounds that more concrete stimuli are learnt more quickly than more abstract ones, if follows that nouns should be learnt more quickly. However, Ambridge and Lieven (2011) convincingly argue that, while this is true for some noun–verb pairs (e.g., *ball* vs *think*), it is not true for others. For example, in the pair *situation* vs *eat*, the noun is less concrete and thus harder to learn. They further argue that the empirical evidence for this phenomenon is weak, as most checklists used by parents contain more nouns than verbs. This bias reflects the statistics of the child's environment: at least in English, the number of nouns is about five times larger than the number of verbs. Nevertheless, the differential number of nouns and verbs creates a methodological problem. Ambridge and Lieven also note that experiments that have found that nouns are learnt more easily than verbs (e.g., Childers and Tomasello, 2002) suffer from the problem that these studies typically use nouns as objects, which means that nouns are selected for their concreteness.

Rather than following innate and universal constraints, it seems that children are sensitive to the semantic characteristics of the language they are learning. This is nicely illustrated in a study by Choi et al. (1999). While English distinguishes between containment (*put in*) and support (*put on*), Korean distinguishes between tight-fit relations (*kkita*) and loose-fit relations (*nehta*). In a preferential looking study with children aged between 18 and 23 months, Choi et al. found that children's eye movement behavior was consistent with the spatial categories of their own language. Thus, the children followed constraints specific to their own language rather than following innate and universal constraints.

Social Pragmatic Theories

As with the previous approach, social pragmatic theories include a variety of specific accounts (e.g., Bruner, 1978; Nelson, 1985; Tomasello, 2003). However, they agree on the key assumption: innate constraints are not necessary for acquiring new words as children can use a number of social cues (e.g., gaze, gestures, movements of the entire body, and facial expressions) to guide their inferences. The concepts of *joint attention* and *understanding of communicative intention* are central in this approach. While these theories are consistent with a fair amount of empirical data, they also suffer from a few weaknesses. For example, as noted by Rowland (2014), children can learn words before they have developed the abilities of joint attention and intention reading. In addition, autistic children learn words to a level similar to that of normally developing children, although they lack precisely these abilities.

Attentional and Associative Learning

A third account argues that word learning simply requires the use of general cognitive mechanisms dealing with attention, memory, and learning (Samuelson and Smith, 1998). Two learning mechanisms are particularly important: associative learning, where an association between two objects is made, and cross-situational learning, where regularities are learnt across situations. While this approach can account for several phenomena, it does not tell the whole story, and it is likely that an account based both on social pragmatics and attentional/associative learning is necessary (Ambridge and Lieven, 2011).

Syntactic Bootstrapping

This approach argues that the meaning of some words, especially verbs, cannot be acquired through a general mechanism

because of their complexity. Therefore, children must use constraints provided by language itself, and they do so by exploiting the information provided by syntax (Landau and Gleitman, 1985). Critics of the approach point out that syntax develops much more slowly than proposed, and thus might be of little use in learning new words (Tomasello, 2003), and that in some languages syntax actually hinders the acquisition of word meaning. For example, Margetts (2008) shows that in Saliba, an Oceanic language spoken in Papua New Guinea, two different verbs express the same meaning (e.g., to give) but are embedded in completely different syntactic structures.

Emergentist Coalition Model

The approaches we have just described focus on one kind of mechanism at a time. A natural extension is to propose an account combining all the mechanisms we have discussed so far. This is what Hirsh-Pasek et al. (2000) did in their emergentist coalition model. The key assumptions are that (1) children are sensitive to a number of cues, including those we have reviewed above; (2) the weight of these cues changes with age; and (3) constraints emerge as language develops. As the model incorporates a large number of mechanisms, it can account for a large amount of empirical data. However, champions of the approaches we have reviewed above retort that the model lacks parsimony, as word learning can be explained by one single mechanism rather than by a large number of them.

Latent Semantic Analysis

Compared to the previous approaches, computational approaches offer a different kind of explanation of word learning. Perhaps the best example is offered by latent semantic analysis (LSA) (Landauer and Dumais, 1997), so we focus on this approach. LSA is a mathematical technique suggesting that word meanings are acquired by exploiting mutual constraints from the language input. With children, these constraints come from interacting with the environment and hearing speech. LSA uses the shortcut of using the paragraphs of a text as a proxy for the episodes we experience in our lives. The central rationale is that the contexts in which a given word does or does not appear powerfully constrain and determine word meanings. Essentially, what LSA is doing is computing correlations between words within different contexts. However, the similarity values estimated by LSA are not simply based upon co-occurrence frequencies, but depend on a deeper statistical analysis (they are 'latent').

Concretely, in its computer implementation, LSA uses as input a corpus of text, where paragraphs define the relevant contextual units. The analysis of word occurrences is carried out across all paragraphs, which makes it possible to compute vectors that represent the meaning and relationships between words. Vectors are mathematically well suited to drawing semantic comparisons between words. To reduce dimensionality, LSA uses singular value decomposition, which is a generalization of factor analysis. Thus, large matrices of word-by-context data (e.g., encyclopedias) can be reduced into 100–500 dimensions. Higher-order co-occurrences are implicitly taken into account by the mechanisms used by LSA. Important contributions of this approach include the description of the acquisition of vocabulary as the construction of a semantic space and the insight that a considerable amount of 'learning' results from indirect induction. However, LSA suffers from two main weaknesses: it does not use syntax or morphology and it cannot handle negation.

Role of Working Memory in Learning Words

Children differ in the speed with which they acquire vocabulary and the number of words they know. Baddeley et al. (1998) have proposed that the phonological loop component of working memory (below, phonological working memory) is the main source of these individual differences. According to these authors, phonological working memory constitutes a bottleneck in the learning of new words. This is due to the fact that, compared to children with low phonological working memory capacity, children with high phonological working memory capacity can maintain more sound patterns in memory. This allows them to learn words more rapidly. Baddeley and colleagues largely support their view using empirical data obtained with the *nonword repetition task* (Gathercole et al., 1994), which they consider to be a reliable measure of the capacity of phonological working memory. In this task, the experimenter presents nonwords of varying lengths and asks children to repeat them back as accurately as they can. One advantage of the nonword repetition task, which uses strings that respect the phonotactic constraints of the language under study, is that it makes it possible to gather a large amount of data whilst at the same time making sure that children do not possess any knowledge of the specific strings that are being used as stimuli.

While influential, this theory suffers from a lack of precision, as it is expressed only verbally. Several computational models have been developed, which provide mechanisms that are better specified. We consider two of these models, one based on connectionism, the other based on the mechanism of chunking.

Gupta and Tisdale (2009) used a 'simple recurrent network' model, consisting of an input layer, hidden layer, context layer, and output layer. In this model, activation patterns represent well-formed phonemic sequences as part of syllabic slots. The model receives input one syllable at a time. Simulations of the nonword repetition test showed that the model captures the overall accuracy patterns, length effects, and error types (i.e., substitution, insertion, and deletion).

Chunking, a mechanism by which units are grouped to build increasingly larger units, is considered a central mechanism of human learning and cognition, including language (Gobet et al., 2001). With respect to the acquisition of vocabulary, Jones et al. (2007) have shown that chunking provides a plausible mechanism for how phonemes are grouped together to form new words. The specific model they use, EPAM-VOC, provides a detailed specification of the way long-term memory (where knowledge of vocabulary is stored) and phonological working memory interact. The model learns by receiving as input child-directed speech and text from storybooks written for children. It captures

key features of children's performance in the nonword repetition task: performance is better for shorter nonwords and for word-like nonwords; performance improves with age; and performance is superior for single consonant nonwords compared to clustered consonant nonwords (Jones et al., 2007). Simulations also suggest that long-term knowledge plays a more important role in explaining individual differences in the acquisition of vocabulary than the capacity of working memory (Jones et al., 2008). Finally, Tamburelli et al. (2012) examine the pattern of children's phonological errors within the syllabic domain. They show that the model captures the fact that children are more likely to make errors at the syllable onset than in the coda position. Thus, the model is able to acquire at least some of the phonotactic rules of English.

Together, the simulations with EPAM-VOC show that the input received by children contains a substantial amount of information, which can be used by relatively simple learning mechanisms to extract the kind of statistical information likely to be useful for acquiring language. They also suggest that developmental changes in the way children perform in the nonword repetition task may reflect differences in the amount of knowledge in long-term memory rather than maturational changes in the capacity of working memory.

Vocabulary Acquisition in Later Language Development

Most developmental psycholinguistics has focused on what happens in the first 5 years of life, perhaps because it is during this period that the greatest challenges are met. One consequence is that there has been little research on vocabulary acquisition in late childhood, adolescence, and adulthood. It is however obvious that there are important changes after early childhood, and that becoming a *proficient speaker* involves much more time than becoming a *native speaker* (Berman, 2007). With respect to vocabulary, there is both a quantitative and a qualitative change. Quantitatively, as seen earlier, individuals learn a considerable number of words during late primary and high school years. Many different sources make it possible to acquire so many words either incidentally or through instruction: school activities, reading books, the Internet, cinema, television, radio, and interaction with peers. Qualitatively, the vocabulary becomes more *marked* (Berman, 2007): it uses longer, more formal, and less frequent words, which often have a highly specialized meaning. In addition, increased use is made of derivational morphology: more words are formed with the help of affixes and compounding (e.g., *un-grateful-ness*). Lexical knowledge becomes more sophisticated, with an understanding and use of synonymy and polysemy. Around the beginning of high school, teenagers become more sensitive to linguistic registers and are able to gracefully and appropriately alternate between casual conversation with all its slang and colloquialisms, and more formal and academic language. In parallel with these changes in vocabulary use, other aspects of language also develop considerably: morphosyntax and metacognition (e.g., critical thinking and use of analogies). This evolution continues well into adulthood, at least with educated individuals.

Conclusion

Like the majority of research on language development, the field of vocabulary acquisition is dominated by the debate about whether language-specific innate constraints are necessary or whether it is sufficient to postulate a combination of general cognitive mechanisms and statistical information in the environment. While the latter theories have gained ground in recent decades, theories based on innate constraints are still championed vigorously. However, as witnessed by the many criticisms we have leveled against most of the theories discussed in this article, our understanding of the way children acquire vocabulary is still poor. As argued by Shatz (2007) for language acquisition in general, what is needed for scientific breakthrough in the field is a better understanding of the information processing abilities underpinning vocabulary acquisition as well as an understanding of the multiple interactions involved, including those between cognitive and social development.

See also: First Language Acquisition, Linguistic Theory of; Language Acquisition; Lexicon; Linguistics: Overview; Second Language Acquisition; Semantic Memory; Semantics; Speech Perception; Speech Production, Psychology of; Syntax (General); Working Memory, Psychology of.

Bibliography

Ambridge, B., Lieven, E.V.M., 2011. Child Language Acquisition: Contrasting Theoretical Approaches. Cambridge University Press, Cambridge.

Baddeley, A.D., Gathercole, S.E., Papagno, C., 1998. The phonological loop as a language learning device. Psychological Review 105, 158–173.

Bates, E., Marchman, V., Thal, D., Fenson, L., dale, P., Reznick, J.S., Reilly, J., Hartung, J., 1994. Developmental and stylistic variation in the composition of early vocabulary. Journal of Child Language 21, 85–123.

Berman, R.A., 2007. Developing language knowledge and language use across adolescence. In: Hoff, E., Shatz, M. (Eds.), Handbook of Language Development. Blackwell Publishing, London.

Brent, M.R., Siskind, J.M., 2001. The role of exposure to isolated words in early vocabulary development. Cognition 81, B33–B44.

Bruner, J.S., 1978. The role of dialogue in language acquisition. In: Sinclair, A., Jarvella, R.J., Levelt, W.J.M. (Eds.), The Child's Conception of Language. Springer-Verlag, New York.

Childers, J.B., Tomasello, M., 2002. Two-year-olds learn novel nouns, verbs, and conventional actions from massed or spaced exposures. Developmental Psychology 38, 967–978.

Choi, S., Mcdonough, L., Bowerman, M., Mandler, J., 1999. Early sensitivity to language-specific spatial categories in English and Korean. Cognitive Development 14, 241–268.

Decasper, A.J., Spence, M.J., 1986. Prenatal maternal speech influences newborns perception of speech sounds. Infant Behavior & Development 9, 133–150.

Ganger, J., Brent, M.R., 2004. Re-examining the vocabulary spurt. Developmental Psychology 40, 621–632.

Gathercole, S.E., Willis, C.S., Baddeley, A.D., Emslie, H., 1994. The children's test of nonword repetition: a test of phonological working memory. Memory 2, 103–127.

Gentner, D., 1982. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In: Kuczaj, S. (Ed.), Language Development: Language, Cognition, and Culture. Erlbaum, Hillsdale, NJ.

Gobet, F., Lane, P.C.R., Croker, S., Cheng, P.C.H., Jones, G., Oliver, I., Pine, J.M., 2001. Chunking mechanisms in human learning. Trends in Cognitive Sciences 5, 236–243.

Golinkoff, R.M., Mervis, C.B., Hirsh-Pasek, K., 1994. Early object labels: the case for a developmental lexical principles framework. Journal of Child Language 21, 125–155.

Gupta, P., Tisdale, J., 2009. Does phonological short-term memory causally determine vocabulary learning? Toward a computational resolution of the debate. Journal of Memory and Language 61, 481–502.

Hirsh-Pasek, K., Golinkoff, R.M., Hollich, G., 2000. An emergentist coalition model for word learning: mapping words to objects is a product of the interaction of multiple cues. In: Golinkoff, R.M., Hirsh-Pasek, K., Bloom, L., Smith, L., Woodward, A., Akhtar, N., Tomasello, M., Hollich, G. (Eds.), Becoming a Word Learner: A Debate on Lexical Acquisition. Oxford University Press, New York.

Jakobson, R., 1941/1968. Child Language, Aphasia and Phonological Universals. Mouton, The Hague, NL.

Jakobson, R., 1962. Why 'mama' and 'papa'? In: Jakobson, R. (Ed.), Phonological Studies, Selected Writings, vol. I. Mouton, The Hague, NL.

Jones, G., Gobet, F., Pine, J., 2007. Linking working memory and long-term memory: a computational model of the learning of new words. Developmental Science 10, 853–873.

Jones, G., Gobet, F., Pine, J.M., 2008. Computer simulations of developmental change: the contributions of working memory capacity and long-term knowledge. Cognitive Science 32, 1148–1176.

Landau, B., Gleitman, L., 1985. Language and Experience: Evidence from the Blind Child. Harvard University Press, Cambridge, MA.

Landau, B., Smith, L.B., Jones, S., 1988. The importance of shape in early lexical learning. Cognitive Development 3, 299–321.

Landauer, T.K., Dumais, S., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.

Lust, B.C., 2006. Child Language, Acquisition and Growth. Cambridge University Press, Cambridge.

Macken, M., 1995. Phonological acquisition. In: Goldsmith, J. (Ed.), The Handbook of Phonological Theory. Blackwell, Cambridge, MA.

Maddieson, I., 1984. Patterns of Sounds. Cambridge University Press, New York.

Margetts, A., 2008. Learning verbs without boots and straps? The problem of give' in Saliba. In: Bowerman, M., Brown, P. (Eds.), Cross-linguistics Perspectives on Argument Structure: Implications for Learnability. Lawrence Erlbaum, New York.

Markman, E.M., 1990. Constraints children place on word meanings. Cognitive Science 14, 57–77.

Nagy, W.E., Herman, P.A., 1987. Breadth and depth of vocabulary knowledge: implications for acquisition and instruction. In: Mckeown, M.G., Curtis, M.E. (Eds.), The Nature of Vocabulary Acquisition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Nelson, K., 1985. Making Sense: The Acquisition of Shared Meaning. Academic Press, New York.

Quine, W.V., 1960. Word and Object. Technology Press of MIT, Cambridge, MA.

Rowland, C.F., 2014. Understanding Child Language Acquisition. Routedge, London.

Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.

Samuelson, L.K., Smith, L.B., 1998. Memory and attention make smart word learning: an alternative account of Akhtar, Carpenter, and tomasello. Child Development 69, 94–104.

Shatz, M., 2007. On the development of the field of language development. In: Hoff, E., Shatz, M. (Eds.), Blackwell handbook of Language Development. Blackwell, Oxford.

Tamburelli, M., Jones, G., Gobet, F., Pine, J.M., 2012. Computational modelling of phonological acquisition: simulating error patterns in nonword repetition tasks. Language and Cognitive Processes 27, 901–946.

Tomasello, M., 2003. Constructing a Language: A Usage-based Theory of Language Acquisition. Harvard University Press, Cambridge, MA.

Vihman, M.M., 1996. Phonological Development. Oxford, Blackwell.

Waldrop, M.M., 1988. A landmark in speech recognition. Science 240, 1615.

Vocational Education

Alison Fuller, Institute of Education, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In a period of historically high youth unemployment the availability and quality of vocational education takes on particular importance. It is seen by policy-makers and researchers as essential for supporting transitions from education to work, but concepts and purposes of vocational education vary. This article draws on international comparisons to explore the extent to which vocational programs are designed to prepare young people for broad vocational areas and citizenship or more narrowly for specific jobs. It identifies the 'societal approach' as a major way of helping to theorize and explain differences in, and similarities between, national vocational education systems.

Introduction

> Any scheme for vocational education which takes its point of departure from the industrial regime that now exists, is likely to assume and to perpetuate its divisions and weaknesses, and thus to become an instrument in accomplishing the feudal dogma of social predestination.
>
> Dewey, 2010: 214–215

While written nearly a century ago, Dewey's concern about the concept of vocational education and its relationship to work remains highly relevant. In the context of contemporary social and economic imperatives, policy-makers and researchers are questioning the extent to which vocational education pathways should prepare young people for specific jobs; more broadly for a changing, dynamic, and global labor market; and/or for active citizenship in twenty-first century democracies (e.g., OECD, 2010; Pring, 2007). Dewey's observation cogently highlights the persistent tensions that run through national and international debates on the purpose(s) of vocational education, who it is for, what it is, and what it is supposed to achieve and for whom. The importance of vocational education takes on added importance in light of historically high levels of youth unemployment in Europe, the US, and other parts of the world, and the known 'scarring effect' (e.g., Morrell et al., 1994) this has on young people.

Improving education and training is clearly only part of the answer to problems that in scale and scope go far beyond this area of public policy. Nonetheless, national and supranational policy-makers are attempting to create and sustain vocational education systems that can support and facilitate youth transitions from education to work and adulthood in a period of economic hardship and intense global competition for the available jobs. One indicator of this has been the recent international revival of interest in apprenticeship, the development of national policies to support the evolution and expansion of programs, and the means of evaluating their quality (e.g., in terms of their 'expansive' and 'restrictive' features) (Fuller and Unwin, 2012).

The purpose of vocational education can then be viewed on a spectrum with training for specific, narrowly defined jobs at one end and, at the other, a more general preparation designed to help young people leads productive and fulfilling adult lives as workers and citizens. The content of vocational provision will vary according to the underpinning aim. A range of key indicators, including the relationship between general and vocational education, assessment and certification, can be identified that offer evidence about the goal and nature of programs. For example, are participants required to pursue academic subjects such as native and foreign languages, science, mathematics, and humanities alongside their vocational studies; or is their education and training focused primarily on the specifics of the vocational area with minimal curriculum extension and enrichment? There are major differences in the concept and scope of general education within the vocational route. In the UK, it is narrowly defined as primarily English and Mathematics, and has previously been referred to as 'core' or 'key' skills model. Andy Green (1998) has provided a comprehensive critique of this approach, its assumptions, and implications. By contrast in other parts of Europe (e.g., Germany, Austria, and Switzerland), it includes a wider range of academic subjects and civic education (Clarke and Winch, 2007).

There are also related questions about the settings in which learning in the vocational route takes place: Are participants primarily located in the workplace with little off-the-job or classroom provision; is it the other way round with full-time attendance in vocational school or college and little opportunity for workplace experience, or is it a structured combination of participation in both workplace and classroom? Moreover, to what extent is the program designed in terms of the teaching and learning of vocational knowledge, including the relevant underpinning scientific, technical, and social theories, or around a list of occupational competences and standards that participants need to show they have met.

Issues of assessment and certification are relevant here too. The purpose and nature of the program will be reflected in how participants are assessed. This may include whether they sit formal written examinations and are examined separately on the general educational element of their program. If so, does this enable them to achieve dual certification for their academic and vocational achievements which facilitates movement between different kinds of programs, or is assessment confined to accrediting the acquisition of vocational competence and occupational expertise, for example, through observations in the workplace? Beyond this, what is the currency of the qualifications participants achieve for progression to higher

 http://dx.doi.org/10.1016/B978-0-08-097086-8.92091-9

education (HE) and entry to Bachelor degrees, or for entry to the skilled labor market? In concerning itself with questions such as these, this article provides an overview of some of the key debates and theories permeating discussions about vocational education, what it consists of and how it is positioned in wider educational and employment landscapes.

Vocational education and training (VET) can, of course, also be used as a route back into employment (e.g., for unemployed adults or women returning to the labor market after raising children), or to more highly skilled work and jobs for those already in employment, but the focus of this article is on initial vocational education for young people.

Why Is Vocational Education Important?

There are various reasons why vocational education matters. At one level the answer is all around us. Building the infrastructure that supports modern life, curing the sick, maintaining and repairing machinery and equipment, providing hospitality and leisure services, technological innovations in telecommunications and more, are only enabled by the availability of people with occupational knowledge, skills, and expertise. The need for vocational education for instrumental reasons, then, is obvious. However, going beyond this and getting to the heart of John Dewey's concern is a call for a holistic and humanistic concept of vocational education that recognizes the inherent and intrinsic values associated with learning for occupational expertise and practice of skilled work, for the individual and wider society (Sennett, 2008). Lorna Unwin has suggested that this broader understanding of vocational education should be reflected in all education (2009).

Most debates about the importance of vocational education and, in particular, the need for its reform, do not start with such philosophical concerns. For the most part, commentators invoke two main and often conflicting discourses. Firstly, in policy terms vocational education has increasingly been associated with economic competitiveness (OECD, 2010; EU, 2002). In the Lisbon Agreement, the European Union (EU) explicitly stated its goal (by 2010) to become "the most competitive and dynamic knowledge-based economy in the world" (Council, 2000: 3) later identifying the "crucial contribution from education and training as factors of economic growth, innovation, sustainable employability and social cohesion" (Council, 2002: 5 both cited in Ertl, 2006). The argument underpinning this and other 'competitiveness' policies is based on the assumption that improving the quality of VET has a positive effect on competitive advantage through generating greater productivity and facilitating innovation. Studies by Sig Prais and colleagues in the 1980s compared the productivity of matched plants in Germany and the UK, concluding that the better performance of German firms could be explained at least in part by the superior vocational preparation of their workers (e.g., Prais et al., 1989). Subsequently, this idea has been developed to suggest that the availability of a skilled workforce allows more productive forms of work organization where employees have more discretion and involvement in the decision making, and the development of higher value-added products and services (Finlay et al., 1998, see also Keep et al., 2006 for a critique of these assumptions). It follows that countries with highly developed and effective VET systems are likely to be more successful in the global economic competition, Germany is probably the country most often given as an example here.

Secondly, vocational education is viewed as a pathway to stable employment for school-leavers who do not progress to HE. In other words, it can be conceived as a 'transition system' for part of the cohort (Raffe, 2008). In some countries the vocational route is also seen as a vehicle for reengaging disaffected young people and social inclusion (Preston and Green, 2008; EU, 2010).

The existence of academic and vocational educational pathways raises the issue of comparative status between the two routes. The dualistic nature of much of the discourse around vocational education (Clarke and Winch, 2007; Pring, 2007; Young, 2004) represented by the academic–vocational, theory–practice, mental–manual, education-training binaries is unhelpful. A point also picked up by the American philosopher Richard Sennett in his recent book *The Craftsman*:

> History has drawn fault lines dividing practice and theory; technique and expression; craftsman and artist; maker and user; modern society suffers from this historical inheritance. But the past life of craft and craftsmen also suggests ways of using tools, organizing bodily movements, thinking about materials that remain alternative, viable proposals about how to conduct life with skills.
>
> Sennett, 2008: 11

Linda Clarke and Christopher Winch highlight how language provides a window on the contrasting values and meanings attributed to vocational education. For example, the German language has three terms that signify three different conceptions of the English word 'training' "first, in the sense of to train an animal … [where German] uses the verb *abrichten* …, second, to train someone to carry out a specific job or activity … *anlernen*; and, finally, the sense of education for an occupation like medicine or carpentry, where German uses *ausbilden*" (2007: 7). The language point is significant because it highlights cultural distinctions in the way VET is conceptualized and understood, and suggests that transferring the characteristics of vocational education from one country to another is far from straightforward. Different ways of conceptualizing and characterizing approaches to vocational education and their consequences for convergence and divergence between systems will be discussed in the next section.

Why Do Different Models of Vocational Education Exist in Different Countries?

There are various ways of organizing and providing vocational education. This can be seen in the diverse models of vocational education that exist in different countries. Given the national and international policy interest in the link between VET and economic competitiveness, researchers have examined the characteristics of systems and ways of grouping countries with similar institutional arrangements. In the main, international comparative research over the past 25 or so years has drawn on original empirical and theoretical work undertaken by French

scholars Maurice et al. (1986) who developed the 'societal approach.'

> The societal approach emphasises the holistic interrelationships among different social and economic institutions, including education and training, the labour market and industrial relations systems, the production system, family structures and cultures, and so on.
> Raffe, 2008: 278

Maurice et al. (1986) analyzed the relationship between education, the labor market, and other civic and social institutions in Germany and France. They identified that the nature of this societal relationship was distinctive in each country. Given strong evidence from the cross-national research literature, including some of the most extensive studies of country differences in the logics that underpin their vocational education systems (e.g., Hannan et al., 1997; OECD, 2010), it is not surprising that the extent to which countries converge or diverge has been a key area of debate (Müller and Shavit, 1998; Raffe, 2008).

A number of country typologies deriving in the main from the comparison of national data sets and cross-national surveys have been developed, drawing on dimensions and concepts that help classify countries into types. A review of the comparative literature suggests that three of these are particularly important and tend to be interrelated, 'logics and linkages,' 'standardization,' and 'stratification.' Firstly, as Ianelli and Raffe (2007) suggest the strength of the links between vocational education and the labor market varies from country to country depending on their industrial histories, civic and democratic culture, and the associated institutions and arrangements that have grown up in the relevant country context. They suggest that countries where the linkages between vocational education and employment are weak (e.g., Sweden) operate in line with an education logic, whereas those where linkages are strong can be viewed as having an employment logic. The nature of the labor market and particularly how it is regulated is a relevant factor here. The employment logic is strong in countries (e.g., Denmark and Germany) where entry to skilled occupations is regulated and restricted to those who have completed an agreed (according to existing institutional arrangements) vocational program (usually in the form of an apprenticeship) and achieved the required qualifications. It is less strong in countries such as the UK and the US which traditionally have been viewed as having more flexible and open labor markets, with access to occupations being less heavily regulated.

Secondly, the concept of standardization refers to the content of vocational curricula and the processes by which they are followed and assessed. The extent to which a vocational pathway is standardized in terms of the curriculum, assessment regime, and qualification(s) has an impact on how well employers understand and recognize the learning and achievements of those who have successfully completed it. Standardization makes it easier for employers to recognize the currency, and to trust the use value of the qualifications (extent to which they can be taken as a reliable proxy for the knowledge and competence award holders) achieved by participants. Allmendinger (1989) identified the importance of standardization, suggesting that young people experience much more straightforward transitions into work in systems with uniform standards. Given the role of standardization in supporting education – work transitions, it follows that countries whose approach to vocational education is characterized by employment logic are also likely to have standardized provision.

Thirdly, stratification refers to the proportion of young people staying in school for as long as the system allows (Allmendinger, 1989). Put another way, if a country's compulsory education system requires students to remain in school until the age of 16 but a further 2 or 3 years of upper secondary schooling are available, then what proportion of the cohort are able or choose to stay on? The more highly stratified the system, the smaller the participation rate in the final post-compulsory years of school and the higher the participation rate in other pathways, for example, apprenticeship. A further consequence of a highly stratified system is the smaller proportion of young people progressing to university. Hence countries with mass participation in HE such as Korea, Japan, and Canada tend to have smaller and weaker apprenticeship systems than countries with lower rates of participation in university such as Germany. The OECD *Education at a Glance* (2012) figures show that approximately 65% of Koreans aged 25–34 years have attained a higher level qualification compared with 25% of Germans in the same age group.

The identification of country typologies suggests that there is a relationship between logics and linkages, standardization and stratification. For example, Kerckhoff (2000) has proposed two 'types.' Type 1 societies (e.g., Germany) can be distinguished by their strong standardization, stratification, and linkages between vocational education and the labor market. In contrast, Type 2 societies (e.g., USA) have been characterized by nonstandardized programs, less stratification, and looser linkages between vocational education and employment. Some countries such as the UK and France are harder to classify and so have been viewed as falling between the two types. This suggests that the dimensions could be usefully perceived as continua, rather than either/or categories. Other typologies have focused primarily on patterns of participation. For example, Andy Green et al. (1999) identify national apprenticeship systems as high, low, or medium participation.

A key factor in the distinctiveness of a country's approach to and organization of vocational education is the role the state plays in the configuration of institutional and societal relationships that underpin its vocational education system. Understanding the role of the state in creating and supporting education – work pathways and transitions shed light on the strength of its interest and involvement in this area of policy (Green, 1990). Drawing on ideas from political economy, David Ashton (2004: 21) argues that in addition to the societal approach (outlined above) VET policy "is also conditioned by the state structures and the ability of the state to administer a central policy initiative." From this starting point he develops a typology of national systems that includes three models. Firstly, the 'free market' model (including the US, the UK, and Canada) is based on a long tradition of the state's 'laissez-faire,' voluntaristic approach to intervention in the relationship between capital and labor. In this model, the role of the state is to provide a legal framework that supports the workings of

market forces and underpins the 'dominance of capital over labor' (Ashton, 2004: 23). It follows that the main contribution of the state, here, has been to create and maintain an education system that provides citizens with basic general education and orientation to work. Thereafter the responsibility for building on this foundation lies with individuals and employers. VET provision can be bought from providers or supplied directly by companies, for example, in the form of on-the-job training.

Secondly and by contrast, in the 'corporatist' model (including Denmark and Germany), the state intervenes in the capital – labor relationship, thereby, producing a social partnership approach (also involving the Trades Unions) to achieving initial industrialization and later economic development. This model is underpinned by shared support for and involvement in the education and training system.

> The achievement of a consensus with the social partners means that this model creates a more institutionally dense environment for training, usually focused on the apprenticeship system.
>
> Ashton, 2004: 24

Thirdly, the 'developmental state' model (including South East Asian economies such as South Korea and Singapore) reflects an approach where, starting from a low industrial and educational base, the state has taken the leading role (since World War II) in promoting industrialization, and in shaping industrial relations and the education and training system.

In drawing attention to the different roles played by the state in the three models, Ashton is able to make further important points about the relationship between national VET systems, work, and employment. Leaving vocational education primarily to employers (free market model) means that provision is more likely to be narrow and task specific to meet the needs of particular organizations (i.e., nonstandardized). Where, as in the corporatist model, the consensus approach produces highly skilled workers (via standardized occupational apprenticeships), the challenge is how to adapt the system to meet new needs flowing from rapid technological innovation. In the case of the developmental state model, where the state plays a highly active not-to-say dominant role, it can push employers 'to adopt new methods of production' and individuals 'to learn the appropriate skills' (p. 34). In the context of an increasingly competitive global economy, then, it is apparently not enough for countries to create and sustain vocational education systems. In addition, the character of jobs created, linked to the mode of production and competitiveness strategies adopted by countries, sectors, and companies can act as a pull factor shaping what is required from the vocational education supply pipeline.

The discussion so far has suggested that there is a relationship between the purpose and nature of the vocational education pathway and employment and occupational opportunities. In other words, insights about the nature of the labor market, the kinds of jobs available, their status and grade, the way skills are constructed and rewarded, can be gained from studying the vocational education system and vice versa. Put more strongly, I would suggest that there may actually be a wash-back effect on the status, purpose, and nature of vocational education from the distribution of good quality (secure, well-paid, full-time permanent, respected status) jobs in national and sectoral labor markets. This raises the question of what kinds of jobs nation-states are preparing their young people for. Lorna Unwin has observed that in the UK politicians and policy-makers make denigrating assumptions about low-grade jobs being performed by individuals who are deficient in knowledge and skills. She criticizes this analysis and the policy 'solution' (which it can be noted is consistent with Ashton's free market model as the state stops short of proactively intervening with employers):

> … at the same time as being dismissive of many occupations, policymakers have become fixated on the individual as the source of the problem, and, hence, millions of pounds are being poured into the acquisition of qualifications without any parallel initiatives to improve job quality.
>
> Unwin, 2009: 2

It follows that enterprises that organize work and production in ways that foster the creation of jobs designed to foster employee discretion, involvement, and contribution are more likely (1) to feed into and support the development of a high-quality initial vocational education system from which to select well-educated and able staff, and (2) to create workplace learning environments capable of fully utilizing and further developing their skills (Fuller and Unwin, 2004; Felstead et al., 2009).

Characteristics of Vocational Education Pathways and Some Implications

In broad-brush terms, vocational pathways can be viewed as school-based, where students follow full-time programs of study linked to a broad vocational area such as business, economics, and commerce in a formal educational institution; apprenticeship, where apprentices are placed (and usually employed) in a company where they receive on-the-job training (usually) coupled with regular off-the-job training (such as day release in college); or mixed, where the vocational system includes both apprenticeship and school-based modes of participation. David Raffe (2008: 285) suggests that pathways can be analyzed along four dimensions:

- Size of vocational pathway in comparison with general education – the proportion of the cohort pursuing each route
- Nature of vocational pathway – particularly extent to which it is school-based, work-based, or a mixture of the two
- Occupational specificity – the degree to which the program is narrowly job specific or more broadly focused on a vocational area
- Relationship between general and vocational pathways – the degree of commonality between the modes of participation, pedagogy, assessment, and certification associated with each, and the ease of movement between them.

These four are key dimensions for illuminating the general character of the pathway but it is also worth identifying some other issues. In addition to the size of the vocational pathway, the main setting in which it is followed, the degree of

occupational specificity and the extent to which there is standardization and permeability between general and vocational pathways, it is important for social justice reasons to ask who is participating in what pathways: How evenly are young people with different social class, race, and gender characteristics represented across the system? Linda Clarke (2007: 63) discusses the role of vocational education in reproducing rather than challenging social inequalities. She argues that the institutional arrangements (discussed earlier in terms of the societal approach) that underpin the shape and structure of VET also reflect historically entrenched 'class and gender relations.' Stereotypical patterns of participation by females and males in vocational courses and apprenticeships continue to persist (Lakes and Carter, 2004) with young women being strongly overrepresented in fields such as hairdressing and child care and underrepresented in engineering and construction, sectors that are dominated by young men. Historically, the ability of institutional gatekeepers (e.g., the guilds, professional bodies) to oversee or even restrict entry (numbers and characteristics) into apprenticeships and occupations can have the effect of maintaining wage levels and job status:

> It [vocational education] is about entry into the labour market; about dividing up the potential labour force into different occupations and skills, each with a distinct quality. And it is about class and gender divisions, acting as a filter to include or exclude particular groups from particular occupations or industries and from acquiring a particular status in society.
>
> Clarke, 2007: 62

The classed nature of who participates in what pathway has been illuminated in highly insightful research by Shavit and Müller (2000). They drew attention to the 'diversion effects' of vocational education systems where most young people do not participate in HE and where, in a strongly stratified system (such as in Germany), it is hard to move between tracks. Their findings indicated that those from lower socioeconomic groups were more likely to be found in the apprenticeship pathway than their more socially advantaged peers. On the other hand, the more regulated nature of the labor market and the involvement of the social partners in designing and supporting apprenticeship meant that it usually provided a strong 'safety net' for participants (male and female) and vehicle for supporting their transitions to skilled jobs and permanent employment. There are few signs that the renewed international interest in expanding the apprenticeship route is, as yet, challenging gendered patterns of participation, although during the 2000s there have been some useful vocational education initiatives sponsored by the EU that have produced some positive outcomes (Fuller and Unwin, in press).

In addition to examining the characteristics of the population participating in vocational pathways, it is important to focus on content and in particular the approach taken to vocational knowledge (Winch, 2010; Bathmaker, 2013). In the UK this has been a strongly contested issue, as researchers and academics have critiqued the marginalization of knowledge in the outcomes-based design of National Vocational Qualifications (NVQs), which have dominated the VET system since the early 1980s (for a critique of NVQs see Grugulis, 2003).

Michael Young (2004) has identified and discussed distinctions between three approaches to knowledge and the vocational curriculum which he has related to different periods in the development of vocational education in the UK since the late nineteenth century. These also serve as useful lenses for analyzing the approach to knowledge in vocational education more widely, including in other countries. Firstly, the 'knowledge-based' approach arose from the perceived need for apprentices in engineering and science-based industries to be taught relevant scientific subject knowledge. It "assumed that the natural sciences were important for their specific content and as a model of reliable, objective knowledge" (Young, 2004: 187). This model was later extended to other vocational areas such as business studies which were taught in further education colleges (the main institutions providing vocational education in the UK).

Secondly, the 'standards-based' approach introduced in the 1980s arose from policy-makers' criticism of the knowledge-based model which they felt had become too academic and removed from what employers required. The new approach used functional analysis to develop occupational standards in line with employers' requirements. Individuals were deemed to be competent (through an accreditation process in the workplace) if they could perform tasks to the specified standards. The principal weakness Young (2004: 191) identified with the first model is that it "failed to consider how this new [scientific subject] knowledge could be recontextualised in the workplace." He went on to argue that the second model "neglected the extent to which only some of the knowledge relevant to particular occupations has its origins in workplaces," and therefore provides an emasculated approach to vocational knowledge. In response, Young identified a third model 'the connective approach' which aimed to overcome these limitations by combining theory and practice, and on and off the job learning. Michael Young and others (e.g., David Guile, 2010; Young, 2008; Winch, 2010) have gone on to develop ideas about the nature of vocational knowledge and expertise and how it needs to be reconceived, positioned, taught, and learned in the context of contemporary vocational and professional education and the dynamics of changing patterns or work organization and production.

Recent studies of the content and inclusion of vocational knowledge in full-time vocational courses and apprenticeship frameworks in England (Brockmann et al., 2010; Fuller and Unwin, 2011a,b; Bathmaker, 2013) have found continuing shortcomings inviting negative comparisons with the quality of vocational education in other countries.

It is also important to consider the value and worth of qualifications attained by participants in vocational education pathways. To what extent do they provide a strong platform for career and educational progression, and to what extent do they articulate and so support horizontal bridging to general educational pathways and vertical ladders to HE? To explore this, I draw on a recent EU Leonardo funded project on 'hybrid qualifications' (HQ) in England, Denmark, Germany, and Austria (Deissinger et al., in press). The

study explores the ability of qualifications positioned as providing both general educational and vocational currency to lead either to employment in the skilled labor market or to a place in HE. The focus of the research reflects the EU policy aims of improving the permeability and progression opportunities within and between national vocational education systems.

In England, Germany, and Denmark there was no well understood or established hybrid pathway that catered for a clear population of young people of upper secondary age. In Germany and Denmark the stratification of the school system into different academic (gymnasium) and vocational institutions (vocational schools and dual system apprenticeship) and tracks led to limited opportunities for movement between them. Moreover, the qualifications attained by young people in each track are not portable such that those who achieve the academic diploma are not qualified to enter the skilled employment in the regulated occupational part of the labor market, and those who successfully complete their apprenticeships do not accrue the qualification currency to access HE. There are experiments underway in both countries to try and develop HQ routes but, in the main, opportunities for vertical progression (either to the labor market or HE) rely on individuals gaining qualifications from both pathways (double qualifications) (Deissinger and Wern, 2011; Jørgensen and Lindvig, 2011).

In England the system is less stratified and standardized but the linkages to the labor market are looser (Davey and Fuller, 2011). Young people populating the middle or hybrid space can progress to HE if they attain a combination of academic and vocational qualifications through full-time study but there is no automatic labor market recognition of these attainments. Alternatively, those following the apprenticeship route are in employment, although the variable quality of programs (and limited labor market and occupational regulation) means that the strength of apprenticeship as a platform for career progression is uneven. The difficulty for apprentices is progression to HE because the currency of the qualifications available in their frameworks is usually insufficient for entry to Bachelor degrees, even those in subject areas that are cognate with the apprenticeship.

In Austria, by contrast with the other three countries, there is a well-understood and established hybrid route within a system that is characterized by stratification and standardization. The hybrid option sits between the country's solely academic pathway followed by approximately 20% of the cohort through the attendance at the gymnasium and its dual system (vocational school and work-based) apprenticeship route followed by approximately 40% of the cohort. Young people (approximately a third of the cohort) participate in the hybrid route where they attend specialist educational institutions known as vocational colleges. The distinctive feature of the vocational college provision is that it leads to a certificate of secondary education (the same Reifeprufung examination as their gymnasium peers) and a VET diploma supplemented with additional specific labor market qualifications. College graduates with these achievements possess the recognized entry qualifications to access the full range of university courses in as well as skilled employment (Aff et al., 2011).

Conclusions

The article began by contrasting a far-reaching, holistic conception of vocational education with a narrow occupationally specific alternative, indicating the range of purposes it may be required to fulfill and the unhelpful persistence of the academic–vocational divide. The ensuing analysis of national vocational education systems has indicated some of the important dimensions along which they can be differentiated and some of the implications for social justice and inclusion, as well as for educational and career progression.

The brief discussion of the HQ study illustrated the uneven value and portability of qualifications, and the scale of the challenge involved in fulfilling the increasingly popular (national governments, EU, OECD) aim of increasing permeability between pathways and progression opportunities from vocational education to HE and skilled employment. Case study comparative research is important here as it provides an effective way of identifying and explaining the diverse characteristics of vocational education systems. Evidence can be generated that shows, on the one hand, how approaches to vocational education reflect deep-seated historical, cultural, economic, and institutional differences between societies and, on the other, how external pressures and economic forces (e.g., related to the negative economic effects of the banking crisis, changing industrial structure and patterns of international labor mobility, and technological innovation) provide a transformed context with which all industrialized countries are grappling. Young people are particularly vulnerable to unemployment and to experiencing barriers to their successful transition to fulfilled adult lives and citizenship. The extant literature suggests that the availability of good quality vocational education can facilitate transitions, but contextual factors mean that what counts as effective, for whom, and in whose interests, is likely to continue to differ.

See also: Communities of Practice; Education and Economic Growth; Education and Employment; Education: Skills Training; Equity and Education; Youth Joblessness.

Bibliography

Allmendinger, J., 1989. Educational systems and labour market outcomes. European Sociological Review 5 (3), 231–250.

Aff, J., Paschinger, E., Rechberger, J., 2011. Illustration, Analysis and Reflection of the Vocational Education on Secondary Level II. Country Report, Austria. Available to download at: http://hq-lll.eu/.

Ashton, D., 2004. The political economy of workplace learning. In: Rainbird, H., Fuller, A., Munro, A. (Eds.), Workplace Learning in Context. Routledge, London, pp. 21–37.

Bathmaker, A.-M., 2013. Defining 'knowledge' in vocational education qualifications in England: an analysis of key stakeholders and their constructions of knowledge, purposes and content. Journal of Vocational Education & Training 65 (1), 87–107.

Brockmann, Michaela, Clarke, Linda, Winch, Christopher, 2010. The apprenticeship framework in England: a new beginning or a continuing sham? Journal of Education and Work 23 (2), 111–127.

Clarke, L., 2007. The emergence and reinforcement of class and gender divisions in vocational education in England. In: Clarke, L., Winch, C. (Eds.), Vocational Education: International Approaches, Developments and Systems. Routledge, London, pp. 62–76.

Clarke, L., Winch, C. (Eds.), 2007. Vocational Education: International Approaches, Developments and Systems. Routledge, London.

Council, 2000. Presidency Conclusions, Lisbon European Council (23 and 24 March 2000). Available online at: http://www.europarl.europa.eu/summits/lis1_en.htm (accessed 24.10.14.).

Council, 2002. Detailed Work Programme on the Follow-up of the Objectives of the Education and Training Systems in Europe (14 February 2002). Available online at: http://www.aic.lv/ace/ace_disk/Bologna/contrib/EU/workprog.pdf (accessed 24.10.14.).

Davey, G., Fuller, A., 2011. Best of Both Worlds or Falling between Two Stools, Country Report 2, England, Hybrid Qualifications Increasing the Value of Vocational Education and Training in the Context of Lifelong Learning. Available to download at: http://www.adam-europe.eu/prj/5213/prj/L%C3%A4nderbericht_2_England.pdf (accessed 24.10.14.).

Deissinger, T., Wern, R., 2011. Germany Country Report 2, Hybrid Qualifications Increasing the Value of Vocational Education and Training in the Context of Lifelong Learning. Available to download at: http://.hq-lll.eu.

Deissinger, T., Aff, J., Fuller, A., Jorgensen, C.H. (Eds.). Hybrid Qualifications – Structural and Political Issues in the Context of European VET Policy, XXX: Peter Lang.

Dewey, J., 2010. Democracy and Education: An Introduction to the Philosophy of Education. Indo-European Publishing, Los Angeles.

Ertl, R., 2006. European Union policies in education and training: the, Lisbon agenda as a turning point? Comparative Education 42 (1), 5–27.

EU, 2010. Education and Training for Social Inclusion: European Success Stories. Publications Office of the European Union, Luxembourg.

Felstead, A., Fuller, A., Jewson, N., Unwin, L., 2009. Improving Working as Learning. Routledge, London.

Finlay, I., Niven, S., Young, S. (Eds.), 1998. Changing Vocational Education and Training: An International Comparative Perspective. Routledge, London.

Fuller, A., Unwin, L., 2004. Expansive learning environments: integrating organizational and personal development. In: Rainbird, H., Fuller, A., Munro, A. (Eds.), Workplace Learning in Context. Routledge, London, pp. 126–144.

Fuller, A., Unwin, L., 2011a. The content of apprenticeships. In: Dolphin, T., Lanning, T. (Eds.), Rethinking Apprenticeships. Institute for Public Policy Research, London, GB, pp. 29–39.

Fuller, A., Unwin, L., 2011b. Vocational education and training in the spotlight: back to the future for the UK's coalition government? London Review of Education 9 (2), 191–204.

Fuller, A., Unwin, L. (Eds.), 2012. Contemporary Apprenticeship: International Perspectives on an Evolving Model of Learning. Routledge, London.

Fuller, A., Unwin, L. The challenges facing young women in apprenticeships. In: Schoon, I., Eccles (Eds.), Gender, Aspirations and Attainment across the Life-course. Cambridge University Press, Cambridge, in press.

Green, A., 1998. Core skills, key skills and general culture: in search of the common foundation in vocational education. Evaluation and Research in Education 12 (1), 23–44.

Green, A., 1990. Education and State Formation: The Rise of Education Systems in England, France and the USA. Palgrave MacMillan, Basingstoke.

Green, A., Wolf, A., Leney, T., 1999. Convergence and Divergence in European Education and Training Systems. Institute of Education, London.

Grugulis, I., 2003. The contribution of national vocational qualifications to the growth of skills in the UK. British Journal of Industrial Relations 41 (3), 457–475.

Guile, D., 2010. The Learning Challenge of the Knowledge Economy. Sense Publishers, Rotterdam, Netherlands.

Hannan, D., Raffe, D., Smyth, E., 1997. Cross-national research on school to work transition: an analytical framework. In: Werquin, P., Breen, R., Planas, J. (Eds.), Youth Transitions in Europe: Theories and Evidence (Documents Séminaires No. 120). CEREQ, Marseille, pp. 409–442.

Ianelli, C., Raffe, D., 2007. Vocational upper-secondary education and the transition from school. European Sociological Review 23 (1), 49–63.

Jørgensen, C., Lindvig, K., 2011. Views and Experiences of Stakeholders in Relation to Hybrid Qualifications, Country Report, Denmark, Hybrid Qualifications Increasing the Value of Vocational Education and Training in the Context of Lifelong Learning. Available to download at: http://.hq-lll.eu.

Keep, E., Mayhew, K., Payne, J., 2006. From skills revolution to productivity miracle – not as easy as it sounds? Oxford Review of Economic Policy 20 (2), 539–559.

Kerckhoff, A., 2000. Transition from school to work in comparative perspective. In: Hallinan, M. (Ed.), Handbook of the Sociology of Education. Kluwer, New York, pp. 453–474.

Lakes, R.D., Carter, P.A. (Eds.), 2004. Globalizing Education for Work: Comparative Perspectives on Gender and the New Economy. Lawrence Erlbaum Associates, Mahwah, NJ.

Maurice, M., Sellier, F., Silvestre, J.-J., 1986. The Social Foundations of Industrial Power: A Comparison of France and Germany. MIT Press, Cambridge, MA.

Morrell, S., Taylor, R., Quine, S., Kerr, C., Western, J., 1994. A cohort study of unemployment as a cause of psychological disturbance in Australian youth. Social Science & Medicine 38 (11), 1553–1564.

Müller, W., Shavit, Y., 1998. The institutional embeddedness of the stratification process. In: Shavit, Y., Müller, W. (Eds.), From School to Work. Clarendon Press, Oxford, pp. 1–48.

OECD, 2010. Learning for Jobs: Summary and Policy Messages. http://www.oecd.org/edu/skills-beyond-school/46972427.pdf (accessed 10.04.13.).

OECD, 2012. Education at a Glance 2012: Highlights. OECD Publishing. http://dx.doi.org/10.1787/eag_highlights-2012-en.

Prais, S., Jarvis, V., Wagner, K., November 1989. Productivity and vocational skills in services in Britain and Germany: hotels. National Institute Economic Review 130, 52–74.

Preston, J., Green, A., 2008. The role of vocational education and training in enhancing social inclusion and cohesion. In: Cedefop (Ed.), Modernising Vocational Education and Training: Fourth Report on Vocational Education and Training Research in Europe. Cedefop Reference series (69-I). Office for Official Publications of European Communities, Luxembourg, pp. 121–194.

Pring, R., 2007. 14-19 and lifelong learning: distinguishing between academic and vocational learning. In: Clarke, L., Winch, C. (Eds.), Vocational Education: International Approaches, Developments and Systems. Routledge, London, pp. 118–132.

Raffe, D., 2008. The concept of transition system. Journal of Education and Work 21 (4), 277–296.

Shavit, Y., Müller, W., 2000. Vocational secondary education. European Societies 2 (1), 29–50.

Sennett, R., 2008. The Craftsman. Allen Lane, London.

Unwin, L., 2009. Sensuality, Sustainability and Social Justice: Vocational Education in Changing Times, Professorial Inaugural Lecture. Institute of Education, London.

Winch, C., 2010. Dimensions of Expertise. Continuum, London.

Young, M.F.D., 2004. Conceptualizing vocational knowledge: some theoretical considerations. In: Rainbird, H., Fuller, A., Munro, A. (Eds.), Workplace Learning in Context. Routledge, London.

Young, M.F.D., 2008. Bringing Knowledge Back in. Routledge, London.

Vocational Interests, Values, and Preferences, Psychology of

Barbara Schneider and Walter G Cook, College of Education, Michigan State University, East Lansing, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by R.V. Dawis, volume 24, pp. 16299–16302, © 2001, Elsevier Ltd.

Abstract

This article presents a brief summary of the classical psychological theories of vocational interest (e.g., the Holland model). Next, it explores how insights from other subfields of psychology (e.g., developmental, positive, and social psychologies) have contributed to new understandings of how vocational interests form. Drawing on recent literature from these different domains, this article provides a synthesis of the current state of research into vocational psychology.

Introduction

Given that most adults spend nearly half of their waking hours at work, the psychology of vocational interests and preferences provides critical insights into the relationships between occupation and personal fulfillment (Robinson and Bostrom, 1994; Robinson and Godbey, 1997). Recent evidence suggests that workers whose occupations are strongly matched with their interests perform better at job tasks than those with weaker matches (Nye et al., 2012). Research also indicates that occupational competence and satisfaction is by in large an important determinant of overall satisfaction and happiness with life choices (Erdogan et al., 2011). Consequently, the study of vocational psychology is relevant for both predicting occupational choice and understanding the determinants of general well-being.

Traditionally, vocational preference theory has focused on matching one's personality type to an appropriate career as determined by the characteristics of the workplace environment including flexibility for making decisions, challenges, and complexity of tasks typical of that career (e.g., Holland, 1997). More recently, researcher emphasis has been to examine more closely how work-related interests form during childhood through adolescence and how these interests are reflected in later education and career choices (Krapp, 2007; Chope, 2011). This change is largely the result of findings that link vocational interests and educational expectations and attainment (Patrick et al., 2011). Empirical results further suggest that those whose occupational aspirations and educational expectations are misaligned have greater difficulty constructing a reasonable life path that leads to job stability and higher wages (Sabates et al., 2011).

This article begins by discussing theories of vocational preference and related empirical research initiatives. Next, several models from educational psychology are introduced that can be applied to understanding individuals' vocational interests. We conclude by describing contributions from positive psychology that have increasingly changed the scope of vocational psychology from one concerned primarily with work preferences and job satisfaction to one that considers overall life-course satisfaction. We take this direction, in part as it is estimated that jobs are likely to change frequently during young adulthood suggesting that vocational interests may be more porous than some have assumed (Bureau of Labor Statistics, 2012).

Foundations of Vocational Psychology

While theories of the source of vocation interest have existed since the 1900s with Parsons (1909); followed by Strong (1943), the basic foundation of vocational psychology can be traced to the work of John L. Holland (1959). (See Gottfredson (1999) for a synthesis of Holland's contributions.) Fundamentally, Holland's theory of job satisfaction, stability, and performance depends on matching one's personality with job environment. Holland identified six personality types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional (abbreviated as RIASEC), each of which can be characterized by a combination of interests, preferred activities, beliefs, abilities, and values. These personality types are linked to work environments that correspond to the same dimensions as the RIASEC categories. *Congruence* occurs when the worker personality type and vocational environment are complementary resulting in the likelihood of increased worker satisfaction compared to other situations where personality type and environment are not compatible. It is important to underscore that the idea of 'fit' between personality and environmental conditions has been adapted and modified to explain congruence and lack thereof in examining the relationships between individual dispositions and the characteristics of various social contexts such as family, peer groups, and classrooms and their reoccurring psychological and social qualities (Cook et al., 2002).

There are numerous measures of vocational interest; two of which have been widely used for assessing the validity of Holland's model including the Vocational Preference Inventory (Holland, 1965) and the Self-Directed Search (Holland et al., 1994). In addition to this classical measure of vocational interests, the Strong Interest Inventory, updated in 2004, provides information that corresponded to Holland's model (Donnay et al., 2004). A recent *meta*-analysis that included 60 different studies based upon Holland's model and similar vocational interest inventories found that vocational interest was a strong predictor of a wide variety of measures of education and work performance (Nye et al., 2012).

In the Nye et al. *meta*-analysis spanning over 60 years, correlations between interests and performance were calculated for four performance criteria – *task performance* (efficacy at completing assigned tasks), *organizational citizenship behavior* (OCB; being helpful coworkers), *persistence* (tenure with an employer), and *counterproductive work behavior* (CWB; actions that detracted from the completion of one's own tasks or the tasks of others). The relationship between these measures of performance were analyzed both with respect to interest by itself and then with indices of congruence. Results show that the positive correlations were stronger for the congruence indices of task performance, OCB, and persistence than interest alone, but the negative correlation with CWB was quite weak. Similar patterns were also observed when academic outcomes – grades and persistence – were used as outcome measures. The authors interpreted this analysis as positive evidence for Holland's theory of congruence.

In addition to Holland there have been numerous measures of vocational interest. Borgen (1986) and Rounds (1995) provide a detailed comparative overview of these and earlier measures of vocational interest that are no longer as widely used, such as the *Kuder Occupational Interest Survey* (Kuder and Zytowski, 1991). For a more recent comprehensive historical review and overview of current advances in the measurement of vocational interest, see Armstrong et al. (2011).

An Educational Psychologist's Conception of Interest

Psychologists use the term 'interest' to describe the enjoyment and willful engagement in an activity or task (Schraw and Lehman, 2001). The important role that interest plays to promote learning was recognized by early psychologists, including Johann Friedrich Herbart, William James, and John Dewey (Hidi, 1990; Schiefele, 1991). Two key constructs from this work include *personal* and *situational* interest. Personal interest is the disposition toward a type of activity based on past experiences and is relatively stable over time; situational interest refers to a temporary state that is prompted by a certain set of circumstances or stimuli. Holland (1997) incorporated aspects from both types of interest in his model of vocational interest. The six personality types are essentially manifestations of personal interest and relatively immutable traits whereas vocational interest is contextually dependent. Today, psychologists typically make a distinction between 'personal interest' and 'situational interest' and consider the construct as having both a cognitive and affective component (Hidi and Harackiewicz, 2000).

The distinction between personal interest and situational interest need not necessarily be dichotomous and discrete. A number of developmental psychological models have sought to integrate the two constructs. One such model is the four-phase model of interest development (Hidi and Renninger, 2006). In their model, while each prior phase is requisite to achieving the next phase, there is no guarantee that subsequent phases will be achieved: Phase One – 'triggered situational interest' – is a response to external stimuli that captures the individual's attention, perhaps because it is novel, unexpected, or unusually intense. Phase Two 'maintained situational interest' – occurs sometime after Phase One and is distinguished by an unusual level of strong concentration over a prolonged period of time. Maintained situational interest requires external stimulation to initially capture the individual's attention; however, attention is maintained as the subject of interest is personally meaningful. Phase Three – 'emerging individual interest' – represents a combination of psychological state and a nascent form of predisposition to seek out the subject of interest. The individual is aware, perhaps subconsciously, that similar past experiences have offered rewarding opportunities for personal fulfillment welcoming an opportunity to re-engage. Phase four – 'well-developed individual interest' – is also both a psychological state and stable personality trait. An individual does not require any external stimuli to promote a desire to undertake an activity, object, or task of interest. Someone who has achieved this level of interest will actively seek out opportunities to engage with the subject of interest, so one might consider this a pure form of personal interest.

The extant literature in vocational psychology has consistently shown that interests stabilize as one transitions from adolescence to early adulthood (Low et al., 2005). This finding is intuitively compatible with the four-phase model of interest development for several reasons. First, the likelihood of encountering a truly novel stimulus decreases as one's frame of reference toward the world increases. Second, as one ages, the more likely it is that previously well-developed interests may crowd out other less valued ones. Since individuals are constrained by limited time and capacity to pay attention, choices on what interests to cultivate compete with one another. The next section provides a framework for describing how individuals prioritize education pathways and perceived vocational opportunities.

Expectancy-Value Theory

Another major contribution from educational psychology that examines the origins of vocational choice is the modern expectancy-value theoretical framework developed by Jacquelynne Eccles, Allan Wigfield, and colleagues (e.g., Eccles, 1983; Eccles et al., 1989; Wigfield and Eccles, 2002; Wigfield et al., 2009). The primary application of this theory is to explain educational outcomes such as the selection of college and career choices based on a series of constructs such as achievement, values, and familial social contexts. The theory posits that opportunity costs are constrained by personal choices regarding the time and effort they invest in specific activities and tasks. Individuals choose between activities based upon future expected benefits and the perceived likelihood of success, which are referred to as *expectancy* and *task value*, respectively.

The construct of expectancy is based upon answers to questions such as, "Can I successfully complete this task?" The foundation of expectancy is past experience with similar tasks, and projected future competency. Expectancies are also based on prevailing social norms, expectations of others such as parents, peers, and teachers, and other external signals. For example, a consistent finding in this literature is that females often experience negative feedback with respect to their competence in mathematics, which becomes integrated into

the formation of their expectancy task value. A complete exploration of these stereotypic responses to particular situations is beyond the scope of this article, but social psychology explores the nature and mechanisms through which these perceptions form and propagate (for a comprehensive review, see Gilovich et al., 2006).

Individuals also construct task values based upon the answer to a question such as "Do I value the future returns of doing this task?" Essentially the individual is assumed to be a rational actor who calculates the expected future benefits of undertaking an activity against the expected future costs subject to information constraints. For example, a young child who fantasizes about being a professional baseball player as an adult is much more likely to want to practice hitting in a batting cage than a peer who does not share that aspiration. It should be noted that as individuals transition from childhood to adolescence and into early adulthood, their task values become more realistic as they obtain more information about the world (Schneider and Stevenson, 1999).

Within this theoretical framework, both expectancy and task value represent necessary, but individually insufficient conditions for motivation. For example, an individual who believes that success is attainable in a particular activity but sees no potential future value in undertaking it will be relatively unmotivated to pursue it. Similarly, an individual who lacks the confidence to succeed will be unmotivated regardless of the value of the perceived goal. When an activity satisfies both conditions, there is a positive feedback in the sense that the individual now has a positive experience that will increase future expectancies for similar activities. Early life educational experiences, termed *affective memories*, are integral to the development of vocational interests. Over time, a collection of positively reinforcing expectancies leads to relatively stable affinities toward those activities that eventually manifest as preferences toward certain academic subjects and career decisions.

The construct of task value can be further refined into four components: *attainment value*, *utility value*, *intrinsic interest*, and *cost belief* (Eccles, 1983). Attainment value represents the nexus between tasks and self-identity. Particularly for children and adolescents, the tasks that one undertakes play a critical role in determining how they perceive their own identity and sense of self. For example, to an adolescent, playing a trumpet in a jazz band not only represents a potentially rewarding activity that is intrinsically motivating, but may represent an essential aspect of how the student perceives him or herself, in this instance as an aspiring musician. Utility value represents the tangible benefits that the individual hopes to realize in the future. For example, a high school student who wishes to become an engineer may persevere through a difficult AP Calculus course because mastery of the material is necessary to gain admission to a competitive undergraduate engineering program. Even if the student does not enjoy the mechanics of solving the problems, he or she may still be motivated if they perceive that successful completion of the course will provide opportunities that the student views as particularly salient for realizing future ambitions.

Intrinsic interest is defined as the genuine enjoyment in undertaking a task that naturally develops from previous experiences. Specifically, within this framework, intrinsic interest develops from affective memories retained from similar tasks in the past. Cost belief refers simply to the anticipated effort required to acquire the level of attainment value that the individual seeks. For example, in the previously mentioned jazz trumpet player, one cost belief might include the amount of time of daily practice that the individual believes is necessary to obtain the level of proficiency to which he or she aspires.

Expectancy-value provides a widely used theoretical framework for describing the development of motivational beliefs, particularly in children, adolescents, and young adults. Numerous empirical studies confirm the validity of the framework within the educational context. Specifically, with respect to one's plans to take courses as well as enroll in them, a great deal of empirical research has shown that attainment value, intrinsic interest, and utility value are stronger predictors than discrete vocational preferences (Eccles, 1983; Meece et al., 1990; Wigfield and Eccles, 1992). In addition, confirmatory factor analysis has empirically established the validity of the constructs (e.g., Eccles and Wigfield, 1995). A thorough review of these findings is presented in Wigfield and Eccles (2000).

Flow Theory

Over the past decade, vocational psychology has increasingly integrated the insights and perspectives from what has become known as positive psychology. One of the foundational theories of this subfield is flow theory. Csikszentmihalyi (1985) describes flow as "The holistic sensation that people feel when they act with total involvement" (p. 36). The state of flow is most easily achieved when an activity presents a match of high skill and high challenge (Csikszentmihalyi, 1997). Beyond fostering contemporaneous engagement in an activity, flow is critical to interest formation because flow represents a positive experience and students will likely repeat activities in which they have previously experienced flow (Csikszentmihalyi, 1997). Csikszentmihalyi and Massimini (1985) posit that the psychic rewards of a flow experience (e.g., an increased level of enjoyment, self-esteem, sense of control) will encourage the student to seek to repeat flow experiences by finding similar opportunities; thus flow, like task value expectancy theory, provides positive feedback that encourages the individuals to continue studying academic subjects that initially engaged them.

Flow represents a theoretical framework that informs not simply the 'in-the-moment' engagement, but also provides insight into how the accumulation of experiences contributes to interest formation over time. In other words, the theoretical framework of flow provides a lens into understanding both how educational experience shapes the formation of intrinsic interests as well as what determines task satisfaction in the present. This dual explanatory power of flow occurs because the essence of the flow experience is the individual's enjoyment of intrinsic rewards simply from undertaking the activity itself (Csikszentmihalyi, 1990), which is perceived as more effective at fostering learning than various forms of extrinsic motivation (Deci and Ryan, 1985). Recent research confirms that intrinsic motivation is positively associated with 'deep learning,' while extrinsic motivation is associated with 'surface learning' (Moneta and Spada, 2009). For this reason, flow theory provides a useful lens into student engagement within an

educational setting (e.g., Shernoff et al., 2003), as well as engagement in work settings (e.g., Eisenberger et al., 2005; Sexton, 2005).

Because flow is necessarily a subjective experience, the quantitative measurement of flow has evolved over time. Moneta (2012) provides a comprehensive historical overview of how methodologies to measure flow developed in the literature over the preceding decades as investigation into flow has advanced from primarily qualitative modes of inquiry to quantitative. In addition, this piece provides an introduction to the most common instrument used in quantitative flow research: the *Experience Sampling Method* (ESM). The ESM involves repeatedly assessing individuals with a questionnaire over a period of time in order to contemporaneously record subjective experiences as well as contextual factors. Flow is achieved when two subjective preconditions are satisfied: high skill and high challenge. The ESM utilizes question stems such as "Indicate how you feel about the main activity" and then presents a nine-point Likert scale to indicate 'challenge of the activity' and 'your skills in the activity' in order to ascertain the activity's relative skill and challenge as perceived by the respondent. The ESM also typically includes additional questions such as "Was this activity interesting?", "How well were you concentrating?", "Did you feel like you were in control of the situation?", and "Were you living up to your expectations?" The responses to these questions are used to assess the quality of other subjective states with which flow is positively associated. Empirical studies of the occurrence of flow in educational settings consistently show that flow is associated with the expected correlates, including sense of control, high concentration, and living up to the expectations of oneself and others (e.g., Shernoff et al., 2003). Hektner et al. (2007) provide a detailed explanation of the design and implementation of ESM studies as well as a comprehensive presentation of technical issues related to measuring flow.

Just as a student naturally gravitates toward academic subjects and tasks that offer a higher propensity for experiencing flow, the flow theory predicts that a worker will be most engaged when partaking in occupational tasks that match both high skill and high challenge. Adults who work fulltime are more likely to experience flow during work activities rather than leisure activities and this finding holds across a wide range of occupational levels and professions (Csikszentmihalyi and LeFevre, 1989; Delle Fave and Massimini, 2005; Sartori et al., 2009). Flow is more readily achieved at work than at leisure. Studies consistently find that flow is not associated with elevated levels of enjoyment and excitement observed in studies of recreational activities. This finding is often referred to as the *work paradox* (Csikszentmihalyi and LeFevre, 1989; Csikszentmihalyi, 1997; Hektner et al., 2007). A number of hypotheses to explain this unexpected finding have been put forth. For example, it has been observed that work tasks center around obligations rather than activities of one's choosing and the loss of sense of control may prevent a worker from experiencing flow (Csikszentmihalyi and LeFevre, 1989). In addition, it is also possible that workers often possess a below average interest in the activity, which interferes with the opportunity to achieve flow despite a task's presentation of high skill and high challenge (Haworth and Hill, 1992). Finally, psychologists of happiness have simply noted that while goals are integral to completing work, they also are often negatively associated with happiness (Rheinberg et al., 2007). For a more comprehensive summary of the literature examining the work paradox phenomenon see, Della Fave et al. (2011).

Other recent flow research indicates that flow is sensitive to contextual or situational factors rather than dispositional. For example, Fullagar and Kelloway (2009) used a hierarchical linear model (Raudenbush and Bryk, 2002) analysis of data from an ESM study to show that about two-thirds of the variance in flow experiences is due to situational factors rather than some personal characteristic or disposition. In addition, by analyzing lagged data they found that flow was predictive of mood, but that mood was not predictive of flow. This distinction is important for vocational psychologists in that it suggests that making workers happier probably will not make them more engaged; rather, more engaged workers tend to be happier (among other positive affective states that an employer would prefer). Similarly, Nielsen and Cleal (2010) found that relatively stable aspects or characteristics of different jobs are not as strong predictors of flow as contextual or situational factors. Demerouti (2006) demonstrates that job characteristics could be predictors of flow as well as performance, but only for conscientious workers, which suggests that an individual's relative conscientiousness may be a moderator for the experience of flow.

The theoretical framework of flow has been applied to test the hypotheses advanced by Kohn and Schooler (1982) that there is a strong relationship with a worker's satisfaction with tasks during the workday and the emotional states that they experience at home. Specifically, the construct of flow has been utilized to test the hypothesis that characteristics of work tasks such as autonomy and intellectual challenge strongly influence the subsequent emotional experiences of the worker at both work and home. This finding corresponds to flow theory in that engagement is determined by a match between high skill and high challenge which is positively associated with subjective states such as a sense of control, interest in the activity, and enjoyment (Sexton, 2005). A key distinction in work tasks is made between primary work activities and secondary work activities, where the worker generally prefers the former to the latter. For the most part, workers seek to maximize the time they spend engaged in their primary activity – while minimizing the time spent engaged in less preferred activities, a condition one could identify as indicative of task value expectancy theory.

Among some social psychologists today, it appears that vocational preferences and values are regarded as less about the personality traits of the individual and more about the individual's experiences and the values he or she attaches to those experiences. It is the repetition and value of experiences, influenced in part by social contexts of the family, peer group, and others that determines career interest, motivation, and expectancy.

Conclusion

The application and integration of insights from educational and social psychology has propelled a secular trend in

vocational psychology toward seeking to understand how the workplace stimulates the worker's overall subjective well-being. Indeed, the growing prominence of positive psychology within the practice of clinical psychology has fundamentally changed the purpose of vocational psychology from a model of matching static preferences to understanding the dynamics of how worker preferences and workplace characteristics form and evolve. The increasing rate of technological innovation in the modern world necessitates that employees as well as employers each endeavor to be more fluid in satiating the emotional well-being of workers in order to maximize their productivity and tenure. The challenge for the next generation of vocational psychologists will be to provide the necessary apparatus to facilitate the realization of workers' avidity toward personal fulfillment.

Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, US Department of Education, through Grant R305B090011 to Michigan State University. This grant supports MSU's doctoral training program in the economics of education. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

See also: Academic Achievement Motivation, Development of; Apprenticeship and School Learning: Lessons from Germany; Avoidance and Approach Motivation: A Brief History; Burnout, Psychology of; Control Behavior: Psychological Perspectives; Decision Making: Nonrational Theories; Eudaemonism; Flow in Motivational Psychology; Gender and Academic Motivation; Interest, Psychology of; Motivation and Actions, Psychology of; Motivation in Australian Aboriginal Populations; Optimism and Pessimism; Passion and Motivation; Personal Projects; Personality and Adaptive Behaviors; Race and Academic Motivation; School Burnout and Engagement: Lessons from a Longitudinal Study in Finland; Schooling: Impact on Cognitive and Motivational Development; Self and Emotional Development in Adulthood and Later Life; Self-Regulation in Adulthood; Sociodigital Revolution: Digital Natives vs Digital Immigrants; Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model.

Bibliography

Armstrong, P.I., Su, R., Rounds, J., 2011. Vocational interests: the road less traveled. In: Premuzic, T., Stumm, S., Furnham, A. (Eds.), The Wiley-Blackwell Handbook of Individual Differences. Wiley-Blackwell, Malden, MA, pp. 608–631.

Borgen, F.H., 1986. New approaches to the assessment of interests. In: Walsh, W.B., Osipow, S.H. (Eds.), Advances in Vocational Psychology, The Assessment of Interest, vol. 1. Erlbaum, Hillsdale, NJ, pp. 31–54.

Bureau of Labor Statistics, 2012. America's Young Adults at 24: School Enrollment, Training, and Employment Transitions between Ages 23 and 24 (Publication No. USDL-12–0216). U.S. Department of Labor, Washington, DC.

Chope, R.C., 2011. Reconsidering interests: the next big idea in career counseling theory research and practice. Journal of Career Assessment 19 (3), 343–352.

Cook, T.D., Herman, M., Phillips, M., Settersten Jr., R.A., 2002. Some ways in which neighborhoods, nuclear families, friendship groups, and schools jointly affect changes in early adolescent development. Child Development 73, 1283–1309.

Csikszentmihalyi, M., 1985. Emergent motivation and the evolution of the self. In: Kleiber, D., Maehr, M.H. (Eds.), Motivation in Adulthood. JAI Press, Greenwich, CT, pp. 93–119.

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. Harper & Row, New York.

Csikszentmihalyi, M., 1997. Finding Flow: The Psychology of Engagement with Everyday Life. BasicBooks, New York.

Csikszentmihalyi, M., LeFevre, J., 1989. Optimal experience in work and leisure. Journal of Personality and Social Psychology 56 (5), 815–822.

Csikszentmihalyi, M., Massimini, F., 1985. On the psychological selection of bio-cultural information. New Ideas in Psychology 3 (2), 115–138.

Deci, E.L., Ryan, R.M., 1985. Intrinsic Motivation and Self-determination in Human Behavior. Plenum Press, New York.

Delle Fave, A., Massimini, F., 2005. The relevance of subjective well-being to social policies: optimal experience and tailored intervention. In: Huppert, F.A., Baylis, N., Keverne, B. (Eds.), The Science of Well-Being. Oxford University Press, Oxford, pp. 379–402.

Delle Fave, A., Massimini, F., Bassi, M., 2011. Work: a paradox in flow research. In: Della Fave, A., Massimini, F., Bassi, M. (Eds.), Psychological Selection and Optimal Experience across Cultures: Social Empowerment through Personal Growth. Springer Science and Business Media B.V., Dordrecht, pp. 155–175.

Demerouti, E., 2006. Job characteristics, flow, and performance: the moderating role of conscientiousness. Journal of Occupational Health Psychology 11, 266–280.

Donnay, D.A.C., Morris, M.L., Schaubhut, N.A., Thompson, R.C., 2004. Strong Interest Inventory Manual: Research, Development, and Strategies for Interpretation. Consulting Psychology Press, Mountain View, CA.

Eccles, J.S., 1983. Expectancies, values, and academic behaviors. In: Spence, J.T. (Ed.), Achievement and Achievement Motivations. W.H. Freeman & Co, San Francisco, CA, pp. 75–121.

Eccles, J., Wigfield, A., Flanagan, C., Miller, C., Reuman, D., Yee, D., 1989. Self-concepts, domain values, and self-esteem: relations and changes at early adolescence. Journal of Personality 57, 283–310.

Eccles, J.S., Wigfield, A., 1995. In the mind of the actor: the structure of adolescents' achievement task values and expectancy-related beliefs. Personality and Social Psychology Bulletin 21 (3), 215–225.

Eisenberger, R., Jones, J.R., Stinglhamber, F., Shanock, L., Randall, A.T., 2005. Flow experiences at work: for high need achievers alone? Journal of Organizational Behavior 26 (7), 755–775.

Erdogan, B., Bauer, T.N., Peiró, J.S., Truxillo, D.M., 2011. Overqualified employees: making the best of a potentially bad situation for individuals and organizations. Industrial and Organizational Psychology 4, 215–232.

Fullagar, C.J., Kelloway, E.K., 2009. 'Flow' at work: an experience sampling approach. Journal of Occupational Health Psychology 82, 595–615.

Gilovich, T., Keltner, D., Nisbett, R.E., 2006. Social Psychology. W.W. Norton, New York.

Gottfredson, G.D., 1999. John L. Holland's contributions to vocational psychology: a review and evaluation. Journal of Vocational Behavior 55 (1), 15–40.

Haworth, J., Hill, S., 1992. Work, leisure and psychological well-being in a sample of young adults. Journal of Community and Applied Social Psychology 2, 147–160.

Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M., 2007. Experience Sampling Method: Measuring the Quality of Everyday Life. Sage Publications, Thousand Oaks, CA.

Hidi, S., 1990. Interest and its contribution as a mental resource for learning. Review of Educational Research 60 (4), 549–571.

Hidi, S., Harackiewicz, J.M., 2000. Motivating the academically unmotivated: a critical issue for the 21st century. Review of Educational Research 70 (2), 151–179.

Hidi, S., Renninger, K.A., 2006. The four-phase model of interest development. Educational Psychologist 41 (2), 111–127.

Holland, J.L., 1959. A theory of vocational choice. Journal of Counseling Psychology 6 (1), 35.

Holland, J.L., 1965. Manual for the Vocational Preference Inventory, sixth rev. Consulting Psychologists Press, Palo Alto, CA.

Holland, J.L., 1997. Making Vocational Choices: A Theory of Vocational Personalities and Work Environments, third ed. Psychological Assessment Resources, Odessa, FL.

Holland, J.L., Fritzsche, B., Powell, A., 1994. Self-Directed Search: Technical Manual. Psychological Assessment Resources, Odessa, FL.

Kohn, M.L., Schooler, C., 1982. Job conditions and personality: a longitudinal assessment of their reciprocal effects. American Journal of Sociology 1257–1286.

Krapp, A., 2007. An educational-psychological conceptualisation of interest. International Journal for Educational and Vocational Guidance 7 (1), 5–21.

Kuder, F., Zytowski, D.G., 1991. Kuder Occupational Interest Survey: Form DD: General Manual, third ed. CTB Macmillan/McGraw-Hill, Monterey.

Low, S.K., Yoon, M., Roberts, B.W., Rounds, J., 2005. The stability of vocational interests from early adolescence to middle adulthood: a quantitative review of longitudinal studies. Psychological Bulletin 131 (5), 713–737.

Meece, J.L., Wigfield, A., Eccles, J.S., 1990. Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. Journal of Educational Psychology 82 (1), 60–70.

Moneta, G.B., 2012. On the measurement and conceptualization of flow. In: Engeser, S. (Ed.), Advances in Flow Research. Springer, New York, pp. 23–50.

Moneta, G.B., Spada, M.M., 2009. Coping as a mediator of the relationships between trait intrinsic and extrinsic motivation and approaches to studying during academic exam preparation. Personality and Individual Differences 46 (5), 664–669.

Nielsen, K., Cleal, B., 2010. Predicting flow at work: investigating the activities and job characteristics that predict flow states at work. Journal of Occupational Health Psychology 15, 180–190.

Nye, C.D., Su, R., Rounds, J., Drasgow, F., 2012. Vocational interests and performance: a quantitative summary of over 60 years of research. Perspectives on Psychological Science 7 (4), 384–403.

Parsons, F., 1909. Choosing a Vocation. Houghton Mifflin Company, Boston.

Patrick, L., Care, E., Ainley, M., 2011. The relationship between vocational interests, self efficacy, and achievement in the prediction of educational pathways. Journal of Career Assessment 19 (1), 61–74.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, second ed. Sage Publications, Thousand Oaks.

Rheinberg, F., Manig, Y., Kliegel, R., Engeser, S., Vollmeyer, R., 2007. Flow bei der arbeit, doch glück in der freizeit. Zielausrichtung, flow und glückgefühle (Flow during work but happiness during leisure time: goals, flow experience and happiness). Zeitschrift für Arbeits- und Organisations psychologie 3, 105–115.

Robinson, J.P., Bostrom, A., 1994. Overestimated workweek-what time diary measures suggest. The Monthly Labor Review 117, 11.

Robinson, J.P., Godbey, G., 1997. Time for Life: The Surprising Ways Americans Use Their Time. Penn State University Press, University Park.

Rounds, J., 1995. Vocational interests: evaluation of structural hypotheses. In: Lubinski, D., Dawis, R.V. (Eds.), Assessing Individual Differences in Human Behavior: New Concepts, Methods, and Findings. Consulting Psychologists Press, Palo Alto, CA, pp. 177–232.

Sabates, R., Harris, A.L., Staff, J., 2011. Ambition gone awry: the long-term socioeconomic consequences of misaligned and uncertain ambitions in adolescence. Social Science Quarterly 92 (4), 959–977.

Sartori, R.D.G., Massimini, F., Della Fave, A., 2009. Well-being in rescuing activities among professionals and volunteers. In: Paper Presented in IX ISQOLS Conference "Quality of Life Studies: Measures and Goals for the Progress of Societies", Book of Abstract, 190. Florence, Italy, July, 19–23.

Schiefele, U., 1991. Interest, learning, and motivation. Educational Psychologist 26 (3 & 4), 299–323.

Schneider, B.L., Stevenson, D., 1999. The Ambitious Generation: America's Teenagers, Motivated but Directionless. Yale University Press, New Haven.

Schraw, G., Lehman, S., 2001. Situation interest: a review of the literature and directions for future research. Educational Psychology Review 13 (1), 23–52.

Shernoff, D.J., Csikszentmihalyi, M., Schneider, B., Shernoff, E.S., 2003. Student engagement in high school classrooms from the perspective of flow theory. School Psychology Quarterly 18 (2), 158–176.

Sexton, H., 2005. Spending time at work and at home: what workers do, how they feel about it, and how these emotions affect family life. In: Schneider, B.S., Waite, L.J. (Eds.), Being Together, Working Apart: Dual Career Families and the Work-Life Balance. Cambridge University Press, Cambridge, UK, pp. 49–71.

Strong, E.K., 1943. Vocational Interests of Men and Women. Stanford University Press, Stanford.

Wigfield, A., Eccles, J.S., 1992. The development of achievement task values: a theoretical analysis. Developmental Review 12 (3), 265–310.

Wigfield, A., Eccles, J.S., 2000. Expectancy–value theory of achievement motivation. Contemporary Educational Psychology 25 (1), 68–81.

Wigfield, A., Eccles, J.S., 2002. The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In: Development of Achievement Motivation, pp. 91–120.

Wigfield, A., Tonks, S., Klauda, S.L., 2009. Expectancy-value theory. In: Wentzel, K.R., Wigfield, A. (Eds.), Handbook of Motivation at School. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 55–76.

Vocational Training

Núria Rodríguez-Planas, IZA – Institute for the Study of Labor, Bonn, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The objective of this article is twofold: first, it reviews the theoretical justifications for the existence of vocational education and training (VET); and second, it summarizes the empirical evidence. After discussing the major empirical identification problems, this article presents international evidence concerning the answers to the following questions: Is VET as effective as general-based education? What are the effects of postponing the tracking of students into vocational schools? Which type of VET best prepares workers for the labor market? What are the returns to a year of apprenticeship training? What is the life-cycle impact of VET? Is informal VET effective? The article concludes with a discussion on policy implications.

Introduction

Vocational education and training (VET) – sometimes also called technical and vocational education and training – refers to qualifying education paths that provide individuals with occupation-specific knowledge and practical skills, independent of the place, content, and education provider. This article focuses on the issue of *initial* VET as in Eichhorst et al. (2013) and Zimmermann et al. (2013), and in contrast to vocation-specific education and training as part of life-long-learning–reviewed, as in Arulampalam et al. (2004) and Bassanini et al. (2005).

VET can be classified into three distinct systems: (1) school-based; (2) a dual apprenticeship system combining school training with a firm-based approach; and (3) informal-based. School-based VET provides youths with a combination of general and occupation-specific knowledge, offered either at the lower or upper secondary level. Accordingly, it is integrated in compulsory schooling as alternative to an academically oriented schooling track, or as part of several postcompulsory education options. Similar to academic education, the skills provided by vocational schools are mostly general given their transferability between employers (Becker, 1964). However, there might be differences in the degree of transferability across occupations – while the VET system in some countries transmits skills that are not restricted to one particular occupation, in others it provides vocational schooling for specific types of occupations (Shavit and Müller, 1998).

Dual VET aims to combine general, transferable skills acquired during class-based VET with structured on-the-job learning and actual work experience within a training company. Apprenticeship graduates acquire occupation-specific skills that render them employable by the training company or other employers, while the overspecialization and low levels of transferable skills generated by the higher share of firm- and occupation-specific training is counteracted by standardized curricula and examinations.

Finally, occupational knowledge is transmitted between generations, within a family or clan in 'traditional apprenticeships' in India and many African countries. Moreover, 'informal apprenticeships' are frequently open to apprentices not belonging to a family or clan. Despite taking place in the informal sector, they have some locally standardized structures and duration, and are based on some contractual agreements between the craftsman and trainee. Apprentices receive no or little remuneration during the training period, and might even have to pay a fee to the trainer. While the training is generally entirely work-based, it may follow an informal training plan. Given the lack of a legal framework to enforce certain training standards, the informality of training may also bear certain risks to the trainee, with particular concerns involving the potential exploitation of children as cheap laborers, gender discrimination based on traditional gender roles, and unenforceable contractual agreements resulting in low training quality or an extension of the apprenticeship duration (ILO, 2011, 2012).

Theoretical Justification for VET

Why Do Governments Offer School-Based Vocational Training?

The government's supply of VET through the educational system can be justified as a means of improving the opportunities of youths who lack the skills demanded in the labor market, or the ability or motivation to continue with higher education (Quintini and Martin, 2006; OECD, 2010a, Middleton and Adams, 1993). Furthermore, individuals might prefer this option to academic education as it implies shorter investment of human capital and facilitates earlier entry into the labor market. Indeed, the close link to work tasks and hands-on-practical experience might represent a source of motivation for practically oriented youths to continue training and remain in school longer, increasing their chances of attaining skilled employment (Shavit and Müller, 1998).

VET is also seen as an instrument to improve labor market mismatch, given that it, ideally, aligns the initial education more closely to particular vocations and tasks demanded by employers. As highlighted by Almeida et al. (2012), the problem of mismatch is often seen as a main source of the high degree of unemployment in developing countries. In such countries, the education of youths with practical-oriented vocational skills is further considered a promising means to create flexible and self-responsible learning attitudes, which might thus better prepare youths for the requirements of the modern work place. Furthermore, given that poor skills and hence the low productivity of firms is often seen as the reason for low levels of development, the investment in vocational

education is often justified as a means to promote a bottom-up labor market transformation. Therefore, the installment of an efficient vocational education system is conceived as an important pillar of the transformation into a knowledge-based economy (King et al., 2007; King, 2011).

In most cases, participation in either vocational or academic courses during school is operationalized by tracking students into the two different education pathways. However, the benefits of such a tracking system are not clear, given that leaving school with vocational qualification often translates into reduced options of further postcompulsory education, and particularly of the academic type. Accordingly, the incentive effect of learning more practice-oriented skills might be mitigated by high costs of later switching to academic education. Although the technical possibility of transferring to academic education might exist, earlier tracking will lead to strongly divergent levels of skills and competences (Woessmann, 2008). Furthermore, through the separation of high- and low-type students, VET might counteract the equalizing potential of vocational education (Shavit and Müller, 2000). Given that only few youths practically manage to enter academic education after vocational schooling (Kogan, 2008), this option is often perceived as a dead-end track and second-choice education for poor academically performing or at-risk youths (Neuman and Ziderman, 1999), and a 'safety net' for early school dropouts and those who are less academically inclined.

In addition, many have highlighted problems with school-based training systems (Woessmann, 2008; CEDEFOP, 2008; among others), with one of the biggest challenges facing school-based VET involving the adaptation of training curricula to the perpetual technological changes of the labor market. Indeed, failure to do so leads to educational mismatch early in workers' careers. The adaptation of school-based education crucially depends on continuous and accurate labor market feedback, as well as the tight coordination among all stakeholders (government, employers, social partners, and educational institutions) in the development and updating of the relevant curricula, as explained by Eichhorst et al., 2013.

Why Would Firms Invest in General Training?

Some of the problems associated with purely school-based training systems, such as skill mismatch, might be solved by the dual VET, given that firm-based training should naturally level off labor supply and demand (Heckman, 1993). Nonetheless, much research has focus on explaining why any employer would provide and finance training. As noted by Becker (1964), while some share of general schooling is crucial in enabling apprentices to enter further employment relationships, firms in a perfectly competitive labor market where workers are paid according to their marginal productivity have no incentives to invest into general schooling, as workers can leave directly after the training period and reap all the benefits of their general skills acquired. Therefore, in order for firms to provide both specific and general training, the cost of general training has to be borne by the worker, which could be implemented by providing either state-funded school-based general education (as discussed earlier) or firm-based general training with workers paying for their training costs. Furthermore, firms could be incentivized to participate if they were able to recoup part of their investments by contractual arrangements ensuring that apprentices accept a wage lower than their marginal productivity during their training period, or if they are able to tie apprentices to the firm beyond the training period (Becker, 1964; Malcomson et al., 2003).

Based on the empirical finding that firms in some countries (such as Germany) invest in general training despite incurring a net cost during the training period (Dionisius et al. 2009), several theories have attempted to explain the training activities of firms (see Wolter and Ryan, 2011 for an excellent overview). In particular, Katz and Ziderman (1990) build a framework in which informational asymmetries regarding the abilities of workers and the quality of training received can lead to sufficient incentives for firms to invest in general training. Alternatively, Acemoglu and Pischke (1998) build an asymmetric information model in which the employer gains monopsony power through learning about their worker's ability during the training period, which incentivizes the employer to offer and pay for training. (In economics, monopsony implies a market form in which only one buyer interfaces with many sellers.) Furthermore, Acemoglu and Pischke (2000) highlight the existence of external certification as a signal to prospective employers that the trainee is willing to invest in their own training. Moreover, Dustmann and Schönberg (2012) discuss that the dual apprenticeship system is likely to be more successful in countries where firms can credibly commit to the provision of training during the apprenticeship, through apprenticeship regulation and external accreditation.

Another set of models aims to explore the deterring effect of poaching, which implies that firms that do not invest in training might hire apprentices from the training firm by *offering* slightly higher wages. Accordingly, several models investigate this hypothesis by exploring exogenous differences in the competition for dual training leavers. Indeed, if firms are able to enjoy some monopsony power arising from industry- and occupation-specific skill requirements, dispersed regional location of firms and lower product market competition, they are found to be more likely to engage in training (Lazear, 2003; Smits, 2007; Gersbach and Schmutzler, 2006).

Empirical Evidence

Identification Problems

Is VET as effective as general-based education? Which type of VET best prepares workers for the labor market? What are the returns to a year of VET? These are some of the questions that researchers have attempted to answer in respect of different countries, as discussed below. As will become apparent, there is no easy answer to these questions given variation both across and within countries across studies. While some such differences are explained by countries' institutional and cultural differences, as well as the available amount of information on workers, jobs, and labor market characteristics in the different data sets used, several identification problems within this literature are difficult to overcome.

Most of this literature compares the employment outcomes of VET students with an alternative group, namely general-based education students or school dropouts, in the same country after controlling for all observable characteristics available. Moreover, it is acknowledged that unobserved heterogeneity may still prevail, given that youths deciding to study VET may have different abilities, tastes, and preferences about work from those who choose an alternative education system or no education. If there are unobserved quality differences between both types of youths, results from cross-sectional studies will reflect an omitted variable bias. For instance, given that VET is frequently intended for youths with lower motivation and ability than those who pursue general-based education, noncausal estimates of the returns to vocational education relative to general-based education will be downward biased (Willis and Rosen, 1979; Tuma, 1994; McCormick et al., 1995). By contrast, the opposite is likely to be true when comparing students from vocational education to school dropouts. Furthermore, a related concern arises due to different occupations requiring a different mix of academic or practical skills. If youths self-select into different occupations based on their skills, evaluating the effectiveness of the different systems becomes a difficult task given that the employment patterns, payment structures, and union coverage in the occupations themselves may not be comparable.

Unable to exploit exogenous changes in the institutional setting, the majority of these country studies conduct descriptive analyses controlling for students' characteristics to capture the expected opportunity of the alternative forms of schooling, including grades and test scores achieved prior to accessing VET or remaining in general education, family background, and local economic conditions. Additional confounders include subjective statements of preferences for VET or academic studies (Bishop and Mane, 2004), subjective self-assessments of ability (Hotchkiss, 1993), and information concerning the vocational orientation of the school captured by full-time vocational teachers and the schooling choice of previous cohorts (Meer, 2007).

In addition to those problems aforementioned, one has to add measurement problems in studies comparing the effectiveness of two types of VET systems across countries. Indeed, cross-country studies analyzing the relative effectiveness of school-based VET and the dual system are frequently biased due to the covariation of other relevant institutional factors, the absence of a unified framework for defining the respective training options, and the difference in data collection and quality (Hoeckel, 2008). While studies exploit the coexistence of the two systems within countries to evaluate their relative effectiveness in an attempt to avoid this problem, nonetheless, as explained by Eichhorst et al. (2013), in most countries one system prevails over the other, with the reason behind this likely to be correlated with the labor market structure, thus adding yet another source of endogeneity.

One way to address the selection problem is to exploit some exogenous change that lengthens or shortens one educational system versus the other. For instance, several researchers have exploited institutional changes increasing the duration of general schooling in the vocational schooling tracks of those respective countries (Oosterbeek and Webbink, 2007; Pischke and von Wachter, 2008; Hall, 2012), whereas others have used an instrumental variable approach (Fersterer et al., 2008). Furthermore, an alternative way to address the endogeneity is to use propensity score matching to address the selection problem (Aedo and Nunes, 2001; Aedo and Pizarro, 2004).

Evaluating School-Based Vocational Training versus General-Based Education: Which One Works Best?

Evidence from the US indicates that, once selection into VET has been accounted for, its participants perform equally or better than high school graduates in terms of employment and wages, also suggesting that these returns have increased over time. For example, Hotchkiss (1993) investigates the short-run effects of vocational schooling on employment and wages of high school graduates in 1980, finding no returns to vocational schooling even after controlling for training-related occupation choice. In contrast, Bishop and Mane (2004) and Meer (2007) investigate the short- to medium-run effects of participating in technical schooling among high school graduates in the US during a longer time period. While the former find youths to greatly benefit from choosing the vocational schooling track both in terms of employment and wages, Meer (2007) finds positive yet small wage increases, with these differences explained by using a more elaborate selection equation. Furthermore, Mane (1999) assesses differences in returns to technical schooling between the 1970s and 1990s more systematically, confirming this positive trend during the 1980s and 1990s.

In contrast, evidence in the UK reveals that academic education leads to higher returns, with Robinson (1997) finding that this is partly explained by academic jobs providing access to better paid occupations. Presenting evidence that lower ability and family background are also important drivers behind lower levels of education in the UK, Dearden (1999) casts doubt on the causality of these earlier findings. Reassessing the earlier findings, Dearden et al. (2002) largely confirm that academic education leads to higher returns, while finding significant heterogeneity regarding the type of VET obtained. In particular, they find that the majority of vocational education programs increase earnings relative to no vocational qualification, especially for low-achieving school-leavers. However, one exception is the UK's National Vocational Qualification (NVQ2), a VET course often undertaken during employment, which seems to have a negative impact on the wages of individuals, and even relative to those with no other qualification (Dearden et al., 2004).

Findings from early studies in developing countries are mixed (see the review by Chung, 1995). For instance, while Psacharopoulos (1994) finds that the returns to vocational secondary education are lower than those from general education, using evidence from 24 countries, when combining data from Psacharopoulos (1994) and Psacharopoulos and Ng (1992) for Latin American and Caribbean countries, Bennell (1996) finds that the differences in both systems' rates of returns are relatively small. According to Bennell and Sergerstrom (1998), and Middleton and Adams (1993), the effectiveness of vocational skills depends on the local demand for labor in the private sector and individuals' ability to find employment in the occupation of training. While most recent studies addressing the selection into VET problem remain

mixed, they seem to indicate that the differences in results lie in the institutional settings (such as the flexibility of labor market regulations) and historical and cultural differences with respect to the organization of work processes (see Almeida et al., 2012; for a thorough literature review in developing countries). For instance, Kahyarara and Teal (2008) investigate the returns from vocational education tracks in Tanzania, finding that the continuously strong preference for academic education is driven by the extremely high returns to high levels of academic education. Furthermore, they also show that lower levels of academic education and vocational education have very similar returns, suggesting that the convex return to education is at least partially driven by the selection of highly educated individuals into large, high-paying firms. Moenjak and Worswick (2003) compare the returns to the two vocational and academic tracks at the upper secondary level in high schools in Thailand, where both tracks qualify individuals to enter higher tertiary education. Moreover, when controlling for positive self-selection using a Heckman selection correction, their results point to a higher return to vocational education. Newhouse and Suryadarma (2011) analyze the returns to vocational versus academic high school tracks in Indonesia, with their results on labor market returns suggesting a change in the returns to vocational education over time. They find significantly higher wage returns for older cohorts yet lower returns for younger cohorts, explained by the decreasing value of vocational skills in the increasingly service-oriented Indonesian economy. Tansel (1994, 1999) and Tunali (2003) provide evidence for the returns to vocational high school education in Turkey, suggesting that the vocational tracks outperform academic tracks in terms of wages for men, and employment probabilities for women.

The Effects of Postponing Tracking of Students into Vocational Schools

Some studies have measured the effects of postponing tracking or extending general human capital education by exploiting exogenous policy changes increasing the duration of general schooling in the vocational schooling tracks, and overall do not find large (if any) effects of such policies on students' wages or employment probabilities. Oosterbeek and Webbink (2007) investigate the increase in duration of the vocational schooling track in the Netherlands in 1975 for 1 year, with the additional year designated only to general schooling. Adopting a difference-in-difference strategy, they analyze the effect of the change on wages 20 years later, and do not find any effect. Similarly, Pischke and von Wachter (2008) exploit the gradual adoption of a 1 year increase in compulsory schooling in the lowest schooling track in Germany between the 1950s and 1970s to study its effects on long-term wages, and likewise find no effects of the policy. Most recently, Malamud and Pop-Eleches (2010) evaluate a Romanian reform that postponed the tracking of students into vocational and academic schools. Using a regression discontinuity design, they find no effects of this reform on university completion, labor market participation, or earnings. While these previous studies consider extensions of general schooling at the lower secondary vocational tracks, Hall (2012) assesses a policy change in Sweden in 1991 that increased the general education content of the vocational schooling at the upper secondary level by 1 year in duration, after which students were eligible to enroll in tertiary education. Exploiting random differences in time and the regional implementation of a policy pilot, Hall (2012) does not find any effects on subsequent study take up, nor any increase in the wages earned up to 16 years after the beginning of upper secondary school. However, he finds that low-achieving students are significantly more likely to drop out of upper secondary education.

Dual System versus School-Based System?

While studies that have exploited the coexistence of the two systems within countries to evaluate their relative effectiveness find that, once selection is accounted for, the dual system seems to offer an advantage of improving early labor market transitions (measured in a variety of ways), evidence seems to indicate that these benefits eventually fade away. For instance, studies by Winkelmann (1996) and more recently Parey (2009) using German data show that participation in the dual apprenticeship improves early labor market attachment and leads to a faster and more structured integration into the labor market, relative to school-based VET. However, these short-term benefits fade away as other education participants find a foothold in the labor market. Moreover, Parey (2009) does not find any significant wage differences across the two types of training.

Similar results on the performance of apprenticeship training versus school-based training are found in studies by Bonnal et al. (2002) for France and Plug and Groot (unpublished manuscript) for the Netherlands. Correcting for the negative selection of youths into the dual apprenticeship, both studies find that apprenticeships perform significantly better in integrating youths into their first employment relationship by the higher labor market relation of their training. However, again this advantage is found to fade over time and is not associated with higher wages. A recent study by Alet and Bonnal (2011) shows that young people integrated into the dual apprenticeship system rather than vocational schooling in France are more likely to successfully complete their final exam and undertake further education.

In countries where the dual apprenticeship is not considered a 'trainer of last resort' for lower skilled individuals, the apprenticeship system is also found to partially act as a mechanism to level the playing field for youths with low school performance. Exploiting information on Programme for International Student Assessment (PISA) test scores to capture differences in ability across students, Bertschy et al. (2009) show that lower performing youths in Switzerland tend to select into less demanding apprenticeship occupations and are thereby penalized in the labor market; however, their initial test score (as a signal of their ability) is no longer important in determining labor market outcomes.

Estimating the Returns to VET Programs

Rather than comparing two systems within a country, several studies have analyzed the returns to apprenticeship training in a particular country. However, selection into this vocational training system and the different types of apprenticeship once

again raises concerns that Ordinary Least Squares (OLS) wage estimates will be biased (Soskice, 1994). In particular, Soskice finds that much heterogeneity is due to firm size, given that the wages for apprenticeship-trained workers increase strongly in the size of the training firm. Several papers have found that the wage returns to apprenticeship training on wages in Germany and Austria range between 15 and 20%, based on OLS estimates (see Krueger and Pischke, 1995; Winkelmann, 1996; Fersterer and Winter-Ebmer, 2003). Given that an apprenticeship lasts around 3 years on average, this implies a return of around 5% a year, which is not far from other forms of school-based education, in colleges or vocational schools. One of the first studies to use an instrumental variables approach to measure the returns of apprenticeships is that of Fersterer et al. (2008), using Austrian data from 1975 to 1998. These authors exploit the different length of apprenticeship period completed for some apprentices in failed firms, and perhaps surprisingly find that the estimated returns for apprentices affected by the firm failure are not very different from the OLS returns in the same sample, suggesting that selection problem is not particularly important in this case.

The Life-Cycle Perspective

A recent line of research has focused on studying the life-cycle impact of vocational education, motivated by the returns to vocational training potentially being quite different at labor market entry than after several years into the labor market, given that skills have to adapt to technological change and be mobile across firms, occupations, and space. Hanushek et al. (2011) use data on the labor market experiences of people at different ages and in 18 different Organisation for Economic Co-operation and Development (OECD) countries, collected in the mid-1990s as part of an OECD-sponsored venture, to compare the life-cycle employment patterns of people with different educational backgrounds. They find a higher initial employment rate for vocational education participants at labor entry, which is reversed by the age of 50, and these results are suggestive of the hypothesis that occupation-specific knowledge outdates faster and thus leads to lower employment opportunities later in life. At the same time, they also find that this pattern is more significant in 'apprenticeship countries' with combined school and work-based education programs (Denmark, Germany, and Switzerland in our data), as opposed to countries with no formal system of vocational education, such as the US. However, more reliable evidence is required concerning the perceived trade-off, given that the causal interpretation of these findings is impeded by occupation-specific segregation of the labor market and the limited availability of long-term panel data.

Alternatively, another recent study by Adda et al. (2011) models the entire career path, starting with the original apprenticeship choice and followed by the period-by-period employment transitions, job mobility, and wages. A structural approach is used to take the dynamics over the life cycle into account, since the two career paths may differ in job attachment, available job opportunities, and wage growth. Using 15 years of German data, they find that apprenticeships lead to more wage growth up front, while wages in the pure on-the-job training (unskilled) grow at a lower rate yet for longer. Overall, they find that wages are higher following an apprenticeship qualification, with the job arrival rates very high and destruction rates very low after some years of experience. These findings contrast with Heckman's suggestions that qualified apprentices are harder to reallocate following a job loss (Heckman, 1993).

What Is the Evidence on the Effectiveness of Informal VET?

Evidence on the effectiveness of informal training is clearly scarce, given that it mostly pertains to developing countries, and also owing to the nature of the training per se, which relies on the informality of both the labor market and training. Nonetheless, the few identified studies inform on the selection problem observed for the other types of training, and seem to indicate that informal training improves trainees' labor market opportunities. However, further research on evaluating traditional and informal apprenticeships is needed.

Compared to informal jobs without training, the informal types of VET are considered to improve the chance of achieving promotion inside informal businesses from unskilled to skilled worker, consequently earning a higher wage. Accordingly, training is essential to become a master craftsman and set up or take over a business in these fields after some years of activity (Aggarwal et al., 2010; Nübler et al., 2009). Furthermore, compared to vocational schooling, which often relies on outdated curricula and lacks the involvement of employers in many developing countries, informal and traditional apprenticeships are coupled with current skill demands and work practices in the informal sector, with informal apprenticeship graduates more easily employed than those from vocational schools. Given that they do not require any formalized entry criteria, informal training often represents the only opportunity for school-leavers to obtain vocational education (Walther, 2011).

In terms of labor market outcomes, evidence from studies on Malawi and Tanzania suggests that most graduates from informal apprenticeships were employed with the same business that had provided training, or with some other employer, with a considerable share of them also being self-employed after some years (Aggarwal et al., 2010; Nübler et al., 2009). Furthermore, an evaluation study of the informal apprenticeship system in Ghana shows an important negative selection of youths into informal apprenticeships, involving youths with low levels of general schooling entering this type of training (Monk et al., 2008). However, when controlling for this negative selection, it is found that individuals with low levels of formal schooling substantially benefit from the informal training, with earning increases of 50%.

Conclusions and Policy Implications

VET represents a valued alternative beyond the core of general education, with the dual system seeming to be more effective at facilitating labor market entry than school-based VET, although this advantage seems to fade over time. While informal training

is effective, not much is known concerning its relative strengths compared with other forms of vocational education.

As explained by Eichhorst et al. (2013) strengthening the vocational part of the educational or schooling system and bringing existing VET systems closer to the current needs of the labor market would help young people with a smoother transition into work in many countries. In particular, vocational education provided in the framework of secondary schooling (vocational schools or vocational tracks) should be modernized and complemented with phases of practical work experience, such as internships or passing the final year with an employer. Employers should also be consulted regarding the design of vocational schooling curricula, which requires a systematic coordination with networks or associations of employers. Furthermore, in order to avoid a negative perception of vocational education as a dead-end option, transition to further education, including tertiary education, should be facilitated. Finally, reducing vocational education fees can help to raise enrollment in some countries. Moreover, in countries with high shares of university graduates facing major difficulties in finding adequate jobs, one significant option is to make academic training more labor market-oriented, incorporating internships with employers into academic curricula to enable the attainment of some experience with current work practices in the private sector. Governments responsible for funding academic education can require public universities to modify academic curricula accordingly.

Providing better training for the informal sector represents a core issue for many developing countries. In countries where traditional or informal apprenticeships are dominant yet mainly confined to traditional crafts, they should be better articulated with the schooling system and formal sector, and also open up to new technologies and occupational change, which requires some recognition of informal employment as part of the economic and social reality in many countries. One first option is to bring societal initiatives aimed at better training closer to the informal sector, family business, and local networks. Accordingly, a concrete step could involve encouraging informal workers and employers to participate in training activities, such as providing informal apprentices with some vocational schooling focusing on more general skills and theoretical aspects. The participation in vocational courses of young people working in the informal sector could be increased by providing incentives to both participants and employers, particularly compensating for hours not worked due to training courses. Moreover, to avoid deterrence, these courses should not be delivered directly by governments, but rather by NGOs, churches, or nonprofit associations with sufficient acceptance and in-depth knowledge of the economic situation in local communities. Involving larger employers or (formal) training centers represents another option, with funding potentially coming from governments and international donors. Furthermore, NGOs, churches, or other nonprofit associations can also facilitate the creation of (informal) associations or networks of informal employers.

Given that traditional or informal apprenticeships tend to be restricted to a number of traditional crafts, raising productivity and potentials for innovation is crucial. Experiences from the African continent show that master craftsmen benefit from skill-upgrading courses to better develop their businesses and become more innovative and productive, and also from better access to technical equipment and capital, which should be made more easily accessible to informal firms. Furthermore, some experiences from sub-Saharan Africa show that ensuring skill recognition outside the local community through some sort of official skill testing open to informal apprenticeship graduates raises the attractiveness of such training courses and enhances mobility on the job market.

See also: Apprenticeship, Anthropological Aspects; Apprenticeships and School Learning: General Considerations; Economic Education; Vocational Education; Vocational Interests, Values, and Preferences, Psychology of.

Bibliography

Acemoglu, Daron, Pischke, Jorn-Steffen, 1998. Why do firms train? Theory and evidence. The Quarterly Journal of Economics 113 (1), 78–118.

Acemoglu, Daron, Pischke, Jorn-Steffen, 2000. Certification of training and training outcomes. European Economic Review 44 (4–6), 917–927.

Adda, Jerome, Dustmann, Christian, Meghir, Costas, Robin, Jean-Marc, 2011. Career Progression and Formal versus On-the-job Training. Mimeo.

Aedo, Cristian, Nunes, S., May 2001. The Impact of Training Policies in Latin America and the Caribbean: The Case of ProgramaJoven. ILADES and Georgetown University.

Aedo, Cristian, Pizarro, M., 2004. Rentabilidad economica del programa de capacitacion laboral de jovenes Chile Joven. Mimeo.

Aggarwal, A., Hofmann, C., Phiri, A., 2010. A Study on Informal Apprenticeship in Malawi. Employment Sector Employment Report 9/2010.

Alet, E., Bonnal, Liliane, 2011. Vocational Schooling and Educational Success: Comparing Apprenticeship to Full-time Vocational High-school. Mimeo.

Almeida, Rita, Richard Behrman, Jere, Robalino, David Alejandro, 2012. The Right Skills for the Job?: Rethinking Training Policies for Workers. World Bank Publications, World Bank, Washington, DC.

Arulampalam, Wiji, Booth, Alison L., Bryan, Mark L., 2004. Training in Europe. Journal of the European Economic Association 2, 346–360.

Bassanini, Andrea, Booth, Alison L., Brunello, Giorgio, De Paola, Maria, Leuven, Edwin, 2005. Workplace Training in Europe. IZA Discussion Papers 1640. Institute for the Study of Labor (IZA).

Becker, Gary, 1964. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press, New York.

Bennell, P., 1996. General versus vocational secondary education in developing countries: a review of the rates of return evidence. The Journal of Development Studies 33 (2), 230–247.

Bennell, P., Sergerstrom, J., 1998. Vocational education and training in developing countries: has the World Bank got it right? International Journal of Educational Development 18 (4), 271–287.

Bertschy, K., Cattaneo, Alejandra, Wolter, Stefan C., 2009. PISA and the transition into the labour market. Labour 23 (Special Issue), 111–137.

Bishop, J.H., Mane, Ferran, 2004. The impacts of career-technical education on high school labor market success. Economics of Education Review 23 (4), 381–402.

Bonnal, Liliane, Mendes, S., Sofer, C., 2002. School-to-work transition: apprenticeship versus vocational school in France. International Journal of Manpower 23 (5), 426–442.

CEDEFOP, 2008. Initial Vocational Education and Training (IVET) in Europe: Review. European Centre for the Development of VET (Cedefop).

Carrero Perez, E., Hakim, G., 2006. Reforming Technical and Vocational Education and Training in the Middle East and North Africa: Experiences and Challenges. European Training Foundation, Luxembourg.

Chung, Y.-P., 1995. Returns to vocational education in developing nations. In: Carnoy, M. (Ed.), International Encyclopaedia of Economics of Education. Pergamon, Oxford, pp. 175–181.

Dearden, Lorraine, 1999. Qualifications and Earnings in Britain: How Reliable Are Conventional OLS Estimates of the Returns to Education?. IFS Working Papers W99/07 Institute for Fiscal Studies.

Dearden, Lorraine, Steven, McIntosh, Michal, Myck, Anna, Vignoles, 2002. The returns to academic and vocational qualifications in Britain. Bulletin of Economic Research 54 (3), 249–274.

Dearden, Lorraine, McGranahan, Leslie, Sianesi, Barbara, 2004. An In-depth Analysis of the Returns to National Vocational qualifications Obtained at Level 2. CEE Discussion Papers 0046. Centre for the Economics of Education.

Dionisius, R., Mühlemann, Samuel, Pfeifer, H., Walden, G., Wenzelmann, F., Wolter, Stefan C., 2009. Costs and benefits of apprenticeship training. A comparison of Germany and Switzerland. Applied Economics Quarterly 55 (1), 5–38.

Dustmann, Christian, Schönberg, Uta, 2012. What makes firm-based VET schemes successful? The role of commitment. American Economic Journal: Applied Economics 4 (2), 36–61.

Eichhorst, Werner, Rodríguez-Planas, Núria, Schmidl, Ricarda, Zimmermann, Klaus F., forthcoming. A roadmap to vocational education and training systems in industrialized countries. Industrial Labor Relations Review.

Eichhorst, Werner, Schmidl, Ricarda, Zimmermann, Klaus F., January 2013. A Roadmap to Vocational Education and Training Systems Around the World. Discussion paper No. 7110. IZA.

Fersterer, Josef, Pischke, Jorn-Steffen, Winter-Ebmer, Rudolf, 2008. Returns to apprenticeship training in Austria: evidence from failed firms. Scandinavian Journal of Economics 110 (4), 733–753.

Fersterer, Josef, Winter-Ebmer, Rudolf, 2003. Are Austrian returns to education falling over time? Labour Economics 10 (1), 73–89. http://ideas.repec.org/a/eee/labeco/v10y2003i1p73-89.html. http://ideas.repec.org/s/eee/labeco.html.

Gersbach, Hans, Schmutzler, Armin, 2006. Does Product-market Competition Decrease Employers' Training Investment? Evidence from German Establishment Panel Data. Ruhr Economic Paper No. 41.

Hall, C., 2012. The effects of reducing tracking in upper secondary school: evidence from a large-scale pilot scheme. Journal of Human Resources 47 (1), 237–269.

Hanushek, Erik A., Woessmann, Ludger, Zhang, Lei, 2011. General Education, Vocational Education and Labormarket Outcomes over the Life-cycle. IZA. Discussion Paper 6083.

Heckman, James J., 1993. Assessing Clinton's Program on Job Training, Workfare, and Education in the Workplace. National Bureau of Economic Research. Working Paper, Nr. 4428.

Hoeckel, K., 2008. Costs and Benefits in Vocational Education and Training. OECD Education, Paris.

Hotchkiss, Julie L., 1993. Effects of training, occupation, and training-occupation match on wage. Journal of Human Resources 28 (3), 482–496.

ILO, 2011. Upgrading Informal Apprenticeship Systems. ILO (Policy Brief).

ILO, 2012. Upgrading Informal Apprenticeship, a Resource Guide for Africa. ILO.

Kahyarara, Godius, Teal, Francis John, 2008. The returns to VET and academic education: evidence from Tanzania. World Development 36 (11), 2223–2242. Elsevier.

Katz, Eliakim, Ziderman, Adrian, 1990. Investment in general training: the role of information and labour mobility. Economic Journal 100 (403), 1147–1158.

King, K., 2011. Eight Proposals for a Strengthened Focus on Technical and Vocational Education and Training (TVET) in the Education for All (EFA) Agenda. Background paper prepared for the Education for All Global Monitoring Report 2012.

King, K., McGrath, S., Rose, P., 2007. Beyond basic education and towards an expanded vision of education – for poverty reduction and growth? International Journal of Educational Development 27 (4), 349–357.

Kogan, I., 2008. Education systems of Central and Eastern European countries. In: Kogan, Irina, Gebel, Michael, Noelke, Clemens (Eds.), Europe enlarged. A Handbook of Education, Labour and Welfare Regimes in Central and Eastern Europe. The Policy Press, Bristol.

Krueger, Alan B., Pischke, Jörn-Steffen, 1995. A comparative analysis of East and West German labor markets. In: Katz, Lawrence F., Freeman, Richard B. (Eds.), Differences and Changes in Wage Structures. University of Chicago Press, Chicago, pp. 405–446 before and after unification. http://ideas.repec.org/h/nbr/nberch/7864.html.

Lazear, Edward, 2003. Firm-specific human capital: a skill-weights approach. NBER Working Paper No. W9679.

Malamud, Ofer, Pop-Eleches, Cristian, February 2010. General education versus vocational training: evidence from an economy in transition. The Review of Economics and Statistics 92 (1), 43–60. MIT Press.

Malcomson, James M., Maw, James, McCormick, Barry, 2003. General training by firms, apprentice contracts, and public policy. European Economic Review 47 (2), 197–227.

Mane, Ferran, 1999. Trends in the payoff to academic and occupation-specific skills: the short and medium run returns to academic and vocational high school courses for non-college-bound students. Economics of Education Review 18 (4), 417–437.

McCormick, A., Tuma, J., Houser, J., 1995. Vocational Course Taking and Achievement. Statistical Analysis Report. National Center for Education Statistics, Washington, DC.

Meer, Jonathan, 2007. Evidence on the returns to secondary vocational education. Economics of Education Review 26 (5), 559–573.

Middleton, Ziderman, Adams, van, 1993. Skills for Productivity: Vocational Education and Training in Developing Countries. Oxford University Press, New York.

Moenjak, T., Worswick, Christopher, 2003. Vocational education in Thailand: a study of choice and returns. Economics of Education Review 22, 99–107.

Monk, C., Sandefur, J., Teal, F.T., 2008. Does Doing an Apprenticeship Pay off? Evidence from Ghana. CSAE Working Paper Series 2008–08. Centre for the Study of African Economies, University of Oxford.

Neuman, S., Ziderman, Adrian, 1999. Vocational education in Israel: wage effects of the vocational-education occupation match. The Journal of Human Resources 34 (2), 407–420.

Newhouse, David Locke, Suryadarma, Daniel, 2011. The value of vocational education: high school type and labor market outcomes in Indonesia. World Bank Economic Review 25 (2), 296–322. World Bank Group.

Nübler, I., Hofmann, Christian, Greiner, Clemens, 2009. Understanding Informal Apprenticeship: Findings from Empirical Research in Tanzania, Employment Sector Employment. Working Paper 32/2009. ILO.

Oosterbeek, Hessel, Webbink, Dinand, 2007. Wage effects of an extra year of basic vocational education. Economics of Education Review 26 (4), 408–419.

Parey, M., 2009. Vocational Schooling versus Apprenticeship Training- Evidence from Vacancy Data. Mimeo.

Pischke, Jorn-Steffen, von Wachter, Till, 2008. Zero returns to compulsory schooling in Germany: evidence and interpretation. The Review of Economics and Statistics 90 (3), 592–598.

Plug, Erik, Groot, W., Apprenticeship versus Vocational Education: Exemplified by the Dutch Situation, unpublished manuscript.

Psacharopoulos, George, 1994. Returns to investment in education: a global update. World Development 22 (9), 1325–1343. Elsevier.

Psacharopoulos, George, Ng, Ying Chu, 1992. Earnings and Education in Latin America: Assessing Priorities for Schooling Investments. World Bank, Washington, DC.

Quintini, G., Martin, S., 2006. Starting Well or Losing Their Way? the Position of Youth in the Labour Market in OECD Countries, 39. OECD social, employment and migration working papers.

Robinson, Peter, 1997. The Myth of Parity of Esteem: Earnings and Qualifications. Discussion Paper 354. London School of Economics: Centre for Economic Performance.

Shavit, Y., Müller, Wieland (Eds.), 1998. From School to Work: A Comparative Study of Educational Qualifications and Occupational Destinations. Clarendon Press, Oxford.

Shavit, Y., Müller, Wieland, 2000. Vocational secondary education. European Societies 2 (1), 29–50.

Smits, Wendy, 2007. Industry-specific or generic skills? Conflicting interest of firms and workers. Labour Economics 14, 653–663.

Soskice, David, 1994. The German training system: reconciling markets and institutions. In: Lynch, Lisa (Ed.), International Comparisons of Private Sector Training. University of Chicago Press. NBER Conference Volume, 1994.

Tansel, Aysit, 1994. Wage employment, earnings and returns to schooling for men and women in Turkey. Economics of Education Review 13 (4), 305–320.

Tansel, Aysit, 1999. General versus Vocational High Schools and Labor market Outcomes In Turkey. Economic Research Forum Working Paper No. 9905.

Tuma, J., 1994. Measuring enrollment and participation in secondary vocational education with high school transcript records. Journal of Vocational Education Research 19 (3), 3–22.

Tunali, Insan, 2003. General vs. Vocational Secondary School Choice and Labor market Outcomes in Turkey, 1988–98. Mimeo.

Walther, Richard, 2011. Building Skills in the Informal Sector. Background Paper Prepared for the Education for All Global Monitoring Report 2012. UNESCO.

Willis, Robert, Rosen, S., 1979. Education and self-selection. The Journal of Political Economy 87 (5), 7–36.

Winkelmann, Rainer, 1996. Employment prospects and skill acquisition of apprenticeship – trained workers in Germany. Industrial and Labour Relations Review 49 (4), 658–672.

Woessmann, Ludger, 2008. Efficiency and equity of European education and training policies. International Tax and Public Finance 15, 199–230.

Wolter, S., Ryan, P., 2011. Apprenticeship, handbook of economics of education. In: Hanushek, Eric (Ed.), Stephen Machin and Ludger Woessmann, vol. 3 (3). Elsevier.

Zimmermann, Klaus, F., Biavaschi, Costanza, Eichhorst, Werner, Giulietti, Corrado, Kendzia, Michael J., Muravyev, Alexander, Pieters, Janneke, Rodríguez-Planas, Núria, Schmidl, Ricarda, 2013. Youth unemployment and vocational training. Foundations and Trends in Microeconomics 9 (1–2), 1–157.

Voluntary Associations, Sociology of

David Horton Smith, Boston College, Retired in Bradenton, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Voluntary organizations have two main subtypes: voluntary associations and voluntary agencies. This article deals with voluntary associations, whereas a different article entitled 'Voluntary Organizations' deals with voluntary agencies. This article reviews association governance and leadership, history, prevalence rates and their determinants, purposive, and analytical types of associations, geographic scope, degree of paid staff dependence, targets of benefits (recipients), types of members, formalization and government recognition, misconduct and deviance in and by associations, government control issues, internal democracy, monomorphic and polymorphic associations, collaboration, intensity of activity, personal contact and relationships, membership requirements, prestige and power, and degrees and types of impact.

According to *Merriam-Webster's Unabridged Dictionary*, the word 'voluntary' means an action of a person's (free) will or choice. Voluntary organizations (VOs) fit with this definition by being organizations that people join and participate in by their own free choice, not because of compulsion by law or force. By common definitions (e.g., Smith et al., 2006), VOs are distinctive in having not-for-profit (nonprofit) primary goals, as contrasted with business organizations (businesses), which have for-profit primary goals. There are various other common names for VOs, the most frequent being nonprofit organizations (NPOs), nonprofits, nongovernmental organizations (NGOs), and civil society organizations (CSOs).

There are two main structural types of VOs: associations and nonassociational nonprofits (Smith, 2000: p. 30). Nonassociational nonprofits, usually called voluntary agencies or nonprofit agencies, are treated in a different article in this *Encyclopedia*, under the title, voluntary organizations. The present article discusses voluntary associations, often simply called associations (or sometimes membership associations).

Many people confuse voluntary associations and voluntary agencies, or incorrectly use the term voluntary organizations to refer just to *one* type or the other, not to *both* as is correct. Others similarly fail to understand how different the two basic VO types are in their basic structure. Simply put, voluntary associations have a structure in which there are members who usually hold the ultimate power, whereas voluntary agencies usually have a board of directors (or trustees) that holds the ultimate power. Voluntary agencies (sometimes called social agencies or welfare agencies) have no members, but they do have paid staff and often volunteers working in volunteer programs as VO departments. Quarter et al. (2001) have shown empirically that membership associations are distinctive in many important ways from nonmember-serving voluntary agencies.

In an association, members usually elect the top leaders (board members, officers) on a regular basis (e.g., annually, biennially, etc.). In a voluntary agency, the board of directors (trustees) is usually self-perpetuating, electing, or appointing new members as needed, sometimes annually or biennially. There is no category of members as a kind of agency affiliate, just employees and often volunteers, as noted. In sum, voluntary associations usually practice a kind of internal democracy, while voluntary agencies rarely if ever practice internal democracy. Voluntary agencies generally are structured more like government agencies or business corporations, as hierarchical bureaucracies with the power at the top. This crucial difference in basic internal structure makes these two subtypes of VOs very different in their operations, organizational culture, resource attraction processes, and impact on participants, among many other effects (Smith, 2000: Part 2). Smith has argued that these two VO subtypes constitute different subsectors of the voluntary nonprofit sector (VNPS), third sector, or civil society sector (Smith, 1991).

Research and Theory about Associations

Research on associations, in the form of historical accounts, goes back at least three millennia, perhaps four to five (Kloppenborg and Wilson, 1996; Smith, 1997). Quantitative empirical research by social scientists began very recently, in the early middle of the twentieth century (Smith et al., 2014: Chapter 1; Smith and Associates, 2015). *Organized* research on associations and other VOs began in 1971, with the founding of the Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA) and its journal, *Nonprofit and Voluntary Sector Quarterly* (Smith et al., 2014: Chapter 1). Many similar associations of researchers and additional academic journals on VOs and on related phenomena of the VNPS subsequently have been founded (Smith, 2013). Another global and interdisciplinary association of researchers was founded in 1992, named the International Society for Third Sector Research (ISTR; Smith, 2013).

The website (ICSERA) of the International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA) lists and updates information on all known researcher associations, academic journals, and research centers in the field (and emerging academic discipline) worldwide. ICSERA is an umbrella association of other researcher associations studying the VNPS, mainly national associations. ICSERA seeks mainly to help build the organized, global research field (and emerging academic discipline) of VNPS research, now termed 'voluntaristics' or 'altruistics' by Smith (2013).

The most comprehensive book summarizing recent research and theory on associations is the *Palgrave Research Handbook*

of Volunteering and Nonprofit Associations (Smith et al., 2014). Contributors from more than 70 nations and all world regions review and assess research on most relevant topics regarding associations and volunteering in associations in 61 chapters and an appendix, including chapters on theory and methodology for studying associations and volunteering in associations. There is no widely accepted general theory of associations, although Smith (2000) has presented an extensive theory of grassroots associations (GAs; see Smith, 2010a). Some middle-level theories are focused on individual volunteering in associations (Smith et al., 2014: Part IV), association prevalence (Smith et al., 2014: Chapter 36), and deviant voluntary associations (DVAs; Smith et al., 2014: Chapter 59).

The History of Associations

For most of the period of existence of our human species (about 150 000–200 000 years), there were no VOs. All societies were small, nomadic, hunting-gathering bands of about 40–70 people, which had no subgroups other than nuclear families of parents and children. The earliest VOs were voluntary associations, originating perhaps 25 000 years ago. Such associations became common about 10 000 years ago, when most human societies had settled down in agricultural villages that had an average of 1500 people, so that subgroups now made sense (Smith, 1997).

For several millennia, associations were mainly social clubs, especially for adult males (Smith, 1997). In ancient agrarian civilizations like Greece and Rome, occupational associations as guilds were added to social clubs as a type of association, as were religious associations (Kloppenborg and Wilson, 1996). Early Christian churches in the Roman Empire were examples of the latter, as were much earlier Egyptian and Greek religious cults (Boulding, 1953; Smith, 1997).

In the second millennium AD/CE, many new types of associations arose, especially in the past two centuries as a result of the Industrial Revolution (Boulding, 1953; Smith, 1997). There has also been a tremendous increase in the numbers of associations in this recent 200-year period – an organizational revolution and especially an associational revolution (De Tocqueville, 1976; Smith, 1997, 2010a).

The Prevalence of Associations and Its Determinants

If the prevalence (frequency in a territory) of associations is plotted on a historical timeline for the human species, there were none until about 25 000 BC and then significant numbers from about 10 000 BC. This rise of associations resulted from the horticultural revolution, in which much smaller, nomadic, hunting–gathering tribes settled down in still small, but significantly larger (estimated 1500 persons), villages to crudely raise crops and domestic animals (Smith, 1997).

When full-scale agriculture was later discovered *c*.3000 BC in Egypt and China, later in Greece, Rome, and Mesopotamia from 1000 BC to the turn of the millennium, economic surpluses permitted large cities and greater differentiation of occupations beyond farming. This much greater division of labor, especially in cities, created a corresponding differentiation of individual and family interests or goals, leading to the formation of guilds for different occupations as economic associations.

Much later, the Industrial Revolution, beginning about 1800 in the West, gave a tremendous impetus to association prevalence (Smith, 1997). In terms of the noted graph (Smith, et al., 2014: Chapter 36), association prevalence has leapt upward exponentially on a global scale since about 1800. Boulding (1953) has referred to this change as the organizational revolution. More recently, other scholars have referred to it as an associational revolution (Smith, 1997, 2010a). These terms highlight the large and rapid increase in association prevalence in the past two centuries plus, especially in all nontotalitarian nations undergoing modernization. The key changes in social structure that foster rapid growth in association prevalence are industrialization, development of mass education systems, expansion of the mass media, growing democratic participation and civil liberties, and construction of an extensive transportation–communication infrastructure.

Not surprisingly, the absolute (raw) number of associations in a nation or other territory is strongly related to the sheer population size in the area studied. Recent research on the nations of the world, however, has shown that the prevalence of associations per capita (or per 1000 population) among the nations of the world can be explained successfully (89% of the variance) by such additional variables as higher gross domestic product per capita, higher average level of formal education, greater presence of civil liberties (especially freedoms of association, assembly, religion, speech, and mass media), longer national experience with democracy, greater number or density of nonassociational organizations (businesses and government agencies) per capita, and greater resource mobilization fostering associations, such as information, consulting support, and research infrastructure organizations (Smith and Shen, 2002; Smith et al., 2014: Chapter 36).

Purposive versus Analytical Types of Associations

Many scholars and laypeople tend to view the primary or official purpose of an association as its most crucial characteristic. Research does *not* support this view. Far more important are the analytical characteristics of an association, such as the geographic scope of its activity, use of paid staff if any, degree of formalization (bureaucratization), types of members (individual persons as members versus organizations versus both), the target of benefits (members versus nonmembers or both), monomorphic versus polymorphic form, and so on (Smith, 2000). This article is divided into many small sections based on a variety of important analytical types or categories, not based on simplistic purposive types of associations (Smith, 2000: pp. 4–5, 232–233).

The Geographic Scope of Association Activities

Associations exist with the widest range of geographic scope, from truly global international associations that have members from nearly 200 nations and territories, down to GAs having members only from a single apartment building or condominium. With such variations in geographic scope come many

other variations in association structures and processes. For instance, international and national associations tend to be much more formalized than GAs, larger in membership size, meet much less frequently in person (annually or less frequently in conferences, if at all), depend more on electronic (especially Internet) communications, have a longer average life span, have their own rented or owned headquarters buildings and offices, have greater income and assets, collaborate more with other organizations, are better known to the general public, and tend to have more power or influence in their societies (Smith, 2010b; Smith et al., 2014: Chapters 38 and 39).

Degree of Dependence on Paid Staff

Because the vast majority of associations everywhere are GAs, they usually have no paid staff at all, even part time. Associations with larger memberships and at supralocal levels of geographic scope (state/provincial, national, and international associations), however, usually have some paid staff, even if only part-time. Very large associations, such as the AARP (American Association of Retired Persons) in America with nearly 40 million members, have hundreds of paid staff. Thus, there are really two distinct analytical types of associations in terms of their operations: The large majority of associations are led and run by volunteers, as amateurs in informal groups. A small minority of associations has one or more full-time paid staff in more formalized associations, with the paid staff doing much or most of the work of the association. In associations with paid staff, the internal dynamics of the association are changed. Paid staff, especially the top executives, have much more influence than members, and often more than the elected volunteer officers of the association. Member passivity and apathy usually contribute to this tendency for dominance by paid staff.

Target of Benefits

The 'target of benefits' of a group is the set of the people or situations that the group intends to help, protect, or improve in some way. Sometimes associations have clear recipients, but not always. Most associations have their members as the main target of benefits, hence being termed 'member-benefit' groups. Nearly all GAs are member benefit associations.

Some associations, especially national or international associations, have specific nonmembers or the general public as the target of benefits (intended recipients). Nonmember benefit associations may also benefit no specific recipients, as with environmental associations and some political or social movement associations. Also, some associations serve both members and nonmembers. There is altruism or serving others involved in both member benefit and nonmember benefit associations. But in member benefit associations, the altruism is limited mainly to fellow members of one's own association.

Examples of member benefit associations include religious congregations, occupation-focused associations (labor unions, farmers' associations, professional associations, business people's groups, trade associations), self-help groups (Alcoholics Anonymous groups, cancer survivors' groups), social clubs, and indulgent leisure groups (e.g., chess clubs, amateur baseball or soccer clubs, transvestite groups). Some examples of nonmember-benefit associations are political parties or associations, social service associations, social movement organizations, and environmental groups.

This internal versus external focus of associations significantly affects group structure and operations. For instance, member benefit associations depend mainly on member dues (required annual payments by all members) or voluntary member donations for their income. By contrast, nonmember benefit associations receive more donations from the general public, fees for services, and sometimes receive grants from foundations.

Individual versus Organizational Members

Associations may have individual persons as members or organizations as members (collective members), or both. Most associations emphasize one or the other type of members, not both. Associations with individual persons as members greatly predominate at all geographic levels of scope, especially at the local level, as GAs. As examples, the Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA) has mainly individual persons as members. By contrast, the International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA) has associations of researchers as collective members. And the International Sociological Association (ISA, http://www.isa-sociology.org) has both 45 national sociological associations as collective members and also about 4500 persons as individual members.

Degree of Legal Formalization and Government Recognition

Associations vary markedly in their degrees of legal formalization and recognition. In most societies, the vast majority of associations, as GAs, have no legal formalization at all. Most GAs are thus informal groups. For instance, in the United States and China, there are many millions of unregistered and unincorporated GAs (Smith, 2000: Chapter 2; Wang, 2011). Nonprofit agencies are much more likely to seek and obtain legal and government recognition than are associations, especially GAs.

In most nations, the law requires larger associations in terms of income, assets, or membership to register with the government. Such registration may lead to or provide legal tax exemption for net revenues. Net revenues are 'surplus revenues' after paying expenses, because there are technically no 'profits' in nonprofit groups. In addition, most nations provide for legal incorporation of an association, so that it can become a 'legal person,' able to have its own bank account, hire employees, own real estate, and so on. In some nations (such as China), the legal incorporation and registration processes are merged, handled by a single government agency.

Degree of Internal Formalization of Association Structure

As small, local, volunteer-based associations, most GAs have loose and informal internal structures. Boards of directors are

rare in GAs, although there usually is one or more key officers (like president, vice president(s), treasurer, secretary) elected by the members annually or less often. The group boundaries of GAs may be vague or loose, so membership is not always clear. In some GAs, whoever attends a given event or meeting is a 'member.' Most GAs keep a formal list of members, however, seeking donations or dues from each one if the person is to remain a member in good standing (official member). In larger GAs and supralocal associations ('above the local level' – at the state/provincial, national, or international level of geographic scope), membership lists are usual. In addition, committees (subgroups of members appointed by the leaders) are charged with specific functions, like arranging events, recruiting members, and generating publicity.

Larger GAs and especially supralocal associations tend to have more internal bureaucracy, usually with three or more levels of structure (e.g., members, committees or departments, key officers, board of directors, the top leader as chief executive officer). These formalized associations tend to have a written constitution, bylaws, or other documents stating explicitly how the association will operate. Such increased internal formalization is more frequent in associations with a paid staff, especially a larger paid staff. Some nations (e.g., China) have legal requirements for significant internal structure and assets before an association can gain legal recognition from the government.

Conventional versus Deviant Goals and Means

Most scholars and the general public think only about conventional, mainstream, socioculturally acceptable associations, ignoring the dark side of associations and of the VNPS. But a small minority of associations in any contemporary society (and in most prior societies) has one or more goals, or a means of achieving their goals, that break the laws or moral rules of their society at the time. These are termed deviant nonprofit groups (DNGs; Smith et al., 2006, 2014: Chapter 59). Volunteers usually run associations, especially GAs, and hence GA leaders and members have no association-related paid jobs to lose. As a result, associations, not voluntary agencies, are the more common kind of DNG, sometimes being willing to take great risks collectively and intentionally. Associations that are DNGs may be termed 'Deviant Voluntary Associations' or 'DVAs.' Voluntary agencies, especially foundations, very rarely take substantial risks by any persisting deviance. When deviance is discovered or made public in a voluntary agency, the people involved usually are caught and punished, at the very least by being fired.

Some common types of DVAs include (1) revolutionary, terrorist, and guerrilla groups as violent political DVAs; (2) doomsday cults (ones that commit mass suicide, like the Peoples' Temple), witches' covens, and polygamous sects as religious DVAs; (3) nudist groups, group marriage, and free-love communes as sexual DVAs; and (4) aggressive hate groups (Al Qaeda, the Ku Klux Klan, Skinheads) as violent DVAs. Some DVAs, like the Nazi Party, fall into more than one of these categories. These DVAs range widely in how serious their deviance is. Nudist groups are harmless, but the Nazi Party in the Third Reich, initially just a DVA with revolutionary goals (1921–33), was directly responsible for the deaths of tens of millions of people in the holocaust and World War II (1933–45).

Because of their societal rule-breaking, DVAs usually have to operate secretly or 'underground' to a significant degree. DVAs carefully avoid legal formalization and government recognition, which would lead to greater government control and often dissolution of the association. DVAs tend to hide their headquarters and meeting locations, their membership lists, and their assets, rarely holding public activities to which outsiders (nonmembers) are invited. Political DVAs, however, may hold public rallies to seek new members and financial support. DVAs may obtain some of their income from illegal activities, but such criminal, income-producing activities are *not* central in DVAs, by definition. (When criminal income-producing activities become central to a group, it is a deviant business, like the Mafia, not a nonprofit DVA.)

DVAs are nearly unique in the high degree of risk and danger they involve for their members and leaders. If caught and convicted by government authorities, participants in DVAs may be imprisoned or even executed. Milder negative outcomes may involve large monetary fines; losing one's job, professional credentials, or reputation; and losing one's family relationships or spouse. Members usually conceal their membership in a DVA from other people, even from close friends and relatives, to the extent possible. All this secrecy is needed to avoid being discovered and punished by government authorities. DVAs generate very high, if often temporary, commitment to the beliefs and goals of the DVA. This strong commitment is necessary to counteract psychologically the substantial danger and risks involved when a person is a member of, or actively participating in, some DVA. Unlike most associations, DVAs can become groups a person would die for. This is rarely true of conventional associations, and almost never true for voluntary agencies. Very powerful group ideologies and deep, close personal relationships with other association members and with leaders usually are involved. Such DVAs become 'primary groups,' like families, unlike most contemporary associations being 'secondary groups.'

Degree of Government Control of Associations

The degree of government control over a specific association or type of association varies greatly. All national governments seek some control over associations operating within their geographic territories, especially for larger associations in terms of members, assets, or income. Totalitarian dictatorships usually eliminate or take over direct control of associations when such regimes assume power, often in a military coup or revolution. This was true in Stalin's Soviet Union, the Nazi Third Reich, and in Mao's China. Nontotalitarian but still authoritarian regimes, with very strong and controlling governments, often allow limited numbers of associations to exist with close government monitoring. But in such nations, the government suppresses opposition political associations or parties, as well as political activities by officially *non*political associations.

More democratic and participatory government regimes allow all kinds of associations to form and exist, so long as they do not break specific laws (e.g., involving treason, violence, disorder, conspiracy, harm to people or property, or theft). The

existence of laws or regulations in a nation guaranteeing civil liberties and strong enforcement of these rules leads to many more associations being formed and being active (greater association incidence and prevalence). In these nations, a genuine civil society may be said to exist. Mere presence of a large voluntary, nonprofit sector with many VOs, as in China now, does not by itself indicate a civil society (Smith, unpublished).

In transitional authoritarian regimes (former totalitarian dictatorships) like the Russian Federation and the Peoples Republic of China, there remain many barriers to the formation, government registration, and independent operation of associations (Guo et al., 2012). Such governments fear active political opposition that could be generated by large associations with any stated goals, not just openly stated political goals. For instance, in China nearly all *national* associations are directly controlled by the party-state, although lower geographic scope associations, especially GAs, now often have substantial operational freedom in making their own decisions. In post-Communist Europe, associations and other VOs remain weak 15 years after obtaining official freedom (Howard, 2003). This weakness results in part from lingering government fears of opposition and also public suspicions regarding nonprofits, volunteering, and associations, based on prior negative experiences under Communist regimes.

Degrees of Internal Democracy

Although voluntary *agencies* have little or no internal democracy, voluntary *associations* tend to have more substantial internal democracy. Guo and Musso (2007) have distinguished *formal* representation (fair, contested elections for leaders) from *participatory* representation (mechanisms for members to participate in key decisions). Especially in more democratic nations, the membership in associations usually has a voice in electing their top leaders annually or less often. In societies with less experience regarding participatory democracy and free associations, however, the association leaders in a given year often decide among themselves, by consensus or voting, who the next year's association leaders will be. Thus, associations can vary greatly in their internal democracy, both within and especially among nations.

The more authoritarian the nation's political regime, the more likely an association will be to have a single, autocratic leader (autarchy) or a small set of top leaders (oligarchy) in power over many years. In authoritarian nations, leaders of supralocal associations often are selected by some official of the government, rather than by the membership of the association. In nations with more democratic government regimes and stronger traditions of civil liberties and democracy, associations usually have more substantial internal democracy at any level of geographic scope. Internal democracy in an association can be defined as having *both* the aspect of formal representation as well as the aspect of participatory representation (Guo and Musso, 2007):

1. Formal representation involves such regular processes as the rotation of specific persons in top leadership positions, contested (two or more candidates) elections of top leaders (key officers, board of directors) by all members annually or less often, and term limits on key officers and board members (such as only two terms in a specific position being allowed for any individual). In more democratic political regimes, there also tends to be better representation among the top association leaders and board members of the membership as stakeholders, relative to the situation in more authoritarian regimes. Even in democratic regimes, there is a tendency toward leadership oligarchy in associations, with the same people remaining powerful over long periods of time. Such leaders often move among different key positions or just have informal power, with no official position at present.
2. Participatory representation involves such regular mechanisms as (1) the provision for holding internal town hall (total membership) meetings to discuss and even vote on major issues, (2) special advisory committees (including many members as well as leaders) focused on discussion and decision making on some key issue(s), and (3) leader-recall elections. Such elections, usually based on the constitution or bylaws of the association, include provisions for members to create and sign petitions for recall elections. In the election itself, members may vote to remove or keep one or more, specific, top leaders whom a sufficient number of members perceive as inadequate in their skills or as acting improperly.

Monomorphic versus Polymorphic Associations

Many associations, especially GAs, are unique, with no linkages to similar associations elsewhere in the same nation or different nations. These are termed 'monomorphic' associations (Smith et al., 2006), having only one form. Many other associations, however, especially GAs, are branches (subunits) of some larger, often supralocal, association. These are termed 'polymorphic' associations, having many (poly-) forms or branches (morphs). Such associations are more frequent in the developed, industrial, and postindustrial nations of the Global North. For instance, a local trade union or religious congregation may be part of a state, provincial, national, or even international association. Similarly, a Boy Scout or Girl Guides troop in a given neighborhood is usually part of a city, state, national, and then international Scouting Association (as an international NGO (INGO)).

This analytical distinction is of substantial practical importance: Polymorphic GAs elicit better public name recognition and more positive attitudes from the public, usually have more prestige, are better able to recruit new members and leaders, attract more donations and gifts, and, as a result, live longer. Such polymorphic GAs receive various kinds of help and support from their larger, usually supralocal, parent associations. In addition, polymorphic GAs are usually able to follow established and successful patterns of internal structure and activities. Unlike monomorphic associations, polymorphic associations do not need to invent everything they do, instead following successful pathways and rules or patterns invented by others in similar GAs (other GA branches).

As in the business sector, GAs with successful models of structure and operations undergo a kind of organizational evolution and dispersion in their own societies of origin,

sometimes also spreading internationally. These successful models of monomorphic GAs thus sometimes are transformed into polymorphic, supralocal associations over time, either combining with similar GAs or establishing a new, supralocal association. In the course of decades or longer, however, even successful polymorphic associations or *types* of polymorphic associations may die out as the needs of people and the circumstances in their society change. For instance, occupational guilds have died out and been replaced by trade unions, representing the workers, not both workers and owners. Putnam (2000) has documented the decline of traditional, polymorphic, membership type of associations in the United States in the last third of the twentieth century. Similarly, Skocpol (2004) has shown a trend toward the more professional, paid-staff-management type of associations in the United States, and the decline of earlier member-based leadership of associations.

Isolated versus Collaborative Associations

Being a polymorphic branch GA of a supralocal association is a *vertical* kind of collaboration with another association. There is also *horizontal* collaboration of a GA with other associations in its own community or locality. Some GAs cooperate with other local GAs on short-term projects of common interest. But some GAs also join long-term networks, federations, or councils of GAs with similar or overlapping goals. For instance, environmental GAs or women's rights and services GAs often join together into such cooperative ventures in a given city or metropolitan area.

As with vertical collaboration in polymorphic GAs, horizontal collaboration by GAs in networks strengthens the GAs involved in various ways: Member GAs gain access to new resources (money and members), achieve broader collective impact in their service or issue area, gain greater public name recognition and prestige, and are more likely to survive longer than isolated GAs. Inter-GA collaboration can lead to fruitful mergers, with the new association being stronger by synergy and its larger size and resources. Many regional, state/provincial, national, and international associations have formed as a result of such fruitful mergers. In unity there is strength, both locally and supralocally. Many association leaders act on this knowledge.

Liveliness and Intensity of Activity

Associations are living, collective entities – one might even say, 'social beings' or 'social organisms.' As such, they have a definite life cycle and an average life span. GAs tend to have a short life span (5–10+ years), whereas national and international associations tend to have a much longer life span (20–30+ years). Knowing that a given association is clearly active and 'alive' is a crucial analytical factor, of *far* greater importance than its purposive type and goals. Many associations gradually fade away and die (disappear, exit, dissolve), or they become comatose and moribund (inactive, but still alive), without many outsiders or nonmembers realizing this. The names of associations may remain on local, state or provincial, national, or even international lists or registries for years, after all group activity has ceased.

For instance, the U.S. Internal Revenue Service (IRS) recently flushed about 400 000 registered charities from its national list of tax-exempt organizations, including many associations. This event resulted from VO inactivity and failure to file annual reports with the IRS. Unlike plants or animals, groups like associations are more difficult to declare 'dead' (defunct; permanently inactive). Associations have no physical bodies and exist mainly in people's minds and in their overt activity patterns. The belief in or desire for an association can persist long after regular, overt activity has ceased, making a declaration of death tricky. Comatose associations can spring back to lively activity if a new and energetic leader gets involved or significant new resources are obtained. This situation is very different than for businesses or even government agencies, where bankruptcy and loss of government funding are clear indicators of death.

The Nature and Degree of Personal Contact

Human groups must involve some degree of interpersonal contact or linkage, by definition (Smith et al., 2006: definition of 'group'). True groups are not mere analytical categories of people, such as women or farmers, as common usage sometimes suggests. How personal contact occurs and how often it occurs are important aspects of associations that also vary markedly. In preliterate societies and thereafter, GAs have long involved regular meetings in person, from daily to weekly. Strong and enduring personal relationships usually were involved, at least with some other members, and often with all or nearly all other members. For most of their existence, GAs have been 'primary groups,' to use a technical term, developing deep friendship relationships often rivaling family ties.

As human societies increased in size and complexity, however, GAs, and especially supralocal associations, usually became 'secondary groups,' with weaker interpersonal ties and fewer such ties in any given association. Personal contact with other members became less frequent as well as less intense or deep in quality. Supralocal associations, with larger geographic scope, have arisen especially in the past two centuries, along with various mass and personal communication media that have permitted mediated personal contact (by mail, telegraph, telephones, cell phones, the Internet, satellites). Supralocal associations usually meet only once (or a few times) a year, in conferences or similar large meetings. Few members know many other members, let alone have deep and personal relationships with them (virtually never with all members).

A new and more recent phenomenon is 'tertiary associations,' as associations in which there are few personal relations or contacts at all with other members, on average. In large, national or international associations, members often merely are people who have paid their annual dues or made a contribution in the past year. For instance, members of the nearly 40-million-member AARP in the United States rarely if ever see or talk with any other members in meetings or conferences. They are listed as members and enjoy membership benefits, such as discounts on certain services or products, but they never see or contact each other personally. This is very different

qualitatively from the deep and primary relationships of GAs in earlier centuries and millennia. In GAs when they were (or are) primary groups, members met (meet) daily or weekly and enjoyed personal contact with deeply meaningful and enduring social relationships.

Clearly, the dynamics of leadership, recruitment, internal structure, activities (if any), member benefits, power, funding, and the like are very different in primary, secondary, and tertiary associations. Perhaps most striking is the huge decline in social or solidarity benefits for members, as one goes from primary to secondary to tertiary associations, historically or analytically. Put another way, only GAs as primary associations retain all of the original social and relationship benefits of associations. Even these 'expressive' benefits have declined in most contemporary GAs, relative to centuries or millennia ago. Secondary and especially tertiary associations have mainly instrumental benefits (achieving some specific goal other than enjoying social relationships per se), with fewer social, solidarity, or personal relationship benefits, if any at all.

Restrictiveness of Membership Criteria

One of the most pervasive and central benefits of associations is their exclusivity. Associations significantly restrict personal relationships, focusing these mainly on other members, at least during group meetings and events, but often outside them as well. This exclusivity often conveys both prestige in the eyes of outsiders (nonmembers) and higher self-esteem in the eyes of members. Humans everywhere value prestige and enjoy self-esteem. As a result, determining the requirements for being accepted as a member of an association is important.

For the earliest associations, as GAs in preliterate horticultural societies, the requirements were simple: be male and of sufficient age (an adult, however defined locally). For most of the history of associations, GAs have been social clubs (Smith, 1997). As time passed and societies grew in size, membership requirements often became more restrictive. In general, associations have higher prestige when they have more restrictive membership requirements or when they mainly have wealthier and more powerful members. More recent social clubs usually have required sponsorship by an existing member of the club (Ross, 1976). Really exclusive social clubs or other exclusive associations often require that each new member be acceptable to all existing members. If a single existing member rejects a new member-candidate (i.e., 'blackballs' the person), the candidate cannot join. This practice emphasizes the highly social nature of such clubs.

Early religious GAs two or more millennia ago in Greece or Rome required new members to believe in a particular god or goddess, or a religious patron or ancestor (Kloppenborg and Wilson, 1996), and in the case of Christian GAs, to believe in Jesus Christ and His resurrection. Early guilds in ancient Rome required that members be involved in a certain trade or occupation, such as metalworking (Ross, 1976). GAs have always had age requirements, excluding infants and young children, and often all nonadult younger people. Until the past century or two, GAs also tended to be exclusively male, although this has changed markedly with the rise of women's rights and the feminist movement worldwide (Smith, 1997; Skocpol, 2004).

A wide variety of purposive types of associations have arisen in the past two centuries, especially in the twentieth century. Many of these have had restrictive membership requirements, but others have been open and inclusive. Large national associations seeking to maximize membership size and donors or dues-payers have been especially inclusive (e.g., the Parent–Teachers Association or the Sierra Club in the United States), often as secondary or tertiary associations. Recent research shows that in any given community in an industrial nation, GAs tend to reflect most or all of the 'demographic niches' for males and females, young and old, rich and poor, educated and less educated, and even major racial-ethnic categories. If a new GA arises in a niche already occupied, it likely will have difficulty surviving for long, given existing local competition. This analytic approach is termed the 'population ecology of associations.'

The Prestige and Power of Associations

In general, GAs have relatively little power or prestige compared with the power and prestige of government agencies and large business firms, let alone compared with old and wealthy families (e.g., the Rockefellers or the Mellons in the United States or the 'nobility' in European nations). But some associations have great prestige, nonetheless. All associations in a community or nation can be placed somewhere on a continuum of combined prestige, power, and wealth – calling the continuum 'dominance.' In U.S. towns and cities, the dominant GAs tend to be civic service associations like Rotary, Kiwanis, and Lions, with mainly high prestige male members, but no longer with males exclusively. Business and Professional Women's Clubs also have high prestige. In the United States, the Junior League has long been a high-prestige women's social service association in communities nationwide.

There are also exclusive social clubs for wealthy and powerful persons residing in urban areas of industrial nations, including so-called country clubs (for golf or tennis) or sporting clubs (for horse racing, cricket, or polo) and other, urban, elite, social and dining clubs. Nationally, very high prestige often is associated with elite professional, academic, or honorary associations. Alumni associations of high-prestige universities tend to have high prestige, as do national academies in the sciences and in the arts and humanities (e.g., the National Academy of Sciences and the American Academy of Arts and Sciences, in the United States). In many high-prestige occupations and professions, there are special academies as professional associations, which honor people with specific outstanding achievements or remarkable careers.

In all of these instances, association prestige as a value tends to follow economic scarcity principles. Associations with the highest prestige are generally the most exclusive in terms of small size and costly membership requirements (including academic degrees). The fewer the available memberships (scarcity), the higher their value is to those seeking to enter or already members. When such memberships also convey material and social benefits, high prestige usually is ensured. Not surprisingly, all high-prestige associations are conventional or mainstream associations, not DVAs.

Nature and Degree of Impact

The dominance continuum for associations referred to deals with public *perceptions* of the value and importance of an association – basically, its reputation. In this section, we consider the *actual impact* of an association on its members, on other associations, and on the larger community or society. Perceived reputation or dominance usually is correlated positively with actual impact, but the relationship is not perfect. This is especially true regarding the impact of an association on its members. Associations with low dominance ratings, including DVAs, may have substantial impacts on their members as active volunteers.

The proliferation and persistence of many kinds of associations in nontotalitarian societies worldwide is one indicator of their actual impact. If association members did not receive significant benefits from their membership(s), associations would not survive, either individually, as specific-purpose types of associations (e.g., self-help groups, labor unions), or as a general form of human group. Recent research (e.g., Smith, 2000: Chapter 9; Smith et al., 2014: Chapter 57) shows clearly that member volunteering in associations leads usually to many positive outcomes for participants. Such positive impacts include better mental and physical health, greater happiness and satisfaction, better social relationships and more social capital, political empowerment and influence in many cases, greater information and certain skills, and sometimes economic and occupational advancement.

Looking at the larger community, many associations have clear and positive impacts on local problems and issues. For instance, the national association in Bangladesh with the acronym BRAC has been extremely successful in helping poor people there and elsewhere in the Global South with their own economic development and poverty issues (Smillie, 2009). This is not an isolated instance, as Fisher (1993) has shown convincingly. Local-development GAs and supralocal, grassroots, support organizations are two keys to sustainable development in developing nations, she has argued. Even when considering related issues in the Global South, associations have important roles to play, such as in helping urban migrants adjust in big cities (e.g., Little, 1965). GAs also may have negative impacts on their members or on their local communities. Delinquent juvenile gangs are one of many possible examples.

At the level of nations, associations have had a major impact throughout the industrial nations and often in Eastern and often in Global South nations as well, both positive and negative. As a beginning, political parties are associations, and have had vast importance and impacts in the politics of most modern nations, including one-party states. Some parties have brought freedoms, but others have brought oppressive dictatorships. Delving deeper, social movements and especially their component social movement organizations are associations that have changed history worldwide in the past two centuries. Social movement associations have significantly helped the situations of slaves, farmers, laborers in factories, women, children, older people, racial-ethnic minorities, the disabled, gays and lesbians, and many other categories of historically disadvantaged people with special needs or problems (Gamson, 1990). Associations have been vital components and preservers of democracy in many nations (Warren, 2001), as manifestations of the participatory essence of civil society. Voluntary *agencies* rarely play this vital role in civil society worldwide, lacking most participatory elements present in associations.

At the global level, it can be argued that INGOs, as international associations, have been global moderating influences, often ameliorating or resolving transnational conflicts, as well as building world culture, for the past century and more (Boli and Thomas, 1999). Even the many intergovernmental organizations (IGOs) of the world, like the United Nations and the Organization of African Unity, are in fact associations of governments as organizational members. These IGOs play key roles in peace and war, providing multinational forums for discussion and delivering certain key services as public goods, including fostering economic and social development. But there are also noxious and harmful INGOs like Al Qaeda that are active on an international scale. Like all other types of human groups (families, businesses, governments), associations mainly do good and useful things, but also can do harmful things at any geographic level (Smith and Eng, 2013).

Conclusion

Associations exist in a bewildering variety of forms, structures, sizes, and geographic scopes. On the basis of the current global population and the estimated global prevalence of seven associations per 1000 population, approximately 56 million associations worldwide exist at present. Of these, about 50 million (roughly 90%) are GAs. An estimated 1 billion people now are members of one or more associations, and even more will be members sometime during their lifetimes. The cumulative monetary value of time contributed by active association members as volunteers was at least US$500 billion in 2005, according to one research-based estimate (Smith et al., 2014: Chapter 44). Associations also have cumulative global incomes and assets in the hundreds of billions of U.S. dollars. Many associations strongly support the economic systems and economic development in their own nations and globally.

There are associations focused on every topical area, idea, belief, issue, and problem in contemporary nations having nontotalitarian political regimes. Beginning mainly 10 000 years ago as social clubs, associations now tackle every type of problem and potential benefit for members or for the larger society, often having significant impacts. Sometimes these impacts are unintentionally negative or harmful to members or nonmembers or both; associations also can be intentionally harmful. The members of associations mainly participate during their leisure time. But such associations can and often do serve economic and occupational needs and goals, as well as conventional political, social movement (activist), religious, health, education, social welfare, self-help, economic development, and indulgent leisure needs and goals.

Most associations in any nation are conventional and law-abiding, but some break the law on a long-term basis, as DVAs. Social movement associations in particular, although initially seen as DVAs, often turn out to be the *positive* 'dark energy' of the voluntary nonprofit sector. Such associations frequently

foster positive social innovation and sociocultural change in their own nations and sometimes in the world as a whole. Governments and businesses mainly *run and control* the world, but certain associations *permanently change* the world, including changing government regime structures and entire national business and economic systems.

Voluntary, nonprofit, membership associations are a persisting and persistent third force to be reckoned with in the world and in every contemporary nation. This is even true in totalitarian dictatorships, where some independent associations usually exist underground in secrecy. Associations as a whole cannot sensibly be ignored as irrelevant, weak, or unimportant. However small specific associations may be, many are in existence at a given time in a given nation, and however much their existence and activities apparently are controlled or suppressed by the government. Associations are a key form of human group that has unlimited potential, as history has demonstrated.

See also: Voluntary Organizations.

Bibliography

www.arnova.org – the Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA).

Boli, J., Thomas, G.M. (Eds.), 1999. Constructing World Culture: International Nongovernmental Organizations since 1875. Stanford University Press, Stanford, CA.

Boulding, K., 1953. The Organizational Revolution. Harper, New York.

De Tocqueville, A., 1976 (1845). Democracy in America. Two Volumes. Knopf, New York.

Fisher, J., 1993. The Road from Rio: Sustainable Development and the Nongovernmental Movement in the Third World. Praeger Publishers, Westport, CT.

Gamson, W.A., 1990. The Strategy of Social Protest, revised ed. Wadsworth, Belmont, CA.

Guo, C., Musso, J.A., 2007. Representation in nonprofit and voluntary organizations: a conceptual framework. Nonprofit and Voluntary Sector Quarterly 36 (2), 308–326.

Guo, C., Xu, J., Smith, D.H., Zhang, Z., 2012. Civil Society, Chinese Style: The Rise of the Nonprofit Sector in China. Nonprofit Quarterly Fall, 20–27.

Howard, M.M., 2003. The Weakness of Civil Society in Post-Communist Europe. Cambridge University Press, Cambridge, UK.

www.icsera.org – the International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA).

www.istr.org – the International Society for Third Sector Research (ISTR).

Kloppenborg, J.S., Wilson, S.G. (Eds.), 1996. Voluntary Associations in the Graeco-Roman World. Routledge, London.

Little, K., 1965. West African Urbanization: A Study of Voluntary Associations in Social Change. Cambridge University Press, Cambridge, UK.

Putnam, R.D., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.

Quarter, J., Sousa, J., Richmond, B.J., Carmichael, I., 2001. Comparing member-based organizations within a social economy framework. Nonprofit and Voluntary Sector Quarterly 30 (2), 351–375.

Ross, J.C., 1976. An Assembly of Good Fellows: Voluntary Associations in History. Greenwood, Westport, CT.

Skocpol, T., 2004. Diminished Democracy: From Membership to Management in American Civic Life. University of Oklahoma Press, Norman, OK.

Smillie, I., 2009. Freedom from Want: The Remarkable Success Story of BRAC, the Global Grassroots Organization that's Winning the Fight Against Poverty. Kumarian Press, Sterling, VA.

Smith, D.H., 1991. Four sectors or five? retaining the member-benefit sector. Nonprofit and Voluntary Sector Quarterly 20, 137–150.

Smith, D.H., 1997. The international history of grassroots associations. International Journal of Comparative Sociology 38 (3–4), 189–216.

Smith, D.H., 2000. Grassroots Associations. Sage Publications, Inc, Thousand Oaks, CA.

Smith, D.H., 2010a. Grassroots Associations. In: Anheier, H.K., Toepler, S., List, R. (Eds.), International Encyclopedia of Civil Society. Springer, New York, pp. 804–810.

Smith, D.H., 2010b. Membership and Membership Associations. In: Anheier, H.K., Toepler, S., List, R. (Eds.), International Encyclopedia of Civil Society. Springer, New York, pp. 982–990.

Smith, D.H., 2013. Growth of Research Associations and Journals in the Emerging Discipline of Altruistics. Nonprofit and Voluntary Sector Quarterly 42 (4), in press.

Smith, D.H., Eng, S., 2013. The Darker Side of Philanthropy. Chapter 19. In: Harrow, J., Jung, T., Phillips, S. (Eds.), The Routledge Companion to Philanthropy. Routledge, London, UK.

Smith, D.H. Review and Assessment of China's Nonprofit Sector after Mao: Is It 'Civil Society' Now and Does that Matter? Department of Sociology, Chestnut Hill, MA, unpublished paper under editorial review.

Smith, D.H., Shen, Ce, 2002. The roots of civil society: a model of voluntary association prevalence applied to data on larger contemporary nations. International Journal of Comparative Sociology 42 (2), 93–133.

Smith, D.H., Stebbins, R.A., Dover, M., 2006. A Dictionary of Nonprofit Terms and Concepts. Indiana University Press, Bloomington, IN.

Smith, D.H., Stebbins, R.A., Grotz, J. (Eds.), 2014. Palgrave Research Handbook of Volunteering and Nonprofit Associations, Palgrave Macmillan, Basingstoke, UK.

Smith, D.H., 2015. Deviant Voluntary Associations: Noxious, Dissenting, and Eccentric Versions of the Dark Energy of the Angelic Nonprofit Sector. David Horton Smith International, Bradenton, FL.

Smith, D. H., Associates, (Eds.), 2015. Handbook of Voluntaristics and Altruistics: The History, Present Scope, and Application of Nonprofit Sector and Voluntary Action Research Knowledge. David Horton Smith International, Bradenton, FL.

Wang, M., 2011. Emerging Civil Society in China, 1978–2008. Brill, Leiden, Netherlands.

Warren, M.E., 2001. Democracy and Association. Princeton University Press, Princeton, NJ.

Voluntary Organizations

David Horton Smith, Boston College, Retired in Bradenton, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Voluntary organizations have two main subtypes: voluntary agencies (VolAgs) and voluntary associations. This article deals with VolAgs, usually nonprofit service organizations with a paid staff, while a different article deals with voluntary associations. The article reviews VolAg history, incidence and prevalence, formation and life cycles, scope and types, goals and missions, governance and management, government relations, resource attraction, impact and outcomes, and research and usable knowledge.

According to Merriam-Webster's Unabridged Dictionary, the word 'voluntary' means an action of will or choice. Voluntary organizations (VOs) fit with this definition by being organizations that people join and participate in by their own free choice, not because of compulsion by law or force. By common definitions, VOs are also distinctive in having not-for-profit (nonprofit) primary goals, as contrasted with business organizations (businesses), which have for-profit primary goals. There are various other common names for VOs, the most frequent being 'nonprofit organizations' (NPOs), 'nonprofits,' 'nongovernmental organizations' (NGOs), or 'civil society organizations' (CSOs).

There are two main structural types of VOs: associations and nonassociational nonprofits (Smith, 2000: p. 30). Associations, usually called voluntary associations (or sometimes membership associations), are treated in a different article in this encyclopedia, under the heading 'Voluntary Associations, Sociological Aspects.' The present article discusses nonassociational nonprofits as a subtype of VOs. The simplest label for this subtype of VOs is 'voluntary agencies' (VolAgs). VolAgs usually have significant respect and prestige in their societies, and are perceived as useful and altruistic.

Many people confuse voluntary associations and VolAgs, or incorrectly use the term 'voluntary organizations' to refer just to one type or the other, not to both, which is correct. Others similarly fail to understand how different the two basic VO types are in their basic structure. Simply put, voluntary associations have a structure in which there are members who hold the ultimate power, while VolAgs usually have a board of directors (or trustees) that holds the ultimate power. VolAgs (sometimes called social agencies or welfare agencies) have no members, but have paid staff and often volunteers.

In a voluntary association, members elect the top leaders (board members, officers, sometimes key executives) on a regular basis (e.g., annually, biennially). In a VolAg, the Board of Directors (Trustees) is self-perpetuating, electing or appointing new members as needed, sometimes annually or biennially. There is no category of members as a kind of affiliate, just employees and volunteers. In sum, voluntary associations practice a kind of internal democracy, while VolAgs usually do not practice internal democracy. VolAgs are generally structured more like government agencies or business corporations, as hierarchical bureaucracies with the power at the top. This crucial difference in basic internal structure makes these two subtypes of VOs very different in their operations, organizational culture, resource attraction processes, and impact on participants, among other effects (Smith, 2000: p. 30). Smith has argued that these two VO subtypes constitute different subsectors of the voluntary nonprofit sector (VNPS), third sector, or civil society sector.

The History of Voluntary Agencies

The earliest VOs were voluntary associations, originating perhaps 25 000 years ago (Smith, 1997). Such associations became common about 10 000 years ago when human societies settled down in agricultural villages after 100 000–200 000 years as nomadic bands of less than a 100 people each. These horticultural villages had an average of 1500 people, so that subgroups now made sense.

Many thousands of years later, in the ancient empires of Egypt, Greece, and Rome, some VolAgs arose, although their independence from the government was questionable in most cases. A reasonably clear example of an ancient VolAg was Plato's Academy in Athens, as an early protocollege (*c.* 387 BC and thereafter), and subsequently Aristotle's Lyceum (founded *c.* 335 BC). There were also ancient universities founded in Asia and the Middle East.

There was an Ancient (or Royal) Library at Alexandria in Egypt from the second century BC until destroyed by fire in 48 BC. Although supported financially by the government of the Pharaohs, the library had some independent VolAg features. Probably the first VolAg hospital was at the Temple of Aesclepius at Epidaurus in Greece (from *c.* 430 BC). Later, a similar Aesclepium was built in Rome (second century BC). VolAg hospitals in Rome were built during the Christian period of the Roman Empire, after Constantine legalized the religion (AD 313 and thereafter). For instance, St Basil founded a Christian hospital in Cappadocia, Turkey (*c.* AD 370), and many others were founded throughout Europe and the Middle East in later centuries of the Middle Ages. Particularly distinctive were Arab hospitals early in the second millennium, which admitted all sick or injured people regardless of religion, race, etc.

Early monasteries, unlike churches, were VolAgs. Buddhist monasteries (termed 'vihara') seem to have arisen first, in the third century BC in India and Bangladesh. Some of these later became early Buddhist universities. Vihara spread to many other Asian nations in later centuries. St Pachomius founded

the earliest Christian monastery in Egypt (*c.* AD 318), with hundreds more being founded there and in Middle Eastern nations subsequently. St Benedict founded the first Western monastery at Monte Casino in Italy (*c.* AD 529), which became a model for hundreds of others in Europe subsequently. Shankara founded Hindu monasteries (termed 'mathas') in India beginning in AD eighth century. Other Hindu sects later also formed mathas in India.

Independent (nongovernmental) elementary and secondary schools as VolAgs (but part of the Catholic Church) were formed in medieval England. First was The King's School, Canterbury (founded in 597), then St Peter's School, York (founded *c.* 627), followed by many others in the 600s and thereafter. The first VolAg secondary schools independent of the church in the United Kingdom were Winchester College (founded in 1382) and Oswestry School (founded in 1407).

Some types of VolAgs as we now view them have arisen mainly in the past millennium. Classic examples are the University of Bologna (founded in 1088), Oxford University (founded in 1096), the University of Paris (founded in 1150), the University of Modena (founded in 1175), and Cambridge University (founded in 1209).

The earliest museums, as collections of artifacts of interest, were private and personal, owned by wealthy collectors. The Ennigaldi-Nanna's museum dates to *c.* 530 BC, founded and led by a Babylonian Princess in Ur (modern day Iraq). Museums open to the public arose in the Renaissance, with the Capitoline Museums of Rome being the first (founded in 1471). The Vatican Museums in Rome came next in 1506. As part of the Roman Catholic Church, this museum was more clearly a VolAg, unlike most subsequent major museums, which were national or municipal government agencies. Many other public museums followed, especially from the 1700s on, some of them being VolAgs.

Almshouses are one of the earliest forms of VolAg social welfare agencies, dedicated to providing lodging, and often food, to the poor, especially the elderly and single poor. The oldest surviving almshouse in the United Kingdom is called the Hospital of St Cross, founded *c.* 1133 in Winchester. Many others were formed subsequently as VolAgs in the United Kingdom and in other European nations (e.g., the Netherlands). Almshouses in the Middle Ages were often called 'hospitals,' giving hospitality/care to the old, poor, and feeble, as well as to people with specific mental or physical health problems. Poorhouses or workhouses were not VolAgs, but were government-run places for the lodging of the poor and elderly, requiring labor from their residents.

Orphanages for children without living or functional parents arose as Christian welfare VolAgs in the first century AD and have existed down to the present as a VO subtype in many nations. In the Middle Ages in Europe, monasteries often acted as VolAg orphanages. Germany and the Netherlands had VolAg orphanages from the late 1500s. The first VolAg orphanage in the United States was founded in New Orleans by the female religious order of Ursulines (1727). Orphanages as VolAgs spread to many nations in Asia, Africa, and Latin America via religious missionaries from about 1600.

There are many other kinds of social welfare VolAgs now in existence, focused on specific disabilities (blindness, deafness, developmental disabilities, low intelligence, physical limitations of the body, etc.), homeless shelters, shelters for battered/abused women or children, immigrant and refugee centers/shelters, earthquake and flood victim shelters, alcohol and drug addict halfway houses, exconvict halfway houses, vocational rehabilitation centers, day care centers for children, etc. The vast majority of current VolAgs have been founded in the past century or so. Although many have religious underpinnings, there are now many others that are secular.

Factors Affecting the Incidence and Prevalence Rates of VolAgs

From the above, it should be clear that the general economic development level of a society has a major effect on VolAg formation (e.g., incidence rate in a territory or society) and existence/growth/decline/demise (e.g., prevalence rate). In preliterate, nomadic, hunting–gathering societies, there were no VOs, including both subtypes. In preliterate horticultural (village) societies, the somewhat higher economic development level permitted somewhat larger populations, owing to primitive agriculture (called horticulture – basically just gardening). However, the VOs formed there were voluntary associations, never VolAgs (Smith, 1997).

When there was more advanced agriculture (with plows, fertilizer, irrigation, etc.), as in major ancient civilizations, much larger populations could be supported in urban places (towns, cities, metropolises, and megalopolises). However, VolAgs were quite rare until the Industrial Revolution, which caused an Organizational Revolution (Boulding, 1953). This higher level of economic development and efficiency provided an economic surplus that some wealthier people used to form VolAgs that they believed would serve segments of the population not served by families, governments, or businesses. In addition, government permissiveness of VolAgs was necessary for their formation and operation in a public (vs. underground) manner. Greater communication–transportation network development and the rise of mass education accompanying the Industrial Revolution also facilitated VolAgs, as did greater population size in a place or society.

Accompanying these larger population densities in ancient civilizations from about 500 BC was the growth of government and business organizations and bureaucracies, with greatly increased differentiation of occupations inside them and independently of these bureaucracies. This greater division of labor led to different economic interests of people in different occupations, which facilitated the formation of occupational voluntary associations (e.g., in ancient Rome; Smith, 1997) and also certain VolAg types.

The Formation and Life Cycles of VolAgs

When one or more founders (we will consider the term 'founder' as potentially multiple henceforth) are considering the formation of a new service organization, such a group can potentially be located in any of three sectors of society: the government, business, or nonprofit sector. Founders usually choose the voluntary nonprofit sector (VNPS) when forming a service organization, since a government agency is unlikely or too time

consuming to form, and since earning a profit is not a paramount goal.

Once the VNPS has been chosen as a societal location, the next founder decision is whether the new group is to be a VolAg or a voluntary association. VolAgs are usually chosen as the appropriate VO form when the new group is to have mainly a social service goal and when paid staff are anticipated to operate the group. The founder also must consider in which nation and city/town/area to have the physical base of the group.

Additional founder decisions that are usually made by VolAg founders (Smith, 2000: Chapter 3) are (1) to have a nonmember-benefit or 'public service' goal, (2) to be conventional or mainstream, rather than deviant or underground, relative to national and local laws, (3) to have some or all paid staff to run the VolAg rather than only volunteers, (4) to have a nondemocratic, corporate-like structure with top-down control by the initial leaders/board, (5) to specify the territorial scope or catchment area of people to be served, (6) to specify what types of recipients are to be served (e.g., the general public, children, the aged, veterans, poor people, or sick or handicapped people), and (7) to specify what services are mainly to be provided (e.g., education, health, exercise, preservation of artifacts or art, funds, food, or lodging).

More variability is found with regard to founder decisions about the specific ideology of the VolAg, the degree of dependence on volunteers (as in a VolAg volunteer program), how to attract and keep volunteers, sources of funding, how to recruit and keep suitable paid staff, linkages or subordination to existing organizations in any sector, whether and where to incorporate as a legal person, whether and where to register formally with the government, whether to seek formal tax exemption from the national government, where and how to have a physical base/office for the group, and the degree of formality vs. informality of internal structure and operation.

VolAgs tend to live longer on average than voluntary associations, because more effort goes into forming and running them, they tend to have more assets and income, and their paid staff want to keep their jobs. VolAgs are also better able to resist economic downturns than businesses because of the leader commitment to their missions (Bowen et al., 1994: p. 99). As with any kind of organization, there is a 'liability of newness' for VolAgs (Bowen et al., 1994: p. 104). Younger VolAgs are more likely to fail/die than older ones. There is a kind of Newtonian momentum to organizations once well established. Wealthier VolAgs are also less likely to fail/die. VolAgs very rarely transition into being businesses, or vice versa, but may become social enterprises as voluntary–business sector hybrids. VolAgs very rarely transition into voluntary associations, or vice versa.

Block (2004) has discussed several major reasons why VolAgs fail: (1) failure to do adequate orientation of board members; (2) a negative or depressed organizational culture/atmosphere; (3) leaders acquiring and using power to benefit themselves, not the VolAg mission; (4) confusion regarding roles/duties of board members, executive staff, and committee members; (5) financial misfortune and loss of funding/income; (6) 'fund phobia,' or fear of asking people for financial contributions; and (7) domination or control of the organization by the founder over an extended time period.

Additional factors in demise can be substantial fraud or abuse by organization leaders or staff, or other significant violation of laws and regulations. VolAgs are as susceptible to crimes and misconduct as are businesses and government agencies (Smith and Eng, 2013; Zack, 2003). Deviance is a human quality, not confined to any one sector of society or type of organizations.

VolAgs need to engage in risk management by purchasing insurance against various kinds of misconduct and implementing internal controls and external audits (Renz et al., 2010: Chapters 22, 23). There are also various kinds of social accountability and social accounting that need to be performed (Renz et al., 2010: Chapter 4).

Scope, Magnitude, and Types of VolAgs

In all contemporary societies where they are present, VolAgs tend to be larger in assets and income on average than voluntary associations in their society. Voluntary associations usually include more participants than do VolAgs, most participants being members and many being active members or associational volunteers. VolAgs usually have more paid staff in the aggregate, compared to associations in a nation. VolAgs are much more likely to own or rent their own office space continuously, compared to voluntary associations, and are much more likely to have a listed telephone number in a directory and to have a Website for the organization. The generalizations above are more accurate for more economically developed societies in terms of average level of education and gross domestic product (GDP) per capita.

Research by Salamon et al. (2004: p. 17) showed that 56% of the full-time equivalents of the nonprofit sector work force are paid staff (vs. volunteers) for the 36 nations studied, comprising 4.4% of the economically active population (p. 16). Most of these paid staff were probably employed in VolAgs, not in voluntary associations. The proportion of the paid work force in VolAgs was substantially higher in developed nations than in developing or transitional nations (pp. 18–19). Salamon et al. (2004: p. 16) also found that VolAgs' expenditures were 5.4% of the combined GDP of the 36 nations. VolAgs are thus a significant global economic force.

The vast majority of VolAgs in most nations fall into the categories of religious congregations and faith-based organizations, nonprofit secondary schools or colleges/universities, hospitals and other health agencies, museums, research centers, social enterprises, social welfare agencies, sports and recreation organizations (especially leagues and federations), art and cultural organizations, advocacy organizations, and foundations. In more developed nations, there are also transnational aid, relief, development, and friendship as well as environmental and advocacy VolAgs. O'Neill (2002) describes and discusses the major types of VOs for the United States, as do many other authors for other nations. Various classification schemes have been suggested as typologies of VolAgs (e.g., Bowen et al., 1994: p. 21; Salamon et al., 2004: p. 12; Smith, 2000: pp. 4–5).

Goals and Missions of VolAgs

Governments usually have as key goals control over the residents in their territories (e.g., maintaining the peace and conformity to laws), maintenance of the current leadership

power structure, generation of tax revenues, and the provision of some public goods (e.g., roads, bridges, water systems, sewer systems, defense, and wars). Businesses have as key goals the generation of a profit for the investors/owners based on income from the sale of private goods or services, longer term growth in company assets and income while protecting capital, greater market penetration, and maintenance of stability of the company labor force, including management.

By contrast, VolAgs may have a wide variety of goals unrelated to control/power and unrelated to assets/income/profits. VolAgs face issues related to such goals, but put primary emphasis on public service or nonmember benefit goals. VolAgs usually develop and change/expand their goals so as to meet some perceived needs of the people they intend to serve, based on an altruistic concern for such recipients of VolAg services. However, VolAgs may also change their goals subsequently to serve the needs of the managers and paid staff of the VO. This can be termed 'mission drift' or 'goal succession' to put it kindly, but 'mission abandonment' or 'organizational maintenance goal' to put it more bluntly and honestly.

When there is a financial surplus of income relative to expenses for a VolAg in a given year, the surplus is not divided among the owners, managers, or original donors of funds, contrary to the practice in most businesses. The surplus (technically, not a profit) is plowed back into the VolAg, but usually benefits the managers and staff, and to a lesser extent recipients of its services in the future. This practice is quite distinctive of VolAgs as contrasted with businesses. Hansmann (1980) terms this, famously, the 'nondistribution constraint.'

VolAgs very rarely produce physical goods, but mainly provide services to their recipients (customers, users), unlike manufacturing businesses. Economists call these services of VolAgs 'club goods.' As such, they stand in marked contrast to the 'private goods' (divisible and usable by individuals) provided for sale by businesses or the 'public goods' (indivisible and usable simultaneously by all or most people residing in a territory) provided to residents by government agencies. Club goods are goods (usually services as goods) that are provided to a limited set of individuals because of their special relationship to the provider organization, such as a particular VolAg. The term 'club goods' originated because economists were analyzing the special benefits (goods) provided by voluntary associations (clubs, broadly viewed) to their members. The club goods concept also applies to the specialized clientele of VolAgs (e.g., hospital patients, museum attendees, university students, and food bank recipients).

VolAgs usually operate mainly with paid staff to accomplish their goals, and have done so as a type of VO for over two millennia. However, beginning in the late 1800s and especially in the past 50 years or so, VolAgs have developed volunteer (service) programs to use unpaid workers to perform some of the personal service activities directed at their goals. Such volunteer programs are not VOs of any kind. Rather, they are departments of other larger organizations that recruit, manage, and utilize volunteers to serve the goals of the larger organization (Smith et al., 2006: p. 244). Importantly, volunteer programs can be departments of a government agency (e.g., volunteers in a national park or government hospital), a business (e.g., volunteers in a for-profit hospital), or in a VolAg (e.g., volunteers in a VolAg museum or social welfare agency like the Red Cross/Red Crescent agencies).

The management/administration of VolAgs is typically a paid staff activity, sometimes very well paid staff (e.g., CEOs of large VolAg hospitals or universities). Volunteer boards of directors are usually parts of the governance structure, but not of the management structures of VolAgs.

Larger and wealthier VolAgs give stipends (essentially part-time pay) to their board members, and often provide expenses for travel to board meetings. The participants in the volunteer program are very rarely, if ever, part of the management, let alone the governance, of the VolAg. Sometimes the Volunteer Program Administrator (who may be paid full-time or part-time, or be a pure volunteer) is a minor part of the VolAg management team. VolAgs very rarely have affiliates who are genuine members, because they are not associations in basic structure.

VolAgs often place their primary goals in the context of an ideology and see the accomplishment of their altruistic goals as a mission, which involves much more commitment than a mere goal. Missions can involve commitment by VolAg staff (paid and unpaid) to goal achievement that transcends economic benefits to the staff, unlike nearly all businesses and government agencies. Thus, VolAgs usually have crucial expressive elements of staff commitment that can override purely instrumental commitments (i.e., commitments based only on salary or other financial compensation).

These expressive, moral, and ideological commitments of VolAg staff members are distinctive to VOs more generally. VO commitments often are even stronger on average in voluntary associations, where generating volunteer commitment among members and leaders is crucial to organizational survival. So-called mission statements of VolAgs may or may not properly formulate such expressive or moral commitments. Many VolAg leaders/managers do not fully understand the importance of these commitments to their existence and growth, and may mistakenly copy blindly the form of mission statements of businesses.

Another distinctive feature of VolAg goals is their sometimes-vague nature. The goals of a business or government agency are usually clear-cut, but VolAg goals are often diffuse, multiple in nature, and hard to pin down. This makes evaluation of the effectiveness of VolAgs more difficult than for businesses, where there is profit/loss as a bottom line. In addition, quite a few VolAgs espouse altruistic goals/missions, but in fact serve the goal of maintaining and expanding the VolAg and payments/compensation to its top executives and other paid staff. This 'goal displacement' implicitly and often gradually replaces seeking of the purported goals of effective and efficient VolAg service to its intended recipients. This situation raises the key issue of 'operational goals' (*de facto*) as contrasted with 'official goals' (*de jure*). Many VolAgs are essentially small (sometimes large) businesses masquerading as VOs to reap the tax benefits involved, according to critics (e.g., Gaul and Borowski, 1993).

Governance, Management, and Administration of VolAgs

In developed nations, the Board of Directors or equivalent (Trustees, Executive Committee, etc.) usually governs VolAgs

by making major policy decisions. This often results from institutional isomorphism in the requirements for a group becoming a legal person or incorporated NPO in many nations. A minimum of three directors is common, although upper limits of size tend to be set around 20 by practicality considerations, not laws. In developing nations, only a President or Board Chair may be legally required, and governance can thus be autocratic.

Even with multiperson boards, VolAg governance can be rather undemocratic for two main reasons: first, VolAg boards are usually self-selecting. Thus, they tend toward oligarchy, with a small subset of members dominating decision making over long time periods. Second, such boards are rarely required (or try) to be representative of recipients, of other similar service providers, or of any other definable category of persons. Hence, VolAg boards are only rarely democratically structured or adequately representative of key constituencies/stake holders (Guo and Musso, 2007).

Nations vary in the degree of control of VolAgs by the national government. All nations have laws requiring government registration of VolAgs, and most require annual reports to the government on expenditures and activities. National governments often limit political activities of VolAgs, especially if they are to keep their tax exemptions. Many nations restrict which VolAgs can register, receive tax exemptions, be legal persons, and operate legally.

This indicates the important role of freedom of association and other civil liberties in the formation and operation of VolAgs. Governments vary also in how closely they regulate VolAgs, with much greater government control in authoritarian nations. VolAgs practice self-regulation to varying degrees.

In most nations there are at least some government-initiated, government-financed, and/or government-controlled VolAgs. There are various special terms for such VolAgs: QUANGOs are quasi-nongovernmental organizations (such as the RAND Corporation in the United States). GONGOs are government organized nongovernmental organizations, common in China (e.g. most national associations).

In authoritarian and one-party nations, the nongovernmental character of VolAgs is usually questionable and variable.

The top administrative leader of VolAgs is usually termed the 'Executive Director' or 'Director,' and is charged with carrying out the policy decisions of the Board. However, the Director has leeway to interpret those decisions, and to make micropolicy decisions regarding implementation practices and timing. The Director usually makes all major operational decisions, including hiring and firing of staff, compensation, roles and tasks, supervision, etc. In essence, the Director answers only to the Board.

When the Director has a close relationship to the Board and its members are strongly committed to the Director, VolAg Boards may fail in their fiduciary duties and let Directors do more or less as they please, to the detriment of the organization (Fishman, 2007).

The top management staff of VolAgs is appointed, not elected as in associations. Leaders are less likely to be charismatic, as contrasted with associations, since economic incentives (salary and fringe benefits) influence staff compliance. Supervision of lower staff is usually high, with only moderate emphasis on human relationships. Priority setting by management tends to be tight and planned, and marketing is carefully managed. Public images of VolAgs are very important, because they strongly affect resource attraction.

The internal structure of VolAgs usually involves formal incorporation with the government as a 'legal person,' formal registration with a government agency, and formal tax exemption with the government taxation agency. VolAgs are usually more bureaucratic in structure than voluntary associations, being more like businesses and government agencies. They are usually monomorphic in character, having only a single location, rather than polymorphic, having two or more branches in different locations (Smith et al., 2006: pp. 146, 178).

The internal processes of VolAgs usually involve continuous daytime operation, often 7 days a week, sometimes for 24 h a day, unlike associations. Paid staff of VolAgs are often quite professional, with higher degrees in their specialty and a strong sense of professionalism. VolAgs often have high prestige and public recognition, sometimes wielding significant power both locally and in their states/provinces, even in their nations. They usually recruit paid staff and volunteers formally, do formal training of staff, and are quite willing to fire staff for inadequate performance, like a business or government agency. VolAgs tend to be more willing to collaborate with other organizations than are associations, and often relate formally to government agencies. Mergers and federations/networks are common among VolAgs, seeking strength in unified and cooperative efforts. There is much published literature for VolAg practitioners on governance and management, both in books and handbooks (e.g., Renz et al., 2010: Parts 2–5).

Resource Attraction by VolAgs

VolAgs are partly similar to service businesses in their resource attraction. Like service businesses, they depend heavily on fees for services – museum entry fees, school and university tuition fees, hospital patient fees, etc. Data reported in 2008 primarily for US VolAgs (Renz et al., 2010: p. 483) indicated that fees accounted for 70% of revenues in all types of VolAgs, especially for health, education, and human services VolAgs. Unlike most fees for service businesses, VolAg fees are sometimes subsidized by government agencies (e.g., Medicare and Medicaid payments to VolAg hospitals in the United States; Renz et al., 2010: p. 483).

VolAgs are especially distinctive from service businesses in receiving donations/gifts from private donors, both living donors and bequests at death as well as foundation grants, accounting for about 12% of revenues in the US data reported in 2008 (Renz et al., 2010: p. 483). Such gifts were most frequent in the US data for international, environment, and arts organizations (ibid.). There is a wealth of practical knowledge about how best to do VolAg fund-raising and other resource attraction (e.g., Renz et al., 2010: Chapters 18, 19). Large and well-respected VolAgs may have very large endowments, which are pools of investments/financial assets that are the long-term residue of gifts. The Bill and Melinda Gates Foundation, for instance, had assets of $37 billion at year-end 2011, while Harvard University had assets of $27 billion as of 30 June 2011. Investment income from endowments accounted for about 5% of VolAg income in US data for 2008 (ibid.).

Unlike associations, VolAgs often obtain external grants or contracts from government agencies, accounting for 9% of all monetary income for US VolAgs in 2008 data (Renz et al., 2010: p. 483). Government direct funding was most common for Human Services and International VolAgs. Government contracting for VolAg services is a growing trend worldwide (Renz et al., 2010: Chapter 21). Voluntary associations differ markedly from VolAgs in getting most of their revenues from dues (regular, usually annual payments to maintain membership) or donations, not from fees, grants, contracts, or investments except in very rare instances.

VolAgs also need to attract human resources in the form of board members, paid staff, and volunteers as unpaid staff. Although recruiting paid staff for VolAgs can be similar to recruiting similar staff for service businesses, recruiting volunteer board members and volunteer service staff is a special process that involves eliciting the altruism of potential workers.

A key problem of VolAgs is retaining volunteers, rather than simply attracting them in the first place. Many VolAgs have high volunteer turnover owing to inadequate volunteer management. Much is known about proper volunteer management to attract and retain volunteers (e.g., Connors, 2012; Renz et al., 2010: pp. 779–780).

Impact and Outcomes of VolAgs

The author has suggested (Smith, 2001) that VolAgs, as key parts of the VNPS, have a variety of major impacts on their societies, varying by nation according to their prevalence, size of paid and volunteer workforce, and freedom to operate. VolAgs help their societies by

- providing many kinds of partially tested social innovations;
- providing new and different definitions of reality and morality, often with advocacy activities;
- preserving traditional ideas, values, and artifacts for the future;
- providing education and training both for occupations and a good life;
- preserving and providing the sense of mystery, wonder, and the sacred;
- permitting individuals to reach the fullest expression of personal capacities and potentials in a limiting environment;
- providing a source of negative feedback and criticism to the dominant institutions of society, often with advocacy activities;
- providing support to the economic system of society;
- providing human services and financial resources to needy people, both domestic and foreign; and
- providing a latent resource that can be mobilized for any new and important goal that arises.

At lower analytical levels, there are four important types of impact of VolAgs: on recipients, on the staff, on the VolAg itself, and on the community. In broad terms, VolAg impact on intended recipients seems most important. Based on evaluation research of various kinds, there is evidence that VolAgs often have significant positive 'main effects.' Schools and universities raise the information levels of students, at least temporarily, and usually improve lifelong earning capacities. Hospitals have positive effects on patient health in the short and often long term. Museums often please and inform their attendees.

However, for some VolAg types, the outcomes are less clear. Evaluating the outcomes of monasteries or of orphanages and several other types of social welfare agencies is more difficult. Appropriate outcome measures for monasteries are hard to define in secular and empirical terms. As for orphanages, many experts now believe that empirical evidence shows orphanages do more harm than good, when compared to other ways of raising children without functional parents. International VolAgs have been particularly ineffective in facilitating economic development in places like Africa, and relief efforts in developing nations often just create more dependency (Smith and Eng, 2013). Foundations mainly tend to support status quo approaches to social and health problems, rather than supporting genuine social innovations.

Evaluation of VolAg impact on the paid staff has received little empirical attention. Research on VolAg volunteer program impacts on volunteer staff has been more extensive. Such volunteering has many positive effects, such as a greater sense of well-being, increased information, more felt social support, improved or maintained health, better subsequent jobs, etc. (Smith, 2000: pp. 196–205). However, there can also be negative impacts on volunteers, such as encountering contagious diseases more frequently, 'volunteer burnout,' and rejection of future volunteering owing to poor experiences with a given VolAg's volunteer management (or mismanagement) practices.

VolAgs themselves benefit from their own effective operations by facilitating the future existence and growth of the organization. Effective VolAg operation also facilitates maintained or increased pay and other compensation for paid staff. Effective operation also benefits volunteer staff, by continuing their volunteer service opportunities. VolAgs benefit their communities by taking some of the load off government-provided services for needy people, thus reducing the tax burden in the long term. Specific VolAgs benefit other local VolAgs by providing a supportive organizational field and often federations or networking opportunities. Certain infrastructure VolAgs specialize in supporting and helping other VolAgs, such as the Foundation Center or Independent Sector in the United States.

Some experts have argued that VolAgs play a significant role in promoting democracy and pluralism. However, this is not supported by empirical evidence. Voluntary associations do clearly contribute not only to democracy and pluralism in their societies but also to corporatism in some societies.

Research on VolAgs and Usable Knowledge

Although research on VolAgs goes back more than two millennia in terms of histories, social science research by sociologists, anthropologists, political scientists, and economists is historically very recent, beginning mainly in the twentieth century. The organized field of voluntary and nonprofit sector research began with the founding in 1971 of ARNOVA (www.armova.org), the Association for Research on Nonprofit Organizations and Voluntary Action. The new

International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA; www.icsera.org) works to draw together all 30+ world region and national voluntaristics/altruistics researcher associations now existing worldwide and to provide global information on various aspects of research (e.g., academic journals, research centers). An overview of the development of VO and VNPS research is presented in Chapter 1 of Smith et al. (2014; see also Smith 2013, Smith and Associates 2015).

There are thousands of books and articles reporting research on VolAgs, many of which are summarized in handbooks like those of Powell and Steinberg (2006), Renz and Associates (2010), Salamon (2012), and Smith and Associates (2015). Smith et al. (2006) have written a dictionary of over 1700 nonprofit terms and concepts, with about 1800 references to relevant research publications. There is also much accumulated usable knowledge directed to practitioners (e.g., Block, 2004; Connors, 2012; Renz and Associates, 2010; Zack, 2003).

See also: Altruism and Prosocial Behavior, Sociology of; Charity and Philanthropy, Economics of; Civil Society, Concept and History of; Organization: Overview; Organizations, Sociology of; Policy Process: Nonprofit Institutions in the United States; Voluntary Associations, Sociology of; Volunteerism, Psychology of.

Bibliography

Block, S.R., 2004. Why Nonprofits Fail. Jossey-Bass, San Francisco, CA.
Boulding, K., 1953. The Organizational Revolution. Harper, New York, NY.
Bowen, W.G., Nygren, T.I., Turner, S.E., Duffy, E.A., 1994. The Charitable Nonprofits. Jossey-Bass, San Francisco, CA.
Connors, T.D. (Ed.), 2012. The Volunteer Management Handbook, second ed. Wiley, New York.
Fishman, J.J., 2007. The Faithless Fiduciary and the Quest for Charitable Accountability 1200–2005. Carolina Academic Press, Durham, NC.
Gaul, G.M., Borowski, N.A., 1993. Free Ride: The Tax Exempt Economy. Andrews and McMeel, Kansas City, MO.
Guo, C., Musso, J.A., 2007. Representation in nonprofit and voluntary organizations: a conceptual framework. Nonprofit and Voluntary Sector Quarterly 36, 308–326.
Hansmann, H., 1980. The role of nonprofit enterprise. Yale Law Journal 89, 835–901.
O'Neill, M., 2002. Nonprofit Nation: A New Look at the Third America, second ed. Jossey-Bass, San Francisco.
Powell, W.W., Steinberg, R. (Eds.), 2006. The Nonprofit Sector: A Research Handbook, second ed. Yale University Press, New Haven, CT.
Renz, D.O., Associates (Eds.), 2010. The Jossey-Bass Handbook of Nonprofit Leadership and Management, third ed. Jossey-Bass, San Francisco, CA.
Salamon, L.M. (Ed.), 2012. The State of Nonprofit America, second ed. Brookings Institution Press, Washington, DC.
Salamon, L.M., Sokolowski, S.W., Associates, 2004. Global Civil Society: Dimensions of the Nonprofit Sector, vol. 2. Kumarian, Bloomfield, CT.
Smith, D.H., 1997. The international history of grassroots associations. International Journal of Comparative Sociology 38 (3–4), 189–216.
Smith, D.H., 2000. Grassroots Associations. Sage Publications, Inc, Thousand Oaks, CA.
Smith, D.H., 2001. The impact of the voluntary sector on society. In: Steven Ott, J. (Ed.), The Nature of the Nonprofit Sector. Westview Press, Boulder, CO, pp. 79–87.
Smith, D.H., 2013. Growth of research associations and journals in the emerging discipline of altruistics. Nonprofit and Voluntary Sector Quarterly 42 (4), 638–656.
Smith, D.H., Associates (Eds.), 2015. Handbook of Voluntaristics and Altruistics: The History, Present Scope, and Application of Nonprofit Sector and Voluntary Action Research Knowledge. David Horton Smith International, Bradenton, FL.
Smith, D.H., Eng, S., 2013. The darker side of Philanthropy. Chapter 19 In: Harrow, J., Jung, T., Phillips, S. (Eds.), The Routledge Companion to Philanthropy. Routledge, London, UK.
Smith, D.H., Stebbins, R.A., Dover, M., 2006. A Dictionary of Nonprofit Terms and Concepts. Indiana University Press, Bloomington, IN.
Smith, D.H., Stebbins, R.A., Grotz, J., 2014. The Palgrave Research Handbook of Volunteering and Nonprofit Associations. Palgrave Macmillan, Basingstoke, UK.
Zack, G.M., 2003. Fraud and Abuse in Nonprofit Organizations. Wiley, Hoboken, NJ.

Volunteerism, Psychology of

Mark Snyder and Alexander Maki, University of Minnesota, Minneapolis, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The psychology of volunteerism has been informed by theoretical, empirical, and practical exploration. This article examines basic and applied scientific research on the antecedents, experiences, and consequences stages of the volunteer process. These stages of volunteerism are considered for individual volunteers, the recipients of volunteer services, the organizations through which volunteers serve, and the wider community in which volunteerism takes place. Theoretical and practical implications of volunteerism are also discussed.

Introduction

Each year, millions of people around the world provide voluntary assistance to individuals, groups, and causes by engaging in such behaviors as tutoring youth, providing companionship to the lonely, raising money for political causes, or organizing neighborhood groups. As diverse as these activities may be, they can all be classified as exemplars of volunteerism – voluntary action that aims to supply services to address the unmet needs of individuals, organizations, and communities.

Although volunteering has much in common with other forms of prosocial behavior, social scientists have identified criteria that make volunteering a distinct domain of prosocial behavior. Specifically, volunteerism is a thoughtful and deliberative behavior requiring the active seeking of opportunities for action, distinct from phenomena (such as emergency helping) that often find individuals in situations where immediate action and assistance are required. Much helping, including emergency helping, is a one-time, short-term effort that has costs only in the present moment, whereas volunteering often lasts for long stretches of time and can incur personal costs (such as time, energy, and resources) that can and do recur and even accumulate over time. This providing of assistance over time is similar to the experiences of caregivers who aid the chronically sick or those with disabilities. However, differences also exist between volunteering and caregiving, as caregiving often includes a sense of obligation or the bonds of a relationship (e.g., familial and marital). Volunteering, on the other hand, is often aimed at aiding a person unrelated or previously unknown to the volunteer. In some circumstances, a volunteer may never even meet the recipients of aid. Volunteerism can also be separated from other activities often found in communities and service organizations. Although some forms of volunteerism may resemble services provided by nonprofit and educational entities, such as classroom support or companionship for adults with disabilities, volunteers are not paid or reimbursed for their time and effort. This is despite the fact that volunteers do commonly work through established nonprofit and educational organizations, which often recruit, train, and place volunteers into service positions. Community activists, who often are volunteers themselves, may have similar goals as volunteers but instead work outside of established organizations to impact society and create change. In sum, it is this unobligated, long-term providing of planful assistance through established organizations without payment that constitutes a behavior as volunteering (for further elaboration on the defining and characteristic features of volunteerism, see Snyder and Omoto, 2008).

In the United States, volunteering has long been, and continues to be, a common form of community support. From September 2010 to September 2011, over 64.3 million people in the United States volunteered at least one time through an organization, equivalent to 26.8% of the country's population (Corporation for National and Community Service, 2011). Those volunteers spent, on average, 51 h per year volunteering for a wide range of causes at the aggregate level, although individual volunteers tended to spend their efforts within just one or two organizations (89%). The most common places to volunteer were in religious (35%) and educational (26.7%) settings; the most common activities of volunteers were fundraising (26.5%), collecting and distributing food (23.5%), and general labor (20.3%). With volunteer service valued at US$21.36 per hour, volunteers provided nearly US$173 billion of services during that year for the 8.1 billion hours of service they provided. Volunteerism is also a common activity in many other regions of the world, and is promoted by numerous organizations including the United Nations through United Nations Volunteers and the celebration of International Volunteer Day (Snyder and Omoto, 2008).

The psychological science of volunteerism has been a productive avenue of scholarly inquiry, not only exploring how volunteerism relates to other prosocial behaviors, but also examining the components of volunteerism that make it a distinctive and important topic worthy of study. Research to-date has considered factors that correspond to the antecedents of volunteerism, the experiences of volunteerism, and the consequences of volunteerism. These three stages of volunteerism have also been explored in studies of the volunteers themselves, the individuals aided by volunteers, as well as the organizations and communities through which volunteerism takes place. Research on these topics has approached the understanding of volunteerism from a variety of perspectives, including demographics, personality, identity, motivations, life experiences, and social contexts. The Volunteer Process Model proposed by Omoto and Snyder (Omoto and Snyder, 1995; Snyder and Omoto, 2008) is one structure for organizing

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22021-7

theory and research on volunteerism, particularly the separate factors that predict when people decide to volunteer and who gets involved in volunteering (antecedents), the variables important to volunteer satisfaction and continued involvement (experiences), and the impact volunteering has on the volunteers themselves, the recipients of volunteer services, and the larger community (consequences).

Antecedents of Volunteerism

A wide array of factors and explanations have been proposed to explain the when, where, why, and who of the initiation of volunteering. At a sociological level, three demographic variables are commonly cited as being tied to volunteering: age, education, and income (e.g., Snyder and Omoto, 2008). Those in middle age, from 35 to 54 years, are the most likely to volunteer, and people in their early twenties are the least likely to volunteer. College graduates are substantially more likely than those without college degrees to spend time volunteering. Income is also positively related to volunteering and, in conjunction with being employed, makes one more likely to volunteer. Other demographic variables that have been noted include gender (women are more likely to volunteer), marital status (married people more commonly volunteer), parental status (those with children are more involved), and race (whites tend to volunteer at the highest rates). Although these demographic differences in rates of volunteering have been found in the literature, it is important to note that they do not necessarily imply causal explanations for the differences.

Moving beyond demographics, numerous psychological approaches have been taken to the scientific study of volunteerism in attempts to discover the underlying foundations and causes of volunteerism. One such approach to understanding why people volunteer is the functional approach (e.g., Snyder and Omoto, 2008). 'Why' questions in psychology are fundamentally questions about motivation, and the functional approach provides answers to the question of why people choose to volunteer. Motivational approaches to human behavior are focused on the reasons people act, why they choose a specific course of action over others, and why they sustain a given behavior. The functional approach focuses on the needs that are being met by an action, the motivations that are addressed, and the functions served by engaging in volunteer behavior.

The functional approach to understanding why people volunteer recognizes that different people may have quite different reasons for engaging in the very same behaviors. This perspective has provided insight into the distinct reasons why people volunteer. Among the motivations of volunteers are expressing personal values, the desire for understanding, to increase self-esteem, to strengthen social ties, to build community, and as a way to improve career possibilities. Research has also discovered that by matching recruitment appeals to the motivations of individuals looking to volunteer, recruiters may effectively boost interest in and intent to volunteer (for a review of research on the motivations for volunteering, see Snyder and Omoto, 2008).

Personality has also been implicated in the initiation of volunteer behavior; specifically, those who are more prosocial in disposition are more likely to seek out and engage in helping behaviors in general, and volunteering in particular (e.g., Penner, 2002). Prosocial personality has often been split into two components: other-oriented empathy and general helpfulness. These two aspects of prosocial personality are not always highly correlated, and relate to different aspects of volunteering. For example, other-oriented empathy has been linked to the number of activities that a volunteer engages in within the context of a particular organization, as well as the amount of time dedicated to that organization and overall satisfaction with the volunteer position; by contrast, general helpfulness has been linked to the overall number of organizations that a volunteer assists.

The role of empathy has also been considered in volunteer contexts that involve members of different social groups. Volunteers often engage in activities that directly aim to benefit members of various groups of people (such as the poor, individuals with disabilities, and people living with human immunodeficiency virus infection/acquired immunodeficiency syndrome (HIV/AIDS)). Sometimes, volunteers intend to assist individuals who are part of the same social group (in-group members), such as when a person who grew up poor volunteers to teach children currently living in poverty. At other times, volunteers serve individuals who are not part of the same social group (out-group members), such as when a person without HIV/AIDS is a companion for an adult living with HIV/AIDS. Research has explored the predictors of volunteering to assist in-group versus out-group members (Stuermer et al., 2005). These findings demonstrate that people tend to volunteer to help in-group members out of empathy and feelings of perceived similarity, whereas volunteers tend to choose to aid out-group members because they having feelings of liking toward the out-group members. Although empathy has been linked to volunteering overall, it may be a better predictor of volunteering to help in-group members as opposed to assisting out-group members.

Social expectations from members of our social networks have also been implicated as an antecedent of volunteering. Drawing on sociological concepts of roles, Grube and Piliavin (2000) posited that the roles that people are expected to fulfill influence their behavior, and this relation has been found to hold true for volunteerism. Specifically, when their friends and associates expect them to volunteer and otherwise get involved in their communities, people tend to seek out and engage in those activities that fulfill these expectations. On the other hand, if members of one's social network do not seem to value giving back to one's communities, it is less likely that people will become involved in volunteer service.

Outside of social expectations, the past experiences of volunteers have also been tied to volunteer behavior. Adolescents with family members who volunteer are more likely to volunteer themselves, as are those coming from families higher in socioeconomic status (Johnson et al., 1998). In addition, teenagers with higher grade point averages and who are more integrated into their communities are more likely to be volunteers. Research on volunteerism has found that, regardless of age, simply having people in one's social network involved in volunteering is a reliable predictor of volunteer initiation, as well as already being involved in one's community in other ways (Omoto and Snyder, 2010).

In addition to these life experiences, there are indications that exposure to crises and traumatic events can lead people to engage in various prosocial behaviors, including volunteerism (for a review, see Omoto et al., 2012). For example, in the aftermath of the terrorist attacks of 11 September 2001, in the United States, there was a spike in prosocial behavior (e.g., Penner et al., 2005). Similar trends have been found for the earthquake that struck Chile in 2010, where rates of donating and volunteering increased in the wake of it (Gabrielsen et al., 2012).

Even when not in times of crisis, factors at the community level have been linked to volunteer behavior initiation (Omoto and Snyder, 2010). Often, volunteer efforts are tailored to the conditions and contexts of neighborhoods and the needs of communities. In fact, many volunteer efforts actually attempt to make changes to the communities in which they are embedded. In this sense, volunteering can be thought of both as an outgrowth of community and as a social force important to the building and improving of communities. One component of communities that has been addressed in empirical research is residential mobility, that is, the extent to which people frequently move in and out of the community. This community-level variable has been tied to prosocial and volunteer behavior, such that those communities that are more stable tend to see more prosocial behavior among its community members (Oishi et al., 2007).

One particularly distinctive type of community that often encourages active involvement can be found online. Web sites that host and share user-generated content are now commonplace on the Internet, with examples of such Web sites being the online encyclopedia Wikipedia and the advice Web site Yahoo! Answers. These sites are places that contain an extraordinarily large amount of content that is developed and added by volunteers without any form of payment. This form of volunteering, often called 'microvolunteering,' has almost everything in common with traditional understandings of volunteerism but instead take place through online communities and organizations with an online presence. For example, recent research on the free movie review Web site MovieLens has explored the relation between typical predictors of volunteer behavior in order to understand the generation of volunteer user content on the MovieLens site (Fuglestad et al., 2012). Results indicate that there is a relation between other-oriented motivations (wanting to help others and the community) and various volunteer activities on the Web site, such as reviewing movies, editing movie information, and responding to questions on the forum board. Those with a history of community involvement and prosocial behavior in real life were also more likely to volunteer on the site. Finally, motivations specifically relating to joining the MovieLens site predicted different patterns of behavior such that utilitarian and personal entertainment reasons for joining related to rating movies but social motivations for joining related to more social volunteer behaviors such as interacting on the forum board.

Clearly, a wide range of factors play important roles in the initiation of volunteer behavior, including demographics, motivations, personality, identity, past experiences, and the community context. Many of these factors have also been implicated at subsequent stages of the volunteer process, as determinants of, for example, the satisfaction of volunteers and their sustained service over extended periods of time (e.g., Snyder and Omoto, 2008).

Experiences of Volunteers

As volunteers continue their service, new factors become relevant to understanding key facets of volunteerism such as volunteer satisfaction and continued volunteer behavior. Factors commonly considered in the literature include the matching of experiences to motivations, development of identity, and creation of community connections. Some of the earliest research on the experiences of volunteers comes from longitudinal studies of volunteer companions for individuals living with HIV/AIDS. Many of these volunteers had high expectations for the quality of the helping relationships between themselves and the individuals with HIV/AIDS, and yet many volunteers later felt that these relationships fell short of their own expectations (Omoto et al., 1998). This inability to have experiences that met their expectations left many volunteers feeling unsatisfied by their volunteer positions.

Research has also revealed that those volunteers who feel that they are gaining the experiences they desired when they started volunteering, and particularly those who feel there is a match between their motivations for volunteering and actual volunteer experiences, were most satisfied as volunteers and had stronger intentions to continue volunteering over time (see Snyder and Omoto, 2008). Some work has also demonstrated that, although people tend to start volunteering for other-oriented motivations, such as values motivations, self-oriented motivations have been found to more closely relate to overall length of volunteer service (Omoto and Snyder, 1995).

Volunteer satisfaction and behavior maintenance have also been tied to the extent to which people self-identify as a volunteer. The effects of identity have been considered in the Role Identity Model (Grube and Piliavin, 2000), which posits that role identity forms because of social expectations; that is, when people believe that others who are important to them want and expect them to engage in a certain behavior, they will develop a particular personal identity relevant to that behavior. This process has been found to hold true for volunteer role identity. Those high in volunteer identity consider volunteering a core component of their sense of self, and derive a portion of their self-esteem from involvement in volunteer service. The Role Identity Model has been linked to the experiences of volunteers over time, particularly how the forming of identity relates to satisfaction and future intentions. The extent to which volunteers form an identity as a volunteer and come to see the behavior of volunteering as important to who they are as a person, the more they are satisfied with their position, the greater the number of hours served, and the greater their intentions to continue to volunteer. Role identity has been linked to other, related prosocial behaviors such as blood and organ donation, but has proven particularly useful for understanding volunteer behavior and experiences.

The role of social expectations during required forms of volunteering, such as service learning in educational settings, also impacts individual volunteer behavior maintenance. Service learning programs that mandate volunteering as part of the curriculum are often designed to encourage positive contributions to a local community at the same time as they instill values of helping others into participants in service learning. However, research indicates that mandating volunteering does not always lead to future volunteering for students. For example, students in service learning classes who perceive the volunteering component of the class as having been under

their own control and intrinsically motivated are more likely to have their intentions to continue volunteering solidified by their service learning experiences than students who perceived their service learning volunteering as something imposed on by an external force (Stukas et al., 1999).

As reviewed by Omoto and Snyder (2010), although matching experiences to motivations and the forming of identity have been linked to volunteer satisfaction and volunteer behavior maintenance, a number of other factors have been linked to outcomes such as organizational support and organizational structure. It has been demonstrated that there are important factors related to the volunteer tasks themselves in regards to predicting volunteer behavior maintenance, such as task achievement of volunteers, the degree to which the volunteer tasks are challenging, and how involved the volunteer is in the organization and in the community. Finally, the existence and ongoing development of time conflicts between personal time and time spent volunteering has been recognized as an important contributor to volunteer behavior maintenance, such that having too many demands of time and energy decreases the likelihood that volunteers will stick with their current service position. Numerous organizational components and experiences have been linked to volunteer behavior maintenance.

Considering the community component of volunteer experiences, volunteers tend to form social and community ties over time as they get to know the recipients of help, other volunteers in the organization, and the staffs of the various organizations in which they serve (Omoto and Snyder, 2010). For example, volunteers who serve as companions for those living with HIV/AIDS tend to meet members of the social networks of the people they are helping, and in return the recipient of the volunteer's assistance also get to know members of the volunteer's community. In fact, the more that volunteers and the people they are helping become involved in each other's social networks, the happier and healthier the recipients of volunteer services tend to be.

However, Snyder and Omoto (2008) have also highlighted the ways in which experiences during the volunteer process are not always positive. Volunteering has been linked to negative impacts on individual's social networks and the individual's own life. It has been shown that sometimes volunteering with a stigmatized group, such as those living with HIV/AIDS, can lead to disruptions in the volunteer's social support. Members of the volunteer's social networks, including family and friends, might shun them or express disapproval of volunteer behavior aimed at supporting traditionally stigmatized groups. These volunteers have reported feeling uncomfortable and stigmatized themselves by their own social networks for helping individuals living with HIV/AIDS. These stigmatizing experiences predict early termination of volunteer service.

Factors that relate to motivations, expectations, identity, and community have all been linked to experiences of volunteerism, including satisfaction with volunteering and intentions to continue to volunteer over time. The question of how volunteering impacts volunteers themselves, the people they help, and the organizations and communities in which they volunteer is addressed in the consequences stage of the Volunteer Process Model.

Consequences of Volunteerism

Research on the consequences of volunteerism has considered its effects on the volunteers themselves, the recipients of volunteer services, and the wider community. Snyder and Omoto (2008) review outcomes that include changes in attitudes and knowledge, improved health and subjective well-being, changes in behavior, and the establishment of community bonds. For example, studies of volunteers who work with people living with HIV/AIDS have revealed that such volunteerism leads to increases in knowledge of safer sex practices, reductions in stereotypical beliefs about individuals with HIV/AIDS, and increases in comfort in discussing issues relevant to HIV/AIDS. Studies of consequences in other volunteer domains have found increases in volunteer self-esteem, self-efficacy and confidence, and even improved academic achievement. In addition, volunteering has been linked to positive health outcomes for volunteers, including increased optimism and longer life, and has been linked to higher subjective well-being for those volunteers who have experienced past traumatic life events.

As reviewed by Snyder and Omoto (2008), the benefits of volunteering extend to the actual recipients of volunteer service, and to the organizations through which volunteers serve. For example, individuals living with HIV/AIDS who have a volunteer providing companionship and home help have been shown to have better psychological functioning as compared to those without a volunteer companion. However, more research needs to be conducted on the impact of volunteer help on aid recipients, to determine, for example, whether receiving volunteer assistance might be threatening to recipients' self-esteem or self-efficacy. More research is also needed on the positive impacts of volunteerism on service organizations, to build upon the aforementioned economic benefits to organizations and the growth of social capital.

The building of community bonds and social networks occurs for volunteers, help recipients, and the member of the organizations through which volunteers serve (Omoto and Snyder, 2010). These community bonds have been linked to not only volunteer behavior maintenance, but also the willingness of community members to engage in other efforts aimed at improving the quality of life for the community, including the amount of money donated in support of causes, attending fund-raisers, and engaging in civic and political activism. In addition, these social networks increase the overall effectiveness of volunteers, as well as serve as means to recruit future volunteers.

Thus, psychological research on the consequences of volunteerism has documented benefits for individual volunteers, the recipients of volunteer services, the organizations through which volunteers serve, and the wider community. For volunteers and recipients, these include health and psychological benefits, as well as increased social capital. For organizations and communities, the benefits include improved capability to address problems and the ability to keep volunteers involved for longer periods of time.

Theoretical and Practical Implications

The lessons that have been learned from the psychological research on volunteerism can be used to assist organizations and communities in their efforts to create successful volunteer programs. Research can inform attempts to improve the recruitment, training, support, and retention of volunteers by guiding the development of recruitment materials, the framing

of volunteer positions, the creation of volunteer activities, and the types of feedback and experiences provided to volunteers. Furthermore, new possibilities for application could be gleaned from findings that relate to the role of matching motivations to messages and experiences, the forming of volunteer identity, building sense of community, and providing appropriate organizational support and volunteer opportunities, all of which could be incorporated into systematic efforts to improve the functioning of volunteer programs around the world. One possibility might be to have organizations provide workshops that intentionally build volunteer identity and sense of community within the organization. Insights from the psychology of volunteerism can be used to help address any number of important and pressing social issues, from poverty in local communities to problems that affect the entire planet like global climate change. The extent to which an issue can be affected and addressed by volunteers implicates the relevance of the psychology of volunteerism.

On a more theoretical level, the extent to which volunteerism can be linked to other prosocial behaviors can benefit the overall understanding of prosocial behavior by allowing the sharing of knowledge between domains of study. The tighter the link between volunteerism and forms of helping, the more useful models of volunteer behavior can be for understanding other forms of helping. The determining of the conditions under which volunteer behaviors are more or less likely to occur can improve the prediction of other prosocial behaviors. For example, by adopting the motivational approach used in research on volunteerism, perhaps studies of helping could consider helper motivation in order to determine the most effective ways for those in need of help to elicit assistance in work or school settings. The usefulness of research on volunteerism holds true not only for helping behaviors that, such as volunteering, are planned, sustained, and lack an obligation, but insights into volunteer behavior also have the potential to inform behaviors that are less commonly researched. These sorts of prosocial behaviors include the informal helping and emotional support one might receive from a barber while getting a haircut, or organizational citizenship behaviors such as whistle blowing. The usefulness of models of volunteer behavior, and the antecedents, experiences, and consequences of volunteerism, could also be applied to realms of behavior that are not always considered strictly prosocial behaviors, such as pro-environmental behaviors like recycling and conservation. Finally, hybrid national service programs, such as the Peace Corps and AmeriCorps, which straddle the line between work and volunteering, also stand to benefit from the knowledge generated by the psychological study of volunteer behavior.

In conclusion, psychological research has uncovered important contributors to the antecedents, experiences, and consequences of volunteerism for the volunteers serving, the recipients of help, and the organizations and communities through which volunteerism takes place. Contributors examined include demographics, personality, motivations, identity, past experiences, and the community context. These factors can help inform theoretical and practical extensions of the research, both in volunteerism in particular, and the wider prosocial literature.

See also: Altruism and Prosocial Behavior, Sociology of; Collective Action; Collective Behavior, Social Psychology of; Helping Intergroup Relations; Levels of Analysis in Social Psychology; Motivation and Actions, Psychology of; Personality and Self; Prosocial Behavior and Empathy; Prosocial Behavior, Cultural Differences In; Social Identity in Social Psychology; Social Movements: A Social Psychological Perspective; Social Psychology; Voluntary Associations, Sociology of.

Bibliography

Corporation for National and Community Service, 2011. Research Brief: Volunteering in America Research Highlights. Available from: http://www.volunteeringinamerica.gov/national (12.05.12.).

Fuglestad, P.T., Dwyer, P.C., Filson Moses, J., Kim, J.S., Mannino, C.A., Terveen, L., Snyder, M., 2012. What Makes Users Rate (Share, Tag, Edit…)? Predicting Patterns of Participation in Online Communities. Paper presented at the Association for Computing Machinery Conference on Computer Supported Cooperative Work. Association for Computing Machinery, Seattle, WA.

Gabrielsen, S., Maki, A., Dwyer, P.C., Snyder, M., Gonzalez, R., Cortes, R., Lay, F., Herrera, P., 2012. Societal Responses to Natural Disasters: Helping in the Aftermath of the 2010 Chile Earthquake. Poster presented at the International Research Conference at the University of Minnesota. University of Minnesota, Minneapolis, MN.

Grube, J., Piliavin, J.A., 2000. Role-identity, organizational experiences, and volunteer performance. Personality and Social Psychology Bulletin 26, 1108–1119.

Johnson, M.K., Beebe, T., Mortimer, J.T., Snyder, M., 1998. Volunteerism in adolescence: a process perspective. Journal of Research on Adolescence 8, 309–332.

Oishi, S., Rothman, A.J., Snyder, M., Su, J., Zehm, K., Hertel, A.W., Gonzales, M.H., Sherman, G.D., 2007. The socioecological model of procommunity action: the benefits of residential stability. Journal of Personality and Social Psychology 93, 831–844.

Omoto, A.M., Gunn, D.G., Crain, A.L., 1998. Helping in hard times: relationship closeness and the AIDS volunteer experience. In: Derlega, V.J., Barbee, A.P. (Eds.), HIV Infection and Social Interaction. Sage, Thousand Oaks, pp. 106–128.

Omoto, A.M., Snyder, M., 1995. Sustained helping without obligation: motivation, longevity of service, and perceived attitude change among AIDS volunteers. Journal of Personality and Social Psychology 68, 671–686.

Omoto, A.M., Snyder, M., 2010. Influences of psychological sense of community on voluntary helping and prosocial action. In: Stuermer, S., Snyder, M. (Eds.), The Psychology of Prosocial Behavior: Group Processes, Intergroup Relations, and Helping. Wiley-Blackwell, Malden, pp. 222–243.

Omoto, A.M., Snyder, M., Hackett, J.D., 2012. Everyday helping and responses to crises: a model for understanding volunteerism. In: Jonas, K., Morton, T. (Eds.), Restoring Civil Societies: The Psychology of Intervention and Engagement Following Crisis. Wiley-Blackwell, Malden, pp. 98–118.

Penner, L.A., 2002. The causes of sustained volunteerism: an interactionist perspective. Journal of Social Issues 58, 447–467.

Penner, L.A., Brannick, M., Webb, S., Connell, P., 2005. The effects of the September 11 attacks on volunteering: an archival analysis. Journal of Applied Social Psychology 35, 1333–1360.

Snyder, M., Omoto, A.M., 2008. Volunteerism: social issues perspectives and social policy implications. Social Issues and Policy Review 2, 1–36.

Stuermer, S., Snyder, M., Omoto, A.M., 2005. Prosocial emotions and helping: the moderating social of group membership. Journal of Personality and Social Psychology 88, 532–546.

Stukas, A.A., Snyder, M., Clary, E.G., 1999. The effects of 'mandatory volunteerism' on intentions to volunteer. Psychological Science 10, 59–64.

von Hayek, Friedrich A. (1899–1992)

Viktor Vanberg, Walter Eucken Institute, Freiburg, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

F.A. von Hayek is widely regarded as one of the most important economists of the twentieth century, as the most influential intellectual figure in the modern revival of classical liberalism, as a preeminent exponent of the Austrian School of Economics, and as the most articulate critic of socialist doctrines. He also deserves recognition as one of the preeminent social philosophers of the past century. In an age of increasing specialization within the social sciences and the humanities, he developed an approach to social analysis and social reform that integrates insights from economics, law, politics, philosophy, psychology, and evolutionary theory into an impressively coherent outlook at the foundations and evolution of social order. Hayek's scholarly writings span more than six decades and include numerous books and articles.

F.A. von Hayek is widely regarded as one of the most important economists of the twentieth century, as the most influential intellectual figure in the modern revival of classical liberalism (Kukathas, 1990), as a preeminent exponent of the Austrian School of Economics (Caldwell, 2004: pp. 17ff.), and as the most articulate critic of socialist doctrines (Caldwell, 1997). He also deserves recognition as one of the preeminent social philosophers of the past century. In an age of increasing specialization within the social sciences and the humanities, he developed an approach to social analysis and social reform that integrates insights from economics, law, politics, philosophy, psychology, and evolutionary theory into an impressively coherent outlook at the foundations and evolution of social order. Hayek's scholarly writings span more than six decades and include numerous books and articles, republished in *The Collected Works of F.A. Hayek* (1988).

Biography

Hayek was born in Vienna on 8 May 1899. Returning from service in the Austro-Hungarian army in World War I, Hayek enrolled at the University of Vienna and earned doctorates in Law (1921) and in Political Economy (1923). Interrupted by a year of study at New York University (1923–24), he worked from 1921 to 1926 as a legal consultant in a government office in Vienna, directed by his mentor Ludwig von Mises. Hayek's reading of Mises' (1922) penetrating critique of socialism and his participation in Mises' famous *Privatseminar* had profound influence on his thinking. Together with Mises he founded, in 1927, the Austrian Institute for Business Cycle Research whose director he was until 1931. When in 1929 Hayek became *Privatdozent* for Political Economy at the University of Vienna his *Habilitation* lecture, 'Gibt es einen Widersinn des Sparens?' (Hayek, 1929); English translation: 'The Paradox of Saving,' (1931a), in which he criticized the theory that business fluctuations are caused by 'underconsumption,' caught the interest of Lionel Robbins, the recently appointed head of the economics department at the London School of Economics (LSE). Recognizing that Hayek's critique was applicable to ideas that John Maynard Keynes had begun to promote in the late 1920s, Robbins invited Hayek to present a series of lectures on the Austrian perspective on business cycles at the LSE in early 1931. The lectures, subsequently published as *Prices and Production* (1931), were so well received that Hayek was offered the Tooke Chair of Economic Science and Statistics at the LSE (Ebenstein, 2001: pp. 172ff.), an appointment he held from 1932 to 1950. In 1950, he left England for the United States in order to take the position of Professor of Social and Moral Sciences at the University of Chicago's Committee on Social Thought. After retiring from Chicago in 1962, Hayek taught as Professor of Economics at the University of Freiburg, Germany, until 1969, when he accepted a guest professorship at the University of Salzburg. During his Salzburg tenure, in 1974, he was awarded the Nobel Prize in Economics. Yet, finding the expectations he harbored in returning to his native Austria disappointed, Hayek decided to return in 1977 as Professor Emeritus to Freiburg, where he spent the remaining years of his life. He died on 23 March, 1992. He is buried in Neustift am Wald, Vienna.

Economics and the 'Knowledge Problem'

Hayek (1967: p. 91) has said that the development of his thought was led 'from technical economics into all kinds of questions usually regarded as philosophical.' It was his early work on monetary theory, capital theory, and business cycle theory that earned him his appointment at the LSE, and it was about these themes that he engaged in controversies with John Maynard Keynes, Piero Sraffa and Frank H. Knight in the early 1930s. Although his 'Austrian' perspective on economic issues initially attracted considerable attention and earned him a reputation as the principal counterweight to Keynes, his fame began to decline in the mid-thirties. When he finally published his treatise *The Pure Theory of Capital* in 1941 it was so much overshadowed by the then dominant Keynesianism that it went almost unnoticed by the economics profession.

It was not only his disappointment at the declining attention he found in the profession that motivated Hayek to move away 'from technical economics.' An important turning point in his scholarly focus occurred in the mid-1930s, when he worked on an English edition of selected contributions to the 'socialist calculation debate,' which included Mises' early (1920) article

on the subject (Hayek, 1935). As Hayek noted in retrospect, when preparing this edition he began to develop a growing interest in the nature and unavoidable limitations of human knowledge; and he addressed the issue of what these limitations implied for theoretical efforts at understanding the social world and practical efforts at organizing social life. The concern with these questions, often referred to as the 'knowledge problem,' was to become the core theme of Hayek's lifework.

In Hayek's own assessment, the critical step in shifting his attention toward the broader issues of social theory and political philosophy was his 1937 article on 'Economics and Knowledge' (1948: Chapter 2), an article in which he addressed what he described as the 'central question of all social sciences,' the question of how the fragments of imperfect knowledge that exist dispersed in individual minds are coordinated in the social process to allow for productive cooperation among multitudes of persons. In his 1945 often cited article on 'The Use of Knowledge in Society' (1948: Chapter 4), Hayek restated and extended his earlier argument, stressing that the "problem of what is the best way of utilizing knowledge initially dispersed among all the people is at least one of the main problems of economic policy – or of designing an efficient economic system" (1948: pp. 78f.). In these two articles, which he considered his most original contributions to economics, Hayek developed a fundamental critique of two bodies of thought prevalent at the time – formal equilibrium economics on the one hand and socialist concepts of central planning on the other – censuring both of them for failing to account for the theoretical and practical difficulties that are posed by the imperfect, fragmentary, and dispersed nature of human knowledge.

In Hayek's view, the models of mathematical equilibrium economics provide only a pseudo-answer to the question that they are meant to illuminate, namely, how markets work. As he put it, with the "assumption of a perfect market where every event becomes known instantaneously to every member" (1948: p. 45), formal equilibrium analysis begs the very question that an empirically meaningful theory of market processes would have to answer, namely how "a solution is produced by the interaction of people each of whom possesses only partial knowledge" (1948: p. 91). Ironically, the failure committed in concepts of central planning is of a very similar kind. By assuming perfect knowledge on part of the planners, advocates of central planning provide only a pseudo-answer to the practical problem of how effective coordination of economic activities can be achieved in the absence of markets. The assumption that all relevant knowledge is available to, and can be administered by, a planning agency, raises the question of how such an agency is supposed to come to terms with the fact that much of the relevant knowledge "never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete … knowledge which all the separate individuals possess" (1948: p. 77), i.e. knowledge that can be activated only by these very individuals.

What is needed, Hayek argued, is a theory that accounts for "the unavoidable imperfections of man's knowledge and the consequent need for a process by which knowledge is constantly communicated and acquired" (1948: p. 91). He called on his colleagues in the economics profession to turn their attention from the formal statements of equilibrium analysis to empirical "propositions about what happens in the real world" (1948: p. 46), propositions about how knowledge is actually acquired and communicated. Whereas his call did not find much response in the profession for a long time, in his own work Hayek himself pursued the research agenda that he had suggested. He specifically addressed the issue of how we acquire knowledge in his work *The Sensory Order – An Inquiry into the Foundations of Theoretical Psychology* (1952a), a most unlikely book for an economist and one of his least known works. However, it is, of great systematic significance for his entire work. In this book – which was in fact based on an early unpublished manuscript from Hayek's student days (Hayek, 2006[1920]) – Hayek developed a theory of the human mind, the essence of which, as has been recognized by renowned experts in the field (e.g., Fuster, 1995: pp. 87ff.), is in amazing accord with modern cognitive and neurological science. Seeking to explain how the human mind establishes its connections "between the input of (external and internal) stimuli and the output of action" (1982: p. 288), Hayek interpreted the working of the mind as an adaptive process of constant classification and reclassification on many levels (sensory perceptions, emotions, concepts), a process in which conjectural internal models of the outside world are adaptively formed and guide actions.

The issue of how knowledge is communicated and advanced in society is at the core of Hayek's theory of the market as a communication system (1976: Chapter 10) and of competition as a discovery procedure (1978: Chapter 12) – a theory that formed the centerpiece of his work. The prices formed in the competitive market process, Hayek argued, signal to all market participants the relevant general information about changes in relative scarcities that they need to know about in order to adjust their behavior adequately, utilizing their unique local knowledge of the particular circumstances of time and place. This type of knowledge could never be communicated to, or utilized by, a central planning agency. Beyond serving thus as a device for utilizing dispersed knowledge, the market serves also as an arena for the continuous competitive exploration of new and potentially better solutions to economic problems, thus inducing the discovery and creation of new knowledge.

Spontaneous Order and Evolution

The relevance of Hayek's work, not only for economics but also for the social and behavioral sciences more generally, is most clearly visible in his theories of spontaneous social order and cultural evolution, which he developed as generalizations and extensions of his arguments on market processes and competition. The thrust of these theories is that human societies are highly complex phenomena and that the processes of social coordination and societal evolution are inappropriately simplified and misconceived if they are approached in the spirit of the mechanistic models of physics (1952b). Criticizing as 'scientism' the misguided ambition to model the social sciences in the image of physics, Hayek argues that "when we turn from mechanical to such 'more highly organized' or essentially complex phenomena as we encounter in the realms of life, mind and society" (1973: p. 41), all we can reasonably hope to

achieve are *explanations of the principle* and *pattern predictions*, rather than explanations or predictions of specific events (1967: Chapter 2). Hayek added that the 'hard' sciences would also be subject to these limitations as their advance "penetrates further and further into complex phenomena" (1967: p. 20). Hayek argued that the theory of evolution by natural selection exemplifies in a paradigmatic manner what is true for all theories of complex phenomena, – namely that they provide "an account of a process, the outcome of which will depend on a very large number of particular facts, far too numerous for us to know in their entirety" (1967: p. 23). He posited that even if we know the general laws according to which complex phenomena operate, our lack of knowledge of all the particular facts by which they are affected prevents us from obtaining anything more than an explanation and prediction of "the abstract pattern the process will follow" (1967: p. 23).

Hayek placed his own social theoretical outlook in the tradition of the Scottish moral philosophers (notably Adam Ferguson, David Hume, and Adam Smith), whom he credited with the discovery of "the twin concepts of the formation of spontaneous orders and of selective evolution" (1988: p. 146). Central to this paradigmatic tradition is a methodological individualism and subjectivism, that is, on the one hand, the notion that social aggregate phenomena are, ultimately, to be explained in terms of the actions and interactions of individual human beings and, on the other hand, the insight that "the analysis of what people do can start only from what is known to them" (1948: p. 44); that is, from their subjective view of the world. Equally central is the notion that the productive coordination and division of labor that we observe in society, as well as the rules and institutions on which societal cooperation is based are not due to deliberate planning and conscious design, but to forces of spontaneous mutual adjustment and evolution. From these premises, Hayek drew important conclusions on the limits of our ability to improve our social condition by deliberate construction.

According to Hayek there are, essentially two categorically different ways in which the actions of individuals can be socially coordinated – either through spontaneous mutual adjustment within a framework of general rules of conduct or through the commands of a central 'planner' (1973: Chapter 2). The respective merits and limits of these *two kinds of order* – *spontaneous order* and *organization* (Vanberg, 1982) – are rooted in the nature of the principal means by which the coordination of actions is achieved, namely by general rules in case of the former and specific commands in case of the latter. Command-based organization is an effective method of social coordination where specific, limited purposes are to be achieved. Because of its reliance on central coordination it is, however, severely limited in its capacity to utilize the knowledge dispersed among its individual participants. By contrast, the advantage of rule-based spontaneous social orders is that, by allowing individuals to pursue their own ends within the general confines defined by the rules, they can utilize and activate localized, dispersed knowledge to a much higher degree. The price that has to be paid for their greater capacity for utilizing knowledge is that the *particular* outcomes that they generate must inevitably remain largely unpredictable. In Hayek's view, the respective merits and limits of the two kinds of order should lead us to conclude that, while organization can be a useful method for solving a multitude of specific, limited tasks, it cannot possibly be appropriate for the extended order of an economy or a society at large. There is ample room for organizations within the spontaneous order of society, but efforts to turn society itself into an organization can only end in disaster.

An important element in Hayek's theory of spontaneous social order is the insight that, while the particular outcomes of a rule-based order remain necessarily largely unpredictable, the general pattern of its outcomes is determined systematically by the nature of its rules. There exists, Hayek emphasizes, a systematic interplay between the 'order of rules' and the 'order of actions' (1967: Chapter 4) – between the nature of the 'rules of the game' on which the interactions of the participating individuals are based and the nature of the action-patterns that emerge from the individuals' behavioral choices within the rules. To the extent that we understand how the *order of rules* affect the *order of actions* we can predict the general nature of the behavioral patterns that will emerge within particular rule-frameworks, just as an understanding of the rules of a sport allow us to predict, if not particular outcomes, the general pattern of moves that we can expect to observe on the playing field. Hayek argued that in fact, our knowledge of the systematic relationship between the order of rules and the order of actions is the primary instrument that we can employ to improve our condition. This knowledge enables us to choose 'rules of the game' that can be expected to generate more desirable overall patterns of outcomes in the various social arenas in which we interact with each other.

Hayek's emphasis on the basic fact of 'the dispersion and imperfection of all knowledge' led him to be highly critical of the mind-set that he labeled 'rational constructivism' or 'constructivist rationalism' (1978: Chapter 1). Under this rubric, he subsumed approaches to issues of social policy and social reform whose recommendations are based on a 'pretence of knowledge' (1978: Chapter 2), that is, on unjustified presumptions concerning the extent of our knowledge and our intellectual powers. The primary targets of Hayek's critique were approaches that, like socialist concepts of central planning, amount to the "claim that man can achieve a desirable order of society by concretely arranging all parts in full knowledge of all the relevant facts" (1967: p. 88). Against such claims, Hayek pointed out that, because it is simply impossible for us to take "conscious account of all the particular facts which enter into the order of society" (1973: p. 13), we need to rely on general rules "as the means to create order in social affairs" (1948: p. 19), and our efforts at building a better society must largely rely on the indirect method of improving the rules and institutions under which we live, instead of on attempts to directly produce desired outcomes through specific interventions.

Hayek stressed that in our efforts to improve the framework of rules and institutions under which we live, we need to be careful to avoid the pitfalls of a pretense of knowledge. Since the desirability of the outcome patterns that particular rules or institutions tend to generate can only be judged appropriately in terms of their working properties over extended periods of time, it would be presumptuous to claim that we can easily design, at the drawing board, rules and institutions that serve us better

than those that have evolved over generations. Just as the process of competition in ordinary markets, by allowing alternative potential solutions to be tried out, helps us to discover how our various 'economic' problems can be solved better than before, so competition in the realm of rules and institutions can serve the function of testing the kinds of rule arrangements that provide more conducive conditions for mutually beneficial cooperation among multitudes of persons. A process of competitive selection can here, just as in ordinary markets, help gradually to evolve solutions that are much more effective than those that any rational design could have ever invented at the outset. In fact, so Hayek argues, the institutions on which our civilization rests, including the rules of the market, should be viewed as the outcome of the process of cultural evolution that extended over innumerable generations and incorporated the experience of many more trials and errors than the wisest and most intelligent rule designer could ever consider. Our efforts at institutional reform and rule design should, therefore, always be informed by a respect for the potential wisdom implicit in inherited institutions, and we should expect further improvement "more from slow experimental piecemeal evolution than from redesign of the whole" (1964: p. 8).

The theory of cultural evolution is the theme to which Hayek devoted much of his later work (see Epilogue in 1979; 1988), and it is this part of his theory that has become a primary focus of discussions on his ideas. A central issue in these discussions is how the notion of competition among institutions – a notion that is at the heart of Hayek's ideas on cultural evolution – is to be specified; under what conditions can such institutional competition be expected to generate 'good' rules, that is, rules with working properties that are desirable for those to whom they apply (Vanberg, 1994a: Chapter 5; 1994b).

Liberalism and Democracy

A central theme in Hayek's political philosophy is the relation between liberalism and democracy (Vanberg, 2011). Hayek pointed out (1978: p. 142) that by its "insistence on a law which is the same for all and the consequent opposition to all legal privilege," liberalism was originally closely allied with the democratic ideal of equal rights in political participation, so that in "the struggle for constitutional government in the nineteenth century the liberal and the democratic movement indeed were often undistinguishable" (1978: p. 142). In Hayek's account, the two movements parted company after the victory of democracy, when the erroneous belief came to prevail "that the adoption of democratic procedures made it possible to dispense with all other limitations on government power" (1979: p. 3), thus fostering a concept of democracy that regards "current majority opinion as the only criterion for the legitimacy of the powers of government" (1978: p. 143). The organizational form in which this concept has become institutionalized (which he describes as 'unlimited democracy') is in Hayek's view incompatible with the liberal ideal of limited government according to which "it is not from a mere act of the momentary majority but from a wider agreement on common principles that majority decision derives its authority" (1960: p. 106). It is also in conflict with the "true content of the democratic ideal" (1979: p. 5) because it inevitably results in a policy that instead of serving the *common interests of the citizenry*, gets entrapped in being lobbied by special interests and granting privileges to them. "The reason is," Hayek noted, "that democratic government, if nominally omnipotent, becomes as a result of its unlimited powers exceedingly weak, the playball of all the separate interests it has to satisfy to secure majority support" (1979: p. 99).

As a remedy for the deficiencies of unlimited democracy Hayek advocates a *constitutional conception of democracy* that insists on the role of a *limiting constitution*, i.e., on the understanding that all democratic power derives from "the consent of the people" (1979: p. 3), and its legitimacy rests "in the last resort on the approval by the people at large of certain fundamental principles underlying and limiting all government" (1979: p. 35). Of utmost importance are, in Hayek's assessment, constitutional limitations that prevent discriminatory politics by effectively restricting the authority of government and the legislature to grant special privileges to interest groups. Hayek argues that rather than weaken the effective power of the democratic state, such restrictions would strengthen its ability to fulfill its true task, namely to advance the *common interest* of its citizens. As a significant step toward an effective limiting constitution Hayek advanced a proposal for constitutional reform (1979: pp. 105ff.) that is based on the diagnosis that a major defect in the prevailing institutional structure of democracy must be seen in the fact that one and the same representative body, the parliament, has been entrusted with two fundamentally different tasks – on the one hand, to establish the 'rules of the game,' or the laws on which society and state are based and, on the other hand, to direct the day-to-day operations of the particular government. As Hayek charges, the inevitable effect of assigning these two tasks to one and the same assembly has been "that the supreme governmental authority became free to give itself currently whatever laws helped it best to achieve the particular purposes of the moment" (1979: p. 101). In this short-sighted approach to legislation, he sees the main reason why such an assembly must become the target of the lobbying efforts of interest groups and finds itself pressured to "use its power to satisfy the demands of sectional interests" (1978: p. 115).

According to the 'model constitution' that Hayek proposed, the two noted tasks are to be assigned to two strictly separated representative assemblies, with the separation to be institutionalized in a manner that can effectively "prevent collusion of the legislative with a similarly composed governmental assembly, for which it would be likely to provide the laws which that assembly needed for its particular purpose" (1978: p. 160). Hayek's ambition to specify in considerable detail the modes in which elections to the legislative assembly ought to be organized had the unfortunate effect that critical reactions to his constitutional proposal almost exclusively concentrated on its subordinate details; whereas his primary concern, the effective separation between genuine legislation and day-to-day government, did not find the attention it surely deserves.

Impact and Significance of Hayek's Work

With his arrival at the LSE in the early 1930s, Hayek became almost instantaneously a well-recognized figure in the

economics profession, engaged in prominent controversies on economic theory with such opponents as John Maynard Keynes and, on the issue of rational economic calculation under socialism, with Oskar Lange and others. When his involvement in the latter debate, induced the shift in his interest toward general issues of social and political philosophy, he increasingly removed himself from standard professional discourse. He turned his attention to a fundamental critique of the philosophical foundations and the political implications of socialist doctrines. This was reflected most prominently in the publication of his most famous book *The Road to Serfdom* in 1944, a book that quickly made him a widely known public figure, in particular in the United States, where a condensed version appeared in the *Reader's Digest* in 1945. While the book earned him public fame, it was met much less favorably by his colleagues in the economics profession, many of whom were, at the time, much more inclined to interventionist views than to Hayek's warnings against socialist concepts of social and economic planning. Feeling the need to form a closer network among the dispersed and often isolated advocates of classical liberalism, Hayek in 1947 brought together a group of like-minded colleagues from various disciplines to a conference at Lake Geneva in Switzerland, which was to become the founding meeting of the Mont Pèlerin Society.

With the rise of Keynesianism in the 1950s and 1960s, Hayek's work fell almost into oblivion, in academia as well as in public debate. When in 1960 his major restatement of the principles of classical liberalism, *The Constitution of Liberty*, appeared, it was widely regarded, even by sympathetic observers, as an outdated treatise, out of step with modern economic and political thought. It came as a surprise to many, including Hayek himself, when in 1974 he was awarded the Nobel Prize in Economics. Even if the fact that he had to share the prize with Gunnar Myrdal (an economist of a quite different orientation) reflected significant ambivalence on part of the Nobel committee, the award was an early indicator of a reawakening interest in classical liberal thought in general, and Hayek's work in particular; an interest that grew significantly in the latter half of the 1970s, as a response to an increasing disillusionment with Keynesian economic policies and the accumulating evidence of the failure of 'real socialism.' In fact, Hayek became very much the symbol of what has been described as the 'modern rebirth of classical liberalism.' His work regained increasing attention in the academic community, his ideas influenced politicians like Margaret Thatcher and Ronald Reagan, and his publications found growing interest among dissidents and reformers in the disintegrating Communist world. Hayek lived long enough to witness the final collapse of the Communist regimes in Europe as a confirmation of his lifelong critique of socialist constructivism.

In modern theoretical economics, Hayek's ideas play an important role in a number of approaches that seek to advance an alternative theoretical outlook to the orthodox equilibrium paradigm. His ideas on the nature of human knowledge and learning are finding increasing attention among scholars who seek a corrective for the deficiencies of the traditional rational choice model. Developments in evolutionary economics, in the new institutional economics, and in the economics of complex adaptive systems, either explicitly draw on, or give support to, core ideas of his work, to his theory of the market as a spontaneous order, his outlook on competition as a discovery procedure, and his evolutionary approach to institutions.

Even if the reception of Hayek's work in social sciences other than economics remains quite limited, his approach to social theory may well come to play a significant role as a paradigmatic contribution toward a theoretically integrated social science. It was one of Hayek's major concerns that the disciplinary fragmentation of modern social science left what should be its central issue, namely "the problem of an appropriate social order," without adequate attention. As he put it, even though this issue "is studied today from the different angles of economics, jurisprudence, political science, sociology, and ethics, the problem is one which can be approached successfully only as a whole" (1973: p. 4). More than any of his contemporaries in the social sciences, Hayek in his own work, made an effort to address the fundamental issue of an 'appropriate social order' from a general social theoretical perspective that cuts across traditional disciplinary boundaries. With his ideas on the two kinds of social order, on the function of rules and institutions, and on cultural evolution he took important steps toward a theoretically integrated social science.

See also: Cultural Evolution: Overview; Econometrics, History of; Economics, History of; Evolutionary Epistemology; Liberalism: Political Doctrine and Impact on Social Science.

Bibliography

Caldwell, B., 1997. Hayek and socialism. Journal of Economic Literature 35, 1856–1890.
Caldwell, B., 2004. Hayek's Challenge – An Intellectual Biography of F.A. Hayek. The University of Chicago Press, Chicago.
Ebenstein, A., 2001. Friedrich Hayek – A Biography. Palgrave, New York.
Fuster, J.M., 1995. Memory in the Cerebral Cortex – An Empirical Approach to Neural Networks in the Human and Nonhuman Primate. MIT Press, Cambridge, MA and London, England.
Hayek, F.A., 1929. Gibt es einen Widersinn des Sparens? Zeitschrift für Nationalökonomie 1, 387–419.
Hayek, F.A., 1931a. The 'paradox' of saving. Economica 32, 125–169.
Hayek, F.A., 1931b. Prices and Production. Routledge, London.
Hayek, F.A. (Ed.), 1935. Collectivist Economic Planning. Routledge, London.
Hayek, F.A., 1944. The Road to Serfdom. Routledge, London.
Hayek, F.A., 1948. Individualism and Economic Order. University of Chicago Press, Chicago.
Hayek, F.A., 1952a. The Sensory Order – An Inquiry into the Foundations of Theoretical Psychology. University of Chicago Press, Chicago.
Hayek, F.A., 1952b. The Counter Revolution of Science. Free Press, Glencoe, IL.
Hayek, F.A., 1960. The Constitution of Liberty. University of Chicago Press, Chicago.
Hayek, F.A., 1964. Kinds of order in society. New Individualist Review 3, 3–13.
Hayek, F.A., 1967. Studies in Philosophy, Politics and Economics. University of Chicago Press, Chicago.
Hayek, F.A., 1973. Law, Legislation and Liberty, vol. 1. University of Chicago Press, Chicago.
Hayek, F.A., 1976. Law, Legislation and Liberty, vol. 2. University of Chicago Press, Chicago.
Hayek, F.A., 1978. New Studies in Philosophy, Politics, Economics and the History of Ideas. University of Chicago Press, Chicago.
Hayek, F.A., 1979. Law, Legislation and Liberty, vol. 3. University of Chicago Press, Chicago.

Hayek, F.A., 1982. The sensory order after 25 years. In: Weimer, W.B., Palermo, D.S. (Eds.), Cognition and the Symbolic Processes, vol. 2. Lawrence Erbaum, Hillsdale, NJ, pp. 287–293.

Hayek, F.A., 1988. The Fatal Conceit – The Errors of Socialism. The Collected Works of F.A. Hayek, vol. 1. University of Chicago Press, Chicago.

Hayek, F.A., 2006[1920]. Beiträge zur Theorie der Entwicklung des Bewußtseins (1920). In: Hayek, F.A. (Ed.), Die sensorische Ordnung. Mohr Siebeck, Tübingen, pp. 199–226.

Kukathas, C., 1990. Hayek and Modern Liberalism. Clarendon Press, Oxford, UK.

Mises, L., 1920. Die Wirtschaftsrechnung im sozialistischen Gemeinwesen. Archiv für Sozialwissenschaft 47, 86–121.

Mises, L., 1922. Die Gemeinwirtschaft. Untersuchungen über den Sozialismus. Fischer, Jena.

Vanberg, V.J., 1982. Markt und Organisation. Individualistische Sozialtheorie und das Problem Korporativen Handelns. Mohr Siebeck, Tübingen.

Vanberg, V.J., 1994a. Rules and Choice in Economics. Routledge, London.

Vanberg, V.J., 1994b. Cultural evolution, collective learning and constitutional design. In: Reisman, D. (Ed.), Economic Thought and Political Theory. Kluwer, Boston, Dordrecht, London, pp. 171–204.

Vanberg, V.J., 2011. Liberal constitutionalism, constitutional liberalism and democracy. Constitutional Political Economy 22, 1–20.

von Humboldt, Alexander (1769–1859)

Tobias Kraft, Institut für Romanistik, Universität Potsdam, Potsdam, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Alexander von Humboldt (1769–1859) made essential contributions to a variety of disciplines and branches of knowledge, namely astronomy, geography, mineralogy, electrophysiology, political sciences, cartography, colonial history, and cultural anthropology. Humboldt's understanding of nature as a system of dynamic and interrelated phenomena, based on empirical studies, puts an end to natural history as a branch of philosophical speculation and stands at the threshold of modern science. Its roots can be traced from his American oeuvre until the *Kosmos*. His abundant reflections on the history of European colonization as well as his attempt to develop global comparative studies of natural phenomena have recently been presented as an important element in the history of globalization discourse.

Humboldt, Alexander von (1769–1859)

Alexander von Humboldt, likely the most famous European naturalist and scientific traveler in the nineteenth century, made essential contributions to a variety of disciplines and branches of knowledge, namely astronomy, physical and applied geography, geodesy, mineralogy, electrophysiology, political sciences, cartography, colonial history, and cultural anthropology. Several fields of study emerge from Humboldt's works: plant (or phyto-)geography, pre-Columbian studies, political geography, and modern climatology. Humboldt's understanding of nature as a system of dynamic and interrelated phenomena, based on empirical studies, puts an end to natural history as a branch of philosophical speculation and stands at the threshold of modern science. Three titles of Humboldt's vast oeuvre stand out the most to represent the literary and aesthetic significance of his works: the *Relation historique* (Personal Narrative) describes his American travels in the genre of scientific travel literature, and the other two, namely *Ansichten der Natur* (Views of Nature) and *Kosmos*, are the key elements of his reformist efforts of democratizing knowledge, inspiring a sway of popular science books in the second half of the nineteenth century.

Today, his reflections on the interconnectedness between human settlement and natural resources are gaining attention of studies on historical climate change, biodiversity, and ecocriticism. His abundant reflections on the history of European colonization as well as his attempt to develop global comparative studies of natural phenomena have recently been presented as an important element in the history of globalization discourse. Humboldt's holistic model of science is considered as a precursor for today's transdisciplinary approaches. It represents an intellectual practice of connecting and intertwining different norms and formations of knowledge, with a coherent pattern traceable from his American oeuvre until the *Kosmos*.

Youth, Education, and European Travels (1769–99)

Not all of this was evident right from the start. Born on 14 September 1769 at Berlin, Humboldt's youth was largely characterized by early failure, physical delicacy, and frustration. Apparently less brilliant and less motivated than his older brother Wilhelm, Alexander had to endure the long hours of instruction from many of, indeed first class, private educators that frequented Schloss Tegel in his early years. At the age of 16, the first visits to Henriette Herz's progressive salons put an end to the deprivation Alexander had felt so strongly, and introduced both the brothers to the brightest minds of the enlightened Berlin. The following years were dedicated to the study of cameralism, economics, natural history, mathematics, anthropology, physics, and languages at Frankfurt/Oder, Göttingen, and Hamburg. Nevertheless, it was another discipline, botany, that attracted his main attention in these years. His first mentor, the Berlin botanist Carl Ludwig Willdenow, introduced Alexander to his collection of exotic specimen from the tropics. Soon, the inspiring sessions with Willdenow boosted his desire for expedition, travel, and first-hand experience. During a stay in Mainz he was introduced to Georg Forster, who would become his travel companion on a voyage through Europe in 1790. It was with Forster, a member of the second Cook expedition, that Humboldt not only witnessed the revolutionary energy of Paris, but also the cosmopolitan spirit of London, where he met Sir Joseph Banks and attended sessions at the Parliament. These weeks were to become crucial for the initiation of the adventurous, the ambitious and the political Humboldt – the restless traveler, prolific author, and future diplomat.

Back in Prussia and after completing his studies of mineralogy at the famous Bergakademie Freiberg (1791–92), Humboldt was all prepared for a promising career as Prussian mining director. His instructive manuals and practical inventions soon led to major improvements in the productivity and working conditions of the mines he supervised. Experiments in electrophysiology and the chemistry of plants were to follow, spreading his name in the scientific community of his days. In search of the 'forces of life' (*Lebenskräfte*), Humboldt developed the idea of the harmony of elements in the inner organization of all living organisms, a concept that fully crystallized in a famous note he would jot down in his American travel diary in 1800: 'Everything is interrelated' (*Alles ist Wechselwirkung*). This conviction marks the epistemic nucleus of the Humboldtian approach, which also shaped the *Kosmos* in the notion of 'nature is unity in diversity' (*Natur ist Einheit in der Vielheit*). Unsurprisingly, the first sketches for a 'physical description of

the world' (*physique du monde/Physikalische Geographie*) date back to these days.

The year 1796 marked a turning point in Humboldt's early career. With the death of his mother, Alexander inherited a large fortune, providing means for realizing his aspiration for scientific expedition. He immediately resigned from his public offices and from there on, his every activity was conceived as preparatory: acquiring only the finest scientific instruments, networking in Paris (with Delamétherie, Borda, Delambre, Laplace, Lamarck, and Cuvier), finding a travel companion (the French physician and botanist Aimé Bonpland), measurements, travels through Spain, new measurements, and finally, soliciting a travel pass through the Americas from King Carlos IV. With diplomatic skill and the backing of the Spanish minister Urquijo, the petition succeeded. On 5 June 1799, Humboldt and Bonpland set off from the port of La Coruña to American shores: the first independent travel expedition in modern times was about to begin.

The American Voyage and the Years in Paris (1799–1829)

"We run around like fools; during the first three days, we did not identify a thing, always dropping one object to pick up another one. Bonpland claims that he will loose his mind if the marvels do not stop soon. But even more beautiful than each of these wonders by itself is the impression of the entirety of this powerful, luxuriant and at the same time so light, cheerful, and mild plant life. I feel that I will be very happy here and that these impressions will continue to cheer me in the future" (Humboldt in Kutzinski et al., 2012).

In this first letter from the *New Continent*, as he liked to phrase it, Humboldt puts in a nutshell the fascination he would feel for the tropics through his entire life. These lines, written to his brother Wilhelm on 16 July 1799 from the Venezuelan port of Cumaná, seem to inscribe Humboldt into the tradition of Spanish chronicles from the conquista days, when America's nature (and people) were miraculous, exotic, and exuberant. But Humboldt's American Voyage was never aimed at 'marvelous possessions' (Greenblatt). The gaze at beautiful wonders soon gave way to thorough science in the field – countless measurements on ships and donkeys, in valleys and on mountain peaks, research in the colonial archives and at archeological sites, and an ever-increasing specimen collection of America's flora and fauna.

The 5-year voyage (1799–1804) had the travelers depart from Spain for a stopover in the Canary Islands to land at the coast of the Venezuelan provinces, traverse the *hinterland* of New Granada, Cuba, Cartagena, the South American Cordilleras, Bogotá, Quito, Lima, Guayaquil, New Spain, Cuba again, and for a last brief stop, the United States, before returning to Europe. He came into contact with all social classes of the Spanish colonial world: Spanish officials, European merchants, indigenous servants, workers and peasants, enslaved Africans, catholic missionaries, and creole elites from all branches of society: politics, economics, culture, and academia. He ascended many of the most prominent mountains and volcanic peaks of the continent, among them Mount Chimborazo near Quito, then considered the highest summit in the world. In 75 days, he traveled 1500 miles in a pirogue, up and down the Orinoco, Meta, and Casiquiare Rivers, lending final proof to the older hypothesis of an intricate connection between the Orinoco and Amazonas network. His interests and research encompassed botany, zoology, climatology, geology, economy, cartography, political geography, and cultural history of the 'New Continent' (**Figure 1**).

After his return to Europe, Humboldt undertook several journeys to complete his American antiquarian research, among them a sojourn in Rome, where he would meet his brother Wilhelm and undertake a 'voyage d'archive' (Bourguet, 2011) through the Vaticane codices in search of Mesoamerican manuscripts. Already an international star of science, Humboldt was faced with a general public curiosity for his travel accounts and the European *République des lettres*, awaiting the results of his scientific findings. The very successful *Ansichten der Natur* (Tübingen 1808) fulfilled the expectations of the former. Their colorful and vivid descriptions of the American landscapes combined literary composition with a naturalist framing, introducing Humboldt's concept of a physiognomy of nature.

Publishing in Paris

Even though the 1st edition of *Ansichten* was all about his American journey, Humboldt did not consider them to be part of the 29 volumes that he would publish in the following decades upuntil 1838, when the last part of Humboldt's history of colonial discovery, the *Examen critique*, was to appear in Paris. His American oeuvre was aimed at his peers and targeted two goals: to deliver specific studies, findings, and measurements to the specialized scholar, and to provide, at the same time, a general and all-encompassing tableau of the physical, political, and cultural character of the tropics. Publishing most of his American works from Paris also meant publishing them in French, a matter of course for Humboldt, who had been raised bilingually from his earliest youth. Needless to say, it is the French capital, where Humboldt's scientific ambitions were to find their most natural habitat. Paris would become the center of his life from 1809 to 1827 and a frequent place to return to, whether in private or on a diplomatic mission with assignments from the Prussian King. The city offered all he needed: the most renowned institutions of science and culture, and the best publishing houses for his demanding task to produce his findings in opulent quarto and folio volumes. Some of Humboldt's plans were to fall short of his own expectations, especially his time frame: instead of 3 years, he needed 30 years for the project, the simultaneous publication of all his writings in both French and German had to be abandoned soon in favor of the French, his *Relation historique* despite taking almost 20 years in the making was never completed, and several intents to publish revised editions of some of his titles failed in the light of more urgent matters.

In his times, only comparable to a work like the *Description de l'Égypte*, Humboldt's *Voyage aux régions équinoxiales du Nouveau Continent* is abundant in its content and, at times, puzzling in its complex network of references. It can certainly best be described by the impact its various publications left on the contemporary academic and political landscape.

As one of its earliest outcomes, Humboldt's *Essay on the Geography of Plants* (Paris, Tübingen 1807, Trans. Humboldt and Bonpland, 2009), became a founding document for the

Figure 1 Map of Humboldt's American travel route. Original file from: Wikimedia Commons (http://en.wikipedia.org/wiki/File:AvHumboldts_Americatravel_map_en.svg) The file is licensed under the Creative Commons Attribution-Share Alike 2.5 Generic license.

comparative study of plants in their spatial and climatic distribution, articulating an integrative science of the earth, which encompasses most of today's environmental sciences (Jackson, 2009). The iconic image of this newly drafted discipline – the 'Physical Tableau of the Equitorial Regions' – was likely its main attraction (**Figure 2**). Much more than an illustration to the text, the 'tableau' functions as a visualization of the discipline itself, its intellectual ambition, and (inter-) disciplinary scope. Both in the book and in the 'tableau' itself, Humboldt first experimented with a conceptual framing of dialogue between text and image that later became the cornerstones of his geographic, political, and cultural atlases (Kraft, 2014). Together with applying geological profiles formerly used in mining to outline the physical character of territories like Spain and Mexico, these visual breakthroughs put Humboldt's name among the most innovative examples of data visualization in the nineteenth century. Humboldt's both theoretical and visual concepts of isothermal lines provided cartographic representation to climate research, an innovation that is still in use today. With Humboldt's isotherms, worldwide comparative climatology became pictorial.

The Political Humboldt

Recently, more attention has been given to the political side of Humboldt's oeuvre. For a Latin American audience, this is, of course, nothing new. His two major writings *Political Essay on the Kingdom of New Spain* (Paris 1808–11/1825–27, Trans. Humboldt, in press) and *Political Essay on the Island of Cuba* (Paris 1826, Trans. Humboldt, 2011) played a major role in the struggle for independence during the first decades of the century. For Humboldt's voluminous text on New Spain, this is even more true as he composed and published his writings and substantial cartographic material on New Spain on the eve of its final decline, that is, just at the moment, when the creole elites were to initiate national liberation from the Spanish dominion. Humboldt, in the history of his reception often coined as the 'second,' that is, the 'scientific discoverer of America,' was thereby enthroned as the programmatic precursor of future Mexico. The Mexican historian and politician Lucas Alamán left no doubt about the foundational character of Humboldt's work when he wrote to Humboldt that the constitutional assembly of 1824 used his *Political Essay* as inspiration and guidance for the tasks ahead. Clearly, both texts – as much his study on New Spain as the one on Cuba – were to set new standards by their concise study, composition and presentation of geography, politics, economics, demography, and social analysis of the colonies. Humboldt's essays gave way to new American self-awareness: up-to-date in its data – gained from proper measurements, from thorough work in the archives, and contributions from his countless correspondents – independent in its judgment and overwhelmingly precise in its cartographical mapping.

In Cuba, matters were different: his work was censured on the spot, and independence far away. The censorship was largely due to Humboldt's crystal clear stand on the issue of slavery, undoubtedly the foremost topic in sociopolitical, economic, and ethical debates in nineteenth-century Cuba. In 1791, the enslaved in the French colony of Saint-Domingue had begun the first successful large-scale insurrection in history, known as the Haitian Revolution (1791–1804). With Saint-Domingue's economy in ruins, the demand for Cuban sugar increased, and so did Cuba's demand for slaves. In addition to infusing the Cuban sugarcane industry with thousands of fleeing French planters and sugar engineers, the Haitian Revolution also generated widespread fears of another revolution. During his two visits to the island (1800–01/1804), Humboldt gained a close impression of the shock-waves that the Haitian Revolution had left among the Cuban oligarchy. He mentions the 'troubles in Saint-Domingue' repeatedly in the *Political Essay on the Island of Cuba*, pointing to the progressive agenda that Havana's merchant guild and Patriotic Society had pursued in the 1790s and again in 1811. But that did not much improve the working conditions of Cuba's enslaved, whose fate – as Humboldt had argued – could only lie in the complete abolition of slave labor. However, the book was to take another turn in its international reception. In 1856, the US journalist, pro-slavery activist, and intimate expert in Cuban affairs, John S. Thrasher, published an English translation of Humboldt's text. In *The Island of Cuba*, Thrasher had made significant changes, omitting Humboldt's section on slavery and making countless emendations that Humboldt himself saw as 'politically motivated.' In the eyes of the English reader, Humboldt had turned from a vehement abolitionist to a mild reformer with sympathies for the ruling oligarchy of sugar planters. That same year, Humboldt complained about Thrasher's manipulations in a public letter that was immediately translated to English and appeared in numerous newspapers across the U.S. The affair was even to influence public opinion during the Presidential campaign between James Buchanan and the Humboldt admirer John C. Frémont. Humboldt's fierce denials in this scandal should have restored his reputation, but in fact, the stain of political opportunism or even reactionary blindness to social problems would remain with the (English-speaking) readership until the recent times.

The Mesoamerican Cultures

What has been called Humboldt's 'imperial gaze,' in other words his contribution to the imperial politics of the late nineteenth century, can also be read as late reflexes of this (at times intentional) misreading. Yet it is not coherent with either his political or his cultural-anthropological writings. On the contrary: Humboldt's political criticism went far beyond that of most of his contemporaries. In his *Political Essay on the Kingdom of New Spain*, he openly condemned the Mexican creole elites and colonial administration for their complete lack of political vision and inaptness to guarantee legal equality to the indigenous populations. In his *Political Essay on the Island of Cuba*, as outlined in the preceding paragraphs he strongly rebuked the agroeconomic complex of mass slavery and monocultural plantations. In both texts, as well as in countless political activities during his lifetime, he argued for political reform both in the Americas and in Europe, in favor of individual freedom, political participation, and liberal economics. Among the first Europeans in nineteenth century to thoroughly study the archeological remains of Mesoamerican cultures, Humboldt made seminal contributions to the cultural history, mythology, and archeoastronomy of the Americas in his 'pittoresque atlas,' the *Views of the Cordilleras and Monuments of the Indigenous*

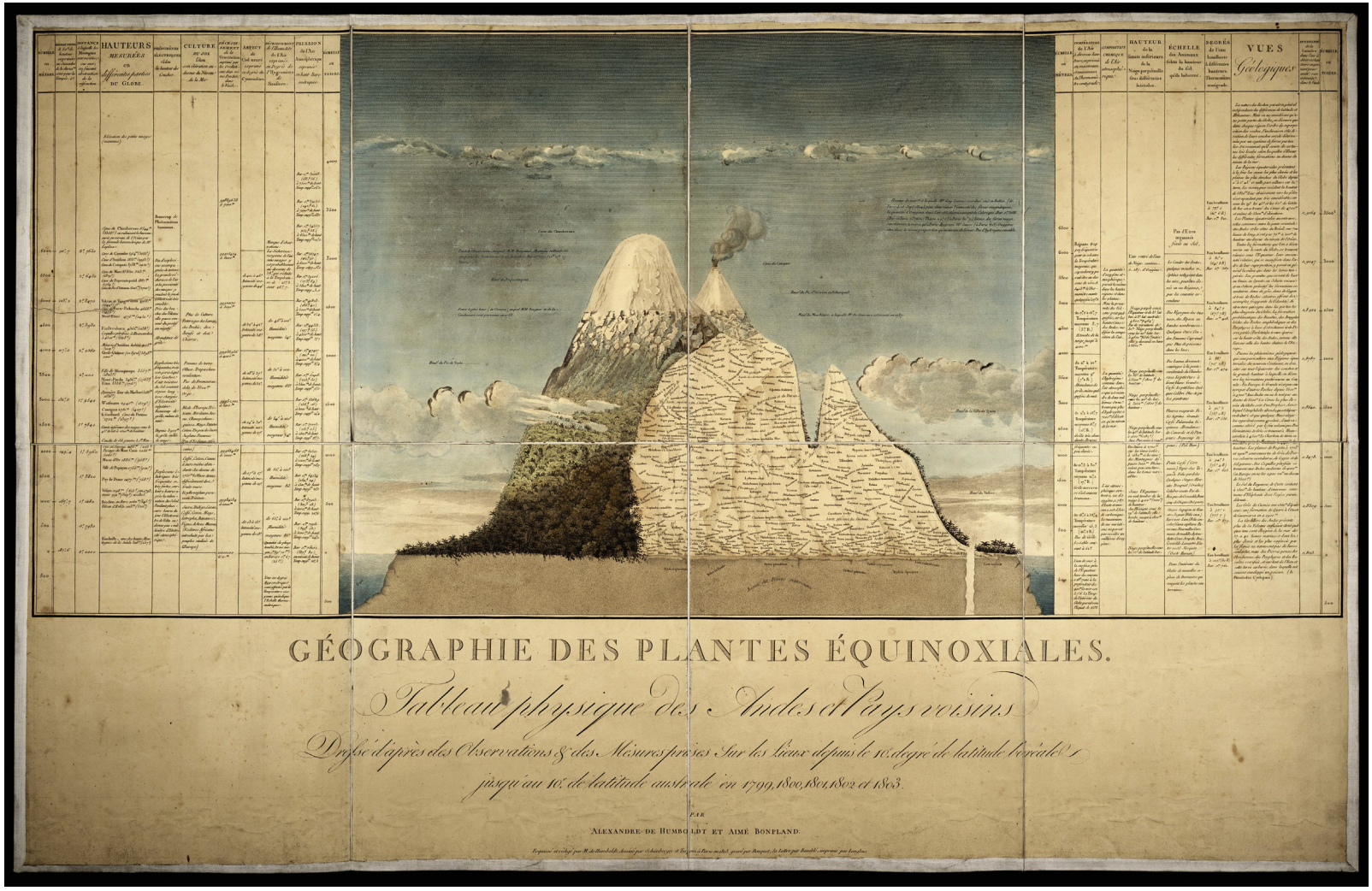

Figure 2 Tableau physique des Andes et Pays voisins.; Humboldt, A. von, Bonpland, A., 1807. Essai sur la Géographie des Plantes accompagné d'un tableau physique des régions équinoxiales. Fondé sur des mesures exécutées, depuis le dixième degré de latitude boréale jusqu'au dixième degré de latitude australe, pendant les années 1799, 1800 1801, 1802 et 1803. Avec un planche. Schoell, Cotta, Paris, Tübingen. Original scan from: Biodiversity Heritage Library (http://biodiversitylibrary.org/page/9869921), Usage is free, since copyright is expired and/or not further specified, see http://www.biodiversitylibrary.org/bibliography/9309#/summary and http://biodivlib.wikispaces.com/Licensing+and+Copyright.

Peoples of America (Paris 1813, Trans. Humboldt, 2012a). As an early example of nineteenth century anthropology and ethnology, Humboldt's keen interest in indigenous cultures helped to shed a new light not just on the remains of perished societies, but also on the originality of Native American art, sparking "a movement to reconsider the aesthetic value of non-European art that eventually led to a new, 'modern' aesthetic'" (Kutzinski et al., 2012). Humboldt's American nature, in consequence, is not a vast territory of unpopulated land ready to become the object of European desires, as some have suggested, but a space of complex histories, layers of time, and language(s), that Humboldt not only reflects upon in his text, but also visually in his American cartography.

The American oeuvre, encompassing 29 volumes with over 1400 copper plates, was to become the largest and the most expensive scientific work a single person had and has ever published. Its impact reached two continents, ruined several publishing houses, and finally, Humboldt's finances as well. This was one of the reasons why he saw himself forced to leave his beloved Paris to return to Berlin in 1827 and take an appointment by his King, Friedrich Wilhelm III, to serve as chamberlain at the Prussian court.

The Russian–Siberian Voyage (1829)

By that time, his plans for an expedition through Russia and Siberia had become very determined. The first sketches of this research program date back to 1812. Already then, Humboldt knew that the major tasks of his scientific masterplan were still ahead. In order to conceive the physical world as a cosmos of forces, natural phenomena, and relations, Humboldt needed an Asian counterpart to his American experience. But unlike during his travels through the 'New Continent,' Humboldt had become too famous and financially too dependent to undertake such a mission on his own. As head of a group of naturalists (Humboldt, Christian Gottfried Ehrenberg, and Gustav Rose), Humboldt had to follow a predefined travel route, with welcoming committees at every station, close surveillance by Cossack horsemen and a clear division of labor, in which his own contribution was strictly limited to mere scientific output. The reputation of Humboldt's political stance over questions of individual liberty, peonage, and social reform seemed too much of a risk for the Russian Emperor Nicholas I. Humboldt's acceptance of these hard-to-bear conditions can only be explained by the credo of his global comparative project, for which the travels to Central Asia were essential. Already 60 years of age, he knew that such an opportunity was improbable to come again. The travels became an Asian *tour de force*, with halts in Königsberg, Riga, Saint Petersburg, Moscow, the Urals, Altai Mountains, a gaze at the Chinese borders, the great Siberian planes, Omsk, the Caspian Sea, the return to Moscow, and, finally, Berlin. In just about 9 months, the expedition had traveled more than 10 000 miles, mostly at night. Restricted as he was in his scientific endeavors, Humboldt focused on magnetic, barometric, and astronomic findings, mining enterprises, and comparative analyses between the American and Asian nature. In the 30 years between his first and the second great voyage, much had changed. As Humboldt stated in a letter to Wilhelm from September 1829, the remarkable pace of his second *grand tour* correlated with the acceleration of modern science, better instruments, and the refinement of methods in all branches of knowledge. Yet despite the modern times, it would take him another 14 years to complete his studies and accommodate the results of the Russian–Siberian Voyage in the three volumes of *Asie Centrale* (Paris 1843). They were to pioneer the era of modern geology on Inner Asia.

The Final Chapter: Scientific Networks and Humboldt's *Kosmos* (1829–59)

Already in 1828, Humboldt had started to build new scientific networks in the Prussian capital. A conference under his direction brought together 600 naturalists and physicians in Berlin, where he initiated numerous activities to propel the city's academic culture and foster new scientific institutions. The goal was as clear as it was ambitious: Humboldt wanted Berlin to become a centerpiece place of modern science. His 'Kosmos lectures' at the University and the Singakademie were to attract crowds of more than 800 people eager to hear Humboldt's lessons on the general aspects of the physical world ('Allgemeines Naturgemälde'). Even though related to the successive appearance of his *Kosmos* from 1845 to (posthumously) 1862, the lectures have to be seen more in the light of raising public awareness for scientific knowledge and education. The idea of the *Kosmos* as a stand-alone work in its own right dated back much earlier, envisioned decades before the lectures in his *Geography of Plants* and fully conceptualized in a letter to his friend Karl August Varnhagen von Ense in 1834. The 'work of my life' was meant to embrace it all and "present an epoch in the mental development of man as regards his knowledge of nature" (Humboldt and Varnhagen von Ense, 1860). Humboldt's goal was to satisfy both: the specialist and the general public, the professor and the student. As the first tranches of *Kosmos* were to appear in 1845, readers' expectations had risen to unprecedented levels. By the time the second volume became available, German booksellers were overrun by public demand and had to commission multiple reprints. British and US-American publishers competed with three simultaneous editions, while rushing to present new and updated translations of the *Ansichten der Natur* and the *Personal Narrative*. *Kosmos* was to become the popular epitome of Humboldtian thinking: modeling nature in its inherent connectivity, its network-like, organic intertwinement, and, finally, in its relation to man. "For the study of nature was, for Humboldt, inseparable from the study of the mind in its material, social, and cultural context. This reflexivity of mind, society, and nature became his overriding argument" (Walls, 2005).

Humboldt was not to see the final volume of his *Kosmos* published, when it finally came out in 1862. Quite literally with the quill in his hands, he died on 6 May 1859, at the age of 89, in Berlin.

The Present: Proliferating Humboldt Studies, New Editions, the American Travel Diaries

Current studies on the life, works, and scientific achievements of Alexander von Humboldt encompass a wide array of disciplines and scholarly fields (history of science, colonial history, postcolonial studies, art history, cultural studies, poetics of science, history of ideas, epistemology, environmental studies,

and ecocriticism). What we might call a rather recent *renaissance* of Humboldt studies is largely due to a decade-long effort (1997–2006) around the bicentennial of the American travels (1799–1804), when a series of large-scale Humboldt exhibits and conferences around the globe (Mexico, Cuba, Venezuela, Germany, Colombia, Ecuador, Peru, Spain, England, and Austria) reminded the scientific and general public of the historic and contemporary relevance of one of Europe's most eminent figures of the nineteenth century. These events set the floor for numerous editorial projects (i.e., Humboldt, 1991, 1998, 2004a,b, 2009b,c, 2011, 2012a,b, in press), which were aimed at recovering the original Humboldtian text as a response to a history of reception, in which Humboldt was largely read and commented upon based on highly abbreviated and often poorly translated editions of his works. These new editions have given the international scientific community, mostly in the German-, Spanish-, and English-speaking world, a restored textual ground, an elevated awareness for the impact Humboldt had on his contemporaries, and for the significance his works have for today's readers. The recent acquisition of Humboldt's American Travel Diaries (December 2013) by a consortium of German public and private institutions under the auspices of the Staatsbibliothek zu Berlin (Prussian Cultural Heritage Foundation) will certainly add to this, as it provides an essential element in fully understanding Humboldt's place in the study of scientific practices and travel literature in the nineteenth century.

See also: Anthropology: Overview; Cartography; Colonialism, Anthropology of; Colonization and Colonialism, History of; Environmental Sciences; Geography: Overview; Mesoamerica, Archaeology of; Nature and Society in Geography; Political Geography; Science, History of; South America: Sociocultural Aspects; von Humboldt, Wilhelm (1767–1835).

Bibliography

Beck, H., 1959. Alexander von Humboldt. In: Von der Bildungsreise zur Forschungsreise 1769–1804, Band 1. Franz Steiner Verlag, Wiesbaden.

Beck, H., 1961. Alexander von Humboldt. In: Vom Reisewerk zum "Kosmos" 1804–1859, Band 2. Franz Steiner Verlag, Wiesbaden.

Bourguet, M.-N., Janvier-Mars 2011. Notes romaines. Le 'voyage d'archive' d'Alexander von Humboldt (1805). Études germaniques 66 (1), 21–44.

Ette, O., Hermanns, U., Scherer, B.M., Suckow, C. (Eds.), 2001. Alexander von Humboldt: Aufbruch in die Moderne. Akademie Verlag, Berlin.

Ette, O., 2001. The scientist as Weltbürger: Alexander von Humboldt and the beginning of cosmopolitics. HiN – Internationale Zeitschrift für Humboldt-Studien 2 (2). http://www.hin-online.de (accessed 13.12.13.).

Ette, O., 2009. Alexander von Humboldt und die Globalisierung: Das Mobile des Wissens. Insel, Frankfurt am Main, Leipzig.

Fiedler, H., Leitner, U., 2000. Alexander von Humboldts Schriften: Bibliographie der selbständig erschienenen Werke. Akademie Verlag, Berlin.

Hein, W.-H. (Ed.), 1987. Alexander von Humboldt: Life and Work. C.H. Boehringer Sohn, Ingelheim am Rhein.

Humboldt, A. von, Varnhagen von Ense, K.A.L.P., 1860. Letters of Alexander von Humboldt, Written between the Years 1827 and 1858, to Varnhagen von Ense. Together with Extracts from Varnhagen's Diaries, and Letters from Varnhagen and Others to Humboldt. Authorized and translated from the German, with explanatory notes and a full index of names. Trübner & Co., London.

Humboldt, A. von, Bonpland, A., 1971–1973. Voyage de Humboldt & Bonpland: Voyage aux régions équinoxiales du Nouveau Continent. Facsimilé intégrale du édition de Paris 1805–1834. Theatrum Orbis Terrarum, Amsterdam, New York.

Humboldt, A. von, 1991. Reise in die Äquinoktial-Gegenden des Neuen Kontinents. In: Herausgegeben von Ottmar Ette. Mit Anmerkungen zum Text, einem Nachwort und zahlreichen zeitgenössischen Abbildungen sowie einem farbigen Bildteil, Band 1. Insel, Frankfurt am Main, Leipzig.

Humboldt, A. von, 1998. Ensayo político sobre la Isla de Cuba (Edición de Miguel Ángel Puig-Samper, Consuelo Naranjo Orovio, Armando García González). Edición Doce Calles, Junta de Castilla y León, Madrid.

Humboldt, A. von, 2004a. Kosmos: Entwurf einer physischen Weltbeschreibung. Ediert und mit einem Nachwort versehen von Ottmar Ette und Oliver Lubrich. Eichborn, Frankfurt am Main.

Humboldt, A. von, 2004b. Ansichten der Kordilleren und Monumente der eingeborenen Völker Amerikas. Aus dem Französischen von Claudia Kalscheuer. Eichborn, Frankfurt am Main. Ediert und mit einem Nachwort versehen von Oliver Lubrich und Ottmar Ette.

Humboldt, A. von, 2009a. Mein vielbewegtes Leben: Der Forscher über sich und seine Werke. Ausgewählt und mit biographischen Zwischenstücken versehen von Frank Holl. Eichborn, Frankfurt am Main.

Humboldt, A. von, 2009b. Kritische Untersuchung zur historischen Entwicklung der geographischen Kenntnisse von der Neuen Welt und den Fortschritten der nautischen Astronomie im 15. und 16. Jahrhundert: Mit dem geographischen und physischen Atlas der Äquinoktial-Gegenden des Neuen Kontinents Alexander von Humboldts sowie dem Unsichtbaren Atlas der von ihm untersuchten Kartenwerke. Mit einem vollständigen Namen- und Sachregister. Nach der Übersetzung aus dem Französischen von Julius Ludwig Ideler ediert und mit einem Nachwort versehen von Ottmar Ette. Insel, Frankfurt am Main, Leipzig.

Humboldt, A. von, Bonpland, A., 2009. Essay on the Geography of Plants. The University of Chicago Press, Chicago, London. Edited with an introduction by Stephen T. Jackson (Sylvie Romanowski, Trans.).

Humboldt, A. von, 2011. Political Essay on the Island of Cuba: A Critical Edition. The University of Chicago Press, Chicago, London. Edited with an Introduction by Vera M. Kutzinski and Ottmar Ette. (J. Bradford Anderson, Vera M. Kutzinski, and Anja Becker, Trans.) With Annotations by Tobias Kraft, Anja Becker, and Giorleny D. Altamirano Rayo.

Humboldt, A. von, 2012a. Views of the Cordilleras and Monuments of the Indigenous Peoples of the Americas: A Critical Edition. Edited with an Introduction by Vera M. Kutzinski and Ottmar Ette. (J. Ryan Poynter, Trans.) Annotations by Giorleny Altamirano Rayo and Tobias Kraft. The University of Chicago Press, Chicago, London.

Humboldt, A. von, 2012b. Vistas de las Cordilleras y Monumentos de los Pueblos Indígenas de América. Traducción de Gloria Luna Rodrigo y Aurelio Rodríguez Castro. Universidad Autónoma de Madrid, Marcial Pons Historia, Madrid.

Humboldt, A. von. Political Essay on the Kingdom of New Spain: A Critical Edition. Edited with an Introduction by Vera M. Kutzinski and Ottmar Ette. The University of Chicago Press, Chicago, London, in press.

Jackson, S.T., 2009. Introduction: Humboldt, Ecology and the Cosmos, in: Essay on the Geography of Plants. Edited with an introduction by Stephen T. Jackson. (Sylvie Romanowski, Trans.). The University of Chicago Press, Chicago, London, pp. 1–46 (this is Humboldt 2009b).

Kraft, T., 2014. Figuren des Wissens bei Alexander von Humboldt. Essai, Tableau und Atlas im amerikanischen Reisewerk. De Gruyter, Berlin, Boston.

Kutzinski, V.M., Ette, O., Walls, L.D. (Eds.), 2012. Alexander von Humboldt and the Americas. edition tranvía, Berlin.

Minguet, C., 1997. Alexandre de Humboldt: Historien et géographe de l'Amérique espagnole (1799–1804). L'Harmattan, Paris, Montreal.

Päßler, U., 2009. Ein Diplomat aus den Wäldern des Orinoko: Alexander von Humboldt als Mittler zwischen Preußen und Frankreich. Franz Steiner Verlag, Stuttgart.

Rupke, N.A., 2005. Alexander von Humboldt: A Metabiography. Peter Lang, Frankfurt am Main, Berlin, Bern u.a.

Sachs, A.J., 2006. The Humboldt Current: Nineteenth-Century Exploration and the Roots of American Environmentalism. Viking, New York.

Walls, L.D., 2005. Rediscovering Humboldt's environmental revolution. Environmental History 10 (4), 758–760.

Relevant Websites

http://www.avhumboldt.de – Alexander von Humboldt Informationen online (German, Spanisch, English, French).

http://www.bbaw.de/forschung/avh – Alexander von Humboldt Research Group at the Berlin-Brandenburg Academy of Sciences and Humanities.

http://www.avhumboldt.de/?page_id=469 – Humboldt Digital – Online Bibliography of digital facsimiles of Humboldt's works.

http://www.avhumboldt.net – The Humboldt Digital Library.

http://sitemason.vanderbilt.edu/site/ld5WVy/hie – HiE – Alexander von Humboldt in English. Project website.

http://www.hin-online.de – HiN – Humboldt im Netz. International Review for Humboldtian Studies (ISSN:1617-5239).

von Humboldt, Wilhelm (1767–1835)

Jurgen Trabant, Free University, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Wilhelm von Humboldt (1767–1835) was a Prussian politician and linguist as well as the founder of political liberalism and the University of Berlin. He founded an anthropological approach to the study of Man, and was also the founder of anthropological comparative linguistics. He studied Basque, Amerindian languages, Kavi, and Austronesian languages.

Wilhelm von Humboldt was a writer, politician, linguist, and language philosopher who developed key concepts in the theory of language and had a strong influence on synchronic and structural descriptive linguistics. He studied at the universities of Frankfurt an der Oder and Göttingen, and maintained lifelong friendships with Goethe and Schiller. He stayed in Paris from 1797 to 1801 and journeyed to Spain and the Basque Country. His subsequent positions included: ambassador to Rome 1802–08, reformer of the Prussian educational system, founder of the University of Berlin 1810, and minister. After resigning from office in 1819, Humboldt carried out linguistic research in Tegel during 1820–35, regularly presenting linguistic papers at the Berlin Academy. He laid down the principles of philosophically oriented anthropological research of the empirical cultural manifestations of humankind. After writing political and esthetic essays, Humboldt concentrated on language. On the basis of Kantian philosophy, he developed a philosophy of language known as the 'formative organ of thought.' He studied many languages, including Basque, Amerindian languages, Sanskrit, Chinese, Japanese, and the Malayo-Polynesian languages to which his major work is devoted. He developed a huge program of descriptive structural linguistics with special emphasis on the 'characterization' of individual languages on the basis of their literature.

Politics

The resignation of Wilhelm von Humboldt, born 22 June 1767 at Potsdam, from his office as the Prussian Minister of Constitutional Affairs at the end of 1819 was a major catastrophe for the political future of Germany. For Humboldt himself it was an act of liberation and a step toward what he regarded as his real vocation, the study of languages, to which he dedicated the rest of his life. Chateaubriand, the French poet and politician, considered this strange scientific activity as the spleen of somebody who had to kill his time as a consequence of the loss of political influence. But it was in fact the achievement of a philosophical and anthropological quest that had started in Humboldt's youth. Studying the languages of the world was not at all something completely extraneous to his political work. The same motivation that would guide his linguistic work had also governed his political activities: it was a quest for the creative center of the human mind and for the achievement of individual felicity and autonomy within the community with other individualities.

This is exactly what had created his political difficulties. At the end of his political career, Humboldt had elaborated a project for the constitutional transformation of Prussia: a proposal to establish a constitutional monarchy with a gradually developing representational system. The French Revolution, which he had welcomed as a decisive step in the history of humanity, had failed in Humboldt's eyes because of its rationalistic one-sidedness: the transformations of France had been governed too radically by abstract rational principles and had not taken into account historically grown entities. For Humboldt, the creation of the new should always be the result of a 'synthesis,' of a sympathetic union of two entities, which, creating a new entity by their union, maintain their identity (the alternative – and lesser – methods of uniting two entities are 'isolation' and 'incorporation,' the former being only the juxtaposition of the two, and the latter, the brutal destruction of their integrity through their unity). Humboldt proposed a 'synthesis' of the traditional prerogatives of the monarch and the interests of the economically active classes (with a consideration also for the traditional rights of the nobility), a 'marriage' that would have transformed Prussia eventually into something new, a constitutional monarchy – and that would have had a model function for the rest of Germany. Not only for the inner organization of Prussia but also for the whole of the area called 'Germany,' Humboldt projected a political structure according to the same principle of a 'synthetical' union: a confederation that would have left quite a lot of autonomy to its members but, nevertheless, would have given important political functions to the whole (like warfare, freedom of movement, and free trade within the union – very much like the United States).

But, in 1819, the times were not propitious anymore for reforms and for the creation of new political structures, as after the defeat of Prussia by Napoleon, which was the beginning of so many happy reforms in Prussia, like the reform of the educational system, the renewal of the municipal administration, and the emancipation of the Jews. Now Napoleon was defeated and the old classes were striking back and were not willing to make any concessions to the bourgeoisie, to the memory of the 'people's war,' or to the experience of freedom and civil rights imported by the Napoleonic armies. Humboldt was the representative of liberal political reforms, and, as such, he had to disappear – in the time of Austria's Metternich, of Hardenberg, and a reactionary king in Prussia.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61054-1

University and Bildung

But if Humboldt did not succeed in realizing his constitutional dreams, he seems to have had an overwhelming success with the establishment of the University of Berlin some years earlier. Humboldt is best and generally known for this activity, which in fact occupied him for little more than a year (from February 1809 to June 1810) and which was only a part of the reorganization of the whole Prussian educational system. Humboldt indeed made the decisive administrative step for the establishment of the University of Berlin. But what was established was only partially Humboldt's University. He planned the university as a state-independent, economically autonomous institution, with an independence that would guarantee the freedom of research and instruction. Therefore, Humboldt requested that the university should be given a donation of land as the economic basis of that autonomy. Instead, the Berlin University was from the very beginning a state establishment, which – it is true – set unprecedented standards by the quality of its new faculty, by new forms of teaching, and by the unity of research and teaching for professors as well as for students. The concept of the research university was successfully realized. As a state-dependent institution it would forge a well-trained loyal elite for the King of Prussia rather than free citizens of the realm of the spirit.

Humboldt's proposals for the establishment of the Berlin University and the other educational institutions became, however, seminal texts for a pedagogical theory whose central concept was that of *Bildung* (educational and cultural development). Following the convictions of German idealistic philosophy and the classical German poets (Schiller and Goethe), Humboldt wanted human individuals to be given every possible means to develop their physical and intellectual capacities. The educational institutions have to grant access to that formation of the human individual. The main aim of *Bildung* is not the acquisition of professional skills but the unfolding of the highest intellectual and artistic aspirations. Ancient Greece as the absolute summit of human development is the cultural reference of that educational program. The concept of *Bildung* certainly gave an ambitious aim to the educational institutions of Prussia and of Germany as a whole. This concept even served as a justification of otherwise unexplainable investments of the state in intellectual enterprises like the humanities, which have no practical ends. *Bildung* was an enormously successful concept as one of the central values of German identity, giving education a very high position in the value system of the nation. But just as Humboldt's University project was only partially realized in the Prussian University, the idea of *Bildung* – a generous and grandiose idea – did not so much create the (also politically) autonomous and free individuals Humboldt was thinking of, but had more of a compensatory function: consoling the German bourgeoisie for its nonparticipation in political power. Hence, all that wonderful *Bildung* did not prevent the German catastrophe.

Anthropology

The establishment of the University of Berlin and *Bildung* secured Humboldt's place in the pantheon of the German nation. But his main activity and – in the long run – also his most successful intellectual achievement, was his linguistic and philosophical work. In this domain, however, Humboldt is not – as some historiographers tend to think – a founding father of the scientific project of historicocomparative linguistics that is linked to the names of Franz Bopp and Jacob Grimm. Bopp and Grimm invented the principles for a historical reconstruction of the unity of Indo-European, Germanic, Romance, Slavic, and other languages and for the comparison of historically related languages. This was to be the successful – romantic – project of the mainstream linguistics of the nineteenth century. Humboldt is, on the contrary, the founder of an alternative linguistic project, which is more the summary of the linguistic ideas of the Enlightenment: the study of the diversity of the languages of the world. *Verschiedenheit*, or diversity, is the key term of Humboldtian linguistics, which was to flourish – at least in part – in the twentieth century, when the historicocomparative paradigm was relieved by synchronic descriptive linguistics.

Humboldt's linguistic project was integrated into a larger 'anthropological' project. 'Anthropology' in the eighteenth century was a term that designated the empirical study of human beings, as opposed to philosophy, which investigates the universal nature of the human race. Humboldt sketched a vast research program for the investigation of the cultural manifestations of humankind. Understanding the creativity of the human mind was the philosophical motive that gave Humboldt's anthropological project its special dynamics. Two events that occurred in his lifetime were decisive for this research into creativity: first, as for any young intellectual of his generation, was the French Revolution, the creation of a completely new, enlightened, political organization. Hence, one of his first publications was an article on the first French Constitution (1791) and for his friends he wrote a booklength political treatise 'On the limits of the activity of the state' (1792), which was printed only after his death and which became a founding document of political liberalism. The other revolutionary event in Humboldt's lifetime was Goethe. Here again, what fascinated the young intellectuals was the 'genius,' the mystery of an unprecedented creativity. Humboldt lived in Jena, as a friend of Schiller, in close contact with Goethe. He studied Immanuel Kant, whose philosophy was to become the framework of his own theoretical developments. Humboldt's work is centered on the Kantian concept of imagination, *Einbildungskraft*, the faculty of the human mind that is responsible for the creation of the new.

Since imagination as the mediation between the senses and the intellect had remained rather mysterious in Kant's philosophy, Humboldt set out to find its roots. He discovered them in the difference between the sexes. The opposition of the male and the female in nature is the very basis of every new creation, not only of physical generation (*Erzeugung*) but also of the most 'sublime creature' of the imagination: thought. The loving embrace of the female and the male generates new living beings in the physical as well as in the intellectual world. Kant was shocked by Humboldt's somewhat pre-Freudian interpretation of his philosophy. But despite Kant's criticism, sexual union, 'matrimony' (*Vermählung*), and love would be the structural model for everything Humboldt set out to investigate: art, politics, education, and language. Just as dialectics is the

fundamental figure of thought in Hegel, the erotic encounter is the basic structure of Humboldt's thought.

After his reflections on the anthropological roots of creativity, Humboldt wrote a book on Goethe (1798), which contains a very modern theory of esthetics: art is not only a creation of the artist's imagination but also a (languagelike) communicative event between the producer of the work of art and the observers, the readers, the listeners who recreate the work of art in their act of reception and thereby accomplish the creative process.

Language

While still in Jena, Humboldt wrote his first reflections on language: 'On thinking and speaking' (1795–96), a short sketch that already contained the main features of his language theory. Humboldt's sojourn in France and in Spain (1797–1801) would be decisive for his making language the very center of his intellectual endeavors. In Paris, he discussed philosophical issues with the *Idéologues*, in whose philosophy language played a crucial role, and here and in his journey to the Basque Country, he encountered the Basque language whose profoundly divergent structural properties intrigued Humboldt. Later, his brother Alexander would bring linguistic material from America, and in Rome, Humboldt had access to the linguistic materials of the Jesuits (Lorenzo Hervás). For years, the Amerindian languages were Humboldt's empirical linguistic objects before he turned to the Malayo-Polynesian languages to which his main – posthumous – work *On the Kavi Language on the Island of Java* (1836–39) is devoted. The introductory volume of that book, *The Diversity of the Human Language-Structure*, completed before his death on 8 April 1835 in Tegel (near Berlin), is Humboldt's major work. With language, Humboldt arrived at the very heart of his general quest for the creation of the new by that mysterious Kantian faculty of *Einbildungskraft*.

Language is not just a means for the communication of thought; it is the very creation of thought. According to the Kantian system, the senses and the intellect create the concept through their synthesis. But for Humboldt, the concept is not created independently of the word: word and concept form an indissoluble synthetic unity; thought is created as sound. Language is the 'formative organ of thought' ('*das bildende Organ des Gedanken*,' 1903–36: VII, p. 53), 'labor of the mind' ('*Arbeit des Geistes*,' 1903–36: VII, p. 46). Like art, language (thought sound) also needs the dimension of the other: it must be heard and understood. Language as a cognitive activity reckons with the other, language is thought in the dimension of the other: *Mitdenken*, or cocognition. But surpassing art, language is a reciprocal creative activity: the hearer becomes the speaker, 'the word formed by me resounds from the mouth of the other' (1903–36: VII, p. 56). This ultimate reciprocity is the completion of the linguistic syntheses. The duality of I and thou is the 'fundamental type' of language.

Humboldt discovered that language is actually at the center of the Kantian system. Humboldt's language philosophy was a linguistic turn of transcendental philosophy. Since Bacon's discovery of the *idola fori*, European philosophy had become gradually aware of the fact that language is not just a device for the communication of thought but first and foremost a cognitive activity. This insight finds a coherent post-Kantian formulation in Humboldt.

The second important moment of that philosophy of language is the insight that the formation of thought, the linguistic 'transformation of the world into the property of the mind' (1903–36: IV, p. 420), always proceeds according to the special ways of historically particular communities, that is, according to particular languages. There is no such thing as language in general, language comes in languages. Thought generated by language, hence, is not uniform and universal but historically and culturally fragmented in different 'worldviews' (*Weltansichten*). The task of language, the generation of thought, is universal, but that task is realized in thousands of different ways in the different languages of humankind.

Now the new – or Leibnizian – moment of that insight is its positive interpretation: from Bacon onward, philosophy has lamented over the particularity of worldviews in the different languages and wanted to get rid of those 'prejudices'; the semantics of natural languages is like 'a mist before our eyes' (Locke) that philosophical enlightenment must discard. Only Leibniz gave an optimistic twist to the discovery of the linguistically generated diversity of thought when he envisaged a study of all the languages of the world in order to know the 'marvelous variety of the operations of our mind.' Herder took that idea from Leibniz, and Humboldt not only integrated it into the Kantian framework but also developed a huge program of descriptive linguistics.

Comparative Study of Languages

All the languages of the world were to be described according to what Humboldt calls their 'inner coherence' (*innerer Zusammenhang*), their particular structure. The languages of the world had mostly been grasped according to the traditional categories of Greek and Latin grammar. But the structural properties of a language can only be captured through its own categories. Humboldt envisaged an encyclopedia of the languages of the world, a new *Mithridates*. He started a whole series of descriptive works on different languages of the world of which the *Mexican Grammar*, the *Mexican Dictionary*, as well as the *Meso-*, *South-* and *North-American Grammars* (Humboldt, 1994, 2000, 2009, 2011, 2013) have been published recently. A second kind of structural studies aimed at a revised form of general grammar (which was also based on the Greco-Latin grammar): comparative investigations of linguistic categories throughout the languages of the world. Humboldt's article on the dual (1827) is an example of that research. After the completion of these structural descriptions, research into the parental relationships of languages (genealogy) or into structural similarities among languages (typology, classification) can be undertaken.

What is often considered as Humboldt's main contribution to linguistics, the so-called Humboldtian 'typology,' is neither central to his linguistic aims nor did Humboldt propose typology as an alternative to the dominant linguistic school, that is, to the historicocomparative studies of languages. On the contrary, in his last book, his unfinished work on Kavi, Humboldt eventually envisaged a comparative study of the

Malayo-Polynesian languages, which would follow Grimm and Bopp's work on the Germanic and Indo-European languages. Humboldtian typology is a linguistic myth created by superficial readings of his texts. Humboldt was very skeptical about the classifications of the Schlegels, and he explicitly refuted the legitimation of classifying languages, since languages are individuals and should be described as such.

The summit or keystone of Humboldtian linguistics is therefore the investigation of the 'character' of languages. This is just the opposite of classification and typology, of grouping similar entities together. It is a tentative attempt to grasp the very individuality of each individual language. Humboldt was not sure whether 'characterization' was still a scientific task. But whatever it is, it has to be tackled since it is the ultimate aim of linguistics to see how the human mind uses the instruments it has created, to investigate the very end of language: speech and discourse. The – necessary – structural description of a language is only 'comparable to its dead skeleton' (1903–36: VI, p. 147), the life of a language is discourse (*verbundene Rede*), a life giving 'character' to that language. Therefore, only languages that have a multifaceted literature yield the possibility of being described as having characters.

Nearly two centuries after Humboldt's sketch of a huge linguistic program, it can be said that the structural description of human languages he envisaged was the main task and achievement of linguistics in the twentieth century. It is not by chance that Humboldt is present in the seminal theoreticians of modern linguistics (Saussure (*see* Saussure, Ferdinand de (1857–1913)), Bloomfield (*see* Bloomfield, Leonard (1887–1949)), and Hjelmslev). The nineteenth century was not very Humboldtian, with its emphasis on diachronical research as well as in its naturalistic research methods. Humboldt's 'keystone' of linguistics, the 'characterization' of languages, was attempted by some literary-minded linguists (e.g., Vossler), but was discredited by nationalistic and ideological interpretations of linguistic structures. Since the structural descriptions of the languages of the world have now virtually been completed, it is not astonishing that in actual linguistics typological attempts are again very important. The so-called Humboldtian typology is often quoted in that context. But it is – even more than the other famous Humboldtian invention, 'Humboldt's University' – rather a myth than a creature of the author. Without any foundation in Humboldt's work is also the widespread conviction that Humboldt discovered fundamental principles of Chomskyan linguistics, like the 'infinite use of finite means.' Humboldtian linguistics as a whole is in direct opposition to that kind of linguistic research: it is a hermeneutical research into the cultural diversity of human languages in historical communities, preferentially through (literary) discourse, and not a naturalistic investigation of the structure of the universal human mind considered as a universal grammar.

See also: Continental Philosophy of Language; Education, History of; Educational Institutions, History of; Instructional Psychology; Language and Literature; Language and Thought: The Neo-Whorfian Hypothesis; Liberalism: Historical Aspects; Liberalism: Political Doctrine and Impact on Social Science; Linguistic Anthropology; Linguistic Turn and Discourse Analysis in History; Linguistic Typology; Linguistics: Overview.

Bibliography

Borsche, T., 1981. Sprachansichten. Der Begriff der menschlichen Rede in der Sprachphilosophie Wilhelm von Humboldts. Klett-Cotta, Stuttgart, Germany.
Borsche, T., 1990. Wilhelm von Humboldt. Beck, Munich, Germany.
Di Cesare, D., 1998. Einleitung. In: Humboldt, W., Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts. Schöningh, Paderborn, Germany, pp. 9–141.
Haym, R., 1856. Wilhelm von Humboldt. Lebensbild und Charakteristik. Gaertner, Berlin, Germany.
Humboldt, W., 1836–39. Über die Kawi-Sprache auf der Insel Java, 3 vols. Akademie der Wissenschaften, Berlin, Germany.
Humboldt, W., 1903–36. In: Leitzmann, A., et al. (Eds.), Gesammelte Schriften, 17 vols. Behr, Berlin, Germany.
Humboldt, W., 1988. On Language. The Diversity of Human Language-structure and its Influence on the Mental Development of Mankind (P. Heath, Trans.). Cambridge University Press, Cambridge.
Humboldt, W., 1994. In: Ringmacher, M. (Ed.), Mexicanische Grammatik. Schöningh, Paderborn, Germany.
Humboldt, W., 2000. In: Ringmacher, M. (Ed.), Wörterbuch der mexicanischen Sprache. Schöningh, Paderborn, Germany.
Humboldt, W., 2009. In: Ringmacher, M., Tintemann, U. (Eds.), Mittelamerikanische Grammatiken. Schöningh, Paderborn, Germany.
Humboldt, W., 2011. In: Ringmacher, M., Tintemann, U. (Eds.), Südamerikanische Grammatiken. Schöningh, Paderborn, Germany.
Humboldt, W., 2013. In: Verlato, M. (Ed.), Nordamerikanische Grammatiken. Schöningh, Paderborn, Germany.
Menze, C., 1965. Wilhelm von Humboldts Lehre und Bild vom Menschen. Henn, Ratingen, Germany.
Meschonnic, H. (Ed.), 1995. La pensée dans la langue. Humboldt et après. Presses Universitaires de Vincennes, Saint-Denis, France.
Müller-Vollmer, K., 1993. Wilhelm von Humboldts Sprachwissenschaft. Ein kommentiertes Verzeichnis des sprachwissenschaftlichen Nachlasses. Schöningh, Paderborn, Germany.
Quillien, J., 1991. L'anthropologie philosophique de G. de Humboldt. Presses Universitaires de Lille, Lille, France.
Scharf, H.W., 1994. Das Verfahren der Sprache. Humboldt gegen Chomsky. Schöningh, Paderborn, Germany.
Sweet, P.R., 1978/80. Wilhelm von Humboldt. A Biography, 2 vols. Ohio University Press, Columbus, OH.
Trabant, J., 1986. Apeliotes oder Der Sinn der Sprache. Wilhelm von Humboldts Sprach-Bild. Fink, Munich, Germany.
Trabant, J., 1990. Traditionen Humboldts. Suhrkamp, Frankfurt am Main, Germany.
Trabant, J., 2012. Weltansichten. Beck, Munich, Germany.

von Ranke, Leopold (1795–1886)

Wolfgang Hardtwig, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 19, pp. 12738–12741, © 2001, Elsevier Ltd.

Abstract

Leopold Ranke is counted along with Johann Gustav Droysen (1808–1884), Barthold Georg Neibuhr (1776–1831) and Theodor Mommsen (1817–1903) as one of the most important founders of modern history in Germany and as the most influential international representative of classical German historism. His own research work and his teaching at the University of Berlin made an essential contribution to the circulation and final establishment of modern source criticism. He pushed ahead the autonomization of history as a science, by distancing himself both from the rhetorical tradition of history as an 'art,' and also from the contemporary historical philosophy, particularly that of Georg Wilhelm Friedrich Hegel. With his systematic work from sources he emphasized the importance of empiricism, and simultaneously made it repeatedly clear in theoretical comments and in his historical practice that the individual event or fact can only be interpreted in its surrounding context. His art of portrayal was essential to his success and to his, at times world recognition in the developing modern international field of history, with which style he has gone down in history as one of the founders of modern historiography.

Leopold Ranke was born in Wiehe an der Unstrut in Thuringia on the December 20, 1795 as the son of the lawyer Gottlob Israel Ranke. His family had long served as evangelical pastors in the area of Saxony between the rivers Saale and Unstrut. From 1809 to 1814 he attended the renowned Pforta school at Naumburg in Saxony, and from the spring of 1814 to the summer of 1818 he studied theology and classical literature in Leipzig. He completed his doctorate in 1817 with a (lost) dissertation on Thukydides. In 1818 Ranke started teaching history at the Fridericianum Gymnasium in Frankfurt an der Oder. In 1825 he was appointed associate professor of history at the University of Berlin, and in 1834 he was promoted to a full professor. In 1841 he was appointed official historian to the Prussian state by Friedrich Wilhelm IV, and in 1854 to the renewed parliament. Ranke was ennobled in 1865, becoming Leopold von Ranke, and in 1867 he was named chancellor of the order *pour le merité*. He rejected countless offers to teach at other universities, at Munich among others, and remained in Bérlin, where he died on the May 23, 1886. From 1880 onwards he was almost blind, and was forced to dictate his later works to assistants.

Ranke's career as a historian was influenced less by the historical science of the time, but rather by his interest in classical literature. His earliest existing studies are on the subject of translation and interpretation of classical Greek texts. Recent research has shown that the German literature of his time, Goethe, Schiller, and Friedrich Schlegel, also had an important influence on his conception of historical writing. It is also known that Ranke admired the literary form of contemporary British and French historiography. Ranke's early years were during the era of French hegemony over Germany and Europe, and the time of the rising German nationalist movement from 1806 onwards. These experiences of contemporary history left lasting impressions on Ranke's thinking. His political sympathies lay with Prussian state conservatism. On behalf of the Prussian government, but with independent control, Ranke published the *Historischpolitische Zeitschrift* from 1832 to 1836, through which the government aimed to exert a moderate conservative influence on the public. Ranke had to give up the journal after a short time because of lack of success; however, he did use the opportunity to publish a number of historical-political essays, in which he developed important methodological principles, and also began to sketch out his later, larger works. The most important of these articles are for example *Über die Zeit Ferdinands I. und Maximilians II.* (On the Time of Ferdinand I. and Maximilian II.), *Frankreich und Deutschland* (France and Germany), *Vom Einfluß der Theorie* (On the Influence of Theory) (all 1832), *Rom 1815–1823* (Rome from 1815 to 1823), *Die großen Mächte* (*The Great Powers*) (both 1833), and *Das politische Gespräch* (*The Political Conversation*) (1836). Ranke placed importance on his good relations with the Prussian court, but was also held in high esteem by the Bavarian King Maximilian II, who had attended his lectures as a student at the University of Berlin, and for whom Ranke held a series of free lectures in Berchtesgaden in the September and October 1854 which were taken down in shorthand. The king had wished for 'a short and condensed overview of the motivating ideas of the various centuries of history from the Christian era on,' and Ranke developed basic characteristics of his historical perception in eight lectures which were then published in 1888 under the title *Über die Epochen der neueren Geschichte* (On the Epochs of Recent History).

Ranke's works total 63 volumes, including nine volumes of a world history which he dictated late in his life but was unable to complete. His first work *Geschichte der romanischen und germanischen Völker von 1494–1514* (History of the Latin and Teutonic Nations from 1494 to 1514) (1824) was a great success. It already showed some of the basic characteristics of his conception of Europe, and was of historiographical importance particularly because Ranke made an exemplary critical analysis of his sources in a separate volume, *Zur Kritik neuerer Geschichtsschreiber* (On the Critical Methods of Recent Historians). In this work he raised the method of textual criticism used in the late eighteenth century, particularly in classical philology to the standard method of scientific historical writing. In 1827 he visited the archives of Vienna, Venice, Rome, and Florence and various German libraries with a stipend of the Prussian government and discovered the since

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61133-9

famous Venetian relations, legation reports from the sixteenth and seventeenth centuries, which became an important basic source for the series of his great historical works beginning in 1832. The experiences of the July Revolution in France in 1830 led him to interpret the modern history of the European states more strongly in perspective of the relation between the present and the past. From 1832 to 1836 he published the three volumes of *Die römischen Päpste, ihre Kirche und ihr Staat im 16 und 17. Jahrhundert* (Ecclesiastical and Political History of the Popes during the 16th and 17th Centuries), and from 1839 to 1847 the six volumes of *Die deutsche Geschichte im Zeitälter der Reformation* (History of the Reformation in Germany). These two works were developed in a single process and complement one another. Further main works followed in 1847–1848, *Preußische Geschichte* (History of Prussia), 1852–1861, *Französische Geschichte* (History of France), and 1859–1868, *Englische Geschichte* (History of England), as well as the incompleted world history of 1881–1888. After the foundation of the German Reich in 1870 and 1871, Ranke partly revised his *Preußische Geschichte* by emphasizing the 'German' role of Prussia in the old Reich more strongly in a second edition. Numerous studies concentrating for the most part on Prussian and Austrian history complete his work. It must be noted that the focal point of these works, with editions and investigations on the Prussian state chancellor von Hardenberg and on Friedrich Wilhelm IV, was in the area of contemporary history.

From 1830 onwards Ranke's thought centers on the history of the European states of the sixteenth and seventeenth centuries, based on his experiences of the time. However, depending on his estimation of the problems and their continuations, he did look back into the late and even the early middle ages, and on the other hand pursued the configurations of the European states not only up to 1789–1795, but also into the beginning of the nineteenth century. We can see from the focal point of the sixteenth and seventeenth centuries, the age of the secularized Renaissance papacy, the reformation and religious wars, the modern process of building states and the crystallization of the European powers system and its functioning between repeated hegemonic claims and equilibrium, that Ranke viewed the European historical process as essentially grounded in the tension between state and religion. The state and the church are buttresses of the system, at the same time however also buttresses of change, as they tend to broaden and reinforce their power so widely that they cause opposition and opposite tendencies, just as do the acting personalities themselves. In Ranke's historical perception however, each state and church, as well as their representatives, have a specific 'task,' with which they contribute to the overall human task of cultural development. The task of historians is to recognise this 'task' and the 'leading ideas,' and to portray at which points the protagonists deviate from them, causing turbulences and even historical catastrophes, and where they return to realizing these leading ideas and thereby to a harmonious equilibrium of the forces. Multiplicity is essential for the workings of this self- regulating system of states and religion, or churches. 'States are "Thoughts of God"' (*Politisches Gespräch*, ed. U. Muhlack, p. 95), and historical research pursues their development over the centuries. Ranke's great historical writings begin at the point when the medieval universalism of the church and the 'Sacrum Imperium' of the Germans started to break down and the specific European culture developed, in which the pretention of religious or state universality repeatedly appeared, but was crushed by the resistance of the state and national individualities. Multiplicity and unity are connected in this history of the great European nations and their summing together in the European powers system in such a way as to enable human energies to be realized more productively than ever before.

The recognition of this historical process has for Ranke, a value in itself. Ranke formulated this thought, which is a plea against liberal progressive thinking, repeatedly; most markedly in the sentence from the lectures '*Über die Epochen der neueren Geschichte*': 'Each age is immediate to God, and its value is not based on that which it brings forth, but in its own existence, in its own self, which however does not rule out that something else emerged from it' (von Ranke et al. 1971, p. 60). Ranke hereby opposes every functionalization of historical events and situations for contemporary interests, yet makes it clear that the present cannot be understood without the historical process and that contemporary action may not ignore the way things have come to pass. For example one can learn from the past that is not possible to re-establish the absolute monarchy or the old importance of the church in the 'Age of Revolution.' Therefore Ranke did not limit himself to a purely contemplative ideal of historical insight beyond all practical applications. In his inaugural lecture in Berlin in 1836, '*Über die Verwandtschaft und den Unterschied der Historie und der Politik*' (On the Relation and the Difference of History and Politics), he declared that an insight into the past as free from prejudice as possible, was a condition for the possibility of reasonable politics in the present. Historical insight explains the conditions under which contemporaries act. However, these must be drawn into the leading ideas of their predecessors, and may not criticize them from the standpoint of their contemporary interests. Ranke also described this concept of historical objectivity in various works, the most well known of which being the formulation: 'We have given history the task of judging the past, of teaching the world for the use of future years: the present attempt does not forbid itself such high tasks: it merely aims to show how it really was' (SW 33, VII).

Ranke founded his ideal of objectivity for the most part on historical theology, not on logical insights. The basis of historical insight was for him a certain knowledge of God: 'History is religion' ... 'At least there is the closest connection between them' (letter to Georg Waitz, December 30, 1841, Briefwerk 313). As Ranke saw God as manifested in all historical appearances, and especially in their inner relations to one another, historical insight represented for him the specific modern and up to date search for the recognition of God. The religious conviction that all historical action is carried by providence meant on one hand a limitation of the possibilities of insight, because the historian can at best guess at God's will. On the other hand this conviction lends Ranke's historical perception a harmonistical character, which has earned sharp criticism in the past decades of historism-critical historical science. His ideal of objectivism was rejected by his Berlin colleague and methodological opponent Johann Gustav Droysen but also by his pupil Heinrich von Sybel and by Heinrich von Treitschke, the protagonist of the politically

active, Prussia-oriented liberal-conservative historiography in late nineteenth century Germany.

In the field of German history of the late nineteenth and early twentieth century, however, Ranke was very highly respected. His rather too uncritical successors unfortunately produced a perception of history which did not take sufficient account of the specific forces of the modern age. As mentioned previously, Ranke himself had always related history to the present and had included the French Revolution as well as the reestablishment of the European system of powers at the Congress of Vienna in his continuity-oriented thinking. He saw in the French Revolution in particular, a dynamic element which gave the stagnating European powers system new energies and ability to integrate, so that European state system could regenerate and reinforce itself through increased nationalization. There was no longer room for industrialization, the modern competitive economy and capitalism, liberalism, even socialism and communism in this historical perception.

Leopold Ranke is counted along with Johann Gustav Droysen (1808–1884), Barthold Georg Neibuhr (1776–1831), and Theodor Mommsen (1817–1903) as one of the most important founders of modern history in Germany and as the most influential international representative of classical German historism. His own research work and his teaching at the University of Berlin made an essential contribution to the circulation and final establishment of modern source of criticism. He pushed ahead the autonomization of history as a science, by distancing himself both from the rhetorical tradition of history as an 'art,' and also from the contemporary historical philosophy, particularly that of Georg Wilhelm Friedrich Hegel. With his systematic work from sources he emphasized the importance of empiricism, and simultaneously made it repeatedly clear in theoretical comments and in his historical practice than the individual event or fact can only be interpreted in its surrounding context. His art of portrayal was essential to his success and to his, at times, world recognition in the developing modern international field of history, with which style he has gone down in history as one of the founders of modern historiography. Ranke expressly aimed to make history into a 'pleasure.' His mastery of specifically narrative forms contributed to bridging the gap between the empirical and specialized historian and the bourgeois readership of the nineteenth century. As mentioned previously, one consequences of this esthetization could also be a harmonizing view of events and situations. However, Ranke's narrative mastership did allow him to lead from the apparently coincidental individual event on to the contemplation of the most general insights, in a way that is very plausible for the historical-political consciousness of the nineteenth century. His stories deal with the responsibility of individuals for their fate, without suggesting that they could control this fate. That is the meaning of the sentence, that the historian is to deal with humanity, 'as it is, explainable or unexplainable: the life of the individual, the generations, the peoples, at times the hand of God above them' (SW 33/34, VIII).

See also: Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Historicism; History: Overview; Objectivity, History of; Religion, History of.

Bibliography

Berding, H., 1971. Leopold von Ranke. In: Wehler, H.-U. (Ed.), Deutsche Historiker, Göttingen, vol. 1, pp. 7–24.
Fulda, D., 1996. Wissenschaft aus Kunst. Die Entstehung der modernen deutschen Geschichtsschreibung 1760–1860. Berlin, pp. 296–410.
Hardtwig, W., 1991. Geschichtsreligion—Wissenschaft als Arbeit—Objektivität. Der Historismus in neuer Sicht. Historische Zeitschrift 252, 1–32.
Hardtwig, W., 1997. Historismus als ästhetische Geschichtsschreibung. Geschichte und Gesellschaft 23, 99–114.
Hinrichs, C., 1754. Ranke und die Geschichtstheologie der Gethezeit. Göttingen/Frankfurt/Main/Berlin.
Iggers, G., 1971. Deutsche Geschichtswissenschaft. Munich, pp. 86–119.
Iggers, G., von Moltke, K. (Eds.), 1973. The Theory and Practice of History, Leopold Ranke. Indianapolis, New York.
Krieger, L., 1977. Ranke. The Meaning of History. Chicago, London.
Metz, K.-H., 1979. Grundformen historiographischen Denkens. Wissenschaftsgeschichte als Methodologie. Dargestellt an Ranke, Treitschke und Lamprecht, Munich.
Muhlack, U., 1988. Leopold von Ranke. In: Hammerstein, N. (Ed.), Deutsche Geschichtswissenschaft um 1900. Stuttgart, pp. 11–36.
Ranke L von, 1897–1890. Sämtliche Werke. vol. 54. [Ed. Dove A] Leipzig, Germany.
Ranke L von, 1926. Werke Gesamtausgabe der Deutschen Akademie. [Ed. Joachimsen P] Munich.
Ranke L von, 1949. Das Briefwerk [ed. Fuchs W P] Hamburg.
Ranke L von, 1953. Die römischen Päpste in den letzten vier Jahrhunderten, with an introduction by V. Friedrich. Baethgen, Stuttgart.
Ranke L von, 1957. 12 Bücher preußischer Geschichte. [Ed. Andreas W] Hamburg, vol. 4.
Ranke L von, 1963. Die großen Mächte, politisches Gespräch. Göttingen.
Ranke L von, 1964–75. Aus Werk und Nachlaß [eds. Fuchs W P and Schieder T|. Munich/Vienna, vol. 4.
Ranke L von, 1988. Und die moderne Geschichtswissenschaft. Stuttgart.
Ranke L von, 1989. Über die Epochen der neueren Geschichte. Darmstadt.
Ranke L von, 1995. Die großen Mähte, politisches Gespräch. Frankfurt/Main/Leipzig.
Vierhaus, R., 1957. Ranke und die soziale Welt. Münster.

Voting and Direct Democracy

Lawrence LeDuc, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In some jurisdictions, citizens have the opportunity to cast a vote on issues rather than only for political parties or candidates. This article considers the factors that shape voting behavior in a referendum or other type of direct vote on issues, examining participation, voter decision making, and the implications of direct issue votes for the health of democracy more generally. The various forms and institutions of direct democracy are considered, as well as the dynamics of initiative and referendum campaigns, the sources of campaign information, and the impact of information and predispositions on opinion formation and voting decisions. Four models of voter decision making are examined, and their relative utility in explaining the outcome of direct issue votes is evaluated.

In some jurisdictions, citizens have the opportunity to cast a vote on issues rather than only for political parties or candidates. In this article, we consider the factors that shape voting behavior in a referendum or other type of direct vote on issues. How do voters make decisions in such circumstances? Do they participate in issue votes at the same rates as in regular elections? What are the main characteristics of initiative and referendum campaigns? And what are the implications of direct democracy for the health of democracy more generally?

Structure and Context

Not all countries practice direct democracy and in those that do it can take on different forms. Initiatives undertaken directly by citizens and referendums called by a government are the two principal institutional models, but there are many variations in these, both in the specific legal requirements and in the frequency of their use.

Varying Forms and Institutions of Direct Democracy

The usage of referendums and initiatives has increased dramatically around the world, both in the number of countries employing such devices and in the number of issues being put to a direct vote in some form (International IDEA, 2008; Altman, 2011; Qvortrup, 2014). There are some established democracies (e.g., Switzerland, Ireland, Italy, Uruguay) in which the institutions and processes of direct democracy are widely used and others in which referendums are employed only occasionally to allow citizens to express their views directly on an important political issue or for more specific purposes such as constitutional amendments or the ratification of international treaties.

Direct democracy comes in many forms, and the issues that arise under different institutional rules and processes are not always the same. A referendum is sometimes mandated by law in countries or states where constitutional amendments require a popular vote, or it may be called by a government or legislature that seeks to test public support for a proposal. Citizen initiatives, on the other hand, are brought forward by members of the public without the sanction of the government. In New Zealand, a nonbinding vote of the people on any issue may be triggered either by citizen initiative or by the government for consultative purposes. Italy employs the *abrogative* referendum or popular veto, in which citizens may force a binding public vote on a law which has already been adopted by the legislature. In many cases, more than one route to a popular vote is possible. Switzerland, one of the countries with the greatest experience in the practice of direct democracy, does not use consultative referendums initiated by the government, but several other types of initiative processes and mandatory constitutional referendums are found there.

In recent years, there have been a number of important referendums in European countries dealing with membership in the European Union or with changes in major European treaties. The rejection by voters in France and The Netherlands of a proposed European Constitutional Treaty ended an attempt in 2005 to create a new constitutional framework for the European Union. Some of the proposals contained in that document were later adopted by the EU member countries in the form of the Lisbon Treaty in 2009, but only Ireland, where approval by the public in a referendum is required by the constitution, held a direct vote on that treaty. There are also newer democracies in which a referendum is often the preferred device for obtaining public approval of a new constitution or for legitimizing a process of fundamental political change such as secession or the restructuring of state institutions. Referendums preceded the declaration of independence of East Timor from Indonesia in 1999 and of South Sudan from Sudan in 2011.

Although there has never been a national referendum on any issue in the United States, the use of direct democracy is widespread in that country at the state and local level. All U.S. states except Delaware provide for amendments to their state constitutions to be submitted to a popular vote, and most states have some type of procedure for putting an issue to a referendum of the people (Cronin, 1989). Twenty-four states have provision for citizen initiatives, in which a group of citizens, upon obtaining a specified number of signatures, may be able to place an issue on the ballot. 'Ballot propositions,' as they are generally called in the U.S., have thus become an

important component of American democracy. In the November 2012 presidential election, there were 174 ballot propositions to be decided in 38 states, including proposals for the legalization of marijuana which were approved by voters in Colorado and Washington and same-sex marriage in Maine, Maryland, and Washington.

How and When Do Initiatives and Referendums Occur?

Governments call referendums for different reasons. In a few cases, such as the referendums that have taken place in Ireland on European treaties such as Nice (2001, 2002) or Lisbon (2008, 2009), the referendums were constitutionally mandated. The same was true of the 1999 Australian referendum on the monarchy, as all constitutional amendments in Australia must be submitted to a national vote. In the case of the Danish (2000) and Swedish (2003) referendums on the Euro, they were *politically* necessary, since the governments in those countries could not have risked adopting the Euro without a public consultation. Similarly, none of the governments of the Nordic countries in 1994 were willing to undertake the historic decision to join the European Union without a referendum (Jenssen et al., 1998). A variety of political calculations often enter into a government's decision to call a referendum on a particular issue (LeDuc and Svensson, 2002). The 1997 Scottish and Welsh devolution referendums, together with a 1998 referendum on local government in London, were in effect all part of the Blair government's wider agenda of institutional reform, but none of these were *legally* required.

Sometimes, the political decisions involved in calling a referendum are part of an electoral strategy. In Quebec, the leadership of the Parti Québécois made a calculation in the course of devising its 1976 election campaign strategy that the commitment to conduct a referendum on the issue of Quebec sovereignty would allow it to decouple the sovereignty question from other election issues, thereby placing it in a more competitive electoral position (Pammett and LeDuc, 2001; Clarke et al., 2004). The party's subsequent victory in the election justified that strategy, but left it with the commitment to hold such a referendum later, which took place in 1980. A second Quebec referendum in 1995 has now made it all but impossible that Quebec sovereignty could ever be achieved without resorting to a third referendum. A similar situation exists in Norway. Having voted twice (1972, 1994) against European Union membership, no Norwegian government would now consider joining the EU without a positive referendum vote.

Government-initiated referendums also frequently take place because a governing party finds itself divided on an important issue. The 1986 referendum in Spain on NATO membership came about for this reason. Similarly, the decision of the Swedish government to hold a referendum on nuclear energy policy in 1980 was taken because all of the major political parties preferred to avoid dealing with that contentious issue in an election campaign, and also because the governing Social Democrats were deeply divided on the issue (Gramberg and Holmberg, 1988). Referendums that come about in this way present a different context for voters than those that arise out of legal necessity or that are part of a partisan political agenda. Partisan cues, which are so important to the voting decision in ordinary elections, may be entirely absent in such situations, or voters may receive mixed messages from well-known partisan political actors (Higley and McAllister, 2002).

Campaigns and Direct Democracy

Earlier literature on direct democracy gave little attention to referendum and initiative campaigns, focusing instead on the institutional rules and norms of direct democracy and on its possible consequences for parliamentary governance and the protection of minority rights. But more recently, the effects of the campaign in providing information to voters on sometimes complex issues and the role of resources and media in affecting outcomes have received greater attention. Referendum and initiative campaigns are *real* campaigns, in which political actors seek to win votes rather than to merely carry on the debate of an issue.

Direct Democracy and Elections

In many instances, a referendum or initiative vote takes place at the same time as an election, usually a regular election for legislative or presidential candidates or other public officials. The practice of holding an election and referendum or initiative vote concurrently is frequently defended on the grounds of minimizing administrative costs, as well as maximizing voter turnout. But one significant disadvantage of this practice is that the referendum or initiative issue may become submerged in the election campaign. On American ballots, propositions do not always receive as much attention as the campaigns for elective offices. Switzerland does not hold referendums to coincide with elections, instead scheduling direct votes on issue proposals three or four times a year, each one typically involving two to five questions.

On balance, the quality of deliberation of an issue is probably better served when debate on an issue does *not* become caught up in the politics of an election campaign. But the risk of holding a separate vote on a single issue is that of lower turnout or voter indifference. This was never a problem with either of the two Quebec referendums, both of which drew significantly higher turnout than the most nearly comparable provincial or federal elections. Neither was it a problem with most of the votes on EU membership that have taken place in recent years or with Australian referendums, in which voting is compulsory. But, low turnout and general voter disinterest in some Irish referendums has not infrequently been a problem in that country, and has sometimes led to the holding of a second referendum on the same issue.

Question Wording

The choice presented to voters in a referendum is typically to vote either YES or NO on a specific question that appears on the ballot. In some instances, the question will be relatively clear and straightforward, as in the 2014 Scottish referendum ("Should Scotland be an independent country?") or the 1994 Norwegian referendum ("Should Norway become a member of the European Union?"). When a governing party has the

opportunity to write the question, it often does so with its own political objectives in mind. The questions in both Quebec sovereignty referendums were written by a government actively seeking a YES vote, incorporating in the wording appealing concepts such as a 'new political and economic partnership' (1995) or the 'equality of nations' (1980). The question used in the 1986 Spanish referendum on NATO membership sought approval under 'the terms set down by the government' but did not specify what those terms were. This facilitated mobilization of supporters of the governing party, many of whom had previously opposed NATO membership. It also enabled the government to structure the debate more around the *terms* that it sought to impose on continued Spanish membership in NATO rather than on the fact of membership itself, a campaign strategy that ultimately proved successful.

The question used in the 1999 Australian referendum on abolishing the monarchy presented a difficult choice to voters because it prominently raised *two* issues – the abolition of the monarchy and the method of election of a new president. In the campaign, it became clear that many voters who favored abolition of the monarchy were also opposed to the provision of the proposal that allowed parliament to elect the president. The Irish referendums on the Nice treaty suffered from a similar problem, in part because the actual question wording in Irish referendums is often technical and legalistic. The question asked voters to "permit the state to ratify the Treaty of Nice and certain related Acts," but said nothing about the specific provisions of those documents. American state ballot propositions often suffer from problems of clarity due to highly technical language or multiple items contained in a proposal. Not infrequently, question wording can be manipulated in various ways to present a different type of choice on the same issue. In Maine, voters in 2012 were presented with a proposal to "allow the state to issue marriage licenses to same-sex couples" while in Minnesota the proposed constitutional amendment stated that "only a union of one man and one woman shall be valid or recognized as a marriage in Minnesota."

Sources of Information

Surveys taken in the aftermath of a referendum campaign regularly show that 'insufficient information' is one of the most common complaints of citizens about the referendum process.

When an issue is partisan, information is likely to flow to voters through the political parties. In the case of citizens' initiatives, which are often undertaken by groups or organizations formed for the purpose, these may become the primary sources of information for voters. Particularly in circumstances such as the European treaty referendums or approval of constitutional amendments, where large and complex packages of proposals are put to a vote, information becomes critical. A similar problem often occurs in the case of U.S. ballot propositions, where complex proposals combined with highly biased sources of information can easily overwhelm voters.

In some jurisdictions, this problem is addressed by having neutral authorities provide 'balanced' information. In Ireland, for example, an independent referendum commission is charged with the responsibility of disseminating unbiased information to the public through its own publications, and of promoting debate and discussion of the issue(s) over the course of the campaign. In California, the Secretary of State is legally responsible for publishing a *Voter Information Guide*, which contains information regarding every proposition on the ballot. Included in the booklet is a summary of the proposal, a statement by the Legislative Analyst regarding its meaning and consequences, and unedited statements by groups supporting and opposing the proposition.

Campaign Regulation

Initiative campaigns in some U.S. states are sometimes criticized as unfair because of the disproportionate amounts of money spent by one side or the other (Sabato et al., 2001). An individual, group, or organization that has been able to bear the high costs of qualifying a proposal for the ballot is often able to spend far more money in a campaign to secure its passage than opponents are able to raise to defeat it. Similarly, access to the media through which voters obtain information on ballot propositions can be a contentious matter. In countries where some media such as television stations are state controlled, parties or groups on one side of an issue may be able to obtain preferential treatment or greater access. In those jurisdictions where all media are privately owned, those with more resources may hold the advantage.

The answer to difficulties posed by the inequality of resources is sometimes seen in tighter regulation of referendum campaigns. Steps can be taken to assure that the amount of money spent in a campaign falls within defined limits, that campaigners for both sides of an issue have sufficient access to the media, and that voters hear the arguments put forward on both sides of an issue. Subsidies are used in some (primarily European) jurisdictions to level the field. Channeling all activities through officially sanctioned 'umbrella' committees may be another way to assure a more evenly balanced campaign. But such regulations are difficult to formulate and sometimes even more difficult to enforce. Ireland has perhaps the most tightly regulated environment, while other jurisdictions such as Switzerland maintain much lighter regulation of campaign activities and finances.

Participation and Engagement

The levels of participation and engagement of citizens voting on an initiative or referendum proposal are sometimes insufficient to provide a true and accurate reading of public preferences or to give full legitimacy to the result. Voting choices may in some circumstances be based on limited information, or even on misinformation transmitted by the opposing sides in a campaign.

Voter Turnout

Turnout fluctuates more widely in referendums than it does in national elections. In general, it tends to be lower, but can sometimes rise to higher levels when a particular issue engages wide voter interest or an intense campaign is waged by interested groups. When turnout is low, the ability to mobilize one's own supporters counts for more. When it rises, it is

generally because the issue itself is perceived as an important one. Some referendums in which turnout registered higher than that of a comparable national election were the 1994 EU membership referendum in Norway, both Quebec sovereignty referendums, the 1992 Canadian constitutional referendum, and the Danish and Swedish referendums on the Euro. Turnout in both Irish referendums on the Nice treaty was very low in comparison with national elections, and the low turnout of only 35% of voters in the first (2001) referendum was widely blamed for the outcome.

Turnout is also typically low in Swiss initiative and referendum votes, sometimes falling below 40%. But turnout also tends to be low in Swiss elections, and referendum participation is frequently higher than electoral participation, depending on the salience of the particular issue being considered. In the case of a 2005 vote on registration of same-sex partnerships, for example, turnout was a full 11% higher than in the 2003 federal assembly election. Other recent Swiss referendums that have drawn higher rates of participation were a 2004 vote on acquisition of citizenship and a widely publicized 2009 initiative banning the construction of minarets (both 54%).

Turnout is also a serious problem in many of the initiative and referendum votes held in the U.S. states. Typically, such items appear on an electoral ballot together with other public offices. But the turnout then depends largely on the election in which the vote is occurring, and not on the propositions per se. Because turnout in U.S. presidential elections tends to be higher than in off year or state elections, items that appear on a presidential ballot achieve higher levels of participation. But turnout on U.S. ballot propositions must also be measured in terms of the total vote cast on the specific item, rather than as a percentage of those going to the polls, since many voters will fail to vote on some or all of the propositions appearing on the ballot. A 'drop-off' of as much as 10% is fairly typical, but for some propositions it can be much higher.

Some jurisdictions impose a turnout quota on referendum votes in order to assure an adequate level of participation. In Italy, a referendum result is ruled invalid unless 50% of voters participate. The relatively low participation rates in Swiss votes have in general seemed to be less of a problem in that country than in some other instances such as the first Nice referendum in Ireland, where low turnout appears to have directly affected the outcome. Whatever the appropriate level might be, it is clear that turnout is connected directly to legitimacy. A vote in which only a small minority of citizens participate may create circumstances under which the legitimacy of the process itself may come under challenge. But, turnout cannot be considered in isolation. It is affected by many other factors including the timing of votes, party mobilization, levels of information, and the nature of the campaign.

Opinion Formation

Not uncommonly, the issues that are presented to voters on a referendum or initiative ballot are ones with which they are for the most part unfamiliar until they are called upon to make a choice. Voters cannot generally be expected to have well-formed opinions on an issue that has not previously been a subject of any broad public debate. Some referendums particularly fitting such a profile are those that involve complex international treaties, large packages of constitutional provisions, technical matters such as bond issues, or various types of institutional change. The 2011 British referendum on electoral reform presented voters with the opportunity to change the electoral system from First-Past-the-Post to an Alternative Vote (AV) model, but few voters would have been familiar with AV prior to the commencement of the debate. In these circumstances, campaigns based on misinformation or distortion are more likely to succeed. Such referendum campaigns are often highly volatile, and frequently result in the rejection of a proposal by a risk-averse electorate that lacks sufficient information regarding the implications of a proposed change.

By contrast, where an issue is a familiar one in the political arena, or where parties take clearly competing positions, the voting decision is easier and tends to be made earlier in the campaign. Here, opinion is much firmer and less subject to rapid change or sudden reversal. Voters will often have strong cues based on partisanship or ideology, and be receptive to arguments presented by familiar and trusted political leaders. The 1995 Quebec sovereignty referendum is a good example of a campaign in which the PQ government knew that it could count on the nearly universal support of the partisan voters who had brought it to power, but also needed the votes of others in order to secure a majority for its sovereignty proposal. The Danish (2000) and Swedish (2003) referendums on the Euro, an issue which had been debated and discussed extensively in national politics, also provide good examples of this pattern, in which the government believed that it could win the referendum by persuading a relatively small number of undecided voters to support the YES side.

Zaller's (1992) model of opinion formation is particularly well suited to the study of public opinion and voting behavior in referendums. As he argues, any process of opinion formation proceeds from an interaction of *information* and *predisposition*. When strongly held predispositions are merely reinforced by the campaign, referendums begin to take on some of the characteristics of elections, in which factors such as party identification or ideological orientation typically play a crucial role. But when parties are internally divided, ideological alignments are unclear, or an issue is new and unfamiliar, voters might be expected to draw more of their information from the campaign discourse. Under these circumstances, the outcome of the contest is unpredictable and information becomes critical.

Voting Behavior in Referendums

The voting decision in referendums on issues which have been discussed extensively in other political arenas, or in which there are strong linkages to partisanship or ideology follows more conventional models of voting choice and tends to be made earlier in the campaign. In contrast, a context in which there is little partisan, issue, or ideological basis from which voters might tend to form an opinion introduces greater uncertainty. In these settings, other models of voting choice involving 'short-cuts' or 'risk aversion' become more applicable.

Models of Voting Choice

The context in which a referendum or initiative vote occurs and the type of issue on which voters are asked to render a decision will act to shape their decision process. When an issue is strongly connected to existing partisan or ideological cleavages, or when it has been the subject of an ongoing political debate in which the public is engaged, the voting choice will be similar to that found in elections. On an issue such as Quebec sovereignty, for example, we would expect predispositions to be strong and the levels of information to be high. Because the voting decision will be easier in these circumstances, we would expect high levels of engagement and participation and relatively modest movement in sentiment over the course of the campaign, since the opinions of many citizens are already well formed and strongly held.

The opposite is more likely to occur in situations where an issue is new and unfamiliar to voters, parties are divided, or the linkage to other political issues is weak. Predispositions will be low, and information flowing from the campaign may act to shape opinions. Such campaigns tend to be much more volatile, sometimes leaving the outcome in doubt until the votes are counted. Turnout is often lower in such circumstances, owing to voter uncertainty or indifference. Proposals are often rejected by voters under these conditions, in part because of low information and/or risk aversion. The 2011 British referendum on electoral reform displayed many of these characteristics, with a turnout of 42% and a decisive NO vote. In a study of that referendum, Whitely et al. (2012) identified and tested four decision models used by citizens to determine whether to participate and/or how to vote. They labeled these rational choice, cognitive engagement, heuristics, and mobilization. Although the terminology employed sometimes varies, these models of choice will be familiar to scholars studying voting behavior in other countries where direct democratic votes are more commonplace (see, for example, Bowler and Donovan, 1998; Buzzer and Marquis, 2002).

Rational Choice

Students of electoral behavior will be familiar with rational choice models, in which recognized self-interest is hypothesized to drive voting decisions. Voters may be expected to support a proposal if they see a benefit to themselves accruing from its adoption or to vote to reject it if they detect a threat to their interests. But for this model to apply, voters must have sufficient information about a proposal in order to determine its possible benefits or costs. Uncertainty or indifference may lead to abstention, which under this model is also a clearly rational decision. Why participate if the benefit is uncertain or the available information regarding its effects is inconsistent?

Applied to the British example, Liberal Democratic voters or those identifying with other smaller parties might have been expected to support the proposed reform, since the potential benefits to their party were more evident. Labour voters, on the other hand, might have been more indifferent to the proposal, since the potential costs or benefits to their party were much less clear. Further, because levels of information regarding the proposed reform were low, and the campaign contained much misinformation, many citizens could reasonably be expected to abstain. Turnout in the referendum in fact might have been even lower than it was had it not been scheduled concurrently with local and regional elections in many parts of the country.

Cognitive Engagement

A cognitive engagement model bears some similarity to rational choice in that voters require some information and knowledge about a proposal in order to come to a voting decision. On an issue such as electoral reform, engaged and politically knowledgeable citizens are likely to have sufficient information. In the study of the 2011 British referendum, Whitely et al. (2012) found that political knowledge was strongly related to a YES vote. But they also argue that disengagement or lack of knowledge, rather than leading to abstention, was more likely to lead to support of the status quo – in this instance meaning a NO vote on the AV proposal.

There is considerable survey evidence in other studies of referendum voting, particularly on certain types of issues, that a combination of risk aversion and lack of information can easily prompt a NO vote rather than abstention (LeDuc, 2009). Respondents to postelection surveys often give 'lack of information' as their reason for a negative choice. We might add that this dynamic is frequently reinforced by the nature of modern media campaigns (de Vreese, 2007). Attacking a proposal is a common campaign tactic, and opponents of a proposal do not necessarily have to make a coherent case against it. Rather, they can attack it on a number of different fronts, delivering the not so subtle message "when in doubt, vote NO." This strategy was particularly evident in the campaigns against the European Constitutional Treaty (Hobolt, 2005).

Heuristics

In addition to seeking information that will allow them to come to a decision, voters also look for cues – sometimes also referred to as 'short cuts' in the voting behavior literature. As Lupia and McCubbins (1998) put it, citizens learn 'what they need to know' in order to be able to cast a vote on a proposal. In the case of the British AV referendum, the Prime Minister's very public support of the NO side communicated a strong message to Conservative voters. In many initiative and referendum campaigns, opinions will often be shaped by the cues provided by political parties, prominent opinion leaders, or groups and organizations engaged in the campaign. This type of knowledge is often sufficient. Voters do not need to know all of the details of a proposal in order to reach a decision.

An important variation on this theme is found in what are generally called 'second-order effects' (Garry et al., 2006). Decisions may be affected by opinions about the particular individuals or groups involved in a campaign rather than the substance of a proposal. A government that calls a referendum may find that the vote becomes a test of its own (un)popularity rather than a vote on the substance of its proposal. Constitutional initiatives undertaken by unpopular governments often fail, partly for that reason, as did the 1992 Canadian Constitutional package (LeDuc, 2002, 2003). Conversely, a popular leader or a newly elected government may be better positioned to build support for an initiative. The Blair government moved quickly following its election in 1997 to bring forward its devolution and local government reform proposals for submission to a popular vote.

Mobilization

As noted earlier (Section Campaigns and Direct Democracy), initiative and referendum campaigns are *real* campaigns, in which parties or groups seek to win votes rather than merely engage in an issue debate. In the British referendum campaign used as an example here, the two rival campaigns (*Yes to Fairer Votes* and *No to AV*) actively sought to communicate their messages to voters through the media and other sources and to motivate their potential supporters. In a short and intense campaign, particularly on an issue with which citizens are relatively unfamiliar, such activity has an impact both on opinion formation and participation. As in an election campaign, getting one's message across and getting voters to the polls can matter a great deal.

Referendums can differ considerably in levels of mobilization, but resources and organization clearly matter. Whitely et al. (2012) found that this factor was of modest importance in explaining the outcome of the 2011 British electoral reform referendum, at least in comparison with other variables in a multivariate analysis of the vote – perhaps because the outcome of that contest was never really in doubt. But in much closer contests such as the 1995 Quebec sovereignty referendum or the 2000 Danish referendum on the Euro, mobilization, both cognitive and organizational, may be of critical importance in determining the outcome. Conversely, the failure of one side or the other to effectively mobilize its potential supporters can also lead to an unexpected outcome, as in the case of the 2001 Irish referendum on the Nice treaty.

Conclusion

Direct issue votes differ widely in terms of the choices presented to voters and the context in which such votes take place. As the studies cited here demonstrate, the decision processes followed by voters in such contests, although now fairly well known, continue to suggest high levels of uncertainty with respect to the outcome of referendum or initiative votes. Voters follow different paths both to a decision to participate and how to vote. As the study of the 2011 British referendum found, there is some evidence in support of all four models of choice discussed here, even within the context of the same referendum. Not all voters reach their decision of whether and how to vote in exactly the same way. But we would expect that, even in different settings with other choice sets, all four models of choice are likely to have some validity, although the relative weight of these alternative decision processes may vary considerably. Direct democracy continues to gain support around the world, and it remains popular with voters in those jurisdictions that conduct such votes regularly. But, as in elections, voting choice continues to be a subject of continuing interest to both scholars and practitioners of democratic politics. Issue votes, when they occur, are closely intertwined with more traditional forms of electoral politics.

Bibliography

Altman, David, 2011. Direct Democracy Worldwide. Cambridge University Press.

Bowler, Shaun, Donovan, Todd, 1998. Demanding Choices: Opinion, Voting, and Direct Democracy. University of Michigan Press, Ann Arbor, MI.

Buzzer, Michael, Marquis, Lionel, 2002. Public opinion formation in Swiss federal referendums. In: Farrell, David, Schmitt-Beck, Rüdiger (Eds.), Do Political Campaigns Matter? Campaign Effects in Elections and Referendums.

Clarke, Harold D., Kornberg, Allan, Stewart, Marianne, 2004. Referendum voting as political choice: the case of Quebec. British Journal of Political Science 34, 345–355.

Cronin, Thomas, 1989. Direct Democracy: The Politics of Initiative, Referendum and Recall. Harvard University Press, Cambridge, MA.

Garry, John, Marsh, Michael, Sinnott, Richard, 2006. Second-order versus issue-voting effects in EU referendums: evidence from the Irish Nice treaty referendums. European Union Politics 6, 201–221.

Gramberg, Donald, Holmberg, Sören, 1988. Preferences, expectations and voting in Sweden's referendum on nuclear power. Social Science Quarterly 67, 379–391.

Higley, John, McAllister, Ian, 2002. Elite division and voter confusion: Australia's republic referendum in 1999. European Journal of Political Research 41, 845–862.

Hobolt, Sara B., 2005. When Europe matters: the impact of political information on voting in EU referendums. Journal of Elections, Public Opinion and Parties 1, 85–109.

International IDEA, 2008. Direct Democracy: The International IDEA Handbook. International IDEA, Stockholm.

Jenssen, Anders T., Pesonen, Pertti, Gilljam, Mikael (Eds.), 1998. To Join or Not to Join: Three Nordic Referendums on Membership in the European Union. Scandinavian University Press, Oslo.

LeDuc, Lawrence, 2002. Referendums and elections: how do campaigns differ? In: Farrell, David, Schmitt-Beck, Rüdiger (Eds.), Do Political Campaigns Matter?: Campaign Effects in Elections and Referendums. Routledge, London, pp. 152–153.

LeDuc, Lawrence, 2003. The Politics of Direct Democracy: Referendums in Global Perspective. Broadview Press, Toronto.

LeDuc, Lawrence, 2009. Campaign tactics and outcomes in referendums: a comparative analysis. In: Setälä, Maija, Schiller, Theo (Eds.), Referendums and Representative Democracy: Responsiveness, Accountability and Deliberation. Routledge, London.

LeDuc, Lawrence, Svensson, Palle (Eds.), 2002. Interests, information and voting in referendums. Special Issue of the European Journal of Political Research 41 (6).

Lupia, Arthur, McCubbins, Matthew, 1998. The Democratic Dilemma: Can Citizens Learn What They Need to Know? Cambridge University Press, New York.

Pammett, Jon H., LeDuc, Lawrence, 2001. Sovereignty, leadership and voting in the Quebec referendums. Electoral Studies 20, 265–280.

Qvortrup, Matt (Ed.), 2014. Referendums Around the World: The Continued Growth of Direct Democracy. Palgrave, Basingstoke.

Sabato, Larry, Ernst, Howard, Larson, Bruce (Eds.), 2001. Dangerous Democracy. Rowan and Littlefield, Lanham, MD.

de Vreese, Claes H. (Ed.), 2007. The Dynamics of Referendum Campaigns. Palgrave Macmillan, London.

Whitely, Paul, Clarke, Harold D., Sanders, David, Stewart, Marianne C., 2012. Britain says NO: voting in the AV ballot referendum. Parliamentary Affairs 65, 301–322.

Zaller, John R., 1992. The Nature and Origins of Mass Opinion. Cambridge University Press, New York.

Voting, Explanations of: Social Class

Geoffrey Evans, University of Oxford, Oxford, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 24, pp. 16311–16314, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The idea of 'class voting' refers to the tendency for citizens in a particular social class to vote for a particular party, political candidate, or groupings of these, rather than an alternative option when compared with voters in another class, or classes. This article overviews past and current comparative research into changes in and explanations of such class-based political behavior and signals its continued significance. Particular emphasis is placed on the implications of choices of method and measurement for estimating the strength of the association between class and vote, and on the importance of developing and testing theories that can account for variation in the strength and form of the class–vote relationship.

The idea of class voting appears straightforward: it refers to the tendency for voters in a particular class to vote for a particular party, political candidate, or groupings of these, rather than an alternative option when compared with voters in another class, or classes. In other words, it describes a pattern of association between class and vote. Yet this simplicity hides considerable intricacy and ambiguity. The definition of social class in and of itself introduces much confusion – class is a much debated, even 'essentially contested,' concept. The issues raised become more complex still when class concepts are translated into measures of class position that allocate voters to classes and when attempts are made to summarize statistically the class–vote association. Explaining patterns of class voting and understanding their implications raises many more unresolved questions. All of these theoretical and methodological issues have at times been the focus of dispute. In addition, however, there has been one major substantive argument that has received more attention than all other questions and which has touched on all of them: has class lost its importance as a source of voters' political preferences?.

Historically, interest in class voting arose in response in part to the failed agenda of Marxism. For classic texts in political sociology electoral politics was an expression of 'the democratic class struggle' (Anderson and Davidson, 1943) that replaced, at least for the short-term, the predicted class-based revolution. This in turn led to a focus on class voting as a dispute between just two classes, the working class and the middle class, and their representatives, parties of the left and right. Early research also relied on data obtained at the level of electoral constituencies. Given the well-known problems with inferring individual vote choices from aggregate data this work had to make strong assumptions about how voters in different classes actually voted. To help overcome this limitation, modern social scientific studies of class voting have focused on its measurement through surveys of voters. From its beginnings in a few national election studies this data base has expanded through time and space to create a large body of evidence on class voting that now informs academic debate.

The Two Class/Two Party Era

The baseline for many current studies was set by Robert Alford's (1967) analysis of trends in class voting in four Anglo-American democracies (Australia, the UK, Canada, and the USA) between 1936 and 1962 using a measure of social and economic position in which occupations were aggregated to form a dichotomous manual/nonmanual class division. He also introduced the most commonly used, cited, and criticized measure of the level of class voting: the 'Alford index.' The Alford index is simply the difference between the percentage of manual workers that voted for left-wing political parties on the one hand and the percentage of nonmanual workers that voted for these parties on the other. This became standard practice in the cross-national and over-time analyses that followed, and which usually concluded that class voting is in decline. Lipset (1981), for example, presented evidence of a downward trend in the UK, Germany, and the USA between 1945 and 1980; while Inglehart (1997) points to evidence of a continuing decline from the early 1980s onwards. Indeed, whether or not class voting in modern industrial societies has declined is, for many commentators, no longer an issue. The question is *why* has it declined? A plethora of explanations have been proposed to account for this outcome (see Goldthorpe, 1996; Manza et al., 1995): some assert that it has lost its importance as a determinant of life-chances and in consequence its role as a source of divergent political interests because the working class has become richer, white-collar workers have been 'proletarianized,' or because of extensive social mobility between classes. Others argue that 'postindustrial' cleavages, such as gender, race, ethnicity, public versus private sector, and various 'identity groups,' are emerging and replacing class-based conflict. Postindustrialism also impacts on class voting through the rise of postmaterial values as a basis of party preference, which has led to the 'new left' drawing its support from the middle classes thus weakening the class basis of left-right divisions; while rising levels of education and 'cognitive mobilization' have supposedly produced voters that are calculative and 'issue oriented' rather than driven by collective identities such as class.

Whatever their strengths and weaknesses – and there are many questions and doubts that remain unresolved in each case – all of these explanations assume that there is indeed a widespread, secular decline in class-based political action which then has to be explained. However, answers to questions about change have been strongly conditioned by choices of method and measurement. The literature on the decline of class

voting that followed Alford's work has been characterized by shortcomings in both these respects.

For example, one problem with the manual/nonmanual distinction is that it obscures variations in the composition of the manual and nonmanual classes. Any changes in their composition may lead to spurious change in estimates of class voting. Thus, if skilled manual workers are more conservative than unskilled workers and the number of skilled workers increases, the difference between manual and nonmanual workers will decline even if the relative political positions of skilled, unskilled, and nonmanual workers remain the same. With this and similar indices the measurement of the class–vote association is open to confounding by changes in the shape of the class structure and in the general popularity of the parties. In other words, this type of index confuses differences in the marginal distributions of the variables with differences in the association it is supposed to measure. The same problem applies to the apparently more sophisticated ordinary least squares regression techniques used in, among others, Franklin, Mackie and Valen's (1992) ambitious 16 nation study that represents, arguably, the culmination of this tradition.

Recent Innovations

In recent years, therefore, the rather atheoretical manual/nonmanual representation of class voting has been to a large degree superseded as researchers have taken into account more nuanced ideas about class structure. Most influential has been the class schema developed by John Goldthorpe and his colleagues (Erikson and Goldthorpe, 1992). First used to study patterns of class voting over time in Britain by Heath *et al.* (1985) this instrument, unlike many other measures of class, has been empirically validated. The main classes identified by the Goldthorpe schema are the petty bourgeoisie (small employers and self-employed), the service class, or salariat (professional and managerial groups), the routine nonmanual class (typically lower grade clerical 'white-collar workers'), and the working class (foremen and technicians, skilled, semi-, and unskilled manual workers). The principal distinction underlying the distinction among the employee classes in the schema is between a service contract in which employees receive not only salaried rewards but also prospective elements – salary increments, job security and pension rights, and, most importantly, well-defined career opportunities – and a labor contract, in which employees supply discrete amounts of labor, under supervision, in return for wages which are calculated on a 'piece' or time basis. As the employment relationship of the service class in relatively advantageous in terms of employment and payment conditions, occupational security, and promotion prospects, its members have a stake in preserving the status quo. In contrast, the disadvantages of the labor contract can explain why the working class provides a basis of support for the redistributive programs of the left.

This shift to greater complexity in the measurement of class has been accompanied by a similar move away from the measurement of political choice as a dichotomy of left versus right (or left versus non-left) to a fuller representation of the voters' spectrum of choice at the ballot box. Apart from its simplicity, the main reason for the use of a dichotomy to represent voter choice seems to have been a desire to make systematic cross-national and over-time comparisons. Unfortunately, the selective nature of what is being compared undermines any true comparability. The problem is analogous to that faced in the analysis of class position. Changes in the composition of composite categories such as 'left' or 'non-left' may lead to spurious changes in estimates of class voting. The use of dichotomies to represent vote choices and social classes also precludes from observation any processes of class–party realignment. The concept of class *re*alignment in voting implies a change in the pattern of association between class and vote without any change in the overall strength of this association – i.e., without class *de*alignment (or, of course, increased alignment). But this cannot be discerned if the distinction between realignment and dealignment is prevented by restricting the numbers of parties and classes to two.

The final innovation of the 1980s and 1990s is in the statistical measurement of the class–vote association, with a move from Alford type indices to ones that are based on odds ratios, which measure the strength of the relationship between class and vote independently of the general popularity of political parties or changes in the sizes of classes. The use of logistic modeling techniques likewise capitalizes on this basic feature of the odds ratio and allows the relationship between more complex class and party systems to be summarized, thus complementing and facilitating greater sophistication in the representation of class and vote.

Research using these advances comes to rather different conclusions than those in the two-class, two-party tradition. Studies in the UK (i.e., Heath *et al.*, 1995) have found little evidence of declining class voting and have concluded that 'trendless fluctuation,' or at most a 'one-off' change, best captures the pattern of association in class voting over time. The USA, by comparison has seen the growth of a new class – vote cleavage – between those who vote and those who do not – as a deficit of working class interest representation that has been associated with increased class differences in participation (Hout *et al.*, 1995). Various studies reported in Evans (1999) show that although there is evidence for a linear decline in left versus non-left class voting (Nieuwbeerta and de Graaf, 1999), it is not typical. For example, when examined over the long-term, levels of class voting in the UK were found to have increased during the 1940s and 1950s before falling back in the 1960s; while the 1990s have seen increased levels of class voting in new post-communist democracies as these societies have undergone the rigors of marketization. Only in certain Scandinavian countries is there robust evidence of a decline from an unusually high degree of class voting to levels similar to those in other Western democracies.

New Directions

The debate on class voting has been strong on evidence and weak on theory. Most scholars have assumed a sociological, relatively deterministic account of the transition to industrial and postindustrial politics, but there are those who have rejected these in favor of more voluntaristic models. Kitschelt (1994), for example, argues that the electoral fortunes of European social democratic parties are largely determined by

their strategic appeals, rather than by secular trends in the class structure – a line of reasoning that echoes Sartori's (1969) influential emphasis on the importance of organization, and especially parties, in the creation of class constituencies. From this perspective there is reason to believe that even in advanced industrial societies class voting might increase as well as decrease. It also implies that the adoption of class-relevant policy programs should be associated with an increase in the class basis of partisanship. Some evidence for this has been presented by Evans *et al.* (1999) who show a relationship over time between left-right polarization in parties' manifestos and the extent of class voting.

This is not to say that sociological changes have no impact, however. Changes in the relative *sizes* of classes have been thought to have implications for party strategy: most importantly, in a change to a 'catch-all' strategy by parties on the left in response to the shrinking class basis of support for those parties (Przeworski and Sprague, 1986). In some political systems such moves leave open the space for left parties to attract support from marginalized working class groups; in others, such as first-past-the-post systems, we might expect that the start-up costs for electorally viable left parties would be too great. Unfortunately, many of these arguments still await rigorous tests using survey analysis. Research that tries to unravel *why* voters in different classes vote differently is in short supply, though Weakliem and Health (1994) provide evidence of the role of rational choice and inherited preferences, while Evans (1993) models the effects for future expected rewards and the likelihood of promotion conditioned by position in the life-cycle. A more explicit link between models of class voting and their dependence on advances in theories of voting behavior more generally is an area where further development might usefully occur.

Editor's Note

For Professor Evans' and his colleagues' current thinking on class voting and for entry points into more current literature on the topic, see: Giedo Jansen, Geoffrey Evans, and Nan Dirk de Graaf, 2013; Line Rennwald and Geoffrey Evans, 2014.

See also: Class Consciousness; Class: Social; Interest Groups, History of; Interest Groups; Interest: History of the Concept; Political Sociology; Social Inequality in History (Stratification and Classes); Social Stratification; Voting, Sociology of.

Bibliography

Alford, R., 1967. Class voting in Anglo-American political systems. In: Lipset, S.M., Rokkan, S. (Eds.), Party Systems and Voter Alignments: Cross-National Perspectives. Free Press, New York, pp. 67–94.

Anderson, D., Davidson, P.E., 1943. Ballots and the Democratic Class Struggle. Stanford University Press, Stanford, CA.

Erikson, R., Goldthorpe, J.H., 1992. The Constant Flux: A Study of Class Mobility in Industrial Societies. Clarendon, Oxford, UK.

Evans, G., Heath, A., Payne, C., 1999. Class: Labour as a catch all party? In: Evans, G., Norris, P. (Eds.), Critical Elections: British Parties and Voters in Long-Term Perspective. Sage, London, pp. 87–101.

Evans, G., 1993. Class, prospects and the life-cycle: explaining the association between class position and political preferences. Acta Sociologica 36, 263–276.

Evans, G. (Ed.), 1999. The End of Class Politics? Class Voting in Comparative Context. Oxford University Press, New York.

Franklin, M.N., Mackie, T.T., Valen, H., 1992. Electoral Change: Responses to Evolving Social and Attitudinal Structures in Western Countries. Cambridge University Press, Cambridge, UK.

Goldthorpe, J.H., 1996. Class and politics in advanced industrial societies. In: Lee, D.J., Turner, B.S. (Eds.), Conflicts about Class: Debating Inequality in Late Industrialism. Longman, London, pp. 196–202.

Heath, A., Jowell, R., Curtice, J., 1985. How Britain Votes. Pergamon, Oxford, UK.

Heath, A.F., Evans, G., Payne, C., 1995. Modelling the class/party relationship in Britain, 1964–92. Journal of the Royal Statistical Society, Series A 158, 563–574.

Hout, M., Brooks, C., Manza, J., 1995. The democratic class struggle in the United States. American Sociological Review 60, 805–828.

Inglehart, R., 1997. Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies. Princeton University Press, Princeton, NJ.

Jansen, Giedo, Evans, Geoffrey, Graaf, Nan Dirk de, 2013. Class voting and left-right party Positions: A comparative study of fifteen western democracies, 1960-2005. Ch. 3 (pp. 46 ff). In: Evans, Geoffrey, Graaf, Nan Dirk de (Eds.), Political Choice Matters: Explaining the Strength of Class and Religious Cleavages in Cross-National Perspective. Oxford University Press, Oxford.

Kitschelt, H., 1994. The Transformation of European Social Democracy. Cambridge University Press, Cambridge, UK.

Lipset, S.M., 1981. Political Man: The Social Bases of Politics, expanded and updated edition. Heinemann, London.

Manza, J., Hout, M., Brooks, C., 1995. Class voting in capitalist democracies since World War II: dealignment, realignment or trendless fluctuation? Annual Review of Sociology 21, 137–162.

Nieuwbeerta, P., de Graaf, N.D., 1999. Traditional class voting in 20 postwar societies. In: Evans, G. (Ed.), The End of Class Politics? Oxford University Press, New York, pp. 23–57.

Przeworski, A., Sprague, J., 1986. Paper Stones: A History of Electoral Socialism. University of Chicago Press, Chicago.

Rennwald, Line, Evans, Geoffrey, 2014. When Supply Creates Demand: Social Democratic Party Strategies and the Evolution of Class Voting. West European Politics 35 (5), 1108–1135.

Sartori, G., 1969. From the sociology of politics to political sociology. In: Lipset, S.M. (Ed.), Politics and the Social Sciences. Oxford University Press, Oxford, UK, pp. 65–100.

Weakliem, D.L., Heath, A.F., 1994. Rational choice and class voting. Rationality and Society 6, 243–270.

Voting, Sociology of

Nonna Mayer, Centre for European Studies, Sciences Po-National Centre for Scientific Research, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Because democracies choose their governments via the ballot box, the bulk of the literature on voting behavior deals with its decisional component, focusing on who votes and for whom. This article presents a comparative assessment of research on electoral choice and electoral turnout, along with the main theories and empirical findings in the field, from the founding sociological, psychological, and economic postwar models to the recent development in the genetics of politics.

Paradigms Explaining Voting Choice

Electoral geography was the first, and for a long time the only, clue for understanding political behavior, linking voters' choices to the social structures and the religious and political traditions of the places where they lived. Since the earliest studies of André Siegfried or Herbert Tingsten, this contextualized or ecological approach, based on the analysis of aggregate electoral data, has gained in statistical sophistication (for the latest trends see Warf and Leib, 2011). Its limitation lies in the fact that the unit of analysis considered is the group – people living in the same constituency, the same district, the same region, etc. – and that the correlations observed at the aggregate level do not necessarily hold true at the level of the individuals, as Robinson's (1950) famous article on the ecological fallacy reminds us. Opinion polls provide a better tool for understanding individuals' behavior, asking them directly how and why they vote. Since World War II, survey-based research has given a new impetus to electoral sociology, especially in the United States, where the pioneer studies and the classic models explaining vote choice were developed.

The Columbia Model

The sociological model originates from the local surveys conducted by Paul Lazarsfeld and his colleagues at the Bureau for Applied Social Research of the University of Columbia. Repeated interviews with a randomly selected panel of voters, first in the district of Erie County, Ohio, during the 1940 Roosevelt–Wilkie presidential campaign, then in the town of Elmira, New York, at the time of the Truman–Dewey contest, revealed, to their surprise, that the campaign had but a limited effect on voters' choices. Mostly, it reinforced long-standing and diffuse political preferences that reflect the norms of the groups they belong to. Constructing an index of political predisposition based on social class, religion, and rural/urban residence, the authors of *The People's Choice* can predict accurately how people vote. Those combining a rural residence, Protestant affiliation, and high social status tend to support the Republicans, whereas those who are working class, non-Protestant, and residing in urban areas vote mostly for Democrats. The main conclusion of their book is that "A person thinks, politically, as he is, socially. Social characteristics determine political preference" (Lazarsfeld et al., 1944). And this social anchoring accounts for the relative stability of voting choices, except among the cross-pressured or poorly integrated voters. However, in his later empirical and theoretical work such as *Voting* (Berelson et al., 1954), Lazarsfeld was to develop a more complex view of action in general and more particularly of voting behavior. (His research evolution is related in Lautman and Lécuyer, 1998.)

The Michigan Model

Reacting against what they see as a social determinism, Angus Campbell and his colleagues at the Survey Research Center (SRC) at the University of Michigan propose a psychological model focusing on individual decision making. Reconstructing the chain of events that converges in what they call the 'funnel of causality,' they give special attention to the attitudes formed toward candidates, issues, and parties. On the basis of nationwide panel surveys conducted during presidential elections from 1948 to 1956, they state in *The American Voter* (Campbell et al., 1960) that party identification is the key factor orienting vote choice. By that they mean an affective attachment to a specific party, formed early in life, generally passed on from parents to children and reflecting individuals' social and religious backgrounds. It functions as a filter coloring their political perceptions and guiding their votes, especially for the least sophisticated: "Like the automobile buyer who knows nothing of cars except he prefers a given make, the voter who knows simply that he is a Republican or Democrat responds directly to his stable allegiance without the mediating influence of perceptions he has formed on the objects he must choose between" (Campbell et al., 1960). Lazarsfeld portrayed voters as social conformists; the Michigan team sees them more as political believers. Yet their conclusions are very similar: average voters are far from the democratic ideal of rational and active citizens. Most of them are not interested in politics, display low levels of ideological awareness, and are unfamiliar with the candidates and the issues at stake. They vote according to long-standing political preferences shaped by their early socialization.

The Economic Approach

The third approach stems from Downs's seminal essay, *An Economic Theory of Democracy* (1957). It postulates, along the lines of the neoclassic economic paradigm, that voters

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32169-9

are rational and that their choices are based on a cost–benefit calculation. Voters cast their votes for the party or the candidate from which they expect the maximum utility, or benefits. The electoral decision between party A in office and opposition party B boils down to a simple equation: the difference between the benefits expected from the victory of party A (Ua) and the benefits expected from party B (Ub). Voters will favor party A if they see the difference between Ua and Ub as positive, party B if the difference is negative, and will not bother to vote if they see no difference. Since Downs's pioneering work, two distinct research fields have developed. The spatial approach of politics has been renewed by directional models (see Merril and Grofman, 1999). Voters won't necessarily choose candidates closer to themselves in a policy space (proximity voting); they might prefer more distant candidates but who are on the same side of an issue as themselves (directional voting), or they might prefer those who could move policy closest to them (discounted proximity). Econometric models, based on the simple idea that voters punish or reward candidates according to the state of the economy, have been sophisticated. Some correlate aggregate electoral results with global economic indicators, such as purchasing power, inflation rates, and unemployment rates. Others use survey data to analyze individual pocket-book voting based on the evaluation of candidates past (retrospective voting) or future performances (prospective voting). Their common goal is to shift attention from the people's characteristics to the supply side of politics, the specific goods offered on the electoral marketplace by competing parties and candidates, in a given election.

The Decline of Cleavage Voting

These three models have been guiding most electoral research agendas.

Class Dealignments

In the wake of the sociological model, attention was first given to the effects of social class and religion, even more so in Europe, where party systems were built historically on the bases of social and religious conflicts and cleavages (see Lipset and Rokkan, 1967). Rose's (1974) classic study of voting patterns in 15 democracies shows that class in Scandinavian countries and religion in the others are by far the most predictive factors explaining postwar electoral behavior, a working-class position predisposing to a left-wing vote and regular church attendance to a right-wing vote. Then, in the 1960s, the Michigan paradigm became the leading model, inspiring nationwide research surveys in Europe (Thomassen, 1994) especially in Britain, where Butler and Stokes (1969) explained long-term alignments of British voters by their party identification, itself rooted in their social backgrounds. What became known as the two-class, two-party model does not fit as well in countries such as France, with its multiparty system. There, Left/Right orientations are the substitute for party identification but are equally rooted in religion and social class (Michelat and Simon, 1977).

Neither model, though, could account for the increasing electoral change that appeared in the United States at the end of the 1960s and in Europe by the end of the 1970s. On the basis of the SRC presidential surveys between 1956 and 1974, the authors of *The Changing American Voter* (Nie et al., 1976) reveal a decline of party ties. The proportion of independent voters, who define themselves as neither Republican nor Democrat, doubled between 1964 and 1974. Among those who still identify themselves with one side or the other, the proportion who declare a strong attachment or have a positive image of a party is declining, and the very link between party preference and voting is weakening. Voters seem more interested in politics, more aware of the issues at stake, and more inclined to choose their candidates according to their political stands, whatever party they belong to.

Parallel studies in Europe detect similar trends of electoral volatility or instability of voting choice (Pedersen, 1979) and partisan dealignment, or gradual moving away from all parties (Crewe and Denver, 1985), which is seen as a structural feature of the emerging postindustrial society. Social and geographical mobility is blurring the traditional class cleavages and loosening community ties. Progress in education and exposure to the media have increased the average levels of political sophistication. The rise of the permissive and individualistic set of values coined as postmaterialist by sociologist Inglehart (*The Silent Revolution*, 1977) encourages a new style of politics, more demanding and protest prone, and promotes new postmaterialist issues – environmental, feminist – that cut through traditional party lines. All these trends converge to make citizens less dependent upon existing parties to make their choices. Voters are seen as being autonomous, strategic, rational, or reasoning, and the developing cognitive sciences enrich studies of political reasoning mechanisms, the way voters process and organize information, the cues they rely on to make decisions, etc. (Sniderman et al., 1991). Short-term issues or candidate-centered voting could well be taking the place of the formerly stable cleavage voting, which had a partisan, religious, or social basis (Franklin et al., 1992).

Class Realignments

But a second research trend has taken an opposite stand, questioning the extent of this dealignment process by casting doubts on the way it is measured. The most heated controversies have developed around the supposed decline of class voting (for a summary, see Evans, 1999). The simplest and most widely used indicator has been Alford's index (1963), which measures the difference between the proportion of manual and nonmanual voters voting for the Left. For instance, if all the British working-class voters supported the Labor party and none in the middle classes supported it in a given election, the value of the Alford index would be 100%, indicating perfect class voting. Measured by this index, the downward trend is undeniable. Class voting in post–World War II elections has decreased, for instance, by almost half in Britain and by more than two-thirds in Germany. But critics underline several weaknesses of such an index. Its dichotomous nature does not take into account the complexity of postindustrial occupational structure or the complexity of party systems. It is statistically

biased because it does not consider the changes in the marginal distributions of the two classes and of the two votes between one election to another. And it measures only dealignment, whereas various processes of electoral realignment along new cleavages, such as private sector vs public sector, self-employed vs wage-earners, or educated vs noneducated voters, seem to be taking place in postindustrial democracies, as shown by multiclass approaches using detailed occupational classifications and sophisticated measures such as odds ratios or log-linear models. If in the long run there is growing evidence of the decline of class voting, it is not linear and differs among countries. In Britain, for instance, from the 1960s to the early 1990s Heath et al. (1991), contrary to previous findings, found no decline in class voting, just 'fluctuations without trend.' In France since the 1980s, the salaried middle classes have tended to support the Left, compensating for the drop in support by the working-class, which is attracted by the extreme right embodied by the National Front (Gougou and Mayer, 2012). In the United States, Hout, Brooks, and Manza disputed Clark, Lipset, and Rempel's claims about the declining political significance of social class by showing the historical realignment in presidential post-War elections since 1968, as professionals and nonmanagerial white-collar workers moved from supporting the Republicans to voting for the Democrats (special issue of *International Sociology*, 8(3), 1993, on class voting). And the multiplication of large-scale electoral survey-data bases allows for systematic comparisons (Knutsen, 2006; Thomassen, 2005; Brooks et al., 2006). A major realignment taking place today in Europe could also be the growing support of the working class, yesterday a stronghold of the left, for the populist radical rights (Oesch, 2008; Rydgren, 2012).

The Resilience of the Religious Cleavage

The effect of religion on votes is even more difficult to measure than the effect of class, owing to the diversity in religious denominations and beliefs from one country to another. Yet roughly, religious integration makes voters more likely to support conservative parties. Abramson et al. (1999) have constructed the equivalent of an Alford index measuring the difference between the proportion of Protestants and non-Protestants voting for the Democrats in the United States. On the basis of a more sophisticated measure based on log odds ratios, Brooks and Manza (2004) confirmed the permanence of an electoral religious cleavage, opposing Protestants, who vote more commonly for the Republicans to Catholics, Jews, and secular voters, who prefer the Democrats. It has just slightly eroded because of the growing heterogeneity of the Protestant group, Republicans losing the support of the more liberal on issues of abortion and sex who tend to turn to the Democrats, contrary to the more conservative Evangelical Christians. More recent studies confirm the growing religious polarization of the American electorate (Putnam and Campbell, 2012).

In Europe, it is the combination of church involvement and religious denomination that influences the votes. On the whole, Christian voters are more likely to vote for the right, and support for the right increases with their degree of church attendance (Broughton and ten Napel, 2000; Knutsen, 2004; Dargent et al., 2009). In Catholic countries and mixed, bireligious contexts such as the Netherlands, the effect of religion is higher than the class effect, whereas in northern Protestant countries and in England, class comes first.

New Social Divides

Beyond the 'old' variables like religion and social class, other dividing lines have gained importance, and they are linked to ethnicity, gender, or age.

The Evolving Gender Gap

The term gender gap was coined at the time of the election of Ronald Reagan in the 1980 American presidential race, when women, for the first time since they had the right to vote, gave more support to the Democrats. Since then this gap has persisted, and there is a robust relationship between gender and vote, still significant after controlling for age, class, race, and religion (Manza and Brooks, 1998). As shown by Inglehart and Norris in *The Rising Tide* (2003), a similar gender gap has developed in most postindustrial democracies, women moving more to the left than men, whatever the indicator used (Giger, 2009). It has also been called the 'modern gender gap,' in opposition to the previous 'traditional' gender gap. In the first phase, just after being granted the right to vote, women tended to be more conservative and give more support than men to right wing parties with traditional views of women's roles. This inversion of the gender gap was attributed to their rising level of education and their massive entrance in the labor market, on the one hand, and to the process of secularization and the emancipating influence of post-materialist values and feminist movements on the other, encouraging autonomy and self-expression, especially among the younger generations, leading Pippa Norris to qualify it as the 'generation gender gap' (Norris, 1996).

Since the 1980s, though, at least in Europe, the electoral boom of far-right anti-immigrant parties has brought about a new cleavage, women appearing more reluctant than men to give them their votes. Hans-Georg Betz (1994: 142–148) was the first to outline the phenomenon in his seminal study of the radical populist rights in Europe and the difference persists. Socioeconomic and attitudinal factors have been put forward to explain this new 'Radical Right gender Gap' (Givens, 2004). Men would more often than women be 'globalization losers', overrepresented in manual unskilled works and more exposed to economic competition with migrants than are women, who are overrepresented in nonmanual jobs and public service. Women, even if a process of secularization is taking place, are still more religious and more frequent church goers than are men, and the xenophobic and antiegalitarian message of the far right parties has been repeatedly condemned by the Christian churches. There is the impact of feminist ideas, slowly progressing at all levels of society, especially among the younger generations, which would prevent women from supporting parties with traditional ideologies, reducing them

to their roles of spouses and mothers. There is an image of verbal and physical violence attached to these parties that would particularly repel women.

But other studies have lately foreseen a possible decline of the radical right gender gap, as in France, where in the 2012 presidential election for the first time, after controlling for age, education, political orientation, and religion, gender had no impact on support for Marine Le Pen, the new chair of the National Front. Religion could be a less protective shield at a time when Islam is presented as a cultural threat to Western Christian values. The economic situation of women has deteriorated, and there is an unskilled female-service proletariat symmetrical to that of the male, manual working class, where economic insecurity fuels anti-immigrant resentment. Last, parties like the Party for Freedom in the Netherlands or the-Norwegian Progress Party are giving a feminist touch to their crusade against Islam, which is presented as a religion of fanaticism and intolerance. They claim they defend the rights of women beaten, mutilated, and ill treated by Muslim men (Akkerman and Hagelund, 2007). In France, Marine Le Pen, the newly elected president of the National Front, is taking a similar turn. This could in the long run considerably expand the electoral base of the far right.

Race and Ethnicity

The electoral impact of race was first studied in the United States, where an African American identity emerged after the Civil War. Black voters first supported the Republican, Lincoln, and his antislavery party. But the New Deal first, then in the 1960s the struggle for civil rights, brought about a major party realignment along the race issue (Carmine and Stimson, 1989), the blacks supporting the Democratic party, champion of civil liberties, and the white Southerners turning to the Republicans. Ever since, African American voters have massively voted for the Democrats, more than 95% supporting Obama in the 2008 presidential election. As shown by Michael Dawson (*Behind the Mule*, 1994), the memory of collective struggles and the feeling of a shared community of destiny unifies the group, and race counts more than class. One finds a similar pattern of ethnic voting among other minorities: in 2008, according to a New York Times/CBS exit poll, 62% of Asians and 67% of Hispanics voted for Obama, compared with 43% of white voters.

In Europe, the field of research into ethnic voting is quickly expanding too (Givens and Maxwell, 2012), even in France, where for a long time, the color-blind republican model was an obstacle to public assertion of ethnic identities. The pioneer survey conducted by Sylvain Brouard and Vincent Tiberj (*As French as Anyone Else?*, 2011) in a sample of French citizens of Maghrebin, African, and Turkish origin, 60% of whom are Muslims, shows massive electoral support for the parties of the Left, controlling for occupation, income, age, gender, and education.

The last Ethnic Minority British Election Study following the 2010 election tells the same story: Black and minority ethnic people, who represent some 8% of the electorate, remain highly supportive of the Left and its egalitarian and universalistic values, with more than two-thirds voting Labor and up to 80% of African and Caribbean people voting Labor.

The Generational Gap

Age cuts out distinct generations whose members are born at the same time, go through the same events, start voting at the same moment and, in certain contexts, can develop a form of group identity. Taking the French case as an example, Louis Chauvel (2006) analyzes the unequal life chances of three distinct generations: the pre-War generation, who went through the world conflict and the depression; the privileged baby boomers, who benefited from a period of peace and prosperity, better educational opportunities, full employment, rising social benefits, and upward mobility; and the post–baby boom generation, who became adults in a context of recession, when the welfare state was in crisis, unemployment was rocketing, income was stagnating, and the risks of downward mobility were increasing. The political consequences are that post–baby boomers stand apart by their low level of conventional involvement in politics (electoral turnout, party proximity, political interest) contrasting with their high potential for protest and contentious politics. In the long run, in an aging society, this could lay the foundations for generational conflicts. Of course, youth does not form a unified group; ethnicity and access to education and jobs draw strong lines of cleavage among them. But the context of global recession, which has hit young people primarily since 2007, could fuel their resentment and amplify the divide.

These lines of cleavage make sense only if one articulates them; the same person has multiple identities, and one of them will be most salient in the specific context of the election. The concept of intersectionality, coined by feminists, is one way to think about these different forms of domination together, linked to race, gender, and class. Another approach is the total social cleavage index used by Manza and Brooks (1997) to capture the global effects of class, religion, race, and gender on presidential votes in the United States, controlling for age, education, and region. It shows, paradoxically, that from 1962 to 1992, the level of the global index remained stable. There is always a social logic to politics (Zuckerman, 2005). What changes is the respective weight of the variables; there was a spectacular rise in racial cleavage after the 1960s, followed by gender.

Cultural Cleavages

Another expanding field of research, started by Inglehart's seminal work on postmaterialism and Herbert Kitschelt's work on the axes of party competition in Western Europe (Kitschelt and McGann, 1995), explores the growing electoral impact of the second dimension of politics. Yesterday, left and right opposed each other on economic issues and on the degree of state intervention and regulation. With the rise of post-materialist values carried by New Left and Green parties, and the backlash against these same values carried by New Radical Rights, the political debate has shifted toward noneconomic issues, such as abortion, sexual minorities, and Christian values in the United States and immigration, multiculturalism, and national identity in Europe. Actually, "Class is not dead – it has been buried alive under the increasing weight of cultural voting," as shown by Van der Waal et al. (2007). Workers who support the far right are in favor of a strong social state, like

the left-wingers, but when the time to vote comes, the main issue in their eyes is immigration, which is seen as an economic, cultural, and political threat. And that is why class is less significant than education when explaining these votes, tightly correlated as they are with the degree of authoritarianism vs libertarianism (Bornschier, 2010). A similar debate is taking place in the United States. Larry Bartels (2008) finds no rise in cultural voting in the working class; on the contrary, he observes a growing gap between high- and low-income people in terms of the level of support for the Democrats.

The Role of Emotions

A last challenge for both the rational choice and the social psychological approaches of voting comes from the neurosciences and a growing field of research revisiting the role of emotions. The pioneer experimental work of political psychologists such as George Marcus in *Affective Intelligence* (Marcus et al., 2000), exploring the emotional foundations of political cognition, showed that anxiety, paradoxically, makes voters more rational. If 'normal' votes rely on habit, going with the preferred party, in a context of crisis or tension, voters tend to reassess their positions, look at candidates and platforms with critical eyes, and eventually vote differently from their party identification. Yet authors such as Drew Westen in *The Political Brain* (2007) and Jonathan Haidt in *The Righteous Mind* (2012) reach the opposite conclusion. Passions, feelings, deeply ingrained values, and moral intuitions count more than rational argument and deliberation when the time comes to take sides. The debate is open.

The Puzzle of Electoral Participation

An analysis of voting choice would not be complete without taking into account the alternative behavior: not voting. Whatever the election, a sizable number of eligible citizens are not registered or choose not to go to the polls. The proportion ranges between less than 5% in countries where voting is compulsory and 40% or more in Switzerland and in the United States.

The Socioeconomic Status Model and Beyond

Whatever the country, the profiles of these nonvoters, revealed by aggregate as well as survey data, are quite similar. Educated voters and members of the upper and middle classes participate more than the uneducated and the working classes, and middle-aged voters go more willingly to the polls than the very young or the very old. What has become known as the socioeconomic status (SES) model (Verba and Nie, 1972) postulates that well-integrated and upper-status individuals are more likely to participate because they have more resources to invest in politics – time, money, information, and networks – and because they are more likely to develop civic attitudes (see Almond and Verba, *The Civic Culture: Political Attitudes and Democracy in Five Nations*, Princeton, NJ: Princeton University Press, 1963) that predispose to do so, such as interest in politics, a sense of duty, and a feeling of political efficacy. Political participation is socially skewed. There are compensating factors, however, such as group consciousness or ethnic awareness. With the development of feminism, the difference between female and male turnout rates has disappeared, and black political empowerment in the United States since the Voting Right Acts has reduced the difference between whites and African Americans, the latter participating even more than whites when differences in SES are taken into account (Verba and Nie, 1972).

Last, as pointed out by Verba et al. (1995), if people do not participate it is not only because they cannot, they lack the necessary skills or money, or they do not want to, because they are not interested; it is also because 'nobody asked them to.' They are isolated from the networks – churches, associations, unions, and formal and informal groups – that mobilize people. They lack social capital, in the sense given to it by Robert Putnam (2000), as a vector of civicness. Conversely, Braconnier and Dormagen (2007) explain the dramatic drop in voting turnout in the former 'red' suburbs of Paris by the disappearance of the Communist Party networks that used to structure and mobilize the neighborhood.

The Puzzle of Participation

On the whole, the increasing education level of the electorate and the growth of the middle classes should gradually raise electoral participation in postindustrial societies. But paradoxically, the opposite is happening. There has been a clear drop in turnout rates in most established democracies (Franklin, 2004).

This participation puzzle has revived interest in rational choice models. Because one vote has but an infinitesimal influence on the electoral outcome, the rational behavior, at first sight, would be to not bother going to the polls, and since Downs, a large body of literature has been devoted to what appeared as the voting paradox. It can be understood only if one extends the notion of rationality and if one admits that voting, like most human actions, is guided not only by interest (instrumental rationality) but also by principles and values (axiological rationality) (Boudon, 1997). At the beginning of the twenty-first century, along the same line, several studies have tried to solve the paradox of nonvoting (for a summary, see Blais, 2000). Some studies focus on the costs and benefits of voting turnout and particularly institutional factors, such as registration formalities, compulsory voting, electoral rules, party competition, and all the political factors that facilitate or make meaningless the act of voting (Norris, 1998). Others, in line with Harold Gosnell's pioneer experiments in Chicago in the 1923 elections (*Getting Out the Votes*: *An Experiment in the Stimulation of Voting*, San Francisco, Greenwood Press, 1977) explore what can best induce citizens to perform their civic duty, intrinsic satisfaction derived from behaving in accordance with the norm, or extrinsic incentives to comply. A large-scale field experiment involving several hundred thousand registered voters used a series of mailings to test these two hypotheses. Turnout was much higher among those who received mailings promising to publicize the turnout to their households or their neighbors than among those who were simply reminded of the civic importance of voting, which demonstrates the decisive importance of social pressure (Gerber et al., 2008).

Another line of research sees nonvoting as a deliberate strategic behavior, expressing political discontent with the electoral choices available, even though those who do not vote have levels of education and political interest that are above average, and they are politically 'in the game' (Jaffré and Muxel, 2000). Studies of electoral turnout detecting nonvoters on official voting records, such as the series conducted in France by the National Institute for Statistics and Economics Studies, more reliable than survey data because voters are reluctant to confess what is considered an uncivic behavior, confirm the rise of intermittent voting, driven by the political context. Along the same lines, the large-scale comparative survey conducted by Mark Franklin (2004) in established democracies shows clearly that the primary factor explaining declining turnout is the lack of political competitiveness. People vote when they think their voices count.

The Challenge of Genopolitics

The latest challenge in the field of electoral sociology comes from a new research subfield, genopolitics, at the crossroad of political science and behavior genetics. An article by Alford et al. (2005) opened the way, claiming that genes count more than environment in explaining ideological orientations, specifically how one becomes a conservative or a liberal. A host of studies has followed, usually based on the methodology of twin studies, comparing the political behaviors and attitudes of genetically identical twins with the same DNA, to nonidentical twins. If the identical pairs show the same political behavior, it would mean that this behavior is heritable, and genes explain its variations. For instance, in a recent article in the *Journal of Politics*, Fowler and Dawes (2008) claim that they have found two genes predicting voting turnout. The study is based on the third wave of the National Longitudinal Study of Adolescent Health (2001–02), from which they extracted a sample of sibling pairs who answered questions, among others, about civic participation, partisanship, and religion and provided genetic material. The results are not totally convincing as yet. A critical review of this study written by Charney and English for the *American Political Science Review* (2012) points to several weaknesses, questioning their indicators of voting turnout, of population stratification, of genotype classification, of independence of cases and controls, and of the ignorance of the recent paradigm shift in the field of genetics.

The three models we started with have changed considerably. Born in the behavioral era, they were reformulated by the cognitive neurosciences, for which affects, group belonging, party identification, or the state of the economy are cues, among others, that help voters to orient themselves in the political space. They might benefit from the new challenge coming from the advance of behavior genetics. Meanwhile, the rising number of emerging or transitional democracies since the mid 1970s, as well as the multiplication of cross-national data bases – Eurobarometers, Central and Eastern Eurobarometers, Latino Barometro, Latin American Public Opinion Project, Afrobarometer, Asian Barometer, Arab Democray Barometer, etc. – provide fresh opportunities to enlarge the scope of electoral studies and test in various political cultures the validity of theories and concepts elaborated in Western democracies. It could also be an occasion to explore what most studies leave aside: the noninstrumental meanings the act of voting can take, from one context and one epoch to another. Although the political outcomes of elections are essential, they also have ritual, expressive, and social functions, as shown by Yves Deloye and Olivier Ihl in their major book, *The Act of Voting* (2008), and its evolution over two centuries.

See also: Citizen Participation; Citizenship, Sociological Aspects of; Class Consciousness; Class: Social; Electoral Geography; Ethnicity, Sociology of; Groups, Sociology of; Minorities; Politics and Participation; Race Identity; Racism, Sociology of; Religion, Sociology of; Values, Sociology of.

Bibliography

Abramson, P., Aldrich, J., Rhode, D., 1999. Change and Continuity in the 1996 and 1998 Elections. Congressional Quarterly Press, Washington, DC.
Akkerman, T., Hagelund, A., 2007. 'Women and children first'. Anti-immigration parties and gender in Norway and the Netherlands. Patterns of Prejudice 41 (2), 197–214.
Alford, R., 1963. Party and Society. Rand McNally & Company, Chicago, IL.
Alford, J.R., Funk, C.L., Hibbing, J.R., 2005. Are political orientations genetically transmitted? American Political Science Review 99, 153–167.
Bartels, L., 2008. Unequal Democracy. Princeton University Press, Princeton, NJ.
Berelson, B.R., Lazarsfeld, P.F., Mac Phee, W.N., 1954. Voting: A Study of Opinion Formation in a Presidential Campaign. Chicago University Press, Chicago, IL.
Betz, H.-G., 1994. Radical Right Wing Populism in Western Europe. Saint Martin's Press, New York.
Blais, A., 2000. To Vote or Not to Vote? The Merits and Limits of Rational Choice Theory. University of Pittsburgh Press, Pittsburgh, PA.
Boudon, R., 1997. Le 'paradoxe du vote' et la théorie de la rationalité. Revue Française de Sociologie 38 (2), 217–227.
Bornschier, S., 2010. Cleavage Politics and the Populist Right: The New Cultural Conflict in Western Europe. Temple University Press, Philadelphia, PA.
Braconnier, C., Dormagen, J.-Y., 2007. La démocratie de l'abstention. Gallimard, Paris.
Brooks, C., Nieuwbeerta, P., Manza, J., 2006. Cleavage-based voting behavior in cross-national perspective: evidence from six postwar democracies. Social Science Research 35 (1), 88–128.
Brooks, C., Manza, J., 2004. A great divide? Religion and political change in US national elections, 1972–2000. The Sociological Quarterly 45, 421–450.
Brouard, S., Tiberj, V., 2011. As French as Anyone Else? Temple University Press, Philadelphia, PA.
Broughton, D., ten Napel, H.M., 2000. Religion and Mass Electoral Behaviour in Europe. Routledge, London.
Butler, D., Stokes, D., 1969. Political Change in Britain: Forces Shaping Electoral Choice. St. Martin's, New York.
Campbell, A., Converse, P., Miller, W., Stokes, D., 1960. The American Voter. Wiley, New York.
Carmine, E., Stimson, J., 1989. Issue Evolution: Race and the Transformation of American Politics. Princeton University Press, Princeton, NJ.
Charney, E., English, W., 2012. Candidate genes and political behavior. American Political Science Review 106 (1), 1–34.
Chauvel, L., 2006. Social generations, life chances and welfare regime sustainability. In: Pepper, D., Culpepper, P., Hall, P., Palier, B. (Eds.), Changing France: The Politics that Markets Make. Palgrave McMillan, Basingstoke.
Crewe, I., Denver, D.T. (Eds.), 1985. Electoral Change in Western Democracies. St. Martin's, New York.
Dargent, C., Duriez, B., Liogier, R. (Eds.), 2009. Religion et valeurs en France et en Europe. l'Harmattan, Paris.
Dawson, M., 1994. Behind the Mule. Race and Class in African American Politics. Princeton University Press, Princeton, NJ.

Deloye, Y., Ihl, O., 2008. L'acte de vote. Presses de Sciences Po, Paris.
Downs, A., 1957. An Economic Theory of Democracy. Harper & Row, New York.
Evans, G. (Ed.), 1999. The End of Class Politics? Class Voting in a Comparative Context. Oxford University Press, Oxford.
Fowler, J.H., Dawes, C.T., 2008. Two genes predict voting turn out. Journal of Politics 70 (3), 579–594.
Franklin, M.N., 2004. Voter Turnout and the Dynamics of Electoral Competition in Established Democracies Since 1945. Cambridge University Press, New York.
Franklin, M.N., Mackie, T.T., Valen, H., et al., 1992. Electoral Change: Responses to Evolving Social and Attitudinal Structures in Western Countries. Cambridge University Press, Cambridge, MA.
Gerber, A., Green, D., Larimer, C., 2008. Social pressure and voter turnout: evidence from a large-scale field experiment. American Political Science Review 102, 33–48.
Giger, N., 2009. Towards a modern gender gap in Europe? A comparative analysis of voting behaviour in 12 countries. Social Science Journal 46, 474–492.
Givens, T.E., 2004. The radical right gender gap. Comparative Political Studies 37 (1), 30–54.
Givens, T.E., Maxwell, R. (Eds.), 2012. Immigrant Politics: Race and Representation in Western Europe. Lynne Rienner Publishers, Boulder, CO.
Gougou, F., Mayer, N., 2012. The class basis of extreme-right voting in France: generational replacement and the rise of new cultural issues (1984–2007). In: Rydgren, J. (Ed.), Class Politics and the Radical Right. Routledge, London.
Haidt, J, 2012. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Pantheon, New York.
Heath, A., Curtice, J., Jowell, A., et al., 1991. Understanding Political Change: The British Voter 1964–1987. Pergamon Press, Oxford, UK.
Inglehart, R., Norris, P., 2003. Rising Tide. Gender Equality and Cultural Change Around the World. Cambridge University Press, Cambridge.
Jaffré, J., Muxel, A., 2000. S'abstenir: hors du jeu ou 'dans le jeu' politique? In: Bréchon, P., Laurent, A., Perrineau, P. (Eds.), Les cultures politiques des Français. Presses de Sciences Po, Paris, pp. 19–50.
Kitschelt, H., Mc Gann, A., 1995. The Radical Right in Western Europe: A Comparative Analysis. University of Michigan Press, Ann Arbor.
Knutsen, O., 2004. Religious denomination and party choice in Western Europe: a comparative longitudinal study from eight countries, 1970–97. International Political Science Review 25 (1), 97–128.
Knutsen, O., 2006. Class Voting in Western Europe: A Comparative Longitudinal Study. Rowman and Littlefield, Lanham, MD.
Lautman, J., Lécuyer, B.P. (Eds.), 1998. Paul Lazarsfeld (1901–1976). La Sociologie de Vienne à New York. L'Harmattan, Paris.
Lazarsfeld, P.F., Berelson, B.R., Gaudet, H., 1944. The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign. Duell, Sloan, and Pearce, New York.
Lipset, S.M., Rokkan, S. (Eds.), 1967. Party Systems and Voter Alignments. Free Press, New York.
Manza, J., Brooks, C., 1997. Social cleavages and political alignments: U.S. Presidential Elections, 1960 to 1992. American Sociological Review 62 (6), 937–946.
Manza, J., Brooks, C., 1998. The gender gap in U.S. presidential elections: When? Why? Implications? American Journal of Sociology 103 (5), 1235–1266.
Marcus, G., Norman, W.R., MacKuen, M., 2000. Affective Intelligence and Political Judgment. University of Chicago Press, Chicago, IL.
Merril, S., Grofman, B., 1999. A Unified Theory of Voting. Directional and Proximity Spatial Models. Cambridge University Press, Cambridge.
Michelat, G., Simon, M., 1977. Classe, Religion et Comportement Politique. Presses de la FNSP/Editions Sociales, Paris.
Norris, P., 1996. Mobilizing the women's vote: the gender–generation gap in voting behaviour. Parliamentary Affairs 49 (2), 333–342.
Norris, P. (Ed.), 1998. Elections and Voting Behavior: New Challenges, New Perspectives. Aldershot, Brookfield, VT.
Nie, N.H., Verba, S., Petrocik, J.R., 1976. The Changing American Voter. Harvard University Press, Cambridge, MA.
Oesch, D., 2008. Explaining workers' support for right-wing populist parties in Western Europe: evidence from Austria, Belgium, France, Norway, and Switzerland. International Political Science Review 29 (3), 349–373.
Pedersen, M., 1979. The dynamics of European party systems: changing patterns of electoral volatility. European Journal of Political Research 7 (1), 1–26.
Putnam, R., 2000. Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York.
Putnam, R., Campbell, P., 2012. The American Grace: How Religion Divides and Unites Us. Simon & Schuster, New York.
Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15, 351–357.
Rose, R., 1974. Electoral Behavior: A Comparative Handbook. Free Press, New York.
Rydgren, J. (Ed.), 2012. Class Politics and the Radical Right. Routledge, London.
Sniderman, P.M., Brody, R.A., Tetlock, P.E., 1991. Reasoning and Choice. Explorations in Political Psychology. Cambridge University Press, Cambridge, UK.
Thomassen, J., 1994. The intellectual history of election studies. European Journal of Political Research 25 (Special issue), 239–245.
Thomassen, J., 2005. The European Voter. A Comparative Study of Modern Democracies. Oxford University Press, Oxford.
Van der Waal, J., Houtman, D., Achterberg, P.H.J., 2007. Class is not dead! It has been buried alive. Politics and Society 35, 403–426.
Verba, S., Schlozman, K.L., Brady, H.E., 1995. Voice and Equality: Civic Voluntarism in American Politics. Harvard University Press, Cambridge.
Verba, S., Nie, N.H., 1972. Participation in America: Political Democracy and Social Equality. Harper & Row, New York.
Warf, B., Leib, J. (Eds.), 2011. Revitalizing Electoral Geography. Ashgate, Farnham, UK.
Westen, D., 2007. The Political Brain. The Role of Emotion in Deciding the Fate of the Nation. Public Affairs, New York.
Zuckerman, A.S. (Ed.), 2005. The Social Logic of Politics. Personal Networks as Contexts for Political Behaviour. Temple University Press, Philadelphia.

Relevant Websites

http://www.electionstudies.org/ – American National Election Studies (ANES).
http://www.cessda.org/ – Council of European Social Science Data Archives (CESSDA).
http://cdsp.sciences-po.fr/ – Centre de données socio politiques (CDSP) de Sciences Po.
http://www.umich.edu/~cses/ – Comparative Study of Electoral Systems (CSES).
http://ess.nsd.uib.no/ – European Social Survey (ESS).
http://ec.europa.eu/public_opinion/index_fr.htm – Eurobaromètres.
http://www.europeanvaluesstudy.eu/ – European Values Study (EVS).
http://www.icpsr.umich.edu/icpsrweb/ICPSR/ – The Interuniversity Consortium for Political and Social Research (ICPSR).
http://www.sscnet.ucla.edu/issr/da/Mobility/framec.htm – International Social Mobility and Politics Data (ICS).
http://www.idea.int/ – International Institute for Democracy and Electoral Assistance (IDEA).
http://www.issp.org/ – International Social Survey Program (ISSP).
http://www.worldvaluessurvey.org/ – World Values Survey (WVS).

Voting: Turnout

Benjamin Highton, University of California, Davis, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Understanding why people vote and why some countries have higher turnout than others requires a consideration of the expressive nature of voting and the costs and benefits it entails. This makes it possible to explain empirical regularities in turnout patterns. Contrary to conventional wisdom, it does not appear that universal turnout would alter election outcomes very much.

Why do some people vote and others do not? Why do some countries have higher turnout rates than others? Would election outcomes change if turnout levels were higher? These are the primary questions that concern students of voter turnout.

Factors That Affect Turnout

With one caveat, a useful approach to explaining turnout rates of either individuals or countries entails considering the costs and benefits of voting. The caveat reflects the observation that any single vote "is lost in a sea of other votes" (Downs, 1957: 246). Paul Meehl clearly identifies the implication of 'this big brute fact,' namely that "my chances of determining who becomes President [or influencing just about any election outcome] are of about the same order of magnitude as my chances of being killed driving to the polls—hardly a profitable venture …. [Voting] is irrational as an instrumental action toward an egocentric end" (Meehl, 1977: 11). To make any headway in explaining voter turnout, one must recognize that there are intrinsic or expressive benefits to be gained from voting that do not depend on whether one's vote is decisive for an election outcome. Without this expanded conception, there can be no coherent explanation for why so many people choose to cast ballots.

The significant factors that affect the costs and benefits of voting may be usefully characterized as either of the institutional or individual variety. There are two primary types of institutional variables. *Laws* that specify the registration process, the type of election districts, and the structure of the national legislature, all influence turnout. *Political characteristics* like the degree of disproportionality in the translation of votes into seats and the number of political parties represent another set of institutional characteristics that affect turnout.

With regard to individual characteristics, although any examination of the vast turnout literature will reveal a plethora of factors, three stand out. Although the size of the difference varies, in virtually every election context the better educated have higher turnout than those who are less educated; older people have higher turnout than the young, although turnout typically dips among the oldest eligible voters; and those who are more interested in politics have higher turnout than those whose interests lie elsewhere. The explanation for these patterns may be found in a consideration of the costs and benefits potential voters face.

The Costs of Voting

Focusing first on the costs that people must overcome in order to vote, empirical research shows that the higher the costs, the lower the turnout. Two primary costs should be considered. First, some countries have 'compulsory voting,' imposing penalties like modest fines for nonvoting. Turnout in these countries is notably higher. The second significant cost of voting is deciding for whom or what to vote. Commonly referred to as 'information costs,' people must gather or be given sufficient information to decide how they will vote. It is a cost that can be easily overlooked, because for "people already well informed, keeping up with politics is scarcely dull, is almost effortless, and the notion that information costs something seems almost peculiar" (Converse, 1975: 97). Yet for many individuals, politics is at most a peripheral concern. For these less interested people, information costs are higher and their turnout is consequently lower.

Information costs for potential voters are also reduced if they are undertaken by someone or something else. In this regard, candidates and political parties are extremely useful. Voter mobilization efforts reduce information costs because they diminish the need for people to gather for themselves the necessary information in order to make their decisions. Consequently, factors that provide candidates and parties incentives for broad based mobilization efforts, like nationally competitive electoral districts, close elections, and lower levels of disproportionality in the translation of votes to seats, all have the effect of boosting turnout by bringing down information costs for potential voters.

The Benefits of Voting

Even in a political context where voting costs are minimized, they are never eliminated; "voting is inherently costly" (Downs, 1957: 265). It is therefore necessary to consider the benefits citizens derive from voting in order to explain why anyone casts a ballot at all. The 'big brute fact' that a single vote has virtually no effect on an election outcome leads to the observation that "most of the action is, in fact, in the intrinsic values of voting per se" (Aldrich, 1993: 266). The notion that one has a duty to vote pervades most democracies. Schooling is one way this norm in inculcated and provides a reason why the better educated have higher turnout. In addition, those who are more

interested in politics gain greater intrinsic satisfaction from voting because they are more engaged with the political system.

One of the most significant new directions in the study of the voting benefits and turnout has been the use of field experiments to analyze voter mobilization – or get out the vote (GOTV) – strategies (Green and Gerber, 2008). This research began with the question of whether in-person voter canvassing is more effective at mobilizing voters than phone calls or direct mail and has proceeded to examine a wide range of tactics and methods that might motivate people to vote. Two of the most important areas of research are partisan versus nonpartisan GOTV appeals and the influence of various forms of social pressure on turnout. Importantly, because this research is based on field experiments, employing random assignment of treatment conditions, the causal claims that follow from are more persuasive than other research on individual motivations based largely on observational data.

The association between institutional characteristics and the benefits of voting is less intuitive, but is empirically evident. Voting is more gratifying when there is a stronger connection between election results and government formation. The relationship is weaker in countries with more political parties where postelection coalition building is far more frequent and complicated, and turnout is lower. Similarly, in countries with unicameral national legislatures, the consequences of election outcomes are more decisive for governing, and people are therefore more likely to perceive elections as meaningful and vote (Jackman, 1987).

The Implications of Nonvoting

In any country without universal turnout, some groups will have higher turnout than others. Compared to their proportion of the population, some will be overrepresented, while others will be underrepresented in the voting population. Because even countries with compulsory turnout fall short of universal turnout, the effects of nonvoting for election outcomes and governmental policy must be considered. The potential political implications are sufficiently significant that they were the topic of Arendt Lijphart's Presidential Address to the American Political Science Association in 1996, "Unequal Participation: Democracy's Unresolved Dilemma."

Two factors determine the consequences of nonvoting. First, the difference between the preferences of voters and nonvoters must be considered. To the extent that what nonvoters want differs from the desires of voters, then the representation of interests in the voting population will be a distortion of the interests of the larger population. If voters and nonvoters have similar preferences, then the degree of distortion will be less. The second important consideration concerns the relative size of the nonvoting population. If the number of nonvoters in an election is small compared to the number of voters, then even if nonvoters have a very different set of preferences, their small number will limit the potential for impact on the aggregate preferences of voters. On the other hand, where nonvoters are more numerous, their potential influence grows.

Several examples illustrate the interplay between nonvoters' preferences, the size of the nonvoting population, and implications of nonvoting. In the United States, after the Voting Rights Act of 1965 and the widespread dismantling of legal and extralegal barriers to black suffrage in the southern states, large numbers of African-Americans entered the voting population. Their preferences were quite different from those of white voters, thereby spurring the political realignment in that region of the country. In contrast, when the United States (in 1919) and Switzerland (in 1971) extended the right to vote to women, similarly large-scale political change did not ensue. Although the enfranchisement of women added larger numbers of voters, their preferences were not especially distinctive, and as a result there were minimal electoral consequences.

In the current political era, the potential implications of nonvoting are greatest in the United States because turnout is comparatively low and the connection between socioeconomic status and turnout is relatively strong. The actual consequences, however, do not appear to be great. Examining presidential and senatorial elections in the United States, Sides et al. (2011: p. 294) report that "universal turnout would tilt the electorate only very slightly in a Democratic direction rather than generating an enduring liberal majority." More generally, analyzing party vote shares across 25 democracies. Bernhagen and Marsh (2007: p. 558) report results "consistent with the bulk of research implying rather mild effects of increased turnout."

To conclude, voter turnout is shaped by both individual and institutional factors. Explaining the connection between these factors and the decision to cast a ballot requires a consideration of the costs and benefits facing potential voters. Because the costs and benefits are not identical for all people, different groups vote at different rates. As a result, there is a possibility that the preferences of voters will not be representative of the entire population. To be sure, this sort of distortion is sometimes evident, but its magnitude is typically modest.

See also: Citizen Participation; Electoral Systems; Political Representation; Politics and Participation; Voting, Sociology of.

Bibliography

Aldrich, John H., February 1993. Rational choice and turnout. American Journal of Political Science 37, 246–278.

Bernhagen, Patrick, Marsh, Michael, 2007. The partisan effects of low turnout: analyzing vote abstention as a missing data problem. Electoral Studies 26, 548–560.

Converse, Philip E., 1975. Public opinion and voting behavior. In: Greenstein, Fred I., Polsby, Nelson W. (Eds.), Handbook of Political Science. Addison-Wesley, Reading.

Crewe, Ivor, 1981. Electoral participation. In: Butler, David, Penniman, Howard R., Ranney, Austin (Eds.), Democracy at the Polls. American Enterprise Institute, Washington, DC.

Downs, Anthony, 1957. An Economic Theory of Democracy. HarperCollins, New York.

Franklin, Mark N., 1996. Electoral participation. In: LeDuc, Lawrence, Niemi, Richard G., Norris, Pippa (Eds.), Comparing Democracies: Elections and Voting in Global Perspective. Sage, Thousand Oaks.

Green, Donald P., Gerber, Alan S., 2008. Get Out the Vote: How to Increase Voter Turnout, second ed. Brookings Institution Press, Washington, DC.

Geys, Benny, 2006. Explaining voter turnout: a review of aggregate-level research. Electoral Studies 25, 637–663.

Jackman, Robert W., June 1987. Political institutions and voter turnout in industrial democracies. American Political Science Review 81, 405–423.

Lijphart, Arend, March 1997. Unequal participation: Democracy's unresolved dilemma. Presidential Address, American Political Science Association, 1996 American Political Science Review 91, 1–14.

Meehl, Paul E., March 1977. The selfish voter paradox and the thrown-away vote argument. American Political Science Review 71, 11–30.

Sides, John, Schickler, Eric, Citrin, Jack, 2011. Who governs if everyone votes? In: Sniderman, P.M., Highton, B. (Eds.), Facing the Challenge of Democracy: Explorations in the Analysis of Public Opinion and Political Participation. Princeton University Press, Princeton, NJ.

Smets, Kaat, Ham, Carolien van, 2013. The embarrassment of riches? A meta-analysis of individual-level research on voter turnout. Electoral Studies 32, 344–359.

Vygotsky, Lev Semenovic (1896–1934)

René van der Veer, Center for Child and Family Studies, Leiden University, Leiden, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

After a brief account of Vygotsky's life, this article discusses four of Vygotsky's fundamental ideas: that (1) higher mental functions are originally shared between persons; (2) inner speech derives from communicative speech; (3) word meanings are not stable in ontogeny and can be influenced by education; and (4) individual task performance can be predicted by characteristics of previous joint task performance. It is shown that these four and interconnected ideas still inspire modern research.

Introduction

In this article the life and work of Lev Semenovich Vygotsky (1896–1934) is discussed. After a brief overview of his life, attention is paid to four of his fundamental ideas: (1) the sociogenetic law, that is, the idea that all higher mental functions are originally shared between persons and only subsequently become private; (2) the idea that the origin and properties of egocentric and inner speech can be derived by studying the development of children's communicative speech; (3) the idea that word meanings undergo a process of development and that children's everyday concepts can be enriched by teaching them scientific concepts in school; and (4) the zone of proximal development, that is, the idea that joint performance is predictive of independent performance, whether in the context of intelligence assessment or in the context of educational settings. It is concluded that Vygotsky's legacy still inspires much of contemporary research in the humanities.

Biography

Lev Semenovich Vygotsky was born on 5 November 1896 in Orsha as the second child to a secular Jewish family and spent his youth in Gomel in Byelorussia. He received private tuition and only entered school at the age of 15 years to follow the last two years of a private Jewish gymnasium. Having finished school, he moved to Moscow to study law at the Moscow Imperial University and humanities at the Shanyavsky University. He finished his studies at Shanyavsky University with a master's thesis on Shakespeare's *Hamlet* and returned to Gomel to engage in a wide variety of activities (Vygodskaya and Lifanova, 1996).

His interests at the time were divided between two areas. First, since his childhood he had been fascinated by literature and theater. During his university years he had pursued these interests, taking courses in literature and linguistics, attending the plays in Stanislavsky's Art Theater and Tairov's Chamber Theater, and writing theater and book reviews for literary journals. Second, taking Pavel P. Blonsky's courses on education and psychology at Shanyavsky University had quickened his interest in problems of a psychological nature (Feigenberg, 1996). His activities in Gomel reflected these two broad areas of interest. During these years (from 1917 to 1924), Vygotsky worked as a literary critic for local newspapers, taught different subjects (e.g., literature, philosophy, and psychology) at various schools of different levels, cofounded a short-lived publishing house and an equally short-lived literary journal, and worked for the Gomel authorities as the official responsible for the cultural life of the town. In this latter capacity he contracted theater groups, held lectures on topics of general interest, and organized poetry evenings (Van der Veer and Valsiner, 1991).

While in Gomel, Vygotsky became interested in the way the formal properties of a literary text evoke aesthetic experiences in the reader. This led him to deepen his interests in psychology, as did his experiences as a teacher of children and adults of highly different abilities and penchants. He installed a simple psychological laboratory at the normal school where he was teaching and conducted his first empirical psychological investigations. His account of these investigations at a congress led to an invitation by Luria to come and work as a psychologist at Moscow University. His Ph.D. thesis *The Psychology of Art* (Vygotsky, 1925/1971) gives a perfect overview of his early views.

Having arrived in Moscow in 1924, Vygotsky again displayed a wide variety of activities. During what was to be the past decade of his life he worked as a clinical psychologist, carried out empirical studies with children, held teaching positions at various universities, cofounded several scientific journals, headed research institutes, worked for publishing houses, and published a couple of hundred books and articles on widely diverging topics. His interests transcended narrow disciplinary boundaries. He began, for example, studying the semiotics of cinematic language with the film director Sergej Eisenstein, he developed a test to study the thinking of schizophrenics, he lectured on the problems of blind and deaf children, he initiated research into memory development, and he developed an interest in what is now called neuropsychology. In 1931, while already a full professor of psychology, he began studying medicine. However, on 11 June 1934 Vygotsky died in Moscow of tuberculosis, a disease that had haunted him for the better part of his adult life.

The writings of Vygotsky were virtually unknown outside Russia until the publication of *Thought and Language* (Vygotsky, 1962) in English stirred international interest. Nowadays a substantial part of his writings are available in English translation (Van der Veer and Yasnitsky, 2011) and the international discussion of the origins, merits, and shortcomings of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61134-0

Vygotsky's ideas has got a new impetus (Miller, 2011; Yasnitsky, 2011; Zavershneva, 2010).

Main Ideas

Given his interest in literature and language, it comes as no surprise that Vygotsky (1962, 1994) attached great importance to language in his psychological thinking. Like other thinkers, he held that the use of language (speech, in his terminology), together with the manufacture and use of tools, is a distinguishing characteristic of humans. In his view, all truly human, higher mental processes involve the use of speech.

From Dialogue to Self-Regulation: The Sociogenetic Law

The original function of language is communication, Vygotsky (1997a) argued. Children and adults communicate their intentions and steer or guide the behavior of the social other. This communicative process is at first very asymmetric in that the adults do most of the guiding and attribute meaning – at first often counterfactually – to the infants' utterances. Very soon, however, children become able to use words to accomplish things and, most importantly, to steer their own behavior. Self-regulation thus originates in dialogue.

A simple example is that of a child learning to cross a street. In a first stage, caregivers will tell the child to watch out carefully and children will simply follow this instruction. In the next stage, the child will say aloud to watch out and thus instruct itself. In the final stage, the child will merely think this instruction. Thus, a process that was originally shared between two persons, an interpsychological process, has become an individual, intrapsychological process. The child is applying to itself what was first applied by others and thereby mastered its own behavior through the use of conventional signs. With Baldwin, Janet, Royce, and others, Vygotsky (1997a) claimed that all higher mental processes are originally shared between persons, and he coined the developmental transition from inter- to intramental processes the sociogenetic law.

This example is linked with another phenomenon that Vygotsky discussed in his writings, namely that of the intersection of thought and speech. In his view, thought and speech have different genetic roots and only merge to form verbal thought or intellectual speech at a later stage in phylogenetic and ontogenetic development. Thus, in human children both preintellectual speech (i.e., babbling) and preverbal thinking (e.g., infants expect to find an object where it was hidden previously) can be observed, while in the phylogenetic domain one sees animals solving simple problems (e.g., using an artifact to reach a goal) and displaying primitive communicative means (e.g., expressive movements). It is only in humans, however, that speech and problem solving merge to form the powerful capacity of verbal thought. Problem solving now no longer proceeds in the visual field, Vygotsky (1997b) argued, but in what he called a 'semantic field.'

Egocentric and Inner Speech

Vygotsky claimed that speech develops in an intrapersonal situation, and subsequently it is used to steer the self-received additional support in studies of the phenomenon of egocentric speech. It was Piaget who first described the phenomenon of children who while carrying out some activity speak for themselves in a way that is often not intelligible to others. Piaget hypothesized that such speech does not take account of the viewpoint of the social other and reflects younger children's general inability to transcend their own point of view. In his view, children are at first unable to see the world from another perspective than their own and only gradually realize, under the pressure of their peers, that theirs is a specific viewpoint.

Vygotsky criticized Piaget's interpretation of the phenomenon of egocentric speech and conducted a series of experiments to argue that children's egocentric speech is social in the sense that it addresses social others (Van der Veer and Valsiner, 1991: 363–369; Vygotsky, 1987). He noted that egocentric speech is absent or greatly reduced when the child is alone (which suggests that it is meant for others) and that its incidence rises when the child is confronted with unexpected problems (which suggests that it has some function in the solution of problems). He also observed that egocentric speech becomes less intelligible as the child grows older. This set of results led him to posit that egocentric speech (1) originates in normal, communicative speech and branches off at a later stage; (2) has as its function to steer the child's behavior when the need arises; and (3) becomes less and less intelligible for the outsider until it becomes proper inner speech. Egocentric speech, then, is an intermediate stage between normal, communicative speech and inner speech. Like communicative speech, it is audible, and like inner speech, it serves to guide the child's thinking.

Vygotsky went on to speculate about the properties of inner speech. He argued that its syntax must be special: it is speech that is abbreviated and tends toward predicativity, that is, omitting the subject of the utterance. He also suggested that inner speech has special semantic properties: personal sense dominates over dictionary meaning. He thus believed to have demonstrated that, in the process of becoming private, speech undergoes a series of transformations.

Word Meaning and Conceptual Development

Above it was said that children use conventional signs to regulate their own behavior. Toward the end of his life, Vygotsky began emphasizing that the meanings of these conventional signs or words change over time. Children use the same words as adults do to designate specific objects but the meaning of these words may be quite different. Thus, in the child the word 'farmer' may evoke picture book images of a person taking care of cattle and growing crops, while adults may think of an entrepreneur who tries to maximize profit. Using various strategies, Vygotsky tried to prove experimentally that a mature understanding of many concepts is only reached in adolescence. In this connection, he introduced a distinction between mature, academic, or scientific concepts, on the one hand, and immature or everyday concepts, on the other hand. Supposedly, the academic concepts express the true nature of things and are linked to other similar concepts. Together they form a coherent network covering a field of knowledge. Academic concepts are taught at school. Everyday concepts, on the other hand, often focus on irrelevant features of some

subject and together form a disconnected whole. Everyday concepts are more or less independently acquired by children in informal settings.

Vygotsky's basic idea was that academic concepts and everyday concepts mutually enrich each other. The everyday concept of a farmer becomes enriched because the child learns the nonapparent fact that farmers take part in a market economy and try to realize certain goals. The child's academic concept of a farmer is enriched because the abstract notion of entrepreneurs dealing with cattle and crops is filled with concrete facts of daily farmer life and thereby gets body and flesh. Thus, academic concepts presuppose everyday concepts, build upon them, but once acquired by the child they alter its everyday concepts in fundamental ways. The result of the interplay between everyday and academic concept is an elaborate and rich notion of the concept of farmer. Vygotsky (1935) speculated that the introduction of systems of academic concepts in school would restructure the child's whole style of thinking; that is, he assumed that the introduction of such concepts in specific subject areas would generalize to other areas. Education leads development, as he used to put it. Instruction in the school setting propels child development along lines that are each time specific for a certain culture or society.

Predicting Mental Growth: The Zone of Proximal Development

In several lectures delivered in 1933, Vygotsky transposed the sociogenetic law to the domain of cognitive assessment and prognosis. He claimed that if one were to measure what the child is able to do in cooperation with someone else (interpsychologically), one would be able to predict what the child can do independently (intrapsychologically) later on. At any given moment, both the child's dependent (e.g., with hints from an adult) and independent performance on an IQ test can be measured. The difference between the two IQ scores yields an indication of the child's cognitive potential. Some children will be able to profit more from adult hints and will improve their independent score more substantially than others. These children show the richest developmental potential, Vygotsky (1935) argued. To measure what children can do independently is to measure what they can do already. To measure what children can do with the help of an adult is to measure what they will be able to do in the near future, in what Vygotsky coined their zone of proximal development. In this way, he criticized the contemporary use of IQ tests and suggested a procedure for the assessment of cognitive potential (Grigorenko and Sternberg, 1998).

Relevance for Contemporary Science

Vygotsky's ideas inspired an enormous and still growing amount of research, both in and outside the former Soviet Union (Miller, 2011).

Wertsch (1985) tried to corroborate Vygotsky's view that self-regulation develops in dialogue by investigating mother–child joint problem solving. He and others have also advanced proposals to fuse Vygotsky's ideas on dialogue and its relevance for self-regulation with those of Bakhtin.

Vygotsky's view of the nature and function of egocentric and inner speech has become a hotly debated topic and has been developed further by others (e.g., A.A. Leontiev, A.R. Luria, and A.N. Sokolov). The Eastern European research is well represented in Sebeok (1968), Slobin (1966), Prucha (1983), and Wertsch (1978). Zivin (1979) gives a representative account of both the research tradition initiated by Vygotsky and Luria and its Western counterparts.

Vygotsky's concept of word meaning and his view of the dialectic interplay between academic and everyday concepts in education have also led to a whole research tradition. An excellent and critical account can be found in Wertsch (1985), while Van der Veer and Valsiner (1991) have traced how Vygotsky's research into the development of concepts was used by Western clinical psychologists to diagnose various groups of mental patients (e.g., schizophrenics and patients suffering from aphasia).

The concept of the zone of proximal development has inspired an immense number of publications. These can be divided roughly into two groups. The first group consists of publications that deal with the influence of education on development. It is analyzed, for example, how classroom discussions can promote a zone of proximal development in students, or what kind of teaching is needed to promote children's use of language for self-regulation. An overview of this kind of work can be found in Moll (1990). The second group consists of publications dealing with the assessment of learning potential. In this approach the zone of proximal development is used as an individual difference metric designed to provide information about individual students; that is, Vygotsky's original idea to measure aided and unaided performance on some IQ test is operationalized, and it is investigated whether such repeated measurements yield important prognostic information (cf. Grigorenko and Sternberg, 1998).

All in all, it can be said that although many of Vygotsky's ideas are debatable (see Van der Veer, 2014) and will have to be discarded or refined, they nevertheless have inspired a great and still growing number of researchers in the humanities. In itself this is remarkable for a man who published very little in languages other than Russian, who died more than 60 years ago, and whose international fame began only 30 years after his death. The fact that Vygotsky's legacy still inspires contemporary research marks him as a true classic thinker of the humanities.

See also: Constructivism and Constructionism: Methodology; Cooperative Learning in Schools; Cultural Psychology; Language and Literature; Luria, Aleksandr Romanovich (1902–77); Pavlov, Ivan Petrovich (1849–1936); Piaget's Theory of Cognitive Development; Piaget's Theory of Human Development and Education; Piaget, Jean (1896–1980); Psychologism, History of; Self-Regulation in Adulthood; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development; Vygotsky's Theory of Human Development and New Approaches to Education; Wisdom, Psychology of.

Bibliography

Feigenberg, I.M. (Ed.), 1996. L.S. Vygotsky: Nachalo Puti. Jerusalem Publishing Centre, Jerusalem.

Grigorenko, E.L., Sternberg, R.J., 1998. Dynamic testing. Psychological Bulletin 124, 75–111.

Miller, R., 2011. Vygotsky in Perspective. Cambridge University Press, Cambridge.

Moll, L.C. (Ed.), 1990. Vygotsky and Education: Instructional Implications and Applications of Sociohistorical Psychology. Cambridge University Press, Cambridge, MA.

Prucha, J., 1983. Pragmalinguistics: East European Approaches. Benjamins, Amsterdam.

Sebeok, T.A. (Ed.), 1968. Current Trends in Linguistics. Soviet and East European Linguistics, vol. 1. Mouton, The Hague.

Slobin, D.I., 1966. Soviet psycholinguistics. In: O'Connor, N. (Ed.), Present-day Russian Psychology; a Symposium by Seven Authors. Pergamon Press, Oxford.

Van der Veer, R., 2014. Lev Vygotsky. Bloomsbury, London.

Van der Veer, R., Yasnitsky, A., 2011. Vygotsky in English: what still needs to be done. Integrative Psychological and Behavioral Science 45, 475–493.

Van der Veer, R., Valsiner, J., 1991. Understanding Vygotsky: A Quest for Synthesis. Blackwell, Oxford.

Vygodskaya, G.L., Lifanova, T.M., 1996. Lev Semenovich Vygotskij: Zhizn', Deiatel'nost', Shtrikhi K Portretu. Smysl, Moscow.

Vygotsky, L.S., 1925/1971. The Psychology of Art. MIT, Cambridge, MA.

Vygotsky, L.S., 1935. Umstvennoe Razvitie Detey V Protsesse Obucheniya. Uchpedgiz, Moscow and Leningrad.

Vygotsky, L.S., 1962. Thought and Language. MIT, Cambridge, MA.

Vygotsky, L.S., 1987. The problem of speech and thinking in Piaget's theory. In: Rieber, R.W., Carton, A.S. (Eds.), The Collected Works of L.S. Vygotsky, Problems of General Psychology, vol. 1. Plenum Press, New York, pp. 53–91.

Vygotsky, L.S., 1994. Tool and symbol in child development. In: Van der Veer, R., Valsiner, J. (Eds.), The Vygotsky Reader. Blackwell, Oxford, pp. 99–174.

Vygotsky, L.S., 1997a. The history of the development of higher mental functions. In: Rieber, R. (Ed.), The Collected Works of L.S. Vygotsky, The History of the Development of Higher Mental Functions, vol. 4. Plenum Press, New York, pp. 1–251.

Vygotsky, L.S., 1997b. Preface to Koffka. In: Rieber, R.W., Wollock, J. (Eds.), The Collected Works of L.S. Vygotsky, Problems of the Theory and History of Psychology, vol. 3. Plenum Press, New York, pp. 195–232.

Wertsch, J.V., 1978. Recent Trends in Soviet Psycholinguistics. Sharpe, New York.

Wertsch, J.V., 1985. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge, MA.

Yasnitsky, A., 2011. Vygotsky circle as a personal network of scholars: restoring connections between people and ideas. Integrative Psychological and Behavioral Science 45, 422–457.

Zavershneva, E.Yu., 2010. The Vygotsky family archive (1912–1934). Journal of Russian and East European Psychology 48, 14–33.

Zivin, G. (Ed.), 1979. The Development of Self-regulation through Private Speech. Wiley, New York.

Vygotsky's Theory of Human Development and New Approaches to Education

Eugene Matusov, University of Delaware, Newark, DE, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article will focus on describing the main tenets of Vygotsky's theory of human development, current critique of these tenets, and new approaches to education stemming from Vygotsky's work and its critique. Vygotsky's legacy is in his emphasis on social, historical, and cultural aspects of development and learning, semiotic mediation, activity, and social interaction. According to Vygotsky, education can lead children's development and learning is shaped by cultural tools and activities in which it is embedded. This article is organized around three major interconnected issues: what develops, what are the processes of development, and what are the directions of development.

Vygotsky's legacy is in his emphasis on social, historical, and cultural aspects of development and learning, semiotic mediation, activity, and social interaction. After about 50 'dark years' of silence, from the early 1930s (immediately after his death) to the end of 1970s, caused in part by the repressive political circumstances in the Soviet Union, social sciences have been experiencing Vygotsky's renaissance. Since the 1970s, Vygotsky has become a 'root theoretician' for a family of approaches in social sciences including sociolinguistics (Heath, 1983), anthropology (Lave and Wenger, 1991), psychology (Cole, 1996; Rogoff, 1990), education (Cazden and Cordeiro, 1992), and literacy studies (Gee, 1996), among others. In fact, it is probably not an overgeneralization to say that currently, developmental psychology and education have been highly influenced by Vygotsky's work. This article will describe the main tenets of Vygotsky's theory of human development, current critique of these tenets, and new approaches to education stemming from a dialogue with Vygotsky's work. This article is organized around three major interrelated issues: what develops, what are the processes of development, and what are the directions of development.

What Develops?

Vygotsky argued that human development is based on intertwining of two lines: (1) natural based on processes of biological maturation and (2) cultural based on mastery and use of cultural means. Psychologically, these two lines are represented by lower and higher mental functions, respectively, such as memory, attention, and intelligence. For example, mechanical memorization of unrelated bits of information mainly involves lower functions of memory. Meanwhile, memorization that is based on mental structuring of information in some hierarchical classes (e.g., axe, saw, and hammer are tools) or a story connecting items to remember involves higher mental functions of memory. For instance, when a child is asked to remember the word 'theater' and a picture of a crab, the child creates a story: "The crab is looking at the stones on the bottom, it is beautiful, it is a theater" (Vygotsky, 1997: p. 181). Vygotsky insisted that higher mental functions are formed by a reorganization of lower mental functions mediated by a cultural sign (e.g., mnemonics). Higher mental functions allow a person to control his or her behavior, "in the elementary form [of memory – EM], something is remembered; in the higher form, humans remember something … the basic characteristic of human behavior in general is that humans personally influence their relations with the environment and through that environment personally change their behavior, subjugating it to their control" (Vygotsky, 1978: p. 51). Thus, according to Vygotsky, what develops is higher mental functions via a mastery of cultural signs (i.e., *semiotic mediation*), leading to a person's development of *self-regulation* and self-control.

Probably the most elaborate application of Vygotsky's stress on cultural mediation as the object of development has been in the 'sign' pedagogy developed by the Russian educational psychologist Davydov (Davydov and Kilpatrick, 1990). Davydov and his colleagues rewrote an entire elementary school curricula (e.g., math, sciences, and language art) representing it as a series of dialectical contradictions of mental actions mediated by cultural signs. They also changed the curriculum sequence putting the abstract concepts – higher level cultural mediators – ahead of concrete application of them (e.g., putting study of fraction and algebra ahead of integer arithmetic). For example, according to Davydov and Kilpatrick (1990), the fraction represents the contradiction between the permanence of the object's length (N) and the variability of its measurement (A), depending on the unit (c) mediating the measurement. Mathematical formula $A/c = N$ is a model of the fraction that allows first-grade students to make comparisons involving any measurement unit.

Yet, a growing literature has implicitly or explicitly criticized Vygotsky for making mental functions the object of development (Rogoff, 1990). This work has argued that culturally mediated functions are not self-contained, but instead embedded in sociocultural activities. Lave (1988) introducing the notion of *situated cognition* demonstrates, for example, that people's computation in school-like math tasks and grocery shopping is not psychologically equal even when the problems are mathematically isomorphic. The problem solving is mediated and structured differently in these two conditions because the problems are embedded in different sociocultural activities with different goals and different social relations. In response to these and other related findings, alternative objects of development have been proposed.

Rogoff (1990) proposes a *sociocultural approach* in which a person's participation in sociocultural activities and communities is the object of development. Development is seen as *transformation of participation* in a sociocultural activity – transformation that is socially valuable. For example, when a child whose parents had previously read books to her starts asking her parents to read a book, this change indicates a transformation of the child's participation in the reading practice. Some have argued that mediation by cultural tools and scripts pointed by Vygotsky is only a part of the developmental story and should be considered together with aspects such as social relations, division of labor, community rules and norms, and relations with other practices and communities (Engeström et al., 1999). These features of social life were often ignored in cultural research and research on child development conducted by Vygotsky and his colleagues (Vygotsky, 1978; Vygotsky and Luria, 1993).

How It Develops?

According to Vygotsky's general genetic law of cultural development, any function in the child's cultural development appears on three planes: natural, social, and psychological. Vygotsky emphasized a dynamic and emergent character of development: each plane emerges on the periphery ('in an embryonic state') of the previous plane (Vygotsky, 1997). First, a function appears on the natural plane, when a child finds him or herself in an environment and can observe the function as used by other people without the child's understanding or active participation in it. For example, in human society, people use gestures in their communication with each other and they often are inclined to (mis)interpret some of child's actions as gestures. When a very young child tries unsuccessfully to grasp a remote object, his or her caregiver may (mis)interpret it as a command gesture to bring the object to the child.

The second plane of development is social – the child is an active participant in the function socially distributed among people, although this participation may not be necessarily conventional or fully grasped by the child on his or her own and requires the involvement of other people for the participation's success. Continuing Vygotsky's example of the development of the index gesture, at some point, the young child notices that she can control adults by stretching her arm toward a desired object – the adults often bring the remote object to the child. Thus, on the social plane, the child actively participates in the social function of controlling other people rather than finding herself in the function fully organized by other people as it was on the natural plane.

Finally, on the third plane, the function transforms from external social into internal psychological. The child's command of others transforms into the index gesture controlling the child's own attention. Now the child's stretched arm controls not others but the child herself; her own attention is mediated by the index gesture. The developmental cycle is completed: the cultural higher mental function of the index gesture that initially exists only outside the child on the natural plane becomes the child's own on the psychological plane for the child's self-regulation (Vygotsky, 1997). According to Vygotsky, development involves nonlinear and systemic processes of changing relationships and organizations of mental functions.

Recently, Vygotsky's so-called general genetic law of development has been challenged on both empirical and theoretical grounds. Thus, Vygotsky's explanation of development of index gesture faces critique (e.g., Kaye, 1982) coming from recent findings that index finger extensions can be observed in infants as young as 2 months old (Fogel and Hannan, 1985; Trevarthen, 1977) and seems to be developing in relative independence from grasping movements (Fogel and Thelen, 1987). It seems to be utilized by the infant first for touching objects at about 6 months and then for an extension of the orienting response at about 11 months (Lempert and Kinsbourne, 1985). During the period between 8 and 12 months, pointing seems to shift from an exploratory to an indicating function (Fogel and Thelen, 1987). It seems that Vygotsky's model of the development of the index gesture might be still relevant for an illustration of how a misunderstanding between child and adult guides child's development (Lock, 1980). Also, Matusov (1998) has challenged Vygotsky's insistence on the mastery of solo activity and autonomous self-regulation as the acme of the human psychological development, suggesting an ethnocentric bias on the part of Vygotsky.

The issue of the relationship between development and education led Vygotsky to introduce the notion of 'the zone of proximal development' (ZPD) as the gap between actual and potential development (Vygotsky, 1978). Vygotsky offered two major interpretations of ZPD. The first interpretation is the ZPD as a gap between the psychological (i.e., completed) plane, what the child can do by him or herself, and the social (i.e., potential) plane, what the child can do with the help of more capable others (adults or peers), of development.

This use of ZPD defines 'teachability' of the child in a specific activity or in problem solving. If an activity or problem can be accomplished by the child with the help of more capable others, this activity or skill is considered possible to teach the child. If, however, the activity or skill cannot be accomplished by the child with the help of more capable others, it is considered not useful to teach to the child. Vygotsky coined the term *formative experiment* to provide methodology for the investigation of ZPD. According to this methodology, a particular higher level psychological function is formed through some kind of instruction with a more knowledgeable other or manipulation of the context and tools used by a child. The successful formation of the psychological function verifies the existence of ZPD and provides insight into the genesis of the function (Griffin et al., 1993).

Unlike Piaget, who believed that instruction should follow development based on biological maturation and a child's active experimentation with the environment, Vygotsky argued that guidance can, should, and does lead the development. Both Piaget and Vygotsky would differently define what is currently called 'developmentally appropriate curriculum.' Piaget insisted that learning is essentially an individual endeavor and that adults can only facilitate it by providing an enriched stimulating learning environment and opportunities for the children to share and discuss their egocentric thinking with each other to promote disequilibrium in the child's

thinking. Adults should not interfere in the child's individual thinking because it can only lead to imposition of the adult's ideas onto the child – what Piaget called 'sociocentrism' (Piaget and Smith, 1995). In contrast, Vygotsky encouraged adults to provide guidance and help and to engage students in activities that are beyond their individual levels of competence ('performance before competence,' Cazden and Cordeiro, 1992).

Currently, there are attempts in education to develop 'scientific instruction' using a medical model of 'educational interventions' (Hargreaves, 1996, 1997). The essence of these efforts is to develop educational diagnostics of students' skill deficits and prescribe an appropriate dose of standardized guidance providing no less and no more help to the student than is required (Newman et al., 1989). For that purpose, some researchers have tried to develop a 'ZPD test' (Dixon-Krauss, 1996; Moran, 1997) – a standardized assessment of a student's teachability. However, it is doubtful that a reliable ZPD test would be possible to develop because, as Newman et al. (1989) demonstrate, the notion of ZPD is relational. Student's teachability depends not only on the student but also on the teacher (and broader communities in which the child participates). Thus, no test of the child alone would accurately determine the child's teachability – the teacher always counts as well as broader and local contexts and activism of the child.

According to this neo-Vygotskian view, the ZPD is applied not only to the student but also to the teacher. Both the teacher and the student try to manage the uncertainty that their joint activity creates. During teacher–student instructional interaction, the student learns how to do the classroom activity, while the teacher learns how to guide the student. Each engages in their own ZPD that mutually constitute each other – the student's learning is shaped by the teacher's guidance, while the teacher's guidance is shaped by the student's ongoing learning.

The second interpretation of the ZPD offered by Vygotsky is about the activity in which the child is involved. Vygotsky argued that at different ages, different activities become leading forces in a child's development. For young children, play is the *leading activity*, "play creates a zone of proximal development of the child. In play a child always behaves beyond his average age, above his daily behavior … As in the magnifying glass, play contains all developmental tendencies in a condensed form and is itself a major source of development … Action in the imaginative sphere, in an imaginary situation, the creation of voluntary intentions, and the formation of real-life plans and volitional motives – all appear in play and make it the highest level of preschool development" (Vygotsky, 1978: p. 102). He introduced the notion of 'leading activity' as an activity that determines the child's development. El'konin and Leont'ev elaborated this notion by creating a scale of leading activities from infancy to adulthood as the base of age periodization in a given society (Griffin and Cole, 1984). The concept of age periodization has become criticized as ethnocentric by several authors (Lave and Wenger, 1991; Matusov et al., 2007a; Rogoff, 1990).

Bruner and his colleagues developed the didactic notion of 'scaffolding' in their attempts to operationalize Vygotsky's ZPD in their study of dyadic interactions involving adults tutoring children (Wood et al., 1976). Scaffolding involves a series of adult strategies that assume responsibility for children's learning, such as focusing the child on the task, simplifying the task to fit the child's current mastery, maintaining the child's motivation for working on the task, managing the child's level of frustration and risk involved in the problem solving, and demonstrating an idealized version of the required act (Rogoff, 1990). Like Vygotsky's ZPD, the scaffolding metaphor is aimed to explain the child's gaining of skills in social interaction that the child may apply later on when working alone. It focuses on how the adult fine-tunes the extent of help he or she provides the child for each successful activity outcome while viewing the child as inept, "One sets the game, provides a scaffold to assure that the child's ineptitudes can be rescued or rectified by appropriate interaction, and then removes the scaffold part by part as the reciprocal structure can stand on its own" (Bruner, 1983: p. 60).

In a critique of the concept of scaffolding, Griffin and Cole (1984) emphasize the role of the child in his or her own development, "The metaphor [of scaffolding] becomes more problematic when we focus not on the execution of a specific task but on the changes in the child" (p. 47). This concept makes it difficult to address the issue of the child's creativity, "If adult support is the universal source of a child's development, then there is a strong sense of theology – children's development is circumscribed by adults' achieved wisdom" (p. 47). If Vygotsky's concept of ZPD seems to be too focused on the child, then the concept of scaffolding seems to be too focused on the role of the adult in guidance. As a consequence, both concepts are limited by dyadic interaction between more and less knowledgeable partners. Besides, cross-cultural research on guidance suggests that scaffolding is not a universal guidance strategy (Rogoff et al., 1993).

Unlike scaffolding, the concept of *guided participation* developed by Rogoff (1990) can be applied in diverse cultural activities because it focuses on the transformation of participation guided not only by more knowledgeable partners but also by sociocultural tools, culturally defined goals and problems, and social arrangements of joint activity: "Interaction with other people assists children in their development by guiding their participation in relevant activities, helping them adapt their understanding to new situations, structuring their problem-solving attempts, and assisting them in assuming responsibility for managing problem solving. This guidance of development includes tacit and intuitive forms of communication and distal arrangements of children's learning environments; it is often not designed for the instruction of children and may not involve contact or conversation. The model is one of routine arrangements and engagements that guide children's increasingly skilled and appropriate participation in the daily activities valued in their culture" (Rogoff, 1990: p. 191).

Lave (1992, April) insists that learning is inherent to activity and an aspect of any activity. Learning occurs even despite expectations and wills of experienced members of community – it is not a matter of whether students learn in school but a matter of what they learn. The students

might actually learn what they were not expected to learn and might not learn what was expected for them to learn. Learning is not an independent activity among other activities but, like development, it is an aspect of any activity in the world.

Lave and Wenger (1991) argue that learning is situated in *communities of practice*. Learning is always a question about membership in the community, about participation in the community practice. A novice is not simply a person who lacks some entities called 'skills' but rather a newcomer who needs to negotiate her or his participation in the community practice (Wenger, 1998). A person engages in ongoing negotiation of membership/participation in different communities of practice. Learning, as a process of negotiation and renegotiation of participation in the community of practice, is often not prime-time community business; rather it is always going on in the periphery of community activity. Because the community is aware of newcomers, the peripheral processes of negotiation and renegotiation of participation have a legitimate character. Newcomers are anticipated and usually (but not always!) organized by the community. An analysis of situated learning allows Lave and Wenger to construct a new productive concept of 'legitimate peripheral participation' that becomes the main definition of learning and development.

Learning as a communal process is both a descriptive and guiding approach (Matusov et al., 2007b). A view of learning as communal processes embedded in communal practices has inspired many educational practitioners and researchers to explore and define new conceptual and empirical forms of guidance that can be used in schools such as *instructional conversations* (Tharp and Gallimore, 1988), *reciprocal teaching* (Brown and Palincsar, 1987), *cognitive apprenticeship* (Rogoff, 1990), *community of learners* (Brown and Campione, 1994; Matusov et al., 2012; Rogoff et al., 1996), *practice- and problem-based learning* (Gijselaers, 1996), *building a professional community* (Matusov et al., 2005), *dialogic inquiry* (Wells, 1999), *dialogic pedagogy* (Matusov, 2009; Sidorkin, 1999; Wegerif, 2007), and *authorial teaching and learning* (Matusov, 2011a). This family of instructional approaches and models shares at least the following important principles: learning is a communal process, learning is embedded in activities and practices in which it occurs, learning involves development and negotiation of new communal identities, students' guided initiation of discourse and definition of problems and goals are crucial for becoming an active member of a community of practice, ownership for guidance and learning should be shared among students and between the students and the teacher, and a community is based on practice and communication. Nonschool sociocultural research on learning shares many of the same principles (Cole, 1996; Heath, 1994; Wenger, 1998).

Finally, influenced by another Soviet scholar Mikhail M. Bakhtin (1896–1975), a philosopher and literary theoretician of the dialogic framework, there have been increasing doubts that the purpose of development and education is in the elimination of gaps between less and more capable people, as Vygotsky seemed to imply in his notion of ZPD. Bakhtin argued that elimination of gaps between consciousnesses is both impossible and undesirable (Bakhtin, 1986, 1999). Matusov (2011b) has come to a conclusion that the important and irreconcilable difference between Vygotsky and Bakhtin comes from their different assumptions about the nature of human consciousness. Vygotsky treated consciousnesses as potentially transparent to each other, while Bakhtin treated consciousnesses as essentially opaque to each other.

Toward What Does the Development Go?

According to Vygotsky, development leads toward a buildup of higher mental functions with mediated cognition increasingly playing the central role among all other psychological functions that promote self-regulation in the individual. Vygotsky's primary focus in his *cultural-historical theory* of development was on how human society reproduces itself. Vygotsky often saw non-Western cultures as 'historical slices' of the past development of Western civilization (Luria, 1976; Vygotsky and Luria, 1993). When the developmental directionality was considered, Vygotsky was, at least, inconsistent with his own tenet that stressed the contextual and activity specificity of intellectual functioning (Cole, 1988; Scribner, 1977). Prioritizing self-regulation, mastery of independent working, literacy, schooling, scientific concepts, universal rationality, decontextualization, universality in societal development, and systematic hierarchical thinking as the pinnacle of individual's development reveals Vygotsky's ethnocentrism (Matusov, 1998, 2011b; Matusov and Hayes, 2000; Rogoff, 1990; van der Veer and Valsiner, 1991; Wertsch, 1985).

A diversity of goals of different cultural communities necessitates defining development in terms of progress toward more responsible participation in specific communities of practice rather than assuming that development is a generic and universal process independent of the goals and institutions of the communities in which an individual develops. As Tharp and Gallimore (1988) argue, "Boys in Micronesia, where sailing a canoe is a fundamental skill, will have a ZPD for the skills of navigation, created in interaction with the sailing masters. A girl in the Navajo weaving community will have experiences in a zone not quite like any ever encountered by the daughters of Philadelphia" (p. 31). At the same time, the developing individual contributes to the further development of the practices (and goals and institutions) of the community.

Griffin and Cole (1984) argue that the ZPD is not only about a gap between the child's present and the adult's past but also between the child's present and the society's future. They criticized the conservative and reproductive nature of learning and development assumed in both Vygotsky's notion of ZPD and Bruner's notion of scaffolding. These authors argue that development and education is "a dialogue between the child and his [unknown – EM] future; [it should not be – EM] […] a dialogue between the child and an adult's [known – EM] past" (Griffin and Cole, 1984: p. 62). There is an increasing tendency in the social sciences to consider the developmental telos as a social construction (Gergen, 1994) and dialogic negotiation (Matusov et al., 2007a) of cultural, institutional, communal, and personal values that have both stable and dynamic components that can conflict with each other

(Lemke, 1995). The historical time when positive values of literacy, school, science, rationality, intelligence, and the Western civilization were unquestionable in the social sciences is gone. Stemming from Vygotsky's cultural-historical theory, new approaches to development and learning are emerging. There has been a growing call for viewing development and education as production and active making of culture, rather than reproduction, focusing on human authorial agency transcending of the given (Bibler, 2009; Lobok, 2012; Matusov, 2011a).

See also: Bruner's Theory of Cognitive Development; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Scientific Concepts: Development in Children; Vygotsky's Sociocultural Theory; Vygotsky's Theory of Cognitive Development; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Bakhtin, M.M., 1986. Speech Genres and Other Late Essays. University of Texas Press, Austin.

Bakhtin, M.M., 1999. Problems of Dostoevsky's Poetics, vol. 8. University of Minnesota Press, Minneapolis.

Bibler, V.S., 2009. The foundations of the school of the dialogue of cultures program. Journal of Russian & East European Psychology 47 (1), 34–60.

Brown, A.L., Campione, J.C., 1994. Guided discovery in a community of learners. In: McGilly, K. (Ed.), Classroom Lessons: Integrating Cognitive Theory and Classroom Practice. The MIT Press, Cambridge, MA, pp. 229–270.

Brown, A.L., Palincsar, A.S., 1987. Reciprocal teaching of comprehension strategies: a natural history of one program for enhancing learning. In: Day, J.D., Borkowski, J.G., et al. (Eds.), Intelligence and Exceptionality: New Directions for Theory, Assessment, and Instructional Practices. Ablex Publishing Corp, Norwood, NJ, USA, pp. 81–132.

Bruner, J., 1983. Child's Talk: Learning to Use Language. Norton, New York.

Cazden, C.B., Cordeiro, P., 1992. Whole Language Plus: Essays on Literacy in the United States and New Zealand. Teachers College Press, New York.

Cole, M., 1988. Cross-cultural research in the sociohistorical tradition. Human Development 31 (3), 137–151.

Cole, M., 1996. Cultural Psychology: A Once and Future Discipline. Harvard University Press, Cambridge, MA.

Davydov, V.V., Kilpatrick, J., 1990. Types of Generalization in Instruction: Logical and Psychological Problems in the Structuring of School Curricula. National Council of Teachers of Mathematics, Reston, VA.

Dixon-Krauss, L., 1996. Vygotsky in the Classroom: Mediated Literacy Instruction and Assessment. Longman Publishers, White Plains, New York.

Engeström, Y., Miettinen, R., Punamèaki-Gitai, R.-L., 1999. Perspectives on Activity Theory. Cambridge University Press, New York.

Fogel, A., Hannan, T.E., 1985. Manual actions of nine- to fifteen-week-old human infants during face-to-face interaction with their mothers. Child Development 56 (5), 1271–1279.

Fogel, A., Thelen, E., 1987. Development of early expressive and communicative action: reinterpreting the evidence from a dynamic systems perspective. Developmental Psychology 23 (6), 747–761.

Gee, J.P., 1996. Social Linguistics and Literacies: Ideology in Discourses, second ed. Taylor & Francis, London.

Gergen, K.J., 1994. Toward Transformation in Social Knowledge. Sage Publications, London.

Gijselaers, W., 1996. Bringing Problem-based Learning to Higher Education: Theory and Practice. Jossey-Bass, San Francisco.

Griffin, P., Belyaeva, A., Soldatova, G., Collective, t. V.-H, 1993. Creating and reconstituting contexts for educational interactions, including a computer program. In: Forman, E.A., Minick, N., Stone, C.A. (Eds.), Contexts for Learning: Sociocultural Dynamics in Children's Development. Oxford University Press, New York, pp. 120–152.

Griffin, P., Cole, M., 1984. Current activity for the future: the Zo-ped. In: Rogoff, B., Wertsch, J.V. (Eds.), Children's Learning in the "Zone of Proximal Development", New Directions for Child Development, vol. 23. Jossey Bass, San Francisco, pp. 45–64.

Hargreaves, D.H., 1996. Teaching as a Research-based Profession: Possibilities and Prospects. Retrieved from: http://www.bera.ac.uk/files/resourcesfiles/educationalresearch/hargreaves_1996.pdf.

Hargreaves, D.H., 1997. In defense of research for evidence-based teaching: a rejoinder to Martyn Hammersley. British Educational Research Journal 23 (4), 405–419.

Heath, S.B., 1983. Ways with Words: Language, Life, and Work in Communities and Classrooms. Cambridge University Press, Cambridge, UK.

Heath, S.B., 1994. The project of learning from inner-city youth perspective. In: Villarruel, F.A., Lerner, R.M. (Eds.), Promoting Community-based Programs for Socialization and Learning, New Directions for Child Development, vol. 63 (Spring). Jossey-Bass, San Francisco, pp. 25–34.

Kaye, K., 1982. The Mental and Social Life of Babies: How Parents Create Persons. University of Chicago Press, Chicago.

Lave, J., 1988. Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life. Cambridge University Press, Cambridge, UK.

Lave, J., April 1992. Learning as Participation in Communities of Practice. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Lave, J., Wenger, E., 1991. Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge, UK.

Lemke, J.L., 1995. Textual Politics: Discourse and Social Dynamics. Taylor & Francis, London.

Lempert, H., Kinsbourne, M., 1985. Possible origin of speech in selective orienting. Psychological Bulletin 97 (1), 62–73.

Lobok, A., 2012. My educational odyssey to dialogic agency-based probabilistic pedagogy. Journal of Russian and Eastern European Psychology 50 (6), 5–8.

Lock, A., 1980. The Guided Reinvention of Language. Academic Press, London; New York.

Luria, A.R., 1976. Cognitive Development, Its Cultural and Social Foundations. Harvard University Press, Cambridge, MA.

Matusov, E., 1998. When solo activity is not privileged: participation and internalization models of development. Human Development 41 (5–6), 326–349.

Matusov, E., 2009. Journey into Dialogic Pedagogy. Nova Science Publishers, Hauppauge, NY.

Matusov, E., 2011a. Authorial teaching and learning. In: White, E.J., Peters, M. (Eds.), Bakhtinian Pedagogy: Opportunities and Challenges for Research, Policy and Practice in Education across the Globe. Peter Lang Publishers, New York, pp. 21–46.

Matusov, E., 2011b. Irreconcilable differences in Vygotsky's and Bakhtin's approaches to the social and the individual: an educational perspective. Culture & Psychology 17 (1), 99–119.

Matusov, E., DePalma, R., Drye, S., 2007a. Whose development? Salvaging the concept of development within a sociocultural approach to education. Educational Theory 57 (4), 403–421.

Matusov, E., Hayes, R., 2000. Sociocultural critique of Piaget and Vygotsky. New Directions in Psychology 18 (2–3), 215–239.

Matusov, E., St. Julien, J., Hayes, R., 2005. Building a creole educational community as the goal of multicultural education for preservice teachers. In: Barnes, L.V. (Ed.), Contemporary Teaching and Teacher Issues. Nova Publishers, Hauppauge, NY, pp. 1–38.

Matusov, E., St. Julien, J., Lacasa, P., Alburquerque Candela, M., 2007b. Learning as a communal process and as a byproduct of social activism. Outlines: Critical Social Studies 1 (1), 21–37.

Matusov, E., von Duyke, K., Han, S., 2012. Community of learners: ontological and non-ontological projects. Outlines: Critical Social Studies 14 (1), 41–72.

Moran, C., 1997. Balanced Approach to Literacy: An Instructional Guide for a Course for Preservice and New Teachers of Literacy. University of California at Santa Cruz, Santa Cruz, CA.

Newman, D., Cole, M., Griffin, P., 1989. The Construction Zone: Working for Cognitive Change in School. Cambridge University Press, Cambridge, UK.

Piaget, J., Smith, L., 1995. Sociological Studies. Routledge, London; New York.

Rogoff, B., 1990. Apprenticeship in Thinking: Cognitive Development in Social Context. Oxford University Press, New York.

Rogoff, B., Göncü, A., Mistry, J., Mosier, C., 1993. Guided Participation in Cultural Activity by Toddlers and Caregivers, vol. 58 (8). University of Chicago, Chicago, IL.

Rogoff, B., Matusov, E., White, C., 1996. Models of teaching and learning: participation in a community of learners. In: Olson, D.R., Torrance, N. (Eds.), The-Handbook of Education and Human Development: New Models of Learning, Teaching and Schooling. Blackwell Publishers Inc, Malden, MA, pp. 388–414.

Scribner, S., 1977. Modes of thinking and ways of speaking: culture and logic reconsidered. In: Johnson-Laird, P.N., Wason, P.C. (Eds.), Thinking. Cambridge University Press, Cambridge, UK, pp. 483–500.

Sidorkin, A.M., 1999. Beyond Discourse: Education, the Self, and Dialogue. State University of New York Press, Albany, New York.

Tharp, R.G., Gallimore, R., 1988. Rousing Minds to Life: Teaching, Learning, and Schooling in Social Context. Cambridge University Press, Cambridge, New York.

Trevarthen, C., 1977. Descriptive analysis of infant communicative behavior. In: Schaffer, H.R. (Ed.), Studies in Mother-Infant Interaction: Proceedings of the Loch Lomond Symposium, Ross Priory, University of Strathclyde, September, 1975. Academic Press, London, pp. 227–270.

van der Veer, R., Valsiner, J., 1991. Understanding Vygotsky: A Quest for Synthesis. Blackwell, Oxford, UK.

Vygotsky, L.S., 1978. Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge, MA.

Vygotsky, L.S., 1997. The Collected Works of L.S. Vygotsky (M.J. Hall, Trans. vol. 4). In: Rieber, R.W. (Ed.), The History of the Development of Higher Mental Functions, vol. 4. Springer, New York.

Vygotsky, L.S., Luria, A.R., 1993. Studies on the History of Behavior: Ape, Primitive, and Child. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wegerif, R., 2007. Dialogic, Educational and Technology: Expanding the Space of Learning. Springer-Verlag, New York.

Wells, C.G., 1999. Dialogic Inquiry: Towards a Sociocultural Practice and Theory of Education. Cambridge University Press, New York.

Wenger, E., 1998. Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge, UK.

Wertsch, J.V., 1985. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge, MA.

Wood, D., Bruner, J.S., Ross, G., 1976. The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry 17, 89–100.

Vygotsky's Theory of Cognitive Development

Alex Kozulin, Feuerstein Institute, Jerusalem, Israel

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Vygotsky's theory of cognitive development is recognized as one of the most innovative psychological theories of the twentieth century. The theory is based on the assumption that culture plays a major role in cognitive development. Each period in child development is associated with a leading activity dominant in a given period. A considerable emphasis is placed on emergent cognitive functions conceptualized through the notion of the zone of proximal development. Instruction and learning are perceived as leading child's cognitive development rather than following it.

History

With wisdom of hindsight, Vygotsky's theory (VT) of cognitive development should be recognized as one of the most innovative psychological and learning theories of the twentieth century. Lev Vygotsky (1896–1934) lived a short but intellectually very intensive life in a period of war, revolution, and social upheaval in Russia. His early interests were in the field of art and literature, and he earned a PhD for his theoretical work on the psychology of art. This interest in art and humanities rather than sciences may explain Vygotsky's conceptualization of psychology as a cultural, rather than natural science. He was able to devote only 10 years (1924–34) to systematic research leading to his theory of cognitive development. Two years after his death of tuberculosis, VT together with other works in the field of applied child psychology (called 'pedology' at that time) became blacklisted by the Soviet government (Kozulin, 1990; Van der Veer and Valsiner, 1991). Only in the late 1950s was the ban lifted and Vygotsky's papers reprinted in Russia. Though some articles by Vygotsky and his colleagues, Alexander Luria and Alexei Leontiev, were published in English in the 1930s, it appears that VT made practically no impact on theories of cognitive development popular in the West at that time. The first book-length publication of Vygotsky's work in English came out in 1962 as an abridged version of *Thought and Language* (later expanded, see Vygotsky, 1934/2012). It was followed by several other translations, culminating in the six-volume set of Vygotsky's *Collected Works* (1987–98). Instead of fading VT became more and more popular with the passage of time. One possible reason is that this growing popularity is associated with the fact that VT provides some tentative answers to the questions that Western psychology and education started asking only recently. At the same time, the development and popularization of VT was seriously complicated by the early demise of its author. We will never know what Vygotsky would have said about translations, interpretations, and further developments of his theory; the author's voice is very much missing in the dialogue on VT.

In what follows, five major aspects of VT will be reviewed together with their educational applications: the sociocultural orientation of VT, the concept of mediation, periods of child development, the relationships between language and thought, and the concept of the zone of proximal development (ZPD).

Sociocultural Orientation

The distinctive feature of VT is its emphasis on culture as the most important factor of cognitive development. Though Vygotsky readily admitted that some basic cognitive processes can be shared by humans and higher animals, he explicitly and deliberately focused his own theory on those cognitive processes that are uniquely human. He called them 'higher mental processes' and associated their development with the involvement of cultural tools in the shaping of human cognition. Culture in VT is not an external envelope or ethnographically specific appearance of human behavior and thinking; culture according to VT is the force that shapes all higher mental processes, such as perception, attention, memory, and problem solving. In the absence of more appropriate terms, we still use the same verbal labels for both basic cognitive processes and culturally shaped higher mental processes though these two groups of processes are very different in their origin, development, and capacity.

Taken as a whole, the VT posed three major objectives for a study of human psychology: reconstruction of the transition from an animal to a human way of thinking and behaving; investigation of the historical change occurring in human mental functions as a result of the introduction of new cultural tools and sociocultural activities; and investigation of the developmental construction of children's and adults' psychological functions in a given society (Vygotsky and Luria, 1930/1993).

Regarding the first objective, Vygotsky relied mainly on the comparison of human behavior and the behavior of apes as reported by other researchers, such as Wolfgang Kohler (see Vygotsky, 1934/2012: pp. 73–85). Vygotsky suggested that nonhuman primates have both some intellectual problem-solving skills and communicative abilities, but that in the apes these two domains remain dissociated. Communicative abilities do not impact on problem solving, while problem solving does not shape interpersonal interaction. In a human child, intensive interaction between these two domains takes place during the second year of life. As a result, speech becomes intellectual, while problem solving acquires the quality of verbal intelligence. Thus, the transition from animal to human cognition was envisaged by Vygotsky as a change in the interaction between different cognitive functions. Vygotsky and his colleagues, however, had no opportunity to investigate this hypothesis in

 http://dx.doi.org/10.1016/B978-0-08-097086-8.23094-8

actual studies with nonhuman primates. The development of Vygotsky's line of reasoning can, however, be identified in the work of Tomasello (1999) and his colleagues. First Tomasello (1999: p. 48) reasserts the main thesis of VT regarding the cultural origin of human higher metal processes: "Following Vygotsky and many other cultural psychologists, I contend that many of the most interesting and important human cognitive achievements, such as language and mathematics, require historical time and process for their realization – even if most cognitive scientists largely ignore these historical processes." He then proceeds to show that in a number of cognitive tasks such as spatial memory, rotation of objects, and estimation of quantities, chimpanzees demonstrate performance comparable to and sometimes exceeding that of two-and-a-half year old children. At the same time children have an obvious advantage in the tasks related to gestural communication, observational learning, and understanding of intentions. Tomasello (1999: p. 213) concludes his analysis by stating that: "Language does not create new cognitive processes out of nothing, of course, but when children interact with other persons intersubjectively and adopt their communicative conventions, this social process creates a new form of cognitive representation – one that has no counterpart in other animal species."

Vygotsky should also be credited with posing an intriguing question regarding possible historical changes in human cognition. Are cognitive functions of people in antiquity, Middle Ages, and the eighteenth century the same as those of people in the twenty-first century? Do the historical changes in cultural tools impact on our cognition? In the absence of a 'time machine,' Vygotsky and Luria decided to rely on a 'quasihistorical' study of cognition in a traditional society that undergoes rapid sociocultural change. Vygotsky and Luria thought that they had found such a historical 'experiment' in the Soviet Central Asia of the early 1930s. The unique sociocultural situation of this region in the late 1920s and early 1930s was determined by a very rapid invasion of Soviet power into an otherwise traditional and mostly nonliterate agricultural society. As a result, people belonging to the same economic and sociocultural group, often even to the same extended family, found themselves under very different sociocultural circumstances. Some of them, especially those in the remote villages, retained all aspects of a traditional nonliterate culture and way of life. Others became involved in new agricultural or industrial enterprises, exposed to the new technology and means of communication, but still without access to systematic formal education. Some of the local people, however, already attended adult literacy courses and even teachers' colleges.

The main conclusions reached by Vygotsky and Luria on the basis of this study were that informants who retain a traditional nonliterate culture and way of life tend to solve problems by using functional reasoning reflecting their everyday life practical experience and reject the possibility of looking at classification, generalization, or drawing conclusions from another; for example, more abstractive point of view. Exposure to modern technology and involvement in jobs based on division of labor tend to increase the subjects' readiness to solve problems both in functional and in verbal–logical ways. It was observed, however, that informants who did not experience formal education rather easily reverted to purely functional reasoning. At the same time, informants who received some form of formal education demonstrated a clear preference for the verbal-logical form of problem solving.

Though the goal of Vygotsky and Luria was to investigate the 'historical' change in human cognition, their Central Asia research is usually interpreted as one of the first cross-cultural studies of cognitive processes. With the wisdom of hindsight, one can distinguish a number of questions that remained unanswered in this initial research. Vygotsky and Luria seem to have grouped together different sociocultural factors such as the acquisition of literacy, formal classroom learning, exposure to modern technology, and participation in labor activities based on the formal division of labor. Each of these factors seems, however, to have a different impact on the construction of cognitive functions and should be investigated separately.

The Central Asia study later inspired the Scribner and Cole (1981) research in Western Africa that demonstrated that literacy and schooling may have a differential cognitive impact. By conducting their research in an African society where literacy in three different languages was associated with different acquisition and application contexts (school, home, and religious institution), Scribner and Cole showed that literacy does not have an overall impact on problem solving but affects specific cognitive functions corresponding to each one of the contexts. Formal education on the other hand has an impact on problem solving in the tasks that resemble those used in school. The emergence of cognitive functions was thus linked to more specific sociocultural contexts and activities. More recent studies conducted in Central American Mayan villages were able to identify the transition from subsistence and agriculture to wage economy and commerce as the main factor leading children from more concrete to more abstractive cognitive representations. At the same time, in more complex tasks that required selecting a strategy for continuation of the model pattern, schooling proved to have the strongest relationship with the choice of a more abstractive and less imitative strategy, with the involvement in the 'new' economy coming second.

Mediation: Physical, Symbolic, and Human

One of the main concepts of VT is that of mediation. Vygotsky and Luria (1930/1993) distinguished three main classes of mediators relevant for the development of cognitive functions: physical tools, symbolic tools, and human mediation. Physical tools (hammer, plow, bulldozer) aim at creating a change in the environment, and, in this sense, they 'address' external natural objects. Each physical tool, however, has a 'handle' corresponding to the physical and mental requirements of a person who operates this tool. Vygotsky hypothesized that the change in physical tools may have a reciprocal effect on the human mental functions required for 'handling' new tools. For a long time, this aspect of VT reasoning received relatively little attention; it is only recently that it has started being investigated in the context of ergonomics and the psychology of human–machine interactions.

The main emphasis of VT is on symbolic tools and their internalization in the form of inner psychological tools. Through acquisition and internalization of symbolic tools, human beings start regulating and shaping their own cognitive and behavioral processes. Child development, according to VT, can be interpreted as the gradual acquisition by the child of increasingly more sophisticated symbolic tools, such as oral speech, reading and writing systems, mathematical symbols, graphic organizers, and pictorial representations. These tools then become internalized and shape the child's cognitive processes. Each culture or subculture has its own system of symbolic artifacts. If a specific tool, for example technical drawing, is unavailable in a given cultural community its members will solve the same problem in an alternative way and their cognitive function of representation will be shaped by these alternative methods rather than psychological tools associated with internalized technical drawing. The availability of a certain symbolic artifact in a given culture, however, does not guarantee its appropriation as a cognitive tool. For example, many graphic-symbolic organizers such as tables, graphs, and charts are perceived by students as external artifacts and are never transformed into their inner psychological tools.

The concept of psychological tools helps to elucidate the distinction made in VT between basic and higher mental processes. For example, the task of recalling the positions of objects on a given surface can be achieved via a basic process of direct memorization and retrieval. However, if objects are placed on a grid-like surface, one can use a culturally developed function of 'mapping' the position of an object with the help of the rows and columns of the grid. Moreover, even when the surface does not have the features of the grid, a person can imagine and project such an imaginary grid on the surface to facilitate the memorization and recall task. While direct memorization relies on a basic cognitive process, the use of an actual or imaginary grid turns the process of memorization into a higher mental activity mediated by a psychological tool. Educational programs based on VT consider acquisition and internalization of a broad range of symbolic tools as an important objective of school education. Moreover, many cognitive education and thinking skills programs developed without the direct influence of VT aim specifically at enriching the range of psychological tools available to children (Kozulin, 1998, Chapter 4).

Regarding human mediation, Vygotsky mentioned two aspects. On the one hand he suggested that every higher mental function in the child's development first appears in the interaction between people (interpsychologically), and then, when internalized, it appears as an inner mental function (intrapsychologically). As an illustration, Vygotsky cited the phenomenon originally described by J.M. Baldwin and then by J. Piaget, namely that a child's cognitive ability to consider different points of view originated in actual arguments between children. The other aspect of human mediation mentioned by Vygotsky is the role of parents and other mentors as mediators of meaning to the children. He illustrated this by describing the evolution of the child's indicatory gesture. Such a gesture, according to Vygotsky, first appears as an attempt of the child to actually grasp some remote object. Adults who witness the child's attempt interpret it as a gesture and help the child to get the object. They thus provide a new meaning to the child's grasping movement. The child then 'accepts' this interpretation, and starts addressing his/her gesture to an adult rather than a desired object. Later on children transform this gesture into a verbal request and then internalize it in a form of command to themselves.

The role of parents and other mentors as mediators was further elaborated by R. Feuerstein (see Kozulin, 1998, Chapter 3). Feuerstein identified specific parameters that turn the situation of interaction between an adult and a child into a situation of mediated learning. By using these parameters, Feuerstein was able to identify the lack of proper mediation during childhood as a source of learning and cognitive problems later in the child's life. The consequences of the lack of human mediation, however, are not irreversible. In a number of case studies, Feuerstein and his colleagues demonstrated how intensive human mediation in late childhood and adolescence may compensate for its lack in earlier childhood.

Periods of Child Development

Vygotsky and his followers in Russia, particularly Daniel Elkonin, created an original model of the periods of child development built around the concept of so-called leading activity (Karpov, 2005). This model assumes that though children are involved in different types of activities, in each period there is one activity that plays the leading role and is crucial for propelling the child's cognitive and emotional development. The above model is explicitly culture-specific and describes periods of child development in a modern industrial society that has a formal educational system. Leading activities and developmental periods in nonliterate and nonindustrial societies and communities could be different.

The periods of child development alternate between those where leading activity has an emotional-interpersonal focus and the periods with cognitive focus. It is important to remember that these are developmental, rather than age periods: some children may stay longer in some of the periods or make a quicker transition to the next period depending on their social situation of development. Thus the ages provided in Figure 1 are approximate and should not be considered as reflecting normative data.

Age (years)	Emotional-interpersonal focus	Cognitive focus
0–1	Emotional interaction with caregivers	
2–3		Object-centered joint activity
3–6	Sociodramatic play	
6–12		Formal learning
12–18	Interaction with peers	
18+		Vocational activity

Figure 1 Periods of child development.

Within each one of the periods, the development of the child's cognitive and interpersonal skills leads to the formation of a new motive that corresponds to the new leading activity in the next developmental period. Thus, for example, the emotional component is central during the earliest period in child development, while the manipulation with objects (toys) plays a subdominant role. Gradually, however, the object-centered joint activity with adults becomes a new motive of the child and at a certain moment assumes the role of the leading activity with emotional contact becoming subdominant. In a similar way, the leading activity of object-centered play gradually prepares the child for the transition to sociodramatic play: If in the first of these two periods manipulation with a toy car is focal, in the second period the imaginary role of a driver becomes the main interest of the child.

Relative to other developmental periods, the period of sociodramatic play and the period of formal learning activity received the major attention in the VT (Karpov, 2005). There are some practical reasons for this emphasis because the period of sociodramatic play is associated with children's school readiness, while the period of formal learning is associated with education in primary schools. In the VT, school readiness is associated with the following abilities developed via sociodramatic play: the ability to flexibly switch roles, enhanced self-regulation, and the development of symbolic thought. To be ready for formal learning in school the child should be able to assume a new role of a student which is different from those of son, daughter, sibling, or playmate. Play according to specific rules enhances self-regulation, while imaginative play strengthens hypothetical reasoning. In what concerns the period of formal learning, the main contribution of VT is in distinguishing between generic learning activity that may appear as an element of play, practical activity, interpersonal interaction, and so on, and a specially designed activity of formal learning that aims at turning a child into a self-regulated learner. The true target of this leading activity is the child himself/herself, who through acquisition of reflective and metacognitive skills becomes an agent of his/her own learning. To be an agent of one's own learning does not mean, however, to learn independently; one of the main goals of primary education according to VT is to teach children how to distinguish between those elements that they can learn themselves and those aspects that require the assistance of teachers and other mentors. In other words, primary school students should be able to distinguish between what they 'know and can do themselves' and what they 'can do only with the help of others.' An elaborate request for assistance is one of the main subjects of this educational approach that will be discussed in greater detail in the last section.

Language and Thought

Among the contributions made by VT to the study of language and thought probably the most important is the broadening of the scope of what is called 'child language.' When Vygotsky and Piaget started their theoretical dialogue child language was considered almost exclusively in its communicative aspect while a child's verbal reasoning was perceived as an 'incomplete' adult thought (see Vygotsky, 1934/2012). VT radically broadened this field by focusing on such phenomena as private speech, silent inner speech, and a child's written language (Wertsch, 1985). An important role here is played by a study of child speech that apparently is not directed at any listener, and which Piaget called 'egocentric.' Piaget considered such a speech to be a by-product of a child's immature egocentric thought. The destiny of the 'egocentric' speech in his opinion was to be replaced by 'socialized' speech attuned to the expectations of the listener. Vygotsky, however, interpreted such private speech as an important developmental tool leading a child toward self-regulation and voiceless verbal thinking.

More recent studies have proved that private speech has many varieties and functions (Winsler et al., 2009). For example, many children at the age of one to three systematically engage in the so-called crib speech. Crib speech takes place at bedtime when toddlers are left in their own room and parents are either not present or at least are not visible to a child. Crib speech includes retelling significant episodes of the child's experience, questions and answers, word play, and 'experimentation' with different language forms. So instead of being a by-product of the child's immaturity, crib speech proved to be an important mechanism for the consolidation of the child's experience, language practice, and acquisition of such social communication forms as dialogue. Crib speech is often much richer than the daytime communicative speech of the same child, and as such provides a glimpse into the child's future development.

Probably the most prominent and well-researched function of private speech is its support of the child's self-regulation and problem solving (Winsler et al., 2009). It has been established that the amount of private speech in 5–7-year-old children is dependent on the subjective difficulty of the task. When the task is too easy or too difficult for a child, the amount of private speech is low. However, when the task is within the child's ZPD the amount of private speech is high. Moreover, the presence of task-relevant private speech is indicative of the child's performance not only with the present but also with future tasks. First graders who made self-guiding task-relevant comments did better at math in the second grade. While private speech is important for problem solving, one expects it to be eventually internalized; i.e., become completely silent inner speech. Children who progressed more rapidly from private to inner speech also showed better sustained attention and were less distractible. Contrarily, children who have learning problems often display private speech that is not task-related (chanting, repetitions, nonguiding comments) for a longer developmental period than children without learning problems. What started as Vygotsky's response to Piaget's notion of 'egocentric' speech has now developed into a broad area of studies on the relationship between different forms of verbalization and the cognitive processes involved in executive functions, self-regulation, and problem solving.

Another important area of research triggered by Vygotsky's critical analysis of Piaget was the distinction between everyday and academic concepts. Vygotsky (1934/2012) suggested that there is a significant difference, even conflict between spontaneous concepts that children acquire through their everyday life experience (e.g., 'The sun rises in the morning') and the corresponding academic concepts essential for

scientific and technological reasoning ('What appears as a sunrise is the result of the rotation of the Earth around its axis'). What is acceptable for everyday functioning cannot serve as a basis for rigorous academic reasoning. Vygotsky and his followers alerted educators to the fact that children do not come to the classroom as a *tabula rasa* but bring with them their preexistent everyday concepts. These concepts in themselves, however, do not lead to proper scientific reasoning. Moreover, they are the basis of many of the students' 'misconceptions' that greatly complicate math and science instruction. The acquisition of academic concepts would not happen without deliberate instructional activity. Such activity should be carried out in a 'dialogue' between the teacher's systemically organized academic concepts and the students' experientially rich but often scientifically invalid spontaneous concepts.

ZPD and Dynamic Assessment

There is little doubt that ZPD became the most popular of Vygotsky's concepts, often to such an extent that in the public mind the entire VT just becomes the theory of ZPD. Quite often, however, the popularity of ZPD fails to be accompanied by a sufficiently deep understanding of this notion. One of the difficulties in 'decoding' the notion of ZPD and its relationship to such practical application as dynamic assessment stems from the fact that Vygotsky and his followers in Russia refer to ZPD in three different contexts – developmental, educational, and directly related to assessment. In the developmental context, ZPD appears as a principle that helps to explain the dynamics of child development. Vygotsky (1935/2011) argued that typical psychological study focuses only on those psychological functions that have already fully matured and as such are displayed by children in their independent activity. What is needed is the method of evaluation of those emergent psychological functions that are still 'invisible' because children cannot yet demonstrate them without assistance. VT stipulates that in joint activity with adults these functions, however, become 'visible' and can thus be studied. The notion of ZPD allows one to see not only the developmental 'present' but also the future of the child.

In the educational context, ZPD helps to explain the relationships between education and child development. Educational process according to VT should include the development of children's cognitive functions as one of its main objectives. Contrary to a popular view that instruction should be based on already fully formed cognitive functions, VT claimed that it should target those emergent functions that can be shaped by education. In this way, students will not only acquire information and specific skills but also form higher mental functions (Kozulin et al., 2003).

In relation to assessment, ZPD was operationalized as a difference between the independent performance of students and their performance under conditions of joint activity with their mentors. In this context ZPD emerges as a precursor of what today is called dynamic assessment (Haywood and Lidz, 2007). Vygotsky (1935/2011) suggested a range of means for turning the assessment process from strictly independent into assisted: presenting a complete problem-solving model, asking leading questions, starting solutions, and then asking the child to continue. What Vygotsky did not provide was a specific step-by-step manual for conducting ZPD assessment. So his contribution in this field remained paradigmatic rather than methodological.

There is still no consensus regarding practical applications of the notion of ZPD. Some authors emphasize the developmental aspect and insist on applying the concept of ZPD only to the qualitative cognitive changes that occur in the transition from one developmental period to the next one. Other authors focus on the assessment context and argue that ZPD is useful for explaining quantitative difference between children's cognitive performance and their learning potential or the degree of their cognitive modifiability (Kozulin et al., 2003).

Taking into account the current popularization of the notion of ZPD it seems important to delineate the theoretical boundaries of this concept. Not every situation that involves a mentor, a child, and a task is related to ZPD, but only those situations that reveal or promote the development of children's psychological functions. A mere transmission of content knowledge or specific skills has nothing to do with ZPD. It would be counterproductive to apply the notion of ZPD to a situation of a teacher simply helping students with specific math or reading problem without setting any cognitive-developmental goals for this activity.

The relationship between the notion of ZPD and the current practice of dynamic assessment is far from unequivocal. Not all dynamic assessment approaches are based on ZPD, and on the other hand, ZPD is not just a theory of assessment. Today various dynamic assessment tests offer a range of possible evaluation formats from intensive one-on-one learning sessions in which the mentor flexibly reacts to the student's performance to learning sequences that are tightly scripted – standardized cues or prompts appear in response to specific answers of the student (Haywood and Lidz, 2007). There is also a variety of tasks used in dynamic assessment; some of them resemble the tasks of standard intelligence tests such as matrices, number series, and analogies, while others use curricular language and math problems. What unites all these approaches and what brings them close to VT's original vision is the need to reach beyond current problem-solving performance into the processes that lead to students' development and change.

Educational Applications

Though some of Vygotsky's ideas have been applied in clinical areas (see Kozulin, 1990, Chapter 6), the main area of VT applications is educational. These applications range from those using a general VT philosophy to specially designed educational programs with their own VT-based curriculum and didactics. Some of the former are only loosely connected to VT theory and just assert the need of taking students' culture into account or focusing not only on current achievement but also on potential. The more structured of VT-based programs and curricula usually implement some of the theoretical concepts that were discussed earlier in this article: the concept of leading activity; internalization of symbolic tools as inner psychological tools; the difference between everyday and

academic concepts; and ZPD as a basis for both assessment and educational action.

For children at kindergarten age, the leading activity is that of sociodramatic play. The VT research identified the following outcomes of this leading activity that contribute significantly to school readiness: development of the function of symbolization both in the verbal and nonverbal sphere that serves as a prerequisite for the acquisition of literacy and other symbolic systems used in school; flexible change of social play roles that helps children to accept their new role as students and reduces their cognitive 'egocentrism'; play by rules develops self-regulation and self-control essential for classroom activities (Karpov, 2005). School readiness is thus interpreted as a development of a broad range of cognitive functions and motives, rather than specific school-related skills. The preschool programs based on the VT model have been developed both in Russia and other countries (Bodrova and Leong, 2007) and have proved to be successful in enhancing children's school readiness and improving their executive cognitive functions.

Probably the best known educational application of VT is the primary school curriculum developed by Vygotskian scholars and practitioners in Russia (Davydov, 2011). The curriculum and its didactics is based on the assumption that children of a primary school age are particularly sensitive to formal learning as a leading activity. The main emphasis is put on the development of students' self-reflection and metacognition using a wide range of symbolic tools. For example, the VT-based math curriculum in the early grades focuses more on general cognitive strategies and skills rather than numerical operations. Numbers are introduced as one of the possible ways to represent the relationships between a selected 'measure' and an object of measurement rather than the operation of counting objects. Various symbolic representations and elements of algebra are introduced much earlier than in other curricular systems. Research has confirmed that students who studied math in primary school according to Davydov's curriculum performed much better with more complex PISA tasks than their peers from elite schools that used a more traditional curriculum (Zuckerman, 2004). The English version of Davydov's math curriculum was also used in a number of schools in the United States (see Kozulin et al., 2003, Chapter 12).

Even greater emphasis on symbolic tools can be seen in the program developed by Kinard and Kozulin (2008) for teaching mathematics in socioculturally heterogeneous classrooms. The authors argued that much of the students' failure in mathematics stems not from a lack of mathematical knowledge, but from underdevelopment of general cognitive skills on the one hand, and insufficient emphasis on mathematical conceptual reasoning, on the other. Their program aims at developing students' cognitive functions and problem-solving skills through appropriation and internalization first of general and then of mathematically specific symbolic tools. Codes, tables, and diagrams are examples of general symbolic tools, while number line, formulae, and coordinate system are mathematically specific tools. This is achieved by the inclusion of special cognitive tasks into the mathematics curriculum and by emphasizing the instrumental role of such mathematical tools as graphs and formulae.

VT curricula also emphasize the difference between everyday and academic concepts. Instead of trying to enrich students' preexistent spontaneous concepts, these curricula aim at developing theoretical concepts that in many respects are opposite to the notions spontaneously formed by students (Davydov, 2011). What unites all curricula based on VT is their emphasis on education as an 'engine' of students' cognitive development. Instead of perceiving the curriculum as an informational content that should be given to students who already have all the prerequisite cognitive functions, the VT-based curriculum aims at actually forming these functions.

The educational application of the concept of ZPD includes a closer integration of assessment and learning processes. Typical educational intervention based on the VT model starts with the evaluation of students' performance in a target learning area (e.g., proportional relationships): some students in a given classroom demonstrate that they have already mastered these relationships, some that they are within their ZPD, and others that they are still beyond their ZPD. This initial evaluation helps to formulate the instructional strategy for students belonging to the three subgroups. After a certain period of intervention, the evaluation is repeated and the cycle of assessment and instructions continues (Zuckerman, 2004). In some respects, this approach has a certain affinity with 'formative assessment,' though the latter usually focuses on curricular content and skills rather than on students' cognitive development.

Theoretical Impact of VT

The continuing and ever growing interest in VT might be related to the fact that one can find in it some tentative answers to the questions that emerged in psychological and educational theory only recently. Through its notion of psychological tools, VT helps to operationalize cognitive differences observed in multicultural classrooms. The distinction between everyday and academic concepts allows us to better understand students' 'misconceptions' and explains the lack of success of educational approaches built around children's everyday experiences. The notion of ZPD has not only paved the way for dynamic assessment but has also helped to turn cognitive development into one of the objectives embedded into almost every one of the national curricula. At the same time, the potential of VT is far from exhausted; many of the VT-based educational applications still remain at the pilot stage awaiting more comprehensive research, development, and implementation.

See also: Bruner's Theory of Cognitive Development; Cognitive Development, Educational Theories of; Cognitive Development: Child Education; Cognitive Development: Mathematics Learning and Instruction; Culture, Cognition and Embodiment; Human Cognition, Evolution of; Implicit Social Cognition; Neo-Piagetian Theories of Cognitive Development; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Piaget, Jean (1896–1980); Vygotsky's Sociocultural Theory; Vygotsky's Theory of Human Development and New Approaches to Education; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Bodrova, E., Leong, D.J., 2007. Tools of the Mind: the Vygotskian Approach to Early Childhood Education, second ed. Merrill/Prentice Hall, Columbus, OH.

Davydov, V.V., 2011. Problems of Developmental Instruction. Nova Publishers, Hauppauge, NY.

Haywood, C., Lidz, C., 2007. Dynamic Assessment in Practice. Cambridge University Press, New York.

Karpov, Y., 2005. The Neo-Vygotskian Approach to Child Development. Cambridge University Press, New York.

Kinard, J., Kozulin, A., 2008. Rigorous Mathematical Thinking: Conceptual Formation in Mathematics Classroom. Cambridge University Press, New York.

Kozulin, A., 1990. Vygotsky's Psychology: A Biography of Ideas. Harvard University Press, Cambridge, MA.

Kozulin, A., 1998. Psychological Tools: A Sociocultural Approach to Education. Harvard University Press, Cambridge, MA.

Kozulin, A., Gindis, B., Ageev, V., Miller, S. (Eds.), 2003. Vygotsky's Educational Theory in Cultural Context. Cambridge University Press, New York.

Scribner, S., Cole, M., 1981. Psychology of Literacy. Harvard University Press, Cambridge, MA.

Tomasello, M., 1999. The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, MA.

Van der Veer, R., Valsiner, J., 1991. Understanding Vygotsky. Basil Blackwell, Oxford, UK.

Vygotsky, L., 1934/2012. Thought and Language, rev. & exp. ed. MIT Press, Cambridge, MA.

Vygotsky, L., 1935/2011. The dynamics of the schoolchild's mental development in relation to teaching and learning. Journal of Cognitive Education and Psychology 10, 198–211.

Vygotsky, L., 1987–1998. In: Rieber, R.W. (Ed.), The Collected Works of L.S. Vygotsky, vols. 1–6. Plenum, New York.

Vygotsky, L., Luria, A., 1930/1993. Studies on the History of Behavior. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wertsch, J., 1985. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge, MA.

Winsler, A., Fernyhough, C., Montero, I. (Eds.), 2009. Private Speech, Executive Functioning, and the Development of Verbal Self-Regulation. Cambridge University Press, New York.

Zuckerman, G., 2004. Development of reflection through learning activity. European Journal of Psychology of Education 19, 9–18.

Vygotsky's Sociocultural Theory

Natalia Gajdamaschko, Simon Fraser University, Vancouver, BC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Lev Vygotsky created a theory of psychological development that has become increasingly influential in recent decades. The theory is viewed by some psychologists and pedagogues as alternative way to understand the development of children and adults in a way that differs substantially from the dominant psychological models that have long been widely accepted, including Piaget.

Vygotskian psychology includes many important concepts that are not found in, or at last are not central to, other psychological theories. Among the most important concepts are (1) the existence of higher psychological functions that differ from lower psychological functions, (2) the mediated nature of higher psychological functions, (3) the existence of internal and external tools that transform the structure of mental function, (4) the central importance of culture and history, as a process of becoming, in development, (5) zone of proximal development.

Overview of Vygotsky's Education and Career

Lev Semenovich Vygotsky was born in the town of Orsha, Vitebskaya Oblast (former Mogilevskaya Gubernya), Belorussia on 5 November 1896 and grew up in Gomel, a larger Belorussian city about 150 miles from his birthplace. There as a young child, he was homeschooled by his father, Semen L'vovich, a banker, and his mother, Cecilia, a teacher by training, plus some tutors hired to assist them. After 5 years of schooling at home, he attended a private single-sex secondary school. As a gifted student, he excelled in all subjects, but was particularly interested in literature and philosophy, subjects for which he maintained an intense interest throughout his life. According to Gita Lvovna Vygodskaya (1925–2010), his daughter and biographer who meticulously cataloged his early life and family history, Lev Vygotsky's interest in arts and humanities was due mainly to the influence of his family, one of the best educated in Gomel.

He graduated in 1913 from secondary school with the highest honors, but his hopes to study philology, his favorite subject, at a university and to become a public school teacher were dashed on the rocks of Tsarist Russia's anti-Semitism. Such a path was not open to Jews in prerevolutionary Russia. Instead, heeding the advice of his parents, Vygotsky decided to study medicine, a more practical profession: As a medical doctor, he could choose to live outside the pale, the boundaries of Jewish settlements in Tsarist Russia. He was admitted to the medical school of Moscow University, but after only a month studying medicine, he transferred to the school of law. A year later, in 1914, while continuing to study law, he, because of his deep interest in literature, art, and philosophy, enrolled in the history-philosophy department of Shanyavsky University. Soon after that, he added psychology to his course of study there.

Vygotsky's first paper, written during his student years, addressed problems of literary criticism. A student graduation paper was titled, 'The Tragedy of Hamlet, Prince of Denmark.' Though not published at the time it was written, it was rediscovered and published 52 years later as a part of second edition of *The Psychology of Art*, in 1968. According to Gita Vygodskaya, the essay was recognized for its originality and unique analysis of Hamlet (Vygodskaya and Lifanova, 1996).

In 1917, Vygotsky graduated from the two Moscow universities and moved back to Gomel to join his family, which had not been spared the harsh impacts of World War I and which would suffer more hardships during the Great October Socialist Revolution and the subsequent civil war. During these times, two of his brothers died.

Though this period of Vygotsky's life is often overlooked by researchers, according to Gita Vygodskaya (1996) Vygodskaya and Lifanova, 1996, the 7 years Vygotsky spent in Gomel after he returned in 1917 were important for his professional and personal development. During this time, he taught at the local pedagogical college, managed a theater, and published a journal of literary criticism. Most importantly, he created his first laboratory in Gomel and began serious study of the psychology of art and the psychology of education. This work led directly to his first major book on psychology, *The Psychology of Art*, which was completed in 1925, but first published in 1965. A second book, *Pedagogical Psychology* was published in 1926, but was later withdrawn from libraries and banned by Communist authorities for more than 60 years. It was republished at the end of the 1980s. While Vygotsky's early writings are often difficult to reconcile with what he wrote during the decade after he left Gomel, they provide researchers with insights into his intellectual development.

In 1924, Vygotsky moved from Gomel to begin an association with the Moscow Institute of Experimental Psychology that would last until his death. He was invited to join the Institute after he made an impressive presentation at the Second Psychoneurological Congress in Leningrad. As a young, unknown researcher who had been working in obscurity far from the center of Russian intellectual life, Vygotsky's presentation so captivated future colleagues, such as Alexander Romanovich Luria, that they wanted him to join them in their work (Cole et al., 2005: p. 39).

At the Institute, Vygotsky, together with his close friends and colleagues Luria and Aleksei Nikolaevich Leont'ev, set out to build a new approach to the study of human development. Their goal was to help create a Soviet psychology by revising and reforming psychology along Marxist lines and building a new, objective materialistic approach to psychology. Their work was influenced by the profound social changes that

followed the Communist revolution, which created a new context for their research. They sought to address the practical needs of the newly formed educational, clinical, and academic systems in the Soviet Union.

As he began his work at the Institute, Vygotsky believed that psychology was in a crisis, divided into two camps that represented different scientific schools that had little interest in mutual understanding. One camp was objective psychology, built on the principles of natural sciences. The leading Russian members of this school were I. M. Sechenov and Ivan Pavlov, who studied reflexes. The second camp included psychologists with phenomenological, idealistic, and subjective orientations.

In response to this perceived crisis, Vygotsky, already ill with the tuberculosis that would kill him, wrote while in a hospital *The Historical Meaning of the Crisis in Psychology*, an attempt to unite the mutually contradictory objective and subjective schools of thought. Although the book was finished in 1927, its publication was not allowed by the Soviet government until 1982. Despite his inability to get the book published, Vygotsky and his colleagues widely spread the ideas in it through their lectures and papers, and his theoretical analysis led to the development of the cultural–historical theory of human development. The elements of this theory are set out in more than 270 articles, book, and monographs he published between 1924 and 1934, the year of his death.

This decade of feverish activity was greatly affected by two developments: the rise of Stalinism and Vygotsky's terminal illness. The increasingly repressive political climate in Moscow caused Vygotsky and his colleagues to relocate to Kharkov to avoid the intense political scrutiny of communist ideologues in Moscow. As a result of the Institute's relocation, Vygotsky spent much time during his last 7 years in trains, traveling between Moscow, Kharkov, and Leningrad.

Vygotsky undertook this travel and kept such an intensive work schedule despite, or perhaps because of, the fact that he knew that tuberculosis would shorten his life. He died in 1934 at the age of 37. His death spared him from the possibility that he might be punished by the Soviet regime as were other leading intellectuals. Although it is impossible to know what would have happened to him if he had lived longer, his likely fate may be indicated by the treatment of his research: his books and papers were banned in the Soviet Union until 1956 as a result of the special decree of Central Committee of Communist Party.

Two years after Vygotsky death, the special decree of Central Committee of Communist Party was issued about pedology on 4 July 1936 (Постановление ЦК ВКП(б) О педалогический извращениях в система накромпросов) was a definitive ruling on what was considered as the orthodox, and thus only acceptable view of the Soviet authorities on the topic. Vygotsky's work was deemed heterodox and thus suffered the fate of being forbidden to be published and republished, withdrawn from libraries and partially destroyed (Vygodskaya and Lifanova, 1996).

Thanks to his family (especially his daughter Gita), colleagues, and friends, most of his written work was preserved in storage and many of his ideas were passed along orally, so that the Stalin regime was not successful in obliterating his contributions to psychology. In fact, Vygotsky is now viewed by many as one of the most influential figures of twentieth-century psychology and pedagogy. Remarkably, the books and articles he wrote more than 80 years ago continue to inspire new researchers and to influence practitioners.

The Essentials of Vygotskian Psychology

With his colleagues Luria and Leont'ev, Vygotsky created a theory of psychological development that has become increasingly influential in recent decades. The theory is viewed by some psychologists and pedagogues as alternative way to understand the development of children and adults in a way that differs substantially from the dominant psychological models that have long been widely accepted, including Piaget.

Because Vygotskian theory offers a substantially different understanding of psychological development, the theory suggests different methods of psychological research and new approaches to pedagogy. Educators who believe that Vygotskian theory accurately reflects the process of development have developed approaches to teaching that differ from those based on the traditional theories underlying pedagogy in the West.

Vygotskian psychology includes many important concepts that are not found in, or at last are not central to, other psychological theories. Among the most important concepts are (1) the existence of higher psychological functions that differ from lower psychological functions, (2) the mediated nature of higher psychological functions, (3) the existence of internal and external tools that transform the structure of mental function, (4) the central importance of culture and history, as a process of becoming, in development, (5) zone of proximal development.

These core ideas of Vygotskian theory, plus other aspects of his theory, are discussed in more detail below.

A New 'Instrumental,' 'Cultural,' and 'Historical' Approach to Psychology

According to Cole et al. (2005), Vygotsky initially liked to describe the new theory of human development is as 'instrumental,' 'cultural,' and 'historical' psychology (pp. 44–45). Each term highlighted a different feature of his new approach to psychology.

The term *instrumental* reflected a fundamental idea that higher psychological functions, that is, cultural functions, are mediated. Unlike basic functions such as reflexes that can be analyzed as simple stimulus–response interactions, cultural functions incorporate in their structure new elements – external and internal tools – that transform the whole framework of mental functioning. External and internal tools are *instruments* that individuals use to modify and master their own behavior. The analysis of these tools became an essential element of Vygotsky's new approach to understanding human development. The method of analyzing tools is known as the double-simulation method.

The different characteristics of basic or natural psychological functions and higher psychological functions are shown in **Table 1**.

The term *cultural* was used to describe Vygotsky's theory because cultural development is a unique direction in the development of a child. Cultural development is concerned

Table 1 Differences in natural psychological functions and higher psychological functions

	Natural psychological functions	*Higher psychological functions*
Origin	Genetically inherited	Socially acquired
Structure	Unmediated	Instrumental, mediated by social means
Functioning	Involuntary	Voluntarily coconstructed and controlled
Relation to other mental functions elements	Isolated	Broad system of functions, not separate

with the 'social situation of development.' It reflects the socially constructed ways that society organizes various types of tasks faced by a growing child. It also refers to the physical and mental tools that society provides to a young child to master those tasks.

Closely connected to the *cultural* aspect of Vygotsky's theory is the *historical* aspect, which is a focus on studying development in the process of change. It incorporates Hegel's dialectic concept that things are constantly changing due to the conflicts and tensions inherent in their contradictory elements. Thus, for Vygotsky, development is constant transformation: it is never stable; rather it is a continual process of becoming. This dynamic concept of history differs greatly from the understanding of history in the West. For them, to study something historically means, by definition, to study past events. Thus, they differentiate between the study of the history of behavior and the study of present-day behavior. For Vygotsky, the two are inseparable in that they are linked together in the process of change that defines development. The present state of a person's development is unknowable without knowing the history of his or her development.

Development by Changing the Interfunctional Structure of Consciousness

Indicated by the importance of the instrumental, cultural, and historical elements of Vygotsky's theory, psychological functions develop throughout childhood by the acquisition of cultural (psychological) tools created by society and by the 'process of becoming' as psychological functions are restructured within the culture. In addition, Vygotsky suggested that development comes by changing the interfunctional structure of consciousness. In this regard, he criticized models that ignore the interdependence of psychological processes and the methods designed to study only individual functions in isolation. Instead of viewing the process of development as a structural change of a single psychological function, Vygotsky suggested that psychology should investigate the changing relationships between psychological functions and their developmental changes. He wrote:

> In the process of development, it is not so much the functions which change (these we mistakenly studied before). Their structure and the system of their development remain the same. What is changed and modified are rather the relationships, the links between the functions.
>
> Vygotsky L.S 1930/1987 On psychological systems, vol. 3, p. 92

Departing from the Old Theories: Development as Cultural Development

Vygotsky's research led him to conclude that the traditional psychology of development could not handle the task of researching cultural development properly because "child psychology did not know the problem of higher mental functions or, what is the same, the problem of the cultural development of the child. For this reason, the central and greater problem of all psychology, the problem of personality and its development, still remains closed" (vol. 4, p. 26). He continued:

> Only a decisive departure beyond the methodological limits of traditional child psychology can bring us to a study of the development of that same higher mental synthesis that, on a solid basis, must be called the personality of the child. The history of cultural development of the child brings us to the history of development of personality (vol. 4, p. 26).

He concluded by observing that many psychologists still analyzed changes in behavior as resulting from changes in natural or basic psychological functions rather than changes in culture. He wrote, "[They] are inclined to … think of [changes in behavior] as facts of habit formation or as intellectual reactions directed toward cultural content" (1997b; vol. 4, p. 92).

Vygotsky's call for a broader study of a child's context stood in stark contrast to theories advocated in the West. The contrast between his own theory and Western thought is exemplified by Vygotsky in his critique of Piagetian theory. Vygotsky believed that Piaget focused too narrowly on identifying both the internal and external sources of a child's development and studying them in isolation. This approach, Vygotsky argued, relied too much on reductionism and ignored culture and history in a child's development, as well as the child itself; that is, the impact of the child's personality on the child's development. He wrote:

> What … is missing in Piaget's perspective is reality and the child's relationship to that reality. What is missing is the child's practical activity. This is fundamental. Even the socialization of the child's thinking is analyzed by Piaget outside the context of practice.
>
> Vygotsky, 1987; vol. 1, p 87

Unlike Piaget, Vygotsky placed the child at the center of his theory and accounted for the mediated nature of cultural development. In the process of development, a child will 'arm and rearm' himself with widely varying tools. Only the acquisition of cultural tools allows the transformation from lower natural functions to higher (cultural) functions:

> If we seriously consider the fact that with the knot tied for remembering, man, in essence, constructs externally a process for remembering, the external object compels him to remember, that is, he

> reminds himself through an external object, and, in this way carries out a process of remembering as if externally, converting it to external activity; if we consider the essence of what occurs here, this one fact can disclose for us all the profound uniqueness of higher forms of behavior.
>
> Vygotsky, 1997; vol. 4, p. 59

> The very essence of human memory consists of man actively remembering with the help of signs. In general, the following might be said about human behavior: in the first place, his individuality is due to the fact that man actively participates in his relationships with the environment and through the environment he himself changes his behavior, subjecting it to his control.
>
> Vygotsky, 1997; vol. 4, p. 59

Warning against a mechanical understanding of the role of 'psychological tools' in a child's activity, Vygotsky explained that not everything could be a tool. If something "did not have the capacity to influence behavior, it could not be a tool" (Vygotsky; vol. 3, p. 87). Following the Hegelian idea that cultural tools, internalized, empower a child in its development, Vygotsky considered the mediated nature of psychological functions to be one of the most important mechanisms of development. A function that a child consciously masters acquires other possibilities of action.

Another aspect of Vygotsky's views on child development that is intimately connected to understanding context in a new way is his idea of development by means of changes in psychological systems. The structural and functional characteristics of cultural functions change not only because new cultural tools are acquired, but also because they enter into a new relationship with each other at every stage of development. Applying this notion to development of different psychological functions, Vygotsky pointed out that the context of development changes because psychological functions are in a constant and ever-changing relationship during the process of a child's development. In this regard, Vygotsky wrote that one of the most fundamental idea of developmental psychology is that there can be no fixed relationship between thinking and speech, for example. As he illustrated in his volume *Thinking and Speech*, at each stage of development we observe a unique and changing set of relationship.

Extending the same idea to the evolving development of imagination, for example, Vygotsky wrote:

> Imagination does not develop all at once, but very slowly and gradually evolves from more elementary and simpler forms into more complex ones. At each stage of development it has its own expression; each stage of childhood has its own characteristic form of creation. Furthermore, it does not occupy a separate place in human behavior, but depends directly on other forms of human activity, especially accrual of experience.
>
> What is substantially new in the development of fantasy during the transitional age is contained precisely in the fact that the imagination of an adolescent enters a close connection with thinking in concepts; it is intellectualized and included in the system of intellectual activity and begins to fulfill a completely new function in the new structure of the adolescent's personality.
>
> Vygotsky, 1998; vol. 5, p. 154

This complex view of the evolving development of thinking, speech, and imagination contrasts with the belief that intellectual change during adolescence is "a simple quantitative accumulation of characteristics already laid down in the thinking of a three year old." According to Vygotsky, the word 'development' does not apply to this explanation of intellectual change (Vygotsky, 1998; vol 5, p. 29).

Vygotsky formulated the first law of the development and structure of higher mental functions, which can be called "the law of the transition from direct, innate, natural forms and methods of behavior to mediated, artificial mental functions that develop in the process of cultural development" (vol. 5, pp. 167–168). He wrote that every function in the child's cultural development appears twice: first on the social level and, later, on the individual level. This applies equally to voluntary attention, to logical memory, and to the formation of concepts. All the higher functions originate as actual relations between human individuals.

The Social Situation of Development

To help understand the dynamics of child development, Vygotsky focused on the social situation of development, by which he meant "the relations between the personality of the child and his social environment at each age level" (Vygotsky, 1998; vol. 5, p. 198). El'konin (1998 as cited in Vygotsky, 1998; vol. 5, p. 312) observed that "For Vygotsky, the central point in considering the dynamic of mental development at this or any period in the life of a child was the analysis of the social situation of development." This concept helped Vygotsky capture the unity of the context and the personality of a child as one unit of analysis.

According to Vygotsky, at the beginning of each age period, an original, exclusive, and unique relation develops between the child and reality (mainly the social reality that surrounds him). He explained, "The social situation of development represents the initial moment for all dynamic changes that occur in development during the given period. It determines wholly and complete[ly] the forms and the path along which the social becomes individual" (Vygotsky, 1998; vol. 5, p. 198).

Based on his notion of the social situation of development, Vygotsky suggested three important elements of the dialectical relationship between context and development.

1. The social situation of development is dynamic. He wrote: "The very initial and essential point in the general determination of the dynamics of age is to understand that the relations between personality of the child and his social environment at each level is mobile" (Vygotsky, 1998; vol. 5. p. 198).
2. The idea of 'crisis in development' should be used to help understand the dynamic and mobile nature of the role of context in development. Context must be treated dynamically and not statically, and Vygotsky proposed to focus on 'crisis in development' as a means to insure that the dynamic nature of context is at the center of its analysis. According to Vygotsky, even if crises of development were not discovered empirically or not noticed in practice, they should nevertheless be 'invented' theoretically.

The context of development, in Vygotsky's view, should follow the Hegelian understanding of uneven, revolutionary

development that includes some stable stages that are replaced by revolutionary jumps. Vygotsky explained the crisis dynamic as follows:

> The crisis of the newborn separates the embryonic period of development from infancy. The one-year crisis separates infancy from early childhood. The crisis at age three is a transition from early childhood to preschool age. The crisis at age seven is a link that joins pre-school and school ages. Finally, the crisis at age thirteen coincides with the turning point in development at the transition from school age to puberty …. Thus, [c]ritical periods alternate [with] stable periods and are turning points in development, once again confirming that the development of the child is a dialectical process in which transition from one stage to another is accomplished not along an evolutionary, but along a revolutionary path.
>
> Vygotsky, 1998; vol. 5, p. 193

3. Attention should be devoted to understanding the changing nature of development during both stable and critical periods of development. Vygotsky believed that following the idea that each stage of development has a unique set of characteristics, the differential study of critical and stable periods would bring us closer to understanding the nature of development. As El'konin (p. 312) wrote, "According to Vygotsky's thinking, the disintegration of the old and appearance of bases of the new social situation of development is the main content of the critical age levels." Unfortunately, this line of Vygotsky's research was not developed much by his followers.

In summary, the concept of the social situation of development helped Vygotsky capture the dynamism of context by pointing out that the social situation of development is unique for each stage of development. It also allowed him to uncover the mechanisms of development by identifying possible contradictions within the social situation of development that would bring in necessary changes in psychological systems of developing cultural function.

In this way, crises help to uncover the inner logic, the essence of development, by highlighting and bringing forward the contradictions between the personality of a child and the context of development that are their source.

Perezhivanie as a Unit of Analysis

In building his analysis, Vygotsky introduced 'perezhivanie' as the 'unit of analysis.' *Perezhivanie* can be translated as experience or experiencing.

Vygotsky advocated the studying complex *perezhivanie* as a product of cultural–historical development mediated by 'cultural tools/artifacts' that are social in origin and are interiorized/appropriated by the child within a particular cultural tradition.

Realization of the sociocultural approach in the study of *perezhivanie* presumes analysis of at least three interrelated questions: What are the specific cultural tools of *perezhivanie*? What are the specifics of the process of their appropriation by a person? What is role of other people in this appropriation and *perezhivanie* of the person? Vygotsky points out that such a study of child's relation to the context of development differs drastically from the phenomenological attempts to study of emotional experience because it is based on the different methodological principles. Perezhivanie as a unit of analysis is closely related to Vygotsky's notion of systemic and semantic structure of child's consciousness.

Vygotsky argued that in order to understand the development of a child's consciousness as a unity of communication and generalization that comes to a child through concept development. In *Thinking and Speech*, Vygotsky stressed that although the objective meaning of a concept could be the origin of conceptual development of a child, in reality, it is impossible to have direct and unmediated teaching of the objective meaning of a concept. Behind the acquisition/appropriation of the culturally given meaning, there is a whole world of personal sense that belongs to the personality of the child. It is not possible to produce directly the internalization of the meaning of the world; it can come only through the layers of mediation and a sense of the meaning in the world of the child.

Conclusion

In his tragically shortened life, Lev Vygotsky, with his colleagues, formulated a complex theory of child development that has in the past couple of decades provided an alternative to the long-prevailing models in the West, stimulated new research of core issues in child development, and inspired teachers and pedagogues in countries throughout the world to try new ways of teaching. Long after it was created, Vygotskian psychology has taken its place among the major theories of how children think and learn.

Elements of this theory have found their way into psychology and pedagogy courses taught at major universities in the Western hemisphere, and it continues to be a major intellectual influence in Russia and other countries of the former Soviet Union. A center for Vygotskian studies is located at the Vygotsky Institute, Russian State University for the Humanities, in Moscow, founded by his daughter and granddaughter, and now headed by a granddaughter, Dr Elena Kravtsova. This center holds an annual conference on research related to Vygotsky. Its Web site is http://psychology.rsuh.ru/.

In the West, several groups explore and interpret Vygotskian psychology. Most noteworthy is a circle created by Prof. Mike Cole and his Laboratory of Comparative Human Cognition at the University of California, San Diego. The group is called XMCA. It was started initially as a listserv for the discussion of Vygotskian psychology, but through the years it has become a key place for Vygotskian scholars throughout the world to collaborate and interact with each other.

In addition to XCMA, the Laboratory of Comparative Human Cognition, under Prof. Coles' leadership, also publishes a quarterly journal, *Mind, Culture, and Activity*, that is devoted largely to Vygotskian-related research. See the following Web site for more information about the Vygotskian discussion list and journal: lchc.ucsd.edu.

Another group addressing Vygotskian psychology is a Special Interest Group (SIG) of the American Education Research Association (AERA). The Vygotskian SIG meets annually at the AERA meeting, organizing different paper

sessions and roundtable discussions, plus it provides professional development opportunities for graduate students and younger faculty members. More information about this group is available at this site: http://www.aera.net/SIG030/CulturalHistoricalResearchSIG30/tabid/13234/Default.aspx.

In all, Vygotsky and his ideas have continuing credibility and influence, offering an alternative way to understand child development and to provide a means for more effective teaching. The future of his ideas depends, to a large extent, on continued research by scholars to test his theory to determine the validity and value of its key ideas.

See also: Bruner's Theory of Cognitive Development; Cognitive Development, Educational Theories of; Concept Formation in Childhood: Cross-Cultural Perspectives; First Language Acquisition, Developmental Psychology of; Language Development, Theories of; Neo-Piagetian Theories of Cognitive Development; Neo-Vygotskian Developmental Theory; Piaget's Theory of Cognitive Development; Pretend Play and Cognitive Development; Self in Culture: Early Development; Self-Development in Childhood and Adolescence; Vygotsky's Theory of Cognitive Development; Vygotsky's Theory of Human Development and New Approaches to Education; Vygotsky, Lev Semenovic (1896–1934).

Bibliography

Cole, Michael, Levitin, Karl, Luria, Alexander, 2005. Autobiography of Alexander Luria: A Dialogue with the Making of Mind. Lawrence Erlbaum Associates, New Jersey, London.

Rieber, Robert W., Carton, Aaron S. (Eds.), 1987. The Collected Works of L.S. Vygotsky. Problems of General Psychology, Including the Volume Thinking and Speech, vol. 1.

Vygodskaya, G.L., Lifanova, T.M., 1996. Lev Semenovich Vygotsky: Zhizn, Deyatel'nost, Shtrihi k portretu. Smusl Publisher, Moscow (in Russian).

Vygodskaya, G.L., Lifanova, T.M., 1999. Lev Semenovich Vygotsky. Journal of Russian & East European Psychology 38, 3–94.

Vygotsky, L.S., 1993. In: Rieber, Robert W., Carton, Aaron S. (Eds.), The Collected Works of L.S. Vygotsky, The Fundamentals of Defectology (Abnormal Psychology and Learning Disabilities, vol. 2.

Vygotsky, L.S., 1997a. In: Rieber, Robert W., Wollock, Jeffrey (Eds.), The Collected Works of L.S. Vygotsky, Problems of the Theory and History of Psychology, vol. 3.

Vygotsky, L.S., 1997b. In: W Rieber, Robert (Ed.), The Collected Works, The History of Development of Higher Psychological Functions, vol. 4. Plenum Publishers, New York.

Vygotsky, L.S., 1997c. In: Rieber, Robert W., Hall, Marie J. (Eds.), The Collected Works of L.S. Vygotsky, Scientific Legacy, vol. 6.

Vygotsky, L.S., 1998. In: Rieber, Robert W. (Ed.), The Collected Works, Child Psychology, vol. 5. Plenum Publishers, New York.

Yaroshevsky, M., 1989. Lev Vygotsky. Progress Publisher.

Relevant Websites

Laboratory of Comparative Human Cognition at the University of California, San Diego. lchc.ucsd.edu (accessed 31.12.13.).

Vygotskian Special Interest Group (SIG) at the American Educational Research Association (AERA). http://www.aera.net/SIG030/CulturalHistoricalResearchSIG30/tabid/13234/Default.aspx (accessed 31.12.13.).

Vygotsky Institute of Psychology in Moscow. http://psychology.rsuh.ru/ (accessed 31.12.13.).

W

Wage Differential and Structure

Robert Drago, Precision Numerics, Springfield, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Johnson, volume 24, pp. 16345–16350, © 2001, Elsevier Ltd.

Abstract

This article discusses five major sources of wage differentials. These include the *skills* required by jobs, such as complexity or investments in human capital; *job location*, which can yield differences for urban as opposed to rural locations; the effects of *job characteristics*, including some jobs involving dangerous conditions; *discrimination or nepotism* related to gender, race/ethnicity or family status; and *rents* due to labor unions or governmental policies such as a minimum wage.

At a point in time in a particular economy, there will be a distribution of wage rates (average total compensation per hour) across all economically active members of the population. This distribution of wages is referred to as the wage structure of the economy. The value of the wage rate of each member of the labor force will depend systematically on a set of observable characteristics of that person as well as on other characteristics that are not observed. This relation between wages and worker characteristics, as well as the distribution of worker characteristics, determines the wage structure of the economy. The average wage of a person with a particular characteristic relative to the wage of a person with a different characteristic (e.g., the wage of a worker in Paris relative to one in Lyon), all other observed and unobserved characteristics held constant (*ceteris paribus*), is called a wage differential. This article discusses the effects of different worker characteristics on wage structure.

Earnings and Wages

There are several different ways of defining the wage rate, and the appropriateness of any particular definition depends on the questions in which one is interested. Table 1 shows the distribution of average hourly gross (pretax) wage rates of a random sample of workers in the United States in 2011. In principle, one would like to include nonwage compensation (employer contributions for fringe benefits and social insurance programs) but such data usually must be collected from employers rather than from individuals. A worker at the 20th percentile of the distribution earned $9.86 per hour, which means that 20% of workers earned less than this amount. One way to summarize the wage structure is to compare the average wage in the highest quintile (from the 80th percentile up) to the average wage in the lowest quintile (from the 20th percentile down).

These estimated averages for the United States in 2011 were, respectively, $28.80 and $9.86, a ratio of 2.92. An alternative to

Table 1 Estimated distribution of hourly wages of full-time workers in the United States in 2011

Percentile	*Hourly wage*
20	$9.86
40	$13.88
50	$16.07
60	$19.03
80	$28.80

Lawrence Mishel, Josh Bivens, Elise Gould, Heidi Shierholz, 2012. The State of Working America, twelfth ed. Cornell University Press, Ithaca, NY, Table 2.5.

Table 2 Ratios of top quintile's share of income to bottom quintile's share for selected countries

Country	*Ratio*
United States	8.5
Canada	5.5
United Kingdom	4.0
France	6.3
Germany	5.4
Italy	4.9
Czech Republic	3.1
Ukraine	3.7
Japan	7.1
China	5.2
Hong Kong	9.5
India	5.0
Egypt	6.9
Morocco	7.0
Colombia	13.9
Brazil	23.1
Kenya	18.2
South Africa	32.1
Cote d'Ivoire	7.2

Deininger, K., Squire, L., 1996. A new data set measuring income inequality. The World Bank Economic Review 10 (1), 565–591, Table 1.

the use of hourly wage rates is the use of average weekly, monthly, or annual earnings of individuals. Total earnings per year for an individual are by definition equal to the average wage times hours per year. Data on individual incomes rather than hourly wages are often more readily available in most countries. An example of this is reported in Table 2 in which the ratio of the average annual income of workers in the top quintile to that in the lowest quintile is reported for a selection of countries. The study from which this study was taken includes data for 108 different countries. The dispersion of incomes within a country generally is higher than the dispersion of wages because annual hours tend to be higher for persons with higher wage rates. Notice that for the United States the ratio of the highest to the lowest quintile in terms of income is 8.5, compared to 5.7 for the wage rates of full-time workers. Table 2 shows that there is substantial variation among countries in the dispersion of incomes – as there would be if the data referred to hourly wages rather than annual incomes. A few facts are illustrated by these data. First, there tends to be substantially less dispersion in individual wages and incomes within industrialized countries than in the developing countries. Second, there is much greater dispersion of incomes in the United States than in other industrialized countries. Third, the degree of wage and income dispersion in the formerly socialist countries is relatively low.

The Determinants of Individual Wages

The hourly wage rate (W) of any individual worker in a particular economy at a moment in time will, in principle, depend on a set of observable variables (such as that person's level of completed schooling) as well as on a set of factors that generally are not observed (such as the individual's innate ability, a characteristic that is difficult to measure and usually not obtained in a census-type survey). Let X represent all the variables determining wages that are observed and U represent all the unobserved determinants. We can then specify what economists call an earnings function in which the value of W for each individual depends on X and U according to

$$W = \beta X + \gamma U + \varepsilon \qquad [1]$$

where ε is a random error term (pure 'luck') that is unrelated to the values of X and U.

β is a vector of coefficients representing the effects of each of the observable variables on wages and γ a vector of coefficients representing the effects of the unobserved variables. If, for example, we thought that there were only two observable determinants of W, years of completed schooling (S) and age (A), βX might be specified as $\beta_1 S + \beta_2 A$, and the average effect of an additional year of schooling on W would be β_1.

In practice most estimated earnings functions feature the replacement of the numerical value of W with its natural logarithm. This means that a unit increase in the value of a particular X variable causes W to increase proportionately by the coefficient on that variable. In the case of years of schooling discussed above, with the semilogarithmic form β_1 represents $(\partial W/\partial)/W$ rather than $\Delta W/\Delta S$, if W is entered in numerical form.

From the earnings function given by eqn [1] it is clear that the wage structure, the distribution of W across the working population, depends on both the joint distributions of X and U (as well as the distribution of ε) and the parameters β and γ. Wage differentials, the *ceteris paribus* difference between the average wages of individuals with one characteristic versus another, are determined by the β parameters.

The various determinants of individual wages, both potentially observable and unobserved, can be classified into five groups. These include:

1. skill characteristics
2. job location
3. job characteristics
4. discrimination/nepotism
5. rents.

We will discuss the variables within these groups in turn.

Skill Characteristics

Generally, it is presumed that in most societies, more skilled workers are more productive and that they earn higher wages (*see* Income Distribution and Inequality: Measurement Issues). Two fairly crude but readily observable variables relating to skill of each worker are their years of schooling and potential labor market experience. Some microdata sets allow a more detailed specification of education (what subjects were studied, how many resources were allocated to each year of schooling, etc.) and experience (actual vs potential years, amount of on-the-job-training, etc.) but this sort of detail is rare. An important source of differences in skill and productivity differences among individuals is the variations originating from unobserved ability and motivation. Although a few surveys have included the scores on simple IQ tests, these factors are either largely or completely ignored; that is, they are considered U rather than X variables. This illustrates a general problem associated with the estimation of earnings functions. The true model is given by eqn [1] above, but the researcher is only able to estimate W as a function of observed variables, that is:

$$W = bX + e \qquad [2]$$

where e now represents variation in W that cannot be explained by the observed variables. We know that for most empirical earnings functions e is very important, for the fraction of the variation in W, jointly explained by the observable variables is usually fairly small – between 25 and 35% (see Juhn et al., 1993). There is thus a large potential for empirical estimates of the effects on W of any individual observed variable to be biased upward or downward, due to the correlation of that variable with unobserved variables. In the very simple model above in which schooling and age are the only observed variables, we want to estimate the 'true' effect of S on W from the model $W = \beta_1 S + \beta_2 A + \gamma U + \varepsilon$, but we are forced because of data limitations to estimate $W = b_1 S + b_2 A + e$. The use of b_1 as an estimate of the true effect β_2 is in general subject to specification error, which arises if S is correlated with γU (holding A constant). It seems reasonable to suppose that persons with relatively high levels of intelligence and/or motivation (1) will tend to have higher wage rates (i.e., $\gamma U > 0$) and (2) be more likely to have relatively high schooling levels. This would suggest that b_1 would be an upward-biased estimate of β_1. Interestingly, however, the consensus of the mass of research

into this question (see the review by Card, 1999) is that conventional estimates of the effect of schooling on wages are probably not significantly biased in an upward direction.

Another feature of empirical earnings functions is that their coefficients may change over time. This is especially true for coefficients concerning skill variables. For example, in the US the relative median weekly earnings of full-time workers 25 years of age and above, for those with a bachelor's degree (16 years of schooling) compared to those with just a high school degree (12 years of schooling), rose from 1.56 in 1996 to 1.75 in 2013 (author's calculations from the Current Population Survey). Why these changes in relative wages occurred has received a great deal of attention from economists, but the answer is still a matter of controversy (see Gordon and Dew-Becker, 2008).

Job Location

A second set of observable variables that are usually included in empirical earnings functions are dummy (one/zero) variables describing the location of the worker within the country. Typically the locational variables represent the size of the area in which the person lives (whether or not, for example, it is a metropolitan area with a population in excess of a million) and indicators of residence in particular areas. For example, in a hypothetical study of the determinants of wages in the Great Britain, one would include a large city dummy variable (L_{large}) and separate dummy variables for persons employed in Greater London (L_{london}), other parts of Southern England (L_{south}), and Scotland (L_{scot}). This specification would be represented as $W = b_1 L_{\text{large}} + b_2 L_{\text{london}} + b_3 L_{\text{south}} + b_4 L_{\text{scot}} + \cdots$ with the other observable determinants of wages (the X's in eqn [2]) included. The estimated coefficient b_2 in this earnings function would represent the average *ceteris paribus* wage differential between a person working in London and another working in a large city in Northern England. If, as is generally the preferred approach, W is entered as a natural logarithm, the estimated coefficient b_2 would reflect the approximate proportionate wage differential associated with employment in London.

What is the economic interpretation of coefficients on location of employment? First, because of higher land prices, it is generally more expensive to live in areas with a highly concentrated population. In order for real wages (nominal wages divided by the price level) to be equal across areas, observed wages must be appropriately higher in expensive areas. This would apply especially to very large cities within individual countries – London, Paris, New York, Tokyo, etc. A second reason for the existence of wage differentials by job location is the preferences of workers concerning the characteristics of different areas. For example, living in a large urban area involves greater stress and poorer air quality than living in a rural area, so, holding the cost of living constant, the typical worker might have to be paid a compensating wage differential to work in the large urban area. Similarly, some regions have better weather and other amenities (like public services) than others, and one would expect that wages would be higher in the more desirable than in the less desirable areas in order for potential migrants to be indifferent between different areas. For the United States, the estimated effect on wages of an area having sunshine 80% of the time versus another area having sunshine only 30% of the time, is, as simple economic theory would predict, a reduction of about 7% (Blomquist et al., 1988). This, of course, presumes that all (or most) workers in the economy have the same preferences for the amenities of different areas. To the extent that preferences differ across the population – some people prefer rural areas (warm weather) to large cities (cold), and others the opposite – predictions concerning compensating wage differentials become somewhat less clear. A final reason for the existence of wage differentials across locations is the effect of changes in the structure of labor demand. Suppose, referring to the example above, that there was for some reason a major boom in Scotland that caused the demand for labor there to increase permanently by 50%. It is unlikely that there would be an immediate large migration of workers from England to Scotland, so wages in Scotland would be bid up significantly. The earnings function for Great Britain based on data collected after the demand shift would yield an estimated coefficient b_{scot} that would now be positive. How rapidly the Scottish/English relative wage would return to one (the value of b_{scot} to return to approximately zero) depends on how long it takes for a portion of the British population to relocate to Scotland.

Job Characteristics

In addition to the location of employment, an important set of determinants of wage rates is the attributes of the job held by each individual worker. We can arbitrarily define each job attribute (the set of which we will call A) as 'bad' – the degree to which it is physically onerous, monotonous, a threat to the worker's health and safety, has low social status, etc. It would be expected that higher values of the elements of A would require a compensating differential. Put differently, holding all other (observed and unobserved) determinants of wages constant, the market equilibrium wage of each job would be higher for 'bad' jobs (those with high values of A) than for 'good' jobs (those with low values of A). Thus, for the specification $W = b_1 A_1 + b_2 A_2 + \cdots$, where the A's are numerical representations of bad characteristics (like the injury rate of the job), we would expect to estimate positive b_is.

There are numerous examples of the operation of compensating differentials. (For an excellent introduction to the topic, see Chapter 8 of Ehrenberg and Smith, 2011.) On the other hand, there are many instances in which the estimated effects of certain job attributes tend to be zero or have the 'wrong' sign (from the point of view of the theory) (see Brown, 1980 for a discussion of these problems). An important set of applications of estimates of job amenities on wages is to the question of the economic value of life. If one of the As (say A_1) is the rate of accidental death associated with the occupation of the worker, then one can infer from the estimated value of $\partial W/\partial A_1$, how much individuals value – apart from the opportunity to earn and consume – the continuation of their lives. This information can, in principle, be used to do cost/benefit analyses of programs to reduce risks to life and health.

An example of wage differentials associated with differences in job characteristics is given in **Table 3**. These results illustrate some of the difficulties in the interpretation of the effect of job attributes. The sample refers to individuals working full time in the United States in 1995, who had received a PhD in science or humanities. The regression equation relates the annual salary of each respondent to years since receipt of the doctorate,

gender, race, and region, as well as the field of receipt of the PhD and the current type of employment. The range of field effects (with a PhD in economics equal to 100) is from 73 for humanities to 105 for specialists in computer science. This means that, holding gender, race, years of experience, type of employment, and region constant, the computer humanities relative salary is 1.44. There are also substantial compensation differences across employer types. For example, PhD's employed in the for-profit business sector earn, *ceteris paribus*, 29% more than those in tenure track academic positions.

There are various explanations of why relative compensation differentials of this magnitude exist – in this case among sets of individuals who have the same quantity of education. First, these differentials may reflect compensating differentials in the sense that some fields (economics, for example) are intrinsically boring relative to other fields (like humanities) and require higher monetary rewards in order to induce persons to study them. Some of the relative compensation levels by employer type – for example, the business/academia comparison – most likely reflect premiums that are necessary to get PhDs to work in jobs that are perceived to be less rewarding in a nonfinancial sense.

A second potential explanation of differentials of this sort is that they reflect, in part, average ability differences of the individuals who are in each field or employment type. This is especially relevant for comparisons of employment type in which there is partial sorting on the basis of ability. For example, the compensation difference between academics in tenure track positions and those in both nontenure track positions and lower education (secondary schools and 2-year colleges) is very likely to reflect practices by which the 'better' PhDs are granted tenure.

A third explanation of these compensation differentials is that some of them reflect temporary market factors and there is

Table 3 Estimated relative salaries of PhD's in the United States by field of doctorate and type of employer, March 1995

Field of doctorate	
Computer science	105
Mathematics statistics	85
Agricultural science	82
Biology	86
Health sciences	97
Chemistry	85
Physical science	88
Economics	100
Psychology	83
Other social sciences	81
Engineering	97
Humanities	73
Type of employment	
Academic tenure track	100
Academic nontenure track	87
Postdoctoral fellowships	72
Lower education	89
Business	129
Nonprofit	111
Federal government	115
State and local government	93

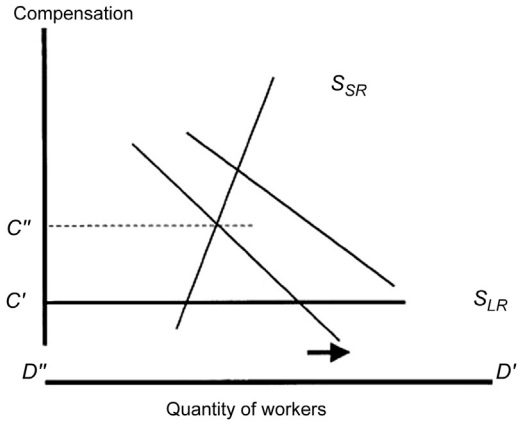

Figure 1 Adjustment of a labor market to demand shock in the short run and the long run.

a presumption that the high relative compensation of the computer science field in Table 3 is an example of this. This phenomenon is illustrated theoretically in Figure 1. The demand for a particular type of labor is assumed to increase such that the demand curve shifts from D′ to D″. In the short run the supply curve, S_{SR}, is quite inelastic, reflecting the fact that the pool of persons qualified to work in this market is fixed. (In the case of computer scientists, some additional labor can be obtained from people not on the labor force or from closely related fields, but these possibilities are limited.) Thus, the short-run effect of an increase in the demand for this type of labor is an increase in its relative compensation level. In the long run, the supply of labor to this market is very or completely elastic, because new entrants to the labor force can choose to train for this type of work. (In the case of computer specialists, young persons are induced to get PhDs in that field rather than in mathematics or physics.) Over time, therefore, the relative compensation level in this market would be expected to return to its initial equilibrium value, C′, rather than stay at its temporary value, C″ in Figure 1.

Analysis of wage differentials of various sorts can proceed along the above lines. A major example of this is the analysis of average compensation across industries. An example of this is Table 4 in which average weekly earnings across broad industrial groups are reported for the United States. Wages in retail trade, leisure and hospitality, and accommodation and food services tend to be relatively low; wages in mining, information, and finance tend to be relatively high. Some of these differentials reflect quality and compositional differences (that are not adjusted for in the table). But some of the differentials reflect differences in the amenities associated with work in the particular industries.

Gender and Ethnicity

A set of observable variables that are generally included in earnings functions are the demographic characteristics of the workers – his/her gender, race, or ethnicity. For example, the inclusion of a one/zero dummy variable for women as one of the *X*s in eqns [1] and [2] is very common, and the inclusion of motherhood dummy variables is becoming more common (for a summary, see Albelda and Drago, 2013). For the theoretically complete model given by eqn [1], which includes *all* observed and

Table 4 Average hourly earnings of employees by industry, United States

Industry	*Hourly wages*
Mining and logging	$30.80
Construction	$26.57
Manufacturing	$24.71
Wholesale trade	$27.94
Retail trade	$16.84
Transportation	$22.80
Utilities	$35.94
Information	$33.37
Financial activities	$30.53
Real estate	$23.82
Professional and business services	$28.95
Health care	$26.80
Accommodation and food services	$12.89

February 2014 estimates from the Current Employment Survey, US Bureau of Labor Statistics.

unobserved influences on wage rates, the interpretation of the (presumably negative) coefficient on women must be in terms of some sort of labor market discrimination against women. Why else would otherwise identical workers receive lower compensation? The same conclusion applies to conclusions concerning potential dummy variables on race and ethnicity for earnings functions in probably every country in the world (African Americans in the United States, French speakers in Quebec, Arabs in France, natives of Greenland in Copenhagen, etc.).

The problem with attributing the negative effects of gender and certain racial and ethnic distinctions entirely to labor market discrimination is that we are only able to estimate the model for which we have data, eqn [2] rather than the theoretically complete eqn [1] above. Thus, there is the possibility that the negative coefficients on gender and ethnicity reflect, wholly or in part, differences in the average value of omitted variables rather than some form of direct labor market discrimination. Further, it is necessary to know exactly what is the cause of an observed wage differential if a society decides it wants to eliminate this differential (see Altonji and Blank, 1999 for an extensive discussion of these issues).

Recent efforts to overcome these deficiencies involve tests where otherwise similar individuals apply for the same job, while diverging on a single characteristic (such as race, age, gender, or motherhood status). Although these tests are imperfect, they are consistent with the continued existence of labor market discrimination (see Chapter 8 of Albelda and Drago, 2013 for a discussion).

Rents

Most of the above discussion has proceeded as if wages are determined in markets that are similar in process to the markets for melons and equities. The fact that the commodities involved in labor markets are human beings requires that this 'model' be modified to some extent. Indeed, a set of institutions has emerged to provide modification of the outcomes that would follow from the unfettered equilibrium of the labor market. Among these institutions is trade unionism, which has, over time and across societies, taken several forms. Government policy – both directly through legislation determining wages and indirectly through its role as employer – also affects the wage structure. At one level the existence of institutions such as trade unions creates another wage differential of interest. In terms of eqns [1] and [2], being a union member or working in a milieu in which wage rates are determined by a collective bargaining contract means that a worker could earn a higher wage than another worker who is not so situated. In this case, $W = \beta_1 U + \cdots$ is the complete wage equation, including all observed and unobserved variables, where U equals one with union representation and equals zero otherwise. We do not observe everything about the workers and jobs, so we estimate the equivalent of eqn [2], $W = b_1 U + \cdots$.

How good the estimator b_1 is of the 'true' effect of unionism β_1 depends, as in the estimation of other wage differentials, on the correlation of the omitted determinants of wages with U. In the United States, the consensus estimate of the average effect of unionism on wages is about 15% (see Lewis, 1986). A more important effect of unionism and government policy, such as a mandated minimum wage, is their influence on other wage differentials. Although the final word is far from in, research suggests that much of the increase in wage inequality in the United States during the last two decades can be attributed to declining unionization and a declining real minimum wage (for a careful discussion, see Lemieux, 2008). That is, unionization and mandated minimum wages tended to equalize wages for many workers not covered by collective bargaining and/or earning above the minimum wage, and the reach of these institutions has declined.

See also: Discrimination, Economics of; Gender, Economics of; Income Distribution and Inequality: Measurement Issues; Inequality: Comparative Aspects; Informal and Underground Economics; Labor Supply; Labor Unions; Labor, Division of; Migration, Economics of; Trade Unions, Economic Behavior of; Wealth Distribution; Work, Sociology of.

Bibliography

Albelda, R., Drago, R., 2013. Unlevel Playing Fields. Dollars and Sense, Boston.

Altonji, J.G., Blank, R.M., 1999. Race and gender in the labor market. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3C. Elsevier Science, Amsterdam, pp. 3143–3260.

Blomquist, G., Berger, M., Hoehn, J., 1988. New estimates of the quality of life in urban areas. American Economic Review 78 (1), 89–107.

Brown, C., 1980. Equalizing differences in the labor market. Quarterly Journal of Economics 94, 113–134.

Card, D., 1999. The causal effect of education on earnings. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. Elsevier Science, Amsterdam, pp. 1801–1863.

Deininger, K., Squire, L., 1996. A new data set measuring income inequality. The World Bank Economic Review 10 (1), 565–591.

Ehrenberg, R., Smith, R., 2011. Modern Labor Economics, eleventh ed. Addison-Wesley, Reading, MA.

Gordon, R.J., Dew-Becker, R., 2008. Controversies about the Rise of American Inequality: A Survey. Working Paper No. 13982. National Bureau of Economic Research, Cambridge, USA.

Juhn, C., Murphy, K.M., Pierce, B., 1993. Wage inequality and the rise in returns to skill. Journal of Political Economy 101 (1), 410–442.

Lemieux, T., 2008. The changing nature of wage inequality. Journal of Population Economics 21, 21–48.

Lewis, H.G., 1986. Union Relative Wage Effects: A Survey. University of Chicago Press, Chicago.

Wallerstein, Immanuel (1930–)

Thomas E Reifer, University of San Diego, San Diego, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Immanuel Wallerstein's landmark multivolume *Modern World-System* and related works transformed the global intellectual landscape, part and parcel of larger changes in the geopolitics, structures of knowledge, and geoeconomics of the modern world. At once a scholarly analysis and a call to intellectual arms, Wallerstein's formation and personal trajectory demonstrates the importance of both intellectual contexts and social networks for paradigm shifts in the structures of knowledge, away from positive social science and toward a new unit of analysis, that of the study of world-systems, specifically the capitalist world-economy with its single and expanding global division of labor, stratified of course by race, ethnicity, class, gender, and nationality. This article provides an overview of Wallerstein's formation and career as a scholar-activist, including family origins, education, networks, historical context, and main achievements. The article includes a brief description of some of Wallerstein's major works and discusses their legacy and lasting impact, including in terms of his call for a unidiscplinary historical social science and possibly even beyond, that with moves toward overcoming what C.P. Snow called The Two Cultures dividing the social sciences, humanities, and the natural sciences, a unique feature of the structures of knowledge of the modern world-system.

Biographical Introduction

Immanuel Maurice Wallerstein, currently a senior research scholar at Yale University, was born on 28 September 1930 in New York City. Wallerstein, today a member of the American Academy of Arts and Sciences, is widely considered to be among the most accomplished critical social scientists of his era, since the publication of his landmark work, *The Modern World-System: Capitalist Agriculture & the Origins of the European World-Economy in the Sixteenth Century* (1974). Wallerstein did both his undergraduate and graduate work at Columbia University – (BA, 1951, MA, 1954, PhD, 1959) and went onto become a tenured professor in its Department of Sociology. As a graduate student, Wallerstein's scholarly and related political interests led him to do extensive field work, travels, and interviews in the Third World, uniquely among the major US social scientists of this period, notably in Francophone West Africa (but traveling throughout the continent), this during the period of colonialism and mobilization was against colonial rule. Wallerstein also spent time in Western Europe, notably Paris, where he met many of the critical diaspora intellectuals from Africa, Latin America, Asia, and the Middle East and his time in France also introduced him to the work and person of Fernand Braudel, one of the chief analysts of capitalism over the *longue duree*, as well as arguably the greatest historian of the twentieth century and the second generation leader of the Annales school, after Lucien Febvre and Marc Bloch, with its emphasis on geography and history. Along with Marxism, the German historical school and Latin American dependency theory, these emphases played critical roles in Wallerstein's career as one the world's foremost scholar-activists and public intellectuals and what Randall Collins called the golden age of (comparative) historical sociology of which Wallerstein has played such an integral part.

Wallerstein came of age during the heyday of post–World War II modernization or development theory as explicated most elaborately by Harvard Sociology's Talcott Parsons, the dominant consensus theorist of the American Century, akin to England's Herbert Spencer during the era of British hegemony, as well as by Walt Rostow (1960) in his *The Stages of Economic Growth: A Non-Communist Manifesto*. In this Whig interpretation of history, states, synonymous with society, are the primary units of analysis and seen as developing along parallel lines from traditional agricultural societies to modern urban-industrial societies based on high mass consumption. At once a cultural explanation of world inequality, modernization theory was simultaneously a political program for 'progress' and a central tenet of US foreign policy in the age of decolonization and national liberation movements. As the US and the Soviet Union squared off in the Cold War superpower confrontation, both the US and USSR touted themselves as embodying modernity – capitalism and communism respectively – telling the newly liberated Third World and those undergoing decolonization that it was best to join sides of one of the superpowers so as to develop and become modern. For Wallerstein, however, a much more basic conflict than this one, of which it was arguably a part, was the more fundamental North–South conflict and it was toward an understanding of and overcoming this divide that he would devote his scholarly career.

As Walter Goldfrank (2000: 153) notes in the most cogent sociological assessment of this Wallerstein's work to date,

> It is difficult to imagine Wallerstein's having come to consciousness outside of New York City … New York was where the action was in business and finance (Wall Street), in fashion and advertising (Fifth and Madison Avenues), in publishing and high culture. Like the London of Marx's maturity, the New York of Wallerstein's youth … was … the prime vantage point for seeing the world as a whole. Wallerstein's work is very much that of the inveterate New Yorker looking for the big picture, the politically savvy New Yorker looking through the ethnic garb of class struggles, the Jewish New Yorker deflating the pretentious claims to legitimacy of an Anglophile New England elite.

And it was at Columbia, too, as Goldfrank also pointed out, under the mentorship of C. Wright Mills, who he later contributed an entry for the *International Encyclopedia of the Social Sciences* (1968), that Wallerstein was likely influenced by Mills "two-front sociological attack on Grand Theory and Abstracted Empiricsm," then deployed against Talcott Parsons, which Wallerstein in turn used against the split in the social sciences between nomothetic universalizing theoretical disciplines (sociology, economics) and idiographic descriptive particularistic empirical ones (history, anthropology). From Mills too, helped along by Braudel, *Annales* and the German historical school, there is the historicism, attention to macro-structures and disdain for positivist US social science, which has characterized all of Wallerstein's work. Also important, from Wallerstein's travels in Africa and Paris, was his introduction to the literature on *negritude* associated with Aime Cesaire, Leopold Senghor (later President of Senegal), Leon Damas, and others. And perhaps even more importantly for Wallerstein, it was in Accra that he met Frantz Fanon, theorist and activist of the Algerian revolution, soon to become famous for his *The Wretched of the Earth*, which Wallerstein helped to get translated into and published in English.

Intellectually, though, along with the Third World radicalism that Fanon personified and which would be central to world-systems analysis, arguably most important were Marx and Engel's, especially their laser-like focus on the division of labor and social conflict between materially based human groups, all of which is at the heart of Wallerstein's world-systems perspective, but with a global eye, rather than the myopic blinders of the nation-state, which still occluded Marx's otherwise world encompassing historical vision of capital. Moreover, Wallerstein was at Columbia University during the world revolution of the 1968, as students occupied parts of the university to protest the Vietnam War and related issues. This was a real wake up call. Wallerstein was one of the few faculty trusted by the students and served as a member of the Executive Committee of the Ad Hoc Committee of Concerned Faculty. The experience pushed Wallerstein to the left politically, led to his departure from Columbia and eventually to his writing and publication of *The Modern World-System*.

As Thomas Kuhn (2012) and Randall Collins (1998) have argued in different but complementary ways, paradigm shifts in the structures of knowledge occur over long-periods. And just as paradigms are not built in a day, neither do they often fall overnight. Though in a variety of ways, key aspects of the main tenets of Wallerstein's world-systems analysis had been adumbrated before, notably by Africa diaspora intellectuals such as W.E.B. Du Bois and Oliver Cox, it was Wallerstein that first put them all together, albeit in a startling original synthesis, with many new features never before articulated. Here, too, the timing was just right. After the promissory note of developmentalism and domestic equivalents such as the War on Poverty gave way to the realities of the Vietnam War some 50 years ago, with social science luminaries such as Walt Rostow, who went onto become President Johnson's National Security Adviser, leading the way, the objectivity of the modernization paradigm was increasingly questioned. The cracks in the liberal consensus and turmoil at universities were part and parcel of a much broader wave of social movements and social conflict in the US and around the world, including over the appropriate paradigm to study social change. Here, following the success of Latin American dependency theory, centered around the Economic Commission on Latin America, and articulations of the development and underdevelopment by Latin American dependency theorists such as Andre Gunder Frank, Wallerstein's *Modern World-System* was a bold frontal attack, akin to Galileo's *Dialogo on the World Systems* and related works, or linguist Noam Chomksy's frontal assault on B.F. Skinner and behaviorism beginning in the 1950s and 1960s.

Most critical and enduring here was Wallerstein's critique of social science's modernization paradigm's unit of analysis, the state, seen as synonymous with society, developing along parallel lines. In sharp contrast, animated by socialist ideals of liberty, fraternity, equality, albeit in sharp contrast with actually existing socialism and global capitalism, Wallerstein reconceptualized social change as taking place within social systems, defined, as in Marx and Engel's *German Ideology*, where they laid out their vision of historical materialism, by the division of labor. But while Marx and Engels still remained trapped in the conceptual tyranny of the nation-state, for Wallerstein's capitalism was always a global system, with a single or increasingly global division of labor – radically stratified by race, ethnicity, class, gender, nation, and geographic location – and indeed as he set out to show, could not have been otherwise.

In what Wallerstein called "the long sixteenth century," following the formation of the regional forerunner of the capitalist world-economy in the Italian city-states, a new modern world-system was born, a European centered capitalist world-economy, based especially on the bringing together of (Latin) American resources, African (slave) labor power, and Asian markets. Here then was the proper unit – totality really – to study social change. Equally as important, Wallerstein argued that there developed a tripartite zonal distribution of the capitalist world-economy and interstate system, within which a polarization of wealth and power took place between rich core states and poor peripheral states, with a middling sector, the semiperiphery, playing an intermediate role. And, as can never be overemphasized, the designations of core, periphery, and semiperiphery are relational categories representing ongoing relational processes of the polarization of wealth and power in the capitalist world-system.

The Modern World-System

As many have noted, Wallerstein will likely be judged by his major work, *The Modern World-System*, now at four volumes, the first three of which have just been republished with new forewords by the University of California (UC) Press. Significantly, Wallerstein's "Introduction" to Volume I of this work is subtitled "On the Study of Social Change," and it is here that the method and research program, which he has followed all of his life was first laid out. The focus here is on the major watersheds in human history, specifically modernity, or the birth of the modern world. Unlike a generation of social scientists with their studied neutrality and objectivity, Wallerstein situated himself as an actor – a scholar to be sure,

but a committed one – in the global system. Wallerstein then relates how the African prism, as one scholar called it, led him to search for and then question what was the appropriate unit for studying social change and to be properly skeptical about "the degree to which society as an abstraction was heavily limited to politico-juridical systems as an empirical reality." So it was that Wallerstein came up with his alternative conceptualization of social systems, as defined by a single division of labor, transgressing the nominally sovereign boundaries of territorially based states and thereby reflecting economic interdependence, as the appropriate totality/unit of analysis for studying social change. The unique aspect of the modern world-system, Wallerstein then argued, was that it was based on the creation of a new type of social system that had never before existed, the capitalist world-economy based on a single division of labor and replete with an interstate system, which became simultaneously the first social system to encompass the entire globe.

After an initial chapter entitled "Medieval Prelude," Wallerstein articulated "The New European Division of Labor: *c*. 1450–1640," thereupon moving to the question of absolutism and then Spain's bid for world-empire, contrasted conceptually with the concept of the capitalist world-economy. Here we see a critical part of Wallerstein's method: analytical concepts and the process they designate are contrasted conceptually and empirically with each other, so that the core states of Western Europe are contrasted with Eastern Europe, the island states of the Caribbean and last but not least the remnants of the Spain's Hapsburg Empire, which largely become part of the semiperipheral zone. As with Marx, wealth and poverty here are conceptually and empirically entwined, albeit here in relational concepts of core, semiperiphery, periphery, and related geoeconomic regions, and in a global rather than a national context.

Here, development and underdevelopment result not from temporal lags in progressing toward modernity but are instead seen as an expression of the oligarchic wealth on which the capitalist world-economy is necessarily based, as Wallerstein's friend and colleague Giovanni Arrighi, author of (2010) *The Long Twentieth Century: Money, Power & the Origins of Our Times*, and (2009) *Adam Smith in Beijing: Lineages of the 21st Century*, has shown. In addition, the tripartite zonal division of the capitalist world-system is seen as concomitant with variegated modes of labor control in these zones, with nominally free wage labor more prevalent in the core and varieties of coerced labor more common in the periphery and semiperiphery. In the remainder of the book, Wallerstein looks at the formation of the core state of the capitalist world-economy, before going onto chronicle the formation of peripheral zones and to contrast these with those areas external to the operations of the world-economy, after which there is a "Theoretical Reprise" reflecting on the book as a whole.

Only in Volume II (1980), subtitled *Mercantilism and the Consolidation of the European World-Economy*, does Wallerstein fully do justice to "Dutch Hegemony in the World-Economy," as one of the early chapters is called, focusing thereafter on interstate competition among the core powers, before turning to peripheries, semiperipheries, and then back to the core. Here was more fully adumbrated the theory and empirical realities of the rise and fall of hegemonic powers in the world-system, in an analysis developed more fully, albeit in radically different ways, by leading world-systems scholar and historian of capitalism, Giovanni Arrighi. Wallerstein's Volume III (1989), *The Second Era of Great Expansion of the Capitalist World-Economy, 1730s–1840s*, offers too, like Volume I, a radical reinterpretation of modern world history. In the first chapter, "Industry and Bourgeoisie," Wallerstein dispenses simultaneously with the notion of both bourgeois revolutions, using the French Revolution as the paradigmatic example, as well as with the Industrial Revolution (the former conceived as anticapitalist despite its almost universal designation as a bourgeois revolution). Here the notion of the British industrial revolution is seen as an expression of British conceit and related ideologies of modernization, concealing the processes of industrialization as a global process centered on England rather than a national British event.

Wallerstein then chronicles the ascendancy of British hegemony in the context of the Napoleonic wars, before going onto deal with the incorporation of Russian and Ottoman empires, as well as India and West Africa, into the capitalist world-economy. The final chapter, "The Settler Decolonization of the Americas," includes analyses of the American and Haitian revolutions and the independence struggles that led to the decolonization of Latin America. Volume IV, subtitled *Centrist Liberalism Triumphant, 1789–1914*, involves in many ways a radical departure from Wallerstein's previous three volumes in its focus on the politics of the world-system rather than on the economics. In the foreword, Wallerstein lays out a prospectus for the rest of his project, namely a Volume V from 1873 to 1968/89 and a possible Volume VI which promises to chronicle the history of the modern world-system right up to the present and to project ahead to somewhere in the mid-twenty-first century, perhaps to 2050.

Over these many decades, too, Wallerstein has published widely, with numerous books, articles, and collections of essays, from his three volume *The African Liberation Reader: Documents of the National Liberation Movements* (with Aquino de Braganca) (1982), the landmark *Antisystemic Movements* (with Giovanni Arrighi and Terence Hopkins) (1989), to *Geopolitics and Geoculture* (1991), *Creating and Transforming Households* (with Joan Smith and others) (1992), *European Universalism: The Rhetoric of Power* (2006), and *Race, Nation, Class* (with Etienne Balibar) (2011). But it is likely that when all is said and done, Wallerstein will be remembered most, after his work on the capitalist world-economy and related inequalities of race, ethnicity, class, and gender, for his work on the structures of knowledge of the modern world-system, part of a larger critique of Eurocentrism and the coloniality of power associated most especially with the work of Edward Said and Peruvian Sociologist Anibal Quijano, respectively. Most important here, along with his *Unthinking Social Science* (2001), the three volume collection of discussions with leading thinkers, *Questioning Nineteenth Century Assumptions About Knowledge* (with Richard Lee) (2010), is *Open the Social Sciences: Report of the Gulbelkian Commission on the Restructuring of the Social Sciences* (1992). Yet it is important to note that all of this work is the outcome of a lifetime spent rejecting and critiquing the premises of mainstream social science and formulating an alternative more holistic vision, drawing on a host of leading albeit iconoclastic scholars such as the institutional economist Karl

Polanyi. Already, at the end of the Introduction to Volume I, Wallerstein (1974: 11) noted:

> When one studies a social system, the classical lines of division within social science are meaningless. Anthropology, economics, political science, sociology – and history – are divisions of the discipline anchored in a certain liberal conception of the state and its relation to functional and geographical sectors of the social order. They make a certain limited sense if the focus of one's study is organizations. They make none at all if the focus is the social system. I am not calling for a multidisciplinary approach to the study of social systems, but for a unidisciplinary approach. The substantive content of the book will, I hope, make it clear what I mean by this phrase and how seriously I take it.

In *Open the Social Sciences*, Wallerstein and his coauthors trace the specificities of modernity's structures of knowledge, most especially what C.P. Snow (2013) referred to as *The Two Cultures* – the unique division between the arts, humanities, and social sciences. Wallerstein chronicles, too, the formation of the modern-day university system in England, France, Germany, Italy, and the US, noting the overwhelming preoccupation of those countries in studying about themselves and the Euro-American world. For as conceptualized by Eric Wolf, there was only *Europe & the People Without History* (2010). Over time, of course, with the emergence of area studies, borders between the disciplines began to break down, leading to widespread calls for the reorganization of the structures of knowledge to which Wallerstein has added his own voice. In subsequent works Wallerstein takes heart from the introduction of historicity, notably the arrow of time, into the natural sciences. Here, Wallerstein finds inspiration from a member of his Gulbelkian Commission, winner of the Nobel Prize in Chemistry for his work on nonlinear (thermo)dynamics, Ilya Prigogine, and the mathematician, Ivar Ekeland.

Simply put, it is hard to think of another work, project really, in the historical social sciences that had a greater impact across disciplinary boundaries than *The Modern World-System*. The work is routinely cited in almost all critical accounts of development or modernization theory and often times by scholars working on particular geographic areas. It permanently altered the intellectual universe of world social science and intellectual culture more generally; not because modernization theory has become a minority position, quite the contrary, especially with recent analyses of *The Ends of History* but instead because alternative critiques can no longer be ignored, and equally as importantly because there are now an increasing body of scholars, inspired in no small part by the work of Wallerstein and his cohorts, presenting cogent scholarly accounts of world history and global inequalities, from the Paleolithic to the present, from a world-systems perspective. Indeed, in many ways, for all its problems, contemporary accounts of globalization – not to mention the current ecological crisis in the age of the Anthropocene – reflect a growing awareness of global interdependence first presented in Wallerstein's work. Before the publication of *The Modern World-System*, European economic history and that of the Third World were almost always presented separately; not so any more, at least as a matter of course.

Here, it is important to note that the intellectual contributions of Wallerstein have been combined with concentrated activity in providing a base for world-systems analysis as an intellectual social movement. With the assistance of his long-time friend and colleague Terence K. Hopkins from Columbia, Wallerstein established a unique graduate program at the State University of New York (SUNY) Binghamton, as well as the Fernand Braudel Center for the Study of Economies, Historical Systems, and Civilizations, with its annual journal *Review*, and related Research Working Groups, making it for some time the key locus for world-systems studies. Simultaneously Wallerstein formed the Political Economy of the World-System section of the American Sociological Association, virtually unique too in its annual conferences and concomitant publication of a yearly volume, with the *Journal of World-Systems Research*, now its official peer-reviewed journal (and the first free open-source one to be established).

Intellectual Influence and Legacy: Lasting Impact and Possible Futures

In the beginning of *The Modern World-System*, Wallerstein noted that it is widely recognized that there have been a series of epochal changes in human history, in the context of discussing the proper spatiotemporal framework for the study of (inequality in) the modern world-system and the need to look at historical capitalism over the *longue duree*.

> One such generally recognized watershed, though one however studied only by a minority of social scientists, is the so-called Neolithic or agricultural revolution. The other great watershed is the creation of the modern world.

> This latter event is at the center of most contemporary social science theory, and indeed, of the nineteenth century as well ….

> … there is no area of study that is not "relevant." For the proper understanding of the social dynamics of the present requires a theoretical comprehension that can only be based on the study of the widest possible range of phenomena, including through all of historical time and space.

To be sure, Wallerstein has not personally heeded his own call to go back into the premodern past to understand the present, though many of his colleagues – one thinks especially of Christopher Chase-Dunn, Director of the Institute for Research on World-Systems at UC Riverside – have. But it would perhaps not be too much to say that by helping to reorient the social sciences and the humanities toward the study of global inequality – something now increasingly being investigated in the natural sciences as well, what with climate change – Wallerstein helped to create the wave of studies that have done exactly this, most especially by analyzing that other major watershed of human history, the Neolithic revolution, indispensable for our understanding of modern inequalities. One thinks here most especially the work of Alfred Crosby, *The Columbian Exchange* (2003), *Ecological Imperialism: The Biological Expansion of Europe, 900–1900* (2004), and Jared Diamond's

Pulitzer Prize winning *Guns, Germs & Steel: The Fates of Human Societies* (2004).

By widening the spatiotemporal horizons of scholars interested in modernity, global inequalities and by helping bring back space and time, as well as geography, history, ecology, climate, and epidemiology into the social sciences, following the tradition of the *Annales* and German historical schools, not to mention by updating Marx's social conflict analysis for a new millennium and in a global vein, Wallerstein has fundamentally reshaped the terrain of global intellectual culture and scholarly inquiry. Moreover, as the virtual inventor of an intellectual social movement, something fairly rare in the structures of knowledge of the modern world, and as an indefatigable scholar and organizer in its knowledge structures (most recently in his capacity as President of the International Sociological Association from 1994 to 1998), Wallerstein has also become a public intellectual cited by both scholars in the ivory tower and scholar-activists in social movements outside its hallowed halls, including in the World Social Forum(WSF). The rise of world-systems analysis is of course part of a broader critique of developmentalism expressed in movements such as the Zapatistas, the WSF, and among Latin American social movements, with their concept of *buen vivir*: a state of living well (Esteva et al., 2013). Proponents of and sympathizers with world-systems analysis can now be found among leading scholars and activists across the globe, whose manifold contributions to the study of the world-system have been amply discussed elsewhere. By opening up the question of the proper unit of analysis for the study of social change and forwarding his own, the single division of labor, Wallerstein has contributed to a veritable revolution in the structures of knowledge, opening up vast new spaces inside and outside of the academy. Turn to any academic discipline in the social sciences or humanities today and you will find Wallerstein's world-systems analysis cited, debated, and discussed.

There are, to be sure, lacuna and unanswered questions. Though it has seldom been commented on, as noted above there was always a conscious analogy, if understated, in Wallerstein's work, to Galileo's *Dialogo*, and to astronomy and physics more generally, as intimated in the introduction to the *Modern World-System*.

> There had only been one "modern world." Maybe one day there would be discovered to be comparable phenomena on other planets, or additional modern world-systems on this one. But here and now, the reality was clear-only one. It was here that I was inspired by the analogy with astronomy which purports to explain the laws governing the universe, although (as far as we know) only one universe has ever existed.

> What do astronomers do? As I understand it, the logic of their arguments involves two separate operations. They use the law of physics, and argue that … these laws hold by analogy for the system as a whole. Second, they argue a posteriori. If the whole system is to have a given state at time y, it most probably had a certain state at time x.

At the time of Wallerstein's writing, not only were the social sciences far from the unidiscplinary historical social science he called for (virtually alone), as they still are today, but also there were even greater barriers separating the natural and the social sciences, as well as the humanities, from each other. Since that time, there has been a noticeable blurring of the lines not only in the social sciences but between these and the natural sciences and humanities as well. Indeed, as many scholars are now becoming increasingly aware of, all forms of life on the planet earth are related to each other and moreover, everything in the universe is related, as Neil Shubin (2009, 2013) has shown in *Your Inner Fish: A Journey into the 3.5 Billion Year History of the Human Body* and more recently in *The Universe Within: The Deep History of the Human Body*. As to the question of how many universes, even this is question in the age of speculation about multi-universes. We can thus see today that Wallerstein was well ahead of his time with his call for a unidiscplinary social science, his drawing attention to the unique modern split between *The Two Cultures* as identified by C.P. Snow and his related trespasses and explorations into the natural sciences as well, borrowing concepts from the latter to analyze the former. Most significant here is Wallerstein's argument, taking from Prigogine's work on chaos theory and nonlinear dynamics of systems moving far from equilibrium and related moments of bifurcation, that we are living in a period of the terminal crisis of the capitalist world-economy and birth of new alternative system(s).

A curious omission here is the nature of humanity itself, as different from other animal species and that of course raises the question of language – only some 50 000 years old – and Chomsky's theory of universal grammar, neglected, with some rare albeit important exceptions by world-system scholars and most social scientists as a whole. And yet, as Wallerstein's own colleague and collaborator Ilya Prigoine (with Isabelle Stengers) noted, "… we need an account of the material world in which it isn't absurd to claim that it produced us" (quoted in Terrence W. Deacon's (2012: 143) *Incomplete Nature: How Mind Emerged from Matter*) and the us, humanity, must certainly include an analysis of how so-called mind, human consciousness, or language, emerged from matter, to use layperson's terms. Such questions, to varying extents, into the nature of language and humans were at the heart of the entwined cognitive and scientific revolutions of the seventeenth century, as recognized by Galileo, Descartes, Locke, Hume, and others, as Emeritus Professor in Linguistics and Philosophy at MIT, Noam Chomsky has argued for many decades, notably in his *Cartesian Linguistics: A Chapter in the History of Rationalist Thought* (2009).

The question of the nature of human and human language have been examined in great detail by Noam Chomsky, including recently in his chapter "The Mysteries of Nature: How Deeply Hidden," in Jean Bricmont and Julie Franckm eds, *Chomsky Notebook* (2009) and "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?," (with Marc D. Hauser and W. Tecumseh Fitch) in Richard K. Larson, Viviane Deprez, and Hiroko Yamakido, eds, *The Evolution of Human Languages: Biolinguistic Perspectives* (2010), the latter calling for interdisciplinary work so as to examine the biolinguistic capacity of humans. Chomsky himself argues that the entire field of cognitive science remains largely at the pre-Galilean stage because of unfounded assumptions and hence has drawn attention to the scientific breakthroughs of Isaac

Newton, with his concepts of action at a distance and gravity, exorcising the machine but not the ghost, at once lowering the expectations for science, thereby paving the way for progress in the field, or normal science. What is especially ironic, here, as has been pointed out by Chomsky and others, is that Newton destroyed the mechanical philosophy and related ideologies of materialism just at the time when these doctrines became ever more influential and to be sure, the legacy of (historical) materialism can be seen in world-systems analysis right up to the present, remnants of the old battle against (Hegelian) idealism.

It remains to be seen of course, whether and to what extent Wallerstein's call for an unidisciplinary social science will be taken up and whether this might be joined with developments in the natural sciences, most especially in terms of our understanding of human language and hence humanity itself as part of natural and social evolution; or whether, and to what extent, the growing realization of the connection of all knowledge and life in the universe, will lead to new revolutions in the structure of knowledge and new understandings of what kind of species we humans are, as one of Noam Chomsky's lectures at Columbia University in the Fall of 2013 was recently titled. That these questions can even be formulated, however fleetingly though, is testament to a life well lived, albeit of course still being lived, of a trail blazer in, and reshaper of the structures of knowledge, who has made and is continuing to make, along with that of his colleagues, enduring contributions to an understanding of the present world-system, so as to encourage greater respect for nature and other human beings and living species, and to changing it as well.

See also: Annales School; Area and International Studies: Political Economy; Capitalism: The History of the Concept; Capitalism; Chomsky, Noam (1928–); Colonization and Colonialism, History of; Dependency Theory; Du Bois, William Edward Burghardt (1868–1963); Global South; Inequality: Comparative Aspects; Marx, Karl (1818–83); Mills, Charles Wright (1916–62); Moore Jr, Barrington (1913–2005); Parsons, Talcott (1902–79); Polanyi, Karl (1886–1964); Race and Racism in the Twenty-First Century; Third World.

Bibliography

Braudel, Fernand, 1992. Civilization & Capitalism, 15th–18th Century, vol. I–III. University of California Press, Berkeley.

Cesaire, Aime, 1994. Aime Cesaire: A Voice for the 21st Century. California Newsreel DVD.

Collins, Randall, 1998. The Sociology of Philosophies: A Global Theory of Intellectual Change. Harvard University Press, Cambridge, MA.

Deacon, Terence W., 2012. Incomplete Nature: How Mind Emerged From Matter. W.W. Norton & Co., New York.

Esteva, Gustava, Babones, Salvatore, Babcicky, Phillip, 2013. The Future of Development: A Radical Manifesto. Policy Press, Bristol.

Goldfrank, Walter L., 2012. Wallerstein's world-system: roots and contributions. In: Babones, Salvatore J., Chase-Dunn, Christopher (Eds.), Routledge Handbook of World-Systems Analysis. Routledge, New York, pp. 97–103.

Goldfrank, Walter L., Summer/Fall 2000. Paradigm regained? The rules of Wallerstein's world-system method. Journal of World-Systems Research 1 (2), 150–195.

Kuhn, Thomas, 2012. The Structure of Scientific Revolutions, 50th anniversary ed. University of Chicago Press, Chicago.

Quijano, Anibal, Wallerstein, Immanuel, 1992. Americanity as a concept, or the Americas in the modern world-system. International Social Science Journal, 549–557.

Shubin, Neil, 2009. Your Inner Fish: A Journey Into the 3.5 Billion-Year History of the Human Body. Vintage, New York.

Shubin, Neil, 2013. The Universe Within: The Deep History of the Human Body. Vintage, New York.

Snow, C.P., 2013. The Two Cultures. Cambridge University Press.

Wallerstein, Immanuel, 1974. The Modern World-System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century. Academic Press, Orlando.

Wallerstein, Immanuel, Collins, Randall, Mann, Michael, Derlugian, Georgi, Calhoun, Craig, 2013. Does Capitalism Have a Future? Oxford University Press.

Wallerstein, Immanuel, Lemert, Charles, Aguirre Rojas, C., 2013. Uncertain Worlds: World-Systems Analysis in Changing Times. Paradigm, Boulder, CO.

Wallerstein, Immanuel, 2012a. World-systems as a knowledge movement. In: Babones, Salvatore J., Chase-Dunn, Christopher (Eds.), Routledge Handbook of World-Systems Analysis. Routledge, New York, pp. 515–521.

Wallerstein, Immanuel, 2012b. Translation of Fernand Braudel's 'Historie et Sciences Sociales: La Longue Duree'. In: Lee, Richard (Ed.), The Longue Duree & World-Systems Analysis.

Wallerstein, Immanuel, 2000. The Essential Wallerstein. New Press, New York.

Walzer, Michael (1935–)

Gerard Boucher, UCD School of Sociology, University College Dublin, Dublin, Republic of Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the life and work of Michael Walzer (1935–present), whose many roles include professor, moral philosopher, critic, editor, activist, public intellectual, social democrat, American Jew, and devoted family man. He is the author or editor of 35 books from 1965 to 2012, 11 in other languages, and 437 articles, essays, and book reviews from 1956 to 2012. Walzer's work is characterized by a concern for equality and diversity within, and between, communities, bound by a shared culture that provides the basis for moral critique, political action, and democratic participation in decision-making and progressive social change.

Michael Walzer was born in New York City in 1935 to first-generation Jewish immigrants from Austria, Galicia, and Belarus, and was raised in the Bronx. His father worked in the fur trade, but went bankrupt during the economic downturn of 1937–38, and subsequently worked in a factory involved in war production until 1944. At 9 years of age, his family moved to Johnstown, Pennsylvania, a steel town east of Pittsburgh, where his father had been offered a job managing a jewelry store (Williams, 2012: 3). His parents were political supporters of the anti-Stalinist left, reading the popular front daily newspaper, PM, and later I.F. Stone's Weekly. At the time, Johnstown was both a strong trade union town dominated by work in Bethlehem Steel, and a New Deal, liberal Democratic Party stronghold. Walzer describes his upbringing as growing up in a 'lefty household' with politics as 'the daily conversation' of everyday life (Michael Walzer: Who Are You?).

At 17, Walzer attended Brandeis University in Waltham, Massachusetts, a nonsectarian, Jewish community sponsored, coeducational institution that had been established in 1948, where he met other 'Jewish red diaper babies and Jewish lefties of different sorts' (Michael Walzer: Who Are You?). During his undergraduate years at Brandeis from 1952 to 1956, Walzer took courses with Irving Howe and Lewis Coser, members of the New York Intellectuals group (Williams, 2012: 3). In 1954, he was invited to the first issue launch party of *Dissent*, a social democratic magazine founded by Howe and other ex-Trotskyists. *Dissent* attempted to define a democratic socialist, third position between Soviet communism and American liberalism (Reiner, 2011: 199). It sought to identify and retain socialist beliefs and practices that remained relevant and useful, particularly in the American context, while incorporating features from liberal democracy such as individual choice, a market economy, and democratic decision-making (Reiner, 2011: 199, 200). This shows an ideological pluralism in Walzer's thought that remains characteristic of his work throughout his life, and is one of the reasons that singular labels such as socialist, liberal, conservative, or communitarian are difficult to apply to his work.

Walzer was asked by Howe to write his first article for *Dissent* in 1956 concerning the CPUSA's (Communist Party USA) response to Krushchev's speech about Stalinism. Working and writing for *Dissent* became a significant part of Walzer's life until his retirement from the magazine in April 2013 (*Dissent's* Sixtieth, 2013). Howe sent Walzer to the American South in 1960 to write about the sit-ins, and he became an active participant in the civil rights movement until Jews and other Whites were invited to leave by Black Nationalists in 1967. He then focused his political activism in the anti-Vietnam War movement and, more broadly, in the American social democratic left (Williams, 2012: 4, 5). While still actively contributing articles, he became an editor at *Dissent* in 1976 (and a contributing editor to *The New Republic* from 1977) and, after Irving Howe's death, coeditor of *Dissent* for 20 years from 1993 (Tracy, 2013: 1).

The influences of democratic socialism, Irving Coser, *Dissent* magazine, and his political activism are most directly expressed in *Political Action* (1971) and *Radical Principles* (1980). *Political Action* is a description and critique of citizen activism in the United States from an insider's perspective following the American invasion of Cambodia during the Vietnam War (1971: 9). Political action involves 'action with or for others' for public reasons to have effects on 'other people as well as ourselves' (1971: 18). Its purposes are to expose injustice and oppression and oppose it through mobilization, revolt, and social change (1971: 120, 121). In the American case, political action addresses the sense of powerlessness experienced by many Americans who believe that 'effective control' of government and the economy has been gradually taken away from citizens (1971: 121). Walzer offers practical advice for political organizations, noting in particular that the specific cause that unites activists reflects only a part of their lives; the other parts of their lives may overlap with each other, opening political space for coalitions or alliances with other groups based on different causes (1971: 32, 40, 41). This shows a concern for organizational pluralism that complements his ideological pluralism. *Radical Principles* (1980) is more focused on the prospects for democratic socialism in the United States. The book is a collection of articles published between 1964 and 1979 in *Dissent* and *The New Republic*. It is dedicated to Irving Howe, Lewis Coser, and 'all my comrades on *Dissent*' (1980: v). The common theme revolves around the question of how the principles of democratic socialism can be maintained and renewed in response to significant changes in the United States during the 1960s and 1970s (Reiner, 2011: 51).

While at Brandeis, Walzer met Judith Borodovko, another student of Howe's, and they married in 1956, after Walzer

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61304-1

graduated summa cum laude with a BA in History from Brandeis. They had two daughters born in 1961 and 1964, and later grandchildren. Along with his work at *Dissent* and his political activism, family and the Jewish community have been central to Walzer's life. His wife in particular became one of the main interlocutors of his work; he calls her, after 60 years together as a couple, "my closest comrade and smartest critic" (*Dissent*'s Sixtieth 2013). Extending beyond yet interrelated with his family, Walzer's "deepest personal connection is with the American Jewish community" (Morton, 2013: 6). This reflects his background as an American Jew who grew up in the aftermath of World War II, the Holocaust, the founding of the state of Israel, and the increasing acceptance of American Jews as an ethnic and religious group in American society. In the preface to *On Toleration* (1997), Walzer notes that "as an American Jew, I grew up thinking of myself as an object of toleration" (1997: xi), and that this background helps to explain his support for religious, ethnic, and national pluralism in the United States and across the world. Walzer also describes himself as having been "a Zionist ever since my bar mitzvah" in 1948 in Johnstown and as a half-immigrant to Israel (*The State of Righteousness* 2012). He has been a member of the Board of Governors of the Hebrew University of Jerusalem since 1974, and is a member of the editorial board of the Jewish Review of Books.

Jewish themes are prevalent in many of Walzer's major works. These include a dedication to those killed in the Holocaust drawn from an inscription at the Yad Vashem memorial in Jerusalem, and an examination of the 1967 Six Day War in *Just and Unjust Wars* (1977); discussions of A Medieval Jewish Community, The Israeli Kibbutz, and The Idea of the Sabbath in *Spheres of Justice* (1983); and a chapter on the biblical prophet Amos in *Interpretation and Social Criticism* (1987). His first book on solely Jewish themes was *Exodus and Revolution* (1985). Fifteen years later, he coedited two volumes on the Jewish Political Tradition on authority (2000) and membership (2003). These were followed by an edited book, *Law, Politics, and Morality* (2006), and Walzer's most recent sole-authored book in English, *In God's Shadow: Politics in the Hebrew Bible* (2012).

After Brandeis, Walzer studied at the University of Cambridge in the 1956–57 academic year on a Fulbright Fellowship. Under the supervision of Geoffrey Elton, he began to research English history in the sixteenth and seventeenth centuries (Reiner, 2011: 4). At the end of the fellowship, Walzer began his doctoral studies in the Department of Government at Harvard University, earning his PhD in Government in 1961. His adviser, Samuel Beer, was a major influence on Walzer with respect to his dissertation topic on the development of Calvinism as a political movement in seventeenth-century England. Beer influenced Walzer's use of historical idealism as a conceptual framework in which "ideas are dependent on particular historical and social contexts and cannot be understood in abstract from those contexts," and as a method of "using history illuminate theory and testing theory with historical cases" (Reiner, 2011: 15, 18). Carl J. Friedrich, Louis Hartz, and Barrington Moore also had an impact on Walzer at Harvard in terms of historical idealism, and comparative historical analysis in Moore's case (Reiner, 2011: 18). These influences are evident in the publication of Walzer's doctoral dissertation as his first book, the *Revolution of the Saints* (1965), and *Regicide and Revolution* (1974), which was dedicated to Beer.

Walzer was hired as an Assistant Professor at Princeton University from 1962 to 1966, before returning to Harvard as a Professor from 1966 to 1980. *Obligations* (1970) was based on lectures that Walzer gave at Harvard between 1966 and 1969 (1970: v). While at Princeton, Walzer became involved with a community of philosophers including Robert Nozick and Thomas Nagel (Reiner, 2011: 5). Nozick and Nagel organized a discussion group, the Society for Ethical and Legal Philosophy (SELF) in 1967, and the journal *Philosophy and Public Affairs*, founded with group members in 1971 (Nagel, 2013: xi). Walzer was an integral member of the group that included, among others, Ronald Dworkin, John Rawls, and Judith Jarvis Thomson, presenting work at their monthly meetings. Nagel describes the group as 'united by two convictions' (Nagel, 2013: xii). The first was 'a belief in the reality of the moral domain' in which 'there are real questions with right and wrong answers'; the second was 'a belief that progress could be made toward discovering the right answers' (Nagel, 2013). Group members applied 'analytic methods of reasoning and argument,' including 'imaginary cases,' to 'examples of conduct and policy' to reveal the 'principles determining right and wrong, justice and injustice' (Nagel, 2013: xii, xiii). In a 2003 interview, Walzer acknowledged the influence of SELF on his work, particularly Nozick and Rawls, stating that he "spent much of the 1960s and 1970s learning to 'do' philosophy" and that, as a result, he made "an effort to write about politics in a more philosophical way" (2007: 307, 308). However, he adds that "I quickly got impatient with the playful extension of hypothetical cases, moving farther…away from the world we lived in" (2007: 308).

Walzer spent much of the 1970s writing *Just and Unjust Wars* (1977), his response to 'the American intervention in Vietnam' and his political activism in the antiwar movement (1997: xix). The book acknowledges the importance of SELF and individual members, particularly Nozick and Thomson (xxvii). It provides an analytical treatment of practical morality in our shared moral world, looking 'to recapture the just war for political and moral theory' (1977: xxii) by secularizing 'Catholic just-war doctrine' (Williams, 2012: 5). Walzer adds that his "decision to work the argument through historical examples was in part a reaction against the hypothetical cases of my friends" in SELF (2007: 308). The book is also an application of the historical idealism and comparative historical analysis that he learned at Harvard. In particular, the book analyzes the case for, and cases of, the justice of war or *jus ad bellum* from justice in war or *jus in bello* (1977: xxiv), emphasizing the human rights of life and liberty as 'something like absolute values' that should provide a moral limit to utilitarian and realist arguments about going to and conducting war (1977: 21).

Walzer supports the doctrine of nonintervention based on the presumption that just wars are defensive wars, and on the right of individuals to exercise collective self-determination as a national community in their sovereign territory (Reiner, 2011: 74, 178). This shows Walzer's belief in international pluralism. However, he reconsidered this position on intervention in later works. In the edited collection, *Arguing About War* (2004), Walzer adds an account of *jus post bellum* or justice

after war to his just war theory, and accepts that he is more willing to condone military interventions for gross human rights violations of life and liberty such as massacres and ethnic cleansing (2007: 251). He is also more willing to accept occupations and nation-building (Reiner, 2011: 178). These revisions to his position on nonintervention are supplemented further in edited collection, *Getting Out: Historical Perspectives on Leaving Iraq* (2009), where he argues for an ethics of withdrawal that takes seriously the dangers and safety of vulnerable people post withdrawal (Reiner, 2011: 191).

In 1980, Walzer was hired as a Professor in the School of Social Science at the Institute for Advanced Study (IAS) in Princeton, New Jersey. In 1986, he was made UPS Foundation Professor at the School which he held until 2007, when he became a Professor Emeritus of the Institute. He joined a small permanent faculty at the School that included the symbolic anthropologist Clifford Geertz (1970–2000) and the heterodox economist Albert Hirschman (1974–85), who together had developed an interpretive style that sought to bridge the divide between social science disciplines. Both were empirically oriented scholars who did not believe in grand theories or abstract models to explain human behavior, focusing instead on understanding the complexity, meaning, and consequences of human actions in collective contexts. Geertz was also a cultural particularist, who did not believe in human universals across cultural groups, and who developed a method of thick description to interpret cultural symbols observed while in use. The IAS, and in particular Geertz and Hirschman, are the last important influences on Walzer's work.

The cumulation and compounding of influences in his work, evident in his early books, is more apparent in some of his major works from his IAS period. This is shown in *Spheres of Justice* (1983), which is dedicated to his deceased father, Joseph, and acknowledges Judith Walzer, Howe, Nozick, Rawls, Geertz, and Hirschman among others, and refers to presentations and discussions at SELF and the IAS. His democratic socialism is clear in the book's purpose "to describe a society where no social good serves or can serve as a means of domination" (1983: xiv). The social goods analyzed in the book are illustrated using situated historical examples following the historical idealism of Beer and his colleagues at Harvard University. There is also an anthropological awareness of human culture and diversity reflecting Walzer's interactions with Geertz at IAS. The emphasis on interpreting 'to one's citizens the world of meanings that we share' (1983) further reflects the influence of Geertz and Hirschman's interpretive style and Geertz's cultural particularism. The topic of distributive justice also indicates the influence of SELF, particularly Rawls and Nozick, who each published major books in the early 1970s on the topic. The main influence, though, is Nozick who taught a course with Walzer at Harvard on Capitalism and Socialism in 1970–71 in 'the form of an argument,' with 'half the argument' expressed in Nozick's *Anarchy, State and Utopia* (1974) and the other half in *Spheres of Justice* (1983: xvii).

While protecting the human rights of life and liberty is the basis of *Just and Unjust Wars*, the distribution of social goods within different spheres of society is the focus of *Spheres of Justice*. These goods include: membership in the political community; security and welfare; money and commodities; office, work, and leisure; education; kinship and love; religion; recognition; and political power. The number and type of goods in a particular community depends on distinct meanings shared by its members based on their collective self-determination, with each good relatively autonomous in its sphere, with distributions between members of each sphere based on internal principles or standards of justice (1983: 10). However, there is not one 'distributive system that philosophy can rightly encompass' that would provide a single standard for just distributions across the spheres (1983: 4, 5). Instead, 'principles of justice' are 'pluralistic in form' and are 'the inevitable products of historical and cultural pluralism' (1983: 6). Balance between the spheres is meant to prevent dominance of one social good, and monopoly of a dominant good by a specific social group or coalition (1983: 11, 12). At the same time, ongoing struggles between groups and coalitions change the meanings of justice and injustice, and the distribution and balance between goods in a community (1983: 12, 13).

While it is a good that can be monopolized and made dominant, political power is 'the regulative agency for social goods generally' (1983: 15). It is "used to defend the boundaries of all distributive spheres, including its own, and to enforce the common understandings of what goods are and what they are for" (1983). The result is a regime of complex equality in which "no citizen's standing in one sphere or with regard to one social good can be undercut by his standing in some other sphere, with regard to some other good" (1983: 19). Democratic participation by community members as active citizens in the ongoing process of engaged argument and collective self-determination is the 'key to the equality of complex equality,' with each citizen 'equally entitled to persuade each other about how we should live our common life' (Reiner, 2011: 100). This societal pluralism extends the range of Walzer's pluralism to encompass ideologies; organizations; religious, ethnic, and national groups; and the international sphere of nation-states.

Walzer revised, added to or modified many of his positions in *Spheres of Justice* in subsequent books. *Interpretation and Social Criticism* (1987) develops the methodology of interpretation used in Spheres. It aims "to provide a philosophical framework for the understanding of social criticism as a social practice" (1987: vii). It also serves as a 'theoretical preamble' (1987) to *Company of Critics* (1988) which applies this perspective to the study of 11 social critics from the United States, Europe, and South Africa from the 1920s to the 1980s. It argues for a process of interpretation that 'accords best with our everyday experience of morality' (1988: 30), and addresses the issue of universalism and particularism. While moral prohibitions 'have been accepted in virtually every society' and 'constitute a kind of minimal and universal moral code' (1988: 24), these provide only a framework for community members to thicken through 'continuous conversation' into 'shared understandings' that develop into a 'moral culture' (1988: 25). Walzer adapts Geertz's concept of thick description here to distinguish universalist from particularist conceptions of morality, and to explain the process by which universal principles become particularist and plural through the continuous, shared conversations of people. While everyone is involved in the interpretation of shared morality, he prefers the 'connected critic' who is 'one of us' and 'who earns his authority' by appealing to 'localized principles' (1988: 39, 40). However, the

interpretations of connected critics have no special authority and must persuade others to prevail. This interpretational pluralism is an ongoing process since there is no single correct interpretation and 'the argument has no end' (1988: 49).

Thick and Thin (1994) extends the argument that there is a thin universal morality and thick cultures in which the meaning of common terms like justice or democracy become particular with each community (1994: 2–5). It is also an example of Walzer's explanatory pluralism accounting for the difference in his focus on 'thin' minimal, human rights in *Just and Unjust Wars*, and 'thick' meanings of distributive justice in specific communities in *Spheres of Justice*. *Thick and Thin* also provides an account of individuals as divided selves in terms of multiple roles, identities, and values (1994: 85). This individual pluralism is better suited for our complex, 'thickly differentiated,' culturally particular societies by allowing more opportunities to act on the various aspects of our divided selves in the spheres of society in everyday life (1994: 101–103).

Politics and Passion: Toward a More Egalitarian Liberalism (2004) is Walzer's most recent revision to *Spheres of Justice*. It focuses on complex equality within a 'domestic society' and critiques liberalism for not adequately accounting for the complexity of people's ordinary lives within a complex society (2004: 140). Its account of associational life does not acknowledge social group, cultural, political, and moral 'involuntary constraints' that many people willingly accept as obligations in their everyday lives (2004: 3–11, 20). Its belief that power is dispersed throughout society downplays 'categorical inequality' that systematically excludes dominated social groups from societal goods (2004: 33). This modifies Walzer's ideal of separated and balanced spheres that distribute goods through internal principles resulting in complex equality at the individual and group levels of society. Addressing this systematic inequality requires both liberal emancipation of individuals from excluded groups, and multicultural empowerment of these groups with the state providing the needed goods to take care of their communities (2004: 33, 37). This combination of emancipation and empowerment more widely disperses power throughout society, and addresses systematic inequality better than liberalism (2004: 41).

After the fall of Soviet Communism, Walzer became interested in the role of civil society movements in Eastern Europe. This is reflected in *Pluralism, Justice and Equality* (1995a), a coedited book with David Miller that includes chapters by critics and supporters of Walzer's theory of complex equality in *Spheres of Justice*, and Walzer's response to these including strengthening the roles of the state and citizen activities in civil society, and in the edited book, *Toward a Global Civil Society* (1995b). He defines civil society as "the space of uncoerced human association and also the set of relational networks – formed for the sake of family, faith, interest, and ideology – that fill this space" (2007: 115). Civil society encourages democratic participation by citizens in associational life, increasing their capacity for self-determination, while satisfying their diverse selves and need for sociability (2007: 123). He claims that the 'strength and vitality of our associations' is 'intimately connected' to the 'quality of our political and economic activity and of our national culture' (2007). Civil society helps to rebalance the dominance of 'market imperialism' and reduce unequal outcomes (2007: 125), becoming a central element in balancing the spheres in societal pluralism. However, "civil society, left to itself, generates radically unequal power relationships, which only state power can challenge" (2007: 129), using its "political power to redistribute resources and to underwrite and subsidize the most desirable associational activities" (2007: 131).

Walzer also became more interested in the relationship between civil society and multiculturalism, particularly in the American context, during the 1990s. *What It Means to Be an American* (1992) examines 'the politics of difference in the United States' (1992: 3). For Walzer, ethnic, religious, and national identities, not class, are the main group commitments across the world (1992: 57). However, he believes "ethnic pluralism is entirely compatible with the existence of a unified republic" (1992: 66) through nonethnic, civic citizenship and a very specific activist state. As in civil society, the state should provide resources through multicultural policies to ethnic, religious, and national groups to strengthen associational life; it should regulate cooperation between different groups to maintain a 'pluralist civil society' based on 'peaceful coexistence' (1992: 6, 7). While the state should not 'foster a particular identity' (1992: 77), it should help to create a unified national identity, for example, through a civics curriculum in the public schools that teaches the history and practice of democracy, including pluralism, in the United States (1992: 17, 18; Reiner, 2011: 172).

On Toleration (1997) continues the discussion of these topics from a conceptual perspective. Walzer argues toleration is made possible by "the peaceful coexistence of groups of people with different histories, cultures, and identities" (1997: 2) and that peaceful coexistence is "clearly an important and substantive moral principle" (1997: 2, 6). However, there are no authoritative moral conclusions that can be drawn from philosophical discussions of toleration, and no form of peaceful existence that is universally valid (1997: 1, 2). Instead, principles of toleration are 'relative to the history and culture of the people whose lives it will arrange,' although there is a thin universal morality of 'minimal coexistence' based on the rights to life and liberty (1997: 4). Based on historical case, he abstracts five 'ideal versions' of regimes of toleration: multinational empires, international society, consociations, nation-states, and immigrant societies (1997: 3, 14–36). At the end of the book, he returns to his goal to meld the best of liberalism, socialism, communitarianism, and other ideologies (1997: 112) into a coherent 'political creed' of democratic socialism (1997) While constantly debatable, this creed generates policies for practical action by committed public intellectuals and citizen activists to achieve through democratic persuasion of their fellow citizens. Multiculturalism in this case is 'a program for greater social and economic equality' that combines a 'defense of group differences and an attack upon class differences' through individual, community, and state action (1997: 111).

The impact of Walzer's work can be quantitatively assessed to a certain extent in terms of awards, recognition, and book sales. Thus, Walzer was awarded the Spinoza Lens in 2008, a prestigious biannual prize for ethics in the Netherlands. He was also recognized as the 61st most prominent public intellectual in 2008 by *Foreign Policy* and *Prospect* magazines (Reiner, 2011: 1). Furthermore, the combined sales of his book, *Just and*

Unjust Wars, approximately equal the total sales of all of his other books (Reiner, 2011: 176). It is more difficult to assess the qualitative impact of his work given his prolific output of books, chapters, articles, essays, and reviews over almost 60 years beginning in 1956. It is also difficult to assess given the range of his topics from war to equality, diversity, and Judaism, the almost continuous reassessment of his own beliefs expressed in his work, the number of significant influences and cumulative effect of these on his work, and the many social roles that Walzer has fulfilled in the family, as a member of the Jewish community, and in work as an activist, editor, professor, and public intellectual. This account of his divided selves suggest that Walzer's impact is more appropriately understood as differentiated according to the answer to the question: Which Walzer for what audience?

The differentiation of Walzer's work across topical spheres is also a product of his rejection of grand theories, and philosophers or others as self-proclaimed authorities. He reaffirmed this position in a recent interview stating "it is useful to have pieces of a theory, but big *meta*-theories more often produce arrogance and authoritarianism in the movement or the state" (Williams, 2012: 6). These pieces of a theoretical approach are evident in the different uses of pluralism in his work, ranging from ideological, organizational, ethnic, religious, national, international, societal, interpretational, explanatory, and individual. However, there is no 'grand' theory of pluralism in a single work that would serve as an introduction and as a focal point to his work. The theory of complex equality in *Spheres of Justice* with its societal pluralism is the closest to a grand theory that Walzer has published. Yet, even without a grand theory, the impact of Walzer's work can be assessed in a more direct, personal way: as an exemplar of a humble man, who no matter how right he thinks he is, allows others to be wrong, and tries to persuade them to his point of view, while not intervening unless life and liberty are at stake. In the long term, it may be that Walzer's impact is in the principle and practice of heroic noninterference as an ethical model to live a good life with others.

See also: Capitalism; Civil Society, Concept and History of; Communitarianism; Democracy: Normative Theory; Equality: Philosophical Aspects; Ethics and Values; Judaism; Justice: Philosophical Aspects; Liberalism: Historical Aspects; Nations and Nation-States in History; Pluralism and Tolerance; Rawls, John (1921–2002); Social Inequality in History (Stratification and Classes); Socialism: Historical Aspects; State, History of.

Bibliography

Kazin, M., Spring 2013. A decent leftist. Dissent 17.
Margalit, A., Spring 2013. Liberal or social democrat. Dissent 21–25.
Morton, B., Spring 2013. The man who knows himself. Dissent 19–21.
Nagel, T., 2013. Foreword. In: Nozick, R. (Ed.), Anarchy, State and Utopia. Basic Books, New York, pp. xi–xviii.
Nussbaum, M.C., Spring 2013. The Mensch. Dissent 18.
Reiner, J.T.D., 2011. Political Thought and Political Action: Michael Walzer's Engagement with American Radicalism. Department of Political Science, University of California, Berkeley (Doctoral Dissertation). http://escholarship.org/uc/item/4x10110b (accesssed 07.07.14.).
Tracy, M., 2013. The dean of "dissent": Michael Walzer on drones, democracy, and his retirement. New Republic. http://www.newrepublic.com/article/112875/michael-walzer-retires-dissent (accessed 07.07.14.).
Walzer, Michael, 1965. The Revolution of Saints: A Study in the Origin of Radical Politics. Harvard University Press, Cambridge, MA.
Walzer, Michael, 1970. Obligations: Essays on Obedience, War and Citizenship. Harvard University Press, Cambridge, MA.
Walzer, Michael, 1971. Political Action. Quadrangle Books, New York.
Walzer, Michael, 1974. Regicide and Revolution. Harvard University Press, Cambridge, MA.
Walzer, Michael, 1977. Just and Unjust Wars. Basic Books, New York.
Walzer, Michael, 1980. Radical Principles. Basic Books, New York.
Walzer, Michael, 1983. Spheres of Justice. Basic Books, New York.
Walzer, Michael, 1985. Exodus and Revolution. Basic Books, New York.
Walzer, Michael, 1987. Interpretation and Social Criticism. Harvard University Press, Cambridge, MA.
Walzer, Michael, 1988. The Company of Critics. Basic Books, New York.
Walzer, Michael, 1992. What It Means to Be an American. Marsilio, New York.
Walzer, Michael, 1994. Thick and Thin: Moral Argument at Home and Abroad. Notre Dame Press, Notre Dame, IN.
Walzer, Michael, 1995a. Pluralism, Justice, and Equality, with Miller, D. Oxford University Press, Oxford.
Walzer, Michael (Ed.), 1995b. Toward a Global Civil Society. Berghahn Books, New York.
Walzer, Michael, 1997. On Toleration. Yale University Press, New Haven, CT.
Walzer, Michael, 2000. The Jewish political tradition. In: Lorberbaun, M., Zohar, N., Lorberbaum, Y. (Eds.), Authority, vol. I. Yale University Press, New Haven, CT.
Walzer, Michael, 2003. The Jewish political tradition. In: Lorberbaun, M., Zohar, N., Lorberbaum, Y. (Eds.), Membership, vol. II. Yale University Press, New Haven, CT.
Walzer, Michael, 2004. Arguing about War. Yale University Press, New Haven, CT.
Walzer, Michael, 2005. Politics and Passion: Toward a More Egalitarian Liberalism. Yale University Press, New Haven, CT.
Walzer, Michael, 2006. In: Law, Politics, and Morality in Judaism. Princeton University Press, Princeton.
Walzer, Michael, 2007. Thinking Politically: Essays in Political Theory. Yale University Press, New Haven, CT.
Walzer, Michael, 2009. In: Mills, Nicolaus (Ed.), Getting Out: Historical Perspectives on Leaving Iraq. University of Pennsylvania Press, Philadelphia, PA.
Walzer, Michael, 2012. In God's Shadow: Politics in the Hebrew Bible. Yale University Press, New Haven, CT.
Williams, J.J., 2012. Michael Walzer's politics, in theory and practice. The Chronicle of Higher Education. http://chronicle.com/article/Michael-Walzers-Politics-in/132041 (accessed 07.07.14.).

Relevant Websites

http://www.dissentmagazine.org/blog/dissents-sixtieth – Dissent's Sixtieth|Dissent, (accessed 07.07.14.).
http://www.nybooks.com/articles/archives/2014/mar/20/imaginary-jews – Imaginary Jews by Michael Walzer|The New York Review of Books, (accessed 07.07.14.).
http://www.ias.edu/people/faculty-and-emeriti/walzer – Michael Walzer|Institute for Advanced Studies, (accessed 07.07.14.).
http://bigthink.com/videos/michael-walzer-who-are-you – Michael Walzer: Who Are You?, (accessing 07.07.14.).
http://www.huffingtonpost.com/michael-walzer/liberal-zionists-speak-out-state-of-righteousness_b_1447261.html – The State of Righteousness: Liberal Zionists Speak Out|Huff Post Religion, (accessed 07.07.14.).

Wanting vs Needing

Mike JF Robinson, Wesleyan University, Middletown, CT, USA
Kent C Berridge, University of Michigan, Ann Arbor, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Wanting (incentive salience) is a psychological process, triggered by rewards and their cues or imagery, which normally provides an 'oomph' that spurs motivation for objects of desire. Typically 'wanting' is in conformity with significant needs. But 'wanting' can sometimes dissociate from needs and from deliberative plans in maladaptive ways. Most dramatically in drug addiction, when amplified excessively by changes in brain reactivity, 'wanting' can have pathological intensity. Other impulse control disorders, such as gambling and overeating, may possibly share a similar compulsive 'wanting.' And even in ordinary daily life, occasional dissociations of 'wanting' from needs might occur. Selective surges in 'wanting' may explain why under these conditions, rewards can be 'wanted' much more than they are needed or even 'liked.'

Want versus Need

In a classic study of delayed gratification, Mischel and colleagues (Mischel et al., 1972, 1989) examined the ability of young preschool-aged children to patiently refrain from consuming a tasty marshmallow placed in front of them, when promised that if they could wait, a second marshmallow would additionally be received when the experimenter returned later. Some children succeeded in resisting the tempting marshmallow in order to earn the extra one, whereas others succumbed to the immediate temptation. Subsequent findings revealed that children who displayed the greatest amount of patience later went on to have better life outcomes as measured by SAT scores, educational attainment, body mass index, and other life measures.

What happens at such moments of temptation? What is the nature of the temptation process that self-control must wrestle with? Why is a marshmallow one can see and smell more difficult to resist than one in the next room? We suggest that one important psychological process in temptation is incentive salience, also known as 'wanting.' Incentive salience is typically triggered by encounters with reward-related cues (or related vivid imagery), and experienced as surges of motivation to obtain and consume the reward that can last even beyond the presence of the cues or imagery. Incentive salience is generated by brain 'mesolimbic' systems that react to reward cues, involving dopamine and other signals that converge on the nucleus accumbens. The intensity of the 'wanting' surge is determined by the reactivity of brain mesolimbic systems at the moment of cue reencounter. This motivation intensity may reach heights that become nearly compulsive if the brain systems are especially highly reactive, which can happen in certain states (e.g., in stress, emotional excitement, relevant appetites) and especially in some individuals (e.g., addicts). Incentive salience is an adaptive psychological response that is crucial to normal life. But sometimes it can go awry, in ways small or large, with consequences for well-being.

The ability to delay gratification in the face of temptation is thought to reflect the sensitivity of an individual's reward system. There frequently are moments of temptation in life for everyone, where temptations can sometimes drive individuals to act in ways that may not best fit their needs. The brain's reward system is designed to make people usually want what they need, but given the right conditions they are prone to wanting in the absence of any need.

Most commonly, intense dissociations between wants and needs may occur when brain reward systems and the attribution of 'wanting' (incentive salience) become hypersensitive. This can happen to moderate degrees temporarily in ordinary life under conditions of appetite, stress, or emotional arousal. Greater degrees of hypersensitivity can occur in more extreme conditions, the most dramatic case perhaps being drug addiction. This is referred to as incentive sensitization and results from repeated assaults to our reward system by addictive substances such as drugs of abuse, highly palatable foods, repetitive gambling, or even chronic stress. Following sensitization of their reward system, an individual may be more vulnerable to the magnetic properties of reward cues, such as the smell of brownies or coffee brewing, the sight of a regular watering hole or a needle for an addict, or the flashing lights of a slot machine in a compulsive gambler. All of these cues will trigger powerful wants that are often separate from an individual's needs, and require the exertion of regular cognitive control to maintain behavior on an appropriate course. In most individuals this may mean merely having a second brownie when you know one was enough, or having another drink at the end of the night. In recovering drug addicts however, this may mean undermining the intention to stay drug free and potentially lead to relapse.

Drug Addiction, Recovery, and Relapse

Addiction may provide a potent example of how incentive salience can rise to compulsive levels in some individuals. Relatively few individuals are susceptible to the degree of elevation in 'wanting' high enough to produce compulsive addiction under typical conditions. For example, most adults have used a potentially addictive drug at least once in their lifetime, if caffeine, alcohol, and nicotine are included in addition to illicit drugs. Yet most do not become addicts. For those who do, none of these drugs satisfy a true need except

when an individual is in pronounced state of withdrawal. Addiction is characterized by compulsive drug seeking, an impairment of social and psychological functions, and/or damage to one's health. Addiction typically involves overwhelming involvement with the addictive reward, a loss of control, and a narrowing of interests. According to a 2010 survey, less than 10% of the US population met the criteria for chronic alcohol or drug abuse disorder, and an even smaller proportion suffered from chronic addiction (Substance Abuse and Mental Health Services Administration, 2011).

A chief problem in treating addiction is chronic or repeated relapse among those who are trying to quit. Even after prolonged periods of withdrawal and abstinence, a high percentage of addicted individuals in treatment programs eventually relapse to drug-taking. For example, in the case of a study of heroin users, relapse rates to reuse after cessation were approximately 60% within 3 months and at least 75% within 12 months (Hunt et al., 1971). For this reason, drug addiction is characterized as a chronic relapsing disorder; relapse is the rule rather than the exception, and often occurs repeatedly.

There are three conventional reasons frequently suggested in psychology and neuroscience to explain addiction and relapse: (1) *Drug Euphoria* – that addicts resume drug-taking to experience the intense pleasure (euphoria) they remember the drug producing; (2) *Overlearning Habits or Predictions* – drug-taking becomes such a well-entrenched habit that relapse is almost inevitable, or that learning becomes distorted in other ways to create false predictions about drug rewards (Everitt et al., 2008; Hyman et al., 2006); and (3) *Withdrawal Escape* – that the withdrawal syndrome that accompanies the cessation of drug intake is so unpleasant an addict would do anything to stop it, and so relapse occurs as an escape from withdrawal (Koob and Volkow, 2010). All three of these explanations certainly play a role in relapse. However, we believe they are insufficient to explain the central problem underlying relapse in addiction, for several reasons.

First, drug pleasure or euphoria certainly accounts for the initial pattern of drug use and abuse, but may have more difficulty accounting for relapse as tolerance can develop to the pleasure. Even addicts who do not find their drugs particularly pleasant anymore may experience increases in drug craving that persist a long time.

It also has been suggested by some learning-oriented scientists that the repeated use of drugs creates a learning disorder, such as making drug-taking an overly ritualized habitual act or creating false expectations of exaggerated reward. Ritualization may be true of the act of drug-taking, but cannot explain the preceding flexible acts of drug seeking during craving. And there is little reason to believe that addicts mispredict the reward value of their drugs, or the consequences of their actions. Learned habits or mispredictions alone cannot account for the excessive motivational attraction of addiction.

Finally, many addictive drugs surely do induce tolerance (when drug is present) and withdrawal (when drug is absent). Withdrawal is typically described as an intense negative emotional state accompanied by dysphoria, anxiety, and irritability, and may indeed be a potent reason why many addicts relapse and take drugs, at least while the withdrawal lasts. Yet withdrawal is a relatively short-lived phenomenon, and decays substantially within days to weeks. By contrast, relapse frequently occurs even after withdrawal is no longer reported, and even in fully 'detoxified' addicts, months after 'recovery.'

In contrast to these suggestions, the Incentive Sensitization Theory (Robinson and Berridge, 1993) originally proposed an alternative explanation that can account for the persistence of relapse, and the independence of addiction from pleasure, withdrawal, or faulty expectations. It may also have applications to some addictions that extend beyond drugs, and proposes an explanation for the dichotomy that sometimes exists between wants and needs. The Theory proposes that relapse frequently occurs as a result of brain changes that lead to intense incentive motivation for drugs. These brain changes generate pulses of incentive salience or 'wanting,' often triggered by encountering drug cues, which may be experienced as feelings of drug craving or may even control behavior implicitly without need of strong conscious feelings. Craving occurs when the process of incentive salience (or core 'wanting'), mediated primarily by subcortical mesolimbic brain systems that use dopamine as an important neurotransmitter, is translated into conscious awareness. At its core, the motivation to take drugs is due to the overattribution of incentive salience to drug-related stimuli. It is important to note that incentive salience is a distinct psychological process from withdrawal and drug pleasure. In some cases, attribution of incentive salience to reward-predicting cues may make the cues 'wanted' as much as the reward itself, irrespective of whether the reward constitutes an actual need. Such cues become motivational magnets, sometimes prompting irrational behavior, such as interactions with the cue specific to those previously seen only during interactions with the reward itself (Davey and Cleland, 1982). As example of irrational cue attraction, crack cocaine addicts can be found inspecting the floor for a white speck that is more likely to be an ordinary pebble than crack cocaine, and can be attracted to pick it up, inspect, and put it in the pipe and even try to light or smoke the noncocaine pebble (a phenomenon that has been called 'chasing ghosts') (Rosse et al., 1993).

Incentive Salience and Utility

In applying incentive salience to daily life choices, it can be useful to consider such explanations in terms of forms of reward *utility* that are important to decision-making: predicted utility, decision utility, experienced utility, and remembered utility (Kahneman et al., 1997). All these forms of utility are critical in the process that determines whether a reward is truly needed, or if it is simply wanted. Predicted utility is an expectation of how much a future drug reward will be liked. Decision utility is the valuation of the drug manifest in choice and pursuit. Experienced utility is how much the pleasant drug is liked when actually taken. Remembered utility is the memory of how pleasant the drug was in the past. Experienced utility is considered the end point of the decision process. It is the state reached after successful attainment of a particular outcome, pertaining to the hedonic evaluation of that outcome. Experienced utility informs to some degree both remembered and predicted utility. However, other signals are needed in order for decisions to actually be made. The incentive sensitization theory suggests that only one of these, decision utility, of which incentive salience is one constituent, need to be distorted to create

a compulsive addiction. If decision utility fails to clearly distinguish between what is wanted and what is actually needed, then addictive and compulsive behavior is given the opportunity to proliferate, with behavior that is persistently guided by urges and cravings rather than cognitive goal-directed action.

Typically, decision utility is determined by predicted and remembered utility. However, predicted and remembered utility may fail to be a perfect representation of experienced utility, because hedonic memories can become distorted, such as when peak-end averages of a hedonic experience outweigh in memory the actual amount of pain or pleasure that was experienced (Kahneman et al., 1997). Drug addicts are also believed to often fail to accurately translate experienced utility into decisions (Bechara, 2005). In general, any distortion in memory or prediction that leads to faulty predictive utility will likewise affect decision utility, producing decisions that fail to maximize experienced utility of chosen outcomes.

However problems in excessive decision utility for addicts and some others facing strong cue-triggered temptations may remain even when their remembered utility and predicted utility for drug consequence are quite accurate. This is where there is a special role for incentive salience, because of the ability of brain mesolimbic systems to become especially highly reactive under some conditions (Berridge, 2012; Berridge and Robinson, 1998). Although 'wanting' typically coheres with 'liking' (hedonic impact) for the same reward, 'wanting' and 'liking' can be dissociated in certain circumstances and by some manipulations, especially those that specifically involve dopamine. Finally, 'wanting' can also be distinguished from learning and predictions about the same reward. For example, 'wanting' triggered by a Pavlovian reward cue can dramatically increase motivation for the reward, even if its previously learned value has not changed (e.g., in hunger, satiety, stress, or drug-related states). Abstinence from smoking for only 24 h can dramatically potentiate neural responses to smoking-related cues (McClernon et al., 2009). This also means that in instances where needs have been satisfied, wants may still occur, particularly when triggered by reward- or drug-related cues.

In this framework, incentive salience 'wanting' is a pure form of decision utility, which is distinct from other forms of utility and in some conditions can decouple from all the others (Berridge and Aldridge, 2008). That is, 'wanting' for an outcome is distinguishable from both experienced utility (hedonic impact or 'liking' the outcome), remembered utility of how nice the outcome was in the past, and anticipated or predicted utility of how nice it will be in the future. For incentive salience, under conditions of dopamine-related stimulation, situations exist where cue-triggered decision utility > remembered utility from the past, and similarly decision utility > predicted utility for future reward value. In other words, it is possible to addictively 'want' what is not expected to be liked, nor remembered to be liked, as well as what is not actually liked when obtained. All of these situations apply to rewards that are needed but become infinitely more problematic when they relate to rewards that satisfy no need and are purely wanted.

Such intense 'wants' may be triggered especially at particular moments, on particular encounters with addictive cues but not on other encounters with the same cues. According to the incentive sensitization theory of addiction (Robinson and Berridge, 1993), attribution of incentive salience to a reward cue becomes exaggerated in addicts due to mesolimbic brain changes that are very long lasting (Paulson et al., 1991). Exaggeration can happen because incentive salience, which makes up part of decision utility, always results from the synergy between two sources: previously learned associations about the reward cue and the current brain state at the time of cue encounter (Berridge, 2012; Zhang et al., 2009). In the case of needs, that brain state can be driven by hunger, thirst, or even salt deficiency, but where drugs are concerned, the presence of drug-related cues or small amounts of the drug itself can spur powerful wants in the form of craving.

How dramatically can raising brain mesolimbic reactivity make 'wanting' detach from, and soar above, memories or learning-based predictions of value? So much so that even a repulsive learned cue for unpleasant sensations can become suddenly and intensely 'wanted' via the activation of mesocorticolimbic circuitry. This can sometimes be useful and meet adaptive needs, which is why the psychological process and brain mechanism of incentive salience evolved. For example, we have shown that rats which learned a revulsion toward a Pavlovian cue for unpleasant Dead Sea saltiness, will on their first reencounter with that cue in a novel salt-need state activate their mesolimbic circuitry in a way that generates intense 'wanting,' instantly transforming the cue into an attractive and powerful motivational magnet that is jumped on and gnawed (Robinson and Berridge, 2013). While adaptive in that case, similar powerful brain transformations may be recruited to varying degrees by addictive drugs, other motivational consumption disorders, and even ordinary emotional states that heighten mesolimbic reactivity to reward cues (including marshmallows). This cue/state transformation helps define what it means to say that addiction hijacks brain limbic circuits of natural reward. Finally, different individuals may be differentially vulnerable to such mesolimbic transformations that generate such moments of intense 'wanting.' Fluctuations in the temptation power for cues, which illustrates the difference between decision utility and predicted utility, hinge on the current neurobiological state factors related to dopamine at the moment of cue encounter. In particular, incentive sensitization suggests that craving and relapse are magnified by a sensitized neural system (mesocorticolimbic dopamine and related systems), which can flip into a super-reactive mode under several conditions. Those include especially states of stress, or if a person tries to 'just take one' hit of their addictive drug (that primes mesolimbic systems to react more powerfully to cues), or in other emotional states that heighten mesolimbic reactivity. At such moments, ordinary stimuli, such as cues associated with rewards, are transformed into potent incentive stimuli, making them attractive and able to trigger an urge to pursue and consume their reward. Under these situations, an individual is the most vulnerable to expressing wants that are most distally removed from true needs.

This type of synergistic modulation of 'wanting' is not limited to addicts. Everyone has experienced at least moderate pulses of incentive salience in their lives that are generated by similar rules. For example, advertisements that pop up on a Web page may prompt the finger to click onto the product. The smell of food as you walk down the street may make you

suddenly feel quite hungry, even if you were not feeling that way moments earlier. But the smell of food as a cue is not constant in its temptation power: if you really have not eaten all day you might find the aroma extremely tempting, whereas you would not if you have recently eaten. The essence of incentive sensitization suggests that addicts encounter fluctuations like this in the temptation power of their drug-associated cues, but that their maximal peaks of temptation are much higher than other people are liable to experience in daily life because of the enduring sensitization of their mesolimbic systems.

The basic mechanisms of the excessive attribution of incentive salience to drugs and drug-related stimuli can even occur as a mostly automatic and unconscious process, creating urges to take drugs whether or not a strong subjective feeling of craving is simultaneously present. Such dissociation between acted-on motivation and confusing subjective feelings is what often renders the compulsive quality of an addict's own behavior astonishing even to him or her. Here, unconscious wants would have the ability to override our cognitive pursuit of needs, sometimes emerging as a sense of loss of control.

Only 'wanting,' not 'liking,' becomes sensitized and consequently more intense on its own, as addiction develops. That is because 'liking' or pleasure has separable, and more restricted brain mechanisms (Berridge and Kringelbach, 2013; Smith et al., 2010). Thus elevated 'wanting' can detach from normal 'liking.' For example, in animal studies, sensitization increases neuronal firing in pathways that code incentive salience as well as the behavioral ability of reward cues to trigger frenzied bursts of effort to obtain the reward (Tindell et al., 2005; Wyvell and Berridge, 2001). Yet sensitization does not increase 'liking' reactions that reflect the hedonic impact of the reward when it actually arrives. Similarly, in humans who are becoming drug-tolerant addicts, incentive motivation to take the drug can grow as they become addicted, so that a single hit of drug can provoke intense urges to take more, even if the person reports the dose of drug no longer gives as much pleasure as initially.

Other Forms of Excessive Wanting Beyond Drugs

Binge Eating and Food Addiction

Beyond purely drug addiction, consequent incentive sensitization may also manifest itself by food bingeing, pathological gambling, hypersexuality, and other compulsive motivations. Unlike drugs, food fulfills a bodily need, yet it can still be wanted to excessive levels. Overeating is a chief cause of growing obesity trends. Could exaggerated 'wanting' or 'liking' play a role in some individuals? Excessive hedonic reactions to food would magnify both 'liking' and 'wanting' above that of a regular individual, thus contributing to binge eating and obesity (Berridge et al., 2009; Davis et al., 2009). Alternatively, changes in 'wanting' alone could be responsible for overeating. Sensitization of mesolimbic dopamine systems by exposure to cycles of bingeing alternating with dieting has been suggested to occur (Avena and Hoebel, 2003a,b). Enhanced sensitivity of the mesolimbic reward system could attribute high levels of incentive motivation to the sights and smells related to food and drive excessive consumption, without necessarily producing comparable levels of 'liking' when the food is itself consumed.

A different set of problems may face some other individuals, who have been suggested to have elevated 'liking' as well as 'wanting' that drives binge eating and overconsumption, and possibly deserve the label of food addicts (Davis et al., 2009; Davis and Carter, 2009). Notably Davis and colleagues have found that certain individuals who are both obese and binge eaters are far more likely to carry both alleles for a gene that codes a gain of function for mu-opioids, and another allele for a gene that may be associated with higher binding for the dopamine D2 receptor. Together these genetic traits have been suggested by Davis and colleagues to combine to simultaneously increase 'liking' and 'wanting' for foods in a manner that strongly promotes binge eating and gives rise to addictive like features, including loss of control and relapse. This may be a rare example where 'wanting' and 'liking' increase in tandem, making it exceptionally hard to adhere to levels of food intake that conform to physiological needs. Similarly, it has been suggested that individuals who carry genes promoting elevated dopamine function might experience stronger cue-triggered urges in response to food cues, making them more liable to develop obesity (Campbell and Eisenberg, 2007).

Conversely, it has been suggested that anorexia nervosa, involving a failure to eat, is related to a reward dysfunction that suppresses the 'wanting' for food despite leaving the 'liking' portion intact (as well as developing an abnormal 'wanting' for body perceptions of self as thin) (Keating et al., 2012). This could be interpreted as a case of insufficient 'want' that fails to achieve the motivation required to fulfill basic needs.

Gambling

Gambling might also involve special recruitment of incentive salience brain systems. Uncertainty may especially promote incentive salience under some conditions, which mirror many of the hallmarks of gambling (Anselme et al., 2013; Linnet et al., 2012; Lobo et al., 2010; van Holst et al., 2010). It has even been shown that not knowing how hard one must work to be rewarded may even sensitize the brain in ways similar to addictive drugs (Singer et al., 2012). This may produce a further example of the dissociation between experienced or remembered utility and decision utility. Individuals sometimes seem driven by cues to gamble, in all cases at a global monetary loss, for only a moderate experience utility. It seems that despite a conscious awareness that, for example, slot machine gambling will repeatedly end with a loss, showing a cognitive awareness that playing will fail to fulfill any monetary need, players want to gamble, sometimes expressly to satisfy an urge to play. Compulsive gamblers may also show other addictions (Zhang et al., 2009).

Consumer Behaviors

Even beyond addiction, situations may arise in everyday life in which incentive salience becomes particularly high to induce moments of strong decision utility to pursue or consume an incentive. Consumer goods can similarly be strong incentives influenced by powerful cues. Some situational factors, like being

in a store, might possibly increase how much consumer goods are 'wanted' while the extent to which they are 'liked' remains unchanged (Litt et al., 2010). Purchasing a product without actually 'liking' the product sufficiently to make the purchase under normal circumstances is likely to induce regret. Accordingly, Litt et al. (2010) suggested that 'wanting' and 'liking' for consumer goods can be driven in opposite directions, almost mimicking reports of drug addiction. When people experienced failure in pursuing desired outcomes (e.g., a 5$ gift card), individuals showed increased willingness to pay for this reward, i.e., 'wanted' the reward more. However, these individuals also happened to 'like' the reward less in the sense that they were likely to more often trade it away for an equivalent prize (5$ gift card from a similar store). Such instances show how consumer goods can be wanted more than they are needed. In cases of excessive consumerism or for people who display hoarding behavior, motivation to possess becomes tantamount and need becomes entirely irrelevant.

Summary

'Wanting' as a process and incentive sensitization helps explain fundamental problems such as addiction and, compulsive pursuits, as well as more mundane temptations such as a dangling marshmallow. In such moments of temptation, cognitive needs and intentions may be set aside either momentarily or sometimes more permanently, to satisfy intense cravings for incentives that are often more 'wanted' than they are needed. Incentive salience operates alongside more cognitive processes of planning and goals that may be aligned to meet major life needs. Often incentive salience operates in congruence with those needs, but not always. When 'wants' and needs diverge, temptations can arise of varied intensities. While temptation can be controlled, that control presents a challenge. Even an addict remains in principle capable of resisting temptation on any single trial, but in practice is likely to succumb to relapse if required to encounter a series of repeated temptations.

Powerful levels of 'wanting,' often in the absence of equivalent 'liking,' are not restricted to addictive drugs. The incentive sensitization theory potentially provides an explanation for why consumer goods and certain foods can also become excessively compelling incentives. In turn, ever-present cues, such as those incorporated in advertisements, can trigger potent 'wanting' peaks that propel susceptible individuals toward overconsumption of material goods and addiction-like disorders such as binge eating or gambling.

Acknowledgments

This article was adapted from an earlier article by Robinson M.J.F., Robinson T.E., and Berridge K.C., *Incentive Salience in Addiction and Overconsumption*, published by MIT Press. Preparation of this article, and research described from our lab, was supported by National Institutes of Health grants DA015188 and MH63649.

See also: Decision Making: Nonrational Theories; Motivation and Actions, Psychology of; Pathological Gambling and Gambling Disorder.

Bibliography

Substance Abuse and Mental Health Services Administration, 2011. Results from the 2010 national survey on drug use and health: summary of national findings. Substance Abuse and Mental Health Services Administration 8, 1458–1463. http://dx.doi.org/10.1038/nn1584.

Anselme, P., Robinson, M.J.F., Berridge, K.C., 2013. Reward uncertainty enhances incentive salience attribution as sign-tracking. Behavioural Brain Research 238, 53–61. http://dx.doi.org/10.1016/j.bbr.2012.10.006.

Avena, N.M., Hoebel, B.G., 2003a. Amphetamine-sensitized rats show sugar-induced hyperactivity (cross-sensitization) and sugar hyperphagia. Pharmacology, Biochemistry, and Behavior 74 (3), 635–639.

Avena, N.M., Hoebel, B.G., 2003b. A diet promoting sugar dependency causes behavioral cross-sensitization to a low dose of amphetamine. Neuroscience 122 (1), 17–20.

Bechara, A., 2005. Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. Nature Neuroscience 8 (11), 1458–1463. http://dx.doi.org/10.1038/nn1584.

Berridge, K.C., 2012. From prediction error to incentive salience: mesolimbic computation of reward motivation. European Journal of Neuroscience 35 (7), 1124–1143. http://dx.doi.org/10.1111/j.1460-9568.2012.07990.x.

Berridge, K.C., Aldridge, J.W., 2008. Decision utility, the brain, and pursuit of hedonic goals. Social Cognition 26 (5), 621–646. http://dx.doi.org/10.1521/soco.2008.26.5.621.

Berridge, K.C., Kringelbach, M.L., 2013. Neuroscience of affect: brain mechanisms of pleasure and displeasure. Current Opinion in Neurobiology 23 (3), 294–303. http://dx.doi.org/10.1016/j.conb.2013.01.017.

Berridge, K.C., Robinson, T.E., 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research. Brain Research Reviews 28 (3), 309–369.

Berridge, K.C., Robinson, T.E., Aldridge, J.W., 2009. Dissecting components of reward: 'liking', "wanting," and learning. Current Opinion in Pharmacology 9 (1), 65–73. http://dx.doi.org/10.1016/j.coph.2008.12.014.

Campbell, B.C., Eisenberg, D., 2007. Obesity, attention deficit-hyperactivity disorder and the dopaminergic reward system. Collegium Antropologicum 31 (1), 33–38.

Davey, G.C., Cleland, G.G., 1982. Topography of signal-centered behavior in the rat: effects of deprivation state and reinforcer type. Journal of the Experimental Analysis of Behavior 38 (3), 291–304.

Davis, C.A., Levitan, R.D., Reid, C., Carter, J.C., Kaplan, A.S., Patte, K.A., et al., 2009. Dopamine for "wanting" and opioids for "liking": a comparison of obese adults with and without binge eating. Obesity (Silver Spring, Md.) 17 (6), 1220–1225. http://dx.doi.org/10.1038/oby.2009.52.

Davis, C., Carter, J.C., 2009. Compulsive overeating as an addiction disorder. A review of theory and evidence. Appetite 53 (1), 1–8. http://dx.doi.org/10.1016/j.appet.2009.05.018.

Everitt, B.J., Belin, D., Economidou, D., Pelloux, Y., Dalley, J.W., Robbins, T.W., 2008. Neural mechanisms underlying the vulnerability to develop compulsive drug-Seeking habits and addiction. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 363 (1507), 3125–3135.

van Holst, R.J., van den Brink, W., Veltman, D.J., Goudriaan, A.E., 2010. Why gamblers fail to win: a review of cognitive and neuroimaging findings in pathological gambling. Neuroscience and Biobehavioral Reviews 34 (1), 87–107. http://dx.doi.org/10.1016/j.neubiorev.2009.07.007.

Hunt, W.A., Barnett, L.W., Branch, L.G., 1971. Relapse rates in addiction programs. Journal of Clinical Psychology 27 (4), 455–456.

Hyman, S.E., Malenka, R.C., Nestler, E.J., 2006. Neural mechanisms of addiction: the role of reward-related learning and memory. Annual Review of Neuroscience 29, 565–598. http://dx.doi.org/10.1146/annurev.neuro.29.051605.113009.

Kahneman, D., Wakker, P.P., Sarin, R., 1997. Back to Bentham? Explorations of experienced utility. The Quarterly Journal of Economics 112 (2), 375–406.

Keating, C., Tilbrook, A.J., Rossell, S.L., Enticott, P.G., Fitzgerald, P.B., 2012. Reward processing in anorexia nervosa. Neuropsychologia 50 (5), 567–575. http://dx.doi.org/10.1016/j.neuropsychologia.2012.01.036.

Koob, G.F., Volkow, N.D., 2010. Neurocircuitry of addiction. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology 35 (1), 217–238. http://dx.doi.org/10.1038/npp.2009.110.

Linnet, J., Mouridsen, K., Peterson, E., Møller, A., Doudet, D.J., Gjedde, A., 2012. Striatal dopamine release codes uncertainty in pathological gambling. Psychiatry Research 204 (1), 55–60. http://dx.doi.org/10.1016/j.pscychresns.2012.04.012.

Litt, A., Khan, U., Shiv, B., 2010. Lusting while loathing: parallel counterdriving of wanting and liking. Psychological Science 21 (1), 118–125. http://dx.doi.org/10.1177/0956797609355633.

Lobo, D.S.S., Souza, R.P., Tong, R.P., Casey, D.M., Hodgins, D.C., Smith, G.J., et al., 2010. Association of functional variants in the dopamine D2-like receptors with risk for gambling behaviour in healthy Caucasian subjects. Biological Psychology 85 (1), 33–37. http://dx.doi.org/10.1016/j.biopsycho.2010.04.008.

McClernon, F.J., Kozink, R.V., Lutz, A.M., Rose, J.E., 2009. 24-h smoking abstinence potentiates fMRI-BOLD activation to smoking cues in cerebral cortex and dorsal striatum. Psychopharmacology 204 (1), 25–35. http://dx.doi.org/10.1007/s00213-008-1436-9.

Mischel, W., Ebbesen, E.B., Zeiss, A.R., 1972. Cognitive and attentional mechanisms in delay of gratification. Journal of Personality and Social Psychology 21 (2), 204–18.

Mischel, W., Shoda, Y., Rodriguez, M.I., 1989. Delay of gratification in children. Science (New York, N.Y.) 244 (4907), 933–938.

Paulson, P.E., Camp, D.M., Robinson, T.E., 1991. Time course of transient behavioral depression and persistent behavioral sensitization in relation to regional brain monoamine concentrations during amphetamine withdrawal in rats. Psychopharmacology 103 (4), 480–492. http://dx.doi.org/10.1007/BF02244248.

Robinson, M.J.F., Berridge, K.C., 2013. Instant transformation of learned repulsion into motivational "wanting". Current Biology 23 (4), 282–289. http://dx.doi.org/10.1016/j.cub.2013.01.016.

Robinson, T.E., Berridge, K.C., 1993. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Research. Brain Research Reviews 18 (3), 247–291.

Rosse, R.B., Fay-McCarthy, M., Collins, J.P., Risher-Flowers, D., Alim, T.N., Deutsch, S.I., 1993. Transient compulsive foraging behavior associated with crack cocaine use. The American Journal of Psychiatry 150 (1), 155–156.

Singer, B.F., Scott-Railton, J., Vezina, P., 2012. Unpredictable saccharin reinforcement enhances locomotor responding to amphetamine. Behavioural Brain Research 226 (1), 340–344. http://dx.doi.org/10.1016/j.bbr.2011.09.003.

Smith, K.S., Mahler, S.V., Peciña, S., Berridge, K.C., 2010. Hedonic Hotspots: Generating Sensory Pleasure in the Brain. Oxford University Press.

Tindell, A.J., Berridge, K.C., Zhang, J., Peciña, S., Aldridge, J.W., 2005. Ventral pallidal neurons code incentive motivation: amplification by mesolimbic sensitization and amphetamine. The European Journal of Neuroscience 22 (10), 2617–2634. http://dx.doi.org/10.1111/j.1460-9568.2005.04411.x.

Wyvell, C.L., Berridge, K.C., 2001. Incentive sensitization by previous amphetamine exposure: increased cue-triggered "wanting" for sucrose reward. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 21 (19), 7831–7840.

Zhang, J., Berridge, K.C., Tindell, A.J., Smith, K.S., Aldridge, J.W., 2009. A neural computational model of incentive salience. PLoS Computational Biology 5 (7), e1000437. http://dx.doi.org/10.1371/journal.pcbi.1000437.

War and Democracy

Hans Joas, Humboldt University, Berlin, Germany; and University of Chicago, Chicago, IL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article is an attempt to discuss possible causal relationships between 'war' and 'democracy.' One can ask: What is the impact of wars on democracies – and what is the effect of a state's democratic nature on its behavior with respect to war? If we add the distinction between 'socialization' and 'selection,' we realize that four key questions arise. All four questions are discussed here on the basis of theoretical reflections taken from the history of social thought and with reference to contemporary developments.

A schematic breakdown of the key questions on the relationship between war and democracy shows that there are two possible arrows of causality to consider. What is the impact of wars on democracies – and what is the effect of a state's democratic nature on its behavior with respect to war? We can add to these familiar questions the distinction commonly made by sociologists since the reception of Darwin in the nineteenth century and expressed most succinctly by Talcott Parsons in *The Social System* (1951). This is the distinction between two types of possible effect – 'socialization' and 'selection,' the former referring to a process of internal transformation and the latter to the varying probability of survival in particular environments – and it gives rise to four key questions:

1. In a belligerent environment, are democracies more likely to survive because they possess unique strengths or are they more likely to be defeated by their less democratic opponents?
2. Do democracies have specific effects on the nature of armed conflict when they engage in it?
3. Do democracies have a greater capacity for peace than other sociopolitical orders or are there qualities specific to democratic societies that put peace at risk?
4. How do wars affect the inner workings of democracies? Do they promote internal processes of democratization or do they threaten these processes, as well as existing democratic structures?

The picture becomes more complex still when we factor in that we cannot use the term 'war' to imply that the entire spectrum of macrolevel violence is covered by conflicts between clearly defined states, states that in principle recognize one another. If we include civil wars and states' suppression or murder of sections of their own populations, then another question arises, namely whether democracies represent insurance against such violence or whether, conversely, such things as ethnic cleansing in fact represent the "dark side of democracy" as Michael Mann (2005) has argued.

A wealth of answers to all these questions can be found in the history of social theory. This does not mean that we have empirically well-founded knowledge on all these questions at our disposal – quite the reverse. The aim here is to present some key ideas on these questions from the history of our subject, relating them to contemporary realities in such a way that we end up with a reasonably clear overview.

Are Democracies Weak?

Given the events of the twentieth century, the first question is probably the easiest one to answer, namely whether democracies are bound to be outdone militarily by their nondemocratic competitors. This supposition was long considered highly plausible, particularly in Germany. In contrast to clear chains of command and highly centralized power, it inevitably seemed that the separation of powers and the complex, necessarily controversial, time-consuming democratic approach to formulating objectives and making decisions could never offer much competition. During World War I, the great sociologically inclined historian Otto Hintze argued, for example, that Germany could not afford to become a democratic republic because of its geographical position in the middle of the European continent. The idea behind this was that only states in an exceptionally secure geopolitical position such as Britain and the USA could in principle do without a standing army. As standing armies always required centralized state administrations and could easily become tools of repression in the hands of monarchs and other rulers, Hintze believed there was an affinity between absolutism and external threats on the one hand, and parliamentarization and a secure position on the other. This idea was then trivialized in anti-British polemics. The phantasm of the *Führerstaat* was so central to fascism and National Socialism that it went hand in hand – as in Hitler's view of America – with a suicidal underestimation of the USA. World War II brought a definitive end to the tendency to underestimate the Anglo-Saxon democracies' capacity for mobilization and staying power. During the Cold War, only a few concerned liberals (such as Samuel Huntington in 1957) suggested that democracies were fundamentally inferior, but not the liberals' opponents. Communist leaders generally expected capitalism to collapse as a result of economic crises and their own technological superiority. And of course they also laid claim to the democratic idea in order to legitimize their own 'people's democracies.' Attempts to resolve empirically the question of the robustness of democracies in war – the most important being the study *Democracies at War* by Dan Reiter and Allan Stam (2002) – conclude that

since 1816 democracies have won three-quarters of the wars in which they were involved. In fact, according to these authors, this success rate increases if we look only at wars begun by democracies; in this case – according to these figures – the success rate is no less than 93%. Dictatorships, meanwhile, lost 4 out of 10 of the wars they had begun. Even when attacked, democracies were more often successful than dictatorships (63 vs 34%). Both these figures and their interpretation are subject to dispute. If the democratic societies are also the economically and militarily superior ones, then their success is not necessarily due directly to the fact they are democratic. Particularly in light of the USA's immense power and leading role, simply finding data that allow us to factor these realities out is a very difficult task. The advocates of the thesis that democracies are more successful in wars reject this criticism by arguing that it reinforces rather than refutes their interpretation of the evidence. This is because they assume that democracies enter into a war of their own accord only if they seem assured of victory. This, they suggest, is precisely why the data reveal a disparity between opponents with respect to military and economic power. According to this interpretation, not only are democratic conditions no disadvantage in war, they are a positive advantage. A democracy tends to have an easier time subjecting the pros and cons of a war to public examination and learning from the war as it progresses than other forms of political order.

But this of course applies only under idealized conditions. Democracies are not immune to internal attempts to emasculate their characteristic forms of critical public examination of the reasons for war and war strategies. In this respect, American democracy in the lead-up to the ongoing war in Iraq was a case in point. Quite deliberately, the country's political leaders constantly changed their position on war aims (WMDs, terrorism, democracy, peace in the Middle East). Congress functioned as a site of argument only to a very limited degree, even in comparison to the British parliament. Major media, particularly television – but also the press, including the New York Times, which later practically begged for forgiveness – accepted official reports bordering on falsification. The media narrowed down the spectrum of rationally possible courses of action to the question of the 'how' of war against Saddam Hussein in such a way that there was almost no discussion of the 'whether.' The aim of the Bush administration's media management was to push a narrative largely detached from the facts: "the greatest story ever sold" (Rich, 2007).

It is true that the progress of the war laid bare the fragility of the reasons given for the war and the incompetence of the planning, but it was, above all, the lack of success that fueled shifts in the public mood. Since Max Weber, sociologists have known that all political orders gain legitimacy through military victories. This was the case as long as Hitler was victorious in war, and even in the case of Stalin the end of World War II brought him many more loyal supporters. Democracies have also become more stable as a result of military victory. But what happens if a democracy starts wars by largely abandoning democratic mechanisms and fails to bring them to a successful close? Will this lead to a reintensification of democracy or to the even more adroit bypassing of democratic principles?

Are Democracies Cruel toward Their Opponents?

Do democracies have specific effects on the nature of any armed conflict they may be involved in? From Rousseau to contemporary journalism, the argument has been made that democrats are not only as cruel as their opponents in wars, but in fact even more cruel. Interestingly, in the work of Rousseau this argument was anchored in a moralization of warfare. In his radical individualism, states are artificial bodies and war is a clash between these artificial bodies. As a result, we may only fight the opposing state in a war, but not its citizens. We have the "right to kill" the defenders of the enemy state "while they are bearing arms; but as soon as they lay them down and surrender, they cease to be enemies or instruments of the enemy, and become once more merely men, whose life no one has any right to take" (Rousseau, 1964: 18). But it is doubtful whether Rousseau can sustain this division between state and people. Especially in the case of wars between well-organized republics – in which the will of the public holds away – the distinction between state and people would be annulled. So we could legitimately annihilate the enemy population in order to win the war. Particularly in the case of warring republics, however, this would make it impossible to humanize war as Rousseau wished to do. Every citizen would become a combatant, making it impossible to distinguish between soldiers and civilians. His pedagogical novel *Émile* shows that Rousseau (1979: 481) was aware of this consequence. He put forward the hypothesis that because of citizens' patriotic identification with their state, the wars of a republic would be waged with greater cruelty than those of a monarchy. Tocqueville gave this idea an interesting new twist in his brilliant analysis of America. According to him, any democratic opening up of the armed forces, in other words the diminution of aristocratic privileges in society as a whole, and armed forces in particular, may make them more aggressive. This is because military performance is best demonstrated in war. As a result, there will probably be a greater readiness to engage in military adventures within an army based on the performance principle. In light of this, Tocqueville believed he could discern a gulf between an army particularly willing and able to wage war and a generally peace-loving democratic polity: "We thus arrive at this singular consequence, that, of all armies, those most ardently desirous of war are democratic armies, and of all nations, those most fond of peace are democratic nations" (De Tocqueville, 2007[1835/40]: vol. 2: 266). Tocqueville's thesis is surely too simple: It merely raises the issue of the socially representative nature of the armed forces and their built-in performance standard without really developing it further. But the argument inherent to this question cannot be lightly dismissed. This problem came into particular focus when the USA entered the World War in 1917, a step it justified with Universalist moral and legal arguments. Carl Schmitt referred here to the "turn to a discriminatory concept of war" (Schmitt, 2003[1938]), asserting that this development intensified war. If he is to be believed, it is easiest to limit the scope of wars waged without illusion exclusively in pursuit of various interests. But if those waging them have a sense of moral superiority and believe they have a mission and morally superior cause, all inhibitions fall away. This is because we are no longer dealing with a battle between opponents on an equal footing, but one

between good and evil, between the preservers of order and those who would threaten it. Like many other elements of his often daring arguments, Schmitt's assertion has proved amazingly tenacious and is still quoted regularly to this day, particularly in polemics against the missionary pursuit of foreign policy. But there is no empirical evidence for it. It is of course true that there is always a danger of merely using morality to disguise interests, and that even if we have a sense of moral justification, we must always weigh up carefully the appropriate means and expectable consequences. Empirically, though, the history of violence in the twentieth century certainly does not show that it was mainly the warring Western nations, couching their arguments in moral Universalist terms, that displayed a tendency toward unbridled violence – quite the reverse. In World War II, American troops tended to commit atrocities on the Japanese rather than the German front, and later in the Vietnam War. In addition to the momentum of unbridled violence, which explains certain massacres, unless we consider them part of a strategy (Greiner, 2007), it was racism rather than a Universalist sense of superiority that played the central role here. The same goes for the violent excesses of the colonial powers, as with the suppression of the Mahdist revolt of 1898 in Sudan by the British. This draws our attention to the enforcement of Western universalism through an ideology of civilizational or even racial superiority. In recent years, Mark Mazower and others have demonstrated (Mazower, 2006) how strongly the history of international law has been pervaded by an exclusionary deployment of the concept of civilization, according to which legal norms are held to apply only among the civilized, but not when dealing with 'barbarians.' Because they will supposedly deploy any and all means to achieve their aims, any means of combating them can be justified. This way of thinking may solidify – as in the case of colonialism – into enduring racism. But there is always a risk of such thinking as a result of the spontaneous or actively managed dehumanization of the enemy. We can see this dynamic of dehumanization at work everywhere – in attempts to justify the lawlessness of Guantanamo, in the refusal to allow that the motives of so-called terrorists are in any way comprehensible, and in the dynamics that led to the sadism of Abu Ghraib. But moral universalism entails a countervailing impulse that assails all particularism, whether of the long-standing or of newly generated variety. Conversely, Schmitt's antiuniversalism was from the outset devoid of any such impulse. This is why he idealized the supposedly limited wars of European history and threw in his lot with Nazi policies on Europe. Furthermore, from a normative point of view, the idea of increasing the scope for violence for moral reasons rests on mistaken reasoning. From such a standpoint, the only thing that would be dangerous would be the moralization of international politics without due regard for institutions and procedures. But this is not the aim of advocates of a Universalist morality. Their goal is to establish legal norms, and organs to enforce this law, on an international level. The prosecution of war crimes and crimes against humanity must be pursued in a lawful way. That American foreign policy in the run-up to the Iraq War was not really Universalist in approach was evident in the arrogant devaluation of existing forms of international conflict resolution.

Are Democracies More Inclined toward Peace?

So our interim finding is that so far democracies have shown specific strengths in wars and by no means act more cruelly than other states because of their democratic character. But are democratic states more inclined toward peace? This is the third of our questions. There was an intensive social scientific debate on this question in the late 1980s and 1990s, though not so much in sociology as in that subdiscipline of political science known as international relations. Unfortunately, dialogue between these disciplines is a rare event. For a time it looked as though a real breakthrough had been achieved here. This theory was traced back to a classical thinker of uncontested stature in the shape of Kant, and it appeared so powerfully corroborated by empirical evidence that we seemed to be dealing with one of those rare empirical laws in the social sciences. Now, a decade after its heyday, the self-assurance of the theory's advocates has dwindled markedly and there is no longer any consensus about its empirical corroboration. Once again I would like to present relevant arguments drawn from the history of theory, along with some of a systematic and empirical nature, though they cannot truly convey the richness of this debate.

As mentioned earlier, Kant's 1795 text *Perpetual Peace* was the key reference point for the debate. Here he argues that war itself contains an innate impulse that eventually leads to peace. His reasoning centered on the idea that all present states have a tendency to become republics, as this is the only way to legitimize power on a broad basis – thus putting in place the crucial preconditions for states to prevail militarily in times of 'wars between peoples.' But if republics proliferate as a result of their success in wars, states' international conduct will change accordingly. This is because republics are controlled by the people, and therefore have no interest in aggressive international adventures. Ultimately, thought Kant, it is the people themselves who have to pay the blood price of all wars. Kant did not claim that republics have a fundamentally greater propensity for peace than other regimes. This propensity relates solely to other republican states. In Kant's eyes, the reason for this is cultural. The peaceful and regulated resolution of conflict within a state gives rise to a culture of peace, which carries over to its relations with other countries. A zone of peace gradually develops – a zone of increasing cooperation between states and mutual trust. A worldwide confederation of states now begins to enter the realm of the possible, a confederation in which it is legal rules and no longer violence that regulate relations between states.

While Kant's text had a tremendous impact in the first few years following its publication, it was later largely dismissed. Often, it was remembered as little more than an example of naive reverie. But this was very unfair. Kant was no dreamer, and even the literary form of his text might have guarded against this misinterpretation; he wrote it as a draft agreement, and all the national rulers had to do in principle was sign it. Some scholars of war presented findings of this stripe during the Cold War, but little attention was paid to them, as they did not seem helpful in overcoming that conflict. Who at the time expected the imminent democratization or even implosion of the Soviet Union? But with the end of the Cold War and a wave of democratization in various parts of the world, the vision of

a new democratic and thus enduringly peaceful world order suddenly became the goal of leaders such as the first President Bush, as well as his successor Clinton and left-wing peace scholars. Critics of this thesis were mostly power-political realists who warned against the moralization of foreign policy.

But it would be misguided to view the thesis of a democratic peace as the only alternative to power-political realism. It is far more productive to draw conclusions in light of the inherent difficulties in corroborating this thesis. Most of its advocates agree with the realists, against a minority in their own camp, that if democracies are more inclined toward peace at all, this is true only with respect to other democracies and not in any general sense. This is by no means an insignificant qualification. It opens our eyes to a danger evident immediately after the appearance of Kant's text. As early as 1796, the young Schlegel pointed out that "there was still material left to wage war over" (Schlegel, 1796; in Batscha and Saage, 1979: 106), as long as despotic states existed. And Görres, a Jacobin in his youth, expressed this more pointedly in 1798 to the effect that he was in favor even of the violent "republicanization of as many despotic (regulative) states as the circumstances of time and place allow" (Görres, 1798; in Batscha and Saage, 1979: 168). Such 'messianic interventionism,' with its inherent propensity for violence, was not limited to German thinkers after Kant, but is also to be found in the work of others such as Thomas Paine. Tocqueville himself got to know democratic society – in America – as a highly expansive way of life. This led him to suspect that democracies could be kept stable only through an expansionary dynamic. This is not messianic interventionism, but does represent an expansionism specific to democracy. He used this argument to endorse the French colonization of Algeria in brutal and unqualified terms.

The credibility of the claim that this amounted to an empirically grounded social scientific law diminished upon closer inspection of the way in which it had been operationalized. Which states should be classified as democratic in the nineteenth century – in view of the fact that universal and equal male suffrage was so rare, let alone its female counterpart? It was near impossible even to define war and involvement in war. There was a clear tendency, whenever an example was put forward that proved the opposite, to define the countries involved as nondemocratic or dispute that the conflict in question represented a war in the first place. Theoretical and methodological consequences soon followed once scholars realized that constant reduction of the number of cases considered valid had led them down an argumentative dead end. Theoretically, the thesis was now expressed in relative rather than absolute terms. So now the claim was not that democracies never wage war on one another, but merely that they do so rarely and less often than other regimes. Methodologically, there was a turn away from comprehensive quantitative surveys of all wars and all democracies toward case studies. These have the advantage of helping us uncover the causal mechanisms that (supposedly) make democracies amenable to peace. We can distinguish a structural argument (control of the government by a populace unwilling to go to war) from a cultural one (negotiations in the absence of violence) in Kant's reasoning. But the structural argument had a weakness that was soon to deprive it of its plausibility after Kant. It was based on the idea that the common people are peace-loving and unmoved by aggressive sentiments or ideologies. The age of nationalism and great political ideologies sufficiently demonstrated the reverse. But the cultural argument, at which hopes of a solid theory were now inevitably directed, also suffered from a congenital defect. It would in fact have lent support to the strong thesis of an unconditional capacity for peace, though this is empirically untenable, rather than the qualified thesis of democracies' relative capacity for peace with respect to other democracies. So the debaters came to two conclusions. First, they recognized that countries' conduct is crucially determined by their perception of one another as democratic, and second, that Kant's notion of the republic is not identical to the notion of democracy and liberal culture. There is a great deal of empirical support for this perception-based argument, as it is always the presumption of trustworthiness that makes trust possible. This argument can be challenged by the fact that perception is not simply objectively given, but may be shaped by processes of definition. Power-political rivals are quickly labeled undemocratic by the leaders of democratic states as and when required, while useful allies are just as swiftly declared 'flawless democrats.' The classic example of such a process of redefinition occurred in connection with American intervention in World War I. In a number of studies, Randall Collins (1999), Ido Oren (1995), and I myself (Joas, 2003) have tried to show that prior to 1914 the German Empire was perceived as culturally advanced and by no means undemocratic in either the USA or Great Britain. It was only when relations deteriorated that Germany's authoritarian characteristics were emphasized and even Kant, Goethe, Beethoven, and Wagner declared pathological figures. For the war to be portrayed as a struggle between democracy and autocracy, from the February Revolution of 1917 at the latest Russia had to be stylized as a democracy and Germany as a special case of deviant modernization. In such contexts, the term 'democracy' is simply too sweeping. The task of sociology is to get closer to social reality beyond the mere attachment of labels to a political order while also investigating the role of the concept of democracy in the construction of antagonistic stereotypes.

The other argument, namely the distinction between liberal and nonliberal democracies, further qualifies the thesis of a democratic peace. If it is not the capacity for democratic control that produces a propensity for peace but rather a liberal culture of tolerance and individualism, then Kantian hopes of perpetual peace depend on the spread of this liberal culture rather than the democratization of the world. In this case, at least up to the present, this theory applies only to the West and essentially to the period since the end of World War II. But is it not possible that quite different factors – such as the tremendous affluence and hegemony of the USA, or European fatigue after two world wars – are responsible for the emergence of a zone of relative peace? Might not the long peace in fact be the cause of the triumphant march of liberal culture? The conclusion here is clear in a negative sense. It is not the case that democracies are markedly superior in terms of their capacity for peace. This is apparent even without factoring in new trends toward the diminution of democratic controls over the armed forces – through the outsourcing of military services to so-called private military companies, for example. The positive conclusion – as peace scholar Harald Müller rightly states – is

that we must "cast light on the specific circumstances in which democracy does in fact engender the peace-maintaining effects claimed for it" (Müller, 2002: 46). Empirically, this opens up the debate within international relations to historical sociology. Normatively, it opens up the debate to a concept of peace not founded on a single dimension of causality.

What Are the Effects of Wars on Democracies?

Kant was interested in the causal relationship between war and democracy (or republic) from both sides of the equation, though this has been paid little notice in the debate on the democratic peace. He had to be, because for him the possibility of establishing a stable "civil constitution" ("bürgerliche Verfassung") (Kant, 1977[1784]: 41) depended on achieving "a lawful relationship with other states." This brings us back to the first of our questions, concerned with the effects – beneficial or detrimental – of wars on democracies. As early as the nineteenth century, sociologists began to develop a wealth of theorems that assumed the existence of a spiral mechanism. There were optimistic and pessimistic variants. Herbert Spencer (1885: 557ff) was one of the optimists. He believed that industrial societies are more peaceful than military ones, and that the dissociation between military service and the status of citizen typical of such societies, in other words the professionalization of the armed forces, in turn strengthens individualism. Friedrich Engels' "Herr Eugen Dühring's Revolution in Science" ("Herrn Eugen Dührings Umwälzung der Wissenschaft (Anti-Dühring)") (1988[1878]: 204) presents a rather similar argument. In his view, the introduction of universal military service and advanced weaponry was driving the states of Europe to financial ruin while arming the workers and peasants: Here again, we have the assumption of a self-reinforcing mechanism leading to a more peaceful world. Interestingly enough, the pessimistic variant of this theory can be found in many forms, particularly in the history of American social science. Notorious social Darwinist, William Graham Sumner warned in 1898 that American imperialism was encouraging militarism and plutocracy and threatening American political culture. In the early days of World War II, Harold Lasswell (1940) set out the horror scenario of an impending "garrison state" in which the people are more or less entirely at the mercy of "specialists on violence." Most spectacularly, in the 1950s, C. Wright Mills advocated the idea that the USA in the twentieth century had broken with its anti-militaristic tradition and that an alliance had formed between economic monopolism and military bureaucracy, an alliance that had come to dominate the very interpretation of reality. He feared that this mind-set and these power relations made it impossible to prevent nuclear war.

With hindsight, however, it is clear that neither the optimistic nor pessimistic prognoses on the basis of these theories have been proved correct. Both variants of the notion of a self-reinforcing mechanism have proved overly mechanistic. It may seem astonishing, but empirically the optimistic predictions proved correct far more often than the pessimistic ones in the nineteenth and twentieth centuries. What Parsons called 'inclusion,' in other words the extension of civil rights to previously excluded sections of the population (blacks, women, young people), often occurred as a consequence of wars, as did the broadening of social rights. Conversely, as Paul Starr (2007) argues, the liberties so often curtailed in wartime were frequently regained once the war was over. Yet, these facts do not really lend support to the optimistic variant for two reasons. First, the causes of successful processes of democratization are quite obviously significantly more complex than it would appear in the simple image of a spiral of reinforcement. Even Spencer and Engels revised their thinking on this point, and we have every reason to take seriously the contingency of the success of democracy. The pessimistic variants at least teach us that wars are not mere interruptions of civilizational progress, but may be constitutive of entire eras, of new social and political structures. This applies in the case of successful steps toward democratization, but especially in cases in which defeat and destabilization enable dictators or totalitarian movements to seize power.

Second, there is no guarantee that the so far often beneficial effects of wars on democracies will persist. The optimistic interpretations are based on two tacit assumptions: that wars last for a limited time, after which, for example, liberties can be enhanced once again, and that they are dependent on the loyalty of large sections of the population, their readiness to support the war effort through reduced consumption, working harder, and perhaps even risking their own lives. But these conditions do not necessarily apply. Systematic attempts may be made to ensure that they do not. During the Cold War, Mills already feared that restrictions would be justified with no prospect of their removal in the foreseeable future. At present, the so-called War on Terror seems to be a textbook case of a war that can never be declared over. The title of the book, *The Forever War* (Filkins, 2007), alludes to this. The Iraq War was something quite new in terms of financing and mobilization. Despite its huge cost, it was accompanied by massive tax cuts, and military strategy there and in Afghanistan was geared toward the minimization of American losses through technological superiority and a military personnel whose composition in no way reflects American society as a whole. So the formerly inclusionary effects of war have now become improbable.

Acknowledgment

This article is based in part on an article published in *European Journal of Social Theory* 15 (2012), 21–34.

See also: Civil Wars; Democratic Theory; Military Sociology; Military, War, and Politics; Peace Processes; Peace; War, Social Causes and Consequences; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Batscha, Z., Saage, R. (Eds.), 1979. Friedensutopien. Kant, Fichte, Schlegel, Görres. Suhrkamp, Frankfurt/M.

Collins, R., 1999. German-bashing and the theory of democratic modernization. In: Collins, R. (Ed.), Macrohistory: Essays in Sociology of the Long Run. Stanford University Press, Stanford, CA, pp. 152–176.

De Tocqueville, A., 2007[1835/40]. Democracy in America. Alfred Knopf, New York.

Filkins, D., 2007. The Forever War. Vintage, New York.
Engels, F., 1988. Herrn Eugen Dührings Umwälzung der Wissenschaft. MEGA I/27. Dietz, Berlin.
Greiner, B., 2007. Krieg ohne Fronten. Die USA in Vietnam (Edition). Hamburger, Hamburg.
Görres, J., 1978. Der allgemeine Friede, ein Ideal. In: Batscha, Zwi/Saage, Richard, 1979. pp. 111–172.
Huntington, S., 1957. The Soldier and the State. Harvard University Press, Cambridge, MA.
Joas, H., 2003. War and Modernity. Polity Press, Cambridge.
Joas, H., Knöbl, W., 2013. War in Social Thought. Hobbes to the Present. Princeton University Press, Princeton, NJ.
Kant, I., 1977[1784]. Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht. In: Weischedel, W. (Ed.), Schriften zur Anthropologie, Geschichtsphilosophie, Politik und Pädagogik, Werkausgabe, Band XI. Suhrkamp, Frankfurt/M, pp. 31–50.
Lasswell, H.D., 1940. The Garrison state. American Journal of Sociology 46 (3), 455–468.
Mann, M., 2005. The Dark Side of Democracy. Explaining Ethnic Cleaning. Cambridge University Press, Cambridge.
Mazower, M., 2006. An international civilization? Empire, internationalism, and the crisis of the midtwentieth century. International Affairs 82 (3), 553–566.
Müller, H., 2002. Antinomien des demokratischen Friedens. Politische Vierteljahresschrift 43 (1), 46–81.
Oren, I., 1995. The subjectivity of the 'democratic' peace. Changing U.S. perceptions of imperial Germany. International Security 20 (2), 147–184.
Parsons, T., Turner, B.S., 1951. The Social System. The Free Press, New York.
Reiter, D., Stam, A.C., 2002. Democracies at War. Princeton University Press, Princeton, NJ.
Rich, F., 2007. The Greatest Story Ever Sold. Penguin, New York.
Rousseau, J.-J., 1964[1762]. Du Contrat Social (Première Version). In: Rousseau, J.-J. (Ed.), Œuvres Complètes, Tome 3. Gallimard, Paris, pp. 279–346. English Translation: The Social Contract (1913). Dent, London.
Rousseau, J.-J., 1979. Émile or on Education. Basic Books, New York.
Schlegel, F., 1796. Versuch über den Begriff des Republikanismus veranlaßt durch die Kantische Schrift "Zum Ewigen Frieden". In: Batscha, Zwi/Saage, Richard, 1979. pp. 93–110.
Spencer, H., 1885. Political Institutions. Principles of Sociology, part 5. Williams & Norgate, London.
Schmitt, C., 2003[1938]. Die Wendung zum diskriminierenden Kriegsbegriff. Duncker & Humblot, Berlin.
Senghaas, D., 2007. On Perceptual Peace. A Timely Assessment. Berghahn, New York and Oxford.
Starr, P., 03 and 12 March 2007. War and liberalism. The New Republic 21–24.
Starr, P., 2010. Dodging a bullet. Democracy's gains in modern war. In: Kier, E., Krebs, R.R. (Eds.), In War's Wake. International Conflict and the Fate of Liberal Democracy. Cambridge University Press, New York, pp. 50–66.

War and Nationalism

Daniele Conversi, IKERBASQUE, Basque Foundation for Science, Bilbao, Spain; and Department of Contemporary History University of the Basque Country UPV/EHU, Leioa, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the relationship between war and nationalism as it developed since the French revolution and the French revolutionary wars from the perspective of international political history. It shows how nationalism as a modern ideology changed the essence of warfare, while the relationship grew in intensity in the long gestation from the Napoleonic campaigns to World War I, reaching its peak in the heydays of totalitarianism between World War I and World War II. Totalitarianism is largely seen as an attempt to preserve the state's radical control over its citizens achieved under war conditions.

The postwar order created the conditions and the institutions for severing the symbiotic relationship between war and nationalism by focusing on interdependence, human rights, etc. However, the excesses of neoliberal globalization have curtailed the representativeness of political institutions, while fomenting instability, which may lead to a new assertion of nationalist conflicts and war. Finally, our traditional notions of war, nations, and nationalism are likely to change when dealing with the catastrophic impact of environmental degradation.

Introduction

While organized warfare has been taking place at least since the Bronze Age (Eckhardt, 1992: 170), nationalism is an entirely modern phenomenon, which can be ascribed back to the French Revolution (1789–99) and, occasionally, to the English Civil War (1642–51), and the American declaration of independence (1776). Because of their impact on both political legitimacy and warfare, the French Revolutionary Wars (1792–1802) are the most clearly identifiable chain of events inaugurating the 'special relationship' between war and nationalism as it shaped the development of the Western system of nation-states. In particular, Paris' introduction of the *levée en masse* (a term combining the meanings of 'uprising' and 'levy') on 16 August 1793 established mass conscription in the defense of the nation as a military model then widely emulated. Despite its uneven spreading, the industrial revolution also contributed to both warfare and nationalism, although the latter often proliferated among elites well before they embraced industrialism.

This article identifies the peak of this relationship in the period between 1789 and 1945, to which most space is therefore dedicated. The last sections briefly discuss both postwar developments and the coming challenges to this relationship.

From the French Revolution to the French Revolutionary Wars and the Napoleonic Campaigns (1789–1915)

As the French revolution degenerated onto a factional bloodbath, appeals to unity became nearly inescapable among ruling elites and quarrelsome Jacobins. After France entered into war (1792), a deeper cycle of conflicts began, so that international violence, rather than revolutionary violence, became the main unitary catalyst. According to the historian David A. Bell, the Battle of Valmy (20 September 1792) saw the first army in human history inspired by nationalism, as throngs of soldiers immolated themselves to shouts of 'Vive la Nation!'(Bell, 2008). Although victory was made possible by casual events such as bad weather, Jacobin propaganda promptly seized Valmy as a foundational myth, unleashing waves of enthusiasm and the belief that fighting in the name of 'freedom' would grant soldiers a sort of immortality and even invincibility. Bell identifies this as the 'first total war,' a concept usually associated with the trenches and the 'human waves attacks' of World War I, but already detectable with the notion of *guerre à outrance* emerged during the Franco-Prussian War (1870–71). At Valmy, for the first time, the sheer number of men ready to sacrifice themselves on the front line became decisive. The unprecedented 'enthusiasm' for mass death was only made possible by the Parisian elites' coherent effort to channel popular emotions by appealing to nationhood and patriotism as an organizing ideology. Despite, and possibly because of, fratricidal struggles among revolutionary elites, populism and patriotism were shared across the political–ideological spectrum. Thus, the first total war was also the first ideological war and the first nationalist war. It provided a further foundational myth to the first modern nation-state with the triumph of a new ideology linked to (positivistic) ideas of modernity and progress.

From July 1791, even before the war began, to July 1794, the French army became the target of a strenuous propaganda effort, with 7 million copies of various revolutionary journals distributed among high- and low-rank soldiers, although most of them could hardly read or write (Lynn, 1996). Mobilized around the sacred defense of *La Patrie* (the Mother/Fatherland), soldiers were hailed as the supreme expression of 'collective will,' while war was described as the finest of national virtues. Before the *levée*, volunteers were drafted in through an array of visual effects and media grandeur, often surrounded by a festival atmosphere punctuated by martial music. For urban elites, mass conscription became *de facto* a 'nation-building' device insofar as nationalism could emerge as the broader interclass ideology suitable to mobilize and control a largely rural population. "The first mass army depended ultimately upon a political revolution whose ideology, redolent of

nationalism, stressed the equality and community of all Frenchmen" (Posen, 1993: 83).

The emphasis on patriotic unity concealed and embellished deep ideological cleavages. This is how nationalism sprouted like a *deus ex machina*, providing the decisive strategic advantage and the common denominator of all ideological forces competing to act in the name of the Republic. Through war mobilization, Parisian riotous elites achieved unified support for what had become one of the most fragmented, ideologically splintered, and identity-fractured countries in Europe. For instance, insofar as he did not control the National Convention and the Committee of Public Safety, Maximilien Robespierre (1758–94) was opposed to war, which he intuitively felt as a potential distractive device through which 'antirevolutionary' forces could seize power (Robespierre, 2007: 30; Scurr, 2006). However, once in full control, he did not hesitate to pursue war further to consolidate his own supremacy. By merging culture, ideology, terror, and war at once, nationalism offered an impromptu combination of consensus and coercion, 'spontaneous' peer pressure and hetero-directed state repression. Although the Reign of Terror's bloodbath reached its peak after November 1793, when the threat of foreign invasion had receded, war abroad coincided with a dramatic increase in repression at home: systematic mass killing by government troops led some historians to identify the *Vendée* massacres (1793–96) as the first modern genocide (Jones, 2010: 6–7; Levene, 2008; Secher, 2003). The 'eliminationist' pattern was replicated, expanded, and 'refined' in other provinces and cities, like Brittany and Lyon. Together with the remoteness of the 'punished' regions, war provided the most suitable cover to carry out ominous crimes and abuses, which were hardly conceivable in pace times and in more dense urban setting.

The Napoleonic campaigns (1803–15) are sometimes seen as a continuation *of* the French Revolutionary Wars, but had an even more direct impact on the spread of nationalism while reinforcing its links with warfare: Napoleon Bonaparte's invasions not only created throughout Europe 'proto-national' institutions, which served as embryos for nascent nation-states, they also spawned unprecedented nationalist reactions against French occupation troops, like in Germany, Russia, Italy, Portugal, and Spain (Esdaile, 2004). The invasions inaugurated and fomented a downward spiral of intensifying aggressiveness, and a Darwinian race for the 'survival of the fittest,' so that a nation's continued existence seemed to be inseparable from expanding state power and strengthening its industrial and military apparatus. In turn, hostilities invoked increasing responses from the enemy, typically embossed by modernist and technocratic justifications.

Fatherland, Barrack, and School

The relationship between nationalism and war is underlined by the role of education as a facilitator and preparatory training for military life and war. Modern states aspired to the long-term molding of national consciousness through educational indoctrination and military mobilization. Of the two, the aim of mass education was more all-pervasive and long-term than mass conscription, ordinary citizens could welcome it more positively, despite heavy fines and even sequestration for nonattendance. The intimate connection between military service, patriotic coaching, and educational standardization is epitomized by Napoleon's 'nation in arms,' with its fusion of barrack and school: Napoleon was both hailed as the greatest military strategist and the 'education Emperor.' He implemented the standardization of the curriculum, instituted the *lycées* system, introduced school uniforms and examination procedures, reinforced the bureaucratic structure, and drafted newly stringent disciplinary standards. For Napoleon, a national elite could be formed by evenly applying military-like discipline and the principle of 'equal' opportunities to education, with copious largesse of newly established scholarships. The main goal was to forge a class of officials capable of administering his empire and leading his armies. The system sought to imbue young pupils with patriotic–militarist virtues and shape them into loyal servants of the central state. Napoleon's own military formation exerted a decisive impact on his confusion between the civic, educational, and military spheres (Colin, 1900).

The Westernization of the World: Nationalism and Conscription

Napoleon's project of a 'nation-in-arms' with its 'citizen-soldiers' was emulated across the ecumene with various degrees of intensity. Mass conscription was thus instituted in a host of countries of both the developed and developing world, from Prussia and Russia to Turkey and Japan, where it often merged with extreme forms of nationalism. In Japan, the new Meiji rulers (1870–81) transformed the *samurai* class system into the Imperial army through nationalism, education, and conscription (Harries and Harries, 1991). Conscription was one of the first measures taken in 1860 by the newly formed Kingdom of Italy: since the start it acted in tandem with the elementary school system to shape and structure a unitary Italian identity. Most of these governments saw conscription and militarization "as an instrument for developing social cohesion and political docility of the masses" (Bond, 1986: 32).

Pax Britannica, Industrial Development, and Empire Building (1815–1914)

The period spanning broadly from Waterloo to World War I, and particularly from 1870 to 1914, is often described as an era of peace and stability. However, just as continental empires collapsed, colonial empires expanded, while some state elites consolidated their grip on power over their rivals. Nation-states strengthened their control over citizens through the twin ideologies of modernism and patriotism, while intensifying the colonial scramble for the unspoiled. War and destruction were simply being exported beyond European borders and, by the late nineteenth century, an unrestrained *laissez-faire* economy had induced a series of 'late Victorian Holocausts' (Davis, 2001), including artificial droughts and floods in which millions perished. Linda Colley noted, "the profit and the price of this hundred-year partial European peace was unprecedented Western, and especially British, freedom to concentrate on global empire. In 1800, the European powers, together with Russia and the United States, laid claim to some 35 percent of the globe's total land area. By 1914 … [their] proportion of the globe … had risen to 84 percent" (Colley, 2002: 311). In

ideological terms, racism accompanied empire building just like nationalism propelled state making, as the two sides of the same coin.

Franco-Prussian Militarism

Strategically located between Prussia and imperial Britain, France was able to combine extreme imperialism and radical nationalism. In a fierce emulative competition, France and Prussia built powerful armies, bureaucracies, and industries. The Franco-Prussian War (1870–71) exerted unprecedented pressures to discipline and assimilate their respective populations, making ordinary people more conform to state directives and control.

Engaging in grandiose project of nation-building and social engineering, official nationalism acquired the capacity to produce that very 'self-disciplined subjectivity' necessary for the state to exert control both within and outside official institutions. This made possible broader forms of organizational discipline, and finally the participation of the masses into industrialized warfare. In a kind of domino effect, the fierce rivalry and mutual emulation between France and Prussia heralded a new era of intense, European-wide.

The Franco-Prussian military model spread across the globe but spared parts of the West, like the Iberian Peninsula, the British Isles, and upper Scandinavia. King Gustavus Adolphus of Sweden (1594–1632) introduced much earlier uniforms and regiments-based conscription districts in Sweden and Finland in which enlisted peasants replaced mercenary soldiers, so that military historians attribute to this important change, the Swedish victory in the Thirty Years' War (1618–48). Britain's imperial makeup turned the navy, rather than the army, into the key military force (Kiernan, 1998): Still in 1850 only about 105 000 armed personnel could be defined as military: that was "less than a third of the size of France's military at that time, less than an eighth of Russia's, and smaller even than the army of Prussia which possessed no colonies at all" (Colley, 2002: 312). Significantly, England/Wales was also one of the last places in Europe to adopt compulsory education (1880). Thus in Britain, "the army never became the school of the nation," but simply functioned as a rite of passage to adulthood, 'making men' more than making British men (Showalter, 2010: 142). This contrasted with Prussia, where the "military service was the precondition not only to full manhood, but also to full citizenship" (Berger, 2008: 620). However, just like in Prussia and France, war became a nation-builder since the British wars with Catholic France drove the distinct Protestant denominations into a closer union, reminding them of what they shared (Colley, 1992). In other respects, and until quite late, the British Empire did not follow the continental drift toward army aggrandizement and mass conscription.

Pupils into Soldiers: Expanding the Franco-Prussian Model

In 1880, France's minister of education, Jules Ferry (1832–93), introduced a set of sweeping reforms to impose compulsory, 'nonclerical' (*laïque*) education, which aspired to nationalize the masses through nationalist indoctrination by raising the public profile of the army. School reforms included courses in military exercises, gymnastics, needlework (for girls), and the actual replacement of religion with the cult of *la Patrie*. Eugene Weber describes the pedagogical catechism imposed throughout the *Hexagone* calling for the child's "duty to defend the fatherland, to shed his blood or die for the commonweal, … to obey the government, to perform military service, to work, learn, pay taxes and so on. At the very start of school, children were taught that their first duty was to defend their countries as soldiers …. Commencement speeches recalled this sacred duty in ritual terms – our boys will defend the soil of the fatherland. The whole school programme turned on expanding the theme" (Weber, 1976: 333). All disciplines were harnessed to this goal: history, literature, geography, and civic education. However, in terms of acculturation, by 1896, the army had proved to be "an agency as potent in its way as the schools" (Weber, 1976: 302).

In reality, after the *defeat* in the Franco-Prussian war, such developments were no longer uniquely endogenous: the nationalist stress on the education–military linkage mimicked developments in rival Prussia. For Peter Paret, "French adopted from Prussia not only true conscription … and the modern general-staff structure, but also the collaboration of the elementary schools and the conscript army to teach nationalism to the masses" (Paret, 1993: 49). The Franco-Prussian mutual contaminations expanded tangentially beside the army into other organizational and bureaucratic areas. Prussian developments were in turn a response to previous French threats and invasions: In 1813, Frederick William III "embarked upon a military mobilization that, for the first time, extended the obligation, and opportunity, to serve in war through all levels of Prussian society. Within the framework of the Revolutionary and Napoleonic eras, the Prussian *Erhebung* … was the movement that came closest to capturing the dynamism … of the French effort twenty years before" (Moran, 2003: 3–4). In other words, conservative Prussia followed the footsteps of revolutionary France, which in turn began a cycle of emulative competition, so that the roles were soon reversed; whereas, France was the initial model for export and even though Prussia had turned into a military state already before the French Revolution, Prussia became the apprentice who outdid the master, turning itself into the military and bureaucratic model to be imitated across the Europe and Asia, including in France.

Similarly, schools elsewhere acted as indoctrinating institutions. Thus, in Japan, primary and secondary schools operated to coach students for the imperial army, while teachers were trained in military barracks (Harries and Harries, 1991). In Turkey, the 'military-nation' based on education became 'foundational myth of Turkish nationalism' (Altõnay, 2004). One decade after France's defeat by Prussia (1871), "one could neatly perceive the notion of a Spartan-style education, entirely devoted to patriotic adulation and where the school became the antechamber of the barrack" (Girardet, 1953: 164). Propaganda through schooling and mass media became essential in the dissemination of militarism and was the key in the process:

> In the 1870s … nearly every French family became acquainted with the nature of army life. The darker aspects of barrack life were pushed into the background; what mattered above all was to prepare for the

> imminent war of revenge against Germany. Officers marked in black the frontiers of the lost provinces on maps of France, and soldiers ending their service often presented a bust of 'Alsace in tears' to their company commander.
>
> Bond, 1986: 35

While conscription was highly unpopular in some areas, with various anticonscript revolts both before and after Napoleon's times (Aaslestad, 2012; Forrest, 2012; Planert, 2012), soldiers' memoirs describe military service in highly commendable, praiseworthy terms (Bond, 1986: 35). The army's popularity was largely founded on the persisting myth of the 'nation-in-arms' as appeared during the late revolutionary period and then regulated by Napoleon (Forrest, 2009). "Prior to 1920, there was no serious discussion of conscientious objection in France. *Insoumis*, rebels and insubordinates, and *réfractaires*, draft dodgers, were a serious problem for the government at times" (Levi, 1997: 186). It is also possible that nationalism exerted a mobilizing function before soldiers joined their battalions and regiments, serving more to tie up civil society to the military, while justifying both military spending and human sacrifice.

The army provided a deeper and faster emotional impact on conscripts than compulsory education could exert on pupils – particularly in wartimes. School took much longer to shape loyal citizens than the drastic, strongly centralized, and hierarchical barrack. Isolation from family and friend, punishment, drill, and round-the-clock propaganda yielded more drastic results than the deeper river's flow of half-day classes.

Class was also an important factor. Through decades of propaganda, schooling, centralization, and bureaucratization, French peasants had incorporated a humiliating sense of class inferiority and often desired to remove the stigma associated with rural mores and illiteracy (Weber, 1976). For some of them, economic incentives for joining the army were significant at a time when rural life was disintegrating, state centralization pressed on, and massive taxation exacted a heavy toll on ordinary citizens. Despite widespread resentment against various forms of state intrusion and repression, the idea that the *Patrie* was being victimized and its soil violated awakened powerful patriotic sentiments. Ordinary citizens were taught to see the soldier as the supreme expression of the collective will, condensing the finest of national virtues: "War itself became an homogenizing experience as soldiers and sailors represented the entire nation and the civilian population endured common privations and responsibilities" (Tilly, 1990: 116).

War, Industry, and Nation: The Egalitarian Impetus

Industrialization and militarism, particularly when supplemented by nationalism, shared the mirage of equal relationships, underpinning new social relations characterized by invocations to 'egalitarianism,' which in turn concealed simultaneous interchangeability and hierarchization. Political cohesion was pursued first within the barracks, then within society at large, while the latter was being radically reshaped by industrialization. The mass army was possibly the most 'egalitarian,' yet the most hierarchical of the major modern institutions; while the rigidity of the chain of command remains inexpugnable, an army can only work on the basis of equal duties, rights, and behavior of its low-level ranks (only within the prison the new *combination* of hierarchy and egalitarianism yielded possibly more appalling human costs). Due to practical reasons, emerging modern armies experienced first and most forcefully the demand for cultural standardization and recruits' sameness. For instance, the U.S. army anticipated desegregation before the heydays of the civil rights movements, Fascism enrolled Italian peasants on the basis of the egalitarian promises, both symbolic (military comradeship) and tangible (pensions, special rights). This 'egalitarian' emphasis in times of war was shared by authoritarian, totalitarian, and democratic systems. In this way, dissent could be controlled through unprecedented, and otherwise unattainable, forms of conformism. But conformism also led to extermination, as replaceability does not value individual life and human uniqueness. In the process, European states led the way by "building up fearsome coercive means of their own as they deprived civilian populations of access to those means," relying mostly on capital and capitalists to reorganize coercion (Tilly, 1990: 68–69). Their impact was so far-reaching that a whole global order emerged in its image.

In the social sciences, "there is virtually no disagreement that the eruption of war almost instinctively increases in-group solidarity and national homogeneity," although it is likely that "macro-level solidarity and group homogeneity exhibited in times of violent conflicts originate outside of these conflicts … (and) homogenisation is a complex process that requires a great deal of long-term institutional work" (Malešević, 2010: 179–180). In fact, before war could complete its task of bureaucratizing and militarizing society, preparations for war had already hard-pressed citizens toward greater forms of homogeneity. But war itself contributed to wiping out many local cultures in all of the belligerent countries, while contributing to the enemy's cultural obliteration. For Eugene Weber, it was the experience of World War I, which proved decisive in diluting local attachments (Weber, 1976).

World War I

State-building, the cultivation of patriotism, colonial aggrandizement, economic expansion, heavy industrialization, and the unprecedented advance of technological progress led to uncontrollable military buildup, which finally exploded with World War I. Industries, economic welfare, and military arsenals expanded until everything short-circuited under the 'European apocalypses' of 1914–18.

Rapid industrial development meant that for the first time peasants could move in larger number to the cities leaving behind millennial traditions. A powerful and influential class of new riches emerged, which often embraced war, nationalism, and modernity with similar enthusiasm and greed. In many consolidated nation-states, like Italy, Germany, and France, modernists and ultranationalists became the most prominent advocates of war, including poets like d'Annunzio and Apollinaire, and art entrepreneurs like Marinetti (Conversi, 2009).

Before the conflagration, Leo Tolstoy (1828–1910) understood well the destructive linkage between nationalism and war:

> To free people from those terrible calamities of armaments and wars, which they suffer now, and which keep growing greater and greater, we do not need congresses, conferences, treaties, or tribunals, but the abolition of that implement of violence that is called government, and from which originates all the greatest calamities of men. To abolish the governments only one thing is needed. It is necessary that men should understand that the sentiment of patriotism, which alone maintains this implement of violence, is a coarse, harmful, disgraceful, bad, and above all, immoral sentiment.
>
> Tolstoy 1901, 1905: 157

This bitter commentary would remain unheeded and submerged by the more vociferous noise of patriotic war mongering. Interstate competition was measured not only in terms of technological progress, but also in the application of new technologies to the practice of war. For instance, the cavalry had remained the key unit until 1916, when war compelled armies to replace horses and cavalrymen with tanks and technically trained drivers, codrivers, gunners, radio operators, and specialist mechanics (Diamond, 2003), therefore amplifying and deepening bureaucratization. In terms of citizens and education, most states continued to further their goal of building homogenous communities in which school and army played a conjunct role in fomenting extreme patriotism and a quasi-religious cult for the Fatherland. The deification and divinization of the state, which first manifested itself under Robespierre, became the new 'normal' throughout Europe, as it was perfected and magnified by Fascism and Nazism.

The Peak of Nationalist Homogenization: From World War I to World War II (1914–45)

In Charles Tilly's words, the modern nation-state's yearning for homogeneity and control spawned 'the most bellicose century in human history' with about 275 wars, 115 million deaths in battle, and over 115 civilian deaths during the whole century (Tilly, 1990: 67). However, these data are still 'optimistic' because they exclude hundreds of millions and more killed by the state, through policies of genocide, politicide, 'classicide,' starvation, population transfer, economic manipulation, and war-related diseases (Mann, 2005). Moreover, "until World War II, more victims of war died of war-borne microbes than of battle wounds" (Diamond, 2003: 197). Tilly (1985) famously describes the modern state as a racket organization built on the taxation of uprooted, terrorized citizens, who were thus offered military 'protection.' More specifically, "the military played a prime role, influencing both the state apparatus and other organizations including, at a later date, business firms. For it was to a large extent in the military sphere that administrative power in the modern guise was pioneered" (1985: 113).

Nationalism, War, and Genocide

The linkage between war and genocide is so deeply established among scholars of all disciplines as to appear indissociable (Markusen and Kopf, 1995). In fact, the two have been defined as the 'Siamese twins of history' (Jones, 2010: 48). There is further consensus that interstate wars provide the ideal circumstance for carrying out atrocities which would be unthinkable in peace times (Bartrop, 2002; Fettweis, 2003; Levene, 2005; Melson, 1992; Shaw, 2003). Under the aegis of total war, pressures toward ethnic and cultural homogenization reached their peak. When the Anglo-French allied forces landed in Gallipoli in 1915, Turkey's military authorities began a 'securitization' campaign against the entire Armenian population, whom they perceived as the West's 'fifth column.' "What turned a war crime into a genocidal act was the context of total war [….] that translated deportation swiftly into the mass slaughter, abuse, and starvation of an entire ethnic group potentially troublesome to an authoritarian regime at war" (Winter, 2003: 208). Under siege by the 'West,' Turkey's elites ended up imitating the West. As a consequence, their military nationalism bred a 'culture of hatred' that demonized Christians and non-Turks. Bartov's (1996, 2001) work on Hitler's war in the east anticipated much of this war-centered approach as the Shoah was also carried out beneath the curtains and under the strains of war.

Fascism, War, and Nationalism

The relationship between war and nationalism, militarism, and patriotism reached its apogee under Fascism. After the calamities of war, the short-lived interwar hiatus was soon thwarted by the rise of Fascism with its futurist political religion, totalitarian symbolism, and a program of radically remaking society (Gentile, 2003). Totalitarianism can be described as a continuation of the state of war (Keegan, 1999; Mondini, 2006). During the war, European, American, and Asian political, military, and cultural elites had experimented an extraordinary control over their societies through mass mobilization. They were generally keen on preserving this power. Most importantly, all-pervasive war censorship was not limited to the war front, but had engulfed the whole of society, and therefore it was never completely lifted after the war. Moreover, the choir of propaganda, chants of victory, and patriotic rhetoric were continued. Despite the enormous sufferings, the cataclysmic destruction, and the obvious strategic errors that cost the lives of thousands, none of these were debated (rater than celebrated) at length, except within marginal sectors of society whose voice could not be heard as widely as necessary. Moreover, the anarchist movement had been annihilated by the war, when most anarchists turned to nationalism, socialism or died. Religious and socialist organizations, which had tried to articulate programs founded on peace-building, had scarce access to the mass media. In short, the war had emboldened elites and these generally had no intention of surrendering the unprecedented control over society they had achieved. In Italy, where an antiwar socialist and Catholic opposition was relatively vocal, both mainstream and fringe media continued their bellicose propaganda. A broad spectrum of public opinion became convinced that Italy's 'sacrifice' had been in vain, since the 'perverse' postwar international order had not accomplished the realization of the nationalists'

maximalist-irredentist goals, in particular the seizure of Istria and Dalmatia. The creation of Yugoslavia and the treaty of Versailles (1918) had been a blow to Italian nationalists, eager to control both shores of the Adriatic. War veterans imbued by expansionist nationalism played a key role in forming the political atmosphere, which propelled Benito Mussolini into prominence. Fascism itself emerged from outside the military sphere; but, once seized power, it promoted increasingly stringent alliances with military cadres. In turn, post-World War I military elites embraced Fascist ideology to fight against antipatriotic 'decadence' and the 'ingratitude' of ordinary Italians (Mondini, 2006). Contrary to Germany, compulsory draft (*leva obbligatoria* or *naja*) had been kept unchanged in Italy since the early days of the unitary state. The Nazis reestablished conscription in March 1935, emboldening the Third Reich's triadic relationship between militarization, nationalism, and mass industrialization.

Postwar Era and Globalization (1945–Present)

The new postwar international institutions and military alliances (like NATO), as well as regional federations (the EU) became an effective way to overcome international conflicts and divisions, so that war and nationalism could be increasingly disjointed and divorced. As we know, the Cold War (1948–89) had frozen most ethnonational conflicts, while threatening a nuclear Armageddon. After the Cold War, the eruption of localized ethnic clashes and regional wars showed that some of these had remained latent. However, other conflicts may have resulted from the rapid meltdown of the old system and the unmediated passage from late Soviet authoritarianism (or Yugoslav benevolent one-party rule) to a regime of wild economic liberalization and the adoption of international rules emanating from outside Europe (IMF, World Bank).

Since the 1980s, neoliberal globalization had been weakening state institutions, a decay and lack of control has been made visible by the interlocking 2008–12 financial and banking crises. Savage neoliberalism has corroded democratic and representative institutions throughout the world, establishing an unprecedented influence of giant multinational corporations and organized crime over the political affairs of nearly every single country. This power reconfiguration also reinforced specific potent institutions within existing states, amplifying bureaucracy, compressing cultural functions, and expanding repressive areas like policing and surveillance. The triumph of 'free-market' ideology has been so complete and devastating that, for the first time in human history, the use of financial capital(s) lies beyond the control of any human actor or institution, let alone the state. In other words, the bulk of free-flowing, boundless financial capital no longer performs an economic function. However, state power did not surrender its monopoly of tax levy and coercion within its territory – and has indeed increased the incarceration of citizens and noncitizens. Nationalism is no less significant now than before, but its relationship with both the state and the military has deeply changed.

On a more positive note, the rise of the doctrine of human rights shifted the focus on individuals and nondominant groups allowing many submerged voices to express themselves by lifting the veil of censorship. In this way, conflicts could be prevented and successfully terminated through consociational or other arrangements, which did not require secession or threaten state breakup, therefore pointing to the possibility of the nation-state.

Future Challenges

It is perhaps too early to assess the impact of unregulated globalization and its likely relationship with new wars, some alimented by the persistence of state nationalism, others by the rise of nonstate nationalism. An area which will not fail to produce groundbreaking research is the relationship between anthropogenic climate change and conflict, including war and nationalism. Against the backdrop of mass human displacement, global emergency may call for a universal supervising body, but crucial decisions may still take place within the hidebound confines of state arenas. Sooner, rather than later, these may include addressing highly taboo questions, like whether or not top carbon dioxide emitters might need to offer shelter to thousands, possibly millions, of environmental refugees. Although these and related issues have been kept at arm's length from the political agenda, they are likely to generate heated public controversy once, and if, released to broader sectors of the public opinion. The rise of higher barriers to stem free human movement lead to an exponential growth of mass policing and surveillance, leading in turn to unprecedented measures of mass expulsions, which may in turn feed extreme violence both by and against the state (Zimmerer, 2014). Worst-case scenarios adumbrate the collapse of entire systems, with those layers weakened by neoliberal globalization, like health and education, falling first. Once climate change begins to hit more dramatically core countries like China and the United States, a descent into international chaos may well push the world into a state of perpetual war alimented and toughened by nationalism. But many predictions are destined to fail, as eco-environmental issues are always complicated by the intrusion of capricious, volatile, arbitrary variables, like uncontainable epidemics and other climate-related health issues. A new relationship between war and nationalism is likely to emerge through a new geopolitics of 'securitization' as sketchily described in this final section.

Conclusions

Dying for the 'fatherland' may not be a contemporary fixation: Horace's epigram *Dulce et decorum est pro patria mori* ('It is sweet and good to die for one's country') expressed a similar mandate during the time of Augustus, the founder of the Roman Empire. What makes the difference with modern times is a superimposed multilayered ideological apparatus, whose upper layer is constituted by nationalism – with various intermediate layers like egalitarianism and

the faith in progress subsumed in it. Although bureaucratization and centralization predated the French revolution, politics and warfare changed forever once infused with nationalism. A wider project of mass self-abnegation and sacrifice could only be conceived by political elites in a post-1789 scenario.

Key to the nation-state's war effort has been its capacity to secure the willingness of its able-bodied citizens to die for a 'higher' ideal. Such power of life and death over most of its population, previously unthinkable, could only occur when nationalism allowed ordinary citizens to visualize the nation as a unified, organic body, the defense of which demanded individual sacrifice.

The second essential element was 'replicability' or replaceability: despite the cult of patriotic heroes, no individual was unique enough not to be fully replaceable by another individual like him. All this was made possible by nationalism's ability to conceal new hierarchies of power behind a rhetoric curtain of egalitarianism. Replicability was largely founded on a previously established stress on national homogeneity encouraging conformism, uniformity, obedience, peer pressure, cultural standardization, and fear of criticism, while conflating nation, ethnicity, and culture. Although these have been systematically cultivated in peace times through nationalism as a mass-mobilizing ideology (Malešević, 2010), it was during the war, and through the war experience, that they could shape society at a deeper level while removing all traces of constructive dissent. Therefore, World War I provided rulers with an opportunity to practice totalitarianism before its 'lessons' could be systematically applied by fascism and communism. In other words, totalitarianism manifested itself first in the trenches as an 'ancillary' aspect of the relation between war and nationalism.

State-builders were often obsessed with shaping the character, virtue, and manners of their citizens. Barrack and school continuously overlapped: the army served to inculcate the values of patriotism and the education system served to prepare the youth for army life. Whenever war and education were joined, nationalism would provide the ideological foundation underpinning the entire bureaucratic military apparatus. The entire armor of human and technological advantages acquired during the prewar period and the Belle Époque through unprecedented accumulation, military growth, free-market expansion, and interstate competition was ultimately unleashed in World War I.

Despite our self-assuring confidence that the postwar order somehow protects us from repeating the tragedies of the past, new and more devastating crises quite possibly loom ahead. The environmental crisis, with its trail of wars, displacements, hunger, epidemics, economic meltdown, and health disasters, may put an unprecedented pressure on existing institutions by spawning new unforeseeable relationships between war and nationalism.

See also: First World War, The; French Revolution, The; Genocide and War; Militarism; Military Sociology; Military, War, and Politics; Nation-State and War; National Socialism and Fascism; Nationalism, Sociology of; Nationalism: General; Nations and Nation-States in History; War Propaganda; War and Democracy; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Aaslestad, Katherine, 2012. Napoleonic rule in German central europe: compliance and resistance. In: Broers, M., Hicks, P., Guimera, A. (Eds.), The Napoleonic Empire and the New European Political Culture. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire/New York, pp. 160–172.

Altõnay, Ayşe Gül, 2004. The Myth of the Military-Nation: Militarism, Gender, and Education in Turkey, first ed. Palgrave Macmillan, New York.

Bartov, Omer, 1996. Murder in Our Midst: The Holocaust, Industrial Killing, and Representation. Oxford University Press, Oxford.

Bartov, Omer, 2001. The Eastern Front, 1941–45: German Troops and the Barbarisation of Warfare. Palgrave, Houndmills, Basingstoke.

Bartrop, Paul, 2002. The relationship between war and genocide in the twentieth century: a consideration. Journal of Genocide Research 4 (4), 519–532.

Bell, David A., 2008. The First Total War: Napoleon's Europe and the Birth of Modern Warfare. Bloomsbury, London.

Berger, Stefan, 2008. Germany. In: Herb, G.H., Kaplan, D.H. (Eds.), Nations and Nationalism. A Global Historical Overview. ABC-CLIO, Santa Barbara, CA, pp. 609–622.

Bond, Brian, 1986. War and Society in Europe, 1870–1970. Oxford University Press, Oxford.

Colin, Jean, 1900. L'education militaire de Napoleon. R. Chapelot & Ce, Paris.

Colley, Linda, 1992. Britons: Forging the Nation 1707–1837. Yale University Press, New Haven, CT.

Colley, Linda, 2002. Captives: Britain, Empire and the World 1600–1850. Jonathan Cape, London.

Conversi, Daniele, 2007. Homogenisation, nationalism and war: should we still read Ernest Gellner? Nations and Nationalism 13 (3), 371–394.

Conversi, Daniele, 2008. 'We are all equals!' Militarism, homogenization and 'egalitarianism' in nationalist state-building (1789–1945). Ethnic and Racial Studies 31 (7), 1286–1314.

Conversi, Daniele, 2009. Art, nationalism and war: political futurism in Italy (1909–1944). Sociology Compass 3 (1), 92–117.

Conversi, Daniele, 2012. Modernism and nationalism. Journal of Political Ideologies 17 (1), 13–34.

Davis, Mike, 2001. Late Victorian Holocausts: El Niño Famines and the Making of the Third World. Verso, London.

Diamond, Jared M., 2003. Guns, Germs, and Steel : The Fates of Human Societies. W.W. Norton, New York.

Eckhardt, William, 1992. Civilizations, Empires, and Wars: A Quantitative History of War. McFarland, Jefferson, NC.

Esdaile, Charles (Ed.), 2004. Popular Resistance in the French Wars : Patriots, Partisans and Land Pirates. Palgrave Macmillan, Basingstoke.

Fettweis, Christopher J., 2003. War as catalyst: moving world war II to the center of Holocaust scholarship. Journal of Genocide Research 5 (2), 225–236.

Forrest, Alan I., 2009. The Legacy of the French Revolutionary Wars: The Nation-in-Arms in French Republican Memory. Cambridge University Press, Cambridge.

Forrest, Alan I., 2012. Policing, rural revolt and conscription in napoleonic France. In: Broers, M., Hicks, P., Guimera, A. (Eds.), The Napoleonic Empire and the New European Political Culture. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire/New York, pp. 49–58.

Gentile, Emilio, 2003. The Struggle for Modernity: Nationalism, Futurism, and Fascism. Praeger, Westport, CT.

Giddens, Anthony, 1987. The Nation-State and Violence. University of California Press, Berkeley.

Girardet, Raoul, 1953. La société militaire dans la France contemporaine, 1815–1939. Plon, Paris.

Harries, Meirion, Harries, Susie, 1991. Soldiers of the Sun: The Rise and Fall of the Imperial Japanese Army. Heinemann, London. New York: Random House.

Jones, Adam, 2010. Genocide: A Comprehensive Introduction, second ed. Taylor and Francis, London.

Keegan, John, 1999. The First World War, first American ed. A. Knopf, New York.

Levene, Mark, 2005. Genocide in the Age of the Nation State: The Rise of the West and the Coming of Genocide. I.B. Tauris, London.
Levene, Mark, 2008. The vendée: a paradigm shift? In: Jones, A. (Ed.), Genocide. SAGE, London.
Levi, Margaret, 1997. Consent, Dissent, and Patriotism. Cambridge University Press, Cambridge.
Lynn, J., 1996. The Bayonets of the Republic: Motivation and Tactics in the Army of Revolutionary France, 1791–94. Westview, Boulder.
Malešević, Siniša, 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.
Mann, Michael, 2005. The Dark Side of Democracy: Explaining Ethnic Cleansing. Cambridge University Press, Cambridge.
Markusen, Eric, Kopf, David, 1995. The Holocaust and Strategic Bombing : Genocide and Total War in the Twentieth Century. Westview Press, Boulder.
Melson, Robert F., 1992. Revolution and Genocide: On the Origins of the Armenian Genocide and the Holocaust. University of Chicago Press, Chicago.
Mondini, Marco, 2006. Between subversion and coup d'etat: military power and politics after the great war (1919–1922). Journal of Modern Italian Studies 11 (4), 445–464.
Moran, Daniel, 2003. Introduction : the legend of the levee en masse. In: Moran, D., Waldron, A. (Eds.), The People in Arms: Military Myth and National Mobilization since the French Revolution. Cambridge University Press, Cambridge, pp. 1–7.
Paret, Peter, 1993. Understanding War: Essays on Clausewitz and the History of Military Power. Princeton University Press, Princeton.
Planert, Ute, 2012. Resistance to napoleonic reform in the grand Duchy of Berg, the kingdom of the Westphalia and the South German states. In: Broers, M., Hicks, P., Guimera, A. (Eds.), The Napoleonic Empire and the New European Political Culture. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire/New York, pp. 148–159.
Posen, Barry R., 1993. Nationalism, the mass army, and military power. International Security 18 (2), 80–124.
Robespierre, Maximilien, 2007. Extracts from 'on the war' (1792). In: Žižek, S. (Ed.), Robespierre. Virtue and Terror. Verso, London, pp. 26–35.
Sassen, Saskia, 2014. Expulsions. Harvard University Press, Cambridge, MA.
Scurr, Ruth, 2006. Fatal Purity: Robespierre and the French Revolution, first ed. Metropolitan Books, New York. London: Vintage.
Secher, Reynald, 2003. A French Genocide: The Vendée. University of Notre Dame Press, Notre Dame, IN.
Shaw, Martin, 2003. War and Genocide: Organized Killing in Modern Society. Polity Press/Blackwell, Cambridge.
Showalter, Dennis, 2010. Germany, britain and warmaking. In: Neilson, K., Kennedy, G. (Eds.), The British Way in Warfare: Power and the International System, 1856–1956. Essays in Honour of David French. Ashgate, Farnham, Surrey, pp. 135–158.
Tilly, Charles, 1985. War making and state making as organized crime. In: Evans, P.B., Rueschemeyer, R., Skocpol, T. (Eds.), Bringing the State Back In. Cambridge University Press, Cambridge, pp. 169–189.
Tilly, Charles, 1990. Coercion, Capital, and European States, AD. 990–1990. Blackwell, Cambridge, MA, Oxford.
Tolstoy, Lev N., 1901. Patriotism and Government. The Free Age Press, London.
Tolstoy, Lev N., 1905. The Complete Works of Count Tolstoy, vol. 23. J.M. Dent & Co., Boston: Estes/London.
Weber, Eugen, 1976. Peasants into Frenchmen: The Modernization of Rural France, 1870–1914. Stanford University Press, Stanford, CA.
Winter, Jay, 2003. Under cover of war: the Armenian genocide in the context of total war. In: Gellately, R., Kiernan, B. (Eds.), The Specter of Genocide: Mass Murder in Historical Perspective. Cambridge University Press, Cambridge.
Zimmerer, Jürgen (Ed.), 2014. Climate change, environmental violence and genocide. Special issue of the The International Journal of Human Rights, 18 (3).

War and Terrorism, Geography of

James A Tyner and Samuel Henkin, Kent State University, Kent, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recognizing that war and terrorism are inherently geographic, we aim to highlight the expanding geographies of war and terrorism. We believe that all-things-military are increasingly permeating the everyday and everywhere. The blurring of 'militarized' and 'civilian' space and the overlap of military technologies and everyday life is a distinctive feature of contemporary society. The objective of our article is to posit an emphasis on critical geographical understandings of war and terrorism by examining the embodiment of war and terror, the nexus of 'nature' and 'war,' the shifting 'place' of war and terror, and the forwarding of peace geographies.

Introduction

The history of geography's engagement with war, militarism, and (more recently) terrorism is long and embattled. As both Mamadouh (2005) and Black (2005) detail, academic geographers have participated in many ways to both the study – and promotion – of war. Operating largely in a context where war was conceived as both 'natural' and 'necessary,' many geographers used their expertise to advise policy makers and military planners (Mamadouh, 2005). Moreover, geographers contributed to the formation of foreign policies which often justified the use of military intervention as acceptable and, indeed, indispensable. This is seen, for example, in the early writings of prominent geopoliticians, including Friedrich Ratzel, Rudolf Kjellén, Halford Mackinder, and Karl Haushofer.

The geographic contribution to war has not gone unchallenged, as exemplified in the remarks of Yves Lacoste, who laments that *la geographie, ca sert, d'abord, a faire la guerre.* (Geography serves, first and foremost, to wage war. Wisner (1986: 212).) Peter Taylor (2004), likewise, sardonically writes that "God invented war to teach Americans geography." Geography, from such a position, is seen as Mars' handmaiden (Wisner, 1986; Cloud, 2002; Barnes, 2008; Kobayashi, 2009; Loyd, 2012). Recent critiques of both warfare as 'natural' or 'necessary' and the broader 'military–industrial–academic' complex have argued that traditional military geographies are linked with geographical determinism that supports and safeguards the sovereign power and territorial integrity of the state (Barnes, 2008). Wisner (1986: 212) is blunt in his assessment, stating that "the reality of geography is that its methods have been used for a long time by the military for targeting places and people." Consequently, numerous geographers have advocated for a more 'peaceful' or 'pacific' geography (Pepper and Jenkins, 1983; Megoran, 2010, 2011; Inwood and Tyner, 2011).

Given that war – and conflict more broadly – is inherently geographic, regardless of one's position, it should come as no surprise that the literature on geographies of war (and now, terrorism) is vast. As an initial read of the literature, it is possible to classify geographic studies of war into three broad approaches: 'traditional' military geography, spatial analytical, and 'critical.' The first has the longest history and is most associated with the subfield of 'military geography.' Here, the focus in on the interactions of *geography* (narrowly defined) and military operations, campaigns, and battles. As Palka and Galgano (2005: v) explain, "Military geography is the application of geographic information, tools and techniques to military problems." Historically, much of this literature addressed either geography as a facilitating or limiting condition to military operations (cf Winters et al., 1998; Harrison and Passmore, 2008) or as a source of conflict and territorial disputes (cf Diehl, 1991). Research along this vein continues, as some geographers have called for a greater application of geographical techniques, such as spatial analysis using geographic information systems to inform both our empirical documentation and understanding of war and terrorism (Siebeneck et al., 2009; O'Loughlin et al., 2010; Laygo et al., 2012) but also, more proactively, as counterterrorism tools (Beck, 2003; Shroder, 2005; Sui, 2008).

A second dominant approach – with strong connections with international relations and political science – employs quantitative methods and spatial analytic techniques in an attempt to analyze the location and dispersion of war and terrorism (cf Mandel, 1980; Most and Starr, 1980; Houweling and Siccama, 1985; O'Loughlin, 1986; Kirby and Ward, 1987). In part, this trend has been spurred by changes in the nature of conflict, namely, the shift from predominantly 'international' wars to 'civil' wars. According to O'Loughlin (2005: 85), more civil wars began in the 1990s than in any other post-World War II decade; moreover, not only are civil wars becoming more commonplace, these conflicts are also lasting longer and are more destructive in their consequences. The salience of geography (and spatial relations more specifically) has greatly informed the theoretical and empirical understandings of civil wars. The scholarly study of civil wars has been largely dominated by large-scale quantitative approaches; these have highlighted the *intrinsic* geography of civil wars, and especially the diffusion of conflict and the importance of spatial proximity to conflict. As O'Loughlin and Raleigh (2007: 495) identify, "one of the most important predictors of whether a country will experience a civil war is whether its neighbors have internal strife." Variously termed 'the spatial lag effect' or 'contagion,' this understanding of conflict is predicated on the observation that "countries are frequently involved in a complex nexus of conflict as wars ebb

and flow across borders due to the sanctuaries offered by neighbors, the placement of rebels in areas occupied by sympathetic minorities, the actions of refugees to undermine the forces that caused their flight, and the access of extraterritorial bases to external supporters and sources of income" (O'Loughlin and Raleigh, 2007: 495).

The quantitative modeling of civil wars has not gone unchallenged, however. Many geographers in particular have cautioned against the uncritical use of the contagion metaphor, nothing that such an approach may too readily elide into a staid geographic determinism (Tyner et al., 2011: 3). Consequently, recent work indicates a move to better conceptualize "how interactions between governments and non-state actors give rise to violence rather than to focus exclusively on the attributes of one actor [or state] in isolation" (Gleditsch and Ruggeri, 2010: 300).

The critique of spatial determinism inherent in many quantitative studies of (civil) wars highlight a third dominant approach to geographies of war and terrorism, this being a 'critical' approach. As Woodward (2005: 720) explains, critical studies seek to understand the "wider geographical imprint on militarism and military activity" (Woodward, 2005: 720). Consequently, the recent proliferation of critical geographies, coupled with a greater interface with other disciplines, has stimulated a more comprehensive – and theoretical – engagement with geographies of war and terrorism. Accordingly, it is not only an epistemological challenge to traditional approaches; it is also an attempt to broad the field; to address heretofore unexplored connections between 'all-things-military' and 'geography' (here defined as both topic and discipline). Although wide in scope, critical geographic approaches to war, terror, and militarism are most often forwarded by those working in the subfields of critical geopolitics and feminist geopolitics (Dowler and Sharp, 2001; Cowen and Gilbert, 2007; Bernazzoli and Flint, 2009, 2010; Dowler, 1998, 2012). Here, the "geographic logics and imaginations undergirding war, geopolitics, security, and reconstruction … have been subject to extensive criticism and theoretical elaboration" (Loyd, 2012: 477). In part, this trend has been motivated by the pervasive, ever-expanding 'war on terror' (Tyner, 2005, 2007; Ingram and Dodds, 2012) and the profuse militarization of society (Woodward, 2005, 2013; Gregory and Pred, 2007; Megoran, 2008; Farish and Vitale, 2011). Critical approaches have also been concerned to study not only geographies of 'war' but also (as discussed below) geographies of 'peace'; these efforts have sparked considerable debate on the definitions and meanings of both 'war' and 'peace' as well as the relationship between these two concepts (Kirsch and Flint, 2011).

Such a preliminary overview calls attention to the fact that as the conduct – and location – of conflict shifts, so too does the geographic study of war and terrorism. For example, recent military conflicts in Afghanistan and Iraq, coupled with the seemingly never-ending 'war on terror' has sparked considerable debate over the temporality of militarized conflict, that war (broadly conceived) has become an enduring feature of the twenty-first century. For Gregory (2011), however, war has become a prominent feature *everywhere*. From the administrative divisions of the earth into military spheres (e.g., the expansion of the United States' military command to cover all areas of the world), to the shadowy *battle-space* of cyber wars, it becomes glaringly obvious that war, terror, and militarism are pervasive. Indeed, feminist geographers (especially) have argued that the militarization of the everyday is central to the extension of state power into the daily and, even intimate, interactions of its governed population (Dowler, 2012: 490; see also Cowen, 2008). In this article we emphasize the expanding geographies of war and terrorism. We take our cue, in part, from the recent work of Derek Gregory and his articulation of 'the everywhere war.' We concur that all-things-military are increasingly permeating the everyday, everywhere; and that there has been a blurring of once distinct areas of 'operation,' between 'military' space and 'civilian' space; and that the overlap of military technologies and everyday life is a distinctive feature of contemporary society.

The Embodiment of War

A guiding principle of our article is to emphasize the *embodiment of war and terror*. It is necessary to not lose sight of the lived dimension of war and terror; that all manner of flora and fauna – and not just 'humans' – have been and continue to be affected (Slim, 2008; Tyner, 2009). As Loyd (2009b) remarks, "war is not healthy for children and other living things." In fact, geography's turn toward the everyday within the context of militarism is in part a reaction to the changing 'nature' of warfare (Ramonet, 2004; Kaldor, 2006; Brooker, 2010). For example, if one considers the totality of armed conflicts that occurred during the late twentieth and early twenty-first centuries, approximately 20 million people were killed and 60 million wounded in declared wars, civil wars, and other forms of mass violence; upward of 80% of these casualties were 'civilians' (Slim, 2008). Such a grim tally sheet highlights that categories previously and uncritically conceived as discreet and distinct are anything but.

No longer confined to 'traditional' or 'conventional' battlefields, warfare in the twenty-first century is all-encompassing. Indeed, war is no longer defined simply as a major armed conflict between states, but rather includes 'low-intensity' conflicts, 'stateless' wars, and 'dirty wars.' In parallel fashion, an entirely new arsenal of euphemisms have concocted in an attempt to minimize the destructiveness of warfare: 'smart-bombs,' 'precision war,' 'bloodless combat,' and 'collateral damage,' to name but a few. In critique of these developments, feminist and critical geopolitics in particular have significantly expanded how geography (as a discipline) approaches war and terrorism and, specifically, have forwarded greater engagement with the 'embodiment' of militarized conflict (Dowler, 1998, 2012; Dowler and Sharp, 2001; Hyndman, 2004, 2007; Fluri, 2009).

Feminist approaches, in part, seek to understand the complex relationship between macrolevel geopolitical policies, practices, and processes with that of the 'everyday.' More specific to geographies of war and terror, feminist perspectives identify, analyze, and critique the formation of various constructed subjectivities at scales usually excluded from conventional understandings of war and terrorism (Dowler, 2012). Both traditional military geographies and spatial analytical

approaches, for example, are often conducted at the national scale and make use of massive aggregate data sets; such approaches tend to provide a disembodied narrative that renders the production and relations of everyday life invisible. Moreover, feminist accounts have called attention to the emotions and embodiment of warfare, to the horrors and ugliness that are often (deliberately) hidden from view. As Griffin (1992: 11) writes, "we can speak about the airplanes and the kinds of bombs. We can even number the dead. But the actual dismemberment of bodies must only be hinted." Feminists and other critical geographers have refused to remain silent, providing embodied accounts of war and violence in contrast to detached, emotionless military histories, or heroic and overly virile glorifications of war (Tyner and Henkin, 2014).

Nature and War

The nexus of 'nature' and 'war' has long been a stalwart of geography. Traditionally, geographers have engaged in what has been described as a 'terrain-and-tactics' approach (Woodward, 2013). Here, the emphasis is directed to understanding how physical geography influences the conduct and context of military operations. Both topography and weather, for example, have long been studied as inhibiting factors that may alter the course of battle. Nature, so conceived, has been something to be overcome and controlled.

More recently geographers have sought to disentangle the complex relations between 'nature' (or, alternatively, the environment) and conflict. As Le Billon (2005: 217) explains "competition over natural resources has figured prominently among explanations of armed conflicts, from Malthusian fears of population growth and land scarcity to national security interests over resources defined as 'strategic' because of their industrial or military use, such as oil and uranium." Research, consequently, has examined the salience of both resource 'abundance' and resource 'scarcity' toward the generation and continuation of armed conflict (Homer-Dixon, 1999; Le Billon, 2001, 2007; Pelosi and Watts, 2001; Le Billon and Cervantes, 2009). This work has proved catalytic in the development of 'environmental security' studies (Dalby, 1992, 2009).

The interface of conflict and the environment is not limited to the presence or absence of resources. Considerable research has focused on the impact of climate change on 'national security' and the potential for armed conflict (Barnett, 2003; Barnett and Adger, 2007; Nordås and Gleditsch, 2007; Gleditsch, 2012). As Barnett (2003: 7) summarizes, climate change "poses significant risks to the livelihoods, culture and health of many millions of people in many different social and ecological and contexts"; such risks are not limited solely to the potential flooding of small island states, but may also include altered climate systems, potentially leading to famine and outbreaks of disease.

Geographers have also directed attention to the impact of both violent conflict (narrowly conceived) and militarism (more broadly) on the landscape. On the one hand, geographic research has documented how the rise of an industrial militarism, as exhibited by the use of landmines and cluster bombs, chemical weapons, and depleted uranium, has left a distinct – and lethal – imprint on the landscape (Wilson, 1988; Hupy, 2006; Davis et al., 2007; Hupy and Schaetzl, 2008; Tyner, 2010; Hupy and Koehler, 2012). Much of this work has adopted a critical perspective, noting that the military detritus of warfare continues to pose health risk to both people and the environment years after the violence has supposedly ended. On the other hand, geographic research has documented (and often critiqued) the production of militarized landscapes (Woodward, 1999, 2001, 2013; Pearson, 2012). This is seen, for example, in the study of military installations (Yamazaki, 2011), training and proving grounds (Watts, 1998; Woodward, 1999), sites of weapons testing, and places of military occupation (Weizman, 2007). A subfield of this literature has engaged in the debates over the 'benefits' of military installations, arguing that such landscapes are beneficial to environmental conservation and preservation (Westing, 1992; Watts, 1998; Havlick, 2007, 2011). Military installations, it is argued, ward off rampant industrialization and urbanization while simultaneously providing refuge to wildlife and plant life. Critics counter, however, that these studies divert attention from the larger environmental impacts of militarization and also obscure the incredible destructiveness of warfare (cf Sanders, 2009).

The Place of War

As the form and function of warfare and terrorist acts change, so too does the place of conflict itself. The increased prevalence of 'stateless' wars has, for example, irrevocably altered our understanding of war, violence, and conflict. 'New' wars are not only fought with new technologies; they are also fought on new terrains. Stephen Graham (2004, 2009, 2010, 2012) and Kenneth Hewitt (1983, 2009), in particular, have forwarded a reworking of both military geography and geopolitics to account for the destruction meted out against cities. Subsumed under the term 'urban geopolitics' this literature has directed attention not only to the targeting of civilians living within urban environments, but also to the deliberate targeting and destruction of cities as a military objective: *urbicide*. Urbicide, on the one hand, entails the deliberate annihilation (killing) of the material, built environment (Coward, 2004, 2008; Fregonese, 2009, 2012; Hewitt, 2009; Ramadan, 2009). Here, violence directed toward cities should not be viewed as 'secondary' or 'collateral' to the targeting of civilians and military personnel; rather, the violence must be seen as a practice of intentional annihilation of the cities themselves (Hewitt, 2009). From this perspective, cities are targeted because of their symbolism as concretized in architecture, but also because cities denote a particular (and presumed 'immoral') way of life. On the other hand, as Coward (2004, 2008) argues, urbicide also entails an attack on plurality; of the perceived cosmopolitanism exemplified by urban areas; in short, an attack on heterogeneous existence. From this vantage point, the destruction of cities is also a destruction of the possibilities of heterogeneity.

Beyond the changing 'terrain' of war and terror, geographers have also considered what we may call the 'immateriality' of militarism. Here, connections are made between the high-tech militarism of both armed conflict (e.g., cyberwar) and the everyday. Louise Amoore (2009), for example, forwards the concept of 'algorithmic war.' Intimately associated with the ever-expanding and never-ending 'war on terror,' Amoore (2009: 50) highlights the emerging geography of securitization and militarization of everyday life; she notes that "from the remote sensing to bodies on a railway platform, to the securing of identity via biometric algorithms, or the profiling of risk at the airport, the practices of the war on terror exceed any clear distinction between military/civil/commercial spheres." She (p. 50) concludes that "what we are seeing is a stitching together of the mundane and prosaic calculations of business, the security decisions authorized by the state, and the mobilized vigilance of a fearful public" (see also Amoore and de Goede, 2005; Amoore, 2006, 2007; Katz, 2007).

Peace and War

Geography as an academic discipline has a long history of peace-related research. Two early efforts include *The Geography of War and Peace* (Pepper and Jenkins, 1985) and *The Quest for Peace: Transcending Collective Violence and War Among Societies, Cultures and States* (Vayrynen et al., 1987). In part, the promotion of more peaceful geographers was in response to the specter of nuclear annihilation. More recently, geographers have critiqued not only the naturalness and necessity of war (Kobayashi, 2009; Loyd, 2009a,b, 2011, 2012; Gregory, 2010; Koopman, 2010, 2011; Ross, 2011; Williams and McConnell, 2011), but have also promoted a more normative vision of geography as an academic discipline (Megoran, 2008, 2010, 2011; Inwood and Tyner, 2011).

Such efforts have not been easy. As Williams and McConnell (2011: 927) highlight, "our understanding of what peace looks like, and how to research it, remains underdeveloped and analysis of its complexities and contradictions problematically lacking." Indeed, considerable discussion has centered around the politics and contestation over the 'meaning' and 'definition' of peace; this work has been spurred, partially, out of a recognition of the false dichotomy between 'war' and 'peace' as well as 'violence' and 'nonviolence.' Drawing inspiration from Galtung's (1969) influential separation of 'direct' and 'structural' violence, many geographers have destabilized the epistemological foundations of a range of artificial, socially constructed binaries, including but not limited to war, peace, violence, nonviolence, conflict, and postconflict (Kirsch and Flint, 2011; Ross, 2011; Loyd, 2012). Flint (2011: 31) accordingly calls for an approach that views the spaces of 'peace' and 'war' as intertwined; he suggests that social relations are basis of construction of both and concludes that instead of discrete categories, it is more productive to think of a continuum of social settings in which power relations and political differences always exist, but in varying degrees of consensus and nonviolence compared to disagreement and violence. Much of this work has, subsequently, drawn attention to the ways in which such false dichotomies fetishize the unremitting processes of the 'production of violence' under the current mode of global militarization (Farish and Vitale, 2011).

For Williams and McConnell (2011: 929; see also Loyd, 2012), it is necessary to understand peace as a social and spatial process, such as "shifts attention towards issues of power and politics and opens up questions on contingency, agency and embodiment …. Subsequently, the spectrum of violence and non-violence is exposed for investigation and peace revealed to be a highly differentiated landscape." By extension, Springer (2011, 2012) has challenged geographers to consider 'violence' also as a process. For Springer (2012: 138) "to treat the material expression of violence only through its directly observable manifestation is a reductionist appraisal" and "ignores the complexity of the infinite entanglements of social relations." It is more appropriate, therefore, to conceive of violence not as 'thing' or 'event' or even 'act' but rather as a process; violence as *moment* implies *momentum*. In other words, the imposition of political-economic structures designed to 'make life' are carried forward with a momentum far beyond the initial act. Violence conceived as such therefore redirects attention toward those actions and relations that are hidden from empirical observation.

Those scholars working in the areas of peace have enlarged the discussion to address other forms of conflict (and terrorism) that have traditionally been neglected in the geographic study of war and terror. Both feminist and critical geographers, in particular, have highlighted the violence directed toward (especially) women's bodies – as manifest, for example, in the geographies of intimate partner violence and rape (Katz, 2007; Fluri, 2009; Tyner, 2012). This research has, by extension, documented how the violence of sexism, racism, classism, genderism, ableism, ageism, homophobia, and xenophobia – to name but a few – continue to be legitimized through the production of militarized and normalized geographical imaginations. As Loyd (2009a: 865–866) concludes, "to understand who is made most vulnerable where and how socially produced harms are naturalized discursively and materially, it is necessary to theorize specific economic, political, and social relations of oppression and domination and how they articulate (or intersect) in particular historical, geographic moments."

See also: Geopolitics; Landscape; Military Geography; Political Geography; Postcolonial Geography; State, Geography of; Territoriality: Geographical.

Bibliography

Agnew, John, Gillespie, Thomas, Gonzales, Jorge, Min, Brian, 2008. Baghdad nights: evaluating the U.S. military 'surge' using nighttime light signatures. Environment and Planning A 40, 2285–2295.

Amoore, Louise, 2006. Biometric borders: governing mobilities in the war on terror. Political Geography 25, 336–351.

Amoore, Louise, 2007. Vigilant visualities: the watchful politics of the war on terror. Security Dialogue 38, 215–232.

Amoore, Louise, 2009. Algorithmic war: everyday geographies of the war on terror. Antipode 41, 49–69.

Amoore, Louise, de Goede, Marieke, 2005. Governance, risk and dataveillance in the war on terror. Crime, Law and Social Change 43, 149–173.
Barnes, Trevor, 2008. Geography's underworld: the military-industrial complex, mathematical modeling and the quantitative revolution. Geoforum 39, 3–16.
Barnett, Jon, 2003. Security and climate change. Global Environmental Change 13, 7–17.
Beck, Richard, A., 2003. Remote sensing and GIS as counterterrorism tools in the Afghanistan war: a case study of the Zhawar Kili region. The Professional Geographer 55, 170–179.
Barnett, Jon, Adger, W.N., 2007. Climate change, human security and violent conflict. Political Geography 26, 639–655.
Bernazzoli, Richelle, Flint, Colin, 2009. Power, place, and militarism: toward a comparative geographic analysis of militarization. Geography Compass 3, 393–411.
Bernazzoli, Richella, Flint, Colin, 2010. "Embodying the garrison state? Everyday geographies of militarization in American society. Political Geography 29, 157–166.
Black, Jeremy, 2005. Geographies of war: the recent historical background. In: Flint, Colin (Ed.), The Geography of War and Peace: From Death Camps to Diplomats. Oxford University Press, Oxford, pp. 19–25.
Brooker, Paul, 2010. Modern Stateless Warfare. Palgrave Macmillan, New York.
Cloud, John, 2002. American cartographic transformations during the Cold War. Cartography and Geographic Information Science 29, 261–282.
Coward, Martin, 2004. Urbicide in Bosnia. In: Graham, Stephen (Ed.), Cities, War and Terrorism: Towards an Urban Geopolitics. Blackwell, Malden, MA, pp. 154–171.
Coward, Martin, 2008. Urbicide: The Politics of Urban Destruction. Routledge, New York.
Cowen, Deborah, 2008. Military Workfare: The Soldier and Social Citizenship in Canada. University of Toronto Press, Toronto.
Cowen, Deborah, Gilbert, Emily (Eds.), 2007. War, Citizenship, Territory. Routledge, New York.
Dalby, Simon, 1992. Ecopolitical discourse: 'environmental security' and political geography. Progress in Human Geography 16, 503–522.
Dalby, Simon, 2009. Security and Environmental Change. Polity, Malden, MA.
Davis, J.S., Hayes-Conroy, J.S., Jones, V.M., 2007. Military pollution and natural purity: seeing nature and knowing contamination in Vieques, Puerto Rico. GeoJournal 69, 165–179.
Diehl, P.F., 1991. Geography and war: a review and assessment of the empirical literature. International Interactions 17, 11–27.
Dowler, Lorraine, 1998. And they think I'm just a nice old lady: woman and war in Belfast, Northern Ireland. Gender, Place and Culture 5, 159–176.
Dowler, Lorraine, 2012. Gender, militarization and sovereignty. Geography Compass 6, 490–499.
Dowler, Lorraine, Sharp, Joanne, 2001. A feminist geopolitics? Space and Polity 5, 165–176.
Farish, Matthew, Vitale, Patrick, 2011. Locating the American military-industrial complex: an introduction. Antipode 43, 777–782.
Flint, Colin (Ed.), 2005. The Geography of War and Peace: From Death Camps to Diplomats. Oxford University Press, Oxford.
Flint, C., 2011. Intertwined spaces of peace and war: the perpetual dynamism of geopolitical landscapes. In: Kirsch, S., Flint, C. (Eds.), Reconstructing Conflict: Integrating War and Post-War Geographies. Ahsgate, Franham, pp. 31–48.
Fluri, J.L., 2009. Geopolitics of gender and violence 'from below. Political Geography 28, 466–486.
Fregonese, Sara, 2009. The urbicide of Beirut? Geopolitics and the built environment in the Lebanese civil war (1975–1976). Political Geography 28, 309–318.
Fregonese, Sara, 2012. Urban geopolitics 8 years on. Hybrid sovereignties, the everyday, and geographies of peace. Geography Compass 6, 290–303.
Galtung, Johan, 1969. Violence, peace, and peace research. Journal of Peace Research 6, 167–191.
Gleditsch, N.P., 2012. Whither the weather? Climate change and conflict. Journal of Peace Research 49, 3–9.
Gleditsch, K., Ruggeri, A., 2010. Political opportunity structures, democracy, and civil war. Journal of Peace Research 47, 299–310.
Graham, Stephen (Ed.), 2004. Cities, War, and Terrorism: Towards an Urban Geopolitics. Blackwell, Malden, MA.
Graham, Stephen, 2009. Cities as battlespace: the new military urbanism. City 13, 383–402.
Graham, Stephen, 2010. Cities under Siege: The New Military Urbanism. Verso, London.
Graham, Stephen, 2012. When life itself is war: on the urbanization of military and security Doctrine. International Journal of Urban and Regional Research 36, 136–155.
Gray, C.H., 1997. Postmodern War: The New Politics of Conflict. The Guilford Press, New York.
Gregory, Derek, 2010. War and peace. Transactions of the Institute of British Geographers 35, 154–186.
Gregory, Derek, 2011. The everywhere war. The Geographical Journal 177, 238–250.
Gregory, Derek, Pred, Allan (Eds.), 2007. Violent Geographies: Fear, Terror and Political Violence. Routledge, New York.
Griffin, Susan, 1992. A Chorus of Stones. Doubleday, New York.
Harrison, S., Passmore, D.G., 2008. The Ardennes campaigns of 1944: some theoretical considerations on the role of topography. In: Nathanail, C.P., Abrahart, R.J., Bradshaw, R.P. (Eds.), Military Geology and Geography: History and Technology. Land Quality Press, Nottingham, pp. 199–206.
Havlick, David, 2007. Logics of change for military-to-wildlife conversions in the United States. GeoJournal 69, 151–164.
Havlick, David, 2011. Disarming nature: converting military lands to wildlife refuges. The Geographical Review 101, 183–200.
Hewitt, Kenneth, 1983. Place annihilation: area bombing and the fate of urban places. Annals of the Association of American Geographers 73, 257–284.
Hewitt, Kenneth, 2009. Proving grounds of urbicide: civil and urban perspectives on the bombing of capital cities. ACME: An International E-Journal for Critical Geographies 8, 340–375.
Homer-Dixon, Thomas, 1999. Environment, Scarcity and Violence. Princeton University Press, Princeton.
Houweling, Henk, Siccama, January 1985. The epidemiology of war, 1960–1980. Journal of Conflict Resolution 29, 641–663.
Hupy, J.P., 2006. The long-term effects of explosive munitions on the WWI battlefield surface of Verdun, France. Scottish Geographical Journal 122, 167–187.
Hupy, J.P., Schaetzl, Randall, 2008. Soil development on the WWI battlefield of Verdun. Geoderma 145, 37–49.
Hupy, J.P., Koehler, Thomas, 2012. Modern warfare as a significant form of zoogeomorphic disturbance upon the landscape. Geomorphology 157/158, 169–182.
Hyndman, Jennifer, 2004. Minding the gap: bridging feminist and political geography through geopolitics. Political Geography 23, 307–322.
Hyndman, Jennifer, 2007. Feminist geopolitics revisited: body counts in Iraq. The Professional Geographer 59, 35–46.
Hyndman, Jennifer, 2010. The question of 'the political' in critical geopolitics: querying the 'child soldier' in the 'War on Terror,'. Political Geography 29, 247–255.
Ingram, Alan, Dodds, Klaus (Eds.), 2012. Spaces of Security and Insecurity: Geographies of the War on Terror. Ashgate, Aldershot.
Inwood, Joshua, Tyner, J.A., 2011. "Geography's pro-peace agenda: an unfinished project. ACME: An International E-Journal for Critical Geographies 10, 442–457.
Kaldor, Mary, 2006. New & Old Wars: Organized Violence in a Global Era, second ed. Polity, Malden, MA.
Katz, Cindi, 2007. Banal terrorism: spatial fetishism and everyday insecurity. In: Gregory, Derek, Pred, Allan (Eds.), Violent Geographies: Fear, Terror, and Political Violence. Routledge, New York.
Kirby, Andrew, Ward, Michael, 1987. The spatial analysis of peace and war. Comparative Political Studies 20, 293–313.
Kirsch, Scott, Flint, Colin (Eds.), 2011. Reconstructing Conflict: Integrating War and Post-War Geographies. Ashgate, Farnham, UK.
Kobayashi, Audrey, 2009. Geographies of peace and armed conflict: introduction. Annals of the Association of American Geographers 99, 819–826.
Koopman, Sara, 2010. Making space for peace: international accompaniment as alter-geopolitics. Antipode 42, 231–235.
Koopman, Sara, 2011. Let's take peace to pieces. Political Geography 30, 193–194.
Lacoste, Yves, 1976. La Geographie, ca Sert, D'abord, a Faire la Guerre. Maspero, Paris.
Laygo, Katrina, Gillespie, T.W., Rayo, Noel, Garcia, Erin, 2012. Drone bombings in the federally administered tribal areas: public remote sensing applications for security monitoring. Journal of Geographic Information System 4, 136–141.
Le Billon, Philippe, 2001. The political ecology of war: natural resources and armed conflict. Political Geography 20, 561–584.
Le Billon, Philippe, 2005. The geography of 'resource wars,'. In: Flint, Colin (Ed.), The Geography of War and Peace: From Death Camps to Diplomats. Oxford University Press, Oxford, pp. 217–241.
Le Billon, Philippe, 2007. Geographies of war: perspectives on 'resource wars,'. Geography Compass 1, 163–182.
Le Billon, Philippe, Cervantes, Alejandro, 2009. Oil prices, scarcity, and geographies of war. Annals of the Association of American Geographers 99, 836–844.
Loyd, J.M., 2009a. A microscopic insurgent: militarization, health and critical geographies of violence. Annals of the Association of American Geographers 99, 863–873.
Loyd, J.M., 2009b. War is not healthy for children and other living things. Environment and Planning D: Society and Space 27, 403–424.
Loyd, J.M., 2011. 'Peace is our only shelter': questioning domesticities of militarization and white privilege. Antipode 43, 845–873.

Loyd, J.M., 2012. Geographies of peace and antiviolence. Geography Compass 6, 477–489.
Mamadouh, Virginie, 2005. Geography and war, geographers and peace. In: Flint, Colin (Ed.), The Geography of War and Peace: From Death Camps to Diplomats. Oxford University Press, Oxford, pp. 26–60.
Mandel, Robert, 1980. Roots of modern interstate border disputes. Journal of Conflict Resolution 24, 427–454.
Megoran, Nic, 2008. Militarism, realism, just war, or nonviolence? Critical geopolitics and the problem of normativity. Geopolitics 13, 473–497.
Megoran, Nic, 2010. Towards a geography of peace: Pacific geopolitics and evangelical Christian crusade apologies. Transactions of the Institute of British Geographers 35, 382–398.
Megoran, Nic, 2011. War and peace? an agenda for peace research and practice in geography. Political Geography 30, 178–189.
Most, Benjamin, Starr, Harvey, 1980. Diffusion, reinforcement, geopolitics and the spread of war. American Political Science Review 74, 932–946.
Nordås, Ragnhild, Gleditsch, Nils P., 2007. Climate change and conflict. Political Geography 26, 627–638.
O'Loughlin, John, 1986. Spatial models of international conflicts. Annals of the Association of American Geographers 76, 63–80.
O'Loughlin, John, 2005. The political geography of conflict: civil wars in the hegemonic shadow. In: Flint, C. (Ed.), The Geography of War and Peace: From Death Camps to Diplomats. Oxford University Press, Oxford, pp. 85–110.
O'Loughlin, John, 2005. The war on terrorism, academic publication norms, and replication. The Professional Geographer 57, 588–591.
O'Loughlin, John, Raleigh, Clionadh, 2007. Spatial analysis of civil war violence. In: Cox, K., Low, M., Robinson, J. (Eds.), The SAGE Handbook of Political Geography. SAGE, Thousand Oaks, pp. 493–508.
O'Loughlin, John, Witmer, F.D., Linke, A.M., 2010. The Afghanistan-Pakistan wars, 2008–2009: micro-geographies, conflict diffusion, and clusters of violence. Eurasian Geography and Economics 51, 437–471.
Palka, E., Galgano, F., 2005. Military Geography: From Peace to War. McGraw-Hill, New York.
Pearson, Chris, 2012. Researching militarized landscapes: a literature review on war and the militarization of the environment. Landscape Research 37, 115–133.
Pelosi, N., Watts, M. (Eds.), 2001. Violent Environments. Cornell University Press, Ithaca.
Pepper, David, Jenkins, Alan, 1983. A call to arms: geography and peace studies. Area 15, 202–208.
Pepper, David, Jenkins, Alan (Eds.), 1985. The Geography of Peace and War. Basil Blackwell, New York.
Perkins, Chris, Dodge, Martin, 2009. Satellite imagery and the spectacle of secret spaces. Geoforum 40, 546–560.
Ramadan, Adam, 2009. Destroying Nahr el-Bared: sovereignty and urbicide in the space of exception. Political Geography 28, 153–163.
Ramonet, Ignacio, 2004. Wars of the 21st Century: New Threats, New Fears. Ocean Press, New York.
Ross, Amy, 2011. Geographies of war and the putative peace. Political Geography 30, 197–199.
Sanders, Barry, 2009. The Green Zone: The Environmental Costs of Militarism. AK Press, Edinburgh.
Shroder, Jack, 2005. Remote sensing and GIS as counterterrorism tools in the Afghanistan war: reality plus the results of media hyperbole. The Professional Geographer 57, 592–597.
Siebeneck, L.K., Medina, R.M., Yamada, Ikuho, Hepner, G.F., 2009. Spatial and temporal analyses of terrorist incidents in Iraq, 2004–2006. Studies in Conflict & Terrorism 32, 591–610.
Slim, Hugo, 2008. Killing Civilians: Methods, Madness, and Morality in War. Columbia University Press, New York.
Springer, Simon, 2011. Violence sits in places? Cultural practice, neoliberal rationalism, and virulent imaginative geographers. Political Geography 30, 90–98.
Springer, Simon, 2012. Neoliberalizing violence: of the exceptional and the exemplary in coalescing moments. Area 44, 136–143.
Sui, D.Z. (Ed.), 2008. Geospatial Technologies and Homeland Security: Research Frontiers and Future Challenges. Springer, Netherlands.
Taylor, P.J., 2004. God invented war to teach Americans geography. Political Geography 23, 487–492.
Tyner, J.A., 2005. Iraq, Terror, and the Philippines' Will to War. Rowman & Littlefield, Lanham, MD.
Tyner, J.A., 2007. America's Strategy in Southeast Asia: From the Cold War to the Terror War. Rowman & Littlefield, Lanham, MD.
Tyner, J.A., 2009. War, Violence, and Population: Making the Body Count. The Guilford Press, New York.
Tyner, J.A., 2010. Military Legacies: A World Made by War. Routledge, New York.
Tyner, J.A., 2012. Space, Place, and Violence: Violence and the Embodied Geographies of Race, Sex, and Gender. Routledge, New York.
Tyner, J.A., Henkin, Samuel, 2014. Feminist geopolitics, everyday death, and the emotional geographies of Dang Thuy Tram. Gender, Place and Culture.
Tyner, J.A., Rice, Stian, Shears, Andrew, 2011. State capacity, scalar politics, and the geographic study of civil wars. International Interactions 37, 1–9.
Vayrynen, R., Senghaas, D., Schmidt, C. (Eds.), 1987. The Quest for Peace: Transcending Collective Violence and War Among Societies, Cultures, and States. SAGE Publications, London.
Watts, S.E., 1998. Short-term influence of tank tracks on vegetation and microphytic crusts in shrub steppe habitat. Environmental Management 22, 611–616.
Weizman, Eyal, 2007. Hollow Land: Israel's Architecture of Occupation. Verso, London.
Westing, A., 1992. Protected natural areas and the military. Environmental Conservation 19, 343–348.
Williams, Philippa, McConnell, Fiona, 2011. Critical geographies of peace. Antipode 43, 927–931.
Wilson, S.D., 1988. The effects of military tank traffic on prairie: a management model. Environmental Management 12, 297–403.
Winters, H.A., et al., 1998. Battling the Elements: Weather and Terrain in the Conduct of War. Johns Hopkins University Press, Baltimore, MD.
Wisner, Ben, 1986. Geography: war or peace studies. Antipode 18, 212–217.
Woodward, Rachel, 1999. Gunning for rural England: the politics of the promotion of military land use in the Northumberland National Park. Journal of Rural Studies 15, 17–33.
Woodward, Rachel, 2001. Khaki conservation: an examination of military environmentalist discourses in the British Army. Journal of Rural Studies 17, 201–217.
Woodward, Rachel, 2005. From military geography to militarism's geographies: disciplinary engagements with the geographies of militarism and military activities. Progress in Human Geography 29, 718–740.
Woodward, Rachel, 2013. Military landscapes: agendas and approaches for future research. Progress in Human Geography. http://dx.doi.org/10.1177/0309132513493219 on-line pre-publication.
Yamazaki, Takashi, 2011. The US militarization of a 'host' civilian society: the case of postwar Okinawa, Japan. In: Kirsch, Scott, Flint, Colin (Eds.), Reconstructing Conflict: Integrating War and Post-War Geographies. Ashgate, Farnham, UK, pp. 253–272.

War: Anthropological Aspects, Historical Development of

Sverker Finnström, Uppsala University, Uppsala, Sweden
Carolyn Nordstrom, University of Notre Dame, Notre Dame, IN, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C.R. Nordstrom, volume 24, pp. 16350–16354, © 2001, Elsevier Ltd.

Abstract

In defining war, this article explores the origin and historical development of armed aggression from the earliest human societies to the present. The fact that war is a relatively recent invention in the span of human existence, arising with complex societies, suggests war as neither a biological imperative nor integral to the human condition. The many forms war takes, from ethnic conflict through conventional militaries to guerrilla warfare are considered in order to understand the nature and culture of war. The twentieth century has been the bloodiest in history, and this piece examines the relationships between violence, society, and the exercise of power that help explain this. The changing philosophies and practices of war over time and society show war to be a complex constellation of economic, cultural, and existential, as well as political factors. The various theoretical approaches to war, from those characterizing premodern societies and the rise of the modern state to the present are discussed. In concluding, the future of war, and the new directions theory might take in best understanding war in the aftermath of the 11 September 2001 attacks, are considered.

Sun-tzu, the famous Chinese military expert, began his book *The Art of War* with the words "Warfare is the greatest affair of state, the basis of life and death, the Way (Tao) to survival or extinction. It must be thoroughly pondered and analyzed." These words, 2500 years later, still hold true: war has grown more deadly, and more damaging to human existence.

In the twentieth century alone, over 250 formally declared wars took over 100 million lives. Undeclared wars, including political repression, communal violence, and genocide took millions more; for example, between 50 and 100 million people have been killed by forces and citizens of states in the twentieth century because of their sociocultural and ethnic belongings. If we expand the definition of war to include such armed conflicts as gang warfare, 'the war on drugs' and, more recently, 'the war on terror,' casualties figures rise, though accurate statistics are not available on these forms of violence.

On entering the third millennium, one-third of the world's countries were engaged in some form of political violence. Whether these conflicts are called war or not often depends more on political rhetoric than on an accepted definition of the term. In addition, approximately two-thirds of the world's security forces use human rights abuses to control their populations. The victims tend to label this violence war or dirty war, while the state tends to classify this as defense or counterinsurgency.

The world has not always been characterized by such high levels of violent warfare. Wars today are longer in duration, more deadly, with visible as well as invisible linkages crisscrossing the globe, and kill higher numbers of civilians than wars of preceding centuries. The reasons behind war and the ways in which it is waged change across cultures and time, leading anthropological observers to talk about 'shadows of war' (Nordstrom, 2004), a cycle of 'global war' without any certain ending (e.g., Whitehead and Finnström, 2013), and of globalized 'war machines' feeding on an underprivileged but increasingly flexible and truly transnational post-Fordian work force (e.g., Hoffman, 2011). Also, the changing characteristics of war demonstrate that organized violence is not a fixed and eternal fact of biology, nor an inescapable feature of a Freudian psyche, but a human practice guided by norms of behavior and codes of conduct situated in cultural and economical values.

Defining War

Neither the two world wars, nor the more recent 'world war' raging in the Democratic Republic of the Congo, nor the several hundred local and regional wars since 1900 have brought us closer to a shared understanding of war. Most scholars accept a basic definition of war as the deployment of violence to force opponents to comply with one's will. Many argue that the number of so-called battle-related deaths is the defining aspect of war, others, especially anthropologists, emphasize the need to include also the consequences of structural violence to the understanding and definition of war. Either way, war is organized, group-level, armed aggression rooted in hierarchies of dominance, which assume winners and losers in a contest over resources, people, and power. Yet war is defined differently by the winners and the losers, by historical perspective, by soldiers and pacifists – and in each case the definitions are more politically charged than factually correct. For example, freedom fighter, terrorist, insurgent, rebel, traitor, soldier, deserter, civilian, perpetrator, and victim are all terms variously applied to the same actors by different groups in different contexts seeking to maximize their own political, economical, and moral justifications. Governments define war in their own interests, and militaries are generally loath to admit strategies that entail civilian or noncombatant casualties, torture, and human rights abuses. In using terms such as 'collateral damage,' military powers instead try to trivialize and relativize civilian casualties. As anthropologists often emphasize, the most basic understanding of war is affected by differential and biased reporting; for example, casualty statistics for World War II vary by millions, depending on the nationality and viewpoint of the researcher. Controlling the definitions of war and its consequences is integral to the waging of war (Sluka, 1992).

The ethnographic study of war and peace has added a new dimension in the understanding of political violence. This academic research has demonstrated that war is a far more complex reality than classical definitions positing a violent contest between two or more armed forces seeking a military, and thus political, victory (Warren, 1993; Nordstrom and Robben, 1995). Rather than to settle with any final and universal definition of war, anthropologists have thus come to emphasize the importance of the comparative ethnographic approach with case studies along the war–peace continuum (e.g., Richards, 2005). Soldiers often battle unarmed civilians and not each other – evident from the ethnic cleansing in former Yugoslavia, especially in Bosnia and Kosovo, or the 2 million deaths in the civil war that lead to the independence of South Sudan. Paramilitaries, private militias, death squads, and roving bands of armed predatory gangs patrol war zones. Some operate at the behest of state forces (Sluka, 2000), while others are independent of all sovereign or rebel control, and they may operate in war as well as in peacetimes. Mercenary forces are a global phenomenon today, and range from informal groups such as European mercenaries fighting in Central Africa to the more formal Executive Outcome organization, comprising former apartheid South African soldiers, who broker with governments as well as rebel groups. Battle zones are also home to looters, sex workers, criminals, and profiteers. War zones are a bazaar of international arms and supplies merchants who reap billions of dollars yearly worldwide. International nongovernmental organizations are found in all war zones today, providing services ranging from conflict resolution to humanitarian and development aid (Nordstrom, 2004). Finally, the fronts of wars remain home to the inhabitants. Regardless of formal military regulations mandating the legal role of women, children, and the aged in war, all of these people fight for survival when they find themselves on the frontlines. Armed or unarmed, women defend homes and towns, children are forced to take up arms and fight, and the aged battle forced sieges. The unscrupulous sell out their neighbors for a few coins, and the altruistic set up medical clinics, schools, and trade routes to provide critical resources under bombardment. War, in short, is a social project that sustains the experience of war, making it and its multiple forms of violence routines among other routines in everyday life (Richards, 2005).

The Development of War

War is a fairly recent invention, in terms of the anthropological expanse of human existence. Humans, as a species, have lived 90% of their history without war. Social hierarchies and concepts of ownership appear necessary for the advent of war. The earliest form of human organization was the band: fluid egalitarian groups of nomads. The archeological record indicates that while interpersonal violence was known in bands – determined by puncture and crushing wounds from weapons – it was limited. It did not reach the level of formalized intergroup violence among contending warriors.

The first indications of organized warfare occur as ownership of animals, goods, and property create divisions within societies. With the historical development of tribal societies and protostates comes a differentiation in power, and the emergence of organized intergroup violence (Ferguson and Whitehead, 2000). Such societies did not have standing armies and military institutions separate from general society; warrior status tended to be open to all able-bodied men, and, less commonly, to women. These early years of war were not necessarily a dangerously lethal activity. For many ethnic groups, preparations for war constituted an elaborate ritual process. The rules of engagement were often well delineated: contending factions would meet in full battle regalia and hurl challenges and possibly weapons. Casualties generally brought a halt to the aggressions. Here, it is the display and enactment of power, and not violence, that defines war. Among some communities – the archeological record suggests these were later developments – fighting was much more lethal, though the intent was seldom, if ever, genocide. The goal was to force surrender and extend control over people, property, or territory.

Formalized military institutions and standing armies develop with the rise of the 'state' as a form of political, economic, and social organization. The term state here is used in its broad anthropological sense – originating some 8000 years ago, and not in the political science definition as developing in the mid-1600s. (The latter will be considered in the next section.) So when chiefdoms are replaced by royal families or other, equally centralized governing bodies, standing armies tend to develop as well, while social, gender, and often ethnic inequalities are increasingly codified in laws of land ownership, labor rights, and inheritance. Dispute resolution becomes formalized into judicial systems, and the legitimate use of force is restricted to state leaders and institutions. Contemporary warfare – fought among contenders for power, privilege, gain, and over identity – emerges. Eventually, as we shall see, such developments find fuel from the context of colonial intervention and annexation.

The Changing Nature of War

Contemporary warfare itself has changed dramatically over time and circumstance, giving lie to any notion that war is a 'natural' social phenomenon or a fixed product of overarching political organization (van Creveld, 1991). The era of the modern state provides a good illustration (Holsti, 1996). In Europe, the end of the Thirty Years War (from 1618 to 1648) coincided with the beginning of the modern state (marked by the Treaty of Westphalia). The Thirty Years War depopulated a large part of Central Europe. It was known for its sheer brutality; writers of the time speak of the wanton killing, torture, plunder, and destruction of anyone and anything who found themselves in the path of the aggressors. The levels of violence are attributed to the enduring impact of religious wars and the Inquisition, to the transformations wrought by urbanization and early industrialization, and to the upheavals marking the shift from kingly rule to the modern state.

Over the following two centuries the nihilism characterizing the Thirty Years War gave way to what has been called the gentlemen's war of the Enlightenment period. Formal warfare during this era often, though certainly not always, followed strict rules of conduct and engagement: soldiers fought soldiers

in hand-to-hand combat on battlegrounds apart from human habitation. This was not a new era of war for humankind: Buddhist and Hindu scriptures 2000 years BCE outlined similar 'gentlemen's wars' in Asia.

While military texts tend to focus on these formal military engagements between two contending armies, another form of warfare developed during this period: colonial repression of conquered peoples. It is commonplace to read in amateur ethnographic accounts of the colonial administrators, that war was a constant occupation among non-Western peoples before colonization, but the reader is most often left in ignorance as to what the colonial administrators really meant with the word 'war.' Rather, here the term 'war' was deliberately kept vague so that it could be used in legitimizing imperial and allegedly honorable agendas. The message of these early colonial reports was that colonization brought peace. The reality was far more violent and saw the first genocides of the twentieth century; for example, the annihilation of the Herero people of present-day Namibia. In many ways, these actions presaged the dirty war of contemporary times – wars that brutally targeted unarmed people in attempts to instill political acquiescence.

The colonial encounter eventually gave rise to another distinct form of war: the guerrilla war, the mainstay of wars for independence worldwide. Guerrilla warfare was developed by nonstate actors challenging financially and technologically superior state forces. Classical guerrilla philosophy – institutionalized in the mid-twentieth century by military strategists such as Mao Zedong, Che Guevara, and Ho Chi Minh – postulates that guerrilla forces, by definition, have the support of the broad population, and it is this that gives them indefatigable strength, crucial resources, and moral political superiority. While in many cases this has proven true, it is by no means always so. Guerrilla or insurgency groups such as the Khmer Rouge in Cambodia, Renamo in Mozambique, the Lord's Resistance Army in Uganda, and the Contras in Nicaragua demonstrated that nonstate forces, sometimes even very small, can also use repressive tactics in an effort to control large populations.

In the aftermath of the 11 September 2001 terrorist attacks on the United States, many politicians came to regard the rules and laws of war as obsolete. A fundamental shift in warfare, with truly global repercussions, took place and torture was sanctioned by the democratically elected US government. Then president George W. Bush declared a 'war on terror' and the US-led invasion of Iraq soon developed into a seemingly permanent yet diffuse state of war and insecurity in Iraq as well as in Afghanistan and along the Afghanistan–Pakistan borderlands. The militarization of societies under occupation has deepened, with violent developments that cross-breed in all kinds of global yet diffuse directions. Many observers again reference the idea of a 'military–industrial complex,' formulated by US president Dwight D. Eisenhower as well as, from the other end of the political spectrum, the Trotskyist theory of a 'permanent arms economy,' which was forwarded to explain the sustained economic growth that occurred in the decades following World War II. Back then, as today, wartime capitalism reproduces itself despite the inherently contradictory nature of the capitalist growth cycle (see, e.g., Whitehead and Finnström, 2013).

In the effort to dedramatize the political rhetoric and reclaim the definitional space as well as dislocate global war from the US homeland, the 'war on terror' was eventually renamed 'overseas contingency operations' by the Obama administration. Again it becomes evident, from an anthropological perspective, that the control over the definitions of war is integral to the waging of war.

The Twentieth Century

Overall, wars were longer in duration, more lethal in the cost to human lives, and more destructive to societal systems and the environment in the twentieth century than in preceding times. This era saw the advent of world wars, nuclear war, and modern paramilitary warfare. The most dramatic development of war concerns the ethics of who may and may not be targeted in war. Over 80% of the casualties in World War I were soldiers. But eventually also a country's citizens were to be seen part of that country's war effort, given their role in producing the means of war, or just as potential supporters of a war effort. When this philosophy was married with developments in military technologies such as aerial carpet bombings, the number of noncombatant casualties raises even more (Lindqvist, 2001). Some 50% of all war-related deaths in World War II were noncombatant casualties. This trend escalates in the last half of the twentieth century. The US military intervention in Vietnam resulted in one of the most massive bomb campaigns in history. More than 80% of all casualties in the Vietnam war were noncombatants. At the start of the twenty-first century, civilians account for 90% of all war-related deaths worldwide.

Even if these comparative figures on war casualties can only be estimated in retrospect, and therefore are notoriously uncertain, they certainly indicate that noncombatants rather than only the military enemy are the explicit target in many of today's wars. Not only has the line between combatant and noncombatant grown indistinct, the line between gender and age in soldiering has too. Women die in equal numbers to men, and more children are killed in war today than soldiers. The use of child soldiers has grown in recent years; at the turn of the millennium over 300,000 existed worldwide. If children armed with AK-47s cynically are forced into war because of their effectiveness and loyalty in combat, another truly scary development of global war is the ever-increasing use of deadly high-tech remote-controlled unmanned aerial vehicles or so-called drones.

The serious rise in civilian casualties parallels an increase in the number and sophistication of international organizations (such as the United Nations, the European Union, or the African Union) and international legal bodies (such as the International Criminal Court and the War Crimes Tribunals) seeking to control destructive wars. Today, there are 70 000 protocols protecting human rights. Even if military strategists tend to describe and dehumanize civilian causalities in terms of unintended collateral damage, they also claim the very space of humanitarianism. Military interventions are no longer always clearly distinguishable from humanitarian interventions and vice versa. This, of course, affects global war and how it is perceived on the ground. For example, in the wake of an increasing number of deadly attacks on their personnel, in 2013 Médecins Sans Frontières decided to leave Somalia after more than 20 years there.

Theorizing War

As the predominate means of waging war changes through time, so too do the definitions and theories of war (Simons, 1999). In premodern times, many, including early Christian and Buddhist societies, saw war as inescapable at times, but not as honorable; the postwar period was crafted as one of atonement. The notion of the honorable war develops in the West with the rise of the 'gentleman's war' of the modern state. Carl von Clausewitz, the famous Prussian military expert, codifies war as 'an extension of politics,' placing war directly in the rational politics of Enlightenment philosophy. Warfare, as rational, became justified. As discussed above, the most dangerous examples of this were seen in colonial conquests, often rationalized under the rubric of the 'evolution of civilization' by scholars.

These Enlightenment philosophies wed with the functionalist and structuralist schools in the early 1900s. Here, theoreticians investigated the 'functions' of war, and placed the causes of war in competition over scarce resources, overpopulation, and the increasing complexity of societies. At the same time, psychological and sociobiological theories were popular. They postulated aggression and self-interest as inherent to humans, and therefore as serving an evolutionary purpose.

One contemporary example of such theories on war and human aggression is anthropologist Napoleon Chagnon's work among the Yanomani people in Venezuela (e.g., 1968). With reference to biology, Chagnon argues that Yanomani men are fierce and warlike and that they wage war over women; that Yanomani men who are killers have more wives and children than men who are not; and that war is the prime mover of cultural and human evolution. Chagnon's thesis on warfare stirred one of the biggest and longest debates in anthropology, with the majority of sociocultural anthropologists vigorously questioning Chagnon's conclusions as well as his research ethics. However, in seeing Chagnon's results as evidence of a universal theory on human nature and war, sociobiologists have generally endorsed his work.

The fact that not all societies engage in war, and that the majority of the people in any society at war do not choose to fight was not addressed by the psychological and sociobiological theories. At the same time, twentieth-century political theory was shaped by the advent of World War I and World War II. After World War I, functionalist theories take an idealist cast that postulates the progress of civilization as one that will finally eschew war. In the wake of the vast destruction of World War II, realist theory replaces idealism as the dominant theory in the social and political sciences. Here, war is seen as a natural effect of competition among sovereign states. In both schools the solution lies in creating strong state and international institutions to wage, and ideally to control, war.

As the period of the World War I and World War II gave way to wars for independence, regional wars, and the Cold War, theories of war underwent another revolution. Critical to this shift is the fact that researchers around the world began to experience political violence directly, whether by intention or by accident. The reigning functionalist theories did not fit their observations (Foster and Rubinstein, 1986). Consequently, by the 1990s the anthropological focus had shifted from such theories on the causes, structures, and functions of war to the study of lived experiences of war, especially coping mechanisms among allegedly innocent victims caught in cross fire. Eventually, anthropologists saw the importance to account also for the stories of perpetrators as well as the larger sociopolitical structures (Löfving, 2005). Today, most anthropologists again emphasize both cause and effect, both (objective) structure and (subjective) agency. War structures the lived experiences, motifs, strategies, and tactics of all kinds of agents of war – victims, perpetrators, profiteers, or bystanders – but as these social agents 'navigate terrains of war' (Vigh, 2006), they also structure war.

The advent of nuclear war gave lie to ideals of 'victory' in war – for the first time all sides to a conflict could perish. Wars such as those fought by the United States in Vietnam and by the Union of Soviet Socialist Republics in Afghanistan, the more recent global terrorism of al-Qaeda and other groups, and subsequently, the war on terror laid to rest old notions of the gentleman's war. Dirty wars such as those in Argentina, and genocides such as those conducted by the Nazis in World War II and by Hutu extremists in Rwanda, as well as the Khmer Rough's immeasurable mass killings in Cambodia challenged notions of the inherent functionality or rationality of war. The increases in noncombatant deaths undermined the claims of sociobiology – it became hard to argue that noncombatant deaths, widespread torture, systematic sexual violence, and the death of children were biologically or socially productive acts (Enloe, 2000). Scholars also began to question Clausewitz's 'truism' that war is (predominately) an extension of politics. Certainly it is political, but the recognition of the vast sums of money made in wildcatting valuable resources in war zones and in selling war supplies worldwide made it necessary to integrate economics with politics in the war equation (Kaldor, 1999). The rise of religious, ethnic, and identity dynamics in contemporary conflicts rendered it necessary to account for social and cultural factors in any politicoeconomic analysis (Rupesinghe and Rubio Correa, 1994). And studies of peaceful societies such as the Semai and the Quakers demonstrated that war is not inevitable, nor basic, to the human condition (Gregor, 1996).

In holistically acknowledging the above-mentioned factors, anthropological theorizing on war and culture tend to emphasize two propositions. First, war and culture cannot be studied as things in themselves, and war is only one among many competing social projects along the war–peace continuum. Second, and following the first proposition, in refuting reductionist explanations of any essence of war (or culture), anthropologists promote globally framed indepth and historicized ethnographic case studies of both war and peace. Today, these are carried out in specific locations as well as in transnational and virtual spaces, and sometimes, if needed, even from a distance (Richards, 2005; Robben, 2010; Whitehead and Finnström, 2013).

The Future of War

The dawn of the third millennium is marked by vast differences in war. While the superpowers spend trillions of dollars on high-technology earth- and space-based weapon systems as well as on the software of anthropological intelligence in the form of the highly contested Human Terrain System (Gonzalez,

2009; Price, 2011; Whitehead and Finnström, 2013), the vast majority of today's war casualties are killed by small arms wielded by nonspecialists. The greatest dangers are the most accessible; there are estimated to be 500 million to 1 billion firearms in use today in the world; a lively international black market sells every conceivable implement of war from AK47s to nuclear materials; recipes for chemical weapons and homemade bombs can be found in basic texts; and computer specialists can potentially hack a drone aircraft or wreck nationwide havoc by disrupting a country's basic infrastructural support systems. There is a growing trend among governments to regard and declare cyber-attacks on their state infrastructures as acts of war, as well as for governments themselves to initiate cyber-attacks on infrastructures of proclaimed and nonproclaimed enemies.

All facets of the war industry are set in global interactions (Castells, 1998). Our theories of war must be revised to address these dynamics defining the contemporary world. Theories of war need to address the complexities of war systems that spend fortunes on technological defenses (at the end of the twentieth century military spending worldwide reached 780 billion US dollars per year) while killing with inexpensive conventional weapons – and will delve into the sociocultural factors and economic gains as well as the political quests that underlie these realities. It can also be contended that the killings on the actual battlefields, wherever they are located, are intimately linked to an emerging virtual space created by news and gaming media as well as the technologies of contemporary military violence – such as airborne attack drones, satellite surveillance, stealth airplanes, and the associated use of politically covert assassination operations. This is the character of war today, which has definitively moved beyond the confines of nation-states and through mechanisms such as 'peacekeeping,' terrorist interdiction, 'target killings,' and 'overseas contingency operations,' and now has a global and increasingly permanent character (Whitehead and Finnström, 2013). Consequently, in recent times the anthropological study of war affirms to a renewed interest in the state, state sovereignty, and related questions (e.g., Kapferer and Bertelsen, 2009).

While employing violence in the pursuit of dominance may continue to fuel war, violence may shift from physical killing to a different order of threat and inequality, and dominance might be reckoned along such nonmilitary factors as economics, environmental control and climate change, social viability, or a set of factors as yet unrecognized. In the long term, we should be prepared for the possibility that war, as we used to know it, may not define future conflict. War has not always been a part of the human condition, and perhaps future changes in sociopolitical organization and ethical systems will render war altogether obsolete. Effective research into the lived experiences, causes, and future of war and peace will hone combinations of theoretical inquiry with ethnography – helping to erase arbitrary distinctions between theory and data (Nordstrom, 1997). The greatest advances will be in rethinking the very meanings of war and aggression, going beyond simple biological and rudimentary social explanations to explore the complex interactions of violence and power, capital and resources, economics, survival, and identity both within and across local, regional, and transnational populations.

See also: Conflict and War, Archaeology of: Behavior and Social Organization; Conflict: Anthropological Aspects; First World War, The; Militaries and Militarization, Anthropology of; Military Geography; Military History; National Security Studies and War Potential of Nations; Peace and Nonviolence: Overview of Anthropological Aspects; Refugees in Anthropology; Second World War, The; State: Anthropological Aspects; Tribe; Violence in Anthropology; War, Sociology of; Warfare in History; Wars among Nation-States: Patterns and Causes.

Bibliography

Castells, M., 1998. End of Millennium. Blackwell, London.
Chagnon, N.A., 1968. Yanomamö: The Fierce People. Holt McDougal, New York, NY.
van Creveld, M., 1991. The Transformation of War. The Free Press, New York.
Enloe, C.H., 2000. Maneuvers: The International Politics of Militarizing Women's Lives. University of California Press, Berkeley, CA.
Ferguson, R.B., Whitehead, N.L. (Eds.), 2000. War in the Tribal Zone: Expanding States and Indigenous Warfare, revised ed. School of American Research Press, Santa Fe, NM.
Foster, M.L., Rubinstein, R.A. (Eds.), 1986. Peace and War: Cross-cultural Perspectives. Transaction Books, Oxford.
Gonzalez, R., 2009. American Counterinsurgency: Human Science and the Human Terrain. Prickly Paradigm Press, Chicago, IL.
Gregor, T. (Ed.), 1996. A Natural History of Peace. Vanderbilt University Press, Nashville, TN.
Hoffman, D., 2011. The War Machines: Young Men and Violence in Sierra Leone and Liberia. Duke University Press, Durham, NC.
Holsti, K.J., 1996. The State, War, and the State of War. Cambridge University Press, Cambridge.
Kaldor, M., 1999. New and Old Wars: Organized Violence in a Global Era. Stanford University Press, Stanford, CA.
Kapferer, B., Bertelsen, B.E., 2009. Crisis of the State: War and Social Upheaval. Berghahn Books, Oxford.
Lindqvist, S., 2001. A History of Bombing. The New Press, New York.
Löfving, S., 2005. Silence and the politics of representing rebellion: on the emergence of the neutral Maya in Guatemala. In: Richards, P. (Ed.), No Peace, No War: An Anthropology of Contemporary Armed Conflicts. Ohio University Press, Ohio, OH, pp. 77–97.
Nordstrom, C., 1997. A Different Kind of War Story. University of Pennsylvania Press, Philadelphia, PA.
Nordstrom, C., 2004. Shadows of War: Violence, Power, and International Profiteering in the Twenty-First Century. University of California Press, Berkeley, CA.
Nordstrom, C., Robben, A.C.G.M. (Eds.), 1995. Fieldwork under Fire: Contemporary Studies of Violence and Survival. University of California Press, Berkeley, CA.
Price, D.H., 2011. Weaponizing Anthropology: Social Science in the Service of the Militarized State. CounterPunch/AK Press Petrolia, Oakland.
Richards, P. (Ed.), 2005. No Peace, No War: An Anthropology of Contemporary Armed Conflicts. Ohio University Press, Ohio, OH.
Robben, A.C.G.M. (Ed.), 2010. Iraq at a Distance: What Anthropologists Can Teach Us about the War. University of Pennsylvania Press, Philadelphia, PA.
Rupesinghe, K., Rubio Correa, M., 1994. The Culture of Violence. United Nations University Press, Tokyo.
Simons, A., 1999. War: back to the future. Annual Reviews in Anthropology 28, 73–108.
Sluka, J., 1992. The anthropology of conflict. In: Nordstrom, C., Martin, J. (Eds.), The Paths to Domination, Resistance and Terror. University of California Press, Berkeley, CA, pp. 18–36.
Sluka, J., 2000. Death Squad: The Anthropology of State Terror. University of Pennsylvania Press, Philadelphia, PA.
Vigh, H., 2006. Navigating Terrains of War: Youth and Soldiering in Guinea-Bissau. Berghahn Books, Oxford.
Warren, K.B. (Ed.), 1993. The Violence within: Cultural and Political Opposition in Divided Nations. Westview Press, Boulder, CO.
Whitehead, N.L., Finnström, S. (Eds.), 2013. Virtual War and Magical Death: Technologies and Imaginaries for Terror and Killing. Duke University Press, Durham, NC.

War Crimes Tribunals, Crimes against Humanity and Transitional Justice

Jamie Rowen, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J. Hagen, volume 24, pp. 16357–16360, © 2001, Elsevier Ltd.

Abstract

Ideas related to war crimes tribunals, crimes against humanity and transitional justice have evolved considerably since the codification of international humanitarian law in the nineteenth century. After the Second World War, a number of new crimes, including genocide, crimes against humanity and war crimes, were codified and adjudicated in international tribunals in Nuremburg and Tokyo. International war crimes tribunals reemerged in the 1990s to deal with the violence that occurred in the break-up of the former Yugoslavia and in Rwanda. Currently, there exist an international treaty-based tribunal, two subsidiary bodies of the United Nations, hybrid tribunals that blend domestic and international law, and domestic tribunals that try international crimes. These tribunals are an important component of transitional justice, an idea that has emerged over the past two decades and offers theoretical and practical tools to the challenges that governments face in their efforts to redress past and ongoing violence.

Contemporary Origins

War crimes and crimes against humanity emerged as a category of international law in the latter half of the nineteenth century, when international humanitarian law, or the body of laws regulating war, was first delineated in international treaties. Prior to this time, those who committed wanton acts of violence during war were treated as traitors, usurpers or enemies (Bass, 2001). The International Committee of the Red Cross (ICRC) played a major role in convening world powers to sign the Geneva Convention of 1864, which regulated the protection of the sick and wounded during war. Although debates that led to the Geneva Conventions and other treaties at the time included topics related to crimes related to *jus ad bellum* (the right to go to war), the Conventions focus on *jus in bello* (legal provisions pertaining to conduct during war) (Sloane, 2009). Several other treaties followed the initial Geneva Conventions of 1864, including the Brussels Declaration of 1874, the Geneva Convention of 1906, and the Hague Conventions of 1899 and 1907. In 1929, after the First World War (WWI), world powers convened and passed another round of the Geneva Conventions, which specified the treatment of prisoners of war and the treatment of the sick and wounded. In 1949, after the Second World War (WWII), state representatives updated the three prior Geneva Conventions and added a fourth. In 1977, they passed Protocols I and II, which specify the treatment of victims of both international and non-international armed conflict.

War crimes are generally considered 'grave breaches' of the Geneva Conventions and Protocol I, which governs international armed conflicts. Legal decisions in war crimes tribunals, discussed below, have extended the definition of 'grave breaches' to include internal armed conflicts, which are governed by Common Article 3, an article found in all four Geneva Conventions, as well as by Protocol II. The ICRC describes grave breaches as including willful killing, torture or inhumane treatment (including biological experiments), willfully causing great suffering or serious injury to body or health, and willfully depriving a protected person of the right to a fair trial if accused of a war crime. The Fourth Geneva Convention also restricts the taking of hostages, unnecessary and extensive destruction of property, and unlawful deportation, transfer or confinement (Henckaerts et al., 2005). The idea of crimes against humanity was articulated in 1915 by Woodrow Wilson in reaction to the Armenian genocide, though the idea that certain actions that violate all humanity existed long before (Bassiouni, 2011).

A number of other treaties clarify regulations around the means and methods of warfare, while declarations and other nonbinding statements form the basis of customary law or customs of war. These various sources comprise the legal foundation of war crimes tribunals and influence current understandings and practices related to transitional justice.

The First and Second World Wars

The first issue addressed at the Paris Peace Conference, which ended WWI, was the punishment of war criminals (Bass, 2001). After WWI, Britain and France attempted to try Kaiser Wilhelm I for the violence that ensued after Germany invaded Belgium and Luxembourg. Article 227 of the Treaty of Versailles contained a clause stating that the Allied powers would arraign the Kaiser "for a supreme offense against international morality and the sanctity of treaties," while Article 228 asked for Germans who had "committed acts in violation of the law and customs of war" to be extradited to London. Neither the Netherlands, which was hosting the Kaiser, nor Germany, where the other defendants resided, participated willingly in the process. The compromise – to hold the trials in Germany's Supreme Court – led to lenient sentences and several acquittals. The Kaiser was not extradited and ultimately died in the Netherlands after the start of WWII. While victims of German aggression felt the trials were too lenient, many Germans felt that they were too harsh. Some argue that these provisions for trials fanned the flames of resentment in Germany, fomenting the aggression that led to WWII.

With the international institutions that were created in the wake of WWII, the modern era of international criminal law began. After Germany's defeat, the allied powers came together to design a process to punish Germany for the horrific violence

it perpetrated. The United Nations War Crimes Commission was established in 1942 and began collecting names of individuals suspected of war crimes (Davidson, 1966). The US had warned Germans about charges for war crimes even before it joined the War. After the war ended in 1945, the allied powers drafted the London Agreement. One of the major concerns had to do with whether the defendants could be charged with waging war. France and Russia were especially concerned about this, but the Americans were influential in their position that Germany could and should be tried for crimes against the peace (Bass, 2001). After negotiations about whether all Nazis, or just their leaders, should be tried (or, as some American government officials expressed, should be hanged), 24 men and six organizations were put on trial at Nuremburg.

The International Military Tribunal (IMT) at Nuremburg has become a model of the international war crimes tribunal (Meron, 1994). It was a treaty-based tribunal, outside the mandate of the newly formed United Nations. It tried the 22 senior Nazi officers for war crimes and crimes against humanity. War crimes were defined as "violations of the laws and customs of war," particularly "murder, ill-treatment of deportation to slave labor," "murder or ill-treatment of prisoners of war," as well as "destruction of cities, towns, or villages, or devastation not justified by military necessity," while crimes against humanity were defined as "murder, extermination, enslavement, deportation and other inhumane acts committed against any civilian population, before or during the war," or "persecutions on political, racial or religious grounds in execution of, or in connection with," war crimes and crimes against the peace, regardless of whether they were legal or illegal under domestic law (London Charter, 1945). The tribunal also applied the doctrine of conspiracy, known primarily in the common law system, to heighten charges, but the judges only allowed for a conspiracy charge to be added to the crime of aggression charge (Bass, 2001).

The trials themselves were a remarkable undertaking, yet procedurally deficient in a number of respects (Davidson, 1966). Thirteen trials were held; all but one were presided over by American judges. Americans represented half of the total workforce in the Tribunal (there were over 600). The defense had no access to the documents held by the prosecution. German lawyers had to specify what they wanted, and they only had a month to prepare their pleas. Twelve of the accused were sentenced to death, 10 were eventually hanged (one died committing suicide the night before his scheduled execution and the other, who was tried *in absentia*, died while trying to escape Germany at the end of the war), seven received prison sentences for periods ranging from 10 to 20 years, and three were acquitted.

Later trials were held for lower ranking officers (Zeck, 1947). Trials known as the Subsequent Nuremberg Proceedings involved another 12 high-ranking officers, and the United States Nuremberg Military Tribunals held trials of 177 additional individuals. Under Allied Control Council Law No. 10, German tribunals were also created to try those accused of war crimes. Over 900 defendants were put on trial in these domestic courts, although there are a number of critiques about the lenient sentences and procedural deficiencies in these domestic tribunals (Ferencz, 1998).

The Nuremburg trials led to a number of important innovations in legal doctrine regarding war crimes. In particular, the Charter explicitly stated that individuals could be held liable for actions they were commanded to undertake, paving the way for both subordinates and state leaders to be held individually liable for actions committed during conflict. Although the crime of conspiracy was only allowed as an additional charge in relation to crimes against the peace (now known as the crime of aggression), the jurisprudence on conspiracy provided a foundation for contemporary doctrines of vicarious liability.

While Nuremburg is often seen as victor's justice, it is usually celebrated as a progressive example of international justice. In contrast, the International Military Tribunal for the Far East (IMTFE), commonly known as the Tokyo Tribunal, is often viewed as the example of postwar justice gone wrong (Danner, 2005). Its establishment was not the product of allied negotiations but rather resulted from the unilateral efforts of US General Douglas MacArthur. The tribunal focused on individuals accused of crimes against the peace, but it also allowed for additional charges related to war crimes and crimes against humanity. Whereas Nuremburg did not adjudicate crimes perpetrated before the start of the war, the IMTFE investigated actions all the way back to 1931, and the judges were much more liberal with conspiracy charges (Boister, 2010) The prosecutor was from the United States, and the 11 judges included representatives from the nine nations involved in the Japanese Instrument of Surrender, as well as India and the Philippines.

The defendants included 28 members of the Japanese military and political establishment, but the Emperor was granted immunity. The trial lasted over 2 years and 3 months, and it included 419 witnesses. The tribunal faced a number of dilemmas related to linguistic and cultural differences between the participants, as well as procedural irregularities (Maga, 2001). The judgments included five separate opinions and two dissenting opinions that highlighted the level of internal discontent with the proceedings. The sentences, decided by a simple majority, led to death by hanging of seven defendants. Interestingly, those who survived the sentencing were all released unconditionally by 1958, after American occupation ended in 1952 and domestic pressure mounted to release not only these defendants, but also the thousands of lesser ranked military personnel prosecuted by foreign nations.

International Criminal Tribunal for Rwanda and International Criminal Tribunal for the Former Yugoslavia

Although there was momentum in the post war years to create a permanent tribunal to deal with international crimes, the Cold War stymied efforts to collaborate on such an endeavor. The next two international war crimes tribunals were designed to be temporary, inspired by failure to stop the devastating violence that ensued in the break up of the former Yugoslavia and the genocide in Rwanda. These tribunals ushered in a new era of international criminal law as the United Nations drew on Chapter VII of its Charter, which enables it to intervene in domestic affairs of member nations in order to maintain international peace and security. Pursuant to this authority, the

UN created two subsidiary bodies to try those most responsible for the devastating violence in these regions.

On 25 May 1993, the UN Security Council established the International Criminal Tribunal for the Former Yugoslavia (ICTY), "for the prosecution of persons responsible for serious violations of international humanitarian law committed in the territory of the former Yugoslavia" (UN Resolution 827). UN Security Council Resolution 827 defined the Tribunal's substantive jurisdiction over 'grave breaches' of the Geneva Conventions, violations of the laws or customs of war, genocide, and crimes against humanity. Notably, unlike the Nuremberg Tribunal, the ICTY had no authority to prosecute states for aggression or crimes against peace. Its mandate specified that the Tribunal was expected 'to put an end to such crimes' and, by bringing those responsible to justice, to 'contribute to the restoration and maintenance of peace' (ICTY Statute).

After the massacre at Srebrenica, later declared by the International Court of Justice to be an act of genocide, the international community intervened to stop the bloodshed and broke the Dayton Peace Accords in 1995. Two years later, the ICTY conducted its first trial of a mid ranking Serb officer. Given the nature of the violence and the leadership structure within the armed forces, the ICTY statute adopted an expansive understanding of individual criminal responsibility to include two types of vicarious liability, meaning liability for crimes that the accused did not personally commit (Danner and Martinez, 2005). Under the theory of command responsibility, a superior officer is individually criminally responsible for the acts of his subordinates when he knew or should have known that his subordinates were going to commit a war crime and failed to prevent such an occurrence. Moreover, a superior officer who learns that a subordinate has committed a war crime and fails to punish the actor is likewise criminally responsible for the offense. The doctrine of liability known as joint criminal enterprise (JCE) enables liability for the "participation in the realization of a common design or purpose." This theory of liability was the foundation of many convictions at the ICTY, and remains a controversial doctrine for its expansive notion of individual responsibility.

The ICTY also tried those accused of crimes committed in Kosovo in 1999. Slobodan Milosevic, widely considered responsible for igniting the violence that ensued in the break up of the former Yugoslavia, was ultimately indicted for his role in the violence in Kosovo, but he died before the conclusion of his trial (Hagan, 2003).

While remarkable for its contributions to international criminal law, there are a number of critiques of the ICTY. Many suggest that the tribunal was too expensive, costing US$2.2 billion over 20 years. Moreover, among the thousands responsible for committing atrocities, only161 individuals have been indicted and only 141 cases have been completed. There have been 18 acquittals, and 36 indictments withdrawn, including six defendants who died before their trials concluded (ICTY Facts and Figures). Although atrocities occurred to all ethnic groups and in all countries, a vast majority occurred in Bosnia and Herzegovina, where Bosnian Serbs waged a campaign of violence against Bosnian Muslims. Various studies point out that opinions of the tribunal reflect the ethnic divides in the region, with Serbs feeling unfairly targeted by the tribunal and Bosnian Muslims feeling that plea bargaining, reduced sentences and acquittals have revictimized them (Delpla, 2007). Given the Tribunal's location in The Hague, there are ongoing questions about its impact on the restoration and maintenance of peace (Nettelfield, 2010).

The International Criminal Tribunal for Rwanda (ICTR) is similar to the ICTY in a number of ways, and it has faced similar challenges related to its procedural and substantive limitations (Møse, 2005). The tribunal had a narrow mandate to "prosecute persons responsible for serious violations of international humanitarian law committed in the territory of Rwanda … between 1 January 1994 and 31 December 1994" (UN Resolution 1503). The Tribunal was established to address the extremist Hutu-led slaying of approximately 800 000 ethnic Tutsis and thousands of politically moderate Hutus. After the violence, the Tutsi-led Rwanda Patriotic Front (RPF) became the ruling political party and asked the United Nations to create the tribunal. Later, the government expressed concern about a number of key provisions, particularly the ICTR's proposed structure, jurisdiction, and location in Arusha, Tanzania, which is far from the site of the violence and thus far from public view (Waldorf, 2006). Rwanda ultimately voted against the resolution that created the tribunal.

In its November 2013 report, the Tribunal declared that it had concluded the substantive work in the trials against 93 individuals. There were 55 first-instance judgments involving 75 accused, 10 referrals to national jurisdictions, three cases that were transferred to the International Residual Mechanism for Criminal Tribunals, two withdrawn indictments, and three defendants who died prior to or in the course of trial. All of the appeals are expected to be concluded by 2015 (ICTR, 2013).

A noticeable difference between the ICTR and the ICTY relates to their respective definitions of international crimes, particularly crimes against humanity (Cryer, 2005). The Rwanda conflict was noninternational in character, so Common Article 3 of the 1949 Geneva Conventions, as well as Protocol II, were applicable. The provisions on genocide in both statutes are the same as the Genocide Convention. Unlike the ICTY statute, the Rwanda Statute does not require that crimes against humanity be connected to armed conflict. Moreover, the ICTR Statute expressly requires that the enumerated inhumane acts be committed against a civilian population "on national, political, ethnic, racial or religious grounds." Unlike other tribunals, which apply the idea of a "serious breach of the Geneva Conventions" to armed conflict, the ICTR statute applied 'serious breach' to Common Article 3 of Geneva Conventions, which applies to noninternational armed conflicts. It also expanded the jurisprudence so that civilians can be held liable for war crimes, and it tied the notion of crimes against humanity to noninternational conflicts (Kamatali, 2005).

The Tribunal is perhaps most notable for its jurisprudence on genocide and sexual violence. It became the first international tribunal to find a defendant guilty of genocide; it also concluded that rape can constitute a crime against humanity and an act of genocide. Moreover, it delineated the difference between sexual assault and rape, concluding that rape need not include intercourse (Amann, 1999). The tribunal also held that media leaders could be held individually accountable if they publish propaganda that incites genocide (Gordon, 2004).

Though generally less controversial than the ICTY, ICTR has been strongly criticized for its inability to adequately address

the causes and the consequences of the genocide (Kamatali, 2005). First, it did not try the Rwandan Patriotic Front (RPF), the current ruling political party, for the group's alleged atrocities. Second, there are a number of concerns about the tribunal's limited temporal mandate to only try crimes committed in 1994. This limited time frame obfuscates the conflict's origins and its ongoing effects in the region, particularly Rwanda's role in the violence that continues to destabilize the Democratic Republic of Congo. Moreover, the trials have been held far away from Rwanda, and thus had a limited impact on how Rwandese population viewed the violence.

The International Residual Mechanism for Criminal Tribunals, or the Residual Mechanism, is the judicial body authorized to take over cases from the ICTY and the ICTR. The Mechanism began its work for ICTR cases in 2012 and for ICTY cases in 2013. This mechanism is supposed to try senior leaders responsible for the violence, handle appeals and charges of contempt, and retrials. But it has a host of other responsibilities too: it is supposed to help enforce sentences, assist national authorities in trying accused parties who are not senior leaders, help with the protection of witnesses in completed cases, aid the ongoing efforts to apprehend fugitives, and manage the Tribunal archives (United Nations Mechanism).

Hybrid Tribunals

By the early 2000s, the ad hoc tribunals were costing the UN nearly 10% of its overall budget, and efforts to domesticate international criminal law came to the fore (Dougherty, 2004). In addition to the ad hoc tribunals, the United Nations has created two hybrid tribunals, the Special Court for Sierra Leone (SCSL) and the Special Tribunal for Lebanon (STL), that blend international and domestic criminal law. While the STL covers transnational terrorism and not war crimes, the creation and operation of the SCSL represents a unique effort to domesticate international war crimes tribunals through the blending of domestic and international practices.

The war in Sierra Leone began in 1991 when the Revolutionary United Front (RUF), supported by Charles Taylor in neighboring Liberia, attempted to overthrow the ineffectual and corrupt government. By the time the conflict officially ended in 2002, an estimated 60 000 people had been killed, nearly 2 million displaced, and 20 000 abducted; nearly 7000 child soldiers had been disarmed. Like Rwanda, the government of Sierra Leone wrote to the UN Secretary General requesting the international body to try rebel leaders for crimes against the people of Sierra Leone and for taking United Nations peacekeepers as hostages (Fritz and Smith, 2001). For 2 years before the conflict officially ended, the UN worked with the government of Sierra Leone to create a tribunal in order to address allegations of horrific acts of violence. The Court had jurisdiction over crimes against humanity, war crimes, defined as violations of Common Article 3 of the Geneva Conventions and Additional Protocol II, which pertain to internal armed conflict, as well as other serious violations of international humanitarian law.

The SCSL is distinct from the ad hoc tribunals in a number of important ways (Dougherty, 2004). The Tribunal incorporated Sierra Leonean laws that prohibit sexual violence against girls and the wanton destruction of property, specifically arson, because the rebels frequently burned buildings. It was located in Sierra Leone, where the violence took place, which addressed concerns that the ad hoc tribunals were too far from the victims. However, unlike the ad hoc tribunals, the SCSL required voluntary donations. Even after the Secretary General halved his initial request for a 3-year operation budget, there was a US$21 million shortfall from the original requests. With first year contributions amounting to a little over 10% of the ICTY's annual budget, the SCSL was severely limited in its operations; it had to rely on donations for the third year in order to continue operating through the second year.

The tribunal has been criticized for its procedural and substantive limitations, particularly the lack of due process afforded the defense through equality of arms and evidentiary standards (Combs, 2010). The tribunal also did not have United Nations Chapter VII powers, meaning that it could not compel state authorities to hand over indicted individuals (Fritz and Smith, 2001). When the SCSL indicted Liberian President Charles Taylor, it struggled to get cooperation from Ghana, where Taylor was attending peace talks to end Liberia's Civil War. Taylor went into exile in Nigeria but was eventually handed over to The Hague in 2006. Although he boycotted the trial, he was eventually found guilty on all counts and sentenced to 50 years in prison.

There are other ongoing efforts to try war crimes and crimes against humanity at a domestic level. Among these efforts, the War Crimes Chamber of Bosnia-Herzegovina is notable for being an early attempt to domesticate international criminal law (Garbett, 2010). In July 2002, the Security Council put forth a strategy to encourage domestic jurisdictions to try lower level perpetrators; it secured €15.7 million for the endeavor. Following these developments, the War Crimes Chamber in Bosnia and Herzegovina was created in 2005. The Chamber is part of the State Court of Bosnia-Herzegovina's Criminal Division. It hears cases referred by the ICTY, but can also bring its own cases under domestic law. In July 2013, the European Court of Human Rights found that two defendants experienced a violation of Article 7 of the European Convention on Human Rights, which prohibits 'punishment without law,' because the Chamber retroactively applied its 2003 Criminal Code for violations that occurred in 1992–93. The Court specifically noted that, because the men were found guilty of war crimes (not only of crimes against humanity), and because they were not found criminally liable for any loss of life, they were entitled to greater procedural and substantive safeguards than the Chamber had provided (Zuvela, 2013).

There have been several other recent attempts to use domestic courts to try leaders accused of international crimes (Roht-Arriaza, 2013). In particular, Guatemala's penal code from 1973 includes a provision on genocide and 'crimes against obligations to humanity,' which refer to the country's obligations under the 1949 Geneva Conventions and were used to indict the country's former president, Efrain Rios Montt, and his defense minister. In May 2013, Montt was found guilty of genocide and war crimes for his leadership during the military campaign that decimated the Ixil Mayan population during an internal armed conflict against Communist rebel groups. The trial was controversial within Guatemalan society; many of the ruling elite supported Montt,

and the current president was implicated in the violence. The Constitutional Court annulled the sentence 10 days after it was handed down, ordered that the trial restart, and later held that, due to a technicality about when her 4-year term officially began, the Attorney General who pursued the case must step down 7 months before her formal appointment ends.

In another problematic example of domestic trials for international crimes, opposition leaders in Bangladesh came to power on a platform that included the prosecution of alleged war crimes during the violence that ensued in the 1971 partition (Luban, 2013). Several of the accused were high-ranking officials in the previous ruling party. The tribunal relied on the 1973 International Crimes Act, which allowed for domestic prosecutions of genocide, crimes against humanity and war crimes. The Tribunal has been heavily criticized for its lack of procedural safeguards, as well as harsh sentences, including the death penalty, in a highly politicized atmosphere.

International Criminal Court

In 1948, the UN General Assembly adopted the Convention on the Prevention and Punishment of the Crime of Genocide, calling for criminals to be tried "by such international penal tribunals as may have jurisdiction." The UN invited the International Law Commission (ILC) "to study the desirability and possibility of establishing an international judicial organ for the trials of persons charged with genocide." In 1950, the ILC, comprising 34 individuals chosen 'as persons of recognized competence in international law' from 'the principal legal systems of the world,' was tasked with creating a draft code of offenses against the peace and security of mankind (Clarke, 2009).

However, after the ILC drafted its statute in 1951, it was difficult to bridge the divides that arose during the Cold War (Schiff, 2008). While formal efforts to create a permanent international criminal tribunal were stalled, the idea lingered. In the 1970s, advocacy groups in the United States and elsewhere began to increase their activities in international affairs. In 1981, the General Assembly asked the ILC to restart its work on a draft code of international crimes, culminating in the Convention Against Torture in 1985. In the late 1980s, the chance to create an international tribunal reemerged. The Prime Minister of Trinidad and Tobago, also the director of a non-governmental organization (NGO) called the Foundation for the Establishment of the International Criminal Court, began to lobby for international coordination to combat drug trafficking. In June 1989, the Minister proposed "the establishment of an international criminal court with jurisdiction to prosecute and punish individuals and entities who engage in, inter alia, the illicit trafficking in narcotic drugs across national borders."

One of the main concerns about a permanent tribunal had to do with the international crimes over which it would have jurisdiction, as well as its relationship with existing international institutions (Roach, 2006). When representatives from the United Nations investigated alleged war crimes in Yugoslavia, they recognized the need for a permanent tribunal to focus on genocide, crimes against humanity and war crimes. In 1993, the International Law Commission offered a draft statute for the ICC. The General Assembly sent it back to the ILC for more discussions, including how to define crimes against humanity in light of the variety of definitions in earlier tribunals (Bassiouni, 2011). Over the next 5 years, coalitions of advocacy groups lobbied heavily for the creation of a permanent tribunal. On 17 July 1998, representatives from a majority of states signed the Rome Statute, the treaty that lays out the ICC's rules and procedures.

The ICC has subject matter jurisdiction over what are now known as the core crimes of international criminal law: war crimes, crimes against humanity, genocide, and the crime of aggression. The statute defines war crimes as 'grave breaches' of the Geneva Conventions, and 'serious violations' of the laws and customs of international armed conflict. It also offers a more limited list of offenses for armed conflicts not of an international nature. It drew on the ICTR definition of crimes against humanity; the acts need not be linked to an international armed conflict, but the actions must be part of state policy in order to qualify (Bassiouni, 2011). Interestingly, the crime of aggression is not enforceable until at least 2017, and not until 30 states ratify amendments that clarify the crime's definition as well as the procedure for referring it to the Court (Trahan, 2011). These amendments include a provision that enables signatories to the Rome Statute to opt out of jurisdiction for acts of aggression.

The United States, along with China, Iraq, Israel, Libya, Qatar and Yemen, voted against the Rome Statute. The US later signed on, but then, under the Bush administration, 'unsigned.' The US government's concerns are related to the crime of aggression, as well as the prosecutor's perceived independence and ability to make decisions that could run counter to US interests (Scheffer, 1999). The ICC prosecutor has the ability to open up cases against countries at her or his discretion so long as the alleged crime was committed on the territory of a state that has ratified the Rome Statute (also known as a State Party), or by a national of a State Party, though he or she cannot bring a case against someone in the custody of a State Party. If the alleged crime does not fall into these categories, an ICC investigation may be commenced by the Security Council, pursuant to Chapter VII of the UN Charter.

There are other jurisdictional hurdles, though. For example, the Court follows the principle of complementarity, meaning that it can only act if a state with jurisdiction is not able or willing to carry out its own proceedings (Clarke, 2009). Moreover, if the Prosecutor initiates an investigation, he or she must ask the Pre-Trial Chamber to approve the investigation. Following approval, victims must be informed and invited to contribute to the investigation. Lastly, the ICC must notify the state with jurisdiction over the alleged crimes, and the state can request a deferral. Investigations or prosecutions may be deferred for a period of 1 year, renewable, at the Security Council's discretion.

A striking innovation by the ICC has to do with its treatment of victims, a group that other tribunals overlooked (Stahn et al., 2006; Stover, 2007). For instance, the Victims' Participation and Reparations Section helps victims to apply for participation in proceedings or reparations, and it provides legal advice where needed. Two other agencies also provide support to victims – The Victims and Witnesses Unit and the Office of Public Counsel for Victims. The Trust Fund for Victims, established separately from the ICC, requires separate contributions but does similar work as the various parts of the ICC.

Soon after the ICC came into existence, the Congo and Uganda referred cases to the Court in order to address their internal armed conflicts. The first trial, in which Congolese rebel leader Lubanga was charged with the recruitment of child soldiers, ended in an unconditional release by the Trial Chamber. Although this judgment was later overturned by the Appeals Chamber and Lubanga was ultimately convicted and sentenced to 14 years in prison, the proceedings highlighted a number of concerns about the Court's capacity to fulfill its mandate. Among other dilemmas facing the prosecution, the first witness who took the stand had previously claimed to be a child soldier under Lubanga's command, but recanted this claim in court (Sarvarian, 2013).

The ICC continues to face criticism for the politically sensitive decisions it must make. The first prosecutor of the ICC made a number of controversial decisions, such as indicting the sitting president of Sudan for genocide in Darfur and not pursuing an investigation in Gaza. All of the 21 cases thus far have focused on defendants from Africa, but to be fair, four of these cases were referred by State parties and three were referred by the Security Council. Furthermore, the Court has recently been pursing investigations in other countries, notably in Colombia and Cambodia.

Transitional Justice

Transitional justice was originally conceived as a way to describe how 'emerging democracies' reckon with former regimes, with a focus on the various kinds of justice that countries may seek (Kritz, 1995; Teitel, 2000). The idea captures both (1) a jurisprudential theory about justice in times of political transition and (2) an associated set of approaches/tools/mechanisms to help ensure accountability, improve survivor well being, and prevent future violence (Gray, 2005). War crimes tribunals are considered one of the many mechanisms of transitional justice; the others include truth commissions, reparations programs, memorials, lustration and democratic institutional reforms.

Though originally discussed in relation to countries undergoing regime change, the idea of transitional justice is now applied to settler regimes, countries in ongoing conflict, and countries with no clear democratic transition (LaPlante and Theidon, 2006; Winter, 2013). Some consider transitional justice as a field, others refer to transitional justice as a discipline within the well-established field of human rights, the difference being that transitional justice focuses on political expediency while human rights focuses on legal accountability (Arthur, 2009; Bell, 2009). Early critics suggest that there is nothing out of the 'ordinary' about transitional justice, arguing that law is never devoid of politics and always reflects trade-offs of expediency (Posner and Vermeule, 2004). However, the idea is now taken for granted when discussing how to ensure justice for past, as well as ongoing, atrocities.

In particular, transitional justice was articulated in relation to concerns that holding trials for former leaders would destabilize newly formed governments (Brito et al., 2001). For example, after experimenting with trials, some of which were criticized for their lenient sentences, Argentina and Chile passed various laws that ensured amnesty for military leaders. The countries also created highly publicized truth commissions. While arguably necessary to ensure political stability at the time, various amnesty laws have been challenged, and even overturned, in recent years and military personnel have been put on trial (McEvoy and Mallinder, 2012).

At the close of the twentieth century, with international attention focused on the end of apartheid and the South Africa Truth and Reconciliation Commission (TRC), the idea of transitional justice became increasingly prominent in scholarly, advocacy, and policy making circles (Teitel, 2003). For years, the various parties debated the trade-offs between achieving truth, or an accurate historical record, and justice, usually referring to legal accountability, in countries undergoing regime change (Zalaquett, 1994). Part of these debates has to do with misconceptions about the South Africa experiences. The TRC did not offer blanket amnesty in exchange for confessions. Rather, the Commission consisted of three panels, one that was responsible for approving amnesty applications based on strict criteria about the political nature of the perpetrator's actions, another that held public events and produced a report on the violence, and another that made recommendations for reparations. While lauded for its contribution to maintaining political stability in South Africa, the Commission has been heavily critiqued for misrepresenting the structural nature of the violence and for the fact that the government has not yet provided full reparations (Mamdani, 2002; Wilson, 2001).

These earlier debates about truth versus justice have largely subsided as a consensus has emerged that both goals are part of the holistic conception of justice that is necessary to redress violence and prevent its recurrence (Roht-Arriaza and Mariezcurrena, 2006). The scholarly literature on transitional justice increasingly focuses on the efficacy of war crimes tribunals, truth commissions, and other approaches (Olsen et al., 2010). In addition, scholars, policy makers, and advocates are paying increased attention to local practices in order to inform their efforts to ensure justice (Shaw et al., 2010). From the community truth-telling panels in East Timor to the cleansing rituals in Uganda, local justice practices are being incorporated into recommendations for redressing international crimes (Roht-Arriaza and Mariezcurrena, 2006).

Of course, these local justice approaches are not without critique; just as there are questions as to whether approaches from one country can be transplanted to another, there remain a number of concerns about whether or to what extent local practices, which are designed for smaller scale disputes, can readily be applied to the kinds of violence that trigger the jurisdiction of international criminal law (Hinton, 2010). The idea of 'local justice' in these cases may actually undermine due process by not providing procedural safeguards for defendants and encouraging potentially false witness statements (Waldorf, 2006). Scholars, policy makers and advocates continue to search for optimal combinations of formal justice mechanisms, like tribunals, and community-based practices that can ensure justice for mass violence and prevent its recurrence.

See also: Human Rights: Legal Aspects; International Law and Treaties; Nazi Law; Torture: Legal Perspectives.

Bibliography

Amann, D.M., 1999. Prosecutor v. Akayesu. Case ICTR-96-4-T. American Journal of International Law 93, 195–199.

Arthur, P., 2009. How "transitions" reshaped human rights: a conceptual history of transitional justice. Human Rights Quarterly 31 (2), 321–367.

Bass, G.J., 2001. Stay the Hand of Vengeance: The Politics of War Crimes Tribunals. Princeton University Press, Princeton.

Bassiouni, M.C., 2011. Crimes against Humanity: Historical Evolution and Contemporary Application. Cambridge University Press.

Bell, C., 2009. Transitional justice, interdisciplinarity and the state of the "field" or "non-field". The International Journal of Transitional Justice 3, 5–27.

Boister, N., 2010. The application of collective and comprehensive criminal responsibility for aggression at the tokyo international military tribunal the measure of the crime of aggression? Journal of International Criminal Justice 8 (2), 425–447.

Brito, A.B.D., Enriquez, C.G., Aguilar, P., 2001. The Politics of Memory: Transitional Justice in Democratizing Societies. Oxford University Press, New York.

Clarke, K.M., 2009. Fictions of Justice: The International Criminal Court and the Challenge of Legal Pluralism in Sub-Saharan Africa. Cambridge University Press, New York.

Combs, N.A., 2010. Fact-Finding without Facts: The Uncertain Evidentiary Foundations of International Criminal Convictions. Cambridge University Press, New York.

Cryer, R., 2005. Prosecuting International Crimes: Selectivity and the International Criminal Law Regime. Cambridge University Press, New York.

Danner, A.M., 2005. Beyond the Geneva conventions: lessons from the Tokyo tribunal in prosecuting war and terrorism. Virginia Journal of Internaitonal Law 46, 83–126.

Danner, A.M., Martinez, J.S., 2005. Guilty associations: joint criminal enterprise, command responsibility, and the development of international criminal law. California Law Review 93 (1), 75–169.

Davidson, E., 1966. The Trial of the Germans: An Account of the Twenty-Two Defendants before the International Military Tribunal at Nuremberg. University of Missouri Press, MI.

Delpla, I., 2007. In the midst of injustice: the ICTY from the perspective of some victim associations. In: The New Bosnian Mosaic. Identities, Memories and Moral Claims in Post-War Society. Ashgate Publishing Company, England, pp. 211–234.

Dougherty, B.K., 2004. Right-sizing international criminal justice: the hybrid experiment at the special court for Sierra Leone. International Affairs 80 (2), 311–328.

Ferencz, B.B., 1998. International criminal courts: the legacy of Nuremberg. Pace International Law Review 10, 203–235.

Fritz, N., Smith, A., 2001. Current apathy for coming anarchy: building the special court for Sierra Leone. Fordham International Law Journal 25, 391–430.

Garbett, C., 2010. Localising criminal justice: an overview of national prosecutions at the war crimes chamber of the court of Bosnia and Herzegovina. Human Rights Law Review 10, 558–568.

Gordon, G.S., 2004. War of media, words, newspapers, and radio stations: the ICTR media trial verdict and a new chapter in the international law of hate speech. Virginia Journal of International Law 45, 139–199.

Gray, D., 2005. An excuse-centered approach to transitional justice. Fordham Law Review 74, 2621–2695.

Hagan, J., 2003. Justice in the Balkans: Prosecuting War Crimes in the Hague Tribunal. University of Chicago Press, Chicago.

Henckaerts, J.-M., Doswald-Beck, L., Alvermann, C., 2005. Customary International Humanitarian Law: Rules. Cambridge University Press, New York.

Hinton, A.L., 2010. Transitional Justice: Global Mechanisms and Local Realities after Genocide and Mass Violence. Rutgers University Press, Piscataway, NJ.

ICTR, 2013. UN Security Council, Letter dated 13 November 2013 from the President of the International Criminal Tribunal for Rwanda addressed to the President of the Security Council S/2013/663(2013).

ICTY, Key Facts and Figures. http://www.icty.org/x/file/Cases/keyfigures/key_figures_en.pdf.

ICTY Statute, United Nations, International Tribunal for the Prosecution of Persons Responsible for Serious Violations of International Humanitarian Law Committed in the Territory of the Former Yugoslavia since 1991. http://www.icty.org/x/file/Legal%20Library/Statute/statute_sept09_en.pdf.

Kamatali, J.M., 2005. From the ICTR to ICC: learning from the ICTR experience in bringing justice to Rwandans. New England Journal of International and Comparative Law 12, 89–104.

Kritz, N.J., 1995. Transitional Justice?: How Emerging Democracies Reckon with Former Regimes. United States Institute of Peace Press, Washington, DC.

LaPlante, L., Theidon, K., 2006. Transitional justice in times of conflict: Colombia's Ley de Justicia y Paz. Michigan Journal of International Law 28, 49–107.

London Charter, August 8, 1945. Agreement for the Prosecution and Punishment of the Major War Criminals of the European Axis, Charter of the International Military Tribunal. Art. 3, 59 Stat. 1544, 82 UNTS 279, reprinted in 39 AJIL 257(1945).

Luban, D., 2013. After the honeymoon reflections on the current state of international criminal justice. Journal of International Criminal Justice 11 (3), 505–515.

Maga, T.P., 2001. Judgment at Tokyo: The Japanese War Crimes Trials. University Press of Kentucky, Kentucky.

Mamdani, M., 2002. Amnesty or impunity? A preliminary critique of the report of the truth and reconciliation commission of South Africa (TRC). Diacritics 32 (3), 33–59.

McEvoy, K., Mallinder, L., 2012. Amnesties in transition: punishment, restoration, and the governance of mercy. Journal of Law and Society 39 (3), 410–440.

Meron, T., 1994. War crimes in Yugoslavia and the development of international law. The American Journal of International Law 88 (1), 78–87.

Møse, E., 2005. Main achievements of the ICTR. Journal of International Criminal Justice 3 (4), 920–943.

Nettelfield, L., 2010. Courting Democracy in Bosnia and Herzegovina: The Hague Tribunal's Impact in a Postwar State. Cambridge University Press, Cambridge, MA.

Olsen, T.D., Payne, L.A., Reiter, A.G., 2010. Transitional Justice in Balance: Comparing Processes, Weighing Efficacy. United States Institute of Peace Press, Washington.

Posner, E.A., Vermeule, A., 2004. Transitional justice as ordinary justice. Harvard Law Review 117 (3), 761–825.

Roach, S.C., 2006. Politicizing the International Criminal Court: The Convergence of Politics, Ethics, and Law. Rowman & Littlefield, Lanham.

Roht-Arriaza, N., 2013. Just a "bubble"? Perspectives on the enforcement of international criminal law by national courts. Journal of International Criminal Justice 11 (3), 537–543.

Roht-Arriaza, N., Mariezcurrena, J., 2006. Transitional Justice in the Twenty-First Century?: Beyond Truth versus Justice. Cambridge University Press, New York.

Sarvarian, A., 2013. Professional Ethics at the International Bar. Oxford University Press, New York.

Scheffer, D.J., 1999. The United States and the international criminal court. American Journal of International Law, 12–22.

Schiff, B., 2008. Building the International Criminal Court. Cambridge University Press, New York.

Shaw, R., Waldorf, L., Hazan, P. (Eds.), 2010. Localizing Transitional Justice: Interventions and Priorities after Mass Violence. Stanford University Press, CA.

Sloane, R.D., 2009. Cost of conflation: preserving the dualism of Jus ad Bellum and Jus in Bello in the contemporary law of war. Yale Journal International Law 34, 47–101.

Stahn, C., Olásolo, H., Gibson, K., 2006. Participation of victims in pre-trial proceedings of the ICC. Journal of International Criminal Justice 4 (2), 219–238.

Stover, E., 2007. The Witnesses: War Crimes and the Promise of Justice in the Hague. University of Pennsylvania Press, Philadelphia, PA.

Teitel, R.G., 2000. Transitional Justice. Oxford University Press, New York.

Teitel, R.G., 2003. Transitional justice genealogy. Harvard Human Rights Journal 16, 69–94.

Trahan, J., 2011. The Rome statute's amendment on the crime of aggression: negotiations at the Kampala review conference. International Criminal Law Review 11 (1), 49–104.

United Nations Mechanism for International Tribunals. http://unmict.org/about.html.

UN Security Council, 1993. Resolution 827, 25 May 1993, S/RES/827 (1993). http://www.refworld.org/docid/3b00f21b1c.html.

UN Security Council, 2003. Resolution 1503, 28 August 2003, S/RES/1503 (2003). http://www.refworld.org/docid/3f535ca64.html.

Waldorf, L., 2006. Mass justice for mass atrocity: rethinking local justice as transitional justice. Temple Law Review 79, 1–87.

Wilson, R., 2001. The Politics of Truth and Reconciliation in South Africa. Cambridge University Press, Cambridge, NY.

Winter, S., 2013. Towards a unified theory of transitional justice. International Journal of Transitional Justice 7 (2), 224–244.

Zalaquett, J., 1994. Balancing ethical imperatives and political constraints: the dilemmas of new democracies facing past human rights violations. Hastings Law Journal 43, 6–16.

Zeck, W.A., 1947. Nuremberg: proceedings subsequent to Goering Et Al. North Carolina Law Review 26, 350–389.

Zuvela, M., 19 November 2013. Bosnian War Crimes Court Frees 10 Convicts, Orders Retrials. Reuters, Sarajevo. Retrieved from: http://www.reuters.com/article/2013/11/19/us-bosnia-warcrimes-idUSBRE9AI0PW20131119.

War, Political Violence, and Effects on Children

Christine E Merrilees, State University of New York, Geneseo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by E. Cairns, volume 24, pp. 16360–16363, © 2001, Elsevier Ltd.

Abstract

The changing nature of war and political violence has increased the likelihood that children will be impacted either directly or indirectly by armed conflict. The study of children in contexts of war and political violence is transdisciplinary, with our knowledge about the effects of these events coming out of psychology, sociology, political science, public health, and anthropology, among other disciplines. Exposure to war and political violence is related to a range of responses including negative, neutral, and even positive adaptation. This article will overview the current knowledge regarding the complexity of children's responses to war and political violence.

Overview

The effects of war and political violence on children are complex and require transdisciplinary study of the multiple overlapping factors that play roles in development of children and adolescents. The idea that children are uniquely affected by armed conflict was brought to light through a study conducted by Graça Machel in 1996. The goal of this study was to document the number of ways that children are uniquely affected by and even targeted in situations of political violence and war. The report notes that in the decade leading up to 1996, 2 million youth had been killed, roughly 6 million had been injured, and countless others had been witness to or forced to participate in the violence associated with armed conflict. The report documents the myriad ways that armed conflict affects children. The most obvious loss is the thousands of children who are killed by bullets, land mines, and bombs among other weapons. Millions more experience the disruption of basic services and the destruction of family and community support systems, leading to physical and psychological trauma and even death. Even for children who do not experience the direct forms of violence brought on by armed conflict, the destruction of community relations, and increased distrust and insecurity that result, have lasting psychological consequences. The study report concludes with a long-term call to action that focuses on the societal and emotional roots of conflict as well as attempts to decrease arms delivered to conflict-ridden areas.

A follow-up study published in 2009 highlights the complexity of issues for children brought on by the nature of war and political violence across the globe. Although the follow-up report acknowledges advances made over the 10-year period, it also highlights the changing nature of war and how these changes increasingly put children in harm's way. "Over the past decade, conflict has impacted children more brutally than ever. They are victims of strikes against schools and of abductions aimed at forcing them to serve as combatants, sex slaves or servants. And in conflict zones their vulnerability is often greatly increased because violence claims their first line of defense – their parents" (UNICEF, 2007).

The study of children in these contexts has increased in number and scope, with an increasing understanding that war and political violence must be understood from multiple perspectives and with awareness that children are not just passive victims, but that their voices and active construction of their experiences must be considered in developing intervention efforts that include both goals of physical and mental well-being. The following article will review the changing nature of war, the myriad social, environmental, and psychological processes that are altered in contexts of war, and the documented effects of war and political violence on children. The section will conclude with a discussion of a move toward including children and adolescents' voices and perspectives in planning and implementing interventions and postconflict agreements.

The Changing Nature of War

The current article incorporates effects of a range of conflicts defined by different terms including political violence, war, ethnic conflict, intrastate conflict, armed conflict, sectarian violence and conflict, among others. Researchers have noted the decrease in interstate wars occurring globally with an increase in intrastate wars involving paramilitary groups and other nonstate actors since the middle of the twentieth century (UNICEF, 2009). The trend away from interstate conflicts involving national armies, and the emergence of smaller intrastate conflicts has increased the likelihood of the deliberate targeting of civilians, use of terrorism, and participation of civilian combatants. The results of these changes have several negative implications for children in particular. Bringing conflict off the battlefield and into communities results in the destruction of basic services, the tearing of the social fabric of communities, and most directly brings the violence associated with war and political violence into the daily lives of youth. As noted in the quote above, the changing nature of war has brought the violence and conflict associated with war and political violence closer to home for children influencing the environment that most directly impacts their well-being, the family system.

Context of War and Political Violence

Despite this recognition that children are uniquely affected by war and political violence, the research documenting these relations was limited for several decades. In the same year as the

original Machel study, Cairns and Dawes called attention to the need for psychological researchers to increase the scope and quality of research being conducted on children in contexts of political violence and war (Cairns and Dawes, 1996). Most of the research conducted to this time had a limited scope with the dominant outcomes of studies in these contexts being post-traumatic stress disorder (PTSD). Cairns and Dawes (1996) also note that research to that point supported the notion that a majority of children exposed to political violence and war did not develop clinical levels of symptomatology. Although results across conflicts suggest there is an increased risk for the development of emotional and behavioral problems in these contexts, positive outcomes are generally seen in most children, barring repeated exposure to severe events without emotional support. These results suggesting most children do not present with clinical problems encouraged researchers to consider resilience processes that may be at work. Research on children and political violence to that point suggested that research has focused too much on individual stress and coping processes focused on negative child outcomes. Cairns and Dawes also discussed the lack of longitudinal research conducted to date, limiting conclusions about the long-term effects of political violence and war as well as limiting our understanding of developmental processes at play for children in these contexts.

Since that call, the studies of children in these contexts has increased in number and scope. Given the noted limitations of the research on children and political violence at that point, Cairns and Dawes made recommendations for future research in this area, including the use of qualitative approaches that help researchers fully grasp the nature of violence as it is experienced and coconstructed by both children and adults, longitudinal research, and research conducted within developmental theory. They also noted the need to conduct research in places where peace agreements are established to better understand the role of systems of trust and reconciliation in establishing the intergroup relations that will surely affect postagreement stability.

This work along with the results showing a variety of both positive and negative responses to political violence motivated the development of theoretical models that help answer questions of developmental process, risk and protective factors, and a focus on resilience processes that can help explain why children in these contexts do not develop the expected high levels of psychopathology. New questions about what constitutes general well-being beyond the absence of psychopathology have also begun to emerge (Barber, 2013). Moreover, researchers are considering questions about cross-cultural differences in subjective well-being to better measure and document the ways that political violence and war affect children. In other words, the somewhat surprising low levels of psychopathology in war-affected communities may be related to the outcomes measured. Were researchers to identify and accurately measure culturally informed processes of well-being, we could gain a clearer picture of how political violence and war impact the lives of children.

War and political violence affect multiple elements of society, subsequently affecting multiple environments in which children develop. War and political violence disrupt basic services such as access to food and water, schools, religious institutions, health care facilities, and they can even cut children off from family members. It is because of these ubiquitous reverberating effects of war and political violence that researchers often adopt a social ecological model to understand the effects of war and political violence on children. The social ecological model describes the multiple overlapping contexts in which an individual lives and interacts (Bronfenbrenner, 1979). These levels include systems with which the developing individual directly interacts, such as the family system, peers, neighborhood systems, and schools. This model also recognizes the interplay between systems that might not directly influence the child, but that influence the systems in which the child is developing. For example, interactions between the school system and parent groups may impact the child through changes to school policy. The social ecological model also includes recognition of cultural systems and historical processes that shape the environment and interpretation of experiences. Finally, this framework recognizes the importance of change over time across all contexts.

Identifying the relevant factors at each of these levels of the social ecology is necessary for a thorough understanding of the effects of war and political violence on child development and well-being. For example, understanding the impact of exposure to community-level violence between conflicting nonstate actors requires understanding of the sociohistorical relations between the groups involved in conflict. The meaning behind the acts of violence between these groups will depend in part on timing, duration, and intensity of the conflict. For example, although there are similarities between intergroup conflicts, there exist critically important histories of relations between Palestinians and Israelis in the Middle East, or between Catholics and Protestants in Northern Ireland, or between Tutsis and Hutus in Rwanda, that must be understood and acknowledged to fully grasp the meaning of violence in these conflicts.

Identifying the full range of the effects of war and political violence on children requires an understanding of how both the child and their environment change over time. The development of mental and behavioral problems such as depression, anxiety, and aggression all take their own developmental course. At the same time, the communities in which children are developing in the face of war and political violence are also changing. Even for children who remain in their homes with intact families, the changes occurring around them will directly affect them or filter down to affect them over time. In some instances, changes will not be seen in children until years after conflict is over. Concern for children in these contexts is usually concentrated during the time that combat is occurring. Cairns and Dawes (1996) suggested that more research must be conducted on children after peace agreements are signed and children are left to deal with the broken environment that war and political violence leave behind. It is during these post-agreement periods, often time years after conflict exposure was experienced, that the consequences of the emotional, behavioral, and cognitive regulation to that exposure will come to fruition.

Mental Health and Psychopathology

Early studies of the effects of war and political violence on children focused on various forms of psychopathology such as

PTSD, depression, and anxiety. A recent review of the literature documented the effects of war on adolescents in multiple conflicts including those in Israel and Palestine, South Africa, Cambodia, Northern Ireland, Sri Lanka, Uganda, Kosovo, Croatia, Bosnia, and Herzegovina, among other places (Barber and Schluterman, 2009). The most frequently studied outcome of armed conflict is PTSD. Given the traumatic nature of events experienced during war, researchers have focused on PTSD with concerns that the violence experienced would disrupt children's ability to cognitively and emotionally process the events they experience leading to maladaptive responding that has long-term consequences for their physical and mental well-being. Relatedly, research has documented the increased risk for internalizing problems such as anxiety and depression for children exposed to war and political violence. Despite these findings across multiple conflicts, researchers have noted the relatively limited effect sizes and the number of studies finding null effects (Barber and Schluterman, 2009; Cairns and Dawes, 1996), suggesting that many children in these contexts show adaptive recovery from their experiences and the ability to use social support and other internal and external resources to overcome their experiences.

Aggression

A common concern regarding youth in contexts of war and political violence is the degree to which living in a violent community increases the likelihood of developing aggression and tendencies toward violence. In addition to the search for links between war and PTSD, depression and anxiety, much of the research on children and war has focused on aggression and related problems as an outcome of exposure to violence in these contexts. Many theoretical models within the social sciences would hypothesize that an increased exposure to violence would breed violent youth through processes such as modeling behavior, desensitization to violence, general increased physiological arousal, decreased moral development, and decreased cognitive processing of interpersonal information (Punamaki, 2009). Feelings of insecurity, distrust, and need for control can instill a sense of fear and defensiveness that motivates individuals to act aggressively as a form of self-protection (Staub, 2006). Although some research does suggest an increased risk for aggression and violence for children exposed to political violence and war, there are a number of individual differences that exist in this relationship. Research across contexts and conflicts has documented the increased risk for child development of aggression when they are exposed both directly and indirectly to war-related violence. At the same time, there are multiple individual- and contextual-level factors that either promote or deter the development of aggression in the context of violent conflict. Similar to the findings for the links between political violence and depression and anxiety, the studies linking war and political violence exposure to aggression in children has found limited and mixed effects.

Researchers are now looking to understand the multiple factors that explain this link along with the conditions under which children are most likely to develop problems with aggression in the face of political violence and war. As discussed above, the role of the family, and parents in particular are key variables in our understanding of the violence-aggression link. "In essence, the human mind is not merely a mirror for external violence; rather, children and adolescents construct their experiences according to the meanings given to them and the various social and psychological preconditions that accompany violence" (Punamaki, 2009: p. 74).

Identifying Risk and Protective Factors

Children affected by war and political violence show a range of outcomes across a large range of war-related situations. Given the heterogeneity of the issues, researchers have focused a great deal of attention on identifying specific factors that increase risk for negative outcomes while at the same time identifying specific factors that decrease risk for these same negative outcomes or those that increase the possibility of adaptive functioning. Risk and protective factors include factors about children themselves such as their gender and age during which the war exposure occurred, as well as factors external to the child such as their family environment, and other environments that contribute to or take away from the child's resources in dealing with the conflict they experience. Although many studies have examined individual characteristics like age and gender with many studies findings differences in responses for boys and girls and for children at various developmental periods, the findings are mixed, limiting the ability to make general statements about any overall impact of gender and age on children's responses to war and political violence. Regarding factors external to the developing child, research has focused on social support broadly defined. Receiving emotional support and guidance from families, schools, religious institutions, community groups among other organizations and individuals helps alleviate the effects of political violence and war on children making these systems critical to maintain and rebuild in the wake of conflict.

Another critical external factor for children exposed to political violence and war is the level of poverty of the family and surrounding community. War and political violence are also most likely to affect children living in poor communities (Cairns, 1996). The accumulation of stressful experiences for children in contexts of political violence has been shown to negatively affect children in poor environments due in part to a lack of financial and material resources, but also due to depleted emotional and family resources. Children living in low socioeconomic status conditions are more likely to experience negative effects of economic repercussions such as inflation, and research in other domains of child psychology have documented links between negative family environments and poverty suggesting that increased stress on the family system through war and political violence is more likely to affect those already in need.

The Role of the Family

It is widely understood that children develop in multiple social contexts with the family being the most proximal influence. Accordingly, much of the research examining for whom and under what conditions war and political violence affect

children examines the role of family members in the relations between violence and child development. The family unit experiences its own disruptions during war, and the break down of family bonds may even be a goal of tactics used in conflict. For example, as Cairns (1996) notes, the breaking of family relations was one of the tactics used by Pol Pot during the Cambodian genocide. In the fortunate circumstances when families do remain intact during war, parents' responses to the war and their children's exposure are thought to impact their children's responses to conflict. Commonly examined variables include parental responses to war, parenting variable such as parental warmth, and maternal and paternal psychopathology.

Several parenting variables have been examined as important to children's processing of stressful contexts. These variables include parental responsiveness, behavioral control, warmth, and acceptance. Research across contexts and conflicts has found that parents' mental health status has an impact on how children experience and react to political violence and war. Research across contexts such as Palestine and South Africa suggest that children are likely to suffer more negative outcomes in the face of war, if their mothers show signs of stressful responses such as posttraumatic stress symptoms (see Barber and Schluterman, 2009).

Intergroup Processes

Given the intergroup nature of many contexts of political violence and war, the effects of exposure to such violence will depend in part on the history of relations between the groups involved. Theorists and researchers across disciplines recognize that shared identification with groups is inherently a part of war and political violence (see Bar-Tal, 2007). Theories stemming from social psychology such as social identity theory and self-categorization theory describe psychological processes involved in the development and sustaining of negative intergroup attitudes.

The social identity paradigm (Tajfel and Turner, 1986) posits that the motivation for group members to positively differentiate their own group from relevant out-groups is partially driven by a need to achieve a sense of positive identity. Research has supported this idea showing that individuals with a stronger sense of identity with their in-group are more likely to have negative attitudes and behaviors directed at out-group members. Experimental studies have shown that the mere act of placing individuals into random groups is enough to create an increased likelihood for favoring in-group members and derogating out-group members. This experimental paradigm is known as the minimal group design. It follows that adding meaning to the groups with which we identify can increase the motivation to act negatively toward out-group members. In other words, if just being placed in an arbitrary group produces a natural motivation to act negatively toward out-group members, adding meaning, family ties, and history to these group memberships may increase the motivation to keep one's group in a higher, more power position than an opposing group. Adding the dimension of violence that is ethnically, religiously, or politically motivated to these naturally occurring intergroup processes clearly intensifies the motivations for the development of negative attitudes and the use of aggression against out-group members in contexts of political violence and war.

At the same time, children and adolescents gain support and meaning from their group and family connections, making the role of bonds to and identifying with the group complex as they relate to positive and negative responses to conflict. Individuals are more likely to give and receive social support from in-group members. Research has also shown that identifying with your ethnopolitical group that is part of the conflict, and understanding one's group's reason for participating in the conflict can help ameliorate the effects of experiencing violence associated with the conflict. This notion that group bonds play a dual role as both a benefit and a burden to positive child development is now being explored (Hammack, 2010). The results from studies across contexts such as Northern Ireland, Palestine, and Bosnia, suggest that the strength of identity with in-group members plays an important and powerful role in determining the behavioral, affective, and cognitive responses to exposure to intergroup conflict and violence for children.

Child Soldiers

Experience with political violence can take many forms. One of these forms in through participating in conflict as a child soldier or other forms of participation with armed groups. Recent estimates suggest that there are roughly 300 000 young people participating in armed forces or armed groups around the world (Betancourt et al., 2013). Youth in armed groups play a variety of roles including child soldiers, as sex slaves, spies, combatants, and laborers among other things. The conditions under which children are recruited and the experiences of child soldiers vary, paralleling the complexity of war experiences for children exposed to war and political violence outside of this particular experience. Children may be forced to participate in violent acts, taken from their homes, off the street or from orphanages, while others are recruited to join armed groups as a means of ensuring they are fed and protected. Child soldiers often go through a training period that may include witnessing and perpetrating killings, threatening, intimidating, or even killing other children in their group, and attacking civilians and villages. The Machel report discusses how the availability of small arms such as assault rifles and AK-47s have made it easier to recruit or force children and adolescents who can carry and use these weapons to join armed forces.

Research on child soldiers indicates that these children are at an increased risk of multiple mental health issues including PTSD, anxiety, and depression. The experiences of child soldiers are also greatly influenced by child gender. Girls often play particular roles as sex slaves, experiencing rape, impregnation, and high risk for sexually transmitted diseases. Girls are also at an increased risk with some studies suggesting lower levels of community acceptance after reintegration into their communities. In some cases, children are forced to take drugs such as amphetamines. Betancourt et al., 2013 recently conducted a review of relevant quantitative literature on child soldiers from 10 countries including Uganda, Sierra Leone, Liberia, Côte d'Ivoire, the Democratic Republic of Congo,

Mozambique, Nepal, Sri Lanka, El Salvador, and Germany. The review further documents the range of traumatizing experiences for youth in armed groups. Across conflicts high rates of children were exposed to violence through beatings, torture, and sexual abuse, and children perpetrated violence through injuring and killing others. Children were also exposed to violent attacks on family members and peers. Variables associated with decreased risk for children were family acceptance, social support, and educational/economic opportunities. In other words, children who have supportive families, opportunities for participation in the community through education and the ability to earn a living are more likely to show improvement in psychological functioning after reintegration. Research documenting the long-term adjustment of former child soldiers is limited and questions remain regarding the benefits of various elements of reintegration programs. The available research does suggest that, similar to the effects of other forms of exposure to political violence, family, and community support are key to helping youth adjust after they are reintegrated into their communities and families (Betancourt et al., 2013).

Treatment and intervention for child soldiers commonly include elements of Western approaches that focus on clinical approaches to trauma along with culturally grounded and community-based approaches. Many scholars in the area agree that the best approach includes multiple elements and supports for formerly recruited children across domains of functioning and society.

Intervention and Prevention Programs

Given the mixed findings regarding the outcomes for children in contexts of war and political violence, there is renewed interest in the question regarding what makes children resilient to these conditions. Resilience is a debated construct with different definitions and prerequisite conditions necessary within each definition. Recent reviews of the empirical literature on children in contexts of war and political violence suggest that the expected strong associations between child exposure to political violence and war and maladaptive outcomes in children are simply not consistently found (Barber and Schluterman, 2009).

As awareness of the needs of children in contexts of war and political violence has increased, researchers, clinicians, and local and international NGOs (nongovernmental organizations) have worked toward developing intervention programs specifically to meet the needs of this age group. Peltonen and Punamaki (2010) recently reviewed available literature in an attempt to evaluate the effectiveness of preventive intervention and treatment programs aimed at children exposed to armed conflict. Programs designed to reduce the negative impact of political violence on children utilized recommended treatment strategies and programs including cognitive behavioral therapy, trauma and grief focused therapy, along with other culturally and developmentally informed interventions including efforts to help alter attitudes out-group members, war, and peace. The limited number of experimental studies included in the review precluded the authors from making strong statements regarding the effectiveness of existing programs. From a subjective standpoint, the authors concluded that there is promise for programs that target youth with exposure to political violence; however, the dearth of experimental studies on the effectiveness of intervention programs limits the ability to make general recommendations for child-focused programming. Moreover, most of the evaluated methods of intervention were based in Western approaches that focus on individual mental health instead of utilizing culturally informed approaches that include traditional methods of understanding and treating mental health problems. There have been many criticisms of the use of individual-focused approaches such as cognitive behavioral therapy in cultures that use communal bereavement and treatment practices.

Although not specifically developed for youth exposed to war and political violence, other approaches based on the promotion of healing and reconciliation have shown positive effects on trauma symptoms and readiness to reconcile in postgenocide Rwanda (Staub et al., 2005). Staub and colleagues have also utilized radio programming to improve intergroup attitudes and behaviors, reflecting the notion that to be efficacious, interventions must target individuals, institutions, and cultural norms (Staub, 2006).

Developmental models must also continue to inform our understanding of how and when intervention should take place. Awareness of sensitive and critical periods of child development across social, cognitive, physical, and emotional domains should be considered in the development of programs for youth affected by war and political violence.

Youth as Victims, Perpetrators, and Agents for Social Change

Much of the recent research on children affected by war and political violence suggests that our understanding of the effects of war and peace-building must consider youths' voices and their active construction of their experiences. Children and young people make up a large portion of the world's population. Children and young people are active participants who can help society work toward positive social change. Moreover, analysis of recent conflict cycles of these intrastate conflicts suggests that conflict-ridden communities are likely to relapse into conflict (UNICEF, 2009). This pattern puts a concentrated focus on children and young people as key populations to target in reducing the relapse of communities into violent conflict. The break down of these cycles of conflict and relative peace must include a commitment by community members to solve disputes in nonviolent means. It is now recognized across international organizations such as the United Nations and nongovernmental organizations working to protect children and young people from the effects of war and political violence, that prevention of conflict or the relapse of conflict is the most promising means of protecting children. For peace-building efforts to be successful, programs must build and support capacities at multiple levels of the community and across domains of life. Sustained peace requires awareness of and programs targeting injustice and inequalities and a commitment to a culture of inclusion.

See also: Afrocentric Approaches to Social Work; Armed Conflict and Social Work: Significance and Implications; Child Protection: International Issues; Child Soldiers as Victims and Actors: Lessons from Ethiopia; Child Soldiers; Collective Behavior, Social Psychology of; Collective Memory, Psychology of; Collective Memory, Social Psychology of; Dehumanization; Emotions and Intergroup Relations; Ethnic Identity, Psychology of; Ethnicity and Ethnic Groups: Historical Aspects; Group Processes, Social Psychology of; Helping Intergroup Relations; Intergroup Relations; Justice: Social Psychological Perspectives; Mass Killings and Genocide, Psychology of; Political Psychology; Self and Identity Development During Adolescence across Cultures; Social Categorization; Social Identity in Social Psychology; Social Justice and Social Inequalities; Stress in Adolescence: Effects on Development; Stress in Childhood: Effects on Development.

Bibliography

Barber, B.K., 2009. Adolescents and War. Oxford University Press, Oxford.

Barber, B.K., 2013. Annual research review: the experience of youth with political conflict—Challenging notions of resilience and encouraging research refinement. Journal of Child Psychology and Psychiatry 54, 461–473.

Barber, B.K., Schluterman, J.M., 2009. An overview of the empirical literature on adolescents and political violence. In: Barber, B.K. (Ed.), Adolescents and War. Oxford University Press, Oxford, pp. 35–61.

Bar-Tal, D., 2007. Sociopsychological foundations of intractable conflicts. American Behavioral Scientist 50, 1430–1453. http://dx.doi.org/10.1177/0002764207302462.

Betancourt, T.S., Borisova, I., Williams, T.P., Myers-Ohki, S.E., Rubin-Smith, J.E., Annan, J., et al., 2013. Research review: psychosocial adjustment and mental health in former child soldiers – a systematic review of the literature and recommendations for future research. Journal of Child Psychology and Psychiatry 54, 17–37.

Bronfenbrenner, U., 1979. The Ecology of Human Development: Experiments by Nature and Design. Harvard University Press, Cambridge, MA.

Cairns, E., 1996. Children and Political Violence. Blackwell, Oxford, UK.

Cairns, E., Dawes, A., 1996. Children: Ethnic and political violence–A commentary. Child Development 67, 129–139.

Freud, A., Burlingham, D., 1943. War and Children. Medical War Books, New York.

Hammack, P.L., 2010. Identity as burden or benefit? Youth, historical narrative, and the legacy of political conflict. Human Development 53, 173–201.

Peltonen, K., Punamaki, R., 2010. Preventive interventions among children exposed to trauma of armed conflict: a literature review. Aggressive Behavior 36, 95–116.

Punamaki, R., 2009. War, military violence, and aggressive development. In: Barber, B.K. (Ed.), Adolescents and War. Oxford University Press, Oxford, pp. 62–80.

Staub, E., Pearlman, L.A., Gubin, A., Hagengimana, A., 2005. Healing, reconciliation and forgiving and the prevention of violence after genocide or mass killing: an intervention and its experimental evaluation in Rwanda. Journal of Social and Clinical Psychology 24, 297–334.

Staub, E., 2006. Reconciliation after genocide, mass killing, or intractable conflict: understanding the roots of violence, psychological recovery, and steps toward a general theory. Political Psychology 27, 867–894.

Steel, Z., Chey, T., Silove, D., Marnane, C., Bryant, R.A., van Ommeren, M., 2009. Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: a systematic review and meta-analysis. Journal of the American Medical Association 302, 538–549.

Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), The Psychology of Intergroup Relations. Nelson Hall, Chicago, pp. 7–24.

UNICEF, 2009. Machel Study 10-Year Strategic Review: Children and Conflict in a Changing World. UNICEF, New York.

UNICEF, 2007. Today's Superheroes Send Wrong Image to Boys, Say Researchers (Press release). Retrieved from: http://www.unicef.org/media/media_41228.html.

War Propaganda

Stacey Scriver, School of Political Science and Sociology, NUI Galway, Galway, Ireland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article examines the development of theories of war propaganda in the twentieth and twenty-first centuries. Beginning with early theorists of war propaganda the article assesses the development of a 'science of propaganda' following World War I, before considering more recent works on war propaganda that focus particularly on the role of informal sources, such as the media. Finally, the article briefly assesses the role of propaganda during the 1991 Gulf War.

While the use of propaganda in situations of war and conflict is neither new nor novel, the systematic development of propaganda as an aspect of war machinery has been a development of the twentieth century. World War I (WWI) brought with it the need to coordinate support for war across large expanses of territory, to bolster the morale of troops, and to weaken the resolve of the enemy. States on both sides of the conflict thus invested in the production of propaganda on a large scale. Since the WWI, war propaganda has become increasingly complex, drawing from fields including marketing, public relations, social psychology, and political theory, and integrating lessons learned from previous war campaigns and scholarship on propaganda. Much of these developments can be identified in Gulf War propaganda. This article examines the evolution of knowledge on war propaganda through the theories of war propaganda that developed during the twentieth-century World Wars, and the later developments in propaganda research that focused on informal sources of propaganda, such as the mass media. Finally, this article will briefly look at the role of propaganda during the Gulf War.

Propaganda and WWI: Developing a Science of Propaganda

Propaganda is a notoriously difficult term to define. It has been variously understood as 'a direct attack against man' (Ellul, 1965), as the 'systematic mistreatment of truth and information and their procedural safeguards' (Cunningham, 2002), 'a mere tool' (Lasswell, 1937) and 'a practical process of persuasion' (Taylor, 1992). Nevertheless, all definitions of propaganda share a common understanding in relation to the purpose of propaganda – to direct public sympathies and attitudes. While the use of propaganda in this sense can be observed in most aspects of social life, from politics to public health campaigns, its use within wartime contexts requires special examination. As all democratic states require, at least, the appearance of public consent to engage in conflict, propaganda serves an essential purpose during war: it may be employed to gather support for entry into war, to maintain support for an on-going war, to justify or legitimize certain actions during war, to direct public sympathies toward some foreign groups or away from others, to dishearten enemy forces, encourage uprising against the enemy government or military, or to develop sympathy among the enemy nation for the invading nation, prior to invasion. The importance of propaganda to war is such that it has been argued that the battle for public opinion is as important during a war as the engagement of soldiers on the front (Hiebert, 2003: p. 243).

Tools and Techniques of Propaganda during the Great War

Propaganda may consist of rhetoric, images, music, national holidays, and other significant cultural symbols which are presented with the intention of directing public opinion. The use of such techniques and tools in situations of war and violent conflict are not new; rather, they have been identified in use since antiquity (Taylor, 2003[1990]) and are likely to have been in use since the beginning of organized warfare. However, a systematic study of techniques for directing public opinion in contexts of war did not develop fully until the beginning of the twentieth century. WWI, and the need to secure public support for the war on all sides, led to the development of modern war propaganda that borrowed techniques from the fields of marketing and public relations (Kingsbury, 2010).

Edward Bernays noted that it was "during the Great War, [that] the Nations realised the necessity of selling their national aims and policies" (Bernays, 1942: p. 6). All sides of the conflict engaged in propaganda as an integral aspect of the war effort and institutionalized the production of propaganda in government offices, including the German Central Office for Foreign Services, the British War Propaganda Board (also known as Wellington House), and the American Committee on Public Information (CPI). German and British propaganda targeted their own populations to raise morale and support for the war and to vilify the enemy, and targeted the enemy population to undermine morale and discourage support. Airdropped pamphlets and radio broadcasts were common tools for the dissemination of the WWI propaganda campaigns for Britain and Germany along with film, children's books, posters, and many other cultural products. However, both countries also targeted the US in the early years of the war as a potential ally (Murphy and White, 2007: p. 17). To this extent, the US acted, before 1917, as "a theatre of official propaganda combat between Britain and Germany" (Messinger, 1992: p. 22).

Unlike German and British war propaganda the target population for the American CPI was primarily US civilians who had shown reluctance to enter the war, although the CPI also distributed propaganda materials in neutral states and behind enemy lines. Drawing knowledge from the fields of public relations and marketing, the CPI disseminated messages that were simple, one-sided, and which portrayed the enemy in a highly prejudiced

fashion. These messages were disseminated via a wide range of materials, from pamphlets and books to educational materials, films, and speeches, including the Four-Minute Men addresses in which individuals delivered prowar/anti-German four-minute speeches at public gatherings. Although effective in achieving the support from the American populace for involvement in the war, these techniques also had a 'dark side' – Germans were largely vilified. Foreshadowing later wars, particularly the War on Terror where patriotism turned French fries into American fries, sauerkraut became known as liberty cabbage, and German-Americans experienced physical attacks and discrimination (Murphy and White, 2007: pp. 19, 20).

WWI brought the importance of propaganda as a tool of warfare to the fore; indeed, it has been described as "a bloody and relentless struggle in which sustaining morale became just as essential for both sides as sustaining the military effort" (Taylor, 2003: p. 176). Following WWI considerable investment was made to improve propaganda techniques, building on lessons learned during WWI. Although World War II (WWII) utilized the same tools for propaganda dissemination as were used in WWI, such as airdropped pamphlets, educational materials and radio broadcasts, a greater sophistication in the directing of public opinion was evident. 'Black operations' were used, for instance where sources of information were obscured in order to manipulate perceptions of events or news stories. During WWII, radio programs were broadcast by British propagandists who posed as German officers in order to spread false information or information to turn the populace against Hitler's regime (See for instance, Sefton Delmer's account of the British black propaganda radio campaign, 'Black Boomerang', 1962). Further, the use of censorship became increasingly important in controlling the flow of information to the mass media, while maintaining the appearance of an independent press. For instance, in Britain all commercial cables were routed to a single point where news was censored before being released to journalists and editors, who were then free to present and comment on the precensored news as they wished (Taylor, 2003: p. 213).

Theories of War Propaganda

In academic circles, WWI initiated an interest in the role of propaganda in war and scholars began publishing analysis of WWI propaganda thus ushering in the development of a 'science' of propaganda. Public opinion specialist and political advisor, Walter Lippmann contributed to the understanding of propaganda a psychological approach which recognized the connection between perception, ideas, and actions (Lippmann, 1998[1922]: pp. 29, 30). He writes, "those pictures [in one's head] which are acted upon by groups of people, or by individuals acting in the name of groups, are Public Opinion with capital letters" (Lippmann, 1998[1922]). Lippmann further argued that the press could not, and should not, be relied upon to deliver the news objectively, rather, he recommended that the press be directed by experts whose role was to organize public opinion for the press (1998[1922]: p. 30). Although criticized at the time for being too instrumentalist, Lippmann's ideas were highly influential; Ellul and Herman and Chomsky, for instance, built on his ideas regarding the mass media, although largely rejected the benefits perceived by Lippmann of elite or 'expert' interaction with the press.

Of the early works by authors including Bernays (1942), Lippmann (1922), and Bruntz (1938), Lasswell's *Propaganda Technique in the World War* (1927a) was among the most influential. Lasswell, building on Lippmann's psychological approach, put forward a value-free conception of propaganda as the "management of collective attitudes by the manipulation of significant symbols" (Lasswell, 1927b: p. 627). Lasswell's understanding of propaganda is significant for several reasons; first, he identifies that propaganda functions through the organization of culturally significant symbols including music, the national flag, art work, and words, in order to organize public attitudes (Lasswell, 1927b: p. 629). Secondly, Lasswell recognized the importance of the propagandist within situations of conflict and war. Indeed, Lasswell wrote that "the problem of the propagandist is to intensify the attitudes favorable to his purpose, to reverse the attitudes hostile to it, and to attract the indifferent, or, at the worst, to prevent them from assuming a hostile bent" (Lasswell, 1927b: p. 629). Thirdly, Lasswell understood propaganda to be deliberate and intentional. Lasswell's conceptualization of the function of propaganda in war had much in common with that of his contemporaries from the worlds of public relations and marketing; for instance, Bernays argued that propaganda involves the primarily elements: to heighten "the morale-unity of your own country, weaken the morale of your enemy and win over the morale of the neutrals" (1942: p. 237).

Lasswell's analysis of WWI propaganda led him to argue that there are six significant factors evident in successful war propaganda:

1. one must fasten guilt to the enemy for creating the contexts of war;
2. one must claim a unity of the people through claims to common history and favor by God and proclaim the inevitability of victory;
3. war aims must be clearly stated that appeal to culturally held ideals, such as freedom, peace, or security;
4. disseminating examples of the enemy's depravity serves to strengthen the belief that the enemy is responsible for the war;
5. unfavorable news should be cast as enemy lies in order to prevent disunity and defeatism;
6. horror stories that cast the enemy in poor light should be told, dehumanizing the enemy and thus justifying violent action (in Bernays, 1942: p. 236).

These factors, in one form or another, have formed the basis of most subsequent analysis of the content of propaganda messages.

German scholars also contributed to the understanding of propaganda. For instance, Hadamovsky, who later became Chief of Staff of the Nazi Party's Central Propaganda Office, wrote *Propaganda and National Power* (1933). Building on the work of Lasswell and others, he contributed a recognition of the potential of ridicule as, he argued, "to laugh at the enemy is as important as to fear his strength" (Hadamovsky, 1933) and the importance of the political sphere, the military and public opinion being presented as a unified will (Hadamovsky, 1933). Thus, he writes "those who direct a war must at the same time direct politics and public opinion" (Hadamovsky, 1933).

Critiquing and Detecting Propaganda: The Seven Devices of Propaganda

The success of American and British propaganda during WWI brought with it a fear of being manipulated by propaganda. Propaganda became largely associated with 'the enemy' as opposed to an expectation of objective reporting and information dissemination from one's own country. In the US, interest grew among the public in developing skills to detect enemy propaganda. The Institute of Propaganda Analysis (IPA) published a framework for the analysis of propaganda in 1937. The framework laid out the following seven common propaganda devices (from Sproule, 2001):

1. *Name-calling* – attaches negative associations through the use of pejorative words and labels "to those individuals, groups, nations, races, policies, practices, beliefs, and ideals which he would have us condemn and reject."
2. *Glittering generalities* – "virtue words" such as "truth, freedom, honor, liberty, social justice, public service, the right to work, loyalty, progress, democracy, the American way, Constitution defender" are used to associate the propagandists' viewpoint with attitudes and ideas that are positively perceived.
3. *Transfer* – "the propagandist carries over the authority, sanction, and prestige of something we respect and revere to something he would have us accept." Sproule notes that "efforts to establish identification between a political project and the audience's reverence for national or religious symbolism" were common forms of transfer (2001: p. 136).
4. *Testimonial* – links the propagandists' position to some specific, favored person or institution, thereby associating their program or ideas with the positive associations already established with the person or institution.
5. *Plain folk* – As Sproule explains, "Persuaders and leaders present themselves as 'just plain folks' to establish an identity with ordinary Americans" (Sproule, 2001).
6. *Card stacking* – overemphasis and underemphasis are used to draw attention to some 'truths' and distract from others, in order to direct public opinion in such a way "that the audience forgets inconvenient information and embraces half truths" (Sproule, 2001).
7. *Band wagon* – The propagandist presents information so that it appears to be already accepted and thus to have people "follow the crowd," because "everybody's doing it" (Sproule, 2001).

Although criticized following WWII as being overly simplistic and adding little to Lasswell's analysis (Sproule, 2007: p. 131), the framework has made a lasting impression through its digestible rhetoric and easy application to propaganda content, and has been reinvigorated more recently through analysis of the Gulf War and the War on Terror.

Propaganda in the New World Order: The Growing Significance of Informal Sources of Propaganda

As in the World Wars, propaganda was an essential tool of the Cold War. While Soviet propaganda remained primarily a product of the State, the U.S. public demonstrated little tolerance for direct state-issued propaganda. Indirect control or covert control of information thus became an increasingly important aspect of propaganda in the US. Television news programs were particularly important tools for propaganda during the Cold War, with the state tightly controlling the flow of information which was reported by ostensibly independent news programs (Bernhard, 1999: pp. 4–6).

Since the 1970s scholarly interest in war propaganda has focused increasingly on the role of the news media and capacity for news corporations and journalists to deliver information independent from the interests of the state (Jowett and O'Donnell, 1992). While a number of authors have studied the role of the media in the dissemination of propaganda during periods of war and conflict, including Ellul (1965), Galtung and Vincent (1992), Jowett and O'Donnell (1992), and Schecter (2003), the Propaganda Model developed by Edward Herman and Noam Chomsky (1988) presents the most significant development in the analysis of the media's role in war propaganda in democratic states.

The Propaganda Model

Written during the Cold War, the Propaganda Model developed by Herman and Chomsky (1988) examines news media behavior in relation to state interests, calling into question the notion of the media as a 'fourth estate' tasked with holding government powers in check (Klaehn and Mullen, 2010: p. 12) and is applicable to states in which the news media is under corporate rather than state control. Rather, Herman and Chomsky argue that the media often reflects the interests of the economic and political elite, although "it is much more difficult to see a propaganda system at work where the media are private and formal censorship is absent" (1988: p. 1). The Propaganda Model sets out a series of five 'filters' which they argue act to cleanse the news of dissenting views, attitudes, and information via money and power. The filters are

1. the size, concentrated ownership, owner wealth, and profit orientation of the dominant mass-media firms;
2. advertising as the primary income source of the mass media;
3. the reliance of the media on information provided by government, business, and 'experts' funded and approved by these primary sources and agents of power;
4. 'flak' (i.e., negative responses) as a means of disciplining the media; and
5. 'anticommunism' as a national religion and control mechanism. (It is reasonable to now argue that anticommunism has been replaced by anti-Islamism as a control mechanism (see, for instance, Brown, 2008)).

Thus, in the context of the Cold War, Herman and Chomsky argue that "these elements interact with and reinforce one another. The raw material of news must pass through successive filters, leaving only the cleansed residue fit to print. They fix the premises of discourse and interpretation, and the definition of what is newsworthy in the first place, and they explain the basis and operations of what amount to propaganda campaigns" (1988: p. 2). Herman and Chomsky's model importantly contributes to the understanding of war propaganda a method for analyzing how an apparently independent mass media acts to reproduce the interests of the state and economic elites. In

terms of war propaganda this is particularly significant as it helps to explain how support for war can be created and maintained by the state, through an 'independent' mass media. As will shortly be examined, the first Gulf War presents one of the most significant examples of the mass media's role in disseminating prowar propaganda.

Propaganda and the Gulf War

As this article has argued, the World Wars established the contexts for the use of propaganda on a massive scale to garner support for war and to demoralize the enemy. It also created the contexts through which successful and unsuccessful propaganda could be studied leading to scholarship on war propaganda that drew from the World War experiences and incorporated knowledge from diverse fields, including public relations, marketing, and political science. This work has produced a compendium of factors evident in war propaganda. Across the various listings of factors, devices, and elements, there are several common aspects identifiable including: the need to establish a clear enemy, to present war as a just cause, to attach sentiments favorable to the war effort to positively perceived attitudes such as patriotism or liberty, and to reach the enemy with information that will act to negate attitudes of resistance and sway the general enemy populace toward views favorable to their enemy (see for instance, Bernays, 1942; Sproule, 2007). Herman and Chomsky (1988) further provide a method for understanding how power and wealth creates filters through which news is cleansed of information that may go counter elite interests. These tools provide the potential to better understand the function and form of war propaganda. In the following section, American propaganda during the Gulf War, disseminated by both the state and the news media, will be considered as it relates to the above factors.

Establishing the Enemy/Attaching Guilt

Harold Lasswell argued that war propaganda must, in the first instance, attach guilt to an enemy as an essential factor in directing public opinion toward prowar sentiments (1927a). Hadamovsky added that ridicule was as important as fear in portraying the enemy in a poor light (1933) while the IPA noted the role of 'name-calling' in war propaganda (Sproule, 2001: p. 136). In the lead up to, and during, the Gulf War Saddam Hussein became the target of ridicule, name-calling, and blame for the war; indeed, Saddam Hussein became perhaps the most vivid symbol of the Gulf War.

From 2 August 1990, the day Iraq invaded Kuwait, numerous articles describing Hussein, his past, his character, even his physical appearance appeared in newspapers in America. A brief scan of the New York Times between 3 and 5 August reveals Hussein as a 'megalomaniac' (Sick, 1990: p. 1), 'paranoid,' 'a mad bull,' 'coldly calculating,' and 'ruthless' (Lewis, 1990: p. 1) and comparisons with Hitler and the Blitzkrieg abounded. In President Bush's Address on the Invasion of Kuwait (1990b), he draws a connection between America's role in defending Kuwait against Hussein's aggression and America's role in defending Europe against Hitler, stating, "As was the case in the 1930's, we see in Saddam Hussein an aggressive dictator threatening his neighbors." The depictions of Hussein that centered on insanity and aggressive irrationality were important not only in creating a negative perception of Hussein but also in cementing support for Bush's position as a wartime leader. Bush further identifies Saddam Hussein and his actions as the cause of the war. He claims that Hussein had 'promised' not to invade Kuwait (Bush, 1990b). In this way, Bush's speech puts the guilt for starting the war squarely on the shoulders of Saddam Hussein and thus largely absolving the American populace of any such guilt.

However, as Hadamovsky noted, ridicule is also an important device of war propaganda. This was also mobilized during the Gulf War in relation to Hussein who featured heavily in cartoons, comedy sketches such as Saturday Night Live, even popular films, like 1991's Hot Shots! featuring a feminized Hussein lounging beside his pool with a cocktail, and in political speeches and newspaper articles. Such a characterization of the leader contributed toward the dehumanization of Hussein and ultimately to the rallying of support for the Gulf War (Conners, 1998: p. 96). The negative portrayal of Hussein was so successful that one commentator described Hussein as "George Bush's secret weapon in bringing about the war" (Mueller, 1993: p. 203).

Just Cause for War

Atrocity stories are a common feature of war propaganda. Such stories help to dehumanize the enemy, linking them with acts and attitudes that are unimaginable to the audience. This creates the conditions in which individuals are more likely to support entering into, or maintaining, a war. Images of past atrocities began to surface in newspapers after Iraq's invasion of Kuwait, including pictures of Kurdish children slaughtered with nerve gas. News agencies disseminating these images neglected to mention that its component parts may have been purchased through agricultural grants from the U.S. and that the helicopters used to spray the gas were American-made and purchased, or that President Reagan blocked the passing of a bill that would have punished similar attacks by cutting Iraq off from U.S. loans, military and nonmilitary assistance, credits, credit guarantees, and items subject to export controls (Hurd and Rangwala, 2001: p. 1), demonstrating an instance of what Sproule calls 'card-stacking' (Sproule, 2001: p. 136). One particularly inflammatory story about Iraqi soldiers pulling infants out of incubators in a hospital in Kuwait and throwing them on the ground to die elicited huge public outrage and was used no less than six times in a single month by George H.W. Bush to point out Saddam's ruthless and despicable nature, before the story was exposed as a lie. Nevertheless, the story was successful propaganda: it has been credited with swaying public opinion toward support of the war (Cohen, 1992: p. 1).

Positive Associations

On 11 September 1990, Bush addressed a joint session of congress. He stated in the opening paragraph of his speech that "following negotiations and promises by Iraq's dictator Saddam Hussein not to use force, a powerful Iraqi army invaded its trusting and much weaker neighbour Kuwait"

(Bush, 1990a: p. 1). The representation of a 'rich' and 'powerful' Iraq invading a 'smaller' and 'peaceful' Kuwait (Bush, 1990b) was used to create the semblance of a just war in which America was defending the weak, thus associating the war with positive notions of honor, liberty, and freedom. In Bush's 11 September 1990 speech he praised the service men in the Gulf, describing them as brave, well-trained, and dedicated and states, "if there was ever a time to put country before self and patriotism before party, the time is now" (Bush, 1990b: p. 2). This statement conflates patriotism with support for the war and dissent with selfishness. He goes on to state that America was defending "civilized values around the world" (Bush, 1990b), invoking the familiar concept of America's grand role as a civilizational force and thus its duty to defend 'civilization' when under threat. Indeed, as Allen et al.'s study of media framing in the Gulf War indicates, media coverage gave primary emphasis to themes that conflated patriotism, militarism, and nationalism (1994: p. 272). Bush's statements also reassured a skeptical population that the war was right, just, and moral. The common use of terms such as patriotism, civilization, liberty, freedom, and honor are, indeed, 'glittering generalities' as described by the IPA (Sproule, 2001: p. 136) which bring with them positive associations that help to obscure the complex and often distressing facts of the war and attach positive sentiment to the position of the speaker. In this way, the use of positive terms in association with the prowar position was an effective technique of Gulf War propaganda.

Demoralizing the Enemy

A further essential aspect of war propaganda, according to Bernays (1942: p. 237), is the demoralization of the enemy. American Gulf War propaganda did not shy from engaging in this aspect of propaganda. Efforts to demoralize the enemy during the Gulf War included the dropping of leaflets that resembled Iraqi Dinar but printed with the title of 'Safe Conduct Pass' informing the recipient that they 'do not have to die' if they lie down their weapons and cease resistance. Radio broadcasts, megaphone announcements, and pamphlets were also used to demoralize the Iraqi army and civilian populace and quell resistance. The effectiveness of the U.S. campaign to demoralize the Iraqi 'enemy' was described by one U.S. Colonel as follows:

> PSYOP [psychological operations] messages persuaded approximately 44 percent of the Iraqi army to desert, more than 17,000 to defect, and more than 87,000 to surrender… radio broadcasts and leaflets, helped cause the defection, desertion and surrender of some 40,000 Iraqis — all without firing a shot.
>
> Jones, 1994: p. 22

The Role of the News Media in Gulf War Propaganda

The Gulf War presents an exceptional opportunity to examine the role of the news media in disseminating state propaganda. Because of the nature of this war that was largely fought in the air and because of military regulations, the press depended on military briefings to gain information (Jacobs, 1992) and news was thus filtered prior to receipt by news-stations, similar to the experience of the news media in Britain during WWII. This limited the press' ability to engage directly with the war, but left them to relate the appearance of the war to an audience at home. However, the importance of appearance in the Gulf War was not limited to reporting on the war itself; securing the public approval for the Gulf War also depended heavily on the *appearance* of approval at home.

Despite massive antiwar protests and even polls that suggested that a plurality of Americans actually supported the notion of allowing Iraq to retain a portion of Kuwait in exchange for peace (Mueller, 1993: p. 203), the media continued to insist that the majority of Americans supported military intervention in Kuwait. Protests received little airtime and when protesters were shown, were portrayed as "untrustworthy, dishevelled, non-conforming 'others'" (Allen et al., 1994: p. 273). In fact, of the 2855 minutes of television coverage of the Gulf crisis between 8 August 1990, and 3 January 1991, only 29 minutes or 1% covered popular opposition (Allen et all., 1994: p. 257). The media suppression of dissent helped consolidate a view of unity within the nation, such that a 'unified will' was presented, conforming to the advice for successful propaganda by Hadamovsky (1933) and Lasswell (1927a) and the notion of the 'band wagon' by the IPA. This theory largely compares with Noelle-Neumann's Spiral of Silence (1984) hypothesis which argues that individuals will self-censor views that they perceive as unpopular so as to avoid social retribution and isolation resulting in a spiral of silence.

The media, in presenting the Gulf War as one that received tremendous approval (Kellner, 1992: p. 1), appears to have played a significant role in actually securing this approval through a band wagon or united will effect. Although the public was initially disinclined to engage in war, on 16 January 1990, polls demonstrate a sudden shift toward support for military action (Allen et al., 1994: p. 255). In reviewing Paletz analysis of the media in the Gulf War through the lens of the Propaganda Model, Herring and Robinson identify where the Propaganda model can be implicitly identified, shown by Herring and Robertson below in square parenthesis:

> Insufficient dedication to the freedom of the press, fear of provoking governmental outrage [the flak filter of the propaganda model], shared frames of reference with governing elites [the ideological filter of the propaganda model], and the pursuit of sales and ratings [the advertising filter of the propaganda model] are among the factors that can help explain the acquiescence to government curbs, no matter how reluctant, of media executives.
>
> Herring and Robinson, 2003: p. 560

In this analysis, the media thus appears as a tool of propaganda. Through the expectation of independence, the public largely accords the mass media a special status as 'testimonial,' thus permitting the dissemination of information as neutral. However, as information passes through various filters, the state and corporate elites (often aligned) strip the news of information harmful to their position. Nevertheless, through reference to their special status, the media dissemination of selected information which reflects elite interests, acts to create the appearance of a unified will which the public will fall in line with. In contexts of war and conflict, the news media is therefore an important source of propaganda, with the capacity, as

described in relation to the Gulf War, to direct public opinion in a significant ways.

The Potential of War Propaganda

This article has provided a brief examination of a selection of the most influential scholars of war propaganda in the twentieth and twenty-first centuries. War propaganda is both a contributor, and a response, to war and conflict and scholarship on war propaganda demonstrates both these impulses: it has attempted to advance the success and possibilities of propaganda to support war efforts and to establish the means to recognize and counter propaganda. While the early twentieth-century theorists of war propaganda analyzed the form and function of propaganda commonly perceiving it as a neutral tool to be used to direct public opinion, propaganda took on an increasingly negative image following WWI and analysis began to focus rather at defending against or deciphering propaganda. Most recently, the behavior of the mass media and non-state sources of war propaganda have been the focus of scholarly inquiry. The growth of social media as a tool for the dissemination of war propaganda will no doubt spawn new theories and models among scholars of propaganda. Nevertheless, what is shared by the approaches to the study of war propaganda is the recognition that propaganda is a significant aspect, and tool, of war.

See also: Military Sociology; Securitization; Violence and Media; War, Sociology of.

Bibliography

Allen, B., O'Loughlin, P., Jasperson, A., O'Sullivan, J.L., 1994. The media and the gulf war: framing, priming and the spiral of silence. Polity 27 (2), 255–284.

Bernays, E.L., 1942. The marketing of national policies: a study of war propaganda. Journal of Marketing 6 (3), 236–244.

Bernhard, N.E., 1999. U.S. Television News and Cold War Propaganda, 1947–1960. Cambridge University Press, Cambridge.

Brown, W., 2008. Regulating Aversions: Tolerance in the Age of Identity and Empire. Princeton University Press, UK.

Bruntz, G.C., 1938. Allied Propaganda and the Collapse of the German Empire in 1918. Stanford University Press, Stanford.

Bush, G.H.W., 11 September 1990a. Toward a new world order, address to a joint session of congress and the nation. Sweet Liberty [www.document]. www.sweetliberty.org/issues/war/bushsr.htm (accessed 01.01.04.).

Bush, G.H.W., 8 August 1990b. Address on Iraq's Invasion of Kuwait [www.document]. http://millercenter.org/president/speeches/detail/5529 (accessed 19.02.12.).

Cohen, M., 28-30 December 1992. What we say goes! How Bush Sr. sold the bombing of Iraq. Counterpunch. Weekend Edition December 28-30. Available at: http://www.counterpunch.org/cohen1228.html (accessed 15.04.08.).

Conners, J.L., 1998. Hussein as enemy: the Persian Gulf War in political cartoons. The International Journal of Press/Politics 3 (3), 96–114.

Cunningham, S.B., 2002. The Idea of Propaganda: A Reconstruction. Praeger, Westport, CT.

Delmer, S., 1962. Black Boomerang: An Autobiography, vol. 2. Secker and Warburg, London.

Ellul, J., 1965. Propaganda: The Formation of Men's Attitudes (K. Kellen, J. Lerner, Trans.). Vintage Books, New York.

Galtung, J., Vincent, R.C., 1992. Global Glasnost: Toward a New World Information and Communication Order? Hampton Press, Cresskill, NJ.

Hadamovsky, E., 2007[1933]. Propaganda und nationale Macht: Die Organisation der öffentlichen Meinung für die nationale Politik (Propaganda and Nation Power: The Organization of Public Opinion for National Politics) (R. Bytwerk, Trans.). Gerhard Stalling, Oldenburg. Available at: http://www.calvin.edu/academic/cas/gpa/hadamovsky.htm.

Hiebert, R.E., 2003. Public relations and propaganda in framing the Iraq war: a preliminary review. Public Relations Review 29, 243–255.

Herman, E., Chomsky, N., 1988. Manufacturing Consent: The Political Economy of the Mass Media. Pantheon Books, New York.

Herring, E., Robinson, P., 2003. Too polemical or too critical? Chomsky on the study of the news media and US foreign policy. Review of International Studies 29, 553–568.

Hurd, N., Rangwala, G., 2001. U.S. diplomatic and commercial relationships with Iraq, 1980–2 August 1990. Casi [www.document]. http://www.casi.org.uk/info/usdocs/usiraq80s90s.html (accessed 10.04.08.).

Jacobs, M.J., February 1992. Assessing the constitutionality of press restrictions in the Persian Gulf War. Stanford Law Review 44 (3), 675–726.

Jones, J.B., 1994. Psychological operations in desert shield, desert storm and urban freedom. Special Warfare 7 (3), 22–29.

Jowett, G.S., O'Donnell, V., 1992. Propaganda and Persuasion, second ed. Sage Publications, Newbury Park.

Kellner, J., 1992. The Persian Gulf TV War. Westview Press, Boulder. Available at: http://pages.gseis.ucla.edu/faculty/kellner/essays/gulfwarch2.pdf (accessed 05.12.12.).

Kingsbury, C.M., 2010. For Home and Country: World War I Propaganda on the Home Front. University of Nebraska Press, London.

Klaehn, J., Mullen, A., 2010. The Propaganda Model and Sociology: Understanding the Media and Society. Paper 5. Sociology Faculty Publications. Available at: http://scholars.wlu.ca/soci_faculty/5.

Lasswell, H.D., 1927a. Propaganda Technique in the World War I. Alfred A. Knopf, London.

Lasswell, H.D., 1927b. The theory of political propaganda. The American Political Science Review 21 (3), 627–631.

Lasswell, H.D., 1937. Propaganda. In: Seligman, E.R.A., Johnson, A. (Eds.), Encyclopedia of the Social Sciences. MacMillan, New York, pp. 521–528.

Lewis, F.L., 4 August 1990. Foreign Affairs: Fruit of Appeasement. New York Times [www.document]. http://query.nytimes.com (accessed 08.04.08.).

Lippmann, W., 1998[1922]. Public Opinion. Transaction Publishers, London.

Messinger, G.S., 1992. British Propaganda and the State in the First World War. Manchester University Press, Manchester.

Mueller, J., 1993. American public opinion and the gulf war. In: Renshon, S.A. (Ed.), The Political Psychology of the Gulf War: Leaders, Publics and the Process of Conflict. University of Pittsburgh Press, Pittsburgh, PA, pp. 199–226.

Murphy, D.M., White, J.F., Autumn 2007. Propaganda: can a word decide a war? Parameters 15–27.

Noelle-Neumann, E., 1984. The Spiral of Silence. A Theory of Public Opinion – Our Social Skin. University of Chicago Press, Chicago.

Perry, R., 2008. On My Honor: Why the American Values of the Boy Scouts Are Worth Fighting for. Stroud and Hall, Macon.

Sanders, M.L., 1975. Wellington house and British propaganda during the first world war. The Historical Journal 18 (1), 119–146.

Schecter, D., 2003. Media Wars: News at a Time of Terror. Rowman & Littlefield, Lanham.

Sick, G., 8 February 1990. Hussein Must Be Stopped. New York Times. Available at: http://query.nytimes.com (accessed 08.04.08.).

Sproule, M.J., 2001. Authorship and origins of the seven propaganda devices: a research note. Rhetoric & Public Affairs 4 (1), 135–143.

Sproule, M.J., 2007. Propaganda and Democracy: The American Experience of Media and Mass Persuasion. Cambridge University Press, UK.

Stark, G., 1998[1930]. Moderne Politische Propaganda (Modern Political Propaganda) (R. Bytwerk, Trans.). Verlag Frz. Eher Nachf, Munich. Available at: http://www.calvin.edu/academic/cas/gpa/stark.htm.

Taylor, P., 2003[1990]. Munitions of the Mind: A History of Propaganda from the Ancient World to the Present Era. Manchester University Press, Manchester.

Taylor, P., 1992. War and the Media: Propaganda and Persuasion in the Gulf War. Manchester University Press, Manchester.

Whitton, J., Larson, A., 1964. Propaganda: towards disarmament. The War on Words 10, 62–82.

War, Social Causes and Consequences

Edgar Kiser, University of Washington, Seattle, WA, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, War, Sociology of, volume 24, pp. 16363–16367, © 2001, Elsevier Ltd.

Abstract

The sociology of war is a central topic in both political and historical sociology, since war is one of the most important policies states can pursue, and the outcomes of wars have often shaped both the formation and the dissolution of states. The literature on war is thus concerned with both its causes and its consequences. Studies of the causes of war can be divided into three broad categories. The first type takes the system as a whole as the unit of analysis and focuses on how characteristics of the interstate system affect the frequency of war. States are the unit of analysis in the second type, which explores the relationships among political, economic, and cultural features of particular states and their propensity to initiate wars. The third type analyzes war as an outcome of choices made by individual and small-group decision making. Theories of the consequences of war tend to focus either on its role in state formation, or on its causal impact on internal revolts and revolutions. Historical sociologists have shown that the frequency, duration, and timing of medieval and early modern warfare were the most important determinants of the size and structure of states. However, just as war can make states, it can break them too. Costly warfare often leads to fiscal crises and state breakdown, facilitating revolutions.

The sociology of war is a central topic in both political and historical sociology, since war is one of the most important policies states can pursue, and the outcomes of wars have often shaped both the formation and the dissolution of states. The literature on war is thus concerned with both its causes and its consequences.

Studies of the causes of war can be divided into three broad categories. The first type takes the system as a whole as the unit of analysis and focuses on how characteristics of the interstate system affect the frequency of war. Debates focus on characteristics of the interstate system that are thought to increase or decrease war, such as global economic cycles, balances of power, and the increasing role of transnational organizations such as the United Nations. States are the unit of analysis in the second type, which explores the relationships among political, economic, and cultural features of particular states and their propensity to initiate wars. Social scientists disagree about the effects of political systems (democracy vs. autocracy) and economic systems (capitalist, socialist, or other) within states on war. The third type analyses war as an outcome of choices made by individual and small-group decision making. There is also no consensus on which model of individual decision making is most appropriate for the study of war. Is the decision to go to war based on a rational calculation of economic costs and benefits, or is it an irrational outcome of distortion in decision making in small groups and bureaucracies?

Theories of the consequences of war tend to focus either on its role in state formation, or on its causal impact on internal revolts and revolutions. Historical sociologists have shown that the frequency, duration, and timing of medieval and early modern warfare were the most important determinants of the size and structure of states (Tilly, 1975, 1990; Ertman, 1997). However, just as war can make states, it can break them too. For example, Skocpol (1979) argues that costly warfare often leads to fiscal crises and state breakdown, facilitating revolutions.

Causes of War: The Interstate System

Most studies of war that take the interstate system as the unit of analysis begin with assumptions from the 'realist' paradigm. States are seen as unitary actors, and their actions are explained in terms of structural characteristics of the system. The most important feature of the interstate system is that it is anarchic. Unlike politics within states, relations between states take place in a Hobbesian 'state of nature.' Since an anarchic system is one in which all states constantly face actual or potential threats, their main goal is security. Security can only be achieved in such a system by maintaining power. In realist theories, the distribution of power in the interstate system is the main determinant of the frequency of war.

Although all realist theories agree on the importance of power distribution in determining war, they disagree about which types of power distributions make war more likely. Balance-of-power theories (Morgenthau, 1967) suggest that an equal distribution of power in the system facilitates peace and that unequal power distributions lead to war. They argue that parity deters all states from aggression and that an unequal power distribution will generally result in the strong using force against the weak. When one state begins to gain a preponderance of power in the system, a coalition of weaker states will form to maintain their security by blocking the further expansion of the powerful state. The coalitions that formed against Louis XIV, Napoleon and Hitler seem to fit this pattern.

Hegemonic stability theory (Gilpin, 1981) suggests exactly the opposite, that unequal power in the system produces peace and that parity results in war. When one state has hegemony in the world system, it has both the incentive and the means to maintain order in the system. It is not necessary for the most powerful state to fight wars, since their objectives can be achieved in less costly ways, and it is not rational for other states to challenge a hegemon with overwhelming power. For example, the periods of British and US hegemony were relatively peaceful and World Wars I and II occurred

during intervening periods in which power was more equally distributed. A related attempt to explain great-power war is power transition theory (Organski, 1968). Power transition theory suggests that differential rates of economic growth create situations in which rising states rapidly catch up with the hegemonic state in the system, and that this change in relative power leads to war.

Debates about power transitions and hegemonic stability are of much more than theoretical interest in the contemporary world. Although the demise of the USSR has left the USA as an unchallenged military hegemon, its economic superiority is being challenged by the European Union and emerging Asian states (Japan in the short run, perhaps China in the long run). If power transition and hegemonic stability theories are correct, this shift of economic power could lead to great power wars in the near future.

Another ongoing debate about systemic causes of war concerns the effects of long cycles of economic expansion and contraction. Some scholars argue that economic contraction will increase war, since the increased scarcity of resources will lead to more conflict. Others have suggested the opposite: major wars will be more frequent during periods of economic expansion because only then will states have the resources necessary to fight. Goldstein's (1988) research suggests that economic expansion tends to increase the severity of great-power wars but that economic cycles have no effect on the frequency of war.

One significant change in the last half of the twentieth century which will require substantial revisions in realist systemic theories of war is the development and increasing power of transnational organizations (such as the United Nations), since their assumption that the interstate system is anarchical may no longer be valid. If the military power of the United Nations continues to grow, it could become more and more effective at preventing wars and suppressing them quickly when they do start. Of course, it remains to be seen whether powerful existing states will choose to cede more power to such institutions.

Theoretical debates about the systemic causes of war have not been resolved, in part because the results of empirical research have been inconclusive. Each theory can point to specific cases that seem to fit its predictions, but each must also admit to many cases that it cannot explain. Part of the problem is that systemic theories have not incorporated causal factors at lower levels of analysis, such as internal economic and political characteristics of states. Since the effects of system-level factors on war are not direct but are always mediated by the internal political economy of states and the decisions made by individual leaders, complete theories of the causes of war must include these factors as well.

Causes of War: Capitalism and Democracy

One of the longest and most heated debates about the causes of war concerns the effects of capitalism. Beginning with Adam Smith ([1776]1976), liberal economists have argued that capitalism promotes peace. Marxists (Lenin, [1917]1939), on the other hand, suggest that capitalism leads to frequent imperialist wars.

The Smithian liberal argument suggests that since capitalism has both increased the benefits of peace (by increasing productivity and trade) and the costs of war (by producing new and better instruments of destruction), it is no longer rational for states to wage war. The long period of relative peace that followed the triumph of capitalism in the nineteenth century and the two world wars that came after the rise of protectionist barriers to free trade are often cited in support of liberal economic theories (but the same facts can be explained by hegemonic stability theory as a consequence of the rise and decline of British hegemony).

In contrast, Marxists (Lenin, [1917] 1939) argue that economic problems inherent in advanced capitalist economies create incentives for war. First, the high productivity of industrial capitalism coupled with a limited home market due to the poverty of the working class result in a chronic problem of 'underconsumption.' Capitalists will thus seek imperial expansion to control new markets for their goods. Second, capitalists will fight imperialist wars to gain access to more raw materials and to find more profitable outlets for their capital. These pressures will lead first to wars between powerful capitalist states and weaker peripheral states, and next to wars between great powers over which of them will get to exploit the periphery.

With the increasing globalization of economics, and the transitions of more states to capitalist economies, the debates about the effects of capitalism, trade, and imperialism on war become increasingly significant. If Adam Smith is right, our future is likely to be more peaceful than our past; but if Marxist theorists are right, we may see economically based warfare on an unprecedented scale.

The form of government in a country may also determine how often it initiates wars. Kant ([1795]1949) argued that democratic states (with constitutions and separation of powers) will initiate wars less often than autocratic states. This conclusion follows from a simple analysis of who pays the costs of war and who gets the benefits. Since citizens are required to pay for war with high taxes and their lives, they will rarely support war initiation. Rulers of states, on the other hand, have much to gain from war and can pass on most of the costs to their subjects. Therefore, when decisions about war are made only by rulers (in autocracies), war will be frequent, and when citizens have more control of the decision (in democracies), peace will generally be the result.

Empirical research indicates that democratic states are less likely than nondemocratic states to initiate wars, but the relationship is not strong (Kiser et al., 1995). Perhaps one reason for the weakness of the relationship is that the assumption that citizens will oppose war initiation is not always correct. Many historical examples indicate that in at least some conditions citizens will support war even though it is not in their economic interests to do so, due to nationalism, religion, ethnicity, or other cultural factors.

Perhaps the most interesting current debate about democracy and war surrounds the proposition that democratic states never fight each other. There is clearly a strong empirical generalization to be explained, since all agree that democratic states rarely fight each other – depending on exactly how 'democracy' and 'war' are defined, some argue they never do (Weart, 1998). However, scholars disagree about the causal mechanism responsible for this association. Some stress the role of political culture, arguing that the norms of toleration and mutual accommodation that prevent conflicts within

democracies from escalating to violence have the same effect in limiting violent conflicts between democracies. These states consider each other part of the same 'ingroup' sharing the same values, and are thus very unlikely to fight. In contrast, others suggest that 'democratic peace' could be the result of rational self-interest. Democratic politicians may simply fear the negative impact that losing a war might have on their prospects for re-election. Further research, probably at the level of detailed case studies that can reveal precise causal mechanisms, will be necessary to resolve this debate.

Causes of War: Decision Making

Few theories of war focus on the individual level of analysis; their assumptions about individual decision making are usually implicit or undeveloped. Notable exceptions include rational choice theories (Bueno de Mesquita, 1981, Bueno de Mesquita and Lalman, 1992) and arguments about organizational and small group decision making (Allison, 1971).

Bueno de Mesquita begins by assuming that the decision to initiate war is made by a single dominant ruler who is a rational expected-utility maximizer. Rulers calculate the costs and benefits of initiating war, and the probability of victory, so wars will be initiated only when rulers expect a net gain from them. These assumptions generate several counterintuitive propositions. For example, common sense might suggest that states would fight their enemies and not their allies, but Bueno de Mesquita (1981) argues that war will be more common between allies than between enemies. Wars between allies are caused by actual or anticipated policy changes that threaten the existing relationship. The interventions of the USA in Latin America and of the USSR in Eastern Europe since World War II illustrate the process. Other counterintuitive propositions suggest that under some conditions a state may rationally choose to attack the stronger of two allied states instead of the weaker, and under some conditions it is rational for a state with no allies to initiate war against a stronger state with allies.

Other analyses of the decision to initiate war focus on how the social features of the decision-making process lead to deviations from rational choice. Allison (1971) argues that standard operating procedures and repertoires within states tend to limit the flexibility of decisions and make it difficult to respond adequately to novel situations. Others focus on the small groups within states (such as executives and their cabinet advisers) that actually make decisions about war. The cohesiveness of these small groups often leads to a striving for unanimity that prevents a full debate about options and produces a premature consensus. In spite of these promising studies, work on the deviations from rational choice is just beginning, and we are still far short of the general microlevel theoretical model of the decision to initiate war.

Consequences of War: State Formation and Bureaucratization

How and to what extent does war affect the formation and the structure of states? Tilly (1975, 1990) argues that 'war made states' in the early modern era. Warfare (along with the repayment of debts from past wars) cost far more than any other state policies in early modern Europe. Moreover, the timing and the nature of war shaped the structures of these developing states.

One of the most interesting ongoing debates concerns whether war facilitates or hinders bureaucratization. Weber ([1922]1978) argued that states involved in military competition with other states (e.g., Western Europe) would be more likely to bureaucratize than those that were more isolated (e.g., China, Japan). Although he does not use this terminology, he basically argues that states not facing the threat of war would be satisfied with existing administrative arrangements, whereas those competing militarily would be forced to adopt the more efficient bureaucratic form.

In contrast to this, Levi (1988) views war as a consistent impediment to bureaucratization. She argues that war raises the discount rates of rulers, causing them to pursue policies that provide immediate gains even if they are costly in the long term. Thus, rulers facing war would be unlikely to pay the high start-up costs of bureaucratization, but would instead do things like selling offices which would make bureaucracy much more difficult to implement.

Ertman (1997) suggests a related argument in which the timing of war is important. When states experienced early sustained warfare (prior to about 1450) they developed patrimonial administrations (due to lack of trained personnel and the dominance of 'cultural models' derived from feudal and Catholic institutions). These institutions were very difficult to bureaucratize due to the power of entrenched officials to block reform. Ertman argues that states that were able to avoid frequent war until later were able to develop more bureaucratic administrations.

Finally, wars that result in severe losses may facilitate bureaucratic reforms. One of the main barriers to bureaucratization is the entrenched officials in the state administration who have both the incentives and the power to block reform (Ertman, 1997). These officials will not be dislodged by most wars, but their power will be broken by a severe loss at war, especially one that results in foreign occupation.

Consequences of War: Revolt and Revolution

Since the classic work of Simmel ([1908]1955, pp. 98–99) and Coser (1956, pp. 19, 95), the conventional wisdom has been that wars decrease the probability of revolts by increasing the internal cohesion of societies (the 'ingroup-outgroup' or 'conflict-cohesion' hypothesis). However, the results of empirical tests of the relationship between revolt and war have been mixed, at best.

In contrast to the 'conflict-cohesion' hypothesis, historical sociologists focusing on the early modern period (Tilly, 1975, p. 74) argue that fighting wars increased the likelihood of revolts. Tilly argues that since subjects had few institutionalized mechanisms available to stop rulers from initiating wars contrary to their interests, the only option they had to try to limit war was revolt. In addition, Tilly (1975, p. 74) notes that since participating in wars weakens rulers, it increases the likelihood of a revolt.

Theda Skocpol (1979) has made the most compelling argument that war is a primary cause of major 'social revolutions.' Her argument about the origins of the French, Russian, and Chinese revolutions stresses the importance of factors that weaken the state, making it more vulnerable to overthrow. The main factor that tends to weaken states is war, especially in the context of strong peasant communities and an alienated dominant class. This work has been the primary inspiration for the development of a 'state-centered' approach to historical sociology.

Conclusion

This short summary has only been able to scratch the surface of the voluminous literature on war. Future research should attempt to link the several topics discussed here, by bringing together the micro and macro causes of war, and by tying the causes more closely to the consequences of war. The increasing development of technologies of mass destruction and the rise of transnational political units will also challenge existing theoretical frameworks. Like most of sociology, the sociology of war is still in its infancy.

See also: Conflict and War, Archaeology of: Weapons and Artifacts; Conflict and War: Anthropological Aspects; Geopolitics; Military Geography; Military History; National Security Studies and War Potential of Nations; Peace-Making in History; Peace; War: Anthropological Aspects, Historical Development of; Wars among Nation-States: Patterns and Causes.

Bibliography

Allison, G., 1971. Essence of Decision. Little Brown, Boston.
Bueno de Mesquita, B., 1981. The War Trap. Yale University Press, New Haven, CT.
Bueno de Mesquita, B., Lalman, D., 1992. War and Reason. Yale University Press, New Haven, CT.
Coser, L., 1956. The Functions of Social Conflict. Free Press, Glencoe.
Ertman, T., 1997. Birth of the Leviathan. Cambridge University Press, Cambridge, UK.
Gilpin, R., 1981. War and Change in World Politics. Cambridge University Press, Cambridge, UK.
Goldstein, J., 1988. Long Cycles. Yale University Press, New Haven, CT.
Kant, I., 1949. Eternal peace [1795]. In: Friedrich, C. (Ed.), The Philosophy of Kant. Modern Library, New York, pp. 430–476.
Kiser, E., Drass, K., Brustein, W., 1995. Ruler autonomy and war in early modern Western Europe. International Studies Quarterly 39, 109–138.
Lenin, V.I., 1939. Imperialism [1917]. International Publishers, New York.
Levi, M., 1988. Of Rule and Revenue. University of California Press, Berkeley, CA.
Morgenthau, H., 1967. Politics Among Nations. Alfred A. Knopf, New York.
Organski, J., 1968. World Politics. Alfred A. Knopf, New York.
Simmel, G., 1955. Conflict and the Web of Group Affiliations [1908]. Free Press, New York.
Skocpol, T., 1979. States and Social Revolutions. Cambridge University Press, Cambridge, UK.
Smith, A., 1976. The Wealth of Nations [1776]. University of Chicago Press, Chicago.
Tilly, C. (Ed.), 1975. The Formation of National States in Western Europe. Princeton University Press, Princeton, NJ.
Tilly, C., 1990. Coercion, Capital, and European States, 990–1990. Blackwell, Oxford.
Weart, S., 1998. Never at War: Why Democracies Will Not Fight One Another. Yale Universtiy Press, New Haven, CT.
Weber, M., 1978. Economy and Society [1922]. University of California Press, Berkeley, CA.

War, Sociology of

Lars B Kaspersen, Copenhagen University, Copenhagen K, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Organized, collective violence has rarely been seen as a driving force in societal development and, consequently, warfare has been a neglected object of investigation. In recent years, sociology have become increasingly interested in war and its implications for society. In particular, scholars within the field of historical and comparative sociology address the issue. This article argues that sociology of war must focus on the relationship between the units waging war and the implications of warfare for societal development. These units are political orders of all kinds (most often we call them states) but inspired by Norbert Elias (1897–1990) here they are called survival units. The core of sociology of war concerns the study of the consequences of war in relation to the stability and/or change of state and society.

Introduction

War seems to be one of the few constants in human history. This has not, however, been reflected in the history of social thought in general nor in sociology in particular. We do find, however, some reflections on collective violence and war in classical social thought (from Thomas Hobbes (1962) to Max Weber (1980, 1991)) but war as such has never been an established object of investigation for sociological research. In the postwar period, in particular, war and collective violence have almost been omitted from the sociological agenda. It is significant that both collective violence within societies and between states are ignored as important social phenomena. In recent years, however, war has gained more interest partly due to 11 September 2001 and the implications (wars in Afghanistan and Iraq, terrorism) and partly due to a renewed interest in the relationship between warfare and state formation processes.

Thus sociology of war as a research field and a subdiscipline has developed fairly recently. How come that war for many decades was almost neglected by sociologists? The omission of war and politically organized violence in sociology and social theory can be seen as a consequence of social theory and sociology being embedded in liberalism and Marxism. Liberalism and Marxism are both worldviews, which promise us a world without violence. Michael Mann (1988) and more recently Hans Joas and Wolfgang Knöbl (2013) have pointed to this explanation. Liberalism argues that an increasing global division of labor and stronger interdependence between nations will reduce the incentives for warfare. Marxism argues along the same line but adds that conflicts will disappear in the communist society when classes are dissolved.

This seems to be a plausible explanation but it must be added that the neglect of violence and warfare in social theory also originates in a specific conceptualizations of society – the key concept in classical social theory. Society is seen as a fusion of its internal elements – as a social organism, a social system, an aggregation of individual social actions, or an aggregation of social institutions. Following from this social change is seen as processes taking place within society. Although Marx argues that class struggle is an important driving force for societal development (Karl Marx and Friedrich Engels, 2002; Karl Marx, 1958) most explanations are related to differentiation, increasing division of labor, processes of rationalization, the changes in the political and economic institutions, changes in demography, or the emergence of new ideas and ideologies. Thus organized, collective violence is rarely seen as a driving force in societal development. Since war, as a social phenomenon, most often takes place between societies/states, warfare has become even less important because most often intersocietal relations play a less important role in a sociological analysis. Moreover, the predominant understanding of society has involved a problematic conceptualization of the state as an entity derived from society. Therefore the state and its activities will always be a reflection of the conflicts and interests dominating in society. The state cannot be conceived as an autonomous actor separated from society pursuing its own (geopolitical) interests. Consequently, a fusion understanding of society and state has led to a neglect of warfare.

As pointed out above in recent years sociology and social theory have become increasingly interested in war. In the following sections a number of issues related to sociology of war will be presented but first we need a clarification of the term 'sociology of war.'

What Is Sociology of War?

A clarification of the term 'sociology of war' requires a look into the two concepts, sociology and war. It is beyond the scope of this article to discuss the concept of sociology in detail. In this context sociology is equal to social theory. Inspired by relational sociology (Norbert Elias, 1978; Pierre Bourdieu, 1977; Christopher Powell and François Dépelteau, 2013; Nick Crossley, 2011; Pierpaolo Donati, 2011; Mustafa Emirbayer, 1997) sociology/social theory is defined as the study of social relations, e.g., face-to-face interaction between a husband and a wife, the relationship between a management team and the employees, and the relationship between states. Following from this, sociology/social theory studies how particular sets of social relations reproduce and change. When particular sets of social relations reproduce over time, they constitute institutions, organizations, and/or actors. Thus a continuous

interaction between states constitutes these states as entities and they become 'real' existing entities. When we are interested in the sociology of war the focus will be the relationship between the units waging war – the units are political orders of all kinds (most often we call them states) but here we will call them survival units (Elias, 1978). The claim is that war is a social practice conducted only by emerging or existing survival units. When workers movement is fighting in the streets against the police, this might be an organized collective violence but not a war. Before we move into a clarification of war we need to specify the key form of social organization in this context – the survival unit.

Survival Units – A Point of Departure for Sociology of War

Human beings are always embedded in social relations and the claim here is that since man erected and walked out of Africa, human beings have been organized in 'survival units.' Survival units are as old as human beings but these are not states in the Weberian sense. A survival unit is a political and social order constituted in relations to other survival units. It can take different forms such as bands, tribes, city-leagues, city-states, or nation-states. Another way to characterize a survival unit is to see it as a process of constantly organizing and reorganizing in order to survive and consequently, with an ability to provide security and material reproduction. According to Elias the world consists of a multiple number of coexisting and competing survival units. A survival unit is always located in a figuration – a set of interdependent relations between several survival units struggling for recognition by other survival units (Elias, 1978). For clarification it is worth mentioning that a figuration can be seen as a variant of a network structure, and some figuration sociologists (Baur and Ernst, 2011) regard figuration analysis as equivalent to modern social network analysis (John Scott, 2000).

So far we have not laid out an argument that determines the difference between human beings and other animals. A struggle for food, shelter, and protection is common to most if not all mammals. Social groups are struggling for recognition (Hegel, 1977, 1991) and honor (Bourdieu, 1977). Social groups of human beings are in this very struggle experiencing stages of being conscious and self-conscious and, consequently, developing an identity. This ability to experience, feel and gain consciousness, self-consciousness, and identity is apparently distinctive human. Identity, in other words, is a human feature, which is as crucial as eating, sleeping, and having a shelter. In our endeavor to achieve identity we have to confront 'the other,' individually as well as a social group. In our daily material life we hunt, gather, grow, and trade food, we build shelter, we make clothes, and during these social practices we struggle and express our identity. Sometimes we have to fight for these material interests and this fight is not merely a fight for material survival. It is also a fight for spiritual and mental survival – in other words, a struggle of identity. Identity is crucial to human beings and it can only be expressed through our social practices and property (Hegel). The key point is that identity can only be achieved by being in a social relation. Identity is dependent upon 'the other.' Someone outside yourself must recognize you. 'The other' is the precondition of identity-making processes. This struggle of identity is in the last instance a life-and-death struggle – which under certain circumstances can lead to war. This conflictual relationship is the center of concern for sociology of war.

Elias argues that these survival units knit "people together for common purposes – the common defense of their lives, the survival of their group in the face of attacks by other groups and, for a variety of reasons, attacks in common on other groups." He continues:

> Thus the primary function of such an alliance is either physically to wipe out other people or to protect its own members from being physically wiped out. Since the potential of such units for attack is inseparable from their potential for defense, they may be called 'attack-and-defence units' or 'survival units'. At the present stage of social development they take the form of nation-states. In the future they may be amalgamations of several former nation-states. In the past they were represented by city-states or the inhabitants of a stronghold. Size and structure vary: the function remains the same. At every stage of development, wherever people have been bound and integrated into units for attack and defense, this bond has been stressed above all others.
> (Elias, 1978: pp. 138–139)

The relationship between survival units (or states) is a confrontation, which generates identities and boundaries. A state becomes a state when it is in a social relationship with another state. It is exactly in the very moment when two or more states are interacting that they constitute each other as states (Hegel, 1991: pp. 359–371).

Hegel explains exactly that the concept of state is unthinkable unless we see a state as a part of a social relationship – vis-à-vis another state. Similarly to an individual who cannot become self-conscious and know that 'I' is me before 'another' has recognized me – 'I' – from the outside, Hegel clearly sees that a state can only become a state with a bounded territory when the boundaries of this territory are drawn from outside. Elias follows a similar kind of relational thinking by stressing the importance of the very relationship for the 'entities' – the survival units. In this very relationship, boundaries are created from 'outside.' The survival units are demarcated by other survival units and not just by the members of the survival unit itself. Their very relationship constitutes states.

Survival units also take on the role as the 'primary' figuration because of their high degree of autonomy, which no other figurations have. Of course, these figurations are also interdependent with other figurations but they have autonomy to the extent they can "fulfil effectively for their members their function as self-reliant and self-regulating defence and survival units" (Elias, 1974: p. xxii). Their high level of autonomy consists in their ability to defend and survive. As long as a unit can defend its own domain of sovereignty (whether this is territory, hunting fields, seaways, or something else), it can be argued that it has the largest possible form of autonomy. If no other unit can encroach upon your domain of sovereignty you can uphold a political order and you are as autonomous as you can be. This is a crucial difference between survival units and other figurations such as families, cities, or companies. Families or companies can very rarely protect themselves from internal and external enemies.

As long as the survival unit is capable of preventing external enemies from encroaching upon its domain of sovereignty, there will be some degree of freedom and autonomy for families, individuals, and businesses.

In looking at the key role of survival units, Elias also emphasizes the problem of violence. He argues that the violent aspect of the survival unit is in the last instance the key dimension of the state/survival unit (1987a: pp. 74–86). In a discussion of the cold war dynamic between the two superpowers, the USA and the former USSR, Elias emphasizes how this social dynamic creates competitive pressures, which continually change and move these states. Their violent potential is the key means to maintain and improve the position of a state in a system of states.

This interstate dynamic is a life-and-death struggle, which consists of the use or threat of the means of violence. There are many different forms of survival units, which are all interdependent because they are competing with each other in a struggle for survival. It is a contest without rules but it is not chaotic. The struggle between survival units is 'a primal contest.' Two survival units struggle for survival – whether it is about prestige, scarce resources, or other things – they are dependent on each other. Each and every step taken by each of the two parties is watched by the other. As enemies they perform a function for each other and each move of one survival unit determines each move of the other survival unit and vice versa (Elias, 1978: pp. 76–80).

This interdependence exists as a precondition for and influences the character and nature of the internal structure of survival units. In modern terms, the organization and structure of society is to a large extent determined by the intensity of the struggle between states. The character and organization of internal social structures is a function of external structures. Their function for each other is in the last resort based on the compulsion they exert over each other because of their interdependence. It is not possible to explain the actions, plans, and aims of either of the two groups if they are conceptualized as the freely chosen decisions, plans, and aims of each group considered on its own, independently of the other group. They can be explained only if one takes into account the compelling force that groups exert upon each other and their bilateral function for each other as enemies (Elias, 1978: p. 77).

The crucial difference between interstate and intrastate relations is the presence and the restrained use of the means of violence. Elias draws upon Max Weber (1864–1920) in his understanding of the survival unit by arguing that the use of violence is an ever-present threat and the normal instrument of last resort in interstate relations. The use of physical violence has almost been eliminated from normal relations within developed states. This is not only indicative of the fundamental difference between the structure of human relationships within states, but also of that of the relationships between states. This significantly means that human beings live simultaneously at two levels whose structure is not only different, but in some respects contradictory. At one level it is strictly forbidden to be violent and to kill people; at another, it is demanded as a duty to prepare for, and to use, violence in relations with other humans (Elias, 1987a: p. 80).

Elias points out that there is a clear difference between, on the one hand, figurations called states or survival units that are interrelated in a state system and, on the other hand, figurations found at the intrastate level such as families, companies, and voluntary associations. The existence of monopolies of the means of violence within states and the nonexistence of such monopolies at the interstate level can clearly explain differences in social structure. A given state-society – a network of functionally interdependent human beings – has a structure of its own. They are bound to each other in specific figurations whose dynamics have a constraining and compelling influence on those who form them.

The existence of a monopoly of physical force within states and its nonexistence in the relationships between states is an example of the firmness of the structure, which interdependent human beings form with one another. It also shows the far-reaching effects, which these structures have on those who form them (Elias, 1987a: p. 79).

We should also be aware that these survival units take on different forms in different periods in history. Over time they have grown in size – from small bands, tribes, city-states, village-states, to modern large-scale states and nation-states.

At this point it should be clear that the survival unit is the center of concern when it comes to war. It is a social and political organization with the capacity to wage war, and the political locus in the survival unit decides on war and peace. Any survival unit deciding on war is situated in a life-and-death struggle to remain or to gain recognition. War, or the threat of war, generates a number of opportunities for changes in the survival unit or the related society. Sociology of war is an attempt to understand the implications of war for survival units – it is about social change related to war!

War

This requires a clarification of the concept of war. War existed before recorded history. Almost all survival units have engaged in some form of combat against others. This does not mean that war is 'natural,' whatever that means. It is not reducible to an instinct for aggression; even if such a well-articulated instinct did exist, there are plenty of other avenues for its expression (Hirst, 2005: p. 51). As Carl Von Clausewitz (1780–1831) says: "War is the continuation of policy by other means." War is the organized pursuit of interest, it involves violence but it is never just a fight between angry people. War is a conduct. It is something that arises in and through structured human interaction. War happens when social groups categorize others as 'enemies' and take organized action against them involving force. A relation of enmity is one in which, as Hobbes says, the will to give battle is known. Fighting may be episodic, but the possibility of conflict is constant (Hirst, 2005: p. 51).

Enmity exists because of a conflict of interest between groups in relation to a mutual process of recognition that cannot be resolved except by combat, in which one group attempts to prevail in some way over the other. "The interests that group people into friends and enemies are various: they can be material or symbolic, groups fight over resources, but they also fight over principles, beliefs, wrongs done to them, or as a means of defining the individual status honour of warriors or of establishing groups prestige" (Hirst, 2005: p. 51).

At its simplest, war consists of combats. Clausewitz starts off with the analogy of two wrestlers each of whom attempts to force the other to submit. From this simple initial definition he is able to construct a complex and comprehensive theory of war. His core definition is: "War, therefore, is an act of violence intended to compel an opponent to fulfill our will." Such acts of violence are not rule-governed, war is not a game. It is complex but unregulated reciprocal action. Each participant in the combat must anticipate and respond to the actions of the other. Hence, war is inherently uncertain, combat cannot be planned. Because war involves two sides each composed of individuals who must cooperate and coordinate, because it takes place in a context of space and time that cannot be given, because it involves a changing physical environment, it is subject to those forms of uncertainty and chance that Clausewitz calls 'friction.' To overcome the enemy and to counter friction, fighters must be resourceful, act together, and have high morale. Other things being equal, it is morale, the will to win and to accept suffering and effort in order to do so, that ultimately decides the outcome. Given forces of equivalent numbers and morale, reciprocal action, the logic of blow and counterblow, means that there is an inherent tendency in war to ascend to extremes, to ever-greater violence, the ultimate state of which Clausewitz calls 'absolute war.' This state of absolute war is never attained precisely because of friction (Hirst, 2005: p. 52).

It is important to make it clear that war is a political phenomenon and the purpose of war is politically determined. By making the distinction between purpose and aim, Clausewitz can separate politics and war. Politics is the subject of the theory, which starts the war and provides the purpose of war. War itself, however, has its own aim.

The following theory cannot tell anything about the character of the political will involved. It only emphasizes that war is a means, and thus only an instrument of politics.

The concept of war can be specified in three 'layers': the 'explosive,' the 'virtual,' and the 'limited' (or 'moderate') war. Clausewitz himself operated with a theory containing two forms of war: explosive and moderate war. From Anders Boserup (1940–90), however, we know that a third form can be deduced from Clausewitz's text: 'the virtual war' (Boserup, 1986).

In explosive war, there is complete polarity between the two opponents, since both have the same unambiguous, universal, and diametrically opposed end. This polarity involves a displacement of the purpose of politics in favor of the real end of war: winning a victory/avoiding defeat (ziel replaces zweck). It is a true duel on a large scale (Clausewitz, 1993: p. 83) in which all means can be used. The aim is clear: to disarm our enemy in order to impose our will. This leads to an explosive and unlimited war where hesitation can mean losses and/or defeat. This state of war is rarely to be found in the real world, but it still throws light on central features of war.

The next form is a modification of explosive war, called virtual war. Although explosive war takes place now and then, it does not happen very often. Consequently, Clausewitz needed a concept, which could explain the fact that there often is a standstill in war. He needed a concept to explain why the polarity between the two opponents breaks down. If explosive war had dominated history, it would follow that a war would never stop before one of the two opponents had been completely defeated by the other opponent because of a complete polarity – you either win or lose. As we know, this is rarely the case. Clausewitz manages to explain the occurrence of pauses in war and the breakdown of polarity between the opponents by introducing two forms of fighting: the offensive and the defensive principles of fighting. The defensive form of warfare is, in principle, always stronger than the offensive war, which explains why wars stop (Clausewitz, 1986[1832]: Chap. 1, book 1: pp. 15–16). Clausewitz points out that the defensive (D) is stronger than the offensive (O) (D > O), but he has no theoretical explanation for it. Here Boserup can be a useful supplement (Boserup, 1986).

Boserup's argument of the superiority of the defensive is rooted in the condition that the offensive only possesses means, which can be freely mobilized ('the army'), whereas the defensive has not only those but also its own purely defensive forces at its disposal. These forces consist of a series of means that are only raised in consequence

> (…) of the progress of the offensive itself and can only be mobilized by this, such as mountains and rivers, the resistance of the civilian population, and the support from those countries that fear the future strength of a victorious attacker.
> (Boserup, 1986: p. 921)

Thus, the defensive has the possibility of strategically combining its bound and freely mobilized forces, whereas the offensive is only capable of operating with freely mobilized forces in its strategy.

The superiority of the defensive is thus the reason why there may be a real pause, and the length of the pause is conditioned by the strength of the defensive. But during this pause

> (…) the war is continued as a virtual war, for the action of war has ceased, it is true, but the war 'goes on' in a particular sense, viz., as an always present possibility that is only not actual for as long as the conditions of the pause are maintained. The pause is only a precarious, temporary balance, conditioned by the forces keeping each other in check, and thus, therefore, even if the war is only 'virtual', it imposes real claims on the two antagonists.
> (Boserup, 1986: p. 915)

Had the defensive not been stronger than the offensive, history would have looked totally different. If the offensive principle of fighting had been strongest, history would have appeared as an uninterrupted state of war that would only have been terminated when the world was united as one global reign.

By developing the concept of virtual war, Boserup emphasizes that the possibility of war is always present but as long as your defense capacity is stronger than the offensive strength of your opponents, war will not be actualized. It will remain 'virtual.'

The limited war is the third form where political motives (zweck) are once again brought into the picture. During the pause, war may

> (…) take on more limited forms where the antagonists only go for small gains that offer themselves, and where both parties may be said, in a certain sense, to be in the defensive!.
> (Boserup, 1986: p. 915)

Clausewitz's concept of the pause opens up for a new understanding of peace. Peace is here understood as a pause based on the superiority of the defensive. It is the inequality of force between the defensive and the offensive that makes a 'peace' possible, during which societal and state units may exist.

The concept of the pause can also be seen in relation to Hegel's struggle of recognition. The pause is the actualization of a mutual recognition between two survival units (alliances). During the pause a genuine mutual recognition can take place and peace occurs, but it is very precarious because the possibility of war is ever present.

It is important to notice that the concept of survival unit in this context is not a set of political and governance institutions only, but a whole and a unity that includes society. It is not, however, a unity constituted of its internal elements.

The German legal and political theorist Carl Schmitt (1888–1985) approves the Hegelian idea that the very struggle between two entities constitutes this relationship and their two identities respectively. Schmitt says that

> (…) the political entity presupposes the real existence of an enemy and therefore coexistence with another political entity. As long as a state exists, there will thus always be in the world more than just one state. A world state which embraces the entire globe and all of humanity cannot exist. The political world is a pluriverse, not a universe. (…) The political entity cannot by its very nature be universal in the sense of embracing all of humanity and the entire world.
> (Schmitt, 1976: p. 53)

However, he takes this point much further when he specifies that the struggle of recognition is mainly a political relation existing partly among states and partly among persons, groups, or movements in civil society within the state territory. Schmitt argues that the differentia specifica of the political is a friend–enemy relation.

The friend–enemy relationship is a mutual struggle of recognition, and in this case it means a relationship dominated by reciprocal action in which "one can be driven by the opponent's moves, if one does not anticipate the competition and drive it instead" (Hirst, 1990: pp. 130–131). It follows that at the most extreme political struggle includes the possibility of war (Schmitt, 1976: pp. 33–37).

Wars follow from enmity. War is the existential negation of the enemy. It is the most extreme consequence of enmity. It does not have to be common, normal, something ideal, or desirable. But it must nevertheless remain a real possibility for as long as the concept of the enemy remains valid (Schmitt, 1976: p. 33).

Politics can lead to war. Schmitt takes us back to Clausewitz, where "war is the continuation of politics by other means." Schmitt, and more distinctly Clausewitz, elucidate the necessity of the means of violence in order to survive as a survival unit. Violence and military means can become the ultima ratio of a survival unit. It is important, however, to make it clear that war is subordinated to politics. Schmitt indicates that the political relationship and the realm of politics determine the purpose (zweck) of war. The political relationship can initiate the war, but the very aim of the war, to achieve victory/avoid defeat, is entirely determined by the logic of war.

Schmitt has demonstrated that the mutual struggle of recognition that constitutes the survival unit is a political relationship, and at the extreme this struggle involves the possibility of war. War is the ultimate struggle of recognition, and the independence and sovereignty of a survival unit can be at stake. The sociology of war presented here emphasizes that a survival unit can never be adequately theoretically understood unless it is situated in a relation with other survival units.

When a survival unit is recognized it is always a preliminary and precarious recognition, and recognition is always at stake. The very struggle always compels the survival unit to provide a defense. Thus the struggle between the survival units makes heavy demands on each survival unit. The survival unit has to produce a strong political will, economic means, and an ideological support from the population. Therefore, the survival unit must possess a society that can provide these means to maintaining a strong defense. Thus the very struggle of recognition and its specific character has a strong impact on the survival unit, the relationship between the structure of the political order and the structure of the society.

Sociology of War: Social Change and Social Reproduction

So far we have outlined sociology of war with a focus on the always potential conflictual relationship between survival units (or emergent survival units) and the consequences of these conflicts (including the preparations for these conflicts) for the internal organization of the survival units. We have discussed the concept of war in order to determine that war is a particular form of organized conduct, which can be applied to 'solve' conflicts. It has also been argued that war is a particular expression of an existential relationship between two survival units/states, which is used – often in the last instance – to preserve sovereignty and remain recognized as an independent political order. Also, it should be clear that war is an instrument for politics, and consequently, war can only be initiated by political agents with decision-making powers, which are located in survival units or emerging survival units. Thus an outcome of the wars in the former Yugoslavia was new survival units/states such as Bosnia, Croatia, and Slovenia (Mann, 2005; Mulaj, 2008; Ramet, 2010). They were not survival units before the wars, but the war turned them into real existing survival units. Following on from this it should be clear that in principle there is no distinctive difference between war and civil war. A civil war is potentially always a war because the opposing wills usually fight for maintaining one survival unit or to provide the possibility for secession in order to create a new survival unit/state.

Sociology of war examines war as a social transformative force. War is a 'state of exception' in which rules, norms, and

values are suspended and/or undermined, and therefore regime change is not unusual and in addition to this other structural changes take place such as new or amended constitutions, new rights and obligations, a new and different state–citizen relationship, the establishment of new core institutions, etc. A potential conflict, an external threat or an actualized war – in other words, virtual wars as well as real wars – can have an effect on the internal structures and create social change.

When a survival unit – in modern terms a state – is confronted with a threat, prepares for a war or takes part in a war, it will go through a number of changes with implications for ideologies, the self-consciousness of state and population, the relationship to the 'other,' and the self-perception of state and society. It stirs up the social and political order and, in particular, the defeating part in a war is often exposed to a more thorough transformation process. Often the defeat leads to new political leaders, maybe even a regime shift. In some cases the victor imposes a new regime on the defeated and at times even a new economic system, e.g., the Eastern European countries after World War II.

Another aspect concerns in the context of war how different political and social fractions and their struggles for power develop within a particular state/survival unit. If oppositional groups seize the power with violent means, we will characterize it as a revolution. An outcome of a war can be a revolution and most major social and political revolutions in world history have taken place in the context of war, e.g., the French Revolution, the Russian Revolution, and the Chinese Revolution (Skocpol, 1979).

A particular tradition within historical sociology and comparative politics examines the relationship between warfare and state-formation processes. There is a tradition often named the bellicose wave within historical sociology going back to Otto Hintze (1861–1940) and Max Weber (1864–1920). The bellicose wave has become institutionalized as a paradigm with scholars such as Charles Tilly (1975, 1992), Michael Mann (1986, 1993), Anthony Giddens (1985), Gianfranco Poggi (1978), Thomas Ertman (1997), Brian Downing (1992), and Norbert Elias (2000). They all emphasize warfare as a driving force for state formation and the increasing state capacity building over time. War compels the states to increase the ability to extract resources, to mobilize manpower, to build fortifications, etc. War provides a necessity of social change. In addition to this some of these scholars claim that an unintended consequence of warfare is the creation of group homogeneity such as patriotism or nationalism. The classical example is how Napoleon's victorious armies generated not only French patriotism and nationalism but nationalism all over Europe, e.g., Prussia.

Among the most well-known and important scholars within this historical sociological tradition addressing these sociology of war issues most directly and comprehensively are Charles Tilly (1929–2008) and Michael Mann (b.1942).

Charles Tilly: Coercion and Capital

Charles Tilly claims that although many other theories in social sciences and humanities contribute with valuable aspects of state formation, most of them are flawed. Tilly here emphasizes two important problems: A widespread tendency to focus on the state and its internal construction without considering the impact of the state–state relation. Another problem is the neglect of the importance of war and war preparations for the structure of the state. Theories in international relations take this into account, but warfare is usually regarded as a result of the individual state actions – not as a consequence of the very relation between the states. These observations are important and very valuable contributions to social theory in general and sociology in particular. This critique provides the backdrop for Tilly's own theoretical framework. Tilly takes his point of departure in the concept of the individual. Men control means of coercion. These men tried to extend the range of people and resources over which they had power. They became conquerors. Conquerors, who managed to exert stable control over the populations in territories and to extract resources, became rulers. Rulers made wars because they now and again faced limits to their rule. War making led to state making. Tilly takes his point of departure in the interdependence and logics between capital and coercion. These two dimensions develop as an interrelated couple. The expansion of coercion is linked to the development of states, whereas capital is linked to cities. He emphasizes coercion and wielders of coercion as the starting point for the process (Tilly, 1992). The state reflects the organization of coercion and capital – at least in its most developed form. When concentration of coercion and accumulation merge, an increasing growth of state power will take place.

What processes push the accumulation and concentration of means of coercion? Tilly is unequivocal in his answer that war involves accumulation and concentration of means of coercion and consequently, state formation. How come war has been en generic aspect of the world history? Because coercion works. "Those who apply substantial force to their fellows get compliance, and from that compliance draw the multiple advantages of money, goods, deference, access to pleasures denied to less powerful people" (Tilly, 1992: p. 70).

In other words, coercion promotes a compliance, which gives access to resources. This logic seems to be a common feature of the European development. A more specific war-provoking logic was present: Everyone (i.e., kings, princes, barons) who controlled means of coercion tried to maintain a 'security area' and a 'buffer zone' to protect the security area. At times they included the buffer zone in the security area, and consequently expanded to maintain a new buffer zone. This logic leads to war and, according to Tilly this logic has been a generic element of the European history since 1990.

Tilly claims that war is the most important activity of the state during most of the history until our time. A general trend has been that the world has seen more wars with higher intensity but fewer wars between great powers. These wars have become even more intensified and cruel.

War has been the most dominant state activity but not the only one. However, the other activities are all connected to warfare. The minimum activities of the state are as follows:

1. State making: The ruler attacks and controls enemies and competitors within the claimed territory. (Here Tilly implicitly accepts that territory is a part of a state definition.)
2. War making: The ruler attacks enemies outside the territory.

3. Protection: Attacking and controlling the enemies of the ruler's allies (inside or outside the territory).
4. Resource extraction: extraction of resources from the population to state making, war making, and protection.

The war initiates a whole range of processes in the state. The state is forced to war making, state making, protection, and extraction, otherwise the war/defense capability weakens and, as a consequence, the sovereignty of the state is at stake.

These state functions generate more functions like adjudication, distribution, and production of goods. A by-product of the war activities is the development of a number of institutions and organizations like courts, legal institutions, financial institutions, tax administration, regional administration, public assemblies, and parliaments. These institutions later became independent forces with their own dynamics in the state and society.

Tilly examines how the forms and structures of the control of the means of coercion changed from a period with small warrior lords and princes and a fragmented sovereignty (my expression) to large territorial domains. This long process bears witness to a development in which the state is strengthened externally by a strong internal pacification.

Increasingly, the state gets control over the means of violence and succeeded in disarming the civilian population and preventing access to weapons. This created a completely new structure of state and society. The perennial threat from private wars, attacks, internal violence, riots, etc., was reduced to a minimum and the state could concentrate on the threat from external enemies.

War drives the state to an extraction of resources. The state developed financial institutions including a tax and revenue administration. Collecting taxes required extended surveillance and control of the population. Therefore, a huge apparatus was necessary. The tax base was nobility, towns, and peasants, and particularly the towns and cities became increasingly more important in time with the development of commerce and the money economy. Here the state could extract many resources.

The evolution of new institutions and organizations usually took place during wartime. After the end of a war, it became very difficult to demolish them and slowly they became permanent. Military organizations, revenue and finance, credit and banking, and legal institutions must all be seen as a direct or indirect effect of the war-making activities of the state.

Michael Mann: War, States, and Power

The British-American sociologist Michael Mann's main thesis is that society is constituted of many overlapping, intersecting sociospatial power networks. Mann provides a theory and history of power relations. We replace society as our primary unit of analysis with power networks. Mann argues that "a general account of societies, their structure, and their history can best be given in terms of the interrelations of the four sources of social power: ideology, economy, military and political (IEMP) relationships" (Mann, 1986: p. 2).

He underlines that these are overlapping networks of social interaction, not levels or dimensions of a single society. They are also organizations, institutional means of attaining human goals (Mann, 1986).

In terms of methodology Mann argues that it is necessary to make a concrete analysis on a sociospatial and organizational level. This involves the problems concerning organization, control, logistics, and communication. "The capacity to organize and control people, materials, and territories, and the development of this capacity throughout history (…) my history of power rests on measuring sociospatial capacity for organization and explaining its development" (Mann, 1986: pp. 2–3).

Mann contributes to sociology of war by bringing in military power. Military power is defined as the social organization of concentrated and lethal violence. 'Concentrated' means mobilized and focused; 'lethal' means deadly (Mann, 2013: p. 2). Military power becomes intertwined with political power in interstate wars. Political power is linked to the state and is defined as the centralized and territorial regulation of social life (Mann, 2013: p. 2). In order to bring war into social theory it requires a reconceptualization of the state. The state wages war – not simply because it is hijacked by the ruling class to fight for its economic interests – but because the state has its own agenda. Consequently, the autonomy of the state is central in Mann's works (Mann, 1984, 1986, 1988, 1990).

Mann's project is an attempt to break with, in particular, the reductionism characterizing liberalistic, Marxist, and functionalist theories, in which the state is reduced to preexisting structures in the civil society. Mann is also drawing upon the legacy from the bellicose thinkers such as Hintze, Weber, Gumplowicz, and Oppenheimer. The state has two dimensions: one has to do with the economic and ideological conditions of civil society and the other with the state's international military conditions. The state responds to two kinds of interests/pressures (both internal and external) generating a particular autonomous 'space' where the state-elite may maneuver and play off the classes against each other.

The origin of autonomous power is related to the necessity of the state. Concerning the necessity of the state, Mann takes his point of departure in a Hobbesian conception of society. The existence of a society is conditioned by the presence of rules that are capable of protecting life and property. These rules originate from a monopoly. Moreover, it is said that states always exist in a system of states, which also imposes certain rules of behavior onto each state. The autonomous power of the state proceeds from this necessity – "necessity is the mother of state power" (Mann, 1984: p. 196).

Only the state has a particular inherent centralized authoritative power over a specific territory. In contrast to economic, ideological, and military groups in civil society, the resources of the state-elite radiate authoritatively outward from a center, which only comes to a halt at a territorially defined border. "The state is, indeed, a place – both a central place and a unified territorial reach" (Mann, 1984: p. 198).

Precisely this property distinguishes the state from other powerful societal groups such as businesses, capital owners, the Church, or the nobility. The autonomous power of the state proceeds precisely from the difference in the organizational structure and sociospatial conditions. "Territorial centralization provides the state with a potentially independent basis of power mobilization being necessary to social development and uniquely in the possession of the state itself" (Mann, 1984: p. 201).

The autonomy of the state therefore consists of its necessity, the plurality of functions and territorial centralization. The power of the state-elite cannot be reduced to the power of any group and it is thus autonomous in relation to civil society.

The development of an active state with great infrastructural power extends territorialization, hereby also the demarcation of society. The development of the European nation-states is used as an example. Two processes in European development contributed to a further degree of centralization and territorialization from the twelfth century onward.

The first condition is the military development, which required more capital (to mercenary soldiers) and later also men skilled in the use of arms (conscripts). Both of these demands could best be fulfilled by territorially centralized states.

The other process concerns the increasingly capitalist economic development. In order to increase growth and wealth, capitalism posed certain requests, which were best met by the territorial states. These states could provide, e.g., military protection against external enemies, a common jurisdiction with legal regulation of property and a guarantee of the currency. Capital was looking for security and conditions of existence provided these states.

Mann, therefore, concludes that the state and its autonomous power is the primary condition of understanding societal development in Europe until today. By his redefinition of the state possessing autonomous power Mann makes it possible to take war and geopolitics seriously. According to Mann the societal development of the world can only be properly understood if we acknowledge that states have their own agenda, in particular, in relation to geopolitics, and thus warfare becomes an important transformative social force in history.

The war (and politics) is regarded as a decisive driving force, and it seems as if war and the struggle for recognition implicitly constitute the state.

Violence and War: Roots and Causes

Obviously, many people ask the question: why do people use violence? Why do states go to war? Since the origin of classical social thought, many attempts to answer questions like these have been made but no clear-cut answer has been provided. Several traditions and theories compete about providing the definitive answer. Thus we find a fierce competition between sociobiological and cultural explanations (Malesevic, 2010: pp. 50–85). Sociobiologists take their point of departure in the assumption that social behavior is rooted in biology and it is an outcome of evolution. In modern versions behavior is related to and determined by certain genetic principles. The core argument is: "human violence is just an extension of animal behavior, which includes aggressive competition over resources or territory with the aim of maximizing one's reproductive success" (Malesevic, 2010: p. 54). Unlike sociobiologists many sociologists and anthropologists stress cultural explanations of violence. War and violence are conceptualized as the product of culture: "while some approaches stress the irreconcilable struggles of different worldviews or theological doctrines as a source of violent action (…), others focus on symbolism, ritualism, and signification as the key features of warfare" (Malesevic, 2010: p. 64).

A competing set of theories emphasizing economic rationality also contributes to an understanding of the roots and causes of war. A simple but strong theoretical perspective can be found in a variety of rational choice models. Rational choice is based upon the assumption that individuals are (purposeful) rational and self-interested. Rational choice theory builds upon the ontology that only individuals and their action exist and social science must take its point of departure in the social actions of individuals. When examining causes of war many rational choice models provide an explanation related to the motivation of the actors involved. War is explained by the economic gains and losses of actors.

Also some of the voices in the globalization debate relate new wars and contemporary violent conflicts to economic globalization processes (Kaldor, 2001; Bauman, 1998). These scholars argue that warfare has become an instrument of economic policy. Thus the war in Iraq in 2003 could be seen as a struggle for material resources.

A central puzzle about war is that wars are costly but nonetheless wars recur. Thus scholars who have studied the origins of wars often have concluded that war can be a rational alternative for leaders who are acting in their states' interest. They find that the expected benefits of war sometimes outweigh the expected costs, however, unfortunate this may be. The two most typical rational choice explanations provided to this question are related to the problem of anarchy or the rational preventive war. The anarchy argument is based upon the anarchical nature of the international system. This is seen as an explanation for the persistence of war. Under anarchy, without a supranational sovereign to make and enforce law, "war occurs because there is nothing to prevent it…. Among states as among men there is no automatic adjustment of interests. In the absence of a supreme authority there is then the constant possibility that conflicts will be settled by force" (Waltz, 1959: p. 188). The argument is related to a fundamental difference between domestic and international politics. Within a well-functioning political order such as a state, organized violence as a strategy is ruled out by the potential reprisals of a central government. In international relations, by contrast, no agency exists that can credibly threaten reprisal for the use of force to settle disputes. The claim is that without such a credible threat, war will sometimes appear the best option for states that have conflicting interests.

The second explanation is related to the notion of preventive war. It is argued that if a declining power expects it might be attacked by a rising power in the future, then a preventive war in the present may be a rational decision.

Conclusion

It is suggested in this article that sociology of war takes as a point of departure that war is a constant in human life and this is an empirical observation. War is a particular conflictual form of social relations when groups (survival units) are struggling for recognition (Lindemann, 2010).

Obviously, we find a great variety of sociobiological explanations of causes of war and the same goes for cultural/sociological and economic explanations. Also theories combining or reconciling the different positions can be found.

This article, however, presents a perspective on the sociology of war, which is not engaged in the question of what causes war.

What triggers a particular war can be accredited to a number of factors, including aggression, demographic changes, economic necessity, or simply misinterpretation of a situation. In other words, a multivariety of causes can create war but it is not possible or within the scope of sociology of war to provide an explanation of the genesis of war and collective violence. It has another advantage, however, because unlike sociology in general this perspective takes war seriously and it provides an explanation for why war stops. This creates a potential for understanding and upholding peace.

See also: Military Geography; Military History; Military Sociology; Military, War, and Politics; War and Democracy; War and Nationalism.

Bibliography

Bauman, Zygmunt, 1998. Globalization. Polity Press, Cambridge.

Baur, Nina, Ernst, Stephanie, 2011. Towards a process-oriented methodology: modern social science research methods and Norbert Elias' figurational sociology. The Sociological Review 59 (2), 117–139.

Boserup, Anders, 1986. Staten, samfundet og krigen hos Clausewitz. In: Berg, Niels (Ed.), Clausewitz, Carl Von, Om Krig, vol. 3. Rhodos, Copenhagen, pp. 911–930.

Boserup, Anders, Neild, Robert, 1990. The Foundations of Defensive Defence. Palgrave Macmillan, London.

Bourdieu, Pierre, 1977. Outline of a Theory of Practice. Cambridge University Press, Cambridge.

Clausewitz, Carl v, 1986. Om Krig. Rhodos, Copenhagen.

Clausewitz, Carl v, 1993[1832]. On War. Everyman's Library, London.

Creighton, Colin, Shaw, Martin, 1987. The Sociology of War and Peace. Macmillan, London.

Crossley, Nick, 2011. Towards Relational Sociology. Routledge, London.

Donati, Pierpaolo, 2011. Relational Sociology. Routledge, London.

Downing, Brian M., 1992. The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe. Princeton University Press, Princeton, NJ.

Elias, Norbert, 1974. Towards a theory of communities. In: Bell, Colin, Newby, Howard (Eds.), The Sociology of Community: A Collection of Readings. Frank Cass & Co., London, pp. ix–xli.

Elias, Norbert, 1978. What is Sociology? Columbia University Press, New York.

Elias, Norbert, 1987a. Involvement and Detachment. Blackwell, Oxford.

Elias, Norbert, 1987b. The retreat of sociologists into the present. Theory, Culture & Society 4 (2–3), 223–248.

Elias, Norbert, 1987c. The changing balance of power between the sexes – a process-sociological study: the example of the ancient roman state. Theory, Culture & Society 4 (2–3), 287–316.

Elias, Norbert, 2000. The Civilization Process. Blackwell, Oxford.

Elias, Norbert, 2001. The Society of Individuals. Continuum, London.

Emirbayer, Mustafa, 1997. Manifesto for a relational sociology. The American Journal of Sociology 103 (2), 281–317.

Ertman, Thomas, 1997. Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe. Cambridge University Press, Cambridge, New York.

Freund, Julien, 1996. Warfare in the Modern World: A Short but Critical History. Plutarch Press, Washington, DC.

Gerth Hans, H., Wright Mills, C., 1985. From Max Weber: Essays in Sociology. Routledge and Kegan Paul, London.

Giddens, Anthony, 1985. Nation-State and Violence. Polity Press, Cambridge.

Hegel, Georg W.F., 1977[1807]. The Phenomenology of Spirit. Oxford University Press, Oxford.

Hegel, Georg W.F., 1991[1821]. Elements of the Philosophy of Rights. Cambridge University Press, Cambridge.

Hirst, Paul, 1990. Representative Democracy and its Limits. Polity Press, Cambridge.

Hirst, Paul, 2001. War and Power in the Twenty-first Century. Polity Press, Cambridge.

Hirst, Paul, 2005. Space and Power. Polity Press, Cambridge.

Hobbes, Thomas, 1962. Leviathan: Or the Matter, Form and Power of a Commonwealth Ecclesiasticall and Civil. E.P. Dutton, New York.

Joas, Hans, Knöbl, Wolfgang, 2013. War in Social Thought. Princeton University Press, Princeton and Oxford.

Kaldor, M., 2001. New and Old Wars: Organized Violence in a Global Era. Polity Press, Cambridge.

Lindemann, Thomas, 2010. Causes of War – The Struggle for Recognition. ECPR Press, Essex.

Malesevic, Sinisa, 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.

Mann, Michael, 1984. The autonomous power of the state: its origins, mechanism and results. Archives europennes de sociologie XXV, 185–213.

Mann, Michael, 1986. The Sources of Social Power. In: A History of Power from the Beginning to AD 1760, vol. 1. Cambridge University Press, Cambridge.

Mann, Michael, 1987. War and social theory: into battle with classes, nations and states. In: Creighton, Colin, Shaw, Martin (Eds.), The Sociology of War and Peace. Macmillan, London, pp. 3–32.

Mann, Michael, 1988. States, War and Capitalism. Blackwell, Oxford.

Mann, Michael, 1990. The Rise and Decline of the Nation-State. Blackwell, Oxford.

Mann, Michael, 2005. The Dark Side of Democracy: Explaining Ethnic Cleansing. Cambridge University Press, Cambridge.

Mann, Michael, 2013. The Sources of Social Power. In: Globalizations, vol. IV. Cambridge University Press, Cambridge.

Marx, Karl, 1958. The eighteenth brumaire of Louis Bonaparte. In: Karl Marx and Friedrich Engels, Selected Works in Two Volumes. Lawrence & Wishart, London.

Marx, Karl, Engels, Friedrich, 2002[1848]. The Communist Manifesto. Penguin, London.

Mulaj, Klejda, 2008. Politics of Ethnic Cleansing: Nation-State Building and Provision of In/Security in Twentieth-Century Balkans. Lexington Books, Washington, DC.

Poggi, Gianfranco, 1978. The Development of the Modern State: A Sociological Introduction. Stanford University Press, Palo Alto, CA.

Porter, Bruce, 1994. War and the Rise of the State: The Military Foundations of Modern Politics. Free Press, New York.

Powell, Christopher, Dépelteau, François, 2013. Conceptualizing Relational Sociology: Ontological and Theoretical Issues. Palgrave Macmillan, New York.

Ramet, Sabrina P., 2010. Central and Southeast European Politics since 1989. Cambridge University Press, Cambridge.

Skocpol, Theda, Evans, Peter B., Ruschmeyer, Dietrich (Eds.), 1985. Bringing the State Back in. Cambridge University Press, New York.

Schmitt, Carl, 1976. The Concept of the Political. Rutgers University Press, New Brunswick, NJ.

Scott, John P., 2000. Social Network Analysis: A Handbook, second ed. Sage Publications, Thousand Oaks, CA.

Skocpol, Theda, 1979. States and Social Revolutions: A Comparative Analysis of France, Russia and China. Cambridge University Press, Cambridge.

Tilly, Charles (Ed.), 1975. The Formation of National States in Western Europe. Princeton, NJ.

Tilly, Charles, 1992. Coercion, Capital, and European States, AD 990–1990. Blackwell, Oxford.

Waltz, Kenneth, 1959. Man, the State, and War: A Theoretical Analysis. Columbia University Press, New York.

Weber, Max, 1980. The national state and economic policy (Freiburg address). Economy and Society 9, 420–449.

Weber, Max, 1991. Gesammelte Politische Schriften. J.C.B. Mohr, Tübingen.

Weber, Max, Ross, Guenther, Wittich, Claus (Eds.), 1978. Economy and Society: An outline of interpretive sociology, 2 vols. University of California Press, Berkeley, CA.

Warfare in History

Michael S Neiberg, United States Army War College, Carlisle, PA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article divides the history of warfare into five periods: the Classical Era (including the wars of Persia, Greece, and Rome); the Middle Ages, including the Mongols and samurai (roughly from AD 500 to the Turkish conquest of Byzantium in 1453); the age of gunpowder (from 1453 to the American and French Revolutions beginning in 1776); the 'long nineteenth century' (from 1776 to 1918); and finally the contemporary period (covering World War II, the Cold War, and after). The article considers war in a variety of geographic settings including Europe, East Asia, and the Middle East. Each of these periods will discuss changes in three dimensions. First and most obviously, each period is associated with a general group of military technologies. This article does not argue for technological determinism, but it does review the major changes in military technologies and the ensuing changes technologies have forced on military organization and purposes. Second, each period also involves a shift in the nature of military organization, including leadership, training, and connections to civilian institutions. Finally, this periodization demonstrates changes in the reasons why societies have gone to war.

This article divides the history of warfare into five periods: the Classical Era (including the wars of Greece and Rome); the Middle Ages (roughly from AD 500 to the Turkish conquest of Byzantium in 1453); the age of gunpowder (from 1453 to the American and French Revolutions beginning in 1776); the 'long nineteenth century' (from 1776 to 1918); and finally the contemporary period (covering World War II and the Cold War). Each of these periods will be defined by changes in three dimensions. First and most obviously, each period is associated with a general group of military technologies. Second, each period also involves a shift in the nature of military organization. Finally, this periodization demonstrates changes in the reasons why people have gone to war.

The Classical Period

Technology

The Classical period deserves consideration on its own merits and in the context of a historical debate over the influence of Greco-Roman warfare on later European patterns. In the field of technology, some historians argue that the Greeks and Romans began a pattern of using superior technology to compensate for having to fight wars against numerically superior enemies. The Persian Wars of the fifth-century BC demonstrated the need for Greek city-states to combat the much larger armies of Persian kings Darius and Xerxes. Needing to slow the Persians, Spartan General Leonidas threw his much smaller force into a mountain pass near Thermopylae in 480 BC knowing that it would be destroyed to a man. Aware that his army would not return, Leonidas took with him only those men who had children to succeed them, an important consideration because the army and the citizenry were closely linked.

To avoid future Thermopylaes, the Greeks developed both offensive and defensive technologies. By the end of the period, the typical Greek soldier wore nearly 70 pounds of armor and carried spears and swords that were superior to Persian weapons. These technological solutions helped the Greeks win, but the unity that the Greek city-states showed against Persia soon devolved into internecine wars. During the most famous of these wars, the Peloponnesian War (431–404 BC), technology again proved critical. In this case, Athenian naval vessels attempted to balance the superior land power of their enemy, Sparta. Defensive technologies evolved in counterpoint as Athens, Syracuse, and other city-states built protective walls strong enough to defy even determined sieges.

The Romans also leaned heavily on technology to defeat more numerous enemies, though many Roman military technologies were not necessarily weapon systems. Building upon the Roman genius for engineering, Roman armies assisted in the construction of a system of roads (even today the Roman roads are roughly contiguous with the European highway system) that permitted the legions to move quickly from one point in the empire to another. The Romans also depended upon defensive works such as trenches and walls (Hadrian's Wall in the UK is the most famous example) to contain their enemies rather than fight them. Some historians argue that the Classical tradition bequeathed to later Western militaries the importance of looking for technological solutions to military problems.

Organization

The Greeks based their military upon the hoplite system. Hoplites (named for their shield, the hoplon) were citizens of the republic and rough social equals. Thus, Greek warfare had no concept directly analogous to the modern division of officers and enlisted men. Greek commanders received their appointment by yearly election. Municipal assemblies also voted on when and where the city-state would go to war. As citizens, hoplites participated fully in these elections.

As nonprofessionals, the hoplites needed a fairly simple tactical system because so little time could be devoted to formal training. The result was a dense concentration called a phalanx. A hoplite's shield protected his own left side and his neighbor's right side. Thus the phalanx could only succeed when every hoplite stood his ground, lest the gap between himself and his neighbor be exploited. Once broken apart, a phalanx soon dissolved, resulting in many casualties by trampling.

 http://dx.doi.org/10.1016/B978-0-08-097086-8.96033-1

The Romans trained and organized much more formally than the Greeks. Groups of about 120 men formed 'maniples,' 35 of which were then used to form legions of approximately 4000 infantry and 300 cavalry (the exact sizes of maniples and legions varied significantly over time). This system foreshadowed the modern creation of platoons, companies, and brigades. The Romans used professional gladiators (named for their short, thrusting sword, the gladius) and evolved a concept of officers and 'other ranks.' Increased sophistication and professional training allowed the Romans to develop a military that used various weapons, including soft iron javelins called pila. The soft iron bent on impact so that enemies could not throw them back. If victorious, however, Roman smiths could recover the pila after the battle and reshape them. The Roman system gave the legions flexibility, endurance, and striking power that few militaries could match.

The Greek and Roman systems of warfare mirrored their societies. Both preferred service by citizens, perhaps another legacy to modern nation-states. The Roman system extended citizenship to conquered people, but the effectiveness of its army noticeably declined as more non-Roman militia entered the Roman military. The citizen model was less prevalent in the armies of autocrats such as Alexander the Great. The Romans also gradually separated wealth from military service by introducing regular pay and even retirement benefits.

Purposes

Among the legacies left by the Greeks and Romans to later styles of warfare is the concept of pitched battle. The Greeks especially tended to favor violent, but short, battles intended to solve disputes in a single afternoon. The bloody nature of Greek warfare appalled many of their opponents, including the Persians, who disliked a style of warfare that left even the winners with high casualties. Despite its violence, Greek warfare did follow certain understood rules, including safe passage for heralds and ambassadors and amnesty for temples and other religious sites.

Roman warfare followed a similar pattern. Wars were generally fought with the single goal of annihilating the enemy. The Roman sack of Carthage in 146 BC is the most famous, but hardly the only example. Roman armies typically pursued retreating enemies with cavalry and often treated prisoners severely. In one case, Julius Caesar kept an opposing general on display in Rome for 6 years before executing him. The Roman preference for fast, deadly, decisive battles differed significantly from other cultures. The Aztecs, for example, believed that the purpose of warfare was to capture prisoners for later sacrifice.

The Middle Ages

Technology

The most important Medieval technology was not a weapon, but the stirrup, which inverted the traditional relationship between cavalry and infantry. With stirrups allowing riders to grasp lances and shields firmly and with newer, more durable horseshoes allowing horses to run faster and farther, cavalry came to the forefront for most of this period. The Mongols proved to be the most successful of the horse-based military systems of the Asian steppes; by the late thirteenth century they controlled an empire that stretched from China to modern-day Ukraine. Their military relied almost exclusively on cavalry.

In Western Europe, cavalry proved to be no less important. For the aristocrats who could afford the extensive support systems that accompanied cavalry, horses meant power. The European aristocracy, as well as its Japanese counterpart, depended upon horses, armor, and swords for offensive power and increasingly larger castles for defensive power. The victory of mounted Norman knights over Anglo-Saxon infantry at the Battle of Hastings in 1066 underscored the value of cavalry. The Catholic Church even tried to ban weapons such as the crossbow that threatened the knight's status, though the church permitted the use of such weapons on infidels.

Technological changes also helped to end the period of dominance by the mounted knight. The Hundred Years War marks an important turning point. At the battles of Crécy (1346), Poitiers (1356), and Agincourt (1415), skilled English archers devastated French knights with the Welsh longbow, a weapon capable of rapidly delivering arrows powerful enough to penetrate armor at ranges up to 400 yards. Lacking effective missile weapons themselves, the French knights suffered terribly. At Crécy, the longbows allowed the English to defeat a French force three times its size, further underscoring the Western pattern of looking to technology to compensate for smaller numbers.

Organization

Military training and military service was concentrated in a relatively small group of wealthy aristocrats. In most Eurasian systems, peasants and serfs were not expected to engage in actual fighting, though they might participate in ancillary operations, particularly as archers or pikemen supporting the cavalry. The challenge to kings and other rulers, then, was to maintain the loyalty and fealty of their aristocratic vassals. In the Third Crusade (1189–92), for example, Frederick I of Germany, Philip II of France, and Richard I of England spent considerable time keeping their unruly knights and barons in check. Powerful, aristocratic military orders, relatively independent of royal authority such as the Knights Templars and the Teutonic Knights emerged from the Crusades as well.

In Japan a similar system developed, though loyalty was ordinarily based more on personal than contractual allegiances. The Japanese samurai class emerged in the twelfth century, replacing the conscript armies that had existed previously. The samurai also came to dominate Japanese politics, supplanting the traditional Japanese aristocracy. The failure of Mongol invasions of Japan in 1274 and 1281 removed any serious outside threat to the home islands, leaving the small samurai class in near-complete charge of military affairs. Civil wars and power struggles did occur, but they rarely drew Japanese peasants into the fighting.

In the Islamic world, the Abbasid Caliphate began in the ninth century to use slave soldiers called mamluks. Originally from the steppes of Asia Minor, the mamluks soon translated their military dominance into political dominance in Egypt and elsewhere in the Middle East. They were the only major Islamic force that proved capable of resisting the Mongols, defeating them in 1260 at the Battle of Ain Jalut in modern-day

Israel. The mamluks, the samurai, the Mongols, and the European aristocracy, though different in many ways, shared a near-monopoly over military force in their societies, which were arranged in feudal, or at least fief-based, political systems. All of these systems were prone to succession crises and internal strife. None proved to be militarily superior to any other.

Purposes

Generally speaking, the motivations and purposes for warfare in this period were either religious, feudal, or both. The Islamic empires acquired tremendous territory, in part through military conquest, without significant technological advantages over their opponents. They did so by wedding religious ideology to their military forces. Under the reigns of Caliphs Umar (634–644) and Uthman (644–656), Muslim armies conquered Spain, North Africa, Persia, and the Indus River valley. Their fleets conquered Cyprus and raided Sicily and Rhodes. Qur'ans accompanied Muslim armies, converting men in the regions they conquered.

The Crusades (1095–1291) also reflect a religious basis for warfare in this period. In the eleventh century, the success of the Seljuk Turks against the Byzantines led Pope Urban II to call the First Crusade (1095–99) to defend Christianity against the incursions of Islam. The Crusades reflect the interplay of religion and feudal obligation. The Pope acted both as a religious leader and (more controversially) as a noble in his own right. The First Crusade demonstrated that only knights could be relied upon to form the center of Christian armies. Peasant armies were too large to feed and too unsophisticated to stand up in battle. Of the 30 000 nonnobles who fought in the First Crusade, 25 000 died. Sending young nobles off to the Middle East also served a political purpose for the Pope, removing warriors and weapons from the constant internal struggles of feudal Europe. Thus by the fifteenth century similar and roughly coequal military systems existed in Western Europe, the Islamic empires, the Asian steppes, and Japan.

The Gunpowder Age

Technology

The Turkish capture of Constantinople in 1453 involved 70 heavy artillery pieces, including a 19-ton piece that fired a 1500-pound projectile more than a mile. The appearance of these weapons was not new. Knowledge of gunpowder had moved from China to the Middle East to Europe. At the Battle of Crécy (1346) the English introduced primitive gunpowder weapons to European warfare. Nevertheless, the dramatic Turkish success at a place Europeans considered sacred marked a watershed in military technology. Ironically, the Islamic empires subsequently proved slower to adopt gunpowder weapons than did their European adversaries. By the end of the eighteenth century, Europeans were the clear masters of gunpowder warfare on land and at sea.

Early small arms, like the arquebus, were clumsy, unreliable, and often more dangerous to the shooter than to his intended victim. Until the widespread adoption of the rifle in the nineteenth century the common command given to musketeers was 'level' rather than 'aim' since one could not expect to hit a specific target at any significant distance. Large artillery was even more dangerous to their users and so large that pieces often had to be cast at the siege site. By the sixteenth century, Europeans had standardized ammunition sizes, developed training systems for musketeers, and invented more reliable firing systems for small arms. These changes made both large and small arms more mobile and effective.

Europeans developed gunpowder weapons more quickly than any other society in this period. The simple availability of raw materials and craftsmen were important prerequisites; men who could cast church bells were in special demand as they could easily cast cannon as well. European nobles also had ample motivation to acquire new weapons for their struggles with other nobles. Finally, the European system proved more amenable to allowing private enterprises to develop and sell new weapons. The mamluks and the samurai showed a particular dislike for guns as they tended to undermine the military monopoly that kept them in power.

Organization

Guns threatened the very existence of a horse-based nobility whose expertise was in swords and lances. The age of the knight, already waning, came to an end. Nonnoble musketeers replaced them. Because small arms were so inefficient when used singly, musketeers had to be arranged in lines and taught to fire and load their weapons in Unison. Armies thus became much larger and more disciplined. Aristocrats who remained in military service became the backbone of an officer class that led armies mostly constituted of peasants.

Holland's Maurice of Orange (1567–1625) and Sweden's Gustavus Adolphus (1594–1632) introduced regular drill and other reforms that eventually led to the classic European linear infantry formation. Initially as many as 10 men deep (later two or three deep) it permitted one line to fire, then move to the back. This system maintained a near-continuous rate of fire, but required intense training (the first drill manuals appeared in 1607) and discipline in order to assure the safe and effective operation of guns. Every officer, furthermore, feared the specter of mutiny. Intense discipline, therefore, characterized most European armies; in many cases, the men feared their officers more than they feared the enemy.

Many European nations also developed powerful sail-based navies in this period. England's Henry VIII, excommunicated from the Catholic Church in 1534, sold Church lands to finance the construction of a fleet. Future monarchs continued to build upon Britain's command of the seas, introducing men-of-war in the seventeenth century that were durable enough to carry large cannon. England's *HMS Victory*, launched in 1759, was built from 2000 oak trees, 27 miles of rope, and four acres of sails. It had 104 guns capable of firing 1100 pounds of iron every 90 s. Sea power became critical to acquiring colonies, protecting merchant marines, and projecting power. Britain dominated the 'Great Age of Sail,' but Portugal, Spain, Holland, France, and others used naval vessels to increase their wealth and status.

Purposes

Three purposes dominated warfare: religion; dynastic ambition; and imperialism. The Thirty Years' War (1618–48)

demonstrated the interplay of the first two. Lutheran–Catholic–Calvinist antagonisms partially explain the bloody nature of the war; Germany proportionately lost more people in the Thirty Years' War than in World War II. Religion, however, did not always determine the allegiances of the participants. Catholic France and Catholic Spain, for example, consistently fought one another for control of Italy and Flanders. Gunpowder weapons also led to a concentration of power as wealthy kings, in command of increasingly larger and more disciplined armies (not to mention artillery able to destroy castles), could compel their restless vassals to accept their rule.

The growth of monarchy meant that warfare often followed the dynastic goals of kings. Louis XIV in France, Peter the Great in Russia, and Frederick the Great in Prussia all used gunpowder-based armies to extend their power inside their kingdoms and enforce their will outside. Often, dynastic wars spilled over into the empires Europeans were in the process of building with those same gunpowder weapons. In the Seven Years' War (1756–63), Britain seized France's Canadian and Indian colonies.

Finally, of course, Europeans translated their gunpowder weapon advantages into overseas empires. In this period, their greatest military advantage over non-Europeans was in areas accessible to their great warships. Navies allowed Europeans to enforce their will along the coastlines of Africa and Asia. On land, guns helped to make the Spanish, French, and British empires in the Americas possible. In Asia, 'gunpowder empires' such as the Ottomans and the Mughals also concentrated political power. Japan, after an initial period of experimentation with guns, effectively banned gunpowder weapons in 1587 because of the threat they posed to samurai dominance.

The Long Nineteenth Century, 1776–1918

Technology

This period, dominated by nationalism and industrialization, marked a tremendous advance in military technologies. Many existing technologies became much more sophisticated. Smoothbore muskets gave way in the nineteenth century to rifles, which have a groove cut in their barrels. Rifling and the conical bullet vastly increased the range and accuracy of small arms fire. After the American Civil War (1861–65) various forms of repeating rifles and machine guns emerged. Artillery also became more deadly. By 1918 the Germans had developed a gun (albeit an inaccurate one) with a range of 80 miles. Naval vessels became more deadly as well. Iron and steel replaced wood as the primary construction material and oil replaced coal as the main fuel. *HMS Dreadnought*, launched in 1906, revolutionized warfare by carrying ten 12-inch guns. *Dreadnought* set off a naval arms race that helped to cause World War I. Winston Churchill noted the importance of Dreadnought-class ships to his nation thus: "the Admiralty asked for six, the Cabinet proposed four, and we compromised on eight."

This period also introduced many new types of weapon systems, including (in rough chronological order) submarines, airplanes, poison gas, and tanks. By the end of this period air forces had developed all major roles for airpower in use today except midair refueling. These technologies brought war directly to civilians as never before. Many important communication technologies such as telegraphs (first used in the Crimean War, 1854–56), railroads, and radio changed the nature of military command and control. Advances in medical technology reduced the number of men who died of wounds and disease.

These technological changes affected Western and Westernized societies (including the USA and, after 1853, Japan) the most. Warfare between them, therefore, reached unprecedented levels of violence. Westernized societies also had tremendous advantages against non-Westernized societies. European imperialists were thus able to use new military technologies to move inland in Africa and Asia. Most visibly, Japan, which Westernized very quickly after 1853, easily defeated China, which did not, in the Sino-Japanese War of 1894–95.

Organization

The most obvious change in the organization of armies and navies in this period was the increasing connection between military service and nationalism. National volunteers and conscripts replaced mercenaries, once common in Europe. Between 1776 and 1783, the American colonies used a militia with both local and national roots to defeat a British army partly dependent upon mercenaries. This contrast partially explains the fame of George Washington's crossing of the Delaware River in 1776: his national army captured 1000 Hessian mercenaries in British service.

The American experience, though, is minor compared to the French Revolution. The logic of the Revolution and the wars that followed implied that national armies must be populated by national soldiers. The *levée en masse* of 1793 established the principle that citizens of all ages, male and female, bore some responsibility for national defense. Napoleon extended this principle and most European armies soon copied the French pattern. Nationalism inspired zeal and a willingness to endure. In 1814 Prussia created the *Landwehr* as a permanent peacetime militia and by the end of the nineteenth century most European nations had introduced national conscription. The militaries of states that were not also nations (Austria–Hungary, the Ottoman Empire) declined as a result.

The two most important military reforms of this period were Napoleon's corps system and the Prussian General Staff. A Napoleonic corps was large enough to operate independently since it contained formations of cavalry, artillery, and infantry that could act in concert. This system, however, depended on capable officers. Napoleon therefore introduced a professional officer corps open to talent in place of the aristocratic one upon which Europe had come to depend. Later in the century, Prussia developed a General Staff system to plan wars. Detailed Prussian planning led to victory in the Wars of German Unification (1864 against Denmark, 1866 against Austria, and 1870–71 against France). These successes led all major Western militaries to copy the general staff idea.

Purposes

Military historians sometimes speak of a continuum between 'limited' and 'total' wars. Because warfare in this period tended to be tied to national goals and national identity, this period represented a general shift toward totality. The Wars of the French Revolution, the American Civil War, and World War I

stand as cases in point. In these wars and others, huge proportions of national income were dedicated to prosecuting wars and fundamental social, political, economic, and cultural patterns were forever altered. The increased suffering (and intentional targeting) of civilians also led to totality.

Advanced technologies also made imperialism (often as an extension of nationalism) cheaper and easier for the Europeans. European militaries moved inland with steam-powered ships and railroads. Machine guns, furthermore, made military contests with natives one-sided. By 1912, the Vickers–Maxim gun could fire 250 rounds per minute. As poet Halaire Belloc noted in a turn-of-the-century (twentieth century) poem, "Whatever happens, we have got/The Maxim Gun, and they have not." Nevertheless, isolated cases such as the Boer Wars (1899–1902) demonstrated that modern weapons in the hands of skilled practitioners could make imperialism extremely bloody and costly. The Boers, white settlers in South Africa of Dutch ancestry, resisted the British with breech-loading guns and smokeless powder. It took the British empire more than two and a half years and almost 500 000 men to subdue less than 90 000 loosely organized Boers. The Boer Wars provided a taste of the changes to come in European military dominance after World War II.

World War I caused 20 000 000 deaths, destroyed four empires, and left in place the conditions for World War II. Only wars fought by national armies could have been so destructive. Warfare in this period combined the productive capacity of industrialization with the inspiration of nationalism. This combination, when matched with the classical legacy for pitched battle, produced a level of carnage that continues to baffle historians today.

The Contemporary Period, 1918–2000

Technology

Although this period has much in common with the previous period, one specific technology divides it from all that had come before. Atomic weapons have changed the very definition of warfare. Some scholars have argued that in the atomic age 'strategy,' as the term was commonly understood before 1945, has no real meaning because atomic weapons do not permit a society to apply a level of force proportionate to desired political ends. In effect, a society only has the military option of obliterating one's enemy and, in the process, taking the risk that a similar fate might befall your own society. Atomic weapons created the post-World War II superpower system that informed not only the Cold War but dozens of regional conflicts as well. The challenge of fighting wars under the nuclear umbrella led to American defeat in Vietnam, Soviet defeat in Afghanistan, and war by proxy as the superpowers tried to fight each other through client states in areas as diverse as Southeast Asia, the Middle East, sub-Saharan Africa, and Latin America.

Of course, more than just atomic weapons are important to this period. World War II saw the development of effective long-range bombers and aircraft carriers, capable of delivering weapons over a greater distance. Radar and sonar provided new technologies to deal with enemy air and submarine attacks. By the end of the war, jet aircraft and rockets were operational, though neither proved decisive. More sophisticated transport planes and landing craft allowed for airborne and amphibious warfare to come into maturity as well.

Since World War II, electronics and computers have revolutionized warfare yet again. Cruise missiles and other 'smart' weapons ideally provide more accurate delivery of weapons and, hopefully, fewer civilian casualties as a result. Satellites, helicopters, sophisticated avionics, networked information systems, and stealth technology once again have provided an upper hand to advanced societies, though the Vietnam and Afghanistan cases prove that a technological advantage is not always enough to ensure victory. Nevertheless, today Western societies still depend upon their technological advantages – much as the ancient Greeks did – to compensate for the smaller numbers of soldiers that they can place on the battlefield.

Organization

Initially, the organization of World War II armies was designed to avoid the totality of World War I. The German *blitzkrieg* system attempted to win 'lightning' victories that would not drain German manpower in the trenches. The Japanese, too, hoped to destroy the American fleet at Pearl Harbor and force the USA into a negotiated settlement. Nevertheless, totality soon emerged as the Axis powers sought to annihilate, not defeat, their enemies. More than ever before, entire societies became mobilized. Women participated directly and indirectly in numbers and ways never before envisioned. The pattern of increasing female participation has continued into the postwar period. The USA armed forces used 35 000 women in the First Persian Gulf War, many of them at or very near the front lines.

The totality of World War II led to the end of Japan and the traditional powers of Europe as independent military actors. The end of European hegemony also meant the virtual end of imperialism. Many militaries resorted to guerrilla warfare to try to defeat more technologically sophisticated enemies. Fighting underneath the technology threshold has proven difficult for more advanced military systems. The Americans had great difficulty with the Vietnamese but few battlefield problems with the Iraqis in 1991 because the latter's military was much more like their own that the former's.

Since the end of the Cold War, the general trend has been to much smaller, but more sophisticated, militaries. Virtually all European states have eliminated or greatly reduced conscription. The USA abandoned conscription during the Cold War (in 1973) partially as a result of Vietnam War protest, but also to acknowledge that sophisticated weapon systems require dedicated professional volunteers to operate them.

Purposes

World War II was a contest of three ideologies: democratic capitalism, fascism, and communism. As a result, armies became deeply politicized. Young German males often left Hitler Youth organizations and entered directly into the Wehrmacht. In the Soviet Union, political commissars served with Soviet military units and had as much decision-making authority as generals. Ideology created what John Dower called a 'war without mercy.' Though he was referring to the

Pacific War between the USA and Japan, the same phrase could be applied equally to the war between Germany and the Soviet Union as well. The Holocaust and the German siege of Stalingrad, among other episodes, show the cost of the war to civilians. The former killed 6 000 000 Jews and other targets of the Nazi state. In the latter, Joseph Stalin refused to order the evacuation of the city even in its most dire moments. In the end, only nine Stalingrad children were reunited with both biological parents.

The Cold War interacted with the general decline of European hegemony to impact regional conflicts. Throughout the Third World, nationalist movements and civil wars became arenas where the superpowers armed clients but, conversely, often went to great lengths to reduce global tensions. In 1956 and 1973, for example, the superpowers brokered truces in the Arab–Israeli wars rather than risk a USA–Soviet confrontation. Regional wars like the Iran–Iraq War (1980–88) also became entangled with the Cold War; the weapons that the superpowers made available lengthened wars that had little relation to the global capitalist–communist struggle.

The end of the Cold War, Israel's 2006 war in Lebanon, the Russian wars in Chechnya, and the American-led wars in Iraq and Afghanistan since 2001 have led some commentators to argue that a new era of warfare is upon us. Led by thinkers like Rupert Smith, they contend that postmodern war will increasingly be fought between states and nonstate actors. They argue that with growing economic interdependence wars between major states is unlikely. Future wars, they contend, will require an entirely new approach to defense and security than that posed by traditional state on state wars. They will be characterized by the decline of the state in a globalized world, increasing involvement of civilians in urban warfare, and the emergence of so-called hybrid threats from groups like Hezbollah or Al Qaeda, which are backed by a state but do not have the standard characteristics of a state. If indeed such thinkers are correct, then we are on the cusp of a new revolution in military affairs.

See also: Arms Control; Conflict and War: Anthropological Aspects; Contemporary History; First World War, The; Folklore; Geopolitics; Home, Anthropology of; Imperialism, History of; Military Geography; Military History; National Security Studies and War Potential of Nations; Peace-Making in History; Peace; Revolutions, History of; Sacrifice; Second World War, The; Violence, History of.

Bibliography

Howard, M., 1976. War in European History. Oxford University Press, Oxford, UK.
Keeley, L.H., 1996. War before Civilization: The Myth of the Peaceful Savage. Oxford University Press, Oxford, UK.
Keen, M. (Ed.), 1999. Medieval Warfare. Oxford University Press, Oxford, UK.
McNeill, W.H., 1983. The Pursuit of Power: Technology, Armed Force and Society since 1000. Oxford University Press, Oxford, UK.
Overy, R., 1995. Why the Allies Won. Norton, New York.
Paret, P. (Ed.), 1986. Makers of Modern Strategy from Machiavelli to the Nuclear Age. Princeton University Press, Princeton, NJ.
Parker, G. (Ed.), 1997. The Thirty Years War. Routledge, New York.
Smith, R., 2007. The Utility of Force: The Art of War in the Modern World. Knopf, New York.
Strachan, H. (Ed.), 1998. World War I, A History. Oxford University Press, Oxford, UK.
Strassler, Robert (Ed.), 1996. The Landmark Thucydides: A Comprehensive Guide to the Peloponnesian War. Free Press, New York.

Wars among Nation-States: Patterns and Causes

Christian Olsson, Université Libre de Bruxelles (ULB), Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by J.S. Levy, volume 24, pp. 16354–16357, © 2001, Elsevier Ltd.

Abstract

This article tries to reframe the question of the conceptual unity and sociological diversity of interstate war while highlighting patterns and causes. The challenge is indeed to be able to account for transformation and heterogeneity without falling into the trap of an all-encompassing definition that would make all mutations a priori consistent with its premises. This will be done by highlighting how the question of the state, war, and the 'interstate' sheds light on the contemporary patterns of interstate war (first section). It is only subsequently that this article can raise the question of the causal pathways and motivating factors associated with past and present interstate wars (second section).

Two dominant ideas pervade contemporary academia concerning interstate war. On the one hand, it is increasingly popular to proclaim its programmed demise (Mueller, 2004). This would be confirmed by the relatively rare occurrences of such wars since the end of Cold War. On the other hand, many traditional international relations (henceforth IR) scholars continue to analyze interstate wars as a permanent and immutable feature of international politics since the Peloponnesian War.

While these two assertions contend for the ability to foretell the future, they unwillingly lend credence to one another. The permanent reinvention of war between states in spite of the abovementioned irenic predictions seems to vindicate the traditional approach. Conversely, because interstate wars are defined by the latter in an asociological and ahistorical way, usually implicitly modeled on past wars, their approaching end can be proclaimed. As opposed to the abovementioned approaches, we will argue for the heterogeneous nature of the phenomenon and hence for an approach that focuses on transformations and diversity. There are good reasons to make such a move. Three interdependent elements, partially linked to contemporary developments, are worth pointing at.

First, contemporary thinking on interstate wars tends to be trapped in a particular imaginary of these wars, linked to a specific time and place, and usually referring to the 'institutionalized wars' of the European eighteenth century (and preferably here the so-called 'great power wars,' i.e., wars seeing the armed forces of two or more 'great powers' fighting each other directly) or the 'total wars' of the twentieth century (Holsti, 1996). It is consequently difficult to see whether the idea of the end of interstate war derives from what interstate war is explicitly said to be by formal definitions or from what it is implicitly expected to look like.

Secondly, the distinction between interstate war and other types of war is increasingly difficult to draw unequivocally. Many wars have an interstate component, but few if any are 'purely interstate.' This might have been the case since the advent of the modern state to some extent. Arguably, it is however increasingly difficult to disentangle 'interstate' from 'intrastate' wars. Already during the Cold War, many civil wars that erupted in the Third World were inseparable from the opposition between the US and the USSR, thus making the idea of bipolar stability dubious at best. Today, many armed conflicts involve government support to insurgencies in third countries, interlocking conflict systems in which two or more governments fight each other directly and indirectly by supporting each other's domestic insurgencies, governments contending by proxy in the context of civil wars in third countries or in mutually contested territories, secessionist conflicts that are seen as civil wars by some and as interstate wars by others, the role of partially recognized states, etc. This does not only beg the question of the external borders of the concept of interstate war. It also raises the issue of the specificity and historicity of interstate war compared to other forms of war: 'nonstate war,' 'intrastate war' (previously 'civil war'), 'extrastate war' (previously 'extrasystemic war') following the current categories of the Correlates of War (COW) data set.

Thirdly, interstate wars are usually denied as wars by governments preferring to refer them as 'intervention,' 'coercive diplomacy,' or 'use of force.' Moreover, the enemy often is not recognized as a government (rather considered an illegitimate regime) or the relation between the designated enemy and the sovereign state on the territory of which it thrives is intentionally blurred as was the case during the Israeli–Lebanon War of 2006. The contemporary challenge is hence also to emancipate academic language from governmental discourse, while simultaneously recognizing that the latter is part of the social reality that academia tries to pin down.

Against this backdrop, this article tries to reframe the question of the conceptual unity and sociological diversity of interstate war while highlighting patterns and causes. The challenge is indeed to be able to account for transformation and heterogeneity without falling into the trap of an all-encompassing definition that would make all mutations a priori consistent with its premises. This will done by highlighting how the question of the state, war, and the 'interstate' sheds light on the contemporary patterns of interstate war (first section). It is only subsequently that this article can raise the question of the causal pathways and motivating factors associated with past and present interstate wars (second section).

Interstate War: From Concept to Patterns

It is difficult to deal with war, the state, and the 'interstate system' separately in order to construct the category of

interstate war: war as we know and represent it is deeply embedded in the historical sociology of the state and it is difficult to think of latter independently of its ability to raise armies and engage them in war against other 'members' of the 'interstate system.' For the sake of clarity, we will however try to disentangle these dimensions in the three next sections, while simultaneously attempting to show how they interact.

The State at War

The concept of the state is often used in at least three conceptually distinct, but theoretically interlinked ways in relation to war: as an organizational-institutional entity; as a sociohistorical configuration; and as a legal entity. Although the definitions flowing from these three approaches are partly overlapping, they do not do so perfectly. Let us see successively these three approaches to the state.

First, when approached as an institutional-organizational concept, the state is often equated with the government and seen as a more or less unitary actor. We hence say the "state has decided…" when the notion of government would be more precise. When identifying the state with the government, one is in fact positing that the state is an institutionalized web of power relations that, at least in international relations, would function as a compact unit headed by the government and differentiated from society. With regards to war powers, the state is then often seen as a political command structure allowing its government to unilaterally impose its decisions on a bureaucracy and military chain of command while avoiding personal rule and societal interference. Whereas political sociologist J.P. Nettl bought into this narrative in 1968, he simultaneously recognized that one has to think of the state as a conceptual variable with different levels of 'stateness' from country to country in domestic affairs (Nettl, 1968). He however thought states to be exempt from such variability in their external relations. His domestic/ international distinction can however not be taken for granted. Not only can different Western states not be considered as invariants from the point of view of interstate war, but also the spread of the modern state beyond the West after decolonization introduced further variation in its empirical forms. As a result, interstate war has become an even more heterogeneous phenomenon varying with the degree of institutionalization, societal interference, or civil–military relations of its protagonists. Hence, while the state as institution was traditionally seen, in war, to imply the allegiance of the population to the government and the former's exclusion from a participation unmediated by the latter, this supposition is increasingly dubious. It is equally noteworthy that most post-1990 interstate wars between neighbors have involved states often described as 'weak' or 'fragile': the Eritrean–Ethiopian War (1998–2000); the South Sudan–Sudan border war (2012); to a lesser extent the Cenepa War between Ecuador and Peru (1995); the Kargil War between India and Pakistan (1999), etc. This highlights the extent to which contemporary interstate wars often occur between states that do not comply with normative definitions of what they are supposed to be.

Secondly, as a sociohistorical configuration, the state is usually defined, following Weber, as a human community or 'institutional association of rule' the governing staff of which successfully claims a monopoly on legitimate violence on a given territory. In this sense, the criterion of statehood lies in the type of organization of violence and social configuration that the government is seen to uphold. Weber's definition is however flexible: it does not involve a monopoly on violence, but the successful claim (and there can be varying degrees of success) to a monopoly on specifically legitimate violence. However, when the infighting between rival armed groups for the violent control over a territory is endemic, durable, and enjoys varying degrees of support, one probably is beyond the reach of this flexibility. It is hence dubious whether there was a Congolese state in this sociohistorical sense during the Congo War II (1998–2003) during which its government was, among others, fighting at least indirectly against other governments (Rwanda, Uganda, Burundi) through diverse proxies.

In Weberian historical sociology, the state is seen as being relatively pacified domestically and generally war-prone externally: the occurrence of civil violence and the capacity of states to wage interstate war are seen as inversely proportional. Indeed, the advent of the state in this sense is linked to the ability of increasingly centralized and bureaucratized power structures to impose a pacified order inside, thus becoming more capable of waging war outside: wars become, at least potentially (but often also actually), more intense and lethal. Interestingly, however, interstate wars become simultaneously less frequent. There is a long-term decline in the number interstate wars since the beginning of eighteenth century and since 1945 they have disappeared from the relations between Western states. The ambiguity between state formation as a process of domestic pacification and as potentially pushing war toward its extremes, both occurring through the increased centralization and bureaucratization of coercion within the state, indeed affects interstate relations as well. States manage to impose order, not only internally, but also on their relations as well, including a stricter separation between war and peace and a lesser chance of nonmilitarized disputes degenerating into full-scale interstate war (Mueller, 2004: 30). It is therefore only an apparent paradox that many of the contemporary interstate wars, rather than being solely triggered by the increase in the capabilities of their protagonists, are in fact linked to the inability, at least on the part of one of them, to control all sustained organized violence on their territory or emanating from it. The US-led invasion of Afghanistan and the Russia–Georgia War are cases in point.

The post-1945 world sheds light on the sociological-historical definition of the state in interesting ways. On the one hand, the outlawing of war (save in exceptional circumstances) by the UN charter further bureaucratizes interstate warfare by submitting it to rational-legal procedures at domestic and multilateral levels (Rasmussen, 2007). On the other hand, there is an increased integration and internationalization of military affairs within organizations such as NATO and other "international 'conglomerates' of state-power" (Shaw, 2002: 89). This effectively limits the concerned governments' autonomy of decision and ability to wage war on their own. It raises the question whether the member states of these 'conglomerates,' in particular in the West, have not in fact pooled their 'monopolies of violence.' In this sense, there are good reasons to believe that dynamics has linked to the internationalization of military organizations account for the

absence of interstate wars between Western states since 1945 at least as much as economic interdependencies or common identities (from which they are inseparable).

Thirdly, from an international law perspective, the state as a legal entity supposes the combination of an effective legal government, a population and a territory. Interstate war, as generally understood, however, implies war against governmental forces and hence a recognition of government as distinct from recognition of statehood. Consequently, interstate war increasingly is denied in political discourse by denying the quality of governmental forces (often by using the notion of regime) to the enemy. For example, Bush Jr. implied that the US was not at war with the Afghan state in 2001 since it was merely trying to destroy its purportedly illegitimate 'regime,' while fighting alongside president in exile Rabanni of the Northern alliance. A similar argument was used during the NATO war in Libya. Typically, political justifications of such denials would contest the legality and/or effectiveness of what might previously have been seen as the government or what others might see as the government. What is at the crux of the matter is that, as previously seen, the UN charter is understood to have outlawed interstate war (save in exceptional circumstances) and the type of objectives that these wars historically have been associated with: territorial expansion. The resulting negative connotation of interstate war implies that it has to be denied when waged.

As shown by what precedes, the meaning of "interstate war" is not devoid of ambiguities, amongst others because the concept of the state is itself polysemic. It therefore needs to be specified, shall it be used rigorously. We will here approach interstate war as a form of organized violence between governments claiming a monopoly on legitimate violence, usually (but not always) with success, on a given territory. We will use the concept of 'stateness' to account for different levels of institutionalization and monopolization of the state thus defined.

War as Organized Violence

According to Hedley Bull, war is "*organized violence carried on by political units against each other*" (Bull, 1977: 184). War, thus defined, predates the advent of the modern state and not all wars involve states. The central aspect we want to emphasize is the one of organized violence. Indeed, it has at least three dimensions: war involves the organization of practices of physical violence, war entails the organization and institutionalization of reciprocal violence, and in war a specific organization resorts to violence and organizes society for this purpose.

On the first aspect, war involves a collective maintained by a more or less hierarchical structure. This organization constrains a group to stand and fight rather than breaking up, fleeing, or engaging in interpersonal violence. The state involves a high level of functional specialization, bureaucratization, and monopolization of violence, and hence a particularly high organizational capacity to extract resources (manpower, money, etc.) both for its armed forces and for their administration. As a consequence, there is a tendency to implicitly assume that interstate war is the 'purest,' because the most organized, form of war. This representation is questionable, if not for any reason because civil wars can be just as organized and 'conventional' as interstate war as shown by the American or Spanish civil wars. In historical sociology, it is however often assumed, in line with this representation, that following Charles Tilly's often-quoted phrase "*wars made the state, and the state made war*" (Tilly, 1975: 42), which is sometimes wrongly assumed to mean that all wars make states and that all states were created by war. As the description of a process that took place at a specific time and place, this description is certainly correct: the territorial wars of early modern Europe and the "elimination struggles" (Elias, 2000) they entailed, created an impetus for sovereigns to increase the extraction of resources and institutional capacity. The resulting bureaucratization gave rise to increasingly similar states. Here war-making and state-making were dynamically interlinked and these wars hence became more and more 'interstate' in nature.

When taken out of history, simplified and transformed into a universal law, this statement becomes highly problematic: war can destroy states, not all wars build states and not all states were built through war. The state as a sociohistoric configuration, however, as opposed to 'all particular states,' emerged through the wars fought at a specific time and place. In this respect, contemporary interstate wars are very different. When engaging Western troops in overseas military operations, they are too asymmetric and the density of the interactions between the 'units' involved too diffuse for the notion of competitive struggles to make sense. When engaging neighboring countries or regional conflict systems, their mode of financing and resourcing (often external rather than internal), and the prior definition of 'national territories' by the principle of intangibility of international borders transforms the process significantly: states can be significantly weakened without disappearing (contra the above-mentioned idea of "elimination struggles"). Organized interstate violence cannot any longer be seen as the central mechanism through which states are consolidated.

The second aspect concerns the organization of the violent encounter between armed forces. War usually involves some degree of agreement between the opposing military forces on rules, norms, and principles. Because the early interstate wars in Europe involved a high degree of interaction between 'units,' because no state managed to become stronger than all the other states combined and because a similar bureaucratic rationality came to define interstate wars, common rules were emerged. Interstate war became increasingly institutionalized between the seventeenth and nineteenth century as a central institution of what was thought of as a shared interstate system. It is still often through this institution that wars are conceived of. As a more or less coherent, legitimized, and commonly accepted international institution, interstate war however has been eroded as previously seen. As traditionally codified in the law of nations, it has lost its unity.

The third aspect is that governments use violence and organize society for this purpose. In this regard, while the increasingly total wars from the French Revolutionary Wars to the World War II violated many of the rules of the 'institutionalized wars' of the previous period (distinction between civilian and combatants, limited aims, etc.) (Holsti, 1996), they gave a very concrete meaning to the state. Not only did it

involve a territorial monopoly on legitimate violence, but also a total mobilization of the population and the economy. In this respect, the post-1945 period, arguably because of the horrors of the World War II and the invention of the nuclear weapon, has marked a significant shift. The Cold War saw the Western bloc experimenting with the concept of 'limited war,' wars that would remain below the nuclear threshold and avoid drawing in the USSR. The post-Cold War world saw the advent of a more radical idea: that war could be a "*spatially and temporally discrete affair without fundamental ramifications for society and politics in the West*" (Shaw, 2005: 2). Today interstate warfare, as waged by Western states, hence mainly remains a strictly governmental affair, minimizing risks for the Western military, economy, population, and for the government itself. The lesser salience of interstate war in Western societies today is at least partially linked to the fact that it has ceased to organize social life when it occurs.

Interstateness as Reciprocity and Similarity

War in principle involves reciprocal acts and a strategic interaction. In strategic thinking, it is understood to imply a same rationality, a common game. The 'interstate' nature of war potentially adds a further level of reciprocity. In principle, states share a same type of institutional setup and social identity, which historically makes the concepts of rivalry and mutual recognition a central feature of interstate war. In the standard realist and neorealist theory in IR interstate war is accordingly described as occurring between 'like units.' Three questions might be raised in this regard.

First, can the assumption of wars between 'like units' be maintained beyond a certain level of difference in sociological constitution and military capabilities between states? Since states are deemed to be 'like units,' interstate wars are usually defined by contradistinction to asymmetric conflicts: wars between 'units' between which there is a considerable qualitative difference in nature. It is however generally assumed that quantitative differences are transmuted into qualitative ones beyond a certain threshold: if differences in capabilities are overwhelming, the weaker part might cease to act 'conventionally,' as a 'like unit.' From this point of view, it is interesting to note that since 1945 there have been no so-called 'great power wars.' This might be explained by deterrent effect of nuclear weapons and/or by the increased integration of the armed forces of the Western 'great powers.' Consequently, contemporary interstate wars involving so-called 'great powers' have been highly unequal from the point of view of capabilities: the Russia–Georgia War, the Gulf Wars, etc. In the case of the 1999 NATO war against Yugoslavia and the NATO–Libya war the level of asymmetry even implied that there was very little in terms of reciprocal violence. This does not formally make these armed conflict less of interstate wars. It however makes them less different from wars considered as 'extrastate wars.'

Secondly, do the governments have to recognize each other's states as like units? From a social sciences perspective, this question could seem irrelevant. The fact that Egypt and Syria did not recognize Israeli statehood has not prevented the Ramadan War (1973) from being seen as an interstate war. The question is however interesting since a war between a government and a self-proclaimed independent state enjoying *de facto* territorial control might have much in common with an interstate war. It hence raises questions regarding the specificity of interstate war. Moreover, since recognition both implies common rules and spurs rivalry, it is a crucial dimension of the traditional view of interstate wars.

Thirdly, do state authorities have to retain formal control of the combat forces or are proxy wars involving nonstate armed groups also potentially 'interstate' provided these groups are informally controlled (but to what extent?) and supplied by governments? Traditionally, interstate wars are seen as involving the direct confrontation between official armed forces bearing their weapons openly and having distinctive signs. Since 1945 at least, there is however a continuum between so-called 'nonstate wars' and 'intrastate wars' on the one hand, 'interstate' wars on the other hand, with governments fighting each other through nongovernmental proxies, while sometimes simultaneously fighting each other directly through their official security forces. For example, in the case of the Nagorno-Karabakh War (1988–1994) there were government-supported volunteers as well as government forces on both sides. The same can be said of the War of Transnistria (1990–92). Here again the interstate nature of many contemporary wars seems to be a question of degree.

On the whole, it would be useful to think of 'interstateness' when approaching contemporary wars. Indeed, very often it is more a question of degree than of a strict distinction between interstate and other forms of war. Of course, it is widely recognized in social sciences that interstate wars vary with the number of states involved (dyadic, triadic, multiparty, 'complex,' systemic wars, etc.), with the objectives pursued (territorial wars, resource wars, wars of rivalry, etc.), with their purported intensity (limited wars vs total wars)… There has however been a tendency to assume that, in the context of these wars, the state is an invariable unit within an immutable interstate system, an assumption that is increasingly challenged.

Pathways to War: Causality and Interstate War

The preceding elements highlight the problematic nature of some of the assumptions in the IR literature on the 'causes of war.' Many correlational studies rest on the idea that the compilation of huge data sets can uncover transhistorical laws since interstate war, and indeed the state itself, essentially would remain the same throughout space and time. Other more mechanism-based approaches have not avoided the same pitfall. While most academics consider that interstate and intrastate wars are different enough for it to justify distinct data sets and theories, we have tried to show that comparable differences (increasingly) affect interstate wars.

Explanatory Strategies and Interstate War

Beyond the question of definition, the heterogeneity of interstate war raises at least two partially interlinked problems in relation to causality, the one of potentially varying combinations of causal mechanisms, and the one of context-dependency. Both are tackled by Charles Tilly in his later

works on explanatory strategies (Tilly and Goodin, 2006). Tilly highlights that the 'processes' likely to trigger a social outcome, in our case interstate war, are usually the result of multiple 'causal mechanisms' that concatenate in different ways depending on context. He hence calls for explanatory modesty and contextualization rather than law-seeking.

In relation to interstate war, one such crucial causal mechanism is arguably the one that links the state's increased (respectively decreased) monopolization of the means of legitimate violence on a given territory to the strengthening (respectively weakening) of its territorial control and extractive capacity. This first causal mechanism is not directly related to interstate war but might become so when combining with other mechanisms. On the one hand, the increased capacity to monopolize violence on a given territory allows turning violence outwards in a more efficient way, thus temporarily enhancing the opportunities of waging war successfully. One process associated with the occurrence of war will then combine the first mechanism with a second mechanism linked to government rationalization of strategic opportunities. The Franco–Prussian War of 1870 might here offer an example. On the other hand, the loss of the monopoly on the use of legitimate violence, or fear thereof, creates a 'domestic security dilemma': the attempt on the part of the governing staff to protect itself is seen as threatening by its political rivals (and vice versa), resulting in domestic turmoil (Ayoob, 1995). This internal unrest, and the associated loss of governmental control over violence occurring at its borders, can spiral out of control, into interstate conflict. The governing staff might under such conditions of civil strife also use interstate war as a means to crack down on the designated 'domestic enemies,' to consolidate the government's grip on the population and to externalize domestic tensions by blaming external enemies (theory of diversionary wars). The Iran–Iraq War (1980–88) is a case in point. Here we have a very different second 'process' or pathway giving prominence to a third causal mechanism ("civil violence" spiralling out of control, into interstate war). The generic processes of 'monopolization of violence' is hence as such an indeterminate, and yet a crucial, causal mechanism for explaining the onset of war.

Similarly, the effect of the bureaucratization of the state on its rational-legal procedures is likely to impact differently on the onset of interstate war depending on the mechanisms with which it interacts in a given context. On the one hand it lays the foundations for routinized practices that, by submitting increasingly important means to increasingly ambitious ends, build up acceptance of high-cost wars as highlighted by First and Second World Wars. On the other hand, it furthers the proliferation of ends and paves the way for the notion that the means to certain ends are also ends requiring means: this logic underpins the contemporary idea that risks of interstate war for populations and economies have to be managed and, as far as possible, minimized (Rasmussen, 2007). Furthermore, when alliances and international organizations breed transnational bureaucracies and form 'international conglomerates of state-power,' war becomes less likely within the latter (Vasquez, 2009; Shaw, 2001).

This mechanism and process-based account is not a plea for a naïvely speculative approach to causality. Rather, we argue that Tilly here sheds light on the limits of the linear logic of dominant approaches to causality. It offers a methodologically compelling explanation for the fact that in the social world, as highlighted by Hegelian and Marxian dialectical thinking, 'more of the same' can actually mean 'something different.' As a causal mechanism grows stronger, the relevant social context is transformed and the mechanisms with which it is likely to interact changes accordingly.

Causal-Explanatory Approaches to Interstate War in IR

Most of the IR literature on interstate war follows what Tilly described as the law-seeking account of explanation "*subjecting robust empirical generalizations to higher and higher-level generalizations, the most general of all standing as laws*" (Tilly and Goodin, 2006: 12). We must here recognize that the notion of causality is multifarious in this literature (Levy and Thompson, 2010). Different theories look for different types of causes: permissive, necessary, sufficient, underlying, proximate causes etc., Moreover there are different conceptions of what can or need to be explained. It is the latter aspect that we will use to structure this quick and necessarily cursive overview. First, we will deal with the theories explaining the permanent recurrence of interstate war. Secondly, we will present the explanatory accounts interested in the varying frequency of interstate wars throughout history. Thirdly, we tackle the attempts to shed light on the most likely causes of the wars that do occur. We will here focus on causal-explanatory theories, leaving more interpretivist-comprehensive approaches for the last section. We will disregard theories emphasizing 'human nature' or individual psychology, as this would engage us in a wider debate on the causes of violence and aggression in general, and we equally disregard theoretical explanation of particular wars.

First, many theories try to explain why interstate war is a recurrent feature of interstate relations. This is the case of the security dilemma that, as systematized by Waltz (Waltz, 1979), lies at the heart of so-called neorealist accounts. John Herz first defined the security dilemma as a "*structural notion in which the self-help attempts of states to look after their security needs tend (…) to lead to rising insecurity for others*" (Herz, 1950: 157). In other words, given the structural absence in international relations of any authority able to guarantee the integrity of all states (the presence of 'international anarchy' in the jargon), the defensive measures taken by a government to increase its national security might be interpreted as evidence of aggressive behavior by other governments. Since the latter will most likely take similar measures in response, the first government will ultimately have triggered a process that will not leave it more secure. Security is consequently seen as a zero-sum if not a negative-sum game. While peace might be the preferred option of all governments, they might under these conditions choose their second best option, to attack, rather than to risk the worst-case scenario: to be attacked. The security dilemma hence explains why, tragically, interstate war might always occur, even if all governments prefer peace, this at least as long as international relations are characterized by distrust and defensive measures are seen as threatening. As such, it does not go further in explaining why some wars occur and others do not. There are however great differences between authors considering that the security dilemma can be mitigated (defensive realism) and others who

consider that it is irreducibly acute in all interstate relations (offensive realism).

The 'bargaining model of war,' many of the foundations of which were laid by historian Geoffrey Blainey, similarly tries to explain why war always looms at the horizon of interstate relations. It mainly emphasizes the role of informational disagreements and misperceptions and uses microeconomic methodologies and theories. Rather than focusing on the international structure ('international anarchy'), it typically analyses the dyadic interactions between states. It posits that a government will only attack another provided two conditions are fulfilled: it hopes to achieve a favorable settlement of the underlying conflict (usually linked to scarce resources) that it cannot achieve through peaceful negotiations; given its superior power, it is sure to win the war and to achieve this favorable settlement through war at a lesser cost than what it expects to gain. If however both governments agree on this power asymmetry, the weaker government will agree on this settlement through peaceful negotiations rather than risking a war which it is bound to loose. Wars hence occur when there is disagreement on the relative power balance between states. James D Fearon's main contribution here is to explain, through the rational-choice perspective, why there might be such disagreements (Fearon, 1995). One of the explanations, 'the private information mechanism,' posits that during the peaceful bargaining process, both of the governments have an interest in concealing information about their military capacity to avoid countermeasures by their adversary. In other words, there are rational incentives to misrepresent the relative power balance, which ultimately makes war a recurrent phenomenon in international relations.

Secondly, there are theories mainly trying to explain the changing frequency of war throughout history. Some of these do not focus on the causes of war per se. This is the case of the so-called 'democratic peace theory' according to which democratic states usually do not wage war against one another. This liberal theory however attempts to explain the historic decline in interstate wars, and especially the disappearance of interstate war between Western states, by the progress of democracy throughout the world. While the statistic correlation between formally democratic regimes and the absence of war between them is more or less established (with some exceptions and reservations), there is no consistently validated theory to explain this. It is hence rather highlighting a correlation than causality. Another comparable idea sees economic interdependence as the main reason for the decline in interstate war in Western Europe during the nineteenth century and its absence since 1945. Economic interdependencies would raise the opportunity costs of war and reduce the relative value of territorial conquest. Marxists have however long argued that the opening of markets can actually be an objective of war rather than an unambiguous source of peace. Furthermore, here as well one can wonder whether there is not in fact correlation rather than causality. One could even argue that it is the absence of war that allows for the increase in economic interdependencies rather than the reverse (Shaw, 2001).

Many of the theories of war interested in the varying frequency of interstate war focus on the role played by the distribution of power in the international system. There is here significant disagreement on the question whether international systems based on a symmetric power balance between states (or coalitions) are less likely to witness war as traditional realists and neorealist theories argue, or if systems based on power-preponderance of one state are more stable as some realist-inspired theories suggest (hegemonic war theory, power transition theory, and long-cycle theory). According to the first category of theories (traditional realists and neorealists), governments tend to be risk averse: since a relative power balance implies that no government can be reasonably certain to win a war, they are altogether deterred from initiating war. There is however a debate between them on the comparative impact of multipolarity and bipolarity. Waltz argues that bipolarity will discourage war by reducing the risk of misinterpretation of the power balance (Waltz, 1979). On the contrary traditional realists Hans J. Morgenthau and Edward H. Carr considered that the reverse is true because multipolarity adds to the uncertainty. According to the second category of theorists (power-transition, long-cycle and hegemonic war theorists) power-preponderance is adverse to war: governments are in fact willing to take risks, but only as long as there is a chance of winning which is the case when there is a relative balance. When there is power-preponderance, however, the powerful state has no reason to challenge the status quo and the weaker ones are deterred from doing so.

Thirdly, there are the theories interested in the likely causes of the wars that occur irrespective of historical permanence or varying frequency. This is the case of the multicausal and not strictly rational choice-inspired 'steps-to-war approach' developed by John Vasquez (Vasquez, 2009). He shows that interstate war often is the result of a sequence of events during which the cumulative policy choices made by governments progressively restrain their policy-options during successive interstate disputes. As the crises escalate, they become more likely to spiral out of control. It is what Vasquez calls the 'realist road to war': the mainly nonintentional process though which the practice of realpolitik, as rationalized by the realist theory of IR and the idea of deterrence through 'power balancing,' is likely to escalate. At the same time, Vasquez's approach illustrates some of the downturns of correlational studies: its inability to produce other than commonsensical propositions. The main findings, concerning interstate wars between relative equals, are indeed that: they mostly occur between neighboring states; involve territorial disputes; are preceded by a history of rivalry involving arms races and repeated confrontations; are associated with unbalanced alliances. One of his conclusions is that territoriality, understood as the tendency of humans to occupy and defend territory, is the main underlying cause of interstate war. This view naturalizes a contingent feature of the state. Since states are by definition territorial, and the modern art of war is inseparable from a territorialized view of the world, it is not astonishing that interstate wars often are triggered by territorial disputes. Vasquez is certainly right that the tendency to contend over territory is not necessarily linked to economic interests or expansionist ideologies. This does however not prove, as he implies, that it is linked to the evolutionary past of the human species.

Ultimately, it is astonishing that the literature on the causes of interstate war has taken the state as an abstract given rather than focusing on its characteristics and specificities as a complex organization constituted by its ability to exert

territorial control in violent interaction with other such organizations. The main underlying cause of interstate wars over territory might lie in the territoriality of states rather than of humans.

Understanding Interstate War: Comprehensive Accounts

Collective ideas, ideologies, motivations, and emotions are arguably important factors in the onset of interstate war. The analysis of the latter however supposes to avoid common pitfalls. Rather than being an 'independent variable,' these ideational and emotional factors are simultaneously conditioned and spurred by social and political structures. The tendency to reduce the state to a representative agent with ideas, emotions, and motivations is also one the downsides of the literature on this subject. At the same time, Weber's idea of legitimized violence highlights that the state is inseparable from a long-term, mainly unintentional, shaping of representations, and also, in case of war, of a short-term, more intentional mobilization of "collective will" (Malesevic, 2010). As opposed to a purely causal explanation of interstate war, a complementary comprehensive approach, giving importance to meaning in shaping social action, is then called for.

Some authors have, from a sociological and constructivist-inspired perspective, focused on the role of 'national standing' and the resulting 'struggles for recognition' in triggering war (Lebow, 2010; Lindemann, 2010). They argue that discursively inflated collective identities and the devaluation of the 'collective other' play a significant role in the onset war. These dimensions are important for at least two reasons. First, modern states are inseparable from the historical process of "domestic pacification" (Elias, 2000) and the correlative devaluation of violence. In this respect, the legitimization of collective mass slaughter supposes a prior dehumanization and denigration of the adversary in order to neutralize this principle (Malesevic, 2010). Secondly, states are highly hierarchical and coercive in spite of their principal egalitarianism rooted in formally egalitarian and universalistic mass ideologies (democracy, nationalism, etc.). The resulting contradictions become acute at times of war when some fight and die and some do not. The projection of social hierarchy unto international relations, through a patriotic narrative (often draped in nationalism), hence becomes instrumental to the war effort. It conveys the idea of a homogenous 'national community' to which everybody's personal 'standing' would be linked and that hence temporarily 'invisibilizes' these hierarchies (Lebow, 2010:181; Malesevic, 2010: 272–273). 'Wars for recognition' become then less of a psychological fact, be it on the collective or individual level, than an element deriving from the nature of the state as simultaneously based on structural inequalities and principal egalitarianism.

How far do these theories go in shedding light on a particular pathway to war? One can indeed consider that the factors that they focus upon highlight less of an underlying reason than a permissive condition of war. Moreover, as we have seen, they have themselves structural causes that shed light on their limits as 'independent variables.' However, several arguments can account for the fact that these ideational factors, often conveyed by official discourses and dominant political ideologies, can come to play a significant role in the onset of war: firstly foreign-policy goals can themselves come to be shaped by these narratives and the values and preferences that they convey, thus leading to emotionally charged interstate interactions (reprisals, shows of force…) in which 'standing' is revealed as an essentially zero-sum game; secondly, the domestic social status of the ruling and military elite might be dependent on the ability to claim success in these quests for national recognition; thirdly, to the extent that the resulting narratives successfully mobilize populations and build up expectations, they create domestic pressures on the governing staff to stay true to them and to act accordingly. This might explain the recurrence of rivalries between relative 'equal states' where ideological questions of 'national standing' are particularly salient and where reciprocation of the actions taken by rivals becomes a matter of 'honor' and 'prestige.' Lebow however considers this motivation of interstate war to be in decline due to the normative devaluation of war: if wars of aggression are morally derided, how could they increase 'national standing'? We would however argue that the current Western narrative on civilizational and moral superiority in military operations that are denied as wars, reproduces this logic of 'standing.' Moreover, the normative devaluation of war does not prevent an attacked party from framing this attack, not only as the actualization of a 'material threat,' but also as a 'denial of recognition,' an 'offense,' and to act accordingly (Lindemann, 2010).

Conclusion

Interstate war is variable in time and space. In this regard, this article tries to draw attention to the crucial contribution of, among others, historical sociology to the study of this phenomenon by contextualizing its patterns and causes. It allows accounting for the specificity of the state-form compared to other forms of political organization, but also to highlight the empirical diversity of particular states and 'interstate systems.' As this article attempts to show, such a sociological approach to interstate war also serve to relativize the artificial distinctions that structure so much of the IR literature on interstate war and its causes: between causes located at individual, governmental, and international systemic levels; between materialist and idealist explanations; between the domestic-internal and the international-external realms of the state.

See also: Conflict and Consensus; Military Sociology; Military, War, and Politics; Nation-State and War; Securitization; Territorial Conflicts; War Propaganda; War and Nationalism; War, Sociology of.

Bibliography

Ayoob, Mohammed, 1995. The Third World Security Predicament: State Making, Regional Conflict, and the International System. L. Rienner, Boulder.

Bull, Hedley, 1977. The Anarchical Society. A Study of Order in World Politics. Columbia University Press, New York.

Elias, Norbert, 2000. The Civilizing Process: Sociogenetic and Psychogenetic Investigations. Basil Blackwell, Oxford.

Fearon, James D., 1995. Rationalist explanations for war. International Organization 49 (3), 379–414.
Herz, John, 1950. Idealist internationalism and the security dilemma. World Politics 2 (2), 171–201.
Holsti, Kalevi J., 1996. The State, War and the State of War. Cambridge University Press, Cambridge.
Lebow, Richard N., 2010. Why Nations Fight: Past and Future Motivations for War. Cambridge University Press, Cambridge.
Levy, Jack S., Thompson, William R., 2010. Causes of War. Wiley-Blackwell, Oxford.
Lindemann, Thomas, 2010. Causes of War. The Struggle for Recognition. ECPR Press, Colchester, UK.
Malesevic, Sinisa, 2010. The Sociology of War and Violence. Cambridge University Press, Cambridge.
Mueller, John, 2004. The Remnants of War. Cornell University Press, London.
Nettl, J.P., 1968. The state as conceptual variable. World Politics 20 (4), 559–593.
Rasmussen, Mikkel Vedby, 2007. The Risk Society and War. Terror, Technology and Strategy in the Twenty First Century. Cambridge University Press, Cambridge.
Shaw, Martin, 2005. The New Western Way of War: Risk-Transfer War and its Crisis in Iraq. Polity Press, Cambridge.
Shaw, Martin, 2002. Globality and historical sociology: state, revolution and war revisisted. In: Hobden, S., Hobson, J.M. (Eds.), Historical Sociology of International Relations. Cambridge University Press, Cambridge, pp. 92–98.
Tilly, Charles (Ed.), 1975. The Formation of National States in Western Europe. Princeton University Press, Princeton.
Tilly, Charles, Goodin, Robert E. (Eds.), 2006. The Oxford Handbook of Contextual Political Analysis. Oxford University Press, Oxford.
Vasquez, John, 2009. The War Puzzle Revisited. Cambridge University Press, Cambridge.
Waltz, Kenneth N., 1979. Theory of International Politics. Mcgraw Hill, New York.

Water, Health and Social Inequality

Linda M Whiteford, University of South Florida, Tampa, FL, USA
Cecilia Vindrola-Padros, London South Bank University, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article reviews writings about water with a focus on health from the social and behavioral sciences. In addition, issues related to the equitable distribution of water resources such as potable water and their health consequences are discussed. A case study of one water-related disease (cholera) is used to exemplify the inexorable consequences of the failure to provide clean water and sanitation equitably. The reader also will be introduced to the magic triangle of water, health, and hygiene captured in the Global WASH (water, sanitation, and health) Campaign to include sanitation as a target of the Millennium Development Goals.

Introduction

Throughout human prehistory and history, access to water has been a key concern and often contentious topic. Foraging peoples structure their annual migrations around access to water. Early agriculture that supported the increasingly large and concentrated populations in Mesopotamia was made possible by the construction of complex and massive irrigation canals. These early canal systems were themselves made possible and controlled by the development of a social structure reflecting a specialization of labor. The large-scale manipulation of water occurred conterminously with the growth of large-scale and densely populated settlements and their concurrent evolving cultures. This is but one small example of how water has always been a critical component of human activities from agriculture, industry, urban development, recreation, but most centrally, to life itself. Water has been channeled, dammed, transported, stored in giant reservoirs, dropped from helicopters, sprayed during riots and on forest fires, but most often it is used for the quotidian aspects of life such as cooking, cleaning, and bathing. And access to a clean and reliable supply of water is primary to good health outcomes (Cairncross, 2003; Cairncross and Valdmanis, 2006; Whiteford and Vindrola-Padros, 2011).

While water has long been recognized as a central element in human health, its partner in reducing the spread of potentially deadly diseases, sanitation, has not shared in the same limelight. That is, however, in the process of changing. In 2001, The Global WASH (water, sanitation, and health) Campaign, initiated by The Water Supply and Sanitation Collaborative Council spearheaded a global advocacy movement around safe water, sanitation, and hygiene. In this article, we include the equitable distribution of sanitation, along with water and hygiene, as a critical component to improved global health outcomes.

The recent and deadly epidemic of cholera in Haiti is a classic example of the consequences of failing to provide water and sanitation necessary to maintain the hygiene required to prevent the transmission of fecal/oral diseases. The cholera epidemic following the 2010 earthquake in Haiti demonstrates how water, sanitation, and health are connected and how their presence (or absence) reflects underlying social inequalities. In this article, we discuss the cholera epidemic in Haiti as one example of a devastating water-borne disease. Other water-borne diseases, such as diarrhea, are also killers. As Figure 1 suggests diarrhea, for instance, kills 1.5 million children every year; and is the leading cause of malnutrition in children under 5 years of age (WHO, 2012a). And the tragedy is

Figure 1 Poster created by the WSSCC as part of the WASH Hurry Up! campaign to raise awareness about the current water and sanitation crisis. Source: Water Supply and Sanitation Collaborative Council (WSSCC), 2012. Resources Hurry Up! http://www.wsscc.org/resources/resource-advocacy-materials/hurry#10 (last accessed 12.08.12.).

that both cholera and diarrhea are preventable and treatable. What makes the Haiti cholera example so powerful is that at multiple times, and in multiple ways, it also could have been avoided. It is an exemplar of government failure to create and maintain water and sanitation infrastructure *so* necessary for the well-being of a population. Whether it is cholera in India in the third century, or again from 1817 to 1920 (Watts, 1997), the South American cholera pandemic in the 1990s (Whiteford and Vindrola-Padros, 2011), or the ongoing endemic epidemic of cholera in sub-Saharan Africa (WHO, 2012b), cholera kills needlessly. With clean water and sanitation, the cholera vibrio does not survive to sicken and kill. Clean water and sanitation must therefore be considered universal human rights, as the UN has argued:

> On 28 July 2010, the UN General Assembly recognized that safe and clean drinking water and sanitation are human rights, essential to the full enjoyment of life and all; other human rights. Subsequently, at its 15th session in September 2010, the UN Human Rights Council affirmed that the right to water and sanitation is derived from; the right to an adequate standard of living and inextricably related to the right to the highest attainable standard of physical and mental health, as well as to the right to; life and human dignity. The combined effect of the two resolutions was to anchor the; right to water and sanitation in the framework of the right to an adequate standard of living, making it legally binding like any other of the rights inscribed in UN treaties. Fundamental to the human rights framework is the concept of progressive realization: Governments cannot solve the drinking water and sanitation situation overnight, but; they must make tangible progress towards the realization of this right. Human rights; principles also define various characteristics against which the enjoyment of the right can be assessed, namely: availability, safety, acceptability, accessibility, affordability, participation, non-discrimination and accountability. A distinctive feature of the human rights framework is the principle of non-discrimination. This; requires looking beyond average attainment and disaggregating data sets to determine whether any sort of discrimination is occurring.
>
> (UNICEF and WHO, 2012)

Access to a clean and reliable supply of water is neither random nor equitably distributed globally. Geography, history, and politics shape the distribution of water across the globe, and even within countries social groups differ in their access to water (Whiteford and Vindrola-Padros, in press). Hygiene, the practice related to washing one's hands and other body parts to reduce the transmission of fecal-oral diseases like cholera or other forms of diarrhea, is strongly connected to the availability and reliability of a source of clean water (Whiteford, 1999). Social and behavioral science research has repeatedly demonstrated how people's attitudes and practices reflect not only cultural rules and expectations, but also the ability to regularly access clean water (Esrey, 1996). During the 1991 cholera pandemic that moved through South and Central America (in research funded through a bilateral agreement between USAID and the Ecuadorian government), researchers found that in areas of water scarcity, water was used and reused, stored, and parceled out. Clean water was first used in cooking, food preparation, and dishwashing, then for personal use like hand washing or bathing, and then as slops for the animals. As part of the Ecuadorian cholera intervention project, soap was introduced along with health education about the need to use soap. The assessment following the intervention showed many successes, but not in the use of soap (Whiteford, 1999; Whiteford and Laspina, 1996a; Whiteford et al., 1996b; Whiteford and Vindrola-Padros, in press). Ethnographic interviews with women in the households uncovered the reason for the failure to use soap in water that needed to be used multiple times and for multiple purposes. 'The pigs do not like soap in their water' women said, 'so we stopped using soap'. If the pigs were less picky, or if the household had more water so the water did not need to be used so many times, the women would have used soap. The lesson is that in water-scarce settings, householders are particularly conscious of the many uses the water must serve, both inside and outside of the house and adapt their practices accordingly.

Water and Health

According to the World Health Organization (WHO), one of the greatest global challenges to world health is the prevention of water-related disease (WHO, 2012c). Close to 1 billion people lack access to an improved water supply, and there is an estimated 2 million annual deaths attributable to unsafe water, sanitation, and hygiene worldwide. In addition, more than 50 countries report the continuing presence of cholera (WHO, 2012c). Schistosomiasis, another water-related disease, infects an estimated 260 million people globally (WHO). The list of water-related diseases and what parts of the globe people are most affected reflects the continuing inequality of access to clean water and sanitation. An estimated 4% of the global disease burden could be prevented by improving water supply, sanitation, and hygiene (WHO, 2012c). "Poor water quality continues to pose a major threat to human health. Diarrheal disease alone amounts to an estimated 4.1% of the total disability adjusted life years global burden of disease and is responsible for the deaths of 1.8 million people every year (WHO, 2004). It was estimated that 88% of that burden is attributable to unsafe water supply, sanitation, and hygiene and is mostly concentrated on children in developing countries" (WHO, 2012c).

Water-related diseases (some of which can be seen in **Table 1** above) are common worldwide and require environmental as well as infrastructural, educational, and behavior change interventions to control their spread. And as humans continuously modify their surroundings by building canals, transforming wetlands into agricultural areas, removing forests, irrigating arid lands, and by bringing new animals – including humans – into contact with the vectors, the type of threat changes, as do the responses. One classic example of this vector/environment/human health picture can be seen in the story of the Aswan High Dam in Egypt. "In 1971, with the completion of the great dam at Aswan and the filling of Lake Nasser behind it, the epidemiology of Schistosomiasism in Egypt changed radically" (Desowitz, 1987).

Schistosomiasis (also known as bilharzia) is a water-related disease caused by parasitic worms that spend part of their life cycle in the water. More than 200 million people are infected globally and the disease is second only to malaria as the most devastating parasitic disease (CDC, 2012). The parasites that cause schistosomiasis live in types of freshwater snails and the infectious form of the parasite emerges from the

Table 1 Partial listing of water-related diseases and their transmission routes

Transmission/route	*Definition*	*Preventive strategies*	*Examples*
Water borne	Person/animal ingests water with pathogen	1. Improve water quality 2. Prevent casual use of unprotected sources	Cholera; typhoid; infectious hepatitis; diarrhea; dysentery
Water washed	Due to inadequate quantity of water for personal hygiene	1. Increase quantity 2. Improve accessibility 3. Improve reliability 4. Improve hygiene	Cholera; bacillary dysentery; bacterial skin sepsis; scabies; fungal skin infection; typhus; trachoma
Water based	Pathogen that spends part of life cycle in water (all parasitic worms)	1. Reduce need for contact with infected water 2. Control snail pop 3. Reduce contamination of surface water	Schistosomiasis guinea worm
Insect vector	Transmitted by insects breeding or biting in or near water	1. Improve surface water management 2. Destroy breeding sites 3. Reduce need to visit breeding sites 4. Use mosquito bed nets	Malaria, yellow fever, dengue, onchoceriasis

Source: Ryan Schweitzer adapted from Mihelcic, J.R., Fry, L.M., Myre, E.A., Phillips, L.D., Barkdoll, B.D., 2009. Field Guide in Environmental Engineering for Development Workers: Water, Sanitation, Indoor Air. American Society of Civil Engineers (ASCE) Press, Reston, Virginia.

Figure 2 Left: *Biomphalaria* sp., the intermediate host for *Schistosoma mansoni*. Right: *Bulinus* sp., the intermediate host for *Schistosoma haematobium* and *Schistosoma intercalatum*. Center: Adults of *S. mansoni*. The thin female resides in the gynecophoral canal of the thicker male. Source: Centers for Disease Control and Prevention (CDC), 2012. Parasites-Schistosomiasis. http://www.cdc.gov/parasites/schistosomiasis/ (last accessed 12.09.12.).

snail, contaminating the surrounding water (see **Figure 2**). People are infected when breaks in their skin come in contact with contaminated freshwater as, for instance, when women wade into contaminated water to wash clothes or engage in other household tasks. The parasites enter the body and become worms that can lodge in the urinary or intestinal tracks and are released back into the water through defecation (CDC, 2012).

In the Egypt case, the Aswan High dam changed patterns of seasonal irrigation and the winter closing of the irrigation canals. Prior to building the dam the schistosome, while endemic to the area was primarily limited to the Delta, and annual relief was provided during the winter season when the canals were allowed to dry out and were dredged to clear them out. As the canals were annually dredged of the accumulated silt, the snails were also removed (Malek, 1975).

But all that changed with the construction of the dam and with the year around activity it made possible. Concomitantly, human activity and fecal pollution (and the distribution of the parasite) increased. As the human population increased, so did its risk of superinfection. Without a doubt the electricity provided by the dam enhanced Egypt's ability to develop the area, but as more than one health advisor has observed: "The beneficiaries of the Aswan project were the snail and the schistosome: they flourish… Unfortunately, the (hydro) electric bill too often contains a high schistosomal surcharge" (Desowitz, 1987).

While the snails may have been a beneficiary of the dam, household and traditional gender roles resulted in women being increasingly exposed to the parasites through increased contact with water contaminated with the snails (Watts, 2004). As the annual dredging of silt (and snail) removal no longer occurred, and as population increased, so did rates of infection, at least in the Delta area. Gender roles and economic status were both implicated in the increased levels of schistosomiasis because certain groups of people (women and people whose livelihoods were attached to the water) were unequally exposed to the contaminated water. This pattern is repeated globally as people in resource scarce countries are more exposed to water-related diseases, and women, poor people, and marginalized or disenfranchised groups are at higher risk of infection.

Gender, economics, geography, and education are all characteristics that correlate to inequalities in access to clean water and sanitation, and thus, are directly tied to increased risk of exposure to water-related disease (Coles and Wallace, 2005; De Alburquerque, 2009; El Katsha and White, 1989; Roark, 2003). Women are particularly vulnerable as they may have to travel considerable distances to bring water from its source to their homes. Likewise, women who must use open defecation practices because of the lack of sanitation often seek secluded areas to protect themselves from being seen. However, that very seclusion may expose women and girls to predatory attacks. These examples show us how water and sanitation are directly tied to power and social inequality (Ahlersa and Zwarteveenb, 2009).

According to a UNICEF and WHO report on the time used to collect water in water-poor areas: "An analysis of data from 25 countries in sub-Saharan Africa, representing 48% of the region's population, reveals that women and girls bear primary responsibility for water collection, at considerable cost in terms of their time. Only a quarter of the population in these countries had water on their premises in 2010, meaning that in 75% of households, water had to be collected from a source some distance from the dwelling. In 71% of all households without water on the premises, women or girls are mainly responsible for water collection. In 29% of households, men or boys assume this task. Further analysis shows that the mean time of one round-trip to collect water is approximately 30 min for both women and men, and is only slightly lower for children (28 min). Each household requires at least one trip per day, but may, in fact, require several trips. The time and energy devoted to water collection is considerable, even based on a one trip per day minimum. In these 25 countries, it is estimated that women spend a combined total of at least 16 million hours each day collecting drinking water; men spend 6 million hours; and children, 4 million hours" (UNICEF and World Health Organization, 2012: p. 31).

One response to the unequal distribution of water and sanitation and their health and well-being consequences is the 1999 Protocol generated by the United Nations Economic Council for Europe (UNECE), The Protocol on Water and Health to the 1992 Convention on the Protection and Use of Transboundary Watercourses and International Lakes (UNECE, 2000).

Article 1 of the Protocol states: "The objective of this Protocol is to promote at all appropriate levels, nationally as well as in transboundary and international contexts, the protection of human health and well-being, both individual and collective, within a framework of sustainable development, through improving water management, including the protection of water ecosystems, and through preventing, controlling and reducing water-related disease" (UNECE, 2000).

According to the document: "… in the European part of the UNECE region alone, an estimated 120 million people, i.e., one person in seven, do not have access to safe drinking water and adequate sanitation, making them vulnerable to water-related diseases, such as cholera, bacillary dysentery, coli infections, viral hepatitis A, and typhoid. Cleaner water and better sanitation could prevent over 30 million cases of water-related disease each year in the region. The 1999 Protocol on Water and Health was negotiated with this in mind.

The main aim of the Protocol is to protect human health and well-being by better water management, including the protection of water ecosystems, and by preventing, controlling, and reducing water-related diseases. The Protocol is the first international agreement of its kind adopted specifically to attain an *adequate supply of safe drinking water and adequate sanitation for everyone* (UNECE, 2000: p. 10, emphasis not in the original) and effectively protect water used as a source of drinking water.

To meet these goals, its parties are required to establish national and local targets for the quality of drinking water and the quality of discharges, as well as for the performance of water supply and wastewater treatment. They are also required to reduce outbreaks and the incidence of water-related diseases.

This Protocol introduces a social component into cooperation on water management. Water resources management should link social and economic development to the protection of natural ecosystems. Moreover, improving the water supply and sanitation is fundamental in breaking the 'vicious cycle of poverty' (UNECE, 2000). The relation between access to clean water and sanitation and good health is clear and has been well understood for decades. However, the global provision of an equitable supply of water and sanitation is elusive. But steps are being taken. Both the Millennium Development Goals (MDGs) and 1999 Protocol provide mechanisms by which to assess and measure improvements in access to clean water and sanitation, and simultaneously (at least the 1999 Protocol) the management of shared resources such as bodies of freshwater. As we will see in the next section, the case of the cholera epidemic in Haiti demonstrates how failed attempts to provide clean water and sanitation allowed the introduction of the cholera vibrio to spread rapidly with deadly results.

Water, Health, and Inequality in Haiti

The 7.0 earthquake that struck Haiti on 12 January 2010 killed 316 000 people and displaced nearly 1.5 million to camps and temporary housing facilities (IOM, 2010). The country was devastated and the large influx of international humanitarian aid was not properly targeted, nor sufficient enough to compensate for the lack of medical resources, destroyed infrastructure, and an ill-functioning water and sanitation system. Internally displaced people (IDPs) were relocated to camps that were not always equipped to handle the large influx of people and many camps did not have water or toilets. The spread of disease was imminent in too many ways.

Since the laboratory confirmation of the first cholera case was made on 21 October 2010 (Mendoza et al., 2011: p. 2177), 250 000 cases, about 125 000 hospitalizations, and 4000 deaths have been reported (Dowell and Braden, 2011; Tuite et al., 2011). The first Haitian cholera case was produced by the contamination of water sources by UN troops who had traveled from Nepal and were stationed near the town of Artibonite (Tuite et al., 2011). Haiti had not experienced cholera for more than a century (Mendoza et al., 2011: p. 1161) and suddenly it had to deal with one of the largest cholera epidemics in the world.

An Epidemic Waiting to Happen

The sharp political, social, and economic inequalities in Haiti set the stage for the rapid spread of the disease, the inability to

contain the bacteria, the late treatment of patients, and the avoidable deaths of thousands of people. As Bell has argued, "problems that occur elsewhere are often more extreme in Haiti, the causes found in the age-old processes of domination and expropriation" (2001: p. 9). Haiti is located in the Greater Antillean Archipelago and shares the Island of Hispaniola with the Dominican Republic. It is divided into 10 departments and Port-au-Prince, the capital, is the largest city (WHO, 2011).

Haitian political administration has been characterized by political instability with 32 coups and a series of brutal dictatorships (WHO, 2011). Pan American Health Organization summarized Haiti's recent socioeconomic situation as follows: "one of the poorest and smallest countries in the world and the least developed in the American region; a high level of corruption and inequality; severe environmental deterioration and deforestation; lack of export industries and a large deficit; an on-going brain drain, primarily to the United States, Canada, and France; most services delivered by NGOs with token supervision and consultation from the Government" (2011: pp. 2–3). With regard to this last point, it is important to highlight the reliance on nongovernmental organizations (NGOs) for the management of the basic services required by the population in Haiti and its negative effect on the role of governmental institutions (Schuller, 2012). International donors and providers of aid have favored foundations and other forms of NGOs when depositing their funds, thus reproducing a privatized form of humanitarian assistance (Schuller, 2012).

Two of the factors most intimately related to the spread of cholera, an inadequate water and sanitation system and poor access to health care, have been severely affected by the political and economic situation outlined before. The public health care system suffered so much deterioration that as of 2010, it delivered only a third of health services (Ivers et al., 2010: p. 2050). Medical care in Haiti has been described as 'an obstacle course, one that places innumerable barriers before poor people seeking care' (Farmer, 1999: p. 145). In 2009, there were three physicians per 10 000 people and 13 hospital beds per 10 000 people (PAHO, 2011). The people who could physically access public services paid for fees as the most basic services had hidden costs (Farmer, 2011). This situation certainly contributed to the low life expectancy (59 for males and 63 for females), high infant (80.3 per 1000) and maternal mortality (523 per 100 000) rates in the country, the highest incidence of tuberculosis in the Americas, and an ongoing HIV epidemic (PAHO, 2011; WHO, 2011).

The water and sanitation system fared no better. Between 2000 and 2004, international aid to the Haitian public sector was significantly reduced, and donors and lending agencies preferred to direct funding to NGOs (Ivers et al., 2010: p. 2048). Water shortage was exacerbated by the earthquake, but it was certainly not a new phenomenon in Haiti. For instance, in 2002, Haiti was ranked last among 147 nations in a water poverty index produced by the World Resources Institute in Washington, DC (Webster, 2011: p. E83). In 2006, it was estimated that only 58% of the population had sustainable access to improved drinking water sources while 19% were able to use improved sanitation (WHO, 2011).

Table 2 The spread of cholera cases in Haiti, 2010

Date	*Event*
14 October 2010	First cholera case in Meille
17 October 2010	First death registered in Mirebalais hospital
18 October 2010	Increase in diarrhea cases reported by the Cuban medical brigades to the Ministry of Public Health and Population
	Haitian investigation team is sent to Meille
19 October 2010	Ten additional cholera cases found in Meille
	The first alert from Bocozel in the Lower Artibonite is received when three children die from diarrhea
20 October 2010	New alerts received from approximately 10 health centers along the lower course of the Artibonite river
21 October 2010	3020 cases, 1766 hospitalizations, and 129 deaths were reported in the Lower Artibonite region
	Epidemiologic confirmation of first case in Meille
22 October 2010	Cholera cases notified in 14 communities in the regions around the Artibonite plain and Port-au-Prince
11 November 2010	Clusters of cases were identified in the North-West Department
14 November 2010	Clusters of cases were identified in the West Department
16 November 2010	Cases confirmed in the Dominican Republic and Florida (U.S.)
21 November 2010	Clusters of cases were identified in the North-East Department (Haiti)

Source: Centers for Disease Control and Prevention (CDC), 2010. Update on cholera-Haiti, Dominican Republic, and Florida, 2010. Morbidity and Mortality Weekly Report 59 (50), 1637–1641; Piarroux, Renaud, Barrais, Robert, Faucher, Benoit, Haus, Rachel, Piarroux, Martine, Gaudart, Jean, Magloire, Roc, Raoult, Didier, 2011. Understanding the cholera epidemic, Haiti. Emerging Infectious Diseases 17 (7), 1161–1168.

In addition, the 2010 earthquake left about 1.5 million people without homes, forcing them to move to temporary settlements, where for many, living conditions worsened (IOM, 2010; PDNA, 2010). In these internally displaced camps, 40% of the camps had drinking water and 30% did not have toilets (Kaussen, 2011: p. 6). As a result of this situation, the cholera bacteria spread at a rapid pace, first in Mirebalais and then in each of the Haitian departments. Trauma victims, those with poor preexisting health status, and nutritionally compromised people could not fight it off and instead became its fatalities. In **Table 2**, we have summarized the different stages of the spread of cholera as outlined by Piarroux et al., 2011, the team of French and Haitian epidemiologists who tracked the dispersion of cholera.

Dealing with the Epidemic

The earthquake that struck the country led to an unprecedented immersion of foreign humanitarian assistance. Foreign assistance took different forms, from the influx of monetary donations to the settlement of medical and assistance personnel in IDPs and Cholera Treatment Centers (CTCs) all over the country. It is estimated that the international community pledged to provide $10 billion in aid to Haiti, but only 2% of

this money had been spent by 2010 (Kaussen, 2011: p. 6). Even though portions of this fund were eventually made available to the country's officials, cases of government corruption were documented, the money was not used effectively, and there was no guarantee that the funds would be used to deal with Haiti's most pressing problems (Farmer, 2011: p. 210).

The coordination of the large number of local and international NGOs working in Haiti after the earthquake and during the cholera epidemic was not without complications. The lack of public oversight of NGOs led to the creation of gaps in services; the distribution of aid was unequal in that a large portion of the aid was focused on the areas that were most affected by the earthquake and not enough attention was paid to the rural areas where the epidemic ultimately emerged; and the overreliance on NGOs for the management of basic services demeaned the role of governmental institutions (Ivers et al., 2010: p. 2048; Schuller, 2010).

By the end of 2011, most public health reports indicated that the cholera epidemic in Haiti had moved toward containment as fewer cases were being reported and treatment was provided to those already suffering from the disease. Several publications, however, cautioned that the improvement of the cholera scenario could be the product of the end of the rainy season in the region and they predicted the disease would probably reemerge once the rain returned (MSF, 2012). That prognosis was also supported by reports that proclaimed that water and sanitation infrastructure had not undergone significant transformations and the living conditions in settlement camps had actually deteriorated (Webster, 2011).

During the first half of 2012, many of these predictions became realities. The rainy season began and the disease continued to spread. Medecins sans Frontieres (MSF) reported in less than 1 month into the rainy season, that the cases of cholera patients in their CTCs nearly tripled. From 16 to 23 April 2012, 134 cholera patients were admitted into MSF's center in Martissant and another 400 patients were admitted to the other MSF centers in Port-au-Prince and Leogane. This organization even had to reopen a CTC in Carrefour to prevent the overburdening of other operating centers in nearby cities (MSF, 2012).

The cholera epidemic in Haiti points to the fact that in areas of the world where water, sanitation, and access to health care cannot be guaranteed for society as a whole, the spread of infectious diseases (like cholera but not limited to it) will remain a latent possibility. The fragile condition of Haitian public services posed a threat to people's lives even before the earthquake struck the country. A natural disaster of this magnitude certainly complicated matters, but Haiti's cholera epidemic was mainly produced by deeper political, economic, and social factors that led to its long-term economic and material impoverishment, dependence on international humanitarian assistance, and lack of coordination of public services.

Social and Behavioral Sciences on Water/Health/Sanitation

The use of water by human populations has been a topic of interest to researchers because of its importance in everyday life. Water is a social commodity in that it 'connects different domains of social life to each other in ways that are not haphazard or accidental because they depend on each other' (Orlove and Caton, 2010: p. 402). In other words, water is not only vital for human existence; it is engrained in all aspects of our society, from gender relations within the household that determine who collects and uses water, to complex economic and political systems that organize distribution at macro levels.

Water has had multiple representations: a quantifiable good that can be commoditized, a scarce resource, a productive tool, and a universal resource (Aiyer, 2007; Derman and Ferguson, 2003). Its connection with all aspects of human life renders water studies varied in approaches; they have focused on the political nature of water, the power relations involved in its control and the inequalities in its distribution (Boelens and Doornbos, 2001; Johnston, 2005). These inequalities are often organized in relation to gender and social class. In relation to gender, "rights and entitlements may shape the ways in which women feel able to participate in particular mechanisms of water access, such as water-point committees or farmers' organizations, how their voices are heard, and how their rights of access are shaped and asserted at the local level" (Cleaver and Hamada, 2010: p. 30). In some societies, gender influences access and use of water in ways that are detrimental to women, excluding them from local property rights and governance bodies, and prohibiting their access to water sources due to religious norms or taboos around menstruation, pregnancy, and childbirth (Cleaver and Hamada, 2010: p. 30). In other cases, women are the ones in charge of the hard labor surrounding water collection, investing energy and time in the collection of a resource that is fundamental for the survival of the household.

The distribution of water is a process involving multiple actors, each with its own needs and interests. Allen et al. have proposed to look at water provision through a 'water supply wheel' (**Figure 3**) where the public, private, and community sectors negotiate different arrangements for water distribution based on policy and needs driven interests. In the developing world, water access is often generated by individual needs and informal mechanisms, rather than by governmental policies, thus leaving specific sectors of the population, usually the resource-poor social classes, without reliable access to potable water (Allen et al., 2006a).

Inequalities in access to safe and potable water have been of particular interest to health researchers as water access, reliability, and quality are strong determinants of health and well-being (Whiteford and Cortez-Lara, 2005; Whiteford and Whiteford, 2005). Systems of water management directly affect the spread of disease and when guided by the principle of water as a commodity, tend to leave marginalized and underprivileged sectors of the population more vulnerable to water-borne, water-based, and water-related infections. Water-borne diseases are caused by ingesting water that is contaminated by human or animal excreta or urine containing pathogenic bacteria or viruses, while water-based diseases are caused by parasites located on intermediate organisms living in water (Montgomery and Eimelech, 2007). These diseases are usually prevalent in areas where people do not have access to running potable water and most frequent other, usually contaminated, sources of water. Water-related diseases are caused by microorganisms with life cycles associated with insects that live or

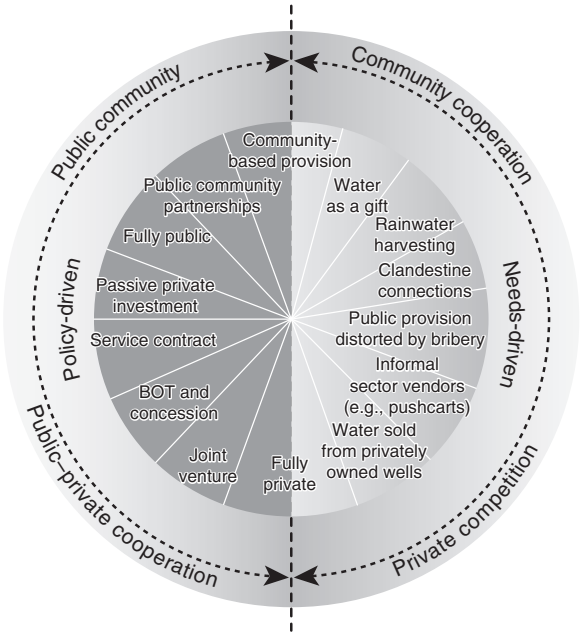

Figure 3 "Water supply wheel" developed to highlight the different types of water distribution. Source Allen, A., Davila, J.D., Hofmann, P., 2006a. The peri-urban water poor: citizens or consumers? Environment and Urbanization 18 (2), 333–351.

breed in water and water collection and storage diseases are caused by contamination that occurs during or after the collection of water (Montgomery and Eimelech, 2007). When access to water is unreliable, people are forced to store water in containers for future use, thus creating breeding grounds for insects that act as disease vectors.

Sanitation has been a relatively unexplored topic by social scientists (Singer, 2012; Whiteford and Whiteford, 2005). Most studies have stemmed from other disciplines such as environmental engineering and public health (O'Loughlin et al., 2006; Rodgers et al., 2007; Wakeman, 1995; McConville and Mihelcic, 2007). This situation is complicated by the fact that policy discussions on the need to consider access to water as a universal human right do not always explicitly include sanitation (Allen et al., 2006a).

When examining practices of waste management, studies have found that this is an aspect of human behavior heavily influenced by social and cultural factors (Kendall, 2005). The management of human waste is influenced by religious beliefs and ideas regarding purity and danger (Alley, 2002). It is not surprising then that latrine programs that have failed to take these beliefs into consideration have been unsuccessful, such as the latrine program directed by the Swedish Army in Somalia during the 1980s where Muslim refugees refused to defecate facing Mecca (Waldman and Williams, 2001) or proposals for latrine programs in Bolivia where participants would not engage in the digging of their waste because it would be considered disrespectful of the Pachamama (deity of mother Earth) (Aruquipa, 2008).

Where defecation takes place, when, and how are all issues inextricably linked to social categories such as gender, class, and ethnicity (Wakeman, 1995). Previous research found that women and young children are the groups that benefit the most from sanitation programs as they are the ones at greatest risk of violence (attacked in 'defecation zones', victims of rape, etc.) when these facilities are lacking (Schuller, 2010; Wakeman, 1995). The distribution, availability, and quality of latrines and/or toilets mirror the distribution of economic resources in the larger society where disadvantaged groups and the lower social classes are often left without sanitation infrastructure or with subpar services (Allen et al., 2006a,b). In many areas of the world, disadvantaged groups are often indigenous groups because they live in marginalized areas without access to basic services such as potable water and sanitation (Whiteford and Vindrola-Padros, in press). This leads to a higher incidence of water-borne and excreta-borne diseases in particular ethnic groups as evidenced by the cholera epidemics of 1992–93 in the Orinoco Delta of Venezuela (Briggs and Mantini-Briggs, 2003) and 1991–93 in Peru and Ecuador (Whiteford and Vindrola-Padros, in press).

When attempts have been made to change or improve sanitation, these interventions have been influenced by community-level behaviors (Yacoob and Whiteford, 1994) and individual preferences (Kendall, 2005; O'Loughlin et al., 2006; Rodgers et al., 2007). Even in cases where latrine programs have been developed in relation to local customs and practices, many times community members misuse or discontinue latrine use due to bad smell, flies, lack of privacy, or the perception that latrines are dangerous (O'Loughlin et al., 2006; Rodgers et al., 2007). Many studies have, therefore, concluded that technological innovation in the form of transforming sanitation infrastructure needs to be adapted to the local context, in sustainable form, and engaging the participation of community members (McConville and Mihelcic, 2007).

Conclusion

Perhaps no other elements are as central to global health and well-being as water and sanitation. And while the relationship between access to a clean and reliable supply of water and health has long been understood, only recently have people involved in development, and global health begun to include sanitation as the twin urgent need. In this article, we used the case of the cholera epidemic in Haiti to demonstrate how the failure to create an equitable system for a water and sanitation infrastructure resulted in avoidably high levels of loss of life.

Political turmoil over the control of water is perhaps as old as civilization itself. Shaped by power relations and reflected in the inequitable distribution of risk that women, the elderly, children, and the poor suffer from water-related diseases, increasingly attention is now being paid to the combination of water and sanitation. The social and behavioral science research clearly identifies the complex relationships among water, health, and inequality; less clear, however, is how to respond to the complex emergency of increased levels of water-related disease globally.

Practitioners say that global health is sexy – curing people from exotic diseases like Ebola, saving babies and young

children from untimely death due to malnutrition, or providing women with safe childbirth. But sanitation is not sexy. It is dirty, smells foul, and makes people uncomfortable. And yet, everyone engages in it. Some people are forced to defecate in the open; others defecate in streams or rivers or back alleys. At the level of the individual, defecation and other sanitation practices reflect personal, cultural, and religious constraints. At the level of the community, they are questions of political will and infrastructure – and in addition, creating and maintaining a system of waste management are expensive.

The problems that beset improving water, health, and social equality are not, however, intractable. Improvements in water and sanitation are occurring and increasingly occurring with attention to equity. The MDGs, like the 1999 Protocol, have focused attention to the need to rectify the global inequities in levels of health, access to water, and now increasingly, access to sanitation. And, while the future itself may not be clear, the road to improved global health is.

See also: Ecology and Health; Hazards and Disaster Research in Contemporary Anthropology; Health Politics and Policy; Health: Anthropological Aspects; Inequality: Comparative Aspects.

Bibliography

Ahlersa, Rhodante, Zwarteveenb, Margreet, 2009. The water question in feminism: water control and gender inequities in a neo-liberal era. Gender, Place & Culture 16 (4), 409–426.

Aiyer, Ananthakrishnan, 2007. The allure of the transnational: notes on some aspects of the political economy of water in India. Cultural Anthropology 22 (4), 640–658.

Allen, A., Davila, J.D., Hofmann, P., 2006a. The peri-urban water poor: citizens or consumers? Environment and Urbanization 18 (2), 333–351.

Allen, A., Davila, J.D., Hofmann, P., 2006b. Governance of Water and Sanitation Services for the Peri-Urban Poor. DFID, London, UK.

Alley, Kelly D., 2002. On the Banks of the Ganga: When Wastewater Meets a Sacred River. University of Michigan Press, Ann Arbor.

Aruquipa, Abraham, 2008. Enfocandose en el Saneamiento en Bolivia. Agua Latinoamerica 8 (3), 1–2.

Bell, Beverly, 2001. Walking on Fire: Haitian Women's Stories of Survival and Resistance. Cornell University Press, Ithaca, NJ.

Boelens, Rutgerd, Doornbos, Bernita, 2001. The battlefield of water rights: rule making amidst conflicting normative frameworks in the Ecuadorian highlands. Human Organization 60 (4), 343–355.

Briggs, Charles L., Mantini-Briggs, Clara, 2003. Stories in the Time of Cholera: Racial Profiling during a Medical Nightmare. University of California Press, Berkeley, CA.

Cairncross, Sandy, 2003. Editorial: water supply and sanitation: some misconceptions. The Lancet Infectious Diseases 3 (5), 275–281.

Cairncross, S., Valdmanis, V., 2006. Water supply, sanitation and hygiene promotion. In: Jamison, D.T., Breman, J.G., Measham, A.R. (Eds.), Disease Control Priorities in Developing Countries. The World Bank, Washington, DC, pp. 771–792.

Centers for Disease Control and Prevention (CDC), 2010. Update on cholera-Haiti, Dominican Republic, and Florida, 2010. Morbidity and Mortality Weekly Report 59 (50), 1637–1641.

Centers for Disease Control and Prevention (CDC), 2012. Parasites-Schistosomiasis. http://www.cdc.gov/parasites/schistosomiasis/ (last accessed 12.09.12.).

Cleaver, Frances, Hamada, Kristin, 2010. 'Good' water governance and gender equity: a troubled relationship. Gender & Development 18 (1), 27–41.

Coles, Anne, Wallace, Tine (Eds.), 2005. Gender, Water, and Development: Cross-Cultural Perspectives on Women. Gordonsville. Berg Publishers.

De Alburquerque, Caterina, 2009. Promotion and Protection of All Human Rights, Civil, Political, Economic, Social and Cultural Rights, Including the Right to Development. Human Rights Council. Report of the independent expert on the issue of human rights obligations related to access to safe drinking water and sanitation.

Derman, Bill, Ferguson, Anne, 2003. Value of water: political ecology and water reform in southern Africa. Human Organization 62 (3), 277–288.

Desowitz, Robert, 1987. New Guinea Tapeworms and Jewish Grandmothers: Tales of Parasites and People. W. W. Norton and Co, New York.

Dowell, Scott, Braden, Christopher, 2011. Implications of the introduction of cholera to Haiti. Emerging Infectious Diseases 17 (7), 1299–1300.

El Katsha, S., White, A.U., 1989. Women, water, and sanitation: household behavioral patterns in two Egyptian villages. Water International 4, 103–111.

Esrey, S., 1996. Water, waste, and well-being: a multicountry study. American Journal of Epidemiology 143 (6), 608–623.

Farmer, Paul, 1999. Infections and Inequalities: The Modern Plagues. University of California Press, Berkeley, CA.

Farmer, Paul, 2011. Haiti after the Earthquake. Public Affairs Books, New York.

International Organization for Migration (IOM), 2010. Significant Drop in Numbers of Internally Displaced in Camps in Haiti. Press briefing (10 December 2010). Accessed at: www.iom.int/jahia/Jahia/media/press-briefing-notes/pbnAM/cache/offonce?entryId=28848.

Ivers, Louise, Farmer, Paul, Alamazor, Charles P., Leandre, Fernet, 2010. Five complementary interventions to slow cholera: Haiti. The Lancet 376, 2048–2051.

Johnston, Barbara R., 2005. The commodification of water and the human dimensions of manufactured scarcity. In: Whiteford, Linda, Whiteford, Scott (Eds.), Globalization, Water, and Health: Resource Management in Times of Scarcity. School of American Research Press, Santa Fe, NM, pp. 133–152.

Kaussen, Valerie, 2011. Do it yourself: International Aid and the Neoliberal Ethos in the Tent Camps of Port-au-prince. NACLA Report on the Americas.

Kendall, Carl, 2005. Waste not, want not: grounded globalization and global lessons for water use from Lima, Peru. In: Whiteford, L., Whiteford, S. (Eds.), Globalization, Water, and Health: Resource Management in Times of Scarcity. School of American Research Press, Santa Fe, NM, pp. 85–106.

Malek, E.A., 1975. Effect of the Aswan High Dam on prevalence of schistosomiasis in Egypt. Tropical and Geographical Medicine 27 (4), 359–364.

McConville, Jennifer R., Mihelcic, James R., 2007. Adapting life-cycle thinking tools to evaluate project sustainability in international water and sanitation development work. Environmental Engineering Science 24 (7), 937–948.

Medecins sans Frontieres (MSF), 2012. Haiti: Cholera Returns to Port-au-Prince. Electronic document. http://www.msf.org.uk/articledetail.aspx?fId=Cholera_returns_to_Haiti_20120427 (accessed May 2012).

Mendoza, Consuelo, Meites, Elissa, Briere, Elizabeth, Gernay, Jacqueline, Morgan, Oliver, Monegro, Nelson Rodriguez, 2011. Preparing health care workers for a cholera epidemic, Dominican Republic, 2010. Emerging Infectious Diseases 17 (11), 2177–2178.

Mihelcic, J.R., Fry, L.M., Myre, E.A., Phillips, L.D., Barkdoll, B.D., 2009. Field Guide in Environmental Engineering for Development Workers: Water, Sanitation, Indoor Air. American Society of Civil Engineers (ASCE) Press, Reston, Virginia.

Montgomery, Maggie A., Eimelech, Menachem, 2007. Water and sanitation in developing countries: including health in the equation. Environmental Science and Technology 41, 16–24.

O'Loughlin, Rosalyn, Fentie, Gashu, Flannery, Brendan, Emerson, Paul M., 2006. Follow-up of a low cost latrine promotion programme in one district of Amhara, Ethiopia: characteristics of early adopters and non-adopters. Tropical Medicine and International Health 11 (9), 1406–1415.

Orlove, Ben, Caton, Steven C., 2010. Water sustainability: anthropological approaches and prospects. Annual Review of Anthropology 39, 401–415.

Pan American Health Organization (PAHO), 2011. Health Response to the Earthquake in Haiti. Electronic document. http://www.paho.org/English/DD/AIS/cp_332.htm (accessed April 2012).

PDNA, 2010. Haiti Earthquake PDNA: Assessment of Damages, Losses, Sectoral and General Needs.

Piarroux, Renaud, Barrais, Robert, Faucher, Benoit, Haus, Rachel, Piarroux, Martine, Gaudart, Jean, Magloire, Roc, Raoult, Didier, 2011. Understanding the cholera epidemic, Haiti. Emerging Infectious Diseases 17 (7), 1161–1168.

Roark, Donelly P., 2003. Women and water. In: Bourne, P.G. (Ed.), Water and Sanitation: Economic and Sociological Perspectives. Academic Press, Orlando, FL.

Rodgers, Ann F., Ajono, Lydia A., Gyapong, John O., Hagan, Maria, Emerson, Paul M., 2007. Characteristics of latrine promotion participants and non-participants:

inspection of latrines and perceptions of household latrines in northern Ghana. Tropical Medicine and International Health 12 (6), 772–782.
Schuller, Mark, 2010. Shattered and Scattered: Haiti's Quake through the Lens of Human Rights. NACLA Report on the Americas.
Schuller, Mark, 2012. Killing with Kindness: Haiti, International Aid, and NGOs. Rutgers University Press, New Brunswick, NJ.
Singer, Merill, 2012. Thinking about waste is not wasted thinking. Anthropology News 53 (8), 5.
Tuite, Ashleigh, Tien, Joseph, Eisenberg, Marisa, Earn, David, Ma, Junling, Fisman, David, 2011. Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions. Annals of Internal Medicine 154, 593–601.
United Nations Economic Commission for Europe (UNECE), 2000. Protocol on Water and Health to the 1992 Convention on the Protection and Use of Transboundary Watercourses and International Lakes. http://www.unece.org/fileadmin/DAM/env/documents/2000/wat/mp.wat.2000.1.e.pdf (last accessed 12.11.12.).
UNICEF and World Health Organization, 2012. Progress on Drinking Water and Sanitation 2012 Update. Electronic document. http://www.who.int/water_sanitation_health/publications/2012/jmp_report/en/index.html.
Wakeman, Wendy, 1995. Gender Issues Sourcebook for Water and Sanitation Projects: Indicators for Water Issues in Gender and Sanitation. International Bank for Reconstruction and Development/The World Bank, Washington, DC.
Waldman, Ronald, Williams, Holly A., 2001. Public health in complex emergencies: toward a more integrated science. In: Williams, Holy (Ed.), Caring for Those in Crisis: Integrating Anthropology and Public Health in Complex Humanitarian Emergencies. NAPA Bulletin, Arlington, VA, pp. 89–111.
Water Supply and Sanitation Collaborative Council (WSSCC), 2012. Resources Hurry Up!. http://www.wsscc.org/resources/resource-advocacy-materials/hurry#10 (last accessed 12.08.12.).
Watts, S., 1997. Epidemics and History: Disease, Power and Imperialism. Yale University Press, New Haven, CT.
Watts, S., 2004. Women, water management, and health. Emerging Infectious Diseases 10 (11), 2025–2026.
Webster, Paul C., 2011. Lack of clean water exacerbates cholera outbreak in Haiti. Canadian Medical Association Journal 183 (2), 83–84.
Whiteford, Linda M., 1999. Water insecurity and infectious disease. International Review of Comparative Public Policy 11, 63–82.
Whiteford, Linda, Whiteford, Scott, 2005. Paradigm change. In: Whiteford, Linda, Whiteford, Scott (Eds.), Globalization, Water, and Health: Resource Management in Times of Scarcity. School of American Research Press, Santa Fe, NM, pp. 3–15.
Whiteford, Linda M., Laspina, C., 1996a. Evaluation of Cholera Control Activities: The BACA Project. Environmental Health Project, Arlington, VA.
Whiteford, Linda M., Laspina, C., Torres, M., 1996b. Cholera Control in Ecuador: Behavior-based Control Activities. Environmental Health Project, CDM, Arlington, VA.
Whiteford, Linda M., Rylko-Bauer, B., Farmer, P. (Eds.), 2009. Global Health in Times of Violence. School for Advanced Research Press, Santa Fe, New Mexico.
Whiteford, Linda M., Vindrola-Padros, C., 2011. The medical anthropology of water. In: Singer, Merrill, Erickson, Pamela (Eds.), Companion to Medical Anthropology. Blackwell Publishers.
Whiteford, Linda M., Vindrola-Padros, C. The Structural Violence of Cholera. Left Coast Press, in press.
Whiteford, Scott, Cortez-Lara, Antonio, 2005. Good to the last drop: the political ecology of water and health on the border. In: Whiteford, Linda, Whiteford, Scott (Eds.), Globalization, Water, and Health: Resource Management in Times of Scarcity. School of American Research Press, Santa Fe, NM, pp. 231–254.
World Health Organization (WHO), 2004. World Health Report 2004 – Changing History. World Health Organization, Geneva.
World Health Organization (WHO), 2010. Cholera-Haiti update. Weekly Epidemiological Record 49 (85), 489–496.
World Health Organization (WHO), 2011. Cholera Country Profile: Haiti.
World Health Organization (WHO), 2012a. Burden of Disease and Cost-Effectiveness Estimates. http://www.who.int/water_sanitation_health/diseases/burden/en/index.html (last accessed 12.08.12.).
World Health Organization (WHO), 2012b. Global health observatory: number of reported cholera cases. http://www.who.int/gho/epidemic_diseases/cholera/cases_text/en/index.html (last accessed 12.10.12.).
World Health Organization (WHO), 2012c. Facts and Figures on Water Quality and Health. http://www.who.int/water_sanitation_health/facts_figures/en/index.html; whqlibdoc.who.int/publications/2008/9789241596435_eng.pdf (last accessed 10.08.12.).
Yacoob, May, Whiteford, Linda M., 1994. Behavior in water supply and sanitation. Human Organization 53 (4), 330–335.

Relevant Websites

http://www.cdc.gov/parasites/schistosomiasis – Centers for Disease Control (CDC).
http://www.fao.org/nr/water/ – FAO–Water.
www.iom.int/jahia/Jahia/media/press-briefing-notes/pbnAM/cache/offonce?entryId=28848 – International Organization for Migration (IOM).
http://www.iwahq.org/1nb/home.html – International Water Association.
http://www.iwmi.cgiar.org/ – International Water Management Institute.
http://www.msf.org.uk/articledetail.aspx?fld=Cholera_returns_to_Haiti_20120427 – Medecins sans Frontieres.
http://www.paho.org/English/DD/AIS/cp_332.htm – Panamerican Health Organization (PAHO) (accessed April 2012).
http://www.globalwaterforum.org/ – UNESCO Global Water Forum.
http://www.unwater.org/ – UNWater.
http://www.wsscc.org/resources/resource-advocacy-materials/hurry#10 – Water Supply and Sanitation Collaborative Council (WSSCC).
http://www.who.int/topics/water/en/ – World Health Organization-Water.
http://www.who.int/water_sanitation_health/publications/2012/jmp_report/en/index.html – World Health Organization-Water, Health, and Sanitation.
http://www.who.int/water_sanitation_health/diseases/burden/en/index.html – World Health Organization (WHO) – Burden of disease and cost-effectiveness estimates (last accessed 12.08.12.).
http://www.who.int/gho/epidemic_diseases/cholera/cases_text/en/index.html – World Health Organization (WHO) – Global health observatory: number of reported cholera cases (last accessed 12.10.12.).
http://www.who.int/water_sanitation_health/facts_figures/en/index.html – World Health Organization (WHO) – Facts and figures on water quality and health.
http://www.wri.org/ – World Resources Institute.
http://www.worldwatercouncil.org/ – World Water Council.

Water Resources and Sustainable Water Management

Jim L Wescoat, Jr., Aga Khan Program for Islamic Architecture, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Water resources generally focus on freshwaters of the world, which include precipitation falling as rain or snow, infiltrating into soils and groundwater aquifers, running off into stream networks, evaporating from lakes and transpiring from vegetation, and flowing back into the deltas and estuaries along the continental margins. The geographic distribution and flow of freshwaters (analyzed through water-budget studies) affect supplies available for human use, and associated problems of surplus, scarcity, and pollution. Psychological and behavioral research has focused on perception of these escalating water concerns about flood risks, drought hazards, and decision-making, while phenomenological research has explored deeper meanings of water experience. During the twentieth century, social scientific research on water evolved from description of regional water problems (e.g., in the Tennessee Valley Authority); to economic and institutional analyses of water rights, water markets, and the 'range of choice' among water management alternatives; and, more recently, to political and cultural investigations of water and related ecosystem management problems. These developments contributed in the late twentieth century toward ideals of sustainable and integrated water resources management, though those conceptions are now being challenged and refined. Influential debates on social aspects of water have involved Karl Wittfogel's arguments about 'hydraulic civilizations' that drew together the key variables of environmental conditions, water use technologies, economic patterns, property regimes, state organization, social classes, and modes of social power. These ideas stimulated intense theoretical and empirical debates, followed by applied research on sustainable water management in many regions of the world, which argues for continuing attention to comparative international research on water resources, including those of transnational and global scales.

Freshwaters of the World

In the social sciences, water resources research generally focuses on freshwaters of the world. Freshwater includes precipitation that falls over the continents as rain or snow, infiltrates into soils and groundwater aquifers, runs off into stream networks that drain the continents, evaporates from lakes and transpires from vegetation, and ultimately flows back into the deltas and estuaries along the continental margins. This framing of the hydrologic cycle has been central to concepts of sustainable water management, which are woven into the fabric of this article, and it has a fascinating if circumscribed history (Linton, 2010). For example, as coastal and marine resources have involved different types of uses and jurisdictions, they are usually treated separately. However, as populations continue to migrate toward urbanizing coasts, the freshwater–saline interface becomes increasingly important and problematic (e.g., saltwater intrusion, accelerated erosion, coupled runoff–storm surge, and sea level rise). Worldwide, coastal urbanization raises fundamental challenges for water research and sustainability; however, it may be defined. Advances in global environmental science may lead toward greater integration of fresh and saline water research.

Freshwaters constitute only 2.5% of the world's water resources, the bulk of which (96.5%) is in the oceans. The largest proportion of freshwaters is frozen in ice caps and glaciers (68.7%), followed by groundwater (30.1%), lakes (0.26%), the atmosphere (0.04%), marshes (0.03%), rivers (0.006%), and biological water (0.003%) (Shiklomanov, 1993; Trenberth et al., 2007). Biological waters are small in volume but have great social significance because they include the water in human bodies, rapid temporal cycling in plants and animals, and large-scale consumptive use by natural and cultivated vegetation. Research on the history and sociology of water physics and chemistry is questioning how water has been conceived, and in some cases argues for pluralistic water science (Chang, 2012; Linton, 2010).

In addition to their physiological and ecological importance, water resources have been central to fundamental debates about property, power, public interest, the role of the state, and sustainability of socio-hydrologic systems (Sivapalan et al., 2014). Those debates, in turn, have influenced the course of environmental science, technology, and law; and those fields have deepened understanding of the distribution and flows of water from one location, physical state, and quality to another.

Water Distribution and Flows

The static percentages of freshwater mentioned above must be qualified in important ways to understand their social significance. First, freshwaters cycle dynamically in space and time. For example, biological waters are replaced (through ingestion, perspiration, and excretion) on timescales of hours to days, while groundwater is only replaced (through discharge and recharge) on timescales of tens of thousands of years. These temporal cycles have profound implications for human water supply and pollution control. Although biological water needs do not vary much within species, they do vary over organismal life spans, and across geographic regions. The distribution of freshwater also varies by latitude and by climatic processes within latitudinal zones, which range from arid to humid and which vary on seasonal, annual, and longer timescales.

These variations have implications for human water use that are analyzed through water budget analysis (Wescoat, 2013). A water budget compares the amount of water supplied by

precipitation with the amount of water consumed by evaporation and transpiration from vegetation and crops – the largest use of water worldwide. The difference between precipitation and evapotranspiration helps define the magnitude and duration of surplus and scarcity, which societies address by acquiring supplemental water, draining excess waters, or moving to more favorable habitats. Groundwater management requires different approaches and in some regions of the world is so little regulated locally and regionally as to be characterized as 'anarchic' (Shah, 2009). To advance water use accounting, some researchers advocate separating 'blue' surface water from 'green' soil moisture water (Falkenmark and Rockström, 2010).

Freshwaters also vary in physical characteristics, chemical composition, and aquatic life. These characteristics reflect the broad climatic and physiographic contexts that shape the landscape of stream networks, groundwater aquifers, and wetlands. Variations in freshwater distribution, characteristics, and landscapes condition human water experience, which is the primary concern of this article (Gleick, 2012). People who live in snow-covered alpine watersheds have different water supplies, quality, and challenges from those living in temperate or tropical watersheds fed by rain. People living in downstream environments face different water opportunities and problems from those upstream. It is well known that downstream groups face water quantity and quality problems generated by upstream groups, but it is appreciated less widely that downstream groups often have the political and economic power to regulate and constrain upstream development, as in the Nile basin. Many engineering projects of the twentieth century failed to meet their intended goals or gain public support, in large part due to inadequate understanding of such basic social and behavioral aspects of water. Increasingly it is recognized that social research represents an important frontier for water resources management, though it remains underfunded relative to other water sciences (National Research Council, 2004).

Private research funding has addressed some flows, most notably those associated with product supply and value chain analysis. Corporations have found this important for logistics optimization, risk management, and brand protection (World Economic Forum, 2011). Research on virtual water trade has likewise developed from rough estimates of export and import embedded in production and transport to more sophisticated physical and economic modeling (Allan, 2011; Konar et al., 2012).

Psychological and Behavioral Aspects of Water

In addition to supporting life, water has distinctive psychological characteristics. At a physiological level, 'thirst' is experienced differently from 'hunger' and other needs (Ramsay and Booth, 1991). Psychological disorders (e.g., water phobias) include fears of immersion, swimming, and drowning. At a social level, perception of water-related risks and adjustments has received attention in behavioral research on flood and drought hazards, beginning in the 1960s. Flood and drought hazards research evolved from an early emphasis on description of disaster losses and recovery strategies to research that seeks to explain decisions to locate in floodplains (White, 1945; Lamond, 2011; National Research Council, 2000). That research concentrated initially on hazards perception, use of scientific forecasts and uncertainty in decision-making, and the behavioral dynamics of alternative policy interventions – research that continues with increasingly strong connections between theory, experimental research, and practice (e.g., Kunreuther and Roth, 1998; Morss et al., 2005). Droughts develop more slowly than floods and are less commonly addressed in disaster research; they involve different processes of perception, behavior, and adjustment (Wilhite, 2005). Similarly, water-related disease hazards range from vectors in aquatic habitats to inadequate water supply for washing, and vectors that breed in wet areas (Thompson, 2002; White et al., 1972).

Reexamination of pioneering studies of water resources and hazards designed to assess the effects of increased knowledge have noted limited progress in reducing economic losses (White et al., 2001). Efforts to develop general decision-making models to address water problems developed a key analytical construct, known as the 'range of choice,' which attributed failings to undue narrowing of the alternatives considered (Wescoat, 1992; Mitchell, 2008). Criticism of behavioral research on water has focused on structural constraints, especially in developing countries. These criticisms led to research on social vulnerability to water problems. Another small yet important line of phenomenological inquiry questions assumptions about human experience of water (Bachelard, 1983[1942]; Cosgrove and Petts, 1990).

Cultural Representations and Interpretations of Water

Social research on the representations of water (e.g., in texts, arts, and popular culture) is scattered across many fields. In the sphere of belief systems, it includes creation stories such as the *Enuma Elish* that begin with a watery chaos from which fundamental divisions emerge. Comparative studies of flood stories (Dundes, 1988), water deities and rituals, and paradise gardens have continuing significance for many cultures (Blair and Bloom, 2009).

In the modern era, numerous travel accounts proceed from source to sea on major rivers of the world. While tourists generally travel downstream, 'imperial' explorers often labored against the current, searching out navigable trade routes and resources (e.g., Rennell on the Ganges and Lewis and Clark on the Missouri River). Some literary works, such as Henry David Thoreau's *Walden*, influenced contemporary and subsequent views of water. There has been less research on water in the visual arts, music, and film, though that is changing. Ethnographic research on water communities and sustainability issues is increasing dramatically, some of it linking historical and ethnographic research (Mosse and Sivan, 2006; Orlove and Caton, 2010).

Water as an Environmental Resource

Questions about managing water as a natural resource, and about whether 'water is different' from other resources, have been central concerns of social scientists. They are pursued in close association with engineering, natural sciences, and law.

Water management has been dominated by engineering sciences, which have addressed social ideas and problems implicitly if not explicitly from antiquity to the present. Debates on the origin of rivers were resolved with the development of watershed science in seventeenth-century France, which made it possible to estimate water supplies. The expansion of irrigation agriculture also advanced with the development of hydroclimatic water budget analysis. Water pollution control evolved with advances in instruments to measure organic and inorganic contaminants.

This is not to say that traditional water management is unsophisticated, impractical, or ineffective. Humoral theories that originated in Greek and Islamic sciences are still practiced effectively in rural South Asia. Hindu temple tanks and modern irrigation works operate in parallel, though not always in coordination, with one another (Lansing, 1991). The contrast between traditional and modern breaks down at many points. Some traditional irrigation societies devised highly sophisticated water diversion and measurement systems that required elaborate forms of social organization (e.g., *qanats* in Persia). Many modern water systems, by contrast, are supply driven with little measurement or management of water. Irrigation efficiencies in some areas of the western US are only 50%, compared with rates of 90% in Israel and Jordan.

The most acute water problems occur in low-income areas subject to civil unrest, state corruption, weak water institutions, and highly skewed tenure regimes. These conditions have arisen in many irrigation systems and peri-urban areas of developing countries. In such situations, the water institutions that do exist are weakly or inequitably implemented.

Laws define aspects of water that society regards as a resource. From antiquity to the present, legal systems have defined certain aspects of water as natural rights, others as social rights, and still others as *res nullius*. Legal struggles reflect changing social values and water demands. Social research on water laws has focused on four major themes: (1) access rights; (2) water rights transfers; (3) institutional fragmentation; and (4) transboundary conflicts. Research on access rights asks who (including animals) has rights to use water, and what are the social and environmental consequences of different access regimes? Research on water transfers has focused on the impacts of property rights transfers from one type or location of water use to another (Saliba and Bush, 1987). Research on institutions has highlighted fragmentation, gaps, and redundancies among water organizations; and it has more recently begun to focus on power relations among organizations and their social and environmental consequences (Emel and Roberts, 1995). Transboundary conflicts encompass water pollution and floods, as well as water scarcity. During the 1990s, a large body of research focused on transboundary water problems in the Middle East (Committee on Sustainable Water Supplies for the Middle East, 1999).

It is not clear what influence social research has had on water laws to date. The clearest examples involve social research on local irrigation institutions, which have led to the development, testing, and refinement of organizational rules and regulations (e.g., for water users associations; Ostrom, 1992). Social research is sometimes cited in water rights case law or commissioned in interstate and international negotiations, but its effect relative to other considerations is difficult to determine (Delli Priscoli and Wolf, 2009). The slow pace of legal reform has led social scientists to concentrate on public policies where legal issues may be at stake, but decisions will likely be made by water users. One of the oldest policy debates revolves around questions of whether 'water is different' from other resources. Most social scientists grant that water has unique physical qualities and uses but some, principally economists, have argued that water is no different from other resources. By this argument, water is a means to ends that has social values which can and must be compared with those of other resources if society and individuals are to make rational decisions to improve human welfare and environmental quality.

These arguments have stimulated detailed empirical research on the economics of water use. Some researchers have quantified the price elasticity of demand for residential water use while others have focused on agricultural water values (Baumann et al., 1998; Griffin, 2006). These sectoral studies were drawn together in models for optimization and multiobjective management of complex river basins in the 1960s. These models sought to calculate trade-offs among competing water uses for agriculture, flood protection, recreation, municipal, and industrial sectors.

The principal internal debate in these efforts pitted optimization models, based on neoclassical economic principles, against simulation models based on behavioral models of decision-making. After sobering recognition of the limited use of these models, and further technological advances, research in the 1990s began to develop decision support systems. That work built upon research in community-based mapping, human–computer interactions, river basin operations, and adaptive environmental assessment and management. More recently, environmental economists have sought to estimate 'nonuse' values of waters that remain in natural watercourses, the value of 'ecosystem services,' and the trade-offs among competing environmental objectives. However, a growing body of research challenges neoliberal water pricing, privatization, and valuation (Bakker, 2010).

As is evident in the examples above, water resources research has been more concerned with pragmatic inquiry, in the philosophical and colloquial senses of the term, than with metatheoretical debates in the social sciences (Wescoat, 1992). It strives to define the social dimensions of pressing water problems and evaluate the consequences of alternative courses of action. Early research focused on river channel problems (e.g., flooding, navigation, and pollution). The emphasis on channels, rather than basins, reflected legal constraints on the role of federal states in natural resources management. A concurrent line of research sought to describe river basin problems and resources. Basin surveys became increasingly important in the early twentieth century in Europe and the US. The Tennessee Valley Authority (TVA), authorized in 1933, sought to address problems of severe economic poverty, land degradation, and bureaucratic fragmentation. It included regional resource surveys, the design of planned communities, and economic development programs. Although the TVA experiment was not replicated in the US, it had growing significance in postwar development programs in many countries (e.g., the Damodar, Gal Oya, Indus, and Mekong river basins; though see Klingensmith, 2007, for a rigorous historical

critique of diffusion and borrowing). International law has adopted, conservatively and tentatively, the concept of a watercourse system that includes linkages among hydrologic processes, but stopping short of land use, land cover, and basin territories (McCaffrey, 2007).

Policy reforms in the 1930s were influenced by social research on water financing, pricing, and decision-making. The TVA framework was broadened to the scale of national water planning in the 1960s. These efforts culminated in the US with the 'Principles and Standards for Planning Water and Related Land Resources,' which established four evaluation criteria: national economic development, regional economic development, environmental impacts, and other social effects (e.g., community impacts). Economic research focused on the first account, while other social scientists focused on the other three accounts.

One line of opposition to these economic and multi-objective approaches argued that water is different from other resources and that it has unique importance in supporting life. These views were advanced both by water developers arguing for irrigation projects and by their environmentalist and anti-dam opponents. Indigenous values often conflict with all of these arguments and approaches, though oppositions between indigenous and modern values are sometimes presumed rather than studied closely (Boelens and Davila, 1998). Although arguments for the inherent natural or cultural value of water have sometimes been self-serving, they effected a major shift from water development to environmental management in the late twentieth century. In some cases, they have disputed the possibility of commensurating different water uses and values (Espeland, 1998). Policy stalemate is sometimes taken to be evidence of incommensurable values. Analysis of institutional problems, including conflict and stalemate, was a major research advance in the 1970s and 1980s. In earlier decades, water organizations increased in scale and scope to address increasingly complex water problems, but they often weakened local institutions. In response, some research focused on local irrigation organizations (Ostrom, 1992), while others assessed bureaucratic behavior and reform (Coward, 1980). As engineering projects failed to meet their planning objectives and had unanticipated social impacts, increasing attention was directed to 'institutional strengthening,' 'capacity building,' and 'participatory development' (Uphoff, 1992). After two decades of research and experimentation, the importance of institutional factors is well established, though the efficacy of institutional alternatives remains a high research priority.

Beyond these economic and institutional approaches, social scientists have examined an array of specific social impacts in ways that have redefined water resources management. Social research on large reservoirs has focused on displacement and resettlement of people whose homes and lands are flooded by reservoirs, examining issues of equity, state performance, and antidam movements that have stopped international aid for projects in India, Turkey, and China.

Recognition that women, especially in poor rural societies, spend disproportionate amounts of time and energy carrying water and managing domestic water and sanitation led to social research on water and gender. Development assistance, by contrast, often addressed men in those societies. Feminist research has shed light on women's understanding of water resources (e.g., Feldhaus, 1995; O'Reilly et al., 2009), and gender was one of four top priorities identified by the UN Commission for Environment and Development in 1992 for improving water management. Movements have organized around environmental problems in aquatic and riparian habitats, including species extinction, wetlands preservation, groundwater contamination, and river restoration. During the 1990s, a large number of watershed movements arose in the US to address issues ranging from nonpoint source pollution to riparian habitat protection and landscape identity. Social scientists have sought to understand how these movements articulate environmental values (intrinsic, instrumental, and inherent) with political strategies, organizational dynamics, and environmental consequences.

A growing line of research focuses on political aspects of water management that behaviorist approaches failed to explain. These political issues include acute water problems faced by the poor, corruption and violence in water administration, inequitable access to water, vulnerability to flood hazards, and devolution of state authority to nongovernmental organizations (Mustafa, 2013). This work entails a shift from institutions to governance and politics (Baviskar, 2007; Mollinga et al., 2010). Brown and Ingram (1987) have sought to explain problems faced by low-income ethnic minorities by contrasting 'community' and 'commodity' values in water (Laurie, 2007). A broader, but fundamentally flawed, political framework was put forward by Wittfogel (1957).

Water and Power: Wittfogel's Theory of Hydraulic Civilizations

Wittfogel argued that irrigation agriculture had a formative role in the development of complex bureaucratic social organization, initially in the large arid floodplain environments of Asia which constituted a distinctively 'Oriental mode of production' that has, throughout history, threatened democratic societies. This contorted web of arguments was refuted soundly on empirical and theoretical grounds. Archaeological research on early irrigation systems showed that large-scale irrigation agriculture often followed complex state formation and had no consistent relationships with 'despotism' as Wittfogel defined it.

Notwithstanding these flaws, Wittfogel's work had three lasting implications for social research on water resources. First, it provoked three waves of research that included leading social scientists. The first wave included archaeologists and social historians who challenged the historical argument (e.g., Butzer, 1976). The next wave of students focused on structures of power in modern irrigation systems (cf. Maass and Anderson, 1978; Worster, 1985). A third generation of students became involved in the applied institutional experiments in irrigation management (Coward, 1980). The latter have shed light on the scales of state–society relations in water management, from community-based systems to state infrastructure and regulation (Zimmerer, 2011). The theoretical contributions of each group were sharpened by their response to, and continuation of, debates provoked by Wittfogel. This history of ideas has been accompanied by a rich body of historical research on water management including international rivers (Tvedt, 2010).

Wittfogel's second legacy was to draw together structural aspects of water management that were previously neglected or treated separately. His framework encompassed environmental conditions, technologies, economic patterns, property regimes, state organization, social class, and modes of social power. Discerning the relationships among these structures of water management remains a major challenge for water researchers. Current theorizations of power go beyond violent conflict to probing analyses of power relations and conflict resolution (Delli Priscoli and Wolf, 2009; Mustafa, 2013; Tal and Rabbo, 2010; Zeitoun, 2011). Research on water sharing and cooperation is limited but highly significant (Kreutzmann, 2000).

Toward Sustainable Water Management

Much if not most contemporary writing on water management concerns its unsustainability. Water policy has developed on a largely separate path from environmental management (Wescoat and White, 2003). Sector-based management of water for agriculture, cities, forests, and environmental quality has fragmented attempts at integrated management. A large body of work seeks to address these problems through a series of impact assessment methods, ranging from simple water stress and water poverty indexes (Falkenmark et al., 1989; Sullivan, 2003) to water footprint analysis of different economic activities (Hoekstra, 2013). At a broader conceptual level, the model of integrated water resources management (IWRM) has been put forward as a sustainable approach from the early 1990s onward by international organizations such as the Global Water Partnership. IWRM has sought to link water sector planning with environmental management and social equity in development and has gained acceptance over the course of a decade, only to facing increasing criticism for being unimplementable in scope and apolitical in substance (Mollinga et al., 2010). The field of sustainability science offers some hope for ways forward, but the evidence tends to rely upon locally sustainable precedents within a context that political ecology defines as inherently global as well as local (e.g., Crate, 2011; Gleick and Palaniappan, 2010). Research that links agro-ecological landscapes and irrigation systems helps bridge the longstanding gap between water and environmental management (Zimmerer, 2012). One of the interesting research challenges is how regions learn from successes in other places and earlier times (Kenney, 2005). While flows of knowledge, technology, and social relations are enormous, they have rarely been addressed with formal comparative methods (Mollinga and Gondhalekar, 2012; Wescoat, 2009).

Conclusion

Of the natural resources, water seems especially prone to social experimentation. Its fluid characteristics make it less subject to private control than land and thus more open to collective control. It is also more subject to social control than diffuse resources like the atmosphere, which has eluded most purposeful human management efforts. Water resources thus lie at a vital intersection between public and private control. The historical interactions between public and private water management and social research on them have yielded insights of broader significance for the social and behavioral sciences.

One important observation on the twentieth century's water research concerns the changing significance of social research in different parts of the world. In the late nineteenth century, US and Australian water specialists traveled around the world for models and lessons of water development in Asia, the Middle East, and Europe. The subsequent dominance of US and European researchers was based on advances in the theory and practice of integrated river basin development. But the most creative social research on water management in the late twentieth century occurred in developing areas that were open to institutional experiments in irrigation, domestic water supply, and sanitation (Lansing, 1991; Shivakoti, 2005; Uphoff, 1992). Transboundary water research has tended to focus on Asian and Middle Eastern water problems. US and European research remained innovative in the fields of floodplain management, urban water management, water quality, and water resources effects of global climate change. An emerging field of socio-hydrology strives to encompass these dimensions of the sustainability of water resource systems.

These observations point toward a number of opportunities for future inquiry, including: (1) collaborative research on transnational water problems; (2) research on social scales of water control from the local to the global; (3) critical assessment of trends from public water management toward private market-based approaches, and post-neoliberal water management; and (4) research on experiments in science-based policy, citizen-science, and adaptive management of complex water ecosystems.

Bibliography

Allan, T., 2011. Virtual Water. I.B. Tauris, London.

Bachelard, G., 1983[1942]. Water and Dreams: An Essay on the Imagination of Matter (E.R. Farrell, Trans.). Pegasus Foundation, Dallas.

Bakker, K., 2010. Privatizing Water: Governance Failure and the World's Urban Water Crisis. Cornell Press, Ithaca.

Baumann, D., Boland, J., Hanemann, W.M., 1998. Urban Water Demand Management and Planning. McGraw-Hill, New York.

Baviskar, A. (Ed.), 2007. Waterscapes: The Cultural Politics of a Natural Resource. Permanent Black, Delhi.

Blair, S., Bloom, J., 2009. Rivers of Paradise: Water in Islamic Art and Culture. Yale University Press, New Haven.

Boelens, R., Davila, G. (Eds.), 1998. Searching for Equity. Conceptions of Justice and Equity in Peasant Irrigation. Van Gorcum, Assen.

Brown, F.L., Ingram, H., 1987. Water and Poverty in the Southwest. University of Arizona, Tucson.

Butzer, K.W., 1976. Early Hydraulic Civilization in Egypt: A Study in Cultural Ecology. University of Chicago, Chicago.

Chang, H., 2012. Is Water H_2O? Evidence, Realism and Pluralism. In: Boston Studies in the Philosophy of Science, vol. 293. Springer, Dordrecht.

Committee on Sustainable Water Supplies for the Middle East, 1999. Water for the Future: The West Bank and Gaza Strip, Israel, and Jordan. National Academy Press, Washington, DC.

Cosgrove, D., Petts, G. (Eds.), 1990. Water, Engineering and Landscape. Belhaven, London.

Coward, E.W. (Ed.), 1980. Irrigation and Agricultural Development in Asia. Cornell University Press, Ithaca.

Crate, S., 2011. A political ecology of 'water in mind': attributing perceptions in the era of global climate change. Weather, Climate and Society 3, 148–164.

Delli Priscoli, J., Wolf, A.T., 2009. Managing and Transforming Water Conflicts. Cambridge University Press, Cambridge.

Dundes, A. (Ed.), 1988. The Flood Myth. University of California Press, Berkeley, CA.

Emel, J., Roberts, R., 1995. Institutional form and its effect on environmental change: the case of groundwater in the Southern High Plains. Annals of the Association of American Geographers 85, 664–683.
Espeland, W.N., 1998. The Struggle for Water: Politics, Rationality, and Identity in the American Southwest. University of Chicago Press, Chicago.
Falkenmark, M., Rockström, J., 2010. Building water resilience in the face of global change: from a blue-only to a green-blue water approach to land-water management. Journal of Water Resources Planning and Management 136, 606–610.
Falkenmark, M., Lundquist, J., Widstrand, C., 1989. Macro-scale water scarcity requires micro-scale approaches: aspects of vulnerability in semi-arid development. Natural Resources Forum 13, 258–267.
Feldhaus, A., 1995. Water and womanhood: religious meanings of rivers in Maharashtra. Oxford University Press, Delhi.
Gleick, P. (Ed.), 2012. The World's Water: The Biennial Report on Freshwater Resources, vol. 7. Island Press, Washington, DC.
Gleick, P.H., Palaniappan, M., 2010. Peak water limits to freshwater withdrawal and use. PNAS 107, 11155–11162.
Griffin, R.C., 2006. Water Resource Economics: The Analysis of Scarcity, Policies, and Projects. MIT Press, Cambridge.
Hoekstra, A.Y., 2013. The Water Footprint of Modern Consumer Society. Routledge, London.
Islam, S., Susskind, L.E., 2013. Water Diplomacy: A Negotiated Approach to Managing Complex Water Networks. RFF Press, Washington, DC.
Kenney, D.S. (Ed.), 2005. Search of Sustainable Water Management: International Lessons for the American West and Beyond. Edgar Elgar, Cheltenham.
Klingensmith, D., 2007. One Valley and a Thousand: Dams, Nationalism, and Development. Oxford University Press, Delhi.
Konar, M., Dalin, C., Hanasaki, N., Rinaldo, A., Rodriguez-Iturbe, I., 2012. Temporal dynamics of blue and green virtual water trade networks. Water Resources Research 48, 1–11.
Kreutzmann, H., 2000. Sharing Water: Irrigation and Water Management in the Hindukush, Karakoram, Himalaya. Oxford University Press, Karachi.
Kunreuther, H., Roth, R.J., 1998. Paying the Price: The Status and Role of Insurance against Natural Disasters in the United States. Joseph Henry Press, Washington, DC.
Lamond, J., 2011. Flood Hazards: Impacts and Responses for the Built Environment. CRC Press, Boca Raton.
Lansing, S., 1991. Priests and Programmers: Technologies of Power in the Engineered Landscape of Bali. Princeton University Press, Princeton, NJ.
Laurie, N. (Ed.), 2007. Pro-poor water? The privatisation and global poverty debate. Geoforum 38, 753–1052. (special issue).
Linton, J., 2010. What is Water? The History of a Modern Abstraction. UBC Press, Vancouver.
Maass, A., Anderson, R., 1978. … And the Desert Shall Rejoice: Conflict, Growth, and Justice in Arid Environments. MIT Press, Cambridge, MA.
McCaffrey, S.C., 2007. The Law of International Watercourses, second ed. Oxford University Press, Oxford.
Mitchell, J.K., 2008. Perspectives on alternatives: differentiation and integration in pursuit of a better fit between society and nature. Progress in Human Geography 32, 451–458.
Mosse, D., Sivan, M., 2006. The Rule of Water: Statecraft, Ecology and Collective Action in South India. Oxford University Press, Delhi.
Mollinga, P., Bhatt, A., Saravanan, V., 2010. When Policy Meets Reality: Political Dynamics and the Practice of Integration in Water Resources Management Reform. LIT Verlag, Berlin.
Mollinga, P.P., Gondhalekar, D., 2012. Theorising Structured Diversity: An Approach to Comparative Research on Water. Resources Management. ICCWaDS Working Paper No. 1. SOAS, London.
Morss, R.E., Wilhelmi, O.V., Downton, M.W., Gruntfest, E., 2005. Flood risk, uncertainty, and scientific information for decision making: lessons from an interdisciplinary project. Bulletin of the American Meteorological Society 86, 1593–1601.
Mustafa, D., 2013. Water Resource Management in a Vulnerable World: The Hydro-hazardscapes of Climate Change. I.B. Tauris, London.
National Research Council, 2000. Risk Analysis and Uncertainty in Flood Damage Reduction Studies. National Academies Press, Washington.
National Research Council, 2004. Confronting the Nation's Water Problems: The Role of Research. National Academies Press, Washington.
O'Reilly, K., Laurie, N., Halvorson, S., 2009. Gender, place, and culture. Special Issue on Gender Geographies of Water 16, 381 ff.
Orlove, B., Caton, S., 2010. Water sustainability: anthropological approaches and prospects. Annual Review of Anthropology 39, 401–415.
Ostrom, E., 1992. Crafting Institutions for Self-governing Irrigation Systems. ICS Press, San Francisco.
Ramsay, D.J., Booth, D. (Eds.), 1991. Thirst: Physiological and Psychological Aspects. Springer, Dordrecht.
Saliba, B.C., Bush, D.B., 1987. Water Markets in Theory and Practice: Market Transfers, Water Values and Public Policy. Westview Press, Boulder, CO.
Shah, T., 2009. Taming the Anarchy: Groundwater Governance in South Asia. RFF Press, Washington, DC.
Shiklomanov, I., 1993. World fresh water resources. In: Gleick, P. (Ed.), Water in Crisis: A Guide to the World's Fresh Water Resources. Oxford University Press, New York, pp. 13–24.
Shivakoti, G.P. (Ed.), 2005. Asian Irrigation in Transition: Responding to Challenges. Sage, Delhi.
Sivapalan, M., Konar, M., Srinivasan, V., Chhatre, A., Wutich, A., Scott, C., Wescoat, J., Rodríguez-Iturbe, I., 2014. Socio-hydrology: use-inspired water sustainability science for the Anthropocene. Earth's Future 2. http://dx.doi.org/10.1002/2013EF000164.
Sullivan, C.A., 2003. The water poverty index: development and application at the community scale. Natural Resources Forum 27, 189–199.
Tal, A., Rabbo, A., 2010. Water wisdom: preparing the groundwork for cooperative and sustainable water management in the Middle East. Rutgers University Press, New Brunswick.
Thompson, J. (Ed.), 2002. Drawers of Water II. International Institute for Environment and Development, London.
Trenberth, K.E., Smith, L., Qian, T., Dai, A., Fasullo, J., 2007. Estimates of the global water budget and its annual cycle using observational and model data. Journal of Hydrometeorology 8, 758–769.
Tvedt, T. (Ed.), 2010. A History of Water. Series II, 3 vols. I.B. Tauris, London.
Uphoff, N., 1992. Learning from Gal Oya: Possibilities for Participatory Development and Post-Newtonian Social Science. Cornell University Press, Ithaca, NY.
Wescoat Jr., J.L., 1992. Common themes in the work of Gilbert White and John Dewey: a pragmatic appraisal. Annals of the Association of American Geographers 82, 587–607.
Wescoat Jr., J.L., 2009. Comparative international water research. Journal of Contemporary Water Research & Education 142, 1–6.
Wescoat Jr., J.L., 2013. Water-conserving design: contributions of water budget analysis in arid and semi-arid regions. In: Chouni, A., Margolis, L. (Eds.), Out of Water. Birkhauser, Berlin.
Wescoat Jr., J.L., White, G.F., 2003. Water for Life: Water Management and Environmental Policy. Cambridge University Press, Cambridge.
White, G.F., 1945. Human Adjustment to Floods. Research Paper No. 29. Department of Geography, University of Chicago, Chicago.
White, G.F., Kates, R.W., Burton, I., 2001. Knowing better and losing even more: the use of knowledge in hazards management. Environmental Hazards 3, 81–92.
White, G.F., Bradley, D., White, A.U., 1972. Drawers of Water: Domestic Water Use in East Africa. University of Chicago Press, Chicago.
Wilhite, D., 2005. Drought and Water Crises: Science, Technology, and Management Issues. CRC Press, Boca Raton.
Wittfogel, K., 1957. Oriental Despotism: A Comparative Study of Total Power. Yale University Press, New Haven.
World Economic Forum, 2011. Water Security: The Water–Food–Energy Climate Nexus. Island Press, Washington, DC.
Worster, D., 1985. Rivers of Empire: Water, Aridity, and the Growth of the American West. Oxford University Press, New York.
Zeitoun, M., 2011. Power and Water in the Middle East: The Hidden Politics of the Palestinian–Israeli Water Conflict. I.B. Tauris, London.
Zimmerer, K., 2011. "Spatial-geographic models of water scarcity and supply in irrigation engineering and management (Bolivia, 1952–2009). In: Turner, M.D., Goldman, M., Nadasdy, P. (Eds.), Knowing Nature, Transforming Ecologies: Science, power, and practice in environmental science and management. University of Chicago Press, Chicago, pp. 167–185.
Zimmerer, K., 2012. Agrobiodiversity and water resources in agricultural landscape evolution. In: Gepts, D., Famula, T.R., Bettinger, R.L., Brush, S.B. (Eds.), Biodiversity in Agriculture: Domestication, Evolution, and Sustainability. Cambridge University Press. Cambridge, pp. 465–473.

Water Wars

Ashok Swain, Uppsala University, Uppsala, Sweden

© 2015 Elsevier Ltd. All rights reserved.

Abstract

It is often said that future wars will be fought over water. These water wars, defined as armed conflicts between two or more states over the scarce water resources, are predicted to take place over the sharing of international rivers. Sharing water is expected not only to increase competition and conflict, but also to contribute to build engagement and cooperation. In recent years, many water treaties have been negotiated and signed. However, these treaties face danger to their survival as allotted water in the existing sharing agreements in most of the cases is unable to meet the increasing demand. Possible impacts of global climate change have also brought further uncertainties to the peaceful sharing of scarce water resources.

Growing Water Scarcity

Water is critical for human survival, economic development, and the environment. Certainly, few other resources affect so many areas of the economy or human and environmental health. Seventy-five percent of the earth's surface is covered by water; however, most of it is saltwater, and much of the remaining quantity is stored in ice caps, glaciers, underground, within soil, in the atmosphere, and in living beings. Excluding lakes, only about 2000 cubic km of fresh water, found principally in rivers, is available for human consumption (Falkenmark, 1990).

More than 80% of this total global run-off is concentrated in the northern temperate zone, which hosts a small portion of the world's population (Postel, 2000). In the tropical and arid areas, where most of the population lives, the remaining water run-off resources are also distributed unevenly. Almost all of the developing countries are in the arid, semiarid, and tropical regions; many of them are facing severe water shortages. The world's population is now increasing by about 78 million people every year, and 95% of this growth is taking place in developing countries of Africa, Asia, and Latin America. This high population growth in the developing countries has multiplied pressure on fresh water resources (Swain, 2001). This problem is further exacerbated in these regions by their efforts toward rapid industrialization, massive urbanization, and agricultural intensification.

Developing countries are increasingly meeting growing water demand by building large dams for water storage, using a canal to divert water from one area to another or extracting ground water. The requirement for hydroenergy and commercial fishing has also contributed toward human intervention in water. Dam building, which has almost become obsolete in North America and Western Europe, is still considered the panacea for water shortage problems in many developing countries. However, dam projects submerge vast areas of land and forest and displace their inhabitants. There are millions of people who have lost their homes and livelihoods due to these large water projects (Swain, 2010).

The developing countries are primarily agricultural economies. To provide food to their growing population and also to achieve food security, these countries use proportionately more water on the agricultural sector than the industrial sector. Many developing countries already face serious problems in meeting rapidly increasing water demands of their population. In many developing countries, urban centers are rapidly increasing. Moreover, some developing countries are also industrializing faster. The expansion of urban areas and industrial sector has further added to the increasing water demand. With greater pressure being placed on the scarce water resources, overexploitation has resulted in acute shortages. Faced with such scarcity, water has increasingly become a source of social tension, bringing further competition and creating conflict within and between countries. However, these water conflicts not yet transformed to water wars, defined as armed conflicts between two or more states over the scarce water resources (Wolf, 1998). For a detailed analysis on conflict structure, please see the article Conflict: Organizational.

Water Scarcity and Conflict

The link between water resources and conflicts can be investigated in at least two different dimensions. First, in a conflict, the deliberate targeting of water storage facilities may be directly responsible for inducing water scarcity or reducing the water quality of the opponent. Thus, water scarcity becomes part of a military strategy and military behavior. Water supply is affected by conflicts, and fresh water resources also have the potential to cause or contribute to the emergence and/or escalation of conflicts among states or human groups.

Dispute over river water sharing usually comes up among the riparian states on three grounds: quantity, quality, and control. The incompatibilities of the last two issues (quality and control) are relatively easier to address with some financial and technical support (Swain, 2001). The quality issue, which had been the cause of disagreement among the riparian states in Europe's Rhine and Danube and North America's Colorado River in the past, has resulted in peaceful and cooperative arrangements. The disagreement over control of the Columbia River and Parana River in the relatively water abundant Americas has been settled for some time.

Water is not easily replaced; so, the problem of its reduced quantity is more difficult to address. The quantity factor in many cases threatens to destroy existing cooperative arrangements and forces the parties to take conflicting positions.

Global water consumption is rising steeply, and the lack of adequate supplies of water is a problem in many parts of the world. Water tables are falling increasingly in every continent. Several countries, most of them in the South, already face serious problems in meeting rapidly increasing water demand. Rivers are one of the most important sources of fresh water for human consumption. In scarcity situation, river water has increasingly become a source of tension as users are worried about the present or future availability of water supplies. Conflicts over water can be observed at different levels of society. Water issues can create new conflicting groups within a state, and in other cases, infuse incompatibility among the states. Even though water disputes are omnipresent they tend to become more complex and difficult when they concern international rivers.

Managing Shared Water Systems

According to the Trans-boundary Freshwater Dispute Database of Oregon State University, approximately half of the global fresh water is available through international basins in the world (UNEP, 2002). Overall, 145 countries have territories that include at least one shared river basin. Water has been frequently called the oil of the twenty-first century. Many believe that the dependence of these poor countries on an external water supply may force them to reorientate their national security concerns in order to protect or to preserve such availability. Such potential for conflict has brought global water issues into the arena of 'high politics.' United Nations (UN) officials and World Bank analysts regularly proclaim that "the previous war was about oil, the next war will be about water" (Morrissette and Borer, 2004).

Several countries are currently in dispute over the sharing of their common rivers. Some on-going international rivers inducing conflict are the Jordan, Nile, Euphrates–Tigris, Danube, Indus, and Ganges. With the exception of the Jordan basin (Cooley, 1984: 3–4), most international water conflicts have not led to physical violence, although the threat of the use of arms in these cases is not uncommon. As early as the mid-1980s, US intelligence services estimated there were at least 10 places in the world where war could break out over the shortage of fresh water supplies, with the majority located in the Middle East (Starr, 1991: 17).

Most developing countries require financial and technical aid and assistance to undertake large water projects. Very few countries can undertake expensive water projects on their own, coming at a heavy economic and political price (e.g., GAP project in Turkey, Three Gorges Project in China, and Narmada Project in India). In the post-Cold War period, it has become increasingly difficult to receive external support for a disputed project in an international basin. The end of the Cold War also stopped the alternative source of borrowing from the Eastern Bloc.

Water scarcity has caused a few minor skirmishes but no war has yet been fought between states. However, wars are very rarely fought over one issue. So, establishing water as the sole reason to cause violent armed conflict between two nation-states is not that easy. Water might have played a contributing role in a war through its contribution to food scarcity, population displacement, and ethnic alignment, which can lead to internal disturbances and political instability resulting in war. So, the real contribution of water scarcity to a war may not be properly examined through a conflict mapping data set, which only captures immediate causes. In-depth studies of individual wars might reveal the real contribution water scarcity has on instigating wars in different parts of the world.

Shared water is expected not only to increase competition and conflict, but also to contribute to building engagement and cooperation among riparian states. Due to mutual dependence, the withdrawal or pollution of river water of one riparian state can potentially not only lead to disputes, but also bring cooperation in the basin. Particularly in the last two decades, several competing riparian countries have moved toward establishing regimes and institutions for cooperation. Shira B. Yoffe and Aaron T. Wolf (1999) count the signing of 145 water-related treaties in the last century. There have been a number of very successful cases of cooperation among riparian countries that address pollution and management issues of their shared waters. The agreements among the riparian countries of the River Rhine, Colorado, and Parana are some of these examples.

Agreements on international rivers have not been limited to addressing water quality or management issues. In the last two decades, several international river basins have also witnessed a trend toward reaching agreements on quantity allocation as well. Competing riparian countries of the Mekong, Jordan, Ganges, Nile, and Zambezi Rivers have signed sharing arrangements in the 1990s. The signing of the agreements on these important rivers in conflict-prone regions has been regularly referred to downplay the possibilities of 'water war' scenarios.

River Water Sharing Agreements

An agreement can be more likely among the contending riparian states over the quantity allocation of a river resource, when there is enough unused water left in the river. Agreement on the Indus River system became a possibility in 1960 between two traditional rivals, India and Pakistan, because nearly 80% of the river water was running into the Arabian Sea, unused by both basin countries. One year before the Indus Agreement, another agreement on the sharing of the Nile River was reached between Egypt and Sudan. The 1959 Agreement become a possibility since a large amount of the run-off remained unallocated from the 1929 Agreement. However, with increasing water demand upstream and less availability of unused water, these rivers have already become a source of serious tension among the major riparian countries. The increasing riparian demand has also raised doubts about the continuation of the existing water sharing arrangements on the Euphrates–Tigris river system. The hope for further exploitation has not only enabled agreements for the Indus, Nile, or Euphrates–Tigris Rivers in the past but also facilitated agreements in recent years. Bangladesh signed the 1996 treaty with India with the hope to build a barrage on the Ganges at Pangsha, downstream of Farakka in Bangladesh. The 1995 Agreement signed among the lower Mekong basin countries became a possibility as the slow-flowing Mekong River

provides a lot of potential for further exploitation. The Zambezi river basin is another example of riparian cooperation based on the hope for the further exploitation.

Water-scarce riparian states have been able to find cooperative solutions in the last two decades rather than enter into violent conflict. Signing of agreements on water sharing may be easy, but the real problem is how to keep the agreement viable. The compliance part poses real challenges. The agreement needs to stand the test of time. Many agreements in recent years have been reached about how the water should be shared. In spite of reaching agreement, riparian discontent has not dissipated as many upstream countries believe they should have complete control over the flow of the rivers and withdraw water according to their demands. In some cases, where the downstream states are often more powerful in economic and military terms, they challenge upstream rights over the river flow, like Egypt, India, and Israel.

Global Climate Change and Uncertainties

Global climate change brings further uncertainties to the smooth functioning, even survival, of these recent international water agreements. With increasing temperatures and rapidly melting glaciers, lesser water supplies will be available to farms and cities during summer months when irrigation demand is high. Some parts of the globe may experience sizable reductions in precipitation, or significant changes in the timing of summer and rainy seasons. Climate change will increase supply side pressure for river water management, and global warming may also contribute to the demand side pressure due to increased demands in domestic, irrigation, industrial, and ecological use (Arnell, 1999).

As climate change can potentially affect water supply and demand patterns, sharing of scarce water resources in the arid and semiarid regions will become the most likely security challenges in the near future. Climate science has been able to provide a basic understanding of how the hydrological cycle will change at the global level, but the predictions of water demand and supplies at the regional and basin level is still far from reaching any consensus. The projected impacts of global climate change over fresh water may be huge and dramatic, but they will not take shape on the same scale in every geographical region. Even within an international river basin, the effects will vary depending on the location. This further enhances the uncertainties and anxieties over the water availability in the shared river systems. In this context, both domestic and international water laws and policies are not well equipped to meet the challenges posed by the global climate change (Eckstein, 2010).

Existing water sharing arrangements between the riparian countries of international rivers in most cases provide some mechanisms to adjust to the run-off variability while agreeing on allocation of fixed quota of water. Of 145 river agreements signed in the last century, approximately 37% dealt with quantity allocations (Wolf and Hamner, 2000). Variability of river water flows can very well reduce the longevity of these river agreements (Drieschova et al., 2009). Usually, the regular water sharing agreements tend to be based upon the assumption that any resulting shortages will be for a short duration only and that they can address the issue with temporary reallocation methods (Tarlock, 1999–2000). However, climate change can not only bring long-term increases or decreases to the average run-off of the river system, but also influence the variability of those flows that require flexibility to be the part of the water sharing framework to cope with emerging situation.

As global climate change brings long-term changes to the volume and pattern of run-off in shared river systems, it becomes crucial to examine the suitability of existing agreements to address this challenge. Climate-related changes might require comprehensive adjustments in the on-going water management structure of international rivers. This comprehensive effort might ask for the water sharing arrangements to be flexible and competent in allocating reduced and surplus water flow, maintaining certain water quality level, sustaining ecosystems, controlling flood, and protecting existing water development infrastructures. Thus, the river sharing arrangements need to have the provisions for information sharing, conflict management mechanisms, flexibility to adjust to the uncertainties, and endeavor for basin-based development strategy.

Basin countries must be under obligation to regularly exchange data and information among each other to monitor and manage changing conditions affecting shared water. In case of any dispute or disagreement over shared water management, there must be provisions available to basin countries to manage them as soon as possible. International river water management regimes and institutions require a flexible mandate to plan, operate, and implement, in order to cope with changing climatic conditions. The emerging unprecedented situation due to changes in climatic patterns requires basin countries to cooperate and act collectively (Swain, 2012).

There is no doubt that climate change poses extreme challenges to water resource management in international river basins in the South. Maarten De Wit and Jacek Stankiewicz (2006) demonstrate the dramatic potential effects of relatively small changes in rainfall due to climate change over the perennial drainage of the river. Moreover, climate change might cause extreme weather events, water shortages, changing sea levels, or melting glaciers that can generate serious threats to critical river water management infrastructure. While the importance of adjustment of flow variability in water sharing is crucial, many of the existing provisions within agreements are not adequate enough to meet the scenarios that global climate change models project. They lack enforcement and generally depend upon 'ideal' riparian behavior in case of eventuality. However, this approach may overcome run-off deficits in the short term, but climate change poses the risk of long-term flow reduction that would severely test existing provisions.

To reach an agreement that meets all competing and fluctuating demands for water in an international basin is in fact a very difficult task. Hydro diplomacy thus needs to adopt a total resource view where river water is seen as a key input for development and growth in the basin. The challenges are limited not only to technical and economic sectors, but also to crucial water sector reforms, which is political by nature. Moreover, the task of hydro diplomacy will no longer be limited to basin-based regimes and institutions; it must achieve effective water management in the face of climate change and must influence the supporting pathways from local, national,

and international policies and practices. In the past, river sharing matters could be effectively covered by a few negotiators trained specifically to deal with water issues. But today, hydro diplomacy needs not only to involve itself in an increasing range of fields (such as energy generation, food production, human rights, and health issues) but also to take into account the possible impacts of climate change (such as precipitation pattern, glacier melting, temperature increase, and rising sea water encroachment on fresh water systems).

Water Resource Governance and Managing Water Scarcity

There have been numerous endeavors to establish and strengthen international institutions and create an international legal framework for the management of international rivers. Global initiatives on the matter of fresh water have brought the international river sharing problem to the fore. International community is engaged in an on-going process toward establishing a common legal framework for the sharing of international watercourses at the global level.

In the first part of this century, the Territorial Sovereignty Doctrine (absolute sovereignty over waters flowing within a country) and the Natural Water Flow Approach (the river belongs to all the riparian states) came up in addressing the issues over the sharing of international rivers. Unfortunately, neither the territorial sovereignty nor the natural water flow approach provided a solution as they were based on an individualistic and anarchical conception of international law. The failure of these two legal approaches led some to think of sharing the rivers on an economic basis. According to this approach, the whole river basin is regarded as an economic unit irrespective of state boundaries and the waters are vested in the community of the users or divided among the co-riparian states. This joint approach includes joint planning, joint construction, joint management, and sharing of expenditure on construction and maintenance.

The idea of a single basin approach is attractive to economists and water engineers because it allows them to consider the international rivers as single hydrological units and plan accordingly. However, there are many difficulties involved in sorting out the externalities among the various riparian nations. The regulation and management of international river basins with so much concentration of power in the hands of nonpolitical commissions is an exception rather than the rule in the interstate practice. Owing to obvious limitations of states actually agreeing to joint development, not many examples are found employing this approach.

All these three above approaches were adopted and implemented in individual cases and due to their various limitations were not feasible for international practice. In the absence of any law to regulate international river systems, the International Law Commission after about 25 years of deliberations has submitted its draft in 1996 for the consideration of the UN General Assembly. Finally, the Convention on the Law of the Non-Navigational Uses of International Watercourses, adopted by the UN General Assembly on 21 May 1997, was submitted to the member states for their ratification. However, the process of ratification is moving along at a very slow pace.

By having both the principles of 'equitable utilization' and 'not to cause significant harm' in the text, the UN Convention was able to obtain a majority support in the UN General Assembly. Stephen McCaffrey describes the Convention as "a basket of Halloween candy: there is something in it for everyone" (McCaffrey, 1998). Even if this Convention is ratified by member states and becomes a legal framework, it will not be sufficient to address the problem of water sharing in different parts of the world. The sharing of international rivers among the riparian countries in different geographical regions is a problem of huge magnitude. Complex water disputes can only be solved by cooperation and compromise, not by a strict insistence on rules of law.

For successful and lasting cooperation on shared waters, there is need for a comprehensive approach to address the water scarcity issue. This comprehensive approach includes a number of measures to be taken at the basin level. The basin-focused measures include treating the river system as a single unit, involvement of both state and nonstate actors in water management, recognition of social and cultural contexts in water use, clear appropriation rules in water sharing, and an information sharing network among the riparian countries (Swain, 1999).

An international river, lake, or aquifer does not, by definition, respect national boundaries and as such, shared water systems should be treated as single units as regards the maximum utilization of their resources. The development of these shared water resources occurs most optimally at the basin-wide level and the whole basin should be regarded as one economic, ecological, and political unit irrespective of state boundaries. The regulation and management of basin organizations should be entrusted to an independent body, which is outside the political control of any single riparian state. Management of international freshwater systems should grow beyond the sphere of national sovereignty to achieve the best possible use of water, i.e., efficiency, equity, and environmental sustainability.

Not only states, but also nonstate water users must be eligible to participate as decision makers in the basin-based organizations. The sustainable use of fresh water requires user participation in all aspects of water policy and management in the basin. In order to construct sustainable basin-based water management institutions, contextual considerations are of the utmost importance. Existing traditions of rain water harvesting, water storing practices, and agricultural patterns are some of the issues to be taken into particular consideration while formulating basin management policy. It is necessary to have a set of clear rules and regulatory measures in the basin regarding water rights and environmental obligation. Basic needs for water must be identified and given priority. In several cases, riparian countries have unequal access to data and information due to differing data accessibility and asymmetric competence to process data. This asymmetric information can be scientific and/or strategic. For the smooth running of a river basin management regime, a functional information-sharing framework is required.

These basin-based initiatives need to be augmented and supported by various nation-state and international measures. Most of the developing countries are exposed to water stress or even water scarcity. The adoption of a supply management

strategy addressing only water shortage in the region is not nearly sufficient. To meet growing demands there is a need to minimize water use, particularly in the agricultural sector. Riparian states may opt for a planned allocation of agricultural activities to improve the productivity of water in their various regions in order to meet the future demand for food. There is a need to restrict and regularize the demand for the increasingly scarce water resources in the basin. The full-cost pricing of the water will create quantity restrictions for competing users. It will also force consumers to use water more efficiently than if there were no price tag on it or if it were available at a highly subsidized price. (For a detailed description of good water management, please see the article Water Resources and Sustainable Water Management).

External intervention and assistance can sometimes facilitate the negotiation of water resource sharing agreements. Riparian countries in the South are unable to establish institutional cooperative arrangements because of their concern regarding existing and future water rights. Mutual suspicion and uncertainties of reciprocal action obstruct constructive engagement. To overcome such obstacles, international actors can possibly provide credible and impartial international assistance to start the process of cooperation. Gradually, it could help to increase mutual trust and confidence among the basin riparian states in order to achieve collective action. Formation of river basin organizations encourages international collaboration and assistance for the river water development. As constraints on the resource grow, the opportunity cost for not cooperating is becoming clearer. The increasing scarcity of available fresh water per capita and lack of financial strength in the developing countries may gradually encourage the basin countries to cooperate in order to achieve an optimal benefit of the shared water. Basin-based development of irrigation, hydropower, water diversion, or flood control projects can provide riparian countries greater net benefits than what they could have achieved through purely state-centric development.

Conclusion

Water resource scarcity can be both the cause and the consequence of armed conflicts. In an armed conflict, as part of military strategy, the deliberate targeting of water projects may be directly responsible for inducing water scarcity or reducing the water quality of the opponent. Water supply is affected and polluted by conflicts, and fresh water resources also have the potential to cause or contribute to the creation of new conflicts. Nearly half of the global fresh water is to be found in more than 260 international basins. Many believe that the serious dependence of many countries on an external water supply may force them to reorientate their national security concerns in order to protect or to preserve such availability.

Shared water is expected not only to increase competition and conflict, but also to contribute to build engagement and cooperation among the riparian states (Wolf, 1998). The last 50 years have witnessed many water treaties being negotiated and signed. However, these water treaties face danger to their survival if they fail to receive support from effective institutional arrangement for proper water. Allotted water in the existing sharing agreements in most of the cases is unable to meet the increasing demand. The scope of further augmentation of river water in the arid and semiarid regions of the world is becoming limited due to possible impacts of global climate change. Global climate change has the potential to alter the water security landscape in a very significant manner in the near future.

Bibliography

Arnell, N.W., 1999. Climate change and global water resources. Global Environmental Change 9, 31–49.

Cooley, J.K., 1984. The war over water. Foreign Policy 54, 3–26.

De Wit, M., Stankiewicz, J., 2006. Changes in surface water supply across Africa with predicted climate change. Science 311 (5769), 1917–1921.

Drieschova, A., Giordano, M., Fishhendler, I., 2009. Climate change, international cooperation and adaption in transboundary water management. In: Adger, W.N., Lorenzoni, I., O'Brien, K. (Eds.), Adapting to Climate Change: Threshold, Values, Governance. Cambridge University Press, Cambridge, pp. 384–398.

Eckstein, G., 2010. Water scarcity, conflict, and security in a climate change world: challenges and opportunities for international law and policy. Wisconsin International Law Journal 27 (3), 409–461.

Falkenmark, M., 1990. Global water issues confronting humanity. Journal of Peace Research 27 (2).

McCaffrey, S., 1998. The UN convention on the law of the non-navigational uses of international watercourses: prospects and pitfalls. In: Salman, M.A., Boisson de Chazournes, L. (Eds.), International Watercourses: Enhancing Cooperation and Managing Conflict: Proceedings of a World Bank Seminar. World Bank, Washington, DC (World Bank Technical Paper No. 414).

Morrissette, J.J., Borer, D.A., 2004. Where oil and water do mix: environmental scarcity and future conflict in the Middle East and North Africa. Parameters 86–101.

Postel, Sandra L., 2000. Entering an era of water scarcity: the challenges ahead. Ecological Applications 10, 941–948.

Starr, J.R., 1991. Water wars. Foreign Policy 82, 17–36.

Swain, A, 1999. Constructing Water Institutions: Appropriate Management of International River Water. Cambridge Review of International Affairs 12 (2), 214–225.

Swain, A., 2001. Water wars: fact or fiction. Futures 33 (8&9), 769–781.

Swain, A., 2012. Global climate change and challenges for international river agreements. International Journal on Sustainable Society 4 (1/2), 72–87.

Swain, Ashok, 2010. Struggle against the State: Social Network and Protest Mobilization in India. Ashgate Publishing Limited, Farnham.

Tarlock, A.D., 1999–2000. How well can international water allocation regimes adapt to global climate change? Journal of Land Use & Environmental Law 15, 423–449.

UNEP (United Nations Environment Programme), 2002. Atlas of International Freshwater Agreements. UNEP, Nairobi.

Wolf, A., 1998. Conflict and cooperation along international waterway. Water Policy 1 (2), 251–265.

Wolf, A.T., Hamner, J.H., 2000. Trends in transboundary water disputes and dispute resolution. In: Green Cross International (Ed.), Water for Peace in the Middle East and Southern Africa. Green Cross International, Geneva.

Yoffe, S., Wolf, A., 1999. Water, conflict and cooperation: geographical perspectives. Cambridge Review of International Affairs 12 (2), 197–213.

Watson, John Broadus (1878–1958)

Peter Harzem[†], Auburn University, Auburn, AL, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16389–16391, © 2001, Elsevier Ltd.

Abstract

John B. Watson, the founder of the behaviorist movement in psychology, was born on January 8, 1878 in Travelers Rest, a poor village in South Carolina. He received his M.A. degree from Furman College, and his PhD from the University of Chicago in 1903. In 1908 he was appointed to chair a new sub department of psychology at Johns Hopkins University where he wrote most of his influential papers. Due to a love affair with Rosalie Rayner, a student at the university, he was compelled to leave academic life in 1920. He and Rosalie were married and Watson developed a very successful second career in advertising. In 1935, at the age of 37, Rosalie died of pneumonia. Watson was never to recover from that misfortune. He lived 25 more years, a sad, depressed man, and died on September 25, 1958. Watson believed psychology should become a science, and advocated behaviorism on the grounds that only behavior could be studied scientifically. He was a prolific researcher into a broad range of phenomena including homing patterns of birds, effects of alcohol on behavior, and acquisition of fear and conditioned reflexes in infants. His influence on contemporary psychology has been profound, although seldom acknowledged.

J. B. Watson the founder of the behaviorist school of psychology was born on January 9, 1878 in Travelers Rest, a poverty stricken village in South Carolina. His father deserted the family when he was 12. His mother, a deeply religious woman, sold what little they had and moved to Greenville, a larger town, in the hope of providing better education for the children. In 1894, through the help of their church, John was admitted to the Baptist Furman College, from which he graduated with an M.A. degree in 1900. He was then admitted to the University of Chicago where 3 years later he became the youngest person to receive a Chicago PhD He remained there until 1908 as Instructor and assistant to James Angell, the eminent philosopher. In 1904 he married Mary Ickes who was the youngest sister of Harold Ickes, the influential Secretary of the Interior in Roosevelt's 'new deal' cabinet. Harold disapproved of the marriage, and may have played a part in the problems Watson was to face later. In 1908 he moved to Johns Hopkins University as Assistant Professor and Chair of Psychology. Almost all of Watson's seminal work was produced in the relatively short span of the ensuing 12 years. In 1920 his academic career came to a premature end when he developed a love affair with a young student, Rosalie Rayner. Anticipating that the affair may lead to unwanted publicity for the university, Watson decided to inform the president of the university. At the interview the president demanded his immediate resignation which Watson instantly wrote. The anticipated scandal broke in newspapers some three months later, when his wife Mary filed for divorce. As soon as the divorce proceedings were complete, on December 31, 1920, Watson and Rosalie were married. Watson was penniless, having left everything he owned to Mary and Rosalie had been disowned by her parents, who disapproved of the marriage. For a time they stayed with friends in New York. Eventually, Watson was offered a position in the advertising agency, J. Walter Thompson Company. In accord with the company policy that all employees must have personal experience of selling, Watson was sent to New Orleans to sell fishermen's boots door-to-door. Upon returning to headquarters his ideas and proposals were so successful that within two years he had risen to the position of Vice President. Having reached the top of his profession, fallen, and risen to the top of another profession, Watson was to suffer yet another tragedy. In 1935 Rosalie died of pneumonia. She was 37. He was never to recover from this loss. He died a disappointed, depressed man on September 25, 1958 at, ironically, the Johns Hopkins Hospital in Baltimore.

In harmony with the spirit of his times Watson believed in the supremacy of science, and its power for resolving our social and personal problems. Psychology, he thought, must become a 'natural science.' His seminal paper, *Psychology as the behaviorist views it*, sometimes referred to as 'the behaviorist manifesto,' was published in 1913. Its essential arguments were that to be a natural science psychology must adhere to the principles that are basic to all sciences, including the principle that what is studied must be public, i.e. observable and replicable by more than one person. Therefore, the subject matter of psychology should be behavior and nothing else. These views came under severe attack from the psychological establishment for two main reasons: first, they implied that the work of the past several decades, based on the method of introspection, had to be discarded or at best, disregarded; and second, they conflicted with the established view that psychology studies the mind, the psyche, feelings, thinking, and the like. If psychology only studied behavior, what was to be made of such concepts? Watson's answer was that these concepts had no place in science. Thus, the controversy over behaviorism arose not as much from the proposal that the science of psychology should study behavior, as it did from Watson's insistence that it should study nothing else.

Watson wrote many scholarly articles, monographs, books, and popular articles developing his view of the behaviorist science of psychology. He believed the function of psychology is to 'predict and control' behavior which, when fully achieved, would result in the good society. He wrote 'behaviorism ought to be a science that prepares men and women for understanding the principles of their own behavior ... I wish I could picture for you what a rich and wonderful individual we

[†] Deceased.

should make of every healthy child if only we could let it shape itself properly and then provide for it a universe in which it could exercise that organization (1925). With reference to the intro-spectionists' efforts to obtain verbal reports of phenomena that occur 'inside' a person, he asserted that such events are no different from events outside the body and they are, just as behavior, fully amenable to objective, scientific investigation. Language is a form of behavior and, therefore, subject to behaviorist study. Personality is the result of 'what we start with and what we have lived through (1919). Since, as far as it was known at the time, what we start with could not be changed, personality could only be developed or changed by appropriately arranging the environment. Watson's statement 'Give me a dozen healthy infants, well-formed, and my own specified world to bring them up in and I'll guarantee to take any one at random and train him to become any type of specialist I might select—doctor, lawyer, artist, merchant-chief, and yes, even beggarman and thief, regardless of his penchants, tendencies, abilities, vocations, and race of his ancestors' is quoted in almost every textbook. However, the very next sentence is 'I know I am going beyond my facts and I admit it, but so have the advocates of the contrary and they have been doing it for many thousands of years (1925). The apparently deliberate omission of this sentence by his critics unfairly suggests an extreme claim that Watson did not make.

After his departure from academic life Watson continued to lecture, supervise research, and write on psychology. Between 1922 and the death of Rosalie he conducted a course on behaviorism at New School for Social Research, corresponded with Bertrand Russell, Yerkes, Titchener and others, many of whom contributed to his course as guest lecturers. His book *Behaviorism* (1925) was originally written as lectures for this course. His research with 'Little Albert' where he and Rosalie conditioned the fear response in the boy to furry objects is often cited, but his research with a large number of other infants is less known. In later studies he supervised the elimination of the fear response in doctoral research by Mary Cover Jones. His views on child rearing were extreme, arguing that the development of emotional ties with infants should be avoided and that parents should not hug and kiss their children. He advised parents to follow this, in a popular book written with Rosalie Watson (1928). Some years later he changed his position on this and expressed regret that he had advocated it.

The roots of behaviorism go back much earlier than Watson. Lloyd Morgan's book, *Introduction to Comparative Psychology* (1894) sets the beginning of 'animal psychology' which necessarily required research into behavior. The assumption that the principles discovered through research into animal behavior will also hold for humans began to gain strength. Further, Pavlov's discoveries of conditioning, Darwin's theory of evolution indicating a continuity across species, and Thorndike's theory of learning based on research into animal behavior contributed to a *Zeitgeist* for the emergence of behaviorism. Watson's doctoral research which was entitled *The psychic development of the white rat correlated with the growth of its nervous system* (published as *Animal Education*) emerged in this context. His major contributions were to articulate the prevailing ideas, explicitly to formulate behaviorism, and to elaborate the ideas and principles that follow from it.

Watson regarded himself primarily a researcher and often referred to himself as 'a laboratory man.' He was interested in broad range of phenomena and in finding new ways of investigating them. It is little known that he was a pioneer of what later came to be established as Ethology. Beginning in 1906 he spent two summers on the Dry Tortugas islands, off Florida Keys, studying the behavior and homing patterns of noddy and sooty terns. Between 1907 and 1915 he published eight papers on the homing behavior and spectral sensitivity of birds. He was also a pioneer of what is now known as behavioral pharmacology. In his last year at Johns Hopkins he had begun to investigate the behavioral effect of ingested alcohol, with the collaboration of Edward Thorndike at Columbia University. The project was abandoned when he resigned. In his second profession in advertising he continued to develop innovative ideas, and pioneered the technique of associating an attractive image with the product even if they are not related intrinsically. The first of these, showing pictures of attractive women in an advertisement for Pond's face creams was very successful, and the technique became established, lasting to the present.

In his relatively short span of years in academia Watson became one of the most influential individuals in psychology. He was president of the American Psychological Association in 1915, editor of *The Psychological Review*, *The Journal of Animal Behavior*, and *The Journal of Experimental Psychology* that he founded. His influence on psychology has been broadly profound and enduring, but seldom recognized. Throughout his career Watson maintained the practice of not replying to his critics. Nevertheless, the controversy over behaviorism continued unabated, and in time detracted from recognition of Watson's scientific researches. This, combined with his departure from academia in mid-career, produced the unfortunate result that all of his ideas about the science of psychology, personality, language, and his researches into conditioning were taken and developed by his contemporaries and by later psychologists, notably by B. F. Skinner, without due acknowledgment. Nevertheless, the signs of Watson's influence in the shaping of contemporary experimental psychology remain evident.

Bibliography

Morgan, C.L., 1894. Introduction to Comparative Psychology. W Scott, London.

Watson, J.B., 1913. Psychology as the behaviorist views it. Psychological Review 20, 158–177.

Watson, J.B., 1919. Psychology from the Standpoint of a Behaviorist. Lippincott, Philadelphia, PA.

Watson, J.B., 1925. Behaviorism. Norton, New York.

Watson, J.B., Watson, R.R., 1920. Conditioned emotional reactions. Journal of Experimental Psychology 3, 1–14.

Watson, J.B., Watson, R.R., 1928. Psychological Care of Infant and Child. W W Norton, New York.

Wealth Distribution

Edward N Wolff, New York University, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The principal assets in the USA in 2007 were homes (33% of total assets), business equity (20%), pension accounts (15%), and stocks and mutual funds (11%). The major sources of data on household wealth are estate tax data, household survey data, wealth tax data, and income capitalization techniques. Wealth inequality is typically measured by the share of total wealth controlled by the richest 1% of families. By this measure, there was a sharp increase in wealth inequality in the USA between 1983 and 1989, a decline from 1989 to 2001, and then a rise to 2010.

Wealth represents a stock of accumulated assets; income represents a flow of current output. Families not only receive income over the course of a year but also save part of their income in the form of housing, savings accounts, stocks, bonds, and the like. Such accumulated savings are referred to as wealth. The first part of this article explains why wealth, like income, has an important bearing on the well-being of families and individuals. The second part develops the concept of household wealth; the third discusses some of the problems inherent in its measurement; the fourth shows how the structure of household wealth has changed over time; and the fifth presents time trends in the inequality of wealth. Data drawn from the USA will be utilized to illustrate parts of the discussion in the text.

Wealth and Well-Being

Why are we interested in household wealth? Most studies use income as a measure of family well-being. Though certain forms of income are derived from wealth, such as interest from savings accounts and bonds, dividends from stock shares, and rent from real estate – income and wealth are by no means identical. Many kinds of income, such as wages and salaries and transfer payments, are not derived from household wealth; and many forms of wealth, such as owner-occupied housing, produce no corresponding income flow.

Moreover, family wealth by itself is also a source of well-being, independent of the direct financial income it provides. There are four reasons for this. First, some assets, particularly owner-occupied housing, provide services directly to their owner. This is also true for consumer durables, such as automobiles. Such assets can substitute for money income in satisfying economic needs. Families with the same money income but differing amounts of housing and consumer durables will have different levels of welfare.

Secondly, wealth is a source of consumption, independent of the direct money income it provides. With the possible exception of consumer durables, assets can be converted directly into cash and thus can provide for immediate consumption needs. This is also true for the equity in owner-occupied housing, because second mortgages and home equity loans are another source of credit.

Thirdly, the availability of financial assets can provide liquidity to a family in times of economic stress (such as occasioned by unemployment, sickness, or family breakup). In this sense, wealth is a source of economic security for the family over and above the income that it directly provides.

Fourthly, large fortunes can be a source of economic and social power that is not directly captured in annual income. Large accumulations of financial and business assets can confer special privileges to their holders. Such large fortunes are often transmitted to succeeding generations, thus creating family 'dynasties.' They can also enable individuals to influence the political process through large donations to candidates running for public office and to lobbyists pursuing the special interests of their clients. Thus, wealth holdings provide another dimension to household welfare over and above income flows.

What Is Household Wealth?

The conventional definition of household wealth includes assets and liabilities that have a current market value and that are directly or indirectly marketable. A typical list of assets includes owner-occupied housing and other real estate; cash, checking, and savings accounts; bonds and other financial instruments; corporate stock; equity in unincorporated businesses; and trust fund equity. This measure will be referred to as 'marketable household wealth' (or simply 'household wealth'), since it represents those assets over which the family or individual has control. This notion corresponds to wealth as a store of value and is used in conjunction with a standard national accounting framework.

Table 1 presents the balance sheet for the US household sector in 2007. The wealth accounts are first divided into two major categories: assets and liabilities. Assets consist of items owned by households which have a positive value to them, while liabilities are debts, which by definition have a negative value. The difference between total assets and total liabilities is called net worth or, simply, wealth. In 2007, total household assets amounted to US$73.5 trillion, total liabilities were US$ 11.3 trillion, and household net worth was US$62.2 trillion.

Assets are divided into three main groups. The first, tangible assets, include stocks of real goods that are both 'reproducible,' such as houses, and 'nonreproducible,' such as land. In 2007,

 http://dx.doi.org/10.1016/B978-0-08-097086-8.71017-8

Table 1 Balance sheet of total household wealth in the USA, 2007 (billions of dollars)

Panel A: Marketable net worth	*Value of component*	*Percent of total assets*
1. Total assets	73 519	100
a. Tangible assets	32 384	44
– Owner-occupied housing	24 112	33
– Other real estate and undeveloped land	8272	11
b. Financial assets	14 807	20
– Liquid Assets: Checking accounts, cash, currency, time deposits, money market funds, CDs, and the cash surrender value of life insurance	4820	7
– Pension Accounts: IRAs, Keogh plans, and other retirement accounts	8918	12
– Bonds and financial securities	1069	1
c. Equities	25 115	34
– Stocks and mutual funds	8669	12
– Unincorporated business equity	14 772	20
– Trust equity	1673	2
d. Miscellaneous assets[a]	1214	2
2. Liabilities	11 271	15
a. Mortgage debt	8413	11
b. Other debt	2859	4
3. Net worth	62 248	85
Memo:		
a. Home equity[b]	15 699	21
b. Pension fund reserves	13 391	18
c. Consumer durables	4468	6

Panel B: Other concepts of household wealth	*Value of component*	*Ratio to net worth*
Marketable wealth	62 248	1.00
Marketable wealth plus consumer durables	66 716	1.07
Marketable wealth minus home equity	46 549	0.75
Augmented wealth	88 547	1.42
a. Pension wealth	7105	0.11
b. Social security wealth	19 195	0.31

[a]Gold and other precious metals, royalties, jewelry, antiques, furs, loans to friends and relatives, future contracts, and miscellaneous assets.
[b]Gross value of owner-occupied housing less outstanding mortgage debt.
Sources: 2007. Survey of Consumer Finances, Available at: http://www.federalreserve.gov/econresdata/scf/scfindex.htm; Federal Reserve Board Flow of Funds; and Wolff, E.N., 2011. The Transformation of the American Pension System: Was It Beneficial for Workers? W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

total tangible assets owned by households were worth US$ 32.4 trillion, which amounted to 44% of total household assets. The major tangible asset owned by households is owner-occupied housing (i.e., houses that are lived in by their owners). This accounted for 33% of total household assets. The second component consists of land, apartment buildings, other rental property, and other real estate held by households. This group made up 11% of all assets.

The second major group is referred to as financial assets. These are assets which function as money or are readily convertible to money. These are divided into three subcategories. The first of these are so-called 'liquid assets.' The most liquid of these are cash, demand deposits (checking accounts), and time and savings deposits, which function directly as means of exchange. Liquid assets also include money market funds, CDs, and the 'cash surrender value' of life insurance funds. Certain life insurance plans (called 'full life insurance') allow individuals to contribute to a savings fund, and the amount the individual can withdraw from this savings fund is referred to as its cash surrender value.

The second subcategory is pension accounts, such as IRAs, Keogh plans, 401(k) plans, and the like. These can be invested in a wide assortment of assets and are referred to as 'tax-deferred' because the accumulations in these accounts are not taxed until they are withdrawn. The third subcategory includes bonds, notes, and financial securities issued mainly by corporations and the government. Such securities are 'promissory notes,' by which the borrower agrees to pay back the lender a certain amount of money (the principal plus interest) at a certain date.

In 2007, financial assets constituted 20% of household assets. Liquid assets as a group amounted to only 7% of total assets, while pension accounts constituted 12%. Bonds and other securities amounted to a mere 1%.

The third major group consists of equities or ownership rights. A corporate stock certificate issued by a company represents ownership of a certain percentage of the company's assets. Unincorporated business equity refers to small businesses (such as a farm or a store) owned directly by individuals, in contrast to corporations which are owned through stock shares. A person's equity in an unincorporated business is the value of the business. The third item, trust fund equity, refers to bank deposits, securities, corporate stock, and other financial instruments, which are held in a special legal arrangement called a trust fund. In a typical trust, the actual assets are managed by a specially named administrator (often a bank), and the income earned from the assets in the trust is remitted to individual beneficiaries. Trust funds held for individuals are counted as part of household wealth.

As a group, equities comprised 34% of total household assets in 2007. Corporate stock shares and mutual funds made up 12% of total household assets, unincorporated business equity 20%, and trust fund equity only 2%.

On the liability side, the major form of household debt is home mortgages. A mortgage is a loan, issued normally by a bank and usually for a period from 15 to 30 years, that is used to finance the purchase of real estate property, particularly houses to live in. In 2007, mortgage debt comprised 75% of total household debt. The remaining 25% consisted of other household debt, including automobile and other consumer loans, credit card debt, and educational loans. Together, total household debt amounted to 15% of the value of total household assets.

Three other entries are shown in the bottom of **Table 1**. The first of these is home equity, defined as the gross value of owner-occupied housing less outstanding mortgage debt. This represents the amount of money a family could net from the sale of the house (after paying off its outstanding mortgage balance). Home mortgages amounted to 35% of the value of owner-occupied housing. As a result, home equity made up a smaller proportion of total assets than the gross value of owner-occupied housing – 21 versus 33%.

The second entry is the reserves (assets) of pension funds. These funds are accumulated mainly by corporations and

governments in order to pay pension benefits for their workers when they retire. In 2007, they amounted to 18% of total marketable household assets. Technically speaking, these reserves are not part of marketable household wealth, since households have no control over the assets in these funds and cannot directly draw on these funds. However, these reserves are held for the benefit of individuals and therefore have some of the characteristics of a trust fund. The third category, called consumer durables, includes automobiles, washing machines and other appliances, furniture, and other long-lasting household equipment. Cars are the major component of household durables, and altogether durables amounted to 6% of total marketable assets.

A theme that regularly emerges in the literature is that there is no one unique concept or definition of household wealth that is satisfactory for all purposes. One definition of household wealth is that it is the sum of marketable wealth and consumer durables. This concept is useful because although consumer durables provide consumption services directly to the household, they are not easily marketable and the resale value of these items is typically far lower than the value of their consumption services to the household. As shown in Panel B, the sum of marketable wealth and consumer durables amounted to US$66.7 trillion in 2007, which was 7% greater than marketable wealth.

A second concept of household wealth defines it as marketable net worth minus net equity in owner-occupied housing. This definition represents a more 'liquid' concept than marketable wealth, since one's home is difficult to convert into cash in the short term. It thus reflects the resources that may be directly available for consumption or various forms of investments in the short term. This concept totaled US$ 46.5 trillion in 2007 – 75% of marketable wealth.

One of the major developments since the end of World War II has been the enormous growth in both public and private pension systems. Even though such pension funds are not under the direct control of individuals, they are a source of future income to families and thus may be perceived as a form of family wealth. Moreover, as Martin Feldstein (1974) has argued, insofar as families accumulate 'traditional' wealth to provide for future consumption needs, the growth of such pension funds may offset private savings and hence traditional wealth accumulation.

A wider definition of household wealth will thus add some valuation of pension rights, from both public and private sources, to marketable wealth. Such a measure provides a better gauge of potential future consumption.

Augmented household wealth is defined as the sum of household marketable wealth, social security wealth, and pension wealth. Social Security wealth is defined as the present discounted value of the stream of future social security benefits, and pension wealth is defined as the present discounted value of future pension benefits (see Wolff, 1992, for a discussion of measurement issues). As is apparent from the definitions, these two forms of wealth are not fungible or marketable, since individuals cannot convert these assets into cash.

What is the rationale for including these two forms of wealth? The main motivation is that from the standpoint of an individual, a future guaranteed flow of income is in many ways like owning a financial asset. Indeed, there are marketable assets called 'annuities' which have precisely the characteristic of providing a steady stream of income after a certain time (or age) is reached. The anticipated social security (or pension) benefits that an individual will receive after retirement is comparable to such an annuity. From the individual's point of view, then, a guaranteed stream of future pension or social security benefits may be as much a form of wealth as money put into a savings account that will be drawn on after retirement.

According to my estimates, total pension wealth in 2007 was US$7.1 trillion and total social security wealth was US$19.2 trillion (see Panel B). Together, they amounted to 42% of marketable net worth. Social Security wealth by itself was thus larger than any other component of household wealth except owner-occupied housing.

Problems in Measuring Household Wealth

There have been four principal sources of data for developing household wealth estimates: (1) estate tax data; (2) household survey data; (3) wealth tax data; and (4) income capitalization techniques. Each has its characteristic advantages and disadvantages.

Estate Tax Data

Estate tax data was the first major source of data used for wealth analysis. When someone dies, that person's assets are said to comprise his or her estate; and estate tax records are actual tax returns filed for probate. Such data have a great degree of reliability since they are subject to scrutiny and audit by the State. Their main limitation in the USA, at least, is that the threshold for filing is relatively high, so that only a small proportion of estates (typically 1% or so) are required to file returns. In the UK the threshold is considerably lower, so that the majority of estates file tax returns. This is also true for some states in the USA, as well as Washington DC, which have their own estate tax requirements.

Another difficulty with this data source is that the sample consists of decedents. Since most researchers are interested in distribution of wealth among the living population, a technique based on 'mortality multipliers' is used to infer the distribution of wealth among the actual population. If mortality rates were the same for each group, then the wealth of decedents would constitute a representative sample of the living population and researchers could use the estate data directly. However, mortality rates are much greater for older people than younger people, so that different 'weights' must be given to the estates filed by decedents of different ages. Mortality rates are also higher for men than for women (i.e., women, on average, live longer than men), and also higher for black individuals than whites. As a result, estates are assigned a weight based on the inverse of the group's mortality rate and these are then used to generate the size distribution of wealth in the living population.

Estimates of wealth inequality based on this technique are, as might be expected, quite sensitive to the precision of the mortality multipliers. The estimates can have a very large standard error, particularly for the young, since there are very few of them in the sample. This means that the results are very

sensitive to to the profile of the deaths in a given year and may therefore not be very reliable for young adults.

There are two other problems associated with this technique. First, insofar as mortality rates are inversely correlated with wealth (i.e., the rich tend to live longer), the resulting multipliers can be biased (see Shorrocks, 1975). Most studies do try to correct for this problem by using mortality rates for the wealthy that are lower than for the population as a whole, due to the longer expected life span of the wealthy. Secondly, the distribution of wealth estimated by this technique is for individuals rather than for families. Changing ownership patterns within families (for example, joint ownership of the family's house) can affect estimated wealth concentration. For example, marital customs and relations have changed over the century. Married women now inherit more wealth and have higher wealth levels than they did in 1900 or 1930. This reduces *individual* concentration even if *household* wealth inequality does not change. For example, between 1929 and 1953, Robert Lampman (1962) reported that the percentage of married women among the top wealth-holders increased from 9 to 18%.

As a result, since most researchers are interested in the distribution of family wealth rather than individual wealth, additional imputations must be performed to infer family wealth from estimates of individual wealth holdings. These typically involve making assumptions about marriage patterns – in particular, the correlation between the wealth holdings of spouses (whether, for example, rich husbands tend to be married to rich wives or poor wives).

Another problem with this data source is underreporting and nonfiling for tax avoidance. Though the returns are subject to audit, the value of cash on hand, jewelry, houseware, and business assets are difficult to ascertain. Their value is typically understated in order to reduce the tax liability of the estate. Moreover, *inter-vivos* transfers (i.e., transfers of wealth between living individuals), particularly in anticipation of death, can bias estimates of household wealth among the living. If older people pass on wealth to their children just before they die, then their estates would tend to underrepresent the wealth of comparably aged individuals still living.

Estate tax data have been used by Atkinson and Harrison (1978) for the UK; and Lampman (1962), Smith (1984), Wolff (1996a), and Kopczuk and Saez (2004) for the USA. The long-term time series on wealth concentration for the UK and the USA are based on estate tax data.

Household Survey Data

Household surveys are questionnaires that are given to a sample of households in a population. Their primary advantage is to provide considerable discretion to the interviewer about the information requested of respondents. Their major drawbacks are that information provided by the respondent is often inaccurate (response error), and, in many cases, the information requested is not provided at all (nonresponse problems). Another problem is that because household wealth is extremely skewed, the very rich (the so-called 'upper tail' of the distribution) are often considerably underrepresented in random samples. An alternative is to use stratified samples, based typically on income tax returns, which oversample the rich. However, studies indicate that response error and nonresponse rates are considerably higher among the wealthy than among the middle class. Moreover, there are problems in 'weighting' the sample in order to reflect the actual population distribution. Household surveys have been used extensively in many countries (see Wolff, 1996b, for a survey).

Income Capitalization Techniques

The third source of wealth data is based on the 'income capitalization' technique, which is usually applied to a sample of income tax returns. The earliest use of this technique on US data was a 1939 study by Stewart. In this procedure, certain income flows, such as dividends, rents, and interest are converted into corresponding asset values based on the average asset yield. For example, dividends are capitalized into corporate stock holdings by dividing dividends reported in an income tax return by the average ratio of dividends to corporate stock in the economy as a whole. This technique when applied to a large sample of tax returns can provide an estimate of the size distribution of wealth.

This source also suffers from a number of problems. First, only assets with a corresponding income flow are covered in this procedure. Thus, owner-occupied housing, consumer durables, and idle land cannot be directly captured. Also, in the USA, state and local bonds cannot be estimated, because this source of interest income is exempt from federal income taxes. Secondly, the estimation procedure rests heavily on the assumption that asset yields are not correlated with asset levels (i.e., for example, that large stock holders will receive the same average return on their stock holdings as small stock holders). Any nonzero correlation between asset holdings and yields can produce biased estimates. Third, the observational unit is the tax return. Various assumptions must be made in order to construct family wealth estimates from tax unit wealth.

Wealth Tax Data

A fourth source is wealth tax return data, which are available in a number of European countries such as Germany, Sweden, and Switzerland. These countries assess taxes not only on current income but also on the stock of household wealth. Though there is typically a threshold for paying wealth taxes, population coverage can be considerably greater than that of estate tax returns. However, the measurement problems are similar to those of estate tax data. The filer has a great incentive to understate the value of his or her assets or even not to report them at all, for tax avoidance. Moreover, the assets subject to tax do not cover the full range of household assets (for example, consumer durables, pensions, and life insurance policies are often excluded). In addition, the observational unit is the tax return unit, which does not directly correspond to the family unit. Wealth tax data have been used extensively by Spånt (1987) for an analysis of wealth trends in Sweden.

Changes in the Composition of Household Wealth

There have been some significant changes in the composition of marketable wealth in the USA over the last three decades.

One of the most salient trends is the rising indebtedness of American families, with the debt–equity ratio leaping from 15.1 to 21.0% between 1983 and 2010 (see Table 2). The principal source of these increases is not rising consumer debt, like credit card balances. In fact, nonmortgage debt as a fraction of total assets fell from 6.8 to 4.5% over the period from 1983 to 2010.

Rather, the primary source is rising mortgage debt, including home equity loans and second mortgages, which climbed from 6 to 13% of total assets. Indeed, mortgage debt as a share of the value of homeowner's property increased from 21 to 41%. Whereas the total market value of homes remained almost constant as a share of total assets over this period, net home equity – the value of the house minus any outstanding mortgages – plummeted from 24 to 18% of total assets.

A second notable change is that pension accounts rose from 1.5 to 15% of total assets between 1983 and 2010. However, this increase almost exactly offsets the decline in liquid assets, from 17 to 6%. As a result, the explosion in the use of various pension type accounts, like IRAs, 401(k) plans, and other thrift plans, rather than increasing overall family savings, appears, instead, to have allowed households to substitute tax-deferred pension accounts for taxable savings deposits.

A third important change is the rise in the share of corporate equities and mutual funds in total assets, from 9 to 11%, despite the drastic drop in the stock market from 2007 to 2010. However, this result does not reflect the full extent of the growth in corporate stock holdings, because stocks are also indirectly held in mutual funds, pension accounts, and trust funds. If we include these, then corporate stocks, both directly and indirectly owned by households, rose from 11 to 18% of total assets.

Table 2 Composition of marketable household wealth in the USA, 1983 and 2010 (percent of gross assets)

	1983	*2010*
Principal residence (gross value)	30.1	31.3
Other real estate (gross value)	14.9	11.8
Business equity	18.8	18.0
Liquid assets[a]	17.4	6.2
Pension accounts[b]	1.5	15.3
Financial securities	4.2	1.8
Corporate stock and mutual funds	9.0	11.4
Net equity in personal trusts	2.6	2.4
Other assets[c]	1.3	1.7
Total	100.0	100.0
Debt on principal residence	6.3	12.9
All other debt[d]	6.8	4.5
Total debt	13.1	17.4
Memo		
Debt/equity ratio	15.1	21.0
Net home equity/total assets	23.8	18.4
Principal residence debt/house value	20.9	41.2
Stocks directly or indirectly owned/total assets[e]	11.3	17.8

[a]Checking accounts, savings accounts, time deposits, money market funds, certificates of deposits, and the cash surrender value of life insurance.
[b]IRAs, Keogh plans, 401(k) plans, the accumulated value of defined contribution pension plans, and other retirement accounts.
[c]Gold and other precious metals, royalties, jewelry, antiques, furs, loans to friends and relatives, future contracts, and miscellaneous assets.
[d]Mortgage debt on all real estate property except principal residence; and credit card, instalment, and other consumer debt.
[e]Includes direct ownership of stock shares and indirect ownership through mutual funds, trusts, and IRAs, Keogh plans, 401(k) plans, and other retirement accounts.
Sources: 1983 and 2010 Survey of Consumer Finances, available at: http://www.federalreserve.gov/econresdata/scf/scfindex.htm.

Recent Trends in Household Wealth Inequality

The final part of the analysis shows trends in wealth inequality in the USA since the early 1980s and illustrates the sensitivity of wealth inequality estimates to different definitions of household wealth.

The calculations shown in Table 3 indicate an extreme concentration of wealth in 2010. The top 1% of families (as ranked by marketable wealth) owned 35.4% of total household wealth. The figures also show that the wealth share of the top 1%, after rising steeply by 3.6 percentage points between 1983 and 1989, fell by 4.0 percentage points from 1989 to 2001, and then increased by 2.0 percentage points from 2001 to 2010.

The inclusion of automobiles in household wealth (the only consumer durable available in the wealth survey data) lessens the estimated share of the richest 1% by about 1 percentage point in each of the years. This reflects both the fact that automobiles are more widely distributed in the population than other assets and also that vehicles are very important in the portfolio of poorer and middle-income families. The data also show similar time trends – a sharp rise in wealth inequality between 1983 and 1989 followed by a decline from 1989 to 2001 and then a modest increase from 2001 to 2010.

Household wealth excluding home equity is even more concentrated than marketable wealth, with the richest 1% (as ranked by this concept of wealth) owning 42% of the total in 2010. The trend in inequality is similar, with the share of the top 1% gaining 4.0 percentage points between 1983 and 1989, losing 7.2 percentage points from 1989 to 2001, and then increasing by 2.4 percentage points in the ensuing 9 years.

Table 3 Wealth inequality trends in the United States, 1983–2010

Year	*Marketable net worth*	*Net worth plus autos*	*Net worth minus net home equity*	*Augmented wealth*
1. Percentage share of total household wealth owned by the richest 1% of households				
1983	33.8	32.6	42.9	NA
1989	37.4	36.1	46.9	20.3
2001	33.4	32.3	39.7	23.2
2007	34.6	33.6	42.7	25.1
2010	35.4	34.3	42.1	24.3
2. Gini index (Gini coefficient x 100)				
1983	79.9	77.9	89.3	NA
1989	83.2	80.7	92.6	66.3
2001	82.6	80.6	88.8	66.1
2007	83.4	81.6	90.8	68.4
2010	87.0	84.8	92.7	68.5

Note that automobiles are the only consumer durable included in the survey data.
Sources: 1983, 1989, 2001, 2007, and 2010 Survey of Consumer Finances, and Wolff, E.N., 2011. The Transformation of the American Pension System: Was It Beneficial for Workers? W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Table 3 also shows the trend in the share of augmented household wealth owned by the top 1% of wealth holders as ranked by augmented wealth from 1989 to 2010 (figures on augmented wealth are not available for 1983). The addition of pension and social security wealth has had a significant effect on measured wealth inequality because pension and social security wealth, particularly the latter, have grown much faster over time than marketable wealth and because social security wealth is more evenly distributed in the population than marketable assets. The gap in the share of the top 1% between marketable wealth and augmented wealth was 17.1 percentage points in 1989 (37.4 vs 20.3%). Unlike marketable wealth, wealth concentration based on augmented wealth shows an increase between 1989 and 2007 and a slight decline from 2007 to 2010.

The Gini index, a broader measure of inequality, which reflects the full distribution of wealth and which ranges from one to one hundred, shows a similar rank order of inequality among the different wealth concepts. However, it also shows somewhat different time trends. According to the Gini coefficient, the inequality of marketable wealth rose sharply from 1983 to 1989, more or less stabilized until 2007, and then increased sharply in 2010. Similar differences are evident for other measures of wealth. These comparisons highlight the sensitivity of wealth trends to different measures of inequality.

Future research might most appropriately be focused on understanding the factors behind changing wealth inequality both in the USA and other advanced nations (household wealth data are not generally available for less developed countries, with the notable exception of China). How have relative wealth holdings changed among age groups, among income classes, across racial and ethnic lines within a country, between families with children and those without children, and across gender lines? How have wealth holdings changed among countries of the world? What are the respective roles of household savings and intergenerational wealth transfers in explaining wealth differences among individuals and between groups? What are the reasons for rising indebtedness among American families and how does this compare to other countries? These are some of the salient research issues that occupy current researchers in the field.

See also: Consumption, History of; Consumption, Sociology of; Equality and Inequality: Legal Aspects; Income Distribution and Inequality: Measurement Issues; Leisure, Sociology of.

Bibliography

Atkinson, A.B., Harrison, A.J., 1978. Distribution of Personal Wealth in Britain. Cambridge University Press, Cambridge, UK.

Feldstein, M., 1974. Social security, induced retirement, and aggregate capital accumulation. Journal of Political Economy 82, 905–926.

Kopczuk, W., Saez, E., 2004. Top wealth shares in the United States: 1916–2000: evidence from estate tax returns. National Tax Journal 57 (2, Part 2), 445–488.

Lampman, R.J., 1962. The Share of Top Wealth-holders in National Wealth, 1922–56. Princeton University Press, Princeton, NJ.

Lipsey, R.E., Tice, H. (Eds.), 1989. The Measurement of Saving, Investment, and Wealth. Chicago University Press, Chicago.

Shorrocks, A.F., 1975. The age–wealth relationship: a cross-section and cohort analysis. Review of Economics and Statistics 57, 155–163.

Shorrocks, A.F., Davies, J.B., 1999. The distribution of wealth. In: Atkinson, A.B., Bourguignon, F. (Eds.), Handbook of Income Distribution. North-Holland, Amsterdam.

Smith, J.D. (Ed.), 1975. The Personal Distribution of Income and Wealth, NBER, Studies in Income and Wealth. Columbia University Press, New York. No. 39.

Smith, J.D., December 1984. Trends in the concentration of personal wealth in the united states 1958–1976. Review of Income and Wealth Series 30, 419–428.

Spånt, R., 1987. Wealth distribution in Sweden: 1920–1983. In: Wolff, E.N. (Ed.), International Comparisons of the Distribution of Household Wealth. Oxford University Press, Oxford, UK.

Stewart, C., 1939. Income capitalization as a method of estimating the distribution of wealth by size group. In: Studies in Income and Wealth, vol. 3. National Bureau of Economic Research, New York.

Wolff, E.N. (Ed.), 1987. International Comparisons of the Distribution of Household Wealth. Oxford University Press, Oxford, UK.

Wolff, E.N., 1992. Methodological issues in the estimation of retirement wealth. In: Slottje, D.J. (Ed.), Research in Economic Inequality, vol. 2. JAI Press, Greenwich, CT, pp. 31–56.

Wolff, E.N., 1996a. TOP HEAVY: A Study of Increasing Inequality of Wealth in America, updated and expanded edition. The New Press, New York.

Wolff, E.N., 1996b. International comparisons of wealth inequality. Review of Income and Wealth Series 42 (4), 433–451.

Wolff, E.N., 1998. Recent trends in the size distribution of household wealth. Journal of Economic Perspectives 12 (3), 131–150.

Wolff, E.N., 2011. The Transformation of the American Pension System: Was it Beneficial for Workers? W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Weber, Max (1864–1920)

Stephen P Turner, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Max Weber was one of the great figures in the history of social science, and is now recognized as a political philosopher. His most famous work, *The Protestant Ethic and the Spirit of Capitalism*, led to a larger life-long comparative project on the relation between capitalist development and religiously conditioned economic ethics. Weber also attempted to define the relations between social science and normative sciences in accordance with the fact value distinction, and to draw out the implications of the nonrationality of value choice for social science, which works with value-related descriptions, and for politics, which involves competing values.

Max Weber was born Karl Emil Maximilian Weber in Erfurt, Thuringia on April 21, 1864. His father, Max Weber (1836–97), was a lawyer from a family of wealthy industrialists and merchants with international connections whose core business was a linen factory in Oerlinghausen, in Westphalia. His uncle transformed this factory, which had been a traditional family concern, into a model modern business. *The Protestant Ethic and the Spirit of Capitalism* contains thinly concealed elements of this history. Weber's father was a younger son in this family. His father pursued a legal and political career, serving in the German Reichstag 1872–84. Politicians and intellectuals, particularly historians, frequented the Weber household in Berlin. His mother, Hélène (née Fallenstein) was highly educated, religious, and active in poor relief.

Weber entered the university in 1882, studying law, as well as economics, ancient history, philosophy, and theology, and entered into the student life of the time, joining a fraternity. He served the first year of officer training in the army at Strasbourg, where relatives of his mother's were prominent at the university. He returned to Berlin in 1884 and concentrated on law and legal history, simultaneously pursuing his legal education, taking exams, and completing military service. Levin Goldschmidt and August Meitzen instructed him in *Staatswissenschaft* and he became a doctoral student under Goldschmidt, also briefly clerking at the Royal District Court of Charlottenburg, and performing military service in Posen.

His dissertation, "Development of the principle of joint liability and the separate fund in the public trading company from the household and trade communities in the Italian cities" (1889), based on hundreds of Italian and Spanish sources, was well received. He joined the *Verein für Sozialpolitik* (Association for Social Policy) out of a concern with the 'Polonization' of the Eastern regions of Germany, and became involved politically in the *Evangelisch-sozialer Kongress*. He continued his academic work, receiving his Habilitation under Meitzen for 'Roman agrarian history and its importance for state and civil law' in 1891. During this period, the *Verein für Sozialpolitik* employed him for a study of "The conditions of the agricultural workers in East Elbian regions of Germany." The study, a huge empirical survey designed as the basis of policy recommendations, marked a new direction in his career. The recommendations were controversial, but made his reputation as an agricultural economist, and enabled him to secure a professorial appointment in *Nationalökonomie* at Freiburg in 1894. From there he was appointed to a similar position at Heidelberg.

A mental breakdown after the death of his father ultimately led him to relinquish his teaching duties, but eventually he returned to an active role as an editor and producing scholar. His major 'methodological' essays (posthumously published as his *Gesammelte Aufsätze zur Wissenschaftslehre* in 1922) included part one of *Roscher and Knies* published in 1903, 'Objectivity in Social Science and Social Policy' in 1904, part two of *Roscher and Knies* and 'Critical Studies in the Logic of the Cultural Sciences' in 1905 and 1906, and, in 1907, *Critique of Stammler*. With the 1904–05 publication of *The Protestant Ethic and the Spirit of Capitalism*, his work began to take several new directions. During this period Weber traveled to the United States, where he spent several months, appearing as an Agricultural Economist at the 1904 St. Louis World's Fair, and visiting American relatives. His experiences were the basis for some of his subsequent writings. He became fascinated by the revolution of 1905 in Russia, learning the language in 3 months in order to follow the unfolding events, and wrote extensively on them, from a point of view that emphasized matters of social structure. Over the next few years he completed his work on ancient agrarian economic life, and agreed to take over the editing of a work which evolved into the text now known as *Economy and Society*, which Weber was never to complete. He became engaged publically in questions of university policy, and also pursued inquiries into industrial psychology in the family textile works, an expression both of his interest in the rationalization of work and his recognition of the role of such psychophysical factors as fatigue in work processes.

Weber also played an active, though characteristically combative role, in two academic organizations, the Social Policy Association and the German Sociological Society. He was one of the founders of the German Sociological Society in 1909, but left soon after because of his views on the separation between scholarship and advocacy. These views, which later became famous, if not well understood, under the label 'value-free science,' were developed and elaborated in a complex and long-running dispute within the Social Policy Association. This dispute began largely as a political dispute over matters of policy but it escalated quickly and was

 http://dx.doi.org/10.1016/B978-0-08-097086-8.61136-4

transformed by Weber and his brother Alfred into an accusation that their opponents in this dispute had confused considerations of value with objective policy considerations, a point that was particularly crucial for Weber, who rejected the bland statism of his monarchical socialist older colleagues. Weber articulated an elaborate rationale for his position that followed from his notion of the nonrationality of ultimate value choice. He also argued that maintaining this separation was part of the duty of university professors. This did not mean that Weber himself refrained from taking on a political role. He was an active participant in a number of political and diplomatic ventures and as a member of private groups attempting to influence policy.

Weber's role in what was to become *Economy and Society* expanded and he became author rather than editor, a task which together with the series of studies that comprised 'The economic ethics of the world religions,' occupied him until his death. Neither project achieved final form. During World War I Weber, a reserve officer, worked as a hospital administrator, until reorganization made his services unnecessary. When the outcome of the war became clear, he wrote a series of articles in the *Frankfurter Zeitung* on the future constitution. During this period he gave two speeches with parallel titles to the Munich Free Students Association, 'Science as a Vocation,' apparently in late 1917, and 'Politics as a Vocation' in early 1919, after which both were published. The speeches were an aggressive and telling challenge to the cultural and political optimism of the time, and in the Weimar era remained relevant, especially to the long-running discussion of the 'Crisis of the Sciences.' After the war Weber served as a member of a group invited to participate in *Reichsamt des Innern* (Office of the Interior) meetings which produced the general outlines of the constitution of what became the Weimar Republic (Mommsen, 1984), became a professor of economics at Vienna in 1918 and Munich in 1919, and unsuccessfully pursued a career in politics. His untimely death, 14 June 1920 in Munich, came when his influence among intellectuals was at its peak.

Weber's Early Work

Weber's habilitation dissertation, although he later dismissed it as a "one of the sins of his youth" (quoted in Tenbruck, 1987), together with an essay on 'The Social Basis of the Decline of Ancient Civilization' (1896) marked him out as an original scholar. The dissertation used his knowledge of medieval land tenure law to illuminate the Roman agrarian situation. It depended on the plausible but unprovable assumption that two of the different methods of mapping used by the Roman Colonial land surveyors corresponded to different types of land tenure: untaxed, roughly measured lease rights from public land, and precisely measured private taxed lots. He argued that the size of the private lots suggested that they were too small for subsistence, and that from the first the smallholders on this land had depended on access to public land to survive, but that the rich, who could pay for leases and work the land with slave labor, eventually cut off this access. He reasoned that the poor but free peasants were forced, by the inadequacy of their holdings, to assume labor obligations to landlords as rent, and later to be legally bound to the soil. When new slaves were no longer available, barracked slaves, who had enabled production for the market, had to be raised in status, freed sufficiently to encourage them to reproduce, making them essentially into serfs. Increasingly, money transactions were replaced by transactions in kind, the coastal commerce on which the slave produced surpluses depended vanished, and the cities diminished in power and in fiscal, and consequently political, importance for the state. The new rural economic order was manorialism *de facto*, so this was in effect an account of the origins of the medieval agrarian order.

The 1903–07 Methodological Essays

Contemporary disputes in ancient history, notably the 'Bücher–Meyer controversy' over a stage-theory of the development of the *oikos*, were formative for Weber. His most important methodological article, 'Objectivity in Social Science and Social Policy,' was published in 1904, on the occasion of the transfer of control of the *Archiv für Sozialwissenschaft und Sozialpolitik* to a team of editors of which he was the leading member. Weber argued that to construct an historical explanation, or even an object of explanation, requires conceptualization, which he examined in terms of the concept of ideal-types that he distinguished from mere classifications that simply sorted objects into categories. Ideal-types typically diverged from reality, or were applied to empirical cases from which they diverged, as in the case of ancient capitalism, but served the purposes of the historian by virtue of the fact that they allowed the historian to ask questions that are meaningful in the present. Thus, while ideal-types were indispensable to the historical disciplines, these disciplines were also prone to a particular kind of error. The mistake in each case is that something that is properly regarded as an idealization of a phenomenon is treated as a real force or a teleological process. In Weber's hands, this was a powerful means of methodological criticism. Not only did it serve to indicate how the developmental models of his near predecessors and contemporaries, including Marx, used the idealization to constitute the phenomenon it explained, it called into question the project of final conceptualizations of history or social life, as well as the cognitive status of such conceptual constructs as 'the state.'

The line between the parts of historical understanding that are relative and those that are nonrelative and factual is difficult to draw. In 'The Logic of the Cultural Sciences,' he spoke of the 'real causal processes' which our concepts give an intelligible form to. In 'Objectivity' he argued that 'genetic ideal-types' are themselves relative to a culture-bound historical interest: the calculation of probabilities is fully objective, but the selection and conceptualization of the conditions cannot be. These are inherent limitations of a social science that must operate with concepts which seek conceptual clarity but whose meaningfulness derives from and remains tied to sources outside the science itself: 'the language of life,' a language which is necessarily bound to a historical moment and its values (Weber, 2012).

This argument takes on a special significance in connection with his account of values themselves. Weber argued that scholarship could at best enable judgments about the adequacy

of means for ends, the consistency of ends, the separation between and the difference between factual and valuative considerations, but could not constitute the basis of either value judgments or the endorsements of policies, at least to the extent that the valuative purposes of the policy could not be given scholarly justification. Ultimate value choices, worldviews, religions, and national cultures were alike in that they were in the end matters of commitment, not reason, and consequently questions of 'superiority' were misplaced.

The Protestant Ethic Essay

These basic methodological ideas, and especially the idea of ideal-types, figure in Weber's 1904–05 essay, *The Protestant Ethic and the Spirit of Capitalism*. The core problem of the essay was to account for the disruption and replacement of traditional economic life, and, he argued, that it was in part the indirect effect of Protestantism. Reformation theology, and particularly Calvinism, altered the problem of salvation. The doctrine held that the Elect, those predestined for salvation, are called to worldly tasks and an ascetic lifestyle: worldly monasticism. But, Weber reasoned, the doctrine of predestination was psychologically intolerable, and produced deep anxiety, which had to be allayed by pastoral assurances that faithful work in a calling was a sign of election, creating a novel and powerful psychological sanction for conduct. The theological premise thus had the indirect effect of producing a personality that, in the appropriate setting, becomes the austere, the abstemious, and the rationalizing capitalist.

There are many controversial features of this account, some of which result from the fact that the doctrinal element of the ideal-type that Weber constructed only imperfectly corresponded to the theological doctrine of much of Protestantism in Europe. Nevertheless, the account is deeply compelling, in part because the depictions of the believer accord so well with a character type that unquestionably did leave a profound mark on Western culture. Weber closed the essay with some prophetic remarks about the course of Western culture, and an important analysis of the cultural situation created by the secularization of the 'ethic.' He observed that the rosy blush of its 'laughing heir,' the Enlightenment, has faded, and that the inexorable demands of the modern, rationally organized economy, enforce the organization of work into 'callings' but strip them of their religious meaning. He commented that even the accumulation of wealth takes on the character of sport, and the people in the machine of capitalism became divided beings, "specialists without spirit, sensualists without heart" (1958a).

In his late lectures on economic history, Weber restated the thesis with some different emphases, notably with respect to the rational organization of free labor, which Weber came to see as especially crucial, and indeed to be the distinguishing feature of modern capitalism in contrast to booty capitalism, trade, and political capitalism, all of which were found in the ancient world. The crucial moment was when "the production of goods was emancipated from all the bonds of inherited intelligence, and came under the dominance of freely roving intelligence" (1961), the moment in which the craftsman was replaced by the worker, and the owner of the firm supplied tools, exacted discipline, and assumed the risks. This change was, Weber thought, a precondition for the mechanization of work with the goal of saving labor costs, and in this sense rationalizing labor. Together with this was the rationalization of the use of capital generally, through rational accounting methods, rational price setting for the purpose of profit, and backed by calculable law. The worldwide expansion, and continued success, of these practices did not rest on the foundation of religious belief. Weber also provided an overview of other factors in the development of capitalism, such as calculable law, and other contingencies, like the availability of coal in Europe, without which, he argued, "the revolution in the means of work … might have stopped and modern capitalism in its characteristic form never would have appeared" (1961).

Economy and Society and the Economic Ethics of the World Religions

In the last decade of Weber's life he produced a major methodological text; 'The Categories of an Interpretative Sociology' (2012), a major systematic work, the incomplete work known as *Economy and Society*, and the bulk of his mature substantive work, 'Economic Ethics of the World Religions.' The essays that make up this work are overtly concerned with the question of why a rational capitalism based on calculability originated exclusively in the West. However, their larger theme is the general problem of the cultural specificity of the West, that is to say the peculiarity, in world terms, of Western political, economic, and legal institutions. But the uniqueness of the West requires an unusual explanation.

The general development of religious ideas corresponds to the development of particular social strata, with peasants having a strong affinity for magic and aristocratic classes for a kind of self-justifying, charismatic self-conception. These affinities tend to be relatively stable and do not work to transform the social structures in which they arise. The problem of explanation, consequently, is to identify the point of openness in which very specific kinds of innovations, namely those that have transformative effects on individuals or groups, might arise. The general promise of religion is health, wealth, and happiness. And the basic affinities to strata express these needs in forms specific to the strata.

The exceptional cases are the salvation religions, which offer rebirth or redemption, and more generally, religions that are doctrinally open to rationalization and that produce potentially transformative innovations through rationalization. Religious rationalization, however, while producing new coherence or consistency in religious doctrines, characteristically serves to produce discrepancies with other 'spheres' from which religious life is differentiating under the same pressure of circumstance. A general problem which all salvation religions must face up to is the problem of the relationship between destiny and merit, the problem of theodicy, which these religions develop in extraordinarily elaborate ways – ways which 'rationalize' the discrepancies by making the fact of evil or bad luck intelligibly consistent with the religious ideology, and also ways that are consequential for practice. The Chinese solution, for example, was to reconcile the two by recourse to popular magical practice, which in turn had

negative consequences: a "deep repugnance to undertaking any change in the established conduct of life because supernatural evils are feared" (1961). Ancient Judaism, in contrast, devised solutions that excluded and repudiated magic, and this carried over into Christianity and consequently into the Western worldview in a way that favored the eventual development of science and technology.

Weber pointed to many similar indirect effects of general institutional conditions, conditions that do not directly determine possibilities, but nevertheless greatly reduce the probability of developments in a particular direction. Rationalization, for example, means the attainment of greater consistency and completeness, but the two take various different forms for intellectuals and practical persons, such as merchants, and the direction that rationalization takes in a given situation or in the face of given problems is governed by many circumstantial contingencies. Weber saw the various forms and processes of rationalization as interacting with one another. Rationalization in one sphere, such as law, might create difficulties that require a response forcing greater rationalization in other spheres, such as the economic, which in turn present problems for the law which are resolved by further legal rationalization, thus producing a kind of dynamic ongoing process of mutually reinforcing rationalization. Occasionally Weber appears to slip into the error of regarding the rationalization process as itself a historical force, but these rare cases ordinarily can be interpreted in causal, and typically 'interest' terms, as when, for example, the rationalization of modern law on the continent is interpreted as driven by the status interests of law professors.

Economy and Society

Economy and Society consists of a typology of forms of social action, especially institutional action, together with extensive historical commentary on the types. Weber said that the claim to be made on behalf of the typology was its usefulness; and the elaboration of the typology is an extended demonstration of the uses to which the types could be put in analyzing actual historical forms, particularly of institutional structure, and characterizing their historical course of development as the product of the basic properties of the type, always with the caveat that these were idealizations rather than 'real' types with a teleological character.

The most famous of these typologies involves beliefs legitimating authority and the forms that were characteristically produced by these beliefs. The categories were charismatic, traditional, and rational–legal. These were ideal types, which almost never, if ever, appeared in reality in a pure form. Traditional authority is based on unwritten rules believed to have been handed down from time immemorial; rational legal authority rested on the belief in the validity of written rules produced according to written procedures, and charismatic authority is the authority of the extraordinary person. Weber had in mind of course such figures as Napoleon and Jesus Christ, but he found many more modest examples in Indian gurus and medieval battle leaders, and in each of these cases obedience rested not on either written or unwritten rules but on the direct influence of the exceptional individual himself. Weber was careful to point out such things as the attenuated charismatic element in the modern monarchy and the jury.

Charismatic authority, Weber observed, is inherently unstable, though its temporary effects, and thus historical consequences, may be great. Gradually the critical charismatic leader must support his followers materially and this marks the beginning of the transformation of charisma into the everyday. Charismatic regimes also face problems of succession, and the 'routinization' of charisma through ritual in order to pass on charisma to successors soon becomes transformed into something akin to traditional authority, or alternatively into a bureaucratic system.

The value of the ideal-type is not so much that it serves to predict as that it enables one to ask relevant questions or identify relevant causal processes. However, this process is itself governed by our cognitive purposes rather than something inherent in the historical world itself. Much of Weber's discussion of the historical case material is indeed about the logical unfolding of implications of basic ideas such as charisma or rational legal authority. However, there is no sense in which this unfolding is causally inevitable. It is rather that concepts like charisma are useful in making processes intelligible precisely because the elaboration of the idea corresponds to the kind of developments that historical agents, employing their own ideas, beliefs, and so forth, actually bring about in the course of their actions.

Weber as a Political Thinker

Weber was self-consciously part of the post-Bismarck generation, and an ardent nationalist who was deeply concerned about the problem of succession of leaders. Bismarck's career had shown how forceful and successful leadership produced popular acquiescence and even enthusiasm, but with the effect of making Parliament weak and irrelevant, and undermining party loyalty. Weber saw this form of popular leadership, exemplified for him by both the unelected Bismarck and the elected Gladstone, as a key to the nature of modern politics: modern regimes were mass democracies in fact, regardless of the forms of constitutions, because they required mass legitimacy and support to be effective. The selection of leaders, and beyond this the administrative machinery of parties, were constrained by the demands of effective electoral politics. He was concerned on the one hand that interest-faction parties, especially parties that inevitably were bureaucratically organized, would be effective electorally, but that this would assure the predominance of party hacks rather than figures who could act boldly as leaders. He responded by suggesting a presidential office in which the president required support in the form of a plebiscite rather than party support, and thus stood above and outside of party structures. Although this analysis supported a liberal and democratic republican form of government, Weber nowhere based his argument for this form of state on traditional liberal constitutional concerns, such as the preservation of rights or the expression of the will of the people, notions he regarded as mythological, but instead based it on the problems of selecting great leaders in the face of the modern need for popular support for the tasks of nation-building. Weber nevertheless had a strong interest in

democratization in the social and cultural sense, and a fascination with and respect for the kinds of popular institutions that he encountered in his American travels, particularly the sectarian congregations.

Many interpreters have noted that the Nietzschean language of greatness that Weber employed, as well as the specific concerns about freedom he expressed, are rooted in a morality of the value of 'personality' that is potentially in conflict with constitutional liberalism, and especially to its egalitarian elements. When he observed that in rational work organizations "each man becomes a little cog in the machine and, aware of this, his one preoccupation is whether he becomes a bigger one," and when he asked if "a portion of mankind (can be kept) free of this parceling out of the soul" (quoted in Mayer, 1956) his concern was with a form of freedom accessible only to the few. Yet Weber's argument proceeds on more than one level. On the one hand he presented the state as a tool whose distinctive and necessary means is violence, and traced the origins of the historical conditions that constrain the possible present forms of state. On the other hand he made the point that the state is a means not necessary merely to the achievement of his own values, but to the achievement of virtually all this-worldly values: only someone like a Tolstoyan Saint has no need of the state. For the rest, there are no alternatives.

Much of Weber's energy went into thinking about international politics. He was an advisor to one of the most persistent politicians of the *Kaiserreich*, Friedrich Naumann, whose book *Mitteleuropa* set out the German problem and its solution through the incorporation of Eastern Europe in a German dominated customs union with common institutions. Although Weber played no role in the writing of the book, it resonates with Weberian insights into the reasons for the superiority of North Germany in the sphere of production. Naumann's account of the dependence of Germany on the agricultural lands of eastern Europe and the implications of economic and export superiority on regional political relations anticipates not only Nazi *Grossraum* policy but the larger problem of creating European institutional forms that accord with German economic superiority. Weber, despite his concerns over Germany being drawn into the political problems of the Austro-Hungarian Empire, participated in 1916 in a working group created to support and inform German policy in the direction of a closer union. He took part in a semiofficial diplomatic initiative to Austria, and made recommendations about the future of 'Congress Poland' and was especially engaged in the Polish question, which he thought should be solved by an independent state dominated by Germany rather than an expansion of Austrian control, and reflected on Bismarck's policy toward the problem. These issues became irrelevant when the tide of war turned against Germany. When it was clear that the war was ending and the question of German guilt was going to play a role in the peace, he assembled a group of professors in Heidelberg to prepare a response, which led to his participation in the Versailles conference as part of the German delegation.

Although Weber's interests in international affairs led to a number of newspaper interventions, and was the basis for his reputation as a political 'Realist,' he never produced a major work in this area. Weber's ideas, however were absorbed by Hans J. Morgenthau, and played a foundational role in the 'Realism' Morgenthau articulated as an American Professor and founder of the modern field of international relations.

Weber's Reception

Weber was not employed as a professor during his most productive years, and exerted relatively little influence through the training of students. His primary personal influence was through a salon that he and his wife held in Heidelberg which attracted many of the best minds, and particularly the best young minds, of the university, across various disciplines. This group included Georg Lukács, Emil Lask, and Karl Jaspers. Jaspers portrayed him as an existentialist hero: "an idealized modern man who resolutely exposes himself to intellectual risks and lives through intellectual tensions passionately without accepting final resolutions."

The Nazi period cut off serious discussion of Weber, who was described in the intellectual circles that supported the regime with the derogatory term 'liberal,' a characterization that shaped the revival of interest in his work in the postwar period. Weber's 'liberal' admirers, such as Jaspers, promoted the idea that Germany would have been saved from Nazism had Weber been accepted as a political leader, and as a consequence, after the Second World War, Weber was reinterpreted as a precursor of the new federal republic and its form of liberal democracy. Mommsen attacked this interpretation in his 1959 dissertation, later revised (1984), which noted Weber's attitude toward the normative premises of conventional liberal constitutionalism and suggested that his devotion to the idea of great leadership placed his thoughts uncomfortably close to the fascist *Führerprinzip*. Mommsen's work produced a controversy that broke out in a visible way in the 1964 commemoration of the centennial of his birth, when the thesis was radicalized by Marcuse and Habermas. Mommsen later stressed that he did not believe that Weber's own intentions had been antidemocratic in the Nazi manner, but that nevertheless "it crossed the boundaries of the traditional understanding of democracy and was amenable to reinterpretation in an anti-democratic sense" (1984).

Weber's thought appears early in the United States, but not as a founding figure of sociology, because he had no 'system' of the sort that other German and American thinkers had, and was only intermittently identified as a sociologist. Economists first translated his work: Frank Knight his *General Economic History*, based on student lecture notes and some added material from earlier work, Talcott Parsons his *Protestant Ethic*, the argument of which was popularized in Britain by R.H. Tawney. After the Second World War, the American social sciences, and especially sociology, were exported, in part as an alternative to Marxism, and Weber's ideas, understood primarily in American terms, became influential in the UK and in Germany itself. The characteristic emphasis was on value-free science and the rejection of ideological thinking, and on specific sociological concepts, such as charisma and bureaucratization, and on Weber's rejection of the Marxist claim of primacy of material over spiritual or institutional explanation. Weberian approaches typically were seen as emphasizing the practical demands of bureaucracy, or the autonomous role of interests other than

economic interests, such as ideal interests. The notion of the Protestant ethic fit with analysis of modernization and development of the sort that were part of the American response to third world and anticolonial movements, and were generalized and applied to such countries as Japan.

Weber came to be regarded as a distinguished forebear of antipositivist sociology by virtue of his use of the terms 'meaning' and 'understanding' to differentiate his own approach to sociology. The refinement and appropriation of Weber's idea of an interpretative sociology by phenomenologists such as Alfred Schütz became one source of a phenomenological approach to sociology. Weberian ideas were also conveyed through the writings of such figures as Karl Mannheim and Norbert Elias, but in forms increasingly distant from Weber's own thought. This set the stage for a series of rediscoveries within sociology of 'the original' Weber so that the 1980s and 1990s saw within sociology, and especially in the work of such writers as Wolfgang Schluchter, and more generally in connection with the *Max Weber Gesamtausgabe* project, a much more disciplined return to Weber's original concerns, and a vindication of the importance to Weber of such basic problems as the problem of Western uniqueness.

Although in his final post and through much of his teaching career Weber was an economist and economic historian, and the early reception of Weber in the United States was facilitated by the major Chicago economist Frank Knight, and by Talcott Parsons, then a junior faculty member in economics at Harvard, Weber's influence on economics and economic history has been limited. But within political science and sociology, by the end of the twentieth century, a great many of the fragments of Weber's thought had become part of the standard conceptual tool kit. It is doubtless a result of Weber's gift as a clarifier of concepts that these tools have shown such an astonishing staying power. In many cases, such as the concepts of charisma and Protestant ethic and the term 'conviction politician' derived from Weber's *Gesinnungsethiker*, the concepts have passed into common usage.

'Weberian' as a description of present scholarship means somewhat different things in different disciplines. In sociology, it means a focus on institutions and institutional forms and an avoidance of abstract assertions about 'society' not grounded in particular specifiable institutions, such as citizenship. In political theory, it means a rejection of value theories that employ Kantian notions of duty and the rational justification of authority or Rawlsian notions of justice in favor of a recognition of the persistence of rationally irreducible conflicts about ultimate ends. In international relations, it means Realism and the recognition that national interests, modified by variant conceptions of national interest and world images, remain the fundamental driving factors in international politics.

See also: Agriculture, Economics of; Ancient Greece and Rome, History of; Authority, Social Theories of; Boas, Franz (1858–1942); Buddhism; Bureaucracy, Sociology of; Capitalism; Civil Society and the Public Sphere: History of the Concept; Civilization, Concept and History of; Cultural Psychology; Culture and Economic Development; Culture: Contemporary Views; Democracy, History of; Economic History; Elites, Anthropology of; Elites: Sociological Aspects; Ethnicity and Ethnic Groups: Historical Aspects; Ethnicity: Anthropological Aspects; Generations, Sociology of; Global History: Universal and World; Historical Thought and Historiography: Classical Period (Especially Greece and Rome); Historical Thought and Historiography: East Asia; Historical Thought and Historiography: Jewish Tradition; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Idealization, Abstraction, and Ideal Types; Judaism; Law: Change and Evolution; Linguistic Anthropology; Macroeconomics, History of From 1933 to Present; Methodological Individualism in Sociology; Migration, Economics of; Migration, Theory of; Migrations, Colonizations, and Diasporas in Archaeology; Organizations, Sociology of; Peasants and Rural Societies in History (Agricultural History); Political Economy, History of; Power in Society; Religion, Sociology of; Religion: Evolution and Development; Sapir-Whorf Hypothesis; Social History; Social Stratification; Society and People: History of the Concept; Sociology, History of; Sociology: Overview; State Formation, Theory of; State, History of; State, Sociology of the; States and Civilizations, of; Status and Role: Structural Aspects; Urban History; Weberian Social Thought, History of; Work, Labor: History of the Concept.

Bibliography

Baier, H., Lepsius, M.R., Mommsen, W.J., Schluchter, W., Winckelmann, J. (Eds.), 1982. Max Weber—Gesamtausgabe. Mohr, Tübingen.

Mayer, J.P., 1956. Max Weber and German Politics: A Study in Political Sociology, second rev. and enlarged ed. Faber and Faber, London.

Mommsen, W.J., 1984. Max Weber and German Politics, 1890–1920 (M.S. Steinberg, Trans.). University of Chicago Press, Chicago.

Mommsen, W.J., Osterhammel, J. (Eds.), 1989. Max Weber and His Contemporaries. Allen and Unwin, London, pp. 234–267.

Radkau, J., 2009. Max Weber: A Biography. Polity Press, Cambridge.

Swedberg, R., 2005. The Max Weber Dictionary: Key Words and Central Concepts. Stanford University Press, Stanford, CA.

Tenbruck, F.H., 1987. Max Weber and Eduard Meyer. In: Mommsen, W.J., Osterhammel, J. (Eds.), Max Weber and His Contemporaries. Allen and Unwin, London, pp. 234–267.

Weber, M., 1946. From Max Weber: Essays in Sociology (H.H. Gerth, C.W. Mills, Trans.). Oxford University Press, New York.

Weber, M., 1951 (Hans H. Gerth, Trans. and Edited). In: Religion of China: Confucianism and Taoism. Free Press, Glencoe, IL.

Weber, M., 1952 (H.H. Gerth, D. Martindale, Trans. and Edited). In: Ancient Judaism. Free Press, Glencoe, IL.

Weber, M., 1958a (T. Parsons, Trans.). In: The Protestant Ethic and the Spirit of Capitalism, student ed. Scribner, New York.

Weber, M., 1958c (H.H. Gerth, D. Martindale, Trans. and Edited). In: The Religion of India: The Sociology of Hinduism and Buddhism. Free Press, Glencoe, IL.

Weber, M., 1961. General Economic History. Collier, New York.

Weber, M., 1963 (E. Fischoff, Trans.). In: The Sociology of Religion. Beacon Press, Boston.

Weber, M., 1976. The Agrarian Sociology of Ancient Civilizations (R.I. Frank, Trans.). New Left Books, London. Humanities Press, Atlantic Highlands, NJ.

Weber, M., 1978. In: Roth, G., Wittich, C. (Eds.), Economy and Society: An Outline of Interpretive Sociology. The University of California Press, Berkeley and Los Angeles, CA.

Weber, M., 1988 (H. Zohn, Trans.). In: Max Weber: A Biography. Transaction Press, New Brunswick, NJ.

Weber, M., 1994. In: Lassman, P., Speirs, R. (Eds.), Political Writings. Cambridge University Press, Cambridge, UK.

Weber, M., 2012 (H.H. Bruun, Trans.). In: Bruun, H.H., Whimster, S. (Eds.), Max Weber: Collected Methodological Writings. Routledge, London.

Weberian Social Thought, History of

Lawrence A Scaff, Wayne State University, Detroit, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C. Turner, volume 24, pp. 16407–16412, © 2001, Elsevier Ltd.

Abstract

Weberian modes of thinking emerged gradually in the twentieth century. Max Weber did not found a school of thought. The posthumous collection of his main work provided a starting point for renewed interest in his ideas. Beginning in the 1920s American scholars discovered Weber's work for reasons of their own, supplemented later by a generation of German-speaking émigrés. The institutionalization of Weberian analysis was a postwar phenomenon, encouraged by professional needs and controversies in the sciences. Today a Weberian paradigm has emerged as one of the principal ways of thinking about society and human affairs.

The formation of the kind of social thought known today as 'Weberian' has an unusual history. The main characteristics of a Weberian approach, perspective, or theory are closely associated obviously with the work of Max Weber (1864–1920), and yet Weber himself never promoted a distinctive school of thought or encouraged a like-minded attachment to any of his leading ideas. Contestation seemed to be his middle name. Weber was not a normal scholar in the ranks of the professoriate. He was academically active for barely 7 years at five different universities at the beginning and end of his adult life, published no books after his dissertation and habilitation, and invested his energies in editorships, scientific essays, handbook articles, newspaper commentaries, encyclopedia entries, polemical tracts, and occasional political writings and speeches. His intended audiences – professional and public, scientific and political – were quite varied. To an outsider his interests could appear unfinished and lacking in focus: an array of topics from the ancient to the modern world, approached from the contrasting perspectives of history, philosophy, law, political economy, economic history, political science, and the new science of sociology. Indeed, Weber's actual published work remained a mere torso of what might have been achieved. In view of his life accomplishments and habits of thought, it seems extraordinary that today we can speak confidently of a distinctive Weberian mode of analysis.

How and why did this development come about? What conditions had to be present for the emergence of a kind of thinking inspired by Weber and his intellectual example? What were the steps taken during the twentieth century that led select groups of scholars to promote Weber's work, to borrow his concepts and ideas, and to extend his questions in new directions? Sometimes it was a matter of direct appropriation, on other occasions the identification of a common problematic traceable to Weber's thinking. Weber liked to speak of a singular concatenation of circumstances in history leading to a particular outcome, and so it was also with Weberian social thought itself: an unusual configuration coalesced to yield what has been called by some today the 'Weberian paradigm.'

Consolidation of the Work

No body of thought can gain paradigmatic status without a textual basis that contains a set of puzzles to be solved and model problems inviting further investigation. When Weber died unexpectedly in 1920 his work was left unfinished and disorganized, the manuscript of *Economy and Society* consisting in large part of notes and scribbled analyses left on his desk. As scholars have long recognized, it is thanks primarily to Marianne Weber (1870–1954), whose biographical comment 'Max Weber's desk is now my altar' (Hanke, 2009), identified in striking prose the commitment to rescue and immortalize her late husband's reputation. Assisted by some of his last students, such as Melchoir Palyi (1892–1970) and Siegmund Hellmann (1872–1942), she compiled and conceptualized the foundation of what came to be understood as the Weberian legacy.

In Marianne's rendition the work consisted primarily of seven edited texts and collected essays and papers. Max Weber himself assisted with the publication only of the first volume of the *Collected Essays in the Sociology of Religion*. Virtually everything else was a product of Marianne's selections and choices as an editor:

- *Collected Essays in the Sociology of Religion*, 3 vols., 1920/1
- *Economy and Society*, 2 vols., 1921/1922; 2nd enl. ed. 1925
- *Collected Political Writings*, 1921
- *Collected Essays on the Theory of Science (Wissen-schaftslehre)*, 1922
- *Collected Essays on Social and Economic History*, 1924
- *Collected Essays on Sociology and Social Policy*, 1924

The sole exception to Marianne's editorial largesse was the Munich lectures on economic history compiled from student notes by Hellmann and Palyi:

- *General Economic History*, 1923

This specification of Weber's interests and topics – religion, economy, law, politics, science, sociology, social relations, and policy – and the ordering of his writings remained the standard until the revisions introduced by the comprehensive critical–historical collected works, the *Max Weber Gesamtausgabe*, that

began publication in 1984 and is only now close to completion. The new collected works promised to establish an authoritative basis for Weber's contributions and to open new avenues for interpretation. But until this century the meaning of the appellation 'Weberian' depended largely on the work Marianne placed in the public domain, supported by the interpretation provided by her own comprehensive biography of her husband's life and work (1926/1988).

The availability of original sources made possible the early reception and translation in the 1920s of Weber's work primarily in two locations: Japan and the United States. In Japan a questioning of capitalist modernity and the role of religion produced an early and independent intellectual perspective, with Weber's *General Economic History* translated in 1927, the same year as the first English translation, and Kajiyama Tsutomu's translation of *The Protestant Ethic and the Spirit of Capitalism* following in 1938. However, for a variety of reasons, including the postwar internationalization of science, the emerging prestige of American universities, and the hegemony of the English language, the appropriation of Weber's work in the United States was especially crucial.

The structural and organizational conditions for the success were set in motion by three sequential developments: first, starting in the 1920s the translations, analyses, and promotional activities of American scholars such as Frank Knight, Talcott Parsons, Edward Shils, and C. Wright Mills; second, the teachings, writings, and translations of the large number of Weimar era émigrés, such as Hans Gerth and Albert Salomon, who fled Nazi Germany in the 1930s and joined universities in the United States; third, the rapid postwar extension of interest in Weber in new directions, expressed in an outpouring of English translations and the institutionalization of his thought in American university curricula and the empirical and theoretical canon of the social science disciplines. These developments in turn provided the basis for the fourth phase: the return of Weberian thought to Germany, and its simultaneous renewal, growth, and extension worldwide.

American Appropriations

The earliest English translation of Weber's work, the student lecture notes compiled as *General Economic History* (1927), was produced by Frank Knight, the founder of the first Chicago school of economics. Knight may seem an unlikely candidate for this honor. But his lifelong interest in Weber grew out of a fascination with the historical origins of modern capitalism and the possible role played by cultural factors and religious beliefs. The Munich lectures seemed to Knight to represent Weber's most mature reflection on these topics. Pursuing similar interests, the young Talcott Parsons was introduced to Weber's work as a graduate student in Heidelberg in 1925. He first read the preface and the essays on *The Protestant Ethic and the Spirit of Capitalism* in the first volume of Weber's *Sociology of Religion*, finding the narrative so compelling that he became immersed in the text 'as if it were a detective story' (Parsons, 1980: p. 39). Studying with Alfred Weber (Max's younger brother), Karl Jaspers, and Karl Mannheim (who was teaching a seminar on Weber), Parsons was soon inspired to write a D.Phil. dissertation topic on capitalism in recent German scholarship. At the same time a proposal arrived to translate a major part of Weber's sociology of religion. With Marianne Weber's support the result after overcoming many obstacles was the book published as *The Protestant Ethic and the Spirit of Capitalism* (1930), in many ways the sacred founding text of Weberian analysis.

The two translations by Knight and Parsons survived for nearly 20 years as the main public sources of knowledge about Weber in English. The textbook quality of the translated Munich lectures, a collection of notes that Weber never saw or corrected, which in any case omitted the conceptual introduction, rendered the book a poor candidate for setting a research agenda. *The Protestant Ethic*, on the other hand, was provocative and controversial from the start. Weber's intellectual daring inaugurated discussion of an alleged 'Weber thesis,' which held in brief that the ascetic Protestant ethic or ethos, oriented toward worldly action and mastery of the world, was one of the important historical factors in promoting the emergence of a 'capitalist spirit.' Weber found this spirit exemplified in the homilies of Benjamin Franklin praising the virtues of personal accomplishment and work in a vocation. He argued further that the social carrier of the ethos was not the compulsory association of a 'church,' but rather the voluntaristic 'sect' in which moral testing of the individual was a condition for membership.

In this early intellectual context, then, to follow a Weberian line of thinking meant essentially (1) to accept as an instructive research hypothesis the proposition that certain kinds of cultural factors can have a profound effect on material economic relationships; (2) to work in the sociology of religion with a basic conceptual distinction between church and sect; and (3) to acknowledge as a methodological principle the use in social analysis of ideal or pure types, such as the construction 'Protestant ethic' or 'capitalist spirit,' as important heuristic devices.

Of course Weber's work encompassed much more than the ideas contained in Knight's and Parsons' translations. Among small circles of scholars efforts were also underway to recover other aspects of Weber's work. In this respect, the most significant fugitive publications emerged at the University of Chicago, where in the 1930s Edward Shils had begun translating numerous Weber texts. His passion for translating Weber focused on selections from the philosophy or theory of science essays collected by Marianne Weber, which included 'Science as a Vocation.' He also attempted the challenging first chapter of *Economy and Society* and 'Politics as a Vocation.' Like Parsons, Shils engaged with Weber's thought for personal edification and out of a sense of intellectual adventure, remarking in retrospect that "reading Max Weber was literally breathtaking. Sometimes, in the midst of reading him, I had to stand up and walk around for a minute or two until my exhilaration died down" (Shils, 1981: p. 184).

Encouraged by Knight, his teacher, Shils began circulating his work to students and faculty in mimeograph format. The texts became important in the effort to define, reform, and integrate the University of Chicago's social science curriculum. Coming to fruition late in the 1930s and supported by the University's President, Robert Hutchins, these pedagogical innovations were the first important institutionalization of a selection of Weber's texts and ideas,

such as the distinction between class and status, the conception of social action, the typology of authority or 'legitimate domination' (*Herrschaft*), and the concept of charismatic leadership.

The introduction of translated Weber texts in the classroom occurred not merely because of professorial interests, but for practical pedagogical reasons in an era when the social sciences were being defined and disciplinary boundaries drawn. In the United States, the debates were intense at Chicago and Harvard, and at public institutions like the University of Wisconsin. Faculty were concerned not only about disciplinary distinctions, but also about countering parochial tendencies in scholarship and integrating knowledge across the social sciences. From this point of view Weber's work offered breadth of coverage, conceptual richness, methodological acumen, and empirical and historical grounding – in a word, a vision of the sociohistorical world and the sciences investigating human affairs. The work's world–historical sweep offered a route to professionalization that could attract enthusiastic support, and it carried the promise of having unquestioned intellectual authority in the battles playing out in the sciences. In the classroom the innovation even took a specific form: armed with Shils's translations, instructors at Chicago addressed the pedagogical questions in the social sciences by placing major portions of Weber's dense prose in the hands even of the uninitiated undergraduates.

This was the kind of educational laboratory in which Weber's ideas began to grow and take root. A more complex and differentiated view of Weber began to emerge: not only the author of the Weber thesis, but also the theorist of bureaucracy and the state, the analyst of legitimacy and authority, the inventor of a new concept of charismatic authority, the defender of an ideal of scientific inquiry free from partisanship, the advocate of a political ethic of responsibility, the theorist of rationality and rationalization processes. Weberian thought started to come of age: it had become more timely, interesting, and challenging.

The early translation of Weber's work into English marked a breakthrough to a new phase in the reception of his ideas. But there is always a politics and sociology of translation. The early work of Parsons and Shils introduced a certain kind of conceptual terminology that assumed a life of its own, often based on the translators' outlook or on the current state of scientific discourse. The act of translation is always an act of interpretation, or more strongly, *mis*interpretation – a sometimes subtle distortion of the original, exaggerating some connotations and deeper intuitions while diminishing others. Citing one famous example, should *Herrschaft* translate as 'authority,' 'domination,' or 'imperative coordination'? The language favored by Parsons and Shils that gained ascendancy in coming decades tended to emphasize the behavioral and causal side of Weber's conceptual syntax, downplaying or avoiding the complexities in concepts like *Entzauberung* (disenchantment) or *Lebensführung* (life-conduct, the way one leads one's life). Subsequently, numerous debates were triggered by such choices and their intellectual consequences: distinguishing the author's actual views from those imposed by the translator, rediscovering an essential but forgotten concept, or reconstructing the theory on an alternative textual basis.

Émigré Perspectives

The second major source of knowledge and interpretation came from those scholars and intellectuals who emigrated from Germany in the 1930s, and who began to contribute to American intellectual life and to staff university social science departments. Some had known Weber personally or experienced him as a teacher and lecturer. They had an effect on the recognition of Weber's work and the growth of interest in some of his key concepts, such as charisma, power, and bureaucracy. In addition to Chicago, in this respect in the 1930s three other universities with prestigious graduate programs became crucial: Harvard, Columbia, and Wisconsin. To this group one should add the New School for Social Research in New York, a university in exile and home to numerous transplanted German-speaking scholars having a deep appreciation for Weber's writings.

Some of the émigrés were especially important for the interpretation, extension, and application of Weber's ideas. Alexander von Schelting is one example. Having met Parsons in Heidelberg, von Schelting later traveled to the United States on a fellowship and renewed their exchange of views. Associated at first with Howard Becker at Wisconsin, he then found a position at Columbia, where he taught a joint Weber seminar with Shils. Von Schelting served importantly as a bridging figure from the Weimar era methodological focus on Weber's work to the reinterpretation of Weber's conception of social action. It was the social action 'frame of reference' that began to preoccupy Talcott Parsons, and for that purpose von Schelting became Parsons's leading authority: the chapters on Weber in *The Structure of Social Action* (1937) bore the imprint of his guidance.

Hans Gerth was another émigré scholar who contributed significantly to the transmission and dissemination of Weber's work. In some ways his role was unique. Associated in Germany with Mannheim, Gerth ended up in 1937 exiled with Mannheim in London in precarious circumstances. Encouraged by Parsons and Shils, in 1940 he was able to find employment in Becker's sociology department in Wisconsin. It was a propitious relocation, for one consequence was Gerth's encounter with two ambitious graduate students: C. Wright Mills and Don Martindale. The Gerth and Mills and Gerth and Martindale partnerships produced some of the most important postwar Weber texts.

At the New School for Social Research many of the émigrés were thoroughly familiar with Weber's work and promoted a broader and more eclectic view of the subject than that espoused by Parsons. In this setting, for example, Albert Salomon's interests took a systematic turn, and in the first issues of Social Research (1934/35) he presented one of the earliest précis of Weberian thought, a survey of the methodology, politics, and sociology for a general audience. His view of the work heralded the emergence of a Weber at variance with the professional sociologist and specialized historical economist put forward previously by American scholars.

Generally speaking, the turn to Weber among the émigrés had to do with several fundamental issues. There was a sense that if one wanted to engage in social research, then one had to come to terms with the figure of Max Weber, widely credited as the most important thinker of their times. As a self-described

outsider, Weber also took on the function of providing orientation to the experience of displacement and homelessness. The rationale for such an appropriation was to be found in the work itself, since Weber could be read to have confronted the specter of capitalist modernity, writ large in America, with an acceptable cosmopolitan and critical sensibility familiar to a European. The result of this émigré perspective was then a different kind of Weber, more attuned to the critical problematics of modern life, the power of the modern bureaucratic state, and the demands placed on the scholar and teacher confronting a world in turmoil. As Franz Neumann formulated the change, previously Weber's work had been reduced to its least inspiring dimensions, but now in America under new conditions it took on a life of its own and broke free of superimposed schemes that tamed its real potential. At Chicago in international relations Hans Morgenthau seemed to prove the point by adapting Weber's political realism to an entire field of inquiry.

In the Columbia University setting émigré scholars like Neumann and Theodore Abel joined others in the kinds of discussions emanating from the New School in Lower Manhattan. Neumann thought Weber's appeal for historically and theoretically grounded inquiry, combined with intellectual sobriety, showed the way for the scholar's vocation. The appeal offered a kind of corrective to unhistorical naïveté and crude, theoretically uninformed empiricism. In his early work Abel also sought to bring Weber's ideas into the orbit of American social science. Columbia provided a perfect setting for the cross-fertilization of these émigré perspectives with the work just emerging in the social sciences, encouraged by well-connected scholars like Robert Merton. One result of such convergence was the postwar seminars on the state and on bureaucracy that convened in 1946, attended by Merton, a newly arrived C. Wright Mills, Daniel Bell, S.M. Lipset, and others – a common forum for historians, sociologists, and political scientists. Minutes of the meetings show a striking reliance on shared knowledge of ideas central to Weber's work, then applied to the development of the modern state in a variety of circumstances, from the Soviet Union's efforts at economic development to the politics of decolonization in the developing world.

Stemming from these discussions, Merton's Reader in Bureaucracy (1952) illustrated one path of development for a new Weberian perspective: no longer disputations over the Weber thesis about ascetic salvation religions, or interpretation of basic concepts, but an application of useful ideas drawn from Weber's texts and extended to novel problems and new research domains. It was this fruitful problem-oriented approach, at times dominated by questions about 'modernization,' in which the ideas could be restated, applied, criticized, elaborated, extended, and renewed.

Postwar Weberians

The third critical event in the United States was the postwar institutionalization of Weber's thought in the canon of social theory, the promulgation of a body of thought that, along with Marx and Durkheim, had to be read and mastered. The record of translations points to the fact that ideas bearing Weber's imprimatur started to become widely circulated only in the 1950s. Considering the textual basis for this development, six publications were especially important:

- Hans Gerth and C. Wright Mills published *From Max Weber: Essays in Sociology* (1946), a student friendly compilation that included a dramatic and informative biographical sketch. It was the Weber reader others had wanted years earlier, now perfectly adapted to classroom use.
- Talcott Parsons published a translation of the first four chapters of *Economy and Society* as *The Theory of Social and Economic Organization* (1947), a project delayed by the war. He consulted a short version of the text written by Shils and von Schelting, and then revised and expanded work begun by H.M. Henderson in England. This was the only part of *Economy and Society* that Weber had prepared for publication shortly before his death. In selecting his own title rather than Weber's, Parsons announced his intention to appropriate Weber for a general theory of society. It was Weber the social theorist that became Parsons's avatar.
- Edward Shils (with assistance from Henry Finch) published *The Methodology of the Social Sciences* (1949), finally releasing for public scrutiny some of his decades-long quest to master Weber's philosophy of science in translation. He introduced readers to commentary on the 'ideal type' and a concept of 'value neutrality' in place of Weber's proposal for 'value freedom' (*Wertfreiheit*).
- Hans Gerth translated and published *The Religion of China* (1951), completing the first volume of Weber's *Collected Essays in the Sociology of Religion*.
- Gerth then teamed with Don Martindale to produce *Ancient Judaism* (1952), the third volume in the *Collected Essays*.
- Finally, Gerth and Martindale published *The Religion of India* (1958), the second volume in the *Collected Essays*.

Completing Weber's three-volume sociology of religion was a major accomplishment, although the original essays were published in translation out of sequence in five different books, a situation that has never been rectified, even though Parsons had warned early-on about the resulting confusion in understanding the structure of Weber's arguments. But Weber's larger comparative historical question – why did modern rational capitalism emerge only in the West? – now came into focus. No one could henceforth write about the sociology of religion or the emergence of modern capitalism without taking Weber's arguments into account.

Regardless of the vicissitudes of partial and piecemeal translations, by 1960 the Anglophone world had a substantial body of Weber's writings, now widely and inexpensively available to scholars, teachers, students, and the general public. At the end of the decade one last addition completed the major projects:

- Guenther Roth and Claus Wittich published *Economy and Society* (1968) in three volumes, the most problematic of unfinished manuscripts, but finally presenting the treatise that many have hailed as the major work of the twentieth century in sociology and social theory.

Having the complete text of *Economy and Society* available was important for establishing the domains of the sociological field of inquiry – the so-called special sociologies of religion,

law, economics, politics, culture, administration, associations, urban life, and so forth – as well as the conception of sociology itself. With the presumed magnum opus in hand, it became possible to conceptualize the Weberian intellectual world, to map its features and to argue about their definition and implications.

Since then there has been continuous expansion and refinement of the textual basis of the Weberian project, casting a much wider net for those interested in the work. There have been some noteworthy surprises: for instance, the presence of a 'structural' perspective in a study like *The Agrarian Sociology of Ancient Civilizations* (1976) that seemed to some to have affinities with neo-marxist class analysis and to confirm the insights of what has been called Weberian marxism. The text actually revealed a vocabulary Weber shared with Marx and many others in German political economy. In addition, the political commentary in Weber's essays on Russia seemed to anticipate subsequent twentieth-century development and illustrated an analytic approach to large-scale social change, if not precisely a theory of revolution. Or Weber's early writings on the stock and commodity exchanges, previously unknown outside specialist circles, demonstrated a hitherto hidden side of Weber as a political economist and economic sociologist concerned with macro-processes, micro-level choices, and their relationship in rule-governed markets. The long-term effect of these postwar retrievals has been a significant reshaping of the Weberian field of inquiry, an extension of the horizon and an expansion of interests into new and uncharted territory.

Homecoming and Renewal

The identification, translation, and analysis of Weber's writings starting in the 1950s, especially in intellectual circles in America, was important not only for the social science disciplines in the Anglophone world, but because the intellectual capital generated by this activity became the basis for the fourth major development in Weberian thought: the reintroduction of Weber's work in Germany, and its transnational renewal under changed circumstances.

While the growth of Weberian perspectives in America was the precondition for the work's return home in the 1960s, it was also the catalyst for the critical reassessment of Weber's contribution and its liberation from limiting preconceptions. The Weber who returned had been domesticated and appropriated by a Parsonian framework consistent with the behavioral revolution in the social sciences. The homecoming was thus laced with controversy, as the efforts in Germany to recover the work achieved public notoriety during the political struggles and academic debates of the 1960s. The abstracted version of Weberian sociology promoted by American scholars like Parsons and in a very different way, Reinhard Bendix, seemed to critics ill-suited to address the burning issues of the day. To be sure, Parsons had channeled Weber's ideas into the postulates of systems theory, structural functional analysis, and the explanation of modernization and social evolution, while Bendix had promoted historical sociology with a Weberian accent and emphasized Weber's appeal for empirical studies free from partisanship. Neither reconstruction was designed to address a looming political crisis. With the disputants talking past each other, the rationale for a distinctive Weberian approach was easily lost in a cacophony of opposing voices.

But the contentious decade of the 1960s turned out to be a prelude to the serious incorporation of Weber's work in the revisionist scholarship of the 1980s and afterward. In Germany evidence was provided by the major and quite different analyses of scholars like Wolfgang Mommsen, Jürgen Habermas, and Wilhelm Hennis. In the social sciences internationally Weberian ideas cast a long shadow over the critique of modernization theory and the work emerging in comparative historical sociology, political and economic sociology, and the investigation of the development of the modern state. These efforts were largely a transatlantic affair, though because of his international fame, the most noticeable contribution came from Habermas.

In his two-volume treatise, *The Theory of Communicative Action* (1984, 1987), Habermas returned to Parsons' theory of social action, which had been derived from a reading of Weber. But instead of accepting Parsons' formulation of Weber's social thought, Habermas set out to develop an ambitious critique of rationalism and rationalization processes. Most significantly, his understanding of the peculiar nature of rationalism in the West was based upon Weber's category of instrumental or purposive rationality (*Zweckrationalität*), that is, the kind of goal-oriented action that aims to maximize an efficient matching of means to ends. Working with Weber's category, he pushed the critique in different directions, arguing in essence that instrumental rationality had run amok in the modern world, invading the human world of everyday experience, violating the natural world, and threatening the very existence of humankind. This sobering analysis served as the platform upon which Habermas launched his lifelong effort to rescue enlightened reason and to find a way out of the cul-de-sac of a fully rationalized and disenchanted modernity.

Habermas' critique, supplemented by the work of others, exemplified an important new direction: for now in social theory Weber had became not the theorist of social action, or the origins of capitalism, but of the world–historical dynamic of rationalization. This was a potent invention. It mean that Weberian thought was linked to the most consequential problem of modern social theory: the radical differentiation and separation of the spheres of value – science, morality, art – and the dissociation or 'decoupling' of the human life–world from the social mechanisms of economic, administrative, and legal control. Weberian social theory now belonged to the Western philosophical tradition. It spoke to the most fundamental questions of the human condition.

Weberian Social Thought Today

The sustained critique of the initial uses of Weber's ideas was essential to liberating Weberian thought and making possible the notion of a distinctive Weberian perspective, approach, theory or 'paradigm.' Today Weberian approaches appear in widely dispersed domains of inquiry, from the significance of religion and the sacred to rational choice theory and formal modeling – so-called 'analytic Weberianism' (Kiser and Baer, 2005). Coherence amid such variety is provided by the sense

that the contemporary Weberian paradigm insists on an approach which is configurational, multicausal, multilevel, and committed to the idea that social and economic action is always embedded in social institutions and practices. It tends also to favor inquiries that are historical and comparative.

The other essential point to make about Weberian social theory is that, following Weber's lead, it is not confined to a single mode of explanation. That is, the advantage of Weberianism is that it can combine (1) macro-level structural analysis that pays attention to economic class and social status; (2) institutional analysis that investigates how and why rules of the game are devised and then constrain human action; (3) cultural analysis that inquires into the values, beliefs, and symbolic expressions which endow action with meaning; and (4) micro-level deployment of models of purposeful or instrumental rational action and its intended and unintended consequences. The choice of a particular mode of explanation depends not on methodological precepts, but on the research problem under investigation. From the Weberian perspective the questions asked take precedent over the toolkit of methods and determine the structuring of an inquiry.

However, today Weberian social thought is not confined only to academic specialties. It has migrated worldwide into the sphere of public opinion, political comment, and cultural interpretation. The problems of the new century have moved on, and Weber's thought has accompanied them. In this regard the thought has remained timely because of its problematic, the questions asked about the forces shaping our modern world. So long as such questions persist, Weberian thought promises to remain edifying, engaged, and capable of renewal.

See also: Bureaucracy and Bureaucratization; Bureaucracy, Sociology of; Bureaucratization and Bureaucracy, History of; Economic Sociology; Rationality in Society; Religion and Economic Life; Religion and Stratification; Religion, Sociology of; Sociological Theory; Sociological Traditions; Weber, Max (1864–1920).

Bibliography

Albert, Gert, et al. (Eds.), 2003. Das Weber-Paradigma: Studien zur Weiterentwicklung von Max Webers Forschungs-programm. Mohr (Siebeck), Tübingen.

Baehr, Peter, 2002. Founders, Classics, Canons: Modern Disputes over the Origins and Appraisal of Sociology's Heritage. Transaction, New Brunswick.

Derman, Joshua, 2012. Max Weber in Politics and Social Thought: From Charisma to Canonization. Cambridge University Press, Cambridge.

Hanke, Edith, 2009. "Max Weber's desk is now my Altar:" Marianne Weber and the intellectual heritage of her husband. History of European Ideas 35, 349–359.

Kiser, Edgar, Baer, Justin, 2005. The bureaucratization of states: toward an analytical Weberianism. In: Adams, J., et al. (Eds.), Remaking Modernity: Politics, History, and Sociology. Duke University Press, Durham, pp. 225–245.

Parsons, Talcott, 1980. The circumstances of my encounter with Max Weber. In: Merton, R., Riley, M. (Eds.), Sociological Traditions from Generation to Generation. Ablex, Norwood, pp. 37–43.

Salomon, Albert, 1934/35. 'Max Weber's methodology,' 'Max Weber's political ideas,' 'Max Weber's sociology'. Social Research 1, 147–168, 368–384; 2, 60–73.

Shils, Edward, 1981. Some academics, mainly in Chicago. The American Scholar 50 (Spring), 179–196.

Scaff, Lawrence A., 2011. Max Weber in America. Princeton University Press, Princeton.

Scaff, Lawrence A., 2014. Weber and the Weberians. Palgrave Macmillan, New York.

Swedberg, Richard, 2005. The Max Weber Dictionary: Key Words and Central Concepts. Stanford University Press, Stanford.

Weber, Marianne, 1926/1988. Max Weber: A Biography (H. Zohn, Trans.). Transaction, New Brunswick.

Weight Regulation

Jane Wardle and Fiona Johnson, Cancer Research UK Health Behaviour Research Centre, University College London, London, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Identifying processes underlying weight regulation in humans poses a challenge. Dramatic increases in body weight worldwide implicate environmental causes for the problem of obesity, but the high heritability of weight and striking interindividual differences in adiposity, even within similar environments, demonstrates genetic influence. One model that incorporates both environments and genes argues that genetic factors influence appetitive traits, which affect vulnerability to overeating in 'obesogenic' environments. Traditional approaches to supporting behavior change for weight regulation are often initially successful, but fail longer term. Strategies that take account of appetitive differences could help to assist with long-term maintenance of weight loss.

Weight regulation in humans is poorly understood. Most people's weights are relatively stable over days, months, and even years, despite considerable day-to-day variation in energy intake and dietary quality. On this basis, fat storage has been assumed to be under homeostatic control. However, defense against weight loss appears to more efficient than that against weight gain. Explanations for this imbalance have tended to focus on human evolutionary past (Prentice et al., 2008). The 'thrifty gene' hypothesis proposes that genes predisposing to energy storage survived into the modern human genome because they provided better survival and reproductive success over times of famine. Others have argued that there have also been pressures against weight gain, because higher body weight could increase the risk of predation. Either way, it is clear that although some people show little tendency toward weight gain and appear to need no deliberate effort to maintain a healthy weight, a tendency for progressive weight gain affects a substantial proportion of the population. Among those who show weight gain tendencies, some also find weight loss, and especially maintenance of weight loss, extremely difficult, and may reach very high levels of adiposity, resulting in the clinical problem of obesity. Hence, until recently clinical obesity was understood as a genetically influenced metabolic abnormality.

However, views about the etiology of obesity have been challenged by the dramatic rises in population weights and rates of obesity worldwide. Starting in the United States in the 1980s, population weights began to rise and the prevalence of obesity increased steadily across the globe (Finucane et al., 2011); now exceeding 90% in some Pacific nations, over 30% in the United States and the Middle East, and over 20% in Australia and many European countries. Even in the developing world where undernutrition is common, obesity rates are rising leading to a dual burden on health care services. This astonishing rise in a biological parameter that was once thought to be primarily under homeostatic control strongly implicates environmental, cultural, or social influences.

The 1950s saw a change of perspective on 'weight problems,' when the American Public Health Association declared obesity to be not just a matter of personal or aesthetic concern but a major health threat. This was on the basis of accumulating evidence of associations between obesity and a wide range of chronic diseases as well as reduced life expectancy. In 1997, the World Health Organization began a consultation that confirmed the gravity of the emerging global obesity crisis and called for investment into effective long-term treatment and preventive strategies. In the resulting report, the key role of behavior, for both prevention and treatment, was recognized (WHO, 2000). The growing obesity crisis highlighted the importance of understanding the process of weight regulation.

Increasing Population Weights: Environmental Factors

The worldwide increases in body weight, and the timing and geographic distribution of weight changes, strongly implicate modern lifestyles in upsetting the balance between energy intake and energy expenditure. Changes in transportation and mechanization have reduced the energy demands of everyday life, and combined with the growth in sedentary, screen-based entertainments, this has had the potential to greatly reduce activity-related energy expenditure. Substantial changes have also taken place in the food environment, with developments in food production, processing, storage, and preparation since the middle of the twentieth century, resulting in highly palatable and energy-dense foods becoming progressively more accessible, more strongly marketed, and cheaper. Whether lower energy expenditure or higher intake is the principal causal factor is debated, but either way it is clear that the modern environment has undermined the capacity of internal regulatory systems to match energy intakes with energy expenditure.

The change in obesity prevalence has also been linked with changes in its demographic distribution. Obesity was once more common among the wealthier groups in society, and in developing countries, the rising obesity rates are still seen primarily in the affluent, urban communities. But in affluent Western countries, the patterning has reversed, with obesity now being more common in people from lower socioeconomic status (SES) backgrounds (Shrewsbury and Wardle, 2008).

Although there is little doubt that changes in the food environment are important, less is known about which particular aspects are primarily responsible for promoting

overconsumption. If humans are, at least in part, opportunistic eaters, accessibility (e.g., availability, visibility or price) would be expected to play a role. Studies have shown that foods displayed prominently – for example, near the front of a food display – are more likely to be selected. In some countries, this evidence has led to supermarkets limiting displays of snack foods for children at the checkout, and to school meal services placing fruits and vegetables prominently on their displays to increase consumption. The sight and smell of highly palatable foods has a well-established influence on appetite. These responses may have a biological foundation in the need to meet future, as well as present, metabolic needs (Johnson, 2013). Food variety also influences consumption, partly at least through the phenomenon of sensory-specific satiety; allowing people to be sated for a particular food but retain an appetite for a different food (e.g., the appeal of dessert). Again this may be a biologically driven adaptation that once served to reduce the risk of micronutrient deficiencies associated with a limited dietary repertoire. External cues may also operate through influences on social norms. For example, the influence of portion sizes on consumption has been attributed to the supplied portion size giving an implicit social norm for the 'right' quantity to consume. This and other environmental influences may be outside consumer awareness (Vartanian et al., 2008), but are potentially modifiable aspects of the food environment, which could reduce the risk of overconsumption.

Individual Variation in Weight and Weight Regulation

The obesogenic food environment has not resulted in uniform increases in weight across the population. This could be partly due to differential exposure to the environment, but the fact that variation in weight is seen even within families who share the same home and local environment suggests that there are substantial individual differences in vulnerability to weight gain. In England, the slimmest 10% of the population in 2003 had a BMI very similar to the slimmest 10% in 1993, but the BMI increase in the top 10% was over 3.5 BMI units, and this pattern was the same in all socioeconomic groups (Wardle and Boniface, 2008). In fact, it is morbid obesity (BMI > 40) that has risen most dramatically, increasing by almost 50% in England in the decade between 2001 and 2011, and by 70% between 2000 and 2010 in the United States. These findings are consistent with the idea that it is the most genetically susceptible individuals who are most vulnerable to the obesogenic environment.

Both thinness and fatness are known to run in families; implicating genes not just in obesity but in the level of adiposity across the weight spectrum. Family studies cannot definitively distinguish shared genes from a shared home environment. However, adoption studies demonstrate elegantly that familial resemblance in weight maps onto biological, more than environmental, relatedness, with adopted children's weights showing no correlation with the weights of their adoptive parents (Stunkard et al., 1990). Twin designs provide a method of estimating of the relative balance of genetic and environmental influence in explaining variation in weight by comparing the extent of resemblance in genetically identical (MZ) and fraternal (DZ) twins. The typical finding in all the early twin studies was that upwards of 70% of the variation in weight could be attributed to genetic influence (Grilo and Pogue-Geile, 1991). Adoption studies have confirmed the same degree of resemblance in twins who are adopted apart as those who are reared together (Stunkard et al., 1990). One concern is that many of the large twin studies on weight were carried out before the 'obesity epidemic' gained traction, and so they may not reflect the impact of powerful environmental forces that tend to increase weight. However, in a large pediatric study of twins born after the onset of the obesity epidemic, the extent of genetic influence was as high or higher as it had been in earlier times (Wardle et al., 2008); implying that expression of obesity risk genes is not diminished by a 'permissive' food environment and may even be enhanced. Both gene–environment interactions (genetically mediated variation in susceptibility to environmental disruption of weight regulation) and gene–environment correlations (genetically mediated variation in the tendency to seek out obesogenic food environments) are thought to play a role in enhancing the toxicity of the obesogenic environment for genetically susceptible individuals.

The high heritability of weight raised hopes that it would be possible to identify the genes influencing weight. In 1997, the discovery of a mutation in the leptin gene in mice that was associated with leptin deficiency and severe, early onset obesity provided the first evidence for a specific obesity gene. The subsequent demonstration of a similar mutation in humans, followed by evidence that leptin replacement in a child with this mutation resulted in weight loss, had a profound influence on the understanding of the biological mechanisms underlying weight. However, the number of families carrying the leptin mutation proved to be small, and, therefore, this abnormality could not explain so-called 'common obesity.' Furthermore, leptin levels in most obese people are high rather than low. With the advent of genome-wide association studies, a common variant in the fat mass and obesity associated (*FTO*) gene was discovered to have a strong association with body fat from age 7 years upwards (Frayling et al., 2007); albeit with a much smaller effect than the leptin gene. The 16% of adults who are homozygous for the risk allele weigh about 3 kg more and have 1.7-fold increased odds of obesity compared with those not inheriting a risk allele. Since that time, many additional genetic variants have been discovered, and the current view is that there are hundreds of gene variants, each with a very small effect size, that underlie the heritability of weight. Individuals inheriting a lower number of high-risk variants would be expected to be thin, while those with a large number would – given an environment that makes it possible for them to get larger amounts of food – be overweight or obese. Generating a polygenic risk score based on the number of high-risk variants out of 32 of the single nucleotide polymorphisms identified to date, already shows a quantitative association with weight (Llewellyn et al., 2014). At present, the total variance in weight explained by known genes is still a fraction of the genetic influence identified in twin studies – a problem known as the 'missing heritability' – although recent analyses using the genome-wide complex trait analysis method show that molecular genetic influences are getting closer to the twin estimate.

Given the evidence that environmental factors influence weight at a population level, what evidence is there that the local environment of the individual is an important determinant of their weight status? The observation that there are strong – and possibly increasing – SES differences in weight is consistent with the idea that aspects of the home environment and local area are important. However, it is also possible that factors such as social norms for eating and weight or nutrition knowledge, are the primary mediators of SES differences; operating by modifying vulnerability to the external environment. There is some evidence that lower SES neighborhoods have more fast-food outlets, less availability of fruit and vegetables, and fewer pleasant green spaces for activity. However, although there has been a lot of interest in the idea that the local environment could play a role in determining the local prevalence of obesity, the findings have been inconsistent. Equally, there has been interest in the idea that the home environment plays an important role in setting children's weight trajectories. To date, comparisons of homes with more versus less availability and accessibility of energy-dense or healthy foods, have not found consistent associations with children's weights, but these studies have been dogged by measurement difficulties. Twin studies with large samples can also provide estimates of the shared environment effect. So far twin analyses indicate an effect of the home environment in the early years, but this dwindles to little or no effect by adulthood. Adoption studies also find that twins reared apart resemble each other as much as twins reared together by adult life (Stunkard et al., 1990); indicating that the long-term influences of the childhood home environment may not be as strong as is often hypothesized.

The fact that the most influential obesogenic environment effects appear to be broader than the home and the local neighborhood does not mean that environmental control is irrelevant. It is entirely appropriate for obesogenic exposures to be controlled and contained in settings for children (home, day care, school), and in health care settings (hospitals) and workplaces. Ultimately, control in multiple local environments could coalesce into a powerful broader force, especially with government support.

Appetitive Behavior as the Link between Genetic and Environmental Influence

One resolution to the paradox of weight being both highly heritable *and* under strong environmental influence is the idea that obesity risk genes do not 'cause' excess weight directly by effects on metabolism, but rather they influence susceptibility to environmental opportunities for overconsumption. In an environment with multiple attractive eating opportunities, people who are most responsive to food cues and least responsive to internal satiety cues will be most likely to overeat (Carnell and Wardle, 2008). According to this model, appetitive traits provide a potential intermediate behavioral phenotype linking genetic and environmental processes.

An appetitive model of obesity was first described in 1968, when Stanley Schachter published a seminal paper identifying common causes for obesity in humans and rats. He used the preload paradigm to examine adjustment of subsequent food intake in relation to the calorie content of a high-energy 'preload' (usually a milkshake or ice cream), and showed that obese adults (or rats) were less accurate at downregulating their appetite after a preload than normal weight individuals (Schachter, 1968). He also showed that obese adults were more responsive than normal weight adults to the palatability of food, increasing their intake relatively more when presented with highly palatable food. This model was the first to emphasize psychological factors – in this case responsiveness to internal and external cues – as key determinants of obesity. However, the model fell out of favor as attention turned to dietary restraint as an important influence on weight (see below).

There has now been renewed interest in appetitive traits as risk factors for obesity – reprising Schachter's model, and emphasizing interactions between the individual and the environment (Carnell and Wardle, 2008); driven in part by developments in our understanding of the genetics of obesity. Importantly, the model was extended to demonstrate that there are not only differences between the eating behavior of obese and normal weight individuals, but also that appetite is quantitatively related to weight across the weight distribution. Studies of children show a more or less linear association between weight status and both food responsiveness and sensitivity to feelings of fullness (satiety responsiveness).

The key finding was that 'obesity genes' influence appetitive processes. All monogenic forms of obesity identified to date are characterized by excessive appetite and diminished satiety. The first gene variant associated with weight in the general population, *FTO*, is highly expressed in the hypothalamus, an area of the brain linked with appetite, and expression of *FTO* is responsive to short-term variation in energy balance from underfeeding or overfeeding. Obesity-related gene variants also predict brain activity in response to food stimuli. The most direct evidence comes from human studies showing that *FTO* is related to satiety sensitivity, and that the effects are seen from early in life. Using a polygenic obesity risk score (described above), clear associations can be seen with satiety sensitivity, and there is evidence that satiety sensitivity may mediate the effect of obesity genes on weight (Llewellyn et al., 2014). These associations provide insight into the pathways that lead from genetic susceptibility, through behavior, to weight; and go some way toward explaining individual differences in weight regulation.

Deliberate Weight Control: The Dietary Restraint Story

Regardless of genetic and environmental influences, weight regulation is usually considered to be primarily the responsibility of the individual. However, messages on how weight can be controlled have often been confused and contradictory. Nowhere is this more apparent than in the longstanding debate over dietary restraint. The concept of dietary restraint was first outlined in 1975, describing an eating pattern consisting of attempted control over eating, but also characterized by regular failures of control. A raft of early studies showed that restrained eaters were less responsive to satiety, more responsive to palatability, and more likely to overeat in response to negative

affect; leading to the idea that chronic restraint itself led to loss of control over eating and a risk of subsequent weight gain (Herman, 1978); and to the adage 'dieting makes you fat.'

This was the era – the 1970s – when anorexia nervosa and bulimia nervosa were seen as the scourge of young women's psychological health, and when many women across the weight spectrum aspired to unhealthily low body weights. Psychological research on appetite and weight regulation in obesity dwindled, and the focus moved to strategies to improve body image and constrain efforts at weight control in young women. However, whether by virtue of the campaigns to improve body image or for other reasons, there were little or no further increases in rates of eating disorders, and the world of weight research was taken over by the dramatic rise in obesity prevalence that began in the late 1980s in affluent Western countries, and over the next 20 years in the developing world. The idea that this was all a reaction to excess dieting lost all plausibility – not least because the rise in obesity was equally strong in men (who certainly do not have the same body image issues as women) and women, and is emerging in countries and cultures where there is no 'slim ideal' as much as those that worship slimness.

One of the problems with the early studies on dietary restraint was that they could not exclude the possibility that a tendency toward overeating was the original trigger for the individual to try to restrict their food intake; which could explain why dietary restraint was associated with loss of control. The first psychometric measures of dietary restraint did not distinguish between the attempt to control eating (now often called restrained eating) and failures of control (called counterregulation or disinhibition), although later approaches to measuring restraint have acknowledged that the two behavioral tendencies are distinct, albeit often cooccurring (Johnson et al., 2012).

Restraint theory offered several models to explain the observed link between restraint and failures of control. The cognitive model proposed that after a high-calorie preload (in the laboratory) or after a dieting transgression (in everyday life), the dieter 'reasons' that they might as well take advantage of the situation and leave dieting until tomorrow. The psychobiological model suggested that dietary restraint interferes directly with an individual's capacity to respond to internal and external cues that influence food consumption. The emotional model suggests that the dieter copes with negative emotional states with dissociation, and as part of this, loses contact with their restrictive intentions.

Despite all these 'explanations' for why restrained eating may be ineffective, the body of clinical research suggests that restraint is actually associated with *less* overeating and *reduced* weight in the long term, and with better weight maintenance (Johnson et al., 2012). Increases in restraint are associated with better treatment outcomes, and formerly overweight individuals who have maintained a significant long-term weight loss tend to show increases, rather than reductions in restraint (Wing and Phelan, 2005). Nonetheless, it is also clear that restraint is difficult, and it is more than plausible that attempts at deliberate restraint trigger biological and psychological processes that tend to oppose weight loss. The key question may therefore be how to help people to exert effective dietary restraint in their daily lives.

Body Weight Perception and Weight Control

Provision of information and advice has been a key element of public-health attempts to address rises in obesity, under the assumption that recognition of unhealthy body weight is one of the prerequisites (although not a sufficient condition) for weight regulation in those with a tendency to gain weight. Several studies have shown that perceived weight status is a better predictor of efforts at weight control than actual weight status. However, promoting recognition of the need for weight control among the overweight and obese population is a conflicted area because raising awareness of overweight could have adverse effects on body image. The dominant discourse in the area of body image has been around normal weight women having excessive concerns about their body weight, and the need to ensure that young women – regardless of body size – grow up with a positive body image. However, recent studies have shown that perceptions of what constitutes a healthy weight are diverging from clinical guidelines, with increasing acceptance and normalization of overweight in all age groups. Many overweight and obese adults are unaware that they are carrying excess weight or that there is a need for weight control and underestimation of body size is increasing (e.g., Burke et al., 2010). Lack of parental awareness of overweight in children has similarly been highlighted as a barrier to healthy lifestyle change at the family level (Parry et al., 2008). Several explanations for inaccurate weight perceptions have been suggested. People are strongly influenced by prevailing norms of appearance and behavior, and the increases in population weights in recent decades may have 'upsized' perceptions of what constitutes a 'normal' body weight. Media coverage of the 'obesity epidemic' is often illustrated with images of extreme obesity (Patterson and Hilton, 2013), which may decrease recognition of more common levels of overweight and obesity. To tackling the growing problem of weight misperception, doctors in the United Kingdom are being increasingly encouraged to assess their patients' BMI and give advice on the need for weight control, and the National Child Measurement Programme is providing feedback to parents about their child's BMI. Accurate understanding of weight-related health risk may be a first step toward encouraging lifestyle changes to counter further weight gain.

Supporting Behavior Change for Weight Control

There is general agreement that simply providing education and advice regarding the components of a healthy diet and the desirable levels of physical activity has limited efficacy as a weight control strategy. Most obese people are broadly aware of the main components of a healthy diet and which foods are high in fat and sugar, although their detailed knowledge may be limited and they may misjudge their own diet quality. They are also aware of the importance of physical activity, even though they may overestimate their activity level. Many – at least at the higher weight levels – are also concerned about their weight; sometimes for health reasons and sometimes for appearance. The success of the commercial weight loss 'industry' and the popularity of books, television programs,

and Web sites on the topic of weight loss, all indicate that public motivation to lose weight is high.

For obese individuals seeking expert assistance with weight control, cognitive behavioral treatment (CBT) has been seen as the 'gold standard.' In essence, it aims to provide people with the skills to control their tendencies toward overeating and to adhere to recommendations for healthy eating and physical activity. It has been aligned particularly with achieving a permanent shift in behavior, and with modest weight loss goals; in contrast to drug treatments or surgery. The key elements of behavioral treatment are 'stimulus control' (reducing exposure to cues for eating, limiting the number of places in which eating occurs), alongside the standard behavioral procedures of self-monitoring, goal setting, and self-reinforcement. The cognitive elements include strategies to modify self-defeating cognitions and address the cognitions that contribute to the lapse–relapse sequence. CBT continues to produce positive results (Van Dorsten and Lindley, 2011).

As CBT principles have become more established and more widely disseminated, many dietary treatments include a behavioral element in parallel with a range of diet or exercise recommendations, including commercial programs such as Weightwatchers. The role of psychologists has largely moved away from delivering obesity treatment to providing advice and training to the wide range of practitioners who can deliver it at lower cost and on a larger scale.

Although the majority of overweight and obese people can achieve significant weight loss with a range of diet and exercise regimens, a much greater challenge is to sustain change over the long term. Long-term weight losers need to adhere to their dietary plans and stay highly active; they need to monitor their weight assiduously, and take steps as soon a deviation occurs. Failure to adhere to planned lifestyle changes is a key factor leading to weight regain in the majority of those who succeed in losing weight (Pagoto and Appelhans, 2013). Part of the difficulty is biological in origin, because weight loss induces physiological changes, which predispose to weight regain. Total energy expenditure appears to be reduced to a greater extent than expected from the loss of body mass (Rosenbaum et al., 2008), and there is evidence for persistent changes in appetite and satiety hormones, and increased neural response to food cues in weight-reduced individuals (Sumithran and Proietto, 2013); factors that conspire to increase the challenge of maintaining changes in eating behavior.

In its early days, there was hope that the control strategies included in CBT programs would equip patients with skills that could be applied long term. In practice, once the active treatment phase of CBT is over, gradual weight regain usually begins. Among the suggestions for achieving longer-term weight control is continuing therapeutic support; consistent with the view that few chronic conditions are successfully treated with brief interventions. Recent results from the Look AHEAD study, which started with a highly restrictive phase followed by a long-term maintenance phase that incorporated behavioral support, showed significant long-term weight change (2014). One issue may be that the primary focus of this intervention was prevention of diabetes, and the differences in treatment goals or the types of participants might mean that effects would be different in treatments focusing primarily on weight. Also, many obese patients are permanently on the lookout for a new treatment that will give better results, so long-term treatment 'merely' to prevent regain is not necessarily an attractive offer.

There is growing interest in other dietary approaches, particularly the use of meal replacements with low or very low calorie diets, to achieve rapid weight loss. This is sometimes delivered in the context of behaviorally based support, as in the Look AHEAD study. The last decade has also seen the rise and fall of the Atkins diet, and regimens of interspersed days of strict restriction and unrestricted eating are now attracting attention. Regardless of the specific dietary goals, some of the core features of CBT, such as self-monitoring, goal setting, and stimulus control would in principle be useful adjuncts; but this has not been a key focus of behavioral research.

There is increasing interest in interventions that can be delivered without the need for specialist therapeutic skills. Many commercial programs are based broadly on CBT and can deliver effective programs without the superstructure costs associated with health systems (Jebb et al., 2011). Self-help versions of CBT programs have been developed, though not fully tested. Of particular interest is the possibility that Web-based services might be a highly cost-effective enhancement to purely self-help treatments (Hedman et al., 2012). Alternative platforms for the delivery of more automated weight management support are also gaining increasing interest. Internet and mobile applications provide the user with the opportunity for an interactive experience without drawing on the costly time of a therapist. Self-monitoring is an intervention component that lends itself well to 'e-health' approaches and which has been repeatedly shown to assist with adherence to lifestyle changes (Ramage et al., 2014). Monitoring and recording fluctuations in weight, eating, and physical activity can help ensure that deviations from desired patterns are noticed and can act as a spur to remedial action.

The Future

Our understanding of the pathways linking genes, environment, and behavior to weight regulation is moving on rapidly. It is now clear that many people who have difficulty with weight regulation are facing a double onslaught of biological and environmental pressures.

It is unlikely that the broader environment will change dramatically in the near future. Governments are reluctant to regulate in areas where public support is low, and it seems unlikely that there will be significant public support for making the food supply less accessible, less palatable, or less affordable. It also seems unlikely that in a market-driven economy, food companies will be motivated to work out ways of making a profit by selling less food. But there could be changes in consumer preferences and pressures that will make it worthwhile for industry to apply technical expertise to producing foods that please the palate, promote satiety, and meet preferences for good value, all the while helping consumers to eat within metabolic needs. Despite this, it seems almost certain that psychological ingenuity will be needed to help people find ways of circumventing their overeating tendencies and enable them to achieve energy balance.

The evidence from recent years that genetic risk for obesity is mediated through higher food responsiveness and lower responsiveness to satiety could guide research into interventions that control these underlying tendencies. Strategies such as careful portion control and slow eating are already used to circumvent poor satiety responsiveness, but there may be other opportunities. For example, strategies to control deployment of attention, used to date in the mental health field, could possibly contribute to helping people control high food responsiveness.

There are some novel behavioral approaches; one of which harnesses habit-formation principles. Habits are behaviors that are performed automatically in response to environmental cues. Most of the health-related literature focuses on how to break bad habits, but creating new 'good' habits could have a role, and in principle could make actions that are difficult to remember easier to perform. The key to habit formation is repetition in a consistent context. Once automatic, the behavior requires less conscious effort to maintain and is more resistant to change than intentional behaviors (Lally et al., 2008). This approach is being tested in a number of behavioral domains including weight control – where the focus is on creating new 'negative energy balance habits' that will be easy to keep up in the long term.

There has also been interest in the possibility of 'training' children or adults in self-regulation, with a focus on intermediate skills such as goal setting, self-monitoring, emotion regulation, and coping strategies. These have had some success in improving skills and self-efficacy, and bringing about positive weight changes (Johnson et al., 2012). There is evidence for a generalizing effect of self-regulation training, with enhancements in one area (e.g., adherence to a physical activity program) improving an individual's ability to self-regulate across the board. It has also been suggested that controlled and supported exposure to tempting situations might contribute to strengthening self-control (Papies et al., 2008). Whether interventions can harness this effect requires further investigation.

Greater genetic knowledge and understanding of the behavioral traits associated with genetic differences may allow more effective tailoring of interventions to the specific vulnerabilities of individuals, and as in other areas of medicine there is a push toward personalized treatment. In the case of obesity, achieving an energy deficit is required regardless of the specific underlying cause, but if genetic information could help highlight individual vulnerabilities, it could be a benefit to those seeking to make changes to their eating and activity behavior. Even more promising is the possibility that a genetic link to weight control advice would provide an element of personalization that might make the information more salient and enhance adherence. Information on genetic risk has also been identified with reduced sense of self-blame, and this could be of benefit to well-being with positive effects on weight control actions.

See also: Body Image; Epigenetic Inheritance; Evolutionary Medicine.

Bibliography

Burke, M.A., Heiland, F.W., Nadler, C.M., 2010. From "overweight" to "about right": evidence of a generational shift in body weight norms. Obesity (Silver Spring) 18, 1226–1234.

Carnell, S., Wardle, J., 2008. Appetite and adiposity in children: evidence for a behavioral susceptibility theory of obesity. American Journal of Clinical Nutrition 88, 22–29.

Finucane, M.M., Stevens, G.A., Cowan, M.J., Danaei, G., Lin, J.K., Paciorek, C.J., Singh, G.M., Gutierrez, H.R., Lu, Y., Bahalim, A.N., Farzadfar, F., Riley, L.M., Ezzati, M., 2011. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. Lancet 377, 557–567.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., Shields, B., Harries, L.W., Barrett, J.C., Ellard, S., Groves, C.J., Knight, B., Patch, A.M., Ness, A.R., Ebrahim, S., Lawlor, D.A., Ring, S.M., Ben-Shlomo, Y., Jarvelin, M.R., Sovio, U., Bennett, A.J., Melzer, D., Ferrucci, L., Loos, R.J., Barroso, I., Wareham, N.J., Karpe, F., Owen, K.R., Cardon, L.R., Walker, M., Hitman, G.A., Palmer, C.N., Doney, A.S., Morris, A.D., Smith, G.D., Hattersley, A.T., Mccarthy, M.I., 2007. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316, 889–894.

Grilo, C.M., Pogue-Geile, M.F., 1991. The nature of environmental influences on weight and obesity: a behavior genetic analysis. Psychological Bulletin 110, 520–537.

Hedman, E., Ljotsson, B., Lindefors, N., 2012. Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Review of Pharmacoeconomics & Outcomes Research 12, 745–764.

Herman, C.P., 1978. Restrained eating. The Psychiatric Clinics of North America 1, 593–607.

Jebb, S.A., Ahern, A.L., Olson, A.D., Aston, L.M., Holzapfel, C., Stoll, J., Amann-Gassner, U., Simpson, A.E., Fuller, N.R., Pearson, S., Lau, N.S., Mander, A.P., Hauner, H., Caterson, I.D., 2011. Primary care referral to a commercial provider for weight loss treatment versus standard care: a randomised controlled trial. Lancet 378, 1485–1492.

Johnson, A.W., 2013. Eating beyond metabolic need: how environmental cues influence feeding behavior. Trends in Neurosciences 36, 101–109.

Johnson, F., Pratt, M., Wardle, J., 2012. Dietary restraint and self-regulation in eating behavior. International Journal of Obesity (London) 36, 665–674.

Lally, P., Chipperfield, A., Wardle, J., 2008. Healthy habits: efficacy of simple advice on weight control based on a habit-formation model. International Journal of Obesity (London) 32, 700–707.

Llewellyn, C.H., Trzaskowski, M., Van Jaarsveld, C.H., Plomin, R., Wardle, J., 17 February 2014. Satiety mechanisms in genetic risk of obesity. Journal of the American Medical Association Pediatrics 168, 338–344.

Look Ahead Research Group, 2014. Eight-year weight losses with an intensive lifestyle intervention: the look AHEAD study. Obesity (Silver Spring) 22, 5–13.

Pagoto, S.L., Appelhans, B.M., 2013. A call for an end to the diet debates. Journal of the American Medical Association 310, 687–688.

Papies, E.K., Stroebe, W., Aarts, H., 2008. Healthy cognition: processes of self-regulatory success in restrained eating. Personality and Social Psychology Bulletin 34, 1290–1300.

Parry, L.L., Netuveli, G., Parry, J., Saxena, S., 2008. A systematic review of parental perception of overweight status in children. Journal of Ambulatory Care Management 31, 253–268.

Patterson, C., Hilton, S., 2013. Normalisation and stigmatisation of obesity in UK newspapers: a visual content analysis. The Open Obesity Journal 5, 82–91.

Prentice, A.M., Hennig, B.J., Fulford, A.J., 2008. Evolutionary origins of the obesity epidemic: natural selection of thrifty genes or genetic drift following predation release? International Journal of Obesity (London) 32, 1607–1610.

Ramage, S., Farmer, A., Apps Eccles, K., Mccargar, L., 2014. Healthy strategies for successful weight loss and weight maintenance: a systematic review. Applied Physiology, Nutrition, and Metabolism 39, 1–20.

Rosenbaum, M., Hirsch, J., Gallagher, D.A., Leibel, R.L., 2008. Long-term persistence of adaptive thermogenesis in subjects who have maintained a reduced body weight. American Journal of Clinical Nutrition 88, 906–912.

Schachter, S., 1968. Obesity and eating. Internal and external cues differentially affect the eating behavior of obese and normal subjects. Science 161, 751–756.

Shrewsbury, V., Wardle, J., 2008. Socioeconomic status and adiposity in childhood: a systematic review of cross-sectional studies 1990–2005. Obesity (Silver Spring) 16, 275–284.

Stunkard, A.J., Harris, J.R., Pedersen, N.L., Mcclearn, G.E., 1990. The body-mass index of twins who have been reared apart. New England Journal of Medicine 322, 1483–1487.

Sumithran, P., Proietto, J., 2013. The defence of body weight: a physiological basis for weight regain after weight loss. Clinical Science (London) 124, 231–241.

Van Dorsten, B., Lindley, E.M., 2011. Cognitive and behavioral approaches in the treatment of obesity. Medical Clinics of North America 95, 971–988.

Vartanian, L.R., Herman, C.P., Wansink, B., 2008. Are we aware of the external factors that influence our food intake? Health Psychology 27, 533–538.

Wardle, J., Boniface, D., 2008. Changes in the distributions of body mass index and waist circumference in English adults, 1993/1994 to 2002/2003. International Journal of Obesity (London) 32, 527–532.

Wardle, J., Carnell, S., Haworth, C.M., Plomin, R., 2008. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. American Journal of Clinical Nutrition 87, 398–404.

WHO, 2000. Obesity: preventing and managing the global epidemic. World Health Organization Technical Report Series No. 894. Geneva: World Health Organization.

Wing, R.R., Phelan, S., 2005. Long-term weight loss maintenance. American Journal of Clinical Nutrition 82, 222S–225S.

Welfare

Lorraine C Minnite and Joan M Mazelis, Rutgers, The State University of New Jersey, Camden, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the social and behavioral sciences, 'welfare' is widely used in at least two distinctive ways: (1) as the complex of state practices and administrative developments known as the welfare state and (2) as the main income support programs for the poor in the United States that serve as the modern successors to the centuries-long tradition of poor relief. Neither is it possible to summarize the extensive scholarly treatments of these subjects in the space allotted here nor is it a simple task to trace the separate meanings of welfare or centuries of experience and ideas about poor relief and disentangle them from the complexities of welfare state development. Instead, we highlight how the politics of welfare policy have shaped how we think about welfare as a societal concern and trace the main contours of the scholarly debates influencing the wide variety of programmatic interventions that constitute welfare state policy in the industrialized West. We begin with a note on origins, when local concerns were paramount; examine key ideas shaping liberal, social democratic, and conservative responses to the emergence of welfare states in market societies; and conclude with a brief overview of some of the main challenges for theory and policy in the contemporary period as economic globalization and supranational organizations come to play much more of a role in shaping policy.

Origins

Welfare states are modern in origin. Historians point to heterogeneous antecedents dating as far back as the Roman Empire or the first poor law of Charlemagne in 779. In Europe, population growth and the transition from feudalism to capitalism in the fifteenth and sixteenth centuries led to alarming increases in vagrancy and beggary. Religious and political elites responded with the development of the first relief systems to deal with the increasing threats to civil order posed by the vagrant poor (Piven and Cloward, 1993 [1971]). Local leaders in England and France at this time sanctioned private charity in the form of almsgiving as a means of personal salvation. The rise of market economies disrupted centuries-old traditional relationships between lords and vassals and the uses of the land upon which the peasantry depended for subsistence (Polanyi, 1968 [1944]). The growth of manufacturing and the rise of commerce dislocated the manorial organization and created an urban proletariat increasingly mobile and vulnerable to cycles in trade and foreign competition (Webb and Webb, 1963 [1927]). Local charity was no longer effective in quelling disturbances of the poor who rose up in reaction to a bad harvest or the seizure of commonly held lands for private pasture. Monarchs and governments responded, as, for example, when King Francis I of France ordered parish leaders to register the poor and to provide for the 'impotent' among them. Similarly, in England, a series of legislative acts of Parliament in the 1530s during the reign of Henry VIII required local officials to register the poor and regulate their begging by issuing personal certificates authorizing them to do so, and to require local parishes to take care of their poor by establishing procedures for the collection and distribution of funds, thus taking control of almsgiving away from religious authorities and establishing secular authority over charity (Piven and Cloward, 1993 [1971]). According to de Schweinitz, the Act Concerning Punishment of Beggars and Vagabonds of 1531 "represented the beginning of definite assumption by government of responsibility for the care of persons in economic distress" (de Schweinitz, 1961 [1943]: p. 21).

In England and Wales, and later, Scotland, public authorities played an early and unusually large role in the production of social welfare. This was accomplished through an increasingly complex system of rules and regulations governing poor relief beginning with the Elizabethan Poor Law of 1601 and continuing in the establishment of a legal regime that dominated for the next nearly 350 years (de Schweinitz, 1961 [1943]). The major features of the 'old poor law' system erected between the sixteenth and eighteenth centuries were a far-flung local system of parish administration of the laws, compulsory poor rates levied on every parish, residency restrictions aimed at stemming the flow of vagrants by requiring the poor to settle in some particular parish as a condition of receiving aid (de Swaan, 1988), differential means of assistance based on distinctions between the deserving ('impotent') and undeserving or able-bodied poor (the 'idle' poor received nothing and were to be whipped through the public streets until they reformed their ways), and a system of two types of relief: (1) 'outdoor' relief, which was the norm and through which the poor were left to reside in their own homes and given a 'dole' of relief in the form of cash, clothing, and food and (2) 'indoor' relief, by which the poor were confined to local alms houses, or later, workhouses; the ill were admitted to hospitals; and the orphans were taken into orphanages. Legislation to significantly reform the old poor law regime was passed in 1834, following recommendations of a Royal Commission. The parish system was largely abolished and a centralized authority created to prevent the wide variation in practices that was a hallmark of the old poor law system; outdoor relief was banned and the building of workhouses for different types of paupers expanded.

It is only with the birth of the modern state in the intellectual ferment of the Enlightenment and the upheavals of the French Revolution that we get the first statements of a governmental obligation to provide for the general welfare

of 'the people.' For example, Thomas Paine's *Rights of Man* (1791–92) declared that, "Civil government does not consist in executions; but in making that provision for the instruction of youth, and the support of age, as to exclude, as much as possible, profligacy from the one, and despair from the other" (Paine, 1999 [1791]: p. 151). Paine proposed a scheme for progressive taxation to support an elaborate plan for state aid to the poor that included a pension system, public works projects for the chronically unemployed urban poor, and supplemental income to support marriage and birth. The French Constitution of 1793, influenced by Paine's ideas, guaranteed a 'right to subsistence' and income redistribution in the form of public relief and public schooling. The modern history of welfare as a concept, policy, and program is one of recurring conflicts between the idea of property rights at the heart of the capitalist mode of production, and assertions of rights to subsistence and human and communal well-being that have shaped the democratic project.

Later, in the waning decades of the nineteenth century, the first framework for anything like a "systematic social concern for the welfare of people who might otherwise lack the basic necessities required for effective functioning within their community" (Goodin and Mitchell, 2000: p. ix) emerged in Germany as state activity organized through a limited program for social insurance. Most scholars date the origins of the modern welfare state to Bismarck's reforms of the 1880s that introduced compulsory insurance against sickness, accidents, and old age (Briggs, 1961). These innovations were followed by similar developments in Sweden and England, and eventually, in the United States with the passage of the Social Security Act of 1935 (Heclo, 1974).

Intellectual Background and Current Knowledge

Heclo notes that the 'welfare objective' of the modern state can take a number of forms, "… the most prominent being investment (e.g., education), consumption (public assistance payments), and compensation (industrial injuries)" (Heclo, 1974: p. 2). Similarly, Briggs argues that the welfare state is one in which organized power is used to insulate individuals and families from the play of market forces through guarantees of minimum income and insurance against the inability to participate in a labor market due to sickness, disability, and old age. The idea of insurance and the use of insurance principles in modern social policy helped to remove the stigma of pauperism and provided a bridge between voluntary methods of self-help and mutual aid and compulsory approaches to social provision that entailed and accompanied the massive growth of governmental intervention in the economy during the middle period of the twentieth century. Significant to the post–World War II development of the welfare state is the expansion of state income support programs from older legacies of relief to income support by right.

The social democratic analysts of the interwar and postwar welfare states provided an influential interpretation and rationale in defense of governmental protection against market failure and public provision of minimum standards of social well-being. Among them, T.H. Marshall's work stands out in shaping ideas about social citizenship, informed by egalitarian ideas about what he called 'social rights.' In a 1950 essay titled 'Citizenship and Social Class,' Marshall argued that rights had evolved from civil rights won in the eighteenth century, such as those necessary for individual freedom (i.e., freedom of speech, the right to own property, and the right to justice), to political rights, including the right to vote won in the nineteenth century, to social rights, a new concept which Marshall defined as "… the right to a modicum of economic welfare and security … the right to share to the full in the social heritage and live the life of a civilized being according to the standards prevailing in the society. The institutions most closely connected with it are the educational system and the social services …" (Marshall, 1950: p. 11). Social rights promised new claims to a standard of life and to 'full membership of a community,' by which Marshall meant equal citizenship (Marshall, 1950: p. 28). Social rights are contiguous with civil and political rights in that they are necessary for the full realization of those rights, but also 'at war' with market principles legitimated by such rights. In other words, they justify governmental regulation of markets as they legitimize welfare policies to expand services in order to ensure that all citizens have the capacities required for full participation and membership in the community. Marshall believed that as the welfare state expanded, social rights ultimately would defeat the divisions of social class.

Marshall's concept of a war between citizenship and social class appears to have been derived from Karl Polanyi's ideas about fundamental contradictions in laissez–faire capitalism and its drive to commodify labor power (Holmwood, 2000). In *The Great Transformation*, Polanyi argued that the premise of a universal psychology of utility maximization and self-interest undergirding the idea of a 'self-regulating' market was a myth, and its invention in both theory and practice a unique, historically specific development that emerged in England in the seventeenth and eighteenth centuries. For Polanyi, a market economy "is an economic system, controlled, regulated and directed by markets alone …" and, "self-regulation implies that all production is for sale on the market and that all incomes derive from such sales. Accordingly, there are markets for all elements of industry, not only goods (always including services) but also for labor, land and money, their prices being called respectively commodity prices, wages, rent and interest" (Polanyi, 1968 [1944]: pp. 68–69). Land, labor, and money were 'fictitious commodities' because they could be subjected to market processes only by denying their social substance outside the constraints of the commodity form. Thus, the problem with the self-regulating market is not a problem with economies or markets *writ large*, as all human societies have economies and markets developed in both premodern and modern economies, which did not limit human survival to market relations. Polanyi's critique of liberal political economy is that the idea that everything can be commodified, including human labor, was antisociety and a dangerous fiction, and by embedding social relations in the economy rather than the other way around, as occurs in other human societies, market economies sow the seeds of their own destruction.

Thus, "the variability of welfare state evolution," writes Esping-Andersen, "reflects competing responses to pressures for de-commodification" (Esping-Andersen, 1990: p. 37). In *The Three Worlds of Welfare Capitalism* (1990), Esping-Andersen advanced the study of welfare states by first rejecting the call

of classical political economists for universal laws in understanding the social and political world. This had led to the misspecification of welfare states based on a misplaced presumption that as states experienced industrialization, urbanization, and modernization they would converge toward welfare state status. For classical theorists, the measure of a welfare state was its level of expenditure on services addressing human needs outside the cash nexus of a market economy. Instead, Esping-Andersen argues that "the welfare state cannot be understood just in terms of the rights it grants …" (Esping-Andersen, 1990: p. 21), and that "expenditures are epiphenomenal to the theoretical substance of welfare states" (Esping-Andersen, 1990: p. 19).

His influential approach offers a more flexible comparative methodology, sorting welfare states into different regime-type clusters based on social democratic, corporatist, and liberal or residual forms of state intervention. The different types of welfare state regime vary across a set of conditions that include (1) the degree to which a welfare state has reached decommodification (which "… occurs when a service is rendered as a matter of right, and when a person can maintain a livelihood without reliance on the market" (Esping-Andersen, 1990: p. 22)); (2) the impact of welfare state programs on social stratification, for "the welfare state is not just a mechanism that intervenes in, and possibly corrects, the structure of inequality; it is, in its own right, a system of stratification … [and] an active force in the ordering of social relations" (Esping-Andersen, 1990: p. 23); and (3) the particular blend of private and public welfare provision reflecting peculiar state–market relations. Here, Esping-Andersen argues, "The division of social protection between public and private provides the structural context of de-commodification, social rights, and the stratificational nexus of welfare-state regimes" (Esping-Andersen, 1990: p. 80).

Welfare states are classified beyond mere effort or generosity in redistribution to take account of their impact on social structure. Corporatist welfare states such as Germany and France have as a goal the preservation of a social order, with paternalistic programs and a larger role for religion in society. Liberal welfare states emerge from classical liberalism's war against traditional social hierarchies and celebration of the free market. Thus, liberal social welfare states such as those in the Anglo-American orbit encourage more interaction with the labor market for purposes of income maintenance and insurance against risk. Finally, social democratic welfare states, built by the efforts of strong labor movements to construct the basis for social solidarity, favor universalism in program design to link policies to a broader goal of social citizenship. Sweden and the Nordic welfare states represent this model.

Myrdal (1944) and Titmuss (1974) had expressed similar ideas about stages of welfare state development. For example, Myrdal writes, "Social reform policies may be conceived of as passing through three stages: a paternalistic conservative era, when curing the worst ills is enough; a liberal era, when safeguarding against inequalities through polling the risks is enough; and a social democratic era, when preventing the ills is attempted" (Myrdal, 1944: p. 152). But welfare capitalism and welfare regime ideas helped to broaden the concept of the welfare state beyond the traditional focus on the transfer sector of the capitalist economy to a larger set of concerns about a wider range of values, socioeconomic institutions, policies, and programs all oriented toward promoting human welfare. Esping-Andersen's work sparked a debate about the variety of national experiences and the preeminence of labor market formations and labor power decommodification as defining of the most significant differences among welfare states. Some claim that he misspecifies the Mediterranean welfare states, others reassess his methodology, and feminist critiques address his neglect of unpaid work (e.g., Arts and Gelissen, 2002; Lewis, 1992; Orloff, 1993; Scruggs and Allan, 2008).

Other lines of research and theorizing about welfare expanded in the 1980s and the 1990s. Of growing importance were analyses that centered on the politics of the welfare state. Piven and Cloward's groundbreaking work on the functions of public welfare (*Regulating the Poor*, 1993 [1971]) presented a theory that reconceptualized the entire history of relief giving in terms of its uses by elites in the larger economic and political order. Rather than a progressive unfolding of a modernizing state apparatus concerned with conferring social rights or promoting human welfare in the face of market instabilities, the history of relief giving for Piven and Cloward was one of expansion and contraction in response to labor conditions among the lower classes. They drew from the rich historical record of peasant uprisings and food riots, factory revolts and occupations, violent opposition to land seizures, vagrancy, and patterns of crime to show how popular resistance to the dislocations of capitalism in effect created the welfare state and shaped the specific forms it took, especially in the United States.

Their controversial thesis stimulated an enormous body of work dealing with questions about the origins of the welfare state, the political process through which policies to deal with poverty are initiated and sustained, the role of parties and interest groups, unions and reform organizations and the institutions that shape them, and the relationship of class mobilization to labor decommodification, among many other topics. Their prescient insights into the importance of conflict to welfare state development and their explicitly political analysis of relief giving have come into even sharper focus with the rise of neoliberalism, the New Right, and policies of welfare state retrenchment in the current period. Before we turn to those important developments, a discussion of the US case is in order.

The US Context

Esping-Andersen characterizes the United States as a 'liberal' welfare state, where "… the state encourages the market, either passively – by guaranteeing only a minimum – or actively – by subsidizing private welfare schemes" (Esping-Andersen, 1990: pp. 26–27). As such, people in the United States typically think of 'welfare' as means-tested government-provided cash assistance, only available to those with incomes below a specified level, although the term also includes social insurance programs such as social security, disability payments, and unemployment compensation. The latter are not means-tested programs and do not carry the same kind or degree of stigma as they represent earned benefits rather than 'relief' (Gordon, 1994; Jansson, 2001; Katz, 1986).

What people usually think of as 'welfare' was born as a federal program as part of the New Deal, but that was not the beginning of assistance provision for poor women and children (Katz, 1986; Teles, 1996). In 1935, the Social Security Act, which contained Aid to Dependent Children (ADC), supplanted what had thus far been state-based mothers' aid programs, usually called widows' pensions or mothers' pensions, perhaps the real beginning of modern welfare (Jansson, 2001; Katz, 1989; Teles, 1996). Those mothers' pension programs originated earlier in the twentieth century (and their roots extend much farther back; see Handler and Hasenfeld, 1997; Katz, 1986) and were designed to provide for widows; most single mothers at the time were widows rather than those who had children outside of marriage, were divorced, or were abandoned by their husbands. Most states developed mothers' pension programs between 1911 and 1919, and nearly all states had them by 1934, but relatively few women received actual benefits; the programs were generally miserly and restrictive. ADC was a federal version of mothers' pension programs, although eligibility standards, grant levels, and other specifics of implementation were left to individual states. ADC provided funds to families with a single parent, and funds only for the children and not for the parent. In 1939, amendments to the Social Security Act extended Old Age Insurance to the widows and children of deceased workers, so a group of those eligible for ADC were now covered by an insurance program and the 'relief' program came to be more for the 'unworthy.' The change meant higher benefits and better treatment for widows and lower benefits and punitive treatment for unwed mothers. Insurance was a benefit program, whereas welfare was a relief program. As mothers' aid or widows' pensions became ADC, some sympathy might have been extended to all ADC recipients, but once ADC recipients no longer included widows, public sympathy for them disappeared (see Katz, 1986; Teles, 1996, for comprehensive treatments).

For the next three decades, ADC remained a program restricted, for all practical purposes, to white women (Gordon, 1994; Neubeck and Cazenave, 2001). The program was renamed Aid to Families with Dependent Children (AFDC) in 1962, and the policy limiting funds to children was amended (Handler and Hasenfeld, 1997; Jansson, 2001; Teles, 1996). The National Welfare Rights Movement, and several court cases and amendments from 1939 through the 1970s, succeeded in expanding the program to all those who qualified financially, regardless of race, and the welfare rolls grew (Gordon, 1994; Handler and Hasenfeld, 1997).

Welfare receipt carries a strong stigma in the United States (Abramovitz, 2000; Edelman, 2012; Ellwood, 1988; Gans, 1995; Gilens, 1999; Gordon, 1994; Hays, 2003; Katz, 1983; Piven, 1996; Piven and Cloward, 1993 [1971]; Rogers-Dillon and Skrentny, 1999; Schram, 1995; Teles, 1996). People tend to blame the persistence of poverty on poor individuals (Handler and Hasenfeld, 1997; Katz, 1983). The emphasis on individual responsibility and material success as a barometer of moral worth necessitates an antipathy toward welfare and those who receive it. Popular American images of people on welfare depict recipients' problems as their own fault (Katz, 1983). Politicians have stirred public outrage about welfare and welfare recipients "with images of steaks in welfare freezers and Cadillacs in the parking lots of public housing" (Katz, 1986: p. 278). The 'welfare queen' image, popularized by Ronald Reagan, is based on myths about women defrauding and abusing the welfare system to become wealthy (Zucchino, 1997).

Much of this welfare stigma is shaped by racial stereotypes (Gans, 1995; Gilens, 1999; Quadagno, 1994; Zucchino, 1997). The white American public generally thinks that most recipients are black (although this is untrue; see, e.g., Hays, 2003), and that blacks are less committed to the work ethic than are members of other groups. Racial stereotypes of blacks as lazy and biased media coverage play a role in generating opposition to welfare (Gilens, 1999). The feminization of poverty and American images and ideas about poor women also depress public support for welfare (Abramovitz, 2000). Poor women, especially single mothers, confront conflicting pressures; they are expected to be providers, working in the paid labor force, and to be the sole caretakers and nurturers of their children.

Six decades after the creation of ADC, The Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) of 1996 replaced the federal entitlement program for low-income families, AFDC, with a state-administered block grant program, Temporary Assistance for Needy Families (TANF) (Moffitt and Ver Ploeg, 1999). PRWORA initiated cuts in assistance, restrictions on eligibility, time limits on receipt of assistance, and work requirements (Handler and Hasenfeld, 1997; Piven, 1996; Quint et al., 1999). PRWORA fundamentally changed welfare from an entitlement available to all who qualify financially to a punishingly restrictive program. Until that time, the typical welfare recipient was already only on welfare for less than 2 years, due in part to the strong pull factor of work in generating positive self-esteem and regard from others (Bane and Ellwood, 1994; Edin and Lein, 1997). At the implementation of welfare reform, a time of economic boom and lower unemployment, the welfare rolls were already falling. And in the 5 years after PRWORA, the number of welfare recipients declined by more than half (Hays, 2003).

By a decade and a half after PRWORA, the percentage of the populace receiving welfare, in the form of cash assistance, remained very low, although the percentage receiving food stamps (the Supplemental Nutrition Assistance Program) ballooned during the 2007–09 recession and beyond, in a time of high poverty and unemployment. In 2010, the number on TANF was about 4.5 million, while the number receiving food stamps was 10 times that – close to 46 million in 2011 (Edelman, 2012; United States Department of Agriculture, 2013; United States Department of Health & Human Services, 2013). There remains a right to receive food stamps, while the right to receive welfare has been virtually eliminated. The low rate of welfare receipt does not reflect lower rates of need, but rather a more limited public safety net of 'welfare' (Edelman, 2012; Hays, 2003).

Conservative Critiques and Welfare Retrenchment

There are many ironies in the contemporary neoliberal assault on welfare state programs, not the least of which is conservatism's characteristic defense of the past. A hallmark of older

European conservative ideologies is the notion that the commodification of human beings is morally degrading and that individuals are not meant to compete with one another but rather subordinate their self-interest to traditional sources of authority (Esping-Andersen, 1990: p. 40). Elites generally feared the march of liberalism, capitalism, and democracy as threatening to their traditional sources of authority and prestige. Feudal notions of the obeisance owed the lord included norms of reciprocity and noblesse oblige owed the peasant. A market-set wage was not the final word on the value of a laborer's contribution when his house burned down or he suffered the loss of a child. In our own time, patronage and clientelism, guilds and fraternal associations, mutual aid societies, and all manner of private charitable giving meant to inspire loyalty, social integration, and the preservation of authority also reflect a disdain toward the commodification and atomization of the individual.

Attacks on public provision of welfare are not new; they date back to the emergence of capitalist forms of production and the earliest relief efforts of the sixteenth century. And at every stage of development, with each new legislative initiative to deal with vagrant paupers, the ill, the infirmed, the aged, widows and children left destitute by the loss of a breadwinner, workers thrown out of work by mechanization, market downturns, and disinvestment, critics of government intervention to buffer or ameliorate these conditions resurrected older arguments about how relief programs were not good for society and worse for the poor themselves.

In the contemporary period, however, the substance of conservative critiques of welfare has little in common with older traditions. Today's conservatives generally favor markets over the state as superior mechanisms for the distribution of all goods and services produced in a society (Friedman, 1962; Hayek, 1960, 1976). Instead of a concern for tradition and the reciprocity of cross-class obligations in society, they claim the values of individual 'freedom' and 'choice' that markets promise to deliver upon. The main arguments are that (1) welfare spending is a drag on productivity and efficiency in the economy; (2) government supports for a social wage deter labor market participation because they provide tantalizing alternatives to work; and (3) both of the main forms of welfare state provision, social assistance and social insurance, have perverse effects on the broader economy, causing the very problems of 'dependency,' poverty, and a lack of class mobility that they are supposed to cure (Anderson, 1978; Gallaway and Vedder, 1986; Gilder, 1981; Mead, 1986; Murray, 1984; Novak et al., 1987).

These ideas were given powerful support with the rise of the New Right in the early 1980s, and especially, in the coming to national power of conservative probusiness, 'free market' regimes in Britain and the United States, where welfare 'reform' has resulted in the dismantling of the main means-tested income maintenance program for the poor, as noted above. Scholars debate the degree to which governments in the advanced industrial and postindustrial countries have retrenched in their provision of the broad mix of welfare state programs developed since 1945, such as old age pensions, unemployment insurance, income maintenance, child and family allowances, and the like. Paul Pierson's influential work on the 'new politics of the welfare state' (Pierson, 1994, 1996, 2001) argues that despite large-scale budgetary pressures, retrenchment has been limited by the strength of public opinion and support for core programs, and because the dynamics of policy 'feedback' or legacies create institutional hurdles to radical change.

It remains to be seen whether these arguments will withstand the effects of the Great Recession, and in particular, the imposition of austerity regimes in the Eurozone. Scholars will continue to debate the extent to which globalization, national integration and regionalization, deindustrialization, and post-industrialization have caused or will cause welfare state retrenchment in the advanced economies (e.g., Castles, 2004; Clayton and Pontusson, 1998; Huber and Stephens, 2001; Lindert, 2004; Scharpf, 1996; Scruggs and Allan, 2008; Swank, 2002, 2005). Meanwhile, developments in the emerging economies of the global South also demand attention. In many developing countries, both centrist and left-leaning governments have answered decades-long popular mobilizations among the lower classes with expansive social policy initiatives known as Conditional Cash Transfers (or CCTs), which tend to take different forms than the main programs of the Western welfare state and older types of poor relief. The programs are flexible in design, but all transfer resources to poor households, usually to mothers, on the condition that parents take active measures to build up the human capital of their children, usually by ensuring that their children are enrolled in and are attending school, and making regular visits to health care providers. Already, CCTs are having positive effects on poverty alleviation (Fiszbein and Schady, 2009). The erosion of the Western welfare state and the emergence of new forms of social protection in the global South present important questions for theory, research, policy, and practice in the forms of social provision we call welfare.

See also: Poverty, Sociology of; Underclass; Welfare Economics; Welfare State, History of; Welfare: Philosophical Aspects.

Bibliography

Abramovitz, M., 2000. Under Attack, Fighting Back: Women and Welfare in the United States. Monthly Review Press, New York.

Anderson, M., 1978. Welfare: The Political Economy of Welfare Reform in the United States. Hoover Institution Press, Stanford.

Arts, W., Gelissen, J., 2002. Three worlds of welfare capitalism or more? A state-of-the-art report. Journal of European Social Policy 12 (2), 137–158.

Bane, M.J., Ellwood, D.T., 1994. Welfare Realities: From Rhetoric to Reform. Harvard University Press, Cambridge.

Briggs, A., 1961. The welfare state in historical perspective. European Journal of Sociology 2 (2), 221–258.

Castles, F.G., 2004. The Future of the Welfare State: Crisis Myths and Crisis Realities. Oxford University Press, New York.

Clayton, R., Pontusson, J., 1998. Welfare state restructuring revisited: entitlement cuts, public sector restructuring, and inegalitarian trends in advanced capitalist democracies. World Politics 51 (1), 67–98.

de Schweinitz, K., 1961 [1943]. England's Road to Social Security: From the Statue of Laborers in 1349 to the Beveridge Report of 1942. A.S. Barnes & Company, Inc., New York.

de Swaan, A., 1988. In Care of the State: Health Care, Education and Welfare in Europe and the USA in the Modern Era. Oxford University Press, New York.

Edelman, P., 2012. So Rich, So Poor: Why It's So Hard to End Poverty in America. The New Press, New York.
Edin, K., Lein, L., 1997. Making Ends Meet: How Single Mothers Survive Welfare and Low-wage Work. Russell Sage, New York.
Ellwood, D.T., 1988. Poor Support: Poverty in the American Family. Basic Books, New York.
Esping-Anderson, G., 1990. The Three Worlds of Welfare Capitalism. Princeton University Press, Princeton.
Fiszbein, A., Schady, N.R., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank Publications, Washington, DC.
Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.
Gallaway, L., Vedder, R., 1986. Poverty, Income Distribution, the Family and Public Policy: A Study Prepared for the Use of the Subcommittee on Trade, Productivity, and Economic Growth of the Joint Economic Committee of the Congress of the United States. U.S. Government Printing Office, Washington, DC.
Gans, H.J., 1995. The War against the Poor: The Underclass and Antipoverty Policy. Basic Books, New York.
Gilder, G., 1981. Wealth and Poverty. Basic Books, New York.
Gilens, M., 1999. Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy. University of Chicago Press, Chicago.
Goodin, R.E., Mitchell, D. (Eds.), 2000. The Foundations of the Welfare State, vol. I. Edward Elgar Publishing, Inc., Northampton.
Gordon, L., 1994. Pitied but Not Entitled: Single Mothers and the History of Welfare, 1890–1935. Harvard University Press, Cambridge.
Handler, J.F., Hasenfeld, Y., 1997. We the Poor People: Work, Poverty, and Welfare. Yale University Press, New Haven.
Hayek, F.A., 1960. The Constitution of Liberty. Routledge & Kegan Paul, London.
Hayek, F.A., 1976. The Mirage of Social Justice. University of Chicago Press, Chicago.
Hays, S., 2003. Flat Broke with Children: Women in the Age of Welfare Reform. Oxford University Press, New York.
Heclo, H., 1974. Modern Social Politics in Britain and Sweden: From Relief to Income Maintenance. Yale University Press, New Haven.
Holmwood, J., 2000. Three pillars of welfare state theory: T.H. Marshall, Karl Polanyi and Alva Myrdal in defence of the national welfare state. European Journal of Social Theory 3 (1), 23–50.
Huber, E., Stephens, J.D., 2001. Partisan Choice in Global Markets: Development and Crisis of Advanced Welfare States. University of Chicago Press, Chicago.
Jansson, B.S., 2001. The Reluctant Welfare State: American Social Welfare Policies – Past, Present, and Future, fourth ed. Wadsworth, Belmont.
Katz, M.B., 1983. Poverty and Policy in American History. Academic Press, New York.
Katz, M.B., 1986. In the Shadow of the Poorhouse: A Social History of Welfare in America. Basic Books, New York.
Katz, M.B., 1989. The Undeserving Poor: From the War on Poverty to the War on Welfare. Pantheon Books, New York.
Lewis, J., 1992. Gender and the development of welfare regimes. Journal of European Social Policy 2 (3), 159–173.
Lindert, P., 2004. Growing Public: Social Spending and Economic Growth since the Eighteenth Century. Cambridge University Press, New York.
Marshall, T.H., 1950. Citizenship and social class. In: Marshall, T.H. (Ed.), Citizenship and Social Class and Other Essays. Cambridge University Press, Cambridge.
Mead, L., 1986. Beyond Entitlement: The Social Obligations of Citizenship. The Free Press, New York.
Moffitt, R.A., Ver Ploeg, M. (Eds.), 1999. Interim Report: Evaluating Welfare Reform: A Framework and Review of Current Work. National Academy Press, Washington, DC.
Murray, C., 1984. Losing Ground: American Social Policy, 1950–1980. Basic Books, New York.
Myrdal, G., 1944. An American Dilemma: The Negro Problem and Modern Democracy. Harper & Bros., New York.
Neubeck, K.J., Cazenave, N.A., 2001. Welfare Racism: Playing the Race Card against America's Poor. Routledge, New York.
Novak, M., Mead, L., Cogan, D., Moran, D., Bernstein, B., Murray, C., Besharov, D.J., Nathan, R.P., Blum, B., Neuhaus, R.J., Carlson, A., Raines, F.D., Horowitz, M., Reischauer, R.D., Kondratas, A.S., Rivlin, A.M., Lenkowsky, L., Ross, S., Loury, G.C., Stern, M., 1987. The New Consensus on Family and Welfare: A Community of Self-Reliance. American Enterprise Institute for Public Policy Research, Washington, DC.
Orloff, A.S., 1993. Gender and the social rights of citizenship: the comparative analysis of gender relations and welfare states. American Sociological Review 58, 303–328.
Paine, T., 1999 [1791]. The Rights of Man. Dover Publications, Mineola, NY.
Pierson, P., 1994. Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment. Cambridge University Press, New York.
Pierson, P., 1996. The new politics of the welfare state. World Politics 48, 143–179.
Pierson, P. (Ed.), 2001. The New Politics of the Welfare State. Oxford University Press, New York.
Piven, F.F., 1996. Welfare and the transformation of electoral politics. Dissent 43, 61–67.
Piven, F.F., Cloward, R.A., 1993 [1971]. Regulating the Poor: The Functions of Public Welfare. Vintage Books, New York.
Polanyi, K., 1968 [1944]. The Great Transformation: The Political and Economic Origins of Our Time, second ed. Beacon Press, Boston.
Quadagno, J.S., 1994. The Color of Welfare: How Racism Undermined the War on Poverty. Oxford University Press, New York.
Quint, J., Edin, K., Buck, M.L., Fink, B., Padilla, Y.C., Simmons-Hewitt, O., Valmont, M.E., 1999. Big Cities and Welfare Reform: Early Implementation and Ethnographic Findings from the Project on Devolution and Urban Change. Manpower Demonstration Research Corporation, New York.
Rogers-Dillon, R.H., Skrentny, J.D., 1999. Administering success: the legitimacy imperative and the implementation of welfare reform. Social Problems 46, 13–29.
Scharpf, F., 1996. Negative integration and the loss of boundary control. In: Marks, G., Scharpf, F.W., Schmitter, P.C., Streeck, W. (Eds.), Governance in the European Union. Sage Publications, Thousand Oaks, pp. 15–40.
Schram, S.F., 1995. Words of Welfare: The Poverty of Social Science and the Social Science of Poverty. University of Minnesota Press, Minneapolis.
Scruggs, L.A., Allan, J.P., 2008. Social stratification and welfare regimes for the twenty-first century: revisiting the Three Worlds of Welfare Capitalism. World Politics 60, 642–664.
Swank, D., 2002. Global Capital, Political Institutions, and Policy Change in Developed Welfare States. Cambridge University Press, New York.
Swank, D., 2005. Globalisation, domestic politics, and welfare state retrenchment in capitalist democracies. Social Policy and Society 4 (2), 183–195.
Teles, S.M., 1996. Whose Welfare? AFDC and Elite Politics. University Press of Kansas, Lawrence.
Titmuss, R.M., 1974. What is social policy? In: Abel-Smith, B., Titmuss, K. (Eds.), Social Policy: An Introduction. Pantheon Books, New York, pp. 23–32.
United States Department of Agriculture, Food and Nutrition Service, 2013. Program Data. http://www.fns.usda.gov/pd/snapmain.htm (retrieved 25.05.13.).
United States Department of Health & Human Services, Office of Family Assistance: An Office of the Administration for Children & Families, 2013. Data & Reports: Caseload Data. http://www.acf.hhs.gov/programs/ofa/programs/tanf/data-reports (retrieved 25.05.13.).
Webb, S., Webb, B., 1963 [1927]. English Local Government: English Poor Law History, Part I. Archon Books, Hamden, Conn.
Zucchino, D., 1997. Myth of the Welfare Queen. Touchstone, New York.

Welfare and Education

Klaus Hurrelmann, Katharina Rathmann, and Julian Kickbusch, Hertie School of Governance, Berlin, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first section examines the entrenchment of educational policy in the tradition of three social welfare policies. These three basic concepts that help guarantee social welfare services each attach a specific value to educational policy.

The second section presents the international results of educational policy in three 'welfare regimes.' The particular political structures contained within are confronted with 'educational outcomes,' i.e., educational quality and equity.

The third section transfers this comparative approach to countries with emerging welfare policies in Latin America. The particular focus of educational policy and integration into the country's social policy determines educational quality and the level of equity.

The Relationship between Social and Educational Policy in Various Welfare Regimes

The educational level of each member of the society is not only relevant on an individual level, but also of great social importance, especially in modern complex industrialized societies. The level of education achieved is primarily significant for the position an individual will have on the labor market, while also having crucial secondary effects upon other important societal issues such as health status throughout the life course and general social capital (Levy and Murnane, 2003). Increased knowledge and enhanced skills ensure a higher chance of attaining an influential, well-paid, and secure position in the workforce, which stands in stark contrast to cases where only lower qualifications have been attained. Education is also one of the most important drivers behind the economic growth of a nation by enhancing general productivity; in addition, the consequences of limited education and insufficient qualifications after leaving school severely limits access to the job market, and in worst-case scenarios may result in downright exclusion. In this respect, the education of all individuals is one of the essential factors in the creation and maintenance of the social status and thus also of the quality of life of all members within the society. In summary, education is fundamental for development, for achieving a cohesive society, for reducing social inequalities, for combating social exclusion, and for combating poverty. Therefore, collectively, education also determines the economic and social prosperity of an entire country (Wößmann and Schütz, 2006).

Policy Traditions in Welfare States: The Role of Education

Consequently, it is all the more surprising that in the light of these correlations only little importance has been attached to educational policy within the political traditions of internationally comparative classifications. The differentiation of various clusters of social welfare policy is based first and foremost on a country's social policy, which in turn is primarily understood as being part of social security policy. A prominent example of this is the typology developed by Esping-Andersen in the 1990s. This approach defines a welfare state as a state whose policies structure both the society, as well as its social and economic relationships. It primarily sees the state in relation to the market and the family structures, and asks how state activities are interlocked with these two institutions in shaping social conditions and providing social welfare benefits. Education is not considered as a relevant component.

Esping-Andersen's classification distinguishes between three 'ideal types' of welfare state regimes (Table 1): The Anglo-Saxon 'liberal' welfare states (i.e., the United States, Canada, and Australia) emphasize the importance of market forces and offer only modest and means-tested benefits. The Central European 'conservative' welfare states (i.e., Germany, France, Italy, Austria, and the Netherlands) link the provision of social security to status differences and traditional family models. The Nordic 'social democratic' welfare states grant social security on the basis of individual social rights and direct their social policies toward achieving more equality, both in the public and private sphere.

As established by Esping-Andersen, welfare regimes could, but do not necessarily need to, include welfare states.

Table 1 Typology of welfare state regimes in developed countries

Welfare state regime types	*Characteristics*
Social democratic (Denmark, Sweden, Norway, Finland, and the Netherlands)	Welfare provision is characterized by universal and comparatively generous benefits; a commitment to full employment and income protection; a strongly interventionist state promotes equality through a redistributive social security system
Conservative (Austria, Belgium, Germany, and France)	'Status differentiating' welfare programs: Benefits are earnings related, administered through the employer, and geared toward maintaining existing social patterns. The role of the family is emphasized and the redistributive capacity is minimal.
Liberal (Canada, Great Britain, and USA)	The state provides minimal welfare benefits. Benefits are modest and often attract strict entitlement criteria, and recipients are usually indigent and stigmatized.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism, vol. 6. Polity press, Cambridge.

Understood as constellations of practices that reallocate resources, welfare regimes may or may not include a well-developed public policy component. Consequently, the study of welfare regimes sheds light on the redistribution of resources, whether or not such redistribution is state-led. If labor markets are well organized and able to absorb the labor force, the less the state will play a role in redistribution, and the more dependent on markets and primary distribution the population will become. Yet when labor markets are highly informal and lead to large inequalities in income, nonmarket-based practices, such as family and community ties, expectations, and obligations, become more significant. Based on Esping-Andersen's work and adapted to Latin America, different authors (considering commodification, decommodification, and defamiliaritzation dimensions) have identified three regional patterns or welfare regimes: Productivist, Protectionist, and Familiarist. In these regimes, states do not necessarily play the same central role in defining welfare regimes as in the developed world; however, they do share some characteristics.

Welfare regimes in Latin American countries exhibit variation in their social policies, for example, in the type, combination, and measure of assistance and benefits they provide. The combination of their social provisions bears important consequences for the socioeconomic opportunities and outcomes of individuals and groups. For example, rapid development in the region has contributed to the erosion of absolute poverty but has simultaneously heightened insecurity and vulnerability. A good system of social policy helps people throughout their life cycle, protecting their first steps, ensuring access to opportunities for education and jobs, covering their risks of disease and disability, and guaranteeing old age with dignity. The welfare policies of Latin American countries include some level of provision for education, health care, old-age pensions, provision of cash transfers, and other safety net provisions. One of the first priorities in these countries has been education policy. In general, Latin American countries spend more or less on education than do Organisation for Economic Co-operation and Development (OECD) countries on average; however, they differ in terms of what they spend the money on and their education learning outcomes. Spending does not necessarily guarantee higher quality; indeed, Latin American countries do not appear to secure the greatest possible value from what is spent.

In this as in most other internationally relevant typologies, education has hardly been taken into direct account. Thus, the significance of education with regard to modern welfare development is completely underestimated. Education has a significant impact not only on the general population's social and economic development, but also on its prospective opportunities and perspectives. This in turn shapes the individuals' future as a crucial resource for the quality of life and the opportunities of social participation. A country's educational policy is thus as integral and decisive for a 'social welfare regime' as its social security policies (hereinafter referred to as 'social policies') (Allmendinger and Leibfried, 2003).

The Combination of Educational and Social Policies in the Three Regimes

The three ideal-typical welfare regimes mentioned here differ with regard to the value they attach to educational policy. In a historical-comparative analysis, Hega and Hokenmaier (2002) outline the importance of education in the development of welfare state traditions. They come to a clear conclusion:

- The Central European, structurally conservative 'welfare regime,' as established, for example, in Germany, traditionally attaches a minor significance to the role of educational policy. The focus lies on historical structures rooted in the social conditions of life. The entire social welfare policy, i.e., the protection of the population's quality of life, is essentially geared to the stabilization of existing structures.
- The market-oriented Anglo-Saxon regime, as established, for example, in Britain, pursues a development approach and is traditionally inclined to extensively invest in the educational sector in order to enable individuals to attain a stronger position on the job market, which minimizes the amount of future social transfer payments required to cover possible risks.
- The Scandinavian regime, as established, for example, in Sweden, traditionally invests in both the education of the individual members of the society, as well as in their protection against possible risks in the course of their lives. This regime is intended to provide all citizens with the opportunity to develop their own potential, and at the same time guaranteeing a high quality of life.

The results achieved by these three different approaches are very clear: the Scandinavian, structurally social-democratic welfare states, invest in both the potential status of their citizens as well as in status maintenance. Consequently, the Scandinavian regime invests the largest amounts both in education as well as in social welfare. At the root of this approach is a model of societal and individual autonomy that aims to provide all citizens with the same publicly funded social security benefits.

The liberal Anglo-Saxon countries invest heavily in the development of their citizens' potential and thus mainly in education and less in social security. The philosophy of this approach is to provide favorable starting conditions, which in combination with the opportunity of good education support the general development of all citizens. The further development of their own chances now lies in the hands of each individual and no longer constitutes a state affair. As a result, social security spending is low.

In structurally conservative, family-oriented Central European countries, a third political variation is typical: Education expenditure as social investment for status acquisition is rather low. The state focuses on status maintenance by enabling its citizens to uphold the particular standard of life traditionally achieved by their family. Consequently, larger amounts are invested in social security than in education.

Current OECD statistics confirm these tendencies (OECD, 2010, 2011). In addition to a comparatively small public percentage, a considerable private segment is available. As a prototype of a social-democratic welfare state, Swedish investments are comparatively lower. However, this is exclusively due to the lower percentage of private education expenditure. When taking only public means into account, Sweden would hold the leading position. Comparing the three country prototypes,

Germany as a conservative welfare state would take the last place in terms of public expenditure on education. Sweden belongs to the leading group of nations with regard to public educational and social expenditures; the United States holds a middle position with regard to educational expenditure but finds itself on the outer margins with regard to social expenditure. Germany takes a higher position with regard to social expenditures but drops to the lower third with regard to education expenditure. The other countries that could be assigned to one of these three welfare regimes do not deviate very much from the patterns established.

Structures of the Various Educational Systems

The varying importance attributed to educational policy affects the structure of each country's educational system. Comparative research has identified a variety of structural aspects within educational systems relevant to educational outcomes (Hofman et al., 2004).

Hofman and colleagues suggest three aspects that characterize the institutional context of schools:

1. The effect of the manner in which public and private education is financed.
2. The administration of public and private education.
3. The institutional effects of school choice and community involvement.

Regarding the type of financing of public and private education, we can see that most education systems include a private and a public sector, but are still funded in different ways. In some countries, the state finances both public and private schools equally, and implements identical policies in these schools. However, there are still differences in quality between public and private education. Other countries combine equal funding and identical policies in public and private schools. Such countries provide opportunities within the public education system to establish or choose grant-supported private schools. The size of the private educational sector may vary widely depending on the amount received from public funds (grant support) and/or from direct financial contributions by parents, corporate donations, or inherited funds.

When we consider the administration of public and private education, we can see that in many countries autonomous school boards administer grant-supported private schools, while traditionally local authorities govern public schools. Consequently, public education depends on local government policy while private education functions more autonomously.

This directly ties into the effect of choice and community involvement in public and private education since many studies trace the discrepancies between public and private schools to environmental differences and their internal social systems. Parental involvement in school affairs (e.g., school choice; parental involvement in schoolwork and in school life, as well as influence on school board decisions) reflects positively on academic performances. Both public and private schools are subject to administrative bureaucratic control, whereas public schools – dependent on government bureaucracy – must submit to extensive administrative control. By contrast, private schools operate more autonomously, foster contact between staff and parents, and develop a common educational basis within the school. Furthermore, the quality of school boards makes a difference. This is also true when student background and school characteristics are taken into account. The cooperation between school board, school faculty, and parents (involved in school life) can have a particularly positive effect on student achievement.

All of these three aspects that affect the structure of educational systems are relevant in defining variations in the effectiveness of the two sectors. According to the analysis, differences in funding possibilities, administrative structures, and community involvement and choice, all affect student achievement. The latter is higher in private than in public education. Consequently, countries such as the Netherlands and Belgium, which have a large private education sector (>30–100%), score higher on quality and equity than Germany and Sweden, which have a smaller private sector involvement (<5%). France and Italy with their medium-sized private sector score in between the two (5–30%).

Countries with a relatively high percentage of students in grant-supported educational institutions perform better in terms of quality when compared to countries that are dominated by the public sector or countries where the proportion of students attending grant-supported institutions is small (<10%). Countries with a relatively high percentage of students attending grant-supported schools tend to perform better than others in terms of equality. The gap in the mathematics performance level between native students with a low socioeconomic status and native students with a high or medium socioeconomic status tends to be smaller in these cases. When we look at the issue of funding and financial basis, we see that education expenditure per capita or as percentage of gross domestic product is not directly related to educational outcomes.

Even though he leaves out Scandinavia in his analysis, Kerckhoff's (2001) attempts to compare the effect educational systems have on social inequalities in educational outcomes is a valuable piece of research to consider. He argues that a society's educational institutions can be described as the 'sorting machine' that stratifies its population. He then identifies four comparative characteristics: stratification, standardization, vocational specificity, and choice. Here, conservative states rank high on stratification and vocational specification while liberal states rank low. This may to a certain extent support existing policy traditions (**Table 2**).

In summary, all studies mentioned point to difficulties experienced in comparing educational systems. This is due to the large policy differences between these systems, and their resulting effect on education. All studies further demonstrate that the structure of educational systems depends on the individual countries' historical and political context.

The Influence of Welfare Policy on Educational Achievements

The characteristics of the prevalent educational policies in the particular types of welfare states are decisive for the results produced in the respective educational system. Nowadays,

Table 2 Characteristics of educational systems and their effect on social stratification processes

Dimensions	*High*	*Medium*	*Low*
Stratification	GER	UK	FR, USA
Degree to which systems have clearly differentiated types of schools whose curricula are defined as 'higher' and 'lower' (…) program offerings in secondary school are associated with different degrees of access to opportunities (for additional advanced schooling)			
Standardization	FR	GER, UK	USA
Refers to the degree to which the quality of education meets the same standards nationwide (…) higher in states that are controlled by the central government			
Vocational specification	GER	UK	FR, USA
Extent to which educational systems offer curricula that are designed to prepare students for particular vocations and award credentials that are vocationally specific			
Choice	FR, USA	UK	GER
Extent to which students have opportunities to make choices			

In accordance with Kerckhoff, A.C., 2001. Education and social stratification processes in comparative perspective. Sociology of Education, 3–18.

there is an international consensus concerning the criteria according to which the standard of the educational balance sheet should be assessed. The following two goals are regarded as benchmarks:

1. The educational level of the population should be as high as possible when compared to the level of qualification and competencies achieved.
2. When compared to the differences in the level of education achieved by the various population groups according to gender and social background, the degree of inequality in education should be as low as possible.

Prior findings indicate that there is an increased probability for an accurate picture if various data sources are used. In other words, the strongest assessment programs will rely on a mix of different measurement tools. Indirect measurement tools include data from surveys of school principals, students, and/or parents. Other statistics include financing information as well as indicators of the quality and equity of education outcomes. Education at a Glance (OECD, 2011), for example, uses several indicators such as completion, repetition, and dropout rates. These measurement tools cannot reflect what students have learned or what they are capable of doing as a result of their education. They can, however, provide information on school characteristics (public/private, autonomy, management), student background (socioeconomic status, gender, ethnicity, immigrant background), and parental influence.

Direct measurement provides us with more evidence regarding the increase in students' knowledge and abilities over a period of time. National and international standardized tests, e.g., Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study, are one type of direct measurement tool. Such standardized tests are designed to assess general education skills or measure the knowledge of specific disciplines. Information derived from this assessment can be used for a variety of evaluative purposes, e.g., to examine quality or equality issues (e.g., socioeconomic differences) in learning outcomes.

Educational Achievements in Different Welfare Regimes

By summarizing all the results of the studies cited, we can arrive at the following conclusions. The three welfare regimes clearly put different emphases on social and educational policies. The Scandinavian countries pursue policies that ensure extensive social and educational benefits. They divide their political authority and financial resources between these two fields. Countries within the conservative regime clearly prioritize social security politics, where educational policy is clearly of lesser importance. The liberal countries focus on 'investive' educational policy, which they regard as a substitute for 'non-investive' social security policy.

The Educational Level of the General Population

Empirically comparable international studies (OECD, 2010) are first utilized to evaluate the success of the three chosen prototypical countries in their attainment of their goal of improving the educational level of the population, i.e., raising the qualification and competence level of the general population:

- Sweden uses a dual combination strategy of focusing on status security and investment in order to ensure that every member of the population attains the status of a top performer. According to the results, the well-balanced investment in (1) educational expenses designed to strengthen the developmental possibilities and the status-acquisition of young members of the population and (2) the social protection that ensures that the status of all citizens is stabilized has stood the test of time, even in crisis situations. This dual strategy that cultivates achievement potential and encourages qualified citizens to choose their own professional path has proven immensely successful.
- By comparison, the political tradition of the market-oriented and liberal United States has had little success. Even though the country has invested a higher percentage of the gross national product in their educational system than Germany, the educational results are not as good comparably. This leads to the conclusion that the educational outcomes in modern welfare states are particularly high if citizens are assured that their status will be secure after the completion of their education. This signalizes to the citizen that the state is concerned with their continued well-being. The United States is one of the countries that does not meet these conditions. Early investments in the educational system are not supported by later investments and can thus not realize their full fruition.
- The policy pursued in Germany leads to moderate results – the investments in status achievement-oriented educational resources are relatively low. A large part of the gross national product is spent on the protection of the status that has been achieved. Only very few resources are devoted to the development of competences and potentials. It could be argued that the generous social security benefits provided to citizens after the completion of their educational training leads to a counterproductive effect – the educational system does not ensure that all opportunities for an optimal status achievement are used.

Obviously, the great success of the Scandinavian regime in attaining the educational target 'quality' is the excellent strategy

of combining status-achievement and status-protection policies. According to Allmendinger and Leibfried (2003), the inherent logic of social security policies can be understood as a retrospective intervention since the adjustment of resources is not available to the unemployed. In contrast to this policy, which can almost be considered as 'rehabilitative,' this educational policy has an empowering 'habilitative' character.

What distinguishes educational policy from social policy is the prospect of intervention. Educational policy, an investment in the developing capacities, primarily of the younger members of the society, opens a chance for them to become an integrated member of the workforce and also provides the opportunity of further education in the course of their career, and thus the opportunity to maintain their professional competence. In this sense, social and educational policies should be understood as two complementary political intervention strategies. If the goal is a comprehensive welfare policy, they must not be played off against one another.

The Degree of Educational Equality According to Social Background

Prior evidence shows that a balanced combination of social and educational policies must go hand in hand with an effective organization of the educational system if the second goal of educational policy, i.e., as much equality as possible in educational quality among population groups from different backgrounds, is to be achieved. On the basis of previous findings, Schütz et al. (2008) have identified three criteria that best pinpoint the greatest effect that social background has on the performance record:

1. The highest possible number of children in the preschool sector.
2. The longest possible stay of the children in these institutions.
3. The differentiation and placement of children in different types of schools according to their achievement follows as late as possible.

Each of these characteristics raises the potential of the educational system to increase the academic achievement level of all children, irrespective of the input received in the parental home. It especially raises the academic performance of children of lower social origins.

Subsequently, the question is how the countries of the three welfare regimes 'perform' when these criteria are taken into account in cross-national studies (OECD, 2010).

- With only a few exceptions, the Nordic countries such as Sweden have the highest number of children in preschool institutions. They also have the longest possible time of stay. This can be attributed to the political philosophy that all citizens should be granted personal autonomy in shaping their lives as early as possible. The defined goal of this policy is to minimize the dependence on social background in children's performance development. Thus, equal social rights are attainable for all citizens. Hence, children are assigned to different types of schools very late.
- By comparison, in the Anglo-Saxon countries and especially in the United States, which is the prototype of a liberal, market-driven country, a considerably lower number of three- to five-year-olds is in public preschool institutions, where parents make a very low contribution or none at all. The ratio is approximately 60%, whereas the ratio in Sweden is nearly 90% (OECD, 2010). The amount of time the average child spends in preschool institutions is also shorter than in Sweden: 3 instead of 4 years. 'Tracking' only occurs late in the students' educational career.
- Similar to the United States, the conservative countries also have a low number of children in preschool institutions and also a shorter period of stay than in Sweden. In contrast to Sweden and the United States, a third factor, the very early 'tracking', i.e., the assignment of students to different school forms, is a defining characteristic for the regime in Germany. In Germany, children are normally assigned to different types of schools at the age of 10 years; in Sweden and the United States, this is only done at the age of 16 years. After that assignment to different types of schools or 'tracking', the developmental impulses offered by a heterogeneous student body are reduced. This negatively impacts the weakest performance groups: in Germany these are the students attending 'Hauptschulen' and 'Förderschulen' (Mutz, 2006; Demeuse et al., 2003). Consequently, student performance is more strongly influenced by social background. The path dependence of German educational policy is apparent here: from its traditional philosophy, German social welfare policy sees the family as the center of children's 'natural' social network. This has to be shielded as much as possible from any state intervention, even that of the educational system.

Table 3 provides us with a summary of the comprehensive typology of welfare regimes in developed countries, while taking their social and educational policies into account for further overview.

The results are thus fairly obvious: of all three educational systems, countries with a social democratic regime offer the best starting position. All other countries that wish to ensure a high level of education of their population can definitely be advised to incorporate these successful structural settings into their political tradition.

The Relevance of Education Policy in Developing Countries

Investment in education is crucial not just for the future welfare of highly developed countries. Also in the developing countries, the aims of education policy are twofold, encompassing both goals of efficient allocation and goals of equitable distribution. The results that have been reported from comparative research in highly developed countries are also valid for developing countries: the efficiency of a school system can be substantially increased by institutional reforms that focus on incentives for all actors in the system to increase the performance of students. Apparently, the combination of accountability (e.g., of teachers) and school autonomy (e.g., external tests with local decision making) is the most promising institutional structure. In addition, competition and choice between schools have a strong potential to shift school systems to a higher level of efficiency and equity.

Table 3 Comprehensive typology of welfare regimes in developed countries

	Welfare state regimes		
Characteristics	*'Social democratic'*	*'Conservative'*	*'Liberal'*
Freedom from servitude for all citizens (decommodification); financial assistance to secure means of subsistence	Extremely generous; universal for all citizens. Goal is the safeguarding of human rights and the freedom of all citizens from pressures of any kind	Generous; applies primarily to citizens in breadwinner roles. Goal is the safeguarding and strengthening of existing social networks	Severely restricted; only applicable to needy members of the population. Goal is the safeguarding of all citizens' productivity and their ability to support themselves
Relationship between the citizens and the state; social order (stratification)	To further social mobility through supporting all citizens, irrespective of background or gender in status attainment and status maintenance	To strengthen the established incorporation of citizens in social (family) networks and stabilize existing social stratification	To establish the self-reliance of all citizens; a targeted promotion of their status-attainment; encourage social mobility and personal responsibility
Relationship between social and educational policy	Combination of status attainment and status safeguarding for all age and population groups during the complete life cycle	Excessive focus on social (status-) safeguarding policy especially for population groups prior to entering a career and during their professional life	Intense advancement of education for age and population groups preparing for professional career; no comprehensive safeguarding of status for those who have been educated
Organization of the educational system	Multicentered with autonomous educational institutions, a unitary school system with few possibilities of choice	Centralized with dependent educational institutions; strongly differentiated school system with early division and selection into various school forms	Decentralized with autonomous but communally dependant educational institutions; late differentiation into various performance groups
Performance record	A high degree of social security; a high level of educational quality and low level of educational inequality	A high degree of social security; medium quality of education; high level of educational inequality	Low level of social security; low level of educational quality; medium level of educational inequality

Furthermore, it is very efficient to invest more money into preprimary education and to avoid early tracking of students into different kinds of schools. Greater competition and choice significantly raises both test scores and the productivity of schools. The effects of school choice are significantly larger for students attending subsidized private schools, and insignificant for students attending public schools facing softer budget constraints (Gallegos, 2004).

Colombia, Chile, and Mexico have been selected as prototypical examples for the connection between different (welfare) policy traditions and specific educational policy profiles in the educational systems that exist in Latin America as well as in other countries of the world (Gough and Wood, 2004). In all three countries, education outcomes are influenced not only by education policy, but also by public policy approaches in general. For education outcomes, an important factor besides education policy is family policy, because it has an impact not only on preschool education but also on health, nutrition, and emotional stability.

Most countries in Latin America have achieved universal primary education and are now witnessing a rapid expansion of secondary education. Latin American governments are committed to providing universal access to basic education, which includes primary and lower secondary education in order to assure a more successful transition from primary to secondary education. In Colombia, Chile, and Mexico, there are no significant gender differences when it comes to educational transition. The transition from primary to lower secondary education in 2005 in the prototypical examples selected was as follows: Chile (male 98% and female 98%), Colombia (90% and 90%), and Mexico (95% and 93%). These countries have experimented with a particularly wide range of governance reforms, such as increased school choice and competition through vouchers, conditional cash transfer programs to poor families, participation of parents and community in decision making, decentralization of government finance, and the use of information form learning assessments. However, these reforms have not unambiguously raised academic achievement standards. Often, they have widened inequalities (UNESCO, 2009, and UNESCO, 2008).

Citizens are encouraged to become wage earners demanded by the commodified labor market. The measures for commodifying the labor force include the enhancement of education and work training programs for qualified workforce. It is expected that education policy and public policy can improve the human capital and enhance the productivity of labor. Particularly, there is a linkage between learning outcomes and labor market needs. Nevertheless, according to PISA 2006, Latin American countries consistently perform at the bottom not only in the quality of outcomes but also in the equity of their distribution (learning opportunities). These low levels of cognitive skills can account for the low productivity, low wages, low social mobility, income and social inequalities, and poor growth performance of Latin America.

Applying the Welfare State Typology to Developing Countries

Not only in highly developed countries, but also in developing and threshold countries, has education become one of the most important drivers of a nation's economic growth because of its role in enhancing productivity. However, structural features of the traditional European and North American welfare states cannot directly be adapted to other parts of the world where states are faced with politically and socially challenging circumstances such as corruption, weak institutions, poverty, as well as imperfect and partial labor markets.

While the states in the advanced countries are trying to decommodify the labor force and reduce the social stratification and social inequalities by social policies, the states in developing countries are still pushing citizens toward being dependent on the labor market by qualifying and therefore commodifying their labor force. Thus, these countries are still far away from being real 'welfare states' by the standards of the highly developed world. In developing countries, a large proportion of the population is unable to deal with social risks by accessing state services or participating in labor markets. For this reason, individuals are required to rely heavily on family and communal arrangements. In general, gender (the importance of female spouses with paid work), unpaid work (the absence of robust labor and state institutions), and families (the presence of extended and compound families, the relevance of international remittances) are important to understand welfare mechanisms (Gough and Wood, 2004).

During the last decade, there have been several studies (e.g., Filgueira, 1998; Huber and Stephens, 2005; Rudra, 2007) that have tried to adapt Esping-Andersen's work to the Latin American context. Rudra (2007), for example, makes an important contribution by focusing on the question whether Latin American countries rely on public policies that predominantly promote access to the market (productivist regimes) or that predominantly protect people from the market (protectionist regimes). The author argues that Latin American countries and their 'welfare states' have a dual role of commodification and decommodification, where access to paid work cannot be taken for granted.

Gough and Wood (2004) point out that in Latin American countries, states in many areas of policy are practically nonexistent, labor markets often exclude the majority of the population, and a great deal of welfare production rests on families and social support networks. In informal welfare regimes, as found in Latin America, most of the population relies on family or community ties. Barrientos (2004) suggests that Latin America has shifted from a conservative-informal to a liberal-informal welfare regime. In the last two decades, collectively shared risks have become few, public policies have diminished, and individuals are increasingly on their own. Thus, the region resembles the liberal regime in developed countries; but unlike them, most countries in Latin American lack solid targeted state programs.

This is based on the valuable work done by Martinez and Vorrend (2009), after his consideration of Esping-Andersen (1990), Filgueira (1998), Gough and Wood (2004), and Rudra's (2007) studies on welfare regimes and welfare states, and based on his study of the complex interaction among the labor market, families, and public policy. Martinez and Vorrend (2009) goes on to identify the three welfare regimes in Latin American countries according to the following typology: protectionist, productivist, and familiarist.

Education Policy in the Productivist Welfare Regime

The productivist regime relies on the market and mainly on individually funded social security. Many individuals, particularly in middle-income sectors, fall between the cracks of targeted public policies and ineffective markets. A prototypical country of this regime is Chile. The country is trying to follow the liberal path and strengthen market forces for the development of the competences and the productivity of its population.

Education policy is part of this policy tradition. Chile tries to promote policy interventions to push forward equity and social mobility, in order to raise educational standards and improve the quality of education delivered. However, the accompanying legislative framework supports the perpetuation of a market-wise organized system that fosters competition and acts against a more equitable distribution of educational opportunity for all. There are inherent tensions here because of the complex relationship between the state and the market over the financing and delivery of education.

The transformative potential of education, its importance to national development and social cohesion, has been reflected in an increased share of public spending. Yet, while state investment has grown, so too has private expenditure as the cost of education is increasingly transferred on to individuals and their families by government financing initiatives. The tendency for children to attend schools with shared funding arrangements is higher among wealthier households, thereby increasing parental contributions to education. The shift from state schools to privately run establishments is continuing as these schools grow in number and their share of the total enrollment increases. The popular perception of private schools as providing a higher quality education than state schools is reflected in the changing distribution of student numbers.

Further, the socioeconomic stratification of the education system distributes access to quality education conditional on the purchasing power of the family. Yet, when the socioeconomic background of the students is factored in, the subsidized private schools perform no better than municipal schools. Consequently, private schools are failing to compensate for the social disadvantage of low-income children and to help in adding value for middle-class children. Therefore, further investment of state resources to subsidized private schools through a voucher system, and the further general expansion of this sector should be challenged as unlikely to provide a positive social return on public investment. The goal of promoting equity through a competitive, market-oriented, and socially stratified system must be questioned in terms of the state's capacity to influence and act in these areas, and the level of accountability to which schools in receipt of public funds are held (Matear, 2007).

Education Policy in the Protectionist Regime

The protectionist regime can be characterized as a state-stratified system whose targeting is based on collectively funding social security. In the case of Mexico, this policy has led to the creation of a fragmented, unequal, and incomplete welfare system that has excluded large sectors of the population, especially the rural workers and informal urban workers. Mexico is the typical example of a nominally universalistic model, modified in reality by fragmentation and embellished by targeted policies of low quality (Martinez and Vorrend, 2009).

In Mexico, there are wide disparities in quality between public and private schools. As a consequence, a higher proportion of middle-income groups has left public schools in search of better quality alternatives. The educational system is divided between those who can afford quality education and those who have to study in public schools.

For financing poverty programs that include an educational component, the Mexican government has increased state spending, often with direct credits from multilateral financial institutions like the World Bank. The most important antipoverty program Oportunidades, a conditional cash transfer program to the poor, provides resources to preferentially support the demand to cover the needs of a defined or targeted population. Oportunidades is a strictly targeted program mounted on universalist institutions like primary and secondary schools. It is a program of direct transfers, in cash and conditioned for selected poor households, and has a triple composition by which it seeks a sense of integration: support for education, health, and nutrition, with the goal of helping homes to overcome the intergenerational transmission of poverty. The educational component seeks to promote school attendance with educational support in grants for each of the sons and daughters studying between third grade of primary school and the third year of secondary school.

Education Policy in the Familiarist Regime

The familiarist welfare regime shows the poorest performance. Social protection and the formation of human capital are minimally decommodified. For education, commodification is the lowest in comparison to the productivist and protectionist countries, and wages are also the lowest. Colombia is an example for this regime. The vast majority of the population depends on familial and communal strategies in the context of exclusive labor markets (Martinez and Vorrend, 2009). This regime envisages a great number of dependent people who need to rely on the commodification of fewer available people.

Over the last two decades, Colombia has made great strides in improving access to education and has made some improvements in internal efficiency. However, it faces challenges such as low completion rates in secondary education, inequality in access and attainment, and poor quality of education. Primary education is mandatory through primary enrollment and is beginning to reach universality. Important gains have been made in secondary enrollment and completion rates; however, serious problems with enrollment and retention at upper secondary grades persist. Some possible reasons behind the high dropout rate at the secondary level include (1) concerns about the high cost of education (distance to schools, cost of tuition and supplies, etc.), (2) increasing opportunity costs for older students (making education less attractive than working), (3) low quality and relevance of learning, and (4) the lack of realistic opportunities to continue education beyond the secondary level. Disparities in enrollment and completion also exist between poor and rich regions and urban and rural areas. This disparity has been associated in part with the political and criminal violence that still prevails in some rural areas (**Table 4**).

Table 4 Typology of welfare state regimes in developing countries

Variables and indicators	*Cluster 1: Productivist Chile*	*Cluster 2: Protectionist Mexico*	*Cluster 3: Familiarist Columbia*
Commodification 1	High	High	Low
Effectiveness of labor markets to incorporate and remunerate the labor force	Highest	Medium	Lowest
Formalization of the labor market (occupied salaried workers)	Highest	Medium	Lowest
Income levels	Highest	Medium	Lowest
% Of population below poverty line	22.7%	28.86%	53%
Remittance (the importance of transnational commodification of the labor force)	0.10% Of GNP	1.05% Of GNP	More than 6% of GNP
Rural population	12%	23%	More than 30%
Decommodification 2	High	High	Very low
Proportion of the EAP occupied in the public sector	Highest	Somewhat lower	Half or less than half of what is found in clusters 1 and 2
Social expenditure in health and education	Highest	Mexico: Between clusters 2 and 3	Consistently much lower than in clusters 1 and 2
Fiscal priority (or effort) devoted to social policy as % of GNP	Less than in cluster 2	Mexico: Lower than in cluster 3 (9.8% in comparison with an average of 19% in cluster 2)	Lowest
Proportion of salaried workers with social insurance	Less than in cluster 2	Highest	Quite low
Private education enrollment 3	Highest (36%)	Lowest (13.46%)	Medium (between 26 and 28%)
Defamiliarization	High	High	High
Extended and compound families	Lowest	Lowest	Highest
Full-time female unpaid work	Central role	Central role	Central role
Nuclear families with spouses devoted to full-time unpaid work (the traditional 'male breadwinner' family model)	Higher	Higher	Lower than in clusters 2 and 3
Demographic transition	Advanced	Somewhat less	Incipient
Population of children under 12 years	25% Or less	25% Or less	30% Or more
Important role of families	Nuclear families (less burdened by dependent family members)	Nuclear families (less burdened by dependent family members)	Extended families (greater number of dependent people who need to rely on the commodification of fewer available people)

GNP, gross national product.

Consequences for the Future Development of Educational Policy

As our analysis has shown, there is an internationally comparable relationship between welfare traditions and educational policy. This is evident in both highly developed as well as in developing countries, and affects the educational results within each specific country and cluster. There are unambiguous correlations between

1. the educational success of a country as measured against the attained quality of educational rate of return and the equal distribution of these returns according to social background
2. and the importance accorded to educational policy in comparison to social policy and the resultant organization of the educational system in the respective country.

The comparison between countries characterized by these various welfare regimes serves to illustrate the relationship between welfare policy and the educational balance sheet, and their unequal distribution. These countries could be seen as 'real-life prototypes' of the three ideal types of welfare regimes, or alternatively, could at least be analytically approached from the perspective of such welfare regimes. At the same time for the case of Latin America, variations across regimes are significant: (1) the Productivist regime performs better in the supply of salaried work, (2) the three regimes have variations in social policy reforms, and (3) the three regimes have variation in allocation of social public expenditure. Our findings provide support that particular kinds of welfare regimes have distinctive educational policy profiles, and that particular kinds of welfare regimes and their educational policy have an influence on their educational outcomes.

Overall, each of the three welfare regimes provides diverse educational scenarios that alter the architecture of social policy. What clearly shines through the findings and research done is that the basic components of education and social policy of countries within the social democratic cluster should be a model and adopted by the countries that fall under the other

regime types. In Latin America, there is also much to be learnt by paying attention to which of the policies within the clusters produces the best results so that they can be adopted as well in other cases. But first, further research will be required in order to analyze how in the face of high path dependency other regimes could adopt certain components from countries with effective social policies, and in what way they could be implemented successfully.

These approaches demonstrate that a comprehensive typology of the welfare regimes should expand the seminal work achieved by Esping-Andersen and should include educational policy alongside social security policy. The research clearly confirmed how effective investment is in the preschool sector, not only in increasing performance, but also in combating inequalities. Our analysis has shown that not only social policy – as a policy that protects quality of life and the attained social status also in crisis situations – is influenced by the deeply rooted power structures and traditions characteristic for a specific welfare system, but also educational policy is significantly determined by the traditions adhered to by the welfare states. Educational policy as the policy that grants status attainment is likewise path dependent: it is impacted by the elementary commercial, legal, and institutional focus that gives social policy its direction.

Acknowledgment

The information about the Latin American countries was compiled by Nancy Cieza. We are very grateful for her assistance.

See also: Skills Training: Public Policy Aspects; Welfare Policy Comparisons.

Bibliography

Allmendinger, J., Leibfried, S., 2003. Education and the welfare state: the four worlds of competence production. Journal of European Social Policy 13 (1), 63–81.

Carrasco, M.R., Torrecilla, F.J.M., 2009. Learning assessment in Latin America. School performance behaviour and trends of Latin American pupils in primary and secondary education. Educational Sciences Journal 9, 31–46.

Demeuse, M., Baye, A., Straeten, M.H., Nicaise, J., 2003. Equity of the European Educational Systems, a Set of Indicators. Synthesis of the Report. European Group of Research on Equity of Educational Systems.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism, vol. 6. Polity press, Cambridge.

Ferragina, E., Seeleib-Kaiser, M., 2011. Thematic review: welfare regime debate: past, present, futures? Policy & Politics 39 (4), 583–611.

Franzoni, J.M., 2008. Welfare regimes in Latin America: capturing constellations of markets, families, and policies. Latin American Politics and Society 50 (2), 67–100.

Gough, I., Wood, G., 2004. Insecurity and Welfare Regimes in Asia, Africa and Latin America: Social Policy in Development Contexts. Cambridge University Press.

Hega, G.M., Hokenmaier, K.G., 2002. The welfare state and education: a comparison of social and educational policy in Advanced industrial societies. German Policy Studies 2 (1), 143–173.

Hofman, R.H., Hofman, W.H.A., Gray, J.M., Daly, P. (Eds.), 2004. Institutional Context of Education Systems in Europe: A Cross-country Comparison on Quality and Equity. Springer Science+Business Media.

Huber, E., Stephens, J.D., 2005. Successful Social Policy Regimes? Political Economy, Politics, and the Structure of Social Policy in Argentina, Chile, Uruguay, and Costa Rica. Conference Documents, University of Notre Dame, Notre Dame.

Hudson, J., Kühner, S., 2012. Analyzing the productive and protective dimensions of welfare: looking beyond the OECD. Social Policy & Administration 46 (1), 35–60.

Hurrelmann, K., 1984. Societal and organizational factors of stress on students in school. European Journal of Teacher Education 7, 181–190.

Kerckhoff, A.C., 2001. Education and social stratification processes in comparative perspective. Sociology of Education, 3–18.

Levy, F., Murnane, R.J., 2003. The skill content of recent technological change: an empirical exploration. The Quarterly Journal of Economics 118 (4), 1279–1333.

Martinez Franzoni, J., Vorrend, K., 2009. The role of distributional coalitions in welfare regimes: Chile, Costa Rica and El Salvador. Social Policy & administration 43, 364–381.

Matear, A., 2007. Tensions between state and market in Chile: educational policy and culture. Revista Europea de Estudios Latinoamericanos y del Caribe/European Review of Latin American and Caribbean Studies, 61–82.

OECD, 2009. PISA – the OECD Programme for International Student Assessment. Brochure PISA. OECD, Paris.

OECD, 2010. Education at a Glance 2010. OECD Indicators. OECD, Paris.

OECD, 2011. Society at a Glance 2011. OECD Social Indicators. OECD, Paris.

Rice, D., 2013. Beyond welfare regimes: from empirical typology to conceptual ideal types. Social Policy & Administration 47 (1), 93–110.

Rudra, N., 2007. Welfare states in developing countries: unique or universal? Journal of Politics 69 (2), 378–396.

Schmitt, C., Starke, P., 2011. Explaining convergence of OECD welfare states: a conditional approach. Journal of European Social Policy 21 (2), 120–135.

Schütz, G., Ursprung, H.W., Wößmann, L., 2008. Education policy and equality of opportunity. Kyklos 61 (2), 279–308.

Scruggs, L.A., Allan, J.P., 2008. Social stratification and welfare regimes for the Twenty-First Century. World Politics 60, 642–664.

Soares, S., Guerreiro Osório, R., Veras Soares, F., Medeiros, M., Zepeda, E., 2009. Conditional cash transfers in Brazil, Chile and Mexico: impacts upon inequality. Estudios Económicos 1, 207–224.

Van de Werfhorst, H.G., Mijs, J.J., 2010. Achievement inequality and the institutional structure of educational systems: a comparative perspective. Annual Review of Sociology 36, 407–428.

Wößmann, L., Schütz, G., 2006. Efficiency and Equity in European Education and Training Systems. Analytical Report for the European Commission prepared by the European Expert Network on Economics of Education. Comisión Europea, Bruselas. http://www.eenee.de/portal/page/portal/EENEEContent/_IMPORT_TELECENTRUM/DOCS/SWP%20EENEE-Part%20060426.

Wood, G., Gough, I., 2006. A comparative welfare regime approach to global social policy. World Development 34 (10), 1696–1712.

Welfare and Work

Robert A Moffitt, Johns Hopkins University, Baltimore, MD, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

All industrialized countries have welfare or social assistance programs intended to provide income support to low-income families. Such programs have the potential to reduce work effort among recipients and the research evidence indicates that such incentives exist and are sometimes large in magnitude. Work disincentives can be ameliorated with financial and nonfinancial incentives such as job search assistance, sanctions, and work requirements. Earnings subsidies are one way to provide financial incentives to those with very low-work capacities.

All advanced industrialized countries offer welfare programs. Often they go by other names, such as social assistance or by special titles (medical assistance, food assistance, and so on), but the common feature of all programs of this type is that they are 'means-tested.' A means-tested program is one that provides benefits or other support only to families or individuals with income or assets below some threshold. There may be other eligibility conditions in addition to low income or assets – such having a disability, having children, or minimum work requirements – and the threshold may vary across individuals and families of different types; but low income is always one of the conditions. Means-tested programs are often contrasted with universal programs, which provide some kind of government benefit to all those in the population, regardless of income, and are often further contrasted with social insurance programs – such unemployment insurance or retirement benefit programs, for example – where the primary condition for eligibility is that the individual has been employed at least to some minimal degree in the past. In social insurance programs, even high-income individuals can receive benefits if they satisfy this criterion. Having said this, however, while eligibility for social insurance programs is not based on low income, the level of benefits paid to recipients are sometimes income-conditioned (i.e., reduced or eliminated if current income rises above certain levels), which creates work issues similar to those arising for welfare programs.

Although the common element of all welfare programs is that they are means-tested, they vary tremendously in other dimensions both within countries and across countries. Some are broad-based, offering benefits to all or at least large swaths of the low-income population, while others are targeted on children, the disabled, the elderly, or other groups, which are judged to be particularly deserving. Some programs are offer 'in-kind' benefits, meaning that specific consumption items such as food, housing, or medical care are provided, as opposed to simple cash payments. Another important distinction is whether the programs are operated through a country's social welfare agency or through its tax agency. While most programs are operated through social welfare agencies, provisions in the tax code, which reduce tax rates or tax burdens specifically for families with low income, are functionally equivalent to a welfare program, and should be considered as such.

Programs in different countries also vary dramatically in their benefit levels and whether they impose behavioral requirements. Table 1 shows that the benefit level in social assistance programs in selected OECD countries ranges from 5% of the average wage in the U.S. to 31.5% in Denmark and Ireland for the head of household, with similar variation for spouses and children. Most countries have some sort of job search requirement and more than half have some kind of work requirement (these requirements will be discussed further below).

This article will discuss how welfare programs relate to work. Since welfare programs provide benefits to those with low income, it is necessarily the case that program recipients will have lower levels of employment and earnings than individuals in the general population. Ideally, governments should address the problem of low employment and earnings of welfare recipients directly by devising education, training, and other programs that increase their work and income levels and hence reduce or eliminate the need for individuals to receive benefits in the first place. However, this is very difficult to do, especially in the short run. Consequently, most programs aim

Table 1 Social assistance programs in selected OECD countries, 2010

	Benefit level as a percent of average wage			*Behavioral requirements*	
	Household	*Spouse*	*Per child*	*Job search*	*Work*
Austria	13.4	8.9	4.3	No	No
Belgium	20.1	6.7	3.5–8.5	No	No
Canada[a]	15.7	11.4	1.3–1.4	Yes	Yes
Denmark	31.5	31.5	10.0	Yes	No
France	15.8	7.9	4.7–6.3	Yes	No
Germany	10.2	9.1	6.1–8.1	Yes	Yes
Ireland	31.5	20.9	4.8	Yes	Yes
Netherlands	33.1	10.0	N.A.	Yes	Sometimes
Norway	13.2	8.7	5.0–8.4	Yes	Yes
Spain[b]	18.2	5.5	3.6	Yes	Yes
Sweden	12.0	7.7	5.8–9.6	Yes	Yes
United Kingdom	9.9	5.7	2.6	Yes	Yes
United States	5.3	4.4	4.2	Yes	Yes

N.A. = not available.
[a]Ontario only.
[b]Madrid only.
Source: http://www.oecd.org/els/benefitsandwagespolicies.htm (accessed 02.03.14.).

instead simply to provide some kind of benefit or support to those with low-employment levels in order to raise their income and consumption to some minimal levels. When doing so, however, there is an incentive for individuals to reduce their work levels to make it more likely that they are eligible for the program or to increase the benefit payments they receive if already eligible. This trade-off between offering needed assistance to low-income families and discouraging them from working has been the central theme of much of the research on this topic.

The discussion below will first review the theory and evidence on this classic trade-off and how the use of financial incentives can ameliorate the trade-off. Following that, the discussion will turn to programs such as job search requirements, work requirements, and active labor market policy programs that attempt to ameliorate the trade-off through nonfinancial methods. A different type of financial incentive, through the use of earnings and wage subsidies, is discussed in the next section. The last section discusses dynamic incentives for human capital formation and related behaviors.

Work Disincentives: Theory and Evidence for Classic Open-Ended Programs

The eligibility and benefit determination for a 'classic' open-ended welfare program – that is, one that simply pays benefits according to the level of income and has no other restrictions or conditions on eligibility or benefit receipt – is illustrated in Figure 1. Letting the monthly benefit be noted by B, the formula for the benefit is $B = G - tE$, where G is the so-called guarantee level, E is the level of the family's earnings, and t is a taper rate (also called a tax rate or benefit–reduction rate). It is assumed that the family has no unearned income. The guarantee is the level of the benefit when a family has no earnings, and hence family income is guaranteed to be at least this high. The horizontal axis in Figure 1 shows the level of earnings (E) and the vertical axis shows the sum of E and B, or total income. When earnings are zero, total income is G. As earnings rise, B is reduced by t for every additional dollar of earnings along line AD. When earnings rises to E*, the benefit falls to zero and the family is no longer receiving benefits.

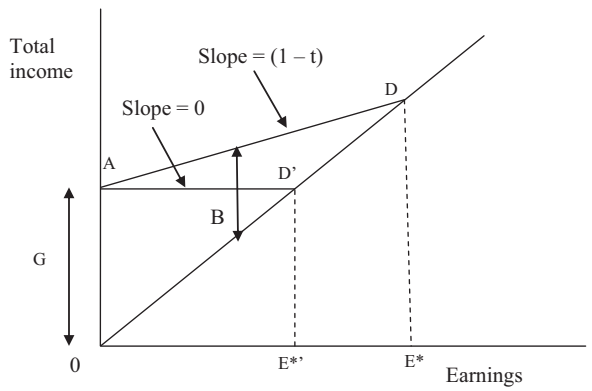

Figure 1 Classic open-ended welfare program.

This program increases family income but also creates work disincentives. Some individuals in families may reduce their earnings to zero in order to receive the maximum benefit, G; others may reduce their earnings in order to receive a higher benefit, B, since B is inversely related to earnings; and some individuals may initially have earnings above E* and hence be ineligible for any benefit at all but may reduce earnings to some level below E* to receive a benefit.

These work disincentives are affected by the level of G and t. A higher G will increase all the work disincentives just mentioned. A higher t will have ambiguous effects on total work disincentives, however. Line AD′ shows the benefit formula when $t = 1.0$, or what is called a 100% taper rate program. Intuitively, one would expect a higher taper rate to increase work disincentives because individuals who increase their earnings now have their benefits reduced dollar-for-dollar, and hence their total income is no higher than before. While this is correct, it is also true that some individuals – specifically, those who were working between E* and E*′ – are no longer eligible for the program and they will increase their work levels and earnings when they leave the program. Some of those with earnings below E*′ will also choose to go off the program and work and earn more, as well, instead of accepting a much lower income by staying on the program. The net work incentives of changing t are ambiguous in sign – they may reduce or increase work effort and earnings (Gruber, 2007; Moffitt, 2003). The work-discouraging effect is most heavily associated with the work of Milton Friedman (1962), who emphasized the first effect, but his analysis ignored the second effect.

There are many variations on this simple benefit schedule across programs and countries. Sometimes the program allows a 'disregard,' or 'exemption' level of earnings, so that benefits do not fall at all for some quantity of earnings. Sometimes a separate upper limit on earnings for eligibility is set below E*, which implies that benefits suddenly are eliminated if earnings go just above that, creating what is called a 'notch' in the benefit schedule. While all of these variations affect the magnitude of work disincentives as well as whom they affect (e.g., individuals with zero income versus those with higher income), the fact that benefits must be reduced, on average, as income rises and that they must fall to zero beyond some income level, means that some work disincentives will remain.

Even if the taper rate is set at a low level, in the hope that, on net, work disincentives will be reduced, the government has to be mindful of the fact that most countries have multiple welfare programs to serve different purposes. It is possible for a family to receive benefits from more than one, e.g., from cash welfare programs and housing subsidy programs. In this case, the total, or 'cumulative' taper rate is the sum of the taper rates of all programs the family receives benefits from (assuming the programs do not count the benefits of other programs as income, which makes the total taper rate not exactly the sum of the individual program taper rates). Even if the individual taper rates are low, therefore, the total taper rate can be high. If the family also has to pay positive taxes in the form of payroll taxes or income taxes, the total taper rate is even higher and possibly even above 100%.

The fact that open-ended welfare programs almost inevitably have some work disincentives does not mean they are not

socially desirable. Rather, they are an unfortunate side effect of accomplishing the socially desirable goal of providing financial support to needy persons and families. Since the goal of welfare programs is to provide assistance to those in need, the existence of work disincentives implies that some of those are not as needy as they appear, for they have reduced their incomes voluntarily because of the presence of the program. Nevertheless, this social cost may be outweighed by the social benefit of increasing incomes.

A relevant phenomenon that has been recognized in the research on welfare programs is that some families who are financially eligible for support (e.g., their incomes are below E*) do not participate and receive benefits. This is often described as a 'take-up rate' that is less than 100%. Why some families do not receive benefits even if they could has been the subject of much discussion. Moffitt (1983) supposed that it might be a function of 'stigma,' meaning a desire not to be a welfare recipient, even if it were to provide an increase in income. However, other research has revealed that nonparticipation can be a result of lack of information about eligibility, or because the time, money, and hassle costs of having to spend time with the welfare bureaucracy makes the benefit not worth applying for (Currie, 2006). The relative importance of different factors affecting the take-up rate are not fully understood.

Average effective tax rates facing low-wage individuals are very high in OECD countries (OECD, 2013). For example, a lone mother with two children in 2011 whose wage is one-third of the average wage in her country faced tax rates ranging from 5% in the U.S. to 127% in Denmark, with an average of 62% across all OECD countries. Tax rates for married couples and for those with different numbers of children were equally high and only slightly smaller for those with higher wages. This is important for the empirical evidence on the work disincentives of open-ended welfare programs indicates that those disincentives do exist. One of the more reliable pieces of evidence comes from a series of experiments in the U.S. which randomized families into groups, which either had no welfare program or offered a welfare program like that in **Figure 1**. The findings indicated that prime-age husbands reduced their work effort by 7% and wives reduced theirs by 17% (Burtless, 1987). Single mothers also reduced their work effort by 17%. Most of the reduction took place by extensions of spells of nonemployment; relatively little took place by reducing hours of work while still working, not surprising given that it is often difficult to adjust hours of work on most jobs or to adjust hours by changing jobs. The findings also showed that increases in benefit levels increased the work disincentives but that increases in the taper rate had very little net effect on work disincentives, with the positive and negative effects canceling each other out.

These findings are consistent in nature, although not always in magnitude, by nonexperimental studies of individual programs in different countries, which resemble that in **Figure 1**, although often having rules differing somewhat from the classic type. The U.S. evidence indicates that welfare programs reduce work effort by 10–50% (Moffitt, 2003: p. 2418). Many similar estimates have been generated in other OECD countries (e.g., Gurgand and Margolis, 2008).

Welfare Program Features to Reduce Work Disincentives

A number of features of welfare programs have been introduced to reduce the magnitude of work disincentives. These include work requirements, job search requirements, the threat of sanctions, and general forms of activity requirements.

In its purest form, a work requirement forces benefit recipients to work some minimum number of hours in order to receive benefits. For an individual with a particular job or wage rate, this implies that benefits cannot be received if the individual is not working at all or has earnings below some cutoff point. The program in **Figure 1** has to be altered by eliminating benefits at all levels of earnings below whatever minimum cutoff the government sets. The effect of work requirements on work disincentives is to unambiguously reduce those disincentives, for individuals who would have worked below the minimum amount will increase their level of work effort as a result. However, the increases can happen in one or two different ways: the individual can either increase his or her work effort to achieve the minimum level and therefore qualify for benefits, or the individual can choose not to do that and not to receive benefits at all. In the latter case, while some individuals who are not working will simply continue not to work, consequently resulting in no change in work effort, some will work more in order to make up for the lack of welfare benefits, thereby increasing work effort. The likelihood that an individual will choose not to receive benefits and to work less than the minimum is correlated with their job qualifications, skill, and wage rate: those with more job skills will more likely comply and receive benefits, while those with lower skills and wage rates will be less likely to comply.

The empirical evidence on work requirements is scanty because they have rarely been implemented in the rather severe form being discussed here. Perhaps the most well-known welfare reform which imposed significant work requirements on a welfare program is that which occurred in the U.S. in 1996. Studies of the evaluations of the effect of that reform show that work requirements increased employment rates by about 10% (Grogger and Karoly, 2005: p. 139). However, much of this effect was achieved by causing women to leave welfare and not receive benefits at all.

Job search requirements can be thought of as a milder form of work requirement. In this case, the benefit recipient is required to engage in job search but not to actually hold a job, if no suitable job can be found. Such a program requires a set of rules specifying how job search has to be conducted. Sometimes the requirement specifies that recipients have to register with the government unemployment agency and participate in looking at job vacancies at the agency on some periodic basis. The requirement may state that some minimum number of employers be contacted every week or month. In other countries and situations, a job search requirement may state that benefit recipients have to attend special sessions at which they are taught job-seeking skills, possibly in addition to contacting employers. Job search requirements can also be combined in sequence with work requirements, by requiring the recipient to search for a job for several weeks but by then also requiring that the individual actually finds a job at the end of that period or risk loss of benefits.

The empirical evidence on how the imposition of job search requirements per se is will affect work levels again rather scanty, despite the fact that it is more common than pure work requirements. Gueron and Pauley (1991) reviewed a number of U.S. job search requirement programs for welfare recipients and found them to have positive, but small, effects on work effort. Job search requirements have also been studied in Unemployment Benefit programs, which have somewhat different populations than do welfare programs (namely, not as disadvantaged and low skill). Nevertheless, all Unemployment Benefit programs require job search by the unemployed individual and there is some evidence that they shorten the length of time spent unemployed (Johnson and Klepinger, 1994).

Sanctions are defined by benefit reductions that are assessed on benefit recipients for not complying with work requirements or job search requirements (or activity requirements, to be discussed next). The sanctions can be partial or full – that is, benefits can be reduced but not eliminated for noncompliance, or the individual and family can be completely discharged from the program. In some cases, sanctions are increased, starting with some benefit reductions for initial noncompliance but then increasing in magnitude if noncompliance persists, possibly eventually resulting in complete benefit elimination Noncompliance means that the work requirements or job search requirements discussed above are not met, depending on the nature of the program. This can therefore mean not obtaining a job (in the case of work requirements), or not conducting a sufficient level of job search, or even not showing up for a special session to teach job search skills.

Theoretically, sanctions should clearly increase work effort relative to not having any sanctions at all. Evidence on this effect is difficult to obtain because all programs impose at least some sanctions, even if mild. The evidence that exists measures instead the effect of increasing or decreasing the level of sanctions. Studies in Western Europe show that sanctions in the unemployment insurance do speed up the transition from unemployment to work (Abbring et al., 2005).

Activity requirement is a broad term that is used to describe any type of requirement that states that the benefit recipient must spend a minimal length of time in some kind of activity. Work requirements and job search requirements and assistance therefore fall into this category, but the category also includes programs in which the individual is required to participate in an educational or training program. These latter programs have a rather different purpose, which is to increase the individual's level of human capital and future earnings. Training programs are, in fact, very common across industrialized countries, although they take many different forms, both informal and formal classroom training. Also included in the active labor market program category are subsidized public and private sector employment and, sometimes, sanction policies. These policies are common in Europe and evaluations of the various types indicate that subsidized public sector employment has little effect if not a detrimental one, while job search assistance and sanction policies have positive ones. Training programs typically have smaller positive effects and few positive effects on youth (Card et al., 2010; Kluve, 2010).

Earnings and Wage Subsidies

A more direct way to increase levels of work among welfare recipients is to actually reverse the problem induced by the taper rate by making the taper rate negative. A negative taper rate is an indirect way of saying that the benefit a recipient receives increases, not decreases, as the individual works more. However, since the program is still a welfare program, meaning that benefits are not provided to those with high income or earnings, benefits must eventually be reduced as income rises. But it is still possible to have a negative taper rate at lower levels of income.

If the benefit schedule is based directly on earnings, we have what is known as an earnings subsidy, as illustrated in **Figure 2**. If the subsidy is calculated as some fraction, s, of earnings, then every extra dollar of earnings raises total income by $(1 + s)$. In the diagram, the subsidy begins to be phased out when earnings reach E1*. The taper rate is labeled as t, which is the fraction by which the subsidy is reduced for each additional dollar of earnings beyond E1*. The subsidy reaches zero when earnings reach E2*.

A rather different program, called a wage subsidy, is one that subsidizes the hourly wage rate rather than earnings. If an individual is being paid $10 per hour, for example, and the wage subsidy is 10%, then the government would pay the individual $1 extra for every hour he works. The wage subsidy is phased out by decreasing the subsidy percent for those workers with higher wage rates, eventually eliminating it altogether at some wage rate picked as the eligibility point.

Most economists prefer a wage subsidy to an earnings subsidy because a wage subsidy does not ever have to be phased out for a given individual and hence there is no need to ever create the work disincentive that a phaseout induces. An individual with a low hourly wage continues to receive a higher benefit if hours of work increase, with no cap. Because an individual with a low hourly wage will never become a high-income person no matter how many hours he works, there is no danger of subsidizing high-income individuals. The work disincentive in a wage subsidy program comes instead from discouraging individuals from taking jobs with higher hourly wages or discouraging individuals from undertaking human capital investments by reducing the net amount by which the wage increases after such investments. This is a less direct work

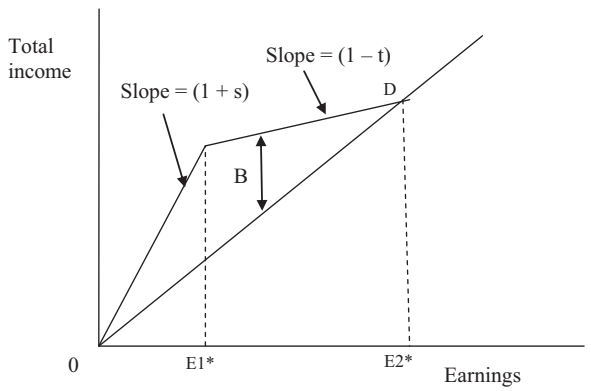

Figure 2 Earnings subsidy program.

disincentive than that created by the earnings phaseout that must be present in all earnings subsidy programs. However, wage rate programs are rarely found for practical reasons. Government measurement of the hourly wage on a job is not easy, especially for workers who are not paid by the hour. In addition, in most forms of wage subsidy, the government has to track the individual's wage job by job, rather than by simply measuring total earnings over some time interval, and this can be difficult administratively if the individual changes jobs frequently, as many low-wage individuals do. Finally, there is an incentive for employers and workers to misrepresent hours of work by paying the individual by the week or month and then overstating the actual hours worked, thereby lowering the measured wage rate and increasing the government subsidy.

The more commonly observed earnings subsidies come in many forms. In the U.S., the main earnings subsidy works through the tax system and provides tax rebates according to family earnings up to a certain level of earnings, then holds benefits fixed over a second interval, then reduces them, eventually to zero, over a third and final earnings interval. In the U.K., the earnings subsidy has a minimum hours-worked requirement, but operates the subsidy through the employer who has to certify that hours are above the minimum. Above that minimum, the subsidy is eventually reduced. An earnings subsidy of this kind will not pay benefits to individuals who work part-time at low hours, unlike the U.S. program, which pays earnings no matter how low hours of work are. The UK minimum-hours program bears some resemblance, in fact, to a pure work-requirement program, which also pays no benefits to individuals who do not work at least some minimum number of hours.

However, this raises a separate important issue, which is whether the earnings subsidy program is a complete replacement, or substitute, for a traditional welfare program or whether it is just a supplement to such a traditional program. In the U.S., for example, no traditional cash welfare program is available at all many individuals – two-parent families, for example, or single childless men and women, no matter low their income. For those types of individuals, an earnings subsidy is not really a welfare program at all because it provides no support to those with zero or very low incomes, whose are the neediest individuals measured by income. But in the UK and in the US for individuals eligible for traditional cash welfare benefits (such as single-mother families), the earnings subsidy is layered on top of the traditional welfare program. In this case, since the traditional welfare program has a positive taper rate and benefits are reduced as earnings rise, the earnings subsidy merely counteracts that effect and reduces the taper rate (e.g., if the welfare program taper rate is 50% and the earnings subsidy rate is 10%, the net taper rate is now 40%). It is possible for the earnings subsidy rate to be so large make the net taper rate negative. In addition, when the earnings subsidy is layered on top of a traditional welfare program, the taper rate is actually increased at higher levels of income. That is because the earnings subsidy must eventually be phased out and, when that happens, the individual faces not only a taper rate on the traditional program benefit but also the phaseout of the earnings subsidy (e.g., if the traditional program taper rate is 50% and the earnings subsidy is phased out at 10%, the net taper rate is now 60%). Thus, work disincentives actually rise for individuals with earnings in higher ranges.

The empirical evidence on the effects of earnings subsidies on work indicates that they generally do have some positive effects on employment. In their summary of the literature on the U.S. program, for example, Hotz and Scholz (2003) find positive effects on the employment rates of single mothers but no effects on that of prime-age males and perhaps a small negative effect on the hours worked of married women. Findings in the UK are very similar, with positive effects on employment for lone mothers (Francesconi and Van der Klaauw, 2007) but small or negligible effects on married men and women as a whole, although married women married to men working low hours experienced positive effects (Francesconi et al., 2009).

Dynamic and Human Capital Effects

Dynamic work effects of welfare programs occur when the individual is forward-looking enough to anticipate work levels in future periods of life (future months, years, etc.) and alters his or her current work effort because of this. One simple effect that can occur is if the individual realizes that he or she can alternate periods of not working and being on the program with periods of working and being off the program in order to achieve a higher level of benefits than he or she could if he or she worked constantly halfway in between. In **Figure 1**, an individual who worked half the time above point E^* and did not work at all half the time would receive average benefits, which are above the schedule if the individual worked all the time halfway in between. This provides an incentive for individuals to cycle on and off the program. Another effect can occur if the individual is able and willing to save any of his or her welfare benefit for future periods when he or she is off welfare (this may be constrained by asset tests in the welfare program, which eliminate eligibility if the recipient has too much money in the bank). In this case, the individual may wish to reduce current consumption by not spending all of the welfare benefit and increase future consumption by saving some of it for later periods when he or she is off welfare. This extra income in later periods may reduce work levels even in those periods. A third dynamic effect can occur if an individual knows that his or her wage rate, or job opportunities more generally, are going to be different in the future than they are currently. If so, the individual may work more in the periods when job opportunities are high, and stay off welfare in those periods, and work less in those periods when job opportunities are low and stay on welfare (in the literature, this is known as 'intertemporal substitution'). This would therefore increase welfare program participation and lower work levels when on welfare. Some effect of this kind might also occur if a welfare recipient's job skills and wage rate increase, on average, over the lifetime (whether this happens for very low-skilled, less-educated individuals is an open question). If so, the individual may intentionally lower levels of work while young and receive welfare benefits, planning to go off welfare and work later when job opportunities are better.

A different set of effects occurs if the welfare program has time limits (the U.S. is the leading example of such programs,

although Unemployment Benefit programs in all industrialized countries are somewhat similar). If the individual is not forward-looking, then he will simply remain on the welfare program until benefits run out and benefits will then fall to zero. However, if he or she anticipates the approach of the time limit, the individual may go off the program and work in order to 'save' the remaining periods of eligibility for periods later in life when job opportunities are particularly bad or when the individual is laid off from his or her job. By 'saving for a rainy day,' work effort is increased in current periods when the individual goes off the program and decreased later when he or she returns to the program.

Human capital effects occur if the individuals consider making investments in their skills either by attending school, undertaking training programs, or simply working in the hopes of raising future wage rates and job opportunities by acquiring work experience and seniority. Welfare programs have two different effects on the incentives to invest in human capital activities that take time (education, training). On the one hand, if the individual anticipates remaining on welfare even after the investment is complete, the 'return' to the investment is lowered because any higher future earnings result in lower future benefits, the amount depending on the taper rate, t. On the other hand, the opportunity cost of taking time off from work in order to attend school or a training program is also reduced by the welfare program – the amount of lost earnings is smaller because benefits would be reduced if the individual were to work instead of devoting time to human capital investment. The net effect on human capital investment is therefore ambiguous in sign.

A less ambiguous effect of welfare programs on human capital investment occurs if the individual is credit-constrained and cannot borrow for the direct costs of education or training or does not have alternative sources of income to support consumption while attending school or training. In this case, the welfare benefit provides, in effect, credit relief and allows the individual to undertake the investment and still have enough money to have at least minimal levels of consumption.

The effects on human capital investment that occur simply by working and obtaining more work experience and seniority (so-called 'learning by doing') are somewhat different, since that type of investment does not require time off from work, as education and training do. The positive taper rate in this case unambiguously reduces the return to work and hence reduces the incentive to work in order to increase future human capital as well. An earnings subsidy does the opposite.

The empirical evidence on dynamic effects is minimal. Keane and Wolpin (2010) estimate a life cycle model with welfare participation but focus more on the effects on the marriage market and racial differences. Sanders and Miller and Sanders (1997) investigate the effect of welfare on human capital accumulation, finding that many individuals do not go onto welfare because it will reduce their future human capital accumulation. There have been a small number of studies of the effect of time limits on labor supply as well, with mixed findings.

See also: Welfare Policy Comparisons; Welfare Reform; Welfare Retrenchment; Welfare State, History of; Welfare.

Bibliography

Abbring, J., van den Berg, G., van Ours, J., 2005. The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. Economic Journal 115, 602–630.

Burtless, G., 1987. The work response to a guaranteed income: a survey of the experimental evidence. In: Munnell, A. (Ed.), Lessons from the Income Maintenance Experiments. Federal Reserve Bank of Boston and Brookings Institution, Boston and Washington, pp. 22–52.

Card, D., Kluve, J., Weber, A., 2010. Active labour market policy evaluations: a meta-analysis. Economic Journal 120, F452–F777.

Currie, J., 2006. The take-up of social benefits. In: Auerbach, A., Card, D., Quigley, J. (Eds.), Public Policy and the Income Distribution. Russell Sage Foundation, New York, pp. 80–148.

Francesconi, M., van der Klaauw, W., 2007. The socioeconomic consequences of 'in-work' benefit reform for British lone mothers. Journal of Human Resources 42, 1–31.

Francesconi, M., Rainer, H., van der Klaauw, W., 2009. The effects of in-work benefit reform in Britain on couples: theory and evidence. Economic Journal 119, F66–F100.

Friedman, M., 1962. Capitalism and Freedom. University of Chicago Press, Chicago.

Grogger, J., Karoly, L., 2005. Welfare Reform: Effects of a Decade of Change. Harvard University Press, Cambridge.

Gruber, J., 2007. Public Finance and Public Policy, second ed. Worth Publishers, New York.

Gueron, J., Pauly, E., 1991. From Welfare to Work. Russell Sage Foundation, New York.

Gurgand, M., Margolis, D., 2008. Does work pay in France? Monetary incentives, hours constraints, and the guaranteed minimum income. Journal of Public Economics 92, 1669–1697.

Hotz, V., Scholz, J., 2003. The earned income tax credit. In: Moffitt, R. (Ed.), Means-Tested Transfers in the U.S. University of Chicago Press, Chicago, pp. 141–197.

Johnson, T., Klepinger, D., 1994. Experimental evidence on unemployment insurance work-search policies. Journal of Human Resources 29, 695–717.

Keane, M., Wolpin, K., 2010. The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, hispanic, and white women. International Economic Review 51, 851–892.

Kluve, J., 2010. The effectiveness of European active labour market programs. Labour Economics 17, 904–918.

Miller, R., Sanders, S., 1997. Human capital development and welfare participation. Carnegie-Rochester Conference Series on Public Policy 46, 1–43.

Moffitt, R., 2003. Welfare programs and labor supply. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. Elsevier, New York, pp. 2393–2430.

Moffitt, R., 1983. An economic model of welfare stigma. American Economic Review 73, 1023–1035.

OECD, 2013. Work Incentives: Inactivity Traps, 2001–2011. http://www.oecd.org/els/benefitsandwagesstatistics.htm (accessed 02.03.14.).

Welfare Economics

Yew-Kwang Ng, Nanyang Technological University, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In this article, the conditions for Pareto optimality are outlined; why the first theorem in welfare economics is true is explained. Despite the required conditions being unrealistic, the first theorem is very important as it shows the working of the market mechanism and serves as a benchmark, allowing us to examine divergences caused by, say, external costs. Coase's argument against the Pigovian taxation of external costs, and of pollution, is based on all-or-nothing comparison ignoring an important asymmetry. Arrow's impossibility theorem and a 90% solution to the paradox of interpersonal cardinal utility are discussed.

What Is Welfare Economics? Preference versus Welfare

Welfare economics is the branch of study that formulates propositions by which we can say that the social welfare in one economic situation is higher or lower than in another. What is social welfare? It is the welfare of individuals in the relevant society. What is individual welfare? The welfare or well-being of an individual may be regarded as her net happiness. Although we typically use the term happiness for immediate feelings and welfare for longer-term well-being, these two are essentially the same for any given time period.

There is a distinction between welfare and utility or preference (*see* Welfare: Philosophical Aspects): in older times, economists used the term utility more in the sense of subjective satisfaction, closer to welfare. However, modern economists use utility just as a representation of preference. In this sense, the utility or preference of an individual may clearly differ from her welfare or happiness due to the following three factors:

1. An individual may ex-ante prefer or choose A over B but her ex-post welfare may turn out to be lower in A than in B due to ignorance or imperfect foresight. Harsanyi (1997) distinguished between actual preference and informed preference.
2. Although probably not accepted by the majority of economists, an individual may have non-affective altruism (Ng, 1999) and prefers A over B due to her concern for the welfare of others over and above any warm-glow effects (which is called affective altruism) that also improves her own welfare (*see* Charity and Philanthropy, Economics of).
3. Ignoring nonhuman welfare and regarding any divergence between preference and welfare other than ignorance and non-affective altruism as irrational, these three factors present an exhaustive distinction between preference and welfare.

Traditional welfare economics mainly discusses preference or utility; although much study has been done on happiness (Diener et al. 2010), attempts to deal with welfare/happiness have not advanced very far (e.g., Kahneman et al., 1997; Ng, 2003).

First and Second Theorems in Welfare Economics

The first theorem in welfare economics, a if not the most important result, states that under certain conditions, a general equilibrium of a market economy under perfect competition results in a most efficient allocation both in resource, inputs, or factors of production, and in final goods (taken to include services). Efficiency here is defined in the sense of Pareto optimality where no one can be made better off without making any other worse off. Perfect competition requires that all sellers and buyers have no influence on prices and take market prices as given. Apart from this (and the implied absence, at equilibrium, of increasing returns; see Ng, 2009), the main additional conditions are perfect relevant information and the absence of real external effects. A consumer needs to know that she can buy the various goods at the given market prices, and her preference for them; a producer of a certain good needs to know the best available method (i.e., the least costly method at the market prices of the inputs) of production. In contrast to real external effects like pollution as discussed in Section Some Welfare Economic Issues of Environmental Protection, the presence of pecuniary external effects does not cause any inefficiency, at least not under the classical conditions of perfect competition (and absence of distortions). Pecuniary external effects are those where the actions of someone affect others through the effect on the market prices. For example, the higher demand for good x by consumer A may increase its price and hence make consumer B (who also likes to consume good X) worse off. This does not cause inefficiency under perfect competition as the loss in consumer surplus by B is offset by the gain in producer surplus for the supplier of x. If conditions for the first theorem are satisfied, the situation is Pareto optimal both before and after the change in preference by A.

The condition of perfect relevant information may be significantly violated for certain goods where many consumers are ignorant of the quality of goods. This explains the need for certain government regulations, on food safety in particular. This is partly because the undesirable effects may not be clear and immediate, and are difficult to be attributed to particular goods. The presence of external effects like pollution is even more important and is discussed in Section Some Welfare Economic Issues of Environmental Protection.

The assumptions of perfect competition, no ignorance, and no pollution are very unrealistic; why else is the first welfare theorem important? This theorem may be misleading if not interpreted carefully. In fact, at least to some extent, the theorem probably has been interpreted incorrectly to support virtually 100% free-marketism (the case against the Pigovian

tax on pollution discussed in Section Some Welfare Economic Issues of Environmental Protection is probably related to this misinterpretation). However, if interpreted properly, the first welfare theorem is one of the most important results in economics. On the one hand, it allows us to focus on the essence of the price mechanism or market coordination in achieving efficiency in resource allocation, production, and distribution of products. On the other hand, it serves as a benchmark to identify the possible sources of inefficiency (e.g., monopoly, pollution, and consumer ignorance) and hence device possible corrective measures (e.g., eliminating administrative monopoly, taxing pollution, and appropriate food safety regulations) to supplement the market.

Why Is the First Theorem True?

To appreciate the essential functioning of the market mechanism, let us consider why the first theorem is true under its assumptions. Before this, consider first the necessary conditions for Pareto optimality. Given the amount of resources and amount of labor, we want to allocate these factors for production of various goods and services in an efficient way, to reach the production possibility frontier so that more output of one good is not possible without reducing that of another. This requires that the marginal rate of technical substitution (MRTS) between any two inputs (factors of production) used in the production of any good X must be the same as the MRTS between the *same* pair of inputs used in any other good, for example good Y. If the MRTS between any given pair of inputs (say, inputs A and B) are different in the production of X and Y, we can reallocate A and B so as to increase the output of either X or Y, or both. Suppose one unit of A can substitute for 3 units of B in the production of X but one unit of A can substitute for only 2 units of B in the production of Y. This suggests that A is relatively (in comparison to B) more productive in X than in Y. If we transfer one unit of A from the production of Y to that of X, the extra unit of A can release 3 units of B from X (while holding output unchanged), but we only need to use 2 units of B for Y to offset the loss of one unit of A there. Thus, we maintain the output levels of both X and Y unchanged, but with one unit of B left unused. Using it in the production of either X or Y, or both may thus increase the output.

The simple point made above may be shown graphically by inverting the isoquant map (not shown; with the amounts of A and B used in Y on the two axes) of Y by 180°, locating the origin diagonally north-east of the origin of the isoquant map of X; the width and height of this Edgeworth box represent the amounts of A and B available, respectively. The tangency points of the two sets of isoquants trace out the locus of productively efficient allocation of the two inputs between X and Y. Along this locus, the slope of the isoquant for X (MRTS between A and B in X) is equal to that in Y. This condition for productive efficiency is $MRTS^X_{AB} = MRTS^Y_{AB}$ for all inputs A and B used in the production of any goods X and Y.

Once goods are produced, we need to distribute them to various consumers efficiently. This condition for efficiency in exchange requires that the marginal rate of subjective substitution (MRS) between any two goods be the same for all consumers consuming this pair of goods, that is, $MRS^J_{XY} = MRS^K_{XY}$ for any two consumers J and K. The reason for this condition is similar to that for productive efficiency. If one unit of X can substitute for 3 units of Y for consumer J while it can substitute for only 2 units of Y for consumer K, transferring X from K to J, compensated by transferring Y from J to K could make both consumers better off.

If we satisfy both conditions (one for productive efficiency and the other for exchange/distributive efficiency), then will it be sufficient to ensure Pareto optimality? No, even if we ignore or assume the satisfaction of the second-order conditions (i.e., ensure a maximum rather than a minimum point), productive efficiency ensures no misallocation of inputs in production; exchange efficiency ensures no misallocation/distribution of the given produced goods to consumers. This is not sufficient for Pareto optimality because the given amounts of goods produced may not be consistent with the preferences of consumers. For simplicity, consider only two goods, food and clothing. We may efficiently produce an excessive amount of clothing and too little or no food; we have a lot to wear but go hungry. In contrast, we may have a lot to eat but go naked! This is far from efficient. To ensure that the amounts of the various goods produced are consistent with the preferences of consumers, we need the top-level optimality condition requiring the equality of the MRS between any pair of goods to be equal to the marginal rate of transformation (MRT, the rate at which one unit of a good may be transformed into another by reallocating inputs in the production): $MRS_{XY} = MRT_{XY}$.

Now consider how the invisible hand of the market mechanism achieves Pareto optimality. This depends on the assumed profit-maximization behavior of producers/firms and the utility-maximization behavior of consumers/individuals. Under the assumptions of perfect competition, no (relevant) ignorance, and no external effects, the maximization actions satisfy all the above three conditions for the Pareto efficiency (assuming the satisfaction of the second-order conditions like convexity of preferences or indifference curves curving the right way, and so on), as explained below.

Under perfect competition, all economic actors (consumers, producers, and input suppliers) have, each acting on her own, no appreciable influence on prices. Ignoring collusion (the presence of which violates the requirement of perfect competition), each takes prices as given (beyond one's control) and maximizes accordingly. Any consumer thus views her budget line as linear and maximizes utility by consuming at the point where the given linear budget line is tangent to her highest indifference curve (ignoring the goods she does not consume). As the slope of the indifference curve is her MRS between the relevant pair of goods and the slope of the budget equals to the price ratio of this pair of goods, we have, at the utility-maximization equilibrium point, $MRS_{XY} = P_X/P_Y$. As the same set of prices is faced by all consumers, and as all consumers who consume this pair of goods satisfy this same equation, the MRS of all consumers of x and y are equal to the same P_X/P_Y, and hence are also all equalized, satisfying the exchange condition as mentioned above.

Taking input prices as given, any producer of any good views its isocost curves/lines as linear (just like the budget line of a consumer). To maximize profit (which equals total revenue minus total cost), he must minimize the total cost of producing any given amount of output (which yields a given

amount of total revenue). Thus, for any output chosen, he chooses an input combination at a point where an isocost line is tangent to that isoquant. This means that the MRTS between this pair of inputs (given by the slope of the isoquant) is equal to the ratio of input prices (the slope of the isocost line). Because all producers of any good are faced with the same set of input prices, we thus have the MRTS between any pair of inputs used in the production of any good being equalized to their input price ratio, and hence equal to each other, satisfying the condition for the productive efficiency.

To maximize profit, a firm produces until its marginal cost (MC) curve just intersects its marginal revenue (MR) curve from below. The MR of a perfectly competitive firm is its horizontal price line or the demand curve for its product. Profit maximization thus entails price = MC. As this is true for all goods, we have $MC_X/MC_Y = P_X/P_Y$ for any pair of goods X and Y. In the absence of external effects, $MC_X/MC_Y = MRT_{XY}$. This is so because if we decrease the production of X by one marginal unit (tiny; the equality here is only approximately true for a large unit), we save inputs costing an amount of MC_X. If we use this amount of inputs saved for additional production of Y, then we can increase the output of Y by an amount that equals MC_X/MC_Y, making this the MRT_{XY}. From the above-mentioned two equations, we thus have $MRT_{XY} = P_X/P_Y$. As the exchange condition above already gives us $MRS_{XY} = P_X/P_Y$, we have $MRS_{XY} = MRT_{XY}$, which is the satisfaction of the top-level condition for the Pareto optimality.

Thus, we see that all the three conditions for the Pareto optimality are satisfied under the conditions stated. In addition to the absence of price-influencing effects and external effects, the exchange condition is satisfied by the utility maximization actions of consumers, and the production condition is satisfied by the cost minimization actions of producers. The top-level condition is satisfied by the interaction of utility maximization of consumers and profit maximization (including cost minimization for given output) of producers, coordinated through the price mechanism. In understanding how the three conditions for Pareto optimality are satisfied, we also understand how the market coordinates resource allocation to produce various goods to satisfy the preferences of consumers efficiently.

Second- and Third-Best Theories

Pareto optimality requires the satisfaction of all the three conditions in the production of all goods. If some of the conditions cannot be satisfied for some goods, is it better to satisfy as many others as possible? The theory of second best (Lipsey and Lancaster, 1956) says that it is, in general, not true. Given the violation of conditions for some goods as additional constraints, in general it is best to deviate from the conditions for all goods (except by coincidence)! This makes piecemeal welfare policies virtually impossible.

Suppose factory A produces good A with huge pollution. The first-best solution is to tax this pollution (see Section Some Welfare Economic Issues of Environmental Protection) according to the marginal damage of pollution. Suppose this factory is owned by a politically powerful person and no one dare tax A. Given this as a constraint, it might be better to deviate from the first-best conditions in all goods, for we may want to encourage goods that are very substitutary (in consumption and/or production) to A and to discourage goods that are very complementary to A to indirectly affect A to partially reduce pollution. This is the intuition of the second-best policies. However, the second-best policies require taking account of all the complicated interrelationships in consumption and production between all goods and inputs; these are informationally impossible and administratively very costly. Taking account of informational and administrative costs, second-best policies are neither feasible nor optimal. In the real world where second-best violations exist and informational and administrative costs are important, it is best to pursue third-best policies. Ng (1977, 2004a) shows that in the absence of sufficient information (Informational Poverty), the third-best policies are simply the same as first-best policies, like price = MC, or MRS = MRT. In the presence of sufficient information (i.e., informational scarcity), the third-best policies falls between the first-best and second-best policies, and account for more important interrelationships only. An example of such a third-best policy is subsidizing train travel (with less congestion and pollution per passenger) to reduce excessive car congestion and pollution, if the direct taxation of the latter is politically or technologically not feasible.

The Second Theorem

The first theorem refers only to efficiency, and it does not say anything about equality. A Pareto-optimal point (i.e., allocation of resources and goods) may involve the existence of extreme richness with extreme poverty. Could an efficient point with higher equality be possible? The second theorem says that any feasible Pareto-optimal point could be sustained as a perfectly competitive general equilibrium with the appropriate initial distribution of endowments (of resources, assets, and earning abilities, possibly including power and connections in a wider analysis). If we want an outcome of more equality, we need a more equal distribution of endowments. On the one hand, a Jackson ends up with a huge income/wealth level because he is endowed with a sexy voice and other talents, allowing him, with some learning/training, to attract thousands/millions of fans willing to pay highly to watch his performance. On the other hand, a Kwang ends up with only a small fraction of that wealth because when he sings, people run away!

If we start with very unequal endowments, could we improve upon the unequal outcomes to achieve higher equality? Yes, but under the traditional analysis, we typically have to sacrifice some efficiency. We may tax the rich more and help the poor. This does not only involve administrative, compliance (on the part of the taxpayers), and policing costs, but also disincentive effects, as it may make both the rich and the poor having less incentives to earn higher incomes. We may also add the costs of generating more rent-seeking activities, not to mention outright corruption. Thus, the second theorem has limited significance, but it does show us that the unequal outcome may largely be traceable to unequal endowments rather than the functioning of the market as such. Some recent studies have shown that some improvements in equality may actually be beneficial to efficiency. With the increase in within-country inequality over

the past few decades, there have been significant increases in public support for more redistribution (IMF, 2014).

The question of efficiency–equality tradeoff alerts us to the limit of Pareto optimality. Most changes make some people better off and some worse off. Therefore, we have to go beyond Pareto optimality, as discussed next.

Social Choice (*See* Public Choice Economics)

How does or should a society decide especially in the presence of conflict of interest? Arrow (1951/1963) asks whether we can have a reasonable rule that allows us to derive our social ranking/ordering of relevant alternatives/social-states from the individual orderings of them. The answer is 'no!' even if only a few very compelling conditions are required: individual freedom to rank, weak Pareto (if all individuals prefer x over y, the society must also so prefer), independence of irrelevant alternatives (ranking of x vs y not depending on z), and non-dictatorship (nonexistence of a single individual who dictates all choices).

This Arrow's impossibility theorem, although appearing to be surprising to many and generating a large amount of debate, may actually be explained quite simply. Since the social ordering has to be derived only from individual orderings (of social alternatives), it effectively ignores the intensities of preferences. A preference situation of:

x, y,……, z (Situation 1)

and another of:

x, ……, y, z (Situation 2)

are taken as exactly the same: the ordering being x above y which is in turn above z. The possible huge difference between the two situations with the intensity of preference of y over z being much higher than that of x over y under situation 1 and the reverse under situation 2 is completely ignored.

True, it is more difficult to obtain information regarding the intensity of preference than regarding just the ordering, especially if the intensities have to be interpersonally compared. Thus, Arrow's problem is not unimportant. If we could make do with easier obtained information, our life would be easier. But the answer is, again, 'no!'; we cannot just rely on individual orderings. The intuition of this impossibility is also clear. If one has a number of children and has to make collective decisions for the whole family, he obviously would not only want to know the ordinal ranking of the alternative options by his children, but also to know how strong their preferences are. Thus, Arrow's impossibility suggests that reasonable social choice should not only be based on individual orderings, but should also take into account the intensities of preferences or cardinal utilities. Moreover, since more than one person are involved, interpersonal comparisons of individual cardinal utilities are also necessary (Sen, 1970).

How do we solve this paradox/dilemma of interpersonal cardinal utilities (that interpersonally comparable cardinal utilities are needed, but they are difficult to obtain)? One option is to improve on the obtaining of the required information, including perhaps the further development of happiness studies (e.g., Diener et al., 2010; Ng, 2013). Before substantial advances on that front, perhaps we may solve the dilemma largely (90%?) by using the willingness-to-pay (information on which is relatively easier to obtain) as the measure of the intensity of preferences. However, this raises the interpersonal or equality issue. The willingness of a rich person to pay $10 000 may be much larger than that of a person of average income to pay $1000, but need not be so in terms of welfare or happiness. This problem is solved by treating a dollar as a dollar but using redistribution to address the equality issue, as it is more efficient to achieve any desired level of equality in real income (utility/welfare) by adopting a purely efficiency-oriented policy in specific issues, and do redistribution in the general tax-transfer system (Ng, 1984). This proposed solution is only about 90% workable since we still need interpersonal comparison of cardinal utilities in determining the optimal tradeoff between efficiency and equality in the general tax-transfer system. But this, at least, frees us from the informationally demanding need in all specific issues (*see* Cost–Benefit Analysis).

Attempts to bypass Arrow's theorem do not seem to be convincing. For example, Little (1952) and Samuelson (1967) argue that the theorem is only relevant to politics, not to welfare economics as the latter does not have to deal with changes in individual preferences. Kemp and Ng (1976) and Parks (1976) show that, even holding individual preferences unchanged, if we insist on using ordinal information only, impossibility results still hold, as generalized by Roberts (1980).

Even if we have interpersonally comparable cardinal utilities, we still have the problem of as to which specific a social welfare function (SWF) should be maximized. In particular, should the society maximize the classical utilitarian SWF of the unweighted (or equally-weighted) sum of individual utilities? Harsanyi (1953, 1955) and Ng (1975) provide compelling arguments for utilitarianism. Harsanyi uses the reasonable conceptual method of impersonal reasoning and reaches utilitarianism compellingly. This is more widely known as the 'veil of ignorance,' a term invented by Rawls (1971) who gave Harsanyi only an obscure footnote. Although he uses the same impersonal reasoning (Hare's rule of universalizability is on a similar principle; see Sen, 1966), Rawls reaches rather inextricably the virtually opposite SWF of maximizing the utility level of the worst-off person. Some people believe that this is based on the principle of extreme risk aversion. There are two important problems with this. First, although one may rationally be risk-averse with respect to some objective/instrumental variables like consumption/income levels due to diminishing marginal utility of consumption/income/wealth, she should not be risk-averse with respect to the ultimate objective like utility/welfare/happiness. Logically, one cannot have diminishing marginal utility of utility! Secondly, even if it were rational to have risk-aversion with respect to the ultimate objective, an extreme degree of risk-aversion is certainly not acceptable. Being willing to sacrifice enormously the welfare of all others including the second worst-off person for a marginal increase in the welfare of the worst-off cannot be rational. Its apparent attractiveness to many philosophers and economists may be explained by the confusion of nonultimate considerations with basic values (Ng, 1990). The extreme emphasis on those at the bottom

cannot be justified on their marginal increases in welfare alone, and must be explained by the indirect effects in the future through improving equality, harmony, etc. This should be achievable by a (even if purely utilitarian) SWF that takes into account effects on all, and also of the future.

Harsanyi asks us to imagine: Suppose you were an unincarnated soul, having 1% chance of becoming (both in objective circumstances and in subjective capabilities, including preferences and the capacity for enjoyment) each of the 100 individuals of a given society. You have to decide before incarnation what SWF would you like the society to maximize in making social choices? If you are rational, you should choose to maximize your expected utility/welfare by requiring the future society to maximize the utilitarian SWF of the equally-weighted sum of individual welfares. This maximizes your expected utility while you are still under the veil of ignorance. People with good moral intuition should find this reasoning compelling. Ng (1975) also argues that our SWF should be utilitarian, based mainly on finite sensibility with weak majority preference: If at least half of the individuals prefer x to y and no one prefers y to x, social welfare is higher in x (see also Economics and Ethics; and Mueller, 2003: Chapter 23).

Occasional pronouncements or lamentations of the death of welfare economics is based on the impossibility of interpersonal comparison of utility/welfare. Salles (2014: p. 8) is correct that "welfare economics is alive and well in multiple guises." Also, Atkinson (2011: p. 157) argues that "welfare economics should be restored to a prominent place on the agenda of economists and should occupy a central role in the teaching of economics."

Many issues related to social choice have been explored. For example, individuals may misrepresent their preferences; could we design voting/aggregation rules that are strategy-proofed? Difficulties abound see pioneering studies by Gibbard (1973), Pattanaik (1973), and Satterthwaite (1975) and a survey by Barberà (2011); on issues of distributive justice, see, for example, Fleurbaey (2008), Gotoh and Dumouchel (2009), Kaplow and Shavell (2002), Kolm (1997, 2005), Moulin (2003), Ng (2000), Roemer (1996), and Sen (2009); also articles Equality; Income Distribution and Inequality: Measurement Issues; on individual freedom and rights, see, Pattanaik and Xu (1990), Dowding and Van Hees (2009), Sen (2002), and Suzumura (2011). Although there are also many important issues that are closely related to welfare economics, only the most important issues are discussed in the following section.

Some Welfare Economic Issues of Environmental Protection (*See* Environmental Economics)

With the threat of global warming, environmental protection and sustainable development have become very important issues. It is widely believed that this is essentially the issue of intertemporal ethics. If we produce, pollute, use up resources, and consume more now, we leave less to the future generations. In fact, why sustainability may be threatened is not mainly due to the disregard of the present generations for the welfare of future generations, but the failure to avoid excessive environmental disruption now. This is essentially a question of I/we versus others; one of external costs, not of intertemporal ethics. In fact, the free market takes care of the intertemporal allocation rather efficiently partly due to the largely sufficient (if not excessive) care of most parents for their own children due to both biological instinct and education/learning. In addition, under sufficient property protection, an old man does not have to hastily exhaust the minerals under his land; he could sell it to younger men who have longer time to use the land more efficiently. In the absence of ignorance and other imperfections, resource utilization will be intertemporally efficient (Hotelling, 1931), as consistent with the first theorem discussed earlier.

In the presence of important external effects like greenhouse gases and other pollution, the first theorem no longer holds, and governmental intervention in the form of Pigovian taxes on external costs in accordance with their marginal damages may improve efficiency. However, Coase (1960), Cheung (2013), and others argue against the Pigovian taxation of external costs, or even the concept of externality as such. This position has been debated. It may be conceded that Coase's point on the two-sided nature of the problem is correct. Restricting pollution benefits the sufferers but affects the polluters. Secondly, if the two parties may negotiate without prohibitive costs to reach an efficient solution, no government intervention is necessary. The externality is 'internalized' through negotiation. We may even concede the so-called Coase's theorem that such an efficient outcome will be reached whether we start with polluter's rights or with amenity rights; in the absence of transaction costs (including game-strategic behavior preventing agreement), an efficient outcome will be reached. However, Coase's case against the Pigovian taxation of external costs, even in the presence of transaction costs preventing mutual agreement, has been argued (Ng, 2007, 2014 on which the following is based; really just textbook with a crucial epsilon) to be based on an unacceptable all-or-nothing comparison only.

The problem may be shown by using the textbook analysis of an external cost in **Figure 1**. The horizontal axis x measures the amount of some activity/pollution/emission that yields a (net) marginal value (MV) to the polluter/emitter measured by the downward-sloping curve AEP (which need not be linear). In the business-as-usual or no restriction option, the polluter undertakes x until the private optimal point P where

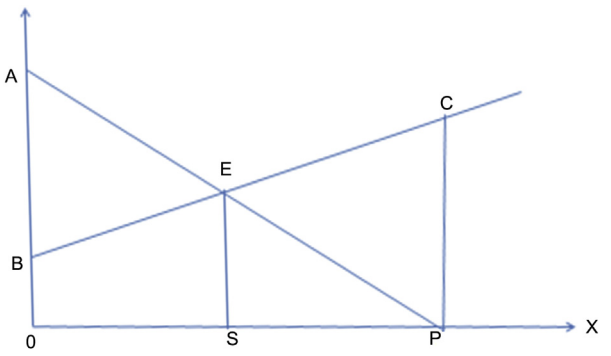

Figure 1 All-or-nothing vs continuous comparisons.

the net MV equals zero. If this activity x imposes no damages on anyone (and ignoring other problems like ignorance and the second best), this private optimal point is also the social optimal point. However, if it imposes damages on some others or the whole society represented by the marginal damage curve BEC, the social optimal point is where the two curves intersect at E where the level of x is at S. Given the damage imposed, point P is not efficient as the amount of pollution over the range SP yields benefits of only triangle ESP, but results in damages of a much bigger area ESPC. The Pigovian solution is to impose a tax on this external cost on the polluter to shift the marginal value curve from AEP downward to pass through point S, motivating the polluter to reduce the level of x from P to S.

Coase criticized the Pigovian tradition for ignoring the reciprocal nature of the problem and concentrated only on the restriction of the polluters' activity that harms the society. But this restriction affects the polluters. The problem is to avoid the more serious harm. A change of approach is necessary. As Coase did not use either maths or graphical analysis, he compared only the free-to-pollute situation P with the no-pollution situation at the origin (Coase, 1960: p. 41). For such an all-or-nothing comparison, Coase correctly concluded that either of the situations may be more efficient, and a case-by-case examination is necessary. In terms of **Figure 1**, this depends on whether the triangle ABE is larger or smaller than triangle EPC. As a general conclusion is not possible, the presumption against external-cost causing activities or in favor of restricting pollution is not warranted.

Although Coase's conclusion is valid for his all-or-nothing comparison, it is not valid against Pigou. Pigou and his followers did not advocate prohibiting pollution outright or eliminating external costs all together. Rather, Pigou just proposed the imposition of taxes on external costs to reduce their levels toward the social optimal ones, or from point P to point S in terms of **Figure 1**. This is an efficient change as, at the free pollution point P, the net marginal value to the polluter is infinitesimal, while the marginal damage to the sufferers of pollution is large (measured by PC in **Figure 1**). This asymmetry in marginal value was missed by Coase in his all-or-nothing comparison which naturally focuses on the relative size of the two triangles ABE and EPC. Coase was correct only against the extreme environmentalists who want to ban pollution altogether but not against Pigovian tradition which was the main objective of that 1960s paper (which was worth half a Nobel Prize!).

Admittedly, the Pigovian case for taxing pollution should be taken only as prima facie, since administrative and other costs of taxation should also be taken into account, not to mention the danger of letting an inefficient and possibly corrupted government having more excuses to raise revenues and to engage in more corruption. However, even if the environmental scientists' prediction on the likely catastrophic outcomes of the business-as-usual option only has 5% probability of being realized, we certainly should address the problem seriously. Even if a report of a time bomb in your flight has only 1% reliability, you probably want to change your flight!

Even if we agree that taxing pollution and emission of greenhouse gases may be desirable, the appropriate amounts of taxes may be difficult to estimate as the damages are typically on the far future. However, in most cases, it is optimal to tax pollution at least in accordance to the marginal costs of reducing the pollutants, which are easier to estimate. Moreover, it would normally raise sufficient revenue to fund the optimal amount of disruption abatement spending (Ng, 2004b).

We have not covered debates such as the communitarian concern of the negative effects of using the market on altruism or morality (*see* Market and Nonmarket Allocation). Also, over the recent decades, studies by psychologists and behavioral economists have shown that individuals are not perfectly rational with consistent preferences (*see* Behavioral Economics; Experimental Economics). In answer, one approach is to regard a single individual as consisting of a number of 'agents' acting at different times or 'frames.' Without sufficient interagent-comparable cardinal utilities, this leads to very wide indecisiveness (Bernheim and Rangel, 2009; Rubinstein and Salant, 2012; Mandler, 2014). Focusing on experienced utility or actual subjective happiness may be a more fruitful approach (e.g., Kahneman et al., 1997; Kahneman and Sugden, 2005; Ng, 2003; Koszegi and Rabin, 2008), especially if we accept the view that happiness is the ultimate value (Ng, 2013).

See also: Behavioral Economics; Charity and Philanthropy, Economics of; Cost–Benefit Analysis; Environmental Economics; Experimental Economics; Income Distribution and Inequality: Measurement Issues; Public Choice Economics; Welfare: Philosophical Aspects.

Bibliography

Arrow, K.J., 1951/1963. Social Choice and Individual Values, second ed. Wiley, New York.

Atkinson, A., 2011. The restoration of welfare economics. American Economic Review 101, 157–161.

Barberà, S., 2011. Strategy-proof social choice. In: Arrow, K.J., Sen, A.K., Suzumura, K. (Eds.), Handbook of Social Choice and Welfare. North-Holland, Amsterdam, pp. 731–831.

Bernheim, B.D., Rangel, A., 2009. Beyond revealed preference: choice theoretic foundations for behavioral welfare economics. Quarterly Journal of Economics 124, 51–104.

Coase, R.H., 1960. The problem of social cost. Journal of Law and Economics 3, 1–44.

Cheung, S.N.S., 2013. The Memorials of Ronald Coase. Conference in Memorial of Ronald Coase, Shenzhen.

Diener, E., Kahneman, D., Helliwell, J., 2010. International Differences in Well-being. Oxford University Press, Oxford.

Dowding, K., Van Hees, M., 2009. Freedom of choice. In: Anand, P., Pattanaik, P.K., Puppe, C. (Eds.), Handbook of Rational and Social Choice. Oxford University Press, Oxford, pp. 374–392.

Fleurbaey, M., 2008. Fairness, Responsibility, and Welfare. Oxford University Press, Oxford.

Gibbard, A., 1973. Manipulation of voting schemes. Econometrica 41, 587–601.

Gotoh, R., Dumouchel, P., 2009. Against Injustice: The New Economics of Amartya Sen. Cambridge University Press, Cambridge.

Harsanyi, J.C., 1953. Cardinal utility in welfare economics and in the theory of risk-taking. Journal of Political Economy 61, 434–435.

Harsanyi, J.C., 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 63, 309–321.

Harsanyi, J.C., 1997. Utilities, preferences, and substantive goods. Social Choice and Welfare 14, 129–145.

Hotelling, H., 1931. The economics of exhaustible resources. Journal of Political Economy 39, 137–175.

International Monetary Fund, 2014. Fiscal Policy and Income Inequality. IMF Policy Paper. http://www.imf.org/external/np/pp/eng/2014/012314.pdf (accessed 14.07.07.).
Kahneman, D., Sugden, R., 2005. Experienced utility as a standard of policy evaluation. Environmental and Resource Economics 32, 161–181.
Kahneman, D., Wakker, P., Sarin, R., 1997. Back to Bentham? Explorations of experienced utility. Quarterly Journal of Economics 112, 375–406.
Kaplow, L., Shavell, S., 2002. Fairness versus Welfare. Harvard University Press, Cambridge, USA.
Kemp, M.C., Ng, Y.-K., 1976. On the existence of social welfare functions, social orderings and social decision functions. Economica 43, 59–66.
Kolm, S., 1997. Modern Theories of Justice. MIT Press, Cambridge, USA.
Kolm, S., 2005. Macrojustice: The Political Economy of Fairness. Cambridge University Press, Cambridge.
Koszegi, B., Rabin, M., 2008. Choices, situations, and happiness. Journal of Public Economics 92, 1821–1832.
Lipsey, R.G., Lancaster, K., 1956. The general theory of second best. Review of Economic Studies 24, 11–32.
Little, I.M.D., 1952. Social choice and individual values. Journal of Political Economy 60, 422–432.
Mandler, M., 2014. Indecisiveness in behavioral welfare economics. Journal of Economic Behavior and Organization 97, 219–235.
Moulin, H., 2003. Fair Division and Collective Welfare. MIT Press, Cambridge, USA.
Mueller, D.C., 2003. Public Choice III. Cambridge University Press, Cambridge.
Ng, Y.-K., 1975. Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. Review of Economic Studies 42, 545–570.
Ng, Y.-K., 1977. Towards a theory of third best. Public Finance 32, 1–15.
Ng, Y.-K., 1984. Quasi-Pareto social improvements. American Economic Review 74, 1033–1050.
Ng, Y.-K., 1990. Welfarism and utilitarianism: a rehabilitation. Utilitas 2, 171–193.
Ng, Y.-K., 1999. Utility, informed preference, or happiness? Social Choice and Welfare 16, 197–216.
Ng, Y.-K., 2000. Efficiency, Equality, and Public Policy: With a Case for Higher Public Spending. Macmillan, London.
Ng, Y.-K., 2003. From preference to happiness: towards a more complete welfare economics. Social Choice and Welfare 20, 307–350.
Ng, Y.-K., 2004a. Welfare Economics: Towards a More Complete Analysis. Palgrave Macmillan, New York.
Ng, Y.-K., 2004b. Optimal environmental charges/taxes: easy to estimate and surplus-yielding. Environmental and Resource Economics 28, 395–408.
Ng, Y.-K., 2007. Eternal Coase and external costs: a case for bilateral taxation and amenity rights. European Journal of Political Economy 23, 641–659.
Ng, Y.-K., 2009. Increasing Returns and Economic Efficiency. Palgrave Macmillan, New York.
Ng, Y.-K., 2013. The Road to Happiness. Fudan University Press, Shanghai.
Ng, Y.-K., 2014. Are Unrealistic Assumptions/Simplifications Acceptable? Some Methodological Issues in Economics. Manuscript.
Pattanaik, P.K., 1973. On the stability of sincere voting situations. Journal of Economic Theory 6, 558–574.
Pattanaik, P.K., Xu, Y., 1990. On ranking opportunity sets in terms of freedom of choice. Recherches Economiques de Louvain 56, 383–390.
Parks, R.P., 1976. An impossibility theorem for fixed preferences: a dictatorial Bergson-Samuelson welfare function. Review of Economic Studies 43, 447–450.
Rawls, J., 1971. A Theory of Justice. Havard University Press, Cambridge, USA.
Roberts, K.W.S., 1980. Social choice theory: the single-profile and multi-profile approaches. Review of Economic Studies 47, 441–450.
Roemer, J.E., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge, USA.
Rubinstein, A., Salant, Y., 2012. Eliciting welfare preferences from behavioral data sets. Review of Economic Studies 79, 375–387.
Salles, M., 2014. Social choice and welfare' at 30: Its role in the development of social choice theory and welfare economics. Social Choice and Welfare 42, 1–16.
Samuelson, P.A., 1967. Arrow's mathematical politics. In: Hook, S. (Ed.), Human Values and Economic Policy: A Symposium. New York University Press, New York.
Satterthwaite, M.A., 1975. Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory 10, 187–217.
Sen, A.K., 1966. Hume's law and Hare's rule. Philosophy 41, 75–79.
Sen, A.K., 1970. Collective Choice and Social Welfare. Holden-Day, San Francisco.
Sen, A.K., 2002. Rationality and Freedom. Harvard University Press, Cambridge, USA.
Sen, A.K., 2009. The Idea of Justice. Allen Lane, London.
Suzumura, K., 2011. Welfarism, individual rights, and procedural fairness. In: Arrow, K.J., Sen, A.K., Suzumura, K. (Eds.), Handbook of Social Choice and Welfare. North-Holland, Amsterdam, pp. 605–685.

Welfare: Philosophical Aspects

Jonathan Riley, Tulane University, New Orleans, LA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Given a distinction between 'welfare' and 'utility,' four different ways of thinking about welfare are discussed, namely: hedonistic conceptions, preference conceptions, resource conceptions, and capability conceptions. A laundered preference approach might be a way to combine insights from the different conceptions. If the preferences can always be represented by utility functions, such an approach can take the form of a nondictatorial welfarist philosophy such as standard utilitarianism. This is conceivable if a consensus exists about interpersonal comparisons of utility. But such a consensus cannot reasonably be expected, suggesting the need for a nonwelfarist philosophy. Yet perhaps a nonstandard welfarism may retain appeal, even if preferences cannot always be represented by utility functions and there is no reliance on interpersonal comparisons.

'Welfare' is not an easy term to define with precision. In the abstract, it refers broadly to the happiness, well-being, or good of an individual, society, political association, and so on. More concretely, it also refers to particular ingredients and sources of happiness or good. Life itself is an essential element of any human being's welfare, for example, and the means of basic subsistence, including food and shelter to some minimal extent, can be viewed as vital sources of welfare whose provision should be guaranteed by government, as in the term 'welfare state.' The sources and ingredients of human happiness are so various, however, that 'welfare' in the philosophical sense must not be limited to basic needs or the means of satisfying them. In this article, four different interpretations of welfare are discussed briefly. 'Welfarist' moral philosophies are also contrasted with 'nonwelfarist' moral philosophies. The first step is to dispel some confusion surrounding the term 'utility' and its relationship to welfare.

Welfare and Utility

Modern economists, decision theorists, and game theorists typically distinguish between welfare and utility. Utility in their terminology refers to the value of a mathematical function – strictly speaking, any function in a family of informationally equivalent functions – that *represents* the coherent preferences of a person or of society such that, for any pair of options in the feasible set of outcomes, a larger value is assigned to the preferred option. The relevant family of informationally equivalent utility functions is determined by the given assumptions of measurability and interpersonal comparability of utility. In this regard, measurability must not be confused with comparability across persons. A measurability assumption (such as ordinal measure, cardinal measure, or ratio-scale measure) specifies a particular degree of precision with which any person's preference ranking can be represented by utility numbers – the more precise, the smaller the family of informationally equivalent utility functions. Such an assumption is logically independent of any comparability hypothesis, according to which the utility representations of different persons' preferences are mutually dependent. Given comparability, the utility function which can be used to represent a person's preferences depends on the utility function selected from the family available to represent another person's preferences. In the absence of comparability, however, there is no constraint of this sort on which members of the relevant families can be used to represent the different persons' preferences.

To illustrate, suppose that utility numbers are ordinal and noncomparable. Given the ordinal measure, a person's utility function is unique only up to positive monotonic transformations. Such transformations preserve the order of relative priority of the utility numbers assigned to the alternatives. Given noncomparability, it is permissible to transform any person's utility function without altering the utilities of other people. Thus, a person's ordinal noncomparable utility ranking generally provides no more information than that contained in his own preference *ordering* – a complete, reflexive, and transitive relation – over the alternatives.

If utility numbers are assumed to be cardinal and noncomparable, however, then a person's utility function is unique up to positive linear (strictly speaking, affine) transformations: the cardinal noncomparable family of informationally equivalent functions is more restricted than its ordinal counterpart, which admits positive nonlinear transformations. A cardinal noncomparable utility ranking represents preference information that is richer than his preference ordering, namely, the relative *intensity* with which he prefers one option to another. If his revealed attitudes toward risk are independently assumed to measure how much subjective importance he assigns to his different options, for instance, the Von Neumann–Morgenstern method of cardinalization is appropriate and expected utility functions are valid cardinal noncomparable representations of his preference intensities (see, e.g., Binmore, 1994: pp. 259–278, 304–315).

To take a final illustration, if utility numbers are interpersonally comparable, then one person's utility function cannot be subjected to any transformation unless every other person's utility function is similarly transformed. A cardinal comparable utility ranking provides information not only about the person's preference intensities, for example, but also about how intense his preferences are in comparison to those of other people. To construct such a utility ranking, however, there must

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63093-3

be a method of making interpersonal comparisons apart from a method of cardinalization (Sen, 1979). For instance, a person might extend sympathy to the other members of his society by adopting their personal preference intensities while imagining himself in their respective positions at any pair of alternatives. He must also be able to render the different persons' intensities dimensionally compatible in order to compare them. By reducing them to a common unit, he could form a suitable extended preference reporting that when he places himself in their respective positions, he believes that one person i's preference for x over y is of greater intensity than another person j's preference for y over x, for any two persons and any pair of alternatives. His personal preferences, a subrelation of his extended ones, might continue to be decisive for his actions. But if he can extend sympathy to others in this way, his preferences can be represented by a cardinal comparable utility function. At the same time, different people may disagree with respect to interpersonal comparisons, and thus form conflicting extended preferences. *A fortiori*, different people may have conflicting personal preferences and conflicting cardinal comparable utility functions.

In general, a utility function encapsulates information about measurability and interpersonal comparability of the underlying preferences. As a device for representing this preference information (including orderings, intensities, and interpersonal comparisons), utility in this technical sense is quite distinct from any substantive idea of welfare. Indeed, utility in this sense can conflict with welfare if the preferences being represented are not valid welfare judgments: larger personal utility values can be assigned to outcomes which, though preferred by the agent, are worse in terms of his welfare, for example, just as smaller aggregate utility values can be assigned to outcomes which, though rejected by society, are better in terms of social welfare. Some have argued that the term utility should thus be confined to the technical meaning so as to avoid needless confusion resulting from the identification of utility with 'good' (e.g., Broome, 1991a,b). Perhaps so. But there are reasons to proceed with caution in this. First, utility continues to be associated with welfare – 'being useful or beneficial' – in our everyday language. Even within economics, this everyday association still surfaces occasionally. More confusion may be caused by trying to banish it than by living with it.

Second, the identification of utility with welfare has a long history within both economics and philosophy (Sen, 1987, 1991). Bentham, J. S. Mill, and other classical utilitarian philosophers held that utility means happiness or enjoyment in the sense of pleasure and absence of pain, for example, and that actions ought to be evaluated in terms of their consequences for general happiness or social welfare, maximization of which is the sole ultimate standard of good (*see* Consequentialism (Including Utilitarianism)). For these thinkers, any action or other thing is 'useful' or 'has utility' if and only if it has a tendency to promote pleasure net of pain. They deny (what is commonly asserted) that something may give much pleasure yet be useless, or that something may be useful even though it tends to promote more pain than pleasure. Earlier thinkers such as Hume and Helvetius also may have held this hedonistic idea of utility if not the distinctive utilitarian doctrine that general utility maximization is the *sole* ultimate test of good conduct.

As Mill (1965: pp. 45–46) elaborated the hedonistic idea, three kinds of utilities can be produced by human actions, namely: 'utilities fixed and embodied in outward objects' (such pleasures are created by 'investing external material things with properties which render these serviceable to human beings'); 'utilities fixed and embodied in human beings' (again, such enjoyments are made possible by 'conferring on human beings qualities which render them serviceable to themselves and others'); and, finally, 'utilities not fixed or embodied in any object, but consisting of a mere service rendered; a pleasure given, an inconvenience or a pain averted, during a longer or shorter time, but without leaving a permanent acquisition in the improved qualities of any person or thing; the labor being employed in producing an utility directly, not (as in the two former cases) in fitting some other thing to afford an utility.' Useful material objects and virtuous or noble human beings are *means* to pleasure. After repeated experience of their consequences, however, these more or less permanent sources of pleasure can also become inseparably associated with the idea of happiness itself, through habit, and thus be seen as concrete ingredients of the good (Mill, 1969: pp. 234–239).

Perhaps because utilitarians were so influential in the development of classical economics, the hedonistic interpretation of utility was also widely accepted by economists prior to the 'ordinalist revolution' of the 1930s and 1940s. Classical economists were preoccupied with the production and distribution of material wealth, however, and so tended to focus their attention on the kind of utilities embodied in material objects susceptible to accumulation. As used by Mill (1965: pp. 48–54), for example, the term 'productive labor' is confined to those types of actions that produce utilities embodied in material objects. Actions are classed as 'unproductive labor' (however useful to society) to the extent that they produce the other two kinds of utilities. Moreover, 'productive consumption' (i.e., investment) includes consumption by productive laborers of things which are necessary to maintain and increase their own health, strength, capacities of work, or population. But it does not include any consumption by unproductive laborers (those who do not produce material wealth) or consumption of luxuries by the productive themselves (however pleasing these forms of consumption are to the community). In short, although they did not deny that other kinds of utilities existed, economists in the classical and early neoclassical tradition (from Smith and Ricardo to Marshall and Pigou) tended to identify utility with material welfare and to be especially concerned about the needs of those members of the community who produced the kind of utilities embodied in material wealth.

In the post–Second World War period, utilitarian ethics have declined in influence and hedonistic conceptions of utility have largely been abandoned by economists and philosophers. Whereas philosophers (including utilitarians) have typically gone on to identify utility with some nonhedonistic idea of welfare, however, economists have generally adopted the technical idea of utility in its purely ordinalist form – utility is merely a convenient mathematical

tool for representing interpersonally noncomparable preference orderings (*see* Rational Choice Explanation: Philosophical Aspects). It is not obvious that philosophers should follow economists in viewing utility as a merely technical term, distinct from welfare. (For a controversial argument that this shift in the meaning of utility within economics – the product of the 'ordinalist revolution' – does not imply 'progress' because it has involved 'losses in understanding human welfare,' in particular, 'material welfare,' including any person's objective material needs; see Cooter and Rappoport, 1984. Further relevant discussion is given by Bruni and Sugden (2007).) But perhaps they should, without necessarily accepting the narrow restriction to pure ordinalism. Even so, it would be incorrect to conclude that because technical utility values are distinct from, and can conflict with, substantive welfare values, there is some inescapable contradiction between utility in the technical sense and welfare. This leads to a third reason for caution in insisting on a complete divorce between utility and welfare, to wit, the technical idea of utility may remain a formal aspect of any adequate conception of welfare if judgments about welfare are a species of preferences that can be represented by utility functions. In that case, technical utility, as a formal component of the idea of good, could be seen as a vestige of earlier utility concepts that identified utility with welfare.

Whether or not judgments about welfare can be represented by utility functions, the question remains: what are such judgments about?

Conceptions of Welfare

At least four different conceptions of welfare continue to be debated in the literature.

Welfare as Pleasure

Although currently out of favor, hedonistic conceptions have not vanished entirely and may be making a comeback (see, e.g., Kahneman et al., 1997; Kahneman, 1999; Kahneman and Sugden, 2005). Hedonism is typically dismissed as crude and naive, however, both as a psychology and as a theory of welfare. Expectation of pleasure including relief from pain is not the sole motivational force in human nature, it is said, and human good involves myriad things other than pleasure. No rational and moral person would be happy, for example, if his life consisted solely of a stream of supreme pleasures guaranteed by artificial means.

Perhaps such arguments are decisive against some versions of hedonism. But, as already discussed, the classical hedonism of Bentham, Mill and others does not deny that many things such as money and virtue are means to pleasure which can also become valued as ends-in-themselves, in which case they are seen as concrete ingredients of the idea of happiness. Classical hedonism is not a purely subjective doctrine. It insists that some kinds of pleasures are embodied in material objects and human qualities, and can only be experienced by people who possess those objects and qualities, which must first be produced. For instance, Mill argues that intellectual and moral pleasures are higher in quality than mere physical sensations of pleasure and can only be experienced by agents who possess the relevant intellectual and moral capacities, capacities that must be developed in part by the agents themselves (Riley, 2009, 2010).

Welfare as Preferences

Perhaps currently most influential, preference-oriented approaches conceive personal welfare in terms of personal preferences and social welfare in terms of social or moral preferences that are constructed in some way from the set of personal preferences. But there is an ambiguity in the key notion of 'preference-satisfaction': it may mean the fulfillment of preference by a suitable shift in the way the world is; or it may mean the satiation of preference by a suitable shift in the subject's conception of the way the world is. Proponents of preference approaches often insist on the latter 'satiation' interpretation because of their implicit subjectivism. Fulfillment without satiation (e.g., posthumous fulfillment of preferences) does not count. From this subjectivist perspective, hedonism may be viewed as a special case in which preference-satiation is equated with a mental state of pleasure including relief from pain. As already mentioned, however, classical hedonism is not a purely subjectivist doctrine. Some feasible pleasures are said to be embodied in material objects and human virtues, for instance, even if they are not at the moment being experienced. In any case, defenders of preference conceptions usually reject hedonism as implausible.

Two versions of the preference approach should be distinguished. One defines 'preferences' in terms of observable 'choice behavior' (see, e.g., Harsanyi, 1992; Binmore, 1994, 1998). A person's preference for x over y is 'revealed' by his choice of x when y is available, independently of his desires or other motivations for the choice. In contrast, the other version defines preferences as 'desires' rather than 'choices' (e.g., Hare, 1981; Griffin, 1986). A preference for x over y implies greater desire-satisfaction (though not necessarily greater pleasure) from x than from y. Nevertheless, the agent could – and perhaps should – still make a counterpreferential choice of y over x in this case. Of course, the notion of 'counterpreferential choice' makes no sense under the choice-oriented variant. But that approach is silent about the motivations people have for making the choices it views as comprising personal or social welfare. It needs to be supplemented by a theory of motivation, not only to explain why people do make those choices but also to justify why those choices ought to be seen as conducive to welfare. For its part, the 'desire approach' does not entirely lack a theory of motivation. But its theory may well be inadequate. It is arguable that people are not always motivated to choose the options that they expect will bring them the greatest satisfaction of their given desires, nor should they be. Perhaps they are, and ought to be, motivated to sacrifice or revise their greatest desires because they have moral commitments (Sen, 1977), for example, or because they wish to seize opportunities for mutual advantage that change from moment to moment (Sugden, 2004, 2010).

However preferences are interpreted, the main appeal of preference conceptions is often taken to reside in their implicit subjectivism: the welfare of any individual or society is judged from the subjective viewpoint of that particular individual or society. The contents of welfare judgments may vary significantly across different persons or social groups (*see* Diversity and Disagreement in Ethics: Philosophical Implications). Nevertheless, the appeal of any purely subjective interpretation of welfare is quite limited. People may be deluded, for example, or ill-informed, or resigned to their customary positions as slaves or 'untouchables' in highly stratified societies (*see* Irrationality: Philosophical Aspects). Their naked personal preferences may thus be poor indicators of personal welfare. Moreover, people may be unable to imagine themselves in others' shoes, or they may clash over interpersonal comparisons, or display undue bias against others' interests. As a result, they may fail to construct moral preferences that can serve as social welfare judgments.

Defenders of preference conceptions often respond to these difficulties by suitably laundering preferences. Conditions are specified that a person's choices or desires must satisfy before they can be accepted as conducive to his own welfare or that of his society. Thus, for example, Harsanyi (1992) suggests that a person's choices are conducive to his own welfare only if they are informed and conform to the expected utility axioms, whereas his choices are conducive to social welfare only if, in addition, they are impersonal. Impersonal choices must conform to the equiprobability model of moral choice and are used only to select an optimal code of moral rules. The person must forget his own position (defined broadly to include family history and genetic makeup) and imagine himself occupying the position of any person in society with equal probability. He must also extend sympathy to each person j by making the same choices j would make from the set of feasible codes while in j's shoes. Given the independent assumption that all persons share a fundamentally similar human nature (including attitudes toward risk used to measure the relative importance of different choices), which is revealed to anybody who fights against his biases by adopting the impersonal perspective of the equiprobability model, he can then infer interpersonal comparisons of cardinal utility which are identical to those inferred by anybody else who goes through this procedure. Finally, he must aggregate impartially over the cardinal comparable utilities he experiences at the respective persons' positions to construct moral preferences, which turn out to be (average) utilitarian in form. Once an optimal moral code is selected, individuals are free to compete over scarce resources by playing their permissible strategies under the rules.

Harsanyi's version of rule utilitarianism is vulnerable to numerous objections (Binmore, 1994, 1998; Riley, 2009). Moreover, even if such hypothetical preferences have some appeal as conceptions of welfare, the subjectivism which is usually thought to be the main strength of preference approaches is lost unless agents affirm the hypothetical preferences under the relevant conditions. To preserve an element of subjectivism, individuals and societies must willingly adopt those preferences before they can be said to be conducive to the welfare of those particular individuals or societies. If actual adoption of the preferences is a necessary condition of good, however, attention is redirected to psychological theories of evolution and education which explain how people develop the intellectual and moral capacities required to make the relevant choices or form the relevant desires (*see* Evolution, Natural and Social: Philosophical Aspects). It remains an open question whether a credible theory of this sort exists. But whether or not human capacities can be developed such that human beings invariably obtain the greatest desire-satisfaction from making moral choices, a credible psychological theory is needed. Some version of hedonism may yet prove plausible in this regard.

Welfare as Resources

In resource-oriented conceptions, personal welfare is defined in terms of an objective list of resources broadly defined, whereas social welfare consists implicitly of the effective production and fair distribution of the relevant resources (*see* Justice: Philosophical Aspects). Rawls (1999, 2005) can be read as offering such a conception, for example. He refers to 'primary goods' as 'things which every rational man is presumed to want' and 'things citizens need as free and equal persons,' and provides a (perhaps incomplete) list that includes 'basic rights and liberties,' 'freedom of movement and free choice of occupation against a background of diverse opportunities,' 'powers and prerogatives of offices and positions of responsibility in the political and economic institutions of the basic structure,' 'income and wealth,' and 'the social bases of self-respect.' Dworkin (2000) spells out another approach of this sort. (Kymlicka (2002) provides an excellent introduction to the wide-ranging critical debate surrounding such liberal theories of justice.)

There is little doubt that human welfare depends on the possession of such valuable things as those listed by Rawls or Dworkin. Data on the extent to which a person possesses these items may provide helpful indicators of different aspects of his good, even if the indicators cannot be combined into a single index of personal welfare. Similarly, data on the social distributions of these things may provide indicators of different dimensions of social welfare, with the same caveat.

Nevertheless, in addition to the specific objections which may be directed against a particular theory such as Rawls's or Dworkin's, resource conceptions face at least two general difficulties. First, it seems that the relevant resources are conducive to welfare only because they are objects of suitably laundered preferences. As Rawls suggests, for example, primary goods are things which every person would prefer to have if he were rational and reasonable. Furthermore, since a person's holdings of one resource (say, political responsibilities) can interfere with his holdings of another (say, wealth), it seems that his preferences must settle which of many different mixes of resources is best for his welfare. Similarly, it seems that moral preferences are needed to settle which among many competing social distributions of resources are permissible if not essential to promote social welfare. If this is right, resources may well be essential ingredients of welfare but they properly figure as objects of preferences.

The second difficulty is that resource conceptions do not take account of interpersonal variation in human capacities to convert resources into the achievement of human functionings (Sen, 1990). This problem leads to another way of conceiving welfare.

Welfare as Capabilities

Sen (1985a,b, 1990, 2004, 2009) argues that human well-being should be conceived in terms of 'capabilities' – or 'positive freedoms' (*see* Freedom: Political) – to achieve valuable human 'functionings' given by some objective list (see also Sugden, 2008; Basu and Lopez-Calva, 2010). He points out that different people may have unequal capabilities to convert resources into achievements. Physically handicapped people may need more income than the able-bodied need to achieve a given level of mobility, for example. People who are genetically vulnerable to particular diseases may need more medicines than the less vulnerable need to achieve a given standard of health. Women may need more food than men need to achieve a given standard of nutrition. Racial minorities may need more educational and job opportunities than white males need to achieve a given level of career success if racist attitudes are prevalent. The elderly may need more wealth than the young need to achieve a given standard of living. These are essentially empirical questions.

Such examples suggest that capability conceptions can remedy a serious deficiency in resource approaches to welfare. But, as with resources, it seems arguable that certain capabilities are properly viewed as essential components of welfare because they are objects of suitably laundered preferences. Some capabilities are in effect things that every rational or moral person is presumed to want or choose. To the extent that anybody lacks these benchmark freedoms, reasonable people may agree that some form of special compensation ought to be provided so that the person can function 'normally,' for example, income subsidies, extraordinary medical assistance, affirmative action in hiring decisions, and so on. Furthermore, beyond the benchmark capabilities which everyone is presumed to need, it seems that a person's preferences must determine which balance of extra capabilities is proper for his welfare since capabilities to achieve some functionings can presumably interfere with capabilities to achieve others. Again, since one person's capabilities might interfere with another's, ethical preferences are needed to settle which of many competing social distributions of capabilities (including benchmark capabilities) is permissible if not essential to promote social welfare.

Combining Insights

Perhaps some version of the preference conception of welfare may have appeal as a way to combine insights from the various approaches to welfare. The upshot seems to be that, to serve as a personal welfare judgment, a person's preferences must be suitably laundered not only to satisfy certain formal conditions of rationality but also certain substantive conditions. In particular, the preferences must demand a personal share of certain all-purpose resources which every rational person is presumed to want, making due allowance for interpersonal variation in capabilities. Moreover, a credible psychology is required to explain how individuals come to form the relevant preferences.

It might be argued, for instance, that a person's welfare is happiness in the sense of pleasure, that a desire for any object or thing is proportional to the pleasure expected – perhaps mistakenly – from it, that mistaken beliefs ought to be corrected, that strongest present desires determine choices, that repeated choices may result in habitual attitudes or dispositions in which the original motivating pleasures are forgotten, and that habitual choice behavior reveals preference rankings that can be represented by utility functions. Other hypotheses may be more plausible, of course. But some such integrated preference-oriented approach appears essential.

To serve as a social welfare judgment, preferences must satisfy, in addition, certain moral conditions, including formal conditions of impartiality or impersonality as well as substantive conditions of fairness in the distribution of all-purpose resources making due allowance for capabilities. A social decision procedure that maps any given set of personal welfare judgments into a social welfare judgment thus understood must be specified. Again, a credible psychology is needed to explain how individuals develop the intellectual and moral abilities required to make and endorse the relevant moral preferences.

If a laundered preference approach has such potentially broad appeal, *welfarist* moral philosophies might adequately capture the phenomenon of human welfare, as Sumner (1996) maintains. To consider this question, suppose initially that personal welfare rankings can invariably be represented by technical utility functions, and that this utility information is all that matters for the construction of moral preferences which can also be represented by utility functions.

Standard Welfarist Philosophies

Standard welfarist philosophies are distinguished by their exclusive reliance on welfare rankings, which I am construing as laundered preferences, and the utility functions that represent them. Perhaps the most well known examples of such philosophies are the standard versions of utilitarianism. But there are many different forms of welfarist philosophies. They can be clarified by means of the concept of a *social welfare functional* (SWFL) (Sen, 1986: pp. 1111–1128). For a given set of options, any SWFL maps a given set of personal utility functions (one for each person in society) into a social or moral preference which can itself be represented by a utility function. Since each person's utility function is an element of some family of informationally identical utility functions as determined by the given assumptions of measurability and comparability, however, the SWFL is required to satisfy an invariance requirement: it must generate the same preferences if any given set of personal utility functions is subjected to transformations permitted by the measurability and comparability assumptions.

Welfarism is a neutrality condition imposed on the form of an SWFL. It requires that if each person's utility values with respect

to one pair of options x and y are equal to his utility values with respect to another pair w and z – subject to transformations permitted by the given measurability and comparability assumptions, then the SWFL must generate the same social or moral preference over the one pair as over the other. The only information that matters for a welfarist SWFL is utility information. Nonutility information has no influence.

Arrow's (1963) well known 'impossibility theorem' shows, roughly, that any welfarist SWFL must be dictatorial if personal utility functions are ordinal and noncomparable. But a dictator only counts his own personal preferences: he sympathizes with nobody else to determine social preferences. (For variations and extensions of Arrow's theorem, including its extension to the case of cardinal noncomparable utilities, see Sen, 1986: pp. 1073–1076, 1114–1115.) A welfarist SWFL that is impersonal in so far as its operation does not depend on any person's identity is possible, however, if interpersonal utility comparisons are possible. As Harsanyi's equiprobability model illustrates, an impersonal observer can be assumed to make the interpersonal comparisons by extending sympathy to each person while in that person's position. The observer is an ideal moral agent who supposedly makes comparisons which all reasonable people with capacities of imaginative sympathy would agree to make. This idea that there is one right way to make interpersonal comparisons is implicit in the design of many other welfarist SWFL's and seems essential to generate incontestable moral preferences. The idea is highly implausible, however, so that there are serious doubts that an adequate philosophy of human welfare can be put into a welfarist form.

Nonstandard Welfarism

Nevertheless, an intriguing possibility is given by the liberal utilitarian philosophy which I have attributed to J.S. Mill (Riley, 1988, 2010, 2012, 2014). As I now understand it, the Millian philosophy is a nonstandard form of welfarism in which a welfarist SWFL is restricted in its operation to a higher kind of preferences associated with the moral sentiment of justice, with justice conceived in terms of a moral and legal code of equal rights and duties for all, whereas individuals are free to pursue their other kinds of preferences (which may be unstable or incoherent) in accordance with their recognized rights and duties because society has no need to map these other kinds of preferences into a social or moral preference ranking. To facilitate this, the neutrality condition imposed on the SWFL must be suitably restricted in scope such that it refers only to the higher kind of preferences and does not apply across the different kinds. In effect, the Millian welfarist SWFL, which generates a moral kind of social preference from any given set of a moral kind of personal preferences, all of which can be represented by utility functions, makes possible an 'enlarged' utilitarian 'art of life' that is compatible with, though it extends beyond, the welfarist morality. The enlarged philosophy prescribes freedom for individuals and groups in domains that are beyond the moral realm of the SWFL, including a purely personal domain as well as a competitive market domain. There is no presumption that individuals will form preferences that can be represented by utility functions in these nonmoral domains.

Although I cannot go into the details here, three further points need to be made about this unusual form of welfarism. First, the different kinds of preferences are ultimately motivated by different kinds of pleasures, some of which are intrinsically more valuable than others are as pleasure, and the different kinds have different sources. In particular, the pleasant feeling of justice is held to be of a higher quality than any competing kinds of pleasure, irrespective of quantity (Riley, 2010, 2012). Codes of equal rights and duties are the only source of this higher kind of pleasure: it is a complex feeling of security which an individual can only get by living with peers under a binding code which he or she has a voice in creating.

Second, competent human beings discover the ranking of the different kinds of pleasant feelings in terms of their quality by experiencing them, a ranking which is confirmed not created by the preferences of the vast majority if not all experienced people. Again, the higher pleasures doctrine is not purely subjective: the higher kinds of pleasures are latent in human nature, awaiting discovery by those who develop their intellectual and imaginative capacities. (Since humans are fallible, some disagreement over the qualitative ranking remains possible.) According to Mill, the different kinds can be distinguished without relying on values other than pleasure, just as colors can be distinguished without relying on anything other than color. (Many philosophers, following Sidgwick and Moore, reject this. Yet Mill may be correct that desiring something and thinking it desirable are just two modes of naming the same psychological fact, i.e., the fact that the person associates some kind of pleasure including relief from suffering with the thing.) In any case, the qualitative ranking of the kinds of pleasures, even if unanimously shared by those who are 'competently acquainted' with the different kinds, is a lexical preference that cannot be represented by a technical utility function. Any higher kind of pleasure is discontinuous in value with any lower kind: any amount of the latter which humans can experience can never be equal in value to even a bit of the former as pleasure.

Third, with respect to the operation of the welfarist SWFL in its limited moral domain, it seems necessary to concede that it is too costly if not impossible to make interpersonal utility comparisons which every competent person accepts. (Interpersonal utility comparisons are not otherwise needed since individuals and groups enjoy liberty outside the moral domain.) Even if such comparisons are abandoned and we are left with purely ordinalist utility information, however, welfarist SWFLs can employ artificial devices to generate a moral preference for society from any given set of the moral kind of personal preferences. Thus, for instance, a Borda SWFL does not rely on interpersonal comparisons and yet generates a social welfare judgment that can be represented by a utility function. It works by employing a rather artificial yet impersonal scheme of points (i.e., ballots) to supplement the utility information contained in any given set of personal preference orderings. (This Borda procedure avoids Arrow's impossibility result by effectively modifying the Arrovian neutrality condition.)

Beyond Welfarism

Although such a nonstandard welfarist philosophy is of considerable interest, the possibility remains that a nonwelfarist moral philosophy – one that relies to some extent on nonutility information, that is, information which is not contained in preference rankings or the utility functions representing them, to produce moral choices for society – may have even more appeal. Such a philosophy might be based on an accurate description of human behavior, for example, even though the goal of maximizing welfare is dropped. Myriad forms of nonwelfarist philosophies are conceivable. Such a philosophy may take the form of an SWFL that does not satisfy the relevant neutrality condition. The nonwelfarist SWFL might make use of nonutility information about the feasible options to generate different social preferences over two pairs of options, even though the utility information relating to the two pairs is identical. One pair might belong to an individual's rightful 'protected sphere' whereas the other does not, for example, or the nature of the options in one pair might involve the oppression of some minority whereas the options in the other pair involve no such oppression. The social preference ranking generated by the SWFL can still be represented by a utility function but the social preferences over the two relevant pairs of options will not be the same even though the personal utility information is identical with respect to those pairs.

Nonwelfarist philosophies may instead eschew the form of an SWFL. Binmore's (1994, 1998) evolutionary cooperative game structure does not attempt to generate social preference orderings of feasible options, for instance, and assigns a special status to the prevailing Nash bargaining equilibrium instead of satisfying a neutrality condition (see also Sugden, 2001; Riley, 2006). To take another example, Sugden (2004a,b, 2011) proposes a coordination game framework in which social outcomes are not deduced from given sets of personal preference rankings but instead reflect mutually advantageous conventions that are induced by individuals on the basis of shared perceptions of salience or 'focal points.' A nonwelfarist philosophy might also take a value pluralistic form (*see* Value Pluralism), in which welfare is only one of plural irreducible values in ethical evaluations. Sen (1985a,b, 2004, 2009) proposes a pluralistic structure in which the value of 'well-being' cannot be reduced to that of 'agency,' for instance. By implication, indeterminacy may arise in all-things-considered ethical preferences if well-being comes into conflict with agency.

See also: Economics and Ethics; Equality: Philosophical Aspects; Ethics and Values.

Bibliography

Arrow, K.J., 1963. Social Choice and Individual Values, second ed. Wiley, New York.

Basu, K., Lopez-Calva, L.F., 2010. Functionings and capabilities. In: Arrow, K.J., Sen, A.K., Suzumura, K. (Eds.), Handbook of Social Choice and Welfare, vol. 2. North-Holland, Amsterdam.

Binmore, K., 1994. Playing Fair. MIT Press, Cambridge, MA.

Binmore, K., 1998. Just Playing. MIT Press, Cambridge, MA.

Broome, J., 1991a. Utility. Economics and Philosophy 7, 1--12.

Broome, J., 1991b. Utility – ideas and terminology – reply. Economics and Philosophy 7, 285–287.

Bruni, L., Sugden, R., 2007. The road not taken: how psychology was removed from economics and how it might be brought back. Economic Journal 117, 146–173.

Cooter, R., Rappoport, P., 1984. Were the ordinalists wrong about welfare economics? Journal of Economic Literature 22, 507–530.

Dworkin, R., 2000. Sovereign Virtue. Belknap Press of Harvard University Press, Cambridge MA.

Griffin, J., 1986. Well-being. Clarendon Press, Oxford, UK.

Hare, R.M., 1981. Moral Thinking. Clarendon Press, Oxford, UK.

Harsanyi, J.C., 1992. Game and decision theoretic models in ethics. In: Aumann, R.J., Hart, S. (Eds.), Handbook of Game Theory, vol. 1. Elsevier, Amsterdam.

Kahneman, D., 1999. Objective happiness. In: Kahneman, D., Deiner, E., Schwarz, N. (Eds.), Well-being: The Foundations of Hedonic Psychology. Russell Sage Foundation, New York.

Kahneman, D., Sugden, R., 2005. Experiential utility as a standard of policy evaluation. Environmental & Resource Economics 32, 161–181.

Kahneman, D., Wakker, P.P., Sarin, R.K., 1997. Back to Bentham? Explorations of experienced utility. Quarterly Journal of Economics 112, 375–405.

Kymlicka, W., 2002. Contemporary Political Philosophy: an Introduction, second ed. Oxford University Press, Oxford, UK.

Mill, J.S., 1965. Principles of political economy. In: Robson, J.M. (Ed.), Collected Works of J.S. Mill, vols. II–III. University of Toronto Press and Routledge & Kegan Paul, Toronto and London, pp. 48–54.

Mill, J.S., 1969. Utilitarianism. In: Robson, J.M. (Ed.), Collected Works of J. S. Mill, vol. X. University of Toronto Press and Routledge & Kegan Paul, Toronto and London, pp. 234–239.

Rawls, J., 1999. A Theory of Justice, second ed. Harvard University Press, Cambridge, MA.

Rawls, J., 2005. Political Liberalism, exp. pbk. edn. Columbia University Press, New York.

Riley, J., 1988. Liberal Utilitarianism. Cambridge University Press, Cambridge, UK.

Riley, J., 2006. Genes, memes and justice. Analyse & Kritik 28, 32–56.

Riley, J., 2009. The interpretation of maximizing utilitarianism. Social Philosophy and Policy 26, 286–325.

Riley, J., 2010. Justice as higher pleasure. In: Kelly, Inin P.J., Varouxakis, G. (Eds.), J.S. Mill: Thought and Influence. Routledge, London.

Riley, J., 2012. Happiness and the moral sentiment of justice. In: Kahn, L. (Ed.), Mill on Justice. Palgrave Macmillan, London.

Riley, J., 2014. Just Happiness. Oxford University Press, Oxford, UK.

Sen, A.K., 1977. Rational fools: a critique of the behavioral foundations of economic theory. Philosophy and Public Affairs 6, 317–344.

Sen, A.K., 1979. Interpersonal comparisons of welfare. In: Boskin, M.J. (Ed.), Economics and Human Welfare. Academic Press, New York.

Sen, A.K., 1985a. Commodities and Capabilities. Elsevier, Amsterdam.

Sen, A.K., 1985b. Well-being, agency and freedom. Journal of Philosophy 82, 169–221.

Sen, A.K., 1986. Social choice theory. In: Arrow, K.J., Intriligator, M.D. (Eds.), Handbook of Mathematical Economics, vol. III. North-Holland, Amsterdam, pp. 1073–1128.

Sen, A.K., 1987. On Ethics and Economics. Basil Blackwell, Oxford, UK.

Sen, A.K., 1990. Justice: means versus freedoms. Philosophy and Public Affairs 19, 111–121.

Sen, A.K., 1991. Utility: ideas and terminology. Economics and Philosophy 7, 277–283.

Sen, A.K., 2004. Rationality and Freedom. Belknap Press of Harvard University Press, Cambridge, MA.

Sen, A.K., 2009. The Idea of Justice. Belknap Press of Harvard University Press, Cambridge, MA.

Sugden, R., 2001. Ken Binmore's evolutionary social theory. Economic Journal 111, F213–F248.

Sugden, R., 2004a. The Economics of Rights, Cooperation and Welfare, second ed. Palgrave Macmillan, Basingstoke.

Sugden, R., 2004b. The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. American Economic Review 94, 1014–1033.

Sugden, R., 2008. Capability, happiness and opportunity. In: Bruni, L., Comim, F., Pugno, M. (Eds.), Capability and Happiness. Oxford University Press, Oxford, UK.

Sugden, R., 2010. Opportunity as mutual advantage. Economics & Philosophy 26, 47–68.

Sugden, R., 2011. Salience, inductive reasoning and the emergence of conventions. Journal of Economic Behavior and Organization 79, 35–47.

Sumner, L.W., 1996. Welfare, Happiness and Ethics. Clarendon Press, Oxford, UK.

Welfare Policy Comparisons

Gosta Esping-Andersen, Universitat Pompeu Fabra, Barcelona, Spain

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The author reviews the three dominant questions within the comparative welfare state literature: one, how does one explain policy variations across countries and time? In particular, what has been the relative importance of political power, industrial modernization, and other historical factors? Two, are welfare states sustainable in an era of heightened globalization? Three, how do such variations influence welfare outcomes, in particular related to income distributions, poverty, family life, and gender relations?

Introduction

Welfare policy comparisons have been a prominent feature of macrolevel social science research since the 1970s. Dominated by political scientists and sociologists, the focus is primarily on the advanced industrial democracies. The literature is guided by two overarching kinds of questions: one, how does one explain the rise of welfare states, their diversity, and, more recently, their adaptation to the new global economic order? Two, how do variations in policy design affect welfare distributions? This article will treat the two questions separately.

It has been difficult to arrive at clear-cut answers to either question because of statistical overdetermination; the number of rival explanations often exceeds the number of cases that can be studied. There are, moreover, conceptual difficulties because scholars do not always compare and explain the same phenomenon. Some compare specific social policies (say pensions), others welfare states, and still others welfare regimes, and the difference is rarely made clear. The presence of a set of social policies does not automatically imply a welfare state, which, following Marshall (1950), implies social entitlements as a matter of citizenship, comprehensive risk coverage, commitments to full employment, and active reduction of inequalities. The welfare state is also an ideological construct, a means to inculcate broad solidarities and national integration. Rather than speaking of more-or-less welfare statism, the literature has increasingly favored typologies of kinds of welfare states. Many scholars use the concept of welfare regimes to denote the broader complex, or mix, of social welfare provision between state, market, and families (for a discussion, see Esping-Andersen, 1999).

Conceptualizing Welfare Policy Models

Early quantitative comparisons, such as Wilensky (1975), compared welfare policy differences through national social spending levels. This was challenged on validity grounds (spending may not reflect welfare effort, and it tells one little about what kind of welfare is being produced and for whom). As a result, research in the past two decades has been greatly concerned with the precise meaning and content of social policies and the welfare state. Most conceptualizations begin with Marshall's (op. cit) emphasis on the rights of social citizenship, and programs are then compared in terms of the strength of entitlements, to what extent they lessen market dependency ('decommodification'), and to what degree rights are universal or differentiated (Flora and Alber, 1981; Korpi, 1983; Esping-Andersen, 1990). More recently has emerged an explicit gender perspective aiming to identify how welfare policies influence gender relations and, more broadly, the revolution of women's roles (O'Connor et al., 1999; Gornick and Meyers, 2003; Esping-Andersen, 2009).

The greater conceptual attention to policy content and objectives led to a keener appreciation of the qualitative differences between national welfare systems. Drawing upon Titmuss (1974), there emerged a voluminous literature on welfare state typologies, initially focused on welfare state models. As research expanded to the interplay among state, market, and family provision, emerged classifications of welfare regimes, such as the liberal, conservative, and social democratic trichotomy in Esping-Andersen (1990) and further refinements, such as Castles' (1993) notion of a unique Antipodean 'wage earners' model,' the identification of a highly familialistic and patronage-based Mediterranean model (Ferrera, 1996; Esping-Andersen, 1999). Some feminist writings have proposed typologies based on the strength of the conventional male breadwinner assumption (Lewis, 1993). With the exception of the latter, most typologies classify nations rather similarly: a Nordic group, distinguished by its universalism, comprehensive coverage, generosity, and a strong commitment to gender equalization; an Anglo - Saxon group, characteristic for its residual public provision, its targeting of benefits, and its encouragement of private welfare; a Continental European group, which stands out for its corporativistic status-segmented social insurance; and, more controversially, a Mediterranean (and, possibly, East Asian) group primarily distinguished by its strong accent on familial welfare provision. As Castles (1993) points out, these typologies resemble 'families of nations,' nation clusters with a deep common historical tradition. Recent years have seen a renewed debate on whether the 'Three Worlds' typology still remains valid in light of decades of policy change and reform. On balance, the answer is largely confirmative (Arts and Gelissen, 2010). This is consistent with Pierson's (1996) influential theory of institutional path dependency – the resiliency of institutions once they are firmly implanted.

Explaining Policy Divergence

The great vitality of comparative welfare state research comes from its address to central theoretical issues in the social sciences. Existing reviews of the literature usually distinguish two broad explanatory thrusts: the convergence thesis, which derives from theories of industrialism and modernization, and the 'politics matter' thesis, which derives from theories of conflict and power mobilization (Esping-Andersen and Van Kersbergen, 1992). In addition, some theorists emphasize the process of nation building and democratization (Flora and Alber, 1981), nations' position in, and response to, the world economy (Cameron, 1978; Katzenstein, 1985), and also the policy-making role that state bureaucracies exert (Skocpol, 1985).

The Modernization Thesis

The earliest systematic welfare state comparisons were chiefly inspired by the theory that industrialization (accompanied by urbanization, the rise of efficient bureaucracy, and demographic aging) erodes traditional social protection mechanisms, such as the family, and makes public responsibilities necessary. As nations industrialize, they will also converge because social risks become more similar, and because the functional necessity of public social security becomes pressing. Main exponents of this thesis include Rimlinger's (1971) as yet unrivaled historiography of the roots of welfare policy in Europe, America, and Russia, and Wilensky's (1975) quantitative demonstration that rich countries gradually converge in terms of the share of gross domestic product spent on welfare. Wilensky provided no direct measure of political variables, and his conclusions were therefore challenged by adherents of the 'politics matter' position.

The 'Politics Matter' Thesis

The resurgence of conflict theory in the 1970s provoked two antitheses to the modernization paradigm. One was Marxian and quite functionalist, arguing that the modern welfare state was not an egalitarian response to new needs, but rather a mechanism, however contradictory, of sustaining the legitimacy of capitalist economies (Offe, 1972). This perspective has produced no real systematic empirical comparison.

The second antithesis argues explicitly that 'politics do matter' in the sense that the nature, scope, and objectives of a welfare state depend primarily on the balance of political power between 'left' and 'right.' Since it is well established that the roots of welfare policy are largely conservative, as in the case of Bismarck, this position is mainly concerned with welfare state variations in mature industrial democracies. The premise is that the collective resources of wage earners can, through parliamentary democratic means, be mobilized so as to establish a kind of welfare state that effectively modifies the distributional outcomes of markets and the decisional power of capital (Korpi, 1983). This thesis lies close to classical social democratic theory and its postulate that a parliamentarian transition from capitalism to democratic socialism is possible (Stephens, 1979). It is accordingly often labeled the 'social democratic' model of the welfare state. Since Sweden epitomizes social democratic power and achievement, the thesis has been criticized for its inherent swedocentrism.

The power mobilization hypothesis has given rise to not only a very large but also conceptually and methodologically heterogenous literature, which, cumulatively, makes a strong case in favor of the underlying thesis. Whether studying individual social programs, such as pensions, sickness benefits, or labor market policies (Myles, 1984; Korpi, 1989, 1991), or more broadly 'welfare states' (Stephens, 1979; Korpi, 1983; Esping-Andersen, 1990; Hicks, 2000), the impact of the political strength of the left is quite robust.

Still, the 'left power' explanation is questioned by several authors who work within the 'politics matter' tradition. Some argue against a simple, linear left-power effect, arguing that real politics are premised on enduring party coalitions. Postwar Christian Democracy, for example, has been seen as a functional equivalent to labor movements, producing similarly generous welfare states. Esping-Andersen (1990, 1999) and, especially, Van Kersbergen (1995) demonstrate, however, that Christian Democracy produces welfare states that may be generous, and that are also strongly corporativistic in design, very 'familialistic,' and transfer biased. They also tend to be biased in favor of upholding the traditional male breadwinner family model (Esping-Andersen, 2009). In a unique synthesis of the modernization and politics-matter theses, Pampel and Williamson (1989) suggest that all governments, left or right, will favor welfare expansion because population aging implies also median-voter aging.

The Impact of Globalization

The literature is quite divided on how economic globalization influences welfare state development. Following Cameron (1978) and Katzenstein (1985), some adhere to the 'compensation' hypothesis, arguing that strong social guarantees are *sine qua non* in small, open economies that face stiff global competition. Others, however, argue that global competitiveness is harmed by generous social provisions and heavy tax burdens, thus putting downward pressures on welfare spending. For an overview, see Glatzer and Rueschemeyer (2005). There is, so far, no clear consensus on which interpretation is most credible because the precise causal connection has been difficult to identify empirically. Recent scholarship does highlight major and far-reaching reforms within advanced welfare states that seem to challenge Pierson's (1996) path-dependency thesis (see especially Palier, 2010). How these are causally related to globalization remains unclear.

Nation Building and Democratization

The puzzle that any 'politics matter' explanation faces is that the first social reformist steps were usually taken by conservative, often antidemocratic forces, aimed at containing working class mobilization (Rimlinger, 1971; Flora and Alber, 1981). Social policy per se cannot therefore be a sufficient indicator of left-political achievements. The issue is really one of

historical sequence. Most would agree that the historical roots of social policy are closely connected to late-nineteenth century nation building, a means to sustain the established order in the face of emerging social, linguistic, ethnic, or religious cleavages. The 'politics matter' paradigm, in contrast, is concerned with the epoch of actual (and programmatically explicit) welfare state construction in the postwar era of full democracy.

The State-Centered Thesis

Modernization theory stresses the functional necessity of efficient bureaucracy for welfare state development because, without it, effective taxation and administration of complex distributional programs is simply impossible. Bureaucracy, once erected, may also come to play an active role in decision making and thus determine policy evolution. Heclo (1974) maintained that Swedish and British welfare policy, although originally rather similar, eventually parted ways because of inherently different bureaucratic policy-making traditions. The independent powers of public officials in framing key policy decisions have received particularly strong emphasis by Skocpol (1985). This so-called state-centered thesis has proved difficult to test empirically against rival arguments. As a result, empirical applications are limited to select case studies that are difficult to generalize and a handful more systematic, quantitative studies (Korpi, 1989).

The Consequences of Welfare States

The literature on the consequences of welfare state differences is, possibly, as huge as the former but the protagonists tend to be different. For one, there is a long-standing tradition in Economics to study the microeconomic effects of social policies on labor supply as well as their macroeconomic effects on inflation, savings, unemployment, or economic growth. This literature, which, in any case, is rarely comparative, will not be reviewed here (for an overview, see Barr, 1993).

Closer to our concerns are studies that examine the impact of welfare state design on distributional outcomes. Most scholars agree that the focus should be on life chance dynamics, but this has been limited by data scarcity. The literature is therefore dominated by cross-sectional analyses of income inequalities, poverty, and, in recent years, also on gender equality.

Poverty and Income Distribution

The conventional assumption has been that equality is positively and linearly related to welfare state generosity. This was implicit in Wilensky (1975), and some early attempts at empirical verification produced positive and significant correlations (Hewitt, 1977). Until the mid-1980s, however, robust comparisons were hampered by problems of cross-national data comparability. Only with the advent of the Luxembourg Income Study and, later, European Union data collection, did genuinely comparable income data become available.

Most studies of the redistributive effects of welfare states (or single social programs) compare pre- and posttransfer tax distributions (representative recent examples include Smeeding (2005) and OECD (2011)). An extensive review of the literature is found in Esping-Andersen and Myles (2009). For a focus on poverty-reduction effects, see Kenworthy (1999) and Nolan and Marx (2009). A few studies involve more explicitly the impact of types of welfare states or social policies on distributional outcomes (Palme, 2006).

Quantitative comparisons come to a surprisingly consensual view: the larger the welfare state, the greater the degree of equality in net disposable income, and the lower the rates of poverty. A less obvious finding is that welfare states targeted to the truly needy do not result in more poverty reduction than does the universalistic approach to social protection (Korpi and Palme, 1998). Reflecting the paucity of comparable panel data, studies on poverty and inequality dynamics are, so far, few. Preliminary evidence does, however, suggest that comprehensive, generous welfare states are substantially more able to guarantee against long-term entrapment in poverty or welfare dependency (OECD, 2011). This appears to hold also for the recent decades of rising economic pressures (Kvist et al., 2012). Most studies conclude that targeted, means-tested policies create strong poverty traps.

The finding that comprehensive, universalistic welfare states are more redistributive than those favoring heavy targeting has provoked debate about the precise mechanisms involved. Some argue that income tests create stigma and lowers the take-up rate. With their notion of the 'paradox of redistribution,' Korpi and Palme (1998) argue that the basic reason is that comprehensive and universalistic policies crowd out inequality-producing private welfare plans.

Almost all research is limited to the effects of government income transfers and taxation and very little is known about the redistributive effects of services. A rare exception is Marical et al. (2006) who show that welfare state services have a substantial redistributive effect, particularly in service-intensive welfare states like the Nordic (see also Esping-Andersen and Myles, 2009).

Gender Inequality

One has witnessed a rapidly growing literature on the impact of welfare states on gender inequalities, in particular via family policies, such as caring services and leave schemes. Specific policy variations are linked to either welfare state typologies (as in the case of Gornick and Meyers, 2003; Stier et al., 2001), or to male breadwinner typologies (Lewis, 1993), and welfare systems are often ranked in terms of their degree of 'women friendliness' (for an overview, see O'Connor, 1996). A number of rather clear-cut empirical conclusions have emerged. One, day care provision is a powerful precondition for women's economic independence, for their capacity to harmonize careers and family, and also for safeguarding against poverty in single-parent families (Gornick and Meyers, 2003). A second finding is that welfare systems that grant women independent entitlements enhance both individual autonomy and gender equality (O'Connor et al., 1999).

A parallel line of research focuses on demographic outcomes, fertility in particular. The core argument is that gender egalitarian policies are becoming a precondition for decisions regarding motherhood (McDonald, 2006; Duvander

and Lappegaard, 2010). While family allowances seem to have no effect, child care provision promotes fertility significantly (Gauthier, 2007; D'Addio and Ercole, 2005; Neyer and Andersson, 2008) and this is also the case for paternity leave schemes, at least in Sweden (Olah, 2003). The impact of maternity leave seems to have less to do with their duration than with inbuilt job-retention guarantees (Neyer and Andersson, 2008). More generally, the connection between welfare state commitments and gender equalization remains causally ambiguous. As argued by Esping-Andersen (2009), one faces a problem of endogeneity to the extent that such welfare reforms mainly emerge when the revolution of women's roles is quite advanced.

See also: Citizenship, Sociological Aspects of; Citizenship: Political; Family Policy; Gender History; Gender, Economics of; Global Public Policy; Income Distribution and Inequality: Measurement Issues; Income Inequality; Industrialization; Poverty Policy; Poverty in History; Poverty, Geography of; Poverty, Sociology of; Power in Society; Welfare State, History of; Welfare.

Bibliography

Arts, W., Gelissen, J., 2010. Models of the welfare state. In: Castles, F., Leibfried, S., Lewis, J., Obinger, Pierson, C. (Eds.), The Oxford Handbook of the Welfare State. Oxford University Press, Oxford, pp. 569–583.
Barr, N., 1993. The Economics of the Welfare State. Stanford University Press, Palo Alto, California.
Cameron, D., 1978. The expansion of the public economy: a comparative analysis. American Political Science Review 72, 1243–1261.
Castles, F.G., 1993. Families of Nations: Patterns of Public Policy in Western Democracies. Dartmouth, Aldershot.
D'Addio, A., Ercole, M., 2005. Policies, Institutions and Fertility Rates. OECD, Paris.
Duvander, A., Lappegaard, T., 2010. Family policy and fertility. Journal of European Social Policy 20, 45–57.
Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Polity Press, Cambridge and Princeton University Press, Princeton, NJ.
Esping-Andersen, G., 1999. Social Foundations of Postindustrial Economies. Oxford University Press, Oxford.
Esping-Andersen, G., 2009. The Incomplete Revolution. Polity Press, Cambridge.
Esping-Andersen, G., Myles, J., 2009. Economic inequality and the welfare state. In: Salverda, W., Nolan, B., Smeeding, T. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 639–664.
Esping-Andersen, G., Van Kersbergen, K., 1992. Contemporary research on social democracy. Annual Review of Sociology 18, 187–208.
Ferrera, M., 1996. The Southern model of welfare in social Europe. Journal of European Social Policy 6, 17–37.
Flora, P., Alber, J., 1981. Modernization, democratization and the development of welfare states in Western Europe. In: Flora, P., Heidenheimer, A.J. (Eds.), The Development of Welfare States in Europe and America. Transaction Books, New Brunswick, pp. 37–80.
Gauthier, A., 2007. The impact of family policies on fertility in industrialized countries. Population Research and Policy Review 26, 323–346.
Glatzer, M., Rueschemeyer, D. (Eds.), 2005. Globalization and the Future of the Welfare State. University of Pittsburg Press, Pittsburg.
Gornick, J., Meyers, M., 2003. Families that Work. Russell Sage, New York.
Heclo, H., 1974. Modern Social Policies in Britain and Sweden. Yale University Press, New Haven.
Hewitt, C., 1977. The effect of political democracy and social democracy on equality in industrial societies. American Sociological Review 42.
Hicks, A., 2000. Social Democracy and Welfare Capitalism. Cornell University Press, Ithaca, New York.
Katzenstein, P., 1985. Small States in World Markets. Cornell University Press, Ithaca, NY.
Kenworthy, L., 1999. Do social welfare policies reduce poverty? Social Forces 77, 1119–1139.
Korpi, W., 1983. The Democratic Class Struggle. Routledge, London.
Korpi, W., 1989. Power, politics and state autonomy in the development of social citizenship. American Sociological Review 54, 309–328.
Korpi, W., 1991. Political and economic explanations for unemployment. British Journal of Political Science 21, 315–348.
Korpi, W., Palme, J., 1998. The paradox of redistribution and strategies of equality. American Sociological Review 63, 661–687.
Kvist, J., Fritzell, J., Hvinden, B., Kangas, O., 2012. Changing Social Equality. Policy Press.
Lewis, J., 1993. Women and Social Policies in Europe. Edward Elgar, Aldershot.
Marical, D., d'Ercole, M., Vaalavuo, M., Verbist, G., 2006. Publicly provided services and the distribution of resources. OECD Social, Employment and Migration Working Paper 45.
Marshall, T.H., 1950. Citizenship and Social Class. Oxford University Press, Oxford.
McDonald, P., 2006. Low fertility and the state: the efficacy of policy. Population and Development Review 32, 485–510.
Myles, J., 1984. Old Age in the Welfare State. Little, Brown, Boston.
Neyer, G., Andersson, G., 2008. Consequences of family policies on childbearing behavior. Population and Development Review 34, 699–724.
Nolan, B., Marx, I., 2009. Economic inequality, poverty, and social exclusion. In: Salverda, W., Nolan, B., Smeeding, T. (Eds.), The Oxford Handbook of Economic Inequality. Oxford University Press, Oxford, pp. 315–341.
O'Connor, J., Orloff, A., Shaver, S., 1999. States, Markets, Families. Cambridge University Press, Cambridge.
OECD, 2011. Divided We Stand. OECD, Paris.
Offe, C., 1972. Advanced capitalism and the welfare state. Politics and Society 4, 479–488.
Olah, L., 2003. Gendering fertility: second births in Sweden and Hungary. Population Research and Policy Review 22, 171–200.
Palier, B., 2010. A Long Goodbye to Bismarck? Amsterdam University Press, Amsterdam.
Palme, J., 2006. Welfare states and inequality: institutional designs and distributive outcomes. Research in Social Stratification and Mobility 24, 387–403.
Pampel, F., Williamson, J., 1989. Age, Class, Politics and the Welfare State. Cambridge University Press, Cambridge.
Pierson, P., 1996. The new politics of the welfare state. World Politics 48, 143–179.
Rimlinger, G., 1971. Welfare and Industrialization in Europe, America, and Russia. Wiley, New York.
Skocpol, T., 1985. Bringing the state back in: strategies and analyses in current research. In: Evans, P., Rueschemeyer, D., Skocpol, T. (Eds.), Bringing the State Back in. Cambridge University Press, Cambridge.
Smeeding, T., 2005. Government programs and social outcomes. Luxemburg Income Study Working Paper 426.
Stephens, J., 1979. The Transition from Capitalism to Socialism. Macmillan, London.
Stier, H., Lewin-Epstein, N., Braun, M., 2001. Welfare regimes, family supportive policies and women's employment along the life course. American Journal of Sociology 106, 1731–1760.
Titmuss, R., 1974. Social Policy. Allen and Unwin, London.
Van Kersbergen, K., 1995. Social Capitalism. Routledge, London.
Wilensky, H., 1975. The Welfare State and Equality. University of California Press, Berkeley.

Welfare Reform

Torben M Andersen, Aarhus University, Aarhus V, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The fundamental question in welfare system design and thus reforms is how to balance concerns for efficiency with equity. The key issue is how to strengthen employment incentives while maintaining social insurance and ensuring an acceptable distributional profile. The main drivers underlying recent reforms are unemployment problems, changing demographic structures, and strained public finances. The trend and issues in recent welfare reforms are reviewed with a focus on how to make high employment consistent with distributional objectives.

Introduction

Welfare reforms are high on the agenda in many OECD countries at the same time as several emerging economies are contemplating how to develop and extend welfare systems.

The use and understanding of the term "welfare reform" is not unequivocal but depends on the welfare regime (Esping-Andersen, 1990). Thus in political discourses in the Nordic countries the term "welfare" is used synonymously with tax-financed welfare state arrangements including both the social safety net and public service provisions. In Anglo-Saxon countries the term "welfare" refers more narrowly to social benefits targeting the poor. In the following, the broad interpretation is used. Likewise a distinction is sometimes made between social welfare benefits and social insurance, where the former is means-tested and the latter event conditioned, cf e.g., Feldstein (2007). This distinction is not clear-cut since the need for benefits is also event conditioned and all welfare arrangements in general include an implicit or social insurance element. Social–implicit insurance is here used in the broad sense of insurance arrangements embedded in tax financed welfare state arrangements.

The basic purpose of the welfare state is to ensure that individuals are protected against social events which can happen through life and which preclude self-support or in other ways affect life conditions (e.g., health) significantly. The causes of such events can be idiosyncratic or aggregate arising from developments in society (business cycles, structural changes etc.). The financial crisis has brought the issue of the importance of collective risk sharing arrangements to the fore in policy discussions.

A main driver of welfare reforms in recent years has been the problem of ensuring the financial viability of welfare systems. Public finances are in most countries under stress due to high debt levels and large budget deficits, not least due to changing demographic structures. The financial sustainability of tax-financed welfare systems is thus under pressure, although it should be noted that the extent of these pressures is not correlated with the size and scope of welfare systems, but rather reflects differences in political abilities and willingness to undertake appropriate reforms in due time.

Lack of financial viability is an obvious reason why reforms are needed, but structural changes in society, including technological developments and globalization in its broad sense, may also change both the benefits and costs of social arrangements. It is not obvious that past designs of welfare systems provide the right answers to current and future challenges. On the one hand the need for welfare arrangements may increase due to a higher pace for structural changes as well as widening inequalities in labor markets, while on the other hand the scope for and consequences of tax financing may be changing.

The use of the term "reform" is often inflated in political discussions where any policy change, small or large, is denoted a reform with the implicit understanding that it is making a change for the better. Analytically it is useful to make a distinction between systemic, parametric, and regulatory/administrative reforms. A systemic reform involves fundamental changes in the principles underlying policies in a specific area, a parametric reform changes certain parameters (e.g., the replacement rate, eligibility condition) within the confines of the existing system, and a regulatory/administrative reform addresses legal, organizational, and administrative issues. There are very few systemic reforms, especially when it comes to the overall structure of welfare state arrangements. Welfare regimes are very persistent. A recurrent theme is whether there is a race-to-the-bottom in the sense that a minimal welfare state will prevail. Assessed in terms of the overall size of the public sector for OECD countries, there is no sign of such an effect, and the relative size of the public sector across countries has been very persistent for decades. There is a continuous stream of small and large parametric and regulatory–administrative reforms, and it is not possible to provide a reliable and simple quantification of these reforms. Especially tax, labor market, and pension reforms have been frequent in recent years. A distinction is sometimes made between macroeconomic policy and structural reforms, where the former addresses aggregate issues via primarily fiscal and monetary policy, while the latter deals with the specific design of tax, social, and labor market policies. This distinction has become more blurred since structural issues, in particular pertaining to the labor market, are core issues in modern macroeconomics.

Welfare reforms capture potentially quite wide policy areas including social, labor, pension, health, care, housing, and migration. In the following, welfare reforms are discussed in relation to the design of the social safety net, and the question of how to reconcile equity and efficiency, that is, social protection and insurance with high employment, is addressed.

Welfare Regimes

Welfare reforms are often controversial and ignite vivid political debates. Reform agendas are frequently labeled as retrenchments of the welfare state or taken to imply that neoliberal views are gaining ground. In these discussions, it is important to distinguish between reforms induced by changes in political objectives for unchanged fundamentals and reforms driven by changes in fundamentals for given policy objectives. If current welfare arrangements are not financially viable, a quest for reforms may derive from a desire to protect basic social objectives. Not all reforms are retrenchments in the sense that social objectives have changed. In addition economic problems and crises may shift political barriers which have postponed reforms and thus become a vehicle for reforms (crisis-driven reforms).

New insights and knowledge may act as a catalyst for policy changes. If evidence shows that incentives have different effects than (perceived) in the past, it is to be expected that this will lead to policy changes for unchanged policy objectives. An important advance in policy design in recent years is the more systematic use of effect studies. This is reflected in a stark contrast between the current approach to reforms compared with the 1970s and 1980s. In that period many countries attempted to solve unemployment problems by reducing labor supply (early retirement, disability, sickness pensions). In recent years reform initiatives have been taking a more direct approach focusing on strengthening incentives for job search and creation. This shift is driven by a concern that passive policies did not address the underlying causes of unemployment problems and that they may create other social problems (stigmatization, benefit dependency etc.). Moreover such passive policies strain public finances. Therefore the recent focus is on reforms that seek to boost labor supply and hence employment both as an objective in its own right but also as a means to improve public finances.

A key challenge for recent welfare reforms has been how to boost employment without compromising distributional objectives. At the risk of oversimplifying it is possible to make a distinction between two different approaches to tackle this policy dilemma: the targeted approach versus the universal approach. The targeted approach aims at ensuring that help is directed to the deserving based on a 'work first' principle; that is, work should pay and low pay may then be supplemented (e.g., via an earned income tax credit) to ensure an acceptable living standard for those willing to work. The universal approach provides benefits as an individual right to ensure a living standard consistent with distributional preferences, and then attaches various conditions to ensure that recipients are directed toward regular work (employment subsidies, active labor market policies, education etc.). The former approach may be associated with Anglo-Saxon countries and the latter with Northern-European and Scandinavian countries. Both approaches have their pros and cons. The 'work first' model tends to penalize those who have difficulties in working, while the European model will reach those groups but tend to reward those who are less willing to work (enjoy leisure).

These properties seem to suggest a straightforward trade-off between efficiency and equity, leaving different policy positions captured by these two different approaches. However, this is contested by comparisons of the two models in terms of employment and distribution. It is well known that employment rates in Scandinavian countries with strong elements of the universal model are as high as in the US. The universal model is often portrayed as paying people for not working (Rogerson, 2007) or as making work unnecessary (Esping-Andersen, 1990). This effectively amounts to assuming that there is a basic income or demo grant, implying an unconditional minimum income for everybody. Basic facts on employment performance contradict this interpretation. This suggests that there are important facts concerning the design of welfare systems which are crucial for labor market performance, and thus important in discussions on welfare reforms.

Incentives and Insurance

Standard economic theory teaches us that there is a fundamental choice to be made between efficiency and equity. The more equity is pursued via social transfers and taxes, the more economic incentives are distorted (Mirless, 1971).

The distortions imply that employment is too low from a social point of view. This is the cost of ensuring a more equitable output. Sometimes policy debates take the view that distortions are to be avoided, but that would imply that equity is sacrificed. Standard economic theory does not imply that distortions should be avoided, but that they should be balanced to the gains from intervention.

The standard trade-off perception, however, does not take market failures into account. If markets are imperfect and incomplete, social insurance or welfare arrangements may be justified on efficiency grounds. This has wide implications, which is captured by the basic insight that redistribution and insurance cannot be separated (see e.g., Varian, 1980). Any scheme, which ex post based on characteristics like work, health, income, etc. provides entitlements or financial contributions, will *ex ante* serve an insurance function to the extent that there is some risk associated with which states of nature will prevail. If agents are risk averse, and private markets fail to provide sufficient insurance, it follows that there is an efficiency argument for the provision of social insurance. Related, the provision of insurance may not only have a direct welfare effect but may be conducive to flexibility and adjustment, e.g., via better job-matches (Diamond, 1981). Hence, up to some level extension of welfare arrangements may thus improve both efficiency and equity. In the second best context the issue of efficiency and equity effects of various policies is much more complicated.

Any insurance arrangement – private or social – suffers from the common pool or moral hazard problem. The provision of insurance distorts incentives since agents do not bear the full benefits/costs of their actions. In principle there is no difference in this respect between a private insurance scheme with an explicit payment of a premium and an implicit tax financed insurance scheme. The moral hazard problem arises due to a common pool problem and is in this sense generic to any type of insurance (although usually termed a distortion when tax financed). However, this does not imply that moral hazard problems are of no importance, the implication is that the incentive problems arising under social insurance should be evaluated taking into account what the realistic alternatives are.

In this respect is it interesting to observe a contrast – or perhaps even an inconsistency – between the micro and macroperspective on welfare arrangements and their financing. To see this, consider first the problem from an individual perspective in a stylized setting. The net gain from finding a job is $(1-t)w-b$, where w is wage income, t the tax rate and b the benefit level (nontaxable) assumed to be available to all not having a job. This net gain can be written $w-wm$, where the marginal effective tax rate ($METR$) is given as $m=t+b/w$; i.e., the sum of the tax rate t and the replacement rate (b/w). The higher m, the lower the net return to work, and this distorts the incentive structure. Next, consider the public budget focusing only on the revenue from labor income taxation and expenditures on benefits. We have that the budget balance can be written $B=twL-(N-L)b$, where L is employment and N the population (all are assumed eligible for benefits if out of job). The budget can be written $B=wmL+bN$, i.e., the effect on the budget of a change in employment L is wm; i.e., the higher the m, the more sensitive the budget is to changes in employment. In the microcontext a high m is considered a problem since it lowers the net return to work, while in a macrocontext it is praised since it captures the automatic budget response which is generally perceived to be stabilizing (hence termed automatic stabilizers). This points to the tension that reforms strengthening the individual incentive structure may reduce overall stabilization power and thus lead to more risk and volatility (see e.g., Knieser and Ziliak, 2002).

However, the size of automatic stabilizers is not a direct result of macrodesign but rather a byproduct of policy choices in relation to tax, social, and labor market policies. The automatic stabilizers are the net outcome of these choices, and therefore they vary substantially across countries. It is well known that there is a tight correlation between the strength of the automatic stabilizers and the extent of welfare arrangements measured by the size of the public sector. It is implied that it is not possible to make a sharp distinction between macroeconomic policies and structural policies.

Employment and Distribution

The issue of how to balance concerns for efficiency (employment) with distributional objectives can heuristically be thought of in terms of the following three elements.

1. Wage determination: Distributional ambitions determine benefit levels in the broad sense of social assistance, unemployment benefits (duration, eligibility, replacement rate) and minimum wages (either explicitly or indirectly via the level of social and unemployment benefits). It is a robust theoretical result that various wage models (competitive, bargaining, efficiency wages, etc.) all imply that wages depend positively on benefit levels. This may be phrased in the sense that benefit levels determine a floor or base for the entire wage distribution. Ambitious distributional objectives may in this sense preclude 'working poor.'

   The standard approach in the literature is to think of wage compression as something which impairs the market mechanism and thus is associated with inefficiencies. This need not be the case. Decentralized wage setting does not necessarily get close to the textbook model of competitive markets (implying that mingling with the wage structure would cause inefficiencies). Due to entry barriers, informational problems, and many other aspects, decentralized markets may be associated with monopoly and monopsony effects which cause wages to deviate from the competitive benchmark model, in particular violating the premise of equal pay for equal work (see Barth and Moene, to appear). A compressed wage structure including minimum wages may also, in the presence of market failures, provide insurance (Dreze and Gollier, 1993, Agell and Lommerud, 1992). Wage setting also involves other aspects related to employment contracts and employment protection legislation.

   Obviously, a compressed wage structure may also reflect concerns for fairness and distributional objectives (see Barth and Moene, 2013). An important issue is the implication of the 'wage floor' created by benefits since it is of crucial importance for those with low qualifications. There is a classical tension between the importance of the wage for those having a job, and its implication for employment and thus unemployment. This has both demand and supply implications.

2. Labor demand: The 'wage floor' matters the most at the lower end of the productivity distribution (the value of the marginal product of labor). Those with low qualifications are thus at risk of being priced out of the market; that is, individuals willing to work might not find an employer because their productivity is below the stipulated wage level. They are involuntarily unemployed. This problem can obviously also arise for business cycle reasons, but the present discussion is confined to structural issues. This may be denoted the demand problem.

   The demand problem is particularly important if trends driven by new technology and globalization tend to reduce the demand for low skilled. This strengthens the dilemma of matching distributional ambitions (e.g., not accepting working poor) with low unemployment for low-skilled.

   Qualifications may depreciate in nonemployment situations; that is, unemployed risk losing qualifications by not maintaining contact to the labor market (depreciating experience, obsolete qualifications, social problems, dwindling network, etc.). As a consequence there may be a path dependence in individual and thus aggregate employment levels, which tends to show up in terms of long-term unemployment (see e.g., Pissarides, 1992).

   The demand problem can in principle be overcome by offering the employer a wage subsidy. However, it is difficult to target such subsidies to the specific individuals facing the above-mentioned barriers, and there is thus a potentially large deadweight loss. Moreover, subsidies are not a solution to the underlying problem created by the lack of qualifications. Therefore, the applied subsidy schemes are usually associated with training requirements.

   The overall demand for labor is obviously business cycle dependent, implying that the strength of the demand problem facing low skilled may vary countercyclically.

3. Labor supply: The basic aim of the social safety net is to cushion income (consumption possibilities) and ensure that loss of job is not causing a dramatic fall in living standard (insurance). Muting the income decrease in case of job loss inevitably implies that the gain from finding a job becomes small (high METR). This is the incentive or supply problem referring to the problem of ensuring sufficiently strong work incentives.
   It is well established that benefits (and additional means-tested supplements for housing, children, etc.) in combination with taxes can cause rather high METRs. In particular cases, they may exceed 100%. Even if this is extreme there is for broad groups a small economic incentive to be active in the labor market (see e.g., OECD, 2012). This raises the pertinent policy question: Does it pay to work?
   The social safety net affects work incentives via the high METRs. This effect does not only relate narrowly to whether people want to work or not, but more widely to how active they are in job search and which types of jobs they are willing to accept. Jobs have numerous characteristics including wages, working hours, qualification requirements, work facilities, work environment, job task, location (transport time/costs), etc. Thus, in most cases the issue is not whether a given individual is willing to work or not, but rather which jobs one is willing to accept and how active one is in job search. A generous social safety net reduces the urgency of finding a job to earn an income.
   It is a concern that reduced job search incentives may lead to permanent benefit dependency by increasing the distance to the labor market for the above-mentioned reasons related to path dependence (see e.g., Ljungqvist and Sargent, 2008). The social safety net was not intended for permanent income support (except for lack of ability to work justifying disability pension) but as a temporary social protection. An important issue for welfare design and reforms is how to maintain short-term insurance without causing a permanent increase in the dependency ratio. (For an overview, see Frederiksson and Holmlund (2006).
   It has been suggested that weak economic incentives to work can be countered by strong work norms, but also that they can be eroded if a sufficiently large share of the population is receiving benefits (Lindbeck, 1995). It may be contested whether strong work norms can be maintained if the economic incentive structure penalizes norm fulfillment.
   Making work pay has featured importantly in recent welfare reforms via earned income tax credits (working tax credit) which may apply to all with a labor income below a certain threshold or be targeted specific groups (single mothers), see e.g., Immervoll and Pearson (2009). The tax credit contributes to make work pay, and in this way strengthens labor supply along the extensive dimension (number of people working). However, to minimize deadweight losses and public finance implications, the tax credit is usually phased out over some income interval, which in turn increases the marginal tax rate, influencing labor supply along the intensive margin. The design of such programs thus involves a trade-off between labor supply along these different margins (see e.g., Saez, 2002).

Tagging or Targeting – Active Labor Market Policy

The above-mentioned discussion points both to qualification and incentive problems as barriers to unemployment. From a policy perspective, the problem is that it is not easy to identify which barrier is binding for a given individual, especially as this may also be state and duration dependent. There is thus a fundamental issue of both targeting (who should be 'treated') and the instrument (which 'treatment').

So-called active labor market policies – also termed workfare, activation, or employment conditionalities – may be a means to address these problems. By active labor market policies is understood that explicit conditions are attached to receiving unemployment benefits or social assistance benefits. These conditions are not requirements to qualify for the benefit, but requirements to maintain it. It is important to note the difference between employment conditionalities and waiting or qualifying conditions. The former refers to conditions associated with receiving benefits; that is, benefits are not provided unconditionally. Waiting or qualifying conditions imply that there are differentiated rights depending on whether these conditions are fulfilled or not. Strong distributional objectives may call the time-consistency of such conditions in doubt. Examples of such conditions include documented active job search, participation in job search courses, education programs, job-training, etc. These conditions are thus different in nature than the eligibility conditions for various forms of disability pensions, where an external health assessment is required. Note that the conditions are two-dimensional involving both the state-of-nature (out of job) and behavioral variables like active job search, participation in education programs etc.

In some countries – in particular the Scandinavian countries – there is a long tradition for using activation requirements in the social safety net, and they have also featured frequently in recent welfare reforms (see e.g., OECD, 2011). These policy reforms have usually been cast in terms of 'rights and duties'; that is, the individual has a right to some benefit if unable to support herself, but also a duty to be active. Oppositely, society has the right to ask the individual to participate in activation programs in return for benefits, but the duty to provide benefits to those unable to support themselves.

To see how active labor market policies may be used to address both the supply and the demand problem note that they may increase the opportunity costs of claiming benefits but also be designed so as to improve the qualifications of the participants.

Active labor market programs thus release a number of effects. Participation in a program may enhance the qualifications of the participants (direct program effect) but also lower their job search during participation (logging-in effect). The increase in the opportunity costs comes primarily via explicit requirements for active job search and the participation requirement. This may imply that certain groups may abstain from claiming the benefit (the screening effect), which makes the system better targeted, and it may imply that reservation

demands for jobs are lowered and job search incentives increased for unemployed (the threat effects), which reduces moral hazard problems. In addition there may be a general equilibrium effect running via wage formation (wage effect) since the outside opportunity for employed is changed if they become unemployed (and eventually participants in an active labor market program).

The political attraction of attaching active labor market conditionalities to unemployment benefits and social assistance is that incentives can be strengthened without necessarily deteriorating the level of support offered by the social safety net. Economic deprivation is not necessary to create incentives! This can also be interpreted in the sense that there is a complementarity between replacement rates and workfare requirements, or that the total package matters in the sense that a given employment level can be supported by either low benefits or relatively high benefits with strong active conditionalities, see e.g., Andersen and Svarer (2014).

Importantly, active labor market policies are resource demanding, and OECD countries spend up to about 1.5% of GDP on such programs. Hence, even though such policies have the attraction that they may work to maintain employment incentives without jeopardizing distributional objectives, this is not coming costlessly.

Two aspects are particularly important in the practical design of active labor market policies, namely timing and program types. Frontloading of workfare requirements will strengthen incentives the most, but it will also be very costly and entail a large deadweight loss from program participation for many who in any case will find a job after a short unemployment spell. This is particularly so in a labor market with a high incidence of short-term unemployment spells. Hence, workfare requirement should be imposed after some duration of an unemployment spell. The group of unemployed is heterogeneous, spanning from some who have the qualifications and experience making them readily employable to some who lack these properties (e.g., due to long-term unemployment) and therefore find it very difficult to get a job. For the former group, help with job search may be sufficient, while for the latter more specific programs may be needed to specifically address the constraints lowering their job finding rate. In some cases, it may be easy to identify these constraints (e.g., if the unemployed lacks specific skills), while in others it may be more difficult and also depend on market conditions (qualifications become obsolete due to structural changes). In the latter case, the avoidance of deadweight losses gives an argument for making workfare programs duration dependent.

These considerations suggest an optimal profile for active labor market policies where the requirements run from general and relatively costless activities to specific and more costly activities, depending on the duration of the unemployment spell. The duration-dependent sequencing of workfare requirements works to minimize deadweight losses and program costs while maintaining the incentive effects and addressing the more specific programs to groups for whom it may make a difference. At the same time it introduces a forward-looking element into the scheme where the stepping up of workfare requirements strengthens work incentives.

There is a vast empirical literature analyzing the effects of active labor market policies. Most of them focus on the direct program effect and the logging-in effect, but some also look at the threat effect. For a recent survey see Card et al. (2010). The general finding is that there is a logging-in effect. Furthermore, it is found that program effects are strongest from private job-training and that there is a threat effect affecting the inflow into benefit schemes. Note that few studies have explicitly assessed the screening and the wage effect.

Active Distribution Policy

The classical discussion of redistribution has its outset in the distribution of market incomes. Is the distribution fair and is there a need for redistribution? However, the distribution of market income is endogenous and dependent on policies, most notably educational policies. Redistribution based on realized marked income may thus be dubbed passive redistribution policy, while educational policy in the broad sense moving in the direction of a more equal distribution of qualifications is an active redistribution policy; that is, it is forming the distribution of market incomes.

Educational policies involve the issue of the general level of qualification of the work force, but the distribution of qualifications is equally important. A main reason for publicly provided education is to reduce the importance of the parents' ability and willingness to pay for education, and in this way to facilitate equal opportunities. The social gradient in education is strong in most countries.

Cross-country data also show that the countries with low income inequality also tend to have low inequality in educational attainments. Countries with extended welfare states like the Northern European countries are also well known for having a substantial public involvement in education, i.e., tax financed and available on a merit basis.

Strong egalitarian preferences rule out working-poor, and thus minimum wages are high. This implies a higher entry threshold in terms of qualifications to find a job. To ensure equal opportunities and to make an egalitarian outcome consistent with a high employment level, it is thus a requirement that the qualification structure is reasonably compressed. To put it differently, it is impossible to maintain a high employment level for a large group of (low) unskilled workers if distributional objectives are ambitious.

It is a well-known fact that the distribution of wages and thus income have become more unequal in recent years, Atkinson et al. (2011). Although the role of various driving factors like technology and globalization is open for debate, the policy issue is how to counteract this. If the premium to qualifications goes up, the immediate response is to shift the supply in the direction of demand. Historically skill-biases in labor demand have to a large extent been counteracted by changes in supply along the qualifications dimensions.

Educational issues are therefore receiving more attention in recent reform initiatives. This applies both to the general level and quality of education, the entry and exit ages from education (including delays, drop-out rates, etc.) and not least the possibilities of reducing the group receiving no or only little education and training of relevance for the labor market. This

involves in particular younger cohorts since it is well known that it is difficult to 'repair' lack of qualifications at a later stage of life. A related policy area is 'life-long' learning to ensure an adaptive labor force.

Concluding Remarks

The frequency of welfare reforms has been high in recent years. This is driven by structural changes and not least demographic changes in combination with strained public finances. Most reforms are parametric or regulatory and rarely involve large systemic changes. Overall policy objectives and hence welfare state arrangements remain rather persistent across countries, but the more specific policy issues and the way to address them are in continuous change.

The informational base for welfare reforms has improved significantly in recent years due to both theoretical advantages in understanding the drivers underlying the fundamental trade-off between insurance and incentives and important empirical advances via the use of microdata and a more systematic use of effect studies as basis for evaluating and designing specific policies.

See also: Earnings Inequality; Income Inequality; Welfare Policy Comparisons; Welfare Retrenchment; Welfare State, History of; Welfare and Work; Welfare.

Bibliography

Andersen, T.M., Holmström, B., Honkapohja, S., Korkman, S., Söderström, H.T., Vartiainen, J., 2007. The Nordic Model – Embracing Globalization and Sharing Risks. ETLA, Helsinki.

Andersen, T.M., Svarer, M., 2014. The role of workfare in striking a balance between incentives and insurance in the labour market. Economica 81, 86–116.

Agell, J., Lommerud, K.E., 1992. Union egalitarianism as income insurance. Economica 59, 295–310.

Atkinson, A.B., Piketty, T., Saez, E., 2011. Top incomes in the long run of history. Journal of Economic Literature 49 (1), 3–71.

Barth, E., Moene, K.O., 2013. Why do small open economies have such small wage differentials? Nordic Economic Policy Review 1, 139–169.

Card, D., Kluve, J., Weber, A., 2010. Active labour market policies: a meta analysis. Economic Journal 120, F452–F477.

Diamond, P.A., 1981. Mobility costs, frictional unemployment, and efficiency. Journal of Political Economy 89, 798–812.

Dreze, J.H., Gollier, C., 1993. Risk-sharing on the labour market and second-best wage rigidities. European Economic Review 37, 1457–1482.

Esping-Andersen, G., 1990. The Three Worlds of Welfare Capitalism. Polity Press.

Feldstein, M., 2007. Rethinking social insurance. American Economic Review 95 (1), 1–24.

Frederiksson, P., Holmlund, B., 2006. Improving incentives in unemployment insurance: a review of recent research. Journal of Economic Surveys 20 (3), 357–386.

Immervoll, H., Pearson, M., 2009. A good time for making work pay? Taking stock of in-work benefits and related measures across the OECD. OECD Social Employment and Migration. Working Papers No 81, Paris.

Knieser, T.J., Ziliak, J.P., 2002. Tax reform and automatic stabilization. American Economic Review 92, 590–621.

Lindbeck, A., May 1995. Hazardous welfare state dynamics. American Economic Review, Papers and Proceedings, 9–15.

Ljungqvist, L., Sargent, T.J., 2008. Two questions about European unemployment. Econometrica 76 (1), 1–29.

Mirless, J., 1971. An exploration in the theory of optimum income taxation. Review of Economic Studies, 175–208.

OECD, 2011. Employment Outlook. Paris.

OECD, 2012. Employment Outlook. Paris.

Pissarides, C.A., 1992. Loss of skill during unemployment and the persistence of employment shocks. The Quarterly Journal of Economics 107, 1371–1392.

Rogerson, R., 2007. Taxation and market work: is Scandinavia an outlier? Economic Theory 32, 59–85.

Saez, E., 2002. Optimal income transfer programs: intensive vs extensive labor supply responses. Quarterly Journal of Economics, 1039–1073.

Varian, H., 1980. Progressive taxation as social insurance. Journal of Public Economics 14, 49–68.

Welfare Retrenchment

Bernhard Ebbinghaus, School of Social Sciences, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Since the economic crisis of the 1970s, welfare state reform is on the political agenda of advanced welfare states facing multiple economic and sociodemographic challenges. The public debate and initial research (in particular, the new politics thesis) focused on the question why welfare retrenchment, the cutting back of public expenditure, has met much public resistance. More recent research studies the variations in welfare restructuring as complex adjustments to multiple challenges, while seeking to explain these changes through compound factors: ideational, institutional, and interest-based explanations.

In public debate 'welfare retrenchment' commonly refers to the restructuring if not curtailment of social policies. Since the economic crises of the 1970s, advanced welfare states have grown to limits facing severe constraints of permanent fiscal austerity and global economic competition. The term 'retrenchment' dates as far back as the British Liberal manifesto of 1906 in referring to avoid an oversized state. During the 1980s it reappeared in public debates on cutting back state activities, in particular public education in the United States, and since the late 1990s, it became more widely used in comparative studies of welfare state reforms (Pierson, 1994).

Most prominently, Paul Pierson claimed that "(t)he politics of retrenchment is typically treacherous, because it imposes tangible losses on concentrated groups of voters in return for diffuse and uncertain gains" (Pierson, 1996: 145). In this view, politicians adopt a 'blame avoidance' strategy by shunning too radical reforms for fear of electoral backlash. This 'new politics' thesis (Pierson, 1996, 2001) has shaped comparative research of the topic (Starke, 2006), becoming a point of reference like Esping-Andersen's regime typology (1990), which also assumed path-dependent institutional inertia. However, there remains much academic debate over the scope and causes of welfare state restructuring (Bonoli and Natali, 2012). Not least, 'retrenchment' is criticized as a too narrow concept for analyzing the cross-national variations of reform capacity and scope. The term focuses merely on cost containment, while there are also more subtle and long-lasting institutional changes that could be described as 'recalibration' (a term used in Pierson, 2001).

Cross-national and case-oriented research has taken different strategies to investigate restructuring of advanced welfare states in The Organisation for Economic Co-operation and Development (OECD) countries. Quantitative studies of welfare retrenchment often tend to measure cutbacks in expenditure, while others also seek to capture restructuring through changes in social rights (both in respect to benefits and services). Outcome-oriented studies evaluate the impact of policy changes on the life chances of social risk groups and the overall level of welfare in a society (for instance, the poverty rate). Politics-oriented studies, both quantitative and qualitative, analyze the processes of reform, in particular the political conflicts involved in changing the status quo in social policies against vested interests. Qualitative policy-oriented case studies investigate which policy instruments are used and how institutional change restructures welfare states: whether reforms are gradual or radical, whether they involve parametric or systematic changes, or whether they have merely short-term or rather long-term impact.

This article focuses on the restructuring, both retrenchment and recalibration, of advanced welfare states since the 1970s. It will first discuss the main external constraints posed by economic globalization and European integration as well as the internal challenges of sociodemographic changes. It then addresses the problem of measuring welfare retrenchment and recalibration by quantitative indicators. The remainder will sketch different theoretical approaches, in particular the new politics thesis as well as theories of institutional change. While the focus of the early literature was on welfare state resilience, more recently, theoretical and empirical analyses turned to how institutional change alters welfare states. While the study of change processes reveals the multifaceted nature of welfare restructuring, the underlying causes need further explanations. Different theoretical frameworks are advanced to explain the mechanism of change: ideas, interests, and institutions. These consider the role of ideas, political institutions, and power relations in welfare state reform processes. The conclusion raises the open research questions and future policy issue is discussed.

Why Are Welfare States in Need of Change?

The 'crisis' of welfare states has been deplored since the economic downturn of the 1970s. While functionalist modernization theory stressed the 'growth to limits' (Peter Flora) of postwar expansion, neo-Marxist accounts saw a 'crises of crisis management' (Claus Offe) that doomed advanced capitalism. The longer unemployment figures ratcheted upward and public debts spiraled out of control, the more the postwar consensus of Keynesian demand-side macroeconomic management became discredited and a 'paradigm shift' toward neoliberal supply side economics took hold (Hall, 1993). Once thought to be the solution to the pressing economic and social problems caused by industrializing economies, since the economic crisis the welfare state has been deplored by its critics as the cause of all problems: an expensive 'nanny state' that fosters inactivity by work disincentives. This neoliberal political turn occurred most

notably in the United Kingdom with the Thatcher government in 1979 and in the United States with the Reagan administration in 1980, both being prime cases of Pierson's (1994) study of welfare retrenchment. Since then the view that welfare states face 'politics in hard times' due to a fiscal crisis of the capitalist state and 'permanent austerity' has become generally shared (Pierson, 2001).

External Challenges: Globalization and European Integration

Welfare states face major challenges, some exogenous through increased international interdependencies and some endogenous through domestic sociodemographic, economic, and political developments (Scharpf and Schmidt, 2000). The economic crisis following the oil shocks in 1973 and 1979 showed the vulnerability of postwar economic growth and its dependence on favorable global 'embedded liberalism.' Since then globalization is seen as a threat to welfare states, not least since the failure of French Keynesianism in one country under the then new Socialist president Mitterrand in 1982. Global competition led to the threat of relocation of production from high labor cost in Western capitalist countries to lower cost countries, first by the Asian Tigers and since the fall of Communism, more widely by emerging market economies. In particular, high labor costs and taxation in advanced economies are seen as major causes of persistent unemployment and low economic growth.

Politically induced liberalization of capital has also increased the threat of tax competition between high-spending welfare states and its leaner competitors. Advanced welfare states have adapted their tax policies by shifting from mobile to immobile taxes, such as tax and payroll taxes as well as value-added taxes (Scharpf and Schmidt, 2000). However, to the degree that welfare states support investment in human capital, foster employment growth, and provide a buffer for open economies, tax competition among OECD countries might not necessarily lead to a downward spiral. Nordic economies, for example, have been capable of combining export-led high-value growth with a social investment strategy financed by high immobile taxes. A social dumping strategy, feared when opening the single European market, has not materialized to the degree that it can endanger welfare state activities. Certainly, the increased importance of international financial markets and the European Monetary Union have made expansionary policies costly and put pressures on nation-states to control if not curtail their public debt. The recent postcrash sovereign debt crisis in Greece, Spain, Portugal, and Ireland has shown that escalating public debt, partly the result of past welfare state expansion, will provide large pressures by foreign debtors to retrench public expenditure.

Within the European Union (EU), political and economic integration also have an impact on national reform processes. The EU limits the capacity of its 'semisovereign' national welfare states to devise their own social policies (Rieger and Leibfried, 2003). The freedom of labor, goods, capital, and services within the European single market increases market pressures, for instance, on social services, while antidiscrimination laws forbid limiting social benefits to citizens, though legal residence requirements are still possible. While the EU requires the transportability of public social benefits for its cross-border mobile workers, it also enforces international competition and market principles in the private provision of social benefits. Nonetheless, the EU or any other supranational organization has only limited competencies in areas of social redistribution as social solidarity remains largely a field of national politics.

Internal Challenges: Demographic and Socioeconomic Changes

Besides these external factors, internal demographic and socioeconomic changes pose further challenges to welfare states; these may be even more important in driving national reform processes (Pierson, 2001). Most importantly, large-scale deindustrialization and the transition to knowledge economies have aggravated labor markets and further accelerated mass long-term unemployment since the 1970s. Not least as a reaction to the employment crisis, expansion of governmental commitments and welfare state maturation expanded access, benefits, and services, yet this has reached levels of welfare state financing that became politically and economically unviable. Thus there are also domestic reasons for soaring public debt and high nonwage labor costs.

The demographic 'time bomb' of aging societies with low birth rate and increasing life expectancy provides a long-term challenge. Aging will increase the costs of pension systems, in particular in pay-as-you-go financed public pensions: ever more pensioners will have to be supported by an ever smaller working population. The EU's old age dependency ratio is projected to double from 26% in 2010 to 53% by 2060. Also, longer life expectancy is seen as a major advancement of functioning health systems and improved public health, the pressures on rising costs for health care and long-term care, particularly for those reaching the Fourth Age with acute risks of physical and cognitive decay. There are also political implications of an ever-older voting population defending their acquired rights, while the small young generation may not have the political clout to overcome intergenerational imbalances.

Further, socioeconomic changes increase the pressure on welfare states in addition to economic downturn following the 2008 global financial crisis. The requirements of today's competitive knowledge society have made those with low skills particularly vulnerable to unemployment and low income. Moreover, changing family, partnership, and household structures require adaptation in various social policies to respond to 'new' social risks. Increased divorce rates have led to a larger share of households with a single parent and children with elevated risks of poverty. More generally, reconciliation of work and family responsibility is needed through a coordination of family policies at communal and workplace level. Continued global migration also remains a challenge to advanced welfare states, requiring social assistance and education policies to enhance social integration and cohesion.

How to Measure Welfare State Change?

Methodologically, the 'dependent variable problem,' that is, the question how to study welfare state change, poses some complexity (Clasen and Siegel, 2007). As Pierson

admits: "Retrenchment is one of those cases in which identifying what is to be explained is almost as difficult as formulating persuasive explanations" (Pierson, 1994: 13). Welfare state restructuring can be studied quantitatively, it can be measured by its output (such as expenditure) and by its outcome (for instance, rise in inequality), or it may also be investigated more qualitatively by the kind of policies undertaken and the politics involved. While an overall perspective based on aggregate data and major principles has been dominant in comparing welfare regimes, changes may occur at varying scope, in different directions, and at specific timing, in particular social policy programs, thus requiring a more disaggregated and policy-oriented analysis.

Public social expenditure is still used as the main indicator for welfare state cutbacks (see **Table 1**), mirroring the earlier modernization studies of welfare state expansion. Commonly used are time series from international organizations, though these are only partly comparable and complete. The OECD Social Expenditure Database, available from 1980 onward, has been the most widely used; it is integrated into the main political indicator series. Eurostat's European System of Social Protection Statistics and the International Labour Organization's Social Security Inquiry also include financing of social expenditure but differ somewhat in detail. To compare social expenditure across time and space, it is common to calculate relative indicators in terms of economic resources (percent gross domestic product), yet changes may then result also from business cycle effects. A comparison across time and major welfare state clusters (see **Table 1**) for longer periods indicates that social expenditure has continued to grow in most countries despite retrenchment efforts since the 1980s, but a rather stable or low growth can be found in the more advanced welfare states. Research has also shown the need to disaggregate these expenditure figures (see contributions in Clasen and Siegel, 2007). Even if there is no decline in overall expenditure, there might be significant changes in its composition, such as from passive to active labor market policy expenditure. Moreover, using social expenditure as a measure of welfare cutbacks may be misleading without taking into account the underlying need, for instance, the rise in aging or unemployment, which increases demand for benefits and services. Thus when social expenditure stagnates in spite of increasing demand, benefits and services will be on average at a lower level unless there are efficiency gains.

Further problems abound in using readily available social expenditure data. Thus only member countries of international organizations, such as OECD or Eurostat, will be considered for international comparison, limiting the set of comparable cases. A particular difficult issue is the inclusion of mandatory and if possible voluntary private expenditure in order to compare the shift from public to private responsibility. Taxation can also claw back social benefits, and taxing of social benefits varies across countries and over time. According to alternative OECD statistics that also include private expenditure and the net benefit effect, the differences between Nordic high spending and the United States as low spender are much narrower than the public gross social expenditure figures suggest.

Studying Changes in Social Rights

Expenditure as a measure of welfare state activity has been heavily criticized, most prominently by Esping-Andersen (1990: 21): "It is difficult to imagine anyone struggled for spending per se." Instead, he stressed the importance of studying social rights, in particular whether social rights are 'decommodified,' that is, benefits are granted independently from the market. For instance, a shift toward greater equivalence between contributions and benefits would indicate recommodification. A major indicator is thus the replacement rate of unemployment, old age, disability pension, or sick pay benefits in relation to past earnings. But the provision of services, for instance, measured by the size of the public sector employment, could be an indicator of welfare production. Indicators for welfare outcome are the prevalence of poverty, inequality, or nonemployment or other indicators of social exclusion (e.g., the EU's Laeken indicators).

More qualitative considerations of social rights and duties study the changing conditions attached to benefits and services. An example is the reliance on benefit conditionality, for instance, active job search requirements for eligibility of unemployment assistance, which stipulate in case of noncompliance sanctions such as the partial or full withdrawal of benefits. Thus the introduction of 'workfare,' tying benefits to activation measures, has been a hallmark of the paradigm shift in labor market policies, particularly in Liberal countries.

Table 1 OECD social expenditure percent gross domestic product, averages 1980–2013

	1980–89	*1990–99*	*2000–07*	*2008–13*	*Growth index*
Continental Europe (1)	22.2	23.8	24.5	25.9	116.9
Southern Europe (2)	15.7	18.9	21.7	25.5	162.9
Eastern Europe (3)	–	19.9	18.9	20.5	–
Nordic Europe (4)	23.1	28.0	25.9	27.4	118.7
United Kingdom	18.4	19.1	19.8	23.5	127.9
Liberal overseas (5)	15.3	18.1	17.4	19.3	126.7
United States	13.4	14.9	15.8	19.2	143.5
Japan	11.1	13.4	17.8	21.5	194.3
OECD total (6)	16.8	19.2	19.5	21.6	128.2

Notes: Index 2008–13 (100 = 1980–89); Country averages: (1) Austria, Belgium, France, Germany, Luxembourg, Netherlands, Switzerland; (2) Greece, Italy, Portugal, Spain; (3) Czech Republic, Estonia, Hungary, Poland, Slovak Republic, Slovenia; (4) Denmark, Finland, Norway, Sweden; (5) Australia, Canada, New Zealand; (6) OECD total includes all member countries.
Source: OECD Social Expenditure Database (http://stats.oecd.org), OECD, Paris 2013.

The retreat of the state from public provision has also led to an increase in private responsibility; therefore, changes in the provision, governance, or regulation of social protection are of importance in understanding welfare state restructuring. However, statistics on the public–private mix in social protection are difficult to obtain and often depend on legal definition, thus even the OECD publishes different statistics depending on the particular publication or database. For instance, in pension policy, collectively negotiated pensions are counted as mandatory (quasi-public) pensions in some statistics, while as private-funded pensions in others.

New Politics Thesis: Welfare State Resilience through Policy Feedback

The new politics thesis claims that welfare state restructuring in times of austerity differs from reforms during the golden age of expansion. While the extension of social rights allowed credit claiming, in times of fiscal constraints politicians face the hard choice of pursuing unpopular measures, which may lead to electoral defeat. Comparing the politics of retrenchment under Reagan and Thatcher, Paul Pierson (1994) claimed that policy feedback led to a surprising resilience of welfare states, observing no overall retrenchment despite its rhetoric. Indeed, politicians would be reluctant to reform since "cutbacks generally impose immediate pain on specific groups, usually in return for diffuse, long-term, and uncertain benefits" (Pierson, 1994: 18). According to prospect theory, voters are risk averse in respect to future gains, but react more severely against current losses. Given this 'negativity bias' (Pierson, 1996), politicians seek strategies of 'blame avoidance' for fear of upsetting voters.

Whatever the fiscal challenges, welfare state retrenchment remains largely unpopular. Past social policies have led to a strong policy feedback in creating a large popular support from below for all those benefiting from welfare state policies: current and potential beneficiaries as well as public sector employees. Comparative research on public attitudes time and again and largely across very different systems has shown the high popularity of welfare states as a major factor for its persistence. In particular, the role of government in guaranteeing social benefits in old age and providing health services is very popular since all expect to profit from it, followed by programs for families with children. But programs for the long-term unemployed and minimum income schemes for the poor are somewhat less popular since these depend on perceptions of deservingness. Thus in respect to public opinion, the distribution of benefit and costs, the degree of universality of benefits, and the nonconcentration of costs matters for widespread public support.

According to the new politics thesis, politicians pursuing blame avoidance strategies can use several instruments (Pierson, 2001): obfuscating, divisive, and compensatory strategies. The strategy of obfuscation (Pierson, 1994: 22) includes gradualism, which is a stepwise transition instead of sudden changes, such as 'grandfathering' rules exempting current pensioners and senior workers from pension reforms, while affecting future cohorts that retire later. Other forms include technocratic regulation or automatic adaptation mechanisms so that politicians are not held directly responsible. For instance, the Swedish pension reform of 1994 introduced an automatic adjustment of benefits to demographic and economic changes. In addition, burden shifting onto other actors, for instance, from central to local governments or from state to private actors could dilute responsibility. Moreover, strategies of 'divide and conquer' could be used to split opposition. For instance, the UK government sold off public housing to current tenants, while cutting back on new social housing projects to accommodate future people in need (Pierson, 1994). Finally, compensation could be used to combine the extension of social rights with cutbacks of current benefits, thus also using 'credit claiming' strategies (Bonoli and Natali, 2012), for instance, cutting back pension benefits for those without work-related contributions but introducing credits for work interruptions for mothers with children. Indeed, politicians might be able to claim credit for such 'vice into virtue' politics, rebalancing the intergenerational contract.

The central argument of the new politics thesis is that politicians are vote seekers; they pursue blame avoidance because they fear electoral backlash. This is particularly the case in two-party systems and first-past-the-post electoral systems. Yet politicians might be also policy seekers; they are ideologically predisposed in favor or against welfare retrenchment, or they see reform pressures as too large to avoid blame. In fact, unpopular and far-reaching reforms have been occasionally undertaken, and this may not always have led to electoral backlash, particularly not if there were few alternatives available. Studies on the consequences of welfare reforms on electoral politics show a rather differentiated picture. Cross-national comparisons show that conservative parties (religious and neoliberal parties) pursue welfare retrenchment by attempting to claim credit among their voters with some success, while left-wing parties claim credit for social investment policies, particularly in Nordic countries.

Reform Dynamics through Ideational and Institutional Change

Claims of a 'frozen' welfare state landscape by Esping-Andersen (1990) and institutional inertia due to policy feedback by Pierson (2001) are challenged in the more recent literature on institutional change. The all-party consensus on the far-reaching pension reform in Sweden or the paradigm shift toward activation in Danish labor market policy were such examples in the 1990s. Even in the least likely case of Conservative welfare states, 'path departures' have been observed across Continental Europe (Bonoli and Natali, 2012). For instance, the Italian pension reform of 1995 led to the major phasing out of previously generous pension arrangement or the 'Hartz IV' reform of 2004 merged unemployment and social assistance in Germany.

Analytically, institutionalist theory moved from a concern about explaining institutional stickiness to explaining path departure or 'regime change' (Weaver, 2010). After parametric reforms have had only limited effect, more radical welfare state reforms led to systemic changes: transforming the basic principles, altering the instruments, and converting the goals of welfare policies. An important impetus for systemic reforms is

negative feedback, these are "consequences of policy that tend to undermine rather than reinforce the political, fiscal or social sustainability of a particular set of policies" (Weaver, 2010: 137). An example of such a vicious circle is the Continental dilemma (Scharpf and Schmidt, 2000): the spiral of increased labor costs through high payroll taxes that drive labor out of work, leading to passive labor market policies such as early retirement, which then drive social security costs further up.

Ideational accounts have been first applied to the paradigm shift from Keynesian to neoliberal economic thinking (Hall, 1993). Ideas could influence the framing of the problem situation, the alternatives available, and the preferred solution. For instance, the EU's Open Method of Coordination is a political process that relies on peer review and learning from others as instruments to cross-border exchange of ideas. While ideational mechanism in the form of learned and shared cognitive scripts plays a role in ordinary situation, more ideologically loaded world views may be activated in a crisis situation where much is unknown. The role of discourse in policy communities, scientific circles, and public opinion have been studied, revealing various strategies of framing debates on the need to reform and the way of reforming. Adopting a discursive institutionalist perspective (see Schmidt in Scharpf and Schmidt, 2000), various case studies have looked at the importance of public discourse and ideational framing has been studied for welfare state reforms. Analytically and empirically, it remains an open question how new ideas diffuse and gain dominance so that they significantly affect the politics of reform and social policy initiatives. Since discursive framing can be used as a political strategy by politicians and interest groups, the dominance of paradigms or the conflict of ideas points ultimately at political factors.

Recent contributions to understand institutional change have led to taxonomies of different change processes, exemplified by historical case studies often drawn from welfare state reforms (Streeck and Thelen, 2005): layering, conversion, and policy drift. The strategy of institutional 'layering' creates a new institution in addition to existing ones, for example, in Germany, the Riester reforms in 2001 and 2004 introduced voluntary personal pensions beside public pensions, which decline in income replacement capacity. The next process is called conversion, the change of the goal or function of an institution; in the German case the Bismarckian pledge to maintain past living standards is given up and public pensions increasingly can only provide a minimum income in old age. While many changes are induced by public policy, welfare state transformation is often the result of policy drift: the nonaction of governments allows deviations on the ground from the functioning of the system. For instance, the rather silent withdrawal of employers from their defined benefit schemes in favor of less costly defined contribution schemes in the United States and the United Kingdom is such a subterranean transformation.

Welfare Restructuring and Political Institutions

Institutionalist political analyses of reform processes highlight the importance of institutional 'veto points' for interest groupings in influencing if not blocking social policy reforms. Ellen Immergut (1991) showed in a comparative study the varying leverage of veto points in the political systems for the medical professions to influence health care policies. Depending on the concentration or fragmentation of political power in political system in general and the specific policy area, governments have more or less possibilities to alter the status quo. In particular, institutional differences between majoritarian and consensus democracy, and between centralized and federalist polity, and the role of constitutional courts have been singled out. Using institutional veto point indices, comparative studies explored whether systems with more (or less) institutional veto points provide a break (or accelerating) impact on welfare retrenchments (Huber and Stephens, 2001).

More formal political theory focuses on the role of veto players, of individual and collective as well as partisan and institutional ones, in changing the status quo. Reform coalitions depend on the cohesion and congruence of the veto players; the lower the congruence and the higher the cohesion of veto players, the more unlikely is a change of the status quo. While during the expansionary phase a system with many veto players may have slowed down social expenditure, it will provide a break on welfare retrenchment. However, veto player analysis is less amenable to comparative analysis of welfare state development than veto point indicators, but requires more detailed policy case studies that explore the positions and properties of veto players.

Welfare Restructuring and Power Relations

Political-power-related explanations have been the hallmark of explanations of postwar welfare state expansions, but they also help to explain shifts in political coalitions supporting or opposing welfare state retrenchment or restructuring. Clayton and Pontusson refute the new politics thesis: "Societal interests and coalitional alignments have changed, but there is no need to invent a new set of analytical categories in order to explain current patterns of welfare-state retrenchment or restructuring" (Clayton and Pontusson, 1998: 98). Thus institutional change of welfare states can be explained by changing power relations and interest coalitions. While Pierson assumed that "the centrality of left party and union confederation strength to welfare state outcomes has declined" (Pierson, 1996: 151), several critics have challenged this view and stressed that old politics, that is, class conflict (Korpi and Palme, 2003) and the strengths of organized labor, still matters.

Conflict theoretical perspectives, most prominently Korpi's power resource approach, has maintained that left (working class) parties and allied trade unions do not only matter for the postwar expansion of welfare states but also for welfare state 'regress' in times of austerity (Korpi and Palme, 2003). Welfare state retrenchment is seen as a remaking of the postwar societal contract to maintain full employment for which partisan politics and welfare institutions matter. Analyzing the decline in net replacement rates of (short-term) social benefits and the number of benefit cuts across OECD countries (1975–95), Korpi and Palme find that partisan government, welfare state institutions, and constitutional veto point matter in addition to globalization and postindustrial development. Welfare

retrenchment is least likely under conditions of left-wing cabinet participation, encompassing welfare states, and corporatist institutions.

In the debate on the impact of globalization on welfare state development, several political economists (Garrett, 1998; Rieger and Leibfried, 2003) argued that welfare states particularly in 'small' economies were expanded in compensation for opening up domestic economies, and that this compensation is still maintained in times of globalization. Comparative research, both quantitative pooled time series and qualitative case studies, showed that social democratic welfare states have been proved to be highly resilient (Garrett, 1998; Scharpf and Schmidt, 2000; Huber and Stephens, 2001). Others argued that social democratic governments would also reform welfare policies, but that these changes would seek a 'new third way' of employment-friendly social investment policies between neoliberal retrenchment and socialist expansion. Major reforms leading to a paradigm shift in labor market policies were enacted under left party participation in the government in Europe, such as the Danish 'flexcurity' and Dutch activation policies in the mid-1990s as well as the German 'Hartz IV' reform of 2004.

Corporatist analyses also pointed at the fact that trade unions have been successful in blocking major reforms through mass protests if not strike action, most notably the pension reforms in Italy in 1994 and France in 1995. To avoid blockage, some governments have sought to negotiate welfare reforms with the social partners, leading in a few cases to far-reaching 'social pacts,' particularly in the run up to the introduction of the Euro (Avdagic et al., 2011). Instead of the old corporatist exchange of wage moderation in return for an extension of social rights, supply side 'competitive corporatism' in times of austerity entailed that unions could influence the process of restructuring but needed to accept some changes of the status quo. Such corporatist concertation, however, remained highly contingent on the willingness and capacity of governments, employers, and unions to strike a deal that could lead to a positive sum outcome for all.

Outlook

Welfare retrenchment, or the conceptually more open welfare restructuring, has been an emerging topic of public debate and scientific analysis. The welfare state as we know it has been considered under attack since the economic crisis of the 1970s. The economic and sociodemographic challenges have been long known, yet readjustments of welfare states were slow in meeting the requirements of globalization and fiscal austerity as well as the changes brought about with an aging society, modern social relations, and knowledge-based societies. While the welfare state remains highly popular and politicians may fear electoral backlash against welfare retrenchment, the need to readjust social policies cannot be denied. Indeed, multiple short-term parametric and also noteworthy long-term systemic changes have been enacted across OECD countries. The financial crash of 2008 and the subsequent sovereign debt crises have indicated that welfare state restructuring if not retrenchment remains on the political agenda. The political costs of inaction may be in the long run as big as those short-term electoral losses for undertaking unpopular reforms. What seem to be missing in recent years are efforts to find broader political and societal consensus on the direction and scope of welfare state restructuring, as had happened in some social pacts in the 1990s.

The current development of research indicates that the dependent variable, the methodological approaches, and explanatory frameworks have become more differentiated. In order to understand the complex process of welfare state restructuring, more detailed specialized studies are needed on different social programs, the public–private mix, the conditions of social benefits, the quality of social services, and the welfare outcomes. However, this has led to a rather bewildering puzzle of confirming and contradicting findings. Moving away from the single theoretical framework of the new politics thesis, alternative explanations are now available across a wide field of ideational, power interest, and institutionalist approaches, albeit without much theoretical integration. Although the phenomena under study has been observed for over three decades, our knowledge about the causes, conditions, mechanism, and outcomes of the dynamics of welfare state restructuring in times of permanent austerity has been still scattered. More systematic synthesis of empirical findings and explanatory successes is needed to allow more forward-looking policy recommendations. In any case, the need for and efforts toward welfare restructuring will stay for a long time to come.

See also: Welfare Economics; Welfare Policy Comparisons; Welfare State, History of; Welfare.

Bibliography

Avdagic, S., Rhodes, M., Visser, J. (Eds.), 2011. Social Pacts in Europe: Reemergence, Evolution, and Institutionalization. Oxford University Press, Oxford.

Bonoli, G., Natali, D. (Eds.), 2012. The New Welfare State in Europe. Oxford University Press, Oxford.

Clasen, J., Siegel, N.A. (Eds.), 2007. Investigating Welfare State Change: The 'Dependent Variable Problem' in Comparative Analysis. Edgar Elgar, Cheltenham, UK.

Clayton, R., Pontusson, J., 1998. Welfare-state retrenchment revisited: entitlement cuts, public sector restructuring, and inegalitarian trends in advanced capitalist societies. World Politics 51 (1), 67–98.

Esping-Andersen, G., 1990. Three Worlds of Welfare Capitalism. Princeton University Press, Princeton, NJ.

Garrett, G., 1998. Partisan Politics in the Global Economy. Cambridge University Press, New York.

Hall, P.A., 1993. Policy paradigms, social learning, and the state. The case of economic policymaking in Britain. Comparative Politics 25, 275–297.

Huber, E., Stephens, J.D., 2001. Development and Crisis of the Welfare State: Parties and Policies in Global Markets. University of Chicago Press, Chicago, IL.

Immergut, E.M., 1991. Institutions, veto points, and policy results: a comparative analysis of health care. Journal of Public Policy 10 (4), 391–416.

Korpi, W., Palme, J., 2003. New politics and class politics in the context of austerity and globalization: welfare state regress in 18 countries, 1975–95. American Political Science Review 97 (3), 425–446.

Pierson, P., 1994. Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment. Cambridge University Press, New York.

Pierson, P., 1996. The new politics of the welfare state. World Politics 48 (2), 143–179.

Pierson, P. (Ed.), 2001. The New Politics of the Welfare State. Oxford University Press, New York.
Rieger, E., Leibfried, S., 2003. Limits to Globalization. Polity Press, Cambridge.
Scharpf, F.W., Schmidt, V. (Eds.), 2000. Welfare and Work in the Open Economy, 2 vols. Oxford University Press, Oxford.
Starke, P., 2006. The politics of welfare state retrenchment: a literature review. Social Policy & Administration 40 (1), 104–120.
Streeck, W., Thelen, K., 2005. Introduction: institutional change in advanced political economies. In: Streeck, W., Thelen, K. (Eds.), Beyond Continuity Institutional Change in Advanced Political Economies. Oxford University Press, Oxford, pp. 1–39.
Weaver, K., 2010. Paths and forks or chutes and ladders? Negative feedbacks and policy regime change. Journal of Public Policy 30 (2), 137–162.

Welfare State, History of

EP Hennock, University of Liverpool, Liverpool, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16439–16445, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

In England, the term 'welfare state' dates from the 1940s. It has stronger associations of paternalism in American English. How is 'welfare state' translated into German? Conflict exists between the terms and *Sozialstaat*. The definition as per Briggs is influenced by the events of the 1940s. The history of the 'Origins of the Welfare State' is illustrated by two versions, the British and the German, that indicate differences in periodization and focus, and suggest the influence of political considerations. A brief reference to the Swedish version illustrates the same point. A general history of 'the welfare state,' as distinct from the history of individual welfare states follows. Is it located in the wider context of the history of the state? Points considered are unification of the law, establishment of bureaucracies, elimination or control of intermediate forms of authority, and its characterization by new systems of domination peculiar to it. The role of representative democracy is also considered. The connection with the development of industrial capitalism is analyzed in terms of the problem of 'externalities' as well as the role of combinations and transfer payments, of public provision of goods and services, and of regulation. The history of welfare states since 1945 with a few generalizations is followed by a look at the effect of major changes in the international economy. The article concludes with an examination of recent influences on the historiography of welfare states such as a gender-conscious historiography, welfare state history 'from below,' studying the 'mixed economy of welfare,' and comparative studies.

The Origin and Use of the Term Welfare State

The term was coined in England in the 1940s. It is to be found in *Citizen and Churchman* (1941) by William Temple, Archbishop of York, where it is contrasted with the Power State or Warfare State. After the publication of the Beveridge Report in 1942, the term soon came to be applied to the hopes for postwar reconstruction and by 1948 was used regularly for the postwar reforms. The meaning of 'welfare' had not suffered the deterioration that had afflicted the word 'charity.' Although Beveridge himself disliked the term for what he called its 'Santa Claus' connotations and preferred the older term 'social service state,' 'welfare state' normally carried a favorable meaning.

Welfare had a different set of connotations in American English. 'Public welfare,' as understood in the United States then and since, excluded the elements of social insurance, with its claims to benefit as of right, which were so prominent in the British construction of the term (Lindeman, 1937). 'Welfare state,' once imported from Britain, thus carried a stronger association of paternalism (Skocpol and Ikenberry, 1983).

In that respect, Germany was closer to the United States. In the early years of the Federal Republic, when many English terms were naturalized into German, it seemed natural to render 'welfare state' with the similar word *Wohlfahrtsstaat*. Such similarities can be misleading. *Wohlfahrt* was associated with charitable gifts, and *Wohlfahrtsstaat* had acquired a pejorative connotation during the crisis of the Weimar Republic, as when in 1932 Chancellor von Papen accused politicians of the early years of the Weimar Republic of attempting to turn the state into a *Wohlfahrtsstaat*, and thereby weakening the moral strength of the nation. The Federal Republic wished to return to the Weimar tradition in social policy without falling foul of the associations of the term *Wohlfahrtsstaat*. One possible course was to use the term in its positive English sense and to resort to *Versorgungsstaat* to convey the pejorative meaning. The alternative was to revive a nineteenth-century term, *Sozialstaat* (Ritter, 1989). The issue has never been settled decisively, although *Sozialstaat* is the more usual.

If the term dates from the 1940s, its definition also owes much to the changes of that period. According to Asa Briggs, "A welfare state is a state in which organized power is deliberately used (through politics and administration) in an effort to modify the play of market forces in at least three directions – first by guaranteeing individuals and families a minimum of income irrespective of the market value of their work or their property; second, by narrowing the extent of insecurity by enabling individuals and families to meet certain "contingencies"…; and third, by ensuring that all citizens without distinction of status or class are offered … a certain agreed range of social services" (Briggs, 1961). These are usually taken to relate at least to medical treatment, housing provision, and education. Briggs added a reference to the best standards available, but it has since been generally accepted that such an aspiration is unrealistic.

The 'Origins of the Welfare State'

Such an emphasis on the 1940s leaves the historian with the task of relating these innovations to what had gone before, i.e., to consider the 'origins of the welfare state.' Three national examples briefly indicate how differently the origins of the welfare state have been constructed.

British Historiography

Since 1961, this has operated in terms of the transformation of old institutions by new principles, forged at moments of political crisis (Fraser, 1973). The first of these institutions was the Poor Law, which established the responsibility of each local parish for the relief of the poor (1603) and was reformed in 1834 on more explicitly deterrent lines in response

to population growth and agricultural change. The subsequent history has been presented in terms of new social problems caused by industrialization, the Victorian origins of collectivism in public health and factory regulation, the crisis of the Poor Law at the end of the nineteenth century, and the consequent Edwardian legislation (1905–14). This put an end to the possibility of a gradual extension of Poor Law services on a nondeterrent basis and provided for needy and deserving groups in other ways. It introduced free school meals, a school medical service, tax-financed old-age pensions for the poor, and national health and unemployment insurance, supported by a national network of labor exchanges. Alongside these measures, 'New Liberalism' articulated a theoretical justification for state intervention in society and economy. The scope of this legislation was limited in respect of population covered and of benefits provided, but it has been regarded as having laid 'the foundations of the British welfare state.'

The interwar years saw housing provision added to the prewar measures, but social policy responded slowly and piecemeal to economic and social problems. It adapted rapidly once World War II had created new problems and a new spirit of citizenship. The postwar legislation, roughly but far from exactly based on the proposals of the Beveridge Report (1942), universalized national insurance, extended its benefits to cover a greater range of contingencies ('from the cradle to the grave'), and established a mainly tax-financed national health service. This was accompanied by tax-financed child allowances, the extension of housing provision, and free secondary education. Fundamental was a government commitment to a policy of full employment, based to some extent on a belief in the efficacy of Keynesian demand management that was, however, far from unequivocal. The transformation of old social services by new principles of universalism had been presented in the Beveridge report as a characteristic British revolution. Historians followed a similar line, which fitted into the then dominant 'Whiggish' historiography with its emphasis on the continuity of British history and the adaptability of its political leaders. In the influential view of the sociologist T.H. Marshall, the welfare state had achieved the extension of citizen rights from civil rights via political rights in the past to a new body of social rights (Marshall, 1950). The blueprint of the welfare state as established in the 1940s was for a long time presented as a major British contribution to postwar social reconstruction in the wider world.

German Historiography

This has located the origins not merely of its own welfare state but of the welfare state as such in the establishment of compulsory insurance for workers by the German Empire under Bismarck in the 1880s. It has both emphasized the pioneering role of Germany and the continuity of its insurance institutions despite drastic political change. This emphasis on the role of Bismarck, notorious for his hostility to the claims of the working-class movement, as the originator of the welfare state has presented historians with the problem of reconciling these two aspects of his policy. Until recently, it was suggested that wooing the workers with the benefits of social policy was an intrinsic aspect of Bismarck's attempt to suppress social democracy. This view had the merit of presenting the welfare state as the achievement of the democratic forces while recognizing that the policies usually were not what these had been demanding. The publication of documents on Bismarck's social policy has cast doubt on this view of his aims and there is a need for reinterpretation (Tennstedt and Winter, 1993; 1995).

The transition from the establishment of institutions with limited scope to the conscious acceptance of the people's welfare as an essential object of the state is located by German historiography in the Weimar Republic. It found expression not only in the text of the constitution but also across a range of its institutions, several of which had developed from war-time innovations. Collective bargaining backed by compulsory arbitration, subsidized housing, the preservation of prewar social insurance for both blue- and white-collar workers, and unemployment insurance as part of labor market policy, all suggest that the Weimar Republic was a welfare state in deed as in word. In German history, the breakthrough to the welfare state was therefore achieved not after World War II but after World War I. However, the Weimar welfare state was unable to live up to its expectations (Abelshauser, 1987). The political acceptance of its new role was insufficiently broadly based. As the earlier quotation from von Papen suggests, repudiation of the Weimar Republic in the years of economic crisis was connected closely with its social policy.

Although marginalized and only just saved from drastic changes in 1942, the social insurance system survived the Nazis. In the immediate postwar years of economic ruin, forced migration, and impoverishment, it faced perhaps its greatest test. All accumulated funds had been lost. In the Russian zone of occupation, it was reconstructed radically along lines originally proposed by the left-wing critics of Weimar policy. It was no more sacrosanct in the West. The French quickly followed the Russian example. Drastic British proposals followed, based partly on Social Democratic ideas from the Weimar years, partly on Beveridgean lines. They involved retrenchment and thereby aroused opposition well beyond the threatened vested interests. Opposition rallied around the defense of the Bismarckian inheritance in social insurance, which was presented as one of Germany's proud contributions at a time when the nation badly needed a boost to national self-esteem. With the onset of the Cold War, the proposals were dropped as the Western occupation powers gave priority to working with German public opinion. It was left to the Federal Republic to deal with the problems of funding social security; under CDU leadership the traditional structures were almost completely reinstated. Significant reform was delayed until 1957. In that year, the introduction of the 'dynamic' pension, linked to average wage levels, represented an imaginative response to the problems of the Federal Republic, not just the problem of inflation but also that of the low pension levels, which contrasted sharply with the rising standard of living enjoyed by the working population. At a time when the legitimacy of the Federal Republic was fragile, this pension reform contributed significantly to the acceptance of the regime. In 1958–61, proposals to remodel the sickness insurance system and contain its spiraling costs were defeated. The Bismarckian prototype had acquired a secure political profile.

Thus, in the German Federal Republic the 1940s was a period of defense and restoration. The emphasis on the

significance of the Bismarckian reforms for the long-term development of the German welfare state and the tendency to identify it narrowly with contributory social insurance were forged in the political conflicts of those years. This emphasis was reinforced in the early 1980s by the centenary celebrations of the Bismarckian reforms. Much of the German historiography of the welfare state was created in the 1980s (Ritter, 1983, 1986, 1989).

These examples indicate that the history of the welfare state has been perceived differently in different countries, in terms of periodization and of emphasis. This reflects not only its actual development in different states but certainly also the political purposes that historiography has served.

The Swedish Model

These points must be made more briefly for the influential Swedish model of the welfare state. There, it was developments in the 1930s that had the significant long-term effects that they singularly failed to have in Britain or Germany. The expansion of the earlier limited social security provisions took a significantly different form. Unlike Germany, where insurance contributions and benefits differed according to levels of income, Swedish benefits were paid, as in Britain, at a flat rate for all. But they differed from both the German and the British provisions in that retirement pensions and sickness insurance were financed from general taxation. They were set at a level that provided pensioners with an adequate income, something that the British pensions, tied to contributions that even the lowest paid could afford, entirely failed to do. Significant differences between the policies of these three welfare states also extended to industrial relations and taxation.

The Swedish welfare state has been presented as the achievement of class mobilization and labor solidarity. Standard accounts have stressed the long period of Socialist-dominated government and cooperation between government and labor organizations in the shaping of labor market policies (Baldwin, 1990).

A General History of 'The Welfare State'

If it exists at all, the history of the welfare state, as distinct from the history of individual welfare states, must be sought on a higher level of generalization. A welfare state is a state of a certain kind, distinguished from earlier kinds by the addition of a new set of objectives and institutions. Its history must be sought within the wider context of the history of the state. It presupposes the unification of law across the state territory, the establishment of bureaucracies, and the elimination or control of intermediate forms of authority. It is, however, also characterized by new systems of domination, consisting of distributing elites, service bureaucracies, and social clienteles. The process by which these objectives were both established and extended to include the whole nation has been driven by the advent of representative democracy, and it also drew on older concerns to maximize the human resources of the state. In addition, these new state objectives have been connected intimately with the development of modern industrial capitalism.

The state became a welfare state because it dealt increasingly with the social consequences of the particular way in which modern industrial capitalism was established. These consequences, often described as externalities, resulted from the narrow definition of the legal obligations of capitalist entrepreneurs, which contrasted greatly with the obligations imposed on entrepreneurs in the older corporatist economy. This emancipation of the entrepreneur was a deliberate act of state policy, undertaken in the interest of increasing 'the wealth of nations' and thus the power of states over against other states. It undoubtedly had that effect. Enterprise took new and unforeseen forms once it was freed from old regulations. To limit obligations toward workers to the short-term purchase of their labor power, toward the community to the payment of local taxes, and to resort to competition without responsibility for those driven out of the market, all this encouraged innovation and facilitated capital accumulation. But it created problems that resulted from economic decisions while being external to the economic process.

Combinations and/or Transfer Payments

These problems could be dealt with in at least two ways. (1) By combination, intended to limit free competition over wages, or to pool the economic resources of individuals, or both (trade unions and mutual aid societies). To what extent that was possible depended on the attitude of the state and on the capacity of individuals for organization. (2) By state action in whatever territorial units – national, provincial, or local – it exercised its power. That meant taxation and a limited redistribution of the resources originally distributed by the market. The administrative capacity that made it possible to introduce graduated taxes and differential tax allowances as instruments of redistribution is an important aspect of this history.

The legalization of combinations to increase wages in turn made it increasingly possible for individuals to satisfy their needs through the market, thereby creating opportunities for new kinds of enterprise. The legalization of mutual aid associations created structures of redistribution, which could operate as alternatives to the bureaucratic structures of the state. In countries in which the development of bureaucracy had not advanced far, or where it was viewed with mistrust, subsidies from taxes became a way of making mutual associations more effective. Such subsidies were not unconditional but subjected the association to a greater or lesser degree of control. An extreme example of such control was the compulsory establishment of combinations on the order of the state, such as the German sickness funds and trade associations to administer accident insurance. The higher the level in the administrative hierarchy on which the state operated to redistribute resources, the greater the degree of uniformity it could impose.

Goods and Services

So far, the focus has been mainly on the redistributive role of the welfare state. The problem of externalities has also been addressed by the public provision of goods and services, which the market supplied only in ways deemed politically unacceptable. Such goods and services, considered necessary for the common good as politically defined, have included the

infrastructure of public health, the provision of healthy and environmentally approved housing, of medical treatment, and of schools, but this list is far from exhaustive. Here, too, subsidies could take the place of full state provision, e.g., in respect of houses, food, or schools.

Regulation

Since externalities had been allowed to arise because the state had regulated the conditions of production in minimal ways, it also began to introduce new regulations. Under such regulations, the market was permitted to supply goods and services only on conditions acceptable to the state. Instead of providing housing, the state regulated the conditions of tenancies. Instead of running its own schools, it regulated and inspected those provided by other bodies. And, as in Thatcherite Britain, it regulated the sanitary infrastructure while relying on the market to provide it. Such regulation could be accompanied by subsidies. However, such regulation suffers from the same considerations that applied to the general conditions of production in the first place. By regulating the actions of the market, the state is liable to hinder innovation and to make the market an inefficient mechanism for economic growth.

Finally, the state has been able to regulate the economy itself in such a way as to minimize externalities. Examples are the state regulation of wages and the regulation of the economy so as to reduce cyclical fluctuations. At all stages of this process, the question has to be faced whether the object is being achieved at the cost of economic innovation and efficiency. For that reason, there are limits to what regulation can achieve.

A radical alternative has been for the state to take over the running of the total economy. That rules out a welfare state, defined as a state that modifies the play of market forces, and it creates a state so different that it is better not to describe it as a welfare state. That is not to deny, however, that 'communist' states provide social services and organize transfer payments in the interest of social security. This article does not deal with the history of states with a command economy, confining itself to the history of what is sometimes called the liberal welfare state.

Even there the importance of the welfare elements of the state can vary according to the extent to which entrepreneurs did in practice externalize their costs. Where company welfare was built into the entrepreneurial ethic, as in Japan, state expenditure on social security has been restricted compared with other advanced industrial countries.

The History of Welfare States since 1945

Despite this attempt and others (Flora and Heidenheimer, 1982) to construct explanatory models of the welfare state as a general phenomenon of European and subsequently extra-European history, the established tradition of nation-state history has led historians to study developments both before and after 1945 in terms of individual states. They have given prominence to the political circumstances under which social policy was constructed and emphasized the importance of national traditions.

Some few generalizations across countries are feasible. Thus, the experience of the 1930s had demonstrated that institutions of social security were seriously vulnerable to periods of mass unemployment. Such periods greatly increased the demands for transfer payments, while also reducing the income from which these were made. This was particularly true of contributory insurance, and it also applied to finance from general taxation. This lesson had been learned, and in the 1940s, macroeconomic policy became a policy component of every state, whether it had passed through the experience of a war economy or, like Sweden, had not. In some countries, this period saw major reforms and extensions in social security, as in Britain, France, and Sweden. In others, major reforms had to wait until the 1960s and early 1970s after the restructuring of the economy had yielded its rewards, as in The Netherlands and Italy.

Even where significant legislative innovation occurred in the immediate postwar years, subsequent expansion of the economy provided the setting for a period of further change in the 1960s and 1970s. Indeed, the most convincing historical generalizations that can be made about the welfare states since 1945 refer to the effect of major changes in the international economy. A long period of economic recovery in Europe led by the 1960s to a remarkable growth in national income. This in turn both raised expectations and facilitated the introduction of more generous provisions. Levels of benefits were raised, flat-rate pensions complemented by graduated ones that mirrored differences in original earnings and appealed to the better off, while those disadvantaged by or excluded from earnings-based insurance systems received compensatory entitlements.

By contrast, the setback to economic expansion in the mid-1970s, triggered by the oil price crisis, caused a fall in employment, except where special measures for job creation had been implemented (as in Sweden). This, in turn, led to rises in social security payments and falls in the revenue available for the purpose. Steady increases in the gross national product, which had spurred ambitious changes in entitlements in the previous decade and a half, could no longer be assumed. Nor had recent rapid growth in social expenditure been due only to deliberate changes in entitlement. It had also been caused by such demographic, cultural, and technological changes as an increase in the proportion of the elderly, a rise in single-parent families, and the growing cost of medical treatment. To the fiscal crisis that occurred in every country sooner or later, even in Sweden, was added a crisis of belief in the assumptions on which welfare states had been based. Yet these assumptions had been deeply implanted in the citizens, and many vested interests created. Measured by the radical debate that broke out, the actual changes in welfare states were not great, even in Great Britain, where the call for the rolling back of the state had been loudest.

Historiography of Welfare States

There are several ways in which the original traditions of welfare state history have been modified.

A Gender-Conscious Historiography

Just as the feminist movement has since the 1960s exerted a political influence on the agenda of policy making itself, so

it has also changed historical perceptions. Feminist historians have highlighted the way in which social policies were shaped by assumptions about the gender divisions of labor, power, and social responsibility. They have shown how little the basic assumption that each adult woman was dependent on a male earner was borne out in reality. Contributory insurance, with entitlements dependent on the performance of waged labor over most of the adult life span, had been at the center of welfare state history. It looked very different through the eyes of feminist historians, but these have not merely confined themselves to criticism. By drawing attention to policies in support of motherhood, they have reconstructed the outlines of welfare state history (Pedersen, 1993; Koven and Michel, 1993). By making historians more aware of the implications of gender, this emphasis has yet unrealized potentials. Male roles, once so taken for granted as to be largely invisible to historians, have in turn become objects of historical study.

Welfare State History 'from Below'

The history of social policy traditionally has been written through the eyes of 'policy makers' and administrators, and occasionally of political philosophers, trade union officials, and radical politicians, authority figures all. Welfare state history has been particularly resistant to the search for history as experienced by those without power and influence. Yet the construction of social policy and its administration is that of a relationship, in which the recipients of 'benefits' are at least half the story. Like all forms of history from below, attempts to recreate their experience has been hindered by their inarticulacy, at least in forms that have survived as records. But once historians search, they are able to find evidence, not merely for recent times but also at least in England going back well into the past (Lees, 1998; Vincent, 1991).

Studying the 'Mixed Economy of Welfare'

One result of paying more attention to the lives of the poor has been to downgrade the importance of state provisions and to draw attention to the mixture of resources that made up the economy of the poor. Since public authorities frequently did not intend to provide those in need with a full subsistence – and, when they did, were not good at judging the amount required – this should be unsurprising. It has made this history less state centered (Katz and Sachsse, 1996; Thane, 2000). This trend has been reinforced by a change of social policy in several countries in the 1980s, Britain and the United States in particular. The replacement of direct state provision of services by contractual relationships with charities and profit-making providers has given prominence to the concept of the 'mixed economy of welfare.'

Third, at a time when the clients of the welfare state are drawn from all citizens, contemporary studies of social policies have come to pay attention to the mixture of resources at the disposal of citizens, particularly of the elderly, whether these were from public, occupational, or commercial sources. Since that mixture differs in different countries, comparative studies have moved from an exclusive preoccupation with state provision toward a concept of welfare regimes more broadly defined (Esping-Andersen, 1990). The influence of comparative studies by social scientists on historiography is the final issue to be considered.

Comparative Studies

The historical study of welfare states was for a long time overparticularized. Attempts to identify categories suited for generalizations were left to social scientists specializing in the quantitative analysis of policy outcomes. This has been a dynamic field of study since the 1970s, but it developed in a particularly ahistorical form. One reason was its dependence on information provided by such international agencies as the United Nations and the Organization for Economic Co-operation and Development. Little of this went further back than the 1960s, so that the historical perspective was limited. Second, the questions asked and the explanatory power of the answers provided were constrained by the availability of such quantitative evidence, which was at first largely limited to figures of expenditure. The narrowness of these comparative studies was only slowly overcome as new evidence became available, e.g., the Luxembourg Income Study in the 1980s, which provided internationally comparable income-distribution data at microlevel, and as researchers began to approach national governments and other agencies directly to construct their own database. Third, the choice of appropriate categories for comparison across different national contexts is notoriously difficult and not always successful.

Evidence of this kind was, in any case, unable to provide explanations that advanced beyond the discovery of correlations. Intentions were ignored, because unsuited to standardization, and important dimensions of history were thereby ruled out of court. The most important result of these correlations was to group welfare state regimes into three main types: liberal, corporatist, and social democratic. It is more obvious that these describe ideal types than that they describe clearly differentiated clusters of indicators. Australians have demurred at the inclusion of their country in the same category as the United States; Germans have felt unhappy at being grouped with Italy and France; and the case of Britain is recognized widely as a hybrid. It is certainly not possible to write anything like a common history of the several welfare states grouped into any one of these types, with the possible exception of the four social democratic ones, who all happen to be Scandinavian.

Yet, by their very limitations, comparative empirical studies have helped to identify so-called anomalies and to encourage the intensive historical investigations of individual cases (Castles, 1989, which includes Japan and Israel). The gulf between the social scientific and the historical study of welfare states has narrowed, as indicated by sharply focused comparative histories of pensions and of public health or by a five-volume history of individual welfare states addressing similar questions and written on a common editorial plan (Baldwin 1990, 1999; Flora, 1986/2000).

See also: Capitalism; State, History of; Welfare Policy Comparisons; Welfare State, History of.

Bibliography

Backhouse, R.E., Nishizawa, T., 2010. No Wealth but Life: Welfare Economics and the Welfare State in Britain, 1880–1945. Cambridge University Press, New York.

Beck, H., 1995. The Origins of the Authoritarian Welfare State in Prussia: Conservatives, Bureaucracy, and the Social Question, 1815–70. University of Michigan Press, Ann Arbor.

Birch, R.C., 1981. The Shaping of the Welfare State. Longman, Harlow.

Börsch-Supan, A., 2011. The Individual and the Welfare State: Life Histories in Europe. Springer, Berlin.

Brown, J., 1995a. The British Welfare State: A Critical History. Mass: Blackwell, Oxford; Cambridge.

Brown, J., 1995b. The British Welfare State: A Critical History. Blackwell, Oxford.

Cohen, S., 2001. Immigration Controls, the Family, and the Welfare State: A Handbook of Law, Theory, Politics, and Practice for Local Authority, Voluntary Sector and Welfare State Workers and Legal Advisors. Jessica Kingsley Publishers, London.

Dutton, P.V., 2002. Origins of the French Welfare State: The Struggle for Social Reform in France, 1914–1947. Cambridge University Press, Cambridge, UK.

Eisenstadt, S.N., Ahimeir, O., 1985. The Welfare State and its Aftermath. Barnes & Noble Books, Totowa, NJ.

Fraser, D., 1984. The Evolution of the British Welfare State: A History of Social Policy since the Industrial Revolution. Macmillan, Basingstoke.

Fraser, D., 2000. The Welfare State. Sutton, Stroud.

Hill, M.J., 1993. The Welfare State in Britain: A Political History since 1945. Edward Elgar, Aldershot.

Hong, Y.-S., 1998. Welfare, Modernity, and the Weimar State, 1919–1933. Princeton University Press, Princeton, NJ.

Jones, M., Lowe, R., 2002. From Beveridge to Blair: The First Fifty Years of Britain's Welfare State, 1948–1998. Manchester University Press, Manchester.

Katz, M.B., 1995. Improving Poor People: The Welfare State, the "underclass," and Urban Schools as History. Princeton University Press, Princeton, NJ.

Kaufmann, F.-X., 2013. Variations of the Welfare State: Great Britain, Sweden, France and Germany between Capitalism and Socialism. Springer, Heidelberg.

Ku, Y., 1997. Welfare Capitalism in Taiwan: State, Economy and Social Policy. Macmillan Press, Houndmills, Basingstoke, Hampshire.

Lloyd, T.O., 1986. Empire to Welfare State: English History, 1906–1985. Oxford University Press, Oxford.

Lloyd, T.O., 1993. Empire, Welfare State, Europe: English History, 1906–92. Oxford University Press.

Moscovitch, A., Albert, J., Carleton University, 1987. The "Benevolent" State: The Growth of Welfare in Canada. Garamond Press, Toronto.

Nadasen, P., Mittelstadt, J., Chappell, M., 2009. Welfare in the United States: A History with Documents, 1935–1996. Routledge, New York.

Noble, C., 1997. Welfare as We Knew it: A Political History of the American Welfare State. Oxford University Press, New York.

Raison, T., 1990. Tories and the Welfare State: A History of Conservative Social Policy since the Second World War. St. Martin's Press, New York.

Rothstein, B., 1998. Just Institutions Matter: The Moral and Political Logic of the Universal Welfare State. Press Syndicate of the University of Cambridge, Cambridge, UK.

Steinmetz, G., 1993. Regulating the Social: The Welfare State and Local Politics in Imperial Germany. Princeton University Press, Princeton, NJ.

Stephen, J.A., 2007. Pick One Intelligent Girl: Employability, Domesticity, and the Gendering of Canada's Welfare State, 1939–1947. University of Toronto Press, Toronto.

Stolleis, M., 2013. Origins of the German Welfare State: Social Policy in Germany to 1945. Springer, Berlin.

Trattner, W.I., 1999. From Poor Law to Welfare State: A History of Social Welfare in America. The Free Press, New York.

Wellbeing and Burnout in the Workplace, Psychology of

Raymond T Lee, I.H. Asper School of Business, University of Manitoba, Winnipeg, MB, Canada

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by S. Sonnentag, volume 24, pp. 16445–16447, © 2001, Elsevier Ltd.

Abstract

Work-related well-being encompasses burnout and its opposite, work engagement. Burnout is chronic psychological strain and somatic difficulties, consisting of the syndromes of emotional exhaustion, depersonalization (cynicism), and low-personal accomplishment (self-efficacy). Three theoretical models that explain well-being and burnout are conservation of resources, job demands-resources and person-job fit. The predictors of well-being and burnout include factors in the work environment and dispositional traits. The outcomes include work-related attitudes, performance, and dysfunctional behaviors. Intervention programs can be either person- or organization-directed, and should concentrate on not just reducing burnout but increasing engagement.

Well-being is living in a state that is, in some sense, good (Warr, 2013). Affective well-being concerns one's feelings, whereas syndrome well-being concerns perceptions, recollections, comparisons, and anticipations, as well as affective response. The state of one's well-being ranges from minimal or temporary changes to long-term degradation of mental and physical health. Work-related well-being is a specific aspect of subjective well-being, comprised of both positive and negative features. Most of the research on workers has focused on their long-term mental health state rather than on momentary fluctuations in mood and affect (Warr, 2013).

Conceptualizations of Well-Being and Burnout

Well-being comprises of two dimensions: displeasure-to-pleasure, and low- and high-mental arousal (Warr, 2013). Intersecting these dimensions yields four distinct states: (1) displeasure-low activation (e.g., cynicism), (2) displeasure-high activation (e.g., anxiety), (3) pleasure-low activation (e.g., job satisfaction), and (4) pleasure-high activation (e.g., job involvement).

Burnout is on the displeasure end that triggers low to moderate arousal levels. It is characterized by chronic psychological strain and somatic difficulties. Occupations that require frequent and intense interpersonal contacts with service recipients are especially at risk for burnout. For them, the relevant syndromes of burnout are *emotional exhaustion*, *depersonalization*, and *lack of personal accomplishment* (Maslach et al., 1996). Emotional exhaustion is the experience of mental fatigue from frequent and intense interpersonal contacts at work. Depersonalization is an unhealthy detachment from others. Lack of personal accomplishment is negative self-concept on the job. Recent evidence suggests that workers in nonservice industries also are at risk for burnout. For them, the relevant syndromes of burnout are *emotional exhaustion*, *cynicism*, and *lack of personal efficacy*. Cynicism is the formation of negative attitudes toward career and work. Lack of efficacy is reduced sense of mastery and effectiveness due primarily to external factors and beyond the individual's control.

As the mirror-opposite of burnout, work engagement is on the pleasure end that triggers moderate to high arousal levels (Warr, 2013). It is characterized by the persistent, positive affective-motivational state of fulfillment (Schaufeli and Enzmann, 1998; Maslach et al., 2001). The three syndromes of engagement are *vigor* (i.e., high energy and resilience), *dedication* (i.e., strong job involvement), and *absorption* (i.e., immersion in one's career/profession), which are relevant for workers across both service and nonservice industries.

Theoretical Models of Work-Related Well-Being and Burnout

The phenomena of work-related well-being and burnout can be understood through Hobfoll and Freedy's (1993) conservation of resources (COR) model. It states that workers strive to acquire, retain, protect, and foster valued resources and minimize any threats of resource loss. Threats to resource loss are usually in the form of role demands along with the energy and efforts expended toward meeting such demands. Workers invest their resources in meeting certain job demands with the expectation of receiving positive outcomes (i.e., rewards) in return and attempt to conserve their resources by making judgments about the resources that would be most helpful at any given moment. Well-being is diminished when valued resources cannot be regained.

Along similar lines, the job demands-resources model states that burnout stems from two parallel processes: (1) the effort required to meet demands; and (2) the resources that facilitate goal attainment (Bakker and Demerouti, 2007). Demerouti et al. (2000) discovered that energy depletion is driven by high demands and leads to emotional exhaustion, whereas the erosion of motivation is driven by a lack of resources and leads to depersonalization. Schaufeli and Bakker (2004) showed that adequate resources increase engagement, whereas both demands and inadequate resources increase burnout.

One's relational transactions in the workplace determine whether burnout or engagement will develop. Maslach and Leiter's (1997) person-job fit model states that incongruity between the worker and the job demands undermines well-being. Person-job mismatches can occur in six distinct areas: (1) workload, typically represented by having too many tasks without sufficient time to complete them, but also by

inadequate job-related skills; (2) control, especially in terms of lacking the resources necessary to do the job; (3) rewards, the existence of both extrinsic and intrinsic motivators; (4) fairness, which concerns mutual respect and confirms individuals' self-worth; (5) community, which provides members with a positive sense of connection to others at work and in society; and (6) values, especially in cases where job requirements are in conflict with one's personal values, or when one is caught between conflicting values within the organization or with society. The number of mismatches will have a cumulative adverse impact on well-being.

Predictors of Well-Being and Burnout

The work-related variables most widely linked to well-being and burnout is the demands of workload, role conflict and role ambiguity, and the resources of control, autonomy (Alarcon, 2011), and social support (Halbesleben, 2006). Alarcon (2011) found the demands to have the strongest relation with emotional exhaustion, supporting the COR model prediction that emotional exhaustion is a form of strain that occurs before the other burnout syndromes. Leiter et al. (2009) found that work overload contributed to exhaustion and cynicism; and more broadly, incongruity in values between the physician and the health care system contributed to exhaustion, cynicism, and low professional efficacy. The impact of value incongruity was stronger for women than for men. In contrast to demands, resources had a strongest relation with reduced personal accomplishment (Alarcon, 2011).

Occupations that involve frequent and intense communications with recipients often entail using emotional labor (Larson and Yao, 2005). As a job requirement, emotional labor becomes a demand (Wallace and Lemaire, 2007), but as means to facilitate job performance, it becomes a resource (Lee et al., 2010). Lee et al. found that among physicians, the hiding of emotions was associated with emotional exhaustion and physical symptoms of strain, suggesting that this was a strategy to reduce interpersonal demands. In contrast, the expression of a variety of emotions was associated with reduced depersonalization and physical symptoms of strain, which would encourage positive patient responses (Larson and Yao, 2005).

Beyond the interpersonal demands specific to the job, the broader context of the social work environment impacts well-being and burnout. Counterproductive work relationships can be expressed through incivility and displaced aggression/bullying. Workplace aggression/bullying has been linked to burnout (Lee and Brotheridge, 2006a) and diminished mental and physical health among those targeted and who witnessed such acts (Lovell and Lee, 2011).

Balancing between work and nonwork demands and resources are particularly challenging in many occupations. Among police officers, excessive workload taxes the resources needed to cope with the demands of law enforcement. This imbalance also has spillover effects on their nonwork roles, which often impact relationships outside of work (Lee and Brotheridge, 2006b). Physicians who experienced work–life conflict experienced emotional exhaustion (Lee et al., 2010), whereas physicians, who found time for interests outside of medicine, their family, and to maintain a healthy and active lifestyle, were able to achieve well-being (Wallace and Lemaire, 2007). Nonwork sources of social support are critical resources for achieving work–life balance, and thereby enhancing well-being and reducing burnout (Halbesleben, 2006).

In addition to work-related variables, dispositional traits have been linked to well-being and burnout (Alarcon et al., 2009). Alarcon et al. found the Big Five Model traits, positive and negative affectivity, self-esteem, self-efficacy, locus of control, optimism, proactive personality, and hardiness were all associated with burnout. Work-related variables and dispositional traits directly, as well as interact with each other, to impact well-being and burnout. For example, Hsieh and Wang (2012) found that the effect of job stress on burnout was moderated by locus of control, where the relationship was weaker for internals than for externals in accounting auditors.

Outcomes of Well-Being and Burnout

Work-related attitudes have been linked to well-being across many occupations (Alarcon, 2011). Consistent with the COR model, the burnout dimension of emotional exhaustion was most strongly associated with job dissatisfaction and reduced organizational commitment, and a moderate relation with turnover intentions. Turnover intentions have also been linked to burnout (Leiter and Maslach, 2009). Leiter and Maslach found that work–life incongruence was associated with the burnout syndrome of cynicism, which in turn, was associated with turnover intentions in nurses. Both burnout and engagement impact job performance and outcomes tied to performance (Maslach et al., 2001). Bakker et al. (2004) found that emotional exhaustion was negatively associated with peer-ratings of in-role performance (e.g., 'Achieves the objectives of the job'), whereas engagement was associated with extrarole performance (e.g., Takes initiative to orient new employees to the department even though not part of his/her job description). Taris (2006) found that emotional exhaustion was strongly negatively associated with customer satisfaction. Burnout also can adversely impact patient safety (Halbesleben et al., 2008). Halbesleben et al. found that burnout was negatively associated with reporting of mistakes that did not lead to adverse events.

Burnout has been linked to substance abuse and various dysfunctional behaviors outside of work (Johnson et al., 2005). Johnson et al. found that in police officers, burnout was associated with alcohol abuse, rigidly applying standards of work to the home, and the frequency of violent behaviors directed toward their spouse. Other researches have found strong ties between burnout and the deterioration of family and social relationships, with adverse spillover and contagion effects on the worker's spouse/partner and family members (Lee and Brotheridge, 2006b).

Intervention Programs

Burnout intervention programs are person-directed, organization-directed, or a combination of the two (Korunka et al., 2013). Person-directed interventions include psychotherapy, counseling, exercises for relaxation, and adaptive skills training.

Communication skills training in the service professions not only reduce burnout but can heighten engagement. Such training includes both verbal and nonverbal communications and the ability to express empathetic concern to recipients (Larson and Yao, 2005). For example, method-acting exercises may help trainees in the health care industry vicariously experience their patients' feelings and motives and facilitate appropriate responses. Ramirez et al. (1995) believe that communication skills training will not only reduce the risk of burnout, but contribute to physicians' satisfaction with their practice, which, in turn, positively influences patient outcomes.

Organization-directed interventions include changes in the work procedures, task restructuring, modifying job demands, increasing job control or the level of participation in decision-making, clarification of role expectations, performance-based incentives, developing a respectful work climate, and cultivating interpersonal relations with colleagues and clients. Research suggests that combined person and organization-directed interventions will lead to significant reductions in burnout, especially emotional exhaustion (Awa et al., 2010). Programs that address the motivational and resource needs of those most at risk for burnout will have the best chance for success. However, they require periodic follow-up interventions to prevent the recurrence of the conditions that led to burnout in the first place (Awa et al., 2010).

Maslach et al. (2001) recommend that intervention programs should concentrate on ways to increase engagement rather than merely seek to reduce burnout because engagement is more closely aligned with the idea of embracing and drawing on the potential of the human spirit to improve the quality of work–life. From this perspective, promoting congruity between individuals and their work entails providing resources that do more than fill existing gaps or offer monetary or material support (i.e., extrinsic motivators). Instead, organizations should present resources that foster intrinsic motivation, and thereby create heightened career involvement, enthusiasm, and significance.

See also: Coping with Job Stress; Occupational Stress Management; Subjective Wellbeing, Psychology of.

Bibliography

Alarcon, G.M., 2011. A meta-analysis of burnout with job demands, resources, and attitudes. Journal of Vocational Behavior 79, 549–562.

Alarcon, G.M., Eschleman, K.J., Bowling, N.A., 2009. Relationship between personality variables and burnout: a meta-analysis. Work & Stress 23, 244–263.

Awa, W.L., Plaumann, M., Walter, U., 2010. Burnout prevention: a review of intervention in programs. Patient Education and Counseling 78, 184–190.

Bakker, A.B., Demerouti, E., 2007. The job-demands resources model: state of the art. Journal of Managerial Psychology 22, 309–328.

Bakker, A.B., Demerouti, E., Verbeke, W., 2004. Using the job-demands resources model to predict burnout and performance. Human Resource Management 43, 83–104.

Demerouti, E., Bakker, A.B., Nachreiner, F., Schaufeli, W.B., 2000. A model of burnout and life satisfaction among nurses. Journal of Advanced Nursing 32, 454–464.

Halbesleben, J.R.B., 2006. Sources of social support and burnout: a meta-analytic test of the conservation of resources model. Journal of Applied Psychology 91, 1134–1145.

Halbesleben, J.R.B., Wakefield, B.J., Wakefield, D.S., Cooper, L.B., 2008. Nurse burnout and patient safety outcomes: nurse safety perception versus reporting behavior. Western Journal of Nursing Research 30, 560–577.

Hobfoll, S.E., Freedy, J., 1993. Conservation of resources: a general stress theory applied to burnout. In: Schaefeli, W.B., Maslach, C., Marek, T. (Eds.), Professional Burnout: Recent Developments in Theory and Research. Taylor & Francis, Washington, DC, pp. 115–133.

Hsieh, Y.-H., Wang, M.-L., 2012. The moderating role of personality in HRM – from the influence of job stress on job burnout perspective. International Management Review 8, 5–18.

Johnson, L.B., Todd, M., Subramanian, G., 2005. Violence in police families: work-family spillover. Journal of Family Violence 20, 3–12.

Korunka, C., Tement, S., Zdrehus, C., Borza, A., 2013. Burnout: Definition, Recognition and Prevention Approaches. http://burnoutintervention.eu/fileadmin/user_upload/BOIT_theoretical_abstract_2705.pdf (accessed 13.03.07.).

Larson, E.H., Yao, X., 2005. Clinical empathy as emotional labor in the patient–physician relationship. Journal of the American Medical Association 293, 1100–1106.

Lee, R.T., Brotheridge, C.M., 2006a. When prey turns predatory: workplace bullying as a predictor of counter-aggression/bullying, coping and well-being. European Journal of Work and Organizational Psychology 15, 352–377.

Lee, R.T., Brotheridge, C.M., 2006b. Crossing the thin blue line: a review of the prevalence, causes, and outcomes of police burnout. Canadian Journal of Police and Security Services 4, 250–262.

Lee, R.T., Lovell, B.L., Brotheridge, C.M., 2010. Tenderness and steadiness: relating job and interpersonal demands and resources with burnout and physical symptoms of stress in Canadian physicians. Journal of Applied Social Psychology 40, 2319–2342.

Leiter, M.P., Frank, E., Matheson, T.J., 2009. Demands, values and burnout: relevance for physicians. Canadian Family Physician 55, 1224–1225, e1–e6.

Leiter, M.P., Maslach, C., 2009. Nurse turnover: the mediating role of burnout. Journal of Nursing Management 17, 331–339.

Lovell, B.L., Lee, R.T., 2011. Impact of workplace bullying on emotional and physical well-being: a longitudinal collective case study. Journal of Aggression, Maltreatment & Trauma 20, 344–357.

Maslach, C., Jackson, S.E., Leiter, M.P., 1996. Maslach Burnout Inventory manual, third ed. Consulting Psychologists Press, Palo Alto, CA.

Maslach, C., Leiter, M.P., 1997. The Truth about Burnout. Jossey-Bass, San Francisco.

Maslach, C., Schaufeli, W.B., Leiter, M.P., 2001. Job burnout. Annual Review of Psychology 52, 397–422.

Ramirez, A.J., Graham, J., Richards, M.A., Cull, A., Gregory, W.M., Leaning, M.S., et al., 1995. Burnout and psychiatric disorder among cancer clinicians. British Journal of Cancer 71, 1263–1269.

Schaufeli, W.B., Enzmann, D., 1998. The Burnout Companion for Research and Practice: A Critical Analysis of Theory, Assessment, Research, and Interventions. Taylor & Francis, London.

Schaufeli, W.B., Bakker, A.B., 2004. Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. Journal of Organizational Behavior 25, 293–315.

Taris, T.W., 2006. Is there a relationship between burnout and objective performance? A critical review of 16 studies. Work & Stress 20, 316–334.

Wallace, J.E., Lemaire, J., 2007. On physician well-being: you'll get by with a little help from your friends. Social Science and Medicine 64, 2565–2577.

Warr, P., 2013. How to think about and measure psychological well-being. In: Sinclair, R.R., Wang, M., Tetrick, L.E. (Eds.), Research Methods in Occupational Health Psychology Measurement, Design and Data Analysis. Routledge, New York, pp. 76–90.

Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences

Sabine Sonnentag, University of Mannheim, Mannheim, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Well-being is a multidimensional concept, with burnout being a specific job-related construct of poor well-being. This article provides definitions of well-being and burnout, and introduces the new concept of work engagement. It briefly presents theoretical models on job-related well-being and burnout. It summarizes empirical studies on workplace and personal factors contributing to job-related well-being and burnout as well as studies on job-related consequences of poor well-being and burnout. Interventions for improving job-related well-being are also described.

Conceptualizations of Well-Being and Burnout

Job-related well-being is a specific aspect of subjective well-being and comprises positive features such as work engagement, job satisfaction, job attachment, job involvement, and job morale; impaired job-related well-being includes experiences such as job tension, depression, burnout, and alienation from work. The well-being concept refers to both minimal and temporary changes in employee mood and more long-term disturbances in mental health (e.g., chronic anxiety or psychosomatic complaints). It is typical for work-related research to examine mental health and well-being in normally working samples who are not psychiatric patients.

Job-related well-being is a multidimensional concept. Based on the conceptualization of arousal and pleasure as two independent dimensions, Warr (1987) proposed a model of job-related well-being comprising three dimensions: the first dimension runs from displeasure to pleasure, the second from anxiety to comfort, and the third from depression to enthusiasm. Additionally, feelings of competence and aspiration can also be regarded as facets of job-related well-being.

Burnout is a specific and severe form of disturbed job-related well-being, originally mainly observed in human service professionals. According to Maslach and Jackson (1981), the burnout syndrome comprises emotional exhaustion, depersonalization (i.e., negative and cynical attitudes toward one's clients or patients), and reduced personal accomplishment as the key dimensions. More recently, Demerouti et al. (2001) have conceptualized exhaustion and disengagement as the two main components of burnout. During the past decades, researchers have observed that burnout not only develops in human service professions but can also occur in all kinds of jobs.

While most research examined negative aspects of well-being, Schaufeli and Bakker (2004) proposed to focus on work engagement as a positive aspect of job-related well-being. Work engagement comprises three dimensions: vigor, dedication, and absorption. It is a positive state of mind that becomes evident in the process of working. Whereas negative aspects of well-being such as burnout seem to be more influenced by job stressors, work engagement tends to benefit from job resources (e.g., job control).

Theoretical Models on Job-Related Well-Being and Burnout

During the past 50 years, researchers have put much effort into the development of theoretical models on the relationship between work situation factors and job-related well-being and burnout. Among the most prominent models are the job characteristics model (Hackman and Oldham, 1976); person–environment fit theory (Harrison, 1978); the job demand–job control model (Karasek and Theorell, 1990), along with its refinements (e.g., Demerouti et al., 2001); the vitamin model (Warr, 1994); and the effort–reward imbalance model (Siegrist, 1996).

In their job characteristics model, Hackman and Oldham (1976) specified five core job characteristics that are assumed to be crucial for high internal work motivation and employee satisfaction: skill variety, task identity, task significance, autonomy, and feedback from the job. These job characteristics are expected to impact critical psychological states (i.e., experienced meaningfulness of the work, experienced responsibilities for work outcomes, and knowledge of the results of work activities), which in turn impact the motivation and satisfaction as outcome variables. Furthermore, Hackman and Oldham proposed that moderators such as growth need strength, contextual satisfaction, knowledge, skills, and abilities have an effect on the relationship between core job characteristics and critical psychological states and also on the relationship between critical psychological states and outcome variables. Humphrey et al. (2007) provided a comprehensive overview over the empirical research on the job characteristics model.

Person–environment (P-E) fit theory (Harrison, 1978) proposes that strain occurs when characteristics of the job environment do not fit the individual. Two types of fit between an individual and the environment can be differentiated: the fit between the demands of the environment and the abilities and competencies of the person (demands–abilities fit), and the fit

between the needs of the person and supplies from the environment (needs–supplies fit). According to P-E fit theory, well-being suffers when environmental demands exceed the person's abilities and when the person's needs exceed the supplies provided by the environment. Meta-analytical evidence shows that demands–abilities fit and needs–supply fit are indeed positively related to well-being indicators (Kristof-Brown et al., 2005).

The job demand–job control model (Karasek and Theorell, 1990) received much attention within work and organizational psychology. In its original version, the job demand–job control model describes two dimensions relevant for job-related well-being: job demands and job control. In this model, job demands refer to high workload and time pressure, and job control includes decision authority and intellectual discretion. The model proposed that the combination of high job demands and low job control becomes evident in so-called high-strain jobs that have severe negative effects on employee well-being. The original model further assumed that a combination of high job demands and high job control results in active jobs that have more effects on employee well-being. Later, other work situation factors, such as social support and informational control, have been incorporated into the job demands–job control model (Karasek and Theorell, 1990).

Demerouti et al. (2001) extended the job demand–job control model into a more general job demands–job resources model. Importantly, Demerouti et al. suggested that in addition to job control, other resources (e.g., feedback, reward, participation, supervisor support) can also buffer the negative effects of job demands on poor well-being or burnout. In addition, this model stated that job resources in itself have motivational power and increase work engagement (Schaufeli and Bakker, 2004).

In his vitamin model, Warr (1994) proposed that job-related well-being is affected by job characteristics in a way similar to the effects of vitamins on physical health. Basically, the vitamin model suggests that some job characteristics have a linear effect, whereas others have a nonlinear effect on well-being. Among the job characteristics assumed to be linearly related to well-being are physical security, a valued social position, and availability of money. Job characteristics assumed to show a curvilinear relationship with well-being include opportunity for personal control, opportunity for skill use, externally generated goals (i.e., demands), variety, environmental clarity, and opportunity for interpersonal contact. More specifically, the vitamin model assumes that well-being increases as these latter job characteristics increase; however, at a very high level, these job characteristics will cause a decrement in well-being.

The effort–reward imbalance model (Siegrist, 1996) builds on the assumption that the effort a person invests in his or her job and the rewards he or she gets are crucial for the person's well-being and health. Effort may result from extrinsic (e.g., obligations and demands of the job) and intrinsic (e.g., resulting from a high need for control or approval) demands. With respect to rewards, money, esteem, and status control (e.g., job stability, career advancement) can be differentiated. In essence, the model proposes that a lack of reciprocity between effort and rewards is experienced as stressful and leads to impairments in health and well-being. Empirical studies tend to support this view (Van Veghel et al., 2005).

Predictors of Job-Related Well-Being and Burnout

Empirical research has identified job stressors and job resources as the core predictors of job-related well-being and burnout. In addition, personality variables can also explain interindividual differences in job-related well-being and burnout. With respect to job stressors (i.e., events and conditions in the work situation that evoke strain reactions), there is evidence that employees who face a high workload and time pressure report a poorer job-related well-being (De Lange et al., 2003) and higher levels of burnout (Crawford et al., 2010). Moreover, so-called hindrance stressors (e.g., hassles, interruptions, situational constraints, organizational politics) are related to poor job-related well-being and burnout (Crawford et al., 2010), with hindrance stressors often showing larger effect sizes than workload or time pressure. Probably, job conditions that make it difficult and effortful for employees to accomplish their tasks are more detrimental than the sheer amount of work. Research further showed that work-related critical incidents (i.e., unexpected and severe events such as the death of a patient a health professional is caring for) predict symptoms of anxiety and depression (de Boer et al., 2011). Moreover, workplace bullying is related to mental health problems such as anxiety and depression as well as burnout (Nielsen and Einarsen, 2012).

Overall, longitudinal studies support the view that the effect of job stressors on poor well-being and burnout is stronger than the effect of poor well-being or burnout on the exposure and perception of job stressors (Sonnentag and Frese, 2012). This pattern of findings speaks for a stronger causal effect from stressors on strains than from strains on stressors. However, with respect to workplace bullying, it turned out that this stressor predicted mental health problems to a degree similar in size as to which mental health problems predicted workplace bullying (Nielsen and Einarsen, 2012). Thus, the experience of workplace bullying might contribute to a vicious cycle, with vulnerable employees suffering most from this type of job stressor.

When it comes to the predictors of work engagement, it is important to differentiate between two types of job stressors. Challenge stressors such as a high workload are positively related to work engagement, whereas hindrance stressors (e.g., situational constraints, hassles, social conflicts) are negatively related to work engagement (Crawford et al., 2010). Combining these findings with the findings on the predictors of poor well-being and burnout, challenge stressors reveal an ambivalent nature: they are related to high work engagement *and* to strain symptoms and burnout. Possibly, they are energizing and – at least under some circumstances – depleting at the same time.

Job resources are conditions at work that help perform well or deal with stressors in a positive way. Job control (i.e., autonomy) and social support belong to the most frequently studied job resources. Studies have shown that job control is positively related to well-being (De Lange et al., 2003; Humphrey et al., 2007) and low burnout levels (Crawford et al., 2010). When employees have some discretion about how to do their job, they experience a better well-being and are less exhausted. Also, social support was found to be a positive factor in the workplace, showing negative

correlations with burnout (Halbesleben, 2006). Employees who receive a high level of social support from their supervisors and coworkers enjoy a better well-being than persons who lack social support. Similarly, job resources are also positively related to work engagement (Christian et al., 2011; Crawford et al., 2010). In addition, leadership variables predict work engagement (Christian et al., 2011).

It has been argued that job resources, particularly job control, buffer the negative effects of job stressors on well-being (e.g., Karasek and Theorell, 1990). Empirical evidence for this buffer hypothesis, however, is mixed (De Lange et al., 2003). It has been suggested that the resources must match the stressors that they should buffer, otherwise they are only of little help (Daniels and de Jonge, 2010).

Job-related well-being and burnout depend not only on employees' job stressors and job resources but also partly on their personality. Alarcon et al. (2009) identified negative affectivity (i.e., low levels of emotional stability) as a strong predictor of burnout. It might be that employees with high levels of negative affectivity see their work environment in a more negative light and are less able to cope with job stressors. In addition, other individual difference variables such as low general self-esteem, low general self-efficacy, and an external locus of control are related to elevated burnout levels (Alarcon et al., 2009). Also, work engagement seems to depend on personality factors. Particularly, conscientiousness, positive effect, and proactive personality are substantially related to work engagement (Christian et al., 2011).

Person-related factors have also been examined as moderators of the relationship between job stressors and poor well-being or burnout. For instance, emotional stability was found to buffer the negative effects of job stressors (Kammeyer-Mueller et al., 2009). Recent research has demonstrated that also processes happening off the job play a role for job-related well-being and burnout. For instance, when employees mentally detach from work during their leisure time (i.e., temporally disengage from the demands of their jobs), they enjoy a better well-being, particularly when job stressors are high (Sonnentag, 2012).

Taken together, job-related well-being and burnout have multiple causes that lie partially in employees' work situation, partially in their personality and their life off the job. It seems that particularly the combination and interplay of various factors are responsible for a person's job-related well-being.

Consequences of Job-Related Well-Being and Burnout

Low job-related well-being and burnout are not only unpleasant experiences but they may also threaten the individual's functioning within the work organization. Research shows that poor job-related well-being is negatively associated with task and contextual performance (Ford et al., 2011). This negative relationship between poor well-being and performance was found for various performance measures (self-reports, supervisor and peer ratings, organizational records), suggesting that the empirical correlations cannot be reduced to a pure perceptional bias. Interestingly, the effect sizes tend to be higher for these more psychological well-being indicators than for physical symptoms (Ford et al., 2011). When looking specifically at burnout, research demonstrates significant correlations between burnout and objective performance indicators as well as customer satisfaction (Taris, 2006). Moreover, burnout also matters for safety issues: employees with elevated burnout levels experience more adverse events, accidents, and injuries (Nahrgang et al., 2011). In contrast to burnout and other impairments of job-related well-being, work engagement was found to be positively related to task and contextual job performance (Christian et al., 2011).

When interpreting these findings, one has to take into account that most studies used cross-sectional designs. Thus, the results do not necessarily imply that poor well-being and burnout have a causal effect on performance. It might be that performance has an effect on job-related well-being or that both well-being and performance are influenced by other factors such a leadership behavior or organizational climate.

Intervention Programs

Different kinds of interventions aim at the protection and improvement of job-related well-being. Most interventions are oriented toward the individual, i.e., to the way an individual perceives or deals with unfavorable work situations. Other individual-oriented interventions are designed to directly influence the individual's psychological and physical well-being, for example, by progressive muscle relaxation, cognitivebehavioral interventions, or meditation techniques. Empirical evaluations of these programs provide evidence for the effectiveness of interventions based on the behavioral–cognitive approach as well as on relaxation or meditation (Richardson and Rothstein, 2008), with behavioral–cognitive interventions showing the largest effect sizes. Interestingly, multi-method approaches (i.e., interventions that combined methods and exercises from a variety of conceptual frameworks) resulted in smaller – albeit still significant – effect sizes (Richardson and Rothstein, 2008), perhaps because they are less successful in giving employees a clear focus and guidance about what to do in a stressful situation. In addition to these specific stress management approaches, interventions that aim at lifestyle changes and that increase efficacy beliefs are also successful in improving well-being (e.g., Chen et al., 2009). For instance, workplace interventions that stimulate increased physical activity are associated with better well-being (Conn et al., 2009).

Another approach to improve job-related well-being focuses on workplace interventions. During such workplace interventions, jobs are redesigned, for instance, by increasing employee job control, by reducing job stressors, or by increasing job resources. Empirical studies show beneficial effects of such workplace interventions on job-related well-being (e.g., Holman et al., 2010). Overall, the effect sizes of all intervention programs vary largely in size, suggesting that they have to be tailored to the specific requirements of the situation and to employees' needs.

See also: Ethical Codes, Professional: Business Codes; Human Resource Management, Psychology of; Leadership in Organizations, Sociology of; Organization: Overview; Organizational Climate; People in Organizations; Stress at Work; Workplace Autonomy; Workplace Safety and Health.

Bibliography

Alarcon, G., Eschleman, K.J., Bowling, N.A., 2009. Relationships between personality variables and burnout: a meta-analysis. Work & Stress 23, 244–263.

Chen, S., Westman, M., Eden, D., 2009. Impact of enhanced resources on anticipatory stress and adjustment to new information technology: a field-experimental test of conservation of resources theory. Journal of Occupational Health Psychology 14, 219–230.

Christian, M.S., Garza, A.S., Slaughter, J.E., 2011. Work engagement: a quantitative review and test of its relations with task and contextual performance. Personnel Psychology 64, 89–136.

Conn, V.S., Hafdahl, A.R., Cooper, P.S., Brown, L.M., Lusk, S.L., 2009. Meta-analysis of workplace physical activity interventions. American Journal of Preventive Medicine 37, 330–339.

Crawford, E.R., LePine, J.A., Rich, B.L., 2010. Linking job demands and resources to employee engagement and burnout: a theoretical extension and meta-analytic test. Journal of Applied Psychology 95, 834–848.

Daniels, K., de Jonge, J., 2010. Match making and match breaking: the nature of match within and around job design. Journal of Occupational and Organizational Psychology 83, 1–16.

de Boer, J., Lok, A., van't Verlaat, E., Duivenvoorden, H.J., Bakker, A.B., Smit, B.J., 2011. Work-related critical incidents in hospital-based health care providers and the risk of post-traumatic stress symptoms, anxiety, and depression: a meta-analysis. Social Science and Medicine 73, 316–326.

De Lange, A.H., Taris, T.W., Kompier, M.A.J., Houtman, I.L.D., Bongers, P.M., 2003. "The very best of the millenium": longitudinal research and the demand-control-(support) model. Journal of Occupational Health Psychology 8, 282–305.

Demerouti, E., Bakker, A.B., Nachreiner, F., Schaufeli, W.B., 2001. Job demands-resources model of burnout. Journal of Applied Psychology 86, 499–512.

Ford, M.T., Cerasoli, C.P., Higgins, J.A., Decesare, A.L., 2011. Relationships between psychological, physical, and behavioural health and work performance: a review and meta-analysis. Work & Stress 25, 185–204.

Hackman, J.R., Oldham, G.R., 1976. Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16, 250–279.

Halbesleben, J.R.B., 2006. Sources of social support and burnout: a meta-analytic test of the conservation of resources model. Journal of Applied Psychology 91, 1134–1145.

Harrison, R.V., 1978. Person-environment fit and job stress. In: Cooper, C.L., Payne, R. (Eds.), Stress at Work. Wiley, New York, pp. 175–205.

Holman, D.J., Axtell, C.M., Sprigg, C.A., Totterdell, P., Wall, T.D., 2010. The mediating role of job characteristics in job design interventions: a serendipitous quasi-experiment. Journal of Organizational Behavior 31, 84–105.

Humphrey, S.E., Nahrgang, J.D., Morgeson, F.P., 2007. Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension of the work design literature. Journal of Applied Psychology 92, 1332–1356.

Kammeyer-Mueller, J.D., Judge, T.A., Scott, B.A., 2009. The role of core self-evaluations in the coping process. Journal of Applied Psychology 94, 117–195.

Karasek, R., Theorell, T., 1990. Healthy Work: Stress, Productivity and the Reconstruction of Working Life. Basic Books, New York.

Kristof-Brown, A.L., Zimmerman, R.D., Johnson, E.C., 2005. Consequences of individual's fit at work: a meta-analysis of person-job, person-organization, oerson-group, and person-supervisor fit. Personnel Psychology 58, 281–342.

Maslach, C., Jackson, S.E., 1981. The measurement of experienced burnout. Journal of Organizational Behavior 2, 99–113.

Nahrgang, J.D., Morgeson, F.P., Hofmann, D.A., 2011. Safety at work: a meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. Journal of Applied Psychology 96, 71–96.

Nielsen, M.B., Einarsen, S., 2012. Outcomes of exposure to workplace bullying: a meta-analytic review. Work & Stress 26, 309–332.

Richardson, K.M., Rothstein, H.R., 2008. Effects of occupational stress management intervention programs: a meta-analysis. Journal of Occupational Health Psychology 13, 69–93.

Schaufeli, W.B., Bakker, A.B., 2004. Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study. Journal of Organizational Behavior 25, 293–315.

Siegrist, J., 1996. Adverse health effects of high effort/low reward conditions. Journal of Occupational Health Psychology 1, 27–41.

Sonnentag, S., 2012. Psychological detachment from work during leisure time: the benefits of mentally disengaging from work. Current Directions in Psychological Science 21, 114–118.

Sonnentag, S., Frese, M., 2012. Stress in organizations. In: Schmitt, N.W., Highhouse, S. (Eds.), Handbook of Psychology, Industrial and Organizational Psychology, second ed., vol. 12. Wiley, Hoboken, pp. 560–592.

Taris, T.W., 2006. Is there a relationship between burnout and objective performance? A critical review of 16 studies. Work & Stress 20, 316–334.

Van Veghel, N., De Jonge, J., Bosma, H., Schaufeli, W.B., 2005. Reviewing the effort-reward imbalance model: drawing up the balance of 45 empirical studies. Social Science & Medicine 60, 1117–1131.

Warr, P.B., 1987. Work, Unemployment, and Mental Health. Oxford University Press, Oxford, England.

Warr, P.B., 1994. A conceptual framework for the study of work and mental health. Work & Stress 8, 84–97.

Western and Non-Western Views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences

Thomas Brisson, University Paris 8, Saint-Denis, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The expression de-Westernization of the Human and Behavioral Sciences refers to two different trends. First, it designates their diffusion outside their Western context of origins, through the establishment of national universities and disciplines, the migrations of scholars, the internationalization of scientific organizations, etc. In the second sense, it stands for the attempts to create Human and Behavioral Sciences not predicated on Western epistemological frameworks. Theses two trends have developed since the nineteenth century and have led to a wide range of situations, with the assimilation of the Western scientific apparatus being linked to their more or less partial rejection.

The expression de-Westernization of the Human and Behavioral Sciences refers to two different trends. In the first sense, it designates their diffusion outside their Western context of origins. This diffusion has consisted in various processes, such as the establishment of disciplines nationally, as well as the material and institutional background that made this creation possible (the establishment of national universities, the migrations of scholars, the internationalization of scientific organizations, etc.). In the second sense, the concept of de-Westernization stands for the attempts to create Human and Behavioral Sciences not predicated on Western epistemological frameworks (see, for instance, Alatas (2006) calls to frame an Asian sociology independently of the 'European Founding Fathers,' Durkheim, Marx, and Weber).

De-Westernization as the global diffusion of Human Science and de-Westernization as a theoretical program have both developed from the nineteenth century on, but have interacted in various ways. They sometimes occurred simultaneously, with the internationalization of human sciences leading to a criticism of their Western apparatus. But the reverse was equally true when the indigenization of European theories did not give rise to any epistemic challenge. Most of the time, a combination of these two trends prevailed. The first encounter with Western theories, under colonial pressure, prompted both their creative appropriation and the denunciation of their imperialistic origins. The same ambivalence continued shaping the approaches of the newly decolonized societies, where the Western scientific apparatus was implemented to build modern societies and denounced in the name of national independence. In this respect, the current situation is in line with process: in a context where Western supremacy is contested, attempts to 'provincialize' European theories are nevertheless struggling with their unprecedented diffusion and political importance (Chakrabarty, 2000).

This conflicting process must therefore be analyzed with reference to its multiple situations. De-Westernization had as many illustrations as there were different national paths, historical situations, and scientific disciplines. Such a diversity of situations notwithstanding, it is possible to gain a better understanding of scientific de-Westernization by focusing on its most salient aspects. De-Westernization took place, first, in a context shaped by the colonial encounter and the disruption of the existing intellectual traditions. The implementation of the Western scientific apparatus outside the boundaries of Europe was then prompted by the efforts to build institutions and to train scientific professionals: the migrations of scholars to the West, or the creation of universities and research centers, are among the most significant achievements of this policy. On a third level, de-Westernization was closely linked to the international and transnational scientific organizations that were set up during the last two centuries.

These three points aim at better characterizing the sociological and historical processes that led to the disputed indigenization of Western sciences. By contrast, two last points scrutinize what functions human sciences fulfilled in non-Western societies. They show how human sciences were instrumental in defining political concepts that acted as rallying cries against perceived forms of scientific and cultural hegemony. In line with this political approach, they finally focus on the idea of 'de-Westernization as a scientific program' and analyze what were their argumentative strategies.

An analysis of the de-Westernization of the human and behavioral sciences would imply tackling all the disciplines in every non-Western country. Needless to say that it was not possible to cover such a variety of situations. The cases of anthropology, psychology, and sociology will receive more attention. Accordingly, China, India, Japan, and specific parts of the Muslim world will be more precisely dealt with. Finally, central historical periods and scientific turning points were given priority, over the homogeneous course of 'normal science.'

(Pre)Colonial Encounters

Western human sciences are now everywhere: officials of African states and NGOs use the same statistics than their North American colleagues. Mental suffering is categorized, in India or in Algeria, according to a conception of the self elaborated by modern European psychology. As for Chinese or Japanese anthropologists, who study the ethnic *Miao* or *Ainu* the way English or German anthropologists mapped the world tribes. If human sciences originally stem from the revolutions that have reshaped Europe at the modern times, they have now spread well beyond the European boundaries.

These changes were linked to the diffusion of European modernity on a global scale and the universal acceptance of Western political ideas, cultural forms, and bodily practices. Yet the acculturation of human sciences followed specific channels. To gain a better understanding of this process, it is useful to differentiate between the diffusion of human sciences in terms of 'content' and 'form.' With regard to their *content*, the acclimation of human sciences implied that Western theories spread outside Europe: for each discipline, canonical authors and their main concepts were translated. Some became extremely influential (see Charles Darwin (1809–82) in the Arab world (Elshakry, 2014) or Herbert Spencer (1820–1903) in China). The European authors who had written on non-Western questions equally received a great deal of attention, as local scholars were interested in the way Western scientists had understood their societies. Joseph Needham's *Science and Civilization in China* practically redefined the understanding the Chinese had of their sciences. Others had a less-enthusiastic reception, such as Renan, whose views on Islam prompted a critical response from Al-Afghani, or Weber's analysis of Buddhism or Confucianism in Asia.

These authors were certainly challenging for non-Western scholars, but to a certain extent only. For people who had rich intellectual traditions, it was possible to dismiss Marx or Comte, or to hold them in low esteem compared to Ibn Khaldun or Confucius. Many local scholars also stressed the numerous parallels between Western and canonical non-Western authors. The content of European human sciences was then easy to appropriate. Such was not the case, yet, with the *form* of human sciences. As 'circulating forms,' they had a 'performativity' (Appadurai, 2013) that completely redefined native intellectual traditions. Human sciences were premised on patterns, sets of relations, and scales of analysis that soon made it impossible to continue using local schemes. Besides, the cogency of European human sciences and their internal coherence made them powerful intellectual tools. For all these reasons, they came to quickly supersede other scientific systems. China is a case in point, where a two-millennia-old system of thinking based on the idea of correlation was replaced, in less than a few decades, by European sciences and their concept of causality (Vandermeersch, 2013).

Moreover, the global acculturation of human sciences had dramatic consequences as it was linked to imperial expansion and came to play a key role in the strengthening of the European rule (Said, 1978). Yet this early process of diffusion was equally fraught with indeterminacy and ambivalence. In the course of the colonial penetration, many local scholars came to collaborate with their European peers. They became familiar with a set of knowledge that cannot be reduced to its imperialistic purposes and they gained a new understanding of their societies (Pouillon and Vatin, 2011). A product of European modernity, human sciences became tools that local scholars could use to make sense of the transformations this very modernity had imposed on them.

Yet not all countries had the same resources to adapt to the changes. Many African or Asian societies, who had resisted state-based organizations, strove to keep the European influence remote. But they were unable, until the recent decades, to set up policies to catch up with the new scientific challenges. Others, on the contrary, had firmly established states and centuries- sometimes millennia-old scholarly traditions. They could therefore more easily react to the new situation.

It was the case in the Ottoman Empire, where the central power in Istanbul had been monitoring the European scientific progress since the eighteenth century. Its administrative structure and the ancient tradition of Muslim scholarship allowed setting up a coordinated policy of intellectual exchanges. It was also the case in Egypt, where a nascent nationalism led Muhammad Ali to open up to the West in the middle of the nineteenth century, in order to strengthen his autonomy toward the Caliphate. Among all the attempts made at this time, yet, Japan stand as the most successful. Despite its 'closed-country policy' that had severed ties with the outside, Japan had been watching the development of Western knowledge and technics, mainly through backdoor contacts with Dutch merchants. Besides, the country had its own tradition of scholarly investigation, which allowed for a swift importation of human sciences: the numerous essays on the particularities of the Japanese language paved the way for modern linguistic, and the administrative accounts of the population for the acclimation of demographic and sociological sciences. In a few decades, Meiji's Japan managed to successfully graft Western human sciences to her German-inspired academic structure. Here again, this was possible only because the former *Shogun* had established a centralized state, which proved decisive when Emperor Meiji (1852–1912) launched the Westernization of the country.

Despite the diversity of reactions, these countries faced more or less the same situation. They had to cope with the disruption of the native cultures brought by the European epistemological systems. There remained classical *Adib* or Confucian scholars in the Arab world or in China, after the West had swept in their countries, but what they could write would never have the same weight. Significantly, they came to be referred through exonymes. The *Ru Jia* became *Confucians*, a label that had been absent from the Chinese culture until the European unified a disparate body of texts under the name of Confucius. *Falsafa*, accordingly, was now referred as *Muslim* or *Arabic Philosophy*, whereas the discipline had always stressed its Mediterranean legacy. In this respect, the introduction of Western knowledge did not put an end to other culture's episteme. Rather it disrupted their unquestioned centrality and made them 'local' or 'alternative.'

Consequently, all countries saw a growing split between 'modern' and 'traditional' scholars. The formers were taught in the West or Westernized schools, and were expected to contribute to the modernization of their countries. The latter, provided they had retained some sort of authority, tried to keep the existing intellectual traditions alive. Yet the divide was not always a clear-cut one until the last part of the nineteenth century. In the Muslim world, the scholars who introduced Western theories had been raised in traditional schools and saw no contradiction between their education and the Western scientific apparatus. As they were supposed to be universal and appropriable by non-Western people, the Ottomans thought it appropriate to reformulate Western theories in the Koranic idiom (Aydin, 2007). Reformist thinkers like Jamal-Eddine Al-Afghani (1838–97) or Muhammad Abduh (1849–1905), accordingly, had no dilemma when they had to combine their training in Muslim exegesis with the new knowledge they

encountered in Europe. It is only with their follower Rachid Rida (1865–1935), at the end of the nineteenth century, that the difference between European and Muslim knowledge came to be seen as unbridgeable. At first a fierce supporter of intellectual modernization, Rida made a radical turn toward traditional scholarship and faith. Westernized intellectuals, at the same time, began to feel alienated from their native culture and promoted projects of Westernization that were more and more incompatible with the ancient intellectual systems.

As the aforementioned examples show, the capacity to oppose Westernization was to a large extent linked to the existence of a strong political apparatus and ancient traditions of knowledge. These very resources could nevertheless have opposite effects, when existing concepts led to misunderstandings and mistranslations. In Japan, when they introduced sociology, scholars had to struggle to frame a concept that adequately conveyed the new meaning of 'society.' They used ancient Chinese ideograms but were unsatisfied as they retained the idea of a loosely affiliated group of persons. It took them years to finally elaborate the concept of *Shakai* (see the article 'Translating Sociology into Japanese, Chinese and Korean Language'). Many countries were not as successful as Japan, though, and faced deep crisis. In China, the strong imperial state and the ancient scholarship paradoxically impeded the reception of European sciences. With its millennia-old scientific system, China could legitimately boast to be one of the major cultures in the world. But this self-confidence had already limited the contacts with the Jesuits at the sixteenth century. Three centuries later, many Chinese scholars and officials were equally blinded to the novelty of European knowledge. They assumed that it was in no way superior to their own and did not bother pondering over its threats.

The problem was not solely intellectual, but political. Several Chinese politicians and scholars from regional background had gained familiarity with European sciences, through contacts with missionaries or during stays in Japan. As they were fully aware of the new situation, they began acknowledging a full-fledged scientific modernization of China. Some local reforms were implemented at the end of the nineteenth century but were stopped by conservative cliques and literati classes in Beijing. The status quo prevailed and China faced a deepening crisis in a context of imperialism. The situation became so serious that imperial officials had to come with a compromise, known as the *Ti-Yong* principle. They differentiated between Chinese knowledge, supposed to shape the moral core and substance (*ti*) of the Chinese society, and Western knowledge, a set of purely functional means (*yong*) designed to modernize China. Yet the Chinese were to discover that Western knowledge and technics were not value-neutral and that it was impossible to acclimate them without affecting the society at large. The failure of the *Ti-Yong* principle worsened the crisis. In 1911 only, as the Empire collapsed, a new generation of politicians and scholars seized power and began modernizing China. They held Confucianism responsible for the precarious situation of China. Confucianism practically lost all its influence and was replaced by Western sciences (Levenson, 1968).

The *Ti-Yong* principle is similar to what almost all non-Western countries tried to implement. In the Ottoman Empire, scholars sought a balance between their Muslim legacy and the new European scientific systems. Some saw no contradiction; others, especially in times of imperialistic encroachments, came to limit the Western influence. The situation was easier for Japanese scholars as Japan had successfully contained the European expansion. Here too, nevertheless, Emperor Meiji promoted a revival of the Confucian values, to provide the society with moral landmarks. Western knowledge was everywhere; yet traditional systems survived, devoid of any epistemological value but with a strong ethical stance. The Japanese had managed to implement what the Chinese had envisioned. Japan had begun reconsidering her traditions already in the eighteenth century and her scholars had more leverage to 'reinvent' a flexible Japanese culture. Non-Western scholars came to Japan, sometimes from as far as Egypt, to understand how the country had achieved a balance between a sweeping scientific modernization and a strong sense of oneself. They later advocated various solutions that illustrate the diversity of the ways human sciences globalized beyond the European frontiers and reshaped non-Western societies.

The Institutional and Material Context of Scientific De-Westernization

Acclimating human sciences in countries where they were hitherto unknown implied implementing structures where these sciences could be developed and taught. Local governments built universities, research centers and created disciplines on the Western model. They also sent scholars and students abroad. Here again, the more autonomous the local rulers, the more leverage they had to set up this structure. In countries where the colonial rule was stronger, the same process occurred through different channels, with the Europeans opening universities that came to play a new role after the independence.

The setting up of (national) universities was the first step to introduce human sciences. Universities were created all over the world, following the European model, at the turn of the nineteenth and twentieth centuries. The University of Tokyo was established in 1877, Peking University in 1898, and Cairo University in 1908. Local rulers created universities in order to introduce Western knowledge and to strengthen their nations. They hired foreign staffs to teach new disciplines. Universities quickly superseded the old educational system. As modern official institutions, they had a prestige ancient schools did not have anymore. Besides, they trained large number of students, who, when they became officials or prominent intellectuals, contributed to legitimate and reproduce the new structure. These transformations affected classical institutions like the Islamic *madrasa* or the Chinese academia. Koranic schools did not cease to exist but were to focus almost exclusively on religious learning. In China, a sweeping change took place as the imperial examination system, which had been formally implemented at the beginning of the seventh century, was suppressed in 1905, a few years after the first modern university was created.

The countries where the colonial rule was more coercive experienced similar changes but at a different pace. The University of Algiers was established in 1909 by the French

colonizers, and the Universities of Calcutta and Bombay/Mumbai in 1857 by the English. The former nevertheless trained very few 'Muslim students' until the middle of the twentieth century: at the end of the 1950s more Algerians were studying at the Sorbonne in Paris than in Algiers. The University of Algiers was an institution committed to teach the Europeans and to exclude the Muslims. It is only after the independence in 1962 that the modern academic system became largely opened to the Algerian population, as it was supposed to contribute to the building of the nation.

In colonized as well as more autonomous countries, universities accommodated research departments where human sciences took roots. Psychology was lectured at the University of Tokyo in 1877 and became part of the curriculum in 1879. The first laboratory of psychology was established at the University of Calcutta in 1905 and at the National University of Peking in 1917. Sociology, accordingly, was institutionalized in the same decades, with Ernest Fenollosa teaching American-styled sociology at the University of Tokyo in 1878 and Masakazu Toyama becoming the first professor of sociology on 1893. In China, sociology had been lectured since 1905 on a rather irregular basis. A department of sociology opened in 1913 at Huijiang University, with mainly foreign academics, and in 1921, the first department with a completely Chinese staff was inaugurated at Xiamen University. In India, sociology developed, in the early decades of the twentieth century as the colonial authorities had to classify the population: a department was established in 1919 at the University of Bombay and several others followed elsewhere.

During the same decades, associations and journals began blooming. China provides an illustration of this global process, where a Fund Board of Chinese Education and Culture, with several social research institutes, was established in 1924. Accordingly, the *Chinese Society for Psychology* was created in 1921 and published a journal called *Xinli*. Chinese anthropologists also set up various societies in the 1920s; in 1930 they merged and became the *Chinese Society for Folklore Studies*, which published its own journal in the subsequent years (Blowers, 2006).

Most of the time, academic disciplines became national. Some attempts to create transnational or regional associations nevertheless took place at this time, for instance, in 1930 when the South-East Asian Association of Sociology replaced the Chinese Association of Sociology. But it is mainly after the Second World War that this trend gained momentum. Besides, the de-Westernization of human sciences that took place in the interwars did usually not lead to challenge their epistemological tenets. True, some Japanese scholars called for distinct Asian human sciences in the 1930s but, in a context of Japanese imperialism, the move was politically motivated. In fact, most of the local scholars involved in the creation of national disciplines during the first half of the twentieth century aimed at replicating European sciences as accurately as possible. Even though they could have a critical stance, they deemed more important to fully understand them and to provide their country with this central outcome of modernity.

As important as universities and associations, was the development of academic journeys and scientific pilgrimages. Local authorities, provided they had enough political leverage, sent students and scholars to the West in order to study the new disciplines. Muhammad Ali sent the first Egyptian mission to Paris in 1826. After 5 years in France, its chaplain, Rifa'a al-Tahtawi (1801–73), wrote the *Takhlis al-Ibriz fi-Talkhis Bariz* (translated as *A Paris Profile*), which became a landmark book in the Arab rediscovery of European culture. Tahtawi was then appointed head of a translation office that contributed to the acclimation of Western philosophy and human sciences. Egyptian students missions continued during the next century and allowed generations of scholars to gain familiarity with European epistemological systems, especially human sciences. Likewise, the Japanese government sent academic missions to Europe and the United States, following Emperor Meiji's declaration that his subjects 'should seek knowledge all over the world.' In 1871, officials and scholars took part in the two-year-long Iwakura Mission. Sixty students were brought along and some of them were left in various countries to complete their studies. When they came back to Japan they set up colleges and universities, translated essays and books; some became influential scholars. Accordingly, the first translators of human sciences in Japan were, directly or indirectly, influenced by the outcome of the mission. As such travels proved crucial to the intellectual modernization of the country, Japan issued more than 10 000 students passports until the end of the nineteenth century.

The Western powers also perceived academic exchanges as an important question of international politics. The United States used a large part of the money remitted by China after the Boxer War to sponsor students' scholarships to the US. Consequently, most of the first Chinese sociologists or psychologists had studied at Columbia or at the University of Chicago and, when they came back to China, they contributed to set up associations and academic departments. If the Chinese imperial authorities had only reluctantly let students go to the United States, the subsequent governments were much more willing to develop exchanges, as they were now supporting the modernization of China's universities (Ye, 2001).

Modern academic pilgrimages clearly show that, at the end of the nineteenth century, intellectual and scientific issue had become a major political concern and that the West was at the center of intellectual exchanges. Granted, nonofficial journeys still existed: when Jamal-Eddine Al-Afghani and Mohammad Abduh toured Europe in the 1880s, they did it without any state support. But this was becoming more and more the exception, as official grants and then the support of private foundations became the main source of mobility. Accordingly, not all journeys brought scholars and students to the West: many continued to go to non-Western countries. For instance, around 10 000 Chinese students went to study to Japan during the first decades of the twentieth century. This was crucial for the development of human sciences in China, for the Japanese had translated many notions that their Chinese peers borrowed. But the journey to Japan was not the goal in itself: the country was seen as the shortest way to European modernity and its most successful illustration.

International and Transnational Networks

To a large extent, the de-Westernization of the human sciences that took place at the nineteenth and the beginning of the

twentieth century was tantamount to a nationalization of European disciplines, in which local authorities played a key role. Yet de-Westernization was also a process in which international and transnational actors took a large part, sometimes even before the implementation of national disciplines. International scientific associations were instrumental in the creation of national associations of human sciences in non-Western countries. But they were not the only ones as nonstates actors were also involved: this is the case of churches on the one hand, and of movements such as Marxist or feminist networks, on the other.

As the previous example of academic journeys highlight, the de-Westernization of human sciences occurred through international contacts. National authorities tried to control the process, but they had to collaborate or to compete with international actors. The internationalization of the scientific life began at the nineteenth century, as an attempt to overcome the fragmentation of the academic community (Guilhot et al., 2008). The first international scientific associations were created before the twentieth centuries but they remained almost exclusively European until the 1950s. The only exception, here again, were the Japanese who took part in international associations from the nineteenth century on. In 1873, at the first International Conference for Oriental Studies, the Japanese delegation was the only non-Western one. Likewise, all the participants to the International Congress of Psychology, which took place between 1889 and 1948, were European and North American, except for the Japanese delegates, starting with Motora Yuijiro in 1905. In 1948 only a Chinese psychologist from Beijing, Oui Tsing-Chen, officially attended the Conference. This does not mean that no non-Western scholars at all attended these conferences, but, most of the time, they did it on a private basis. Besides, international political cooperation came to a halt in the 1920s and this impinged on the scientific life. Several countries where modern academic life was burgeoning, like India, Egypt, or China, could therefore not participate in the international scientific exchanges.

International cooperation resumed after 1945 under the supervision of the UNESCO and its International Social Science Council. The International Union of Anthropological and Ethnological Sciences was created in 1948, the International Sociological Association and the International Political Science Association in 1949, and the International Union of Psychological Science in 1950. Their goal was not only to bring together the existing national associations, but also to actually implement them where they did not exist. The non-Western nations had an important place in this process: they were becoming independent and needed scientific expertise to build modern states and societies. The UNESCO, moreover, played a crucial role in the promotion of non-Western cultures, and published several of their classics in literature and philosophy. It was committed to promote the equal dignity of all cultures and did so with the help of renowned social scientists (see Lévi-Strauss et al., 1952): human sciences, which had been instrumental in the colonization of non-Western people, had the opportunity to critically reflect on their function. For the last decades, the UNESCO and its associations have allowed non-Western scientists to play a more important role in the scientific life; it is still the case today: its *World Social Sciences Reports* still give non-Western voices the place they deserve and, under its supervision, non-Western scholars have taken part in international conferences and set up Asian, Arab, or South American research associations.

Likewise, nonstates actors have played an important role. Churches were instrumental in the acclimation of European disciplines at the nineteenth century. American missionaries taught psychology in China decades before it was part of the curriculum of national universities there. Admittedly their goal was not to introduce a scientific discipline per se but to foster a new conception of the self that, they hoped, would lead to a new work ethic. Other religious missions, like in Egypt, were truly involved in the dissemination of modern sciences and have been acting as intermediaries between the Egyptian scholars and their Western colleagues. Non-Western scholars and officials took the role of Western churches seriously. In Japan, where Japanese converts and Western missionaries had been instrumental in the creation of some of the major universities, the religious issue prompted a heated debate. Fukuzawa Yukichi (1835–1901), the father of Japanese scientific modernity and the founder of Keio University, was so convinced of the key role of religion that he advocated the formal conversion of Japan to a Christian state.

Other intellectual networks took an important part in the acclimation of human sciences and the empowerment of non-Western voices. Despite Marx's ethnocentric views, which can be found in some of his earlier texts, Marxism quickly spread in non-Western countries. Its influence grew beyond the political arena and reached the field of human sciences. One of the main reasons for this success is to be found in the ambivalent position of Marxism with regard to Western modernity, as it was both a direct outcome of this modernity and critical of its many shortcomings. This ambivalence was appealing to non-Western scholars who were aware that modernization was inevitable but sought a way to reassess the power relations underlying the diffusion of Western modernity. Indian historiography relied heavily on the Marxist point of view to deconstruct the English history of India. Whereas colonial works had stressed the role of local elites and had dismissed endogenous processes, Marxism helped a new generation of Indian scholars to pay attention to lower-social classes and the antagonisms that had shaped the building of the Indian nation. In Japan, likewise, Marxism gained influence in the human sciences in the 1910s and the 1920s as the Japanese were translating the debate on the birth of capitalism launched by Max Weber or Werner Sombart. It helped them to account for the class conflicts that were latent at this time and to address their country's special position in a system of global economic asymmetries. The military government ruthlessly cracked down on Marxists but they reflourished after 1945, as they had critical resources to analyze the rise of Japanese militarism as well as the numerous inequalities that the Japanese 'economic miracle' had not removed (Barshay, 2007).

The feminist movement played a similar role. At the juncture of politics and academy, it allowed for a quick dissemination of theoretical texts and empowered dominated voices from the South. Feminism has long history but was decisively reshaped in the West in the 1960s; it then became global through an internationalization of the political fight and the worldwide influence of American gender studies (Marx Ferre and Tripp, 2006). This has favored the transdisciplinary

circulation of theories, drawing from sociology, psychology, literary studies, etc. Worldwide connections were obvious, for gender-based dominations seemed to supersede determinations such as national and cultural identities. Yet the rise of 'global feminism' has not completely displaced the opposition between Western and non-Western countries. Many feminists from the South asked whether the narratives from the North could accurately describe the experience of women there. Some radically called into question the very possibility for such textual apparatus to give access to the subaltern women's voice (Spivak, 1988). The diffusion of feminist theories therefore gave rise to epistemological debates in which non-Western scholars intended to challenge the Western bias at the heart of human sciences: some radically dismissed any European-styled theory whereas others pleaded for pragmatic solutions.

Both Marxist and feminist networks highlight the role played by transnational diaspora in the diffusion of human sciences. Many non-Western Marxists have been moving between their country and the West, where they studied or, for some of them like the Arab Marxists, found shelter. This is also the case with feminism, Gayatri Spivak (1942–present), a Calcutta born scholar who teaches in New York, being probably the most famous. But transnational connections have allowed for important exchanges in other disciplines. Many of the Indian scholars, who founded the Subaltern and Postcolonial Studies, have taught and lived in the West. Their position not only contributed to the hybridization of Western and Indian history; it also allowed them to take a critical distance with their country's mainstream historiography. As they had more freedom to reevaluate the nationalistic discourse and the modernist bias on which Indian Marxist history was premised, the Subaltern historians reassessed the role of social classes previously dismissed as 'backward.' Likewise, many Arab scholars who took part in the postcolonial reevaluation of Orientalist history were Diaspora intellectuals, such as Anouar Abd-el-Malek (1924–2012), a Paris-based Egyptian Marxist, or Edward Said (1935–2003), who lived and taught in the United States. These Diaspora intellectuals were deeply familiar with the languages and culture of the West and, at the same time, felt 'out of place' there: 'outsiders as insiders' they turned this ambivalent position into a critical awareness of the European knowledge.

Politics of De-Westernization: 'Nation,' 'Socialism,' 'Culture'

Western human sciences prompted the global diffusion of political concepts and references, which, in turn, allowed for de-Westernization programs. The ideas of 'nation,' 'socialism,' and 'culture' have been among the most labile and influential ones: acclimated, they allowed for critical reflections on what human sciences could and should be outside of their original context.

The nation could not have been 'imagined' without human sciences. Indian rulers could not have turned a collection of kingdoms into a nation without the English statistics. Far from challenging the categories of the colonizers, the Indian administration endorsed their view of the society. Whereas castes had been rather fluid entities, they became fixed status into which Indian citizens were to be classified. Indian sociologist played a key role in this process. The discipline developed as an outcome of colonial anthropology without reassessing its underlying culturalist bias. The founding fathers of Indian sociology reified anthropological entities and developed an understanding of the Indian society based on conservative Hinduism and Brahmanism (Patel, 2010). India became a nation organized according hierarchies sociology had largely reconstructed.

Accordingly, many new independent East-Asian nations used colonial statistics to launch their modernization. In socialist countries, governments and anthropologists used colonial censuses to target 'backward' people: ethnic groups were to be identified and assimilated to the new society (Van Bremen et al., 2005). Ethnicity was to a large extent a creation of colonial administrators, for there was no homogeneous population behind the categories. Yet both the new national states and, surprisingly, the tribes themselves, began using these categories. The latters understood what they could gain from the statistical identification, as it gave them leverage to oppose assimilation policies or, on the contrary, to claim a fair share of the national income.

Socialism provides a second illustration of an idea that, once rooted outside the West, both favored the acclimation of human sciences and influenced de-Westernization programs. As said before, Marxist networks acted as early channels of scientific diffusion. After 1945, numerous non-Western intellectuals sided with the Communist parties, which had been involved in the anticolonial fight. Those who did not live in socialist countries used Marxism to address scientific asymmetries. In the Arab world, scholars such as Anouar Abd-el-Malek or Mohamed Talbi (1921 to present) used Marxist epistemologies to critically reread colonial scholarship. In Japan, likewise, Marxism allowed sociologists to oppose the American functionalist paradigm and offered critical perspective that enhanced the scientific creativity.

Besides, as many non-Western countries joined the socialist nations, socialism shaped the frame into which these states modernized and opposed Western hegemony. Various brands of Marxism provided the guidelines of the new human sciences. China offers the most striking illustration of this process. In 1952 most human sciences were targeted as 'bourgeois sciences.' The scholarly activity became state-controlled in the name of the edification of socialism. Scientists were discriminated, especially in paroxystic moments such as the '100 Flowers' or the Cultural Revolution, and saw their autonomy significantly restrained, which was detrimental to theoretical innovation. Sociology, for instance, practically shut itself to critical approaches and statistics became the only legitimate way to account for the society.

Not all human sciences were dismissed though. Linguistics was crucial to the elaboration of a socialist nation, as the renewal of the Chinese script was necessary for the largest part of the population to have access to the written language. Chinese scholars were nevertheless asked to severe ties with Western theories. They found inspiration in Soviet linguistics but, when the relations between the two countries deteriorated, had to find other paradigms. They rediscovered Chinese treatises and began elaborating their own brand of linguistics. International contacts resumed in 1979 when Deng Xiaoping

(1904–97) opened a new era of modernization and decided that the country had to catch up with Western sciences. Yet this did not lead to a complete re-Westernization of human sciences. Chinese scholars were still committed to find proper ways to account for their culture, as well as to take the specificities of socialism into account. This is, for instance, what led some sociologists to launch research programs premised on the idea of a specific 'socialist civilization' (Merle, 2004).

'Culture,' accordingly, decisively influenced de-Westernization programs. The idea of culture, as European ethnologists had elaborated it, spread through a variety of channels. Chinese ethnologists, for instance, created their own notion by drawing from traditions as different as the *Kulturkreislehre*, the Durkheimian School or the American diffusionism. All countries quickly assimilated the notion of culture as, in the nineteenth century, the Western scientific gaze was grasping them under this category. The idea was highly appealing to local rulers and scholars: it allowed them to assert a sense of fundamental difference with the West at a time of threatening Westernization. Folklorist educated in the West played a key role. They went to the remotest parts of the country to gather old tales, forgotten idioms and customs, and reconstructed a cultural legacy that was crucial for the nationalist narrative. Countries that had been among the first modernizers were also the first to assess the ancient nature of their culture. In Egypt, anthropologists reconstructed the pristine figure of the *Fellah*, as being immemorially linked to the Egyptian land. In Japan, likewise, the Westernization of the country went hand in hand with a rediscovery – which was partly a re-creation – of Japanese culture. Some recipes, martial arts, or drawing techniques had already been codified as 'typically Japanese' at the eighteenth century. This trend gained momentum with the importation of European anthropology and many social activities came to be redefined as belonging to the Japanese cultural legacy.

Culture easily allowed for counter-hegemonic critics. As it gave a scientific basis to the difference between Europe and other countries, non-Western scholars could legitimately claim that they were in the best position to study 'their own culture.' This argument became powerful after the decolonization, when Western researchers did not get an easy access to the field anymore and local ones pressed for significant changes in the anthropological discipline. More recently, 'culture' became a key word for new de-Westernization programs. Countries like Iran or Singapore reframed their political opposition to the West in culturalist terms. In the 1990s, the Singaporean government launched the 'Asian values' campaign to assert Asia's different political and cultural standards. The campaign led to attempts to build alternative sciences. Some Singaporean sociologists called to reframe Asian sociology, for, they argued, the European founders of the discipline were of little help to understand Asia's specific problems at the end of the twentieth century. Accordingly, the Iranian leaders tried to promote scientific and cultural de-Westernization programs after the Islamic Revolution. Under the label of the *Jahad-e-Daneshgahi* (scholarly Holy War) they intended to Islamize the Western epistemic frame. European human sciences had been acclimated under the Shah, when the country was to be modernized at any price. The revolutionaries logically sought to remove this symbol of cultural alienation. They had the universities' curricula rewritten and tried to find Muslim sources to reframe human sciences. The program nevertheless felt short of its promises. Iranian scholars, even the most loyal to the Revolution, were committed to European theories: they had already found ways to conciliate their Muslim culture with the promises of Western scientific enlightenment. Some paid lip service to the cause of the Islamic scientific revolution but, when it was possible, began to openly acknowledge the necessity to take Western human sciences seriously. As the reformist movement gained momentum, they finally came with some of the most innovative propositions to blend Islamic and Western theories (Mahdi, 2010).

Intellectual and Argumentative Strategies of De-Westernization Programs

De-Westernization programs prompted as many intellectual and argumentative strategies as there were different countries and time periods involved. Nevertheless, it is possible to gain a better understanding of these programs by subsuming them into four main (nonexhaustive) categories.

De-Westernization programs arose, first, from the *early reactions to European imperialism and intellectual hegemony*. In the Muslim world, from the 1850s on, scholarly reactions to the disruption of local knowledge followed two lines of reasoning that are still prevalent today. Some intellectuals claimed that Western scholarship was unable to grasp the specificity of the Islamic culture; the fact that it had been brought though imperial conquests was another important argument to disqualify this type of knowledge. Other scholars, on the contrary, argued that it was necessary to make the difference between the political acclimation of these theories and their heuristic value. For them, they could help the Muslims to get a better understanding of their civilization and to rebuild Islam. Conservative and reformist scholars differed when it came to what should be done to reshape the intellectual life, even though all agreed that they had to take human sciences seriously.

In Japan, as well, scholars began to reassess Western episteme in the first decades of the twentieth century. The country had gone through one of the swiftest scientific acculturations: the followers of the founders of Japanese human sciences began pondering on what had happened. Some argued that they should elaborate theoretical frames that fit the specificity of their society and, sometimes, browsed ancient Asian texts to find conceptual inspiration. Some scholars even developed the idea of an encompassing pan-Asian intellectual unity, from Tokyo to Istanbul, but it did not manage to challenge the European systems, as it was premised on the assumption of Japanese imperial supremacy.

In almost all countries, de-Westernization programs took place as Western epistemologies were reframing local knowledge. Yet these early attempts have no real unity except for the fact that they happened in the same historical context. Besides they had little direct effect, as scholars thought more important to understand Western human sciences than to challenge them. They were nevertheless important, because they show that non-Western intellectuals were perceptive of the epistemic reconfigurations of modernity and tried to curb them: they developed

arguments that, in the following decades, became resources their followers tapped into.

A second set of de-Westernization program was linked to the *decolonization of human sciences*. The implementation of national universities and research centers was an important step, and non-Western people, who had been the object of Western human sciences, became active subjects in their production. Accordingly, the nationalization of disciplines implied a rupture with their European origins and favored the recognition of scholars from postcolonial nations. There was a strong legitimacy to do so: human sciences had been instrumental in the colonial rule and a new scientific apparatus was therefore necessary to understand decolonized societies. Scholars began rereading the texts Western researchers had written on their country, with dramatic consequences sometimes. Orientalism practically ceased to exist when non-Western researchers asked why it had been incapable, during almost two centuries, to integrate the scholars coming from the very cultures it was studying. Anthropology, likewise, had to reassess its ethnocentric bias. Yet this critical introspection reinforced the discipline: Western and non-Western scholars came to work together and improved their understanding of non-Western societies. Such collaboration was easy to develop, for many Western anthropologists had taken part in the anti-colonial fight and knew the shortcomings of colonial sciences. This example also shows that, despite some attempts to fully replace Western epistemologies, the scientific decolonization, rather, led to reconsider them. The existing disciplines were not abolished, nor the type of rationality they were premised on: but, in terms of content, the way they accounted for non-Western cultures was significantly different.

Scientific decolonization was a strong motto in the new independent nations, but it soon faltered. Postcolonial states relied on modern knowledge to modernize their societies and could not completely severe ties with Western-styled sciences. Besides, the disenchantment with political decolonization impinged on the scientific one. The very scholars who had been truly committed to de-Westernization began to reconsider their positions. In the West, yet, intellectuals of the Diasporas, such as Edward Said and some Indian scholars based in the United States, revived the project of scientific decolonization under the guise of the postcolonial studies. They rephrased the debate that took place in the 1960s into the theoretical frame of the American academia at the then end of the twentieth century. They found new spaces for (post)colonial voices, at the margins of literary studies and human sciences, and could claim that scientific decolonization was still an unfinished journey (if not an unfinishable one).

A third type of de-Westernization programs has been premised on what can be called *priority disputes*. As they acclimated human sciences, non-Western scholars dug into local ancient texts to find equivalents of European concepts. This led them to 'go back to the classics' and to claim that the supposed novelty of human sciences could be traced back to the past of other civilization. Fei Xiaotong (1910–2005), one of the founders of Chinese sociology, compared Durkheim and Confucius in the 1920s: for him, Durkheim's notion of society, understood as a set of collective determinations, was similar to the emphasis that the Confucians had put on the commonality of human life. In the Arab world, likewise, sociologists compared European sociology with Ibn Khaldun's (1332–1406) theories and saw in the latter the true father of sociology. The comparisons did not always have the same function. In the first case, the contrast between Confucius and Durkheim was meant to introduce the latter's ideas to China: as Chinese readers were familiar with Confucius, the commentators of Durkheim hoped that they would gain a better understanding of the French sociologist. With Ibn Khaldun, the strategy had a distinct political tone. It implied that the modern sciences, which were supposed to express European superiority, were not as modern and not as European as they looked like. The Arab and Islamic civilization, the argument goes, had the same kind of intuitions earlier and deserved credit for their role in the rise of scientific modernity. Such priority arguments often mixed the two strategies to some extent though: Ibn Khaldun was also used to introduce modern sociology and Confucius to politically oppose the claim of European superiority.

Among all disciplines, psychology has probably been the most extensively reassessed in an historical light. Asian countries have traced the roots of *their* psychology to Confucius and the Buddhist Canon. The Indians have located it in the classical texts of Hinduism and Muslim scholars have written countless essays on psychology and the Koran. All claimed that the topics psychology apparently deals with (desire, fear, death, etc.) had already been addressed in their own traditions. Yet there are deeper reasons for this reconstruction. Western psychology was introduced in its experimental form, with the idea that a true understanding of the human mind has more to do with laboratory experiments than with symbolic questions. This hardcore epistemology came under strong attacks soon, including in the West, as it treated the psyche like any other material being. An alternative psychology soon emerged, which paid more attention to the role of language and history. For non-Western psychologists, it was an opportunity to reassess the role of culture in the formation of the self. They argued that the mind is structured by linguistic categories, which are not the same as in all languages. Besides, they claimed that family patterns, religions, or power relations variously shape the psyche. Their argument was scientifically and politically sound. It allowed them to challenge modern psychology and to rebuff the claim that the West had the key to the universal self.

A fourth set of de-Westernization programs has focused on the idea of *alternative or multiple modernity* (Parameshawar Gaonkar, 2001; Eisenstadt, 2000). As human sciences are a typical outcome of modernity, non-Western scholars argued that different types of modernity, and their corresponding scientific patterns, are possible. In its most radical version, this program intended to challenge the epistemological basis of human sciences. The aforementioned Singaporean program was premised on this line of reasoning; so were the ones some Chinese scholars developed from the 1990s on. They often originated in Asian universities, as many Asian countries had become as economically powerful as the West and could argue that they had developed their own modernity. For them, Asian modernity was different from the Western one: the Chinese and Japanese people were held less individualistic than the Europeans, and were supposed to have given rise to a significantly different type of society. The classical concepts that the founding fathers of human sciences had elaborated one century

before in Europe were therefore inadequate to the current Asian situation. Some Chinese sociologists, for instance, claimed that classical sociology was incapable to grasp China's entry into modernity, as it did not occur according to the liberal and capitalist scheme, but through a socialist and culturally distinct one.

These programs had different outcomes. It proved difficult to prompt a complete shift of paradigm and Western notions continued to loom in the background of the new human sciences. Besides, many non-Western scholars were skeptical. For them, total de-Westernization was inspired by conservative politics and, on a scientific level, was not as creative as some had hoped. Some scientists argued that many positive outcomes of modernity were still missing in Asia and that it was too early to dismiss the Western experience. Other scholars began to work on more realistic de-Westernization programs: for them, the point was not to replace the European human sciences but to accommodate them to non-Western situations. The European scientific apparatus should be retained because it is flexible enough to be transformed from the inside, with non-Western scholars creating new concepts to account for their cultures. Current developments in Chinese sociology point to that direction: Chinese scientists try to hybridize Western concepts and to promote new ones that can grasp the specific modernity of China.

To some extent, the debate about alternative modernities has echoed the reflections on postmodernity that developed in North American universities in the 1980s. Not only the Japanese or Chinese, who studied there, but Iranian scholars too, have been influenced by these questions. Postmodernity implied that the Western epistemic centrality had been overcome. Besides, postmodern thought has been premised on the notion that no narrative is universal, not even the scientific one. European Western sciences, accordingly, are one narrative among others, with no intrinsic superiority. Many conservative Iranian scholars, albeit politically adverse to the United States, took part in the debate and concluded that the modern scientific apparatus is nothing but the result of the specific cultural conditions of Europe. The fact that American scholars had originally elaborated postmodern thinking, seemed to give more credit to the idea. If the postmodern debate has bloomed in the field of literary studies more than in the human sciences, it was nevertheless significant enough to legitimate the idea that the latter were an outcome of Western modernity with no claim to general validity.

The de-Westernization of human sciences is as old a program as their very diffusion of the human sciences. Yet it has gained momentum in the last decades, as the centrality of the West was contested. Following the postcolonial debates, the scientific community came to better associate Western and non-Western researchers. Accordingly, there has been a realization that Western epistemologies, despite their worldwide diffusion, were not the only way to account for the reality: different epistemic patterns used to prevail among non-Western cultures and could reframe alternative scientific systems one day. It is nevertheless too early to conclude that these programs have significantly challenged the global scientific asymmetries. The United States and Europe are still hegemonic, with around 90% of all the articles of human sciences written between 1998 and 2007. Despite the significant rise of the publications by Asian and South American scientists (respectively +74 and +46%), the Western share has recently increased. Moreover local and regional collaborations have decreased, as more and more non-Western scholars now work with their Western colleagues and/or in a Western language (Gingras and Mosbah-Natanson, 2010). Countries like India or China, who intend to become major scientific players, continue to train their scholars in Western universities and have them write in English.

This ambivalent situation indicates that the idea of de-Westernization is today at a crossroad. On the one hand, the exhortation to 'provincialize Europe' and the promotion of epistemic pluralism, are enjoying unprecedented influence. On the other hand, Western scientific hegemony has reinforced and, statistically speaking, appears barely contested. The current critics against the West might hint less at the impending end of the European and North American scientific centrality, than at its continuity in a new period of globalization. The situation also exemplifies the different outcomes of the de-Westernization programs launched in the last decades. The most radical attempts to replace Western human sciences have not managed to actually implement alternative epistemologies. Others sought to reframe human sciences from the inside and were more successful. They argued that, as Western modernity has reshaped the world, it is important to use the tools this modernity has brought to make sense of the current situation. Besides, they stressed that Western human sciences are politically cogent, for they allow critical insights, including on what has made the West dominant. In this respect, the apparently abiding centrality of Western human sciences may hide significant reconfigurations at a deeper level.

See also: Anthropology: Overview; Area Studies, History of; Ethnocentrism; Globalization and World Culture; Scientific Disciplines, History of; Social Science Learned Societies and Professional Associations; Translating Sociology into Japanese, Chinese and Korean Languages.

Bibliography

Alatas, Syed Farid, 2006. Alternative Discourses in Asian Social Sciences. Responses to Eurocentrism. Sage.

Appadurai, Arjun, 2013. The Future as Cultural Fact, Essays on the Global Condition. Verso.

Aydin, Cemil, 2007. The Politics of Anti-Westernism in Asia ,Visions of World Order in Pan-Islamic and Pan-Asian Thought. Columbia Unviersity Press.

Barshay, Andrew, 2007. The Social Sciences in Modern Japan, the Marxian and Modernist Tradition. University of California Press.

Blowers, Geoffrey, 2006. Origins of scientific psychology in China. In: Brock, Adrian C. (Ed.), Internationalizing the History of Psychology. New York University Press, pp. 94–111.

Chakrabarty, Dipesh, 2000. Provincializing Europe, Postcolonial Thought and Historical Difference. Princeton University Press.

Eisenstadt, Shmuel (Ed.), 2000. Multiple Modernity. Transaction Publishers.

Elshakry, Marwa, 2014. Reading Darwin in Arabic, 1860–1950. University of Chicago Press.

Gingras, Yves, Mosbah-Natanson, Sebastien, 2010. Where are social sciences produced?. In: World Social Sciences Report. UNESCO, ISSC, pp. 149–154.

Guilhot, Nicolat, Heilbron, Johan, Jeanpierre, Laurent, Spring 2008. Towards a transnational history of the social sciences. Journal of History of Behavioral Sciences 44 (2), 146–160.

Levenson, Joseph, 1968. Confucian China and Its Modern Fate. University of California Press.
Lévi-Strauss, Claude, et al., 1952. Race and History. UNESCO.
Marx Ferre, Myria, Tripp, Aili Mari (Eds.), 2006. Global Feminism, Transnational Women's Activism, Organizing and Human Rights. New York University Press.
Mahdi, Ali Akbar, 2010. Sociology in Iran : between politics, religion and western influence. In: Patel, Sujata (Ed.), The ISA Handbook of Diverse Sociological Tradition. Sage Studies in International Sociology, pp. 268–277.
Merle, Aurore, 2004. Vers une sociologie chinoise de la 'civilisation communiste'. Perspectives Chinoises (online article).
Parameshawar Gaonkar, Dilip (Ed.), 2001. Alternative Modernities. Duke University Press.
Patel, Sujata, 2010. At crossroads : sociology in India. In: Patel, Sujata (Ed.), The ISA Handbook of Diverse Sociological Tradition. Sage Studies in International Sociology, pp. 280–291.
Pouillon, Francois, Vatin, Jean-Claude, 2011. Après L'Orientalisme : L'Orient Crée Par L'Orient. Karthala.
Said, Edward, 1978. Orientalism. Pantheon Books.
Spivak, Gayatri, 1988. Can the subaltern speak? In: Nelson, Carry, Grossberg, Lawrence (Eds.), Marxism and the Interpretation of Culture. University of Illinois Press, pp. 66–111.
Van Bremen, Jan, Ben-Ari, Eyal, Alatas, Syed Farid, 2005. Asian Anthropology. Routledge.
Vandermeersch, Léon, 2013. Les deux raisons de la pensée chinoise: divination et idéographie. Gallimard.
Ye, Weili, 2001. Seeking Modernity in China's Name, Chinese Students in the United States, 1900–1927. Stanford University Press.

White Collar Crime: Recent Trends and Debates

Michael L Benson, University of Cincinnati, Cincinnati, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Recent trends in white-collar crime include a shift in theoretical orientation to opportunity and choice perspectives, a rising interest in the psychology of white-collar crime, research focused on the role of white-collar crime in the global financial crisis, and a continuation of research interest on white-collar offenders' use of neutralizations and accounts. Debated issues include the efficacy of different regulatory styles in controlling white-collar crime, gender and white-collar crime, and the applicability of low self-control theory to white-collar offenders.

Introduction

Although the problem of how white-collar crime should be defined is an old one, no summary of developments in this area, recent or otherwise, can avoid it. Definitions of white-collar crime can be divided into two main groups – offender based and offense based. Following Sutherland (1949), the offender-based approach defines white-collar crime in terms of the social, economic, and occupational characteristics of offenders. One well-regarded definition, for example, defines white-collar crime as "violations of law to which penalties are attached that involve the use of the violator's significant power, influence, or trust in the legitimate economic or political institutional order for the purpose of illegal gain or to commit an illegal act for personal or organizational gain" (Reiss and Biderman, 1981). This definition specifically requires that the offender have and use 'significant power, influence or trust in the legitimate economic or political institutional order' in committing his or her violation of the law. It follows that individuals lacking these characteristics cannot be white-collar offenders. Offense-based approaches ignore the characteristics of the offender and focus instead on the nature of the act. The most well-known offense-based definition was proposed by Edelhertz (1970), who defined white-collar crime as an "illegal act or series of illegal acts committed by nonphysical means and by concealment or guile to obtain money or property, to avoid loss of money or property, or to obtain business or personal advantage." The key element in this definition is the stipulation that the offense be committed by 'nonphysical means and by concealment or guile.' Because the characteristics of the offender are not included as part of the definition, individuals from all walks of life could, in theory, commit a white-collar crime.

Both types of definitions have their respective strengths and weaknesses. The offender-based approach draws attention to the serious offenses committed by high status individuals that are all too often ignored in mainstream criminology, but it also arbitrarily restricts focus to just these offenses while ignoring similar offenses committed by individuals from less elevated social backgrounds. The offense-based approach not only avoids this arbitrary restriction, but it also broadens the conceptual terrain of white-collar crime to such a degree that the term loses its significance as a focus of theory and research. In practice, the offense-based approach often misses the crimes of the powerful and substitutes in their place small-time offences, such as frauds by appliance repairmen and employee theft by retail clerks. Benson and Simpson (2009) note that although there is a long-running and vigorous debate over which definition is best, the two approaches are not necessarily mutually exclusive. The white-collar offenses committed by high status individuals rely on the very means identified by Edelhertz's (1970) offense-based definition – concealment, guile, and deception. Researchers following Edelhertz's approach could study high status individuals if they would take the time and effort to find them. This often does not happen reflects a failure of the researchers rather than a logical weakness of the offense-based approach. Nevertheless, the debate is not yet resolved and any summary of recent developments in white-collar crime must recognize that different definitions are used in the field, and that this influences the types of findings and trends that one observes.

Trends

Recent trends in the study of white-collar and corporate crime can be divided into two broad categories: (1) developments in substantive areas of focus and (2) evolution in theoretical perspectives. With regard to substantive areas of focus, scholarship in white-collar crime often tends to follow current events and to be directed toward the latest social, political, or economic scandals, because such scandals often involve white-collar type offending in one way or another. Thus, spurred by the global financial collapse, a considerable amount of research attention has been focused recently on national and international financial- and market-based crimes that are thought to have contributed to, if not actually caused, the current crisis. Theoretically, scholarship on white-collar crime has been evolving in directions that have obvious implications for each other and may eventually merge. First, there is the increasing dominance of the opportunity or choice perspective over the traditional criminological theories as the preferred explanatory framework for understanding white-collar offending. Secondly, there has been a surprising flurry of interest in the psychological and individual level attributes of white-collar offenders compared with other types of offenders and law-abiding citizens. Finally, research on how white-collar offenders' use of neutralizations and accounts may be shaped

by gender and social class has reinvigorated this traditional area of interest.

Mortgage Fraud, Financial Crimes, and the Global Financial Crisis

The financial crisis that hit the United States in 2008 as a result of the collapse of the housing market has provoked considerable and ongoing commentary (Levi, 2010a; Benson, 2010; Shover and Grabosky, 2010). It seems relatively clear that the global financial crisis was sparked in part by massive fraud and unconscionable risk-taking in the mortgage and finance industries during the 1990s and early 2000s (Financial Crimes Enforcement Network, 2006; Partnoy, 2003). But it is unclear whether the financial crisis itself necessarily led to more white-collar offending. Levi (2010a), for example, argues that the financial crisis was not really global. Rather, it hit some nations much harder than others. He further contends that rather than sparking more white-collar crime, the financial crisis may simply have brought to light what was already happening. The distribution and potential collateral consequences of mortgage fraud has provoked research on ecological variations in the prevalence of mortgage fraud in neighborhoods (Carswell and Bachtel, 2009), as well as research on whether the high rates of bankruptcies that resulted from the collapse of the housing market led to increased street crime in disadvantaged neighborhoods (Baumer et al., 2012).

Much of the research on mortgage fraud has focused on the subprime mortgage market, where fraud seemed to be most rampant. Findings indicate that several different factors were involved in the explosion of mortgage fraud, including inadequate regulation, the proliferation and overuse of alternative loan products, and a general lack of accountability in the mortgage industry (Nguyen and Pontell, 2010). These factors permitted unscrupulous mortgage brokers and profit-hungry lenders to engage in predatory lending, a form of lending in which loan makers and their associates use deceit to push financially ill-advised loans on unsuspecting borrowers. Such loans end in default at a much higher rate than traditional loans.

Although mortgage fraud was rampant, by itself this type of fraud might not have had such far-reaching and detrimental financial effects had it not been accompanied by other developments. The real problem involved changes in the mortgage and financial industries, in particular the practice of pooling mortgage loans and treating them as securities. These mortgage-backed securities or collateralized mortgage obligations (CMOs) as they are also known, were put together by large Wall Street financial firms and sold to investors. Under this system, banks and other lenders no longer had much incentive to carefully screen loan applications and to approve only those that were financially viable. Rather, it made more sense to make as many loans as possible regardless of the financial status of the borrowers because even bad loans could be off-loaded to the market for CMOs. Unfortunately, CMOs are so complex that it is exceedingly difficult for even highly experienced investors to evaluate the risks associated with them, making it easy for those who market CMOs to fraudulently misrepresent or conceal their true value. This apparently is exactly what happened in many cases; when the housing bubble finally collapsed in 2008 and borrowers began to default, many CMOs lost significant value overnight, threatening the financial security of their holders.

A central theme of all of the recent research on the role of white-collar crime in the various financial crises that have rocked the US economy in the past two decades has been the inadequacy of oversight and regulation of the corporate sector. Regardless of whether the focus is on fraud in mortgages, mutual funds, investment banks, energy companies (Enron), or telecommunication giants (WorldCom), a root cause of the scandals in all of these different arenas has been the lack of effective regulation and supervision and the over-reliance on 'private market discipline' to control abuses (Levi, 2010a). In effect, ineffective regulation creates criminal opportunities for white-collar offenders.

Opportunities and Choices

Although virtually all of the major theoretical perspectives in criminology have been applied at one time or another to white-collar or corporate crime (for a brief review, see Benson and Simpson, 2009), most have not fared well in regard either to their explanatory scope or their ability to motivate a sustained body of empirical investigations. To apply traditional criminological theories to particular instances of white-collar crime often requires 'extensive conceptual acrobatics' (Simpson, 2013); and after the theory is fitted to its procrustean bed, its usefulness as a general explanatory tool seems diminished. Thus, the field has witnessed a number of one-shot efforts in which traditional theories are applied once to a particular case or sample but the results are never replicated or extended by others in any sustained way. One exception to this general pattern involves the 'low self-control' theory, which has been the subject of a notable number of empirical and theoretical investigations, most of which have been highly critical of the theory. Since the scholarship around the low self-control theory has been largely confrontational (as opposed to confirmatory), it is discussed in the section on current debates.

However, unlike the largely desultory approach to white-collar crime prevalent in traditional criminological circles, the opportunity or choice perspectives have gained traction and appear to be generating sustained interest among white-collar crime scholars (Shover and Hochstetler, 2006; Benson and Simpson, 2009). The double-barreled nomenclature used here – 'opportunity or choice' – reflects the differing points of emphasis found in this perspective. As advanced by Benson and Simpson (2009), the 'opportunity' perspective draws attention to situational conditions that either facilitate or impede particular types of white-collar offending. This perspective rests on the assumption that the presence of criminal opportunity is an essential causal element in all offending. A criminal opportunity arises whenever a suitably motivated person or group of persons recognizes or perceives that a suitable target is not adequately guarded. In regard to white-collar crime, the term 'suitable target' is not meant to be restricted to a person who can be attacked or a physical object that can be stolen or damaged. Rather, targets are conceived more broadly and flexibly to include situations in which an individual or group of individuals can pervert or distort

legitimate economic business or professional activities so as to illegally benefit themselves. For example, the buying and selling of securities is a legitimate economic activity that can be perverted when a corporate insider trades shares of his or her own company on the basis of nonpublic material information. Here, the activity of trading securities is the target. If that activity is not well regulated and supervised (i.e., guarded) by governmental entities, then it becomes a suitable target by means of which a potential offender can take advantage of others to enrich himself or herself.

In regard to white-collar crime, criminal opportunities arise or become available to potential offenders as a result of their holding certain occupational positions. The occupational position gives the offender 'legitimate access' to some situation or opportunity in which the offender can advance his or her own interests at the expense of other individuals, organizations, or the market in general (Benson and Simpson, 2009; Felson and Boba, 2010). For example, by virtue of their positions in the health care system, physicians have access to various public and private health care insurance systems. Because of this legitimate access, physicians can submit fraudulent claims for payment for health care services rendered (Sparrow, 1996). As another example, by virtue of their positions, the owners or highly placed executives in certain types of companies can conspire together to set prices or rig bids in violation of antitrust laws (Geis, 1977).

The degree to which occupational positions create or make available opportunities for white-collar crime depends on the laws, regulations, and rules governing occupations, industries, and types of economic activities. As laws or regulations evolve or change, opportunities for white-collar crime expand or contract. White-collar crime scholars have investigated numerous examples of the interplay between changing regulatory regimes (and especially deregulation) and the expansion or maintenance of criminal activities in various industries or services, including the thrift industry (Calavita et al., 1997), the health care system (Sparrow, 1996), the accounting profession (Coffee, 2002), and recently the mortgage industry (Nguyen and Pontell, 2010), to name a few. Indeed, recent and historical research on white-collar crime suggests a general conclusion presaged by Sutherland (1949) – that some form of white-collar crime can be found in all industries and occupations and the degree to which it is present depends, to a great extent, on the results of the political process that produces the structure and level of enforcement of regulatory regimes and criminal laws (Shover and Hochstetler, 2006).

Criminal opportunities only result in actual crimes, however, when someone decides to take advantage of them. Thus, the second term in the 'opportunity or choice' perspective draws attention to the reality that opportunities are not self-activating. Rather, they must be 'acted upon' by actors who make choices. Contemporary theory conceptualizes these choices as being made by cognitively astute rational actors, who weigh the costs and benefits of both legal and illegal lines of action before deciding to engage in white-collar crime (Shover and Hochstetler, 2006; Paternoster and Simpson, 1993; Piquero et al., 2005). Shover and Hochstetler (2006) argue that tempted or criminally predisposed individuals are attracted to what they call situations of 'lure.' Lure is similar to what routine activity theorists call suitable targets or criminal opportunities, but it is a broader, more economically oriented concept because it is not limited to static persons or things that an offender can attack or steal. Rather, it includes situations that provide actors with opportunities to advance their own interests – that is, to get ahead in some way, whether it is in a personal financial perspective or in the sense of improving the bottom line of a business, at little or some acceptable level of cost. For example, illegal undocumented immigrants are a source of lure, because they could be exploited by unscrupulous employers as a cheap and compliant source of labor who can be made to work under conditions and rates of pay that would not be tolerated by naturalized or native-born workers (Shover and Hochstetler, 2006). Contemporary research focuses on how the evolution and enforcement of criminal laws and administrative regulations, as well as the structure and standard practices of industries and markets, create criminal opportunities, and on how those opportunities affect or pattern the choices made by actors (Shover and Routhe, 2004).

Psychology and Individual Differences

Over the past two decades, criminology has taken a decisive turn toward individual differences, especially in regard to psychological and cognitive characteristics, as important explanatory factors in the etiology of criminal behavior. The importance of individual differences has been made apparent by the increasingly relevant research of biosocial criminologists, as well as the theories and findings of life-course and developmental criminologists. Yet, as Levi (2010b) notes, until quite recently, the practitioners of the biosocial perspective on crime showed little interest in white-collar crime or white-collar offenders; and white-collar crime scholars have returned the favor by showing little interest in life-course, developmental, or biosocial theory (Piquero and Benson, 2004). Indeed, white-collar offenders have long been assumed to be basically normal people who do not suffer from the psychological or personal pathologies that seem so common among street offenders (Coleman, 1989; Spencer, 1965; Sutherland, 1949). For many years, the only psychological or personality characteristic of white-collar offenders that provoked much interest was their seemingly total resistance to any attempt to define their activities as crimes (Benson, 1985; Maruna and Copes, 2005). Other than their tendency to rationalize and excuse their crimes, the psychological makeup of white-collar offenders was not thought to be an important dimension of white-collar crime.

The assumptions that white-collar offenders are psychologically normal and that individual differences have little or nothing to do with white-collar crime are now being called into question. A small but growing body of theory and research has begun to suggest that certain psychological characteristics may both facilitate white-collar crime and be more common among white-collar offenders than the general population. Oddly enough, this surge of interest in individual differences and psychology may have been sparked, at least in part, by the contention that there is nothing special about the etiology of white-collar crime and that for theoretical purposes, the whole concept of white-collar crime should be abandoned (Hirschi and Gottfredson, 1987). The developers of the general

theory of crime argued that white-collar crime is caused by the same factor that causes all other forms of crime – low self-control. Although most white-collar crime scholars vigorously dispute this claim (Benson and Moore, 1992; Yeager and Reed, 1998), it did lead some researchers to consider whether other psychological characteristics might nonetheless play a role in white-collar offending. For example, the psychological characteristic of 'desire for control' has been implicated in corporate offending (Piquero et al., 2005). The desire for control refers to a general wish or need to be in control of everyday events in life. Piquero et al. (2005) suggest that the strength of an individual's desire for control, influences the interpretation of rational choice considerations in regard to crime; and that the desire for control may lead to white-collar offending by businesspeople in an effort to reduce uncertainty in the business environment.

There are theoretical reasons for thinking that certain types of people may be more amenable to white-collar crimes than others. For example, theoretically it can be argued that since almost all white-collar crimes depend in one way or another on deception (Benson and Simpson, 2009), people who are capable of and are not troubled by the deceiving of others, will be better suited to white-collar crimes than people who have a difficult time fooling others or who feel the pangs of conscience when they lie and deceive. Although there are to date no studies that have explicitly examined deception as a trait related to white-collar crime, studies have found differences between white-collar offenders and others on traits that conceivably may be related to deception. An early study by Collins and Schmidt (1993) compared a sample of persons who had been convicted of white-collar crimes to a matched control sample of noncriminal white-collar employees on several personality scales. Offenders differed from nonoffenders on two dimensions: social conscientiousness and social extraversion. Conscientious individuals are dependable, responsible, rule-abiding, honest, motivated to high job performance, and committed to civic and social values. White-collar offenders seem to be less likely to display these traits than their law-abiding counterparts. Social extraversion refers to one's popularity and ability to navigate well in social situations. Collins and Schmidt (1993) speculate that people with outgoing personalities may compete well for higher level jobs where they are exposed to white-collar criminal opportunities.

More recently, efforts have been made to apply standard personality assessment tools to white-collar offenders (Walters and Geyer, 2004; Ragatz et al., 2012). One of the best-known assessments, the Psychological Inventory of Criminal Thinking Styles, has been used to compare the criminal thinking styles of white-collar offenders with that of other types of offenders. Before discussing the results of these investigations, however, it is important to note that this research suggests that care is needed in identifying whom to count as a white-collar offender. Indeed, there appear to be two types of white-collar offenders: (1) individuals who commit only white-collar crimes, called 'white-collar only offenders' and (2) individuals who commit both white-collar and non-white-collar crimes, called 'white-collar versatile offenders.' White-collar versatile offenders appear to be distinctly different from their counterparts who commit only white-collar type offenses, in regard to their lifestyles and criminal thinking patterns (Walters and Geyer, 2004; Ragatz et al., 2012). Ragatz and colleagues found that white-collar offenders tend to score lower than non-white-collar offenders on criminal lifestyle measures, but they score higher on measures of psychopathology and psychopathic traits. Indeed, a number of researchers have begun to focus on so-called successful psychopaths (Gao and Raine, 2010; Babiak et al., 2010; Babiak and Hare, 2006). Unlike most psychopaths who suffer from cognitive deficits, successful psychopaths manifest both psychopathic traits (glibness, manipulativeness, callousness, and lack of emotion) and enhanced neurobiological functioning and cognitive skills. In theory, these individuals have the cognitive and social skills necessary to use covert and deceitful methods to pursue their goals in ways that are not obvious to others and that sometimes involve undetected white-collar crimes. In addition, successful psychopaths are not burdened by moral qualms about obeying the law or by normative expectations not to hurt others unnecessarily. They are clever and unconstrained by ethical scruples. While this line of research is still in its infancy, it nevertheless adds an important new dimension to the choice perspective by providing a theoretical means to explain why similarly situated persons make different choices in regard to white-collar crime opportunities.

Neutralizations and Accounts

Even though it was originally developed to account for juvenile delinquency, neutralization theory has for many years had a receptive audience among those who study organizational and white-collar offending (Maruna and Copes, 2005). Investigations and explications of particular types of white-collar crime often devote considerable attention to detailing the linguistic mechanisms used by the perpetrators to justify, excuse, or account for their particular offenses (Benson, 1985; Conklin, 1977). Recently, however, researchers have begun to explore new questions about neutralizations that extend beyond merely identifying and describing their use by white-collar offenders. For example, researchers have begun to explore how the construction and use of neutralizations and accounts are influenced by gender and social class (Klenowski et al., 2011; Vieraitis et al., 2012; Willott et al., 2001). Stadler and Benson (2012) compared the use of neutralizations by white-collar offenders to those of other types of offenders and found surprisingly few differences. The significance and generalizability of the findings of this new body of research for white-collar crime, however, is questionable for methodological reasons. Some of the studies are based on relatively small samples of offenders who have been convicted of relatively pedestrian forms of white-collar crime, such as minor kinds of fraud, low-level embezzlements, workplace related offenses, and identity theft. Others are based on scenario studies of MBA students. Absent from these investigations are the professionals and corporate executives responsible for the most serious white-collar offenses, such as health care fraud, accounting fraud, antitrust violations, and securities violations. There remain only a very few studies of neutralization practices among upper class elite offenders.

Debates

Prevention and Control

Since many of the most serious forms of white-collar crimes are committed by or within business organizations, a significant amount of scholarly attention has focused on the most effective way to control these organizations. The regulation of business activities has a long and controversial history in the United States and in many other developed nations. In the United States, over the past few decades there has been a shift away from what is known as 'command and control' regulation to 'responsive regulation' (Shover and Routhe, 2005; Simpson, 2013). Command and control regulation takes a legalistic approach to the problem of controlling business organizations. In this style of regulation, regulatory agencies promulgate rules and then mechanically enforce them by imposing sanctions for noncompliance. Regulators see regulated entities as being driven solely by the amoral pursuit profit and as adversaries who will comply with regulatory directives only under the threat of state-imposed sanctions. While there is some evidence that the threat of sanctions works, the command and control approach also appears to foster the development of cultures of resistance and a minimalistic approach to compliance among regulated entities. Over time, the antagonistic relationship between regulators and the subjects of regulation raises the costs of oversight and reduces its effectiveness.

The command and control approach to regulation, with its emphasis on punishing violations in order to achieve greater compliance through deterrence, began to fall out of favor in the 1980s. Taking its place was a more flexible approach to enforcement called responsive regulation (Ayres and Braithwaite, 1992). This approach to regulation assumes that most business firms want to comply with regulations and fail to do so only because they are either not well-informed about their regulatory obligations or are simply unable to comply because of lack of technical or financial resources. In other words, firms are considered to be trustworthy and willing to self-regulate if treated with fairness and understanding by regulators. Noncompliance is seen as evidence of a problem that needs to be fixed rather than as a violation that needs to be punished. The role of the regulator is to educate and assist firms in their efforts to comply, and as long as firms are seen as making efforts in good faith toward compliance, sanctioning is unnecessary. Sanctions may eventually be imposed on firms who are obstinate in their noncompliance and who refuse to cooperate with regulators , but under responsive regulation, control through punishment is viewed as a last resort rather than the first option.

As with the command and control approach, the responsive regulation approach also has its success stories but the relative effectiveness of the two approaches is still a matter of debate. Critics of the flexible approach argue that it gives corporations too much influence over regulatory regimes and in the end fails to adequately protect the public interest. On the other hand, deterrence-based regulation is criticized for alienating the business community and for stifling productivity and innovation, which also harms the interests of the general public.

Gender

The study of women and white-collar crime represents both an emerging trend and an old debate shifted to a new domain. In the 1970s, the subject of women and crime became a hotly debated issue. Some scholars saw signs that the 'gender gap' in crime (that is, the difference in the prevalence of men and women in involvement in crime), which has always been large, was in the process of diminishing as a result of the Women's Movement and the greater social and economic equality that it was bestowing on women (Simon, 1975; Adler, 1975). As women attained social equality with men, it was thought that they would start to behave more like men, including getting more involved in crime. Others said that except for minor property offenses, such as shoplifting, the gender gap was not closing at all; and also argued that women's offenses were more often committed out of desperation rather than for enhanced social power.

A similar sort of debate is now going on in regard to women and white-collar crime. As with almost all other forms of crime (excluding only prostitution), women are less involved in white-collar crime than men. However, as more and more women enter and assume leadership positions in the corporate world, their level of involvement may increase. Dodge (2009) argues that in order to rise in organizational hierarchies women have had to become more like men in regard to aggressiveness and risk-taking. In her view, women's involvement in white-collar crime is and will continue to be on the rise as women's occupational opportunities increase. Other scholars are not so sure. The opposing view holds that even though women have made great strides in the workplace, they are nevertheless still marginalized at the higher levels of the corporate world. Even when women occupy positions of power, they are still usually excluded from the social networks that orchestrate and carry out large-scale corporate crime schemes. Some research suggests that, as with street crimes, women play subsidiary and minor roles in most corporate crimes. In addition, some argue that for both social and biological reasons, women are guided by different focal concerns than men and as a result, they are less likely than men to engage in white-collar crime even when they have the opportunity (Steffensmeier et al., 2013). When women do commit white-collar crimes, research suggests they usually do so for different reasons than men and they interpret their crimes differently from men (Daly, 1989; Klenowski et al., 2011).

Applicability of Low Self-Control Theory

Low self-control theory was developed by Gottfredson and Hirschi (1990). Although it has been applied primarily in regard to ordinary street crimes and offenders, Hirschi and Gottfredson (1987) controversially claimed that it also applied to white-collar crime and to white-collar offenders. In brief, the theory posits that at the individual level, the fundamental cause of all criminal behavior is low self-control. Low self-control refers to a behavioral propensity to pursue one's individual desires and pleasures with little consideration for long-term consequences or for the effects of one's behavior on others. People with low self-control are impulsive and insensitive, and when they encounter a criminal opportunity that

can satisfy their desires, they are likely to take advantage of it without concern for the potential long-term consequences of their actions. Hirschi and Gottfredson (1987) made a couple of controversial claims regarding white-collar crime, including (1) that is relatively rare and (2) that the demographic distribution (age, sex, and race) of these crimes is the same as it is for other forms of crime, which would imply that they all result from similar underlying causal processes.

With only a few exceptions (Herbert et al., 1998), the response to Gottfredson and Hirschi has been largely negative (Steffensmeier, 1989; Benson and Moore, 1992; Yeager and Reed, 1998). Critics have argued that Gottfredson and Hirschi based their assertions on flawed analysis of faulty data. Further, they have argued that white-collar crime is not rare but widespread and that its demographic distribution looks nothing like that of ordinary street crimes (Benson and Kerley, 2000; Wheeler et al., 1988). Finally, the organizational location of much white-collar crime renders the whole focus on individual level characteristics misguided (Yeager and Reed, 1998; Simpson and Piquero, 2002). Organizational or corporate crimes often involve groups of people who play different roles and who may participate in the offense for different periods of time. Indeed, the people involved in organizational criminal conspiracies may change over time, such that while any particular offense may be continuous over time, the offenders involved in it are not. In these cases, it seems difficult to attribute individual involvement in corporate crime, which varies over time, to stable underlying traits which supposedly remain constant over time. Rather, for organizationally based crimes, the causes of involvement seem more likely to arise out of externally generated pressures for performance rather than out of individual levels of self-control.

Conclusions

The study of white-collar crime has long been marginalized in criminology and it is likely to remain so in the future (Simpson, 2013). Nevertheless, recent developments suggest there is some potential for it to move closer to the mainstream. These developments pertain most directly to shifts in theoretical focus. As white-collar crime scholars have adopted the 'opportunity and choice' perspective to organize their research, their investigations have become more relevant to and congruent with current trends in mainstream criminology. Opportunity theory as applied to white-collar crime fits well with the approaches taken by situational crime prevention specialists, in that it draws attention to opportunity structures as important causes of criminal events. In regard to choices, the greater focus on individual differences and on life-course trajectories in white-collar crime also resonate well with the concerns of more traditional criminologists and with the rising biosocial perspective. Indeed, one could argue that white-collar crime scholars are playing an important role in expanding and deepening the study of opportunities, choices, and individual differences. For example, within traditional criminology, criminal opportunities are conceived in terms of the availability of physical targets and lack of guardianship, but white-collar crime investigations suggest that conceptualizations of opportunity need to be broadened to recognize that they can also arise out of legitimate economic and occupational activities. Another example of the potentially salutary outcomes of recent white-collar crime scholarship involves research on individual differences, especially those related to psychological and cognitive characteristics. Within traditional criminology, the temperamental characteristics of offenders are usually conceived as deficits or negative attributes. Offending is seen to arise out of the offender's cognitive or neuropsychological deficits, which make it difficult for him or her to succeed in conventional institutions such as school, work, and marriage. Recent research on white-collar offenders, however, raises the intriguing possibility that some offenders may not suffer from neuropsychological deficits. Rather, they may have enhanced cognitive functioning that actually facilitates white-collar offending. If this line of research continues to bear fruit, it suggests that the connection between temperament and crime will need to be reconceived.

See also: Control Theories and Crime; Crime and Class; Crime and Gender; Crime, Sociology of; Crime: Whitecollar; Delinquency, Sociology of; Economics and Crime; Life Course Criminology; Routine Activities, Criminal Opportunities, Crime and Crime Prevention; Self-Control Theory and Crime; Social Learning Theory and Crime.

Bibliography

Adler, Freda, 1975. Sisters in Crime: The Rise of the New Female Criminal. McGraw-Hill, New York.

Ayres, Ian, Braithwaite, John, 1992. Responsive Regulation: Transcending the Deregulation Debate. Oxford Socio-Legal Studies. Oxford University Press, New York.

Babiak, Paul, Hare, Robert D., 2006. Snakes in Suits: When Psychopaths Go to Work, first ed. Regan Books, New York.

Babiak, Paul, Neumann, Craig S., Hare, Robert D., 2010. Corporate psychopathy: talking the walk. Behavioral Sciences and the Law 28, 174–193.

Baumer, Eric P., Wolff, Kevin T., Arnio, Ashley N., Chiapputo, Joseph, 2012. Assessing the Link between Foreclosure and Crime Rates: A Multilevel Analysis across Neighborhoods in Large U.S. Cities. Technical Report. National Institute of Justice.

Benson, Michael L., 1985. Denying the guilty mind: accounting for involvement in a white-collar crime. Criminology 23, 583–608.

Benson, Michael L., Moore, Elizabeth, 1992. Are white-collar and common offenders the same: an empirical and theoretical critique of a recently proposed general theory of crime. Journal of Research in Crime and Delinquency 29, 251–272.

Benson, Michael L., Kerley, Kent R., 2000. Life course theory and white-collar crime. In: Pontell, Henry N., Shichor, David (Eds.), Contemporary Issues in Crime and Criminal Justice: Essays in Honor of Gilbert Geis. Prentice Hall, Saddle River, NJ, pp. 121–136.

Benson, Michael L., Simpson, Sally S., 2009. White-Collar Crime: An Opportunity Perspective. Routledge, New York.

Benson, Michael L., 2010. Evolutionary ecology, fraud and the global financial crisis. In: Rosenfeld, Richard, Quinet, Kenna, Garcia, Crystal (Eds.), Contemporary Issues in Criminological Theory and Research: The Role of Social Institutions. Wadsworth, Belmont, CA, pp. 299–306.

Calavita, K., Tillman, R., Pontell, H.N., 1997. The savings and loan debacle, financial crime, and the state. Annual Review of Sociology 23, 19–38.

Carswell, Andrew T., Bachtel, Douglas C., 2009. Mortgage fraud: a risk factor analysis of affected communities. Crime, Law and Social Change 52, 347–364.

Coffee Jr., John C., August 2002. Understanding enron: it's about the gatekeepers, stupid. The Business Lawyer 57, 1–2.

Coleman, James W., 1989. The Criminal Elite. St. Martin's Press, New York.

Collins, Judith M., Schmidt, Frank L., 1993. Personality, integrity, and white-collar crime: a construct validity study. Personnel Psychology 46, 295–311.
Conklin, John E., 1977. Illegal but Not Criminal: Business Crime in America. Prentice-Hall, Englewood Cliffs, CA.
Daly, Kathleen, 1989. Gender and varieties of white-collar crime. Criminology 27, 769–794.
Dodge, Mary, 2009. Women and White-Collar Crime. Women in Criminal Justice Series. Prentice Hall, Upper Saddle River, NJ.
Edelhertz, Herbert, 1970. The Nature, Impact and Prosecution of White-Collar Crime. U.S. Department of Justice, Washington, DC.
Felson, Marcus, Boba, Rachel, 2010. Crime and Everyday Life, fourth ed. Sage, Los Angeles.
Financial Crimes Enforcement Network, 2006. Mortgage Loan Fraud: An Industry Assessment Based on Suspicious Activity Report Analysis (cited June 18, 2010). Available from: http://www.fincen.gov/news_room/rp/reports/pdf/MortgageLoanFraud.pdf.
Gao, Yu, Raine, Adrian, 2010. Successful and unsuccessful psychopaths: a neurobiological model. Behavioral Sciences and the Law 28, 194–210.
Geis, Gilbert, 1977. The heavy electrical equipment antitrust cases of 1961. In: Geis, Gilbert, Meier, Robert F. (Eds.), White-Collar Crime: Offenses in Business, Politics, and the Professions, revised ed. The Free Press, New York, pp. 117–132.
Gottfredson, Michael R., Hirschi, Travis, 1990. A General Theory of Crime. Stanford University Press, Stanford, CA.
Herbert, Carey, Green, Gary S., Larragoite, Victor, 1998. Clarifying the reach of a general theory of crime for organizational offending: a comment on Reed and Yeager. Criminology 36, 867–884.
Hirschi, Travis, Gottfredson, Michael R., 1987. Causes of white-collar crime. Criminology 25, 949–974.
Hirschi, Travis, Gottfredson, Michael R., 1989. The significance of white-collar crime for a general theory of crime. Criminology 27, 359–371.
Klenowski, Paul M., Copes, Heith, Mullins, Christopher W., 2011. Gender, identity, and accounts: how white collar offenders do gender when making sense of their crimes. Justice Quarterly 28, 46–69.
Levi, Michael, 2010a. Fraud vulnerabilities, the financial crisis, and the business cycle. In: Rosenfeld, Richard, Quinet, Kenna, Garcia, Crystal (Eds.), Contemporary Issues in Criminological Theory and Research: The Role of Social Institutions. Wadsworth, Belmont, CA, pp. 269–292.
Levi, Michael, 2010b. Individual differences and white-collar crime. In: Cullen, Francis T., Wilcox, Pamela (Eds.), Encyclopedia of Criminological Theory. Sage, Thousand Oaks, CA, pp. 469–473.
Maruna, Shadd, Copes, Heith, 2005. What have we learned from five decades of neutralization research? Crime and Justice 32, 221–320.
Nguyen, Tomson, Pontell, Henry N., 2010. Mortgage origination fraud and the global economic crisis: a criminological analysis. Criminology and Public Policy 9, 591–612.
Partnoy, Frank, 2003. Infectious Greed: How Deceit and Risk Corrupted the Financial Markets. Times Books, New York.
Paternoster, Raymond, Simpson, Sally S., 1993. A rational choice theory of corporate crime. In: Clarke, Ronald V., Felson, Marcus (Eds.), Routine Activity and Rational Choice. Transaction, New Brunswick, NJ, pp. 37–58.
Piquero, Nicole Leeper, Benson, Michael L., 2004. White-collar crime and criminal careers: specifying a trajectory of punctuated situational offending. Journal of Contemporary Criminal Justice 20, 148–165.
Piquero, Nicole Leeper, Lyn Exum, M., Simpson, Sally S., 2005. Integrating the desire-for-control and rational choice in a corporate crime context. Justice Quarterly 22, 252–280.
Ragatz, Laurie L., Fremouw, William, Baker, Edward, 2012. The psychological profile of white-collar offenders. Criminal Justice and Behavior 39, 978–997.
Reiss, Albert J., Biderman, Albert D., 1981. Data Sources on White-Collar Law-breaking. U.S. Dept. of Justice, National Institute of Justice, Washington, DC.
Shover, Neal, Grabosky, Peter, 2010. White-collar crime and the great recession. Criminology and Public Policy 9, 429–433.
Shover, Neal, Hochstetler, Andrew, 2006. Choosing White-Collar Crime. Cambridge University Press, New York.
Shover, Neal, Routhe, Aaron S., 2004. Environmental crime. In: Tonry, Michael (Ed.), Crime and Justice: A Review of Research, vol. 32. University of Chicago Press, Chicago, pp. 321–371.
Simon, Rita James, 1975. Women and Crime. Lexington Books, Lexington, MA.
Simpson, Sally S., 2013. White-collar crime: a review of recent developments and promising directions for future research. Annual Review of Sociology 39, 301–331.
Simpson, Sally S., Piquero, Nicole Leeper, 2002. Low self-control, organizational theory, and corporate crime. Law and Society Review 36, 509–548.
Sparrow, Malcolm K., 1996. License to Steal: Why Fraud Plagues America's Health Care System. Westview Press, Boulder, CO.
Spencer, John C., 1965. White-collar crime. In: Grygier, Tadeusz, Jones, Howard, Spencer, John C. (Eds.), Criminology in Transition. Tavistock Publishers, London, pp. 233–266.
Stadler, William A., Benson, Michael L., 2012. Revisiting the guilty mind: neutralization of white-collar crime. Criminal Justice Review 37 (4), 495–512.
Steffensmeier, Darrell J., 1989. On the causes of 'white-collar' crime: an assessment of Hirschi and Gottfredson's claims. Criminology 27, 345–358.
Steffensmeier, Darrell F., Schwartz, Jennifer, Roche, Michael, 2013. Gender and 21st century corporate crime: female involvement and gender gap in enron-era corporate frauds. American Sociological Review 78, 448–476.
Sutherland, Edwin H., 1949. White-Collar Crime. Holt, Rinehart and Winston, New York.
Vieraitis, Lynne M., Piquero, Nicole Leeper, Piquero, Alex R., Tibbetts, Stephen G., Blankenship, Michael, 2012. Do men and women differ in their neutralizations of corporate crime? Criminal Justice Review 37, 478–493.
Walters, Glenn D., Geyer, Matthew D., 2004. Criminal thinking and identity in male white-collar offenders. Criminal Justice and Behavior 31, 263–281.
Wheeler, Stanton, Weisburd, David, Waring, Elin, Bode, Nancy, 1988. White-collar crime and criminals. American Criminal Law Review 25, 331–357.
Willott, Sara, Griffin, Christine, Torrance, Mark, 2001. Snakes and ladders: upper-middle class male offenders talk about economic crime. Criminology 39, 441–466.
Yeager, Peter Cleary, Reed, Gary E., 1998. Of corporate persons and straw men: a reply to Herbert, Green and Larragoite. Criminology 36, 885–897.

White, Gilbert (1911–2006)

James L Wescoat Jr., Massachusetts Institute of Technology, Cambridge, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Gilbert F. White (1911–2006) is variously known as the "father of natural hazards research," "father of floodplain management," and all-around leader of the environment-society branch of geography in the United State during the twentieth century (Wescoat, 2006). These father-figure references convey the regard in which White has been held for his mentoring of two generations of colleagues and students as well as for his pioneering research on water resources, natural hazards, and environmental management. White's seven decades of creative work, beginning with policy analyses in the Franklin D. Roosevelt administration in the 1930s and concluding with reflective essays on environmental stewardship in the early years of the twenty-first century has produced a rich historiography of appraisals. The first section of this article reviews the historiography of previous assessments of White's research, from an early festschrift (Kates and Burton, 1986) to theoretical debates (Wescoat, 1992), an autobiographical essay (White, 2002), and memorial reviews following his death (Kates, 2011). The second section of the article rereads White's research contributions in different periods of his career, underscoring the relationships among research subjects, ethical concerns, and contributions to the field of environmental geography.

Historiography of White's Contributions to Water and Hazards Research

Commentators have reviewed the major themes in White's work in several different ways. His most widely perceived contributions were in the fields of water resources and natural hazards research. However, the first broad assessment that came in the form of a festschrift presented White's research contributions under 12 broad headings (Kates and Burton, 1986: p. xi):

- Public water supplies
- Floodplain management
- Natural resources
- International resources and environment
- Arid lands
- River basins
- Decision making
- Geographic education
- Weather and climate modification
- Environmental perception
- Environmental hazards
- Energy

This list appears diffuse, but the festschrift editors boiled down their appraisal to three main scholarly contributions to water, hazards, and environmental research: (1) the concept of human adjustment; (2) the theoretical and practical range of choice; and (3) the role of perception in decision-making (Kates and Burton, 1986: p. xii). They showed that these three ideas were interrelated with one another in a research framework that focused on decision-makers' perceptions about the range of choice among alternative adjustments to resource scarcity and hazards. This research framework helped bring about a shift from thinking of flood losses as 'Acts of God' to regarding them as consequences of individual and collective action.

As elaborated below, this early appraisal followed shortly after intense criticisms of White's natural hazards research, articulated in part as book reviews of *The Environment as Hazard* (Burton et al., 1978), which characterized his approach as behaviorist, scientistic, atheoretical, politically silent about the roles of power relations and injustice, and insufficiently critical of technocratic ideologies and methods of water engineering (e.g., Hewitt, 1983). Ironically, White's earliest research had directly challenged those structural engineering approaches to flood control on empirical, economic, and policy grounds (e.g., White, 1942/1945); but it had not addressed explicitly the underlying social and political structures and processes that drove them, such as race, gender, poverty, and class. A significant yet less vocal scholarly criticism questioned White's emphasis on applied policy-relevant geography compared with basic research (White, 1972).

A philosophical appraisal addressed these challenges by examining White's contributions in relation to four main themes in the pragmatism of John Dewey, citing direct and indirect influences (Wescoat, 1992):

Dewey	*White*
Precariousness of existence	Hazards perception
Logic of inquiry	Expanding the range of choice
Learning from experience	Behavioral geography
Discourse and democracy	Public policy

White had mixed reactions to this theoretical interpretation, questioning its utility for people and places (pers. comm., 1993 onward). He grew more positive about theoretical debates over time as graduate students and others worked more consciously with or in opposition to pragmatist approaches (e.g., White, 1997). White continued to challenge scholars about what difference their work can make for environmental and human well-being, which is a core query in pragmatism. White and colleagues also gave increasingly explicit attention to societal

driving forces in human vulnerability to natural hazards, and that research has become increasingly mainstreamed in the field (Cutter, 2006; White et al., 2001; Wisner, 2003).

The most deeply personal study of White's work and life came in a biography by Robert Hinshaw (2006) that was completed in the year White died. Written by a scholar, Quaker, and multidecadal friend, Robert Hinshaw's *Living with Nature's Extremes: The Life of Gilbert Fowler White* strives to present the inner deliberations as well as outward manifestations of White's research, gleaned from numerous interviews and personal diaries. It is perhaps no accident given the complexity of this task that the volume has *three* concluding chapters: 'Gilbert White: the Man', 'Gilbert White: The Legacy', and 'Epilogue'. Personhood is a vital dimension of all scholarly lives, but it has been particularly underscored in White's case. Hinshaw's portrait focuses on moral convictions, social values (humility, service), and personal relationships (respect, kindness, generosity, loyalty, veracity). White's main substantive legacy was presented as the stewardship of floodplains, along with effective ways of communicating and institution-building. His paradoxical qualities were framed as optimism and faith in the face of evil.

White (2002) himself identified six periods of his career and four research themes that carried through most of them.

1. Early influences – university, ranching, and Quaker service
2. Periods of development
   a. The New Deal
   b. Vichy's France and Germany
   c. American Friends Service Committee
   d. Haverford College
   e. University of Chicago
   f. University of Colorado

He described his four main scientific and policy interests as: (1) floods and other natural hazards; (2) integrated water management; (3) environmental effects of water management; (4) postaudits; and (5) professional geographers. The key ideas in these works were human adjustment to natural hazards, alternatives in water management, and environmental effects of natural resources management. He pursued these studies, "in a spirit of trying to be helpful to needy fellow humans by cooperative nonviolent methods" (White, 2002: p. 364), and noted that his greatest personal satisfaction lay in observing the contributions of former students. His final comment was that "a basic judgment of geography is what it contributes cooperatively in fashioning a sustainable earth" (White, 2002: p. 364).

After White died in 2006, a series of memorial essays again weighed his contributions, increasingly in terms of ethical and religious values as well as scientific content (Burton, 2008; Kates and Burton, 2008; Murphy, 2008; Palm, 2008; Wescoat, 2006). Kates and Burton (2008) organized one of these essays under the headings of local studies, national comparisons, and global challenges, evoking an important work in White and Hass's (1975) career on natural hazards. The most recent memorial essay published by the National Academy of Sciences by White's long-time protégé and colleague Robert Kates (2011) linked five substantive research fields and associated normative aims:

1. Early emphasis on public water supplies
2. Natural hazard reduction
3. Making peace by managing water
4. Science in practical service to humanity
5. Human development and environment

Kates concluded his biographical memoir with a description of the glass artistic memorial by White's daughter Mary in the Boulder Creek floodplain.

Rereading Gilbert White's Contributions to Environmental Geography

This rereading of White's research takes Mary White's artistic memorial to her father as its starting point (http://www.colorado.edu/hazards/gfw/GFW_memorial.html). It stands as a simple individual upright glass column, sited along Boulder Creek near the headgate of an historic irrigation ditch. Its thin pillar of blue-green glass tiles tapers from about a foot in width at its pedestal to several inches in width at its peak. Heights of historic and projected floods are demarcated in the glass, and it has interior lighting that shines through the glass at night. It thus conveys the form as well as philosophy of a creative individual in a social environment. This is an apt perspective for reading White's work.

White's (2002) autobiographical essay conveyed a sense of unity of purpose in geographic inquiry, which he described as 'fashioning a sustainable earth'. His diverse contributions and pluralistic worldview could lead one to interpret his work as wide-ranging, even eclectic, but a stronger argument can be made for the unity of its contributions to what may be termed environmental geography. To say this is to go beyond the twentieth-century hyphenated framings of the man–land, environment–society, nature–society, and human–environment branches of geography, and to suggest that White worked toward a simpler approach that wove together natural resources, natural hazards, and environmental management.

Early Influences and Studies from Chicago to Wyoming

White cited three main formative experiences, first, as a student in John Dewey's Laboratory School and subsequently in the Department of Geography at the University of Chicago; second, in summers working on a family ranch in Wyoming; and third, in his college introduction to Quakerism. All three emphasized the unity of experience within an expanding appreciation of the vast diversity of the world. Dewey (1929, 1934/1958, 1938) likewise focused on environmental experience as the basis for breaking down dualisms between mind and body, nature and culture, and other formalisms in philosophy and education. He regarded organism–environment relationships not so much as a dialectic as a complex process of experimental inquiry striving toward ends-in-view. He wrote that organisms, including human beings, exist not just within but rather as part of and because of their environment (Dewey, 1934/1958).

At that time, during the 1930s, the University of Chicago was one of the early leading geography departments in the US Although Dewey had long since moved to Columbia University, some of his ideas persisted both in education and

in ecological modes of inquiry across the natural and social sciences. White's advisor, Harlan Barrows (1923), published an influential essay on 'Geography as Human Ecology', at a time and place where ecological ideas and methods were widespread in the University. It was a time of interdisciplinary ecological research among botanists, urban sociologists, and landscape architects as well as geographers (Entrikin, 1980).

Ecological modes of social thought concerned White throughout his career. White was a bit late for their flowering in the early twentieth century, and a bit early for their revival in the early twenty-first century, though he would build upon the former and contribute to the latter. Although in some universities in the United States ecological approaches to social issues went into eclipse in mid-century, displaced by other analytical and hermeneutical methods, various strands of urban ecology, cultural ecology, periodically flourished.

White was exposed to Quakerism through a visit by Haverford philosophy professor Rufus Jones to the University of Chicago (pers. comm. *c.*2000). Jones was a mystical Quaker philosopher who became a particularly close personal mentor when White was president of Haverford College in the late 1940s. Jones' conception of experience and being was more akin to the pragmatism of William James' *Varieties of Religious Experience* than that of John Dewey (James, 1902). Indeed, he argued that, "We need to come back from thin behaviorist theories and scant pragmatist conceptions of mind to recover the mind's true amplitude and scope of creative activity" (Jones, 2001: p. 57). Jones' environmentalism linked soul, nature, and cosmos. For example, in writing laced with geographic metaphor, he wrote:

> We do not know how far our own margins of being reach. We cannot completely map the full area that belongs to us. No one can with certainty draw the boundary between himself and the beyond himself, any more than we can tell where the tidal river ends and the ocean begins, but we unmistakably feel on occasions that tides from beyond our own margins sweep into us and refresh us.
> (Jones, 2001: p. 14)

Jones' environmentalism was thus boundless. While White did not write in that mystical mode, he drank it in and sought to live it more in actions than words (White, 2002: p. 364).

Early Contributions to Natural Resources Planning and Policy in Washington, DC

White's first professional experience and writings were not in the academy but rather in government service during the Roosevelt administration. It was not an uncommon career path among geographers at the University of Chicago in the 1930s – some went to the Tennessee Valley Authority and others to various policy and planning centers. With Barrows' recommendation, White was appointed as a temporary assistant first to the Mississippi Valley Committee and then to full time service for New Deal advisory bodies such as the Natural Resources Committee and Natural Resources Planning Board. White's work in Washington concluded with 2 years of policy studies in the Bureau of the Budget.

- Geographer with the Mississippi Valley Committee; National Resources Committee; and National Resources Planning Board, 1934–40
- Staff Member, Bureau of the Budget, Executive Office of the President, 1940–42

In substantive terms, this involved studies of integrated approaches to land, water, flood hazards, and energy resources. White published several scientific papers during the 1930s primarily in water resources and planning journals. Most of his natural resources writing contributed to multidisciplinary committee reports, e.g., river basin surveys, which remain excellent examples of applied scientific writing. Other writings were memoranda for and responses to requests for information from President Roosevelt and others in his administration. These publications followed Harlan Barrows' discipline of succinct issues-driven writing supported by data and judicious weighing of alternative courses of action.

White was increasingly active in Quaker affairs in Washington, so that when the US entered World War II, he left the administration to do relief work in France, but not before finally completing his doctoral dissertation. White's dissertation on *Human Adjustment to Floods* (1942/1945) had an enduring impact on water resources research and has proven to be one of his most influential works (Mcdonald et al., 2012). It began with the paradoxical observation that the more the federal government spent on flood control works the greater the flood losses. It was not simply that flood control had no effect on loss reduction, but rather that it increased the likelihood of increased losses from major flood events. This has subsequently been referred to as the 'levee effect', which encourages floodplain development by increasing perception of flood protection, which is true for high-frequency low-magnitude flood events, but which poses increased physical and economic hazards in the event of levee failure during low-frequency high-magnitude events. With the dissertation completed, White did service as a conscientious objector in France, where he was captured and interred for 19 months in Baden–Baden, Germany, during which time he and others taught and took classes, including Russian, which proved important for his research on US-USSR scientific collaboration during the Cold War era.

Academic Administration and Resumption of Resource Policy Studies at Haverford College

Repatriated in a prisoner exchange in 1944, White returned to Philadelphia for 2 years of humanitarian work with the American Friends Service Committee on Bengal famine relief, voluntary service, peace activism, and other concerns published as pamphlets before and during his presidency at Haverford College, a position he held from 1946 to 1955.

During that period he made several important research contributions. The first was to natural resources studies of the new United Nations, including the first international conference on the subject at Lake Success, New York. International river basin management inspired in part by the Tennessee Valley Authority (TVA) model was one of the main themes at that conference. In an influential review article, White (1957) wrote that while "no two rivers are the same," it is useful to

make broad comparisons of development trends, such as the transition from single- to multipurpose storage reservoirs. He also led integrative studies of arid land management sponsored by UNESCO, at a time of major drought in the US (White, 1956).

Additionally, White resumed his work on US national water policy studies contributing to a presidential commission on *A Water Policy for the American People*, which continues to deserve attention for its recommendations at a time when national water policy has eroded (Christian-Smith and Gleick, 2012; President's Water Policy Commission, 1950). Finally, it was not until this time that be began to publish articles in geography journals, the first of which was, *Toward an Appraisal of World Resources* (White, 1949).

Less well known from this period are White's short essays on issues, trends, and creative approaches to undergraduate college education, including international studies. This combination of voluntary service, national and international natural resources policy, as well as college leadership was heady but difficult to sustain. Prompted by a challenge from anthropologist Margaret Mead that he would not be able to sustain scholarly research, White decided to take his first full time academic position at the University of Chicago's Department of Geography in 1956, at the age of 41, where he remained for 13 years.

Integrating Natural Resources and Natural Hazards Research from the University of Chicago

The 1960s marked the first formal academic research activities of White's career, and they included recommitment to geographic research. One of White's first publications at Chicago was an essay on the department's approach to graduate study in geography, which described early diagnostic appraisals of students and potential advisor–advisee relationships (White, 1958).

Funded collaborative research with doctoral students constituted a new mode of inquiry. In addition to its empirical contributions to knowledge about patterns and processes of floodplain occupancy in the United States, vis-à-vis flood control and flood damage research, these studies addressed sensitive issues of land use planning in water resources research (White et al., 1958, White, 1961a, 1964a). Another innovation involved methods for eliciting and comparing natural hazards perceptions across different social groups, ranging from floodplain occupants to water planners, which were compared with hydrologic risk estimates, as partial explanations for hazards exposure and behavior.

Generalizing from this work, White's theoretical contribution in this period was a sixfold model of natural resource decision-making (White, 1961b). The six key components were: (1) resource assessment; (2) expansion of the range of choice; (3) assessment of technologies; (4) assessment of economic efficiency; (5) spatial linkages; and (6) social guides. While resource assessment was a classical form of geographic inquiry, expanding the range of choice was one of White's most significant transdisciplinary contributions (Wescoat, 1987; Mitchell, 2008). The range-of-choice concept was advanced in part through a pair of National Academy of Sciences studies that White chaired on *Alternatives in Water Management*, and *Water and Choice in the Colorado River Basin: An Example of Alternatives in Water Management* (National Research Council, 1966, 1968).

This framework was subsequently applied to issues ranging from flood hazards and forest management in the US to domestic water supply in East Africa and comparative international river basin planning around the world. The East African study stands out for its enduring classification of water-related diseases by transmission paths, developed by coauthor David Bradley, for its combination of meticulous field research and compassionate reporting authored by Anne U. White, and for its subsequent replication by geographer John Thompson and colleagues in East Africa (Thompson et al., 2001).

In summary, this was a period of theorization and conceptual model development. It diverged from utilitarian and systems analysis models of the time and has subsequently been characterized as advancing a pragmatist logic of inquiry in geography (Wescoat, 1992).

Integrating Resources, Hazards, and Environmental Management at the University of Colorado

White had a lifelong passion for the semiarid western region of the United States, so that moving to the University of Colorado at Boulder was an attractive final career move, in part for its active Quaker community, which included economist Kenneth Boulding who helped recruit White to join the university. White had deep differences at that time with the University of Chicago administration's treatment of student Vietnam War protestors (Hinshaw, 2006: pp. 111–113; White, 1964b).

The University of Colorado presented a final major career challenge that would yield his most significant institutional contributions to the natural hazards field. White was recruited to direct the Institute of Behavioral Science which housed interdisciplinary research groups on environmental and social issues. There he codirected studies such as the first national *Assessment of Research on Natural Hazards* (White and Haas, 1975); edited a volume on *Natural Hazards: Local, National, Global* (1975); and coauthored the volume on *The Environment as Hazard* mentioned earlier in this article (Burton et al., 1978). The 1970s were thus a productive expansion from floodplain research to the broader interdisciplinary field of natural hazards research before White officially retired in 1978. The most influential and lasting institutional accomplishment of this period was establishment of the Natural Hazards Research Applications, and Information Center (or the Natural Hazards Center) as part of the Institute of Behavioral Science. In addition to its clearinghouse function, the Natural Hazards Center convenes an annual summer workshop of several hundred participants to forge long-term relationships between academic and practitioner communities. It established predisaster 'quick response grants' to support immediate research travel to disaster areas to better understand relief phenomena, and it gave special support for doctoral students and young scholars.

White's research would continue for another 25 years, during which time it shifted to a major emphasis on environmental effects of resource use, from the local scale of

Boulder County to early international studies of climate change with the Scientific Committee on Problems of Environment (SCOPE), environmental effects of nuclear war, nuclear waste disposal at the Yucca Mountain site in Nevada, a diagnostic study of the Aral Sea basin crisis, nonpoint source depletion, and other concerns. During the 1980s White served as editor for *Environment* magazine. He was deeply concerned that water and hazards research were not well integrated with environmental inquiry (Wescoat and White, 2003).

Although oversimplifying the continuities in White's career, it is reasonable to note that the 1960s stand out for research on natural resource management, the 1970s for natural hazards, the 1980s for environmental issues, and the 1990s for his essays on the interrelationships among these fields. As noted in his autobiographical essay, he described the role of geography as helping "fashion a sustainable earth," and concluded that "With that aim, the geographic discipline has solid performance, and expanding opportunities." (White, 2002: p. 364). In his final years, White referred most often to his research field as geography and as environmental management. It seems reasonable therefore to join these two terms together as environmental geography.

Legacy: Local, National, and Global

As Gilbert White's many honors and accomplishments are listed on a website dedicated to his work at the University of Colorado at Boulder's Natural Hazards Center, his many medals, honorary degrees, and lifetime achievement awards will not be listed here (see Hinshaw, 2006: pp. 310–313; http://www.colorado.edu/hazards/gfw/). Instead, it may briefly be mentioned that there are fellowships in his name at the Resources for the Future, the University of Colorado; a named chair at Arizona State University; and an annual memorial lecture at the National Academy of Sciences; to name a few. His local influence is marked by the floodplain memorial discussed above, along with a trail named after his wife Anne U. White for her work in Boulder. White mentioned personal satisfaction in his students who themselves have supervised cadres of doctoral students who have become in turn environmental professionals and faculty members primarily in departments of geography (e.g., a couple of third generation examples include Dow and Downing, 2011; Mustafa, 2013; Showalter and Lu, 2010).

In the years following White's death in 2006 through 2012, the Web of Science Citation Index indicates that his work has been cited some 324 times at an increasing rate from 24 to 59 citations per year. Over the longer term, White's most cited works (>100) have been: *The Environment as Hazards* (Burton et al., 1978 – 233 citations); *Drawers of Water* (White et al., 1972 – 157 citations); *Assessment of Research in Natural Hazards* (White and Haas, 1975 – 149 citations); and *Human Adjustment to Floods* (White, 1942/1945 – 140 citations). It is interesting to note that with the exception of his doctoral dissertation, all of these studies were collaborative works. The continuing influence of White's research on water, hazards, and environment, including a dissertation published more than a half-century ago, is unique among twentieth century environmental geographers in the United States and perhaps globally.

See also: Nature and Society in Geography.

Bibliography

Barrows, H.H., 1923. Geography as human ecology. Annals of the Association of American Geographers 13, 1–14.

Burton, I., 2008. Gilbert White: progress in geography. Progress in Human Geography 32, 448.

Burton, I., Kates, R.W., White, G.F., 1978. Environment as Hazard. Oxford University Press, Oxford.

Christian-Smith, J., Gleick, P., 2012. A Twenty-First Century U.S. Water Policy. Oxford University Press, New York.

Cutter, S., 2006. Hazards Vulnerability and Environmental Justice. Earthscan, London.

Dewey, J., 1929/1958. Experience and Nature. Dover, New York.

Dewey, J., 1934/1958. Art as Experience. Capricorn Books, New York.

Dewey, J., 1938. Logic: The Theory of Inquiry. Southern Illinois University Press, Carbondale.

Dow, K., Downing, T.E., 2011. The Atlas of Climate Change, third ed. University of California Press, Berkeley, CA.

Entrikin, J.N., 1980. Robert Park's human ecology and human geography. Annals of the Association of American Geographers 70, 43–58.

Hewitt, K. (Ed.), 1983. Interpretations of Calamity: From the Viewpoint of Human Ecology. Allen and Unwin, London.

Hinshaw, R.E., 2006. Living with Nature's Extremes: The Life of Gilbert Fowler White. Johnson Books, Boulder, CO.

James, W., 1902/1982. The Varieties of Religious Experience. Penguin Books, Harmondsworth, UK.

Jones, R., 2001. In: Walters, K. (Ed.), Rufus Jones, Essential Writings. Orbis Books, Maryknoll.

Kates, R.W., 2011. Gilbert F. White, a Biographical Memoir. National Academy of Sciences, Washington, DC.

Kates, R.W., Burton, I., 1986. Geography, Resources and Environment. Selected Writings of Gilbert F. White, vol. I. University of Chicago Press, Chicago.

Kates, R.W., Burton, I., 2008. Gilbert F. White, 1911–2006: local legacies, national achievements, and global visions. Annals of the Association of American Geographers 98, 1–8.

Macdonald, N., Chester, D., Sangster, H., Todd, B., Hooke, J., 2012. The significance of Gilbert F. White's 1945 paper 'human adjustment to floods' in the development of risk and hazard management. Progress in Physical Geography 36, 125–133.

Mitchell, J.K., 2008. Perspectives on alternatives: differentiation and integration in pursuit of a better fit between society and nature. Progress in Human Geography 32, 451–458.

Murphy, A.B., 2008. Gilbert F. White (1911–2006): the extraordinary career of Gilbert White. Progress in Human Geography 32, 445–447.

Mustafa, D., 2013. Water Resource Management in a Vulnerable World: The Hydro-Hazardscapes of Climate Change. I.B. Tauris, London.

National Research Council, 1966. Alternatives in Water Management. National Academies Press, Washington, DC.

National Research Council, 1968. Water and Choice in the Colorado River Basin: An Example of Alternatives in Water Management. National Academies Press, Washington, DC.

Palm, R., 2008. Recollections of Gilbert F. White in Colorado. Progress in Human Geography 32, 459–461.

President's Water Policy Commission, 1950. A Water Policy for the American People. U.S. Government Printing Office, Washington, DC.

Showalter, P., Lu, Y. (Eds.), 2010. Geospatial Techniques in Urban Hazard and Disaster Analysis. Springer, Dordrecht.

Thompson, J., Porras, I.T., Tumwine, J.K., Mujwahuzi, M.R., Katui-Katua, M., Johnstone, N., Wood, L., 2001. Drawers of Water II. Summary Volume. International Institute for Environment and Development, London.

Wescoat Jr., J.L., 1987. The 'practical range of choice' in water resources geography. Progress in Human Geography 11, 41–59.

Wescoat Jr., J.L., 1992. Common themes in the work of Gilbert White and John Dewey: a pragmatic appraisal. Annals of the Association of American Geographers 82, 587–607.

Wescoat Jr., J.L., 2006. Gilbert F. White (1911–2006): wisdom in environmental geography. The Geographical Review 96, 700–710.
Wescoat Jr., J.L., White, G.F., 2003. Water for Life: Water Management and Environmental Policy. Cambridge University Press, Cambridge.
White, G.F., 1942/1945. Human Adjustment to Floods. University of Chicago. Department of Geography Research Paper No. 29, Chicago.
White, G.F., 1949. Toward an appraisal of world resources. Geographical Review 39, 625–639.
White, G.F. (Ed.), 1956. The Future of Arid Lands. American Association for the Advancement of Science, Washington, DC.
White, G.F., 1957. A perspective of river basin development. Law and Contemporary Problems 22, 157–184.
White, G.F., 1958. Introductory graduate work for geographers. The Professional Geographer 10, 6–8.
White, G.F. (Ed.), 1961a. Papers on Flood Problems. University of Chicago Department of Geography Research Papers, No. 70, Chicago.
White, G.F., 1961b. The choice of use in resource management. Natural Resources Journal 1, 23–40.
White, G.F., 1964a. Choice of Adjustment to Floods. University of Chicago, Department of Geography Research Paper No. 93, Chicago.
White, G.F., 1964b. Vietnam: the fourth course. Bulletin of the Atomic Scientists 20, 6–10.
White, G.F., 1972. Geography and public policy. The Professional Geographer 24, 101–104.
White, G.F., 1974a. Domestic Water Supply: Right or Good? Proceedings of Ciba Foundation Symposium on Human Rights in Health. Associated Scientific Publishers, Amsterdam, pp. 35–59.
White, G.F. (Ed.), 1974b. Natural Hazards: Local, National, Global. Oxford University Press, New York.
White, G.F., 1997. Classics in human geography revisited: human adjustment to floods. Progress in Human Geography 21, 248–250.
White, G.F., 2002. In: Gould, P., Pitts, F.R. (Eds.), Autobiographical Essay. Geographical Voices: Fourteen Autobiographical Essays. Syracuse University Press, Syracuse, New York, pp. 341–364.
White, G.F., 2013. Website, Including Full List of Publications. www.colorado.edu/hazards/gfw.
White, G.F., Calef, W.C., Hudson, J.W., Mayer, H.M., Shaeffer, J.R., Volk, D.J., 1958. Changes in Urban Occupancy of Flood Plains in the United States. University of Chicago, Department of Geography Research Paper No. 57, Chicago.
White, G.F., Bradley, D.J., White, A.U., 1972. Drawers of Water: Domestic Water Use in East Africa. University of Chicago Press, Chicago.
White, G.F., Haas, J.E., 1975. Assessment of Research on Natural Hazards. MIT Press, Cambridge, MA.
White, G.F., Kates, R.W., Burton, I., 2001. Knowing better and losing even more: the use of knowledge in hazards management. Environmental Hazards 3, 81–92.
Wisner, B., 2003. At Risk: Natural Hazards, People's Vulnerability and Disasters. Routledge, London.

Relevant Websites

http://www.iwr.usace.army.mil/Library/MaassWhiteReferenceRoom.aspx – Arthur Maass – Gilbert F. White Reference Room. U.S. Army Corps of Engineers.
http://www.colorado.edu/hazards/gfw/ – University of Colorado, Boulder; Natural Hazards Center.

White-Collar Workers, History of

M Prinz, Emsdetten, Germany

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16496–16500, © 2001, Elsevier Ltd., with revisions made by the Editor.

Abstract

The category of the 'white-collar' or professional nonindustrial worker has permeated the social sciences since the World War II, although the origins of the term can be found in earlier German thought. White-collar has frequently been opposed to blue-collar or manual/industrial labor, and the distinction has been central to discussions of the nature and likely future of advanced capitalist societies. More recently the usefulness of the term has been questioned, and the class-based nature of social relations opposed by analyses emphasizing the extent to which white- or blue-collar labor reveals little about the social structure and social attitudes.

In most countries the term 'white-collar' worker refers to a group of dependant workers, which compared to manual or 'blue-collar' workers, are occupied not directly in the process of manufacturing. Its delimitation and its specific social profile in relation to blue-collar workers and public servants, however, remain ambiguous and controversial. White-collar workers were introduced into the realm of modern social science at the end of the nineteenth century. The most intensive debates took place in German-speaking countries from about 1890 up to the 1930s against the background of the formation of strong labor movements (largely blue-collar) and in the context of a political system, dominated by conservative elites eager to defend the status quo and in search of numerically important allies among the general population. After World War II the once clear distinction between white- and blue-collar workers were largely blurred and scientific as well as political interest in the group as a whole has faded. It is one chief characteristic of debates in international social science on white-collar workers ever since that they tend to relate only to subgroups as the technical staff ('new worker'), the engineers ('technocracy'), managers ('managerial revolution'), etc. Any wholesale white-collar sociology appears to be a thing of the past.

Preliminary Remarks

Although now in many developed industrial societies white-collar workers represent, at least in numerical terms, the single most important group within the general workforce, their delimitation and their specific social profile in relation to blue-collar workers and public servants remain ambiguous and controversial (Kocka, 1980, 1981; Lockwood, 1989). With hardly any other social stratum was and is the complex interrelation between the actual appearance of a group, its self-definition and its treatment as an object of investigation so clear as with the white-collar employees.

The sociological interest in these groups has a long tradition, which dates back into the late nineteenth century. It arose mainly for two reasons: (1) the exceptional rapid growth of the various groups, labeled as 'white-collar workers,' *Angestellte*, *employŭs*, *impiegati*, *empregados*, etc., and (2) the fact that these groups, with regard to some objective features of their working conditions, aspects of their legal status, as well as their consciousness and their behavior, differed markedly from other social groups already established. Interest was nowhere so strong as in the German-speaking countries, where, as a field of the emerging social sciences, a special *Angestelltensoziologie* developed. This peculiarity was mainly due to the existence of a radical labor movement and of a strong bureaucratic tradition. The institution of the privileged 'civil servant' served as a role model for many white-collar workers and attributed much to the relative depth of the distinction in relation to the blue-collar workers in the respective countries. However, comparative historical research shows that the terms 'white-collar workers,' *Angestellte*, *employŭs*, *impiegati*, and *empregados* are in no way interchangeable. The delimitation and, particularly, its social 'meaning' vary strongly historically and from country to country. Though much of the special social profile of the various groups labeled as 'white-collar' has faded away, there still remain some important differences, especially in relation to blue-collar workers.

Definition by Functional Criteria

In many countries in everyday language as well as in scientific usage the term white-collar worker refers to a group of dependent workers, which compared to manual or blue-collar workers, is occupied not directly in the process of manufacturing. Rather it is situated in functional as well as in spatial terms *before* the process of manufacture, *beside* it and *behind* it. One encounters white-collar workers in all economic sectors, even agriculture, at least, as it is organized commercially and as a large-scale business. However, the actual domain of the employees is the service sector that is, trade, banking, insurances, where they always made up the predominant part of the workforce. In addition, large groups of employees are to be found in industry and the public sector. Since the end of the nineteenth and above all in the course of the twentieth century employees have increased their share in all sectors considerably in all developed societies compared to that of manual workers and public servants (**Table 1**).

Table 1 White-collar workers in US, Great Britain, and Germany 1882–1950 (in percent of the total labor force)

	USA	*Great Britain*	*Germany*
1882			6.1
1900	12		8.5 [1895]
1910	15	18.7	11.7 [1907]
1921	19 [1920]	21.2	16.0 [1925]
1931	22 [1930]		17.1 [1933]
1939	24 [1940]		18.9
1950	30.9	30.9 [1951]	19.9
1970			35.7
1987			50.1

Source: Behringer, P., 1985. Soziologie und Sozialgeschichte der Privatangestellten in Großbritannien. Frankfurt, Germany, p. 515; Kocka, J. 1980. White-collar Workers in America: 1890–1940: A Social-political History in International Perspective. Sage, London, p. 42f; Kunig, p. 172. Note: These numbers include white-collar employees in private businesses as well as in the public sector.

With respect to their productive functions white-collar workers differ from manual workers in that they – varying with the economic sector in which they are employed – perform directing, overseeing, constructive, analyzing, administrative, commercial, and information-processing tasks. These criteria may lead, in many cases, to a clear distinction between manual workers and white-collar employees in private enterprises. They are hardly suitable for a clear distinction between officials or civil servants (*Beamte*) and employees (*Angestellte*) in the public sector. The functions of these two categories normally overlap. But the upper ranks of the hierarchy are staffed by employees with the status of civil servant (Humphreys, 1958). *Beamte* and *Angestellte* differ by legal status rather than by types of functions.

Origin and Growth

The origins of some white-collar professions may be traced back to the premodern period. Among them figure, for instance, the commercial employees, which at that time still had a good chance to reach the status of a self-employed salesman later in life. Nevertheless as a mass stratum one has called white-collar employees rightly the 'latest child of industrial capitalism.' Two main developments contributed to the rapid growth of the group: (1) the shifting balance between the 'first' (agriculture) and the 'second sector' (industry) in favor of the 'third,' the service sector (Fourastiй, 1963) – a development that took place first in the Great Britain while other industrial nations followed suit in the course of the nineteenth and twentieth century; (2) the rise of 'Organized Capitalism' (Hilferding, 1981) since the end of the nineteenth century, a phase characterized by the growing importance of organization, planning, information-gathering, bureaucratic control, and the creation of structures within and outside industry to serve these ends. The relative growth of the share of the employees within the public sector, however, is determined not only by changes in the function of the modern state but also to a large extent by the political will either to expand or to restrict the spread of particular privileges attached to the status of civil servant (*Beamte*). Such privileges are, for example, the protection against dismissal and pension schemes, etc.

As recent social data show, even in the most advanced contemporary industrial societies the expansion of white-collar jobs has not yet come to an end. But there are clear signs that the rate of expansion slowed down significantly in the second half of the twentieth century. As with all social growth, the creation of white-collar jobs as a strategy to tackle problems arising from the development of economy and society as a whole is subject to the law of diminishing returns. While in the early stages of industrial capitalism white-collar employees were instrumental in the process of rationalizing production – a fact that caused widespread resentment among manual workers – in further stages they were themselves subjected to strategies of cost-cutting and rationalizing. This development began very soon after the turn of the twentieth century with the introduction of specific office machines as the typewriter, the Hollerith machine – an early form of large-scale data processing – and other changes in the structure and organization of office work. This process definitely has gained a new quality in recent years. In particular, the spreading of the computer, the communication revolution, and the integration of office work into controllable networks has opened new chances for quantitative control of employee activities once considered to be typical only for manual work. Whether this tendency is outweighed by other trends toward 'lean management' and more autonomy on the job remains to be seen.

Status and Privilege

The Swedish–German sociologist Croner (Croner, 1951) developed his 'theory of delegation,' according to which, most early white-collar jobs in industry emanated through a process of division of labor and of delegation out of functions originally performed by the employer. This theory has been questioned by empirical research. Especially in modern industries many tasks performed by white-collar workers actually represent new functions without any historical precedent. It seems highly questionable that *all* early functions classified as white-collar carried with them high prestige. Especially, in the area of small-scale shop keeping there existed extremely exploitive working conditions for clerks with low pay, no job security, long working hours, etc. (Anderson, 1976; Crossick, 1977). Nevertheless, it holds true that many white-collar occupations enjoyed at least some form of privilege in comparison with manual workers, and continue to do so. Some of the differences resulted from the specific nature of their work that could not be subjected easily to quantitative methods of control. One of the traits that distinguished most – not all – white-collar jobs from blue-collar functions was a fixed income, the 'salary,' that, at least in principle, did not vary with working hours or the amount of work carried out. Because of the fact that knowledge of business transactions within the specific firm constituted the main qualification of many white-collar employees, the accumulation of such individualized knowledge gave members of the group a higher job security

than manual workers, especially in the early stages of industrialization. Even in later days, the office never became the place for a general 'hire and fire' climate. As many positions of white-collar workers were integrated into large, delicately differentiated bureaucratic hierarchies, aspirations for upward mobility never vanished completely. Besides, in many countries there were, traditionally, legal privileges attached to certain white-collar occupations. These entitlements included protection against dismissal, paid vacation, continuation of wage payment in the case of illness, company pensions schemes, etc.

The idea that the class of white-collar workers as a whole would face 'proletarianization,' as put forward especially by writers in the Marxist tradition, has given way to a much more nuanced picture. At earlier stages of the development legal privileges attached to certain types of white-collar workers were generalized among larger groups. Moreover, the French sociologist Michel Crozier has argued that the consequences of rapid rationalization were often absorbed by the constant influx of female workers. Though the equation 'proletarianization-feminization' does not come out, there is some degree of truth in it. New styles of management – the movement toward 'lean management' since the 1980s, for instance – have amplified autonomy and reopened scope, at least, for certain groups among the white-collar population. On the other hand, there can be no doubt that a process of narrowing and even blurring of differences between blue-collar and white-collar workers took place in the course of the twentieth century. Important groups within the white-collar population, especially clerical workers and sales personnel, have been overtaken by skilled manual workers in terms of income, job security, even prestige. Many of the rights once attached exclusively to certain white-collar positions have, step-by-step, been granted to the workforce as a whole, making white-collar workers, at least in some countries such as Germany, important, if involuntary, harbingers of the modern social state. Though, on average, white-collar workers still appear to be privileged in relation to the manual workforce, the 'average' white-collar worker largely has become a statistical fiction due to the immense variety of positions and professions that nowadays characterize these categories.

Organizational Behavior

Depending on the structure of the trade union movement within the respective countries, white-collar workers tend to organize exclusively or are integrated into general unions (Behringer, 1985; Bain and Pollins, 1965; Bain, 1972; Jenkins and Sherman, 1979). In most cases union membership among white- and blue-collar workers overlaps at least to a certain degree. The main area in which both groups differ is trade union membership in general. White-collar workers are considerably less well organized than manual workers with the one exception of the public sector. Hence, their growing weight among the workforce represents one of the most important factors in the crisis of the trade union movement in developed countries in general.

National Differences and Peculiarities

While in most industrialized countries the division between white- and blue-collar workers was never clearly marked or valid only in regard of relatively small groups, in some European countries such as Austria, Czechoslovakia, Switzerland, and Germany until the middle of the twentieth century, blue- and white-collar workers constituted distinguished entities, separated in many aspects, legal as well as organizational and ideological. Social legislation, especially pensions schemes, had been enacted before and after the World War I. They had an important impact, strengthening existing differences between white- and blue-collar workers. The dividing line between the two groups structured large parts of the fabric of these societies – legislation, trade union movement, self-help organizations, political affiliations, etc. By the 1950s, the distinction between white- and blue-collar had begun to erode gradually, a development that was supported by the expansion of former white-collar privileges to the general workforce. Another important aspect in which German-speaking countries before the World War II differed from other societies was the sheer size of the legally defined white-collar population, which included sales personnel as well as general managers.

Social Theory

White-collar workers were introduced into the realm of modern social science at the end of the nineteenth century. The starting point was twofold: (1) the notion of a growing number of occupations, employments, and functions that shared many traits of the working class while at the same time differing in important aspects from manual workers as well as from the higher social strata; and (2) the exceptional rapid growth of these categories in the decades around the turn of the century, which even outdid that of the manual workers.

The most intensive debates on white-collar workers took place in German-speaking countries from about 1890 up to the 1930s against the background of the formation of strong labor movements (largely blue-collar) and of a political system, dominated by conservative elites eager to defend the status quo and in search of numerically important allies among the general population. In the first stage of the debate, which took place before the World War I, status quo-oriented scholars identified white-collar workers as part of a new socially stable and conservative 'middle estate' and, at the same time, encouraged the state to engage in a social-protective policy (e.g., white-collar specific insurances and labor laws) while Marxist theorists predicted a general proletarianization of these groups. The 1920s saw a convergence of opinions, reinforced by the skeptical notion that the 'new middle class' in German-speaking countries seemed to figure prominently among the mass base of National Socialism. The label of a potential for right-wing extremism has stuck to the group ever since, not least through the intellectual influence of German social scientists, many of whom were forced to emigrate in the early 1930s (Fromm, 1984; Speier, 1986). The notion showed up again

in many theories of fascism developed in the 1950s and 1960s, generalizing what was seen to be the experience of the interwar years (Lipset, 1960). In contrast, social historians who took up the debate in the 1970s tried to demonstrate the peculiarities of the German and Austrian experience (e.g., Kocka, 1980) while latest research has challenged the notion of a predominantly right-leaning, 'new middle class' in the 1920s altogether (Falter, 1991).

The debate over the role of white-collar workers in the mass base of National Socialism represented the climax and at the same time a turning point. Interest in interpretations of the white-collar group as a whole has subsided (but see Schulz, 2000) in Germany as well as in other countries where it never figured prominently.

Not surprisingly, with regard to the heterogeneity of the group, social science attributed contradictory attitudes and social functions to these groups, often departing from specific traits of subgroups. One important theme over the decades has been the impact the expansion of white-collar occupations has had on the emancipation of women, whether it offered new opportunities for social mobility or whether it only helped to reproduce traditional inequalities (Suhr, 1930). There has been much research on the impact of rapid technological change on office work through the decades, some with more pessimistic, others with more optimistic overtones. Studies of the service sector, for instance, stress its relative resistance to change as a result of the need to maintain a workforce even when there is no constant demand, a fact, which has been labeled 'guard labor.' Other scholars have pointed to the partial regain of autonomy on the workplace as part of the automation movement in the 1970s, which was said to have led to the formation of a 'new worker' (Mallet, 1963). Actually, the debate referred to members of the technical staff supposedly holding key positions for potential labor disputes. Another important context was the rise of modern consumer society since the World War I in Europe and the United States. The concentration of large numbers of white-collar employees in larger cities led to the identification of them with the urban population as a whole, regarding them as pioneers of modern consumption and new forms of leisure (Coyner, 1977). This notion dates back to the influential study first published in 1924, of the German social writer, Siegfried Kracauer (Kracauer, 1998) and translated into many languages ever since. It has been taken up in various studies of the historic development of modern consumption and leisure, leaving out, with regard to the white-collar population, those members of the group living in smaller and middle-size cities.

Though, in the second half of the twentieth century, interest for white-collar employees in the social sciences has flared up from time to time, there seems to be no future for any revival of a wholesale white-collar worker sociology as it developed in some European countries during the interwar years. One of the most important prerequisites, the existence of a self-conscious socialist working class, no longer exists. With its disappearance the interests of social science, policy makers, and the different groups that make up this strata to develop a broad definition based upon the distinction between 'us,' the 'white-collar,' and 'them,' the 'blue-collar workers,' seem to have vanished. What remains is a remarkable gap between the numerical importance of the group in question and its relative neglect by social scientists in the last decades.

See also: Bourgeoisie and Middle Classes, History of; Income Distribution and Inequality: Measurement Issues; Professions and Professionalization, History of; Professions, Sociology of; Social Inequality in History (Stratification and Classes); Social Mobility, History of; Work, History of; Work, Labor: History of the Concept; Working Classes, History of.

Bibliography

Anderson, G.I., 1976. Victorian Clerks. University Press, Manchester.
Bain, G.S., 1972. The Growth of White-collar Unionism, second ed. Clarendon Press, London.
Bain, G.S., Pollins, I.I., 1965. The history of white-collar unions and industrial relations: a bibliography. Bulletin of the Society for the Study of Labour History 11, 20–25.
Behringer, P., 1985. Soziologie und Sozialgeschichte der Privatangestellten in Großbritannien. Frankfurt, Germany.
Bonham, J., 1954. The Middle Class Vote. Faber of Faber, London.
Coyner, S.J., 1977. Class consciousness and consumption: the new middle class during the Weimar republic. Journal of Social History 10, 310–331.
Creese, G.L., 1999. Contracting Masculinity: Gender, Class and Race in a White-collar Union, 1944–1994. Oxford University Press, Toronto, Ontario.
Croner, F., 1951. Tjänstemannakåren i det moderna Samhället Stockholm (German, revised ed. Frankfurt 1954).
Crossick, G. (Ed.), 1977. The Lower Middle Class in Britain 1870–1918. St. Martins Press, New York.
Crozier, M., 1971. The World of the Office Worker. (D. Landau, Trans.). University of Chicago Press, Chicago.
Falter, J.W., 1991. Hitlers Wðhler. Beck, Munich.
Fourastiй, J., 1963. Le grand espoir du XXe siècle. Gallimard, Paris.
Friedrichs, D.O., 1996. Trusted Criminals: White Collar Crime in Contemporary Society. Wadsworth Pub. Co., Belmont.
Fromm, E., 1984. The Working Class in Weimar Germany – a Psychological and Sociological Study. Bierg, Leamington Spa, UK (B. Weinberger (Ed.), Trans. and with an introduction by gang, W. Bonss).
Gerstl, J.E., Ilutton, S.P., 1966. The Engineers: The Anatomy of a Profession. A Study of Mechanical Engineers in Britain. Tavistock Publications, London.
Giddens, A., 1981. The Class Structure of the Advanced Societies, second ed. Hutchison, London.
Hilferding, R., 1981. Finance Capital: A Study of the Latest Phase of Capitalist Development (Ed. with an introduction by Bottomore, T., from translations by M. Watnick, and S. Gordon). Routledge, London.
History of Women in the United States, 1993a. Pt 2, Saur, Munich.
History of Women in the United States, 1993b. Pt 1, Saur, Munich.
Humphreys, B.V., 1958. Clerical Unions in the Civil Service. Blackwell & Mott, London.
Jenkins, C., Sherman, B., 1979. White-collar Unionism. The Rebellious Salariat. Routledge, London.
Kater, M.H., 1983. The Nazi Party – a Social Profile of Members and Leaders, 1919–1945. Blackwell, Oxford, UK.
Kocka, J., 1980. White-collar Workers in America: 1890–1940: A Social-political History in International Perspective. Sage, London.
Kocka, J. (Ed.), 1981. Angestellte im europðischen Vergleich. Die Herausbildung angestellter Mittelschichten seit dem spðten 19. Jahrhundert.
Koenig, M., 1991. Die Angestellten unterwegs. Vom Berufsstand zur modernen Gewerkschaft 1890–1990. Bund-Verlag, Kuln.
Kracauer, S., 1998. The Salaried Masses–duty and Distraction in Weimar Germany. Verso, New York (Q. Hoare, Trans., and with an introduction by I. Mulder-Bach).
Lederer, F., 1979. Kapitalismus, Klassenstruktur und Probleme der Demokratie in Deutschland 1910–1940. Ausgewðhlte Aufsðtze von Emil Lederer herausgegeben von Jurgen Kocka. Vandenhoeck und Ruprecht, Goettingen, Germany.
Linder, M., 2004. "Time and a Half's the American Way": A History of the Exclusion of White-collar Workers from Overtime Regulation, 1868–2004. Fǎnpihuà Press, Iowa City, Iowa.
Lipset, S.M., 1960. Political Man: The Social Bases of Politics, first ed. Doubleday, Garden City, NY.

Lockwood, D., 1989. The Blackcoated Worker – a Study in Class Consciousness, second ed. University Press, Oxford, UK.
Mallet, S., 1963. La Nouvelle Classe Ouvrière Paris (Spanish: La nueva condicion obrera. Madrid 1969).
Mazsu, J., 1997. The Social History of the Hungarian Intelligentsia, 1825–1914. Social Science Monographs. Boulder, Colorado.
Mills, C.W., 1951. White-collar – the American Middle Classes. Oxford University Press, New York, London.
Parker, D.S., 1993. The Rise of the Peruvian Middle Class: A Social and Political History of White-collar Employees in Lima 1990–1950. Stanford University.
Purcell, A.D., 2009. White collar Radicals: TVA's Knoxville Fifteen, the New Deal, and the McCarthy Era. University of Tennessee Press, Knoxville.
Reader, W.J., 1966. Professional Men: The Rise of the Professional Classes in 19th Century England. Weidenfield & Nicolson, London.
Schulz, G., 2000. Die Angestellten seit dem 19. Jahrhundert. Oldenbourg, Munchen, Germany.
Sobel, R., 1989. The White-collar Working Class. From Structure to Politics. Proeger, New York.
Speier, H., 1986. German White-collar Workers and the Rise of Hitler. Yale University Press, New Haven and London.
Suhr, S., 1930. Die weiblichen Angestellten. Zentralverband der Angestellten, Berlin.
Touraine, A., 1969. La société post-industrielle. Paris.
Wilson, R.G., 1999. Disillusionment or New Opportunities?: The Changing Nature of Work in Offices, Glasgow 1880–1914. Ashgate, Aldershot.

Whiteness and White Supremacy

June Ying Yee, School of Social Work, Ryerson University, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The profession and practice of social work is steeped in whiteness and white supremacy. Whiteness can be viewed as the normalization of helping those who are marginalized via a set of practices and techniques that, in the end, work to maintain the status quo of failing to recognize systemic and structural inequalities that privilege white people. A review of how the social work profession, primarily in North America and around the world, has been shaped by the historical and global evolvement of whiteness and supremacy is considered.

Introduction

Historically, social work practice in North America evolved from two different movements: the Settlement House and Charity Organization Society (COS). The COS started in 1869 in England. In 1877, the United States promoted scientific philanthropy, which emphasized the need to rehabilitate poor people through individual case work (Lundy, 2004). Mary Richmond's first book, titled *Friendly Visiting among the Poor*, provided a manual on how to promote individual change among poor people in an advancing and growing industrial society. At the same time, Jane Addams (1860–1935) promoted social change through the establishment of settlement houses. Addams believed that the state should help poor people, whereas Richmond argued that those who were more fortunate could provide charity work and help for poor people. The approaches of each of these movements signaled a difference in their understanding of how to work with those who are marginalized in society, with the former providing a more individual level focus and the latter providing a social justice emphasis. Nonetheless, both the COS movement and Settlement House practices of intervention were aligned with the dominant values of society, that is, whiteness (Bradley, 2009; Dominelli, 2010; Lundy, 2004; Smith, 1999; Thobani, 2007).

Whiteness can be viewed as the normalization of helping those who are marginalized via a set of practices and techniques that, in the end, maintains the status quo of privileging white people and failing to recognize systemic and structural inequalities. More recent scholarship in social work (Todd, 2011) examines the performative acts of whiteness that simultaneously positions white social workers into a place of 'goodness' with good intentions, while also imposing techniques and practices onto clients that may not be very helpful, if not harmful. From an anticolonial perspective, an understanding of whiteness has helped to shed light on the harmful colonial practices inflicted on particular marginalized communities such as black and Aboriginal peoples. For example, the child welfare system in North America has an overrepresentation of black children and Aboriginal children in care (Chand, 2000; D'Souza, 1994; Delfabbro et al., 2010; Dettlaff and Rycraft, 2008; Trocmé et al., 2004). Questions are being asked about the role and relationship between whiteness, colonialism, and structural inequalities in creating such outcomes.

The recent social work literature (Baines, 2011) has emphasized how social inequalities experienced by service users are often transformed into social problems that social workers are expected to treat as technical problems to problem-solve and fix. In doing so, social control of marginalized people takes place while simultaneously concealing the social inequalities that arise from gender, race, and class inequalities. For example, the use of risk assessment models on clients in the Canadian child welfare system (Krane and Davies, 2000) has been viewed as a technical instrument used to gain social control over service users. Krane and Davies (2000) make a compelling argument that risk assessment models are set up to show the 'dangerousness' of clients' behaviors from the perspective of the dominant group. The consequence is that families are not well-supported or helped at the structural level and, more significantly, marginalized people are put under more scrutiny by public officials and the welfare state than those who hold privilege and power in society (Aronson and Smith, 2011; Baines, 2011, Dumbrill, 2010). This article discusses the role of whiteness and white supremacy in shaping not only the professional identity of social work but also the practice of social work. The article starts with an in-depth and detailed discussion of what is whiteness and white supremacy, and how our understandings of whiteness has shifted and changed over time.

What Is Whiteness?

In North America, the concept of whiteness is a fluid, dynamic, and situationally specific social and cultural process that positions those who are white, in terms of skin color, into a place of power and privilege. Conversely, those who are black, in terms of skin color, are positionally located into a place of powerlessness, marginalization, and oppression. Differential treatment can result in other markers that exclude such as religion, language, gender. However, in this context, whiteness primarily focuses on race as the entry point through which to understand exclusion. The operation of whiteness is complex because the white race is mediated by other forms of identity such as class, gender, disability, and sexuality. Within one's identity, a person can have sites of privilege and oppression operating at the same time (Carniol, 2005; Styhre and Eriksson-Zetterquist, 2008; Yee and Wagner, 2013). Due

to this complexity, such forms of intersecting identities can cause a person who is white not to recognize the power of whiteness within his/her own identity and, instead, to turn to places of innocence and shame. This examination of emotions, known as affective economies, is a burgeoning area in whiteness studies and is paramount to showing why whiteness is a difficult topic to discuss (Ahmed, 2012; Harding and Pribram, 2004; Hunter, 2010). This place of innocence is known as the crisis of whiteness (Bonnett, 2000; Todd, 2011; Wagner, 2010). To date, in the Western world, white people have not been able to reconcile nor confront the historical and present-day enactments of racism that are routinely inflicted upon Aboriginal peoples and people of color. For example, although antiracism is known to be a social movement working toward the dismantlement of institutional systems that benefit white people, we still see the repositioning of white antiracism activists as the 'good whites' (Todd, 2011; Wagner, 2010). Or, another example is how many white feminists, although aware of the way oppression works across gender, may still have difficulty recognizing and discussing their own white privilege while discussing very clearly the dynamics of gender inequality (Lorde and Rich, 1981; Todd, 2011; Wagner, 2010).

Around the Western world, power and privilege is given to those who hold a white racial identity. Even though the definition of white, as a racial identity, is socially constructed, past and present-day political, social, and economic arrangements reaffirm the power of the white race. It is no accident that many examples exist of the differential and negative treatment of those who are nonwhite and Aboriginal. This process of differential treatment is known as racialization. Behind all processes of racialization rests the ability of those who hold power through the white race to grant power or take away power from people. In the Western world, the white race has garnered power through acts of colonialism that are ultimately rooted in notions of superiority and used to justify the economic exploitation of those who are nonwhite. The legacy of colonialism inflicted upon Aboriginal peoples and people of color continues into present day institutional systems and structures without people being aware of it happening. The way it manifests is not always readily apparent or detectable. This is because the enactment of racism itself has been routinized, and naturalized through the strategies and processes of whiteness. Whiteness predates the modern era and Kincheloe (1999) suggests that whiteness arose in Western European societies, during the Enlightenment period, with the construction of a white male rational subject who was able to transcend time, space, and history. Power and privilege is granted to white males through three social and cultural processes (Gabriel, 1998: p. 13): (1) exnomination, that is, the ability to shape the norms and values of everyday life and a failure to ethnicize one's own self; (2) naturalization, that is, seeing the representation and embodiment of white culture as the referential social norm without having to ever define one's own self; and (3) universalization, that is, taking for granted that the worldview of white people is the way to understand everything through its own historical and sociopolitical vantage point. These three social and cultural processes are the strategies and processes of whiteness.

A common theme emerges, which is the invisibility and normalization of social inequality taking place throughout the Western world. If society fails to see the processes of racialization occurring, then those in positions of power and dominance are able to continue to prevail with impunity. This practice of failing to see or notice social inequality is worldwide, but can be applied and transferred to different contexts. For example, although Green et al. (2007) discuss the operation of whiteness in South Africa, they make the insightful comment that "white people [are able] to distance themselves from racism and racial privilege through a range of discourse and practices, which also [simultaneously] protects and secures the dominance and normativeness of whiteness." The ultimate effect is that white people are not required to deal with the realities of white supremacy, which is at the root of the social and cultural process of whiteness (Schick, 2012). According to Schick (2012: p. 13), white supremacy can be defined as

> reasserting white space is a performative act that accomplishes white supremacy and white identity; it also demonstrates and confirms the white racial knowledge of how to do this. "White racial knowledge limits and defines what whites are prepared to hear about a subordinate group and still 'allow a self-concept of innocence to continue' Nelson, 2002: p. 224."

Specifically, in the United States in the 1990s, a proliferation of scholarly studies on whiteness studies increased the consciousness of Americans to the insidious ways in which power and dominance continues to exclude black people from equal participation in society. To illustrate this in the United States, the phenomenon of 'white flight' takes place when a large number of white people leave a racially mixed neighborhood to a more racially white neighborhood due to an irrational fear and anxiety of black people. This fear and anxiety is fueled by the material reality of a legacy of slavery, which effectively denigrated and dehumanized black people. Such a historical past set a material foundation for creating present-day stereotypical assumptions about black people's ability, merit, and worth in the economic labor market. Therefore, in the United States, the economic value of white privilege is dependent on the continued denigration and exclusion of great number of blacks in the United States from the paid labor market even after the abolition of slavery (Baldwin, 2012). This pattern of segregation results in white people benefitting from better chances in obtaining a job, higher wages, capital, and job mobility. Another example is the normalization of white segregated neighborhoods that came through real estate developments that independently incorporated the upkeep and maintenance of such neighborhoods (Baldwin, 2012). This privatization of payment of services effectively allowed people living in these economically exclusive neighborhoods to no longer contribute to general municipal services. The recipients of general municipal services are often associated with low-income neighborhoods, where a lot of nonwhite people may live. This racial divide in the spatial environment of neighborhoods, therefore, not only reflects a fear of blacks but also bolsters the economic value of whiteness (Baldwin, 2012). As evidenced by this example, new forms of colonization are being recreated under the guise of capitalism, merit, and entitlement to private property. In the Western world, capitalism is cherished as a system of fairness, equality, and opportunity but

in reality benefits mainly white people and hides the unequal power relations. Ultimately, this way of knowing the world is created and recreated through the mobilization of strategies and processes of whiteness that are historically undergirded by white supremacy. In the next section, a more detailed discussion on white supremacy is provided.

What Is White Supremacy?

The roots of white supremacy begin with the European white colonizers gaining power and control over the economic, political, and social institutions of the world. The global impact of racial colonialism and justification of capitalist systems to exploit, marginalize, and colonialize people of color and Indigenous peoples are taken for granted as an invisible norm that hardly anyone examines for its current day consequences. One could argue, therefore, that the rise of the modern era of capitalism and past colonial expansion across the globe has secured the dominance and power of the white race with the consequent negative suppression, erasure, and decimation of the lifeworlds of Aboriginal peoples and people of color. Rabaka (2007: p. 8) draws comparisons between the past and present by noting that, in the past, white supremacy enabled whites to use 'racial colonialism and colonial capitalism' to rob Aboriginal peoples and people of color of their autonomous knowledge base while, in the present, we see how 'white hegemonic views and values, white conquest, and racialization colonization' are a reward to whites as evidenced by the inordinate amount of power and privilege granted to them (Allen, 2001; Gordon, 1997; Thobani, 2007).

Therefore, it is important to ensure that any analysis of whiteness moves beyond understanding white supremacy as just the inherent belief in the biological superiority of white people. It is a misnomer to understand white supremacy as only about the extreme hate groups, such as the Klu Klux Klan in the United States. Rabaka (2007: p. 3) argues that white supremacy is "not simply a global and social phenomenon, but a personal and political one as well" that intricately combines three interpretations at the same time: (1) Often race is essentialized into binary understandings of black and white with the result of creating valuations that mislead most to conflate race with culture. For example, Green et al. (2007: p. 397) note that "in colonial contexts, whiteness was represented by orderliness, rationality and self-control, while non-whites were represented by chaos, irrationality, violence, and a lack of self-control" (Kincheloe and Steinberg, 1998); (2) The Marxist analysis of racial inequality, which argues, from a political economy viewpoint, that the 'other' was marginalized for profit and gain of the capitalist system; and (3) Race as a social construction emerged in the modernist era due to Europe's desire for global dominance. This colonial expansion coincided with the rise of science and the Enlightenment period of the white, male rational subject. Regardless of which perspective one favors the resultant outcome is the need to recognize the contemporary reproduction of white supremacist values and practices in the profession of social work. Social work practice operates in the context of processes of whiteness, racism, and colonialism. Therefore, it is important to discuss the evolution of whiteness over time in society before one can begin to understand the ways in which social work practice is entrenched with the social and cultural process of whiteness and white supremacy.

Three Waves to the Study of Whiteness

Whiteness studies (Delgado and Stefancic, 1997; Fine, 1997; Frankenberg, 1997), also known as critical white studies, have crossed into several transdisciplinary boundaries, including ethnic studies, cultural studies, and critical geography. According to Twine and Gallagher (2007), three waves of whiteness studies can be delineated in the scholarly literature: (1) the works of W.E.B. Dubois and other US black scholars; (2) the works of critical legal theorists, feminist theorists, and black theorists; and (3) the development of empirical research using various research methodologies that delve into the operations of whiteness in different contexts whether it be through personal narratives or critiques of current institutional practices.

It is important to note that the ideas that form our current understandings of the white race, and its ideological formations, were central to the scholarly writings of US black scholars for over a century (Twine and Gallagher, 2007) and form the first wave of whiteness. Twine and Gallagher (2007) note that W.E.B. Dubois made three central observations about race and whiteness that lay the theoretical foundation for critical whiteness studies. First, because white laborers from southern and eastern Europe in the Black Reconstruction period in America (1860–80) embraced the identity of the white race, rather than forge a solidarity with freed slaves, these white laborers were able to acquire social and material advantages that only further buttressed the creation of the power of the white racial identity. In other words, white workers were able to gain admission to the social and material privileges of the class that formed the dominant group via a white racial identity, rather than seeing themselves as marginalized like black workers. Although these white workers received a low wage for their labor, they were seen as holding better social status and opportunities than blacks. These white workers who were previously seen as unambiguously white were now seen as part of the privileged white race. Most whites are not consciously aware of their own racial prejudice and such unawareness helps to support white supremacy. W.E.B. Dubois argues that white supremacy is premised on the failure of whites to see that institutional racism exists, despite overwhelming data that shows the racial inequality of blacks. Second, given this historical foundation of not recognizing the discrimination and differential treatment experienced by blacks, at a structural level, W.E.B. Dubois continues his argument by noting that threats to the nationalism of the country, or rather white culture, are coming from the poor culture of blacks. These cultural deficit arguments deflect attention away from the center of white dominance and, in particular, fail to notice or scrutinize its operations of power. Therefore, instead of scrutinizing the source of power, more attention is paid to the marginality and threat of blacks to the economic stability of the US economy. Thirdly, W.E.B. Dubois recognized that this problem of racism toward blacks otherwise known as the 'color line' extended beyond the US border and was a worldwide, international phenomenon.

The second wave of whiteness studies builds on the previous work of US black scholars by challenging the works of mainstream sociologists writing in the area of race. They include the works of feminist scholars, black critical legal theorists, and black theorists. They criticized race scholarship in the United States by noting how white scholars dominated the thinking and approaches of race theory, which were limited to understanding racism as an individual psychological problem, rather than a systemic one. One example of viewing racism primarily at the individual level is through the work of Myrdal (1944). Although Myrdal (1944) did examine the widespread racism that affected every aspect of US society as well as the many Americans' strong belief that equal opportunity is available to all regardless of race, Myrdal (1944) did not go far enough in recognizing that white Americans, did indeed, receive material and economic rewards due to the racial inequality of blacks. But, more importantly, Myrdal (1944) also failed to see the significance of how deeply embedded white supremacy organized America's entire institutional arrangements to the point that people were in denial of this possibility. Consequently, structural interpretations on race were not explored until the second wave of whiteness studies.

Therefore, the goal of the second-wave whiteness writers was to show the invisibility of the white race as the taken-for-granted referential norm that operates not only in US institutions and systems, such as in the courts, but also in the everyday popular discourse and culture of society. For example, Morrison (1922) criticized literary works for defining and equating the word 'American' as synonymously meaning 'white' while all other races were marked as different and less than. Furthermore, Twine and Gallagher (2007) note how the solidification of white supremacy is exemplified in the different historical analysis of various authors' works (Allen, 1994; Jacobsen, 1998; Roediger, 1991). These show how those who are ambiguously white were able to align themselves with the workings of whiteness and, in turn, reap the benefits of white privilege.

Currently, researchers and scholars are in the third wave of whiteness studies. The emphasis of the third wave of whiteness is to examine how people learn and unlearn whiteness along with the complex and different ways in which whiteness can be deployed as a performative strategy whereby those who are white are able to retreat into a place of innocence and non-culpability with acts of racism (Ahmed, 2012; Warren, 1999; Arai and Kivel, 2009). In particular, researchers are interested in using self-reflexivity as a way to understand how whiteness operates in how researchers come to know things and inevitably look at how the workings of whiteness are made invisible in research methods and institutional practices. This is challenging to do since even antiracist authors inevitably reproduce their own power relations while attempting to undo whiteness. In particular, greater insight is also provided on the ways in which white supremacy is undergirding much of antiracism work. Berg (2012) provides four examples of how feminist critical race studies (Thompson, 2003; Wiegman, 1999) reproduce white supremacy while attempting to undo whiteness. First, when white scholars write about whiteness, they run the risk of appropriating the earlier works of scholars of color and Aboriginal peoples and, as a consequence, those who read their work will see them in a favored light, whereas for people of color (Srivastava, 2005; Fellows and Razack, 1998) who speak about the same issues experience violence, harm, and negative consequences to their own personal selves. This difference and impact is hardly discussed in the literature. Second, liberal white people who adopt color-blind approaches to race do not always see how they, too, benefit from an overall system that is stepped in white supremacy and are adept at disassociating themselves from the extremist forms of racism. And, third, even when white feminists or profeminists pursue an antiracist identity, they inevitably recenter themselves, in the process, by claiming a position of innocence. In claiming innocence, these white antiracist activists ignore their own complicity in the reproduction of white supremacy. The literature (Schick, 2012; Todd, 2011) being developed in this area is known to be tackling the problematic of whiteness.

Although earlier scholarship on whiteness was primarily focused in the United States, more recent scholarship is exploring the impact of the dominance of European powers as producing Eurocentric thought not only in international contexts, but also in local contexts. Certainly, European colonization in South Africa, South America, Africa, Australia, New Zealand, North America, and the Caribbean has greatly influenced the global and economic order of the world. Still forthcoming is more research and scholarship that puts forth a global and comparative analysis on the workings of the ways in which whiteness and white supremacy operate together (Allen, 2001; Bhattacharyya et al., 2002).

Professional Identity of Social Work Steeped in Whiteness and White Supremacy

Historically, the social work profession struggled to create a unique professional identity of its own that carried legitimacy and merit in comparison to other professions, such as the medical profession. Flexner, a well-known and recognized expert on professional schools, remarked that social work did not seem to be a profession because of its failure to demonstrate a specific knowledge base with unique activities and techniques of its own with specialized professional expertise. For example, social workers were seen as helpers, but not as expert professionals. To gain legitimacy as a profession, Richmond published the book *Social Diagnosis* based primarily on the medical profession's scientific approach of diagnosing, and treating a client's presumed individual pathology. This approach to social work was in alignment with the already preexisting value base of the COS, which includes principles of administration and efficient organization, a central feature of scientific philanthropy. It is thus, no accident, that the setup of social service organizations, today, is primarily based on this model of organizational practice on how to serve clients. Examples of social service agencies that follow this task-centered model approach to practice include child welfare and family service agencies, boy and girls clubs, counseling agencies, and drop-in centers.

As discussed earlier, the scientific model is a part of the Enlightenment movement that places enormous value on the white, male rational subject who garners power and dominance in the creation of knowledge about the 'other.' However, most social workers are unaware of the predominance of this

colonialist and imperialist underpinning of the social work profession because of the effective deployment of strategies and processes of whiteness that operate in everyday social work practice (Yee and Dumbrill, 2003; Yee, 2005). More importantly, this overemphasis on individual counseling and use of the scientific model has caused social workers to shift away from studying the social conditions and systemic causes of people's marginalization in society. In fact, Lundy (2004) argues that psychodynamic theory dominated social work practice from the 1920s through to the 1960s. This history of whiteness, in the context of global imperialism, has white women leading the way in this work. Roger (2000) has argued that although psychotherapy was a form of resistance by white women to male patriarchy, this therapeutic approach still reinforced the regulation of all women and, worse, regulated people of color and Aboriginal peoples. In response, feminism, antiracism, and Aboriginal worldview perspectives have attempted to expose the hidden and not-so-evident forms of male patriarchy, racism, and colonialist acts that are play out within institutional practices.

White women, in particular, played an important role in this historical development of the social work profession. During the time of the COS movement, white middle-class women out of empathy wanted to help the 'deserving poor.' Their understanding of the plight of the marginalized has been critiqued as coming from a place of middle-class moralizing. Many of these bourgeois white women were also Christian missionaries who have been labeled by critical white studies scholars as do-gooders who wanted to help (Roger, 2000). Psychotherapeutic practice with people of color and Indigenous peoples could not adequately deal with issues of racism and colonialism and carried its own built-in whiteness. For example, why is it that child welfare frontline practice is dominated by new graduates who are young, white females (Fallon et al., 2003; Pon et al., 2011)? In doing so, we see the ways in which whiteness is secured in these hiring practices. Power granted to white females effectively positions them as the "'knowers' of the helping profession" (Roger, 2000: p. 138). Over time, white women will possess an investment in this type of practice making it much more difficult for them to undo their own whiteness. This contradictory stance and tension resides in the psyches and practices of white women.

Given these critiques about whiteness and white supremacy in the social work profession, more recent scholarship has also examined the possibility of exploring a post-race society. A post-racial society is one where people's experiences do not match their racial identity (Steyn and Conway, 2010). In other words, to essentialize race, including the racialization of white people, restrictively takes away from the ways in which strategies and processes of whiteness are fluid, malleable, and mobile. That various groups of people can benefit from whiteness and, in turn, bolster and support white supremacy shows the complexity of how power and dominance operates in a global world. For example, Shaw's (2006) work is aimed at decolonizing whiteness. Too often, whiteness studies are focused solely on 'white' ethnicity with an Anglo focus. Ironically, such metanarratives about white dominance may inadvertently recenter whiteness. As noted by Shaw (2006: p. 865), "whiteness is a process that not only racialised and excludes, but de-ethnicizes and *includes* otherwise ethnicized group, as required." Similarly, Twine and Gallagher (2007) argue that for us to understand the operation of whiteness, there needs to be more research done on the ways in which ethnic minorities are co-opted into strategies of whiteness that in the end, maintain white supremacy.

Conclusion

If we apply research and scholarship on how to end white supremacy, there may still be possibilities on how to create real change to the profession of social work. This requires, though, a deeper and theoretical understanding of the ways in which whiteness and supremacy undergirds the social work profession. Rabaka (2007) argues that white supremacy will not end until we allow for knowledge making that comes from people of color and Aboriginal peoples to inform the values of what is needed in social work. In doing so, this is not about culture, but about actual life worldviews that teaches everyone about alternative systems that are outside of a colonizing and white supremacist viewpoint. Currently, in the social work profession, more research is emerging about the importance of allowing the service users voice to be heard. Whether service users will be allowed to garner institutional power that actually changes the values and norms of the organization is yet to be seen. Because service users come from a diverse set of backgrounds along the sites of race, gender, class, and disability, we may yet be able to bring in fresh perspectives that challenge the status quo of whiteness and white supremacy in the profession of social work.

See also: Antisocial Behavior; Culturally Appropriate Interventions in Social Work; Ethnicity, Sociology of; Indigenous Social Work; Oppression; Racism, Sociology of; Social Work Theory.

Bibliography

Addams, J., 1922. Peace and Bread in Times of War. Macmillan, New York.
Ahmed, S., 2012. On Being Included: Racism and Diversity in Institutional Life. Duke University Press, Durham, NC.
Allen, T.W., 1994. The Invention of the White Race: Racial Oppression and Social Control. Verso, London.
Allen, R.L., 2001. The globalization of whites supremacy: toward a critical discourse on the racialization of the world. Educational Theory 51 (4), 467–485.
Arai, S., Kivel, B.D., 2009. Critical race theory and social justice perspectives on whiteness, difference(s) and (anti)racism: a fourth wave of race research in leisure studies. Journal of Leisure Research 41 (4), 459–469.
Aronson, J., Smith, K., 2011. Identity work and critical social service management: balancing on a tightrope? British Journal of Social Work 41 (3), 432–448.
Baines, D. (Ed.), 2011. Doing Anti-Oppressive Practice: Social Justice Social Work. Fernwood Publishing, Halifax, Nova Scotia, Canada.
Baldwin, A., 2012. Whiteness and futurity: towards a research agenda. Progress in Human Geography 36 (2), 172–187.
Berg, L., 2012. Geographies of identity I: geography – (neo)liberalism – white supremacy. Progress in Human Geography 36 (4), 508–517.
Bhattacharyya, G., Gabriel, J., Small, S., 2002. Race and Power: Global Racism in the Twenty-First Century. Routledge, London.
Bonnett, A., 2000. Whiteness in crisis. History Today 50 (12), 38–40.
Bradley, K., 2009. Poverty, Philanthropy and the State: Charities and the Working Classes in London, 1918–79. Manchester University Press, Manchester.

Carniol, B., 2005. Analysis of social location and change: practice implications. In: Hick, S., Pozzuto, R., Fook, J. (Eds.), Social Work: A Critical Turn. Thompson Educational Publishing, Toronto, pp. 153–166.

Carniol, B., 2010. Case Critical: Social Services and Social Justice in Canada, sixth ed. Between the Lines, Toronto.

Chand, A., 2000. The over-representation of Black children in the child protection system: possible causes, consequences and solutions. Child and Family Social Work 5 (1), 67–77.

Delfabbro, P., Hirte, Craig, Rogers, N., Wilson, B., 2010. The over-representation of young Aboriginal or Torres Strait Islander people in the South Australian child system: a longitudinal analysis. Children and Youth Services Review 32 (10), 1418–1425.

Delgado, R., Stefancic, J. (Eds.), 1997. Critical White Studies: Looking behind the Mirror. Temple University Press, Philadelphia.

Dettlaff, A.J., Rycraft, J.R., 2008. Deconstructing disproportionality: views from multiple community stakeholders. Child Welfare 87 (2), 37–58.

Dominelli, L., 2010. Social Work in a Globalizing World. Polity, Malden, MA.

Dubois, W.E.B., 1935. Black Reconstruction, 1860–1880. Harcourt, New York.

Dubois, W.E.B., 1969. The Souls of Black Folk. Signet Classic, New York.

Dubois, W.E.B., 1970. The Philadelphia Negro: A Social Study, third ed. Schocken Books, New York.

Dumbrill, G.C., 2010. Power and protection: the need for a child welfare service users' union or association. Australian Social Work 63 (2), 194–206.

D'Souza, N., 1994. Indigenous child welfare or institutionalised colonialism? Social Alternatives 13 (1), 32–34.

Fallon, B., MacLaurin, B., Trocmé, N., Felstiner, C., 2003. A national profile of child protection workers. In: Kufeldt, K., McKenzie, B. (Eds.), Child Welfare: Connecting Research, Policy, and Practice. Wilfrid Laurier Press, Waterloo, pp. 41–52.

Fellows, M.L., Razack, S., 1998. The race to innocence: confronting hierarchical relations among women. The Journal of Gender, Race and Justice 1, 335–353.

Fine, M. (Ed.), 1997. Off White: Readings on Race, Power, and Society. Routledge, New York.

Frankenberg, R. (Ed.), 1997. Displacing Whiteness: Essays in Social and Cultural Criticism. Duke University, Durham.

Gabriel, J., 1998. Whitewash: Racialized Politics and the Media. Routledge, London.

Gordon, L.R., 1997. Her majesty's other children: sketches of racism from a neocolonial age. Rowman and Littlefield, Lanham, MD.

Green, M., Sonn, C., Matsebula, J., 2007. Reviewing whiteness: theory, research and possibilities. South African Journal of Psychology 37 (3), 389–419.

Harding, J., Pribram, D., 2004. Losing our cool? Following Williams and Grossberg on emotions. Cultural Studies 18 (6), 863–883.

Hunter, S., 2010. What a white shame: race, gender, and white shame in the relational economy of primary health care organizations in England. Social Politics: International Studies in Gender, State, and Society 17 (4), 450–476.

Jacobson, M.F., 1998. Whiteness of a Different Color: European Immigrants and the Alchemy of Race. Harvard University Press, Cambridge.

Kincheloe, J.L., 1999. The struggle to define and reinvent whiteness: a pedagogical analysis. College Literature 26 (3), 162–194.

Kincheloe, J.L., Steinberg, S.R., 1998. Addressing the crisis of whiteness: reconfiguring white identity in a pedagogy of whiteness. In: Kincheloe, J.L., Steinberg, S.R., Rodriguez, N.M., Chennault, R.E. (Eds.), White Reign: Deploying Whiteness in America. Macmillan, Basingstoke, pp. 3–29.

Krane, J., Davies, L., 2000. Mothering and child protection practice: rethinking risk assessment. Child and Family Social Work 5 (1), 35–45.

Lorde, A., Rich, A., 1981. An interview with Audre Lorde. Signs 6 (4), 713–736.

Lundy, C., 2004. Historical developments in social work. In: Lundy, C. (Ed.), Social Work and Social, Justice. Broadview Press, Peterborough, pp. 19–47.

Morrison, T., 1922. Playing in the Dark: Whiteness and the Literary Imagination. Harvard University Press, Cambridge.

Myrdal, G., 1944. An American Dilemma. Harper and Brothers, New York.

Nelson, J., 2002. The space of Africville: Creating, regulating, and remembering the urban "slum". In: Razack, S. (Ed.), Race, space, and the law: Unmapping a white settler society. Toronto: Between the Lines, pp. 211–32.

Pon, G., Gosine, K., Phillips, D., 2011. Immediate response: addressing anti-native and anti-black racism in child welfare. International Journal of Child, Youth and Family Studies 2 (3/4), 385–409.

Rabaka, R., 2007. Soul of white folk: W.E.B. Du Bois's critique of white supremacy and contributions to critical white studies. Journal of African American Studies 11 (1), 1–15.

Richmond, M., 1899. Friendly Visiting among the Poor. Macmillan, New York.

Roediger, D.R., 1991. The Wages of Whiteness. Verso, New York.

Roger, K., 2000. "Making" white women through the privatization of education on health and well-being the context of psychotherapy. In: Calliste, A., Dei, G., with the assistance of Margaida, Aguiar (Eds.), Anti-racist Feminism: Critical Race and Gender Studies. Fernwood Publishing, Halifax, pp. 123–142.

Shaw, W.S., 2006. Decolonizing geographies of whiteness. Antipode 38 (4), 851–869.

Schick, C., 2012. White resentment in settler society. Race, Ethnicity and Education 1, 1–15 ifirstarticle.

Smith, M.K., 1999. University and Social Settlements, and Social Action Centres, the Encyclopaedia of Informal Education [Online] Available at: http://www.infed.org/mobi/university-and-social-settlements (accessed 11.08.13.).

Spade, D., 2013. Intersectional resistance and law reform. Signs 38 (4), 1031–1055. Intersectional.

Srivastava, S., 2005. 'You're calling me a racist?' The moral and emotional regulation of antiracism and feminism. Signs: Journal of Women in Culture and Society 31 (1), 29–62.

Steyn, M., Conway, D., 2010. Introduction: intersecting whiteness, interdisciplinary debates. Ethnicities 10 (3), 283–291.

Styhre, A., Eriksson-Zetterquist, U., 2008. Thinking the multiple in gender and diversity studies: examining the concept of intersectionality. Gender in Management: An International Journal 23 (8), 567–582.

Thobani, S., 2007. Exalted Subjects: Studies in the Making of Race and Nation in Canada. University of Toronto Press, Toronto.

Thompson, A., 2003. Tifany, friend people of color: white investments in antiracism. Qualitative Studies in Education 16 (1), 7–29.

Todd, S., 2011. That power and privilege thing: securing whiteness in community work. Journal of Progressive Human Services 22 (2), 117–134.

Trocmé, N., Knoke, D., Blackstock, C., 2004. Pathways to the overrepresentation of Aboriginal children in Canada's child welfare system. Social Service Review 78 (4), 577–600.

Twine, F.W., Gallagher, C., 2007. Introduction: the future of whiteness: a map of the 'third wave'. Ethnic and Racial Studies 31 (1), 4–24.

Wagner, A., 2010. Crisis…What crisis? The revitalization of academic feminism. Canadian Social Work Review 27 (2), 187–204.

Warren, J.T., 1999. Whiteness and cultural theory: perspectives on research and education. The Urban Review 31 (2), 185–200.

Wiegman, R., 1999. Whiteness studies and the paradox of particularity. Boundary 26 (3), 115–150.

Yee, J.Y., 2005. Critical anti-racism praxis: the concept of whiteness implicated. In: Hick, S., Pozzuto, R., Fook, J. (Eds.), Social Work: A Critical Turn. Thompson Educational Publishing, Toronto, pp. 87–104.

Yee, J.Y., Dumbrill, G., 2003. Whiteout: looking for race in Canadian social work practice. In: Al-Krenawi, A., Graham, J. (Eds.), Multicultural Social Work in Canada: Working with Diverse Ethno-racial Communities. Oxford University Press, Toronto, pp. 98–121.

Yee, J.Y., Wagner, A., 2013. Is anti-oppression teaching in Canadian schools of social work a form of neo-liberalism? Social Work Education 32 (3), 331–348.

Whorf, Benjamin Lee (1897–1941)

Joseph L Subbiondo, California Institute of Integral Studies, San Francisco, CA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Benjamin Lee Whorf (1897–1941) advanced a theory of linguistic relativity in which he maintained that one's language substantially affected one's worldview. His 'principle of linguistic relativity,' as he referred to it, has often been identified as the Sapir–Whorf Hypothesis because Edward Sapir (1884–1939) inspired Whorf's development of the concept. In addition to his version of linguistic relativity, Whorf introduced the allophone that enhanced the study of phonology and he formulated covert grammatical categories, which explained the role of grammar in modeling reality. In his initiatives, Whorf drew upon his studies of the Hopi and Maya languages and cultures; asserted a close interrelationship among language, culture, and thought; integrated wisdom and esoteric traditions; challenged core tenets of Western science; and incorporated interdisciplinary research into linguistics. In critiquing many traditional linguistic and scientific assumptions of his era, Whorf's writings foreshadowed much current research in cultural studies, cognitive psychology, anthropological linguistics, and consciousness studies.

The most detailed biography of Whorf was written by John B. Carroll as the introduction to his book collection of Whorf's writings (Whorf, 1956). Whorf was born on 24 April 1897 in Winthrop, Massachusetts, a suburb of Boston. He was the eldest of three sons born into a prominent family who descended from the first wave of English settlers to the American colonies. In his primary and secondary education, Whorf developed a lifelong interest in science, particularly in chemistry; and in 1918, he earned a Bachelor of Science degree in Chemical Engineering at the Massachusetts Institute of Technology (M.I.T.). In 1920, Whorf and Celia Inez Peckham married and they had three children.

A year after his graduation from M.I.T., Whorf began his career in fire prevention engineering at the Hartford Fire Insurance Company (HFIC). He was among the first generation of fire safety inspectors in the insurance industry, and his fire prevention consultation was highly sought after by a wide variety of manufacturing companies. Whorf's pioneering specialty and acknowledged expertise not only brought him deserved professional recognition in the growing field of fire prevention, but it also attracted many clients to the HFIC. Because Whorf contributed substantially to its business growth, the HFIC retained his services by regularly granting him long- and short-term leaves to pursue his scholarly research in linguistics. Whorf apparently valued this unconventional career arrangement with HFIC because he never left the company to accept an academic appointment, and he remained in the employ of HFIC until his death in 1941. Whorf's professional standing as a chemical engineer as well as his education in the sciences provided him with a capacity for scientific inquiry and analysis not typically found in a linguist. By drawing on his science vocabulary, he consistently introduced linguists to scientific words and metaphors as he described and analyzed the complex systems and structures of language.

Whorf met Sapir in 1928, and they became research collaborators in 1931 when Sapir left the University of Chicago to join the faculty at Yale University, which is located in Whorf's hometown of Hartford, Connecticut. Whorf enrolled in the first course in American Indian linguistics taught by Sapir at Yale, and he was inspired by Sapir to study Hopi, a language related to Aztec (Darnell, 1990). Although Whorf studied with Sapir, the two never collaborated on a presentation or a publication on linguistic relativity. Thus, the term Sapir–Whorf Hypothesis is a misnomer. In fact, while Sapir was interested in the relationship among language, thought, and culture as was his teacher Franz Boas (1858–1942), he was not nearly as committed to the notion of linguistic relativity as Whorf. For Whorf, his extensive research on linguistic relativity was clearly the core passion of his scholarly life.

Whorf engaged in extensive anthropological research by working in the renowned Native American collection at the Hartford Public Library. He renewed his interest in Mexican languages, especially Aztec and Maya. In 1929, he became a member of the Linguistic Society of America (LSA); and in the first year of his membership, he presented a paper on Maya lexical stems at the annual meeting of the LSA. Soon after this presentation, he received his first research grant to study the Mayan ruins in Mexico.

With the exception of Sapir's courses at Yale, Whorf was self-taught in linguistics. While he was never appointed to a faculty position, Whorf was not an amateur scholar in the traditional sense: he lectured frequently at Yale and other universities, taught Sapir's course at Yale when Sapir was on sabbatical leave, participated in a broad range of academic circles and organizations, held distinguished honorary appointments and offices, received credible academic grants, and presented and published scholarly papers and reports. As evidence of Whorf's prominence in the linguistic community, George Trager, eminent professor at Yale as well as Whorf's collaborator and friend, wrote an obituary for Whorf in *Language*, the journal of the LSA.

Whorf's research focused on two general areas: the Hopi and Mayan cultures and languages, and the integration of language, thought, and culture. For the most part, he connected the two areas because he usually drew on his research on Hopi and Maya to exemplify aspects of his theory of linguistic relativity. As for the former area, the Hopi and Maya languages provided him with many illustrative examples of non-Western world views that supported his persistent critique of the assumptions of Western science. As for the later area, Whorf argued that the interplay of language, thought, and culture enabled the process of mental growth in the human development. He regarded

mental growth as essential for the quality of cognitive development required to solve human problems, both present and future. Whorf insisted that the survival of the human species as well as that of the environment was totally dependent on the evolution of thinking and cognition.

In 'An American Indian Model of the Universe,' Whorf (1956) provided an example to explain the bond between language and thought in his theory of linguistic relativity: he pointed out that the Hopi language, unlike Western languages, did not have traditional verb tenses. He used this distinctive grammatical feature of the Hopi language to develop his case against the notion of linguistic and/or cultural universals. Whorf indicated that Hopi speakers who knew only the Hopi language and the cultural ideas of their society did not have the same concepts of time as those held by speakers of English. Specifically, Whorf stated that a Hopi speaker did not have an understanding of time as a smooth flowing continuum in which everything in the universe proceeded at an equal rate from past to present to future, or in reverse order. While his example of verb tenses has often been disputed, even ridiculed, by various linguists, it has not been disproven.

Whorf noted that Hopi did not contain words, phrases constructions, or verbal expressions that directly referred to the concept of time as well as to the notions of past, present, or future. Rather than viewing Hopi speakers as trapped in flawed metaphysics and speaking a deficient language, Whorf argued that the Hopi had a consciousness of time and space that was simply different than that of a speaker of English. Based on this conclusion, Whorf claimed that the Hopi viewed the universe through a different lens than speakers of other languages. This observation supported Whorf's thesis that there were ways of perceiving the universe other than those enshrined in traditional Western science. Because Whorf believed that worldviews were determined culturally and linguistically, he argued that one could understand the worldviews of others only by opening oneself to other cultures and languages. He asserted that the structure of English, like that of all languages, layered its distinct metaphysical systems on the world.

In 'A linguistic Consideration of Thinking in Primitive Communities,' Whorf (1956) explained his principle of linguistic relativity by focusing on what he believed were the cognitive and rational processes that affected thinking. He indicated that anthropological linguists were the most knowledgeable scholars of these processes because in their research they explored those aspects that were primarily related to language and culture. In particular, he cited Boas, Sapir, and Leonard Bloomfield (1887–1949) – scholars whom he considered among the leaders in this emerging area of research.

In addition to the work of the anthropological linguists, Whorf recognized the research of those in the field of psychology in general and the work of Carl Jung (1875–1961), in particular. Whorf claimed that psychologists contributed significantly to advancing his theory of linguistic relativity. In particular, he was drawn to Jung who maintained that of the four psychic functions; two were rational (thinking and feeling) and two irrational (sensation and intuition). Throughout his writings, Whorf referred to Jung's psychic functions to support his idea that the study of language was the most appropriate path leading to a thorough and reliable understanding of thinking.

Whorf believed that linguists were primarily driven by their persistent search for meaning, and he believed that the discipline would eventually draw upon more scholarship from psychology and cultural studies. He emphasized that importing data from other disciplines into linguistics would add, not detract, from its scientific status. He was strongly convinced that linguistics would always retain its scientific precision in its analysis of language structure. He envisioned a day when linguists would eventually become aware of the many psychological undercurrents that have made an impact on individuals and groups throughout history. To support his thesis, Whorf presented a brief history of linguistics in which he emphasized the psycholinguistic patterning of language as an essential part of culture. He contended that the study of language from the ancients to the twentieth century could be characterized as developing merely formal and overt structures. He indicated that a debt of gratitude was owed Boas because in his *Handbook of American Indian Languages* (1911), he was the first linguist to analyze a language on its own terms and not force a traditional and classical model on it.

In 'A Linguistic Consideration of Thinking,' Whorf (1956) clearly expressed his notion that human thinking, like other human capacities, was continuing to evolve. He noted that an evolution of thinking was necessary in order for the human species to preserve itself as well as its environment. He argued that a history of mental growth could be found in the history of each language because languages have their own systems for discerning, selecting, and thinking. Similar to Owen Barfield (1898–1997) who in his *History in English Words* (1926) stated that an evolution of consciousness could be found in the history of the lexicon of a language (Subbiondo, 2003), Whorf vigorously defended the notion that language was the ultimate inroad to understanding an evolution of thinking, and he compared the status of research of mental growth in his day to that of pre-Linnaean botany.

In order to advance his theory of linguistic relativity, Whorf developed a case for a new scientific paradigm in 'Language, Mind, and Reality' (Whorf, 1956). Whorf's paradigm was based on an acceptance of many and varied multiple logical systems underling the grammatical structures of the world's languages. Whorf's principal criticism of Western science was not based on its conclusions but on its methodological analyses of the universe. He was convinced that Western methods of analyses were severely limited by the primary assumption of Western scientists that there was a universal logic. Whorf declared that the way for Western scientists to transcend their notion of a universal logic was that they needed to recognize the existence and validity of alternative logical systems. He held that if one could accept alternative logical systems, one could appreciate other worldviews, which could provide fresh insights into all aspects of reality.

In 'Science and Linguistics,' Whorf (1956) argued that because each language had its own logical system, there were as many logical systems as there were languages. He pointed out that in their studies of the background phenomena of language, comparative grammarians discovered that the notion of language universals was false because no feature was common to all languages. He maintained that the grammatical structure and the lexicon of each language did more than shape the expression of ideas, it shaped the content of ideas as well. For Whorf, ideas were dependent upon the language in which they were conceived and expressed.

Whorf documented the fact that comparative grammarians in studying nonIndo-European languages (he cited Semitic,

Chinese, Tibetan, and Hopi) discovered that each language had its distinct worldview. He illustrated his point by specifically referring to Hopi to show that measurements of time and velocity were different to a Hopi speaker than to a speaker of Western languages. Moreover, he cautioned that by privileging the worldview of speakers of Indo-European languages, one was encouraged to overlook the unique panorama of worldviews embedded in other languages. Moreover, Whorf held that by privileging Western languages as the sole model for all languages contributed to what he perceived to be sheer intellectual arrogance that prevented people from being receptive, authentically curious, and sufficiently detached toward expanding their perceptions and thoughts.

In 'Linguistics as an Exact Science,' Whorf (1956) argued that science could make measurable progress, if it would focus not on changing the facts but on changing how one thought about the facts. He specifically cited the findings of Bloomfield who asserted that language played an indispensable role in advancing scientific progress. Bloomfield explained that scientific research started with a set of sentences that recorded specific observations. After extensive experimentation, the findings would become part of the scientific record when they were expressed in another set of sentences, which then served as the basis for further research through more experimentation. Like Bloomfield, Whorf understood that there could be neither science nor scientific progress without language.

Whorf concluded that because linguistics was an experimental science, it fully deserved the academic prestige and financial support that was provided by physics and chemistry. He noted that similar to their colleagues working in the other sciences, linguists collected data, experimented with human subjects, and formulated verifiable conclusions. Moreover, he maintained that the subject matter of linguistics, language, shaped as well as expressed the totality of human agreement and learning.

In 'Language and Logic,' Whorf (1956) focused his attention on the nature and structure of alternative logical systems in various language and their role in shaping diverse worldviews. He argued that differences in languages were reflected in different interpretations of facts. Whorf believed that one understood facts according to the worldviews embedded in one's language. Drawing on his knowledge of chemical engineering, Whorf compared languages to the two types of mixtures of material substances: mechanical and chemical. He pointed out that a mechanical mixture of a material consisted of adding any ingredient that did not change the appearance of the material. A chemical mixture, on the other hand, was made up of mutually suited materials that overtly changed the appearance of the material. After considerable discussion, Whorf classified English as a mechanical language and Shawnee as a chemical language. He pointed out that the logical system of English like its Indo-European relative, Greek was based on a mechanical mixture; and he facetiously added that its logical system would have been chemical if Aristotle were a Shawnee.

Whorf believed that reality in its fullest dimensions could not possibly be explained by the logical system of any one particular language, no matter how expansive its logical system. He maintained that the generations of Western cultures have expressed no more than provisional analyses of reality even though each generation had resolutely determined that its analysis was final. Whorf argued that plausible correctives to this 'final' analysis were in the various non-Western languages and these languages had also evolved over the centuries independent of Western languages. As a consequence, the speakers of non-Western languages arrived at a different, but equally logical, analysis of reality. Whorf maintained that the principal benefit of understanding the alternative logical systems of non-Western languages was that they could advance scientific inquiry. It was possible, in Whorf's view, that non-Western languages could lead to plausible understandings of complex scientific issues that would be inexpressible in Western languages. Whorf stressed repeatedly throughout his writings that languages differed not only in their constructions of sentences but also in the imaginable content of their sentences. In Whorf's time as today, there was a strong allegiance to the infallibility of universal logical reasoning; and any suggestion to the contrary, as was clearly the case for Whorf's principle of linguistic relativity, was often met with fierce and dismissive resistance.

In 'Language, Mind, and Reality,' Whorf wrote what has been regarded by many as his most controversial essay because in it he proposed an overarching view of linguistics that included references to Indic philosophy and spirituality that were contained in esoteric wisdom texts. The essay was published a year after his death in the *Theosophist*, the publication of the Theosophical Society. The Society was initially formed in New York City in 1875 by Helena Blavatsky among others to explore mystical teachings and practices. After a few years, the Society moved its headquarters to Chennai (formerly Madras), India. In time, the Theosophists studied a full range of Eastern religions and forms of spirituality, especially Hindu and Buddhist teachings and practices. The Society gradually formed a global community that was open to all people regardless of race, creed, gender, caste, and color. In light of the social mores of the time and the prevailing legal support for discrimination, the Theosophists were a group offering a radical social agenda of complete inclusion. In addition to the study of world religions, the group also encouraged the study of philosophy and science.

The breadth and depth of Whorf's lifelong reading of esoteric texts influenced his thinking about language and shaped his theory of linguistic relativity far more than scholars have typically acknowledged (Joseph, 2002). As a young man, Whorf discovered the work of Antoine Fabre d'Olivet (1768–1825), an obscure French philologist and mystic who taught that the root ideas of ancient Judaic spiritual thought were embedded in Hebrew letters. Whorf was fascinated by the work of d'Olivet because he was very interested in biblical studies. It can be safely argued that Whorf's inspiration to study language was religious. Whorf was convinced that d'Olivet's method of biblical study was similar to the structural approach that dominated linguistics in his day. Whorf's reading of d'Olivet prompted him to study Hebrew in order to better appreciate and evaluate the legitimacy of d'Olivet's claims.

The archival collection at Yale contains extensive materials that document Whorf's lifelong interest in the work of the Theosophists in particular and in spiritual and religious traditions and practices in general. Whorf's study of esoteric wisdom was aligned with his lifelong personal and professional commitment to understanding multiple logical systems. In

light of the emerging interest in Eastern thought currently, Whorf's considerable attraction to the study of mysticism is more acceptable to scholars today than it was in his own day. The fact that this aspect of Whorf's research was not mentioned in any detail by Carroll in his biography of Whorf strongly suggests that his supporters were understandably concerned that Whorf and his work might not be taken seriously if his connection to the Theosophists were common knowledge. Scholars in this century warn that it would be a mistake to treat Whorf's discussion of Eastern ideas superficially – they had a significant impact on his work (Lee, 1996).

Drawing on the notion of the Indic mantra as a formulaic expression in language, Whorf argued that language structures and grammars represented cosmic patterns and arrangements. Whorf was well aware that this perspective drawn from the Yoga tradition was new to Western scientists and would probably be rejected outright; nonetheless, he insisted that the notion of the mantra was based on unimpeachable evidence. He pointed out that an individual's thoughts were controlled by unconscious laws of patterning, which affected intricate logical systems. These logical systems channeled reasoning and built the structure of consciousness in each person as well as in every culture. While Whorf recognized differences in logical analyses spawned differences in worldviews, he insisted that all humans had equal analytical capacity.

In this final article, Whorf summed up the central theme of the essays written at the end of his life: namely, the study of language was critical to the study of thinking and cognition. Whorf argued that there was an urgent need for linguists not only to learn diverse languages but also to understand the diverse worldviews that flowed from these languages. He claimed that only through clearly unbiased research of language, one could discover in non-Western languages profound, effective, and scientific expressions unknown and unthinkable to those who speak Western languages.

From the 1950s to the 1990s, the transformational generative grammar of Noam Chomsky with its emphasis on language universal dominated linguistic theory in the United States and elsewhere. Because Whorf's theory of linguistic relativity was antithetical to the generative grammarian's theory of language universals, his work was disputed if not dismissed by leading generative linguists (Pinker: 1994). In fairness one could argue that Whorf was often read out of context and quoted in isolated statements. Nonetheless after Whorf's death, most anthropological linguists as well as psycholinguists often referred to Whorf's work as their starting point and either supported or rejected in part or in whole his principle of linguistic relativity (Lucy, 1992). With the current interest in consciousness studies and the focus on language in studying the nature and structure of consciousness, Whorf's work has become of interest once again to many contemporary scholars (Gumperz and Levinson, 1996).

Although Whorf engaged himself in an ambitious conceptual enterprise, his theories and lines of reasoning have become more accessible to scholars in recent times than when they were originally developed (Lee, 1996). For example, it is common knowledge today that people in non-Western cultures view the world differently than do those in Western cultures. While how much language affects culture may be debatable, most scholars would agree that culture shapes worldviews. This conclusion which was strongly argued by Whorf has come about mainly due to an extraordinary increase of late in studies of non-Western cultures by Western scholars and drawing from them fresh perspectives that can be applied in many areas of research (Nisbett, 2003).

Whorf's writings foreshadow the current trend in the academy to integrate interdisciplinary research into mainstream scholarship. References to Whorf's work can be found in disciplines outside linguistics including philosophy, psychology, and cultural studies. In addition to interdiciplinary research, there are many more disciplinary blends today than there were in Whorf's time: for example, psycholinguistics (a term first used by Whorf). The accelerating increase in interdisciplinary studies in the present century has contributed in large measure to advances that could not have been made possible or even imagined within the confines of traditional academic disciplines. For example, the study of language and consciousness could not have flourished without the convergence of biology, linguistics, neuroscience, philosophy, and psychology (Wardy, 2000). In many aspects, the intedisciplinarity of current scholarly research reflects a breadth of integrated learning that Whorf envisioned nearly a century ago.

See also: Bloomfield, Leonard (1887–1949); Boas, Franz (1858–1942); Chomsky, Noam (1928–); Gender and Language: Cultural Concerns; Historical Thought and Historiography: Current Trends; Historical Thought and Historiography: Modern History (Since the 18th Century) – the West; Intellectual History; Jung, Carl Gustav (1875–1961); Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects; Meaning and Rule Following; Multiculturalism; Psychohistory; Religion, History of; Sapir, Edward (1884–1939); Science, History of; Self: Philosophical Aspects; von Humboldt, Wilhelm (1767–1835).

Bibliography

Boas, F., 1911. Handbook of American Indian Languages. Government Printing Office, Washington, DC.

Darnell, R., 1990. Edward Sapir: Linguist, Anthropologist, Humanist. University of California Press, Berkeley.

Gentner, D., Goldin-Meadow, S. (Eds.), 2003. Language in Mind: Advances in the Study of Language and Thought. the MIT Press, Cambridge, MA.

Gumperz, J., Levinson, S. (Eds.), 1996. Rethinking Linguistic Relativity. Cambridge University Press, Cambridge.

Joseph, J., 2002. From Whitney to Chomsky: Essays in the History of American Linguistics. John Benjamins, Amsterdam.

Lee, P., 1996. The Whorf Theory Complex: A Critical Reconstruction. John Benjamins, Amsterdam.

Lucy, J., 1992. Language Diversity and Thought: A Reormulation of the Linguistic Relativity Hypothesis. Cambridge University Press, Cambridge.

Nisbett, R., 2003. The Geography of Thought. The Free Press, New York.

Subbiondo, J., 2003. Lexicography and the evolution of consciousness: a study of Owen Barfield's history in English words. Historiographia Linguistica 30, 407–427.

Subbiondo, J., 2005. Benjamin Lee Whorf's theory of language, culture, and consciousness: a critique of western science. Language and Communications 25, 149–159.

Wardy, R., 2000. Aristotle in China: Language, Categories, and Translation. Cambridge University Press, Cambridge.

Whorf, B., 1956. The Relation of Habitual Thought and Behaviour to Language. In: Carroll, John B. (Ed.), Language, Thought, and Reality: Selected Writings by Benjamin Lee Whorf. MIT Press, Cambridge, MA, pp. 134–159.

Williams Syndrome

Joanne S Camp, Goldsmiths, University of London, London, UK
Annette Karmiloff-Smith, Centre for Brain and Cognitive Development, School of Science, Birkbeck, University of London, London, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by A. Karmiloff-Smith, volume 24, pp. 16500–16504, © 2001, Elsevier Ltd.

Abstract

Williams syndrome (WS) is a rare neurodevelopmental disorder, caused by a hemizygous deletion of 26–28 genes on chromosome 7. It is of particular interest to neuroscience because of its uneven cognitive profile and the hope of mapping genotype/phenotype relations. The article describes important aspects of the WS neural and sociocognitive profile from infancy to adulthood.

In the final decades of the twentieth century, cognitive neuroscience paid particular attention to a group of disorders known as microdeletion syndromes. These are neurodevelopmental disorders in which the missing genes on a chromosome can be identified and the resulting gene–behavior relations explored. They differ from syndromes like autism and dyslexia, which are likely to involve a very large number of genes of small effect. One microdeletion disorder that caused particular excitement among cognitive neuroscientists is Williams syndrome (WS). This is because WS results in an uneven cognitive profile. Language and face processing are seemingly spared, whereas other higher cognitive functions (e.g., spatial cognition, number, planning, and problem solving) are seriously impaired. Initial characterizations of the syndrome at the cognitive level seemed to hold the promise of relatively straightforward gene-cognition mappings. WS was, and continues to be by some researchers, hailed as the prime example of intact, independently functioning cognitive modules in the face of general intellectual impairment (e.g., Bellugi et al., 2000; Landau et al., 2006; Musolino et al., 2010; Pinker, 1994; Tager-Flusberg et al., 2003). However, more recent evidence challenges this view, favoring instead a more dynamic, neuroconstructivist approach to genetic disorders (Karmiloff-Smith et al., 2012a). In this article, data on the neural and cognitive processes underlying the purportedly spared domains will be examined, together with a focus on the infant phenotype and the developmental trajectory leading to the phenotypic outcome.

Williams Syndrome: Genetic Profile

Williams syndrome is a rare neurodevelopmental disorder occurring in approximately 1 in 20 000 live births (Pober, 2010; although see Strømme et al., 2002 for an estimate of higher prevalence), involving a hemizygous deletion of 26–28 genes on chromosome 7q11.23 (Pober, 2010). Clinical symptoms are wide-ranging, including dysmorphic facial features, supravalvular aortic stenosis (SVAS), dental abnormalities, gastrointestinal and endocrine problems, and hypersensitivity to sound (Pober, 2010) as well as an uneven cognitive profile (Mervis and John, 2010). The gene for elastin (*ELN*) within the deleted region (Ewart et al., 1993) remains the only clear-cut genotype–phenotype mapping, contributing to SVAS and general connective tissue problems (Pober, 2010). Another of the deleted genes, Limkinase1 (*LIMK1*), is expressed in the brain and was initially claimed to be implicated in spatial cognition deficits (Frangiskakis et al., 1996). However, Gray and colleagues assessed two partial deletion patients (only *ELN* and *LIMK1* deleted) on a range of spatial abilities, comparing them to two participants with the full WS deletion who presented with strong verbal abilities. The individuals with WS continued to display the characteristic WS profile of weak spatial skills, while the partial deletion patients did not, leading to the conclusion that the *LIMK1* gene is either not implicated in spatial cognition, or if involved, it is through interaction with other genes in the Williams Syndrome Critical Region (WSCR) (Gray et al., 2006). Currently, four telomeric genes in the WSCR, all transcription factors, are thought likely to be the most critical contributors to the WS sociocognitive phenotype (Karmiloff-Smith et al., 2012b), involving complex cascading effects. Clearly, the way in which genes in the deleted region impact upon the eventual phenotype is complex and dynamic, involving widespread, low-level and interactive effects of gene expression over time (Karmiloff-Smith, 1998), rather than a direct one-to-one genotype–phenotype mapping.

Williams Syndrome: Phenotypic Profile

The pioneering work of Bellugi and collaborators initially pinpointed some clear-cut dissociations in the cognitive architecture of WS. Language and face processing appeared to be preserved in the face of both general retardation and particularly serious problems with visuospatial cognition. These apparent dissociations continue to be highlighted by some authors (Bellugi et al., 2000; Tager-Flusberg et al., 2003), while most researchers in the field of WS have been fairly cautious about their claims, referring to relative strengths and weaknesses rather than absolute ones (Karmiloff-Smith, 1998; Mervis and John, 2010). By contrast, secondary sources cited in writings by linguists, developmental psychologists, neuropsychologists of adult brain damage, and philosophers have often used WS to bolster claims about innately specified,

independently functioning modules, some of which are purportedly intact and others impaired (e.g., Clahsen and Almazan, 1998; Pinker, 1994). This stems from the view that the pattern of behavioral performance found in the phenotypic outcome is a direct window on the innate, modular structure of the brain's cognitive architecture (Baron-Cohen, 1998; Temple, 1997).

Advances in brain imaging technology have facilitated examination of the structure and function of the WS brain, revealing that it is qualitatively different from normal brains (see Karmiloff-Smith, 2010 for discussion) in terms of brain anatomy (Bellugi et al., 2000), brain biochemistry (Meyer-Lindenberg et al., 2006; Rae et al., 1998), and computational processing (Mills et al., 2000). WS brains show reduced brain stem and overall brain volumes, with normal proportional cerebellar and superior temporal region volumes (Reiss et al., 2004). The complexity of cortical folds is increased in WS relative to controls. Abnormalities of the corpus callosum and hippocampus (Meyer-Lindenberg et al., 2006) have also been reported. Attempts have also been made to link specific brain abnormalities to particular strengths and deficits in the WS sociocognitive phenotype. For example, Meyer-Lindenberg et al. (2006) aimed to identify separable neural subsystems for visuospatial cognition, social behavior and memory, which they deemed to be under genetic control. The approach taken in this and other studies (e.g., Germine et al., 2011; Musolino et al., 2010) tends to treat the atypically developing brain as if it were a normal adult brain with parts intact and parts impaired, ignoring the developmental history of WS in terms of the dynamic role of genetic mutations in interaction with environmental input in fostering overall brain growth over developmental time. Interestingly, although most research on WS brains has focused on adults and adolescents, a recent MRI study of older children with WS found increased complexity as well as reduced surface area and cortical volumes compared to healthy controls (Fahim et al., 2012). Studies of the early stages of WS brain development are rare (but see Boddaert et al., 2006), yet they are critical if we are to understand the relationship between the WS developing brain and the cognitive phenotype, because both progressively change over developmental time (Karmiloff-Smith, 2010).

We now examine language and the social domains for which some authors have claimed that the syndrome eloquently illustrates selectively intact processing alongside general intellectual impairment. However, even in cases where the phenotypic outcome in adults seems to display intact performance, in-depth analyses now suggest that individuals with WS process inputs via different neural and cognitive processes compared to healthy controls (Karmiloff-Smith, 1998; Karmiloff-Smith et al., 2004).

Language Development

Notable is the fact that although WS language ends up relatively fluent in the adult phenotype, the time course of WS language acquisition is significantly delayed (Mervis and John, 2012). Although at 2 years of age WS expressive vocabulary size is larger than that of chronological age (CA)-matched toddlers with Down syndrome (DS) (Mervis and John, 2012), the onset of vocabulary in toddlers with WS is seriously delayed relative to norms on the Communicative Development Inventory (Fenson et al., 1993). Moreover, infant precursors to language, such as babbling, are delayed in WS (Masataka, 2001). Finally, early WS vocabulary production precedes referential pointing, the opposite to the normal order of acquisition (Laing et al., 2002; Mervis and John, 2012).

Despite these early delays, it is notable that older children and adults with WS perform well on tests of receptive vocabulary, such as the Peabody Picture Vocabulary Test – Revised (PPVT-R; Dunn and Dunn, 1981). Indeed, PPVT-R scores in WS are generally above the level expected for their mental age (MA) (e.g., Brock, 2007). However, while better than expected for MA and strength relative to other areas of language, WS vocabulary falls well below CA (Mervis and John, 2012). Moreover, vocabulary in WS is not simply delayed, but also atypical. Language with a spatial component poses particular difficulties for individuals with WS (Phillips et al., 2004), while for young children with WS, more general conceptual/relational vocabulary knowledge (rather than specifically spatial knowledge) is weaker than their concrete receptive vocabulary (Mervis and John, 2012). Several studies now suggest that there may be an imbalance between phonology and semantics in WS (Karmiloff-Smith, 1998; Thomas and Karmiloff-Smith, 2003). For example, when WS participants monitor sentences for a target word, they do not show sensitivity to certain sentential violations, suggesting that semantic information may become available too slowly to be integrated with the online processing of syntax (Tyler et al., 1997). A study of reading in WS came to similar conclusions about the weak role of semantics (Laing et al., 2001). The control group found concrete, imageable words much easier to read than abstract words, but the WS participants failed to benefit from the imageability of the concrete words, displaying equal levels of (poorer) reading for both concrete and abstract words. In general, imageability effects have been shown to be weaker in people with WS than in typical development. A study by Vicari and his colleagues demonstrated that, compared to normal controls, word learning is superior in WS if the auditory presentation of a word is accompanied by the simultaneous presentation of a photograph depicting the object (Vicari and Carlesimo, 2006), whereas healthy controls easily create images mentally when they hear words. There is also evidence for lower levels of audiovisual integration in WS than typical development (TD) (Böhning et al., 2002). Taken together, a variety of studies suggest that, unlike typical development, phonological representations weigh more strongly than semantic representations in WS language acquisition over developmental time. In the domain of syntax, WS adults tend to display patterns typical of much younger children (Mervis and John, 2012). Earlier it had been claimed (Clahsen and Almazan, 1998) that individuals with WS have intact regular past tense formation of verbs, alongside impaired associative lexical processes. More recent research challenges this claim, in that once verbal MA is taken into account, no selective deficit for irregular verbs emerges in WS (Brock, 2007; Thomas and Karmiloff-Smith, 2003).

Despite the growing database on WS language, some continue to claim that WS syntax is intact (e.g., Musolino et al., 2010; Pinker, 1994). There are, however, a number of lines of evidence to challenge this. First, when WS vocabulary and syntax are compared to norms for healthy individuals, WS

vocabulary levels are usually better than their syntactic levels on various standardized tasks, suggesting an imbalance between the two (Brock, 2007; Mervis and John, 2012; Vicari and Carlesimo, 2006). Second, even in very simple imitation tasks, participants with WS show impairments with complex syntactic constructions such as embedded relative clauses (Grant et al., 2002; Vicari and Carlesimo, 2006). Moreover, even in an area of relatively simple syntax – grammatical concord – which normal French-speaking children acquire easily and very early in development, individuals with WS show deficits even in adulthood. Although the WS participants learned the local gender marker (correct article) for a nonce term easily (in fact, more easily than controls), their capacity for gender agreement across sentence elements, such as grammatical concord on adjectives or pronouns, was seriously impaired (Grant et al., 2002). Studies of Italian-speaking children (Volterra et al., 2003) have also revealed that grammatical gender presents a particular problem, with WS children displaying errors that are never encountered in normal development at any age. Finally, several studies (e.g., Karmiloff-Smith et al., 1997; Mervis and John, 2012; Vicari and Carlesimo, 2006; Volterra et al., 2003) suggest that the problems that people with WS have with semantics and syntax are often camouflaged by their good verbal memory.

These various findings are consistent with the hypothesis that the WS language system is not only seriously delayed, but also atypical, because it has developed under different constraints (Thomas and Karmiloff-Smith, 2003). In sum, the relatively impressive WS language found in adulthood may well be acquired through alternative, compensatory routes involving different neural processes.

Social Development

Despite reports that WS face processing is intact (Bellugi et al., 2000; Tager-Flusberg et al., 2003), in-depth studies of face processing suggest that individuals with WS use different strategies from normal controls (Karmiloff-Smith et al., 2004). Several studies (Deruelle et al., 2003; Karmiloff-Smith, 1998; Karmiloff-Smith et al., 2004) have replicated earlier work, revealing normal or near normal behavioral scores on standardized tasks like the Benton Facial Recognition Test (Benton et al., 1983) and the Rivermead Behavioural Memory Test (Wilson et al., 1985). However, experimental approaches alongside the standardized tasks have seriously challenged the notion that the WS behavioral success in face processing tasks is subserved by normal cognitive and neural processes. Where normal controls use predominantly configural strategies for processing upright faces, and featural strategies for processing inverted faces, individuals with WS tend to process both upright and inverted faces featurally. While subsequent studies have found good scores on holistic face processing in WS (Tager-Flusberg et al., 2003), a distinction should be drawn between holistic face processing and configural face processing: the former involves the gluing together of facial features (and hairline) into a gestalt (first-order information), without necessarily computing the spatial distances between features, whereas the latter involves the ability to differentiate faces based on sensitivity to the spatial distances among internal features, i.e., second-order relational information (Karmiloff-Smith et al., 2004). While there is indeed evidence of typical-looking holistic face processing in WS, albeit only in adulthood (Deruelle et al., 2003; Tager-Flusberg et al., 2003), a series of experiments demonstrated that configural face processing is difficult for individuals with WS and that it does not develop with age in the same way as for controls (Karmiloff-Smith et al., 2004). This is also supported by evidence from brain imaging studies. When electrophysiological measurements were taken while older children and adults with WS processed faces in an event-related potential study, they tended to process them bilaterally or predominantly with the left hemisphere, whereas healthy controls showed a right hemisphere bias for faces (Mills et al., 2000). Furthermore, an EEG study of face processing by Grice et al. (2001) yielded patterns of gamma-band activity in individuals with WS that were very different from controls. While the controls showed regular gamma bursts tied to stimulus onset, gamma-band activity in WS was almost flat. This suggests that binding of perceptual features in WS is impaired, thus contributing to their difficulties with configural processing. All of these findings point to atypical face processing in WS, despite the normal behavioral scores on standardized tests.

Face processing is not the only area of the social domain that is atypical in WS. Individuals with WS are often described as having a characteristic social, over-friendly personality, i.e., a 'hypersocial' profile at all ages from infancy to adulthood (Doyle et al., 2004; Jones et al., 2000). This adversely affects WS adults' rating of stranger approachability (Jones et al., 2000; Martens et al., 2008). Moreover, they also have difficulties with social interactions and peer relationships (Campos and Sotillo, 2012; Stinton and Howlin, 2012; Stojanovik, 2006), and their communicative language reveals serious impairments in pragmatics (Laws and Bishop, 2004). Importantly, Jawaid et al. (2011) highlight the social vulnerability and high levels of abuse and social isolation experienced by individuals with the disorder, in part due to atypical gaze behaviors and over-friendliness. In contrast, and seemingly at odds with their eagerness to interact in social situations, anxiety disorders occur at a high level in WS (Mervis and John, 2010; Stinton and Howlin, 2012).

Cognitive Development

Full-scale IQ is within the mild-to-moderate range for most individuals with WS (Martens et al., 2008). However, given the uneven language/spatial profile in WS (Mervis and John, 2010), it is more informative to measure verbal and nonverbal skills separately for this group (for discussion, see Mervis and John, 2010). Indeed, the uneven cognitive profile in WS makes it especially important to carefully consider the implications of experimental designs when selecting matched groups (Hudson and Farran, 2012), which can be surmounted by using the developmental trajectories approach (Thomas et al., 2009).

In the domain of memory, the verbal/visuospatial discrepancy observed in WS (e.g., Phillips et al., 2004) is also reflected in their performance in short-term memory tasks, although only in adolescents and adults, and not in younger children (Vicari and Carlesimo, 2006). Executive function abilities are generally poor (Menghini et al., 2010; Rhodes et al., 2010).

Similarly, visuospatial construction is especially impaired in WS (Landau et al., 2006), even when compared to their already poor level of nonverbal ability (Phillips et al., 2004). Moreover, Vicari and Carlesimo (2006) provide evidence that unlike DS and TD groups, WS visuospatial memory is poor even after their perceptual skills are taken into account.

Individuals with WS also experience considerable difficulty with number. Infants with WS can discriminate between small numbers, but have more difficulty discriminating large approximate magnitudes (Van Herwegen et al., 2008). They also display atypical scanning patterns when processing numerical displays (Karmiloff-Smith et al., 2012a). Furthermore, while somewhat older children with WS can learn to count, many do not understand cardinality, i.e., the final count word represents the total number in a set. When they are asked, 'how many?' just after counting a set, they merely start to count the objects all over again. Moreover, the ability to understand cardinality is predicted by language levels in WS, whereas it is predicted by spatial scores in typical control children (Ansari et al., 2003), suggesting different developmental trajectories.

Drawing practical implications from experimental work is important, such as finding one's way around the environment. Farran and colleagues recently studied navigational skills and asked individuals with WS to learn a real-life route. They found that although the WS participants were able to learn the route after practice, they were less successful at creating a mental map, which would have allowed them to work out a short cut (Farran et al., 2010).

Finally, John and Mervis (2010) demonstrated that children with low levels of sensory deficits were less impaired in executive functioning, temperament, adaptive functioning, and problem behaviors than those with high levels of sensory deficits. Attention is also a capacity that also shows an atypical developmental course in WS (Beckenridge et al., 2012). These findings constitute examples of low-level general impairments, which have cascading effects on the outcomes of specific cognitive domains.

Concluding Comments

The WS brain is 20% smaller than normal brains and qualitatively different in terms of brain anatomy, brain function, brain biochemistry, and computational processing. This holds throughout embriogenesis and postnatal brain development and means that interaction with environmental stimuli is likely to be subtly different. Given a different brain, it is unsurprising that even when overt behavior looks as if it were normal, as in some aspects of WS language and face processing, these skills actually turn out to be underpinned by cognitive and neural processes that are different from the normal case. Early low-level impairments may underlie later-developing skills, and performance can and does change over developmental time (Karmiloff-Smith, 1998; Karmiloff-Smith et al., 2012a). Delineating changes in developmental trajectories over time is critical to understanding any neurodevelopmental disorder (Thomas et al., 2009).

Williams syndrome is an excellent model for the neurocognitive study of genetic disorders, because of its known genetic origins and its unusual cognitive profile. The syndrome is especially important because of the way in which in-depth research highlights the need to go beyond both observable behavior and static descriptions of snapshots of development, to charting the full course of neurocognitive trajectories from infancy onward.

See also: Cognitive Development in Infancy, Neural Mechanisms of.

Bibliography

Ansari, D., Donlan, C., Thomas, M.S.C., Ewing, S.A., Peen, T., Karmiloff-Smith, A., 2003. What makes counting count? Verbal and visuo-spatial contributions to typical and atypical number development. Journal of Experimental Child Psychology 85, 50–62.

Baron-Cohen, S., 1998. Modularity in developmental cognitive neuropsychology: evidence from autism and gilles de la Tourette syndrome. In: Burack, J.A., Hodapp, R.M., Zigler, E. (Eds.), Handbook of Mental Retardation and Development. Cambridge University Press, Cambridge, pp. 334–348.

Beckenridge, K., Atkinson, J., Braddick, O., 2012. Attention. In: Farran, E.K., Karmiloff-Smith, A. (Eds.), Neurodevelopmental Disorders across the Lifespan: A Neuroconstructivist Approach. Oxford University Press, Oxford, pp. 119–133.

Bellugi, U., Lichtenberger, L., Jones, W., Lai, Z., St George, M., 2000. The neurocognitive profile of Williams syndrome: a complex pattern of strengths and weaknesses. Journal of Cognitive Neuroscience 12, 7–29.

Benton, A.L., Hamsher, K. de S., Varney, N.R., Spreen, O., 1983. Contributions to Neuropsychological Assessment. Oxford University Press, New York.

Boddaert, N., Mochel, F., Meresse, I., Seidenwurm, D., Cachia, A., Brunelle, F., Lyonnet, S., Zilbovicius, M., 2006. Parieto-occipital grey matter abnormalities in children with Williams syndrome. NeuroImage 30, 721–725.

Böhning, M., Campbell, R., Karmiloff-Smith, A., 2002. Audiovisual speech perception in Williams syndrome. Neuropsychologia 40, 1396–1406.

Brock, J., 2007. Language abilities in Williams syndrome: a critical review. Development and Psychopathology 19, 97–127.

Campos, R., Sotillo, M., 2012. Mental state understanding and social interaction. In: Farran, E.K., Karmiloff-Smith, A. (Eds.), Neurodevelopmental Disorders across the Lifespan: A Neuroconstructivist Approach. Oxford University Press, Oxford, pp. 279–295.

Clahsen, H., Almazan, M., 1998. Syntax and morphology in Williams syndrome. Cognition 68, 167–198.

Deruelle, C., Rondan, C., Mancini, J., Livet, M., 2003. Exploring face processing in Williams syndrome. Cognitie, Creier, Comportanent 7, 157–171.

Doyle, T.F., Bellugi, U., Korenberg, J.R., Graham, J., 2004. "Everybody in the world is my friend" hypersociability in young children with Williams syndrome. American Journal of Medical Genetics Part A 124A, 263–273.

Dunn, L.M., Dunn, L.M., 1981. Peabody Picture Vocabulary Test – Revised. American Guidance Service, Circle Pines, MN.

Ewart, A.K., Morris, C.A., Atkinson, D., Weishan, J., Sternes, K., Spallone, P., Stock, A.D., Leppert, M., Keating, M.T., 1993. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. Nature Genetics 5, 11–16.

Fahim, C., Yoon, U., Nashaat, N.H., Khalil, A.K., El-Belbesy, M., Mancini-Marie, A., Evans, A.C., Meguid, N., 2012. Williams syndrome: a relationship between genetics, brain morphology and behaviour. Journal of Intellectual Disability Research 56, 879–894.

Farran, E.K., Blades, M., Boucher, J., Tranter, L.J., 2010. How do individuals with Williams syndrome learn a route in a real-world environment? Developmental Science 13, 454–468.

Fenson, L., Dale, P.S., Reznick, J.S., Thal, D., Bates, E., Hartung, J.P., Pethick, S., Reilly, J.S., 1993. MacArthur Communicative Development Inventories: User's Guide and Technical Manual. Singular Publishing Group, San Diego, CA.

Frangiskakis, J.M., Ewart, A.K., Morris, C.A., Mervis, C.B., Bertrand, J., Robinson, B.F., Klein, B.P., Ensing, G.J., Everett, L.A., Green, E.D., Pröschel, C., Gutowski, N.J., Noble, M., Atkinson, D.L., Odelberg, S.J., Keating, M.T., 1996. *LIM-kinase1* hemizygosity implicated in impaired visuospatial constructive cognition. Cell 86, 59–69.

Germine, L., Cashdollar, N., Düzel, E., Duchaine, B., 2011. A new selective developmental deficit: impaired object recognition with normal face recognition. Cortex 47, 598–607.

Grant, J., Valian, V., Karmiloff-Smith, A., 2002. A study of relative clauses in Williams syndrome. Journal of Child Language 29, 403–416.

Gray, V., Karmiloff-Smith, A., Funnell, E., Tassabehji, M., 2006. In-depth analysis of spatial cognition in Williams syndrome: a critical assessment of the role of the LIMK1 gene. Neuropsychologia 44, 679–685.
Grice, S.J., Spratling, M.W., Karmiloff-Smith, A., Halit, H., Csibra, G., de Haan, M., Johnson, M.H., 2001. Disordered visual processing and oscillatory brain activity in autism and Williams syndrome. NeuroReport 12, 2697–2700.
Hudson, K., Farran, E.K., 2012. Executive function and motor planning. In: Farran, E.K., Karmiloff-Smith, A. (Eds.), Neurodevelopmental Disorders across the Lifespan: A Neuroconstructivist Approach. Oxford University Press, Oxford, pp. 165–181.
Jawaid, A., Riby, D.M., Owens, J., White, S.W., Tarar, T., Schulz, P.E., 2011. 'Too withdrawn' or 'too friendly': considering social vulnerability in two neurodevelopmental disorders. Journal of Intellectual Disability Research 56, 335–350.
John, A.E., Mervis, C.B., 2010. Sensory modulation impairments in children with Williams syndrome. American Journal of Medical Genetics Part C: Seminars in Medical Genetics 154C, 266–276.
Jones, W., Bellugi, U., Lai, Z., Chiles, M., Reilly, J., Lincoln, A., Adolphs, R., 2000. Hypersociability in Williams syndrome. Journal of Cognitive Neuroscience 12, 30–46.
Karmiloff-Smith, A., 1998. Development itself is the key to understanding developmental disorders. Trends in Cognitive Sciences 2, 389–398.
Karmiloff-Smith, A., 2010. Neuroimaging of the developing brain: taking "developing" seriously. Human Brain Mapping 31, 934–941.
Karmiloff-Smith, A., Grant, J., Berthoud, I., Davies, M., Howlin, P., Udwin, O., 1997. Language and Williams syndrome: how intact is "intact"? Child Development 68, 246–262.
Karmiloff-Smith, A., Thomas, M., Annaz, D., Humphreys, K., Ewing, S., Brace, N., Van Duuren, M., Pike, G., Grice, S., Campbell, R., 2004. Exploring the Williams syndrome face-processing debate: the importance of building developmental trajectories. Journal of Child Psychology and Psychiatry 45, 1258–1274.
Karmiloff-Smith, A., D'Souza, D., Dekker, T.M., Van Herwegen, J., Xu, F., Rodic, M., Ansari, D., 2012a. Genetic and environmental vulnerabilities in children with neurodevelopmental disorders. Proceedings of the National Academy of Sciences of the United States of America 109, 17261–17265.
Karmiloff-Smith, A., Broadbent, H., Farran, E.K., Longhi, E., D'Souza, D., Metcalfe, K., Tassabehji, M., Wu, R., Senju, A., Happé, F., Turnpenny, P., Sansbury, F., 2012b. Social cognition in Williams syndrome: genotype/phenotype insights from partial deletion patients. Frontiers in Psychology 3 (168), 1–8.
Laing, E., Butterworth, G., Ansari, D., Gsödl, M., Longhi, E., Panagiotaki, G., Paterson, S., Karmiloff-Smith, A., 2002. Atypical development of language and social communication in toddlers with Williams syndrome. Developmental Science 5, 233–246.
Laing, E., Hulme, C., Grant, J., Karmiloff-Smith, A., 2001. Learning to read in Williams syndrome: looking beneath the surface of atypical reading development. Journal of Child Psychology and Psychiatry 42, 729–739.
Landau, B., Hoffman, J.E., Kurz, N., 2006. Object recognition with severe spatial deficits in Williams syndrome: sparing and breakdown. Cognition 100, 483–510.
Laws, G., Bishop, D.V.M., 2004. Pragmatic language impairment and social deficits in Williams syndrome: a comparison with Down's syndrome and specific language impairment. International Journal of Language & Communication Disorders 39, 45–64.
Martens, M.A., Wilson, S.J., Reutens, D.C., 2008. Research review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype. Journal of Child Psychology and Psychiatry 49, 576–608.
Masataka, N., 2001. Why early linguistic milestones are delayed in children with Williams syndrome: late onset of hand banging as a possible rate–limiting constraint on the emergence of canonical babbling. Developmental Science 4, 158–164.
Menghini, D., Addona, F., Costanzo, F., Vicari, S., 2010. Executive functions in individuals with Williams syndrome. Journal of Intellectual Disability Research 54, 418–432.
Mervis, C.B., John, A.E., 2010. Cognitive and behavioral characteristics of children with Williams syndrome: implications for intervention approaches. American Journal of Medical Genetics Part C: Seminars in Medical Genetics 154C, 229–248.
Mervis, C.B., John, A.E., 2012. Precursors to language and early language. In: Farran, E.K., Karmiloff-Smith, A. (Eds.), Neurodevelopmental Disorders across the Lifespan: A Neuroconstructivist Approach. Oxford University Press, Oxford, pp. 187–204.
Meyer-Lindenberg, A., Mervis, C.B., Berman, K.F., 2006. Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour. Nature Reviews Neuroscience 7, 380–393.
Mills, D.L., Alvarez, T.D., St George, M., Appelbaum, L.G., Bellugi, U., Neville, H., 2000. Electrophysiological studies of face processing in Williams syndrome. Journal of Cognitive Neuroscience 12, 47–64.
Musolino, J., Chunyo, G., Landau, B., 2010. Uncovering knowledge of core syntactic and semantic principles in individuals with Williams syndrome. Language Learning and Development 6, 126–161.
Phillips, C.E., Jarrold, C., Baddeley, A.D., Grant, J., Karmiloff-Smith, A., 2004. Comprehension of spatial language terms in Williams syndrome: evidence for an interaction between domains of strength and weakness. Cortex 40, 85–101.
Pinker, S., 1994. The Language Instinct. Penguin, Harmondsworth, UK.
Pober, B.R., 2010. Williams–Beuren syndrome. New England Journal of Medicine 362, 239–252.
Rae, C., Karmiloff-Smith, A., Lee, M.A., Dixon, R.M., Grant, J., Blamire, A.M., Thompson, C.H., Styles, P., Radda, G.K., 1998. Brain biochemistry in Williams syndrome: evidence for a role of the cerebellum in cognition? Neurology 51, 33–40.
Reiss, A.L., Eckert, M.A., Rose, F.E., Karchemskiy, A., Kesler, S., Chang, M., Reynolds, M.F., Kwon, H., Galaburda, A., 2004. An experiment of nature: brain anatomy parallels cognition and behavior in Williams syndrome. The Journal of Neuroscience 24, 5009–5015.
Rhodes, S.M., Riby, D.M., Park, J., Fraser, E., Campbell, L.E., 2010. Executive neuropsychological functioning in individuals with Williams syndrome. Neuropsychologia 48, 1216–1226.
Stinton, C., Howlin, P., 2012. Adult outcomes and integration into society. In: Farran, E.K., Karmiloff-Smith, A. (Eds.), Neurodevelopmental Disorders across the Lifespan: A Neuroconstructivist Approach. Oxford University Press, Oxford, pp. 103–115.
Stojanovik, V., 2006. Social interaction deficits and conversational inadequacy in Williams syndrome. Journal of Neurolinguistics 19, 157–173.
Strømme, P., Bjørnstad, P.G., Ramstad, K., 2002. Prevalence estimation of Williams syndrome. Journal of Child Neurology 17, 269–271.
Tager-Flusberg, H., Plesa-Skwerer, D., Faja, S., Joseph, R.M., 2003. People with Williams syndrome process faces holistically. Cognition 89, 11–24.
Temple, C.M., 1997. Cognitive neuropsychology and its application to children. Journal of Child Psychology and Psychiatry 38, 27–52.
Thomas, M.S.C., Annaz, D., Ansari, D., Scerif, G., Jarrold, C., Karmiloff-Smith, A., 2009. Using developmental trajectories to understand developmental disorders. Journal of Speech, Language, and Hearing Research 52, 336–358.
Thomas, M.S.C., Karmiloff-Smith, A., 2003. Modeling language acquisition in atypical phenotypes. Psychological Review 110, 647–682.
Tyler, L.K., Karmiloff-Smith, A., Voice, J.K., Stevens, T., Grant, J., Udwin, O., Davies, M., Howlin, P., 1997. Do individuals with Williams syndrome have bizarre semantics? Evidence for lexical organization using an on-line task. Cortex 33, 515–527.
Van Herwegen, J., Ansari, D., Xu, F., Karmiloff-Smith, A., 2008. Small and large number processing in infants and toddlers with Williams syndrome. Developmental Science 11, 637–643.
Vicari, S., Carlesimo, G.A., 2006. Short-term memory deficits are not uniform in Down and Williams syndromes. Neuropsychology Review 16, 87–94.
Volterra, V., Caselli, M.C., Capirci, O., Tonucci, F., Vicari, S., 2003. Early linguistic abilities of Italian children with Williams syndrome. Developmental Neuropsychology 23, 33–58.
Wilson, B., Cockburn, J., Baddeley, A.D., 1985. Rivermead Behavioural Memory Test. Thames Valley Test Co., Hampshire, UK.

Relevant Websites

http://www.autourdeswilliams.org/j/ – Autour des Williams.
http://www.sindromewilliams.org/ – Williams Syndrome Association, Spain.
http://www.sindromediwilliams.org/ – AISW Associazione Italian Sindrome di Williams Onlus.
http://www.syndromedewilliams.be/ – Association belge du syndrome de Williams.
http://www.w-b-s.de/ – Bundesverband Williams Beuren Syndrom e.V.
http://www.williams-syndrome.org/ – Williams Syndrome Association.
http://www.williams-syndrome.org.uk/ – Williams Syndrome Foundation, UK.

Wilson, William Julius (1935–)

Peter I Rose, Kahn Liberal Arts Institute, Smith College, Northampton, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

William Julius Wilson, Lewis P. and Linda L. Geyser University Professor at Harvard University, is one of the best known and widely cited American sociologists of the late twentieth and early twenty-first centuries. His major books, *Power, Racism and Privilege* (1973), *The Declining Significance of Race* (1978), *The Truly Disadvantaged* (1987, revised edition 2012), *When Work Disappears* (1996), and *More than Just Race* (2010), all address an abiding personal and professional concern with and critical analysis of two issues, race and poverty.

The sociologist William Julius Wilson, currently Lewis P. and Linda L. Geyser University Professor at Harvard University, has long made outstanding contributions to the study of social inequality and urban poverty in the United States.

Wilson has been given many different labels but the one he claims to feel most comfortable with is 'social democrat.' He especially identifies himself with the approach of late Michael Harrington, Chairman of the Democratic Socialists of America, coiner of the expression 'neo-conservative,' and author of *The Other America: Poverty in the United States* (1962).

The Other America, published in the heady days of the new Kennedy Administration, was for many an exposé of the plight of millions of Americans without jobs, without status, and often without hope. Wilson's far more extensive studies of the same phenomenon indicate that the chasm between those who make it in America and those left behind is as wide as it was 50 years ago. His research and conclusions have influenced two more recent presidents of the United States, Bill Clinton and Barack Obama, and many others inside the government and in the private sector, as well as thousands of students of urban sociology and social policy in this country and around the world.

Early Life and Career

Wilson's personal awareness of the inequities in our society may well have been generated during a childhood spent in the small town of Blairsville in Derry Township in western Pennsylvania where he was born in the depth of the Great Depression on 20 December 1935. He was one of six children whose father – who had worked as a coal miner and in steel mills – died when William was 12 years old. To support the family, his mother sought employment as a housemaid. In various interviews he has described the situation as living in rural poverty, part of the time on relief (social welfare). He studied first at the small, private, historically black AME (African Methodist Episcopal) affiliated Wilberforce University in Ohio, from which he received a Bachelor of Arts in 1958. He went on to do graduate work, first at Bowling Green University (MA, 1961), then to Washington State University in Pullman, Washington, from which he received a Ph.D. in sociology and anthropology in 1966. In those varied environments and during his brief stint doing military service, Wilson was exposed to a considerable cross section of America's people. These experiences doubtlessly had an influence on the perspectives he developed on the nuances and contradictions of social life, leading him to eschew commonplace pigeonholing in his long and active profession as a social scientist.

His teaching career began in Amherst, Massachusetts, where he started as an Assistant Professor at the flagship campus of the University of Massachusetts. While teaching and starting work on the book on the growing socioeconomic gap in African-American communities, he was an active member of the Department of Sociology and of a Five College Committee working on ways to bring Afro-American Studies to each of the campuses in that consortium (Amherst, Hampshire, Mt. Holyoke, Smith colleges, and the University of Massachusetts). In the winter of 1971–72 he published a thoughtful article about "Issues and Challenges of Black Studies," in the *Journal of Social and Behavioral Scientists*. The following fall he moved to the University of Chicago.

While Wilson would later conduct comparative studies looking at race and poverty issues in a number of other metropolitan areas, as was the case during the heyday of Robert E. Park's 'Chicago School of Sociology,' Chicago, the city described in reports such as *The Negro in Chicago: A Study of Race Relations* (1922) by Charles S. Johnson, and such books as Frederic M. Thrasher, *The Gang* (1927), Louis Wirth's *The Ghetto* (1928), Harvey Zorbaugh's *The Gold Coast and the Slum* (1929), and St. Clair Drake and Horace Cayton's *Black Metropolis: A Study of Negro Life in a Northern City* (1945), Chicago would be the major site and anchor point of much of Wilson's own research.

He twice served as Chairman of the University of Chicago's Department of Sociology (1978–81 and 1984–87), spent the next academic year as a Fellow at the Center for Advanced Study in the Behavioral Sciences in Stanford, California, and the fall of 1983 as Langston Hughes Visiting Professor at the University of Kansas in Lawrence, Kansas. Upon his return to Chicago, he was appointed Lucy Flower Distinguished Service Professor. In 1990, he became a University Professor as well as the Director of the Center for the Study of Urban Inequality in the University of Chicago's School of Public Policy. He gave up both positions when he joined the Harvard faculty in 1996. Cambridge, Massachusetts has been his home base ever since.

William Julius Wilson's books, articles, and lectures on the roots and persistence of urban poverty have enlightened many

fellow academics and practitioners in the fields of urban economics, social policy, and grassroots social work. While addressing some of the most complex of all the persisting domestic problems in the United States, his writings have a clarity of structure and presentation that offer important grounding for social policy makers and a useful model for graduate students in the social sciences and for those already in the disciplines like sociology, economics, and political science reluctant to eschew arcane language lest they be considered unscientific.

This is not to say that everything Wilson has written or that all of his studies are flawless in their construction, or that his conclusions and opinions based on his studies do not open themselves to critical assessment by others who have looked at similar issues and come to somewhat different conclusions about causes, effects, and remedies. Yet by almost every measure Wilson has altered perceptions of class and race in American society and offered insights that seriously challenge many assumptions about these two critical variables and their impact – especially on city dwellers – over the four decades he has been studying, analyzing, and commenting on these phenomena using a truly integrated historical, economic, and sociological approach.

Wilson's earliest publications – one in the *Pacific Sociological Review* (1964); a second in the *American Sociological Review* (1967) – were not about race or class but concerned with sociological theory and theory construction. He was particularly interested in varied approaches to the study and empirical measurement of norms and values. The first time he published anything on what many think of as *his* subject was an essay, "The Quest for a Meaningful Black Experience on White Campuses," in the literary journal, *The Massachusetts Review* (1969). From then on almost all of his papers, articles, or books have been on race, class, and related matters.

Illustrative of his proclivity for developing paradigms and integrating sociological theory and social concerns, his contribution the book, *Nation of Nations*, published by Random House in 1972, compared and contrasted several dominant themes in approaching what he would later call 'the underclass.' He called his essay "Race Relations Models and Ghetto Behavior." Wilson had already begun to address these matters while working on his first book, *Power, Racism and Privilege: Race Relations in Theoretical and Historical Perspectives*, published by Macmillan in 1973.

During that same time period, Wilson joined me as a co-editor of an already contracted volume of original essays triggered by the famous Japanese film, *Rashomon*, in which different witnesses to a brutal crime were asked to present their views of what they thought they saw and remembered. The idea was to invite a number of black and white social scientists to offer their perspectives on what we were all witnessing in American society circa 1970. Wilson signed on, as did my colleague, the late political scientist, Stanley Rothman. Together we gathered a group of writers and – in some cases, those of persons in assigned roles: black and white professors, black integrationists, city cops, welfare mothers, school teachers, white ethnics, black militants, etc.

Through Different Eyes: Black and White Perspectives on American Race Relations was published by Oxford University Press in 1973.

One of the most important extras was Bill Wilson's sensitive contribution to the Epilogue of the book, an essay called "The Significance of Social and Racial Prisms." There he observed that while it is true that blacks and whites of the general public tend to view American race relations through different eyes, it is likewise the case that within each racial cohort, perceptions vary according to income, occupation, education, and place of residence. And even here, he noted

> There is no simple one-to-one relationship between individuals' views on racial matters and their respective positions in the social structure because racial attitudes are in major measure shaped by an important intervening variable – feelings of apprehension that out-group racial members plan to undermine, encroach on, or challenge in-group claims to certain rights and privileges. People of high social status, for example, tend to be less virulent in their racial attitudes because their position in society is more secure and they have fewer fears that their interests or prerogatives will be jeopardized by out-group racial members.

These sentiments – many reflecting on the widening gap in Black America – would echo and reecho in what he would publish throughout the rest of the twentieth century and into the second decade of the twenty-first century.

The Race–Class Debate

The first of those new works, published in 1978 by the University of Chicago Press, proved to be one of Wilson's most controversial books – not the least because of its stark title, *The Declining Significance of Race*. In a lengthy essay, "The Declining Significance of Race: Revisited and Revised" in *Daedalus* in 2011, a year before the University of Chicago Press published a third edition of the book, Wilson recently reflected on what he had written more than three decades earlier.

His essay begins with a well-known quotation from George Bernard Shaw: "It is better to be criticized and misunderstood than to be ignored." Wilson then restates his basic arguments, noting, for example, that "changes in the system of production and in government policies have affected, over time, black/white access to rewards and privileges as well racial antagonisms" with two significant results: the deteriorating economic conditions of poor blacks and the concurrent improvement of middle-class blacks during three significant eras in American race relations, the last being the years since World War II when "the black class structure began to take on some of the characteristics of white class structure and that economic class gradually became more important than race in determining the life chances of individual African Americans."

In his paper he addresses the concerns of certain critics, such as Jennifer Hochschild, who felt he did not pay as much attention to generational issues as he should have. He quotes Hochschild's comment that "One has not really succeeded in American unless one can pass on the chance for success to one's children." (The sentence may have reminded some readers of the quip that "In the U.S. it is easy to climb the ladder of success if Daddy owns the ladder.") But Wilson claims that there is ample evidence that middle-class blacks have not only been

able to pass their successes on to their children but have advantaged them in other ways as well.

Wilson also responds to Michael Hout, who had pointed out, while not gainsaying Wilson's basic thesis, that much of the mobility that took place among middle-class blacks was in the public rather than the private sector, and that much more weight should be attributed to 'affirmative action' programs implemented in the 1960s to account for widening rather than narrowing the gap.

Wilson mentions the criticisms of others who had argued that his thesis does not jibe with hard data on differentials between whites and blacks in job security and income in the years since the first edition of *The Declining Significance of Race* was first released, even though the majority noted that 'class does matter.' He concedes that there is some justification in what they say and notes that the biggest changes that have occurred in the number of government jobs available to blacks and whites and the waning enthusiasm for 'affirmative action' in both public and private sectors and that the black class structure increasingly reflects gender differences, especially among younger blacks, as males have fallen behind females on a number of socioeconomic indicators: employment rates, the imbalanced ratio of females to males who complete high school, and average income, with some of the sharpest discrepancies at the lower end of the income hierarchy.

Withal, at the end of his self-critical review Wilson noted that, "If I were writing the *Declining Significance of Race* today, I would provide more balance in my policy recommendations by placing much greater emphasis on the need to *strongly* and *continuously* embrace, as well as advance, both race- and class-based solutions to address life chances for people of color." (Italics are Wilson's.)

Considering "The Truly Disadvantaged"

What Wilson had variously referred to as an 'underclass' and 'the ghetto poor' is graphically captured in the title of his next book, *The Truly Disadvantaged*, also published by the University of Chicago Press (1987). A new edition would be released in 2012.

Addressing "The Inner City, the Underclass, and Public Policy," it was – and remains – a landmark contribution to the literature. Some reviewers said it was Wilson's *magnum opus*. When first published, it was far more widely discussed by politicians, pundits, and the public than his book, *The Declining Significance of Race*, which was read and debated mostly in academic circles.

In the pages of data-rich *The Truly Disadvantaged*, Wilson zeroed in on the interplay among three critical variables: race, employment, and education. The volume is in many ways a clear extension and refinement of concerns underlying his basic premise about changes within the black community and the wider society.

Based on a close examination of both hard and soft data, the author argued that increasing general prosperity and a lessening of racial discrimination enhanced the opportunities of many members of the middle class, including many upwardly mobile African-Americans. They not only began to move up, they also moved out, leaving a great void for those left behind in the inner-city ghettos. More and more young people, especially young men, suffered from the departure of leaders and role models. With fewer and fewer job opportunities, many were drawn into illicit activities, exacerbating the situation and contributing to a breakdown of values, familial and communal ties, and a dramatic increase of various forms of urban plight and pathology, both social and psychological.

Not surprisingly, as he probed more deeply, Wilson found himself revisiting many of the issues raised and enumerated on nearly 20 years earlier by Daniel Patrick Moynihan in his controversial report, *The Negro Family: A Case for National Action* (1964). This subject was also addressed the same year in Kenneth Clark's book, *Dark Ghetto* (1964).

In Wilson's far more detailed study, he also examined the character of the ghetto, the fragility of families, the dramatic increase in one-parent female-headed homes, and the estrangement of both old and young in what both Moynihan and Clark had called a "tangle of pathology." But more than either Moynihan or Clark, Wilson zeroed in on the economic structure and the depressing situation where employment prospects were rapidly disappearing in the very metropolitan areas where for decades blacks, having been displaced by the mechanization of agriculture in the south – and, in no small measure, Jim Crow restrictions – had sought and often found employment with labor-intensive jobs in the north. Then, as new trends in production and the collapse of certain core industries – steel production, automobile assembling, and lighter forms of manufacturing – accelerated, they were again harshly affected. But this time, for workers, especially those without much education, there was no place to which to flee. The ghetto that had become their home often became their jail. This subject would be addressed in much greater detail in Wilson's next major book.

What Happens When Work Disappears

In *When Work Disappears: The World of the New Urban Poor*, published by Alfred A. Knopf in 1996, Wilson presents a case for understanding the plight of the poor based on his analyses of a series of neighborhood studies, detailed interviews and surveys. That data had three major sources: research conducted under the aegis of UPFLS (the Urban Poverty and Family Life Study), the Woodlawn and Oakland Neighborhood Study, and an examination on adolescent social development and the effects of growing up in high-risk neighborhoods.

In the first pages of the book, the author acknowledges and thanks his many colleagues and students at the University of Chicago whose data-sets gave grounding to the examination and analysis of two principal changes in recent urban history: the flight of whites to the suburbs with the subsequent diminution of the tax-base, and, particular institutional changes, particularly the sharp reduction of entry-level opportunities for poor people as businesses and industries also moved out. Concomitant with these causal factors, the author points to their effects not only on the life chances of those left behind in the inner-city but to the demoralizing reality of having fewer and fewer positive role-models within the community and what Wilson sees as the ebbing away of a sense of personal responsibility by those too injured by their plight to mobilize

even the few resources to help themselves. Here he returns to a consideration of the successes and failures of affirmative action programs – programs that clearly had enhanced mobility for those ready, willing, and able to avail themselves of their promise but seemed to do little to attack or mitigate the situation of those left behind and suffering the most alienation in the anomic depths of the black ghettos. (*Note*: It is here where Wilson contributes his own views to the still on-going debates about the implicit bias in affirmative action programs which at present clearly serve the needs of a particular segment within the communities of designated minorities – Asians, Latinos, Native Americans, and especially African–Americans – who are ready, willing, and generally able to take full advantage of such institutionalized gateways. He, along with many others, favors a shift from an exclusively race-based to a more class-based approach.)

Wilson argues that it is not sufficient simply to understand what happens when work disappears, especially in communities that are and have long been terribly vulnerable to patterns of discrimination, poor public services, and low expectations, but it is also necessary to address the extent to which too many inhabitants internalize their marginalized status and act accordingly. As any number of people who have commented on this work have indicated, there is a difference between those who were in fact members of a solid blue-collar working class and found themselves out on the streets and those who, for any number of reasons, never having held steady jobs, were already there.

In a final chapter in *When Work Disappears*, titled "A Broader Vision: Social Policy Options in Cross-National Perspective," Wilson offers some important comparative perspectives on extant programs that simultaneously address the problems of economic growth, joblessness and wage inequality, and programs worthy of emulating in this country. He looks to other societies and their welfare and education systems, especially, in the latter case, those that stress and enthusiastically support school-to-work programs. He ends by offering several concrete policy prescriptions for what to do to address America's problems, many related to general economic conditions and their impact and many a clear remnants of slavery's persisting legacy.

Among a number of policy recommendations regarding the United States, most of which seem to stress the clear requirement that coalitions are needed to move forward, Wilson presses for race-neutral programs geared to jobs creation, further expansion of the earned income tax credits, public school reform, child care programs, and universal health insurance. Recognizing that such will take time – and much intense politicking to achieve their aims, he argues that some matters cannot wait. Most pressing are labor-intensive programs of public service jobs for all workers, and not only those in the inner city with low skills and little or no work experience. This, says Wilson, would help to avoid stigmatization and might broaden the political base of support for the more far-reaching proposals. Such a new New Deal campaign would have as its immediate goal to provide training, jobs, and hope. Finally, linking the discussion of recommended ways of remediating deep-rooted and contemporary problems, the author loops back to his core argument: "My framework for long-term and immediate solutions is based on the notion that the problems of jobless ghettos cannot be separated from those of the rest of the nation."

Here he opened himself once again to criticism by those who have long claimed that the African-American experience is so unique that it must be separated from that of others. He was (is) prepared to weather the storm as he had done earlier in his career when under fire for the use of some highly charged words, not the least, 'underclass,' and more significantly for persisting in stressing the growing significance of class over race in American society.

Wilson's Analyses and Social Policy

Many who reviewed *The Truly Disadvantaged* echoed the sentiments of Robert Greenstein, former Director of the Center on Budget and Policy Priorities in Washington, DC, and published as "Prisoners of the Economy" in the *New York Times* on 25 October 1987. "[The book is] both a rejoinder to what Mr. Wilson regards as flawed conservative critiques and a call to liberals and civil rights leaders to broaden their agendas beyond what he terms the War on Poverty-vision and 'race-specific' policies such as affirmative action." Like Greenstein, many sociologists and economists would also claim that *The Truly Disadvantaged* far surpasses many other attempts to explore the issues of race and poverty. Some, however, would come to agree with Greenstein's statement that "the reader is left with one disappointment. The book's policy prescriptions are less well developed than its powerful analysis of what has gone wrong."

Others said similar things in reviews of *When Work Disappears*, a book that contains far more detailed discussion of social policy than Wilson's previous ones. The critics of *When Work Disappears* tended to fall into one of two camps: those who felt that somehow Wilson had abandoned his own community and, like Moynihan before him, seemed to put an undue burden on the victims and not enough on the system; and those who, like Greenstein, felt that he was spot-on in terms of the problems but rather naive in terms that, given the political climate, his recommendations could bring about the far-reaching structural changes. Either way, it seemed there was a tendency to 'kill the messenger,' the messenger being Wilson himself.

Among those who most closely examined the Wilson's writings up to and including the publication of *When Work Disappear*s is the sociologist Jack Niemonen. In a book-length study called *Race, Class, and the State in Contemporary Sociology: The William Julius Wilson Debates* (2010), Niemonen examined a number of issues that appear and reappear throughout the body of Wilson's scholarly oeuvre, most but not all still focused on 'The Race–Class Debate.'

Perhaps Niemonen's most critical chapter reflects the sentiments many had long expressed. He called it "Wilson's Achilles Heel," that is, Wilson's seeming failure to adequately come to terms with the continuing significance of racially-based segregation. His charge reprises the negative reaction of the Association of Black Sociologists way back in 1978, the year he was awarded the American Sociological Association's Sydney Spivack Award "for recognition of a significant sociological work in intergroup relations over the previous five-year

period." It is a matter that, as already noted, is a leitmotif in the writings of those who dismiss his basic premises.

Niemonen's own conclusion seemed to damn Wilson with faint praise. Like Greenstein and others, he starts by acknowledging the important role that Wilson played in catalyzing a reassessment of the relationship of race and class, and a renaissance of theoretical constructs related to social stratification in general – something that was already being done by several sociologists. Niemonen allows that Wilson's early writing, particularly the principal thesis in *The Declining Significance of Race*, not only led to a number of studies to test it, including those extensive surveys by Wilson's teams, but also to Wilson's modification of it in subsequent essays and in later editions of the book that started the whole debate:

> Without a doubt, Wilson's contribution to the sociology of race relations ranks as extraordinary and productive. Nevertheless, I have to conclude that his attempt to bridge the racial divide through affirmative opportunity and economic growth falls short. An explicit recognition of the state's role in structuring life-chances in the industrial and modern industrial periods is missing.

But Wilson does this. Whether emphasizing continuing discrimination against African-Americans that makes it even harder for them to compete or broader societal changes related to the shifting demands of the economy, excessive power of large corporations and legislative logjams caused by bittern political polarization which contribute to difficulties not only for the never-hads but also for the once-hads, he has long argued for innovative programs and much greater government intervention. He has also been quite specific about the areas most in need of such assistance.

The gnawing question seems not to be the inappropriateness of Wilson's solutions relating to employment and education, but a conundrum faced by people like Neimonen – and himself – who understand the problems and have concrete solutions but find implementation of truly effective societal-changing policies difficult owing to the fact that public support and funding are dependent on the willingness of those who govern to commit themselves to making dramatic structural changes. At the present time, the prospect of such happening is highly unlikely.

More Recent Work

"Well, there goes the neighborhood!" is a sardonic expression long associated with urban succession often said to be caused by the invasion of "others" (Irish, Italians, Jews, Latinos, Asians, and especially African-Americans) who are perceived as threats to a way of life and to property values. Wilson and a group of colleagues set forth to examine the veracity of the phrase in four sections of Chicago, each an ethnic enclave. The study resulted in another book, *There Goes the Neighborhood: Race, Ethnic, and Class Tensions in Four Chicago Neighborhoods and Their Meaning for America*, written with Richard Taub and based on the research of several teams of graduate students and assistants. Published by Alfred A. Knopf in 2006, it offers a twenty-first century reprise of the type of sociological research conducted by those in the Chicago School under the leadership of Robert E. Park in the 1920s and 1930s. As was the case nearly a century ago, the focus was on different communities; now, as then, the emphasis is on stability and change. *There Goes the Neighborhood* also offers personal commentaries and rich statistical data that provides ample evidence – some quite startling, of persisting prejudices and considerable defensive insulation in the four working- and lower-middle class neighborhoods mainly inhabited by whites, blacks, Latinos, and others, respectively. This major study of Chicago today reminds the reader of the seminal line in Nathan Glazer and Daniel Patrick Moynihan's 50-year-old book about New York City, *Beyond the Melting Pot*, when the authors bluntly stated, "The point about the melting pot … is that it did not happen. At least not in New York and, *mutatis mutandis*, in those parts of America which resemble New York."

From the researches and reflection of William Julius Wilson and his colleagues, today's Chicago resembles New York of the 1950s in many significant ways. And, in words first used by the sociologist, Milton M. Gordon, in *Assimilation in American Life* (1964), structural assimilation, especially among those in the working-class in American cities is still a long way off.

There Goes the Neighborhood provided important insight in many of the matters on which William Julius Wilson had been drawn to examine over the years. This is again reflected in his most recent book, which, in itself, ties together many loose ends.

More Than Just Race: Being Black and Poor in the Inner City, published by Norton in 2009, part of a series of volumes under the editorship of Wilson's Harvard colleague, Henry Louis Gates, is a brief but surprisingly comprehensive portrait of urban sociology and anthropology, most of it focused on studies and theories about social conditions in the inner city in post–World War II America. The major emphasis is on the inequities built into the social structure that exacerbate cultural styles and responses on the part of those in both the dominant groups and in the minorities, especially blacks.

The Legacy of William Julius Wilson

Like the late historian, Stanley M. Elkins, author of *Slavery: A Problem in American Institutional and Intellectual Life* (1959), and the late Senator Daniel Patrick Moynihan, author of many books and reports, not the least, *The Negro Family: A Case for National Action* (1964), William Julius Wilson has often found himself in the middle of disputes about his contentions – and his intentions, his research, his analyses, and his use of the term 'underclass' as a shorthand expression for the urban poor. (One of the best examples of his explication of the concept and an explanation of the word as an appropriate one for that on which he was so keenly focused appeared in "The Underclass: Issues, Perspectives, and Public Policy" in a special issue of the *Annals of the American Academy of Political and Social Science* in January 1989.)

Controversies about his work often motivated Wilson to rethink certain suppositions and declamations. And he did. Sometimes he modified positions taken, more often he qualified them, countering challengers with compelling arguments, often presenting solid empirical evidence to back his claims. In

reference to a late edition of *The Declining Significance of Race* in a *New York Times* book review several years ago, the sociologist Paul Starr called it "easily one of the most erudite and sober diagnoses of the American black situation. Students of race relations and anybody in the policy-making position cannot afford to bypass this study." Nor, it should be added, should they overlook William Julius Wilson's other work, all elaborations, refinements, and variations on his basic theme.

Thoughtful, thorough, and innovative, it is not without reason that this mild-mannered man with a steely mind is seen by many as the very model of a modern sociologist, traits that have led to his garnering of many honorary doctorates, fellowships including a coveted MacArthur 'genius award,' invitations to lecture at hundreds of colleges and universities, and invitations to serve on innumerable boards and panels relating to his principal concerns.

A major theme in the domestic agenda of the second term of American President Barack Obama was to address social inequality, urban poverty, and the growing rifts not only within the black community but throughout the United States. These are the very issues that have long been and remain at the very core of William Julius Wilson's research and analyses, and his repeated calls urging action be taken to enhance access to the mainstream of society for those most disadvantaged by structural and cultural factors.

See also: Affirmative Action, Sociology of; Ghetto; Mobility, Social; Poverty, Sociology of; Racism, Sociology of; Social Inequality in History (Stratification and Classes); Social Stratification; Social Structure; Urban Sociology.

Bibliography

Glazer, N., Moynihan, D.P., 1964. Beyond the Melting Pot: The Negroes, Puerto Ricans, Jews. Italians, and Irish of New York City. MIT Press, Cambridge, MA.

Gordon, M.M., 1965. Assimilation in American Life: The Role of Race, Religion and National Origins. Oxford University Press, New York.

Greenstein, R., 25 October 1987. Prisoners of the Economy. New York Times.

Moynihan, D.P., 1964. The Negro Family: A Case for National Action, U.S. Government Printing Office, Washington, DC.

Niemonen, J., 2002. Race, Class, and the State in Contemporary Sociology: The William Julius Wilson Debates. Lynne Rienner Publishers, Boulder, CO.

Rose, P.I., Rothman, S., Wilson, W.J. (Eds.), 1973. Through Different Eyes: Black and White Perspectives on American Race Relations. Oxford University Press, New York and London.

Smelser, N.J., Wilson, W.J., Mitchell, Faith (Eds.), 2001. America Becoming: Racial Trends and Consequences in the United States. National Academy Press, Washington, DC.

Wilson, W.J., 1973. Power, Racism and Privilege: Race Relations in Theoretical and Sociohistorical Perspectives. Macmillan Company, New York (Reprinted in a paperback edition by The Free Press, 1976).

Wilson, W.J., 1978. The Declining Significance of Race: Blacks and Changing American Institutions. University of Chicago Press, Chicago (Second and enlarged edition, 1980; third edition –with a new Afterword – 2012.).

Wilson, W.J., 1987. The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy. University of Chicago Press, Chicago. Second edition – with a new Afterword – 2012. (Also published in French, *Les Oubliés de l'Amérique*, trans. by Ivan Ermakoff, Paris: Desclée de Brouwer in 1994; in Japanese, Akashi Shoten, Tokyo, 1999; and in Chinese, Shanghai People's Publishing House, Shanghai, 2008.).

Wilson, W.J., 1996. When Work Disappears: The World of the New Urban Poor. Alfred A. Knopf, New York (Vintage paperback edition, 1997; also published in Japanese, Akashi Shoten, Tokyo, 1999.).

Wilson, W.J., 1999. The Bridge over the Racial Divide: Rising Inequality and Coalition Politics. University of California Press, Berkeley, CA.

Wilson, W.J., 2001. Soziale Ungleichheit in Den USA. Plädoyer für eine multiethnische Bündnispolitik Passagen Verlag, Vienna.

Wilson, W.J., Taub, 2006. There Goes the Neighborhood: Racial, Ethnic, and Class Tensions in Four Chicago Neighborhoods and Their Meaning for America. Knopf, New York.

Wilson, W.J., 2009. More than Just Race: Being Black and Poor in the Inner City. Norton, New York.

Wilson, W.J. (Ed.), 1989. The Ghetto Underclass: Social Science Perspectives. Annals of the American Academy of Political and Social Sciences (Updated edition, 1993, Sage Publications, Newbury Park, CA.).

Wilson, W.J. (Ed.), 1993. Sociology and the Public Agenda. Sage Publications, Newbury Park, CA.

Wilson, W.J., McFate, K., Lawson, R. (Eds.), 1995. Poverty, Inequality and the Future of Social Policy: Western States in the New World Order. Russell Sage Publications, New York.

Wilson, W.J., Tienda, M. (Eds.), 2002. Youth in Cities: A Cross-national Perspective. Cambridge University Press, New York.

Wilson, W.J., 2011. The declining significance of race: revisited and revised. Daedalus 140 (2), 55–69.

Wilson, W.J., 1971. Racial relations models and ghetto behavior. In: Rose, P.I. (Ed.), Nation of Nations: The Ethnic Experience and the Racial Crisis. Random House, New York, pp. 259–278.

Wisdom, Psychology of

Judith Glück, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by U.M. Staudinger, volume 24, pp. 16510–16514, © 2011, Elsevier Ltd.

Abstract

Psychological wisdom research is a relatively young and fast-growing field. This article first gives an overview of the history of psychological wisdom research and work on laypeople's theories of wisdom. Then, psychological conceptions of wisdom are reviewed, and approaches to measuring wisdom are described. The relationships of wisdom to age, including developmental theories of wisdom, and to other psychological constructs such as intelligence and personality are reviewed, followed by approaches to fostering wisdom in the short and long run. Finally, new developments and important open questions are discussed.

Introduction

Ever since human beings began to reflect upon themselves and life in general, they have arguably been striving for wisdom – for ways to deal with the fundamental questions of being human, such as our awareness of our own mortality, and for understanding what it means to live a good life. Long before the science of psychology entered the stage, the Greek term 'philosophia' was coined to describe people's longing and search for wisdom, and questions of wisdom have been central not only to the philosophical, but also the religious sciences and other fields of the humanities for a long time (cf Curnow, 1999; Osbeck and Robinson, 2005).

Within psychology, wisdom has become a major topic only recently. Some early classics in the life-span developmental literature (e.g., Erikson, 1959; Hall, 1922) discussed wisdom as an ideal late-life outcome, but empirical wisdom research did not really begin before the 1980s. Then, wisdom research grew slowly in the context of a general trend toward differentiated conceptions of aging, but remained limited to a very small number of research groups until about 2000. In recent years, the number of theoretical conceptualizations, empirical research approaches, and fields of application has steeply increased, and currently wisdom is being investigated in many different contexts and research traditions. This interest in wisdom is probably related to social, political, and demographical developments that create broad concerns for issues of positive aging as well as for general societal and ethical questions of how human lives should be lived.

Lay Theories of Wisdom

To get a first grasp of the multifaceted and somewhat elusive concept of wisdom, early psychological wisdom research investigated laypeople's conceptions (e.g., Clayton and Birren, 1980; Sternberg, 1985; overview in Bluck and Glück, 2005). Lay conceptions are often referred to as 'implicit theories' of wisdom (Sternberg, 1985), although the methods used to investigate them tend to be rather explicit, using, for example, lists of characteristics that laypeople rate concerning their centrality for wisdom. This research derived some core components of wisdom that were also important in shaping psychologists' emerging theories. For example, Clayton and Birren (1980) identified a cognitive, reflective, and affective component of wisdom – a structure that is still seminal to psychological conceptions of wisdom (Ardelt, 2003).

Later research on lay theories mostly focused on individual differences in wisdom conceptions, e.g., across cultures (Takahashi and Overton, 2005; Yang, 2001) or age groups (Glück and Bluck, 2011). One main distinction is between cognition-oriented, analytical conceptions of wisdom that focus on life knowledge and complex thinking and more integrative conceptions that include noncognitive characteristics such as concern for others. Cross-cultural research has found that typical 'Western' conceptions of wisdom tend to emphasize cognition, whereas 'Eastern' conceptions include noncognitive aspects such as humility, empathy, and emotion regulation as equally central (Takahashi and Overton, 2005). While these differences reflect general tendencies across cultures, there are also individual differences within cultures. For example, middle-aged and older adults view affective aspects as more central to wisdom than young adults do (Glück and Bluck, 2011).

Psychological Definitions of Wisdom

With the growing number of researchers in the field, the number of definitions and operationalizations of wisdom is also increasing, with different approaches emphasizing different aspects. In the following, four current approaches are described in some detail; **Table 1** gives an overview of a larger number of conceptions.

The Three-Dimensional Wisdom Model

Monika Ardelt's Three-Dimensional Wisdom Model (Ardelt, 2003) defines wisdom as a personality characteristic that consists of three components. The cognitive component is based on the motivation and ability to understand the human condition and its core issues, especially intra- and interpersonal matters. This motivation leads individuals to acquire rich and deep factual and procedural knowledge. The reflective

Table 1 Overview of current definitions of wisdom

Authors	*Definition of wisdom*
Ardelt (2003)	Integration of cognitive (ability to understand life), reflective (taking many different perspectives), and affective (positive emotions and behavior toward others) dimensions.
Baltes and Staudinger (2000)	Expertise in the fundamental pragmatics of life entailing factual knowledge, procedural knowledge, life-span contextualism, value relativism, and acceptance and management of uncertainty.
Brown (2004)	Wisdom develops through the interaction of an orientation towards learning, relevant experiences, and interactions with others. These three factors lead the individual to learn from life, that is, to reflect and integrate experiences and apply what has been learned.
Brugman (2006)	Eudaimonic life in the face of uncertainty, involving (meta-)cognition (acknowledging uncertainty), personality and affect (emotional stability despite uncertainty), and behavior (ability to act in the face of uncertainty).
Glück and Bluck (2013)	Affective, self-related, and cognitive integrity acquired through the interaction of life experiences with four resources (openness, reflectivity, emotion regulation, and sense of mastery).
Brown and Greene (2006)	Self-knowledge, understanding of others, judgment, life knowledge, life skills, and willingness to learn.
Grossmann et al. (2010)	Six criteria for wise reasoning: perspective-shifting to take different viewpoints, recognition of the likelihood of change, flexibility of predictions, recognition of uncertainty and the limits of knowledge, search for conflict resolution, and search for compromise.
Jeste et al. (2010)	Six components of wisdom common to published definitions: prosocial attitudes/behaviors, social decision-making/pragmatic knowledge of life, emotional homeostasis, reflection/self-understanding, value relativism/tolerance, acknowledgement of and dealing effectively with uncertainty/ambiguity.
Knight and Laidlaw (2009)	Result of a self-concept and life narrative that incorporates lifespan contextualism, accumulated life experience ('knowing how'), tolerance of uncertainty, and a balancing of dialectics such as emotion and reason.
Levenson et al. (2005)	Self-transcendence, building on self-knowledge, non-attachment, and integration.
Mickler and Staudinger (2008)	Realizing one's own potential while considering the well-being of others and society, combining rich self-knowledge, heuristics of growth and self-regulation, interrelating the self, self-relativism, and tolerance of ambiguity.
Sternberg (2007)	Use of successful intelligence, creativity, and knowledge as mediated by values to seek to reach a common good by balancing inter-, intra-, and extrapersonal interests in a short- and long-term perspective by adapting to, shaping, and selecting environments.
Webster (2007)	Drawing upon life experience to facilitate the optimal development of oneself and others, entailing life experience, emotion regulation, reminiscence/reflectiveness, openness, and humor.
Yang (2008)	Integration (of ideas, interests, modes, of operation, and personality traits that are ordinarily considered as separate or conflicting), embodiment (taking action to realize one's ideals), and accomplishing positive effects brought forth by acting to the acting self and to others.

component is defined as a general willingness to look at phenomena from different sides and take multiple perspectives. Wise individuals also take different perspectives on themselves, which fosters self-criticism and self-insight. The affective component is defined as an empathetic, caring attitude toward others that extends beyond friends and family to include humankind and the world at large. Together, these three components characterize individuals who are highly interested in learning and growing from experiences and interactions and in improving not just their personal situation but the world at large. Such individuals become sources of knowledge and support for others and live rich, satisfactory lives (Ardelt, 2011a).

Wisdom as Self-Transcendence

Michael R. Levenson (Levenson et al., 2005) argues that the core of wisdom is self-transcendence. A self-transcendent individual is independent of external self-definitions and therefore able to perceive others and the world as they really are, without a need for self-enhancement or self-confirmation. Therefore, self-transcendence encompasses caring relationships and feelings of union with others. Drawing on Trevor Curnow's (1999) philosophical work on culturally universal characteristics of wisdom, Levenson argues that self-transcendence develops through three stages. Self-knowledge is a clear awareness of the internal and external sources of one's self. Detachment is an awareness of the transience and provisional nature of external sources. Integration is the acceptance and inclusion of all self-aspects, overcoming self-illusions, and accepting one's own weaknesses. Individuals who are able to integrate and accept their desirable as well as undesirable sides and the independence of this inner core from external rewards are able to develop self-transcendent relations to others and the world.

The Berlin Wisdom Model

The Berlin Wisdom Model, developed by Paul Baltes and colleagues from the 1980s on (overview in Baltes and Staudinger, 2000), was for a long time the sole 'gold standard' in conceptualizing and measuring wisdom. It defines wisdom as expertise in the fundamental pragmatics of human life: just as some people become experts in mathematics or chess, the model assumes that some individuals are motivated to learn about what it means to be human and how to deal with the difficult, existential issues of life. Through reflecting upon direct and indirect experiences with such issues, they develop wisdom-related knowledge, which is characterized by five criteria. Factual knowledge is knowledge of the nature and variability of personalities, developmental trajectories, and relationships. Procedural knowledge includes general and

specific strategies and heuristics for dealing with one's own and others' life problems. Life-span contextualism means awareness and consideration of the fact that people's views and behavior are strongly influenced by contextual factors and developmental stages. Value relativism is awareness and acceptance of individual and cultural differences in value orientations. Perhaps most importantly, recognition and management of uncertainty means an awareness of the inherent unpredictability and uncontrollability of human life and the ability to deal with it.

Wisdom as Self-Distanced Reasoning

Igor Grossmann and colleagues have recently developed a new approach to investigate wise reasoning. They define it as "the use of certain types of pragmatic reasoning to navigate important challenges of social life" (Grossmann et al., 2010, p. 7246). These types of pragmatic reasoning involve dialectical thinking and intellectual humility and manifest themselves in the ability to take different viewpoints, recognize the limitations of knowledge and the likelihood of change, make flexible predictions, and search for conflict resolution and compromise. Grossmann et al. (2010) showed that these indicators increase with age for reasoning about group conflicts and increase or stay stable for reasoning about individual conflicts. Grossmann et al. (2013) found that wise reasoning, but not intelligence, is associated with various indicators of well-being and even with longevity.

Grouping Definitions of Wisdom

The four models described above emphasize different aspects of wisdom. While Ardelt and Levenson consider wisdom as a mature stage of personality development characterized by self-reflection and self-transcendence, the Berlin model and Grossmann's approach emphasize wise forms of thinking and life knowledge. As other conceptions can be categorized along similar lines, Staudinger et al. (2005) suggested to distinguish conceptions of 'personal wisdom,' which is acquired through personal life experiences and concerns one's own life, from conceptions of 'general wisdom,' which pertains to human life in general. The existing psychological definitions of wisdom can be grouped into these two categories quite neatly (see also Staudinger and Glück, 2011). However, personal and general wisdom may also reflect different perspectives on the same phenomenon: perhaps a self-reflective way of perceiving and thinking about experiences leads individuals to develop self-transcendence and concern for others as well as certain types of general knowledge and ways of thinking (Ardelt, 2003).

Table 1 lists current psychological definitions of wisdom and their components, illustrating the commonalities and differences among the conceptions. Some commonalities at least across the majority of definitions are the following. (1) Wisdom is a quality that manifests itself in challenging situations of human life, such as interpersonal and intrapersonal conflicts, confrontation with mortality, and difficult decisions. (2) Wisdom includes a broad and deep body of knowledge, obtained through direct and indirect experience. (3) Wisdom encompasses an awareness of the multiplicity of perspectives and the uncertainty of outcomes. (4) Wisdom includes an orientation toward a larger common good. Wisdom researchers differ in how much importance they assign to these aspects, particularly (just like laypeople) with respect to concern for a larger common good, which is central to some wisdom definitions but not included at all in others.

Measuring Wisdom

The differences in definitions of wisdom could be considered largely as a matter of emphasis, but they have resulted in marked differences in how wisdom is measured. Wisdom definitions that center on reflective or affective aspects have mostly been translated into self-report scales, whereas cognition-centered definitions have generated performance measures. An overview of the most common measures is given in the following.

Self-Report Scales

A number of self-report wisdom scales have been developed and tested for reliability and validity. These include the 39-item Three-Dimensional Wisdom Scale (3D-WS; Ardelt, 2003), which measures Ardelt's cognitive, reflective, and affective dimension. The Self-Assessed Wisdom Scale (SAWS; Webster, 2007) consists of 40 items measuring openness, emotional regulation, humor, critical life experience, and reminiscence and reflectiveness. The Adult Self-Transcendence Inventory (ASTI; Levenson et al., 2005) measures wisdom as self-transcendence. The most current version consists of 34 items, 10 measuring alienation and 24 measuring self-transcendence. The Wisdom Development Scale (WDS; Brown and Greene, 2006) consists of 79 items measuring eight components: self-knowledge, altruism, leadership, judgment, life knowledge, life skills, emotional management, and willingness to learn. The 20-item Brief Wisdom Screening Scale (Glück et al., 2013) was empirically derived by selecting those items from the 3D-WS, ASTI, and SAWS that had the highest loadings on a common factor.

Advantages and Disadvantages of Self-Report Scales

The clear advantage of self-report wisdom scales is that they are easy to administer and score. The main potential problem is that participants may distort responses intentionally or unintentionally. While this is true for all self-report assessments of positive constructs, it may be particularly problematic with wisdom because wisdom entails critical self-reflection: truly wise individuals may describe themselves less positively than people with strong self-illusions. Using reverse-coded items, as is the case in the 3D-WS, may alleviate the problem, but may also have negative effects on reliability and increase cognitive load. Correlations to social desirability have been found for the 3D-WS and the SAWS (Ardelt, 2003, 2011b; Taylor et al., 2011).

Performance Measures of Wisdom

The face-validity issues with self-report scales suggest the use of performance tasks. Participants are presented with open-ended

problems that tap wisdom-related competencies, and their responses are rated with respect to wisdom-related criteria. The most prominent performance measure is certainly the Berlin Wisdom Paradigm (overview in Baltes and Staudinger, 2000). Participants are presented with brief descriptions of difficult life problems such as "A 15-year-old girl wants to move out of her family home right away" or "In reflecting over their life, people sometimes realize that they have not achieved what they had once wanted to achieve." They are asked to think aloud about what could be considered and done in such a situation. Responses are transcribed and rated with respect to five criteria: factual knowledge, procedural knowledge, life-span contextualism, value relativism, and recognition and management of uncertainty.

A performance measure of personal wisdom was developed by Mickler and Staudinger (2008). The *Bremen Wisdom Paradigm (BrW)* is in many respects parallel to the Berlin Wisdom Paradigm, but it focuses on the wisdom that participants have developed with respect to their own self and life experiences. Participants are asked to think aloud about themselves as a friend, including their behavior in difficult situations, reasons for their behavior, and their strengths and weaknesses. Responses are transcribed and rated with respect to five criteria: rich self-knowledge, heuristics of growth and self-regulation, interrelating the self, self-relativism, and tolerance of ambiguity.

Grossmann et al. (2010) developed a new paradigm for measuring wise reasoning. Participants are presented with stories about conflicts between social groups, such as ethnic tensions or the assimilation of immigrants, in foreign countries. They are interviewed about what they think will happen in those countries and why. Response transcripts are coded for perspective-shifting to take different viewpoints, recognition of the likelihood of change, flexibility of predictions, recognition of uncertainty and the limits of knowledge, search for conflict resolution, and search for compromise.

Advantages and Disadvantages of Performance Measures

All performance measures that are currently available consist of open-ended tasks, which mean that they require high efforts for transcription and coding. Attempts to develop closed-ended questions have been unsuccessful to date, probably because it is far easier to recognize than to produce a wise solution to a problem.

A potential problem with performance measures is that they may overemphasize cognitive aspects of wisdom: giving a wise response to a theoretical problem faced by a fictitious person may require wisdom-related knowledge, but it may not tap aspects of self-reflection, emotion regulation, and empathy that the same problem would require if it occurred in real life (Ardelt, 2004). Some new approaches to measuring wisdom try to increase participants' self-immersion in different ways; for example, by using videos of real conflicts (Thomas and Kunzmann, 2013) or autobiographical narratives (Glück et al., 2005). Other possibilities are to include the perspective of other persons by studying wisdom nominees (Ardelt, 2003; Baltes et al., 1995; Glück et al., 2013; Yang, 2011) or to collect peer ratings of wisdom (Redzanowski and Glück, 2013).

Relations of Wisdom to Other Variables

In the following, the relationships of wisdom to relevant psychological variables (intelligence, personality, affect and well-being, and identity-related constructs) and chronological age are reviewed.

Psychological Correlates of Wisdom

Table 2 summarizes findings about correlations of measures of wisdom to four groups of variables: intelligence, personality, affect and well-being, and identity-related constructs. Some of the constructs are conceptually so close to wisdom that they are components in some conceptions of wisdom but treated as correlates by others. For example, Webster's SAWS (Webster, 2007) includes subscales named 'openness to experience' and 'emotional regulation.'

As **Table 2** shows, the correlations between measures of wisdom and intelligence tend to be relatively low. They are somewhat higher for performance-based measures of wisdom, which may be partly due to shared method variance. Conceptually, one would expect to find positive correlations between wisdom and crystallized intelligence and no or negative correlations between wisdom and fluid intelligence, but the empirical findings only partly support this assumption. While the SAWS and Grossmann's wise reasoning measure show the expected pattern of correlations, the 3D-WS, BWP, and BrWP are positively related to fluid intelligence. It is an open question, however, whether linear correlations are a good measure for the relations of wisdom to other variables. For example, wisdom might require a certain minimum level of intelligence but be unrelated to intelligence above that level.

With respect to personality, the one Big Five dimension that has turned out to be related to wisdom in almost all studies is openness to experience. People higher in openness are more likely both to encounter new ideas, experiences, and contexts, and to learn and grow from them throughout their life. Apart from openness, there is a clear divergence between self-report measures of wisdom, which tend to be correlated to various personality dimensions, and performance measures, which are mostly uncorrelated to personality. This may be the case partly because most self-report measures are based on conceptions of wisdom that include personality aspects and partly due to shared method variance, as all studies listed in **Table 2** measured personality by self-report.

There is an ongoing controversy about the relationship between wisdom and well-being. While some authors argue that wisdom is closely related to life satisfaction and positive affect (e.g., Ardelt, 2011a), others believe that wisdom entails an awareness of the difficult and negative sides of human life that may prohibit high levels of happiness (Kunzmann and Baltes, 2003; Staudinger and Glück, 2011). Conceptually and empirically, wisdom is clearly more closely related to eudaimonic well-being, the awareness of meaning in one's own life, than to hedonic well-being, the simple enjoyment of a pleasurable life (Kunzmann and Baltes, 2003; Le, 2011; Webster et al., 2013).

Relationships to identity-related and experiential variables tend to be relatively low but consistently positive across measures of wisdom: wisdom is related to higher levels of

Table 2 Correlations of measures of wisdom to other psychological constructs

	Intelligence	*Personality*	*Affect/well-being*	*Identity/self/life experience*
3D-WS (Ardelt, 2003)	Inductive reasoning: 0.22[c] Vocabulary: n.s.[c]	Openness to experience: 0.59[c], 0.20[f] Environmental mastery: 0.40[b] Personal growth: 0.52[b], 0.41[c] Purpose in life: 0.61[a], 0.45[b] Self-acceptance: 0.49[b], 0.37[c] Autonomy: 0.41[b] Positive relations: 0.48[b]	General well-being: 0.45[a] Life satisfaction: 0.33[f], 0.16[m] Positive affect: 0.14[m] Negative affect: −0.29[m] Depressive symptoms: −0.59[a] Emotion regulation: 0.63[c], 0.29[m]	Critical life experiences: n.s.[b] Self-efficacy: 0.33[c] Self-forgiveness: 0.27[b]
ASTI (Levenson et al., 2005)	Inductive reasoning: n.s.[c] Vocabulary: n.s.[c]	Neuroticism: −0.28[g] Extraversion: 0.16[g] Openness to experience: 0.44[c], 0.16[f], 0.20[g] Agreeableness: 0.23[g] Conscientiousness: 0.20[g] Personal growth: 0.22[c] Self-acceptance: 0.33[c]	Emotion regulation: 0.50[c] Alienation: −0.25[g] Life satisfaction: n.s.[f]	Self-efficacy: 0.34[c]
SAWS (Webster, 2007)	Inductive reasoning: −0.15[c] Vocabulary: +0.18[c]	Neuroticism: −0.19[l] Extraversion: 0.35[l] Openness to experience: 0.41[c], 0.46[l] Environmental mastery: 0.17[b] Personal growth: 0.51[b], 0.28[c], 0.55[k] Purpose in life: 0.24[b] Self-acceptance: 0.43[b], 0.17[c] Autonomy: 0.32[b] Positive relations: 0.34[b]	Emotion regulation: 0.32[c] Hedonistic mental health: 0.30[l] Eudaimonic mental health: 0.44[l]	Self-efficacy: 0.38[c] Self-forgiveness: 0.25[b] Ego integrity: 0.23[j] 0.45[k] Generativity: 0.44[j] Pleasurable-life orientation: n.s.[k]
BWP (Baltes and Staudinger, 2000)	Inductive reasoning: n.s.[c], 0.17[h], 0.28[i] Vocabulary: n.s.[c], 0.33[h], 0.34[i]	Openness to experience: 0.37[c], 0.24[h], 0.42[i] Conscientiousness: n.s.[h] Agreeableness: n.s.[h]	Life satisfaction: 0.20[h] Emotion regulation: 0.28[c] Affective involvement: 0.28[e] Pleasant affect: −0.17[e], n.s.[h] Negative affect: −0.13[e], n.s.[h] Subjective well-being: n.s.[h]	Self-efficacy: n.s.[c] Life-insight orientation: 0.23[e] Pleasurable-life orientation: −0.42[e] Ego development: n.s.[h] Self-concept maturity: 0.19[h] Life events reflected: n.s.[h]
BrWP (Mickler and Staudinger, 2008)	Inductive reasoning: 0.30[h] Vocabulary: 0.24[h]	Openness to experience: 0.27[h] Conscientiousness: n.s.[h] Agreeableness: n.s.[h]	Life satisfaction: n.s.[h] Negative emotions: n.s.[h] Positive emotions: n.s.[h] Subjective well-being: n.s.[h]	Ego development: 0.26[h] Self-concept maturity: 0.28[h] Life events reflected: 0.19[h]
Grossmann wise reasoning	Processing speed: −0.25[d] Vocabulary: 0.25[d] Digit span: n.s.[d] Comprehension: n.s.[d]	Neuroticism: n.s.[d] Extraversion: n.s.[d] Openness to experience: n.s.[d] Agreeableness: 0.27[d] Conscientiousness: n.s.[d]	Life satisfaction: 0.17[d] Positive affect: n.s.[d] Negative affect: −0.27[d] Depressive brooding: −0.33[d]	

[a]Ardelt, M., 2003. Development and empirical assessment of a three-dimensional wisdom scale. Research on Aging 25, 275–324.
[b]Ardelt, M., 2011b. The measurement of wisdom: a commentary on Taylor, Bates, and Webster's comparison of the SAWS and 3D-WS. Experimental Aging Research 37, 241–255.
[c]Glück, J., König, S., Naschenweng, K., Redzanowski, U., Dorner, L., Strasser, I., Wiedermann, W., 2013. How to measure wisdom: content, reliability, and validity of five measures. Frontiers in Psychology 4. http://www.frontiersin.org/personality_science_and_individual_differences/10.3389/fpsyg.2013.00405/abstract.
[d]Grossmann, I., Na, J., Varnum, M.E.W., Kitayama, S., Nisbett, R.E., 2013. A route to well-being: intelligence versus wise reasoning. Journal of Experimental Psychology: General, 142, 944–953.
[e]Kunzmann, U., Baltes, P., 2003. Wisdom-related knowledge: affective, motivational, and interpersonal correlates. Personality and Social Psychology Bulletin 29, 1104–1119.
[f]Le, T.N., 2011. Life satisfaction, openness value, self-transcendence, and wisdom. Journal of Happiness Studies 12, 171–182.
[g]Levenson, R., Jennings, P.A., Aldwin, C., Shiraishi, R.W., 2005. Self-transcendence, conceptualization and measurement. International Journal of Aging and Human Development 60, 127–143.
[h]Mickler, C., Staudinger, U.M., 2008. Personal wisdom: validation and age-related differences of a performance measure. Psychology and Aging 23, 787–799.
[i]Staudinger, U.M., Lopez, D., Baltes, P.B., 1997. The psychometric location of wisdom-related performance: intelligence, personality, and more? Personality and Social Psychology Bulletin 23, 1200–1214.
[j]Webster (2003).
[k]Webster (2010).
[l]Webster, J.D., Westerhof, G.J., Bohlmeijer, E.T., 2013. Wisdom and mental health across the lifespan. The Journals of Gerontology, Series B: Psychological Sciences.
[m]Zacher, H., McKenna, B., Rooney, D., 2013. Effects of self-reported wisdom on happiness: not much more than emotional intelligence? Journal of Happiness Studies 14(6), 1697–1716.

identity development, as manifested in variables such as ego development, integrity, and generativity, and in an orientation toward personal growth and learning. Interestingly, correlations between life experiences and wisdom have not been investigated very often, but they tend to be low to insignificant. These findings support the notion that wisdom is not fostered by any particular experiences per se but by the way individuals deal with them, reflect upon them and their own role in them, and integrate them into their life story (Glück and Bluck, 2013).

Wisdom and Age

How is wisdom related to age? Laypeople (Bluck and Glück, 2005; Glück and Bluck, 2011) and wisdom researchers (Jeste et al., 2010; Staudinger, 1999) seem to agree in two points: First, wise people tend to be old rather than young, but second, wisdom does not generally come with age. These two seemingly contradictory statements can be reconciled by assuming that wisdom is a relatively rare phenomenon that emerges in only a small number of individuals – those who have the necessary cognitive, affective, and motivational resources and encounter contexts and experiences that foster learning and growth. Empirically, relations to age differ across measures of wisdom. The Berlin Wisdom Paradigm is unrelated to chronological age in adulthood (Staudinger, 1999; Pasupathi et al., 2001). Grossmann's measure of wise reasoning (Grossmann et al., 2010) is positively related to age, whereas the Bremen Personal-Wisdom Paradigm and the 3D-WS are negatively related to age. It may be worthwhile to also look for nonlinear relations, as recent work has found an inverse U-shaped relation of the SAWS with age (Webster et al., 2013). How a particular measure is related to age probably depends on which facets of wisdom it emphasizes. While some aspects, such as relativistic thinking, affective positivity, or accumulated life experience may increase with age, others, such as fluid intelligence or self-concept complexity, decrease in older age. Wisdom is a relatively rare optimal constellation of these facets, which implies that wise individuals may be able to compensate losses in some domains better than other people do. For example, the idea of wisdom as expertise suggests that the development of wisdom involves the acquisition of strategies and heuristics that can overcome cognitive limitations resulting from losses in fluid intelligence in old age.

Theories of the Development of Wisdom

Why do a few individuals develop high levels of wisdom in the course of their lives, while so many others, faced with similar experiences, do not seem to learn or grow from them? The fact that high levels of wisdom are rare suggests that the development of wisdom requires specific constellations of resources and experiences. In the context of the Berlin Wisdom Paradigm, Baltes and Staudinger (2000) suggested that the development of wisdom involves characteristics of the individual such as intelligence, creativity, or openness to experience; expertise-specific factors such as the availability of mentors or a motivational disposition to strive for wisdom; and relevant experiential contexts such as parenthood, certain professional roles, or old age.

Somewhat more specifically, the MORE Life Experience Model of the development of wisdom (Glück and Bluck, 2013) proposes that wisdom develops through an interaction of psychological resources with challenging life experiences. The four most relevant resources are Mastery, Openness, Reflectivity, and Emotion Regulation/Empathy (MORE). They influence what life challenges people encounter and how they appraise them, how they deal with challenges, and how they integrate past challenges in their life story. Individuals who are higher in the MORE resources are more likely to deal with a life-changing experience in a way that fosters both the resources themselves and learning and growth toward wisdom.

Fostering Wisdom

How can individual and societal wisdom be fostered? This question has obvious societal relevance. Psychological work suggests that it is indeed possible ways to enhance wisdom by targeted interventions. Two different lines of theoretical and empirical work focus on long-term and short-term interventions. Long-term interventions are mostly discussed with respect to educational systems: are there ways in which schools or universities could teach for wisdom (Brown, 2004; Ferrari and Potworowski, 2008; Sternberg, 2001)? Although little empirical research has been published, it seems likely that especially cognitive facets of wisdom, such as relativistic thinking and critical self-reflection, can be practiced in educational settings.

Another kind of long-term intervention that may foster wisdom is psychotherapy. Generally, it could be argued that any psychotherapy aims at improving wisdom-related capacities such as self-reflection, emotion regulation, and empathy, and it would certainly be worthwhile to investigate the effects of psychotherapy on wisdom. More specifically, a 'wisdom therapy' has been developed for treating embittered participants (Linden et al., 2011). Based largely on the Berlin wisdom model, this approach focuses on a cognitive reframing of the life experiences that caused embitterment by taking different perspectives and roles and considering different persons, contexts, and values.

Short-term wisdom interventions have been investigated experimentally both by the Berlin wisdom group and by Igor Grossmann and colleagues. All wisdom interventions that have proven successful to date instructed participants to transcend their own viewpoints: Wisdom-related knowledge can be fostered by imagining to discuss the task with a close other person (Staudinger and Baltes, 1996), as well as by taking a distanced rather than self-immersed perspective on an issue (Kross and Grossmann, 2012). Interestingly, an instruction to "try to give a wise response" elicited higher levels of wisdom-related knowledge only in participants who were relatively wise to begin with (Glück and Baltes, 2006).

These findings also imply a certain degree of state-dependence of wisdom: if a simple instruction is sufficient to elicit higher levels of wisdom in people than they usually show, the question arises to what degree situational contexts in general are conducive or obstructive to wisdom. For example, high emotional involvement in a situation may have negative effects on wisdom.

Current Developments and Important Next Steps

Current developments of wisdom research take two different directions. On the one hand, wisdom research is applied in societally relevant contexts such as leadership or medical ethics. On the other hand, the conceptions and paradigms of genuine wisdom research are currently expanding in several new directions.

The Societal Relevance of Wisdom Research

When laypeople are asked to nominate wise figures, many nominees (e.g., Mahatma Gandhi, Jesus Christ, Martin Luther King, or Nelson Mandela; Paulhus et al., 2002) fit into our idea of 'leaders,' that is, they have inspired millions of followers and improved the lives of people fundamentally. In this vein, it has been argued that wisdom is a crucially important characteristic of economical and political leaders, although for the time being, there is more theoretical than empirical work on wise leadership (e.g., Kilburg, 2006; McKenna et al., 2009; Sternberg, 2007; Yang, 2011). However, the idea that if only politicians and managers were wiser, the world would be a better place may be too simple. For example, the wisdom of managers is not independent of organizational culture (Limas and Hansson, 2004). The goals and structural requirements of leading a state or a big company may not or only partly be compatible with characteristics of wisdom such as striving for a common good rather than maximizing profit (Staudinger and Glück, 2011).

Wisdom has also been investigated in the context of medical ethics, which touches upon some of the fundamental issues of human existence. For example, Barilan (2009) argues that a critical and reflective moral attitude contributes to wise decisions in the context of truth-telling to terminally ill patients. Kaldijan (2010) discusses wise clinical judgment on the basis of philosophical conceptions of practical wisdom.

New Developments in Psychological Wisdom Research

As mentioned before, psychological wisdom research is a fast-growing field. New paradigms are developed for measuring wisdom (e.g., Grossmann et al., 2010; Redzanowski and Glück, 2013; Thomas and Kunzmann, 2013), fostering wisdom (e.g., Kross and Grossmann, 2012; Linden et al., 2011), and analyzing data beyond linear correlations (e.g., Webster et al., 2013). We are learning more about the development of wisdom (e.g., Glück and Bluck, 2013; Thomas and Kunzmann, 2014). Still, a large number of important questions are still open. First, most research has been conducted in modern Western societies – which aspects of our conceptions of wisdom are universal, and which are dependent on cultural, historical, and societal contexts? Second, while intraindividual predictors of wisdom such as personality or intelligence are well-studied, little research has investigated the role of social and societal resources for the development and manifestation of wisdom. Third, how can wisdom be measured in more ecologically valid ways? We hope that these questions and many others will be tackled by the wisdom researchers of the next decades.

See also: Coping across the Lifespan; Coping, Personality and; Emotional Regulation; Emotions and Aging; Erikson's Theory of Psychosocial Development; Intelligence: Central Conceptions and Psychometric Models; Leadership; Openness to Experience; Personality Assessment; Personality, Trait Models of; Self and Emotional Development in Adulthood and Later Life; Subjective Wellbeing, Psychology of.

Bibliography

Ardelt, M., 2003. Development and empirical assessment of a three-dimensional wisdom scale. Research on Aging 25, 275–324.

Ardelt, M., 2004. Wisdom as expert knowledge system: a critical review of a contemporary operationalization of an ancient concept. Human Development 47, 257–285.

Ardelt, M., 2011a. Wisdom, age, and well-being. In: Schaie, K.W., Willis, S.L. (Eds.), Handbook of the Psychology of Aging, seventh ed. Elsevier, San Diego, CA, pp. 279–291.

Ardelt, M., 2011b. The measurement of wisdom: a commentary on Taylor, Bates, and Webster's comparison of the SAWS and 3D-WS. Experimental Aging Research 37, 241–255.

Baltes, P.B., Staudinger, U.M., 2000. Wisdom: a metaheuristic (pragmatic) to orchestrate mind and virtue towards excellence. American Psychologist 55, 122–136.

Baltes, P.B., Staudinger, U.M., Maercker, A., Smith, J., 1995. People nominated as wise: a comparative study of wisdom-related knowledge. Psychology and Aging 10, 155–166.

Barilan, Y.M., 2009. Responsibility as a meta-virtue: truth-telling, deliberation and wisdom in medical professionalism. Journal of Medical Ethics 35, 153–158.

Bluck, S., Glück, J., 2005. From the inside out: people's implicit theories of wisdom. In: Sternberg, R.J., Jordan, J. (Eds.), A Handbook of Wisdom: Psychological Perspectives. Cambridge University Press, Cambridge, pp. 84–109.

Brown, S.C., Greene, J.A., 2006. The wisdom development scale: translating the conceptual to the concrete. Journal of College Student Development 47, 1–19.

Brown, S.C., 2004. Learning across campus: how college facilitates the development of wisdom. Journal of College Student Development 45, 375–390.

Brugman, G.M., 2006. Wisdom and aging. In: Birren, J.E., Schaie, K.W. (Eds.), Handbook of the Psychology of Aging, sixth ed. Elsevier, San Diego, CA, pp. 445–475.

Clayton, V.P., Birren, J.E., 1980. The development of wisdom across the lifespan: a reexamination of an ancient topic. In: Baltes, P.B., Brim, O.G. (Eds.), Life-Span Development and Behavior, vol. 3. Academic Press, San Diego, CA, pp. 103–135.

Curnow, T., 1999. Wisdom, Intuition, and Ethics. Ashgate, Aldershot, UK.

Erikson, E.H., 1959. Identity and the Life Cycle. International University Press, New York.

Ferrari, M., Potworowski, G., 2008. Teaching for Wisdom: Cross-Cultural Perspectives on Fostering Wisdom. Springer, New York.

Glück, J., Baltes, P.B., 2006. Using the concept of wisdom to enhance the expression of wisdom knowledge: not the philosopher's dream, but differential effects of developmental preparedness. Psychology and Aging 21, 679–690.

Glück, J., Bluck, S., 2011. Laypeople's conceptions of wisdom and its development: cognitive and integrative views. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 66B, 321–324.

Glück, J., Bluck, S., 2013. The MORE life experience model: a developmental theory of personal wisdom. In: Ferrari, M., Weststrate, N. (Eds.), The Scientific Study of Personal Wisdom. Springer, New York, pp. 75–98.

Glück, J., Bluck, S., Baron, J., McAdams, D., 2005. The wisdom of experience: autobiographical narratives across adulthood. International Journal of Behavioral Development 29, 197–208.

Glück, J., König, S., Naschenweng, K., Redzanowski, U., Dorner, L., Strasser, I., Wiedermann, W., 2013. How to measure wisdom: content, reliability, and validity of five measures. Frontiers in Psychology 4. http://www.frontiersin.org/personality_science_and_individual_differences/10.3389/fpsyg.2013.00405/abstract.

Grossmann, I., Na, J., Varnum, M.E.W., Kitayama, S., Nisbett, R.E., 2013. A route to well-being: intelligence versus wise reasoning. Journal of Experimental Psychology: General 142, 944–953.

Grossmann, I., Na, J., Varnum, M.E.W., Park, D.C., Kitayama, S., Nisbett, R.E., 2010. Reasoning about social conflicts improves into old age. PNAS Proceedings of the National Academy of Sciences of the United States of America 107, 7246–7250.

Hall, G.S., 1922. Senescence: The Last Half of Life. Appleton, New York.

Jeste, D.V., Ardelt, M., Blazer, D., Kraemer, H.C., Vaillant, G., Meeks, T., 2010. Expert consensus on the characteristics of wisdom: a Delphi method study. The Gerontologist 50, 668–680.

Kaldjian, L.C., 2010. Teaching practical wisdom in medicine through clinical judgement, goals of care, and ethical reasoning. Journal of Medical Ethics: Journal of the Institute of Medical Ethics 36, 558–562.

Kilburg, R.R., 2006. Executive Wisdom: Coaching and the Emergence of Virtuous Leaders. American Psychological Association, Washington.

Knight, B.G., Laidlaw, K., 2009. Translational theory: a wisdom-based model for psychological interventions to enhance well-being in later life. In: Bengston, V., Silverstein, M., Putney, N.M., Gans, D. (Eds.), Handbook of Theories of Aging, second ed. Springer, New York, pp. 693–705.

Kross, E., Grossmann, I., 2012. Boosting wisdom: distance from the self enhances wise reasoning, attitudes, and behavior. Journal of Experimental Psychology: General 141, 43–48.

Kunzmann, U., Baltes, P., 2003. Wisdom-related knowledge: affective, motivational, and interpersonal correlates. Personality and Social Psychology Bulletin 29, 1104–1119.

Le, T.N., 2011. Life satisfaction, openness value, self-transcendence, and wisdom. Journal of Happiness Studies 12, 171–182.

Levenson, R., Jennings, P.A., Aldwin, C., Shiraishi, R.W., 2005. Self-transcendence, conceptualization and measurement. International Journal of Aging and Human Development 60, 127–143.

Limas, M.J., Hansson, R.O., 2004. Organizational wisdom. International Journal of Aging and Human Development 59, 65–103.

Linden, M., Baumann, K., Lieberei, B., Lorenz, C., Rotter, M., 2011. Treatment of posttraumatic embitterment disorder with cognitive behaviour therapy based on wisdom psychology and hedonia strategies. Psychotherapy and Psychosomatics 80, 199–205.

McKenna, B., Rooney, D., Boal, K.B., 2009. Wisdom principles as a meta-theoretical basis for evaluating leadership. The Leadership Quarterly 20, 177–190.

Mickler, C., Staudinger, U.M., 2008. Personal wisdom: validation and age-related differences of a performance measure. Psychology and Aging 23, 787–799.

Osbeck, L., Robinson, D., 2005. Philosophical theories of wisdom. In: Sternberg, R., Jordan, J. (Eds.), A Handbook of Wisdom: Psychological Perspectives. Cambridge University Press, Cambridge, pp. 61–83.

Pasupathi, M., Staudinger, U.M., Baltes, P.B., 2001. Seeds of wisdom: adolescents' knowledge and judgment about difficult life problems. Developmental Psychology 37, 351–361.

Paulhus, D.L., Wehr, P., Harms, P.D., Strasser, D.I., 2002. Use of exemplar surveys to reveal implicit types of intelligence. Personality and Social Psychology Bulletin 28, 1051–1062.

Redzanowski, U., Glück, J., 2013. Who knows who is wise? Self- and peer-ratings of wisdom. The Journals of Gerontology, Series B: Psychological Sciences 68, 391–394.

Staudinger, U.M., 1999. Older and wiser? Integrating results on the relationship between age and wisdom-related performance. International Journal of Behavioral Development 23, 641–664.

Staudinger, U.M., Baltes, P.B., 1996. Interactive minds: a facilitative setting for wisdom-related performance? Journal of Personality and Social Psychology 71, 746–762.

Staudinger, U.M., Dörner, J., Mickler, C., 2005. Wisdom and personality. In: Sternberg, R.J., Jordan, J. (Eds.), A Handbook of Wisdom: Psychological Perspectives. Cambridge University Press, New York, pp. 191–219.

Staudinger, U.M., Glück, J., 2011. Psychological wisdom research: commonalities and differences in a growing field. Annual Review of Psychology 62, 215–241.

Staudinger, U.M., Lopez, D., Baltes, P.B., 1997. The psychometric location of wisdom-related performance: intelligence, personality, and more? Personality and Social Psychology Bulletin 23, 1200–1214.

Sternberg, R.J., 1985. Implicit theories of intelligence, creativity, and wisdom. Journal of Personality and Social Psychology 49, 607–627.

Sternberg, R.J., 2001. Why schools should teach for wisdom: the balance theory of wisdom in educational settings. Educational Psychologist 36, 227–245.

Sternberg, R.J., 2007. A systems model of leadership: WICS. American Psychologist 62, 34–42.

Takahashi, M., Overton, W.F., 2005. Cultural foundations of wisdom: an integrated developmental approach. In: Sternberg, R.J., Jordan, J. (Eds.), A Handbook of Wisdom: Psychological Perspectives. Cambridge University Press, Cambridge, pp. 32–60.

Taylor, M., Bates, G., Webster, J.D., 2011. Comparing the psychometric properties of two measures of wisdom: predicting forgiveness and psychological well-being with the self-assessed wisdom scale (SAWS) and the three-dimensional wisdom scale (3D-WS). Experimental Aging Research 37, 129–141.

Thomas, S., Kunzmann, U., 2013. Age differences in wisdom-related knowledge: does the age relevance of the task matter? The Journals of Gerontology, Series B: Psychological Sciences 69, 897–905.

Thomas, S., Kunzmann, U., 2014. Age differences in wisdom-related knowledge: does the age relevance of the task matter? The Journals of Gerontology Series B: Psychological Sciences 69, 897–905.

Webster, J.D., 2003. An exploratory analysis of a self-assessed wisdom scale. Journal of Adult Development 10, 13–22.

Webster, J.D., 2007. Measuring the character strength of wisdom. International Journal of Aging & Human Development 65, 163–183.

Webster, J.D., 2010. Wisdom and positive psychosocial values in young adulthood. Journal of Adult Development 17, 70–80.

Webster, J.D., Westerhof, G.J., Bohlmeijer, E.T., 2013. Wisdom and mental health across the lifespan. The Journals of Gerontology, Series B: Psychological Sciences 69, 209–218.

Yang, S.-Y., 2001. Conceptions of wisdom among Taiwanese Chinese. Journal of Cross-Cultural Psychology 32, 662–680.

Yang, S.-Y., 2008. Real-life contextual manifestations of wisdom. The International Journal of Aging and Human Development 67, 273–303.

Yang, S.-Y., 2011. Wisdom displayed through leadership: exploring leadership-related wisdom. The Leadership Quarterly 22, 616–632.

Zacher, H., McKenna, B., Rooney, D., 2013. Effects of self-reported wisdom on happiness: not much more than emotional intelligence? Journal of Happiness Studies 14 (6), 1697–1716.

Relevant Websites

http://wisdomresearch.org – Wisdom Research. The University of Chicago, USA.
http://www.wisdompage.com/ – The Wisdom Page.

Wisdom: Philosophical Aspects

Wolfgang Welsch, Friedrich Schiller University, Jena, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The first part of the article retraces shifts in the understanding of wisdom from antiquity to modern times. Having originally, as with the Seven Sages of Greece, betokened practical–political expertise, 'wisdom' for Plato became a specifically theoretical notion, and for Aristotle the highest praxis as philosophy. In Hellenistic and Roman philosophy, the sage was once again conceived of as an expert in practical matters of human life. In modern times, when philosophy came to be associated increasingly with science, wisdom still remained important on the practical side but was also criticized more and more. The second part of the article focuses on a contemporary understanding of wisdom. It is shown that particularly concepts in recent theoretical – not practical – discourse exhibit traits fruitful for wisdom. Nietzsche, Quine, and Wittgenstein, among others provide elements suitable for a new, epistemologically inspired, understanding of wisdom centered around circumspection, awareness of plurality, limits of justification, and competence in dealing with uncertainty. From this a renewal of philosophy's role as a powerful source of wisdom could result.

Wisdom is not a central topic of contemporary philosophy. Today's philosophers seem to confirm Nietzsche's observation: "When philosophers meet among themselves they start casting off all sorts of wonderful rubbish; above all … they hang up 'the love of wisdom' like stuffy robes of office" (Nietzsche, 1980a: p. 511). Common sense, on the contrary, still expects wisdom to be provided by philosophy, and the discipline's social reputation is partly based on this expectation. But as contemporary philosophy seems scarcely inclined or able to fulfill this demand, common sense draws on forms of wisdom from nonoccidental cultures or from esoterics. While 'wisdom' is a term rarely used in contemporary philosophy, the literature about wisdom from the esoteric and the anthropological–cultural side is flourishing.

There is nothing wrong with sensing deficiencies in the academic field and turning to other sources instead. Yet in the second part of this article, it will be demonstrated that contemporary philosophy can in fact provide a great many elements of wisdom appropriate for our times. The first part is devoted to an accurate reconstruction of the main historical turns in the philosophical understanding of wisdom.

Philosophy – A Concept Introduced by Plato

As is commonly known, philosophy is related through its name to wisdom: 'philo-sophy' (in Greek: *philo-sophia*) means 'love of wisdom' or 'quest for wisdom' (in Greek: *sophia*). But this concept of *philosophia* corresponds neither to the original understanding of philosophical thinking nor to that of wisdom. The concept of *philosophia* was first created by Plato in the first half of the fourth century BC.

The Pre-Platonic Understanding of Philosophy and Wisdom

Philosophers before Plato neither deployed the term 'philosophy' nor described their enterprise with respect to wisdom (*sophia*). Rather, they characterized their doctrines as 'histories.' Moreover, the Greek understanding of wisdom – as paradigmatically expressed in the topical talk of the 'Seven Sages of Greece' – focused on practical, not theoretical knowledge. In everyday language, a wise person was just an expert of some kind or other (a craftsman, artist, doctor, military leader, etc.); in order to be included in the catalog of the Seven Sages his work had, in addition, to be of extraordinary political relevance. So Thales, the only ancient philosopher ever to figure among the Seven Sages, was included in this group not as a philosopher, but as an astronomer and engineer who brought military success to his community by predicting an eclipse and redirecting a river.

Plato's Turn

What, then, is the specific sense of Plato's introduction of the concept of *philo-sophia* – an innovation so successful that his concept is commonly taken to be *the* concept of philosophy altogether?

Three shifts are characteristic. First, the emphasis in the understanding of *sophia* as well as *philosophia* is put on theory instead of praxis. Philosophers observe the nature of things not for the sake of practical usage but for that of pure cognition. The new concept of philosophy is strictly opposed to the old practical aims of *sophia*. Second, a border line is drawn between philosophy and wisdom: philosophers can only *strive* for wisdom – without ever actually reaching wisdom. This is why philosophy is defined as *philo*-sophy (*love* of wisdom), not as wisdom proper (*sophia*). 'Philosophy' implies an element of modesty and limitation. Thirdly, this limitation cannot be overcome – complete wisdom is reserved for the Gods only. By philosophizing we do not surpass philosophy to actual wisdom.

Strategic Implications

Plato's innovation was directed not only against the traditional and the everyday understanding of wisdom but above all against his contemporary competitors, the Sophists. In everyday language at Plato's time – where compositions with 'philo' had

 http://dx.doi.org/10.1016/B978-0-08-097086-8.63094-5

long since become popular – 'philosophizing' (*philosophein*) was a term without any extravagant meaning; it just meant studying, educating oneself, being inclined to reflective activity. So Plato's declaration that wisdom is impossible for humans and reserved to the Gods alone discredited not only the mythical Seven Sages but also his compatriots' devotion to philosophical activity and above all the Sophists, who at the time of Plato roved from city to city offering, for payment, lessons in rhetoric, virtue, poetry, mathematics, music, or astronomy. As their name (*Sophists*) indicates, they were considered to be bearers of wisdom (*sophia*). The idea that they were in fact frauds – Plato was constantly trying to discredit the Sophists as dealers in fake knowledge, and as being the morally and intellectually dubious counterparts of the real philosopher – was the logical result of Plato's redefinition: If wisdom is altogether unattainable for humans, then the Sophists who masquerade as teachers of wisdom, must obviously be frauds.

Philosophy's Claim to Represent True Praxis and Wisdom

The introduction of modesty and limitations in matters of wisdom with the new concept of philosophy was an efficient strategy in distancing the Sophists, but once this was achieved the proponents of the new concept soon turned to more ambitious claims. Philosophy was successively considered capable of representing the utmost fulfillment of both practical and theoretical wisdom.

Philosophical Theory as the Highest Praxis

The move from a practical to a theoretical emphasis was meant as an empowerment of philosophy in matters of true praxis. The traditional concept of praxis, so characteristic of the original understanding of wisdom, was criticized for having in fact provided only a concept of pragmatic utility, one ignoring the very kind of praxis pertaining to the full accomplishment of human life. This higher and true sense of praxis was to be fulfilled precisely by philosophy in the new sense of theoretical activity. This tendency is obvious already in Plato, who always conceived of philosophy as the best way of human life, with respect to both the philosophizing individual and societies complying with philosophical insights. And Aristotle (in his emphatic eulogy of theory in *Nicomachean Ethics* X: pp. 7–9) finally openly declared the theoretical life to be the best and highest form of human praxis and life altogether. Significantly enough, he explained the supremacy of theoretical life by contrasting its self-sufficiency with the purposeful orientation of ordinary practical life: according to him, philosophical theory is the only activity "loved for its own sake; for nothing arises from it apart from the contemplating, while from practical activities we gain more or less apart from the action."

Philosophy as Wisdom Proper

Although Plato had initially declared wisdom to be reserved for the Gods, he made the border between human philosophy and divine wisdom permeable. Ultimately philosophizing should result in a *homoiosis theo*, a "becoming like the divine" (Plato, *Theiatetos*, 176 b 1 f). Aristotle definitely dropped the Platonic reserve. According to him, philosophy and wisdom are one – his appraisal of philosophical *theoria* in *Nicomachean Ethics* is one of *sophia*. In theoretical life, he says, we realize "the divine in our midst" and hence our "true Self." With Aristotle, '*philosophia*' is no longer a modest term inferior to *sophia*. Hegel's phrase that philosophy should "lay aside the title "*love* of knowing" and be *actual* knowing" (Hegel, 1977: p. 3) is Aristotelean in spirit.

Reaccentuating the Practical

Theory's Open Flank

But philosophy's effort to completely absorb the older practical aspects of wisdom has an open flank. In his theoretical attitude the philosopher may well possess the utmost theoretical knowledge – but this is not to possess all kinds of knowledge: genuinely practical insight, ethical knowledge, and situational prudence (*phronesis*) are lacking. And although theory may represent the highest form of praxis, it cannot as such replace all forms of praxis. Hence, specifically practical knowledge – lying outside theoretical *sophia* – requires respect and clarification in addition. This applies all the more, given that practical knowledge is (as, quite properly, Aristotle was the first to demonstrate) of an intrinsically different type from theoretical knowledge. So the new concept of philosophical theory, as impressive as it may be, remains in need of supplementation. Genuinely practical knowledge (which is so decisive for all our everyday orientation) requires a different approach.

Hellenistic and Roman Philosophy: The Return of the Sage as a Practical Expert

Following Greek philosophy's classical period, a return to conceiving the sage as an expert in practical matters took place in Hellenistic and Roman philosophy (with the Epicureans and Stoics). The sage once again became a figure of outstanding practical knowledge and exemplariness, knowing how to lead the best life. From now on wisdom was seen as the archetype of accomplished living. Theoretical capacities were to serve this goal.

This understanding of wisdom prevailed even in modern times. So Montaigne defined the sage as someone who appreciates that "our great and glorious masterpiece is to live appropriately" (Montaigne, 1958: p. 851). Kant too declared that "the teacher of wisdom through doctrine and example," "the practical philosopher," "is the real philosopher" (Kant, 1992: p. 537). Even today, we still expect a wise person to embody the right mode of human living.

Middle Ages

The Christian philosophy of the Middle Ages had trouble with wisdom. In the Bible, the wisdom of this world had been declared to be foolishness before God (*Paul*, 1 *Cor*. 3:19; similar 1 *Cor*. 1:20 and *Romans* 1:22). This position, by consequence, called into question the very possibility of reasonable reflection and hence of Christian philosophy altogether. At the least it imposed severe limitations. Some philosophers tried to settle the conflict by developing the doctrine of 'double truth.' But distrust of purely

human knowledge – be it wisdom, reason, or science – prevailed. Luther was later to renew the accusation that reason is "the devil's whore" (Luther, 1964: p. 164, cf also Luther, 1967: p. 82).

Philosophy in League with Science – Not with Wisdom

In modern times, philosophy's ally has been science, not wisdom. When philosophers, occasionally, do still use the term 'wisdom' what they actually have in mind is science. So Descartes uses 'wisdom' as just another name for the sum total of science (Descartes, 1998: p. 65). Even Kant, though affording wisdom the highest rank, emphasizes that "science … is the narrow gate that leads to the *doctrine of wisdom*" (Kant, 1993: p. 171). Husserl was to resume this modern emphasis on science instead of wisdom by declaring "science has spoken, from now on wisdom has the task of learning" (Husserl, 1965: p. 63).

Modern Philosophy's Practical Intentions

But modern philosophy has a strong practical side too. It comes in two flavors: as technical purposiveness along the lines of science, and as moral–practical orientation taking up older elements of wisdom.

Technical–Practical Side

From the start modern science was addressed to technical ends. It was in this respect that Bacon and Descartes commended it. Marx then claimed that philosophy should become altogether practical – and thus realize itself by vanishing as philosophy. Even if this project has failed, the prospect of technically reshaping reality has proven widely successful.

Moral–Practical Side

Complementary to this scientific–technical strain, to which wisdom is irrelevant, practical aspects of wisdom survive in modern philosophy's other strain, the practical one devoted to ethics and morals. It is here, if anywhere, that 'wisdom' can be a substantial term in modern times. Descartes had already conceived of ethics as "the final stage of wisdom" (Descartes, 1983: p. XXIV), and Kant, having defined wisdom as "the idea of the practical use of reason's complete conformity to law" (Kant, 1983), afforded this practical wisdom the highest rank, declaring "the practical philosopher" to be "the real philosopher" (Kant, 1992: p. 537). In the moral–practical respect wisdom retains the highest esteem.

Enlightenment philosophers, as well as their opponents, agreed that true morality – and by consequence wisdom in its practical sense – is the most important thing for us humans. When Rousseau, in contrast to the 'rationalists,' emphasized the simplicity of moral knowledge, even this claim to simplicity was shared by a great many philosophers of the period. In this they could even see the great advantage of practical over theoretical knowledge: "It took centuries to know part of nature's laws. A single day suffices to recognize man's duties" (Voltaire, 1967: p. 342). With respect to this simplicity distinguishing practical wisdom, Wolff, D'Holbach, and Voltaire proclaimed Confucius, the exemplary 'simple sage,' to have been the wisest person ever.

But obviously this modern line of revalidating practical wisdom was not very fortunate. Why else would we feel disoriented and lament the absence of wisdom appropriate to our times and problems?

Protest against Wisdom – Nietzsche

In the course of time modern philosophers came to protest more and more against all traditional understanding of wisdom. Nietzsche is the most prominent example.

He generally suspected the traditional ideal of wisdom to have been simply a symptom of weakness. The old philosophers' "claim that they possessed wisdom" appears to him to have always been "a screen behind which the philosopher saves himself because he has become weary, old, cold, hard – as a premonition that the end is near, like the prudence animals have before they die: they go off by themselves, become still, choose solitude, hide in caves, and become *wise*" (Nietzsche, 1974: p. 315). The Stoic is the paradigm case for Nietzsche. His wisdom consists in making himself insensible of the tensions of life and acquiring an armor-plated attitude. Hence Nietzsche's ruthless formula: "The stoic type. Or: the perfect ox" (Nietzsche, 1980b: p. 125).

The attitude of the 'genuine philosopher' would, in Nietzsche's opinion, be completely different: he "lives "unphilosophically" and "unwisely," above all imprudently, and feels the burden and the duty of a hundred attempts and temptations of life—he risks himself constantly, he plays the wicked game" (Nietzsche, 1989: p. 125). "The wisest person would be the richest in contradictions, who has, at it were, organs of touch for all kinds of person: and in their midst his great moments of grandiose harmony – in us too the lofty coincidence!" (Nietzsche, 1980a: p. 182).

Ultimately, however, 'wisdom' remains an improper expression in Nietzsche's view – it is too strongly impregnated by its traditional meanings. Only in a critical manner can we still refer to or draw on wisdom. The first possibility is exemplified by Nietzsche's phrase: "There is more reason in your body, than in your best wisdom" (Nietzsche, 1966: p. 146). The second is represented by a striking claim of Nietzsche: "My wisdom has long gathered like a cloud; it is becoming stiller and darker. Thus does every wisdom that is yet to give birth to lightning bolts … For these men of today I do not wish to be light, or to be called light. These I wish to blind. Lightening of my wisdom! put out their eyes!" (Nietzsche, 1966: p. 401).

Languid wisdom or lightning wisdom, wisdom as a sedative or as dynamite – this is Nietzsche's alternative in matters of wisdom. If wisdom is referred to, one should be on one's guard as to which type of wisdom is involved.

Epistemological Contours

Shifting from Practical to Theoretical Philosophy

The disappointment with contemporary philosophy's neglect of wisdom is flawed by one-sidedness. One assumes elements

of wisdom must be found in the field of practical philosophy, where it would indeed be hard to discover any. Conversely, however, insights of contemporary theoretical philosophy provide a great many elements giving contours to a contemporary concept of wisdom. My thesis is that the design of wisdom today is to be drawn from the epistemological development in recent decades.

Recent Insights: Theoretical Reason with the Characteristics of Practical Reason

Shifting to epistemology in matters of contemporary wisdom makes all the more sense given that in recent decades theoretical reason has been shown to be characterized by traits formerly considered typical of practical reason alone. Since Aristotle, the difference between the two types of reason had been considered fundamental: theoretical reason considers permanent things, whereas practical reason deals with changeable things. Correspondingly, theoretical reason demands distance and provides perfect cognition, whereas practical reason requires involvement and allows only context-dependent cognition. In recent decades, however, theoretical reasoning has been shown to bear the same basic traits as practical reasoning: theoretical cognition too represents relative, not absolute cognition; it too is context-bound and its validity is tied to specific (axiomatic, historical, social, etc.) frameworks. In short, theoretical reasoning also exhibits the characteristics of practical reasoning. This explains why the shape of contemporary wisdom can be inferred from innovations on the theoretical side.

Elements of Modern Epistemology Relevant for a Contemporary Concept of Wisdom

As this view is unusual, I will set out in the following sections the relevant elements in some detail.

From Nietzsche to Quine: The Uncertainty of Fundaments

Nietzsche can be regarded as a forerunner of the contemporary condition in having pointed out that our theoretical cognition is of a piece with practical and pragmatic cognition: "one may certainly admire man as a mighty genius of construction, who succeeds in piling up an infinitely complicated dome of concepts upon an unstable foundation, and, as it were, on running water. Of course, in order to be supported by such a foundation, his construction must be like one constructed of spiders' webs: delicate enough to be carried along by the waves, strong enough not to be blown apart by every wind" (Nietzsche, 1979: p. 85). In the twentieth century Neurath said quite similarly, "we are like mariners who must rebuild their ship on the open sea, without ever being able to disassemble it in dock and reassemble it with the best components" (Neurath, 1932–1933: p. 206). Popper's diagnosis reads: "just when we believed that we were standing on firm and safe ground, all things are, in reality, insecure and unstable" (Popper, 1992: p. 65); "the basis is unstable" (Popper, 1969: p. 76). Quine later adopted Neurath's phrase as the motto of *Word and Object* (Quine, 1960: p. VII).

Scientific Pluralism (Quine)

The inevitability of scientific pluralism is a decisive insight of contemporary epistemology that was first formulated by Quine. Even in the strong realm of science, no one, singularly valid theory is to be expected, but a multitude of equally valuable alternative candidates has to be taken into account. This is a consequence of the empirical under-determination of theory shown by Quine in his seminal essay "Two Dogmas of Empiricism" from 1951: "Total science, mathematical and natural and human, is … underdetermined by experience" (Quine, 1980: p. 45). Empirical tests can only tell us that *something* is in need of being altered in the theory, but not exactly *which element* is to be changed. "Any statement can be held true come what may. Conversely, by the same token, no statement is immune to revision" (Quine, 1980: p. 43).

In *Word and Object* (Quine, 1960), Quine resolutely drew the consequences of scientific pluralism: "we have no reason to suppose that man's surface irritations even unto eternity admit of any one systematization that is scientifically better or simpler than all possible others. It seems likelier … that countless alternative theories would be tied for first place" (Quine, 1960: p. 23). "We can never do better than occupy the standpoint of some theory or other, the best we can muster at the time" (Quine, 1960: p. 22).

The Impossibility of Any Meta-Position (an Insight Common to Hermeneutics, Poststructuralism, and Analytic Philosophy)

This also means that it is impossible to assume some meta-position that would allow us to regard the multitude of scientific theories from a superior point of view and select the 'truly right' one among them. There is no God's-eye standpoint, no view from nowhere – neither in theoretical nor in practical matters. The impossibility of such a meta-stance is a decisive insight of recent philosophy that is common to all (superficially so variant) strands: to Gadamer's hermeneutics (according to which we cannot achieve a metahistorical standpoint toward history), to Lyotard's poststructuralism ("Simplifying to the extreme, I define postmodern as incredulity toward meta-narratives," Lyotard, 1984: p. xxiv), and, of course, to analytic philosophy, where this insight was first formulated not by Quine but by Wittgenstein, who declared it his "leading principle" that there is no "metaphilosophy" (Wittgenstein, 1974: p. 116).

Wittgenstein: Limits of Justification

Wittgenstein has developed insights crucial to the present understanding of cognition and also relevant to every conception of contemporary wisdom altogether.

First, Wittgenstein demonstrated limits of justification with respect to the certainties of common sense. These are neither based on cognition nor could they ever be secured by cognitive justification. For these certainties are by themselves constitutive of all our activities of justification: "All testing, all confirmation and disconfirmation of a hypothesis takes place already within a system. And this system is not a more or less arbitrary and doubtful point of departure for all our arguments: no, it belongs to the essence of what we call an argument. The system

is … the element in which arguments have their life" (Wittgenstein, 1972: p. 16).

Second, Wittgenstein showed how this structure affects all our decisions: "Nothing we do can be defended absolutely and finally. But only by reference to something else that is not questioned" (Wittgenstein, 1984: p. 16). Decisions necessarily have recourse to something which they take for granted: "The questions that we raise and our doubts depend on the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn" (Wittgenstein, 1972: p. 44). "If I want the door to turn, the hinges must stay put" (Wittgenstein, 1972: p. 44).

So far this means that not only the validity of scientific theories (Quine), but also of our everyday certainties and all our argumentation, cannot be conclusively secured; they possess only conditional validity. In the regress of justifications one does not reach some incontrovertible ground, but only postulates, which in principle are once again questionable. While admitting good and varied justifiability, there is no such thing as an absolute, ultimate justification.

Finally, Wittgenstein has also shown how all this applies even to the whole framework of justification and argumentation (to 'rationality'): "We can't talk of reasons for thinking … We can describe the game of thinking, but not the reasons why we think. "Reason" only applies within a system of rules … It is nonsense to ask for reasons for the whole system of thought" (Wittgenstein, 1980: p. 88). What Wittgenstein wants to draw our attention to is that the quest for reasons (e.g., for reasons for the whole business of reasoning) suggests leaving behind this game, but obviously it does not do so: it asks the very question that typifies this game itself and thus remains within its realm.

Hence there is no possibility of justification that absolutely vouches for itself. This is expressed in the following sentence (which, significantly enough, sounds like the dictum of a sage): "You must bear in mind that the language-game is so to say something unpredictable. I mean: it is not based on grounds. It is not reasonable (or unreasonable). It is there—like our life" (Wittgenstein, 1972: p. 73).

Epistemic Contemporariness and Wisdom

These epistemic advances, which characterize the present state of affairs in philosophy, surprisingly also correspond to contemporary standards of everyday consciousness. The latter is molded to a high degree by the acknowledgment that all validity is relative and by an awareness of the plurality of modes of experience, cultures, forms of life, and even worlds. Owing to this correspondence, the new epistemic philosophical standards are also apt to serve as maxims for contemporary everyday orientation. They match our expectations and needs in matters of wisdom, and they do so in a manner specifically appropriate to the contemporary situation.

The present understanding of wisdom typically implies circumspection, awareness of limitation, and competence in dealing with uncertainty. The sage is someone who is attentive to neighboring and commonly neglected aspects. She is aware of the context and brings it to bear regulatively. Whereas an ignorant person confidently attributes himself a grasp of the whole, the wise person objects to such totalization. She recognizes the limited nature of all singular perspectives, is aware of differing viewpoints, and attentive to the open potential for additional alternatives. She warns against misrecognizing particularities, infringements, and overbearingness. She advocates justice at the level of detail and openness on the whole. So the wise person is an expert in uncertainty too. Wisdom is particularly needed in situations where justice is required with respect to diverging demands without a meta-rule being available from which one could deduce the right decision. In these cases, even though it be cognized, but only found, the wise person will have a sense for what is right. Sometimes, however, it will be important just to fend off the wrong thing while leaving open what's right. Such composure also distinguishes the wise person.

Contemporary epistemology, which emphasizes the limits of all justification and demands their respect, is thus properly tailored to the aspects of wisdom just set out. It inherently works toward a consciousness that bears traits of wisdom in high measure.

This is why a state of consciousness that has incorporated those epistemic standards and that knows how to practice them can provide orientation for our time and satisfy our demands in matters of wisdom. Whoever has made this genuinely philosophical possibility her own will be in no great need of extra-philosophical and esoteric doctrines of wisdom that have become so popular in recent years.

See also: Pluralism and Tolerance; Relativism: Philosophical Aspects; Value Pluralism; Wisdom, Psychology of.

Bibliography

Aristotle, W., 1984. Complete Works of Aristotle. Princeton University Press, Princeton, NJ.
Burkert, W., 1960. Platon oder Pythagoras? Zum Ursprung des Wortes Philosophie. Hermes 88, 159–177.
Curnow, Trevor, 2010. Wisdom in the Ancient World. Duckworth, London.
Descartes, R., 1983. The Principles of Philosophy. Reidel, Dordrecht, The Netherlands.
Descartes, R., 1998. Regulae ad Directionem Ingenii. Rodopi, Amsterdam.
Diogenes, L., 1931. Lives of Eminent Philosophers, vol. 2. Harvard University Press, Cambridge, MA.
Diogenes, L., 1972. Lives of Eminent Philosophers, vol. 1. Harvard University Press, Cambridge, MA.
Goodman, N., 1978. Ways of Worldmaking. Hackett, Indianapolis, IN.
Hall, Stephen S., 2010. Wisdom: From Philosophy to Neuroscience. Knopf, New York.
Hegel, G.W.F., 1977. Phenomenology of Spirit. Clarendon Press, Oxford, UK.
Husserl, E., 1965. Philosophie Als Strenge Wissenschaft. Klostermann, Frankfurt, Germany.
Kane, R., 2010. Ethics and the Quest for Wisdom. Cambridge University Press, Cambridge.
Kant, I., 1983. Anthropologie in Pragmatischer Hinsicht. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany (Anthropology from a Pragmatic Point of View. Southern Illinois University Press, Carbondale, IL (1978)).
Kant, I., 1992. The Jäsche Logic. Lectures on Logic. Cambridge University Press, Cambridge, UK.
Kant, I., 1993. Critique of Practical Reason. MacMillan, New York.
Luther, M., 1964. Werke. Kritische Gesamtausgabe (Weimarer Ausgabe). Wider die himmlischen Propheten, von den Bildern und Sakramenten, vol. 18. Böhlau, Weimar, Germany.
Luther, M., 1967. Werke. Kritische Gesamtausgabe (Weimarer Ausgabe). Wider der Vernunft Hindernis, vol. 6. Böhlau, Weimar, Germany.

Lyotard, J.-F., 1984. The Postmodern Condition: A Report on Knowledge. University of Minnesota Press, Minneapolis, MN.
Montaigne, M., 1958. The Complete Essays of Montaigne. Stanford University Press, Stanford, CA.
Neurath, O., 1932–1933. Protokollsätze. Erkenntnis 3, 204–214.
Nietzsche, F., 1966. Thus Spoke Zarathustra. The Portable Nietzsche. Viking Press, New York.
Nietzsche, F., 1974. The Gay Science. Vintage Books, New York.
Nietzsche, F., 1979. On truth and lies in a nonmoral sense. In: Breazeale, D. (Ed.), Philosophy and Truth. Selections from Nietzsche's Notebooks of the Early 1870's. Humanities Press, New York.
Nietzsche, F., 1980a. Nachgelassene Fragmente. Juli 1882 bis Herbst 1885. In: Colli, G., Montinari, M. (Eds.), Sämtliche Werke, Kritische Studienausgabe in 15 Bänden, vol. 11. Deutscher Taschenbuch Verlag, Munich, Germany.
Nietzsche, F., 1980b. Nachgelassene Fragmente. Herbst 1885 bis Anfang Januar 1889. In: Colli, G., Montinari, M. (Eds.), Sämtliche Werke. Kritische Studienausgabe in 15 Bänden, vol. 13. Deutscher Taschenbuch Verlag, Munich, Germany.
Nietzsche, F., 1989. Beyond Good and Evil. Vintage Books, New York.
Plato, K., 1961. The Collected Dialogues of Plato. Pantheon Books, New York.
Popper, K.R., 1960. The Logic of Scientific Discovery. Hutchinson, London.
Popper, K., 1969. Logik der Forschung. Mohr, Tübingen, Germany.
Popper, K., 1992. The Logic of the Social Sciences. In Search of a Better World. Routledge, London, New York.
Quine, W., 1960. Word and Object. MIT Press, Cambridge, MA.
Quine, W., 1980. Two Dogmas of Empiricism. From a Logical Point of View. Harvard University Press, Cambridge, MA.
Toulmin, S., 1990. Cosmopolis. University of Chicago Press, Chicago, IL.
Voltaire, 1967. Article on philosophie. Dictionnaire philosophique. Garnier, Paris.
Welsch, W., 1989. Weisheit in einer Welt der Pluralität. In: Oelmüller, W. (Ed.), Philosophie und Weisheit. Schöningh, Paderborn, Germany, pp. 214–259.
Welsch, W., 1995. Vernunft. Die zeitgenössische Vernunftkritik und das Konzept der transversalen Vernunft. Suhrkamp, Frankfurt, Main, Germany.
Welsch, W., 2011. Weisheit. In: Welsch, W. (Ed.), Immer nur der Mensch? Entwürfe zu einer anderen Anthropologie. Akademie, Berlin, Germany, pp. 139–167.
Wittgenstein, L., 1972. Über Gewissheit—on Certainty. Harper, New York.
Wittgenstein, L., 1974. Philosophical Grammar. University of California Press, Berkeley, LA.
Wittgenstein, L., 1980. Wittgenstein's Lectures. Cambridge, 1930–1932. Blackwell, Oxford, UK.
Wittgenstein, L., 1984. Culture and Value. University of Chicago Press, Chicago, IL.

Witchcraft

Peter Geschiere, University of Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Till far into the twentieth century two nodal points in the study of witchcraft stood out: the work by historians on witch trials in early modern Europe, and anthropologists' studies of its role in local tensions in Africa and Melanesia. The differences between the two settings were so glaring that some authors objected to the use of the same term. Yet, in both settings similar themes emerged, even though these were interpreted in quite a different way. Common interests were the relation of witchcraft to the forces of modernity (the State, the Market), but equally its ongoing link to the intimacy of family and home; also its relation to (a)moral power, to healing, and to class differences. Moreover, over the last decades the geography of witchcraft studies was constantly widening: seminal studies appeared concerning different parts of Asia, Latin America, and other regions of the world. In 2004, historian Wolfgang Behringer published a 'global history' of witches and witch-hunts, challenging anthropologists to finally stop their endless bickering over right or wrong terms, so that comparisons could at last take off. The differences between the approaches by historians and anthropologists should not be exaggerated, but precisely the confrontation between the two can help to focus on new directions for research around this challenging topic that seems to refuse to be domesticated.

Until far into the twentieth century, the academic study of witchcraft was concentrated in two research fields clearly separated in time and space: historians' studies of the enigmatic outburst of witches' trials in early modern Europe, and anthropological studies of the role of witchcraft in notably African and Melanesian societies during colonial and post-colonial periods. The contrasts between the historical and anthropological approaches can serve as an entry point for exploring different trends in witchcraft studies in interaction with rapidly changing contexts.

Historians and Anthropologists: Differences and Correspondences in Approach

In her introduction to a conference volume commemorating the 30th year anniversary of Evans-Pritchard's classical study *Witchcraft, Oracles and Magic among the Azande* (southern Sudan), Douglas (1970: p. xiii) noted with some apparent surprise the deep chasm between these two fields of study: "The anthropologists of the 1950s developed insights into the functioning of witch beliefs which seemed about as relevant to the European experience as if they came from another planet. Dangerous in Europe, the same beliefs in Melanesia or Africa appeared to be tame, even domesticated; they served useful functions and were not expected to run amuck." In retrospect, one might wish the second half of this quotation to be still true. People's growing preoccupation over the last decades in many parts of Africa and Melanesia with new forms of witchcraft seems to reflect, on the contrary, widespread worries that it has 'run amuck' in those regions as well.

Yet, Douglas's trenchant observation did highlight striking differences between the trajectories of witchcraft studies by historians and anthropologists and, therefore, offers a challenging starting point for a survey of the topic. The differences between the two fields are, indeed, so great that some authors even objected to the use of the same term. Crick (1979), for instance, warned that the term witchcraft was so saturated with implications stemming from European history that anthropologists working on other areas had better avoid the term altogether. One problem with such a purist stance is that this English term (or the corresponding French, Spanish, and Portuguese words) has been adopted widely by people throughout the Southern Hemisphere. Another consideration might be that despite huge variations the representations concerned revolve around basic conceptions that are surprisingly similar. In many parts of the world notions that are now commonly translated as 'witchcraft' refer not only to hidden human aggression, but also to more specific representations about people being able to double themselves, leaving their bodies at night and flying off to nocturnal meetings with their accomplices.

Especially anthropologists, confronted by a bewildering proliferation of imaginaries in this field, have tried to clarify things by formulating strict definitions and distinctions. The best known is Evans-Pritchard's proposition – on the basis of the distinctions the Zande, the subjects of his first book, supposedly made – to distinguish witchcraft and sorcery: the first referring to an inborn quality, the second to the use of acquired means. But even for neighboring areas in Africa this distinction proved difficult to maintain in view of the fluidity and dynamics of witchcraft discourses. More recently, some authors (see Geschiere, 1997, 2013) suggested that this search for clear-cut definitions and classifications risked to make anthropologists blind to the surprising dynamics of these representations and their chameleonic qualities – which are, moreover, crucial to understand how they can adapt themselves so easily to modern changes. In this treacherous field of study, researchers may have to resign themselves to working with fuzzy notions that exhibit a common core, but at the same time engender highly variable and constantly changing elaborations.

The differences between the historians' and the anthropologists' versions of witchcraft are illustrative of the widely varying implications these notions can acquire. In general, historians paid more attention to the dynamics of these

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12165-8

representations. After all, a central question to them was why persecutions of witches – quite rare throughout what was sometimes called the 'dark' Middle Ages – acquired epidemic forms in several parts of Western and Central Europe, precisely at the dawn of modern times (sixteenth and seventeenth centuries). Historians like Robert Mandrou or Norman Cohn explained this as the effect of a longer process of the diabolization of local witchcraft beliefs. Popular worries about perpetrators of evil within the local community (*maleficium*) increasingly became linked to ecclesiastically inspired visions of a Satanist cult. In the confessions of people accused of witchcraft – increasingly obtained under torture by the henchmen of ecclesiastical and secular courts – the Devil emerged as the leading figure at their nocturnal meetings. Some historians related the quite sudden proliferation of witch-hunts after 1500, therefore, to the disciplinary action of the State and the Church that ever more affected the local communities. Muchembled (1978), for instance, saw a clear distinction in the archives between two discourses: the one of the witnesses who accused their neighbors of destroying corps, cattle, and even human lives; and the one of the judges who forced the accused to relate all this to their dealings with Satan. Thus, Muchembled construed a direct link between the witch trials, on the one hand, and, on the other, the rise of the absolutist State and the religious struggles inspired by Reformation and Counter-Reformation; he saw the witch trials as an emanation of the 'forced acculturation' of popular culture by a 'triumphant centralization.' However, in the beginning of the 1970s other historians, mainly working on England, had already advanced another explanation, emphasizing rather the increased willingness of notably rural people to accuse their fellow villagers. Thomas (1971), for instance, related this to the deepening inequalities within rural communities. Richer villagers increasingly resisted older obligations to share with their poorer fellowmen, translating feelings of guilt about this into witchcraft accusations. In this view, State and Church were rather drawn into these affairs by pressures from below.

Thus, there seemed to be a more or less clear contrast between a continental pattern – that for some time also affected Scotland due to dynastic links with Denmark – in which State and Church authorities played a key role with their obsession with the Devil; and a quite different one, prevailing in England, in which locals pressurized the authorities to do something about the *maleficium* of the witches in their midst. More recently, however, this contrast was nuanced by several authors (see Briggs, 1996), who emphasized that on the continent as well, tensions and accusations within the local communities played a key role, while direct involvement of State or Church authorities in the persecution of witches was the exception – certainly in the most absolutist States (France, Spain) – rather than the rule. Moreover, they signaled that in England as well, Satan came to play an increasing role in the witches' confession, be it belatedly and less prominently. Yet, despite such different emphases, historians' interpretations have in common the premise that witchcraft cannot be analyzed as a purely local issue but has to be studied in relation to the intervention of outside agents, notably the State and the Church. After all, their main sources – court archives – were produced by the latter. An attempt by Ginzburg (1990) to relate the imaginary of the witches' Sabbath to a set of folk beliefs – going back to Neolithic times as some sort of substratum throughout Eurasia – received a lukewarm reception and clearly fell outside the historians' main paradigm.

Anthropologists, in contrast, for a long time continued to address witchcraft as a local issue, trying to resolve its riddles by in-depth studies of local relationships and conceptions. But the anthropological perspective as well was subject to constant turns, some of which were quite surprising. An obvious question is why especially Africa and Melanesia loomed so large in anthropological studies of witchcraft. However there was a change in this respect over the last decades due to publication of seminal studies on witchcraft in different parts of Asia, Latin America, and other parts of the world (cf for instance Whitehead and Wright, 2004; Siegel, 2006; Buyandelger, 2013). In retrospect it seems that other notions – like shamanism, voodoo, magic – were used to refer to ideas and practices that had much in common with what was referred to elsewhere as witchcraft. There was also substantial turns in the anthropological perception of these ideas and practices. Evans-Pritchard's Azande study, already mentioned above, is often quoted as a classic in the field. Yet it hardly proved to be a trendsetter. Its focus was on issues of cognition – on why the Azande combined very commonsensical explanations with interpretations in terms of occult aggression. But it inspired in the 1950s and 1960s, a series of micro-political studies in which witchcraft was analyzed – in line with the functionalist ideas that dominated the discipline at the time – in relation to social order and integration. In a series of in-depth monographs of local societies in what at the time was called British Central Africa, anthropologists of the 'Manchester school' studied witchcraft accusations as a form of social control. Typical for this approach was Max Marwick's characterization of witchcraft as 'a social strain gauge': as a measure of sociopolitical tensions, but also as a way to restore social order. It was this interpretation that produced the in retrospect somewhat amazing view of it as 'tame' and 'domesticated' (see quote from Douglas above). However, especially toward the end of the last century, this conception of witchcraft as a conservative force, defending the status quo, was increasingly felt to be too limited in its focus on the local context. Anthropologists also were forced to study these supposedly 'traditional' discourses in the context of broader changes.

Witchcraft and Modernity

It is questionable whether the functionalist view of witchcraft, as restoring social order, applied even during the colonial period. At the time researchers may have underestimated to what extent the imposition of the *Pax Britannica* (or *Gallica*) played a role in this respect. But it is clear that this view hardly prepared anthropologists to confront the quite spectacular dynamics of representations of the occult after Independence (in most African countries around 1960), especially in the more modern sectors of society. It is difficult to decide whether

accusations and rumors, let alone actions, in terms of witchcraft increased after Independence (just as the old debate among anthropologists on whether witchcraft 'increased' or 'decreased' after the colonial conquest had no conclusive ending). But it is quite clear that it came out into the open, in many parts of postcolonial Africa. Witchcraft became a hotly debated issue, not only in the street rumors of *Radio Trottoir*, but also in newspapers, on TV, etc. It was striking, moreover, that rumors and references to the occult forces proliferated especially in more modern contexts: in relation to novel forms of entrepreneurship, national politics, sport, and the new institutions for health care and education. The view, to many still self-evident, of witchcraft as a 'traditional' relict that will automatically disappear with modernization – 'witchcraft will disappear with electric light' as a Dutch missionary in Cameroon told me in 1971 – became difficult to maintain in consideration of the popular tendency to address precisely such modern developments in terms of these occult discourses. Neither did the anthropological image of witchcraft, as a local phenomenon, proves to be of much use for understanding the spectacular increase in scale of these rumors that acquired national or even transnational dimensions. No wonder that after the 1960s, anthropologists – also afraid of being accused of primitivizing Africa and denying its rapid modernization – tended to abandon the topic to theologians and journalists. The then-current anthropological approach in terms of social order proved to be of little help in analyzing the new transformations of these local beliefs.

In many parts of Africa, however, there was increasing popular concern about new forms of witchcraft that often were explicitly related to the *nouveaux riches* who were supposed to accumulate wealth by exploiting the labor of their witchcraft victims. Such rumors made a contrast between older forms of witchcraft in which the victims were 'eaten,' and the new one of 'selling people,' – that is, turning them into some sort of zombies to sell them to work on 'invisible plantations.' The fear of such new threats triggered a frantic search for new forms of protection – for instance, urgent appeals to the State authorities to intervene. An unsettling example of how pressing the witchcraft issue was becoming in situations of rapid change was the lethal witch-hunts that marked the end of Apartheid in the northeastern part of South Africa – young men, encouraged by *nganga* (witch-doctors) and claiming to act as local representatives of the African National Congress, executing dozens of 'witches' (Niehaus, 2001). The half-hearted implications of State authorities in such witch-hunts in many parts of the continent made it all the more urgent to gain deeper insight into the modern dynamics of witchcraft discourses. In this context, historical studies of the witch-hunts in early modern Europe served for some anthropologists as an eye opener, since the former had always emphasized the crucial role of external powers and wider political developments – the state and ecclesiastical courts impinging on peasant societies as representatives of the emergence of the absolutist state – in the explosion of witchcraft rumors and accusation at the local level.

No wonder that since the early 1990s there came a new wave of witchcraft studies in Africa and Melanesia, notably by anthropologists concentrating on the relation with popular views of 'modernity' – that is, the dreams of the modern, in all its varieties, that effected ever wider strata of the population (cf Comaroff and Comaroff, 1993; Geschiere, 1997; Ashforth, 2005). This linking of witchcraft and modernity, according to some even a new 'paradigm,' implied attention to particular aspects that could explain the continuing resilience of these local beliefs – notably their circular character, their moral ambiguity, and their striking capacity to relate local realities of the family and the home to global processes of change. The circularity of witchcraft discourses is expressed most vividly in the paradox that the main protection against witchcraft often is to be sought in the very same field. In many parts of Africa, for instance, the *nganga* (healer) is supposed to be able to offer protection and healing only because of his (or her) overdeveloped witchcraft. In some societies there is a consistent effort to conceptually or even institutionally separate the *nganga*'s (or the chief's) occult power from the evil forces of the witches. Then witchcraft seems to be, indeed, unequivocally evil. But in practice, such distinctions between positive and negative forms of occult power are often highly precarious and controversial. In many African societies, for instance, *nganga* remain highly ambivalent figures – healers but at the same time suspect because of their dealings with extremely dangerous forces. The *nganga* is, therefore, also a striking example of the moral ambiguity of these forces. Basically these forces are certainly evil, but – as in the *nganga*'s case – they can be canalized and used in a constructive way. In this more positive sense, witchcraft – it is in such a context that this English translation of African notions becomes, indeed, problematic – can even be seen as a precondition for accumulating wealth and power.

This applies most strikingly to the new modes of enrichment that seem to bring out ever more starkly the ambivalence of witchcraft conceptions: the rumors, mentioned above, about the new witchcraft of wealth that turns its victims into zombie-laborers express horror about such treacherous attacks, but also fascination with the new riches thus acquired. It is these ambiguities that give witchcraft discourses their elasticity and their surprising capacity to relate the global to the local. In most parts of Africa and Melanesia these discourses still are rooted strongly in the local setting, notably in the intimacy of kinship: witches are supposed to have a special hold over their kin or in any case to work through an ally 'within.' Yet, the same discourses are invoked to cope with the unsettling modern developments: the vagaries of the world market, the dazzling riches of the happy few, and the recurrent mishap of the many in their pursuit of the new forms of wealth. In such contexts, witchcraft rumors can attain truly global dimensions referring, for instance, to the circulation of zombies on a world scale or to the appropriation of extremely potent 'medicine' from Europe or India.

Such anxieties about hidden aspects of modern developments are certainly not special to Africa or Melanesia (cf the seminal studies by Taussig, 1987; Kapferer, 1997 on, respectively, Amazonia and Sri Lanka). The new anthropological interest in the paradoxical link between witchcraft and modernity inspired also alternative readings of developments in what are thought to be typically 'modern' societies. In this context there was increasing interest, among Africanists and others, in anthropologist Favret-Saada's challenging studies of

the resilience of witchcraft practices in the French countryside (1977). Truly innovative elements in her analysis were the emphasis on how words can become charged, and also on the inevitability of the researcher herself becoming involved in the struggles for protection against secret attacks. In a more direct relation to modern forms of politics and communication, Geschiere (in Meyer and Pels, 2003) drew attention to intriguing parallels between witch-doctors and 'spin-doctors' in respective African and American politics – both manipulating esoteric expertise and 'public secrets' that shroud the exercise of power. In her detailed analysis of rumors on Satanism and child abuse in present-day England, La Fontaine (1998) demonstrated the relevance of anthropological witchcraft studies on Africa and Melanesia in such a modern context. The broader interest of these new directions in the study of witchcraft, often still seen as the archetype of the traditional, was precisely that they helped to debunk simplistic opposition between modern and traditional, science and magic, transparent and opaque, or even more simply between the West and the rest. Clearly, even in the West modernity, witchcraft retained its own forms of enchantment and opacity.

Paradigms and Critiques

However, after 2000 a rapid reaction against this emphasis on modernity set in, especially in African studies. More than among historians, anthropological approaches to witchcraft seem to progress – or in any case continue – in a jerking movement since it has become common practice that scholars construe the work of their predecessors into a 'paradigm.' Apparently an attraction of the latter is that it automatically implies limitations, and therefore the challenge to break out of these. As noted before, Douglas contrasted the vision by her contemporaries of witchcraft as a system of social control to Evens-Prichard's epistemological approach. The anthropologists of the 1990 saw their emphasis on the 'modernity' of witchcraft as a breakthrough of the paradigm of classical studies with their emphasis on social control within a limited local context. But especially over the last few years younger anthropologists evoked a 'witchcraft and modernity paradigm' that would have determined the publications of the 1990s. The switch in the 1990s would have imposed a view in which witchcraft had to be studied as a product by modernity, so that the longer historical trajectories of these notions and practices were ignored (see, for instance, Henry and Tall, 2008; or Stroeken, 2010). It remains to be seen whether any of the scholars to whom this 'paradigm' was attributed had such limited ideas. Critics often quote in this context Jean and John Comaroff's notion of 'occult economies' as a by-product of 'millennial capitalism' (Comaroff, 2000). But it is quite clear that their insistence on the generality of such occult economies in different parts of the world (and not only in Africa – still a very welcome contribution to undermine tenacious ideas about this continent's otherness) does not preclude close attention to the specific forms such occult economies acquire as products of specific histories. Moreover, it is hard to believe that anyone would ever think witchcraft to be only about modernity (indeed, in present-day contexts in many parts of Africa it seems to become an almost empty signifier that can refer to anything). Amidst the clamor of such critiques, Todd Sanders' formulation is a very wise one: "African witchcraft may well be part *of* modernity, but by no means needs to be *about* modernity" (2003: p. 338). The merit of the switch in African witchcraft studies in the 1990s may have been that it raised the question how local discourses on occult forces could be so easily grafted upon incisive modern changes. But this link with people's visions of modernity should certainly not become a final truth. Most scholars – also the authors who supposedly initiated this paradigm of witchcraft and modernity – will agree that there are many more aspects to be taken into account.

In certain respects this turn in anthropological debates around 2000, resembled the turn that took place a decade earlier in the debates among historians about early modern Europe. As noted above, in the 1990s ever more historians turned against the emphasis by Robert Muchembled and others on the 'civilizing offensive' by State and Church against popular culture as an explanation for the unleashing of bloody persecutions in the sixteenth and seventeenth centuries. The trend in historical studies became also to see the role of Church and State in a broader context, taking a wide array of varying factors into account to understand outburst of persecutions. A particularly convincing example of such a broader approach is Thomas Robisheaux' beautiful and disquieting monograph on *The Last Witch of Langenburg* (2009). In this study of the last burning of a witch in Württemberg (Germany – 1672) the author succeeds, through a creative use of all sorts of sources, to evoke such a vivid and detailed image of the series of events that led to the execution of Anna Schmieg, the miller's wife, that many anthropologists will envy him for such a complete picture. Striking is especially how the role of the local state and the representatives of the Lutheran church turns out to be only one element – next to local customs, pressures from the community, and so on – of a collective moral effort of frightening intensity to make the woman confess of her dealings as a witch with the devil. The hold of the latter over the community could only be broken if the woman could be forced to confess. Apparently any means – torture but also constant moral admonition – was permitted for this.

Themes

Underneath these somewhat noisy transitions from one 'paradigm' to another, themes emerged that can be of great value for further developments in the field. One promising direction already mentioned above was the emphasis on the circularity of witchcraft discourse, and the ambiguity in the role of healers – implying that the protection against these occult forces had to be looked for in the field of these forces themselves. Related is the emphasis on the capacity of these representations to relate the global to the local – to explain new developments in terms that are rooted in the local realities of family and home. Pursuing these lines of thought, there was growing interest in the ambiguous relation between witchcraft and the intimacy of home. For a long time anthropologists, especially those working in Africa, had

already stressed the close link between witchcraft and kinship. As noted, witches are supposed to have a special hold over their close relatives, and the instigators of an occult attack are always first sought within 'the house.' However, with the increasing scope of witchcraft rumors – now addressing issues that are far beyond the seemingly closed world of the village, studied by anthropologists in the 1950s – people's tendency to continue linking witchcraft conspiracies to the local realities of home and village is striking. This is certainly not a simple, unambiguous continuity. On the contrary, in many parts of Africa, for instance, people are now anxious about new forms of witchcraft because these would surpass the old limits of kinship (apparently witches can nowadays attack anybody, kin or nonkin). However, on closer inspection, cases often continue to refer ultimately back to the family and kinship – albeit in new, dynamic versions of what can still be interpreted as kinship. Indeed, the hypertrophied tenor of present-day witchcraft fantasies in many African contexts seems to reflect a constant stretching of kinship, almost to a breaking point. Kinship terms continue to be used to bridge increasing social and geographical distances. Yet, it is clear that this stretching with new forms of mobility puts the concomitant kinship norms under heavy stress. Even transcontinental migrants feel now persecuted by the dangers of home. So, rather than take the link between witchcraft and kinship (or more largely, intimacy) for granted, the challenge might be to follow the ambiguities of this link in rapidly changing contexts – and also in its changing implications for the issue of trust and the question as to how one can ever succeed in building confidence in one's intimates if they have such a dangerous hold over you. This link between occult dangers and intimacy – and its problematic implications for trust – is not special to Africa; on the contrary, it returns in many parts of the world, all be it in different forms following different historical trajectories. In such a comparative view, people's obsession with witchcraft in Africa or Melanesia is no longer that exceptional (see further Geschiere, 2013).

Other topics looming large in the study of witchcraft are the relations with healing and, closely connected to this, with the new churches, notably those of evangelical signature. The central role of issues of witchcraft in the rapid spread of Pentecostal churches in many parts of the world – in contrast to many older churches, Pentecostals offer a direct cure for witchcraft – is only one example of witchcraft fears being central in the impressive flowering of prophets and spiritual healers in Africa and elsewhere. Mary (2009) shows in a most interesting comparison of the role of prophets in Gabon, Benin and Ivory Coast different trajectories in the way these *visionnaires* deal with witchcraft and development, in relation to different political contexts. In a challenging vision, Joseph Tonda (2005) tries to definitively break away from the conceptual opposition between the forces of modernity, on the one hand, and the logic of witchcraft, on the other, by evoking the figure of *Le Souverain moderne*. This amoral figure dominates postcolonial forms of governmentality in postcolonial Africa, emerging from an osmosis – a 'magma' – of all these elements. It is precisely by realizing the full complexity of *Le Souverain moderne* that scholars can finally break out of the premise of a 'great divide' opposing the forces of modernity (or 'the work of God') and *l'esprit sorcellaire*.

Terms and Their History

Another difference between historians and anthropologists is that the latter keep worrying about the terms to be used in academic analysis. With some historians such terminological hesitations led to a certain impatience. Behringer (2004: pp. 3 and 216) reproaches anthropologists, for instance, that their constant worrying about terms blocked comparative studies. It is clear that in this marshy field a definitive terminological purity can never be reached. All terms adopted by academics will be polluted by their colloquial use in addressing highly ambiguous and polyinterpretable phenomena. However, more recently, debates about terminology among anthropologists went beyond quibbles over definitions and distinctions by focusing on the historicity of the terms involved. One can agree that all these notions of Western origin (witchcraft, sorcery, *sorcellerie*, magic, and so on), offer partial translations of local notions that are often much broader in their implications. Yet, rather than try to put a ban on such notions – which is redundant in view of the great currency of these terms, a more interesting approach might be is to try and follow the detailed history of how these notions became attached to local terms, and what the implications of these translations were (see for an early example of this approach B.Meyer's monograph on 'translating the devil' in Ghana – 1999).

An inspiring example is provided by the recent book by Andrea Ceriana Mayneri (2014) who – due to a most creative use of missionary archives, combined with intensive fieldwork on the spot – reconstructs how in the 1930s the word *sorcellerie* was imposed on local notions among the Banda in the Central African Republic due to precarious misunderstandings between missionaries, civil servants, and interpreters. Clearly this misunderstanding proved to be a 'working' one, since from then on the word *sorcellerie* deeply marked the evolution of the Banda imaginary on these occult forces. In a similar, but more general vein, historian Florence Bernault (2009) explores the switch in the late colonial literature on French Equatorial Africa from *fétichisme* to *sorcellerie*. Important is, in her view, that the word *sorcellerie*, like witchcraft, retains certain premises colored by Western academy, in a time that ongoing secularization was seen as the normal evolution. In this line of thought *sorcellerie*/witchcraft would refer to the supposed impossibility to separate politics and religion in African societies, which thus were marked as some sort of counterpoint to modern, secularizing society. In Bernault's view, academics should be more conscious of the fact that the term still carries such connotations.

One can wonder whether the dynamics that terms – also academic ones – acquire through their daily use in society, is not of importance in this context. The general appropriation of these notions by Africans has set in motion certain shifts. Similarly, on the side of academics, more recent uses of the term may be marked by the general realization that modernity is not bringing a general tide of secularization. The 'return of religion' – even in the supposedly secularized societies of Europe – may shift also the implications terms like witchcraft had to the people who originally imported those notions to Africa and other parts of the world. Due to such shifts the resilience of what people call 'witchcraft' in, for instance, Africa may also affect people's views in

societies that consider themselves thoroughly modernized (is this the explanation of continuously rising interest in phenomena like witchcraft in America and Europe?). Yet, such comments do not preclude the great value of a more historical approach to the issues of terminology that keep plaguing the study of witchcraft. Following in detail the introduction of this and parallel notions and the effects of their appropriation by the population can help us to get at the heart of the issues raised by this mesmerizing and at the same time highly productive cluster of images, rumors, accusations, and practices.

The same authors emphasize also that more attention among academics to what they take on board by adopting certain terms – also when taking them over from general parlor in the societies concerned – can make them more sensitive to their own role in the construction of witchcraft as a topic. One can wonder, though, whether in such reflections anthropologists have not a tendency to overestimate their own impact. In the field, the words of 'innocent' anthropologists – that is, those who have not been initiated – have not much power when witchcraft is concerned. After all, what can they see if they have not even the 'second pair of eyes'? Initiated anthropologists on the other hand – and many see initiation also for the researcher as a precondition for acquiring any knowledge in this esoteric field – are often supposed to be in a particularly dangerous situation since as neophytes they will be an easy target for ill-meaning colleagues. The power of the scientific discourse (which, indeed, constantly marks local people's formulations on witchcraft as a different kind of 'science') does not come from the anthropologist, but rather from representatives of powerful institutions like the medical services, the school, maybe the theologian and or the developer. The anthropologist is – certainly while in the field – often caught in an uneasy position between the two discourses.

Yet, it is clear that further analysis of the power of the words that we – not only the informants but also the academics – use, is one of the more promising directions in the field. How terminological issues can be placed at the heart of the analysis through an ethnographic approach, rather than by raising issues of definitions, is shown from a different perspective and with great theoretical sophistication by James Siegel's monograph (2006) on the 'naming of the witch' in the terrible outbursts of violence against witches in post-Suharto Java (end of the 1990s). Criticizing anthropologists' tendency to shy away from the full horrors of such episodes by falling back on a reductive approach – explaining witchcraft as an expression of something else – Siegel proposes rather to follow in detail what people do when they start using the word witch ([*dukun*] *santet*) that can trigger such wild violence. For him, the word is a desperate attempt to name something un-namable. Making a daring link between Kant's notion of the sublime and Freud's of the uncanny, he analyzes the word 'witch' as failing to close off an experience of the uncanny/sublime. Thus the great uncertainties and the violence it can trigger. Again, the challenging way in which Siegel relates witchcraft to experiences of the uncanny – for him a category that is of any time and any place, and always manifests itself in unpredictable way – can serve to place the experiences captured (often only partially) by a notion like witchcraft, at a more general level, making it more than just a sign of some sort of 'otherness' of the society or the people concerned.

Underneath all the changes and reorientations in the academic debates about witchcraft, some constant concerns remain. A crucial challenge to both historians and anthropologists – although the latter seem to wrestle with this more explicitly, due to their direct confrontations with the realities at stake – is how to face the strong presence of these imaginaries in the societies concerned, without contributing to 'othering' them – as irrational, prescientific, and nontransparent, or whatever peculiarities can serve to make these societies the 'Other' of the supposedly modern West. Or, to put it more directly, ignoring the generality of popular preoccupations in many parts of, for instance, Africa – as for early modern Europe – is an impossibility. But how then can we address them without 'primitivizing' the continent or the earlier period? Promising directions for facing this challenge can be found in the recent tendency to confront terminology, not in order to arrive at purer definitions, but rather in a historical and ethnographic sense by following the particular trajectories of these problematic terms in everyday life. Another way to confront this challenge might be through a broader comparative perspective, showing that the issues addressed by a notion like witchcraft emerge in maybe different but still corresponding forms in societies all over the world.

See also: Borders, Anthropology of; Cannibalism; Family: The Anthropology of the Concept and Its History; Kinship in Anthropology; Magic, Anthropology of; Religion, Phenomenology of; Religion: Evolution and Development; Religion: Morality and Social Control; Shamanism; Spirit Possession, Anthropology of.

Bibliography

Ashforth, A., 2005. Witchcraft, Violence and Democracy in South Africa. University of Chicago Press, Chicago.
Briggs, B., 1996. Witches & Neighbours. Fortuna Press, London.
Behringer, W., 2004. Witches and Witch-hunts. A Global History. Polity Press, Cambridge.
Bernault, F., 2009. De la modernité comme impuissance – Fétichisme et crise du politique en Afrique équatoriale et ailleurs. Cahiers D'Etudes Africaines 49 (3), 747–774.
Buyandelger, Manduhai, 2013. Tragic Spirits : Shamanism, Socialism, and the Neoliberal State in Mongolia. University of Chicago Press, Chicago.
Ceriana, M.A., 2014. Sorcellerie et prophétisme en Centrafrique – L'imaginaire de la dépossession en pays banda. Karthala, Paris.
Comaroff, J., Comaroff, J. (Eds.), 1993. Modernity and Its Malcontents. Chicago University Press, Chicago.
Comaroff, J., 2000. Millennial Capitalism: First Thoughts about a Second Coming. Public Culture 12 (2), 291–344 (Comaroff, Comaroff (Eds.), Millennial Capitalsim and the Culture of Neoliberalism, Special issue).
Crick, M., 1979. Anthropologists' witchcraft: symbolically defined or analytically undone? Journal of the Anthropological Society of Oxford 10, 139–146.
Douglas, M. (Ed.), 1970. Witchcraft Confessions and Accusations. Tavistock, London.
Favret-Saada, J., 1977. Les mots, la mort, les sorts. Gallimard, Paris.
Geschiere, P., 1997. The Modernity of Witchcraft, Politics and the Occult in Post-colonial Africa. University Press of Virginia, Charlottesville, VI.
Geschiere, P., 2013. Witchcraft, Intimacy and Trust – Africa in Comparison. University of Chicago Press, Chicago.
Ginzburg, C., 1990. Ecstasies: Deciphering the Witches' Sabbath. Hutchinson Radius, London.

Henry, C., Tall, E.K., 2008. La sorcellerie envers et contre tous. Cahiers d'Etudes Africaines 189 (90), 11–35 (Henry, Tall (Eds.), Territoires Sorciers, special issue).
Kapferer, B., 1997. The Feast of the Sorcerer. Chicago University Press, Chicago.
La Fontaine, J.S., 1998. Speak of the Devil, Tales of Satanic Abuse in Contemporary England. Cambridge University Press, Cambridge, UK.
Mary, A., 2009. Visionnaires et prophètes de l'Afrique contemporaine. Karthala, Paris.
Meyer, B., 1999. Translating the Devil – Religion and Modernity Among the Ewe in Ghana. Edinburgh University Press, Edinburgh.
Meyer, B., Pels, P. (Eds.), 2003. Magic and Modernity, Interfaces of Revelation and Concealment. Stanford University Press, Stanford, CA.
Muchembled, R., 1978. Cultures populaire et cultures des élites. Flammarion, Paris.
Niehaus, I., 2001. Witchcraft, Power and Politics: Exploring the Occult in the South African Lowveld. Pluto, London.
Robisheaux, T., 2009. The Last Witch of Langenburg. Murder in a German Village. Norton & Co, New York.
Sanders, T., 2003. Reconsidering witchcraft: postcolonial Africa and Analytical (un) certainties. American Anthropologist 105 (2), 481–486.
Siegel, J., 2006. Naming the Witch. Stanford University Press, Stanford, CA.
Stroeken, K., 2010. Moral Power: The Magic of Witchcraft. Berghahn Books, Oxford.
Taussig, M., 1987. Shamanism, Colonialism and the Wild Man. Chicago University Press, Chicago.
Thomas, K., 1971. Religion and the Decline of Magic. Weidenfeld & Nicolson, London.
Tonda, J., 2005. Le souverain moderne. Le corps du pouvoir en Afrique centrale (Congo, Gabon). Karthala, Paris.
Whitehead, N.L., Wright, R. (Eds.), 2004. Darkness and Secrecy: The Anthropology of Assault Sorcery and Witchcraft in Amazonia. Duke University Press, Durham, NC.

Wittgenstein, Ludwig (1889–1951)

Eike von Savigny, Universität Bielefeld, Bielefeld, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16517–16522, © 2001, Elsevier Ltd.

Abstract

Ludwig Wittgenstein (1889–1951) was a leading figure in twentieth-century philosophy of language and philosophy of mind. In his Tractatus Logico-Philosophicus, he devised a theory of logical truth based on a naturalistic metaphysics and a referential and speaker-oriented theory of the meaning of declarative sentences. Since its fundamental postulates turned out to be inapplicable to ordinary language, Wittgenstein made a fresh start, in his Philosophical Investigations, by grounding linguistic meaning in a sign's role in social interaction. Relevant ways of using signs have to be socially established; therefore, meaning is essentially public rather than intentional. Applied to first person sensation language, this resulted in public accessibility and social determination of subjective psychological states, all the more so since Wittgenstein treated verbal expressive behavior as a special case of expressive behavior in general. While very influential in philosophy, his ideas took little substantial effect in other areas. Semantic theories of meaning were influenced by the Tractatus, whereas the later ideas are still waiting to be transformed into research programs.

Life

Ludwig Josef Johann Wittgenstein was born in Vienna on April 26, 1889 as the youngest of eight children of Karl and Leopoldine (née Kalmus) Wittgenstein. His father was a protestant, his mother a catholic, and the children were brought up in catholic faith. Wittgenstein was of partly 'Jewish' descent (three-quarters according to Nuremberg laws), and he accepted this classification to the point of self-attributing character traits that counted as Jewish in the Habsburg monarchy of the turn of the century. (Weininger's *Geschlecht und Charakter* impressed him deeply.) The family had become one of the wealthiest of the Habsburg Empire in the iron and steel business, and young Ludwig got his early schooling from private teachers until his father decided to send him to the *Realschule* (technical gymnasium) at Linz in 1903, no doubt hoping him to become an engineer, where he passed his '*Matura*' in 1906.

The Wittgenstein home in Vienna was very *fin de siècle*. On the positive side, the house was open to artists and composers, the family exercising generous Maecenatism; Ludwig read intensively in his youth, and there is documented influence from Goethe, Schopenhauer, Heinrich Hertz, Karl Kraus, Wolfgang Köhler, and others. On the negative side, the children were attended more by staff than by their parents, and the sons had definitely self-destructive tendencies. At least two of Ludwig's brothers, possibly a third one, committed suicide, a tendency which Ludwig had to combat himself for most of his life. Looking back at his childhood, he considered it to have been unhappy.

Having graduated from the Linz *Realschule*, Ludwig went to the Technische Hochschule Charlottenburg (now Technische Universität Berlin) where he studied mechanical engineering and was awarded his 'Diplom' in 1908. He went to Manchester in the same year to do research in aeronautics; however, he began to develop a growing interest in pure mathematics and read Bertrand Russell's *Principles of Mathematics* of 1903 and Gottlob Frege's *Grundgesetze der Arithmetik* of 1893. Fascinated by these books' enterprise of deriving mathematics from logic, he visited Frege in Jena in 1911 who encouraged him to study with Russell in Cambridge. He was a student in Cambridge from 1911 to 1914, but although attending some classes in philosophy (notably with George Edward Moore), his work had more or less the character of an extended dialogue with Russell on questions of logic and the foundations of mathematics.

During his Cambridge time, Wittgenstein experienced the first of a number of deep emotional friendships with young men. Although he was homosexual, the relation with David Pinsent probably remained platonic until it was ended by the latter's death in World War I. One love relationship, with Francis Skinner in Cambridge in the 1930s, is known to have been fulfilled sexually, but Wittgenstein broke it off, just as if unfit for happiness in human bonds, and in addition, deeply ashamed of his disposition.

He volunteered for the Austrian army in 1914 and served, with distinction, throughout the war; extant diaries testify to ongoing work that was to lead up to his only philosophical book published during lifetime, the *Logisch-Philosophische Abhandlung*, which was actually finished in summer 1918 and first appeared, very poorly edited, in Ostwald's *Annalen der Naturphilosophie* for 1921. The book became famous only after an English translation, under the title *Tractatus Logico-Philosophicus*, had been published in 1922 when Wittgenstein, after having trained as an elementary school teacher, was teaching children in villages in Niederösterreich – evidently in order to live a morally meritorious life. However, he had to resign in 1926 because of punishing a schoolboy too severely, and he would probably have given way to despair had not his sister, Margarete Stonborough, asked him to build her new Vienna house together with his friend, the architect Paul Engelmann. (The building now houses the Bulgarian embassy.)

Philosophical contact with Cambridge had never been cut off completely. Back in Vienna, Wittgenstein had some philosophical conversations with members of the 'Vienna Circle,' notably Moritz Schlick and Friedrich Waismann. Thus, although a lecture by L. E. J. Brouwer is said to have given the last push, it was not completely out of the blue that Wittgenstein returned to Cambridge in 1928. Having given away his inherited fortune to his brothers and sisters after the war, he had to apply for a fellowship (submitting the now famous

Tractatus as a doctoral dissertation). This was his way of earning his living until his death, except for holding a chair in Cambridge from 1939 to 1947 and living on savings now and then. He held regular classes in Trinity College, laconically entitled 'Philosophy' or 'Philosophy for Mathematicians.' A difficult person and not popular with most of his colleagues, his teaching success with a small number of students was enormous and accounted for his burgeoning philosophical ideas to be spread far ahead of the posthumous publication of his constantly rewritten notebooks and typescripts.

Except during wartime, Wittgenstein regularly went to Vienna in summer and for Christmas to see his family. Also, he several times went to Skjolden on the Sognefjord in Norway where he had had a little solitary house built before World War I. In 1939, with the assistance of John Maynard Keynes, he succeeded in applying for British citizenship which allowed him to travel in '*Großdeutschland*' even as an Austrian 'Jew' and to arrange for two of his sisters to be declared 'Aryan' in exchange for transferring large valuta parts of the family money to the German Reichsbank. During World War II, he served different jobs outside the university that he considered more important than doing philosophy.

Wittgenstein's second book, the *Philosophical Investigations*, was more or less finished in 1946 – finished in the sense of Wittgenstein himself considering it as printable. But a general mood of dissatisfaction with philosophical results seems to have got an increasing grip on him. Even though he continued writing and discussing philosophical questions, he was suspicious of finished answers. Asking the proper questions was, in his view, the important thing – an attitude which he had developed in the 1930s and which he expressed more and more sharply. In this mood, Wittgenstein did not press for publication. Living outside Cambridge for much of his remaining time, he alternately sought company and solitude until cancer of the prostate, diagnosed late in 1949, had worsened so much that traveling became impossible. He died in Cambridge on April 29, 1951.

Philosophical Work

Wittgenstein's early (*Tractatus*) philosophy can be regarded usefully as an attempt at answering two questions: Why are logical truths true? How can we recognize that they are? He did this by inventing a theory of meaning which, in turn, was based on an ontological theory of the most general features of reality. Starting from the observation that logical truths, as he had come to know them from the work of Gottlob Frege and Bertrand Russell, are always syntactically complex, Wittgenstein first stipulated that there is but one way of constructing complex sentences from components, viz., by combining the components in a complex which is a 'truth function' of the components (if every single component is either true or false, this is sufficient for the complex sentence to be definitely true or definitely false). The components may be complex themselves; but according to a second stipulation, the combinatory process originally starts from components that are elementary sentences, i.e., sentences which are not truth-functional complexes of anything else. If a complex sentence has been constructed from n elementary sentences, there are 2^n different combinations of each of them being either true or false; this is the postulate of mutual independence of the elementary sentences which is characteristic of 'Logical Atomism.' If the complex sentence comes out as true in every such combination, then it is logically true (e.g., 'If it is neither raining nor snowing, then it is not raining' is true, no matter whether or not it is raining and whether or not it is snowing). This is manifest for anybody who understands the complex sentence; thus logical truth 'shows itself,' and both questions above are answered if every elementary sentence is either true or false. Here a third stipulation comes in: Every elementary sentence is a 'picture' of an elementary 'state of affairs' which either exists, in which case the elementary sentence is true, or does not exist, in which case the corresponding elementary sentence is false. Existing elementary states of affairs make up the 'world' as it is. By a picture, Wittgenstein means an isomorphic correlation of the simple signs (which he called 'names') whose concatenations form the elementary sentences and of simple 'objects' which make up the elementary states of affairs. Given a set of elementary states of affairs and a set of elementary sentences, normally there will exist several possibilities to correlate simple signs and simple objects while preserving isomorphy; the actual correlation, according to a further stipulation, is chosen by the speaker who correlates sentences with states of affairs by thinking a state of affairs as a sentence's sense. Since 'names' occur in many different sentences, this thinking of the senses of whole sentences reduces the alternative correlations and thus determines certain objects as the meanings of given names which can form other sentences; so the understanding of new sentences can be explained.

So far, Wittgenstein's stipulations add up to a referential and speaker-oriented theory of the meaning of sentences that explains logical truth. However, he added restrictions two of the most provocative of which are the following: First, the states of affairs that make up the world and are necessary for language to be meaningful are all natural ones; there are no moral or aesthetic states of affairs. Evaluative pronouncements are, therefore, nonsensical. Second, the fact that a state of affairs is correlated with a sentence as the latter's sense, or the fact that a speaker correlates sentences with states of affairs (and in consequence, names with objects) by this or that 'projection rule' is not itself a state of affairs, which means that any description of this correlation is nonsensical too. Wittgenstein accepted this consequence by declaring all *Tractatus* sentences nonsensical, even if helpful, in the book's last propositions. On which Frank P. Ramsey is said to have commented 'What you cannot say you cannot say, and you cannot whistle it either' (an allusion to Wittgenstein's popular gift of whistling entire pieces of classical music). A third restriction often goes unnoticed, viz. that the *Tractatus* theory of meaning is restricted to declarative sentences, excluding questions, imperatives, optatives, and interjections. Wittgenstein hoped these might be incorporated into his framework by proper analysis, without, however, attempting to do so; the same holds for his failure to even look for real examples of elementary sentences.

In a period of hard work from 1929 until about 1936 which it has become customary to call 'intermediate,' Wittgenstein renounced Logical Atomism, i.e., the supposition that elementary sentences are mutually independent; all candidates (e.g., simple color statements) proved to exhibit meaning

relations with other ones of the same kin (what is green is not red). So the idea of systematic dependency being co-responsible for meaning emerged. This fell in line with the new idea of devising a nonrepresentational (nonsymbolic) conception of meaning: pieces in chess have a meaning (a role), without representing anything, simply in virtue of their place in the rules of the game. If linguistic elements played a role in a system of rules for their use in social interaction, this might well suffice for their meaningfulness. However, the appeal to rules implied a problem of its own. If rules are identified by means of their linguistic expressions, there must be meaningful expressions for there to be meaning (an obvious interpretation regress), and if rules have to be agreed upon there must be rules to confer validity on the agreement procedures (an obvious ruling regress known from constitutional and international law). Sticking to the idea of meaning to be found in systems of rules, Wittgenstein invented one of his most famous tools, the 'language-game,' which he elaborated in the posthumous masterpiece of his 'late philosophy,' the *Philosophical Investigations*. A language game may be understood as a set of regularities in behavior, involving several people, where the use of certain elements (to be identified as linguistic on further inspection) on the one hand depends on what has been done and what the other circumstances are, and on the other hand is relevant for further things to be done subsequently, either by the 'speaker,' 'audience,' or by third parties. These regularities in behavior have to be rich enough to differentiate, e.g., statements from commands; thus 'The window is open,' if playing the role of a command, will evoke responses different from when it is used as an answer to a question.

Commands do not qualify as such simply by being followed by someone else's doing something. They occur only where the speaker has the required kind of authority, and the addressee, rather than actually doing something, has to do it. (This is why utterances with given meanings are possible only in a social setting that provides for the characteristic rights and duties–languages are embedded in 'forms of life.') Both facts can, in the event, be analyzed as certain rules being followed by the people concerned, and it is here that Wittgenstein's solution to the double-regress problem about rules sets in. Rather than asking how to identify or how to establish rules, he asked how behavior looks like if it is rule-following behavior. If behavior is rule-following, the rule it follows can be extracted from it in the following way. There is a description of the behavior that makes use of the rule although the people who follow the rule do not make any use of the rule themselves. With Wittgenstein, a behavior *B* follows the rule *R* if *B* conforms to *R*, if it is learnable and if it goes without saying, i.e., is mutually expected among the people concerned. This is not new to present-day sociology, of course, but it served Wittgenstein to cut off both regresses sketched above. Nobody among the people who follow the rule need be able to identify it, and it need not be enacted, or agreed upon.

Anchoring linguistic meaning in practices that are complex in such a way that they can truly be described as rule-following has the consequence that there is no linguistic meaning outside a community. Whether or not an utterance is a command depends on others to expect a certain person now to do something. Unless they expect the person to do so, the speaker has not uttered a command. Meaning an utterance in a certain way is not something a speakers achieve by exhausting their very own mental resources and capacities (it is not due to their capacity of 'intending' in a traditional sense). Saying something meaningful, as well as meaning something, depend on others to be prepared to understand the speaker in a given way. This implies one of Wittgenstein's most famous theses, viz. that there cannot be a language private to one speaker, i.e., a language whose signs have their meanings no matter whether or not others understand them or react to them. However, Wittgenstein argued his critique independently, thus providing indirect evidence for his social account. The only rule available to a 'private meaner' would stipulate something like: '*S* means *x* if and only if I remember *S* to have meant *x* in the past'; since what is to be explained, viz. '*S* means *x*,' occurs in the explanation, the only available rule turns out to be void.

Whereas this famous 'private language argument' has attracted much philosophical debate, a side effect has proved to be more fruitful. Wittgenstein used the language of sensations as his example and so proved, by implication, that it is not possible to report on one's own sensations by way of privileged inner observation. (For in an unproblematic situation of observation, using the right words is a test of the observer's linguistic competence; if others cannot interfere, reports on one's own sensations would have private meanings in the sense defined above.) The idea that there is privileged knowledge of one's own sensations (and other 'conscious' psychological states like moods, emotions, impressions, and the like) originates in the fact that speakers have a special authority in expressing them, and Wittgenstein took two ingenious steps. First, he pointed out that the speaker's privilege is just one case among others of utterances (often prefaced by 'I know' as in 'I myself know best what is good for me') which are privileged according to our social rules rather than epistemically. Second, he interpreted them, rather than as reports, as being meaningful in the same way as nonverbal expressive behavior is, viz. in virtue of playing a recognizable role in a behavioral pattern. Thus, 'Ouch' or 'I am in pain' are meaningful insofar as they occupy the place of crying in the series injury/crying/winning attention (a series that exhibits nicely the structure, which is characteristic of commands, of authority/imperative/expected behavior, or the structure of precondition/utterance/expected reaction that characterizes meaningful utterances in general).

By parity of reasoning, the symmetry of verbal and nonverbal expressive behavior implies a last important result. Just as verbal utterances owe their meanings to others' reactions, so nonverbal behavior expresses what it does in virtue of others' socially established response. Now if one cannot make head or tail of the notion of an inexpressible psychological state (as Wittgenstein argued at length), a person's psychological state is what he or she could express at the moment, and then what that person is feeling, thinking, wishing, etc. is co-defined by how others would react should he or she exhibit expressive behavior. Wittgenstein is, thus, the first philosopher to have touched the topic of what has come to be called 'Social Anti-individualism.' In an unduly neglected paper, Fleming (1978) has likened Wittgenstein's view of the interpretation of expressive behavior to the interpretation of pictures. Pictures owe their content to a civilization's way of seeing them. Their content is manifest because it is defined by established

interpretation. Just so, established responses to other persons' behavior are operative in determining their psychological situation. This is what distances Wittgenstein from Logical Behaviorism. Rather than being inferred from behavior, psychological facts are directly observable; and rather than just being ascertained by observation, they are dependent on external interpretation.

Although Wittgenstein developed his 'middle' and later philosophy in opposition to and correction of his *Tractatus* ideas, there is considerable continuity both as regards content and methodology. In point of content, Wittgenstein is sure to be the first philosopher in history who developed his philosophical ideas from thoughts about how it is that the elements of language have meaning. In this respect, our linguistic competence plays a role, with Wittgenstein, not unlike that of our epistemic abilities in Kant's far-reaching deductions outside epistemology. As regards methodology, Wittgenstein insisted, both in the *Tractatus* and in later writings including the *Philosophical Investigations*, that good philosophy has a peculiar task, i.e., that of clarifying obscurities that have been and are constantly being produced, either when scientists misunderstand what they are doing or when scientific or everyday questions and experiences are misunderstood and linguistically distorted by bad philosophy. This continuity has often been stressed, but for at least two reasons it should not be overestimated. First, the task of clarifying was all that was left over for philosophy in the *Tractatus* since its theory of meaning reserved all meaningful statements for the natural sciences. Second, the assignment was a piece of explicit methodology which Wittgenstein did not obey in practice. In contrast, Wittgenstein's later conception of meaning admits to philosophy every useful language-game (e.g., that of describing language-games), the assignment of the task of clarifying being motivated by his experience of how much linguistic distortion his own early philosophy had suffered from; and clarification meant method rather than methodology–there is sure to exist no other corpus of philosophical writing so densely packed with repeated demonstrations of how translating philosophical questions and theses into everyday language provides clarity. As an original contribution to our picture of what philosophy can achieve, Wittgenstein developed this clarificatory way of handling philosophical thinking into intellectual therapy. The real aim is achieved when the philosopher, by way of rectifying his distorted use of language, has come to see that there has never been any philosophical problem except for the riddles he got entangled with by falling prey to the temptation of highbrow speaking. Wittgenstein went so far as to say that after such therapy, there was no place in philosophy for controversial theses or theories. Whereas many take these declarations at face value, others have argued that in view of the substantial and controversial achievements of his later philosophy, the declarations should be either interpreted in a restrictive way or taken as overstating an otherwise valuable methodological warning.

Impact

Wittgenstein's influence on the intellectual life of the twentieth century has not been uniform or coherent; early and late philosophy, written work and teaching contact, work and personality have appealed to different interests. The general public has been and still is mostly interested in his disconsolate life; following the arbitrary turns of fashion, he has been canonized in a way that would, to his utter disgust, have nowadays turned him into an unwilling victim of talk shows. Furthermore, Wittgenstein's prose style has been all the rage since the *Tractatus* was published, and whereas every Tom, Dick, and Harry arranged their texts in decimal notation before 1953, quoting aphorisms from the *Philosophical Investigations* became imperative for feuilletonists in Anglo-Saxonia from that date, and in the German speech area after 1960 when Suhrkamp had timidly published a joint edition of both of Wittgenstein's masterpieces (together with a '*Beiheft*' in order to secure respectability for an unknown Austrian–British philosopher by documenting the interest of poets like Ingeborg Bachmann).

As regards more serious questions, research has been affected differently by the *Tractatus* and by subsequent work, and by the latter differently depending on whether people were confronted with Wittgenstein's teaching in Cambridge or with his writing. The *Tractatus* gained a great popularity soon after the publication of the bilingual edition in 1922, and it exerted considerable influence on some of the members of the 'Vienna Circle' who had gathered around Moritz Schlick and were striving for a really scientific philosophy. The most enduring effect was to be the 'verification theory of meaning' which, as stated in the *Tractatus*, identified any proposition's meaning with the set of distributions of 'true' and 'false' to its elementary sentences which make the proposition true; elaborated by the circle's most influential member, Rudolf Carnap, it made its way into logic, philosophical semantics (the present-day truth-condition semantics), and is even influential in linguistic semantics. Karl Raimund Popper's disclaimers notwithstanding, the *Tractatus* theory that only those sentences have sense which risk to wreck on experience is a clear forerunner of his falsifiability theory of cognitive significance, influential in sociology and psychology since the 1940s.

In his philosophizing since 1929, Wittgenstein influenced some bright Cambridge students who learnt, from his live example, that being sober, honest, and clear was more important than, and at least as attractive as, producing high-faluting and sloppy nonsense. Following the advice to rephrase any philosophical sentence in everyday idiom, they began to form a veritable movement, Ordinary Language Philosophy, which had its high tide in the 1950s and early 1960s. The most famous and most important representative was Gilbert Ryle, Oxford based but on friendly terms with Wittgenstein, whose seminal *Concept of Mind* of 1949 inaugurated Logical Behaviorism, to be superseded by Functionalism; so it is a safe guess that without Wittgenstein, present-day research on the mind-body problem would look entirely different. Further effects resulted from the publication of the *Philosophical Investigations*, jointly in German and English, in 1953; the impact in philosophy was heavy, but surprisingly small in linguistics (the reason being, perhaps, that John L. Austin's speech act theory won the day). At present, the most interesting developments, in the behavioral sciences outside philosophy, seem to be going on in Cognitive Science and in Ethnomethodology. In Cognitve Science, Hubert Dreyfus and others have argued forcefully that intelligent achievements

cannot be accounted for by supposing that subjects are following explicit rules but have to follow them implicitly in Wittgenstein's sense. This has been condensed into the prediction that the architecture of the brain's working is to be regarded as connectionist rather than computational. In Ethnomethodology, Peter Winch and others have suggested that without a Wittgensteinian conception of forms of life, an adequate understanding of alien societies cannot be attained.

On the whole, however, Wittgenstein's influence, on the behavioral sciences, has remained marginal outside philosophy. In particular, social anti-individualism is still awaiting proper conversion into research strategies, just as is an adequate use of Wittgenstein's conception of language games in linguistic pragmatics.

See also: Atomism and Holism: Philosophical Aspects; Communicative Competence: Linguistic Aspects; Ethnomethodology, General; Individualism versus Collectivism: Philosophical Aspects; Interactionism, Symbolic; Language and Society; Logical Positivism and Logical Empiricism; Meaning and Rule Following; Science and Technology Studies, Ethnomethodology of; Semantics; Truth and Credibility in Science.

Bibliography

Baker, G.P., Hacker, P.M.S., 1980. An Analytical Commentary on the 'Philosophical Investigations,' vol. I: Wittgenstein, Understanding and Meaning. Blackwell, Oxford, UK.

Baker, G.P., Hacker, P.M.S., 1985. An Analytical Commentary on the 'Philosophical Investigations,' vol. II: Wittgenstein, Rules, Grammar and Necessity. Blackwell, Oxford, UK.

Drudis-Baldrich, R., 1992. Bibliografia Sobre Ludwig Wittgenstein. Aporia, Madrid.

Fleming, N., 1978. Seeing the soul. Philosophy 53, 33–50.

Fogelin, R.J., 1976/1987. Wittgenstein. Routledge and K. Paul, London and New York.

Frongia, G., McGuinness, B.F., 1990. Wittgenstein: A Bibliographical Guide. Blackwell, Oxford, UK.

Glock, H.-J., 1996. A Wittgenstein Dictionary. Blackwell, Oxford, UK.

Griffin, J.P., 1964. Wittgenstein's Logical Atomism. Oxford University Press, Oxford, UK.

Hacker, P.M.S., 1990. An Analytical Commentary on the 'Philosophical Investigations,' vol. III: Wittgenstein, Meaning and Mind. Blackwell, Oxford, UK.

Hacker, P.M.S., 1996. An Analytical Commentary on the 'Philosophical Investigations,' vol. IV: Wittgenstein, Mind and Will. Blackwell, Oxford, UK.

ter Hark, M., 1990. Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology. Kluwer, Dordrecht, The Netherlands.

Kenny, A., 1973. Wittgenstein. Allen Lane, London.

Lapointe, F.H., 1980. Ludwig Wittgenstein: A Comprehensive Bibliography. Greenwood Press, Westport, CT.

McDonough, R.M., 1986. The Argument of the Tractatus. State University of New York Press, Albany, NY.

McGuinness, B.F., 1988. Wittgenstein: A Life, Band I, Young Wittgenstein. University of California Press, California.

McGuinness, B.F., Schulte, J. (Eds.), 1989. Logisch-Philosophische Abhandlung. Frankfurt a.M., Suhrkamp.

Monk, R., 1990. Ludwig Wittgenstein, The Duty of Genius. Free Press, London.

Mounce, H.O., 1981. Wittgenstein's Tractatus: An Introduction. Blackwell, Oxford, UK.

Philipp, P., 1996. Bibliographie zur Wittgenstein-Literatur. Wittgenstein Archives, Bergen, Norway.

von Savigny, E., 1988–1996. Wittgenstein's 'Philosophische Untersuchungen.' Ein Kommentar für Leser. Frankfurt a.M., Klostermann.

St. Hilmy, S., 1987. The Later Wittgenstein: The Emergence of a New Philosophical Method. Blackwell, Oxford, UK.

St. Shanker, V.A., Shanker, S.G., 1986. A Wittgenstein Bibliography. Croom Helm, London.

Schulte, J., 1989/1992. Wittgenstein. State University of New York Press, Albany, NY.

Schulte, J., 1993. Experience and Expression. Oxford University Press, Oxford, UK.

Stern, D., 1995. Wittgenstein on Mind and Language. Oxford University Press, Oxford, UK.

Wittgenstein, L., 1922/1961. Tractatus. Routledge, London.

Wittgenstein, L., 1953/1967. Philosophical Investigations. Blackwell, Oxford, UK.

Wittgenstein, L., 1984. Werkausgabe. Frankfurt a.M., Suhrkamp.

von Wright, G.H., Nyman, H., Schulte. J.(Eds.), in press Philosophische Untersuchungen. Frankfurt a.M., Suhrkamp.

Women and Criminal Justice

Loraine Gelsthorpe, Institute of Criminology, University of Cambridge, Cambridge, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article outlines some general themes concerning perceived judicial paternalism toward women offenders, punishment for offences against gender roles, and the increasing use of imprisonment for women, as well as highlighting some positive policy developments concerning women and criminal justice and what we know about 'what works' with women. The article includes brief concluding reflections on what has still to be addressed in terms of recognizing women offenders as victims and not just offenders. Women's experiences of criminal justice, wherever they are, may also reflect complex intersections of injustices that characterize their lives.

The aim of this article is to provide an overview of issues relating to women offenders and criminal justice. There are three main sections:

1. general themes, which relate to women, crime, and criminal justice;
2. concerns about the increasingly high rate of imprisonment of women in a number of countries; and
3. policy developments concerning women and criminal justice and what we know about 'what works' with women.

The article includes brief concluding reflections on what has still to be addressed.

Introduction

In all countries where statistics are collected in routine fashion, the majority of convicted offenders and those given court sentences are men. Broadly speaking, offending by women differs in a number of important ways from offending by men: it is less common, less frequent, and generally less serious. In western countries, especially, the USA, Australia, Canada, and New Zealand, women who offend largely commit property-related offences (bearing in mind different legal terms and definitions in an international context).

What Do We Know about Women Offenders and Their Criminogenic Needs?

Historically, the study of crime has been male-centered and the criminal justice system's treatment of women has been imbued with assumptions about men. When the female offender has not been ignored, she has been pathologized and treated as abnormal. It is only recently that we have begun to recognize more nuanced accounts of women's pathways into crime, with acknowledgment that women do not offend as much as men and that their offending behavior is primarily of a less-serious nature. There have been periodic moral panics about women's crime, particularly in relation to violent crime and 'girl gangs,' but it is hard to sustain the arguments in this direction in light of the research evidence and statistical analysis of patterns of crime (Chesney-Lind and Pasco, 2013; Sharpe, 2012). Poverty features highly in women offenders' lives (Carlen, 1988; Wacquant, 2009a). Moreover, there have been consistent messages from the research literature on women offenders and from experienced service providers and service users (Sheehan et al., 2007) that women offenders tend to have a history of unmet needs in relation to sexual and violent victimization, physical and mental health, housing and income, and training and employment. Substance misuse and child care responsibilities often compound these problems. Desistance studies relating to pathways out of crime serve to highlight the complexity of factors relating to women's pathways into crime, while pathways out of crime point to the need for broadly based provision that can be individually tailored. Service providers in prisons and in probation/community corrections as well as researchers support claim that there may well be indirect relationships between abuse and mental health, labor market participation, and substance abuse, all of which are associated with risk of reoffending. In other words, 'victimization' creates psychological sequelae, which can lead to offending behavior (Sheehan et al., 2011).

What Do We Know about Women and Criminal Justice?

In a number of countries very little is known about possible differences in the treatment of men and women, but equally, several authors have reviewed and/or researched the respective treatment of women and men by the courts, although few offer straightforward support for 'chivalry' or leniency toward women even though this is an oft quoted claim about the treatment of women in criminal justice systems (see Gelsthorpe, 2001). Many British and American researchers, for example, point to the complexities in sentencing. In the UK, researchers have observed that women conforming to conventional roles are treated more leniently than those who do not. Hedderman and Gelsthorpe (1997) have pointed two differences drawn between 'troubled' and 'troublesome' women, the former being more deserving of leniency than the latter, though with little to distinguish them in reality. In the USA, Daly (1994) found that it was children and the family, rather than just women themselves, who were the focus of 'chivalry' or 'judicial paternalism' as it has been termed, as the courts sought to support and conserve the fabric of society.

Women's low share of recorded criminality has significant consequences for those women who do offend. They are often

 http://dx.doi.org/10.1016/B978-0-08-097086-8.28100-2

seen to be 'doubly deviant,' for example, for having transgressed not only social norms but gender norms too. As a result they may feel that they are doubly punished – especially when informal sanctions and stigmatizing processes are taken into account. There is double jeopardy in women offenders being seen as worse than men; they are both failing individuals and failed women. One striking example of the impact of stigma, mainly on a female group of the families of serious offenders is offered by Condry (2006) who observes that relatives of serious offenders in the UK are shamed and that this stigma has its roots in ideas about familial blame and contamination.

There is particular concern relating to risk/needs confusions in dealing with women offenders. Correctional authorities have become increasingly concerned with assessing the risks and needs of offenders in recent years. This move toward risk/needs assessment can be seen as a contemporary manifestation of two residual concerns: predicting future offending and determining 'what works.' However, as Canadian researchers Hannah-Moffat and Shaw (2003) have highlighted, with international reverberations, the research behind these assessment tools rarely considers the gendered, racialized, or otherwise stratified characteristics of risk and need. Moreover, women offenders are typically characterized as scoring high on 'needs' and low on 'risk' but that in practice, sentencers sometimes confuse the two with consequential up-tariffing of women in the sentencing system or in terms of 'risk' coming to dominate overall approaches to women offenders. One such example concerns the 1990 Task Force on Federally Sentenced Women in Canada, entitled *Creating Choices*. The Task Force recommended the closure of the only women's penitentiary and replacement with regional facilities, which would recognize the low risk, specific needs and experiences of women, but alongside new initiatives, there has been a growth in high security arrangements, and the use of assessment tools based on men (Hannah-Moffat and Shaw, 2003).

Women's Imprisonment

One significant concern in recent years relates to the number of women in prison across a range of legal jurisdictions (relative to the seriousness and frequency of their crimes). The World Female Imprisonment List is based on statistics gathered from 187 prison systems in independent countries and dependent territories. It shows the percentage of women and girls within each national prison population. The world list is compiled from a variety of sources – in most cases from the national prison administration of the country concerned or from the relevant ministry of justice or equivalent (see: http://www.prisonstudies.org/info/worldbrief/wpb_about.php).

At a general level, in 2006, it was reported that more than half a million women and girls were held in penal institutions throughout the world, either as pretrial detainees (remand prisoners) or having been convicted and sentenced (Walmsley, 2006). But more specifically, 31.3% of the prison population in Monaco are women, 19.1% in Hong Kong (China), 8.7% in the USA, 8.3% in the Russian Federation, 8% in Chile, 7.5% in Spain and Australia, 5.9% in New Zealand, 5.5% in Scotland (UK), and 5.3% in Norway. In Scotland women represent about 5.5% of the overall prison population, but this is 4.6% in England and Wales and only 3% in Northern Ireland (Ministry of Justice, 2012; Berman, 2009). Interestingly, in African countries women constitute a much smaller percentage of the total in prison, although women may be subject to other forms of control and coercion of course (ICPS, 2013).

Increases in Women's Imprisonment

Women's imprisonment has increased dramatically in most Western jurisdictions at least; certainly in England and Wales and in the USA and Australia over the past 15–20 years as shown by increases in the numbers of women given custodial sentences (Kruttschnitt and Gartner, 2003). (N.B. There are a number of ways in which rates of imprisonment can be analyzed. The number of receptions or admissions to custody provides an indication of how many individuals are incarcerated over a given period of time – usually 12 months.) In Australia, the number of female prisoners increased by 57% between 1999 and 2009 (Australian Bureau of Statistics, 2009). The increase in women's imprisonment has been particularly dramatic in the USA. Between 1986 and 1995, the number of women imprisoned for drug offences in the USA increased by 888% for example, with the increase being especially marked in states which had introduced severe penalties for drug offences and particularly among black women (Frost et al., 2006).

There are similar increases reported in New Zealand (Harpham, 2010), Spain (Alameda, 2005) and, according to a Council of Europe survey completed by the Quaker Council for Europe, in Finland, Greece, the Netherlands, Germany, and Poland too; in half of the 28 states surveyed there had been increases in the number of women in prison (Quaker Council for European Affairs, 2007).

Data suggest that between 1996 and 2009 the women's prison population in England and Wales increased from 2300 to 4300 women and from 4.1 to 5.1% of the total prison population (Berman, 2009). Similarly, the average female prison population in Scotland almost doubled between 1999/2000 and 2008/09, from 210 to 413 and there was also a huge increase in the number of women remanded into custody (McIvor and Burman, 2011). (There are some notable exceptions to the pattern. In Canada, for example, Gartner et al. (2009) suggest that the female prison population overall has remained relatively stable, although with a marked increase in federal (state/national) prisons rather than in provincial (local) prisons.) Moreover, there is need for care in interpreting increases because large percentage increases can mask relatively small changes in numbers. But an important caveat here is that the increases around the world have been sufficiently large to cause consternation among policymakers, politicians, and academics alike.

Why Have There Been Increases in Women's Imprisonment?

There is seemingly little evidence that more women are being imprisoned because of an increase in serious offending. Rather, it seems that a complex range of factors are responsible, including legislative and policy changes that have impacted on women, as well as there being increasingly punitive responses to women who offend and suggestion that prisons may have become 'modern day social services' for

women in the absence of appropriate provision in the community. In the United States, for example, Kruttschnitt and Gartner (2003) suggest that the growth in the state prison population is due to an increase in arrests (particularly for drugs offences), increase in the length of sentences, and an increased likelihood of being returned to prison when in breach of the conditions of parole. Indeed, the 'war on drugs' might well be interpreted as a 'war on women,' with minority women being disproportionately represented in those imprisoned for drug offences (Chesney-Lind and Pasco, 2013). Thus, the introduction of mandatory minimum sentences for drug offences has had a direct connection with the increase in the number of women in prison. One problem here is that seemingly 'gender-neutral' sentencing laws often fail to distinguish between drug traffickers and drug 'mules' (couriers) who are often persuaded to carry drugs across jurisdictions with the promise of money for food and shoes for children.

There has been a significant development in England and Wales in this regard, with the Sentencing Council implementing changes to sentencing practice so as to recognize the vulnerability of some 'drug mules' (who tend to be women). The change in approach to 'drug mules' forms part of new comprehensive rules on drugs offences from the Sentencing Council for England and Wales (2012).

Other commentators have pointed to increases in arrest rates. Schwartz and Steffensmeier (2007), for example, point to and then explain an increase in female arrest rates in the United States by referring to an increased tendency to arrest juvenile offenders. New technologies for detecting drug use (such as urinalysis) may also have given rise to an increase in parole violations in the United States for failing random drug tests, with women being more likely than men to be imprisoned for a technical parole violation of this sort than for further offending (Chesney-Lind and Pasko, 2013).

Notwithstanding difficulties in discerning whether or not there have been *real* changes in girls' and young women's violent behavior or simply changes in public and police responses, it has also been suggested that the increased rate of imprisonment in the United States can be accounted for by increases in violent behavior among women (Yin, 2008). A similar conclusion was reached in relation to increases in the use of custody for women in Australia (Gelb, 2003), although it should be noted that there is no corresponding change in self-reported studies (Zahn et al., 2008). In England and Wales, there have been concerns about increases in girls' and young women's violence, although it is suggested that there is more media hype than real increases (Sharpe, 2012).

In England and Wales, the reasons for increases in women's imprisonment seem similarly complex. Gelsthorpe and Morris (2002) investigated changes in the nature and seriousness of women's crime, changes in the type of woman sentenced to imprisonment (in terms of previous convictions and so on), a move toward equal sentences for men and women, as well as changes in sentencing patterns, and increases in the length of women's prison sentences. Their overall conclusion was that women were committing more offences in 2002 than in 1992 but that there had been only a slight change in the *nature* of women's offending, mainly with respect to drugs offences (although mostly involving possession of drugs rather drug selling – in other words, not the most serious drugs offences). The 'type of women' imprisoned remained much the same as in 1992: most women were criminally unsophisticated, serving their first custodial sentence for property-related offences. They also noted that a general increased punitiveness (reflecting successive governments' 'tough on crime' stances) and the increased use of short custodial sentences may have impacted on women in particular.

Concerns about the Treatment of Women in Prison

Prison conditions continue to be of concern, from high security levels to strip searching, from the paucity of programs and appropriate training and work opportunities for women to the poor provision for women with children (with relatively few mother and baby units or appropriately situated play areas for young visitors). In England and Wales, Scotland, and Northern Ireland, respective reports signify that nearly two-thirds of women sentenced to imprisonment are responsible for children less than 18 years of age (see The Prison Reform Trust, 2012). We might add to these concerns, those about visiting distances for the families of imprisoned women in England and Wales (women's prisons being unevenly scattered throughout the country). Other concerns revolve around women's unmet health needs and the high rate of self-harm in prisons – not only in the UK (Plugge et al., 2006) but in other countries too (see, for example, World Health Organization data: van den Bergh et al., 2011). In sum, the deleterious effects of imprisonment are indisputable; women's general vulnerability makes the experience worse – not only for themselves but for their children too.

The United Nations Rules for the Treatment of Women Prisoners and Noncustodial Measures for Women Offenders (the *Bangkok Rules*) were approved by the Third Committee of the General Assembly at its 65th session in New York on 15 October 2010 (UN General Assembly, 2010). The rules not only acknowledge that many of the existing prison facilities worldwide were designed for male prisoners, but recognize that women prisoners are one of the vulnerable groups who have specific needs and requirements. The rules also emphasize that when deciding on pretrial measures for a pregnant woman or where a woman is a child's sole or primary caretaker, noncustodial measures should be preferred wherever possible, with custodial sentences only being considered where the offense is serious or violent. Rule 6 highlights the need for "comprehensive health screening to determine primary health care needs," including (under Rule 6e) "Sexual abuse and other forms of violence that may have been suffered prior to admission." Moreover, Rules 10, 11, 12, and 13 draw attention to the need for gender-specific health care and mental health care. Specific mention is made (in Rule 13) that "Prison staff shall be made aware of times when women may feel particular distress, so as to be sensitive to their situation and ensure that the women are provided adequate support." Under Rule 25 (1) it is stated that "Women prisoners who report abuse shall be provided immediate protection, support and counselling, and their claims shall be investigated by competent and independent authorities, with full respect for the principle of confidentiality. Protection measures shall take into account specifically the risks of retaliation." And under Rule 26 we learn

that "Women prisoners' contact with their families, including their children, their children's guardians, and legal representatives shall be encouraged and facilitated by all reasonable means. Where possible, measures shall be taken to counterbalance disadvantages faced by women detained in institutions located far from their homes." Yet there is wide recognition that countries fall short of these rules (see, for example, Malloch and McIvor, 2013).

In her searching review of vulnerable women in the criminal justice system in England and Wales in 2007, requested by the government following the death of six women in rapid succession in a single prison in England, Baroness Corston lambasted prison conditions and noted that what was needed was "a distinct, radically different, visibly-led, strategic, proportionate, holistic, woman-centred, integrated approach" (Corston, 2007).

Research and Policy Developments

The history of work with offenders is not replete with success, but the research base (particularly the meta-analyses) developed since the early 1990s in the UK and other parts of Europe (drawing especially on work in Canada) now strongly supports the position that effective work with offenders is possible. But there has been relatively little attention to 'what works with women offenders' among policymakers in different countries, leading to the claim that women are "correctional afterthoughts" (Ross and Fabiano, 1986). While the focus of campaigns has commonly been in the high numbers of women in prison, the lack of gender sensitivity in prisons, and lack of attention to women's specific needs, in the Western world in particular, there have been recent attempts to focus more directly on what we know about women offenders and their needs, and based on this, to divert them from crime in general and prison in particular.

Gender Responsiveness

Notwithstanding various developments in programs designed to address offending behavior in the 1990s, there are controversies regarding their suitability for all offenders (women, black and ethnic minorities, young offenders, offenders with mental health problems, and very high risk and psychopathic offenders). Offender treatment programs are typically designed for white, male, adult offenders, with only minor adaptations for other groups of offenders.

Indeed, the criminogenic needs that have emerged from research on men have typically been applied to women offenders. The problem is encapsulated very well in the reference to women as "correctional afterthoughts" (Ross and Fabiano, 1986). However, as indicated above, it has been argued that the concepts of 'risk' and 'need' are themselves 'gendered,' and that differences between men and women should be taken into account in devising intervention programs. There are also claims that what works for men will work for women too. But this is not so, and even promoters of the popular cognitive skills programs now acknowledge that there may be an issue around the responsivity of female offenders to cognitive skills programs.

One recent study in England and Wales has tried to take things forward by exploring women's lower rate of completion on the community-based General Offending Behaviour Programme. The study indicates that despite some similarities, the predictors of program completion not only vary for men and women, but also operate differently between them. The findings support the 'gender-responsiveness' position that men and women should be approached differently, and suggest, moreover, that men are more likely to engage in instrumental compliance and women more likely to achieve normative compliance (relating to compliance achieved though self-interested calculation and compliance based on values and the legitimacy of interventions, respectively) (Martin et al., 2009).

Attempts to facilitate gender responsiveness in programs and interventions have been hampered because the state of research and program development for women generally lags behind that for men, but we are able to deduce what is likely to work with women from educational and other areas of research as well as from emerging research findings. From educational research we learn most women prefer to learn in collaborative, rather than competitive, settings. If we put this alongside evidence, which suggests that women-only environments facilitate growth and development, we can see that the evidence adds up to a need to work with women in nonauthoritarian cooperative settings, where women are empowered to engage in social and personal change. In addition, 'women-specific' factors such as health care, child care, and mental health together with factors relating to race and gender combined need to be addressed (see Gelsthorpe, 2010).

New Initiatives

On the basis of analysis of work in Canada, Blanchette and Brown (2009) take us further in advocating the integration of a number of gender-informed theories and methodologies in responses to women offenders. Specifically, they recommend gendered pathways, the use of relational theory, strengths-based approaches, and the use of positive psychology, all of which are critical frameworks for intervention with women. We might add to this, the need for such interventions to be sensitive to 'trauma.'

There have been some important new initiatives that reflect this kind of thinking. In Scotland, we can turn to Centre 218, which has its origins in reviews following a series of suicides in Scotland's only prison for women. The Centre is based on the idea that female offenders should be able to get 'time out' of their normal (and perhaps chaotic and stressful) environment without resorting to 'time in' custody. The Centre thus serves as a diversion from prosecution and as an alternative to custody, and more generally it offers particular support – for detoxification, support and outreach to health, social work, or housing services. The ethos is therapeutic in intention and there is much emphasis on providing a safe environment for the women (Beglan, 2013).

There have been some interesting developments in relation to women and criminal justice in England and Wales too, including major funding and effort to divert women from both crime and custody (Hedderman, 2011; Gelsthorpe, 2010). Holistic services based in women's community centers have been set up in order to support women offenders (see Women's Breakout: http://www.womensbreakout.org.uk/). Of course,

there are questions to ask: do the developments reflect a new sense of responsibility toward disadvantaged women or might they be interpreted as an extension in the network of control and regulation (Wacquant, 2009b). What is perhaps at issue now is how the centers can be sustained in a market economy of justice and how they can provide hard evidence of positive impact beyond consumer views (Gelsthorpe and Hedderman, 2012).

In the United States, the National Institute of Corrections in cooperation with the University of Cincinnati (University of Cincinnati, 2013) has created a series of new gender-specific risk/need assessments for adult, women offenders. The assessments include: (1) a full instrument, the Women's Risk/Needs Assessment, which assesses both gender-neutral and gender-responsive factors and affords separate forms for probation, prison, and prerelease and (2) the Women's Supplemental Risk/Needs Assessment which is designed to supplement existing risk/needs assessments. The Women's Supplemental Risk/Needs Assessment is also available in separate forms for probation, prison, and prerelease populations.

Elsewhere in the United States, there are developments relating to gender-responsive programming in prisons (Lawston, 2013), and in Victoria, Australia, developments regarding diversion from prison and better community provision in relation to drug and alcohol support, family reunification, debt management, and independent living skills for example (Sheehan, 2013).

It is obviously impossible to provide illustration of all the new developments across different jurisdictions, but there is emerging evidence that provision for women offenders in the community should be women-only to foster safety and a sense of community and to enable staff to develop expertize in work with women; integrate offenders with nonoffenders so as to normalize women offenders' experiences and facilitate a supportive environment for learning; foster women's empowerment so they gain sufficient self-esteem to directly engage in problem-solving themselves, and feel motivated to seek appropriate employment. Thus what is required is the utilization of what is known about effective learning styles with women; a holistic and practical stance to helping women to address social problems which may be linked to their offending; links with mainstream agencies, especially health, debt advice, and counseling; capacity and flexibility to allow women to return for continued support and development where required, and arrangements for women to have a supportive milieu or mentor to whom they can turn when they have completed any offender-related programs. In addition, it is important to provide women with practical help with transport and child care so that they can maintain their involvement in the center or programme (Gelsthorpe, 2010).

At the same time, and thinking particularly of developing countries here, 'improvements' may amount to no more than the institution of a single family visit each year for women in prison. There have been no wide sweeping reforms, merely 'pockets' of progress.

Concluding Reflections

Increasingly, political swings and fiscal controls exacerbate the challenges in addressing women offenders' needs in appropriate and proportionate fashion. Across different jurisdictions and among academics, policymakers, practitioners, and campaigners there has been growing recognition that prison is not an appropriate place for women in conflict with the law. The prison has often been a focus for concerns about human rights violations, but a human rights perspective alone has not brought the much needed reforms (see, for example, Moore and Jemphrey, 2013). Moreover, too little attention has been given to human rights dimensions of community sanctions (regarding potential intrusion into family life or the lack of services that are specifically tailored for women).

What is recognized is that women in conflict with the law are often victims of abuse, subject to domestic violence, and poor. Power relationships shape vulnerability and pathways into crime. Some judicial and public support has been garnered for decarceration and community-based alternatives in some countries, but there is much further to go. Moreover, steadily increasing recognition of women offenders' victimization in some western countries is one thing, recognition of women's victimization as they are caught up in human trafficking and yet prosecuted for being illegal migrants or without relevant papers is quite another (Hales and Gelsthorpe, 2012; Troshynski, 2012).

At the beginning of the twenty-first century, new crimes and concerns have emerged: terrorism, cybercrime, human trafficking, and other human rights violations for instance, all of which have produced new ways of women being involved in crime as offenders, or indeed, as victims – a topic not covered here, except in passing (see Gelsthorpe and Larrauri, 2013). Moreover, it is always a challenge to draw upon experiences in non-UK and nonwestern countries, largely because the topic of women and criminal justice has received limited academic attention, but it is important to conclude with recognition that women's experiences of criminal justice, wherever they are, may reflect complex intersections of injustices that characterize their lives.

See also: Criminal Justice and Social Work Practice.

Bibliography

Almeda, E., 2005. Women's imprisonment in Spain. Punishment and Society 7 (2), 183–199.

Australian Bureau of Statistics, 2009. Prisoners in Australia, 2009. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4517.0Main+Features12009 (accessed 16.08.13.).

Balfour, G., Comack, E. (Eds.), 2007. Criminalizing Women. Fernwood Publishing, Halifax.

Beglan, M., 2013. The 218 experience. In: Malloch, M., McIvor, G. (Eds.), Women, Punishment and Social Justice. Human Rights and Penal Practices. Routledge, Abingdon, pp. 152–166.

van den Bergh, B., Gatherer, A., Frazer, A., Moller, L., 2011. Imprisonment and women's health: concerns about gender sensitivity, human rights and public health. Bulletin of the World Health Organization 89, 689–694. http://www.who.int/bulletin/volumes/89/9/10-082842/en/ (accessed 01.09.13.).

Berman, G., 2009. Prison Population Statistics. House of Commons Library, London.

Blanchette, K., Brown, S., 2006. The Assessment and Treatment of Women Offenders. John Wiley and Sons, Chichester.

Carlen, P., 1988. Women, Crime and Poverty. Open University Press, Milton Keynes.

Carlen, P. (Ed.), 2002. Women and Punishment. The Struggle for Justice. Willan Publishing, Cullompton.

Chesney-Lind, M., Pasko, L., 2013. The Female Offender. Sage, Thousand Oaks, CA.

Condry, R., 2006. Stigmatised women: relatives of serious offenders and the broader impact of crime. In: Heidensohn, F. (Ed.), Gender and Justice: New Concepts and Approaches. Willan, Cullompton, pp. 96–120.
Corston, J., 2007. The Corston Report: A Report by Baroness Jean Corston of a Review of Women with Particular Vulnerabilities in the Criminal Justice System. Home Office, London.
Daly, K., 1994. Gender, Crime and Punishment. Yale University Press, New Haven.
Frost, N., Greene, J., Pranis, K., 2006. Hard Hit: The Growth in the Imprisonment of Women, 1977–2004. Institute on Women and Criminal Justice, New York, NY.
Gartner, R., Webster, C.M., Doob, A.N., 2009. Trends in the imprisonment of women in Canada. Canadian Journal of Criminology and Criminal Justice 51 (2), 169–198.
Gelb, K., 2003. Women in prison: why is the rate of incarceration increasing? In: Paper Presented at Evaluation in Crime and Justice: Trends and Methods. ACT, Australia, Canberra.
Gelsthorpe, L., 2001. Critical decisions and processes in the criminal courts. In: McLaughlin, E., Muncie, J. (Eds.), Controlling Crime. Sage in association with the Open University, London, pp. 101–156.
Gelsthorpe, L., 2010. What works with women offenders. In: Herzog-Evans, M. (Ed.), Transnational Criminology Manual, vol. 3. Wolf Publishers, Nijmegen, NL, pp. 223–240.
Gelsthorpe, L., 2011. Working with women offenders in the community: a view from England and Wales. In: Sheehan, R., McIvor, G., Trotter, C. (Eds.), Working with Women Offenders in the Community. Willan, Cullompton, pp. 127–150.
Gelsthorpe, L., Hedderman, C., 2012. Providing for women offenders: the risks of adopting a payment by results approach. Probation Journal 59 (4), 374–390.
Gelsthorpe, L., Larrauri, E., 2013. Gender and crime in Europe. In: Body-Gendrot, S., Hough, M., Klarezsi, K., Levy, R., Snacken, S. (Eds.), The Routledge Handbook of European Criminology. Routledge, Abingdon, pp. 188–203.
Gelsthorpe, L., Morris, A., 2002. Women's imprisonment in England and Wales. Criminal Justice 2 (3), 277–301.
Hales, L., Gelsthorpe, L., 2012. The Criminalisation of Women. Institute of Criminology, University of Cambridge, Cambridge.
Haney, L., 2010. Offending Women. University of California Press, Berkeley.
Hannah-Moffat, K., Shaw, M., 2003. The meaning of 'risk' in women's prisons: a critique. In: Bloom, B. (Ed.), Gendered Justice: Addressing Female Offenders. Carolina Academic Press, Durham, NC, pp. 45–68.
Harpham, D., 2010. Offender Volumes Report 2009. Department of Corrections, Wellington, New Zealand.
Hedderman, C., 2011. Policy developments in England and Wales. In: Sheehan, R., McIvor, G., Trotter, C. (Eds.), Working with Women Offenders in the Community. Willan, Cullompton, pp. 26–44.
Hedderman, C., Gelsthorpe, L. (Eds.), 1997. Understanding the Sentencing of Women. Home Office Research Study 170. Home Office, London.
International Centre for Prison Studies, 2013. http://www.prisonstudies.org/info/worldbrief/wpb_stats.php (accessed 01.09. 13.).
Kruttschnitt, C., Gartner, R., 2003. Women's imprisonment. In: Tonry, M. (Ed.), Crime and a Review of Research, vol. 30. University of Chicago Press, Chicago, IL, pp. 1–80.
Kruttschnitt, C., Savolainen, J., 2009. Ages of chivalry, places of paternalism. European Journal of Criminology 6 (3), 225–247.
Lawston, J., 2013. Prisons, gender responsive strategies and community sanctions: the expansion of punishment in the United States. In: Malloch, M., McIvor, G. (Eds.), Women, Punishment and Social Justice. Human Rights and Penal Practices. Routledge, Abingdon, Oxon, pp. 109–120.
Malloch, M., McIvor, G. (Eds.), 2013. Women, Punishment and Social Justice. Routledge, Abingdon, Oxon.
Martin, J., Kautt, P., Gelsthorpe, L., 2009. What works for women?: a comparison of community-based general offending programme completion. British Journal of Criminology 49, 879–899.
McIvor, G., Burman, M., 2011. Understanding the Drivers of Female Imprisonment in Scotland. Report No. 02/2011. Scottish Centre for Crime and Justice Research, Glasgow.
Messina, N., Grella, C., Burdon, W., Prendergast, M., 2009. Childhood adverse events and current traumatic distress: a comparison of men and women durg-dependent prisoners. Criminal Justice and Behaviour 34 (11), 1385–1401.
Ministry of Justice, 2012. Statistics on Women and the Criminal Justice System 2011: A Ministry of Justice Publication under Section 95 of the Criminal Justice Act 1991. Ministry of Justice, London.
Moore, L., Jemphrey, A., 2013. Women's imprisonment in Northern Ireland: social exclusion, punishment and breaches of rights. In: Malloch, M., McIvor, G. (Eds.), Women, Punishment and Social Justice. Human Rights and Penal Practices. Routledge, Abingdon, Oxon, pp. 27–38.
National Institute of Corrections, 2013. Women Risk Needs Assessment. http://www.uc.edu/womenoffenders.html (accessed 01.09.13.).
Player, E., 2005. The reduction of women's imprisonment in England and Wales: will the reform of short prison services help? Punishment and Society 7 (2), 419–439.
Plugge, E., Douglas, N., Fitzpatrick, R., 2006. The Health of Women in Prison Study Findings. Department for Public Health, University of Oxford, Oxford.
Prison Reform Trust, 2012. Bromley Briefings Factfile. Prison Reform Trust, London.
Quaker Council for European Affairs, 2007. Women in Prison: A Review of the Conditions in Member States of the Council of Europe. Quaker Council for European Affairs, Brussels.
Ross, R., Fabiano, E., 1986. Female Offenders: Correctional Afterthoughts. McFarland, Jefferson, NC.
Schwartz, J., Steffensmeier, D., 2007. The nature of female offending: patterns and explanation. In: Zaplin, R.T. (Ed.), Female Offenders: Critical Perspectives and Effective Interventions, second ed. Jones and Bartlett, Boston, MA, pp. 43–76.
Sentencing Council for England and Wales, 2012. Courts Issues with New Guideline for Sentencing Drug Offenders. http://sentencingcouncil.judiciary.gov.uk/media/674.htm (accessed 01.09.13.).
Sharpe, G., 2012. Offending Girls: Young Women and Youth Justice. Routledge, Abingdon, Oxon.
Sheehan, R., 2013. Justice and community for women in transition in Victoria, Australia. In: Malloch, M., McIvor, G. (Eds.), Women, Punishment and Social Justice. Human Rights and Penal Practices. Routledge, Abingdon, Oxon, pp. 342–360.
Sheehan, R., McIvor, G., Trotter, C. (Eds.), 2007. What Works with Women Offenders. Willan Publishing, Cullompton.
Sheehan, R., McIvor, G., Trotter, C., 2011. Working with Women Offenders in the Community. Willan Publishing, Cullompton.
Troshynski, E., 2012. Human trafficking. In: DeKeseredy, W., Dragiewicz, M. (Eds.), Routledge Handbook of Critical Criminology. Routledge, Abingdon, Oxon, pp. 342–360.
United Nations, 2010. United Nations Rules for the Treatment of Women Prisoners and Non-custodial Measures for Women Offenders (The Bangkok Rules). GA/SHC/3980. General Assembly, United Nations.
Wacquant, L., 2009a. Prisons of Poverty. University of Minnesota Press, Minneapolis.
Wacquant, L., 2009b. Punishing the Poor: The Neoliberal Government of Social Insecurity. Duke University Press, Durham.
Walmsley, R., 2006. World Female Imprisonment List. International Centre for Prison Studies, King's College, London.
Yin, S., 2008. Are US Girls Becoming More Violent? Population Reference Bureau, Washington, DC.
Zahn, M.A., Brumbaugh, S., Steffensmeier, D., 2008. Violence by Teenage Girls: Trends and Context. US Department of Justice, Washington, DC.
Zaplin, R. (Ed.), 1998. Female Offenders: Critical Perspectives and Effective Interventions. Aspen Publishers, Gaithersburg, MD.

Women and Psychiatry

Carol Cooperman Nadelson, Harvard Medical School, Boston, MA, USA
Carol Ann Bernstein, New York University School of Medicine, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by C. Nadelson, volume 24, pp. 16522–16528, © 2001, Elsevier Ltd.

Abstract

In this article, we review the history of women in health care from several thousand years BC through the present day. Although women practiced medicine as early as 4000 BC, European medical education did not begin until the ninth century. We analyze the trajectory of women's roles in medicine, throughout the middle ages until the present time, culminating in a focus on women in eighteenth- and nineteenth-century American and European Medicine. We conclude by focusing specifically on the roles of women in psychiatry and the challenges which continue to be experienced by women in leadership in Western medicine in the twenty-first century.

In this article we will consider women in medicine and psychiatry, including the history of women's roles and contributions. We will provide data on the status of contemporary women physicians and discuss the challenges they face.

The History of Women in Health Care

Throughout history, women have been healers and caretakers, playing multiple roles as pharmacists, nurses, midwives, abortionists, counselors, physicians, and 'wise women,' as well as witches. As early as 4000 BC, there were women who studied, taught, and practiced medicine. The first known woman physician, Merit Ptah, practiced in Egypt in 2500 BC. According to Chinese records, in 1000 BC, women engaged in typical physician-like activities.

The Roman, Soranas, believed that women were divinely appointed to care for women and children. His criteria for practitioners included literacy, anatomic understanding, a sense of patient responsibility, and ethical concerns, particularly regarding confidentiality. Roman women physicians, however, were rarely permitted to perform in the same capacities and positions as men.

Women founded and staffed early hospitals in 260–220 BC. The deaconesses of the early Christian church developed the first public health systems (Shryock, 1959). By the eighth century, Arab medicine began to flourish. Since Muslim men were forbidden from having contact with women who were not family members, women cared for women. In Islamic Spain, women received medical education and were surgeons before the twelfth century.

European medical education formally began in Salerno, Italy, in the ninth century. By the tenth and eleventh centuries, women were thought to have studied there (Corner, 1937). When medical degrees were first awarded in CE 1180, however, they were awarded only to men. Women received special diplomas until the fifteenth century. A woman physician, Trotula, was an important eleventh-century academic and teacher at the University of Salerno. She wrote texts on obstetrics and gynecology and headed a department of women's diseases. Her contributions were significant because, at that time, women provided all obstetrical care and many had no training. Many of her accomplishments, and those of other women, however, were attributed to their husbands and sons. An early twelfth-century abbess, Hildegard of Bingen, was a healer, philosopher, composer, and scientific scholar who provided scientific data on nervous system regulation and an understanding of normal and abnormal psychology.

Women were ambivalently received when they expanded their traditional midwifery roles, and they generally were prevented from obtaining medical training or degrees. In England, the need for physicians during the Black Death of 1349 opened the way for women to practice medicine. Constanza Calenda, in 1423, became the first woman to formally obtain a medical degree. She lectured in medicine at the University of Naples. In the fourteenth and fifteenth centuries women practiced medicine in several European countries, but their practices were usually limited to bleeding, administration of herbs and medicines, and reducing fractures, in addition to midwifery. In some countries, women did practice as physicians, surgeons, orthopedists, and ophthalmologists, and it was generally believed that women were better able to treat women's diseases. However, there were also women physicians who served kings, royal families, and even the military.

Despite this apparent success, by the end of the fifteenth century the movement to exclude women from practicing medicine gained momentum and most were specifically limited to functioning as midwives. This coincided with the misogyny of Johann Sprenger and Heinrich Kraemer, in the *Malleus Maleficarum* of 1486, which laid out much of the philosophy of the inquisition. Witch-hunting focused on the 'spiritual and mental' inferiority of women. Even when active witch hunts subsided, their effects remained and women were eliminated from many medical roles. By the seventeenth century, when examinations and licensing procedures were required for physicians, women were rarely allowed to sit for required examinations (Nadelson and Notman, 1995).

Women in Eighteenth- and Nineteenth-Century American and European Medicine

In colonial America, the healing role of women was critical to the survival of the colonies, since medical care of families was the

responsibility of women. There were no medical schools and relatively few trained physicians in the colonies. Training was obtained by going abroad or by apprenticing to a practitioner.

In 1765, Dr John Morgan founded the first 'regular' American medical school at the University of Pennsylvania. By excluding women, the tradition of barring women from formal medical training in the United States began. Still, many women set up 'irregular' practices without diplomas, having apprenticed in homeopathic or eclectic traditions. Toward the latter part of the nineteenth century, US women began to study in the few US and European schools that admitted them, including some in Switzerland, Italy, Germany, Austria, England, Ireland, and Scotland.

In 1847, Elizabeth Blackwell became the first woman accepted to a 'regular' US medical school. When she received her MD degree from the Geneva Medical School, the New York State Medical Association censured the school. In 1850, Harriet K. Hunt, an 'irregular' practitioner in Boston, applied for admission to Harvard Medical School. She was admitted, but denied her seat when the all-male class threatened to leave if women (and minorities) were admitted. It was not until 1946 that Harvard Medical School finally admitted women.

By the mid-nineteenth century, three all-female medical colleges were founded in the United States: the Women's Medical College of Pennsylvania, one in Boston (now Boston University), and another in Cincinnati. Several US universities, including Syracuse, Michigan, Iowa, and California also accepted women (Luchetti, 1998).

During the Civil War, Dr Mary Edwards Walker was awarded the Medal of Honor, the most prestigious military award, for her performance as a military surgeon. She was, however, subjected to enormous criticism and the medal was rescinded in 1917 after a change in guidelines limited it to men in combat. It was reinstated in 1977 (Luchetti, 1998). In 1878 the National Eclectic Medical Association became the first medical society in the United States to accept women for membership, but full membership in the American Medical Association continued to be closed to women until 1915.

Negative attitudes about women physicians were supported by a 'new science' rationalizing the inferior status of women on biological grounds. This escalated early in the nineteenth century and persisted into the twentieth century. A book published in 1873 by a prominent Harvard professor, Dr Edward Clarke, concluded that higher education for women produced 'monstrous brains and puny bodies' as well as a host of other abnormalities in them and their offspring. His views were successfully refuted by the eminent feminist, physician, and scientist, Dr Mary Putnam Jacobi. The road continued to be difficult even for those women who managed to obtain training. Ann Preston, founder of the Women's Medical College of Pennsylvania, described the entry of women onto the wards of the Pennsylvania Hospital, "We entered in a body, amidst jeerings, groaning, whistlings, and stamping of feet by the men students … On leaving the hospital, we were actually stoned by those so-called gentlemen" (Alsop, 1950).

Denied hospital appointments, US women began to open their own hospitals and clinics. In 1857, Elizabeth Blackwell, her sister Emily, and Marie Zakrzewska established the New York Infirmary for Women and Children. In 1862, Zakrzewska founded a similar hospital in Boston. They also founded the Women's Medical College of the New York Infirmary in 1868. Believing that women received an inferior education, they sought to correct the problem by establishing rigorous entrance exams and longer clinical training than other medical schools.

By the end of the nineteenth century, there were many important successes for women. Dr Mary Putman Jacobi was the first woman elected to membership in the New York Academy of Medicine in 1880. In 1875, Emmeline Cleveland, Dean of the Pennsylvania Women's Medical College, became the first nonmilitary woman physician in the United States to perform major surgery.

The role of women in medicine, however, continued to be hotly debated. In 1881, Rachel Bodley, Dean of the Women's Medical College of Pennsylvania, surveyed the 244 living women graduates of the school and found that, despite persistent belief that training for women was wasteful, the overwhelming majority were in active practice. Those who had married reported that their profession had no adverse effect on their marriages, nor had marriage interfered with their work. Many early women physicians came from medical families and they joined family members in practice. By 1876, 0.8% of US physicians were women. By 1890, 35 US medical schools admitted women, and in 1900, 6% of physicians were women (Luchetti, 1998).

In Europe, 1866 saw Elizabeth Garrett Anderson open a dispensary for women and children in London during a cholera epidemic. People were grateful for medical attention, even from women. In 1874 Sophia Jex-Blake, Elizabeth Garrett Anderson, and others opened the London School of Medicine for Women. The medical hierarchy threatened to blackball any male who taught there, and no teaching hospital would allow women to practice. For several years women could not get the needed training for registration in medicine. In 1877, the Royal Free Hospital in London allowed women access to its facilities. The Royal College finally opened its exams to women in 1910.

In Canada the situation was no different. Dr James Barry (1797–1865), Inspector General of the British Army Medical Department and subsequently Inspector General of Hospitals for Canada, was discovered to be a woman after her death. She was the first woman doctor to practice medicine in Canada (Hacker, 1974).

Women in Twentieth-Century Medicine

Despite the prominence and contributions of many women, negative attitudes about women in medicine persisted. In 1905, the President of the Oregon State Medical Society stated:

> hard study killed sexual desire in women, took away their beauty, brought on hysteria, neurasthenia, dyspepsia, astigmatism and dysmenorrhea … Educated women could not bear children with ease because study arrested the development of the pelvis at the same time it increased the size of the child's brain and therefore its head. This causes extensive suffering in childbirth.
>
> Bullough and Voght, 1973

In 1900, women comprised 10% of the students at 18 medical schools, 50% at Boston University, and 40% at Tufts University. Johns Hopkins opened its medical school in 1893

with the financial help of women, who exacted the condition that women be admitted on the same terms as men, based on qualifications. By 1917, Dr Florence Sabin became the first woman full professor at that institution (Luchetti, 1998).

Abraham Flexner, in his Carnegie Commission Report on the status of US medical education (1910), stated: "Medical education is now, in the United States and Canada, open to women upon practically the same terms as men. If all institutions do not receive women, so many do, that no woman desiring an education in medicine is under any disability in finding a school to which she may gain admittance." Further, he stated, "Now that women are freely admitted to the medical profession, it is clear that they show a decreasing inclination to enter it."

Flexner's Report established medicine as an academic discipline in the United States, with high standards for training and practice. This resulted in the closing of many medical schools, often those that admitted women. Thus, the number of female physicians in the United States leveled off at between 4% and 7% until the 1970s. This was a lower percentage than in most other Western countries.

Other countries continued to report higher percentages of female physicians. In 1965, for example, 7% of US physicians were women, while in the Soviet Union 65%, in Poland 30%, in the Philippines 25%, in the German Federal Republic 20%, in Italy 19%, and in the United Kingdom and Denmark 16% of physicians were women (Lopate, 1968).

Medicine had generally been viewed as a male profession in the United States more than in most other countries. Some hypothesize that this occurred because medicine had higher prestige and income than many other professions. It is also likely that science requirements of US universities were not considered appropriate for women. Since the practice of medicine represented a melding of roles and responsibilities, the caretaking functions of physicians fit the 'traditional' views of women's roles, but the technological and instrumental aspects did not. Clearly different cultures have conceptualized these issues differently.

The choices of specialty have remained relatively constant since women were admitted to medical schools in the United States. Women have characteristically entered pediatrics, internal medicine, psychiatry, family practice, and, more recently, obstetrics and gynecology, anesthesiology, dermatology, radiology, pathology, and other medical subspecialties. The surgical fields have remained sparsely populated by women and, although the numbers have changed over the last few decades, the percentages have moved very slowly.

Regardless of specialty, academic and administrative appointments as well as other leadership positions have been almost exclusively held by men. In the 1970s, the fact that women tended to primarily assume clinical roles was used as an argument for increasing their numbers in medical school. It was believed that they could be counted on to meet the direct health care needs of expanding populations, and men could assume specialty roles.

The large numbers of women in medicine in many countries have tended to be relegated to low prestige and clinical positions. Women, as in the United States, were rarely elected to membership in national academies of medicine or science. In the former USSR, women were in positions that reduced the differentiation between doctors and paramedical workers. They were the *feldschers* (semiprofessional health worker or physician's assistant) and this was considered to be a female role. In China, the 1950s brought the 'barefoot doctors' who were local health workers and many were former midwives (Nadelson and Notman, 1995).

The large increase in numbers of women physicians in the early 1970s in the United States coincided with the civil rights, antiwar, and women's movements. At the beginning of the twenty-first century almost 50% of graduating medical students and over 20% of all physicians were women. By 2011–12, women represented 47.8% of medical school graduates and almost 30% of physicians (AAMC, 2012).

There continues to be significant disparity in the leadership roles of women in medicine. Although 37% of full-time medical school faculty are women, up from 29% in 2001, women remain underrepresented in the professorial ranks. Despite the fact that the percentage of women at each rank has increased about 7% on average over the past 10 years, women continue to be overrepresented in the lower ranks, e.g., 54% of instructors are women while only 20% of full professors are women. In 2012, only 14 of the 119 permanent medical school deans (12%) were women, up from 2004 when 10% were women (AAMC, 2012).

Although women make up increasing proportions of residents in each specialty, the distribution of women among various specialties has shown little fluctuation. The highest percentages of women in 2010–11 were in obstetrics and gynecology (82% of residents) and the lowest are in the surgical subspecialties (15.4% in neurosurgery and 13.6% in orthopedics). The percentage of women residents in psychiatry in 2011 is slightly above the average at 55.5% up from 50.7% in 2001 (AAMC, 2012).

Women in Psychiatry

Mental illnesses have always been viewed differently than other medical illnesses, because they were generally not seen as medical. Early historical references to the care of the mentally ill, dating from classical times, framed mental illness in terms of religion. Many civilizations had sanctuaries where religious rites were used to treat people. As pagan faiths lost favor, there was deterioration in the quality of treatment afforded to the mentally ill. Mental illness was increasingly viewed as caused by supernatural powers, evil spirits, or demons. It took many centuries before institutional care was comparable to what had existed in ancient times (Zilboorg, 1941).

The first 'morotrophium' or house for 'lunatics' was recorded in the Byzantine world in the fourth century CE. Pilgrimages were made by those suffering to shrines believed to eject evil spirits. One of the early morotrophiums (house for lunatics) was named for St. Dymphna. It was a home for mental healing run by women in CE 1200.

Over the next several centuries, asylums were established in many parts of Europe and the Middle East, often with the support and leadership of women. In the sixteenth century, the Sisters of Charity took in the mentally ill when hospitals refused to admit them (Tomes, 1994). By the seventeenth century, some French institutions began to train nurses to care

for these patients. In general, however, care was punitive and often inhumane. A seventeenth-century Swiss woman physician, Barbara von Roll, pioneered in the treatment of mental illness and opened the field of 'psychosomatic medicine' by documenting the connection between physical and mental health (Brooke, 1997).

By the eighteenth century mental illness was receiving greater attention in Europe and the United States. Although the focus was primarily on nosology and categorization, some therapeutic approaches included bloodletting, purgatives, dousing in ice water, and placement of the patient in a spinning chair until the patient lost consciousness. Benjamin Rush, the founder of US psychiatry, was an advocate of the latter method since he believed that congested blood in the brain caused mental illness, and the movement of the spinning chair would relieve it (Alexander and Selesnick, 1966).

During the nineteenth century, wives of US physicians entertained asylum patients in their homes, to 'calm a fevered mind' (Tomes, 1994). These women were often responsible for the 'moral' atmosphere of the institutions and advocated for respect and attention in order to promote recovery. Dorothea Dix was a pioneer in her efforts to care for the mentally ill and reform inhumane practices.

In the late nineteenth century, about 200 women physicians staffed US asylums (Lunbeck, 1994). In 1880, Alice Bennett became the first woman chief of a women's division at Norristown State Hospital in New Jersey. Bennett's unpopular and outspoken opposition to mechanical restraint and neglect of the chronically insane eventually caused her to lose her position (Barton, 1987).

By the twentieth century, increasing numbers of women worked with the mentally ill. The development of psychoanalysis in Europe attracted many women who were important in the development of the field. In 1923, the American Psychoanalytic Association admitted two women, Lucille Dooley and Marion Kenworthy to the organization. During this time Karen Horney began to challenge psychoanalytic theory, and a school of psychoanalysis evolved around her work (Menninger and Nemiah, 2000).

Over the course of the next few decades, many notable women psychoanalysts emigrated from Europe to the United States. Prior to World War II, 30% of European and 20% of US psychoanalysts were women (Chodorow, 1986). They contributed substantially to the growth and influence of psychoanalysis in the United States. Among the early women leaders were Clara Thompson, Hilda Bruch, Helene Deutsch, Phyllis Greenacre, Judith Kestenberg, Marianne Kris, Margaret Mahler, Edith Jacobson, Stella Chess, Frieda Fromm-Reichmann, and many others who pioneered as researchers and clinicians. Grete Bibring, a noted psychoanalyst, became the first woman to head a department of psychiatry in the United States (Beth Israel Hospital in Boston) and to become a full professor at Harvard Medical School.

Psychosomatics evolved in the 1930s with significant contributions of women, such as Therese Benedek and Helen Flanders Dunbar. Jimmie Holland was instrumental in founding the field of psychiatric oncology. Marion Kenworthy was notable in exerting an influence on the military, successfully facilitating an increase in the number of mental health units on army bases (Deutsch, 1959). In addition, Dr Kenworthy was a founding member and president of the Group for the Advancement of Psychiatry and became the first woman president of the American Psychoanalytic Association in 1958. She was elected as the first woman vice president of the American Psychiatric Association (APA) in 1965. Viola Bernard was a leader and founder of many psychiatric organizations and made pioneering contributions to child and social psychiatry (Menninger and Nemiah, 2000).

After World War II, women leaders emerged in organizations and academia, including, Eveoleen Rexford, Helen Beiser, Jeanne Spurlock, Mildred Mitchell-Bateman, and June Jackson-Christmas. It is notable that the latter three were African-American. Much of the leadership in the development of child psychiatry in the United States came from women, including many of those mentioned above. Melanie Klein developed internationally accepted theory concerning early life psychodynamic processes and Anna Freud contributed significantly to the understanding of ego development. In 1967 Evelyn Ivey became the first woman president of the American College of Psychiatrists.

As in other fields the numbers of women increased to almost 20% of psychiatrists in the 1970s but they continued to hold few leadership positions. Concern about women in leadership positions escalated enough to stimulate the APA to establish a Task Force on women, headed by Nancy Roeske. This led to the establishment of the APA Committee on Women. In 1973 Rebecca Solomon became the first woman appointed to the Editorial Board of the American Journal of Psychiatry. In 1974 the APA appointed Jeanne Spurlock to be the first Deputy Medical Director of Office of Child Psychiatry, Minority Affairs, and Women in Psychiatry. This was followed in 1976 by the appointment of Carolyn Robinowitz as Deputy Medical Director of the Office of Education.

In the 1980s and 1990s, women's leadership roles increased. In 1981, Judith Gold was elected as the first woman president of the Canadian Psychiatric Association and in 1984 Carol Nadelson was elected the first woman APA president (Sabshin, 2008). Since then, of the subsequent 24 APA Presidents, 7 have been women.

In 1984 Judith Rappaport became the head of Child Psychiatry at the National Institutes of Mental Health of the United States. Neuroscientist Nancy Andreasen became the first woman Editor in Chief of the American Journal of Psychiatry in 1992.

Other women leaders in psychiatry emerged during this time period in other parts of the world. In Australia, Beverly Raphael, the first woman to chair a psychiatry department in the country, became the first woman president of the Royal Australia New Zealand College of Psychiatry. Sarah Romans became the first woman Psychiatry Chair in New Zealand. Felice Lieh Mak, Chair of Psychiatry in Hong Kong, was the first woman elected President of the World Psychiatric Association, and Astrid Heiberg became a Professor of Psychiatry and the Undersecretary of Health in Norway, as well as President of the Norwegian Red Cross and later the International Red Cross.

Women psychiatrists who pioneered in research included Barbara Fish, who in 1955 began a long-term follow-up study of schizophrenic children. She was followed by Lissy Jarvik who was in the forefront of work on Alzheimer's disease, Nancy

Andreasen who pioneered the use of neuroimaging in psychiatric illness, Gabrielle Carlson and Paula Clayton whose work on affective disorders paved the way for future understanding of these illnesses, and Judy Rapoport who initiated research on early onset schizophrenia, hyperactivity, and endocrine function (Menninger and Nemiah, 2000). Nora Volkow MD, a leading researcher in substance abuse, was appointed Director of the National Institute on Drug Abuse in 2003.

Female Leadership in Medicine and Psychiatry: the Twenty-First Century

Although 55% of the trainees in psychiatry in the United States in 2013 are women, they remain in junior faculty positions in academic psychiatry as they do in the rest of academic medicine. They constitute 13% of US psychiatry chairs (AAMC, 2012). Women often lack research training and the kind of mentorship that will enable them to advance their careers. They are, less often than their male colleagues, offered research opportunities during training or at entry level academic positions, and they tend to work in clinical, teaching, and administrative positions that are time-consuming and do not facilitate the development of the necessary skills to enable them to progress academically. The pressures are greater for women interested in academic careers since they are in their peak childbearing years when they must also commit themselves to their career development (Nadelson and Notman, 1995).

Women in academic psychiatry, as in other fields of medicine in the United States, continue to be less likely to have peer-reviewed grants or publications than men, independent of parental status. A study of career development award recipients (K08 and K23 awards) emphasized that the career trajectories of men and women in academia differ. Both men and women award recipients tended to remain in academic medicine but women were less likely to receive R01, independent investigator grants and they published fewer papers (Jagsi et al., 2011). A 2011 paper supporting the differences in career trajectories indicated that women's publication numbers increase and exceed men's as they advance in their careers, but they continue to hold fewer leadership positions (Reed et al., 2011). These authors indicate that the disparity data suggest that mid-career assessments of academic productivity should not be a measure of future potential. Regarding academic rank these authors also found, as have other reports, that throughout their careers women were slower at achieving associate and full professor rank than men and they were less likely to attain divisional, departmental, institutional, or national leadership positions during the course of their careers. The usual reasons given for the lack of female leadership are that the numbers of women available are not sufficient, and/or that women are not motivated because of family responsibilities. The evidence has suggested, however, that men and women in similar cohorts have different career trajectories, and that women have not been promoted at the same rate as men, despite comparable qualifications (Tesch et al., 1995). Moreover, a 2011 analysis of the gender of editorial board members of 60 top-ranked medical journals of 12 Thomson Reuters Web of Knowledge Journal Citation Reports categories revealed that only 15.9% of editors in chief and less than one-fifth of all editorial board members were women (Amrein et al., 2011). Clearly, the numbers of qualified women have increased substantially but they are not being appointed. From a study of 4500 women physicians reviewing experiences of gender-based harassment, especially in historically male specialties, More et al. note that these early themes still resonate (More et al., 2009).

Another issue that has generated considerable discussion has been salary differential between men and women in academia as well as other fields. A study of gender differences in mid-career physician researcher salaries found that these differences do exist, when controlling for specialty, institutional characteristics, academic productivity, academic rank, and work hours (Jagsi et al., 2012). While there are many possible explanations for this disparity, gender discrimination or inequality remains an important consideration.

Another area of interest regarding the careers of men and women in academic medicine has been attrition. Levine et al. reported in 2011 that women leave academic medicine at higher rates than men. They suggest several determinants of this effect, including: lack of role models for family balance and career development, frustration regarding research funding and mentorship, work life balance, and institutional environments which discourage collaborative work and collegiality.

Despite contrary data, women's commitment, qualifications, and capabilities continue to be questioned and doubted more than men's. Discrimination, subtle harassment, and demeaning of women continue to exist. A study investigating gender bias demonstrated that when male and female names were placed on identical applications for the same positions, the male-named applicants were viewed as more competent, hired, and offered higher starting salaries than the female named applicants (Moss-Racusin et al., 2012). This type of study has been repeated over and over for many years with the same results, implying subtle and perhaps unconscious bias (Isaac et al., 2009). These authors also postulate that bias was eliminated when there was a commitment to the value of credentials before the review of applicants and when 25% or more of the applicants were women.

A summary of the issues faced by women as potential leaders begins with the difference in resources and opportunities offered to them as they begin their careers. Many studies repeat that women have greater problems with work/life balance, lack of family friendly policies and career flexibility, gender bias, and reduced support from senior leaders and mentors (Westring et al., 2012). Although we often assume that these issues are primarily experienced in the United States, studies from other countries echo the data and conclusions we have described (Goldacre et al., 2012; Riska, 2011; Van den Brink, 2011).

Psychiatry has a rich history of female contributions and leadership. If the field is to continue to evolve and expand its knowledge base, the culture must change. Women must play a vital role in the evolution of the field and innovations in research. Given that more than 50% of US trainees are women, their talents and potential contributions are essential to future progress.

See also: Alcohol-Related Disorders; Child and Adolescent Psychiatry, Principles of; Depression; Education in Psychiatry; Family and Psychiatric Illness: Family Therapy; Hysteria: History and Critiques; Perinatal Psychiatry; Somatoform Disorders; Therapeutic Community.

Bibliography

Alexander, F.G., Selesnick, S.T., 1966. The History of Psychiatry. Harper and Row, New York.

Alsop, G.F., 1950. History of the Women's Medical College of Pennsylvania. Lippincott, Philadelphia, PA.

Amering, M., Schrank, B., Sibitz, I., 2011. The gender gap in high-impact psychiatry journals. Academic Medicine 86 (8), 946–952.

Amrein, K., Langmann, A., Fahrleitner-Pammer, A., et al., 2011. Women underrepresented on editorial boards of 60 major medical journals. Gender Medicine 8 (6), 378–387.

AAMC, Joliff, L., Leadley, J., Coakley, E., et al., 2012. Women in U.S. Academic Medicine and Science: Statistics and Benchmarking Report 2011–2012. Association of American Medical Colleges.

Barton, W.E., 1987. The History and Influence of the American Psychiatric Association. APPI, Washington, DC.

Bonner, T.N., 1992. To the Ends of the Earth. Harvard University Press, Cambridge, MA.

Brooke, E., 1997. Medicine Women. Quest, Wheaton, IL.

Bullough, V., Voght, M., 1973. Bulletin of the History of Medicine 47, 66–82.

Chodorow, N., 1986. Varieties of leadership among early women psychoanalysts. In: Dickstein, L.J., Nadelson, C.C. (Eds.), Women Physicians in Leadership Roles. American Psychiatric Press, Washington, DC, pp. 45–54.

Corner, G.W., 1937. The rise of medicine at Salerno in the twelfth century. In: Lectures on the History of Medicine: A Series of Lectures at the Mayo Foundation and the Universities of Minnesota, Wisconsin, Iowa, Northwestern and the Des Moines Academy of Medicine, 1926–1932. Saunders, Philadelphia, PA, pp. 371–399.

Deutsch, A., 1959. The Story of GAP. Group for the Advancement of Psychiatry, New York.

Flexner, A., 1910. Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Technology. Carnegie Foundation, New York. Reprinted 1960 Washington Science and Health Publications, Arno Press, New York.

Goldacre, M., Goldacre, R., Lambert, T., 2012. Doctors who considered but did not pursue specific clinical specialties as careers: questionnaire surveys. Journal of the Royal Society of Medicine 166–176.

Hacker, C., 1974. The Indomitable Lady Doctors. Clarke, Irwin, Toronto, ON, Canada.

Isaac, C., Griffin, L., Carnes, M., 2010. A qualitative study of faculty members' views of women chairs. Journal of Women's Health 19 (3), 533–546.

Isaac, C., Lee, B., Carnes, M., 2009. Interventions that affect gender bias in hiring: a systematic review. Academic Medicine 84 (10), 1440–1446.

Issac, C., Chertoff, J., Lee, B., et al., 2011. Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. Academic Medicine 86 (1), 59–66.

Jagsi, R., DeCastro, R., Griffith, K.A., et al., 2011. Similarities and differences in the career trajectories of male and female career development award recipients. Academic Medicine 86 (11), 1415–1421.

Jagsi, R., Griffith, K., Stewart, A., et al., 2012. Gender differences in the salaries of physician researchers. Journal of the American Medical Association 307 (22), 2410–2417.

Levine, R.B., Lin, F., Kern, D.E., et al., 2011. Stories from early-career women physicians who have left academic medicine: a qualitative study at a single institution. Academic Medicine 86 (6), 752–758.

Lopate, C., 1968. Women in Medicine. Johns Hopkins University Press, Baltimore, MD.

Luchetti, C., 1998. Medicine Women. Crown, New York.

Lunbeck, Elizabeth, 1994. The Psychiatric Persuasion: Knowledge, Gender and Power in Modern America. Princeton University Press, Princeton, NJ.

Menninger, R., Nemiah, J., 2000. American Psychiatry after World War II. APPI, Washington, DC.

More, E., Fee, E., Parry, M., 2009. Women Physicians and the Cultures of Medicine. The Johns Hopkins University Press, Baltimore, MD.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., et al., 2012. Science faculty's subtle gender biases favor male students. PNAS Early Edition 1–6.

Myers Jr., S.L., Fealing, K.H., 2012. Changes in the representation of women and minorities in biomedical careers. Academic Medicine 87 (11), 1525–1529.

Nadelson, C.C., Notman, M.T., 1995. Women as health professionals. In: Reich, W.T. (Ed.), Encyclopedia of Bioethics, second ed. Macmillan, New York, pp. 1713–1720.

Reed, D.A., Enders, F., Lindor, R., et al., 2011. Gender differences in academic productivity and leadership appointments of physicians throughout academic careers. Academic Medicine 86 (1), 43–47.

Riska, E., 2011. Gender and medical careers. Maturitas 68, 264–267.

Sabshin, M., 2008. Changing American Psychiatry. APPI, Washington, DC.

Shryock, R.H., 1959. The History of Nursing: An Interpretation of the Social and Medical Factors Involved. Saunders, Philadelphia, PA.

Tesch, B., Wood, H.M., Helwig, A.L., Nattinger, A.B., 1995. Promotion of women physicians glass ceiling or sticky floor? Journal of the American Medical Association 273, 1022–1025.

Tomes, N., 1994. The Art of Asylum Keeping: The Thomas Story Kirkbride Origins of American Psychiatry. University of Pennsylvania Press, Philadelphia, PA.

Van den Brink, M., 2011. Scouting for talent: appointment practices of women professors in academic medicine. Social Science & Medicine 72, 2033–2040.

Villablanca, A.C., Beckett, L., Nettiksimmons, J., et al., 2011. Career flexibility and family-friendly policies: an NIH-funder study to enhance women's careers in biomedical sciences. Journal of Women's Health 20 (10), 1485–1496.

Westring, A.F., Speck, R.M., Sammel, M.D., et al., 2012. A culture conducive to women's academic success: development of a measure. Academic Medicine 87 (11), 1622–1631.

Zhuge, Y., Kaufman, J., Simeone, D., et al., 2011. Is there still a glass ceiling for women in academic surgery? Annals of Surgery 253 (4), 637–643.

Zilboorg, G., 1941. A History of Medical Psychology. Norton, New York.

Women in Buddhism

Lisa J Battaglia, Samford University, Birmingham, AL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by R.M. Gross, volume 2, pp. 1386–1390, © 2001, Elsevier Ltd.

Abstract

This article focuses on strategies and approaches to the study of women in Buddhism, women in early Buddhism (particularly as represented in the Pāli canon), women in Mahāyāna Buddhism, and recent developments in the Buddhist women's ordination movement, particularly as manifested in the Theravāda cultures of Southeast Asia. Attention is given to Buddhist attitudes toward and representations of women, the status of women as expressed through Buddhist textual discourses, and the concerns and perspectives of contemporary Buddhist renunciant women amidst the ordination controversy.

The subject 'Women in Buddhism' spans a period of over 2500 years and encompasses different Buddhist locales and traditions. From texts highlighting the Buddha's attitude toward women in fifth century BCE in India to current debates over women's higher ordination in Buddhist schools, women have been the subject of attention and concern throughout the history of Buddhism. They have also been actors and participants in the development of Buddhism. This article will focus on strategies and approaches to the study of women in Buddhism, women in early Buddhism (particularly as represented in the Pāli canon), women in Mahāyāna Buddhism, and, lastly, recent developments in the Buddhist women's ordination movement, particularly as manifest in Theravāda cultures of Southeast Asia.

Women in Buddhism: Strategies and Approaches

Historically, Buddhism has been a male-dominated enterprise in terms of the chiefly male authorship and perspectives of its classic texts and the predominantly male leadership and orientation of its institutions. British scholars in the late nineteenth and early twentieth centuries commenced the 'first wave' of scholarship on women in Buddhism. C.A.F. Rhys Davids (1857–1942), Bimala Churn Law (1862–1969), and Isaline Blew Horner (1896–1981) were textual scholars who studied the role of women in Buddhism and proposed that women played definitive roles in Buddhist history and Pāli literature embodied and produced a remarkable degree of religious egalitarianism. In accordance with early Western scholarship on Buddhism, they prioritized texts and consulted primary sources (e.g., the Pāli canon) to locate their subject of study. These pioneering studies championed Buddhism as emancipatory for women in contrast to the (perceived) oppressiveness of the Brahmanical culture in which it arose. Brahmanism's own developments and advances vis-à-vis women's status and representation during the period of Buddhism's inception did not receive scholarly attention. By and large, Buddhism was deemed liberating for women and inherently egalitarian because it allowed women to renounce the subordinate, socially conditioned roles of wife and mother and to assume the gender-free spiritual path of the monastic. Any discrimination or sexism within Buddhist textual and institutional structures was attributed to foreign imposition (i.e., seen as a residue of traditional Indic conceptions of women), propositions that are widely disputed in Buddhist Studies today.

The 'second wave' of scholarship grew out of the Women's Liberation Movement and feminism of the 1960s and 1970s. Equipped with the language, rhetoric, conceptual frameworks, and institutional advances spawned by the feminist movement, a new cohort of Buddhist scholars revisited the work begun by their predecessors (Murcott, 1991; Blackstone, 1998). Many of these scholars observed evidence of egalitarianism in Buddhism and upheld the essential teachings of Buddhism as incompatible with gender hierarchy and discrimination against women (Minamato, 1991; Carmody, 1991; Gross, 1993). Sexist or discriminatory elements in the Buddhist record were seen as cultural accretions or later accumulations on an originally egalitarian core. Similarly, some scholars differentiated between two levels of Buddhist teachings – the core teachings of Buddhism which are free from gender bias (and thus the springboard for feminist collaboration) and the mundane level of teachings which are affected by social context and thus amenable to varying manifestations of gender discrimination (Kabilsingh, 1991). Other scholars have located a 'new' egalitarian Buddhism, replete with women leaders and teachers, in the convert communities and Dharma centers established in America since the 1960s (Friedman, 1987; Boucher, 1988; Dresser, 1996). Recreating Buddhist institutions free from hierarchy, gender inequity, and sexism has also been the concern of Asian Buddhist immigrants in Western contexts (Gregory and Mrozik, 2008: p. 12). 'Second-wave' scholarship tended to hierarchize Buddhist schools and periods in terms of their positive or negative attitudes toward women. Later developments of Buddhism, namely the Mahāyāna traditions, were seen as increasingly egalitarian and emancipatory for women in contrast to the (perceived) conservatism and sexism of early Indian (equated with Theravāda) Buddhism. Many studies focused on images of women in Mahāyāna texts, revered female figures in Tibetan Buddhism, and feminine symbolism in Buddhist Tantra (Paul, 1979; Klein, 1995; Shaw, 1994). Critics note that the characterization of Mahāyāna as 'pro-women' overlooks vicious tirades against women in Mahāyāna texts (Wilson, 1996: pp. 7, 8), dismisses positive characterizations of women in Theravāda Buddhism (Lang, 1986: p. 64), and displaces the Brahmanical-Buddhist continuum of earlier

scholarship onto a Theravāda–Mahāyāna continuum – in which Mahāyāna is championed primarily for not being Theravādin (Walters, 1994: pp. 361, 362). Resisting the tendency to view Buddhist schools and periods as uniformly or univocally 'pro-women' or 'anti-women,' prominent scholars recognize a diversity of voices in Buddhist texts, leading to varied outcomes, such as inclusive salvation (androgynous or egalitarian), institutional androcentrism, and misogynistic asceticism (Sponberg, 1992).

The 'third wave' of scholarship on women in Buddhism is heavily informed by poststructuralist and postcolonial theory, gender and queer theory, men and masculinities studies, and third wave feminism. Studies in this most recent wave seek to take women's subjectivity seriously and to construct interpretive models of representation that at once privilege indigenous perspectives and engage in critical analysis. Some studies employ ethnographic research and focus on women's experiences in very specific Buddhist contexts and localities (Arai, 1999; Brown, 2001; Gutschow, 2004; Falk, 2007). Other studies have shifted from 'women' to 'gender' as the subject of investigation (recognizing that both women and men are impacted by gender roles and 'perform' gender), including the examination of men and masculinities in Buddhism (Faure, 2003; Powers, 2009). Still other studies are beginning to explore the complexity of gender identities, particularly those identities that fall outside of the male/female polarity (Gyatso, 2003). Works by Cabezón (1992, 1998), Zwilling (1992), Schalow (1992), and others have also established queer theory as a dynamic field of inquiry for scholars and practitioners of Buddhism.

The varying strategies and approaches of scholars attest to the diversity of women's images, roles, experiences, and concerns in different Buddhist contexts. These varied methodologies and assessments also demonstrate that "the history of Buddhism is made to reflect the concerns of the scholars reconstructing it" (Wilson, 1996: p. 7).

Women in Early Buddhism

In Pāli scriptures, the path espoused by the Buddha as most conducive to the attainment of *nirvāṇa* is the path of renunciation, referred to as 'going forth' from home to homelessness. 'Going forth' involves renunciation on three levels: first, physically renouncing the home, that is, leaving the householder life to pursue the monastic life; second, psychologically renouncing the home, that is, leaving behind desire and attachment; and third, ontologically renouncing the home, that is, leaving *saṃsāra*, the cycle of death and rebirth characterized by suffering (Collins, 1982: p. 169). A key concern of early Buddhists was whether women can, or should, renounce the home and 'go forth' into the monastic life. Buddhist texts voice apprehension about the appropriateness of women renouncing the world (particularly leaving the home and family and eschewing the roles of wife and mother) and doubt regarding the possibility of women cultivating detachment and dispassion. At the same time, women's spiritual capacity is acknowledged and women are regarded as ontologically equal to men in matters of the *Dharma* (the Buddha's teachings). In the well-known Pāli account of the founding of the community of Buddhist nuns (*bhikkhunīs*), the Buddha permits women into the monastic order, citing women's ability to realize the highest Buddhist goal of liberation. Early Buddhist literature affirms women's capability of attaining *nirvāṇa* (and other spiritual goals) and provides many examples of highly accomplished female devotees in meditation, morality, and wisdom. Most notably, the *Therīgāthā* (*Verses of the Elder Nuns*), the enlightenment poetry of the early Buddhist nuns, documents the spiritual achievements of women and their ability to assume spiritual authority from Buddhism's inception. The *Therīgāthā* is the only work in the Pāli canon attributed to the authorship of women and comprises the oldest known collection of women's literature. The position of the nun is further strengthened by the early Buddhist concept of the fourfold society. According to Buddhist scriptures, the Buddha sets forth the prescription for an ideal Buddhist community replete with four assemblies: monks, nuns, laymen, and laywomen. All four divisions are seen as essential to the well-being of the religion and are said to 'illuminate the Order' (*Aṅguttara Nikāya*).

This optimistic view of renunciant women is curbed by the Buddha's proclamation that the creation of an order of nuns will dramatically hasten the decline of his teachings in the world. Moreover, as a condition for their ordination, renunciant women are obligated to observe Eight Heavy Rules (*Garudhammas*), with the first rule requiring a senior nun to bow down before a monk ordained even one day. The Eight Heavy Rules are highly controversial on two fronts. First, they clearly attempt to relegate women to a subordinate role. Second, the historicity of the Eight *Garudhammas* is questionable. Evidence suggests that the Eight Heavy Rules were formed after the establishment of the nuns' order and, as later interpolations (whether in the Buddha's lifetime or in later centuries), were interjected into the founding story of the *bhikkhunī* order. Beyond the *Garudhammas*, nuns are required to observe more precepts than monks as set forth in the monastic code of discipline (*Vinaya*). Whether more precepts necessarily translates into sexist discrimination has been thrown into question by those who see these rules as operating on women's behalf and for women's protection, and as consistent with an affirmation of the spiritual equality of women and men (Chung, 1999: p. 45). Nonetheless, the male gender is consistently hierarchized above the female gender in early Buddhism.

Early texts confer women's ability to attain the status of an *arhat* (one who has attained enlightenment). However, they are more ambivalent about other spiritual attainments. According to the list of women's 'five hindrances' which appears repeatedly in Buddhist sources, a woman cannot occupy the position of 'Buddhahood,' among four others, in the cosmos. Thus, in the early tradition, Buddhas are never female. While the *Jātakas* testify to the myriad incarnations of the Buddha in previous lives, including animal forms, female corporeality is conspicuously absent. To a large degree, the decided maleness of the Buddha stems from the pan-Indian notion of the 32 Marks of a Great Man (*Mahāpuruṣa Lakṣaṇa*), including the mark of having a penis concealed in a sheath. It must be noted here that the indisputable affirmation of women's ability to attain *nirvāṇa* probably held much greater significance than

women's exclusion from Buddhahood since early Buddhist sources stress enlightenment as the only feasible goal.

Gender is arguably irrelevant in matters of spiritual pursuit. Monks (*bhikkhus*) and nuns (*bhikkhunīs*) follow the same path to the same liberation. Birth in a male or female body neither abets nor hinders spiritual accomplishment: "Whoever has such a vehicle, whether a woman or a man, shall indeed, by means of that vehicle, come to nirvāṇa" (*Saṃyutta Nikāya*). Indeed, the *Dharma* appears to be 'gender-free.' The classic Buddhist doctrine of *anātman* (Pāli, *anattā*) – frequently translated as 'no soul,' 'no self,' or 'egolessness' – is common to all forms of Buddhism and has strong implications for gender analysis. If there is no permanent self or soul attached to an individual's existence, then there cannot be any permanent, real, or lasting identity based on gender. Early Buddhist texts recognize the fallacy of clinging to gender identities and distinctions, as evidenced in the *Saññoga Sutta* which suggests that attachment to gender identity is a form of bondage. Likewise, in the *Soma Sutta*, the *bhikkhunī* Soma links gender discrimination to the work of Māra and gender neutrality to the realization of *anātman*. The Pāli canon suggests that the very concept of gender differentiation can serve as a hindrance to attaining enlightenment.

The Buddhist adulation of 'genderlessness' mirrors the early Buddhist myth of the origin of society found in the Pāli *Agañña-sutta* (*Knowing the Beginning*). This cosmogony depicts a Golden Age where identical, asexual, ethereal beings dwell blissfully in the heavens. Upon becoming greedy for food, they begin eating and gradually become more corporeal and grosser in form, ultimately differentiating into male- and female-sexed beings. The emergence of the sexes leads to further moral degeneration in the form of passion, lust, and jealousy. Androgyny and asexuality characterize the once perfect state of blissful luminosity and, in effect, Buddhist monastics seek to return to this genderless, passionless ideal. In terms of appearance and behavior, Buddhist monks and nuns exercise androgyny with shaven heads, baggy robes, and identical lifestyle. Likewise, one of the most important distinguishing characteristics of monks and nuns is that they are celibate. For male and female monastics alike, transgressing the vow of chastity is an offense that carries the penalty of expulsion. While there is a tendency toward androgyny and soteriological equality, gender is not obsolete on the Buddhist path. According to the Pāli canonical texts, proper male and female gender must be officially confirmed at the time of ordination and ambiguously gendered or inter-sexed persons are banned from entering the *Saṅgha* (monastic community).

A widespread Buddhist belief echoed in early texts is that being born a woman is the result of bad karma. Buddhist scriptures articulate the woes of female incarnation, including "leaving relatives behind to go to a husband, menses, pregnancy, giving birth and having to wait upon a man" (*Saṃyutta Nikāya*). It is debatable whether these judgments are evidence of misogyny or evidence of the adverse conditions under which women lived. Either way, while women are often depicted aspiring to be reborn as men (and succeed in achieving this 'higher' rebirth by performing meritorious acts), the men of Buddhist lore never express the desire to be reborn as women (though they sometimes incur a female rebirth as a result of bad deeds). This popular and long-standing attitude persists in some Buddhist cultures today. This ethos of female inferiority has adversely affected the order of nuns throughout history. Because nuns are regarded as lesser 'fields of merit,' lay donors prefer to give alms to monks to accrue more merit. Recent data, however, suggests an improved situation for Buddhist renunciant women, including increased institutional support and wider recognition as worthy religious persons by monks and laity (Seeger, 2009).

The relevance or irrelevance of gender is further complicated by the constraint of male celibacy and the threat women pose to this celibacy. Certain texts of the Pāli canon attribute to women an uncontrolled sexual appetite and inherent licentiousness. Women are further characterized as uncontrollable, envious, greedy, and weak in wisdom (*Aṅguttara Nikāya*). A particularly salient image is that of woman as temptress. Buddhist texts attribute inordinate sexual (and destructive) powers to women. Women are described as 'snares of Māra' (*Aṅguttara Nikāya*) and members of the *Saṅgha* are advised to 'keep wide awake' in the face of feminine wiles (*Dīgha Nikāya*). Women's beauty holds men in thrall and leads wayward monks astray; to avoid temptation monks are instructed to contemplate the transience and foulness of the female body. Just as the allure of the female form threatens men's spiritual progress, so attachment to one's own beauty is an obstacle to women on the spiritual path. 'Vain nuns' like Rūpa Nandā of the *Therīgāthā* are described as being intoxicated with their own beauty and having 'fierce affection' for their bodies. Enlightenment can only be achieved by contemplating the impurity and impermanence of the female body, often one's own. In canonical and commentarial texts, women's bodies and bodily processes are described as polluted and defiling much more frequently than are men's, and women are not depicted contemplating horrifically transformed, disfigured, dead, or sleeping male bodies (Lang, 1986; Wilson, 1995, 1996; Blackstone, 1998).

Analyses suggest that women tended to fare better as models of lay practice (versus monastic practice), particularly as supporters and nurturers of the Buddhist tradition (Willis, 1985). Some scholars posit that the exaltation of laywomen parallels the declining status of *bhikkhunīs* (Paul, 1985: p. 79). Others conjecture that laywomen are presented much more positively than nuns in Buddhist stories because laywomen conform to socially prescribed female roles (Allione, 1986: p. 8; Falk, 1980: p. 222; Neumaier-Dargyay, 1995: p. 164). Viśākhā, the most lauded woman in the Pāli texts, is commended for her immense generosity, unwavering piety, knowledge of the *Dharma*, unsurpassed beauty, and abundant fertility. While lay-benefactors like Viśākhā figure prominently and positively in early Buddhist literature, other laywomen are portrayed negatively: namely, wives trying to regain their monastic husbands (e.g., the story of Sudinna in the *Vinaya Piṭaka*), mothers mourning the loss of their sons to the renunciant life, and beautiful women presenting themselves as sexual obstacles to abstinent monks. On the one hand, the scriptures praise laywomen, as when the Buddha of the *Etadagga-vagga* commends 10 laywomen for their qualities – among them being learned, developing loving kindness, exhibiting undivided pleasantness, attending to the sick, and reaching the *jhanas* (meditative states). On the other hand, Buddhist texts warn monks against the

filial and social obligations that entrap the householder. Interestingly, nuns are oft depicted voicing welcome emancipation from domestic drudgery and filial duties.

Women in Mahāyāna Buddhism

Mahāyāna Buddhism, which began to develop about 500 years after the Buddha, expounded upon earlier Buddhist doctrines and heralded new ideas regarding women and gender ideology. The wide range of attitudes toward women (as described above) in early Buddhism continues to be found in later Mahāyāna traditions. It has been suggested that Mahāyāna revalorized the female gender and that the religious ideals of universal salvation, emptiness, and undiscriminating wisdom signaled greater (or renewed) openness to women and a more equitable sociopolitical order. However, there is no substantial evidence indicating that women fared better in Indian Mahāyāna Buddhism than in the mainstream tradition. This, of course, points to the broader issue of distinguishing between symbolic representation and flesh-and-blood reality, as well as between textual materials and the practical religion of people at large.

With the advent of Mahāyāna came a new religious goal: that of becoming a *Bodhisattva* (and eventually a Buddha). In contrast to the earlier ideal of the *arhat*, the *Bodhisattva* (literally, 'enlightenment-being') vows to remain in *saṃsāra* to work for the complete enlightenment of all sentient beings. As an immensely compassionate being, the *Bodhisattva* postpones his or her enlightenment but is destined in the future to become a Buddha. The *Bodhisattva* path is open to both men and women, and all Mahāyānists are encouraged to pursue it. Mahāyāna texts are noted for their gender-inclusive language and sūtras are frequently addressed to 'good sons' and 'good daughters' of the *Dharma*. However, while the 'good son' is often addressed independently, the 'good daughter' is usually addressed in association with the 'good son.' Nonetheless, the view of the good daughter is decidedly optimistic. According to the *Lotus Sūtra*, the good daughter is capable of awakening to the thought of enlightenment, receiving a prediction of Supreme, Perfect Enlightenment, and teaching the *Dharma*, among other feats.

While the *Bodhisattva* ideal appeals to the masses and points to the universal availability of enlightenment, the path of the *Bodhisattva* is understood to be strenuous and to require an enormous span of time over many rebirths. There is disagreement among Mahāyāna texts as to what level of *Bodhisattvahood* a woman can attain in female form. Some sūtras deny women's entrance in Buddha-lands, while others accept women as lower-stage *Bodhisattvas*; still others accept women as advanced *Bodhisattvas* and imminent Buddhas. The *Pure Land Sūtra* (*Sukhāvatīvyūha*), for example, dictates that one must be a male in order to be born in the land of Amitābha, implying that a woman must wait for another existence as a man in order to progress on the *Bodhisattva* path. In contrast, many Mahāyāna sūtras describe virtuous and capable women who may become *Bodhisattvas* and even Buddhas in current lives, which were begun as female. Employing a narrative theme known as the 'transformation of sex' (*parivṛttavyañjana*), these texts open up the possibility of a woman attaining *Bodhisattvahood* or Buddhahood in this lifetime by undergoing sexual change. Transformation of sex sūtras follow a typical narrative pattern. A debate over the spiritual capabilities of women ensues between a wise and eloquent woman (a *Bodhisattva*) and a highly respected male elder (often a monk-hero or *Bodhisattva*). Responding to the challenge posed by the man that a woman cannot (in her female form) attain Buddhahood, the female *Bodhisattva* instantaneously changes her sex and transforms into a man (prior to immanent Buddhahood). In some transformation of sex sūtras, such as the *Vimalakīrtinirdeśa-sūtra* (*Discourse on the Teaching of Vimalakīrti*), the female protagonist engages in a playful display of illusion and ultimately refuses to transform her sex on the basis that such distinctions are illusory and empty of any inherent reality. These texts have been interpreted in a variety of ways. It has been suggested that sexual transformation represents a symbolic transition from a state of imperfection and attachment to sensual existence (represented by the female body) to a state of mental perfection, desirelessness, and eventual enlightenment (represented by the male body) (Paul, 1979). Others contend that the significant point is not that the woman becomes a man, but rather that the woman had already become a *Bodhisattva* and had already made her resolution to attain Buddhahood while in her female form, thus equipping her with the power to change her sex (Gross, 1993). Still others analyze the Mahāyāna sex-change sūtras through the lens of poststructuralist feminist theory and view the transformation of sex episodes as theatrical testaments to the fluidity and performativity of gender (Owen, 1997). Notably, it is always women who transform (or refuse to transform) into men; men are never depicted transforming (or being challenged to transform) into women – indeed, it appears this would be unthinkable.

To complicate the matter further, some Mahāyāna texts state that womanhood and Buddhahood are mutually exclusive. In other words, a Buddha will never appear in female form and a woman cannot be a Buddha. In the *Lotus Sūtra*, one of the most popular and influential Mahayana texts, Śāriputra tells a young woman (the Nāga princess) that she cannot possibly achieve her goal of Supreme, Perfect Enlightenment because women cannot realize five types of status, among them the status of an irreversible *Bodhisattva* (and implicitly a Buddha). The possibility of female Buddhas is clearly a vexing issue. With regard to a female Buddha, the Mahāyāna texts do not explicitly name any. Nonetheless, the Mahāyāna tradition depicts female *Bodhisattvas* of the most advanced order, such as Tārā and Kuan-yin, both of whom are the objects of major worship and veneration cults. Queen Śrīmālā "Who Had the Lion's Roar (of a Buddha)" (*Śrīmālādevīsiṃhanāda-sūtram*) comes closest, perhaps, to an implied image of female Buddhahood. According to this sūtra, Queen Śrīmālā converts many to Buddhism, is praised for her wisdom and compassion, eloquently expounds the *Dharma*, and receives the prediction that she will become a Buddha. The level of *Bodhisattvahood* which she represents became the subject of many subsequent commentaries. Some suggest Queen Śrīmālā was an irreversible *Bodhisattva*; others claim she was really a male appearing as a female; yet others demote her to a first-stage *Bodhisattva*. Like the discrepancies over Queen Śrīmālā's status, the question of whether women were ever recognized

as potential or imminent Buddhas, remains unanswered (Paul, 1979).

The personification of the 'Perfection of Wisdom' (*Prajñāpāramitā*) as feminine emerges as another oft-cited 'pro-female' feature of Mahāyāna. *Prajñāpāramitā* is worshipped as a female goddess and extolled as 'the mother of the Buddhas' (since Buddhas are born from wisdom). Some claim that she owes her femininity to a grammatical coincidence since the Sanskrit word for wisdom, '*prajñā*,' takes the feminine gender. Whether accidental or deliberate, the pursuit of wisdom – personified as female – has been the subject of much Buddhist art and has developed into a major metaphor and spiritual discipline in Mahāyāna and Vajrayāna Buddhism. The female *prajñā* (often symbolized by a lotus or a ritual bell) is typically paired with the male *upāya*, a masculine noun meaning 'compassionate skillful means' (often symbolized as a *vajra* or thunderbolt). The attainment of perfect enlightenment rests on the symbolic union of these two complementary forces. Tantric Buddhism, in particular, explicitly envisions Buddhahood as the perfect integration of these male and female energies. Visually, this is often depicted by male Buddhas and *Bodhisattvas* in sexual union with female consorts, also known in Tibetan as *yab–yum* ('father–mother') images. This highly sexual gender symbolism is unique to Tantric philosophy and practice, and it has received accolades as a dramatic valorization of the feminine as well as condemnation as the sexualization of women for the spiritual advancement of men.

The Contemporary Buddhist Women's Ordination Movement

A major issue for Buddhist women today is the status of women in Buddhist monastic institutions. The order of fully ordained nuns (Pāli: *bhikkhunī*, relevant to Theravāda Buddhists; Sanskrit: *bhikṣuṇī*, relevant to Mahāyāna and Tibetan/Vajrayāna Buddhists) is not uniformly present in Asian Buddhist settings. Currently, women can be ordained as *bhikṣuṇīs* in the Mahāyāna traditions of China, Korea, Taiwan, and Vietnam as well as in Chinese diaspora communities in countries such as Indonesia, Malaysia, and the Philippines. However, the *bhikṣuṇī* category is not available to women in the Tibetan/Vajrayāna traditions of Bhutan, the Indian Himalayas, Mongolia, Nepal, and Tibet (Mrozik, 2009). Likewise, women have been barred from *bhikkhunī* status in the Theravāda traditions of South and Southeast Asia, until recent developments. Largely in response to these status disparities and the unequal treatment of female Buddhist renunciates, an international Buddhist women's movement has emerged (hereafter referred to as the '*bhikkhunī* movement'). Broadly speaking, the *bhikkhunī* movement seeks to revive the lineage of Buddhist nuns where it has been broken, to reinstate the *bhikkhunī* order where it has died out, and to introduce the community of Buddhist nuns where it has been absent. Founded in 1987, Sākyadhitā: International Association of Buddhist Women has been instrumental in the global *bhikkhunī* movement. Sākyadhitā international conferences bring together scholars, practitioners, and activists with the aim of promoting gender equity in Buddhism and equal opportunities for education and training for Buddhist women worldwide. Though the *bhikkhunī* movement has been critiqued as a western feminist imposition, attempts have been made in Theravāda countries to revive the *bhikkhunī* order, from the early twentieth century up to the present day.

The Pāli canon recounts that the Buddha allowed women to be ordained into the *Saṅgha* and history attests to a thriving order of *bhikkhunīs* on the Indian subcontinent and beyond. However, the Theravāda *bhikkhunī* order vanished some 1000 years ago. While Sri Lanka is generally regarded as the only Theravāda country where a *bhikkhunī* order once flourished (the order of nuns appears to have lapsed in Sri Lanka in the eleventh century), scholarship suggests that *bhikkhunīs* existed in the Theravāda country of Burma until about the thirteenth century (Kawanami, 2007). There is no evidence pointing to any *bhikkhunīs* in Thai history, though isolated attempts have been made in the past to establish the order. Overall, it is widely believed that *bhikkhunīs* never existed in all of Southeast Asia. On both local and international levels, there is an ongoing controversy over the possibility of reviving the vanished ordination lineage of Theravāda nuns. This controversy hinges on the legitimacy of Higher Ordination (*upasampadā*) for women, an ordination which confers the status of '*bhikkhunī*.' According to traditionalist readings of the Pāli *Vinaya*, a female ordinand must seek Higher Ordination from a dual *Saṅgha*, that is, from the *bhikkhunī Saṅgha* and the *bhikkhu Saṅgha*. In the absence of an existing *bhikkhunī Saṅgha*, proper ordination cannot be granted.

There are a multitude of arguments for and against the revival of the Theravāda *bhikkhunī* lineage. Representative arguments for 'pro-revivalists' include: the Buddha's example of ordaining women; the Buddha's prescription for a fourfold society, of which *bhikkhunīs* are an essential component; the doctrinal premise of the equal potential of women and men to attain enlightenment; and various appeals to the secular ideals of human rights and religious freedom. Representative opinions of 'anti-revivalists' include: the illegitimacy of reinstating the Theravāda *bhikkhunī* lineage via a dual ordination procedure that involves Mahāyāna *bhikṣuṇīs*; adherence to the official teaching that once extinguished, the *bhikkhunī Saṅgha* cannot be revived; the corollary belief that only a perfectly enlightened Buddha (and implicitly the next Buddha) can reconstitute a broken *bhikkhunī Saṅgha*; and opposition to changing or manipulating the training rules outlined in the *Vinaya*.

What is perhaps most interesting is how Theravāda Buddhist women have responded to this challenge. Chatsumarn Kabilsingh, a leading proponent of the *bhikkhunī* movement in Thailand, took Higher Ordination in 2003 outside her native Thailand in Sri Lanka and became *Bhikkhunī* Dhammanandā. To fulfill the dual-*Saṅgha* ordination rule and revive the *bhikkhunī Saṅgha*, Dhammanandā advocates that female ordinands turn to Chinese nuns of the uninterrupted Dharmaguptaka lineage to orchestrate *upasampadā*. Critics, however, question the legitimacy of 'cross- *Saṅgha* ordination'

and the intactness of the ordination lineage (Seeger, 2008). Apart from religious matters, many perceive the *bhikkhunī* lineage as foreign and 'non-Thai.' Moreover, many religious women do not want to follow in Dhammanandā's path and prefer to practice Buddhism in ways that are more commonly recognized and accepted in their local communities. Overall, the *bhikkhunī* movement has enjoyed greater success in Sri Lanka (however not without controversy), in large part because a *bhikkhunī* lineage is part of Sri Lankan history, can be perceived as local, and appeals to issues of national and ethnic identity (Mrozik, 2009).

In Thailand, Burma, Sri Lanka, and other traditionally Theravāda countries, women have pursued the renunciant life even though the *bhikkhunī* lineage has died out or been altogether absent. New categories of female renunciation have emerged in these societies and have persisted for several centuries: namely the *mae chis* of Thailand, the *thilashin* of Burma, and the *dasasilmātās* of Sri Lanka. Though each is distinctive, these alternative forms of renunciation share common features. They observe eight to ten precepts, shave their heads, don monastic robes (of varying colors), and observe celibacy. Concerns about these alternative renunciant institutions include: their liminal status between lay and monastic forms of practice (Jordt, 1988; Falk, 2000); their marginalized, unsupported, and extremely low status (Kabilsingh, 1991); their valuation as a poor substitute for the *bhikkhunī* ideal (Tomalin, 2006); and doubt as to whether they can factually or potentially be real alternatives to the missing *bhikkhunī* order (Seeger, 2009). Many of these renunciant women are not 'interested' in becoming *bhikkhunīs* (Battaglia, 2007; Falk, 2007; Kawanami, 2007). Reasons for resisting *bhikkhunī* ordination include doubts about lineage purity, and the validity of the ordination procedure; concern that attaining *bhikkhunī* status would place nuns more strictly under the control of the *Saṅgha*; skepticism that the *bhikkhunī* role is viable or necessarily carries more social or spiritual efficacy; fear of social censure and loss of lay support; and allegiance to their own categories of renunciation as worthy and legitimate. The fact that some women in these alternative orders do not want full ordination or do not staunchly support the revival of the *bhikkhunī* lineage is a source of tension in the contemporary *bhikkhunī* movement.

This is a particularly germane time for the study of women in Buddhism. Myriad visions of female religiosity gain recognition and momentum, particularly with the prominence of global Buddhist–feminist networks, the role of the internet and digital devices in information dissemination and accessibility, and the increasing scholarship and print-media devoted to female practitioners of Buddhism and Buddhist women's communities. It remains to be seen how the *bhikkhunī* movement will play out in Southeast Asia and beyond, and how the continued involvement of women will change the face of Buddhism, both locally and globally.

See also: Buddhism; Globalization and Religion; Monasticism: West and East; Religions of East Asia; Religions of India; Women's Religiosity.

Bibliography

Allione, Tsultrim, 1986. Women of Wisdom. Arkana, New York.

Arai, Paula Kane Robinson, 1999. Women Living Zen: Japanese Sōtō Buddhist Nuns. Oxford University Press, New York.

Battaglia, Lisa, 2007. Women Who Have Gone Forth: Gender and Religious Identity among Buddhist Nuns in Thailand (Ph.D. dissertation). Vanderbilt University, Nashville, TN.

Blackstone, Kathryn R., 1998. Women in the Footsteps of the Buddha: Struggle for Liberation in the Therīgāthā. Curzon Press, Surrey, Great Britain, pp. 59–81.

Boucher, Sandy, 1988. Turning the Wheel: American Women Creating the New Buddhism. Harper & Row, San Francisco, CA.

Brown, Sid, 2001. The Journey of One Buddhist Nun: Even against the Wind. State University of New York Press, Albany, NY.

Cabezón, José Ignacio (Ed.), 1992. Buddhism, Sexuality, and Gender. State University of New York Press, Albany, NY.

Cabezón, José Ignacio, 1998. Homosexuality and Buddhism. In: Leyland, Winston (Ed.), Queer Dharma: Voices of Gay Buddhists, vol. 1. Gay Sunshine Press, San Francisco, CA.

Carmody, Denise Lardner, 1991. Religious Woman: Contemporary Reflections on Eastern Texts. Crossroad, New York.

Chung, In Young, 1999. A Buddhist view of women: a comparative study of the rules for bhikṣuṇīs and bhikṣus based on the Chinese prātimokṣa. Journal of Buddhist Ethics 6, 1–62.

Collins, Steven, 1982. Selfless Persons: Imagery and Thought in Theravāda Buddhism. Cambridge University Press, Cambridge.

Dresser, Marianne (Ed.), 1996. Buddhist Women on the Edge: Contemporary Perspectives from the Western Frontier. North Atlantic Books, Berkeley, CA.

Falk, Nancy Auer, 1974. An image of woman in old Buddhist literature: the daughters of Māra. In: Plaskow, Judith, Arnold, Joan (Eds.), Women and Religion. Scholars Press for the American Academy of Religion, Missoula, MT.

Falk, Nancy Auer, 1980. The case of the vanishing nuns: the fruits of ambivalence in ancient Indian Buddhism. In: Falk, Nancy Auer, Gross, Rita M. (Eds.), Unspoken Worlds: Women's Religious Lives in Non-Western Cultures. Harper & Row, San Francisco, CA, pp. 207–224.

Falk, Monica Lindberg, 2000. Women in between: becoming religious persons in Thailand. In: Findly, Ellison Banks (Ed.), Women's Buddhism, Buddhism's Women: Tradition, Revision, Renewal. Wisdom Publications, Boston, MA, pp. 37–57.

Falk, Monica Lindberg, 2007. Making Fields of Merit: Buddhist Female Ascetics and Gendered Orders in Thailand. University of Washington Press, Seattle, WA.

Faure, Bernard, 2003. The Power of Denial: Buddhism, Purity, and Gender. Princeton University Press, Princeton, NJ.

Friedman, Lenore, 1987. Meetings with Remarkable Women: Buddhist Teachers in America. Shambhala, Boston, MA.

Gregory, Peter N., Mrozik, Susanne (Eds.), 2008. Women Practicing Buddhism: American Experiences. Wisdom Publications, Boston, MA.

Gross, Rita M., 1993. Buddhism after Patriarchy: A Feminist History, Analysis, and Reconstruction of Buddhism. State University of New York Press, Albany, NY.

Gutschow, Kim, 2004. Being a Buddhist Nun: The Struggle for Enlightenment in the Himalayas. Harvard University Press, Cambridge.

Gyatso, Janet, 2003. One plus one makes three: Buddhist gender, monasticism, and the law of the non-excluded middle. History of Religions 43, 89–115.

Jordt, Ingrid, 1988. Bhikkhuni, thilashin, mae-chii: women who renounce the world in Burma, Thailand, and the classical Pali Buddhist texts. Crossroads: An Interdisciplinary Journal of Southeast Asian Studies 4 (1), 31–39.

Kabilsingh, Chatsumarn, 1991. Thai Women in Buddhism. Parallax Press, Berkeley, CA.

Kawanami, Hiroko, 2007. The *bhikkhunī* ordination debate: global aspirations, local concerns, with special emphasis on the views of the monastic community in Burma. Buddhist Studies Review 24 (2), 226–244.

Klein, Anne Carolyn, 1995. Meeting the Great Bliss Queen: Buddhists, Feminists, and the Art of the Self. Beacon Press, Boston, MA.

Lang, Karen Christina, 1986. Lord death's snare: gender-related imagery in the Theragāthā and the Therīgāthā. Journal of Feminist Studies in Religion 11 (2), 63–79.

Minamato, Junko, 1991. Buddhist attitudes: a woman's perspective. In: Becher, Jeanne (Ed.), Women, Religion, and Sexuality: Studies on the Impact of Religious Teachings on Women. Trinity Press International, Philadelphia, PA.

Mrozik, Susanne, 2009. A robed revolution: the contemporary Buddhist nun's (bhikṣuṇī) movement. Religion Compass 3 (3), 360–378.

Murcott, Susan, 1991. The First Buddhist Women: Translations and Commentary on the Therigatha. Parallax Press, Berkeley, CA.

Neumaier-Dargyay, Eva K., 1995. Buddhist thought from a feminist perspective. In: Joy, Morny, Neumaier-Dargyay, Eva K. (Eds.), Gender, Genre, and Religion: Feminist Reflections. Wilfrid Laurier University Press, Calgary.

Owen, Lisa Battaglia, 1997. Toward a Buddhist feminism: Mahayana *Sutras*, feminist theory, and the transformation of sex. Asian Journal of Women's Studies 3, 8–51.

Paul, Diana Y., 1979, 1985. Women in Buddhism: Images of the Feminine in the Mahāyāna Tradition. University of California Press, Berkeley, CA.

Powers, John, 2009. A Bull of a Man: Images of Masculinity, Sex, and the Body in Indian Buddhism. Harvard University Press, Cambridge, MA.

Schalow, Paul Gordon, 1992. Kūkai and the tradition of male love in Japanese Buddhism. In: Cabezón, José Ignacio (Ed.), Buddhism, Sexuality, and Gender. State University of New York Press, Albany, NY, pp. 215–230.

Seeger, Martin, 2008. The *bhikkhunī*-ordination controversy in Thailand. Journal of the International Association of Buddhist Studies 29 (1), 155–183.

Seeger, Martin, 2009. The changing roles of Thai Buddhist women: obscuring identities and increasing charisma. Religion Compass 3/5, 806–822.

Shaw, Miranda, 1994. Passionate Enlightenment: Women in Tantric Buddhism. Princeton University Press, Princeton, NJ.

Sponberg, Alan, 1992. Attitudes toward women and the feminine in early Buddhism. In: Cabezón, José Ignacio (Ed.), Buddhism, Sexuality, and Gender. State University of New York Press, Albany, NY, pp. 3–36.

Tomalin, Emma, 2006. The Thai *bhikkhuni* movement and women's empowerment. Gender & Development 14 (3), 385–397.

Walters, Jonathan S., 1994. A voice from the silence: the Buddha's mother's story. History of Religions 33, 358–379.

Willis, Janice D., 1985. Nuns and benefactresses: the role of women in the development of Buddhism. In: Haddad, Yvonne Yazbeck, Findly, Elisson Banks (Eds.), Women, Religion, and Social Change. State University of New York Press, New York, pp. 59–85.

Wilson, Liz, 1995. The female body as a source of horror and insight in post-Ashokan Indian Buddhism. In: Law, Jane Marie (Ed.), Religious Reflections on the Human Body. Indiana University Press, Bloomington, pp. 76–99.

Wilson, Liz, 1996. Charming Cadavers: Horrific Figurations of the Feminine in Indian Buddhist Hagiographic Literature. The University of Chicago Press, Chicago.

Zwilling, Leonard, 1992. Homosexuality as seen in Indian Buddhist texts. In: Cabezón, José Ignacio (Ed.), Buddhism, Sexuality, and Gender. State University of New York Press, Albany, NY, pp. 203–214.

Women's Religiosity

Grace Davie, University of Exeter, Devon, UK
Tony Walter, University of Bath, Bath, UK
Darren E Sherkat, Southern Illinois University, Carbondale, IL, USA

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16532–16534, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

In Western societies, women are consistently shown to have stronger religious beliefs and engage in religious participation and devotion more frequently than men. Sociologist and psychologists have explored several possible explanations for these differences, including women's higher levels of family involvement, lower rates of labor force participation, higher rates of deprivation, and greater health risks earlier in life due to childbirth.

In Western societies, women are more religious than men on every measure of religiosity. Why this should be has become a major interest of sociologists of religion. Yet, the sociology of religion only belatedly included gender on its agenda (Woodhead, 2000), and feminists have been more concerned with why there are so few women in the priesthood than with why there are so many in the pews. Scholars have also been concerned with documenting the historical patriarchal stance of Christianity, the pre-Christian spirituality of women, and the spirituality of hitherto ignored but highly gifted individuals, rather than documenting the ordinary religiosity of millions of churchgoing women. Feminist theories that religion is bad for women need to come to terms with, but in fact often ignore, women's opting for formal religion in greater numbers than men. This article reviews the possible explanations for women's greater religiosity, focusing particularly on mainline churchgoing in the West in the twentieth century.

Data

Whether it is a matter of churchgoing, private prayer, or the content of religious belief, women appear more religious than men, and have been for at least 200 years (Douglas, 1978). Within Western Christianity, mainline churches demonstrate the greatest fall off in attendance by men; independent, Pentecostal, and orthodox churches have been able to resist the trend. The extent to which secularization is caused by men reducing religious observance more than women varies by country (Woodhead, 2000; Freese, 2004).

The dimension of religious life where the difference between women and men is biggest is private devotion – prayer and Bible reading. This raises the question whether men engage in religious practices when they are publicly acceptable or even required, but tend not to bother with private devotions when there is no social pressure. It is possible that secularization, defined as the decline in the social importance of religion, tends at the same time to lead to religion's feminization. Islam and Judaism, in which religion is still part of public life, involve men more than women in public worship. Studies in the West have also focused on how men's irreligiosity may be a function of their inherent tendency to engage in more risky behaviors (Miller and Hoffmann, 1995; Miller and Stark, 2002; Freese, 2004).

Independent of degree of belief, what men and women believe differs. Women, if they are asked to describe the God in whom they believe, concentrate rather more on a God of love, comfort, and forgiveness, while men tend to refer to a God of power, planning, and control (Walter and Davie, 1998).

Material and Social Deprivation

Neo-Marxist deprivation compensation theories see religion as a way in which oppressed people cope with their misery and suffering. If, as feminists argue, women are more oppressed than men, we would expect more women than men to turn to religion. There certainly are examples of this, especially, in the early periods of industrialization both historically in the West and in less-developed societies. In advanced industrial societies, however, there is less evidence for this theory of female religiosity, religious participation being related to parenthood and employment status rather than to measurable indices of gendered deprivation.

Lack of Status

Women's systematic exclusion from better paid jobs, from political office, and from other parts of the public world has reduced their social status. Christianity's proclamation that God loves every individual equally may be more immediately attractive to those who are not accorded respect in society. Once inside the church, however, women may find themselves even more deprived of status and power vis-à-vis male members. Yet women remain in greater numbers than men. Ozorak (1996) found that many women valued their churches for the social and emotional support, which existed irrespective of their exclusion from formal power; for those more concerned with connection and relationship than with hierarchy and power, the female fellowship of the church empowered them even though the office-holding men monopolized the glory.

Conservative churches and synagogues that emphasize traditional gender roles within the family appeal to many women: they provide status for the wife and mother, and provide hope that conversion may turn men into responsible husbands and fathers.

Guilt

A number of psychological studies show that women experience more guilt feelings than men. If one of the main purposes of Christianity is to assuage guilt, then we would expect it to attract those who feel a need for forgiveness. Some churches actively foster guilt feelings, which they then offer to resolve. Churches have often portrayed women as more guilty, more profane, than men. Much of this is linked to sexual guilt and a double standard enforcing women's chastity. Insofar as Christianity induces guilt, we may say that religion not only compensates for a societally created deprivation of status and self-respect, but actually fosters the very deprivation that it then compensates for. This is perhaps the nearest to a feminist theory of why women should be attracted to a religion that oppresses them. Though there is considerable historical evidence of this process, and studies show that members of exclusivist sectarian groups express negative views of nonmarital sexuality even though members often engage in premarital, extramarital, and homosexual relations (Sherkat, 2014).

Birth and Child Rearing

Religion is supposed to provide comfort against anxiety and existential terror; psychological studies point out that women generally have higher levels of fear than men (Miller and Hoffmann, 1995; Miller and Stark, 2002). This could be related to their more direct involvement in the inherently scary businesses of birth and death and their day-to-day responsibility for child rearing (Sherkat, 2008). Men have comparable familiarity with sickness and death only on active service during wartime – a time when they may be more likely to turn to religion. There have been very few studies into the spirituality of childbirth, so it is too early to say whether experiencing the 'miracle' of birth accounts in part for female religiosity. Much more research has been done into religiosity in relation to child rearing. The general pattern is that having children to look after raises men to the religious level of women; it is possible that the vulnerability and dependency of young children raises religious issues that men can normally avoid. It is, therefore, parenthood, not motherhood, that raises questions of existential anxiety and a desire to involve children in religion.

Workforce Participation

A major thread in theories of secularization is that the more people are immersed in the institutions of modernity, notably by working for bureaucratic and or capitalist organizations, the more secular they are likely to become. Women who stay at home are, therefore, protected from the secularizing influence of the modern marketplace. This hypothesis (classically formulated by Luckmann) has led to a number of empirical surveys aiming to determine whether, and if so why, women who go out to work are less religious than those who stay at home.

The findings of these studies vary by country. Taking a paid job in, for example, Australia where there is a secular ethos in the workplace is likely to have a more secularizing effect than taking a job in the American Bible Belt. Since both data and hypotheses remain inconclusive, a constructive avenue of research might be to look at the type of paid work done by women. Women's paid work is far more likely than men's to be of a caring nature, involving skill in interpersonal relations and 'emotional labor.' If what protects women (and some men) from the secularity of modernity is not family life but interpersonal care, then those employees engaged in caring work should be compared with those engaged in other kinds of work. If research were to confirm that women paid to care are more religious than those paid to do other things, we would still be left with a puzzle: are they more religious because the caring nature of the work protects them from secularity, or do they enter caring work because they were more religious to start with?

An Ethic of Connection

Gilligan (1982) considers that women operate according to a flexible ethic of compassion and connection rather than abstract justice. Woodhead (2000) has used this insight to critique the 'privatization thesis,' namely that in an alienating, plural modern world, people retreat to the private sphere where they find identity, meaning, and perhaps religion. She suggests that this thesis makes sense from a male perspective, but women are less involved in the alienating public sphere. For them, religion has less to do with providing meaning and more to do with articulating the social relationships. Traditional churches do this by encouraging the family, but they can also encourage other kinds of relationships. A religion whose founder emphasized love and compassion is more likely to appeal to Gilligan's women. Indeed, male clergy have often struggled to make Jesus appeal to masculine men, and research has shown that gay men are relatively religious when compared to heterosexual men or lesbians (Sherkat, 2002). This is echoed in Thompson's (1991) finding that what is correlated with Christian religiosity is not gender but a 'feminine' worldview, that is, one that is affectionate, sympathetic, sensitive to the needs of others, eager to soothe another's feelings, tender, and loving toward children.

Death

The extent to which religion is a way of helping humans deal with death has been extensively debated. What is clear is that women encounter death disproportionately as carers, and also as mourners. In the times of high infant mortality, women may have been more affected than their husbands by the deaths of their children, though men were more likely than women to lose a spouse – as nineteenth-century gravestones bear ample witnesses. Today, with women living longer than men and marrying younger, wives are very likely to outlive their husbands. Spousal bereavement is now a much more common experience for women than for men. Whether or not this helps to explain women's greater religiosity, it seems very likely to lie at the root of the most common form of afterlife belief in the modern West, namely a belief in reunion with loved family members. Women who place a high value on personal relationships are unlikely to give them up when death intervenes, and this may possibly account for women being more likely than men to experience the presence of one who is dead. Life

after death is central to Christianity, and insofar as women have a greater interest in there being life after death, we may postulate that some women's religiosity may be rooted in their experience of death and bereavement. On the other hand, it may be the other way around – that more women than men believe in life after death simply because of their general religiosity.

See also: Religion and Health: Clinical Considerations and Applications; Religion, Family, and Kinship; Religion: Mobilization and Power; Religion: Morality and Social Control; Women in Buddhism.

Bibliography

Beit-Hallahmi, B., Argyle, M., 1996. The Psychology of Religious Behavior, Belief and Experience. Routledge, London.

Douglas, A., 1978. The Feminization of American Religion. Alfred Knopf, New York.

Francis, L., 1997. The psychology of gender differences in religion: a review of empirical research. Religion 27, 81–96.

Freese, J., 2004. Risk preferences and gender differences in religiousness: evidence from the world values survey. Review of Religious Research 46, 88–91.

Gilligan, C., 1982. In a Different Voice. Harvard University Press, Cambridge, MA.

Miller, A.S., Hoffmann, J., 1995. Risk and religion: an explanation of gender differences in religiosity. Journal for the Scientific Study of Religion 34, 63–75.

Miller, A.S., Stark, R., 2002. Gender and religiousness: can socialization explanations be saved? American Journal of Sociology, 1399–1423.

Ozorak, E., 1996. The power but not the glory: how women empower themselves through religion. Journal for the Scientific Study of Religion 35, 17–29.

Sherkat, D.E., 2014. Changing Faith: The Dynamics and Consequences of Americans' Shifting Religious Identities. NYU Press, New York.

Sherkat, D.E., 2008. Beyond belief: atheism, agnosticism, and theistic certainty in the United States. Sociological Spectrum 28, 438–459.

Sherkat, D.E., 2002. Sexuality and religious commitment: an empirical assessment. Journal for the Scientific Study of Religion 41, 313–323.

Thompson, E., 1991. Beneath the status characteristics: gender variations in religiousness. Journal for the Scientific Study of Religion 30, 381–394.

Walter, T., Davie, G., 1998. The religiosity of women in the modern West. British Journal of Sociology 49, 640–660.

Woodhead, L., 2000. Feminism and the sociology of religion: from gender-blindness to gendered difference. In: Fenn, R. (Ed.), The Blackwell Companion to the Sociology of Religion. Blackwell, Oxford, UK, pp. 67–84.

Women's Status and National and International Security

Valerie M Hudson, Texas A&M University, College Station, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Until the turn of the twenty-first century, mainstream scholars and professionals in national and international security have not viewed the well-being and security of women as important. Empirical literature has emerged, which demonstrates that this linkage is significant. This article surveys the literature, examining in particular issues of war, peace, and stability of nations in relation to women's status and security.

Introduction

The study of national and international security has historically centered on relations between governments and particularly on the prevention of great power war. It has traditionally included topics such as the causes of war and war termination, empires and world order and has expanded in more recent decades to include study of civil wars, humanitarian intervention, and other issues.

Notably lacking in most literature in Security Studies is attention to women's security and the role of women in promoting security. And yet, in the policy-making community, we hear statements by world leaders that there is a strong relationship between the situation and security of women, on the one hand, and the security of the nations in which they live. For example, in 2010, then-Secretary of State Hillary Clinton asserted that, "the subjugation of women is a direct threat to the security of the United States" (AFP, 2010). In 2006, then-Secretary General of the United Nations, Kofi Annan, opined, "The world is starting to grasp that there is no policy more effective in promoting development, health and education than the empowerment of women and girls. And I would venture that no policy is more important in preventing conflict, or in achieving reconciliation after a conflict has ended" (Annan, 2006). What does the social science literature tell us about the relationship between the security of women and national and international security?

First a note on the measurement of the security and well-being of women. There are two primary approaches to such measurement. One is to measure the difference between men and women on measures deemed important to security and well-being. So, for example, one might examine the difference in years of schooling between men and women within a society, or the difference in the representation of men and women in national legislatures. These measures are often characterized as measures of 'gender equality' or 'gender inequality'. Other studies use direct measures of the level of the security and well-being of women. For example, researchers might use the prevalence of rape and sexual assault against women in a society to examine women's physical security. Both approaches are justifiable, and the type of measure used will typically be noted in the survey below.

War and Peace

Studies focusing on the role of gender equality and interstate conflict highlight the relationship between international violence and domestic gender equality (Caprioli, 2000, 2003; Caprioli and Boyer, 2001; Tessler and Warriner, 1997). In general, countries characterized by gender inequality are more likely to be involved in interstate disputes and more likely to rely on violence to settle those disputes.

Monty Marshall and Donna Ramsey (1999) conducted what may be the first quantitative study of feminist international relations theories regarding state behavior. The authors tested the feminist claim that issues of gender are important in predicting the use of force during interstate disputes. They included gender empowerment in their multiple regression, first to test whether gender equality helped predict state use of force, and second as a means of broadening our understanding of democracy within the democratic peace literature by using gender equality as one indicator of the entrenchment of democratic values. Although the authors make no strong causal inferences based on their analysis, they do find a robust relationship between gender empowerment and the willingness of states to use force. The authors further argue that the use of violence institutionalizes discrimination and inhibits further gender equality. Marshall and Ramsey (1999: p. 32) conclude "that gender theory and analysis may be successfully integrated with traditional international relations theory and quantitative research methods".

Adding to the quantitative feminist international relations literature, Caprioli (2000) drew upon the empirical findings of the public opinion surveys discussed above to test the potential impact of domestic gender equality on state behavior internationally. She used logistic regression to test the relationship between several measures of women's social, economic, and political equality and the escalation of violence during militarized interstate disputes. Caprioli concludes that higher levels of domestic gender equality result in less emphasis on military action in settling international disputes.

Caprioli and Boyer (2001) extend Caprioli's earlier work to assess the impact of domestic gender equality on state's international crisis behavior. The authors provide an empirically based descriptive analysis of the behavior of states with female leaders as primary decision makers during times of

Certain sections of this article are reprinted with publisher's permission from Hudson, Valerie, Charli Carpenter, Mary Caprioli, 2010. Gender and Global Security. In: Robert A. Denemark (Ed.), The International Studies Encyclopedia. Blackwell, New York.

international crises. They then use logistic regression to test the relationship between domestic gender equality and the level of violence exhibited during international crises, concluding that the severity of violence in a crisis does decrease as domestic gender equality increases.

In a later study, Caprioli (2003) further tested the relationship between domestic gender equality and state behavior during militarized interstate disputes. In this piece, she included an analysis relating gendered structural inequality to domestic norms of violence. Using logistic regression to assess the role of domestic gender equality in predicting the likelihood of a state using force first during interstate disputes permits a more rigorous test of the author's earlier work by isolating the effects of reciprocated violence. Caprioli concludes that higher levels of gender equality lower the level of state aggression during interstate disputes. In other words, states with higher levels of gender inequality are more likely to use force first in interstate disputes.

Patrick Regan and Aida Paskeviciute (2003) extended the analysis of gender equality and state use of force internationally beyond the state level by focusing on whether the gender distribution of political power at the societal level influences the willingness of the ruling elite to engage in interstate disputes. The authors conclude that women's access to the political arena helps predict the likelihood of a state engaging in interstate disputes and war. They also explore the extent to which fertility rates and rates of women's employment are associated with the use of force by a state. Provocatively, and in contrast to the democratic peace theory, they argue that pairs of nations with low fertility rates (representing a higher status for women in society) do not fight each other (see also Fukuyama's gendered alternative explanation of the democratic peace thesis (1998)). The authors offer a policy prescription suggesting that support for family planning facilities can facilitate more peaceful interstate relations.

Caprioli and Peter Trumbore, using established cross-national data sets such as the Correlates of War find that states characterized by norms of gender and ethnic inequality as well as human rights abuses are more likely to become involved in militarized interstate disputes and in violent interstate disputes, to be the aggressors during international disputes, and to rely on force when involved in an international dispute (Caprioli and Trumbore, 2003a,b, 2006). David Sobek and coauthors confirm Caprioli and Trumbore's findings that domestic norms centered on equality and respect for human rights reduce international conflict (Sobek et al., 2006).

Hudson and coauthors examine the links between the security of women and the security of states, using data from the WomanStats Project, the Global Peace Index, and other data sources. Hudson and Den Boer (2004) research the extent to which women are missing from Asian populations, finding that in some of the most populous Asian nations, including China and India, which accounts for approximately 40% of world population, that birth and childhood sex ratios are becoming more abnormal and masculinized over time. They ask what effects the creation of a subgroup of the young adult male population who will not be able to become heads of household will have on national security. They find that 12–15% of the young men in this age-group in China and India will be 'bare branches', the Chinese term given to young men who became in effect surplus during historical time periods of prevalent female infanticide. Probing historical and aggregate empirical work, they find that societal instability, primarily in the form of violent crime rates, increase significantly as a result, and that the strategies used by governments to mitigate this growing instability may have regional security ramifications.

Hudson et al. (2008/2009) investigate whether there is a relationship between the level of violence against women in society, and state levels of peacefulness and compliance with international norms. They find a strong relationship, and in multivariate analysis this relationship is more predictive of state peacefulness and norm compliance than levels of democracy, levels of wealth, or presence of Islamic civilization. Furthermore, Hudson et al. (2012b) find a strong relationship between inequities in family law within a society and its level of violence against women. Putting these two articles together, it can be argued that variables related to the status of women may be much more important in understanding high politics than Security Studies has heretofore understood.

Hudson et al. (2012a) broaden their focus to examine how three particular aspects of female subordination – lack of physical security, lack of equity in family law, and lack of parity in collective decision making – undermine not only the security of women, but also create authoritarian, unstable, and bellicose states. Hudson et al. (forthcoming) are also able to trace how systematic female subordination encourages the growth of clan-based societies that undermine not only democratic rule, but also national economic growth and levels of national stability.

In sum, this body of work demonstrates that the promotion of gender equality goes far beyond the issue of social justice and has important consequences for national and international security. Using conventional methodologies, these authors are able to demonstrate a real linkage between the security of women and the security of states. Indeed, this linkage may be stronger than that posited for the democratic peace, or the capitalist peace. It may be more important for national and international security what family law looks like in a given state than how highly it is rated in the Freedom House index.

Female Leadership and Use of Force

Alice Eagly and her coauthors have found important differences in leadership style between men and women (Eagly et al., 2003). Performing a meta-analysis of 45 studies of male/female leadership in industrialized nations, they contrast three styles of leadership: transformation, transactional, and laissez-faire. Using the multifactorial leadership questionnaire studies, they associate transformational leadership with instilling respect, purpose, optimism, creativity in subordinates, with an emphasis on mentoring. Transactional leadership is a style that emphasizes interaction based on appeals to the subordinate's self-interest. Laissez-faire style was associated with a lack of involvement on the part of the leader, or a failure to take responsibility.

The authors note that the interplay between leadership roles and gender roles may be significant: "The tendency for the demands of the female gender role and leader roles to be contradictory can also foster differing behavior in female and

male leaders." One reason that gender roles have different implications for female and male leaders is thus that inconsistency often exists between the predominantly communal qualities that perceivers associate with women (e.g., friendly, kind, unselfish) and the predominantly agentic qualities that they generally believe are necessary to succeed as a leader (e.g., assertive, masterful, instrumentally competent).

"People's beliefs about leaders are thus more similar to their beliefs about men than women" (Eagly et al., 2003: p. 572). The authors go on to note that there is a backlash against women leaders who behave in a more directly agentic manner; that is, women who display high levels of competence, or exert authority over men, or use a dominant style of communication are disliked and mistrusted (p. 573). As a result, female leaders must 'temper' their leadership "with sufficient displays of female-typical communal behaviors" (p. 573). The authors note that "negative reactions can be lessened when female leaders are careful to also display warmth and lack of self-interest by, for example, expressing agreement, smiling, supporting others, and explicitly stating an interest in helping others reach their goals" (p. 573).

The meta-analysis included studies with people in both the United States (53%) and other industrialized countries (47%). The authors find that women are more likely to adopt a transformational leadership style than men, and that as a result, women's performance as leaders is on average more effective than that of men.

Natalie Florea et al. (2003) empirically tested hypotheses inspired by feminist research regarding differences in negotiating styles between men and women, using data from a large teaching simulation, GlobalEd, conducted by their university involving students from junior high to college levels. The authors found that gender affected negotiating style based on evidence indicating that all-female groups behaved differently from all-male groups and that mixed gendered groups behaved differently from both single sex groups. Women tended to negotiate by appealing to higher principles such as justice or the consequences on participants (see also Gilligan, 1982).

This research has recently been extended in Boyer et al. (2009). Their simulation allows for the formation of negotiating groups that are all-male, all-female, and mixed gender. While students are aware of the gender composition of their own group, they are blind to the gender composition of the groups with which they are negotiating. There is also a 'mediator', called the SimCon, whose gender is unknown to the negotiating groups. They find that the addition of women alters the behavior of the group. Message volume is higher when females are involved either as negotiators or as mediators. They also find that for junior high students, the presence of a female SimCon increases collaborative and creative styles of negotiating, but for the high school students, a female mediator increases the prevalence of conflictual and assertive styles. They conclude "gender plays a significant role in social and political interaction" (p. 45).

Rose McDermott and Jonathan Cowden examine sex differences in aggression with the context of a simulated crisis game (McDermott and Cowden, 2002). In these experiments with college students, all-female pairs proved significantly less likely than all-male pairs to spend money on weapons procurement or to go to war in the face of a crisis. In further research, McDermott and her coauthors find that in simulation, males are more likely to display overconfidence prior to gaming, and are more likely to use unprovoked violence as a tactic (Johnson et al., 2006). King et al. (1991) had undergraduate students engage in prisoners' dilemma games played over computer. Some students were told the sex of their opponents, while others were not told. The researchers coded the strategies used as being competitive, cooperative, or tit for tat. When their opponents used competitive strategies, students who did not know the sex of their opponent were three times more likely to assume their opponents were male. There is also some evidence from simulation studies that when negotiators take place in the presence of male onlookers, negotiations become more contentious, with outcomes more zero-sum in nature than if negotiators are watched by females (Pruitt et al., 1986). Conversely, this same study finds that when negotiators are negotiating on behalf of women (who may or may not be present), the negotiations are more protracted and difficult than when the negotiators' constituents are male. These types of simulations, despite their constraints, permit the inclusion of sex-based psychological variables in theories concerning the micro processes by which gender differences might affect resulting state security processes and outcomes.

National Stability and Governance

Caprioli (2005) extends her aforementioned findings relating gender equality to interstate behavior to internal or domestic-level conflict. This research tests the implications of feminist theories by providing an analysis of structural violence and the role of gender inequality and discrimination in nationalist uprisings to assess this variable's potential role in predicting intrastate violence. Caprioli uses logistic regression to examine the impact of gender inequality on the likelihood of intrastate conflict and concludes that domestic gender equality reduces the occurrence of intrastate violence. These findings were replicated by Erik Melander (2005). In other words, states characterized by gender inequality are more likely to experience intrastate conflict, confirming the basic link between gender inequality and intrastate conflict.

A considerable amount of case study research also suggests that gender discourses and identities also become a means, during civil conflicts, to create and maintain ethnic, sectarian, or other in-group/out-group identities (Yuval-Davis, 1997; Elshtain, 1987). Because of the role women come to play in such conflicts as symbols, targets, and packages which can be used to send messages between men of contending groups, gender norm policing within in-groups tends to become heightened during periods of conflict. Implications for those concerned with the dampening of civil or ethnic conflict are to avoid policies that reinforce disparate treatment of women by 'their' men within a country context and to promote the opportunities for women to participate and organize collectively on the basis of a shared identity as women (Handrahan, 2004).

Many international institutions nominally include a commitment to the development of democracy worldwide – a sensible thing to do considering the evidence that states with more democratic institutions are less likely to be involved in conflict with other democratic nations (Maoz and Russett,

1993; Ray, 1995). This project cannot be effective if the gender dimensions of democratic state-building are overlooked. Political variables at the state level have been related to the situation of women, most specifically levels of corruption. For example, in a study of 80 countries, a negative correlation between indices of corruption and indices of women's social and economic rights was found (Kaufmann, 1998). Because decreases in political corruption increase investment and growth, gender equity thus exerts additional influence on economic growth (World Bank, 2001). According to an Inter-Parliamentary Union survey of 187 women holding public office in 65 countries (2000), women's presence in politics increases the amount of attention given to social welfare, legal protection, and transparency in government and business, and 80% of respondents said that women's participation restores trust in government. An analysis by Swers concerning the United States concurs in its general conclusions (Swers, 2002).

Furthermore, there is a strong and statistically significant association between high levels of gender inequality and authoritarianism (Betzig, 1986). Betzig finds that polygyny, in particular, is highly predictive of authoritarian government. This is so because polygyny creates an underclass of less-privileged young males who will never become heads of households due to polygyny-induced scarcity of women. Along the same lines, in an empirical analysis of 157 countries, M. Steven Fish (2002) disconfirms the notion that Islamic societies per se are disproportionately involved in conflict or disproportionately suffer from authoritarian rule. Rather, Fish uncovers two indicators that better explain the variance of these variables in the Islamic world: sex ratio and the literacy gap between males and females. Fish finds that models incorporating these two variables are significantly correlated to authoritarianism in Islamic countries. He hypothesizes that the oppression of females – one of the earliest social acts observed by all in the society – provides the template for other types of oppression, including authoritarianism, in Islamic nation-states.

National Economic Development and Women

A substantial literature links the treatment of women to important state-level variables. Scholarly attention to the link between women and the state arguably began in the field of development. As early as 1970, Ester Boserup argued that omission of gender aspects of development led to project failure. Since her pioneering work, we have seen fruitful waves of successive research in the paradigms of Women in Development, Women and Development, and Gender and Development (Rathgeber, 1990; Chen, 1992; Jacobson et al., 1992; Sen, 1989; Chowdry and Nair, 2002), and the establishment of strong cross-national linkages between gender variables and economic variables.

Examples of the impact of women and women's labor on development are plentiful. According to a recent study by the World Economic Forum, the lower the gender gap within a country, the higher its global competitiveness ranking, and the higher its level of gross domestic product (GDP) per capita (2007: p. 21). The World Bank notes that in nearly all countries, women work longer hours than men, given their double day of productive and reproductive activities, and that women dominate nonmarket activities in all countries (World Bank, 2001). Studies show that an increase in women's income increases the share of household money spent on education, health, and nutrition, while simultaneously decreasing the share of household income spent on alcohol and cigarettes (Hoddinott and Haddad, 1995). A series of striking studies show that the effect of female income on child survival is nearly 20 times larger than that of male income (Thomas, 1990; Thomas et al., 1997). Micro-credit loans to women, pioneered by Nobel Peace Prize winner Muhammad Yunus, provide greater returns to families than micro-credit loans to men (Yunus and Jolis, 2003). It has also been shown that the proportion of the workforce that is female bears a statistically significant relationship to the level of national economic growth (Klasen, 1999) and performance (King and Mason, 2001).

As a result of these research results, regional and global development planners now routinely address the role of women in successful development. For example, the UN Millennium Goals include the goal of empowering women, and occasioned 'Women Watch 2000' (later renamed The World's Women), an effort to more closely monitor the status of women on a cross-national basis to formulate nontraditional indicators of development (UN, 2005). The path-breaking Arab Human Development Report of 2002 identified the generally low status of women in Arab societies to be one of the four major variables retarding growth and advancement therein. This argument was elaborated in the report of 2005, which argues, "the rise of women is in fact a prerequisite for an Arab renaissance, inseparably and causally linked to the fate of the Arab world" (AHDR, 2005: p. 24).

As development became more integrated with a gender perspective, other state-level phenomena also began to be seen through a gender lens, with important results. Analysis of the effects of women's education was a natural next step. For example, the World Bank has concluded that low investment in female education reduces a country's overall output. One study found that if South Asia, Sub-Saharan Africa and the Middle East, and North Africa had invested earlier in closing gender gaps in education, their income per capita would have increased by 0.5–0.9 from 1960 to 1992 (World Bank, 2001). A study of 90 nation-states concluded that a 1% increase in female to male primary enrollment rates increases GDP growth by 0.012% (Eteve-Volart, 2000). In addition, the more education a mother has, the lower her children's mortality rates, even after controlling for household income and socioeconomic status (Jejeebhoy, 1995; Schultz, 1993). A study from the Philippines concluded that a mother's education contributed more to her children's health status than household income (Barrera, 1990), and another cross-national study of 63 nation-states determined that women's education was the single most important factor in levels of malnutrition over a 25-year period (Smith and Haddad, 2000).

The world is now beginning to recognize that the status of women often substantially influences important aspects of the states in which they live (Hunt, 2007). Many now readily acknowledge the empirically established relationship between the situation of women, and state welfare and even state prosperity. An emerging empirical literature is able to

demonstrate the linkage between the security of women and the security of the states in which they live. We are fast approaching the day when it will be impossible to credibly comment on national and international security affairs without a knowledge of the situation and status of women.

See also: Boserup, Ester (1910–99); Conflict: Anthropological Aspects; Data Bases and Statistical Systems: Security and Conflict; Feminism in Anthropology; Gender History; Gender: Anthropological Aspects; National Security and Domestic Terrorism: The US Case; Violence in Anthropology.

Bibliography

AFP, 12 March 2010. Subjugation of Women Is Threat to US Security: Clinton. Agence France-Presse. www.google.com/hostednews/afp/article/ALeqM5ib71t2EuWmwDahxNKDaHxcEbfcYA (accessed 26.12.12.).

Annan, Kofi, 30 August 2006. No Progress More Effective Than Empowerment of Women. http://www.un.org/News/Press/docs/2006/sgsm10370.doc.htm (accessed 26.12.12.).

Arab Human Development Report, 2005. Towards the Rise of Women in the Arab World. UNDP. http://www.palestineremembered.com/download/UNDP/EnglishVersion/Ar-Human-Dev-2005.pdf (accessed 26.12.12.).

Barrera, A., 1990. The role of maternal schooling and its interaction with public health programs in child health production. Journal of Development Economics 32, 69–91.

Betzig, Laura, 1986. Despotism and Differential Reproduction: A Darwinian View of History. Aldine de Gruyter, New York.

Boyer, Mark A., Urlacher, Brian, Hudson, Natalie B., Niv-Solomon, Anat, Janik, Laura, Butler, Michael J., Brown, Scott W., Ioannou, Andri, 2009. Gender and negotiations: some experimental findings from an international negotiation simulation. International Studies Quarterly 53, 23–47.

Caprioli, Mary, 2000. Gendered conflict. Journal of Peace Research 37, 51–68.

Caprioli, Mary, 2003. Gender equality and state aggression: the impact of domestic gender equality on state first use of force. International Interactions 29 (3), 195–214.

Caprioli, Mary, 2004. Democracy and human rights versus women's security: a contradiction? Security Dialogue: Special Issue Gender and Security 35 (4), 411–428.

Caprioli, Mary, 2005. Primed for violence: the role of gender inequality in predicting internal conflict. International Studies Quarterly 49, 161–178.

Caprioli, Mary, Boyer, Mark A., 2001. Gender, violence, and international crisis. Journal of Conflict Resolution 45, 503–518.

Caprioli, Mary, Trumbore, Peter F., 2003a. Hierarchies of dominance: identifying rogue states and testing their interstate conflict behavior. European Journal of International Relations 9 (3), 377–406.

Caprioli, Mary, Trumbore, Peter F., 2003b. Ethnic discrimination and interstate violence: testing the international impact of domestic behavior. Journal of Peace Research 40 (1), 5–23.

Caprioli, Mary, Trumbore, Peter F., 2006. Human rights rogues in interstate disputes, 1980–2001. Journal of Peace Research 43 (2), 131–148.

Caprioli, Mary, Hudson, Valerie M., McDermott, Rose, Ballif-Spanvill, Bonnie, Emmett, Chad F., Stearmer, S. Matthew, November 2009. The WomanStats project database: advancing an empirical research agenda. Journal of Peace Research 46 (6), 1–13.

Caprioli, Mary, Douglass, Kimberly Lynn, 2008. Nation building and women: the effect of intervention on women's agency. Foreign Policy Analysis 4 (1), 45–65.

Chen, Martha, 1992. A matter of survival: women's right to work in India and Bangladesh. In: Nussbaum, Martha, Glover, Jonathon (Eds.), Women and Culture and Development: A Study of Human Capabilities. Clarendon Press, Oxford.

Chowdry, Geeta, Nair, Sheila, 2002. Power, Postcolonialism and International Relations: Reading Race, Gender and Class. Routledge Press, New York.

Eagly, Alice H., Johannesen-Echmidt, Mary C., van Engen, Marloes L., 2003. Transformational, transactional, and laissez-faire leadership styles: a meta-analysis comparing women and men. Psychological Bulletin 129 (4), 569–591. http://psycnet.apa.org/index.cfm?fa=search.displayRecord. http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=2003-06077-007 (accessed 26.12.12.).

Eichenberg, Richard, 2003. Gender differences in public attitudes toward the use of force by the United States, 1990–2003. International Security 28 (1), 110–141 (Summer).

Elshtain, Jean Bethke, 1987. Women and War. Basic Books, New York.

Eteve-Volart, Berta, 2000. Sex Discrimination and Growth. IMF Working Paper WP/00/84. International Monetary Fund, African Department, Washington, DC.

Finlay, Barbara, Love, Gayle, 2006. Gender differences in reasoning about military intervention. Psychology of Women Quarterly 22 (3), 481–485.

Fish, Steven M., October 2002. Islam and authoritarianism. World Politics 55 (1), 4–37.

Florea, Natalie B., Boyer, Mark A., Brown, S.W., Butler, M.J., Hernandez, M., Weir, K., Meng, L., Johnson, P.R., Mayall, H.J., Lima, C., 2003. Negotiating from Mars to Venus. Simulation and Gaming 34, 226–248.

Gilligan, Carol, 1982. In a Different Voice: Psychological Theory and Women's Development. Harvard University Press, Cambridge, MA.

Handrahan, Lori, 2004. Conflict, gender, ethnicity and post-conflict reconstruction. Security Dialogue 35, 429–445.

Hoddinott, J., Haddad, L., 1995. Does female income share influence household expenditure patterns? Oxford Bulletin of Economics and Statistics 57 (1), 77–97.

Hudson, Valerie M., Boer, Andrea Den, 2004. Bare Branches: The Security Implications of Asia's Surplus Male Population. MIT Press, Cambridge, MA.

Hudson, Valerie M., Ballif-Spanvill, Bonnie, Caprioli, Mary, Emmett, Chad F., 2012a. Sex and World Peace. Columbia University Press, New York.

Hudson, Valerie M., Bowen, Donna Lee, Nielsen, Perpetua Lynne, 2012b. What is the relationship between inequity in family law and violence against women? Approaching the issue of legal enclaves. Politics and Gender 7 (4), 453–492 (Winter).

Hudson, Valerie M., Donna Lee Bowen, Perpetua Lynne Nielsen. Marriage Law and Custom in the Creation of Clan Governance, forthcoming.

Hudson, Valerie M., Caprioli, Mary, Ballif-Spanvill, Bonnie, McDermott, Rose, Emmett, Chad F., 2008/2009. The heart of the matter: the security of women and the security of states. International Security 33 (3), 7–45.

Hunt, Swanee, 2007. Let women rule. Foreign Affairs 86 (3), 109–120.

Inter-Parliamentary Union, 2000. Politics: Women's Insight. http://www.ipu.org/iss-e/women.htm (accessed 26.12.12.).

Jacobson, Jodi L., et al., 1992. Gender Bias: Roadblock to Sustainable Development. Worldwatch Institute, Washington, DC.

Jejeebhoy, Shireen J., 1995. Women's Education, Autonomy, and Reproductive Behavior: Experience from Developing Countries. Clarendon Press, Oxford.

Johnson, Dominic, McDermott, Rose, Barrett, Emily, Cowden, Jonathan, Wrangham, Richard, McIntyre, Matthew, Rosen, Stephen, 2006. Overconfidence in wargames: experimental evidence on expectations, aggression, gender, and testosterone. Proceedings of the Royal Society (Biology) 273, 2513–2520.

Kaufmann, D., 1998. "Challenges in the Next Stage of Corruption," in New Perspectives in Combating Corruption. Transparency International and the World Bank, Washington, DC.

King, Elizabeth M., Mason, Andrew D., 2001. Engendering Development. World Bank, Washington, DC.

King, Wesley C., Miles, Edward W., Jane, Knista, 1991. Boys will be boys (and girls will be girls): the attribution of gender roles stereotypes in a gaming situation. Sex Roles 15, 607–623.

Klasen, Stephan, 1999. Does Gender Inequality Reduce Growth and Development? Evidence from Cross-Country Regressions. World Bank, Washington, DC. http://siteresources.worldbank.org/INTGENDER/Resources/wp7.pdf (accessed 26.12.12.).

Maoz, Zeev, Russett, Bruce, September 1993. Normative and structural causes of democratic peace, 1946–1986. American Political Science Review 87, 624–638.

Marshall, Monty G., Ramsey, Donna, 1999. Gender Empowerment and the Willingness of States to Use Force. Unpublished research paper. Center for Systemic Peace. Available at: http://www.systemicpeace.org/CSPpaper2.pdf (accessed 26.12.12.).

McDermott, Rose, Cowden, Jonathan, 2002. The effects of uncertainty and sex in a crisis simulation game. International Interactions 27 (4), 353–380.

Melander, Erik, 2005. Gender equality and interstate armed conflict. International Studies Quarterly 49 (4), 695–714.

Pruitt, Dean G., Carnevale, Peter, J.D., Forcey, Blythe, van Slyck, Michael, 1986. Gender effects in negotiation: constituent surveillance and contentious behavior. Journal of Experimental Social Psychology 22, 264–275.

Rathgeber, Eva M., 1990. WID, WAD, GAD: trends in research and practice. The Journal of Developing Areas 24 (4), 489–502.

Ray, James Lee, 1995. Democracy and International Conflict: An Evaluation of the Democratic Peace Proposition. University of South Carolina Press, Columbia.

Regan, Patrick M., Paskeviciute, Aida, 2003. Women's access to politics and peaceful states. Journal of Peace Research 40, 287–302.

Schultz, T.P., 1993. Investments in the schooling and health of women and men: quantities and returns. Journal of Human Resources 28 (4), 694–725.

Sen, Amartya, 1989. Women's survival as a development problem. Bulletin of the American Academy of Arts and Sciences 43, 14–29.

Smith, Lisa, Haddad, Lawrence, 2000. Overcoming Child Malnutrition in Developing Countries, 2020 Vision Briefs 64, Food Policy Research Institute. http://ideas.repec.org/p/fpr/2020dp/30.html (accessed 26.12.12.).

Sobek, David, Abouharb, M. Rodwan, Ingram, Christopher G., 2006. The human rights peace: how the respect for human rights at home leads to peace abroad. The Journal of Politics 68 (3), 519–529.

Swers, Michele, 2002. The Difference Women Make: The Policy Impact of Women in Congress. University of Chicago Press, Chicago.

Tessler, Mark, Warriner, Ina, 1997. Gender, feminism, and attitudes toward international conflict. World Politics 49, 250–281.

Thomas, D., 1990. Intrahousehold resource allocation: an inferential approach. Journal of Human Resources 25, 635–664.

Thomas, D., Contreras, D., Frankenberg, E., 1997. Child Health and the Distribution of Household Resources at Marriage. RAND and UCLA, California.

UN Department of Economic and Social Affairs, 2005. Statistics and Indicators on Women and Men. http://unstats.un.org/unsd/demographic/products/indwm (accessed 26.12.12.).

World Bank, 2001. Engendering Development: Through Gender Equality in Rights, Resources, and Voice. World Bank Policy Research Report. Oxford University Press, Washington, DC.

World Economic Forum, 2007. The Global Gender Gap Report. http://www.weforum.org/pdf/gendergap/report2007.pdf (accessed 26.12.12.).

Yunus, Muhammad, Jolis, Alan, 2003. Banker to the Poor: Micro-Lending and the Battle against World Poverty. PublicAffairs, New York.

Yuval–Davis, Nira, 1997. Gender and Nation. Sage Publications, London.

Word Order

Jan Rijkhoff, Aarhus University, Aarhus C, Denmark

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by A. Siewierska, volume 24, pp. 16552–16555, © 2001, Elsevier Ltd.

Abstract

The way constituents are ordered in a linguistic expression is determined by general principles and language-specific rules. This article is mostly concerned with general ordering principles and the three main linguistic categories that are relevant for constituent order research: formal, functional, and semantic categories. The general principles appear to be motivated by cognitive considerations, which are deemed to facilitate language processing and which can all be regarded as manifestations of iconicity: nonarbitrary relations between the form and the content of a linguistic expression. Three major iconic ordering principles are the principles of Domain Integrity, Scope, and Head Proximity.

Introduction

The way words or groups of words are ordered in any particular language is always meaningful, so that when a language allows for alternative ordering patterns, each variant should be regarded as communicatively significant. For instance, the English sentence 'The woman called the doctor' does not mean the same as 'The doctor called the woman.' To give another example, in Spanish one can find both *mi amigo viejo* (lit. 'my friend old') and *mi viejo amigo* (lit. 'my old friend'), but *mi amigo viejo* (with the noun preceding the adjective) means 'my elderly friend' ('my friend who is old'), whereas *mi viejo amigo* (with the adjective preceding the noun) expresses emotional content: 'my old/dear/longtime friend.' Even in so-called 'free word order' languages, it is not the case that each possible order is fully grammatical or acceptable under all circumstances (Dik, 1997, pp. 18, 394; Heine and König, 2010, p. 87).

Speech is inherently linear as we can only pronounce one word at a time, but from a crosslinguistic perspective it appears that some ordering patterns are selected on a language-specific basis to structure discourse, whereas in other cases the linear organization of speech appears to be motivated or constrained by general cognitive considerations (e.g., for communicative purposes or to facilitate language processing), which would explain why we find certain ordering tendencies in languages across the globe. Thus, whether the object argument precedes the verb (as in Turkish) or follows it (as in English) in the unmarked declarative sentence of a language is probably not due to any general ordering preference, as in this particular case there is no obvious reason why one order should be preferred over the other in human communication. Furthermore, Object–Verb order and Verb–Object order are more or less evenly distributed in languages across the globe (which might also be due to chance, of course).

On the other hand, there are several reasons why the subject argument commonly precedes the object argument in the languages of the world: 'Subject before Object' mirrors the fact that language users prefer to process GIVEN before NEW information, i.e., WHAT IS BEING TALKED ABOUT (Topic or Theme) before WHAT IS BEING SAID ABOUT IT (Focus or Rheme). The insight that constituents which help the hearer to interpret the rest of the clause normally appear at the beginning is commonly attributed to the Prague School of Linguistics in the first half of the twentieth century, which divided the communicative structure of the sentence into a thematic and a rhematic part, but statements to the same effect can already be found in Weil (1844, p. 19).

As may be clear from these introductory paragraphs, the current article discusses word order from a functional–typological perspective, which means that, in addition to formal aspects, due attention is also paid to semantic and functional (pragmatic) factors.

Before we continue, we should mention two issues. First of all, since word order studies are concerned with the order of INDIVIDUAL WORDS as well as GROUPS OF WORDS that function as a single unit in a larger linguistic expression, it would be more appropriate to use the term 'constituent.' For this reason 'constituent order' will be the preferred term in the remaining part of this article.

Secondly, the current contribution is restricted to constituents that are properly contained in a linguistic unit or domain (section 'General Tendencies in Constituent Ordering') like a clause or a noun phrase (NP). Not all studies on constituent order indicate whether the constituents they discuss are actually an integral part of a clause or an NP (rather than e.g., an afterthought or an apposition), but failing to make this distinction gives an unreliable picture of the true ordering possibilities within a phrase or clause. For example, a complete clause often contains one or more extraclausal constituents, as in (with extraclausal constituents in italics):

1. *Well*, what do you think, *Michelle*?
2. *By the way*, I shouldn't have said that.

In the case of phrases, a distinction is often made between 'tight' and 'loose constituency.' An extreme case of 'loose constituency' is attested in Kalkatungu (Australia). In this language "there are in fact no NPs, but [...] where an argument is represented by more than one word we have nominals in parallel or in apposition. [...] Each word is a constituent of the clause [...]" (Blake, 1983, p. 145).

In other languages only postposed elements are regarded as loosely connected or 'appositional' elements. For example, in the Caucasian language Georgian a postposed modifier behaves more like a separate NP, or apposition, in that it

inflects for the full set of cases and numbers (Testelec, 1997, p. 652). This is illustrated in (3b).

Georgian

3 a. *am* *or-Ø* *lamaz-Ø* *kal-s*
that:OBL two-DT nice-DT woman-DAT
'to those two nice women'

b. *kal-eb-s* *lamaz-eb-s*
woman-PL-DAT nice-PL-DAT
'to the nice women'

Thus, constituents such as 'well' and 'Michelle' in (1), 'by the way' in (2), and Georgian *lamaz-eb-s* (nice-PL-DAT) in (3b) will not be taken into account.

In addition to offering a brief outline of main developments in the history of constituent order research (section 'Traditional Constituent Order Typology'), this article discusses some problems and recent developments in this area, in particular the way in which constituents are identified and compared across languages. Since certain semantic categories (like basic color terms) or form-based categories (word classes such as adjectives or prepositions) are not attested in all languages, and since the same constituent may have different discourse functions (and conversely, since the same function can be fulfilled by constituents belonging to different form classes), we need semantic, formal as well as functional categories to account for constituent order phenomena in all the world's languages (section 'Form and Function in Constituent Order'). Constituent order studies have been mainly concerned with the relative order of (1) subject, object, and verb in the clause and (2) certain adnominal modifiers (typically adjectives, numerals, and demonstratives). The present survey is somewhat biased toward the NP, as constituent order in this phrase has been investigated in considerable detail in some large crosslinguistic studies (Hawkins, 1983; Dryer, 1992; Rijkhoff, 2004; Cinque, 2005).

Investigating Constituent Order across Languages: Some General Issues

Constituent order is not equally relevant for all languages. The study of constituent order presupposes that there are constituents that appear in certain linear patterns, but what is expressed as a free constituent in one language can be expressed as a morphologically bound form in another languages. This is particularly true in the case of incorporating or polysynthetic languages, as illustrated by this example from the Australian language Tiwi (Blake, 2006, p. 591):

Tiwi

4. *Pitiwuliyondjirrurlimpirrani*
Pi-ti-wuliyondji-rrurlimpirr-ani.
3 PL-3 Sg.fem-dead.wallaby-carry.on.shoulders-PT.HABIT
'They would carry the dead wallaby on their shoulders.'

It is also possible to find polysynthesis in phrases. Since in Nivkh (Siberia, Russia) "dependent-head synthesis is obligatory with nominal heads" (Mattissen, 2002, p. 152), demonstratives, numerals, and adjectives all appear as bound forms, as in (Mattissen, 2002, p. 156):

Nivkh

5. *t'-vila-Gan*
2Sg-big-dog
'your big dog'

A crosslinguistic comparison of constituent order can also be affected by the fact that languages do not necessarily share the same set of categories. For example, the Amazonian language Pirahã is deemed to have no color terms and adnominal demonstratives are absent in Hixkaryana (another Amazonian language from Brazil). Furthermore it appears that true numerals are not attested in all languages and that in some languages the expression for a cardinal number must be accompanied by a gesture; in fact in some languages (e.g., the Papuan language Kobon) the gesture alone can suffice.

Then there is the problem that the translational equivalents of some form or construction can be too varied for crosslinguistic comparison (section 'The Employment of Semantic Criteria in Constituent Order Research'). For example, a distinct class of true adjectives is lacking in many of the world's languages, in particular outside Europe. Such languages typically employ phrasal or clausal structures headed by a stative verb ('be.pretty$_V$,' 'be.silent$_V$') or an abstract noun ('prettiness$_N$,' 'silence$_N$') to express adjectival notions in the NP, as shown in these examples from Kiribati (Kiribati, Micronesia) and Makwe (Palme, Mozambique).

Kiribati (Ross, 1998, p. 90)

6 a. *E* *tikiraoi.*
3Sg.S be.pretty
'She/He is pretty.'

b. *te* uee *ae* *e* *tikiraoi*
ART flower REL 3Sg.S be.pretty
'a pretty flower'

Makwe (Devos, 2008, p. 136)

7. *muú-nu* *w-á=ki-búúli*
NC1-person PP1-GEN=NC7-silence
'a silent person' (lit. 'person of silence')

Similar things can be said about the various ways numeral modifiers are expressed across the globe, as in quite a few languages cardinal numerals are regarded as morphological verbs or nouns. Notice also that many African languages (e.g., Igbo, Hausa) only have a minor, closed class of nonderived adjectives.

Another reason why certain constituent orders may not be found in all the world's languages has to do with the fact that speakers of a language fail to realize certain ordering sequences in normal discourse, even though these patterns are deemed grammatically possible. According to Lyle Campbell (personal communication), there would be nothing ungrammatical about an NP containing a demonstrative, a numeral, and an

adjective in Pipil (El Salvador), in which case the order would be [dem num A N], but apparently this never happens in actual discourse. Some languages have structural constraints that prevent certain constituents from appearing simultaneously in the same phrase or clause. Thus, a lexical NP can accommodate only one modifier in Yimas (New Guinea). Any other modifying material must take an agreement suffix and occur as an appositional phrase (Foley, 1991, pp. 184–188).

When such factors (polysynthesis, apposition, co-occurrence restrictions, etc.) are taken into consideration, quite a few languages are excluded from a crosslinguistic investigation of constituent order. For example, because of the issues mentioned above (but also because of e.g., 'dependency reversal'; cf Malchukov, 2000), the relative order of demonstrative, numeral, adjective, and head noun could only be established for less than half of the languages in a representative sample of the world's languages (Rijkhoff, 2004, p. 327–329).

Traditional Constituent Order Typology

Especially since Greenberg (1963), constituent order typology has mainly been concerned with two issues:

1. classification of languages on the basis of the relative order of the verb (V) and its arguments subject S and object O in a transitive clause (which gives six logically possible patterns: SOV, SVO, VSO, OSV, OVS, VOS – see the section Classification of Languages in Terms of 'Basic Constituent Order');
2. establishing correlations between ordering patterns (e.g., between the order of verb and object on the one hand and between noun and relative clause on the other; see below on Constituent Order Correlations).

A further section is concerned with some problems in Traditional Constituent Order Typology.

Classification of Languages in Terms of 'Basic Constituent Order'

In traditional constituent order typology, it was initially assumed that V, S, and O are labels for categories which are relevant for all languages and that each language uses one of the six logically possible ordering patters of V, S, and O as the 'basic word order.' Based on a large-scale investigation, Tomlin (1986) proposed that some 45% of the world's languages have SOV as the basic order and that in 42% SVO and in 9% VSO is the basic order (for discussion, see Dryer, 2005; Song, 2012). Only a very small percentage of the world's languages have Object before Subject as the basic order: 3% of the languages are deemed to employ VOS order and 1% OVS order. There were no OSV languages in Tomlin's sample, but the Amazonian language Nadëb, for instance, has since been claimed to have OSV as the unmarked order (Weir, 1994, p. 292).

In spite of these results, the classification of languages in terms of the basic order of V, S, and O can be problematic, especially from a crosslinguistic perspective (Song, 2012, pp. 14–15). For example, the notion 'subject' is deemed irrelevant for 'topic prominent' languages, languages without an active–passive opposition and languages with an ergative–absolutive alignment system (Dik, 1997, pp. 247–289). It has also been claimed that there are languages without a distinct class of verbs (Rijkhoff and van Lier, 2013).

Furthermore, 'basic order' is commonly defined as "the order that occurs in stylistically neutral, independent, indicative clauses with full noun phrase (NP) participants, where the subject is definite, agentive and human, the object is a definite semantic patient, and the verb represents an action, not a state or an event" (Siewierska, 1988, p. 8). However, clauses in which both arguments of a transitive verb are expressed as definite lexical phrases (as in '*The dog chased the cat*') are very infrequent in natural language, especially in the spoken variety (Siewierska, 1988, pp. 8–14).

Constituent Order Correlations

The first serious investigation of correlations between constituent order patterns is undoubtedly Greenberg's (1963) seminal crosslinguistic survey, which led to the formulation of implicational universals. Greenberg classified languages on the basis of the unmarked order of nominal subject (S), verb (V), and nominal object (O), which resulted in three groups: languages with basic order VSO, SVO, or SOV. This is shown in **Table 1**, which is based on Hawkins's (1983) more recent study. Notice that here 'V-1' ('Verb-initial') is used, rather than 'VSO' as in Greenberg's original publication (A = adjective, G = genitive/possessor phrase).

Table 1 shows, among other things, that most SOV languages have both the adjective and the genitive preceding the noun (type 11), whereas most V-1 and SVO languages have these modifiers following the noun (types 1 and 5).

Greenberg (1963) also formulated 45 'universals of grammar with particular reference to the order of meaningful elements'; for example:

Universal 16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb.

Universal 17. With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.

Universal 18. When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise.

In the same article Greenberg (1963, pp. 96–104) offered tentative explanations for some of his findings, suggesting that several ordering sequences could be accounted for in terms of two competing motivations: Dominance and Harmony. For example, a language that has the head-initial pairs VS–VO–NA–NG (where V and N are the head of the

Table 1 Classification of languages (based on Hawkins, 1983; '—' signals that both A and G precede or follow the head noun N)

1. V-1 & N—	38	5. SVO & N—	56	9. SOV & N—	21
2. V-1 & ANG	13	6. SVO & ANG	17	10. SOV & ANG	0
3. V-1 & —N	2	7. SVO & —N	19	11. SOV & —N	98
4. V-1 & GNA	0	8. SVO & GNA	17	12. SOV & GNA	55
	53 lgs.		109 lgs.		174 lgs.

clause and the NP, respectively) is 'harmonic' and the same goes for languages with the head-final pairs SV–OV–AN–GN. The notion 'dominance' can be exemplified with Hawkins's Universal VI (see (9)), as summarized in Table 2 (Rel = relative clause). Since GN order occurs with both NRel and RelN order, GN is regarded as the dominant order (Tigre, mainly spoken in Eritrea, seems to be an exception with RelN and NG order). Greenberg added, however, that his theory was far from complete and emphasized that disharmonic and nondominant (or: 'recessive') patterns should not be ignored.

Subsequently Lehmann (1973) and Vennemann (1975) offered more radical accounts for Greenberg's findings, essentially reducing his three-way typology (VSO, SVO, SOV) to two 'word order types' (see Song, 2012, pp. 17–21, 29–54) for detailed discussion of two-way ordering typologies). Lehmann's *Fundamental Principle of Placement* suggested it is possible to predict certain ordering pairs (such as the order of adjective and noun), if one knows that the language has OV or VO as the basic order. According to Vennemann's *Principle of Natural Serialization* constituents are either Operators or Operands, which tend to be serialized either with Operator before Operand or with Operand before Operator. Examples of Operators would be 'object' and 'adverbial,' both of which have 'verb' as their Operand. Even though there are many languages with ordering patterns that deviate from Lehmann's or Vennemann's ideal types, such languages were largely ignored as they were deemed to be in the process of changing from one type to another due to internal development or contact.

Greenberg's original tripartite typology (Table 1) was restored in Hawkins's (1983) monograph on constituent order universals. Using a sample containing over 300 languages, Hawkins formulated some new, often exceptionless universals (but see Dryer (1992) and Rijkhoff (2004, p. 227) for counterexamples). For instance (Hawkins, 1983, pp. 64, 83):

8. If a language has OV order, then if the adjective precedes the noun, the genitive precedes the noun; i.e., OV ⊃ (AN ⊃ GN).
9. If a language has noun before genitive, then it has noun before relative clause; i.e., NG ⊃ NRel (equivalently: RelN ⊃ GN).

He then demonstrated that various implicational universals could be collapsed into statistical implications with a preposition or postposition as ultimate antecedent (Hawkins, 1983, pp. 75, 83):

10. Prepositional Noun Modifier Hierarchy:
    Prep ⊃ ((NDem ∪ NNum ⊃ NA) & (NA ⊃ NG) & (NG ⊃ NRel))

Table 2 Schematic representation of Universal VI

	NG	*GN*
NRe1	+	+
Re1N	–	+

(cf. Hawkins, 1983, p. 83).

11. Postpositional Noun Modifier Hierarchy:
    Postp ⊃ ((AN ∪ RelN ⊃ DemN & NumN) & (DemN ∪ NumN ⊃ GN))

He also attempted to account in a more principled way for the many languages with ordering patterns that do not quite fit the 'ideal' two-way classification, which had been the focus of Lehmann's and Vennemann's proposals. For example, the *Heaviness Serialization Principle* (Hawkins, 1983, pp. 90–91) deals with the fact that many languages have adnominal modifiers on both sides of the head noun (cf. also the *Mobility Principle* – Hawkins, 1983, pp. 93–97).

Almost a decade later Dryer (1992) used an even bigger sample (625 languages) in an attempt to investigate which constituent order pairs actually correlate with the order of object and verb (OV/VO), because assertions about such correlations had never been substantiated by systematic research (but cf Perkins, 2001, p. 432). Dryer concluded there are indeed several constituent order pairs that more or less correlate with the OV/VO distinction (Table 3), among them the pair HEAD NOUN–POSSESSOR NP (NG/GN) and the pair HEAD NOUN–RELATIVE CLAUSE (NRel/RelN). However, he did not find evidence for a correlation between OV/VO order and, for instance, the order of noun and adjective (AN/NA) or the order of noun and demonstrative (demN/Ndem).

Thus, we tend to find NG and NRel order in VO languages and GN and RelN order in OV languages. To account for the correlation pairs, Dryer proposed the *Branching Direction Theory*, according to which there is tendency for phrasal categories (such as G and Rel) to either precede or follow a member of a nonphrasal category (such as N, A, and other 'verb patterners'). The theory makes no ordering predictions about nonbranching categories like demonstratives or numerals. However, the theory failed to account for some correlation pairs, such as the order of verb and manner adverb (both of which are nonphrasal). Another problem of the *Branching Direction Theory* concerns the categories that are used to make the crosslinguistic comparison (see Rijkhoff (2004, pp. 278–291; Rijkhoff, 2009) and Song (2012, pp. 29–54 for discussion). This is discussed in sections 'The Employment of Semantic Criteria in Constituent Order Research' and 'Categorial Confusion: Mixing 'Semantic' and Formal Categories.'

Some Problems in Traditional Constituent Order Typology

From a crosslinguistic or typological perspective, the notion 'basic constituent order' is problematic for various reasons,

Table 3 Some of Dryer's correlation pairs

Verb patterners	*Object patterners*	*Example*
Verb	Manner adverb	*Ran + slowly*
Adposition	Noun phrase [NP]	*on + the table*
Copula verb	Predicate	*is + a teacher*
'want'	Verb phrase [VP]	*wants + to see Mary*
Noun [N]	Genitive [G]	*father + of John*
Noun [N]	relative clause [Rel]	*movies + that we saw*
Adjective [A]	Standard of comparison	*taller + than Bob*

Dryer, 1992, p. 108; see also Dryer, 2009.

casting some doubt on the assumption that all languages have a basic order. Here we will focus on two problems: (i) the establishment of a basic constituent order (section The Notion 'Basic Constituent Order') and (ii) the employment of semantic categories to identify constituents for comparative (i.e., crosslinguistic) syntactic purposes (section 'The Employment of Semantic Criteria in Constituent Order Research'). The section 'Categorial Confusion: Mixing 'Semantic' and Formal Categories' deals with the problem of categorial confusion and 'The Relevance Of Meaning- And Form-Based Categories For Constituent Order Research' shows that both meaning-based and form-based categories are relevant in the area of constituent order research, provided they are kept separate from each other.

The Notion 'Basic Constituent Order'

Even if it would always be possible to find for each language a representative number of stylistically neutral (active, positive), independent, indicative clauses that contain a transitive verb and lexical subject and object arguments (section 'Investigating Constituent Order across Languages: Some General Issues'), there would still be quite a few languages that, statistically speaking, display no clear preference for just one of the six possible orderings of V, S, and O. To give some examples: in Dutch, where the finite (main or auxiliary/Aux) verb normally appears after the subject in the main clause (SV_{finite} …), the basic order would be SVO in the absence of an auxiliary verb (*De kat ving the vogel* 'The cat caught the bird'), but SAuxOV if the clause contains a finite auxiliary verb, as in: *De kat heeft*$_{Aux}$ *de vogel gevangen*$_{V}$ (lit. 'The cat has the bird caught') 'The cat has caught the bird.' Notice also that in subordinate clauses the whole verbal cluster appears after S and O (SOAuxV or SOVAux), as in *Ik hoorde dat de kat de vogel heeft*$_{Aux}$ *gevangen*$_{V}$/ *gevangen*$_{V}$ *heeft*$_{Aux}$ (lit. 'I heard that the cat the bird has caught/caught has').

To give another example, Ket (a language isolate spoken in Siberia) is typically classified as having two 'basic' orders (SVO/SOV). Furthermore, since this language makes considerable use of incorporation, many transitive verbs appear without lexical subject or object arguments. Constituent order is also claimed to be remarkably free in Yimas, "even in comparison to the relative freedom exhibited by other Papuan languages. Within the clause, there is no set word order pattern at all (in fact, the majority of clauses consist of just a verb, so word order is not much of a question in these cases anyway). […] Word order in Yimas is so free that NPs need not even form a constituent" (Foley, 1991, p. 4).

These remarks also highlight another problem touched upon earlier: that it is rather unusual to find a transitive verb with two full (i.e., lexical) subject and object arguments where the subject is definite, agentive, and human, the object is a definite semantic patient (Siewierska, 1988, p. 8), particularly in normal spoken discourse. As was already mentioned in the section 'Investigating Constituent Order across Languages: Some General Issues,' in polysynthetic languages (but also in other languages) the arguments of a transitive verb are often not expressed as free constituents, but by clitics or affixes, if at all. These problems could indicate that in quite a few languages basic constituent order has been established on the basis of an elicitation procedure, i.e., at the explicit request of the linguistic field worker who provides the native speaker informant with a constructed example, which is, of course, far from ideal.

The Employment of Semantic Criteria in Constituent Order Research

In a crosslinguistic investigation of constituent order patterns, it is important that the constituents involved are sufficiently similar to be identified and compared across languages (Song, 2012, pp. 10–15). Due the enormous degree of morphological and syntactic diversity that is attested in the world's languages, form-based definitions of constituents have generally been avoided in constituent order studies. For example, to say that a noun is a word that is inflected for number is quite irrelevant for the many languages across the globe in which number marking is absent.

Because formal categories are on the whole too language specific, typologists have usually followed Greenberg (1963, p. 74), who applied semantic criteria to identify the various kinds of constituents in his crosslinguistic investigation of ordering patterns. Thus, his semantic category Adjective essentially covers all the forms and constructions that translate as an adjective in English, even if a language does not have a dedicated class of adjectives, but instead uses abstract nouns such as 'bigness$_{N}$' (as in certain Chadic and Bantu languages), or stative verbs like 'be.big$_{V}$' (as in e.g., many SE Asian languages) to express adjectival notions.

Two decades after Greenberg's publication, Hawkins (1983, p. 12) also claimed that semantic criteria 'suffice to make the cross-linguistic equation' and when Dryer (1992, p. 120) investigated 'Greenbergian word order correlations,' he also identified 'different categories largely on the basis of semantic criteria' (more on this in the section 'Categorial Confusion: Mixing 'Semantic' and Formal Categories'). Thus, their studies suggested that, by employing semantically defined categories, the same constituent ordering patterns (e.g., the order of adjective and noun) can be investigated in all the world's languages, even when it is clear from the grammatical description of a language that the constituents involved belong to different form classes, whose members can have their own, form-based ordering preferences (see the section 'The Relevance of Meaning and Form-Based Categories for Constituent Order Research' and 'Different Constituents, Same Function'). In other words, whereas formally defined categories are usually too narrow for crosslinguistic research into constituent order, semantically defined categories can be too wide: they may include forms or constructions that are too varied to allow for crosslinguistic comparison of constituent ordering patterns. In fact, this was already acknowledged in Greenberg's (1963) investigation of language universals. In his discussion of comparative constructions (as in English '*Mary is TALLER THAN BILL*'), he found that the semantic category Comparison is expressed rather differently in languages across the globe. As later shown by Stassen (1984), there are at least five major construction types to express the semantic category of Comparison. For example, whereas English uses an adjective followed by the standard introduced by the marker *than* (as in the example above: Adjective/*taller* – Marker/*than* – Standard/*Bill*), Duala (a Bantu language from Cameroon) employs two predicates: one

expressing the adjectival notion (here *kolo* 'big'), the other is a verb meaning 'to exceed' or 'to surpass' (Stassen, 1984, p. 151).

Duala

12. *Nin ndabo e kolo buka nine.*
this house it big exceed that
'This house is bigger than that.'

Even though the 'exceed construction' constitutes the second largest type in Stassen's 110-language sample, Greenberg writes that he could only formulate *Universal 22* if he left this construction out of consideration.

"*Universal 22*. If in comparisons of superiority the only order or one of the alternative orders is standard-marker-adjective, then the language is postpositional. With overwhelmingly more than chance frequency, if the only order is adjective-marker-standard, the language is prepositional" (Greenberg, 1966, p. 111).

> A number of languages are not entered in [Table 8 on the same page – JR] because they utilize a verb with a general meaning 'to surpass.' This is particularly common in Africa (e.g., Yoruba): 'X is large, surpasses Y.' Loritja, an Australian language which has 'X is large, Y is small,' is likewise not entered
>
> Greenberg, 1963, p. 88

Two further problems concerning the employment of so-called semantic (or 'conceptual') criteria in traditional constituent order research are, firstly, that the semantic criteria are not explicitly listed and, secondly, that in some cases constituents are essentially identified and compared on the basis of functional rather than semantic criteria. For this reason, the notion 'semantic' as employed in traditional constituent order research will henceforth appear in single quote marks. Notice also that semantic categories are not always distinguished from conceptual categories (especially in cognitive linguistics), but there are good reasons why the two categories should not be conflated. Thus, the reason why adjectives and relative clauses can sometimes be grouped together is not because they belong to the same semantic category (see section 'The Relevance of Meaning and Form-Based Categories for Constituent Order Research' and **Table 4** on genuine semantic features like '±animate' and categories like 'color'), but rather because members of both formal categories can serve as a qualifying modifier in the NP, specifying a more or less inherent property of an entity, such as its age, weight, or size (section 'Form and Function in Constituent Order').

Table 4 The same function (Qualifying Modifier) expressed by members of different formal categories (Adjective, Prepositional Phrase, Relative Clause)

	Functional category: Qualifying Modifier (in the NP)		
Semantic categories	*Formal Categories Adjective*	*Prep. Phrase*	*Relative Clause*
Size	*a big N*	*a N of enormous size*	*a N that is rather big*
Value/Quality	*an expensive N*	*a N of great value*	*a N that is pretty cheap*
Age	*a young N*	*a N under age 16*	*a N that is much too young*
Color	*a black N*	*a N of blackness*	*a N that is pitch black*

Categorial Confusion: Mixing 'Semantic' and Formal Categories

In spite of the problems concerning the application of 'semantic' criteria in traditional constituent order research (section 'The Employment of Semantic Criteria in Constituent Order Research'), these criteria continue to be used in syntactic typology, including large-scale projects such as *The World Atlas of Language Structures* (Rijkhoff, 2009a). Thus, Dryer essentially treats the translational equivalents of English adjectives in all languages as members of the 'semantic' category Adjective, but admits that "there are many languages in which what I call adjectives are really verbs, and 'adjectives' modifying nouns are really just a kind of relative clause" (Dryer, 1992, p. 196, fn. 12). However, Dryer's correlation pairs also include Relative Clause (Rel) as an independent category (**Table 3**). If relative clauses can belong to two 'semantic' categories (Adjective and Relative Clause), one cannot properly interpret the *Branching Direction Theory* (section 'Constituent Order Correlations'), where adjectives and relative clauses are treated as members of distinct, nonoverlapping formal categories. Relative clauses are branching or phrasal categories and 'object patterners'; adjectives are nonbranching categories and the order of Adjective and Noun does not correlate with the order of Object and Verb. Whereas the correlations pairs in **Table 3** are presumably all 'semantic' categories (Dryer, 1992, p. 120), the two categories referred to in the *Branching Direction Theory* (viz. phrasal/branching and nonphrasal/nonbranching categories) are unambiguously syntactic (i.e., formal) categories. In sum, it remains unclear what happened with the phrasal or branching members of the 'semantic' category Adjective in the *Branching Direction Theory*.

The same confusion between 'semantic' and formal categories is found in Hawkins (1983), where Greenbergian 'semantic' categories (which may contain members of different form classes) are used to identify constituents and compare ordering patterns in many different languages (Hawkins, 1983, pp. 9–12). However, some of the same category labels (Adjective, Genitive, Relative Clause) are also used when constituents are discussed as part of the Heaviness Hierarchy, which is based on formal criteria (Hawkins, 1983, pp. 90–91):

$$\text{Rel} \geq_R \text{Gen} \geq_R \text{A} \geq_R \text{Dem/Num}$$

"where $\geq_R$ means: 'exhibits more or equal rightward positioning relative to the head noun across languages.' That is heavier noun modifiers occur to the right."

One constituent is heavier than another, if one or more of the following factors imposes an ordering between them: (1) length and quantity of morphemes, (2) quantity of words, (3) syntactic depth of branching nodes, and (4 inclusion of dominated constituents. The different degrees of 'heaviness' would explain why relative clauses, which are comparatively 'heavy,' are more likely to appear after the head noun in prepositional languages than adjectives, which are relatively 'light.' In other words, names for 'semantic' categories (A, Rel etc.) are now used as labels for categories defined on

the basis of formal criteria. Here, too, it remains unclear how verbal and nominal expressions of adjectival notions (see examples (6) and (7)) are categorized in terms of heaviness. Presumably the original members of the single semantic category Adjective are now distributed over several formal categories (including at least A and Rel).

We will see in the section 'Form and Function in Constituent Order' that the confusion between 'semantic' and formal categories in constituent order studies is at least partly due to two problems: (1) the 'semantic' criteria used in traditional constituent order research are in fact sometimes functional criteria in that constituents are occasionally identified and compared on the basis of the function they serve in a linguistic expression (e.g., QUALIFYING MODIFIER) and (2) the same function (e.g., QUALIFYING MODIFIER) can be fulfilled by constituents that belong to different form classes (e.g., adjective, prepositional phrase, relative clause), and vice versa, that the same constituent (e.g., relative clause) can have different functions in actual discourse.

The Relevance of Meaning- and Form-Based Categories for Constituent Order Research

Both semantic and formal criteria are needed to account for the way constituents are linearized in a linguistic expression, but it is very important that they are kept separate. For example, the semantic features '±definite' or '±animate' (often in tandem with other meaning-based or pragmatic notions like 'prominence,' 'topicality,' or 'salience') can have various effects on the grammar of a language (Dik, 1997, pp. 34–41), including the way constituents are linearized. Thus, in some languages the animacy hierarchy determines the order of the two object phrases in clauses with a ditransitive verb (Heine and König, 2010). The following examples are from Kinnauri (Nepal). In (13a) both objects (child, maternal aunt) refer to animate entities, in which case the direct object precedes the indirect object. In (13b), however, the order of the objects in determined by animacy: the animate entity (uncle) precedes the inanimate entity (letter) in the unmarked order (Kittilä, 2006, p. 2).

Kinnauri

13. a. *amas anu chan-u ane-pən rano*
mother:ERG she:REFL child-ACC maternal aunt:ACC send
'The mother sent her son to the maternal aunt'

b. *gəs än bapu-pən cithi cemigduk*
1Sg:ERG 1Sg:GEN uncle-ACC letter write:S
'I have to write a letter to my uncle'

Alamblak (Papua New Guinea) provides us with another example of the relevance of semantics for constituent order. In this language, color adjectives must precede the noun, whereas adjectives belonging to other semantic categories can occur on either side of the head, although some (e.g., adjectives denoting a physical property) tend to precede the noun and others (e.g., those denoting 'value') usually follow the noun. To give another example, in Maybrat (Bird's Head Peninsula, New Guinea) the genitive precedes the noun (GN) in the case of inalienable possession, but follows it (NG) in the case of alienable possession.

Formal factors clearly play a role when, for example, constituent order is determined by the way a constituent is formally expressed, which in its turn is connected with the internal complexity of a constituent (see above on 'morphosyntactic weight' or 'heaviness'). Whether a constituent takes the form of a single word (adjective, adverb), a phrase (NP, PP). or a clause (relative clause) can (co-) determine its position in a linguistic expression (see the section 'Domain Integrity'). The role of structural complexity was already captured in 1932 by Behaghel's *Gesetz der wachsenden Glieder* ('law of growing members'), according to which constituents are arranged in their order of relative increasing complexity (cf 'heavy shift' in transformational grammar and the Principle of Increasing Complexity in Dik's Functional Grammar). Compare

14. a. ?I gave *the paintings that my uncle left to me as part of his inheritance* to her.
b. I gave to her *the paintings that my uncle left to me as part of his inheritance*.

The version with the long constituent in clause-final position (14b) is preferred over the one with the short constituent ('to her') at the end (14a). The same holds for heavy constituents in the NP: the short *of*-phrase in (15a) can appear immediately after the head noun, but the preferred position of the longer *of*-phrase is at the end of the NP, as in (15b).

15. a. the painting *of him* that she found
b. the painting that she found of *her grandfather wearing that strange jacket*

In sum, it is necessary to distinguish between meaning-based and form-based categories, as both play their own distinct role in structuring discourse through constituent order. For a more complete and satisfactory account of constituent order, however, the discourse function of a constituent in actual speech also needs to be taken into consideration; this is discussed in the section 'Form and Function in Constituent Order.'

Form and Function in Constituent Order

An important reason why 'semantic' and formal categories have been mixed up in some constituent order studies is that these studies did not distinguish between semantics and pragmatics, i.e., between the coded meaning of a constituent and the discourse function of a constituent in an actual utterance (section 'Categorial confusion: mixing 'semantic' and formal categories'). This means that functional categorization is essentially context dependent (see e.g., the introduction about 'Subject before Object'). Furthermore, not all constituent order studies took into account that there is often no direct correspondence between the form and the function of a constituent. The same constituent (e.g., a relative clause) can have different functions in interpersonal communication (section 'Same Constituent, Different Functions'), and vice versa, constituents belonging to different formal categories (e.g., adjectives and relative clauses) can be used in the same function (section 'Different Constituents, Same Function'). In either

case we see that the discourse function of a constituent in actual speech may put certain constraints on its ordering possibilities.

Functional Modifier Categories

The functional, instrumental approach to categorization can be said to have its roots in the Prague School of Linguistics, which has greatly influenced current functional theories of grammar. Functional categorization is not directly concerned with the formal or semantic properties of a constituent, but rather with the actual job of a linguistic form or construction in the process of verbal communication. Here we will only deal with some FUNCTIONAL MODIFIER CATEGORIES that are relevant for both NPs and clauses (for details, see Rijkhoff, 2008a,c). They are briefly characterized in (16a–e), where they are listed according to scope increase. Thus, a CLASSIFYING MODIFIER such as 'presidential' in 'the presidential election' only has the head constituent ('election') in its scope, whereas a DISCOURSE-REFERENTIAL MODIFIER like a definite or indefinite article has all other constituents of an NP in its scope (see also the section 'Different Constituents, Same Function' on the Scope Principle).

16. Five functional modifier categories in NPs and clauses (with examples of English adnominal modifiers)
    a. CLASSIFYING MODIFIERS further specify what KIND of entity (subcategory) is denoted by the head constituent (as in e.g., 'SOLAR energy,' 'a VIRAL infection,' 'a man OF THE STAGE');
    b. QUALIFYING MODIFIERS specify more or less inherent properties ('qualities') of an entity (e.g., 'a BLUE hat,' 'children UNDER AGE 16,' 'a house THAT WAS RATHER BIG');
    c. QUANTIFYING MODIFIERS specify quantitative properties (quantity, number, cardinality) of an entity (e.g., 'TWO cars,' 'SEVERAL books');
    d. LOCALIZING or ANCHORING MODIFIERS specify locative properties of an entity in spatial, temporal, or cognitive dimension. By 'anchoring' an entity in the world of discourse, the speaker makes the referent identifiable for the hearer (e.g., 'THIS car,' 'the house ON THE CORNER,' 'That's the man WHO STOLE MY BIKE');
    e. DISCOURSE-REFERENTIAL MODIFIERS specify discourse-referential (pragmatic) properties of an entity (e.g., 'THE/A car').

Since there is often no one-to-one correspondence between the form and the function of a constituent, the same constituent can have different functions in actual speech. Thus, an adnominal demonstrative is a LOCALIZING MODIFIER in its deictic function, but a DISCOURSE-REFERENTIAL MODIFIER when it merely marks definiteness (section 'Different Constituents, Same Function'). Similarly, a relative clause is a QUALIFYING MODIFIER when it serves to express a more or less inherent property (example (16b), but a LOCALIZING/ANCHORING MODIFIER when it is used to make the referent of an NP identifiable (as in example (16d)). Conversely, constituents that belong to different formal categories (e.g., adjectives and relative clauses) can be used in the same function (e.g., QUALIFYING MODIFIER – see **Table 4**). The many-to-many correspondence between the form and the function of constituents and its relevance for constituent order is discussed in the sections 'Same Constituent, Different Functions' and 'Different Constituents, Same Function.'

Same Constituent, Different Functions

Membership of formal categories such as Adjective, Relative Clause, or Complement Clause is generally not a very reliable indicator of the discourse function of a constituent in actual speech. For example, the adjective 'civil' is a CLASSIFYING MODIFIER in 'civil rights' (implying e.g., that it cannot be modified by a degree adverb: *very civil rights), whereas 'civil' in 'a civil reply' serves as a QUALIFYING MODIFIER (which means it can be modified by a degree adverb: 'a very civil reply') (Halliday, 2004, pp. 320–322; Bisetto, 2010).

There are many examples of constituents whose actual function in an utterance is marked by constituent order. Berbice Dutch Creole is one of the languages in which the same constituent can be used as a definite article or as an adnominal demonstrative modifier, but as an article it precedes the noun and when it has a deictic (localizing) function it follows the noun. When a numeral appears before the noun in Russian (num N) it expresses cardinality (*dvadcat' knig* ' twenty books'), but when the order is reversed (N num) the numeral has an approximative meaning (*knig dvadcat'* 'approximately twenty books'). In Turkish and many other languages the same constituent can occur as an indefinite article or as the numeral 'one' (in Turkish *bir*). The position of *bir* in the NP is determined by its actual modifier function. Similar things can be said about adnominal possessor phrases or genitives and relative clauses. For example, in the Nepalese language Kham a qualifying relative clause appears immediately before the noun, whereas a localizing relative clause normally precedes the numeral and clusters with other localizing modifiers in the order (see the section 'Functional Modifier Categories'): [dem Rel num Rel N].

Different Constituents, Same Function

The fact that the same function can be fulfilled by constituents that belong to different form classes is seen in **Table 4**, which shows that English adjectives, prepositional phrases, and relative clauses can all be used as a descriptive or QUALIFYING MODIFIER, which specifies a more or less inherent property of an entity.

According to the Scope Principle (section Scope), constituents that belong to the same functional modifier category occupy the same position before or after the head constituent. This was shown to be the case in the Nepalese language Kham (section 'Same Constituent, Different Functions'), where qualifying relative clauses (but not localizing relative clauses) occupy the same position as true adjectives in other languages (cf the ordering patterns in (31)).

When a language has alternative forms of expression for the same function (**Table 4**), the choice is often determined by linguistic genre, pragmatic considerations, or internal complexity (section 'The Relevance of Meaning- and Form-Based Categories for Constituent Order Research'). This can be seen in (18), where the 'heavy' postnominal qualifying

relative clause contains coordinated constituents (both examples from the Internet).

17. *I am looking for a* RATHER BIG$_A$ *house* ….
18. *Indeed there was a house* THAT WAS RATHER BIG AND VERY PRETTY$_{REL}$.

General Tendencies in Constituent Ordering

Globally observed constituent ordering tendencies are generally regarded as manifestations of iconicity: nonarbitrary relations between 'form' and 'content' (Haiman, 1985). Iconicity in grammar concerns any formal manifestation of the (largely unconscious) effort to maintain a close correspondence between, on the one hand, nonlinear, multidimensional MENTAL representations of entities (as well as relations between these entities) and, on the other hand, the way these representations are mapped onto a linear, one-dimensional structure in speech. Constituent order is considered iconic whenever it reflects aspects of the conceptual or semantic content, which typically manifests itself as 'iconicity of distance,' where conceptual or semantic distance matches with spatial distance. Especially among functional linguists, it is widely assumed that there is direct causal connection between iconicity in grammar and increased communicative efficiency.

Under the narrow view of iconicity, the notion refers to the tendency of linguistic expressions to mirror the temporal or logical sequential order of events. This can be seen in (19a–b), where the coordinated clauses in the a-sentence (but not in the b-sentence) reflect the actual order of events.

19. a. They met in Konstanz and married in Amsterdam. (unmarked)
    b. ?They married in Amsterdam and met in Konstanz. (marked)

Another instance of iconicity in the narrow sense is the order of clauses in conditional sentences (e.g., 'If he comes, I'll leave'), where the condition (*protasis*) precedes the consequence (*apodosis*) as the unmarked order in all languages (Universal 14 – Greenberg, 1963, p. 84).

In its broad sense, iconicity also refers to any correspondence between linguistic expressions and "generally accepted perceptions of dominance and salience and/or semantic dependencies" (Siewierska, 2006, p. 646). Defined this way, iconicity accounts for several of the general ordering tendencies mentioned earlier, such as the strong universal preference for Subject arguments to precede Object arguments (subjects tend to be agentive, definite, more salient and higher on the person and animacy hierarchy than objects) and even the tendency to place long and complex constituents after relatively short, less complex constituents. The use of informationally and structurally 'light' constituents correlates with topicality and high identifiability: placing light constituents before heavy constituents would iconically reflect the preferred way of processing information.

Behagel's famous first law (1932, p. 4) already captured the insight that constituent order is often an iconic reflection of the way we perceive or conceptualize entities or relations between entities to facilitate language processing.

20. Behaghel's First Law
    Das oberste Gesetz ist dieses, daß das geistig eng Zusammengehörige auch eng zusammengestellt wird ('The principal law is this: that what belongs together mentally is also placed close together').

It is possible to formulate more precise iconic subprinciples on the basis of Behaghel's first law: (1) the Principle of Domain Integrity (PDI), (2) the Principle of Head Proximity (PHP), and (3) the Principle of Scope (PoS) (Rijkhoff, 2004, 2008b). These principles are briefly discussed below, mostly illustrated with examples involving NP-internal constituents, as the effects of the various principles are better seen at the level of the phrase than in the larger clausal domain.

Domain Integrity

The PDI can be formulated as follows:

21. Constituents prefer to remain within the boundaries of their proper domain; embedded domains (subordinate clauses or phrasal modifiers such as adnominal possessive NPs) prefer to be in the periphery of their embedding domain (e.g., main clause, argument NP), so as to avoid possible discontinuity among other constituents of the embedding domain.

The notion 'domain' stands for an integral, hierarchically organized linguistic unit and can be regarded as the functional equivalent of 'constituency' in syntactocentric theories of grammar. The PDI is perhaps the most obvious interpretation of Behaghel's first law and accounts for the fact that discontinuity is a marked phenomenon in languages across the globe. This principle sums up ideas that have been proposed in various forms, namely that constituents belonging to the same conceptual or semantic unit (for instance, because they help to describe or refer to the same object or event) are processed faster and easier when they appear in a continuous, uninterrupted sequence. Some languages, such as the western Austronesian languages, do not allow discontinuous phrases at all, but when a language does allow a constituent to appear outside the boundaries of its proper linguistic domain, this is commonly due to some pragmatic factor (e.g., contrastive focus) or complexity ('heaviness'). For example, in Ngiti (Zaire) highly emphasized numerals may appear in clause-initial position (Kutsch Lojenga, 1994, p. 355).

Ngiti

22. [*árù*]$_{num}$ *ma* *m-ɔ̀kɛ̀rɛ̀* [*itsu*]$_{NP}$
    eight 1.SG SC-cut:PERF.PRES tree
    'I have cut EIGHT trees'

In Dutch 'heavy' adnominal modifiers, such as relative clauses, may appear in clause-final position:

Dutch

23. *Ik heb net* [*de man*]$_{NP}$ *gezien* [*die gisteren mijn fiets gestole heeft*]$_{Rel}$
    I have just the man seen who yesterday my bike stolen has
    'I just saw the man who stole my bike yesterday'

Since appositional constituents are not part of an integral domain, it is less unusual for them to appear discontinuously (see the Introduction). It appears that loose, appositional relationships between constituents that together form a conceptual or semantic unit historically precede the employment of hierarchically organized linguistic domains (Van de Velde, 2009).

Head Proximity

The PHP basically states that languages prefer to have a short distance between lexical heads of hierarchically ordered domains (e.g., between head 'N' of subordinate domain 'NP' and head 'V' of superordinate domain 'clause').

24. Constituent ordering rules conspire to keep the heads of different, hierarchically ordered domains as closely together as possible.

This general ordering principle is based on the processing strategy, which basically holds that domain identification (e.g., of a clause or NP) plays a central role in language processing. One of the most reliable strategies to identify a linguistic domain is to find the constituent that is the head of the domain, because the head is typically (1) the only obligatorily constituent in a domain and (2) the most informative constituent of a domain. In languages that have distinct classes of verbs and nouns, verbs typically serve as the head of a clausal domain and nouns as the head of a phrase (incl. prepositional phrase; notice that adpositions are not regarded as heads in e.g., Functional (Discourse) Grammar).

The PHP could account for several ordering phenomena (Rijkhoff, 2004, pp. 261–312), including some of the constituent order correlations observed in Greenberg (1963) and Dryer (1992), such as (notice that, due to the pragmatic status of Subject, SVO languages are comparable to V-initial languages regarding head proximity; see Rijkhoff (1986, pp. 1–12 for details):

- the tendency to avoid embedded domains (like relative clauses or genitives) appearing between the head of the clause V and the head of the phrase N (cf **Table 1** and Dryer's *Branching Direction Theory*; dots symbolize adnominal modifiers);

25. a. V $[N....]_{NP}$ .... (verb-initial)
    b. .... $[....N]_{NP}$ V (verb-final)

- Greenberg's Universal 16 (see the Section Constituent Order Correlations):

26. a. Aux V .... (verb-initial)
    b. .... V Aux (verb-final)

- Behaghel's (1932) *Gesetz der wachsenden Glieder* ('law of growing members'), according to which long constituents in the clause appear at a greater distance from the head of the clause V than short constituents, which would result in a better overall ratings in terms of 'head proximity.'

27. Aux V $[...]_{NP}$ $[....]_{NP}$ (verb-initial)

Notice, however, that in terms of head proximity the long-before-short order is preferred in verb-final languages:

28. $[....]_{NP}$ $[...]_{NP}$ V Aux (verb-final)

While it is true that some verb-final languages have a preference for 'long-before-short' (Hawkins, 1994, p. 152), which is preferred in terms of head proximity, the structure in (28) also shows that in verb-final languages the PHP is in conflict with Behaghel's *Gesetz der wachsenden Glieder*, which predicts a general preference for short-before-long. This fundamental conflict between the two ordering principles could explain why so many verb-final languages are also characterized as 'free word order' languages, as there is often no clear statistical preference for a 'basic word order' in these languages. This conflict may also be the reason why some verb-final languages allow constituents to appear after the verb (also known as 'leaking').

The constituent orders exemplified in (27) and (28), i.e., short-before-long in verb-initial and SVO languages and long-before-short in verb-final languages are also predicted as the preferred patterns by Hawkins's *Principle of Early Immediate Constituents*, which says that structures with shorter recognition domains are easier to process (a detailed comparison between the *Principle of Head Proximity* (PHP) the *Principle of Early Immediate Constituents* and the *Branching Direction Theory* can be found in Rijkhoff (2004, pp. 261–312); see also Song (2009) and Song (2012, pp. 234–303). Hawkins's *Principle of Early Immediate Constituents* is similar in spirit to the PHP with regard to ordering preferences among short and long constituents in a sentence, but the PHP seems to cover more ground in that it could also account for certain seemingly disparate grammatical phenomena, in addition to the ordering tendencies mentioned above. For example, one could regard the two following phenomena as strategies that reduce the overall ratio of $head_N$-to-$head_V$ proximity in a clause (Rijkhoff, 2004, pp. 311–312).

1. Keep the number of NPs in a clause to a minimum: the fewer the number of NPs in a clause, the better will be the overall head proximity ratio. This can be seen in Cayuga and other polysynthetic languages, where ordinary discourse can consist largely of verbs.
2. Increase the number of verbs in a clause: the more lexical verbs a clause contains, the shorter will be the overall distance between heads V and N. This is exemplified in the Papuan language Kobon: if a sentence would contain many arguments and adjuncts, speakers have a strong tendency to divide them over two or more clauses (which may contain an identical predicate) rather than including them all in one sentence.

Scope

The PoS states that differences in scope among modifiers in a certain domain is iconically reflected in the actual linguistic expression:

29. Modifiers tend to occur next to the part of the expression that they have in their scope.

Scopal relationships between modifiers in a domain can be captured in meaning-based, layered representations of linguistic expressions as employed in, for example, Van Valin's (1993) *Role and Reference Grammar*, Dik's (1997) *Theory of Functional Grammar* and its successor *Functional Discourse Grammar* (Hengeveld and Mackenzie, 2008). In a layered representation it is clearly indicated which part

('layer') of a domain is within the scope of members of a certain functional modifier category. As was shown in (16), at least the following functional modifier categories can be used to analyze both NPs and clauses (Rijkhoff, 2008a, 2008c): (1) CLASSIFYING MODIFIERS, (2) QUALIFYING MODIFIERS, (3) QUANTIFYING MODIFIERS, (4) LOCALIZING or ANCHORING MODIFIERS, and (5) DISCOURSE-REFERENTIAL MODIFIERS.

Modifier categories include both grammatical and lexical elements (cf (16)), covering all the 'dependents' in a domain that are not arguments or complements. In a formal, meaning-based representation, these modifiers are distributed over 'nested' layers around the head constituent, reflecting differences in semantic scope. Conventionally, grammatical modifier categories (LOCALIZING/ANCHORING MODIFIERS such as demonstratives or tense markers) are represented on the left, whereas lexical modifiers (e.g., QUALIFYING MODIFIERS in the form of an adjective, a prepositional phrase, or a relative clause; see **Table 4**) are represented on the right. The various layers are shown in a strongly simplified form in **Figure 1** (notice that clauses have at least two more outer layers of modification to accommodate propositional and illocutionary modifiers; see Van de Velde, 2012 on NP internal mood).

The iconic relationship between the distance of a functional modifier category relative to the head in the layered representation of a linguistic expression and constituent order is most clearly seen in the case of simple (nonphrasal) modifiers in the NP, because here the PoS does not compete with Domain Integrity (Section Domain Integrity), which says that the preferred position of an embedded domain (e.g., Genitive, Relative Clause) is in the periphery of its embedding domain. Behaghel's (1932) *Gesetz der wachsenden Glieder* ('law of growing members') is another competing ordering principle that is relevant here (see above). Furthermore, embedded domains tend to be more versatile in functional terms than non-embedded domains (section 'Same Constituent, Different Functions'; Rijkhoff, 2009b). On the assumption that deictic adnominal demonstratives (dem) typically serve as LOCALIZING/ANCHORING MODIFIERS, cardinal numerals (num) as QUANTIFYING MODIFIERS, and genuine, nonderived adjectives (A) as QUALIFYING MODIFIERS, there are 24 logically possible combinations in an NP where the noun is modified by a member of all three modifier categories (as in English 'those three black boxes'). Only eight of these ordering patterns iconically reflect scopal relations among the three modifiers and it appears that each of these 'iconic' constituent orders is attested (Rijkhoff, 2008b, p. 802):

30. dem num A N (Dutch) — num A N dem (Berbice Dutch Creole)
    dem num N A (Spanish) — num N A dem (Hmong Njua)
    dem A N num (Zande) — A N num dem (Sango)
    dem N A num (Bambara) — N A num dem (Oromo)

In each ordering pattern in (30) the adjective is adjacent with the noun and the demonstrative is in the periphery, mirroring the layered representation in **Figure 1**. It has been argued that claims about the occurrence of noniconic patterns only constitute apparent counterexamples, because in such cases at least one of the following statements is true (Rijkhoff, 2004, pp. 272–273; 2008b):

31. a. constituents treated as semantic numerals or adjectives are formally speaking embedded domains headed by stative verbs or abstract nouns (see the Section Domain Integrity and heaviness);
    b. constituents are expressed as bound rather than free elements, which means their expression is a matter of morphology rather than syntax (Section Investigating Constituent Order Across Languages: Some General Issues);
    c. constituents are in apposition (rather than fully integrated constituents; see the Introduction).
    d. constituents are assigned a special pragmatic function like Focus, indicating we are dealing with a marked pattern (see, for example, the marked position of adjectives in Hungarian NPs).

Recently Cinque (2005) has argued that 14 orders are attested: the eight iconic patterns listed in (31) plus the six noniconic ones in (32). Interestingly, the noniconic orderings only involve postnominal modifiers (cf. the Introduction on apposition):

32. dem N num A — A N dem num — N dem A num
    N dem num A — N A dem num — N num A dem

However, since Cinque (2005, p. 315) employed semantic criteria to identify constituents (section 'The Employment of Semantic Criteria in Constituent Order Research'), he also included constituents that are characterized by one or more of the properties listed in (31). For example,

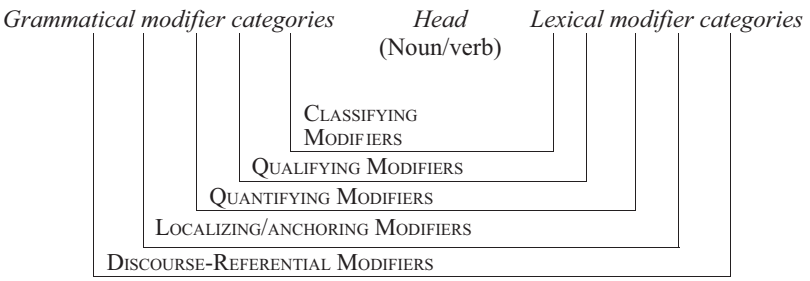

Figure 1 Some functional modifier categories in a layered representation of NP/clause structure reflecting differences in scope (Rijkhoff 2008a, 2008c).

- in Arbore (Cinque: N dem num A) demonstratives are suffixes and numerals are regarded as a subset of nouns;
- in Sampur (Cinque: dem N num A) adjectives are basically verbs heading a relative construction.

In sum, when we exclude apparent counterexamples like those mentioned under (31), it remains to be seen how many noniconic ordering patterns in (32) actually occur (in fact, there are just a handful of languages with noniconic orders in Cinque's study).

Arbitrary Patterning

Notice finally that even noniconic or 'arbitrary' ordering patterns can aid in language processing. Any consistent form of linear organization can be used to structure discourse. For example, speakers of the Peruvian language Iquito mark the distinction between REALIS and IRREALIS clauses solely by means of word order constraints (Beier et al., 2011). It is easier to identify constituents and process the linguistic expression they are part of when constituents with the same functional specification occupy the same position in a linear structure. Thus, whether a language has OV or VO order is essentially arbitrary, but knowing that one's language normally has the object argument before the verb provides the language user with interpretation or decoding strategies. Similarly, whether a language has AN or NA order is essentially arbitrary, but knowing which order is used in a language facilitates information processing.

Abbreviations

1 First person
2 Second person
3 Third person
A Adjective
ACC Accusative
ART Article
AUX Auxiliary verb
COMP Complementizer
DAT Dative
DT Dative-transformative
Dem Demonstrative
ERG Ergative
FEM Feminine
G/GEN Genitive, possessor phrase
HABIT Habitual
N Noun
NC1 Noun class 1
NC7 Noun class 7
NP Noun phrase
Num Cardinal numeral
O Object
OBL Oblique
PT Past tense
PERF Perfective
PL Plural
PP1 Class 1 pronominal prefix
PREP.PHRASE Prepositional phrase
PRES Present tense
REFL Reflexive
Rel Relative clause
S Subject
SC Subject concord
Sg Singular
V Verb
V-1 Verb-initial.

Acknowledgment

I am grateful to Eva van Lier for helpful comments on an earlier version.

See also: Comparative Method in Linguistics; Dependency in Language; Functional Approaches to Grammar; Grammatical Relations; Iconicity in Linguistics; Information Structure in Linguistics; Linguistic Fieldwork; Linguistic Typology; Markedness in Linguistics; Parts of Speech.

Bibliography

Behaghel, O., 1932. Deutsche Syntax/4. Wortstellung, Periodenbau. Winter, Heidelberg.
Beier, C., Hansen, C., Lai, I.-W., Michael, L., 2011. Exploiting word order to express an inflectional category: reality status in Iquito. Linguistic Typology 15, 65–99.
Bisetto, A., 2010. Relational adjectives crosslinguistically. Lingue e linguaggio 9 (1), 65–85.
Blake, B.J., 1983. Structure and word order in Kalkatungu: the anatomy of a flat language. Australian Journal of Linguistics 3 (2), 143–175.
Blake, B.J., 2006. Australian languages. In: Brown, Keith (Ed.), Encyclopedia of Language & Linguistics, second ed. Elsevier, Oxford, pp. 585–593.
Cinque, G., 2005. Deriving Greenberg's universal 20 and its exceptions. Linguistic Inquiry 36 (3), 315–333.
Devos, M., 2008. A Grammar of Makwe. Lincom Europa, München.
Dik, S.C., 1997. The Theory of Functional Grammar – Part 1: The Structure of the Clause (second revised edition, edited by Kees Hengeveld). Mouton de Gruyter, Berlin.
Dryer, M.S., 1992. The Greenbergian word order correlations. Language 68 (1), 81–138.
Dryer, M.S., 2005. Order of subject, object and verb. In: Dryer, M.S., Haspelmath, M. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich (Chapter 81). Available online at: http://wals.info/chapter/81 (accessed on 06.03.13.).
Dryer, M.S., 2009. The Branching Direction Theory of word order correlations revisited. In: Scalise, S., Magni, E., Bisetto, A. (Eds.), Universals of Language Today. Springer, Berlin, pp. 185–208.
Foley, W.A., 1991. The Yimas Language of New Guinea. Stanford University Press, Stanford, CA.
Greenberg, J.H., 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (Ed.), Universals of Language. MIT, Cambridge, MA, pp. 73–113.
Haiman, J. (Ed.), 1985. Iconicity in Syntax. Proceedings of a Symposium on Iconicity in Syntax, Stanford, June 24–26, 1983. Benjamins, Amsterdam (Typological Studies in Language 6).
Hale, K.L., 1976. The adjoined relative clause in Australia. In: Dixon, R.M.W. (Ed.), Grammatical Categories in Australian Languages. Australian Institute of Aboriginal Studies, Canberra, pp. 78–105.
Halliday, M.A.K., 2004. An Introduction to Functional Grammar, third ed. Arnold, London.

Hawkins, J.A., 1983. Word Order Universals: Quantitative Analyses of Linguistic Structure. Academic Press, New York.
Hawkins, J.A., 1994. A Performance Theory of Order and Constituency. Cambridge University Press, Cambridge.
Heine, B., König, C., 2010. On the linear order of ditransitive objects. Language Sciences 32, 87–131.
Hengeveld, K., Mackenzie, J.L., 2008. Functional Discourse Grammar. Oxford University Press, Oxford.
Kittilä, S., 2006. Object-, animacy- and role-based strategies: a typology of object marking. Studies in Language 30 (1), 1–32.
Kutsch Lojenga, C., 1994. Ngiti: A Central-Sudanic Language of Zaire. Köppe, Köln.
Lehmann, W.P., 1973. A structural principle and its implications. Language 49 (1), 42–66.
Malchukov, A.L., 2000. Dependency Reversal in Noun-Attributive Constructions: Towards a Typology. Lincom, München.
Mattissen, J., 2002. Dependent-head synthesis in Nivkh – with an outlook on polysynthesis in the Far Northeast. In: Evans, N., Sasse, H.-J. (Eds.), Problems of Polysynthesis (Studia Typologica 4). Akademie Verlag, Berlin, pp. 135–166.
Perkins, R.D., 2001. Sampling procedures and statistical methods. In: Haspelmath, M., König, E., Oesterreicher, W., Raible, R. (Eds.), Language Typology and Linguistic Universals: An International Handbook, vol. 1. Walter de Gruyter, Berlin, pp. 419–434.
Rijkhoff, J., 1986. Word order universals revisited: the principle of Head Proximity. Belgian Journal of Linguistics 1, 95–125.
Rijkhoff, J., 2004. The Noun Phrase. Oxford University Press, Oxford [Expanded Pb of the 2002 Hb edition].
Rijkhoff, J., 2008a. Layers, levels and contexts in functional discourse grammar. In: García Velasco, D., Rijkhoff, J. (Eds.), The Noun Phrase in Functional Discourse Grammar. Mouton de Gruyter, Berlin, pp. 63–115.
Rijkhoff, J., 2008b. Descriptive and discourse-referential modifiers in a layered model of the noun phrase. Linguistics 46 (4), 789–829.
Rijkhoff, J., 2008c. Synchronic and diachronic evidence for parallels between noun phrases and sentences. In: Josephson, F., Söhrman, I. (Eds.), Interdependence of Diachronic and Synchronic Analyses. Benjamins, Amsterdam and Philadelphia, pp. 13–42 [Studies in Language Companion Series 103].
Rijkhoff, J., 2009a. On the (un)suitability of semantic categories. Linguistic Typology 13 (1), 95–104.
Rijkhoff, J., 2009b. On the co-variation between form and function of adnominal possessive modifiers in Dutch and English. In: McGregor, W.B. (Ed.), The Expression of Possession. Mouton de Gruyter, Berlin, pp. 51–106.
Rijkhoff, J., van Lier, E. (Eds.), 2013. Flexible Word Classes: Typological Studies of Underspecified Parts of Speech. Oxford University Press, Oxford.
Ross, M., 1998. Proto-Oceanic adjectival categories and their morphosyntax. Oceanic Linguistics 37 (1), 85–119.
Siewierska, A., 1988. Word Order Rules. Croom Helm, London.
Siewierska, A., 2006. Word order and linearization. In: Brown, K. (Ed.), Encyclopedia of Language & Linguistics, second ed. Elsevier, Oxford, pp. 642–649.
Song, J.J., 2009. Word order patterns and principles: an overview. Language and Linguistics Compass 3 (5), 1328–1341.
Song, J.J., 2012. Word Order. Cambridge University Press, Cambridge.
Stassen, L., 1984. The comparative compared. Journal of Semantics 3 (1/2), 143–182.
Testelec, Y.G., 1997. Word order variation in some SOV languages of Europe. In: Siewierska, A. (Ed.), Constituent Order in the Languages of Europe. Mouton de Gruyter, Berlin, pp. 649–679.
Tomlin, R., 1986. Basic Word Order: Functional Principles. Croom Helm, London.
Van de Velde, F., 2009. The emergence of modification patterns in the Dutch noun phrase. Linguistics 47 (4), 1021–1049.
Van de Velde, F., 2012. A structural-functional account of NP-internal mood. Lingua 122, 1–23.
Van Valin Jr, R.D. (Ed.), 1993. Advances in Role and Reference Grammar. Benjamins, Amsterdam.
Vennemann, T., 1975. Analogy in generative grammar: the origin of word order. In: Heilmann, L. (Ed.), Proceedings of the Eleventh International Congress of Linguists, Bologna-Florence, August 28–September 2, 1972, vol. II. Il Mulino, Bologna, pp. 79–83.
Weil, H., 1844. De L'ordre des Mots dans les Langues Anciennes Comparées aux Langues Modernes. Translated by C.W. Super as The Order of Words in the Ancient Languages Compared with that of the Modern Languages (1978). Benjamins, Amsterdam.
Weir, E.M.H., 1994. Nadëb. In: Kahrel, P., van den Berg, R. (Eds.), Typological Studies in Negation. Benjamins, Amsterdam, pp. 291–323.

Relevant Websites

http://typo.uni-konstanz.de/rara/intro/ – Das Grammatische Raritätenkabinett – A Leisurely Collection to Entertain and Instruct.
http://typo.uni-konstanz.de/archive/intro/ – The Universals Archive.
http://wals.info/ – The World Atlas of Language Structures (WALS).

Work and Industry, Sociology of

Tony J Watson, University of Nottingham, Nottingham, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The sociology of work and industry studies the patterning of work activities and work-related institutions in industrialized societies, taking into account ongoing changes in technology, global relations, labor markets, work organizations, managerial practices, and employment relations. It observes the ways in which these changes are related to patterns of inequality in modern societies and to the experiences of individuals and families as well as the ways in which workers challenge, resist, and make their own contributions to the shaping of social worlds. These concerns are consistent with but more extensive than what was previously labeled 'industrial sociology.' This term is currently rarely used. The former industrial sociology's essential concern with the social implications of industrialization can nevertheless be carried forward within a coherent but expanded 'sociology of work and industry.'

Beyond 'Industrial Sociology'

The most obvious reason to use the term 'the sociology of work and industry' rather than 'industrial sociology' is that the latter term is nowadays rarely used. There are various factors contributing to this change. At the heart of it is a shift of focus onto 'work' rather than on undertakings which might be labeled 'industrial.' This has come about for a further variety of reasons. Especially significant is the fact that erstwhile industrial sociologists have, in recent decades, become increasingly sensitive to gender aspects of social life and have recognized that a great of vital laboring effort is expended, especially by women, outside formalized 'industrial' workplaces. Reinforcing this trend has been a reluctance to analyze nonfactory-like workplaces such as medical centers, charities or sports clubs in the same way that industrial undertakings tended to be analyzed.

Although some writers have suggested that contemporary society is somehow 'postindustrial,' there are still grounds for the retention of the adjective 'industrial' in the field of study with which we are concerned given that the societies which are studied have all at some time become industrialized or, alternatively, are facing the prospect of industrialization in the future. We might further note that the term 'industry' has by no means disappeared from everyday language. It is not only used to refer in a generalizing way to an institutional slice of society ('the industrial sector') but to particular areas of formal economic activity – the retail industry, the agricultural industry, creative industries, the leisure industry, the car industry, the mining industry, and even the education industry and the health industry. Some of these expressions may be more amenable to sociological use than others. Sociologists of work have made valuable contributions to our understanding of the growing 'creative industries,' for example, in a way perhaps not significantly different from the way industrial sociologists used to write about the mining industry. What is important, however, if we are going to retain a notion of 'industry,' is that we go beyond a focus on the factory which the descriptor 'industrial' has tended to connote (together perhaps with extractive and transport undertakings). It is interesting to note that in the previous edition of the *International Encyclopedia of Social and Behavioral Sciences*, Berg wrote that "Industrial sociology denotes the systematic study of the cultural and historical forces that foster the development of manufactories, and the diverse organizational structures, infrastructures, and market structures with which this development is associated" (Berg, 2001: p. 7351; my emphasis).

Berg's chapter on industrial sociology has a valuable historical emphasis and her emphasis on the centrality of the factory to processes of industrialization is undeniable. It should remind us that the factory is still a highly significant aspect of people's working lives, if we look at work across the world. Such a global perspective also shows that the work carried out by women and men in less developed economies has important implications for the work and lives of people located in the vastly more affluent societies with which these establishments are intimately involved. But many contemporary sociologists of work would argue that, even with the long and hard hours worked in factories of developing economies, work for people in these establishments, life-dominating as they appear to be, remains an activity to be carried out in and around the home as well as at the rows of machines in giant manufacturing sheds. It is reasonable to speculate that the acceptance of this fundamentally important insight not only results from the growing gender-sensitive nature of contemporary sociology mentioned earlier but also from a growing interest among sociologists in human beings not just as members of social categories (factory worker, accountant, nurse, say) but as 'whole persons,' with biographies, homes, sexualities, and ethnicities, as well as particular 'jobs.'

An interest in looking beyond simple job categories was present within the older industrial sociology tradition but is now an increasingly central feature of research in the newer and broader sociology of work tradition, an enterprise which combines a concern with global and societal shifts in working patterns with an interest in the lives of the people caught up in such shifts. This trend can be interpreted as one in which sociologists are fulfilling, whether consciously or not, the promise implicit in Mills' notion of the *sociological imagination*, with its call for social scientists to examine the connections between the 'patterns of people's own lives' and 'the course of world history' (1970: p. 9). This would mean that a study of

people engaged in, say, unpaid 'intern' work roles would bring together an analysis of changing global and societal patterns in such things as labor flexibility, employer power, roles of trade unions and professional associations, and legal regulation with attention to the implications for different individuals undertaking these roles, in light of their particular class and educational background, gender, ethnicity, and self-identities.

Changing Conceptions of Work

The trends identified so far in the development of the contemporary sociology of work and industry all raise important questions of just how sociologists might most usefully conceptualize 'work.' From what has been observed so far, it is clear that we are seeing a broadening of the 'work' category of human activity. Industrial sociologists tended simply to regard work as the task-based activities for which people are paid by an employer, client, or customer. This is a conception of work which is often used in everyday conversation when, for example, people are referred to as being 'in work' or 'out of work.' But difficulties can arise in such conversations as in the following ethnographic snippet:

> I hear you retired earlier in the year.
> Yes, that's right. It's really nice not to get up early every morning to drive to work.
> But don't you miss work?
> Well, not really, I'm working as a volunteer in the local library two days a week. And I must admit that we neglected our garden for years. It's taking an awful lot of work to sort it out.

This is a perfect parallel with the dilemma facing sociologists of work and industry who, if they did not broaden what is to be treated as work, would have to ignore activities which people still regard as 'work,' even though they involve no wage, salary, or fee. However, sociologists of work are increasingly anxious not to ignore unpaid activities which involve the completion of tasks through the expenditure of labor. Some researchers have gone as far as suggesting this broadening of our conceptions of work in creating a 'new sociology of work' (Pettinger et al., 2006). These researchers are significantly influenced by Glucksmann's (2006) concept of a 'total social organization of labor' (TSOL).

A TSOL approach, or something like it, studies the ways in which "all the labor in a particular society is divided up between and allocated to different structures, institutions and activities" (Glucksmann, 2005: p. 67). Thus, if we were to look at a society in which bread is important in people's diets, we would not only consider the historical and cultural context of bread production, we would go on to look at the various ways in which bread is produced across the society as a whole (and perhaps beyond) – some in factories, some in bakers' shops, some in homes, and some in groups of baking hobbyists – all the time relating these activities to patterns of class, gender, taste, income, and so on. The fact that it is often difficult to distinguish which are work activities and which are 'nonwork' activities is not a problem to a researcher adopting a perspective of this kind. We might ask, for instance, whether or not we should regard as 'work' the hours of effort expended on restringing all the guitars of an amateur guitar band undertaken by one member of that band. This is not, however, something to worry about unduly. We would be better considering how the members of the band themselves regard this task and asking just how this particular task allocation came about. The fact that there are often blurred lines between work and nonwork is something to examine in any particular research situation. Despite this, it is helpful to take as a starting point for research in the sociology of work and industry a basic concept of 'work.' This would retain the notion of task-based activity and would combine this with the notion of the shaping of people's lives. Thus, we can usefully conceptualize work as the carrying out of tasks which enable people to make a livelihood, in the sense of economic well-being, the maintenance of social relationships, and the achievement of a coherent sense of self.

The concept of work suggested here entails attention to an enormous range of topics and issues. To map this territory, it is helpful to divide researchers' and theorists' interests and activities into five subterritories or dimensions of the sociology of work and industry.

Major Dimensions of the Sociology of Work and Industry

Work and Global Social–Political–Economic Change

The discipline of sociology emerged as an intellectual reaction to the massive changes in European societies (Nisbet, 1970) brought about by the sixteenth-century Reformation, the seventeenth- and eighteenth-century Enlightenment, and the late eighteenth- and nineteenth-century Industrial and French Revolutions which further shook and often overturned these institutions. Industrial sociologists focused on the industrialization element of these changes but, true to their sociological priorities and their reliance on the powerful analyses and critiques of Durkheim, Marx, and Weber, they showed how the massive expansion in the use of emerging complex technologies, bureaucratic structures, and large-scale concentrations of workers in and around factories could only be understood as part of fundamental changes in the societal distributions of power and the cultural and economic shifts coming about with what Weber called a 'spirit of capitalism.'

Sociologists continue to debate on the extent to which the form of industrial capitalist society which emerged in the nineteenth century and came to dominate large parts of the world in the twentieth century is being superseded by, for example, postindustrial or postmodern societies. Sociologists of work and industry played their part within wider debates while focusing in particular, we might say, on the ways in which the division of labor is changing in the world. This builds on the concerns of the key figures of early sociology. The two most important contributions here were, first, Durkheim's (1984) recognition that the division of labor of a society was central to its form of social solidarity ('mechanical' in societies with simple divisions of labor and 'organic' in societies in which solidarity came about through the interdependence of a multiplicity of occupations in a society). Second, we have Marx's distinction between

a social division of labor, where work tasks are allocated at the level of society into occupations and trades and a technical division of labor in which specialization occurs within an occupation or broad work task. Durkheim's concerns are central to the current interest in 'professions' while Marx's analysis is vital to the notion of the 'capitalist labor process' which is used in the study of a considerable range of work activities and managerial arrangements observing its distinctive approach to the design, control, and monitoring of work tasks and activities by managers acting as agents of the capital-owning class in order to extract surplus value from the labor activities of working people.

In studying changing divisions of labor in contemporary times, particularly in the light of rapidly changing information and computer technologies, considerable attention has been paid to the role of knowledge or 'information,' with modern societies sometimes being characterized as information or network societies (Castells, 2000). Different, but not unrelated, is the changing balance between manufacturing work and service work with some researchers utilizing a novel concept of the McDonaldization of work (Ritzer, 1993). This deals with the way in which industrial manufacturing (or 'Taylorist') principles of deskilling, mechanization, and rationalization are applied, not just to fast-food businesses but to work in retailing, banking, and many other 'service work' areas of employment.

The changes in social organization which brought into being both the institutions of industrialism and the discipline of sociology began in Europe but have implications now for almost every part of the world. The much contested concept of *globalization* has emerged as part of intellectual attempts to make sense of this development. Perhaps the simplest way to understand 'globalization' is to treat it as general trend in which the political, economic, and cultural activities of people in different parts of the world increasingly influence each other and become interdependent. The sociology of work and industry, with its roots in the thinking of the classical sociologists who first struggled to come to terms with European industrialization and its accompanying transformations, is now coming to terms with changes and trends on this much larger scale.

Work Organizations

If we build upon the insights into changing divisions of labor associated with industrialization introduced by Durkheim and Marx, it is possible to suggest that there are two fundamental principles of work structuring in industrialized societies. One is the occupational principle of work structuring (to be considered later) and the other is the organizational principle in which work is patterned as the outcome of institutional arrangements in which some groups of people conceive of and design work and then recruit, reward, coordinate, and control the efforts of other groups. Sociological study of work organizations not only examines them as parts of the wider social organization of societies, it also recognizes that organizational structures involve a great deal more than the formal arrangements represented by managers' organization charts, rules books, and operating procedures. This accords with Weber's emphasis on the unintended consequences of human social actions and means that unofficial/informal aspects are studied alongside official or formal arrangements. Despite the power of such ways as looking at organizations, certain writers have argued that postbureaucratic organizations are now emerging. These are variously characterized as 'entrepreneurial,' 'networked,' or 'culture-led.' Sociologists have reacted to this type of thinking by arguing, in effect, that although there may be new variants of bureaucracies appearing, this does not warrant a denial of analyzing such arrangements as bureaucracies (e.g., Reed, 2005).

Sociologists generally are continuing to develop forms of analysis which give balanced attention to human agency and to the structural and cultural circumstances, and contingencies which both constrain and enable human initiative taking in the organizational setting. This has been reflected in a shift of emphasis by sociologists of work and industry away from the so-called contingency theory in which an organization's contingencies were, implicitly at least, seen to determine organizational forms to one in which the contingent circumstances of any organization (such as its ownership, market circumstances, technology, and type of people employed) are recognized as being mediated by the 'strategic choices' made by managers (Child, 2005). Yet these managers are by no means members of unified groups, as is demonstrated by studies of organizational *micropolitics* (Burns, 1961). Managers are frequently in competition with each other in their careers and researchers adopting a *negotiated order perspective* (a notion beginning in the symbolic interactionist theoretical tradition; Strauss, 1978) apply this insight much more broadly. Negotiated orders are the patterns of activity emerging over time in an organization as an outcome of the interplay of the various understandings, interests, initiatives, and reactions of the individuals and groups (including and beyond those of managers) who are involved in that organization.

It is within this broad theoretical emphasis, together with a strong interest in cultural as well as structural aspects of organizations, that sociologists of work and industry have examined changing managerial and corporate initiatives in the managing of work organizations. The work of sociologists studying work and work organizations has often been inspired by a perceived need to counter the assumptions and analyses of managerial thinkers and practitioners, this perhaps most famously occurring with the investigations made at the Hawthorne plant of the Western Electrical Company in Chicago in the 1920s and 1930s (Roethlisberger and Dickson, 1939). These investigations countered the assumptions and recommended practices of the 'scientific management' movement or 'Taylorism' (Taylor, 1911). The spokesperson of the counter movement, the so-called human relations school, inspired by the Hawthorne studies, was Elton Mayo, an individual influenced by Durkheimian thinking. But where Durkheim hoped for the achievement of 'social integration' through an emphasis on occupational moral communities, Mayo looked to work organizations and, especially, to work groups for the fostering of positive social sentiments and moral stability in society.

This broadly critical role for the sociology of work and industry continues to be significant as it relates a variety of corporate and organizational initiatives to broad processes of *work restructuring*, in which patterns of work experience,

organizational, and occupational activity change as part of the general economic, political, and cultural shifts occurring across the world. This again involves countering such discourses as those of team working, lean production, delayering, flexible working, and a managerialist 'human resource management.' Research shows that high commitment/indirect control human resourcing and work design practices (using an organic or flexible bureaucratic style) are not the outcome of a straightforward 'progressive' movement among employers which will simultaneously achieve greater business effectiveness and all round fairness, as much 'Human Resource Management' writing implies. Managers, in fact, are equally ready to pursue rigidly bureaucratic low-commitment and direct control practices when the circumstances allow it (say when they are using a simple technology requiring low worker skills in a context of low-state regulation of employment practices).

Occupations and Occupational Structure

The work organization dimension of the sociology of work and industry outlined earlier deals with one of the two main principles of work structuring operating in industrialized societies: the organizational principle, with bureaucracy (which might be relatively rigid or relatively flexible) at its heart. Industrialization, the bureaucratizing of work and the shift to a complex technical division of labor, has increasingly dominated sociological analysis of the organizational principle. But it has not totally supplanted attention to the second principle. This is the much older occupational principle of work restructuring, the principle underlying the social division of labor (Marx) or the mechanical division of labor (Durkheim). Analyses of aspects of the occupational principle concentrate on the ways in which particular work tasks are done, taking as a starting point the carrying out of a particular type of work activity such as building houses, running businesses, driving taxis, or catching fish. Attention is then paid to the implications for workers and society of the existence of groups of people regularly undertaking similar tasks.

Because a person's occupation generally plays a part in their economic position and their social status, the sociological study of occupations necessarily links strongly with the sociology of inequality. It examines how continuous shifts in the social division of labor affect the life chances of people in particular occupations. In particular, close attention is paid to trends of market segmentation and occupational areas that associated with nonstandard or 'contingent' employment including part-time and temporary work, home-working and 'teleworking.' But there are also occupations which do not involve employment. Hence sociologists of work are increasingly looking at self-employment, work in the informal economy, domestic work, and 'volunteering.' Connected to all of these, and also more broadly, are questions of gender inequality, and sociologists of work often combine their research efforts with researchers who have a primary interest in gender matters. The same point applies to issues of ethnicity and work.

A long tradition in the occupational dimension of the sociology of work and industry is that of the in-depth study of specific occupations including managers and supervisors; knowledge, cultural, and creative workers; and 'dirty' or deviant occupations like prostitute, mortuary attendant, and drug dealer. When particular occupations are studied, the sociology of work has a range of occupational features to examine: processes of occupational recruitment and socialization; occupational careers; and occupational identities, occupational cultures, ideologies, and occupational communities. In the study of all of these, especially with regard to newly emergent occupations, the sociological perspective encourages the linking of occupational matters to changes covered in the dimensions of the sociology of work and industry reviewed earlier. This is especially the case with the extent to which occupational members may organize themselves to gain the relative autonomy associated with occupations having undergone or undergoing 'professionalization.'

Work Meanings, Experiences, and Identities

Throughout the history of the sociology of work and industry, with its roots in concerns about social transformations, attention has been paid to the fact that just what work means to people in their lives, varies from one historical period to another, from one society to another, from one occupational or organizational setting to another. Weber's concern with the protestant ethic and the spirit of capitalism mentioned earlier is clearly an inspiration here at the broader societal level. Moreover, a general social concern inspired, in the twentieth-century sociology of work, studies of worker *job satisfaction* and its opposite *worker alienation*. From the 1960s, these issues tended to be conceptualized in terms of people's *work orientations* (Goldthorpe et al., 1968), this shift being inspired by sociologists' dissatisfaction with the reductionist and often managerially biased psychologistic notion of 'work motivation.' This saw a shift toward studying the meanings that people attach to their work in the light of how they are managing their whole lives and its value lies in the way it connects issues in personal lives and social circumstances to the ways in which they think about and act in the work setting. This is fully in the spirit of the sociological imagination, linking the personal to the organizational and occupational and, indeed, to the societal. It also fits with the trend toward achieving a balance between attention to 'agency' and to 'structure' in sociology generally.

In the past two or three decades, the insights of the earlier shift of focus toward work orientations has been further developed by a strong concern with the connections between work and *identities* – the category of 'identity' covering both the *self-identities* (individuals' own notions of who and what they are) and *social identities* (cultural, discursive, or institutional notions of who or what any person might be). Human beings are seen as continually 'working' on their notion of self – doing *identity work* (Watson, 2008) – in part through a cognitive dialog with the discursive resources available or imposed upon them by the social world in which they live – notions of what a manager is, what a police officer, a waiter, a retail worker, a sex worker is. Research is increasingly showing how complex these processes are in the light of various ambiguities, stresses, tensions, and contradictions inherent in different work roles. Attention is increasingly paid to emotional aspects of work (including the part played by sexuality in the workplace). Recent research has examined the distinctive pressures on people involved in 'aesthetic labor.' Alongside these recent

developments, attention continues to be paid to long-standing concerns of the sociology of work such as to the changing ways in which people relate their working and their nonwork lives and to the people's experience of unemployment.

Conflict, Challenge, and Mischief at Work

Work organizations are not big consensual human machines rationally and neutrally directed by their designers and managers. They are settings in which there are multiple and often competing interests. This view is vital to the concept of the work organization as a negotiated order, a notion which was introduced earlier. The implication of this is that conflict will be ever present in workplaces alongside, and in tension with, the cooperative efforts without which the organization or workplace would not continue to exist. The conflict aspect of work arrangements and activities generally thus runs through all the dimensions of the sociology of work and industry. It is nevertheless helpful to treat the study of conflict and challenge as a dimension of the sociological study of work in itself. One immediate justification of this lies in recognition of the trend whereby specialist academic disciplines such as 'industrial relations' or 'employment relations' have tended be subsumed into work sociology, something which might be welcomed as bringing a much fuller contextualizing perspective to studies of the collective organization of workers and their bargaining with employers. The benefit of this is apparent in studies which examine the reducing influence of trade unions in societies, something inevitably related to the world changes and processes of globalization. Also, by bringing into analyses the concepts of contradiction, paradox, and unintended consequences of classical sociology, stronger light can be thrown onto the innovations and adaptations in work activity regulation following the reduction in the influence of traditional trade unions.

Given the pervasiveness of conflict of interest and orientation in workplaces, it has been wholly inadequate to attempt to study these within the frameworks of industrial and employment relations. Many of the challenges presented to those in charge of work operations arise from individuals – sometimes alone, sometimes in small groups – asserting their interests, their relative autonomy, their pride, and their sense of 'self.' Notions such as 'organizational misbehavior' (Ackroyd and Thompson, 1999) or 'organizational mischief' (Watson, 2012) have usefully pulled together the considerable range of activities in workplaces which, first, should not happen according to the official rules and structure of the workplace and, second, contain an element of challenge to the dominant interests of those in charge of the work organization. This type of framework can incorporate the long-standing industrial sociology interest (originally in the Hawthorne studies mentioned earlier) in how workers develop job control strategies and 'make out' in the light of managerial efforts to control them. Studies regularly appear of the role that humor plays in the workplace; the causes and consequences of sexual harassment and bullying; fiddling and cheating; and sabotage, rule manipulation, and 'defense of self' tactics developed by service workers in the light of the potential subservience and humiliation involved in serving customers or clients.

Conclusion

In the world of academic endeavor, as in the physical world, it is not possible to create a single, objective map of any given territory. The best one can do is to offer a map of the territory which the map-maker feels is the most useful one possible for the map-reader. The present chapter's mapping of the territory of the sociology of work and industry, with the utilizing of the notion of five 'dimensions,' has been produced in this spirit. Underlying the accounts of the various dimensions of the sociology of work and industry is the argument that they are unified and integrated by their application of the sociological imagination to the task of intellectually coming to terms with, and offering insights to the nonacademic public with regard to, the tensions and difficulties created for individuals and societies of living in an industrialized world.

See also: Bureaucracy and Bureaucratization; Capitalism: Global; Class: Social; Conflict: Organizational; Corporate Culture; Industrialization, Typologies and History of; Information Society; Interactionism, Symbolic; Labor, Division of; Organizations: Negotiated Orders; Social Identity in Sociology.

Bibliography

Ackroyd, S., Thompson, P., 1999. Organizational Misbehaviour. Sage, London.
Berg, I., 2001. Industrial sociology. In: Smelser, N.J., Baltes, P.B. (Eds.), International Encyclopedia of the Social and Behaviour Sciences. Elsevier, Amsterdam, pp. 7351–7356.
Burns, T., 1961. Micropolitics: mechanisms of institutional change. Administrative Science Quarterly 6, 257–281.
Castells, M., 2000. Materials for an exploratory theory of the networked society. British Journal of Sociology 51, 5–24.
Child, J., 2005. Organization: Contemporary Principles and Practice. Blackwell, Oxford.
Durkheim, E., 1984. The Division of Labour in Society (W.D. Halls, Trans.). Macmillan, London.
Glucksmann, M., 2005. Why "work"? Gender and the "total social organisation of labour". Gender, Work and Organisation 2, 63–75.
Glucksmann, M., 2006. Shifting boundaries and interconnections: extending the "total social organisation of labour". In: Pettinger, L., Parry, J., Taylor, R., Glucksmann, M. (Eds.), A New Sociology of Work? Blackwell/Sociological Review, Oxford, pp. 19–36.
Goldthorpe, J.H., Lockwood, D., Bechhofer, F., Platt, J., 1968. The Affluent Worker: Industrial Attitudes and Behaviour. Cambridge University Press, Cambridge.
Mills, C.W., 1970. The Sociological Imagination. Penguin, Harmondsworth.
Nisbet, R., 1970. The Sociological Tradition. Heinemann, London.
Pettinger, L., Parry, J., Taylor, R., Glucksmann, M. (Eds.), 2006. A New Sociology of Work?. Blackwell/Sociological Review, Oxford.
Reed, M., 2005. Beyond the iron cage? Bureaucracy and democracy in the knowledge economy and society. In: Du Gay, P. (Ed.), The Values of Bureaucracy. Oxford University Press, Oxford, pp. 115–140.
Ritzer, G., 1993. The McDonaldization of Society. Pine Forge, Thousand Oaks, CA.
Roethlisberger, F.J., Dickson, W.J., 1939. Management and the Worker. Harvard University Press, Cambridge, MA.
Strauss, A., 1978. Negotiations. Wiley, New York.
Taylor, F.W., 1911. The Principles of Scientific Management. Harper, New York.
Watson, T.J., 2008. Managing identity: identity work, personal predicaments and structural circumstances. Organization 15, 121–143.
Watson, T.J., 2012. Sociology, Work and Organisation. Routledge, London.

Work, Anthropology of

Jim Weil, Science Museum of Minnesota, St Paul, MN, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by G. Spittler, volume 24, pp. 16565–16569, © 2001, Elsevier Ltd.

Abstract

Work encompasses direct subsistence pursuits, paid employment, and other activities contributing to survival and enhancement of individual and group life. Such efforts may be coerced, onerous, and deficient in rewards, but they can also bring fulfillment. Assumptions and conceptual breakthroughs that have shaped the study of work within anthropology are outlined. Work is identified as a prominent theme in classic ethnographic accounts. The genealogy of anthropological research more explicitly focused on work is traced. Recent scholarship examining the margins of work at its intersections with unemployment, underemployment, and leisure is reviewed, along with reflexivity in the work of anthropologists themselves.

Disciplinary and Subdisciplinary Orientations

Anthropologists study work in many ways, often inadvertently while more concerned with other themes. Work encompasses direct subsistence pursuits and paid employment, along with other activities contributing to survival and enhancement of individual and group life. While the efforts may be coerced, onerous, and deficient in rewards, they can also bring fulfillment. Much of what people think and do is affected by understandings, intentions, and identities derived from work. Considering all aspects of work, anthropologists address fundamental attributes of what it means to be human.

This article begins by outlining assumptions and conceptual breakthroughs that have shaped the study of work within anthropology. Next it explores treatments of work as a prominent theme in classic ethnographic accounts. The discussion traces the genealogy of research more explicitly focused on work, expanding and deepening our understanding of its centrality in social life. A survey of the current proliferation of literature is not possible, but major topics are identified. Much recent scholarship examines the margins of work at its intersections with unemployment, underemployment, and even leisure. Concluding remarks integrate the reflexivity that arises when the analytical framework extends to the work of anthropologists themselves.

Terminology

Given the pervasive and varied uses of the term, any attempt to define work narrowly would be futile. In an early delineation of the anthropology of work as a field of study, Sandra Wallman (1979) grounded work in energy expenditures, including thoughts ('psychic energy'), which produce both material and immaterial things of value. This and other behavioral criteria lay a foundation for the anthropology of work, but additional considerations are necessary to specify why certain activities in which people engage in a given setting are considered work and others are not.

What meanings do anthropologists apply to work and related terms? In English, *labor* usually contextualizes work socially. As indicated by its use for the birthing process, labor refers to biological as well as social reproduction. In this article, the focus is on work but with labor always part of the picture. Thus, *employment* is a condition of labor rather than work as such. *Occupation* and *livelihood* are terms, which reflect configurations of productive pursuits with some degree of social integration and permanence. *Career* is a problematic designation for conglomerations of roles conveying a sense of continuity, identity, and competence. Anthropologists have used this term to refer to the work of self-sufficient food collectors and food producers as well as individuals and groups in technologically advanced societies with complex divisions of labor (Goldschmidt in Gamst, 1995; Orr, 1996).

Whether work entails direct inputs to survival and well-being through *production* and *exchange* of goods and services or through less direct forms of *reproduction*, it involves culturally specific overlays on more fundamental pursuits. For example, consumption practices may blur the boundary between work and leisure, as with shopping, which may be a recreational pastime. This raises the inferential problem of cognitive criteria in relation to observable behavior. Ambiguities and nuances in terminology are illustrated below by examples, which interrelate the personal, interpersonal, behavioral, and institutional aspects of work.

Anthropological Departures

Anthropology's scientific (and humanistic) legacy of understandings about work draws from three conceptual domains, arguably fostering more holistic approaches than are usually encountered in other disciplines. The first is *natural economy*, an evolutionary construct for distinguishing degrees of local self-sufficiency based on use values, including vestiges of domestic modes of production in the present day. The second, *political economy*, concerns power relations that shape and are shaped by access to and control of resources, labor (and labor markets) in particular. The third, *moral economy*, pertains to ethical, spiritual, and legal codes that influence work through ideas about obligations, justice, and ultimate ends. Anthropologists have been in the forefront in documenting and assessing the local implications of the global integration of these three dimensions of economic life (e.g., Narotzky, 1997; Collins, 2003; Hakken, 2003).

Eric Wolf's (1983) ironically titled and frequently cited *Europe and the People without History* can be read as a comparative

analysis of the incorporation of localized and regionally connected labor regimes in the making of the modern world. Synthesizing ecology and political economy in his analysis, he shows how the (re)organization of work and expropriation of labor in widely dispersed zones with different kinds of resources were crucial to European colonial expansion and the development of capitalism. Wolf's global ethnohistorical survey highlights the centrality of work in processes of change in forms of culture, which he defines as social constructions built up from forces generated through established ways of mobilizing social labor, i.e., *modes of production*.

Disputes regarding evolutionary versus historical change may conflate incongruent scales of time, space, and changing human–environment relationships. Although industrial societies were not neglected as anthropologists positioned themselves among other social and behavioral scientists studying work (Applebaum, 1984; Gamst, 1995; Mollona et al., 2009), anthropology has been characterized by a broader comparative interest in what it means to be human in all times and places. An archaeologist seeking evidence of livelihoods in two-million-year-old chipped stone tools can make common cause with an ethnographer who observes the tasks and interaction styles of computer hardware and software engineers. Anthropologists also have been attentive to current events, for example, through ethnographic research that illuminates the assumptions and practices of Wall Street traders that contributed to the financial meltdown of 2008. The context for anthropology of work in recent years has been the neoliberal world economy and the precarious labor market that is one of its defining characteristics.

Although most anthropological research about work has been ethnographic, understanding of the multiple facets of work is strengthened when the concerns and expertise of more than one subdiscipline are brought to bear. Biological anthropologists have examined anatomical and physiological correlates of work in terms of energetics and health indicators. Physical structures and capacities of the human body, beyond their original importance in the course of human evolution, remain a factor in ongoing cultural change. Archaeologists encountering the material remains of work cannot directly answer questions about associated behavior patterns and meanings. However, from one era to another in any given area, similarities and differences in flora, fauna, raw materials, tools, and work sites generate insights about continuity and change in subsistence patterns and other features of economic life. Archaeologists who design their investigations of wealth and power in relation to production, exchange, and consumption can trace continuities and discontinuities in the political economy of work as well. Stylistic analysis, yet another avenue of archaeological investigation, offers otherwise inaccessible insights about the ways a distinctive *ethos* can shape work and its products over time.

Ethnographic Trajectories

Anthropology's long ethnographic tradition integrates the conditions, procedures, and products of work as parts of an entire way of life. These findings have enabled ethnologists – through comparative cross-cultural interpretations and explanations – to trace both recurrent patterns and significant discontinuities holistically. Work is a polysemic term. *Embodiment* refers to its physicality and its experiential aspects: the first in the neuromuscular manipulation of materials and tools, the procedures carried out, and the corporal effects of these efforts, and the second in the satisfactions, sufferings, struggles, and sense of identity of participants in work-related roles and activities. *Engagement* refers to the ways individual roles and interpersonal interactions are organized. Work rarely originates in isolation. Individuals benefit from preparatory nurturing and training, which applies not only to established routines – as with apprenticeship – but also to changes brought about by innovators and entrepreneurs. *Embeddedness* refers to the norms and social matrix that facilitate, constrain, and otherwise shape individual and group efforts. Economic anthropologists have long argued over the degree to which self-interest and rationality are culturally contingent, asking whether the meaning of work varies from one social context to another. Recent trends in the study of work have brought the subjective dimension to the forefront, building on a long tradition of biographical studies (e.g., Mintz, 1974 [1960]). Thus, the cumulative ethnographic tradition in anthropology has yielded richly nuanced descriptions, interpretations, and explanations of work.

Food Collection

Detailed ethnographic accounts of work appeared at least as far back as the depictions and explanations of arctic hunting by Franz Boas (1964 [1888]) in the late nineteenth century. By the mid-twentieth century, a sense of urgency propelled quantitative and qualitative research on the food quest of the few remaining small and relatively self-sufficient gathering and hunting societies in marginal habitats. This most basic human form of provisioning is often described as 'foraging,' but 'food collection' – whereby a few individuals split off in different directions to find and bring food and other necessities back to share with the rest of the group – differs fundamentally from the browsing and grazing of other species. The search for origins in the evolution of work in early human societies is distinct from arguments over historical continuities and discontinuities in structurally similar subsistence patterns occurring from ancient times into the recent past.

Gathering, Hunting, and Quality of Life in the Kalahari Desert. Among many fine-grained ethnographic accounts of food collectors, Richard B. Lee's 1979 study of *The !Kung San: Men, Women, and Work in a Foraging Society* (excerpted in Applebaum, 1984) stands out. Incorporating research strategies of biological anthropologists, Lee quantified the mix of tasks and considered the efforts expended in obtaining foods in relation to their nutritional values. Differences between the time spent by women gathering plant foods, and occasionally capturing small animals, and by men hunting larger game revealed a marked division of labor. The implications of this documentation are profound, both to discern a gender axis for assessing other elements of social organization and for inferences about cultural change. Men devoted more time to the food quest, but women spent more time on work overall, when other recurrent domestic tasks were taken into account. Lee found the rhythm of women's work steadier, distinguishing 'subsistence' (including travel) from 'manufacture and

maintenance,' 'child care,' and 'housework' (with cooking mostly done by women, for instance, but with larger game cooked by men). He decided not to categorize planning for the hunt as work, since reports of game sightings and consultation of oracles took place during the relaxation of pleasurable evening gatherings. While relevant to efforts to establish primordial baselines for work and leisure, this distinction is also useful for comparative discussions of the fuzziness of such boundaries in many settings today.

Evolutionary and Historical Perspectives. The much disputed assumption of an 'ethnographic present' beyond the here and now of a given study is relevant in the search for persistent features of the natural economy. To what extent, if any, can subsistence patterns documented in the mid-twentieth century be extrapolated as continuous from early human societies? At stake are not only debates about the significance of continuities in environmental relations and historical political economy, but also controversies over quality of life, cultural autonomy, and international human rights today (Povinelli, 1993; Narotzky, 1997; Lee in Nash, 1998; Gordon and Douglas, 2000).

Time allocation and health indicators among relatively autonomous gatherers and hunters raise questions about pressures and motivations for the development of food production systems. Evidence suggests that food collectors averaged fewer hours per day than members of agricultural and industrial societies in meeting their subsistence needs until population growth or displacement left them with an inadequate resource base (Johnson in Applebaum, 1984). Beginning in the second half of the twentieth century, however, indigenous peoples have been struggling to retain or regain rights to territories and subsistence regimes. For instance, 'labor action' by Belyun Aboriginal Australians, especially women, led to successful legal contestations of their land claims (Povinelli, 1993).

Food Production

Work has been central to major social transformations, none more momentous than the agricultural revolution. With the domestication of plant and animal species, greater productivity and population growth led to more competition for resources, increasingly complex settlement patterns, and corresponding institutional responses.

Gardening and the Work of Magic in the Trobriand Islands. Published in 1935, Malinowski's magisterial *Coral Gardens and their Magic: A Study of the Methods of Tilling the Soil and of Agricultural Rites in the Trobriand Islands* (excerpted in Applebaum, 1984) is an early, meticulous account of crop cultivation in Melanesia. It merits attention not only for the thorough descriptions of all the physical elements one would expect – soils, plants, tools, and techniques of horticulture – but also for Malinowski's intricately detailed explanations and interpretations of associated activities. From the perspective of a coherent anthropology of work, the ritual tasks of the specialist in magic who coordinates the work schedule are as relevant to the production process as physical manipulation of the environment (Spittler, 2008). Once crops are growing, as Malinowski puts it, "Nature now has to do most of the remaining work." In accord with Malinowski's functionalist theory, then, the magician's work parallels other tasks required to fulfill the population's survival needs.

Family Farming in Nigeria, Switzerland, and China. Robert Netting's *Smallholders, Householders: Farm Families and the Ecology of Intensive Sustainable Agriculture* (1993) surveys continuities and changes in the organization of agricultural work during the second half of the twentieth century in societies where marketing of crops and dairy products supplemented subsistence farming. Drawing on his own cultural ecology research in Nigeria in the 1960s and Switzerland in the 1970s, and extrapolating from research in other areas, Netting concludes that industrialized agriculture is not destined to replace (peasant) family farming everywhere. He treats households as work groups, with their members motivated to apply their knowledge and skills to the range of tasks required in localized agrarian social formations in ways that are efficient and environmentally sound. He documents their inputs through time allocation and their outputs in products for autoconsumption and the market. Netting's description of sustainable cultivation practices of the Kofyar of Nigeria, with hand tools and the inclusion of adults and children in the household division of labor, testifies to the possibilities for maintaining integration, security, and fulfillment in domestic farming in contrast to the fragmentation, insecurity, and alienation of agricultural wage labor. He finds similar work patterns in Swiss dairy farming families (cf Minge-Kalman in Applebaum, 1984) and contrasts these two ethnographic cases with the People's Republic of China, where collectivization countered venerable inequalities and vulnerabilities of the prior agrarian system, but where the regime eventually permitted greater household autonomy to reemerge.

From the Agricultural to the Industrial Revolution. Ethnographic studies of subsistence practices have necessarily taken mixed economic activities, also including fishing and herding, into account. Expanding networks of production and exchange have generated hybrid forms of work and multiple livelihoods, although degrees of self-sufficiency persist at the household and community levels. Economic downturns result in reversion to former practices, such as household gardening and inter-household barter, in situations characterized by exclusion or retreat from outside employment. Short-term and long-term economic and political reversals reinforce the significance of Netting's research on variable and incomplete transitions to large-scale, industrialized food production. Transitions from crafts production to technologically more advanced manufacturing are similarly incomplete.

Manufacturing

Manufacturing refers to the work of making things, whether with simple tools by hand (literally) or with heavy machinery or fully automated production systems. In the remote past, artisan roles developed through part-time or full-time specialization. Through the industrial revolution, artisan work was largely, although never entirely, replaced by the 'work' of machines. Anthropologists have studied a vast array of low-technology industries that persisted in some places such as Bolivia where large-scale manufacturing had not been established (Buechler and Buechler, 1992). Already in the early twentieth century, however, research in factories set the stage for new directions in the anthropology of work.

Workforce Organization in a Massachusetts Shoe Factory. Having first carried out ethnographic fieldwork under Radcliffe-Brown's guidance in Australia, W. Lloyd Warner was attuned to the way human behavior, including work, is structured according to status and role relations. Turning to research in the USA, he played a peripheral part in one of the pioneering experiments in applied industrial ethnography in the late 1920s. This was the flawed Hawthorne Project at the Western Electric plant in Cicero, Illinois, an attempt to apply the principles of Frederick Winslow Taylor to establish scientific labor control protocols. Warner contributed to an implicitly Fordist finding: the productivity of factory workers depended as much on morale as on rational organization of the assembly line. Warner went on to pursue his broader goal of developing anthropological concepts and methods relevant to the theoretical and practical problems posed by modern communities in urban-industrial settings. Newburyport, Massachusetts, was chosen for the renowned 'Yankee City' study he directed in the 1930s and 1940s. One resulting publication analyzes the composition and performance of the workforce in a shoe factory (Warner and Low, 1947). It traces this industry's long history, showing how job categories reflected divisions of gender, ethnicity, and national origin in the community at large. This enhanced the solidarity of immediate work groups but inhibited development of a shared working-class identity. The adversity of the Great Depression of the 1930s fostered worker solidarity, mobilized through a labor union in the 'natural history' of a strike. Marietta Baba (2009, cf Gamst in Gamst, 1995) attributes key innovations in the anthropological study of work to Warner, establishing protocols for the organizational research that came to the forefront in the closing decades of the twentieth century.

From Handicrafts to Heavy Industry and Beyond. Scattered through classic ethnographies – exemplified in the accounts by Boas, Malinowski, and Lee – are descriptions and analyses of the materials, processes, and arrangements of work involved in the making of tools and other objects. In subsistence-oriented societies, manufacturing was usually part of the mix of tasks performed in most if not all households, although some raw materials or finished products may have been obtained through exchange networks. Long-term industrialization and urbanization normalized full-time occupational specialization, and the factory worker became an iconic labor figure (Mollona et al., 2009). Much research has addressed technological and organizational changes in forms and conditions of industrial work, both where the human role persists, as in *maquiladora* assembly processes, and in more automated plants, where robots carry out tasks previously done by humans. Where manufacturing by hand has been revalorized in industries such as ethnic and tourist arts, anthropologists turn to new methodological and explanatory frameworks to accommodate its changing circumstances and meanings.

Coalescence of Anthropologies of Work

The preceding survey attests to the prominence of work as a persisting theme in anthropological research since the inception of the discipline, albeit more self-consciously and overtly for some scholars than others. When the *anthropology of work* emerged in the 1970s, the first programmatic statements embraced a comprehensive concern with work, conceptualized, and studied as broadly as possible. Anthropological investigations have ranged from direct subsistence practices through the gamut of other livelihoods, while also exploring the equivocal and previously neglected transitional areas between work and nonwork. These developments reflect reorientations in anthropology as a whole and its constituent subdisciplines. The authors of the case studies reviewed below have benefitted from participation in a collaborative cohort of practitioners investigating ever more detailed and kaleidoscopically shifting patterns in the world of work.

From Ethnography to Ethnology

By the late 1970s, anthropologists were making their shared interests in work more explicit. This newly recognized specialty was demarcated in publications generated from a groundbreaking symposium sponsored by the Association of Social Anthropologists in England (Wallman, 1979), sessions organized by the Society for the Anthropology of Work at meetings of the American Anthropological Association in the USA (e.g., Gamst, 1995; Nash, 1998), and a research group formed in Germany (Spittler, 2008). Along with theoretical and topical developments, methodological advances have supported more systematic comparative studies (e.g., the bioenergetic and time-allocation research previously mentioned). The emergence of the anthropology of work coincided with – and was stimulated by – the neoliberal turn in global political economy. Labor issues raised by deregulation, privatization, flexibility, and other attributes of *structural adjustment* presented ethnographic researchers with new interpretive and explanatory challenges (Nash in Nash, 1998; Collins, 2003; Ferguson in Mollona et al., 2009).

Since 1980, articles, reviews, and visual accounts have been disseminated through the *Anthropology of Work Review*, the journal of the Society for the Anthropology of Work, as well as other journals in many countries. Two of the leading publishers of books concerning the anthropology of work have been the State University of New York Press (formerly through its 'SUNY Series on the Anthropology of Work') and Cornell University Press. Given the limitations placed on references cited here, several are edited collections of articles that include interdisciplinary contributions and excerpts from exemplary monographs. This sampling highlights the theoretical development and ethnographic diversification of the anthropology of work, allowing for reference to only a few examples of the most recent scholarship. Readers interested in the themes explored here should track them in articles and the book review sections in the major journals.

Contesting the Boundaries of Work

Descriptions and analyses of work sites and situations, processes, and relations transcend the conceptual limitations of earlier models in anthropology and conventional assumptions in society at large. Ethnographic accounts published through the mid-twentieth century typically recorded a substantial residue of knowledge, beliefs, and practices in relatively stable

settings. From that period onward, cultural flux became the norm. These trends are particularly salient in reassessments of the development of women's work (Leacock et al., 1986).

Recognition of the mobility, subjectivity, and agency of workers responding to (or initiating) disruptions of long-established routines has led to reconsiderations of earlier scholarship. The so-called 'Bushman myth' controversy concerning the relevance of mid-twentieth-century food collectors' way of life for understanding early humans already has been mentioned. Nevertheless, critiques of our predecessors for writing in a timeless *ethnographic present* may attribute to them greater naïveté or bias than warranted. The shift of emphasis from continuity to change in anthropological research does not mean that preexisting configurations of skills, attitudes, and customs (*habitus*) necessarily become obsolete or irrelevant to ongoing survival strategies (Gamst on 'the web of rules' in Gamst, 1995).

June Nash's *We Eat the Mines and the Mines Eat Us: Dependency and Exploitation in the Bolivian Tin Mines* (excerpted in both Applebaum, 1984, and Mollona et al., 2009) demonstrates the resilience of Andean beliefs and practices in combination with Western religious and political ideologies. For example, to fortify themselves to withstand the rigors and danger of their work underground, tin miners both make offerings of coca to appease the syncretized 'devil' spirit (El Tío, or uncle) who holds sway over the underworld, including the mines, and chew the leaves as a mild stimulant. Leadership in protests for better working conditions was largely left to women, accustomed to the public work of managing community matters. Such examples of ethnic solidarity help explain how in 2006, an indigenous labor organizer from the federation of coca cultivator *sindicatos* could rise to become president of Bolivia, where neofeudal labor relations had persisted through the mid-twentieth century.

Beyond the ostensibly seamless blend of work and other activities in the natural economies of the imagined past, the conceptual and empirical boundaries of work have been effaced in at least three ways. A growing body of research addresses *unacknowledged work* supporting the lives of individuals and groups, spilling over from times and places designated by employment or applied to the quest for livelihoods lost or never secured. Under unstable economic conditions, *work in flux* has become the new norm, manifest through displacement, mobility, and improvisation. Anthropologists have become acutely aware of the *fuzzy boundaries of work* as it blends into a wide array of activities that support everyday life, political actions, aesthetic expressions, and leisure pastimes.

Diverse activities merge when workers create militant movements not only to claim material benefits as rights, but also to affirm identities as agents of their own destinies through art and other aesthetic expressions (Calagioni et al., 1992). This *affective* dimension of work has increasingly received attention, notably in Aihwa Ong's research on spirit possession as a response to women's factory work discipline in Malaysia and Carla Freeman's study of lifestyle enhancements, including professional attire, as a motivation for women's employment as highly regimented data entry technicians in Barbados (both in Mollona et al., 2009). The following topics illustrate other ways in which the boundaries of work are breached, both in lived experience and in ethnographic analysis.

Work without wages. From a cultural evolutionary perspective, paid employment emerged after millennia of diverse forms of work not formally distinguished from – on the contrary, integrated with – other activities and institutions. The legacy of slavery has been studied ethnohistorically and ethnographically, for instance, in the cases of the vestiges of colonial relations in the underpaid work of household domestics in sub-Saharan Africa (Hansen, 1989) and the field labor of Jamaicans in the ethnically stratified workforce on banana plantations in the Caribbean area (Bourgois, 1989). The historical emergence of the sector recognized as the *economy*, eventually entailing the commoditization of labor all over the world, left housework and others kinds of unpaid work as a problematic residual category (Wadel in Wallman, 1979). One collection of case studies by anthropologists and sociologists distinguishes labor for one's own domestic group and self-employment as two kinds of work not usually remunerated with wages (Collins and Giminez, 1990). Unpaid internships have become part of the neoliberal career path, although more traditional apprenticeships still can be found (Coy, 1989; Buechler and Buechler, 1992). Much unpaid work also remains hidden in the realm of consumption. Shifts to 'self-service' in retail sales can be seen as increases in involuntary unwaged work. From pumping one's own gas to scanning groceries and shopping from home, tasks associated with consumption have displaced many job categories (only partially replaced by supporting occupations to handle sales by phone or online).

Global Factories and International Labor Migration. The launching of the neoliberal era scrambled workforce designations, leading to unprecedented intersections of class, ethnicity, nationality, age, and especially gender. An early collection of case studies by Frances Rothstein and Michael Blim (1992) compares and contrasts localized manifestations of the 'global factory' in Mexico, China, India, Italy, Spain, the UK, Egypt, South Africa, the USA, and Puerto Rico. One trend has been the establishment of factories, often assembly plants in designated export-processing zones, in former colonies and Third World nations. A groundbreaking set of case studies collected by June Nash and María Patricia Fernández-Kelly (1983), with several featuring the roles of women in the electronics and apparel industries, focuses on workforce inequalities resulting from freer global flows of capital and labor. With the rise of transnational manufacturers powerful enough to resist regulation by overseas governments, a key factor in this shift of production away from metropolitan centers (runaway shops) has been the availability of cheaper and less-resistant labor. In the USA, in particular, the corresponding factory closings and the resulting unemployment – or the threat of factory closings – have been studied ethnographically (Nash in Gamst, 1995; Collins, 2003). Along with seasonal farm work, immigrants have occupied the onerous, low-paying niches that remain in manufacturing and processing (such as meat-packing) plants (Lamphere et al., 1994, Griffith, 2006).

Learning from Silicon Valley. The name 'Silicon Valley' evokes the rise of informatics, high-tech frontiers, and settings that foster creativity and productivity. New formulations of work regimes in the San Francisco Bay Area – emerging with the development and marketing of the first personal computers in the 1970s – have been replicated in other parts of the world. The emphasis on 'knowledge work,' including development

and management of highly automated manufacturing processes, supersedes production models originally developed for heavy industry. J.A. English-Lueck's *Cultures@SiliconValley* (2002) portrays lives and lifestyles built around work that spills over from cubicles and labs to eating and recreational areas and, with much of the interaction online, into homes as well. Research in Silicon Valley has contributed new insights and focused not only on elite knowledge workers and their families, many of whom are Asian immigrants, but also on other occupational niches. The ethic of raising the quality of work by eliminating rigid protocols and allowing for creativity is pursued in a study of photocopy machine service workers by Julian Orr (1996). Developing the concept of *work practice*, he explains how technicians repair the machines by triangulating symptoms of malfunction, statements of clients, and the maze of standardized procedures. They share a subculture of discourse-based expertise not recognized in their formal job descriptions. Christian Zlolniski's (2006) research on Mexican immigrants in this high-tech zone focuses on subcontracted janitors but also looks at workers in the informal sector: street vendors, care takers, and even providers of professional services (e.g., dentistry). It includes the work of social activism – political struggles for union recognition, the provision of public services, and legal residency – at the lowest stratum of this famously affluent regional economy.

Workscapes. David Hakken, in *The Knowledge Landscapes of Cyberspace* (2003), envisions possibilities for 'resocializing' work through automated information technology. He calls for 'ethnographic knowledgescaping' as a way of studying how people in principle can work together without physical proximity or hierarchies of control. We can extend these ideas beyond informatics to the increasing fluidity of low-tech as well as high-tech work as people adapt their skill sets and work settings to transcend (or not) the disappearance of countless occupational niches and the emergence of new ones. Thus, while ethnographers still conduct fieldwork on farms, in factories, in offices, and at other relatively self-contained sites, deterritorialized modes of research – whether by following workers from place to place or tracking their activities in cyberspace – are increasingly coming to define the anthropology of work. A few more examples of new occupations – or old ones studied in innovative ways – are weavers and woodcarvers in Mexico, offshore call-center workers in India, second-hand clothing merchants in Zambia, tourism workers in Barbados, Chinese workers in a Korean factory in China, open-source software 'hackers' in California, Indian migrant laborers in Bahrain, child laborers in Brazil, Filipina domestic workers in Hong Kong, drug dealers and sex workers in New York, therapists for victims of workplace harassment in Italy, the work of young men looking for work in Ethiopia, and foreign news correspondents around the world.

Reflexivity and the Work of Anthropology

With the acceleration and intensification of globalization at the time the anthropology of work coalesced as an intentional research specialization, concern about the history of anthropology was also growing. This has encouraged a critical perspective on the work of anthropologists themselves (Fabian, 1991). Participation in colonial administration and other government programs has generated much controversy over the years. The ethical and scholarly issues that arose over 'embedding' anthropologists in military units in Iraq and Afghanistan are an extreme example of debates about entanglement in armed conflict that have divided the discipline in the USA since the Vietnam War but go back to the formative stages of field research under the auspices of nineteenth-century imperial states. Another area of debate has been the engagement of anthropologists in the proliferation of socioeconomic 'development' programs, with research ostensibly carried out under the premises of value-free science.

Contemporary anthropologists continue the legacy of fieldwork in marginal or peripheral settings from the bottom-up perspective of people whose working lives have been increasingly uprooted and displaced (Nash, 1998). With a shift to over half of professional anthropologists working outside of academia and other traditional venues such as museums, it has become less uncommon to be hired by the constituencies traditionally studied, carrying out research, for instance, on behalf of indigenous groups or labor unions (Field and Fox, 2007; Durrenberger and Reichart, 2010). Now anthropologists are likely to conduct research in urban neighborhoods and institutional settings – medical facilities, schools, offices, restaurants or bars, or more diffuse workspaces – rather than in relatively discrete communities. Anthropologists are also more likely to study groups with which they themselves identify (autoethnography). Whatever the setting and focus, they enter the field less as relatively powerful strangers able to act on their own volition than as guests who must justify their presence to the people they presume to study (Gellner and Hirsch, 2001).

The adoption of practice theory to assess the relationship between structure and the agency of anthropologists in situations in which investigations occur has also drawn attention to research ethics, including dissemination and use of findings. To the extent that such reflexivity transcends narrow or simplistic research strategies, practitioners of the anthropology of work can be characterized more by their attitude or angle of approach than by a clearly demarcated set of principles and methods that constitute an institutionalized subdiscipline. A central concern with work leads in many directions related to elements of social life encountered in its evolutionary roots, institutional manifestations, political economy, and expressions of cultural identity.

See also: Apprenticeship, Anthropological Aspects; Craft Production, Anthropology of; Development: Social-Anthropological Aspects; Economic Anthropology; Energy, Anthropology of; Exchange in Anthropology; Families and Households, Formal Demography of; Fishing Societies; Gender, Economics of; Horticultural Societies; Hunter-Gatherer Societies, Archaeology of; Hunting and Gathering Societies: Anthropology; Industrial Societies; Informal Sector; Irrigation Societies; Kula Ring, Anthropology of; Labor History; Labor, Division of; Modes of Production; Pastoralism in Anthropology; Peasants in Anthropology; Political Economy in Anthropology; Postindustrial Societies; Work, History of; Work, Sociology of.

Bibliography

Applebaum, Herbert (Ed.), 1984. Work in Non-market and Transitional Societies and Work in Market and Industrial Societies. State University of New York Press, Albany.

Baba, Marietta, 2009. W. Lloyd Warner and the anthropology of institutions: an approach to the study of work in late capitalism. Anthropology of Work Review 30, 29–49.

Boas, Franz, 1964[1888]. The Central Eskimo. University of Nebraska Press, Lincoln.

Bourgois, Philippe, 1989. Ethnicity at Work: Divided Labor on a Central American Banana Plantation. Johns Hopkins University Press, Baltimore.

Buechler, Hans, Buechler, Judith Maria, 1992. Manufacturing against the Odds: Small-scale Producers in an Andean City. Westview, Boulder, CO.

Calagioni, John, Francis, Doris, Nugent, Daniel (Eds.), 1992. Workers' Expressions: Beyond Accommodation and Resistance. State University of New York Press, Albany.

Collins, Jane, 2003. Threads: Gender, Labor, and Power in the Global Apparel Industry. University of Chicago Press, Chicago.

Collins, Jane, Giminez, Martha (Eds.), 1990. Work without Wages: Comparative Studies of Domestic Labor and Self-employment. State University of New York Press, Albany.

Coy, Michael (Ed.), 1989. Apprenticeship: From Theory to Method and Back Again. State University of New York Press, Albany.

Durrenberger, Paul, Reichart, Karaleah (Eds.), 2010. The Anthropology of Labor Unions. University Press of Colorado, Boulder.

English-Lueck, J.A., 2002. Cultures@SiliconValley. Stanford University Press, Stanford, CA.

Fabian, Johannes, 1991. Time and the Work of Anthropology. Harwood, Chur, Switzerland.

Field, Les, Fox, Richard (Eds.), 2007. Anthropology Put to Work. Berg, Oxford.

Gamst, Frederick (Ed.), 1995. Meanings of Work. State University of New York Press, Albany.

Gellner, David, Hirsch, Eric (Eds.), 2001. Inside Organizations Anthropologists at Work. Oxford, Berg.

Gordon, Robert, Douglas, Stuart, 2000. The Bushman Myth: The Making of a Namibian Underclass, second ed. Westview Press, Boulder, CO.

Griffith, David, 2006. American Guestworkers: Jamaicans and Mexicans in the U.S. Labor Market. The Pennsylvania State University Press, University Park.

Hakken, David, 2003. The Knowledge Landscapes of Cyberspace. Routledge, London.

Hansen, Karen Tranberg, 1989. Distant Companions: Servants and Employers in Zambia, 1900–1985. Cornell University Press, Ithaca, NY.

Lamphere, Louise, Stepik, Alex, Grenier, Guillermo (Eds.), 1994. Newcomers in the Workplace: Immigrants and the Restructuring of the U.S. Economy. Temple University Press, Philadelphia.

Leacock, Eleanor, Safa, Helen, et al., 1986. Women's Work: Development and the Division of Labor by Gender. Bergin and Garvey, New York.

Mintz, Sidney, 1974 [1960]. Worker in the Cane: A Puerto Rican Life History. W. W. Norton, New York.

Mollona, Massimiliano, de Neve, Geert, Perry, Jonathan (Eds.), 2009. Industrial Work and Life: An Anthropological Reader. Berg, Oxford.

Nash, June (Ed.), 1998. Twenty years of work anthropology: a critical evaluation. Anthropology of Work Review 18 (2).

Nash, June, Fernández-Kelly, María Patricia (Eds.), 1983. Women, Men, and the International Division of Labor. State University of New York Press, Albany.

Netting, Robert, 1993. Smallholders, Householders: Farm Families and the Ecology of Sustainable Agriculture. Stanford University Press, Stanford, CA.

Narotzky, Susana, 1997. New Directions in Economic Anthropology. Pluto Press, London.

Orr, Julian, 1996. Talking about Machines: An Ethnography of a Modern Job. Cornell University Press and ILR Press, Ithaca, NY.

Povinelli, Elizabeth, 1993. Labor's Lot: The Power, History, and Culture of Aboriginal Action. University of Chicago Press, Chicago.

Rothstein, Frances, Blim, Michael (Eds.), 1992. Anthropology and the Global Factory: Studies of the New Industrialization in the Late Twentieth Century. Bergin and Garvey, New York.

Spittler, Gerd, 2008. Founders of the Anthropology of Work: German Social Scientists of the 19th and Early 20th Centuries and the First Ethnographers (Ruth Shubert, Trans.). Transaction Publishers, Piscataway, NJ.

Wallman, Sandra (Ed.), 1979. Social Anthropology of Work. Academic Press, London.

Warner, W. Lloyd, Low, J.O., 1947. The Social System of the Modern Factory: The Strike, a Social Analysis. Yale University Press, New Haven, CT.

Wolf, Eric, 1983. Europe and the People without History. University of California Press, Berkeley.

Zlolniski, Christian, 2006. Janitors, Street Vendors, and Activists: The Lives of Mexican Immigrants in Silicon Valley. University of California Press, Berkeley, CA.

Work, History of

J Ehmer, University of Salzburg, Austria

© 2015 Elsevier Ltd. All rights reserved.

This article is reproduced from the previous edition, volume 24, pp. 16569–16575, © 2001, Elsevier Ltd, with revisions made by the editor.

Abstract

The article brings together two themes. First, it discusses the history of the concept of work, from the philosophers of Antiquity and the Judeo–Christian tradition to modern Western thought. Second, the article concentrates on the history of work as practice. In doing so, three perspectives are employed, which predominate in the historiography of work from its very beginning until present: work as production and application of technology, work as economic activity, and work as culture. Finally, the article discusses the relation between work and gender as a field in which concepts and practices of work are closely interrelated.

The concept of work designates certain human activities; however, those activities that are considered to be work and those that are not, vary among different cultures and epochs. Each historical society has its own definition of work, and a variety of different conceptions are juxtaposed to one another within each society. How work is defined is closely associated with the value placed on it, and this in turn is connected with workers' positions in society. The history of work must therefore bring together two themes: the history of the meaning of work and the history of work as a practice. The relation between gender and work cuts across these two themes and shows how strongly they are connected.

History of the Concept of Work

The state of research on the history of the concept of work and the meaning of work suffers from three shortcomings. First, it is based upon the reflections of the intellectual elite. The ways in which other social groups – above all, the workers themselves – understood and evaluated work have not left much of a trace in historical source material, nor has much research been done on this tissue. The history of the concept of work is a history of ideas, and – in very few approaches to this field – a history of mentalities as well (Le Goff, 1980). Second, research has tended to ascribe a unitary – or, at least, a predominant – concept of work to the respective historical society under investigation. Contradictory or conflicting attitudes have been almost neglected. Third, the history of the concept of work has remained Eurocentric. Despite the pronounced interest in the work ethic of non-European cultures and religions that was already evident in the work of Max Weber, and in spite of the growing interest in global history, a systematic intercultural comparison remains a desideratum (Godelier, 1980).

Concepts of Work in Preindustrial Societies

The history of the concept of work in Western historiography begins as a rule with the great philosophers of Greek Antiquity (Aristotle, Plato) (Conze, 1972). They employed the term work (*ponos*, *poiesis*) to designate work done manually, and described thereby activities that required physical exertion and strength, that presupposed the use of tools, and the aim of which was to turn out a particular product. The meanings that coalesced to define the word *ponos* included labor, toil, hardship, and suffering. This concept of work had a parallel in the Indo-Germanic languages, which as a rule contain several terms for work and working: *labor* and *opus* in Latin, *travailler* and *ouvrier* in French, *labor* and *work* in English. The first word stem means great effort, travail, and affliction, activities dictated by necessity or constraint. The second word stem refers to the results of this effort – the attainment of a certain goal, the creation of a product.

Most historians agree that philosophical and political thought in Antiquity was characterized by contempt for work. Work was an evil, the activity of slaves, while the élite engaged in art, philosophy, and politics. The connection between abhorrence for manual labor and slavery is confirmed by the contrast provided by non-European cultures. Slave labor did not play an important role in the historical societies of East Asia, in whose great religious-philosophical systems – foremost among them Buddhism and Confucianism – work (the fieldwork of peasants and the handiwork of artisans) had a positive connotation. But European Antiquity was also characterized by contradictory attitudes toward work. Critical schools of thinking such as the Cynics esteemed physical and mental labor and contrasted them to idleness, which they regarded as a vice. In ancient Rome as well, particularly during the Republic, peasant farm work performed by free men was not seen in a totally negative light.

In the Judeo-Christian tradition, the ambivalence of the concept of work is expressed with particular clarity. In the Old Testament, God himself appears as a worker, and man was originally meant to care for God's creation. "And the Lord God took the man, and put him into the garden of Eden to dress it and to keep it" (Genesis 2:15). However, with man's fall from grace, work became penance and punishment: "In the sweat of thy face shalt thou eat bread" (Genesis 3:19). The New Testament carried forth this ambivalence. Christian writings of the Middle Ages constitute an "ideological arsenal containing weapons for the defence of every position, in favour both of labour and nonlabour … ranging from … contempt and condemnation to a tendency to respectful evaluation" (Le Goff, 1980: p. 77). From the tenth century on, a new trifunctional or tripartite social schema took hold; it divided society into 'priests' (*oratores*), 'warriors' (*bellatores*), and 'workers' (*laboratores*). This model assigned all working people to the lowest level of the social hierarchy, but nevertheless made them an

indispensable segment of society. The Reformation strengthened the tendencies to enhance the status of work. Luther and Calvin understood work as serving God. The work that each Christian performed was considered a 'calling' or a 'divine calling,' and thus assumed a quasi-religious character. The adherents of certain Protestant denominations – above all, the Puritans – regarded success they achieved in their work as confirmation that they numbered among God's elect. Since the publication of Max Weber's study (1904/1905), it is generally accepted that the higher valuation of work during the Reformation legitimated and promoted the spread of a bourgeois work ethic, even if it has important roots in medieval thought.

Concepts of Work in Modern Western Society

In early modern Europe, work was redefined in two respects. The first definition exhibits anthropological traits. Work came to appear as active human intervention in nature for the purpose of assuring the ongoing existence of the human species. At the same time, work appeared to be a social process based upon collaboration among human beings. This new idea of work played a central role in Enlightenment philosophy and the writing of cultural history done during this period. Man was seen as ruling over nature; division of labor and cooperation as the foundation of human socialization; and increasing domination of nature and further division of labor as the very epitome of social progress. Tools were considered the basis upon which human dominion over nature rested. Man thus appeared to be the 'tool-making animal,' as Benjamin Franklin put it. In the nineteenth century, Karl Marx and other thinkers of the early Industrial Age carried on this intellectual tradition. In his classic formulations in Volume I of *Das Kapital*, Marx described this dual character of work: "Labor is, in the first place, a process in which both man and Nature participate, and in which man of his own accord starts, regulates and controls the material re-actions between himself and Nature …. By thus acting on the external world and changing it, he at the same time changes his own nature" (Marx, 1867(1968): p. 192).

The second redefinition originated in the seventeenth century in the field of political economy, which linked the concept of work with the idea of value. Work was defined as an activity that created value, and value-producing activities were considered to be all those that were for pay or that generated income. This led to equating work with paid labor. In industrial capitalism, paid labor became the very epitome of work.

These new definitions were associated with the greatly enhanced status of work. "The modern age has carried with it a theoretical glorification of labor and has resulted in a factual transformation of the whole society into a laboring society" (Arendt, 1958: p. 4). Paid labor became the basis of self-esteem, social recognition, and political participation – in short, the admission ticket to civil society. However, attitudes toward work were ambivalent in modern society as well. The classic works of political economy of the eighteenth and nineteenth centuries, beginning with Adam Smith, feared the brutalizing and stultifying effects of extreme division of labor and repetitive mass production. For Marx, work in capitalism was exploitation and alienation; his communist utopia vacillated between self-realization by means of work and liberation from it. These ambivalences were expressed in increasingly multifaceted views of work. The esteem accorded to work referred to a specified type of work in contrast to its polar opposite, which was assessed negatively; mental verses physical work, planning and supervision verses actually performing tasks, skilled verses unskilled work, a steady job verses 'casual labor,' 'proper' work verses housework. A complex hierarchy of work is an essential component of the modern concept of work.

Work as Practice

Which activities were taken into consideration by scholars compiling the history of work as practice is closely connected to the development of the concept of work. Three perspectives emerge in the historiography of work: work as production and application of technology; work as economic activity; and work as culture.

Work as Production

The conception of work as dominion over nature that had emerged in the seventeenth century began to exert an influence on the history of work in the nineteenth century. Benjamin Franklin's characterization of man as the 'tool-making animal' was cited approvingly by Marx. Tools, or, in more general terminology, instruments of labor, were for Marx, fundamental to the periodization of history and a yardstick of human progress: "It is not the articles made, but how they are made and by what instruments, that enables us to distinguish different economic epochs" (Marx, 1867(1968): p. 194). Proceeding on this basis, Marxian historians and archeologists developed a world historical perspective on work that focused on two (or three) 'revolutions in the forces of production' (Childe, 1936; Bernal, 1954).

The first of these was called the Neolithic Revolution. Its point of origin was the Near East (Southern Anatolia, Mesopotamia, the Jordan Valley) and between about 8000 BC and 3000 BC it spread over large portions of South, East, and Southeast Asia, North Africa, Europe, and Central America. The essence of this revolution was the transition to agriculture and animal husbandry as the most important basis of human existence, and in the emergence of agrarian (farming) societies. According to this view, this was the beginning of goal-oriented human intervention in nature and was thus, in an alternate formulation, the beginning of work. "Early men did not work at all in the true sense … real work, steady work, labor for one's livelihood, came into being when agriculture was invented" (Williams, 1976: p. 335).

The second great revolution in the forces of production was, from this perspective, the Industrial Revolution, which began during the second half of the eighteenth century in Western Europe and encompassed large parts of the world over the course of the nineteenth and twentieth centuries. Now, the tools employed in performing work were no longer operated manually by the human hand, but rather by a machine that was driven by artificially produced energy. Thus, production became independent of the skillfulness, endurance, and physical strength of human beings. In capitalist fantasies, as in Andrew Ure's *Philosophy of Manufacturers*, 1835(1969), machinery even functioned as the most perfect means to

control and to discipline the workforce, as the final victory over the obstinate spirit of the workers. Factory, mass production, and conveyor belt became the symbols of work in industrial society.

The increasingly important role played by microelectronics and the computer in industrial production since the 1960s appeared, in this perspective, as the third great revolution in the history of work. Now, human beings withdrew from the process of controlling production and turned over mental activities as well to a machine, the computer. Automation, or flexible communication between humans and computers – a 'man–machine dialogue' – seems now to be the defining characteristic of a new world of work. The amount of human labor necessary for direct production is on the decline. The production process is increasingly becoming the activity of a 'collective worker' (Marx), in which technology and science play an important role.

This 'grand narrative' was subjected to critique and developed further by scholars working in the history of technology, a field that became over the course of the twentieth century an independent academic subdiscipline (cf 62117). Its early practitioners were engineers and technicians with their own narrow understanding of technology, who were not interested in human agency, but in the interaction of material objects. As early as the 1930s (Bloch, 1967), however, economic, social, and cultural historians were beginning to get involved in this field, and making the effort to conceive of technology as an integral part of society, and thus to conceive of work in an all-encompassing sense. In contrast to Marxist 'history of the forces of production,' they emphasized the numerous small technical innovations made in all historical epochs. An evolutionary perspective predominated in the history of technology. What the two approaches have in common is the tendency to represent the history of work as developing along a path rising in linear fashion, and as the history of progress.

Work as Economic Activity

At around 1900, with the newly awakened interest in economic and cultural history, there was also increased demand for a history of work, which then proceeded to emerge as a not-very-clearly delimited subdivision of general economic and social history of various different cultures and epochs. Since then, the history of work as an economic activity has been focusing on work as wage labor; on working hours, wages, and working conditions; and on the history of the modern work ethic and work discipline.

Wage labor was widespread in many preindustrial societies. In Europe, it took on increased significance from the late Middle Ages. Over the course of the Early Modern Period, the population became increasingly dependent upon wage labor. In the economically most highly developed regions of Europe, the 'proletarization' of the populace was a precondition and not a consequence of the Industrial Revolution. In the industrialization process of the nineteenth century, wage labor in its 'pure' form ultimately established itself as the predominant type of labor relationship. Among the features that typically characterized it were the standardization and regulation of work, including a high intensity of labor, a regular working rhythm, and a highly pronounced working discipline; a high degree of specialization of work along occupational lines; a clear demarcation of time spent at work and time off (over the course of a day, a week, a year, and a lifetime); the separation of the workplace from home; and a complex system of institutional and legal regulation.

The development of wage labor, however, was not a linear process. Recent research has shown that diversity of labor relationships was characteristic of the period from the late Middle Ages to the eighteenth century. A growing number of men and women earned their daily bread from a combination of agriculture, crafts and trades, and commercial activities, and by combining self-employment with wage labor. A high degree of flexibility and increased readiness to undertake labor migration were preconditions called for by this mixed economy. Flexibility, however, was not only a voluntary practice, but was enforced by the instability of the economy, by necessity and misery. The French historian Fernand Braudel suggested that commercial capital and industrial capital, small workshops and large-scale enterprises, rural home industry, and centralized manufacturing during this period should not be regarded as sequential phenomena, but rather as a pool of possibilities, as coexisting labor relationships and modes of production, among which men and women could choose and switch.

An ongoing point of discussion is the question of why men and women of the Early Modern Period increased the intensity of their work. Was this attributable to the exigencies of hardship in an increasingly polarized or proletarianized society, as mercantilist economists of the seventeenth century presumed? Was it the capitalist entrepreneurs who used shop regulations and penalties to turn up the pressure upon their employees (Thompson, 1967(1993))? Was the transformation of the world of work simply a part of a general disciplining process – including workhouses, prisons, and schools – to which men and women were forced to submit (Foucault), or to which they – as part of the process of civilization (Elias) – subjected themselves? Or was it the transformation of material culture – namely, the brave new world of consumer goods that expanded so rapidly in the Early Modern Period – that made working harder acceptable, or even desirable? Are the beginnings of consumer society thus also the historical roots of the laboring society (de Vries, 1994) (see Consumption, History of)? Which role was played by the changing perception of time, by which – from the fourteenth century on – Europeans began to adapt to the abstract rhythm of mechanical clocks? No matter how these questions are answered, it is clear that the role of the Industrial Revolution and mechanized factory labor as factors giving rise to the modern world of work have been devalued, whereas more emphasis is now placed on the long history of capitalism.

The industrialization process, too, became a field of revisionism. As it is conceived now, it was not represented by the factory alone; handicrafts, skilled work done by hand, and small-scale enterprises continued to play a key role (Sabel and Zeitlin, 1997; Samuel, 1977). The diverse branches of the service sector, such as transportation and shipping, retailing, education, entertainment, and the widely varied world of office work would very soon be employing more men and women than the manufacturing sector. In most industrial states, a majority of the workforce has never been employed in the secondary (industrial) sector. And ultimately, factory work has

proved to be more varied than it had long appeared to outside observers.

The concentration of research on the 'ideal type' of paid labor in modern society has caused too little attention to be paid to forms of labor that deviate from that model. This applies, for instance, to various forms of unfree or forced labor. In European history since the Middle Ages, slave labor has not been widespread, but as a basis of the Atlantic economic system from the sixteenth to the nineteenth century, it was indeed an essential component in the emergence of industrial capitalism. However, forced labor was widespread in the totalitarian political systems of the twentieth century. Seen from a global perspective, 'free' and 'pure' wage labor has (still?) not become the predominant form of work. Today, both agriculture as well as a combination of different activities and forms of work – as was characteristic of European history during the Early Modern Period – provides work and a living to a majority of the world's population.

Work as Culture

Since the 1980s, a new way of looking at the history of work has become widely accepted. A growing number of historians consider work as culture. As far as the concrete practice of research is concerned, this means a microhistorical perspective focusing on the workplace itself. Particular attention has been paid to two types of workplaces: first, the small workshops of artisans in the Early Modern Period and in the nineteenth century, that represent, on one hand, 'traditional' work, and, on the other hand, their increasing integration into capitalist relationships of production (Kaplan and Koepp, 1986); and secondly, the factories of industrial enterprises of the late nineteenth and the early twentieth centuries, as symbols of the modern industrial world of work (Lüdtke, 1993). The chief focus of these microhistorical studies is upon day-to-day working processes themselves and on social relationships in the workplace. They show the enormous diversity of practical labor and thus constitute a much-needed corrective to the generalizations yielded by the macrohistorical perspective. Scholars investigating the cultural history of work have also attempted to reconstruct the meaning of work for, 'common, everyday people' by using 'ego documents' such as autobiographies and diaries; iconographic and pictorial sources; and oral history interviews (cf Passerini, 1982).

There has also been great interest in the symbolic and ritual forms in which solidarities and conflicts are carried out and resolved on the workplace (cf Darnton, 1984). In investigating these phenomena, the microhistorical perspective also extends beyond the immediate confines of the workplace. Work gives rise to social relationships that retain their significance beyond the workplace, the term 'world of work' is used to connote this expanded meaning.

Often, work's significance becomes clearly evident only when it is absent. The implementation of the working norms of industrial capitalism also led to a redefinition of being out of work. Between 1880 and 1910, a new term describing involuntarily not working, became widespread: unemployment (*Arbeitslosigkeit*, *chômage*) (Topalov, 1994). The first microanalysis of work and unemployment that became one of the classics of twentieth-century social science is the study *Die Arbeitslosen von Marienthal*. In this small Austrian factory town in 1932, the textile mill that dominated the village's economic life shut down, so that practically the entire population was thrown out of work overnight. As a result, the rhythms of everyday life became slower and slower, and the unemployed were able to summon less and less energy for activities that they had previously pursued enthusiastically during their free time (Jahoda et al., 1933).

This draws attention to the fact that leisure is, on the one hand, the antithesis of labor, although on the other hand, it is also part and parcel of the world of work. In preindustrial societies, there was no strict dividing line between working and not working. The peasants of the Occitanian village Montaillou, who in about 1300 were interrogated by envoys of the inquisition about their way of life, interrupted their labors at the slightest opportunity – to chat with neighbors, to hang around the village square, to have some fun playing a game (LeRoy Ladurie, 1975). It was only Industrial Society that drew a clear distinction between working and not working, thereby changing the character of not only the world of work but of leisure as well. The concept of leisure refers less and less to merely not working and rather more to an ensemble of activities pursued by specific classes at certain times, such as games, sports, watching TV, and going to bars or symphony concerts. In this sense, leisure seems to some historians to be a cultural practice that was not 'invented' until the Early Modern Period (Burke, 1995).

Work and Gender

The relation between work and gender is a field of scholarship in which the reciprocities between the concept and practice of work, between work as economic activity and as culture, can be seen with particular clarity. Early studies of women's work conducted up into the 1970s almost exclusively treated gainful employment on the part of the women. Initial interest was focused upon factory labor, and then increasingly on 'specifically feminine' occupations like maid, governess, midwife, teacher, secretary, and salesgirl. The aim of this research was to reintegrate women into the male-defined world of paid labor. Beginning in the 1970s, feminist historians increasingly criticized the prevailing modern concept of work. Housework became the subject of theoretical discussions and of empirical investigations. Increasingly, studies on the history of women's work focused both on gainful employment and domestic activities (Tilly and Scott, 1978; Simonton, 1998).

The issue of the division of labor between the sexes has become a much-favored field of interdisciplinary research. Anthropologists, historians, and sociologists have shed a great deal of light on the variability of the division of labor. One and the same activity can be a typical man's job in one culture and woman's work in another. Which work is done by men and women does not depend on the 'nature' of their respective sex, but upon socially constructed gender roles. Recent theoretical approaches oppose biological or 'biosocial' as well as cultural determinism. Emphasis is in the variety of options taken in the sexual division of labor and the contesting definition and redefinition of norms.

In European history of the late Middle Ages and the Early Modern Period, the economic significance and the symbolic meaning of the respective activities assumed decisive explanatory power. One research approach emphasizes the exclusion of women from market-oriented activities in agriculture, crafts, and trades. Work that promised to yield profits, was learned and practiced according to formalized rules, coalesced into formalized occupations, and had organized politically along occupational lines became the domain of men. Early 'classical' studies of women's work and some recent research proceed under the assumption that up to the beginning of the modern period women also took part in artisanal market production and were organized in guilds – either in exclusive women's guilds or mixed guilds (Wiesner, 1989). According to this perspective, it was not until later that there occurred a gradual exclusion of women, be that as a consequence of 'male bonding,' the spread of capitalism, or the Industrial Revolution. Other historians have stressed the continuities of the gender-based division of labor since the twelfth century, which placed women in the domestic sphere or in low-skilled, low-status, and low-paying occupations (Bennett, 1988). Most historians emphasize the fluctuating boundaries between men's and women's work, which are continuously redefined and renegotiated (Honeyman and Goodman, 1998).

Another research approach emphasizes the cooperation between men and women. In this perspective, the 'working couple' formed the economic backbone of the agrarian and the artisanal economy (Pahl, 1984; Wunder, 1992). This can be seen even more clearly in the 'family economy' of the lower classes. Women played a central role in efforts to exploit and combine as many of the available income opportunities as possible. The 'industrious revolution' that preceded the industrial one in the view of many historians, was based on the cooperation of men and women in the family economy (de Vries, 1994). Conflict and cooperation are not necessarily antagonistic but rather complementary perspectives on the gender-specific division of labor in preindustrial Europe.

With respect to the nineteenth century, the focus of research has been on the 'male breadwinner model' and the emergence of 'separate spheres,' which made paid labor the exclusive domain of the male and, conversely, explained domesticity and care of children as being appropriate to the female nature (Janssens, 1998; Hausen, 1981). This was an effective model, even if there are diverging assessments of its effective range and the extent to which it actually did determine practice. Indeed, it was a short-lived model. The proportion of women pursuing paid labor began to climb again in the 1960s, including highly qualified professional work and lifelong careers. The history of the division of labor between the sexes shows that "the division of labor is an effect of the social hierarchy, not its cause" (Godelier, 1980: p. 170).

The feminist critique of the concept of work in modern society has considerably widened the boundaries of the history of work. Since generally applicable definitional criteria are not in sight, they have to be developed from case to case. "In this widest sense" the English historian Penelope J. Corfield wrote, "'work' is whatever people consider as work … the boundaries between 'work' and 'non-work' are in practice very often obscure" (Corfield and Keene, 1990: p. 208).

See also: Alienation, Sociology of; Alienation: Psychosociological tradition; Artisans and Guilds, History of; Capitalism; Industrialization, Typologies and History of; Labor Unions; Labor, Division of; Slavery as Social Institution.

Bibliography

Arendt, H., 1958. The Human Condition, second ed. 1998. University of Chicago Press, Chicago.

August, A., 2007. The British Working Class, 1832–1940. Pearson Longman, Harlow, England.

Behan, T., 1997. The Long Awaited Moment: The Working Class and the Italian Communist Party in Milan, 1943–1948. P. Lang, New York.

Belshaw, J.D., 2002. Colonization and Community: The Vancouver Island Coalfield and the Making of the British Columbian Working Class. McGill-Queen's University Press, Montréal.

Bennett, J.M., 1988. History that stands still: women's work in the European past. Feminist Studies 14, 269–283.

Berger, S., 2000. Social Democracy and the Working Class in the Nineteenth and Twentieth Century Germany. Longman, Harlow, England.

Bernal, J.D., 1954. Science in History. Watts, London.

Bloch, M., 1967. Land and Work in Medieval Europe (Selected papers by Marc Bloch). Routledge and Kegan Paul, London.

Burke, P., 1995. The invention of leisure in early modern Europe. Past and Present 146, 136–150.

Childe, V.G., 1936. Man Makes Himself. Watts, London.

Clark, A., 1995. The Struggle for the Breeches: Gender and the Making of the British Working Class. University of California Press, Berkeley.

Conze, W., 1972. Arbeit. In: Brunner, O., Conze, W., Koselleck, R. (Eds.), (Geschichtliche Grundbegriffe. Lexikon zur politisch)-sozialen Sprache in Deutschland. Bd. 1, Klett-Cotta, Stuttgart, Germany.

Corfield, P.J., Keene, D. (Eds.), 1990. Work in Towns 850–1850. Leicester University Press, Leicester, UK.

Darnton, R., 1984. The Great Cat Massacre and Other Episodes in French Cultural History. Basic Books, New York.

Donkin, R., 2010. The History of Work. Palgrave Macmillan, Houndmills, Basingstoke, Hampshire.

Dubofsky, M., 2000. Hard Work: The Making of Labor History. University of Illinois Press, Urbana.

Eggebrecht, U., Flemming, J., Meyer, G.v., Müller, A., Oppolzer, A., Paulinyi, A., Schneider, H., 1980. Geschichte der Arbeit. Vom Alten Ägypten bis zur Gegenwart. Kiepenbheuer & Witsch, Köln, Germany.

Ellis, P.B., 1996. A History of the Irish Working Class. Pluto Press, London.

Frader, L.L., Rose, S.O., 1996. Gender and Class in Modern Europe. Cornell University Press, Ithaca, NY.

Freund, B., 1995. Insiders and Outsiders: The Indian Working Class of Durban, 1910–1990. Heinemann, Portsmouth, NH.

Godelier, M., 1980. Work and its representations: a research proposal. History Workshop Journal 10, 164–174.

Guharaya, S., 2007. Calcutta Tramwaymen: A Study of Working Class History. Progressive Publishers, Kolkata.

Hausen, K., 1981. Family and role-division: the polarisation of sexual stereotypes in the nineteenth century—an aspect of the dissociation of work and family life. In: Evans, R.J., Lee, W.R. (Eds.), The German Family. Croom Helm, London, pp. 51–83.

Hindman, H.D., 2002. Child Labor: An American History. M.E. Sharpe, Armonk, NY.

Honeyman, K., Goodman, J., 1998. Women's work, gender conflict, and labour markets in Europe 1500–1900. In: Shoemaker, R., Vincent, M. (Eds.), Gender and History in Western Europe. Arnold, London.

Jahoda, M., Lazarsfeld, P., Zeisel, H., 1933. Die Arbeitslosen von Marienthal. Ein soziographischer Versuch. S. Hirzel Verlag, Leipzig, Germany.

Janssens, A. (Ed.), 1998. The Rise and Decline of the Male Bradwinner Family?. Cambridge University Press, Cambridge, UK.

Joyce, P. (Ed.), 1987. The Historical Meanings of Work. Cambridge University Press, Cambridge, UK.

Kaplan, S.L., Koepp, C.J. (Eds.), 1986. Work in France. Representations, Meaning, Organization, and Practice. Cornell University Press, Ithaca, NY.

Kessler-Harris, A., 2003. Out to Work: A History of Wage-Earning Women in the United States. Oxford University Press, Oxford, England.
Kirk, N., 1998. Change, Continuity and Class: Labour in British Society, 1850–1920. Manchester University Press, Manchester, UK.
Le, B.P., 1999. A Short History of the U.S. Working Class: From Colonial Times to the Twenty-first Century. Humanity Books, Amherst, NY.
Le Goff, J., 1980. Time, Work & Culture in the Middle Ages (Pour un autre Moyen Age: temps, travail et culture en Occident). The University of Chicago Press, Chicago.
LeRoy Ladurie, E., 1975. Montaillou, Village Occitan de 1294 á 1324. Editions Gallimard, Paris.
Lockman, Z., 1994. Workers and Working Classes in the Middle East: Struggles, Histories, Historiographies. State University of New York Press, Albany.
Lüdtke, A., 1993. Eigen-Sim. Fabrikalltag, Arbeitererfahrungen und Politik vom Kaiserreich bis in den Faschismus. Ergebnisse Verlag, Hamburg, Germany.
MacDowell, L.S., Radforth, I.W., 2000. Canadian Working Class History. Canadian Scholars' Press, Toronto.
Marx, K., 1867. Das Kapital. Kritik der politischen Ökonomie. Dietz Verlag, Berlin. Marx-Engels Werke Band 23(1968).
McIvor, A., 2001. A History of Work in Britain, 1880–1950. Palgrave, Houndmills, Basingstoke, Hampshire.
McLeod, H., 1993. Religion and Irreligion in Victorian England: How Secular Was the Working Class? Headstart History, Bangor.
Pahl, R.E., 1984. Divisions of Labour. Basil Blackwell, Oxford, UK.
Pahl, R.E. (Ed.), 1988. On Work. Historical, Comparative and Theoretical Approaches. Basil Blackwell, Oxford, UK.
Passerini, L., 1982. Work ideology and working class attitudes to fascism. In: Thompson, P. (Ed.), Our Common History: The Transformation of Europe. Pluto Press, London, pp. 54–78.
Sabel, C.F., Zeitlin, J. (Eds.), 1997. World of Possibilities. Flexibility and Mass Production in Western Industrialization. Cambridge University Press, Cambridge, UK.
Samuel, R., 1977. The workshop of the world: steam power and hand technology in mid-Victorian Britain. History Workshop Journal 3, 6–27.
Schneider, D., 1994. Trade Unions and Community: The German Working Class in New York City, 1870–1900. University of Illinois Press, Urbana.
Silbey, D., 2005. The British Working Class and Enthusiasm for War, 1914–1916. Frank Cass, New York.
Simonton, D., 1998. A History of European Women's Work. 1700 to the Present. Routledge, London.
Thompson, E.P., 1967[1993]. Time, work discipline and industrial capitalism. In: Thompson, E.P. (Ed.), Customs in Common. Penguin Books, London, pp. 352–403.
Tilgher, A., 1931. Work. What It Has Meant to Men Through the Ages (*Homo faber*) (Homo faber. Storia del concetto di lavoro nella civiltà occidentale). George G. Harrap, London.
Tilly, L.A., Scott, J.W., 1978. Women, Work, and Family. Holt, Rinehart and Winston, New York.
Topalov, C., 1994. Naissance du Chômeur 1880–1910. Édition Albin Michel, Paris.
Ure, A., 1835[1969]. The Philosophy of Manufactures; or, an Exposition of the Scientific, Moral, and Commercial Economy of the Factory System of Great Britain. 3rd edn. C. Knight, London; B. Franklin, New York.
Usner, D.H., 2009. Indian Work: Language and Livelihood in Native American History. Harvard University Press, Cambridge, MA.
Van, H.C.E., Schaffner, H.A., 2003. Work in America: An Encyclopedia of History, Policy, and Society. ABC-CLIO, Santa Barbara, CA.
de Vries, J., 1994. The industrial revolution and the industrious revolution. The Journal of Economic History 54, 249–270.
Weber, M., 1904/1905. (English edn. 1976) The Protestant Ethic and the Spirit of Capitalism. Scribner, New York.
Wiesner, M.E., 1989. Guilds, male bonding and women's work in early modern Germany. Gender & History 1, 125–137.
Williams, R., 1976. Keywords. A Vocabulary of Culture and Society. Fontana Press, London, p. 335.
Wunder, H., 1992. Er ist die Sonn', sie ist der Mond.' Frauen in der Frühen Neuzeit. C. H. Beck, Munich, Germany.

Work, Labor: History of the Concept

Bénédicte Zimmermann, Ecole des Hautes Etudes en Sciences Sociales, Paris, France

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Work refers as much to the content and the performing of an activity, to the production of economic goods and values in capitalist societies, as to the production of social identities and hierarchies. Since the late nineteenth century, capitalist societies have essentially been built on the two latter meanings, reducing work to labor, that is to abstract work that can be quantified and measured in time and money. The history of the concept is closely linked to the constitution of a social order on the basis of standards of class and nation in the framework of the welfare state, as well as to the production of its principles of intelligibility by the social sciences.

The concept of work has at least three different but complementary meanings in contemporary Western societies. As an action, it refers to the content and the performing of an activity; as a production of value, it is part of the organization of capitalist societies; finally, as a status, it constitutes a linchpin of social organization. Work, as the mainspring of economic activity, is also at the crux of the functioning of the political and social order. Affecting the slightest aspects of life in society, it can be viewed as a 'total social fact' (Mauss, 1966/1950). This pivotal position has made work a notion that has long been taken for granted as universal, to the detriment of its changing meaning in time and space. Neither has work always been a pillar of the social and political order in Occidental societies, nor does work similarly serve this function in any society. In Asian or African countries, the perception of work, its role in society and, therefore, the history of the concept took quite a different road (Kocka and Offe, 2000).

Yet its contemporary meaning in capitalist societies is the result of an eminently historical process, which made out of it a yardstick for economic and social organization. Although various meanings of the word 'work' existed in the nineteenth century (Cottereau, 1994), it is on the particularly restrictive concept of labor, in the sense of quantifiable productive activity, that Western capitalist societies have been built. The history of the reduction of the concept of work to that of labor, that is, to abstract work that can be measured in time and money, is also the history of the constitution of a social order based on wage labor in the framework of the welfare state. It is closely linked to the building of a collective order on the basis of standards of class and nation, as well as to the production of its principles of intelligibility by the social sciences.

Labor, conceptualized in political economy after Adam Smith under the angle of the production of value, became as a mediator between the individual and the collective, a focal point of sociological analysis. Hence, the transformations of productive activity and their consequences in industrial society have strongly contributed to shaping the development of the social sciences, but the descriptive as well as prescriptive discourse of the latter has also played an essential role since the early twentieth century in establishing the social sphere on the basis of work. The *Verein for Sozialpolitik* in Germany and the Fabian society in England provide cases in point for such an intertwining between social reform purposes and social sciences (Rueschemeyer and Skocpol, 1996).

The history of the concepts of work and labor will be examined from this angle of a dual process in which the social sciences and the societies they study mutually engender one another. Thus delimited, the history will not be a 'total' one, but situated in time and space, that is, focused on the question of the reduction of the concept of work to that of labor in capitalist European societies from the nineteenth century onward. At the turn of the twentieth century, this history overlaps with that of the statistic and legal codification of labor into social status, providing access to citizenship within the welfare state. Finally, since the 1980s, it turned into a reopening of the concept, closely tied to a reassessment of paradigms of action in European social sciences, after the dominant role played by functionalism and structuralism.

The Semantic Reduction of Work to Labor

Labor, more precisely wage labor, a dominant form in social-economic relationships in today's societies, has not always had this value as a standard. In the Middle Ages, when idleness was the privilege of the aristocracy, labor was synonymous with pain, poverty, and social devaluation. Without going back over the history of work (*see* Work, History of), it is important to underline the contradictory heritage that characterizes the notion in the nineteenth century. On the one hand, work was associated with the idea of alienation. This association finds its origins in the Greek conception of slavery, the function of which was to satisfy the vital needs of reproduction of the human species (Arendt, 1958). It was further developed by Marx's critique of wage labor, a modern form of alienation through work. But on the other hand, as Marx himself noted it, work is a source of individual accomplishment and self-expression. This alternative conception, fueling the Protestant ethic and its formalization of an occupation (*Beruf*) as a vocation (Weber, 2010/1920), as well as Hegel's conception of work as the essence of man (Hegel, 1977/1805), considerably fed the ideology of the welfare state. The tension between alienation and liberation was to mark the concept of labor in the long term and run through theoretical debates in the social sciences until today (Méda, 1996).

The Invention of Abstract Work

The contradiction between the anthropological conception of work as the essence of human existence and the reality of alienating labor provided the basis of the Marxist critique of the political economy (Marx, 1968/1867); a critique that marked a major turning point in the conceptualization of work (Noguera, 2011). Political economy produced an initial swing of the concept of work toward that of labor. This swing occurred in the context of liberalization of work, consecrating the idea of the market through the promulgation of a set of legal provisions abolishing hereditary dependency, establishing the free circulation of individuals, freedom of commerce, and free competition. On the theoretical level, Adam Smith's work (1776) consecrated labor as a value, as a source of national wealth, and saw freedom of work as the means to maximize this wealth. From Adam Smith to John Stuart Mill, British theorists of political economy made labor a product, a commodity sold according to the laws of supply and demand, on a market held to be the dominant form of regulation. Determined by the principles of economic liberalism, this conception makes out of work a commodity having an exchange value. According to the principle of individual freedom, each person was theoretically supposed to be able to sell this commodity as he or she wishes by way of freely agreed contracts. The contractual mediation made out of work a general and legal category. It meant the invention of abstract work, quantifiable and measurable in time and money: the invention of labor as a founding principle of capitalism (Biernacki, 1995).

Marx embraced this conceptualization of abstract work, but introduced two new elements. The first was labor power. Marx viewed labor as much from the perspective of the production of economic value as from the angle of the labor power (*Arbeitskraft*) required to achieve it. Doing so, he introduced a significant shift in the problematics of work. Work is no longer reduced to the mere economic dimension of producing and circulating goods, it is henceforth associated as well with the condition of the producer in a capitalist market system. The second contribution of the Marxist critique is the introduction of non-work, a notion previously absent from political economy (Marx, 1982: p. 208). Seen from the angle of the economic dependence of the worker, who can only meet his needs through the sale of his labor power, labor cannot be envisaged independently from non-work and its individual consequences. Marx thus contributed to a conceptualization of labor as dependent work, that is, seen from the producer's rather than the product side. This Marxist critique of political economy provided, in the second half of the nineteenth century, a fertile ground for the development of the social sciences. One of their contributions has been to objectify non-work by breaking it down into categories of social organization such as old age, disability, illness, and unemployment.

Wage Labor: A New Expression of Dependent Work

The concept of work used for this purpose of social objectivation meets the dual criteria of quantification and dependence. The reformist social sciences of the early twentieth century and the new categories they contributed to producing thus reinforced the process of dissociation between on the one hand work as a creative activity, characterized by the independence of the producer and the quality of the product, and on the other hand abstract work associated with dependence and quantity. This social differentiation grew more marked up until the 1960s, along with the spread of wage labor which was legally codified – by the means of labor contracts and labor legislation – and statistically codified, in the form of socio-professional categories, as a response to standardization and mass production in economic activity.

The concept of employment, at the heart of the legal codification of labor, is informed by a dual meaning: from a macroeconomic standpoint it means trading a given quantity of work for a given salary on the market; from the political and social standpoint, it means exchanging individual subordination – the worker's subordination to his employer – for security, the security offered by labor legislation and social protection. From this standpoint, employment is as much determined by the nature of economic activity as by the nature of the social relationships that regulate this activity. Employment can thus be defined as socialized labor, including, in response to the Marxist critique, protection for the producer. But this protection is highly variable geographically, reflecting the different avenues of building societies around the integrative virtues of work (Esping-Anderson, 1990). As Biernacki (1995) has shown in the case of England and Germany, the historical forms of objectivation of labor specific to each country – as a product for the English, as labor power for the Germans – affect the way in which the concept of labor is still operating nowadays in the production of the political and social order.

Labor and the Constitution of the Social Order

With the welfare-state, labor has become the nerve center of the social world. As participation in the common good, it was used at the end of the nineteenth century in the definition of a link that binds together individuals in the broadened framework of the nation-state. Consequently, it has been codified as a vector of individual belonging to the collective, insuring the group's cohesion.

The social sciences, as institutionalized at the turn of the twentieth century, played an essential role in this codification, indexing the social order on labor and defining labor's relationships as a matter of public concern (Durkheim, 1964/1893; Weber, 2010/1920). During the first half of the twentieth century, producing general categories aiming to organize the industrial society at the level of the nation-state was of major concern for most prominent social scientists. Stating scientific principles of homogenization and generalization ran parallel with the organization of society on a large scale, using labor as a fulcrum for standardization and social regulation within the framework of national institutions.

Labor as a Status

The close connection between work and social status is one of the main features of the concept of labor. This type of connection doubtless already existed in feudal societies and guild organizations, but its specificity since the late nineteenth

century lies in its relation to demands of social justice, political equality, and democracy. Labor has become the democratic means of access to social status and political integration, a function that for centuries was ensured by property and heredity (Castel, 1995). As a provider of income and the principle on which the redistribution of wealth is based, labor has become the central social relationship in capitalist societies. It contributes to determining social hierarchy and individuals' position within it, as well as shaping forms of citizenship and personal identities. Therefore lack of work, categorized as unemployment by the welfare state, does not only mean a loss of income likely to be compensated by social policies. Its consequences for individuals reach far beyond the shortening of economic resources. As shown by Jahoda, Lazarsfeld, and Zeisel in their ground-breaking study (2002/1933), unemployment causes a range of other disorders which bring to the fore the psychosocial functions of work. These "latent functions" (Jahoda, 1982), which become most relevant when one is unemployed, encompass collaboration with others, access to social interactions, constitution of social identity, as well as structuration of time and making sense of one's own life and role in society. The centrality of work in the constitution and organization of society, was not, however, a foregone conclusion even two centuries ago (Arendt, 1958). But it has become such a natural fact within welfare-based societies that it is today taken for granted (Zimmermann et al., 1999).

Labor, Class, and Nation

The socialization of work and non-work fulfills a double purpose in the political project of the welfare-based nation-state: providing security for wage workers through social protection and labor legislation and, thereby, consolidating the political legitimacy of the nation-state. This conjunction considerably increased the social value attached to paid and dependent work. The expression 'social security' suggests to what extent the socialization of non-work is as much a question of collective as of individual security. Collective, that is, social, security is based in the national welfare-state on the dichotomy between capital and wage labor, according to a scenario in which the social regulation of labor is thought of as a partner of economic regulation of the market and a basis for economic development. The encounter of the nation-state, economic organization, and labor institutions under the auspices of the welfare state thus appears as the fruit of a historical compromise between capital and labor. This compromise, at the basis of Fordism, is founded on the norm of full employment. From such a perspective, the social order can be seen as a collective order produced on the basis of wage labor and codified by a set of rights and duties attached to the salaried class (Castel, 1995).

The history of this collective order is, from 1890 to 1960, one of establishing a coherence between individual identities and economic and political practices on the basis of the principles of class and nation (Wagner, 1994). It is a movement of widespread standardization aiming to reduce uncertainty as well as the variety of possible actions and interpretations. Statistical and legal methods were the preferred instruments of this standardization and establishment of equivalences, going hand in hand with a dominance of the functionalist and structuralist paradigms within the social sciences. Through the institution of nomenclatures, statistics gave shape to a homogenous national body that can be observed and planned thanks to the presumed objectivity of figures. By granting social and political rights guaranteed by the state, legal codification stipulated individual belonging to this body. But during the two last decades of the twentieth century, these standardization principles have been put to the test by increasing demands of flexibility and the diversification of contractual status of workers, especially wage earners.

Transformations in Work and the Reconfiguration of the Social Sciences

At the turn of the twenty-first century, the quest for generality of the early twentieth century has been replaced by a concern for diversity. This holds true for economic and social life as well as for the social sciences. The issue has turned into producing new equivalencies between situations and persons, able to place increasingly individualizing practices into a collective order that guarantees equity. The requirement for flexibility, which spread from the financial markets and economic sphere to the whole political and social life, provides an exemplary illustration. Whereas flexibility is for the employer a powerful tool for dealing with the ups and downs of the economy and the necessity to adapt to versatile markets, it introduces for the employee uncertainty that the rhetoric of risk and predictability at the basis of the welfare-state is no longer able to govern and master (Evers and Nowotny, 1987).

One of the current social and political challenges is thus to make individual pathways secure in the new context of uncertainty that weighs on work (Supiot, 1999). But the categories established on the norm of full employment and the model of society that goes with it hardly manage to integrating these imperatives for diversity and flexibility. It is more generally the reduction of the concept of work to that of labor that is incriminated, to such an extent that an examination of the ongoing transformations entails a need to rediscover the diversity of meanings of work and to recast the relationship between the different facets of people at work: quite at the same time producer of value, creative individuals, and social beings.

In 1958 already, Arendt undertook a radical critique of this reduction that led, in her mind, to labor invading the political sphere, even crushing the latter in favor of the 'social.' Against the institution of labor as a general category of human action and its colonization of all the other fields of human activity, Arendt reminds us that it is merely one aspect among others of *Vita activa*. Doing so, she emphasizes a major differentiation between:

1. Labor, an activity of *anima laborans* governed by the human's needs of subsistence and reproduction, producing ephemeral things destined to be consumed;
2. Work, an activity of *homo faber*, producing lasting objects aimed at stabilizing human life; and
3. Action, *praxis*, which produces things that are inseparable from the actor and which only occur through interaction. Action thus constitutes for Arendt, the eminently public and political aspect of *V. activa*; a sphere that should not be contaminated by the invasive logic of labor and consumption.

Whereas Arendt deplored the progressive absorption first of work, then of action, by the reductionist concept of labor, the topicality of the relationship between work and action has since been renewed. Far from Arendts' claim in favor of a clear-cut dissociation between the sphere of labor, the one of work and the public sphere, their articulation has become a major political and social issue; notably because of the transformations of work, of labor statuses, and the growing part taken by the services and tertiary sector in the economy. In relation with the challenge of analyzing these transformations, the concept of work has opened up toward action, that is, work as activity. Several movements converged within the social sciences in favor of such a broadening of the concept of work. In 1982, Ralf Dahrendorf made a plea in favor of a conceptual shift from a labor society (*Erwerbsgesellschaft*) to an activity society (*Tätigkeitsgesellschaft*). Referring himself to Arendt's categories, he promoted an enlarged understanding of work integrating informal work, namely, domestic work and moonlighting (Dahrendorf, 1983). The feminist movement (Kittay, 1999) defended the same cause, initiating a debate on how to define and measure household work (Eichler and Albanese, 2007). Other authors seek for criteria of definition of non-market activities, like the "third person criterion" according to which "an economic activity is one which can be done by a third person without reducing its final utility value" (Hawrylyshyn, 1977). But the question whether one should or not limit the understanding of work to economic activity, that is to an activity that can be converted into an amount of money according to market rules, remains a crucial point of debate.

The shift toward a problematic of work as activity happened concomitantly with a revival of action theories in the social sciences, particularly the rediscovery of the pragmatist vein of the Chicago School. Pragmatism and interactionism have been, during the whole twentieth century, a strength of analyses and an answer to functionalism in American sociology of work, with scholars like E.C. Hughes (1958), A. Strauss, and their followers. Continental European social sciences have instead developed, as an alternative to the dominant functionalist paradigm of the 1960s, a structuralist approach mainly concerned with the status conferred on individuals by their jobs and the ensuing inequalities. The 1990s can insofar be characterized by a pragmatist turn in European social sciences (Joas, 1996/1992).

This turn contributed to spelling out the issue of work as action and activity, therefore taking into account a greater diversity of meanings of work (women's domestic work, voluntary and civic work, informal work, etc.) and exploring their moving frontiers with wage labor. It furthermore lead to envisaging work from the perspective of the interactions and conventions that preside over coordination among people at work (L'économie des conventions, 1989). Such an approach pays attention to the content of work in itself, to the kind of activity, products, people, and situations involved (Boltanski and Thévenot, 1989). After the major movement of standardization that characterized the history of the concept of labor and sealed its hegemony, the concept of work today is being redesigned to integrate the nature and the quality of what it produces, as well as work situations and people's capabilities (Zimmermann, 2011). It opens up to the analysis of a plurality of configurations of action (Salais and Storper, 1997/1993; Boltanski and Chiapello, 1999) taking account of the anthropological dimension of work.

By the same token, the uncertainties that weigh on national employment institutions, under the pressure of globalization and the constitution of supranational political entities (Habermas, 1999), have found considerable resonance in social sciences debates. After the supremacy of the national level, alternative scales of regulation and organization of the collective sphere are being reassessed, both from a practical and theoretical standpoint. In these matters too, the connection between economic and political transformations, on the one hand, and paradigmatic changes in the social sciences, on the other, is a close one. Once passed the postmodernist fever announcing altogether the end of work and history (Lyotard, 1984/1979), the twenty-first century opened an era of revising both the categories of organization of society built around work and nation, and their principles of intelligibility as they were defined by the social sciences. The challenges for a new conceptualization of work are numerous. Named globalization, sustainability, human development and capability, growing inequalities, ecology, etc., they require to rethinking work not only as an economical good, but as a human activity tightly intertwined with the other aspects of human life at both the individual and collective level.

See also: Alienation, Sociology of; Arendt, Hannah (1906–75); Capitalism; Class: Social; Differentiation: Social; Durkheim, Emile (1858–1917); Labor History; Labor Unions; Marx, Karl (1818–83); Smith, Adam (1723–90); Social History; Weber, Max (1864–1920); Welfare State, History of; Work, Anthropology of; Work, History of; Work, Sociology of; Working Classes, History of.

Bibliography

Arendt, A., 1958. The Human Condition. The University of Chicago Press, Chicago.
Biernacki, R., 1995. The Fabrication of Labor: Germany and Britain, 1640–1914. University of California Press, Berkeley, CA.
Boltanski, L., Chiapello, E., 1999. Le nouvel esprit du capitalisme. Gallimard, Paris.
Boltanski, L., Thévenot, L., 1989. Justesse et justice dans le travail. Presses Universitaires de France, Paris.
Castel, R., 1995. Les métamorphoses de la question sociale. Une chronique du salariat. Fayard, Paris.
Cottereau, A., 1994. Théories de l'action et notion de travail. Note sur quelques difficultés et quelques perspectives. [Theories of action and the notion of work–Difficulties and prospects]. Sociologie du Travail Sp.lss.Sl 36, 73–89.
Dahrendorf, R., 1983. Wenn der Arbeitsgesellschaft die Arbeit ausgeht. In: Matthes, J. (Ed.), Krise der Arbeitsgesellschaft? Verhandlungen des 21. Deutschen Soziologentages in Bamberg 1982. Campus, Frankfurt a. Main, pp. 25–37.
Durkheim, E., 1964/1893. Division of Labor in Society. Free Press of Glencoe, New York.
Eichler, M., Albanese, P., 2007. What is householdwork? A critique of assumptions underlying empirical studies of housework and an alternative approach. Canadian Journal of Sociology 32 (2), 227–258.
Esping-Anderson, G., 1990. Three Worlds of Welfare Capitalism. Polity Press, Cambridge, UK.
Evers, A., Nowotny, H., 1987. Über den Umgang mit Unsicherheit. Die Entdeckung der Gestaltbarkeit von Gesellschaften. Suhrkamp, Frankfurt/Main.
Habermas, J., 1999. Der europäische Nationalstaat unter dem Druk der Globalisierung. Blätter für Deutsche und Internationale Politik 4, 425–436.

Hawrylyshyn, O., 1977. Towards a definition of non-market activities. Review of Income and Wealth 23 (1), 79–96.
Hegel, G.W.F., 1977/1805. Phenomenology of Spirit. Clarendon Press, Oxford, UK.
Hughes, E.C., 1958. Men and Their Work. Free Press, Glencoe, IL.
Jahoda, M., 1982. Employment and Unemployment: A Social-Psychological Analysis. Cambridge University Press, Cambridge.
Jahoda, M., Lazarsfeld, P., Zeisel, H., 2002/1933. Marienthal: The Sociographie of an Unemployed Community. Transaction Publishers, Piscataway, NJ.
Joas, H., 1996/1992. The Creativity of Action. University of Chicago Press, Chicago.
Kittay, E., 1999. Love's Labor: Essay on Women, Equality and Dependency. Routledge, New York.
Kocka, J., Offe, C., 2000. Geschichte und Zukunft der Arbeit. Campus, Frankfurt/New York.
L'économie des conventions, 1989. Revue Economique 40 (2) (special issue).
Lyotard, J.-F., 1984/1979. The Postmodern Condition. A Report on Knowledge. Manchester University Press, Manchester.
Marx, K., 1968/1867. In: Marx, K., Engels, F., Werke (Eds.), Das Kapital. Kritik der politischen Ökonomie, vol. 23. Dietz, Berlin.
Marx, K., 1982. Oekonomisch-philosophische Manuskripte. In: Marx, K., Engels, F. (Eds.), Gesamtausgabe, vol. 2. Dietz, Berlin.
Mauss, M., 1966/1950. Sociologie et anthropologie. Presses universitaires de France, Paris.
Méda, D., 1996. New perspectives on work as value. International Labor Review 135, 633–643.
Noguera, J.A., 2011. Le concept de travail et la théorie sociale critique. Travailler 26 (2), 127–160.
Rifkin, J., 1995. The End of Work: The Decline of the Global Labor Force and the Dawn of the Post-Market Era. G.P. Putnam's Sons, New York.
Rueschemeyer, D., Skocpol, T. (Eds.), 1996. States, Social Knowledge and the Origins of Social Policies. Princeton University Press, Princeton, NJ.
Salais, R., Storper, M., 1997/1993. Worlds of Production. The Action Frameworks of the Economy. Harvard University Press, Cambridge, MA.
Smith, A., 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. Whiteston, Dublin, Ireland.
Supiot, A. (Ed.), 1999. Au-delà de l'emploi. Transformations du travail et devenir du droit du travail en Europe. Flammarion. Rapport pour la Commission Européenne, Paris.
Wagner, P., 1994. A Sociology of Modernity. Liberty and Discipline. Routledge, London/New York.
Weber, M., 2010/1920. The Protestant Ethic and the Spirit of Capitalism. Oxford University Press, New York.
Zimmermann, B., 2011. Ce que travailler veut dire. Une sociologie des capacités et des parcours professionnels. Economica, Paris.
Zimmermann, B., Didry, C., Wagner, P. (Eds.), 1999. Le travail et la nation. Histoire croisée de la France et de l'Allemagne. Maison des Sciences de l'Homme, Paris.

Work Motivation

James M Diefendorff and Gina A Seaton, The University of Akron, Akron, OH, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The study of work motivation focuses on understanding why individuals pursue particular activities at work as well as the amount of effort exerted and duration of persistence in pursuing activities over time. Work motivation research attempts to understand how individuals allocate their limited resources to competing demands at work, describing when and why specific activities are pursued. In this article, we elaborate on the role of goals as a central organizing mechanism through which motivated action is observed and discuss a variety of approaches aimed at understanding motivated behavior at work.

Motivation is often defined as an internal force that directs, energizes, and sustains behavior over time and across changing circumstances (Kanfer, 1990). As such, motivation is observed in the choices that individuals make regarding which goal to pursue (i.e., direction), the amount of effort they invest in attaining a particular goal (i.e., the vigor or intensity of action), and how long they are willing to pursue a goal (i.e., persistence). Stated differently, motivation pertains to how individuals allocate their limited resources to competing demands over time, describing when and why specific activities attract individuals' attention and limited resources. The study of work motivation represents an attempt to understand motivated behavior in the context of organizations, which introduces a variety of boundary conditions and unique contingencies that can impact basic motivational processes (e.g., compensation systems, externally derived goals, job characteristics, distinct social relationships, norms, and organizational culture; see Diefendorff and Chandler, 2011).

Before describing specific motivation constructs and theories, it is important to distinguish motivation from other influences on behavior. It has become axiomatic in the organizational behavior literature that behavior is the result of three factors: ability, the situation, and motivation. Thus, when determining whether motivation is operating, it is necessary to rule out the influence of ability and the situation on action. An individual may be very motivated but fail to perform because of a lack of ability or because of defective tools or equipment. Alternatively, a person may have little desire to perform a behavior, but may do so because a strong situation coerces action (e.g., an assembly line) or because he or she has a great deal of knowledge and experience and can perform the task easily.

Much of the research on work motivation takes the concept of goal as the central organizing mechanism through which motivated action is observed. We begin our discussion of work motivation by describing goals, goal-setting theory, and constructs thought to reside in the 'hub' of work motivation (Locke, 1991) as well as the self-regulatory processes involved in striving for goals over time (Carver and Scheier, 1998). Next, we turn our attention to describing two approaches aimed at characterizing the structure and phases of behavioral self-regulation. Finally, we discuss content theories of motivation that focus on why individuals pursue different activities and the ways that these reasons for goal pursuit impact self-regulation. The discussion in this article is intended to touch on some of the dominant approaches to understanding work motivation. As such, it is necessarily brief and incomplete. More detailed and comprehensive reviews of work motivation can be found elsewhere (e.g., Diefendorff and Chandler, 2011; Latham, 2007; Lord et al., 2010; Kanfer, 1990; Kanfer et al., 2008).

Goals and the Motivational Hub

Goals

Goals can be defined as "internal representations of desired end-states" (Austin and Vancouver, 1996: p. 338), with these desired states varying along several dimensions, including importance/commitment (i.e., low vs high), difficulty (challenging vs easy), specificity (specific vs general), temporal range (short vs long), connectedness complexity (isolated, simple vs connected, complex), and level of consciousness (conscious vs unconscious). Further, goals are shaped by both personal characteristics and situational factors. Goals are the dominant focus of motivation research, and their importance can be seen in their pervasive presence in theories of motivation. Some authors argue that goals guide and organize nearly all human activity (e.g., Locke and Latham 1990).

An important tenet of many theories of motivation is that goals are hierarchically arranged (e.g., Carver and Scheier, 1998; Lord et al., 2010). Near the top of this hierarchy exist superordinate goals, which tend to (1) extend relatively far into the future, (2) be relatively stable, and (3) be grounded in core aspects the self, such as personality, needs, motives, and values. Near the bottom of the goal hierarchy exist action goals that represent concrete behaviors individuals engage in at any given point in time. This level of the goal hierarchy is often the focus of employees' attention and regulatory resources during the workday.

As one moves down the goal hierarchy, each successive layer of subordinate goals is more tactical, concrete, and short term. Each subgoal can be considered a strategy (i.e., a means) for achieving the goal above it in the hierarchy. One's confidence for attaining a goal is influenced by the ease with which the subgoals needed to reach the higher order goal can be performed. Additionally, the value of a lower-level goal is determined, in part, by the strength of its relationship to important

 http://dx.doi.org/10.1016/B978-0-08-097086-8.22036-9

higher level goals. Thus, goals are best understood as embedded in a goal hierarchy with their characteristics being tied to their position in the goal hierarchy and the strength of their ties to other goals in the hierarchy.

Goal-Setting Theory

Given the large number of ways that goals can differ, research has examined how different goal dimensions affect performance. One approach to systematically studying the effects of different goal dimensions is goal-setting theory (Locke and Latham, 2002). Goal-setting theory is one of the most influential theories in the study of work motivation, describing how goal characteristics influence performance through the mechanisms of attentional focus, effort, persistence, and strategy development. A key finding from goal-setting research is that difficult, specific goals that are accepted result in better performance than do-your-best or easy goals (Locke and Latham, 2002). Goals with these characteristics are beneficial in that they reduce ambiguity and require individuals to expend more effort, better focus their attention, and persist longer over time. Moreover, as tasks become more complex, difficult and specific goals also produce more strategy development, which can enhance performance.

There are several boundary conditions and caveats surrounding the prescription of assigning difficult and specific goals. First, goal difficulty is determined relative to each individual's capability, with optimal effort exertion occurring when there is a good match between the demands of the task and an individual's skill (Locke and Latham, 1991). Thus, an assigned goal that is determined to be normatively difficult may overshoot the optimal difficulty level for some individuals and undershoot it for other individuals, both of which can result in lower motivation. Second, research has shown that goal setting must be accompanied by feedback that allows individuals to monitor their progress and make necessary adjustments to their goal-directed behavior (Locke and Latham, 1990). Goal environments that are not feedback rich can prevent goal-setting interventions from producing their intended benefits. Finally, research has shown that difficult, specific goals may harm performance when they are assigned for novel or complex tasks as they may inhibit needed strategy development (e.g., Kanfer and Ackerman, 1989). When first working on a task (i.e., early in the skill acquisition process), not assigning a goal may actually be beneficial as it can facilitate learning and strategy development, resulting in a better long-term performance.

An Expanded Motivational Hub

Locke (1991) described the motivational hub as containing goals, performance, and self-efficacy, and suggested that this is "where the action is" (p. 296) for goal pursuit. Building on these ideas, we contend that expectancy theory constructs (expectancy, valence, and instrumentality) and affect may also be a part of this motivational hub, as these dynamic constructs are closely tied to goals and have been shown to impact goal selection and goal striving. We describe each of these constructs and their links to work motivation below.

Valence, Instrumentality, and Expectancy

According to Vroom's (1964) valence, instrumentality, and expectancy (VIE) theory, when individuals are deciding which activity to pursue they evaluate the expectancy, valence, and instrumentality of the possible activities. These three assessments are thought to combine to represent the overall motivational force for a goal. The goal with the highest motivational force is then adopted for pursuit. *Expectancy* is the perceived likelihood that expending effort will lead to goal attainment (Van Eerde and Thierry, 1996), *valence* is the desirability, attractiveness, importance, or anticipated satisfaction with outcomes associated with a goal (Van Eerde and Thierry, 1996), and *instrumentality* is the perceived likelihood that goal attainment will lead to desired outcomes (i.e., if I perform at a particular level, I will receive the associated rewards; Vroom, 1964). Each of these judgments contributes to an individual's motivation for a given task; if any of the three judgments are zero, no motivation is said to exist.

VIE theory variables have received a great deal of attention in the literature, though support for the theory varies as a function of the level of analysis. Van Eerde and Thierry's (1996) meta-analysis found weak support for the theory when the tests were at the between-person level of analysis (i.e., comparing how individuals differ in expectancy, valence, and instrumentality for an action). However, the results were stronger for predicting preferences and effort when the analyses were restricted to the more appropriate within-person level of analysis (i.e., comparing how actions differ in expectancy, valence, and instrumentality within the person). In other words, empirical tests that matched the underlying theory (i.e., choosing among goals vs examining who is more or less motivated to pursue a particular goal) provided stronger support for the theory.

Self-Efficacy

Self-efficacy is a person's belief that he or she can perform the actions needed to attain a specified level of performance (Bandura, 1997), and is conceptually related to the expectancy concept in VIE theory. Self-efficacy is shaped by a variety of factors, with past experiences having the strongest influence (Bandura, 1997). Self-efficacy is a center piece of Bandura's (1997) social cognitive theory and was incorporated into Locke and Latham's (1990) goal-setting theory. As a result, self-efficacy is one of the most commonly studied predictors of task-specific motivation. In general, self-efficacy has been shown to have positive effects on a variety of motivational and performance outcomes across a large number of studies and research contexts (e.g., Stajkovic and Luthans, 1998). However, a stream of research in recent years has challenged the causal role of self-efficacy in positively influencing performance, with some researchers finding that self-efficacy is primarily a reflection of past performance (and does not *cause* future performance; Heggestad and Kanfer, 2005) and that the causal role it plays in influencing effort and performance can sometimes be negative (e.g., Vancouver and Kendall, 2006).

Recent studies have attempted to reconcile these conflicting findings. For instance, Vancouver et al. (2008) theorized and found support for a nonmonotonic and discontinuous model in which self-efficacy has positive effects up to the decision to pursue a task and then has negative effects as self-efficacy

further increases. That is, an increase in self-efficacy can lead individuals who would not pursue a goal to suddenly switch and decide to pursue the goal (producing a positive effect). However, after the person has adopted the goal, further increases in self-efficacy may lead individuals to become overconfident and, as a result, not exert as much effort in the pursuit of the goal, resulting in a negative effect of self-efficacy on performance.

Other research has also attempted to further clarify these conflicting effects. For instance, Schmidt and DeShon (2009) found that challenging circumstances led to a positive relationship between self-efficacy and performance (i.e., very confident individuals rose to the challenge following poor performance), whereas less challenging situations resulted in a negative relationship (i.e., very confident individuals did not exert as much effort as individuals with low confidence following successful performance). Further, Schmidt and DeShon (2010) found that ambiguity moderated the self-efficacy and performance relationship, such that under conditions of high ambiguity a negative relationship was observed, but under low ambiguity a positive relationship was observed. Together, these studies suggest that the self-efficacy and performance relationship is complex and that there are conditions under which high self-efficacy can be beneficial or harmful. More work on this interesting and complex topic is needed.

Affect

Motivation and affect are inextricably linked. Affect is thought to influence motivation both indirectly through cognitive processes and directly through unconscious processes. Seo et al. (2004) present a theoretical model in which these processes are explicated. In relation to indirect effects, they theorized that when evaluating goals or performance (especially in ambiguous situations) individuals use their current affective state to make judgments regarding expectancy, utility, and progress. Positive affect is thought to result in more favorable goal-related judgments as compared to negative affect. Consistent with this idea, Erez and Isen (2002) found that for more difficult performance goals, individuals in a positive mood had higher expectancy, valence, and instrumentality than individuals in a neutral mood. Further, Carver and Scheier (1998) argued that positive affect may signal that goal progress is adequate, leading individuals to reduce their subsequent effort. In contrast, negative affect may signal that goal progress is not sufficient and that greater goal-direct effort is needed.

Seo et al. (2004) also theorized that there are situations in which motivational processes are not mediated by conscious judgments. Instead, they argue that affect itself can be a motivational force that directly energizes an individual's goal-directed behaviors. Indeed, research has suggested that positive and negative affect trigger different motivational orientations, with positive emotions showing greater alignment with approach motivation and negative emotions show greater alignment with avoidance motivation (e.g., Neumann and Strack, 2000). Additionally, some research has shown that unconscious affective priming can shape goal-directed behavior (Aarts et al., 2007). The continued integration of affect and motivation processes represents an exciting direction for work motivation research.

Self-Regulation Approaches to Understanding Work Motivation

Now that we have described goals and constructs in the motivational hub, we turn our attention to the regulation of goal-directed behavior over time. Motivation is a dynamic process in which individuals self-regulate their goal-directed activities (Diefendorff and Chandler, 2011). Consistent with the definition of motivation provided at the beginning of this article, the self-regulation of goal-directed activities occurs over time and across changing circumstances. As such, many models of work motivation adopt a self-regulation perspective to characterize this process. Self-regulation may be described in structural terms that explain the interrelationship of self-regulation components or in terms of distinct phases in the self-regulation process. We elaborate on each of these perspectives below.

Structure of Self-Regulation

The basic structure of self-regulation can be characterized using the principles of control theory (e.g., Carver and Scheier, 1998). At the center of control theory is the negative feedback loop. As shown in **Figure 1**, the negative feedback loop includes a standard (reference value or goal a person is trying to attain), a comparator (for matching the standard against performance), an input function (for sensing current performance), and an output function (for changing behavior). The negative feedback loop operates as a continuous process in which individuals monitor and sense discrepancies between current and desired states and engage in behaviors to reduce any observed discrepancies. If no discrepancy is sensed, the person is expected not to make any changes to behavior. If performance is lower than the goal (i.e., a negative discrepancy), the output function is engaged, resulting in greater effort expenditure so as to bring performance into line with the goal and reduce the discrepancy. If performance is higher than the goal (i.e., a positive discrepancy), effort on the focal goal may be reduced and resources may directed to another goal that is awaiting attention. This discrepancy detection and reduction process is thought to operate continuously and automatically over time as conditions in the environment change (Carver and Scheier, 1998).

Phases of Self-Regulation

In addition to describing the cyclical process of self-regulation, some researchers have broken self-regulation into distinct phases and examined how motivational, cognitive, and affective processes differ across these phases. One common way to divide the process is into the phases of goal setting, goal striving, and goal revision (see Diefendorff and Lord, 2008, for more discussion of self-regulatory phases).

Goal Setting

Self-regulation begins with goal selection (or commitment to an assigned goal). The act of choosing or accepting a goal creates a discrepancy that individuals are motivated to reduce. Selecting an action goal involves choosing among different tasks to pursue (e.g., finish reading a book vs finish writing a report) or the level of performance to aspire to within a task

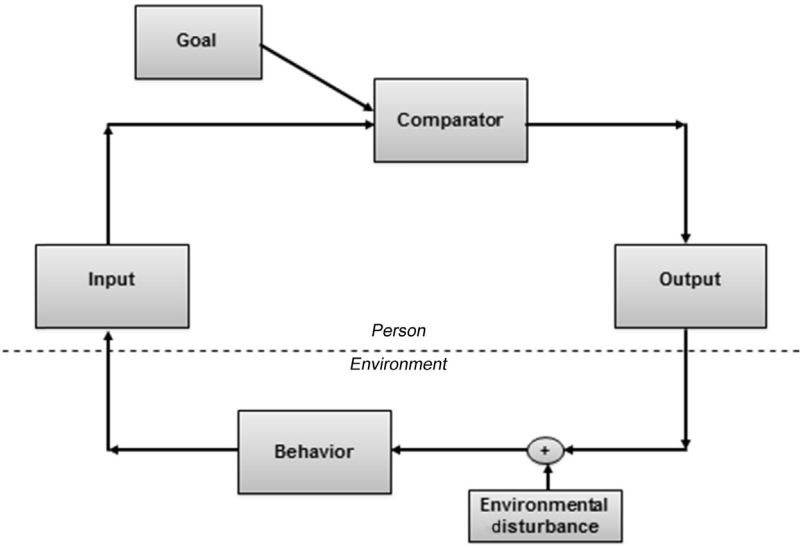

Figure 1 Structural model of self-regulation: the negative feedback loop.

(e.g., write five pages vs write ten pages), which can be done according to the VIE theory principles outlined above. The selection of an action goal can involve both conscious and nonconscious processes that take into account stable and dynamic aspects of the person (e.g., long-term motives, short-term affect) and situation (e.g., organizational culture, fluctuations in work processes).

Goal Striving

After an action goal is chosen, individuals must put forth effort to attain the goal. Effort refers to the amount of mental or physical resources devoted to a task. The more effort allocated to a task, the more the person is said to be motivated. Effort has been operationalized in a variety of ways including cardiovascular reactivity, resource allocation, and subjective experience of exerting effort. Some researchers have argued that persistence (another hallmark of motivation) essentially reduces to the same concept of effort in that both pertain to the amount of resources dedicated to a task during a period of time (Dalal and Hulin, 2008).

Individuals regulate their effort during goal striving by not only comparing their current state against the desired state, but also by matching their current rate of discrepancy reduction against their desired rate of discrepancy reduction (Carver and Scheier, 1998). In effect, individuals compare their actual rate of progress with a desired rate of progress and experience distinct affective states based on this comparison. If progress is faster than expected, individuals may experience a sense of calm or happiness, which leads to a reduction in effort on the task and allows the person to devote resources to other things. In contrast, if progress is slower than expected, the individual may experience anxiety and increase effort on the task. Goal striving continues until the goal is reached or a predetermined stop time is reached (i.e., a deadline).

Goal Revision

At the end of a goal-striving episode, individuals once again compare their performance against the goal to determine whether the goal was attained (Carver and Scheier, 1998). If the goal was attained, individuals may set a new goal within the same domain or decide to pursue a goal in a different domain. If the goal was not attained, individuals must make a choice on how to proceed. For example, individuals may decide to keep the goal at its current level and try again, they may decided to revise it (upward or downward), or they may abandon the goal altogether and pursue a new and/or different goal. Of course, in many work environments where competing demands are plentiful, individuals may also decide to table the goal and momentarily direct attention to another activity, returning later to the decision of whether and in what way to revise the goal.

In general, research investigating goal revision has suggested that individuals with large, negative discrepancies (i.e., one's performance was below the goal) tend to revise their goal downward in future goal-striving episodes, whereas individuals with positive discrepancies (i.e., one's performance was above the goal) tend to revise their goals upward. In addition, these relationships may be affected by both the time left to pursue the goal and the causal attributions that individuals make for performance (Donovan and Williams, 2003).

Content Approaches to Understanding Work Motivation

In contrast to focusing on the motivational hub or the structure and phases of self-regulation, *content theories* of motivation emphasize the ways in which motivation may differ between individuals and the effects of these differences on self-regulation. Content theories focus on why individuals pursue activities and how the reasons for goal pursuit alter the ways in which self-regulation proceeds. Several theories may fall into the category of content approaches to motivation, but a few have been frequently applied to organizational contexts and are discussed in more detail here: Deci and Ryan's (2000) self-determination

theory (SDT), Dweck's (1986) goal-orientation theory, and Higgins' (1997) regulatory focus theory.

Self-Determination Theory

SDT (Ryan and Deci, 2000) emphasizes the satisfaction of basic human needs for autonomy, competence, and relatedness as a key driver of motivated behavior. In contrast to many theories of motivation, which emphasize the quantity of motivation (e.g., goal-setting theory), SDT emphasizes the idea that qualitative differences in motivation also matter. Specifically, a key contribution of SDT is that it describes a continuum in which goals may be pursued for different reasons, ranging from being intrinsically motivating at one extreme to being extrinsically motivating at the other end. The theory suggests that these different motivations can all lead individuals to act, but they may have different implications for performance and well-being.

According to SDT, intrinsically motivating tasks are interesting, enjoyable, and spontaneously pursued by individuals. Such motivation is experienced as autonomous and self-determined (Deci and Ryan, 2000). In contrast, extrinsically motivated tasks are pursued because of external contingencies, but can be experienced as controlling or autonomous depending on the extent to which the contingencies have been internalized by the individual. As such, SDT describes several different types of extrinsic motivation. External regulation is the least internalized form of extrinsic motivation with individuals pursuing tasks because of rewards or punishments present in the environment. Introjected regulation is the next form of extrinsic motivation, with individuals pursuing goals because they have internalized and come to self-administer the rewards and punishments for the goal. Often, individuals who experience introjected motivation pursue tasks because they would feel guilty if they did not. Identified regulation reflects a more internalized form of extrinsic motivation whereby individuals understand the value of the goal and pursue it based on that understanding, but they have not come to personally value the goal and the reasons for its pursuit. Integrated regulation reflects the most complete internalization of an extrinsic goal whereby individuals have come to personally value the goal and experience it as important. The goal is not inherently interesting or fun (so it is not intrinsically motivating), but it is valued, important, and experienced as autonomously motivating. Finally, in contrast to the motivated behaviors described above, SDT also recognizes that some behaviors may be amotivated, which means that individuals are not able to provide a reason for why they engage in them.

Ryan and Deci (2000) argued that some goals are more consistent with satisfying an individual's basic needs than others. In particular, striving for autonomous goals may lead to greater need satisfaction and better well-being, whereas striving for controlled goals may lead to the thwarting of one's needs and worse well-being. A key finding from research on SDT is that more autonomous forms of motivation are associated with better performance on interesting or complex tasks. However, differences between autonomous and controlled motivation are not observed for boring or mundane tasks (Koestner and Losier, 2002). Further, autonomous motivation is positively associated with well-being and job attitudes. Thus, both forms of motivation may lead individuals to perform a behavior, but autonomous motivation better satisfies psychological needs and leads to greater well-being compared to controlled motivation.

Goal Orientation and Implicit Person Theories

Goal orientation refers to differences in how people interpret and respond to achievement situations (Dweck, 1986). Research has identified two distinct goal orientations: performance and mastery. Someone high in performance goal orientation focuses on demonstrating ability and defines success or competence by comparing oneself to others, either by demonstrating competence relative to others (performance prove orientation) or by avoiding displays of incompetence relative to others (performance avoid orientation; Dweck, 1986). In contrast, an individual with a mastery goal orientation focuses on developing skills and abilities and defines success or competence against internal standards (Dweck, 1986). The commonly accepted view is that a mastery orientation is associated with an adaptive behavioral pattern (e.g., selecting challenging tasks, setting difficult goals, persisting in the face of obstacles), whereas performance approach can be adaptive or maladaptive depending on the circumstances, and performance avoid is maladaptive.

Mastery and performance goal orientations are grounded in implicit person theories (Dweck, 1986). According to this view, individuals vary in the degree to which they view attributes and abilities as malleable (as opposed to stable and enduring). A belief that human attributes are malleable and that effort and experience can improve one's standing reflects an incrementalist implicit person theory. A belief that human abilities are fixed and that effort and experience cannot improve the situation reflects an entity implicit person theory (Dweck and Grant, 2008). Research has shown that implicit theories are associated with differences in reactions to failure. For entity theorists, failure is viewed as a reflection of the self and thus, is experienced as threatening. Accordingly, these individuals tend to engage in strategies aimed at preventing or hiding failure (e.g., choosing easy tasks, self-handicapping, or withdrawing from tasks; Dweck and Grant, 2008). In contrast, incremental theorists view failure as an opportunity to learn (Dweck and Grant, 2008). As a result, incremental theorists enjoy learning new things because they view them as a way to increase their ability. Consistent with these ideas, Robins and Pals (2002) found that entity theorists adopted performance goals and displayed less adaptive behavioral patterns when faced with challenge, whereas incremental theorists adopted learning goals and displayed more adaptive behavioral patterns when faced with challenge.

Regulatory Focus Theory and Approach/Avoidance Motivation

According to regulatory focus theory, goal-directed behavior operates differently for individuals high in promotion focus compared to individuals high in prevention focus (Higgins, 1997). A promotion focus refers to a motivational strategy that is approach oriented and concerned with positive outcomes, accomplishments, gains, and aspirations. A prevention focus, on the other hand, refers to a motivational strategy that is avoidance oriented and concerned with safety,

duties, and avoiding negative outcomes. For prevention-focused individuals, goals are seen as obligations, rather than desired standards. In this way, promotion-focused individuals seek to reduce discrepancies between actual and 'ideal' selves, whereas prevention-focused individuals seek to minimize discrepancies between their actual and 'ought' selves.

Individuals with these different motivational foci also exhibit distinct behavioral patterns and experience different emotions in the goal-striving process. For example, individuals with a promotion focus tend to experience eagerness when goal striving, feel joy when goals are attained, and feel sadness when goals are not attained (Brockner and Higgins, 2001). On the other hand, prevention-focused individuals are cautious during goal striving, feel relaxed when goals are attained, but feel nervous when goals are not attained (Brockner and Higgins, 2001).

Consistent with the approach/avoid distinction outlined by regulatory focus theory, Gray (1982) theorized that behavior is derived from two biologically based motivational systems: the behavioral activation system (BAS), an appetitive or approach system, and the behavioral inhibition system (BIS), an aversive or avoidance system. High BAS sensitivity is reflected by a heightened reactivity to signals of reward or nonpunishment and a perceptual readiness and strong emotional responsiveness to such stimuli (Gray, 1982). High BAS sensitivity is associated with under arousal, impulsivity, a strong drive to attain goals, and the experience of positive emotions. High BIS sensitivity, on the other hand, is reflected by responsiveness to signals of punishment or nonreward and a strong sensitivity to and perceptual readiness for negative stimuli. High BIS sensitivity is associated with low levels of behavioral activity and the experience of negative emotions. Consistent with this theory and research, Elliot and Thrash (2002) found that BIS sensitivity, neuroticism, and negative affectivity loaded on a common avoidance latent trait, and BAS sensitivity, extraversion and positive affectivity formed a common approach latent trait. Additionally, the authors found that these latent traits predicted the selection of achievement goals (i.e., mastery, performance approach, and performance avoid) in theoretically consistent ways (Elliot and Thrash, 2002).

Conclusion

The study of work motivation has seen significant advances in recent years in our understanding of motivational processes, antecedents, and outcomes. In this Encyclopedia article, we have highlighted what we believe to be some of the most influential concepts and perspectives in this area of study. However, it is important to note that we have only scratched the surface of empirical findings pertaining to the field of work motivation. Indeed, researchers have explored a myriad of issues that fall outside the scope of this article.

Although the field has made significant strides in terms of understanding motivation at work, there is also still much to be learned. Some new and exciting directions to take the field of work motivation include examining: (1) the role of conscious (i.e., self-attributed) and unconscious (i.e., implicit) motivational processes, (2) the substantive role of time in shaping motivational processes (e.g., effects over time and within-person dynamics), (3) group motivation, and (4) integration of multiple theoretical perspectives, such as the investigation of structural models of self-regulation with content theories like SDT. The advances in work motivation research outlined in this article, coupled with emerging themes in the study of motivation, make it clear that this area of study is vibrant and will remain so in the coming years.

See also: Avoidance and Approach Motivation: A Brief History; Emotions and Work; Expectancy-Value-Cost Model of Motivation; Mastery Learning; Motivation and Actions, Psychology of; Self-Determination Theory; Self-Efficacy; Self-Regulated Learning.

Bibliography

Aarts, H., Custers, R., Holland, R.W., 2007. The nonconscious cessation of goal pursuit: when goals and negative affect are coactivated. Journal of Personality and Social Psychology 92, 165–178.

Austin, J.T., Vancouver, J.B., 1996. Goal constructs in psychology: structure, process, and content. Psychological Bulletin 120, 338–375.

Bandura, A., 1997. Self-Efficacy: The Exercise of Control. Freeman, New York.

Brockner, J., Higgins, E.T., 2001. Regulatory focus theory: implications for the study of emotions at work. Organizational Behavior and Human Decision Processes 86 (1), 35–66.

Carver, C.S., Scheier, M.F., 1998. On the Self-Regulation of Behavior. Cambridge University Press, New York.

Dalal, R.S., Hulin, C.L., 2008. Motivation for what? A multivariate dynamic perspective of the criterion. In: Kanfer, R., Chen, G., Pritchard, R. (Eds.), Work Motivation: Past, Present, and Future. Taylor and Francis Group, New York, pp. 63–100.

Deci, E.L., Ryan, R.M., 2000. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry 11 (4), 227–268.

Diefendorff, J.M., Chandler, M.M., 2011. Motivating employees. In: Zedeck, S. (Ed.), Handbook of Industrial and Organizational Psychology. American Psychological Association, Washington, DC, pp. 65–135.

Diefendorff, J.M., Lord, R.G., 2008. Self-regulation and goal striving processes. In: Kanfer, R., Chen, G., Pritchard, R. (Eds.), Work Motivation: Past, Present, and Future. Taylor and Francis Group, New York, pp. 63–100.

Donovan, J.J., Williams, K.J., 2003. Missing the mark: effects of time and causal attributions on goal revision in response to goal-performance discrepancies. Journal of Applied Psychology 88, 379–390.

Dweck, C.S., 1986. Motivational processes affecting learning. American Psychologist 41, 1040–1048.

Dweck, C.S., Grant, H., 2008. Self-theories, goals, and meanings. In: Shah, J.Y., Gardner, W.L. (Eds.), Handbook of Motivation Science. The Guilford Press, New York, pp. 405–416.

Elliot, A.J., Thrash, T.M., 2002. Approach-avoidance motivation in personality: approach and avoidance temperaments and goals. Journal of Personality and Social Psychology 82 (5), 804–818.

Erez, M., Isen, A.M., 2002. The influence of positive affect on the components of expectancy motivation. Journal of Applied Psychology 87, 1055–1067.

Gray, J., 1982. The Neuropsychology of Anxiety. Oxford University Press, Oxford, England.

Heggestad, E.D., Kanfer, R., 2005. The predictive validity of self-efficacy in training performance: little more than past performance. Journal of Experimental Psychology: Applied 11, 84–97.

Higgins, E.T., 1997. Beyond pleasure and pain. American Psychologist 52, 1280–1300.

Kanfer, R., 1990. Motivation theory and industrial and organizational psychology. In: Dunnette, M.D., Hough, L.M. (Eds.), Handbook of Industrial and Organizational Psychology. Consulting Psychologists Press, Palo Alto, CA, pp. 75–170.

Kanfer, R., Ackerman, P.L., 1989. Motivation and cognitive abilities: an integrative/aptitude treatment interaction approach to skill acquisition. Journal of Applied Psychology 74, 657–690.

Kanfer, R., Chen, G., Pritchard, R.D., 2008. Work Motivation: Past, Present, and Future. Taylor and Francis Group, New York.

Koestner, R., Losier, G.F., 2002. Distinguishing three ways of being highly motivated: a closer look at introjection, identification, and intrinsic motivation. In: Deci, E.L., Ryan, R.M. (Eds.), Handbook of Self-Determination Research. University of Rochester Press, New York, pp. 101–121.

Latham, G., 2007. Work Motivation: History, Theory, Research, and Practice. Sage Publications, Inc., Thousand Oaks, CA.

Locke, E.A., 1991. Goal theory vs. control theory: contrasting approaches to understanding work motivation. Motivation and Emotion 15, 9–28.

Locke, E.A., Latham, G.P., 1990. Work motivation and satisfaction: light at the end of the tunnel. Psychological Bulletin 90, 125–152.

Locke, E.A., Latham, G.P., 1991. Self-regulation through goal setting. Organizational Behavior and Human Decision Processes 50, 212–247.

Locke, E.A., Latham, G.P., 2002. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. American Psychologist 57, 705–717.

Lord, R.G., Diefendorff, J.M., Schmidt, A.M., Hall, R.J., 2010. Self-regulation at work. In: Fiske, S.T. (Ed.), Annual Review of Psychology. Annual Reviews, Chippewa Falls, WI, pp. 543–568.

Neumann, R., Strack, F., 2000. Approach and avoidance: the influence of proprioceptive and exteroceptive cues on encoding of affective information. Journal of Personality and Social Psychology 79, 39–48.

Robins, R.W., Pals, J.L., 2002. Implicit self-theories in the academic domain: implications for goal orientation, attributions, affect, and self-esteem change. Self and Identity 1, 313–336.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Schmidt, A.M., DeShon, R.P., 2009. Prior performance and goal progress as moderators of the relationship between self-efficacy and performance. Human Performance 22, 191–203.

Schmidt, A.M., DeShon, R.P., 2010. The moderating effects of performance ambiguity on the relationship between self-efficacy and performance. Journal of Applied Psychology 95, 572–581.

Seo, M., Feldman, B.L., Bartunek, J.M., 2004. The role of affective experience in work motivation. Academy of Management Review 29, 423–439.

Stajkovic, A.D., Luthans, F., 1998. Self-efficacy and work-related performance: a meta-analysis. Psychological Bulletin 124, 240–261.

Van Eerde, W., Thierry, H., 1996. Vroom's expectancy models and work-related criteria: a meta-analysis. Journal of Applied Psychology 81, 575–586.

Vancouver, J.B., Kendall, L.N., 2006. When self-efficacy negatively relates to motivation and performance in a learning context. Journal of Applied Psychology 91 (5), 1146–1153.

Vancouver, J.B., More, J.M., Yoder, R.J., 2008. Self-efficacy and resource allocation: support for a nonmonotonic discontinuous model. Journal of Applied Psychology 93 (1), 35–47.

Vroom, V.H., 1964. Work and Motivation. Wiley, New York.

Work, Sociology of

Frederick C Gamst, University of Massachusetts, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Presented is the centrality of work in the human existence, including work's personal and social dimensions. Anglo-Saxon folk and intellectual thought about work exclude much laboring activity. Reviewed are work's purpose, gain, control, temporal synchronizations, and monetization. Work is an adult's principal mooring in society. Discussion includes employee participation, social control of 'progress,' division of labor, gender, industrial policy, computerization, alienation, deskilling, training, outsourcing, unionization, retirement, the 'new economy,' and the 'the end of work.'

Work is central to the panhuman existence. Work is above all a sociocultural phenomenon. Humans engaged in work are primarily *Homo sociologicus* rather than *Homo economicus* (Dahrendorf, 1983 [1968]). For providing livelihood, generating creations, and producing other outcomes, work concerns culturally patterned mental, physical, and emotional exertions. The sociology of work investigates the opportunities and constraints of work. Investigation covers the broad, patterned context for work as it encompasses individuals, supports communities, shapes nations, and binds countries. Workers in modern society spend much of their waking time on the job and in its commutation (Gamst and Helmers, 1991).

'Modern society' has some degree of an ideological and organizational basis of bureaucratic industrial capitalism or socialism, both influenced by the global market. Modern bureaucracies have both efficiencies and inefficiencies for controlling work. In modern society, to varying extents purpose for an organization is negotiated both internally from its members and externally from outside organizations including those governmental. Work organizations include purposeful for-profit businesses, governmental units, labor unions, schools, churches, charities, advocacy groups, and military bodies. A work organization has a 'formal' aspect, where action accords with its rules, procedures, and policy, and an 'informal' aspect, where members' loose alliances and adaptable partnerships influence performance of actions. Regarding everyday life in all societies, the sociology of work is an instructive subject for extensive reflection in conceptualization and manifold relevancy in practice.

In the twin disciplines of sociology and sociocultural anthropology, the study of work has long been a central interest, and it pervades much of the disciplines' published research (Abbott, 1993; Gamst, 1977, 1995; Kalleberg and Berg, 1987; Wallman, 1979). In this article the product of both disciplines has the label 'sociological.' 'Work' has perhaps the widest meaning of any word in English. It is a culture-bound, scholarly concept and folk idea. This article examines value-laden meanings and dimensions of 'work.' Every society (and subsociety) has particular patterns of rewards for work and distributions of gains from work. Not all work has factors of production institutionalized in price-making markets for labor, products, and land.

Selected History of the Sociology of Work

Few ethnographic monographs about nonwestern societies exclude information on work. Into the twenty-first century, the twin disciplines have depicted the centrality of work in community studies in the United States and Canada. Employment and its access are a vital interest of the fieldworker studying a community, often from an ecological perspective. Pioneering research on work and workers in North America apart from community studies includes that of Whiting Williams, beginning in 1918, on the human relations of employees and employer (Williams, 1918). Earlier, Emile Durkheim (1947 [1902]: 13) discussed rules that "fixed the responsibilities of employers and workmen, as well as the duties of employers toward each other." Durkheim's ideas have expanded into a French sociology of work (Bouvier, 1994; Tanguy, 2008). Castillo (1997) cautions against relying solely on English-language studies of work and reviews other pertinent sources (see also Helmers, 1993; Spittler, 2008).

With the advent of the experiments and observations regarding workers in the Hawthorne plant of Western Electric, by 1930, the twin disciplines entered their phase of the social science of industrial work, then often having a social psychological bent (Carey, 1967). Studies of work received legitimization in 1946, when the American Sociological Association formed its Section on Industrial Sociology and again, in 1951, when Delbert Miller and William Form wrote the first textbook, *Industrial Sociology* (Miller and Form, 1951). Originally, work-studies were a variety of empirical emphases not having a characteristic conceptual framework or methodology. Workers display the ordinary range of social behavior.

Organizations of work, from corporations to cliques on their jobs to peasant and forager work groups, are components of the greater human society. In studying modern societies, work relations are part of most conceptualizations of social structures and role information. Researchers frequently collect information in a workplace. Sociological study of work gradually diverged into a number of conceptual and methodological directions. Nevertheless, organizations remain a focus of the study of work (Perrow, 1986).

Occupational sociology's main interest is in studies of work (Pavalko, 1988; Volti, 2008). Occupations are purposive and are enacted by individuals or groups – either a singleton craftsperson, artist, or consultant or a socially linked number of like positions. Occupational hazards are a growing concern of research (Nichols, 2005).

Research on work broadened to encompass many emphases and overlapping subtopics, from gender issues to division of labor. Work research gradated into investigations conducted across disciplines, such as on organizations, social control,

industrial relations, and role theory. Organizational role theory examines role behavior in the workplace, often to increase productivity. Given all the research, what is 'work'?

Work as Culture-Bound Idea

Knowledge does not exist supra culture; thus, in the English-speaking realm, one must first examine the Anglo-Saxon cultural context of work. Language does not determine cultural outcomes. But it is a 'color filter' through which we perceive the world in expressions of our native technologies of work and their procedures, values and norms, produced property, and distributions of the fruits of labor. Work is the broadest English word for both the process of doing something and the state of something done. Work is viewed ambivalently. It is a moral necessity and the deity's curse of travail upon humankind ('in the sweat of thy face…'). As etymologically rooted and in folk views, the conception of work is broadly viewed as concerning all endeavors.

Many social scientists, however, attempt an analytically useful narrowing of the definition to include the creating of goods and services valued in a society. But valued by whom? As Melville Herskovits (1952) explains, the phenomena of value regarding things exchanged are understandable only as a part of a specific culture or a subculture. Adriano Tilgher (1958) says that because economic value is always relative, so is assessing the work creating it.

The social value of work, then, depends upon the socialized view of the observer, varying within a specific society. Managers of antebellum southern railroads and their corporate slaves had differing views of the nature of value created.

As with other issues in the sociology of work, for example the management-oriented studies spawned by Hawthorne, the definition of value of work resolves into whose ox is being gored or fed (Carey, 1967). A societal disruption, however, could be eufunctional for at least part of a society. Thus, an intensive strike after which workers lose their employment is not always dysfunctional to labor overall. Locomotive engineers and switchmen decisively lost the bitter strike on the Burlington road in 1888. Management's 'victory,' however, proved so costly to shareholders that other railroads began to recognize unions. If not definable by value, work has useful perspectives in various dimensions.

The Dimensions of Work

Work has a 'personal dimension' in defining the self, necessarily in relation to roles of others, and in obtaining positive affect from community members (Watson, 2008: pp. 251–256; Budd, 2011: pp. 143–161; Schemien, 2012). This dimension of identity has a strong hold on all humans, in nonmarket and market societies. In Africa, the Wayto hippopotamus hunter seeks his villagers' recounting of his prowess as a killer of game; the Qemant plowman seeks community recognition as a good husbandman. In American inner cities, for some streetcorner men, job experiences and consequent job fears constantly intimidate them. Apprehensive about an opportunity to work, they stand overwhelmed by experience (Liebow, 1967). Work greatly affects the integration of personality. The personal dimension does not exist apart from the social dimension.

Work's 'social dimension' concerns interpersonal relations as much as productive output. Most behaviors at work are comprehensible at the collective level. Individual self-interest is dampened by behavioral prescriptions, including values, norms, attitudes, and etiquette. Moreover, in modern society, with its atrophied bonds of kinship and limited ties of community, work in groups provides opportunities for the close interactions that all primates require, including humanly exchanging positive affect and releasing tensions in a 'support group.'

Each society structures work in a construction of time. Even in so-called timeless foraging and agrarian societies, space could not be used effectively without some kind of time reckoning for coordinating work. A Wayto fishes at the appropriate hour and hunts during seasonal movements of game. A Qemant plowman looks at the sun and reads time to within half an hour, sufficient for his community's needs. To coordinate industrial transportation, North American railroads developed standard time zones. Factory shifts must be organized by the time clock. For modern needs, scientists created the microsecond. Timing in industrial societies refers to temporal modes, including frequency, synchronization, and duration. Without such timing, society's work becomes dysfunctionally independent instead of functionally interdependent. The clock, more than the steam engine, was the serpentine machine in the garden of agrarian societies.

Anglo-Saxon folk and intellectual thought about work exclude much laboring activity. Some activities are not considered work when the time and exertion seem sporadic and little planned, the activities are only somewhat related or are unrelated to an institutionalized product (purpose), or the activities are commonplace. For utility as theory, a concept of human work must be valid across cultures and not bound by Anglo-Saxon patterns. When looking for the meaning of 'work,' it must be determined whether work must be purposeful and gainful and whether livelihood is viewed cross-culturally.

Work as Concept

A biological exertion involving energetic activity in motor acts coupled to thought has expression in culturally patterned work. A purely mental activity such as planning a task is work. Work as solely a motor act, apart from related thought, probably does not occur.

Because work is social, even when done in isolation from other humans, it is enmeshed in the collective conventions of a society. A person's work activities ordinarily intertwine with activities of others in role groups for producing, exchanging, and consuming goods and services. A work role is a cluster of attributes and expectations for a work status, a social position in a society occupied by a person. Even when work has performance distant from other persons in a group, it is enacted regarding the group.

Should exertion be 'purposeful' to qualify as work? Purpose overlaps with relative social value. The issue remains: purpose as defined from whose perspective? Canadians each perform

the same exertion and learned technique in felling a tree as an employed lumberjack, a tree farmer, a thieving 'timber pirate,' and a vandal. In economic parlance, the first two persons purposefully 'add value,' the third and fourth do not. To the extent that defining work depends on purpose, this characteristic is culturally relative.

Work could be defined as 'gainful' (simply, producing a gain) such as a livelihood earned in a nonmarket, transitional, or market economy (Applebaum, 1984). Broadly considered, gainful work does not require market exchange or money. Work is not always monetized and market-oriented (Williams, 2005). It could be coerced, as in slavery. The gain could be either an increment or a saving of something valued, incorporeal or material. It could be generalized reciprocity: a shaman reciprocates séances for sundry goods and services from community members. It could be collection-redistribution: a chief collects part of all goods and services produced and redistributes some to the chiefdom's members in kind and as political and other administration.

In modern societies, an individual has multiple statuses for 'gainful' work. A woman might be a homemaker rewarded by familial gratifications, a corporate executive paid in the market, a community volunteer given 'positive affect,' a coordinator of a food cooperative compensated in kind, and a member of a congregation striving for 'eternal life.' 'Livelihood' work supports a human's life (one's biological requisites) and life way (one's sociocultural requisites). With such a broad conception of livelihood, fewer kinds of exertion have exclusion as work.

Work in Modern Societies

Although a conception of work must be cross-culturally valid, much of its study concerns modern societies, which are industrial or industrializing, including an empirical, scientific, secular worldview and having dominant market exchange. In industrial society, few remain primary food or craft producers. People consume goods and services created outside their local communities. With an infinitesimally particularized division of labor, they gainfully work mainly for increments and savings to enhance their standards of consumption. Technological devices enormously multiply exertion in work with a consequently great personal productivity. By hoe, a Plains Indian, tilled 0.1 of an acre annually, but with machines, a Plains farmer tills 2000 acres annually. Although a concept of work does not depend on money, in today's market economy, it is the acme of activity. Sufficient money allows great independence from kinship bonds, community ties, and role expectations.

Because monetarily paid work is a central identifier of the modern self, the characteristic used as a person's overall social label and assessment is occupational status. After the ritualistic 'How do you do?' comes the probing 'What do you do?' Work furnishes a person's basis of social power, reputation based on achievement, and economic and other welfare.

In its concomitants, loss of modern work could mean the severing of one's principal mooring in society. The largest medium of monetary savings, the house, can be lost. In the United States, vital health insurance is terminated, and retirement credits are not accrued. Furthermore, the job delimits the kind of neighborhood along with the personal safety and property security obtainable. It determines the quality of housing inhabited. It influences the education and opportunity for children. It also restricts who will be friends, neighbors, and co-workers. Above all, the job allows a reassuring, predictable routine in an otherwise hectic life.

Social Control of Work

Coherence in scientific study rests on a central notion. In the twin disciplines, the notion might be 'social control,' their most widely used concept but about which no agreed definition exists. The term could mean an underlying social process, a social mechanism delimiting behavior, or a scholarly model for analysis. Most conceptions, however, concern norms (a culture's prescriptive 'shoulds,' 'oughts,' and 'don'ts') and the ways in which social organizations change while preserving some societal stability (Gibbs, 1994).

It is useful to conceive work organizations as 'contingent,' open systems in which no single managerial model is universally most efficient. Management suitably adjusts a work organization to the changing contingencies of the market, political pressures, peoples' demands, and technology by adjusting operations to its particular environment. Managers' adjustments form an organization's characterizing kinds of effectiveness by responding to contingencies (Donaldson, 2001).

Despite much writing about change in business organizations, McCabe (2007) finds that their patterns continue to be much the same as those of the past. Control maintains social relations that are not equally beneficial to the differing interests of all the involved actors, as depicted in the innumerable studies of work and reactions to controllers of work. Work organizations have formal and informal purpose and structuring of tasks and responsibilities. Informal social relations conflict variously with formal relations.

'Industrial relations' constitute contests for social control between employers and employees (sometimes with direct government involvement). Social control of work has a number of forms. These include bureaucratic, by formal rules and procedures; technological, by devices limiting or monitoring action; and collectively occupational, by grouped workers defining, in some part, the terms, conditions, and content of work. A pivotal question in studying control is, how are workers forced to act in ways not serving self-interest or worker-group interests? And what is the ill-defined boundary of their alternative, serving societal and private corporate interests?

Employee Participation

In the 'new industrial relations,' the strong control of work by managers is reclad by 'cooperative' programs, such as quality circles, team concept, and employee involvement. For unionized workers, participation in the programs necessarily broadens the character of their labor–management relations. A cooperative program, if fully developed and controlled through bilateral negotiations, could foster mutual benefits.

Participation programs, however, are sometimes created in trendy public relations (naively?) overlooking the opposing goals and highly consequential exercises of social control in the relations between labor and management. A large business organization possesses great social power, alone and in coordination with similar firms united in associations having great influence on government.

Unorganized employees, and sometimes those organized, are at a great disadvantage regarding power relations and abilities to exercise social control in the market. In its traditional adversarial role of countering union social control through cooperative programs, management can usurp union responsibilities and pose as a benefactor of employees. Cooperation could allow a transition from powerful collective to powerless individual bargaining for employees. In true employee participation, management must surrender some of its social control of workers. Genuine cooperation in work is fully bilateral, thus a partnership. It can be effected only by near equals in social control. This has not yet been the case. The global market pressures on wages and conditions of employment are inexorably downward.

Social Control of Technological 'Progress'

Business practice in North America includes a sectarian, 'laissez-faire' ideology promoting technological 'progress.' A tenet of such practice is the view that workers, their families, and their fellow community members must suffer the consequences of progress without redress or complaint. Technology, however, is not a neutral force or an impartial hand for advancement in an industry, for shaping society or for influencing an individual's life. Technological change in work receives guidance from visible hands, according to the goals of private firms, public agencies, and other organizations, including unions and advocacy groups. In business, only rarely does management effect social control solely through technology. Ultimate control comes through the bureaucratic mechanisms of a firm, thus from managerial norms. The high value North Americans place on technological change masks what seldom changes: the social controls shaping technology and its policy and, thereby, society.

Division of Labor and Gender

In foraging societies, occupational Jacks- and Jills-of-all-trades exist–hunters and gatherers. As technologies related to plant and animal domestication free some persons from primary food getting, a true division of labor, with occupational specialization, develops in addition to the universal natural division of labor by age and sex.

Despite the true division – including specialization of work technology within segments of a society or separation of a particular technology into its component techniques, or both – division by sex remains everywhere. A specific society assigns to each biological sex a number of gender roles that are mainly learned culture. The roles are normative results of socialization, including those regarding work. Not only do the conventional roles delimit what is 'male' and 'female' but also what is appropriate work for each sex.

Because a great overlapping range of somatic and mental characteristics exists for males and females, a similar range of tasks performed equally well by each sex obtains. Cross-culturally, almost no work roles are exclusively men's roles. Among the Hopi, women do the construction of multistory houses; among the Dahomey, women were the assault troops; and in many African societies, women have apical political statuses. Whatever a man can do, a woman can do equally well, with few exceptions that are related mainly to somatic strength.

Nevertheless, occupational and work-role sex typing is dominant even in industrial democracies. At Christmas-time, a boy customarily receives a pink nurse's kit and a girl, a black doctor's kit? Such socialization for conventional aspiring to statuses of work is multistranded throughout the fabric of a culture. Some feminist scholarship explores the relations among capitalism, patriarchy, and job segregation. Although these neo-Marxist studies are illuminating, they do not explain similar segregation in the patriarchal, plow-agricultural societies. These might additionally feature other sex typing, such as clitorectomies, infibulations, and familial honor-killings of females.

In industrial societies, occupational sex typing of women limits their kinds of work, levels of compensation, and ranges of beneficial concomitants of work. Restricted compensation opportunities foster women's dependence on men, with increased exploitation by men. Moreover, work prejudice against kinds of people restricts societal efficiency.

Focus on the workplace ignores the importance of an unpaid wife at home supporting a successful husband at work. Such economic exchanges between social unequals are at odds with the folk category of the self-made man. Not only is the workplace a site of male exploitation of women, sexual and otherwise, but it is also a locale for behaviors from flirting to consensual sex. Not just sexual harassment but additionally sexual pursuits have a workplace arena (Williams et al., 1999).

The Changing Nature of Work

In industrial societies, the future of work could include not enough of it to permit, for all, the earning of middle-class compensation (Simpson, 1989). Competition in the global market engenders a massive deindustrialization, with the loss of adequate jobs for those with high-school levels of education or less. Globalization (see Edgell, 2006: pp. 182–203; Volti, 2008: pp. 39–58) integrates national economies into one price-making, capitalist whole, even for so-called socialist states. Return on capital is the ultimate arbiter of whether a business fails or succeeds and jobs are extant. Generally, with rationalization and automation of work a country's standard of consumption rises.

A new industrial revolution dawns in additive manufacturing (so-called 3-dimensional printing) in which a digital materials 'printer' precisely forms an object by successively layering material into final shape (Grenda, 2012; Anderson, 2013). This fledgling technology replaces the traditional subtractive machining or forging of an object to shape. Developers can produce concept designs of objects in a few hours. Thus, production industries face radical transformations with redundancies in the world of work and its traditional tasks.

Production workers in developed countries must compete with global cut-rate labor. In the North American job market, employee loyalty has become a one-way street. Overnight, jobs can be eliminated in the thousands within a single firm, sometimes after its leveraged buyout. Some terminations are psychologically brutal.

Industrial Policy

Industrial policy involves a national plan and government involvement, ranging from pronounced to weak, for helping industries, including increasing and protecting their jobs. Industrial sectors and government agencies are coordinated in selected activities, and the former receive protections, grants, loans, and tax incentives for particular businesses. The developing counties benefit from such policies, and the developed countries have gained from it across time (Altenburg, 2011). In the United States, the Internet, satellite technology, and various forms of transportation benefit from industrial policy. However, especially when labeled as such, the policy is ideologically controversial, including regarding growth in jobs.

Computerization

Computer control and the handling of information and electronic communication both deskill and eliminate jobs. The blitz of combined computing and instant electronic communications change work and life ways. Watson aptly summarizes these changes as including computer-enhanced products; rapid design and testing of products; production of things by automatically controlled tools, robotic devices, and production lines; administration and monitoring of work; and control and determining of distribution of goods and services. Professionals tap enormous ranges of databases in medicine, law, engineering, and other professions, thereby using the accumulated experiences of numerous peers (Watson, 2008: pp. 168–171). Consequent to the computerization of work is its intensification and a narrowing of the scope for initiative (Gallie et al., 2004).

With computerization, products enter into a new age of custom specification. The worker produces and the consumer receives precisely the product desired. Information services have automation, often frustrating consumers, ranging from hierarchically organized call centers to a firm's voice mail directions for connections. Nevertheless, through the Internet, a surfeit of information is available.

Computer-controlled supply chains components of manufacturing and its jobs are global. These parts have selection for effectiveness through their design and fabrication by many specialists around the world. Production in one country depends on components from another. After the tsunami of 2011, Japanese factories could not produce certain automotive components, thereby disrupting the manufacture of automobiles globally.

An issue for computerized operations is, to what extent does reduction of task content reduce operator responsibility for control of a procedure, including, under operational circumstances, at the edges of the operating range and the environment? What are the boundaries between human and computer responsibilities and, does the human have sufficient time, observability, and directability for handling the procedure, as in aircraft, boat, and locomotive control? As long as an operator is responsible for outcomes and can lose any certification, he/she must have effective authority and realistic ultimate control over solving a range of problems in the human–computer interface (Gamst and Gavalla, 2005: pp. 99–109).

Alienation in Industrial Work

The Marxian-rooted idea of 'alienation' (estrangement) in unskilled and semiskilled work concerns subjective feelings about paid toil. Such lack of job satisfaction occurs in capitalist and socialist societies. Isolation in the work group, a feeling of powerlessness, and insufficient remuneration are dimensions of work alienation. This finds expression in the North American workers' anthem "Take This Job and Shove it." However, even assembly-line workers might find satisfaction of the job supported by their employment (Blauner, 1964; Edgell, 2006: pp. 28–47).

Deskilling and Upskilling

Automation and reorganization of work created a controversy about the existence of 'deskilling' under capitalism, a continuous lowering of levels of skill throughout a national economy, leading to an unskilled proletariat. The notion of 'upskilling,' in a society characterized as knowledge-based and post-private-property, criticizes the idea of deskilling (Edgell, 2006: pp. 48–72).

Neo-Marxist 'deskilling' is a form of social control of workers through a minute division of labor, furthering organizing work into simple, repetitive tasks involving few mental reflections. Although a good number of crafts and some professions have had deskilling through the present, no deskilling throughout an overall industrial economy exists as yet. Professions, often protected by their gatekeepers, have specialized knowledge requiring extensive education and occupational training for their power-wielding roles (Volti, 2008: pp. 97–116).

Training and Education

In modern societies, opportunity for 'good' jobs requiring slight instructional background diminishes. Operational redundancy in targeted jobs continues. Technical backgrounds necessary for employment in the United States and Canada raise questions about the requisite kinds of preparation for jobs and careers. Politicians at the highest levels mindlessly advocate everyone 'going to college,' a course sometimes leading to a low-compensation retail position as an 'associate.' Values and norms for 'blue-collar' work wither against those for earning a BA in art history, liberal studies, or comparative literature. Students avoid as too demanding STEM (science, technology, engineering, and mathematics) studies. Rejected as undemocratic is the education and training found in Germany's trade tech emphases coupled to meaningful, job-acquiring apprenticeships.

Rudderless national educational policy has costs for individuals and society. Several North American railroads, however, prepare future employees for rewarding jobs at a community college or a corporate campus. Some German firms introduce their job instruction in the United States.

Education of managers and workers plus legal remedies can foster diversity and lessen its impediments in the workplace (Volti, 2008: pp. 137–156). Diversity is a concern in most developed countries (DiTomaso et al., 2007).

Outsourcing and Offshoring

Businesses sometimes reduce jobs and foster an erosion of remuneration by having work done outside the firm more cheaply and, when in another country, sometimes with fewer workers' rights. 'Outsourcing' of work has a business contracting to have some of its functions, often regarding information and administrative services, performed by an independent firm, located in or out of country. 'Offshoring' is 'outsourcing' when a business moves functions, such as manufacturing or information and administrative services, from the home country to another country (Davis-Blake and Broschak, 2009). China and Mexico are common locations for offshoring production and India for the two services plus assessing medical information. Threatened are even the developed world's jobs in the STEM professions.

The Union Label?

Trade unions contest management for social control in some workplaces. In America, union membership declined from 26.7% of the workforce in 1973 to 13.6% in 2011. This declining unionization has lowered wages, in part through employers' aggressive positions hostile to unions and through changes in applying and processing labor law, thus contributing to union weakening, including its means of increasing remuneration. The weakening of unions contributes to the faltering wages of the middle class. The union premium on wages and benefits compared to those of nonunion members is 17.3% for men and 9.1% for women. Unionized workers have 28.2% greater likelihood to have health insurance coverage provided by the employer and a 53.9% greater likelihood of receiving a pension from the employer (EPI, 2012: pp. 1–3). Unionization provides an ever declining wage advantage to American workers.

The future of the labor movement, however, is clouded (Durrenberger, 2007). In the United States and Canada, trade unions have diminished in union density, organizing success, level of labor actions, and political influence. Reasons for this decline include union sloth; the law as enacted and then interpreted by the courts; globalized competition; management-labor 'cooperative' programs; and a neo-liberal philosophy embraced by the large political parties. For union success, new forms and policies in the labor movement must be created including, multi-employer bargaining, removal of legal impediments to organizing, and coordination with community, advocacy, religious, and ethnic organizations to further mutual goals (Clawson and Clawson, 1999).

Retiring from Work

Macroeconomic changes affect the retirement income previously earned by workers. When government policy maintains a near-zero percent interest for loans, it harms projections of pension funds and impoverishes retirees relying on interest-bearing accounts. Policy for near-zero rates promotes spending rather than prudent saving. Beginning in 2007 real interest rates (after inflation) became negative. In the United States alone, about 10 000 workers retire every day. Is neediness a reality of their future?

Working in the New Economy

During the 1990s, changing the world of work, a 'New Economy' dawned based on computerization, communication and information technologies, e-commerce, and outsourcing (Grint, 2005: pp. 355–384; Perrucci and Perrucci, 2007). These changes include dissolving the iron-bound business bureaucracy and proliferating work that is unstable and without long-term careers and their benefits. Market competition eliminates personnel through the new technologies. Greater productivity results in using fewer workers (Sennett, 2006).

Concurrent with the New Economy is dissipation of the shared prosperity for American workers. Hedrick Smith holds that multiple enacted and adjudicated US policies favoring the ultra-rich retard economic growth, jobs, and remuneration for ordinary workers (Smith, 2012). Mollona notes that constructs and generalizations relating to 'post industrial' work do not entirely fit ethnographic observations in both industrial and developing countries. Overall, managerial policies for work conceptualized in Taylorism and Fordism persist; white-collar workers experience Braverman's deskilling and alienation, and class relations remain in various forms (Mollona, 2009a). At times, "The 'modern' system of flexible production is built upon a network of Victorian sweatshops hidden beneath the post-industrial façade of shopping malls, highways and leisure centers" (Mollona, 2009b: pp. xiii–xiv).

The End of Work?

The era of lifetime jobs gradually erodes, replaced by insecure 'contingent work,' including part-time, contracts and temporary persons – all are workers who are not full-time with full benefits. Conditions of employment not obtained by contingent workers include job training and career advancement with a firm. Such workers become employed only because of an immediate demand for them. Some people desire contingent work.

Rifkin (1995) and others pessimistically predict 'end-of-work' circumstances through technological and organizational rationalizations of work, including an irreversibly growing future unemployment and the loss of the centrality of work in peoples' lives. They foresee a crisis of lack of work in modern societies. Wilson (2012) and others optimistically counter this pessimism. Wilson holds that full employment with equality and a revitalized labor movement is possible given appropriate legislation and changes in corporate power. The issue is, how are these effected in the competitive global market and with neoliberal legislators fervent in many developed countries? Industrial policy cannot be free of values and norms. Accordingly, the future of modern work poses profound questions.

See also: Alienation, Sociology of; Apprenticeship, Anthropological Aspects; Capitalism: Global; Career Development: Multilevel Perspective; Community Sociology; Community Studies: Anthropological; Gender, Economics of; Horticultural Societies; Human Resource Management, Psychology of; Hunting and Gathering Societies: Anthropology; Industrial–Organizational Psychology: Science and Practice; Innovation and Technological Change, Economics of; Irrigation Societies; Job Analysis and Work Roles, Psychology of; Labor, Division of; Organizations, Sociology of; Pastoralism in Anthropology; Performance Appraisal and Evaluation; Professions in Organizations; Slavery: Comparative Aspects; Stress at Work; Strikes; Trade Unions, Economic Behavior of; Vocational Interests, Values, and Preferences, Psychology of; Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences; Work and Industry, Sociology of; Workplace Autonomy; Workplace Deviance; Workplace Safety and Health.

Bibliography

Abbott, A., 1993. The sociology of work and occupations. Annual Review of Sociology 19, 187–209.

Altenburg, T., 2011. Industrial Policy in Developing Countries: Overview and Lessons from Seven Country Cases. Deutsches Institut für Entwicklungspolitik, Bonn.

Anderson, C., 2013. Makers: The New Industrial Revolution. Crown Business, New York.

Applebaum, H. (Ed.), 1984. Work in Non-market and Transitional Societies. SUNY Press, Albany, NY.

Blauner, R., 1964. Alienation and Freedom: The Factory Worker and His Identity. University of Chicago Press, Chicago, IL.

Bouvier, P., 1994. Le travail [Work], second ed. Presses Universitaires de France, Paris.

Budd, J., 2011. The Thought of Work. Cornell University Press, Ithaca, NY.

Carey, A., 1967. The Hawthorne studies: a radical criticism. American Sociological Review 32, 403–416.

Castillo, J., 1997. Looking for the meaning of work. Work and Occupations 24, 413–425.

Clawson, D., Clawson, M., 1999. What has happened to the U.S. labor movement? Union decline and renewal. Annual Review of Sociology 25, 95–119.

Dahrendorf, R., 1983 [1968]. Homo Sociologicus. Routledge, London.

Davis-Blake, A., Broschak, J., 2009. Outsourcing and the changing nature of work. Annual Review of Sociology 35, 321–340.

DiTomaso, N., Post, C., Parks-Yancy, R., 2007. Workforce diversity and inequality: power, status, and numbers. Annual Review of Sociology 33, 473–501.

Donaldson, E., 2001. The Contingency Theory of Organizations. Sage, Thousand Oaks, CA.

Durkheim, E., 1947 [1902]. The Division of Labor in Society (G. Simpson, Trans.), second ed. Free Press, Glencoe, IL.

Durrenberger, E.P., 2007. The anthropology of organized labor in the United States. Annual Review of Anthropology 36, 73–88.

Edgell, S., 2006. The Sociology of Work: Contingency and Change in Paid and Unpaid Work. Sage, Los Angeles, CA.

EPI (Economic Policy Institute), 2012. Unions, inequality, and faltering middle-class wages. Brief No. 342.

Gallie, D., Felstead, A., Green, F., 2004. Changing patterns of task discretion in Britain. Work, Economy and Society 18, 243–266.

Gamst, F.C. (Ed.), 1977. Golden Anniversary Special Issue on Industrial Ethnology. Anthropological Quarterly 50(1) (whole issue).

Gamst, F.C. (Ed.), 1995. Meanings of Work: Considerations for the Twenty-First Century. SUNY Press, Albany, NY.

Gamst, F.C., Helmers, S., 1991. Die kulturelle Perspektive und die Arbeit [The Cultural Perspective and Work]. Zeitschrift für Ethnologie 116, 25–41.

Gamst, F.C., Gavalla, G., 2005. Hazard Survey of Remote Control Locomotive Operations on the General System of Railroads in the United States. Brotherhood of Locomotive Engineers and Trainmen, Cleveland. Additionally, an e-edition at: http://www.ble-t.org/pr/pdf/BLETRCOHazardSurvey.pdf.

Gibbs, J., 1994. A Theory about Control. Westview, Boulder, CO.

Grenda, E., 2012. Printing the Future: The 3D Printing and Rapid Prototyping Source Book. Castle Island, Arlington, MA.

Grint, K., 2005. The Sociology of Work, third ed. Polity, Malden, MA.

Helmers, S. (Ed.), 1993. Ethnologie der Arbeitswelt: Beispiele aus europaischen und auereuropa ischen Feldern [Ethnology of the Work World: Examples from European and non-European Locations]. Holos Verlag, Bonn, Germany.

Herskovits, M., 1952. Economic Anthropology. Knopf, New York.

Kalleberg, A.L., Berg, I., 1987. Work and Industry: Structures, Markets, and Processes. Plenum, New York.

Liebow, E., 1967. Tally's Corner: A Study of Negro Streetcorner Men. Little, Brown, Boston, MA.

McCabe, D., 2007. Power at Work: How Employees Reproduce the Corporate Machine. Routledge, London.

Miller, D., Form, W., 1951. Industrial Sociology. Harper and Row, New York.

Mollona, M., 2009a. Made in Sheffield: An Ethnography of Industrial Work and Politics. Berghahn, New York.

Mollona, M., 2009b. General introduction. In: Mollona, M., De Neve, J., Parry, J. (Eds.), Industrial Work and Life: An Anthropological Reader. Berg, New York.

Nichols, T., 2005. The Sociology of Occupational Injury. Taylor and Francis, New York.

Pavalko, R., 1988. Sociology of Occupations and Professions, second ed. Peacock, Itasca, IL.

Perrow, C., 1986. Complex Organizations: A Critical Essay, third ed. Random House, New York.

Perrucci, R., Perrucci, C., 2007. The Transformation of Work in the New Economy. Oxford University Press, New York.

Rifkin, J., 1995. The End of Work. Putnam, New York.

Schemien, M.-D., 2012. Der Mensch und die Arbeit: Identitätsprobleme in der Post-moderne. Akademiker Verlag, Saarbrücken.

Sennett, R., 2006. The Culture of the New Capitalism. Yale University Press, New Haven, CT.

Simpson, I.H., 1989. The sociology of work: where have all the workers gone? Social Forces 67, 563–581.

Spittler, G., 2008. Founders of the Anthropology of Work: German Social Scientists of the 19th and Early 20th Centuries and the First Ethnographers. LIT Verlag, Berlin.

Smith, H., 2012. Who Stole the American Dream? Random House, New York.

Tanguy, L., 2008. Retour sur l'histoire de la sociologie du travail en France: place et rôle de l'Institut des sciences sociales du travail. Revue Française de Sociologie 49, 723–761.

Tilgher, A., 1958 [1930]. Work: What It Has Meant to Men through the Ages. Henry Regnary, Chicago, IL.

Volti, R., 2008. An Introduction to the Sociology of Work and Occupations. Pine Forge Press, Los Angeles.

Wallman, S. (Ed.), 1979. The Social Anthropology of Work. Academic Press, New York.

Watson, T., 2008. Sociology, Work, and Industry, fifth ed. Routledge, New York.

Williams, C., 2005. Unraveling the meanings of underground work. Review of Social Economy 63, 1–18.

Williams, W., 1918. Human Relations in Industry. U.S. Department of Labor, Washington, DC.

Williams, C., Guiffre, P., Dellinger, K., 1999. Sexuality in the workplace: organizational control, sexual harassment, and the pursuit of pleasure. Annual Review of Sociology 25, 73–93.

Wilson, S., 2012. The Struggle Over Work: The 'End of Work' and Employment Alternatives in Post-industrial Societies. Routledge, London.

Working Classes, History of

M Van der Linden, International Institute of Social History, Amsterdam, The Netherlands

© 2015 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16579–16583, © 2001, Elsevier Ltd, with revisions made by the Editor.

Abstract

This article discusses the emergence of the concept of the working class in Western Europe during the nineteenth century; the classical patterns of interpretation (structural, agency-focused, and combinations of these two approaches); the rise of 'peripheral' working classes, their historiography, and their conceptual implications; and the new insights offered by Third World historians for the analysis of advanced countries.

The Working Classes: Origins of the Concept

The concept of the working class emerged in Western Europe during the nineteenth century. Originally, the term was used as a plural, denoting a heterogeneous conglomerate of social groups performing various forms of wage labor. The working classes were considered to be a fraction of a larger 'proletariat,' that is, the estate of people without property, beyond honor. The wage earners were only a part of this amorphous mass. According to the French nobleman Granier de Cassagnac (1838: p. 30), the proletariat formed 'the lowest rank, the deepest stratum of society,' that consisted of four groups: 'the workers, the beggars, the thieves, and public women.' "The worker is a proletarian, because he works in order to live and earns a wage; the beggar is a proletarian, who does not want to work or cannot work, and begs in order to live; the thief is a proletarian, who does not want to work or beg, and, in order to make a living, steals; the prostitute is a proletarian, who neither wants to work, nor beg, nor steal, and, in order to live, sells her body." As capitalist relations crystallized more and more, the desire arose among the 'respectable' workers to distinguish themselves from the rest of the 'proletariat.'

Over the course of the nineteenth century, social theorists and the emerging labor movements began to redefine these notions: (1) the working classes became homogenized as a singular working class; (2) this working class was identified with the proletariat; and (3) this proletariat/working class was marked off in three ways: from the petty bourgeoisie, from the so-called lumpenproletariat (the criminal and unsteady 'dangerous classes'), and from the unfree workers (slaves and serfs). An early example of this development was the *Manifesto of the Communist Party* of 1848, which emphatically distanced the proletariat/working class on the one hand from the 'lower middle class, the small manufacturer, the shopkeeper, the artisan, the peasant,' and on the other hand from the "'dangerous class,' the social scum, that passively rotting mass thrown off by the lowest layers of old society" (Marx and Engels, 1973: p. 77). A similar sharp distinction was drawn between 'free' workers and the slaves in the South of the United States and in some colonies.

Moreover, the notion of a working class was a gendered abstraction, based on the assumption that workers' families exist mainly on the wages earned by the male head of the household and perhaps the additional income of other household members. From 1900 or so scholars of various theoretical orientations seemed to have reached an agreement on what a worker is. In 1926 the German social-christian sociologist. Briefs summarized this *communis opinio* as follows: "The worker is *personally free*, i.e., his physical and spiritual-moral powers are completely at his own disposal. […] He *has no property*, i.e., he has no exclusive material power over capital as a secure basis with relative permanency. […] He has neither a stock of consumer goods that enable him to live, nor permanent interests of capital. […] He lives in economic circumstances in which means of subsistence can be obtained only through *economic returns*. […] He is forced to offer personal capacities with an economic exchange value in return for means of subsistence" (Briefs, 1926: p. 149).

Origins in the 'Core Countries'

Detailed surveys of the living and working conditions of the working classes date from the first decades of the nineteenth century. A relatively early example was the Tableau de l'état moral et physique des ouvriers employés dans les manufactures de coton, de laine, et de la soie, written by Villermé (1840), a medical doctor with a great interest in the social and physical condition of 'the people.' Many writings of this kind were intended partly as moral and political indictments of social abuses. One of the best known instances is Engels's (1845) *Die Lage der arbeitenden Klasse in England* (*The Condition of the Working Class in England*), which gave a minute description of the working classes in English industry, mining, and agriculture. During the second half of the nineteenth century social surveys became a quite common genre, especially in industrializing Western Europe. Highlights include the many studies of Frédéric Le Play and his school which covered all of Europe, beginning with *Les ouvriers européens* (1855); Henry Mayhew's *London Labour and the London Poor* (1851–61) with the significant subtitle 'a cyclopaedia of the conditions and earnings of those that will work, those that cannot work, and those that will not work'; Charles Booth's *Life and Labour of the People of London* (1892–97); Alphons Thun's detailed monograph *Die Industrie am Niederrhein und ihre Arbeiter* (1879) on the industrial and working classes in West Germany; and numerous publications of the German Verein für Sozialpolitik (Social

Policy Association), including Max Weber's famous study of agricultural laborers in East Prussia (1894).

All these works were mainly contemporaneous, though they have by now themselves become important historical sources. Truly historical studies of working classes followed at the beginning of the twentieth century. A pathbreaking effort was the trilogy written by John Le Breton Hammond and his wife Barbara Bradby. *The Village Labourer, 1760–1832* (1911), *The Town Labourer, 1760–1832* (1917), and *The Skilled Labourer, 1760–1832* (1919). Together these studies provide a rich survey of the British working class during the Industrial Revolution. In other countries similarly wide-ranging works usually were published much later, generally after the World War II.

Classical Debates

The study of the working classes has always been central to social and labor history. But different aspects have been studied and emphasized.

The concept of the working class has been used in many different ways. Somewhat simplifying, one might distinguish three main patterns of interpretation. The structural approach uses the notion of the working class in a restrictive sense as a term merely characterizing the common socioeconomic location of a group of wage earners. The historical study of a particular working class then boils down to the reconstruction of their joint 'class position.' How this class position is analyzed depends on theoretical assumptions (Marxist, Weberian, or otherwise) and on the various ways in which research is operationalized (models of social stratification).

A second approach focuses on agency and emphasizes culture, mentalities, consciousness, and social and organizational behavior. Numerous variations on this theme have developed over time. Perhaps the most influential approach was presented by Edward P. Thompson in his *The Making of the English Working Class* (1963). Here, he argued that. "The notion of class entails the notion of historical relationship. Like any other relationship, it has a fluency which evades analysis if we attempt to stop it dead at any given moment and anatomize its structure." Class, according to Thompson is an outcome of collective experience, and workers are active participants in their own class formation.

While the first approach has been criticized for its 'objectivism' and the second approach for its 'culturalism,' a third approach considers working-class formation as 'an outcome of temporal conjunctures between multiple causal structures' (Sewell, 1990: p. 73). In this view, class formation results from a combination of economic and technological changes on the one hand and of cultural, political, and discursive developments on the other hand.

The debate on working-class formation is paralleled by an extensive debate on working-class devolution: Does it still make sense to speak of working classes in advanced capitalist countries? Or have these classes dissolved? And if there has been a process of devolution, how and when did it start? Naturally, everything depends again on one's conceptualization of the working class. What does one consider as decisive: structural issues (free wage labor, etc.), collective behavior (e.g., trade unions, strikes), or cultural features? Whichever position one takes, it is evident that the condition of wage earners in advanced societies has changed profoundly during the twentieth century. Research has not given equal attention to all these developments. Especially the cultural shifts after the World War II remain underexplored.

Most early contributions to working-class history focused on work, industrial relations, and labor markets. At a later stage the field of research was expanded and historians began to pay serious attention to other aspects of working-class existence, including labor processes, family life, religion, etc. Over the course of time the first approach has become more sophisticated and the second approach has gained influence. The historiography of the workers as a class is now increasingly merged with the historiography of race, ethnicity, and gender. While earlier historians considered race and gender as mere adjuncts of class, it is now generally becoming accepted that class, race, and gender are thoroughly interpenetrating, and that they shape each other. As a result of these developments, working-class history increasingly overlaps into women's studies, ethnic studies, anthropology, institutional economics, sociology, or social psychology.

Now that there exists a huge and still growing body of literature on the various historical aspects of working-class existence first attempts are made to integrate the history of various aspects of working-class life and to write comprehensive syntheses. The most spectacular example of such an endeavor is the ongoing project on the *History of the Workers and the Labor Movement in Germany since the End of the Eighteenth Century*, coordinated by Gerhard A. Ritter (see Kocka, 1990, 1991; Ritter and Tenfelde, 1992; Schneider, 1999; Winkler, 1984, 1985, 1987).

The Rise of the 'Periphery'

The growth of working classes outside the advanced capitalist countries has, especially since the World War II, caused an increased interest in the working classes of Africa, Asia, and Latin America. From a contemporary point of view this field was first explored by anthropologists and students of industrial relations, who were interested in the problems accompanying the integration of rural migrants into the emerging wage-earning economies. Early examples were A.L. Epstein's ethnographic *Politics in an Urban African Community* (1958), on copper miners in Northern Rhodesia, and industrial relations studies driven by the question how processes of social change could be controlled and how the 'commitment' of the workers could be intensified (e.g., Moore and Feldman, 1960; see also Breman, 1999).

Simultaneously, historiographical studies emerged; initially these were often based on official sources and looked at the workers through the eyes of employers and authorities. Such was for instance the case in J. Norman Parmer's thorough *Colonial Labor Policy and Administration* (1960) on the rubber plantation industry in Malaya during the last decades before the World War II. Since the 1960s and 1970s the historiography of the 'peripheral' working classes has changed too. First, there was an upsurge of classical labor movement history. Many of the major early studies of this genre were written by radical

intellectuals, like Jean Chesneaux's *Le mouvement ouvrier en Chine de 1919 à 1927* (1962) or Guillermo Lora's *Historia del movimiento obrero Boliviana* (1967–70). Second, gradually the perspective broadened and everyday life, mentalities, etc. of working classes received more attention, just as had been the case in the historiography of the high-income countries. Early attempts in this direction were still heavily influenced by Western models of historical development. Morris D. Morris's *The Emergence of an Industrial Labor Force in India: A Study of the Bombay Cotton Mills, 1854–1947* (1965) is a major example of this approach, despite its rich empirical basis. Later attempts often tried to develop a less Eurocentric approach. Pathbreaking works include Charles van Onselen's *Chibaro* (1976) on mine labor in Southern Rhodesia or Ranajit Das Gupta's *Labour and Working Class in Eastern India* (1994) on plantation workers, miners, and textile workers in Assam, Bengal, and elsewhere. These new developments have been covered by a series of consecutive collections of essays (Sandbrook and Cohen, 1975; Gutkind et al., 1978; Cohen et al., 1979; Munslow and Finch, 1984; Agier et al., 1987; Amin and Van der Linden, 1997). While the traditional historiography of the high-income countries focused primarily on artisanal and factory labor, the new working-class historiography of the peripheral countries reflected the peculiarities of colonial exploitation and unavoidably paid much attention to the labor in agriculture (including plantations), mining, and the transport sector (docks, railways).

Global Insights

The rise of working-class history in low-income and medium-income countries has made necessary a rethinking of old concepts. First, in the periphery, the larger part of the wage earners is not 'free' in an economic sense. According to the classic definition, wage earners, as 'free persons' are free in a double sense because they alone market their own labor as a commodity and have no other commodity for sale. A 'free' worker is, strictly speaking, someone without obligations to any particular employer. He or she can cancel employment at will and apply for a job with another employer. In reality, however, such a completely uncommitted relationship with the boss is rare, occurring, for example, among some forms of casual labor. Many workers are tied to their employer, for example, through debts or nonmaterial obligations. There can also be all sorts of cultural or geographic reasons why it is difficult to change employers. Somewhat better-placed personnel can be tied to a firm through superannuation provisions. Frequently, the accommodation of workers is part of the labor contract, etc. In short: what formally seems like a free labor relation is in practice often much less free. A 'free' worker, moreover, does not possess his or her own means of production at all. But in reality there are many types of labor relations in which the employee contributes at least part of the means of production himself. One thinks here of factory workers who own (part of) their own tools, but one can also think of agrarian share-cropping arrangements. Still clearer is the case of self-employed workers, who formally are wholly independent, but who in fact often form a very flexible labor force for enterprises. We can also mention here the case of home workers.

Second, there is wage labor integrated throughout in households and families whose survival always remains partly dependent on other activities. Not only does a great deal of subsistence labor take place every day in every family (especially, but not exclusively performed by women), but there are usually also several forms of activities remunerated with money. Already in the early 1970s Allen observed, for example: "Cash crops and wage employment are simply alternative methods of coping with the encroachment of capitalism on the traditional way of life. A fall in the demand for labour increases the emphasis on subsistence or cash-crop production, whereas a drought or a bad harvest, for instance, forces Africans to enter the labour market" (Allen, 1972: p. 181). Within these arrangements with multiple coping methods the 'roles' different family members have are often not fixed permanently, but instead signify a transient social relationship, which can be replaced rather quickly by other sources of income.

Third, the new working class does not mainly or exclusively exist in the industrial sector. The agricultural sphere is proportionally more important in this sphere rapidly advancing proletarianization has created a large stratum of agricultural laborers and share tenants.

Fourth, in the so-called Third World and even in the East- and Southeast Asian newly industrializing countries, permanent employment relationships have always been the exception rather than the rule. Insecure employment relationships, underemployment, and long-term unemployment are the reality there.

Implications for Future Research

The new insights offered by Third World historians have implications for the advanced countries as well.

First, the dividing line between wage labor and small entrepreneurship is much more obscure than was originally thought. In both the 'Third' and 'First' worlds there are gray zones of (ostensible) self-employment, in which individuals formally work on their own but in reality are dependent on one or two customers. This type of self-employment has been widespread in peripheral capitalism for decades, but lately, it has been increasing rapidly in the centers as well, such as in the construction business. Generally, there seems to be a historical connection in highly developed capitalism between phases of rising unemployment, on the one hand, and stagnation or the growth of 'small-scale' self-employment, on the other.

Second, the border between wage laborers and the marginal groups is not nearly as obvious as older theories would have us believe. In times of desperation, regularly employed wage earners may indeed exhibit behavior characteristic of the *Lumpenproletariat*, such as begging and prostitution, as a way to survive. They will also steal at the workplace, which represents thoroughly criminal behavior from the entrepreneurial viewpoint. Yet, the experience in the Third World has drawn our attention to a lower social class that can be characterized neither as the industrial reserve army in a Marxist sense nor as

the classic *Lumpenproletariat*. Inevitably we need to come up with a new concept to apply to insecure employment relationships in this respect.

Third, the concept of the 'free' laborer is less precise than is usually assumed. Fourth and last, the strict differentiation commonly made in advanced countries between urban and rural life must be revised. Contrary to what modernization theories like to tell us, the ties of urban migrants to their home villages often do not weaken but strengthen over time. Most likely, the reason for this is the lack of social insurance systems. Therefore, the villages become safety nets during times of economic crisis. Recently, though, many villages have not been able to fulfill this task, since the rise of a commodity economy has drained their capacity for subsistence.

What it means to be classified as 'working class' is changing; the category is becoming especially multifarious. We could even flip our intellectual problem and ask why a 'full-time proletarian' in capitalism should have existed or can now even exist at all. From the entrepreneurial standpoint, part-time proletarians are cheaper by far. Part-time proletarians (who belong to a household in which there are other sources of income, e.g., from agricultural work or from self-employment) lower the bottommost, barely acceptable wage threshold because they have other forms of real income (Wallerstein, 1983).

See also: Alienation, Sociology of; Bourgeoisie and Middle Classes, History of; Class Consciousness; Class: Social; Labor History; Labor Unions; Labor, Division of; Work, History of; Work, Labor: History of the Concept; Work, Sociology of.

Bibliography

Agier, M., Copans, J., Morice, A. (Eds.), 1987. Classes ouvrires d'Afrique noire. Karthala-Orsthom, Paris.

Allen, V.L., 1972. The meaning of the working class in Africa. The Journal of Modern African Studies 10, 169–189.

Amin, S., Van der Linden, M. (Eds.), 1997. "Peripheral" Labour? Studies in the History of Partial Proletarianization. Cambridge University Press, Cambridge, UK.

August, A., 2007. The British Working Class, 1832–1940. Pearson Longman, Harlow, England.

Behan, T., 1997. The Long Awaited Moment: The Working Class and the Italian Communist Party in Milan, 1943–1948. P. Lang, New York.

Belshaw, J.D., 2002. Colonization and Community: The Vancouver Island Coalfield and the Making of the British Columbian Working Class. McGill-Queen's University Press, Montréal.

Berger, S., 2000. Social Democracy and the Working Class in the Nineteenth and Twentieth Century Germany. Longman, Harlow, England.

Breman, J., 1999. Industrial labor in post-colonial India, part 1/part 2. International Review of Social History 44, 249–300, 451–483.

Briefs, G., 1926. Das gewerbliche proletariat. In: Grundriss der Sozialökonomik, Part IX. Mohr (Paul Siebeck), Tubingen, Germany, pp. 142–240.

Clark, A., 1995. The Struggle for the Breeches: Gender and the Making of the British Working Class. University of California Press, Berkeley.

Cohen, R., Gutkind, P.C.W., Brazier, P. (Eds.), 1979. Peasants and Proletarians: The Struggles of Third World Workers. Hutchinson, London.

Ellis, P.B., 1996. A History of the Irish Working Class. Pluto Press, London.

Frader, L.L., Rose, S.O., 1996. Gender and Class in Modern Europe. Cornell University Press, Ithaca, NY.

Freund, B., 1995. Insiders and Outsiders: The Indian Working Class of Durban, 1910–1990. Heinemann, Portsmouth, NH.

Granier de Cassagnac, A., 1838. Histoire des classes ouvrières et des classes bourgeoises. Desrez, Paris.

Guharaya, S., 2007. Calcutta Tramwaymen: A Study of Working Class History. Progressive Publishers, Kolkata.

Gutkind, P.C.W., Cohen, R., Copans, J. (Eds.), 1978. African Labor History. Sage, Beverly Hills, CA.

Kirk, N., 1998. Change, Continuity and Class: Labour in British Society, 1850–1920. Manchester University Press, Manchester, UK.

Kocka, J., 1990. Weder Stand noch Klasse. Unterschichten um 1800. Dietz, Bonn.

Kocka, J., 1991. Arbeitsverhltnisse und Arbeiterexistenzen. Grundlagen der Klassenbildung im 19. Jahrhundert. JHW Dietz, Bonn.

Le, B.P., 1999. A Short History of the U.S. Working Class: From Colonial Times to the Twenty-First Century. Humanity Books, Amherst, NY.

Lockman, Z., 1994. Workers and Working Classes in the Middle East: Struggles, Histories, Historiographies. State University of New York Press, Albany.

MacDowell, L.S., Radforth, I.W., 2000. Canadian Working Class History. Canadian Scholars' Press, Toronto.

Marx, K., Engels, F., 1973. Manifesto of the communist party (1848). In: Marx, K. (Ed.), Political Writings, The Revolutions of 1848, vol. 1. Penguin Books, Harmondsworth, UK, pp. 67–98.

McLeod, H., 1993. Religion and Irreligion in Victorian England: How Secular was the Working Class? Headstart History, Bangor.

Moore, W.E., Feldman, A.S. (Eds.), 1960. Labor Commitment and Social Change in Developing Areas. Social Science Research Council, New York.

Munslow, B., Finch, H. (Eds.), 1984. Proletarianisation in the Third World. Croom Helm, London.

Parmer, J.N., 1960. Colonial Labor Policy and Administration. A History of Labor in the Rubber Plantation Industry in Malaya, c. 1910–1941. Augustin, Locust Valley, NY.

Ritter, G.A., Tenfelde, K., 1992. Arbeiter im Deutschen Kaiserreich 1871–1914. Dietz, Bonn.

Sandbrook, R., Cohen, R. (Eds.), 1975. The Development of an African Working Class: Studies in Class Formation and Action. Longman, London.

Schneider, D., 1994. Trade Unions and Community: The German Working Class in New York City, 1870–1900. University of Illinois Press, Urbana.

Schneider, M., 1999. Unterm Hakenkreuz. Arbeiter und Arbeiterbewegung 1993–1939. Dietz, Bonn.

Sewell Jr., W.H., 1990. How classes are made: critical reflections on E.P. Thompson's theory of working-class formation. In: Kaye, H.J., McClelland, K. (Eds.), E.P. Thompson: Critical Perspectives. Polity, Cambridge, MA, pp. 50–77.

Silbey, D., 2005. The British Working Class and Enthusiasm for War, 1914–1916. Frank Cass, New York.

Thompson, E.P., 1963. The Making of the English Working Class. Gollancz, London.

Villermé I, L.-R., 1840. Tableau de l'état physique et moral des ouvriers employés dans les manufactures de coton, de laine et de soie (Introductions J.P. Chaline, F. Dérnier). Etudes et Documentation Internationales, Paris.

Wallerstein, 1983. Historical Capitalism. Verso, London.

Winkler, H.A., 1984. Von der Revolution zur Stabilisierung. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1918–1924. Dietz, Bonn.

Winkler, H.A., 1985. Der Schein der Normalität. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1924–1930. Dietz, Bonn.

Winkler, H.A., 1987. Der Weg in die Katastrophe. Arbeiter und Arbeiterbewegung in der Weimarer Republik 1930–1933. Dietz, Bonn.

Working Conditions

Françoise J Carré, Center for Social Policy, J.W. McCormack Graduate School of Policy and Global Studies, University of Massachusetts Boston, Boston, MA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Working conditions range from the employment relationship to the environment in which workers perform their tasks. This article addresses how working conditions are shaped by regulation. Most commonly, regulation establishes labor standards, that is, sets minima. Regulation can shape structures for negotiation that affect working conditions. Most narrowly, regulation is understood as emanating from government through laws as well as administrative rules. Social norms along with worker relative bargaining power also affect working conditions. Some see labor standards as a cost and/or as distorting labor markets; for others, standards reflect social norms and power relations between employers and workers.

Working conditions range from the terms under which workers enter the employment relationship to the environment in which they perform their tasks. This article addresses how working conditions are shaped by regulation. The most common purpose of the regulation of working conditions is to establish labor standards, that is, set minima for wage, hours, or health and safety conditions. Also, regulation can establish a framework for the formation of structures for negotiation that contribute to shaping working conditions. Most narrowly, regulation is understood as emanating from government through laws as well as administrative rules (executive). Courts help enforce laws. Also, the relative bargaining power of labor unions and other organizations affects the ability of workers to bargain for improvements, to obtain thorough enforcement, and to gain new rights through legislation. Lastly, social norms along with worker relative bargaining power are reflected in company personnel policies and these also affect working conditions. Some observers see labor standards as an externally imposed cost and/or as having a distorting effect on labor markets. Others see regulation as a reflection of social norms and of power relations between employers and workers. In this view, regulation may hinder, add costs to, or facilitate greater efficiency in production; its effects are not unidirectional. Three recent trends – the emergence of new employment arrangements, weakened enforcement, and the expansion of cross-national production – challenge the ability of national governments to regulate working conditions and require further research.

Historically, working conditions have varied widely across countries and across time within each country in terms of their quality – from the workers' vantage point. Working conditions have been subject to regulation to varied degrees, a situation that accounts for variability in their quality.

Working conditions are part of employment systems that have been shaped by actions, rules, and regulations emanating from several spheres of society. These include the legal, judiciary, administrative, and collective bargaining spheres as well as personnel policies and informal social norms. All these spheres shape the 'way of doing business' (Piore, 1990) as well as working conditions for workers. In industrialized societies, the regulation of working conditions has primarily affected people in waged or salaried employment, that is, those who are in a relationship, recognized legally as economically dependent, with an employer who usually controls the means of production. In recent years and in some countries, however, some labor regulations have been applied to the self-employed.

Working Conditions, Definition

Working conditions consist of the following elements:

1. *The terms of the employment relationship*. They include whether the employment relationship is durable or for short term, formal or informal. The terms of employment may be explicit or implicit.
2. *Access to employment*. This access may be equitable or discriminatory. In the latter case, groups of workers may be excluded from employment or have differential access based on their race/ethnicity, gender, age, or other characteristic.
3. *The workplace itself*. Work may be performed in a single site, multiple sites, or from the worker's home.
4. *The wage level*.
5. *Time at work and away from work*. This includes the daily, weekly, and yearly number of work hours as well as their scheduling. It also includes provisions for time away from work due to illness, vacation, or leave for family or medical reasons or for training.
6. *The supervisory structure and mechanisms*. This structure may or may not include procedures for due processes in disciplining, discharges, and layoffs. It may entail varied degrees of protection from supervisory arbitrariness.
7. *Health and safety conditions in the workplace*. These include occupational health, that is, the ways in which the performance of work tasks affects workers' physical health and well-being, and environmental health, that is, whether the work environment in which the work is performed has negative effects on physical health.
8. *Voice at work*. This encompasses access to worker representation and the shape of representation structures. These structures may include representative organizations such as

labor/trade unions or worker associations as well as participatory structures such as workplace councils.

9. *Social protection/social insurance benefits for the individual and their dependents that have been tied to waged employment and are administered through payroll systems.*

Some of the elements of working conditions listed above have a direct relationship to employment and the performance of work, e.g., health and safety conditions at work. Others have been historically tied to employment with a particular employer in the United States and to labor market participation in Western European countries but do not have to be. For example, health insurance coverage can also be provided as a universal benefit regardless of work activity.

Systems of Employment and Working Conditions across Countries

National systems of employment can be viewed as a whole that provides degrees of regulation and protection to workers and employers in a variety of ways (Jacoby, 1982; Boyer, 1986; Piore, 1986, 2010). Systems of employment can be seen as a conjunction of 'spheres'; they encompass legal, executive/administrative, judiciary, collective bargaining, and employer personnel policy 'spheres' as well as social norms and customs. Different countries have a varied mix of reliance on these spheres to affect working conditions. Some rely on collective bargaining approaches, whereas others favor the action of the legislative and executive spheres combined. The regulation of working conditions can therefore be achieved through these spheres in a different manner across countries as well as across time in one country.

The Purpose of Regulation on Working Conditions

The regulation of working conditions has several purposes. Most commonly, it establishes labor standards, that is, sets minima for wage, hours, or health and safety conditions that employers are mandated to respect and that are enforced through prohibition and penalties. This is the narrowest role of regulation.

Regulation can enable the formation of structures for negotiation, mediation, or problem solving that contribute to shaping working conditions in varied ways and with diverse outcomes across industrial sectors and occupations (Weiler, 1990). One example of such regulation is the federal and state statutes governing worker representation and collective bargaining (e.g., US National Labor Relations Act of 1935).

Sources of Regulation

Most narrowly and conventionally, regulation is understood as emanating from government (national, state, local) through laws and statutes (legislative branch), as well as administrative rules (executive). There is a range of enforcement philosophies and mechanisms across countries. For example, in the United States, separate agencies have been established over time to enforce different standards. The Wages and Hours Division in the US Department of Labor enforces the Fair Labor Standards Act of 1938 and related legislation on minimum wages and overtime pay for hourly workers, while the Occupational Health and Safety Act sets minima for safe working conditions. Through inspectors and administrative courts, violations are verified and penalties issued for responsible employers. In contrast, France and to some extent Marocco and Mexico rely on a professional body of labor inspectors to monitor and be the first level of assessment of a wide range of labor standards and health and safety conditions (Piore, 2010). Enforcement mechanisms and their effectiveness play a key role in the shaping of working conditions. Conversely, the work setting affects the effectiveness of enforcement. For example, if workplaces are small, scattered, and numerous, it is difficult for an enforcement agency that relies on visits by inspectors to monitor work conditions and enforce a specific law.

The court system is another feature of the regulation of working conditions. Courts contribute to enforcing existing laws and statutes. For example, an employer responsible for hazardous working conditions that have resulted in worker death may be sued. Sometimes, in response to cases initiated by workers, courts issue decisions in areas where government legislation and executive enforcement have not acted yet. In this case, court actions precede legislative action. A number of countries (e.g., France, Germany) rely on peer representation labor courts (employer, worker/union representatives), which address labor enforcement issues that cannot be mediated by labor inspectors. These courts rely on a mix of mediation and adjudication processes.

Working conditions are also affected by the relative bargaining power of workers and their organizations. In turn, this bargaining power is shaped partly by the provisions made in legislation for worker voice and representation. For example, the US National Labor Relations Act of 1935 established a mechanism for representation for many workers in the post–World War II years and also shaped the form that labor unions took. In some European countries, where there is administrative extension of terms of a collective agreement to whole sectors, workers are directly affected by conditions in union collective agreements with leading firms. The relative bargaining power of labor unions and worker organizations affects the ability of workers to bargain for improvements, the thoroughness of enforcement of existing laws, and the ability to obtain new laws.

Lastly, social norms and customs along with worker relative bargaining power are reflected in personnel policies and employer practices; these also affect working conditions. In countries with a limited role for government regulation and with low incidence of collective bargaining, such as the United States, many aspects of working conditions are shaped by personnel policy rather than law or collective bargaining and thus vary greatly across employers.

In sum, working conditions are shaped by the totality of action in these respective spheres and by the interaction of actions from these spheres. For example, in the United States, provisions in collective bargaining agreements or innovative personnel policies may serve as a model for later state or federal legislation. In France, national sectoral bargaining in lead industrial sectors on a particular aspect of working conditions often precedes national legislation. (National

sectoral bargaining occurs between large employers and their federations, on the one hand, and national unions and their federations, on the other hand, and is sometimes mediated by government.) In some countries, national sectoral bargaining on a particular issue (e.g., implementing shift work) is used sometimes as an alternative to government regulation.

The Long-Standing Debate on the Regulation of Working Conditions

The role and benefits of the regulation of working conditions has been the subject of long-standing debate. One view sees regulation, particularly regulation on labor standards, as an externally imposed cost and/or as having a distorting effect on the operation of labor markets. This view stems from the neoclassical school of economics whose core perspective – presented here in stylized fashion – is that markets are self-regulating and that prices must be allowed to vary freely in order to be efficient allocators of human and capital resources. This view also tends to assume that there is no significant conflict or power differential between workers as a group and employers as a group. Rather, workers and employers are individual decision makers who make choices unhampered by power differentials. Critics of minimum wage laws have argued that the laws push wages above the 'market clearing' level and thus cause unemployment. Other analysts have taken a more benign view of their impacts on firm decisions and employment levels. The research literature has centered on the empirical measurement of the relative magnitude of effects on employer costs and employment levels (Dube et al., 2010; Houseman, 1998).

Another view sees the regulation of working conditions and its effects as a special case in economic analysis because labor markets are deemed the least self-regulating among all markets. In this view, the regulation of labor markets cannot easily be analyzed with the conventional tools of economic analysis.

Consistent with this view, another perspective sees regulation as a reflection of customs and social norms and of power relations between employers and workers – all of which vary across countries (Piore, 1986; Gautié and Schmitt, 2010). In this view, regulation may hinder, add costs to, or facilitate greater efficiency in production; the effects of the regulation of working conditions are not conceived as unidirectional. They are expected to vary widely depending on the contents of regulation, the structure of particular industries, and their production organization. For example, minimal labor standard on wages, hours, or health and safety have been argued to discourage 'sweatshops' and favor other business strategies that are less onerous to workers (Piore, 1990). In sweatshops, workers are unprotected because employers are insulated from the economic impact of variations in productivity among workers and over time thanks to the piece rate system (workers are paid for measurable units of output). A minimum wage rate that is high enough to create employer concern with productivity may compel a change in a firm's practice and working conditions; it may deter the use of low-productivity workers such as young children.

Recent debates on the regulation of working conditions in North America and parts of Europe have been colored by the general policy trend toward deregulation of markets. Laws on hours and scheduling, and those on worker representation structures, have been questioned. Advocates of deregulation have argued that existing regulations fit an old order of production organization, and market structure, and must be removed. Others have called for their replacement with different standards and representation systems that provide equivalent protections for workers but through new means that are more appropriate for contemporary forms of production and work organization.

Recent Trends and Implications

Diversified Employment Arrangements

In the post–World War II period, the contents and enforcement of the regulation of working conditions in late industrialized countries have been constructed, albeit in different ways, around a certain notion of the employment relationship as a relatively long-term, stable attachment to a single employer with the work tasks being performed on the employer's work site. While there were always exceptions to this notion – for example, farm labor, construction, expert consulting, or domestic labor – the bulk of postwar regulatory attention to working conditions has been structured around this employment arrangement.

In the future, most observers expect less employer commitment to long-term employment relationships, less worker attachment to particular employers, and greater lateral mobility (across firms, industries, and possibly occupation). This is due to changes in the organization of production and employment structures within firms; increasingly, firms recruit workers from the outside labor market rather than relying on internal training and promotion. Career paths for workers will become less predictable and may span several companies and industries (Arthur and Rousseau, 1996; Cappelli et al., 1997).

The manifestations of these transformations include the emergence and establishment of varied forms of employment that are explicitly short term (limited-duration hiring, temporary work), may be brokered by an intermediary (a for-profit or nonprofit temporary/staffing service), or may be the result of contracting arrangements between two businesses (e.g., workforce of a janitorial subcontractor), thus creating a triangulated employment relationship. Independent contracting has also become more common – a situation where individual workers are treated as self-employed under the law, that is, not wage/salary employees, while they are economically dependent on a relationship with a sole purchaser of their services.

Short-term employment relationships may render existing regulations and their enforcement ineffective; workers may not stay long enough in a particular job to bring a complaint about a violation of a labor standard. Triangulated employment relationships entail ambiguities as to employer responsibility for compliance with particular laws. Independent contracting means that there is no employment relationship and therefore no option for regulation under most labor and employment law, leaving solely collective organization for bargaining and mobilization for policy advocacy as alternative venues for

affecting working conditions. In practice, the financing and implementation of social protection benefits such as health insurance and pension through employment relationships are weakened and, in countries with employer-based systems, rendered ineffectual.

As employment relationships change in these ways, the contents and means of the regulation of working conditions will likely change. The regulation of working conditions will require a reconceptualization and analysis.

Although employment relationships change in these ways, they still remain the primary means through which core social protection benefits continue to be provided. For example, in the United States, which has relied on social protection benefits being provided by employers either voluntarily (fostered by tax deduction incentives) or through rare mandates, regulatory attention has focused on mechanisms to better enable workers to meet their care responsibilities. For example, family leave – time off for caring for a child or sick relative – has been the subject of US legislation (1993), of continued activism at the state and federal level, and of innovative benefit policies by some large employers.

Enforcement of Regulation

Observers have also noted that related changes in the organization of production and firm practices have changed the context for enforcement of legal statutes on working conditions, thereby restricting reliance on legal mechanisms of regulation. The production of goods and services through networks of firms with varied ownership structure and a diversity of interfirm relationships (alliance, franchise, subsidiary, subcontracting) has created labor environments where workers experience difficulty knowing how to exert their statutory rights, and traditional enforcement mechanisms are less effective (Weil, 2010). Developed countries have witnessed difficulties with standards enforcement. Notably in the United States, rates of violation of wage and hours laws, of workers' compensation, or health and safety rules have been high in several industries in a number of urban markets (e.g., restaurants, retail, delivery services, garments (Theodore et al., 2012)). Enforcement challenges have led scholarship to explore models of enforcement, contrasting administrative enforcement of single statutes (e.g., US system) to the interpretation and enforcement of labor standards as a whole through a professional corps of inspectors such as in France, Morocco, or the Dominican Republic (Piore, 2010).

The Regulation of Working Conditions and Globalization of Production

International economic integration has continued to grow in recent years. Its main features include the reduction or removal of trade tariffs and of restrictions on capital flows either by policy change or simply through the spread of new communication technologies, which put distant production areas directly in competition. It has resulted in greater cross-border capital mobility, the freer movement of goods, and the development of production processes ('production chains') that span national boundaries. This pattern holds particularly for manufacturing and, increasingly, for service activities such as call centers or records transcription (Batt et al., 2009; Whitman, 1999).

In this context, the issue of the regulation of working conditions will take on new facets requiring further consideration and research. First, the adaptation and strategic responses of firms in industrialized countries to minimal labor standards were developed during the post–World War II period when mass production was in expansion, and domestic markets were relatively sheltered from foreign competition. Growing competition from goods and services produced in countries with dramatically lower wages and fewer regulations of working conditions (e.g., Bangladesh, Cambodia), as well as the increased opportunities for firms to expand their production processes cross-nationally, are continuing to change the parameters that affect working conditions for workers in both industrialized and developing countries. These developments challenge the ability of national and local governments to regulate working conditions in ways that ensure effective enforcement.

Second, in developing countries, the growth of activities linked to the production processes of cross-national firms has given rise to extensive subcontracting production networks where monitoring is difficult. It has also favored the extension of forms of employment (e.g., off-site contract labor such as home-based light assembly) that are not subject to existing national regulation of working conditions, or that take place in settings and on terms in which monitoring and enforcement are difficult (see research on 'informal employment').

Additionally, the ability of countries to enforce the regulation of working conditions through labor standards in those developing countries with low wages and low-quality working conditions is weakened by the ability of cross-national firms to shift production to another low-wage country with a workforce of similar skill level. Over the past decade, very large, visible, multinational corporations have addressed the issue of labor standards enforcement in their long, diverse, subcontractor chains through voluntary codes of conduct and 'ethical trade' certifications, including monitoring by the corporation, nongovernmental organizations, or other recognized third party (Locke et al., 2007 for a review). The degree and quality of labor standards enforcement achieved through these mechanisms varies greatly, depending not only on investment in monitoring but also on the productive quality of the component firms of the production network. Going forward, some multinational corporations are exploring other approaches directly focused on how the production process itself (work organization and technology use) generates poor working conditions (Locke et al., 2007).

Outlook and Future Research Issues

These trends raise several concerns. For workers, there is increasing risk of worsening working conditions, that is, increased difficulty with accessing jobs subject to the full extent of regulations and lower odds of benefiting from thorough enforcement of their rights and protections at work. The quality of employment is likely to decline – if policy thinking and national regulatory approaches do not adapt to these changes.

Three groups of issues will need to be considered in future research. First, the parameters of firms' choices will need to be documented in the context of particular countries and for specific activities, contrasting capital-intensive and labor-intensive manufacturing as well as comparing service activities among themselves.

Second, future research on regulation will increasingly examine the sphere of international organizations (e.g., World Trade Organization, Organization for Economic Cooperation and Development, International Labor Organization, European Union), the process of international negotiation of international conventions on working conditions, and the impacts of these conventions, first, on national regulation efforts and, second, on firm strategies and practices. Research has developed and will continue to explore ways to achieve implementation of labor standards that are adaptable to national institutional and social settings as well as specific industry contexts. This research is relevant to working conditions in developed as well as developing economies.

Lastly, for policy analysis and discussion, the core question will be what *should* be the complex of regulation of working conditions that is most compatible with new forms of cross-national production, the productive use of new production and communication technologies, and the well-being of workers in industrialized and developing countries.

See also: Economic Sociology; Labor Market Policy; Regulatory Agencies; Welfare Policy Comparisons.

Bibliography

Arthur, M.B., Rousseau, D.M. (Eds.), 1996. The Boundaryless Career: A New Employment Principle for a New Organizational Era. Oxford University Press, New York.

Batt, R., Holman, D., Holtgrewe, U., 2009. The globalization of service work: comparative institutional perspectives on call centers. Industrial & Labor Relations Review 62 (4), 453–488.

Boyer, R., 1986. La flexibilité du travail en Europe. Editions La Découverte, Paris.

Cappelli, P., Bassi, L., Katz, H.C., Knoke, D., Osterman, P., Useem, M., 1997. Change at Work. Oxford University Press, New York.

Doeringer, P.B., Piore, M.J., 1971. Internal Labor Markets and Manpower Analysis. D.C. Heath, Lexington, MA.

Dube, A., Lester, T.W., Reich, M., 2010. Minimum wage effects across state borders: estimates using contiguous counties. The Review of Economics and Statistics 92 (4), 945–964.

Gautié, J., Schmitt, J. (Eds.), 2010. Low-Wage Work in the Wealthy World. Russell Sage Foundation, New York.

Houseman, S., 1998. The effects of employer mandates. In: Freeman, R.B., Gottschalk, P. (Eds.), Generating Jobs: How to Increase Demand for Less-Skilled Workers. Russell Sage Foundation, New York, pp. 154–191.

Jacoby, S.M., 1982. The duration of indefinite employment contracts in the United States and England: an historical approach. Comparative Labor Law 5, 85–128.

Jacoby, S.M., 1997. Modern Manors: Welfare Capitalism since the New Deal. Princeton University Press, Princeton, NJ.

Locke, R.M., Fei, Q.I.N., Brause, A., 2007. Does monitoring improve labor standards? Lessons from Nike. Industrial & Labor Relations Review I 61 (1), 3–31. Article 1. Available at: http://digitalcommons.ilr.cornell.edu/ilrreview/vol61/iss1/1.

Piore, M.J., 1986. Perspectives on labor market flexibility. Industrial Relations 25, 146–166.

Piore, M.J., 1990. Labor standards and business strategies. In: Herzenberg, S., Perez-Lopez, J.F. (Eds.), Labor Standards and Development in the Global Economy. Government Printing Office, Washington, DC, pp. 35–49.

Piore, M.J., 2010. Flexible Bureaucracies in Labor Market Regulation. The Idea of Labour Law International Workshop, April 8–9, rev. September. St Catharine's College, Cambridge, UK.

Theodore, N., Bernhardt, A., DeFlippis, J., Milkman, R., Heckathorn, D., Auer, M., Gonzalez, A.L., Narro, V., Perelshteyn, J., Polson, D., Spiller, M., 2012. Under the radar: tracking the violation of labor standards in low-wage industries in the US. In: Warhurst, C., Carré, F., Findlay, P., Tilly, C. (Eds.), Are Bad Jobs Inevitable Trends, Determinants, and Responses to Job Quality in the Twenty-first Century. Palgrave, Basingstoke, UK, pp. 208–223.

Weil, D., 2010. Improving Workplace Conditions through Strategic Enforcement. Boston University School of Management Research Paper No. 2010-20, Boston.

Weiler, P.C., 1990. Governing the Workplace: The Future of Labor and Employment Law. Harvard University Press, Cambridge, MA.

Whitman, M.N., 1999. New World, New Rules: The Changing World of the American Corporation. Harvard Business School Press, Boston, MA.

Working Memory, Neural Basis of

Michael Petrides, Montreal Neurological Institute, McGill University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Working memory, which refers to the temporary storage of recently perceived and interpreted information for further processing and cognitive manipulation, depends on distributed neural systems involving multiple brain regions that include several prefrontal and higher-order temporal and parietal cortical areas. There is considerable evidence that the temporary storage of information depends primarily on modality-specific and multisensory temporal and parietal cortical areas. Distinct prefrontal areas in interaction with these posterior storage areas support various aspects of control processes such as the monitoring of the contents of working memory (known as the epoptic process), rule-based attentional selection, active-controlled retrieval, inhibitory control, etc. The mid-dorsolateral prefrontal region, which underlies the monitoring of the contents of working memory, in interaction with certain areas of the posterior parietal cortex, forms a fronto-parietal circuit for the manipulation of information in working memory.

The present article focuses on research that aims to identify the role of various areas of the cerebral cortex in what has been termed 'working memory' at the cognitive level of analysis. In experimental psychology, the term 'working memory' has been used to refer to the temporary storage of information that is necessary for cognitive processing underlying the comprehension and learning of new information, reasoning, mental arithmetic, the mental manipulation of information, etc. This theoretical construct emerged in the 1970s out of earlier experimental behavioral work on short-term memory (see Baddeley, 1986, 2007), although the term 'working memory' was used as early as 1960 by Pribram, Miller, and Galanter in an attempt to deal with cognitive processing related to the planning of action. The term working memory has been defined variously to refer not only to the transient maintenance of information (short-term memory), but also to the further processing, including manipulation, of that information. For instance, one of the most popular working memory models in cognitive psychology assumes the existence of a central executive system supplemented by 'slave' systems that are specialized for the maintenance and processing of verbal, visuospatial, and episodic information (Baddeley, 1986, 2007). Such an expanded concept of a working memory system becomes almost tantamount to the online processing of information that occupies current awareness. Naturally, any system defined so broadly will be the result of neural processing across multiple areas of the brain. Investigations of working memory impairments in human patients with relatively restricted brain damage and in monkeys with precise excisions of specific cortical areas have begun uncovering some of the key cortical structures involved in working memory. It has been shown that modality-specific impairments in the temporary storage and processing of incoming and recalled information result from damage to particular cortical areas in the temporal and parietal lobes. Impairments on working memory tasks after lesions to prefrontal cortical areas are considered to reflect deficits in executive processing that is necessary to allocate attention to different types of information, inhibitory control, and to monitor the contents of working memory, the latter process also referred to as epoptic processing (Petrides, 2013).

The Prefrontal Cortex and Working Memory

There is considerable evidence that the prefrontal cortex plays a major role in working memory. The prefrontal cortical region, which lies anterior to the motor/premotor frontal cortex, has expanded enormously in the human and nonhuman primate brains and comprises many distinct areas that differ in terms of their cytoarchitecture (i.e., cellular composition) and their connections with other parts of the brain (**Figure 1**; Petrides, 2005). These anatomical facts imply that there must be differences in the local neuronal computations occurring within these areas since both the local circuitry, reflected in cytoarchitectonic structure, and their connections with other brain areas are distinct. The first evidence that the prefrontal cortex plays a major role in working memory came from research in the 1930s by Jacobsen and coworkers, who examined the performance of monkeys with lesions restricted to the prefrontal cortex on the delayed response task (Jacobsen, 1936). This work showed that frontal cortical damage rendered animals incapable of performing the delayed response task which was introduced by Hunter in 1913 to study transient memory in animals. This task typically is conducted in the following manner. The experimenter attracts the monkey's attention and, in full view of the animal, a reward (e.g., a piece of food) is placed under one of two identical stimuli (e.g., cups, plaques, etc.). The stimuli are then hidden from the animal's view by an opaque screen and, after a delay of a few seconds, the screen is raised and the animal is allowed to choose one of the two stimuli to obtain the reward if, of course, the choice is correct (**Figure 2**). In interpreting the impairment on delayed response in monkeys with frontal cortical damage, Jacobsen (1936: p. 52) argued that the behavior of such animals "instead of being directed by a balance of immediate, recent and temporally remote stimulation, is dominated by present external stimuli and the stable habit systems derived from earlier training; the mnemonic influence of past experience is lacking."

The demonstration that adequate performance of the delayed response and related tasks is impaired severely by lesions of the prefrontal cortex has been the starting point of several investigations into its possible role in working memory.

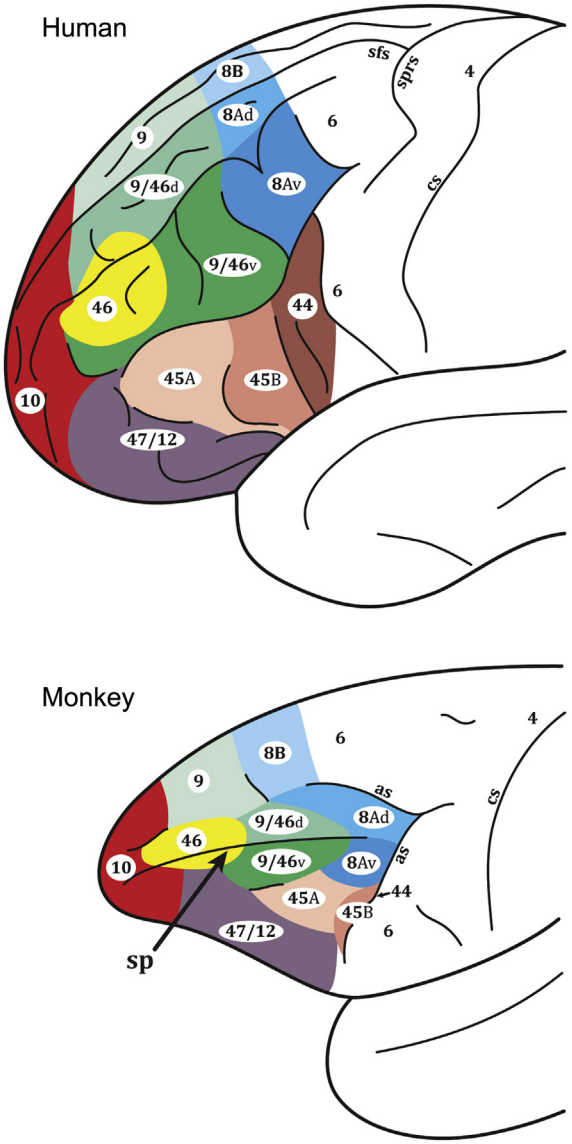

Figure 1 Cytoarchitectonic maps of the prefrontal cortex on the lateral surface of the human and the macaque monkey brains by Petrides and Pandya (see Petrides, 2005). Note that the white region in front of the central sulcus represents the motor (area 4) and premotor (area 6) cortex. Several prefrontal cortical areas can be defined based on differences in cellular architecture. Note that the cortex within and surrounding the sulcus principalis (see arrow) is often referred to as area 46 in the older literature (e.g., Goldman-Rakic, 1996), but Petrides and Pandya have subdivided old area 46 into area 46 proper (yellow) and related areas 9/46d and 9/46v (green) based on comparative cytoarchitectonic analysis of human and macaque monkey prefrontal cortex. as, arcuate sulcus; cs, central sulcus; sfs, superior frontal sulcus; sp, sulcus principalis, also known as principal sulcus (arrow); sprs, superior precentral sulcus.

The lesion investigated by Jacobsen involved the entire prefrontal cortex. Subsequent research on monkeys by Mishkin, Pribram, Gross, Weiskrantz, and others has demonstrated that (1) lesions restricted to the lateral surface of the prefrontal cortex impair delayed response performance, (2) the delay is critical for the impairment to emerge, and (3) the impairment on the classical spatial version of the delayed response task can be localized around the sulcus principalis, a prominent sulcus that runs longitudinally in the middle of the lateral frontal cortex of the monkey and where area 46 lies (see Petrides, 1994, for review of this literature and illustrations of the experimental paradigms used). Note that, in the older literature, the definition of area 46 in the monkey includes area 46 proper and related area 9/46 as more recently defined by Petrides and Pandya on the basis of comparative cytoarchitectonic analysis (**Figure 1**). Thus, in older and some recent lesion and recording studies, area 46 is used to refer to both areas 46 and 9/46.

By the early 1970s, the method of recording the activity of isolated neurons in monkeys as they perform various cognitive tasks was perfected and a number of studies appeared that recorded the activity of single neurons in the prefrontal cortex during delayed response performance. Fuster and Alexander (1971) in the United States and Kubota and Niki (1971) in Japan were the first to record such neuronal activity and demonstrated that certain neurons in the lateral prefrontal cortex respond specifically when a particular location is indicated as the one to be remembered and continue to discharge during the delay period until the choice has been made. This early work and several later studies by Goldman-Rakic, Miller, and coworkers clearly established the existence within the prefrontal cortex of sustained neuronal activity during the delay period of various working memory tasks (see Fuster, 2008, for review).

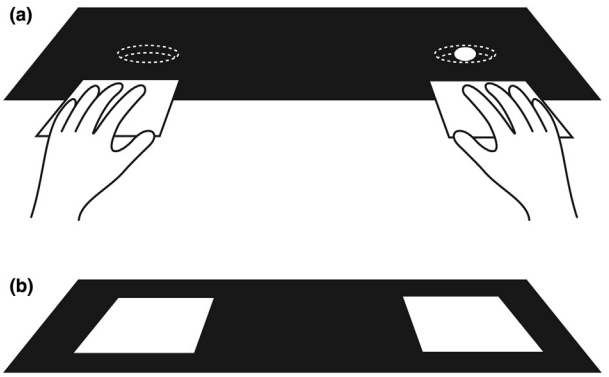

Figure 2 In full view of the monkey, one of two foodwells, selected randomly, is baited with a reward (a) and then the two foodwells are covered by two identical stimuli (here square white plaques). (b) An opaque screen is lowered and, after a short delay, the screen is raised to allow the monkey to uncover one of the two plaques to find the reward. Uncovering the foodwell where the reward had been placed before the foodwells were covered and the delay imposed provides the monkey with a reward. Damage to the region of the sulcus principalis (see **Figure 1**, arrow) results in a severe impairment on this task. In the older literature, this impairment was interpreted as a location memory impairment resulting from damage to the cortex of the sulcus principalis (area 46). However, Petrides (1994, 2005) has shown that the problem arises from the requirement to keep track (i.e., monitor) the constantly changing location of the stimuli and that the impairment is not restricted to spatial stimuli, but is a general monitoring impairment in working memory that involves both spatial and nonspatial stimuli.

Thus, both the lesion and the single-neuron research indicated the involvement of the prefrontal cortex in some aspect of working memory. But what exactly is the role of the prefrontal cortex in working memory? In one of the major theoretical interpretations of the above data, Goldman-Rakic (1996) assumed that the prefrontal cortex is a critical region for the short-term maintenance of processed information and considered the different prefrontal cortical areas as specialized working memory modules for the maintenance of different types of information for the purpose of future action. For instance, it was argued that the dorsolateral prefrontal area 46 (i.e., areas 46 and 9/46 in the newer definition; **Figure 1**) is specialized for spatial working memory (i.e., the maintenance of location-related information), whereas the ventrolateral prefrontal area 45 is a specialized module for nonspatial visual working memory (i.e., the maintenance of object-related information; Goldman-Rakic, 1996). An alternative theoretical model of the role of the different lateral prefrontal cortical areas in working memory was proposed by Petrides (1994, 1996), based on research that further analyzed the nature of the impairment on various working memory tasks by human patients and monkeys with prefrontal lesions. Petrides argued that the existence of delay neurons in the prefrontal cortex, although implying that some neuronal computations are occurring there during the delay in working memory tasks, does not require the conclusion that the primary role of the prefrontal cortical areas is the short-term maintenance of information. For instance, neurons with sustained activity during the delay in working memory tasks can be observed in virtually every part of the cerebral cortex, as well as in several subcortical structures (see Fuster, 2008, for review). In addition, there was the major puzzle that patients with damage to the lateral prefrontal cortical region can perform well on many standard short-term memory tests such as simple recognition of recently experienced stimuli, digit span, and immediate story recall. Petrides was able to demonstrate that patients with lesions that invaded the mid-dorsolateral prefrontal region (areas 46 and 9/46) and monkeys with lesions restricted to this particular part of the prefrontal cortex are impaired on working memory tasks in which successful performance cannot be based simply on the maintenance of information in short-term memory but requires, in addition, the executive process of monitoring that information (the epoptic process). In one of these tasks, the self-ordered working memory task, the subjects are presented with different arrangements of the same set of stimuli and, on each trial, they have to select a different stimulus until all have been selected once (**Figure 3**). Since the same stimuli are constantly recurring in the same context, there are no differences in the familiarity or the contextual occurrence of the stimuli that often support memory performance and, therefore, successful performance requires the subjects to monitor closely (keep track of) those stimuli that have already been selected and differentiate them from the constantly recurring and equally familiar stimuli that have not yet been selected. Several experiments have shown that, if familiarity or contextual cues are available to support working memory performance, patients and monkeys with mid-dorsolateral prefrontal lesions (areas 46 and 9/46) can perform normally (see Petrides, 1994, 1996). Similarly, if the monitoring requirements are minimal and memory performance is taxed by increases in delay, monkeys with area 46 and 9/46 lesions perform normally. In other words, impairments after area 46 and 9/46 lesions appear when events in working memory (e.g., the choices made by the subject in a particular task) must be monitored closely. Importantly, the deficit in the monitoring of events in working memory after lesions to areas 46 and 9/46 was not restricted to spatial information but was observed in nonspatial working memory tasks requiring the monitoring of object information (Petrides, 1994, 1996, 2013).

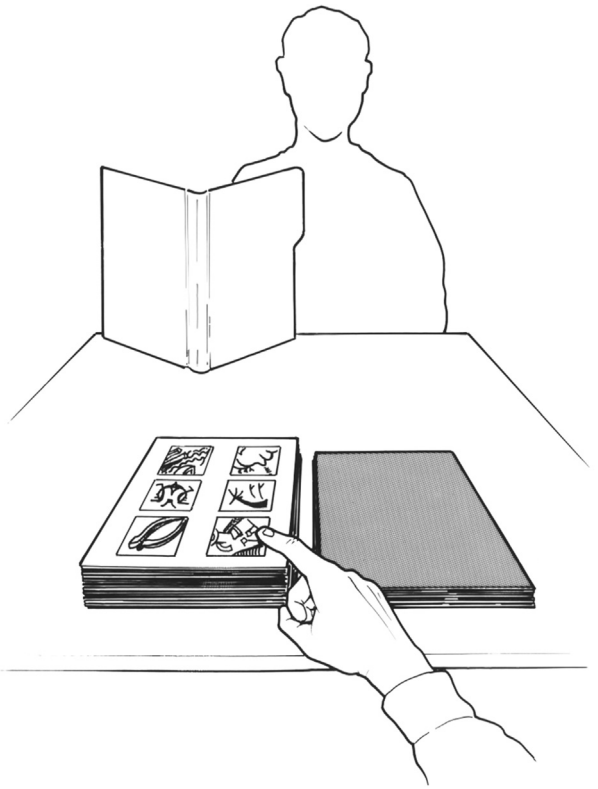

Figure 3 In this self-ordered pointing working memory task, the patient has to select one of the designs that appear on the card, turn over to the next card on which the same designs will appear but in different positions and select one of the designs that he/she had not previously selected, and so on. Importantly, the same designs appear on each card but in random arrangements. Thus, there are no differences in the familiarity or the contextual occurrence of the stimuli to support memory performance. Since the same, equally familiar designs appear all the time and in the same context, the patient can only perform well by keeping a mental record of the designs that he/she has already selected and differentiate them from the equally familiar designs that remain to be selected. This monitoring process (also known as the epoptic process) has been shown to depend on the mid-dorsolateral prefrontal cortex (areas 46 and 9/46) in human patient and macaque monkey lesion research and in functional neuroimaging studies with normal human subjects (see Petrides, 1996, 2005, 2013, for review).

Further evidence for the importance of the mid-dorsolateral prefrontal cortex to the monitoring of information within working memory is provided by functional neuroimaging studies on normal human subjects that have consistently

shown increased activity in the mid-dorsolateral prefrontal region (areas 46 and 9/46) when the monitoring requirements in working memory tasks are high, regardless of whether the stimulus material is spatial or nonspatial (see Petrides, 1996, 2013). When the monitoring requirements (e.g., keeping track of choices across trials) in working memory tasks are minimized, activity is not observed in the mid-dorsolateral prefrontal region, but in other parts of the prefrontal cortex depending on the nature of the control process being tapped. For instance, in a study by Courtney et al. (1998) in which monitoring and manipulation requirements in working memory were minimal, activity related to the mnemonic maintenance of spatial information was observed in the caudal superior frontal sulcus (sfs) close to its intersection with the superior precentral sulcus (sprs) where the frontal eye field lies (see **Figure 1** for location of sfs and sprs). The location-related memory activity was found to be immediately anterior to the frontal eye field (most probably in area 6). This finding is consistent with the conclusion of Petrides (1994, 1996) that simple spatial short-term memory is not the primary role of the mid-dorsolateral prefrontal region (areas 46 and 9/46) which is involved in monitoring the constantly changing status of multiple events in working memory, i.e. the contents of current awareness working memory (the epoptic process) for both spatial and nonspatial information. It is also consistent with the suggestion from several previous studies that posterior dorsolateral frontal cortex, where Courtney and colleagues observed increased activity during the delay in a simple spatial short-term memory task, plays a critical role in the top-down regulation of spatial attention both in the presence and in the absence of the spatial stimuli and, as such, particular parts of it can be expected to contribute to the maintenance of location-related information through top-down focusing of attention during a delay (Petrides, 1996, 2005). In other studies of attentional selection of visual stimuli in which monitoring and manipulation requirements were minimized, activity increases were observed again in posterior dorsolateral frontal areas (e.g., Petrides et al., 1993; Bledowski et al., 2009).

The Temporal and Parietal Cortex and Working Memory

Several studies on patients who had sustained lesions limited to posterior cortical regions have documented impairments in short-term memory, thus providing evidence that the temporary maintenance of information may be occurring in particular temporal and parietal cortical regions (see Vallar and Shallice, 1990). For instance, there are several reports of patients with a selective impairment in auditory–verbal short-term memory after lesions to the inferior parietal region of the language-dominant left hemisphere. These patients may be able to maintain and repeat single numbers, letters, or words presented auditorily, but their span (i.e., the number of items that they can maintain after a single presentation) is severely limited (e.g., Warrington and Shallice, 1969). Later on, with the emergence of functional neuroimaging, strong evidence was provided that the anterior part of the inferior parietal lobule, the supramarginal gyrus, is a critical cortical region for the phonological aspects of working memory (Paulesu et al., 1993; Church et al., 2011). For instance, in the fMRI study of reading by Church et al. (2011) in which the length of words and lexicality were manipulated, activity in the supramarginal region of the language-dominant left hemisphere was consistent with increased phonological processing. Interestingly, it was the posterior part of the supramarginal gyrus (area PFG) which was showing modulation of its activity in relation to phonological processing, namely the part that is known to be most strongly connected with area 44 (Broca's region) in the inferior frontal region (e.g., Margulies and Petrides, 2013). Thus, there is now strong evidence that the left supramarginal gyrus in interaction with ventrolateral prefrontal area 44 may underlie the phonological loop in working memory.

Patients with severely impaired short-term memory spans for visual–verbal material usually have lesions in the occipito-temporal region of the left hemisphere (e.g., Shallice and Saffran, 1986). De Renzi et al. (1977) have investigated visuospatial span using the Corsi blocks test. In this test, the patient observes the examiner tap, in sequence, a number of randomly arranged blocks and has to reproduce this sequence. Patients with right posterior hemisphere lesions were particularly impaired on this task, suggesting that the occipito-parietal cortex in the right hemisphere may be critical for the maintenance of visuospatial information (De Renzi et al., 1977). More recently, Della Sala and Logie (2002) have provided evidence from patients with brain damage that visuospatial and object working memory can be dissociated, a finding that is consistent with the demonstration of two separate streams of visual information processing, an occipito-temporal stream for visual object information and an occipito-parietal stream for spatial and action processing (Ungerleider and Mishkin, 1982; Goodale and Milner, 1992).

The notion that posterior parietal and lateral and ventral temporal cortical regions may play a role in the maintenance of information in working memory receives additional support from work in nonhuman primates. For instance, there is considerable evidence that the anterior part of the inferior temporal cortical region, which is known to be the final stage in the occipito-temporal neural system underlying visual object perception, is also critically involved in short-term visual memory. This evidence comes both from investigations of the effects of lesions (Mishkin, 1982) and from studies in which the activity of single neurons was recorded in monkeys as they performed visual short-term memory tasks (e.g., Fuster and Jervey, 1982; Miller et al., 1991). Indeed, a lesion study in the macaque monkey dissociated the contribution of the anterior inferotemporal region from that of the mid-dorsolateral prefrontal region (46 and 9/46) in visual object working memory (Petrides, 2000). Increasing the delay during which information had to be maintained in visual working memory (while keeping the monitoring requirements minimal) impaired performance in monkeys who had lesions to the anterior inferotemporal region, but not in those with lesions to the mid-dorsolateral prefrontal region. By contrast, increasing the number of stimuli that had to be monitored (while keeping the delay minimal) produced an impairment after mid-dorsolateral prefrontal lesions, but not after anterior inferotemporal lesions. This double dissociation between the effects of these two lesions on the same visual object working

memory task provides the strongest possible evidence of the selective contribution of these areas in working memory (Petrides, 2000). In addition, lesions restricted to certain parts of the superior temporal cortex in the monkey have been shown to lead to a serious impairment in auditory short-term memory (e.g., Colombo et al., 1990). Thus, selective impairments in the maintenance of information can be obtained from damage to specific parts of the lateral and ventral temporal and posterior parietal association cortical regions: Auditory impairments after superior temporal cortical lesions, visual object impairments after inferior temporal cortical regions, visuospatial impairments after occipito-parietal lesions, phonological impairments after language-dominant inferior parietal lesions, etc.

Prefrontal, Temporal, and Parietal Cortical Interactions for Working Memory

How can the evidence for impairments in working memory from posterior parietal and temporal association cortical regions, as well as from lateral prefrontal cortex, be conceptualized? In the theoretical model proposed by Petrides in 1994, neural activity in higher-order temporal and parietal cortical areas underlies not only perceptual processing and long-term storage of information but also short-term maintenance of recently processed information which has been provided by the environment and/or recalled (i.e., reactivated) from long-term storage (**Figure 4**). The temporary storage of information is viewed as an integral part of neuronal computation in modality-specific and multimodal temporal and posterior parietal cortical regions, i.e., particular areas that are parts of auditory, visual, tactile, spatial, vestibular, kinesthetic, multisensory and, in the human brain, verbal (i.e. phonological, lexical, and semantic) processing streams (Petrides, 1994, 1996, 2013). On the other hand, the prefrontal cortical areas are considered to play a role in working memory as distinct executive processing modules that allocate attention to different types of information and monitor the contents of information in these posterior cortical regions. Thus, working memory can be viewed as the multicomponent workspace of current cognitive activity, i.e., the processing and manipulation of information that occupies current awareness, and is based on local computations occurring within specialized posterior cortical regions for the online maintenance and processing of specific information in interaction with different prefrontal cortical regions that contribute various aspects of executive processing. The posterior lateral frontal region (such as area 8 and rostral premotor cortex) regulates the attentional focus and selection of stimuli, the ventrolateral prefrontal region (areas 45 and 47/12) underlies the active controlled retrieval of information, and the mid-dorsolateral prefrontal region (areas 46 and 9/46) monitors the contents of working memory, a process referred to as the epoptic process (Petrides, 1994, 1996, 2013).

Several investigators studying patients with brain damage have explored the special contributions of the various prefrontal regions in the executive control of cognitive processing (e.g., Robbins, 1996; Shallice, 2002; Stuss and Alexander, 2007; Petrides, 2013) and, although these explorations may not have always been in the context of working memory, the specific executive contributions of the various prefrontal areas demonstrated are likely to be reflected in working memory processing, depending on the requirements of the task, the strategy utilized by the subjects, etc. The same applies to functional neuroimaging studies of working memory (see next section).

A Fronto-Parietal Circuit for the Manipulation of Information in Working Memory

The mid-dorsolateral prefrontal region (areas 46 and 9/46; **Figures 1** and **4**) is strongly linked with the cortex of the inferior parietal lobule and the nearby intraparietal sulcal cortex, consistent with the fact that concomitant increases in activity within the mid-dorsolateral prefrontal cortex and the posterior parietal cortex are often observed in many functional neuroimaging studies of working memory. In the theoretical model of Goldman-Rakic (1996), the connections of the posterior parietal cortex with area 46 are viewed as the neuroanatomical substrate of the provision of spatial input to a prefrontal area specialized for the maintenance of spatial information in working memory. In the model of Petrides (1994, 1996, 2013), these parieto-frontal connections are viewed as not only bringing spatial information but primarily providing the means by which areas 46 and 9/46, a specific prefrontal region monitoring the contents of working memory (epoptic processing), can interact with critical components of the posterior parietal region for the cognitive manipulation of information. In other words, in the Petrides model, the parieto-frontal interactions are viewed as the basis of the mental manipulation of information in working memory. Evidence has recently been provided for such a fronto-parietal circuit for the mental manipulation of stimuli in working memory (see Petrides, 2013 for review). For instance, in an event-related fMRI experiment, Champod and Petrides (2007) were able to show that the posterior parietal cortex is involved in the manipulation (operationally defined as the mental rearrangement) of information in working memory, whereas the mid-dorsolateral prefrontal cortex monitors the information that is being manipulated. Thus, there is critical interaction between the mid-dorsolateral prefrontal cortex and certain parts of the posterior parietal cortex during the manipulation of information, as one would expect since the current status of information in working memory is constantly changing as the information is being manipulated. More recently, neuropsychological evidence has been provided that lesions to the posterior parietal region can produce impairments on tests involving the manipulation and rearrangement of information in working memory, but not on working memory tests requiring only rehearsal and retrieval processes or long-term memory (Koenigs et al., 2009). Although the lesions in the latter study were focused on the superior parietal region, lesions in patients inevitably damage nearby areas (e.g., intraparietal sulcal cortex) and therefore, the exact posterior parietal areas contributing to the mental manipulation of information require further definition.

Sometimes the difference in the theoretical positions of Goldman-Rakic (1996) and Petrides (1994, 1996) is incorrectly framed in terms of domain-specific versus domain-nonspecific

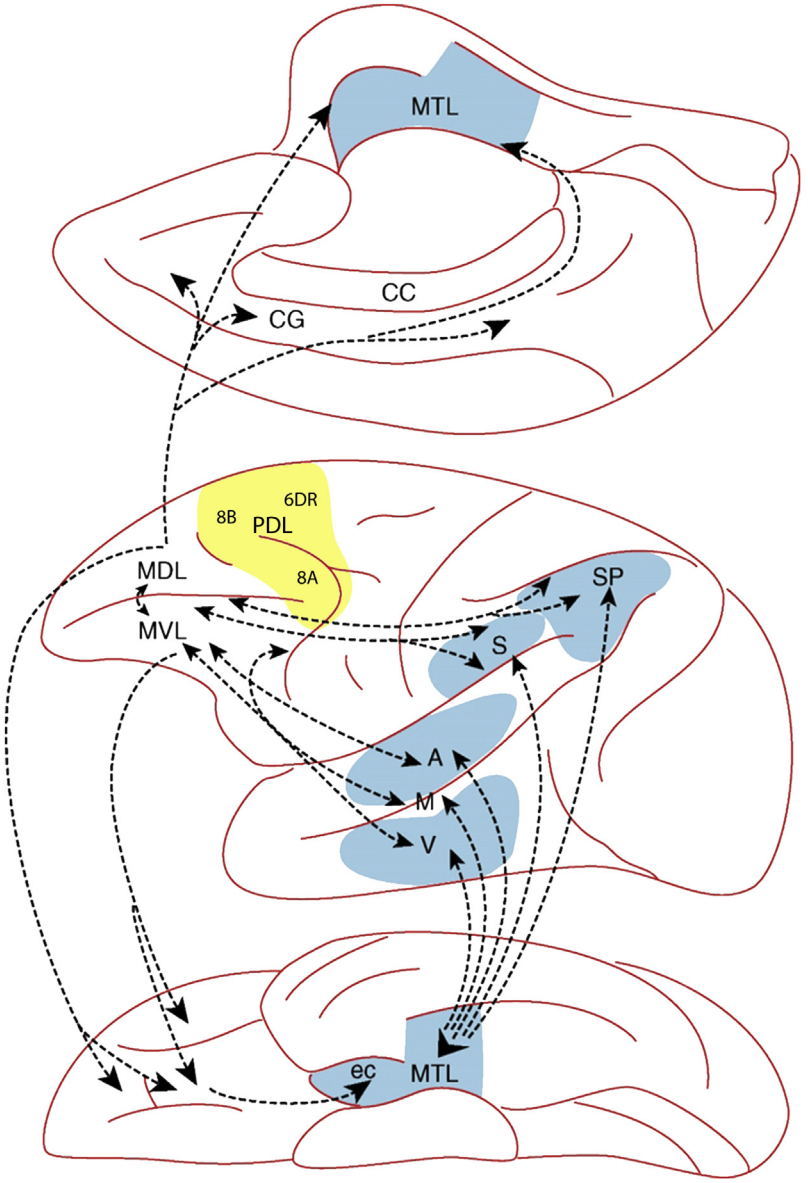

Figure 4 Schematic diagram of the model of the neural basis of working memory proposed by Petrides (1994) and which provides a framework to conceptualize the available experimental findings. Specialized areas in posterior parietal and temporal cortex that are part of somatosensory/somatomotor (S), spatial (SP), auditory (A), visual object (V), or multisensory (M) processing streams are critical for the short-term maintenance of processed information. In the language-dominant hemisphere of the human brain, there are additional specialized somatomotor verbal, auditory verbal, and visual verbal processing areas that underlie phonological and other language-related working memory processing (not shown in the schematic diagram above). Interactions with the mid-ventrolateral prefrontal region (MVL) permit active controlled retrieval of information from posterior cortical areas and interactions with mid-dorsolateral prefrontal (MDL) cortex permit the monitoring of information in working memory. Furthermore, interactions of certain frontal areas (e.g., MDL) with posterior parietal areas permit the cognitive manipulation and the monitoring of the manipulated information in working memory. Interactions between the posterior dorsolateral (PDL) frontal region (in yellow), which includes areas 8 and rostral area 6, with posterior parietal, occipital, and posterior temporal areas permit attentional selection (including conditional rule-based attentional selection) of processed stimuli and contribute to their short-term maintenance through top-down focusing of attention during delay periods. The connections of the posterior dorsolateral frontal region with posterior parietal and temporal areas are not shown so as to not crowd the diagram (see Petrides, 2005). CC, corpus callosum; CG, cingulate gyrus; ec, entorhinal cortex; MTL, medial temporal limbic region.

activations in the frontal region. It should be remembered here that the original debate between Goldman-Rakic and Petrides was not about the general existence or nonexistence of domain-specific dorsal versus ventral activations in the prefrontal cortex. The issue was about two specific prefrontal areas: dorsolateral area 46 (later subdivided into areas 46 and 9/46, see **Figure 1**) and ventrolateral area 45. Goldman-Rakic argued that these two areas could be conceptualized as performing information maintenance functions in working memory, with the fundamental difference being that area 46 maintains spatial information and area 45 maintains nonspatial visual object information (see Wilson et al., 1993;

Goldman-Rakic, 1996). By contrast, Petrides (1994, 1996) argued that area 46 is a critical module for monitoring (tracking of) the contents of working memory, *regardless of whether the material is spatial or nonspatial* and that area 45 is part of the ventrolateral prefrontal region that is involved in the active control of the retrieval of information from memory. Petrides made this argument on the basis of lesion studies to area 46 (which was later divided into areas 46 proper and 9/46) that yielded impairments on both spatial and nonspatial object working memory tasks provided the monitoring of the contents of working memory was taxed. Later, single-neuron recording studies showed that neurons coding location or object information were mixed in the part of the lateral prefrontal region where area 46 lies (Rao et al., 1997).

Often discussions of changes in activity in lateral frontal cortex in functional neuroimaging studies (partly influenced by the debate between Goldman-Rakic and Petrides) are couched in terms of dorsal versus ventral or domain-specific versus domain-nonspecific activations. Would increases of activity in prefrontal cortex be domain-specific or domain-general? Both types of activations should be expected in the prefrontal region during the performance of working memory tasks in functional neuroimaging studies since multiple executive requirements are likely to be recruited depending on the particular requirements of the task, the strategy adopted by the subjects, the control tasks against which activations are assessed, and several other factors (see, Owen, 1997; D'Esposito et al., 1998; Nystrom et al., 2000; Sala et al., 2003; Wager and Smith, 2003). Thus, the fundamental question would appear to be the *specification of the nature of the prefrontal executive process* that is operating in interaction with domain-specific (auditory, visual, tactile, spatial, verbal, etc.) and/or multisensory posterior parietal and temporal cortical areas, rather than a pure classification of prefrontal activations into dorsal (spatial) versus ventral (nonspatial; see Petrides, 1996, 2005, 2013).

In conclusion, there is evidence that the temporary storage of information depends primarily on modality-specific and multisensory temporal and posterior parietal cortical areas. Distinct prefrontal areas in interaction with these posterior storage areas support various aspects of higher-order control processing such as the monitoring of the contents of working memory (the epoptic process), rule-based attentional selection, active-controlled retrieval, inhibitory control, etc. The mid-dorsolateral prefrontal region, which underlies the monitoring of the contents of working memory, in interaction with certain areas of the posterior parietal cortex forms a fronto-parietal circuit for the manipulation of information in working memory.

See also: Cognitive Control (Executive Function): Role of Prefrontal Cortex; Prefrontal Cortex Development and Development of Cognitive Function; Prefrontal Cortex; Working Memory, Psychology of.

Bibliography

Baddeley, A.D., 1986. Working Memory. Oxford University Press, Oxford.

Baddeley, A.D., 2007. Working Memory, Thought and Action. Oxford University Press, Oxford.

Bledowski, C., Rahm, B., Rowe, J.B., 2009. What works in working memory? Separate systems for selection and updating of critical information. Journal of Neuroscience 29, 13735–13741.

Champod, A.-S., Petrides, M., 2007. Dissociable roles of the posterior parietal and the prefrontal cortex in manipulation and monitoring processes. Proceedings of the National Academy of Sciences USA 104, 14837–14842.

Church, J.A., Balota, D.A., Petersen, S.E., Schlaggar, B.L., 2011. Manipulation of length and lexicality localizes the functional neuroanatomy of phonological processing in adult readers. Journal of Cognitive Neuroscience 23, 1475–1493.

Colombo, M., D'Amato, M.R., Rodman, H.R., Gross, C.G., 1990. Auditory association cortex lesions impair auditory short-term memory in monkeys. Science 247, 336–338.

Courtney, S.M., Petit, L., Maisog, J.M., Ungerleider, L.G., Haxby, J.V., 1998. An area specialized for spatial working memory in human frontal cortex. Science 279, 1347–1351.

Della Sala, S., Logie, R.H., 2002. Neuropsychological impairments of visual and spatial working memory. In: Baddeley, A.D., Kopelman, M.D., Wilson, B.A. (Eds.), The Handbook of Memory Disorders, second ed. Wiley, Chichester, Chap. 13, pp. 271–292.

De Renzi, E., Faglioni, P., Previdi, P., 1977. Spatial memory and hemispheric locus of lesion. Cortex 13, 424–433.

D'Esposito, M., Aguirre, G.K., Zarahn, E., Ballard, D., Shin, R.K., Lease, J., 1998. Functional MRI studies of spatial and nonspatial working memory. Cognitive Brain Research 7, 1–13.

Fuster, J.M., 2008. The Prefrontal Cortex, fourth ed. Academic Press, New York.

Fuster, J.M., Alexander, G.E., 1971. Neuron activity related to short-term memory. Science 173, 652–654.

Fuster, J.M., Jervey, J.P., 1982. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. Journal of Neuroscience 2, 361–375.

Goldman-Rakic, P.S., 1996. Architecture of the prefrontal cortex and the central executive. Annals of the New York Academy of Science 769, 71–83.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends in Neurosciences 15, 20–25.

Jacobsen, C.F., 1936. Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. Comparative Psychology Monographs 13, 1–60.

Koenigs, M., Barbey, A.K., Postle, B.R., Grafman, J., 2009. Superior parietal cortex is critical for the manipulation of information in working memory. Journal of Neuroscience 29, 14980–14986.

Kubota, K., Niki, H., 1971. Prefrontal cortical unit activity and delayed alternation performance in monkeys. Journal of Neurophysiology 34, 337–347.

Margulies, D.S., Petrides, M., 2013. Distinct parietal and temporal connectivity profiles of ventrolateral frontal areas involved in language production. Journal of Neuroscience 33, 16846–16852.

Miller, E.K., Li, L., Desimone, R., 1991. A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254, 1377–1379.

Mishkin, M., 1982. A memory system in the monkey. Philosophical Transactions of the Royal Society of London, Series B 298, 85–95.

Nystrom, L.E., Braver, T.S., Sabb, F.W., Delgado, M.R., Noll, D.C., Cohen, J.D., 2000. Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. Neuroimage 11, 424–446.

Owen, A.M., 1997. The functional organization of working memory processes within human lateral frontal cortex: the contribution of functional neuroimaging. European Journal of Neuroscience 9, 1329–1339.

Paulesu, E., Frith, C.D., Frackowiak, R.S.J., 1993. The neural correlates of the verbal component of working memory. Nature 362, 342–345.

Petrides, M., 1994. Frontal lobes and working memory: evidence from investigations of the effects of cortical excisions in nonhuman primates. In: Boller, F., Grafman, J. (Eds.), Handbook of Neuropsychology, vol. 9. Elsevier, Amsterdam, pp. 59–82.

Petrides, M., 1996. Specialized systems for the processing of mnemonic information within the primate frontal cortex. Philosophical Transactions of the Royal Society, Series B 351, 1455–1462.

Petrides, M., 2000. Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. Journal of Neuroscience 20, 7496–7503.

Petrides, M., 2005. Lateral prefrontal cortex: architectonic and functional organization. Philosophical Transactions of the Royal Society, Series B 360, 781–795.

Petrides, M., 2013. The mid-dorsolateral prefronto-parietal network and the epoptic process. In: Stuss, D.T., Knight, R.T. (Eds.), Principles of Frontal Lobe Function, second ed. Oxford University Press, New York, Chap. 7, pp. 79–89.

Petrides, M., Alivisatos, B., Evans, A.C., Meyer, E., 1993. Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. Proceedings of the National Academy of Sciences USA 90, 873–877.

Pribram, K.H., Miller, G.A., Galanter, E., 1960. Plans and the Structure of Behavior. Holt, Rinehart and Winston, New York.

Rao, S., Rainer, G., Miller, E., 1997. Integration of what and where in the primate prefrontal cortex. Science 276, 821–824.

Robbins, T.W., 1996. Dissociating executive functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London, Series B 351, 1463–1470.

Sala, J.B., Rämä, P., Courtney, S.M., 2003. Functional topography of a distributed neural, system for spatial and nonspatial information maintenance in working memory. Neuropsychogia 41, 341–356.

Shallice, T., 2002. Fractionation of the supervisory system. In: Stuss, D.T., Knight, R.T. (Eds.), Principles of Frontal Lobe Function. Oxford University Press, New York, pp. 261–277.

Shallice, T., Saffran, E.M., 1986. Lexical processing in the absence of explicit word identification: evidence from a letter-by-letter reader. Cognitive Neuropsychology 3, 429–458.

Stuss, D.T., Alexander, M.P., 2007. Is there a dysexecutive syndrome? Philosophical Transactions of the Royal Society of London, Series B 362, 901–915.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.

Vallar, G., Shallice, T. (Eds.), 1990. Neuropsychological Impairments of Short-term Memory. Cambridge University Press, Cambridge.

Warrington, E.K., Shallice, T., 1969. The selective impairment of auditory verbal short-term memory. Brain 92, 885–896.

Wilson, F.A.W., O'Scalaidhe, S.P., Goldman-Rakic, P.S., 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 260, 1955–1958.

Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. Cognitive and Affective Behavioral Neuroscience 3, 255–274.

Working Memory, Psychology of

Robert H Logie, University of Edinburgh, Edinburgh, Scotland, UK

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Working Memory refers to a set of cognitive functions that allows humans to keep track of what they are thinking or doing moment to moment. It can also manipulate and process information such as when planning, learning, retrieving stored information, and transforming the information it holds. This article describes the theories that have been developed to explain these human abilities and some of the evidence for these theories from empirical studies of healthy adults and children and of individuals who suffer from cognitive deficits as a result of brain damage.

Definition of WM

Working Memory (WM) can be thought of as the desktop of the human brain, a set of cognitive functions that allows humans to keep track of what they are thinking, what they are doing, or where they are moment to moment. It holds information long enough to facilitate making a decision, learning a new word, mentally imaging the layout of a home, navigating around the world, or remembering what to do next. It can also manipulate and process information such as when planning, learning, retrieving stored information, and transforming the information it holds. This article describes the theories that have been developed to explain these human abilities and some of the experimental evidence for buttressing these theories that have been derived from empirical studies of healthy adults and children and of individuals who suffer from cognitive deficits as a result of brain damage.

Why Is the Concept of WM Useful?

The concept of WM has figured prominently in theories and explanations for human on-line cognition, as well as for offering a basis for understanding how humans accomplish a wide range of everyday tasks. It can be thought of as a mental tool that aids survival; if an organism is to interact successfully with its environment, then clearly there must be some means to record, on a temporary basis, its own most recent actions, keep track of rapid changes in the environment, have some means to plan actions for the immediate future, and to link changes in the environment with stored knowledge gained from past experience. This gives the organism a significant advantage over other organisms that are entirely driven by automatic responses to external stimuli that signal basic needs. Examples of how this mental tool is used in everyday interaction with the environment highlight its utility. Humans use WM every time they have to solve a problem, to perform a mental calculation, to recall where an object has just been laid down, to navigate around an unfamiliar environment, or to repeat an unfamiliar foreign word that has just been heard.

From these examples, it is clear that the concept is much broader than the more traditional idea of short-term memory. The concept of WM arose from the need to provide a theoretical account of a wide range of everyday tasks that require both temporary memory and on-line cognition. It is also much broader than what some people might think of as conscious experience. Our conscious experience is the result of how our brain interprets the information in our environment and stored information about past experiences. So, not everything that is available to conscious inspection necessarily reflects the actual functioning of on-line cognition, and not all functional aspects of on-line human cognition are available to conscious inspection (Pylyshyn, 1973). WM includes aspects of on-line cognition that are not necessarily available to conscious inspection as well as some elements that might be.

Until the early 1970s the focus of research on temporary memory and on-line cognition involved experimental studies of memory for random sequences of words, letters or digits, with the everyday example of recalling telephone numbers while dialing. However, a mental faculty would not have evolved to deal with these relatively artificial tasks, and therefore the concept of WM is intended to encapsulate on-line cognition and temporary verbal and visuospatial memory in both laboratory tasks and in everyday life.

Theoretical Views of WM

Most memory researchers consider WM to be a useful concept that uses relatively few assumptions to explain a wide range of experimental evidence as well as important aspects of everyday cognition; it is a system with limited capacity, which loses information over periods of seconds, whose contents are subject to displacement by new input, and which interacts and is supported by other parts of the cognitive system such as stored knowledge or the products of perception. Most theories of WM include these features; however, there remains a debate as to its precise characteristics. Some theories view memory as a largely unitary system with WM as the sum total of what is currently activated from the store of knowledge, skills, expertise, and past events. Others view WM as separate from the long-term store of knowledge and events accumulated over the lifetime. However, even among the theories that view WM as separate from the long-term store, there is some amount of discussion as to whether the cognitive functions of WM are supported by a single flexible system that provides both temporary storage and on-line processing, or that comprises multiple components, each specialized for a particular function, but that work

together to support task performance. There are multiple books as well as thousands of published research reports on different aspects of WM. As at the time of writing, books by Baddeley (2007) and Osaka et al. (2007) offer the broadest coverage, with more recent shorter reviews in Baddeley (2012) and Logie (2011).

The first type of theory, that memory is a unitary system and that WM is simply the currently activated knowledge, can account for a wide range of data from experimental psychology with healthy adults, including the ability to bring rapidly to mind information related to areas of personal expertise. Theories of this kind can very successfully explain how chess masters can play several games simultaneously, expert waiters can recall details of meal orders from large groups, and why burglars can recall better than householders, or indeed the police, details of houses related to their crime. Expertise extends the apparent capacity of WM, but there is some recognition among proponents of this type of theory of the need for a temporary memory system to deal with completely novel or unfamiliar material. There is also a need to have some kind of memory system that can be updated moment to moment: For example, in order to understand what you are reading now, you have to remember what you read immediately before, and to continually update your memory for what you have just read as you progress through the text. Something similar is needed to keep track of a conversation, or of cars nearby when driving on a busy motorway. In the distant past when humans did not have supermarkets and had to hunt for their food or to escape a predator, there would have been a need to continually update memory for where the animal is as it moves through the trees and bushes if there is any chance of catching something to eat or to avoid being eaten. The same would still be true for many animals and also some humans today. In fact, anything that we do throughout our waking hours that involves moment to moment changes in our environment requires a WM that can keep track of those changes, and humans would not have survived very long if they did not have some form of WM.

The view that WM is no more than the sum total of currently activated experience and memories for past events also runs into difficulty when considering dissociations that are apparent from studies of healthy adults performing more than one task at the same time. There is substantial evidence that healthy adults can perform two quite demanding tasks concurrently such as walking and talking. In research under more controlled conditions this can be studied by, for example, hearing and recalling a random sequence of numbers, such as 3-7-4-6-7-9, and, at the same time using a stylus to follow a randomly moving target on a computer screen. There is very little impact on the ability to recall the number sequence correctly or to track the moving target when they are performed together compared with when each task is performed on its own, and this can be accomplished without any particular expertise or training. Complementary studies have shown that placing a heavy demand on temporary memory places few, if any restrictions on retrieval of stored knowledge about past events or that is linked to expertise. These findings suggest that the human cognitive system may comprise several components, each specialized for particular kinds of tasks, which may operate in parallel even if our conscious experience is of only one component operating at a time. The findings also point to the idea that WM is best considered as being separate from, but interacting with, the system that stores previously acquired knowledge of the world and of past experiences.

The unitary view of memory also has difficulty explaining patterns of impairment and sparing in individuals who suffer from cognitive deficits as a result of brain damage. The cognitive impairments are often selective, for example, with very poor performance on immediate memory tasks (such as repeating back a random sequence of digits) coupled with normal performance on long-term retention (such as recalling events that occurred weeks, months, or years previously). The converse dissociation also occurs, namely normal memory for something that happened within the last few seconds contrasting with very poor retrieval of details about events that happened a few hours, days, or months previously, or difficulties in learning something new. Theories assuming that WM comprises temporary activation of relevant information from long-term memory would predict that if long-term memory access is impaired then temporary activation of long-term memory should also be impaired. Theories that assume WM to be a functionally separate part of the cognitive system can account more readily for the different patterns of cognitive impairment observed following brain damage, as well as for the data patterns from healthy adults, and therefore may be more useful.

The Case for Multiple Memory Systems

During the first half of the twentieth century, the prevailing view in North America was of human memory as primarily the result of observable learning of associations, and theories of learning dominated. During the same period in Europe, memory as a cognitive concept was more prevalent, for example in the work of Bartlett, Broadbent and Zangwill. By the mid-1950s, the latter started to become the more generally agreed view, and the spread of a cognitive approach to memory on both sides of the Atlantic gave rise to major new avenues of research. Very quickly it became clear that memory was best thought of as comprising at least two elements, with one focusing on temporary or short-term retention and the other on long-term retention. This reflected a dichotomy proposed originally in the seventeenth century by the British philosopher John Locke who suggested that the 'faculty of retention' comprised 'contemplation' or 'keeping an idea in mind,' and that this was separate from 'memory' or the 'storehouse of ideas.' This distinction appeared again in the late nineteenth century when the North American philosopher and psychologist William James referred to primary memory as holding information in mind which he contrasted with retention as storing records of life experiences and acquired knowledge.

During the 1960s, the contrast between primary memory and what then became known as secondary memory was demonstrated experimentally. For example, when healthy adult volunteers are asked to retain sequences of words for immediate recall in any order, omissions tend to occur among words from the middle part of the list and much better recall is observed for items close to the beginning or close to the end of each list. If instead, the volunteers are asked to recall the items following a delay, performance on the last few items is no better than it is for items in the middle of the list. However,

initial items in the list retain their advantage despite the delay. These contrasting patterns were thought to support the idea that two memory systems are reflected in recall of different parts of the list: the early list items are thought to be rehearsed and transferred into a longer term memory system, or secondary memory. As the list progresses, rehearsal becomes more difficult and therefore few items are transferred to a long-term store. The items in the final segment of the list are presumably held in a temporary memory system, or primary memory, whose contents are lost unless recall is immediate. The advantage for the items in the early segment of the list is known as the 'primacy effect,' and the advantage for the final items with immediate recall is known as the 'recency effect.'

Further support for the dichotomy comes from people who, following brain damage, suffer from specific types of memory impairment. One type of such patients is characterized by an inability to learn new material, a condition referred to as anterograde amnesia. The damage usually involves an area of the brain known as the hippocampus. When such people are tested for immediate free recall of word lists, their pattern of recall lacks the advantage for the initial items in the list (i.e., no primacy effect) but they show a normal recency effect. This suggests damage to the memory system responsible for retention of the early items but normal functioning of the memory system responsible for the recency effect. These people can also keep track of normal conversations over a few minutes, but just an hour later will have forgotten that the conversation took place. So, people with anterograde amnesia seem to have a primary memory that functions normally, but a longer-term memory that is malfunctioning. Another type of memory impairment shows the opposite effect in which there is great difficulty remembering and repeating back sequences of letters, numbers and words, but completely intact use of longer-term memory. The fact that ability to remember verbal material over short periods of time can be impaired independently from the ability to undertake new learning, suggests that these reflect two different kinds of memory systems in the brain. That is, primary memory is not simply activated long-term memory.

Developing the Concept of WM

The idea of primary and secondary memory provided a useful framework to account for the accumulating body of experimental data in support of a two-component memory system. However the traditional study of primary memory was limited by the fact that virtually all of the research relied on experiments with presentation and recall of words, letters, or digits. It is clear that we can also retain information about the appearance of objects and scenes, shapes, colors, and sequences of movements, as well as names and verbal descriptions. Therefore, from the late 1960s, primary memory began to be seen as a system that could support temporary retention of different kinds of material as well as offering a means to process and manipulate that material. The dominant or 'modal' concept then began to be identified as a shortterm, or WM processing and storage system. This 'modal model' retained the notion that information was transferred from the senses, via WM to secondary, or long-term memory. In this sense, WM was seen as a form of gateway, or temporary waiting space between input from the outside world and stored knowledge. It also incorporated the assumption that WM has a maximum capacity that could be used by storage, by processing, or by elements of both.

However a number of features of this model came under pressure from the 1970s onwards. The recency effect, once thought to be a major source of support for a temporary memory system, also was demonstrated in longer-term recall. For example when recalling visits to movies, one can recall where the car was parked, or attending rugby matches, individuals often can recall details of very recent events, even if these had occurred several days or weeks before, but have rather poorer recollection of similar experiences that happened previously. This demonstrated that recency was not unique to a temporary memory system. From this came the suggestion that memory loss functioned a little like observing telephone poles as they recede into the distance. Memory for the last visit to a restaurant a few days previously can be vivid and detailed, just as the closest telephone pole can be seen very clearly. Memory for the second last, or third last or fourth last restaurant visited becomes progressively less clear with increasing numbers of restaurant visits, just as it becomes more difficult to distinguish between different poles the more intervening poles are in view. This suggested that forgetting came from memories interfering with one another, and that recent events are remembered well because there are fewer other memories to cause interference. This would then apply to memory for events that happened a few seconds ago as well as memories for events that happened a few days or weeks previously. Also, there were demonstrations that semantic knowledge could be used to help support short-term retention. For example immediate free recall of real words is better than recall of nonsense words, suggesting that the stored knowledge about the real words (such as their meaning) can support primary memory, and so primary memory might simply comprise activated secondary memory.

There was, and still is, some debate as to whether research findings such as those just described might undermine the view that primary memory is separate from secondary memory. However, the separate memory systems view is still needed to explain the different kinds of memory problems found in people with contrasting forms of brain damage. So, patients with a dense amnesia typically show a normal short-term recency effect, but do not show long-term recency effects. One resolution to this debate is that better memory for recent events might be a property of any human memory system (short or longterm). So, showing recency effects over short and long periods does not mean that these effects reflect the operation of the same memory system. The evidence for short-term semantic coding showed that the memory systems interact, with any one memory task probably drawing on several different aspects of the cognitive system depending on task demands, rather than reflecting a single memory system.

Although the modal model survived the discovery of long-term recency and semantic information in immediate memory tasks, another core assumption of the model was challenged by the discovery that information from vision or hearing could reach long-term memory without being available in WM. Evidence for use of semantic codes in tasks thought to require WM suggested that information from the senses might first activate stored knowledge related to the sensory input before it reaches WM. So, this suggests that the flow of information is

the other way round from that suggested by the modal model, and activation of stored knowledge following sensory input acts as the route in to WM. Again, the insight came from studies of brain-damaged individuals: one famous case is that of patient 'PV' who, following a stroke in the sylvian region of the left hemisphere, presented with a very severe deficit in retaining sequences of words or digits for immediate recall, but had no difficulty in retrieving details of autobiographical events and proved well able to learn from new experiences. The pattern for PV and other similar patients reported subsequently suggests that access to long-term memory does not depend on normal function in short-term verbal retention, either for new learning or for recall of past events (reviews are given in Vallar and Shallice, 1990).

Additional evidence about the characteristics of WM comes from patients who suffer from a disorder known as unilateral spatial neglect. This arises most commonly from damage to the right parietal cortex of the brain, and such patients typically are unable to report details of scenes or objects to their left, despite having normal vision. A few patients show unilateral neglect only for information in their mental image or immediate memory for scenes. For example, one patient 'NL' (Beschin et al., 1997) could quite readily report details of everything in his immediate environment. However, if he was asked to close his eyes, he could only report from memory details that were on his right. Moreover, if asked to recall details of a familiar scene such as the town square in his home town he could recall only those details that were on his imagined right. This was not a problem of accessing his long-term memory, because when asked to imagine himself standing at the opposite end of the square, he could then report details that were now on his imagined right but had been omitted before from a different imagined viewpoint. This pattern demonstrates a clear dissociation between the processes of perception, which for NL were perfectly normal, the contents of his long-term memory, which again was normal, and the operation of his visual WM, which was dramatically impaired. This further suggests that WM does not act as a gateway between perception and long-term memory as assumed by the modal model, but instead deals with information that has been activated from long-term memory.

Studies of healthy adults have also demonstrated that the contents of WM are interpreted, for example in studies of memory for ambiguous figures. Figure 1 can be interpreted in one of two ways, either as a duck or as a rabbit, and if we continue to inspect the figure then our interpretation 'flips' between one and the other. However in some ingenious experiments, Chambers and Reisberg (1985 – see Cornoldi et al., 1996) showed this figure for only a brief time to a large number of volunteers who were asked to report from their memory what the figure depicted. All of the volunteers reported it as either a duck or a rabbit, but none of them could then reinterpret their image in memory and see the figure as anything other than their first interpretation. Nevertheless, they could do so when asked to draw their image on paper moments later. So, looking at the ambiguous figure allows a switch between seeing it as a rabbit or as a duck, but the representation of the figure in WM is linked closely with the initial interpretation of what we saw. In other words, the contents of WM are interpreted before they enter WM. This offers quite convincing evidence that perception feeds into long-term memory, which activates relevant knowledge for manipulation with WM (for discussions see Cornoldi et al., 1996; Logie, 1995, 2011).

Figure 1 Example of the duck-rabbit ambiguous figure.

A second assumption of the modal model was that of WM as a single, flexible, limited capacity system that could undertake both temporary storage and some processing of the material being stored. This theme remains in one contemporary, and influential strand of WM research that has focused on measures of individual differences in cognitive capacity. In a typical task, healthy adult volunteers are asked to read a series of sentences. The requirement is to recall the last word of each sentence after all the sentences have been presented. The number of sentences is gradually increased until the participant can no longer accurately recall all of the final words, and the maximum number of words that can be recalled is taken as a measure of the individual processing and storage capacity. The measure, in various forms, has been shown to correlate well with other measures of cognitive ability such as reading comprehension or mental arithmetic performance, and with general intelligence (e.g., Kane and Engle, 2002). These correlations are then used to argue that WM is a mental capacity for processing and storage that can be applied to a wide range of both laboratory and everyday cognitive tasks. In this sense, the sentence 'span' (or WM capacity) measure has been quite useful, and offers the basis for a standardized test of cognitive ability. Some researchers have even suggested that it might replace standard IQ tests.

Although the above approach has significant value as a measure of individual differences in mental capacity, the underlying theoretical assumptions are that there is a single flexible cognitive system supporting the memory performance as well as the processing of material, such as reading the intervening sentences. One possible alternative view is that the WM capacity task actually relies on several different aspects of cognition, and that this same range of cognitive functions is employed for other complex cognitive tasks such as reading comprehension. A further limitation is that the test provides only a measure of memory in the context of processing (retention of the final words while reading). Typically, there is no measure taken of performance on the processing component of the task. In a small number of tests of the assumption that the same system supports memory and ongoing processing, a separate measure has been taken of processing capacity by asking volunteers to verify each of the sentences for their semantic content without a memory requirement. Memory ability was then measured independently by asking volunteers to recall lists of individual words without any requirement to verify sentences. Participants were then asked to combine verification of each sentence with recall of all of the final words for the sentences, and the task demands were set at the maximum that had been determined for each task component separately. If WM were a single flexible, limited capacity system,

we might expect performance to be very poor when these two demanding task elements were combined. Results have shown that despite the high processing demand and high memory demand, performance levels on both elements of the task remain very close to those obtained when the memory task and the processing task are performed independently. Moreover, there are typically poor correlations between the processing task scores and the memory task scores, and a combined score that includes both processing and memory scores shows much higher correlations with other measures of mental ability than does the score for recall of the final sentence words alone.

These results point to the notion that temporary memory might be served by different aspects of the cognitive system from those used to support on-line processing. This argument in turn suggests that WM might be best thought of as the orchestrated operation of multiple cognitive instruments, each one with a specific role for temporary memory or for processing. For any given task, there is a selection of which instruments are needed. Some of these instruments might be needed to perform at the limits of their capacity. Other instruments might be needed to perform the task, but they make a contribution that is well within the limits of their capacity. So, looking at the differences between people in the maximum score that they can obtain on a complex cognitive task will reveal something about the overall mental ability of each individual, but will not reveal much about the cognitive instruments or components that are available, or how those components work together to maximize task performance. For more detailed discussion of this point see Logie (2011).

WM as a Multiple Component Workspace

The most influential model of a multicomponent WM was originally proposed in 1974 by two British scientists, Alan Baddeley and Graham Hitch. The model included some elements of the modal model by incorporating processing and memory, however these were viewed as distinct. Three components were originally proposed to provide respectively temporary verbal storage, temporary visuospatial storage and a coordinating or 'executive' function. The verbal storage component was originally named the articulatory loop, although subsequently it has been referred to as the phonological loop, and to consist of two components, a passive phonologically based temporary store and a mental verbal rehearsal process. The visuospatial component was originally referred to as the visuospatial scratch pad, but has subsequently been thought to comprise a specialist temporary memory store (or visual 'cache') for retaining the visual appearance of objects and scenes linked to a system (an 'inner scribe') for retaining, and possibly rehearsing, sequences of movements and more dynamic properties of scenes. The coordinating mechanism was referred to as the central executive, but in more recent years has been thought of as a collection of different executive functions that collectively can support switching between tasks, updating the contents of WM, inhibiting irrelevant information, the coordination of multiple task performance, the allocation of attention and the implementation of strategies and heuristics used in problem solving, and learning and retrieval.

A schematic representation of a recent version of the multiple component model is shown in Figure 2. The diagram illustrates the various specialist subcomponents in addition to the suggested flow of information from perception through a long-term knowledge base into WM. In the latter sense the contents of WM comprise some of the products of perception and of subsequent mental processing. WM can then be considered as a multicomponent workspace that stores and manipulates information, which has been activated from the long-term store, rather than as a primary route into long-term memory from the environment. Detailed discussion of the evidence for the features of the model can be found in Logie and van der Meulen (2009), Logie (2011), and Baddeley (2012).

This conception of WM fits neatly with the notion that people are strategic in how they use their cognitive functions. For example, in recalling a list of words, this can be accomplished by rehearsing the sounds of the words, thinking about their meaning, organizing them into groups, remembering what the words look like when printed, or imagining the objects or concepts to which the words refer. As a result, a phenomenon such as the phonological similarity effect (see above) with visually presented words only appears if the participants rely on mentally rehearsing the sounds of words

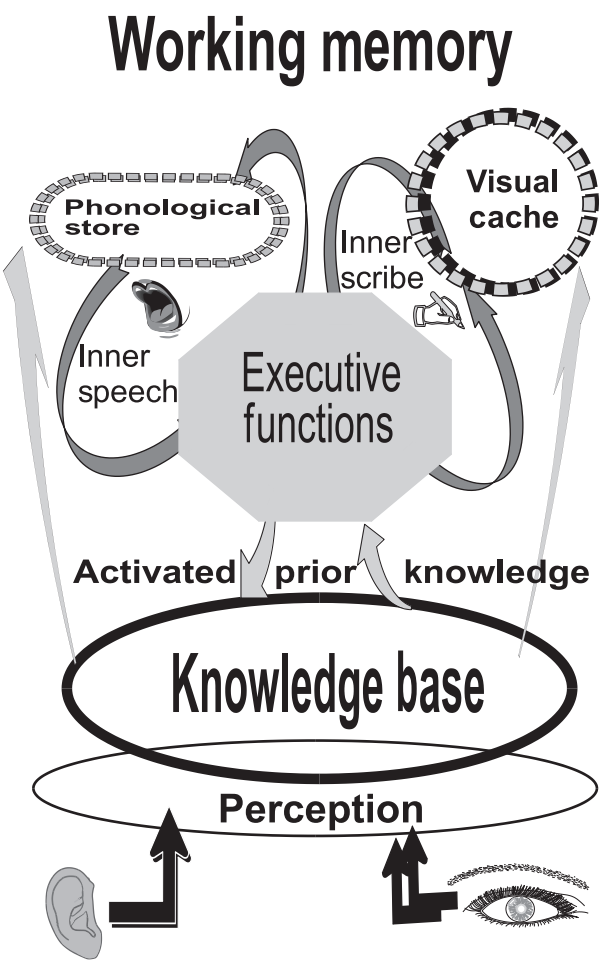

Figure 2 Schematic diagram of a recent version of the multiple component model of working memory.

prior to recall. Observing the phonological similarity effect can then indicate how the people are remembering the words. However, there is evidence that some participants in these experiments spontaneously use other strategies that might rely on visual or semantic codes, and for them the phonological similarity effect does not appear (e.g., Logie et al., 1996). The same argument holds for other well-established phenomena such as the recency effect. Different people might perform the same task in different ways. Moreover, as people get older, some aspects of WM remain relatively intact (primarily verbal memory ability), whereas other aspects of WM start declining in early adulthood and continue to decline throughout life (primarily visual memory ability). It also appears that older people might perform WM tasks in different ways than do younger people (see Johnson et al., 2010). The differential changes across age, and differential strategic use of different aspects of WM would not arise if the material did not activate knowledge stored in long-term memory prior to it being held in WM for the period until recall is required. Nor, would it arise if there is only a single, general purpose WM system. The evidence is easier to explain if we assume that WM comprises a range of specialist cognitive functions that can act together when performing a cognitive task.

Applying the Concept of WM

The idea of a multiple component WM has proved useful in understanding some of the cognitive deficits suffered by individuals with brain damage as well as accounting for important aspects of healthy cognition. Discussed above are some of the insights into cognitive deficits arising from the use of the model, and how studies of people with such deficits help the understanding of normal cognition. There have been several other areas of application, including the use of the phonological loop in explaining important aspects of vocabulary learning and of mental arithmetic. For example, one feature of the phonological loop is that it can retain a recently heard word, mentally rehearse the word, and then generate the word as articulate speech. A number of studies have examined the ability of young children to repeat aloud new or nonsense words that they have just heard. Their ability to do so correlates well with their general competence in language and predicts their language ability several years later. Patients with severe verbal short-term memory deficits also appear to have difficulty acquiring new foreign vocabulary. These findings have led to the suggestion that the phonological loop is a cognitive function that has evolved to support the acquisition of language, and that has the emergent property that it can support temporary verbal memory.

Among the acquired skills that appear to depend on the phonological loop are counting and mental arithmetic. Keeping track of a running total in counting, relies on subvocal rehearsal of the counting sequence, while keeping track of partial solutions in mental arithmetic again appears to rely on subvocalization of the subtotals involved, unless the solution is already well learned. For example, for most people the answer to $4 + 5 = ?$ would be available without the need for mental calculation, because it would be recalled as an arithmetic fact. However the answer to $23 + 19 + 48 + 34 = ?$ would require a series of operations with temporary memory for the subtotals (held in the phonological loop) as each part of the sum is completed. Evidence for these conclusions arises from studies in which volunteers perform counting or mental arithmetic while suppressing articulation through repeating an irrelevant word. Accuracy in counting and in arithmetic is disrupted by this procedure but is not disrupted by performing other irrelevant tasks such as tapping the table in a regular pattern or even when being presented with irrelevant speech sounds. The phonological loop therefore seems to support accuracy in counting and arithmetic.

The visuospatial resources of WM can support a wide range of other everyday cognitive tasks such as remembering where we are on the page while reading, maintaining and updating a representation of objects in our immediate environment, mental scanning to plan a route around a familiar town, or imagining the configuration of furniture in an empty room. The visual cache appears to be able to support memory for the visual appearance of letters and words, and to support temporary memory for combinations of colors and shapes, as well as visual patterns. We are not necessarily aware of the contents of this visual temporary memory, but we can retrieve information from it when it is needed. The visual cache does not seem to be directly linked with visual perception, and can store visual information that has been activated from hearing the name of an object. So, for example, hearing the words 'red circle,' 'blue cross,' 'green triangle,' would activate stored visual knowledge about those words, allowing a temporary visual memory to be created comprising those particular shape–color combinations. A temporary visual memory could likewise be created by seeing a red circle, a blue cross and a green triangle. The temporary visual memory system (the visual cache) that holds the information appears to be the same, regardless of whether the information was heard or seen. A second temporary representation could also be created based on the sounds of the words, and this could be held in the phonological loop as a sequence of word pairs. Again, it would not matter if the word sounds had been heard, or had been activated from stored knowledge when looking at the colored shapes. A third temporary memory for the same material could involve the creation of a meaningful integrated mental image that includes all three colored shapes in a recognizable object, for example in the form of a colorful ice cream cone. This would more directly involve activation of stored knowledge and additional cognitive functions that allow manipulation of the information to support creative thinking. When asking someone to remember which colored shapes they saw or heard, their memory could come from any or all of these three types of temporary memory representation. So, WM then can function as a set of different cognitive functions acting together when confronted by tasks in experiments and in our daily lives.

Conclusion

WM is an evolving concept, which may be subject to change in the light of new evidence or questions as to its detailed characteristics. It is broader than the concept of short-term memory in that it not only includes short-term retention of information, but also includes the selection, processing and manipulation of

that information. It encompasses separate short-term memory systems for retaining different kinds of information such as the sounds of words, the appearance of shapes or colors or their combination, and sequences of actions and movements. Crucially it allows understanding of a great deal of evidence and phenomena from laboratory tasks performed by healthy individuals, from studies of individuals with specific forms of brain damage, and from everyday experience.

See also: Amnesia: General; Hippocampus and Related Structures; Prospective Memory; Short Term Memories, Theories of; Short-Term Memory: Psychological and Neural Aspects; Spatial Navigation; Temporal Distinctiveness Models of Memory; Visual Imagery; Visual Memory, Psychology of; Working Memory, Neural Basis of.

Bibliography

Baddeley, A.D., 2007. Working Memory, Thought and Action. Psychology Press, Hove, UK.

Baddeley, A., 2012. Working memory, theories models and controversy. The Annual Review of Psychology 63, 12.1–12.29.

Beschin, N., Cocchini, G., Della Sala, S., Logie, R.H., 1997. What the eyes perceive, the brain ignores: a case of pure unilateral representational neglect. Cortex 33, 3–26.

Cornoldi, C., Logie, R.H., Brandimonte, M.A., Kaufmann, K., Reisberg, D., 1996. Stretching the Imagination: Representation and Transformation in Mental Imagery. Oxford University Press, New York.

Johnson, W., Logie, R.H., Brockmole, J.R., 2010. Working Memory Tasks Differ in Factor Structure Across Age Cohorts: Implications for Dedifferentiation. Intelligence 38, 513–528.

Kane, M.J., Engle, R.W., 2002. The role of prefrontal cortex in working memory capacity, executive attention, and general fluid intelligence: an individual differences perspective. Psychonomic Bulletin and Review 9, 637–671.

Logie, R.H., 1995. Visuo-spatial Working Memory. Lawrence Erlbaum Associates, Hove, UK.

Logie, R.H., 2011. The functional organisation and the capacity limits of working memory. Current Directions in Psychological Science 20 (4), 240–245.

Logie, R.H., Della Sala, S., Laiacona, M., Chalmers, P., Wynn, V., 1996. Group aggregates and individual reliability: the case of verbal short-term memory. Memory and Cognition 24 (3), 305–321.

Logie, R.H., van der Meulen, M., 2009. Fragmenting and integrating visuo-spatial working memory. In: Brockmole, J.R. (Ed.). The Visual World in MemoryPsychology Press, Hove, UK, pp. 1–32.

Osaka, N., Logie, R.H., D'Esposito, M. (Eds.), 2007. The Cognitive Neuroscience of Working Memory. Oxford University Press, Oxford, UK.

Pylyshyn, Z.W., 1973. What the mind's eye tells the mind's brain: a critique of mental imagery. Psychological Bulletin 80, 1–24.

Vallar, G., Shallice, T. (Eds.), 1990. Neuropsychological Impairments of Short-Term Memory. Cambridge University Press, Cambridge, UK.

Workplace Autonomy

Linda C Forbes, Ancell School of Business, Western Connecticut State University, Danbury, CT, USA
John M Jermier, College of Business & Patel College of Global Sustainability, University of South Florida, Tampa, FL, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Autonomy, the condition or quality of being self-governing or free from excessive external control, has been a vital concept in theory and research on human growth, development, well-being, and motivation. To demonstrate the widespread significance afforded the concept of workplace autonomy (and the crucial practical consequences associated with it), in this article, we review the historical foundations, significant models, and emerging streams of research in this area of study. This article leads us to conclude that both individual and collectivist autonomy are highly beneficial for individuals, organizations, communities, and societies.

For decades, autonomy has been a vital concept in theory and research on human growth, development, well-being, and motivation. Autonomy can be defined as the condition or quality of being self-governing or free from excessive external control. According to Immanuel Kant, the eighteenth century German philosopher often credited with laying the groundwork for modern philosophy, autonomy is important to human beings because it is the foundation of human dignity and the source of all morality (Hill, 1991). The dignity (poise and self-respect) of an adult is usually based on possessing some meaningful degree of control in making the decisions that deeply affect his or her life. Likewise, the source of a person's morality lies in the volitional self that rises above socially conditioned desires and partisan attachments in order to adopt impartial ethical principles and standards of conduct.

More recently, scholarly interest in the concept of autonomy has led to extensive research on workplace autonomy, the domain of life often seen as most essential to human accomplishment, enrichment, and fulfillment. To demonstrate the widespread significance afforded the concept of workplace autonomy (and the crucial practical consequences associated with it), in this article, we review the historical foundations, significant models, and emerging streams of research in this area of study.

Autonomy and the Individual

During the 1960s, amid massive social upheaval in many of the advanced industrial nations of the world, organizational and management theorists began to turn their attention to the sources of discontent in the workplace, manifesting in "accelerating absenteeism, job turnover, shoddy workmanship, sabotage, and the greatest outbreak of wildcat strikes since the 1930s" (Greenberg, 1975: p. 192). Several now classic texts called attention to the neo-Marxist thesis of alienation and the notion that lack of control over the product and process of work would lead to intense and widespread feelings of powerlessness, meaninglessness, social isolation, and self-estrangement (e.g., Blauner, 1964).

In one of the first systematic attempts to address the phenomenon of work alienation and freedom, Hackman and Oldham (1976) developed the Job Characteristics Model, a framework that focused attention on autonomy and four other key factors involved in designing enriched work. Work designed to be complex and challenging (characterized by high levels of autonomy, skill variety, identity, significance, and feedback) was theorized to promote high intrinsic motivation, job satisfaction, and overall work performance. Of these characteristics, autonomy is arguably the most studied. Decades of research on this widely cited model have shown that job scope or complexity, an additive combination of autonomy and the four other job characteristics, (1) is correlated significantly with more objective ratings of job characteristics, (2) may be reduced to a primary factor consisting of autonomy and skill variety, and (3) has substantial effects on affective and behavioral reactions to work, mostly indirectly through critical psychological states such as experienced responsibility for the outcomes of the work. There is some evidence suggesting that the positive effects of complex and challenging work have thresholds, but the main lesson we can draw up from this line of research is that the experience of autonomy at work has positive consequences ranging from higher job performance to enhanced general well-being. Studies continue to support the model across nation-states, industries, jobs, and teams (see DeVaro et al., 2007).

Another influential model and line of ongoing research in which autonomy at work is central was launched by Karasek's (1979) study of national survey data from Sweden and the USA. In this research, he proposed the Job Demands–Job Control Model that included a thesis about the causes of physical and mental strain and health problems. Job demands include requirements for working fast and hard, having too much to do and too little time to do it, conflicts among demands, and personal conflicts with one's role. Job control is specified in terms of decision latitude and relates to the theme mentioned earlier about possessing some meaningful degree of control over decisions that deeply affect one's life. Karasek (1979) proposed two dimensions to job control (authority to make decisions on how to work and having say over what happens; and latitude to use a variety of skills on the job). These dimensions parallel Hackman and Oldham's (1976) two central concepts of autonomy and skill variety. The main proposition in Karasek's original model was that when jobs are

simultaneously high in demands and low in decision latitude, physical and mental health problems are more likely to occur – the 'high strain job' (Karasek, 1979: p. 288). Demanding jobs that permit autonomy through high decision latitude are challenging and create opportunities for arousal, learning, and healthful response, thus mitigating the stressful effects of demands and generating the potential for job satisfaction and life fulfillment.

As with the Job Characteristics Model, research on Karasek's (1979) framework has underscored the importance of autonomy at work while also revealing some of the intricacies involved in explaining the specific processes through which autonomy exerts its positive effects. Research on Karasek's original Job Demands–Job Control Model has been mixed. Many studies reported that demands have negative effects regardless of the degree of job control, but in other studies, the predicted interaction effect has been found. In response, researchers began to develop the model further by examining the role of individual differences moderating the importance of decision latitude. For example, the role of self-efficacy and the role of proactive personality have been explored. It was found that high strain was mitigated when decision latitude was exercised by more self-confident and proactive individuals.

Karasek and others have developed an enhanced model that includes job demands, control, and support. Recognizing the long history of research demonstrating the importance of social support and its effect on well-being and productivity (e.g., from coworkers and supervisors), the enhanced model included social support as a third dimension (Karasek and Theorell, 1990). In a recent meta-analysis involving 106 studies, Luchman and Gonzalez-Morales (2013: p. 45) found that all three dimensions (demand, control, and social support) contributed (each in their own way) to the experience of work, thereby supporting the model and its dimensions. Their findings support the conclusion that the highest levels of well-being in work environments are characterized by high control (autonomy), low demands, and supportive supervisors and coworkers. Research on the model of job control, demands and support underscores the importance of autonomy and other characteristics of jobs as essential to understanding the quality of work experiences.

A prominent approach to theorizing perceptions of autonomy, the internal–external locus of control model of human personality, proposed that individuals perceive that their actions can affect the events in their lives, but vary in the degree to which they believe forces outside their control typically influence the events (Langer, 1977). Some people (externals) have a relatively consistent tendency to perceive that the outcomes in their lives are governed by external forces, such as powerful others and social conditions, fate, luck, or chance. Others (internals) have a relatively consistent tendency to believe they control their own destiny; they perceive themselves as masters of their own fate and as in control of the actions and events that affect their lives. This model has been researched extensively, leading to conclusions that people who do not see themselves as self-governing (externals) are handicapped in numerous ways (e.g., less effective performance after controlling for cognitive ability and conscientiousness – Hattrup et al., 2005; greater psychological strain from workplace conflict, Dijkstra et al., 2011). Recent research, however, suggests that a more nuanced cultural approach to understanding locus of control may be warranted. In a meta-analysis integrating 40 years of locus of control research, Cheng et al. (2013) found evidence that suggests that external locus of control does not have the same negative connotations in collectivist societies that it does in individualist societies. They propose that in collectivist cultures such as Mainland China, "a sense of control may not be confined to control originated from oneself but may also encompass control derived from one's social network members" (p. 176).

Among the models of human growth and development that are centered on autonomy, the most theoretically sophisticated approach has been developed around the basic psychological needs of autonomy, competence and relatedness, and the concepts of self-regulation and intrinsic motivation (Ryan and Deci, 2000; Niemiec et al., 2010). Self-determination theory (SDT) proposes "higher behavioral effectiveness, greater volitional persistence, enhanced subjective well-being, and better assimilation of the individual within his or her social group" result when individuals act from motivations that emanate from the inner self (intrinsic motivation) rather than from sources of external regulation (Ryan and Deci, 2000: pp. 72–73). Although not without controversy (due to the entrenched belief that tangible rewards primarily govern behavior), research in a wide array of social conditions has demonstrated that individuals develop a perceived external locus of causation (i.e., a sense of diminished autonomy) when tangible rewards are administered contingently, based on evaluated performance at a task. For self-determination theorists, it is the experience of an external locus of causation (or the belief that one's actions are controlled by external forces) that undermines the most powerful source of natural motivation and that (when chronic) also can lead to stultification, weak self esteem, anxiety and depression, and alienation. Thus, health and well-being as well as effective performance in social settings are closely related to the experience of autonomy. Interestingly, Ryan and Deci (2000: p. 74) point out "autonomy refers not to being independent, detached, or selfish, but rather to the feeling of volition that can accompany any act, whether dependent or independent, collectivist or individualist."

A related framework for understanding workplace autonomy may be found in the widely circulated literature dealing with psychological empowerment. This framework builds on job characteristics research, SDT, and models of participation in decision-making and power equalization in the workplace. It specifies cognitions of perceived control and experienced psychological states that are inherent to; autonomous work. In a foundational article, Spreitzer (1995) defined psychological empowerment as a process manifesting in four primary cognitive states: (1) meaning, (2) competence, (3) self-determination, and (4) impact. In combination, these key cognitive states lead to high intrinsic motivation, work engagement, creativity and innovative behavior, organizational citizenship behavior, and quality work performance as well as personal benefits for employees such as reduced psychological strain (Seibert et al., 2011). Psychological empowerment research runs parallel with concepts of empowering leadership (e.g., Zhang and Bartol, 2010) and these concepts may be traced back to research conducted decades earlier that focused

on the beneficial effects of democratic, participative, and even laissez-faire leadership. What is especially noteworthy about research on psychological empowerment is that it is a highly integrative approach to theorizing, drawing together elements of workplace autonomy that have been conclusively linked to important organizational and personal outcomes.

Autonomy and Groups and Teams

The theme of workplace autonomy also runs through over six decades of research on groups and teams in organizations. Of primary concern here are the streams referred to by Moldaschl and Weber (1998) as the sociotechnical systems theory approach (STS) and the lean management approach. Both approaches focus on teams that are, to varying degrees, autonomous or self-organizing. Self-organizing teams have been theorized as having processes of self-management and collaborative teamwork. In self-organizing teams, members experience relatively high autonomy by collectively controlling, for example, the pace of work, task allocation, performance evaluation, recruitment and selection of new members, training, and even reward distributions. As noted by several researchers, because groups can undertake much larger pieces of work than individuals can, the potential to experience positive outcomes through heightened control is greatly enhanced.

Beginning with Trist and Bamforth's (1951) classic study of autonomy among long wall coal miners, through to recent reviews and meta-analyses of research in the STS tradition, positive results have been reported. Despite 'substantial variance' in research findings regarding the consequences of autonomous work groups (Guzzo and Dickson, 1996: p. 326), positive impacts on motivation and productivity (especially performance quality and cost savings), attitudes and perceived quality of work life, and (to a lesser extent) attendance and retention are routinely reported. Failures of some STS projects and the fact that STS interventions in organizations have not been used as widely as initially anticipated have been attributed to poorly designed (imbalanced) programs and internal political conflicts.

The lean management approach, used initially by Japanese manufacturers such as Toyota, differs from the STS approach primarily in the degree of autonomy experienced by team members. The lean management approach tends to be more supervisor-centered as team leaders are appointed rather than elected, decision-making is consultative rather than consensual, and the domain of decision-making authority for the team is more limited. Ilgen et al. (1993) noted that when team-based structures are created in organizations, there is a tendency to retain an overarching hierarchical structure and to require teams to report to leader-managers in a very formal and traditional way. In situations where teams are teams mostly in name, autonomy is very limited, leading to negative results. When interventions empower teams to become more self-directed and autonomous, results tend to be positive across a number of indicators, including financial performance (cf Latham and Pinder, 2005).

While the body of research on teams is rapidly mounting, with much of it supporting the benefits of autonomous or semiautonomous teams, scholars continue to call for a better understanding of the causal mechanisms that produce beneficial results and for complex models examining micro and macro contingencies (Delarue et al., 2008). They continue to explore new areas of research, such as the relationship of individual autonomy to team autonomy, responses to conflict in teams, team dynamics in service delivery firms, intergroup coordination across autonomous work teams, and shared leadership in teams.

Many scholars are quick to point out that work is often done by groups that are merely labeled 'teams' by managers but that lack the defining qualities of teams (see Procter and Burridge, 2008). Others point out that teams are not a panacea for producing high organizational performance or team members' well-being (e.g., D'Innocenzo et al., in press). Moreover, despite the positive results often attributed to autonomous and self-organizing teams, one cannot assume that the approach frees members from hierarchical control that automatically creates emancipated practice. Some writers contend that while workers in the typical self-managed team situation might feel a measure of influence, upon closer inspection, it is only an illusion of self-control. This theme is developed further by Ezzamel and Willmott (1998) and others who show that teams, through peer pressure and behavioral regulation, can intensify work and coerce their members in ways that are perhaps even more tyrannical than are the forms of control exercised by managers.

Autonomy and Organizational Democracy

The final stream of research on workplace autonomy that we will consider is best labeled organizational democracy or workplace democracy and is more familiar to researchers in Europe than in North America (see the academic journal *Economic and Industrial Democracy* (1980) – current volumes, for a review of this approach). The case for large-scale, organizational democratization is frequently linked to arguments asserting various benefits of autonomy realized through power equalization and participation in decision-making. Participatory structures are advanced as an antecedent to human psychological and social development (through, for example, opportunities for self-expression, creativity, achievement, or voice in how one's work is done) and as a formula for increasing organizational efficiency, productivity, and competitiveness.

Organizational democracy continues to be a popular theme, especially in the international management literature. Contemporary debates raise a range of provocative questions about the most fundamental issues. For example, Kokkinidis (2012) raised serious questions about whether, in contrast to direct workplace democracies, representative forms only serve to legitimize hierarchical structures. Harnecker (2009) also pushed on the limits of existing workplace democracies by asserting that workplace democracies such as cooperatives are not sufficient for meeting the social needs of communities if they are not embedded in the democratic planning processes of their surrounding communities.

The tradition of organizational democracy is important as a beacon for genuine social and economic reform because it helps sharpen distinctions among the varieties of macro-factors operating in the world that shape self-regulating organizational

systems and their effectiveness. Consistent with this approach is the conclusion that self-directed, autonomous teams tend to be more effective in nations other than the USA. Dobbin and Boychuck (1999), for example, found that national employment systems shaped the degree to which autonomy was realized (relatively high in Nordic countries and relatively low in the USA, Canada, and Australia). Why should this be so? Studies of workplace autonomy from the standpoint of organizational democracy provide some answers to this question and open the field up in at least two important ways. First, they suggest how, for example, individualistic and collectivist nation-state cultures shape attitudes and other reactions toward team-based, democratic organizing, in effect constraining and prefiguring the effectiveness of any intervention aimed at enhancing workplace autonomy. This point also corresponds with Guzzo and Dickson's (1996) statement that the effects of team-based organizing are highly situationally dependent and extends it into the realm of nation-state contingencies. Second, they suggest a need to develop further programs of inquiry that go beyond the psychology of autonomy, programs that describe and explain the emergence of objectively different structures of control that are anchored in political–economic, legislative, and cultural factors (see Forbes and Jermier, 1995). This is important because so long as studies of workplace autonomy rely exclusively on micro thinking and psychological models, the politics of 'empowerment' in any given culture at any given time will probably be limited to regimes of illusory control.

The streams of research on workplace autonomy reviewed in this article all support the idea that being truly self-governing and free from excessive external control produce positive outcomes for individuals, groups, and work organizations. The more general thesis, that autonomy promotes human growth and development as well as positive health and well-being, is also supported. Autonomy sometimes is considered to be synonymous with individualism and egoistic willfulness but there is increasing evidence supporting the idea that collectivist autonomy is equally beneficial. The concept of workplace autonomy has received considerable attention from researchers in several traditions of study. While more research in each of the traditions is needed, a more pressing need is to begin integrating insights from all the traditions to develop a general theory of autonomy at work.

See also: Alienation, Sociology of; Alienation: Psychosociological Tradition; Autonomy, Philosophy of; Group Processes in Organizations; Group Processes, Social Psychology of; Job Analysis and Work Roles, Psychology of; Self-Efficacy; Self-Regulation in Adulthood; Work, Anthropology of; Work, Sociology of.

Bibliography

Blauner, R., 1964. Alienation and Freedom. University of Chicago Press, Chicago.

Cheng, C., Cheung, S., Chio, J.H., Chan, M.S., 2013. Cultural meaning of perceived control: a meta-analysis of locus of control and psychological symptoms across 18 cultural regions. Psychological Bulletin 139, 152–188.

Delarue, A., Van Hootegem, G.V., Procter, S., Burridge, M., 2008. Teamworking and organizational performance: a review of survey-based research. International Journal of Management Reviews 10, 127–148.

DeVaro, J., Li, R., Brookshire, D., 2007. Analyzing the job characteristics model: new support from a cross-section of establishments. International Journal of Human Resources Management 18, 986–1003.

Dijkstra, M.T.M., Beersma, B., Evers, A., 2011. Reducing conflict-related employee strain: the benefits of an internal locus of control and a problem-solving conflict management strategy. Work & Stress 25, 167–184.

D'Innocenzo, L.D., Mathieu, J.E., Kukenberger, M.R. A meta-analysis of different forms of shared leadership-team performance relations. Journal of Management 40, in press.

Dobbin, F., Boychuck, T., 1999. National employment systems and job autonomy: why job autonomy is high in the Nordic countries and low in the United States, Canada and Australia. Organization Studies 20, 257–291.

Ezzamel, M., Willmott, H., 1998. Accounting for teamwork: a critical study of group-based systems of organizational control. Administrative Science Quarterly 43, 358–396.

Forbes, L.C., Jermier, J.M., 1995. Industrial democracy in Europe: a review essay. Organization Studies 16, 1080–1086.

Greenberg, E.S., 1975. The consequences of worker participation: a clarification of the theoretical literature. Social Science Quarterly 56, 191–209.

Guzzo, R.A., Dickson, M.W., 1996. Teams in organizations: recent research on performance and effectiveness. Annual Review of Psychology 47, 307–338.

Hackman, J.R., Oldham, G.R., 1976. Motivation through the design of work: test of a theory. Organizational Behavior and Human Performance 16, 250–279.

Hattrup, K., O'Connell, M.S., Labrador, J.R., 2005. Incremental validity of locus of control after controlling for cognitive ability and conscientiousness. Journal of Business and Psychology 19, 461–470.

Harnecker, C.P., 2009. Workplace democracy and social consciousness: a study of Venezuelan cooperatives. Science & Society 73, 309–339.

Hill Jr., T.E., 1991. Autonomy and Self-Respect. Cambridge University Press, Cambridge, UK.

Ilgen, D.R., Hollenbeck, J.R., Sego, D.J., Major, D.A., 1993. Team research in the 1990s. In: Chemers, M.M., Ayman, R. (Eds.), Leadership Theory and Research. Academic Press, San Diego, CA, pp. 245–270.

Karasek Jr., R.A., 1979. Job demands, job decision latitude, and mental strain: implications for job redesign. Administrative Science Quarterly 24, 285–308.

Karasek, R.A., Theorell, T., 1990. Healthy Work: Stress, Productivity and the Reconstruction of Working Life. Basic Books, USA.

Kokkinidis, G., 2012. In search of workplace democracy. International Journal of Sociology and Social Policy 32, 233–256.

Langer, E.J., 1977. The psychology of chance. Journal for the Theory of Social Behavior 7, 185–208.

Latham, G.P., Pinder, C.C., 2005. Work motivation theory and research at the dawn of the twenty-first century. Annual Review of Psychology 56, 485–516.

Luchman, J.N., Gonzalez-Morales, M.G., 2013. Demands, control, and support: a meta-analytic review of work characteristics interrelationships. Journal of Occupational Health Psychology 18, 37–52.

Moldaschl, M., Weber, W.G., 1998. The 'three waves' of industrial group work: historical reflections on current research on group work. Human Relations 51, 347–388.

Niemiec, C.P., Ryan, R.M., Deci, L., 2010. Self-determination theory and the relation of autonomy to self-regulatory processes and personality development. In: Hoyle, R.H. (Ed.), Handbook of Personality and Self-regulation. Wiley-Blackwell, West Sussex, UK, pp. 169–191.

Procter, S., Burridge, M., 2008. Teamworking and performance: the extent and intensity of teamworking in the 1998 UK Workplace Employee Relations Survey (WERS98). International Journal of Human Resources Management 19, 153–168.

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 68–78.

Seibert, S.E., Wang, G., Courtright, S.H., 2011. Antecedents and consequences of psychological and team empowerment in organizations: a meta-analytic review. Journal of Applied Psychology 96, 981–1003.

Spreitzer, G.M., 1995. Individual empowerment in the workplace: dimensions, measurement, validation. Academy of Management Journal 38, 1442–1465.

Trist, E., Bamforth, K., 1951. Some social and psychological consequences of the long wall method of coal-getting. Human Relations 4, 3–38.

Zhang, X., Bartol, K., 2010. Linking empowering leadership and employee creativity: the influence of psychological empowerment, intrinsic motivation, and creative process engagement. Academy of Management Journal 53, 107–128.

Workplace Deviance

Rebecca Bennett, Louisiana Tech University, Ruston, LA, USA
Shelly Marasi, Eastern New Mexico University, Portales, NM, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Workplace deviance is a serious and costly problem for organizations. This encyclopedia article outlines the growing body of scholarly research investigating these counterproductive workplace behaviors. Researchers have identified two main dimensions underlying deviant workplace behaviors. Numerous antecedents increase the likelihood and affect the target of workplace deviance. Multiple techniques deter, reduce, and/or prevent employees from participating in workplace deviance. However, these methods may come at great cost, whether financial or motivational. On the other hand, workplace deviance may sometimes be functional, for either employees and/or the organization. Proactive strategies are suggested.

For centuries, workplace deviance has been a serious issue for nearly all organizations (Case, 2000; Coffin, 2003) with prior research suggesting that the majority of employees (between 50 and 75%) have engaged in some type of deviant behavior (Bennett and Robinson, 2000; Harper, 1990; Harris and Ogbonna, 2002; Slora, 1991). The costs associated with workplace deviance have been estimated to be rather significant, with annual estimations being in the millions for US organizations (Case, 2000; Harris and Ogbonna, 2006). Workplace deviance is defined as "voluntary behavior that violates significant organizational norms and in so doing threatens the well-being of an organization, its members, or both" (Robinson and Bennett, 1995: p.556). Behaviors fitting this definition have also been referred to as counterproductive work behaviors (Fox and Spector, 1999), antisocial work behaviors (Giacalone and Greenberg, 1997), workplace aggression (Baron and Neuman, 1996; Folger and Baron, 1996), and organizational misbehavior (Vardi and Wiener, 1996).

Typology

The typology of workplace deviance developed by Robinson and Bennett (1995) used multidimensional scaling techniques to identify the underlying vast array of deviant workplace behaviors: (1) whether the potential harm was serious or minor and (2) the entity that was the target of the deviant act (employees or other individuals associated with the organization, such as customers and suppliers, which is referred to as *interpersonal deviance*, or the organization itself, which is referred to as *organizational deviance*). The overlap of the two orthogonal dimensions resulted in four different forms of workplace deviance. *Political deviance* involves minor deviant acts directed toward individuals at work, such as gossiping about coworkers; whereas, *personal aggression* entails serious deviant acts targeted toward individuals at work, such as sexual harassment. *Production deviance* refers to minor deviant acts aimed at the organization, such as leaving work early or arriving to work late; whereas, *property deviance* involves serious deviant acts targeted toward the organization, such as sabotaging equipment (Robinson and Bennett, 1995). Refer to Figure 1 for a complete overview of the typology, including examples of the four different types of workplace deviance.

Assessment

One of the earliest and most employed assessments for workplace deviance is Bennett and Robinson's (2000) Workplace Deviance scale, which consists of 19 items. This scale allows researchers to independently assess interpersonal deviance (intentional, potentially harmful behaviors that violate significant organizational norms and are directed at other individuals at work, which combines political deviance and personal aggression) and organizational deviance (intentional, potentially harmful behaviors that violate significant organizational

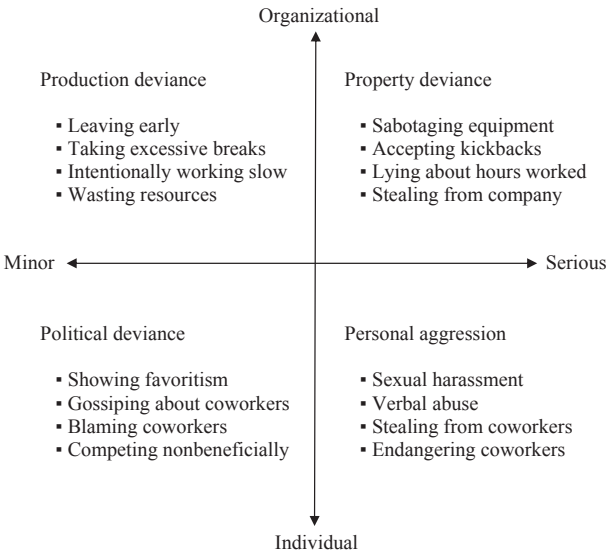

Figure 1 Typology of workplace deviance. Taken from Robinson, S.L., Bennett, R.J., 1995. The typology of deviant workplace behaviors: a multidimensional scaling study. Academy of Management Journal 38, 555–572.

norms and are directed at the organization itself, which combines property deviance and production deviance). For an overview of the Workplace Deviance scale refer to **Table 1**.

Another workplace deviance measure is the Counterproductive Work Behavior Checklist (CWB-C) that was developed by Paul Spector and his colleagues (Spector et al., 2006), which consists of 45 items which can be scored using two different formats. The first format replicates Bennett and Robinson's (2000) the two-factor structure of interpersonal deviance (referred to as CWB-P) and organizational deviance (referred to as CWB-O). The second format categorizes the 45 items into five dimensions: abuse, employee theft, sabotage, withdrawal, and production deviance. Additionally, Spector et al. (2006) constructed a shorter version of the original 45-item CWB-C measure. The shorter version of the CWB-C scale consisted of 32 items and had the ability to be scored using the same two different formats as the original 45 items scale.

Stewart and colleagues (Stewart et al., 2009) transformed Bennett and Robinson's (2000) Workplace Deviance scale into a nonself report measure of workplace deviance, which consists of 15 items and categorizes the items into interpersonal deviance (referred to as personal aggression) and organizational deviance, which can be further divided into production deviance and property deviance.

Meta-analyses of workplace deviance have confirmed that although interpersonal deviance and organizational deviance are highly correlated (Berry et al., 2007; Dalal, 2005), their differential relationships with important antecedents (e.g., personality traits) suggest that they are separate constructs and should be measured as such, especially if the research question calls for such a distinction. Berry et al.'s (2007) meta-analysis also suggested some differential effects in studies depending on which measure of workplace deviance was used (e.g., interpersonal or organizational deviance measures).

Table 1 Workplace deviance scale

Interpersonal deviance ($\alpha = 0.78$)
Made fun of someone at work.
Said something hurtful to someone at work.
Made an ethnic, religious, or racial remark at work.
Cursed at someone at work.
Played a mean prank on someone at work.
Acted rudely toward someone at work.
Publicly embarrassed someone at work.
Organizational deviance ($\alpha = 0.81$)
Taken property from work without permission.
Spent too much time fantasizing or daydreaming instead of working.
Falsified a receipt to get reimbursed for more money than you spent on business expenses.
Taken an additional or longer break than is acceptable at your workplace.
Come in late to work without permission.
Littered your work environment.
Neglected to follow your boss's instructions.
Intentionally worked slower than you could have worked.
Discussed confidential company information with an unauthorized person.
Used an illegal drug or consumed alcohol on the job.
Put little effort into your work.
Dragged out work in order to get overtime.

Reproduced from: Bennett, R.J., Robinson, S.L., 2000. Development of a measure of workplace deviance. Journal of Applied Psychology 85, 349–360.

Although these measures have been validated as measuring general workplace deviance, some studies have modified the measures to better represent a specific workplace deviance context for their specific research hypotheses. For instance, Darrat et al., 2010 created modified scales of organizational, interpersonal, and customer-directed deviance for outside sales' employees, Mitchell and Ambrose (2007) developed a measure for supervisor-directed deviance, and Posey et al. (2011) created deviance scales specifically for measuring computer abuse. Of course a plethora of measures for other individual deviant behaviors, such as theft, sabotage, and workplace incivility, already exist. The type of question being asked should predicate which measure is used.

History

Scholarly interest in workplace deviance has analytical roots dating back to the Industrial Revolution (Klotz and Buckley, 2013). Interpersonal deviance was first widely publicized during the eighteenth century when abusive supervision and harassment of employees was prominent in the workplace. Organizational deviance can be traced to the Luddite Rebellion during the nineteenth century, where employees destroyed machinery that replaced certain job functions.

Nonetheless, in spite of awareness and interest in employee misbehavior, and prior to the development of the broad bidimensional workplace deviance measure (Bennett and Robinson, 2000), empirical research on employee misbehavior involved analyzing only one or two specific counterproductive, antisocial, or deviant behaviors, such as employee theft (Greenberg, 1990), sexual harassment (Gutek, 1985), sabotage (Giacalone, 1990), absenteeism (Goodman and Atkins, 1984), withdrawal (Gupta and Jenkins, 1980), withholding effort or social loafing (Kidwell and Bennett, 1993; Latane et al., 1979), and soldiering (Taylor, 1903). Of course there is still a place for the study of individual deviant behaviors, such as with deviant behaviors involving national security information leaks, and consequently research continues to be done on individual antisocial behaviors. However, the broader focus on multiple behaviors that can be defined as 'harming the organization' allows researchers to investigate the common causes of such diverse organizational deviance behaviors as employee theft (Case, 2000), time banditry (Ketchen et al., 2008; Martin et al., 2010), and cyberloafing (Blau et al., 2006; Chidambaram and Tung, 2005; Lim, 2002). Considering these behaviors to have an underlying root cause provides theoretical integration as well as more efficient and productive research findings than studying individual deviant behaviors independently.

Potential Antecedents

Past research has investigated numerous situational, personal, environmental, and other factors in hopes of determining which strongly influence or predict workplace deviance. The main focus of workplace deviance antecedents appears to be situational or personal factors, with some studies integrating these two factors in conceptual models (e.g., Colbert et al.,

2004; Henle, 2005; Judge et al., 2006; Kantur, 2010; Skarlicki et al., 1999).

Situation-based Perspective

The situation-based perspective views workplace deviance as an outcome of organizational circumstances, actions, and/or treatment. For instance, abusive supervision (Tepper et al., 2001), boredom (Spector et al., 2006), group norms (Siehl, 1987), job (dis)satisfaction (Mangione and Quinn, 1975), and injustice perceptions (e.g., Colbert et al., 2004; Greenberg, 1990; Skarlicki and Folger, 1997), specifically interpersonal injustice (e.g., Ambrose et al., 2002; Colquitt et al., 2001; Flaherty and Moss, 2007; Henle, 2005; Judge et al., 2006) have all been identified as significant situational antecedents of workplace deviance.

Person-based Perspective

The person-based perspective infers an explanation for a person (or employee's) behavior from their personality, characteristics, or emotions rather than from situational or environmental factors. For instance, stress (Fox et al., 2001), negative emotions (Lee and Allen, 2002; Spector and Fox, 2002), negative affectivity (Richards and Schat, 2011), cognitive ability (Dilchert et al., 2007), low moral standards (Merriam, 1977), narcissism (Penney and Spector, 2002), age (Berry et al., 2007; Ng and Feldman, 2008), gender (Hershcovis et al., 2007), conscientiousness, and agreeableness (Mount et al., 2006; Salgado, 2002) have all been found to significantly impact workplace deviance.

Managing Workplace Deviance

There are several techniques managers and organizations can utilize to deter, reduce, or prevent employees from engaging in workplace deviance. The first thing organizations should do to help manage workplace deviance is to require the human resource department or hiring managers to screen applicants by having them complete a personality, cognitive ability, and/or integrity test since lower levels of cognitive ability (Dilchert et al., 2007), integrity (Lucas and Friedrich, 2005), conscientiousness, and agreeableness (Mount et al., 2006; Salgado, 2002) have been associated with workplace deviance. Another good screening method is to have the applicant sign a release form allowing the organization to contact the applicants' former employer(s) and professional reference(s) to identify past work ethic, attitudes, and antisocial behaviors in the workplace.

After the screening process, organizations can continue to manage workplace deviance by implementing different policies, strategies, and courses of actions. For instance, organizations can ensure employees are not wasting valuable work time and effort by monitoring employees electronically, such as by screening their e-mails, histories of internet browsing, and individual key strokes (Klotz and Buckley, 2013). Organizations can also attempt to decrease or abolish workplace deviance by engaging in different strategies, such as enhancing employee and supervisor training, employer–employee relations, communications, and perceptions of organizational justice. Additionally, organizations can adopt policies which discourage or prohibit employees from participating in certain deviant behaviors, such as writing unflattering or damaging information on social media web sites, by stating they will terminate employment for policy violators. However, in order for the organization or manager to be considered fair (and to enhance organizational justice perceptions), the consequences should match the seriousness of the deviant behavior (Robinson and Bennett, 1995). For instance, serious deviant acts (such as, stealing from the organization or sexual harassment) should have similarly stringent consequences and they should be harsher than those given for minor deviant acts (such as, leaving work early or gossiping about coworkers). Punishments should also be delivered in private, consistently, and according to the established organizational process (Bennett, 1998).

Although organizations may desire to reduce or eliminate workplace deviance, these techniques come with several costs. For instance, the financial costs associated with having job applicants complete various screening tests (e.g., personality and integrity tests) are notable, especially if there are many applicants to screen. Therefore, organizations need to identify whether the specific technique is cost efficient and will significantly reduce workplace deviance such that the costs associated with implementing the technique are less than the costs associated with workplace deviance (Levy and Tziner, 2011). Otherwise, when the costs associated with workplace deviance are less than the costs associated with the technique, the organization should choose a different technique that produces a better cost–benefit analysis. Additionally, some techniques (e.g., electronically monitoring employees) may lower employee morale, loyalty, organizational trust and commitment, while increasing aggression, job dissatisfaction, and other undesirable attitudes or behaviors, which may potentially increase workplace deviance.

Recommendations

Organizations should be wary of the techniques they use to manage workplace deviance, especially since most employees believe they are 'entitled' to privacy (e.g., to not being monitored) or have the right to behave deviantly to some extent (Klotz and Buckley, 2013). Organizations and their managers should be advised that the costs associated with some techniques (e.g., cameras and other monitoring equipment) may actually outweigh the costs associated with workplace deviance (e.g., a few lost moments due to an employee surfing an unauthorized web site), not to mention the impact the technique may have on employee attitudes and behaviors. For instance, employee monitoring has been shown to occasionally backfire and to result in increased levels of workplace deviance because of employees' reactance to the perceived privacy invasion and the sense of being controlled (Posey et al., 2011). Additionally, when punishments for engaging in deviant behaviors are imposed, organizations and managers should put forth effort to ensure they are fair and match the severity of the deviant act. It is probably better to be proactive in creating an organizational culture which offers employees respect, fair processes, just pay, meaningful work, autonomy, and involvement, factors which are related to reduced levels of

counterproductive workplace behavior, rather than to be reactive in responding punitively to observed employee aggression.

Organizations should also be aware of the possibility that workplace deviance may have functional outcomes for employees and organizations (Robinson and Bennett, 1995). For instance, employees may be able to relieve stress or tension by engaging in certain deviant behaviors (Klotz and Buckley, 2013; Krischer et al., 2010; Robinson and Bennett, 1995), such as using work time to make personal phone calls to ensure a family member's well-being, thereby allowing them the ability to focus more on their job and thus, increase performance. Additional examples of 'functional' deviance may include employee gossip assisting in establishing workplace norms, rule-breaking leading to innovative processes and improved efficiency, and employee conflict or harassment (such as abusive supervision) causing functional turnover. Additionally, more noticeable deviant acts (e.g., theft or sabotage) may provide warning signs to organizations and managers that employees are likely to engage in additional and possibly more severe deviant acts and thereby giving the organization an opportunity to identify and implement different techniques to reduce or eliminate these deviant behaviors before they increase or escalate.

Conclusion

Workplace deviance is a serious and costly problem for organizations. There are numerous variables that increase the likelihood of employees engaging in deviant behaviors. Therefore, managers should be aware of the selection tools that might restrict those prone to deviant behaviors from joining the organization. Managers should also be aware of their own behaviors (e.g., monitoring or bullying) that may provoke employees to behave deviantly, especially since most employees already believe they are 'entitled' to participate in deviant behaviors to some extent. There are multiple techniques that organizations and managers can use to deter, reduce, or abolish workplace deviance. However, these methods may come at a price, whether it is monetary or at the expense of valuable employee attitudes, such as loyalty and trust. Additionally, even though organizations attempt to minimize or eliminate these deviant behaviors, some of them may prove useful and productive for employees and organizations. The best strategy is probably to develop an organizational culture based on fairness, mutual respect, employee involvement, and autonomy. All of these factors have been shown to be related to lower levels of workplace deviance.

See also: Organizational Justice; Personality and Values at Work; Supervision, Abusive.

Bibliography

Ambrose, M.L., Seabright, M.A., Schminke, M., 2002. Sabotage in the workplace: the role of organizational justice. Organizational Behavior and Human Decision Processes 89, 947–965.

Baron, R.A., Neuman, J.H., 1996. Workplace violence and workplace aggression: evidence on their relative frequency and potential causes. Aggressive Behavior 22 (3), 161–173.

Bennett, R.J., 1998. Taking the sting out of the whip: reactions to consistent punishment for unethical behavior. Journal of Experimental Psychology: Applied 4 (3), 248.

Bennett, R.J., Robinson, S.L., 2000. Development of a measure of workplace deviance. Journal of Applied Psychology 85, 349–360.

Berry, C.M., Ones, D.S., Sackett, P.R., 2007. Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. Journal of Applied Psychology 92, 410–424.

Blau, G.G., Yang, Y.Y., Ward-Cook, K., 2006. Testing a measure of cyberloafing. Journal of Allied Health 35, 9–17.

Case, J., 2000. Employee Theft: The Profit Killer. John Case & Associates, Del Mar, CA.

Chidambaram, L., Tung, L.L., 2005. Is out of sight, out of mind? an empirical study of social loafing in technology-supported groups. Information Systems 16, 149–169.

Coffin, B., 2003. Breaking the silence on white collar crime. Risk Management 50, 8.

Colbert, A.E., Mount, M.K., Harter, J.K., Witt, L.A., Barrick, M.R., 2004. Interactive effects of personality and perceptions of the work situation on workplace deviance. Journal of Applied Psychology 89, 599–609.

Colquitt, J.A., Conlon, D.E., Wesson, M.J., Porter, C.O., Ng, K.Y., 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology 86, 425–445.

Dalal, R.S., 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. Journal of Applied Psychology 90 (6), 1241.

Darrat, M., Amyx, D., Bennett, R., 2010. An investigation into the effects of work-family conflict and job satisfaction on salesperson deviance. Journal of Personal Selling and Sales Management 30, 239–252.

Dilchert, S., Ones, D.S., Davis, R.D., Rostow, C.D., 2007. Cognitive ability predicts objectively measured counterproductive work behaviors. Journal of Applied Psychology 92, 616–627.

Flaherty, S., Moss, S.A., 2007. The impact of personality and team context on the relationship between workplace injustice and counterproductive work behavior. Journal of Applied Social Psychology 37, 2549–2575.

Folger, R., Baron, R.A., 1996. Violence and hostility at work: a model of reactions to perceived injustice. In: Bulatao, E.Q., VandenBos, G.R. (Eds.), Violence on the Job: Identifying Risks and Developing Solutions, American Psychological Association, Washington, DC, pp. 51–85.

Fox, S., Spector, P.E., 1999. A model of work frustration–aggression. Journal of Organizational Behavior 20 (6), 915–931.

Fox, S., Spector, P.E., Miles, D., 2001. Counterproductive work behavior (CWB) in response to job stressors and organizational justice: some mediator and moderator tests for autonomy and emotions. Journal of Vocational Behavior 59, 291–309.

Giacalone, R., 1990. Employee sabotage: the enemy within. Supervisory Management 35, 6–7.

Giacalone, R.A., Greenberg, J.E., 1997. Antisocial Behavior in Organizations. Sage Publications, Inc.

Goodman, P., Atkins, R., 1984. Absenteeism: New Approaches to Understanding, Measuring and Managing Employee Absence. Jossey-Boss, San Francisco, CA.

Greenberg, J., 1990. Employee theft as a reaction to underpayment inequity: the hidden costs of pay cuts. Journal of Applied Psychology 75, 561–568.

Gupta, N., Jenkins, G., 1980. The Structure of Withdrawal: Relationships Among Estrangement, Tardiness, Absenteeism, and Turnover. National Technical Information Service, Springfield, VA.

Gutek, B.A., 1985. Sex in the Workplace. Jossey-Bass, San Francisco, CA.

Harper, D.C., 1990. Spotlight abuse – save profits. Industrial Distribution 79, 47–51.

Harris, L.C., Ogbonna, E., 2002. Exploring service sabotage: the antecedents, types, and consequences of frontline, deviant, antiservice behaviors. Journal of Service Research 4, 163–183.

Harris, L.C., Ogbonna, E., 2006. Service sabotage: a study of antecedents and consequences. Academy of Marketing Science 34, 543–559.

Henle, C.A., 2005. Predicting workplace deviance from the interaction between organizational justice and personality. Journal of Managerial Issues 17, 247–263.

Hershcovis, M.S., Turner, N., Barling, J., Arnold, K.A., Dupre, K.E., Inness, M., et al., 2007. Predicting workplace aggression: a meta-analysis. Journal of Applied Psychology 92, 228–238.

Judge, T.A., Scott, B.A., Ilies, R., 2006. Hostility, job attitudes, and workplace deviance: test of a multilevel model. Journal of Applied Psychology 91, 126–138.

Kantur, D., 2010. Emotional motives and attitudinal reflections of workplace deviant behavior. The Business Review, Cambridge 14, 70–77.

Ketchen, D.J., Craighead, C.W., Buckley, M.R., 2008. Time bandits: how they are created, why they are tolerated, and what can be done about them. Business Horizons 51, 141–149.

Kidwell, R.E., Bennett, N., 1993. Employee propensity to withhold effort: a conceptual model to intersect three avenues of research. Academy of Management Review 18, 429–456.

Klotz, A.C., Buckley, M.R., 2013. A historical perspective of counterproductive work behavior targeting the organization. Journal of Management History 19, 114–132.

Krischer, M.M., Penney, L.M., Hunter, E.M., 2010. Can counterproductive work behaviors be productive? CWB as emotion-focused coping. Journal of Occupational Health Psychology 15, 154–166.

Kulas, J.T., McInnerney, J.E., DeMuth, R.F., Jadwinski, V., 2007. Employee satisfaction and theft: testing climate perceptions as a mediator. The Journal of Psychology 141, 389–402.

Latane, B., Williams, K., Harkins, S., 1979. Many hands make light the work: the causes and consequences of social loafing. Journal of Personality and Social Psychology 37, 822–832.

Lee, K., Allen, N.J., 2002. Organizational citizenship behavior and workplace deviance: the role of affect and cognitions. Journal of Applied Psychology 87, 131–142.

Levy, T., Tziner, A., 2011. When destructive deviance in the workplace becomes a liability: a decisional behavioral model. Quality & Quantity: International Journal of Methodology 45, 233–239.

Lim, V.K., 2002. The IT way of loafing on the job: cyberloafing, neutralizing and organizational justice. Journal of Organizational Behavior 23, 675–694.

Lucas, G.M., Friedrich, J., 2005. Individual differences in workplace deviance and integrity as predictors of academic dishonesty. Ethics & Behavior 15, 15–35.

Mangione, T.W., Quinn, R.P., 1975. Job satisfaction, counterproductive behavior, and drug use at work. Journal of Applied Psychology 60, 114–116.

Martin, L.E., Brock, M.E., Buckley, M.R., Ketchen, D.J., 2010. Time banditry: examining the purloining of time in organizations. Human Resource Management Review 20, 26–34.

Merriam, D., 1977. Employee theft. Criminal Justice Abstracts 9, 380–386.

Mitchell, M.S., Ambrose, M.L., 2007. Abusive supervision and workplace deviance and the moderating effects of negative reciprocity beliefs. Journal of Applied Psychology 92 (4), 1159.

Mount, M., Ilies, R., Johnson, E., 2006. Relationship of personality traits and counterproductive work behaviors: the mediating effects of job satisfaction. Personnel Psychology 59, 591–622.

Murphy, K.R., 1993. Honesty in the Workplace. Brooks/Cole, Belmont, CA.

Ng, T.W., Feldman, D.C., 2008. The relationship of age to ten dimensions of job performance. Journal of Applied Psychology 93, 392–423.

Penney, L.M., Spector, P.E., 2002. Narcissism and counterproductive work behavior: do bigger egos mean bigger problems? International Journal of Selection and Assessment 10, 126–134.

Posey, C., Bennett, R., Roberts, T., Lowry, P.B., 2011. When computer monitoring backfires: invasion of privacy and organizational injustice as precursors to computer abuse. Journal of Information System Security 7, 24–47.

Richards, D.A., Schat, A.C., 2011. Attachment at (not to) work: applying attachment theory to explain individual behavior in organizations. Journal of Applied Psychology 96, 169–182.

Robinson, S.L., Bennett, R.J., 1995. The typology of deviant workplace behaviors: a multidimensional scaling study. Academy of Management Journal 38, 555–572.

Salgado, J.F., 2002. The big five personality dimensions and counterproductive behaviors. International Journal of Selection and Assessment 10, 117–125.

Siehl, E.W., 1987. Garment workers: perceptions of inequity and employee theft. British Journal of Criminology 27, 174–190.

Skarlicki, D.P., Folger, R., 1997. Retaliation in the workplace: the roles of distributive, procedural, and interactional justice. Journal of Applied Psychology 82, 434–443.

Skarlicki, D.P., Folger, R., Tesluk, P., 1999. Personality as a moderator in the relationship between fairness and retaliation. Academy of Management Journal 42, 100–108.

Slora, K.B., 1991. An empirical approach to determining employee deviance base rates. In: Jones, J. (Ed.), Preemployment Honesty Testing: Current Research and Future Directions. Qurorum Books, Westport, CT, pp. 21–38.

Spector, P.E., Fox, S., 2002. An emotion-centered model of voluntary work behavior: some parallels between counterproductive work behavior and organizational citizenship behavior. Human Resource Management Review 12, 269–292.

Spector, P.E., Fox, S., Penney, L.M., Bruursema, K., Goh, A., Kessler, S., 2006. The dimensionality of counterproductivity: are all counterproductive behaviors created equal? Journal of Vocational Behavior 68, 446–460.

Stewart, S.M., Bing, M.N., Davison, H.K., Woehr, D.J., McIntyre, M.D., 2009. In the eyes of the beholder: a non-self-report measure of workplace deviance. Journal of Applied Psychology 94, 207–215.

Taylor, F.W., 1903. Shop Management. Harper & Brothers, New York.

Tepper, B.J., 2000. Consequences of abusive supervision. Academy of Management Journal 43, 178–190.

Tepper, B.J., Duffy, M.K., Shaw, J.D., 2001. Personality moderators of the relationship between abusive supervision and subordinates' resistance. Journal of Applied Psychology 86, 974–983.

Vardi, Y., Wiener, Y., 1996. Misbehavior in organizations: a motivational framework. Organization Science 7 (2), 151–165.

Workplace Safety and Health

Simo Salminen, Finnish Institute of Occupational Health, Promotion of Occupational Safety, Helsinki, Finland

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Workplace safety and health refers here to all practical actions taken at the shop floor level to promote the safety and health of workers. Such actions should be theory driven because accident theories help to pinpoint dangerous factors in the company, and they offer directions for practical solutions. The focus of these actions is the behavior and attitudes of the workers subject to the highest risk of occupational accidents and diseases.

Definition of Workplace Safety and Health

Workplace safety and health refers here to all practical actions taken at the shop floor level to promote the safety and health of workers. Such actions should be theory driven because accident theories help to pinpoint dangerous factors in the company, and they offer directions for practical solutions. The focus of these actions is the behavior and attitudes of the workers subject to the highest risk of occupational accidents and diseases.

Short History of Occupational Safety and Health

The history of occupational safety and health starts in the fourth century BC, when Hippocrates described the occurrence of lead poisoning among miners. Later, another Roman physician, Galen, wrote about various occupational diseases (Fraser, 1989). However, the first written source to mention a disease as actually being occupational in origin was Agricola's *De Re Metallica* (1556), which ends with a note on diseases and accidents among miners. At that time one woman could be married to as many as seven miners, one after the other. Eleven years later (1567) Paracelsus published *Von der Bergsucht und anderen Bergkrankheiten*, describing the diseases of smelter workers and metallurgists.

Finally, in 1700, Ramazzini published a book entitled *De Morbis Artificum Diatriba*, in which he examined the health conditions of workers in over 50 occupations. Because of his success, he is called the father of occupational medicine (Carnevale et al., 2012).

Observations on occupational diseases did not lead to legislative action until 1802, when the Health and Morals of Apprentices Act was confirmed in England. This act was then improved by the Factory Acts in the middle of the nineteenth century. The first compulsory social security system for wage earners was also set up about that time, in Germany in 1883. The Workmen's Compensation Act followed in Britain in 1897 providing for automatic compensation to an injured worker. In 1919 occupational safety and health became a truly worldwide concern when the International Labor Organization was founded in Geneva to establish International Conventions and International Recommendations for the development of standards and procedures for occupational health and accident prevention (Fraser, 1989).

Theories on Occupational Accidents

A feature common to different definitions of an occupational accident is the unexpected and unintentional nature of the incident. When physical injury is added, an occupational accident can be defined as an unexpected and unintentional series of events leading to the physical injury of a person at work. The other precondition for an occupational accident is a contractual relationship between the employer and the employee (Figlio, 1985).

Until the end of the nineteenth century people thought that accidents were brought on by their sins or were 'God's will.' This idea was based on the fact that most hazards were caused by natural phenomena. People had very limited means with which to resist the forces of nature. The onset of industrialization in the twentieth century meant that production moved from homes to factories. Since people had built these environments, most of the hazards were also man-made. Thus the idea of accident prevention fell on fruitful ground.

Theories on Accident Causes

The first scientific theory attempting to explain accidents was the theory of accident proneness. Greenwood and Woods (1919) insisted that 'accidents occur to a limited number of individuals who have a special susceptibility to accidents' (preface). Accident proneness was presumed to be a stable personality trait. The theory assumed that some people have an inherent tendency to be involved in accidents. The need for the term 'accident proneness' was so strong that it was independently found both in England (accident proneness) and in Germany (Unfallneigung) (Burnham, 2008).

The theory of accident proneness has been criticized on both theoretical and empirical grounds. The studies published have displayed conceptual confusion surrounding the meaning of accident proneness. Accident proneness is understood as a unitary trait, as a general characteristic, and as an innate and unmodifiable characteristic, and it is thus used as an explanatory concept. The statistical criticism of the accident proneness theory culminates in the question of whether the entire population has the same risk to be involved in an accident or not. Because the accident proneness theory is based on an incorrect assumption of homogeneous exposure to risk, it should be substituted by the concept of accident liability, which is based on the assumption of the inequality of exposure. Also the

personnel selection method based on the accident proneness has been ineffective and showed a review of 26 studies (Guastello, 1993).

According to the danger factor theory, an accident happens when a worker and a danger factor meet so that the worker injures himself or herself. The accident is a stochastic event, the probability of which is determined by the number of danger factors. The most critical danger factors are those with the highest energy content (Skiba, 1973). This theory is mostly based on practical experience.

A more sophisticated version of the danger factor theory is the ergonomic approach to the causes of accidents. The ergonomic approach assumes that disturbances in the flow of information increase the risk of accident occurrence. The exchange of information between a worker and his or her environment or between a worker and coworkers or supervisors is a precondition to accident avoidance (Saari, 1984).

According to deviance theory (Kjellén, 1984), a chain of events leading to an accident starts, when in the production process happens something which deviates from the normal process. In the concluding phase, the loss-of-control of the production system and the inadvertent release energy are activated. In the injury phase body starts to absorb energy, which is nearby the danger factor theory.

The normal accident theory assumes that accidents are inevitable or 'normal' in some types of technological systems (Perrow, 1984). Accidents generally begin with failures of one or more lower level components of the organization. Then they escalate to higher levels of organization. The examples of this kind 'normal' accident come from nuclear power industry.

A sociological theory of occupational injuries insisted that injuries are produced by social relations of work. The theory assumes that social relations at workplace generated errors, which lead to occupational injuries. Increased autocontrol of workers and management involvement in safety are the means to prevent occupational injuries (Dwyer, 1991).

The Berkeley group from the University of California studied safety in the high reliability industries like air traffic control systems, nuclear power plants, and submarines. They are complex technological systems that put a high premium on reliability because their operating environments seldom offer them a second chance. The high reliability theory emphasizes the following conditions as being necessary, but not sufficient, for ensuring reliability: a strategic prioritization of safety, careful attention to design and procedures, a limited degree of trial and error learning, redundancy, decentralized decision making, continuous training often through simulation, and strong cultures that encourage vigilance and responsiveness to potential accidents (Weick et al., 1999).

Comparing these two theories, Shrivastava et al. (2009) have insisted that theories look at the same phenomenon from different points of time. High reliability theory asserts that accidents are avoidable if the organizations take enough pains to make workplace safe. Thus accident-free environments are possible in the long run. Normal accident theory holds accidents in complex and tightly coupled system inevitable. Accidents can occur at every minute in the organization.

The latest theory of occupational injuries is the resilience theory starting from environmental sciences. The theory stressed the organization's ability to anticipate and avoid unpredictable challenges. A resilient organization is flexible in such a way that operations can return to the normal process with only minimum damage. Occupational injuries are according to resilience theory due the variation of human performance. When the human performance cannot stabilize, the technology should build such as allowing a person to make mistakes (Hollnagel et al., 2006).

Theories on Risk Taking

The recent theories related to accidents are those on risk taking. They have been developed in research on traffic safety. Most of the authors have thought that their theories can be applied to risk-taking situations in work life as well.

The risk compensation theory assumes that people have a constant level of risk that they are ready to accept. If the risk level decreases due to safety measures (e.g., seat belts in cars), people tend to adjust their goals (e.g., try to arrive sooner at their destination) so that the risk level returns to exactly the same level as before. The effects of safety measures are thus eliminated in the long run (O'Neill, 1977). Because the assumption of exact compensation is criticized, the concept of risk compensation is often replaced by the concept of behavioral adaptation.

According to the risk homeostasis theory, the risk of traffic accidents is determined by the homeostatic system, which is analogous to the room temperature system. The main assumption of the theory is that the target level of risk is the only factor determining accident rates. Perceived benefits and costs of risk taking and cautious behavior are the factors controlling the level of target risk. The risk homeostasis theory argues that the effect of a safety measure depends on its motivational nature for the people who use it. The effect of a measure (e.g., mandatory seat belt use), which does not motivate its users will be eliminated in the long run as the drivers take more risks. However, safety can be enhanced by measures that motivate people to desire safety (Wilde, 1982). If such motivation can be achieved, the effect of the measure may be greater than what technical calculations show.

A complementary theory for the risk homeostasis theory is the risk motivation theory (Trimpop, 1994). This model assumes that both personality and situational factors influence one's risk perception, which is divided into physiological, emotional, and cognitive components. Each one has its own target level of risk, which contributes to the total target level of risk. Based on the rational calculation of costs and benefits, a risk taker is motivated to act. These acts influence the environment, which then gives either positive or negative feedback. The actor can then perceive the risks in a new and better way.

The theory of zero risk received its name from the assumption that drivers of motor vehicles adapt to the risks involved in driving to the point that they do not generally feel any risk in a traffic situation or their subjective risk assessments approach zero. People avoid the feeling of risk just as they avoid pain. There is a risk threshold, however, above which risk is experienced as aversive. A driver tends to satisfy his or her motives in traffic, as well as in other areas of life. This need to satisfy motives pushes him or her to drive faster and more hazardously. Both excitatory and inhibitory motives influence

the decision making of a driver. The most hazardous excitatory motives are 'extra motives' (i.e., those outside traffic, such as the saving of time and effort) which prompt the driver to increase speed. The increased risk of an accident is related to the strength of these extra motives (Näätänen and Summala, 1976).

If the psychological portrayal of a human being with respect to accident theories is analyzed, a change can be seen. The first theories assumed that a human being is a passive accident victim who has an inherent tendency to be involved in accidents. In the more recent theories, a human being is an actor who actively searches for information from the environment. The theories, however, accept that a human being is not perfect and does make mistakes and errors. As technology becomes more reliable, a human being will be the greatest risk factor for himself or herself.

The Behavior Modification Approach

The approach of behavior modification is based on an operant learning behavior. Behavior is seen as a function of its relation to various prior and current events. The preceding events are called antecedents, and the events following the performance are consequences. If an act has increased in rate after a particular consequence has reliably followed it, the consequence is reinforcing. But if the rate has decreased, the consequence is aversive. Lastly, if a behavior disappears when a prior reinforcing consequence is continued, the situation is called extinction (Sulzer-Azaroff et al., 1994).

Interventions

The training of workers is usually the first form of intervention included in a behavior modification program. The aim of training is to teach safe work habits. The training must be as illustrative as possible, for example, showing photographs of workers using safety glasses. It is also possible to show unsafe scenes and ask workers to point out the unsafe actions. The important point in training is to repeat the safe habits in different ways for the workers. Several studies, however, have shown that training alone is not sufficient to improve safety performance (Sulzer-Azaroff et al., 1994).

The alternative method for traditional classroom teaching is an informational safety campaign. The first prerequisite for an effective informational campaign is to set a clear objective with a specific message that is positive, simple, plausible, and comprehensible. The campaign material should be designed to attract attention. Posters should be placed at crucial points of action, for example, on the wall of the canteen at a construction site. The effect of a campaign is usually short-lived. The types of intervention used in informational safety campaigns have not influenced the number or the seriousness of accidents.

The other alternative method is to utilize small groups in accident prevention. This idea is based on experiences from quality circles, which are problem-solving groups largely used in the manufacturing industry. As in the case of quality circles, small accident prevention groups consist of members of the same organization meeting regularly to solve safety-related problems. In a Finnish shipyard the small groups had from 4 to 13 members, and they met 5–11 times during 1 year. The two tasks of the groups were to eliminate the obstacles to good order and safety and establish new housekeeping practices – behavior modification. The program was successful both in improving housekeeping and in reducing accidents (Saarela, 1990).

Based on a social psychological approach, a discussion method was used in one Finnish company to reduce work-related traffic accidents. Every team had three sessions: in the first one, drivers gathered problems related to driving during work hours. These problems were gathered to a report and based on it, drivers made suggestions to solve these problems on the second round. In the third round, drivers made decisions on commitments to change their driving behavior. Then drivers wrote down the common decision, signed it and put the paper in their own pocket. After this process the number of traffic-related accidents decreased significantly by 72% (Salminen, 2008).

Goal setting is the other main method to improve performance. The first requirement for successful goal setting is that the workers who are the subjects of the intervention participate in the setting of the goals. Difficult goals inspire better performance than easy goals do. The best goals are difficult, but achievable. The utmost goal should be divided into subgoals that are easy to understand. The first goals should be rather concrete, such as 'we should use hard hats all the time at the construction site' (Locke and Latham, 1990).

Positive Reinforcement

Positive reinforcement refers to an increase in the frequency of a response following the presentation of a positive reinforcer. Whether or not a particular event is a positive reinforcer is determined empirically. The effectiveness of the reinforcer depends on several factors. Reinforcement given immediately after the response is more effective than the feedback given after a long delay. The greater the amount of a reinforcer contributes, the more frequent the response presents. Reinforcers that are highly valued lead to greater performance. A continuous schedule of reinforcement, where a response is reinforced each time it occurs, is the most effective (Kazdin, 1975).

Positive reinforcement should be given soon after the observation of a desired behavior. It ought to be given regularly, for example, once a week. The feedback should be visible and given to all workers at the same time. One way to fulfill these requirements is to place a special information board on the wall so that a curve illustrates the development of the safety score on the experimental area from one observation to another. One example of such an information board is the TUTTAVA-board developed at the Finnish Institute of Occupational Health (**Figure 1**), TUTTAVA being a Finnish acronym for a program to help improve production, quality, and safety.

One form of positive reinforcement is the use of a token economy. A token is a tangible object, for example, a lottery ticket, which can become a conditioned reinforcer. The use of tokens may shorten the delay between the behavior and the consequence, reinforce the behavior any time and any place, and provide a visible record for evaluation. The weakness of the token economy is that it pictures behavior as a simple series of stimuli and reactions. The token economy is, however, rather inexpensive for companies to operate (Brown, 1978).

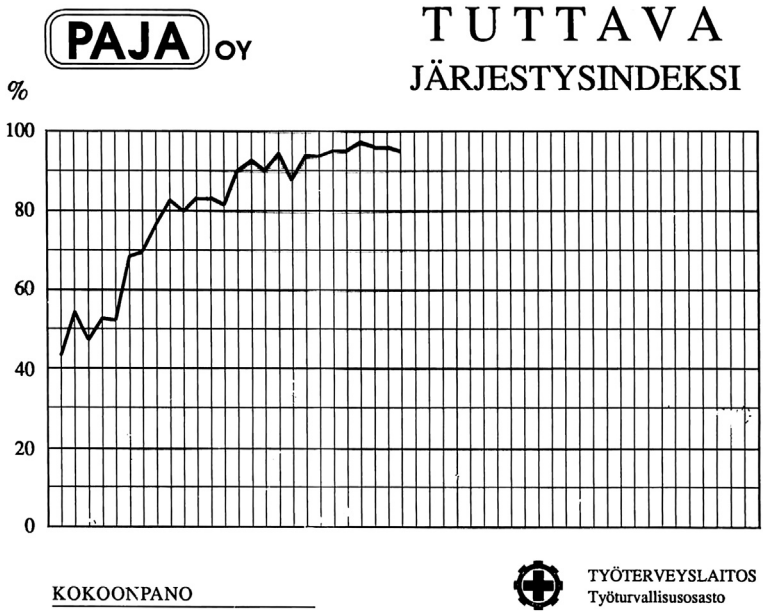

Figure 1 TUTTAVA information board.

Measurement of Safety Behavior

It is useful to start a behavior modification program with accident analysis. The analysis of accidents during the previous 3–5 years pinpoints the hazardous sites of the workplace. It is worthwhile classifying accidents as unpreventable and preventable, that is, as whether or not workers could avoid a similar accident if they changed their behavior. The researcher should then discuss the actions that could prevent accidents with the on-site personnel. Using the results of this analysis, the researcher collects 10–20 safety practices that should be phrased positively (Komaki, 1986).

The next step is to prepare the measurement of safety performance. Generally an observation method is used for this purpose because the object of measurement is the external behavior of the subjects. The researchers should select observation periods at random. When the observation is done at different times of the day, the workers behave normally and do not avoid the most hazardous tasks. The observation can be done from a single observation point or by walking through the observation area; the latter method is more common.

The observation sheet includes, at a minimum, a list of the desired safety practices. Usually it is constructed with only a dichotomized classification. The behavior of the observed worker is either safe or unsafe, as defined by the criterion of behavior included in the observation sheet. The third possible category is 'not observed,' under which the practice does not occur (Komaki, 1986).

One example of an observation method is ELMERI, ELMERI being the Finnish acronym for a workplace safety and health observation method. It has been developed through cooperation with safety authorities, safety managers, and safety representatives from selected workplaces. The observation method consists of 26 items, divided into 7 groups: (1) safety behavior, (2) order and tidiness, (3) machine safety, (4) industrial hygiene, (5) ergonomics, (6) walkways, and (7) first aid and fire safety. For each item, the observer makes a decision as to whether the item is correct according to the safety criteria defined in the handbook of the observation method. The alternatives for the observer are 'correct,' 'not correct,' or 'not observed.' The 'not observed' alternative is used, for example, when there are no machines with permanent means of access in the workplace. It takes about 15 min to observe one workstation with the ELMERI method, and only a few hours of training are needed to reach an acceptable level of interobserver reliability. It has been shown that a high safety index from ELMERI is related to a low accident rate (Laitinen et al., 1998).

The walkthrough observation rounds should be repeated once a week during 5–10 weeks. The suggested time for each round is 30–60 min. The minimum number of observations during one observation round is 100. The results of the observations are calculated as a safety score as follows:

$$\text{safety score} = \text{correct}/(\text{correct} + \text{not correct}) \times 100.$$

The higher the value of the safety score, the better the safety situation within the workplace (Komaki, 1986).

The reliability of the observation is based on the interobserver agreement. Therefore, two observers should independently assess the behavior of the same workers. It is desirable to conduct reliability checks when both the observers and the observed workers are unaware that checks are being conducted. The reliability is calculated by dividing the smaller frequency by the bigger frequency and multiplying by 100. The reliability should be at least 80% to be acceptable (Kazdin, 1975).

Experimental Designs

The effect of the intervention is demonstrated by the experimental design. It has two tasks. First, it should show that the change in behavior is permanent. Second, it should prove that

the change in behavior is due to the intervention. There are three main types of experimental designs.

Reversal or ABAB Design

The reversal design demonstrates the effect of the intervention by alternating the presentation and removal of the program over time. The purpose of the design is to demonstrate a functional relationship between the target behavior and the intervention. The first task in a reversal design is to measure the baseline rate of behavior, which describes the behavior in a normal situation before the intervention. The baseline period (referred to as phase A) is continued until the rate of the response becomes stable.

Then, in the experimental phase (referred to as phase B), the intervention is carried out. This phase continues until the behavior reaches a stable level or diverges clearly from the baseline level. Now the change in behavior is evident, but the cause of change is unclear.

In the reversal phase (phase A) the intervention is withdrawn. The target behavior usually returns to or near the original baseline level. The purpose of the reversal phase is to determine whether the behavior would have remained unchanged if the intervention had not been introduced. When the behavior reverts to the baseline, it is possible to reinstate the intervention (phase B). The design is called the ABAB design because the phases A and B are alternated (Kazdin, 1975).

Control Group Design

Another way to demonstrate the effect of intervention is the classical experimental design with experimental and control groups. The intervention is done with one group (the experimental group) and not done with the other group (the control group). Immediately before and after the intervention the behavior of the workers is assessed. The basic requirement of the study design is the similarity of the groups. The best procedure to control for systematic differences between the groups is to allocate the workers to groups at random (Kazdin, 1975).

At the shop floor level, a random classification of workers is not always possible for production reasons. Then two departments, which are as similar as possible, are selected as experimental and control groups. In addition to nonrandomness, overlapping of the information between the departments may be a problem in the comparison of departments.

Multiple-Baseline Design

The third experimental design to show the effect of interventions is the multiple-baseline design. The aim of this design is to demonstrate that the change in behavior is associated with the introduction of the contingency at different time points. This design requires the use of two or more interventions. After the baseline reaches a stable rate, one of the interventions is introduced, while baseline conditions are continued for other interventions. When the rates are stable once again, the second intervention is introduced. The effects of each intervention are established only after it has been introduced.

One problem that can occur with the multiple-baseline design is that changes in behavior take place too early. If behavior changes before the intervention are introduced, it is not clear whether the intervention is responsible for the changes or not (Kazdin, 1975). Another problem is that the effects of the different interventions could be confused.

Analysis of Intervention Data

Intervention studies often require the use of statistical tests appropriate for single-case experimental designs. Conventional t and F tests can be used in the statistical analysis of intervention data if the data are not serially dependent. This practice requires that the autocorrelation (i.e., correlation between data points separated by different time intervals) not be significant. The other statistical method is to use tests for time series. The best for a short time series ($n < 10$) is Kendall's tau, which analyzes the mean trend.

It is possible to complement the quantitative data of the effects of the intervention study by qualitative data. Qualitative methods are the most useful for analyzing the meanings and symbols of the culture or subculture. In an intervention study, the most useful qualitative method is perhaps a discourse analysis. With this method, a researcher can analyze the possible change in the attitudes of the workers participating in the intervention.

In the reports of behavior modification studies, graphs should be used to illustrate both the study design and the results. It is easy to show the development of a safety score with a line graph. In some cases, it is possible to use the same graph to give feedback to workers and to illustrate the results in the report.

High-quality intervention studies help to avoid the wasting of time, money, and effort. Shannon et al. (1999) have presented eight criteria for an intervention study of high quality. First, the objectives of the study or its underlying hypotheses are clearly presented. In the experimental design, the subjects are randomly allocated to experimental and control groups. The study has external validity and therefore its results can be generalized to other workplaces. The outcome measurement with acceptable reliability and validity should be appropriate to the objective of the intervention. Qualitative data can be used to supplement quantitative data. Diffusion of intervention manipulation from the experimental group to the control group (overlapping) or improper randomization is the main threat to internal validity. Appropriate statistical analysis should be applied in the analysis of the data set. The conclusions should address the program objectives and the limitations of the study and be supported by the analysis. These criteria are useful both in the assessment of other studies and in the planning of one's own study.

Future Recommendations for Workplace Safety and Health

The fatality rate in Finland has decreased from 7.2 to 1.2 deaths per 100 000 workers from 1975 to 2009 (Statistics Finland, 2009). Most of this positive development is due to technical solutions to prevention and improved safety inspections. Nowadays, most of the technical solutions are in use and simple accidents can be prevented. Only the more complex

accidents need to be tackled; they are due to the interaction between technical failure and human error.

Occupational accidents cause disturbances in production. When a serious accident occurs, all workers stop their work and go to help. According to Finnish calculations of accident costs, this participation forms the highest costs to the employer. Some customers therefore demand that the number of occupational accidents should stay under the acceptable limit.

Traffic research in Sweden has produced an idea called Vision Zero. According to this idea, no one should be killed or sustain an injury resulting in permanent impairment. Some workplaces, at least in Finland, have already accepted the idea of Vision Zero. In other words, they follow the principle that no accident is acceptable and that all accidents are preventable. The zero accident forum is a voluntary-based community of the companies, which wanted to reduce their accidents. In the forum sessions and intranet they can change their experiences about the accident prevention and pick up the best practices.

In the European Union, national legislation on workplace safety is based on Framework Directive 89/391/EEC. The implementation of the national legislation has been rather successful in the Finnish workplaces. Another successful example is the Occupational Safety System in Germany, through which stakeholders can provide preventive services.

These trends increase the expectations for safety modification behavior. Workplaces require more effective ways of changing the attitudes of workers to a more positive direction and activating them toward work safety. On the other hand, a deeper understanding of human behavior based on the advances of social and behavioral sciences makes it possible to respond to these increased requirements.

See also: Ethical Codes, Professional: Business Codes; Human Resource Management, Psychology of; Organization: Overview; Organizational Climate; People in Organizations; Stress at Work; Wellbeing and Burnout in the Workplace, Psychology of; Workplace Autonomy.

Bibliography

Brown, R.L., 1978. Adapting token economy systems in occupational safety. Accident Analysis and Prevention 10, 51–60.

Burnham, J.C., 2008. Accident proneness (Unfallneigung): a classic case of simultaneous discovery/construction in psychology. Science in Context 21, 99–118.

Carnevale, F., Mendini, M., Moriani, G., 2012. Bernardino Ramazzini Medical and Physiological Works. Cierre Edizioni, Verona.

Dwyer, T., 1991. Life and Death at Work. Plenum Press, New York.

Figlio, K., 1985. What is an accident? In: Weindling, P. (Ed.), The Social History of Occupational Health. Croom Helm, London.

Fraser, T.M., 1989. The Worker at Work. Taylor & Francis, London.

Greenwood, M., Woods, H.M., 1919. The Incidence of Industrial Accidents upon Individuals with Special Reference to Multiple Accidents. Medical Research Committee, Industrial Fatigue Research Board, Report No. 4. Her Majesty's Stationery Office, London.

Guastello, S.J., 1993. Do we really know how well our occupational accident prevention programs work? Safety Science 16, 445–463.

Hollnagel, E., Woods, D., Leveson, N., 2006. Resilience Engineering. Ashgate, Burlington.

Kazdin, A.E., 1975. Behavior Modification in Applied Settings. Dorsey Press, Homewood, IL.

Kjellén, U., 1984. The deviation concept in occupational accident control – I: definition and classification. Accident Analysis and Prevention 16, 289–306.

Komaki, J.L., 1986. Promoting job safety and accident prevention. In: Cataldo, M.F., Coates, T.J. (Eds.), Health and Industry. A Behavioral Medicine Perspective. Wiley, New York.

Laitinen, H., Rasa, P.-L., Räsänen, T., Nykyri, E., 1998. An observation method as a tool for workplace risk assessment. Technology, Law and Insurance 3, 57–61.

Locke, E.A., Latham, G.P., 1990. A Theory of Goal Setting & Task Performance. Prentice Hall, Englewood Cliffs, NJ.

Näätänen, R., Summala, H., 1976. Road-user Behavior and Traffic Accidents. North-Holland, Amsterdam.

O'Neill, B., 1977. A decision-theory model of danger compensation. Accident Analysis and Prevention 9, 157–165.

Perrow, C., 1984. Normal Accidents: Living with High Risk Technologies. Princeton University Press, Princeton, NJ.

Saarela, K.L., 1990. An intervention program utilizing small groups: a comparative study. Journal of Safety Research 21, 149–156.

Saari, J., 1984. Accidents, and disturbances in the flow of information. Journal of Occupational Accidents 6, 91–105.

Salminen, S., 2008. Two interventions for the prevention of work-related road accidents. Safety Science 46, 545–550.

Shannon, H.S., Robson, L.S., Guastello, S.J., 1999. Methodological criteria for evaluating occupational safety intervention research. Safety Science 31, 161–179.

Shrivastava, S., Sonpar, K., Pazzaglia, F., 2009. Normal accident theory versus high reliability theory: a resolution and call for an open system view of accidents. Human Relations 62, 1357–1390.

Skiba, R., 1973. Die Gefahrenträgertheorie. Forschungsbericht Nr. 106. Bundesanstalt für Arbeitsschutz and Unfallforschung, Dortmund.

Statistics Finland, 2009. Accidents at Work 2009. Official Statistics of Finland, Labour Market 2009. Tilastokeskus, Helsinki. http://www.stat.fi/til/ttap/2009./ttap_2009_2011-11-30_tau_001_fi.html (retrieved 22.11.12).

Sulzer-Azaroff, B., Harris, T.C., McCann, K.B., 1994. Beyond training: organizational performance management techniques. Occupational Medicine: State of the Art Reviews 9, 321–339.

Trimpop, R.M., 1994. The Psychology of Risk Taking Behavior. North-Holland, Amsterdam.

Weick, K.E., Sutcliffe, K., Obstfeld, D., 1999. Organizing for high reliability. Research in Organizational Behavior 21, 81–123.

Wilde, G.J.S., 1982. The theory of risk homeostasis: implications for safety and health. Risk Analysis 2, 209–225.

Workplace Stress

Michael Sliter and Zhenyu Yuan, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article defines workplace stress and discusses major historical models for explaining stress, namely the Institute for Social Research model and the job demands–control model. Next, we discuss contemporary theories of workplace stress, including conservation of resources, job demands–resources, and the effort–reward imbalance model, followed by a discussion of primary, secondary, and tertiary workplace interventions. We continue by examining contemporary issues and advances in stress research. Finally, we conclude by calling for future research in several highlighted areas, such as technology-related stressors, diversity, and eustress.

Stress is one of the most universal experiences in the workplace. In fact, it is such a pervasive issue that for decades, a large portion of workplace research has consistently been dedicated to investigating the antecedents and consequences of workplace stress. Despite this volume of literature, the term 'stress' is still often misused in both research and colloquial contexts. At its core, workplace stress can refer to the overall process by which employees are met with demands, and their subsequent physical, cognitive, and emotional reactions to those demands. The demands, themselves, are correctly referred to as stressors, whereas the outcomes of, or reactions to, the stressors are correctly referred to as strains. A variety of workplace stressors have been heavily investigated, including workload, role conflict, role ambiguity, organizational constraints, interpersonal conflict, and time pressure, to name a few. Similarly, research has investigated a host of strain outcomes, including burnout, physical symptoms/health (e.g., illness; back pain), disengagement, withdrawal, depression, and other well-being variables.

The extant literature has generally made a compelling case that workplace stressors can not only impact individuals through strain outcomes, but also organizational effectiveness (such as individual and group performance, absenteeism, and turnover). Given the negative impact of the stressor–strain process in the workplace, researchers have endeavored to create theoretical models to determine how this process operates, as well as to determine aspects of the individual and the workplace that could potentially exacerbate or buffer the impact of workplace stressors. In this article, we will briefly describe two historical models of workplace stress, followed by discussing three contemporary models. Next, we describe work conducted on workplace stress interventions, and finish by discussing contemporary issues in workplace stress research and areas for future research.

Historical Models of Workplace Stress

ISR Model

Though several general models of stress were proposed in the mid-1900s (e.g., General Adaptation Theory; Selye, 1956), workplace stress was not explicitly investigated until the 1960s and 1970s. Perhaps the first stress model that explicitly investigated stress in the workplace originated from the University of Michigan's Institute for Social Research (ISR), and therefore became known as the ISR model (French and Kahn, 1962; Kahn and Byosiere, 1992). The ISR model is inclusive, and consists of several steps or factors that could potentially impact a person's experience of stress at work. The first step involves aspects of the objective work environment and broadly contains anything in an employee's surroundings that could be considered a stressor. Temperature, noise, hours worked, number of coworkers, physical space, and similar environmental factors would be encompassed by the first step of the ISR model.

The second step concerns the psychological environment – that is, how an employee *perceives* the objective workplace environment. This perception dictates whether environmental variables are perceived as 'threatening', or as a stressor, and leads to immediate cognitive, behavioral, or physiological activation. This activation, over time, can result in employee strain outcomes (e.g., burnout, depression) and other negative effects, such as reduced performance and withdrawal. The ISR model recognizes that perception of stressors are not unilateral across all people, and hence takes into account individual differences, including demography, personality, and quality of interpersonal relations. Any of these factors could impact this model at any point, including initial perception of the stressors, immediate reactions to the stressors, and long-term effects of these reactions.

The ISR model has not been empirically tested, but rather has served as an important guide for researchers interested in stress. The influence of the ISR model can be seen in other older stress models (job demands–control (JDC); facet model; Beehr and Newman, 1978), as well as in contemporary models of stress (e.g., job demands–resources (JD-R)). This model, therefore, laid an important groundwork for further theoretical development of workplace stress models.

JDC Model

A second model that played a major role in the development of modern workplace stress theories is the JDC model (Karasek, 1979). The JDC model distinguished two elements of the work environment to explain job strain. Job demands refer to the taxing aspects of work whereas job control taps into the discretion individuals have over job-related decision making.

Individuals experience strain when facing a combination of high job demands and low job control (strain hypothesis). On the contrary, when high job demands are matched by a high level of job control, people feel activated at work (learning hypothesis). The buffer hypothesis was later derived from the JDC model and states that job control can buffer the negative effect of high job demands (Van der Doef and Maes, 1999). The model was later expanded into the job demands–control–support (JDCS) model by taking into consideration the important role of support (Johnson and Hall, 1988). Accordingly, the buffer hypothesis is concerned with the synergistic effects of high control combined with high support. Historically, the JDCS model has been perhaps the most influential model in job stress research, though it is not without limitations. For instance, although considerable empirical support was found for the strain hypothesis, the buffer hypothesis has received limited empirical support (Van der Doef and Maes, 1999). The empirical support for the buffering effect of control in the JDC model is mainly limited to cross-sectional studies that measured job demands and job control at a comparable level of specificity. The synergistic moderating effect of job control and support, as stated in the JDCS model was largely unsupported. Moreover, the model was also criticized for neglecting other important facets of the work environment (Bakker and Demerouti, 2007).

Contemporary Models of Stress

Across the past 20 or so years, there have been major advances in theoretical models designed to better explain occupational health. Though a discussion of all models is beyond the scope of this article, we will elaborate upon three of the most-researched models: effort–reward imbalance (ERI), conservation of resources (COR), and JD-R.

ERI Model

A first contemporary theory used in the workplace stress literature is the ERI model, which has gained popularity in recent years, particularly in Europe. Originally established in sociological-medical literature (Siegrist et al., 1986), the ERI model takes into account the relationship between efforts and rewards associated with the workplace. Efforts are actions taken to respond to workplace demands – both stressors and obligations – and rewards represent benefits distributed by the organization, including money, promotion opportunities, belonging, and esteem/appreciation. The ERI posits that individuals evaluate the rewards they receive in exchange for their efforts to determine whether the rewards are sufficient. A long-term imbalance in the effort–rewards relationship – where work, in general, is characterized by high effort and low rewards – might cause strain reactions, such as burnout and physical symptoms. Furthermore, commitment is thought to impact the strain outcomes of this imbalance, where people who are highly committed (or 'overcommitted') to the job will experience worse strain outcomes than those who are less committed.

The ERI model shares many of the characteristics of Equity Theory (Adams, 1965), which also posits that employee effort (inputs) should match rewards (outputs). However, Equity Theory has primarily been applied in order to explain employee behaviors, such as counterproductive work behaviors, whereas the ERI model – originating from a medical tradition – has been primarily focused on explaining health-related outcomes, particularly objectively measured strains. Indeed, imbalances in the effort/reward ratio, where efforts exceed rewards, have been linked in numerous studies to cardiovascular incidences, cardiovascular risk symptoms (e.g., hypertension; cholesterol), sickness absence, unhealthy behaviors (e.g., smoking cigarettes and drinking alcohol), and mental health (depression and burnout). The most frequently recognized limitation of the ERI model is that the effort–reward balance/imbalance is assessed using a scale that assesses a ratio between these factors, and is therefore difficult to evaluate from a psychometric perspective (though efforts have been made to address this concern; see Siegrist et al., 2004). For the most recent review of the literature on the ERI model, see Van Vegchel et al. (2005).

Conservation of Resources

The COR theory takes an ecological approach to stress research by incorporating environmental and internal processes into the concept of resources (Hobfoll, 1989, 2001). Resources refer to elements that are valued by individuals or factors that are instrumental for attaining the valued elements. They can be delineated into objects (e.g., food), personal characteristics (e.g., sense of humor), conditions or states (e.g., marriage), or energy resources (e.g., time). The fundamental tenet of COR theory is that individuals are motivated to obtain, retain, protect, and foster their resources. Accordingly, psychological strain will occur: (1) when resources are threatened, (2) when resources are lost, or (3) when individuals fail to gain resource return proportionate to their resource investment. Moreover, individuals are more sensitive to resource loss than resource gain. Individuals will need to invest resources in order to cope with threatened or actual resource loss and to reap resources. Therefore, people with fewer resources are less able to invest their resources and more prone to strain, leading to a resource loss spiral and the adoption of a defensive strategy to protect their resources. On the contrary, people with more resources are more likely to focus on protecting their resources and attaining more resources, thus resulting in a resource gain spiral.

In comparison to other job stress models, the COR theory offers a comprehensive account of stress by taking into account both contextual and self-related resources. Moreover, the COR theory also gives explicit explanations for the conditions that could lead to and result from stress, thus rendering a more dynamic perspective in stress research. As such, the theory has been applied to and supported by stress studies ranging from traumatic stressful events to chronic job stressors (Hobfoll, 2001; Hobfoll and Shirom, 1993). Although sometimes criticized for being too general in terms of conceptualization and theorization (Hobfoll, 2001), the COR theory will continue to provide unique insights into workplace stress research.

Job Demands–Resources

The JD-R model (Demerouti et al., 2001) contends that job aspects fall into two general categories (i.e., job demands and

job resources) and provides a parsimonious model to explain workplace stress. Specifically, job demands refer to the physical, psychological, social, or organizational characteristics of the job that take substantial physical and/or psychological efforts to address at the expense of physical and/or psychological costs. Common examples of job demands include work overload and role conflict. Job resources refer to the physical, psychological, social, or organizational aspects of the job that are instrumental for work goal achievement, promote personal learning and growth, and/or reduce job demands and their associated negative impacts. Examples of job resources include coworker support and job autonomy.

The JD-R model elaborated on the JDC model by classifying a broad range of job characteristics into two overarching categories (Bakker and Demerouti, 2007), presumed to impact both employee strain and beneficial outcomes (e.g., engagement). Coping with job demands can exhaust one's energies and result in resource depletion, a mechanism similar to the well-established stressor–strain relationship. This mechanism is the health impairment or energetic process of the JD-R model. Job resources, due to their motivational nature, can lead to higher levels of engagement at work, thus activating a motivational process. In addition to the main effects of job demands and job resources, it was further proposed that job resources could buffer the negative impact of job demands, which is similar to the buffering effects of job control and support in the JDC model (Bakker and Demerouti, 2007).

More recently, the importance of personal resources has received increasing attention in the JD-R model (Schaufeli and Taris, 2014). Personal resources refer to the psychological characteristics that are associated with one's resiliency and perceived capability to exert control and impact the environment, including factors such as attribution style, hardy personality traits, and self-efficacy, among others. The idea behind this expansion of the JD-R model is that personal resources could influence the way people perceive job characteristics, the way people react to perceived job characteristics, and the way people feel at work (Schaufeli and Taris, 2014). Moreover, job characteristics can also impact personal resources, thus rendering a dynamic role for personal resources in the JD-R model.

The challenge and hindrance stressor framework has also recently been integrated into the JD-R model to provide a finer-grained account of different types of job stressors (Cavanaugh et al., 2000; Crawford et al., 2010). Challenge job demands, although appraised as stressful demands, tend to promote goal achievement, learning, and growth because they are perceived as generally nonthreatening hurdles to overcome in pursuit of valued outcomes. High workload, for instance, would be a challenge job demand because it is stressful but usually associated with desired outcomes of recognition and reward once the necessary tasks are accomplished. Hindrance job demands, on the other hand, impede goal achievement, learning, and growth (Crawford et al., 2010). For example, administrative paperwork could be viewed as a hindrance job demand because fulfilling the demand usually gets in the way of achieving work goals. Challenge job demands can be both motivating and exhausting since individuals have to deploy their resources to meet the challenge. Hindrance job demands, on the other hand, are associated with lower levels of motivation and higher levels of strain.

The heuristics nature of the JD-R model adds to its flexibility to different kinds of occupational contexts, which, on the other hand, limits its generalizability such that research finding concerning a specific job demand might not apply to *all* demands (Schaufeli and Taris, 2014). Moreover, the model only specifies two psychological processes through which job characteristics influence job strain, leading many researchers to incorporate other stress models to provide a fine-grained examination. With its limitations being said, the JD-R model is still one of the most influential theoretical frameworks for job stress research. Empirical evidence provides strong support for the basic contention and the two processes of the model. As is evidenced by the current review, the JD-R model has also generated more refined inquiries into specific components of the model such as the role of personal resources and important distinction between challenge stressors and hindrance stressors. The JD-R model will continue to stimulate more scientific endeavor to better understand workplace stressors (Schaufeli and Taris, 2014).

Workplace Stress Interventions

Significantly more empirical attention has been paid to identifying the negative effects of stressors than has been paid to reducing either the incidence, or the effects, of these stressors. Nonetheless, over the past 20 years, a good deal of evidence has accumulated that indicates the relative effectiveness of a variety of programs in reducing workplace stress, with some of these interventions guided by theoretical models. We will briefly discuss such programs, highlighting primary, secondary, and tertiary intervention strategies.

Primary Interventions

Primary interventions focus on prevention or elimination of stressors in the workplace through the utilization of work redesign and organizational change and development. For instance, if poor management appears to be causing strain within individuals at work, then a primary intervention could involve management restructuring. Alternately, if a job involves a good deal of physical work resulting in lower back pain (e.g., nursing), lift-assisting machines could potentially be added to reduce this stressor (therefore reducing the strain). By Tatrgeting the source of workplace stress, these primary interventions can, in theory, also eliminate the strains that result from the stressors.

Empirical research on primary intervention is relatively sparse, as changing the workplace can be costly and is not always practical or possible. However, a careful search of the literature does yield a small, but informative, group of studies that demonstrate the effectiveness of primary intervention strategies. For instance, primary interventions involving changes to the work environment, redesigning individual jobs, and training have been linked to decreased levels of employee depression and resulting absenteeism (Kawakami et al., 1997). The evidence is even more abundant that primary interventions

focused on work redesign can reduce physical strain – particularly back pain – associated with heavy lifting in the workplace (Tveito et al., 2004).

Organizations have also begun to focus on increasing autonomy or participation in the workplace as potential primary inventions. As per the JDC and JD-R models of stress, increasing autonomy and participation could work to both reduce stressors and their associated costs. Such interventions typically involve increasing decision latitude and skill discretion of individual employees without adding duties. That is, employees maintain the same tasks in their job description, but are allowed more discretion in determining how to complete the tasks. These interventions have been shown to have beneficial effects on employees, particularly when autonomy and participatory behavior are combined with environmental changes (e.g., Bond and Bunce, 2001).

Despite evidence that primary interventions can be effective, efforts such as work redesign and increasing employee autonomy and participation are not always possible or desirable from the viewpoint of management. For instance, for jobs with inherently high levels of job demands (executives, police officers, and firefighters), only so much can be done to redesign the job. Many of the major stressors in these positions are simply unavoidable. As such, the majority of research has investigated interventions designed to mitigate the effects of stressors either through stress management tactics (secondary interventions) or through healing the negative effects of stressors (tertiary interventions).

Secondary Interventions

Secondary interventions focus on teaching employees to manage stressors at work, emphasizing and teaching the use of effective coping strategies. That is, they involve teaching employees what to do when faced with a stressor. Secondary interventions most often manifest as stress management programs which can vary widely in terms of content, from time management to progressive muscle relaxation and meditation. Because of this variety, and because individual differences play an important role in individual level stress management techniques, it is difficult to make a unilateral judgment about the effectiveness of secondary interventions. However, a good deal of evidence has accumulated for individual aspects of stress management programs. For instance, significant evidence supports the benefits of meditation and relaxation training in reducing strain associated with stressor exposure. Alternately, training programs that have focused on adaptive coping strategies – time management, organization, and problem-focused coping – have shown mixed results (e.g., Macan, 1996), though it is important to note that there is little systematic research investigating the effectiveness of these types of stress management programs.

Finally, health promotion programs – a particular type of stress management program that promotes healthy living through diet, exercise, and other processes (sleep habits) – are designed to promote resiliency in employees through increasing well-being and vitality so that stressors are less likely to lead to negative outcomes. Health promotion approaches have been shown to be relatively effective in reducing strains associated with work stressors. The research in this area also indicates that the benefits of health promotion programs might cross to nonwork domains, benefiting general life satisfaction, and personal relationships. For an excellent meta-analysis investigating the outcomes of different types of stress management programs, see Richardson and Rothstein (2008).

Tertiary Interventions

Whereas secondary interventions focus on helping individuals to identify, manage, and cope with their stressors, thereby reducing the negative effects, tertiary interventions focus on healing the negative effects after they have already occurred. The most common tertiary interventions provided by organizations are employee assistance programs (EAPs). These programs (now primarily externalized by contracted EAP vendors) typically offer a range of services, including counseling and referral services, where an employee can receive a limited number of mental health services free of cost (to the employee). Oftentimes, the benefits provided by an EAP extend to the immediate family of an employee, as well.

As with stress management programs, it is difficult to evaluate the effectiveness of EAPs, given that the programs differ greatly in the quality of the services, as well as the manner in which services are provided. Additionally, many of the available studies rely on single organization samples (i.e., case studies), and are reported by the providers of the EAPs, which might indicate bias. Furthermore, the largest group of studies that investigate EAPs focus on return-on-investment for organizations as opposed to employee outcomes. The results of empirical studies that actually investigate the effects of EAPs on employee strain are mixed, with some showing that EAPs have no positive effect on health or absenteeism and others have finding increased job satisfaction and lowered health care costs (implying a healthier workforce). Overall, these inconsistent results point to the importance of more systematic research on the effects of EAPs, with an emphasis on determining moderators that might contribute to these varied findings.

Contemporary Issues in Workplace Stress Research

Measurement Issues and Advances

Historically, workplace stress literature has been relatively limited by methodological concerns. Namely, the majority of this research has been self-report and cross sectional in nature, particularly before the twenty-first century. This, of course, severely limits the conclusions that can be drawn about the relationship between stressors and strains, where reverse causality (strains leading to stressors) is a possible explanation for findings, and where common method concerns are exacerbated. For instance, in a cross-sectional study linking interpersonal conflict and burnout, stress theories (e.g., JD-R) would suggest that the conflict *results* in burnout, but, due to the cross-sectional nature of the data, it may be that more burnt-out people are more likely to initiate conflict with others.

However, there have recently been strides in the methodology of workplace stress research. Specifically, the long-ignored call for more longitudinal studies on the stressor–strain

relationship has been recently heeded, and much of the stress literature now published involves two or more time points of data collection. This allows researchers to demonstrate temporal contiguity – that the stressors precede the strains. The time lag between waves of data collection can provide insights into both the short- and long-term implications of stressors. For instance, a study that examined the stressor–strain relationship across several years can provide insight into potential negative effects associated with prolonged exposure to stressors. Shorter term studies, however, can provide insight into the immediate reactions to stressors, and provide a more microview of the phenomenon.

Along those lines, diary studies have become more popular (though still relatively infrequent) in workplace stress research. Diary studies allow researchers to survey people multiple times within a relatively short period of time. For instance, employees could respond to measures once per day for a month. Assessments are typically conducted through communication technology (CT), such as e-mails, smart phones, or tablets (which may be provided by the researchers). A similar method – experience sampling – assesses people multiple times within a single day. Both of these methods have provided more insight to the daily, within-person impact of stressors. These methods have been particularly effective in understanding how mood can impact the perception of stressors, and how emotions can mediate the stressor–strain relationship (see Harris et al., 2003, as an example).

There has also been a push to understand how work stress might be more than just an individual level phenomenon. That is, given that people at work are often nested within a work group, within a department, within an organization, it is likely that the experience of stress is not independent of these levels. Based on this literature, group-level variables (such as workplace climate or group efficacy) could potentially have a direct effect on individual and group reports of strain, and may also moderate the stressor–strain relationship. The immediate benefit of multilevel research is that these studies provide additional perspectives and guidance to organizations seeking to develop interventions. As mentioned earlier, most interventions tend to be focused on the individual (e.g., relaxation training, EAPs), but results from multilevel studies could potentially inform effective, group-level interventions, though little research has been conducted in this area.

In terms of assessment of stressors and strains, researchers are more frequently moving away from self-report and using coworker or supervisor reports of stressors to gain perspective on how stressors are largely the result of perception. That is, an employee might perceive their workload as high while a supervisor reports that same workload as relatively light, which offers some insight into individual differences in perception. Additionally, some workplace research actually utilizes physiological measures of strain, such as cortisol levels, blood pressure, heart rate, and diagnosed disease conditions. These more objective measures of strain allow researchers to better understand how stressors affect the body, and allow more concise and accurate validation of self-report strain measures. Typically, such studies find small, but significant relations between self-reported stressors and physiological measures.

Finally, the pursuit of different methodologies for assessing and understanding workplace stress has led to more sophisticated statistical analyses. Early stress research was marked by relatively simple tests of the stressor–strain relationship. Recently, there has been a push to incorporate analyses such as Structural Equation Modeling, Hierarchical Linear Modeling, and Latent Growth Modeling to answer more sophisticated questions about the relationships among stressors and strains, as well as antecedents, mediators, and moderators of these relationships. Overall, this push for more methodological and statistical sophistication has led to a more informed body of literature in the workplace stress area, primarily though allowing contemporary models (JD-R; ERI) to be tested more dynamically.

Contemporary Topics in Workplace Stress

With the way that the workplace has continued to change across the past 20 years, researchers are beginning to pay attention to a variety of new (or sometimes ignored) workplace stress issues. One relatively new topic is the computerization of the workforce and the impact of CT (e-mail, text messages), especially the persistent use of smart phones. More employees work in front of computers than ever before, and researchers are beginning to investigate the impact of this computer use, both in terms of the physical and the psychological impact. Particularly, new research is beginning to investigate the impact of 'cyber incivility' and 'cyber bullying', where people accidentally, or intentionally, mistreat each other using computer-mediated communication (e.g., e-mails, text messages; Lim and Teo, 2009). This preliminary research has shown that cyber incivility can have a strong, negative impact on employee strain, perhaps even stronger than face-to-face incivility. However, this area is still new, and there are many questions that have yet to be answered.

Along those lines, the persistent use of CT is beginning to spark interest in the work-life conflict literature. That is, the boundaries between the work and nonwork domains are becoming more permeable, leading to more work-life conflict. The research is quite new in this area, and researchers are actively attempting to determine which employees – in terms of individual difference and preferences – tend thrive or stagnate in occupations with these blurred boundaries. More generally speaking, a great deal of attention has been paid to the impact of work-life balance on employee strain outcomes in the past 10 years, and we expect this literature to continue to grow. More people than ever are in the workforce – particularly women – and approximately 5% of the population work multiple jobs (with this number growing). This indicates that, for many employees, work-life conflict may be increasing, and researchers are working to determine factors to lessen, or at least alter perceptions of, this conflict.

In addition to technology-related and work-life stressors, the diversification and globalization of the workforce has led to concerns about stressors and strains experienced by different groups. Specifically, there is a growing body of literature examining how women and underrepresented minorities (i.e., racioethnic and sexual orientation status) might experience a disproportionate number of stressors in the workplace. Particularly, these stressors tend to be interpersonal in nature,

where these groups experience higher levels of outright harassment, as well as more 'modern' forms of subtle discrimination such as incivility. These stressors have been shown to relate to strain in these groups, and people who share multiple minority statuses (e.g., minority females) may experience worse strain as a result of compounding forms of harassment (Raver and Nishii, 2010). As such, researchers have been investigating methods for reducing harmful stressors for at-risk populations. Climate-based interventions may be effective, such that organizations that strive to create a climate for civility, climate for women, and climate for diversity seem to have lower levels of harassment of these at-risk groups. However, research in this area is still primarily associative, and certainly merits further investigation.

A final burgeoning topic in the workplace stress literature is eustress, or so-called good stress. The most commonly used definition of eustress in the workplace literature is that eustress represents a positive perception of stressors (whereas 'distress' represents a negative perception). This perception may be influenced by a variety of factors, including individual differences and characteristics of the stressor (e.g., timing, whether it is self-imposed or external, etc.). The concept of eustress has received limited empirical attention in the research, and may be most clearly lacking in the organizational literature, though some research seems to indicate benefits from eustress. Particularly, research on the challenge-hindrance distinction of stressors seems to indicate that challenges may be indicative of eustress, which could potentially reduce strain outcomes (Van den Broeck et al., 2010). Similarly, research in demanding occupations, such as health care workers, indicate that some nurses react to stressors with feelings of energy, hope, and gratitude (Simmons and Nelson, 2001). The factors that predict eustress are still unclear, and are a clear direction for future research.

Future Directions and Conclusions

Though workplace stress research has certainly made strides across the past 20 years, it is far from complete. First, in terms of contemporary theories of workplace stress, researchers need to continue testing, and refining, these models. For instance, the JD-R model was designed to be a *descriptive, heuristic* model, which posits that demands and resources, in general, impact workplace outcomes. The model is broad and generalizable, but does not specify psychological processes that link demands and resources to strain outcomes. For a summary of the JD-R model, including its limitations and calls for future research, see Schaufeli and Taris (2014). In terms of COR theory, the field is moving away from discussing personal resources in general, and toward assessing specific resources, and how they might impact the stressor–strain relationship. Indeed, these resources could potentially be incorporated into the JD-R model, combining COR and JD-R theories, and researchers are beginning to discuss this possibility. In fact, the incorporation of multiple workplace stress theories in a single study can allow for greater theoretical precision, and result in better overall understanding of observed and predicted relationships. Researchers should also continue investigating how changing technology, including CT, can impact individual outcomes. Along these lines, this technology may have a differential impact on employees based various demographic factors. One such factor may be age, as some research has demonstrated that older workers are less skilled with technology and require more time and training to learn new technologies. Given that the workforce is rapidly aging in many countries, it is important to understand the implications of technology use on this group. More generally, topics related to workplace stress and aging should continue to be investigated.

Finally, given that diversity and inclusion-related issues are an increasingly important topic in workplace stress literature, there is a need to continue identifying stressors faced by at-risk groups, especially subtle stressors. It is also important to explicitly investigate interventions that might be effective in eliminating these stressors, with a particular emphasis on climate-related interventions. Finally, researchers are beginning to incorporate positive psychology topics, such as eustress, into the workplace stress research. These topics certainly merit investigation, particularly for identifying what types of people may thrive under stress. Positive psychology is perhaps dually beneficial, as aspects can be incorporated into the study of any of the other stress-related topics discussed.

To conclude, workplace stress research has gained significant momentum over the past two decades, and clear strides have been made in developing theories, conducting empirical research, and creating interventions. However, our understanding of workplace stress is still far from complete, and we urge researchers to continue investigating the important areas that we highlighted for research, as well as other important topical areas.

See also: Emotions and Work; Job Satisfaction; Organizational Climate; Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences; Workplace Safety and Health.

Bibliography

Adams, J.S., 1965. Inequity in social exchange. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, vol. 2. Academic Press, San Diego, pp. 267–299.

Bakker, A.B., Demerouti, E., 2007. The job demands-resources model: state of the art. Journal of Managerial Psychology 22, 309–328.

Beehr, T.A., Newman, J.E., 1978. Job stress, employee health, and organization effectiveness: a facet analysis, model, and literature review. Personnel Psychology 31, 665–699.

Bond, F.W., Bunce, D., 2001. Job control mediates change in a work reorganization intervention for stress reduction. Journal of Occupational Health Psychology 6, 290–302.

Cavanaugh, M.A., Boswell, W.R., Roehling, M.V., Boudreau, J.W., 2000. An empirical examination of self-reported work stress among U.S. managers. Journal of Applied Psychology 85, 65–74.

Crawford, E.R., LePine, J.A., Rich, B.L., 2010. Linking job demands and resources to employee engagement and burnout: a theoretical extension and meta-analytic test. Journal of Applied Psychology 95, 834–848.

Demerouti, E., Bakker, A.B., Nachreiner, F., Schaufeli, W.B., 2001. The job demands-resources model of burnout. Journal of Applied Psychology 86, 499–512.

French, J.R.P., Kahn, R.L., 1962. A programmatic approach to studying the industrial environment and mental health. Journal of Social Issues 18, 1–47.

Harris, C., Daniels, K., Briner, R.B., 2003. A daily diary study of goals and affective well-being at work. Journal of Occupational and Organizational Psychology 76, 401–410.

Hobfoll, S.E., 1989. Conservation of resources: a new attempt at conceptualizing stress. American Psychologist 44, 513–524.

Hobfoll, S.E., 2001. The influence of culture, community, and the nested-self in the stress process: advancing conservation of resources theory. Applied Psychology: An International Review 50, 337–369.

Hobfoll, S.E., Shirom, A., 1993. Stress and burnout in the workplace: conservation of resources. In: Golombiewski, R.T. (Ed.), Handbook of Organizational Behavior. Dekker, New York, pp. 41–61.

Johnson, J.V., Hall, E.M., 1988. Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population. American Journal of Public Health 78, 1336–1342.

Kahn, R.L., Byosiere, P., 1992. Stress in organizations. In: Dunnette, M.D., Huogh, L.M. (Eds.), Handbook of Industrial and Organizational Psychology, second ed., vol. 3. Consulting Psychologist Press, Palo Alto, pp. 571–650.

Karasek, R.A., 1979. Job demands, job decision latitude, and mental strain: implications for job redesign. Administrative Science Quarterly 24, 285–308.

Kawakami, N., Araki, S., Kawashima, M., Masumoto, T., Hayashi, T., 1997. Effects of work-related stress reduction on depressive symptoms among Japanese blue-collar workers. Scandinavian Journal of Work, Environment, and Health 23, 54–59.

Lazarus, R.S., Folkman, S., 1984. Stress, Appraisal, and Coping. Springer, New York.

Lim, V.G., Teo, T.S.H., 2009. Mind your E-manners: impact of cyber incivility on employee's work attitude and behavior. Information and Management 46, 419–425.

Macan, T.H., 1996. Time-management training: effects on time behaviors, attitudes, and job performance. Journal of Psychology 130, 229–236.

Raver, J.L., Nishii, L.H., 2010. Once, twice, or three times as harmful? Ethnic harassment, gender harassment, and generalized workplace harassment. Journal of Applied Psychology 95, 236–254.

Richardson, K.M., Rothstein, H.R., 2008. Effects of occupational stress management intervention programs: a meta-analysis. Journal of Occupational Health Psychology 13, 69–93.

Schaufeli, W.B., Taris, T.W., 2014. A critical review of the job demands-resources model: implications for improving work and health. In: Bauer, G.F., Hämmig, O. (Eds.), Bridging Occupational, Organizational and Public Health. Springer, Netherlands, pp. 43–68.

Selye, H., 1956. The Stress of Life. McGraw-Hill, New York.

Siegrist, J., Siegrist, K., Weber, I., 1986. Sociological concepts in the etiology of chronic disease: the case of ischemic heart disease. Social Science & Medicine 22, 247–253.

Siegrist, J., Stark, D., Chandola, T., Godin, I., Marmot, M., Niedhammer, I., et al., 2004. The measurement of effort-reward imbalance at work: European comparisons. Social Science and Medicine 58, 1483–1499.

Simmons, B.L., Nelson, D.L., 2001. Eustress at work: the relationship between hope and health in hospital nurses. Health Care Management Review 26, 7–18.

Tveito, T.H., Hysing, M., Eriksen, H.R., 2004. Low back pain interventions at the workplace: a systematic literature review. Occupational Medicine 54, 3–13.

Van den Broeck, A., De Cuyper, N., De Witte, H., Vansteenkiste, M., 2010. Not all job demands are equal: differentiating job hindrances and job challenges in the job demands–resources model. European Journal of Work and Organizational Psychology 19, 735–759.

Van der Doef, M., Maes, S., 1999. The job demand-control (-support) model and psychological well-being: a review of 20 years of empirical research. Work & Stress 13, 87–114.

Van Vegchel, N., De Jonge, J., Bosma, H., Schaufeli, W., 2005. Reviewing the effort-reward imbalance model: drawing up the balance of 45 empirical studies. Social Science & Medicine 60, 1117–1131.

World Republic

Otfried Höffe, Universität Tübingen, Tübingen, Germany

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The article starts from the diagnosis of globalization. If its implications are taken seriously, it is argued, the establishment of a world legal order and eventually of a world republic is demanded from a moral point of view. In discussing eight possible objections to the concept of a world republic, the concept's more detailed characteristics and implications emerge clearer into focus.

Our age of globalization asks for the establishment of a world legal order, ultimately even for a world state or a world republic, since 'world republic' means a legal and social community which extends over the whole world. If one compares the world republic to a ship, one might think of Antoine de Saint-Exupéry's saying: "créer le navire c'est exclusivement fonder le pente vers la mer," ("building a ship simply means to create the longing for the sea": 1959: p. 826). Political scientists examine the lumber for the ship. Their naive counterpart, an idealistic utopia, on the one hand, offers suggestions of how one might travel the sea without any lumber at all. Philosophy, on the other hand, develops concepts for a feasible 'not yet': it sketches a realistic vision of how one might carry over the common heritage of humanity, the right, and, furthermore, the political achievement of the modern era, liberal democracy, to the age of globalization. Since who has reasonable hopes, does not shrink from possible difficulties.

The most important philosopher who advocates a world republic is Immanuel Kant in his *Perpetual Peace*. Communitarianism is a modern objector to the idea of a world republic, arguing that a world republic would threaten the political and social integrity of the single states. Cosmopolitans such as Charles Beitz (1979), however, consider the single nation-states as a historical steppingstone, and advocate a homogenous world republic that eventually will incorporate and thus dissolve the single states.

An apt justification starts out with a diagnosis of globalization (1), since everybody talks about globalization today; what the term actually refers to, anyway, often remains unclear. It follows – as a therapy – the world republic (2), which needs to be defended against some misconceptions. Finally, we will come through the discussion of eight objections (3), to the following insight: "Créer un ordre juridique mondiale, c'est avant tout fonder la pente vers la paix et la justice globale" ("To establish a world legal order means, first and foremost, to create the longing for global peace and global justice").

Diagnosis: Globalization

Strangely, it is still common to picture globalization in a reductionist economic perspective; one only thinks about economic and financial markets, at most also about labor markets, tourism, and environmental protection. The truth is, however, that the sciences, the school and university system, and all facets of culture globalize as well. One should not forget that long before the first computer was built the writings of Plato, Aristotle, and Kant were distributed over the world, also the works of Homer, Dante, Shakespeare, and Goethe, the tales from *One Thousand and One Nights* and texts of eastern wisdom. Likewise, all around the world people listen to Bach, Beethoven, and Mozart, Jazz, and certain pop music. All around the world people admire Indian temples, Japanese Zen-gardens, and European castles and churches. All around the world people study quantum theory, mathematical analysis, and the building blocks of life. And long before all these various products of arts and sciences, various religions have spread all over the world.

Against the reductionist economic perspective one might also point to the noneconomic causes of the economic globalization: political decisions and technical innovations. Even the economic globalization itself is not something completely new. Already at school we have heard about international trade routes like the Silk Road, the emerging world trade zone with world market prices and world trading centers like Alexandria in Hellenistic times, and a dense trade system in the times of gold currency.

New are, however, elements that are based on innovations like broadcast satellites and the Internet: that information can be exchanged within seconds and events can be perceived almost simultaneously worldwide. The Internet also bears a huge democratic potential. It does not entail equal distribution of usage, but nevertheless treats all persons, enterprises, and states equally. Furthermore, its information bypasses the censorship of autocratic states, exerts pressure for democratization, and facilitates worldwide protest against violations of human rights. Moreover, there is an economic benefit: a person who travels the Internet reduces his or her environmental footprint and increases his or her legal security, because at least his or her life and limb are not at risk.

The Internet, however, is also of use for global terrorism and organized crime. Since there are also innovations in the sector of military technology, the first dimension of globalization, the multifaceted 'community of cooperation,' couples with a second dimension which is, unfortunately, again rich of facets, the 'global community of violence.' And in the burdens that follow from international competition, we recognize the third dimension, the global community of solidarity, the worldwide 'community of misery and suffering.' By now the large movements of refugees and the high amount of other

migratory movements make it apparent that neither the consequences of civil wars nor of economic as well as legal and political underdevelopment stop at national borders.

Therapy: World Republic

Since all three dimensions demand for global action, this observation is linked to the question of how to respond. Let us approach this question topically and draw on some widespread assumptions.

Humanity knows two basic models of visionary power for laying out the structure of communal life, liberal democracy, and a contest of powers. The first model says that (1) we should use law instead of violence; (2) law should be based on human rights; which in turn demands that the (3) establishment of public institutions (4) by the people who are governed by these institutions. Moreover, (5) it is called for commitment in civic society. These five elements – law, human rights, statehood, more democracy, and civic society – are categorical and universal. They are elements of a morality of law. Our human desire for right and justice is combined with the call of a duty of the morality of right.

A liberal democracy consisting of these five elements of course considers a strong 'statism,' which tries to solve all tasks by coercive intervention, to be foolish. Instead, such a democracy leaves room for the free play of powers, competition in economic, but also political, academic, and cultural spheres. Because of the creativity, efforts, and willingness to take on risks that is released in the market, one expects an age-old dream of humanity to become true. According to the prophetic word, "their swords will be turned into plough-blades, and their spears into vine-knives" (Jesaja 2,4), military force shall be turned into economic and cultural power and there shall be peace combined with scientific development and economic and cultural prosperity.

Because of their universal scope the five elements mentioned not only bind particular political communities, but also the international relations. Hence, the idea of a worldwide peace and legal order emerges in which societies and in particular individuals prosper through economic, academic, and cultural competition.

Political philosophy acknowledges also the second vision, the wealth promising competition. Political philosophy, however, rejects its absolutization, the neoliberal replacement of politics by markets. There is a lot that the world society might leave to free competition and contingent evolution. Anyway, the framework of this competition needs to set likewise in a particular state and the world society. It is called for a world legal order, that is bound to human rights and democracy – not only but also because interstate violence, the wildfire that flares up every once in a while, can only be erased that way. The price for globalization shall not be paid with a regression in liberal democracy.

One might ask: 'Why not?' The response is evident, however. As far as the pragmatic reason, the promised multifaceted wealth, speaks in favor of globalization liberal democracy is, because of its being an element of the morality of right, of higher value and cannot be sacrificed to wealth which is a lower ranked value. As far as globalization is legitimated functionally by promoting freedom and democracy – in other words, if globalization is supposed to serve democracy – it would be irrational to sacrifice democracy for democracy. Perhaps such a proceeding is dishonest: one talks about democracy and actually means one's own power, either like some neoliberal economists the supremacy of economic and financial markets or like politicians of the contemporary hegemonic power the supremacy of their own country.

If one gives priority, on the contrary, to liberal democracy, the spreading of its basic elements over the global sphere is inevitable. In the same way that a particular society coerces the economic and cultural market forces to follow the framework of human rights and democracy, global demands for action call for a global legal order of a democratic layout, that is to say a liberal world democracy, including a global civic society, in a nutshell: a world republic. (For a detailed design of the world republic and a profound substantiation: (Höffe, 1999/2007); for a discussion of this approach: (Gosepath and Merle, 2002); cf as well Lutz-Bachmann and Bohrman, 2002.)

Eight Objections

Though the global call for action in combination with law and democracy urges the establishment of a world republic, such a world republic breaks with familiar ideas. As a consequence, skepticism spreads. We can counter this skepticism with a new method: a modest form of Hegel's determinate negation. By way of the discussion of eight objections the vision world republic reveals her modest, hence, realistic profile.

Priority of Statehood?

According to a first objection, statehood is given a priority over its alternatives – market, 'governance without government,' and a network of global civic constitutions – that are denied the power and capability to coordinate.

The response to this objection is based on a modest understanding of statehood: for the very reason that what law means, that is the impartial enforcement of common rules, does not remain a pious hope, the enforcement should neither be entrusted to one of the parties affected nor to contingent evolution. On the contrary, nonprivate, that is to say public powers, have to take on responsibility without privileging the existing structure of public power. Statehood means nothing more and nothing less than the notion of a nonprivate responsibility. A responsibility for shaping legal norms, for enforcing them, and to settle disputes. Moreover, I do not picture global statehood as statist, but as state centered. Since the global civic society, the cosmopolitan society is a part of the world republic, nongovernmental organizations need to be given possibilities for cogoverning in certain limits.

Two Models

According to the second objection, the vision of a world republic asks too much of humanity. Though there have been tendencies toward globalization for a long time, humanity never established a global order. Why should we be able to do what mankind has not achieved in the past?

The response to this question is twofold. On the one hand, the global call for action increased enormously. To fight environmental damage, terrorism, and organized crime, as well as poverty, underdevelopment, and oppression global agreements are indispensable – in any case much more important than the spread of philosophy and science.

On the other hand, there are at the minimum two examples of a successful supranational law. Against the danger of cultural bias, the two examples are taken from two very different legal cultures. The first example stems from outside of the Western culture.

It is the Iroquois Peace Coalition and League of Nations that was founded presumably in the sixteenth century (see Morgan, 1851; Tooker, 1978). The second example is an ancient European model, the *ius gentium*, developed in the Roman Empire, which in that time was a highly complex multicultural community.

The idea of a worldwide, cosmopolitan order is at least as old as the cynic Diogenes of Sinope. When he, driven from his homeland, calls himself a 'citizen of the world' (cf Diogenes Laertius, VI 63), he has no legal institutions in mind. The same is true with the second founder of the Stoa, Chrysippus; his cosmopolis does not yet appear as the legally ordered community, but as an open space of action, not qualified further (*Stoicorum veterum fragmenta* II 32; 328). The Roman *ius gentium*, on the other hand is actually law – and this is our first factor. The same applies to the peace coalitions of the Iroquois. The desire for peaceful coexistence might be as old as humanity itself. Only in a ceremonially sealed coalition, however, does it maintain its legal bindingness.

This does not originate from fulfilled wishes, but – secondly – from an urgent call for action. The Iroquois had been tired of war, and in Rome, one had to regulate the increasing trade traffic between citizens of various communities and cultures. Therefore, the new law is subsidiary in the original sense: it helps in an emergency, namely, whenever the previously competent institutions no longer suffice. Also, it does not extend 'imperialistically' to all legal areas, but acts with a wiser modesty. The Iroquois limit themselves to the international domain, as their alliance belongs to an international law. In Rome it has to do with the then-urgent call for action regarding business and trade law, while constitutional law and administrative policy, as well as religious law and case law are excluded (cf Kaser, 1993). The Roman *ius gentium* is primarily an international commercial law, an *ius gentium commerciale*.

Because the other legal domains remain untouched, the new law seeks a completion of what is already known, not its replacement. It is, thirdly, complementary. Rome might have been able to exercise its hegemonic power more freely, though perhaps at the price of loosing a good portion of its trade traffic. Whether out of legal and moral, or, more probably, out of pragmatic reasons, a once again modest Rome returns, fourthly, to a consensus, even to a double one, that is, to a 'theoretical' and to a lived, 'practical' consensus. Instead of imposing a law upon others, one appeals, on the one hand, to the agreement of all cultures that is indicated in the expression *gentium*: it makes use of an interculturally valid ground of obligation. For example, the obligations (including sales, rent, management, etc.) are grounded upon the requirement to keep one's word, the *fides* that bind all human beings equally, independently of their ethnic and religious affiliation, as well as of their citizenship. On the other hand, one seizes the trade and traffic requirements that were customary in Rome's international market. Because there was no apparent hegemonic power among the Iroquois, they proceeded consensually as well.

The fifth factor is already implicit in the fourth factor. It lies in the renunciation of particularities, in the indifferent vis-à-vis difference: the Iroquois oppose their differences, while Rome is as good as indifferent to all cultural differences.

One characteristic of Rome deserves its own mention. The Roman *ius civile* contains religious formal requirements to which the concerned citizenry is bound, but which are unnecessarily complicated. The *ius gentium*, however, separates itself from religious obligations. It features, sixthly, a secularization (though this is missing with the Iroquois) that is not only connected with simplification, but also, seventhly, with a concentration on the heart of the legal transaction: on the wills of the parties concerned.

The parties concerned count as equal: this is the eighth success factor. Whether rich or poor, whether Egyptian, Roman, or Jew, whether pious or irreligious, the *ius gentium* treats everyone equally, though, of course, it does not yet cover every legal domain. The legal equality primarily reigns in contract law, though not in family law or law of inheritance. While princes still rule in Europe 'by the grace of God,' the Iroquois, long before the first declaration of human rights, the *Virginia Bill of Rights*, add, 'Equal, without the superiority of the chiefs.'

Because all of these factors come together – aside from (1) the character of law, there is also (2) subsidiarity and (3) complementarity, (4) the search for both a theoretical and a practical consensus, (5) neutrality toward cultural particularities, (6) secularization, (7) concentration on the legal transaction itself, and (8) the (extensive) equal rights of all concerned – the *ius gentium* is the model for the world in the sense that it is compatible with the world, that is, capable of governing in an agreeable way the coexistence of cultures from the entire world.

A further, ninth factor opposes 'globalist' parochialism. Notably, it is not an international law, but a Roman authority that supports the *ius gentium* jurisdiction of the foreign praetor. This reveals itself as an apparently paradoxical law, a world law that can neither establish a world organization nor can be established by it: a national world law.

Hence, the second objection can be answered not only by reference to the increased necessity for action but also by facts on the ground.

Ungovernability?

A third objection is: the world republic is such a giant project that it will remain ungovernable because of its size and complexity. The answer to this objection increases the modesty in the idea of a world republic.

For Liechtenstein (with its 28.500 inhabitants), Switzerland (with eight million inhabitants) is a giant and the United States (with 340 millions inhabitants) a monstrosity. If, however, a society like the United States can be governed, though it is 10 000 times as big as Liechtenstein and still 40 times as big as Switzerland, the third objection cannot be a knockdown

argument. We cannot regard the third objection as an absolute veto, but only as a relative veto, nevertheless as a constructive one: the vision of a world republic remains permissible, even necessary – under the premise that it is able to prevent ungovernability as well as overcompensation, excessive bureaucratization, or a police state. For this overcompensation is the vision that frightens most people, not a liberal world republic.

For spelling out how the detailed set up of the world republic is to be like, political imagination and experience are needed. Both already suggest two things. On the one hand, the world republic should not be understood in too statist terms, since a global civic society is one of its essential building blocks. On the other hand, it seems to be recommended to slide in political units of continental and subcontinental size. Following the model of the European Union, most problems could be solved 'at home' and only problems that call for global action would need to be handed over to the global order. These problems, however, range from fighting terrorism and organized crime – when it crosses the borders of the smaller unions – over the establishment of a global peace order that guarantees peace within and between the unions, and social and ecological minimum criteria, perhaps also a world competition regulatory body and a global supervision of banks, to the establishment of world courts, like the International Criminal Court. The response to the third objection is anyway: world civic society plus larger regional intermediate units.

A Threat to Human Rights?

According to a fourth objection, a world state puts at risk the great achievements of civilization, namely, human rights and civil rights. For up to now, only the individual state has succeeded in guaranteeing such rights. This argument is correct not only in its underlying normative prerequisite – the obligation of every state to uphold human rights and civil rights – but also in its empirical assertion. However, only one-third of this claim is true.

Without a doubt, in the Western world, it is predominantly states that protect human rights and civil rights (although in the case of Europe the analysis needs to be qualified so as to include the European Convention on Human Rights, the European Court of Human Rights, and the European Commission on Human Rights). All citizens who obtain such protection only from international organizations tend to fare embarrassingly badly. However, the second third of the truth reminds us that Western states themselves have actually threatened these rights at some point in the past: France pursued the Huguenots; the United States is founded by religious refugees who escaped the intolerance prevailing in Great Britain at the time, and the United States itself permitted slavery well into the middle nineteenth century. The last third of the truth consists of the principle of subsidiarity: where human rights and civil rights are already protected – either state-internally or through human rights conventions established by larger regional units – the world republic can and should stay in the background. Yet, in case of massive human rights violations or genocide, such restraints are only warranted if the ('humanitarian') intervention creates even more evil than before; a general restraint, however, is not justifiable (see Höffe, 1999/2007: ch. 14.3).

Hence, the fourth objection, once more, does not have the force of an absolute veto, but only that of a constructive veto. It confirms the principles of federalism and subsidiarity: the primary protection of the law remains the responsibility of the individual states, which is why they deserve the rank of primary states. Such status entails an original (not derived) authority to legislate as well as an original responsibility for the finances. The world republic is only a complementary state, a secondary state, and, in cases where a regional intermediate level is introduced also, more modestly still a tertiary state. At any rate, the world republic does not constitute a single, unitary state that extends across the entire globe, but a state of states: a state of multiple nations.

According to the principle of global subsidiarity, the world republic is not legitimized from above, but democratically through the citizens and individual states. Yet, it assumes only those remaining responsibilities that the primary republics and larger regional units alone cannot master at all or not nearly as well. In the first case, the world republic has a compulsory responsibility; in the second case, it has an optional responsibility. Added to that, the 'principle of care and circumspection' applies: they must not endanger the degree of qualified democracy that individual states and larger regions have already attained.

States as well as larger regional unions that are committed to liberal democracy possess legitimacy, which implies no obligation to dissolve. In some places, one likes to declare the nation state as outdated. The truth is, however, that despite its perversions it became the worldwide dominant model because of its important accomplishments. For instance, religious and economic freedoms of individuals follow from its separation of state and society. Economy and the material wealth created by it as well as modern administration is, in turn, hardly conceivable without the extent of social integration that we find in nation-states. The renewal of the European education and science, the increase of the educational level of all citizens, and the establishment of a welfare state are also accomplishments rooted in the nation-state. Furthermore, the state admits the thought of basic and human rights, advances the sovereignty of the people and establishes a parliament. Moreover, it abolishes serfdom, feudal privileges, and the legal inequality of women. Finally, the supplement, the civic society grows – sometimes in opposition to the established governmental institutions – first and foremost within states and still helps these to fulfill many tasks and be at the same time more citizen-oriented and efficient.

For reasons like that philosophy gives the world republic a subsidiary and complementary rank. And because of the so-defined rank of the world republic, philosophy is not dreaming a sweet dream that modifies the reality of the world according to its needs. The world republic does not promise, anyway, the all-encompassing well-being which religions call 'salvation' and rightly do not expect for this world. The world republic that is called for is an ideal, and humanity is morally obliged to realize it. Fortunately, humanity already brought its establishment on its way to: the mere coexistence of states and their working against each other has been replaced by a dense net of economic, academic, and cultural, even political, and sometimes even ecological cooperation.

To a large part, all this still amounts to a political 'too little,' because an executive power is missing. And the singular, the single ultraminimal state, has given way to a plural: a regional as well as thematically ('sectoral defined legal regimes') complex multiplicity. Moreover, a common state power is not emerging yet. It might even remain an open question whether the world republic needs the establishment of a unified head and a monolithic center. A part of the agreements mentioned starts to take over legislative tasks by means of binding statements in such a way that one might speak of the beginnings of and the equivalents to legislation, a 'soft legislation.' In cases where international inspections and sanctions are planned, executive power emerges, a 'soft executive power,' which carries the cutback of national sovereign rights further forward. It is extended, in cases where additional international arbitral authorities or even courts of justice, a 'soft jurisdiction,' are established. If we take all these elements together, we already see the first beginnings of a 'soft world republic.'

Global Peace by Means of Global Democratization?

According to the fifth objection, there is a more simple way of achieving the protection of human rights: democratization of all states. In line with the thesis 'global peace by means of global democratization' one could be content with a global democratization policy. The world republic would be dispensable, according to this reasoning.

Undeniably, liberal democracy protects human rights already domestically. It seems to be appropriate, however, that a global human rights commission is established that checks this protection like the European Commission of Human Rights especially in cases where the protection of human rights did not already become natural because of a long practice of human rights standards. Most notably, the states themselves need protection: of their territorial integrity and their political and cultural self-determination.

On the danger of a war of aggression political science did take on Kant's famous thesis: republics, meaning pretty much liberal democracies, do not tend to engage in wars of aggression (*Perpetual Peace*, 1. Definitive Article, VIII 351). History, however, gives us reason for being skeptical: the young French Republic inflicted war on Europe and pursued a clear interest to rule. The United States expanded to the West without taking the rights of the native inhabitants into account. Furthermore, they annexed Texas and incorporated Arizona, Nevada, Utah, California, and New Mexico after their war with Mexico. Nor did Great Britain's development toward a republic hinder it from having plans to become a world power, the expansion of the Commonwealth.

One of the reasons for this is that Kant's argument, the well-considered self-interest, is correct, but has not the wished momentousness, since it is not *always* an argument against war. Wars that happen far away are not as important to the citizens and wars against apparently weaker enemies are even less important to them. Furthermore, wars may distract the public from domestic problems. Moreover, there is money in wars. Finally, there are legal problems below the war threshold. Consequently, the fifth objection has again only the power of a constructive veto: the legal protection and the protection of peace that is already brought about by the worldwide democratization remains in its hands. However, states also have a claim to a legal solution of conflicts like individuals as opposed to a solution by power. Hence, a worldwide legal order with public powers is called for.

Deficit in a Global Common Sense of Justice?

According to the sixth objection, a world legal order presupposes something that is in fact missing: a globally common sense of justice.

It is apparent that even the West is lacking a common sense of justice. The amount that is paid for indemnity in civil lawsuits, for instance, is much higher in the United States than in Europe and there are significant differences concerning the question what is to be admissible as evidence in criminal trials. Even stronger differences become evident in the attitudes toward death penalty, corporal punishment, and the treatment of dissidents. However, besides these apparent differences, there are in fact substantial similarities: equality and impartiality are globally recognized as essential in the application of law. The same is true of procedural rules like 'one shall listen to the other party as well' or the presumption of innocence. Furthermore, the same basic goods are legally protected by almost all legal orders: life and limb, property and honor. Moreover, it is neither allowed to manipulate weights and measures nor to falsify documents. Additionally, the environment is protected, for instance, by the prohibition of poisoning wells and the institution of protective forests. On top of that, the human rights treaties of the United Nations prove many more similarities. It is 'only' lacking the commitment to enforce these similarities impartially and effectively. The constructive veto in this case is, therefore, almost mundane: the global sense of justice needs some more time to be given complete expression. The existing similarities, anyway, are already big enough to make the establishment of world courts possible: the International Court of Justice, the International Tribunal for the Law of the Sea, and the International Criminal Court – despite the veto of the United States.

The world courts, again, are governed by the principle of subsidiarity. They do not replace the national jurisdiction. They simply extend it to a new level, via the jurisdiction of a greater region like the European Court of Justice to a global jurisdiction. Apart from that the 'national' level needs to change also and it is called for the establishment of a 'national world criminal law,' which is completely compatible with human rights standards.

Global Public

An indispensable institution of a functioning community is a political public (see Höffe, 1999/2007: ch. 4.3), a requirement that gives rise to problems that are well known at the level of large-scale regional unions such as the European Union. These difficulties are exacerbated even further in the case of a world state, a fact that allows us to forward another, seventh, objection: a world state without a world public, a global public, is unreasonable.

Fortunately, attempts toward a global political public can already be attested. They are found, for instance, in the media reports on violations of rights all over the world, which induces

reactions by the 'general public' virtually everywhere. Kant was able to state for his own time that "a violation of rights in one place is felt throughout the world" (Perpetual Peace, 3rd definitive article: VIII 60). Yet, he pays attention neither to the danger of a selective, biased perception of rights violations, nor to the fact that opposing voices can also be heard: collective prejudices, national superciliousness and even appeals to break the law and violate peace. Moreover, an effective global public has far more tasks to perform than merely function as a medium for the outrage of rights violations.

According to three stages of the principle of the 'public,' global politics is, firstly, to happen in a public manner so that all citizens are able to engage in it. Secondly, every citizen is to be eligible to vote and is to be able to stand for office at the ultimately decisive institution, the legislature. Thirdly, the citizenry is to be no temporary electorate, possibly complemented by permanent public opinion polling; rather, politics is to be exercised by the entire people, even if it is done so professionally only by a few and limited to the parliament as the House of Representatives. In a participatory democracy, the general public is a critical institution the entire political sphere has to answer to – not only the government, but also the parliament, the judiciary and the constitutional court.

Let us focus on legislature as an example: similar to the individual state, the world state, too, is in need of a variety of controversies that initiate, prepare and accompany parliamentary debates, and later trigger further amendments if so required. In order to ensure that these discussions actually take place, numerous transnational as well as supranational institutions are required. Without a doubt, a great many international nongovernmental organizations already exist, some of which even have considerable influence. They are in a position to convince (in democracies even to coerce) governments, corporations, and other institutions to enter into a dialogue with the general public and to accept minimum requirements of both a formal and substantial nature. At times, they even manage to organize civil protest across borders. Nevertheless, most of these institutions – the media; political, cultural, and ecclesiastical academies; grass-roots movements; parties and associations – are, so far, still organized 'nationally,' within the confines of a particular individual state. At global level, by contrast, a considerable deficit has to be attested.

In the process of overcoming this deficit and by way of establishing an effective global public, it is important that all the various interests, voices, and languages command roughly the same level of attention. The legal requirements are generally known and uncontested: a global public needs the freedom of the press, the freedom of speech, and the freedom of assembly. These negative freedoms are to be complemented by a positive freedom (of the press), for instance in the form of a press antitrust law. Such measures facilitate the creation of a multitude of competing media, political academies, and grass-roots initiatives. A ban on monopolies, oligopolies, and cartels in other spheres is justifiable only on competitive grounds. In the sphere of the political public, by contrast, additional democracy-promoting reasons can be found: the positive freedom of the press is indispensable for a functioning participatory democracy. A sufficient level of educational attainment is, for the same reason, also required, so as to establish at least a minimum level of literacy among citizens. For cultures that do not rely on written texts, alphabetization is no legal-moral precept per se, but in a world society in which the media play such an influential role it becomes an inevitable constraint.

Yet, even when the negative and positive freedom of the press is secured, the various interests do not even remotely command the same level of public attention. Of the host of dangers that exist I mention only one: the hegemony of a particular language. This follows because a living language is not merely an interest-neutral means of transport. The media of the hegemonic language find a far greater resonance, so that the pictures, symbols, myths, and interests of the accompanying cultural space occupy a privileged position. What is more, cultures that are based on written texts receive much greater weight compared to those of oral provenance.

In individual states, the public is an institutionally varied, if not vague phenomenon: it occurs on different levels, possesses no sharp boundaries, and covers all shades of expertise and impact, from professional academic discourse at one end of the spectrum to the tabloid press at the other. Since the public uses different media and institutions to address and make sense of different subject matter, it exists not in the singular but in the plural, as a colorful bouquet of overlapping publics. Many of them have an aristocratic tendency, in that professional politicians as well as particular media and associations command an overwhelming influence.

World society is, by virtue of its greater size and heterogeneity, greatly marked by these phenomena. The global public is, nevertheless, relatively well developed particularly in areas such as human rights. For there is a global discourse on human rights that is conducted by a heterogeneous intellectual 'community' comprising individual scientists, philosophers, media commentators, human rights organizations, professional associations, the media, and public opinion in important states, as well as relevant conferences and publications. Of similar great significance is the interaction between the high courts. Even states themselves must not to be neglected, because within the United Nations and other international organizations they are the entities that ratify human rights conventions and, in so doing, enable the legislative evolution of human rights on a global level.

No matter how optimistic or pessimistic one's judgment is with regard to the development of a global public, the seventh constructive veto suggests itself in the form of a dependency: as long as there is no effective global public, it is irresponsible to establish a subsidiary and federal world republic.

A global public is indispensable for yet another task: the development of a common political identity that encompasses a consciousness of world citizenship and a critical world memory (see Höffe, 1999/2007: ch. 12.2). The consciousness of world citizenship, in turn, is a prerequisite for global plebiscites. Provided that the notion of worldwide plebiscites and referenda is generally persuasive, it is only defensible if there exists a consciousness of world citizenship as well as a global political public within which the relevant issues are discussed in a thorough and fair manner. Here, too, a threat emerges that requires countermeasures: in order to prevent that the decisions of the six to eight most populous states crush the interests of all others, the example of Switzerland may be followed and a dual majority be demanded: global

plebiscites require a majority by both the states and the world population.

The Danger of Leveling

According to the eighth and final objection that I will discuss here, in times of globalization we face the threat of a leveling that we need to counter with strengthening our particularities, so that the cultural richness of our world and the identity of each individual which is bound to it is maintained.

It is true that many national states and also greater regions have their common history, a particular tradition, culture, and language – or a defined multilingualism – and follow certain shared understandings of how a good communal life is to be organized. States do in fact have a right to that, a right to national particularities and difference. This right substantiates the principle of subsidiarity in the world republic and contradicts the (hyper-)globalist proposal to replace nation-states by a world republic following the model of ancient Rome which downgraded states to provinces of a homogene world empire.

The following constructive veto agrees with the objection, but again only with one third of it: people do have a right to collective particularities like history, tradition, religion, language, culture, and common understandings of the good. Since their plurality increases the wealth of humanity and since the identity of each and everyone depends on them, there exists an interest to take advantage of this right and to counter the danger of leveling that is, for instance, brought about by a culture that acts imperialistic.

According to the second third, the existing states are, however, not an end in themselves. They are units that exist for the sake of human beings. Therefore, they can be combined newly and dissolved for their sake. And according to the last third, neither states nor foreign citizens are released from the commitment to the maxim of the rule of law and democracy. For the purpose of strengthening the relevant principle here, federalism, and countering the danger of a democratic deficit it is called for a dual strategy: on the one hand, one should not unnecessarily hand over responsibilities. The burden of proof lies with the one who demands handing them over. On the other hand, the new authorities and institutions need to be tied to the lower levels. In particular, they need to be accountable to them.

For the legitimation of a democratic world order, there are three strategies at hand. According to the exclusive legitimation through citizens, the world state is established through the will of a global citizens body, which includes the entire world population. Since this strategy appeals to individuals as the agents to whom justification is owed, one might think this strategy to be appropriate. However, against this line of reasoning speaks the right to form and uphold an individual state combined with the fact that the interests of a group, also the ones of a state, are not reducible to the sum of the interests of its members.

Since the nation-states represent the interests of each and every single citizen as well as the interests of the citizenry as a whole, one might argue in favor of an exclusive legitimation through the states. Against this line of reasoning speaks, however, that there are 'cross border' affiliations like religion, language, and profession, sophisticated hobbies and political and social interests that are concentrated by organizations like Amnesty International or Doctors Without Borders. To give an example: one might be German speaking, yet Muslim and Swiss, anyhow chess player and mountaineer.

Consequently, a third combined strategy seems to be apt. According to this strategy, the world republic is legitimized by the citizens as well as the states, and all power of the world state stems from a twofold citizenry: the society of all humans and of all states. This dual strategy needs to be embodied in the organizations of the world republic. In the manner of the draft to the Constitutional Treaty of the European Union, one might call it a 'union of states and peoples.' The highest institution, the world legislator as a world parliament, is to consist of two houses: a World Congress, the representation of the citizens, and a World Senate, the representation of the states. Their exact constitution and the difficulties that may rise from the establishment of greater regions do not need to be of concern for us now. That Liechtenstein is to be given less weight than India or China is evident, but which weight they are given exactly is a matter of political decision making. May be the radical of the inhabitants might be a solution: for one specified unit of inhabitants one is given one vote, for the quadruple number of inhabitants two, for the ninefold three votes and so on.

Conclusion

It is a universal imperative of right, which speaks in favor of a world republic: conflicts between humans shall be solved not by private opinion or violence, but only by law and its protection though the state. In a world republic we are world citizens, cosmopolitans, but not in the exclusive sense of the term as Hegel criticized it (*Elements of the Philosophy of Right*, section 209) that is hostile to nation-states. With a feeling of moral superiority the exclusive cosmopolitan says: I am not German, French, or Polish, but merely a world citizen, a cosmopolitan. In this case the cosmopolitan right replaces the 'national' civil right; one is a cosmopolitan *instead* of being a citizen of an individual state. The complementary world republic steps back from this either/or rhetoric of 'either nation state or cosmopolitan world order.' The civil right of the world republic does not dissolve the national civil right, but supplements it. Furthermore, it adds regional subunits.

There is going to be a new and hitherto unknown multiple citizenship. The longing for this combined with the longing for global peace and global justice might inspire humanity to realize the vision of the world republic: whether one is primarily German, French, or Italian and a citizen of Europe only secondarily will be decided by the democracies of Europe within the next decades. Primarily, however, one is one of the two, secondarily, the other and tertiary, one is a world citizen: a citizen of the subsidiary and federal world republic.

See also: Contractarianism; Democracy: Normative Theory; Equality: Philosophical Aspects; Freedom: Political; Justice: Philosophical Aspects; Public Reason.

Bibliography

Beitz, C., 1979. Political Theory and International Relations. Princeton University Press, Princeton, NJ.

Doyle, M., 1983. Kant, liberal legacies, and foreign affairs. Philosophy and Public Affairs 12, 205–235, 323–353.

Gosepath, S., Merle, J.-C. (Eds.), 2002. Weltrepublik. Globalisierung und Demokratie. München.

Hegel, G.W.F., 1821/1991. Elements of the Philosophy of Right (H.B. Nisbet, trans.). Cambridge/New York (Orig.: Grundlinien der Philosophie des Rechts. In: Werke in 20 Bänden, vol. VII. Suhrkamp, Frankfurt, M).

Höffe, O. (Ed.), 1995. Immanuel Kant. Zum ewigen Frieden. Akademie Verlag, Berlin.

Höffe, O., 1996. Vernunft und Recht. Bausteine zu einem interkulturellen Rechtsdiskurs. Frankfurt, M.

Höffe, O., 1999/2007. Democracy in an Age of Globalization (D. Haubrich and M. Ludwig, trans.). Dordrecht (Orig.: Demokratie im Zeitalter der Globalisierung. München (1999)).

Kant, I., 1795/1991. Perpetual peace. A philosophical sketch (H.B. Nisbet, trans.). In: Reiss, H.S. (Ed.), Political Writings. Cambridge/New York, pp. 93–130 (Orig.: Zum ewigen Frieden. Ein philosophischer Entwurf. In: Gesammelte Schriften (Ed.), Königlich Preußische Akademie der Wissenschaften, vol. VIII. Berlin. pp. 341–386).

Kaser, M., 1993. Ius Gentium. Köln.

Lutz-Bachmann, M., Bohrmann, J. (Eds.), 2002. Weltstaat oder Staatenwelt? Für und wider die Idee einer Weltrepublik. Frankfurt, M.

Morgan, L., 1851. League Oft He Ho-dé-no-sau-nee or Iroquis. Rochester.

Russett, B., 1992. Grasping the Democratic Peace. Princeton University Press, Princeton, NJ.

Saint-Exupéry, A. de, 1959. Citadelle. Paris.

Tooker, E., 1978. The League of the Iroquis. Its history, politics, and ritual. In: Trigger, B.G. (Ed.), Handbook of North American Indians, vol. 15. Smithsonian Institution, Washington, DC, Northeast, pp. 418–441.

World Systems Theory

Daniel Chirot, University of Washington, Seattle, WA, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

World systems theory (WST) was created by Immanuel Wallerstein in the 1970s, though earlier Marxist theories first proposed similar ideas. WST claims that rich core capitalist societies succeed by exploiting poorer peripheral ones. In between are semiperipheral societies, a precarious global middle class. The periphery therefore can only advance through global revolution that will end the world capitalist system. WST's holistic approach that studies societies in the context of the entire world rather than in isolation is valuable. If freed of its Marxist extremism and questionable economic assumptions that have marginalized it, this approach can help us better understand the nature of social change.

Introduction

The essence of World systems theory (WST) is that the modern world, dominated by capitalism, has since the sixteenth century increasingly become an integrated whole in which single countries cannot be usefully analyzed without placing them within the context of the entire global economic and political system. Though its intellectual antecedents date back to the late nineteenth and early twentieth centuries, WST's contemporary version was pioneered by Immanuel Wallerstein (1974, 1979, 2011) whose goal it has been to overcome narrow disciplinary specialization in the social sciences. In his many works, Wallerstein has sought to explain unequal development and wealth between societies in the modern capitalist world and to understand the cyclical patterns of expansion and contraction that characterize this world system. WST stresses that poorer, or 'peripheral' societies have been systematically underdeveloped by the rich and powerful 'core,' and they stand little chance of prospering without world revolution. Between peripheral and core countries are 'semiperipheral' ones that act as a kind of global middle class stabilizing the world capitalist system, but because semiperipheral countries are subject to the greatest systemic strains, that is where revolutions are most likely.

Wallerstein's vision is strongly Marxist and emphasizes the exploitative nature of capitalism; but instead of relying only on the internal class structures of individual societies, he has theorized that societies as a whole fall into broad categories that mirror the more classical Marxist classes. So, for him, the 'ruling class' in the world capitalist system consists of the ruling classes in the most powerful core countries, and those countries compete among themselves for hegemony. Poorer societies, therefore, are not that way because they are 'underdeveloped' so much as because they are kept poor by the exploitative world capitalist system's core economies. So WST attempts to debunk what is called 'modernization theory,' which claimed that all societies can follow a similar path toward greater economic prosperity and modernity as the one followed earlier by the more advanced Western societies.

Wallerstein has further posited that a world capitalist system can only function smoothly if there is one dominant core power, and that in times when there is conflict between potential hegemonic powers, the system is unstable. This is all the more the case because capitalism is subject to major periodic economic crises, just as Marxists have always claimed.

Because of its ideological radicalism, WST was very appealing when it appeared in the 1970s at a time when American and many other social scientists around the world were trying to understand America's war in Vietnam and the seemingly inexorable rise of world communism. Since then however, WST has been marginalized, especially since the collapse of the Soviet Union and the dramatic changes that have transformed China while bringing it solidly into this world capitalist system. The communist Soviet Union and China (as well as, for a time, North Korea) were the societies Wallerstein looked to as models for the future world socialist system that was supposed to be fairer and more egalitarian than the ruling world capitalist system. He also saw what used to be called 'third worldism,' that is, the rise of radically socialist former colonies and their alliance in an anticore, revolutionary bloc as something that would lead to the collapse of the world capitalist system. Of course, that is now all history, as it has hardly come to pass.

It has to be said that there are very few if any major economists or any serious political scientists who would accept WST's basic premises about how the world economy functions or where it is headed. Despite the obvious return of a major global economic crisis since 2008, only a very small minority of analysts believe that today's economic troubles are likely to lead to worldwide socialism or even to the partial revival of communist nonmarket command economies.

Nevertheless, if WST can be divorced from its ideological bias it has the potential to become a valuable tool for understanding social and political changes. Its emphasis on holistic global analysis and its awareness of the dangers inherent in an unbalanced capitalist world system that lacks a stabilizing hegemonic power have as much relevance as ever. If anything, the world is even more globalized than in the past. Indeed, many of WST's assumptions have been absorbed into work by some sociologists, geographers, ethnographers, and historians. Without resorting to crude Marxist doctrines, this kind of analysis could even offer some interesting ideas about the world economic situation.

The Intellectual History of the Theory

The notion that capitalist growth and therefore successful industrialization can only occur through the ruthless exploitation of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.32172-9

an impoverished working class was basic for Karl Marx, but he concentrated primarily on the internal workings of class conflict within rather than between different societies. Nevertheless, Marx also observed that capitalist expansion was a global phenomenon that had a particularly unsettling effect on colonial or semicolonial societies. This became an important aspect of later Marxist theories of international relations. The works of Rosa Luxemburg and V.I. Lenin as well as those of many non-Marxist critics of late-nineteenth and early twentieth century European imperialism stressed the internationally exploitative nature of capitalism and its seeming need to control the ever growing markets and resources. Lenin's theory of *Imperialism* (1939), written in 1916, claimed that World War I was the logical outgrowth of the competition between capitalist powers for colonies, because without the surplus extracted from dependencies, the internal contradictions of capitalism would produce growing worker immiseration and eventually, revolutions. The consequence of this theory, if it were to be believed, would be that revolutions outside the core capitalist societies can deprive the core of easily exploitable human and material resources, and therefore lead to worldwide revolution. This belief remains at the heart of Wallerstein's ideas, except that the direction of class conflict for him is determined by the global struggle between exploiting and exploited societies. For Wallerstein, no revolutionary movement within a single society, or even several of them, can succeed in bringing down capitalism unless it sparks a global revolution.

There is a great irony, therefore, in the fact that WST's other intellectual parent was early twentieth-century right wing nationalism. Joseph Love's work (1996) traces the development of nationalist economic theories of underdevelopment from the German Friedrich List in the nineteenth century, through the Romanian fascist–corporatist Mihail Manoilescu, and to the Latin American right in the 1930s. The relative backwardness of these societies (they were what Wallerstein would call 'semiperipheral' at the time they produced the relevant theorists) was explained mostly by the dominant market position of the most developed capitalist powers of those times, first England, and later, the major West European powers and the United States. The implications were that in order to develop, the poorer societies had to shield themselves from the world market, to somehow exclude that world system. Nationalism justified such policies as essential in order to protect the self-interest and integrity of the less developed nations.

World War II discredited fascist–corporatist notions of development, but the theory retained its attractiveness, and it gradually migrated leftward. In Latin America, Raul Prebsich, Celso Furtado, and Fernando Henrique Cardoso became the chief spokesmen for what came to be called 'Dependency Theory.' The explanation of backwardness remained essentially the same, that the developed capitalist powers impeded the development of those who were more backward. But now it no longer seemed necessary to carry out protectionist theories within the framework of an authoritarian, ultranationalist, fascist–corporatist state (Cardoso and Faletto, 1979). It should be noted that Cardoso, who eventually became Brazil's president, abandoned most of the implications of his theory long before achieving power in his own county. Meanwhile, the theory became increasingly popular within the revolutionary left.

The problem was that a high degree of protectionism in an increasingly powerful and advanced world economy required heavy state intervention to block market forces and popular demand for the consumer goods produced by the world capitalist system. Therefore, some sort of authoritarian, state centralizing mechanism was required to carry out such policies rigorously enough to permit escape from the world system. If the theorists were right, the capitalist core, now led by the United States, would use its economic, diplomatic, and even military power to force open recalcitrant nations. This was the theoretical explanation of American intervention in Vietnam and throughout the Third World. To motivate their people and mount a defense, the leaders of anticapitalist regimes would have to use nationalism as the glue binding their people together in struggle, because they would be unable to provide the kind of consumer rewards offered by world capitalism. For those hostile to right wing, authoritarian nationalism, this led inescapably to the conclusion reached by Wallerstein and his followers, that Leninism in its most autarkic, closed Stalinist or Maoist forms was the only way to protect antisystemic nations until the world revolution achieved its goal of destroying the world capitalist system and replacing it with a more equitable, redistributive world socialist system.

This created an intellectual quandary recognized by many of WST's early followers who did not relish the ideological conclusions reached by Wallerstein and those who remained with him. Could something be saved from the comparative, holistic methodological approach advocated by Wallerstein without having to accept the extreme leftism he and his closest followers, such as Christopher Chase-Dunn (1989), were also promoting? The problem became all the more acute in the late 1980s and 1990s as communism collapsed and Marxism–Leninism was largely discredited throughout the world.

Today, the main challenge to the hegemony of an American dominated world capitalist system comes from the kind of state centered capitalism championed by China that accepts the existence of that system and operates within it, but with a somewhat different set of rules. If China succeeds in its ambitions, in WST terms, it would become a new hegemonic power, but certainly not one eager to end capitalism. On the other hand, the anticapitalist leftist populism of a figure such as Hugo Chavez of Venezuela is so muddled and ad hoc that it hardly presents much of threat to the existing world system.

The Uses of WST as a Supplement to but Not a Substitute for Comparative Economics

Despite its shortcomings, a world systemic approach to the study of social change remains a powerful antidote to narrow specialization. It also addresses problems that will remain with us far into the future, and it offers insights absent in other theoretical models.

While many economists now recognize that perfectly free trade policies imposed on poorer economies may indeed have detrimental effects, and even that the globalization of the late twentieth to early twenty-first centuries may have harmed richer economies, few if any accept the notion that autarky is a reasonable development policy, whatever the type of political regime (Stiglitz, 2006; Amsden, 2007). The import-substitution

policies followed by many Latin American countries from World War II until the 1980s, as well as by India and many other countries, slowed economic growth, led to entrenched, corrupt domestic interest groups that blocked innovation and reform, and ultimately provoked economic and financial crises that threatened economic catastrophe. On the other hand, trade protectionism and developmental policies that fostered export-led industrialization in East Asia produced robust economic growth, and, generally, it is now accepted that working within the capitalist world system but with good governance and sound domestic development policies works well (Ray, 1998; Rodrik, 2003). In contrast, the truly extreme forms of withdrawal from the world capitalist system, that is the autarkic communist states, have all produced economic disasters.

Economic historians have shown that in the nineteenth century, at least, if not before, capitalist development did not rely on colonial markets, so that the economic argument in favor of imperialism turned out to be largely specious (Bairoch, 1993). This brings into doubt yet another of the fundamental assumptions of WST. Yet, no matter what their strengths, there remains something unsatisfying about purely economic models of development.

In the hands of comparativists who place the societies they analyze in a global context and understand capitalism's transformative power, but who also recognize the importance of domestic social structures and the variations caused by specific historical experiences, world systemic ideas help account for some of the differences in rates of economic modernization. Peter Evans (1995) has done this in explaining the success of certain state-led industrialization projects. Though somewhat suspicious of core power, especially American hegemony in the contemporary world capitalist system, Evans recognizes that development within that system remains possible for those who know how to combine effective bureaucratic leadership with local entrepreneurial talent, and who remain oriented to succeeding within the world system and using its advanced technology and market opportunities rather than revolting against capitalism.

Historical studies such as the one by Resat Kasaba (1988) also show how WST can provide important insights as long as it is not used dogmatically. He proves that in the mid-nineteenth century local merchants within the Ottoman Empire were able to use growing trade relations with the capitalist world to strengthen their own position against both the Ottoman state and the interests of the hegemonic core powers. But later, the Ottoman bureaucracy, helped by some powerful foreign, core financial institutions regained its strength, and this reduced the economic power of the domestic merchant class. Far from being 'compradors,' that is, subservient tools of the capitalist core (which is what conventional WST and most Marxists contemptuously call them), these local entrepreneurs might have spearheaded an economic and social modernization that would have made Ottoman society more open, tolerant, and progressive. The world capitalist system did indeed play an important role in peripheralizing late Ottoman society, but not in the expected way, and only because it was successfully manipulated by an increasingly nationalistic, Turkish, anticommercial state elite.

When East-Central European economic histories are compared, it turns out that contact with the growing capitalist world economy centered in northwestern Europe and later Germany did indeed play an important role in the past, but not in the way predicted by WST. The more closely associated these economies were with the core capitalist ones in the past, the more advanced they have been in the modern era, and this holds true to this day (Chirot, 1989; The Economist, 2012).

In other words, it is certainly very useful to take into account a particular society's position in the world capitalist system. Global trade, information, and political networks unquestionably play an important role in how various societies develop. Exploitation of colonial economies did hurt many of them in the long run (Bairoch, 1993). Extracting resources, particularly oil, from countries with weak governance institutions and corrupt power elites clearly sets those peripheral economies back, or at least stymies their economic growth and general prosperity. This is most evident in a number of African cases, but also in oil-rich former Soviet Republics, including Russia. Good comparative studies, however, show that there is no strict causal relationship between being a peripheral resource exporter and slow development. On the contrary, it is the interaction between global position and domestic institutions that is the key to whether or not balanced economic growth occurs (Humphreys et al., 2007).

Long-term business cycles used to be studied by economists, most famously, by Joseph Schumpeter, and more recently by Walt Rostow (American Economic Association, 1944; Rostow, 1971). This tradition practically died out in economics in the late twentieth century. Business cycle theorizing came to be, instead, an attempt to explain away short-term changes, not systemic long-term ones (Hartley et al., 1998). Here, WST and especially Immanuel Wallerstein have taken up the challenge and claimed to see very long-term waves of growth and decline over the many centuries of the world capitalist system. How WST measures and interprets such cycles can be questioned, but it is reasonable to assume that major technological changes over time usher in drastic social and political as well as economic changes. The development of cheaply produced cotton textiles brought about the first phase of the industrial revolution. Subsequent major inventions, from the railroads to the advanced use of chemistry and the technological leap that produced widespread usage of electricity, underlay the rise of Great Britain, then France, Germany, and the United States in the nineteenth and early twentieth century. This was followed by the age of automobiles, then the electronics revolution, and now a transformation just getting under way in communications and the use of biotechnology. At each stage, certain countries led the way and gained power, but each major new technology also unsettled old social institutions and produced crises because of failures to adapt in time. While many analysts might not agree, WST would predict that the world economic crisis that began in 2008 is a function of just such a shift and the failure of both Americans and Europeans to properly adapt, at least so far (Chirot, 2011).

Theoretical Limitations and Methodological Promises

WST's methodological approach reaffirms an old but often neglected concept. All societies have operated within larger

systems, and without placing them in that context, our understanding of how they work is severely limited. There can be no general laws of social behavior that do not take this into account.

Beyond this, WST has shed light on the nature of modernization, even though that is a term it avoids in order to show it rejects the functionalist theories proposed by 'Modernization Theory' in the 1950s and 1960s. WST has shown that there is indeed a capitalist world system, and its rules, largely determined by the core powers, affect everyone. Furthermore, this capitalist world is itself an evolving, changing structure. It is subject to periodic waves of prosperity followed by economic downturns. Even economists need to take such large systemic factors into account in order to perfect their models. So do the other social scientists who, by and large, limit comparisons to discrete cases rather than incorporating system-wide variables.

A world systemic approach to social science, however, must admit the limitations of the theory as it now stands. Aside from its tendency to neglect economic theory, mentioned above, WST has also relegated technology and science to an almost insignificant role. Thus, Wallerstein and most other followers of WST refuse to admit that Western European capitalism's early dynamism depended on the religious and scientific transformations that occurred during the Renaissance, Reformation, and Enlightenment. Max Weber remains a better guide to this process of change than any of the contemporary practitioners of WST, and by neglecting him, and the cultural issues he raised, they have weakened their paradigm.

Ideas generated by the core, from science and technology to various political ideologies such as nationalism and Marxism, have affected the rest of the world as much as the economic and military power of the core. These ideas have been transformed in many ways within the periphery and have stirred up particularly disruptive ideological currents of resentment and antimodernism in what Wallerstein has called the semi-periphery. There is no theoretical reason why WST could not use its model to expand our understanding of how this has worked. In fact, one of Wallerstein's most devoted followers, Georgi Derlugian (2005) has used the analysis of culture as well as more conventional WST to explain the catastrophic developments in much of the Muslim part of the former Soviet Union.

There is no need to continue to adhere to the simplistic materialism so many of its practitioners now follow. Ideas, scientific innovation, and even changing notions of social and political morality, including the unexpected increase in the role of religion throughout much of the world, can be analyzed in world systemic ways to great advantage.

When he first presented his theories, Wallerstein declared himself to be a disciple of Fernand Braudel and the French *Annales* School. That Braudel himself (1979/English translations 1981–84) was influenced by Wallerstein is evident in the vast comparative sweep of entire regions, and ultimately the globe, that he treated as a system of interdependent parts.

In a sense, WST has succeeded well enough in influencing historical studies of social, economic, and political changes that its ideas appear in some very influential, major studies. Some cite Wallerstein directly, while others do not but use ideas and sources influenced by his ideas.

For example, the historian Anthony Reid (1988/1993) has presented a daringly new vision of Southeast Asian economic, cultural, and political history by putting it into the kind of comparative and global perspective that WST has always championed.

Ken Jowitt's analysis of the failed attempt to create a Leninist world system (Jowitt, 1992) also adopts a world systemic approach without admitting it. He points out that the communists actually were more successful at creating a cultural and ideological world system of their own, with a powerful core struggling to keep its peripheries under control, than at creating a system based on economic links. He then predicts that the collapse of Leninism will have a globally destabilizing effect because it will leave the resentful losers in the world capitalist system without any constructive or predictable revolutionary alternative.

More recently, Kenneth Pomeranz (2000) in his pathbreaking explanation of why China fell behind the West, particularly Great Britain, in its economic development has incorporated WST ideas. For him, the explanation does not lie so much in England's development of modern ideas, but rather in the combination of the fortunate placement of its coal and its ability to exploit its colonial empire and domination of international trade routes. To be sure, these conclusions are themselves controversial and can be questioned, but the success of Pomeranz' shows that WST ideas continue to be spreading.

Even a superb recent history textbook by Jane Burbank and Frederick Cooper (2010) incorporates many WST ideas without directly citing Wallerstein. Their book makes the interaction between major empires the key to understanding the history of the modern era.

WST's potential is far from exhausted. Needless to say, many leftist social scientists such as David Harvey (2010) have incorporated WST ideas into their work, but the important point is that this approach has influenced much analytical work that is not specifically on the far left. Flexibly used, it still has lot of new insights to offer, and it is likely that in the twenty-first century, in response to increasing globalization, continuing inequality, uneven prosperity, and worldwide economic crises, it will continue to develop in ways that will take it well beyond its sectarian origins.

See also: Capitalism; Center–Periphery Relationships; Class: Social; Dependency Theory; Economic Sociology; Finance, Geography of; Globalization and Religion; Globalization, Anthropology of; Historical Sociology; Imperialism, History of; Imperialism; Marxism in Contemporary Sociology; Modernization, Sociological Theories of; Nationalism, Sociology of; Revolution, Theories of.

Bibliography

American Economic Association (Ed.), 1944. Readings in Business Cycle Theory. Blakiston, Philadelphia, PA.

Amsden, A., 2007. Escape from Empire: The Developing World's Journey through Heaven and Hell. MIT Press, Cambridge, MA.

Bairoch, P., 1993. Economics and World History: Myths and Paradoxes. University of Chicago Press, Chicago, IL.
Braudel, F., 1979/1981–1984. Civilization & Capitalism 15th–18th Century, vol. 3. Harper & Row, New York.
Burbank, J., Cooper, F., 2010. Empires in World History. Princeton University Press, Princeton, NJ.
Cardoso, F.H., Faletto, E., 1979. Dependency and Development in Latin America. University of California Press, Berkeley, CA.
Chase-Dunn, C.K., 1989. Global Formation: Structures of the World-Economy. Basil Blackwell, Cambridge, MA.
Chirot, D. (Ed.), 1989. The Origins of Backwardness in Eastern Europe. University of California Press, Berkeley, CA.
Chirot, D., 2011. How Societies Change. Sage, Thousand Oaks, CA.
Derlugian, G., 2005. Bourdieu's Secret Admirer in the Caucasus. University of Chicago Press, Chicago, IL.
The Economist, 2012. Pocket World in Figures. Profile Books, London.
Evans, P.B., 1995. Embedded Autonomy: States and Industrial Transformation. Princeton University Press, Princeton, NJ.
Harvey, D., 2010. The Enigma of Capital: And the Crises of Capitalism. Oxford University Press, Oxford, UK.
Hartley, J., Hoover, K., Salyer, K. (Eds.), 1998. Real Business Cycles. Routledge, New York.
Humphreys, M., Sachs, J., Stiglitz, J., 2007. Escaping the Resource Curse. Columbia University Press, New York.
Jowitt, K., 1992. New World Disorder: The Leninist Extinction. University of California Press, Berkeley, CA.
Kasaba, R., 1988. The Ottoman Empire and the World Economy: The Nineteenth Century. State University of New York Press, Albany, NY.
Lenin, V.I., 1939. Imperialism, the Highest Stage of Capitalism: A Popular Outline. International Publishers, New York.
Love, J.L., 1996. Crafting the Third World: Theorizing Underdevelopment in Romania and Brazil. Stanford University Press, Stanford, CA.
Pomeranz, K., 2000. The Great Divergence. Princeton University Press, Princeton, NJ.
Ray, D., 1998. Development Economics. Princeton University Press, Princeton, NJ.
Reid, A., 1988/1993. Southeast Asia in the Age of Commerce, 1450–1680, vol. 2. Yale University Press, New Haven, CT.
Rodrik, D. (Ed.), 2003. Search of Prosperity. Princeton University Press, Princeton, NJ.
Rostow, W.W., 1971. Politics and the Stages of Growth. Cambridge University Press, Cambridge, UK.
Stiglitz, J., 2006. Making Globalization Work. W.W. Norton, New York.
Wallerstein, I.M., 1974. The Modern World System. Academic Press, New York.
Wallerstein, I.M., 1979. The Capitalist World Economy: Essays. Cambridge University Press, Cambridge, UK.
Wallerstein, I.M., 2011. The Modern World System, vol. 4. University of California Press, Berkeley, CA.

Worldview

Rik Pinxten, Ghent University, Ghent, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In anthropology worldview studies have long been linked to a mentalist approach to culture, especially from Boas to Lévi-Strauss, with sophisticated analyses in so-called cognitive anthropology. A modest materialistic view emerged as a critique, with its most powerful elaboration in cross-cultural. Also, an evolutionary view on man–environment emerged, speaking of 'sphere' instead. Applications can be found today in such technical subdisciplines as ethnomathematics.

Finally, with the virtual disappearance of worldview studies in anthropology, the concept is found in philosophy of science and in religious studies.

The Notion of Worldview

The literature on worldviews leaves one with a feeling of uneasiness. On the one hand some philosophers and scholars in religious studies claim that the progressively scattered field of knowledge of the twentieth century points to the importance of worldview as a unifying force for the organization and deployment of scientific research. It would act like a logical theory in the project of Enlightenment. On the other hand, explicit use of the term *worldview* in anthropology is decreasing over the years, to the extent that some have announced the end of it (Beine, 2010).

Hence, presenting a critical assessment of the notion for an encyclopedia is far from easy. This article will focus on the use of worldview in anthropological studies first, and expand from there to a larger interdisciplinary discussion.

Anthropological Studies

Worldview has been associated over a long period with the culturally idealistic approaches in anthropology. Starting with Franz Boas, anthropologists have been working in the perspective of the 'psychic unity of humanity,' meaning that all human beings would share the same basic categories of thought, but also of sentiment and volition. For example, time, space, causality, and such would be universally shared, although particular shapes and contents of them could be distinguished in diverse cultural traditions. In a general sense, this perspective is based on a philosophical program of mentalism: researchers took for granted that 'in the human mind' or 'in immaterial culture' situated in the mind, such categories can be found. Kearney (1984) states, in a rare overview of research on this issue, that this emphasis on the 'mental' across cultural borders is probably due to the close links between linguistics and anthropology in the American case: language structures were presupposed to exist in the mind, and hence by extension all other structures and processes that organize the experiences of an outside world would be situated there too. The encompassing phenomenon, which would synthesize concepts, views, and expectations about the way the world is, is the worldview of a cultural group: a worldview is the shared software of a cultural group, which processes input from the senses, stores its information, and serves as the basis for action for the collectivity (after Kearney, 1984: chapter 2).

Most certainly, Claude Lévi-Strauss reinforced the mentalism in anthropological studies by starting almost exclusively from a linguistic perspective with his structuralist anthropology (Lévi-Strauss, 1958). In his theory, the deep structures, which matter most for the study of any domain of culture in any part of the world, are to be found in the human mind. Their universal form is uncontested in structuralism until the 1980s of the past century. In the view of Lévi-Strauss sense data, concepts, principles, and systems of thought and action are structured logically, much like the data in a computer. In his study of thinking and knowing in non-Western cultures, nor in his systematic comparative analyses of myth Lévi-Strauss questions the mentalist presupposition: in his perspective it is 'all in the mind.' He does not doubt either that the motor function (or the generative force) of the deep structures can be found in the human mind. Thus, in the famous *La pensée sauvage* (1962) the primacy of classification logic in the human 'hardware' is an unquestioned point of departure, which allows the author to sketch human thinking worldwide as the universal mapping and structuring of inputs from external reality into highly comparable worldviews. The differences he finds are interpreted by means of structural nuances, which are reduced to the workings of the very same universal ironclad logic in the mind. In his numerous studies of mythology a similar intellectual move can be perceived: in these very elaborate analyses Lévi-Strauss (1980) tries to show how the same inborn and hence universal logic is at work in the structuring of human imagination throughout the world. In that sense, the structuralist school of Lévi-Strauss illustrates very strongly what Kearney criticizes as a mentalist view on culture. Although Lévi-Strauss hardly ever uses the term *worldview* to indicate the ensuing encompassing structures that he finds (or builds) across cultures, his use of deep structures in the mind of humanity seems to amount to the same effect.

In the period of the 1960s until the 1980s, a similar, but more differentiated, subdiscipline emerged in American anthropology, known as ethnoscience. It grew out of folk taxonomies – a range of subdisciplines from ethnobotany to ethnozoology – and was later generalized to cognitive anthropology. In an overview article by one of the scholars in this branch of anthropology (and linguistics), Werner and

Schoepfle (1987) explain that this approach aims at mapping in a systematic way what amounts to be the knowledge of different cultural traditions. In practice, ethnographers trigger terms and meanings from a set of informants while presupposing that all human beings actually work in their mind with a preconceived (or inborn) classification logic. Doing fieldwork then amounts to 'filling in' by means of the informants the formal structures of classification about different domains of reality that are in themselves unquestioned, because all human minds are believed to work in a rather exclusive way within the frame of universal classification logic. Of course, different informants will be competent in different reality domains. Hence, particular taxonomies will be filled in more by some and less by others. The ethnographer is, in Werner's view, the one who can construct the ideal knowledge system that is the sum of all competent data and classification thinking in one particular cultural group. Werner calls this the 'synthetic informant model' or SIM: it holds the knowledge of the 'omniscient informant.' No one member of a community has the total classification system of knowledge, all terms, and all concepts that are formulated in the model. Only the virtual 'omniscient informant,' built up as a product of research with many particular informants, adding on bits and pieces along the way, will constitute the overview of knowledge available in the community. The anthropologist thus constructs a sort of superstructure of worldview items that is supposed to represent all partial worldview items in real and particular members of the community. The SIM is warranted, according to the researchers, because it synthesizes in an encompassing model what is present in a scattered way throughout individual cultural subjects. This thought is intriguing, since it carries to its limit the a prioris that I indicated in Lévi-Strauss's approach. Based on the presupposition of the universality of taxonomic thinking as the preferential or even unique way of mapping reality, the 'omniscient informant' of SIM thus becomes something like the necessary and natural 'omniscient informant' of SIM, and thus becomes something like the necessary and master discourse for a community. Although not one person in any particular community will actually incorporate the model, the anthropologist pretends that the construct as SIM would be warranted and truthful because it would be the most encompassing instantiation of the deep logic at work. It is clear that this approach carries the philosophical idealism of former scholars to an extreme: actual flaws and patchwork patterns in empirical work do not hinder the scholar because the deep structure logic that the researcher projects into human worldviews guarantees the validity of the virtual model.

Over the years critical assessments have been developed, yielding more varied and nuanced descriptions and models. On the one hand, different logics have been recognized in non-Western cultures, e.g., inference next to classification logic. Thorough experimental work has been done (e.g., Cole, 1996). On the other hand, knowledge and learning have been looked at much more as contextualized processes. This made for perspectives on worldviews as differentiated and dynamic ways of production and use of knowledge, rather than fixed or inborn software structures (Lave, 1990). Hence, the worldview as such is not the focus of attention, but rather the processes of acquisition, production, use and change of data, models, and folk theories. Along the way, the term *worldview* silently disappeared from the anthropological literature. A quick survey of textbooks over the past decade teaches us that almost none of them mention the concept in their index. A rare exception is Bonvillain (2006), who uses the term exclusively in the context of struggles by minorities to claim their rights in a context of marginalization: "[Worldview is the] culture-based, often ethnocentric, way that people see the world and other peoples." (Bonvillain, 2006: 35).

However, a second approach should be mentioned. It is the materialistic view on culture (in the line of M. Harris and a small group of Marxists) elaborated by Kearney. In his book on worldview, Kearney (1984) explicitly turns against the philosophically idealist approach focused on in the previous sections of the present contribution. Although he dwells on historical roots most of the time (Boas and others) and is limited in his scope to work before 1980, he makes a deep critique of the mentalism in most of the old work, which is largely valid for the broader cognitive anthropological studies of a later date.

Kearney refers succinctly to the work of the Russian psychologist Vygotsky. The so-called sociohistorical approach of this early Marxist psychologist was an example of situating thinking, and hence worldview, in sociohistorical contexts. With a school of researchers (Luria, Leontiev, and many others, see Wertsch, 1990), Vygotsky developed an approach on the cognitive world that emphasized social functions as well as the practical workings of concepts, models, and worldview. Kearney picks this up and develops a materialistic theory on worldview that recognizes static aspects (like classification systems) but at the same time looks at subjects in their manifold interrelatedness: they are related to their self, and to others. On top of that they are, through learning processes, related to concepts and pragmatic procedures that are transferred by groups or communities: notions of dealing with time, space, causality, and so on. Kearney emphasizes that the ecological environment has impact on worldviews as well, for example, cosmological processes induce views about time, while the daily experience of the geographical environment will serve as foundation for a 'carpentered space' (in the Western worldview, see Campbell, 1989) or else a 'habitat lifespace' in some oral traditions (Pinxten et al., 1983).

What Kearney did not know, apparently, is that the Vygotsky school became very influential in the West, basically by the rediscovery, translation, and subsequent introduction in the Anglo-Saxon world of his writings and a lot of the publications of his collaborators by scholars such as J. Wertsch and M. Cole. Not only did a series of books get translated and discussed but at least two English language journals were launched (*Activity, Mind and Society* and *Culture & Psychology*). Moreover, a whole network of groups and institutional programs was developed in the past two decades, stretching from Russia over Europe to the United States and Japan. So, although Kearney's proposal to work out a materialistic perspective on worldview, with links to the early Marxist psychology of Vygotsky, was not picked up in anthropology, a strong group of psychologists now work out a materialistic view, combining cognition, culture,

and an action perspective in thorough and cross-culturally oriented research. However, the notion of worldview, still so central in Kearney's book, is not used anymore in this line of 'cultural psychology' (which is the actual title of Cole, 1996). Rather, the focus is on types of activity, types of contexts, mathematical skills, time (e.g., Vaalsiner, 2006), and so on.

Worldview and Sphere

A special case of recent date is the work of T. Ingold. This influential and prolific anthropologist combines an evolutionary approach with a deep interest in an environmental focus on culture and learning. The old work of ecologists such as von Uexkühl, von Bertalanffy, and others studied social and cultural patterns and forms in the ecological context of the subjects. The term *worldview* is seldom used, but such concepts as 'Umwelt' or ecosystem seem to be conceptually similar or overlapping with what was understood by worldview in the social sciences and philosophy. The fact that these authors came from biological sciences may account for not having them brought into contact with the worldview perspective of the humanities. *The Perception of the Environment* by Ingold (2004) is, in my view, the one great exception of recent date in that respect. Ingold did thorough ethnographic research with the Saami people in Finland. In this major book, he expands his scope to encompass all hunter-gatherer cultures. A central issue he introduces has to do with cosmology, or the broader notion of 'worldview' (without using the term in any central way). Ingold explains how hunter-gatherer peoples live within nature, whereas at least Western peasant cultures share a worldview of humans living vis-à-vis nature. The latter have an objectifying view: humans look upon all other creatures as if from an outsider position. Ingold examines how the globalization discourse of today's social sciences can only be developed and even understood in this latter worldview: the whole world is constructed as the context that exists outside of human beings, and hence the natural environment is seen as a globe by humans, who distance themselves from it by conceiving as a whole existing independent and outside of the observer. In the worldview of hunter-gatherers, the environment is seen as a sphere, that is to say an environment that is like a habitat, with all creatures (including humans) living within the whole. The consequences of these different stands on nature or environment are tremendous: in the hunter-gatherer culture humans see themselves on an equal level with animals and other natural phenomena, as is shown in the 'discourse' between animals and humans in such cultures (e.g., illustrated in the rock paintings and drawings Ingold discusses). In the Western view on reality as a globe, humans see the environment as the objective other, which is somehow estranged from human beings and can hence be manipulated by them. My reading of this new line of research is that, by linking the cultural and social anthropological models (again) with the broader ecological perspective, Ingold in fact opens up a new line of thinking on worldview.

Applications

Ingold did link his approach to culture with natural sciences, especially with biology. On the other hand, in the work cited as well as elsewhere (Ingold, 2010) he shows a keen interest in the production and use of artifacts, including what is generally referred to as art. In that sense he allows for applied anthropology to be linked with worldview. In a very specialized field, outside of anthropology, this interdisciplinary link between the worldview focus and applied anthropology is booming in recent research (of the first and second decade of this twenty-first century). In what is now known as ethnomathematics – linked in name and by means of some researchers in that field with the cognitive anthropological approach, but mainly emanating from mathematics education circles – the relevance of worldview is keenly discussed. The founding father of this subdiscipline, the Brazilian mathematician U. D'Ambrosio (1985) has made multiple references to the living conditions, the cultural practices and beliefs, the social aspects of schooling and of learning in general, and how all of these impact the success or failure of mathematics education.

A group of sociologists and anthropologists carried out research in the same perspective. One of the influential scholars in this regard is the Danish mathematician Skovsmose who successfully launched the concepts of 'background knowledge' (BK) and 'foreground knowledge' (FK) (Skovsmose, 1994). BK refers to the preschool knowledge a child brings to school from her culture: concepts and models about reality, linguistic categories, learning styles, and strategies to build and use knowledge. FK then encompasses all those concepts and skills the teacher can add on to the BK of each child to introduce it by means of insightful and relevant (to the child) steps to a further level of formal thought. Here again the notion of 'worldview' as such is hardly ever used in the literature. But to my mind the combination of BK and FK seems to add up to the content of what researchers in the humanities and the social sciences understood by worldview (Pinxten and François, 2011). The whole engagement with ethnomathematics is to understand what formal thought (and mathematical thinking as an integral part of it) amounts to in different traditions of thinking and learning. From this knowledge the mathematics educator will then try to develop learning procedures and curriculum material that will allow an opportunity to counter the serious dropout rates of minority groups and lower social groups.

Religion, Ideology, and Worldview

In the past decades the concept of worldview is abundantly and almost exclusively used in religious contexts. Not surprisingly it then also has a normative ring.

Already in the Marxist perspective on worldview, which was introduced by Kearney (1984) action and action strategies were an integral part of the concept (his Chapter 2). That focus broke away from the purely objectifying 'idealist' view of structuralists and cognitive anthropologists. In a broader philosophical view we find worldview linked with the old German notion of *Weltanschauung*, with a decidedly normative aspect: it refers

to the way the world is known and is (to be) interacted with. Most scholars refer to Wilhelm von Humboldt (1767–35) as the originator of this notion. Today, philosophers and scholars in religious studies continue on this line of research. Thus, the well-known scholar in comparative study of religion Ninian Smart gives it a central place in his study on human beliefs (Smart, 2000), while other Christian scholars diminish the role of the cognitive or knowledge dimension of worldview in order to underline the prominence of religious and existential dimensions. Thus Sire (2004) states that worldview is "a commitment, a fundamental orientation of the heart, that can be expressed as a story or a set of presuppositions (assumptions which may be true, partially true, or entirely false) which we hold (consciously or subconsciously, consistently or inconsistently) about the construction of reality, and that proved the foundation on which we live and move and have our being" (Sire, 2004: 15–16).

The conservative philosopher of science and of religion McGrath (2004) approvingly cites MacIntyre in his disbelief about the project of Enlightenment and calls for a reuse of a worldview notion that includes knowledge but places it firmly in a broader, not only rational but clearly religious (Christian), foundation.

Whatever the chances for success may be for such a re-emergence of a religious worldview notion, the general criticism on a detached and purely cognitive perspective on worldview and knowledge has been picked up by numerous religious scholars outside of anthropology.

An interesting research group of the past decades was started by the Belgian philosopher and logician Leo Apostel. He was trained by the philosopher of rhetoric Chaim Perelman and by the logician Rudolph Carnap. From the latter's Vienna Circle work he borrowed the ideal of a 'unified science.' Apostel worked extensively with Jean Piaget, adopting the focus on logically coherent development of knowledge from him. Finally, he was close to the research group of Ilya Prigogine, who combined ontological interests with thermodynamics. Apostel embraced some of the criticism on Enlightenment thinking and started out to work on an ethically responsible approach to science (parallel to I. Stengers, M. Serres, B. Latour, and others today). This critical position did not reject rationality or a strongly scientific approach to knowledge but looked at science as a form of contextualized thinking and acting. Apostel engaged in extensive collaborative research with scholars from different disciplines and also took care to involve researchers from various life stance and religious positions. In the course of this work he developed the idea that scientific knowledge was too scattered and hence vulnerable to particular and even particularistic interests. At the same time the group around Apostel recognized that science is produced by human beings (taking in the criticism by Kuhn, 1962, and sociologists of science) and that the latter were in need of a synthetic and self-critical worldview to found their intuitions and heuristics. He founded a research center at the Free University of Brussels (the Center Leo Apostel), which is entirely devoted to the study of worldviews in philosophy and science. He thus merged the 'unified science' ideal of the Vienna Circle with the pluralistic and interdisciplinary approach he explored in the 1970s and 1980s of the past century. In one of his last publications (Apostel and Vanderveken, 1991) the project is outlined as a major research program for the whole of science, recognizing at the same time that science is a human endeavor and thus in need of more than only rationalistic principles. In a sense, the criticism on Enlightenment thinking by conservative thinkers and religious scholars (such as McGrath) is taken up as a fair critique. It is answered by the exploratory research in worldviews: a new, scientifically screened and tested worldview is needed, according the group, in order to overcome the failures and inconsistencies of the old Enlightenment project. The center was successful in attracting research grants, producing PhDs, and publishing worldwide. It certainly acquired a high status by organizing a series of pathbreaking symposia, followed by an intriguing book series with a first-class academic publisher, under the inviting title Einstein Meets *Magritte*. The series title in itself highlights the programmatic perspective of the worldview group in Brussels: both science and art, both cognition and vision. Over the years a plethora of great names has become attached to the initiatives of this group, all of them engaging themselves with the worldview notion in one way or another (e.g., see Aerts et al., 2005).

Whether the program will be successful in the end is an open question. It is relevant here to mention the initiative since it focuses squarely and uniquely on worldview. In their perspective science needs a unifying worldview today. This worldview will certainly be cognitive but does entail ontological, religious-moral, and political stands as well. Because scientific rigor is combined with societal engagement, it presents an intriguing positive alternative to the merely ideological worldview concepts of the religions. In that sense the Brussels worldview program is linked with the older anthropological perspective (explicit use is made of anthropological material, but also of the contemporary sociology and ethnography of science with Latour and others, Apostel and Vanderveken, 1991). On the other hand the ecological and evolutionary thinkers (e.g., von Bertalanffy) as well as genetic psychologists (e.g., Piaget) are recognized as foundational for the contemporary project. A question that remains is to what extent the group is aware and critical of the transcendental implications of former philosophies of science (such as Kant and others) and is able to break away from that deep-seated pretention and move to a genuine comparative and possibly pluralistic perspective on worldview (see Aerts et al., 2005, 2011; Note et al., 2009; Vanderbeken et al., 2011).

Conclusion

In this article one finds a short introduction to the notion of worldview in anthropology. The use of this notion as it were stopped after structuralism. Except for the original but rather unique views of Ingold, the perspective as such seems to have become absent from the discipline. The Marxist interpretation of worldview is essential in Kearney's work and is somewhat differently continued in the Western follow-up of Vygotsky's school in what is now called 'cultural psychology.' More ideological positions with explicit use of the term *worldview* will be found today in religious documents. Finally, a single interdisciplinary group, emerging from philosophy, takes worldview as its primary focus (Apostel).

Bibliography

Aerts, D., D'Hooghe, B., Note, N. (Eds.), 2005. Worldviews, Science and Us. Redemarcating Knowledge and Its Social and Ethical Implications. Worldscientific Publishing, New Jersey.

Aerts, D., D'Hooghe, B., Pinxten, R., Wallerstein, I. (Eds.), 2011. Worldviews, Science and Us. Interdisciplinary Perspectives on Worlds, Cultures and Society. Worldscientific Publishing, New Jersey.

Apostel, L., Vanderveken, J., 1991. Wereldbeelden. Pelckmans, Antwerp.

Beine, D., 2010. The End of Worldview in Anthropology? SIL Electronic Working Papers.

Bonvillain, N., 2006. Cultural Anthropology. Prentice Hall, New Jersey.

Campbell, D.T., 1989. Methodology and Epistemology for Social Science. Chicago University Press, Chicago.

Cole, M., 1996. Cultural Psychology. Harvard University Press, Cambridge, Massachusetts.

D'Ambrosio, U., 1985. Ethnomathematics and its place in the history and pedagogy of mathematics. For the Learning of Mathematics 5 (1), 44–48.

Ingold, T., 2004. The Perception of the Environment. Routledge, London.

Ingold, T., 2010. Ways of mind-walking: reading, writing, painting. Visual Cultures 25 (1), 15–23.

Kearney, M., 1984. World View. Chandler & Sharp, Novato.

Kuhn, T., 1962. The Structure of Scientific Revolutions. Chicago University Press, Chicago.

Lave, J., 1990. Cognition in Practice. Cambridge University Press, Cambridge.

Lévi-Strauss, C., 1958. Anthropologie Structurale. Plon, Paris.

Lévi-Strauss, C., 1962. La Pensée Sauvage. Plon, Paris.

Lévi-Strauss, C., 1980. Les Mythologiques. Plon, Paris.

McGrath, A., 2004. The Science of God. T&T Clark, London.

Note, N., Fornet-Betancourt, R., Estermann, J., Aerts, D. (Eds.), 2009. Worldviews and Cultures. Einstein Meets Magritte, vol. 10. Springer, New York.

Pinxten, R., van Dooren, I., Harvey, F., 1983. Anthropology of Space. Explorations into Natural Philosophy and Semantics of the Navajo. University of Pennsylvania Press, Philadelphia.

Pinxten, R., François, K., 2011. Politics in an Indian canyon. Some thoughts on the implications of ethnomathematics. Educational Studies in Mathematics 78 (2), 261–273.

Sire, J., 2004. The Universe Next Door: A Basic Worldview Catalog. Intervarsity, London.

Skovsmose, O., 1994. Toward a Philosophy of Critical Education. Kluwer Academic Publishers, Dordrecht.

Smart, N., 2000. Worldviews: Crosscultural Explorations of Human Beliefs. Cambridge University Press, Cambridge.

Vaalsiner, J., 2006. Once upon a time…again. Culture and Psychology 12 (2), 139–143.

Vanderbeken, R., LeRoy, F., Stalpaert, C., Aerts, D. (Eds.), 2011. Drunk on Capitalism. An Interdisciplinary Reflection on Market Economy, Art and Science; Einstein Meets Magritte, vol. 11. Springer, New York.

Werner, O., Schoepfle, M., 1987. Systematic Fieldwork, vols. 2. Sage, London.

Wertsch, J., 1990. Vygotsky and the Social Formation of Mind. Harvard University Press, Cambridge, Massachusetts.

Wright, Sewall (1889–1988)

James F Crow, University of Wisconsin, Madison, WI, USA

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 23, pp. 16616–16619, © 2001, Elsevier Ltd.

Abstract

Sewall Wright was the leading US population and evolutionary geneticist. Along with R. A. Fisher, he converted breeding into a quantitative science. With his coefficient of inbreeding, now standard textbook material, he provided a simple algorithm for measuring the genetic consequences of inbreeding. In evolution, in addition to developing much of the theory of quantitative population genetics, he advocated his 'shifting balance theory.' He visualized a large population, broken up into partially isolated subgroups, within which random changes of gene frequencies could occur. Occasionally, such a subpopulation would drift into a state of higher fitness, whereupon it would increase and send migrants to upgrade the whole population. He viewed this as the best way for a population to bypass the evolutionary dilemma of not being able to go from one highly adapted state to a higher one without passing through maladapted intermediates. Wright's most important contribution to statistics is the method of path analysis, by which standardized partial regression coefficients are used to measure the relative strength of causal pathways. This is his major contribution to the social sciences.

Sewall Wright was born 21 December 1889, in Melrose, Massachusetts, and died in Madison, Wisconsin, 3 March 1988. Before his third birthday, the family moved to Galesburg, Illinois, where his father, Phillip G. Wright, had accepted a teaching position at Lombard College. Wright graduated from Lombard College in 1911 and, after a year at the University of Illinois, went to Harvard University as a student of William E. Castle. After receiving the doctorate in 1915, he was employed as senior animal husbandryman by the US Department of Agriculture (USDA) in Beltsville, Maryland. He remained there until 1925, when he accepted a position as Professor at the University of Chicago. Upon retirement in 1954 he moved to the University of Wisconsin, where he stayed for the rest of his long life. Wright remained intellectually and physically active until his death at age 98, caused by his slipping on an icy sidewalk during one of his regular long walks.

Early Life

Wright was a precocious child. He could read and kept a diary before starting school. He also learned arithmetic and astonished his first-grade teacher by extracting cube roots. Later he and his brothers printed the first poems of Carl Sandburg, who was a student in their father's composition class. At Lombard he took several courses from his father, including calculus. But, despite his later reputation as a formidable mathematician, Wright never took advanced courses; his mathematics was largely self-taught. His interest in genetics was kindled by reading Punnett's article on Mendelism in the Encyclopedia Britannica.

Between his junior and senior years of college, he used his mathematical and surveying skills to work on a railroad construction in the Standing Rock Reservation in South Dakota. It was an exciting time for him. In his 90s he enjoyed telling about the rough frontier life with cowboys, Indians, and outlaws, and he still recalled a large number of words in the Sioux dialect. This work was cut short by a lung infection. While confined to a caboose, he used the time to study Tait's book on quaternions. As a result of the damaged lung, he later had trouble getting life insurance, something that he found increasingly amusing as he continued to live far beyond the usual life-expectancy.

Wright's father, Philip, was a man of many talents. He taught courses in astronomy, mathematics, economics, surveying, English composition, and physical education. He loved poetry and music, and was disappointed that his son did not take to them. Later he moved to the Brookings Institute where he published a number of books on economics. Wright had two brothers who also attained distinction. Quincy became a leader in international law and Theodore was an aeronautical engineer and acting president of Cornell University.

A full length biography of Wright is available (Provine, 1986), with full details of his life and scientific accomplishments. His scientific papers span the period from 1912 to 1988, an astonishing 76 years. His bibliography comprises more than 200 articles, including a four volume set published during his retirement (Wright, 1968–78).

Animal Breeding

Wright's thesis at Harvard analyzed interaction of coat colors in guinea pigs, a subject which he pursued throughout his USDA and Chicago years. He enjoyed guinea pigs and spent many hours taking care of them. Only the non-availability of an animal house prevented his continuing this after his move to Wisconsin. While at Harvard, he also began his theoretical studies of population genetics. On moving to the USDA, he inherited a colony of guinea pigs that had been inbred for many generations. His analysis of the effects of inbreeding and crossbreeding set a standard for careful analysis; the paper could be written today with hardly a word changed. During this time he invented the inbreeding coefficient, with a simple algorithm now standard textbook material (Wright, 1922a). Also during this period, he formulated the basic principles for analysis of quantitative traits. Wright, along with R. A. Fisher (*see* Fisher,

Ronald A (1890–1962)), converted animal breeding into a quantitative science.

Wright was especially proficient in manipulating correlation and regression coefficients. One of his earliest papers (Wright, 1917) foreshadowed the analysis of covariance. His major statistical contribution was developing the method of path analysis. This is a way of measuring the relative contribution of different causal paths to a quantity of interest, such as size or performance. This was presented in a diagram in which causal paths were designated by arrows and correlations by curved, two-headed arrows (Wright, 1921). A path coefficient, measuring the influence of a path, is a standardized partial regression coefficient. Wright formulated a set of simple rules that made the analysis easy to apply. This method enjoyed great popularity among livestock breeders, although recently it has been largely replaced by more sophisticated, computer-driven techniques that permit analysis of large bodies of data, and measurement of the precision of the estimates.

The technique is most useful for nonexperimental situations and in recent years has found its greatest use in the social sciences. Wright himself used the method in the late teens to analyse 510 correlations of 42 variables involving corn yields and prices and hog production. He could not get this monumental paper published because an animal husbandryman was not supposed to know about economics. It was not published until several years later, and only after the intervention of Henry A. Wallace, son of the then Secretary of Agriculture.

Evolution

After moving to Chicago, Wright changed his emphasis from animal breeding to evolution, although he still retained his interest in the physiological genetics of guinea pigs. Three people, R. A. Fisher (1929) and J. B. S. Haldane (1932) in Britain, and Sewall Wright (1931) in the USA are mainly responsible for the 'new synthesis,' the welding of Mendelian inheritance and Darwinian natural selection into a coherent mathematical theory. Wright's particular contribution was the 'shifting balance theory.' He thought that a major difficulty in evolution was the frequent inability of a well-adapted organism to evolve into a better adapted one without passing through maladaptive intermediates. In Wright's metaphor, a population at an adaptive peak could not move to a higher peak without passing through less well adapted intermediate stages.

Wright thought that a way out of this dilemma was to have a population broken up into partially isolated subgroups. Within a small subgroup random gene-frequency changes can occur and occasionally these may produce a combination of genes that is close to that of the genotypes at the higher peak. Then, selection could take over and carry the population to the peak. This highly adapted subpopulation would then expand and send migrants to upgrade the whole population, and the process could start over.

Wright's theory became very popular among biologists. Its emphasis on gene interactions and its offering a way of putting together adaptive combinations of genes that might be deleterious in other combinations had a wide appeal. On the other hand, the theory was less popular with mathematical theorists. Wright's model required a delicate balance of population size, selection, and migration, which would be rare. It also required that the population spend much of its time, during the trial and error period, in a less than optimally adapted state. Finally, there is a question as to whether the theory is really needed; perhaps mass selection usually suffices for adaptive evolution.

Wright's theory was steadily refined during the rest of his life, but the essence remained unchanged. The theory became almost an obsession with him and he delivered many talks on the subject and wrote paper after paper. It is fair to say that, as to the importance of his theory, the jury is still out. Wright's latest paper (1988) was spirited defense of his idea, while still noting the importance of the contributions of Fisher and Haldane, and of Kimura (*see* Kimura, Motoo (1924–94)) who argued for the importance of mutation-driven, neutral evolution of molecular differences. He concluded that 'all are valid.'

Retirement Years

After moving to Wisconsin, Wright spent the first few years finishing the analysis of his extensive guinea pig studies. These were masterful analyses, but by that time the molecular revolution had changed the nature of developmental genetics, and Wright's methods were out of vogue. When this was completed, in his 70s he began his monumental four volume set, the last volume of which was published when he was 88 (Wright, 1968–78). Although his shifting balance theory is conspicuous throughout, the book is far more than this. It summarizes, in four densely packed volumes, not only his life's work, but the whole field of population genetics. For the last decade of his life Wright's eyesight became progressively worse, and he used a reader that projected a greatly enlarged picture onto a television screen. He continued to read and to enjoy conversations, but in his last years he found it increasingly difficult to keep up with molecular advances and turned more to reading history and biography.

As a person, Wright was quiet, shy, and introverted. He had no small talk, and was hard to converse with. Yet, once the conversation turned to a subject on which he had an interest – and there were many – he could go on at great length. His lectures were always full of details and regularly ran overtime. He had a charming sense of humor and was excessively generous with his time. His students and colleagues held him in great affection.

Social Sciences and Philosophy

Wright's greatest influence in the social sciences is his method of path analysis. It was taken up by several social scientists, such as Dudley Duncan in sociology and Arthur Goldberger in economics, and has become a standard part of the statistical methodology. One of Wright's earliest applications of path analysis used adopted children to separate genetic from environmental influences on IQ scores, and there have been many follow-up studies extending his methods to more complex analyses of behavioral traits with better data. Finally, with the increasing interest in behavioral evolution, especially

among psychologists, mathematical theories of evolution are becoming increasingly common and Wright takes a place in this, along with Fisher and Haldane.

Wright was unusual among biologists in taking an active interest in philosophy. He had a somewhat Leibnitzian view. He disliked any idea of emergence. He found no sharp borders between different levels of complexity, such as between embryo and adult or between mind and no mind. Thus if the mind does not emerge by magic, it must trace its development into a continuous process back to the embryo, the egg and sperm, and ultimately the DNA molecules. He called his concept 'dual-aspect panpsychism.' Mind is everywhere; so is matter.

The biological community either disagreed with Wright's view, or was indifferent. Most regarded mind as a natural outgrowth of appropriately organized matter. Wright held on to his dualistic view, however, and found a few kindred spirits among philosophers, particular his long time colleague at the University of Chicago, Charles Hartshorne.

Bibliography

Fisher, R.A., 1929. The Genetical Theory of Natural Selection. Oxford University Press, Oxford, UK.

Haldane, J.B.S., 1932. The Causes of Evolution. Longmans Green, London.

Provine, W.B., 1986. Sewall Wright and Evolutionary Biology. University of Chicago Press, Chicago.

Wright, S., 1917. The average correlation within subgroups of a population. Journal of the Washington Academy of Sciences 7, 532–535.

Wright, S., 1921. Correlation and causation. Journal of Agricultural Research 20, 557–585.

Wright, S., 1922a. Coefficients of inbreeding and relationship. American Naturalist 56, 530–538.

Wright, S., 1922b. The effects of inbreeding and crossbreeding in guinea pigs. Bulletin 1090, 1–63; 1091, 1–60.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–139.

Wright, S., 1968-78. Evolution and the Genetics of Populations, vol. 4. University of Chicago Press, Chicago.

Wright, S., 1988. Surfaces of selective value revisited. American Naturalist 131, 115–123.

Writing, Evolution of

Denise Schmandt-Besserat, University of Texas, Austin, TX, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Writing – a system of graphic marks representing the units of a specific language – was invented independently in the Near East, China, and Mesoamerica. The cuneiform script, created in Mesopotamia, present-day Iraq, c. 3200 BC, was the first form of writing. It is also the only writing system that can be traced to its earliest prehistoric origin. This antecedent of the cuneiform script was a system of counting and recording goods with clay tokens. The evolution of writing from tokens to pictography, syllabary (written symbols representing syllables), and alphabet illustrates the development of information processing to deal with larger amounts of data in ever greater abstraction.

The three writing systems that developed independently in the Near East, China, and Mesoamerica shared a remarkable stability. Each preserved over millennia features characteristic of their original prototypes. The Mesopotamian cuneiform script can be traced furthest back into prehistory to an eighth millennium BCE counting system using clay tokens of multiple shapes. The development from tokens to script reveals that writing emerged from counting and accounting. Writing was used exclusively for accounting until the third millennium BCE, when the Sumerian concern for the afterlife paved the way to literature by using writing for funerary inscriptions. The evolution from tokens to script also documents a steady progression in abstracting data, from one-to-one correspondence with three-dimensional tangible tokens, to two-dimensional pictures, the invention of abstract numbers and phonetic syllabic signs, and finally, in the second millennium BCE, the ultimate abstraction of sound and meaning with the representation of phonemes by the letters of the alphabet.

Writing is humankind's principal technology for collecting, manipulating, storing, retrieving, communicating, and disseminating information. Writing may have been invented independently three times in different parts of the world: in the Near East, China, and Mesoamerica. In what concerns this last script, it is still obscure how symbols and glyphs used by the Olmecs, whose culture flourished along the Gulf of Mexico c. 600–500 BCE, reappeared in the classical Maya art and writing of 250–900 CE as well as in other Mesoamerican cultures (Marcus, 1992). The earliest Chinese inscriptions, dated to the Shang Dynasty, c. 1400–1200 BCE, consist of oracle texts engraved on animal bones and turtle shells (Bagley, 2004). The highly abstract and standardized signs suggest prior developments, which are presently undocumented.

Of these three writing systems, therefore, only the earliest, the Mesopotamian cuneiform script, invented in Sumer, present-day Iraq, c. 3200 BCE, can be traced without any discontinuity over a period of 10 000 years, from a prehistoric antecedent to the present-day alphabet. Its evolution is divided into four phases: (1) clay tokens representing units of goods were used for accounting (8000–3500 BCE); (2) the three-dimensional tokens were transformed into two-dimensional pictographic signs, and like the former tokens, the pictographic script served exclusively for accounting (3500–3000 BCE); (3) phonetic signs, introduced to transcribe the name of individuals, marked the turning point when writing started emulating spoken language and, as a result, became applicable to all fields of human experience (3000–1500 BCE); and (4) with two dozen letters, each standing for a single sound of voice, the alphabet perfected the rendition of speech. After ideography, logography, and syllabaries, the alphabet represents a further segmentation of meaning.

Tokens as Precursor of Writing

The direct antecedent of the Mesopotamian script was a recording device consisting of clay tokens of multiple shapes (Schmandt-Besserat, 1996). The artifacts, mostly of geometric forms, such as cones, spheres, disks, cylinders, and ovoids, have been recovered from archaeological sites dating 8000–3000 BCE (**Figure 1**). The tokens, used as counters to keep track of goods, were the earliest code – a system of signs for transmitting information. Each token shape was semantic, referring to a particular unit of merchandise. For example, a cone and a sphere stood, respectively, for a small and a large measure of grain, and ovoids represented jars of oil. The repertory of some 300 types of counters made it feasible to manipulate and store information on multiple categories of goods (Schmandt-Besserat, 1992).

The token system had little in common with the spoken language except that, like a word, a token stood for one concept. Unlike speech, tokens were restricted to one type of information only, namely real goods. Unlike spoken language,

Figure 1 Envelope, tokens, and corresponding markings from Susa, Iran. Courtesy Musée du Louvre, Département des Antiquités Orientales.

the token system made no use of syntax. That is to say, their meaning was independent of their placement order. Three cones and three ovoids, scattered in any way, were to be translated 'three baskets of grain, three jars of oil.' Furthermore, the fact that the same token shapes were used in a large area of the Near East, where many dialects would have been spoken, shows that the counters were not based on phonetics. Therefore, the goods they represented were expressed in multiple languages. The token system showed the number of units of merchandise in one-to-one correspondence, in other words, the number of tokens matched the number of units counted: x jars of oil were represented by x ovoids. Repeating 'jar of oil' x times in order to express plurality is unlike spoken language.

Pictography: Writing as Accounting Device

After four millennia, the token system led to writing. The transition from counters to script took place simultaneously in Sumer and Elam, present-day western Iran when, around 3500 BCE, Elam was under Sumerian domination. It occurred when tokens, probably representing a debt, were stored in envelopes until payment. These envelopes made of clay in the shape of a hollow ball had the disadvantage of hiding the tokens held inside. Some accountants, therefore, impressed the tokens on the surface of the envelope before enclosing them inside, so that the shape and number of counters held inside could be verified at all times (Figure 1). These markings were the first signs of writing. The metamorphosis from three-dimensional artifacts to two-dimensional markings did not affect the semantic principle of the system. The significance of the markings on the outside of the envelopes was identical to that of the tokens held inside.

About 3200 BC, once the system of impressed signs was understood, clay tablets – solid cushion-shaped clay artifacts bearing the impressions of tokens – replaced the envelopes filled with tokens. The impression of a cone and a sphere token, representing measures of grain, resulted, respectively, in a wedge and a circular marking, which bore the same meaning as the tokens they signified (Figure 2). They were ideograms – signs representing one concept. The impressed tablets continued to be used exclusively to record quantities of goods received or disbursed. They still expressed plurality in one-to-one correspondence.

Pictographs – signs representing tokens traced with a stylus rather than impressed – appeared about 3100 BCE. These pictographs referring to goods mark an important step in the evolution of writing because they were never repeated in one-to-one correspondence to express numerosity. Besides them, numerals – signs representing plurality – indicated the quantity of units recorded. For example, '33 jars of oil' were shown by the incised pictographic sign 'jar of oil,' preceded by three impressed circles and three wedges, the numerals standing, respectively, for '10' and '1' (Figure 3). The symbols for numerals were not new. They were the impressions of cones and spheres formerly representing measures of grain, which then had acquired a second, abstract, and numerical meaning. The invention of numerals meant a considerable economy of signs since 33 jars of oil could be written with 7 rather than 33 markings.

In sum, in its first phase, writing remained mostly a mere extension of the former token system. Although the tokens underwent formal transformations from three- to two-dimensional and from impressed markings to signs traced with a stylus, the symbolism remained fundamentally the same. Like the archaic counters, the tablets were used exclusively for accounting (Nissen and Heine, 2009). This was also the case when a stylus, made of a reed with a triangular end, gave to the signs the wedge-shaped 'cuneiform' appearance (Figure 4). In all these instances, the medium changed in form but not in content. The only major departure from the token system consisted in the creation of two distinct types of signs: incised pictographs and impressed numerals. This combination of signs initiated the semantic division between the item counted and number.

Logography: Shift from Visual to Aural

About 3000 BCE, the creation of phonetic signs – signs representing the sounds of speech – marks the second phase in the evolution of Mesopotamian writing, when, finally, the medium parted from its token antecedent in order to emulate spoken language. As a result, writing shifted from a conceptual framework

Figure 2 Impressed tablet featuring an account of grain, from Godin Tepe, Iran. Courtesy Dr Cuyler Young, T., Royal Ontario Museum, Toronto.

Figure 3 Pictographic tablet featuring an account of 33 measures of oil, from Godin Tepe, Iran. Courtesy Dr Cuyler Young, T., Royal Ontario Museum, Toronto.

Figure 4 Economic cuneiform tablet. Courtesy Texas Memorial Museum, The University of Texas at Austin.

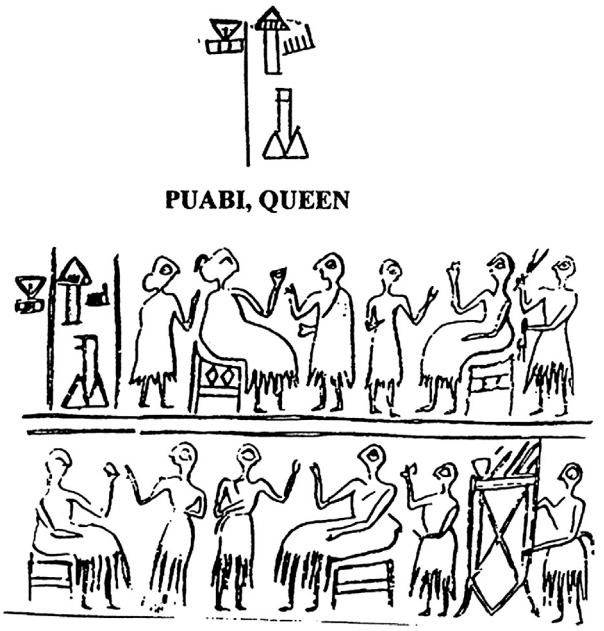

Figure 5 Name and title of Puabi carved on a seal recovered in the Royal Cemetery of Ur (U10939) (Source: Pierre Amiet, 1980. La Glyptique Mésopotamienne Archaique. Editions du CNRS, Paris, Pl. 90: 1182).

of real goods to the world of speech sounds (Malafouris, 2010). It shifted from the visual to the aural world.

With state formation, new regulations required that the names of the individuals who generated or received registered merchandise were entered on the tablets. The personal names were transcribed by the mean of logograms – signs representing a word in a particular tongue. Logograms were easily drawn pictures of words with a sound close to that desired (e.g., in English the name Neil could be written with a sign showing bent knees 'kneel'). Because Sumerian was mostly a monosyllabic language, the logograms had a syllabic value. A syllable is a unit of spoken language consisting of one or more vowel sounds, alone, or with one or more consonants. When a name required several phonetic units, they were assembled in a rebus fashion. A typical Sumerian name 'An Gives Life' combined a star, the logogram for An, god of heaven, and an arrow, because the words for 'arrow' and 'life' were homonyms. The verb was not transcribed, but inferred, which was easy because the name was common.

Phonetic signs allowed writing to break away from accounting. Inscriptions on stone seals or metal vessels deposited in tombs of the 'Royal Cemetery' of Ur, c. 2700–2600 BCE, are among the first texts that did not deal with merchandise, did not include numerals, and were entirely phonetic (Schmandt-Besserat, 2007). The inscriptions consisted merely of a personal name, 'Meskalamdug,' or a name and a title, 'Puabi, Queen' (Figure 5). Presumably, these funerary texts were meant to immortalize the name of the deceased, thereby, according to Sumerian creed, ensuring them of eternal life. Other funerary inscriptions further advanced the emancipation of writing. For example, statues depicting the features of an individual bore increasingly longer inscriptions. After the name and title of the deceased followed patronymics, the name of a temple or a god to whom the statue was dedicated, and in some cases, a plea for life after death, including a verb. These inscriptions introduced syntax, thus bringing writing yet one step closer to speech.

After 2600–2500 BCE, the Sumerian script became a complex system of ideograms mixed more and more frequently with phonetic signs. The resulting syllabary – system of phonetic signs expressing syllables – further modeled writing on to spoken language (Rogers, 2005). With a repertory of about 400 signs, the script could express any topic of human endeavor. Some of the earliest syllabic texts were royal inscriptions, and religious, magic, and literary texts.

The second phase in the evolution of the Mesopotamian script, characterized by the creation of phonetic signs, not only resulted in the parting of writing from accounting but also its spreading out of Sumer to neighboring regions. The first Egyptian inscriptions, dated to the late fourth millennium BCE, belonged to royal tombs (Baines, 2007). They consisted of ivory labels and ceremonial artifacts such as maces and palettes bearing personal names, written phonetically as a rebus, visibly imitating Sumer. For example, the Palette of Narmer bears hieroglyphs identifying the name and title of the Pharaoh, his attendants, and the smitten enemies. Phonetic signs to transcribe personal names, therefore, created an avenue for writing to spread outside of Mesopotamia. This explains why the Egyptian script was instantaneously phonetic. It also explains why the Egyptians never borrowed Sumerian signs. Their repertory consisted of hieroglyphs representing items familiar in the Egyptian culture that evoked sounds in their own tongue.

The phonetic transcription of personal names also played an important role in the dissemination of writing to the Indus Valley where, during a period of increased contact with Mesopotamia, c. 2500 BCE, writing appears on seals featuring individuals' names and titles (Parpola, 1994). In turn, the Sumerian cuneiform syllabic script was adopted by many Near

Eastern cultures who adapted it to their different linguistic families and in particular, Semitic (Akkadians and Eblaites); Indo-European (Mitanni, Hittites, and Persians); Caucasian (Hurrians and Urartians); and finally, Elamite (Black, 2008) and Kassite. It is likely that Linear A and B, the phonetic scripts of Crete and mainland Greece, c. 1400–1200 BCE, were also influenced by the Near East.

The Alphabet: The Segmentation of Sounds

The invention of the alphabet about 1500 BCE ushered in the third phase in the evolution of writing in the ancient Near East (Sass, 2005). The first, so-called Proto-Sinaitic or Proto-Canaanite alphabet, which originated in the region of present-day Lebanon, took advantage of the fact that the sounds of any language are few. It consisted of a set of 22 letters, each standing for a single sound of voice, which, combined in countless ways, allowed for an unprecedented flexibility for transcribing speech (Powell, 2009). This earliest alphabet was a complete departure from the previous syllabaries. First, the system was based on acrophony – signs to represent the first letter of the word they stood for – for example, an ox head (*alpu*) was 'a,' a house (*betu*) was 'b' (Figure 6). Second, it was consonantal – dealing only with speech sounds characterized by constriction or closure at one or more points in the breath channel, like b, d, l, m, n, and p. Third, it streamlined the system to 22 signs, instead of several hundreds.

The transition from cuneiform writing to the alphabet in the ancient Near East took place over several centuries. In the seventh century BCE, the Assyrian kings still dictated their edicts to two scribes. The first wrote Akkadian in cuneiform on a clay tablet, and the second Aramaic in a cursive alphabetic script traced on a papyrus scroll. The Phoenician merchants established on the coast of present-day Syria and Lebanon, played an important role in the diffusion of the alphabet. In particular, they brought their consonantal alphabetic system to Greece, perhaps as early as, or even before, 800 BCE. The Greeks perfected the Semitic alphabet by adding letters for vowels – speech sounds in the articulation of which the breath channel is not blocked, like a, e, i, o, u. As a result, the 27-letter Greek alphabet improved the transcription of the spoken word, since all sounds were indicated. For example, words sharing the same consonants like 'bad,' 'bed,' 'bid,' and 'bud,' could be clearly distinguished. The alphabet did not subsequently undergo any fundamental change (Solomon, 2012).

The Modern Alphabets

Because the alphabet was invented only once, all the many alphabets of the world, including Latin, Arabic, Hebrew, Amharic, Brahmani, and Cyrillic, derive from Proto-Sinaitic. The Latin alphabet used in the Western world is the direct descendant of the Etruscan alphabet (Bonfante and Bonfante, 2002). The Etruscans, who occupied the present province of Tuscany in Italy, adopted the Greek alphabet, slightly modifying the shape of letters. In turn, the Etruscan alphabet became that of the Romans, when Rome conquered Etruria in

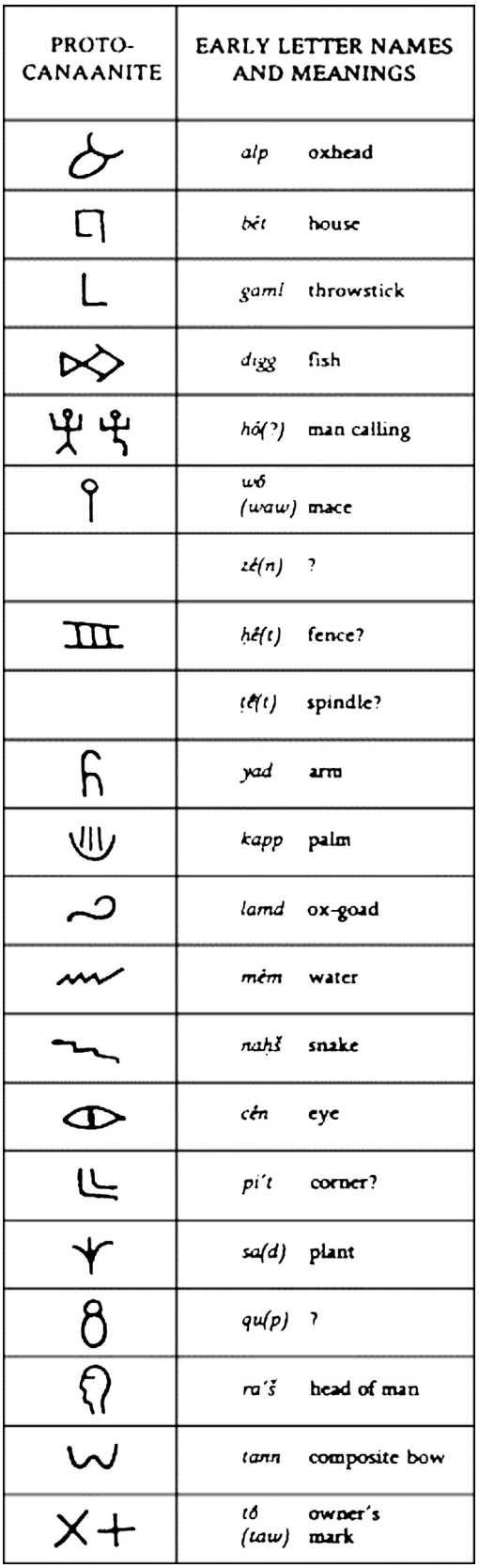

PROTO-CANAANITE	EARLY LETTER NAMES AND MEANINGS	
	alp	oxhead
	bét	house
	gaml	throwstick
	digg	fish
	hó(?)	man calling
	wó (waw)	mace
	zé(n)	?
	ḥé(t)	fence?
	ṭé(t)	spindle?
	yad	arm
	kapp	palm
	lamd	ox-goad
	mém	water
	naḥš	snake
	cén	eye
	pi't	corner?
	sa(d)	plant
	qu(p)	?
	ra'š	head of man
	tann	composite bow
	tó (taw)	owner's mark

Figure 6 Proto-Sinaitic alphabet (Source: Michael Roaf, 1990. Cultural Atlas of Mesopotamia. Equinox, Oxford, p. 150).

the first century BCE. The alphabet followed the Roman armies. All the nations that fell under the rule of the Roman Empire became literate in the first centuries of our era. This was the case for the Gauls, Angles, Saxons, Franks, and Germans who inhabited present-day France, England, and Germany.

Charlemagne (800 CE) had a profound influence on the development of the Latin script by establishing standards. In particular, a clear and legible minuscule cursive script was devised, from which our modern-day lower case derives. The printing press invented in 1450 dramatically multiplied the dissemination of texts, introducing a new regularity in lettering and layout. The Internet catapults the alphabet into cyberspace, while preserving its integrity.

Writing: Handling Data in Abstraction

Beyond the formal and structural changes undergone by writing in the course of millennia, its evolution also involved strides in the ability to handle data in abstraction. At the first stage, the token system antecedent of writing, already abstracted information in several ways. First, it translated daily-life commodities into arbitrary, often geometric, forms. Second, the counters abstracted the items counted from their context. For example, sheep could be accounted independently of their actual location. Third, the token system separated the data from the knower. That is to say, a group of tokens communicated directly specific information to anyone initiated in the system. This was a significant change for an oral society, where knowledge was transmitted by word of mouth from one individual to another, face to face. Otherwise, the token system represented plurality concretely, in one-to-one correspondence. Three jars of oil were shown by three tokens, as it is in reality. At the same time, the fact that the token system used specific counters to count different items was concrete – it did not abstract the notion of item counted from that of number. (Certain English numerical expressions referring to particular sets, such as twin, triplet, quadruplet and duo, trio or quartet, are comparable to concrete numbers.)

When tokens were impressed on the envelopes to indicate the counters enclosed inside, the resulting markings could no longer be manipulated by hand. In other words, the transmutation of three-dimensional counters into two-dimensional signs constituted a second step in abstraction. By doing away with tokens, the clay tablets marked a third level of abstraction since the impressed markings no longer replicated a set of actual counters. The invention of numerals, which separated the notion of numerosity from that of the item counted, was a crucial fourth step in abstraction. The signs expressing the concept of oneness, twoness, etc. allowed plurality to be dealt with in fully abstract terms. In turn, the phonetic units marked a fifth step of abstraction, since the signs no longer referred to the objects pictured, but rather the sound of the word they evoked.

Phonetics allowed writing to shift from a representational to a conceptual linguistic system. That is to say, it enabled writing to leave the realm of real goods in order to enter the world of words and the ideas they stand for. Finally, the process that started with ideograms expressing concepts and phonetic signs referring to the sound of monosyllabic words reached the ultimate segmentation of meaning with letters. As Marshall McLuhan (Moos, 1997) defined it, the alphabet consists of semantically meaningless letters corresponding to semantically meaningless sounds. The alphabet brought data handling to a final double-stepped abstraction.

Conclusion: The Stability of Writing Systems

The origin of the Chinese script and the development of Mesoamerican writing are still obscure. The Mesopotamian script, however, offers a well-documented evolution over a continuous period of 10 000 years. The system underwent drastic changes in form, gradually transcribed spoken language more accurately, and handled data in more abstract terms. The most striking universal feature of all writing systems, however, is their uncanny endurance, unmatched among human creations. The Chinese script never needed to be deciphered because the signs have changed little during the 3400 years of its recorded existence (Xigui, 2000). It also always remained ideographic, merely inserting rebus-like phonetic complements in some characters. The Mesoamerican Maya phonetic glyphs preserved the symbolism initiated by the Olmecs in the previous millennium (Coe and van Stone, 2005). Finally, when the last clay tablet was written in the Near East, c. 300 CE, the cuneiform script had been in use for three millennia. It replaced an age-old token system that had preceded it for over 5000 years; it was replaced by the alphabet, which we have now used for 3500 years.

See also: Brain, Evolution of; Chinese Literacy Acquisition: A Multidimensional Puzzle; Cultural Evolution: Overview; Evolution and Language: Overview; Number Systems, Evolution of; Semantic Similarity, Cognitive Psychology of; Spelling, Psychology of; Visual Word Recognition, Neurocognitive Psychology of; Writing Instruction; Writing Systems, Psychology of; Writing Systems.

Bibliography

Bagley, R.W., 2004. Anyang writing and the origin of the Chinese writing system. In: Houston, S.D. (Ed.), The First Writing. Cambridge University Press, Cambridge, pp. 190–249.

Baines, J., 2007. Visual and Written Culture in Ancient Egypt. Oxford University Press Oxford, Cambridge University Press, Cambridge.

Black, J., 2008. The obsolescence and demise of the cuneiform writing in Elam. In: Baines, J., Bennet, J., Houston, S. (Eds.), The Disappearance of Writing Systems. Equinox, London, pp. 45–72.

Bonfante, G., Bonfante, L., 2002. The Etruscan Language, revised ed. Manchester University Press, Manchester.

Coe, M.D., van Stone, M., 2005. Reading the Maya Glyphs. Thames and Hudson, London.

Malafouris, L., 2010. Grasping the concept of number: how did the sapient mind move beyond approximation. In: Morley, I., Renfrew, C. (Eds.), The Archaeology of Measurement. Cambridge University Press, Cambridge, pp. 35–42.

Marcus, J., 1992. Mesoamerican Writing Systems. Princeton University Press, Princeton, NJ.

Moos, M.A. (Ed.), 1997. Marshall McLuhan. Essays. In: Media Research. Overseas Publishers Association, Amsterdam.

Nissen, H.J., Heine, P., 2009. From Mesopotamia to Iraq. University of Chicago Press, Chicago, IL.

Parpola, A., 1994. Deciphering the Indus Script. Cambridge University Press, Cambridge.

Powell, B.B., 2009. Writing: Theory and History of the Technology of Civilization. Wiley Blackwell, London.

Rogers, H., 2005. Writing Systems, A Linguistic Approach. Blackwell, London.

Salomon, R., 2012. Some principles and patterns of script change. In: Houston, S.D. (Ed.), The Shape of Script. Sar Press, Santa Fe, NM, pp. 119–133.

Sass, B., 2005. The Alphabet at the Turn of the Millennium, The West Semitic Alphabet ca. 1150–850 BC – The Antiquity of the Arabian, Greek and Phrygian Alphabets. Tel Aviv University, Tel Aviv.

Schmandt-Besserat, D., 2007. When Writing Met Art. University of Texas Press, Austin, TX.

Schmandt-Besserat, D., 1996. How Writing Came About. University of Texas Press, Austin, TX.

Schmandt-Besserat, D., 1992. Before Writing, vols. 2. University of Texas Press, Austin, TX.

Xigui, Q., 2000. Chinese Writing, The Institute of East Asian Studies. The University of California, Berkeley, CA.

Writing Instruction

Steve Graham, Mary Lou Fulton Teachers College, Arizona State University, Tempe, AZ, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D. Galbraith, volume 24, pp. 16625–16629, © 2001, Elsevier Ltd.

Abstract

This article presents evidence-based practices (EBPs) for teaching writing to school-aged students. These practices were drawn from over 20 meta-analyses examining true- and quasi-experiments testing specific writing interventions. The EBPs are presented within the context of two primary theories that have been influential in shaping our understanding of writing and writing development. The social/contextual view of writing emphasizes the importance of context in writing development, including the role of social, cultural, historical, and institutional factors. The cognitive/motivational view stresses the importance of developing students' writing skills, strategies, knowledge, and motivation. Recommendations and caveats for applying the evidence-based writing practices identified here are examined.

In today's world, writing is essential to success. At school, teachers commonly use writing to assess and promote students' learning (National Council of Teachers of English, 2004). At work, white- and blue-collar workers increasingly use writing as part of their job (National Commission on Writing in America's Schools and Colleges, 2004, 2005, 2006). At home, writing is now part of everyday life thanks to e-mailing, blogging, texting, and other forms of digital composing (Hillocks, 2006).

While writing is widely used, there has been a long-standing concern about the quality of writing instruction students receive at school (e.g., Sheils, 1975). This concern has its genesis in two related findings. First, many students do not develop strong writing skills. Even in an affluent country like the United States, only about one-third of students become proficient writers (National Center for Educational Statistics, 2012). Second, little time is devoted to teaching writing in most schools (Applebee and Langer, 2011; Gilbert and Graham, 2010; Kiuhara et al., 2009), and teachers indicate that they are not adequately prepared to teach writing (Brindle, 2012).

This is unfortunate, as writing is an extremely complex skill that is not easily mastered (Graham, 2006a). Good writing depends on the ability to regulate a variety of processes, strategies, and mental operations when planning, drafting, evaluating, and revising text. The creation of well-written text further involves accessing different kinds of knowledge, such as knowledge about the topic, the intended audience, and genre. Writers must apply transcription and translation skills such as handwriting, typing, spelling, and sentence construction to realize their intentions. Because writing is a self-directed task, it also involves engagement and persistence, and it is shaped the writers' motivational state, as attributes such as efficacy, anxiety, and attitudes can enhance or impede its execution. Moreover, writers must intelligently apply these resources across situations, occasions, and contexts over time.

One approach for addressing the quality gap in writing instruction in schools is through the use of evidence-based practices (EBPs; Cook et al., 2012). These are "instructional approaches shown through high-quality research to result in generally improved student outcomes" (Cook et al., 2012: p. 495). This is the approach adopted here, as I draw on 21 past meta-analyses of writing interventions studies employing true or quasi-experimental designs (Andrews et al., 2006; Bangert-Drowns, 1993; Bangert-Drowns et al., 2004; Gersten and Baker, 2001; Goldberg et al., 2003; Gillespie and Graham, 2014; Graham, 2006b; Graham et al., 2012a; Graham and Harris, 2003; Graham et al., 2011, 2013b; Graham and Hebert, 2011; Graham et al., 2012b; Graham and Perin, 2007a,b; Graham and Santangelo, 2014; Graham et al., in press; Hillocks, 1986; Morphy and Graham, 2012; Sandmel and Graham, 2011, 2013) to identify EBPs for writing. In true-experiments, students are randomly assigned to the writing treatment and control condition. Such randomization does not occur in quasi-experiments.

With meta-analysis the magnitude and the direction (positive or negative) of the effects of a writing practice are averaged across all pertinent studies. I share here writing EBPs that resulted in positive and statistically significant effects in improving the quality of students' writing across studies testing a specific writing intervention. We present these EBPs within the context of a view of writing development that emphasizes both the social/contextual as well as cognitive/motivational aspects of writing (Graham et al., in press).

Writing Development

Our understanding of how writing develops is incomplete, but enough is known to be certain that the road to competence depends on the context in which writing takes place as well as what the writer brings to the task cognitively and motivationally (Graham, 2006b). The work of Russell (1997) aptly illustrates the social/contextual view of writing development. He emphasizes the importance of two structures in his model of writing. One is *activity system*, which specifies how actors (a student, pair of students, student and teacher, or class perceived in social terms and taking into account the history of their involvement in the activity system) use concrete tools, such as paper and pencil or word processing, to accomplish an action leading to an outcome. For instance, a desired outcome, such as the collaborative creation of a text explaining the idea of balance, is

accomplished in a problem space where the actors use writing tools in an ongoing interaction with others (peers and teachers) to shape the text that is being produced over time in a shared direction.

The second structure in Russell's (1997) model is *genre*: a "typified way of purposefully interacting in and among some activity system(s)" (p. 513). These genres become stabilized via regularized use by and among students, creating a predictable approach for writing within a classroom (e.g., plan, draft, revise, edit, and publish). Genres are conceived as temporarily stabilized structures, however, because they are subject to change depending on the context. For example, a new student entering a classroom with established and stabilized genes (e.g., Edit before publishing your paper) will likely appropriate some of the routinized practices of his classmates. The youngster may, in turn, create changes in existing genres by applying writing routines learned at his old school (e.g., Get feedback on your writing plan before editing) and deemed as valuable by his new teacher, classmates, or both.

In contrast to the social/contextual view of writing, the cognitive/motivational approach concentrates primarily on the individual writer and the mental and affective processes involved in composing text, as illustrated in a model of skilled writing developed by Hayes (1996). He identified the mental moves, cognitive processes, and motivational resources writers draw on as they write. This includes mental processes such as text interpretation, reflection, and text production that writers use to create a representation of the writing task; develop a plan to complete it; draw conclusions about the audience and possible writing content; use cues from the writing plan or text produced so far to retrieve needed information from memory; turn these ideas and information into written sentences; and evaluate plans and text and modify them as needed. It also includes long-term memory which contains a variety of resources, such as topic knowledge and schemas, for carrying out particular writing tasks, as well as working memory which serves as a temporary place for holding ideas for writing and carrying out mental operations that require the writer's conscious attention. The application of these mental operations and cognitive processes is influenced by a variety of motivational factors, as writers' goals, predispositions, beliefs, and attitudes shape what they do.

Together, the social/contextual and cognitive/motivational viewpoints provide needed structure for designing an effective writing program. One, it is important to create a writing context that is supportive. This includes developing classroom routines that make writing an interesting and enjoyable activity as well as using instructional procedures and writing tools that support students as they write. Two, it is equally important to make sure students acquire the skills, strategies, and knowledge they need to become skilled writers. EBPs for obtaining these goals (based on the meta-analyses referenced earlier), including the use of writing as a tool for learning, are presented next.

Create a Supportive Writing Context

As students develop as writers, they learn to write for many purposes (e.g., entertain, persuade, learn, inform, communicate), while composing in a planful, thoughtful, reflective, and collaborative manner, through the use of multiple technologies (e.g., pen and paper, word processing) and modalities (written text combined with visual and auditory material). Such applications of writing are not only demanding mentally, but require learning and working with others in a productive manner. This makes it especially important to develop a writing environment where teachers support students and their writing efforts and students support each other in a positive and motivating manner.

One EBP, the process approach to writing (see Sandmel and Graham, 2011), includes a number of mechanisms for creating a supportive writing context. This includes creating a pleasant and positive writing environment; writing for real audiences and authentic purposes; creating routines that emphasize what students need to do as they write (i.e., plan, draft, evaluate, revise, and edit); stressing personal responsibility and ownership of writing projects; encouraging self-reflection and evaluation; promoting high levels of student interactions to support writing; and providing students with personalized individual assistance and feedback, as well as brief instructional lessons, as needed.

This general approach to creating a supportive writing environment can be complemented by the application of other more specific EBPs, some of which are instantiations of the practices included in the process approach to writing. One, students' writing can be strengthened by creating routines that ensure they write frequently (e.g., Graham et al., 2012b). The basic premise underlying this EBP is that students must write frequently if they are to prosper and grow as writers. This allows developing writers to become comfortable with writing and further hone their writing skills.

Two, students are likely to produce better text when the goals for their writing are clearly established (e.g., Graham et al., 2012a). Goals work best when they are clear, specific, and reasonably challenging. For instance, an appropriate goal for 10 to 12-year-old students might include: addressing both sides of an argument, providing three or more reasons to support your point of view and countering at least two reasons supporting the opposing view when writing an argument.

Three, developing a supportive writing environment requires some consideration of the tools students use to compose. Many schools just use pencil and paper as the writing medium, instead of more modern writing tools, such as a word processor and other digital writing tools which may make it possible to compose through multiple modalities. These tools have many advantages over paper and pencil (e.g., Bangert-Drowns, 1993; Morphy and Graham, 2012). For instance, electronic text produced via word processing is legible, and it can easily be deleted, added, rewritten, or moved. Other software which support writers, such as spell checkers or speech synthesis, may be included as part of this writing tool. Word processor can be connected to the Internet and other programs where students can gather material for what they write, as well as share their text with others.

Four, writing development can also be facilitated by having students work together as they compose (e.g., Graham and Perin, 2007b). This includes working together to plan, draft, revise, and edit their compositions. A common characteristic of studies that tested this collaborative approach was that students were provided with specific directions for what to do when working together and then taught how to carry out these

procedures. To illustrate, in one study (Yarrow and Topping, 2001), teachers applied a systematic instructional routine (i.e., exploration, demonstration, clarification, practice, monitoring, coaching, and reinforcement) to teach students how to successfully work with a peer as they planned, drafted, revised, and edited their papers.

Five, students' writing development can be enhanced through formative assessment practices (Graham et al., 2011). When teachers monitor students' progress as writers, they can adjust classroom practices to meet the collective as well as the individual needs of their students. When students evaluate their own writing, writing improves. When students receive feedback about their writing, writing improves. When students are partners in writing assessment, giving and receiving peer feedback, students' writing improves.

Help Developing Writers Acquire Needed Skills, Knowledge, and Strategies

Skilled writing requires mastery or facility with a wide range of skills (e.g., handwriting, typing, spelling, grammar, and usage), knowledge (e.g., attributes of specific types of text such as persuasive, informative, and narrative), and strategies (e.g., planning, drafting, revising, and editing text). How can developing writers acquire competence in these areas? It is clear, based on the accumulated evidence from the available meta-analyses, that explicitly teaching writing skills, knowledge, and strategies is advantageous, as such instruction makes students better writers (Hillocks, 1986; Graham et al., 2012b; Graham and Perin, 2007b). While most of the EBPs presented next involve explicit instruction, it is important to keep in mind that students also acquire skills, knowledge, and skills through less direct practices too. For example, students learn much of what they need to know about spelling as a result of reading and writing (Graham, 2000). The challenge is to know what combination of explicit and implicit instructional practices should be applied when?

Teach Foundational Writing Skills to Beginning Writers

Skilled writers rarely think about handwriting, typing, or spelling when they compose. These skills are executed correctly and with little or no conscious attention. Such mastery is important, as having to devote conscious attention to these skills can interfere with other writing processes (Scardamalia and Bereiter, 1986). Explicitly teaching such skills to elementary grade students improves the overall quality of their writing (Graham et al., in press), with the added bonus that teaching spelling improves students' reading skills too (Graham and Santangelo, 2014).

Writers invest considerable energy into transforming their ideas into the words and syntactic structures that convey their intended meanings. This includes constructing sentences as well as using appropriate grammar, punctuation, capitalization, and so forth. An EBP for improving such skills is sentence-combining (Andrews et al., 2006; Hillocks, 1986). This involves the teacher modeling how to combine two or more sentences into a more complex one, followed by student practice combining similar types of sentences.

Increase Developing Writers' Knowledge for and about Writing

Three types of knowledge that developing writers need to acquire are knowledge about their writing topic, knowledge about genre, and knowledge about the words they use to convey meaning (Olinghouse et al., 2012; Olinghouse and Wilson, 2013). An EBP for addressing the issue of content knowledge is for teachers to put into place practices that help students gather ideas and information to write about (e.g., Graham and Perin, 2007a). Examples of such activities include locating possible writing content through brainstorming, reading, drawing, the Internet, and/or by using graphic organizers.

An EBP for acquiring genre knowledge is to teach students about the basic characteristics of specific types of text (Graham et al., 2012b). For example, teachers and students might read a series of stories together in order to define and identify the basic elements (e.g., setting, episode) included in such narratives. A second EBP for acquiring genre knowledge is to provide students with model text for a specific type of writing and ask them to emulate it (e.g., Hillocks, 1986).

Although not as well studied as most other EBPs, teaching students specialized vocabulary to use in their writing can improve what students' write (Graham et al., in press). For example, teaching students words about outer space can improve what they write about this topic. When writers use such vocabulary correctly in their compositions, they show their understanding of the topic and express their ideas in a way that is likely to be valued and accepted by their readers, including their teacher (Olinghouse and Wilson, 2013).

Teach Developing Writers Strategies for Planning, Drafting, Revising, and Editing

Skilled writers use strategies for planning, drafting, revising, and editing to help them create and improve what they write (Zimmerman and Reisemberg, 1997). These strategies range from more general strategies such as brainstorming, semantic Webbing, or editing which can be applied across genres to strategies that are designed for specific types of writing. For instance, when writing a story, the writer may generate possible ideas for writing by thinking about the characters, the setting, the main character's goals, action to achieve those goals, characters' reactions, and how the story ends (see Graham and Harris, 2005; Harris et al., 2008 for other examples).

When developing writers are taught strategies for planning, drafting, revising, and editing the overall quality of their text improves (Graham et al., 2012b). Strategy instruction involves teachers explaining the purpose and rationale of the strategy, modeling how to use it, and providing students with assistance in applying the strategy with the goal of independent and effective application. Such instruction is improved, however, when students are also taught the knowledge or skills they need to effectively use the target writing strategy, made aware of the

gains that result through its use, attribute these gains to effort and strategy use, and set goals to learn and apply the new strategy (see Harris et al., 2008).

Use Writing to Support Students' Reading and Learning

Writing about material read, presented in class, or obtained through other means may be advantageous for multiple reasons (Graham and Hebert, 2011). Writing about such material requires that the writer decides what ideas are most important and how they are related one to another. The permanence of writing makes it possible for the writer to review, reexamine, critique, and even construct new understandings of the extracted material. When writers put these ideas into their own written language it can make them think more carefully about what the ideas mean.

The available evidence indicates that when students write about material they are learning in class their learning increases (Bangert-Drowns et al., 2004; Graham and Perin, 2007b). Likewise, writing about material read enhances students' understanding and retention of such material (Graham and Hebert, 2011). This was true across disciplines as varied as the language arts, science, social studies, and mathematics. Example of writing activities that promoted student learning included writing short answers to questions, taking notes, writing a summary, and writing more extended responses such as explaining how something learned can be applied or relates to one's own life.

Students with Special Writing Needs

Learning to write can represent a major challenge for students with a disability, as many of these youngsters write well below grade level (National Center for Educational Statistics, 2012). While we know less about how to teach writing effectively to students with disabilities than to their classmates without disabilities, enough writing intervention research has been conducted with students with learning disabilities that five EBPs have been identified (Gersten and Baker, 2001; Gillespie and Graham, 2014; Graham et al., in press; Morphy and Graham, 2012).

Four of the five identified EBPs for students with learning disabilities were also EBPs for developing writers in general. These included implementing a process approach for writing instruction; establishing specific, challenging, but realistic goals for students' writing; teaching strategies for planning, drafting, revising, and using word processing as the primary medium for composing. This last EBP was especially effective when other software, which provided support for planning and revising was included as part of the program. Such word-processing programs are less likely to be effective, however, if students' typing skills are slow (MacArthur and Graham, 1987).

The fifth EBP for students with learning disabilities was dictation. This primarily involved students dictating some part of their composition (e.g., plan, draft) into a tape recorder, which was later transcribed by the student or someone else. There are obvious limitations to dictation, especially when someone else must transcribe it, which limits its implementation. Speech-to-text synthesis word-processing programs provide a workable alternative, but they do not yet provide fully accurate transcription, especially for younger developing writers.

Collectively, the intervention research conducted with students with learning disabilities demonstrates that teachers can improve the writing of students who experience considerable difficulty with writing and learning. It is imperative that these students, and all other students for that matter, receive the best writing instruction available in order to maximize their writing growth.

Caveats

This article provides guidance on possible writing practices and interventions that teachers can apply to help students become better writers. It draws on scientific testing of writing interventions through true- and quasi-experiments to identify EBPs that improve the quality of students' writing across multiple studies. The advantage of such an approach is that it provides a relatively trustworthy method for identifying effective teaching practices in writing, as such studies provide evidence on whether the instructional procedure produced the desired impact, the observed effects are representative, and how much confidence can be placed in them (Graham, 2010).

It must be noted, however, that there are other sources of evidence for identifying effective practices for developing writers. These include the practices applied by highly effective teachers as well as practices identified through other forms of research, such as single subject design studies (Horner et al., 2005). These various forms of evidence are not incompatible with one another, and many of the EBPs presented here are commonly used by highly effective teachers (Graham and Perin, 2007c) and identified as effective using single subject research methodology (Rogers and Graham, 2008).

One challenge in implementing writing EBPs into everyday classroom practices involves teacher preparation. Many teachers, but especially content area teachers, indicate that their preparation to teach writing is inadequate (e.g., Kiuhara et al., 2009). If the EBPs identified here are to become a positive and routine aspect of classroom life, then the professional development teachers receive in college, through their schools or school district, from their profession, and through their own efforts must provide them with the knowledge and skills needed to implement them effectively.

A second challenge in implementing the practices identified here is that we do not know what combination of writing EBPs is most effective. This is likely to depend on a variety of factors including who the students are, where they are at in their development, the setting in which instruction takes place, and the expertise of the teacher. This challenge should not be underestimated, as a student's writing development is uneven across different types of writing (Graham et al., 2011), and students likely benefit from different forms of support at different points in their development.

A third challenge to implementing writing EBPs revolves around the different organizational structures or formats for writing that exist in schools today. Writing instruction can occur within the context of the language arts classroom, a content area

classroom, or with a learning specialist, such as a special education teacher. These options are not exclusive, however, as they can be combined in various ways. For example, a language arts and a learning specialist might use subject matter, such as science or social studies, as a focal point for writing instruction. To illustrate, they might collaborate in teaching strategies for planning and writing persuasive essays using text students read in a concurrent social studies class. Furthermore, a social studies teacher and the language arts instructors could align their curricula, giving students assignments that systematically connect writing and content instruction. To illustrate, in a study by De La Paz (2005), a history teacher taught a historical reasoning strategy using historical documents, and the language arts teacher taught strategies for writing argumentation essays, using the documents applied by the history teacher.

The effectiveness of such various formats has neither been tested or compared one to another, so it is not certain how well the EBPs identified here would fair in these different formats. Before implementing a writing EBP, a careful analysis of the organizational structure or format within which it well be placed should be undertaken, with the aim of identifying factors that will facilitate and impede their effectiveness.

While research investigating writing instruction spans more than 100 years of study, it must be noted that there are many gaps in our knowledge about how to teach writing effectively. Some writing practices have not been scientifically tested at all and others have been tested sparingly. Over time, other effective writing practices and interventions will be identified. Thus, the EBPs presented here provide a starting point, not an end point, for developing effective programs for developing writers.

It should further be recognized that our knowledge about teaching writing effectively to all students is incomplete. For example, very little writing intervention research has been conducted with students learning to write in a second language. In fact, virtually none of the research included in the meta-analyses that formed the basis for the EBPs presented here was conducted with these developing writers. This is also the case for very young writers (3–5 years old) and students with disabilities other than learning disabilities. Likewise, we know much less about teaching writing to struggling writers than typically developing ones.

Finally, the use of writing EBPs likely hold their greatest promise for helping teachers and schools develop students who are skilled and confident writers if they are implemented as part of a comprehensive reform of writing instruction. To address the long-standing concerns about students' writing and the neglect of writing instruction in many classrooms, policy makers and the educational community need to develop new and better policies that establish unambiguous, challenging, and realistic plans for improving writing instruction and students' writing. These plans should specify methods and incentives to ensure that teachers are prepared to teach writing effectively; the teaching of writing is the responsibility of all teachers; students write frequently; students use writing as a tool to support learning across the curriculum; sufficient time is devoted to teaching of writing at all grades; and students increasingly use twenty-first century digital writing tools.

We encourage teachers to learn more about the writing EBPs presented here before applying them, and to use each practice judiciously with good common sense. Teachers should monitor the success of these practices as they apply them and make needed adjustments accordingly. Just because an intervention was effective in multiple research studies does not guarantee that it will be effective in all other situations. There is rarely, if ever, an exact match between the conditions in which the intervention was tested, and the conditions in which it is subsequently implemented by teachers. The safest course of action, therefore, is for teachers to monitor if the applied practices are working in her or his classroom.

See also: Instructional Psychology; Learning and Instruction: Social-Cognitive Perspectives; Literacy Education; Meta-Analysis, History of; Meta-Analysis: Theory; Systematic Reviewing and Meta-Analysis.

Bibliography

Andrews, R., Torgerson, C., Beverton, S., Freeman, A., Locke, T., Low, G., Robinson, A., Zhu, D., 2006. The effects of grammar teaching on writing development. British Educational Research Journal 32, 39–55.

Applebee, A., Langer, J., 2011. A snapshot of writing instruction in middle and high schools. English Journal 100, 14–27.

Bangert-Drowns, R.L., 1993. The word processor as an instructional tool: a meta-analysis of word processing in writing instruction. Review of Educational Research 63, 69–93.

Bangert-Drowns, R.L., Hurley, M.M., Wilkinson, B., 2004. The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis. Review of Educational Research 74, 29–58.

Brindle, M., 2012. Examining the relationships among teachers' preparation, efficacy, and writing practices. Unpublished Dissertation, Vanderbilt University, Nashville, TN.

Cook, B., Smith, G., Tankersley, M., 2012. Evidence-based practices in education. In: Harris, K.R., Graham, S., Urdan, T. (Eds.), APA Educational Psychology Handbook, vol. 1. American Psychological Association, Washington, DC, pp. 495–528.

De La Paz, S., 2005. Teaching historical reasoning and argumentative writing in culturally and academically diverse middle school classrooms. Journal of Educational Psychology 97, 139–158.

Gersten, R., Baker, S., 2001. Teaching expressive writing to students with learning disabilities. Elementary School Journal 101, 251–272.

Gillespie, A., Graham, S., 2014. A meta-analysis of writing interventions for students with learning disabilities. Exceptional Children 80, 454–473.

Gilbert, J., Graham, S., 2010. Teaching writing to elementary students in grades 4 to 6: a national survey. Elementary School Journal 110, 494–518.

Goldberg, A., Russell, M., Cook, A., 2003. The effect of computers on student writing: a meta-analysis of studies from 1992 to 2002. The Journal of Technology, Learning, and Assessment 2. Retrieved from: http://escholarship.bc.edu/jtla/vol2/1/.

Graham, S., 2000. Should the natural learning approach replace traditional spelling instruction. Journal of Educational Psychology 92, 235–247.

Graham, S., 2010. Teaching writing. In: Hogan, P. (Ed.), Cambridge Encyclopedia of Language Sciences. Cambridge University Press, Cambridge, UK, pp. 848–851.

Graham, S., 2006a. Writing. In: Alexander, P., Winne, P. (Eds.), Handbook of Educational Psychology. Erlbaum, Mahwah, NJ, pp. 457–478.

Graham, S., 2006b. Strategy instruction and the teaching of writing: a meta-analysis. In: MacArthur, C., Graham, S., Fitzgerald, J. (Eds.), Handbook of Writing Research. Guilford, New York, pp. 187–207.

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., Olinghouse, N., 2012a. Teaching Writing in Elementary School: A Practice Guide. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Graham, S., Gillespie, A., McKeown, D., 2013a. Writing: importance, development, and instruction. Reading & Writing: An Interdisciplinary Journal 26, 1–15.

Graham, S., Harris, K.R., 2005. Writing Better: Teaching Writing Processes and Self-regulation to Students with Learning Problems. Brookes, Baltimore, MD.

Graham, S., Harris, K.R., 2003. Students with learning disabilities and the process of writing: a meta-analysis of SRSD studies. In: Swanson, L., Harris, K.R., Graham, S. (Eds.), Handbook of Research on Learning Disabilities. Guilford, New York, pp. 383–402.

Graham, S., Harris, K.R., Hebert, M., 2011. Informing Writing: The Benefits of Formative Assessment. Alliance for Excellence in Education, Washington, DC.

Graham, S., Harris, K.R., McKeown, D., 2013b. The writing of students with LD and a meta-analysis of SRSD writing intervention studies: Redux. In: Swanson, L., Harris, K.R., Graham, S. (Eds.), Handbook of Learning Disabilities, second ed. Guilford Press, New York, pp. 405–438.

Graham, S., Hebert, M., 2011. Writing-to-read: a meta-analysis of the impact of writing and writing instruction on reading. Harvard Educational Review 81, 710–744.

Graham, S., McKeown, D., Kiuhara, S., Harris, K.R., 2012b. A meta-analysis of writing instruction for students in the elementary grades. Journal of Educational Psychology 104, 879–896.

Graham, S., Perin, D., 2007a. Writing Next: Effective Strategies to Improve Writing of Adolescent Middle and High School. Alliance for Excellence in Education, Washington, DC.

Graham, S., Perin, D., 2007b. A meta-analysis of writing instruction for adolescent students. Journal of Educational Psychology 99, 445–476.

Graham, S., Perin, D., 2007c. What we know, what we still need to know: teaching adolescents to write. Scientific Studies in Reading 11, 313–336.

Graham, S., Santangelo, T., 2014. Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. Reading & Writing: An Interdisciplinary Journal 27, 1703–1743.

Graham, S., Harris, K.R., Santangelo, T. Research-based writing practices and the common core: meta-analysis and meta-synthesis. Elementary School Journal, in press.

Harris, K.R., Graham, S., Mason, L., Friedlander, B., 2008. Powerful Writing Strategies for All Students. Brookes, Baltimore, MD.

Hayes, J., 1996. A new framework for understanding cognition and affect in writing. In: Levy, M., Ransdell, S. (Eds.), The Science of Writing: Theories, Methods, Individual Differences, and Applications. Erlbaum, Mahwah, NJ, pp. 1–27.

Hillocks, G., 2006. The Testing Trap: How State Writing Assessments Control Learning. Teachers College Press, New York.

Hillocks, G., 1986. Research on Written Composition: New Directions for Teaching. National Council of Teachers of English, Urbana, IL.

Horner, R., Carr, E., Halle, J., McGee, G., Odom, S., Wolery, M., 2005. The use of single-subject research to identify evidence-based practice in special education. Exceptional Children 71, 165–180.

Kiuhara, S., Graham, S., Hawken, L., 2009. Teaching writing to high school students: a national survey. Journal of Educational Psychology 101, 136–160.

MacArthur, C., Graham, S., 1987. Learning disabled students composing under three methods of text production: handwriting, word processing, and dictation. Journal of Special Education 21, 22–42.

Morphy, P., Graham, S., 2012. Word processing programs and weaker writers/readers: a meta-analysis of research findings. Reading and Writing: An Interdisciplinary Journal 25, 641–678.

National Center for Education Statistics, 2012. The Nation's Report Card: Writing 2011 (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington, DC.

National Commission on Writing in America's Schools and Colleges, 2004. Writing: A Ticket to Work … or a Ticket Out. College Board, New York.

National Commission on Writing in America's Schools and Colleges, 2005. Writing: A Powerful Message from State Government. College Board, New York.

National Commission on Writing in America's Schools and Colleges, 2006. Writing and School Reform. College Board, New York.

National Council of Teachers of English, 2004. NCTE Beliefs about the Teaching of Writing. Retrieved from: http://www.ncte.org/print.asp?id= 118876&node=367.

Olinghouse, N., Graham, S., Gillespie, A., February 2012. The Role of Discourse and Content Knowledge in the Narrative, Persuasive, and Informational Writing of Fourth Grade Students. Paper presented at the Pacific Coast Research Conference, Coronado, CA.

Olinghouse, N., Wilson, J., 2013. The relationship between vocabulary and writing quality in three genres. Reading & Writing 26, 45–66.

Rogers, L., Graham, S., 2008. A meta-analysis of single subject design writing intervention research. Journal of Educational Psychology 100, 879–906.

Russell, D., 1997. Rethinking genre in school and society: an activity theory analysis. Written Communication 14, 504–554.

Sandmel, K., Graham, S., 2011. The process writing approach: a meta-analysis. Journal of Educational Research 104, 396–407.

Santangelo, T., Graham, S., February 2013. Meta-analyses of Handwriting and Spelling Instruction. Paper presented at the Pacific Coast research Conference, Mission Beach, CA.

Scardamalia, M., Bereiter, C., 1986. Written composition. In: Wittrock, M. (Ed.), Handbook of Research on Teaching, third ed. MacMillan, New York, pp. 778–803.

Sheils, M., 8 December 1975. Why Johnny Can't Write. Newsweek, p. 58.

Yarrow, F., Topping, K.J., 2001. Collaborative writing: the effects of metacognitive prompting and structured peer interaction. British Journal of Educational Psychology 71, 261–282.

Zimmerman, B., Reisemberg, R., 1997. Becoming a self-regulated writer: a social cognitive perspective. Contemporary Educational Psychology 22, 73–101.

Writing Systems

Roger D Woodard, The University of Buffalo, Buffalo, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Writing systems, means of graphically recording language, are of three fundamental types: (1) *logographic systems* which consist of nonphonetic symbols, called *logograms*, which represent entire words; (2) *syllabic*; and (3) *segmental* writing systems which comprise *phonetic* symbols representing syllable-units and individual sounds, respectively. Segmental scripts can be further divided into two types: those representing only consonants (*alephbeths*) and those spelling both consonants and vowels (*alphabets*). Writing first appears among the Sumerians of southern Mesopotamia in the late fourth millennium BC and reveals itself in Egypt not long thereafter. Both the Sumerian and Egyptian systems are partially logographic and partially phonetic (the former syllabic, the latter segmental). The first attested script that consists solely of symbols for individual sounds (consonants) is called *Proto-Sinaitic*, used to write the Canaanite language of a people living in Egypt early in the second millennium BC. Proto-Sinaitic is one form of the more widely attested *Proto-Canaanite* script, ancestor of the Phoenician alephbeth, which is itself the source of the Greek alphabet, appearing in *c*.late ninth/early eighth century BC (the Greeks had earlier written with the syllabic scripts called *Linear B* and the *Cypriot Syllabary*). The Phoenician script is also the source of the Aramaic writing system, which in turn probably gave rise to the early scripts of India. Writing appears independently in China by *c*.1200 BC. Chinese writing is parent to Japanese (by the end of the ninth century AD); early Korean was written with Chinese characters, but the native Han'gul appeared in perhaps the fifteenth century AD. Writing also develops independently in Mesoamerica, as among the Maya by at least the third century AD.

The term 'writing system' specifically denotes a set of symbols that is used for the graphic (written) recording of language. Its use is commonly synonymous with that of the term 'script,' and the two are used interchangeably; however, the notion 'writing system' may additionally include the conventions (or spelling practices) which dictate how the symbols of such a system can be used. Forms of graphic expression which may have some semantic content – such as cave drawings, petroglyphs, icons, and even sophisticated picture messages – but which do not or could not record a linguistic utterance directly, are thus excluded from the realm of *writing*. This is not to suggest that all writing systems are capable of recording language with equal precision, but that by definition writing systems are fundamentally linguistic.

Types of Writing System

Writing systems are of three fundamental types. Logographic systems consist of symbols, called logograms, which represent entire words. In opposition to the two remaining types, a logographic system is said to be non-phonetic to the extent that its minimal elements record the pronunciation of an entire word, rather than representing the individual phonetic components (consonants and vowels) that make up the word. The reality of the phonetic opacity of such a system becomes obvious when one works with an ancient language recorded with logograms. It may be possible to identify the meaning of a logogram by its context and other clues, but if the word which it represents is not also attested by a more revealing, phonetic writing system, its pronunciation will remain uncertain or unknown.

In contrast to logographic scripts, syllabic, and segmental writing systems are said to be phonetic. A syllabic system consists of symbols, syllabograms, which have the value of a single syllable (although the syllables of a word spelled with a syllabic script may not actually line up with the syllabograms used in a one-to-one fashion). Syllabaries most commonly possess symbols of the value V (vowel) and CV (consonant + vowel); CCV, CVC, VC, among still other permutations, also occur, though less frequently.

Segmental writing systems are those which are composed of symbols representing individual sounds and are themselves of two types. Consonantal scripts (or alephbeths) have symbols for consonants only. Alphabets are characterized by both consonant and vowel symbols.

The relationship of the three types of writing systems to one another is in large measure of a historical, developmental nature. Consequently, the ensuing discussion will roughly follow a chronological scheme. As will be seen, most writing systems are not purely logographic or phonetic, but display some degree of hybridization.

The Birth of Writing

On the basis of presently available evidence, it appears that human beings first began to record language in southwest Asia and northeast Africa. Until quite recently, the very earliest examples of writing were held to be found among the archaeological remains of ancient Mesopotamia. However, a recent hypothesis would place the appearance of writing in Egypt at approximately the same time as its first attestations in Mesopotamia.

Mesopotamian Writing

Writing appears among the Sumerians of southern Mesopotamia in the late fourth millennium BC, and it is the Sumerians who are commonly credited with the invention of writing, although it appears to be attested not much later in Syria (at Tell Brak) and in Iran (in texts preserving the poorly

understood language of Proto-Elamite). The events which led to the advent of Sumerian writing are of course unknown, although there are various theories regarding its appearance. In the view of some scholars, the Sumerian invention of writing was a deliberate act, probably occasioned by the needs of palace administrators, and most likely in the city of Uruk where it is first attested. Others would see the process as being more of an evolutionary, in effect accidental, nature.

Regardless of the nature of the crucial step which led to the appearance of a system of writing on the Sumerian Plain, it seems undeniable that a precursor of the script is to be found in the clay tokens which had been in use for record keeping in Mesopotamia since at least 8000 BC. Some of these tokens actually resemble the commodity that they depict; others are of an arbitrary shape. By the latter half of the fourth millennium BC, some tokens had begun to be used as stamps, being impressed into moist clay tablets to produce administrative ledgers. Other tokens, those of more complex design, began to be etched freehand in the clay using a stylus. Schmandt-Besserat (1996) has argued that the appearance of Sumerian logographic writing was a natural evolutionary step following from the practice of inscribing token images (a process linked, she argues, to the development of symbols for abstract numerals).

As suggested by their conjectured association with tokens, the earliest Sumerian symbols are *pictographic*; in other words, the symbols have a picture-like appearance. The curved lines of the pictographic characters began, however, to give way to more angular, conventionalized shapes. Such conventionalization of symbols is a common process in the evolution of writing systems, but in this instance that process and the direction of development were dictated by the medium and instruments of writing. The chief medium used throughout the history of Sumerian writing was clay. Sumerian scribes first incised their pictographic characters in wet clay tablets using pointed styli. These, however, were replaced by a stylus with a triangular-shaped head that was simply pressed down into the clay, allowing for more rapid and efficient production of the symbols. Characters produced in this way have a characteristic wedge shape (quite pronounced by the middle of the third millennium BC), and are hence called cuneiform (from the Latin *cuneus*, 'wedge').

The value of early Sumerian symbols is logographic: each symbol represents an entire word. This logographic system is characterized by extensive homophony and polyphony. Homophony refers to the existence of multiple logograms having the same pronunciation, as they spell phonetically identical but semantically different words (homophones). Polyphony is, in a sense, the very opposite; the term denotes the practice of using a single logogram to spell two or more conceptually related but phonetically different words. For example, a single logogram is used to spell both DINGIR 'god' and AN 'heaven' (by Sumerological convention, Sumerian words are transcribed using capital letters).

A subset of the Sumerian logograms did double duty as semantic markers, called determinatives. These were written in conjunction with another logogram and were used, in effect, to provide the reader with a clue as to the meaning of that logogram, although their use is both conventionalized and optional. For example, the logogram GIŠ, 'wood,' is used as a determinative with logograms denoting items made of wood and tree names, such as ^{giš}ERIN, 'cedar' (determinatives are commonly transcribed as superscript). Similar devices also occur in the logographic Egyptian and Maya scripts, and in a somewhat different form in Chinese writing as well.

Sumerian is an agglutinating language with complex word structure involving the joining of multiple affixes in sequence. Logographic spelling thus obscures much word-internal morphology. Perhaps, it was chiefly this inherent shortcoming that led to a significant typological alteration in the Sumerian logographic writing system – the introduction of syllabic symbols (by about 3000 BC). The method employed in devising such symbols was that of a rebus. Just as in an English rebus, a picture of the sun, for example, could be used to 'spell' the word 'son,' so Sumerian logograms began to be used to represent syllables having the same sound. Sumerian scribes used the syllabic symbols not only for representing grammatical elements but also as phonetic complements, syllabic characters used to distinguish between polyphonous logograms. For example, the logogram KA was used to spell 'mouth' as well as related notions such as 'word,' 'speak,' and 'voice'; attaching the phonetic complement *-ma* (i.e., KA-*ma*) signals that the logogram is being used to spell INIM 'word.'

Late in the third millennium BC, texts begin to appear in the Mesopotamian archaeological world that are written in the cuneiform script of the Sumerians, but which are composed in a completely different language, the Semitic language of Akkadian. The Assyrians and Babylonians, speakers of Akkadian, greatly extended the use of syllabic symbols, but continued to utilize Sumerian logograms and determinatives as well. The Akkadian cuneiform script was transmitted to other peoples of the Near East, providing writing systems for the Hittites and Luvians, the Hurrians and Urartians, the Elamites, and, in a significantly streamlined form, the Persians. The latest known Akkadian cuneiform texts were produced in the first century AD.

Egyptian Writing

While the pictographic symbols of the Egyptian hieroglyphic script are clearly graphically distinct from those of Sumerian, and the internal operation of the two scripts shows appreciable differences, similarities are undeniable. Both possess a logographic component; both have a phonetic component, which is commonly used for phonetic complementation; both expand the use of symbols through the rebus principle; both use a subset of logograms as determinatives. As it has appeared that the Egyptian script is first attested a bit later than the Sumerian (about 3100–3000 BC), as there is evidence for a precursor to the Sumerian logograms, and as Egypt was otherwise being influenced by Mesopotamia at the end of the fourth millennium BC, it has been commonly conjectured that at least the idea of writing, if not a writing system itself, came to the Egyptians from Mesopotamia. One possible difficulty with this scenario is that the Egyptian writing system appears well developed at its earliest attestation, perhaps suggesting some earlier, undocumented period of use.

From the Abydos tomb complex in Upper Egypt, small ('postage-stamp-size') tags of bone and ivory have been recovered on which are carved symbols that in some cases bear

similarity to later-known hieroglyphs. Günter Dreyer (1998) has dated these tags to around 3400–3200 BC and interpreted the symbols as writing. The dating of the tags and, perhaps more crucially, the interpretation of their markings continues to be explored.

The most striking difference between the Egyptian and Sumerian systems is the nature of the phonetic component. In Egyptian writing the phonetic symbols do not have syllabic value, but represent consonants – either single consonants or sequences (although not necessarily phonetically contiguous strings) of two or three consonants. Twenty-six monoconsonantal symbols occur (including two variant forms), representing the individual consonants of the language; there are almost 100 biconsonantal symbols, and somewhat fewer triconsonantal. Biconsonantal and triconsonantal phonetic symbols are often themselves accompanied by one or more monoconsonantal phonetic complements. For example, the triconsonantal symbol *nfr*, spelling the Egyptian word meaning 'good,' is commonly used in conjunction with one or two monoconsonantal complements, as in the spelling sequence *nfr f r*.

Consonantal Writing Systems

During the early second millennium BC, a Semitic people living and laboring in Egypt developed a script for their own language by exploiting a single component of the Egyptian writing system. With the inspiration provided by the monoconsonantal Egyptian symbols, these people designed a fully phonetic and greatly simplified writing system which consists entirely of symbols for individual consonants. The first evidence of the script, labeled Proto-Sinaitic, came to light in the early twentieth century with Sir Flinders Petrie's excavations of Egyptian turquoise mines at Serabit el-Khadim in the Sinai; hence the name. Quite recently, additional examples of Proto-Sinaitic writing have been discovered near the ancient city of Thebes, along the Farshût Road at Wadi el-Hol, dating *c.* 1850 BC (Darnell et al., 2005). The evidence from Wadi el-Hol suggests that development of the script was significantly motivated by written-communication needs among Semitic-speaking mercenaries serving in the Egyptian army.

The method that the adapters employed in turning the Egyptian script to their own use was dubbed the acrophonic principle by the British Egyptologist Sir Alan Gardiner. The values assigned to the phonetic characters of this earliest Semitic script – characters of pictographic form – were determined by the initial consonant of the Semitic word for the symbol depicted, rather than by any phonetic value associated with a corresponding Egyptian symbol. For example, the Egyptian hieroglyphic symbol resembling waves of water and used in the Egyptian system as a monoconsonantal symbol with the value /n/was assigned the value /m/ in the Proto-Sinaitic script, as the word for 'water' in the West Semitic language of that script began with /m/. The Proto-Sinaitic character for /n/, on the other hand, was based on the Egyptian symbol representing /dž/, a hieroglyph depicting a serpent, and received its Proto-Sinaitic value from the initial consonant in the Semitic word for 'snake.'

Proto-Sinaitic is one particular form (the earliest by the interpretation of an origin within Egypt) of a more broadly, if meagerly, attested consonantal script called Proto-Canaanite, found at several sites in Syria–Palestine. This Proto-Canaanite script is the immediate ancestor of the Phoenician script (appearing in the eleventh century BC), which is in turn the source of the ancient Hebrew (*c.*ninth century) and Aramaic (*c.*eighth century) writing systems (the Hebrew language began to be written with the Aramaic script during the sixth century BC, although the Samaritan sect continued to use the old Hebrew letters). The Aramaic script gave rise to the writing system of Syriac (*c.*first century AD, and late enough to be influenced by the Greek alphabet) and, through the particular Aramaic script used by the Arab Nabateans, to the Arabic writing system (*c.*fourth century AD).

Proto-Canaanite writing also served as the inspiration and model for the script of the ancient Levantine city of Ugarit. The Ugaritic writing system, however, departs markedly from the other Canaanite scripts in that its symbols are of cuneiform shape. It thus represents something of a hybrid writing system, combining the consonantal nature of Proto-Canaanite with the graphic form of the Mesopotamian scripts, although some scholars have argued that the shapes of the Ugaritic consonantal characters were themselves modeled on the Proto-Canaanite symbols but executed with the cuneiform stylus. The Ugaritic writing system also differs from other Canaanite scripts to the extent that it does not consist wholly of consonantal characters, but includes in its graphemic inventory three CV syllabic symbols, representing a glottal stop followed by the vowels /a/, /i/, and /u/.

Yet a different Semitic writing tradition would descend from the Proto-Canaanite script. In the south of the Arabian peninsula a distinct consonantal script had appeared by at least the sixth century BC, and certainly had begun to develop several hundred years earlier. An interesting feature of this Old South Arabian script is that the order of its letters departs from that of Phoenician and the consonantal script of Ugarit. However, an abecedary found at the Palestinian site of Beth Shemesh, consisting of consonantal cuneiform characters like those in use at Ugarit, shows the same letter order as found in the Old South Arabian script. The consonantal script of South Arabia spread northward to be adapted for use by early North Arabian languages (distinct from Arabic) and westward to the horn of Africa, where it was employed for writing the Semitic language of Ethiopic. Interestingly, the consonantal script of Ethiopic had undergone a significant metamorphosis by the fifth century AD, becoming a syllabary of almost 200 CV symbols.

Mediterranean Syllabaries

Crete

In roughly the same period that a Semitic people resident in Egypt were fashioning a consonantal script out of Egyptian hieroglyphs, three writing systems were emerging among the Minoans of Crete. Probably the first to appear was the pictographic script called Cretan Hieroglyphic. Dating back to about the eighteenth century BC is a second Minoan script, attested on only a single document, the Phaistos Disk. A third script, also appearing in the eighteenth century and likely to be descended from Cretan Hieroglyphic, is the one that was labeled Linear A by Sir Arthur Evans, the British archaeologist

who excavated the Minoan city of Knossos early in the twentieth century. The name reflects the linear nature of this script's symbols in contrast to the pictographic quality of symbols in the other two Minoan scripts. None of these ancient Minoan scripts has yet been deciphered by modern investigators, although recent work on Linear A may suggest that it records an Indo–European language of the Anatolian subfamily (an idea first proposed in the mid-twentieth century by the British classicist and linguist L.R. Palmer). Cretan Hieroglyphic is possibly a syllabic script, as is probably that of the Phaistos Disk; Linear A is certainly a syllabary.

In his excavations at Knossos, Evans unearthed documents written in yet a different script, but one that bears an unmistakable similarity to Linear A: this writing system was accordingly given the name Linear B. In the early 1950s the Linear B script was successfully deciphered by Michael Ventris, a British architect, working in collaboration with the Cambridge classicist John Chadwick. The language of the Linear B documents turned out to be Greek – that dialect used by the Mycenaean Greek culture of the second-half of the second millennium BC. The identification of the language of the Linear B script was made particularly difficult by the seeming unsuitability of the script for spelling that language. The great majority of the symbols of this syllabic writing system are of the type CV. As Greek has many word-internal CC sequences and word-final consonants, spelling practices were devised which in some cases required the use of CV symbols for spelling consonants only (with no following vowel), and in some cases resulted in a consonant not being spelled at all. The basic strategy for spelling consonant sequences is one that co-opts a phonological property of the Greek language, namely the relative sonority of consonants. It is the general case that a consonant is written when it precedes a consonant of greater sonority (such as a stop occurring before a fricative), utilizing the CV symbol whose vocalic component is identical with the vowel which follows the consonant sequence, but is not written if the ensuing consonant is of less sonority (such as a fricative before a stop). In either case, the second consonant is written (with the appropriate CV symbol). Word-final consonants are not written and this practice appears to be a further expression of the sonority-based strategy.

Cyprus

There is evidence for writing on the eastern Mediterranean island of Cyprus as early as the middle of the second millennium BC. The writing system in use in that period is preserved in distinct forms called Archaic Cypro-Minoan and Cypro-Minoan I, II, and III (although the last type is actually attested in the documentary remains of the city of Ugarit, on the neighboring Levantine coast, and not on Cyprus itself). In spite of the fact that none of the three has yet been deciphered, the formal similarities between these scripts and those of Minoan Crete suggest that the Bronze Age Cypriot writing system had its origin in the Cretan scripts.

Cyprus was a key locale lying along the trade routes which stretched from Greece to the east, and certainly for this reason there had been a Greek presence in Cyprus during the Mycenaean era; however, the Greek population of Cyprus increased dramatically as refugees fled the conflagration which brought an end to the Mycenaean civilization in the Greek homeland. The earliest remains of Greek writing in Cyprus come from the middle of the eleventh century BC, and the system of writing used was that of the Cypriot Syllabary. This syllabary, while probably based graphically upon Cypro-Minoan I is clearly closely affiliated with the Mycenaean script. Like Linear B, the Cypriot Syllabary is composed predominantly of symbols having the value CV. The same problems of representing word-internal consonant sequences and word-final consonants which Linear B scribes faced also plague the Cypriot Greek writing system; and the Cypriot solutions to these problems are virtually identical to those utilized by the Mycenaeans, being based upon the relative sonority of Greek consonants. The notable difference between the two systems in this regard is that the Cypriot writing system makes much more limited use of the CV-deletion strategy than does Linear B. In most instances in which the first of two consonants is of greater sonority, the Cypriot strategy is to write that consonant (and not to refrain from writing it, as in Linear B) using a CV symbol whose vocalic component is identical to the vowel that precedes the consonant sequence.

The Alphabet

The Greek Origin

It was the Greeks who created the first segmental writing system designed to represent not only consonants but vowels as well. The raw material for the Greek alphabet was provided by the consonantal script of the Phoenicians. As there was both a significant Phoenician presence on the island of Cyprus and a vibrant tradition of Greek literacy, Cyprus provides a *de facto* likely candidate for the cultural center that played a crucial role in the creation of the Greek alphabet, an event that likely occurred in the early eighth or late ninth century BC. There are, moreover, particular features of the early Greek alphabet which are not motivated within that segmental script, but which are motivated within the Cypriot Syllabary. For example, the biconsonantal symbol *xi*, representing the sequence of /k/ +/s/, is unnecessary within a segmental script that has both a *k*-symbol (*kappa*) and an *s*-symbol (*sigma*). Within the Cypriot Syllabary, however, *ks*V symbols are idiosyncratically required by the strategy of consonant cluster spelling which is based upon the relative sonority of consonants, co-incident to Greek phonological structure (Woodard, 1997b). Cypriot Greeks appear to have transferred this and other systemic features from their operationally complex syllabic script to the newly devised alphabet. For a likely scenario in which the adoption of such a simplified segmental means for writing Greek would be highly valued, one might look to a setting in which literate Cypriot Greeks and non-literate Ionian Greeks served jointly in a mercenary capacity in the context of Assyrian military operations in Syria–Palestine (Woodard, 2010).

For vowel symbols, the Greeks adapted Phoenician consonantal characters that were not required for writing Greek. Thus, Phoenician *'alep*, symbol for a glottal stop (which sound Greek did not possess phonemically) was assigned the Greek value /a/ (the Greek letter *alpha*). Phoenician *he*, spelling the glottal fricative /h/, was used by the Greek adapters to record mid front (*e*) vowels (*epsilon*; for writing Greek /h/, Phoenician

ḥet was employed, the symbol for a voiceless pharyngeal fricative /ħ/). To write mid back (*o*) vowels (*omicron*), the Greeks tapped Phoenician *ʿayin*, which represented a voiced pharyngeal fricative /ʕ/. Phoenician *yod*, spelling the palatal glide /y/, and *waw*, representing /w/, were assigned the Greek values /i/ (*iota*) and /u/ (*upsilon*), respectively.

The Etruscan Alphabet

Following rapidly upon its creation, the alphabet was widely disseminated throughout the Greek world, developing many local varieties. The Phrygians of central Anatolia acquired and adapted the Greek alphabet not long after its creation. In the south of Italy, Greek colonists passed their western Greek alphabet on to a non-Indo-European people of Italy, the Etruscans, by the late eighth century BC. As the Etruscan language lacked voiced stops, the Greek letters representing such sounds, *beta* (/b/), *delta* (/d/), and *gamma* (/g/), were not needed for that purpose. The first two of these letters would disappear from the Etruscan alphabet, although *gamma* was retained and used for spelling the voiceless stop /k/ when it occurred before a front vowel. The Etruscans used Greek *kappa* for /k/ before /a/ and Greek *qoppa* for /k/ before the high back vowel /u/. In turn, the Etruscans transmitted the alphabet to other peoples of Italy.

The Roman Alphabet

Among those places to which the Etruscan writing system came was the city of Rome – a singly significant step in the history of the alphabet. For a time, the Romans continued the Etruscan practice of using three different symbols to represent the velar stop /k/ in its various phonetic contexts. However, *C* (from Greek *gamma*) was eventually used regularly for /k/ as well as for /g/ (*K*, Greek *kappa*, continued to be used for a very few words and *Q*, Greek *qoppa*, was employed before the letter *V* for spelling /k^{w}/). By the third century BC, a variant form of *C* had been introduced, namely *G*, for writing the voiced velar stop /g/. However, as *G* was assigned the alphabetic position held by Greek *zeta* (spelling the early Greek sequence /zd/, later /z/), which was not required for spelling Latin, it has been argued that *zeta*, not *gamma*, provided the raw material for the creation of the new letter *G*. Later, in the first century BC, with the influx of much Greek vocabulary, both *Y* and *Z* (spelling Greek /ü/ and /z/ respectively) were appended to the Roman alphabet.

The Alphabet Abroad

The alphabet would make its presence known, directly and indirectly, throughout the European continent and beyond. The fragmentary remains of Celtic writing from continental Europe are attested in alphabetic scripts – Etruscan, Roman, or Greek (aside from the Hispano-Celtic inscriptions written in a curious script which is part syllabic, part alphabetic). Moreover, the 'native' writing systems of both the ancient Celts of Britain and Ireland, and of the Germanic peoples of Europe are alphabetic traditions, although formally distinct from both the Greek and Roman scripts, and strikingly so in the case of Celtic Ogham, with its characters of oblique and right-angle lines. Nevertheless, it is undoubtedly the case that both Ogham and the Germanic runes found their inspiration and model in the Mediterranean alphabets. When in the fourth century AD, the Gothic Bishop Wulfila translated portions of the Bible into his Germanic language, he used an alphabetic script of his own design, based chiefly on the Greek alphabet.

The spreading of the Christian faith and the effort to ground believers in the biblical texts were – far beyond Wulfila – fundamental factors in the evolution of alphabetic writing systems. The Roman Catholic Church played a key role in the introduction and eventual establishment of the Roman alphabet across much of the European continent. Carried by the colonizing European nations, this writing system would be transported around the world, being utilized for writing European languages on all six of the populated continents, and adapted for spelling indigenous languages of Asia and Africa, as well as providing raw material for the symbols of the Cherokee syllabary of North America.

In the ninth century AD, Greek missionaries led by Saint Cyril and Saint Methodius developed writing systems, based on their own Greek alphabet, for translating the Bible into the language of south Slavic converts. One of these scripts, the Cyrillic alphabet, thrived and is the predecessor of the alphabetic scripts of various modern Slavic languages, most notably Russian.

The alphabets of both Armenian and Georgian are traditionally attributed to the fifth-century cleric, Saint Mesrop. These writing systems were based at least partially on the Greek alphabet, but their creator may have also drawn upon a Semitic consonantal script.

Writing in South and East Asia

The Indian Syllabaries

The earliest attested writing system of India is the script of the Indus Valley civilization, dating back to *c.*2500 BC. Many of the symbols are of a pictographic nature, and the number of symbols identified suggests that this writing system is logographic, possibly partially syllabic. It is, however, presently poorly understood (and an object of academic controversy) and does not appear to have been the source of the Indian syllabaries that appear in the third century BC. These scripts, Kharoshthi and Brahmi, probably developed from the Aramaic consonantal writing system. Of the two, only Brahmi appears to have survived late antiquity; its perhaps best-known form is Devanagari, the writing system of Sanskrit, as well as that of Hindi and other modern Indo-Aryan languages.

Devanagari differs notably from the Mediterranean and Mesopotamian syllabic writing systems in that rather than having a distinct symbol for each possible consonant + vowel combination, Devanagari consists of a basic set of CV symbols having the vowel value *a* (i.e., C*a* symbols). These CV symbols can be assigned different vowel values (C*i*, C*u*, etc.) by the addition of diacritic strokes (this is essentially the same strategy utilized by the Ethiopic syllabary, which traces its origins to a Semitic consonantal writing system, as apparently do the Indic scripts). Devanagari also has the capacity to transform basic CV symbols into ones representing more complex consonantal components (CCV, CCCV, etc.) through the use of

conventionalized ligatures. Such ligatures consist essentially of a C*a* symbol (representing the final member of the consonant cluster) to which a form, often highly abbreviated, of one or more other C*a* symbols is attached. Word-final consonants are spelled utilizing the appropriate C*a* symbol beneath which a diacritic mark, called *virama*, is placed, revealing to the reader that the symbol is to be read with no vowel component.

The Indian syllabaries were spread through much of Central and Southeast Asia, giving rise to writing systems for numerous languages. Not only does the Tibetan script trace its roots to India, but also the writing systems of Javanese, Thai, Burmese, Lao, and Khmer.

Chinese Writing

The earliest evidence for writing in China is provided by the Shang dynasty oracle-bone inscriptions of *c*.1200 BC. There is no clear evidence that writing developed in China under the influence of any then existing writing system; as in Mesopotamia, and perhaps in Egypt, writing seems to have appeared de novo. Like the earliest Mesopotamian writing system, the early Chinese script is pictographic in form and logographic in operation (i.e., each picture-like symbol represents an entire word). Moreover, as with both Mesopotamian and Egyptian scripts, the Chinese writing system is one that both multiplied the value of its characters utilizing the rebus principle and employed orthographic markers in order to minimize graphic ambiguity. In China, however, these markers were commonly graphically combined with the associated characters to produce compound symbols of two components – one signaling pronunciation, the other meaning. The fundamental logographic nature of Chinese writing has been preserved to the present day, and the inventory of logographic characters has steadily increased over time. While approximately 60 000 distinct characters are currently identified, about 2500 occur in common use.

Japanese Writing

The modern Japanese writing system is quite complex, probably the most so of modern scripts. The earliest attested examples of Japanese writing (*c*.eighth century AD) are executed in Chinese script. However, by the end of the ninth century, two Japanese syllabaries had appeared, evolving out of the practice of using certain Chinese characters as phonetic symbols. Both of these syllabaries remain in use alongside *kanji*, essentially logographic symbols originating in Chinese characters; in the contemporary script, a single such symbol can have multiple values. The syllabary called *katakana* is utilized for scientific terms, for foreign loans, for indicating emphasis, among still other uses. Usages of modern *hiragana* are varied as well: it is, for example, the syllabary used to spell grammatical particles and to represent affixes associated with *kanji*.

Korean Writing

As in the case of Japanese, the Korean language was first written with Chinese characters. It was in the fifteenth century AD, according to Korean tradition, that the Korean alphabetic writing system of Han'gul was invented. Credit for this script – the world's only alphabetic writing system which may not trace its roots to the Greek alphabet – is given to Sejong, a Korean king. Han'gul is a highly efficient and phonetically transparent script. Its various consonantal symbols are purported to have been designed to represent schematically the articulatory position of the tongue in producing the consonant sounds of Korean. While Han'gul is alphabetic, consonant and vowel characters are written together in such a way as to form orthographic blocks corresponding to phonetic syllables.

Writing in Mesoamerica

In addition to China, on the one hand, and Egypt and the Near East, on the other, there is at least a third region of the world, where human beings appear to have designed a means of writing without prior knowledge of any such thing, namely Mesoamerica. While the study of ancient Mesoamerican

Table 1 The Greek alphabet

Symbol	Value	Symbol	Value
Α, α	a	Ξ, ξ	ks
Β, β	b	Ο, ο	o
Γ, γ	g	Π, π	p
Δ, δ	d	Μ	s
Ε, ε	e	Ϙ	k
Ϝ	w	Ρ, ρ	r
Ζ, ζ	zd	Σ, σ	s
Η, η	ē	Τ, τ	t
Θ, θ	t^h	Υ, υ	u
Ι, ι	i	Φ, ϕ	p^h
Κ, κ	k	Χ, χ	k^h
Λ, λ	l	Ψ, ψ	ps
Μ, μ	m	Ω, ω	ō
Ν, ν	n		

Table 2 The Ugaritic cuneiform consonantal script

Symbol	Value		Symbol		Value
	å	[ʔa]		d̲	[ð]
	b			n	
	g			ẓ	[ð']
	ḫ	[x]		s	
	d			ʿ	[ʕ]
	h			p	
	w			ṣ	[s']
	z			q	[k']
	ḥ	[ħ]		r	
	ṭ	[t']		t̲	[θ]
	y			ǵ	[ɣ]
	k			t	
	š			ỉ	[ʔi]
	l			ủ	[ʔu]
	m			s̀	[s] (?)

Table 3 The basic CV symbols of Devanagri

Symbol	Value	Symbol	Value	Symbol	Value
क	ka	च	ca	ट	ṭa
ख	kha	छ	cha	ठ	ṭha
ग	ga	ज	ja	ड	ḍa
घ	gha	झ	jha	ढ	ḍha
ङ	ṅa	ञ	ña	ण	ṇa
त	ta	प	pa		
थ	tha	फ	pha		
द	da	ब	ba		
ध	dha	भ	bha		
न	na	म	ma		
य	ya	श	śa		
र	ra	ष	ṣa		
ल	la	स	sa		
व	va	ह	ha		

writing is still at an early stage, significant strides have been made since the 1970s. Of the several Mesoamerican scripts currently known, the best attested and best understood is that of the Maya, who were writing by at least the third century AD. In its fundamental characteristics, the Maya writing system is much like those of the Sumerians and Egyptians. The script's pictographic symbols, or glyphs, are logographic and syllabic in value. The latter type is used both in conjunction with logograms as well as for the phonetic spelling of words. In addition, the Maya system makes use of determinatives, like the Sumerian and Egyptian systems, to provide semantic guidance to the reader (Tables 1–3).

See also: Anthropological Writing; Reading Nonalphabetic Scripts; Writing Instruction; Writing Systems, Psychology of; Writing, Evolution of.

Bibliography

Coulmas, F., 1989. The Writing Systems of the World. Blackwell, Oxford, UK.

Coulmas, F., Ehlich, K., 1996. A Bibliography on Writing and Written Language. De Gruyter, Berlin.

Daniels, P.T., Bright, W. (Eds.), 1996. The World's Writing Systems. Oxford University Press, Oxford, UK.

Darnell, J.C., Dobbs-Allsopp, F.W., Lundberg, M.J., McCarter, P.K., Zuckerman, B., Manassa, C., 2005. Two Early Alphabetic Inscriptions from the Wadi El-Hôl: New Evidence for the Origin of the Alphabet from the Western Desert of Egypt. American Schools of Oriental Research, Boston.

Diringer, D., 1968. The Alphabet: A Key to the History of Mankind, third ed. Funk and Wagnalls, New York.

Driver, G.R., 1976. Semitic Writing from Pictograph to Alphabet, rev. ed. In: Hopkins, S.A. (Ed.), Oxford University Press, Oxford, UK.

Dreyer, Günter, 1998. Umm el-Qaab I: Das prädynastische Königsgrab U-j und seine frühen Schriftzeugnisse. Verlag Philipp von Zabern, Mainz am Rhein.

Gelb, I.J., 1963. A Study of Writing, second ed. University of Chicago Press, Chicago.

Günther, H., Ludwig, O. (Eds.), 1994. Schrift und Schriftlichkeit. De Gruyter, Berlin.

Hooker, J.T., 1990. Reading the Past: Ancient Writing from Cuneiform to Alphabet. The British Museum, London.

Robinson, A., 1995. The Story of Writing. Thames and Hudson, London.

Sampson, G., 1985. Writing Systems: A Linguistic Introduction. Stanford University Press, Stanford, CA.

Schmandt-Besserat, D., 1996. How Writing Came About. University of Texas Press, Austin, TX.

Senner, W.M. (Ed.), 1989. The Origins of Writing. University of Nebraska Press, Lincoln, NE.

Woodard, R.D., 1997a. Writing systems. In: Comrie, B., Matthews, S., Polinsky, M. (Eds.), The Atlas of Languages. Bloomsbury, London.

Woodard, R.D., 1997b. Greek Writing from Knossos to Homer. Oxford University Press, Oxford, UK.

Woodard, R.D., 2002. Cambridge Encyclopedia of the World's Ancient Languages. Cambridge University Press, Cambridge, UK.

Woodard, R.D., 2010. Phoinikēia Grammata: an alphabet for the Greek language. In: Bakker, E. (Ed.), Companion to the Ancient Greek Language. Wiley-Blackwell, Oxford, UK.

Woodard, R.D., 2014. The Textualization of the Greek Alphabet. Cambridge University Press, Cambridge, UK.

Writing Systems, Psychology of

David R Olson, Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Writing systems are cultural inventions for the storage and communication of information. Although only some 5000 years old and invented ab initio four times – Sumeria, Egypt, China, and Mesoamerica – writing systems have become universally adopted and now serve a fundamental set of social and personal uses and are considered by some as mankind's most remarkable technical achievement. In spite of the ubiquity of writing, its psychological study is only now taking shape. This introduction sketches some of these lines of research.

The writing systems of the world differ importantly in how they relate to spoken language. Tokening and pictographic (semasiographic) scripts relate to meanings or intentions directly with little regard to the spoken form in which that meaning could have been expressed. So-called full writing systems represent properties of the spoken language but in completely different ways. Morphophonemic (logographic) scripts represent the words or morphemes of the language; syllabic scripts represent the syllables of the language whether or not they also represent word boundaries. Alphabetic scripts represent, with varying degrees of precision, the phonemes of the language, but also by means of spaces may represent the words of the language. Not only do these differences have an effect on learning to read, they also have an important effect on the ways in which one thinks about language and consequently about the world and the mind. Writing systems provide models for thinking about speech in addition to providing a distinctive mode of communication.

Although the psychological study of *reading* and the psychology of *language* are well-established domains of research and theory, the *psychology of writing systems* and their uses, *literacy*, are relatively recent additions to the cognitive sciences. The central goal of the psychology of writing systems is to bring these two, hitherto independent, fields of study into conjunction, allowing both to be seen in a new light.

Writing systems have been neglected in psychology for two reasons. First, twentieth-century linguistics has focused almost exclusively on speech, consigning writing merely to a way of recording. The study of writing was left to specialist domains such as epigraphy and classical studies. Nonetheless, important studies of writing were produced, that by Gelb (1963) providing the foundation for more recent work (Houston, 2004; Daniels, 2009). Second, only the alphabetic writing system was the object of psychological studies of reading, leaving the impression that other writing systems were primitive and scarcely worth careful study.

In the period since the 1960s, the two fields, the psychology of reading and the psychology of language, have begun to converge to produce several lines of new research and theory represented by several handbooks and two new journals *Written Language and Literacy* and *Writing Systems Research*. Line of research make up the major sections of this article.

Relation between Speech and Writing

Making marks which can serve mnemonic and communicative purposes is as old as human culture itself. Writing systems are a subset of such marks which, typically, represent aspects of linguistic form. Sampson (1987) provided an abstract classification of types of writing system. Some systems, *semasiographic* systems, represent meanings without any direct link to spoken form. Historical examples include the so-called 'picture writing' of American Indians, more modern ones include systems of highway signs and mathematical symbols. Some writers (Coulmas, 1989; Harris, 1995) consider such nonspeech sign systems as full writing systems, claiming that what makes a writing system is its adequacy in conveying information visually rather than its purported links to speech.

Logographic scripts are linguistically based but the spoken elements represented are meaningful units, primarily words and morphemes. A clear example of such a script is Chinese in which morphemes which sound identical are represented by entirely dissimilar graphic characters, although some theorists (de Francis, 1989) insist that any full writing system is phonologically based, at least to some extent. Robertson (2004: p. 25) defines logographic signs as any sign, whether icon, index or symbol, that identifies a relation between a written sign and 'the object of the spoken sign.' A logograph, then, even if it resembles an object or event, is not a picture of an event but a representation of language.

Syllabaries are scripts which represent spoken patterns by signs for syllables. Interestingly, syllabaries are now thought to be modern inventions rather than historically step toward the invention of the alphabet (Trigger, 2004: pp. 64–65). Yet syllables are easily discriminated units of speech that play a role in children's learning to read an alphabet (Adams et al., 1998). In the Ancient Middle East, Aramaic, Hebrew, and Arabic writing systems sometimes described as syllabaries are more correctly described as consonantaries or unvocalized alphabets because of the absence of vowels. Thus the syllables *ka*, *ke*, *ki*, *ko*, for example, are written with the same graphic sign representing the consonant /k/.

Finally, *alphabets* employ a separate mark to represent the consonant and vowel segments of a syllable, roughly, the phonemes of the language. These categories of writing systems are generic types; no living writing system is based exclusively

on one type of sign but rather combinations of word signs, morpheme signs, syllabic signs, and phonemic signs. Thus in English script, signs represent both phonemes by means of letters and words by means of spaces between letter groups and sentences by means of punctuation. Japanese writing uses Chinese logographs (Kanji) to represent the stems of content words and a completely different syllabic script (Kana) to represent grammatical morphemes and inflections (Taylor and Taylor, 1995).

In one sense, the history of scripts reflects the discovery of increasingly subtle aspects of speech, going from meanings to analytical constituents of speech, namely phonemes. Although Gelb (1963) saw an evolutionary force as work that moved logograms to phonograms and eventually to the alphabet, most researchers (Gaur, 1987; Sampson, 1987; Olson, 1994; Harris, 1995; Trigger, 2004; Daniels, 2009) now argue that while writing systems have evolved to more adequately represent speech, they have done so by using iconic signs, indexical signs, and phonological signs. The Greek invention of the alphabet, once thought of as the work of genius is now seen as the result of borrowing a script invented for one language to represent a completely different language. The consequence of such borrowing was the discovery of previously ignored properties of speech, word signs, syllable signs, and signs for phonemes a pattern that bears some similarity to the processes children undergo in learning to read and write.

Writing Systems and Reading

As different types of writing systems bear distinctively different relations to speech, learning to read and write them call upon different cognitive activities. Semasiographic systems, signs that represent meanings rather than linguistic forms make no sharp distinction between drawing and writing in that they allow variable readings. Nineteenth-century Canadian Blackfoot Indians, for example, depicted 'the year that horses got drowned' by means of a circle (representing a lake) and four stick drawings of horses superimposed on the circle. Such a chronicle records an important cultural event but not the linguistic form of its description. Ancient Mesoamerica, namely the Mixtec, Aztec, and Mayan systems invented before the conquest long thought to be picture writing are now recognized as writing systems based primarily on logographs representing units of meaning that are then borrowed to represent units of sound, the so-called rebus principle (Houston, 2004).

All other so-called full writing systems map on to or represent spoken forms and learning to read each of them requires a somewhat different orientation to the script. Scribner and Cole (1981) found that their Vai subjects in Liberia in learning to read a syllabic script became adept at analyzing their speech into syllabic units but failed to analyze their speech into words. Others have found that the ability to manipulate phonemes was tied to learning to read an alphabet. Read et al. (1986) found that Chinese readers of traditional character scripts could not detect phonemic segments whereas those who could read Pinyin, an alphabetic script representing the same language, could do so.

Ample empirical evidence exists to show that it is acquaintance with an alphabetic script that is critical to the acquisition of awareness of the phonological properties of the language (Morais et al., 1979; Vernon and Ferreiro, 1999). Concepts of words as lexical entities, similarly, have been shown to be tied to concepts of words as written entities (Olson, 1994). Concepts of sentences, needed for the mastery of punctuation, are acquired at even more advanced levels of literacy. Conceptual development in children is, in part, the consequence of the acquisition of these systems for representing language (Vygotsky, 1986).

Much confusion about the cognitive and social implications of writing and reading arises from a failure to recognize that modes of reading and interpretation do not inevitably arise from simply learning to identify words and letters. The history of writing is largely a history of inventing devices, such as punctuation and text structures, as well as rules for interpretation that have taken sometimes millennia to develop. A dictionary itemizing all the words of a language is a recent literate invention. The very idea of fixity of a text is not inevitably attributed to writing and printing but to a specialized set of conventions for interpreting texts and institutional structures for preserving them (Johns, 1998). However, once procedures had been developed for the systematic interpretation of Scriptural texts, it was found that such procedures were also suitable for the interpretation of the 'book of nature.' Scientific epistemology, it has been argued, is a form of applied hermeneutics (Olson, 1994; Killeen and Forshaw, 2007).

Theories of Reading Alphabetic Scripts

Assumptions about the relation between script and language have given rise to considerable, well-publicized, differences of opinion and differences in reading theory. One group emphasized the recognition of meaning of written expressions in ways analogous to the recognition of meaning in oral forms. That is, written language is seen as an alternative to spoken language each with their peculiar and distinctive perceptual and cognitive processes but with a common goal of meaning construction, hence the label Whole Language.

The other group, credit for which is often given to Alvin Liberman and Isabel Liberman of the Haskins Group at Yale University, is that reading is essentially a metalinguistic activity in which readers must find mappings from the written form to aspects of their implicit knowledge of speech (Olson, 1994; Adams et al., 1998). Evidence on the importance of word and phoneme segmentation, two forms of metalinguistic knowledge, to learning to read and write tends to favor the metalinguistic theory of reading (Stanovich, 1986; Lundberg et al., 1988), that is, the theory that reading is not a matter of mapping signs to meanings but rather a matter of mapping visual signs to aspects of oral speech and hence on to meaning. It is this metalinguistic theory of reading that helps to explain not only why learning to read brings aspects of speech into consciousness but also why reading and writing may contribute to a new and distinctive mode of thought in certain social and institutional contexts.

Writing Systems and Thinking

Nineteenth-century theorists took the view that writing and literacy had the general effect of making thought more logical

and formal. The recognition of the fact that there are many types of writing systems and that even those different types may be read in very different ways (Saenger, 1997) has made such general claims problematic, as has the dispute as to the meaning of literacy, some identifying it with the basic skills of reading and writing (Scribner and Cole, 1981) and others with the evolved products of particular literate cultural traditions (Goody, 1968; Olson, 1994).

Although there is wide agreement on the 'psychic unity of mankind' ("They are just like you and me," as Jack Goody once put it), and there is wide agreement that all languages are *in principle* capable of expressing all nuances of meaning, nonetheless patterns of thought vary widely from culture to culture and context to context, so much so that it is not uncommon to despair of finding interesting cognitive differences attributable to writing or to a written tradition. Mathematics, the epitome of a written tradition, is sometimes seen as basically equivalent to practical computations, so-called street math.

On the other hand, cultural historians have asserted the cultural and psychological implications of writing and literacy, arguing, for example, that the medium of representation is the most important aspect of the message (McLuhan, 1962), that writing separated the knower from the known (Ong, 1992), that scriptal changes caused the origins of modern silent reading (Saenger, 1997), and that writing allowed the formulation of modern conceptions of logic and mathematics (Goody, 1968).

The metarepresentational theory of reading gives new credence to these more radical views. Writing and reading a linguistically based script, on this view, are an exercise not only in meaning, but also in consciousness and discourse about language, that is, about units of sound, units of meaning, words and structural units, sentences. Literacy on this view is seen as instrumental to discourse about word meaning, sentence meaning, and to other aspects of form which may be *mentioned* rather than *used* in discourse. This metarepresentational shift invites talk not only about particular words and sentences independently of their use, but also about other metarepresentational properties, such as *truth*, *logical form*, and *aesthetic goodness*. Furthermore, it allows mathematics to be clearly distinguished from 'street math,' the former being concerned with number rather than quantity. In the former, a number is expressed in the singular: *five* is more than *four*. In the latter, quantity is expressed as plural: I have seven apples, *five* are red. Only the former is metarepresentational, that is, mentioned rather than used, and consequently an object of thought in its own right.

Furthermore, writing is instrumental to thinking in general as a form of metalinguistic knowledge, that is, knowledge 'about' the lexical, grammatical, and logical properties of the language. Vocabulary knowledge, for example, is both greatly extended by reading (Anglin, 1993), and reflective knowledge about words, as required for judgments of synonymy, for example, serves as a major aspect of measured intelligence in a literate society (Stanovich, 1986). Formal reasoning tasks which require judgments of validity rather than truth or agreement, too, appear to be related to literacy, especially the specialized literacy of schooling (Ong, 1992; Olson, 2009).

To account for such effects ranging from phonological awareness to awareness of sentence meaning (as opposed to what a speaker means by the sentence), Olson and Oatley (2014) has proposed that writing be seen as quoted speech, and that reading and writing be viewed as learning to deal with quoted utterances rather than with ordinary direct utterances. This associates writing with a special feature of language known as reflexivity and as roughly equivalent to the use – mention distinction and language – metalinguistic distinction. Thus, in writing and reading one is not merely using an expression but quoting and thereby mentioning a linguistic form. The special interpretation called for in such a context depends upon the special properties of the form, its 'linguistic meaning' or 'sense,' rather than its referential meaning as well as allowing for judgments of similarity of meaning and judgments of logical entailments. Some writers such as Havelock (1982) and Ong (1992) argue that "all elaborate, linear, so-called 'logical' explanation depends on writing" (Ong, 1992: p. 312). Admittedly, not every culture and/or every reader exploits these distinctions as they rely heavily on particular institutional arrangements such as schools, professions, and academic disciplines. However, no reader reading a word- and sentence-segmented alphabetic text can fail to acquire the special metalinguistic concepts of sound, word, and sentence. The remainder comes with more specialized study, normally in schools.

Writing as Normative Practice

Writing does not merely transcribe speech but presents speech in terms of an idealized model. When learning to write, one is taught not to transcribe speech but rather to edit the speech by removing false starts, correcting errors, and adding needed qualifiers to make the transcription more complete, correct, and formal than the more spontaneous oral utterance. Writing in a literate culture is a normative practice, which is to say it involves the correct, authorized forms of words, sentences, and texts. Admonitions to 'speak grammatically' are in fact admonitions to speak as one would write. Although linguists abhor so-called prescriptive grammars, claiming that all utterances of a native speaker are grammatical, schools are regularly charged with teaching 'proper English.' In this way, writing comes to serve as a model for speaking.

Reading texts and composing them provide a special forum for coordinating the content of a domain with the syntactic and rhetorical demands of the written models available. van Dijk and Kintsch (1983) showed how readers coordinate linguistic form with conceptual structure in interpreting texts and Scardamalia and Bereiter (1986) showed how in writing, students coordinated their ideas with available discourse forms.

Thinking for Speaking and Writing

Slobin (1996) has suggested that the relation between language and thought be reconsidered in terms of thinking for writing and thinking for speaking as any writer will attest. Biber (1988) found important linguistic differences between spoken and written discourse especially in the informational density allowed through complex noun phrases and in the richness

and diversity of lexical choice. He attributed these differences primarily to the opportunities for editing and revision allowed by writing. Hayes (2009) has examined this composing and editing process empirically. Stromquist and Verhoeven (2004) compared the oral and written narrative accounts of a social event produced by children, adolescents. The written accounts by the oldest subjects included a vastly larger range of speech act and mental verbs to describe interactions. Berman and Ravid (2009) examined student's oral and written productions of expository texts and again found that the differences, primarily in the use of causal links, appeared primarily in the written discourse of older, college-age subjects. Snow and Uccelli (2009) have examined the complexity of structures that must be learned in mastering academic prose. These topics are examined in detail in three recent Handbooks, Bazerman (2008), Olson and Torrance (2009), and Beard et al. (2009).

Writing and Knowledge

Although all of the peoples of the world know a great deal about nature, themselves, and the social world, writing systems in the modern world have largely pre-empted such local knowledge, overwriting it with institutional forms of knowledge based on textual archives and specialized interpretive practices. Havelock (1982) suggested that writing was responsible for divorcing the knower from the known, allowing the latter to accumulate and become the subject of research and revision when the social conditions were suitable. Contemporary forms of knowledge are without exception text based, with access to these forms limited to those willing to undergo years of specialized training in the formulation and interpretation of the principle documents of the tradition. Stock (1983) called such traditions 'textual communities' and Smith (1990), a sociologist, characterizes the organization of power in contemporary society in terms of objectivized forms of text-based knowledge: "The primary mode of action and decision in … business, government, the professions … are textual. The realities to which action and decision are oriented are virtual realities vested in texts and accomplished in the distinctive practices of reading and writing" (Smith, 1990: pp. 61–62). Psychologists have only begun to explore the implications of living and working in such text-based environments.

Literacy and Social Development

Because literacy plays such a prominent role in modern societies, it is often assumed that the route to social development is through teaching people to read and write, a view supported by UNESCO. Research and practice have shown that to bring about cultural and social transformation, literacy must be seen as an activity embedded in social and cultural practice. Literacy, bureaucratic institutional structures with explicit procedures and accountability, and democratic participation are mutually reinforcing. Rather than being seen simply as a goal, literacy has come to be seen as a means to fuller participation in the institutions of the society, whether in law, science, or literature (Street, 1984), as well as a means for their transformation.

See also: Chinese Literacy Acquisition: A Multidimensional Puzzle; Language and Thought: The Neo-Whorfian Hypothesis; Speech Production, Psychology of; Writing Instruction; Writing Systems; Writing, Evolution of.

Bibliography

Adams, M.J., Treiman, R., Pressley, M., 1998. Reading, writing and literacy. In: Damon, W., Seigel, I.E., Renninger, K.A. (Eds.), Handbook of Child Psychology, Child Psychology in Practice, fifth ed., vol. 4. Wiley, New York.

Anglin, J.M., 1993. Vocabulary development: a morphological analysis. Monographs of the Society for Research in Child Development 58 (10), 1–166.

Bazerman, C., 2008. Handbook of Research on Writing: History, Society, School, Individual, Text. L. Erlbaum Associates, New York.

Beard, R., Myhill, D., Riley, J., Nystrand, M. (Eds.), 2009. The Sage Handbook of Writing Development. Sage, London.

Berman, R., Ravid, D., 2009. Becoming a literate language user: oral and written text construction across adolescence. In: Olson, D.R., Torrance, N.G. (Eds.), The Cambridge Handbook of Literacy. Cambridge University Press, Cambridge, UK.

Biber, D., 1988. Variation across speech and writing. In: Olson, D.R., Torrance, N.G. (Eds.), The Cambridge Handbook of Literacy. Cambridge University Press, Cambridge, UK.

Coulmas, F., 1989. The Writing Systems of the World. Blackwell, Oxford, UK.

Daniels, P., 2009. Grammatology. In: Olson, D.R., Torrance, N.G. (Eds.), The Cambridge Handbook of Literacy. Cambridge University Press, Cambridge, UK.

van Dijk, T.A., Kintsch, W., 1983. Strategies of Discourse Comprehension. Academic Press, New York.

de Francis, J., 1989. Visible Speech: The Diverse Oneness of Writing Systems. University of Hawaii Press, Honolulu, HI.

Gaur, A., 1987. A History of Writing, paperback ed. British Library, London (Original work published 1984).

Gelb, I.J., 1963. A Study of Writing, second ed. University of Chicago Press, Chicago.

Goody, J., 1968. Literacy in Traditional Societies. Cambridge University Press, Cambridge, UK.

Harris, R., 1995. Signs of Writing. Routledge, London.

Havelock, E., 1982. The Literate Revolution in Greece and Its Cultural Consequence. Princeton University Press, Princeton, NJ.

Hayes, J.R., 2009. From idea to text. In: Beard, R., Myhill, D., Riley, J., Nystrand, M. (Eds.), The Sage Handbook of Writing Development. Sage, London, pp. 65–79.

Houston, S., 2004. Writing in early Mesoamerica. In: Houston, S. (Ed.), The First Writing: Script Invention as History and Process. Cambridge University Press, Cambridge, UK.

Johns, A., 1998. The Nature of the Book: Print and Knowledge in the Making. University of Chicago Press, Chicago.

Killeen, K., Forshaw, P.J. (Eds.), 2007. The Word and the World: Biblical Exegesis and Early Modern Science. Palgrave/Macmillan, London.

Lundberg, I., Frost, J., Petersen, O., 1988. Effects of an extensive program for stimulating phonological awareness in preschool children. Reading Research Quarterly 23, 264–282.

McLuhan, M., 1962. The Gutenberg Galaxy. University of Toronto Press, Toronto, Canada.

Morais, J., Cary, L., Alegria, J., Bertelson, P., 1979. Does awareness of speech as a sequence of phones arise spontaneously? Cognition 7, 323–331.

Olson, D.R., 1994. The World on Paper. Cambridge University Press, New York.

Olson, D.R., 2009. A theory of reading/writing: from literacy to literature. Writing Systems Research 1, 51–64.

Olson, D.R., Torrance, N.G. (Eds.), 2009. The Cambridge Handbook of Literacy. Cambridge University Press, Cambridge, UK.

Olson, D.R., Oatley, K., 2014. The quotation theory of writing. Written Communication 31, 4–26.

Ong, W., 1992. Writing is a technology that restructures thought. In: Downing, P., Lima, S.D., Noonan, M. (Eds.), The Linguistics of Literacy. John Benjamins, Amsterdam, pp. 293–319.

Read, C.A., Zhang, Y., Nie, H., Ding, B., 1986. The ability to manipulate speech sounds depends on knowing alphabetic reading. Cognition 24, 31–44.

Robertson, J.S., 2004. The possibility and actuality of writing. In: Houston, S. (Ed.), The First Writing: Script Invention as History and Process. Cambridge University Press, Cambridge, UK, pp. 16–38.

Saenger, P., 1997. Space between Words: The Origins of Silent Reading. Stanford University Press, Stanford, CA.

Sampson, G.R., 1987. Writing Systems: A Linguistic Introduction, revised ed. Hutchinson, London.

Scardamalia, M., Bereiter, C., 1986. Research on written composition. In: Wittrock, M. (Ed.), Handbook on Research on Teaching, third ed. Academic Press, New York, pp. 778–803.

Scribner, S., Cole, M., 1981. The Psychology of Literacy. Harvard University Press, Cambridge, MA.

Slobin, D., 1996. From "thought and language" to "thinking for speaking". In: Gumpertz, J.J., Levinson, S.C. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, Cambridge, UK, pp. 70–96.

Smith, D., 1990. Texts, Facts, and Femininity. Routledge, New York.

Snow, C., Uccelli, P., 2009. The challenge of academic language. In: Olson, D.R., Torrance, N.G. (Eds.), The Cambridge Handbook of Literacy. Cambridge University Press, Cambridge, UK, pp. 112–133.

Stromquist, S., Verhoeven, L., 2004. Relating Events in Narrative. In: Typological and Contextual Perspectives, vol. 2. Erlbaum, Mahwah, NJ.

Stanovich, K.E., 1986. Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly 21, 360–407.

Stock, B., 1983. The Implications of Literacy. Princeton University Press, Princeton, NJ.

Street, B., 1984. Literacy in Theory and Practice. Cambridge University Press, Cambridge, UK.

Trigger, B., 2004. Writing systems: a case study in cultural evolution. In: Houston, S. (Ed.), The First Writing: Script Invention as History and Process. Cambridge University Press, Cambridge, UK.

Taylor, I., Taylor, M.M., 1995. Writing and Literacy in Chinese, Korean and Japanese. John Benjamins, Amsterdam.

Vernon, S.A., Ferreiro, E., 1999. Writing development: a neglected variable in the consideration of phonological awareness. Harvard Educational Review 69, 395–415.

Vygotsky, L.S., 1986. Thought and Language. MIT Press, Cambridge, MA.

Wrongful Convictions

Marvin D Free, Jr., University of Wisconsin–Whitewater, Whitewater, WI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

The earliest recorded cases of wrongful conviction in the United States are the Salem witch trials (1692–93). Yet interest in the formal study of wrongful convictions did not emerge until the early 1930s and research into this topic did not begin to accumulate substantially until the advent of DNA testing in the late 1980s. Some of the more commonly reported factors associated with known wrongful convictions are witness error, police and prosecutorial misconduct, false confessions, use of informants/snitches, forensic errors/junk science, perjury by criminal justice officials, insufficient evidence to support a conviction, and ineffective assistance of counsel. Plea bargains may additionally contribute to the wrongful conviction of innocents. Most known wrongful convictions involve murder and rape. Recent research discloses the presence of both gender and racial differences. African Americans in particular appear to be overrepresented among those who have been falsely convicted.

Scholars frequently distinguish between *legal* innocence and *factual* innocence when analyzing wrongful convictions. The former suggests that the individual is cleared of the charges as a result of procedural errors that violated his or her constitutionally protected rights, whereas the latter implies that the individual is innocent either because he or she did not commit the crime or no crime was actually committed. Consequently, data on wrongful convictions vary somewhat from one study to another. In practice, these two definitions tend to blur, because convictions characterized by serious constitutional issues or due process errors enhance the likelihood that factually innocent persons will be found guilty.

Although no knowledgeable scholar would deny the existence of wrongful convictions, a lack of consensus exists regarding the prevalence of wrongful convictions in the United States. Although much of this variation results from differing definitions of wrongful convictions, the use of different methodologies additionally contributes to these variable statistics. Both subjective and objective approaches are used by researchers. Some researchers, for instance, ask criminal justice personnel such as prosecutors, defense attorneys, judges, and sheriffs to estimate the percentage of cases in which innocents were wrongly convicted. Using this method, Huff et al. (1996) found that 72% of their respondents thought that fewer than 1% of all criminal cases resulted in wrongful convictions. An additional 20% of their subjects believed that wrongful convictions occurred in 1–5% of all criminal cases. More recently, Poveda (2001) utilized inmate self-reports to establish the frequency with which innocent defendants are found guilty. That analysis revealed that 15.4% of the inmates in the three states sampled reported having been wrongfully convicted. Although this method undoubtedly exaggerates the problem, it should be noted that the likelihood of denying guilt varied by type of offense. Those convicted of sexual assault or some other sex offense that typically relies heavily on victim and eyewitness accounts were the most likely to report being wrongly convicted. In contrast, those who were convicted of drug-related or property offenses were the least likely to express their innocence.

Some scholars focus on a specific case in which a factually innocent individual was pronounced guilty. This case study method, although providing some depth to our understanding of the wrongful conviction process, makes generalization impossible. Still other researchers extrapolate the prevalence of wrongful convictions from known wrongful convictions. This method has resulted in an estimated error rate as low as 0.027%. More typically, however, scholars report estimates that range from approximately 3–5% of all felony cases (Gould and Leo, 2010). The scope of this problem becomes self-evident when these statistics are translated into an approximation of wrongfully convicted individuals. Using a conservative figure (0.5%) to estimate the number of wrongful convictions, one finds that almost 7500 innocent people may have been falsely convicted of index crimes (murder and nonnegligent manslaughter, forcible rape, aggravated assault, robbery, burglary, larceny-theft, motor vehicle theft, and arson) in 2000. And when the research on wrongful convictions is disaggregated by race, people of color are disproportionately represented among the cases.

The formal study of wrongful convictions in the United States is commonly associated with the publication in 1932 of *Convicting the Innocent: Sixty-Five Actual Errors in Criminal Justice* by Yale law professor Edwin Borchard. Nevertheless, research in this area proceeded rather sporadically until the publication of a study by Hugo Bedau and Michael Radelet in 1987 in the *Stanford Law Review*. In that article the authors identify 350 cases in which defendants were erroneously convicted, 23 of which culminated in execution. Interest further intensified when DNA testing began to be used to exonerate innocent persons. Over 300 individuals have been exonerated through DNA testing since 1989. Approximately 62% of the DNA exonerees are African American.

A Brief History of Wrongful Convictions

The first recorded cases of wrongful conviction in the United States occurred prior to the founding of the country. The Salem witch trials during the late seventeenth century were influenced by the Puritan view that they were chosen by God to create a place of perfection on earth. Consequently, religion permeated all aspects of Puritan social life and prohibited the

rejection of these precepts by anyone. Any abnormal behavior was envisioned as an act of the devil that required immediate action. Puritans also endured high mortality rates and ongoing conflicts with the indigenous population, which further promoted fear, hysteria, and revenge. Although the first accusations of witchcraft were against women who lived on the margins of society, the hysteria soon spread to include more respectable members of society. By the termination of these events, almost 30 people had died and at least 100 more were confined in cramped prison cells. Some were tortured for months at a time. Moreover, at least seven men were charged with being wizards and performing witchcraft (Hill, 1995). The hysteria that prevailed at the Salem witch trials points to a variety of factors that continue to be found in many wrongful convictions. The suggestibility of children to events that never occurred, the inability of a coerced confession to result in the truth, and a rush to judgment are all potential factors in wrongful convictions today.

The first known wrongful murder conviction in the United States transpired during the early part of the nineteenth century in Vermont. When Russell Colvin disappeared from Manchester, Vermont in 1812, Jesse and Stephen Boorn, his two brothers-in-law, were suspected of murder, as they had frequently voiced their displeasure over their sister's spouse. Seven years later a series of events – including buried bone fragments believed to be the remains of Colvin that were uncovered by a dog – led to Jesse Boorn's arrest. A warrant was also issued for his brother, who had moved to New York. While in jail, Jesse met Silas Merrill, a forger, who falsely testified that Jesse had confessed to clubbing Colvin during an argument. Merrill further contended that Jesse admitted to hiding Colvin's remains with the assistance of his brother and his uncle, Amos Boorn, who allegedly had cut Colvin's throat with a knife. In exchange for his testimony, Merrill was released from custody. Fearing a death sentence, Jesse confessed to the crime but placed the primary blame on Stephen, because he was in another state. Unexpectedly, however, Stephen returned to Vermont to refute the allegations against him. At this point, Jesse recanted his testimony. In a miscalculation designed to avoid the death penalty, Stephen also falsely confessed to the murder but insisted that the altercation was in self-defense. Before the trial the physicians who had verified that the bone fragments were human recanted their earlier statements. Nevertheless, confessions from both defendants and statements from witnesses that the two brothers had been seen arguing with Colvin on the day of his disappearance carried the day and both were found guilty and sentenced to death. When the Vermont General Assembly convened to review the case, Jesse had his sentence commuted to life. With Stephen's impending execution scheduled for 28 January 1820, it took a series of fortuitous events to prove his innocence. An article appearing in the 28 November 1819 issue of the New York Evening Post inspired Tabor Chadwick to compose a letter to the newspaper in which he acknowledged knowing a man named Russell Colvin when he lived in New Jersey. His letter was then read by a former Vermont resident, James Whepley, who successfully cajoled an unwilling Colvin to return to Vermont in time to prevent Stephen from being executed (Warden, n.d. (a)). This early nineteenth-century case clearly illustrates the potential harm that can arise from relying on questionable forensics, jailhouse snitches, and confessions obtained under duress, factors that continue to be associated with many wrongful convictions in the twenty-first century.

The year 1989 marks the first time that DNA evidence was used to exonerate a convicted inmate. This Illinois case involved a 16-year-old female who claimed that she had been abducted on 9 July 1977, by three men, one of whom raped her in the backseat of an automobile. A medical examination disclosed that her underpants contained several hairs and a large stain that was erroneously thought to be a seminal stain but that actually was a combination of semen and vaginal fluid. The rape kit also included a vaginal swab that was preserved as evidence. A sketch of the alleged perpetrator portrayed a young clean-cut white man with long hair. The victim later identified Gary Dotson as her assailant while viewing photographs. When he was arrested by the police 5 days after the alleged rape he sported a prominent mustache, which he could not have grown since the incident. Perjured forensic testimony, prosecutorial misconduct, and an ineffective public defender resulted in Dotson being found guilty and being sentenced in 1979 to concurrent sentences of 25–50 years for rape and aggravated kidnapping. In 1985 the victim recanted her accusations after admitting that she concocted the story to disguise the fact that she had engaged in a sexual relationship with her boyfriend earlier and was afraid that she might be pregnant. Illinois Governor James Thompson subsequently commuted Dotson's sentence to time served and placed him on parole. In 1987 Dotson was charged with a parole violation. Although the State Attorney's Office eventually dropped the charges, the Illinois Department of Corrections placed a 'parole hold' on Dotson, which required him to remain in jail while awaiting the Illinois Prison Review Board's decision. At the hearing, Dotson had his parole revoked and his original sentence reinstated. However, Dotson obtained his release from prison later that year when Governor Thompson agreed to give him 'one last chance.' Nonetheless, just 2 days later he was rearrested for an altercation resulting from a refusal to pay a contested bill at a tavern. Although the charges were later dropped, the Prisoner Review Board found Dotson guilty of a 'technical' violation of his parole and ordered him back to prison for 6 months. In August 1988, DNA testing, known as polymerase chain reaction (PCR), positively excluded him as contributor of the semen and positively included the victim's boyfriend. On 14 August 1989, all charges against Dotson were finally dropped and on 9 January 2003, he received a full pardon (Warden, n.d. (b)).

In 1993 DNA evidence was used to prove that Kirk Bloodsworth was innocent of a 1984 rape and murder conviction, making him the first person on death row in the United States to be exonerated. The incident that led to his conviction in 1985 involved the rape and murder of Dawn Hamilton, a 9-year-old child in Baltimore County, Maryland. The primary evidence used to obtain the conviction was the testimony of five witnesses who placed Bloodsworth in proximity to the crime as well as questionable forensic evidence that supposedly matched his shoes to marks on the child's body. After the Maryland Court of Appeals overturned his conviction in 1986, he was retried and found guilty again. This

time, however, Bloodsworth received two life sentences. In 1992, Centurion Ministries of Princeton, New Jersey agreed to review his case and sought DNA testing utilizing PCR technology. Tests conducted by Forensic Science Associates and confirmed by the Federal Bureau of Investigation ruled out Bloodsworth as the perpetrator. He gained his freedom on 28 June 1993, and received a full pardon from Maryland Governor William Schaefer in December 1994. In spring 2003, new stains were discovered while reviewing evidence from the cold case. These stains, which were DNA tested and compared against a national DNA database, culminated in the charging of Kimberly Ruffner with Hamilton's murder on 5 September 2003 (Warden, n.d. (c)).

Sources of Data on Wrongful Convictions

Although a number of websites contain some information on wrongful convictions, the data sources discussed in this section represent the best-known and most respected websites currently available. For those desiring an international database on wrongful convictions, the Forejustice website (http://forejustice.org) includes over 3650 cases from 102 countries dating back to the fifth century. Conveniently arranged by either location or last name, the serious researcher will find this site a trove of invaluable information both current and historical. Although the amount of information on any individual case is limited, the inclusion of cases from outside the United States makes it possible to examine wrongful convictions from different geographical locations.

With 626 documented cases in mid-2013, the United Kingdom is second only to the United States in number of reported wrongful convictions. The earliest wrongful convictions in the United Kingdom are from the sixteenth and seventeenth centuries, when almost 90 individuals were wrongfully convicted of witchcraft. An analysis of a more contemporary period (1970s and later) reveals that murder and manslaughter cases represent the single largest category among the false convictions. The dominance of this offense is illustrated by the fact that all of the reported wrongful convictions for rape and sexual assault, robbery, terrorism, and drugs fail to approach the number of innocents falsely convicted of murder and manslaughter.

Australia is another country with a substantial number of reported wrongful convictions. There were 108 Australian cases in mid-2013. As previously observed in the United Kingdom, murder (36) is the most common offense leading to a false conviction in Australia. The 14 wrongful rape and sexual assault convictions make this crime the second most common among the wrongful convictions. With 12 cases, wrongful assault convictions comprise the third most frequently occurring category.

An analysis of Canada during this same time period discloses 79 cases, the largest number of which involve murder. Wrongful murder convictions constitute approximately 42% of the cases from this country. And if manslaughter and infanticide cases are included, then this category swells to nearly half of the reported wrongful convictions in Canada. Rape and sexual assault cases are the second most common offense, mirroring the findings of both the United Kingdom and Australia.

Unfortunately, the quantity and quality of the information available for each country varies, making cross-cultural comparisons problematic at best. Twenty-six countries on the Forejustice list contain only a single wrongful conviction. Moreover, countries that tend to impede the free exchange of information derogatory to the state have few reported false convictions. For instance, there is only 1 case each from Cuba and the Russian Federation, 2 cases from North Korea, and 10 cases from China.

Of special significance on the Forejustice website is a link to *Justice Denied*, a magazine founded in 1998 by the Justice Institute, a nonprofit organization. Available in both electronic and print formats, *Justice Denied* is devoted exclusively to wrongful conviction cases. Particularly helpful is a free online index of every published article beginning with issue one. There is additionally an online bookstore with over 80 titles of books related to this subject matter.

Since the 1999–2000 academic year, Northwestern University's School of Law has been home to the Center on Wrongful Convictions (CWC). With a paid staff of five, the CWC maintains a list of exonerated individuals from state and federal courts plus the District of Columbia. Unfortunately, many of the entries are limited to the names and perhaps relevant dates of the wrongful convictions. An exception is the list of Illinois exonerations, which contains considerable information on virtually all of the cases in that state. In December 2000 the CWC began collaborating with the Innocence Project to further the development of innocence projects in other jurisdictions. The Center partnered with the University of Michigan's School of Law in May 2012 to launch the National Registry of Exonerations (NRE). Later that same year it established the first ever Women's Project. This project emphasizes both litigation and research involving falsely convicted women. Initially, the research will concentrate on cases in which wrongful convictions were obtained through false confessions and circumstantial evidence. More information about the CWC is available on its website, which can be accessed at www.law.northwestern.edu/wrongfulconvictions.

The Innocence Project is affiliated with Yeshiva University's Benjamin Cardoza School of Law in New York. Founded in 1992 by Barry Scheck and Peter Neufeld, the home page states that "The Innocence Project is a national litigation and public policy organization dedicated to exonerating wrongfully convicted individuals through DNA testing and reforming the criminal justice system" (www.innocenceproject.org). Over 300 post-conviction exonerations in the United States arranged in alphabetical order are profiled on the website. Additionally, there is a brief discussion of non-DNA exonerations with a link to a more thorough analysis of five cases.

Information on wrongfully convicted death row inmates in the United States is available at the Death Penalty Information Center (DPIC) (www.deathpenaltyinfo.org). The DPIC provides the user with an enumeration of over 140 individuals who were sentenced to death but were exonerated through acquittal of all charges related to the crime that placed them on death row, dismissal of such charges by the prosecution, or a full pardon based on evidence of innocence. This list, which is

continuously updated, includes all individuals who have been released since 1973.

The jointly sponsored NRE is the most recent entrant into the wrongful conviction arena. A product of well-respected law schools from Northwestern University and the University of Michigan, information about the NRE list of exonerees is available online at www.law.umich.edu/special/exonerations/Pages/about.aspx. A rapidly growing database, the NRE enumerates almost 1200 exonerations since 1989. This is an especially user-friendly website as the list of exonerees is arranged so that cases can be sorted by name, state, county, date of conviction, date of exoneration, whether DNA was used in the exoneration, and contributing factors. Narratives of each of the cases can be accessed as well.

Factors Associated with Wrongful Convictions

It is beyond the scope of this article to fully address the causes of wrongful convictions, as they are many and varied. Nevertheless, the disproportionality of people of color in the criminal justice data suggests the presence of racial discrimination in both legislation and enforcement of laws. Selective enforcement of the law, racial profiling, prejudice and discrimination, and poverty continue to contribute to the overrepresentation of minorities in the criminal justice system. Drug prohibition in the United States has had a particularly deleterious effect on the nonwhite population, beginning with laws that prohibited the Chinese from importing opium during the nineteenth century and continuing with the focus on crack cocaine during the mid-1980s. Although some researchers have analyzed these issues, much of the literature on wrongful convictions has concentrated instead on factors that directly result in the conviction of an innocent defendant. The remainder of this section examines these factors.

Witness error is a major problem in known wrongful convictions. Mistaken eyewitness identification was a factor in 43% of all exonerations in the United States from 1989 to 2012 (Gross and Shaffer, 2012). Psychological, systemic, societal, and cultural variables can affect one's perception. Although most eyewitness misidentification is unintentional, research has demonstrated that cross-racial identifications are most problematic. Incidents in which white subjects are attempting to identify black subjects are especially vulnerable to misidentification.

Police and prosecutorial misconduct are factors in many documented wrongful convictions as well. A study of DNA exonerations by the Innocence Project revealed that police misconduct was present in half of the cases and prosecutorial misconduct occurred in 42% of the wrongful convictions (Scheck et al., 2000). The organizational culture of the police and district attorney's office can produce a climate that enhances the likelihood of wrongful convictions if expedience and winning at all costs are emphasized. Police and prosecutorial misconduct can manifest itself in numerous ways. For instance, the police may 'coach' the witness during the lineup; employ brutality, force, or deceit to obtain a confession; plant evidence; mishandle physical evidence; or threaten potential witnesses for the defense. Similarly, prosecutorial misconduct may result from improper behavior during grand jury proceedings and the dismissal of potential jurors because of their race, ethnicity, or gender. Prosecutors sometimes use known false or misleading evidence, suppress exculpatory evidence, and withhold information that the witness for the prosecution received immunity or other rewards for testifying. Use of improper closing arguments, badgering hostile witnesses, mischaracterizing the facts and evidence of the case, and mishandling the evidence are additional potential examples of prosecutorial misconduct. The probability of a wrongful conviction is further enhanced when police and prosecutors fail to adequately investigate other potential perpetrators of the crime and ignore evidence that fails to corroborate their preconceived views.

False confessions are more likely to occur in cases in which the defendant is a juvenile, has a mental deficiency, or has been interrogated over a sustained period of time. A 2012 study by the NRE disclosed that juveniles were five times more likely to falsely confess than adults. Further, mentally disabled individuals were nine times more likely to falsely confess than individuals without such deficiencies (Gross and Shaffer, 2012). Duress, coercion, intoxication, a misunderstanding of the law, fear of violence by the police, actual harm by the police, the threat of a harsh sentence if a confession is not forthcoming, and a misunderstanding of the situation also affect the probability of a false confession. The extent to which false confessions contribute to the wrongful conviction of innocents varies in the research. A report that examined wrongful convictions in Illinois, for example, found that false confessions played a role in about 53% of those convictions (Warden and Fredrickson, 2012). In contrast, false confessions accounted for only 16% of the exonerations in the United States as a whole from 1989 to 2012 (Gross and Shaffer, 2012).

Although the Innocence Project suggests that over 15% of DNA exonerations involve the use of informants/snitches, about 46% of all death row wrongful convictions examined by the Center on Wrongful Convictions involved the testimony of witnesses with incentives to lie. In many of these cases the jury was unaware that the informant/snitch had an agreement with the prosecutor to either get a charge reduction or to be released in exchange for incriminating testimony (Warden, 2005). Moreover, racial disparity in drug enforcement may be exacerbated through the use of informants in minority communities. Because law enforcement typically focuses its attention on high-crime urban communities where people of color are more likely to reside, and given the incentive that informants have to lie, inhabitants of these areas are at greater risk of "false accusations, mistaken warrants, erroneous raids, and wrongful convictions" (Natapoff, 2009, p. 113).

In the late 1980s the utilization of DNA evidence to exonerate the innocent and to convict the guilty represented a milestone in the discovery of wrongful convictions. Although DNA analysis has benefitted from extensive research, other forensic techniques such as hair analysis, bite mark comparisons, firearm tool mark analysis, and shoe print comparisons have not received similar scrutiny, yet they have been used to convict innocent individuals. Even cases involving DNA evidence may be subject to contamination, distortion of findings, mislabeling, as well as other issues. Nor is this problem

isolated to a few incidents. The Federal Bureau of Investigation, the US Army Criminal Investigation Laboratory, as well as laboratories in various states have all been accused of the improper handling of forensic evidence in the recent past. According to the Innocence Project, unvalidated or improper forensic science is a factor in approximately half of the DNA exonerations.

Perjury by criminal justice officials, insufficient evidence to support a conviction, and ineffective assistance of counsel complete the list of commonly reported factors in the wrongful conviction literature. Ineffective assistance of counsel is undoubtedly underreported in the literature because of its subjective nature. Nonetheless, it is potentially problematic since a strong defense can deter a wrongful conviction. Insufficient funding, lack of mechanisms for monitoring the quality of legal representation, an unmotivated defense counsel, the presumption of guilt that pervades the criminal justice system, and the difficulty of proving the presence of ineffective counsel on appeal can influence this process.

Although not as frequently identified in the literature, plea bargains may lead to wrongful convictions. A plea bargain may look attractive to an innocent defendant with a prior record or one who does not qualify for a public defender. And with their limited resources, innocence projects are more likely to devote their attention to cases involving extended sentences or the death penalty, thereby excluding many plea-bargained cases.

Wrongful Convictions in the United States

Before discussing wrongful convictions it should be noted that many false convictions go undetected. Organizations that defend wrongfully convicted clients have limited financial resources and tend to concentrate on the most egregious cases, thereby potentially ignoring many cases in which a defendant is falsely convicted but the circumstances do not lend themselves to an easy legal remedy. Moreover, cases in which wrongfully accused defendants receive probation or suspended sentences are less likely to be perceived as warranting the expenditure of precious capital by organizations that defend the innocent. Defendants with few financial resources may also be more willing to accept erroneous guilty verdicts. And because wrongly convicted defendants may possess a prior criminal record, there is frequently little outcry from the public to correct these miscarriages of justice. With these caveats in mind, an analysis of known wrongful conviction cases is in order.

A comprehensive review of known wrongful convictions in the United States was conducted by Samuel Gross and Michael Shaffer at the NRE in 2012. They examined 873 exonerations from January 1989 through 1 March 2012. Some of the more salient findings appear below.

- Almost all of the exonerations involve male defendants. Only 7% of those exonerated were female.
- Minorities comprise the largest number of exonerations. Half of the people exonerated were African American. Whites, Hispanics, and American Indians or Asians represented 38, 11, and 2% of the exonerated population, respectively.
- Although only 37% of the exonerations are the result of DNA evidence, the use of DNA in wrongful conviction cases is increasing.
- Wrongful convictions are not distributed evenly throughout the United States. Illinois, New York, Texas, and California led the United States in exonerations.
- The vast majority of exonerations are for the offenses of murder and rape. Almost half of the exonerations were homicides, whereas sexual assaults accounted for over one-third of the exonerations.

The study additionally investigated factors that contribute to false convictions. As other scholars have observed, wrongful convictions frequently result from the culmination of multiple factors. Slightly over half of the wrongful convictions included perjury or false accusation. Mistaken eyewitness identification (43%) and official misconduct (42%) were the second and third most important contributors to wrongful convictions. Approximately one-fourth of the cases disclosed the presence of false or misleading forensic evidence, whereas false confessions were a problem in 16% of the false convictions. The investigation further revealed that the frequency of these factors varied according to the type of offense. For instance, mistaken eyewitness identification was the primary contributing factor in both sexual assault and robbery exonerations, although perjury or false accusation was the most common contributing factor in homicide exonerations.

African Americans and Wrongful Convictions

Minorities in general and African Americans in particular are overrepresented at every stage of the criminal justice system. On any given day, for instance, 10% of all African American men in their 30s are either in jail or prison. Because blacks comprise the largest racial/ethnic category of exonerated individuals in the United States, this section examines black wrongful convictions in greater detail.

A small, albeit growing, segment of the literature on wrongful convictions focuses on African Americans. In *Convicted but Innocent: Wrongful Conviction and Public Policy*, Huff et al. (1996) found that African Americans were overrepresented in wrongful convictions. The authors argue that witness misidentification (especially in cases involving white witnesses and black suspects) and prejudice by police and prosecutors, along with other factors, were responsible for this disparity. More recently, an examination of US exonerations from 1989 to 2003 documented racial disparities, particularly in rape cases and those with juvenile defendants. Although the data did not permit a determination of the reasons for these disparities, the authors speculate that cross-racial misidentification in many of the interracial rape cases contributed to the false convictions in addition to racial bias and discrimination. They further note the potential negative impact of "a dual system of juvenile justice …, one track for white adolescents, a separate and unequal one for black adolescents" (Gross et al., 2005, p. 551).

Marvin Free and Mitch Ruesink (2012) provide a detailed analysis of 343 wrongfully convicted African American men in their book, *Race and Justice: Wrongful Convictions of African*

American Men. Classifying their cases according to the most serious offense that led to the wrongful conviction, they found that slightly over half of the false convictions were for murder/attempted murder. Nearly one-third of their wrongful convictions involved rape and sexual assault. However, fewer than 12% were drug-related and fewer than 5% were robbery cases. Although wrongful convictions occurred in 31 states plus the District of Columbia, the cases were unevenly distributed. With 52 false convictions, Texas headed the list, followed closely by Illinois, which had 47 wrongful convictions. Ohio was the third-highest state with 26 cases.

There were 174 murder/attempted murder cases in the study. Most commonly, the case involved a single charge of murder. When additional charges were present they usually involved rape/sexual assault and/or robbery. Typically multiple factors contributed to the failed justice. Witness error (42.5%), police misconduct (38.5%), prosecutorial misconduct (36.2%), and use of informants/snitches (33.3%) were the most common factors associated with these wrongful convictions.

As noted earlier, rape/sexual assault ($n = 109$) was the second most common offense category. An analysis of these cases revealed that almost 70% of these wrongful convictions included at least one additional charge, the most frequent of which was robbery. Of particular note was the significance of witness error in rape/sexual assault wrongful convictions. Witness error was a contributing factor in almost 93% of the cases. And in many instances it was the primary or only factor that was responsible for the erroneous conviction. Forensic errors were a distant second place, being found in over 30% of the convictions for rape/sexual assault. Police misconduct (about 23%) and prosecutorial misconduct (about 15%) were the third and fourth most common factors identified in the study.

Women and Wrongful Convictions

Most of the wrongful conviction literature focuses on wrongly convicted men, because most known wrongful convictions are for murder and rape. Thus, false convictions tend to be framed as a male problem. Yet the Bureau of Justice Statistics reveals that the number of incarcerated women in the United States increased almost sixfold from 1980 through 2011 and 1 in every 56 women will be incarcerated during her lifetime. Although only 7% of US exonerations from 1989 through 2012 involved women – a figure comparable to the distribution of women in the prison population – women are more likely than men to be incarcerated for crimes for which there is seldom biological evidence that can be subjected to DNA testing. Consequently, many cases in which women have been falsely convicted are overlooked. Although the literature in this area is still in its infancy, a sparse but slowly emerging body of research suggests that meaningful gender differences exist in wrongful convictions. For instance, false female convictions are more likely to be based entirely on circumstantial evidence than false convictions of men, according to Rob Warden, the executive director of the CWC. There is also evidence that suggests that female defendants who are commonly accused of harming their children or other intimates may be more easily induced to provide false confessions during police interrogations than their male counterparts.

Research on wrongly convicted women in the United States by Konvisser (2012) uncovered 138 exonerations and pardons involving women since 1900. Her study further revealed that 70% of the female exonerees were white. African American women and Hispanic or Latina women comprised 28 and 4% of the cases, respectively. Many of the women in her sample had been falsely convicted of nonexistent crimes. Murder (33%) was the most common offense with which the wrongly convicted women had been charged. An additional 15% had been charged with some form of child abuse or neglect. Drug-related charges were found in 12% of the cases. The remainder of the cases included a potpourri of diverse crimes.

An earlier study in this area examined more contemporary wrongful convictions. Utilizing two large electronic databases Ruesink and Free (2005) developed a list of wrongly convicted women in the United States since 1970. They then compared their sample with a random sample of male false convictions taken from the Forejustice database. Their results revealed that male wrongful convictions were more likely to include charges of murder and rape, whereas female wrongful convictions were more likely to include charges of child abuse. When factors associated with wrongful convictions were analyzed, again gender differences were present. Although witness error and overzealous or unethical police and prosecutors were the first and second most common factors associated with known wrongful convictions in general, the rank order of these factors was reversed in female wrongful convictions. The importance of unethical police and prosecutors was particularly pronounced in female false convictions, being present in 86% of the cases. An analysis of the racial identity of their subjects disclosed that 37% of the falsely convicted women were black. Racial differences were also detected. African American women were more likely to be wrongly convicted of murder and drug-related offenses, whereas white women were more likely to be wrongly convicted of child abuse. And finally, the two most common factors associated with false convictions among black women were perjury by criminal justice officials (53%) and prosecutorial misconduct (40%). In contrast, the two most common factors leading to wrongful convictions among white women were prosecutorial misconduct (35%) and coerced confessions (27%).

See also: Crime, Race and Ethnicity; Criminal Justice, Sociology of; Critical Criminology; Race: Prespectives and Crime; Sentencing, Trends, and Disparities: Crime; Social Justice and Social Inequalities.

Bibliography

Bedau, H., Radelet, M., 1987. Miscarriages of justice in potentially capital cases. Stanford Law Review 40, 21–179.

Free, M.D., Ruesink, M., 2012. Race and Justice: Wrongful Convictions of African American Men. Lynne Rienner, Boulder, CO.

Gould, J., Leo, R., 2010. One hundred years later: wrongful convictions after a century of research. The Journal of Criminal Law and Criminology 100, 825–868.

Gross, S., Jacoby, K., Matheson, D., Montgomery, N., Patil, S., 2005. Exonerations in the United States 1989 through 2003. Journal of Criminal Law and Criminology 95, 523–560.

Gross, S., Shaffer, M., 2012. Exonerations in the United States, 1989–2012. Report by the National Registry of Exonerations http://www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf (accessed 07.01.13.).

Hill, F., 1995. A Delusion of Satan: The Full Story of the Salem Witch Trials. Doubleday, New York.

Huff, C., Rattner, A., Sagarin, E., 1996. Convicted but Innocent: Wrongful Conviction and Public Policy. Sage, Thousand Oaks, CA.

Konvisser, Z., 2012. Psychological consequences of wrongful conviction in women and the possibility of positive change. DePaul Journal for Social Justice 5, 221–294.

Natapoff, A., 2009. Snitching: Criminal Informants and the Erosion of American Justice. New York University Press, New York.

Poveda, T., 2001. Estimating wrongful conviction. Justice Quarterly 18, 689–708.

Ruesink, M., Free, M.D., 2005. Wrongful convictions among women: an exploratory study of a neglected topic. Women & Criminal Justice 16, 1–23.

Scheck, B., Neufeld, P., Dwyer, J., 2000. Actual Innocence: Five Days to Execution and Other Dispatches from the Wrongly Convicted. Doubleday, New York.

Warden, R., 2005. The Snitch System: How Snitch Testimony Sent Randy Steidl and Other Innocent Americans to Death Row. Center on Wrongful Convictions, Chicago, IL.

Warden, R., n.d. (a). America's First Wrongful Murder Conviction Case. Center on Wrongful Convictions, Chicago, IL. http://www.law.northwestern.edu/wrongfulconvictions/exonerations/vtBoornSummary.html (accessed 03.01.13.).

Warden, R., n.d. (b). Gary Dotson: The Rape that Wasn't – The Nation's First DNA Exoneration. Center on Wrongful Convictions, Chicago, IL. http://www.law.northwestern.edu/wrongfulconvictions/exonerations/ilDotsonSummary.html (accessed 02.01.13.).

Warden, R., n.d. (c). Kirk Bloodsworth. Center on Wrongful Convictions, Chicago, IL. http://www.law.northwestern.edu/wrongfulconvictions/exonerations/mdBloodsworthSummary.html (accessed 03.01.13.).

Warden, R., Fredrickson, R., 2012. The Role of False Confessions in Illinois Wrongful Conviction Cases. Center on Wrongful Convictions Special Report http://www.law.northwestern.edu/wrongfulconvictions/issues/causesandremedies/falseconfessions/FalseConfessionsStudy.html (accessed 17.02.13.).

Relevant Websites

http://www.law.northwestern.edu/cwc – Center on Wrongful Convictions.

http://www.deathpenaltyinfo.org/innocence-and-death-penalty – Death Penalty Information Center (Innocence and the Death Penalty).

http://forejustice.org – Forejustice.

http://justicedenied.org – Justice Denied Magazine.

http://www.law.umich.edu/special/exoneration/Pages/about.aspx – National Registry of Exonerations.

Wundt, Wilhelm Maximilian (1832–1920)

Lothar Sprung, Humboldt University, Berlin, Germany

© 2001 Elsevier Ltd. All rights reserved.
This article is reproduced from the previous edition, volume 24, pp. 16644–16647, © 2001, Elsevier Ltd.

Abstract

Wundt was a pioneer of modern psychology. The article begins with a brief biography and characterization of the historical context in which Wundt lived. It proceeds with an overview of his work and influence. His major work is introduced: the fundamentals of his 'individual psychology' and his 'ethnopsychology' are presented with their subject matter and methods. The following section describes Wundt's contributions to the genesis of the disciple of modern psychology and situates his life's work within the history of recent psychology. The question as to his transitory or enduring influence is posed in the final section, which offers a brief evaluation of his work and influence from today's standpoint. This section also contains a brief characterization of some of Wundt's important pupils and their continuation of his work in their home countries.

Life and Times

Wilhelm Maximilian Wundt was born on August 16, 1832 in Neckarau, Germany. From 1844, he attended secondary schools in Bruchsal and Heidelberg. He enrolled at Tüingen for medical study in 1851, then went to Heidelberg, where he earned his doctorate in 1856 with a work 'On the Behavior of Nerves in Inflamed and Degenerated Organs.' In Heidelberg, neuroanatomist Friedrich Arnold and the clinician and pathologist Ewald Hasse guided his interest toward neuroanatomical and neurophysiological problems. During the summer of 1856, he worked under Emil Du Bois-Reymond (1818–96) at the famous physiological laboratory founded in Berlin by Johannes Peter Müller (1801–56). Returning to Heidelberg, he became Hasse's assistant. He completed postgraduate work in physiology in 1857.

In 1858, Hermann Helmholtz (1821–94) was made professor at Heidelberg. Wundt was made professor of Anthropology and Medical Psychology in 1864. For 16 years, he worked alongside Helmholtz at the Physiological Institute in Heidelberg. During this period, his orientation shifted from experimental physiology to physiological psychology, and finally to philosophy. From 1864 to 1868 he was a representative to parliament in the state of Baden, where he advocated social integration for workers. He was appointed professor of Inductive Philosophy (philosophy of science) in Zurich, Switzerland in 1874. On October 18, 1875 he accepted a professorship in Leipzig in Philosophy ('with a natural scientific orientation'). In Leipzig, he founded the world's first Institute for Experimental Psychology, which soon attained international renown. Students came to Leipzig from many countries, including Bulgaria, the UK, Japan, Canada, Poland, Romania, Russia, and Hungary, and especially from the USA. Many students went on to establish institutes in their home countries, using the Leipzig Institute and its course of study as their model. In 1881, Wundt founded one of the world's first journals of psychology, *Philosophical Studies*. Wundt died on August 31, 1920 in Groβbothen near Leipzig, and was buried in the professors' section of the South Cemetery in Leipzig.

What kind of an era did Wundt live in? In the year of his birth (1832), the celebrated writer of German classical literature Johann Wolfgang von Goethe died. In 1848, the revolution that sought to achieve democratic process in Germany failed. In 1871, the Second German Reich was born after the violence of Prussia's wars against Denmark (1864), Austria (1866) and France (1870–1), and Austria was excluded from the German states' centuries-long common political development. By the time Wundt died (1920), the Second German Reich had lost World War I in 1918 and the Weimar Republic had been established. Looking back, Wundt wrote in his memoirs: 'Lived experience is the most immediate of what the gods have bestowed, but knowledge is the better portion of what he has been granted' (Wundt, 1920, p. iv).

Work and Influence

An Overview

Early in his career, Wundt focused on the experimental analysis of physiological and then psychophysiological problems. Somewhat later, he attempted to provide a philosophical foundation for his experimental work. The titles of his works published in Heidelberg vividly illustrate his path: *Contributions to the Science of Muscle Movements* (1858), *'Contributions to the Theory of Sense Perception'* (1859–62), *Lectures on the Mind of Humans and Animals* (1863), *Textbook of Physiology* (1864), *Physical Axioms* (1866), *Handbook of Medical Physics* (1867), and *Principles of Physiological Psychology* (1874). In 1902, one of his first biographers summarized: 'Evidently, it was involvement with questions of sensory physiology that first awakened in Wundt the need for a philosophical orientation' (König, 1902, p. 23).

In Heidelberg, Wundt also developed the fundamentals of his ethnopsychology *(Völkerpsychologie)*, stimulated by the work of Moritz Lazarus (1824–1903) and Hajim Steinthal (1823–99) in their *Journal for Ethnopsychology and Linguistics* published from 1859. Decades later in Leipzig, Wundt elaborated these fundamentals into a 'general psychology' and a 'general developmental psychology.' After 1875, he wrote the works that would bring him world renown as a methodologist, experimental psychologist, and ethnopsychologist. This progression, too, can be traced in the titles of his books: *System of Philosophy* (1889), *Logic: An Investigation of the Principles of Knowledge and Methods of Scientific Research* (finally a three-volume work, 1906–8), *Introduction to Philosophy* (1901), and the 10-volume work *Ethnopsychology* (1900–20). Between 1908

and 1911, he also expanded *Principles of Physiological Psychology*, his standard work on experimental individual psychology,' to three volumes. His students in Leipzig carried out the psychological experiments that would fill the 20 volumes of his *Philosophical Studies* and those of the subsequent *Psychological Studies*.

Wundt's Major Work – Individual Psychology and Ethnopsychology

Wundt's major work was devoted to psychology, which was for him the basis of philosophy. His psychology had a twofold basis, consisting of a scientifically oriented experimental 'individual psychology' and a humanistically oriented comparative 'ethnopsychology.' About his individual psychology, he wrote, 'The concept of individual psychology covers investigations whose subject are the psychological processes of individual human consciousness, inasmuch as these possess a *typical*, generally valid significance for normal consciousness' (Wundt, 1908, p. 162). He added, with regard to methodology, 'In truth, there are ... for individual psychology ... *two* aids available: casual self-observation and experimental method. The former assumes both a preparatory and a supplementary function; the latter, however, is the sole tool for analysis of elementary psychological processes' (Wundt, 1908, p. 163). 'Ethnopsychology,' in contrast, had' ... the task of complementing individual psychology with information about the general mental products of social life' (Wundt, 1908, p. 163). ' The principle domains of ethnopsychology ... are: *language, art, myth* and *customs*. They signify for the consciousness of cultures what imagination, feeling, and will signify for individual consciousness' (Wundt, 1908, p. 232). Clarifying the main aspects of ethnopsychology he continues, 'So the psychological value of language consists primarily in expressing laws of *thought*, therefore various forms of language correspond to specific steps in the development of this basic mental function. In the same way, *myth* is intimately connected with the activities of fantasy in *art*' (Wundt, 1908, p. 232). Finally, he describes the psychological value of (social) customs, writing, 'As myth does for art and religion, so do *customs* possess a central significance in their relation to two other important areas of life' (Wundt, 1908, p. 233). 'One of these areas is *society* ... It is society that first makes possible the formation of customs, the prevailing embodiments of norms of will within a society ... The other area deriving from customs is *law*' (Wundt, 1908, p. 234). He adds finally: 'Language, art, myth and customs are thus not the sole objects of ethnopsychology, but they are the ones that from the beginning express social life' (Wundt, 1908, p. 234). On methodology in ethnopsychology, he wrote, 'The fundamental method of ethnopsychology ... is ... *comparison* in the ways it is generally used in the humanities' (Wundt, 1908, p. 240). And he adds, '... ethnopsychology ... is conducted ... by assembling for comparison a collection of facts about psychological viewpoints, and through this close and direct connection to psychological analysis, comparison itself leads to the establishment of psychological laws . In this manner, the correct use of the comparative method here gains a significance similar to that of experimental procedure in individual psychology' (Wundt, 1908, p. 240).

With these ideas, Wundt created a conceptual framework that would have great consequence in the history of psychology: a scientific psychology of elementary psychological phenomena and a humanistic psychology of higher psychological phenomena. Scientific psychology is an 'individual psychology' since it gets its data from the experimental investigation of individuals. Humanistic 'ethnopsychology' is a 'general psychology,' since it examines general psychological objects. As such, it is also a 'general developmental psychology' since those objects also develop throughout history. Methodologically, ethnopsychology gets its data from the ethnological analysis of historically developed products of mind (i.e., language, art forms, myths, customs). From these, conclusions can be drawn about the psychological processes involved in their emergence. Such processes include thinking, feeling, or willing. The view that a methodical humanistic approach was the only way to study higher psychological processes was momentous in the history of psychology. According to this concept, there could be no experimental psychology of thinking, emotion, or volition. From today's standpoint it should be noted that Wundt did recognize that psychological methodology could not be solely experimental. It must also be a quasi-experimental and historical methodology. On the other hand, his view that higher psychological processes were inaccessible to experimentation proved to be false. Wundt's methodological quarrel at the beginning of the twentieth century with members of the Würzburg school around Oswald Külpe (1862–1915), which concerned the role of experimentation in the psychology of thinking, had already shown that his critique of their method as 'interrogation experiments' was untenable.

Wundt and the Genesis of the Discipline of Modern Psychology

Wundt made fundamental contributions to the genesis of the discipline of modern psychology in the nineteenth and twentieth centuries, in Germany and beyond. This genesis may be described generally in terms of six phases and their corresponding developmental features.

First, Wundt had a lasting impact on the development of the subject matter and methods of empirical psychology since the beginning of the nineteenth century with his 'individual psychology.' This is also true to a certain extent for his 'ethnopsychology,' if historically developed products of mind are regarded as historical empirical experience.

Second, Wundt was influential in institutionalization, that is, founding institutes for experimental psychology and anchoring them in the university system, as well as other forms of institutionalization, such as the founding of scholarly journals and scholarly societies. Active in scientific policy making, Wundt made a definitive contribution to institutionalization by founding the world's first psychological laboratory, the Institute for Experimental Psychology at the University of Leipzig in 1879. He also founded one of the first experimental journals in psychology, *Philosophical Studies* in 1881.

Third, Wundt can be said to have contributed to the temporary division into major schools during the period between about 1890 and 1940, inasmuch as his institute became a model for psychological institutes in many countries. But he did not found an actual school in the strict sense of the Berlin school of Gestalt psychology, the Würzburg school of cognitive psychology, or the Vienna school of psychoanalysis.

The fourth phase is the development of applied psychology, that is, through the inclusion and development of fields of praxis outside the universities and higher education after about 1890. Wundt had an ambivalent, rather negative relationship to the development of applied psychology. In his opinion, psychology should attempt a transition to practical applications only after its foundation has been systematically constructed (Wundt, 1913).

Fifth, with his twofold psychology and with numerous students from many countries, Wundt made a definitive contribution to the development of a pluralistic system of psychology and to the gradual integration of the major schools after about 1930.

The final phase is professionalization, that is, the establishment, social anchoring, and state certification of standards for academically educated psychologists active outside of the university system around the time of World War I. With his early work in developing a curriculum for the education and training of psychologists, Wundt contributed indirectly to professionalization. This curriculum, with further contributions from other psychologists, was adopted in 1941 in the form of the first state diploma examination regulations in Germany.

The genesis of the discipline described above was initially, in the nineteenth century, determined primarily by experimental psychology. This constituted the first phase of methodological development in which experimental methods, the technical apparatus and the mathematics of experimental physics and of experimental physiology were transferred to psychology. Wundt's 'individual psychology' is an expression of this development. It thus becomes understandable why many of the 'pioneers' of modern psychology, Wundt's contemporaries in nineteenth-century Germany, were trained as experimental physicists or experimental physiologists. They include Ernst Heinrich Weber (1795–1878), anatomist and physiologist, Gustav Theodor Fechner (1801–87), doctor and physicist, and Hermann von Helmholtz (1821–94), physiologist and physicist.

During the second half of the nineteenth century a second phase of methodological development began. More or less standardized nonexperimental methods were added step by step to the experimental procedures. These methods were initially analytical documentary procedures that Wundt called methods of 'comparison' within his 'ethnopsychology.' Later they also included the preexperimental empirical methods based on the paradigms of assessment, achievement, arrangement, and interpretation. These methods contributed significantly to the development of psychodiagnostic procedures. Both groups of nonexperimental procedures today constitute elements of 'historiographic' and 'quasi-experimental' methodology in psychology. Wundt's work and its influence may be categorized within these two phases of nineteenth-century methodological development: his experimental 'individual psychology' falls within the first phase and his nonexperimental 'ethnopsychology' within the second.

Review and Outlook – What was Transitory, What Endures

Four achievements of Wundt's life's work must be especially emphasized. First, he contributed, in the politics of science, to the genesis of the discipline of modern psychology, that is, to its independence as an empirical psychology with its own subject matter and methods. Second, he constructed a system of psychology as an experimental 'individual psychology' and as a humanistic 'ethnopsychology.' Third, he founded the world's first Institute for Experimental Psychology, which gave impetus to the institutionalization of psychology and fourth, he founded one of the first scientific journals for psychology, *Philosophical Studies*.

Critically it should be noted that Wundt's concept of the experimental inaccessibility of higher psychological processes played a limited role for a time in the discipline's development. However, even some of his contemporaries did not share this idea. They included Gustav Theodor Fechner (1801–87), Carl Stumpf (1848–1936), Hermann Ebbinghaus (1850–1909), Georg Elias Müller (1850–1934) and Oswald Külpe (1862–1915). Nor did many experimental psychologists of the next generation, such as Narziß Ach (1871–1946), Karl Bühler (1879–1963), Max Wertheimer (1880–1943) or Otto Selz (1881–1944), share Wundt's view.

Wundt's writings and activities attained worldwide influence, especially because of his many students. It was them, most of all, who spread the tradition of experimental 'individual psychology.' Granville Stanley Hall (1844–1924), who established the second psychological laboratory in the USA at Johns Hopkins University in Baltimore in 1883 and in 1887 founded the *American Journal of Psychology*, was a prime example. Another was James McKeen Cattell (1860–1944), Wundt's first assistant in Leipzig, who occupied the first chair in psychology in the USA at the University of Pennsylvania in 1887. He became known primarily as the cofounder of the field of psychological diagnostic testing. Edward Bradford Titchener (1867–1927) who left Oxford, England in 1890, spent two years in Leipzig and went on to promulgate Wundt's teachings as structuralism at Cornell University in the USA. Another student was Dimitrij Nikolaevich Uznadze (1887–1950), founder of the Georgian school of psychology and known for his theory of set. Notable among Wundt's German students were Emil Kraepelin (1856–1926), one of the founders of modern psychiatry, psychopharmacology, and experimental psychodiagnostics; Ernst Meumann (1862–1915), founder of experimental pedagogy; Theodor Lipps (1851–1914), an exponent of experimental esthetics; Hugo Münsterberg (1863–1916), founder of industrial psychology and cofounder of applied psychology ('psychotechnics'); and Karl Marbe (1869–1953), one of the founders of forensic psychology.

See also: Empiricism, History of; Ethnology; Hall, Granville Stanley (1844–1924); Kraepelin, Emil (1856–1926); Scientific Disciplines, History of.

Bibliography

Bringmann, W.G., Tweney, R.D. (Eds.), 1980. Wundt Studies. Hogrefe, Toronto, Canada.

Bringmann, W.G., Lück, H.E., Miller, R., Early, C.E. (Eds.), 1997. A Pictorial History of Psychology. Quintessence, Chicago, IL.

Eckardt, G., Sprung, L. (Eds.), 1983. Advances in Historiography of Psychology. VEB Deutscher Verlag der Wissenschaften, Berlin, Germany.

König, E., 1902. W. Wundt als Psycholog und als Philosoph [W. Wundt as Psychologist and Philosopher], second ed. Frommanns, Stuttgart, Germany.

Popplestone, J.A., White McPherson, M., 1999. An Illustrated History of American Psychology, second ed. The University of Akron Press, Akron, OH.

Sprung, L., Sprung, H., 1981. Wilhelm Maximilian Wundt—ancestor or model? Zeitschrift für Psychologie 189, 237–246.

Sprung, L., Sprung, H., 1999. Rückblicke auf ein schwieriges Jahrhundert—Zur Geschichte der Psychologie im 20. Jahrhundert in Deutschland. In: Hacker, W., Rinck, M. (Eds.), Zukunft gestalten. Bericht über den 41. Kongreß der Deutschen Gesellschaft für Psychologie in Dresden 1998. Pabst, Lengerich, Germany, pp. 123–143.

Viney, W., Brett King, D., 1998. A History of Psychology. Ideas and Context, second ed. Allyn and Bacon, Boston, MA.

Wundt, W.M., 1863/1990. Vorlesungen über die Menschen- und Thierseele [Lectures on the Mind of Humans and Animals] vol. 2. VEB Deutscher Verlag der Wissenschaften, Berlin, Germany.

Wundt, W.M., 1900–09. Völkerpsychologie [Ethnopsychology] vol. 10. Engelmann, Leipzig, Germany.

Wundt, W.M., 1906–8. Logik. Eine Untersuchung der Prinzipien der Erkenntnis und der Methoden wissenschaftlicher Forschung [Logic: An Investigation of the Principles of Knowledge and Methods of Scientific Research] vol. 3. Enke, Stuttgart, Germany.

Wundt, W.M., 1908–11. Grundzüge der Physiologischen Psychologie [Principles of Physiological Psychology], sixth ed. Engelmann, Leipzig, Germany.

Wundt, W.M., 1910–21. Kleine Schriften [Small Papers] vol. 3. Engelmann, Leipzig, Germany. pp. 515–43.

Wundt, W.M., 1913. Die Psychologie im Kampf ums Dasein [Psychology in the Struggle of Existence]. In: Wundt, W.M. (Ed.). 1921. Kleine Schriften, vol. 3. Kröner, Stuttgart, Germany.

Wundt, W.M., 1920. Erlebtes und Erkanntes [Lived Experience and Knowledge]. Kröner, Stuttgart, Germany.

X

Xenophobia and Nativism

Peter Hervik, Aalborg University, Aalborg, Denmark

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In everyday terms, 'xenophobia' is now used to refer to dislike of foreigners, such as being anti-immigration, antiforeign, and/or anti-different groups, while 'xenophobia' in the social sciences has imported semantic changes that reflect mega-events such as 9/11, the Danish Muhammad cartoon crisis of 2005–06, and the bombing and killings on 22 July 2011 in Norway. In this new development, xenophobia (including Islamophobia) and nativism embrace the idea of cultural incompatibility and the naturalization of xenophobic attitudes following from it. Out of this understanding of xenophobia as a natural reaction, antimigration springs as a nativist necessity with a right to defend 'one's culture' while either legitimizing racism or denying that cultural and political self-defense can be racist.

Introduction

In its broadest sense, xenophobia refers to 'fear of strangers' and is associated with sentiments of antimigration. In recent years, xenophobia has emerged as an exclusionary ideology, associated not least with indiscriminate anti-Islamic beliefs and practices. From being used predominantly in North America for the last couple of hundred years, the term 'nativism' now also appears in Europe and elsewhere. It refers broadly to a favoring of established inhabitants over newcomers that eventually leads to the marginalization and exclusion of the latter.

The meanings of these two terms overlap in their association with bigot and exclusionary ideologies, which operate (like all ideologies) with explicit constructions of positively represented and defended in-groups and attacks on negatively rendered out-groups. Conceptually they can be approached as varieties of cultural racism – along with anti-Semitism; anti-black, anti-Roma, and anti-Gypsy racism; and others – as they construct distinct cultural identities that are seen as inherited, rather than as acquired through naturalization.

The word xenophobia comes from the Greek 'xenos,' meaning both the stranger and the guest, and 'phobos,' meaning fear; therefore, the term is literally 'fear of stranger.' Anatole France may have been the first to use the term 'xenophobia' in Europe, in his novel *Monsieur Bergeret à Paris* (1901).

In many languages, xenophobia is used as the supposedly less morally reprehensible word for 'racism' (Stolcke, 1995), while the more correct meaning we can find in the German term *Fremdenfeindlichkeit*, or hostility toward strangers. In this manner, the meaning of xenophobia holds the double sense of 'fear of the stranger' and 'hatred of strangers.' Although the stranger whom one fears can be welcomed as a guest, the stranger whom one hates cannot.

The hated stranger may be constructed as opposed to one's own identity and therefore negated. This negation, in Balibar's (2005) words, stems from a collective self-understanding as the other's other: the collective self receives its identity from being the other of the external collective other. In Huntington's (1996: p. 21) thesis, the 'hated stranger' is constructed within the opposition of "Unless we hate what we are not, we cannot love what we are." This resembles the formulation by Carl Schmitt (1996 [1932]), a prominent German ideologue during the Nazi period and one of the founding father figures of neoconservatism: "We don't know who we are, if we don't know our enemies."

The distinction between the feared guest and hated enemy may help us to understand that the presence of 'phobia' in 'xenophobia' is almost exclusively metaphoric. The phobia part of xenophobia carries the risk of treating the phenomenon as an illness and a form of irrationality (Rattansi, 2005: p. 295), even when used as a metaphor. While the psychological point of view associates phobias with physiological symptoms such as uncontrollable anxiety, shortness of breath, and chest pain, in the anthropological perspective, xenophobia refers to social and cultural meanings that these phobias represent.

While nativism has been present in practice since the sixteenth century in North America, it is most strongly associated with the early nineteenth century (the era of immigration), when the first book on nativism, *Political Nativism in New York State*, appeared in 1901. Earlier, Samuel Morse, a founding figure of American nativism, published the two books *The Foreign Conspiracy against the Liberties of the United States* (1834) and *The Imminent Dangers to the Free Institutions of the United States through Immigration* (1835). The emergence of nativism by the beginning of the nineteenth century was shaped by decades of anti-Catholicism, the source of antebellum nativism, which was possibly nourished by claims that Catholics were primarily loyal to the pope, and therefore could not be loyal citizens in the new United States but were instead

coming to subvert American liberty and democracy (Billington, 1938). A few years later, the newly born nativist ideology yielded a physical manifestation in Philadelphia in 1844, when nativist riots broke out. In the same year, the American Republic Party was founded and soon included the nativist Know Nothing Party.

Historian John Higham (in Michaels, 1995: p. 2) defined nativism in the US context as "intense opposition to an internal minority on the grounds of its foreign (i.e., 'un-American') connections." In anthropology, Ralph Linton and Irving Hallowell (1943: pp. 230–240) discussed nativism as "any conscious, organized attempt on the part of a society's members to revive or perpetuate selected aspects of its culture." The conscious, organized attempt was understood to lift the effort above a 'normal process' of what could best be seen as ethnocentrism. The stress on consciousness also conveys awareness that "there are cultures other than its own and that the existence of its own culture is threatened" (Linton and Hallowell, 1943: pp. 230–240). Thus, the semantic scope of nativism ranges from an ethnocentric belief of a community (or part of a community) to an ideology of social and nationalist movements.

In sum, xenophobia and nativism are both exclusionist ideologies claiming the rights to defend one's cultural difference. Xenophobia's fear and hatred of the stranger embody a call for cultural self-defense. Similarly, the nativist reliance on the neoracist maxim that different cultures are of equal value but essentially incompatible with each other (Hervik, 2011) is one that constructs a "culturally defined in-group," explained by Walter Benn Michaels (1995: p. 80) as follows: "If to belong to a culture you must be able to encounter it in the form of memory, as a tradition that is already yours, then no one can ever cease to belong to his or her culture or begin to belong to some other culture." This is where the right to claim cultural rights is born.

Current Usages

The usage of 'xenophobia' and 'nativism' is widespread in everyday language and currently under heavy scrutiny in social science research. In everyday discourse, xenophobia is now used to refer to dislike of foreigners or to holding beliefs that are anti-immigration, antiforeign, and/or anti-different groups, while xenophobia in the social sciences has imported semantic changes that reflect recent mega-events and comprehensive changes.

From a social psychological perspective, xenophobia is approached as a person's uncontrollable fear of the difference of people, who are often associated with racial, cultural, or religious difference. This version of xenophobia highlights the anxiety disorder, the irrational and literally 'phobic' fear; in this case, the central feature of xenophobia is, like with all social phobias, an excessive fear of being observed or scrutinized by unfamiliar persons (Furmark, 2000). The social psychological approach has declined in importance next to the drastic growth, in the social sciences and humanities, of findings relating to terrorist attacks and the contestation of Muslim cultural integration. In these approaches, xenophobia and nativism are seen as yet more sociocultural constructions, where dislike (if not hatred) of others is learned through socialization process, especially by way of news media coverage and popular culture.

With events such as the World Trade Center bombing in 1993, the bombings of the American embassies in Kenya and Tanzania in 1998, the 9/11 attacks in the United States in 2001, the bombings in Madrid in 2003, the bombings in London in 2005, the Muhammad Cartoon Crisis beginning in Denmark in 2005–06, and the so-called 22/7 massacre in Norway in 2011, as well as a series of smaller scale events and subevents, it became clear that definitions of xenophobia and nativism had to move beyond reducing these acts of cultural racism to simply discriminatory attitudes or sentiments. Instead, these definitions must include the level of social action, or practice, where state-supported policies and social movements played a role. In the aftermath of these events and what they represent, the conceptual implications and the analysis of the events from anthropological and sociological perspectives present us with a huge and urgent challenge.

Anchoring definitions of cultural racism in its various forms can also be seen at the level of situated social interaction. An incident in Denmark is useful for illustration. Two black men were denied entrance to a nightclub there on the basis of their visible difference from other customers, an appearance that the security guards associated with sub-Saharan African men. However, the moment they realized that the guests were African-Americans, they were immediately welcomed. In this case, the rejection can hardly be reduced to xenophobia in the strict social psychological sense, with an emphasis on 'phobia.' An analysis of institutional support of acts of exclusion must be included: the nightclub's policies and its employees' actions, municipal and state policy on intervening or not intervening, the role of the police, and so on.

Another peril in the study of xenophobia (which can also be illustrated by the nightclub incident) is when an incident is being talked about both as xenophobic and as racist discrimination. We may note, for illustrative purposes, that the term is regularly used to describe intense dislike of foreigners or migrants as well as the racist act of rejection on the basis of one's skin color and association with the non-Western world. The term may appear as less morally reprehensible than racism, since everyday uses of 'race' and 'racism' are negatively loaded words in most countries, especially in the German-speaking countries of Central Europe. Here, the words *Rasse* and *Rassismus* are avoided on all levels. Instead, 'xenophobia' has been preferred to describe the same phenomena. Yet, the phenomena are largely the same, and the consequences for the discriminated are the same.

Anthropology has performed a tremendous amount of work on critiquing the essentialization of cultural differences, yet this effort has not substantially changed public opinion and public consciousness. While anthropologists continue to point out that the "world is much more complicated and complex than presented in the media" (and possibly by others) (see, e.g., Ferguson, 2006), cultural identity is still very frequently treated as fixed and irreducible, and biological or cultural. Most anthropologists today agree with John Comaroff (1996: p. 166) that there cannot be a universal theory of xenophobia, nationalism, nativism, Islamophobia, or ethnicity per se, "only a theory of history capable of

elucidating the empowered production of difference and identity." If social worlds are constructed differently, that is, if they are relative, they can still be compared through the appropriate canons of comparison while avoiding conflating cultural relativism with moral relativism. Nevertheless, the essentialization or absolutization of cultural differences remains ubiquitous.

Research on the media coverage of 22 July 2011 in Norway, where the ethnic Norwegian terrorist Anders Behring Breivik went on a killing spree at a summer camp of young political activists whom he judged as immigration friendly, resulting in 77 deaths, has drawn attention to another issue in the study of xenophobia and, particularly in this case, Islamophobia. The Breivik case showed the tendency to reduce acts of terror to individual pathologies. The pathologizing comes easily through the association of xenophobia with phobias, but there is nothing 'natural' about most xenophobia. Instead, it is 'naturalized.' Therefore, acts such as Breivik's should not be explained as the outcome of a lone wolf (Boisen and Hervik, 2013). Instead, lone-wolf claims need to be analyzed for their social function and as communicative acts that serve to deflect attention from ideologies that a person such as Breivik also shares, such as the conspiracy of 'Eurabia.' Reducing xenophobia, nativism, and racism to individual pathologies is taking these phenomena out of history, deflecting attention from immediate circumstances and structural conditions.

Unlike xenophobia, nativism has primarily been studied as a phenomenon in North America. Thus, the ideas that allowed for nativism to appear in this particular context are related to the history of a 'settler nation.' In this new nation, 'native-born Protestants' regarded newcomer Catholics as a threat to their 'original' values and 'society's norms' and therefore opposed immigration harshly. Hans-Georg Betz (2007) argues that the concept of nativism can be applied to a European context, particularly in relation to the anti-Islamic arguments recently used by radical and extreme-right parties and movements. At least two mutually reinforcing anti-Islamic arguments are set forth here: Islam is an alien religion, fundamentally incompatible with Western Europe's secular values and way of life; and, given its 'totalitarian' claims and aspirations, Islam is a fundamental threat to individual freedom and liberal democracy. As in nineteenth-century American nativism, a people or civilization is seen as "inherently different to the members of another civilization"; something within Islam and in the nature of Muslim people, this argument goes, makes them inherently more prone to pursue their objectives with violent means (Betz, 2007: p. 47).

Betz's introduction of the concept of nativism to the European context raises the issue of what units comprise the 'native' group at hand. We can conceptualize it as religious, national, ethnic, or racial. Or, since these are also political entities, what ideological group, or what belief system, do the nativists claim to be their out-group adversary? Are these units recognized as standing in a relationship of superiority and inferiority? What kinds of fissures can be discerned among them?

Fatima El-Tayeb (2011) has found, for instance, that in Europe a racialized understanding of proper Europeanness and Westerners in a nativist manner continues to exclude certain migrants and their descendants. Rather than restricting the term 'migrant' to that first generation who physically migrated across borders, the racialized European understanding represents the second and third generations as 'migrants' as well, although they are in effect minority citizens. These processes of externalization (or social exclusion), where second-, third-, and fourth-generation citizens are continuously racialized and ethnicized as foreign and inferior, now threaten to further polarize and radicalize society. Also, citizens who are seen as sympathizers of unwanted migrants and the 'annoying difference' they represent are also being placed in question.

The new perception of xenophobia or nativism is to a high degree associated with the emergence and drastic increase of Islamophobia, particularly in Europe but also in North America. A closer look at this development can help us further clarify the terms 'xenophobia' and 'nativism.' In attempts to reduce these to a core of negative attitudes or beliefs, it may be argued that following terrorist attacks, it is predictable if not obvious that Islamophobia would rise. This, however, does not happen mechanically. It is helped along by organized agents, who are willing to capture the situation and use the mediatization of society to produce moral panics and a politics of fear, in order to support their organizations and initiatives.

Islamophobia

In its own terms, Islamophobia is yet another concept that refers to 'xenophobic' reactions to the Islamic faith and to Muslims. The xenophobic association would yield a 'fear of Islam,' but it is in reality a xeno-hostility – a hostility toward Islam and Muslims. The notion of 'phobia' as a social phobia does not seem to be present. So, whereas xenophobia embraces both fear and hatred, Islamophobia is a more aggressive and predatory entity (Appadurai, 2006).

Initially, Islamophobia was defined as 'an unfounded hostility toward Islam, and therefore fear or dislike of all or most Muslims' (The Trust Commission on British Muslims and Islamophobia, 2004 (originally written in 1997)). The term gained its prevalence in a public discourse after the 9/11 attacks in 2001, but it also spread through the discussions of the UK Runnymede Trust Report on Islamophobia in 1997, and the Danish Muhammad cartoon conflict in 2005–06. However, there is no full agreement on what 'Islamophobia' refers to. During the Danish Muhammad cartoon crisis, after a Danish newspaper had published cartoons of the Prophet Muhammad, the Free Speech society and its organized anti-Islamic sympathizers grew strong, having both network relations and shared core values with the neoconservatives on the east coast of the United States (Hervik, 2008). In the wake of the Breivik murders in Norway, it has become clear that Islamophobia was organized in the so-called Counter-Jihad movement, which has likewise extended economically and ideologically to North America. That movement as well as the Free Speech Society subscribes to the notion of the Eurabia conspiracy, which was first formulated by author Oriana Fallaci and the historian Bat Ye'or in 2004 and 2005 (see Strømmen and Indregard, 2012: p. 23). According to this conspiracy, Europe is being taken over by Muslims in collaboration with their European multiculturalist and left-wing intellectual collaborators. These Muslims will introduce shari'a law and turn Europe into a caliphate. Immigration, collaboration, and

high rates of fertility are the means to do this (Bangstad, 2013). Accordingly, the understanding of the conspiracy is constructed both as a populist criticism of the elite and as a racialization of Muslims. Conspiracy theorizing, however, can also have effects beyond the ideological content itself and the communication about elite betrayals: it serves as an operational spur to urgent, extraordinary, or violent action (van Buuren, 2013).

The term 'Islamophobia' itself is rather misleading: it literally refers to the fear of a religion, while in fact it is a fear of Muslims and a hostility legitimized because they adhere to Islam. The term indicates religious discrimination, but intolerance toward racial, class, and cultural identity is also closely related to it. In this way, Islamophobia does not contain one simple meaning but, rather, encompasses different discourses that define the sociopolitical situation of Muslim minorities in Europe (Cesari et al., 2006: p. 8). Christopher Allen (2010: p. 195) describes Islamophobia not as a 'phobia' as such, but rather as a form of governmentality, or an ideology that reduces the meaning of Islam and Muslims to an anti-other jeopardizing neoliberal Western values.

Matti Bunzl has argued that Islamophobia and anti-Semitism hold two separate historic origins and are therefore two phenomena that have little in common. Anti-Semitism arose strongly in the nineteenth century with the development of the idea of "race, which gave the Jews an immutable biological destiny. Jews could never belong to the national German community, no matter their strivings for cultural assimilation" (Bunzl, 2005: p. 502). But critics see it differently, and obviously both anti-Semitism and anti-Islamic racism can be found earlier than Bunzl had found it (Smith, 1996; Gardell, 2011). Gingrich (2005) directs our attention to Edward Said's work on Orientalism to show that Islamophobia was present much earlier than the recent emergence suggested in Bunzl's claim. He also reminds us about Frontier Orientalism, which considered Austria one of the historical conflict zones (besides Spain, Italy, Russia, and the Balkan countries). This is not missed by the extremist-right anti-Islamic networks in Europe, particularly the 'Gates of Vienna' blog, to which terrorist Anders Behring Breivik was much attracted, as well as the blog of Peder Are Nøstvold Jensen (aka Fjordman), Breivik's biggest single inspiration. The 'Gates of Vienna' blog bears the slogan: "At the siege of Vienna in 1683 Islam seemed poised to overrun Christian Europe. We are in a new phase of a very old war."

In response to Bunzl, John Bowen (2005) insists that Islamophobia has been well known in France since the Algerian War (1954–62) and deeply rooted in Algerian colonization (from 1830 onward); it persists in reactions to continuing violence in North Africa, Lebanon, and elsewhere. These attitudes, he argues, involve a mix of condescension toward Islam, postwar hatreds, and ordinary racism.

Against defining Islamophobia merely as an attitude, as "indiscriminate negative attitudes or emotions directed at Islam or Muslims" (Bleich, 2011: p. 1581), Brodkin (2005), in her reaction to Bunzl's argument, insists that Islamophobia is seen as state supported and also supported by social movements (2005). Nina Glick Schiller (2005) also criticizes Bunzl for his thin material and for neglecting the doctrine of War on Terror announced by President G.W. Bush following the 9/11 attacks. Her argument is that Islamophobia is not, as Bunzl claims, replacing ideas of race but is a variant of racialized thinking. The very rhetoric and the attitudes she encounters in Europe are based on ideas of immutable cultural difference that now use 'culture' for 'race' only as old wine in new bottles.

Analysts of Islamophobia have attempted to distinguish between hostility to Islam and to Muslims. Whereas the past may have seen Islam as the faith of the enemy, today the target is different: "The attack is now against not Islam as a faith but Muslims as a people" (Halliday, 1999: p. 898). But the distinction between hatred of Islam and hatred of Muslims is hard to sustain. Intolerance of Muslim prayers, clothes, eating habits, living arrangements, and so on, often supported by state policies, reveals an attack on cultural identity. In this perspective, Islamophobia does not necessarily indicate a secular form of discrimination but, rather, a social–political discourse of anti-Muslim sentiments and intolerance (Erdenir, 2010: p. 29). Since 9/11, Muslims have increasingly become stigmatized, a metonym of all kinds of troubles in Western societies. Not only are they seen as a financial burden and blamed for crime, violence, and fundamentalism, but also they embody a challenge to the neoliberal state and the values associated with it. As Evans (2011: pp. 1752–1753) puts it: "What makes Islam exceptional in the eyes of the neoliberal order is its vast membership, its potential as a global mass movements, and its distinctive and complex tradition of religious belief, political ideology, social norms, world outlook, and philosophical method." Islamophobia is a xenophobic reaction, but it is also much more. The ideology of a Eurabian threat and the compendium by the terrorist Breivik are two of the sources among many that show what John Esposito (2011: p. xxvi) has called "the nexus of Islamophobia, multiculturalism, and Muslim-West relations," with nationalism and Eurocentrism as the two nativist nodes. Islamophobia does not necessarily indicate religious ideology and anti-Muslim discourse; rather, it is based on cultural incompatibility. Muslim cultural identity ends up being the enemy, or rather one of the enemies. This notion of cultural incompatibility alludes to the 'new racism' ideology, which will be discussed in section Cultural Incompatibility – The New Racism Discourse and includes confrontational stances and attacks on multiculturalists and others.

Cultural Incompatibility – The New Racism Discourse

Like all emotions, xenophobic feelings must be studied in context and therefore also be seen through a constructionist lens that includes the social evaluation of the rights to a given feeling (Lutz and Abu-Lughod, 1990). In recent years, the growth of anti-Islam sentiments has yielded a claim that xenophobia, besides being chosen instead of the 'reprehensible' racism, is increasingly regarded as justifiable and legitimate. In the aftermath of the Danish Muhammad conflict, interviewees in Hervik's (2011) study agreed that xenophobia has become legitimate in Danish society, due to the association of Islam with terrorism in Danish news media coverage; a large section of Danish society now considered Muslim values and Danish values as incompatible. This tendency to naturalize xenophobia can be seen as a consequence of the high percentage of Europeans (ranging

from 40 to 60%), who see greater interaction with Muslims as a threat (Betz and Meret, 2009).

The idea that xenophobic attitudes and reactions are 'natural' is one of the key features of the neoracism (cultural racism and cultural fundamentalism) that arose in the United Kingdom (Barker, 1981) and the Nouveau Droit that arose in France (Taguieff, 1988) in the late 1970s and early 1980s. In a new rhetorical shift, the political right now acknowledges the equality and equal rights of people of different cultures, but at the same time it emphasizes the 'natural' belonging of people to certain places in the world. Liisa Malkki (1992: p. 27) has found such naturalization in people's unreflective use of plant metaphors and kinship talk when they discuss identity and belonging: "People often thought of themselves as being rooted in place and as deriving their identity from that rootedness." This supports the 'national order of things' as the 'natural order of things.' The theory of human nature embedded in neoracism, then, is that people can be found in the 'wrong' places, where they do not belong 'naturally' (as a consequence of being 'uprooted'); therefore, deriving from this notion, xenophobic reactions will be the expected and natural consequence of human nature (Barker, 1981; Stolcke, 1995). Out of this understanding of xenophobia as a natural reaction, antimigration springs as a nativist necessity and a right to defend 'one's culture,' while either legitimizing racism or denying that cultural and political self-defense can be racist. In the Nordic context, in the last 15 years, people have increasingly believed that some differences were so great that they could not be overcome. The widely different newcomers could never be integrated and should therefore be 'sent home' (Hervik, 2011).

New racism in Britain has drawn its strength from the defense of a mythic British or English way of life in the light of political and social crises, and what could also be seen as nativism, since it promotes an in-group of established citizens. An in-group or 'we-group' promotes itself through its opposite, which it negates. This group does its identity work by opposing the 'cultural other,' or 'itself as the other's other,' which, as I argued, lies behind Huntington's (1996) narrative of 'clashes of civilizations.' The attention to religious (and at the same time cultural) artifacts such as the Muslim headscarf, or to physical characteristics, becomes an indicator of cultural racism. The 'other' is culturally incompatible.

Bunzl (2005) has raised important questions about anti-Semitism and Islamophobia in both a diachronic and a synchronic comparative perspective. His focus and material are not so broad, and his research was conducted before such mega-events as the Muhammad cartoon conflict and 22/7 in Norway. For his part, Bleich (2011), a political scientist, comes out strongly calling for tight definitions, so that he can do comparative analysis and work with causes. However, Barth (1999) has convincingly argued that comparison has come to deal less with actual phenomena and more with definitions, conceptualizations, and texts. This is the shortcoming of Bleich as well, who does not include the level of practice – with all its complexities – or the integral support of institutions.

What kind of revisions would Bleich, Bunzl, and the commentators have made if they had been able to include what happened in Norway on 22 July 2011 and the Muhammad cartoon crisis? Research into these events has revealed that Islamophobia and nativism are not the only things that were at stake. Feminists, multiculturalists, and cultural Marxists are equally as much the objects of hatred as the Muslims. The identification of some of these other enemies comes out loud and clear in the 'cultural war of values' strategy in Danish politics (Hervik, forthcoming) and in radical right-wing political programs. They also combine with a populist discourse, where rhetoric of war serves to naturalize violence and the slip from politics into war, such as war and terror and defending Europe against the emergence of Eurabia by any means. The terrorist Breivik killed young multiculturalists and cosmopolitans at their summer camp in Norway, but the collective categories were dealt with ideologically in writing long before that. They were categories selected as adversaries in the 'war,' and they were 'elected' or 'selected' for elimination (Balibar, 2005).

In the systems of belief underlying the actions of Breivik and the network of extremists and radicals who share most of his values, xenophobia, nativism, and Islamophobia combine with violent and/or ideological attacks of adversaries. In future research, much thicker contextualization is needed in anthropological research on how these manifestations of Islamophobia relate to anti-Semitism (not least among people who are not Muslims or Jewish), and how Islamophobia in people's reasoning relates to antimulticulturalism, antifeminism, left wingers, and labelings of the 'political correct.' We need to theorize the intrinsic relations of combined forms of exclusionary ideologies, discriminatory practices, and identification patterns.

See also: Afrocentric Approaches to Social Work; Cosmopolitanism; Cultural Relativism, Anthropology of; Ethnicity: Anthropological Aspects; Ethnocentrism; Intercultural Communication: An Anthropological Perspective; Multiculturalism; Race and Racism in the Twenty-First Century.

Bibliography

Abu-Lughod, Lila, Lutz, Catherine A., 1990. Introduction: emotion, discourse, and the politics of everyday life. In: Lutz, Catherine A., Abu-Lughod, Lila (Eds.), Language and the Politics of Emotion. Cambridge University Press, Cambridge.

Allen, Christopher, 2010. Islamophobia. Ashgate, London.

Appadurai, Arjun, 2006. Fear of Small Numbers: An Essay on the Geography of Anger. Duke University Press, Durham, NC.

Balibar, Etienne, 2005. Difference, otherness, exclusion. Parallax 11 (1), 19–34.

Bangstad, Sindre, 2013. Eurabia comes to Norway. Islam and Christian Muslim Relations 24 (3), 369–393.

Barker, Martin, 1981. The New Racism: Conservatives and the Ideology of the Tribe. Junction Books, London.

Barth, Fredrik, 1999. Comparative methodologies in the analysis of anthropological data. In: Bowen, John, Petersen, Roger (Eds.), Critical Comparisons in Politics and Culture. Cambridge University Press, Cambridge.

Betz, Hans-Georg, 2007. Against the 'green totalitarianism': anti-Islamic nativism in contemporary radical right-wing populism in western Europe. In: Liang, Christina S. (Eds.), Europe for the Europeans: The Foreign and Security Policy of the Populist Radical Right. Ashgate, London.

Betz, Hans-Georg, Meret, Susi, 2009. Revisiting Lepanto: the political mobilization against Islam in contemporary Western Europe. Patterns of Prejudice 43, 3–4.

Billington, Ray A., 1938. The Protestant Crusade, 1800–1860: A Study of the Origins of American Nativism. Quadrangle Books, Chicago.

Bleich, Erik, 2011. What is Islamophobia and how much is there? Theorizing and measuring an emerging comparative concept. American Behavioral Scientist 55, 1581–1600.

Boisen, Sophie, Hervik, Peter, December 2013. Danish media coverage of 22/7. Nordic Journal of Migration Research 3 (4), 197–204. Theme issue, guest editors Susi Meret and Peter Hervik.
Bowen, John, 2005. Commentary on Bunzl. American Ethnologist 32, 524–525.
Brodkin, Karen, 2005. Xenophobia, the state, and capitalism. American Ethnologist 32, 519–520.
Bunzl, Matti, 2005. Between anti-Semitism and Islamophobia: some thoughts on the new Europe. American Ethnologist 32, 499–508.
Cesari, Jocelyne, et al., 2006. Securitization and Religious Divides in Europe: Muslims in Western Europe after 9/11. Report Submitted to the Changing Landscape of Citizenship and Security 6th PCRD of European Commission, Paris.
Comaroff, John L., 1996. Ethnicity, nationalism, and the politics of difference in an age of revolution. In: Wilmsen, Edwin N., McAllister, Patrick (Eds.), The Politics of Difference: Ethnic Premises in a World of Power. University of Chicago Press, Chicago.
El-Tayeb, Fatima, 2011. European Others: Queering Ethnicity in Postnational Europe. University of Minnesota Press, Minneapolis.
Erdenir, Burak, 2010. Islamophobia qua racial discrimination. In: Triandyfillydou, Anna. (Ed.), Muslims in 21st Century Europe Structural and Cultural Perspectives. Routledge, London.
Esposito, John, 2011. Introduction. In: Esposito, John, Kalm, Ibrahim (Eds.), Islamophobia: The Challenge of Pluralism in the 21st Century. Oxford University Press, Oxford.
Evans, Tony, 2011. The limits of tolerance: Islam as counter-hegemony? Review of International Studies 37, 1751–1773.
Ferguson, James, 2006. Global Shadows: Africa in the Neoliberal World Order. Duke University Press, Durham, NC.
Furmark, Tomas, 2000. Social Phobia. From Epidemiology to Brain Function. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 95, Uppsala.
Gardell, Mattias, 2011. Islamofobi. Leopard Förlag, Stockholm.
Gingrich, Andre, 2005. Anthropological analyses of Islamophobia and anti-Semitism in Europe. American Ethnologist 32, 513–515.
Halliday, Fred, 1999. 'Islamophobia' reconsidered. Ethnic and Racial Studies 22, 892–902.
Hervik, Peter, 2008. The original spin and its side effects: freedom speech as Danish news management. In: Eide, Elizabeth, Kunelius, Risto, Phillips, Angela (Eds.), Transnational Media Events. The Mohammed Cartoons and the Imagined Clash of Civilizations. Nordicom, Gothenburg.
Hervik, Peter, 2011. The Annoying Difference. The Emergence of Danish Neo-nationalism, Neoracism, and Populism in the Post-1989 World. Berghahn Books, New York and Oxford.
Hervik, Peter (forthcoming). Cultural war of values: The proliferation of moral identities in the Danish public sphere. In: Tripathy, Jyotirmaya (Eds.), Becoming Minority: How Discourses and Policies Produce Minorities in Europe and India. Sage Publications, New Delhi, India.
Huntington, Samuel, 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.
Linton, Ralph, Hallowell, Irving, 1943. Nativistic movements. American Anthropologist 45, 230–240.
Malkki, Liisa, 1992. National geographic: the rooting of peoples and the territorialization of national identity among scholars and refugees. Cultural Anthropology 7, 24–44.
Michaels, Walter B., 1995. Our America: Nativism, Modernism, and Pluralism. Duke University Press, Durham, NC.
Rattansi, Ali, 2005. The uses of racialization: the time-spaces and subject-objects of the raced body. In: Mirji, Karim, Solomos, John (Eds.), Racialization. Studies in Theory and Practice. Oxford University Press, Oxford.
Schiller, Glick N., 2005. Racialized nations, evangelizing Christianity, police states, and imperial power: missing in action in Bunzl's new Europe. American Ethnologist 32, 526–532.
Schmitt, Carl, 1996 (1932). The Concept of the Political. University of Chicago Press, Chicago (original work published 1932).
Smith, David N., 1996. The social construction of enemies: Jews and the representation of evil. Sociological Theory 14, 203–240.
Stolcke, Verena, 1995. Talking culture: new boundaries, new rhetorics of exclusion in Europe. Current Anthropology 36, 1–24.
Strømmen, Øyvind, Indregaard, Sigve, 2012. Den nye høyreekstremismen. In: Indregaard, Sigve (Ed.), Motgift. Akademisk respons på den nye høyreekstremismen. Flamme forlag, Oslo.
Taguieff, Pierre-André, 1988. La Force du Préjugé. La Découverte, Paris.
The Runnymede Trust Commission on British Muslims and Islamophobia, 2004. Islamophobia: Issues, Challenges and Action: A Report. Trentham Books, London.
van Buuren, Jelle, December 2013. Spur to violence? Anders Behring Breivik and the Eurabia conspiracy. Nordic Journal of Migration Research 3 (4). Theme issue.

Relevant Website

http://gatesofvienna.net – Gates of Vienna (accessed 13.01.05).

Xenophobia: Social Psychological Aspects

Margarita Sanchez-Mazas, Université de Genève, Genève, Switzerland
Laurent Licata, Université Libre de Bruxelles (ULB), Bruxelles, Belgium

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Xenophobia refers to a fear of the stranger that has taken on diverse forms throughout history and is conceptualized according to different theoretical approaches. Often confounded with racism in modern times, the concept of xenophobia is rooted in universal fears of the unfamiliar, but also in specific historical traditions and regulations regarding the place devoted to outsiders within host societies. With the rise of the nation-state and the development of migration, xenophobia depicts immigrants, asylum seekers, or populations of immigrant descent as posing a material or identity threat to nationals.

Very often the word xenophobia is simply used as a synonym of prejudice – with no clear distinction with the term of racism. Racism is a form of prejudice toward individuals or groups based on the socially constructed notion of 'race', which defines social categories on the basis of differentiating phenotypical markers such as skin color. In contrast, coming from the Greek words *xenos* = stranger and *phobos* = fear, xenophobia denotes a hostility toward the stranger perceived as a threat. However, the notion of stranger is ambiguous, ranging from total otherness to the foreigner next door.

The Greek root *xeno*s may evoke the stranger as a general category, referring to the uncommon or the unfamiliar, as well as a political dimension that, rather than a universal psychological reaction to extraneity, is primarily concerned with rules and regulations that codify access to a territory and demarcate the civic body from the outsiders. Therefore, xenophobia may be conceptualized at different levels of analysis, depending both on the targeted stranger and the theoretical frame of reference. Nevertheless, the core notion of threat seems to be consistently associated with the phenomenon and is present in most theoretical approaches of its social psychological roots. Indeed, reactions to the stranger have been addressed both on their positive side as welcoming behaviors toward the outsider as a guest or partner for exchange, and on the negative side as an aversion toward the outsider as a menace. However, the fundamental ambivalence attached to the figure of the stranger (Simmel, 1908/1984) fades away in the case of xenophobia. This is unambiguously a response to the alleged dangers a given society faces when confronted with someone who is not 'one of us'.

Hence, xenophobia is not simply an attempt to demarcate oneself from the other. Underlying the phenomenon, particular classifications allow to distinguish members of the community from all those who belong to the outside world (as were the barbarians for the Greeks or in modern times citizens of other nations) or from those who become, temporarily or permanently, and under a variety of forms, 'strangers from the inside' (as were the Wogs for the Greeks or are the contemporary immigrants). These 'foreigners living among us' can be considered as the typical targets of xenophobia. The relationship with them may vary considerably, between acceptance and rejection. Xenophobia is an explicit hostility against those who are considered as intruders, usurpers, or vectors of disease.

As a social psychological phenomenon, this hostility has often been understood either as a decontextualized fear of otherness, or as a form of prejudice against an out-group somewhat confounded with racism. Yet, the more recent forms of enmity against immigrants, which have developed under the very name of xenophobia within national contexts recruiting 'guest workers', have received less research attention. This may be due to the focus of Western social psychology on the study of universal psychological processes. However, these can hardly account for the genuine xenophobic reactions that arise when members of a receiving society are entitled with formal rights and prerogatives that are denied to the members of an incoming group. Recent research has addressed anti-immigrant attitudes in a variety of ways, reflecting the increasing complexity of immigration issues that are concerned with a larger array of targets in the current context of globalization. Within this turn, it becomes difficult to disentangle xenophobia, as a fear of the stranger, from a large variety of responses to the presence and arrival of people whose presumed difference from national majorities evokes material as well as value, cultural and symbolic threats.

The Fear of the Other

Throughout history, outsiders have been associated with risks. Unclean foods, unusual hygienic customs, contagion, or epidemic spread stand as potential dangers motivating avoidance of intergroup encounters. The Jewish people, for instance, have experienced persecution on the basis of their alleged propagation of disease from the Middle Ages, where they were blamed for poisoning the water, to the contemporary era in which anti-Semitic propaganda has likened them to nonhuman carriers of germs and infections. A number of models link the recurrent and widespread disease-related reactions to outsiders to adaptive behaviors bearing safeguard and protection for the group against potential harmful contacts. The evolutionary approach attempts to explain how contemporary psychological experiences, and in particular xenophobic attitudes, may reflect ancestral adaptive mechanisms eliciting affective and behavioral reactions aimed at keeping the stranger at a distance (Faulkner et al., 2004). While evolutionary models identify perceived risk of becoming ill as

 http://dx.doi.org/10.1016/B978-0-08-097086-8.24031-2

a central feature, another line of thought draws upon psychodynamic theory for understanding blaming responses as a means to protect the identity of the self by denying one's own vulnerability to illness and linking health and safety risks to the 'other' (Joffe, 1999).

Either at a collective or at an individual level, several social psychological models have addressed fundamental processes underlying perceived threats – whose fear of disease is but one significant case in point – and eliciting anxiety, stigmatization, and prejudice (Stangor and Crandall, 2000). Yet, along these lines, research has stressed a universal and decontextualized version of the stranger, equating it with the transhistorical figure of the Other. Moreover, recent accounts from genetic theory derive in-group favoritism and out-group derogation from the operation of genes and hormones (Dawkins, 2006).

While these borrowings from other scientific domains, such as biology, may contribute to an understanding of an indiscriminate and panhuman hatred of strangers, it fails to grasp the peculiar social, political, and symbolic underpinnings of contemporary xenophobia. These are more accurately accounted for by classic social psychological models, even though, most often, the phenomenon remains confounded with neighboring notions such as prejudice, discrimination, or outgroup derogation.

The Societal Context of Xenophobia

Xenophobia should be situated in its wider societal context. According to Scheffer (2011), the cycle of interethnic relations following immigration generally comprises three steps: avoidance, conflict, and accommodation. "On arrival, migrants tend to keep themselves to themselves, partly as a result of the attitude of avoidance they detect in the society around them. In the years that follow, migrants and their children struggle to claim a place for themselves in the new country, and this leads to rivalry and strife. The question of how everyone can live together becomes unavoidable. If a satisfactory answer is found, the descendants of the original migrants will be absorbed more or less smoothly into society" (Scheffer, 2011: p. 6). Xenophobia can be seen as a symptom of the avoidance and/or conflict steps, and interpreted as reactions of host community members facing societal changes often experienced as threatening. Migrants often experience this perception of threat as well, which may lead some of them to withdraw in their own community. This withdrawal then contributes to strengthen host community members' mistrust.

With the transformations of migration flows, the arrival of immigrants and asylum seekers from more distant non-Western regions, but also with the settling down of reunited families and the increase in the numbers of second or third generation residents born or grown up in their families' host country, discourses about immigrants tend to include a series of themes related to values, culture, or habits. Migration is not only more diverse in terms of cultural origins, but groups of migrants tend to assert their cultural identity through struggles for recognition involving ostensible signs or demands for cultural or religious arrangements (Sanchez-Mazas, 2004). Furthermore, former societal divisions along political lines during the cold war have given way to a framing of antagonisms in cultural terms, for example through the notion of 'clash of civilizations' (Huntington, 1996). The turning of xenophobia into more radicalized forms of hostility toward the culturally different foreigners jeopardizes not only harmonious multicultural relations within nation-states, but also the construction of a cosmopolitan worldwide society.

However, these themes do not replace, but rather merge with more traditional arguments evolving around the actual or alleged foreign status of the immigrant offspring. This combination of delegitimation registers has important implications in the recent evolution of xenophobia toward more complex and combined forms of anti-immigrant attitudes. The fact that the populations of immigrant descent are still viewed as foreign regardless of their formal citizenship allows a wider range of anti-immigrant prejudices compared to prejudice against indigenous ethnic minorities. Thus, in France, access to formal citizenship has not prevented relegation of 'youngsters from the suburbs' to a second-class citizenship, hence feeding among them feelings of denial of recognition, sometimes giving rise to violent outbursts. In addition, anti-immigration discourses include both blatant and subtle expressions of prejudice (Pettigrew and Meertens, 1995), the latter being often associated with apparently rational arguments about 'tolerance threshold', scarcity of jobs, or welfare abuse that legitimize the expression of 'politically incorrect' opinions against them. Hence, the stranger, now reconstructed as an Other reluctant to integrate, and therefore as a menace to Western culture, is the target of rhetorical positions against 'foreign overpopulation', 'invasion', or 'flood' (i.e., the expression 'migration tsunami' used by French far-right leader Jean-Marie Le Pen).

Classical xenophobic themes involving notions of foreign invasion and overgrowth have gained momentum during the last decades, as asylum policy has been placed on the political agendas of Western nations. Official discourses concerned with national sovereignty, State control over borders, and antiterrorist policy after 11 September 2001 have targeted immigrants and asylum seekers, in such a way that threats posed by incoming refugees and other immigrants were associated with fears regarding domestic security. Concerns with the volume of asylum seekers have been addressed through expeditious processing of applications, tight restrictions upon the right to work, removal of support, or detention for failed asylum seekers. The representation of the refugees as claiming for asylum on a fraudulent basis, searching for illegal work, or being engaged into criminal activities such as drug dealing and prostitution, represents a renewed vision of the undeserving and unwanted foreigner (Sayad, 1991). In addition, the liberalization of Western economies has produced a category of precarious illegal workers who can easily be marked out as stealing jobs or lowering wage rates. Hence, nowadays, rejected asylum seekers are the preferred target of xenophobic discourses. These discourses lie at the core of right-wing populist politicians' strategies for conquering power.

Realistic Threats

Several major social psychological theories addressing the psychological roots of prejudice propose key concepts that

might be highly relevant for the study of xenophobia. Thus, a structural approach conceptualizes intergroup prejudice as reflecting the modalities of intergroup relations that can be framed in terms of competition over scarce resources or asymmetrical economic power. This *Realistic Conflict Theory* (Sherif, 1967) assumes that intergroup conflict, but also intergroup cooperation, stem from material interests that provide the infrastructure of reciprocal representations. However, images not only evolve according to the transformations of the nature of interactions, they also actively contribute to the unfolding and the justification of these interactions. The xenophobic rhetoric depicting immigrants as taking advantage of jobs, housing, or welfare opportunities originates from the perceived realistic threat foreign workers or foreign unemployed pose to nationals. Empirical research dealing with xenophobia against 'guest workers' in Switzerland showed that negative attitudes toward foreign people went hand in hand with the perception of intergroup relations in terms of negative interdependence between groups. That is, these interactions were perceived as a competitive relationship between Swiss nationals and foreigners, in a zero-sum game where 'what the other group gains, my group loses' (Mugny et al., 1991). This competitive zero-sum belief has also been identified as a mediator in the relationship between social dominance orientation – the individual preference for hierarchy within social systems – and negative attitudes toward immigrants (Esses et al., 1998).

Concern over material threats may be a relevant criterion for differentiating xenophobia from other types of prejudice primarily based on feelings of in-group superiority. The impact of alleged economic menaces on attitudes toward immigrants has been found both in Europe (Pereira et al., 2010) and North America (Esses et al., 1998). Indeed, arguments blaming foreign populations for benefitting from undue advantages at the expense of nationals are pervasive, particularly in those countries where immigrants, deprived of citizen rights, nonetheless benefit from the welfare system. Thus, people who see themselves and their national group as relatively disadvantaged in comparison with immigrants experience resentful feelings. This kind of reaction, typical of xenophobia, perfectly fits the assumptions of *Relative Deprivation Theory*. Because, unlike *Realistic Conflict Theory*, it emphasizes the subjective evaluation of this disadvantaged comparison (Pettigrew et al., 2008), this perspective helps explain the prevalence of xenophobia under different objective circumstances as well as the sensitivity of public opinion to anti-immigration stances about the foreign 'profiteer'.

Identity Threats

Social Identity Theory complements the *Realistic Conflict Theory* by adding a symbolic dimension tapping people's identities and belief systems (Tajfel and Turner, 1986). It is based on the notion that social identities are part of the self-concept, and that people are motivated to derive a clear and positive image of their social identities through comparisons with other social categories. The explanation of prejudice, including prejudice against immigrants, is often derived from *Social Identity Theory* predictions (Esses et al., 2005). Symbolic motives underlie intergroup attitudes and behaviors and account for the tendency to favor one's own group because it serves and secures its members' sense of identity. Thus, this model clarifies some observed discrepancies in receiving societies between interpersonal relations with out-group members and political stances or voting behavior. Hence, people may support xenophobic political parties while at the same time describe friendly relationships with immigrants. Social identity theory may explain this gap according to the notion that attitudes and behaviors held at an intergroup level are driven by social identity motives – i.e., maintain or acquire a positive social identity as a group member – that differ from interindividual motives, such as personal attraction.

The combined effects of perceived material and symbolic threats are very much plausible in the development of xenophobic attitudes. But perception that their national, subnational, or supranational identity is threatened by divergent social practices, values, or cultural arrangements introduced by immigrants may be sufficient to trigger prejudice against them among majority-group members. Muslims, in particular, are portrayed as posing such a value threat, as shown by examples of xenophobic political parties condemning the use of the Muslim scarf, the presence of minarets, or the selling of halal meat. Indeed, Velasco Gonzalez et al. (2008) found that prejudice against Muslim immigrants among Dutch adolescents was predicted by symbolic threats and negative stereotypes, but not realistic threats.

Perception of the values of immigrants as being incompatible with those of the receiving society implies the vision of nationals as a homogeneous group sharing common values (Esses et al., 2005). The view that immigrants undermine national values suggests that xenophobia correlates with nationalism. However, nationalism has proved to be a determinant of anti-immigration attitudes only to the extent that national identification is framed in ethnic – rather than civic or political – terms that is, with an emphasis put on ancestral values, shared culture, and blood ties (Pehrson et al., 2009). Since nations vary in the importance they attribute to different dimensions of national identity and citizenship (e.g., the German vs the French models and the corresponding jus sanguinis and jus solis), different approaches to the presence of people of foreign origin are to be distinguished, ranging from cultural assimilation – imposing strong cultural obligations – to multiculturalism – promoting culture maintenance among immigrant minorities. Besides, identification at other levels than the nation could also inspire xenophobia. For example, a study conducted in Belgium showed that identification with Europe, rather than national identification, positively correlated with a tendency to wish there were fewer foreigners in Belgium (Licata et al., 2011).

Xenophobia and Attitudes toward Immigration Issues

Xenophobia can be expressed through informal prejudice against individuals in day-to-day interactions. However, it can also inspire political position-taking regarding immigration issues and policies (Licata et al., 2011). Citizens of receiving countries may adopt different attitudes toward entry criteria set for immigrants, which define who can enter and live in the

in-group territory. Entry criteria can either be based on category membership (ascribed criteria), such as nationality or religion, or on individual merit (acquired criteria), such as education or professional skills. Green (2009) showed, across 20 European countries, that citizens of countries with a poor economic situation (i.e., experiencing realistic threat) expressed stronger support for the use of ascribed immigration criteria than citizens of wealthier countries. She also showed that individual perceived threat (realistic and symbolic) was more predictive of support for acquired criteria in wealthy than in poor countries. This suggests that xenophobic attitudes toward foreigners are mainly based on a social category distinction inspired by economic threat in poorer countries, whereas they tend to be based on a meritocratic ideology and pressure to normative conformity on individual immigrants in wealthier nations.

The way different categories of immigrants are named and represented in public discourse and the media, particularly on the moral dimension, heavily weighs on xenophobic attitudes. Esses et al. (2008) showed that negative attitudes toward refugees and the Canadian refugee policy were influenced by representations of refugees as immoral (not respecting the asylum procedure, cheating, etc.). As a case in point, the expression 'illegal immigrants' or 'undocumented' immigrants have different connotations. 'Illegal' contains an accusation of criminality and, through internal attributions, portrays the immigrant as active and guilty. In contrast, 'undocumented' refers to bureaucratic procedures. Suggesting external attributions, it presents the immigrant as a passive victim of the system. Not surprisingly, in the USA, different segments of the population tend to use the two expressions: 'illegal' is often used in legislative debates, whereas 'undocumented' is often used by immigrant rights groups (Deaux, 2008).

Of particular relevance is the framing of immigration as a personal choice or as a constraint, which affects attitudes toward integration policies. Verkuyten (2006) showed that presenting immigration as a personal choice, rather than as 'lack of choice', led to more negative attitudes toward multiculturalism in the Netherlands. This held true both for work-related immigrants (Moroccans or Turks) and for refugees and asylum seekers.

Research conducted in Germany also showed that negative attitudes toward immigrants – both blatant and subtle forms of prejudice – predicted the endorsement of ideologies of assimilation or segregation (versus integration) of immigrants (Zick et al., 2001). Conversely, Ward and Masgoret (2008) showed in New Zealand that both endorsing a multicultural ideology and having contacts with immigrants decreased perception of threat. A lower perceived threat was then associated with more positive attitudes toward immigrants, which, in turn, predicted more endorsement of positive immigration policies (higher number and greater diversity of immigrants).

Studies conducted in Québec and in California (Bourhis et al., 2010) showed that respondents who endorsed more restrictive acculturation orientations toward immigrants (assimilationism, segregationism, or exclusionism) had more conflictual relationships with them than those who favored more welcoming orientations. Studies carried out in Québec further revealed that more restrictive acculturation orientations were endorsed with respect to devalued – visible minorities – but not with valued immigrants – Christian White immigrants from Europe.

Reducing Xenophobia

Different strategies have been proposed for reducing prejudice. They are generally based on the social psychological theories that have been mobilized to identify its causes and can be applied to xenophobic instances of prejudice. However, most research on prejudice reduction has been conducted in laboratories with artificial groups (Paluck and Green, 2009), or focused on racist prejudice against Black people, and cannot be blindly extrapolated to xenophobia. There exist, however, interesting exceptions.

Research assessing Allport's (1954) contact hypothesis has consistently confirmed that intergroup contact reduces prejudice, provided it takes place on an equal footing, involves cooperation (positive interdependence), allows for the development of close relationships, and receives institutional support. For example, a longitudinal study with school students from three European countries – Germany, Belgium, and England – showed that contact between host majority members and immigrant minorities' members had a positive effect on majority members' subsequent attitudes toward immigrants (Binder et al., 2009). But the reverse relation was also true: prejudice reduced opportunities for contact. These effects were strengthened if immigrant friends were seen as typical members of their group. In addition, reduction in intergroup anxiety mediated this effect. Interestingly, however, this positive effect of contact was not found among immigrant participants.

Building on *Social Identity Theory*, scholars have explored the potentially beneficial effects of different solutions based on social categorization strategies (Paluck and Green, 2009): The recategorization strategy aims at reducing prejudice by decreasing the salience of a social identity while emphasizing a superordinate level of identification that includes both the in-group and the out-group; the crossed categorization strategy is based on the idea that making members of both groups aware that they share a belonging to a third group should improve their relations; finally, the dual identification strategy tries to make a superordinate category salient, while at the same time maintaining identification with the subgroup. Each strategy was found to yield significant reduction in prejudice, although mainly in laboratory settings using artificial groups and with prejudice being measured as in-group bias. A study conducted in Australia by Nickerson and Louis (2008) in real settings brought support for the dual identification strategy. It showed that attitudes toward asylum seekers were the most positive among participants who both identified with their nation and as human beings.

Prejudice against foreigners has also been investigated in the framework of models addressing social norms and conformity. Conformity to group norms proves to depend on perceived in-group threat and on the extent to which the norm provides an appropriate answer to such a threat. Falomir-Pichastor et al. (2004) observed that Swiss nationals conformed to an antidiscrimination in-group norm (i.e., they reduced discrimination against foreigners) when the

out-group was perceived as nonthreatening (i.e., when foreigners were not perceived as taking jobs nor increasing nationals' unemployment). However, the same norm did not obtain any influence, and a counterconformity effect even tended to appear, when the out-group was perceived as threatening. Hence, adhering to antidiscrimination norms endorsed by the in-group tends to decrease prejudice, except when immigrants are perceived as threatening economic or symbolic interests.

Xenophobia vs Xenophilia

Xenophobia and racism are difficult to oppose since these phenomena are subject to transformations and often become confused. Interestingly, however, unlike racism and other related notions, xenophobia has an antonym: Xenophilia, meaning sympathy for the stranger. Thus, xenophobia can be considered as the opposite of hospitality. In a variety of historical situations, foreigners have been marked out as not belonging to the political community. Yet, embodying different characters – the merchant, the pilgrim, the beggar, the fairground stallholder – they held a designed place among local citizens and natives. According to the roman distinction between *urbs* – the material town – and *civitas* – the civic city – they brought novelty to the former while being excluded from the latter. While xenophobia reflects the refusal to incorporate the foreigner as a citizen, it develops in the framework of a bond or even interdependence between outsiders and the host society.

Indeed, whereas the total stranger triggers fears and hostile reactions, the foreigner can become familiar when his or her otherness is mastered through categorization processes, social practices, and institutional controls that define his or her status within a given community. Besides antagonism, the relationship with the stranger is also open to alliance and reciprocity, sometimes represented through rituals or recognition signs. In premodern times, for example, the traveler receiving hospitality would give his host half of a piece of clay, a *symbolon*, and keep the other half for himself, in such a way that, when the host traveled in turn, he could be recognized abroad as a guest rather than a stranger.

Nowadays, the rising of xenophobia conceals opposing trends in most Western societies toward the development of multiculturalism, the adoption of antidiscrimination laws and diverse forms of support to immigrants and asylum seekers claiming for their rights. Opposing xenophobia through institutional devices, volunteer support, or political mobilization contributes to shape today's complex and contradictory forms of the relationships with foreigners. Contemporary manifestations of xenophilia involve political and ideological positions as well as individual and collective behaviors, which still realize nowadays some of the fundamental dimensions of human experience, such as persons mobility, intergroup contact, mutual social influence, and integration of novelty.

See also: Dehumanization; Essentialism; Ethnicity and Migration in Europe; Immigration: Social Psychological Aspects; Intergroup Relations; Meritocracy and Tokenism; Racism: Social Psychological Perspectives; Social Dominance Orientation; Social Identity in Social Psychology; Social Psychology; Stigma, Social Psychology of.

Bibliography

Allport, G.W., 1954. The Nature of Prejudice. Addison-Wesley Pub. Co., Cambridge, MA.

Binder, J., Zagefka, H., Brown, R., Funke, F., Kessler, T., Mummendey, A., Maquil, A., Demoulin, S., Leyens, J.P., 2009. Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis among majority and minority groups in three European countries. Journal of Personality and Social Psychology 96, 843–856.

Bourhis, R.Y., Montaruli, E., El-Geledi, S., Harvey, S.P., Barrette, G., 2010. Acculturation in multiple host community settings. Journal of Social Issues 66, 780–802.

Dawkins, R., 2006. The Selfish Gene. Oxford University Press, Oxford.

Deaux, K., 2008. To be an American: immigration, hyphenation, and incorporation. Journal of Social Issues 64, 925–943.

Esses, V.M., Dovidio, J.F., Semenya, A.H., Jackson, L.M., 2005. Attitudes towards immigrants and immigration: the role of national and international identity. In: Abrams, D., Hogg, M.A., Marques, J.M. (Eds.), The Social Psychology of Inclusion and Exclusion. Psychology Press, New York, pp. 317–337.

Esses, V.M., Jackson, L.M., Armstrong, T.L., 1998. Intergroup competition and attitudes toward immigrants and immigration: an instrumental model of group conflict. Journal of Social Issues 54, 699–724.

Esses, V.M., Veenvliet, S., Hodson, G., Mihic, L., 2008. Justice, morality, and the dehumanization of refugees. Social Justice Research 21, 4–25.

Falomir-Pichastor, J.M., Munoz-Rojas, D., Invernizzi, F., Mugny, G., 2004. Perceived in-group threat as a factor moderating the influence of in-group norms on discrimination against foreigners. European Journal of Social Psychology 34, 135–153.

Faulkner, J., Schaller, M., Park, J.H., Duncan, L.A., 2004. Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes. Group Processes & Intergroup Relations 7, 333–353.

Green, E.G.T., 2009. Who can enter? A multilevel analysis on public support for immigration criteria across 20 European countries. Group Processes & Intergroup Relations 12, 41–60.

Huntington, S.P., 1996. The Clash of Civilizations and the Remaking of World Order. Simon & Schuster, New York.

Joffe, H., 1999. Risk and 'The Other'. Cambridge University Press, Cambridge.

Licata, L., Sanchez-Mazas, M., Green, E.G.T., 2011. Identity, immigration, and prejudice in Europe: a recognition approach. In: Schwartz, S.J., Vignoles, V.L., Luyckx, K. (Eds.), Handbook of Identity Theory and Research. Springer, New-York, pp. 895–916.

Mugny, G., Sanchez-Mazas, M., Roux, P., Perez, J.-A., 1991. Independence and interdependence of group judgments: xenophobia and minority influence. European Journal of Social Psychology 21, 213–223.

Nickerson, A.M., Louis, W.R., 2008. Nationality versus humanity? Personality, identity, and norms in relation to attitudes toward asylum seekers. Journal of Applied Social Psychology 38, 796–817.

Paluck, E.L., Green, D.P., 2009. Prejudice reduction: what works? A review and assessment of research and practice. Annual Review of Psychology 60, 339–367.

Pehrson, S., Brown, R., Zagefka, H., 2009. When does national identification lead to the rejection of immigrants? Cross-sectional and longitudinal evidence for the role of essentialist in-group definitions. British Journal of Social Psychology 48, 61–76.

Pereira, C., Vala, J., Costa-Lopes, R., 2010. From prejudice to discrimination: the legitimizing role of perceived threat in discrimination against immigrants. European Journal of Social Psychology 40, 1231–1250.

Pettigrew, T.F., Meertens, R.W., 1995. Subtle and blatant prejudice in Western Europe. European Journal of Social Psychology 25, 57–75.

Pettigrew, T.F., Christ, O., Wagner, U., Meertens, R.W., van Dick, R., Zick, A., 2008. Relative deprivation and intergroup prejudice. Journal of Social Issues 64 (2), 385–401.

Sanchez-Mazas, M., 2004. Racisme et xénophobie. Presses Universitaires de France, Paris.

Sayad, A., 1991. L'immigration ou les paradoxes de l'altérité. Kinsek/Méridiens, Paris.
Scheffer, P., 2011. Immigrant Nations. Polity Press, Cambridge.
Sherif, M., 1967. Group Conflict and Co-operation: Their Social Psychology. Routledge & K. Paul, London.
Simmel, G., 1908/1984. Digressions sur l'étranger. In: Grafmeyer, Y., Joseph, I. (Eds.), L'Ecole de Chicago. Aubier, Paris, pp. 53–61.
Stangor, C., Crandall, C.S., 2000. Threat and the social construction of stigma. In: Heatherton, T.F., Kleck, R.E., Hebl, M.R., Hull, J.G. (Eds.), The Social Psychology of Stigma. Guilford Press, New York, pp. 62–87.
Tajfel, H., Turner, J.C., 1986. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), The Psychology of Intergroup Relations. Nelson-Hall, Chicago, pp. 7–24.
Velasco Gonzalez, K., Verkuyten, M., Weesie, J., Poppe, E., 2008. Prejudice towards Muslims in the Netherlands: testing integrated threat theory. British Journal of Social Psychology 47, 667–685.
Verkuyten, M., 2006. Multicultural recognition and ethnic minority rights: a social identity perspective. European Review of Social Psychology 17, 148–184.
Ward, C., Masgoret, A.M., 2008. Attitudes toward immigrants, immigration, and multiculturalism in New Zealand: a social psychological analysis. International Migration Review 42, 227–248.
Zick, A., Wagner, U., Van Dick, R., Petzel, T., 2001. Acculturation and prejudice in Germany: majority and minority perspectives. Journal of Social Issues 57, 541–557.

Y

Youth Culture, Anthropology of

Vered Amit, Concordia University, Montreal, QC, Canada

© 2015 Elsevier Ltd. All rights reserved.

Abstract

In considering the effects of globalization and urbanization in many parts of the world, anthropologists' attention has been drawn to the implications of these processes for a large but often marginalized global cohort of young people. To address these circumstances, ethnographic studies in this field have also expanded from a primary focus on peer cultures to include more concerted accounts of intergenerational relations as well. Contemporary research in anthropology is therefore likely to situate young people within a complex set of networks, processes, and activities, including a more varied treatment of the relationship between youth and culture.

For much of the history of anthropology, as for other social sciences, scholarly attention toward youth was shaped by a paradigm of socialization and a focus on adolescence as a life cycle stage. Since the mid-1990s, this orientation has shifted to include both a more concerted focus on young people in their own right as well as a more diversified set of approaches. This shift reflected the conjuncture of several different influences: the earlier seminal work of what has been called the Birmingham School, whose studies of youth subcultures helped to establish cultural studies as an interdisciplinary field of interest; the rethinking of notions of culture and power within anthropology; and increasing attention to the effects of globalization and urbanization on an expanding global cohort of young people.

As the number of studies focusing on young people has increased, the parameters of ethnographic studies in this field have also expanded from a primary focus on peer cultures to include more concerted accounts of intergenerational relations as well. Contemporary research in anthropology is therefore likely to situate young people within a complex range of relations, processes, and activities. As a result, the relationship between youth and culture has come to be treated in more varied ways including attention to the role of young people as cultural producers as well as to the involvement of a broader range of agents in the construction of age and generational categories and distinctions.

Defining Youth

The range of persons anthropologists include in the category of youths has often varied depending on the cultural context in which this position is being defined and on what question the researcher is asking in respect to this field. In a rare early example of a systematic ethnographic focus on youths, during the 1920s Margaret Mead conducted research centering on a group of girls, ranging in age between 10 and 20 years, who were residing in three villages on the Samoan island of Tau (Mead, 1928/1968). In carrying out this research, Mead was concerned to answer the question of what 'coming of age' meant in Samoa and how this differed from the experience of growing up in the USA. She claimed that unlike the emphasis on storm and stress, which then (and has since) dominated American characterizations of adolescence, growing up in Samoa was a simple, easy matter, a process which involved no special conflicts or dramatic transitions. Beyond the obvious signs of physical maturation, she suggested that there were few differences between Samoan girls still approaching puberty, those who were just passing through this transition, and those already several years beyond it. Given this emphasis, it is hardly surprising that Mead was not concerned with defining the boundaries of adolescence too precisely or distinguishing it sharply from either childhood or young adulthood.

In contrast, Alice Schlegel and Herbert Barry's (1991) much later survey of adolescence in ethnographic accounts of 186 nonindustrial societies was more expressly directed at delimiting this stage cross-culturally. Schlegel and Barry argued that while the social duration of adolescence varied cross-culturally and between genders, it was universally recognized as a social stage of the life cycle and tended to fall between the ages of 11 and 17 or so. When social adulthood was delayed far beyond puberty, a further stage intervened between adolescence and adulthood, a stage that Schlegel and Barry called 'youth.' While adolescence was universal, the stage of youth was only recognized in a minority of societies.

However, this distinction between adolescence and youth as two distinct social stages is not otherwise common in the anthropological literature on youth culture. More often this literature is likely to draw attention to the ambiguity of notions of childhood and youth. In part, these ambiguities reflect

 http://dx.doi.org/10.1016/B978-0-08-097086-8.12168-3

variations in the construction of the life course in shifts across both space and time. But in part, these ambiguities also reflect analytical orientations. For example, a cross-cultural collection of articles on youth cultures edited by Vered Amit-Talai and Helena Wulff (1995) included groups ranging in age between 10 and 30 years with considerable variation in the scope of autonomy and social responsibilities represented among them. However, this collection, unlike Schlegel and Barry's or Margaret Mead's studies, was not directed at defining youth (or adolescence) as a stage of human development. Rather it was concerned with the ways in which young people creatively engage with, are affected by, and in turn help to shape the statuses, images, products, styles, and orientations identified, sometimes very broadly, with youth. In short, it was concerned with the experiences and attributions of youth rather than when and how youths become adults.

While later studies have been much more interested in processes of 'coming of age' they have noted that the movement from childhood to youth to adulthood is ambiguous, situationally defined and often contested (Durham, 2007: p. 118). So collections of case studies are still likely to include the wide range of ages that appeared in the Amit-Talai and Wulff volume. In order to avoid taking for granted definitions of youth that may fit into particular historical moments or places, Deborah Durham (2004: p. 592) adapts a linguistic term to introduce the concept of youth as a "social shifter," a concept that has both an indexical and referential function. Treated in this way, the concept of youth does not so much define an age grade as raise questions about the relationships between different age categories and statuses. However, Karen Tranberg Hansen (2008: p. 5) cautions against the frequent slippage between terms like child and youth that occurs in the burgeoning literature on young people as well as in the delineation of age categories by international development and nongovernmental organizations. The "difference between these terms matters," Hansen argues (Ibid).

The Cross-Disciplinary Antecedents of Youth Culture Studies

While Margaret Mead was very successful as a popularizer of anthropology, she was not equally successful in provoking fellow anthropologists to continue her focus on adolescents. For many years following Mead's work in Samoa, youths largely remained marginal figures in the anthropological record. Concerted interest in this field was much more evident in other social science disciplines than it was in anthropology. During the 1930s, a little over a decade after Mead's initial research in Samoa, the sociologist William Foote Whtye began participant observation fieldwork in the North End of Boston. Whyte's (1943/1981) classic ethnographic study of the organization of peer relations among second-generation young Italian men in a 'slum area' exerted a seminal influence primarily on the development of US urban studies. Yet it was and remains one of the most thorough and sympathetic portraits of youth street life ever produced, effectively countering the notion of social disorganization that had dominated the US sociological literature on inner cities until that point. Not long after Whyte's study was published, the concept of youth culture appears to have been introduced by another American sociologist, Talcott Parsons (1942/1964) and was later taken up more systematically by James Coleman (1961) to denote the circumscription of American youth peer interaction.

It was, however, the work of British cultural theorists during the 1970s and early 1980s which did most to develop youth culture as a central field of research on class, subculture, and consumption. During this period, the Center for Contemporary Cultural Studies (CCCS) at the University of Birmingham, sometimes referred to as the 'Birmingham School,' launched an ambitious collaborative study of British youth subcultures (Hall and Jefferson, 1976). For the CCCS scholars, youth appeared to offer a special vantage point from which to consider the more general dislocation and fragmentation of the British working class as the structure of Britain's system of production, labor force, income distribution, and lifestyles was transformed over the course of the post–World War II period. The emergence of a variety of highly stylized British youth subcultures (i.e., punks, mods, skinheads) appeared to crystallize the growing importance of mass culture, consumption, and leisure practices as bases for increasingly fragmented working class identities. The gaze of the CCCS was self-consciously selective, focusing on young British nonconforming white working class males. The larger population of 'ordinary' working class youths who accommodated to rather than visibly resisted institutional controls, middle-class youths, and girls, with some notable exceptions, were not accorded similar attention, gaps which were well noted by later critics. The members of the 'Birmingham School' were also criticized for reading ideological undertones of class resistance and consciousness into what sometimes appeared to be fairly trivial instances of adolescent rebellion. In spite of an emphasis on case studies, the CCCS published work was notably slim in empirical findings. Here again, there was a notable exception in Paul Willis' (1977/1981) ethnography, Learning to Labour, which followed a group of boys, the 'lads,' through their last year of high school and onto the industrial shop floor. Yet, in spite of these limitations, the Birmingham School's focus on popular culture, consumption practices, resistance and cultural agency, and their attention to youth as a nexus for broader social change, continue to resonate in research on youth culture both within anthropology as well as in other disciplines.

Youth Culture: A Move Away from Socialization

The mid-1990s is often identified as a turning point in attention to youth within anthropology. But this turn really encompasses two shifts rather than one. The first turn marked a determined move away from socialization approaches. In volumes such as Vered Amit-Talai and Helena Wulff's (1995) Youth Cultures: A Cross-Cultural Perspective as well as in volumes dealing with a rethinking of approaches to childhood (James and Prout, 1990; Stephens, 1995), socialization approaches were criticized for entailing less of a focus on youths than on the adults they will eventually become or the

adults attempting to direct and socialize them. Instead an argument was made for attention to youth agency, which was interpreted as an insistence on respecting the capacity of juveniles to be active agents creatively shaping their environment even as they contend with a variety of legal, political, economic, and social constraints imposed on them by virtue of their age and liminal status.

The emphasis on agency emerged within the context of a broader reorientation of anthropology away from the notion of unitary and enduring cultural formations that was more common in Mead's day to a conception of culture that could more easily accommodate internal diversity, fragmentation, and transience. This made it easier for anthropologists to be comfortable with a notion of youth as eliciting specific kinds of knowledge and cultural competence that are different from rather than incomplete versions of adult cultures. It also rendered more persuasive the claim that youth cultures can have broader analytical and social significance even when they are highly ephemeral and situationally specific. During the 1980s and 1990s, critical reappraisals of the conventions of ethnographic accounts, in particular of the ways in which the textual authority exercised by anthropologists can silence or distort the perspectives of the people they are studying, provided a broader context for a reconsideration of the ways in which ethnographic representations had effaced the presence and experiences of children and youths. In a cogent illustration of the dialectical nature of intellectual exchange, an essay by George Marcus (1986) in a volume that was particularly influential in this broader anthropological reassessment, focused on Paul Willis' seminal CCCS ethnographic study of working class boys in school. In turn, volumes of this kind and the critical introspection they prompted influenced subsequent anthropological studies and representations of both children and youth.

But in spite of their conceptual aspirations, the legacy of the Birmingham School as well as the emphasis on agency that was introduced within anthropological work on this subject all too often combined to produce a narrowed emphasis on one domain of young people's lives, their relations with age peers. On the one hand, the Birmingham School tended to treat youth culture as a particular form of working class subculture manifested in peer interactions standing "in opposition to other cultural forms [which] took shape most fully in the domain of leisure beyond the institutional controls exerted in more formal realms" (Amit, 2001: p. 147). On the other hand to the extent that the later 1990s insistence on youth agency sought to compensate for socialization approaches that were criticized as being too adult centered, it too often tended to emphasize peer interactions among children and youths (Ibid). As a result, while the efflorescence of interest in youth during the 1990s did produce a flurry of ethnographic studies, many of these were focused on youth peer relations and popular culture (for example, Pilkington, 1994; Thornton, 1995; Alexander, 1996; Schade-Poulson, 1999). While such an interest is not necessarily problematic in and of itself, when it begins to crowd out consideration of other involvements and relationships it portends three possible consequences. First, as Jennifer Cole has noted, this approach "runs the risk of fetishizing and reifying the category of youth" (2004: p. 573). Second, it decontextualizes the lives of young people, painting a "picture of youth as an entity, which is socially and culturally detached from the surrounding world" (Christiansen et al., 2006: p. 16). Third, because it focuses on such a narrow range of youthful activities, it also threatens the marginalization of the study of youth within anthropology and related disciplines (Amit, 2001). Ironically, therefore, a field of research that had been inspired and shaped by the Birmingham School's orientation toward youth as a vantage point from which to consider broader societal transformations was increasingly circumscribed as a narrow domain of inquiry separate from these larger contexts.

Youth, Generation, and Context

In the 2000s, this narrowed treatment of youth cultural production was challenged by a second shift in this field, which featured a re-embedding of young people's lives across a variety of contexts. A major impetus for this reorientation came from anthropological studies conducted in various African locales. In Africa, anthropologists were encountering a situation in which young people predominated numerically but were often marginalized in terms of access to employment, education, health, and social services as well as political and economic influence (De Boeck and Honwana, 2005). This marginalization occurred within the context of processes of globalization including IMF-imposed structural adjustment programs, the influx of international aid, and the creation of free trade zones (Christiansen et al., 2006; De Boeck and Honwana, 2005; Moyer, 2004; Cole, 2004). The resulting shrinkage in the state, decline in social services, and rising unemployment (Moyer, 2004: p. 120) had an especially severe impact on youths even as they were also targeted as potential consumers (Cole, 2004: p. 573). But the exclusions of young people were further heightened by elite hierarchies predicated on generational divisions. As Jean and John Comaroff (2005: p. 24) have pointed out, these generational divisions are social and political rather than chronological. Indeed as Meyer Fortes argued in 1984, chronological age and generational categories "may be wildly out of sync" (Fortes as cited in Durham, 2004: p. 591). Nicholas Argenti further notes that "African political hierarchies are said to be gerontocratic not because men (and to some extent women) accumulate power as a function of growing old but, on the contrary, because they can only grow old to the extent that they have successfully accumulated wealth and power" (2007: p. 8).

Caught between the workings of globalization and gerontocratic hierarchies, young people in many African locales have found previous avenues for mobility blocked. Whereas schooling was once considered a means of upward mobility, even young people who have acquired educational credentials can no longer count on gaining preferred white collar jobs but are instead facing un- or underemployment (Durham, 2004: p. 601; Moyer, 2004; Cole, 2004: p. 576; Argenti, 2007: pp. 8–9). The result has been that many young people are not able to acquire the means to establish their own households,

assume responsibility for dependents or exert politicoeconomic influence within more extended family or communal networks. Moreover, without the capacity to assume these responsibilities and authorities, they are not able to achieve a secure adulthood. They are, in short, 'stuck' in the categorical status of youth.

In their interpretations of these kinds of tensions, anthropologists have turned to some older work on issues of generational distinctions and transmission such as Karl Mannheim's (1952) essay on 'The Problem of Generation.' In doing so, they have emphasized the *relational* nature of generational and life course categories, the "specific circumstance that generate and emphasize one generational category over another" (Christiansen et al., 2006: p. 10). Such an approach insists that "youth cannot be analytically separated from the network of social relations in which they are embedded" (Cole, 2004: p. 585) and these social relations include families, elders, and juniors as much as generational peers.

If the marginalization of young people in Africa has been sufficiently acute to make generation an inescapably "dominant line of cleavage" on this continent (Comaroff and Comaroff, 1999: p. 284), many of the processes of exclusion to which anthropologists have drawn attention in this region, are also reflected in the circumstances of youths in other locales as well. Thus in a comparison of young people in Recife (Brazil), Hanoi (Vietnam), and Lusaka (Zambia), Karen Tranberg Hansen has noted that these cities do not afford many of their young inhabitants a decent livelihood (2008: p. 3). Young men studying in Meerut College in North India find themselves remaining in education for long periods after they have tried and failed to get the middle-class jobs (Jeffrey, 2010). Like their contemporaries in many African countries, these Indian students had aspired to positions in government bureaucracies, a form of work that they had grown up viewing as secure and well paid (Jeffrey, 2010: p. 471) but a shortage of employment and a dramatic increase in the numbers of young, educated people has meant that most of them will not be able to secure these kinds of white collar positions. Craig notes that similar kinds of developments have been observed in other postcolonial settings (2010: p. 467) while in their introduction to a volume examining the circumstances of young men in different parts of the world, Vered Amit and Noel Dyck (2012: pp. 11–12) point out that high levels of youth unemployment have extended across *all* world labor markets even as evidence has also accumulated about the uncertain or limited rewards of educational qualifications in many parts of the globe. Hence Karen Tranberg Hansen argues that these circumstances may well be contributing to a "convergence of global youth experiences" (2008: p. 4) in spite of sometimes-stark differences in the respective age profiles of different regions. Amit and Dyck observe that these developments have rendered the efforts of many young people "to conform to dominant and pervasive orthodoxies a double-edged sword: on the one hand, if they do not acquire any or much formal schooling, they risk being viewed as failures; on the other hand, if they do attend school but do not succeed in achieving the kind of 'good jobs' that this education is assumed to enable, then they are also likely to be viewed as failing" (2012: p. 14).

According to Jean and John Comaroff, a global feature of the contemporary world "is a sense of crisis surrounding the predicament of juveniles" (2005: p. 20). Viewed in these terms, the circumstances of young people cannot be slotted away into a segregated domain of activities. Instead, anthropologists working in many different parts of the world increasingly view the contradictions exposed in the construction of youth as a historical and relational category (Durham, 2000: p. 114) as well as in the positioning and experiences of the people attributed this status, as poignantly revealing of the more general – global, national, and local – contexts within which they live.

Extending and Diversifying the Study of Youth

There is a sense in which youth studies in anthropology have returned to the original ambition of the Birmingham School of treating youth as a vantage point from which to examine general social transformations. But more recent efforts seem to have expanded the scope of this standpoint to encompass a much wider range of youth involvements, relationships, and cultural constructions. But if one conceptual emphasis is simply replaced by another, for example if a stress on agency is simply supplanted by one on generation, then the scope of youth studies will once again, be more limited than it needs to be. Fortunately, there are indications that a dynamic interchange is occurring between previously distinct approaches. Taking "age itself as an analytic" (Cole and Durham, 2007: p. 14) allows scholars to deconstruct and re-examine some of the truisms that are often associated with childhood or youth, for example the notion that 'children are the future' (Cole and Durham, 2008: p. 21). In critically examining the construction of different age, life course and generational categories, youth as a status, distinction or set of practices, is treated relationally and contextually. Hence much of the reinvigoration of the classic concept of generations has not rejected the importance of attention to agency but has sought to critically interrogate it within different spheres and activities (Durham, 2000).

Furthermore, elements of this kind of reinvigoration and interchange have appeared across a wide range of youth studies. Thus recent studies by Brent Luvaas of young fashion entrepreneurs and 'indie pop' bands in Indonesia consider the development of youth styles in respect to processes of globalization (Luvaas, 2009, 2010). If Cole and Durham are concerned to pick apart truisms about the future, Luvaas is equally concerned to question the tendency to automatically associate youthful stylistic appropriations with resistance to imported commodities and tastes. Some socialization or developmental approaches have recently featured notions of agency by extending concepts of "active subjectivity" to children as well as adults (Kontopodis and Newnham, 2011: p. 72) or have noted the impact of globalization and urbanization on the 'reorganization of lifecourse cultural models' (Worthman, 2011: p. 432).

Today, considerations of youth that focus variously on style, socialization, agency, generation, life course, groupings, and practices have not necessarily merged but they are much more likely to overlap as the specific questions being posed and processes being examined require. One of the effects of the diversification and extension of the ways in which youth is being

approached is the deconstruction of a concept like 'youth culture.' Youths are certainly being considered as cultural producers but youth is as likely to be approached as a cultural construction in and of itself. The relationship between youth and culture thus raises a wide scope of questions rather than an already demarcated field of investigation.

See also: Adolescence, Sociology of; Childhood: Anthropological Aspects; Youth Culture, Sociology of; Youth Gangs; Youth Movements.

Bibliography

Alexander, C.E., 1996. The Art of Being Black: The Creation of Black British Youth Identities. Clarendon, Oxford, UK.

Amit, V., 2001. The study of youth culture: why it's marginal but doesn't need to be so. Europaea/Journal of the Europeanists VII (1–2), 145–157 (Special issue on Youth).

Amit, V., Dyck, N., 2012. Pursuing respectable adulthood: social reproduction in times of uncertainty. In: Amit, V., Dyck, N. (Eds.), Young Men in Uncertain Times. Berghahn Books, New York and Oxford, pp. 1–31.

Amit-Talai, V., Wulff, H. (Eds.), 1995. Youth Cultures: A Cross-cultural Perspective. Routledge, London.

Argenti, N., 2007. The Intestines of the State: Youth, Violence, and Belated Histories in the Cameroon Grassfields. The University of Chicago Press, Chicago and London.

Christiansen, C., Utas, M., Vigh, H.E., 2006. Introduction to. In: Christiansen, C., Utas, M., Vigh, H.E. (Eds.), Navigating Youth, Generating Adulthood: Social Becoming in an African Context. Nordiska Afrikainstitutet, Uppsala, Sweden, pp. 9–28.

Cole, J., 2004. Fresh contact in Tamatave, Madagascar: sex, money and intergenerational transformation. American Ethnologist 32 (4), 573–588.

Cole, J., Durham, D., 2008. Globalization and the temporality of children and youth. In: Cole, J., Durham, D. (Eds.), Figuring the Future: Globalization and the Temporalities of Children and Youth. School for Advanced Research Press, Santa Fe, pp. 3–23.

Cole, J., Durham, D., 2007. Age, regeneration and the intimate politics of globalization. In: Cole, J., Durham, D. (Eds.), Generations and Globalization: Youth, Age and Family in the New World Economy. Indiana University Press, Bloomington and Indianapolis, pp. 1–28.

Coleman, J.S., 1961. The Adolescent Society. Free Press of Glencoe, New York.

Comaroff, J., Comaroff, J., 2005. Reflections on youth: from the past to the post-colony. In: Honwana, A., De Boeck, F. (Eds.), Makers and Breakers: Children and Youth in Postcolonial Africa. Africa World Press, Trenton, NJ and Asmara, Eritrea, pp. 19–30.

Comaroff, J., Comaroff, J.L., 1999. Occult economies and the violence of abstraction: notes from the South African postcolony. American Ethnologist 26 (2), 279–303.

De Boeck, F., Honwana, A., 2005. Introduction: children and youth in Africa: agency, identity and place. In: Honwana, A., De Boeck, F. (Eds.), Makers and Breakers: Children and Youth in Postcolonial Africa. Africa World Press, Trenton, NJ and Asmara, Eritrea, pp. 1–18.

Durham, D., 2000. Youth and social imagination in Africa: introduction to parts 1 and 2. Anthropological Quarterly 73 (3), 113–120.

Durham, D., 2004. Disappearing youth: youth as a social shifter in Botswana. American Ethnologist 31 (4), 589–605.

Durham, D., 2007. Empowering youth: making youth citizens in Botswana. In: Cole, J., Durham, D. (Eds.), Generations and Globalization: Youth, Age and Family in the New World Economy. Indiana University Press, Bloomington and Indianapolis, pp. 102–131.

Hall, S., Jefferson, T. (Eds.), 1976. Resistance Through Rituals: Youth Subcultures in Post-war Britain. Hutchinson, London.

Hansen, K.T., 2008. Introduction: youth and the city. In: Hansen, K.T., Dalsgaard, A.L., Gough, K.V. (Eds.), Ulla Ambrosius Madsen, Karen Valentin and Norbert Wildermuth, Youth and the City in the Global South. Indiana University Press, Bloomington and Indianapolis, pp. 3–23.

James, A., Prout, A. (Eds.), 1990. Constructing and Reconstructing Childhood: Contemporary Issues in the Sociological Study of Childhood. Farmer Press, London, New York, and Philadelphia.

Jeffrey, C., 2010. Timepass: youth, class and time among unemployed young men in India. American Ethnologist 37 (3), 465–481.

Kontopodis, M., Newnham, D.S., 2011. Building bridges in dialogue with the future: an introduction. ETHOS 39 (1), 71–75.

Luvass, B., 2009. Dislocating sounds: the deterritorialization of Indonesian indie pop. Cultural Anthropology 24 (2), 246–279.

Luvaas, B., 2010. Designer vandalism: Indonesian indie fashion and the cultural practice of cut 'n' paste. Visual Anthropology Review 26 (1), 1–16.

Mannheim, K., 1952. The Problem of Generation, Essays in the Sociology of Knowledge. Routledge and Kegan Paul, London.

Marcus, G.E., 1986. Contemporary problems of ethnography in the modern world system. In: Clifford, J., Marcus, G.E. (Eds.), Writing Culture: The Poetics and Politics of Ethnography. University of California Press, Berkeley, CA, pp. 165–193.

Mead, M., 1928/1968. Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization. Dell, New York.

Moyer, E., 2004. Popular Cartographies: youthful imaginings of the global in the streets of Dar es Salaam, Tanzania. City & Society 16 (2), 117–143.

Parsons, T., 1942/1964. Essays in Sociological Theory. Free Press, Glencoe, IL.

Pilkington, H., 1994. Russia's Youth and Its Culture: A Nation's Constructors and Constructed. Routledge, London.

Schade-Poulsen, M., 1999. Men and Popular Music in Algeria: The Social Significance of Raï. University of Texas Press, Austin.

Schlegel, A., Barry III, H., 1991. Adolescence: An Anthropological Inquiry. The Free Press, New York.

Stephens, S. (Ed.), 1995. Children and the Politics of Culture. Princeton University Press, Princeton, NJ.

Thornton, S., 1995. Club Cultures: Music, Media and Subcultural Capital. Polity Press, London.

Whyte, W.F., 1943/1981. Street Corner Society: The Social Structure of an Italian Slum. University of Chicago Press, Chicago.

Willis, P., 1977/1981. Learning to Labour: How Working Class Kids Get Working Class Jobs. Columbia University Press, New York.

Worthman, C.M., 2011. Inside-out and outside-in? Global development theory, policy, and youth. ETHOS 39 (4), 432–451.

Youth Culture, Sociology of

Ryan Moore, CUNY-Queensborough Community College, New York, NY, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by M. Buchmann, volume 24, pp. 16660–16664, © 2001, Elsevier Ltd.

Abstract

The study of youth culture has a long history in sociology and the social sciences. The pioneering Chicago School of Sociology was especially concerned with young people as they engaged in delinquency and deviance. Youth culture was enabled by social changes like the extension of public schooling and the growth of consumer culture, and during the 1960s young people would become powerful agents of political and cultural change. The upheavals of the 1960s proved to be a pivotal turning point that transformed sociological approaches to youth culture. Subsequent studies have emphasized the symbolic acts of resistance perpetrated by an increasingly diverse range of subcultures.

The notion of a distinct youth culture and the problems of young people have been enduring topics of study among social scientists. The sociology of youth culture developed through the empirical investigation of issues specific to young people, initially in the psychosocial concept of adolescence, later in studies of juvenile delinquency and deviance, and finally in the analysis of subcultures and their symbolic forms of resistance. The construction of 'youth' as an intermediary phase of the life cycle and a source of identity is inextricably linked to social factors, particularly the extension of schooling and the growth of consumer culture. The rise of youth culture thus parallels the international hegemony of American culture and the mass media, which became most firmly entrenched after World War II but was clearly discernable by the 1920s. The power of youth culture reached its zenith during the 1960s through the hippie counterculture and the New Left, and by the end of the decade young people in many parts of the world had become politically involved as students in social movements. After the 1960s, many scholars began to assess youth culture in terms of its capacity to resist power and create social change. The ethnographic study of subcultures and various forms of youth culture continues to be a vital interdisciplinary field for contemporary scholars.

Adolescence and the Origins of Youth Culture

The notion of adolescence was initially developed by the pioneering psychologist G. Stanley Hall. In a two-volume work published in 1904, Hall defined adolescence as a stage of life between childhood and adulthood, one that social institutions should preserve as a period of self-exploration. The experience of adolescence was shaped by a mixture of biological and sociological processes. Adolescence was characterized by emotional volatility and conflict, which Hall called 'storm-and-stress.' Hall had been strongly influenced by the Romantic writers of the eighteenth century, including Rousseau and Goethe, the latter of whom had written the classic tale of young heartbreak and suicide in *The Sorrows of Young Werther*. Hall argued that the turbulence of adolescence was especially acute among American youth, who were confronted with the lure of vice in the industrializing cities while being deprived of the rites of passage that marked the transition to adulthood in preindustrial societies. Hall advocated for extended years of schooling and other institutional reforms which would allow adolescents to explore themselves while delaying their entry into adult society (Arnett, 1999).

This 'in-between' phase of the life cycle was enabled and extended as a result of the expansion of public schooling for older youths during the early decades of the twentieth century (Nasaw, 1979). Extended schooling concentrated teenaged boys and girls together in institutions, which, unlike the military, could not control all of their behavior and were not typically segregated by sex. High school and college students began to create an age-specific culture centered on their schools, mainly through extracurricular activities and clubs, such as after-school dances, glee clubs and choruses, student government, and school newspapers. At American colleges during the 1920s, the Greek system of fraternities and sororities prevailed as the dominant social grouping with exclusive hierarchies of status and prestige; these quickly became linked to collegiate sports and especially the growing popularity of college football (Fass, 1977). Competition for status among high school and college students created pressure for social conformity along with heightened alertness toward shifts in fashion and fads. These social anxieties about keeping up with peers could be exploited by commercial forces and advertisers who were just beginning to recognize the value of young people as consumers. Youth culture and consumerism would be closely intertwined from this time on, with consumption becoming the primary means of either conforming to peer groups ('fitting in') or competing for the status of nonconformist hipness ('standing out').

While the extension of schooling helped create youth culture by sequestering age groups, youth culture was also developing among working-class youth outside of schools, in the city's streets and spaces of entertainment. In American cities, millions of young people from immigrant families would form a vanguard of cultural change, discarding the ethnic culture and traditions in favor of a modern, Americanized identity shaped by consumer culture and the mass media. Department stores, dance halls, and amusement parks were among the places of commercialized leisure where working-class youth converged and created a cultural style

distinctive to their age. These urban spaces of amusement and consumption presented young working-class women, in particular, with unprecedented opportunities to evade the strictures of patriarchy imposed in their immigrant households (Peiss, 1986). Above all, movie theaters would prove to be the most popular gathering place for young people, as nearly half of all American adolescents were going to the movies at least once a week in 1922. Beginning in the 1920s, the movie studios competed for this coveted audience by releasing different films with 'youth' in their title; one of the most popular and provocative was *Flaming Youth* (1923), which included a 'skinny-dipping' scene of naked bodies shot in silhouette (Savage, 2008).

The segregation of young people into a distinct social group enabled the formation of subcultures characterized by symbolic resistance to social norms. Flappers were a pioneering subculture, which emerged in the 1920s, composed of young women who defied the dominant forms of femininity through their distinctive fashion and association with the hedonism of the 'Jazz Age.' The flapper's signature was a short, bobbed hairstyle, a modern and androgynous alternative to the long, flowing hairstyles that prevailed among Victorian women. The flappers were also associated with shorter skirts and the heavier use of cosmetics, stylistic markers that suggested more assertive attitudes toward sexuality and overall greater independence of self. Like all the other youthful subcultures, which followed, the flappers would be exposed to the larger public through the mass media, who invariably represented them through a prism of sensationalistic anxieties about changing gender roles and sexual behavior. Nonetheless, the 1920s was a decade of major changes among young people who were increasingly engaged in a range of premarital sexual activities and had begun the practice of going on dates without chaperones. Though sensationalized, the images of flappers in the mass media – dancing boisterously to jazz, smoking, and drinking in public, etc. – were a reflection of significant changes in gender roles and sexual norms occurring in the 1920s (Dumenil, 1995).

Deviance and Juvenile Delinquency: Sociology and the Chicago School

In the nineteenth century, authorities of various stripes constructed the social category of juvenile delinquency in connection with a host of social problems arising from urbanization and immigration. This diagnosis of delinquency informed the groundbreaking research of the Chicago School of Sociology beginning in the 1920s, as delinquents and gangs were studied through ethnographic methods and analytically connected to the social ecology of urban life. As the city of Chicago emerged as a focal point for movements toward social reform, the concern for young people led to the creation of the Juvenile Protection Association (JPA), founded by Jane Addams and her colleagues in 1901. Over the next several decades, the JPA supported numerous studies of social problems surrounding urban youth, including prostitution, drug use, child labor, poverty, and racism. The collection of data about young people and the push for social reform were connected with the larger Progressive Movement, particularly those reformers known as the Child Savers who had a crucial role in inventing and defining the category of juvenile delinquency (Platt, 1969). Sociologists at the University of Chicago later inherited these depictions of delinquency, and their studies of youth and gangs would link them to the larger problems of social disorganization in industrial cities.

The Chicago School of Sociology produced its pioneering studies through methodologies that mixed observation and documentation, interviews and oral history, and the study of official statistics. These studies were informed by a broader perspective on urbanization, developed by Robert Park and Ernest W. Burgess, which considered the city as a sort of ecological system, where people competed for resources, adapted to their environment, and formed a kind of equilibrium in their social lives. This social ecology model of urbanization was applied in the groundbreaking studies of gangs by Thrasher (1927) and juvenile delinquency by Shaw and McKay (1942). Thrasher maintained that the attachment to territory was paramount for gang formation, as illustrated by the fact that gangs were ethnically homogenous in homogenous neighborhoods but could also be ethnically mixed if they formed in mixed neighborhoods. He observed that gangs tended to develop from the playgroups of children, and that their bonds of loyalty to one another became stronger as adults and authorities expressed their disapproval. Shaw and McKay built upon the ecological model by stressing the effects of social disorganization with respect to juvenile delinquency, which they saw as symptomatic of the weakening of social bonds in urban neighborhoods. Consistent with Burgess' model of how urbanization develops in zones of concentric circles, Shaw and McKay demonstrated that neighborhoods nearest to the city's center tended to have the highest rates of delinquency, regardless of which ethnic group inhabited them at a given time.

Beyond the Chicago School, youth became a central topic of concern for sociologists during the 1930s, and Talcott Parsons was the first to examine youth culture in an article published in 1942 (Brake, 1985). Parsons (1942) maintained that youth culture was defined by its reversal of the values of adult society, namely hard work and conformity, in favor of consumption, leisure, and mild hedonism. The 'glamour girls,' 'swell guys,' and 'good companions' who were the most esteemed individuals within youth culture were characterized by physical attractiveness, athletic success, and popularity among peers. While Parsons noted the ways in which youth culture inverted mainstream values, he confidently asserted that it had an important function in facilitating the transition from childhood to the social roles of adulthood. The youth culture, which Parsons analyzed, clearly revolved around high schools, rather than city streets and neighborhoods, and he neglected to examine class differences within youth culture. Hollingshead's (1949) ethnographic study of high schools revealed how social class, mediated by family and neighborhood, shaped the status hierarchies of adolescent cliques. Others, including Coleman (1961), downplayed the role of class in favor of a pluralist model, which maintained that youths had access to diverse sources of social status, with Coleman adding that all these status groups within youth culture were united by an antagonism to the educational values of the school.

Juvenile delinquency would reemerge as a volatile issue in American society during the 1950s, and again sociologists

produced a series of studies that documented and analyzed these social problems associated with youth. In 1952, public concern about juvenile delinquency led the US Children's Bureau to create the Special Juvenile Delinquency Project. The US Senate Subcommittee on Juvenile Delinquency was formed in 1953, and the following year it held hearings (led by Sen. Estes Kefauver) to investigate whether comic books were negatively influencing young people and provoking delinquent behavior. These hearings in the Senate expressed growing concerns about the influence of the mass media and commercial entertainment over young people, from comic books to movies to popular music (Gilbert, 1986). In turn, movie studios appealed to the fascination of the public in general and young people in particular with gangs and juvenile delinquency, prompting the controversial releases of films including *The Wild One*, *Blackboard Jungle*, and *Rebel Without a Cause*. Such imagery also played an important role in the evolution of rock 'n' roll in the 1950s, and it was here that young people's attraction toward the defiance of the working-class delinquent would become most apparent.

The sociological analysis of juvenile delinquency centered on the concept of anomie, which Robert Merton (1938) broadened beyond its original formulation in the works of Durkheim. Whereas Durkheim generally defined anomie as the disconnection of individuals from society, Merton revised it to describe the disaffection of those who lack opportunities for upward mobility. Merton's conception of anomie would be fundamental to the development of his strain theory of deviance. As the ideologies of meritocracy and egalitarianism suggested that opportunity was equally available to all, the realities of inequality and social reproduction frustrated those who could not advance through the legitimate institutions of mainstream society. Merton developed a typology of deviance based on the motives of deviant actors with respect to 'cultural goals' and 'institutionalized means.' Other sociologists later added that understanding the motivation was not sufficient for a full explanation of deviance, as one must also explain how people learn to engage in deviant behavior through their association with other deviants (Cloward and Ohlin, 1960).

Albert K. Cohen (1955) built upon Merton's strain theory of anomie and deviance in a study of Chicago-area gangs. Cohen argued that gangs develop among young people who share a similar 'problem of adjustment' in a society divided by class. Gangs tended to form among working-class youth who could not or would not adjust to middle-class values, particularly the middle-class values that promote education as a path to upward mobility. If delinquent, working-class boys faced a 'problem of adjustment,' then the gang or delinquent subculture represented a 'collective solution' by upholding alternative norms and criteria of status. While rejecting schooling, deferred gratification, and other middle-class norms, delinquents formed subcultures that valued masculine toughness, group loyalty, and the ability to negotiate violence. Gangs and delinquent subcultures depend on regular social interaction and an intense degree of group solidarity, which serve to insulate them from the negative judgments of mainstream society. Typically, their values and norms directly invert those of mainstream society, to the extent that "the delinquent's conduct is right by the standards of his subculture, precisely because it is wrong by the norms of the larger culture" (1955: p. 28).

Counterculture and Student Movements in the 1960s

The 1960s were a pivotal moment when young people brought about significant changes to the political systems and cultural norms of societies in many parts of the world. The student activism of the New Left and a more all-encompassing counterculture grew in the US during the 1960s, and oppositional youth culture and student movements would become international phenomena during the riotous events of 1968 and the turbulent years that followed. In the US, college students played a leading role in the Civil Rights Movement during the early 1960s, and opposition to the Vietnam War later in the decade incited the student movement's expansion on an international scale. The counterculture's influence proved to be more diffuse, as an antiauthoritarian style and sensibility spread among young people in search of personal liberation. Although the individualism of the counterculture often conflicted with the collective efforts of the New Left, the youthful culture of opposition also significantly shaped the theatrical tactics and egalitarian ethos of those movements for social change. While the link to radical politics was largely severed after the 1960s, the antinomian culture, which developed during that decade, particularly its music, would exert a lasting influence on various youth subcultures that followed.

The number of students enrolled at institutions of higher education multiplied at its greatest rate during the 1960s. Competition with the Soviet Union led the US government to increase federal spending for higher education through the passage of the National Defense Education Act of 1958. A substantial increase in the birth rate following the end of World War II – America's 'baby boom' – meant that massive numbers of students would enroll at colleges and universities beginning in the mid-1960s. The baby boom generation was poised to benefit from US economic and geopolitical dominance in the decades after World War II, with the ranks of the white middle class swelling and enjoying an unprecedented growth in their standard of living. The growing numbers of young people with increasing disposable incomes made them into a coveted demographic for advertisers, the mass media, and the culture industry, which expanded their appeals to an emerging notion of the 'teenage' market, especially through movies and pop music. The baby boom generation was a celebrated product of American prosperity, but in passing through institutions of higher education it was poised to inherit a vertically managed system linking industry, government, and the military with a consumerist lifestyle centered in recently developed suburbs.

Throughout the 1960s, college students would be key actors in movements for social change, with universities often becoming battlegrounds of conflict. African-American students started the sit-in movement that spread through the South in 1960 and led to the formation of the Student Nonviolent Coordinating Committee the following year. Increasing numbers of White students also started to get involved with the Civil Rights Movement in the early 1960s. After issuing its *Port Huron Statement* in 1962, Students for a Democratic Society emerged as the leading radical organization for students concerned with a host of issues, including American foreign policy and nuclear weapons along with the domestic struggle for civil rights and economic equality. Students could also mobilize in

opposition to the repressive organization and bureaucratic impersonality of the university itself, as evident in the Free Speech Movement at the University of California at Berkeley in 1964. With the escalation of the Vietnam War during the second half of the 1960s, college campuses would become contested spaces subject to teach-ins, demonstrations, occupations, walkouts, and strikes. In the late 1960s, youth began to adopt an oppositional form of collective identity, most directly expressed in the popular saying, 'don't trust anyone over 30.'

Students were at the forefront of an international wave of protest, which swept around the world in 1968, involving both capitalist and communist societies. During the Prague Spring, students had a leading role in liberalizing state control over civil society, which lasted until they were crushed by Soviet forces that August. Meanwhile, students joined forces with the working class to nearly topple the French government in May 1968, leaving behind revolutionary graffiti and slogans, which appealed to desire and the imagination. Particularly during the Spring of 1968, numerous incidents of student protest occurred in Spain, Poland, Italy, Yugoslavia, and West Germany, but also in Algeria, Brazil, Japan, and Pakistan; students were key actors in China's Cultural Revolution, and student demonstrations ended with a horrifying massacre by the military in Mexico City (Katsiaficas, 1987). The global New Left distinguished itself from previous generations with its more horizontal style of organization and participation, and within the student movements political protest was closely intertwined with cultural rebellion and personal freedom. Student protest reached its peak in the US during May 1970, in the days following the invasion of Cambodia and the National Guard shootings at Kent State and Jackson State, which prompted students and faculty at more than 500 colleges and universities to go on strike (Katsiaficas, 1987: p. 120).

The youth culture of the 1960s was profoundly shaped by these movements for social change, while in turn the New Left's political radicalism incorporated much of the style and sensibility of the counterculture. Keniston (1968) and Roszak (1969) argued that youthful radicalism and the counterculture were shaped by their opposition to a totally administered society or 'technocracy' where capital, the state, and mass consumption were fundamentally intertwined. College students quickly discovered that higher education served as a crucial link in this chain of institutions. In response, young people upheld the pursuit of freedom and pleasure against the systemic demands for conformity, instrumental rationality, and deferred gratification. The affluence and security enjoyed by the middle class presented many young people with the opportunity to 'drop out' and experiment with the process of developing an identity and searching for authenticity. Advertisers and media sought to profit from the enormous value of young consumers by incorporating some of the imagery, music, and rebellious attitude of the counterculture; appeals to freedom and nonconformity were thus steered toward consumerism (Frank, 1997).

The vanguard of youth culture in the 1960s was the hippie movement, infamous for its collective pursuit of sex, drugs, and rock 'n' roll but otherwise consistent with some enduring American values of individualism (Miller, 1991). The liberalization of sexuality picked up where the 1920s left off, enabled by innovations in birth control but also by the continuing transformation toward a culture of consumption and individual choice. By the end of the 1960s, the mixture of New Left militancy and countercultural open-mindedness had helped spark the women's and gay rights movements. The hippie counterculture used drugs as tools for the expansion of consciousness and perception, not simply among individuals but as a kind of religious sacrament in the practice of communal joy. Most influential of all was the music scene, which evolved from folk to electrified rock yet continued to be a powerful vehicle of protest and the basis for youthful identities. Rock music appropriated not only the sounds of rhythm and blues but also the defiantly 'cool' bravado of urban African-Americans. Rock music soon proved to be ripe for commercial cooptation, and in becoming big business its symbolic acts of rebellion were increasingly severed from movements for social change.

Resistance and Subcultures: The Birmingham School

In the aftermath of the 1960s, scholars began to examine youth cultures in terms of resistance to hegemonic power, defining them as subcultures engaged in symbolic struggles over culture and style. This approach was developed by interdisciplinary scholars at the Centre for Contemporary Cultural Studies at Birmingham University during the 1970s, and has since come to be known as the 'Birmingham School' theory of subcultures. In contrast to the Chicago School, the theoretical point of departure for the Birmingham School was Marxism, and they conceptualized subcultures as expressions of symbolic resistance from working-class youth against the dominant classes. The Birmingham School revised the orthodox Marxism's notion of a determining base and determined superstructure, thus allowing them to theorize that culture could be a relatively autonomous means of resistance instead of just social reproduction. Youth culture disconnected from political radicalism during the 1970s, but the Birmingham School theorized that subcultures continued to represent indirect, symbolic challenges to the hegemonic powers, particularly through style and the use of signs.

Birmingham School scholars argued that the subcultures of working-class youth maintained a complicated relationship with the 'parent culture' of the working class at large. The mainstream parent culture of the working class includes the elements of resistance and hostility to authority, but overall it is more conventional and accommodating to hegemonic society (Hall and Jefferson, 1976). Youth subcultures are never totally separate from this parent culture, as both are responding to the conditions of social life in the working class. A central idea developed by the Birmingham School was that subcultures represent 'imaginary' or 'magical' solutions to these social-structural issues, exemplified in the way that skinheads sought to recover a lost sense of working-class community by adopting a proletarian style (Cohen, 1972). The demolition of older neighborhoods and construction of high-rise housing projects destroyed the fabric of working-class communities, which had come together in street life or spaces like the local pub and the corner store. At the same time,

a conflict of cultural values between 'traditional working-class Puritanism' and a 'hedonistic culture of consumption' had intensified in the years after World War II through the advent of television and mass culture. Cohen argued that the mod subculture reflected working-class parent culture in their argot and rituals, but their dress and music represented the hedonistic image of the consumer: new and modern, colorful and glamorous. Skinheads, on the other hand, reacted against these social changes by trying to defend their ideal of traditional working-class culture against the hedonistic consumption represented by mods and hippies. Skinheads reasserted an image of masculine chauvinism and Puritanism through their signature style, particularly their workboots, suspenders, and cropped hair, all of which represented them as a 'caricature of the model worker.'

Resistance among youth subcultures is a central theme in one of the most influential studies to emerge from the Birmingham School, Paul Willis' *Learning to Labor* (1977). Willis' ethnographic study examined different groups of working-class youth at school, giving special attention to the rebellious 'lads' who were immersed in a culture of opposition against authority, teachers, their conformist peers, and the whole system of education itself. The lads see their school as an institution of social control, and Willis maintains that they reject the dominant ideology that education is a meritocratic means of upward mobility. The lads have also inherited a working-class 'shopfloor culture' in which masculinity is associated with manual labor, while in turn office work and mental labor are seen as effeminate. Willis argues that the resistance among lads in their schools ultimately serves to reproduce their class position as manual laborers. As teenagers, the lads are eager to drop out and get a job, and it is only as they get older that they will realize how much drudgery and exploitation is involved with manual labor. Willis concludes that this is how working-class kids get working-class jobs, as the culture of opposition to school authorities leaves the lads uneducated and unskilled; ironically, their resistance ultimately reproduces class relations.

Written as punks were defiling the cultural order of British society in 1976–77, Dick Hebdige's *Subculture* (1979) was a similar influential study of youthful resistance that also emerged from the Birmingham School. Here again Hebdige argued that subcultures issue an indirect challenge to hegemonic power through their use of signs. Tracing the development of subcultures from the teddy boys to the mods and rockers to the skinheads and punks, Hebdige presents their semiotic forms of stylish resistance in the social context of a breakdown of postwar Britain's liberal consensus. Subcultures take banal objects, like a safety pin, which seem to have no symbolic meaning, and they appropriate those objects to give them new meanings and uses for which they were not intended. Hebdige (1979: p. 90) argued that "subcultures represent 'noise' (as opposed to sound)" and that they serve "not only as a metaphor for potential anarchy 'out there' but as an actual mechanism of semantic disorder: a kind of temporary blockage in the system of representation."

Within the Birmingham School, some feminist scholars challenged the sexist assumptions behind these conceptualizations of subcultures. McRobbie and Garber (1976) critiqued the Birmingham School for their lack of attention to girls and their silence about masculinity in subcultures. At the heart of the problem is the way that subcultures are defined, as they focus on public spaces where boys have greater access rather than the private spaces to which most girls are confined. Families place greater restrictions on girls and those who spend a lot of time with boys in public are likely to get a bad reputation for sexual promiscuity. Yet McRobbie and Garber maintained that girls still participate in the practices of subcultures, like listening to music and using style as a means of self-expression and experimentation, but the difference was that their activities often take place in more private, domestic spaces. McRobbie and Garber labeled this as the 'bedroom culture' of girls, and they argued that it was imperceptible to male ethnographers of subcultures, like Willis and Hebdige, who focused on the activities of boys in public spaces.

Postsubcultures?

More recently, some scholars have suggested that the Birmingham School theory of subcultures is outdated and needs to be replaced. Many have argued that subcultures are less unified and distinct than the Birmingham School suggests, and moreover that the conflict between subcultures and dominant society is not a simple matter of hegemony and resistance. Such critics have nominated terms like 'neotribes' and 'postsubcultures' to supplant the original concept of subcultures (Bennett, 1999; Muggleton and Weinzerl, 2003). A more fluid, diverse notion of subcultures has become necessary as the variety of styles and identities has multiplied and combined into new hybrid forms during the past few decades. Youth cultures have become increasingly fragmented, overlapping, and ephemeral, and the boundaries separating them from mainstream society are less clear than they used to be. Advances in media and communications technologies have facilitated the multiplication of subcultures and their combination into hybrid forms, while niche marketing allows them to be immediately commodified and assimilated into mainstream society.

Other scholars continue to acknowledge the important contributions of the Birmingham School, while maintaining a critical stance, as they utilize the concept of subcultures in their ethnographic studies of youth culture (Haenfler, 2010; Moore, 2010). The increasing variety of subcultures has prompted scholarly studies of hip-hop (Rose, 1994), extreme metal (Kahn-Harris, 2007), straight-edge punk (Haenfler, 2006), club cultures (Thornton, 1995), and goth (Hodkinson, 2002). They have also drawn attention to the role of young women in subcultures, especially female participation in punk (Leblanc, 1999) and the feminist version of DIY punk known as 'riot grrrl' (Kearney, 2006). While media technologies and niche marketing blur the boundaries between subcultures and popular culture, young people continue to construct identities based on authenticity and a sense of collective difference from mainstream society.

See also: Adolescence, Sociology of; Alienation, Sociology of; Anomie; Counterculture: 1960s and Beyond; Counterculture: The Classical View; Generations, Sociology of; Identity in Childhood and Adolescence; Youth Movements.

Bibliography

Arnett, J.J., 1999. Adolescent storm and stress reconsidered. American Psychologist 54, 317–326.

Bennett, A., 1999. Subcultures or neo-tribes? Rethinking the relationship between youth, style, and musical taste. Sociology 33, 599–617.

Brake, M., 1985. Comparative Youth Culture: The Sociology of Youth Subcultures in America, Britain, and Canada. Routledge, New York.

Cloward, R., Ohlin, L.E., 1960. Delinquency and Opportunity. Free Press, Glencoe, IL.

Cohen, A.K., 1955. Delinquent Boys: The Culture of the Gang. Free Press, Glencoe, IL.

Cohen, P., 1972. Subcultural conflict and working-class community. In: Gelder, K., Thornton, S. (Eds.), The Subcultures Reader. Routledge, New York, pp. 90–99.

Coleman, J.S., 1961. The Adolescent Society: The Social Life of the Teenager and Its Impact on Education. Free Press, Glencoe, IL.

Dumenil, L., 1995. The Modern Temper: American Culture and Society in the 1920s. Hill and Wang, New York.

Fass, P., 1977. The Damned and the Beautiful: American Youth in the 1920s. Oxford University Press, New York.

Frank, T., 1997. The Conquest of Cool: Business Culture, Counter-Culture, and the Birth of Hip Consumerism. University of Chicago Press, Chicago.

Gilbert, J., 1986. A Cycle of Outrage: America's Reaction to the Juvenile Delinquent in the 1950s. Oxford University Press, New York.

Haenfler, R., 2006. Straight Edge: Clean-Living Youth, Hardcore Punk, and Social Change. Rutgers University Press, New Brunswick, NJ.

Haenfler, R., 2010. Goths, Gamers, & Grrrls: Deviance and Youth Subcultures. Oxford University Press, New York.

Hall, S., Jefferson, T. (Eds.), 1976. Resistance through Rituals: Youth Subcultures in Post-War Britain. Unwin Hyman, London.

Hebdige, D., 1979. Subculture: The Meaning of Style. Routledge, New York.

Hodkinson, P., 2002. Goth: Identity, Style and Subculture. Berg, Oxford.

Hollingshead, A.B., 1949. Elmtown's Youth: The Impact of Social Class on America's Adolescents. Wiley, New York.

Kahn-Harris, K., 2007. Extreme Metal: Music and Culture on the Edge. Berg, Oxford.

Katsiaficas, G.N., 1987. The Imagination of the New Left: A Global Analysis of 1968. South End Press, Cambridge, MA.

Kearney, M.C., 2006. Girls Make Media. Routledge, New York.

Keniston, K., 1968. Young Radicals: Notes on Committed Youth. Harcourt, Brace, and World, New York.

Leblanc, L., 1999. Pretty in Punk: Girls' Gender Resistance in a Boys' Subculture. Rutgers University Press, New Brunswick, NJ.

McRobbie, A., Garber, J., 1976. Girls and subcultures. In: Gelder, K., Thornton, S. (Eds.), The Subcultures Reader. Routledge, New York, pp. 112–120.

Merton, R.K., 1938. Social structure and anomie. American Sociological Review 3, 672–682.

Miller, T., 1991. The Hippies and American Values. University of Tennessee Press, Knoxville.

Moore, R., 2010. Sells like Teen Spirit: Music, Youth Culture, and Social Crisis. NYU Press, New York.

Muggleton, D., Weinzerl, R., 2003. The Post-Subcultures Reader. Berg, New York.

Nasaw, D., 1979. Schooled to Order: A Social History of Public Schooling in the United States. Oxford University Press, New York.

Parsons, T., 1942. Age and sex in the social structure of the United States. American Sociological Review 7, 604–616.

Peiss, K., 1986. Cheap Amusements: Working Women and Leisure in Turn-of-the-Century New York. Temple University Press, Philadelphia.

Platt, A., 1969. The Child Savers: The Invention of Delinquency. University of Chicago Press, Chicago.

Rose, T., 1994. Black Noise: Rap Music and Black Culture in Contemporary America. Wesleyan University Press, Hanover, NH.

Roszak, T., 1969. The Making of a Counter Culture: Reflections on the Technocratic Society & Its Youthful Opposition. University of California Press, Berkeley.

Savage, J., 2008. Teenage: The Prehistory of Youth Culture, 1875–1945. Penguin, New York.

Shaw, C.R., McKay, H.D., 1942. Juvenile Delinquency and Urban Areas. University of Chicago Press, Chicago.

Thornton, S., 1995. Club Cultures: Music, Media and Subcultural Capital. University Press of New England, Hanover, NH.

Thrasher, F., 1927. The Gang: A Study of 1,313 Gangs in Chicago. University of Chicago Press, Chicago.

Willis, P., 1977. Learning to Labor: How Working Class Kids Get Working Class Jobs. Columbia University Press, New York.

Youth Gangs

Ross Deuchar, University of the West of Scotland, Hamilton, UK

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by C.S. Taylor, volume 24, pp. 16664–16668, © 2001, Elsevier Ltd.

Abstract

The origins of youth gangs can be traced back to the eighteenth century, but a consensus on gang definitions has never emerged and wide-ranging theories about the nature of gang culture abound. Although some evidence suggests a proliferation of gang violence in many countries outside of the United States, criminologists need to be aware of the way in which media and political influences generate 'gang talkers.' By focusing only on the small minority of prolific gang members who are involved in most of the violence, law enforcement agencies may have more success in preventing the issues from occurring.

Introduction

Youth gangs have been a part of American culture since the eighteenth century, and elements of subcultural issues among youth can be traced back to the nineteenth century in the UK. But, the word 'gang' is often used loosely by those in the media and political circles. Young people living in deprived communities frequently feel that local residents assume youth gatherings and friendship groups are gangs when, in fact, they are not. There is thus confusion about what gangs are among journalists, politicians, and the general public alike.

Where they do exist, gangs are diverse in American and British society, and not all gangs are the same. And defining what actually constitutes a gang can be problematic. Drawing upon those theories which suggest that crime is a social product, this article explores the commonly recognized causes of gang culture, and begins to examine the difficulties associated with defining youth gangs. It examines the diverse ethnic membership of youth gangs in America, and explores some of the differences in those gangs found in the UK as well as providing a brief history of gang culture on both sides of the Atlantic. The role of the media and political rhetoric is examined in terms of the romanticism, mythmaking, and 'moral panics' that can emerge, and potential solutions are examined within the context of focused deterrence strategies being implemented across the USA and (more recently) in Scotland. Implications for future research and law enforcement practice are considered. Following the conclusion, an extensive reference list is provided to assist with further review of the topic.

The Causes and Nature of Gang Culture

In the late nineteenth century, Emile Durkheim argued that a state of normlessness or 'anomie' may occur in the lives of those who become exposed to unlimited socially generated aspirations combined with a lack of realistic opportunities for achieving these goals (Durkheim, 1897/1952). Later, Robert Merton (1938) took these arguments further when he drew attention to the way in which elements of social and cultural structure exert a 'strain' on some people to engage in nonconformist conduct. He argued that 'aberrant conduct' might arise where there is a disassociation between culturally defined aspirations and 'socially structured means' of achieving these goals (Merton, 1938: p. 674). In Merton's view, large sections of the American population found themselves unable to achieve their aspirations through conventional, legitimate means. Accordingly, the disjunction between means and goals led to a propensity for deviation to emerge (see also Deuchar, 2013).

Albert Cohen (1955) subsequently highlighted that the groundwork for violent behavior includes the existence of unresolved tensions and the drifting together of individuals who experience social exclusion (Deuchar, 2013). In order to deal with these issues, Cohen argued that many young people (and young men in particular) may participate in delinquent subcultures. In so doing, it was felt that young people may engage in innovation, whereby conflict and frustration are eliminated by rejecting conventional means of achieving success and by creating alternatives. Alternatively, they may engage in rebellion, where extreme forms of frustration and marginalization lead to attempts to introduce a 'new social order' (Merton, 1938: p. 678).

Since the beginning of the last century, 'gangs' have been the focus of sociological research and have been commonly cited as manifestations of delinquent youth subcultures. Puffer (1912) was one of the earliest authors in the USA to explore the nature of the gang. He described it as a 'play group,' and argued that gang fighting tended to be centered around personal disputes related to the upholding of a 'code of honor' or could be about territorial disputes with other gangs (Puffer, 1912: pp. 67–68). In early work conducted as part of the Chicago School of Sociology, Frederic Thrasher (1927: p. 57) argued that gangs were characterized by behavior such as 'milling,' 'movement through space,' and conflict that results in 'solidarity' and 'attachment to a local territory.'

More recently, Elijah Anderson's (1999: p. 32) pioneering ethnographic study into urban street life in Philadelphia, the USA, highlights that the inclination to violence "springs from the circumstances of life among the ghetto poor – the lack of jobs that pay a living wage, limited basic public services … the stigma of race, the fallout from rampant drug use." Young men who feel disrespected by others out on the street need to avenge their honor, and require 'running buddies' or 'homies'

that can be depended upon to back them up (Anderson, 1999: p. 73).

The Eurogang Project is a thematic network for comparative research on violent youth groups across Europe and the USA. Two of its members, Frank Van Gemert and Mark Fleisher, have argued that most gangs do not actually share the stereotypical characteristics associated with hierarchical ranks and initiation ceremonies. In order to facilitate gang research across borders, the Network has created a generic definition of a gang as being "any durable street-oriented youth group whose involvement in illegal activity is part of its group identity" (Van Gemert and Fleisher, 2005: p. 12). Indeed, Hagedorn (2008) has highlighted that the deindustrialization process across inner-city areas in the USA and the wider world has led to more violence being motivated by economic impulses. However, other evidence suggests that gang culture in the USA has become associated more with issues of disrespect and the norms and narratives of the street, as opposed to purely being linked to drug markets and economic enterprise (Anderson, 1999; Deuchar, 2013).

Consensus of definition with respect to 'gang' culture does not exist today, with criteria varying across jurisdictions. Many researchers have created their own versions of what constitutes a gang (Klein, 1968; Monti, 1993; Van Gemert and Fleisher, 2005). In fact, every organization that comes in contact with gangs has created an operational definition of the term to suit its own needs. Many recognize that there are different patterns of behavior associated with different types of gangs. Although some young gang members earn a living through drug trafficking, sales of illegal weapons, and other criminal activities, others join gangs for noncriminal reasons. For example, a young person may join a neighborhood gang to get protection in their neighborhood; another may join to gain status, respect, or the type of social support he or she would otherwise not receive at home or in the community; and still others may join for the camaraderie. Whatever the reason, gang members in America are as diverse as the gangs themselves.

Emergence and Growth of American Gangs

Although sociological research into gang culture emerged much later, the earliest record of youth gangs in the USA may have been as early as the eighteenth century. Some of these gangs were known as Smith's Vly Gang, the Bowery Boys, the Broadway Boys, and the Long Bridge Boys (Osman, 1992). These gang members were not necessarily criminals who committed violent crimes, although they did fight with rivals. Redfield (1941) believed the first gangs in the USA migrated from Mexico in the nineteenth century after the Mexican Revolution in 1813. It was around this time that the first criminal gangs were formed. These were youth gangs involved in criminal activity, due in part to the economic situation and the population increase in urban areas.

The Irish immigrants were the first to form criminal gangs in New York City in the 1820s. The 1850s saw an increase in gang membership as urban government corruption was rampant. During the early years after the civil war ended, migration patterns increased for industrial centers like New York, Chicago, and Detroit. The Irish, Jewish, and Italian immigrants who came were impoverished and formed gangs based on ethnicity. It was also at this time that drugs were introduced to gang life. The early part of the twentieth century was dominated by economic turmoil. Due to the increase in population of the urban areas and the decline in the economy, the gap between the rich and poor grew. In the 1920s, Thrasher (1927) estimated that there were over 1000 youth gangs operating within Chicago. He argued that the gang provided a sense of refuge and belonging for young men in socially disorganized communities and was a "substitute for what society fails to give…it fills a gap and affords an escape" (Thrasher, 1927: p. 33).

With the growth of immigrant populations, gang culture continued to expand within the USA. In the early 1940s, large numbers of Puerto Ricans entered New York City. This fact, along with a growing African-American population from the south, contributed to the large minority populations in northern cities. While Irish, Italian, Jewish, and other Eastern European ethnic groups were establishing their communities, Puerto Ricans and African-Americans, including strong pockets of black West Indians, became a strong presence. Racial conflict was clear in northern big cities such as Detroit, Michigan, where one of the worst race riots in American history took place in 1943. Groups of white youth gangs roamed the city attacking black citizens (Sheldon et al., 1997). Around the same time in Los Angeles, the Zoot Suit Riots of 1943 were underway. In these riots, white residents and visiting soldiers harassed and beat up young Chicano men who dressed in the popular zoot suit style of clothing (Moore, 1978).

Hagedorn (2008) highlights that the crushing of revolutionary 1960s social movements by the state in Chicago led to gang culture growing stronger. Further, two of the most violent American gangs – the *Crips* and the *Bloods* – were founded in Los Angeles, California, in the late 1960s and largely comprised African-American membership. These gangs were known to be involved in murders, robberies, and drug dealing and had intense rivalry between them. In time, the gangs migrated throughout the USA and members can be found in numerous States and prison populations. The gangs of the 1970s, with greater use of guns, were more lethal in their encounters than previously (Horowitz and Schwartz, 1974). By the 1980s, the urban ghettos became overcrowded and the youth gangs transformed themselves into entrepreneurial organizations with the crack cocaine epidemic of that decade (Sanchez-Jankowski, 1991). Indeed, Hagedorn (2008) argues that, in today's world, the spaces of the prison and the ghetto coincide, and that street gangs often have strong links to prison gangs. He argues that 'many of Chicago's supergangs' were formed behind bars, and that gangs often use prison to advance their interests.

In addition to these insights, Hagedorn (2008: pp. xxiv–xxv) also raises three fundamental issues about gang culture. First, he argues that gangs are not a unique form, but one of many kinds of armed groups that occupy the slums, including ethnic militias, drug posses, vigilantes, and political extremists. Second, gangs are shaped by racial and ethnic oppression as well as poverty: as marginalized ethnic groups have lost their faith in 'progress' due to the continued prevalence of racism, many have turned toward the 'nihilistic power of the gangsta.' And third, we can best overcome the negative impact of gang culture by encouraging 'cultural resistance identities' and linking them to

social movements. Accordingly, it is important to draw upon the positive aspects of gang culture as a means of finding potential solutions to the problem of violence. In so doing, we must also recognize that the issues associated with gangs are ones that stem beyond the confines of the USA.

Gangs in the UK and the Wider World

Even although the phenomenon of the gang was traditionally seen as an exclusively American entity, claims have emerged that gang culture is proliferating in other countries (Hagedorn, 2008; Hallsworth and Young, 2008). For example, there have been reports about an increase in Moroccan youth gangs who are strongly associated with hip-hop music and *gangsta rap* (Van Gemert and Fleisher, 2005); evidence of Pakistani gangs in Oslo that perform criminal acts and where membership is passed on from father to son (Lien, 2005); and gang leaders who are viewed as heroes to young boys in Jamaica, Rio de Janeiro, and Central and South America (Deuchar, 2009). Indeed, Hagedorn (2008) argues that the proliferation of gangs in Africa and Latin America is due to the combined influence of urbanization, immigration, and social marginalization.

In the UK, it has long been recognized that gangs do not exist solely within the context of criminality. In the 1960s, Yablonsky (1967) classified the type of gangs in evidence in inner-city neighborhoods within the UK into several categories. Firstly, there were 'social' gangs, where a certain degree of responsible social interaction may take place. Secondly, 'delinquent' gangs were primarily organized to carry out illegal acts such as burglary and assault for purposes of profit. And, thirdly, he defined 'violent' gangs as those organized around violence as a means of 'self-protection and defense' or to 'channel aggression' (Yablonsky, 1967: p. 174, and see also Deuchar, 2013). Academics associated with the Eurogang network argue that, unlike the USA, youth gangs in the UK have fluid membership and flexible boundaries and are often involved in less criminally oriented activity (Van Gemert and Fleisher, 2005; Klein, 2008).

One city with a long history of violent gang culture is Glasgow in Scotland. The earliest recorded gangs in Glasgow date back to the late nineteenth century, with the emergence of the *Penny Mob*s. Like their New York City counterparts, these gangs were formed following the large-scale Irish migration that resulted from the famines of the mid-nineteenth century. By the early twentieth century, the smaller gangs had converged and the focus of the gang violence was stimulated by the tensions that arose following the Irish settlement and the subsequent founding of the religiously divided Old Firm soccer clubs (Rangers and Celtic FC). The formation of two notorious gangs in the city – the *Billy Boys* (a Protestant gang) and the *Norman Conks* (from Norman Street, where many Roman Catholics lived) characterized the sectarian divisions within the city at the time (Davies, 2007).

During the post–World War II period, the issue of gang culture in the west of Scotland was rarely mentioned in the popular press. However, from 1965 on, an increasing number of articles on juvenile crime began to focus once again on the subject of gangs and violence, and it was reported that Glasgow had the worst gang problem in the UK (Bartie, 2010). Research into Scottish gangs was carried out by James Patrick (1973), whose ethnographic study identified the presence of numerous territorial gangs in Glasgow, which were centered primarily on violence. For young Glasgow boys, the lure of the gang was seen to be driven by the Glasgow culture of 'self-assertion and rebellious independence against authority as a means of attaining masculinity' (Patrick, 1973: p. 170). Today, it has been estimated that Glasgow is home to approximately 170 street gangs, with the vast majority of their activity centered around territorial disputes (Deuchar, 2009). It has long been recognized that the gang provides some young men in Glasgow with a way to "accrue compensatory social capital, and to experience social bonding, status, identity and the expression of a narrowly-defined but widely accepted version of masculinity" (Deuchar, 2009: p. 144). The 'hyper place attachment' found in Glasgow's deprived neighborhoods is driven by lack of opportunities for employment, social stigmatization, and oppression (Kintrea et al., 2011: p. 68). The same type of territorial rivalry has been identified in other communities in England, although some have suggested that the violence found in large inner-city communities in London and Manchester is more closely associated with drug dealing and firearms use as opposed to the territorial knife crime culture found in Glasgow (Stelfox, 1998; Bennett and Holloway, 2004).

Some academics believe that the current interest in gang culture in the UK is attributable to the same type of 'moral panics' that surrounded the emergence of other youth subcultures such as the 'mods and rockers' in the 1960s and the punks in the 1970s. Hallsworth and Young (2008) argue that many of the young people typically labeled as 'gang members' simply do not conceptualize the peer groups with whom they associate as 'gangs.' Some young people may simply regard their peer networks as 'friendship groups' or 'young teams,' while resisting the gang label (Deuchar, 2009: p. 9). Further, Hallsworth (2011: p. 184) highlights that, even if 'gang' members commit offenses it is often "not evident that the offense in question is motivated by 'gang' membership in and of itself." For academics, it is important to move beyond the gang 'gaze.' Street violence, Hallsworth argues, can be normless and unregulated and often lacks a social orientation. In focusing exclusively on American stereotypes of gang culture and violence, he argues that so-called 'gang talkers' lose sight of the wider issues that lead to street violence (Hallsworth and Young, 2008).

The Role of the Media

The media has been developing a biosketch of a gang member since the Zoot Suit Riots of 1943 and using it to keep fear of diversity in the minds of many. The Watts Riots of 1965 gave the media more ammunition for its exaggerated view of a gangster. Thanks to their portrayal, many people today believe that a gang member is a Latino or African-American illiterate youth, who comes from a female-headed household, in an impoverished urban area. American cinema has a long history of gangster movies. From 'West Side Story' to the more hardcore portrayal of criminal street gangsters today, the cinema has literally been a training ground for gang wannabes. Classic movies such as 'The Godfather' give an excellent

example of Thrasher's (1927) evolution of street gangs, rising from neighborhood playgroups and evolving into successful criminal street gangs. From the portrayal of rebelling youth (i.e., 'Rebel Without A Cause,' 'Blackboard Jungle,' 'Wild Ones,' etc.) to star-crossed love (i.e., 'West Side Story') and the hard-core reality of street gangsters (i.e., 'Menace II Society,' '187,' 'Heat,' 'Boyz in the Hood'), the youth of America did not have to look far for gangster role models.

In recent years within the UK, there has been an explosion of films and media reports on gang culture, which have both romanticized and celebrated gang violence. For example, while films like 'Green Street' celebrate violent soccer gangs, the 2010 Peter Mullan directed movie, 'Neds,' romanticizes 1970s gang violence in Glasgow. Ross Kemp, an English actor and journalist, presented a BAFTA award-winning television documentary series on gangs, while another recent documentary fronted by English actors and singers Gary and Martin Kemp claimed that there is more gang violence in Glasgow than in any other part of the UK. As a consequence of this intense media publicity, public perceptions about gang culture and violence are on the increase even while crime rates are falling.

Increased political focus on gang culture has also led to the demonization of young people. For example, following the English riots in 2011, Prime Minister David Cameron launched an 'all-out war on gangs and gang culture,' describing it as a 'criminal disease' that infects streets across the UK (Hallsworth and Brotherton, 2011: p. 3). Criminal gangs were seen to be key players in the riots, while young people were viewed as being 'out of control' (Hallsworth and Brotherton, 2011: p. 2). And yet, it was later found that fewer than 10% of those who participated in the riots outside of London were known to be gang members.

Society, and in particular some media and political forces, has been ready to assign responsibility for brutal acts exclusively to so-called violent gangs. But the reality is that gang members run the spectrum of angels to demons, and very few of those we might describe as gang members are hardcore, violent individuals. Many may hang out in street-oriented social groupings but never participate in violence. Indeed, researchers have found that most are peripherally involved in violence; more important, there is evidence to show that only a tiny percentage of gang members are responsible for the violence. This leads us to conclude that specific strategies need to be adopted that target the violence initiated by a very small number of prolific offenders.

Strategies for Preventing Gang Violence

Over the years, there have been numerous attempts to try and combat youth violence associated with gangs in the USA. Although suppression strategies such as zero tolerance policing have resulted in some positive benefits in cities such as New York, they have also brought with them a reduction in public perceptions of police legitimacy and equity (Burke, 2004). But, in more recent years law enforcement agencies in the USA have adopted strategies associated with focused deterrence and problem-oriented policing (Kennedy, 2009). First used as part of the Boston Ceasefire in the 1990s, these initiatives are based on the fundamental presumption that the majority of serious violence is committed by a small number of individuals who are often organized (to some degree) in groups or gangs. Gang members are compelled to attend offender notification meetings (or 'call-ins') where they are told that, if a violent incident occurs, law enforcement will punish those who commit violence as well as their associates or 'gangs.' But, service and community partners also offer help in the form of social and job services for those who are willing to move away from violence (see Deuchar, 2013).

The principles were used to form the basis of the Cincinnati Initiative to Reduce Violence (CIRV) from 2007 onwards. Recognizing that the majority of the violence within Cincinnati was caused by a small number of individuals, the initiative drew upon a range of stakeholders including political leaders, law enforcement agencies, academics, medical professionals, Street Advocates, social service providers, and community and business leaders. Targeted gang members (who were mostly young African-Americans males in their 20s) were routinely brought into Hamilton County Court as a condition of their probation or parole and exposed to some powerful messages. First, law enforcement officials warned them that the violence must stop, and drew upon a 'pulling levers' approach to ensure that any and all legal resources were brought into play to punish individuals and their associates who continued to involve themselves in violence (Kennedy, 2009; Engel et al., 2013). Second, community voices appealed to the young men to leave violence behind: reformed offenders and mothers of victims provided messages that were both emotive and effective. And, finally, social service providers described the alternatives available to a violent lifestyle and promised streamlined social service assistance to those willing to transition to a prosocial lifestyle (Deuchar, 2013; Engel et al., 2013).

A quantitative evaluation of CIRV found a statistically significant 42% reduction in gang member involved homicides in Cincinnati over a 5-year period (Deuchar, 2013; Engel et al., 2013). In 2008, the initiative was implemented in Glasgow, Scotland, with some adaptations to suit the local gang problems. The majority of those targeted in Glasgow were young white men, with an age range of 14–19 years. In a similar way to the Cincinnati initiative, the Glasgow CIRV aimed to reduce youth violence through a mixture of enforcement and suppression strategies, the offering of services and programs and the moral voice of the community (Deuchar, 2013). In the final evaluation of the Glasgow initiative, there was found to be a 46% reduction in violent crime in the areas where the initiative was deployed. In particular, significant decreases were recorded in the type of behaviors targeted by the program, such as weapon carrying and gang fighting.

Deuchar (2013) highlights that, on both sides of the Atlantic, the CIRV initiative 'complemented and supported offenders' journeys toward desistance' by providing them with positive institutional influences, supportive human relationships, and an opportunity to take on generative activities. The strategies associated with focused deterrence have also been applied in numerous settings across the United States, and the initiatives have consistently led to positive results (NNSC, 2013). In Scotland and the wider UK, there is also an increasingly common focus on adopting a prevention approach to tackling gang violence through partnership working and the need to provide pathways out of violence.

Conclusion

This article has drawn upon theoretical perspectives that suggest that crime is a social product, and identified the way in which classic criminologists have drawn attention to issues of social disorganization, 'anomie' and strain to explain the reasons that some young people may participate in gang violence. It has provided a historical overview of the research that has been conducted into gang culture on both sides of the Atlantic, exploring the relationship with ethnicity and economic decline. The problems associated with gang definitions have been explored, and the motivation and behavior of gang members in and beyond the context of criminality examined. The article has identified the role that the mass media and political rhetoric can play in romanticizing gangs and/or demonizing young people, and has discussed the potential solutions that have been framed around the need to be clear about who the small minority of prolific violent offenders actually are, alongside the potential for early intervention within the context of focused deterrence.

In the twenty-first century, society needs to define the term 'gang,' taking into account the facts that there are different types of gangs ranging from noncriminal to criminal; loosely knit to highly structured; coming from diverse ethnic backgrounds; experiencing different reasons for joining; as well as choice of activities pursued. In spite of the challenges associated with gang definitions, some consensus must be reached. We also need to ensure that researchers, politicians, and the media avoid demonizing and stigmatizing youth, through differentiating between violent gang members and those who simply become attracted to gang culture for innocuous, social reasons. And law enforcement agencies around the world need to take note of the positive outcomes that have emerged from focused deterrence strategies that concentrate their efforts on those who are responsible for most of the violence. By identifying who the really violent gang members are within a specific neighborhood, focusing efforts on providing them with alternatives to violence while avoiding stigmatizing other young people, perhaps we can find a way of building safer, trusting, and supportive communities.

Bibliography

Anderson, E., 1999. Code of the Street. W.W. Norton and Co., New York.

Bennett, T., Holloway, K., 2004. Gang membership, drugs and crime in the UK. British Journal of Criminology 44 (3), 305–323.

Bartie, A., 2010. Moral panics and Glasgow gangs: exploring 'the new wave of Glasgow hooliganism', 1965–1970. Contemporary British History 24 (3), 385–408.

Burke, R.H., 2004. Introduction. In: Burke, R.H. (Ed.), Hard Cop, Soft Cop: Dilemmas and Debates in Contemporary Policing. Willan, Cullompton, pp. 1–22.

Cohen, A.K., 1955. Delinquent Boys: The Culture of the Gang. The Free Press, New York.

Davies, A., 2007. Glasgow's reign of terror: street gangs, racketeering and intimidation in the 1920s and 1930s. Contemporary British History 21 (4), 405–427.

Deuchar, R., 2009. Gangs, Marginalized Youth and Social Capital. Trentham, Stoke on Trent.

Deuchar, R., 2013. Policing Youth Violence: Transatlantic Connections. IOE Press, London.

Durkheim, E., 1897/1952. Suicide. Routledge, London.

Engel, R.S., Skubak-Tillyer, M., Corsaro, N., 2013. Reducing gang violence using focused deterrence: evaluating the Cincinnati Initiative to Reduce Violence (CIRV). Justice Quarterly 30 (3), 403–439.

Hagedorn, J., 2008. A World of Gangs: Armed Young Men and Gangsta Culture. University of Minnesota Press, Minneapolis, MN.

Hallsworth, S., 2011. Gangland Britain? Realities, fantasies and industry. In: Goldson, B. (Ed.), Youth in Crisis: 'Gangs', Territoriality and Violence. Routledge, London, pp. 183–197.

Hallsworth, S., Brotherton, D., 2011. Urban Disorder and Gangs: A Critique and a Warning. Runnymede, London.

Hallsworth, S., Young, T., 2008. Gang talk and gang talkers: a critique. Crime Media Culture 4 (2), 175–195.

Horowitz, R., Schwartz, G., April 1974. Honor, normative ambiguity and gang violence. American Sociological Review 39, 238–251.

Kennedy, D.M., 2009. Deterrence and Crime Prevention: Reconsidering the Prospect of Sanction. Routledge, New York.

Kintrea, K., Bannister, J., Pickering, J., 2011. 'It's just an area – everybody represents it': exploring young people's territorial behaviour in British cities. In: Goldson, B. (Ed.), Youth in Crisis? 'Gangs', Territoriality and Violence. Routledge, London, pp. 55–71.

Klein, M.W., 1968. Impressions of juvenile gang members. Adolescence 3, 53–78.

Klein, M., 2008. Foreword. In: Van Gemert, F., Petersen, D., Lien, I.L. (Eds.), Street Gangs, Migration and Ethnicity. Willan, Collumpton, pp. xi–xv.

Lien, I.-L., 2005. Criminal gangs and their connections: metaphors, definitions and structures. In: Decker, S.H., Weerman, F.M. (Eds.), European Street Gangs and Troublesome Youth Groups. Alta Mira, Lanman, MD, pp. 31–50.

Merton, R., 1938. Social structure and anomie. American Sociological Review 3, 372–682.

Monti, D.J., 1993. Origins and problems of gang research in the United States. In: Cummings, S., Monti, D.J. (Eds.), Gangs: Origins and Problems of Gang Research in the United States. State University of New York Press, Albany, NY, pp. 3–26.

Moore, J.W., 1978. Homeboys: Gangs, Drugs and Prison in the Barrios of Los Angeles. Temple University Press, Philadelphia, PA.

National Network for Safe Communities (NNSC), 2013. Saving Lives, Saving Communities. http://www.nnscommunities.org/ (accessed 15.05.13.).

Osman, K., 1992. Gangs. Lucent Books, San Diego.

Patrick, J., 1973. A Glasgow Gang Observed. Eyre Muthuen, London.

Puffer, J.A., 1912. The Boy and His Gang. Houghton Mifflin Co, Chicago, IL.

Redfield, R., 1941. Folk Culture of Yucatan. University of Chicago Press, Chicago, IL.

Sanchez-Jankowski, M.S., 1991. Islands in the Street: Gangs and American Urban Society. University of California Press, Berkeley, CA.

Sheldon, R.G., Tracy, S.K., Brown, W.B., 1997. Youth Gangs in American Society. Wadsworth, Belmont, CA.

Stelfox, P., 1998. Policing lower levels of organized crime in England and Wales. The Howard Journal of Criminal Justice 37 (4), 393–404.

Thrasher, F.M., 1927. The Gang. University of Chicago Press, Chicago, IL.

Van Gemert, F., Fleisher, M.S., 2005. In the grip of the group. In: Decker, S.H., Weerman, F.M. (Eds.), European Street Gangs and Troublesome Youth Groups. Altamira Press, Lanham, MD.

Yablonsky, L., 1967. The Violent Gang. Pelican, Middlesex, PA.

Youth, Geography of

Tracey Skelton, National University of Singapore, Singapore

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Youth geographies are complex and highly dynamic. Young people are everywhere in our social worlds and are a significant part of the social fabric of everyday living. As a subdiscipline of geography, the focus on young people is relatively new but the theoretical and empirical work has developed rapidly and our awareness of what young people are doing, how they are doing what they do, and, importantly, where they are doing it is expanding and deepening. This article provides an insight into aspects of these developments and a particular focus on young people and political, urban and development geographies.

Within academia, young people in the past have tended to be invisible, neglected, or focused on only because of their deviant or spectacular qualities. Their 'in-betweeness,' neither child nor adult, has always been an unsettling position; they do not quite fit and consequently are rendered as a source of concern and in need of strict attention. Similar to media focus on young people, particularly in the postwar era in the Western world, this age-group were either not present or represented in extreme ways and linked with violence, gangs, teenage pregnancy, unemployment, disruption, rebellion, and moral panics. In the mid-twentieth century, scholarship on young people tended to be academically located within psychology, criminology, and education. Psychology investigated social, emotional, and biological development to understand how youth developed into adults and how interventions might keep them on the right track to adulthood. Criminology focused on what appeared to be increasing lawlessness and the so-called 'juvenile delinquency.' Education studies focused on the ways in which young people were not succeeding or were disruptive within the school system, particularly in the case of boys. Hence, the dominant emphasis was generally negative, treating young people as a group to be cured, contained, or coerced into becoming the right kind of members of society. Youth (particularly young males) were studied because they were identified as being abnormal, a problem, or failing.

However, more positive and hopeful social science work with young people began to emerge in the late 1960s and 1970s, and this was particularly strong within the United Kingdom, located at the Center for Contemporary Cultural Studies at the University of Birmingham (Holt, 2009; Skelton and Valentine, 1998). Influenced by Marxism (and later by feminism), scholars focused on youth cultures and the ways in which they were resisting and challenging the social structures and orders that they were growing up in. There was recognition that young people had agency, creativity, and the potential for change within society. It was argued that young people's position within age structures and life courses as 'in-between' or 'on the edge' was what brought about their resistive and inventive activities, cultures, and attitudes. For Marxist and feminist scholars who wanted to identify possibilities for, and work toward social change in the direction of greater equality, youth were seen to hold some of that possibility.

In the 1980s and 1990s social scientists recognized the problematic neglect of young people within their purview, not least because of the important roles they play and will play within society. Key actors in social sciences disciplines began to observe the lack of focus on young people and gather data about their lives in much more of an everyday way, rather than focus solely on the problematic or flamboyant. In 1995 one of the first texts to focus on youth cultures within anthropology was published (Amit-Talai and Wulff); the first book on the geographies of youth appeared in 1998 (Skelton and Valentine); and the highly influential 'International Journal of Youth Studies' was first published in the same year. The international journal of 'Children's Geographies,' with its remit of 'advancing interdisciplinary understanding of younger peoples lives' was launched in 2003 and Nicola Ansell's monograph, 'Children, Youth and Development' appeared in 2005. There were a range of other significant publications within youth studies and cultural studies throughout this time but they are beyond the remit of this article, which will focus specifically on geographies of youth.

Why is making young people visible within the social sciences generally, and its disciplines specifically, important? There are several reasons; here I offer just four that demonstrate the imperative for inclusion. First, it is about justice; no particular social group should be excluded from social science scholarship and the benefits that it can bring. Second, it is about logic; young people make up a significant part of the world's population – in 2000 those aged 0 to 24 years constituted 47.5% of the world's population. If we do not understand their lives, how can we make any claims to knowledge about society? Third, it is about impact; when young people are invisible from study and research they are easier to marginalize and exclude; they are missing from debates, policy making, planning, legislation, and this further deepens their relegation. Fourth, it is about curiosity; what, where, how, and why do young people live their lives differently in relation to other age groups and also within their age grouping? What do these complexities bring to disciplinary approaches?

Geographies of youth emerged as a significant subdiscipline within the broader discipline of geography essentially from the mid-1990s onward. It is strongly influenced by other disciplines and theoretical traditions within the social sciences, such as sociology, cultural studies, feminist theory, youth studies, and social theory and, as it gains critical mass, this geographical work influences other social sciences. A central tenet of the geographies of youth is the focus on specific geographical

 http://dx.doi.org/10.1016/B978-0-08-097086-8.72114-3

concepts, namely place, space, and spatiality, and the ways in which young people interact with their social and physical environments (Evans, 2008). Investigating ways in which young people use places, create spaces, and practice spatialities are central and distinctive aspects of youth geography research and analysis. There remain problems of wider integration, however, as geographies of young people are still relatively absent in the more established subdisciplines of the subject (Skelton, 2010 for analysis of political geography; Skelton and Gough, 2013 for discussion of its absence in urban studies) but this is a changing dynamic as the larger subdisciplines respond to the challenge of previous neglect.

Defining 'Youth' and Why 'Geographies' in the Plural?

Age is a slippery concept and identity. It is mapped onto bodies in different ways in different cultures and geographies. Across the globe, physical bodies are labeled with age and such aged identities come with a sense of what a person should be doing; what they should have achieved; what kinds of legal rights they have; and judgments of appropriate behavior. However, the same age does not mean the same thing in the multitude of societies; age is a relational concept. Nevertheless, in almost all societies there are markers of age upon which social expectations are laid down. There are several rights of passage related to age: first/last day at school; first paid job; first sexual partner; learning to drive; religious rituals such as Confirmation, Bar/Bat Mitzvah, or Ameen; reaching the age of majority; entering a marriage/civil union; parenting a first child; first independent home. Some of these are legally constructed and others are social expectations. Certain rights of passage that may be significant to young people (first love, first sexual encounter, first alcoholic drink, first cigarette, first drive) are often condemned socially and outlawed legally if they take place when the person is too young. Many of these restrictions are based on both protection and/or control of young people. Nevertheless, what is evident is that young people do not go through all of these stages, and that even within the same culture they do not do the same things at the same age (hence the importance of the use of the plural 'geographies'). Indeed globally there seems to have been a slowing down or even complete disruption of youth transitions toward adulthood as they may extend their education, fail to get jobs, be unable to support their own families, and remain dependent on the older generations of their families longer (Jeffrey, 2010; Katz, 2004).

Within most work on geographies of youth, authors will define the age boundaries they are referring to; but these are debated, as are the terms used such as adolescent, teenager, young people, younger people, and youth. The United Nations (UN) also tends to slippage and overlap (Skelton, 2007). 'Youth' are defined as 15 to 24 years old; hence children would be those under 14 years. However the United Nations Convention on the Rights of the Child (UNCRC) defines children as those under 18 years or the age which states identify as the age of majority, which is when, in most cases, children legally become adults (United Nations 'Youth' website). However, youth as a category has considerable valence and while the age of majority throughout most of the world is 18 years, it is clear that young adults still face similar issues to those they faced before they became 18 years. Generally, within geography the definitions adopted for 'youth' include teenagers (13 years plus), which can be qualified as younger or older teens; and 'young people' often relates to ages between 16 and 25 years. There are also geographers who write generically about children and young people at a disciplinary scale but most of their empirical research will involve participants in one group or the other, young people or children. It is important to recognize that there are distinct differences in life experiences, competencies, and agencies between a child of 5 years, a teenager of 15 years, and a young adult of 25 years. Within the generic children's and young people's geographies, concern has been voiced that young people are not as central as children and childhood. In this article and the scholarship it considers, young people are older teenagers and young adults up to the age of 25 years; hence the term 'geographies' of youth.

The term 'geographies,' that is geography in the plural, is deliberately used in this article to remind readers of the intense complexities and contradictions that exist within the concept of youthful geographies. There is considerable diversity of youth experiences in different geographies and these intersect with social identities and differences such as race, ethnicity, gender, sexuality, educational status, bodily abilities, and social class. Early 'geography of youth' tended to be Anglo-American centric, and scholars from, and research in, this region remain predominant in the subdiscipline. However, increasingly, geographies of youth from all over the world are expanding, complicating, and enriching this youthful subdiscipline (Cahill and Katz, 2008; Hansen et al., 2008; Holt, 2011; Jeffrey and Dyson, 2008). The term 'geographies' reminds us of the pluralities of meaning, practices, and constructions of youth and youthhood. This allows the subdiscipline of geographies of youth to challenge Western hegemonic constructions of what is, and who are youth – both within academic geography and also in terms of policies and practices that tend to universalize a singular type of youthhood that emanates from the global North. Some definitions of young people and youth simply do not make sense in different parts of the world. The term 'geographies' reminds us that things are just not that simple and that young people's lives are highly complex, differentiated, and dynamic. This article provides insight into a geographical scholarship that is working to be genuinely international in order to explore beyond, and critique the Anglo-American contexts that have been relatively predominant.

Selected Thematic Research in Geographies of Youth

Despite its relative youthfulness as the 'new kid on the block' in geography, geographies of youth has grown rapidly and developed a varied scholarship that has both theoretical import and provides empirical evidence. This article will, by necessity, have to be selective and will focus on three broad areas of geographical work with a strong focus on young people, specifically those that connect with the more established subdisciplines of political geography (geographies of youthful politics), urban studies (young people and the city), and development studies (young people and development). However, there are other significant focused thematic areas emerging in geographies of

youth (that also often include children). These include young people in relation to public space; social practices and institutions (health, employment, education, training); transport, mobilities, and migration; housing and homelessness; religion; embodiment and intergenerationality; social identities. These identified collectivities are loosely defined, not mutually exclusive nor inclusive of all areas of research; they are listed here in order to provide insight into key thematic areas and spheres of scholarship in youth geographies.

Before I consider the three areas of work where other significant subdisciplines of geography (political, urban, and development geographies) connect with youth geographies, let us briefly consider aspects of a generic geographical focus on young people, i.e., in the context of spaces, places, and scales. Such approaches form the foundation for almost all geographical work on young people.

Young People, Spaces, Places, and Scale

Place, space, scale, and spatialities are central concepts in human geography. 'Where' young people are is a central theme within youth geographies. Their 'whereness,' their spatialities, and the ways in which they make, destroy, and transform spaces and places, have constituted some of the key elements of the subdiscipline; it is all highly geographical. Youth geography work has focused on young people's experiences, behaviors, and practices in the key places of social life: home; school; workplace; streets and open spaces; spaces of leisure, religion, and consumption; virtual space. It has demonstrated the ways in which young people are significant and competent social and spatial actors and agents. There has also been exploration of the spatial–temporal roles that young people play in spaces. City streets, supermarket car-parks, disused buildings, empty rural roads, and parks can be transformed into play-scapes, gang sites, illicit spaces, and/or social spaces by young people at different times of the day and in different seasons. Perhaps the most problematic use of space by young people relates to homeless street youth. There is both anxiety and serious concern over the threat and need that such young people pose (this overlaps considerably with development geographies of youth too). However, for some young people living on the streets, there is an agency and freedom that they resist relinquishing (Bordonaro, 2012).

The concept of scale has been debated in relation to the scalar approach of geographers to young people's lives. Too narrow a focus on the microworlds of young people's everyday worlds in spaces and places extracts them from the wider world context; what then can geographies of young people say about broader issues and how can they contribute to geographical knowledge? Adult practices in society have long closed young people's worlds down with the intention of containing and taming them; some geographical scholarship has pursued a similar approach. An important intervention by Nicola Ansell (2009) has been to reject the notion of scale as hierarchical and to see it, rather, as networked and part of a flat ontology. With such an approach to scale, more complex, accurate, and broader understandings of young people's lives can be garnered. We can examine the ways young people interact and engage with issues, such as politics, changes in education, and unemployment, that transcends scale. Such issues go beyond the local/global binary; they spill out over regional and national boundaries and the consequences touch down in different geographies in different ways. Young people are affected by such issues but they are also part of their creation and transformation, and such activities also transcend scale. Young people migrate to find work or education; they challenge destructive influences and demand improvements; they connect with global movements based within social media. Rethinking scale allows geographers to capture the complexities of young people's geographies as they are lived and practiced beyond the micro-spaces and places they are presumed to reside in.

Young People and Political Geography

Geographical work focusing on young people and politics has tended to lag behind such work in sociology, political sciences, and youth studies but it is currently catching up fast, with special issues and books being published that specifically focus on the geographies of young people's politics (Jeffrey and Dyson, 2008; Kallio and Häkli, 2013; Philo and Smith, 2003 and many more journal articles).

Geographers have begun working with the concept of citizenship, mostly to expand its meaning and hence broaden debates about children's and teenagers' roles in citizenship. In a political–legal rights construct, citizenship is something that those below a certain age cannot make legal claim to; an individual has to reach the age of majority as legally established within the nation state or other legal territory to be recognized as a full citizen. Consequently young people below this age have often been ignored in debates about formal politics (sometimes written as Politics with a capital 'P'); not least because they cannot vote and are not yet part of the electorate. Consequently they are understood not to have any influence in formal politics. Geographers and other social scientists have argued that this is a very limited and limiting approach to citizenship; that in fact, citizenship can be more than a formally recognized political and legal status. Citizenship is bound up with civic duties, social activism, and community care and support. It can be bound up with informal politics (or politics with a small 'p'). There is an expectation of good citizenship in almost all societies and what researchers have shown is that young people play significant roles in citizenship through their care, actions, and social labor for their peers, their families, and their communities, long before they reach the age of legal majority. Throughout the world, young people take responsibilities as citizens of their communities and nations in a range of spaces and ways, regardless of their lack of political–legal status as citizens (Weller, 2007).

Part of the interest in destabilizing the concept of citizenship in order to force a recognition of the citizenship-type energy and action of young people has come about through a vigorous interest in participation. This has some connection to the UNCRC (see below for further discussion), which clearly identified the rights that young people should have to participate in processes and decision-making that affect their lives and those of their communities. Geographers are interested in the ways that young people participate (or don't) and

where and how they demonstrate such participation. Hence work has emerged on school councils, youth parliaments, youth-led resistance, and youthful action around issues such as the right to work, environmental protection, and right to education. Geographers have clearly demonstrated that young people, in many different spaces and places and from different social groups, are participating in their societies in complex and dynamic ways – they are most certainly 'doing' politics. Most of this might be construed as informal politics; but when given the opportunity, young people get involved in the action and responsibilities of formal politics through institutions such as youth parliaments and trade unions.

Young people's political activities within their communities are highly varied in terms of context, motivation, and practices. For some young people of Mexican parentage living on the United States side of the border/la frontera, their politics is about language; and they help their parents engage and connect with social, legal, and political institutions because they translate for them. Some US states have rendered this citizenship-like, political–social, supportive activity illegal, as it can burden the young people or place them in situations they should not be part of at such a young age. However, young people continue to resist the legislation, arguing that they have a right to help their families gain the support they deserve and that they are already exposed to various issues within their communities. Their actions as translators between Spanish and English have therefore become a political and resistive act (see Aitken and Plows, 2010). In other communities, young people seek paid employment in all manner of ways to contribute to their families, especially at times of economic crisis. This can bring them into conflict with legislation against underage labor, but throughout the world, many young workers are forming international, working-youth organizations in order to make their case not only for the right to work but also their right to protection as workers. Young people also demonstrate political action and responsibility to challenge social injustice, and set up or get involved in action and community groups against violence, racism, homophobia, sexism, and abuse, for example. Activities related to such political deeds take place somewhere, whether in public space, within institutions such as schools and youth groups, and/or through social media; geographers are therefore alert to the ways in which spaces and places play significant roles in the structure, temporality, and spatiality of such exploits of youthful politics.

Geographical scholarship about young people's politics has made young people's actual and potential political agency, subjectivity, competence, and capacity visible. Geographers have demonstrated the variety of ways, in an assortment of geographies, in which young people are being political and doing politics in order to create change. Geographies of youthful politics are important aspects of political processes and increasingly, the more well-established subdiscipline of political geography is slowly coming to realize that young people have a role to play. When young people come into the remit of political geography they force a rethinking of the foundational concepts and definitions of the subdiscipline, because young people often do politics differently from adults. It is also, as noted above, about social justice. Young people face marginalization and exclusion within the political arena and consequently geographers and geography work to expose such omissions and write young people's political geographies into geographical and social sciences scholarship (Skelton, 2013b).

Young People and Urban Geography

Over half of the world's young people live in urban places and so they are in and of the city; they experience the city, and they are part of making the city. Young urbanites are important actors in all aspects of urban dwelling and, in turn, their own lives are shaped, influenced, supported, and thwarted by city processes and spatial politics, such as segregation, provision, facilities, and neglect. It is important to understand the mechanisms and processes (economic, political, cultural, and social) through which young people are marginalized, excluded, or included as residents of the city; as well as to consider the ways in which young people resist and rework urban inequalities to improve their city lives (Hörschelmann and van Blerk, 2012). Young people are often the most entrepreneurial, creative, and energetic members of their communities. The city is the site of possibility for their inventiveness, which might be related to finding sources of income (vending on the streets); creating designs and entertainment (street art or busking); hustling for jobs; or learning from artisan workers by watching and copying. Greater political and financial investment directed toward urban youth should be an important component of urban planning and policy; in reality this age group is the most neglected.

Geographers have shown that young people use the city in very different ways to adults. Young people are everywhere present in the city despite attempts by planners and policy actors to plan them out of public and commercial spaces; contain them in age-specific institutions; and exclude them from spaces designed for adults. From skateboarding to cycling and from parkour to playing, young people make use of urban spaces in almost impossible, unimagined, yet inventive ways, as countless YouTube videos of stunt cyclists and parkour groups demonstrate (YouTube links 1 and 2). In the city, young people enter places where they are not supposed to be and use spaces in ways that planners and architects could never have predicted. Such adaptability and audacity often brings young people into conflict with other users of the city and also urban authorities, notably security guards and the police. In more progressive and youth-friendly cities, such activities can be a starting point of greater inclusion and participation in planning and creating spaces for young people in the urban landscape, for example, as in Curitiba, Brazil.

In response and as a challenge to the apparent Western-centric approaches to geographies of young people in the city, Hansen et al. (2008) provide an interdisciplinary analysis of life in cities of the South, notably Recife in Brazil, Hanoi in Vietnam, and Lusaka in Zambia. Partly this is about redressing a scholarly imbalance but it is also about the demographics and processes of urban spaces and urbanization in these continents. Whereas Northern/Western cities are experiencing an aging population as older people make up more and more of the population, in Southern cities young people are the growing population and this has significant cultural, political, and socioeconomic impacts and consequences. Hansen et al.

demonstrate that while there are similarities between the young people in Latin-American, sub-Saharan African, and South East Asian cities, and young people in Western cities, particular configurations and geographical relations of history, global processes, economic and political structures, cultural practices, and urban planning, create uncertainties, fears, and hardships of particular intensity in Southern cities. What constitutes suitable or possible transitions for young people; how adulthood is defined and recognized; and what resources exist to support young people varies enormously between cities of the South and the West; between cities within the South; and within different parts of individual cities.

What perhaps is distinctive in the current era is that urban young people in different parts of the world probably know more about the lives of urban youth in other parts of the world than ever before, because of global media, social on-line media, and for some groups, increased mobility and access to international travel. Some of that knowledge, particularly knowledge relating to travel, is particularly one-way, as privileged Western youth take their 'gap years' between school and university to travel, often to cities of the world much poorer than their own. For young people of the South their mobilities may take them to the West but for very different reasons, most notably migration for work or study, legally or illegally. Others may arrive in the West as refugees or asylum seekers, literally moving for their lives. Many young migrants or refugees may be surprised and disappointed to find themselves only able to work and reside in the poorest parts of Western cities, not so very different from their home cities.

Mobilities and immobilities within cities is an area of geographical research that has demonstrated that not only do young people use the city in different ways, but they also move into, out of, and through the city in different ways to adults. Additionally mobilities and immobilities are differentiated between groups of young people, notably in relation to gender, race, and socioeconomic status. Skelton (2013a) demonstrates that in a city like Auckland where cars are relatively cheap and the city is designed for the car user, young people with wealthy parents can gain access to, or own, cars from the age of 17. They experience considerable independent mobility within the city but tend to use their cars to travel to places they know and to engage with people of the same educational and socioeconomic class. On the other hand, young Pasifika people have the lowest rates of car access and rely heavily on public transport (a relatively expensive and scarce commodity in Auckland). They tend to walk a lot in their neighborhoods to see their friends. They too spend time in places they are familiar with and in the company of other islander youth, but these spatialities of socioeconomic status are very different from each other. However, some young people use the transport possibilities in order to explore different parts of their city deliberately, to see what other places are like and with the hope of encounters with people different from themselves; this was particularly true of young women. With friends, some wealthier girls used their cars to escape the sameness of the neighborhood and traveled to other parts of the city for leisure and excitement; poorer Pasifika girls used buses and trains to explore the downtown sites of the marina areas and Queenstreet, with their bars and restaurants, high-end shops, and municipal facilities such as art galleries. Consequently these girls cross urban, social, and cultural boundaries in order to make use of the city and to learn more about other spaces, places, and people. Such mobility can be an important part of developing an urban cultural capital that may enhance their future training and employment opportunities, as they develop confidence in and about the city and what it might have to offer. This demonstrates the ways in which race, gender, and socioeconomic status intersect in young people's lives in ways that directly affect how they access and utilize the urban spaces they dwell in (see Reynolds, 2013; Allen and Hollingworth, 2013 for examples of ethnic identity, social capital, and immobility/mobility in the United Kingdom).

Young People and Development Geography

Geographers have made significant interventions and contributions to placing young people in development geography and wider development studies and agendas, as well as contributing to actual development policy and practice. A focus on young people (and children), within geography, is not just about social justice or the possibilities for intellectual challenges to established conceptualizations of development; it is also about very practical factors, such as numbers and where young people are. Third World countries (also known as the 'majority world,' and 'lesser developed countries') account for 90% of 0–17 year olds and 85% of 15–24 year olds (Ansell, 2005). Hence the majority of the world's young people live in so-called 'developing contexts' and it is inevitable that they play important roles in development processes and practices.

Initially, work focused on the adding of young people (and children) into development studies perspectives; it was about making young people visible within development discourse and practice. More recently, geographers have raised the importance of concomitant rethinking of economic development that places the West as central, and of stage-assumptive child development that places adults as central (Aitken et al., 2008). In more practice and policy contexts, development geographers continue to recognize the complexities and often very deleterious effects, of major problems such as poverty, natural hazards, or state/civil violence in young people's lives. Particularly influential work has been done around HIV/AIDS development work where social reproductive responsibilities, burdens of care, household maintenance, and livelihoods have fallen on the shoulders of young people, in the absence of adult family members due to ill health or death (van Blerk and Ansell, 2007).

Other development issues have included the ways in which young people are victims of violence, particularly state violence and civil war, including the associated forced mobility of refugees. Development geographers have been alert to the ways in which young bodies have been commoditized as laborers, sex workers, and/or trafficked people. Young people, especially young men, have been recognized as perpetrators (as well as victims) of youth violence and creators of social anxieties relating to the formation of gangs and their powerful abilities to create disruption and social fear (Jones, 2008). Work with, and about, young people in a development context has allowed insights into the ways in which global processes – globalization, neo-liberalism, international debt, privatization, structural adjustment, securitization, international agreements, war,

forced migration – touch down into and damage young lives in extreme ways. All of these processes have implications for their productive social and economic capacities in the future. There are also cases where global processes, particularly those with a more social justice imperative such as some development approaches and NGO work, facilitate and enhance young people's resilience and ability to rework their lives into something positive and possible.

Increasingly, and particularly in the decades since the launch of the UNCRC in 1989, the participation of young people in processes and decisions that affect their lives has been encouraged alongside the more established international agreements around provision and protection. Partly through this influence and partly through new academic and political thinking, young people have been recognized as actors and agents within development in their own right; they are progressively seen as part of the sociospatial processes at play, rather than being perceived as passive recipients. However, it is important to resist the universalization tendency based on neoliberal individualism that has emerged through the UNCRC and to bear in mind that not all young people's lives are the same, between and within different geographies. Consequently, young people will have variable access to resources and differential potential to develop various capacities to achieve what they believe to be best for them in their lives and for their communities. Nevertheless the recognition of young people as development agents has been an important shift in approach for development geographers. It is not so much about making the world better for the next generation but about working with current generations to make the world better for now; it is also about not imposing adult agendas on young lives – this is part of adulthood. Young people make important contributions within development and this has to be recognized and worked with. Geographers within development studies have risen to this task through their work with young people.

Conclusion

Youth geographies is an exciting subdisciplinary area of geography and connects across much of the social sciences. Youth geographies also refers to the lived geographies and spatialities of young people in space and place. Young people are increasingly being recognized as, and have for some time been performing the roles of agents within contexts of politics, cities, and development (along with other aspects of everyday life). It has taken time for the academics to catch up and explore further what young people do, where they are, and how they live. Fortunately for young people themselves, their families, and their communities, young people have been living, doing, and being geographically active always.

See also: Aging and Older People, Geography of; Children's Geographies; Development Geographies; Gender and Feminist Studies in Geography; Globalization: Welfare and Social Work; Migration: Anthropological Perspectives; Place in Geography; Political Geography; Race and Racism, Geography of; Scale in Geography; Sexuality and Space; Social Movements, Geography of; Transportation Geography; Urban Geography.

Bibliography

Aitken, S., Lund, R., Kjörholt, A.T., 2008. Why children? Why now? In: Aitken, S., Lund, R., Kjörholt, A.T. (Eds.), Global Childhoods: Globalization, Development and Young People. Routledge, London, pp. 3–14.

Aitken, S., Plows, V., 2010. Special issue: young people, border spaces and Revolutionary Imaginations. Children's Geographies 8, 4.

Allen, K., Hollingworth, S., 2013. 'Sticky subjects' or 'cosmopolitan creatives'? Social class, place and urban young people's aspirations for work in the knowledge economy. Urban Studies 50 (3), 499–517.

Amit-Talai, V., Wulff, H. (Eds.), 1995. Youth Cultures: A Cross-cultural Perspective. Routledge, London.

Ansell, N., 2005. Children, Youth and Development. Routledge, London.

Ansell, N., 2009. Childhood and the politics of scale: descaling children's geographies? Children's Geographies 33 (2), 190–209.

Bordonaro, L.I., 2012. Agency does not mean freedom: Cape Verdean street children and the politics of children's agency. Children's Geographies 10 (4), 413–426.

Cahill, C., Katz, C., 2008. Young Americans: geographies at the crossroads. Environment and Planning A 40, 2809–2813.

Evans, B., 2008. Geographies of youth/young people. Geography Compass 2 (5), 1659–1680.

Hansen, K.T., Dalsgaard, A.L., Gough, K.V., Madsen, U.A., Valentin, K., Wildermuth, N., 2008. Youth and the City in the Global South. Indiana University Press, Bloomington.

Holt, L., 2009. Youth/Youth cultures. In: Kitchin, R., Thrift, N. (Eds.), International Encyclopedia of Human Geography. Elsevier, pp. 283–287.

Holt, L., 2011. Geographies of Children, Youth and Families: An International Perspective. Routledge, London.

Hörschelmann, K., van Blerk, L., 2012. Children, Youth and the City. Routledge, London.

Jeffrey, C., 2010. Geographies of children and youth I: eroding maps of life. Progress in Human Geography 34 (4), 496–505.

Jeffrey, C., Dyson, J. (Eds.), 2008. Telling Young Lives: Portraits of Global Youth. Temple University Press, Philadelphia.

Jones, G.A., 2008. Children and development II: 'youth' violence and juvenile justice. Progress in Development Studies 8, 345–348.

Kallio, K.P., Häkli, J. (Eds.), 2013. Children and young people's politics in everyday life. Space and Polity 17 (1), 1–16 (special issue).

Katz, C., 2004. Growing up Global: Economic Restructuring and Children's Everyday Lives. University of Minnesota Press, Minneapolis, MN.

Philo, C., Smith, F.M. (Eds.), 2003. Political geographies of children and young people. Space and Polity 7 (2), 99–115 (special issue).

Reynolds, T., 2013. 'Them and us': 'Black neighbourhoods' as a social capital resource among black youths living in inner-city London. Urban Studies 50 (3), 484–498.

Skelton, T., 2007. Children, young people, UNICEF and participation. Children's Geographies 5 (1–2), 165–181.

Skelton, T., 2010. Taking young people as political actors seriously: opening the borders of political geography. Area 42 (2), 145–151.

Skelton, T., 2013a. Young people's urban im/mobilities: relationality and identity formation. Urban Studies 50 (3), 467–483.

Skelton, T., 2013b. Young people, children, politics and space: a decade of youthful political geography scholarship 2003-13. Space and Polity 17 (1), 123–136.

Skelton, T., Gough, K.V., 2013. Young people's im/mobile urban geographies: introduction. Urban Studies 50 (3), 455–466.

Skelton, T., Valentine, G. (Eds.), 1998. Cool Places: Geographies of Youth Cultures. Routledge, London.

van Blerk, L., Ansell, N., 2007. Alternative care giving in the context of AIDS in southern Africa: complex strategies of care. Journal of International Development 19 (7), 865–884.

Weller, S., 2007. Teenagers' Citizenship: Experiences and Education. Routledge, New York.

Relevant Websites

http://www.tandfonline.com/toc/cchg20/current#.UarG45XUhUQ – Children's Geographies Journal Homepage (accessed 02.06.13.).

http://social.un.org/index/Youth/FAQs.aspx – United Nations, Youth: Social Policy and Development Division, FAQs. (accessed 30.05.13.).

http://www.youtube.com/watch?v=Z19zFlPah-o&hl=en-GB&gl=SG – YouTube 1 Stunt Cycling (accessed 25.07.13.).

http://www.youtube.com/watch?v=xhG4F1ijD1k&hl=en-GB&gl=SG – YouTube 2 Parkour (accessed 25.07.13.).

Youth Joblessness

David NF Bell, Stirling Management School, University of Stirling, IZA and CPC, Stirling, UK
David G Blanchflower, Dartmouth College, Hanover, NH, USA; Stirling Management School, University of Stirling, Stirling, UK; and Peterson Institute for International Economics, IZA, CESifo and NBER

© 2015 Elsevier Ltd. All rights reserved.

Abstract

We present evidence showing that in advanced countries there has been a rise in both youth unemployment, a decline in youth employment, alongside as a rise in underemployment for those youngsters lucky enough to have a job. We also document marked differences across countries in the size of the youth cohort, which in most western countries rose by around 10% between 2000 and 2010 but is then set to decline sharply. We examine youth wages and discuss evidence that suggests that long spells of unemployment create permanent scars.

Across most OECD countries, young people were especially hard hit by the onset of the Great Recession. (The OECD has a major program on youth unemployment; see http://www.oecd.org/youth.htm. See also its Jobs for Youth Review http://www.oecd.org/employment/theoecdjobsforyouthreview.htm with reports on youth labor markets in 16 countries – Australia, Belgium, Canada, Denmark, France, Greece, Japan, Korea, the Netherlands, New Zealand, Norway, Poland, the Slovak Republic, Spain, the United Kingdom, and the United States. See also ILO (2013) and work at the European Commission – 'Youth employment,' http://ec.europa.eu/social/main.jsp?catId=1036.) This was consistent with the predictions of the existing literature, which suggested that youth unemployment rates are more cyclically volatile than adult rates, and so it turned out. The definitive work on youth unemployment is contained in three National Bureau of Economic Research (NBER) volumes that documented the struggles the young experienced during downturns (Freeman and Wise, 1982a; Freeman and Holzer, 1986; Blanchflower and Freeman, 2000). In a series of earlier papers, we have updated this work to examine the youth unemployment problem in the Great Recession across countries (Bell and Blanchflower, 2011a,b,c), as well as other work on overall unemployment (Bell and Blanchflower, 2010a,b). This article builds on that work.

Of note is the fact that during the Great Recession, youth unemployment has been a major issue in the Southern European countries, especially Greece, Italy, Portugal, and Spain (Dolado et al., 2013; Eichhorst and Neder, 2014), but also in France, the UK, and Sweden. It has been much less of a problem in Germany, Austria, and the Netherlands. (For recent work on youth unemployment in France and Germany, see Cahuc et al. (2013); on Germany, see Rinne and Zimmermann (2012); and Bell and Blanchflower (2014) for Greece.) First, we set out the recent facts and then we turn to the findings from the literature. We show that, in advanced countries there has been a rise in both youth unemployment, a decline in youth employment, alongside as a rise in underemployment for those youngsters lucky enough to have a job. In developing countries, youths have generally fared much better. We also document marked differences across countries in the size of the youth cohort, which in most Western countries rose by around 10% between 2000 and 2010, but is then set to decline by around 20% by 2020. In contrast, countries in Eastern Europe saw declines of around a quarter in the number of young people between 2000 and 2010; by 2020, their numbers are expected to have declined by two-fifths. In further contradistinction, the cohort size among North African countries will have risen by a third.

The Recent Evidence

Table 1 presents the most recent data we have available on youth unemployment rates, here defined as being for those under the age of 25. It is apparent, as noted above, that they are especially high in Southern Europe, with over half of the youth workforce unemployed in Spain (54%) and Greece (53%), but also very high in Italy (43%) and Croatia (42%), and averaging 22% in the EU and 23% in the Euro Area. We also present unemployment rates for older adults and in the final column the ratio of the two. The United States has a relatively low youth unemployment rate of 13% well below that of Europe (22%), although the ratio compared to the older age group is about the same – 2.7 and 2.4 – respectively. Interestingly, the four countries with the highest ratios around four – Romania; Sweden, Italy, and the UK – are not those with the highest rates of youth unemployment. The concern is that high relative rates might be a problem when youth compare themselves to adults, given we know from happiness research that relative things matter (Luttmer, 2005). It is instructive that there have been riots in both the UK and Sweden. (In May 2013, there were riots in the suburbs of Stockholm, apparently stoked by long-term unemployment, http://www.theguardian.com/world/2013/may/23/swedish-riots-stockholm and riots in London, which spread to other UK cities in the summer of 2011 http://www.theguardian.com/uk/series/reading-the-riots.) **Figure 1** reports time series of youth–adult unemployment rates for four countries with high ratios – Italy, Sweden, and the UK with the United States plotted for comparison purposes. The extremely high ratio of around seven is especially notable in Italy, along with its subsequent decline, whereas the UK and Sweden saw a steady increase, which was not seen in the United States.

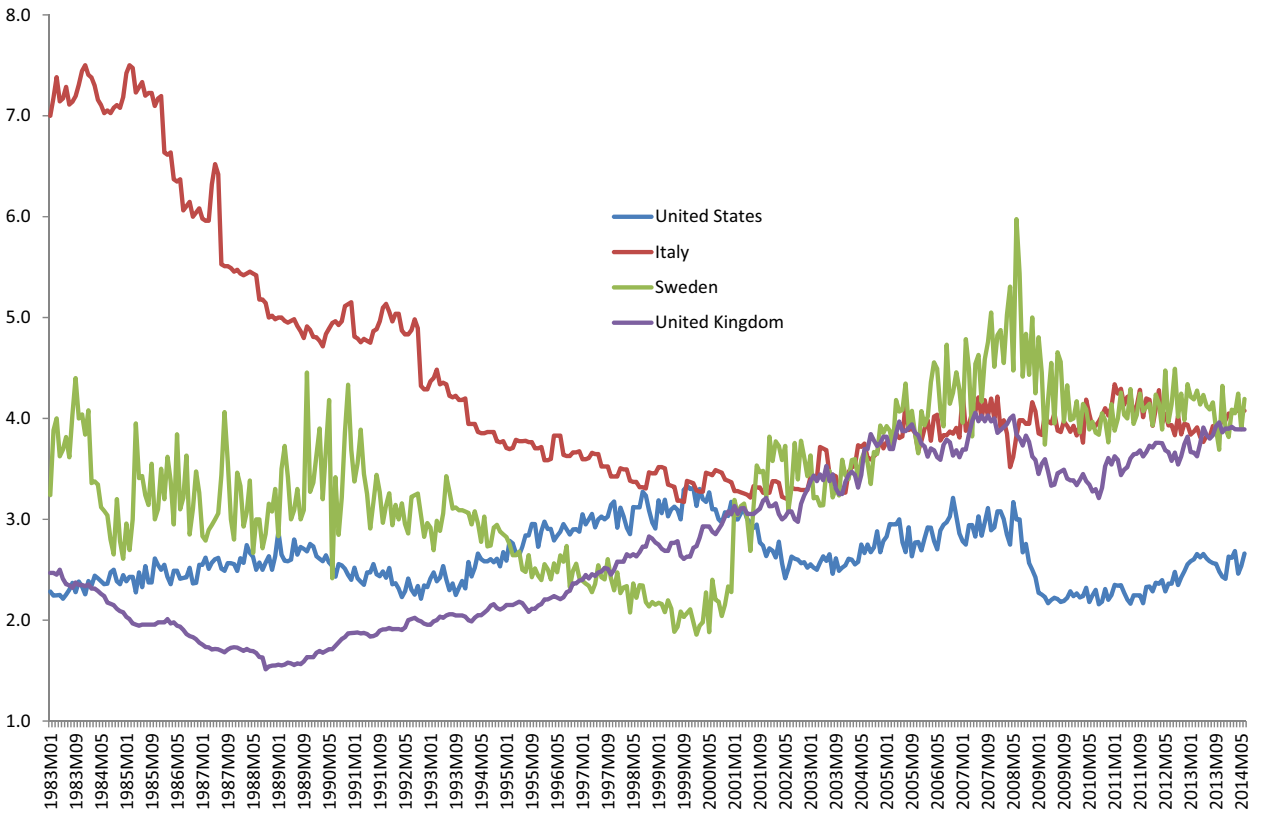

Figure 1 Ratio of youth–adult unemployment rates, 1983–2014 (Italy, Sweden, the United States, and the UK).

Table 2 reports on levels of youth employment to population rates, which can be a more accurate indicator of youth joblessness than unemployment rates, given low levels of youth participation rates in some countries, which are also reported in the table, for 2008 and 2013. The higher the proportion of youngsters in education, the lower will be the participation rate.

It turns out that we get a very similar picture of youth joblessness from the employment rates as we do from unemployment rates. Employment rates for the young fell during the recession in most advanced countries and are lower in 2013 than they were in 2008; the three main exceptions are Sweden, Switzerland, and Germany where they are unchanged. In contrast, employment rates are higher in Chile, Columbia, Israel, and Turkey. Among advanced countries, participation rates of the young are only up in Switzerland (+7ppt); Sweden (+1ppt), suggesting that the young have not rushed to increasingly expensive education as a refuge from joblessness.

So in most European countries, both employment and participation rates are down since the onset of recession for the young and some sharply so. Examples of the differences between 2008 and 2013 rates for some selected countries with participation rates with employment rates in parentheses are as follows as percentage points: Denmark −10 (−12); Canada −4 (−5); the UK −4 (−7); the United States −4 (−5); Germany −2 (0); Finland −5 (−6); Spain −10 (−21); Ireland −13 (−18); France 0 (−2); Portugal −6 (−12); Italy −1 (−6); and Greece −2 (−12). It turns out that the three countries with the highest employment rates are also the ones with the highest participation rates – Iceland, the Netherlands, and Switzerland. The correlation coefficient is 0.96. The three countries with the lowest employment rates – Italy, South Africa, and Greece are close to the bottom in the ranking of participation rates, ranking 33, 38, and 34, respectively. As a consequence, unemployment rates are also up in the advanced countries but less so in the developing world.

A phenomenon that we have also observed in the recession is that not only those without jobs, the jobless, have been impacted, but also the youngsters who have held jobs. Table 3 shows the rise in the share of involuntary part-time employment in total employment of the young, which has more than doubled in many countries since including Greece, Portugal, the Netherlands, the UK, the United States, and especially Spain that has seen nearly a threefold increase. Young workers are disproportionately underemployed (Bell and Blanchflower, 2011c).

These rises in unemployment, and falls in employment and participation rates that have occurred in many countries have come at a time of large youth cohorts. So, a rise in the supply of young people may exacerbate further the evidence found by Kahn (2011) that graduating from college in a recession permanently lowers a cohort's earnings. Table 4 reports on such changes in cohort size for a number of countries since 2000 and reports the change in the size compared with that in 2000 over the period 2004–2020. We can calculate 16–24 cohort size in 2020 by extrapolating from the number of 6- to 14-year-olds 10 years earlier, assuming no further immigration, which is a stretch of course. It is apparent that some countries had big increases in the size of their youth cohorts between 2000 and

Table 1 Youth and adult unemployment rates, 2014 ranked by youth unemployment rates

	<25 years Unemployment rate (1)	*25–74 years Unemployment rate (2)*	*Ratio (1)/(2)*
Spain	53.5	22.4	2.4
Greece	53.1	25.4	2.1
Italy	43.7	10.5	4.2
Croatia	41.7	13.7	3.0
Cyprus	35.4	13.2	2.7
Portugal	33.5	12.5	2.7
Slovakia	32.3	12.2	2.6
Romania	24.4	5.7	4.3
Sweden	23.9	5.7	4.2
Poland	23.6	8.2	2.9
Belgium	23.2	7.2	3.2
Ireland	23.2	10.7	2.2
Bulgaria	23.2	10.9	2.1
Euro area (18)	23.1	10.4	2.2
France	22.4	8.9	2.5
Slovenia	22.3	9.3	2.4
European Union (28)	22.0	9.0	2.4
Hungary	20.5	7.1	2.9
Finland	20.2	7.2	2.8
Latvia	19.7	10.6	1.9
The United Kingdom	17.9	4.6	3.9
Lithuania	17.7	9.8	1.8
Estonia	17.4	6.3	2.8
Czech Republic	16.7	5.3	3.2
Turkey	16.2	7.7	2.1
Luxembourg	14.8	5.7	2.6
Malta	13.3	4.3	3.1
The United States	13.3	5.0	2.7
Denmark	12.6	5.5	2.3
The Netherlands	10.5	6.1	1.7
Iceland	9.4	4.2	2.2
Austria	9.0	4.3	2.1
Germany	7.8	4.8	1.6
Norway	7.3	2.5	2.9

Notes: data are for May for Estonia, Greece, Hungary, and Norway; for April for the UK and Turkey; and March for Latvia and Romania.
Source: Eurostat.

2008 when the Great Recession started, including the UK, Sweden, and the United States, whose cohort size was 10% bigger. In contrast, cohort size was sharply smaller in Spain (−30%), Italy (−21%), and Greece (−25%). By 2020, youth cohort size will be at least 10% smaller in most countries except the United States, where it will continue to grow. By 2020, the youth populations of North Africa and Latin America will have grown by over a quarter, whereas that of Western Europe will have shrunk by 10%. The youth population in most East European countries, including Poland and Russia, will likely fall by around two-fifths over a 20-year period from 2000 to 2020. The Arab Spring occurred in countries that had growing youth cohorts and high youth unemployment rates.

We now turn to examine five questions about youth unemployment from the literature. The concern is that spells of unemployment experienced during the Great Recession will create permanent scars for a *Lost Generation*.

Table 2 Youth ages 15–24 employment and participation rates (%)

	Employment rates		*Participation rates*	
	2008	*2013*	*2008*	*2013*
Australia	65	59	71	67
Austria	56	54	61	59
Belgium	27	24	33	31
Canada	60	55	68	64
Chile	27	30	34	36
Columbia	37	44	48	55
Czech Republic	28	26	31	32
Denmark	66	54	72	62
Estonia	37	33	41	41
Finland	46	40	55	50
France	31	29	38	38
Germany	47	47	53	51
Greece	24	12	30	28
Hungary	20	20	25	27
Iceland	72	70	79	79
Ireland	47	29	54	41
Italy	24	18	31	30
Japan	41	40	45	43
Korea	24	24	26	27
Latvia	37	30	43	39
Luxembourg	24	22	29	26
Mexico	44	42	48	47
The Netherlands	67	62	71	70
New Zealand	56	50	63	60
Norway	58	52	63	58
Poland	27	24	33	33
Portugal	34	22	41	35
Russia	36	34	42	40
Slovak Republic	26	20	32	31
Slovenia	38	27	43	34
South Africa	17	13	31	26
Spain	40	19	52	42
Sweden	42	42	53	54
Switzerland	62	62	67	68
Turkey	30	32	38	40
The United Kingdom	56	49	66	62
The United States	51	46	59	55
OECD countries	43	40	49	47

Notes: employment rate is employment/population; participation rate is (unemployed + employed)/population or can be defined as workforce/population.
Source: OECD.

The Literature on Youth Joblessness

There is a large literature on youth unemployment, drawing on the experience of recent decades. In this section, we examine this literature and attempt to answer eight vital questions regarding the workings of the youth labor market and the consequences of failing to successfully make the transition from school to work.

Is Youth Unemployment Cyclical or Structural?

Clark and Summers (1982), in their classic study of the dynamics of youth joblessness, argue that the problem of teenage unemployment arises from a shortage of jobs. "Aggregate demand has a potent impact on the job prospects and market experience of teenagers" (1982: p. 230). Freeman and Wise (1982b), for example, found in their study of youth

Table 3 Share of involuntary part-time employment in total employment, 15- to 24-year-olds

	2008	*2010*	*2012*
Australia	13	16	16
Austria	2	2	1
Belgium	6	5	4
Canada	7	10	10
Czech Republic	0	1	1
Denmark	3	5	4
Finland	7	9	8
France	7	8	8
Germany	4	3	2
Greece	5	8	11
Hungary	1	2	3
Ireland	2	10	13
Italy	9	13	17
Japan	7	10	7
Luxembourg	1	1	3
The Netherlands	2	2	4
New Zealand	9	11	11
Norway	1	2	2
Poland	1	2	3
Portugal	4	6	8
Slovak Republic	1	4	4
Spain	8	14	23
Sweden	9	11	11
The United Kingdom	4	7	9
The United States	3	5	6
G7 countries	4	7	6
OECD countries	3	5	5

Note: The figure for UK is for 2007 and not 2008
Source: OECD.

joblessness in the 1970s that it was concentrated, by and large, among a small group who lacked work for extended periods of time. Over half of the male-teenage unemployment they examined was among those who were out of work for over 6 months, a group constituting less than 10% of the youth labor force and only 7% of the youth population. Freeman and Wise reported that the youths who make up the relatively small group that was chronically without work had distinctive characteristics. They were disproportionately black, disproportionately high-school dropouts, and disproportionately residents of deprived areas.

Blanchflower and Freeman (2000) identified one basic pattern in the job market for young workers: the disproportionately large response of youth employment or unemployment to changes in overall unemployment. They argued that the sensitivity of youth employment and unemployment to the overall rate of unemployment dominate sizable demographic and structural changes favorable to youth in determining how youths fare in the job market. OECD (2008) confirmed this conclusion "Youth unemployment rates are more sensitive to business-cycle conditions than the adult unemployment rate and this high-sensitivity tends to decline progressively with age." Consistent with this argument, there is evidence that young people do well in booms. Freeman and Rodgers (2000) analyzed the 1990s boom in the United States and found that it substantially improved the position of noncollege-educated young men, especially young African-Americans who are the most disadvantaged and troubled group in the United States. Young men in tight labor markets experienced a substantial boost in both employment and earnings. Adult men had no gains and their earnings barely changed even in areas where unemployment rates were below 4%.

There is also evidence of state dependence in youth unemployment. In this context, state dependence means that the experience of one spell of unemployment of itself increases the probability of further spells. This may be because on the demand side, employers are less willing to hire those with a record of unemployment or that, on the supply side, the experience of unemployment discourages individuals from job search activity.

How Has Changed School Enrollment Impacted on Youth Unemployment?

Most OECD countries have tried to extend the duration of formal schooling in recent decades. This delays entrance into full-time employment, but not necessarily part-time work. Table 5

Table 4 Cohort size of those aged 16–24 based on current population, compared with size of cohort in 2000 (%)

	2004	*2006*	*2008*	*2010*	*2012*	*2014*	*2016*	*2018*	*2020*
Australia	1	2	2	0	−1	−4	−8	−11	−13
France	0	0	−1	−2	−2	−3	−3	−2	0
Germany	3	4	3	0	−4	−9	−13	−16	−19
Greece	−10	−19	−25	−29	−32	−33	−34	−35	−35
Italy	−12	−18	−21	−23	−25	−27	−29	−30	−31
Poland	2	−4	−9	−16	−21	−26	−31	−36	−39
Russia	9	12	7	−3	−14	−27	−35	−38	−38
Spain	−14	−24	−30	−34	−38	−41	−43	−43	−43
Sweden	3	7	10	11	9	2	−5	−11	−14
The United Kingdom	6	9	10	9	6	2	−3	−8	−11
The United States	5	8	10	11	11	9	7	4	2
Western Europe	1	1	1	−1	−3	−6	−8	−9	−10
North Africa	13	18	20	22	23	25	27	29	32
Latin America	8	13	16	18	20	22	24	24	25
Southeast Asia	5	7	9	10	12	13	14	14	15

Source: Eurostat.

Table 5 Trends in enrollment rates – students as a percentage of the population of the age group

	Ages 15–19		*Ages 20–29*	
	2008	*2010*	*2008*	*2010*
Austria	79	78	22	25
Belgium	92	93	29	30
Chile	74	75	21	25
Czech Republic	90	90	21	24
Denmark	84	85	37	38
Estonia	84	87	26	28
Finland	87	87	43	42
France	84	84	19	20
Germany	89	89	28	31
Greece	83	83	29	40
Hungary	89	92	25	25
Iceland	84	88	35	38
Ireland	90	96	18	21
Israel	64	65	21	22
Italy	82	83	21	21
Korea	89	86	28	30
Luxembourg	75	77	10	13
Mexico	52	54	11	12
The Netherlands	90	91	29	30
New Zealand	74	79	29	30
Norway	87	86	29	29
Poland	93	93	30	30
Portugal	81	86	23	24
Slovak Republic	85	85	19	21
Slovenia	91	92	33	34
Spain	81	84	21	24
Sweden	86	86	33	36
Switzerland	85	85	23	24
Turkey	46	56	13	20
The United Kingdom	73	77	17	18
The United States	81	82	23	26

Source: OECD Education at a Glance 2012 Highlights.

shows there are marked variations in the proportion of young people who are students. This ranges from a high of 96% in Ireland for 15- to 19-year-olds to a low of 54% in Mexico, and a high of 42% for 20- to 29-year-olds in Finland compared to a low of 13% in Luxembourg. These levels reflect a belief in most industrial countries that greater investment in human capital is required to maintain competitiveness, but they may also reflect lowering costs of education as average family sizes fall. Yet again, it may be a defensive strategy, with enrollment increasing when the labor market deteriorates. It is especially notable in the Table 5 that the rise in the proportions of young people who are students as a response to youth joblessness especially in Greece among 20- to 29-year-olds, that Bell and Blanchflower (2014) found were especially impacted by a lack of jobs. The rise in attendance rates across countries is much more marked among 20- to 29-year-olds, partly because of the already high rates existing for the younger group. Rates are up by at least 3% points, between 2008 and 2010, for 20- to 29-year-olds in Austria, Germany, Greece, Iceland, Ireland, Luxembourg, Spain, Sweden, and the United States.

Increased schooling affects youth unemployment in two ways. First, so long as increased educational attainment is increasing the 'employability' of the young, it should increase the success rate of their job search activity. On the other hand, delayed job market entry caused by additional schooling may reduce the average 'employability' of those that are economically active, since those selected into additional schooling are likely to be more able. This effect may be difficult to disentangle within aggregate statistics, since some of those taking additional schooling may also be economically active and therefore classified as either employed or unemployed.

Have Youth Wages Been Too High, so Increasing Youth Unemployment?

There appears to have been a secular decline from the early part of this century in relative youth wages, some recovery around 2007–08 followed by further decline during the recession. Employment among older workers has increased during the recession, even though these data suggest that their relative wages have been increasing compared with those of young people in the UK and the United States. There is no prima facie evidence of younger workers pricing themselves out of jobs.

To examine this further, we looked at microwage data for the United States and the UK to see how the relative wages of the young had changed over the recession. For the UK we took Labour Force Survey (LFS), while for the United States we have data from the Merged Outgoing Rotation Group (MORG) files of the Current Population Survey (CPS). For both countries, we have hourly earnings data since 2001 and can compare the earnings of youngsters – we present data for 16 to 19, 20 to 24, and 25- to 29-year-olds – to earnings of prime-age workers aged 40–44 for 2001, 2004, 2008, and 2013. Interestingly, the ratios and the changes are broadly similar across the two countries.

	2001	*2004*	*2008*	*2013*
The United Kingdom				
16–19/40–44	41%	41%	37%	36%
20–24/40–44	61%	60%	53%	52%
25–29/40–44	84%	84%	74%	70%
The United States				
16–19/40–44	41%	40%	38%	37%
20–24/40–44	56%	54%	53%	49%
25–29/40–44	80%	77%	76%	73%

We find a big decline in the ratio of youth in each of the three categories post-2004. In the United States, the 16- to 19-year-olds earn 37% (36%) compared to 41% (41%) in 2001, whereas the 25- to 29-year-olds had a bigger absolute decline from 80% (84%) in 2001 to 73% (70%) with the UK numbers in parentheses. For both the UK and the United States, it appears that youth wages have been flexible downward in the face of declining labor demand.

Do Active Labor Market Programs Work? (Hint: No!)

In response to the substantial increases in youth unemployment during the 1980s and early 1990s, there was a significant increase in policy initiatives targeted at unemployment in

general and youth unemployment in particular. These culminated in the OECD Jobs Study (1994), which argued that policies to improve the working of the labor market were critical for reducing high levels of youth unemployment. It suggested that:

> A progressive shift of resources is needed from passive income support to active measures. Active labor market policies improve access to the labor market and jobs; develop job-related skills; and promote more efficient labor markets.
>
> OECD Jobs Study, 1994

With OECD encouragement, many member countries introduced a range of active labor market policies (ALMPs). These were extensively evaluated to determine their labor market effects. Card et al. (2009) carry out a meta-analysis of ALMPs. They examined 199 programs drawn from 97 studies over the period 1995–2007. The programs themselves are heterogeneous covering classroom and work experience, job search assistance, subsidized public and private employment programs. In the short run, the number of programs with a significantly positive outcome only exceeds those with a significantly negative outcome by a small margin. Over longer time horizons, the balance is more in favor of positive outcomes.

There are also significant differences between countries in the types of programs, which succeed or fail. There is no upward trend over time in positive outcomes. The evaluations contain little information on costs, making it difficult to establish how affordable large-scale expansion of successful programs might be. The Card et al. study supports Grubb and Martin's (OECD, 2001) conclusion that, "one of the most disappointing conclusions from the evaluation literature is that almost all evaluations show that special measures are not effective for disadvantaged youths." There is very little evidence from around the world that ALMPs work well in recessions.

What Impact Does Youth Unemployment Have?

There is a wealth of literature showing that unemployment is a stressful life event that directly reduces individual well-being. This applies not just to the unemployed, but also may affect the employed due to increased anxiety over job security. Unemployment increases susceptibility to malnutrition, illness, mental stress, and loss of self-esteem, and increases the risk of depression. The unemployed also appear to be at higher risk of committing suicide and of poor physical health outcomes later in life. Low levels of happiness and health have predictive power for subsequent outcomes (Blanchflower, 2009).

Youth unemployment also has particularly adverse social impacts. Higher unemployment is associated with increases in burglaries, thefts, and drug offenses. Unemployment is often part of the cycle where involvement in crime reduces subsequent employment prospects and consequently increases the probability of participating in crime. There is new evidence that even young people who choose to go to college or university are hurt if they enter the labor market during a recession. Kahn (2010) has shown that graduating from college during a recession has large, negative, and *persistent* effects on wages. Lifetime earnings are substantially lower than they would have been if the graduate had entered the labor market in good times. Furthermore, cohorts who graduate in worse national economies tend to end up in lower-level occupations. Giuliano and Spilimbergo (2014) suggest that the period of early adulthood (between 18 and 25) seems to be the age range during which people are more sensitive to macroeconomic conditions. They find that being exposed to a recession before age 17 or after age 25 has no impact on beliefs about life chances. However, youngsters growing up during recessions tend to believe that success in life depends more on luck than on effort; they support more government redistribution, but have less confidence in public institutions. Recessions seem to adversely affect youngsters' beliefs.

Further, unemployment while young, especially of long duration, appears to be associated with permanent scars rather than the temporary blemishes that result for older workers (Ellwood, 1982). The majority of older workers get over spells of unemployment reasonably quickly while youngsters do not as they struggle to find a toehold in the labor market. The scarring effect of youth unemployment has two components: first, for the young, a spell of unemployment does not end with that spell; it raises the probability of being unemployed in later years. Second, early spells of unemployment also carry a wage penalty. These effects are much larger than for older people experiencing unemployment. Mroz and Savage (2006) find that a 6-month spell of unemployment at age 22 results in an 8% lower wage at 23 and even at ages 30 and 31, wages are 2–3% lower than they otherwise would have been.

In Bell and Blanchflower (2011a), we found that early adulthood unemployment in the UK creates long-lasting scars, which affect both wages and happiness. Our results showed significant effects at the age of 50 from early adulthood unemployment, but none from recent unemployment experience. Bell and Blanchflower (2011b) extended that work and found evidence that spells of youth unemployment have harmful impacts on a number of outcomes, including health and job satisfaction, many years later. A spell of unemployment when young continues to have a negative impact in later life. Our evidence is consistent with the view that negative impacts on wages, life satisfaction, health and job satisfaction may persist well into the future.

Conclusions

The young have taken it on the nose in the Great Recession in most European countries and especially so in Southern Europe where youth joblessness has soared. This has not been true everywhere though; the young have done well both absolutely and relatively in Germany and Austria where youth unemployment rates have remained in single digits. Relative rates have been high in a number of other countries including the UK and Sweden where we have observed serious social disorder and riots. The consequence of these high rates of joblessness is that many youngsters have not been able to strike out on their own. This has meant that there has been a marked rise since 2008 in the proportion of youngsters who continue to live with their parents even into their late-twenties.

Youth unemployment rates are highly cyclical so the likelihood is that as economies recover on average, the young will be the ones who benefit the most having lost the most in the downturn. But there will be losers. We have also found that the young are faced with a double whammy, when they have jobs they tend to be underemployed. They are also more likely to be in a part-time job when they would like a full-time job. We have also found evidence that their earnings relative to adults has also fallen since the Great Recession started.

The biggest concern is that there is evidence that permanent scarring can occur when young people experience long spells of unemployment and are unable to make a foothold in the labor market. The concern is that these young people may end up as a lost generation and then it is too late.

See also: Career Development: Multilevel Perspective; Data Bases and Statistical Systems: Labor and Work; Disposable Youth in America in the Age of Neoliberalism; Employment and Labor, Regulation of; Informal Sector; Labor Supply; Unemployment and Mental Health; Youth, Geography of.

Bibliography

Bell, D.N.F., Blanchflower, D.G., 2014. Youth unemployment in Greece: meeting the challenge. Paper presented at Peterson Institute conference on Greek Unemployment, New York 3rd/4th April 2014.

Bell, D.N.F., Blanchflower, D.G., 2011a. Young people and the Great Recession. Oxford Review of Economic Policy 27, 241–267.

Bell, D.N.F., Blanchflower, D.G., 2011b. Youth unemployment in Europe and the United States. Nordic Economic Policy Review 1, 11–38.

Bell, D.N.F., Blanchflower, D.G., January 2011c. Youth underemployment in the UK in the Great Recession. National Institute Economic Review R1–R11.

Bell, D.N.F., Blanchflower, D.G., March 2010a. Recession and unemployment in the OECD. CESifo Forum 1, 14–22.

Bell, D.N.F., Blanchflower, D.G., October 2010b. UK unemployment in the Great Recession. National Institute Economic Review 214, R3–R25.

Blanchflower, D.G., 2009. International evidence on well-being. In: Krueger, Alan B. (Ed.), Measuring the Subjective Well-being of Nations: National Accounts of Time-Use and Well-being. NBER and University of Chicago Press.

Blanchflower, D.G., Freeman, R.B. (Eds.), 2000. Youth Employment and Joblessness in Advanced Countries. University of Chicago Press and NBER, Chicago, IL.

Cahuc, P., Carcillo, S., Rinne, U., Zimmermann, K.F., 2013. Youth unemployment in old Europe: the polar cases of France and Germany. IZA Journal of European Labor Studies 2 (18), 2–18.

Card, D., Kluve, J., Weber, A., 2009. Active labor market policy evaluations: a meta-analysis. Economic Journal 120, F452–F477.

Clark, A.E., Oswald, A.J., 1994. Unhappiness and unemployment. Economic Journal 104 (424), 648–659.

Clark, K.B., Summers, L.H., 1982. The dynamics of youth unemployment. In: Freeman, R.B., Wise, D.A. (Eds.), The Youth Labor Market Problem: Its Nature, Causes, and Consequences. University of Chicago Press and NBER, Chicago, IL.

Congressional Budget Office, November 2004. What Is Happening to Youth Employment Rates?.

Dolado, J.J., Felgueroso, F., Fuentes, A., Jansen, M., 2013. Youth Labour Market Performance in Spain and its Determinants: A Micro-level Perspective. OECD Economics Department Working Paper #1039, Paris, France.

Eichhorst, W., Neder, F., 2014. Youth Unemployment in Mediterranean Countries. IZA Policy Paper #80.

Ellwood, D., 1982. Teenage unemployment: permanent scars or temporary blemishes? In: Freeman, Richard B., Wise, David A. (Eds.), The Youth Labor Market Problem: Its Nature, Causes and Consequences. University of Chicago Press, Chicago, pp. 349–390.

Freeman, R.B., Holzer, H.J. (Eds.), 1986. The Black Youth Employment Crisis. University of Chicago Press and NBER, Chicago, IL.

Freeman, R.B., Rogers, J., 2000. Area economic conditions and the labour market outcomes of young men in the 1990s expansion. In: Cherry, R., Rodgers, W.M. (Eds.), Prosperity for All: The Economic Boom and African Americans. Russell Sage Foundation, New York.

Freeman, R.B., Wise, D.A. (Eds.), 1982a. The Youth Labor Market Problem: Its Nature, Causes, and Consequences. University of Chicago Press and NBER, Chicago, IL.

Freeman, R.B., Wise, D.A., 1982b. The youth labor market problem: its nature causes and consequences. In: Freeman, R.B., Wise, D.A. (Eds.), The Youth Labor Market Problem: Its Nature, Causes, and Consequences. University of Chicago Press and NBER, Chicago, IL.

Grubb, D., Martin, J., 2001. What Works and for Whom: A Review of OECD Countries' Experiences with Active Labour Market Policies. OECD Office of Labour Market Policy Evaluation, Working Paper 2001:14.

Giuliano, P., Spilimbergo, A., 2014. Growing up in a recession. Review of Economic Studies 81 (2), 787–817.

ILO, 2013. Global Employment Trends for Youth: A Generation at Risk. Geneva. http://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/documents/publication/wcms_212423.pdf.

Kahn, L.B., April 2010. The long-term labor market consequences of graduating from college in a bad economy. Labour Economics 17 (2), 303–316.

Luttmer, E., 2005. Neighbors as negatives: relative earnings and well-being. Quarterly Journal of Economics 120, 963–1002.

Mroz, T.A., Savage, T.H., 2006. The long-term effects of youth unemployment. Journal of Human Resources XLI, 259–293.

OECD, 1994. The OECD Jobs Study: Facts, Analysis, Strategies. OECD, Paris.

OECD, 2008. Off to a Good Start? Youth Labour Market Transitions in OECD Countries, OECD Employment Outlook, 2008, Paris.

Rinne, U., Zimmermann, K.F., 2012. Another economic miracle? The German labor market and the great recession. IZA Journal of Labor Policy 1 (3), 1–21.

Youth Movements

Ken Roberts, University of Liverpool, Liverpool, England, UK

© 2015 Elsevier Ltd. All rights reserved.
This article is a revision of the previous edition article by M.M. Braungart, volume 24, pp. 16668–16671, © 2001, Elsevier Ltd.

Abstract

These are amorphous, multicentered movements among young people who regard themselves, and are regarded by others, as an upcoming, usurping generation. Such movements are likely to form in periods of major historical change. This article discusses the main youth movements that have been formed during and since the midtwentieth century: the student movements of the 1960s and the 1970s; movements that challenged colonial and neocolonial rulers in less-developed countries; young people's involvement in the revolutions that saw the end of communism in East-Central and South-East Europe in 1989, and in the subsequent 'color revolutions' in some former Soviet republics; the still evolving movements that resist neoliberalism and its impacts (*Anti-Globalisation*, the *Indignados*, *Occupy*); and the *Arab Spring* of 2011. It is concluded that youth movements are most successful in mobilizing dissent while they lack clear aims and leaders; that while they retain the support of security services, incumbent political elites can usually resist these challenges; and that youth movements are liable to fragment when they begin to engage in conventional democratic politics.

Definitions and Theoretical Approaches

Before 1939, a youth movements article would have been mainly about movements led by adults and aimed at young people – Boy Scouts, Girl Guides, the Boys Brigade, and the numerous sport and recreation clubs that had young people as a main if not the sole target group (see Evans, 1965). By 1939, over 1000 colleges in North America were training leaders for movements such as these (Pangburn, 1940). These movements were adjuncts to the school systems that had been created in urbanized industrial societies. The main targets were 'adolescents' who were experiencing a life stage said to be characterized by inner stress and turmoil (Hall, 1904). Young people were considered 'at risk' when beyond the control and protection of families and schools, that is, during after-school hours and holidays, and also between leaving school and settling in adult occupations and marriages. These youth movements are still very much alive, but they have been joined and overshadowed by movements created and led by young people themselves. To avoid confusion, the older movements would now be better described as 'youth services.'

Social movements differ from organizations and associations. Movements are relatively amorphous: they arise in different places more or less simultaneously. There are no official leaders, although spokespersons may be recognized, and links between groups may be formed. Social movements are products of 'society,' formed 'bottom-up' rather than 'top-down' at the behest of governments. 'Youth' has to be defined as a relationship (with elders), and since the midtwentieth century, 'youth movements' have been those in which substantial numbers of participants have seen themselves, and have been regarded by others, as an upcoming, usurping generation.

In Europe, teenagers began to be aware of themselves as a distinct age group from the 1870s onward (Savage, 2007), and youth cultures (using that term) have been formed since the 1920s (Fowler, 2008). These are not youth movements, but they can become the seedbed for movements with broad social and political objectives. Such movements can be, but will not necessarily be formed in any place where the age group congregates: armed forces and universities have been common breeding grounds. The original *Young Turks* were a prototype. They were young army officers who led a rebellion against the Ottoman monarchy in 1908 and ruled Turkey until the end of the First World War. Subsequently all groups of young upstarts have been likely to be labeled 'Young Turks.' In the 1920s and the 1930s, young people were mobilized in Europe's then ascendant communist and fascist movements. Reflecting on this experience, Karl Mannheim (1952) argued that periods of major historical change were likely to see the birth of new political generations who were unable to relate to the policies and worldviews of existing political elites. Since then, the formation of successive youth movements has been studied more systematically, starting with the student movements of the 1960s up to, so far, the Arab Spring and the Occupy movements that erupted in 2011.

Findings

Student Movements in the 1960s

The student movements of the 1960s and the 1970s have been models for future young and older observers of young people in the West. They have structured expectations about the likely features of successor movements, and in doing so have tended to mislead. The student movements of the 1960s and the 1970s had origins in the civil rights protests that began in the United States in the 1950s and the peace (anti-H bomb) marches and demonstrations in Europe. During the 1960s, the protests spread across both continents, and more thinly in other parts of the world, focusing on opposition to American involvement in the Vietnam War, but the movements adopted a wider range of new issues – second wave feminism, environmentalism, and global justice. All these were embraced within a call for greater personal freedom. The young protesters were unimpressed by the futures that were laid out before them: careers in dull bureaucracies and family lives in 'little boxes.' Political messages were blended with anthems from the Rolling

Stones, Bob Dylan, Joan Baez, hippie dress, the recreational use of drugs, and the sexual liberation that accompanied the marketing of the contraceptive pill from 1961. In Paris, in 1968, student protests drew in trade unions and the French government felt threatened (very briefly). This wave of protests was the work of what Ronald Inglehart (1977) called the postscarcity generation. The postwar baby boomers were the first young people to grow up with no experience of even the threats of poverty and unemployment. Jobs and rising standards of living were taken for granted, and as the generation grew to political maturity and became majorities of voters and politicians, 'quality of life' issues were pushed up political agendas (Reich, 1972; Roszak, 1970). Only minorities of students were involved in any protests during the 1960s and the 1970s (Blackstone et al., 1970) but there are senses in which the activists spoke for their entire generation. As they grew older, many aspects of the counterculture of the 1960s became mainstream.

Social science learned a great deal from studies of this particular generation. Their biographies confirmed Karl Mannheim's proposition that political events experienced during youth, when a cohort is first becoming politically aware, make a profound and lasting impression (Schuman and Corning, 2000). Afterward individuals know whose and which sides they are on and can respond to new issues accordingly. However, we have also learned that basic political orientations can remain fluid well beyond the normal student years (Burnett, 2000). Also, we now know that youthful ideals may be retained but are not necessarily acted on throughout adulthood. Many who rejected materialism in the 1960s built lucrative careers in business, management, and the professions (Fendrich, 1977), and this generation became the adult voters and politicians who ushered in the neoliberal era. Those who pursued careers in politics often made U-turns in response to opportunities and circumstances encountered later on. These were comparable turnarounds to those of the young fascists and communists of the 1930s who helped to build the post-1945 social democracies in Western Europe.

Anticolonial and Neocolonial Movements

The student protesters and hippies of the 1960s may be the West's favored model, but in what were then becoming known as third world, developing or underdeveloped societies, youth movements intent on changing these countries were simultaneously being formed in locations other than city streets and university campuses, and in some cases were seizing power. These movements sought independence, first from colonial powers, then afterward from 'neocolonial' elites to whom power was passed without much else changing. The young activists were often young army officers. They included Gamal Abdel Nasser (1918–70) who was among the young Egyptian army officers who overthrew King Farouk in 1952, Muammar Gaddafi (1942–2011) who was also part of a group of young army officers when they overthrew Libya's King Idris in 1969, and Saddam Hussein (1937–2006) who was part of the group that seized power in Iraq in the 1968 rebellion. Armies as well as universities can nurture new political generations. The original *Young Turks* were from Ottoman Turkey's army.

In other colonies and excolonies, in the decades during and following the Second World War, new generations were formed in political parties that eventually propelled their leaders into power. Examples include a young Robert Mugabe (1924–till date) who became Secretary-General of the Zimbabwe African National Union in the 1960s (in what was then Southern Rhodesia) and was elected to office in the first elections held under universal adult suffrage in 1980. Nelson Mandela (1918–2013) became a founder member of the African National Congress Youth League in 1944, and after 27 years in prison, became the first elected president of postapartheid South Africa in 1994. This particular new political generation, formed when the activists were all young, had to wait a long time before achieving political power.

In the twenty-first century, some countries' leaders waged war against others who would have been regarded as kindred radical spirits in the 1960s. Each side would no doubt argue that the other had changed beyond recognition since then.

Revolutions in Eastern Europe

Two political generations triumphed in 1989. For several years before then, young people who were living under communism had been able to signal support for change by playing Western popular music and wearing Western fashions (Pilkington, 1994). In 1989, they demonstrated on the streets of Poland, the German Democratic Republic, and Czechoslovakia. They celebrated the outcome of the June 4 election in Poland, then iconically on November 9 when the Berlin Wall was opened (by the East German authorities). However, these revolutions were really due to the efforts of an older political generation whose formative political experiences were during and around the time of the Prague Spring in 1968 when attempted reforms in Czechoslovakia were crushed by Soviet tanks. Dissent went underground but was not extinguished, and in 1980, a free trade union called *Solidarity* was formed in Poland, which refused to disband or even go underground despite the imposition of martial law and the imprisonment of its leaders. A key event was the appointment of Mikhail Gorbachev as head of the Soviet regime in 1985. He had once been another young person who had recognized imperfections and been frustrated by the Soviet version of communism. He wanted to reform the system, and his key policies were *glasnost* (freedom of expression) and *perestroika* (reform). He informed communist rulers in East-Central Europe that the Soviet army would not intervene should they lose the support of their people. This warning was intended to reinvigorate the communist systems. In the event, Gorbachev's measures did not reform communism but led to its collapse. Poland's rulers were obliged to hold round table talks with *Solidarity* and agree to free elections. The outcome on June 4 (the true history-making date in 1989) surprised everyone. *Solidarity* won all but one of the seats that were up for election. *Solidarity* was surprised. Poland's communists were devastated and faded into the background before reorganizing themselves as social democrats the following autumn. By then, Poland had a postcommunist government. The possibility of change had been demonstrated. Hungary accelerated its liberalization program and opened its border with Austria. This led to the Trabant exodus. East Germans motored through Czechoslovakia into Hungary, then

through Austria and into West Germany. During September, crowds began marching in Leipzig and other East German cities. During the autumn, crowds assembled and remained in Prague's Wenceslas Square. On December 25, the Ceausescus were summarily executed in Bucharest and communism had ended throughout East-Central and Southeast Europe except in Bulgaria (which held its first free election in June 1990) and Yugoslavia where change was more protracted (see Roberts, 2012a). None of the former communist countries' new leaders were from the youth who had taken to the streets in 1989. *Solidarity*'s main spokesperson and the country's first postcommunist present was a middle-aged electrician, Lech Walesa (1943–). In Czechoslovakia, the spokesperson for *Civic Forum* and the first elected president was an aging playwright, Vaclav Havel (1936–2011).

There was no immediate blossoming of civil societies in Eastern Europe in the early 1990s. The young people who had taken to the streets and squares returned to their homes, schools, and jobs (or unemployment). They did not know it, but in this respect their behavior would prove prototypical for a new pattern of youth political engagement, and they had demonstrated to themselves and to others that change was possible. In the countries that eventually became European Union members, it was possible for the new generation to be elected into office, but this was not the case in most of the new independent states that were created when the Soviet Union was dissolved in 1991. In most of these countries, the communist rulers simply rebranded themselves as nationalists and democrats. New recruits into the political elite in Dneipropetrovsk (a major city in Ukraine) tended to be from the Komsomol (the communist party youth organization) rather than rock 'n roll (Zhuk, 2010). However, the events of 1989 had demonstrated to young people elsewhere that change was possible. Young people in Belgrade stormed the parliament building in 2000 and hastened the fall of President Milosevic. Then came the 'color revolutions': the Rose Revolution in Georgia in 2003, the Orange Revolution in Ukraine in 2004–05, and the Tulip Revolution in Kyrgyzstan in 2005 (see Beachain and Polese, 2010; Collin, 2007). Once again, immediately after these apparently seismic events, the crowds of young people quit the streets and melted away. With hindsight, it became clear that their achievements were better described as regime change rather than true revolutions.

Resisting Neoliberalism

A series of new generations has been announced since the student movements of the 1960s and the 1970s subsided: Generation Y, the ecstasy generation, the Internet generation, and so on. These have not been true generations in which youth have mobilized in movements to change their societies. They have merely headlined new features of the youth life stage. The first true successor generation to the student movements is still in formation and it is being formed in opposition to the neoliberal policies that have spread globally since the 1980s, promoted and enforced by transnational organizations, mainly the International Monetary Fund (IMF), the World Bank (WB), the World Trade Organization (WTO), and the European Union (EU). The core neoliberal policies are free markets and free movement of goods and capital (but not people), and the removal of state subsidies and protection from extractive and manufacturing industries and even public utilities and services, which are thereby exposed to competition from local and international profit-seeking companies.

Anti-Globalisation was the first mass resistance movement. Its first major demonstration was at an IMF/WB meeting in Berlin in 1988. This was followed by a mass protest at a 1989 Paris meeting of the G7 (heads of government in the world's seven largest economies), and an IMF/WB meeting in Madrid in 1994. However, the truly mammoth *Anti-Globalisation* demonstrations were at a meeting of the WTO in Seattle in 1999, and then at a G8 meeting in Genoa in 2001. These were the first mass mobilizations in which new technologies (mobile phone, e-mail, and Internet Web sites) were noted as playing significant roles. All authorities found these events difficult to police since there were no leaders to negotiate with or to arrest. We should note that *Anti-Globalisation* did not mobilize global youth. When the movement was at its height, youth in Eastern Europe were supporting their governments' efforts to attract direct foreign investment.

The *Generation of €1000* did not become a youth movement, but it spread awareness across Europe that young people's salaries were declining in real terms (when they could obtain jobs), and that the generation was at risk of being unable to live as well as its parents. These fears are proving well founded (Roberts, 2012b). *Generation €1000* was the title of an Italian book by Antonio Incorvio and Alessandro Rimassa that was published on the Internet in 2005 and instantly 'went viral.' The book subsequently became a film. In 2005, €1000 per month was the typical starting pay being offered to university graduates across Southern Europe and sometimes in Northern European countries as well. The change since 2008, the year of the banking crisis, has been a further depression of youth salaries. These dipped to under €500 in Greece in 2011.

During 2011, thousands of young Greeks joined a series of demonstrations in Athens against the austerity policies that were being imposed by 'the troika' (the IMF, the European Central Bank, and the European Commission), but these were not specifically youth mobilizations. Then in 2011, Spain became the second EU member state (Greece had been the first) in which youth unemployment rose above 50%. In Spain, this austerity sparked the *Indignados* demonstrations that spread throughout the country and into Portugal (Castells, 2012). Research among Spain's *Indignados* suggested that they were developing new ways of practicing politics. They were not interested in joining or forming political parties or even voting in elections. Some claimed that 'the process is the product,' meaning that the demonstrations were ends in themselves. Their preferred ways of 'participating' were via interactive Web sites and at mass meetings, which worked toward a consensus. Some were also involved in pioneering alternative economic practices (Conill et al., 2012). It was the same in the *Occupy* movement that erupted when the *Indignados* demonstrations were fizzling out in the autumn of 2011. The first *Occupy* event was *Occupy Wall Street* whose 'camp' in Zuccotti Park commenced on 17 September 2011. By the end of October, there were *Occupy* demonstrations in 95 cities in 82 countries, and in over 600 communities across the United States. The *Occupy* slogan was 'We are the 99%.' By the end of 2011, this series of events was drawing to a close. Research among

Indignados and *Occupy* demonstrators, and also among participants in the *Arab Spring* (Honwana, 2013, and see below), produced similar findings as regards the motivations and behavior of young protesters. Low turnout at elections and miniscule political party membership figures should not be read as signs of young people's political apathy. Rather, as in Eastern Europe in 1989 and as context for the subsequent 'color revolutions,' indignation and outrage can be there all the time before suddenly erupting in short-lived outbursts of protest (Benedicto, 2008).

The Arab Spring

This is the name given to a series of mass protests that began at the start of 2011. Young people were very high profile, more so than in any of the instances discussed above except the student movements of the 1960s and the 1970s. This was due to the demography of cities in the 'global south' (Hansen, 2008). The countries have young populations, and their young people have been moving into cities despite facing chronic shortages of housing and employment.

The spark that ignited the *Arab Spring* was struck on 17 December 2010 in Sidi Bouzid, a small town in Tunisia, when a 26-year-old street vendor, Mohamed Bouazizi, set fire to himself when instructed to cease trading because he did not have a license. He died on 4 January 2011, and this sparked 29 days of sustained protest, beginning in Tunis, the capital city, and ended only when President Ben Ali fled (Honwana, 2013). By then, protests had started in Jordan (January 14) and Oman (January 17), followed by a 'Day of Rage' called for January 25 in Cairo's Tahrir Square, which became the site of a sustained protest that ended only when President Mubarak stepped down on February 11 (Osman, 2013). Protests began in Yemen on January 27, in Bahrain on February 14, in Benghazi (Libya) over the death of a human rights lawyer on February 15 (Vanderwalle, 2012), in Morocco on February 20, and in Syria on March 15. There were shorter lived outbursts in some other countries.

The role of the new media in the rapid spread of these protests is disputed. *Facebook* and *Twitter* are said to have been crucial in drawing the initial crowds, *YouTube* could spread anger by posting images of police brutality, often recorded on mobile phones (Shahine, 2011). However, at the beginning of 2011, only around a quarter of Egypt's population was using the Internet and less than 6% used *Facebook* (Noueihed and Warren, 2013). *Al Jazeera*, the television channel, was probably crucial in creating a transnational public sphere and in amplifying messages with new media origins. It is noteworthy that wherever the protests led to elections, the 'Twitterati' were likely to be shocked to discover that they were nowhere near 50% let alone 99%.

The outcomes of these protests differed from country to country. Protests were quickly suppressed or ended with minor concessions in Bahrain, Jordan, Oman, and Morocco. There were changes of government in Tunisia, Egypt, and Yemen, and in Libya after intervention by North Atlantic Treaty Organization forces, the death of Gaddafi on 20 October 2011, and the election of a General National Congress in July 2012. In Syria, the protests in Damascus during March 2011 led to a prolonged civil war. In Egypt, elections in December 2011/January 2012 made the Moslem Brotherhood the largest group in the new parliament, and the presidential election in July 2012 produced a Moslem Brotherhood winner, Mohamed Morsi. The parliament was dissolved by Egypt's judiciary, and in July 2013, President Morsi was deposed in an army coup, which was followed by the arrest of Moslem Brotherhood leaders and the shooting of hundreds of Moslem Brotherhood street protesters. Tahrir Square was the site of further demonstrations, supporting and celebrating the army's latest actions.

Assessment, Prospects, Future Directions

There are several conclusions that can be drawn from a survey of youth movements since the 1950s.

- These movements are most successful in mobilizing dissent, anger, despair, and outrage, while they have no clear aims or leaders. This has applied in all the movements reviewed above.
- Incumbent regimes can face down prolonged and massive youth protests for as long as they retain the support of their armed forces. Eventually the crowds of protesters dwindle. Presidents and monarchs are toppled only when their security melts away, as happened in Eastern Europe in 1989, in all the 'color revolutions,' and in Tunisia and Egypt during the Arab Spring.
- Elections are disastrous for youth protest movements. Once parties are formed, leaders elected, and policies declared, protesters tend to fragment, while others return to 'the party of the couch.' On election day, another faction of the 'couch party' is likely to emerge, and then protesters find that they do not represent 'all' or even 'most' of their people.
- The new media may be able to accelerate the spread of protests, but revolutions of 1989 happened with underground printed media.
- Up to now there have been no global youth movements. This is despite necessary preconditions having arisen – the Internet which can host 'virtual' global communities, and the instant global transmission of sound and images so that, for example, 9/11 is now etched into the memories of people in countries all over the world (Edmunds and Turner, 2005). It is also despite the scale of present-day youth migration, some permanent, but in other cases by young people who intend to return home after completing courses or after earning and saving money, which means that local youth scenes and cultures are often cosmopolitan (Beck and Beck-Gernsheim, 2009). The revolutions in 1989, the 'color revolutions,' the *Indignados*, *Occupy*, and the *Arab Spring* have all been concentrated within if not confined to specific groups of countries.

Many governing regimes have held onto power throughout long periods and have seen off the formation of successive political generations and associated youth movements. The Alouite monarchy has ruled Morocco since 1666, although with far less power today than in 1666. Russia's political elite has held power since the time of Lenin. Splits and power transfers have been within the elite. There have been no revolutions from below, and President Putin's young supporters vastly outnumber the more visible opposition, which tends to

be concentrated in major cities and comprises mainly highly educated young people from intelligentsia families (Dafflon, 2009; Lyytikainen, 2013). There are Central Asian countries whose communist leaders in Soviet times became their independent countries' first presidents and still hold office (Kazakhstan, Tajikistan, Uzbekistan). Modern dynasties are being created. Heydar Aliyev became president of Azerbaijan in 1993 (he was formerly part of the communist elite) and was succeeded by son Ilham in 2003. Hafez al-Assad became president of Syria in 1971 and was followed by son Bashar in 2000.

Older-type adult-led youth services still flourish. Major political parties have youth sections. Many governments run schemes that recruit young people for voluntary service at home and abroad. The European Union has projects that draw the continent's youth together in citizenship-building projects (Kovacheva, 2000; Loncle et al., 2012). A lesson is that adaptation and evolutionary change have been and remain alternatives to revolutions, and both can be products of movements formed and led either by adults or by young people.

See also: Civil Rights Movements; Collective Behavior, Sociology of; Counterculture: 1960s and Beyond; Ethnic, Racial, and Nationalist Movements; Identity Movements; Movements, Social; Youth Culture, Sociology of.

Bibliography

Beachain, D.O., Polese, A., 2010. "Rocking the vote"; new forms of youth organisations in Eastern Europe and the former Soviet Union. Journal of Youth Studies 13, 615–630.

Beck, U., Beck-Gernsheim, E., 2009. Global generations and the trap of methodological nationalism for a cosmopolitan turn in the sociology of youth and generation. European Sociological Review 25, 25–36.

Benedicto, J., 2008. Young people and politics: disconnected, sceptical, an alternative, or all of it at the same time? In: Benedicto, J., Blasco, A.L. (Eds.), Young People and Political Participation: European Research. Young People's Studies Magazine, 81, June, Observatory of Youth in Spain, Madrid, pp. 13–27.

Blackstone, T., Gales, K., Hadley, R., Lewis, W., 1970. Students in Conflict: LSE in 1967. Weidenfeld and Nicolson, London.

Burnett. J., 2000. Time for a change: a contribution to the sociology of generations. Paper Presented at British Sociological Association Conference, York.

Castells, M., 2012. Networks of Outrage and Hope: Social Movements in the Internet Age. Polity Press, Cambridge.

Collin, M., 2007. The Time of the Rebels: Youth Resistance Movements and 21st Century Revolutions. Serpent's Tail, London.

Conill, J., Castells, M., Cardenas, A., Servon, L., 2012. Beyond the crisis: the emergence of alternative economic practices. In: Castells, M., Caraca, J., Cardoso, G. (Eds.), Aftermath: The Cultures of the Economic Crisis. Oxford University Press, Oxford, pp. 210–248.

Dafflon, D., 2009. Youth in Russia – The Portrait of a Generation in Transition. Swiss Academy for Development, Biel.

Edmunds, J., Turner, B.S., 2005. Global generations: social change in the twentieth century. British Journal of Sociology 56, 559–577.

Evans, W.M., 1965. Young People in Society. Blackwell, Oxford.

Fendrich, J.M., 1977. Keeping the faith or pursuing the good life. American Sociological Review 42, 144–157.

Fowler, D., 2008. Youth Culture in Modern Britain, c1920–c1970. Palgrave Macmillan, Basingstoke.

Hall, G.S., 1904. Adolescence. Appleton, New York.

Hansen, K.T., 2008. Introduction: youth and the city. In: Hansen, K.T., Dalsgaard, A.L. (Eds.), Youth and the City in the Global South. Indiana University Press, Bloomington, pp. 3–23.

Honwana, A., 2013. Youth and Revolution in Tunisia. Zed Books, London.

Inglehart, R., 1977. The Silent Revolution. Princeton University Press, New Jersey.

Kovacheva, S., 2000. Keys to Youth Participation in Eastern Europe. Council of Europe, Strasbourg.

Loncle, P., Cuconato, M., Muniglia, V., Walther, A. (Eds.), 2012. Youth Participation in Europe. Polity Press, Bristol.

Lyytikainen, L., 2013. Gendered and classed activist identity in the Russian oppositional youth movement. Sociological Review 61, 499–524.

Mannheim, K., 1952. The problem of generations. In: Essays on the Sociology of Knowledge. Routledge, London.

Noueihed, L., Warren, A., 2013. The Battle for the Arab Spring: Revolution, Counter-revolution and the Making of a New Era. Yale University Press, New Haven.

Osman, T., 2013. Egypt on the Brink: From Nasser to the Moslem Brotherhood. Yale University Press, New Haven.

Pangburn, W.M., 1940. Play and recreation. Annals of the American Academy of Political and Social Science 212, 121–129.

Pilkington, H., 1994. Russia's Youth and Its Culture. Routledge, London.

Reich, C.A., 1972. The Greening of America. Penguin, Harmondsworth.

Roberts, K., 2012a. 1989: so hard to remember and so easy to forget. In: Leccardi, C., Feixa, C., Kovacheva, S., Reiter, H., Sekulic, T. (Eds.), 1989 – Young People and Social Change After the Fall of the Berlin Wall. Council of Europe, Strasbourg.

Roberts, K., 2012b. The end of the long baby boomer generation. Journal of Youth Studies 15, 479–497.

Roszak, T., 1970. The Making of a Counter Culture. Faber, London.

Savage, J., 2007. Teenage: The Creation of Youth Culture. Chatto and Windus, London.

Schuman, H., Corning, A.G., 2000. Collective knowledge of public events: the Soviet era from the Great Purge to glasnost. American Journal of Sociology 105, 913–956.

Shahine, S.H., 2011. Youth and the revolution in Egypt. Anthropology Today 27, 1–3.

Vanderwalle, D., 2012. A History of Modern Libya. Cambridge University Press, Cambridge.

Zhuk, S.I., 2010. Rock and Roll in the Rocket City: The West, Identity and Ideology in Soviet Dneipropetrovsk. John Hopkins University Press, Baltimore.

Youth, Music, and Peace Building

Mwenda Ntarangwi, Calvin College, Grand Rapids, MI, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

This article explores the ways through which youth mobilize music to advance peace in their communities, specifically emphasizing that music is not only used to bring people together for social harmony and reconciliation but also for protest and resistance in order to create a 'culture of peace.' It further seeks to establish the impetus that propels youth into persistent participation in peace building especially through an analysis of the interrelatedness of concepts of peace building and culture of peace as they interact with youth identity. Using examples of the role played by new media in shaping such identity, the article illuminates the power of music in building some kind of virtual coalition that promotes peace. As an expressive medium that often defies physical and cultural barriers, music provides both a platform for advancing peace and an analytical tool through which to understand sociocultural dynamics within which peace building can occur.

Introduction

The twenty-first century has brought with it a world where violence has become banal; where armed political conflict has become the most common mode of solving disputes; where unequal access to resources has subdivided the world into new fiefdoms based on ethnic and racial identity, religion, gender, age, and social class; and where economic competition and exploitation have led to both marginalization and privilege. But it has also brought greater access to information and information sharing; a wider awareness of social, political, and economic realities of people in faraway locations; sensitivity to indigenous rights; and a growing participation of youth in public life. These features of the century allow for a more complex analysis of lived experiences of people around the world, sometimes in real time, as well as projecting what life ought to be by using multiple forms of representation including through music.

Advanced technological processes witnessed in the last few decades have allowed youth to use music and other forms of media to express their lived realities and their aspirations of what their lives should entail. Peace building is one such phenomenon but it is shrouded in a number of extensions of meanings that cannot be overlooked. What is 'peace' and 'peace building,' and how do both concepts get mobilized by youth through music?

Harmonizing Definitions

Peace

Scholars have grappled with the meaning of peace, often seeking to provide a comprehensive definition to a concept that is ever elusive and complex. Definitions range from 'negative peace,' seen as 'absence of violence' and 'positive peace,' considered to be the 'integration of human society,' (Galtung, 1964: p. 1) to peace as 'absence of violence in all its forms – physical, social, psychological, and structural' (Reardon, 1988: p. 19). Historically, however, peace has been understood in reference to the absence of violence, hostility, or war. The Nobel Peace Prize, for instance, was instituted to recognize and honor those seeking to abolish or reduce standing armies as well as for promoting congresses or meetings that may lead up to signing peace treaties (Lundestad, 2001). In the academy, Peace Studies, which started in the mid-twentieth century with the first program in the United States, created in 1948 at the then Manchester College in Indiana and focused on understanding how peace can be reached through conflict resolution.

Many contemporary scholars operate with an understanding that peace is "not just the absence of war but also the presence of the conditions for a just and sustainable peaceful society, including access to food and clean drinking water, education for women and children, security from physical harm, and other inviolable human rights" (Kroc Institute, 2012). As a result scholars have expanded how they study and understand peace so as to allow for possibilities of seeing people in all categories actively participating in peace building rather than as a state or government endeavor. It is in line with this approach that scholars such as John Paul Lederach have sought to pay attention to the conditions that ought to be present for peace to prevail. Lederach argues that for peace to be attained there has to be the cultivation of 'moral imagination.' Moral imagination, he argues, "is cultivated through four capacities – first, the capacity to imagine ourselves in a web of relationships that include our enemies; second, the ability to sustain a paradoxical curiosity that embraces complexity without reliance on dualistic polarity; third, the fundamental belief in and pursuit of the creative act; and finally, the acceptance of the inherent risk of stepping into the mystery of the unknown that lies beyond the far too familiar landscape of violence" (Lederach, 2010: p. 5).

Peace building becomes a process embedded in reflections about self and the other as well as in creative ways of relating to others one has had no prior relationship with or with whom such a relationship is based on limited contact or even stereotypes.

Contributing to this expanded definition of peace, Olivier Urbain expands on Galtung's earlier definition of peace to show how empathy, creativity, and nonviolence can be understood as components of peace only when placed within contexts that would otherwise negate peace. He notes that creativity is needed to enable people unlock their powers of

 http://dx.doi.org/10.1016/B978-0-08-097086-8.64119-3

imagination that enables them to articulate similar goals with those with whom they may be in conflict; that creativity can work best when paired with empathy, otherwise it can be used to manipulate, swindle, or trick people; and, that conflict transformation comes when creativity and empathy work together (Urbain, 2008: pp. 4–5). When considered together, all these multiple definitions of peace and the similarly diverse approaches to peace building, all exhibit common threads that connect them, including references to violence, harmony, and human thriving. The pursuit of peace, therefore, comprises an active agenda that requires sensitization, negotiation, imagination, creativity, human flourishing, and long-term commitment. Peace building is not an event but a process.

Peace Building

To cultivate and maintain a culture of peace individuals, communities, and governments ought to proactively pursue it through a strategic process that builds the right circumstances for its thriving. Peace building, as Lesley Pruitt notes, is an active process that is pursued in order to establish 'positive peace.' Pruitt uses the phrase 'peace building' to "refer to bottom-up approaches aimed at preventing, reducing, transforming, and assisting people to recover from all forms of violence, including structural violence that has yet to result in widespread civil turbulence" (2013: p. xv). Institutionally, peace building has been championed by the United Nations (UN) through many of its departments, especially United Nations Educational Scientific and Cultural Organization (UNESCO). At its 1995 General Conference UNESCO introduced the concept of a 'culture of peace,' which it used to refer to "values, attitudes and behaviors that reflect and inspire social interaction and sharing, based on the principles of freedom, justice and democracy, all human rights, tolerance and solidarity" (Adams, 2003: p. 34). UNESCO's commitment to peace is also well represented in its motto, which reads, "building peace in the minds of men and women."

Elise Boulding, who attended the 1995 UNESCO meeting, carried out extensive work on peace and education and has provided many insights into the concept of the 'culture of peace' (2000). She argues that approaches that cultivate and acknowledge differences articulated through individual behavior and existing social structures as well as finding acceptable ways of appreciating uniqueness and similarity lead to a culture of peace. As a social scientist and activist, Boulding injected into the study of peace the need to understand the world before changing it and as such rejecting any essentialist positions taken on how people or cultures are or behave and instead asserting that humans have the potential for both peaceable and warlike behaviors, for a culture of violence as well as a culture of peace. Her 'rewriting' of history to bring in stories not represented in the dominant historical narratives of conquest allowed her to show how including the untold histories of women, the nurturing of families, the celebration of events, and all forms of human creativity represented an existing culture of peace that had been overlooked by an aggressive focus on the culture of war and conquest (Boulding, 1976, 1988, 1989). It is in this inclusive approach to peace building that youth, through the use of music, find a place to engage with the process.

Youth, Music, and Peace Building

In conflict-prone societies and those wrestling with recurrent election violence, internal resource-based tensions as well as those teetering on the brink of political violence, there are clear needs to cultivate a culture of peace. Many youth have turned to music as their platform from which to appeal to and convince their fellow citizens to build such a culture of peace. As a result music has been used to heal fractured relationships torn apart by ethnic-based violence and police brutality and to prevent a reoccurrence of violence or repeat confrontations. Youth are also very much aware that for real peace to prevail, socioeconomic and political challenges have to be addressed effectively. As they sing about peace and appeal for peace, youth challenge those in positions of power as well as their own peers to create conducive circumstances for all to thrive. Building and sustaining a culture of peace is critical for youth who not only are in most cases the majority in their societies but also are negatively affected by processes that undermine peace.

For many youth, pursuing peace is an act that is simultaneously voluntary as it is coerced. On the one hand, because youth are caught up between childhood and adulthood, they often find themselves unrepresented in areas of decision making that produce violence and social strife, often thrust into it because they are vulnerable. On the other, given their expanded understanding of the world they live in, their increased education opportunities and access to mass media, they no longer are content to wait in the fringes of society as others define their lives. Instead they take charge and seek to not only define themselves but also redefine their societies. These two realities of youth identity provide a complex lens through which to understand their attraction to and participation in peace building. Different phenomena have aided in this position taken up by youth including global processes that have pushed the sources and implications of social action beyond state borders, providing unprecedented advantages to youth interested in building peace in their communities.

Transformations in communication and information sharing have opened up opportunities for shaping individual ideas about the world and the local. As the emergence and expansion of youth-led initiatives such as *Ushahidi* – an open source project that allows users to crowd source crisis information that is then sent via mobile phones – has shown how youth social networks provide opportunities for action through the circulation of information about ongoing activities, existing organizations, people to contact, and a reduction of the practical cost attached to participation. Other forms of media such as music also allow youth to draw on a transnational and transregional orientation, to assess the social, political, and cultural conditions of their time. With socially and politically charged songs, artists give voice to the concerns of previously ignored groups and assert the legitimacy of those concerns in shaping opinions and policies on a national and regional level. The use of hip-hop in democratic movements in Tunisia and Senegal, for instance, shows the role youth play in shaping political outcomes in their societies (Honwana 2012). These new ways of communicating provide youth with social networks that enable identity mobilization and the gradual development of social movements that build platforms for positive change.

How does involvement in such social networks affect individual identity and behavior and how does it shape not only policy but national and international responses to peace building? How do individual networks shape other important forms and features of collective action? What is the connection between social networks and individual participation in social movements given that social ties are critical in mobilizing collective action? What are the mechanisms and dynamics that induce people to become involved in collective action and how does that provide a holistic explanation of the process of individual participation? If we take social ties to be important in shaping collective action which could lead to social movement and that participants are recruited through preexisting social ties, what are the ties that bind youth in championing peace? Similar to Tunisia and Senegal, the 2011 political power harnessed by Egyptian youth through social networks mobilized by social media and music has been recognized for its role in bringing down a longtime political entity. Tahrir Square, the center of the revolts in Egypt, provided fertile ground for new songs as well as highlighting old ones as artists sought for democracy and freedom through such songs as *Sout Al Horeya* (Voice of Freedom), *Matloob Zaeem* (Leader Needed), and *Ya El Midan* (O Square) (Watch Soud Al Horeya here https://www.youtube.com/watch?v=PAEt6QJJi-c, Matloob Za3eem here https://www.youtube.com/watch?v=bk2JIohWdHY, and Ya El Midan here https://www.youtube.com/watch?v=VQTmvIKI-f8 and here https://www.youtube.com/watch?v=ljVTj9yu-ns for English subtitles). These movement-based claims on the political system that speak the language of inclusion and representation are also examples of how youth are linked and even immersed in virtual/transnational communities and networks that have political and economic consequences on their lives.

The Egyptian revolts, along with other forms of the 'Arab Spring,' provide a good illustration of the vitality and creativity youth bring to public protest movements and which then shape political and social trends in their societies. Through activism sparked by their generation's increasingly interconnectedness brought by social media and technology, these youth, like many of their counterparts in other parts of the world respond to the reality of low wages, high unemployment, and poor governance, all closely tied to economic issues. Unemployment and poor governance that are mostly experienced by youth by virtue of their position in society fuel discord and negatively affect peace. Manifestations of such discord are also exemplified by the 'Occupy Wall Street' movement are organized by youth in major cities of Europe and North America, seeking to arrest a social system that has disenfranchised many. The youth involved in these movements mobilized through music and chants, demanding more social justice and equality for themselves as well as others across many boundaries (Watch video of Lupe Fiasco's song 'The End of the World' that was the Occupy Wall Street anthem here http://www.youtube.com/watch?v=PDDCmHr0G2E and here http://www.youtube.com/watch?v=E8VcUxpvzxQ for Jeremy Gilchrist's song 'We are the 99%'). Indeed, youth-led movements such as those in the United States and countries in Europe and Africa speak the language of inclusion and representation and seek the provision of services for real social transformation that ultimately lead to broad flourishing and prosperity for all.

Since peace is a condition that is necessarily perceived to be long lasting, many social movements do not always sustain peace because they are often mobilized through spontaneous political protests in which collective action is directed at a specific entity deemed to be impeding on the participants' identified rights or privileges. As soon as the target is met such social movements diminish. Others slowly wear out because the targeted problems are so entrenched in the system that they require a major structural shift to accomplish. Even in cases where targets are met, lasting change is not always attained as in the case of Egypt where youth that were initially involved in the movement were mostly sidelined in the new political dispensation as the Muslim Brotherhood took political power. This is the disillusionment that triggered the ensuing Egyptian demonstrations against president Morsi that escalated in 2013, leading to his ouster by the military. Such processes highlighted here may not be the best strategy for upholding peace but they reflect the reality of the fragile nature of peace building in the world. Music has the potential to be the thread that unites and holds together social movements across time and space because of its ability to endure beyond the immediate social circumstances within which it is produced. Music can be recorded, played back, shared across different social platforms, and even reconstituted to capture newer meanings. It is this versatile nature of music that allows it to be variably used by youth for peace building. In the following section, I share some examples from different parts of the world of how music has played an important role in reflecting and shaping peace building.

India

In 1969 John Lennon, still a member of the Beatles, released a single titled 'Give Peace a Chance' in Montreal, Canada. The song soon became a major hit around the world and especially in the United States where antiwar activists opposed to the Vietnam War used it. In 2012 two young men calling themselves 'Still Rhyming' from Nagaland, Northeast India performed a hip-hop song titled 'Give Peace a Chance.' The two artistes that form Still Rhyming are 'Rhymez Disguised' and 'Doak 17,' and perform hip-hop and R&B music, mostly in English. They are the first hip-hop group from Northeast India to be featured on Vh1 Hit Factory music show in India and when the video for the song was released in August of 2012, it won the 'best music video award' at the 2012 annual music awards of Nagaland held in September of the same year. The video starts with images of the terrorist attacks on the World Trade Center in New York in 2001, refugees in the Congo scrambling for relief food, and reports of suicide bombers detonating two bombs following each other. The two artistes are shown flipping through the television channels displaying these world images and asking, "Have you ever thought why are we facing this? It ain't supposed to be the way we see it." The artists then offer a solution, saying, "I wish we get solutions, I know it's a little bizarre, but if you want peace and love then prepare for war against evil ways" (Watch 'Give Peace a Chance' here https://www.youtube.com/watch?v=iNUubAvkuk8 and go here http://www.metrolyrics.com/give-peace-a-chance-lyrics-still-rhyming.html for the song lyrics). The video is shot in a space with a background of

multiple colors and many television sets lined up behind the musicians but the most prominent identifiable image is the picture of Malcom X placed at the right-hand corner.

Nagaland, a state in northeastern India, has a population of almost 2 million according to 2011 census data and is neighbor to Myanmar (Burma) to the east. It is also one of the only three predominantly Christian states in India with Christians constituting 90% of the population. Nagaland's position as an immediate neighbor to Myanmar has made visible to these musicians the many political challenges faced by many of its residents and shaped the music content of Still Rhyming. Moreover consistent with India's persistent caste system Nagaland also has its recognized scheduled caste communities, leading Still Rhyming to sing about its role in dividing up people. They wonder where humanity is when two lovers are put apart by caste and creed based on 'old beliefs.' They call for peace to be given a chance to make the world a better place. This reference to local cultural realities, the projection of transnational images, and reference to world catastrophes makes hip-hop such an important youth phenomenon for twenty-first century peace building. It allows them to creatively articulate their message with the aid of images that have the ability to remind, reflect, and reference phenomenon and events relevant to a cross section of population.

Israel and Palestine

Israel and Palestine represent one of the most prominent conflicts of the Middle East and despite many attempts to broker peace violence has almost become the sole identity of the relationship between the two societies. But that is not the only reality of the region. As Stein and Swedenburg (2005) show, a focus on the dynamic forms taken by popular culture allows us to move beyond the binary oppositions and preponderance of violence that tend to dominate much literature and reporting on the Israeli–Palestinian conflict. Music has been used historically to express various sentiments regarding the relationship between Palestine and Israel. Songs were sung in the mid-twentieth century urging Palestinians and other Arabs to liberate their stolen homeland (Massad, 2005). Today we see a different kind of approach to the conflict with youth from both societies engaging in peace building through music. One such group is named 'Heartbeat' that exemplifies the role music can play in bringing the two warring sides together.

Composed of youth from both Israel and Palestine, Heartbeat performs and tours together playing music, holding camps, recitals, and workshops, and developing ensembles that promote peace. The group was founded in 2007 by the efforts of Aaron Shneyer supported by a Fullbright-MTVu fellowship (Heartbeat 2012). One of their songs titled *Bukra Fil Mishmish*, for instance, explores the connections that exist between the two societies but which are torn by constructed differences. The song's title loosely translated means, 'when pigs fly' in reference the perceived impossibility of their coming together and making something happen (Video of the song can be watched here http://www.youtube.com/watch?v=7Xw6NiUb44o#t=68). The song is presented in different languages and styles of music to honor the different identities represented by the group members. The group comprises Guy Gefen from Rehovot (Israel), Muhammad 'Moody' Kablawi from Haifa (Palestine), and Siwar Monsour from Tira (Israel), among others. In an interview with Missy Daniel for PBS at the Busboys and Poets in Washington, DC, Guy says that, "peace is not only something that the governments will sign on a paper and that's it, we'll have peace." Peace can be created through people without any governments getting involved (Watch video and read interview transcript here http://www.pbs.org/wnet/religionandethics/2013/07/03/heartbeat/15189/). The song *Bukra Fil Mishmish* presents a balanced representation of the voices and styles of the three artists mentioned here with each not only contributing to the lyrics but also to a different music style. Guy sings about breaking down the walls, taking down the flags, and understand that they are all humans. Moody sings about the need to make one's anxiety disappear because liberating self leads to freedom and Siwa asks to be given the chance to play her violin. Research has shown that when individuals or groups in conflict share space and listen to each other, positive relationships emerge that could lead to peace through empathy and tolerance for ambiguity (McNair 2003; Ramirez 2007). The music and programs that Heartbeat uses to bring together Israeli and Palestinian youth have great potential for healing the Israeli–Palestinian conflict. Admittedly, this will not come quickly or easily.

Kenya

The Kenya electoral scene has been marred by violence since the advent of multiparty politics in 1991 but the most discussed is the 2007 postelection violence, which led to international intervention through Kofi Annan acting as a peace broker for the two competing sides represented by the incumbent Mwai Kibaki and the challenger Raila Odinga. In the 1992 and 1997 elections, violence emanated from frustrations of, in part, many feeling like they did not have power to make a difference in the politics of the day due to rigging and other corrupt practices. Indeed, in the 1992 election, close to 1 million youth were denied national identity cards needed to register to vote (Mutua, 2008). The 2013 elections resulted in a different outcome, however, with no acts of violence reported primarily due to strategic processes for peace laid down prior to elections including those undertaken by youth. Having gone through a major bout of postelection violence in 2007 that went on for weeks into early 2008 and leading to more than a thousand people dead and thousands displaced, Kenyans were keen on avoiding a similar scenario in the 2012/13 election period. There were a number of issues that could have triggered violence including the election date being moved from November 2012 to March 2013 and the final presidential election results delayed for days before being decided by the Supreme Court. But Kenyans remained calm through it all determined not to repeat the same violent responses witnessed in 2007. After the 2007 postelection violence Kenyan musicians played a major role in spreading messages of peace. A well-known artist named Eric Wainana, for instance, composed a song titled 'Daima' (forever) which extols Kenyans to live in peace and reflect the symbolized identities of the colors of Kenya's flag – black for their skin color, red for the blood shed for independence, green for their land, and white for

peace (Watch 'Daima' here http://www.youtube.com/watch?v=Dh3oKHJ7KZQ accessed 28.12.13. The Web site also has the song lyrics in the original Swahili language as well as an English translation). A group called Wakenya Pamoja (Kenyans together) comprising many young Kenyan musicians along with some pastors composed a song titled 'Wakenya Pamoja' in January 2008 in the midst of violence calling for peace (Watch video here http://www.youtube.com/watch?v=1y7O857Q1jk and here for one with English subtitles http://www.youtube.com/watch?v=KiVrP9AAMUg accessed 27.12.13.). And as Kenyans waited to vote in the 2013 elections a musician named Juliani, with a number of other artists, performed a song titled 'Mama Kenya' (Mother Kenya) in which they asked for peace while highlighting the vulnerability of women and children during election violence (See video of 'Mama Kenya' here http://www.youtube.com/watch?v=rj0mBVjNJXQ accessed 28.12.13.).

In Kenya as it is in other parts of Africa no segment of the population has adversely experienced the realities of violence and unemployment, as have the youth who comprise the majority of the population. Kenyan youth under the age of 30, for instance, constitute up to 70% of urban populations with many of them being directly involved in unemployment and poverty. They have, for many years, also been appropriated by politicians to foment chaos to the advantage of the politicians while others have been co-opted into militias to terrorize communities, including their own. And yet Kenyan youth have been keenly involved in processes that bring about peace especially through the composition and performance of music as seen in the few examples shared here. It is, however, in creating better contexts for youth to thrive that could lead to permanent peace. In 2008 the US Agency for International Development and the international development organization Mercy Corps came together with youth in Kenya to form 'Yes Youth Can,' a volunteer-based group of youth all over Kenya. The group brings together Kenyan teens and young adults to form village-level 'bunges' – youth councils – that, among other things, discourage violence, encourage civic activism, and provide young people with training and resources to start new businesses and nonprofits (Docksai, 2013). This group has helped many youth find meaningful employment and connect with their communities in ways that help build a lasting culture of peace. A few months before the 2013 elections, for instance, many youth from the Yes Youth Can group were vigilant in their areas of residence, keeping watch for any incidents that might fan violence and reported them to local and national authorities via their cell phones (Docksai, 2013).

Mozambique

On 5 February 2008, angry mobs brought all business to standstill in Maputo, the capital city of Mozambique. The mob was rioting, taking over the city streets in protest against a 50% hike in private minibus fares. This focus on transportation resonated with many city residents. A majority of them use minibuses as their most common means of transportation (See this story at http://freemuse.org/archives/1223 accessed 16.12.13.). Ultimately the rioters were able to get their way – forcing the government to not raise the price of petrol that had prompted the fare hike. Two months later on 30 April 2008 Mozambican rapper Edson da Luz (commonly known as Azagaia) was taken to the Maputo City attorney's office where he was questioned for an hour (Docksai, 2013). His crime? Composing a song titled *Povo no Poder* (people in power) in which he calls those in power corrupt thieves that have ruined the country, and threatens to barricade roads, stop buses from running, and respond to any police violence with even more violence (See video here http://www.youtube.com/watch?v=RhSKixT-n0w accessed 27.12.13.). Maputo city authorities saw a direct connection between Azagaia's song and the riots and brought him in to gauge the extent of his instigation.

Azagaia's music finds fertile space for social relevance among the large population of youth who are also unemployed. In 2011, an estimated 66.6% of Mozambique's total population was younger than 24 years with 21% of them falling in the 15–24 years of age category (See, www.cia.gov/library/publications/the-world-factbook/geos/mz.html accessed 26.12.13.). As Honwana's ethnographic research shows, 'corruption, nepotism, and favoritism' are the reason for lack of active 'involvement in politics,' among youth in Mozambique, leading them to feel excluded from being part of the political machinery and process unless they become 'yes men' (2012: pp. 112–13). Such exclusion leads to frustration and cynicism that in turn broods anger that is represented in Azagaia's music and the riots that took place in Maputo. For Mozambique to continue enjoying any form of peaceful existence, these social realities faced by youth as well as their expressed sentiments captured through such media as music, cannot be ignored.

Senegal

In the same year Azagaia was arrested, a Senegalese rap group called Kuer Gui released an album titled *Nos Connes Doleances*, whose song 'Coup 2 Guelle,' (The video can be watched here http://www.youtube.com/watch?v=ZLrTLPrUodQ (with English captions) and here http://www.youtube.com/watch?v=Q-6H3YCx3x4 (without captions)) castigated the corrupt and inept government of president Abdoulaye Wade and blamed him for a '7-year expensive reign' that had led to youth 'drowning in hunger and unemployment' (Honwana 2012: p. 119). Senegalese youth have almost always been at the center of social change. In 1988 they took to the streets to protest election results because of the lack of transparency in the voting process. In the 1990s youth again presented themselves as a generation sensitive to their environment when they decided to clean up the city of Dakar under the philosophy of a movement termed 'set setal' (order and cleanliness). The movement was geared at cleansing society through the restructuring of individual lives and neighborhoods and advocating for the 'sanitation' of public and social 'morals.' Through graffiti, frescoes, and other symbolic images, the youth made their presence on the political scene known with their new motto: order and cleanliness (Diouf, 1996).

With the withdrawal of the state from providing social services following structural adjustment policies, the youth became even more aware of the difficult situation they faced economically. This trend continued into the 2000s with

rising prices of staple foods, and fuel, manifesting themselves not only in people's individual lives but also through power cuts, floods in Dakar's suburbs, crisis in the educational system, and the health sector. These conditions led to the birth of another youth movement known as 'Y'en a Marre' (literally meaning 'we are fed up'). The movement brought together youth within the large Dakar hip-hop community, civil society members, journalists, and others opposed to the corrupt and inefficient regime of Abdoulaye Wade and gained even more momentum in its opposition to Wade's desire to tamper with the constitution in order to extend his eligibility to run for presidency a third time. Y'en a Marre was officially born in January 2011 and found its name from a song with the same title by a musician from Ivory Coast (but exiled in Mali) named Tiken Jah Fakoly (See video here http://www.youtube.com/watch?v=KVUwpFk3ZpA). His other songs including one titled 'Promesse Bla Bla' and 'Quitte Le Pouvoire,' that castigate leaders who not only fail to fulfill their promises to their electorate but also insist on staying in power too long (See video of 'Promesse Bla Bla' here http://www.youtube.com/watch?v=Lso_VLgxuhc and of 'Quitte Le Pouvoire' here https://www.youtube.com/watch?v=6y7Y2UqgckY). In December 2007, the year that the song was released, Tiken Jah Fakoly was declared *persona non grata* in Senegal because he had criticized Abdoulaye Wade's government. The hip-hop group Keur Gui was part of Y'en a Marre group and used their music to mobilize youth to participate in the installation of regime change through voting (Honwana 2012).

Pulsating into the Future

Music has the ability to enable peace building by promoting healing individually and collectively and by facilitating the rebuilding and restoring of relationships in a nonthreatening manner (Slachmuijlder, 2005). Music is also able to build peace by helping develop dialogue between communities (Pruitt, 2011) and stimulating social engagement that can create cultural awareness and appreciation (Heble, 2003). Such initiatives ultimately lead to peace building (Lederach and Lederach, 2010). As a genre of expressive culture, music has a number of qualities that provide it with this ability for peace building including its nonthreatening identity, portability, and flexibility. Music making itself as well as its performance creates an environment where participants (composers, singers/performers, and listeners/viewers) feel comfortable sharing their thoughts, ideas, and dreams with each other and even strangers.

Music welcomes different interpretations of the meaning and style it presents, it can carry different messages depending on who is involved, when and where, and no one is fearful of being ridiculed for interpreting a song wrong. Songs are open to multiple interpretations and renditions. Indeed, today's popular music often samples songs from the past and gives them new meanings and melodies. Peace building through such a genre allows participants to feel welcome and to apply their creativity in fruitful ways that not only seek a better life for all but also the changes necessary to make it happen.

See also: Artists, Competition and Markets; Conflict and Conflict Resolution, Social Psychology of; Conflict: Education and Youth; Peace Movements; Peace Processes; Peace and Nonviolence: Anthropological Aspects; Peace-Making in History; Peace; Violence and Media; Youth Culture, Anthropology of.

Bibliography

Adams, D., 2003. Early History of the Culture of Peace: A Personal Memoire. Available at: http://www.culture-of-peace.info/history/page34.html (accessed 22.12.13.).
Ardizzone, L., 2007. Gettin' My Word Out. SUNY press, New York.
Boulding, E., 1976. The Underside of History: A View of Women through Time. Halsted, New York.
Boulding, E., 1988. Building a Global Civic Culture: Education for an Interdependent World. Teachers College Press, New York.
Boulding, E., 1989. One Small Plot of Heaven: Reflections on Family Life by a Quaker Sociologist. Pendle Hill Press., Philadelphia, PA.
Boulding, E., 2000. Culture of Peace: The Hidden Side of History. Syracuse University Press.
Clay, A., 2006. "All I need is one mic": mobilizing youth for social change in the Post-Civil Rights Era. Social Justice 33 (2), 105–121.
Diouf, M., 1996. Urban youth and Senegalese politics: Dakar 1988–1994. Popular Culture 8, 225–249.
Docksai, R., July–August 2013. Kenya's youth take charge. The Futurist 47 (4). Available at: http://www.wfs.org/futurist/2013-issues-futurist/july-august-2013-vol-47-no-4/kenya%E2%80%99s-youth-take-charge (accessed 26.12.13.).
Felice, C.D., Wisler, A., 2007. The unexplored power and potential of youth as peace-builders. Journal of Peace and Conflict Development (11), 1–28.
Fischlin, D., Heble, A. (Eds.), 2003. Human Rights, Resistant Sounds, and Politics of Music Making. Black Rose Books, Montreal.
Frith, S., 1996. Music and identity. In: Hall, Stuart, Gay, Paul du (Eds.), Questions of Cultural Identity. Sage Books, London, pp. x–z.
Galtung, J., 1964. An editorial. Journal of Peace Research 1 (1), 1–4.
Heble, A., 2003. Take two/rebel musics: human rights, resistant sounds, and the politics of music making. In: Fischlin, Daniel, Heble, Ajay (Eds.), Rebel Musics: Human Rights, Resistant Sounds, and the Politics of Music Making. Black Rose Books, Montreal.
Honwana, A., 2012. The Time of Youth: Work, Social Change, and Politics in Africa. Kumarian Press, Sterling, VA.
Kroc Institute, 2012. What is Peace Studies? Kroc Institute for International Peace Studies at University of Notre Dame. http://kroc.nd.edu/about-us/what-peace-studies.
Lederach, J.P., 2010. The Moral Imagination: The Art and Soul of Building Peace. Oxford University Press.
Lederach, J.P., Lederach, A.J., 2010. When Blood and Bones Cry Out: Journeys through the Soundscape of Healing and Reconciliation. University of Queensland Press, Brisbane.
Love, N., 2002. Singing for our lives: women's music and democratic politics. Hypatia 17 (4), 71–94.
Lundestad, G., 2001. The Nobel Peace Prize, 1901–2000, Amprdquo. Available at: http://www.nobelprize.org/nobel_prizes/themes/peace/lundestad-review/ (accessed 21.12.13.).
Massad, J., 2005. Liberating songs: Palestine put to music. In: Stein, Rebecca, Swedenburg, Ted (Eds.), Palestine, Israel and the Politics of Popular Culture. Duke University Press, Durham, NC, pp. 175–201.
McEnvoy, S., 2000. Communities and peace: Catholic youth in Northern Ireland. Journal of Peace Research 37 (1), 85–103.
McEnvoy, S., 2001. Youth as social and political agents: issues in post-settlement peace building. In: Kroc Institute Occasional Paper #21:OP(2), pp. 1–40.
McIntyre, A., Thusi, T., 2003. Children and youth in Sierra Leone's peace-building process. African Security Review 12 (2), 73–80.
McNair, R., 2003. The Psychology of Peace: An Introduction (Psychological Dimensions to War and Peace). Praeger Publishers, Westport, CT.
Mutua, M., 2008. Kenya's Quest for Democracy: Taming the Leviathan. Lynne Rienner Publishers, London.
Pruitt, L., 2011. Creating a musical dialogue for peace. International Journal of Peace Studies vol. 16 (1), 81–103.
Pruitt, L., 2013. Youth Peacebuilding: Music Gender and Change. State University of New York Press, Albany.

Ramirez, M.J., 2007. Peace through dialogue. International Journal on World Peace 24 (1), 65–81.
Reardon, B., 1988. Comprehensive Peace Education: Educating for Global Responsibility. Teachers College Press, New York.
Slachmuijlder, L., 2005. The Rhythm of Reconciliation: A Reflection on Drumming as a Contribution to Reconciliation Processes in Burundi and South Africa. Brandeis University, Waltham, MA.
Stein, R., Swedenburg, T. (Eds.), 2005. Palestine, Israel and the Politics of Popular Culture. Duke University Press, Durham, NC.
Storr, A., 1992. Music and the Mind. HarperCollins Publishers, London.
Urbain, O., 2008. Music and Conflict Transformation: Harmonies and Dissonances in Geopolitics. I. B. Tauris & Co. Ltd, London.

Relevant Websites

http://www.heartbeat.fm/ – Heartbeat unites Israeli and Palestinian youth to build trust and transform conflict through the power of music. Founded in 2007 under a Fulbright-MTVu fellowship, Heartbeat empowers Israeli and Palestinian musicians by creating opportunities and spaces for musicians from both sides to hear each other and work together to build trust between their communities. Since 2007, over 100 youth musicians have come together through a variety of programs including retreats, workshops, camps, overseas exchanges, and the development of ensembles. Heartbeat's core program focuses on the establishment and facilitation of ensembles of young Arab and Jewish musicians (14–21 years of age) who come together weekly for sustained music-based dialogue programs in multiple cities. During these weekly sessions, Heartbeat youth explore each other's cultures and narratives in safe, politically neutral settings that support creativity and communication. Developing skills in improvisation, songwriting, performance technique, music theory, communication, and leadership while engaging in dialogue about their daily lives and the conflict, participants raise their critical consciousness, and harness their agency to build trust between their communities.

http://www.dmeforpeace.org/discuss/designing-music-based-conflict-transformation – The Website explains some objectives and goals that peace building practitioners might integrate into the design of music-based conflict transformation projects, and are derived from cases in Europe, South America, and Africa.

http://www.yorksj.ac.uk/education–theology/faculty-of-etrs/cris/cris-resources/creative-media-arts-and-peace.aspx – This Website is a project based at The Centre for Religion in Society (CRiS) was established in the Faculty of Education and Theology at York St John University in 2008. CRiS provides a forum where scholars, in particular from the disciplines of Theology and Religious Studies, but also from other disciplines in the humanities and social sciences, as well as relevant practitioners and professionals, come together to look at the role of religion in the cultural, social, and political life of society as well as its role in conflict resolution, peace building, and reconciliation. The aim of the Centre is to enhance the national and international profile of the Faculty and the University, thereby increasing opportunities for joint research projects and securing external funding.

http://www.yett.org/index.php?option=com_content&view=frontpage&Itemid=1 – YETT is a youth networking organization committed to the full participation of young people in sustainable development through advocacy and capacity building of youth organizations in Zimbabwe. It exists primarily to serve the youth and youth organizations to ensure that the youth are involved in all aspects of development without resorting to violent means. YETT works with youth-led organizations that work directly with marginalized young people in urban and rural areas.

Z

Zooarchaeology

Jack M Broughton, University of Utah, Salt Lake City, UT, USA

© 2015 Elsevier Ltd. All rights reserved.

This article is a revision of the previous edition article by D.K. Grayson, volume 24, pp. 16691–16695, © 2001, Elsevier Ltd.

Abstract

Zooarchaeology is the subdiscipline of archaeology that focuses on the analysis of animal remains from archaeological contexts. An integral part of prehistoric archaeology since the emergence of the discipline in the nineteenth century, zooarchaeological analyses can provide information on the chronological position of archaeological sites, on the nature of past landscapes, on human interactions with those landscapes, on human diet, and on a wide variety of past human behaviors that would otherwise be invisible to us. Such analyses can also provide information relevant to the conservation and management of living animals.

Introduction

Zooarchaeology is the subdiscipline of archaeology that focuses on the analysis of animal remains from archaeological contexts. The term has a less euphonious twin, archaeozoology, which has less currency in English than it does in other languages. While some have distinguished among the kinds of approaches that might be treated under these two terms (Reitz and Wing, 2008), research identical in form and content is published under both names.

A thorough history of zooarchaeological research has yet to be written. It would be a daunting task to do so, since highly relevant literature in a diverse variety of languages extends well back into the nineteenth century. Zooarchaeology has continued to be an inherent part of archaeological research for well over a century for a wide variety of reasons: faunal data can help place archaeological sites in time, provide detailed information on past human diets, provide detailed knowledge of the environments in which past peoples lived and on human impacts on those environments, and, less securely, provide insights into the nature of the social organization of the people involved.

Even though zooarchaeology has always been an integral part of prehistoric archaeology, zooarchaeology as a distinct and recognizable subdiscipline extends no further back than to about 1970. Prior to that time, most research dealing with animal remains from archaeological sites was conducted by people whose primary training and interests lay outside of archaeology itself, usually within zoology or paleontology. The research done by these practitioners was frequently of superb quality and enduring value, but was often poorly integrated with other aspects of the archaeological research involved.

This situation began to change in the mid-1960s, as archaeology as a whole began to shed its largely descriptive roots and turned toward more explicitly scientific approaches to the past. Within a decade or so, there came into existence a set of scientists with intensive training in both biology and archaeology whose research focused on zooarchaeological matters, and who began developing techniques, methods, and theory relevant to answering those questions (e.g., Grayson, 1984; Klein and Cruz-Uribe, 1984). Today, most well-regarded archaeology PhD programs in the USA now include zooarchaeology as part of their training program.

Compared with other object-based research realms in archaeology, including lithic and ceramic analysis, zooarchaeology benefits tremendously from the fact that the units studied are provided by biology, and that their biological meaning is never in doubt. These units tend to be of two prime sorts. First, there are taxonomic units, drawn from the Linneaen system – species, genera, and families, for instance. Second, there are anatomical units – the separate parts of the vertebrate skeleton, for instance, or of the insect exoskeleton. As a result, while zooarchaeologists might argue over the identification of a particular specimen – and making secure taxonomic identifications by comparing ancient bone fragments with known modern comparative skeletons is the core of zooarchaeological lab work (Driver, 2011) – they do not argue about the meaning of their classification systems (*see* Archaeology: Philosophy and Science). This is often not the case in lithic or ceramic analysis – as anyone who has worked with European Neanderthal stone tool assemblages or with ceramic Poverty Point objects in southeastern USA can attest.

Vertical Applications

Much zooarchaeological research is directed toward understanding change through time, and relies heavily on the

analysis of data from stratified archaeological sites or from sites which, if not stratified, can be arrayed in time. As a result, such research has a strong temporal, or vertical, component, and can differ in approach from research that focuses on a particular site or set of sites at a precise point in time, and is thus primarily spatial or horizontal in emphasis. These different approaches may be driven largely by the research problems that intrigue a particular investigator, but they may also be driven by the kinds of sites available for analysis. Because vertical zooarchaeological research tends to focus on change through time, it often draws more heavily on biological theory than do horizontally oriented analyses, which tend to depend more heavily on less well-developed social theory. Even though this is the case, vertical analyses often have horizontal components, horizontal analyses often have vertical components, and either approach can make use of either kind of theory.

In recent years, vertical zooarchaeological research has revolutionized our understanding of the relationship between prehistoric peoples and their biotic environments. It has done this by showing the impacts that even small-scale societies can have on their faunal landscapes, and by showing that those impacts can force changes in the behavior of the people themselves.

Those impacts can be dramatic. For instance, it has long been clear that people caused the extinction of some 11 species of large, flightless moas in New Zealand within a few hundred years of the human colonization of these islands some 1000 years ago. While it is not known what combination of human hunting, habitat alteration (largely by burning), and human-introduced predators (rats and dogs) caused moa extinction, human activities were certainly to blame.

During the past several decades, however, zooarchaeological work throughout Polynesia has shown that bird extinctions followed human arrival in island settings with deadly consistency. While large, flightless birds were particularly vulnerable, smaller, flighted species were lost in high numbers as well (Steadman, 2006), and over half the species that comprised the Polynesian avifauna may have been lost after the initial human colonization of these islands. As is the case with New Zealand's moas, the exact causes of each extinction is unknown, but some combination of human hunting, human-induced vegetational change, and human-introduced predators was involved (Kirch, 2005; Steadman, 2006).

Zooarchaeological research has detected impacts of this sort, though not necessarily of this magnitude, in other parts of the world as well. In some of these places, the branch of behavioral ecology known as foraging theory has been used to understand not only why particular species declined in abundance under conditions of human predation, but also the impact of such declines on the people who caused them.

Among other things, foraging theory is designed to answer two apparently simple questions. First, given that resources are not often distributed randomly across the landscape but are instead clumped into patches (like ducks at oases), which of those patches will a predator choose to visit and how long should it stay there? Second, given that multiple prey types, or species, exist in a given patch, which of these prey types should a predator bother to pursue upon encountering it? The former issues are treated with what are called the patch-choice and patch-use models, the latter with what is called the prey-choice or diet-breadth model, but combinations of the two also exist. Developed in biology for nonhuman predators – but inspired by microeconomic theory from economics – foraging theory has seen wide application in contemporary human contexts, with impressive successes (Winterhalder and Smith, 2000; Nettle et al., 2013).

In recent years, it has begun to be applied to archaeological contexts as well (Bird and O'Connell, 2006). These applications have typically assumed that human prey choice is driven by energy returns, and that up to a point, larger animals provide higher returns than smaller ones. From this, it follows that vertical analyses of archaeological faunal assemblages should show that these animals face the steepest decline in the face of human predation. Zooarchaeological research has shown that they do just that.

In the San Francisco Bay area of California, for instance, it has been shown that such large-bodied animals as elk and deer formed an ever-decreasing part of human diets between 3000 and 2000 years ago, while such smaller and more costly animals as sea otters became increasingly important. After 2000 years ago, this pattern changed, and deer began to become more common in the diet, but only because sea otters themselves had become rarer, and it now paid to go further afield to hunt deer. The increasing use of distant deer-hunting patches was documented through analyses of patterns in anatomical part representation that showed an increase in field processing. In addition, the carcass portions that were returned to villages were subjected to more intensive processing – to judge from bone fragment size and stone tool damage – to apparently squeeze the maximum yield out of the declining availability of animal tissues. As all this was happening, sturgeon, the largest fish available in the area, not only declined in abundance through time, but even declined in average age and size. A wide range of large-sized bird species such as geese and larger shorebirds also declined in abundance over time compared to smaller and more elusive birds (Broughton, 1999, 2013).

The moa case, introduced earlier, provides another example. It has been realized for many years that after moas became extinct, human diets in New Zealand became dominated by small prey species, including fish and shellfish. Recently, Nagaoka (2005) has used foraging theory to provide the details of this process.

The prey-choice model predicts that, as resources that provide high-energy returns decrease in abundance, diets will broaden and the number of prey types included in the diet should increase. Using data from the Shag Mouth archaeological site on the east coast of New Zealand's South Island, Nagaoka has shown that, as the abundance of moas decreased through time here, not only did higher cost, smaller vertebrates become more important in the diet, but the number of species included in the diet increased as well. She has also shown that as moa populations dwindled, the places people went to procure food changed substantially. As time went on, for instance, people began to bring only the meatiest parts of moas back to Shag Mouth – as with deer mentioned earlier – suggesting that they had to go further and further to hunt those moas that still existed. In addition, during the earliest occupations of the site, most prey species

came from either inland or coastal settings. Only late in the history of the site did people begin to exploit offshore resources, mainly barracuda, heavily, and this only after the human activities had caused depletion of larger vertebrates available nearer-by. All this occurred within a period of less than 200 years, between about AD 1300 and AD 1500.

While the trends from these examples have been documented with standard zooarchaeological data (bone-based patterns in species abundance, anatomical portions, age/size structure, bone fragmentation), more recent work has applied information from isotope chemistry (geo-sourcing) and ancient DNA derived from animal bone to evaluate, for instance, long-distance hunting and population declines of favored game species (Grimstead, 2011; Broughton et al., 2013). These approaches are based on the facts that animals record in their bony tissues chemical signatures (e.g., ratios of different kinds of strontium) unique to their local environments and that genetic diversity is closely tied to the population size of animals. Both these chemical signatures and DNA can be readily extracted from animal bones tens of thousands of years old and such data are becoming increasingly important in zooarchaeological research.

In the last several decades, cases of resource depression such as those from San Francisco Bay and New Zealand described earlier, have been documented in a wide range of times and places all over the world (e.g., Kirch, 2005; Stiner and Munro, 2011). Most importantly, the implied declines in foraging efficiency (overall calories generated per hour of hunting) and increasing risks of food shortfalls have been linked to a wide range of other trends in human behavior and biology. These include, but are not limited to, increases in the degree of violence and warfare, elevated levels of disease and dietary stress, the emergence of social hierarchies, and even to such monumental phenomena as regional abandonments and the collapse of past civilizations (Broughton et al., 2010; Emery, 2007; Schollmeyer and Driver, 2013).

However, patterns of long-term stability in animal resource use and apparent sustainability have also been observed (Campbell and Butler, 2010; Etnier, 2007). Anthropogenic depressions seem to occur in contexts characterized by high human population densities fueled by high levels of subsistence productivity derived from either a rich array of wild resources or agriculture. Understanding in more detail the variables influencing apparent sustainability versus resource depression is an important current issue in zooarchaeology.

Zooarchaeologists have thus shown some of the kinds of impacts that even small-scale societies can have on the faunal landscape, and the influence that those impacts can in turn have on the people themselves. These kinds of theory-driven studies are fairly new. More traditional are zooarchaeological analyses that show the influence that 'natural' (i.e., nonanthropogenic) environmental changes can have on human adaptations, but these make far heavier use of, and greater contributions to, biological theory than they did in even the fairly recent past.

For instance, the site of Grotte XVI in southwestern France contains a series of archaeological deposits that date to between 36 000 and 12 000 years ago. Zooarchaeological analyses of the large mammal bones from these deposits showed that the percentage of reindeer that people hunted increased steadily through time, from 47% of all the large mammals at about 36 000 years ago, to 94% of those mammals at about 12 000 years ago. Comparing the dated archaeological record at Grotte XVI with a reconstructed summer temperature curve for southern France demonstrates that the increase in reindeer abundance was associated with a decrease in summer temperature (Grayson et al., 2001). Biogeographic theory shows that the challenges posed to mammals by temperature extremes, either hot or cold, can lead to mammalian communities that are dominated by a small number of species, and Grotte XVI, as well as other sites, show this process in operation through time.

Thus, vertical zooarchaeological analyses have illustrated in rich detail the complexities of human relationships with their faunal landscapes. In some cases, people caused the abundances of the species on which they preyed to decline dramatically, and both people and prey paid the price. In others, people were forced to react to environmental changes for which they were not responsible, and the results can be readily seen in their diets, and probably in other aspects of their lives as well.

Horizontal Applications

Horizontal approaches to zooarchaeology focus on fine-scale analyses of faunal material across space within, and occasionally between, sites, though other aspects of faunal data play a role here as well. These distributions are usually used to test hypotheses about social organization, hunting patterns, or other aspects of past human behavior, or to infer aspects of such behavior from the distributions themselves.

Work at the *c*12 500 year-old site of Verberie, located in the Oise River Valley, north of Paris, provides a simple yet compelling example of such an analysis. Unlike Grotte XVI, which is a cave, Verberie is an open-air site consisting of a series of thin depositional layers, each of which represents a distinct occupation of short duration. These occupations vary in extent from 45 to 250 m^2, and are marked by stone-lined fireplaces, around which stone artifacts are concentrated. The site is also quite rich in well-preserved bone, most of which is of reindeer (Audouze and Enloe, 1997; Enloe, 2004). The Verberie bones are distributed across each occupational layer, but many are found concentrated around relatively bone-free, circular zones. These concentrations tend to be marked by parts of the reindeer skeleton that do not have much meat associated with them in the living animal, suggesting that they had been left behind after butchering and removal of the meatier parts. In addition, these concentrations are located several meters from the hearths, which are, in turn, located several meters from one another. Taken as a whole, however, the occupation floors at Verberie contain all parts of the caribou skeleton. Since reindeer are large animals that are not likely to have been transported far when whole, the Verberie occupations would appear to represent hunting camps.

One of the most powerful tools zooarchaeologists have available for studying social organization is provided by what is known as refitting. There are two sorts of bone refitting, the first of which involves reassembling broken bones. Refits of this sort

can tell us about the stratigraphic integrity of a site, since fragments of the same bone found in different strata demonstrate that specimens have moved up or down the deposits. They can also provide much more precise information as to the nature of the faunal assemblage as a whole, since small shaft fragments often cannot be identified to species on their own, but can be identified once reassembled (e.g., Marean and Kim, 1998).

In the second approach, developed by Enloe (2004), detailed measurements and morphological comparisons allow bones that are highly likely to have come from the same individual animal to be refit to one another. If such refit body parts are found on spatially disparate parts of an undisturbed occupation surface, they have the potential of telling us about such things as food sharing and other aspects of past social organization.

Bone refitting of this second sort done by Enloe at Verberie shows that bones belonging to the same animal either did not travel very far from the areas in which they are concentrated (and at which initial butchering apparently occurred), or, if they did, traveled to, but not between, the hearths. This situation is quite different from that found at the roughly contemporaneous site of Pincevent, located not far from Verberie. Here, bones traveled from hearth to hearth, suggesting food sharing among the people using those hearths (Enloe, 2004). All this, along with a wealth of other evidence from the site, suggests that Verberie, unlike Pincevent, served as a briefly occupied hunting camp – a place from which hunting was conducted and to which the results of the hunt were brought for processing. Enloe was also able to use tooth eruption data to show that in all likelihood the Verberie reindeer had been killed during the Fall, and thus presumably during migration.

The work done at Verberie provides a remarkably detailed view of what happened at this particular place at about 12 500 years ago. Small groups of people came here to hunt reindeer during autumn. Hearths were constructed and stone tools fashioned or refashioned around those hearths; animals were hunted nearby, then brought back to be butchered, with the meat transported elsewhere.

Not all horizontal archaeology is as firmly grounded as that which has been conducted at Verberie and Pincevent. A good deal of it, in fact, is hardly compelling at all, consisting largely of poorly controlled assertions about the behavioral meaning of a particular set of faunal material. However, when horizontal zooarchaeology is well done, as at Verberie and Pincevent, the degree of detail it can provide about ancient human behavior can be close to astonishing.

Taphonomy

The differences between vertical and horizontal approaches to zooarchaeology are often in degree, not kind. Multiple horizontal analyses at a single stratified site can be combined to look at change through time, and many vertical analyses examine the distribution of specimens across space as well.

Both approaches require that we understand the nonhuman factors that may have impacted the faunal material that is under study. Carnivores, for instance, can introduce bones into sites occupied by people, and can modify both the condition and the location of bones deposited by people. Specimens can be destroyed by a roof-fall in a cave or by chemical action anywhere; they can be moved by water, trampled by large animals (including people), attacked by roots, and, in general, be affected by a depressingly broad variety of processes that can dramatically transform the assemblage that was originally deposited. Because of this, the area known as taphonomy, the study of the processes that intervene between the time an animal dies and scientists study what is left of it in the lab, has become extremely important in zooarchaeology (Lyman, 1994, 2008; Munoz and Mondini, 2008). Most zooarchaeologists are proficient at conducting taphonomic analyses. Without such analyses, there can be a high risk of interpreting as archaeologically meaningful patterns that have been produced in other ways.

Applied Zooarchaeology

As summarized earlier, the primary mission of zooarchaeology is to reconstruct and explain variation in past human societies and behavior from analyses of animal remains from archaeological sites. While tremendous insights into the human past have been gained from zooarchaeological research, knowledge germane to modern conservation biology and wildlife management has increasingly emerged from this work as well (Lauwerier and Plug, 2004; Wolverton and Lyman, 2012). These latter fields are often focused on understanding 'baseline' or 'benchmark' conditions for particular animal species, communities, or ecosystems that modern management efforts – policies and programs – are geared toward maintaining or restoring. 'Applied zooarchaeology' is thus the application of zooarchaeological data to these issues relevant to conservation.

Many applied zooarchaeology studies have provided information on the presence and abundance of particular animal species in regions where data on them are lacking or incomplete in historical records. Other studies forecast future faunal change, by showing how past climate changes impacted animal populations and comparing those with climatic changes – such as global warming – that are modeled to recur in the future.

One of the more high-profile cases of applied zooarchaeology involves the question of whether mountain goats are native to the Olympic National Park in Washington State (Lyman, 1995). The question has still not been resolved definitively, but detailed zooarchaeological research has clearly shown that the currently available faunal record from the Olympic Mountains is too meager to allow the detection of mountain goats even if they were present prehistorically (Lyman, 1995, 2008). More definitive examples include the documentation of Sandhill Crane breeding colonies in western Washington that were unknown to exist there from historical records (Bovy, 2012) and the finding that sea otters native to the Oregon coast prior to their eradication by nineteenth-century fur trade hunters belonged to the California subspecies rather than the Alaskan variety that had been the source of previously unsuccessful reintroductions (Valentine et al., 2008).

In perhaps the greatest success story of applied zooarchaeology, analysis of the Trappers Point archaeological fauna played a key role in the documentation of a 7000-year-old pronghorn migration route in western Wyoming. That analysis enhanced public awareness of pronghorn migrations that ultimately resulted in conservation easements from public and private landholders to help preserve the migration corridors. Most recently, over 10 million dollars in funding was provided by the Wyoming Department of Transportation to fund the construction of a series of fences, overpasses, and underpasses to allow for the safe passage of pronghorn herds across the highways that currently cross their ancestral migration paths (Sanders and Miller, 2004).

In short, the thriving subdiscipline of zooarchaeology provides us with information that can be acquired in no other way on past human diet, on human interactions with the environment, on those environments themselves, and on a surprising variety of past human behaviors that are themselves invisible but that leave faunal residues from which they can now at times be fairly confidently inferred. While this information is vital to understanding broad-scale patterns in past human behavior and evolution, it can also provide insights that are valuable to the conservation of animals and their environments today and in the future.

See also: Archaeology, Theory in; Archaeology: Philosophy and Science; Bioarchaeology; Environmental Archaeology; Ethnoarchaeology; Europe, Archaeology of; Food Production, Origins of; Hunter-Gatherer Societies, Archaeology of; Intensification and Specialization, Archaeology of; North America, Archaeology of; Pacific Islands, Archaeology of.

Bibliography

Audouze, F., Enloe, J.G., 1997. High resolution archaeology at Verberie: limits and interpretations. World Archaeology 29, 195–207.

Bird, D.W., O'Connell, J.F., 2006. Behavioral ecology and archaeology. Journal of Archaeological Research 14, 143–188.

Bovy, K.M., 2012. Zooarchaeological evidence for Sandhill Crane (*Grus canadensis*) breeding in northwestern Washington state. In: Wolverton, S., Lyman, R.L. (Eds.), Conservation Biology and Applied Zooarchaeology. University of Arizona Press, Tucson, pp. 23–41.

Broughton, J.M., 1999. Resource Depression and Intensification During the Late Holocene. University of California, San Francisco Bay. Anthropological Records, 32.

Broughton, J.M., Beck, R.K., O'Rourke, D.H., Rogers, A.R., Coltrain, J., 2013. A late Holocene population bottleneck in California tule elk (*Cervus elaphus nannodes*): provisional support from ancient DNA. Journal of Archaeological Method and Theory 20, 495–524.

Broughton, J.M., Cannon, M., Bartelink, E.J., 2010. Evolutionary ecology, resource depression, and niche construction theory: applications to central California hunter–gatherers and Mimbres-Mogollon agriculturalists. Journal of Archaeological Method and Theory 17, 371–421.

Campbell, S.K., Butler, V.L., 2010. Archaeological evidence for resilience of Pacific Northwest salmon populations and the socioecological system over the last ~7,500 years. Ecology and Society 15, 17.

Driver, J.C., 2011. Identification, classification, and zooarchaeology. Ethnobiology Letters 2, 19–39.

Emery, K., 2007. Assessing the impact of ancient Maya animal use. Journal of Nature Conservation 15, 184–195.

Enloe, J.G., 2004. Equifinality assemblage integrity and behavioral inferences at Verberie. Journal of Taphonomy 2, 147–165.

Etnier, M.A., 2007. Defining and identifying sustainable harvests of resources: archaeological examples of pinniped harvests in the eastern North Pacific. Journal of Nature Conservation 15, 196–207.

Grayson, D.K., 1984. Quantitative Zooarchaeology. Academic Press, Orlando.

Grayson, D.K., Delpech, F., Rigand, J.-P., Simek, J., 2001. Explaining the development of dietary dominance by a single ungulate taxon at Grotte XVI, Dordogne, France. Journal of Archaeological Science 28, 115–125.

Grimstead, D.N., 2011. Contradictions and compliments: the use of geochemistry and body part utility analysis to detect non-local procurement strategies during the late Holocene in northern California. In: Glassow, M.A., Joslin, T. (Eds.), Exploring Methods of Faunal Analysis: Perspectives from California Archaeology. Cotsen Institute Press, Los Angeles, pp. 33–52.

Kirch, P.V., 2005. Archaeology and global change: the Holocene record. Annual Review of Environment and Resources 30, 409–440.

Klein, R.G., Cruz-Uribe, K., 1984. The Analysis of Animal Bones from Archaeological Sites. University of Chicago Press, Chicago.

Lauwerier, R.C.G.M., Plug, I. (Eds.), 2004. The Future from the Past: Archaeozoology in Wildlife Conservation and Heritage Management. Oxbow Books, Oxford.

Lyman, R.L., 1994. Vertebrate Taphonomy. Cambridge University Press, Cambridge, UK.

Lyman, R.L., 1995. Determining when rare (zoo-)archaeological phenomena are truly absent. Journal of Archaeological Method and Theory 2, 369–424.

Lyman, R.L., 2008. Quantitative Paleozoology. Cambridge University Press, Cambridge, UK.

Marean, C.W., Kim, S.Y., 1998. Mousterian large mammal remains from Kobeh Cave: behavioral implications for Neanderthals and early modern humans. Current Anthropology 39, S79–S113.

Munoz, A.S., Mondini, M., 2008. Neotropical zooarchaeology and taphonomy. Quaternary International 180, 1–4.

Nagaoka, L.A., 2005. Declining foraging efficiency and moa carcass exploitation in southern New Zealand. Journal of Archaeological Science 32, 1328–1338.

Nettle, D., Gibson, M.A., Lawson, D.W., Sear, R., 2013. Human behavioral ecology: current research and future prospects. Behavioral Ecology 24, 1031–1040.

Reitz, E.J., Wing, E.S., 2008. Zooarchaeology, second ed. Cambridge University Press, Cambridge, UK.

Sanders, P.H., Miller, M.E., 2004. Archaeological evidence of pronghorn (*Antilocapra americana*) migration in the Upper Green River Basin of Wyoming: implications for wildlife management. In: Lyman, R.L., Cannon, K.P. (Eds.), Zooarchaeology and Conservation Biology. University of Utah Press, Salt Lake City, pp. 164–177.

Schollmeyer, K.G., Driver, J.C., 2013. Settlement patterns, source-sink dynamics, and artiodactyl hunting in the prehistoric U.S. Southwest. Journal of Archaeological Method and Theory 20, 448–478.

Smith, E.A., Winterhalder, B., 1992. Evolutionary Ecology and Human Behavior. Aldine de Gruyter, New York.

Steadman, D.W., 2006. Extinction and Biogeography of Tropical Pacific Birds. University of Chicago Press, Chicago.

Stiner, M.C., Munro, N.D., 2011. On the evolution of diet and landscape during the Upper Paleolithic through Mesolithic at Franchthi Cave (Peloponnese, Greece). Journal of Human Evolution 60, 618–636.

Valentine, K., Duffield, D.A., Patrick, L.E., Hatch, D.R., Butler, V.L., Hall, R.L., Lehman, N., 2008. Ancient DNA reveals genotypic relationships among Oregon populations of sea otter (*Enhydra lutris*). Conservation Genetics 9, 933–938.

Winterhalder, B., Smith, E.A., 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five years. Evolutionary Anthropology 9, 51–72.

Wolverton, S., Lyman, R.L., 2012. Conservation Biology and Applied Zooarchaeology. University of Arizona Press, Tucson.

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITED BY

JAMES D. WRIGHT

International Encyclopedia of the
SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

This page intentionally left blank

INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES

SECOND EDITION

EDITOR-IN-CHIEF
JAMES D. WRIGHT
University of Central Florida, Orlando, FL, USA

VOLUME 26

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015 Elsevier Ltd. unless otherwise stated. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein, Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalog record for this book is available from the British Library

Library of Congress Catalog Number:
A catalog record for this book is available from the Library of Congress

ISBN (print): 978-0-08-097086-8

For information on all Elsevier publications visit our website at store.elsevier.com

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

Publisher: Lisa Tickner
Acquisition Editor: Lisa Tickner
Content Project Managers: Claire Byrne, Michael Nicholls, Gemma Tomalin, Mark Harper
Designer: Maria Ines Cruz

Printed and bound in United States of America

CONTENTS TO VOLUME 26

Editorial Board *vii*

Subject Classification Index *1*

Subject Index *37*

This page intentionally left blank

EDITORIAL BOARD

Editor-in-Chief

James D. Wright
University of Central Florida, Orlando, FL, USA

Senior Editorial Associates

Amy M. Donley
University of Central Florida, Orlando, FL, USA

Donald C. Barrett
California State University, San Marcos, CA, USA

Area Editors

John Berry
Queen's University, Kingston, Canada; and Higher School of Economics (HSE), Moscow, Russia
Clinical Psychology
Applied, Industrial, and Organizational Psychology
Development Psychology
Social Psychology
Personality Psychology
Motivational Psychology
Psychiatry
Social Work

Stefan Ecks
University of Edinburgh, Edinburgh, UK
Ethics of Research
Anthropology
Archaeology
Health

Guillermina Jasso
New York University, New York, NY, USA
Demography
Sociology
Studies of the Life Course
Sexuality

Kenneth C. Land
Duke University, Durham, NC, USA
Institutions and Infrastructure of the Social and Behavioral Sciences
Statistics
Mathematics and Computer Sciences Applications
Logic of inquiry, Data Bases, and Research Design
Criminology

Barbara Prainsack
Kings College London, London, UK
Evolutionary Sciences
Genetics, Behavior, History, and Society
Religious Studies
Science and Technology Studies
Law

Peter Schmidt
Justus-Liebig-Universität Giessen, Giessen, Germany; and Higher School of Economics (HSE) Moscow, Russia
Environmental and Ecological Sciences
Education
Political Science
Labor Studies
Media Studies and Mass Communication
War, Peace, Violence, and Conflict

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, Scotland, UK
Biographies
History
Philosophy
Contemporary Cultural Concerns

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Memory: Cognitive and Neuroscientific Aspects
Linguistics A
Linguistics B
Neuroscience of Language
Behavioral Neuroscience
Cognitive Neuroscience
Cognitive Psychology

James Wright
University of Central Florida, Orlando, FL, USA
Area, Development, and International Studies
Gay, Lesbian, Bisexual, and Trans-sexual Studies
History of the Social and Behavioral Sciences
Applied Social and Behavioral Sciences

Henry Wai-Chung Yeung
National University of Singapore, Singapore
Economics
Geography
Management, Organizations, Business, Marketing, and Finance
Urban Studies and Planning
Public Policy

Section Editors

Helmut Anheier
Hertie School of Governance, Berlin, Germany
Institutions and Infrastructure of the Social and Behavioral Sciences

Neal M. Ashkanasy
The University of Queensland, Brisbane, QLD, Australia
Applied, Industrial, and Organizational Psychology

Sarah B. Barber
University of Central Florida, Orlando, FL, USA
Archaeology

Donald C. Barrett
California State University, San Marcos, CA, USA
Biographies
Gay, Lesbian, Bisexual, and Trans-sexual Studies

Dinesh Bhugra
Kings College London, London, UK
Psychiatry

Tora Bikson[†]
RAND, Santa Monica, CA, USA
Ethics of Research

Phillip Bonacich
University of California, Los Angeles, Los Angeles, CA, USA
Mathematics and Computer Sciences
Applications

Dominic Boyer
Rice University, Houston, TX, USA
Anthropology

Peter Bryant
IE Business School, Madrid, Spain
Management Organizations, Business, Marketing, and Finance

Graciela Cabana
University of Tennesee, Knoxville, TN, USA
Genetics, Behavior, History and Society

Ingrid Callies
Leem, Paris, France
Ethics of Research

Stefano Cappa
Vita-Salute San Raffaele University, Milano, Italy
Cognitive Neuroscience

Xenia Chryssochoou
Panteion University of Social and Political Sciences, Athens, Greece
Social Psychology

Harold Clarke
University of Texas, Dallas, TX, USA
Political Science

[†] Deceased.

Henri Cohen
Institute of Cognitive Science, University of Quebec, Montreal, QC, Canada
Cognitive Psychology

Daina Crafa
McGill University, Montreal, QC, Canada
Developmental Psychology

Lena Dominelli
Durham University, Durham, UK
Social Work

Tosha Dupras
University of Central Florida, Orlando, FL, USA
Archaeology

Jacquelynne Eccles
University of Michigan, Ann Arbor, MI, USA
Motivational Psychology

Irma T. Elo
University of Pennsylvania, Philadelphia, PA, USA
Demography

Dafna Feinholz
Social and Human Sciences Sector, UNESCO, Paris, France
Ethics of Research

Marina Fischer-Kowalski
Alpen-Adria University, Vienna, Austria
Environmental and Ecological Sciences

Christian Fleck
University of Graz, Graz, Austria
History of the Social and Behavioral Sciences

Andrew D. Foster
Brown University, Providence, RI, USA
Demography

Qiang Fu
Duke University, Durham, NC, USA
Statistics

Rosann Greenspan
University of California, Berkeley, CA, USA
Law

Xin Guo
The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
Statistics

Rachel Hammersley
Newcastle University, Newcastle, UK
History

Ulf Hannerz
Stockholm University, Stockholm, Sweden
Anthropology

Susan Hanson
Clark University, Worcester, MA, USA
Geography

Jeffrey K. Hass
Faculty of Economics, St. Petersburg State University, St. Petersburg, Russia; and University of Richmond, Richmond, VA, USA
Area, Development, and International Studies

Anke Hassel
Hertie School of Governance, Berlin, Germany
Public Policy

Andreas Hess
University College, Dublin, Ireland
Biographies

Gonia Jarema
Institut Universitatire de Gériatrie de Montréal, Montréal, QC, Canada
Linguistics B

Heidi Keller
University of Osnabrueck, Osnabrueck, Germany
Developmental Psychology

Bryan Kolb
University of Lethbridge, Lethbridge, AB, Canada
Behavioral Neuroscience

Kay Levine
Emory University, Atlanta, GA, USA
Law

Michael Lynch
Cornell University, Ithaca, NY, USA
Science and Technology Studies

Sinisa Malesevic
University College, Dublin, Ireland
War, Peace, Violence, and Conflict

Chrysostomos Mantzavinos
University of Athens, Athens, Greece
Philosophy

Jean-Christophe Marcel
Université Paris IV, Paris, France
Biographies

Douglas S. Massey
Princeton University, Princeton, NJ, USA
Sociology

John A. Mathews
Macquarie University, Sydney, NSW, Australia; and LUISS Guido Carli University, Rome, Italy
Management Organizations, Business, Marketing, and Finance

Catherine McBride
Chinese University of Hong Kong, Hong Kong, China
Education

Gustavo Mesch
University of Haifa, Haifa, Israel
Media Studies and Mass Communication

Alex Mesoudi
Durham University, Durham, UK
Evolutionary Sciences

Barbara Miller
George Washington University, Washington, DC, USA
Contemporary Cultural Concerns

Melinda Mills
University of Oxford, Oxford, UK
Studies of the Life Course

Wolfgang H.R. Miltner
Friedrich Schiller University, Jena, Germany
Clinical Psychology

Stephen L. Morgan
Johns Hopkins University, Baltimore, MD USA
Logic of Inquiry, and Research Design

Daniel Muijs
University of Southampton, Southampton, UK
Education

Thomas Nechyba
Duke University, Durham, NC, USA
Economics

Karen Parker
University of Delaware, Newark, DE, USA
Criminology

Henrike Rau
National University of Ireland, Galway, Ireland
Environmental and Ecological Sciences

Kees van Rees
Erasmus University, Rotterdam, The Netherlands
Culture and the Arts

Richard D. Roberts
Professional Examination Service, New York, NY, USA
Personality Psychology

Debra J. Rog
Westat, Rockville, MD, USA
Applied Social and Behavioral Sciences

Sergio Della Sala
University of Edinburgh, Edinburgh, UK
Memory: Cognitive and Neuroscientific Aspects

Katariina Salmela-Aro
University of Helsinki and Jyvädkylä, Helsinki, Finland
Motivational Psychology

Yoshimichi Sato
Tohoku University, Sendai, Japan
Sociology

Darren Sherkat
Southern Illinois University, Carbondale, IL, USA
Religious Studies

James D. Sidaway
National University of Singapore, Singapore
Geography

Johannes Siegrist
University of Duesseldorf, Duesseldorf, Germany
Health

Marianne Stewart
University of Texas, Dallas, TX, USA
Political Science

Claus Vögele
University of Luxembourg, Luxembourg
Health

Linda Waite
University of Chicago, Chicago, IL, USA
Sexuality

William Shi Yuan Wang
Chinese University of Hong Kong, Hong Kong, China; and University of California at Berkeley, USA
Linguistics A

Richard Whatmore
University of Sussex, Sussex, UK; and University of St. Andrews, St. Andrews, Scotland, UK

Harry Whitaker
Northern Michigan University, Marquette, MI, USA
Neuroscience of Language

Fulong Wu
University College London, London, UK
Urban Studies and Planning

Henry Yeung
National University of Singapore, Singapore
Economics

Karl Zimmerer
Pennsylvania State University, Pittsburgh, PA, USA
Environmental and Ecological Sciences

Klaus F. Zimmermann
Institute for the Study of Labor (IZA) and Bonn University, Bonn, Germany
Labor Studies

This page intentionally left blank

SUBJECT CLASSIFICATION INDEX

Anthropology

Advocacy in Anthropology
Affect and Emotion, Anthropology of
Africa, Sociocultural Overviews: West Africa
Age: Anthropological Aspects
Americas, Sociocultural Overviews: Central America
Americas, Sociocultural Overviews: Mexico
Americas, Sociocultural Overviews: North America, Native Americans
Americas, Sociocultural Overviews: Caribbean
Anarchism
Ancestors, Anthropology of
Anthropological Research, Ethics of
Anthropological Writing
Anthropology and History
Anthropology at Home
Anthropology: Overview
Apprenticeship, Anthropological Aspects
Arctic: Sociocultural Aspects
Art, Anthropological Aspects of
Asia, Sociocultural Aspects: China
Asia, Sociocultural Overviews: Korea
Asia, Sociocultural Overviews: South Asia
Australia, Sociocultural Overviews: Australian Settler Society
Australian Aboriginal Society and Culture: An Overview
Authenticity, Anthropology of
Belief, Anthropology of
Big Man, Anthropology of
Birth, Anthropology of
Body: Anthropological aspects
Borders, Anthropology of
Brokers and Brokerage, Anthropology of
Business Anthropology
Cannibalism
Cargo Cults
Carnival
Caste
Center-Periphery Relationships
Central Africa: Sociocultural Aspects
Central Asia, Anthropology of
Childhood: Anthropological Aspects
Citizenship, Anthropology of
Civilizations
Clan
Cognitive Anthropology
Collective Memory, Anthropology of
Colonialism, Anthropology of
Color Classification and Symbolism
Common Sense, Anthropology of
Communalism
Communes, Anthropology of
Community Studies: Anthropological
Comparative Method in Anthropology
Conflict: Anthropological Aspects
Cosmopolitanism
Craft Production, Anthropology of
Creolization: Sociocultural Aspects
Cultural Critique: Anthropological
Cultural Heritage
Cultural Relativism, Anthropology of
Culture Shock
Culture: Contemporary Views
Dance, Anthropology of
Death, Anthropology of
Development: Social-Anthropological Aspects
Diaspora
Diffusion: Anthropological Aspects
Discourse, Anthropology of
Drinking, Anthropology of
East Africa: Sociocultural Aspects
Economic Anthropology
Ecumene
Education: Anthropological Aspects
Elites, Anthropology of
Energy, Anthropology of
Ethnicity: Anthropological Aspects
Ethnocentrism
Ethnogenesis in anthropology
Ethnography
Ethnohistory
Ethnology
Europe: Sociocultural Overview
Everyday Life, Anthropology of
Exchange in Anthropology
Expertise, Anthropological Approaches to
Factionalism
Family: The Anthropology of the Concept and its History
Feminism in Anthropology
Fetishism
Fieldwork in Social and Cultural Anthropology
Finance, Anthropology of
Fishing Societies
Folklore
Food in Anthropology
Friendship, Anthropology of
Functionalism in Anthropology
Gender: Anthropological Aspects
Genealogy in Anthropology
Generation in Anthropology
Ghetto
Global South
Globalization, Anthropology of
Governmentality
Hazards and Disaster Research in Contemporary Anthropology
Health: Anthropological Aspects
Hegemony: Anthropological Aspects
Home, Anthropology of
Honor and Shame
Horticultural Societies
Households in Anthropology
Human Rights, Anthropology of
Humor, Anthropology of
Hunting and Gathering Societies: Anthropology
Hybridity
Identity in Anthropology
Indigeneity
Industrial Societies
Inequality: Comparative Aspects
Informal Sector
Infrastructures of Society, Anthropology of
Intercultural Communication: An Anthropological Perspective
Internet and Social Media: Anthropological Aspects
Interpretation in Anthropology
Irrigation Societies
Japan: Sociocultural Aspects
Kinship in Anthropology
Knowledge, Anthropology of
Kula Ring, Anthropology of
Land Tenure
Law: Anthropological Aspects
Liberalism, Anthropology of
Liminality
Mafia, Anthropology of
Magic, Anthropology of
Markets: Anthropological Aspects
Maroons in Anthropology
Marriage
Materiality and culture

Matrifocality
Middle East, North Africa: Sociocultural Overview
Migration: Anthropological Perspectives
Militaries and Militarization, Anthropology of
Modernity: Anthropological Aspects
Modes of Production
Money: Anthropological Aspects
Multiculturalism, Anthropology of
Multisited Field Studies
Music: Anthropological Aspects
Names and Naming
National Character
Nationalism: General
Navigation
Nongovernmental Organizations: Anthropological and Historical Aspects
North America: Sociocultural Aspects
Nostalgia, Anthropology of
Oceania, Sociocultural Overviews: Melanesia
Oceania, Sociocultural Overviews: Polynesia and Micronesia
Organizational Culture, Anthropology of
Orientalism
Pastoralism in Anthropology
Patron-Client Relations, Social and Anthropological Study of
Peasants in Anthropology
Performance: Anthropological Aspects
Personhood, Anthropology of
Play, Anthropology of
Political Anthropology
Political Ecology: Anthropological Perspectives
Political Economy in Anthropology
Popular Culture
Postcoloniality
Postindustrial Societies
Potlatch in Anthropology
Poverty, Culture of
Practice: An Anthropological Approach
Primitive Society
Property: Anthropological Aspects
Psychological Anthropology
Queer Anthropology
Race and Racism in the Twenty-First Century
Reflexivity in Anthropology
Refugees in anthropology
Religion, Anthropology of
Reproduction and Cultural Anthropology
Ritual
Roma People
Sacrifice
Science and Technology, Anthropology of
Secrecy, Anthropology of
Senses, Anthropology of the
Shamanism
Slavery: Comparative Aspects
Socialist Societies: Anthropological Aspects
Sociality In Anthropology
South America: Sociocultural Aspects
Southeast Asia: Sociocultural Aspects
Southern Africa: Sociocultural Aspects
Spirit Possession, Anthropology of
Sport, Anthropology of
State: Anthropological Aspects
Structuralism
Symbolism in Anthropology
Syncretism
Taboo
Third World
Time, Anthropology of
Totemism
Tourism
Tradition, Anthropology of
Tribe
Trickster Ethnography
Urban Anthropology
Values, Anthropology of
Violence in Anthropology
War: Anthropological Aspects, Historical Development of
Witchcraft
Work, Anthropology of
Worldview
Xenophobia and Nativism
Youth Culture, Anthropology of

Applied Social and Behavioral Sciences

Alcohol Interventions: Disease Model vs. Harm Reduction
Applied Criminology and Criminal Justice
Applied Demography
Applied Sociology
Case Studies
Cost-Benefit and Cost Effectiveness Analyses in Evaluation Research
Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices
Data Bases and Statistical Systems: Applied Social Research
Empowerment Evaluation
Ethnography in Applied Social Research
Evidence-Based Practice
Experimental Design: Randomization and Social Experiments
Field Experiments
Focus Groups
Forecasting
Gender and Women's Studies, applied research on
Homelessness in the United States
Immigration, Applied Research on
Implementation Science
Intimate Partner Abuse, Applied Research on
Logic Models
Mixed Methods
Needs Assessment
Participatory Action Research in Social Research
Policy Analysis
Prevention Research
Propensity Scoring
Qualitative Research
Quasi-Experimental Designs
Realism, in Applied Research
Social Experiments
Social Network Analysis
Societal Impact Assessment
Survival Analysis: Whether, and If So, When?
Systematic Reviewing and Meta-Analysis

Applied, Industrial, and Organizational Psychology

Climate, Justice, and Power

Affirmative Action, Sociology of
Affirmative Action: Comparative Policies and Controversies
Affirmative Action: Empirical Work on its Effectiveness
Corporate Culture
Gender and Women's Studies, applied research on
Negotiation and Conflict, The Psychology of
Organizational Climate in the Work Setting
Organizational Justice
Power, Politics, and Influence in Organizations
Subjective Wellbeing and Culture

Emotions, Cognition, and Personality

Agreeableness
Authoritarianism
Conscientiousness
Coping, Personality and
Core Self-Evaluations
Dark Triad: The "Dark Side" of Human Personality
Emotion in Cognition
Emotion, Perception and Expression of
Emotional Intelligence and Competencies
Emotional Regulation
Emotions and Work
Extraversion
Five Factor Model of Personality, Assessment of
Five Factor Model of Personality, Facets of
Five Factor Model of Personality, Universality of
Genetic Factors in Organizational Psychology
Managerial and Organizational Cognition
Neuroticism
Openness to Experience
Personality and Creativity
Personality and Economics
Personality and Values at Work
Personality Assessment, Faking and
Personality Assessment, Forced-Choice
Personality Assessment: Overview
Self-Efficacy
Self-Esteem
Stress: Psychological Perspectives
Subjective Wellbeing, Psychology of
Workplace Stress

Industrial-Organizational Psychology: General

Consumer Psychology
Cross-Cultural Training in Industrial and Organizational Psychology
Diversity in organizations
Indigenous Management Styles

Industrial–Organizational Psychology: Science and Practice
Peace Promotion, Psychology of
Program Evaluation
Technological Change and the Sociotechnical System, Applied Psychology of
Transportation, Psychology of
Volunteerism, Psychology of
Absorptive Capacity (of Organizations)
Administration in Organizations
Agency Theory
Behavioral Theories of Organization
Big Five Factor Model, Theory and Structure
Corporate Governance
Customer Satisfaction
Development: Organizational
Ethical Codes, Professional: Business Codes
Experimental Design: Randomization and Social Experiments
Family Business
Field Experiments
Greening of Business
Human Resource Management, Psychology of
Information and Knowledge, Organizational
Interorganizational Relationships and Networks
Job Design and Evaluation: Organizational Aspects
Learning: Organizational
Modern Management: Origins and Development
Myths and Symbols: Organizational
Narrative Approaches to Organizations
Network Analysis
Network Paradigm: Applications in Organizational Science
Organizational Behavior, Psychology of
Organizational Climate
Organizational Control
Organizational Decision Making
Organizational Errors: Catastrophic
People in Organizations
Qualitative Research
Quasi-Experimental Designs
Rational Choice and Organization Theory
Research and Development in Organizations
Social Business
Social Capital in Organizations
Social Innovation
Social Network Analysis
Strategic Management
Strategizing
Stress at Work
Structural Contingency Theory
Systematic Reviewing and Meta-Analysis
Wellbeing and Burnout in the Workplace, Psychology of
Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences
Wisdom, Psychology of
Workplace Autonomy
Workplace Safety and Health
Realism, in Applied Research

Leadership and Teams
Authentic Leadership
Group Processes in Organizations
Leader–Member Exchange Theory
Leadership in Organizations, Sociology of
Supervision, Abusive
Teamwork and Team Performance Measurement

Performance and Satisfaction
Behavior Analysis, Applied
Cultural Intelligence and Competencies
Human Factors and Ergonomics
Innovation
Job Analysis and Work Roles, Psychology of
Job Satisfaction
Money and Finances, Psychology of
Organizational Citizenship Behavior
Organizational Commitment
Pay, Compensation and Performance, Psychology of
Pro-Environmental Behavior
Shadow Work: Measurement
Work Motivation
Workplace Deviance

Personnel and Careers
Aging and Work
Career Development: Multilevel Perspective
Performance Appraisal and Evaluation
Personnel Selection, Psychology of
Professions in Organizations
Promotions
Psychological Contracts
Unemployment: Structural
Wage Differential and Structure

Vocational Training
Vocational Training

Archaeology

Chiefdoms, Archaeology of
Civilizations
Conflict and War, Archaeology of: Behavior and Social Organization
Conflict and War, Archaeology of: Weapons and Artifacts
Cultural Heritage
Food Production, Origins of
Historical Archaeology
Household Archaeology
Human Language, Evolution of
Hunter–Gatherer Societies, Archaeology of
Intensification and Specialization, Archaeology of
Language Families, Archaeology and History of
Migrations, Colonizations, and Diasporas in Archaeology
Prehistoric Art
Religion: Evolution and Development
Religious Belief, Evolution of
Ritual and Symbolism, Archaeology of
Settlement and Landscape Archaeology
States and Civilizations, Archaeology of
Trade and Exchange, Archaeology of
Writing, Evolution of
Museums

Research Methods
Archaeometry
Bioarchaeology
Ceramics in Archaeology
Chronology, Stratigraphy, and Dating Methods in Archaeology
Cultural Resource Management: Conservation of Cultural Heritage
Environmental Archaeology
Ethnoarchaeology
Forensic Archaeology
Geoarchaeology
Geospatial Applications in Archaeology
Lithics and Archaeology
Paleoethnobotany
Survey and Excavation (Field Methods) in Archaeology
Zooarchaeology

Theoretical Perspectives
Archaeology and Cultural Memory
Archaeology, Politics of
Archaeology, Theory in
Archaeology: Philosophy and Science
Civilization, Concept and History of
Cognitive Archaeology
Cultural Evolution: Overview
Cultural Evolution: Theory and Models
Diffusion: Anthropological Aspects
Ethnic Identity and Ethnicity in Archaeology
Evolution, Natural and Social: Philosophical Aspects
Gender, Feminism, and Sexuality in Archaeological Studies
Marxist Archaeology
State Formation, Theory of
State, History of

World Regions
Ancient Egypt, Archaeology of
Ancient Greece and Rome, History of
Arctic Archaeology and Prehistory
Australia and New Guinea, Archaeology of
Classical Archaeology
Europe, Archaeology of
Mesoamerica, Archaeology of
Near Middle East/North Africa Studies: Society and History
North America, Archaeology of
Pacific Islands, Archaeology of
South America, Archaeology of
South Asia, Archaeology of
Southeast Asia, Archaeology of

Area, Development and International Studies

Africa's Environment: A Stressed Biogeographical and Cultural Landscape
African Legal Systems
African Studies: Culture
African Studies: Health
African Studies: History
Agrarian Political Economy

American Studies, Twentieth Century History of
Area and International Studies: Cultural Studies
Area and International Studies: Political Economy
Area and International Studies: Stakeholders
Development and the State
Development and Urbanization
East Asian Studies: Culture
East Asian Studies: Economics
East Asian Studies: Gender
East Asian Studies: Politics
Eastern European Studies: Culture
Eastern European Studies: History
Eastern European Studies: Society
Ethical Turn in International Development
Global Development: Markets and Wellbeing
Near Middle East/North Africa Studies: Culture
Near Middle East/North Africa Studies: Gender
Near Middle East/North Africa Studies: Politics
Near Middle East/North Africa Studies: Religion
Near Middle East/North Africa Studies: Society and History
Peasantry in the 21st Century
Post-socialism: Political Economy
Russian and Post-Soviet Studies: Culture
Russian and Post-Soviet Studies: Education
Russian and Post-Soviet Studies: Environment
Russian and Post-Soviet Studies: Gender
Russian and Post-Soviet Studies: Politics
Russian and Post-Soviet Studies: Society
Sustainable Agriculture

Behavioral Neuroscience

Alzheimer's Disease, Neural Basis of
Anxiety and Fear Conditioning, Neural Basis of
Attention, Neural Basis of
Attention-Deficit and Hyperactivity Disorders, Neural Basis of
Brain Damage: Neuropsychological Rehabilitation
Brain Development: Ontogenetic Neurobiology of
Brain Stimulation Reward
Brain: Response to Enrichment
Cerebral Cortex
Cognitive Development in Infancy, Neural Mechanisms of
Computational Neuroscience
Cortical Areas Engaged in Movement: Neuroimaging Methods
Cortical Development: Transplantation and Rewiring Studies
Dysexecutive Syndromes
Electroencephalography: Basic Principles and Applications
Emotion, Neural Basis of
Epigenetics of brain plasticity and behavior
Epilepsy
Eye Movement, Control of: Oculomotor control
Eyeblink Classical Conditioning
Fear Conditioning: Overview
Fear: Psychological and Neural Aspects
Fear-Potentiated Startle
Food Preferences, Psychology and Physiology of
Habituation
Hunger and Eating, Neural Basis of
Hypnosis, Neural Basis of
Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of
Hypothalamus
Ion Channels and Molecular Events in Neuronal Activity
Kindling
Limbic System
Long-term Potentiation and Long-term Depression
Motor Cortex
Near-Infrared Spectroscopy
Neural Plasticity
Neural Representations of Direction (Head Direction Cells)
Neural Synchrony as a Binding Mechanism
Neuroendocrinology
Neurons and Dendrites, Integration of Information in
Neurotransmitters
Occipital Lobe (Visual Cortex): Functional Aspects
Olfactory System
Optogenetics
Parietal Lobe
Prefrontal Cortex
Prefrontal Cortex Development and Development of Cognitive Function
Premotor Cortex
Recovery of Function: Dependency on Age
Sleep: Neural Systems
Taste and Smell, Psychology of
Temporal Lobe
Topographic Maps
Vestibular System
Vestibulo-ocular Reflex, Adaptation of the
Visual Streams: Dorsal and Ventral
Williams Syndrome

Biographies

Addams, Jane (1860–1935)
Albert, Hans (1921–)
Alexander, Jeffrey C. (1947–)
Allport, Gordon W (1897–1967)
Annales School
Arendt, Hannah (1906–75)
Aron, Raymond (1905–83)
Baltes, Paul B (1939–2006)
Barthes, Roland (1915–80)
Beauvoir, Simone de (1908–86)
Beck, Ulrich (1944–)
Becker, Howard (1928–)
Bell, Daniel (1919–2011)
Bellah, Robert N (1927–2013)
Bendix, Reinhard (1916–91)
Benedict, Ruth (1887–1948)
Benhabib, Seyla (1950–)
Bentham, Jeremy (1748–1832)
Berlin, Isaiah (1909–97)
Bielefeld School of History
Binet, Alfred (1857–1911)
Bleuler, Eugen (1857–1939)
Bloch, Mark Leopold Benjamin (1886–1944)
Bloomfield, Leonard (1887–1949)
Boas, Franz (1858–1942)
Bobbio, Norberto (1909–2004)
Boltanski, Luc (1940–)
Boserup, Ester (1910–99)
Bourdieu, Pierre (1930–2002)
Bowlby, John (1907–90)
Broadbent, Donald Eric (1926–93)
Bruner, Jerome Seymour (1915–)
Burckhardt, Jacob (1818–97)
Burke, Peter (1937–)
Butler, Judith (1956–)
Cambridge School of Intellectual History
Campbell, Donald Thomas (1916–96)
Cardoso, Henrique (1931–)
Caro Baroja, Julio (1914–95)
Castells, Manuel (1942–)
Cattell, Raymond Bernard (1905–98)
Chomsky, Noam (1928–)
Coleman, James Samuel (1926–95)
Collins, Randall (1941–)
Communitarians
Comte, Auguste (1798–1857)
Coser, Lewis A. (1913–2003)
Dahl, Robert A (1915–2014)
Dahrendorf, Ralf (1929–2009)
Darwin, Charles Robert (1809–82)
de Finetti, Bruno (1906–85)
Deutsch, Karl W (1912–92)
Dewey, John (1859–1952)
Dilthey, Wilhelm (1833–1911)
Drucker, Peter (1909–2005)
Du Bois, William Edward Burghardt (1868–1963)
Dubos, René Jules (1901–82)
Durkheim, Emile (1858–1917)
Durkheimians, The
Edgeworth, Francis Ysidro (1845–1926)
Eisenstadt, Shmuel N. (1923–2010)
Eliade, Mircea (1907–86)
Elias, Norbert (1897–1990)
Elster, Jon (1940–)
Erikson, Erik Homburger (1902–94)
Evans-Pritchard, Sir Edward E (1902–73)
Feyerabend, Paul (1924–94)
Fisher, Irving (1867–1947)
Fisher, Ronald A (1890–1962)
Foucault, Michel (1926–84)
Frankfort School: Institute for Social Research
Freud, Sigmund (1856–1939)
Freyre, Gilberto (1900–87)
Fromm, Erich (1900–80)
Galton, Sir Francis (1822–1911)
Garfinkel, Harold (1917–2011)

Gauss, Carl Friedrich (1777–1855)
Geertz, Clifford (1926–2006)
Gellner, Ernest (1925–95)
Giddens, Anthony (1938–)
Goffman, Erving (1922–1982)
Goldthorpe, John H. (1935–)
Gorz, Andre (1923–2007)
Habermas, Jürgen (1929–)
Halévy, Élie (1870–1937)
Hall, Granville Stanley (1844–1924)
Harlow, Harry Frederick (1905–81)
Hart, Herbert Lionel Adolphus (1907–92)
Hebb, Donald Olding (1904–1985)
Hegel, Georg Wilhelm Friedrich (1770–1831)
Heider, Fritz (1896–1988)
Helmholtz, Hermann Ludwig Ferdinand von (1821–94)
Hempel, Carl Gustav (1905–97)
Henry, Louis (1911–91)
Hilberg, Raul (1926–2007)
Hintze, Otto (1861–1940)
Hirschman, Albert O. (1915–2012)
Hochschild, Arlie (1940–)
Honneth, Axel (1948–)
Hotelling, Harold (1895–1973)
Hume, David (1711–76)
Huntington Samuel P. (1927–2008)
Hurst, James Willard (1910–97)
Husserl, Edmund (1859–1938)
Illich, Ivan (1926–2002)
Jackson, John Hughlings (1835–1911)
Jacobs, Jane (1916–2006)
Jakobson, Roman (1896–1982)
James, William (1842–1910)
Janet, Pierre (1859–1947)
Jeffreys, Sir Harold (1891–1989)
Jung, Carl Gustav (1875–1961)
Kant, Immanuel (1724–1804)
Key, Valdimer Orlando (1908–63)
Keynes, John Maynard (1883–1946)
Kimura, Motoo (1924–94)
Klein, Melanie (1882–1960)
Kohlberg, Lawrence (1927–87)
Köhler, Wolfgang (1887–1967)
Kraepelin, Emil (1856–1926)
Kuhn, Thomas S. (1922–96)
Lacan, Jacques (1901–81)
Laplace, Pierre Simon, Marquis de (1749–1827)
Lashley, Karl Spencer (1890–1958)
Latour, Bruno (1947–)
Lazarsfeld, Paul Felix (1901–76)
Levi-Strauss, Claude (1908–2009)
Lewin, Kurt (1890–1947)
Llewellyn, Karl Nickerson (1893–1962)
Lorenz, Konrad (1903–89)
Lotka, Alfred James (1880–1949)
Luhmann, Niklas (1927–98)
Luria, Aleksandr Romanovich (1902–77)
Malinowski, Bronislaw (1884–1942)
Malthus, Thomas Robert (1766–1834)
Mann, Michael (1942–)
Mannheim, Karl (1893–1947)
Marr, David (1945–80)
Marshall, Alfred (1842–1924)
Marshall, Thomas Humphrey (1893–1981)
Maruyama, Masao (1914–96)
Marx, Karl (1818–83)
Mauss, Marcel (1872–1950)
Mead, George Herbert (1863–1931)
Mead, Margaret (1901–78)
Memmi, Albert (1920–)
Merton, Robert K (1910–2003)
Mill, John Stuart (1806–73)
Mills, Charles Wright (1916–62)
Moore Jr , Barrington (1913–2005)
Muller, Hermann Joseph (1890–1967)
Mumford, Lewis (1895–1990)
Myrdal, Alva (1902–86)
Myrdal, Gunnar (1898–1987)
Needham, Joseph (1900–95)
Neumann, John von (1903–57)
Neyman, Jerzy (1894–1981)
Nietzsche, Friedrich (1844–1900)
Nisbet, Robert (1913–96)
Notestein, Frank Wallace (1902–83)
Olson, Mancur (1932–98)
Pareto, Vilfredo (1848–1923)
Parsons, Talcott (1902–79)
Pavlov, Ivan Petrovich (1849–1936)
Pearson, Karl (1857–1936)
Pestalozzi, Johann Heinrich (1746–1827)
Piaget, Jean (1896–1980)
Plato (428 BCE - 347 BCE)
Polanyi, Karl (1886–1964)
Popper, Karl Raimund (1902–94)
Putnam, Hilary (1926–)
Quetelet, Adolphe (1796–1874)
Rawls, John (1921–2002)
Ricardo, David (1772–1823)
Ricoeur, Paul (1913–2005)
Riesman, David (1909–2002)
Robinson, Joan (1903–83)
Rogers, Carl Ransom (1902–87)
Rokkan, Stein (1921–79)
Rorty, Richard (1931–2007)
Rousseau, Jean-Jacques (1712–78)
Sandel, Michael (1953–)
Sapir, Edward (1884–1939)
Sarmiento, Domingo Faustino (1811–88)
Sauer, Carl Ortwin (1889–1975)
Saussure, Ferdinand de (1857–1913)
Savage, Leonard J (1917–71)
Schmitt, Carl (1888– 1985)
Schumpeter, Joseph A (1883–1950)
Schütz, Alfred (1899–1959)
Searle, John (1932–)
Sherrington, Sir Charles Scott (1857–1952)
Shils, Edward (1910–95)
Shklar, Judith N (1928–92)
Simmel, Georg (1858–1918)
Simon, Herbert (1916–2001)
Skinner, Burrhus Frederick (1904–1990)
Skocpol, Theda (1947–)
Smith, Adam (1723–90)
Sorokin, Pitirim Alexandrovich (1889–1968)
Spencer, Herbert (1820–1903)
Sperry, Roger Walcott (1913–94)
Stevens, Stanley Smith (1906–73)
Stigler, George Joseph (1911–91)
Strauss, Anselm (1916–96)
Sumner, William (1840–1910)
Taylor, Charles (1931–)
Thorndike, Edward Lee (1874–1949)
Tocqueville, Alexis de (1805–1859)
Touraine, Alain (1925–)
Tversky, Amos (1937–96)
Veblen, Thorstein (1857–1929)
Vienna Circle: Logical Empiricism
von Hayek, Friedrich A. (1899–1992)
von Humboldt, Alexander (1769–1859)
von Humboldt, Wilhelm (1767–1835)
von Ranke, Leopold (1795–1886)
Vygotskij, Lev Semenovic (1896–1934)
Wallerstein, Immanuel (1930–)
Walzer, Michael (1935–)
Watson, John Broadus (1878–1958)
Weber, Max (1864–1920)
Whorf, Benjamin Lee (1897–1941)
Wilson, William Julius (1935–)
Wittgenstein, Ludwig (1889–1951)
Wright, Sewall (1889–1988)
Wundt, Wilhelm Maximilian (1832–1920)

Clinical Psychology

Eating Disorders
Psychopathy

Behavioral Change and Modification Principles and Models
Classical Conditioning Methods in Psychotherapy
Concepts and Methods of Cognitive Therapies
Constructivist Theory and Psychotherapy
Counseling and Psychotherapy: Ethnic and Cultural Differences
Counseling Psychology
Managed Care and Psychotherapy in the United States
Motivation in Clinical Interventions
Projective Methods in Psychology
Psychotherapy and Hermeneutics
Reinforcement, Principle of
Relapse Prevention Training
Therapist–Patient Relationship
Training in Clinical Psychology in the United States: Practitioner Model
Training in Clinical Psychology in the United States: Scientist–Practitioner Model
Virtual Reality in Psychotherapy

Clinical Disorders
Bereavement and Complicated Grief across the Lifespan
Clinical Psychology, Psychiatry and Homosexuality
Mass Trauma: Psychopathological Effects across the Life Span
Obsessive-Compulsive Disorder across the Life Span
Panic Disorder and Agoraphobia Across the Lifespan

Pathological Gambling and Gambling Disorder
Phobias across the Lifespan
Psychodynamic Therapy: Short-Term
Sex Therapy, Clinical Psychology of
Sleep Disorders across the Life span
Social Phobia Across the Lifespan
Torture and its Consequences, Psychology of

Diagnosis and Assessment
Diagnostic Processes in Clinical Psychology and Psychotherapy
Epidemiology, Concepts and Methods of
Personality Assessment

Psychological Therapies
Behavior Therapy: Background, Basic Principles and Early History
Behavior Therapy: The Second and Third Waves
Experiential Psychotherapy
Exposure therapies and stress inoculation: A brief overview
Feminist Psychotherapy
Forensic Psychology
Group Psychotherapy, Clinical Psychology of
Hypnosis
Imaging Methods in Clinical Psychology
Imaging Methods in Psychotherapy
Intellectual Functioning, Assessment of
Interpersonal Psychotherapy
Multimodal Therapy in Clinical Psychology
Narrative Exposure Therapy

Research and Treatment Process and Outcome
Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues
Evidence-Based Psychotherapy
Manual Based Treatments
Psychological Treatment, Effectiveness of
Psychological Treatments Delivered Nontraditionally
Psychological Treatments, Empirically Supported
Psychological Treatments: Randomized Controlled Trials
Psychotherapy Integration
Psychotherapy Process Research

Cognitive Neuroscience

Methods
Bayesian Models in Neuroscience

Neuropsychology
Agnosia
Agraphia (Acquired Dysgraphia)
Alzheimer's Disease, Neural Basis of
Amnesia: General
Aphasia
Apraxia
Asymmetry of Body and Brain
Attention-Deficit and Hyperactivity Disorders, Neural Basis of
Autism spectrum disorders
Autobiographical Memory
Bilingual Aphasia
Brain Asymmetry
Brain Damage: Neuropsychological Rehabilitation
Cerebellum: Cognitive Functions
Cognitive Control (Executive Function): Role of Prefrontal Cortex
Confabulation
Consciousness, Neural Basis of
Declarative Memory, Neural Basis of
Dementia: Psychiatric Aspects
Dreaming, Neural Basis of
Dysexecutive amnesia
Dysexecutive Syndromes
Dyslexia, Developmental
Dyslexias, Acquired and Agraphia
Emotion, Neural Basis of
Face Recognition, Psychological and Neural Aspects
Hallucinations, Neural Basis of
Handedness and Language, Relationship between
Hemispatial Neglect, Neural Basis of
Hippocampal Amnesia
Human Spatial Orientation, Neural Basis of
Interhemispheric Interaction in Language and Cognitive Processes
Language Disorders, Treatment and Remediation of
Language Processes in Delirium and Dementia
Language Processing, Functional Magnetic Resonance Imaging of
Lateralization of Language as Demonstrated by Brain Imaging Procedures
Learning and Memory, Neural Basis of
Linguistic Approaches to Agrammatism
Luria, Aleksandr Romanovich (1902–77)
Navigation in Virtual Space: Psychological and Neural Aspects
Neural Basis of Language Development
Neural Basis of Reading
Neural Plasticity
Neural Representations of Direction (Head Direction Cells)
Neural Representations of Objects
Neurolinguistic Processing of Psychological Verbs
Neurolinguistics, History of
Neurological Approaches to Agrammatism
Phantom Limbs
Phlological Disorders
Post Traumatic Amnesia
Pragmatic Disorders
Prefrontal Cortex
Prefrontal Cortex Development and Development of Cognitive Function
Premotor Cortex
Prosopagnosia
Recovery of Function: Dependency on Age
Semantic Memory
Short-Term Memory: Psychological and Neural Aspects
Sign Language, Disorders of
Speech Production, Neural Basis of
Sperry, Roger Walcott (1913–94)
Split Brains and Human Consciousness
Visual Perception, Neural Basis of
Visual Word Recognition, Neurocognitive Psychology of
Working Memory, Neural Basis of
Working Memory, Psychology of
Aging and Memory

Social Neuroscience
Anxiety and Fear Conditioning, Neural Basis of
Brain Stimulation Reward
Cooperative Game Theory
Decision Making, Psychology of
Embodied Social Cognition
Empathy: A Social Neuroscience Perspective
Experimental Economics
Fear: Psychological and Neural Aspects
Neural Correlates of Social Development
Neural Foundation of Morality
Neuroeconomics
Neuroesthetics
Neuroethics: Ethics of Science and Science of Ethics
Neuromarketing
Neuroscience of Education
Psychopathy
Self and Brain
Social Neuroscience of the Twenty-First Century
Behavioral Economics

Theory and Methods
Cognitive Neuropsychology, Methods of
Cognitive Neuroscience
Computational Neuroscience
Cortical Areas Engaged in Movement: Neuroimaging Methods
Electroencephalography: Basic Principles and Applications
Embodiment Theory
Functional Brain Imaging of Language processes
Functional Magnetic Resonance Imagery, Analysis of
Intracarotid Sodium Amytal Test and Language
Magnetic Resonance and Molecular Imaging in Psychiatry
Mind, Theories of
Mirror Neurons, Theory of
Near-Infrared Spectroscopy
Neuropsychological Testing
Placebo Effect
Transcranial Direct Current Stimulation and Language

Cognitive Psychology

Cognitive Styles and Learning Styles
Decision Making, Psychology of
Inferences in Discourse, Psychology of
Numbers and Number Sense
Perception of Action Space: Using Multiple Frames of Reference

Problem Solving and Reasoning, Psychology of
Psycholinguistics: Overview
Psychology of Inferences
Sentence Comprehension, Psychology of
Serial Order, Representation of
Space, Linguistic Expression of
Tacit Knowledge, Psychology of

Cognitive Neuroscience
Face Recognition, Psychological and Neural Aspects
Mental Imagery: Visual Cognition
Multisensory Integration, Principles of
Navigation in Virtual Space: Psychological and Neural Aspects
Visual Word Recognition, Neurocognitive Psychology of

Developmental Psychology
Intentionality in Language and Communication, Emergence of
Scientific Concepts: Development in Children

Evolutionary Sciences
Human Cognition, Evolution of
Human Language: Evolutionary Precursors

Linguistics A; Linguistics: B
Prototype theory in linguistics

Linguistics: B
Language and Thought: The Neo-Whorfian Hypothesis
Logic and Linguistics
Syntax—Phonology Interface
Syntax—Semantics Interface

Contemporary Cultural Concerns

Alternative Food Movements
American Conservative Movement
Busyness
Cash Transfers
Child Soldiers
Children, Rights of: Cultural Concerns
China as a Rising Aid Donor
Climate Change and Global Poverty
Collaborative Consumption
Conflict: Education and Youth
Cultural Rights and Culture Defense: Cultural Concerns
Culture and Economic Development
Culture of Poverty: Critique
Cyberbullying
Deconstruction
Disposable Youth in America in the Age of Neoliberalism
Eating Disorders
Electronic Waste
End of Life Choices
Environmental Conservation and Development: Critical Perspectives
Ethical Consumers
Female Genital Mutilation
Female Quotas in Parliament: Do They Make a Difference?
Feminisms and Acculturation around the Globe
Food Deserts in US Cities
Gender and Language: Cultural Concerns
Global Violence Against Women
Globalization and Human Trafficking
Globalization and Transnational Corporations
Hazing in Fraternities and Sororities
Health in Low-Income Countries
Honor Killing: Multidimensional and Multilevel Perspectives
Human Terrain System (United States): Critique
Indigenous Knowledge and Intellectual Property Rights
Integration and Multiculturalism: Focus on Western Europe
Justice, Conflicts, and the Justice of Conflict Resolution
McDonaldization
Medicalization: Sociological and Anthropological Perspectives
Missing girls: A Globalizing Issue
Morality 'East' and 'West': Cultural Concerns
New Media, Political Mobilization, and the Arab Spring
Obesity: Cultural and Biological Factors
Online Pornography
Oral and Literate Expression
Oversharing: The Eclipse of Privacy in the Internet Age
Religious Nationalism and the Secular State
Sexual Politics in the Arab World
Slow Food as a Social Movement
Social Resilience in Post-Tsunami Japan
Street Children: Cultural Concerns
Surveillance Studies
Tattoos and Body Modification
Teens, Gender, and Self-Presentation in Social Media
Tourism, the Millennium Development Goals, and Gender Equality
Transnational Citizenship
Violence, Criminalization, and the Homeless
Water, Health and Social Inequality
Women's Status and National and International Security
Youth, Music, and Peace Building

Criminology

Age and Crime
Biosocial Approaches: Crime
Collective Efficacy and Crime
Communities and Crime
Community Supervision and Diversion in the United States
Control theories and Crime
Crime and Class
Crime and Gender
Crime in the Rural Context
Crime Trends and Debates
Crime, Desistance, and Persistence
Crime, Geography of
Crime, Race and Ethnicity
Crime, Sociology of
Crime: Knowledge about and Prevalence
Crime: Organized
Crime: Whitecollar
Criminal Justice, Sociology of
Criminology: Development and Current Status of the Discipline
Critical Criminology
Cyber Crime: Identity Theft
Data Bases and Statistical Systems: Crime
Data Bases and Statistical Systems: Crime Measurement
Delinquency, Sociology of
Deterrence Theory: Crime
Deterrence: Sanction Perceptions
Domestic Violence: Sociological Perspectives
Drug Control: International Aspects
Drugs, Decriminalization of: Cultural Concerns
Drugs: Illicit Use and Prevention
Economics and Crime
Fear of Crime
Feminist Criminology and Gender Studies
Gangs, Sociology of
Genocide: Social and Economic Aspects
Guilt
Guns and Criminals
Hate Crimes
Homicide
Immigration and Crime in the United States
Imprisonment: Sociological Aspects
International and Comparative Criminology
Intersectionality and Crime
Law and Society
Life Course Criminology
Mafias, Criminology of
Mass Imprisonment and Its Consequences
Media, Crime and Crime Waves
Misdemeanor Offenses
National Security and Domestic Terrorism: The US Case
Opiate Use and Abuse, History of
Personality and Crime
Police Use of Force
Police, Sociology of
Policing
Pornography
Prevention of Crime and Delinquency
Prison Life, Sociology of: Recent Perspectives from the United Kingdom
Prisons and Imprisonment
Property Crime: Victimization and Trends
Punishment
Punishment: Comparative Politics of
Quantitative Methods and Crime
Race: Prespectives and Crime
Rape and Sexual Coercion
Recidivism
Re-entry and Reintegration After Incarceration
Rehabilitation of the Offender
Repetitive Sex Offenders, Treatment of
Routine Activities, Criminal Opportunities, Crime and Crime Prevention

School Safety
Self-Control Theory and Crime
Sentencing, Trends and Disparities: Crime
Sex Crimes
Social Justice and Criminal Justice
Social Justice and Social Inequalities
Social Learning Theory and Crime
Spatial Dynamics and Crime
Strain Theories and Crime
Transnational Organized Crime
Victimless Crime
White Collar Crime: Recent Trends and Debates
Wrongful Convictions
Youth Gangs

Culture and the Arts

Art and Socialisation
Artists, Competition and Markets
Avant-garde Art and Artists
Classifications in Popular Music
Collective Identity
Commemoration
Cultural Capital and Education
Cultural Mediators and Gatekeepers
Cultural Participation, Trends in
Cultural Policies in East Asia
Cultural Policy Regimes in Western Europe
Cultural Production in Networks
Culture and Actor Network Theory
Culture and Economy
Culture and Institutional Logics
Culture and Networks
Culture Policy Regimes in China
Culture, Cognition and Embodiment
Culture, Production of
Culture, Production of: Prospects for the Twenty-First Century
Embodiment and Culture
Epistemic Cultures
Formal Methods of Cultural Analysis
Globalization and World Culture
Internet and Culture
Leisure and Cultural Consumption: The European Perspective
Leisure and Cultural Consumption: US Perspective
Markets: Artistic and Cultural
Miracles, Cultural Logic of
Narrative Networks
Nation-State as Symbolic Construct
Networks and Cultural Consumption
Networks and Meaning
Organizations and Culture
Social Inequality in Cultural Consumption Patterns
Symbolic Boundaries
Video Cultures: Television

Demography

Abortion, Demography of: United States
Adoption, Demography of
Adult Mortality in Developing Countries
Adult Mortality in Eastern Europe and the Former Soviet Republics
Adult Mortality in Industrialized Societies
Aging and Health in Old Age
Assimilation of Immigrants
Assortative Mating in the Marriage Market
Biomarkers in Demographic Research
Children, Value of
Cohabitation: United States
Conditional Cash Transfers: Influence on Marriage and Fertility
Demographic Impact of Disasters
Demographic Measurement: General Issues and Measures of Fertility
Demographic Measurement: Nuptiality, Mortality, Migration, and Growth
Demographic Models
Demographic Techniques: Data Adjustment and Correction
Demographic Techniques: Family Reconstitution
Demographic Techniques: Indirect Estimation
Demographic Techniques: Inverse Projection
Demographic Techniques: LEXIS Diagram
Demographic Techniques: Small-area Estimates and Projections
Demography: History Since 1900
Differential Aging
Dowries and Other Marital Transfers
Early Life Influences on Health and Mortality in Adulthood
Ethnic Intermarriage
Event History Analysis in Continuous Time
Evolutionary Demography: A Synthesis of two Population Sciences
Experimental Methods in Survey Research in Demography
Families and Households, Formal Demography of
Family Planning Programs: Feminist Perspectives
Family Planning: Program Effects
Family size preferences
Family theory: Competing perspectives in social demography
Family Theory: Economics of Childbearing
Family Theory: Economics of Intergenerational Relations
Family Theory: Economics of Marriage and Divorce
Fatherhood
Fertility and Culture: Anthropological Insights
Fertility and Social Networks
Fertility Change: Quantum and Tempo
Fertility of Single and Cohabiting Women
Fertility Theory: Embodied-capital Theory of Life History Evolution
Fertility Theory: Theory of Intergenerational Wealth Flows
Fertility Transition: China
Fertility Transition: Cultural Explanations
Fertility Transition: East Asia
Fertility Transition: Economic Explanations
Fertility Transition: Latin America and the Caribbean
Fertility Transition: Middle East and North Africa
Fertility Transition: South Asia
Fertility Transition: Southeast Asia
Fertility Transition: Sub-Saharan Africa
Fertility Trends in the Formerly Socialist Countries of Europe
Gender, Time-Use, and Fertility Recovery in Industrialized Countries
Gene-Environment Interplay
Hutterite Demography
Immigration: Health and Mortality
Infant and Child Mortality in Industrialized Countries
Infant and Child Mortality in the Less Developed World
Infertility: Demographic Aspects
Internal Migration in Industrialized Countries
Internal Migration: Developing Countries
International Migration
International Migration and Remittances
International Migration by Ethnic Chinese
International Migration by Ethnic South and Southeast Asians
Jewish Diaspora
Microsimulation in Demographic Research
Migrants and Migration in Europe
Migration and Fertility
Migration Out of Europe
Migration to and from Latin America
Migration, Theory of
Migration: Cultural Aspects
Mortality and the HIV/AIDS Epidemic
Mortality Crossover
Mortality Differentials: Selection and Causation
Mortality of the Oldest-Old
Mortality, Biodemography of
Mortality, Epidemiological, and Health Transitions
Mortality: The Great Historical Decline
Multilevel Models in Demography
Multistate Transition Models in Demography
Obesity and Mortality
Old Age: Definitions, Theory, and History of the Concept
Period and Cohort Analysis in Demography
Population Change in the Former Communist States of Europe and Asia
Population Composition by Race and Ethnicity: North America
Population Cycles, Formal Theory of
Population Dynamics: Classical Applications of Stable Population Theory
Population Dynamics: Mathematic Models of Population, Development, and Natural Resources
Population Dynamics: Momentum of Population Growth
Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems

Population Dynamics: Theory of Nonstable Populations
Population Dynamics: Theory of Stable Populations
Population Dynamics: Two-Sex Demographic Models
Population Forecasts
Repartnering and Stepchildren
Risk Screening, Testing, and Diagnosis: Ethical Aspects
Second Demographic Transition
Sex Selective Abortion
Sexuality: Cultural Aspects
Twin Studies in Demography

Developmental Psychology

Behavior and Personality Development
Behavioral Inhibition and Social Withdrawal across Cultures
Behavioral Problems, Effects of Parenting and Family Structure on
Behavioral Problems, Effects of Socio-Economic Status on
Defiant Behavior During Adolescence across Cultures
Defiant behavior during infancy and early childhood
Motivational Development, Systems Theory of
Motivational Development: Cross-Cultural Perspectives
Neonatal Temperament: From Phenomenology to Mechanism
Personality Changes during Adolescence across Cultures
Personality Development: Systems Theories
Self in Culture: Early Development
Self-Development in Childhood and Adolescence
Self-Regulation During Adolescence: Variations Associated with Individual - Context Relations
Self-Regulation During Early Childhood Across Cultures, Development of
Temperament and Human Development
Temperament Development, Theories of

Brain and Cognitive Development
Academic Performance, Effects of Socio-Economic Status on
Aging and Memory
Aging Mind: Facets and Levels of Analysis
Autobiographical Memory during Early Childhood and Cultural Variations
Breast-feeding: Effects on Cognitive and Neural Development
Childhood Memories
Cognitive Development during Infancy and Early Childhood across Cultures
Cognitive Development, Educational Theories of
Concept Formation in Childhood: Cross-Cultural Perspectives
Conscious Control during Childhood, Development of
Early Memory Development and Cultural Variations
Epistemic doubt during adolescence
Equivalence and Transfer Problems in Cross-Cultural Research
Executive Functions during Childhood, Development of
Face Perception Development During Early Childhood
Infant Gaze Behaviors
Life Management, Developmental Psychology of
Logical and Hypothetical Reasoning in Adolescence, Development of
Mathematical Concepts during Early Childhood Across Cultures, Development of
Memory Development during Infancy and Early Childhood across Cultures
Mirror Neurons, Theory of
Motor Skills: Development in Infancy and Early childhood
Neonatal Social Isolation: Effects on Neural Function
Problem Solving during Infancy and Early Childhood, Development of
School Achievement as Moderated by Sociodemographic Status: The US Case
Scientific Concepts during Childhood, Development of
Short- and Long-term Memory during Infancy and Early Childhood across Cultures
Spatial Cognition during Infancy and Early Childhood Across Cultures, Development of
Stress in Adolescence: Effects on Development
Stress in Childhood: Effects on Development
Stress in Infancy and Early Childhood: Effects on Development
Stress: Effects on Neonatal and Child Development
Tactile Processing

Language Development
Biological Preconditions for Language Development
Child-Directed Speech: Influence on Language Development
First Language Acquisition, Developmental Psychology of
Human Language, Evolution of
Joint Attention during Infancy and Early Childhood Across Cultures
Language Development, Theories of
Language Development: Emergentist Theories
Language Development: Evolutionary Theories
Language Development: Influence of Socio-Economic Status
Multilingual Language Development
Neurophysiological Markers of Language Development
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Methods in Developmental Psychology
Cross-Cultural Research Methods in Psychology
Cross-Sectional Methodologies in Developmental Psychology
Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions
Qualitative Methodology in Developmental Psychology

Phases of Human Development
Adolescence, Sociology of
Adulthood: Dependency and Autonomy
Childhood and Adolescence: Developmental Assets
Childhood: Anthropological Aspects
Cultural Views of Life Phases
Infancy and Human Development
Transition to Adulthood, Cultural Differences in

Social and Cultural Development
Adolescent Health and Health Behaviors
Adolescent Sexual Risk
Adoption, Demography of
Adoption: Domestic, International and Global Perspectives
Attachment during Infancy and Early Childhood: Cultural Variations, Development of
Autism spectrum disorders
Child and Adolescent Psychiatry, Principles of
Child Care and Child Development
Child Neglect
Childcare and Development across Cultures
Early Emotional Development and Cultural Variability
Emotional Development, Effects of Parenting and Family Structure on
Empathy during early Childhood Across Cultures, Development of
Friendship during Adolescence and Cultural Variations
Friendship during Childhood and Cultural Variations
Friendship during Infancy and Early Childhood and Cultural Variations
Gender Development: Cultural Differences
Goal-Oriented Behaviour Among Human and Non-Human Primates
Grandparent–Grandchild Relationships
Imitation during Infancy and Early Childhood
Intentionality during Infancy and Early Childhood, Development of
Longitudinal Analyses of Sexual Development through Early Adulthood
Marital Status, Divorce, and Child Development
Moral Development, Cultural Differences In

Moral Development, Theories of
Mother-Child Communication: Cultural Differences
Neonatal Bonding
Neonatal Crying Behaviors
Neonatal Imitation
Neural Correlates of Social Development
Parenting Attitudes and Beliefs across Cultures
Play During Infancy and Early Childhood: Cultural Similarities and Variations
Pretend Play and Cognitive Development
Prosocial Behavior during Adolescence
Prosocial Behavior during Childhood and Cultural Variations
Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations
Pro-social Behavior, Effects of Parenting and Family Structure on
Self and Identity Development during Adolescence across Cultures
Social and Emotional Development in the Context of the Family
Social Competence During Adolescence Across Cultures
Social Referencing During Infancy and Early Childhood Across Cultures
Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 months of Age: A Systematic Narrative Review
Theory of Mind during Infancy and Early Childhood Across Cultures, Development of
Values Across Cultures, Development of

Theories of Human Development
Aging, Theories of
Bruner's Theory of Cognitive Development
Ecocultural theories of development
Eisenberg's theory of prosocial reasoning
Erikson's Theory of Psychosocial Development
Gibson's Theory of Perceptual Learning
Human Development, Theories of
Integrated Attachement Theory
Meltzoff's "Like Me" Theory
Neo-Piagetian Theories of Cognitive Development
Neo-Vygotskian Developmental Theory
Piaget's Theory of Cognitive Development
Vygotsky's Sociocultural Theory
Vygotsky's Theory of Cognitive Development
Gene-Culture Coevolution

Economics

Field
Agriculture, Economics of
Art and Culture, Economics of
Behavioral Economics
Environmental Economics
Expectations, Economics of
Experimental Economics
Gender, Economics of
Health Economics
Housing Economics
Informal and Underground Economics
Information, Economics of
Innovation and Technological Change, Economics of
International Trade: Economic Integration
Law and Economics
Migration, Economics of
Old and New Institutionalism in Economics
Public Choice Economics
Regulation, Political Economy of
Search, Economics of
Welfare Economics

History
Alternative Schools of Economic Thought
Econometrics, History of
Economic History, Quantitative: United States
Economics, History of
Financial Crises, History of
Macroeconomics, History of From 1933 to Present

Method
Cost-Benefit Analysis
Data Bases and Statistical Systems: Economics (Macroeconomics)
Econometric Software
Economic Growth: Measurement
Income Distribution and Inequality: Measurement Issues
Poverty: Measurement and Analysis

Policy
Antitrust Policy: Lessons from the US
International Trade: Commercial Policy and Trade Negotiations
Monetary Policy since the 2007–2009 Financial Crisis

Theory
Cooperative Game Theory
Economic Growth
Game Theory: Noncooperative Games
Regulation, Economic Theory of
Transaction Costs and Property Rights

Topic
Auctions
Charity and Philanthropy, Economics of
Corporate Taxation
Discrimination, Economics of
Dispute Resolutions in Economics
Economic Education
Economic Transformation: From Central Planning to Market Economy
Foreign Investment: Direct
Income Taxes
Infrastructure Investment
Insurance
Labor Supply
Market Structure and Performance
Regulation, Empirical Analysis of
Retirement, Economics of
Safety, Economics of
Sports, Economics of
State and Local Taxation
Stock Market Predictability
Trade Unions, Economic Behavior of
Transportation, Supply, and Congestion
Unemployment: Structural
Value of Life, Economics of
Wage Differential and Structure
Wealth Distribution

Education

Academic Self-concept and Achievement
Adult Education and Lifelong Learning: The US Experience and Beyond
Aesthetic Education
Affordable Private Schools in Developing Countries
Anthropological Approaches to the Study of Education: The United States and Beyond
Apprenticeships and School Learning: General Considerations
Bilingual Education: International Perspectives
Business Education
Class Size
Classroom Assessment
Classroom Climate
Classroom Level in Educational Effectiveness Research
Co-constructivism in Educational Theory and Practice
Coeducation and Single-Sex Schooling
Cognitive Development: Child Education
Cognitive Development: Mathematics Learning and Instruction
Cognitive Styles and Learning Styles
Communities of Practice
Comparative Research in Education: IEA Studies
Computer Games and Academic Achievement
Computer-assisted Instruction
Continuing Professional Development
Cooperative Learning in Schools
Creativity Training
Cross-Cultural Study of Education
Cross-National Comparisons in Education: Findings from PISA
Data Bases and Statistical Systems: Education, Statistical Systems
Developmental Behavioral Genetics and Education
Distance Education
Doctoral education
Early Education and Care in the World Regions
Education (Primary and Secondary Schools) and Gender
Education and Economic Growth
Education and Employment
Education and Income Distribution
Education and Learning: Lifespan Perspectives
Education for the Gifted and Talented

Education in Latin America and the Caribbean: Systems and Research
Education in South Asia
Education in Sub-Saharan Africa
Education in the Middle East and North Africa
Education, Economics of
Education: Skills Training
Educational Effectiveness Theory and Research: Recent Advances
Educational Effectiveness, The Field of
Educational Evaluation: Overview
Educational Institutions, History of
Educational Learning Theory
Educational Media: Potentials for Learning
Educational philosophy
Educational Planning Worldwide
Educational Policy, Mechanisms for the Development of
Educational Productivity
Educational Research, The Field of
Educational Sociology
Educational Systems: North America
Environments for Education
Equity and Education
Ethnicity and Educational Achievement
Evolutionary Approach to Education Science
Extended schools
Families as Educational Settings
Family and Schooling
Foreign Language Teaching and Learning
Gender and Education
Globalization in Education
Group Processes in the Classroom
Health Education and Health Promotion
Higher Education Management
Higher Education Market
Higher Education Research
Information and Communications Technology in Education
Instructional Design
Instructional Psychology
Intelligence in Cultural, Social and Educational Context
Intelligence, Prior Knowledge, and Learning
International Organisations and Education
Intrinsic Motivation: A Cultural Perspective
Leadership Education: Theory and Practice
Learning and Instruction: Social-Cognitive Perspectives
Learning Theories and Educational Paradigms
Learning to Learn
Legal Education
Lifelong Learning and Its Support with New Technologies
Literacy Education
Mathematics Education
Media and Child Development
Medical education
Metacognitive Development: Educational Implications
Moral Development and Education
Multicultural Education
Networking for School Improvement
Open Education: Learning and Teaching
Parents and Teachers as Partners in Education
Personality Development and Temperament
Piaget's Theory of Human Development and Education
School (Alternative Models): Ideas and Institutions
School as a Social System
School Business Management
School Choice
School Effectiveness Research
School Improvement
School Learning for Transfer
School Management
School to School collaboration
Schooling: Total Impact of
Schools, Micropolitics of
Science Education
Second Language Learning
Self-Concepts: Educational Aspects
Self-Efficacy: Education Aspects
Sex Role Development and Education
Situated learning (Learning In Situ)
Social Inequality and Schooling
Social Studies Education
Socialization and Education: Theoretical Perspectives
Socioeconomic Status and Academic Achievement
Special Education
Sustainability: A Dialog of Values
Tangible User Interfaces in Learning and Education
Teacher Behaviours and Student Outcomes
Teacher Education
Teacher Expertise
Teacher Leadership
Teacher Stress and Burnout: Methodological Perspectives
Teaching as a Profession: United States
Teaching Economics
Teaching for Metacognition
Vocational Education
Vygotskij's Theory of Human Development and New Approaches to Education
Writing Instruction

Environmental and Ecological Sciences

Conceptual Approaches

Animal Rights in Research and Research Application
Annales School
Beck, Ulrich (1944–)
Bioethics
Bioethics: Philosophical Aspects
Cultural Evolution: Overview
Ecofeminism
Environmental Determinism
Environmental Sciences
Gaia
Gene-Culture Coevolution
Global History: Universal and World
Human Evolutionary Genetics
Illich, Ivan (1926–2002)
Innovation and Technological Change, Economics of
IPAT (Impact, Population, Affluence and Technology)
Kuznets Curves
Limits to Growth
Materiality and culture
Nature and Society in Geography
Political Ecology: Anthropological Perspectives
Population Cycles, Formal Theory of
Primitive Society
Risk, Sociology and Politics of
Social Ecology: The Chicago School
Space and Social Theory in Geography
Sustainability: A Dialog of Values
Sustainable Development
Sustainable Development: An Economic Perspective
Sustainable Lifestyles
Sustainable Practices
Sustainable Production and Consumption

Interdisciplinary Research Area

Adaptation, Fitness, and Evolution
Agricultural Science and Technology
Agriculture, Economics of
Agroecology and Agricultural Change
Climate Change: Economics
Communication and Transportation, History of
Development Studies
Ecological Economics
Ecological Social Work: Shifting Paradigms in Environmental Practice
Ecology and Health
Energy and Society
Environment and Development
Environment and Health
Environmental Cognition, Perception, and Attitudes
Environmental Economics
Environmental Education and Education for Sustainable Development
Environmental Policy: Protection and Regulation
Environmental Sociology
Evolutionary Sociology
Food, Sociology of
Green Social Work
Human Behavioral Ecology
Industrial Ecology
Land Change: The Merger of Land Cover and Land use Dynamics
Modernization and Modernity in History
Political Economy of Renewable Energy
Social Ecology
Sociobiology and Sociology: A New Synthesis
Sociobiology: Overview
Sustainability and Sustainability Science
Climate, History of
Colonization and Colonalism, History of
Environmental History
Historical Demography
Industrialization, Typologies and History of

Issues
Alternative Food Movements
Anthropocene Epoch
Carrying Capacity of the Environment
Climate Change and the Media
Consumer Policy
Consumption, Geography of
De-growth
Development and Urbanization
Ecotourism
Electronic Waste
Environmental Health and Safety: Social Aspects
Environmental Justice in the United States
Environmental Movements
Environmental Risk
Environmental stress and health
Ethical Consumers
Fertility and Culture: Anthropological Insights
Fire, Human Use, and Consequences
Food Security and 'Green Revolution'
Geographies of Food: Politics, Place, Power Reprised
Green Parties
Hazards and Disaster Research in Contemporary Anthropology
Indigenous Knowledge and Intellectual Property Rights
Intergenerational Justice
Irrigation Societies
Maritime Networks, History of
Mineral Resources, Limits to: The Case of Peak Oil
Natural Disasters: Health-Related Aspects
Nomads and Nomadism in History
Oil Age: Energy in Transition
Resource Geography
Slavery: Comparative Aspects
Slow Food as a Social Movement
Social Resilience in Post-Tsunami Japan
Sustainable Agriculture
Sustainable Transportation
Time Use and Resource Consumption
Water Resources and Sustainable Water Management
Water Wars
Climate Change
Climate Change and Global Poverty
Collaborative Consumption
Consumer Psychology
Demographic Impact of Disasters
Eco-cities and Sustainable Urbanism
Environmental Archaeology
Environmental Conservation and Development: Critical Perspectives
Food Production, Origins of
Hunting and Gathering Societies: Anthropology
Modes of Production
Nongovernmental Organizations: Anthropological and Historical Aspects
Pastoralism in Anthropology
Peasantry in the 21st Century
Postindustrial Societies
Pro-Environmental Behavior
Russian and Post-Soviet Studies: Environment
Social Work and Disasters
Sustainable Development and Social Work
Urbanization
Work, Anthropology of

Methods and Measurement
Demographic Measurement: General Issues and Measures of Fertility
Demographic Measurement: Nuptiality, Mortality, Migration, and Growth
Energy Flow Analysis
Environmental Attitudes and Behavior: Measurement
Environmental Planning
Geographic Information Systems and Remote Sensing
Human-Environment Interactions: Case Studies
Land Use Indicators
Material Flow Analysis
Participatory Action Research in Environmental and Ecological Studies
Spatial–Temporal Modeling

Ethics of Research

Animal Rights in Research and Research Application
Archiving: Ethical Aspects
Bioethics
Censorship in Research and Scholarship
Cluster Randomized Trials, Ethical Implications of
Common Rule: US Regulatory Protections for Human Research Participants
Ethical Dilemmas, Research and access to treatment
Ethical Practices: Institutional Oversight and Enforcement
Ethics for Biomedical Research Involving Humans: International Guidelines
Euthanasia
Fieldwork: Ethical Aspects
Gender and Social Research
Informed Consent of Research Subjects
Intellectual Property Rights: Ethical Aspects
Internet Research: Ethical Concerns
Media Ethics
Mentoring: Supervision of Young Researchers
Objectivity of Research: Ethical Aspects
Placebo-Controlled Trials, Ethics of
Privacy of Individuals in Social Research: Confidentiality
Research Ethics, Cross-Cultural Dimensions of
Research Ethics: Research
Research Funding: Ethical Aspects
Research Publication: Ethical Aspects
Surveys and Polling: Ethical Aspects
Therapeutic Effectiveness Assessment: Approaches in Contemporary Research

Evolutionary Sciences

Emotion, Evolution of
Evolution: Diffusion of Innovations
Evolutionary Demography: A Darwinian Renaissance in Demography
Evolutionary Economics
Evolutionary epistemology
Evolutionary medicine
Human Behavioral Ecology
Incest Prohibition, Origin and Evolution of
Kinship, Evolution of
Morality: Evolution of
Number Systems, Evolution of
Religious Belief, Evolution of
Technology, Evolution of
Writing, Evolution of

Cultural Evolution
Cultural Evolution: Overview
Cultural Evolution: Phylogeny versus Reticulation
Cultural Evolution: Theory and Models
Gene-Culture Coevolution
Social Learning and Culture in Nonhuman Organisms

Evolution and Language
Evolution and Language: Cultural Transmission
Evolution and Language: Overview
Evolution and Language: Phylogenetic analyses

Evolution of Cooperation
Human Cooperation, Evolution of
Kin Selection
Evolutionary Psychology
Cognition, Evolution of
Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future
Evolutionary Developmental Psychology
Evolutionary group dynamics
Evolutionary social psychology

Evolutionary Theory
Adaptation, Fitness, and Evolution
Comparative Method in Evolutionary Studies
Darwinism
Epigenetic inheritance
Evolution, History of
Evolutionary Theory, Structure of
Genotype and Phenotype
Microevolution
Sociobiology: History of
Sociobiology: Overview

Human Evolution
Body, Evolution of
Brain, Evolution of
Primates, Evolution of

Mating Human
Mate Choice, Evolution of
Marriage Systems, Evolution of

Gay, Lesbian, Bisexual and Trans-sexual

Clinical Psychology, Psychiatry and Homosexuality
Femininities
Masculinity, Scripts of
Queer Theory and Intersectionality
Sexual Motivation
Sexuality Education
Queer Anthropology
Feminist Theory: Radical Lesbian
Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals
Gay and Bisexual Latinos in US Gay Culture
Gay and Lesbian Movements
Gay and Lesbian Tourism
Gay Men's Emancipation Movement in Germany: 1897–1933
Gay, Lesbian, and Bisexual Youth
Gays, lesbians and sports: No longer on the sideline
Heterosexism and Homophobia
Lesbian and Gay Activism in Latin America
Lesbian, Gay, Bisexual and Transgender Aging
Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures
Lesbian, Gay, Bisexual and Trans-sexual Drag Culture
Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview
Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex
Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures
Lesbians: Social and Economic Situation
Medicine and Health for Sexual Minorities
Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations
Queer Anarchism
Queer Geography
Same-Sex Sexualities in Turkey
School Environment for LGBTQ/Sexual Minority Youth
Sexual Minorities and Sexual Citizenship
Sexual Minorities in China
Sexual Minorities in Modern Russia
Sexual Minorities in South Africa
Sexual Minorities in the Maghreb
Sexual Minority Families of Choice and Family Networks: We are family
Sexual Orientation: Biological Influences
Sexual Orientation: Historical and Social Construction
Transgender, Transexual, and Gender Variant Individuals

Genetics, Behavior, History and Society

Anthropology, Genomics, and Human Variation: National Roots
Behavioral Genetics: Psychological Perspectives
Biobanking: Ethical Issues
Biocitizenship
Bioethics in the Post-genomic era
Bioethics: Genetics and Genomics
Bioethics: Philosophical Aspects
Community Genetics
Direct-to-Consumer Personal Genetic Testing
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute
Eugenics as a Basis of Population Policy
Eugenics as an International Movement
Eugenics in Asia
Eugenics in Europe
Eugenics in the Americas
Eugenics, History of
Familial Studies: Genetic Inferences
Forensic Genetic Databases: Ethical and Social Dimensions
Gene-Environment Interactions in Health and Well-Being
Genes and Behavior: Lessons from the Fruit Fly
Genetic Admixture
Genetic Ancestry Testing
Genetic Counseling: Historical, Ethical, and Practical Aspects
Genetic Engineering
Genetic Screening for Disease
Genetic Studies of Behavior: Methodology
Genetics and Anthropology
Genetics and Disaster Victim Identification
Genetics and Economic Behavior
Genetics and Forensics
Genetics and Indigenous Communities: Ethical Issues
Genetics and Social Justice
Genetics and Society
Genetics and Sociology
Genetics and the Media
Genetics of Personality
Genetics, Disease, and Reproduction
Genetics: Legal Aspects
Genetics: The New Genetics
Genographic Project
Genomics, Ethical Issues in
Human Evolutionary Genetics
Human Genome Diversity Project: History
Human Genome Project: History and Assessment
Human Genomics in Asia
Human Mate Choice
Incest, Inbreeding, and Their Consequences
Intelligence, Genetics of: Cognitive Abilities
New Genetics and Race
Primate Behavior: Significance for Understanding Humans
Primate Groups and Their Correlates
Race: Genetic Aspects
Sexual Orientation: Genetic Aspects
Social Epigenetics
Twin Studies and Epigenetics

Geography

Cartography

Cartographic Visualization
Cartography
Citizen Mapping
Cognitive Maps
Dynamic Mapping in Geography
Tactile Maps in Geography
Thematic Maps in Geography
Topographic Maps

Core Concepts

Access: Geographical
Diffusion: Geographical Aspects
Equality and Inequality: Legal Aspects
Equity in Law
Kant, Immanuel (1724–1804)
Knowledge, Anthropology of
Landscape
Location, Absolute and Relative
Networks, Geography of
Place in Geography
Region: Geographical
Scale in Geography
Spatial Equity
Spatial Interaction
Territoriality: Geographical
Time–Space in Geography

Data, Methods, and Analysis

Census Microdata
Censuses, Population: Comparative International Aspects
Censuses: Current Approaches and Methods
Censuses: History and Methods
Common Rule: US Regulatory Protections for Human Research Participants
Comparative Studies: Method and Design
Computers: Impact on the Social Sciences
Contextual Studies: Methodology
Data Bases and Statistical Systems: Overview
Demographic Techniques: Small-area Estimates and Projections
Ecological Fallacy, Statistics of
Environmental Attitudes and Behavior: Measurement
Fieldwork: Ethical Aspects
Geographic Information Systems
Geographic Information Systems and Remote Sensing
Geospatial Applications in Archaeology
Global Positioning System
Grounded Theory: Methodology and Theory Construction
Mathematical Models in Geography
Qualitative Methods in Geography
Remote Sensing
Spatial Analysis in Geography
Spatial Association, Measures of
Spatial Autocorrelation
Spatial Data
Spatial Data Infrastructures
Spatial Decision Support Systems
Spatial Optimization Models
Spatial Pattern, Analysis of
Spatial Sampling
Spatial–Temporal Modeling
Citizen Science
Confidentiality and Statistical Disclosure Limitations

Ecological Inference
Indigenous Knowledge: Science and Technology Studies
Neighborhood Effects
Spatial Statistical Methods

Economic
Actor-Network Theory
Agriculture, Economics of
Behavioral Economics
Behavioral Economics, History of
Bounded Rationality
Business Process: Outsourcing
Capitalism: Global
Castells, Manuel (1942–)
Collaborative Consumption
Commodity Chains
Consumption, Geography of
Cost-Benefit Analysis
Culture and Economic Development
Culture and Economy
Data Bases and Statistical Systems: Labor and Work
Decision and Choice: Bounded Rationality
Decision and Choice: Luce's Choice Axiom
Development and Employment: A Mutual Relationship
Development and the State
Development Geographies
Division of Labor: History of the Concept
Ecological Economics
Economic Geography
Economic Restructuring: Geographic Aspects
Economic Sociology
Education and Economic Growth
Education and Employment
Entrepreneurship and Economic Growth
Entrepreneurship, Geography of
Finance, Geography of
Foreign Investment: Direct
Global Development: Markets and Wellbeing
Globalization and Transnational Corporations
Hotelling, Harold (1895–1973)
Industrial Cluster Dynamics
Industrial Geography
Innovation
Innovation and Employment
Innovation and Technological Change, Economics of
International Business: Geographic Aspects
International Trade: Commercial Policy and Trade Negotiations
International Trade: Economic Integration
Internationalization: Patterns in Business firms
Labor, Geography of
Land Tenure
Local Economic Development
Location Theory
Market Geographies
Marketing Strategies
Network Analysis
Nonprofit Sector, Geography of
Old and New Institutionalism in Economics
Organizational Path Dependence
Pareto, Vilfredo (1848–1923)
Service Economy, Geography of
Shadow Work: Measurement
Social Capital
Social Networks
Trust, Sociology of
International Migration and Remittances
Environmental Economics
Evolutionary Economics

Nature-Society
Africa's Environment: A Stressed Biogeographical and Cultural Landscape
Agroecology and Agricultural Change
Boserup, Ester (1910–99)
Carrying Capacity of the Environment
Climate Change
Climate Change and Global Poverty
Climate Change and the Media
Climate Change: Economics
Community resilience to disasters
Disasters in Social, Cultural and Political Context
Ecofeminism
Energy and Society
Energy, Geography of
Environmental Conservation and Development: Critical Perspectives
Environmental Education and Education for Sustainable Development
Environmental Planning
Environmental Policy: Protection and Regulation
Environmental Risk
Environmental Sociology
Fire, Human Use, and Consequences
Fishing Societies
Hazards and Disaster Research in Contemporary Anthropology
Human-Environment Interactions: Case Studies
Industrial Ecology
IPAT (Impact, Population, Affluence and Technology)
Irrigation Societies
Land Change: The Merger of Land Cover and Land use Dynamics
Land Use Indicators
Limits to Growth
Malthus, Thomas Robert (1766–1834)
Mineral Resources, Limits to: The Case of Peak Oil
Natural Disasters: Health-Related Aspects
Nature and Society in Geography
Participatory Action Research in Environmental and Ecological Studies
Political Economy of Renewable Energy
Resource Geography
Sauer, Carl Ortwin (1889–1975)
Sustainability and Sustainability Science
Sustainable Agriculture
Sustainable Development
Sustainable Lifestyles
Sustainable Practices
Sustainable Production and Consumption
von Humboldt, Alexander (1769–1859)
Water Resources and Sustainable Water Management
Water Wars
Water, Health and Social Inequality
White, Gilbert (1911–2006)
Environmental Sciences

Political
Citizen Participation
Citizenship: Political
Civil Rights Movements
Civil Society and the Public Sphere: History of the Concept
Civil Society, Concept and History of
Elections: Apportionment
Electoral Geography
Electoral Systems
Environmental Movements
Ethnic Cleansing, History of
Ethnic Conflict, Geography of
Genocide: Social and Economic Aspects
Military Geography
Movements, Social
Neoliberalism
Organization, Geography of
Political Geography
Political Representation
Postcolonial Geography
Postcolonial Science Studies
Social Movements and Political Violence
Social Movements, Geography of
State, Geography of
State, Geography of
Surveillance and Privacy, Geography of
Territorial Conflicts
War and Nationalism
War and Terrorism, Geography of
Geopolitics
Global Governance

Population
Aging and Older People, Geography of
Demographic Models
Demography: History Since 1900
Diaspora
Ecology of Aging
Global Migration
Internal Migration in Industrialized Countries
Internal Migration: Developing Countries
International Migration
Population Aging: Economic and Social Consequences
Population Geography
Public Opinion: Social Attitudes
Refugees: Geographical Aspects
Migration and Development
Demographic Impact of Disasters

Social
Asceticism
Bourdieu, Pierre (1930–2002)
Children's Geographies
Class: Social
Communities and Crime

Disability, Geography of
Discrimination: Racial
Environmental Justice in the United States
Integration and Multiculturalism: Focus on Western Europe
Queer Geography
Race and Racism, Geography of
Social Action and Social Justice
Social Geography
Social Justice and Social Inequalities
Youth, Geography of

Themes
Applied Geography: A US perspective
Crime, Geography of
Cultural Geography
Ecology and Health
Education, Geography of
Environment and Health
Environmental Cognition, Perception, and Attitudes
Geographies of Food: Politics, Place, Power Reprised
Health, Geography of
Historical Geography
Law and Geography
Medical Geography
Poverty, Geography of
Public Health as Social Science
Rural Geography
Science, Geography of
Sexuality and Space
Sexuality, Theories of
Spatial Cognition
Spatial Navigation

Theoretical Perspectives
Critical Race Theory
Critical Realism in Geography
Critical Theory
Critical Theory: Contemporary
Deconstruction
Dependency Theory
Determinism: Social and Economic
Development Theory in Geography
Discourse Studies
Ecumene
Environmental Determinism
Gaia
Gender and Feminist Studies in Geography
Geographic Thought, History of
Geography: Overview
Giddens, Anthony (1938–)
Globalization: Geographical Aspects
Information Society, Geography of
Knowledge, Geography of
Marxist Archaeology
Marxist Geography
Realism, in Applied Research
Regional Geography
Regional Science
Situated Knowledge, Feminist and Science and Technology Studies Perspectives
Space and Social Theory in Geography
Spatial Thinking in the Social Sciences

Transport
Automobility
Communication and Transportation, History of
Ecotourism
Mobility and Travel Activity Patterns
Physical Activity and Health
School Choice
Sustainable Transportation
Tourism, Geography of
Transportation Geography
Transportation Planning
Transportation, Psychology of
Transportation, Supply, and Congestion

Urban
Cities and the Financial Crisis
Cities Under Postsocialism
Cities: Capital, Global, and World
Cities: Internal Structure
City Marketing
Community Sociology
Concentrated Disadvantage
Critical Planning Theory
Data Bases and Statistical Systems: Environment and Urban Studies
Desegregation
Development and Urbanization
Eco-cities and Sustainable Urbanism
Food Deserts in US Cities
Gated communities
Gentrification
Ghetto
Growth Management and Urban Planning in the United States
Homelessness in the United States
Human Ecology
Inner City
Jacobs, Jane (1916–2006)
Just City
Neighborhood
Neighborhood Revitalization and Community Development
New Urbanism
Planning and Development
Planning during Post-Socialism
Planning Ethics
Planning Issues and Sustainable Development
Planning Theory: The Good City and Its Governance
Postmodern Urbanism
Race: Ethnicity and Urban Planning
Regional Planning
Residential Segregation in the United States
Residential Segregation: Geographic Aspects
Residential Segregation: Recent Trends
Resilient City
Right to the City: A Liberal-Democratic Perspective
Segregation Indices
Segregation, Demographic Effects of
Shrinking Cities
Smart Cities
Social Ecology: The Chicago School
Spatial Dynamics and Crime
Spatial Mismatch
Suburbanization and Suburbanism
Urban Geography
Urban Growth Models
Urban Informality: The Production and Regulation of Space
Urban Internal Spatial Structure
Urban Life and Health
Urban Life: Future Developments
Urban Neoliberalism
Urban Places: Statistical Definitions
Urban Policies in Asia
Urban Policy in Europe
Urban Policy in North America
Urban Poverty in Neighborhoods
Urban Regeneration
Urban Sprawl
Urban Studies: Overview
Urban System in Geography
Urban Violence
Urban Violence and Militarization
Urbanization
Urbanization in Africa
Urbanization in China

Health

Adolescent Sexual Risk
Adult Mortality in Developing Countries
Adult Mortality in Eastern Europe and the Former Soviet Republics
Adult Mortality in Industrialized Societies
African Studies: Health
Aging and Health in Old Age
AIDS: Acquired Immune-deficiency Syndrome
Alcohol Interventions: Disease Model vs. Harm Reduction
Alcohol-Related Disorders
Alzheimer's Disease, Neural Basis of
Antidepressant Drugs
Anxiety and Anxiety Disorders
Behavior Therapy: Background, Basic Principles and Early History
Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues
Behavior Therapy: Psychiatric Aspects
Bereavement and Complicated Grief across the Lifespan
Bioethics
Bipolar (Manic-Depressive) Disorder
Burnout, Psychology of
Child and Adolescent Psychiatry, Principles of
Data Bases and Statistical Systems: Health
Death and Dying, Psychology of
Death and Dying, Sociology of
Dementia: Psychiatric Aspects
Depression
Disability: Sociological Aspects
Drugs: Illicit Use and Prevention
Early Life Influences on Health and Mortality in Adulthood
Eating Disorders
Emotions and Health
Environment and Health

Epilepsy
Family and Psychiatric Illness: Family Therapy
Genetic Screening for Disease
Genetics, Disease, and Reproduction
Geriatric Psychiatry
Health Care: Legal Issues (United States)
Health Economics
Health Education and Health Promotion
Health in Low-Income Countries
Infant and Child Mortality in Industrialized Countries
Infant and Child Mortality in the Less Developed World
Limbic System
Managed Care and Psychotherapy in the United States
Medical education
Medical Geography
Medical Sociology
Medicalization: Sociological and Anthropological Perspectives
Medicine and Health for Sexual Minorities
Medicine, History of Western
Mental Disorders, Epidemiology of
Mental illness, Genetics of
Mental Illness: Family and Patient Organizations in Europe
Mentally Ill, Public Attitudes towards
Mortality and the HIV/AIDS Epidemic
Mortality Crossover
Mortality Differentials: Selection and Causation
Mortality of the Oldest-Old
Mortality, Biodemography of
Mortality, Epidemiological, and Health Transitions
Mortality: The Great Historical Decline
Nosology in Psychiatry
Obesity and Mortality
Occupational Health and Safety, Regulation of
Organizational Justice
Panic Disorder
Personality Disorders
Placebo Studies (Double-Blind Studies)
Placebo-Controlled Trials, Ethics of
Post-traumatic Stress Disorder
Psychiatric Care, Economics of
Psychopharmacotherapy, Costs of
Psychopharmacotherapy: Side Effects
Psychosomatic Medicine
Psychotherapy and Pharmacotherapy
Recovery of Function: Dependency on Age
Research Ethics, Cross-Cultural Dimensions of
Research Ethics: Research
Research Funding: Ethical Aspects
Research Publication: Ethical Aspects
Schizophrenia
Schizophrenia, Treatment of
Schizophrenia: Metabolic Consequences
Self-Directed Care and Individual Budgets for Older People: Lessons from the UK
Self-Esteem
Sexually Transmitted Infections: Social Network Analysis
Social Justice and Social Inequalities
Social Stratification
Sociology and Psychiatry
Somatoform Disorders
Stigma, Social Psychology of
Stress at Work
Stress in Adolescence: Effects on Development
Stress in Childhood: Effects on Development
Stress in Infancy and Early Childhood: Effects on Development
Stress: Effects on Neonatal and Child Development
Stress: Psychological Perspectives
Subjective Wellbeing and Culture
Subjective Wellbeing, Psychology of
Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model
Suicide
Suicide, Sociology of
Therapist–Patient Relationship
Value of Life, Economics of
Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences
Working Conditions
Workplace Autonomy
Workplace Safety and Health

Psychology

Adolescent Health and Health Behaviors
Behavioral Medicine
Bereavement
Body Image
Cancer Screening
Cancer: Psychosocial Aspects
Cardiac Disease, Coping with
Cardiovascular Reactivity and Recovery
Childhood Cancer: Psychosocial Aspects
Chronic illness, psychosocial coping with
Chronic Pain: Models and Treatment Approaches
Co-morbidity of Mental and Physical Conditions
Control Beliefs: Health Perspectives
Coronary Heart Disease: Psychosocial Aspects
Depression, Pessimism, and Health
Diabetes: Psychosocial Aspects
Drug Use and Abuse: Psychosocial Aspects
Epigenetics
Evidence-Based Medicine
Explanatory style and health
Gastrointestinal Diseases: Psycho-social Aspects
Gender and Physical Health
Gynecological Health: Psychosocial Aspects
Health Behavior, Psychosocial Theories of
Health Behaviors
Health Behaviors, Assessment of
Health Outcomes, Assessment of
Health Psychology
Health Psychology in Childhood
Health Risk Perception
Health: Definitions
HIV Risk Interventions
Hypertension: Psychosocial Aspects
Illness and Dyadic Coping
Illness Behavior and Care Seeking
Infectious Diseases: Psychosocial Aspects
Interoception
Men's Health
Mental Health
Mindfulness
Oral Health, Psychology of
Pain, Health Psychology of
Patient Adherence to Health Care Regimens
Personality and Health
Personalized Medicine
Physical Activity and Health
Physical Activity and Health Promotion
Psychoneuroimmunology
Rehabilitation (Cardiac): Psychosocial Aspects
Resilience
Respiratory Disorders: Psychosocial Aspects
Rheumatoid Arthritis: Psychosocial Aspects
Self-Efficacy and Health
Sexual Risk Behaviors
Sleep and Health
Social Support and Recovery from Disease and Medical Procedures
Startle Reflex and Health
Stress and Cardiac Response
Stress, Coping and Health
Sun Exposure and Skin Cancer Prevention
Symptom Perception, Awareness and Interpretation
Weight Regulation

Sociology

Addictions: General Considerations
Alcohol Use among Young People
Alzheimer's Disease: Behavioral and Social Aspects of
Caregiver Burden
Comparative Health Systems
Compression of Morbidity
Coping with Job Stress
Culture as a Determinant of Mental Health
Disability: Social and Psychological Aspects
Doctor–Patient Interaction in the West: Psychosocial Aspects
Ecology and Health
Environmental Health and Safety: Social Aspects
Environmental stress and health
Functioning, Disability and Health, International Classification of
Gender and Health Care
Globalization and Health
Healing Practices: Alternative and Complementary
Health and Illness: Mental Representations in Different Cultures
Health Care Markets: Theory and Empirical Evidence
Health Care, Rationing of

Health Insurance: Economic and Risk Aspects
Health Interventions, Community-based
Health Literacy
Health Promotion in Schools
Health Research, Qualitative
Health Surveys
Health-Related Support and Self-help Mutual Aid Groups
Hospitals and Health: Psychosocial Aspects
Human Development and Health
Injuries and Accidents: Psychosocial Aspects
Justice and Health
Life Course Epidemiology
Medical Profession
Migration and Health
Natural Disasters: Health-Related Aspects
Nonphysician Health Professionals
Occupational Health
Occupational Stress Management
Public Health as Social Science
Race: Ethnicity and Health
Religion and Health: Clinical Considerations and Applications
Retirement and Health
Smoking and Health
Social Epidemiology
Social Exclusion, Social Deprivation and Health
Social Integration, Social Networks, and Health
Socioeconomic Status and Health
Socioeconomic Status and Health Care
Unemployment and Mental Health
Urban Life and Health
Violence as a Public Health Problem

History

Absolutism, History of
American Revolution, The
Ancient Greece and Rome, History of
Aristocracy, Nobility, and Gentry, European History of
Art History
Artisans and Guilds, History of
Balance of Power, History of
Biography: Historical
Body, History of
Bourgeoisie and Middle Classes, History of
Bureaucratization and Bureaucracy, History of
Business History
Capitalism
Chinese Revolutions: Twentieth Century
Citizenship, Historical Development of
Civil Society, Concept and History of
Civilization, Concept and History of
Classical Legacies
Climate, History of
Colonization and Colonalism, History of
Communication and Transportation, History of
Communism, History of
Comparative History
Confederations, History of
Conservatism: Historical Aspects
Consumption, History of
Contemporary History
Crowds in History
Cuban Revolution, The
Cultural History
Data Bases and Statistical Systems: Archives and Historical Databases
Democracy, History of
Dictatorship in History
Economic History
Education, History of
Emotions, History of
English Revolution, The
Enlightenment
Environmental History
Ethnicity and Ethnic Groups: Historical Aspects
Family and Kinship, History of
Feudalism
First World War, The
Freedom, History of
French Revolution, The
Frontiers in History
Gender History
Generations in History
Genocide, Historical Aspects of
Global History: Universal and World
Historical Demography
Historical Thought and Historiography: Christian Tradition
Historical Thought and Historiography: Classical Period (especially Greece and Rome)
Historical Thought and Historiography: Current Trends
Historical Thought and Historiography: East Asia
Historical Thought and Historiography: Islamic Tradition
Historical Thought and Historiography: Jewish Tradition
Historical Thought and Historiography: Modern History (since the 18th century) - the West
Historical Thought and Historiography: South Asia
Historical Thought and Historiography: Southeast Asia
Historical Thought and Historiography: Sub-Saharan Africa
Historical Thought and Historiography:Indigenous Cultures in the Americas
Historicism
History and Literature
History and the Social Sciences
History from Below, the History of Everyday Life, and Microhistory
History, Professionalization of
History: Overview
History: Representation within and Functions of
History: Theories and Methods
Human Rights, History of
Imperialism, History of
Industrial Relations, History of
Industrialization, Typologies and History of
Intellectual History
Interest Groups, History of
Isms and History
Labor History
Laicization, History of
Liberalism: Historical Aspects
Life Course in History
Linguistic Turn and Discourse Analysis in History
Literacy and Illiteracy, History of
Maritime Networks, History of
Media and History
Medicine, History of Western
Meiji Revolution, The
Mexican Revolution, The
Middle Ages, The
Migration History
Military History
Modernization and Modernity in History
Music, History of
National Liberation Movements
National Socialism and Fascism
Nationalism, Historical Aspects of: Africa
Nationalism, Historical Aspects of: Arab World
Nationalism, Historical Aspects of: East Asia
Nationalism, Historical Aspects of: South Asia
Nationalism, Historical Aspects of: The West
Nations and Nation-states in History
Nomads and Nomadism in History
Oral History
Parliaments, History of
Peasants and Rural Societies in History (Agricultural History)
Plagues and Diseases in History
Political History: History of Politics
Political Parties, History of
Political Thought, History of
Poverty in History
Professions and Professionalization, History of
Property Ownership
Property Rights, History of
Prosopography (Collective Biography)
Psychohistory
Public Sphere: Eighteenth-Century History
Public Sphere: Nineteenth and Twentieth-Century History
Quantification in History
Racism, History of
Reformation and Confessionalization
Religion, History of
Renaissance
Revolutions of 1989–90 in Eastern Central Europe
Russian Revolution, The
Science, History of
Second World War, The
Security: Commercial Transactions
Slaves and Slavery, History of
Small States, History of

Social History
Social Inequality in History (Stratification and Classes)
Social Mobility, History of
Social Movements, History of: General
Socialism: Historical Aspects
Sport, Historiography of
State, History of
Subaltern History
Time, Chronology, and Periodization in History
Urban History
Vietnamese Revolution, The
Visual Turn
Welfare State, History of
White-collar Workers, History of
Work, History of
Working Classes, History of

History of the Social and Behavioral Sciences

Actors, opportunity structures for
Academic Careers in Comparative Perspective
Academic Research and Employment: Recent Changes in Europe and the United States
Associations, Scholarly
Awards, public recognition and evaluation
Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography
Ethical Guidelines and Codes of Conduct in Social and Behavioral Research
Gatekeepers in Social Science
Scholars, Private Lives of
Social Scientists as Experts and Public Intellectuals
Social Science Professions and Professionalization

Concepts, Ideas, and Theories, History of
Eugenics, History of
Measurement Theory: History and Philosophy
Old Age: Definitions, Theory, and History of the Concept
Psychophysical Theory and Laws, History of
Signal Detection Theory, History of
AGIL, History of
Anomie: History of the Concept
Capital: History of the concept
Capitalism: The History of the Concept
Civil Society and the Public Sphere: History of the Concept
Class: History of the Concept
Community and Society: History of the Concepts
Conceptual History: Begriffsgeschichte
Definition of the situation: History of the Concept
Division of Labor: History of the Concept
Elites: History of the Concept
Emotion: History of the Concept
Equilibrium: History of the Concept
Geist (Spirit): History of the Concept
Gift: History of the Concept
Habit: History of the Concept
Habitus: History of a Concept
Happiness: History of the Concept
Ignorance, History of Concept
Individuals in Society: History of the Concept
Intellectuals: History of the Concept
Intelligence: History of the Concept
Interest: History of the Concept
Introspection: History of the Concept
Metaphor, Role in Social Thought: History of the Concept
Modernity: History of the Concept
Moral Entrepreneur
Motivation: History of the Concept
Political Religions: History of the Concept
Profits
Reification: History of the Concept
Secularization: History of the Concept
Self: History of the Concept
Social Structure: History of the Concept
Social, The: History of the Concept
Society and People: History of the Concept
Solidarity: History of the Concept
State and Society: History of the Concept
Translating Sociology into Japanese, Chinese and Korean Languages
Uncertainty: History of the Concept
Unconscious: History of the Concept
Value: History of the Concept
Work, Labor: History of the Concept

Controversies
Determinism: Social and Economic
Egalitarianism
Gender: Gendering of Categories
Keynesianism and the Question of State Interventionism
Mind-Body Dualism
Nature-Nurture Controversy, History of
Objectivity, History of
Progress: History of the Concept
Psychologism, History of
Science and Politics: Value Neutrality
Utilitarian Social Thought, History of

Disciplines, History of
Anthropology, History of
Biomedical Sciences, History and Sociology of
Clinical Psychology in Europe, History of
Clinical Psychology in North America, History of
Cognitive Psychology: History
Criminology, History of
Demography, Early History of
Demography: History Since 1900
Discipline formation in the social sciences
Econometrics, History of
Economics, History of
Law: History of its Relation to the Social Sciences
Linguistics, History of
Macroeconomics, History of From 1933 to Present
Macroeconomics, History of up to 1933
Microeconomics, History of
Political Economy, History of
Political Science, History of
Political Theory, History of
Psychoanalysis, History of
Psychology, History of (Early Period)
Psychology, History of (Twentieth Century)
Psychotherapy, History of: Psychiatric Aspects
Science, History of
Scientific Disciplines, History of
Sociology, History of
Specialization and Recombination of Specialties in the Social Sciences
Statistics, History of

Fields of Research
Applied Social Research, History of
Area studies, history of
Behavioral Economics, History of
Communication Research and Media Studies, History of
Cultural Sociology, History of
Epistemic Communities and Collaborative Research, History of
Science and Technology Studies, History of

Impact and Consequences
American and French revolutions: Impact on the Social Sciences
Cold War: Impact on Social and Behavioral Sciences
Computers: Impact on the Social Sciences
Counter-Insurgencies and the Social Sciences
Enlightenment: Impact on the social sciences
History and the Social Sciences
Liberalism: Impact on Social Science
Scientific Revolution, History and Sociology of
Social and Behavioral Sciences under Dictatorship
Social Question: Impact on Social Thought
Western and Non-Western views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences

Institutions and Practices
Censuses, Population: Comparative International Aspects
Censuses: History and Methods
Encyclopedias, Handbooks, and Dictionaries
Ethical Questions in Social and Behavioral Sciences, history of
Funding of Social Science, History of
Journals, Scholarly
Museums
Peer Review: Organized Skepticism
Publishing: Academic and University Presses
Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance

Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct
Social Science and the Media
Social Science and Universities
Social Science Learned Societies and Professional Associations
Statistical Analysis Packages
Survey Research Centers and Companies

Methodologies, History of
Case Method and Case Study Collections
Causal Analysis, History of
Correspondance Analysis, History of
Empirical Social Research, history of
Heuristics, History of
Interviewing, History of
Meta-analysis, History of
Methodology of the History of the Social and Behavioral Sciences
National Account Systems
Network Analysis, History of
Qualitative Methods, History of
Quantification in the History of the Social Sciences
Quasi- and Field experiment
Sample Survey Methodology, History of
Semantic Differential
Small Group Research, History of
Unobtrusive Methods, History of

Research Programs and Paradigms
Alternative Schools of Economic Thought
Experimentation in Psychology, History of
Field Theory
Functionalism, History of
Hermeneutics, History of
Human Sciences, History of
Institutionalism and Neo-institutionalism: History of the Concepts
Intellectual History
Linguistic Turn
Marxism in Contemporary Sociology
Positivism, History of
Pragmatist Social Thought, History of
Prospect Theory
Social Ecology: The Chicago School
Social Mechanism
Sociobiology: History of
Weberian Social Thought, History of

Research Topics, History of
Alienation, History of
Attitude: History of Concept
Conspicuous consumption and distinction, history of
Discourse Studies
Ethnocentrism and Prejudice: History of the Concepts
Human Genome Diversity Project: History
Human Genome Project: History and Assessment
Instinct, History of
Knowledge Society, History of
Labeling Theory, History of
Life-style, History of the Concept
Mass Society: History of the Concept
National Character, History of
Obedience (Research), History of
Reference Group, History of
Science: Constructivist Perspectives, History of
Social Control and Deviance, History of
Spatial Thinking in the Social Sciences
Unintended Consequences: History of the Concept

Institutions and Infrastructure of Social and Behavioral Sciences

Census
Census Microdata
Censuses, Population: Comparative International Aspects
Censuses: Current Approaches and Methods
Censuses: History and Methods

Databases
Data Bases and Statistical Systems: Communications and Media
Data Bases and Statistical Systems: Crime and Delinquency
Data Bases and Statistical Systems: Culture
Data Bases and Statistical Systems: Demography
Data Bases and Statistical Systems: Economics (General)
Data Bases and Statistical Systems: Economics (Statistical Systems)
Data Bases and Statistical Systems: Education, General
Data Bases and Statistical Systems: Environment and Urban Studies
Data Bases and Statistical Systems: Health
Data Bases and Statistical Systems: International Comparative Research
Data Bases and Statistical Systems: Labor and Work
Data Bases and Statistical Systems: Law
Data Bases and Statistical Systems: Linguistics
Data Bases and Statistical Systems: Political Behavior and Elections
Data Bases and Statistical Systems: Political Science (General)
Data Bases and Statistical Systems: Security and Conflict
Data Bases and Statistical Systems: Sociology

Experimental Laboratories
Experimental Laboratories: Psychological
Experimental Laboratories: Social and Economic

Institutions, Organizations and Communication
Associations, Scholarly
Case Method and Case Study Collections
E-Communications Platforms and E-Learning
Encyclopedias, Handbooks, and Dictionaries
Epistemic Communities and Collaborative Research
Epistemic Communities and Collaborative Research, History of
Ethical Guidelines and Codes of Conduct in Social and Behavioral Research
Journals, Scholarly
Museums
Publishing: Academic and University Presses
Social Science and Civil Society
Social Science and the Media
Social Science and Universities
Social Science Professions and Professionalization
Survey Research Centers and Companies

Social Science Infrastructure
Social Science Infrastructure: Africa
Social Science Infrastructure: Asia
Social Science Infrastructure: Central and South America
Social Science Infrastructure: East Asia and the Pacific (Research and Teaching)
Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations)
Social Science Infrastructure: European Union (Resourcing and Funding)
Social Science Infrastructure: European Union (Social Science Academies and Related Organizations)
Social Science Infrastructure: North America (Research and Teaching)
Social Science Infrastructure: North America (Resourcing and Funding)
Social Science Infrastructure: North America (Social Science Academies and Related Organizations)
Social Science Infrastructure: European Union (Research and Teaching)

Labor Studies

Development
Child Labor
Development and Employment: A Mutual Relationship
Informal Work in Developing Countries
Labor Market Institutions and Labor Market Outcomes in Transition Economies
Migration and Development

Family and Demography
Aging and the Labor Market
Assimilation of Immigrants
Children and Families in Social Work
Family and Work
Female power at work
Gender and Women's Studies, applied research on
Gender, Economics of
Household Economics
Immigration, Applied Research on
International Migration
International Migration and Remittances
Migration, Economics of
Old Age Work Participation
Retirement, Economics of

Value of Life, Economics of
Youth Joblessness

Institutions
Choice, Empowerment, and Social Work
Job Displacement Insurance
Labor Market Policy
Minimum Wages
Mismatch on the Labor Market
Poverty Policy
Shadow Work: Measurement
Strikes
Unemployment: Structural
Vocational Training
Wage Differential and Structure
Welfare and Work
Welfare Reform
Working Conditions
Discrimination, Economics of
Economic Education
Poverty: Measurement and Analysis
Trade Unions, Economic Behavior of
Vocational Education

Labor Performance
Beauty and the Labor Market
Behavioral Economics
Discrimination: Racial
Earnings Inequality
Education and Employment
Entrepreneurship and Economic Growth
Ethnic Identity and Work
Green Jobs
Happiness and Work
Income Distribution and Inequality: Measurement Issues
Informal and Underground Economics
Innovation and Employment
Labor Supply
Language and Labor Market Success
Obesity and Labor Market Performance
Promotions
Time Use Research: Recent Developments
Work, Sociology of

Law

Administrative Law: The United States and Beyond
Affirmative Action: Comparative Policies and Controversies
Affirmative Action: Empirical Work on its Effectiveness
Appeals: Legal
Bankruptcy
Business Law
Children and Law
Civil Law
Civil Liberties and Human Rights
Civil Rights
Class Actions, Legal
Common Law
Compliance and Legal Authority
Constitutional Courts
Constitutionalism, Comparative
Contracts: Legal Perspectives
Corporate Law: Publicly Traded Corporations
Courts and Adjudication
Crime Policy in the United States
Criminal Liability and Criminal Defenses
Criminal Procedure: Adversarial and Inquisitorial Legal Systems
Critical Legal Studies
Critical Race Theory
Death Penalty
Deterrent Effect of Police and Prisons
Discovery: Legal Aspects
Discrimination and the Law
Disputes, Social Construction and Transformation of
Domestic Violence: Legal Perspectives
Due Process in Law
East Asian Legal Traditions
Empirical Legal Studies
Employment and Labor, Regulation of,
Equality and Inequality: Legal Aspects
Equity in Law
European Union Law
Expert Testimony
Family Law in the United States
Feminist Legal Theory
Gender and the Law
Globalization: Legal Aspects
Health Care: Legal Issues (United States)
Human Rights: Legal Aspects
Immigration, Emigration, and Citizenship
Indigenous Rights
In-House Counsel
Injustice: Legal Aspects
Insurance and the Law
Intellectual Property: Law in Context
International Arbitration
International Law and Treaties
Judges
Judicial Review
Juries
Justice and Law
Justice, Access to
Juvenile Justice: International Law and Children's Rights
Latin-American Legal Systems
Law and Aging
Law and Courts in Authoritarian States
Law and Democracy
Law and Development
Law and Economics: Empirical Dimensions
Law and Emotion
Law and Everyday Life
Law and Geography
Law and Neuroscience
Law and People with Disabilities
Law and Risk
Law and Society: Development of the Field
Law and Technology
Law and Terrorism in the US and Beyond
Law as an Instrument of Social Change
Law as Constitutive
Law Firms
Law, Autopoiesis in
Law, Bioethics, and Biotechnology
Law, Mobilization of
Law, Sociology of
Law: Change and Evolution
Law: Economics of its Public Enforcement
Law: History of its Relation to the Social Sciences
Law: The Socio-Legal Perspective
Laws, Conflict of
Lawyers, Regulation of
Lawyers: Social Organization of the Profession
Legal Culture and Legal Consciousness
Legal Formalism
Legal Issues: Public Opinion
Legal Personality
Legal Pluralism
Legal Positivism
Legal Process and Social Science: United States
Legal Realism
Legal Reasoning and Argumentation
Legal Reasoning: Computational Models
Legal Scholarship
Legal Systems, Classification of
Legal Systems: Private
Legal Traditions of Southeast Asia
Legalization: The Expansion of Law's Domain
Lex Mercatoria
Litigation
Mediation, Arbitration, and Alternative Dispute Resolution (ADR)
Middle Eastern and North African Legal Traditions
Natural Law
Nazi Law
Negotiation and Bargaining: Role of Lawyers
Occupational Health and Safety, Regulation of
Organizations and the Law
Paralawyers: Other Legal Occupations
Plea Bargaining: A Comparative Analysis
Political Lawyering
Political Trials
Postcolonial Law
Poverty Law: United States
Privacy: Theoretical and Legal Issues
Property: Legal Aspects
Property: Legal Aspects of Intergenerational Transmission
Prosecution
Prostitution and Sex Work
Public Interest Law: The United States and Beyond
Punishment: Social and Legal Aspects
Race and the Law
Regulation and Administration
Remedies and Damages: Legal Aspects
Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective
Reproductive Rights in Affluent Nations
Rights: Legal Aspects
Rule of Law (and Rechtsstaat)
Scientific Evidence: Legal Aspects

Sexual Harassment: United States and Beyond
Sexual Orientation and the Law
Social Insurance: Legal Aspects
Socialist Law
South Asian Legal Traditions
Special Education in the United States: Legal History
Sub-Saharan African Legal Traditions
Supreme Courts
Torts in Law
Torture: Legal Perspectives
Trials
War Crimes Tribunals, Crimes against Humanity and Transitional Justice

Linguistics A

Biological Preconditions for Language Development
Child-Directed Speech: Influence on Language Development
Continental Philosophy of Language
Evolution and Language: Cultural Transmission
Evolution and Language: Overview
Evolution and Language: Phylogenetic analyses
First Language Acquisition, Developmental Psychology of
Foreign Language Teaching and Learning
Gender and Language: Cultural Concerns
Human Language, Evolution of
Human Language: Evolutionary Precursors
Intentionality in Language and Communication, Emergence of
Language and Thought: The Neo-Whorfian Hypothesis
Language Development, Theories of
Language Development: Emergentist Theories
Language Development: Evolutionary Theories
Adverbial Clauses
Areal Linguistics
Birdsong and Vocal Learning during Development
Chinese Literacy Acquisition: A multidimensional puzzle
Color Terms, Linguistics of
Control, syntactic
Corpus Linguistics
Deixis
Dialectology
Evidentials in Linguistics
Grammatical Relations
Historical Linguistics: Numerical Methods
Iconicity in Linguistics
Indo-European Languages
Language Policy
Linguistic Presupposition
Linguistic Typology
Linguistics: Incorporation
Markedness in linguistics
Phonetics, Articulatory
Phonology
Pidgin and Creole Languages
Possession, linguistic
Reading Nonalphabetic Scripts
Sapir-Whorf Hypothesis
Sign Language
Sign Language: Psychological and Neural Aspects
Speech Errors, Psychology of
Speech Perception
Speech Production, Psychology of
Speech Recognition and Production by Machines
Suprasegmentals
Syntax (general)
Tense, Aspect, Mood and Evidentiality, Linguistics of
Valency and Argument Structure in Syntax
Writing Systems
Writing Systems, Psychology of

Linguistics: B

Anaphora
Case, Morphological: Linguistics
Classifiers, Linguistics of
Code Switching: Linguistic
Communicative Competence: Linguistic Aspects
Comparative Method in Linguistics
Dependency in Language
Diglossia
Etymology
First Language Acquisition, Linguistic Theory of
Functional Approaches to Grammar
Generative Grammar
Gesture in Linguistics
Grammatical Agreement
Grammaticalization
Information Structure in Linguistics
Internal Reconstruction in Linguistics
Kinship Terminology
Language and Ethnicity
Language and Gender
Language and Literature
Language and Poetic Structure
Language and Social Class
Language and Society
Language Contact
Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects
Lexical Semantics
Lexicology and Lexicography
Lexicon
Linguistic Anthropology
Linguistic Fieldwork
Linguistics: Overview
Literacy and Orality
Montague Grammar
Morphology in Linguistics
Morphophonology
Movement Theory and Constraints in Syntax
Multilingualism
Onomastics
Parts of Speech
Politeness and Language
Pragmatics, Linguistic
Quantifiers
Quantitative Linguistics
Second Language Acquisition
Semantic Similarity, Cognitive Psychology of
Semantics
Sociolinguistics
Spelling, Psychology of
Vocabulary Acquisition
Word Order

Logic of Inquiry and Research Design

Analytical Sociology
Archival Methods
Case Study: Logic
Case Study: Methods and Analysis
Case-Oriented Research
Causal Counterfactuals in Social Science Research
Classical (Psychometric) Test Theory
Classification, Understandings of in the Social Sciences
Comparative Studies: Method and Design
Constructivism and Constructionism: Methodology
Content Analysis
Content Validity
Contextual Studies: Methodology
Control Variables in Research
Cross-Cultural Research Methods in Sociology
Data Bases and Statistical Systems: Overview
Data Collection: Interviewing
Dimensionality of tests: methodology
Ecological Fallacy, Statistics of
Emergent properties
Endogeneity: Methodology
Ethnomethodology, General
Event History Analysis: Applications
Event Sequence Analysis: Methods
Experimenter and Subject Artifacts: Methodology
Explanation: Conceptions in the Social Sciences
External validity
Field Observational Research in Anthropology and Sociology
Generalizability Theory
Generalization: Conceptions in the Social Sciences
Grounded Theory: Methodology and Theory Construction
Growth Curve Analysis
Hypothesis Testing: Methodology and Limitations
Indicator: Methodology
Inequality Measurement
Internal Validity
International Social Survey Programme
Interpretive Methods: Macromethods
Interpretive Methods: Micromethods
Interrupted Time Series Quasi-Experimental Designs

Large-Scale Social Surveys, Analysis of
Longitudinal Data: Event–History Analysis in Discrete Time
Longitudinal Research: Panel Retention
Mechanism-Based Causal Analysis
Mediating Variable
Moderator Variable: Methodology
Nonequivalent Group Designs
Observational Studies: Outcome-based Sampling
Panel Surveys: Uses and Applications
Participant Observation
Postmodernism: Methodology
Problem Selection in the Social Sciences: Methodology
Qualitative Methods, History of
Quantitative Cross-national Research Methods
Questionnaires: Cognitive Approaches
Questionnaires: Collection of Subjective Expectations Data
Random Assignment: Implementation in Complex Field Settings
Regression Discontinuity Design
Reliability and Validity Assessment: New Approaches
Reliability: Measurement
Replication
Sample Surveys: Cognitive Aspects of Survey Design
Scaling and Classification in Social Measurement
Selection Bias, Statistics of
Single-Case Designs: Methodology
Statistical Analysis, Special Problems of: Outliers
Structural Equation Modeling
Surveys: Extreme Response
Surveys: Question Wording and Response Categories
Symbolic Interaction: Methodology
Thick Description: Methodology
Time Use Research Methods
Triangulation: Methodology
Visual Methods in the Social Sciences

Management, Organizations, Business, Marketing and Finance

Management Domains
Administration in Organizations
Advertising: General
Authentic Leadership
Business Education
Business Ethics
Business History
Business Law
Business Models
Business Platforms
Business Process: Outsourcing
Career Development: Multilevel Perspective
Corporate Governance
Corporate Law: Publicly Traded Corporations
Corporate Taxation
Customer Satisfaction
Emotions and Work
Ethical Codes, Professional: Business Codes
Greening of Business
Innovation
International Business
Internationalization: Patterns in Business firms
Job Design and Evaluation: Organizational Aspects
Management Accounting
Management: General
Marketing Strategies
Occupational Health and Safety, Regulation of
Occupational Stress Management
Stockholders' Ownership and Control
Strategic Management
Stress at Work
Venture Capital
Work Motivation
Workplace Autonomy
Workplace Deviance
Workplace Safety and Health
Workplace Stress

Management Processes
Bankruptcy
Budgeting and Anticipatory Management
Business and Society: Social Accounting
Corporate Finance: Financial Control
Human Resource Management, Psychology of
Multinational Corporations
Strategizing
Supply Chain Management

Organizational Domains
Bureaucracy and Bureaucratization
Corporate Culture
Cyclical Industrial Dynamics
Development: Organizational
Dominant Designs
Ecology: Organizations
Economic Transformation: From Central Planning to Market Economy
Entrepreneurship and Economic Growth
Entrepreneurship and Organizations
Environmental Economics
Evolutionary Economics
Family Business
Globalization and Transnational Corporations
Hierarchies and Markets
Industrial Cluster Dynamics
Industrial–Organizational Psychology: Science and Practice
Information and Knowledge, Organizational
Information, Economics of
Innovation and Technological Change, Economics of
International Business: Geographic Aspects
Labor Market Institutions and Labor Market Outcomes in Transition Economies
Labor Unions
Leadership in Organizations, Sociology of
Managerial and Organizational Cognition
Modularity and Organizations
Myths and Symbols: Organizational
Organization: Overview
Organizational Behavior, Psychology of
Organizational Climate
Organizational Climate in the Work Setting
Organizational Commitment
Organizational Culture, Anthropology of
Organizational Ecology
Organizational Errors: Catastrophic
Organizational Justice
Organizational Path Dependence
Organizations and Culture
Organizations and the Law
Organizations, Sociology of
Organizations: Authority and Power
People in Organizations
Professions in Organizations
Psychological Contracts
Public Administration: Organizational Aspects
Public Sector Organizations
Social Business
Social Capital in Organizations
Social Innovation
Special Economic Zones and Cluster Dynamics: China
Technology and Organization
Voluntary Organizations
Vygotsky's Theory of Cognitive Development
Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences

Organizational Processes
Absorptive Capacity (of Organizations)
Agency Theory
Automation: Organizational Studies
Behavior Analysis, Applied
Behavioral Economics
Boundaries and New Organization Forms
Commodity Chains
Design and Form: Organizational
Diversification, Relatedness, and the New Logic of Co-Creation
Group Processes in Organizations
Group Processes, Social Psychology of
Industrial Dynamics: Patenting Perspectives
Interorganizational Relationships and Networks
Learning: Organizational
Negotiation and Bargaining: Organizational Aspects
Network Paradigm: Applications in Organizational Science
Neuromarketing
Organizational Citizenship Behavior
Organizational Control
Organizational Decision Making
Organizational Emergence and Firm Formation
Research and Development in Organizations
Strategic Adaptation

Theories of Organizations
Behavioral Theories of Organization
Business Anthropology
Closed and Open Systems: Organizational
Industrial Ecology
Innovation, Theory of
Institutionalism and Neo-institutionalism: History of the Concepts
Institutions
Modern Management: Origins and Development
Modernization, Sociological Theories of
Narrative Approaches to Organizations
Network Analysis
Networks in Economics
Neuroeconomics
New Institutionalism in the Analysis of Complex Organizations
New Managerialism and New Public Sector Management
Organizational Fields
Organizations: Negotiated Orders
Organizations: Unintended Consequences
Rational Choice and Organization Theory
Schumpeter, Joseph A (1883–1950)
Structural Contingency Theory
Weber, Max (1864–1920)
Weberian Social Thought, History of

Mathematics and Computer Sciences Applications

Algorithm
Algorithmic Complexity
Artificial and Natural Computation
Artificial Intelligence: Combinatorial Search
Artificial Intelligence: Connectionist and Symbolic Approaches
ArtificIal Intelligence: Genetic Programming
Auditory Scene Analysis: Computational Models
Autonomous Agents
Axiomatic Theories
Bayesian Theory, History, Applications, and Contemporary Directions
Best Worst Scaling: Theory and Practice
Classical Mechanics and Motor Control
Clustering and Cohesion in Networks: Concepts and Measures
Cognition: Knowledge Spaces
Cognitive Modeling: Connectionist Approaches
Computational approaches to model evaluation
Computational organizational theory
Computerized test construction
Concept Learning and Representation: Models
Connectionist Models of Concept Learning
Constraint-satisfaction Models
Core-periphery Patterns in Networks
Decision and Choice: Bounded Rationality
Decision and Choice: Economic Psychology
Decision and Choice: Heuristics
Decision and Choice: Luce's Choice Axiom
Decision and Choice: Random Utility Models of Choice and Response Time
Decision and Choice: Sequential Decision Making
Deductive Reasoning Systems
Diffusion and Random Walk Processes
Discrete-State Models of Information Processing
Dynamic decision making
Face Recognition Models: computational approaches
Fourier Analysis
Game Theory
Information Processing Architectures: Fundamental Issues
Information Theory
Kinship, Formal Models of
Knowledge Representation
Learning: Mathematical Learning Theory, History of
Logics for Knowledge Representation
Majorization and Stochastic Orders
Markov Decision Processes
Markov Processes in Learning Spaces
Mathematical Models in Philosophy of Science
Mathematical Psychology
Mathematical Psychology: History
Measurement Theory: Conjoint
Measurement Theory: Conjoint Analysis Applications
Measurement Theory: History and Philosophy
Measurement, Representational Theory of
Memory Models: Quantitative
Model Testing and Selection, Theory of
Models of Influence in Networks
Models of Power in Exchange Networks
Motor Control Models: Learning and Performance
Network Centrality, Measures of
Network Models of Tasks
Networks in Economics
Nonstandard Reasoning
Optimal Control Theory
Ordered Relational Structures
Partial Orders
Problem Solving and Reasoning, Psychology of
Problem Solving and Reasoning: Case-Based
Procrustes Analysis
Psychometrics
Psychometrics: Combinatorial Data Analysis
Psychometrics: Preference Models with Latent Variables
Psychophysical Theory and Laws, History of
Scientific Discovery, Computational Models of
Self-Organizing Dynamical Systems
Sensation and Perception: Direct Scaling
Signal Detection Theory
Signal Detection Theory, History of
Social Network Algorithms and Software
Social Networks, Algebra of
Social Simulation: Computational Models
Statistical Pattern Recognition
Stochastic Models
Structural Balance Theory: Recent Developments
Structural Equivalence: Meaning and Measures
Time Perception Models
Virtual Reality and Spatial Cognition
Vision, High-Level Theory of
Vision, Low Level Theory of
Visual Space, Geometry of

Media Studies and Mass Communication

Advertising Agencies
Advertising and Advertisements
Advertising, Control of
Advertising: Effects
Advertising: General
Agenda Setting, Media Effects on
Audience Measurement
Audiences, Media
British Cultural Studies
Broadcasting: Regulation
Celebrity
Citizen Journalism
Communication, Twostep Flow of
Community and Media
Computer Mediated Communication
Documentary and Ethnographic Film
Entertainment
Film and Video Industry
Film History
Film: Genres and Genre Theory
Freedom of the Press
Hegemony and Cultural Resistance
Human-Computer Interfaces
Identity Offline and Online
Information Society
International Advertising
International Communication: History
Journalism
Journalism and Journalists
Libraries
Mass Communication: Normative Frameworks
Mass Media and Cultural Identity
Mass Media and Sports
Mass Media, Political Economy of
Mass Media, Representations in
Mass Media: Introduction and Schools of Thought
Media and Social Movements
Media Effects
Media Effects on Children
Media Events
Media Imperialism
Media Talk Shows
Mobile Communication
Moral Panics
Narrative, Sociology of
New Media and Democracy in the Arab World
New Media and Social Capital
New Media and the Digital Divide
New Media, News Production and Consumption

News Interview
News: General
Online Dating
Photography as a Medium
Political Advertising
Political Communication
Postal Systems
Printing as a Medium
Public Broadcasting
Public Sphere and the Media
Publishing as Medium
Radio as a Medium
Religion and New Media
Reputation
Rhetorical Analysis
Science and Media
Science Communication
Semiotics
Social Media
Social Protest and New Media
Telegraph
Telephone
Television: General
Television: History
Terror and the Internet
Textbooks
Visual Images in the Media

Memory: Cognitive and Neuroscientific Aspects

Aging and Memory
Amnesia: General
Amnesia: Psychogenic
Associative Learning in the Cerebellum
Autobiographical Memory
Change Blindness
Childhood Memories
Confabulation
Context Dependent Memory
Declarative Memory, Neural Basis of
Dysexecutive amnesia
Episodic memory
Eyewitness testimony
False memories
Flashbulb Memories
Future Thinking
Genetic Approaches to Memory
Hippocampal Amnesia
Hippocampus and Related Structures
Implicit Memory
Learning and Memory, Neural Basis of
Lexical Consolidation
Memory Binding
Memory in the Bee
Non declarative (procedural) memory
Post Traumatic Amnesia
Proactive Interference
Prospective Memory
Recognition
Reconstructive Memory, Psychology of
Reinforcement: Neurochemical Substrates
Retrieval
Retroactive Interference
Semantic Memory
Short Term Memories, Theories of
Short-Term Memory: Psychological and Neural Aspects
Sleep and Memory
Source Monitoring
Spatial Navigation
Synaptic Mechanisms and Protein Synthesis in Memory
Temporal Distinctiveness Models of Memory
Transient Epileptic Amnesia
Transient Global Amnesia
Visual Imagery
Visual Memory, Psychology of
Working Memory, Neural Basis of
Working Memory, Psychology of

Motivational Psychology

Academic Achievement Motivation, Development of
Academic Engagement
Academic Motivation and Performance: Task Value Interventions
Affect-Regulation Motivation
Apprenticeship and School Learning: Lessons from Germany
Avoidance and Approach Motivation: A Brief History
Broaden-and-build Theory of Positive Emotions
Burnout, Psychology of
Control Behavior: Psychological Perspectives
Coping across the Lifespan
Decision Making: Nonrational Theories
Dimensional comparisons and their consequences for self-concept, motivation and emotion
Disgust, Psychology of
Emotion and Expression
Emotions, Sociology of
Epistemological Beliefs and Motivation
Eudaemonism
Expectancy-Value-Cost Model of Motivation
Flow in Motivational Psychology
Gender and Academic Motivation
Grit
Health Self-Regulation, Motivational and Volitional aspects of
Interest, Psychology of
Internal/External Frame of Reference Model
Leisure Activities Choices among Adolescents
Mastery Learning
Motivation and Actions, Psychology of
Motivation and Emotion: Reflections on the History of an Attributional Approach
Motivation in Australian Aboriginal Populations
Motivation in Youth Sport and Physical Activity: Developmental Perspectives
Motivation, Familial Influences on
Motivation, Learning, and Instruction
Motivation: Life Course and Sociological Perspectives
Optimism and Pessimism
Passion and Motivation
Personal Projects
Personality and Adaptive Behaviors
Personality and Risk Taking
Race and Academic Motivation
School Achievement: Motivational Determinants and Processes
School Burnout and Engagement: Lessons from a Longitudinal Study in Finland
Schooling: Impact on Cognitive and Motivational Development
Self and Emotional Development in Adulthood and Later Life
Self in Culture: Early Development
Self-Determination Theory
Self-regulated Learning: Theories, Measures, and Outcomes
Selfregulation in Adulthood
Social and Economic Change: Psychological Challenges for Individuals
Socio-digital Revolution: Digital Natives vs Digital Immigrants
Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model
Teacher Motivation
Temperament and Motivation
Test Anxiety and Academic Achievement
Vocational Interests, Values, and Preferences, Psychology of
Wanting vs Needing
Adulthood: Dependency and Autonomy
Data Collection: Interviewing
Emotion: History of the Concept
Emotions, Psychological Structure of
Erikson, Erik Homburger (1902–94)
Growth Curve Analysis
Health Behavior, Psychosocial Theories of
Large-Scale Social Surveys, Analysis of
Lewin, Kurt (1890–1947)
Longitudinal Data: Event–History Analysis in Discrete Time
Longitudinal Research: Panel Retention
Mobility, Social
Norms
Personality and Self
Power
Power in Society
Questionnaires: Collection of Subjective Expectations Data
Racial Relations
Self and Identity Development during Adolescence across Cultures
Self-Esteem
Social Influence, Sociology of
Social Psychology, Sociological
Subjective Wellbeing and Culture
Subjective Wellbeing, Psychology of
Values, Social Psychology of

Neuroscience of Language

Agraphia (Acquired Dysgraphia)
Aphasia
Asymmetry of Body and Brain
Bilingual Aphasia

Brain and Behavioral Lateralization in Animals
Classical Tests for Speech and Language Disorders
Dichotic Listening and Language: Overview
Discourse and Indentity
Dyslexia, Developmental
Dyslexias, Acquired and Agraphia
Eye Movements in Reading
Fluency and Voice, Disorders of
Functional Brain Imaging of Language processes
Gender: Grammatical
Hallucinations, Neural Basis of
Handedness and Language, Relationship between
Interhemispheric Interaction in Language and Cognitive Processes
Intracarotid Sodium Amytal Test and Language
Language Acquisition
Language Disorders, Treatment and Remediation of
Language Learning Impairment
Language Processes in Delirium and Dementia
Language Processing, Functional Magnetic Resonance Imaging of
Lateralization of Language as Demonstrated by Brain Imaging Procedures
Linguistic Approaches to Agrammatism
Linguistic Synesthesia
Negation, Linguistics of
Neural Basis of Language Development
Neural Basis of Reading
Neurolinguistic Processing of Psychological Verbs
Neurolinguistics, History of
Neurological Approaches to Agrammatism
Neuroprosthetics
Neuropsychological Testing
Number, Grammatical
Phantom Limbs
Phological Disorders
Pragmatic Disorders
Proper and Common Names: Impairments of Anomia
Serial Verb Constructions
Sign Language, Disorders of
Speech Processing, The Cortical Organization of
Speech Production, Neural Basis of
Split Brains and Human Consciousness
Transcranial Direct Current Stimulation and Language

Personality Psychology

Agreeableness
Authoritarianism
Big Five Factor Model, Theory and Structure
Chronotype
Conscientiousness
Coping, Personality and
Core Self-Evaluations
Cultural Intelligence and Competencies
Dark Triad: The "Dark Side" of Human Personality
Emotion in Cognition
Emotion, Perception and Expression of
Emotional Intelligence and Competencies
Emotional Regulation
Emotions and Aging
Emotions and Health
Emotions, Psychological Structure of
Emotions: Methods of Assessment
Extraversion
Facial Emotion Expression, Individual Differences in
Facial Perception
Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis
Factor Analysis and Latent Variable Models in Personality Psychology
Five Factor Model of Personality, Assessment of
Five Factor Model of Personality, Facets of
Five Factor Model of Personality, Personality Disorder
Five Factor Model of Personality, Universality of
Gender Differences in Personality and Social Behavior
Health Risk Appraisal and Optimistic Bias
Honesty and Humility
Hope and Hopelessness
Implicit Association Test
Intelligence: Assessments of
Intelligence: Central Conceptions and Psychometric Models
Intelligence: Historical and Conceptual Perspectives
Interpersonal Attraction, Psychology of
Minnesota Multiphasic Personality Inventory
Moral Reasoning in Psychology
Neuroticism
Openness to Experience
Person Perception, Accuracy of
Personal intelligence and competencies
Personality and Creativity
Personality and Economics
Personality and Educational Outcomes
Personality and Marriage
Personality and Religion
Personality and Self
Personality Assessment, Faking and
Personality Assessment, Forced-Choice
Personality Assessment: Overview
Personality Differences and Development: Genetic and Environmental Contributions
Personality Theory and Psychopathology
Personality, Biological Models of
Personality, Cognitive Models of
Personality, Evolutionary Models of
Personality, Trait Models of
Personality: Historical and Conceptual Perspectives
Prejudice and Discrimination
Procrastination
Self-Concept: From Unidimensional to Multidimensional and Beyond
Selfconscious Emotions, Psychology of
Self-Efficacy
Self-Esteem
Self-Regulated Learning
Sensation Seeking: Behavioral Expressions and Biosocial Bases
Sibling-Order Effects
Situational Judgment Test
Social Beliefs and Attitudes: Personality and the Isms
Social Intelligence and Competencies
Stress: Psychological Perspectives
Subjective Wellbeing and Culture
Subjective Wellbeing, Psychology of
Temperament
Time Use and Gender
Values, Social Psychology of
Wellbeing and Burnout in the Workplace, Psychology of
Wisdom, Psychology of

Philosophy

Analytical Marxism
Atomism and Holism: Philosophical Aspects
Autonomy, Philosophy of
Business Ethics
Causation: Physical, Mental, and Social
Causes and Laws: Philosophical Aspects
Consequentialism (Including Utilitarianism)
Continental Philosophy of Language
Contractarianism
Conventions and Norms: Philosophical Aspects
Counterfactual Reasoning, Qualitative: Philosophical Aspects
Counterfactual Reasoning, Quantitative: Philosophical Aspects
Critical Rationalism
Critical Theory
Democracy: Normative Theory
Deserts, Philosophical Concept of
Diversity and Disagreement in Ethics: Philosophical Implications
Economics and Ethics
Economics, Philosophy of
Equality: Philosophical Aspects
Ethics and Values
Evolution, Natural and Social: Philosophical Aspects
Evolutionary Approaches to Human Behavior: Philosophical Aspects
Free Will and Action
Freedom: Political
Functional Explanation: Philosophical Aspects
Hermeneutics
Historical Explanation, Theories of: Philosophical Aspects
Idealization, Abstraction, and Ideal Types
Identity and Identification
Ideologies, Institutions, and the New Institutionalism

Individualism versus Collectivism: Philosophical Aspects
Intentionality and Rationality
Intercultural Legal Discourse: Philosophical Aspects
Intergenerational Justice
Irrationality: Philosophical Aspects
Joint Action
Justice: Philosophical Aspects
Kantian Ethics and Politics
Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects
Law and Morality: A Continental European Perspective
Law and Morality: An Analytical Perspective
Libertarianism
Logical Positivism and Logical Empiricism
Market and Nonmarket Allocation
Meaning and Rule Following
Mental Causation
Methodological Individualism: Philosophical Aspects
Multiculturalism
Naturalism
Objectivity: Philosophical Aspects
Person and Self: Philosophical Aspects
Personal Identity: Philosophical Aspects
Phenomenology: Philosophical Aspects
Philosophy of Law
Physicalism and Alternatives
Pluralism and Tolerance
Postmodernism: Philosophical Aspects
Practical Reasoning: Philosophical Aspects
Pragmatism: Philosophical Aspects
Prisoner's Dilemma, One Shot and Iterated
Public Reason
Rational Choice Explanation: Philosophical Aspects
Reference and Representation: Philosophical Aspects
Reflective Equilibrium
Relativism: Philosophical Aspects
Republicanism: Philosophical Aspects
Responsibility: Philosophical Aspects
Scientific Explanation
Self: Philosophical Aspects
Shared Belief: Philosophical Aspects
Social Constructivism
Social Ontology
Social Properties (Facts and Entities): Philosophical Aspects
Social Science, The Idea of
Trust: Philosophical Aspects
Value Pluralism
Virtue Ethics
Welfare: Philosophical Aspects
Wisdom: Philosophical Aspects
World Republic

Political Science

Agendas: Political
Alliances: Political
Appeasement: Political
Attitudes, Political and Civic Culture
Attitudes, Political and Public Opinion
Attitudes, Political: Authoritarianism and Tolerance
Bounded Rationality
Campaigning: Political
Church and State
Citizenship: Political
Communitarianism
Congress: United States
ConservŠatism: Theory and Contemporary Political Ideology
Consociationalism
Constitutions
Delegation of Power: Agency Theory
Deliberation and Democracy
Democratic Theory
Democratic Transitions
Dictatorship
Economic Evaluations and Electoral Consequences
Efficacy, Political
Elections: Apportionment
Electoral Systems
Executive Government
Federalism
Feminist Political Theory and Political Science
Generations: Political
Governments
Green Parties
Ideological Constraint: History and Current Status of the Concept
Immigration: Political Aspects
Imperialism
Implementation: Political
Institutionalism
Interest Groups
Islamic Fundamentalism
Issue Evolution
Judicial Behavior
Judicial Independence
Liberalism: Political Doctrine and Impact on Social Science
Majoritarianism and Majority Rule
Marxism-Leninism: The ideology of Twentieth-Century Communism
Mass Media and Politics
Modernization
National Security Studies and War Potential of Nations
Oligarchy, Iron Law of
Opinion Formation, Theory of
Parliamentary Government
Party Identification and Nonpartisanship
Party Systems
Pluralism
Political Money and Party Finance
Political Parties: Communist
Political Parties: Extreme Right
Political Parties: Manifestoes
Political Protest and Civil Disobedience
Political Representation
Political Repression
Political Science and Empirical Implications of Theoretical Models
Political Science and Polling
Political Tolerance
Politics and Participation
Politics and Socialization
Politics, Formal Modeling of
Pooled Cross-Sectional and Time Series Analyses, in Political Science
Populism
Power
Presidency, United States
Presidential Government
Proportional Representation
Public Bureaucracies
Public Interest
Realism/Neorealism
Sanctions in Political Science
Social Capital
Social Democracy
Socialism
State Formation, Theory of
Totalitarianism
Trust, Political
Voting and Direct Democracy
Voting, Explanations of: Social Class
Voting: Turnout

Psychiatry

Change Blindness
Cross-Cultural Research Methods in Sociology
AIDS: Acquired Immune-deficiency Syndrome
Alcohol-Related Disorders
Antidepressant Drugs
Anxiety and Anxiety Disorders
Behavior Therapy: Psychiatric Aspects
Bipolar (Manic-Depressive) Disorder
Borderline (Patient) Personality
Caregiver Burden
Cerebellum: Cognitive Functions
Child and Adolescent Psychiatry, Principles of
Child Neglect
Childcare and Development across Cultures
Childhood Cancer: Psychosocial Aspects
Children, Rights of: Cultural Concerns
Chronic illness, psychosocial coping with
Chronic Pain: Models and Treatment Approaches
Classical Conditioning Methods in Psychotherapy
Cognitive Behavioral Therapy
Cognitive Neuropsychology, Methods of
Co-morbidity of Mental and Physical Conditions
Concept Formation in Childhood: Cross-Cultural Perspectives
Concepts and Methods of Cognitive Therapies
Cultural Psychiatry
Culture as a Determinant of Mental Health
Culture-specific Psychiatric Disorders
Deinstitutionalisation in Low and Middle Income Countries
Deinstitutionalization, Psychiatry of
Dementia: Psychiatric Aspects
Depression
Differential Diagnosis in Psychiatry
Education in Psychiatry
Endocrinology and Psychiatry

Ethics and Psychiatry
Evidence-Based Psychiatry
Family and Psychiatric Illness: Family Therapy
Forensic Psychiatry
Geriatric Psychiatry
Health Psychology in Childhood
Hypnosis, Psychiatry of
Hysteria: History and Critiques
Liaison and Consultation Psychiatry
Magnetic Resonance and Molecular Imaging in Psychiatry
Mania
Mental Disorders, Epidemiology of
Mental Illness, Etiology of
Mental illness, Genetics of
Mental Illness: Family and Patient Organizations in Europe
Mentally Ill, Public Attitudes towards
Military and Disaster Psychiatry
Normality and Mental Health
Nosology in Psychiatry
Obsessive–Compulsive Disorder
Panic Disorder
Panic Disorder and Agoraphobia Across the Lifespan
Perinatal Psychiatry
Personality Disorders
Perversions, Sexual (Paraphilias)
Philosophy, Psychiatry, and Psychology
Placebo Studies (Double-Blind Studies)
Post-traumatic Stress Disorder
Prophylaxis in Psychiatry
Psychiatric Assessment: Negative Symptoms
Psychiatric Care, Economics of
Psychiatry and Religion
Psychiatry: Anthropological Aspects
Psychiatry: Informed Consent and Care
Psychoanalysis: Current Status
Psychology of the Real Self: Psychoanalytic Perspectives
Psychopharmacotherapy, Costs of
Psychopharmacotherapy: Side Effects
Psychosomatic Medicine
Psychotherapy and Pharmacotherapy
Public Psychiatry and Public Health Mental Systems
Schizophrenia
Schizophrenia, Treatment of
Schizophrenia: Metabolic Consequences
Sleep Disorders: Psychiatric Aspects
Social Intervention in Psychiatric Disorders
Sociology and Psychiatry
Somatoform Disorders
Somatotherapy, History of
Suicide
Therapeutic Community
Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts
Women and Psychiatry
Cognitive Aging
Cognitive Dissonance
Discrimination: Racial
Childhood and Adolescence: Developmental Assets
Cognitive Control (Executive Function): Role of Prefrontal Cortex
Cognitive Development during Infancy and Early Childhood across Cultures
Cognitive Development in Infancy, Neural Mechanisms of
Cognitive Development, Educational Theories of
Cognitive Development: Child Education
Cognitive Neuroscience
Cross-Cultural Psychology
Cross-Cultural Research Methods in Psychology
Eating Disorders

Public Policy

Agenda Setting, Public Policy in
Budgeting: Political Science Aspects
Civil Service
Consumer Policy
Corruption: Political and Public Aspects
Defense and Security Economics
Family Policy
Financial Regulation
Global Governance
Global Public Policy
Governance
Health Politics and Policy
Immigration Policy
Internet and Privacy
Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks
Labor Market Policy
Pension Policy
Policy Analysis: Evidence Based Policy-Making
Policy Cycle
Policy Knowledge: Foundations
Policy Knowledge: New Liberalism
Policy Knowledge: Universities
Policy Networks
Policy Process: Nonprofit Institutions in the United States
Poverty Policy
Public Goods: International
Public Policy
Public Policy Schools
Public Policy: Ethics
Quality of Government
Regulatory Agencies
Skills Training: Public Policy Aspects
Think Tanks
Trade Policy
Welfare and Education
Welfare Policy Comparisons
Welfare Retrenchment
Working Conditions

Religious Studies

Buddhism
Catholicism
Catholicism and Gender
Charisma and Charismatic
Christianity in Asia
Christianity in Latin America
Christianity in the English-Speaking West
Christianity, Origins of
Civil Religion
European Religion from the Atlantic to the Black Sea
Evangelical, Revivalist and Pentecostal Religions
Globalization and Religion
Islam: Asia
Islam: Middle East
Islam: Sub-Saharan Africa
Islamic Law
Judaism
Judaism and Gender
Liberal Christianity
Millennialism
Mission: Proselytization and Conversion
Monasticism: West and East
Mysticism
Myth in Religion
New Religious Movements
Orthodox Churches
Pilgrimage
Prophetism
Religion and Abuse
Religion and Economic Life
Religion and Health: Sociological Considerations
Religion and Politics in the United States
Religion and Sciences
Religion and Stratification
Religion and Youth
Religion, Phenomenology of
Religion, Psychology of
Religion, Sociology of
Religion: Evolution and Development
Religion: Family and Kinship
Religion: Mobilization and Power
Religion: Morality and Social Control
Religion: Nationalism and Identity
Religion: Psychiatric Aspects
Religions of East Asia
Religions of India
Religiosity
Religiosity and Prejudice against Minorities
Religious Fundamentalism: Cultural Concerns
Religious Organizations
Science and Religion
Secularization
Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns
Women in Buddhism
Women's Religiosity

Science and Technology Studies

Agricultural Science and Technology

Conceptual Issues in Science and Technology Studies
Archaeology: Philosophy and Science
Biocitizenship

Bioethics: Philosophical Aspects
Citizen Science
Enlightenment: Impact on the social sciences
Ethical Dilemmas, Research and access to treatment
Experiment, in Science and Technology Studies
Genomics, Ethical Issues in
Ideology, Sociology of
Ignorance, History of Concept
Indigenous Knowledge: Science and Technology Studies
Knowledge Society, History of
Meaning and Rule Following
Medicalization: Sociological and Anthropological Perspectives
Modernity
Modernity: Anthropological Aspects
Modernity: History of the Concept
Multiculturalism
Multiculturalism, Anthropology of
Multisited Field Studies
Norms in Science
Objectivity: Philosophical Aspects
Peer Review: Organized Skepticism
Practical Reasoning: Philosophical Aspects
Reference and Representation: Philosophical Aspects
Reflexivity in Anthropology
Reflexivity in Science and Technology Studies
Reification: History of the Concept
Relativism, Cognitive
Relativism: Philosophical Aspects
Scientific Controversies
Scientific Explanation
Scientific Revolution, History and Sociology of
Tacit Knowledge, Psychology of
Technological Determinism
Thick Description: Methodology
Trust, Sociology of
Truth and Credibility in Science

Key Topics on Science, Technology, and Biomedicine
Academic Careers in Comparative Perspective
Bioethics in the Post-genomic era
Bioethics: Genetics and Genomics
Biotechnology
Classification, Understandings of in the Social Sciences
Climate Change
Computers: Impact on the Social Sciences
Cybernetics
Darwinism: Social
Data Bases and Statistical Systems: Archives and Historical Databases
Energy and Society
Environmental Sciences
Eugenics, History of
Expert Testimony
Expertise, Anthropological Approaches to
Galton, Sir Francis (1822–1911)
Genetic Engineering
Genetics and Anthropology
Genetics and Society
Health: Anthropological Aspects
Human Genome Diversity Project: History
Human-Computer Interfaces
Intellectual Property: Law in Context
Intellectuals, Sociology of
Internet and Culture
Internet and Privacy
Internet and Social Media: Anthropological Aspects
Law and Neuroscience
Law and Technology
Law, Bioethics, and Biotechnology
Museums
Nanotechnology: Social and Cultural Aspects
Patenting in the Biological Sciences
Peer Review and Quality Control
Professions, Sociology of
Race: Genetic Aspects
Religion and Sciences
Risk, Sociology and Politics of
Science and Development
Science and Media
Science and Religion
Science and Technology, Social Study of: Computers and Information Technology
Science Communication
Science Education
Science, New Forms of
Scientific Concepts during Childhood, Development of
Scientific Concepts: Development in Children
Scientific Discovery, Computational Models of
Scientific Evidence: Legal Aspects
Stem Cell Research and Cloning
Technology and Organization
Technology Transfer
Technology, Evolution of
Universities and Science and Technology: Europe
Universities and Science and Technology: United States
Visualization in Science and Technology

Perspectives on Science and Technology
Actor-Network Theory
Bioethics
Bourdieu, Pierre (1930–2002)
Constructivism and Constructionism: Methodology
Cultural Studies of Science
Culture and Actor Network Theory
Engineering Studies
Epistemic Cultures
Ethnography
Ethnomethodology, General
Evolutionary epistemology
Feminism in Anthropology
Feyerabend, Paul (1924–94)
Field Observational Research in Anthropology and Sociology
Functionalism, History of
Garfinkel, Harold (1917–2011)
Gender, Feminism, and Sexuality in Archaeological Studies
Goffman, Erving (1922–1982)
Husserl, Edmund (1859–1938)
Innovation and Technological Change, Economics of
Innovation, Theory of
Institutionalism
Knowledge, Anthropology of
Kuhn, Thomas S. (1922–96)
Labeling Theory, History of
Laboratory Studies: Historical Perspectives
Latour, Bruno (1947–)
Logical Positivism and Logical Empiricism
Mannheim, Karl (1893–1947)
Medical Sociology
Mumford, Lewis (1895–1990)
Participant Observation
Popper, Karl Raimund (1902–94)
Positivism, History of
Postcolonial Science Studies
Postmodernism in Sociology
Schütz, Alfred (1899–1959)
Science and Technology Studies, Ethnomethodology of
Science and Technology, Anthropology of
Science, Geography of
Science, History of
Science, Public Engagement with
Science, Sociology of
Science: Constructivist Perspectives, History of
Scientific Knowledge, Sociology of
Scientometrics
Situated Knowledge, Feminist and Science and Technology Studies Perspectives
Social Constructionism
Social Constructivism
Social History
Social Science, The Idea of
Sociology and the New Institutionalism
Standpoint Theory, in Science
Strong Program, Sociology of
Technological Change and the Sociotechnical System, Applied Psychology of
Technology, Social Construction of
Visual Sociology
Visual Turn

Politics of Science and Technology
Archaeology, Politics of
Ecofeminism
Ethical Practices: Institutional Oversight and Enforcement
Foucault, Michel (1926–84)
Gender and Technology: From exclusion to inclusion?
Pharmaceutical Industry: Political Economies of Drugs and Knowledge
Science and Politics: Value Neutrality
Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct
Science and Law
Science and Social Movements

Science and Technology Studies: Experts and Expertise
Science and The State
Technology and Social Control
Technology Assessment

Social-Historical Studies of Science and Technology
Biomedical Sciences, History and Sociology of
Chemical Science: History and Sociology
Empiricism, History of
History and the Social Sciences
Human Genome Project: History and Assessment
Human Sciences, History of
Intelligence: History of the Concept
Medicine, History of Western
Modernization and Modernity in History
Nature-Nurture Controversy, History of
Physical Sciences: History and Sociology
Scientific Disciplines, History of
Sociology, History of

Sexuality

Adolescent Sexual Risk
Erotic Capital
Femininities
Heterosexual Expression at Older Ages
Heterosexuality
Heterosexuality in Africa
Heterosexuality in China
Hooking Up and Casual Sex
Intimate Partner Violence
Marital Quality and Health
Masculinity, Scripts of
Methodological Issues in the Study of Sexuality
Parallel Relationships
Popular Sexual Cultures in Western Europe (1930s to the Present)
Satisfaction and Commitment in Long-term Heterosexual Relationships
Sex Life Challenges: The Finnish Case
Sexual Behavior and Social Networks
Sexual Debut
Sexual Dysfunction, Epidemiology of
Sexual Motivation
Sexuality and the Internet
Sexuality Education
Sexuality in Australia
Sexuality Over the Life Course
Sexuality, Theories of
Sexually Transmitted Infections: Social Network Analysis

Social Psychology

Applied Social Psychology
Health Behavior, Psychosocial Theories of
Health Risk Perception
Volunteerism, Psychology of

Attitudes
Attitude Formation and Change
Attitude Measurement
Attitudes and Behavior
Attitudes, Political and Civic Culture
Attitudes, Political and Public Opinion
Attitudes, Political: Authoritarianism and Tolerance

Beliefs
Belief in a Just World
Collective Beliefs, Sociological Explanation of
Common Sense, Anthropology of
Shared Belief: Philosophical Aspects

Collective Action
Collective Action
Collective Behavior, Social Psychology of
Deindividuation, Psychology of
Social Movements, History of: General
Social Movements: A Social Psychological Perspective
Collective Behavior, Sociology of
Identity Movements
Social Protest
Movements, Social
Social Movements and Political Violence

Collective Memory
Collective Memory, Social Psychology of

Concepts
Essentialism
Future Time Perspective and Motivation
Inequality, Social
Justice: Social Psychological Perspectives
Legitimacy, Sociology of
Levels of Analysis in Social Psychology
Mobility, Social
Norms
Power
Power in Society
Prisoner's Dilemma, One Shot and Iterated
Risk: Social Psychological Perspectives
Social Categorization
Social Stratification
Social, The: History of the Concept
Solidarity: History of the Concept
Status and Role, Social Psychology of
Values, Psychology of
Values, Social Psychology of
Social Comparison, Psychology of
Social Dilemmas, Psychology of
Social Dominance Orientation
Reification: History of the Concept
Self-Efficacy
Self-Esteem
Subjective Wellbeing and Culture
Subjective Wellbeing, Psychology of
Trust, Sociology of

Conflict
Conflict and Conflict Resolution, Social Psychology of
Conflict and Consensus
Conflict Mediation
Ethnic Conflicts
Genocide and War
Justice, Conflicts, and the Justice of Conflict Resolution
Mass Killings and Genocide, Psychology of

Culture-Immigration
Acculturation
Collectivism and Individualism: Cultural and Psychological Concerns
Cross-Cultural Psychology
Cultural Influences on Interpersonal Relationships
Cultural Psychology
Culture and Emotion
Culture and the Self: Implications for Psychological Theory
Ethnicity and Migration in Europe
Immigration: Social Psychological Aspects
Indigenous Psychology
Migrants and Migration in Europe
Multicultural Education
Values Across Cultures, Development of
Multiculturalism
Integration and Multiculturalism: Focus on Western Europe

Discipline
Political Psychology
Social Psychology
Social Psychology, Sociological

Emotions
Emotion: History of the Concept
Emotions and Intergroup Relations
Emotions, Psychological Structure of
Emotions, Sociology of
Emotion in Cognition

Ethics
Ethical Guidelines and Codes of Conduct in Social and Behavioral Research

Group Processes
Group Processes, Social Psychology of

Identity
Ethnic Identity, Psychology of
Identity and Identification, Social Psychology of
Personality and Self
Self: History of the Concept
Self: Philosophical Aspects
Social Identity in Social Psychology
Collective Identity
Self and Identity Development during Adolescence across Cultures
Self in Culture: Early Development

Influence
Persuasion Theories
Social Influence, Social Psychology of
Social Influence, Sociology of

Intergroup Relations
Helping Intergroup Relations
Intergroup Relations
Meritocracy and Tokenism
Nationalism, Sociology of
Affirmative Action, Sociology of

Interpersonal Relations
Love and Intimacy, Psychology of
Prosocial Behavior and Empathy
Prosocial Behavior, Cultural Differences in

Method/Statistics
Social Psychology: Research Methods
Personality Assessment: Overview

Prejudice
Dehumanization
Desegregation
Discrimination: Racial
Minorities
Prejudice and Discrimination
Prejudice in Society: Psychological Perspectives
Prejudice in Society: Sociological Perspectives
Racial Relations
Racism, History of
Racism, Sociology of
Racism: Social Psychological Perspectives
Sexism
Xenophobia: Social Psychological Aspects

Social Cognition
Decision Making, Psychology of
Embodied Social Cognition
Heuristics in Social Cognition
Implicit Association Test
Implicit Social Cognition
Social Cognition
Stereotypes in Social Psychology

Social Psychologists
Allport, Gordon W (1897–1967)
Bruner, Jerome Seymour (1915 -)
Campbell, Donald Thomas (1916–96)
Erikson, Erik Homburger (1902–94)
Heider, Fritz (1896–1988)
Lazarsfeld, Paul Felix (1901–76)
Lewin, Kurt (1890–1947)
Mead, George Herbert (1863–1931)
Tversky, Amos (1937–96)
Goffman, Erving (1922–1982)

Social Relations
Aggression, Social Psychology of
Conformity: Sociological Aspects
Cooperation and Competition
Leadership
Obedience (Research), History of
Obedience: Social Psychological Perspectives
Solidarity, Sociology of
Stigma, Social Psychology of
Tyranny

Theories
Attributional Processes: Psychological
Authoritarian Personality
Cognitive Dissonance
Cognitive Psychology: History
Critical Psychology
Deprivation: Relative
Interactionism, Symbolic
Representations, Social Psychology of
Self-Categorization Theory
Social Constructionism
Social Psychological Theory, History of
Social Simulation: Computational Models
Social Constructivism

Social Work

Adoption: Domestic, International and Global Perspectives
Afrocentric Approaches to Social Work
Alcohol Use and Abuse
Antisemitism
Antisocial Behavior
Armed Conflict and Social Work: Significance and Implications
Asylum Seekers
Bologna Process
Charity and Philanthropy: Overview
Child Neglect
Child Protection
Child Protection: International Issues
Child Soldiers as Victims and Actors: Lessons from Ethiopia
Children and Families in Social Work
Choice, Empowerment, and Social Work
Communication Skills in Social Work
Community care in England
Community resilience to disasters
Community Work
Criminal Justice and Social Work Practice
Critical Reflexivity and Social Work Practice
Critical Social Work Practice
Culturally Appropriate Interventions in Social Work
Disabilism
Disability and Disabling practices
Disability Movement
Disasters in Social, Cultural and Political Context
Ecological Social Work: Shifting Paradigms in Environmental Practice
Elder Abuse
Elder Care
Emergency relief and social work responses
Ethnicity, Race, Oppression and Social Work: The Canadian Case
Fathers and Fathering: Social Work Perspectives
Feminist Research and Social Work
Gendered Care and Social Work Practice
Global Agenda for Social Work and Social Development
Globalization: Welfare and Social Work
Green Social Work
Harmony and Social Work Practice
Health Social Work
Human Rights and Social Work
Immigration, Citizenship, and Integration: Social Work Connections
Indigenous Child Welfare Practice: The Canadian Case
Indigenous Social Work
International Association of Schools of Social Work
International Council on Social Welfare
International Issues in Child Protection
International Social Problems: A Systems Perspective
International Social Work
International Social Work Student Exchanges
Islamic Social Work: Hope and the Commitment to Change
Learning Disability and Social Work
Legal Bases of Social Work
Liberation Theology in Social Work
Maori Social Work
Marxist Approaches to Social Work
Masculinity and Social Work Practice
Mental Health and Social Work
Mental Health, Mentalism and Sanism
Migration and Social Work
Migration, Ethnicity, Aging and Social Work Practice
Motherhood
Narrative in Social Work
Neocolonialism and Social Work
Neoliberalism
New Managerialism and New Public Sector Management
Oppression
Orientalism, Oppression, and Contemporary Social Work
Post-Traumatic Stress in Social Work
Probation, Offenders and Risk
Professional Development
Psychiatric Social Work in the United Kingdom
Psychosocial Work
Queer Theory and Intersectionality
Radical Social Work
Recognition, Politics of: Perspectives from Social Work and Sociology
Reflexivity
Relational Social Work
Research and Theory Building in Social Work
Residential Child Care
Responsibility in Social Work
Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice
Risk Assessment in Social Work
Roma People: Discrimination and Social Work Practice
Rural Social Work
Self-Directed Care and Individual Budgets for Older People: Lessons from the UK
Sex offenders
Sexuality and Social Work
Social Action and Social Justice
Social Enquiry and Action Research for Social Work
Social Protest Movements and Social Work Practice
Social Work and Disasters
Social Work and Psychodynamic Thought
Social Work and Refugees
Social Work Education

Social Work Ethics
Social Work in China
Social Work in Eastern Europe
Social Work in Organizations
Social Work in Russia
Social Work Theory
Social Work with Racialized Groups: Frameworks for Practice
Spirituality
Structural Adjustment in Africa: Implications for Social Work
Sustainable Development and Social Work
Trade Unions and Social Work: Lessons from Canada
Trafficking of Women and Children in Latin America
United Nations and its Agencies
Whiteness and White Supremacy
Women and Criminal Justice

Sociology

Action, Theories of Social
Adolescence, Sociology of
Affirmative Action, Sociology of
Alienation, Sociology of
Alienation: Psychosociological Tradition
Altruism and Prosocial Behavior, Sociology of
Anomie
Asceticism
Attempted Utopias and Intentional Communities
Authority, Social Theories of
Biography and Society
Bureaucracy, Sociology of
Capitalism: Global
Charisma, Social Aspects of
Citizenship, Sociological Aspects of
Civil Rights Movements
Class Consciousness
Class: Social
Collective Behavior, Sociology of
Collective Beliefs, Sociological Explanation of
Collective Memory, Pyschology of
Collective Memory, Sociology of
Communitarianism, Sociology of
Community Sociology
Community, Social Context of: The US Case
Concentrated Disadvantage
Conformity: Sociological Aspects
Consumption, Sociology of
Control, Social
Conversation Analysis: Sociological
Cooperation: Sociological Aspects
Counterculture: 1960s and Beyond
Counterculture: The Classical View
Crime, Immigration and Ethnicity
Critical Theory: Contemporary
Culture, Sociology of
Darwinism: Social
Death and Dying, Sociology of
Dependency Theory
Deprivation: Relative
Differentiation: Social
Diffusion in Sociological Analysis
Disability: Sociological Aspects
Discrimination: Racial
Divorce, Sociology of
Dramaturgical Analysis
Economic Sociology
Elites: Sociological Aspects
Emotions, Sociology of
Equality of Opportunity
Ethnic, Racial, and Nationalist Movements
Ethnicity, Sociology of
Evolutionary Sociology
Exchange: Social
Family as Institution
Fashion, Sociology of
Food, Sociology of
Generations, Sociology of
Groups, Sociology of
Historical Sociology
Human Ecology
Identity Movements
Ideology, Sociology of
Illiteracy, Sociology of
Income Inequality
Industrialization
Inequality Analysis: Overview
Inequality, Social
Inner City
Innovation, Theory of
Institutions
Integration, Social
Intellectuals, Sociology of
Interactionism, Symbolic
Interests in Sociological Analysis
Intersectionality
Knowledge, Sociology of
Labor, Division of
Legitimacy, Sociology of
Leisure, Sociology of
Literary and Philosophical Utopian Thought
Luhmann's Social Systems: Meaning, Autopoiesis and Interpenetration
Macrosociology-Microsociology
Marriage and the Dual-Career Family
Marxism in Contemporary Sociology
Medical Sociology
Methodological Individualism in Sociology
Migration and Development
Migration and Social Capital
Migration: Motivations
Minorities
Mobility, Social
Modernity
Modernization, Sociological Theories of
Money, Sociology of
Movements, Social
Nation, Sociological Aspects of
Nationalism, Sociology of
Neighborhood Effects
Neurosociology
Neutrality, Axiological
New Institutionalism in the Analysis of Complex Organizations
Norms
Organizational Ecology
Organizations, Sociology of
Phenomenology in Sociology
Police: A Sociology of Knowledge Approach
Political Sociology
Positivism, Sociological
Postmodernism in Sociology
Postsocialist Societies
Poverty, Sociology of
Power in Society
Prejudice in Society: Psychological Perspectives
Prejudice in Society: Sociological Perspectives
Prison life, Sociology of
Professions, Sociology of
Psychoanalysis in Sociology
Public Opinion: Social Attitudes
Race Identity
Racial Relations
Racism, Sociology of
Rational Choice Theory in Sociology
Rationality in Society
Relativism, Cognitive
Residential Segregation in the United States
Rumors and Urban Legends
Rural Sociology
Secular Religions
Segregation Indices
Segregation, Demographic Effects of
Shame and the Social Bond
Slavery as Social Institution
Social Changes: Models
Social Classification
Social Identity in Sociology
Social Influence, Sociology of
Social Justice: Historical and Theoretical Considerations
Social Networks
Social Psychology, Sociological
Social Stratification
Social Structure
Social System
Societies, Types of
Sociobiology and Sociology: A New Synthesis
Sociocybernetics
Sociological Theory
Sociological Traditions
Sociology and the New Institutionalism
Sociology, Epistemology of
Sociology: Overview
Solidarity, Sociology of
Spatial Mismatch
Sport, Sociology of
State, Sociology of the
Status and Role, Social Psychology of
Status and Role: Structural Aspects
Subculture, Sociology of
Suicide, Sociology of
Time Use Research: Recent Developments
Time Use Research: Subjective Time Use
Tradition: Social
Trust, Sociology of
Underclass
Urban Sociology

Urban Violence
Values, Sociology of
Visual Sociology
Voluntary Associations, Sociology of
Voting, Sociology of
Welfare
Work and Industry, Sociology of
Work, Sociology of
World Systems Theory
Youth Culture, Sociology of
Youth Movements

Statistics

Invariance in Statistics
Order Statistics

Bayesian Statistics
Bayesian Graphical Models and Networks
Bayesian Statistics
Bayesian Theory, History, Applications, and Contemporary Directions
Markov Decision Processes
Markov Processes in Learning Spaces
Monte Carlo Methods and Bayesian Computation: Importance Sampling
Monte Carlo Methods and Bayesian Computation: Overview

Computational Models
Agent Based Modeling, Statistics of
Computation: Expectation-Maximization Algorithm
Markov Chain Monte Carlo Methods
Markov Models and Social Analysis
Social Simulation: Computational Models

Data Analysis
Binary Response Models and Logistic Regression
Correspondence Analysis
Event Sequence Analysis: Methods
Exploratory Data Analysis: Univariate Methods
Factor Analysis and Latent Structure Analysis: Overview
Factor Analysis and Latent Structure: IRT and Rasch Models
Forecasting
Functional Data Analysis
Functional Magnetic Resonance Imagery, Analysis of
Growth Curve Analysis
Large-Scale Social Surveys, Analysis of
Logic Models
Log-Linear and Marginal Models
Longitudinal Data
Longitudinal Research: Panel Retention
Mathematical Models in Philosophy of Science
Mediation, Statistical
Meta-analysis: Theory
Mixed Methods
Model Selection and Model Averaging
Model Testing and Selection, Theory of
Moderator Variable: Methodology
Multivariate Analysis: Classification and Discrimination
Multivariate Analysis: Discrete Variables (Correspondence Models)
Multivariate Analysis: Discrete Variables (Loglinear Models)
Multivariate Analysis: Discrete Variables, Overview
Negotiations: Statistical Aspects
Panel Surveys: Uses and Applications
Quantitative Cross-national Research Methods
Regression Discontinuity Design
Replication
Sequential Statistical Methods
Statistical Analysis, Special Problems of: Outliers
Statistical Clustering
Structural Equation Modeling
Systematic Reviewing and Meta-Analysis

Data Analysis: Demography
Demographic Models
Demographic Techniques: Data Adjustment and Correction
Demographic Techniques: Family Reconstitution
Demographic Techniques: Indirect Estimation
Demographic Techniques: Inverse Projection
Demographic Techniques: LEXIS Diagram
Demographic Techniques: Small-area Estimates and Projections
Multistate Transition Models in Demography
Period and Cohort Analysis in Demography
Population Dynamics: Mathematic Models of Population, Development, and Natural Resources
Population Forecasts

Data Analysis: Econometrics
Econometric Software
Econometrics, History of

Data Analysis: Geography
Geographic Information Systems
Mathematical Models in Geography
Spatial Decision Support Systems
Remote Sensing

Data Analysis: Network
Clustering and Cohesion in Networks: Concepts and Measures
Models of Influence in Networks
Models of Power in Exchange Networks
Network Analysis
Network Centrality, Measures of
Network Models of Tasks
Network Paradigm: Applications in Organizational Science
Networks in Economics
Social Network Algorithms and Software
Social Network Analysis
Social Network Models: Statistical
Social Networks, Algebra of

Data Analysis: Political
Politics, Formal Modeling of
Pooled Cross-Sectional and Time Series Analyses, in Political Science

Data Analysis: Psychology
Mathematical Psychology
Mathematical Psychology: History
Psychometrics
Psychometrics: Classical Test Theory
Psychometrics: Combinatorial Data Analysis
Psychometrics: Item Response Theory
Psychometrics: Preference Models with Latent Variables

Data Collection and Measurement
Census Microdata
Censuses, Population: Comparative International Aspects
Censuses: Current Approaches and Methods
Censuses: History and Methods
Data Bases and Statistical Systems: Communications and Media
Data Bases and Statistical Systems: Crime
Data Bases and Statistical Systems: Crime and Delinquency
Data Bases and Statistical Systems: Crime Measurement
Data Bases and Statistical Systems: Culture
Data Bases and Statistical Systems: Demography
Data Bases and Statistical Systems: Economics (General)
Data Bases and Statistical Systems: Economics (Macroeconomics)
Data Bases and Statistical Systems: Economics (Statistical Systems)
Data Bases and Statistical Systems: Education, General
Data Bases and Statistical Systems: Environment and Urban Studies
Data Bases and Statistical Systems: Health
Data Bases and Statistical Systems: International Comparative Research
Data Bases and Statistical Systems: Labor and Work
Data Bases and Statistical Systems: Law
Data Bases and Statistical Systems: Linguistics
Data Bases and Statistical Systems: Overview
Data Bases and Statistical Systems: Political Behavior and Elections
Data Bases and Statistical Systems: Political Science (General)
Data Bases and Statistical Systems: Security and Conflict
Data Bases and Statistical Systems: Sociology
Data Collection: Interviewing
Demographic Measurement: General Issues and Measures of Fertility
Demographic Measurement: Nuptiality, Mortality, Migration, and Growth
Economic Growth: Measurement
Government Statistics
Indicator: Methodology
Inequality Measurement

International Social Survey Programme
Measurement Error Models
Nonsampling Errors
Observational Studies: Outcome-based Sampling
Observational Studies: Overview
Questionnaires: Cognitive Approaches
Questionnaires: Collection of Subjective Expectations Data
Random Assignment: Implementation in Complex Field Settings
Sample Surveys: Cognitive Aspects of Survey Design
Sample Surveys: Model-based Approaches
Scaling and Classification in Social Measurement
Spatial Data
Spatial Data Infrastructures
Spatial Sampling
Survey Research Centers and Companies
Surveys: Extreme Response
Surveys: Question Wording and Response Categories
Time Use Research Methods
Visual Methods in the Social Sciences

Data Structure and Design
Crossover Designs
Multidimensional Scaling I
Multidimensional Scaling II

Estimation
Compliance
Confidentiality and Statistical Disclosure Limitations
Errors in Statistical Assumptions, Effects of
Estimation: Point and Interval
Likelihood in Statistics
Likelihoods that Eliminate Nuisance Parameters
Multiple Comparisons, Statistics of
Population Size Estimation and Capture–Recapture Methods
Probabilistic Grammars and their Applications
Quantile Regression
Queues
Reliability and Validity Assessment: New Approaches
Reliability: Measurement
Resampling Methods of Estimation
Robustness
Screening and Selection
Statistical Identification and Estimability
Statistical Pattern Recognition
Statistical Sufficiency
Stochastic Models
Stochastic Optimization
Test Theory: Applied Probabilistic Measurement Structures

Experiment
Experimental Design: Bayesian Designs
Experimental Design: Compliance
Experimental Design: Large-scale Social Experimentation
Experimental Design: Overview
Experimental Design: Randomization and Social Experiments
Experimental Economics
Experimental Laboratories: Psychological
Experimental Laboratories: Social and Economic
Experimental Methods in Survey Research in Demography
Interrupted Time Series Quasi-Experimental Designs
Optimal Experimental Design
Quasi-Experimental Designs
Social Experiments

Graphical Methods
Graphical Markov Models
Graphical Methods, Analytic
Graphical Methods: Presentation

Hypothesis Testing
Dimensionality of tests: methodology
Hypothesis Testing in Statistics
Hypothesis Testing: Methodology and Limitations
Hypothesis Tests, Multiplicity of
Significance, Tests of
Statistical Distribution Models: Goodness of Fit Tests

Inference
Causal Counterfactuals in Social Science Research
Causal Inference: Overview
Ecological Fallacy, Statistics of
Ecological Inference
Endogeneity: Methodology
Fiducial and Structural Statistical Inference
Instrumental Variables in Statistics and Econometrics
Latent Structure and Causal Variables
Longitudinal Causal Inference
Mechanism-Based Causal Analysis
Mediating Variable
Propensity Scoring
Selection Bias, Statistics of
Simultaneous Equation Estimates (Exact and Approximate), Distribution of
Simultaneous Equation Estimation: Overview

Linear Models
Analysis of Variance and Generalized Linear Models
Generalized Linear Mixed Models
Linear Hypothesis
Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)
Linear Hypothesis: Regression (Basics)
Linear Hypothesis: Regression (Graphics)
Linear Mixed Models
Method of Moments and Generalized Method of Moments

Missing Data
Missing Data
Multiple Imputation

Mixture Models
Hierarchical Models: Random and Fixed Effects
Mixture Models in Statistics
Statistical Analysis: Multilevel Methods

Probabilities and Distributions
Distributions, Statistical: Approximations
Distributions, Statistical: Special and Continuous
Distributions, Statistical: Special and Discrete
Elicitation of Probabilities and Probability Distributions
Martingales
Probabilistic Thinking
Probability and Chance: Philosophical Aspects
Probability Density Estimation
Probability Theory: Formal
Probability Theory: Interpretations

Semi- and Nonparametric Methods
Nonparametric Regression
Nonparametric Statistics: Advanced Computational Methods
Nonparametric Statistics: Asymptotics
Nonparametric Statistics: Rank-based Methods
Nonparametric Statistics: Records
Semiparametric Models

Spatial Analysis
Event History Analysis
Event History Analysis in Continuous Time
Event History Analysis: Applications
Longitudinal Data: Event–History Analysis in Discrete Time
Spatial Analysis in Geography
Spatial Association, Measures of
Spatial Autocorrelation
Spatial Cognition
Spatial Equity
Spatial Interaction
Spatial Optimization Models
Spatial Pattern, Analysis of
Spatial Statistical Methods
Spatial–Temporal Modeling
Survival Analysis: Introduction
Survival Analysis: Overview
Survival Analysis: Whether, and If So, When?

Time Series
Time Series: Advanced Methods
Time Series: ARIMA Methods
Time Series: Cointegration
Time Series: Cycles
Time Series: Economic Forecasting
Time Series: General
Time Series: Seasonal Adjustment
Time Series: State Space Methods
Time Series: Unit-roots and Nonstandard Limiting Distributions

Studies of the Life Course

Adult Mortality in Developing Countries
Adult Mortality in Eastern Europe and the Former Soviet Republics
Adult Mortality in Industrialized Societies
Aging and Health in Old Age
Differential Aging
Early Life Influences on Health and Mortality in Adulthood
Event History Analysis in Continuous Time
Fertility Theory: Embodied-capital Theory of Life History Evolution
Fertility Theory: Theory of Intergenerational Wealth Flows
Mortality of the Oldest-Old
Mortality: The Great Historical Decline
Old Age: Definitions, Theory, and History of the Concept
Second Demographic Transition

Adulthood
Adulthood: Dependency and Autonomy
Fertility Theory
Fertility Theory: Theory of Life History Evolution
Midlife Psychological Development
Partnership Formation and Dissolution in Western Societies
Social Relationships in Adulthood

Age and Aging
Age Policy
Age Stratification
Age Structure
Age, Sociology of
Aging Mind: Facets and Levels of Analysis
Aging, Theories of
Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences
Cognitive Aging
Ecology of Aging
Population Aging: Economic and Social Consequences

Childhood and Adolescence
Childhood and Adolescence: Developmental Assets
Identity in Childhood and Adolescence
Infancy and Childhood: Emotional Development
Prenatal and Infant Development: Overview, Current Trends, Future Directions
Socialization in Infancy and Childhood
Transition to Adulthood

Life Course
Intergenerational Solidarity
Life Course and Event History Analysis
Life Course: Sociological Aspects
Life Table
Linked Lives

Lifespan
Human development, Successful: Psychological Conceptions
Human Development, Theories of
Interpersonal Trust across the Lifespan
Lifespan Development, Theory of
Lifespan Development: Evolutionary Perspectives

Older adulthood
Death and Dying, Psychology of
Old Age and Centenarians: The Human 'Warranty Period'
Retirement and Encore Adulthood: The New Later Life Course

Urban Studies and Planning

City Planning
Advocacy and Equity Planning
Citizen Participation
Critical Planning Theory
Eco-cities and Sustainable Urbanism
Environmental Planning
Gentrification
New Urbanism
Planning and Development
Planning during Post-Socialism
Planning Ethics
Planning Issues and Sustainable Development
Postmodern Urbanism
Property Rights and Urban Planning
Race: Ethnicity and Urban Planning
Real Estate Development
Regional Planning
Right to the City: A Liberal-Democratic Perspective
Smart Cities
Suburbanization and Suburbanism
Urban Places: Statistical Definitions
Urban Planning (Western), History of
Urban Planning in China
Urban Planning, Competitions and Exhibitions
Regional Planning, History of

Urban Studies
Assimilation of Immigrants
Cities and the Financial Crisis
Cities Under Postsocialism
Cities: Capital, Global, and World
Cities: Internal Structure
City Marketing
Community Studies: Anthropological
Economic Transformation: From Central Planning to Market Economy
Gated communities
Gender and Place
Ghetto
Growth Management and Urban Planning in the United States
Informal and Underground Economics
Inner City
Integration, Social
Just City
Landscape
Neighborhood
Planning Theory: The Good City and Its Governance
Postsocialist Societies
Poverty, Culture of
Residential Segregation in the United States
Resilient City
Rural Planning: General
Shrinking Cities
Social Capital
Strategic Planning for Public and Nonprofit Organizations
Underclass
Urban Design and Public Space
Urban Entreprenerualism
Urban Growth Models
Urban Informality: The Production and Regulation of Space
Urban Internal Spatial Structure
Urban Life: Future Developments
Urban Megaprojects
Urban Planning: Methods and Technologies
Urban Policies in Asia
Urban Policy in Europe
Urban Policy in North America
Urban Sprawl
Urban Studies: Overview
Urban Violence
Urban Violence and Militarization
Urbanization
Urbanization in Africa
Urbanization in China
Welfare Retrenchment

Urban Studies/City Planning
Housing and Cities
Neighborhood Revitalization and Community Development
Regional Planning: China
Transportation Planning
Urban Neoliberalism
Urban Poverty in Neighborhoods
Urban Regeneration

War, Peace, Violence and Conflict

Age and Crime
Aggression, Social Psychology of
Agonism
American Revolution, The
Arms Control
Aron, Raymond (1905–83)
Cannibalism
Child Soldiers
Civil Wars
Cold War, The
Collins, Randall (1941–)
Conflict and Conflict Resolution, Social Psychology of
Conflict and Consensus
Conflict and Socioemotional Development
Conflict and War, Archaeology of: Behavior and Social Organization
Conflict and War, Archaeology of: Weapons and Artifacts
Conflict and War: Anthropological Aspects
Conflict Mediation
Conflict: Anthropological Aspects
Conflict: Education and Youth

Conflict: Organizational
Coser, Lewis A. (1913–2003)
Crime and Class
Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices
Crime and Gender
Crime in the Rural Context
Crime Policy in the United States
Crime Trends and Debates
Crime, Desistance, and Persistence
Crime, Race and Ethnicity
Crime, Sociology of
Crime: Organized
Criminal Justice, Sociology of
Criminology, History of
Critical Criminology
Crowds in History
Cuban Revolution, The
Death, Anthropology of
Dehumanization
Deterrence
Elias, Norbert (1897–1990)
Ethnic Cleansing, History of
Ethnic Conflict, Geography of
Ethnic Conflicts
First World War, The
Genocide and War
Genocide, Historical Aspects of
Genocide: Social and Economic Aspects
Geopolitics
Guns and Criminals
Holocaust, The
Homicide
Honor Killing: Multidimensional and Multilevel Perspectives
Insurgency
Justice, Conflicts, and the Justice of Conflict Resolution
Law and Terrorism in the US and Beyond
Mann, Michael (1942–)
Mass Killings and Genocide, Psychology of
Mexican Revolution, The
Militaries and Militarization, Anthropology of
Militarism
Military Geography
Military History
Military Sociology
Military, War and Politics
Nation-State and War
Peace
Peace and Nonviolence: Anthropological Aspects
Peace Keeping
Peace Movements
Peace Processes
Peace Promotion, Psychology of
Peace-Making in History
Policing in Divided Societies
Punishment
Punishment: Comparative Politics of
Religion and Abuse
Revolution
Revolution, Theories of
Revolutions of 1989–90 in Eastern Central Europe
Revolutions, History of
Revolutions, Sociology of
Securitization
Simmel, Georg (1858–1918)
Slavery and Violence
Slavery as Social Institution
Slavery: Comparative Aspects
Slaves and Slavery, History of
Social Movements and Political Violence
Social Protest
Spencer, Herbert (1820–1903)
Sumner, William (1840–1910)
Territorial Conflicts
Terrorism
Urban Conflicts
Urban Violence
Vietnamese Revolution, The
Violence and Media
Violence in Anthropology
Violence, Criminalization, and the Homeless
Violence, History of
Violence: Public
War and Democracy
War and Nationalism
War and Terrorism, Geography of
War Crimes Tribunals, Crimes against Humanity and Transitional Justice
War Propaganda
War, Political Violence, and Effects on Children
War, Social Causes and Consequences
War, Sociology of
War: Anthropological Aspects, Historical Development of
Warfare in History
Wars among Nation-States: Patterns and Causes
Weber, Max (1864–1920)
Youth Gangs

This page intentionally left blank

SUBJECT INDEX

Notes

Cross-reference terms in *italics* are general cross-references, or refer to subentry terms within the main entry (the main entry is not repeated to save space). Readers are also advised to refer to the end of each article for additional cross-references - not all of these cross-references have been included in the index cross-references.

The index is arranged in set-out style with a maximum of three levels of subheading. Major discussion of a subject is indicated by bold page numbers. Page numbers suffixed by t and f refer to Tables and Figures respectively. *vs.* indicates a comparison.

The index entries are presented in word-by-word alphabetical sequence in which a group of letters followed by a space is filed before the same group of letters followed by a letter. For example, entries beginning 'air density' are alphabetized before 'aircraft.' Prefixes and terms in parentheses are excluded from the initial alphabetization.

Where index subentries and sub-subentries pertaining to a subject have the same page number, they have been listed to indicate the comprehensiveness of the text.

A

A* algorithm 2:32
A posteriori physicalism 18:108–109, 18:112
A priori class probabilities 23:411–413
A priori knowledge 7:569
A priori physicalism 18:107–109, 18:112
A Toroidal LHC Apparatus (ATLAS) collaborations at CERN 18:104
AA *see* Alcoholics Anonymous (AA)
AAA *see* American Anthropological Association (AAA)
AAAP *see* American Association of Applied Psychology (AAAP)
AAAS *see* American Association for the Advancement of Science (AAAS)
AACP *see* American Association of Community Psychiatrists (AACP)
AAG *see* Association of American Geographers (AAG)
AAI *see* Adult Attachment Interview (AAI)
Aalen additive hazards model 23:774
Aalen–Johansen estimator 8:315
AALS *see* Association of American Law Schools (AALS)
AAM *see* Active Appearance Model (AAM)
Aanvullend Voorzieningengebruik Onderzoek (Amenities and Services Utilization Survey) 5:447
AAPOR *see* American Association for Public Opinion Research (AAPOR)
AAR *see* Absence of Attachment Representations (AAR)
AARP Experience Corps 1:150
AAU *see* Addis Ababa University (AAU)
AAUP *see* American Association of University Professors (AAUP)
AAUW *see* American Association of University Women (AAUW)
AAVE *see* African American Vernacular English (AAVE)
ABA *see* American Bar Association (ABA); Applied behavior analysis (ABA)
ABAB design, for workplace safety and health 25:731
ABAD *see* Apni Beti Apta Dhan (ABAD) program
Abatement effect 23:825
AB/BA design *see* Two-period two-treatment design
Abbāsid Empire 12:747
Abbasids (caliphal dynasty) 11:63–64
Abbott, Jennifer, works, *The Corporation* 6:620–621
ABC *see* Active, Beautiful and Clean (ABC) Waters program; Activity-based costing (ABC); Approximate Bayesian computation (ABC); Attachment and Biobehavioral Catch-up (ABC)
AB5C *see* Abridged Big Five Dimensional Circumplex (AB5C)
ABC: The Alphabetization of the Popular Mind 11:586–587
A-B-C (action, belief, and consequence) model 19:375–376
A-B-C (attitude, behavior, conditions) model 7:703
Abdel-Malek, Anouar, works, *Contemporary Arab Political Thought* 16:384
Abducens nerve 8:628–629
Abduction 19:61, 19:64
  inference 18:809
  and legal reasoning 13:776–777
  in semiotics 21:595
Abduh, Muhammad 12:730
Abdul-Kadir, Abdullah bin, works, *Kisah Pelayaran Abdullah ke Kelantan* 11:84
Abelard, Peter 3:774–775, 19:408
Abelmann, Nancy 2:73
  on Korean villages 2:72–73
ABF *see* American Bar Foundation (ABF)
Abhidhamma Pitaka 2:888
ABI *see* Auditory brainstem implants (ABI)
Ability 7:276
  beliefs 8:504–505
    and academic motivation 19:790–792
    racial stereotypes 19:790–791
  and emotion regulation 7:454
  racial stereotypes about 19:790–791
  scales 17:753–754
    general scales 17:754
    specific area scales 17:753–754
  versus skill 22:372
  social comparison 22:210–214
Abject dependency 22:485–486
Ableism 6:453, 6:456
  and disablism 6:448
  ideology of 6:476

ABM *see* Antiballistic Missile (ABM) Treaty
ABMs *see* Agent-based models (ABMs)
Abnormal 5:474
Abnormal children, study of 2:597
Abnormal grief 2:793
Abnormal illness behaviors 11:599
  hysteria 11:509
Abnormal involuntary movement scale 19:463
Abolafia, Mitchell, on corporate cultures 5:536
Abolitionism, peaceful *versus* violent 22:40
Aboriginal art
  in Australia 11:773
  copyright 11:773–774
Aboriginals 11:762–763, 22:220, 24:509
  Australia *see* Australian Aborigines
  in Canada *see* Canadian Aboriginals
  groups 1:251–252
  identity movements 11:544–545
  populations 5:151
  see also Indigenous peoples
Aborted commemoration 4:239
Abortion-dividend thesis 5:200
Abortions
  in affluent nations 20:469–470
  Boltanski on 2:750
  ad culture 9:2–4
  and death 5:872
  demography, in United States 1:1–5
    changes 1:4–5
    data sources 1:1–2
    disparities 1:5
    incidence and abortion rates by subgroup 1:2–4, 1:2t–3t
    lifetime incidence 1:4
    unintended pregnancy 1:1, 1:3–5
  ethics 2:613, 2:615
  in formerly socialist European countries 9:99–101, 9:103, 9:103f
  issue, evolution of 12:767
  in Japan 12:796–797
  in Latin America and Caribbean 9:71
  legalized 5:200
  and reproduction 20:450–453
  rights 20:457
  sex selective 15:608–609, 15:611–612, 21:686–688
  in South Asia 9:81–82
Abraham, Karl 9:416
Abrams, Lynn 17:285
Abridged Big Five Dimensional Circumplex (AB5C) 4:659–660, 9:232, 9:237
Abromowitz, Mimi, works, *Taxes are a Women's Issue* 8:902
ABS *see* Asian Barometer Survey (ABS)
Absence of asymmetry 2:806
Absence of Attachment Representations (AAR) 12:231
Absenteeism 1:423, 14:423
  job satisfaction 12:835
  and obesity 17:70
Absolute centralization index (ACE) 21:412
Absolute chronology 7:695–696
Absolute cognition 23:595
Absolute constructions 11:827
Absolute dating (archaeology) 23:744
Absolute deterrence 6:255–256, 6:262
Absolute fit indices 8:688
Absolute frequencies, survey questions 23:767, 23:769
Absolute Idea 10:737
Absolute income poverty measures 18:742
Absolute judgments, in forced-choice personality assessment 17:841, 17:843, 17:848
Absolute knowledge 10:737
Absolute location 14:284–286
  geographical referencing 14:285
  spatial relations 14:285–286
Absolute monarchy 1:6
Absolute moral prohibitions 16:322
Absolute negative rights 16:322
Absolute neutralization 18:77
Absolute personal risk perceptions 10:704
Absolute poverty 18:756–757
Absolute power 1:7–9
  configuration of 1:9–10
Absolute rates 15:642, 15:644
Absolute Return for Kids (ARK) 1:270
Absolute rights 16:322
Absolute spatial terms (language) 13:295
Absoluteness 1:9
Absolutism 20:227
  conservative perspective 1:6
  early modern 1:6–7, 1:10
  and historiography 1:6–7
  and liberalism 14:18
  Marxist perspective 1:6
  and morality 15:807
  in practical dynamics 1:6
  phenomenal 8:167
  practice and theory 1:7–9
  revisionist perspective 1:6
  and state 23:364
  terminology 1:9–10
  theoretical level 1:6
  typology 1:9–10
Absolutist ethical objectivism 6:592
Absolutization, of sovereignty 22:848
Absorbing demographic events 5:740
Absorption
  school engagement 21:59–60
Absorptive capacity 1:12–19
  articles in social science field 1:13f
  bibliometric analysis 1:12
  competitive advantage 1:12–13, 1:15, 1:17
  conceptualization of 1:14–16, 1:16t
  and dynamic capabilities 1:15–17, 1:17f
    integration into conceptualization 1:17
  implications and outlook 1:17–18
  innovation management 1:13
  intellectual core of 1:12, 1:14
  knowledge acquisition 1:13–16, 1:18
  knowledge assimilation 1:14–16, 1:18
  knowledge exploitation 1:14–16, 1:18
  knowledge management 1:12–13
  knowledge transformation 1:14–16, 1:18
  and learning process 1:15, 1:17f
  operational capabilities 1:15–16
  organizational learning 1:12–13, 1:15–16
  research themes 1:14f
Abstention rates, alcohol consumption 1:517
Abstinence violation effect 20:218
Abstinence-only sex education 21:826
Abstinence-plus sex education 21:826
Abstract concepts and words (ACWs) 7:423–424
Abstract constructs, in generalization 9:840, 9:844
Abstract mappings 16:507–508
Abstract realism 5:89
Abstract rhythmic structures 13:275
Abstract rules
  imitation of 11:623f
  spelling as 23:271–272
Abstract self-evaluations, self-development 21:492
Abstract thought 21:268
Abstract work, invention of 25:676
Abstract-deductive theory 16:254
Abstracted empiricism 15:537
Abstractions (finance) 9:172–175
Abstractions (mental process) 11:516, 16:506–507
  and intelligence 12:316
  reasoning 5:933–935
    formal languages 5:933–934
    inferential relationship 5:934
    syntactic inference 5:934–935
  and reification 20:203–205
Abstractions (writing), handling data in 25:765
Abu Ghraib prison 13:901
Abu-Hanifah 12:735
Abu-Lughod, Janet, works, *Before European Hegemony* 16:382
Abu-Lughod, Lila, works, *Writing Women's Worlds* 8:901
Abuse
  and addiction 1:97, 1:100
  alcohol *see* Alcohol abuse
  in asylums 6:47–48
  child *see* Child abuse
  and childhood stress 23:570, 23:572–574
  clergy malfeasance 20:256
  clergy sexual misconduct 20:256–257
  community-based agencies, challenge for 20:256
  of discretion 1:824
  domestic violence and communities of faith 20:254
  dynamics of 20:253–254
  of elder *see* Elder abuse
  managing 20:688–690
  prevalence and severity of 20:253
  and religion 20:253–257
  and religious battered women 20:254–255
  and religious leaders 20:254, 20:256–257
    challenge for 20:255–256
  and religious survivors 20:253, 20:255
  responding to 20:255
  spiritual 20:253

Abusers
in intimate partner abuse 12:678–679
programs for 6:635
Abusive practices, in psychiatric care 6:42–43
Abusive supervision 23:701–708
conceptual facets of 23:701
consequences of 23:701–702
attitudes 23:701–702
family well-being 23:702
victim personal 23:702
work behaviors 23:701–702
mechanisms linking 23:702–703
emotions 23:703
relational theories 23:702–703
self-regulation 23:703
methodological issues in 23:704–705
cross-sectional research designs 23:705
inadequate model specification 23:705
self-reported data 23:705
similar constructs 23:704–705
predictors of 23:703–704
subordinate characteristics 23:704
supervisor characteristics 23:703–704
research directions in 23:705–706
cross-cultural research 23:706
multilevel research 23:705–706
prevention and management 23:706
third-party observers 23:705
AC *see* Adjunct Condition (AC); Alternative certification (AC) programs
ACA *see* Affordable Care Act (ACA)
Academia Sinica, Taiwan 22:639–640
Academic achievement
and academic climate 1:45
and academic self-concept 1:54–63
and class size 3:751–754
coeducation schooling 3:928
and computer games 4:498
and country income 1:43–44
cross-cultural studies 5:327–331
effect of future orientation on 9:564
and ethnicity 1:44
expectancy-value model 1:37–40
expectancy-value-cost model 8:503–504, 8:506–508
funding 1:46
and gender 9:677–686
I/E model 12:425–426, 12:429
and immigrant status 1:44
intervention 1:46–47
and openness to experience 17:232
role of parents in 1:40–41
role of teachers in 1:44–45
and self-concepts 21:471–472
effects 21:471–472
empirical evidence 21:471
significance of temperament in 24:185–186
single-sex schooling 3:928
and socioeconomic segregation 1:45–46
and socioeconomic status 22:924
test anxiety and 24:244–249
utility value interventions 1:38–40
Academic achievement motivation 1:20–25, 21:50, 21:52
achievement goals 1:21f
competence-related/efficacy beliefs 1:21–22
change in mean level of 1:21–22
children's understanding of 1:21
development of children's goals 1:23
development of motivational problems 1:23–24
alleviating learned helplessness 1:24
anxiety intervention programs 1:24
learned helplessness 1:24
test anxiety 1:23–24
interest theories 1:22
developmental changes 1:22
intrinsic motivation theories 1:23
developmental changes 1:23
subjective task values 1:22
development of 1:22
success 1:20
Academic anthropology 1:772
Academic associations 22:661
changing landscapes for 22:662–664
digital revolution 22:662–663
federal funding 22:664
foundation funding 22:663–664
public funding 22:663–664
and communications research 4:278, 4:281
future in social sciences 22:664–666
growth 22:661–662
opportunities 22:661–662
shifting context for 22:662
Academic buoyancy, and motivation in Australian Aboriginals 15:902–903
Academic careers 1:49–50
in comparative perspective 1:26–30
academic training 1:26–27
employment contracts 1:27–28
hierarchies 1:28
policy 1:29–30
reforms 1:29–30
tenure 1:28–29
Academic climate
and academic achievement 1:45
and SES 1:45
Academic colonialism 11:898
Academic concepts, Vygotsky on 25:313–314
Academic criminology 1:832
Academic development
evolutionary perspective 8:394–395
brain mechanisms 8:395
cognitive mechanisms 8:395
implications for instructional activities 8:396–397
motivation to learn 8:395–396
Academic disciplines 6:491, 22:589, 22:989
of anthropology and state 23:349–350
and area and international studies 1:943–944
art 2:18–19
in colonialism 4:218–221
degree of coherence of 23:227–228
of economics 7:93–96
and ethics 8:58
folklore 9:280–281
fragmentation and specialization 23:225
human rights and 11:376–377
human–computer interaction as 11:403–404
hybridization 23:226–228
institutionalization of division into subfields 23:226
military sociology as 15:515–516
peer review 17:685–688
recombination of specialties across 23:226–228
reference works by 7:588–592
anthropology 7:589–590
cognitive sciences and related fields 7:590
economics 7:590–591
political science 7:591
psychology 7:591–592
sociology 7:588–589
related to anthropology of education 1:730–731
restricted disciplines 23:226
schools of 23:226
scientific disciplines 21:287–290
versus specialties 23:226
synthesis or recombination 23:228
Academic disengagement 19:791
and adolescent health 1:124
Academic disidentification 19:791
Academic distinctions, and public recognition 2:333
Academic emotions 6:434
Academic engagement 1:31–36
Academic fields, contribution to dramaturgical analysis 6:647
Academic freedom 1:28–29
Academic functioning 16:725
Academic involvement, and adolescent health 1:124
Academic learning, and social work education 22:777–778
Academic libraries 14:45, 14:49
Academic literature, perspectives on financial regulation in 9:193–194
Academic material 24:56
Academic motivation
in Australian Aboriginal students 15:901–903
and gender 9:677–681
cross-national trends 9:679
increased focus on male students in US studies 9:677–678
large-scale assessments 9:677
theoretical frameworks 9:679–680
variations within gender groups in US studies 9:678
and race 19:790–795
achievement values and immigrant paradox 19:794
African-Americans and stereotype threat 19:791
Asian-Americans and model minority stereotype 19:791
beliefs about ability 19:790–792
beliefs about effort 19:792–794

Academic motivation (*continued*)
characteristics of valued peers 19:793–794
coping with discrimination 19:792
entity *versus* incremental theories about intelligence 19:791–792
ethnicity 19:790–795
expectancy 19:790–795
motivational intervention 19:794–795
oppositional identity 19:793
racial stereotypes about ability 19:790–791
self-affirmation 19:794
task value interventions for 1:37–42
Academic Network of European Experts in Academia 6:465
Academic outcomes
and bullying of LGBTQ youth 21:87
media use and 15:36–37
Academic performance
with agreeableness 17:788
with conscientiousness 17:788
effects of socio-economic status on 1:43–48
and emotional stability 17:789
and extraversion 17:789
Academic press *see* University press
Academic problems, and computer games 4:499–500
Academic professionalism 22:593
Academic programs, in women's studies 9:738–739
Academic psychology 1:853
Academic publishing, and peer review 17:686–687
Academic research 10:867
and employment 1:49–53
academic careers 1:49–50
academic work 1:49–51
academic work contracts 1:49
externally sponsored research 1:49
external research funding 1:49–50
funding in higher education 1:51
network governance 1:51
New Public Management (NPM) 1:49
postdoctoral experience 1:50
postdoctoral phase 1:50
postdoctoral scholars 1:50
project-based, external funding in academe 1:50–51
project work 1:51
research funding 1:51–52
research projects 1:49–52
sponsored projects 1:51
tenure 1:50–51
third-party funding 1:49
growth of 20:497
surveys 23:748–749, 23:751
system 22:669–670
Academic science and technology 24:758–759
knowledge transfer 24:760
management capacity 24:760–761
market-based funding 24:760
problem orientation 24:760
Academic self-concept (ASC) 21:460–465
and academic achievement 1:54–63
big-fish-little-pond effect 1:59–61
causal ordering 1:56–57
educational implications 1:56–57, 1:59–61
empirical evidence 1:58–59
frame of reference 1:57–59
global self-esteem *versus* multidimensional self-concept 1:54–55
hierarchical multidimensional structure 1:54–56
importance of multidimensionality and positive psychology 1:57
internal/external model 1:57–59
low correlations among self-concept domains 1:55–56
mechanism 1:60
motivation 1:60
other outcomes and moderators 1:59–60
reflected glory 1:60
self-enhancement and skill development models 1:56–57
simultaneous use of external and internal frames 1:58
social comparison 1:59–61
integration of I/E model with other models 12:429–430
and school comparison 22:214
Academic skepticism 6:589
Academic Spring 21:248
Academic stereotypes 8:734
Academic stratification system 9:619–620
Academic Taiwan 4:278
Academic teaching of history 11:121
Academic training 1:26–27
Academic tribes 21:132
Academic work 1:49–51
contracts 1:49
Academics, role in higher education management 10:845–848
Academie des Sciences 21:177–178
Académie des sciences morales et politiques 22:990–991
Académie française 5:460–461, 7:165, 19:611
Academies 22:661–662
Academization 17:440
of social work 22:799–800
Academy of Criminal Justice Sciences (ACJS) 8:916
Academy of Science of South Africa (ASSAf) 22:608
Academy of the Social Sciences in Australia (ASSA) 22:638–639
Acalculia 17:509–510
ACASI *see* Audio-Computer Assisted Self-Interviewing (ACASI)
Acataphasia 16:683
ACC *see* Adaptive cruise control (ACC); Anterior cingulate cortex (ACC); Asia Cohort Consortium (ACC)
ACCCRN *see* Asian Cities Climate Change Resilience Network (ACCCRN)
Accelerated failure time models 23:774
Accelerated long-term forgetting (ALF) 24:554, 24:557–558
biology of 24:558
clinical and neuropsychological features 24:557–558, 24:558f
cognitive architecture 24:558
Acceleration
in gifted education 7:160
human growth data 9:514
and time use 24:373, 24:376–378
Accelerator mass spectrometry (AMS) 1:905
radiocarbon dating 3:566, 9:304
Accentuation, self-categorization theory 21:456, 21:458
Acceptable gay public spheres 21:809
Acceptable loss function 2:551
Acceptance
community-based health interventions 10:647
in social work practice 10:530, 10:533
Acceptance and commitment therapy (ACT) 2:420, 2:432, 4:544
for childhood cancer 3:462
for depression 19:375–376
for generalized anxiety disorder 19:374
Acceptance region, in hypothesis testing 11:491
Access
consciousness 4:675
to culture
and cultural policy 5:461–463
to education 7:918–922
to employment 25:698
to health care 10:596–597
to journals 12:866–867
to public facilities and resources 23:149–150
Access to justice 12:919–927, 13:601
conditional fees 12:922
contingent fees 12:922
court-awarded fees 12:922
future issues 12:926
government-paid lawyers 12:920–921
hotlines 12:925
legal aid 12:919–923, 12:925–926
poverty law 18:737, 18:739
private attorneys for lower income people without government subsidy 12:921–923
right to counsel 12:920–922
self-help assistance 12:925–926
strategies 12:919–920
voluntary and mandatory pro bono legal services 12:922–923
without supplying free/government paid lawyers 12:923–926
comprehensive fully integrated systems 12:925–926
less expensive, nonlawyer representatives 12:924–925
out-of-court assistance to unrepresented litigants 12:925
redistribution of functions and responsibilities 12:923–924

simplification of procedures 12:923–924
simplification of substantive legal rules governing disputes 12:924
technological substitutes for legal expertise 12:925
Access-and-legitimacy perspective, of organizational diversity 6:597
Access-based theories 18:949
Accessibility 4:871
and disability 6:467, 6:469
of early childcare 3:397
in fatherhood 8:841
food 9:285–288
international public goods 19:540–541, 19:543
measures 24:825–826
and social categorization 22:187–188
and spatial interaction 23:155–156
and transportation 24:616–617, 24:619–620
Accessorial liability 17:144
Accessory olfactory bulb (AOB) 17:198
Accessory olfactory system (AOS) 17:198
Accessus ad auctores 10:780
Accessus ad auctores 10:780
Accident causation theories 12:130–131
Accidental communities 4:370
Accidental diaspora 8:160
Accidental regularity 3:266–267
Accidents
and adolescent health 1:121, 1:126
and injuries
behavioral modifications for safety 12:133
causes 12:131–132
demographic factors 12:131–132
environmental modifications for safety 12:133
epidemiological approach 12:130
health and safety education 12:133
individual differences 12:131
life stress and work-related stress 12:132
prevention 12:132–133
psychological and sociological approach 12:130–131
risk-taking behavior 12:132
laws 13:448–449
evaluation under economic approach 13:448–449
incentives toward safety 13:448
insurance and compensation 13:448
litigation-related costs 13:448
occupational 25:727–729
causes 25:727–728
risk taking 25:728–729
proneness 12:131, 25:727
unintended consequences 24:745, 24:748
Accominotti, Fabien
on assortative matching 2:55
on movement networks 5:469
Accommodating correlation, in generalized linear mixed models 9:845–847
Accommodation 2:82–83, 14:87–88, 18:126, 18:784, 22:564, 23:230–231
cognitive development 4:52–53
in error processes 23:230–231
and life span development 14:103–104
of outliers 23:389–390
rejection of neonatal imitation 16:484
Accommodative coping, of chronic illness 3:550–551
Accomplishments 7:276
life phase 5:508–509
Account system, advertising agencies 1:192–193
Accountability 3:921, 6:65, 7:579, 10:295, 13:721–722, 16:770, 17:201, 19:210, 21:101–102, 23:739–740, 24:675–676
and administrative law 1:109–110
and deliberative democracy 6:65
downward 16:770–771
electoral 18:463
and empowerment evaluation 7:579
and governance 10:295
of government, and social work 13:721–722
and independence to public 19:212
managed care 14:464
movement, and education 7:294–295
and new managerialism 16:770
of nonprofit organizations 16:917
and oligarchy 17:201
peer review 17:685–686, 17:688
and political representation 18:442–443
in prosecution 19:210
psychological treatments 19:382
responsibility as 20:589–590
for results 16:770–771
in school management 21:101–102
single-case designs 21:988
and surveillance studies 23:739–740
of teachers in United States 24:75, 24:78
and trust 24:668–671
of universities 18:307, 18:309–310
upward 16:770–771
Accountable Care Organizations 10:616
Accountable government 23:370
Accounting
equations 23:468–469
financial 4:929
management 14:466–470
standards, in material flow analysis 14:761–762
Accounting device, writing as 25:762
Accounting identity 9:319–320
Acculturation 1:68–74, 2:84–85, 5:154, 5:314, 8:905–911, 9:319–320, 15:408, 15:472, 15:474–477, 17:533
adaptation and 1:73
affective changes 1:70
behavioral changes 1:70
change and 1:69
cognitive changes 1:70
continuous first-hand contact 1:68
definition 1:68
developmental issues 1:70–71
and diffusion 6:403–404
and ethnicity 8:159
frameworks and theoretical perspectives 1:69–70
group *versus* individual-level phenomenon 1:68–69
and immigration 11:676, 11:679–680
minority 8:145
model 21:363
mutuality of influence 1:68
outcomes 1:72–73
and parenting 17:496–497
personality and individual factors 1:71
processes 1:71–72
psychology, ethnic identity 8:113, 8:115
resource centers 20:295–296
strategies in ethnocultural groups and larger society 1:71f
and stress 1:70
and syncretism 23:882
use of term 1:68
Accumulation
capitalist 3:105–106, 3:108
capitalist and class struggle 6:242
models
correlated exposures or chain of risk model 14:69–71, 14:70f
independent exposures and direct effects 14:69–70, 14:70f
models in practice 14:71–72
single exposure, duration/dosage of 14:69–70, 14:70f
urban, in China 24:974–976
Accumulative prediction error (APE) 4:475
Accumulator models, comparison with diffusion models 6:400
Accuracy 10:302, 14:285
diffusion models 6:399
econometric software for 7:5
of flashbulb memory 9:255–256
of person perception 17:739–741
degree 17:740
difficulties in research 17:739
errors 17:739–740
good judge 17:740
good target 17:740
good trait 17:740–741
heuristics 17:740
importance 17:739
information 17:741
moderators 17:740–741
PERSON model 17:741
realistic accuracy model 17:741
realistic conceptualization 17:740
resolving difficulties in research 17:740
social judgment 17:740
social relations model 17:740
theoretical development 17:741
stereotypes 23:448–449
Accuracy assessment of census data 3:293–294
Accuracy Standards 19:119
Accuracy–effort trade-off, in heuristics 10:831–832
Accursed death 5:866

Accusative alignment 10:328
Accusative cases in positive environments 16:413
ACD *see* Armed Conflict Dataset (ACD)
ACE *see* Absolute centralization index (ACE); Average causal effect (ACE)
ACE model in twin studies, estimates and variance decomposition 24:704–705
Acetaldehyde enzyme alleles 17:238–239
Acetylation 7:849–850
  histone 7:840–841
Acetylcholine (ACh) 9:1012, 16:749–752
  role in neural plasticity 16:621–622
  and sleep 22:89
Acetylcholinesterase 16:62
ACh *see* Acetylcholine (ACh)
Achbar, Mark, works, *The Corporation* 6:620–621
Achebe, Chinua, works, *Arrow of God* 1:300
Achenwall, Gottfried 23:423
Achieved identity, effect of future orientation on 9:565
Achievement 6:431–432, 7:272–276
  academic *see* Academic achievement
  assessment of 24:245
  during competition/cooperation 4:857–859
  equality of opportunity 7:904
  expectancies 21:471–472
  and fatherhood 8:842
  and feedback 21:473
  goals 24:247
  and grit 10:397, 10:399–400
  identity and 21:433–437
  and impact of test anxiety on 24:245–246
  and just world belief 2:490
  motivation 8:395
    and economic development 5:528–529, 5:531, 5:533
    and personality 17:831–832
    relation of epistemological beliefs to 7:889–891
  and motives 15:952–955
  and self-concept 21:462–463
  tests 7:272–273
  values, and epistemological beliefs 7:890–891
Achievement Emotions Questionnaire (AEQ) 24:245
Achievement gap 21:43–45
Achievement goal theory (AGT) 15:929–931, 24:247
  goals and temperament 24:187–188
  implications for learning and instruction 15:930–931
  teacher motivation 24:67–68
Achilles' heel phenomenon 15:206
Achondroplasia 9:1019
ACHSSW *see* Australian Council of Heads of Schools of Social Work (ACHSSW)
Acid rain 7:794
ACJS *see* Academy of Criminal Justice Sciences (ACJS)
Ackelsberg, Martha A., on anarchism 19:748
Ackerman, Bruce 4:705
Ackerman, Edward 22:571
ACLEC *see* Advisory Committee on Legal Education and Conduct (ACLEC) Report
ACLED *see* Armed Conflict Location and Event Dataset (ACLED)
ACLS *see* American Council for Learned Societies (ACLS)
ACM *see* Anticult movement (ACM)
ACOLA *see* Australian Council of Learned Academies (ACOLA)
Acolytes 15:576, 15:579–580
Acoustic signaling 16:475
Acoustic-phonetics
  decoding of speech sounds 23:246
  features 23:243
  response selectivity 23:244–245
  and speech processing 23:246
Acoustics 23:243
Acquaintance rape 19:888–891
Acquired brain injury, neuropsychological rehabilitation for 2:817
Acquired characteristics, inheritance of 8:383–385
Acquired dysgraphia 1:450–455
Acquired dyslexias 6:730, 6:732–738
Acquired economic status 11:896
  groups with 11:896
Acquired immune deficiency syndrome (AIDS) 1:488–494, 18:174, 21:793
  in Africa 1:284
  antiretroviral therapy (ART) 1:491–493
  CCR5 for 1:489–492, 1:489f
  CD4 for 1:488–492, 1:489f
  and central nervous system 1:488, 1:492
  clinical pathology 1:488–490
  demographic microsimulation 15:345
  etiology 1:488–490
  and gay/lesbian movements 9:634
  and heterosexual sexuality 10:805–806
  HIV associated neurocognitive disorders (HAND) 1:491–492
  and homosexuality, in China 21:736–737
  illness progression 1:491
    acute infection or seroconversion 1:491
    clinical and immunological 1:490–491
    viral latency 1:491
  immunology 1:488–490
  and international social work student exchanges 12:566
  in Latinos 9:630
    influence on gay/lesbian activism 13:861
  and lesbians 15:111
  neuropsychiatric and psychosocial issues 1:492–493
  and older LGBT relationships 13:864–865
  policy and ethical issues 1:493–494
  priority dispute 21:281
  and prospective memory 19:275–276
  and reproductive/sexual rights 20:459
  and same-sex sexualities 20:870, 20:872–873
  and sexuality 21:822–823
  testing issues and infection stages 1:490
  and transgender communities 24:552
  treatment of HIV infections 1:491–492
Acquired pragmatic disorders 18:791–794
Acquired prosopagnosia (AP) 19:250–251
Acquired stuttering (AS) 9:272, 9:272t
  stutterers 9:266
Acquired variations, epigenetic inheritance 7:834, 7:837
Acquis communautaire 18:693–694
Acquisition editors 19:626–627
Acquisitions
  consumer psychology 4:738
  fear conditioning 1:811–812
  of food preferences 9:297–298
  planning, in language policy 13:354
  of sign language 21:929
  in spatial cognition 23:112
Acquitted of a crime
  and guilt 10:455–456
  jury 10:455–456
  meaning 10:455–456
ACR *see* Association for Consumer Research (ACR)
Acrolect 18:133–134, 18:136–137
ACS *see* Acute coronary syndrome (ACS); American Community Survey (ACS)
ACSI *see* American Customer Satisfaction Index (ACSI) model
ACT *see* Acceptance and commitment therapy (ACT); Affect control theory (ACT); Assertive community treatment (ACT); Attentional control theory (ACT); Automatic Control of Thought (ACT)
ACT (authoritarianism, conservatism, and traditionalism) model 2:263
Act of Parliament 12:897
Act of Prohibition of Private Monopolization 3:31
Act on the Modernization of the Law of Obligations 3:671
ACT UP *see* AIDS Coalition to Unleash Power (ACT UP)
Acta Museorum Agriculturae 17:671
Acta Sanctorum 11:72–73
Actable idea 6:646–647
Actants
  Latour on 13:415–417
ACTFL *see* American Council on the Teaching of Foreign Languages (ACTFL)
ACTH *see* Adrenocorticotropic hormone (ACTH); Adrenocorticotropic releasing hormone (ACTH)
Acting white (perception) 19:793
Actio 20:651
Action citizen science projects 3:632
Action disorganization syndrome 1:877–878
Action Erntefest (harvest festival) 16:260
Action regulation theory 12:813
Action representation, like me theory 15:132
Action research (AR)
  community members 22:270
  micromethods 12:651

and nongovernmental organizations 16:859–860
and social justice 22:269–273
and social research 22:269, 22:272–273
for social work 22:269–274
antioppressive research 22:272–273
community-based action research 22:271–272
participatory action research 22:270–271
Action sounds, representation of 23:246
Action systems 1:381–382
analysis, AGIL for 1:390
human condition 1:390
Action tendencies, emotional expression 7:484
Action theory 14:87, 17:315
models 3:445
in sociology 1:75–76
Actionalism 24:459
Action-awareness, and self 21:419–421
Action-phase model
of developmental regulation 4:815–816, 4:815f
of life span development 14:104, 14:106
Action(s)
action knowledge 13:93
archaeology 8:162
and cognition, in pragmatism 18:809–810
and cognitive development 2:882
control 1:874, 10:713
and motivation 15:937–938
and discourse 14:203
and embodied cognition 7:420–422
and free will 9:404
goal-directed 10:268, 10:271–272
and habit 10:475–478
reflective model 10:478
identification in conversation analysis 4:850–854
integration of 12:243–244
joint actions 12:839–843
knowledge-based 13:108–110
and legal reasoning 13:776–780, 13:782
Luhmann on 14:384–385, 14:394
methodological individualism 15:308–309, 15:313
neurosociology of 16:739–740
-oriented cognition building 1:106–107
and performance 17:708–711
planning, health behavior 10:578–579
potentials 16:703–705
psychology of 15:887–893
attributions 15:887–890
behavior change interventions 15:892
concept of goals 15:888
current goal research 15:892
dealing with hindrances 15:891–892
determinants of effective goal pursuit 15:888–889
dwelling 15:889–890
fantasies 15:889–891
fantasy realization theory 15:889
implementation intentions 15:891–892
impulsivity 15:892
indulging 15:889–890
instigating determined goal pursuits 15:889–891
intention–behavior relation 15:891
mental contrasting 15:889–890, 15:892
self-completion theory 15:889
self-efficacy 15:887–888
self-regulation of goal pursuit 15:889–892
strategic automaticity 15:891–892
rationality of 12:337, 15:310–311
readiness 5:544–545
and facial expressions 5:545
reasons for 18:776–779
regulation, and coping 4:891–892
relative deprivation 6:216–217
ritual 20:717, 20:719
and self-efficacy 21:504–508
in situ 17:530
social networks 16:575–576
in social systems 14:384–385
symbolic 16:181
symbolic interactionism 12:352–354
traditional views on 1:76–77
interpretative paradigm 1:76–77
neoclassical economics and homo economicus 1:77
voluntaristic theory and homo sociologicus 1:76
understanding, and mirror neurons 15:583
work as 25:678
Action-sets 14:424
Activated opinions 2:195
Activation
function 16:117
of labor market 13:185
measures, of inferences 19:395–396
by opsins 17:270–271
Activation metaphor 21:350
Activation threshold hypothesis 2:576
Active, Beautiful and Clean (ABC) Waters program 24:896–897
Active agency of social work community 22:792
Active aging 6:380–381
and social change 22:154
and WHO 6:382
Active Appearance Model (AAM) 8:658–659
Active articulators 18:66–67
Active birth control 9:77
Active bystanders, role in mass killings/genocide 14:705
Active cooperation, and Honesty-Humility 11:170
Active distribution policy, welfare reforms 25:519–520
Active euthanasia 8:281–284
Active intermodal mapping (AIM), neonatal imitation 16:485, 16:487
Active labor market policies (ALMPs) 12:827, 13:178–181, 13:184–186, 18:750, 25:835
welfare reforms 25:518–519
Active learning 3:176, 6:897
Active leisure, flow in 9:263
Active life expectancy 1:328
Active media sports audiences 14:724
Active membership 17:532
Active parliaments 17:515–517
Active pluralism 3:681–682
Active sensor systems 20:419
Active Shape Model (ASM) 8:658–659
Active society, unemployment 24:736
Active tolerance 18:240
Active use impacts, in cost–benefit analysis 5:49
Active-type languages 14:212
Activism
and advocacy 1:223
in civil society 10:203
and community-based health interventions 10:648
and disability 6:451
ecofeminist 6:849
feminist 8:902–903
sex worker 19:281, 19:283
Activists
partisan 12:766–767
social work 19:870–871
Activities 18:125–126, 18:783–785
bioarchaeology 2:601–602
above neck 2:601–602
bone structure 2:601
degenerative pathology 2:601
and disability 6:467–469
estimating biological relatedness 2:602
forms of 18:785–786
patterns 24:415–416
autism spectrum disorders 2:276
and structure 18:785
see also Practice
Activities of daily living (ADLs) 1:395, 4:464
in adulthood 1:181, 1:183
Activity system 24:616–617, 24:619, 24:621
writing instruction 25:767–768
Activity theory 17:180, 18:784
Activity-based costing (ABC) 14:466–467, 14:469
Activity-based models 24:619
Activity-specific parenting 15:925
Activity–travel patterns 15:637–639
Activos 9:628–629
Actor-focus (AF) construction 23:890, 23:892
Actor-network theory (ANT) 1:80–84, 9:726–727, 16:569–570, 20:96
of brokers/brokerage 2:872–873
central problems 1:82–83
constitution of collective worlds 1:83
generalized agency and the critique of modern reason 1:82
material semiotics, inscriptions, translation, and networks 1:82–83
and culture 5:523–527
cultural analysis 5:526
pragmatist research 5:525–526
sociotechnical approach 5:523–524
cybernetics 5:645–646
ethnomethodology 21:196–197

Actor-network theory (ANT) (*continued*)
of gender–technology relationship 9:726–727
and innovation 12:171–172
intellectual foundations and historical origins 1:80–82
constructivism in sociology of sciences 1:81–82
engineering and innovation 1:80–81
Parisian poststructuralism 1:81
and knowing 13:697
Latour on 13:415–417
and materiality 14:765, 14:769–770
meaning and 16:562–563
and organization geography 17:304
philosophical developments 1:83–84
and politics 21:218
and postcolonial science studies 18:654
science and technology studies 21:203
scientific controversies 21:284
and scientific knowledge 21:310–311
and social constructivism 22:230–232
and social movements 22:439
and sociality 22:855–856
of society–technology relationship 24:136
and sociology of science 21:265
in STS anthropology 21:182
Actor-oriented statistical models, for social networks and consumption 16:556–557
Actor–partner interdependence models 16:558
Actors 21:134–135
assumptions 22:195–196
behavior, and network structure 16:520
child soldiers as 3:439–444
conflict forms 4:621–622
diffusion 6:413–414
in formulating education policy 7:267–268
governmental 7:267
nongovernmental 7:267
philanthropic foundations 7:267, 7:270
health policy 10:663–664
inhabiting social networks 5:558–559
inner-directed and tradition-directed *versus* other-directed 5:558–559
as people who remember 5:559
rational *versus* practical 5:558
integration of 12:243
and international problems 12:547
and labor history 13:171
political ecology 18:352–353
in social and behavioral sciences 15:320–321
and structures, in social change 22:195
Acts of citizenship 3:639
Acts of identity 6:499
Acts of Paul and Thecla 3:238
Acts on objects
infants' imitation of 11:621–622
implications for theory 11:621–622
Actual malice 9:395
Actual self 4:725
Actual sequence explanation 14:878
Actually existing socialism 22:834
Actually existing unsustainability 7:735
Actual–potential awareness 14:391
Actuarial methods 14:89, 14:351
Actuarialism 12:218–219
Actus reus, defenses negating 5:220–221
Acumen's Human Centered Design model 12:546
ACUPCC *see* American College &; University Presidents' Climate Commitment organization (ACUPCC)
Acupuncture 10:562
Acute coronary syndrome (ACS) 20:187–188, 20:190
coronary heart disease 4:920
Acute dystonic reaction 19:466
Acute illness, complementary and alternative medicine for 10:565
Acute pain 3:556t
Acute stress disorder (ASD) 15:497, 23:563
ACWs *see* Abstract concepts and words (ACWs)
Acyclic graphical models 2:365
AD *see* Alzheimer's disease (AD); Analog to digital (AD) conversion
Ad hoc jurisdiction 12:510
Ad infinitum 9:406
ADA *see* American Diabetes Association (ADA); Americans with Disabilities Act (ADA)
Adaboost algorithm 16:118
ADAM *see* Arrestee drug abuse monitoring (ADAM)
Adamec, Bob 13:32
Adams, Herbert Baxter 18:467
works, *The Johns Hopkins Studies in Historical and Political Science* 18:467
Adams, John 5:232, 12:897–898
on republicanism 20:477
Adams, Richard, on energy development 7:622–623
AdamWalsh Child Protection and Safety Act 2006 20:434
Adaptability
of communicative competence 4:299
of infants 16:493–494
of language 18:799–800
Adaptation, Goal-Attainment, Integration, Latent-Pattern-Maintenance (AGIL) 22:748
before 1:381–382
for analysis of action systems, up to human condition 1:390
and cybernetic hierarchy 1:385–387
differentiation 1:388–389
early 1:382–385
functionalism 1:382, 1:393
and generalized media of interchange 1:389–390
inclusion 1:382, 1:388–389
Parsons 1:381
pattern variables 1:382–387
post-Parsonian developments 1:390–393
social evolution 1:388–389
societal community 1:388–389
and study of societies 1:388–389
Adaptation function of hope 11:198
Adaptationism 1:88–89, 8:400–402, 22:904–905, 22:910
Adaptationists 22:908–909
and language development 13:321
Adaptations 1:85–90
and acculturation 1:72–73
AGIL model 22:748
alternatives 1:87
in anthropology 9:940
comparative evolutionary psychology 4:394–395, 4:398
in comparative method 4:417–418
concept of 1:85–86
elaborations 1:85–86
and coping 4:902
Darwinism 5:683–684
to dentures 17:281–282
to environmental and social process 10:547–548, 10:550–551
and evolution 8:383
evolution of body 2:723, 2:725–726
evolutionary demography 8:417
evolutionary developmental psychology 8:420–424, 8:427
evolutionary group dynamics 8:442, 8:444–446
and fitness compared 1:85–86
functional explanation of 9:520–522
future research 1:89
and human ecology 11:283, 11:285–286
and human evolutionary genetics 11:291, 11:293–294
of institutions 17:392–393
limitations 1:87
misconceptions 1:88
organizational 3:894–895, 17:375, 17:377
psychological adaptation to childhood cancer 3:459
to requirements at work 17:134
and successful aging 23:667–669, 23:671
and transportation 24:627–628
to unemployment 10:516–517
Adapted mind 22:893–894
Adaptive agent models 22:701–702
Adaptive and integrative risk governance model 7:787f
Adaptive behavioral learning 16:639
Adaptive behaviors 17:766–769, 22:704
Adaptive capability, of vestibulo-ocular reflex 25:70–71
Adaptive competence, instructional psychology 12:201–202
Adaptive cruise control (ACC) 12:133–134
Adaptive design optimization (ADO) 4:475, 4:478
Adaptive developmental regulations 21:548
Adaptive expectations (AE) 8:510–512
Adaptive expertise 4:799–801
Adaptive hypermedia (AH) 14:93–94
Adaptive leadership 13:668
Adaptive learning approach, for expectations 8:513–515
Adaptive life span 17:173
Adaptive market hypothesis (AMH) 23:487

Adaptive organizations, strategic adaptation 23:502–503
Adaptive performance, and personnel selection 17:937
Adaptive preferences 7:403
Adaptive radiation model 24:131
Adaptive sampling 23:190f
Adaptive shame 21:477
Adaptive skepticism 7:883
Adaptive Social Behavior Inventory 22:946
Adaptive social networks, and evolution 8:390
Adaptive tests 7:930, 19:420
  computerized 4:512
Adaptive toolbox 5:912
Adaptive value, of animal social learning 22:405–406
Adaptivity, and human development 11:272, 11:274
Adat law 13:757, 13:813, 13:815–816
ADC *see* Aid to Dependent Children (ADC)
Addams, Jane 1:91–96, 8:612, 22:133, 22:698, 25:569
  and anthropology 1:94–95
  biography 1:91–92
  on citizenship 1:95
  on democracy 1:93–95
  and Dewey 1:93–96
  early years 1:92
  epistemology of 1:93
  and James 1:93–96
  and Mead 1:93–94
  Nobel Peace Prize for 1:91–92
  on peace 1:91
  and political science 1:95
  pragmatism of 1:93
  settlement house movement of 1:91, 1:93–95
  and sociology 1:94
  and Spencer 1:92, 1:94
  works
    Hull House Maps and Papers 1:92–93
Addiction 1:97–102, 21:816, 23:335, 24:276–277
  and abuse 1:97, 1:100
  to alcohol consumption 1:97–98, 1:100
  background 1:97
    history 1:97
    importance 1:97
  Becker on 2:403–404
  behavioral addiction 1:99
  brain stimulation reward 2:842–843
  cannabis 1:97–98, 1:100
  to cigarette smoking 1:97–98, 1:100–101, 22:126–127
  definition 1:97–100
  and dependence 1:97, 1:100
  disorder
    substance-related disorder 1:97
  disorders 1:99–100, 1:98f
    addictive disorders 1:99–100
    relapse prevention 20:217–219
  to drug use 1:97–100
  elements 1:97–99, 1:98f
    behavior 1:98–99
    conditions 1:98
    reward 1:98–99
    stimulus 1:97–98
  to gambling 1:97–99
  habit 1:98–99
  non-substance-related disorder 1:97, 1:99–101
  and overweight 1:97, 1:100–101
  and public health 1:100–101
  severity of 1:100
  understanding 1:97
  wanting/needing 25:351–354
Addiction industry, research funding 20:524
Addiction Potential scale (APS) 15:567
Addiction Severity Index Gambling score 17:595
Addictive drugs
  that were legal in the United States 17:239–240
    heroin sniffers 17:240
    laudanum 17:239
    lessons learnt 17:241
    opium smoking 17:239–240
    response to opiate use 17:240–241
Addictive power of consumption 13:853
ADDIE *see* Analysis, design, development, implantation, and evaluation (ADDIE) model
Addio Pizzo (anti-mafia campaign) 14:424–425
Addis Ababa University (AAU) 22:608
Addison, Joseph 15:11–12
Additional Member System (AMS) 7:349
Additive acculturation 2:85
Additive bilingualism 2:580–581
Additive conjoint ISOP (ADISOP) model 24:253
  linear 24:254
Additive drift rate variability 5:904
Additive factor methods 12:80
  task network models 16:541, 16:543
Additive genetic effects 9:919–920
Additive genetic variance 8:715, 12:298
Additive manufacturing 12:149–150
Additive representations
  conjoint measurement 14:863
    specific theories 14:863
    uniqueness 14:863
Additive welfare functions 11:905
Additivity 4:874
Additivity hypothesis, for persuasion 17:945
Address terms 13:61
Adel, German 1:965
Adelante (magazine) 9:630–631
Adequacy, job displacement insurance 12:823–825
Adequate stimulation, in therapies 21:689
Adequate Yearly Progress (AYP) 21:123
ADHD *see* Attention deficit hyperactivity disorder (ADHD)
Adherence, of patients 17:598–602
  assessment 17:599–600
  to behavioral medicine treatment 2:467
  to cardiac rehabilitation 20:190
  between combined pharmacotherapy and psychotherapy 19:496–497
  compliance 17:598
  depression 17:599
  factors affecting 17:598–599
  health literacy 17:599
  improvement 17:600
  incentive-based engagement strategies 17:600–601
  information-motivation-strategy model 17:600
  interpersonal factors 17:599
  nonadherence 17:598–601
  and patient behavior 10:677–678
  physician–patient relationship 17:598–600
  psychiatric prophylaxis 19:203
  terminology 17:598
Adherent mass distributions 5:859
Adhesion contracts 4:810
ADISOP *see* Additive conjoint ISOP (ADISOP) model
adīth 12:747–749
Adiuretin 16:650
Adjacency 22:474
Adjacency list 22:454
Adjacency matrix 16:572, 22:454, 22:478
Adjectival passives 16:672
Adjectives 11:826, 17:553–557
Adjudication 5:103–107, 13:616, 14:265, 24:635–638
  appeals 5:106
  arbitration 5:103, 5:105–106
  business and organization of courts 5:105–106
  civil law 5:104–106
  common law 5:104–106
  discretion 5:106
  dispute resolution 5:103
  disputes 6:560–561, 6:565
  judicial independence 5:104–105
  mediation 5:103, 5:105–106
  policy making 5:103–104, 5:106
  procedures 5:104, 5:106
  trials 5:103, 5:105–106
Adjudicative facts 13:768–769
Adjudicator 20:417
Adjunct condition (AC) 16:4
Adjuncts 9:875, 25:1
Adjustable contracts 12:941
Adjusted *p*-values 16:84–85
Adjustment
  census 3:306
  in childhood cancer
    between dysfunction and resilience 3:462–463
    long-term outcome 3:462
  in chronic illness 10:675–676
  coefficients, hypotheses on 24:326
  of data 6:126
  definition of 5:335t
  and early friendships 9:450
  in human resource management 11:348, 11:352
  minority 8:145–146

Adjustment (*continued*)
U-curve theory of 5:335, 5:335t
Adjustment heuristic 10:826–827
of decision making 5:920
Adler, Alfred 2:701, 13:639
and Freud 19:503
lifestyle by 14:112–113
psychotherapy theories 19:503
Adler, F., works, *Sisters in Crime* 8:913–914
ADLs *see* Activities of daily living (ADLs)
Administered society 1:541
Administration 2:919
in organizations 1:103–108
action-oriented cognition building 1:106–107
mechanisms 1:106
moral community building 1:106
societal dynamics and 1:107–108
and regulation 20:161–164
as science 1:103
statistical coordination 23:426
strategic behavior in 1:104–106
of voluntary agencies 25:264–265
Administrative classifications 3:799
Administrative data 5:737, 10:304–305
for migration statistics 5:736
Administrative elites 3:5
Administrative ethics, public policy 19:583
Administrative graft 5:367
Administrative information systems 24:759–760
Administrative justice 13:429
Administrative law 1:109–114, 13:819
courts and judicial review 1:111–112
and democracy 1:109–110
and governance 1:113
Hurst's contribution 11:439
legislative and executive oversight 1:112–113
Administrative records 5:760, 6:163, 10:303, 10:306
census 3:306
labor statistics 5:799
of small areas 6:163, 6:165
for time use 24:380
Administrative research 13:640
Administrative Science Quarterly 21:323
Administrative sources, for census 3:297
Administrative units, small areas 6:167–168
Administrators
cross-cultural research, feedback from 7:932
discretion 1:110
Admissibility
of expert testimony 8:583–584
of scientific evidence 21:208–209, 21:298–300
Admissible observations 9:835
ADMIXTURE analysis 9:892–893
ADNI *see* Alzheimer's disease neuroimaging initiative (ADNI)
ADO *see* Adaptive design optimization (ADO)
Adolescence 3:451, 20:298
abstract religious thinking 20:300
and age 1:318
alcohol use in 1:513–516
apprenticeship/school learning 1:861–863
arts participation 2:9–13
biological explanations of 6:10
and conflict 4:634
coping during 4:889
defiance during 6:7–11
development and health 11:267–269
effect of stress on 23:562–569
development of logical and hypothetical reasoning in 14:306–312
developmental assets 3:445–450
developmental assets during 3:445–450
disengagement from religion 20:301
and education 4:635
epistemic doubt during 7:881–886
Erikson on 7:935–937
ethnic minority 8:114
extracurricular activity involvement and PYD prediction 3:447–449
friendship during
see also Friendship
and gangs 9:595
and hazing 10:556–557, 10:559
health-related behaviors in 10:681
identity in 11:537–542
importance of religion 20:299
life span development 14:109–110
logical and hypothetical reasoning in 14:308–310
mass trauma events in 14:748
and media 15:3
and parental motivation 15:922–923
personality changes during
see also Personality
prosocial behavior during
see also Prosocial behavior
recovery of functions during 20:57–58
religious service attendance 20:299
role across cultures 9:432–433
and same-sex youth 9:649, 9:651–653
school, influence of 20:302
self and identity development during 9:434–435, 21:432–438
self-development in 21:492–497
early adolescence 21:493
later adolescence 21:493
mid adolescence 21:493
self-regulation during
see also Self-regulation
sexuality development in 21:841–842
agency 21:841
cultural and social differences 21:841–842
effect of divorce 21:842
sexuality education 21:825, 21:828, 21:830–831
social competence during 22:216–221
social media use 24:146–150
transition to adulthood *see* Transition to adulthood
trust across life span 12:638
and youth culture 25:813–814
Adolescence gene–environment interactions 7:845
Adolescence-limited offenders (ALs) 5:139, 14:60
Adolescent justice 13:493–494
Adolescent sex offending/offenders 21:662
Adolescent Transition Program (ATP) 22:413
Adolescents 4:666–668, 21:842
age 1:115–116, 1:118–119
body image 2:735–740
death attitudes 5:861
development of 1:116–119
as actors 1:119
and computer games 4:498
endeavors 1:119–120
family's role in 1:116
neighborhood 1:118–119
peer group's role in 1:116–117
and schools 1:117–118
work 1:118
fertility of 4:558
in Latin America and Caribbean 9:69
girls, marriage and fertility of 4:558
health behaviors 1:121–127
accidents 1:121, 1:126
academic involvement or disengagement 1:124
affective factors 1:122
aggressiveness and hostility 1:124
buffering effects 1:121–125
changes in focus over time 1:122
definition 1:121
demographic variables 1:123
early theoretical models 1:121–122
emphases in current research 1:122–125
environmental variables 1:123
family relationships 1:123
future directions 1:125–126
intellectual context 1:121
life events, emotional distress, and coping motives 1:124
obesity, dietary intake, and physical exercise 1:125
peer relationships 1:125
personality factors 1:121
protective factors 1:121–126
racial/ethnic discrimination 1:125
risk factors 1:121–126
self-control and dysregulation 1:123
self-regulation 1:122
sexual risk behavior 1:121–123, 1:125
specific attitudes and efficacies 1:124–125
substance use 1:121–125
suicide 1:121–122, 1:124, 1:126
temperament dimensions 1:123
violence 1:121–123
HIV infection 11:138–140
impact and epidemiology 11:138
improving future prevention efforts 11:140
intervention efficacy 11:139
selected predictors of risk 11:139
impact of war/political violence on 25:389–392
initiation 5:509

interactions with peers 16:515–516
leisure activities choices among 13:829–836
life stages 1:115–116
mental health 19:299, 19:303–304
and political socialization 18:510–511
sex offending/offenders 21:662
sexual risk 1:128–135
background 1:128–129
biological factors 1:129–130
demographic correlates 1:129
direct rewards 1:130–131
direct sanctions 1:130–131
ecological factors 1:132–133
historical trends 1:128–129
and institutions 1:133
and neighborhood 1:132–133
opportunities 1:132
psychological factors 1:129–130
social context 1:130
social support 1:131–132
vicarious rewards 1:130
vicarious sanctions 1:130
sexuality 18:534, 18:536–538
sociology 1:115–120
soldiers 3:433–434
stress in
antisocial behavior 23:564
anxiety 23:563
cross-cultural differences 23:567–568
depression 23:562–563
substance use 23:564
suicide 23:563–564
physiological and physical health-related consequences 23:565–566
and violence 25:108
Adolphus, Gustavus 25:365
Adoption 1:136–147, 3:474
of African-American children 1:137, 1:139
causes and justifications 1:138–139
changing sources of adoptable children 1:138–139
infertility 1:138
child abandonment 1:137, 1:139
contemporary issues in 1:146
and cultural transformation 1:139–140
definitions 1:136–137
demography 1:137–138
data and collection issues 1:137–138
standardization of measures 1:138
designs, for genetic studies of behavior 9:934–935
domestic 1:136–137, 1:144
evolution of 1:143–144
and family 1:139–140
fertility 1:137–139
formal/informal 1:136
foster care 1:136–139
humanitarianism 1:137
and immigration 1:138
intercountry 1:145
international 1:136–137
irregular 1:137
kin/strangers 1:137
law 1:140
in Native America 1:137
order of 1:142
of orphans 1:137–139
perspectives 6:408–409
plenary/simple 1:136
public/private 1:136
and race 1:137
and reproduction 20:451–453
secrecy and openness in 1:144–145
social construction of 1:142–143
and social work 3:474–475
Adoption Act 1976 1:143, 1:146
Adoption and Children Act 2002 1:143–146
Adoption studies 2:451, 9:1010–1011
of sexual orientation 21:779
Adoption–diffusion theory 1:473
Adoptions Act 11:765
Adoptive families, emotional development of children in 7:442
Adorno, Theodor W. 3:776, 5:499, 7:378, 8:215, 14:735, 19:331, 20:203
on critical theory 5:295–296
and Frankfurt School 9:382–383
works
The Authoritarian Personality 2:197–198, 9:459, 14:745
Dialectic of Enlightenment 4:752, 5:295, 14:745
ADP *see* Approximate dynamic programming (ADP)
ADR *see* Alternative dispute resolution (ADR)
Adrain, Robert 9:624
Adrenal cortex, hormones functions 16:651
Adrenal gland 11:484–486, 11:488–490
Adrenal medulla, hormones functions 16:651
Adrenalin 16:749, 16:751, 17:585
neuroendocrinology 16:651
Adrenocorticotropic hormone (ACTH) 11:476–477, 23:577
Adrenocorticotropic releasing hormone (ACTH) 7:606
ADS *see* Adult-directed speech (ADS); Atrocities Documentation Survey (ADS)
Adult Attachment Interview (AAI) 2:139, 2:792, 12:231, 16:470
Adult conversation 10:111
Adult education 1:148–152
emergent population of learners 1:148
future directions 1:150–151
leadership education 13:656–657
lifelong learning
and andragogy 1:148–149
benefits and opportunities 1:149–150
skills training 22:20
see also Education
Adult entertainment 13:882
Adult family processes 20:293–294
divorce 20:293
fertility behavior 20:293–294
marriage and cohabitation 20:293–294
Adult female, marriage and fertility of 4:556–557
Adult intellectual development 4:10
Adult learning, leadership education 13:660
Adult life development in cognitive aging 1:401–403
Adult life events, and desistance/persistence 5:142
Adult Literacy and Life Skills Survey 5:847
Adult morality
in eastern Europe and former Soviet republics 1:161–169
excess mortality shifting to older ages 1:167–168, 1:167f
fluctuations and mortality at working ages 1:163–167, 1:166f
health crisis 1:161–165, 1:162f, 1:164f–165f
indirect estimation of 6:145
Adult mortality
data sources and estimates 1:153–157
censuses 1:156
surveys and survey series 1:156–157
vital registration (VR) systems 1:153–156
decline in 15:868, 15:872
future challenges 1:159
in industrialized societies 1:170–180
America 1:171, 1:179–180
causes of death to changes in 1:175–178
current levels 1:171
current-specific trends 1:173–175
driving factors of past trends 1:179–180
eastern Asia 1:171
eastern Europe 1:171
future expectations 1:179–180
life expectancy 1:171, 1:173–180, 1:173f, 1:176f–178f
Oceania 1:171, 1:179–180
southeast Asia 1:171
trends and patterns of sex differences in 1:178–179
trends in age pattern 1:175
Western Asia 1:171
Western Europe 1:171, 1:179
levels and trends in developing world (1950–2010) 1:157–159
heterogeneity 1:157–158
heterogeneity in change by sex 1:158–159
rate of 1:158–159, 1:158f–159f
Adult mortality in developing countries 1:153
Adult Participation in Learning Surveys 5:847
Adult play scaffolding, similarities in effects of 18:221
Adult prisons/jails and juvenile facilities 21:663
Adult protection 13:722
Adult research 2:660–661
Adult support differences, during toddlerhood 18:221–222
Adult ties, redefinition of 8:781–782
Adult years 7:935–937
Adult-directed speech (ADS) 3:399

Adult-directed speech (ADS) (*continued*)
and child-directed speech (CDS) 3:400
Adultery 9:729
Adulthood 2:351
age policies 1:323
autonomy and dependency in 1:181–184
activities of daily living 1:181, 1:183
behavioral dependency 1:182
challenge in midlife 1:182
complexities 1:181–182
dependency–support script 1:182
in old age 1:182–183
physical dependency 1:182
research directions 1:183–184
structured dependency 1:182
body image 2:737–739
conflict during 4:596, 4:598
death attitudes 5:861
and ego identity 5:861
and terminal illness 5:861
early life influences on health and mortality in 6:752–758
education in 7:141–142
generativity *versus* stagnation 11:539
identity in 11:538–539
integrity verses despair 11:541
intimacy *versus* isolation 11:539–541
life phase 5:512
life span development 14:110
older, sexualities in 21:843
prosocial behavior and empathy in 19:216
psychosocial development 7:941
self and emotional development in 21:424–431
self-regulation in 21:557–562
transition to *see* Transition to adulthood
trust across life span 12:638–639
young, sexualities in 21:842–844
Adult.lit 16:874, 16:867t
Adults
age differences in intelligence 1:428–430
differential age-gradients of cognitive mechanics and pragmatics 1:430, 1:430f
increase in variability and covariation 1:430
arts education 1:234
basic education 7:122–123
common sleep disorders in 22:73
circadian rhythm sleep disorders 22:75
insomnia 22:73
parasomnias 22:75
sleep-disordered breathing (SDB) 22:73–74
development of 15:369
and health 11:267–269
theoretical approaches 7:492–493
emotional development 12:5
emotions in 12:3
interpretations, and intentionality of children 12:350–351
mass trauma events in 14:747–748
mortality crossover 15:849
and new media 2:218
OSA treatment 22:73–74
physical activity guidelines for 18:95t
pragmatic disorders in 18:789–794
relationship, improvement of 20:435
sex offending/offenders 21:662
sexuality 21:666
single 8:770
sleep pattern changes 22:70–71, 22:71t
sleep-related breathing disorder 22:73
sleep-related disorders in *see* Sleep-related movement disorders
social care assessment in 12:504
source monitoring in 23:33–34
Advance organizers 13:710
Advanced care directives, end-of-life choices 7:597–598
Advanced certification, for teachers in United States 24:79
Advanced cognition 6:747
Advanced computational methods 16:878–884
Bayesian methods and Dirichlet measures 16:878–880
bootstrapping and Gibbs sampling 16:880–883
testing for concordance 16:880–881
testing independence 16:881
Advanced democratic countries, political participation outside 18:506–507
Advanced economics, governance of consumer policy in 4:734–735
Advanced economies, heterogeneous fertility decline in 9:776
Advanced marginality 24:818–819
Advanced practice nurses 16:910–912
Advanced practice registered nurses (APRNs) 16:910
Advanced producer services (APSs) 21:648, 21:651–652
Advanced psychosexual disorder, homosexuality as 20:872
Advanced Research Projects Agency Network (ARPANET) 12:600, 15:12
Advanced Research Projects of the Army (ARPA) 21:188
Advanced societies 9:19–20, 13:72
Advanced Very High Resolution Radiometer (AVHRR) 10:65
Advantage, interest as 12:386
Advantaged groups, and meritocracy 15:241–242
Advantageous selection, health insurance 10:641–642
Adverbial clauses 1:185–188
distribution of 1:186
functions of 1:187
beyond sentence 1:187
interpretation of 1:187–188
major functions of 1:185
ordering of 1:186
semantic types of 1:186–187
spectrum of possible circumstances 1:185
speech act 1:187
structure of 1:185–186
Adverbs 14:295–296
Adversarial allegiance in expert testimony 8:585–586
Adversarial and inquisitorial procedure, eclipse of 5:231
Adversarial attorney 8:585
Adversarial 'case or controversy' proceeding 12:899
Adversarial legalism 4:735, 24:637
Adversarial systems, scientific evidence and 21:297–300
Adversary, in trials 24:637
Adverse effects 19:462
of pornography 18:617–618
of systemic cocaine 19:463
Adverse environmental factors, and HPA axis activity 11:479
Adverse selection 1:345, 12:104
health insurance 10:615–616, 10:641–642
Adversity
coping 4:888–889, 4:891, 4:893
psychological stress 23:584–587
Advertisements 1:194–200
definition 1:194
form 1:195–198
function 1:194–195
mode of work 1:198–199
Advertising 1:194–200, 1:208–217, 20:650–652
anthropological structuralism 1:216
anthropology of 3:8–9
audience measurement 2:211, 2:213
branding 1:213–217
and brand-oriented society 1:209
consumer psychology 4:738, 4:740, 4:742
context of 1:208
control of 1:201–207
critique of 1:215–216
as cultural industry 1:213–214
cultural turn 1:216–217
definition 1:194
development of 1:214–215
economic effects 1:208–209
elaboration likelihood model (ELM) 1:211
ethics of 8:64
ethnographic work on 1:208
failures 1:201–202
extent of 1:202
form of 1:195–198
function 1:194–195
future of 1:211–212
and globalization 1:214–215, 1:217
and glocalization 1:214–215
and hegemony 1:210
individual effects 1:210–211
information processing tradition 1:210–211
international 12:455–461
in Korea 2:75
and mass communication 1:211–212
mass media 1:217
measuring effectiveness of 20:878
and media effects on children 15:38–39
memory-strategy usage deficits 1:211
in neuromarketing 16:701

online 12:589
online behavioral advertising 1:217
as part of integrated marketing communication 1:213
postmodern reflexiveness 1:217
regulation/deregulation 1:213, 1:215–216
semiological structuralism 1:216
in social media 1:213, 1:217
in social theory 1:216–217
sociocultural effects 1:209–210
special audiences 1:211
and Western Marxism 1:215–216
without sales 1:209
Advertising agencies 1:189–193
account system 1:192–193
compensation 1:191–192
full service agencies 1:190–191
organization 1:191
origins and early developments 1:189–190
Advertising breaks 14:721–722
Advertising control
alternative system for 1:202–203
co-regulation 1:202
factors affecting 1:203–204
impacts 1:204
mechanisms 1:202
strengths and weaknesses of advertising regulation and self-regulation 1:203, 1:203t
theory 1:202
Advertising Deontology Jury 1:206
Advisory and research team (A&RTeam) 21:93–94
Advisory Committee on Legal Education and Conduct (ACLEC) Report 13:735
Advocacy
in anthropology 1:223
and antisocial behavior 1:794
coalitions
concept of 24:297
and issue networks 12:772
and communication 4:290
criticisms 1:226
in disasters 22:757
and elder care 7:336–339
ICSW 12:497–498
issues, in autism spectrum disorders 2:279
land tenure 13:236
and motherhood 15:885
of nonprofit organizations 16:915–916
planners 1:218
planning 1:218–222, 18:204
concept of 1:220
and disadvantaged 1:221
historical background 1:218
and institutional innovations 1:221
issues 1:222
organizations 1:220–221
use of term 1:218
rebuttal 1:226–227
refugees 22:771, 22:773
role in community-based health interventions 10:648
and sanism 15:202
social work 2:111–112
and statehood 23:350
Advocacy groups 10:543, 12:371
Advocacy networks 7:268
The Advocate (newspaper) 9:638, 13:882
Advocates, publics and meanings as 17:243–244
AE *see* Adaptive expectations (AE)
AEA *see* American Economic Association (AEA); American Enterprise Association (AEA)
AECM *see* Alternating expectation–conditional maximization (AECM) algorithm
Aedes aegypti mosquito 7:815, 7:818
AEDPA *see* Antiterrorism and Effective Death Penalty Act 1996 (AEDPA)
AEDs *see* Antiepileptic drugs (AEDs)
AEI *see* American Enterprise Institute (AEI)
AEQ *see* Achievement Emotions Questionnaire (AEQ)
AERA *see* American Educational Research Association (AERA)
Aerial photography 3:174, 10:53–54
Aerodynamics 8:838
Aeromobility, cultural geography 5:412
Aeroturbation 10:54
Aesthesiometers 8:567f
Aesthetic codes 3:803
Aesthetic discourse space, popular music 3:803
Aesthetic education 1:229–235
aesthetic experience 1:230–231
aesthetic understanding 1:230–231
contemporary practices in 1:231
and creativity 1:231
and culture 1:229–232
Dewey on 1:230
and empathy 1:230–231
Greene on 1:230–231
history of 1:229–230
releasing social imagination 1:230–231
research 1:231, 1:233
training aesthetic perception for democracy 1:230
in United States 1:232–234
impact of reduced provisions for arts education 1:234
reduction in instructional time 1:232–234
Aesthetic judgment 13:3
Aesthetic knowledge, and organizational symbolism 16:181
Aesthetic mobility 5:609
Aesthetic normativity 24:821
Aesthetic pleasure 25:156
Aestheticization 5:609
Aesthetics 3:548, 13:270, 17:231, 18:824
African 1:278
art 2:16–17
and creativity 17:710
cultural 5:727–729
film 9:165, 9:168
Humboldt on 25:279, 25:282–284
and kinship 13:40
Aesthetigram 1:231
AET *see* Affective events theory (AET)
AFCENA 1:307–308
AFDC *see* Aid to Families with Dependent Children (AFDC) program
AFE *see* Arab Foundation for Freedoms and Equality (AFE)
Affairs, incidence and trends 17:486–488
Affect 1:241, 21:365, 24:174–175, 24:398–401, 25:682
and aggression 1:379
as analytical object 1:236
attunement, neonates 16:487
of cinema 6:621, 6:623
consumer psychology 4:738–740
and decision making 5:921–922
and emotions 7:527, 10:94–95
and empathy 7:550
and ethnography 1:236–239
four-field approach 1:236–237
and happiness 24:401
heuristics 10:826–827
as information 7:462
and decision making 5:921–922
job satisfaction 12:833–834, 12:836–837
negative 24:174–177
overall life satisfaction 24:401
positive 24:174, 24:176–177
and procrastination 19:76–77
and risk 20:713
risk-taking behavior 17:811
social well-being ratings in ATUS 2010 24:398–401
and sociality 22:856–859
studies in anthropology 1:237–238
emergent practices 1:238–239
Affect control theory (ACT) 7:537, 22:539
and emotions 7:537
Affect dysregulation 2:759
Affect misattribution procedure (AMP) 11:715–716
Affect program (AP) theory 7:468
of basic emotions 7:468
of facial expressions 7:457–459, 8:668
contemporary alternatives to 7:459
Affect theory (sociality) 22:856
Affect–cognition priming, social comparison 22:211
Affective and social development
coeducation schooling 3:929–930
single-sex schooling 3:929–930
Affective behaviors 7:484–485
Affective changes of acculturation 1:70
Affective citizenship 21:729–730
Affective climate, organizational level 7:509–510
Affective components, test anxiety 24:244
Affective disorders 14:441–443, 15:213–214
molecular genetics 15:213–214
molecular imaging studies 14:441–443
structural and fMRI studies 14:441
twin studies 15:213
Affective empathy 17:776–777, 19:218
Affective events theory (AET) 7:507, 12:834

Affective factors, of adolescent health 1:122
Affective labor 1:239
Affective memories, in vocational psychology 25:241
Affective neutrality (pattern variable) 1:382
Affective self 21:419–420
Affective social competence 7:485
Affective states 1:241–243, 1:245
Affective turn 25:211–212
Affectively valenced stimuli, startle modulation by 23:332
Affective-motivational personality constructs 17:829
Affectivity
  individual differences and genetic factors 9:920–921
  pattern variable 1:382
Affect-regulation motivation 1:241
  definitional issues 1:241–242
  and emotions 1:241
  future research 1:246
  and moods 1:241
  pro- and contra-hedonic affect-regulation motivation 1:241–243
    age-related differences 1:244
    cognitive demands 1:243–244
    cultural differences 1:245
    individual differences 1:244–246
    instrumental-affect perspective 1:242–243
    mixed-affect perspective 1:242
    personality 1:245
    self-esteem 1:245
    trait-related differences 1:245
Affectual solidarity 12:403–405
Afferent agraphia 6:736
Afferent neuronal connections of hypothalamus 11:486
Afferent neuroprosthetics 16:714–718
Afferents, vestibular 25:65
Affiliation
  communities of 4:387–388
  and ethnicity 8:102, 8:104
  group 10:423
  and neonatal bonding 16:469
Affiliation-intimacy, and motives 15:952–953
Affiliative motivational states 16:612–613
Affinal kinship terms 13:61
Affine transformation 16:36
Affirmative action 1:256–268
  aims and justifications for 1:250–251
  American-style 1:249–250
  as an umbrella concept 1:248
  cognate terms and 1:248
  collective mentoring by 9:619
  comparative policies and controversies 1:248
  consequences of 1:266
  definition 1:262
  degree of control of corporations 1:253
  in education 1:253, 1:260, 1:262–267
  in employment 1:252–253, 1:262–267
  future research 1:259–260
  gerrymander apportionment 7:343
  in government contracting 1:260
  historical sociology 1:262
  interactions between measures and 1:253
  legal constraints on 1:254
  measures 1:248–250
  mechanisms for pursuing 1:260
  and organizational diversity 6:598
  in political representation 1:253
  for procuring public contracts 1:253
  public opinion on 1:267
  and race 19:808
  social movements 1:264, 1:267
  symmetrical and asymmetrical 1:252
  through contract compliances 1:257–259
  Title VII of the Civil Rights Act and 1:256–257
  as a tool in land reform 1:253
  use of term 1:256
  Wilson on 25:586–587
Affirmative discovery
  in England 6:519–520
  in United States 6:517–518
Affixes (linguistics) 15:820–821
Affluence 4:752
  and property crime 19:164
Affluent nations, reproductive rights in 20:468–476
Affluent society, and consumption 4:759
Affluent Worker study 10:279
Affordable Care Act (ACA) 6:879, 10:599, 10:642, 15:101, 18:430–431, 22:33–34
Affordable food, securing 10:389
Affordable Health Care Act 13:505
Affordable private schools
  in developing countries 1:269–271
  in Africa 1:269
    children's achievement in 1:270
    developments in recent years 1:270
    fees in 1:269
    longetivity of 1:269–270
    number and proportion of 1:269
    teacher salary and qualifications in 1:270
Affordances 16:739
  and embodied cognition 7:422
  perceptual learning 10:129
al-Afghani, Jamal al-Din 24:603
Afghanistan
  education in 7:195
  military operations in 13:534
AFM *see* Atomic force microscope (AFM)
AFMs *see* Alternative food movements (AFMs)
AFNs *see* Alternative food networks (AFNs)
AFP *see* Australian Federal Police (AFP)
Afri Twin project 21:116
Africa 7:36
  affordable private schools in 1:269
  Arab nationalism in 16:271, 16:273
  Atlantic slave trade from 22:56
  child mortality in 12:22
  citizenship in 3:659
  civil society in 18:390
  climatic variability 3:868–869
  and colonialism 22:57
  confederation 4:575
    Pan-Africanism and postcolonial experiences 4:570–571
  customary law 23:654
    distinct features 23:654–655
    survival of 23:655
  diasporas 1:286, 6:353
    study of 11:2
  environment of 1:292–299
    agriculture and food security 1:293–294
    biodiversity 1:295–296
    coastal zones 1:296–298
    drinking water 1:297f
    ecosphere 1:294–298
    ecosystem degradation 1:295–296, 1:298
    extreme weather events associated with climate change 1:295f
    global environmental change and climate change 1:294–295
    human sphere 1:292–294
    knowledge generation 1:298
    land and land use 1:293
    Millennium Development Goals 1:295–298, 1:297f
    policy and policy-making processes 1:298
    population, migration, and the environment nexus 1:292–293
    poverty 1:292–294, 1:296–298
    regional development planning 1:298
    research and development 1:298
    resource use, indigenous knowledge, and livelihoods 1:293
    water resources 1:296
  feminism in 8:909
  fetishism in 9:105–110
  gender in 9:781–784
  genital surgeries in 5:480
  heterosexuality in 10:812–818
  history 1:288–289
  and household slavery 22:57
  human rights 11:376
  impact of slavery 22:57
  legal systems 1:272–275
  national liberation movements 16:241–245
  origins of food production in 9:301
  peace building and music in 25:845–846
  sexuality in 10:814
  slaves/slavery in 1:286
    history of 22:54–57
  social ontogeny 6:840
  social science infrastructure 22:604–614
  state building 1:287, 1:289
  structural adjustment in 23:598
    adjustment with transformation 23:600–601
    Cold War 23:598–599
    conditionalities 23:600
    conflict 23:598–599, 23:601–602
    debt burden 23:600–601
    donor nations 23:601–602
    implications for social work practice 23:602

International Monetary Fund 23:598–603
neocolonialism 23:599, 23:601–603
neoliberal economics 23:598, 23:602
origins of economic crisis 23:598–600
programs for 23:599–602
World Bank 23:598–603
sub-Saharan *see* Sub-Saharan Africa
systematic racism 22:56–57
urbanization in 24:965–972
counterurbanization 24:970–971
in globalizing world economy 24:966–967
history 24:965–966
informal sector 24:967–968
levels 24:966f
migration 24:965–966, 24:969–972
and population growth 24:969–971
primary commodities 24:967, 24:972
prospects of 24:971–972
urban economies and livelihoods, and built environment of city 24:967–969
West Africa *see* West Africa
Africa Command (AFRICOM) 15:486–487
Africa Health Strategy 2007–2015 1:293
Africa-EU Peace and Security Partnership 17:633–634
African American Vernacular English (AAVE) 18:136
African Americans 12:140–141, 12:143, 16:18, 22:320
adolescents 8:770, 21:841–842
children
adoption of 1:137, 1:139
sexuality in 21:841
Civil Rights Movement 3:691–694, 16:18–19
household 3:204
Maroons 14:591–594
and musical categories 3:801
scholars of critical race theory 5:267
and stereotype threat 19:791
syncretism of 23:882
and urban violence 24:949–951
and wrongful convictions 25:789–790
youths, schooling experiences of 9:680
see also Blacks
African Charter on Human and Peoples' Rights 11:800, 13:720
African Charter on the Rights and Welfare of the Child
of 1990 20:582
of 1999 13:720
African Commission 11:800
African Cross-Border Initiative 12:574
African Development Bank 21:152
African Health System 1:284
African Jamaicans 11:2
African languages 1:276, 11:824
African Law Studies 13:758
African library 22:604
African National Congress (ANC) 16:268, 19:849, 21:745–746, 23:78
African project of Princeton's Office of Population Research 1:281
African Renaissance 5:532–533
African self consciousness scale 19:834
African studies 16:266–270
on culture 1:276–279
historical overview 1:276–277
phases in 1:276
today 1:278–279
on health 1:280–285
African Health System 1:284
AIDS epidemic 1:284
AIDS era 1:283–284
future of 1:285
good health 1:280
health levels 1:280–281
improvement of health 1:283
life expectancy, at birth 1:282t
life expectancy, increases in 1:282t–283t
past mortality trends 1:282
research literature 1:284–285
in second half of twentieth century 1:282
well being 1:280
and World Health Organization 1:280, 1:285
history 1:286–291
African Studies Association (ASA) 22:662
African tricksters 24:652–654
African Union (AU) 17:633
African witchcraft 25:607
AFRICOM *see* Africa Command (AFRICOM)
AFRN *see* Australasian Agri-Food Research Network (AFRN) conference
Afro-Asian Conference (1955) 3:314
Afrobarometer 5:793, 5:849, 19:567
Afrocentricity
and identity 19:834
and social work 1:305–309
AFCENA 1:307–308
approach to 1:306
black freedom movement 1:305
black social work movement 1:305–306
early beginnings 1:305
Eurocentricity 1:306–307
proof of 1:308
seeking new knowledge 1:305–306
theory and practice 1:306–307
AFSCET *see* Association Française de Science et des Systèmes Cybernétiques, Cognitifs et Techniques (AFSCET)
Afterlife 5:867–868
After-school child care 3:392
Aftertastes (senses) 21:618
After-the-event insurance 14:267
Aga Khan Foundation 5:417
Agamben, Giorgio 3:776
Agar 15:136
AGCOM *see* Autorità; per le granzie delle communicazioni (AGCOM)
Age
and accident liability 12:131–132
and affect-regulation motivation 1:244
and alcohol consumption 1:513–514
anthropological aspects 1:318–322
bureaucratic rationality 1:334
change design, of cognitive aging 4:10
and chronotype 3:570
and crime 1:310–317
behavioral trajectories and group-based analysis 1:313–314
career criminals and population heterogeneity 1:312
criminal careers 1:310–312
debate about research 1:311–312
delinquency 1:310, 1:312–314
developmental criminology 1:311–313
deviance 1:310, 1:312
dimensions in DLC criminology 1:310–311
life course 1:310–314
longitudinal study 1:310–313
maturation 1:312–313
state dependence/persistence and desistance 1:312–313
cultural significance of 1:334
and cyberbullying 5:641
determinations, archaeometry 1:905–906
different conceptualizations of 1:420
discrimination and regulation, older workers 17:186–187
and enriched environment 2:838
estimation, in archaeology 3:562
and fertility estimation 6:145
at first birth
in formerly socialist European countries 9:99–102, 9:100f
in Latin America and Caribbean 9:69–70, 9:70t
in life history theory 9:41
of first exposure to pornography 18:614
and free time 13:846
and generation 9:853–855
and health 15:100
and homicide 11:165–166
and inequality 11:897
and intellectual competence 4:12–13, 4:13f
and interviewing 5:853
and joking 11:418
and leisure 13:846, 13:855
LEXIS diagram for 6:156
and life course 14:76–78
and linked lives 14:231, 14:234
and long-term heterosexual relationships 20:915–916
at marriage 8:767
and migration 15:397–399, 15:403, 15:405, 15:435, 15:398f, 15:402t
misreporting of, and mortality crossover 15:849–850
and obesity 17:76–77
of offenders 20:199
of onset of illicit drug use 6:680
and organizational diversity 4:739, 4:741, 6:596, 6:598
and parenting 2:475
patterns
of adult mortality 1:175
of AIDS-related mortality 15:839–841, 15:840f
of demographic events 6:123–124
in demographic processes 6:133–134
of Hutterite fertility 11:446

Age (*continued*)
of oldest-old mortality 15:864–865
social and cultural contexts 9:867–868
personality-religion links 17:804–805
population distribution, evaluation and adjustment of 6:129–133
and powerlessness 1:548
promotion and 19:136
and psychosocial development 7:941
-related differences, and questionnaires 19:772
and second language learning 21:365
and self-concept 21:461–462
and sexual development
menarche 14:330
sexual debut 14:330
and sleep 22:59–60
sociology of 1:327–332
-specific homicide offending trends 5:194–196
standardization 1:341
mortality 6:121
stratification 1:329–330, 1:333–337
age structure 1:335–336
cohort analysis 1:333–336
economics and generational equity 1:336
effects of stratum size 1:335
as feature of populations 1:333–334, 1:333f
as feature of social structure 1:334
interstratum inequality 1:335–336
interstratum variation in intrastratum inequality 1:336
life course 1:334–336
and life expectancy 1:335
paradigm 1:334–335
policy 1:334–336
and social theory 1:334–336
structure, for population forecasts 18:593–594
systems, and hidden knowledge 1:320–321
and value of life 25:18–20
and wisdom 25:595
work–age patterns 13:194
and workplace attitudes 1:423–424
and workplace outcomes 1:420–421
and youth culture 25:808–809, 25:811
Age, period, and cohort (APC) analyses 22:275, 22:278
of cognitive aging 4:10–11
data collection strategies 4:11
sequential strategies 4:11
hierarchical models 17:731–732
linear regression models for 17:728–730
conventional approaches and limitations 17:729–730
methodological process 17:728–729
theoretical problems 17:730
model-based demographic forecasting 17:732
Age dependence, organizational ecology 6:890–891
Age differences
in emotional experience 7:490–491
in emotional expressions 7:491
in emotional regulations 7:491–492
and emotion-related physiology 7:491–492
in prosocial moral reasoning 7:317
Age Discrimination in Employment Act 1967 6:531
Age distribution
of death 6:151
population dynamics 18:580
Age groups
generational conflict 9:857
in human society 9:857
Age heaping 14:89
data adjustments 6:129
Age of acquisition 21:350, 21:353–354
Age of Colonialism 12:363
Age of criminal responsibility 12:960–961
Age of degenerative diseases 15:859
Age of development 3:99–100
Age of Discovery 12:363
Age of Enlightenment 7:657, 12:641
Age of globalization 10:247
consequences 10:247
definitions of 10:247
see also Globalization
Age of pestilence and famine 15:859
Age of preservation 4:203–204
Age of receding pandemics 15:859
Age of Revolution 25:291
Age of the image 25:169
Age of World Trade 12:363
Age policy 1:323–326
fall of 1:325–326
government by age 1:323
implementation of 1:324–325
to comprehensive age policies 1:324–325
to coordinated age policies 1:324–325
for individual's biography 1:323–324
modern state 1:323
old age insurance 1:324
perverse effects of 1:325
retirement pensions 1:324
rise of 1:325–326
for shaping life course 1:323–324
Age set 16:840
Age structure 1:327–328, 1:338–343, 15:940–942
age-dependency ratio 1:340
age-specific growth rate 1:340–342
age-standardization 1:341
and crime 5:196–197
and demographic transition 1:327, 1:338–340, 1:342
and demography 6:177
and development 1:327
impact of HIV/AIDS on 15:842
mathematical expressions of 1:341–342
measurement 1:340–341, 1:340t, 1:341f
population aging 1:338–339
and population dynamics 1:338–340, 1:339f–340f
and population growth 1:342
population momentum 1:342
population pyramid 1:338–340, 1:339f
stable equivalent 1:342
stable population 1:342
Age-based thresholds 1:323–324, 1:326
Age-based time metric for sexual development 14:333–334
Age-comparative design, of cognitive aging 4:10
Age–crime curve 5:196
Age-dependency ratio 1:340
Age-dependent recovery 20:57–58
Age-graded theory 14:60–61
Ageism, and elder care 7:336
Agencification 19:597
Agency 1:903, 2:314, 4:381, 5:566, 7:16–17, 11:108–109, 11:816, 13:205, 13:729, 15:940, 25:826
art 2:16–17
autonomous agency 2:318
choice as 3:520–521
in cultural studies 13:730–731
cybernetics 5:645–648
and education 7:147–148
and fetishism 9:109–110
free 2:314–318
Giddens on 10:135–136, 10:138
health economics 10:615
human 15:941
in hypnosis 11:458–460
idea of 5:587
and identity 7:118, 11:532–535
individual 13:729
infants as 12:9–10
and institutional change 17:393
institutional logics 17:393
and institutionalism 12:184
in legal consciousness studies 13:731
and liberalism 14:13–14
and life course 14:80–83
and linked lives 14:232
moral 2:314
and perceptual learning 10:132
and political ecology 18:351, 18:353–354
practical reasoning 18:776–781
principle of 15:941
regulatory 20:181–186
responsible, conditions for 20:588–589
and revolution 20:622–623
and ritual 20:722, 20:724
and self-efficacy 21:504f
in self-regulated learning 21:535, 21:537–538
sense of 21:420–421
sexual 21:841
slaves/slavery, history of 22:54, 22:56
social psychology 22:538–540
and social structure 22:724–725
and street children 23:536
and structure 21:531
and working classes 25:695
and youth culture 25:809–812
Agency costs 4:942
Agency for Healthcare Policy and Research (AHCPR) 19:383
Agency for Healthcare Research and Quality (AHRQ) 17:371–372, 24:282

Agency for Science, Technology and Research (A*STAR) 11:319
Agency for women 9:732–733
Agency freedom 9:401
  in politics 11:559
Agency law 3:29
Agency models *see* Agency theory (AT)
Agency of Cultural Affairs (Bunkachô, Japan) 5:417
Agency theory (AT) 1:344–350, 6:59, 8:775, 14:113–115, 19:892, 19:894
  and civil service 3:697–700
  exemplary applications and contributions 1:347
    complete contracting 1:348
    criticisms and limitations 1:347–349
    defense of 1:349
    given and state independent utility/risk functions 1:347–348
    intrinsic motivation 1:348
    linear models 1:348
    practical usefulness 1:348–349
    self-interest/opportunism 1:348
  foundations
    asymmetric information 1:344–345
    core insights and predictions 1:345–347
    delegation and conflict of interest 1:344
  in human development 14:103–104
  organizations 17:311
Agenda 21 7:813, 23:796, 23:813
Agenda building TA 24:126–128
Agenda setting 1:351–356, 15:29, 18:288
  attention 1:357–361
  attributes 1:352–355
  Baumgartner and Jones' model 1:359
  for climate change 3:858
  and comparative public policy 1:360
  consequences of 1:354–355, 15:29
  effects of
    explaining 1:353–354
    second level 1:352–353
    third level 1:353
  first studies 1:357–358
  and framing 1:361
  issue characteristics 1:358
  issue definition 1:361
  Kingdon's model 1:358
  and mass media
    effects 22:345
    function 14:714
    impact on 15:29
  media agendas 1:351–352
    sources 1:354
  need for orientation 1:354
  policy monopolies 1:359, 1:361
  policy windows 1:358–359
  public agendas 1:351–352
  punctuated equilibrium model 1:359
  recent trends 1:360–361
  research 1:357
  theory 1:355
  venue shopping 1:359–360
Agendas, political 1:362–366, 18:761–762
Agent-based models (ABMs) 1:367–372, 1:672, 14:418–419, 21:532, 22:702
  approximate Bayesian computation 1:371
  cellular automata 1:367–368
  criticisms 22:703–704
    artificiality aspect 22:703
    defense against 22:703
    principle of emergence in complex systems 22:704
    rationality and adaptive behavior 22:704
    unwrapping 22:704
  demographic microsimulation 15:344–345
  disease models 1:370
  emulators 1:371
  evolution of language 8:369
  family and household demography 8:729
  of geospatial data 10:102
  hierarchical models 1:368–369
  history 1:367–370
    Epstein and Axtell's artificial societies 1:368–369
    Kauffman's random networks 1:368
    rabies in raccoons 1:369–370
  individual level adaptation 22:703
  individual-based models 1:367
  limitations of 1:370–371
    distances between models 1:371
    inference 1:371
    verification and validation 1:370–371
  mathematical models 14:786
  neighborhood segregation model 22:702
  of segregation 20:552
  for studying diffusion of behaviors 6:415
  Sugarscape 1:368–370, 1:370f
  TRANSIMS 1:367, 1:370–371
Agent-based simulations 4:191
Agentes Juveniles en Salud (Young Health Activists) 6:43
Agent-first heuristic, of psychological verbs 16:672
  comprehension heuristic 16:672
Agential realism 21:203–204
Agentic qualities 18:126
Agentic state, in obedience 17:63–64
Agentic theory, of socialization and education 22:863
Agent-initiated surveillance 23:737
Agentive functions 21:344
Agent-neutral conception 4:680
Agent-neutral values 4:680
Agents 1:536
  of justice 12:947–948
  responsibility of 20:587–588
  of ritual 20:719
  in simulation environments 4:487–488, 4:487f
  of socialization 18:509–510
Age-related differences, contra-hedonic motivation 1:244
Age-related expectations, for life management 14:86–87
Age-segregated social structures 7:137
Age–sex variation 6:116
Age-specific birth rates (ASBRs) 9:99, 9:102f
Age-specific fertility rates (ASFRs) 6:110–113, 9:318–319, 15:397–399
Age-specific growth rates 1:340–342, 6:134
Age-standardized death rate (ASDR) 1:341
Age-structured cycles 18:581–582
Agglomerations
  economies 7:64, 11:858
  and entrepreneurship 7:687–688
  and regional science 20:147–149
  and urban spaces 24:858
Agglomerative clustering algorithms 23:393
Aggregate 12:210
  economy 12:211
  risk 12:210–211
Aggregate matching function 21:334–335
Aggregate production functions 3:92
Aggregate social entities 21:5
Aggregated data
  and ecological fallacy
    equivalence conditions for 6:866–867
  inverse projection 6:149
Aggregated user representations (online information) 4:508
Aggregate-level data 19:566
  assessment and adjustment of 6:128–135
    adjustments over time 6:134–135
    age patterns 6:133–134
    completeness 6:128–129
    parity data 6:129
    population distribution by age 6:129–133
  dara generating processes 2:448
Aggregate-level opinions 19:566
Aggregate-level studies, of homicide 11:163–165
Aggregation 7:915
  and macroeconomics 14:400–404
Aggregation of narratives 16:214–215
Aggregators 13:621
Aggression 1:373–380, 5:642, 14:367, 19:451–453, 22:156, 24:643, 5:639t
  to abusive supervision 23:701
  and agreeableness 1:467
  of animals 2:800–801
  appetitive 24:439
  archaeology of war and conflict 4:600–606
  in children
    parenting and 2:474
    studies on 3:882
  definitions 1:373
  and deindividuation 6:39
  directed against ego 23:677
  due to shame 21:875
  early developmental trends in 6:12
  electronic aggression 5:638, 5:640–641
  facilitative 24:439, 24:439f
  gender differences in 9:753–754
  Heider on 10:753
  hostility 1:374–375
  mediators 1:379
  peer aggression 5:639t
  and prosocial behavior 19:229
  research findings 1:377–379
    anonymity 1:378–379
    arousal 1:379
    callous unemotional personality traits 1:377

Aggression (*continued*)
cognition 1:379
development and stability over time 1:377
emotion/affect 1:379
environmental factors 1:378–379
environmental stressors 1:378
gender differences 1:377
genetic predispositions 1:378
hormones 1:378
impulsivity, executive control and self control 1:377–378
intelligence 1:378
person factors 1:377–378
personality traits, Big Five 1:378
provocation 1:378
social rejection 1:379
substances 1:379
trait anger 1:377
violent environment 1:378
violent media 1:378
weapons 1:378, 1:378f
research methodologies in social psychology 1:376–377
brain scanning techniques 1:377
laboratory assessments 1:376
nonexperimental research 1:376
observations 1:376–377
risk factors 1:379–380
role in friendships during adolescence 9:435–436
and self-conscious emotions 21:476, 21:478–479
and self-esteem 21:524
social psychological theories of 1:374–376
arousal 1:374–375
cognitive labeling and excitation transfer 1:374–375
cognitive neoassociation theory 1:375
early social psychological theories 1:374–375
frustration-aggression hypothesis 1:374
general aggression model 1:375–376, 1:375f
information processing and script theories 1:375
learning theories 1:374
mainstream cognitive theories 1:375
theories outside social psychology 1:374
societal aggression and violence 1:379–380
types and characteristics of 1:373–374
and war/political violence 25:391
Aggressive behavior 1:594
and media 15:4
Aggressive driving, transportation 24:623–624, 24:626
Aggressive joking 11:417
Aggressiveness
and adolescent health 1:124
and personality psychology 17:910–911
proactive 9:436–437
reactive 9:436
AGIL model *see* Adaptation, Goal-Attainment, Integration, Latent-Pattern-Maintenance (AGIL) model
AGIL schema 17:522
Agincourt Health and Demographic Surveillance System 12:434
Aging 2:350–351, 7:602, 17:179–180, 25:590
adolescence *see* Adolescence
Ca^{2+} in 12:713–714
changing profile of, and urban migration 1:321
in China 2:69
and fertility transition 9:47–48
and comparative health systems 4:401
coping during 4:889
differential aging 6:378
disengagement theory of 20:599
ecology of 6:884–889
economics of 20:599
and emotion-related physiology 7:491–492
emotions 21:430
and emotions 7:490–495
and epigenetics 24:688
experience of maturation and 1:319–320
family and ambivalence toward 1:318–319
and feminist social work 8:938
and fertility 9:4
and generation 9:854
geography of 1:407–412
emerging critical themes 1:409–410
emerging critical theory 1:410–411
places 1:409
spaces 1:407–409
grandparent–grandchild relationships 10:336
and health 6:378–381
in old age 1:394–397
and health care 15:100, 15:104
heterosexual expression and 10:796–800
influence in adulthood 6:756–757
and intelligence 12:294
international perspectives 6:379–380
and job displacement insurance 12:829–830
junk theory of 17:175
and labor market 1:413–419
dependency ratio 1:413, 1:418
labor force projections 1:414–415, 1:414t–415t
older workers 1:413–414, 1:417–418
productivity 1:417–418
retirement and extending working lives 1:415–417, 1:415t
and law 13:420–426
growing interest in and development of the field 13:424
history of 13:420–421
mapping the field of 13:421–422
tool for social change 13:425
LGBT
see also Lesbian, gay, bisexual, transgender (LGBT individuals)
life course 1:318–319
and memory 1:398–406
adult life development in cognitive aging 1:401–403
brain imaging 1:399–400, 1:403–404
dementia 1:398–399
episodic memory 1:400–403
genetics 1:399, 1:401, 1:404
memory disorders 1:398
memory systems and cognitive functioning 1:399–401
perceptual representation system 1:399
procedural memory 1:399
semantic memory 1:399–402
variability in cognitive performance 1:403–404
working memory 1:400–401
minority ethnic elders 15:440–445
and morbidity compression 4:463
and mortality 15:845
of navigational abilities 23:167–169, 23:169f
old age work participation 17:182–189
pension policy 17:690, 17:694–695
and population changes, in Eurasia 18:550–552
population needs, response to 7:339–340
process 18:542–543
and prospective memory 19:276–277
and self-regulation 21:560
stigma of 23:465–466
successful 23:667–672
theories of 1:435–439, 6:381
Baltes' SOC theory 1:435
coconstructive perspective 1:437
dementia 1:437–438
Erikson's stage model 1:435–436
evolutionary development 1:435
executive stage 1:436
gains and losses 1:435
generativity 1:435
integrity *versus* despair crisis 1:435–436
intimacy crisis 1:435
legacy-creating stage 1:437
reintegrative stage 1:436–437
reorganizational stage 1:436–437
responsible stage 1:436
Schaie and Willis' stage theory of cognition 1:436–437
stagnation 1:435
successful, normal, and pathological aging 1:437–438
unsuccessful aging 1:438
and vestibular system 25:69
virtual spatial navigation 16:351–352
warranties 17:172–176
and work 1:420
age-related effects 1:420
career development 1:421–422
innovation 1:422
resistance to change 1:422
training performance 1:421
in workplace 12:132
Aging brain 2:795–798
adult age differences in intelligence 1:428–430
differential age-gradients of cognitive mechanics and pragmatics 1:430, 1:430f

increase in variability and covariation 1:430
brain reserve 2:797
brain–cognition–personality relations 2:796
cognitive neuroscience 2:797
cognitive plasticity 2:796–797
cognitive reserve 2:795, 2:797
computational neuroscience 1:428, 1:431–433
deficits in basic information-processing mechanisms 1:430–432
attentional and inhibitory mechanisms 1:431
processing speed and variability 1:431
resource-reduction account 1:431–432
working memory 1:430–431, 1:431f
epigenetics 2:795
interdiscipline integration 1:428
cognitive aging phenomena 1:428–429, 1:429f
proposals to integrate studies of brain, cognition, and behavior 1:428–429, 1:429f
life span theory of cognitive development 1:428, 1:432
neurogenesis 2:795–796
psychometrics of human abilities 1:428–432
Aging enterprise 1:325
Aging in Manitoba Study 23:671
Aging mind 1:428
aging brain of 1:432
attenuated neuromodulation 1:432
reduced hemispheric asymmetry 1:432
integrating facets and levels of 1:432–433
Aging out of crime process 1:311–312
Aging population
caregiving 3:135–137
needs, social work's capacity to respond to 7:339–340
Aging society 18:543
AGL *see* Artificial grammar learning (AGL)
Agnicayana 20:862
Agnosia 1:440–443, 16:632, 25:203
apperceptive 1:440–441
associative 1:440–441
and brain damage 1:440–441, 1:443
knowledge retrieval levels 1:440
knowledge types 1:440
meaning of term 1:440–441
visual 1:441–442
definition 1:441
prosopagnosia 1:441–442
Agnosticism, *versus* pragmatism 7:798–799
Agnostics 20:356
Agnotology, and ignorance sociology 11:582–583
Agomelatine 1:779
Agon (Greek) 1:445
Agones 1:444
Agonism 1:444–449
agon as democratic narcissus 1:444–445
against antagonism 1:447–448
bounded state of agon 1:445–446
and conflict/consensus 1:446–448, 4:591
contestation 1:447
deliberation 1:444, 1:446, 1:448
and democracy 1:444–448
and liberalism 1:446, 1:448
pluralization against pluralism 1:446–447
politics 1:444–449
and poststructuralism 1:446
staging the contest 1:446
violence 1:444–446, 1:448
Agora 1:448
Agoraphobia 2:793, 3:766, 17:469, 17:474–481
behavioral model 17:476–477
biological model 17:477
clinical implications 17:479
cognitive model 17:476
comorbidity 17:476
definition 17:474
development and avoidance behavior 17:475–476
diagnostic criteria 17:474
from DSM-IV and DSM-5 17:475
epidemiology 17:475
exposure therapies for 8:606
life events and stress 17:476
mechanisms of change 17:477
anxiety sensitivity 17:477
panic appraisal 17:477
medical conditions 17:476
and panic disorder 17:476
patients, fear-potentiated startle in 8:865
prevalence 17:475
treatment 17:477–478
cognitive restructuring 17:478
exposure in vivo 17:478
guidelines 17:477–478
introceptive exposure 17:478
pharmacological treatment 17:478
psychoeducation 17:478
relaxation and breathing training 17:478
Agouti viable yellow allele 7:835
AGRA *see* Alliance for a Green Revolution for Africa (AGRA)
Agrammatic aphasia 16:672
Agrammatism 1:818, 14:176, 16:683–684, 16:690
Broca's aphasia with 16:690
difficulties with 16:690
inquiry into 14:177
challenges 14:181
linguistic approaches to 14:176–182
neurological Approaches to 16:690–697
syntactic deficits 16:690
trace-deletion hypothesis 14:177–181
working memory and 16:693–694
Agraphia 1:450–455, 6:732–738
allographic dysgraphia 1:453–454
apraxic agraphia 1:453–454
assessment principles 1:454–455
callosal agraphia 1:454
cognitive models of spelling 1:453
cognitive neuropsychological approach 1:451–454
central dysgraphia 1:453–454
peripheral dysgraphia 1:454
processing models 1:452–453
deep dysgraphia 1:453
deficits in classical aphasiology 1:450–451
forms of 6:734–736
graphemic buffer dysgraphia 1:452–454
lexical agraphia 1:453
neuroanatomy 1:455
oral language 1:450
phonological dysgraphia 1:453, 1:455
in scripts 6:736
surface dysgraphia 1:453, 1:455
written language 1:450
acquired impairments 1:450–451
Agrargeschichte 17:669–670
The Agrarian History of England and Wales (project) 17:669–671
Agrarian life 2:148–149
Agrarian societies
change in, and development geography 6:314, 6:316–317
encroachments of land 13:235
energy flow analysis 7:628–630
history
Europe 17:670
peasants 17:669–670
United States 17:670
in Japan 12:794–795
political economy of 1:456–462
anti-imperialism and transitions to socialism 1:459
colonial experiences 1:457–458
corporate power and technology 1:460
development and underdevelopment (1950s–1970s) 1:458–459
ecology 1:460–461
fates of peasants and small farmers 1:461
founding moment 1:456–457
globalization and development (1970s onwards) 1:459–460
precapitalist agrarian formations 1:457
transitions to capitalism 1:457
in West Africa 1:300
social solidarity 1:301
Agrarian South: Journal of Political Economy 1:457
Agrarianization, fire and 9:197–198
Agrawal, Kanu 24:819–820
Agreeableness 1:463–470, 9:236, 16:493–494, 17:777, 17:787–788, 17:802, 17:914–915
with ability of students 17:788
with academic performance 17:788
and adolescent personality 17:864–866
in Big Five 1:463–464, 16:743
and conflict 4:598
of emotions 1:465–466
and empathy 1:466, 1:468–469
and Honesty-Humility 11:169–170
individual differences 1:463–466
measurement 1:463–464
normal and abnormal variants of facets of 9:243t

Agreeableness (*continued*)
opponent process theory 1:468–469, 1:468f
and personality traits 1:463–466, 17:922–923
and social behaviors 1:465–468
aggression 1:467
conflict resolution 1:467
cooperation and competition 1:465–466
friendship 1:465
helping 1:466
prejudice 1:467–468
social desirability artifact 1:464
theoretical accounts of 1:464–465
Agreed-upon languages 11:820
Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) 12:267, 12:269
Agreement tendencies, powerlessness 1:545
Agreements, joint action 12:842–843
Agribusiness 23:808
Agricultural Adjustment Administration 20:421
Agricultural and Resource Inventory Surveys through Aerospace Remote Sensing 20:422
Agricultural change 1:484–487
Agricultural crime 5:159–160
Agricultural education 1:481
Agricultural extension 1:481
Agricultural growth 2:766
Agricultural history 17:669–675
in Europe 17:670
in peasants 17:669–670
Agricultural History 17:670
Agricultural History Review 17:669–670
Agricultural labor 1:478
Agricultural museums 17:670
Agricultural organizations 12:372
Agricultural politics 1:482
Agricultural production 9:307
Agricultural research 1:481
Agricultural revolution 11:648, 11:908
Anthropocene 1:723
Agricultural sciences
and technology 1:471–476
Agricultural sciences and technology 1:479
biotechnologies 1:473–474
botanical and zoological gardens 1:471
climate change impacts on 1:475
division of labor between public and private financing 1:472–473
economic studies 1:472–473
in European nations 1:472
funding 1:473
Green Revolution technologies 1:472–473
historical studies 1:471–472
induced innovation, theory of 1:472
institutional funding 1:473
nanotechnologies 1:474
philosophical studies 1:474–475
plant breeding 1:474
public–private partnerships 1:472–473
robotic technology 1:474
significance of 1:471
sociological studies 1:473–474
standards 1:474
technical innovations 1:471
technology treadmill theory 1:472
technoscience 1:474
tomato production 1:473
Agricultural systems, in archaic states 23:372
Agriculturalists, East African 6:774–775
Agriculture
in Africa 1:293–294
agricultural politics 1:482
in America 20:803
in Andes 23:38
and archaeology of language families 13:337, 13:340–342
and climate change 3:849–850
commercial agriculture 20:805
demand for 1:479–480
in economic development 1:477
and environmental archaeology 7:695–697
farms 1:477–479
and food production 9:301–304
government and 1:480–482
commodity programs 1:480–481
environmental policy 1:482
food policy 1:482
resource policy 1:482
rural development policy 1:481–482
Green Revolution approach 20:804–805
horticulture 11:202, 11:204–205
and land use indicators 13:239–240, 13:242
markets 1:479–480
modernization of 1:458
and poverty 18:730
and production 1:479, 12:329–331
and rural planning 20:802, 20:804
rural sociology 20:814–819
sustainable 20:805–806
in Third World countries 20:802
in twentieth century 23:807–809
economic 23:808
environmental 23:807–808
origins 23:809
social 23:808–809
see also Sustainable agriculture
work anthropology 25:664
Agriculture, economics of 1:477–483
Agri-food knowledge production 10:78
Agroecology 1:472, 1:484–487, 17:665
agroecological practices 1:485–486
background 1:485
changes in agriculture 1:486–487
and irrigation 12:726
as movement 1:486
science of 1:485
in teaching 1:486
Agroforestry 1:486
Agro-industrial enterprises 1:639
Agronomic regimes, and horticulture 11:202–204
Agro-pastoralists 16:840
AGS *see* American Geographical Society (AGS)
AGT *see* Achievement goal theory (AGT)
AH *see* Adaptive hypermedia (AH)
AHCPR *see* Agency for Healthcare Policy and Research (AHCPR)
AHD *see* Authoritarian ROLAuthorized heritage discourse (AHD)
AHEAD *see* Study of Assets and Health Dynamics among the Oldest Old (AHEAD)
Ahistoricism, in Maghreb 21:751
Ahlstrom, Sydney 2:146
Ahmadinejad, Mahmoud 9:657, 16:778–779, 23:48
Ahmedabad Municipal Corporation (AMC) 24:898
AHP *see* Analytical hierarchy process (AHP)
AHRQ *see* Agency for Healthcare Research and Quality (AHRQ)
AHTUS *see* American Heritage Time Use Study (AHTUS)
Ahwahnee principles 16:810, 16:810t
AI *see* Artificial intelligence (AI)
AIC *see* Akaike's information criterion (AIC); Anterior insular cortex (AIC)
AICP *see* American Institute of Certified Planners (AICP)
Aid for Families with Dependent Children program 8:783
Aid in dying 8:282
Aid to Dependent Children (ADC) 25:478
Aid to Families with Dependent Children (AFDC) program 9:15, 18:725, 20:541, 25:478
AIDS *see* Acquired immune deficiency syndrome (AIDS); HIV infections
AIDS Coalition to Unleash Power (ACT UP) 1:494, 2:666, 9:634, 19:748–749
AIDS Emergency Fund 13:887–888
AIDS Memorial Quilt (NAMES Project) 9:634
AIDS/HIV activism 23:327–328
Aid-the-King movement 25:95
AIES *see* All India Educational Survey (AIES)
AIM *see* Active intermodal mapping (AIM)
AIMs *see* Ancestry informative markers (AIMs)
Ainsworth, Mary
contribution to attachment theory 2:792, 12:229–231
disorganized attachment 12:230–231
and Main 12:230–231
Strange Situation Procedure 12:230
Ainsworth-Darnell, James W., on educational outcomes 5:378
Ainu identities 12:794–795
AIP *see* American Independent Party (AIP); Anterior intraparietal (AIP) area
aIPS *see* Anterior intraparietal sulcus (aIPS)
AIR *see* American Institutes for Research (AIR)
Air France flight 447, loss of 11:300
Air of Injustice study (2002) 7:758
Airbnb (social exchange system) 4:142, 8:487
Airline reservation system 24:764–765

Airplane crashes, forensic archaeology for 9:336
Airport
delays, and congestion 24:629, 24:631, 24:633–634
runways, pricing 24:632–633
Airy, George Biddell 21:226
A&IS *see* Area and international studies (A&IS)
Aït Sabbah, Fatna 21:751
works, *La femme dans l'inconscient musulman: Désir et pouvoir* 21:751
Aitken, Alexander Craig 9:626
Aka fathers 9:744
Aka people, childcare of 3:396
Akaike's information criterion (AIC) 15:647–648, 15:653, 15:656–657, 18:606, 23:401–402
Akan 24:648, 24:653–654
Akass, Kim, works, *Sex in the City* 2:219
Akathisia 19:466
Akerlof, George 17:782–783
Akers, Ronald 22:409–410
social learning theory of 6:70, 22:410–412
definitions 22:411
differential association 22:410
differential reinforcement 22:411–412
imitation 22:410–411
works
Social Learning and Social Structure: A General Theory of Crime and Deviance 22:413
Akhbar 11:61
Akhmatova, Anna 20:823
Akrasia 15:892, 25:22–23
AKUF *see* Arbeitsgemeinschaft Kriegsursachenforschung (AKUF)
AL *see* Antennal lobes (AL)
Al Capone approach 13:532
Alamblak, gender assignment in 9:795–796
Alameda County 20:269
Al-Arabiya 16:776
Alarmed discovery 1:353–354
Alaska 1:923
mounds 1:923
village aggregations 1:923
Alaska Native Claims Settlement Act 1:930, 13:236
Alaska Native Knowledge Network 14:187
Alaskan Area Specimen Bank 9:966
ALBA (Bolivarian Alliance for the Peoples of Our America) 23:48
Albanian Church 17:435
Albanian language 11:824, 13:303–305
Albert, Hans 1:495
biography 1:495–496
bridge principles of 1:498
on critical rationalism 1:495–496
criticism of religion 1:499
on ethics 1:498–499
on fallibilism 1:495–496
and Frankfurt School 1:496–497
on national economy and law 1:497–498
on noncognitivism 1:498
on positivist dispute 1:496
problem of objectivity 1:496–497
on relationship between natural sciences and social sciences 1:496–497
ultimate justification problem 1:496
on unity of reason 1:498
on value judgment 1:497
works
Münchhausen Trilemma 1:496, 1:498
Treatise on Critical Reason 1:495–496
Albert, Mathieu, on epistemic cultures 7:877
Albert the Great 20:285
Alchemists 14:12
Alchemy 3:371
Alcmaeon 19:406
Alcohol 19:463, 21:610
among young people 1:513–516
anthropology of drinking 6:658–662
consumption prevalence 1:513–514
dependence 3:766–767
effect on prospective memory 19:275
and myelopathy 1:512
negative consequences 1:514
and normative psychosocial development 1:514–515
trends across age 1:513–514
universal and selective prevention attempts 1:515
Alcohol abuse 1:508, 1:513, 1:517–522, 20:839
binge drinking 1:519–520
brief interventions 1:520–521
drug use 1:517, 1:519, 1:521
harm 1:518–519, 1:521
hazing and 10:557–558
health problems 1:518–519
heavy episodic drinking 1:519–520
recovery 1:521
social problems 1:519
treatment for 1:501
understanding alcohol misuse 1:518
Alcohol consumption
addiction 1:97–98, 1:100
and social networks 16:557–558
Alcohol embryopathy 1:509
Alcohol idiosyncratic reaction 1:510
Alcohol interventions 1:501–506
aftercare 1:503
and Alcoholics Anonymous (AA) 1:503
Center for Alcohol Studies 1:501
cognitive behavioral therapy 1:502
disease model 1:501–502
and HAMS 1:503–504
harm reduction 1:502
motivational enhancement theory 1:502–503
National Epidemiologic Survey on Alcohol and Related Conditions 1:505
project MATCH 1:502
public health model 1:502
and SMART Recovery 1:503
and twelve-step facilitation 1:503
Alcohol misuse
and child neglect 3:413
and coronary heart disease 4:920
Alcohol use 1:517–522, 6:671, 10:583–584
control 6:670
health prevention programs 10:670
and homicide 11:166
and sexual development 14:328, 14:330–331
criminal justice system 1:520
cultural context to consumption 1:517
female drinkers 1:520–521
global patterns of consumption 1:517–518
health and social services and individual behavior change 1:520–521
social work practice 1:521
population level policy responses 1:520
availability 1:520
pricing 1:520
taxation 1:520–521
types of alcoholic beverages 1:518
World Health Organization 1:517, 1:519
Alcoholics Anonymous (AA) 1:503, 8:57, 15:218, 21:380
Alcoholism 1:501–502, 17:236, 17:240, 22:485–486
as disease 1:501–502
and homophobia 15:113
treatment for 1:501–502
Alcoholism Project M 1:501–502
cognitive behavioral theory 1:502
implications of 1:504
motivational enhancement theory 1:502–503
and twelve-step facilitation 1:502–503
Alcohol-related disorders 1:507–512
driving ability and accidents 1:510
effects 1:508–509
behavioral changes 1:508
on body and health 1:509
on brain 1:508
cognitive function and mental processes 1:508
fetal alcohol syndrome 1:509
morbidity and mortality 1:509
physical dependence 1:508
physical withdrawal 1:508–509
tolerance 1:508
genetics 1:507–508
liver cirrhosis 1:509
metabolism and pharmacology 1:507
and nervous system 1:510–512
abuse 1:508
alcohol idiosyncratic reaction 1:510
alcohol intoxication 1:510
alcohol psychosis 1:510–511
alcoholism 1:507–512
alcohol-related myelopathy 1:512
autonomic disorders 1:511
central pontine and extrapontine myelinolysis 1:512
cerebellar atrophy 1:511
cerebral vascular diseases 1:512
corpus callosum atrophy 1:512
delirium 1:510
Marchiafava–Bignami syndrome 1:512
movement disorders 1:512
myopathy 1:511

Alcohol-related disorders (*continued*)
organic brain syndrome, encephalopathy, and dementia 1:511
pathological intoxication 1:510
polyneuropathy 1:511
prevalence 1:509–511
seizures 1:511
sleep disorders 1:512
tobacco–alcohol amblyopy 1:512
psychiatric complications 1:510
treatment programs 3:767
ALCOVE, concept learning 4:655–656, 4:656f
al-Dahik, Abu Musab 24:233
Aleatory explanation model 8:596–597
Aleks (software) 4:8–9
Alemannic dialect 16:68
Aleph *see* Aum Shinrikyo cult
Alexander, Cheryl, on cigarette smoking 16:558
Alexander, Christopher 15:718
Alexander, Jeffrey C. 1:523–528, 5:501
biography 1:523
on civil society 1:526–527
on civil sphere 1:526–527
cultural sociology program 1:525–526
cultural theories of action and materiality by 1:527
on internalization 5:578–579
theories of 1:524–525
works
Bringing Democracy Back In 1:527
Citizen and Enemy as Symbolic Classification 1:526
Civil Religion in America 1:523
The Civil Sphere 1:526
Cultural Pragmatics: Social Performance Between Ritual and Strategy 1:527
Cultural Sociology or Sociology of Culture? Toward a Strong Program 1:526
Culture and Political Crisis: Watergate and Durkheimian Sociology 1:525
The Discourse of American Civil Society 1:526
The Elementary Forms of Religious Life 1:525
The Promise of a Cultural Sociology: Technological Discourse and the Sacred and Profane Information Machine 1:525
Real Civil Societies 1:526–527
Theoretical Logic in Sociology 1:523
Twenty Lectures 1:524
Alexander, Michelle 5:267
Alexander II, Czar 24:238
Alexander III, Pope 15:578
Alexander Technique 24:726
Alexias 6:732, 12:420, 16:685
forms of 6:732–734
in other scripts 6:736
Alexithymia 11:454
and health 17:793
ALF *see* Accelerated long-term forgetting (ALF)
Alfarabi, Abu Nasr 18:487
Alford, Robert R. 17:393
Alford index, class voting 25:299
Alfred P. Sloan Foundation 22:659
Algebra 4:66, 4:71t
children's performance in 4:67t
impact of 17:59
of social networks 22:478–484
Boolean addition 22:478–481
dynamic multiple social network 22:481–483
left dioid 22:483
monoid 22:479–481, 22:483, 22:480t–481t
multiple social network 22:478–481
relational composition 22:478–480, 22:483
semigroup 22:479–481, 22:483
time-ordered composition 22:478, 22:481–483
Algebraic dynamical model 14:797
Algebraic structure, kinship 13:59
Algeria, legal system in 15:362–363
Alghero (Italy), mortality decline in 15:871–872
Algonquian languages 2:705–706, 11:840, 14:221
Algorithmic ancestry estimation 9:888–889
Algorithmic complexity 1:534–538
and cognition 1:536–537
defining intelligence 1:537
language acquisition 1:536
measuring intelligence 1:537
sequential decision making 1:536–537
definitions 1:534–536
results 1:534–536
Algorithmic information theory 1:534
Algorithmic probability 1:535–536
induction 1:536
sequence prediction 1:536
Algorithmic randomness 1:535
Algorithms 1:529–533, 16:118
artificial intelligence 1:530, 1:532
autonomous agents 2:309–310
breadth-first search 22:458, 22:458f
for classification 1:531–533
clique enumeration 22:456, 22:456f
complexity of 1:530–531
computational neuroscience 4:480–482
for computerized test construction 4:511
core numbers 22:455–456, 22:455f
decision 14:320
development of 16:518–519
and Entscheidungsproblem 1:530
evaluation of 22:454
genetic 2:30
graph representations, social network 22:454–455
for learning 1:531, 23:415
maximum structural equivalences 22:457f
Monte Carlo methods 1:531
neural network 1:530–531
NP-completeness 1:532
optimal experimental design 17:256, 17:259, 17:261–262
probability density estimation 19:31
regular equivalences 22:457–458, 22:457f
reasoning 14:319
rule 1:529, 1:531–533
social networks 22:454–460
for solving Markov decision processes 14:574
sort vertices by degree (small to large) 22:455, 22:455f
source dependencies 22:459f
statistical clustering 23:392–393
hierarchical clustering 23:393
object trees 23:393
partitions 23:392–393
triad census 22:456–457, 22:457f
Turing machine 1:530
variations in neural constraint-satisfaction 4:720–721
Ali, Ben 18:424
ALI Test *see* American Law Institute Test (ALI Test)
Alien color effect 14:195–196
Alien hand syndrome 12:421
Alien Tort Claims Act 1789 6:564, 18:502
Alienated labor, and powerlessness 1:547
Alienation 1:539, 1:551, 4:269–270, 14:647, 17:442
and anomie 1:714–716
Arendt on 1:551, 1:554
and class consciousness 3:740, 3:742
concept of 14:656
conceptualizations of 1:541
in critical sociology 1:541
cultural alienation 1:553–554
in drama 1:540
economical alienation 1:552–553
and exchange 8:476–480
Gorz on 10:288
history of 1:539–543
Hochschild on 11:146
implications 1:539–540
in industrial work 25:691
Jaspers on 1:551, 1:554
Kula ring 13:142
and labor 1:552–553
in literature 1:540
Marx on 1:551–555
and modernity 11:865
and money 1:553
Nietzsche on 1:554
philosophical alienation 1:551–552
and private property 1:553
psychosocial tradition 1:544–550
and reification 20:203–205
research 1:541–542
root of 1:539
and shame 21:872–873
Simmel on 1:551, 1:553
and social philosophy 1:540
social psychological consequences 1:541
meaninglessness 1:541
normlessness 1:541
powerlessness 1:541
self-estrangement 1:541
social isolation 1:541
in social sciences 1:540–541
sociological alienation 1:553–554

sociology of 1:551–555
subjective alienation *see* Subjective alienation
Tönnies on 1:551, 1:553
and workplace autonomy 25:718–719
Alienation from world 1:963
Alienation syndrome 1:542
Alien-limb syndrome 5:25
Alignment, in constructive technology assessment 22:875
Alim 11:62
Alimentation (nutrition) 9:312
Alinsky, Saul 4:380–381
on social work 22:135
Alive–dead, degrees of 5:871
Alker, H.R., works, *World Handbook of Political and Social Indicators* 6:267
All in the Family program 14:738
All India Educational Survey (AIES) 5:847
All-affected principle 24:591–593
Allah 12:755–756
strengthening relationship with 12:756–757
ALLBUS *see* Allgemeinen Bevölkerungsumfragen der Sozialwissenschaften (ALLBUS)
AllClear ID Alert Network 5:634
All-day schools 8:613–614
Alleanza Nazionale (AN) 18:411–412
Allegoresis 10:780
Allele–environment interactions 9:829–831
conceptual models 9:830f
Alleles 2:450, 9:933, 11:289–291, 15:210
Allen, G., works, *Physiological Aesthetics* 16:657
Allen, Woody 5:927
works, *Deconstructing Harry* 5:927
Allen calculus 14:320
Alley, R.S., works, *The Producer's Medium* 24:165
All-female interaction 13:257
Allgemeine Staatslehre tradition 18:466, 18:468
Allgemeinen Bevölkerungsumfragen der Sozialwissenschaften (ALLBUS) 5:845, 12:550
Allgemeines Künstlerlexikon (encyclopedia) 2:23
Alliance Christianity 8:287
Alliance for a Green Revolution for Africa (AGRA) 9:309
Alliance for Employment Equity 24:493
Alliance for the Mentally Ill (AMI, Massachusetts) 15:217
Alliance Israélite Universelle 24:602–603
Alliance theory 3:890, 14:600
Alliances
age of 1:561
and balance of power 2:345–348
and conflict 4:625
formation of 1:556–557
inherent good faith models 1:561
international effects of 1:560–561
national effects of 1:561
performance of 1:557–560
political 1:556–561
and psychotherapy 19:517–519
of small states 22:109–110
therapeutic *see* Patient–therapist relationship)
Allied health professionals 16:908
Allied Social Science Associations 20:146
Allies 21:368
LGBT friends as 9:456
Alliteration 13:266–269, 13:276–277
Allocation
of care *see* Health care; rationing
housing 11:241–243
market and nonmarket 14:538–543
market geographies 14:546
of punishment 19:634–635
in schools 21:57
Allocation concealment, in placebo-controlled double-blind studies 18:159–160
Allocation rules 16:116–117
sample-based 16:117–118
Allocative efficiency 1:797, 5:47–48
cost–benefit analysis 5:47–48, 5:50, 5:52–53
Allocentric frame of reference, *versus* body-centered frame of reference 23:119
Allocentric frames 17:704
Allocentric representation 17:704
Allocentric spatial information 11:386–389
Allocentric spatial terms (language) 13:295
Allochronism 11:181
Allocortical entorhinal region 1:595
Allographic agraphia 6:735
Allographic dysgraphia 1:453–454
Allometry, brain evolution 2:830, 2:832–833
Allomorph 15:820
Allomothers 5:510
Alloparents 5:512
All-or-none model for paired-associate learning 6:524–525
Allostasis 11:420
resilience 20:555–556
Allostatic load 23:586
Allouache, Merzak, works, *Bab el-Oued* 21:753
Allport, F.
on collective behavior 4:152–153, 4:155
works
Social Psychology 4:152
Allport, Gordon W. 1:562–565, 3:243, 13:911–912
biography 1:562
conception of personality traits 1:563
consistent and forceful perspective of psychology 1:563
contributions 1:564–565
on eclectic balance 1:562–563
education 1:562
formulation of central explanatory problems 1:563
on prejudice 1:563–564
report on Gestalt developments 1:562
as teacher and mentor 1:564
theory of nature of prejudice 23:462–463
theory of values 25:41–42
works
A Handbook of Social Psychology 7:591
The Nature of Prejudice 1:563–564, 2:256–257, 12:409
All-Russian Nature Protection Society 20:832–833
Alltagsgeschichte 11:42, 11:108
critiques of 11:110–112
Allure, of sport 23:304–305
Alluvium 7:693t
Allyship, transgender 24:553
ALM *see* Audio-lingual method (ALM)
Almeling, Rene, on commodification 5:538
ALMO (software) 23:378
Almond, Gabriel Abraham 2:192, 6:267, 13:803, 23:226
works
The Civic Culture Revisited 2:191
The Politics of the Developing Areas 16:371–372
ALMPs *see* Active labor market policies (ALMPs)
Almshouses 25:262
Alnus glutinosa 7:692
Alper/Narens Theorem 17:299
Alpers, Svetlana 25:207
Alpha rhythm, in electroencephalography 7:356–357
Alphabet
abroad 25:777
Etruscan 25:777
Greek 25:776–778, 25:778t
Roman 25:777
Alphabetic principle, in visual word recognition 25:214, 25:217
Alphabetic script 14:242–243
as a metaphor 14:243
5-Alpha-reductase deficiency 21:760
Alpines 19:801
ALQ *see* Authentic Leadership Questionnaire (ALQ)
Al-Qaeda 24:228, 12:744, 13:528, 16:247–248, 16:807, 20:400, 20:623
associated movements 16:247–248
individuals influenced by, in United States 16:249–250
network 20:403
Al-Qaeda and associated movements (AQAM) 16:247, 16:250
ALs *see* Adolescence-limited offenders (ALs)
Altai, nomadism in 16:838
Altemeyer, Bob 2:263
authoritarian personality 2:257–258
RWA conception 2:263
views on authoritarianism 2:263
Alter Mittelstand (middle classes) 2:786
Alteration, of control beliefs 4:820
Alteration of Sex Description and Sex Status, 2003 (South Africa) 21:744
Altered gene expression 22:485–486

Altered state of consciousness (ASC) 21:868–869
and spirit possession 23:282
Alter-geopolitics 10:94
Alterglobalization movement 1:684–685, 10:186–189
Alterity 16:183
Alternating expectation–conditional maximization (AECM) algorithm 15:627
Alternating gaze 12:28
Alternating treatment design 21:986–987
Alternations 15:827, 15:830–832
Alternative care
child care 20:535
social work 3:474
Alternative certification (AC) programs 23:214
for teachers in United States 24:76
Alternative communities 4:266
Alternative conceptions 21:232, 21:275
Alternative diagnosis 10:563
Alternative dispute resolution (ADR) 6:560–561, 12:934–936, 13:591, 13:609, 15:70–74, 17:399–400
applications 15:72–73
controversies 15:73
distortions and deformations 15:73
evaluation and empirical verification of effectiveness 15:73
future of 15:73–74
process of 15:71
theory and history of 15:71–72
Alternative educational institutions 22:590–591
Alternative energy, and sustainable transportation 23:847
Alternative fit indices 23:618
Alternative food movements (AFMs) 1:566–572
background 1:566
challenges and considerations 1:570–571
community supported agriculture 1:566–567, 1:570
concentrated animal feeding operations 1:568–571
farmers' markets 1:566–567, 1:570–571
food justice 1:570
genetically modified organisms 1:569–570
local food movement 1:566–568
permaculture 1:566–567
plant-based diets 1:568–569
and sustainability 1:568, 1:570–571
Alternative food networks (AFNs) 10:76–78, 10:81
Alternative healing practices *see* Complementary and alternative medicine (CAM)
Alternative hypothesis 11:491, 11:494–495
for significance tests 21:959–961
Alternative masculinities 9:644–645
Alternative models 2:562
Giant three (Extraversion, Neuroticism, and Psychoticism) 2:562
Pavlovian theory-based model of temperament 2:562
population censuses 3:292–293
Zuckerman's alternative five factorial model 2:562
Alternative modernities 15:689, 15:693–694
and de-Westernization 25:548–549
in South America 23:47–48
Alternative optima model 9:830–831
Alternative philosophies 14:256
Alternative planning models 24:617
Alternative reality 11:458–460
Alternative religions 16:805
Alternative schools 21:39–41
background and conception 21:39
developments since the 1960s 21:40
of economic thought 1:573–578
recent trends and perspectives 21:40–41
structural and organizational characteristics 21:39–40
Alternative spirituality 20:360–362
Alternative technology development 21:177
Alternative treatment 12:450–451
Alternatives, and long-term heterosexual relationships 20:912–913, 20:916
Alteroception, neonatal imitation 16:487
Altertumswissenschaft 3:760
Althoff, Friedrich 21:167
Althusser, Louis 5:499, 14:646, 22:723
Altman, Irwin, works, *Small Group Research* 22:103–104
Altria Corporation 3:447
Altruism 1:579–583, 3:358–359, 3:364, 4:520, 5:405, 5:678, 11:260–261, 12:389, 22:892–893, 22:900, 22:909–910, 25:19–20
and adolescent personality 17:866
competitive 8:443–444
definition 1:579
and empathy 19:218
experimental economics 8:550
and friendship 9:425–428, 9:430–431
and Honesty-Humility 11:169–170
impure 1:579
and interests 12:393–394
and intergenerational transfers 8:826–827
and kin selection 13:26–30
and methodological individualism 15:309
and morality 16:606, 16:612
and parenting 19:246–248
problem of 15:811–813
group selection 15:813
kin selection 15:811–812
reciprocal altruism 15:812
sexual selection 15:812–813
and prosocial behavior 19:228–230, 19:235
pure 1:579
Sorokin's work 23:25–26
and value of life 25:19
Altruism theory 12:404
Altruistic activities, religion and 20:265
Altruistic behaviors 7:319, 12:394
Altruistic donations 2:606
Altruistic helping 19:221–222
Altruistic punishment 16:606
Altruistic reciprocity 4:865
cooperation 4:865
Altruistic suicide 1:719, 23:683
Alvares, Santiago, works, *Now* 6:622–623
Alzheimer's Association 10:105
Alzheimer's disease (AD) 6:76–79, 10:500, 10:890, 12:420, 14:363–364, 15:210–211, 15:214–215, 15:441, 16:718, 18:792, 20:516–517, 22:485–486
adaptation in socioecological context 1:585–586
and amnesia 1:645–646
anosognosia 1:587
aphasia 1:585, 1:587
caregiver burden 1:587–589
caregivers care for 3:135–136
caregiving 1:587–588
cognitive deficits 1:585
coping with chronic disease 1:586–589
course of 6:78
cytoskeletal pathology 1:592–593
NFT stages (I–VI) 1:594–595
pretangle stages 1:593
relationships between age and 1:592–593
dementia 1:584–589
and early parental death 6:756
evolution and 1:595
geriatrics 10:104–107
globular forms 1:591–592
intraneuronal tau lesions, developmental sequence of 1:592–595
language processes in 13:362
life span developmental theory 1:585–586
LTP/LTD 14:363–364
and memory binding 15:144
and memory loss 1:584–585, 1:588
mild dementia in old age 6:78
mortality from 1:395
neural basis of 1:591–596
neurodegenerative disorders 1:584–585, 1:587–588
neurodegenerative process in 1:591–592
appearance and composition of neuritic plaques 1:592
bandlike formations 1:591–592
extracellular deposition of abnormal b-amyloid protein 1:591–592
intraneuronal formation of abnormal tau protein 1:591–592, 1:592f
onset of 1:586–587
phenomenology and diagnosis 1:584–585
placebo effect for 18:155
prevalence of 1:595
and prospective memory 19:276
psychosocial effects 1:585–586
disease progression and the necessity of care 1:587–588
at individual and family level 1:586–589
last stage 1:588–589
onset of disease 1:586–587

stage of insecurity and emerging losses 1:586–587
quality of life 1:589
religion and 20:267
source monitoring in 23:34
Alzheimer's disease neuroimaging initiative (ADNI) 14:444
Alzheimer's studies 14:443–444
functional cerebral glucose and activation studies 14:443
molecular imaging studies 14:443–444
structural studies 14:443
AMA *see* American Medical Association (AMA)
Amacrine cells 25:158–159
Amae (Japanese) 11:792
Amalgamation 15:572
ethnogenesis 8:177
Amanita muscaria 16:753–754
Amartya, Sen 23:798
Amateur productions 17:221
Amateur's ignorance 11:579
Amateurs in science 3:631–633, 3:635–636
Amazon 3:47, 23:36, 23:39–40, 23:43–44
art of Marajoara 23:39–40
ceramics 23:40
cities 23:40
conditions for agriculture 23:40
history 23:39–40
natural environment 23:40
in twenty-first century 23:39–40
Amazonian archaeology 23:39–40
Amazonian languages 3:812
Amazonian people, kinship 13:38–39, 13:41
Ambidexterity 12:148
Ambidextrous leadership, I–O psychology 11:883
Ambient vision 23:293–294
Ambiguation 1:641
Ambiguity 20:790
negotiated orders 17:412
and uncertainty 24:717
Ambiguity resolution 21:623–624
Ambiguous Utopias and dystopias 14:257–258
Ambitious teaching 24:76
Ambivalence 12:406, 17:12
attitude 2:172
conflict and 12:402–403
in international policy 20:534–535
prejudice as 18:834–835
in sexism 21:695–696, 21:698
toward aging 1:318–319
toward fashion 8:838
Ambivalence Toward Men Inventory 21:694
Ambivalent emotions 21:680–681
Ambivalent racism 19:865
Ambivalent sexism 18:828–829
and prejudice 18:835
Ambivalent Sexism Inventory 21:697
Ambulatory assessment 22:548
AMC *see* Ahmedabad Municipal Corporation (AMC)
aMCC *see* Anterior midcingulate cortex (aMCC)
AMEF *see* Ansar al-Mujahideen English Forum (AMEF)
Amenorrhea 6:823–824
Amer, Sahar 21:752
American academics 24:763
American Academy of Pediatrics 25:100
American Academy of Physician Assistants 16:911
American Academy of Psychotherapists 20:738
American and French Revolutions 2:671
American Annual Cyclopaedia and Register of Important Events: Embracing Political, Civil, Military, and Social Affairs 7:591
American Anthropological Association (AAA) 1:225, 1:737, 1:740–742, 3:6, 8:898, 11:361, 11:392, 16:856, 19:753–754, 22:661, 25:34
Code of Ethics 1:226–227, 1:736
Committee on Research Problems and Ethics 8:56
Council on Anthropology and Education 7:146
Principles of Professional Responsibility 8:56
Resolution on the Freedom of Publication 8:55–56
Standing Committee on Ethics (COE) 8:56
Statement on Human Rights 1:737
American Anthropologist 7:622, 11:361–362
American anthropology 14:12
American Arbitration Association 12:470
American Association for Geriatric Psychiatry 10:105
American Association for Public Opinion Research (AAPOR) 23:752–753
American Association for the Advancement of Science (AAAS) 1:475, 9:970–971, 22:675
American Association for the Cure of Inebriety 17:240
American Association for the Promotion of Social Science 5:238–239
American Association of Applied Psychology (AAAP) 24:525–526
historical background 3:879–880
American Association of Community Psychiatrists (AACP) 19:590
American Association of Educational Research 4:776
American Association of University Professors (AAUP) 1:28–29
Statement of Principles on Academic Freedom and Tenure 1:28–29
American Association of University Women (AAUW) 9:685
American Bar Association (ABA) 13:620–621
Canons of Professional Ethics 13:622
Center for Professional Responsibility 13:624
law schools accredited by 13:620
Model Rules of Professional Conduct 13:622
American Bar Foundation (ABF) 13:516
Civil Justice Infrastructure Mapping Project 13:595–596
American Bear magazine 13:890–891
American behavioralism 18:469
American Bill of Rights 5:228
American Board of Commissioners for Foreign Missions 15:614–615
American Board of Medical Specialties 14:1
American business unionism 13:201
American civic nationalism 22:889
American civil justice system 16:419
American Civil Liberties Union 19:555–556
versus Myriad Genetics 17:588–590
American civil litigators 16:418
American civil religion 3:680, 18:438–439
definitional debates 3:680–681
diversity and pluralism 3:681–682
global ocntext 3:683
ideological polarization 3:682–683
American civil rights doctrine 5:267
American Civil War 22:444, 24:155
American College 24:762–763
American College & University Presidents' Climate Commitment organization (ACUPCC) 7:774–775
American Commission on Human Rights 11:799
American Community Survey (ACS) 5:695, 10:375
American conservatism 11:433
American Convention on Human Rights 11:799
American Council for Judaism 12:872
American Council for Learned Societies (ACLS) 1:933–934
American Council of Christian Churches 20:397–398
American Council on the Teaching of Foreign Languages (ACTFL) 9:330–331
American critical realism 20:21
American cultural anthropology 12:794
American Customer Satisfaction Index (ACSI) model 5:630–631, 5:631f
American Cyrus Field 24:155
American Declaration of Independence 3:674
American Declaration of the Rights and Duties of Man 11:799
American Declaration on Human Rights (1969) 13:720
American democracy 6:335
American Diabetes Association (ADA) 21:21
American Economic Association (AEA) 7:23, 22:661
American Economic Review 23:459
American Educational Research Association (AERA) 7:280–283
Special Interest Group (SIG) 25:333–334
American Enterprise Association (AEA) 18:323
American Enterprise Institute (AEI) 18:323

American eugenics 8:225–226
American exceptionalism 1:599, 7:664, 12:899–900, 19:127
American Federal Arbitration Act 15:71
American federal corporate income tax 4:946
American federalism 13:818
American Folklore Society 1:756
American Geographical Society (AGS) 1:846–847
American hegemony 13:200
American Heritage Time Use Study (AHTUS) 24:384
American identity 19:834
  black group evaluation 19:834
  black separatism 19:834
  closeness 19:834
American Independent Party (AIP) 18:410–411
American Indian linguistics, work on 20:900
American industrial society 11:860
American Institute of Certified Planners (AICP) 19:821, 24:879
  Code of Ethics and Professional Conduct 18:195
American Institutes for Research (AIR) 7:283
American institutionalism 1:573–574
  evolutionary economics 8:432–433
American Intergovernmental Coordination Act 7:773
American Jewish assimilation 12:877
American Jews
  and Christians, difference between 12:872
  dati, meaning of 12:873
  levels of religious observance among 12:872–873
  religious preferences of 12:873
American Journal of Cultural Sociology 1:526
American Journal of Psychology 25:794
American Journal of Public Health 15:116
American Journal of Sociology 7:560, 19:803, 21:964–965
American kinship terminologies 13:64f
American law 13:819
American Law and Economics Association 7:554
American Law Institute, Restatement of the Law 4:248–249
American Law Institute Test (ALI Test) 9:348
American legal process school 13:793
American legal realism 13:792, 14:276–277
  Idiosyncracy Wing of Realism 14:277
  legal indeterminacy 14:276
  ways of interpreting precedent 14:276
American legal studies 5:266
American linguistics 2:705–706
American Medical Association (AMA) 8:63, 25:100
  codes of ethics 8:65
  The Principles of Medical Ethics 8:53
American model of therapeutic community 22:631, 24:276–277
American National Election Studies (ANES) 5:825, 5:831, 5:833, 7:26–28, 24:659
American Nurses Association 16:910
  nursing profession, definition of 16:909
American Pain Society 17:451–452
American Peace Society 17:644
American Physical Therapy Association 16:908–909
American Planning Association (APA) 19:821
American Planning Association (APA)/American Institute of Certified Planners
  Code of Ethics and Professional Responsibility 1:221
American pluralism 2:86–87, 18:233
American political science, influence on East Asian politics 6:798, 6:800, 6:802
American Political Science Association (APSA) 18:467, 18:470, 21:893, 22:661
American Political Science Review 18:467, 18:470
American politics 4:317, 4:321, 22:31, 22:33–35
  conservatism 1:602–608
  study of 13:11
American pragmatism 14:835–836
  and Veblen 25:54
American Progressive tradition 13:12
American Protestant fundamentalists 20:396–397
American Protestant missionaries 24:602–603
American Protestantism
  case study 20:397–398
  in Japan 12:796
American Psychiatric Association (APA) 8:65, 17:28, 17:949, 21:736, 24:549, 25:625
  codes of ethics 8:65
  see also Diagnostic and Statistical Manual of Mental Disorders (DSM)
American Psychoanalytic Association 25:625
American psychoanalytic influence 3:886
American Psychological Association (APA) 4:776, 5:69–70, 19:416, 20:514–515, 21:736, 24:517, 24:525–526
  Clinical Section foundation 3:879–880
  Ethics Code 20:515, 20:518
  Guidelines for Psychological Practice With Girls and Women 8:929–930, 8:933
  A History of Psychology in Autobiography 6:496
  Standing Committee on Ethical Standards for Psychology 8:55
American Psychological Society (APS) 3:881–882
American Psychologist 15:897
American Public Health Association 15:116, 19:556
American public opinion 13:748
American Recovery and Reinvestment Act 2009 10:372
American Research and Development (ARD) 25:60
American research university 24:762–763
American Revolution 1:597–601, 1:609–614, 4:705, 7:657, 7:661–662, 13:217, 14:18, 20:639–640, 20:895
  Arendt on 1:597, 1:599–600
  Burke 1:598–599
  confederation and constitution 1:612–613
  and freedom 1:597–598
  Furet on 1:600
  historiography 1:613–614
  human rights 1:597, 1:599
  Marx on 1:598–599
  in perspective 1:612
  and religion 1:598–600
  Tocqueville on 1:597–600
  and war 1:611–612
American Samoa 17:161–162
The American Scholar (magazine) 21:884–885
American Sign Language (ASL) 21:928
  linguistic principles of 21:942–943
    deaf culture 21:942
    language 21:942
    role of right hemisphere in 21:945
    sign language acquisition 21:943
    structure of 21:942–943
  spatialized syntax 21:929f
  spatially organized (three-dimensional) morphology in 21:929f
American slave law 12:933–934
American Social Science Association (ASSA) 6:487, 18:321, 22:698, 22:990–991
American socialism 5:38
American socialization 13:747–748
American Society for Ethnohistory (ASE) 8:192–193
American Society for Photogrammetry and Remote Sensing 20:419
American Society of Criminology (ASC) 5:234, 5:239, 8:916
American Society of Human Genetics (ASHG) 9:902
American Sociological Association (ASA) 22:661
  Committee on Professional Ethics (COPE) 8:55
American Sociological Review 2:507, 15:248–249
American sociology 19:677, 22:430
American soldier study 20:73–75
American Standard distribution 2:138
American Statistical Association 20:877
American Stop Smoking Intervention Study (ASSIST) 6:878
American structural–functional anthropology 22:722–723
American studies
  of education 7:299
    crosscultural study 7:299
  of society 1:615–620
    nineteenth-century 1:617–618
    scholars 1:615, 1:619–620
    seventeenth- and eighteenth-century 1:616–617
    students 1:615

terms and periods 1:615–616
twentieth-century 1:618–619
American Telephone and Telegraph *see* AT&T
American Time Use Survey (ATUS) 5:746, 24:383–384
American Uniform Commercial Code 4:810
American universities 22:593
American War of Independence 12:221
American Way of Life 2:732
American Wood-Paper Patent case 17:585
American Wood-Paper Patent case 17:585
Americaness of political ideas 21:893
Americanization 4:754, 10:228, 14:708, 19:126
of anomie 1:715–716
criminology and 5:240, 5:242
of the governing regimes 12:898
immigrants 1:218
Americans for Medical Progress Educational Foundation 1:704
Americans with Disabilities Act (ADA) 1:256, 6:446, 6:531, 8:46, 11:882, 13:544, 24:550–551
Americans with Disabilities Amendments Act 2008 6:532
American-style critical theories 22:962
Americas
adult mortality in 1:171, 1:175, 1:179–180
anthropology at home in 1:755
Caribbean *see* Caribbean
civil religion in 2:508
cultural critique in 5:383–384
early foundations for archaeological survey in 21:654
impact of Cold War in 4:122–123
labor leadership 15:535
origins of food production in 9:302–303
Central and South America 9:303
Eastern North America 9:303
Mexico 9:303
southwestern United States 9:303–305
public mood 2:195
religion in 2:508
school of anthropology 5:484
sociocultural overviews
Mexico 1:626–632
urban poverty in 24:915
AmericasBarometer 5:849
Amerindian myths 13:906–907
Ames, H.B., works, *The City Below the Hill* 22:601
AMH *see* Adaptive market hypothesis (AMH)
Amhara National Democratic Movement (ANDM) 3:440
Amherst seminar 13:518
AMI *see* Autobiographical Memory Interview (AMI)
Amici curiae 1:824
Amin, Samir
works
Accumulation on a World Scale: A Critique of the Theory of Underdevelopment 16:371–372
The Arab Nation 16:383
Amineptine 1:779
Amino acid changes 13:25
Amish 2:147
culture 20:269, 20:396
English 18:137
Amitriptyline 1:779, 22:85
Ammon, Otto 5:689–690
on social Darwinism 5:689–690
Amnesia 1:644–650, 4:560–562, 4:564, 11:709
artificial grammar learning in patients with 13:85
childhood 6:761
and memory development 15:147, 15:151–153
dysexecutive amnesia 6:718–724
forensic psychology 9:352
hippocampual *see* Hippocampual amnesia
and medial temporal lobe memory system 5:923
posttraumatic 18:636–639
psychogenic 1:651–658
serial order memory 21:639–640
source monitoring 23:28, 23:30–31, 23:33
transient 1:651–658
transient epileptic amnesia *see* Transient epileptic amnesia (TEA)
transient global amnesia *see* Transient global amnesia (TGA)
Amnesic disorders 1:651
behavioral characteristics, general forms 1:657t
due to prevailing psychological mechanisms 1:651–652
memory impairment in 1:651
Amnesic syndrome 24:562–563
Amnestic prosopagnosia 19:251–254
Amnesty International 1:579–580, 3:677
Amniocentesis 9:994–995
Amobarbital, intracarotid sodium amytal testing 12:689–695
Amodal semantics 21:574–577
Amoral familism 12:193, 22:177
AMOS (psychometric test) 3:787–788
Amotivation 21:488
Amour de soi 20:769
Amour propre 20:769
AMP *see* Affect misattribution procedure (AMP)
AMPA receptors 12:711–713
synaptic delivery disruption
glucocorticoids 16:490–491
whisker experience-dependent 16:490
Ampere, Andre Marie 24:153
Amphetamines 6:674, 6:685
Amplification
of interoception
explanatory style 8:598–599
of minimal effects 17:881
Amplitude gain 24:334
AMPM *see* Automated Multiple-Pass Method (AMPM)
AMS *see* Accelerator mass spectrometry (AMS); Additional Member System (AMS)
AMSs *see* Automatic Milking Systems (AMSs)
Amsterdam Stock Exchange Association 22:979–980
Amsterdam Treaty of 1997 7:771
Amsterdam–Nijmegen Everyday Language Test 3:794
Amtscharisma 3:355
AMY1 (salivary amylase gene) copy number 9:815
Amygdala 1:653–654, 8:870, 8:872, 10:890, 11:613–614, 14:125, 16:494, 16:603, 16:640, 16:739, 18:39–40, 22:485
and anxiety 1:804–807
and anxiety-like behaviors 16:496
and dreaming 6:653–655
and emotion 7:479
and fear conditioning 1:811–815, 8:849–850, 8:871–872
and fear-potentiated startle 8:861–865
functional magnetic resonance imaging (fMRI) of 16:494
Amygdaloidotomy 23:18t
Amygdalotomy 23:18
Amyloid precursor protein (APP) 1:591–592
Amyotrophic lateral sclerosis 12:420
AN *see* Alleanza Nazionale (AN)
Anabaptists 2:145
Anabolism, at work 17:133–134
Anal intercourse and HIV transmission 15:835
Anal sex, unprotected 21:798
Análise Psicológica 13:902
Analog to digital (AD) conversion 7:353–354
Analogical reasoning 21:292
Analogs 25:171
in EITMs 18:455, 18:457–458
error-in-variables regression 18:458
expectations and uncertainty 18:457–458
formal and applied statistical analogs 18:455–456
linking formal and applied statistical analogs 18:457
relating behavioral analogs to empirical analog for measurement error 18:458
tools and components for behavioral analog 18:457–458
unification and evaluation 18:456
human–environment relationships 11:407–408
Analogy
legal reasoning 13:776–783
and metaphors 15:289
totemism 24:453, 24:457
uses of equilibrium for 7:913–914
Analyse des Données 5:9
Analysis, design, development, implantation, and evaluation (ADDIE) model 12:198
Analysis of covariance (ANCOVA) 8:190, 19:140–141

Analysis of covariance (ANCOVA) (*continued*)
-based approaches, observed variables 4:840–841
control variables 4:840–841
nonequivalent group designs 16:852–853
Analysis of variance (ANOVA) 1:659–664, 5:357, 13:690, 14:812, 16:86, 16:134, 20:248, 20:250, 22:548–549
higher-order 1:661–663
for moderation 15:673
one-way 1:659–660
Pearson model 17:661
sampling 20:878
two-way 1:660–661
Analysis phase, ADDIE 12:198
Analysis scale 21:1–2
Analytic category, gender as 9:784
Analytic epidemiology 15:179
Analytic generalization 3:199
Analytic induction 23:857
social work-specific research 20:504–506
Analytic intelligence 12:317
Analytic issues in testing moderation 15:672–673
Analytic memos, in qualitative research 19:685
Analytic philosophy 1:665–667, 20:761–762, 20:765
of education 7:259
of language 10:786–787
logical empiricism/logical positivism 14:314, 14:317
and pragmatism 18:810
and wisdom 25:601
Analytic pragmatism 18:806
Analytic statement 14:317
Analytical capacity, required for health information systems 5:786
Analytical concept
of equilibrium 7:914–915
Third World as 24:300–301
Analytical conflict theories 22:956
Analytical hierarchy process (AHP) 23:138
Analytical Marxism 1:665–667
Elster on 7:402
Analytical perspectives of phenomenology 20:344–346
formal phenomenology 20:345–346
hermeneutics 20:345
structuralist phenomenology 20:345
Analytical philosophy 10:773
Taylor on 24:40–41
Analytical primacy 1:75
Analytical psychology, role of Jung in 12:904–905
Analytical rigor 23:459
Analytical sociology 1:668–673, 14:417–419, 14:875–877
agent-based modeling 1:672
analytical realism 1:668–669
computer simulation 1:672
Elster on 7:402
mechanism-based explanation 1:668–672
rational choice theory 1:669, 1:671–672
self-fulfilling prophesy 1:669–671
social explanada and individual action 1:671–672
social mechanisms 1:669–670
and social mechanisms 22:415, 22:419
structural individualism 1:671
theories of middle range 1:670–671
theory of action 1:672
Analytical techniques
and public policy schools 19:577–578
transference 24:546–547
Analytical theory of probability 13:382–383
Analytic-deliberative process, for environmental risks 7:786, 7:788–789
Analytics, definition of 24:107
Ananda Marga (new religious movement) 16:806
Anansesem 24:653
Anaphora 1:674–679
binding theory 1:676
cataphora 1:674
defined 1:674–675
endophora 1:674
exophora 1:674
grammatical *versus* pragmatic approaches 1:676–679
logophoricity 1:678
minimalist analysis 1:676
neo-Gricean pragmatic theory 1:677–679
reference-tracking system 1:674–676
reflexivity analysis 1:677
typologies 1:675–676
Anaphoric account 14:192
Anaphoric reference 6:53–54
Anarchism 12:273–274, 22:834–835, 24:597
Chomsky on 3:526
ethnography of 1:683–686
alterglobalization movement 1:684–685
authority 1:680–683
capitalism 1:680–681, 1:685
development 1:682–683, 1:686
direct action 1:684–686
direct democracy 1:681, 1:685–686
existing anarchism 1:686
global uprisings 1:685–686
hierarchy 1:682–684
Marxism 1:680–681, 1:683–684
mutual aid 1:680–682
neoliberalism 1:684
primitives 1:681–682
social movement 1:680, 1:683–684, 1:686
Spanish anarchism 1:683–684
utopia 1:681
queer *see* Queer anarchism
revolutionary tradition 1:680–681
classical anarchism 1:680–681
Spain and violent demise of classical anarchism 1:681
as theory 1:681–683
critical theory 1:683
prospecting in ethnographic record 1:683
stateless societies and 1:681–683
Anarchitecture 18:189
Anarchy, and realism 20:26–29
Anatomical asymmetries 13:398
and language functions 13:403–404
Anatomical basis, of neonatal crying 16:476–477
Anatomical connectionism 18:45, 18:55
Anatomical differences, anomia 19:149–151
Anatomical locus of selective attention 2:163–164
Anatomically modern humans 11:293
Anatomy
brain stimulation reward 2:842–843
of temporal lobe 24:206, 24:206f
Anaxagoras 20:284
Anaximander 20:284
Anaya, James 11:771–772
ANC *see* African National Congress (ANC)
Ancestor worship, East Asian studies 6:786–787
Ancestora, clan 3:731
Ancestors
anthropology of 1:688–690
beliefs 1:688–689
death/dying 1:688–689, 5:865
kinship 1:688–689
and religion 1:688–689
Ancestral Hominin behavior, early primate models of 18:891–892
Ancestral language 4:423–426
Ancestral State Reconstruction and Phylogenetic Prediction 4:417–419
dinosaur genome size 4:417–418
feeding time and food processing evolution 4:418–419
Ancestry 9:898–899
biogeographical 9:899
deep ancestry *versus* near ancestry 9:898–899
estimation 9:902
and ethnic intermarriage 8:118–119, 8:121, 8:118t
and genealogy 1:761–762, 9:806–810
Genographic Project 10:22, 10:24–25, 10:28
and genomics 10:37
and new genetics 9:1018
Ancestry informative markers (AIMs) 9:899–900
Ancestry painting 9:889
AncestryDNA 9:988
Anchor companies 22:172
Anchored instruction 12:199
Anchored poverty 18:756–757
Anchoring heuristic 10:826–827
of decision making 5:920
Anchoring homosexuality, definition 16:862
Anchoring social representation 4:198, 20:447
Ancien régime 19:613
Ancien Regime in Europe 17:516
Ancient DNA 2:602
Ancient economy, guilds 2:48

Ancient Egypt 1:691
archaeology
development of 1:691–692
faience, blue frits, and glass 1:694–695
funerary 1:692–693
material culture study 1:693–695
metal 1:695
pottery 1:694
settlement 1:693
stone 1:694
chronology 1:691, 1:691t
geography 1:691
Ancient geography 16:334
Ancient Greece
Athens 1:698
city-state 1:697–698
early Greece 1:697
fifth and fourth centuries BC 1:698–699
Hellenistic Age 1:699
history of 1:697–702
and individual-collective relationship 11:816
peacemaking in 17:638
psychopathology theories 3:877
and Rome 1:699–700
Sparta 1:698
Ancient Greeks 13:799
Ancient human migrations, detecting 9:892
Ancient India 11:78–79
European scholarship 11:78
Indian contribution 11:78
Indology after World War II 11:78–79
Orientalism (Said) 11:78
traditions 11:78
Ancient Judaism 19:197
Ancient Rome, peacemaking in 17:638
Ancillary statistics 23:420–422
ANCOVA *see* Analysis of covariance (ANCOVA)
AND gates 16:540–541
Andean archaeology 23:37–39
Andean divinities 3:535
Andersen–Gill (Poisson) model 8:297
Anderson, Benedict 7:147, 22:616–617
works
The Chrysanthemum and the Sword 7:145
Imagined Communities: Reflections on the Origins and Spread of Nationalism 2:871–872, 11:86
Anderson, Chris, works, *The Long Tail* 12:588
Anderson, Elijah 6:72
Anderson, Elizabeth Garrett 25:623
Anderson, Hans Christian 9:637–638
Anderson, John Robert, works, *Human Associative Memory* 4:105
Anderson, Nels, works, *The Hobo* 7:561
Anderson, Perry 1:752, 9:672–673
Anderson, Robert 24:144
Anderson–Darling test 23:398, 23:403–404
Anderson-Levitt, K., on educational anthropology 7:146–147
Andes 23:36–39, 23:43–48
agriculture 23:37–38
architectural features 23:39
ceramics 23:38
civilizations 23:37
domestication of animal and plants 23:37–38
earthworks 23:38
empire 23:36
formative societies 23:38
history 23:38
Inca state 23:39
late period 23:39
marine resources 23:37
Middle period 23:38–39
production activities 23:38
societies 23:37
state societies 23:38
urbanism 23:38–39
ANDM *see* Amhara National Democratic Movement (ANDM)
Andragogy, lifelong learning and 1:148–149
Andreasen, Nancy 25:625
Andreas-Salomé, Lou 21:740
Andrew W. Mellon Foundation 22:659
Androcentric archaeology 9:765–766
Androcentrism 23:326–327
Androgen, prenatal 21:774–776
Androgynous gender expression 13:870
Androgyny 25:630
meanings of 8:944
Anecdotal evidence 10:454
ANES *see* American National Election Studies (ANES)
Angelbeck, Bill, on anarchism 1:682–683
Angelic Reiki 10:562
Angell, James 19:413
Anger 4:145, 4:148–149, 7:515, 7:534–535
and behavior during infancy and early childhood 6:12–13
and cardiac disease 20:188
and coronary heart disease 2:420
and disgust 6:547–548
due to shame 21:875
and health 17:793
health risks of 7:497–498
in intergroup relations 7:502–503
Anger management 20:436
Angiopathy, and diabetes 6:337
Anglican Methodism 10:492
Anglicanism 12:762
Anglo conformity 11:637
Anglo faith traditions 3:541
Anglo-America 3:540
Anglo-American
adoption *see* Adoption
adversary tradition 1:823
analytical philosophy 1:665
common law world 13:585–587
concepts of personality disorder 17:892
corporate governance 4:935–936
criminal justice practitioners 18:940
debate, sub-Saharan Africa 11:90
law, in Latin America 13:412–413
law firms model, globalization of 13:579–580
legal philosophy 13:551
legal scholarship 5:254
media traditions 3:612
news journalism 16:821
relations 1:610–611
succession system 19:174
Anglocentricism 9:201
legal order 23:59
Anglo-Hindu law 23:59–60
Anglophone countries 11:741
Anglo-Saxon approaches 12:371
Anglo-Saxon eugenics 8:220
Anglo-Saxon 'liberal' welfare states 25:481
Anglo-Saxon patterns 25:688
Ang-marking 23:890
Angola, sociocultural aspects 23:77–79
Angular gyrus 13:371
Anh, Dao Duy, works, *Short History of Vietnamese Culture* 11:85
Animacy hierarchy 14:210
Animal behavior 25:449
example for Markov decision processes 14:573
Animal Behaviour 14:369
Animal breeding
and genealogy 9:807
Wright's contribution to 25:758–759
Animal laboratory, reinforcement and 20:208–209
Animal magnetism, Mesmer's theories 19:502
Animal models
of anxiety 1:808
conditioned emotional responses 1:808
ethological models 1:808
of human amnesia 5:923
of memory and learning 9:905–907
of stuttering 9:269–272
birds' and humans' errors 9:271–272
'rule to repeat' in birdsong 9:271
Zebra finches 9:270–271
Animal phobia, and fear-potentiated startle 8:864
Animal reminder disgust 6:547
Animal rights
versus animal welfare 1:703
concept of 1:703
extremists, in United States 16:249
history 1:703–704
in research 1:703–707
animal rights movements 1:705
attitudes 1:704, 1:706
contemporary protection activities 1:704–705
coverage of species 1:706
guidelines for use of animals 1:705
legal protections 1:703
moral issues 1:703
policy 1:704–705
sentience 1:705
Animal slaughter law 5:495
Animal social learning 22:401–402
adaptive value of 22:405–406
affordance learning 22:403
conservation and animal welfare, implications for 22:406
context-dependent biases 22:406
cultural intelligence hypothesis 22:406

Animal social learning (*continued*)
history of research into 22:402
imitation 22:402–404, 22:406
innovation 22:405
laboratory studies of 22:402–404
primates 22:402, 22:405
and social intelligence 22:406
studies of animal culture 22:404–405
transmission chain studies 22:404–405
Animal studies 5:931–932
Animal welfare 1:703, 20:799
Animal Welfare Act 1:704, 1:706
Animal–environment reciprocity 10:128
Animalization 6:31
Animals
brain/behavioral lateralization in 2:799–805
cognition of 13:125, 14:367–368
comparative evolutionary psychology 4:395–397
domestication 17:578
ecology 22:264
ethics for 2:617
ethology 17:532
geographies 20:798–799
hypnosis 11:465, 11:468
needs, and herders 17:579
problem solving 13:123–125
role of cerebellum in eyeblink conditioning 2:93–95, 2:93f–94f
sacrifice 17:577–578
virtual spatial navigation in 16:353–354
Animals Australia 1:705
Animate objects 21:269
Animation 3:165
as dynamic maps 6:716
Animé 5:457
Animism
in Arctic 1:929–930
and fetishism 9:106–107
hunting/gathering societies 11:430
Anisotropic scaling 19:80
Anisotropy 23:192
Ankersmit, F.R. 11:15
works, *Narrative Logic* 11:15
Anna O *see* Breuer, Josef
Annales 2:704, 25:341
writing of history 11:40–41
sciences of man 11:40
Annales d'histoire économique et sociale 4:765, 11:40
Annales School 1:708–713, 5:420–421, 11:42, 11:76
and Bloch 1:708–712
and Braudel 1:708–712
and Burke 1:712, 2:924, 2:927–928
cliometrics 1:708–710
deep structures of history 1:708–710
and Febvre 1:708–712
geohistory 1:710–711
historical change 1:708–709
historical positivism 1:709
historiography 1:708–709
human agency 1:708–709, 1:711–712
labor history 13:170
social sciences in 11:76
Annals of Quauhtitlan 11:57
Annals of Regional Science 20:147
Annals of Statistics 24:345
Annals of the American Academy of Political and Social Sciences 20:421
Annals of the Association of American Geographers 20:421
Annexation
minority-group status 15:570
policies 10:449
Annotation (linguistics) 4:951–952, 5:812
types, of corpora 5:813–814
Annotation corpus 4:951
Annual cash income 18:742
Annual conferences 22:662, 22:664–665
Annual Review of Ecology and Systematics 22:899
Annual Review of Psychology 2:349
Anomalies 25:532
Anomalous commemoration 4:238
Anomia 1:818, 1:820, 4:93, 18:57, 19:147–152
anatomical differences 19:149–151
lexical level 19:147, 19:150
neuropsychological dissociations 19:148–149
problem of difficulty 19:149
proper names in clinical practice 19:151
prosopanomia 19:148
pure reference 19:147–148
semantic level 19:147–149
systems 19:151
temporal lobe 19:150–151
token reference 19:147–148
type reference 19:147
Anomia scales 1:715–716
Anomic aphasia 19:148, 1:820t
Anomic suicide 1:719, 23:683
Anomie 1:540–541, 1:714–717, 6:701–702, 13:571, 15:247–249
and alienation 1:714–716
Americanization of 1:715–716
Bell on 2:503
and delinquency 6:69–70
extensions or reformulations in 6:71–72
Durkheim's concept 1:714–715
etymology 1:718
history of 1:718–721
in classical sociology 1:719
contemporary travels and adventures 1:721
and deviant behavior 1:719–721
before sociology 1:718–719
institutional anomie 1:716
recent departures 1:716
and social control 22:238
and suicide 1:714–716
use of term 1:718
Anomie theory 5:191–192
of deviance 5:205
Anonymity 14:446, 21:814–815, 23:736, 23:739
and aggression 1:378–379
Anonymous group (digital activism) 22:514
Anonymous helping 19:221–222
Anopheles mosquitoes 7:815, 18:175
Anorexia nervosa 6:822–825
psychotherapy for 19:376
Anorgasmia 21:714, 21:718
Anosmia 17:199–200, 17:199t
Anosognosia 1:587, 10:500, 17:507
A–Not-A task 21:946
A-not-B task 8:491–492, 8:491f
ANOVA *see* Analysis of variance (ANOVA)
Anovlar (hormonal contraceptive) 18:536
ANPO *see* United States–Japan Security Treaty (ANPO)
ANS *see* Approximate number system (ANS); Autonomic nervous system (ANS)
Ansar al-Mujahideen English Forum (AMEF) 24:231
Anselm 19:408
Answer set programming (ASP) 5:937
Answerability 20:589
Answering and resisting 16:817
ANT *see* Actor-network theory (ANT)
Antagonism 1:447–448, 11:575, 16:460
Antagonistic acculturation 6:403–404
Antagonistic classes 3:746
Antagonistic forces 8:837
Antagonistic pleiotropy 8:451, 17:902
Antagonists 22:486–487
Antebellum nativism 25:796–797
Antebellum Southern slave plantation 22:55
Antecedents 5:312, 19:368
of cultural intelligence 5:435–436
of knowledge society 13:107
reinforcement 20:208
Antenarratives 16:195
Antennal lobes (AL) 15:156
Anterior capsulotomy 23:18t
Anterior cingulate cortex (ACC) 10:889, 16:601–602, 16:644, 16:701, 18:817, 23:866, 16:602f
interoception 12:614–615
Anterior cingulotomy 23:18t
Anterior commissure 12:416
Anterior communicating artery aneurysm, rupture of 6:722
Anterior dorsal thalamus 16:626
Anterior insula 16:603, 16:603f
interoception 12:614–615
Anterior insular cortex (AIC), and empathy 7:543
Anterior interpositus, eyeblink conditioning 8:636–637, 8:639
Anterior intraparietal (AIP) area 5:23
Anterior intraparietal sulcus (aIPS) 5:26
Anterior midcingulate cortex (aMCC), and empathy 7:543
Anterior pituitary
endocrine cell types of 11:484f
hormones of 11:485t
Anterior temporal lobe, and semantic memory 21:574, 21:576–577
Anterior thalamic peduncle 13:679–680
Anterior-posterior partitioning, central nervous system 2:118

Anterograde amnesia 1:644, 1:651, 1:653, 10:876–879, 10:881, 24:562, 24:565–568
Anterograde memory 1:441, 1:644–645
Anterolateral temporal cortical regions 13:681
Anthony, Susan B. 22:444
Anthropic principle 20:288
Anthropocene, and postcolonial geography 18:645–646
Anthropocene epoch 1:722–727
  case studies 1:724–727
    atmosphere 1:724–725
    biodiversity and land-use change 1:726–727
    ocean acidification 1:725–726
  domestication 1:723
  eutrophication 1:724
  fire 1:722–723
  global warming 1:726–727
  Great Acceleration 1:724
  pre-Anthropocene impact/early anthropogenic hypothesis 1:722–723
    agricultural revolution 1:723
    prehistory 1:722–723
  stage 1 1:723–724
  stage 2 1:724
Anthropocentrism 8:168, 9:567–568, 18:194, 18:196
  comparative evolutionary psychology 4:395, 4:397
Anthropogenic climate 7:817
  change in 3:854
Anthropogenic global warming 23:833
Anthropogenic hazards 6:481
Anthropogeographie 7:723, 23:199
Anthropoid primates 15:966–967
Anthropolo 3:722–723
Anthropological capacity 14:834
Anthropological development, social-anthropological aspects 6:301–306
Anthropological ethnography, and morality 15:807
Anthropological jurisprudence, of cultural rights 5:494–496
Anthropological linguistics *see* Linguistic anthropology (LA)
Anthropological method, and apprenticeship 1:865, 1:867
Anthropological structuralism, and advertising 1:216
Anthropological theory, human rights 11:360, 11:365
Anthropological writing 1:740–745
  applied anthropology 1:742
  and emotion 1:741–742
    and anthropological writing 1:741–742
  ethnography 1:740–743
  journal article 1:740
  monograph 1:740, 1:742
  myopia of monopoly 1:740
  new genres 1:742–743
  new liveliness 1:741–742
  new readerships 1:740–741
  popular writing *versus* scholarly writing 1:741–742
  public engagement 1:740–742
  social justice 1:740
  text 1:741–743
Anthropologists
  as advocates 1:223
  as expert witnesses 1:225–226
Anthropology 1:758–759, 1:762, 1:772–777, 8:199–200, 9:671–672, 10:738, 17:179, 18:83, 19:313
  Addams on 1:94–95
  advocacy in 1:223
  of affect *see* Emotion
  and age 1:318–322
  of Amazon 23:36
  ancestors of 1:688–690
  Andes on 23:36
  of apprenticeship 1:865–869
  of art 2:15–20
  aspects of psychiatry 19:313–319
  of authenticity 2:251–254
  of bases 15:486–487
  of belief 2:493–497
  of big man 2:567–573
  of birth 2:697–699
  of borders 2:760–764
  of boundaries 1:772–773
  Bourdieu on 2:777–778
  British schools 1:767
  of brokers and brokerage *see* Brokers/brokerage
  of business *see* Business anthropology
  of cannibalism 3:85–86
    theory 3:86–88
  of capitalism 18:365
  of care 25:123–124, 25:126
  of cargo cults 3:139–143
  of Caribbean 1:639–640
  of Central Asia
  see also Central Asia
  of childhood 3:451
  of China 2:65
  of citizenship 3:637–642
  of classical legacies 3:775–776
  cognitive *see* Cognitive anthropology
  cold war agendas for 25:122–123
  of colonialism 4:218–222
  of common sense 4:1–5
  of communes 4:266–272
  community studies 4:369–371
  comparative anthropology 1:774–775, 2:567, 2:569–571
  comparative method in *see* Comparative method in anthropology
  of conflict
  see also Conflict
  of consciousness 7:147
  of contemporary 1:749
  of the contemporary 1:749
  core of 1:773–775
  cosmopolitan, Kant's 5:42–43
  counterinsurgencies 5:99–102
  of craft production 5:109–110
    learning and apprenticeship 5:110
    Marxist studies 5:109
    material culture and folklore 5:109
    rebirth of material culture studies 5:109–110
    social, ethnographies in 5:109
  and creolization 5:118–119
  cultural 4:824–825
    syncretism 23:882
  cultural, ethics in
  see also Ethics
  cultural critique 5:382
  of cultural heritage 5:414–418
  and cultural history 5:420, 5:422–423
  of cultural relativism 5:484–487
  and culture 5:617–618, 5:620
  culture in, African Studies 1:278
  and culture of poverty 5:590–594
  current primate studies in 18:892–894
    cognition 18:894
    functional morphology 18:893–894
    intelligence 18:894
    social behavior 18:892–893
  of cyberculture 12:600
  of dance 5:666–670
  of death 5:870–873
  definition of 1:68, 6:620
  and development 1:775–776
  of diffusion 6:402–405
  of discourse 6:503–509
  of drinking 6:658–662
  economic *see* Economic anthropology
  of education 1:728–733, 7:144–148
    agency 7:147–148
    assumptions 7:147
    children 7:144–148
    contributions 7:148
    ethnography 7:144–148
    fieldwork 7:147–148
    historical overview 7:144–147
    in Japan 7:145–146
    learning 7:144–147
    methodology 7:147–148
    psychology 7:145, 7:147
    schools 7:145, 7:148
    scope of study 7:144
    socialization 7:144–148
    structure 7:147
    teaching 7:144, 7:146–147
  of elites 7:386–389
  of emotions *see* Affect
  and end of colonialism 1:768
  of energy
  see also Energy
  ethical questions in 8:55–56
  of ethnicity 8:102–104, 8:148–151
  ethnicity in 1:769–770
  ethnogenesis in 8:175–177
  ethnography *see* Ethnography
  and ethnohistory 1:746–747
  ethnology as formative period of 8:198–199
  of everyday life 8:323–326
  exchange in 8:476–481
  of experience, and liminality 14:136

Anthropology (*continued*)
explanations of ethical diversity 6:589–590
of family 8:809–814
history of 8:738–744
feminism in
see also Feminism
of fertility and culture 9:1–5
and fertility transition 9:52
field observational research in 9:135–139
and fieldwork 1:773–776, 9:149–152
of finance 9:171–176
food and cuisines in 9:290–295
foundations 1:765–768
four-field approach
in American 1:766
in French 1:768
Freyre on 9:421
of friendship 9:425–431
functionalism in 9:532
future of 1:776–777
future primate studies 18:894–895
Geertz on 9:661–665
of gender
see also Gender
genealogy in 9:806–812
generation in 9:853–856
genetics in 9:939–944
adaptation 9:940
human variation 9:939–940
population genetics 9:940–941
of global south
see also Global south
and globalization 1:776, 10:233–238
and governmentality 10:308–311
of health 10:571–576
of hegemony 10:747–749
and hermeneutics 1:747
historical consciousness 1:748
and history 1:746–750, 4:181–182
ethnohistory 1:746–747
hermeneutics 1:747
natural history 1:746
history of 1:765–771
at home 1:773–774
home in 11:155–158
honor in 11:181–182
of horticulture 11:202, 11:206–207
households in 11:228–232
of human behavior 8:399–405
and human rights 11:377
of human rights 11:360–366
Humboldt, A. on 25:279, 25:284
Humboldt, W. on 25:287–288
of humor 11:416–417
of hunter-gatherer societies 11:424
of hunting and gathering societies 11:428–431
identity in 11:532–536
of industrial societies
see also Industrial societies
of intercultural communication *see* Intercultural communication
interpretation in 12:641–644
of Japan 12:794
kinship in 13:36–43
of knowledge 1:865–869, 13:70–73, 20:227–228
advanced societies 13:72
classification 13:70–72
cognitive style 13:70, 13:72
communication 13:70, 13:73
cosmology 13:70–72
education 13:72–73
elementary societies 13:70–73
ethnography 13:71
human development 13:70
intellectuals 13:72
knowledge and power 13:72–73
late nineteenth and early twentieth centuries 13:70–71
modernity 13:72
psychology and brain science 13:71–72
tradition 13:70–73
of Korea 2:71–72
of Kula ring 13:139–143
and labor history 13:170–171
of law 13:535–541
Lévi-Strauss on 13:904–910
of liberalism *see* Liberalism
of mafia 14:422–428
of magic *see* Magic
Malinowski's contributions to 14:453–454
maritime 9:228–229
Maroons in 14:591–595
media audiences 2:216
medical 2:496
and medicalization 15:106
of medicine 15:121
of Melanesia 17:154
metaphors and metonymy 15:290–291
methods 9:941–943
biotechnological breakthroughs 9:942
chromosome mapping 9:942
mitochondrial DNA 9:942–943
repetitive DNA 9:943
stains 9:942
Y chromosomes 9:943
of migration
see also Migration
of militaries and militarization 15:483
modernity in 15:687
of money 15:746–749
of morality 1:737–738
of multiculturalism 16:28–33
of Native Americans 1:636–637
and natural history 1:746
of nongovernmental organizations 16:856–860
of nostalgia 17:34
of organizational culture 17:346–351
pastoralism in 17:577–583
of peace and nonviolence 17:624–630
as peace founder 17:624
peasants in 17:676
performance in 17:708–715
of personhood 17:932–935
perspectives in 1:769
physical 19:704
of play 18:214–217
of political ecology
see also Political ecology
political economy in 18:362–366
of politics 1:768, 18:335–341
positivism 1:748
postcolonial science studies 18:653
and postcolonialism 1:770
potlatch in 18:714–718
practical uses of 1:775–776
of property *see* Property
psychological 1:768, 19:359–365
public 1:225–226
queer
see also Queer anthropology
reflexivity in 20:88–92
refugees in 20:105–110
of religion 2:496, 20:385–386, 20:385f
rethinking 2:569–570
of rumors 20:790–791
of science and technology
see also Science
of secrecy 21:374–379
of senses
see also Senses
social Darwinism 5:689
of social exchange 8:483
social network analysis in 16:528–529
of socialist societies
see also Socialist societies
and society (STS), at home
see also Home
sociocultural 1:772
and sociology 1:766
of Spanish anarchism 1:683–684
specializations in 1:770
specialties 1:775
of spirit possession 23:282
of sport 23:299–306
of statehood 23:347–352
structural functionalism 1:768–769
structuralism in 1:768, 23:626–627, 23:630
studies of worldview 25:753–755
subdisciplines 1:775
sub-Saharan Africa 11:89–90
symbolic 1:768, 23:860–865
temporality 1:747–749
terminologies 1:772–773
of time 24:406–408
of tourism 24:465
of tradition 24:503–507
tribes in 24:645–646
of trouble 1:226
in twentieth century, second half of 1:768–770
urban 24:774–782
of values 25:33–35
Victorian 1:766
violence in 25:122–127
of war 4:614–619, 25:377–381
of Western societies 1:770
witchcraft 25:604–605
of work 25:662

Anthropology and Medicine 19:311
Anthropology Research Group on Homosexuality (ARGOH) 19:753–754
Anthropometry 22:527
  quantification 19:704
Anthropomorphic concepts, autonomous agents 2:310
Anthropomorphic (psychological) sense of fear 8:868
Anthropomorphism 6:36, 20:390
Anthropophagy, cannibalism 3:83–88
Anthroposophical medicine 10:563
Anthroposphere 9:197
Antiabortionists 16:247
Antialcoholism campaign 1:161, 1:163–168
Anti–anti relativism 5:485–486
Anti-Arab racism
  implications of 17:430
  and Orientalism 17:429–431
Antiballistic Missile (ABM) Treaty 1:976–977
Antibiotic cycling 8:454–455
Antibiotic management, in hospitals 8:454–455
Antibiotics 15:869
  research 6:694
Anti-Black racism, in social work education and practice, Canada 8:154–155
Anticategorical intersectionality 21:810
Anticategorical research 12:664
Anticipated agreement hypothesis 14:713–714
Anticipated reactions 18:760, 18:762
Anticipation 18:51–52, 18:55, 24:125–127
  and feedback 24:126–127
  of future developments 24:127
  philosophy of 2:894–895
  of social aspects and impacts 24:126
  and symptom perception 23:868–869
  and unemployment 10:516–519
Anticipation ratio 18:57, 18:60
Anticipatory coping 4:903
Anticipatory nausea and vomiting (ANV) 2:419
Anticipatory socialization, reference group 20:73–74
Anticlassification principle 7:894
Anticlericalism 3:536–537, 15:328
  and Enlightenment 7:664–665
Anticollectivism 22:313
Anticolonial movements 25:838
Anticolonial nationalism 16:264
Anticolonialism
  in Canada 8:154
  and indigenous social work 11:807
  and nationalism 16:264
Anti-Comintern Pact 1:557, 21:369
Anticommunism 25:397
  Bell on 2:498–499, 2:502
  and conservatism 1:604
  counterinsurgencies 5:100–101
Anti-consumerism 8:25–26
Anticonvulsants, for bipolar disorder 2:687–688
Anticore 4:871
Anti-Counterfeiting Trade Agreement signed in 2011 and 2012 12:572
Anticraving drugs, with psychopharmacotherapy 19:461
Anticult movement (ACM) 16:808
Anticyclone 9:571
Anti-death penalty lawyers and policy makers 5:875
Antidepressants 1:778–783, 4:31, 17:595–596, 18:8
  for bipolar disorder 2:685, 2:687–688
  with different pharmacological profiles 1:779
  ethical, methodological, and clinical issues 1:781
  future prospects 1:781–782
  history 1:778–779
  indications 1:780–781
    bulimia nervosa 1:780
    depressive disorders 1:780
    generalized anxiety disorder 1:780
    obsessive–compulsive disorder 1:780
    panic disorder 1:780
    premenstrual syndrome 1:780
    social phobia 1:780
  mechanism of action 1:779–780
    blockade of enzymes that inactivate monoamines 1:779
    blockade of one or more types of synaptic monoamine transporters 1:779
    stimulation and/or blockade of receptor subtypes 1:779–780
  monoaminooxidase inhibitors 1:778
  placebo-controlled double-blind studies of 18:161
  with psychopharmacotherapy 19:460
  for schizophrenia 21:19
  selective serotonin reuptake inhibitors 1:778–779
  side effects of 19:463–467
  for sleep disorders 22:85
  therapeutic principles 1:780–781
  tricyclic antidepressants 1:778
Antidiscrimination
  housing 11:244
  racial 19:809–810
Antidiscrimination laws
  economics of gender 9:760–761
  and gender variant individuals 24:551
  against homophobia 10:793
  for people with disabilities 13:498
  against racism 5:266
  in United States 6:530–533
Antidiscriminatory practice, and criminal/social justice 5:211–212
Antidopaminergic drugs, for schizophrenia 21:13–14
Antidoping system 6:671
Anti-Drug campaign 6:681
Anti-egalitarianism 2:255
Antiepileptic drugs (AEDs) 24:555
Antiessentialism 22:223–224
  and racial identity 19:810
Anti-Europe 8:273–274
Antievolutionism 20:287
Antifeminism 25:800
  and psychoanalysis 22:766
Antifetishist tradition 9:109
Antifeudal–anticompradore political movements 16:243
Antiforeigner sentiments 17:652
Antiformalism 20:22
Antifoundationalism 18:806
  and cultural relativism 5:484
Antifraud rules, corporate governance 4:935
Antiglobalisation 25:839
Antiglobalization movements 11:451, 16:858
  and global south 10:186–187
  and youth movements 25:839
Anti-Homosexuality Bill, December 2013 (Uganda) 21:770
Antiidentity movement 24:552
Anti-immigration 25:797
Antiinflammatory drugs, for schizophrenia 21:13–14
Antiintellectualism 12:275, 12:760, 16:458–459
Anti-intraception 2:255
Antimafia
  campaign 14:433
  strategies in Italy 14:434
Antimonies of global south 10:188–190
Antimulticulturalism 25:800
Antinaturalism, and hermeneutics 10:778–779, 10:781–782
Antinaturalists 16:326–327
Antinepotism rules, and scholars 21:37–38
Antinuclear movements 3:711, 21:176
  and green parties 10:381
Anti-oppression, silencing of antiracism by 8:153–154
Antioppressive practice (AOP) 3:522, 15:201–202, 16:458, 17:248–249, 21:805, 22:520
  call to action 17:250–251
  colonial social work 16:458–459
  for diablism 6:450
  and harmony 10:531
  and masculinity 14:682
  and Orientalism 17:429–431
Antioppressive research (AOR) 22:271–273
Anti-oppressive social work 6:456, 14:32
Antioppressive theory 17:248
Antipassivization 25:2
Antipathy 12:760–761
Antiphilosopher's philosopher, Rorty as 20:761
Antipode: A Radical Journal of Geography 10:85, 14:649, 14:674
Antipornography 17:220
Antipornography Civil Rights Ordinance 13:884–885
Antipositivism 22:997
Antipoverty
  effects, of policy interventions 18:745
  law 18:734–736
  policies 18:747
  programs/movements 18:733, 18:735–739
  strategies 18:735

Antipriming 11:711–712
Antipsychiatry 19:587
Antipsychiatry movement 19:302–303
Antipsychologism, Husserl on 11:441–442
Antipsychotics 19:201, 19:463–465
  anticholinergic effects 19:464
  atypical 19:465, 19:468
  for bipolar disorder 2:688
  chronic effect of 19:466
  and deinstitutionalization 6:48–49
  drugs 19:464, 19:466–467
  medications 19:465–466
  negative symptoms 19:292
  for schizophrenia 21:12, 21:17–20, 21:22–23, 21:18t–19t
  side effects of 19:463, 19:466–467
Antiquarianism 1:891–892, 23:63–64
  development of 24:504
Antique Greek political philosophy 4:337
Antiquity 2:644, 11:36, 14:244, 17:948
  cannibalism 3:85
  and Christianity 11:38
  and classical archaeology 3:760–763
  and economic growth 5:528–529
  historiography of 11:38
  and prosopography 19:256–257
  and scientific revolution 21:319–320
Antiquity of war 4:615–616
Antiquity/city states (iron age)
  genocide, history of 10:9–10
Antiquity/feudalism (bronze age)
  genocide, history of 10:8–9
Antiquus 15:700
Antiracism
  in Canada
    efforts to advance antiracism in social work 8:153
    silencing by anti-oppression perspectives 8:153–154
  and culturally appropriate interventions 5:516–518
Antiracist community work theories 4:380
Antiracist legal reform 19:809–810
  antidiscrimination 19:809–810
  antisubordination 19:810
  cultural discrimination 19:810
  implicit bias 19:810
  integrationism and separatism 19:809
Antiracist social work, in Roma 20:757–758
Antirationalism 12:760
Anti-relativism 9:675
Antirepresentationalism 20:762–763
Antiretroviral therapy (ART) 1:491–493
  and AIDS-related mortality 15:836
  for HIV/AIDS 15:836–837, 15:839–843
Anti-Rightist Campaign 3:516
Antisanist practice 15:202–203
Anti-school hypotheses 25:170
Antiscientific criticisms 1:735
Antiseizure drugs (ASDs) 7:853–855
Anti-semitism 1:784–789, 4:764, 11:151, 16:260, 18:411–412, 18:414, 18:840, 19:852–854, 25:799–800
  and authoritarian personality 2:255
  connecting 1:786–788
  dealing with past 1:785
  disconnecting 1:786–788
  ethical turn 1:785–786
  and political parties 18:411–412, 18:414
  in social work 1:784–785
Anti-Slavery Society
  Agency Committee (UK) 22:40
Antisocial behavior order (ASBO) 1:791, 1:793, 19:633–634
Antisocial behaviors (ASBs) 5:234–235, 19:241, 23:564
  advocacy 1:794
  care ad control functions 1:792–793
  in children 2:470
  defining 1:790–792
  deviance 1:790, 1:794
  discipline 1:792
  ethics 1:792–794
  evolutionary developmental model and 8:427
  implications for social work practice 1:793–794
  labeling 1:791
  nature/nurture 16:343
  nuisance 1:790–793
  and psychopathy 19:451–453
  social control 1:790, 1:792–793
  street-level bureaucrats 1:794
Antisocial personality disorder (ASPD) 6:7, 17:892, 17:894, 17:896, 19:451
Antisocial personality/lifestyle impulsivity, control of 20:437
Antisocial potential (AP) 14:61–62
Antisocial workplace behaviors 25:722–724
Antisodomy laws, sexual orientation 21:770
Anti-stigma interventions 15:232
Antisubordination 9:730
  inequality 7:896, 7:899
  and race 19:810
Antisyncretism 23:882
Antiterrorism actions 13:532
Antiterrorism and Effective Death Penalty Act 1996 (AEDPA) 11:637, 13:531
The Antithesis of Pitirim Sorokin 23:26
Antitrust 1:796–803
  and Chicago School 1:799
  goals 1:797
  in high technology and dynamic network industries 1:800–801
  historical developments 1:797–800
  international 1:801–802
  leveraging 1:801
  public choice approach 1:800
  standard setting 1:801
  strategic economic behavior 1:799–800
  structure-conduct interventionist period 1:798–799
Antitrust law 3:31
  Clayton Act 1:797
  FTC Act 1:797
  Sherman Act 1:796–797
Antiurban bias thesis 18:182
Antiurbanization sentiments, in Africa 24:962–963
Antiveiling laws 20:331
Anti-Vietnam war, and counterculture 5:83
Antizionism 1:787
Antoinistes 16:806
Antonio, Emile de, works, *In the Year of the Pig* 6:623
Antonovsky, Aaron 10:612
ANV *see* Anticipatory nausea and vomiting (ANV)
Anxiety 1:804–810, 8:349, 8:748–749, 8:872, 17:469, 17:471, 17:872–874, 23:563, 24:244, 8:348t
  amygdala 1:804–807
  animal models of 1:808
    conditioned emotional responses 1:808
    ethological models 1:808
  and cancer 3:73–75
  and cardiac disease 20:188
    coping 3:114
  cholecystokinin tetrapeptide 1:804, 1:808–809
  corticotropin-releasing factor 1:804, 1:806–809
  definition 1:811
  and diabetes 6:340
  differentiation of 1:805
    fear 1:805
    stress 1:805
  expectancy-value-cost model 8:505–506
  of failure 24:244
  in functional GI disorders 9:608–609
  and grief 2:537–539
  and HIV infection 1:492–493
  and hopelessness 11:199
  human models 1:808–809
    challenge studies 1:809
    state and trait anxiety 1:808
  in inflammatory bowel disease 9:608
  intervention programs, and academic achievement 1:24
  mental–physical comorbidity 4:390–393
  and neonatal temperament 16:495, 16:497–498
  neural and behavioral mechanisms of 1:814–815
  neural basis of 1:811–817
  neuronal basis of 1:806–808
    neuroanatomical structures 1:806–807
    transmitter systems 1:807–808
  neuropeptides 1:804, 1:807–808
  neurotransmitter 1:804, 1:806–807
  noradrenalin 1:804, 1:807, 1:809
  panic disorder *see* Panic disorder
  and personality 17:872–874
  and physical activity 18:89–90
  and placebo effect 18:153–154, 18:156
  prenatal 23:556–560
  psychotherapy for 19:374–375
    generalized anxiety disorder 19:374
    panic disorder 19:374
    posttraumatic stress disorder 19:374–375
  religion and 20:264
  and respiratory disorders 20:576–579
  rheumatoid arthritis 20:657
  rumors and 20:789–790

serotonin 1:804, 1:806–808
sleep abnormalities in 22:83
and social-emotional development of children 22:944, 22:946–947
and socioeconomic status 2:478
sources in humans 1:805–806
genetic and biological disposition 1:806
life events 1:806
social and environmental influences 1:806
understanding using fear conditioning 1:811
Anxiety disorders 1:804–810, 3:766, 11:617–618, 15:194, 17:593, 23:11–14, 23:333, 25:138
behavior therapy for 2:426–429
cognitive behavioral therapy for 25:138
exposure therapy in 17:476–477
fear conditioning 8:852–853
and fear-potentiated startle 8:860
generalized anxiety disorders 1:805
imaging of 11:605, 11:607, 11:609
and neuroticism 16:744–745
and obsessive-compulsive disorder 17:120–121, 17:124–125
panic attacks and panic disorder 1:805, 17:468–473
perinatal 17:722
phobias 1:805
short-term psychodynamic therapy for 19:345–346
Anxiety sensitivity (AS) 22:495
and panic disorder/agoraphobia 17:477
Anxiolytics, for schizophrenia 21:19
Anxious attachment 22:566–567
Anzaldua, Gloria 19:747
ANZASW *see* Aotearoa NZ Association of Social Work (ANZASW)
ANZSOG *see* Australia and New Zealand School of Government (ANZSOG)
ANZUS (Australia, New Zealand, United States) Treaty 1:558
AOB *see* Accessory olfactory bulb (AOB)
AOIR *see* Association of Internet Researchers (AOIR)
AOL, business process outsourcing 3:46
AOP *see* Antioppressive practice (AOP)
AOR *see* Antioppressive research (AOR)
AOS *see* Accessory olfactory system (AOS)
Aotearoa, Maori social work 14:505–511
Aotearoa NZ Association of Social Work (ANZASW) 14:508
Social Work Registration Board 14:507–508
A-over-A principle 16:4
AP *see* Acquired prosopagnosia (AP); Affect programs (AP); Antisocial potential (AP)
APA *see* American Planning Association (APA); American Psychiatric Association (APA); American Psychological Association (APA); Association for the Advancement of Psychoanalysis (APA); Association of American Planners (APA)
Apache Kinship Systems 2:506
Apartheid 16:29, 16:31–32, 23:78–79
and desegregation 6:218, 6:220, 6:223–227
APC
see also Age, period, and cohort (APC) analyses; Article processing charge (APC)
APE *see* Accumulative prediction error (APE); Associative–propositional evaluation (APE) model
APEC *see* Asia Pacific Economic Cooperation (APEC)
APEL *see* Assessment of Prior Education and Learning (APEL)
Apel, Karl-Otto 4:794, 18:810
language philosophy 4:790–791, 4:794–795
Apes, goal-directed behavior 10:271
Aphasia 1:585, 1:587, 1:818–821, 4:92, 12:782, 13:406, 21:944, 23:294
acute phase after brain damage 13:406
agrammatic 16:672
and anomia 19:148
assessment of 3:793
bedside and screening tests 3:793
cognitive neuropsychological model of spoken picture naming 3:793f
comprehensive examinations and aphasia batteries 3:793–794
specific aspects of language 3:794
bilingual 2:574–579
chronic phase after brain damage 13:406
definition 1:818, 21:944
functional consequences of 1:821
history 1:818–819, 16:678–687
and paradigmatic and syntagmatic axes of language 15:291
psycholinguistic approaches to 1:819–821
sign language 21:934–936
Aphasia Diagnostic Profiles 3:793–794
Aphasia therapy
approaches to 13:330–333
efficacy of 13:330–333
Aphasiology 13:926, 16:678
Aphemia 13:329, 23:251
Apirana Mahuika et al. v. New Zealand 11:799
APIs *see* Application Programming Interfaces (APIs)
APLIC *see* Association of Population Libraries and Information Centers (APLIC)
Aplysia 10:481–483
Aplysia californica, learning and memory of 23:876
Apni Beti Apta Dhan (ABAD) program 4:558
Apo di tutti capi 14:422
Apocalyptic thought, Christian historiography 11:33–34
Apocrypha 3:548
APoE *see* Apolipoprotein E (APoE)
Apolipoprotein E (APoE) 6:441–442
genotype and cognitive functioning 9:831f
Apolitical citizenship 3:651–652
Apollo Program 11:315
Apollonian child 7:147
Apologetics 9:657–658
conception of dictatorship 6:372
Apoptosis, in brain development 2:824–825
Apostasy 12:742, 20:363–364
Apostles of Infinite Love 16:807
Apostolic fathers 3:545, 3:547–548
Apostolic succession 3:545, 3:547–548
Apotimema 21:401
APP *see* Amyloid precursor protein (APP); Aß; precursor protein (APP)
Appadurai, A. 3:317
works, *The Social Life of Things: Commodities in Cultural Perspective* 5:109–110
Appalachian Regional Commission (ARC) 20:143
Appalachian Trail (US) 1:219, 16:138
Apparent distance 25:199
Apparent time hypothesis 22:949
Appeals
and adjudication 5:106
appellate jurisdiction, principle of 1:824–825
as a central feature of court 1:822
in criminal cases 1:822–823
discretionary review of 1:822
distinguished from other proceedings 1:822
immediate 1:823–824
interlocutory 1:823–824
legal 1:822–825
as a matter of right 1:822–823
nature of 1:822
principle of restraint 1:823–824
record on 1:824
standing to 1:823
Supreme Courts and 23:722–723
use of term 1:822
Appearance motivation, and sun exposure 23:696–699
Appeasement 1:826–831, 21:370
balance of power 1:826, 1:828–829
balancing 1:829
challenging conventional wisdom 1:827
of Chamberlain's policy 1:826–830
conciliation 1:826–828, 1:830
definition 1:826
deterrence 1:826–828, 1:830
engagement 1:830–831
ethics of 1:828–829
foreign policy 1:826, 1:830–831
grand strategy 1:829
Hitler and 1:826–830
incentives 1:828, 1:831
inducements 1:826, 1:828, 1:831
lessons of 1930s 1:826–827
Munich 1:826–827
and peace 1:826, 1:828
success conditions 1:827–828
and war 1:826–830
Appellate jurisdiction, principle of 1:824–825
Apperceptive agnosia 1:440–441
Apperceptive prosopagnosia 19:251–254
Appetite

Appetite (*continued*)
and food preferences 9:296
taste and smell 24:26–27, 24:30
Appetitive aggression 24:439
Appetitive behavior, and weight regulation 25:470–471
Appetitive motivation 2:327, 2:329–330
Appetitive–aversive distinction 2:329–330
Apple
diversification of 6:586, 6:588
strategic management of 23:510
Applebaum, William 1:846
Applicant perceptions 17:851
Application Programming Interfaces (APIs) 12:867
Application representation 11:402
Application studies, econometric software 7:3
Applications, core–periphery patterns 4:915–916
Applicative affix 25:3
Applicative constructions 25:5
Applied anthropology 1:742, 1:775, 16:234
ethnohistory 8:192
Applied behavior analysis (ABA) 2:407–411, 2:413
application, principal areas of 2:410–411
clinical problems 2:410
community psychology 2:410
education 2:410
forensic applications 2:410–411
industrial/organizational behavioral management 2:410
larger sociocultural applications 2:411
sport psychology 2:411
behavior therapy 2:407–408, 2:410
behaviorism 2:410
definition 2:407
historical background 3:880–881
history 2:408
from laboratory to applications 2:407–408
reinforcement 2:407–411
single-subject designs 2:407–408
social learning 2:407–408, 2:410
techniques 2:408–410
avoidance training 2:408–409
behavioral self-control 2:410
biofeedback 2:408
cognitive procedures 2:410
contingency management 2:409
environmental design 2:409
escape training 2:408
extinction 2:409
modeling 2:410
Premack principle 2:409
punishment 2:408
shaping 2:409
systematic desensitization 2:408
token economy 2:409
Applied criminology 1:832–838
crime 1:832–833
offenders 1:833–834
victims 1:834–836
policy, developments in 1:836
restorative justice 1:835–836
Applied cross-cultural psychology 5:314–315
acculturation 5:314
health 5:315
intercultural communication and training 5:315
intercultural relations 5:314–315
national development 5:315
work and organizations 5:315
Applied demography 1:839–844
de facto population estimates 1:841
in private sector 1:840
professional recognition 1:841–842
in public sector 1:840–841
subsets 1:839
training 1:842
in United States 1:839
Applied economics 1:854
Applied ethics 2:625–626
public policy 19:580–581
and virtue ethics 25:150–151
Applied geography 1:845–846, 10:87
American Geographical Society (AGS) 1:846–847
careers in 1:846–847
client driven 1:845
early leaders 1:846
future of 1:848
historical context 1:846–848
institutionalization 1:846–847
in largest organizations 1:847
themes 1:848
US Bureau of Census Geography Division 1:847
Applied health 10:573
Applied mathematics, statistics in 23:423
Applied probabilistic measurement structures 24:250–255
construction of interval scales 24:253
construction of ordinal scales 24:252–253
integration of linear models 24:254
isotonic probabilistic (ISOP)
additive conjoint ISOP (ADISOP) 24:253–254
complete ADISOP (CADISOP) 24:253–254
linear ISOP 24:254
measurement structure 24:252–253
item response theory 24:250
making deterministic axiomatic structures probabilistic 24:251–252
Applied psycholinguistics 14:228
Applied psychology, of technological change and sociotechnical system 24:104–108
Applied research 7:415
on crime and criminal justice 5:127–131
criminology and criminal justice 5:127
corrections 5:129
crime monitoring/surveillance 5:127–128
definition 5:127
evaluation 5:128
evidence-based practices 5:129
for-profit research 5:128
government funding in 5:128
implementation 5:129
intervention 5:129
law enforcement/policing 5:129–130
nonprofit research 5:129
personnel studies 5:130
university-based research 5:128–129
what works 5:129
on immigration 11:635–640
infrastructure 22:670–671
on intimate partner abuse 12:677–684
realism in 20:20–25
single-case designs 21:985
Applied social research 1:850–857
Booth on 1:851–853
data bases and statistical systems 5:692–699
disciplines 1:855–856
DuBois on 1:853
Durkheim on 1:851–852
economics 1:854–855
Enlightenment 1:850, 1:854, 1:856
ethnography in 8:184–191
Freud on 1:851, 1:853
Galton on 1:851–853
individualism 1:850
Marx on 1:851–852
Munsterberg on 1:854
origins 1:850–851
political science 1:855
psychology 1:853–854
Quetelet on 1:851–852
Smith on 1:850, 1:854
social reform 1:851–852
social theory 1:850–851
sociology 1:851–853
Weber on 1:850–852
William on 1:850, 1:853
Applied social science, and area studies 1:949
Applied sociology 1:858–860
assessment of 1:858–859
clinical sociology 1:858–859
community-based research 1:858–859
definition 1:858
and engagement 1:858–860
evaluation 1:858–859
intervention 1:858–859
participatory-action researchers 1:858–859
public sociology 1:858–860
role of sociologists 1:858
sociological practice 1:858–860
sociological practitioner 1:858–859
translational sociology 1:858–859
use-based application/research 1:858
Applied statistics
in EITMs 18:455–456
for economic voting 18:456–457
linking with formal analogs 18:457
procedures that fail to identify mechanisms 18:454–455
Applied zooarchaeology 25:852–853
Apportionment 7:342–344, 19:206–207

census 3:302, 3:304–306
districting 7:342–344
political geography 7:342–343
Appraisal theory, of emotions 7:474
Appraisals 4:895–896, 5:544–545, 7:483–484
coping 4:887–890, 4:893
and coping 4:903
dimensions of 7:483t
and emotions 7:526–531
models 7:461–462
support 12:403
technique 17:717–718
Appreciating the value of thought (level of consciousness) 4:670
Appreciation process 7:786
Apprentices 2:47
Apprenticeship 1:861–864, 1:870–873, 1:910, 2:47, 15:679, 18:786
adolescence 1:861–863
anthropological aspects 1:865–869
as anthropological field method 1:867
coaching 1:864
cognitive apprenticeship 1:863–864, 1:871
communities of practice 1:871–872
craft production 5:110
craftwork 1:865–869
cross-cultural comparison 1:869
exploration 1:863–864
fieldwork 1:866–867, 1:869
industrialization 11:869
learning as 22:6
and learning forms 1:867–869
modeling 1:864
participant observation 1:867–869
problem-solving 1:871–872
Reformpädagogik, situated learning 1:870–871
articulation and reflection 1:871
cooperation 1:871
as enculturation 1:871
multiple perspectives 1:871
by solving authentic problems 1:871
scaffolding 1:864
school instruction as 1:863–864
school-to-work transition 1:862
self-regulation 1:863
situated learning 1:861, 1:864
skill-based learning 1:865, 1:868–869
social organization form 1:866–867
as solutions for problems with school learning 1:870
specialist training 1:866
and transition to adulthood 24:580
and vocational education 25:232, 25:234–237
vocational training 1:861–862
Approach–avoidance motivation 2:326–331
appetitive 2:327, 2:329–330
approach–withdrawal distinction 2:329–330
aversive 2:327, 2:329–330
biology 2:326–327, 2:329–330
cognition 2:326–327
conceptualization 2:327–329
definition 2:326
hedonism 2:326–327, 2:329–330
history 2:326–327
and social withdrawal 2:456, 2:458
terminology 2:329–330
theories of personality 17:859–860
use of term 2:330
valence 2:326–330
withdrawal 2:328–330
work motivation 25:684–685
Approach-oriented coping, for cardiac disease 3:113
Appropriability, and innovation 12:165–166
Appropriate beach behavior 19:867–868
Appropriate dispute resolution 15:70
Appropriateness 4:298–299
Appropriateness of play differences, during infancy 18:219–220
Approval-oriented moral reasoning 7:315–318, 7:316t
Approved Social Worker (ASW) 19:302–303
Approximate Bayesian computation (ABC) 1:371
bootstrap 16:90
Approximate dynamic programming (ADP) 23:478–479
Approximate inference, significance tests 21:959
Approximate magnitudes 17:46, 17:48–51
Approximate number system (ANS) 17:50–51
Approximate replication 20:441
Approximate sufficiency 23:421–422
Approximations
combinatorial data analysis 19:432
clustering 19:434
seriation/scaling 19:433
variable selection 19:434–435
normal 6:571–572
Apps 6:899
Apraxia 1:874–880, 17:509
clinical examination of 1:874–876
communicative gestures 1:876–877
defective tool use 1:877–879
sign language 21:936–937
signs 6:76
of speech 23:251
use of term 1:874
Apraxic agraphia 1:453–454, 6:735–736
Après la crise, 2010 24:463
Apriorism 22:986
Apriorist approach to categories of thought 13:70–71
APRNs *see* Advanced practice registered nurses (APRNs)
APS *see* Addiction Potential scale (APS); American Psychological Society (APS); Association for Psychological Science (APS)
APSA *see* American Political Science Association (APSA)
APSs *see* Advanced producer services (APSs)
Aptitude 21:365
Aptitude-by-Treatment Interactions (ATI) 4:113, 12:327–328
Apukhtin, Aleksei 21:741
AQA *see* Association for Queer Anthropology (AQA)
AQAM *see* Al-Qaeda and associated movements (AQAM)
AQS *see* Attachment Q-Set (AQS)
Aquainas St. Thomas 18:32–33
Aquinas, St. Thomas 3:771–772, 7:567, 10:484, 12:933, 13:762, 13:858, 16:320, 18:436, 20:285
Works, *Summa Theologiae* 3:771–772, 10:484, 16:320, 16:323–324
AR *see* Action research (AR); Augmented reality (AR); Auto-regressive (AR) model
Arab Barometer 5:793
Arab countries 7:210–211
Arab Democracy Barometer 5:849
Arab Federation of Iraq and Jordan 4:575–576
Arab Foundation for Freedoms and Equality (AFE) 21:795
Arab League 9:73
Arab marriage 23:628–629
Arab Muslims 8:160, 12:727–728
culture 12:738
Arab nationalism 15:355, 16:271–274
in Africa 16:271, 16:273
contingency and continuity 16:271
Arabism 16:271–272
pan-Arabism 16:273
Islam 16:271–273
in Middle East 16:271, 16:273
modernity 16:271–272
national identity 16:271, 16:273
religion 16:271–273
Arab nationality, defined 20:340
Arab Spring 3:202, 4:634, 6:46, 6:98, 7:919, 9:782, 10:189, 11:416–417, 12:744–745, 14:717, 16:374–375, 16:798–804, 20:311, 20:625, 21:189–190, 21:791, 24:452, 25:840, 25:844
FEMEN-inspired postings of nude photos 21:794
generations 9:870
and media 15:12, 16:778–779
and social media 12:602, 16:800
Egypt 16:801–802
Saudi Arabia 16:802–803
Tunisia 16:800–801
social media in 18:423–424
Arab world 15:354
confederation, during cold war 4:575–576
sexual politics in 21:791–796
Arabic (language, literature) 11:61–65
and diglossia 6:419, 6:421–422
in Europe 17:423–424
Arabidopsis thaliana 7:835
Arabism 16:271–272
Arab–Israeli conflict 10:745, 16:363
Arabization policy 10:18
Arab-Jew, Memmi as 15:137

Arapesh, gender assignment in 9:798
Ararat 2:148
ARAS *see* Ascending reticular activating system (ARAS)
Arbeitsgemeinschaft Kriegsursachenforschung (AKUF) 5:836
Arbitrage 2:437–438
Arbitrariness 11:512
  axiological neutrality 16:757–758
Arbitrary control 4:828
Arbitrary detention 13:532
Arbitrary patterning, in word order 25:655
Arbitrary power 20:781–783
Arbitration 6:558–559, 15:70–74
  and adjudication 5:103, 5:105–106
  conventional 6:558
  empirical evidence 6:559
  final-offer 6:558–559
Arbitration centers 12:470
Arbitration clauses 13:917
Arbitrator
  becoming an 12:472–473
  judge *versus* 12:879
ARC *see* Appalachian Regional Commission (ARC)
Arc consistency, constraint-satisfaction problem 4:717
Arc de Triomphe (Paris, 1806–1836) 3:771
ArcGIS (Environmental Systems Research Institute) 10:58, 10:97
  software 23:138, 23:140
ARCH *see* Autoregressive conditional heteroscedastic (ARCH) models
Archaeobotanist 7:694–695
Archaeobotany 7:694–695
Archaeological archives 7:692–693
Archaeological culture 1:900–901
Archaeological evidence of household inequality 11:221
Archaeological history 1:891–896
  antiquarianism 1:891–892
  disunity of science 1:895
  early calls for science 1:891–892
  epistemic settlements 1:893–895
  new archaeology 1:892–893
  philosophical pluralism in 1:895–896
  philosophy of science 1:892–893
  science studies 1:895
Archaeological prospecting 10:53
Archaeological science 7:692–693
Archaeology 14:670, 23:63
  Arctic 1:921–926
  of Australia and New Guinea 2:229–233
  bioarchaeology 2:600–603
  of chiefdoms 3:376–382
  chronology, stratigraphy, and dating methods in 3:562–567
  classical archaeology 3:760–763
  cognitive archaeology 4:23–29
  of conflict and war 4:600–613
  correlation with language family prehistory 13:340
  of cultural evolution 5:394–395, 5:397–399
  and cultural/national memory 1:881–884
  current orientations 1:896–897
  data collection in 23:63
  dating systems in 23:744
  definition 23:742
  dispersal histories with claimed correlations 13:340–342
  environmental 7:692–698
  ethics 23:745
  and ethnic identity/ethnicity 8:102–105
  ethnoarchaeology *see* Ethnoarchaeology
  of Europe 8:269–272
  excavation 23:745
  feminist 8:901
  forensic *see* Forensic archaeology
  gender, feminism, and sexuality in 9:764–767
  geoarchaeology 10:53–57
  and geographic information system 10:97, 23:744–745
  geospatial analysis 23:744–745
  geospatial data
    agent-based modeling 10:102
    analytical activities 10:99–102
    chained analysis 10:101
    data collection platforms 10:98–99
    online visualization tools 10:97–98
    predictive modeling 10:101–102
    spatial analysis 10:100–101
    terrain analysis 10:101
  historical 11:1
  history 23:742
  history of engagement 1:891–896, 23:63
    antiquarianism 23:63–64
    early calls for science 23:63–64
  of hunter-gatherer societies 11:423–427
  importance of 23:742–743
  of intensification and specialization 12:329–332
  of language families 13:337–343
  LiDAR 23:744–745
  lithics 14:260
  Marxist 14:670–673
  as memory 1:883
  versus memory 1:882–883
  of Mesoamerica 15:252–265
  migrations, colonizations, and diasporas in 15:472–482
  modern developments 23:745–746
  and national agendas 1:883–884
  of North America 17:15
  of Pacific Islands 17:446–450
  paleoethnobotany in 17:458–461
    human evolution and emergence of modern behavior 17:458–459
    Iceman and His environment 17:460
    origins of food production 17:459–460
    ritual and feasting 17:460–461
  philosophical pluralism in 1:895–896
  of philosophy and science 1:891–898
  politics of 1:885–890
  of ritual 20:721–726
  settlement and landscape 21:654–658
  sites 23:743
  as source of information on human prehistory 13:340
  of South America 23:36–42
  of South Asia 23:50–57
  of Southeast Asia 23:63
  of Southeast Asian 23:63–64
  of states and civilizations 23:371–375
  studies 23:743
  survey in 23:744–745
  of symbolism 20:721–726
  theory in 1:899–904
  of trade and exchange 24:481–484
  typological schemes 23:63–64
  of war 4:600–607
  zooarchaeology 25:849–853
Archaeomagnetic dating 3:566–567
Archaeometry 1:905–908
  age determinations 1:905–906
  ceramics 1:907
  metallurgy 1:906–907
  paleodiets 1:907
  source of artifacts 1:906
Archaeozoology 25:849
Archaic Caucasoids 12:794
Archaic epoch 3:760–761
Archaic period, North America 17:16
Archaic religion 2:507–508
Archaic society, transition to 20:315
Archaic states 4:603–604, 17:449
  and irrigation 12:725
  see also States
Archeological and ethnographic cultures, phylogenetic studies of 5:397–398
Archeology and iconography of music 16:156
Archers 25:415
Archetype, tricksters 24:648–649, 24:651–652, 24:654–656
Archimean axioms 14:863
Archimedes 19:66
  works, *The Sand Reckoner* 17:907
Archipel 11:87
Architect state 5:464
Architectures
  artificial intelligence, hybrid models 2:38–39
  and gender 9:711–712
  of human mind 12:314f
  of intelligence 12:317
  knowledge representation 13:100–103
    frame-based systems 13:102–103
    large-scale systems 13:103
    logic-based systems 13:100–101
    semantic networks 13:101–102
  Mumford on 16:137–138, 16:140
Archival labors 1:910
Archival methods 1:909–915
Archival studies of jury case 12:908
Archives 5:700–703, 17:284–285, 17:287
  challenges to professional traditions 1:912–913
  collection, preservation, and cataloging methods 1:911–912
    in bureaucratized institutions 1:911

construction of statistical series 1:912
data 5:791, 5:817–820
database 5:831–832
electronic information technology 1:913
future of 1:914
of Indigenous collection 11:774
Internet 5:702
modern archives since French Revolution 5:701
modern public 1:909–910
in European states 1:909
France as the center 1:910
French Revolution and 1:909–910
technology of state power 1:909
museums 5:702
national 5:847
preservation and information technology 5:702
professional cultures of archivists 1:912
professional history 1:910–911
records as historical sources 5:701–702
as research sites 1:909
terminology and history 5:700–701
Archives of Family Medicine 15:116
Archiving 14:187–188
by-products of social and behavioral research 1:916–917
and disseminating 17:289
ethics 1:916–920
censoring potentially controversial or offensive material in collections 1:919
consent for data sharing 1:918
due credit assignment 1:919
effective program archives 1:919
future 1:920
ownership of research and development by-products 1:919–920
protecting integrity of selection process 1:917
protecting respondents' confidentiality 1:917–918
special considerations for clinical trial and biomedical data 1:918–919
tension between fidelity and usability 1:919
timing of release of information 1:919
linguistic databases 5:814–816
Archivists, professional cultures of 1:912
Archons 6:61–62
Arctic
ecological definitions of 1:921
sociocultural aspects 1:927
animism 1:929–930
and Christian missionary activity 1:930
circumpolar North 1:927–931
climate change 1:927–928, 1:931
contemporary developments 1:930–931
ecology and economy 1:928
geopolitics 1:927, 1:931
human development 1:927
hunting and gathering 1:928–929
indigenous and colonial histories 1:927–928
indigenous inhabitants 1:927
kinship 1:929
local-level politics 1:931
religion and worldview 1:929–930
self-determination 1:930
shamanism 1:929–930
sharing of resources 1:928–929
social and political organization 1:928–929
Subarctic 1:927–929
Arctic archaeology 1:921–926
Arctic Mongoloids 1:759
Arctic Small Tool tradition (ASTt) 1:922–923
Arcuate fasciculus 23:251–252
Arcuate nucleus 11:421–422
Arcuate/superior longitudinal complex in speech production 23:251–252
ARD *see* American Research and Development (ARD)
Ardener, Edwin 7:146
Area and international studies (A&IS)
stakeholders, in United States 1:943–948
academic disciplines 1:943–944
cold war 1:946
consumers 1:944–945
demand and supply 1:945–946
educational institutions 1:943–947
Fulbright-Hays 1:946
governments 1:943–947
Higher Education Act 1:943, 1:945–948
Higher Education Act (HEA), Title VI 1:943–948
investors 1:946–948
producers 1:943–944
Area demand, and land use indicators 13:242
Area deprivation and health 22:285–286
Area of residence 22:933
Area studies 1:933–937, 1:949–954
applied social science 1:949
cultural studies *see* Cultural studies
in Eastern Europe 6:805–808
Eastern European studies 6:814
foreign policy 1:949–953
history 1:933–934
intelligence 1:949–952
interdisciplinarity 1:949, 1:951, 1:953
language studies 1:933–935
political economy 1:938–942
in Soviet Union 1:949–951
diversification during Cold War 1:950–951
oriental studies in interwar period 1:949–950
Sputnik 1:933
in United States 1:951–954
criticisms and newer trends 1:953–954
diversification and liberalization 1:952–953
early Russian/Soviet studies 1:951–952
intellectual trends 1:953
postwar, institutional takeoff 1:951
World War II 1:951
Area symbols 24:265–266
Area V1
interleaved pathways through 17:129
neurons in 17:128
retinotopic projection to 17:127–128
Area V2, interleaved pathways through 17:129
Area-based interventions 24:921
Areal differentiation 20:122–123
in regional geography 20:122–123
Areal linguistics 1:955–960
definition 1:955, 1:957–958
examples 1:955–957
Balkans 1:955
Baltic 1:956–957
Ethiopia 1:957
Mesoamerica 1:956
Northwest Coast of North America 1:956
South Asia (the Indian Subcontinent) 1:955–956
kinds of linguistic areas 1:959
linguistic areas *versus* borrowing generally 1:959–960
linguistic concerns 1:959
for linguistic reconstruction and subgrouping 1:958
proposals of distant genetic relationship 1:958–959
Area-related land use indicators 13:238–239
Arendt, A., on work/labor 25:677–678
Arendt, Hannah 1:554, 1:753, 1:961–964, 2:519, 2:522–523, 10:871, 18:438–439, 19:616, 22:828–829
on alienation 1:551, 1:554, 1:963
on American/French Revolutions 1:597, 1:599–600
on authority 2:272
on bureaucracy 1:962–963
conception of 'being in the world&rsquo 1:963
on democracy 1:963
on dictatorship 1:961–962
on European Societies 1:961–962
and Marx 1:963
methodological approach to theory and thinking 1:961–962
on Nazi Germany 1:963
on political freedom 1:962–963
as political thinker of morality 1:963–964
on politics 1:962
on power 1:963
on public sphere 3:704
on radical and banal evil 1:963
and social 22:828–829
on society 1:962
on totalitarianism 1:963
and Weber 1:962–963
works
The Human Condition 22:828–829
Arens, Hans, works, *Sprachwissenschaft* 14:217
ARFIMA *see* Auto-regressive fractionally integrated moving average (ARFIMA) model
Argentina 22:627
community participation in health 4:402–403

Argentina (*continued*)
gay/lesbian activism in 13:859–862
transgender activism in 13:861–862
Argentine Confederation 4:574
Argentine Social Forum of Health and the Environment 4:402–403
Argentinian Dirty War 9:336–337
Arginine vasopressin (AVP) 11:476–477, 23:577
ARGOH *see* Anthropology Research Group on Homosexuality (ARGOH)
Argument incorporation 14:221–222
Argument nominalizations 23:891–892
gap in 23:892
marginal NP-use of 23:892–893
Argument structure of predicate 16:672–673, 25:1–6
Argumentation 13:787, 13:789–791
legal 13:776–783
social interaction 13:673–674
and thinking 20:653
Argumentative strategies of de-Westernization programs 25:547–549
Arguments 23:901–902
explanation as 8:594–595
Arguments to the best explanation 11:15
Arid zone, in Australia 2:231–232
Ariely, Gal, on national pride 16:312
Arìes, Philippe, works, *Western Attitudes toward Death* 7:595
Ariga, Nagao, works, *Ethnic Evolutionism* 24:585
ARIMA *see* Autoregressive integrated moving average (ARIMA) models
Aristocracy 1:965, 15:769–770
ancient-world origins 1:965–966
and celebrity 3:275
early modern state 1:967
European 1:968
European history of 1:965–969
French revolution 1:968
mafias 14:423
as rule of the best 1:966
and social mobility 22:427–428
triumph and twilight of 1:968–969
urban 1:966
Aristocratic armies 4:603–604
Aristophanes 16:155–156
Aristotelian ideal 3:645
Aristotelian notion of constitutionalism 4:706
Aristotelianism 20:286
Aristotelian natural philosophy 20:285
Aristotle 3:771–772, 3:776, 11:885, 11:890
on artificial worlds 22:703
on citizenship 3:643–644
on civil society 3:707–708
contributions to 19:407
democratic theory of 6:80, 6:90–92
on environmental determinism 7:720
on ethics 8:74–75
on friendship 9:426–427
on habit 10:476
on historians 11:126
and historical thought 11:36–37
on inductions 19:63
on irrationality 12:719
on justice 12:933–936
on knowledge 11:126
on neurolinguistics 16:680–681
notion of asceticism 2:60
philosophy 18:31–32
on poetry 11:126
on political citizenship 3:650–651
psychology 19:407–409
on religion 20:285
in Renaissance 1:967
on republicanism 20:477
on semiotics 21:592
on slaves/slavery, history of 22:55
on solidarity 23:1
on virtue ethics 25:147–150
on wisdom 25:599, 25:601
works
De anima (*On the Soul*) 18:32
Ethica Nikomacheia 18:32
Nichomachean Ethics 4:706, 6:232, 6:589, 12:932, 22:886
On Rhetoric 15:289
Politics 3:63–64, 3:643, 4:706, 7:721, 12:932, 13:437, 13:799, 22:878
on written language 14:243
Arithmetic, political 23:423
Arithmetic items 9:835
Arithmetic mean 20:731
ARK *see* Absolute Return for Kids (ARK)
Ark of Taste 22:95
Arkheia 1:909
Arkhipov, Nikolai 21:742
Arm actions, cortical areas involved in 5:26–27
superior parieto-occipital cortex, precuneus, and mIPS 5:26–27
ARMA *see* Autoregressive moving average (ARMA) models
Armamentarium, theoretical 11:247–248
Armchair anthropology 9:149, 17:715
Armed Conflict Dataset (ACD) 5:836
Armed Conflict Location and Event Dataset (ACLED) 5:836
Armed conflicts 10:2–5
impact on children 25:389–391, 25:393
nation-states 25:420, 25:423
safeguarding children abused in 3:426
social identity, societal narratives, and ideology 17:655
and social work 1:970–975
asylum seekers 1:970, 1:973
background information 1:970
child soldiers 1:970–972
clients 1:972–973
ethical dilemmas 1:973
experiencing 1:970–971
impact on social workers 1:973
internally displaced people 1:970
long-term impact 1:971
positive outcomes 1:971
refugees 1:970–973
relevance to humanitarian activities 1:972–973
research 1:972
training and educational initiatives 1:973–974
UCDP–PRIO 8:97–98, 8:98t
Armed forces 15:502–504
demobilization, *versus* postsocialist transition 25:123
psychiatry *see* Military psychiatry
Armed Forces and Society (*AF&S*) 15:515
Armed phase 15:327
Armed struggle, and child soldiers 3:441–443
Armenian Apostolic Church 17:438
Armenian Christians 20:328–329
Armenian language 11:824
Armoniene v. Lithuania (2009) 18:950
Armor, warfare 25:414–415
Arms 15:506
Arms control 1:976–978
Antiballistic Missile Treaty 1:976–977
Cold War 1:976
disarmament 1:976–977
horizontal proliferation 1:976
noncompliance 1:976
nuclear club 1:976
Nuclear Non-proliferation Treaty 1:976–977
Strategic Arms Limitation Talks 1:976–977
Strategic Arms Reduction Talks 1:977
Arms race 6:2, 6:5
concept 8:450–451
metaphor 8:450
Armshell, Kula ring 13:140–142
Army
privatization of 6:3
professional influence on military history 15:506
Army Alpha test 3:879
Army Beta test 3:879
Army psychiatrists 6:48
Arnheim, Rudolf 2:24, 15:15
Arnold, Edwin, works, *The Light of Asia* 2:891
Arnold, Matthew 14:691
works, *Gender and Education* 7:117–118
ARNOVA *see* Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA)
Aromatase cytochrome P450 (CYP19) 21:781
Aron, Raymond 1:979–983, 13:108, 18:438–439, 22:699
biography 1:979–980
career 1:979
on international diplomacy 1:980, 1:982–983
on philosophy of history 1:980
on political theory and ideology 1:981–982
positions held by 1:979
on sociology and sociological theory 1:980–981
on war and international relations 1:982–983
works 1:980–983

Democracy and Totalitarianism 1:982
Eighteen Courses on Industrial Society 1:981
Manuscripts of 1844 1:981
The Opium of the Intellectuals 1:979
Politics and History 1:980
Arousal 15:947–948
and aggression 1:374–375, 1:379
and emotions 7:520–523
empathic 19:216–218
and sleep 22:88–90
Arousal system 15:948
ARPA *see* Advanced Research Projects of the Army (ARPA)
Arpajon 2:703
ARPANET *see* Advanced Research Projects Agency Network (ARPANET)
Arranged marriage, and dowries 6:642
Arrangement, marital 14:602
Ar-Raniri, Nuruddin, works, *Garden of Kings* (*Bustan as-Salatin*) 11:83
Arrest 5:185–186
for domestic violence 6:626
Arrestee drug abuse monitoring (ADAM) 5:715–716
Arrighi, Giovanni
works
Adam Smith in Beijing: Lineages of the 21st Century 25:342
The Long Twentieth Century: Money, Power & the Origins of Our Times 25:342
Arrival stories 24:291–292
Arrow, Kenneth J.
impossibility theorem of 19:896–897, 25:500
works
Social Choice and Individual Values 7:90, 18:484
Arrow–Debreu model 17:190
ARROWSMITH (data mining tool) 21:294–295
Arson 15:592–593
ART *see* Antiretroviral therapy (ART)
Art
prehistoric 18:824–827
Art districts, in China 5:598–600
Art history 2:17–19, 2:21–26, 20:426–428
character of time 2:21–23
image history 2:24–25
to mannerist Renaissance 20:428–429
material, language, visual studies 2:25
and Renaissance 20:427–428
revolution of medievalists 20:426
space: globalization 2:23–24
visuality 25:207–210, 25:212
Art museums, marketization of 14:566
Art of human portraiture 2:645
Art of living 10:522
Art of Marajoara 23:39–40
Art of oratory 20:650
Art production, of Australian aboriginals 2:237
Art worlds 5:466–467, 14:565
avant-garde 2:322–323
Becker's contribution 2:404–405
A&RTeam *see* Advisory and research team (A&RTeam)
Artefactual field experiment 9:129
Arthritis, and physical activity 18:88, 18:90
Article processing charge (APC) 12:866
Articles (journal)
electronic articles 12:867
numbers 12:865
retractions, growth in 12:864
Articles of Confederation 1:612
Articles of UNCRC 12:957
Articulation 7:15–16, 10:743, 13:900–902, 18:65
concept 13:900
illustrations 13:900–902
language acquisition 13:246
and political economy 18:364
Articulatory phonetics 18:65
Artifact discrimination, electroencephalography 7:356
Artifact emotions 16:658–659
Artifact manufacture *see* Craft production
Artifact provenance, geoarchaeology 10:53, 10:56
Artifact sourcing studies 10:100–101
Artifacts 14:766–770
archaeological, physical analysis of 10:56
archaeometry 1:906
corporate culture 4:924
earlier clues about 8:572
and ethics in collision 8:574
experimenter 8:571–573
minimizing potential 8:573–574
and organizational symbolism 16:181
scientific 25:221–224
in scientific inquiry 8:571–572
in social and behavioral 8:571–572
subject 8:571, 8:573
Artifactual mechanism 15:851
Artificial computation 2:27–31
dynamics of biological neurons 2:28–29
genetic algorithms 2:30
mathematical notion of computation 2:27
neural networks 2:28–30
in neural systems 2:28
Turing machine 2:27–28
Artificial feeding 2:848t
Artificial genes 10:39
Artificial grammar learning (AGL) 13:83–84
finite state grammars 13:83–84
learning and classification 13:84
in patients with amnesia 13:85
Artificial intelligence (AI) 2:32–40, 4:84, 4:716, 5:646, 5:934, 13:98, 13:784, 14:320, 21:187, 21:291–292, 23:416
algorithms 1:530, 1:532
cognition 2:35, 2:39
connectionist 2:37–38
knowledge representation 2:37–38
learning 2:37
constraint-satisfaction problems 2:33–34
emergence of 4:103–104
emulated world of 21:187
genetic programming 2:41–45
hybrid models 2:38–39
architectures and representations 2:38–39
learning 2:39
knowledge representation 13:98–99
and law 13:784, 13:786–787, 13:789–790
neural networks 2:35, 2:37, 2:39
paradigms 2:35
principles 2:309
propositional *versus* imaginal representations 4:106–107
research 19:60
single-agent path-finding problems 2:32
symbolic 2:35–37
criticisms 2:37
learning 2:37
representation and search 2:35–36
two-player games 2:33
Artificial language-learning experiments 8:360
Artificial limb 15:969
Artificial neural networks 22:703
Artificial neural systems, reasoning 5:938
Artificial randomizers 8:567
Artificial societies 1:368–369
Artificial worlds, simulation of 22:703
Artisanal and craft production *see* Craft production
Artisanal labor 2:46–47
Artisans 1:868, 2:46–51
ambiguous journeyman 2:47
in class society 2:47–48
concepts 2:46–48
in corporate society 2:47
definitions 2:46
enlightenment philosophy 2:47
historiography 2:49–50
history 2:46–48
in industrial revolution 2:49–50
Marx concept 2:47
in modern economic 2:46
in neolithic revolution 2:46
in *petit bourgeois* 2:47
in political economy 2:47
and poverty 18:729
in proletarian 2:47
revisionism in 2:49–50
role of women 2:50
skill 2:46
in towns 2:47
work 2:47
working class 2:48
Artist craftsmen 5:111
Artistic aesthetics, and STS 21:183
Artistic charisma 3:355
Artistic classification systems 3:799–800, 22:336
Artistic creation 5:730
support, and cultural policy 5:462–463
Artistic genius 6:428
Artistic legitimation 3:800
Artistic market 14:564–568
art worlds 14:565
artistic reputation 14:565–566
cultural industry 14:564–567

Artistic market (*continued*)
and culture 14:567
gatekeeper 14:565, 14:567
high culture model 14:565
history 14:564–565
and Internet 14:567
long-tail retailing strategy 14:567
marketization of art museums 14:566
and neoliberalism 14:566
new technologies 14:567–568
privatization 14:566–567
production of culture 14:564–565
uncertainty of 14:565
visual art markets 14:566–567
Artistic reputation 14:565–566
Artistic revolution, concept of 2:322
Artists 2:52–59
assortative matchings 2:55–56
avant-garde 2:320–325
careers
labor markets and 2:3, 2:56–57
and social networks 5:468–469
of culture 5:731–732
and job stratification 2:52–53
as lone genius 5:466
quality uncertainty 2:53–55
star jobs and Paretian distribution of earnings 2:53
ARTs *see* Assisted reproductive technologies (ARTs)
Arts 5:441, 18:824
Aboriginal art 11:773–774
African Studies 1:276, 1:278
anthropology of 2:15–20
aesthetics, ethnoaesthetics, and universals 2:16–17
agency 2:16–17
challenge 2:15–16
contemporary anthropology and disciplinarity 2:18–19
in contemporary world 2:17–19
crafts 2:15, 2:18
ethnic and tourist arts 2:17
ethnography 2:16–18
heritage 2:17
history 2:17–19
linguistic approaches 2:16
mainstream arts 2:17
material culture 2:18
meaning 2:16–17
metropolitan art world 2:17–18
postmodernity 2:15
psychological approaches 2:16
sociopolitical approaches 2:16
world art system 2:17–18
avant-garde 2:320–325
and brain asymmetry 23:296
and cultural production 5:608–611
and culture 5:618
and cybernetics 5:647–648
databses/statisitical systems for 5:727–729, 5:732–733
degradation, mass communication 14:691–692
and disgust 6:548
economics of 2:1–6
aesthetics 2:1, 2:4–5
art history 2:1, 2:4
copyright 2:1–3, 2:5
cultural and creative industries 2:5
cultural heritage 2:4–5
labor markets and artists' careers 2:3
performing arts 2:4
public support 2:2–3
visual arts 2:3–4
ethnopsychology 25:793
film as 9:165–166
folk 9:283
iconographic representations in 4:612–613
in MENA 16:361–362, 16:364–365
Mumford on 16:137–138
music *see* Music
participation 13:848–849
and photography 18:83–84
postmodern methodology as 18:681–682
and socialization 2:7–14
adolescence 2:9–13
arts education 2:11–12
cultural capital 2:7–9
cultural competence 2:9
cultural mobility model 2:8, 2:7f
cultural participation 2:10–12
cultural reproduction model 2:8, 2:7f
empirical evidence 2:10–12
family socialization 2:10–11
general education 2:12
information theory 2:9
methodological challenges 2:9–10
parents 2:7–13
school socialization 2:11
status motives 2:9
status theory 2:9
sociology of 5:728
symbolic boundaries 23:852
use of term 2:15
works of 11:100
Arts and Crafts movement 5:111
Arts Councils (Great Britain) 5:464
Arts education
arts socialization 2:11–12
impact of reduced provisions for 1:234
Arts et métiers 2:47
Arts events, attendance at 5:376
Arts integration 1:233
Arts lessons, and cultural capital 5:376
Arts participation
cultural consumption 13:847–849, 13:848t
definition 5:447
socialization 2:7–14
ARTstor (image library) 2:24
Arturo Soria y Maya 24:863
Artwork, meaning of 2:322
ARV *see* Antiretroviral (ARV) drugs
Arya Samaj 20:374
AS *see* Acquired stuttering (AS); Anxiety sensitivity (AS)
Aß precursor protein (APP) 14:363–364
ASA *see* African Studies Association (ASA); American Sociological Association (ASA); Attraction–selection–attrition (ASA) model
Asad, Talal 8:291
on public symbols and private discipline 23:863–864
Asad al-Jihad 2 24:231
ASAT *see* Automatic speech attribute transcription (ASAT)
ASBO *see* Antisocial behavior order (ASBO)
ASBRs *see* Age-specific birth rates (ASBRs)
ASBs *see* Antisocial behaviors (ASBs)
ASC *see* Academic self-concept (ASC); Altered state of consciousness (ASC); American Society of Criminology (ASC)
Ascended Masters 16:807
Ascending arousal system 22:81f
VLPO projections to 22:81f
Ascending reticular activating system (ARAS) 8:624, 16:750–751, 22:88–89
Ascertainment bias 11:291
Ascetic movements 6:822
Asceticism 2:60–64, 3:238
monastic asceticism 2:61–62
and mysticism 2:62–63
origins of 2:60–61
social theories of 2:62–63
use of term 2:60
Asch, Solomon 19:414
Aschaffenburg, Karen, on cultural capital 5:377–378
Ascribed honor 11:896–897
caste/race 11:896–897
ethnicity 11:897
sex/gender 11:897
Ascription
equality of opportunity 7:904
migration history 15:450
Ascription by others 8:103–104
ASD *see* Acute stress disorder (ASD)
ASDR *see* Age-standardized death rate (ASDR)
ASDs *see* Antiseizure drugs (ASDs); Autism spectrum disorders (ASDs)
ASE *see* American Society for Ethnohistory (ASE)
ASEAN *see* Association of Southeast Asian Nations (ASEAN)
Asena, Duygu 20:873
Asexual reproduction 2:617
totemism 24:455
ASFRs *see* Age-specific fertility rates (ASFRs)
ASHA *see* American Speech-Language-Hearing Association (ASHA)
Ashanti Confederacy 4:575
Ashb, Ross 14:383
ASHG *see* American Society of Human Genetics (ASHG)
Ashgate (Surrey, UK and Vermont) 7:587
research companion series 7:588
Ashmolean Museum in Oxford 16:142
Ashmore, M., works, *The Reflexive Thesis* 20:94
Ashoka, Emperor 2:887–888
Asia

central *see* Central Asia
child mortality in 12:21–22
China *see* China
confederation 4:574–575
counterinsurgencies in 5:100–101
cultural capital research in 5:379
in global economy 7:36
human genomics in 11:318–322
Asia Cohort Consortium 11:319
biopiracy 11:320
collaboration 11:320–321
contribution 11:318–319
current challenges 11:319–321
ELSI 11:319–321
funding 11:319–320
future perspectives 11:321
genetic diversity 11:318–319
genomic medicine 11:318–321
local concerns 11:319
PASNP 1.0and 2.0 11:319
publics 11:321
regulation 11:320
trust 11:319–321
impact of Cold War in 4:125
impact of family status on labor market in 8:758–759
international social work in 12:559–560
Islam in 12:727–730
Korea *see* Korea
national liberation movements 16:241, 16:243–245
nostalgia in 17:35–36
population change in former communist states of 18:545–552
social and behavioral science institutions in 22:615–620
Asian Development Bank 22:619–620
Benedict Anderson 22:616–617
big data analysis 22:617–618
colonial research societies 22:616
contributions of 22:619–620
defining community and state 22:616–617
drivers of R&D 22:616–618
government-linked research centers 22:619–620
maps 22:617
national icons 22:619–620
national museums 22:617
national narratives 22:616–617
national population census 22:616–617
national think tanks 22:615, 22:619–620
private research organizations 22:619
R&D centers 22:618–619
Royal Asiatic Society 22:616
smart power 22:615–616
tertiary institutions 22:618
third world development programs 22:617
United Nation agencies 22:619
universities 22:617–618
urban policies 24:894–899
diversities 24:894–895
in East Asia 24:895–896
in South Asia 24:898
in Southeast Asia 24:896–898
values in 25:27
see also South Asia; Southeast Asia
Asia as method 2:607, 18:656
Asia Cohort Consortium (ACC) 11:319
Asia Minor 3:546
Asia Pacific Economic Cooperation (APEC) 12:574, 12:578, 13:621
AsiaBarometer survey 5:793, 5:849, 19:567, 22:635
Asian Agri-History 17:672
Asian Cities Climate Change Resilience Network (ACCCRN) 20:564
Asian Development Bank 21:152, 22:619–620
Asian Drama research project 2:765
Asian eugenics 8:230–231
Asian financial crisis 6:790–791
Asian flu 18:174
Asian Journal of Social Psychology 11:793
Asian laws 13:815–816
Asian neoliberalism 24:898–899
Asian Perspectives 23:64
Asian tigers 6:278, 6:791, 12:358
Asian values 5:528, 13:816
Asian women 8:909
Asian-Americans
and model minority stereotype 19:791
teens 21:841–842
Asianism 16:277
Asian-model of autocratic paternalism 4:707
Asians
British *see* British Asians
ethnic composition in United States 18:557t
Asiatic societies 12:724
Asilomar Conference on Recombinant DNA (1975) 9:1017
ASL *see* American Sign Language (ASL)
ASM *see* Active Shape Model (ASM)
As-novice *versus* as-expert researchers 4:771
Asociación de Trasvestis Transexuales y Transgenérico 13:861–862
Asociación por la Identidad Trasvesti Transexual 13:861–862
Asokan edict 23:55f
ASP *see* Answer set programming (ASP)
ASPD *see* Antisocial personality disorder (ASPD)
Aspen Institute 22:663
Asperger's disorder 2:275
Aspers, Patrik
on cultural industries 5:536
on value 5:538–539
Aspirations 21:490
ASPO *see* Association for the Study of Peak Oil (ASPO)
ASSA *see* Academy of the Social Sciences in Australia (ASSA); American Social Science Association (ASSA)
al-Assad, Bashar 24:451–452
ASSAf *see* Academy of Science of South Africa (ASSAf)
Assault 15:592
Assault weapons 10:457
Asselin, Olivier 25:211
Assemblage theory 5:279–280, 24:948
Assemblages 10:94–95, 18:151
of relations 20:91
Assembly, test 7:930
Assertion 12:95–96
normativity of content (NCA) 14:844, 14:847–848
Assertive community treatment (ACT), for psychiatric disorders 22:384
Assertively egalitarian societies 11:894–895
Assertives 18:797
Assessment of Prior Education and Learning (APEL) 7:920–921
Assessment-led learning, and curriculum differentiation 4:114
Assessments 15:283–284, 22:757
assessment for learning, use of term 4:796
Asset pricing, stock market 23:483
Asset purchases/sales, ongoing usefulness of 15:732
Asset specificity, hierarchies and markets 10:842
Assets
racial difference in 6:543
settlement assets 19:177–179, 19:181–182
sex as 10:816
Asset-seeking investments 12:581
ASSH *see* Australian Society for Sports History (ASSH)
Assigned sex 21:678
Assignment of form by inferencing 13:268–269
Assignment of homework 4:545–546
Assignment variables 20:153
Assimilation 1:71, 1:73, 5:677–678, 8:159, 14:87–88, 18:126, 18:784–785, 20:790, 22:212
American pluralism 2:86–87
in American sociology 2:82–84
and bilingual education 2:581, 2:584
and cognitive development 4:52–53, 18:120, 18:122
and dimensional comparisons 6:433–434
in education 22:214
and ethnic intermarriage 8:118–119, 8:121–122
and ethnicity 8:143, 8:145–146, 8:159
evolution 2:82
of immigrants 2:82
in immigration 2:86–87, 11:676, 15:429, 15:431, 15:459
incorporation processes 2:84–85
acculturation 2:84–85
identification 2:84–85
integration 2:84–85
Jewish Diaspora 12:801
and life span development 14:103–104
migration 15:436–437
minority-group status 15:572
paradoxes 2:86
process 5:154
prospects 2:86
racial relations 19:846
rejection of neonatal imitation 16:484

Assimilation (*continued*)
and residential segregation 20:542
sociological concept of 2:82
upward assimilation theory 22:212–213
in Western Europe 12:235–237, 12:239–241, 12:236t
and work 8:107–112
Assimilation effects 6:433–434, 20:883–884
in attitude measurement 2:178–179
of I/E model 12:428–429
impact of education environments 21:473
Assimilation of injustice 2:489
Assimilation theory, and learning 13:710
Assimilationism 11:679, 12:916
Assimilative and accommodative coping 14:87
ASSIST *see* American Stop Smoking Intervention Study (ASSIST)
Assistants, in academics 1:27
Assisted dying 7:600
Assisted human reproduction, in affluent nations 20:472–473
Assisted reproductive technologies (ARTs) 12:47–48, 20:453, 20:468–469, 20:471–474, 23:442
Assisted suicide 7:600
Assmann, Jan, on memory 1:882
Association cortex 5:22–23, 17:506
brain, evolution of 2:827–828, 2:830
Association for Consumer Research (ACR) 4:739, 4:742
Association for Psychological Science (APS) 19:416
Association for Public Policy Analysis and Management 22:662
Association for Queer Anthropology (AQA) 19:754
Association for Research on Nonprofit Organizations and Voluntary Action (ARNOVA) 25:254, 25:266–267
Association for the Advancement of Psychoanalysis (APA) 9:461
Association for the Cure of Inebriety 17:240
Association for the Study of Peak Oil (ASPO) 15:554
Association Française de Science et des Systèmes Cybernétiques, Cognitifs et Techniques (AFSCET) 22:912
Association International des Musées d'Agriculture 17:671
Association Internationale des Sociologues de Langue Française 22:676
Association measures, epidemiology 7:829
Association of American Geographers (AAG) 1:847
Association of American Law Schools (AALS) 7:555, 13:734–735
Section on Clinical Legal Education's database 19:559
Association of American Planners (APA) 18:195
Association of Black Social Workers 8:153
Association of Caribbean Social Work Educators 12:475
Association of Critical Heritage Studies (Gothenburg, Sweden) 5:414
Association of Geographic Information Laboratories in Europe 10:60
Association of German Geographers 7:166
Association of Greek Psychologists, establishment 3:874
Association of Internet Researchers (AOIR) 12:607
Association of Medical Superintendents of American Institutions for the Insane 17:28
Association of Population Libraries and Information Centers (APLIC) 1:917
Association of Schools of Allied Health Professions 16:908
Association of Southeast Asian Nations (ASEAN) 2:347–348, 9:84–85, 12:575–576, 13:812, 23:70, 23:72–73, 24:223
Association psychology 2:700–701
Association studies, of sexual orientation 21:781
Associational community 22:854
Associational power 13:94
Associational solidarity 12:403, 12:405
Associationism, theory of 24:980–981
Associationist psychology 13:125
Associations 4:90–91, 12:371–372, 12:791, 22:720
anthropology 1:756
community sociology 4:361–363
informal 12:58
International Association of Legal Science 2:90
International Economic Association 2:90
International Political Science Association 2:90
International Sociological Association 2:90
International Union of Anthropological and Ethnographical Sciences 2:90
International Union of Psychological Science 2:90
national academies 22:596–597
quantitative memory models 15:163–164
rejection of neonatal imitation 16:484
scholarly 4:475–479, 22:596
voluntary associations *see* Voluntary associations
Associative agnosia 1:440–441
Associative functions 17:297
Associative learning 11:717–718, 15:158
eyeblink conditioning 8:635–637, 8:639–640
role of cerebellum in 2:92
and semantics 25:228
Associative memory, and hippocampal amnesia 10:877, 10:879, 10:882–884
Associative prosopagnosia 19:251–254
Associative Recall Index 16:725
Associative–propositional evaluation (APE) model 2:172–173, 11:716–718
Assonance 13:276
Assortative matchings 8:120
and artistic labor 2:55–57
Assortative mating 5:402, 8:715, 11:336, 11:742, 11:890, 12:298–299
in marriage market 2:100
behavioral interpretation of local log odds in marriage matching 2:104–106
Choo on 2:100, 2:104–106
educational marital matching 2:102t–103t
ethnic marital matching 2:100t, 2:103t
local log odds in education matching 2:102t
measuring endogamy 2:100–102
positive assortative matching 2:102–103, 2:106
rationalizing endogamy 2:105–106
Siow on 2:100, 2:104–106
social exchange 2:103–104, 2:106
trade-offs in multidimensional matching 2:103–104, 2:106
microevolution 15:340
Assumption and Deterrence Act 5:635
Assumption into Heaven 3:234–235
Assumption-free heritability estimation 8:717
Assumptions 11:496
capture–recapture methods 18:606–607
realism 20:26
Assurance game, protest as 18:421–423
Assurances, Common Rule 4:252, 4:256
A*STAR *see* Agency for Science, Technology and Research (A*STAR)
Astereognosis 17:507–508
Asthma
comorbid psychological symptoms in 20:576–577
interventions in 20:578–579
social aspects in 20:577
Aston research program 24:115
Astounding Science Fiction (magazine) 16:190
As-treated estimate 4:446, 4:450
Astronomical hazards 6:481
Astronomy 18:100
ASTt *see* Arctic Small Tool tradition (ASTt)
ASW *see* Approved Social Worker (ASW)
Asylum policies 2:109
Asylum seekers 2:108–114, 20:105, 20:107–108
advocacy and social work 2:111–112
and armed conflict 1:970, 1:973
in Australia 2:109–110
background 2:108
border security *versus* human security 2:108–109
human rights
norms 2:108
violations 2:110–111
international human rights approach 2:111
islands of places of banishment 2:110
migration history 15:446, 15:450
racism 2:109
refugees 2:108–112
role of social workers 2:110–111

social exclusion due to 22:288
social work ethics 2:109, 2:111–112
use of immigration detention 2:109–110
and xenophobia 25:803, 25:805–806
Asylum seeking, and migration 15:380
Asylums 19:587
decline of
public mental health systems 19:587–589
public psychiatry 19:587–589
and immigration policy 11:662, 11:664, 11:666
rise of 6:47–48
middle ages 19:586
renaissance 19:586–587
Asymmetric agrammatic comprehension 14:176
Asymmetric dominance 19:444
Asymmetric information 1:344–345, 12:103–104
about health insurance markets 10:641–643
adverse and advantageous selection 10:641–642
moral hazard 10:642–643
and ignorance 11:579
principal-agent models 12:104
Asymmetric learning, promotions 19:137–138
Asymmetrical affirmative action 1:252
Asymmetries 16:413
in intergenerational justice 12:396–397
in pathological populations 13:404
social representations 20:445
Asymmetry of body and brain 2:115–130
axis formation and positioning of gut, heart, and central nervous system 2:116–118
central nervous system development, anterior-posterior partitioning 2:118
CNS asymmetry 2:125–127, 2:127f
cortical specification 2:118–120, 2:119f–120f
dyslexia 2:115–116, 2:125–126
genetics 2:115–116, 2:126–127
handedness 2:115, 2:121, 2:126–128
human embryology 2:116–120, 2:117f–118f
language 2:115–116, 2:120, 2:125–128
laterality 2:115–116
left-right asymmetry body (organs) 2:123–124
mirror movements 2:115
molecular basis 2:121–128, 2:122f
molecular pathways setting stage 2:124–126, 2:125f–126f
processes laying ground work for asymmetry 2:122–124, 2:123f–124f
substrates underlying cerebral dominance 2:126–128
Asymptomatic sufficiency 23:421
Asymptotic analysis, cointergration 24:328
asymptotic distribution of estimators 24:328
asymptotic distribution of rank test 24:328
Asymptotic distribution 16:885–890
Markov chain Monte Carlo methods 14:569
simultaneous equations
single-equation 21:976–978
whole system 21:976
survival analysis 23:779
Asymptotic independence results 17:293
Asymptotic inference, unit roots time series 24:363, 24:367
Asymptotic tests 11:493
Asymptotic theory 17:293
asymptotic independence results 17:293
central order statistics 17:293
extreme order statistics 17:293
intermediate order statistics 17:293
L-statistics 17:293
Asynchronous 6:567–568
Asyntactic comprehension 16:691–692
AT *see* Agency theory (AT)
Ataman, Kutlug, works, *Two Girls* 20:873
Ataque de Nervios ("Attack of Nerves") 5:624
Ataturk, Mustapha Kemal 13:218
ATE *see* Average treatment effects (ATE)
Athabaskan, linguistic work on 20:899
Atheistic materialism 20:286
Atheists 20:356
Athenian democracy 6:61–62
Athens 1:698
democracy in 6:80
property rights 19:190–191, 19:194
Athletes
gay 9:656, 9:658–659
lesbian 9:656–659
Athletics, anthropology of 23:299–306
ATI *see* Aptitude-by-Treatment Interactions (ATI)
Atkinson, A.B., works, *Elsevier Handbook of Income Distribution* 11:737
Atkinson, P., on interviews 12:674
Atkinson index 11:905–906
Atkinson's measures 11:886
Atlan, Henri 14:383
Atlanta Lesbian Feminist Alliance 15:112
Atlanta University Studies 7:560
Atlantic, challenges to Central Africa 3:313–314
Atlantic capitalism 20:746
Atlantic Revolutions 1:612–613, 20:647
Atlantic Telegraph Company 24:155
Atlantic world, slaves/slavery history 22:55–57
ATLAS *see* A Toroidal LHC Apparatus (ATLAS)
ATLAS detector 7:874
Atlas Foundation 24:298
Atlas linguistique de France (ALF) 6:348
Atlas ti, software package 22:545
Atluri, Tara, on sexual citizenship 21:731
Atmosphere 9:567
Anthropocene 1:724–725
ATMs *see* Automatic telling machines (ATMs)
Atomic age, warfare 25:418
Atomic bomb 21:368, 21:884
Atomic Bomb Casualty Commission 8:231
Atomic Energy Act 3:282
Atomic fallout 16:16
Atomic force microscope (AFM) 25:222–223
Atomic mechanisms 22:416–417
Atomism 2:131–135, 15:309, 15:311
and holism
arguments 2:134–135
characterization 2:132–134
historical background 2:131–132
types of 2:131
Taylor on 24:39
Atomistic individualism 4:308
Atomistic orientation of liberalism 17:735
Atoms 20:286
Atoms of kinship 23:628
ATP *see* Adolescent Transition Program (ATP); Attitudes toward pornography (ATP)
Atrocities Documentation Survey (ADS) 10:17–18
Atrocity stories, war propaganda 25:398
Atropa belladonna 16:753–754
AT&T 24:154–155, 24:763
AT&T Mobility v. Concepcion 3:734–735
Attached specialists 3:379
Attachment 2:758, 6:747, 16:513–514
and adolescent friendship 9:432–434, 9:436
childhood 3:454–455
development during infancy and early childhood
Bowlby's attachment concept 2:137
context-informed attachment research 2:141–142
cross-cultural attachment research 2:140
cultural variations 2:140–142
emotional development 2:137
empirical approaches 2:137–140
intergenerational continuity 2:139
mentalization 2:139
parenting 2:141
psychobiology 2:138–139
research methods and problems 2:139–140
roots of attachment theory 2:136–137
situating attachment theory 2:136
Harlow on 10:526
integrated attachment theory 12:228
and love 14:377–380
maternal sensitivity 2:138–139, 2:141
and neonatal bonding 16:469–472
and organizational commitment 17:334, 17:336–338
and prenatal/infant development 18:854, 18:856, 18:859
and psychodynamic thought 22:762, 22:765
and social referencing 22:560
Attachment, national 16:311
Attachment and Biobehavioral Catch-up (ABC) 12:233
Attachment behavior system 2:791–792

Attachment behavior system (*continued*)
cross-cultural studies 2:793
implications of 2:793
normative development and parenting 2:793
psychopathology 2:793
quality of 2:792
social policy issues 2:793
Attachment Q-Set (AQS) 12:231
Attachment Q-sort 2:139
Attachment security 21:490
Attachment strategies, of children 2:141
Attachment theory 20:351, 21:426
and child neglect 3:414
childcare 3:396
friendship formation 9:432
of socialization and education 22:862–863
see also Integrated attachment theory
Attachment-promoting signals, and neonatal crying 16:475
Attainment value (task value) 1:38, 7:890–891
expectancy-value-cost model 8:504–507
in vocational psychology 25:241
Attempted utopias 2:144
American enlightenment 2:145
architectural experiments 2:151–152
Christian (and Islamic) roots of utopian thinking 2:144–145, 2:147–148
communal experiments of 1960s and 1970s 2:152–153
communal utopians 2:146–147
ecovillages 2:151–152
extant Anabaptist utopian communities 2:147
futurism 2:151
literary influence on 2:150
and New World 2:145–146
and pietist intentional communities 2:145–146
reformation 2:145
religious revivalism on American soil 2:146
Shakers 2:146–147
technology 2:151
utopians in the American West 2:147–148
ATTEND bereavement model 5:862
Attending (communication skill) 4:288
Attention 7:247–248, 7:250–252, 12:418–419, 16:635–637, 24:174, 24:176, 24:245, 7:248f
in agenda-setting 1:357–361
change blindness 3:344–349
education of 10:131
and emotions 7:461–465
and hypnosis 11:468–470
imaginal neglect 25:165–166
and intelligence 12:316–317
and memory binding 15:141–145
neural basis of 2:163
anatomical locus of selective attention 2:163–164
bottom-up and top-down biases 2:165
control of visual attention 2:167
dorsal stream 2:163–166
multiple visual pathways 2:163
neural mechanism of selection 2:166–167
nonspatial feature-based selection 2:166
receptive field 2:163–167
saccade 2:166
space-based selection 2:165–166
subject of selection 2:165–166
ventral stream 2:163–165
visual cortex 2:163–165, 2:167
visual search task 2:166–167
visual selective attention 2:164–165
and neuromarketing 16:698, 16:700
selective, in concept learning 4:539–540
and short-term memory 21:904–906
and social cognition 22:206–207
and symptom perception 23:867–868
virtual spatial navigation for 16:351
Attention and Concentration Index 16:725
Attention and Executive Functions domain in a Neuropsychological Assessment-II 8:495
Attention deficit hyperactivity disorder (ADHD) 2:453, 2:810, 9:444, 11:605, 12:420, 15:3–4, 17:82–83, 18:820–821, 19:304, 21:634, 23:557–558, 23:560
neural basis of 2:155–162
connectivity studies 2:158–159
cortical and subcortical neurocircuitry dysfunction 2:157–158
cortical dysfunction 2:157
diffusion tensor imaging 2:158
dopaminergic system 2:159
functional connectivity 2:158–159
functional imaging studies 2:156–159
historical background 2:155
magnetic resonance imaging 2:156–159
neurobiological basis 2:155–159
neuropsychological impairments 2:155
neurotransmitter systems 2:159–160
noradrenergic system 2:159–160
other interactions with neurobiology 2:160–161
pharmacogenetics 2:160
structural connectivity 2:158
structural imaging studies 2:155–156
and procrastination 19:77
-related behaviors, and media 15:4–5
and stimulating drugs 3:882
Attention problems, and computer games 4:498–499, 4:501
Attention regulation, and temperament 24:185–186, 24:188
Attention tasks 3:571
Attention to causal indicators 11:751
Attentional alexia 6:734
Attentional bias 17:874, 22:496–497
Attentional blink (vision) 3:348
Attentional control theory (ACT) 17:874
Attentional flexibility 2:865
Attentional learning, and semantics 25:228
Attentional mechanisms, and aging 1:431
Attentional modulation, of startle 23:332
Attentional significance, cardiac defense 23:542–543
Attention-dependent input selection 16:637
Attention-dependent synchronization 16:637
Attention-shifting strategy 7:699–700
Attenuated neuromodulation 1:432
Attenuated psychosis syndrome 21:12–13
Attenuation 14:850–851
correction, classical true score model 19:425–426
hypothesis, for persuasion 17:945
of syntax 13:267
of whisker–barrel function 16:491
Attitude change 4:292
fertility 9:24
Attitude measurement 2:176–182
assimilation effects 2:178–179
behavioral observation 2:181
contrast effects 2:178–179
explicit self-reports 2:178–180
attitude scale formats 2:179
evaluating attitude objects 2:178–179
implications 2:179
question answering process 2:178–179
reducing socially desirable responding 2:179–180
implications and future directions 2:182
implicit associations test 2:180–181
implicit attitude measures 2:180–181
context effects 2:181
low tech measures 2:181
response competition measures 2:180–181
response facilitation measures 2:180
response time measures 2:180–181
priming 2:180
psychophysiological measures 2:181
rating scales 2:179
response alternatives 2:178–180
Attitude scales 2:176–177
for sexism 21:697
Attitude strength 2:185–186
Attitude theory 4:740
Attitude–behavior consistency 2:183–189
contemporary themes in research 2:186–189
deliberative and spontaneous behaviors 2:187–188
implicit/explicit attitude measures 2:188
processes underlying moderators 2:188–189
elaboration likelihood model 2:186
heuristic-systematic model 2:186
in history 2:183–184
low associations 2:184–186
attitude strength 2:185–186
measurement explanations 2:184
role of attitude in behavior 2:184–185
MODE model 2:187–188
planned behavior theory 2:185, 2:187–188
reasoned action theory 2:184–185, 2:187–188
Attitude–behavior gap *see* Ethical purchasing gap

Attitude-reinforcing content 16:796
Attitudes 2:175–177, 4:721, 12:381–382, 15:228–234, 17:334, 17:336–339, 20:356, 21:365
about drugs 6:687–688
and abusive supervision 23:702
and accessibility 2:186, 2:188–189
and adolescent health 1:124–125
ambivalence 2:186
to animals in research 1:704, 1:706
APE model 2:172–173
and certainty 2:186
change 2:170
classical conditioning 2:169–170
and conflicts 4:582–584, 4:587–588
and creativity training 5:115
culture as 9:51–52
Darwin on 2:175
definition 25:28
of drivers in transportation 24:625–626
environmental 7:801
environmental cognition 7:709–711
in Europe 15:229–230
evaluative conditioning 2:169–170
and food preferences 9:297–298
formation 2:169–170
genetic factors 9:921
in organizational research 9:921
and twin research 9:921
and habit 10:478
and health behavior 10:578, 10:580
and health behaviors 10:585–586
of homophobia 10:790
Hovland's base model 2:170–173
ambivalence 2:172
distraction 2:171
fear arousal 2:171
implicit and explicit attitude measurement and change 2:172–173
message design, elaboration, and counterargumentation 2:170
minority *versus* majority sources of influence 2:170–171
misdirection 2:172
mood 2:171–172
overload 2:171
source 2:170
and implicit social cognition 11:714, 11:716–718
importance of 2:186
job satisfaction 12:833–837
of journal authors 12:865
of LGBTI commercial sex 13:883–884
majority 8:143–144
message-learning model 2:170
metacognitive model 2:172–173
in North America 15:230
operant conditioning 2:169
and opinion formation 2:175–177, 17:243–244
parenting, across cultures 17:493
persuasion 2:169–173
and persuasion 17:943
change 17:941–945
political 18:459, 18:462–463
Second Demographic Transition 21:346–347
social *see* Social attitudes
and social psychology 22:527–530
structure 2:186
Thomas on 2:175–176
Thurstone on 2:176–177
toward disability 6:467, 6:469
toward environment and development 7:809–813
from 1800 to 1940s 7:810
after 1940s 7:811–812
after 1970s 7:812–813
toward immigration issues, and xenophobia 25:804–805
toward money 15:736–737
toward nanotechnology 16:191
toward science and science education 21:229–230
and values 25:36, 25:39, 25:41–43, 25:45, 25:49
Attitudes toward pornography (ATP) 18:616
Attitudes Toward Women Scale (AWS) 21:697
Attitudinal model 23:725
for judicial behavior 12:886–889
Attitudinal research, qualitative 19:777–778
Attitudinal restructuring 16:430
Attitudinal-value-behavioral authoritarianism 2:199
Attorney fees
arrangements for 14:267
class actions 3:733, 3:737
discovery 6:518, 6:520
Attorneys 13:629
public evaluation of 13:747–748
Attraction, interpersonal *see* Interpersonal attraction
Attraction–selection–attrition (ASA) model 17:382, 17:699
Attractiveness 24:792–793
sexual 21:760
social influence 22:350
Attractors, constraint satisfaction 4:720–722, 4:720f
Attributable risk, epidemiology 7:830
Attribute agenda setting 15:29
Attribute disclosure 4:578
Attribute knowledge 21:575–576
Attributes 10:59–61, 14:871
agenda setting 1:352–355
associated with adulthood 7:941
conjoint analysis 14:865–866
conjoint measurement 14:862
of parents 5:377
of self-concepts 21:470
perspectives on 21:470
Attributing blame to organization, relational theories 23:703
Attribution theory 4:814, 10:751–752, 15:895–896, 17:314, 21:50–51, 22:529–530
conflict 4:583
of customer satisfaction 5:631
leadership in organizations 13:665–666
of motivation 15:937
on motivation 15:936
race–crime relationship 19:843
Attributional approach
emotion and motivation 15:894–898
Attributional processes
background 2:201
future directions 2:204
Heider on 2:201–202
recent theoretical developments 2:203–204
attribution as communication 2:204
attribution as folk theory of mind 2:203–204
research after Heider 2:202–203
attribution as causal judgment 2:202–203
attribution as trait inference 2:202
Attributional processes, psychological 2:201–204
Attributions
in collectivism/individualism 4:208
explanatory style 8:598–599
heuristics 10:826–828
and motivation/actions 15:887–890
responsibility as 20:589
self-concept, and unconscious mental activity 24:723
Attributive possession 18:630–631
Attrition 5:738, 9:131, 12:452, 14:91, 14:354, 16:61
teacher, in United States 24:78–79
Attrition bias 19:738–739
ATUS *see* American Time Use Survey (ATUS)
ATUS time diaries 13:848
At-will contract 7:571
Atypical development, social referencing 22:560
Atypical employees 7:572–573
Atypical language representation, predictors of 12:693–694
Atypical lateralization 2:810
Atypical neuroleptics 19:459
AU *see* Organization of African Union and African Union (AU)
Au Bon Marché 4:753
Aubert, Vilhelm, works, *In Search of Law* 13:437
Auburn prison 18:938–939
Auburtin 18:46–47
Auction games, experimental economics 8:550
Auctioneer 2:206
Auctions 2:205–210
analyzing 2:206
application to financial markets 2:209
Bayes–Nash equilibrium 2:206, 2:208–209
common value model 2:207–208
general value model 2:208
independent information variables 2:207
strictly affiliated information variables 2:207–208
winner's curse 2:207

Auctions (*continued*)
definitions 2:205–206
double auctions 2:208–209
Dutch auction 2:205–208
empirical results 2:208
English auction 2:205–208
game-theoretic model 2:206
private value model 2:206–207
independent private values 2:206–207
strictly affiliated private values 2:207
ranking 2:207–208
sealed-bid 2:205–209
Audience 5:442–445, 24:165–167
analysis, and hegemony 10:745–746
British cultural studies 2:853
-centered effects of mass media 14:713–714
desired communication needs of 15:32
of documentary/ethnographic film 6:620, 6:623
Internet 12:590–591
cultural information regimes 12:590–591
cultural participation 12:590
voice of consumers 12:591
and language 22:950–951
media 2:216–222, 15:11, 15:15–16
anthropology 2:216
British Cultural Studies 2:216–217
feminist 2:218–219
future research directions 2:220–221
global audiences 2:219–220
new media 2:218
postmodernism 2:217–218
qualitative research 2:217–219
reality television 2:219–221
social class 2:217, 2:221
third generation research 2:217–218
of media events 15:48–52
media imperialism 15:57
media sports 14:719–724
responses to media content 21:165
scientific 21:225, 21:227
television 25:82–83
Audience discussion format 15:61–62
Audience discussion programs 15:61
Audience engagement 16:147
Audience measurement 2:211–215
advertising 2:211, 2:213
diaries 2:212–214
in digital media environment 2:212
error sources in 2:213–214
institutional aspects of 2:214
marketing 2:211
meters 2:212–214
methods 2:212–213
nonresponse error 2:213–214
ratings 2:211, 2:213–214
research 2:211–214
response error 2:213–214
sampling error 2:212–213
servers 2:212–214
TV ratings 2:214
uses of 2:211–212
Audience research, contribution of Lazarsfeld to 13:638
Audience segmentation, community-based health interventions 10:647–648
Audience tuning, and collective memory 4:190
Audio-Computer Assisted Self-Interviewing (ACASI) 15:316
Audio-lingual method (ALM) 9:327–328
Audiovisual Media Services Directive (AVMS Directive) 2:862
Audio–visual tactile maps 24:11–12
Audit
data quality 4:577–578
and risk 20:686, 20:688
Audit Bureau of Circulations 2:211
Audit studies 9:130
Audition, language acquisition 13:245–246
Auditory brainstem implants (ABI) 16:715
penetrating electrodes (PABI) 16:715
speech recognition 16:715
Auditory cortex 13:404, 23:243
development 5:32–33
and language 11:324
Auditory dorsal stream 23:245–247, 23:247f
representation of action sounds 23:246
unified function o 23:246–247
Auditory Gaussian noise 21:949
Auditory hallucinations 2:810, 10:500
Auditory input processing, by temporal lobe 24:207
Auditory midbrain implant 16:715
Auditory neuroprostheses 16:714–715
Auditory processing
and language learning impairment 13:346–347
problems, in dyslexia 6:729
Auditory scene analysis, computational models 2:223–228
Auditory ventral stream 23:244–245
Auditory verbal hallucinations 10:513
Aufklärung 15:683
Augmented reality (AR) 6:900
Ambient Wood project 24:22
Computer Assisted Instruction (CAI) 4:495
tangible user interfaces (TUI) 24:21–22
August Revolution 25:96
Augustine of Hippo 19:408, 20:285, 23:1–2
works, *City of God* 17:908
Augustinianism, personality psychology in 17:907–908
Aum Shinrikyo cult 12:797, 16:806–807, 20:370–371
Aura 10:501
Aurelius, Marcus, works, *Meditations* 19:408
Auroville 2:151–152
Auschwitz concentration camp 5:416–417
Ausnahmslosigkeit der Lautgesetze 6:348
Austerity agendas 24:491–494
and health 10:201
Austin, Alice Constance 1:218–219
urban planning 1:219
Austin, G.A. 4:104
Austin, J.L., linguistic philosophy of 10:536–537
Austin, John
legal positivism of 10:537, 13:764
utilitarian social thought of 24:982–983
works
The Province of Jurisprudence Determined 4:249
Austin-Broos, Diane 1:640
Austral 17:158–159
Australasia, Christianity in 3:540, 3:543–544
Australasian Agri-Food Research Network (AFRN) 10:76–77, 10:79, 10:81
Australia 3:540
Aboriginal Land Rights Act 13:236
Anglican Church 3:543
archaeology of 2:229–233
arid zone 2:231–232
colonization 2:229–231
dating of arrival of humans 2:229–231
Holocene continuities and changes 2:232–233
horticulture 2:231
hunter–gatherer 2:231
intensification 2:230–232
island Melanesia before Lapita 2:230
megafauna 2:231–232
Pleistocene 2:229–233
rock art 2:232–233
routes of entry of humans 2:230
Sahul 2:229–231
Sunda 2:230
Tasmania 2:232
watercraft 2:230
asylum seekers in 2:109–110
changes in religious identification 3:544
Disability Discrimination Act 1992 6:446
elder care in 7:339–340
emergency relief and social work responses 7:427–432
community 7:427–431
connectedness 7:427, 7:431
family support team 7:428–430
leadership 7:429–431
offshore disasters or terrorist events 7:430
policy responses 7:428
resilience 7:428, 7:431
strengths-based interventions 7:428, 7:431
Victorian bushfires 7:430–431
impact of family status on labor market in 8:758–759
income inequality in 22:327
indigenous mobilization and political process in 11:757–759
liberal settler societies in 14:13
religious profile 3:543–544
sexuality in 21:833–839
abortion 21:835
childbearing 21:834–835
contraception 21:834
contraceptive behavior 21:834
culture 21:833–834
homosexuality 21:837
illness 21:834

indigenous sexual health 21:836–837
sex education 21:835
sex work 21:837
sexual health 21:834
young people 21:835–836
sociocultural overviews of 2:239–244
middle classes and environmental crisis 2:243–244
middle classes and political economy 2:242–243
migration, urbanization, and multiculturalism 2:240–242
prosperity and paradox 2:239
rural change 2:239–240
tax system 23:340–342
wrongful convictions in 25:787
Australia and New Zealand School of Government (ANZSOG) 3:176–177
Australian Aborigines 8:176–177, 9:962–963, 9:965–966
and culture 2:234–238
art production and sport 2:237
economics and kinship 2:235
education 2:236–238
employment 2:238
ethnic identity and pride 2:236
European impacts and aftermath 2:234–235
future prospects 2:237–238
government policies 2:235, 2:237–238
health 2:238
indigenous minorities 2:234
land rights 2:237
language 2:236–237
origins 2:234
religion and worldview 2:236
resistance and protest 2:234–235, 2:237
social organization 2:234
sovereignty issues and relations with national and local governments 2:237
discrimination of 5:151, 5:155
indigenous management styles of 11:784–785
languages 23:889
motivation in 15:899–906
academic buoyancy 15:902–903
academic motivation of students 15:901–903
advances in theory and research 15:901
deficit models 15:899
expectancy value theory 15:900–901
nature of educational disadvantage 15:900–901
positive psychology 15:899–901, 15:903, 15:905
reciprocal effects model 15:903
self-concept research 15:903–905
totemism 24:453–457
worldviews 11:785
Australian Community Care Needs Assessment of 2007 12:504
Australian Council of Heads of Schools of Social Work (ACHSSW) 2:112
Australian Council of Learned Academies (ACOLA) 22:638
Australian Developmental Screening Test 2:850
Australian Federal Police (AFP) 7:429
Australian High Court 18:396
Australian higher education market 10:857
Australian Institute of Criminology 5:128
Australian knowledge bases 21:154
Australian languages 3:814
Australian Longitudinal Study of Aging 15:846–847
Australian National Research Council 3:6
Australian Society for Sports History (ASSH) 23:307–308
Australian SunSmart program 23:698
Australia-New Zealand Closer Economic Relations Trade Agreement (1983) 12:569
Australopithecus afarensis 2:827
Australopithecus sediba 17:458
Austria
political economy 18:360
and World War I 9:212
Austrian economics 22:696
Austrian Freedom Party (FPÖ) 18:410–411, 18:413–415, 18:610–611
Austrian School 21:128
Austrian school of economics 1:574–576
Austrian Societies for Population Policy 8:220
Austro-Hungarian monarchy 3:597, 6:812
AustroMarxism 16:223, 16:292, 19:326
Austronesians 8:208
expansion of 17:446–447
family 17:446–447
language families 11:822, 13:337–340, 13:342
languages 2:705, 8:370, 8:373–375, 13:342, 23:892
in Oceania 17:159
Auteur theory, and film genres 9:160–161
Authentic leadership 2:245–250, 13:664, 13:668
authentic followership 2:248–249
core components 2:247
balanced processing of information 2:247
internalized moral perspective 2:247
relational transparency 2:247
self-awareness 2:247
development 2:249
early conceptions 2:245–246
historical overview 2:245–246
psychological well-being 2:247–248
theoretical and empirical antecedents 2:248
theoretical and empirical outcomes 2:248–249
theory 2:246–249
authenticity 2:246–247
work engagement 2:248
Authentic Leadership Questionnaire (ALQ) 2:248
Authentic pride 21:478
Authenticity 2:246–247
authenticity 2:253
and collectivities 2:252
and constructionism 2:252–253
cultural 2:252–253
of cultural heritage 5:414–416
cultural production 5:610
of culture 5:606
endangered 2:252
food/cuisine 9:292–294
indigenous control of 2:253
and modern individualism 2:251–252
movements, in Central Africa 3:314
negotiation 2:252–253
politics of 2:253
of qualitative research 19:688
and reality 2:251
of self/person 17:737
and sincerity 2:252
and world culture 10:228–229
Authentization 16:459
Authorial intention, political thought 18:488
Authoritarian 10:313–314
broadcasting regulation 2:859–860
Authoritarian aggression 2:255, 2:257, 2:259
Authoritarian character 9:459–460
Authoritarian leadership style, abusive supervision 23:703
Authoritarian parenting 3:385, 22:219
and socialization of children 22:867
Authoritarian parents 7:439, 22:157
Authoritarian personality 2:255–261, 9:459, 19:126, 19:564–565, 22:533
Altemeyer's 2:257–258
alternative theories of 2:256–257
beyond 2:259–260
critiques against 2:258–259
original authoritarian personality theory 2:256
original theory of 2:255–256
second 2:258
Shils on 21:885
Authoritarian personality syndrome 18:429
Authoritarian political systems 18:468
Authoritarian regimes, policing and 18:267–268
Authoritarian ROL 13:432–433
Authoritarian states
law and courts in 13:427–434
science in 21:214
Authoritarian submission 2:255, 2:257, 2:259
Authoritarianism 2:197–200, 13:427, 18:388–389, 25:639, 25:641
Altemeyer's reconceptualization 2:263
and conservatism 2:263–264
criticisms 2:197–198
democratic transitions 6:98
developmental origins of 2:264–265
dictatorship 6:373, 6:375–376
early criticism 2:262–263
and environment in Soviet Union 20:832, 20:834
Fromm on 9:459–460, 9:462
and gender 2:265–266
as group phenomenon 2:266–267
and information processing 2:265
in international perspective 2:267

Authoritarianism (*continued*)
and just world belief 2:487–488
on left 2:264
limitations and problems 2:198
origins 2:197, 2:262
and personality 2:265
and racism 19:866
and personality 17:803–804
recent approaches 2:198
religiosity and prejudice against minorities 20:381–382
right-wing authoritarianism
scale, measurement 2:263
versus social dominance orientation 2:264
and social and behavioral sciences 22:140, 22:144
and threat 2:266
versus totalitarianism 24:448–452
and well-being 2:265
Authoritative parenting 3:385, 22:157
and socialization of children 22:867
Authoritative parents 7:439
Authorities 3:29, 16:323
American theories 2:271
and anarchism 1:680–683
Bendix on 2:512–513
British theories 2:270
bureaucratic 2:913–914
and conservatism 4:689
culture and 2:273
financial supervision 9:191, 9:193–194
French theories 2:269–270
global governance 10:160
Italian theories 2:270
and legal reasoning 13:776–779, 13:781–782
legitimacy and 2:272
legitimacy of 13:825–827
obedience and 2:272–273
in organizations 17:405–410
critical approaches 17:407–408
and domination 17:407–408
Foucault on 17:408–409
managerialist approaches 17:405–407
organizational discourse 17:409
reconstruction 17:408–409
pervasiveness of obedience to 17:69
and power 2:272–273, 10:294–295, 18:765–766, 18:768
rationalization of 15:697
religion and new media 20:273–274
in religious organizations 20:408–410
forms of governance 20:409
sources 20:409
social influence 22:350
of state, source of 11:817
theories and ideologies 2:269
Weberian concepts 2:271–272, 25:459, 25:461
charisma 2:271–272
rational-legal authority 2:271
traditional authority 2:271
Authority constraints 15:638
Authority maintenance rituals, in police use of force 18:258–259
Authority maintenance theory 18:258
Authority of scientists 21:261–262, 21:266
Authority-based policy instruments 19:572
Authorized heritage discourse (AHD) 5:415
Authors
as a fly in the soup 6:622
as a fly on the wall 6:622
postcolonial 'other' as 6:622–623
as propagandist 6:622
as reconstructing authentic other 6:621
reflexive 6:623
as reflexive roving eye 6:621–622
as social educator 6:621
Authors (journal) 12:864
behavior, perceptions, and attitudes 12:865
Authors (university press) 19:626–629
Authorship 24:258
and authority 5:505
influence on 7:587
Autism 3:329–330, 12:420–421, 15:212, 18:821, 19:304, 23:556–557, 25:189
anthropology of 22:857–858
like me theory 15:133
and mirror neurons 15:582–583, 15:585
nature/nurture 16:343
social relationships 22:857
theory of mind deficit in 15:541–542
Autism spectrum disorders (ASDs) 2:275–281, 2:810, 3:402, 9:444, 10:513, 11:625, 12:27, 12:29–30, 18:790, 22:560
Asperger's disorder 2:275
causes 2:277
child psychiatry 2:276–277
clinical features 2:276–277
co-occurring conditions 2:276–277
restricted, repetitive patterns of behavior, interests, or activities 2:276
social communication and social interaction 2:276
clinical management 2:278–279
diagnosis and assessment 2:278
treatment 2:278–279
cultural issues 2:279
current issues and directions in research 2:279–280
developmental psychology 2:275
disability 2:275, 2:279
effects on sign language use 21:938–939
epidemiology 2:275–276
evolution of concept 2:275
human rights 2:279–280
intellectual disability 2:275–278, 2:280
joint attention in children with 12:846
language delay in 13:347
neuroscience 2:275
onset and developmental course 2:277–278
rights/policies/advocacy issues 2:279
social cognition 2:277
social intelligence 2:275
and social referencing 22:560
special education 2:278–279
Autistic spectrum conditions 12:420–421
Autobiographic memory *see* Episodic memory
Autobiographical amnesia 24:554–555, 24:558–559
Autobiographical insideness 6:886
Autobiographical memories 1:646, 4:560, 4:566–567, 5:564, 24:565–566
changes in childhood and adolescence 21:898
cultural differences in 2:286
definition 2:282
development of 2:289–295, 15:147, 15:149, 15:151–153
distribution of memories across life span 2:284–285, 2:284f
and early memory development 6:761
emergence in preschool years 21:898
episodic memory 2:282, 2:286
example 2:282–283
false memories 8:709, 8:712
imagery 2:285–286
incidental autobiographical memory 21:898
involuntary memory 2:286
loss of 24:560
measurement 2:283–284
neuroimaging 2:286
olfactory cuing of 2:285
overgeneral memory and depression 2:285–286
phenomenological properties of 2:285
Proust on 2:282–286
questionnaires 19:773
reminiscence bump 2:284–286
and self 21:422
Autobiographical Memory Interview (AMI) 24:559
Autobiographical narratives, mother–child coconstructions of 15:877
Autobiographies 6:495–496, 15:136, 20:95
Autocannibalism 3:86
Autocorrelation, computational auditory scene analysis 2:223–226
Autocovariances, and spectral density, relationship between 24:332
Autocracies 6:368
media in 22:584–585
Autocratic societies, and group violence 14:704
Autoethnographies 6:495–496, 12:674, 17:532, 19:677–678, 22:539
qualitative developmental psychology 19:672
Autogestion 20:675–676
Autoimmune disorders 20:271
Automated deduction 5:935–938
artificial neural systems 5:938
clausal form, resolution, tableaux, and connection method 5:935–936
complexity, systems, and performance 5:937–938
first-order features 5:936
variants, specializations, and extensions 5:936–937

Automated Multiple-Pass Method (AMPM) 10:589
Automated performance measurement 24:100–101
Automated reasoning 5:935
Automated theorem proving, reasoning for 5:938
Automatic and impulsive behaviors 10:624
Automatic appraisal mechanism 7:468
Automatic cognition 7:416–418
Automatic control 2:296
Automatic Control of Thought (ACT) 15:162
Automatic determinants 10:623–624
Automatic evaluation effect 2:328
Automatic Milking Systems (AMSs) 1:474
Automatic passive resonance (inferences) 19:396
Automatic processing 12:79
  in abnormal illness behavior 11:599
Automatic programming, genetic programming 2:41
Automatic speech attribute transcription (ASAT) 23:263
Automatic telling machines (ATMs) 21:651
Automatic thoughts 4:542, 4:730–731
Automaticity
  in mathematics learning 4:67–68, 4:68t
  mindfulness 15:549–550
  and prejudice 18:835–836
Automation
  applications 2:296–298
  in banking sector 2:297–298
  concept of 2:296
  and division of labor 13:159
  essentials of 2:296
  of human work 21:187
  human-centered design 2:296
  impact on composition of the workforce 2:299–300
  of office 2:297
  organizational studies 2:296–302
  for professionals 2:297
  risks of 2:300
  role of the human operator and 2:300
  in services 2:297
  social effects of 2:298–300
Automatism 5:221–222
Automobile cluster 11:830
Automobile dependence, sprawl 24:934–936
Automobile mismatch 23:158
Automobiles, urban settlements 24:836–837
Automobility 2:303–308, 15:637–638
  automotive emotions 2:306
  complexity theory 2:303, 2:305
  cultural geography 5:412
  future of 2:306–308
  identity 2:305–306
  living 2:305–306
  peak travel 2:306–307
  politics 2:303–305
  private car 2:303–304, 2:307
  sociotechnical system 2:304, 2:307
  subjectivity 2:305, 2:307
  system/regime 2:303–305
  transport policy 2:306
  transportation 2:303–307
Automorphism group, of ordered relational structures 17:298–299
Automorphisms, representational theory of measurement 14:859
Automotive emotions 2:306
Autonoetic awareness, and episodic memory 7:856, 7:858
Autonoetic consciousness, and memory development 15:151
Autonomic disorders, and alcohol 1:511
Autonomic measures, of emotions 7:521
Autonomic nervous system (ANS) 14:126–128, 19:445–446
  and hypertension 11:454–455
Autonomic side effects 19:464
Autonomization 24:851
Autonomous
  in democracy 6:85–86
  moral orientation of 15:775
Autonomous agents 2:309–313
  algorithms 2:309–310
  anthropomorphic and mentalistic concepts 2:310
  applications for 2:310
  construction 2:309
  cooperating agents 2:310
  definition 2:309
  design principles 2:309–310
  in electronic trade 2:311
  emergent behavior 2:310
  ethical considerations 2:311–312
  implementation 2:310
    and biological entities 2:312
  information filtering and access 2:311
  multiagent systems 2:310
    as models of the behavior of social systems 2:312
  ownership of personal data and privacy 2:312
  performance evaluation 2:311
  reactive system 2:309
  and robots, distinction between 2:309
  simulators 2:311
  subsumption architecture 2:310
  in user interfaces 2:311
  user–agent communication 2:311
Autonomous functionalism 9:472
Autonomous Lesbian Feminist Group 13:860
Autonomous motivation 21:486–488, 24:68
  FTP and FO 9:563, 9:565
Autonomous speech 10:507f
Autonomous-related self, and parenting 17:495–497
Autonomy 2:314–319, 3:483, 9:732, 9:909–910, 10:522–523, 11:765, 15:931, 21:487–488, 21:586, 24:68, 25:734–736
  and adolescent leisure activities 13:832
  in adulthood 1:181–184
  of artists in China 5:600
  and bioethics 13:557–560, 13:563
  in close relationships 21:490
  definition of 1:181
  ethics of 13:5–8
  in genetic testing 20:699–700
  informed consent 12:110–111, 12:113–114
  of legal system 13:567
  of migration 1:686
  and morality 15:807–809
  professions 19:95–98
  public sector organizations 19:599–601
  of religions 3:575–576
  and research ethics 20:510
  of self-organizing dynamical systems 21:529
  social work in Russia 22:810–812
  of society 22:553–554
  in teaching 24:77
  treatment research and access 8:33
  workplace 25:718–721
Autonomy, philosophy of 2:314–315
  dialogical approaches 2:318
  external moral 2:318–319
  Kantian conception of 2:315
  neo-Kantian theories 2:317–318
  procedural theories of 2:315–317
    authenticity as historical nonalienation 2:316–317
    endorsement 2:316
    identification 2:316
    self-discovery 2:317
  self-regarding attitudes 2:318
  social conditions 2:318–319
Autonomy claim 15:948
Autonomy rights, of children 3:488–489
Autonomy support 21:489–490
Autonomy system 15:948
Autopoiesis 9:541, 12:342, 13:554, 14:390–395, 21:530, 22:914–915
  capacity of an entity 13:554
  evolution of social systems and 13:555
  functionally differentiated society and 13:554
  in law 13:554–556
  of legal system 13:554–555
  normative closure and cognitive openness 13:554–555
  in social systems 14:386
  theory 13:552
Autoproduction 5:610
  of culture 5:605–606
Autoregression formulation 24:323
Autoregressive conditional heteroscedastic (ARCH) models 24:312–313, 24:338
Auto-regressive fractionally integrated moving average (ARFIMA) model 24:365
Autoregressive integrated moving average (ARIMA) models 5:652, 9:320–321, 17:732, 24:316–321, 24:335, 24:337–338, 24:347–348, 24:351–353

Autoregressive integrated moving average (ARIMA) models (*continued*)
components 24:356–357
intervention analysis and outlier detection 24:319–320
model building 24:318–319
objectives 24:316
seasonal models 24:319
time series with heterogeneous variances 24:320
in time series quasi-experiments 12:652–655
Auto-regressive (AR) model 24:362
AR(p) models 24:364–365
Autoregressive moving average (ARMA) models 23:260, 24:317–318, 24:341, 24:343
Autoregressive spectral estimators 24:335
Autorità per le granzie delle communicazioni (AGCOM) 5:705
Autoscopic hallucinations 10:499
Autosegmental representations 18:78, 23:717
Autosomal genetic admixture
at individual level 9:888–889
complexity of human demographic history 9:890–894
sex-biased admixture 9:889–890
at population level 9:887–888
Autosomes, Genographic Project 10:23
Autotelic personality 9:262
Autotrophic organisms 7:627
Auxiliary residuals 24:359
Availability heuristic 10:825–826
of decision making 5:919
Availability problem, in visual word recognition 25:215
Available-case analysis 13:389
Avant-garde, modernity 15:682–683, 15:685
Avant-garde art/artists 2:320–325
art worlds 2:322–323
Bourdieu on 2:322–324
contemporary relevance of 2:324
culture 2:320–324
Duchamp on 2:320
Frankfurt School 2:321
origin and evolution of term 2:320–321
Outsider art 2:322
and popular culture 2:323–324
Postmodernism 2:321–322
sociological studies 2:322–323
theories 2:321–322
transgressive art 2:321
Avant-garde markets 14:566
Average causal effect (ACE) 3:253, 8:528
bounds using instrumental variable 8:529
Average income differentials, international migration 12:513
Average lexicograhic value 4:871
Average linkage 23:393
Average theoretical RT 12:78–79
Average treatment effects (ATE) 9:131
estimators 24:707
in randomized controlled trials 8:560
Average weekly hours (manufacturing) 9:317
Average weekly jobless claims for unemployment insurance 9:317
Averaging, risk 20:707
Aversion, and cognitive dissonance 4:76–78
Aversion therapy 3:766–767
Aversive leadership 23:704t
Aversive motivation 2:327, 2:329–330
Aversive racism 10:759–760, 18:842, 19:865
and prejudice 18:834
Avery, Sewell S. 7:375
AVHRR *see* Advanced Very High Resolution Radiometer (AVHRR)
Avian flu 18:174
Avineri, Shlomo, on Middle East society 16:382
AVMS Directive *see* Audiovisual Media Services Directive (AVMS Directive)
AVN Media Network 18:614
Avoidance 24:642
behaviors 2:467
and conflict 4:622
in development of psychopathology 3:766
and emotions 7:523
Avoidance behavior, in panic disorder/agoraphobia 17:475–476
Avoidance coping
with cardiac disease 3:113
and gender 9:705–706
Avoidance motivation
see Approach–avoidance motivation
Avoidance training 2:408–409
Avoidant attachment by infants 16:470, 22:566–567
AVP *see* Arginine vasopressin (AVP)
Awards 2:332–337
honorary degrees 2:335–336
international 2:334–335
typology 2:333–334
Awareness 4:673–674
of death/dying 5:866
of language 2:583–584
of newborn 16:482–483
and procedural learning 16:848–849
of self 21:440–441
detecting of I 21:440–441
feeling self-consciousness 21:441
see also Self–awareness
tactile, higher order processing of 24:15–16
Awas Tingni community, indigenous rights of 11:799
Awle 8:880
AWS *see* Attitudes Toward Women Scale (AWS)
Aws tribe 12:732
Axel Springer Germany (2012) 55 EHRR 6 18:950–951
Axelrod, R., works, *The Evolution of Cooperation* 22:512–513, 22:702
Axial age 20:313–314, 20:388
civilizations 3:726, 7:323, 20:314–315
expansion of 3:726–727
internal transformation 3:728
multiplicity 3:727–728
Axial partitioning, MDS 16:44f
Axial religions 2:511
Axiological neutrality 16:755
arbitrariness 16:757–758
axiological commitments 16:757
evaluation 16:757–758
is/ought separation and value exclusion from science 16:755
untenable fact/value separation
logical grounds 16:755
practical grounds 16:755–756
value commitments 16:757
value freedom 16:756–757
value judgments 16:755
value neutrality 16:757–758
Wertfreiheit on 16:757
and working scientists 16:757–758
Axiological rationality 12:395
and interests 12:394–395
Axiology 8:76
Axiom of quality 5:558
Axiom of rationality 3:358
Axiomatic approach 5:895, 5:897
Axiomatic equilibrium analysis 7:915–916
Axiomatic expected utility theory 7:89
Axiomatic foundations of probability theory, by Kolmogorov 19:33–34
Axiomatic reasoning 13:776–779
Axiomatic theories 2:338–344
of choice 24:679–680
Greek geometry 2:338–339
Hilbert's 2:339
historical background 2:338–339
axiomatic geometry in nineteenth century 2:339
ingredients of standard axiomatic practice 2:339–340
axioms as defining theories 2:340
independence of axioms 2:340
Padoa's principle for proving independence of primitive concepts 2:340
primitive concepts 2:339
isomorphism of models 2:340–341
invariance theorems for axiomatized theories 2:341
quantum decision theory 2:341
mathematical models 14:792
Pasch's 2:339
set-theoretical models 2:338, 2:340–342
subjective probability 2:338–339, 2:341–342
subjective utility 2:338–339, 2:341–342
theories with standard formalization 2:341–343
negative results about axiomatizability 2:342–343
positive results about axiomatizability 2:342
zero-sum games 2:339, 2:341

Axiomatic utility theory 7:88
Axiomatizability
negative results about 2:342–343
positive results about 2:342
Axioms 4:872–877
basic 4:872–873
conjoint measurement 14:862–863
independence and cancellation conditions 14:862–863
ordering 14:862
solvability and Archimean axioms 14:863
as defining theories 2:340
dummy player 4:872–873
independence of 2:340, 4:874
inequality 11:902–904
invariance 4:874
monotonicity 4:873–874
nullifying player 4:873
punctual 4:872
relational 4:872
representational theory of measurement 14:858
trivial-game 4:872
variable-population 4:875–877
Axis mundi 20:345
Axis powers 21:371
Axon 16:703–706
Axtell, R., on artificial societies 1:368–369
Ayamara linguistic community 13:298
Ayer, A.J. 6:593
Ayer, Alfred Jules 18:626
works, *Language, Truth and Logic* 6:488
Ayittey, George B.N. 4:575
AYP *see* Adequate Yearly Progress (AYP)
Ayurveda 10:568–569, 21:618
Ayutthaya 13:813–814
Azande kingdoms, ethnographic research of 8:290–291, 8:293
AZT *see* Zidovudine (AZT)
Aztec Empire 4:574
Aztec Triple Alliance 15:262–263

B

BA *see* Basal (BA) of the amygdala; Behavioral archaeology (BA)
Baba, Sathya Sai 3:351, 16:806–807
Baba, Shirdi Sai 3:351
Baba Sali 24:605
Babad 11:83
Babad Ing Sangkala 11:83
Babad Tanah Jawi 11:83
Babbage, Charles 4:515–517, 6:602
on division of labor 6:603–604
Babbling 13:246, 21:943
Babelian myth 6:351–352
Babkin, Boris 10:732
Baby boomers 9:866–867, 25:815
cohorts 13:855
and youth culture 25:815
BAC *see* Blood alcohol concentration (BAC)
Baccalauréat 22:325
Bacchanalia 3:145
Bach, Johann Sebastian, works, *Matthew Passion* 17:908
Bach, Lee 21:971–973
Bachelard, Gaston 14:383
Bachelor of Science in Nursing (BSN) 16:910
Bachelors 21:687–688
Bachelor's degree, Bologna Process 2:742–745
Bachofen, Johann Jakob 1:766, 13:802
on division of labor 6:601
Bacillus thuringiensis 1:475
Back projection 6:150
Back propagation 15:964, 16:117
Background expectancies 9:601
Background knowledge (BK) 25:755
Backlash, in sexism 21:698
Backpropagated action potential (bAP) 16:706
Backpropagation 4:86
concept learning 4:655–656
Backpropagation algorithm 2:37
Back-to-the-city movement *see* gentrification
Backtracking 7:484
constraint-satisfaction problem 4:718
Back-translation 5:320
Backward citation rate 20:499
Backward induction 9:586–589
Prisoner's Dilemma 18:934
Backward induction argument 19:902
Backward inferences 12:39, 19:395
Backward inverse projection 6:150–152
Backwardness, world systems theory 25:749
Backwards martingales 14:630
Bacon, Francis 11:573, 14:254, 15:576, 19:409, 21:213
works, *New Atlantis* 14:254
Bacon, Lloyd, works, *The Singing Fool* 9:168
Bacon, Roger 20:285
BACON (data discovery tool) 21:294
Baconianism 23:690–691, 23:693
Bacon-Smith, Camille, on media audiences 25:82–83
Bacteria resistance 8:454
Bacteriocin gene 13:30
Bad death 1:688
versus good death 5:872–873
Bad enough parenting 3:412
Bad science 8:927
BADC *see* Black Action Defense Committee (BADC)
BADD *see* Bothered About Dungeons and Dragons (BADD)
Baddeley, Alan 25:715
cognitive modules 3:328–329
Baehr, P.R., works, *The Protestant Ethic and the Spirit' of Capitalism* 21:389
Bagby, Philip 20:396
Bagehot, Walter 5:688, 22:697
works, *English Constitution* 2:270
Bagging 19:142, 23:415
Bahamas 1:638
Bahrain, legal system in 15:363
Bailey, Abigail 13:471
Bailey, F.G. 1:768–769, 2:871
on brokers 2:870
works
Strategems and Spoils 2:871
Bailment 19:195
Baker, A.R.H., works, *Geography and History* 11:21
Baker, Lee 17:22
on race in North America 17:23
Baker, Ted, on bricolage 17:393–394
Bakhtin, Mikhail Mikhailovich 10:745, 15:292
Bakr, Abu 12:732
Bakufu, overthrow of 15:126–127
Bakunin, Mikhail 1:680–681, 22:837, 24:728–729
Bal, Mieke 25:209–210
Balance
and factions 8:685
macro/microsociologies 14:414–415
vestibular system 25:63, 25:67–69
Balance as bias thesis 3:856
Balance norm 15:20
Balance of power 2:347–348, 16:255
alliance 2:345–348
and appeasement 1:826, 1:828–829
background 2:345–346
Concert of Europe 2:346
global successors 2:346–347
Congress of Vienna 2:346
and deterrence 6:247
diplomacy 2:345, 2:347
in East Asia 6:791
and economy 2:347–348
empire 2:345–347
in Europe 2:346
and nationalism 2:346
of nations 2:346–347
and peace 2:345–347
permanent feature of international politics 2:348
and realism 20:26–29
and war 2:345–348
Balance theory 10:751–752, 23:605–606
in social network analysis 16:525, 16:530–531
studying adjustment toward balance 23:607–608
Balanced corpus 4:950
Balanced crossover designs 8:538–539
Balanced processing of information, and authentic leadership 2:247
Balanced reciprocity 7:15
Balanced risk awareness, and societal impact assessment (SIA) 22:876
Balanced states, structural balance theory 23:606–607
Balance-of-power formulations 1:556–557
Balancing dilemma 3:678
Balancing selection (personality) 17:900–902
Balancing strategy, appeasement 1:829
Balandier, Georges 1:767–768
Balázs, Béla 15:15
Baldassar, L., on migration in Australia 2:241

Baldassarri, D., on social network analysis 5:560
Baldwin, James, works, *Giovanni's Room* 21:742
Baldwin, James Mark
  on circular reactions theory 16:483–484, 16:487
  on infant intentionality 12:341, 18:114–115
  on neonatal imitation 16:482–484
  works
    Dictionary of Philosophy and Psychology 7:592
Bale, C., works, *Guide for the Modern Bear* 13:891
Bales, Robert Freed
  on small group research 22:101–103
  works
    Interaction Process Analysis 22:102–103
Balfour, Colonel 1:612
Bali, Indonesia, homosexuals in 9:638
Balibar, Etienne 16:285
Balinese calendars 24:407
Balinese cockfight 24:292–293
Balint, Michael 19:504
Balkan Slavic 8:354
Balkanization 5:493, 18:104, 22:795
Balkans 1:955
  language contact 13:303–306
  and World War I 9:212
Ballew v. Georgia 13:768
Ballot 7:349
Ballot propositions, in direct democracy 25:293–296
Ballot structure 7:349
Balloting, proportional representation 19:207–208
Ballroom drag 13:876–877
Baltes, Paul B. 2:349–352
  behavioral science, contributions to 2:351
  honors 2:350–351
  life 2:349–350
  as scientist and scholar 2:350–351
    generativity 2:350
    institutional initiative 2:350–351
    interdisciplinarity 2:350
    internationality 2:350
    prudence 2:350–351
    societal interest 2:350
  selection, optimization, and compensation (SOC) theory of 1:435
  social science, contributions to 2:351
Baltic coast 15:480f
Baltic subgroup 11:824
Baltics 1:956–957
Balzac, Honore, works, *La Comédie Humaine* 7:876
Bamberger, Clinton 18:735
Bamyeh, Mohammed, works, *The Social Origins of Islam* 16:384–385
Banaba 17:158
Banach-space geometry 14:633
Banaji, Jairus, on agrarian political economy 1:457
Banaji, Mahzarin R.
  on cognition 7:416, 7:418
  on social network actors 5:559
Banal evil 1:963, 17:63–64
Band society, hunting/gathering societies 11:429
Band wagon 25:397
Bandalier, Adolf 1:743
Bandelj, Nina, on cultural wealth of nations 5:539
Bandit problems
  sequential experiments 21:636–637
  stochastic optimization of 23:477
Banditry 14:422–423
Band-pass filters 24:334–335
Bandpassed noise (BPN) 23:244–245
Bandura, A.
  self-efficacy model 8:396, 21:511
  and social cognitive theory 8:731–732
  social learning theory 22:816
Bandwagon effect 4:696
Bandwagoning 1:556
Bangladesh 12:729, 23:61
  archaeology of 23:50–57
  education in 7:195
  fertility transition in 9:80–83
  sociocultural aspects *see* South Asia; sociocultural aspects of
  urban policies in 24:898
Bangladeshis, in British *see* British Asians
Banishments, and punishment 19:659
Banister, H., works, *A Handbook of General Experimental Psychology* 7:591
Bank Wiring Observation Room (BWOR) 3:4–5
Banking
  automation 2:297–298
  development of 19:180
Banking crises 9:184–185, 9:188–189
Bankoku Taiji policy 15:127
Bankruptcy 2:353
  equality principle of 2:353
  fresh-start principle of 2:353–354
  law, content of 2:354–356
    equality principle implementation 2:354–355
    fresh-start principle implementation 2:355–356
Bankruptcy Act 1542 3:32
Bankruptcy law 3:32
Banlieues 4:525–526, 4:528
al-Banna, Hasan 24:602–603
Banter 11:417
Bantu 3:312–314, 6:773–776
Bantu kingdoms 6:774
Bantu languages 21:647
  gender assignment in 9:796
Banya 21:741
Bao concept 25:31
BAP *see* Body Action and Posture coding system (BAP)
bAP *see* Backpropagated action potential (bAP)
Baptism 16:185–186
Baptismal theology 3:237
Baptist community 13:470
Baptist Missionary Society 15:614–615
Baptists 3:542
Bar charts 10:351–353, 10:356, 10:360–364, 10:368, 16:540, 10:351f–352f, 16:540f
Barack, Obama 4:155
Barbados 1:640–641
Barbarian Europe 8:272
Barbarism 3:721
  Sarmiento on 20:908, 20:910
Barber, Bernard 17:12
Barber, James David, works, *Presidential Character* 18:861
Barber, Sir Michael 22:637
Barbiturates 6:674, 6:685
Barclay, Harold, works, *People Without Government* 1:682
Barclay, William 9:388
Bare life theme, and governmentality 10:310
Bare normativity of meaning (BNM) 14:843–846
Barebone's Parliament 7:649
Barfield, Thomas, works, *Dictionary of Anthropology* 7:589–590
Bargaining 4:805–806, 9:578, 13:606, 16:431, 16:433, 16:435, 21:336
  and exclusion 16:415–416
  and jointness 16:415–416
  models 23:589–590
    accepting union's wage demand 23:589
  negotiated orders 17:411–412
  organizational aspects 16:415–417
    to create institutions 16:417
    diversity 16:415–416
    future research directions 16:417
    institutional designs for different goods 16:416–417
    metaregimes, regimes, and international institutions 16:416
    mixed-motive situations 16:415
    pure conflicts of interest 16:415
    and role of institutions 16:416–417
    and types of goods 16:415–416
    zero-sum situations 16:415
  and power 15:667–669
  role of lawyers in 16:418–422
  trade unions 24:500–501
    power 24:498
Bargaining power 4:811, 15:73
  inequalities of 18:228–229
  plea bargaining 18:226–231
Bargaining sets (cooperative game theory) 4:869, 4:877
Bargaining theory 7:90
Bargaining value 9:579
Barker, Roger 10:751
Barker hypothesis 14:67–68
Barlett, Sir Frederic 4:189
Barley, Nigel, works, *The Innocent Anthropologist* 11:417
Barley, Stephen R. 17:393
  on organizations and culture 17:390
Barlow's unified protocol 4:546
  awareness and tolerance of physical sensations 4:546

cognitive flexibility 4:546
emotion-driven behaviors 4:546
exposure exercises 4:546
present-focused emotion awareness 4:546
Barnabas 3:548
Barnard, Chester 21:971, 21:973–974
Barnard, Jaco, on sexual minorities in South Africa 21:749
Barnes, J.A. 16:562
Barnes Akathisia rating scale 19:463
Barnett, Bernice 12:659
Baroja, Julio Caro 3:150
cultural history, contributions to 3:150–151
cultural sociology, contributions to 3:150
education and training 3:150–151
in Sahara 3:152–153
in Spain 3:152–153
study of Basque and Navarrese culture 3:153
study of culture applied to Iberian peninsula 3:151–152
works
Análisis de la cultura: Etnología, historia, folklore (*Cultural analysis: Ethnology, history, folklore*) 3:151
Biografías y vidas humanas (*Biographies and human lives*) 3:153
De la superstición al ateismo (*From superstition to atheism*) 3:152–153
El carnaval (*The carnival*) 3:152–153
El estío festivo (*The festive summer*) 3:152–153
El laberinto vasco (*The Basque labyrinth*) 3:153
El mito del carácter nacional (*The myth of national character*) 3:152–153
Estudios mogrebies (*Maghreb studies*) 3:152
Estudios saharianos (*Sahauri Studies*) 3:152
Estudios Vascos (*Basque studies*) 3:153
Etnografía historica de Navarra (*Ethnografic history of Navarre*) 3:153
Inquisición, brujería y criptojudaismo (*Inquisition, witchcraft and crypto-Judaism*) 3:152
La aurora del pensamiento antropológico (*The golden age of anthropological thinking*) 3:153
La casa en Navarra (*Homes in Navarre*) 3:153
La estación de amor (*The season of love*) 3:152–153
La vida rural en Vera de Bidesoa (*Rural life in Vera de Bidesoa*) 3:151
Las falsificaciones de la historia en relación con la de España (*Historical falsifications in relation to Spain*) 3:153
Las formas complejas de la vida religiosa (*Complex forms of religious life*) 3:152–153
Los Baroja (*The Barojas*) 3:153
Los Brujas y su mundo (*Witches and their world*) 3:152
Los fundamentos del pensamiento antropológico moderno (*The foundations of modern anthropological thought*) 3:153
Los hombres y sus pensiamientos (*Men and their thoughts*) 3:153
Los judios en la España moderna y contemporanea (*Jews in modern and contemporary Spain*) 3:152
Los pueblos de España (*The peoples of Spain*) 3:151
Los pueblos del Norte de la Peninsula Ibérica (*The northern people of the Iberian, Pensinsula*) 3:151
Los Vascos y su historia a través de Garibay (*The Basque and their history according to Garibay*) 3:151, 3:153
Materiales para una historia de la lengua vasca en su relación con la latina (*Documents for a history of the Basque language's relation to Latin*) 3:151
Razas, pueblos y linajes (*Races, peoples and lineages*) 3:152
Ritos y mitos equivocos (*Rites and false myths*) 3:152–153
Ser o no ser vasco (*To be nor not to be Basque*) 3:153
Terror y terrorismo (*Terror and Terrorism*) 3:153
Tres estudios etnograficos relativos al País Vasco (*Three ethnographic studies related to the Basque Country*) 3:151
Vidas magicas e inquisición (*The lives of magicians and the inquisition*) 3:152
Barometer series of data collections 5:833–834
Barometer surveys 5:793, 5:848–849
Afrobarometer 5:849
AmericasBarometer 5:849
Arab Democracy Barometer 5:849
Asian Barometer Survey 5:849
Eurobarometer 5:848–849
Latinobarometer 5:849
New Democracies Barometer 5:849
Barone, E. 17:504
Baroquization of military technology 24:142–143
Barracks, during Napoleon era 25:364
Barraclough, Geoffrey 4:764–767
works
Introduction to Contemporary History 4:764–766
The Origins of Modern Germany 4:764
Barratt, Alfred 7:114
Barrel cortex 16:489–491
Barrett, Elizabeth, works, *Stranger with a Camera* 6:621
Barrierhood 16:4
Barriers to entry 14:550
Barriers to insulin treatment questionnaire (BIT) 6:338t
Barrios 4:525–526
and urban poverty 24:915–916
Barron, F., works, *No Rootless Flower* 5:113
Barrows, Harlan H. 1:846
on human ecology 22:266
Barry, James 25:623
BARS *see* Behaviorally anchored ratings scales (BARS)
Barter, and exchange 8:478, 8:480
Barth, Fredrik 1:768–769, 2:871, 8:139
on ethnicity 8:149–151
on knowledge 13:73
Barthes, Ronald 2:357–362, 5:499–500
communication studies of 2:357
on consumer society 2:361
on continental philosophy of language 2:359, 2:361
contribution to visual sociology 2:357, 2:360
cultural studies of 2:361
formative years 2:357–358
influence in social and behavioral sciences 2:361–362
on linguistics 2:357–359, 2:361
popular culture studies of 2:357–358, 2:361
on poststructuralism 2:360–361
semiology and myth of 2:358–359
on semiotics 2:357–362
and social theory 2:357–358, 2:361
on sociology of fashion 2:357, 2:360–361
on sociology of leisure 2:361
on structuralism 2:359–360
works
Camera Lucida: Reflections on Photography 2:360
Elements of Semiology 2:358–360
Empire of Signs 2:360
The Fashion System 2:359–361
Image-Music-Text 2:360
A Lover's Discourse: Fragments 2:360–361
Michelet 2:361
Mourning Diary 2:360–361
Mythologies 2:358
The Neutral 2:360–361
On Racine 2:359–360
The Preparation of the Novel 2:360–361
The Representation of Forms 2:361–362
Roland Barthes by Roland Barthes 2:360–361
The Rustle of Language 2:359
The Structuralist Controversy 2:360
Writing Degree Zero 2:358–360
Bartholomew, D.J., works, *The General Linear Latent Variable Model* 8:693
Bartlett, Frederic 4:194
Bartlett's test 14:147
Bartok, Bela 1:754
Barton, Russell 6:48
Barts Explanatory Model Inventory (BEMI) 10:567–568
BAS *see* Behavioral activation system (BAS); Behavioral approach system (BAS)
Basal cell carcinoma 23:696–697
Basal forebrain, and wake 22:89
Basal ganglia 23:252
and nondeclarative memory 16:847–849
Basal (BA) of the amygdala 8:870
Basal radial glial cells (bRGCs) 2:821–822
Bas-Breton 16:298
Bascom, John 10:495
BASE *see* Berlin Aging Study (BASE)

Base and superstructure, determinism and 6:242
Base of the pyramid approach (BoP) 22:171
  defined 22:172
Base of the pyramid/inclusive business (BoP/IB) 22:168, 22:171–172
Basel Mission 15:614–615
Baseline measurements, in crossover designs 5:349, 5:351, 5:353
Baseline model, rational choice theory 19:901
Base-rate neglect 12:720
Bases, military 15:486–487
Base-superstructure model 10:744
Bashfulness, and self 21:441
Basic color terms (bcts) 4:232
  Berlin and Kay's definition 4:232–233
    practical problems 4:233
    theoretical issues concerning 4:233
Basic communism 16:138–139
Basic demography *see* Academic demography
Basic distribution theory 17:291–292
Basic Educational Movements 14:33
Basic emotion models 7:461
Basic emotion theory 7:528–531
Basic emotions 5:543–544, 6:747, 7:484
  affect program theory of 7:468
  facial and vocal expressions 5:543
  physiological response patterns 5:543–544
  as universal building blocks 5:544
Basic encounter group 20:738–739
BASIC I.D. (multimodal therapy framework) 16:72–73
Basic income guarantee (BIG) 18:749
Basic Interpersonal Communicative Skills (BICS) 2:583
Basic needs, in human development 6:310–311
Basic propositions 19:527
Basic statistical model 10:836
Basic v. Levinson 3:734–735
Basic/academic demography 1:839
  subsets of 1:839
Basic-level hypothesis, of semantics 13:934–935
Basiland Cassianto (monk) 15:725
Basilect 18:133–134, 18:136–137
Basilectalization 18:134–135
Basket of goods 18:723
Basolateral amygdala (BLA), interoception 12:618
Basque 3:184, 3:186
  anthropology, Baroja on 3:150, 3:152–153
BASR *see* Bureau of Applied Social Research (BASR)
BASRC *see* Bay Area School Reform Collaborative (BASRC)
Bastard feudalism 1:966–967
Bastard Nation 1:144
Bastian, Adolf 1:766
Bastian, Henry Charlton 16:683
BAT *see* Bilingual Aphasia Test (BAT)
Batatu, Hanna, works, *The Old Social Classes and the Revolutionary Movements of Iraq* 16:383
Batch data evaluation, in stochastic optimization 23:474
Batch mode, punch cards 23:377
Bates, Robert 6:277
Bateson, Gregory 1:743, 13:71, 14:383, 19:505
Bateson, P.P.G., works, *Growing Points in Ethology* 22:900
Bateson, William 8:386, 17:661
Bath School, scientific controversies 21:283
Batres, C.M., works, *La Arqueología de la América Andina* 23:36–37
Battered babies 3:473
Battered wives 6:633–634
Battered women's syndrome 6:625–627, 6:630, 6:636
Batterer intervention programs (BIPs) 20:254
  for intimate partner abuse 12:678
Batterie Imagination et Perception (BIP) 25:166–167
Batteries
  for aphasia 3:793–794
  for language disorders 3:795
Battersea and Wandsworth Trade Union Council's Organising Centre 13:163–164
Battery (assault) 15:592
Battery electric vehicle (BEV) 2:307
Battle fatigue 15:497
Battle of Marne (1914) 9:213
Battle of Seattle (1999) 19:747
Battle of Somme (1916) 9:214
Battle of Tannenberg (1914) 9:213
Battle of the Nile 1:301–302
Battle of the Sexes, example of Nash equilibria 9:577
Battle of Valmy 25:363
Battle of Verdun (1916) 9:214
Battle of Ypres (1917) 9:214–215
Battlefield 15:502, 15:504
Battleship curves 3:325
Battuta, Ibn 12:737
BAU *see* Business as usual (BAU)
Baudelaire, Charles 1:540, 15:684
  works, *Ein Lyriker im Zeitalter des Hochkapitalismus* 15:684
Baudrillard, Jean 5:245, 15:682
Bauer 17:669
Baum, D., works, *Smoke and Mirrors: The War on Drugs and the Politics of Failure* 15:27
Bauman, Karl E., on cigarette smoking 16:558
Bauman, Zygmunt 9:462, 22:994
Baumann, Shyon, on popular music 3:800
Bauman's continuum 19:821
Baumeister, Reinhard 24:863
Baumgarten, Alexander, on aesthetic education 1:229
Baumgartner, Frank R., model of agenda-setting 1:359
Baumol, W., on artists 2:58
Baxi, Upendra 13:515
Bay Area School Reform Collaborative (BASRC) 16:552
Bayat's analysis 24:819
Bayes, Reverend Thomas 8:17, 14:809, 20:928
Bayes decision rule 23:261–262, 23:412–413
Bayes factor 4:476–477
Bayes intervals 8:19
Bayes' Law 17:104
Bayes risk function for learning 23:414
Bayes theorem 2:375, 2:378–382, 12:799–800, 14:147, 16:117–118, 16:826, 19:19, 19:779
  applications prior to 1990 2:380–381
  brief history 2:379–380
  conditional probability 2:378–379
  Gibbs sampling 2:381
  inverse probability 2:379–381
  likelihood 2:379
  Markov chain Monte Carlo 2:381
  mathematical psychology 14:809
  Metropolis–Hastings algorithm 2:381
  normalizing constant 2:379
  posterior distribution 2:379–381
  prior distribution 2:379–380
  proportionality 2:379
  shrinkage 2:380
  since 1990 2:381
  subjective opinion elicitation 7:382
  subjectivity 2:380
  updating 2:378–379, 2:381
  versions of 2:378
Bayesian computation 15:758–762
  Markov chain Monte Carlo methods in 15:763–767
Bayesian equilibrium 16:432–433
Bayesian estimation 10:445, 22:469
  linear hypothesis 14:147
Bayesian geostatistics 23:195
Bayesian graphical models and networks 2:363–367
  DAG models for expert systems and decision analysis 2:365
  definition and basic properties 2:363–365
  learning DAG models from data 2:366
  pointers to literature 2:366
Bayesian inference 2:368–369
  likelihood in statistics 14:117–118
Bayesian information criterion (BIC) 15:627, 15:653, 15:656–657, 23:401–402, 23:618–619
Bayesian method of moments (BMOM) method 21:982
Bayesian methods
  for multiple comparisons 16:85–86
  phylogenetic methods 8:372–373
    data 8:372
    finding the most probable trees 8:372–373
    model 8:372
  spatial statistics 23:195
    geostatistics 23:195
    hierarchical models 23:195

Bayesian models in neuroscience 2:368–372
Bayesian procedures
  sensitivity of 8:11
Bayesian simulation methods 15:606
Bayesian spatial statistics 23:195
  geostatistics 23:195
  hierarchical models 23:195
Bayesian statistics 2:373–377, 2:380, 8:695, 14:805, 16:85–88, 16:878–880, 17:259–260, 20:928, 21:271, 23:379–380
  application in neuroscience
    Bayesian inference 2:368–369
    behavioral models 2:369–372
    Gaussians 2:369
  Bayesian updating model 6:253
  calculation of posterior distribution 2:375–376
  capture–recapture studies 2:380, 18:605–606
  cognitive development 4:38
  discrete-state models 6:528
  experimental designs 8:521–526
    applications 8:524
    Bernoulli distributions 8:521
    breast cancer 8:524
    calculations 8:523–524
    choice of utility functions 8:522–523
    example 8:524–525
    Monte Carlo experiments 8:523–525, 8:525f
    optimal designs 8:521–522
    prior distributions 8:523
    software 8:524
  fiducial/structural inference 9:124–127
  Fisher's studies 9:224
  generalized linear mixed models 9:849
  hierarchical models 1:369, 10:836–839, 23:195
  identification 23:408
  inference 2:373–376
  item response theory 8:699–701
  Markov chain Monte Carlo methods 2:375–376, 14:569
  mathematical psychology 14:805
  meta-analysis 15:274–275, 23:912–913
    confidence intervals 15:275
    robust standard errors 15:275
    statistical power of tests 15:275
  model assessment 2:376
    model checking 2:376
    model comparison 2:376
    sensitivity analysis 2:376
  model checks and criticisms 15:654–655
  model of learning 5:880–881, 5:883, 6:253, 8:514
    principles of 8:359–360
  model selection 15:656–657
  model specification 2:373–374
  objective inference 2:374
  phylogenetic methods 8:372–373
    data 8:372
    finding the most probable trees 8:372–373
    model 8:372
  prior distribution 2:373–376
  probability 2:373–376
  queues 19:786–788
  selection of models 4:475–477
  sequential experiments 21:636
  simulation 2:375–376
  subjective prior distribution 2:374
  test, for hypothesis 11:492
Bayesian techniques, Pearson's use of 17:662
Bayesian treatment of identification 23:405–406
Bayes–Nash equilibrium, auctions 2:206, 2:208–209
Bayh-Dole Act 1985 2:667, 24:765
Bayley Scale of Infant Development (BSID) scores 2:849–850, 7:929–930
Bayley Scales of Infant and Toddler Development 3:795
Bayley Short Form-Research Edition (BSF-R) mental scale 2:850
Baylon, Pasqual 15:579
Baym, Nancy, works, *Personal Connections in the Digital Age* 2:218
Bazerman, M.H., works, *Cognition and Rationality in Negotiation* 16:431
Bazin, André 15:15
B2B (business to business) commerce 8:487
BBC *see* British Broadcasting Corporation (BBC)
BBSC *see* Benefit/Burden Scale for Children (BBSC)
BC *see* Bounded confidence (BC) model of social influence
BC Programming Language *see* BCPL (BC Programming Language)
BCA *see* Benefit–cost analysis (BCA)
BCAFN *see* British Columbia Assembly of First Nations (BCAFN)
BCI *see* Brain–computer interaction (BCI); Brain–computer interfaces (BCI)
BCPL (BC Programming Language) 7:2
BCS *see* Benefit–cost analysis (BCS)
BCT *see* Behavioral couples therapy (BCT)
BDAE *see* Boston Diagnostic Aphasia Examination (BDAE)
BDDs *see* Binary decision diagrams (BDDs)
BDH *see* Bono de Desarrollo Humano (BDH) cash transfer
BDI-II *see* Beck Depression Inventory-II (BDI-II)
BDNF *see* Brain-derived neurotropic factor (BDNF)
Bdnf gene 7:842–844
BDP *see* Kurdish Peace and Democracy Party (BDP) (Turkey)
BDR *see* Behavioral decision research (BDR)
BDS *see* Boycott, divestment, and sanctions (BDS) movement
BDSM *see* Bondage, domination/discipline, submission/sadism, masochism (BDSM) communities
BDT framework *see* Business Data toolset (BDT) framework
BDTC *see* British Dependent Territory Citizens (BDTC)
BE *see* Biological Economies (BE) research program
BEA *see* Bureau of Economic Analysis (BEA)
Beachy, Robert 9:644–645
Beacon model 20:793
Beagle, H.M.S. 20:287
BEAGLE (Bound Encoding of the Aggregate Language Environment) 21:581–582
Beal Report (American Anthropological Association) 8:56
Beam search techniques 23:262
Bean v. Southwestern Waste Management Corp. 7:756–757
Bear hugs 13:890
BEAR Magazine 13:890
Bear Nation (documentary) 13:891
Bear Paws of Iowa 13:890
Bear Run (documentary) 13:891
Bear runs 13:891
Bear subculture 13:887, 13:889–891
  critiques of 13:891
  current state of 13:891
  gender and sexuality 13:890–891
  history of 13:889–890
Bear weekends 13:891
Bearman, P.S. 16:211
  on narrative networks 16:212–216
Beat-hip bohemianism 5:81
Beats (youth culture) 5:80
Beatson, Alexander, works, *Tracts relative to the Island of St Helena* 7:748
Beattie, Geoffrey 10:110
Beauchamp, T.L., works, *Principles of Biomedical Ethics* 2:632–633
Beaufort, Frances Anne 7:113
Beaumont, Harry, works, *The Broadway Melody* 9:168
Beauty
  definition 2:383–384
  as erotic capital 8:1
  and labor market 2:383–391
    cognitive/noncognitive skills 2:385–386
    country findings 2:388–389
    customer discrimination/productivity 2:385
    effect of beauty on wages 2:389t
    empirical estimation methods 2:386–387
    employer discrimination 2:384–385
    explanations of beauty premium and painless penalty 2:384–386
    gender differences 2:387–388
    human capital model 2:386
    occupational sorting 2:385
    stereotypes 2:384
  measure and alternative measures 2:383–384
Beauty premium 2:384–386
Beauvoir, Simone de 2:392–395
  biography 2:392
  on fixed destiny for women 2:393
  against French atrocities in Algeria 2:392
  institution of the family 2:394
  on mother–child relationship 2:393

Beauvoir, Simone de (*continued*)
on women's blocked situation 2:393–394
work on feminist theory 2:392
works
Being and Nothingness 2:392
Bodies That Matter 2:394
Gender Trouble 2:394
The Mandarins 2:392
The Second Sex 2:392–394
Beaux arts 5:451
Beccaria, Cesare 5:232, 5:875, 6:255, 13:802
on division of labor 6:602–603
works
Dei delitti e delle pene 7:657, 7:659, 13:802
On Crimes and Punishments 5:875, 13:570, 24:443
Beck, Ulrich 2:396–400, 14:114, 22:994
biography 2:396–397
on cosmopolitanism 2:396–399
on environment 2:397–399
in Europe 2:396–399
in Germany 2:396, 2:398
on globalization 2:396–399
on individualization 2:397–398, 14:114
on methodological nationalism 2:396–399
on political economy 2:397–398
on reflexive modernization 2:397
on risk 2:396–399
on risk society 2:397
on second modernity 2:396–399
impact of 2:398–399
works
Brave New World of Work 2:397
on zombie categories 2:399
Beck Depression Inventory 18:89
Beck Depression Inventory-II (BDI-II) 6:343–344
Becker, Ernest, works, *The Denial of Death* 7:595
Becker, Gary 1:77, 7:12, 12:389, 17:781–782
on division of labor 6:605
on economics of childbearing 8:821, 8:823
on economics of gender 9:760
criticism of 9:761
on household economics 11:224–225
on human capital 22:331
treatment of risk preference 13:571–572
works
Crime and Punishment: An Economic Approach 13:570–571
The Economics of Discrimination 6:536
Becker, Howard S. 2:401–406, 3:71–72, 13:150, 21:265, 25:193
on art worlds 2:404–405, 5:466
on careers and social networks 5:468
cases 2:405
on consumption 16:555
and Goffman 2:401, 2:405
and Hughes 2:401, 2:404
intellectual influences of 2:401
on labeling theory, drugs, and addiction 2:403–404
on student's perspectives 2:401–402
on symbolic interactionism 12:352–354
on teaching and social status 2:402–403
and Warner 2:401–402
and Whyte 2:401–403
works
Art Worlds 5:502, 5:605
The Other Side 13:151
Beckert, Jens, on value 5:538–539
Beckham, David 23:315–316
Becoming
organizational learning 13:697
in psychotherapy 20:740
BED *see* Binge eating disorder (BED)
Bed nucleus of the stria terminalis (BNST) 8:862
Bede, Venerable 20:285, 24:413–414
works, *Historia ecclesiastica gentis Anglorum* 24:411
BEE *see* Best Evidence Encyclopedia (BEE)
Beer–Lambert law 16:388
'Before–after' comparison 12:450
Before–after studies, evidence-based medicine 8:328–329
Beginning of life 2:615
controversies over 5:872
Beginning writers, teaching foundational writing skills to 25:769
Begriffsgeschichte 2:784–785, 4:548–554, 12:257–258, 18:486, 18:489–490
Behaghel's first law 11:512
Behaghel's second law 11:513
Behar, Ruth, works, *Translated Woman* 8:901
Behavior analysis 2:413
applied 2:407–411
Behavior Assessment System for Children 16:726
Behavior change interventions
and motivation/actions 15:892
personalized medicine 17:929
Behavior change techniques
for self-efficacy 21:511
Behavior changes 10:622–624, 15:909–912, 21:650–651
for accident prevention 12:133
and alcohol 1:508
and alcohol misuse 1:520–521
and infectious diseases 12:32–33
and innovation 12:171
and sexual development 14:328–329
stages of 15:909
supporting for weight control 25:472–473
and sustainability lifestyles 23:828–829
and sustainable practices 23:834
Behavior genetics
of sexual orientation 21:779
of socialization and education 22:862
and temperament 24:191–193, 24:195
Behavior modification approach
personalized medicine 17:926, 17:928
for workplace safety and health 25:729–731
analysis of intervention data 25:731
control group design 25:731
experimental designs 25:730–731
interventions 25:729
multiple-baseline design 25:731
positive reinforcement 25:729
reversal or ABAB design 25:731
safety behavior, measurement of 25:730
Behavior observation techniques 10:591
Behavior prediction, and implicit social cognition 11:717
Behavior problems
childhood 2:477
effects of socioeconomic status on 2:477–480
importance 2:478
development of 2:470
and fatherhood 8:842
role of genetic factors on 2:473–474
Behavior Rating Inventory of Executive Function (BRIEF) 16:726
Behavior Rating Inventory of Executive Function for age from 18 to 90 (BRIEF-A) 16:726
Behavior Rating Inventory of Executive Function for ages of 2 and 5 (BRIEF-P) 16:726
Behavior Rating Inventory of Executive Function self-report forms (BRIEF-SR) 16:726
Behavior therapy 2:407–408, 2:410, 2:412–417, 2:431–436, 8:518, 11:612–614, 18:41–42, 24:288
acceptance and commitment therapy 2:420
anxiety disorders 2:426–429
biopsychosocial model 2:419–420
classical conditioning 2:413–414
for clinical disorders 2:419–421
clinical practice guidelines 2:418
cognitive behavioral therapy 2:419–423, 2:431–434
cognitive therapy 2:426, 2:429, 2:431–433
conditioning 2:426
couple therapy 2:428
depression 2:426
desensitization 2:426–427
development 3:874
evidence-based practice in psychology 2:418–419
exposure treatments 2:428
future of 2:429
group therapy 2:426–427
historical background 3:880–881
history 2:413–416, 2:426–427
medical model 2:422
mindfulness 2:431, 2:433–434
operant conditioning 2:415
operant techniques for chronic problems 2:428–429
principles 2:413–416
psychiatric aspects 2:426–430
for psychosomatic disorders 19:489
reinforcers which increase desirable behaviors 2:429
negative reinforcers 2:429
positive reinforcers 2:429
reinforcers which reduce undesirable behaviors 2:429

punishment 2:429
response cost 2:429
relaxation 2:426–427
rule-governed behavior 2:432
second-wave 2:431–432
standardization 2:418, 2:421
teaching new skills 2:428
third-wave 2:432–434
treatments 2:419t
unresolved issues 2:421–423
Behavioral activation 2:431, 11:615
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
for depression 19:375–376
Behavioral activation system (BAS) 8:624, 17:828, 17:922, 21:850–851, 25:685
Behavioral activation therapy 4:544–545
Behavioral addiction 1:99
Behavioral analog
in EITMs
relating to empirical analog for measurement error 18:458
tools and components for 18:457–458
Behavioral analytics, technological change 24:107
Behavioral approach system (BAS) 17:794, 17:860
Behavioral approaches
for physical activity 18:96–97
to politics study 5:656–659
Behavioral archaeology (BA) 1:892–893
Behavioral aspects, of Alzheimer's disease 1:584
Behavioral assessment 16:726
Behavioral changes
of acculturation 1:70
Behavioral community psychology 2:410
Behavioral components
future orientation 9:562
test anxiety 24:244
Behavioral concepts, in EITMs 18:456
Behavioral contract 10:429
Behavioral control 4:820
modes of 20:211–212
Behavioral costs 7:700
Behavioral couples therapy (BCT) 8:748, 8:847
for depression 19:376
Behavioral cultural intelligence 5:433
Behavioral decision making *see* Judgment and decision making
Behavioral decision research (BDR) 2:445–446
framing effects and prospect theory 2:445–446
preference reversals and construction of preferences 2:445
Behavioral decision theory, consumer psychology 4:739
Behavioral dependency, in adulthood 1:182
Behavioral disengagement, cardiac disease, coping with 3:112
Behavioral disorders 19:457–458
in USA 19:457t
Behavioral ecology 14:370, 18:892–893, 18:895, 22:903, 22:909
and hunter-gatherers 11:425
Behavioral economics 2:437–449, 2:626, 5:879, 5:888–889, 8:480, 8:548, 9:582, 9:588, 11:259, 13:458–459, 15:334, 17:780
behavioral decision research 2:445–446
framing effects and prospect theory 2:445–446
preference reversals and construction of preferences 2:445
bounds of human nature 2:438–439
consumer policy 4:734
directions for 2:441
Elster on 7:402
expected utility theory 2:443–444
early criticisms of 2:444–445
finance 2:439–440
and genetics 9:951
Homo economicus 2:437–438
market forces 2:437
new 2:446
old 2:444–445
and procrastination 19:76
psychology 2:443–448
savings 2:440–441
second-wave 2:446–448
modeling dynamically inconsistent preferences 2:447
modeling prosocial preferences 2:447
recent criticisms of 2:447–448
as separate science 2:443
and sustainability lifestyles 23:831
theoretical innovations 2:438
time discounting 2:447
unbounded rationality 2:437–438
unbounded selfishness 2:438
unbounded willpower 2:438
Behavioral engagement 1:31
Behavioral exercises 4:545
Behavioral expressions 21:608–610
drinking 21:609–610
driving 21:609
drugs 21:609–610
entertainment preferences 21:610
sex 21:609
smoking 21:609–610
sports 21:609
vocations 21:610
volunteering 21:608
Behavioral factors, of gender gap in life expectancy 9:704–705
Behavioral game theory 9:587–590, 13:750, 16:430–431
bounds on human rationality 9:587–588
empirical regularities 9:588
learning 9:589–590
modeling player's thinking processes 9:589
players' motivations 9:588–589
Behavioral genetics 2:267, 2:450–455, 9:826, 12:297, 22:282–283
adoption studies 2:451
and crime 2:672–673
DNA sequence 2:450
and economics 9:952–953
environmental influences 2:450–453
and fear-potentiated startle 8:863–864
genetic influences on behavior 2:452
genome-wide association 2:452–453
heritability 2:452
identification of specific genes and pathways 2:452–453
Mendel's studies 2:450
research 2:453–454
twin studies 2:451–452
Behavioral genomics 2:623
Behavioral geography 16:337, 22:302, 22:303
Behavioral habituation 10:481
Behavioral health, and religion 20:265–266
Behavioral implications, personality-religion links 17:803–804
Behavioral Inattention Test, for hemispatial neglect 10:767
Behavioral Indicators of Conscientiousness scale (BIC) 4:660
Behavioral inhibition (BI) 2:456–462, 22:495–497, 24:191, 24:193–196
biological correlates of 2:458
development of 2:457–458
emotional and cognitive outcomes of
internalizing problems 2:460
self-concept and coping 2:460
social cognition 2:460–461
measures of 2:457
in social contexts
cultural influence of 2:458
in parent–child relationships 2:458–459
in peer relationships 2:459–460
Behavioral inhibition system (BIS) 8:624, 17:794, 17:828, 17:860, 17:922, 21:850–851, 25:685
Behavioral interpretation of local log odds in marriage matching 2:104–106
Behavioral interventions, medical model for 18:162
Behavioral lateralization, in animals 2:799–805
Behavioral learning 16:639–642, 16:644
competition and cooperation 4:856–857
Behavioral life cycling hypothesis 15:737
Behavioral marker 3:833
Behavioral mechanisms of anxiety 1:814–815
Behavioral medicine 2:463–469, 17:793
assessment 2:466–467
avoidance 2:467
cognitive styles 2:467
coping and adherence to treatment 2:467
lay representations of illness and treatment 2:466–467
resources 2:467
biopsychosocial model 2:463–464
chronic physical conditions 2:465–468
definition 2:463
evolution 2:464
current state of art 2:464
historical development 2:464

Behavioral medicine (*continued*)
interventions 2:467–468
health promotion and support in disease management 2:468
treatment of comorbid mental disorders 2:467–468
morbidity and mortality in the twenty-first century 2:464–465
behavioral risk factors 2:465
consequences for health care and social systems 2:465
epidemiology 2:464–465
socioeconomic status and health 2:465
psychological factors and physical disorders 2:465
concepts and disciplines 2:465–466
interactions 2:465
psychoneuroimmunology 2:464
psychophysiology 2:464
for respiratory disorders 20:578
stress management programs 2:468
Behavioral modeling applications 23:416–417
Behavioral models 2:369–372
collective decision making 2:371–372
of panic disorder/agoraphobia 17:476–477
sensorimotor integration 2:371
sensory integration 2:370
visual perception 2:370–371
Behavioral modules 14:367
Behavioral neuroscience 2:369–372, 4:98
Behavioral outcome, emotional intelligence as 7:446
Behavioral pharmacology, Watson contributions to 25:449
Behavioral physiology 14:366–367
Behavioral plasticity, and successful aging 23:669
Behavioral pride *see* Authentic pride
Behavioral regularities, norms as 17:6–7
Behavioral regulation, reinforcement as 20:209–210
Behavioral rehabilitation 10:716t
Behavioral reinforcement 20:215
Behavioral reports, questionnaires 19:775
Behavioral research, and environmental cognition 7:710–711
Behavioral revolution
and Cold War 4:122
United States presidency 18:862
Behavioral risk factors
reduction of 20:270
smoking 20:270
Behavioral risks, cardiac rehabilitation 20:188–189
Behavioral sciences
Baltes' contributions to 2:351
human resource management 11:351
on water and health 25:433–434
Western views of 25:541–550
Behavioral scientists
roles in personalized medicine 17:928–930
change management 17:929–930
creating teams 17:930
interventions requiring behavior change 17:929
risk perception 17:930
Behavioral self-control 2:410
Behavioral specificity, brain stimulation reward 2:845
Behavioral strategy 23:511
Behavioral studies
of children with focal injuries 16:587–588
of eyeblink conditioning 8:635–636
Behavioral theories
modern trends in 13:706
of organizations 2:481–486
coordination 2:482
discussion 2:485
evolutionary theory 2:482–485
institutional theory 2:484–485
learning 2:482–483
learning theory 2:482–483
management 2:481
motivation 2:482
organization theory 2:481
organizational goals 2:481–482
population ecology 2:481–482, 2:484
theoretical review 2:481–482, 2:482t
transaction cost theory 2:485
and public health 19:545–546, 19:550
role of self-efficacy in 21:509
Behavioral theory of the firm (BTOF) 2:481
Behavioral theory of timing (BeT) 24:308–309
Behavioral therapies 20:59
cognitive behavioral therapy 4:30–32
for depression 19:375–376
Skinner's 22:28–29
Behavioral traits prediction, and genetics 10:37
Behavioral trajectories, heterogeneity in 1:313–314
Behavioral viewpoint, market geographies 14:547
Behavioral weight loss (BWL) treatment 19:377
Behavioral-based conflict, and work–family conflict 8:759
Behavioral–cognitive approach 25:539
Behavioralism, American 18:469
Behavioralists, view on political institutions 12:182
Behaviorally anchored ratings scales (BARS) 24:100
Behaviorally spontaneous confabulation 4:560–561
Behavior-centered approach, Sapir–Whorf hypothesis 20:905
Behavior-dependent honesty 12:637
Behavior-dependent trust 12:637
Behavior-enacting trust 12:637
Behaviorism 2:410, 7:238, 8:483, 12:813, 13:694, 13:702–703, 14:367, 16:340–341, 17:609, 18:125, 21:847–848, 21:953, 22:986
basic assumptions of 13:704
Broadbent on 2:856–857
Chomsky on 3:527
and cognitive psychology 4:103–104
Computer Assisted Instruction 4:495
and concept of introspection 12:703–704
in education 12:67–68
habit 10:477–478
and instinct 12:179
and learning 13:704
and life management 14:85
and psychology 19:413–414
Skinner on 22:25–30
Watson on 25:448–449
Behaviorist learning theory 23:214
Behaviorist orientation of linguistics 2:706–707
Behaviorist psychology 4:292, 21:231
Behavioristic movement in psychology 17:609
Behaviorization of power 15:800
Behaviors 15:228
addiction 1:98
adolescent health 1:121–127
asking questions about 19:772–774
facilitating recall 19:773–774
and attitude 2:183–189
avoidance and approach motivation 2:326–331
classrooms, and gender 9:683
climate change-related 3:858
cognition, evolution of 4:1–2
and collective efficacy 4:169–172
and commitment 17:334–339
consistency across demographically heterogeneous workgroups 17:699–700
consistency across organizational cultures 17:698–699
consumer
complexity of 4:738
evolution of scholarly research on 4:739–740
introduction of psychology to 4:740
reasons for studying 4:738–739
coping, and chronic illness 3:549
distinctions 3:550–551
customer *see* Customer behavior
delinquent 6:68
determinants of 10:622–623, 10:626
determined by self-efficacy 21:511–512
of disgust 6:546, 6:548
drug-using 6:685
ecology *see* Human behavioral ecology
effect of parenting and family structure on *see also* Family structure; Parenting
empathy-related, neuroevolution of 7:542–543
epigenetics in 7:842–844
evidence for lexical consolidation 13:919–921
evolution and heritability of 12:175–176
gender-specific 9:695
genetic factors 9:921–922
genetics 6:320–325
habit 10:475–478
health 10:582–587
and HPA axis 11:478–479

human, perceptual learning and 10:132–133
agency 10:132
flexibility 10:132–133
prospectivity 10:132
hysteria 11:509
individual, patterning of 5:578
infant gaze behaviors 12:27–31
interictal behavioral dysfunction 13:35
of journal authors 12:865
judicial 12:885–890
modern, emergence of 17:458–459
moral 7:319, 15:788–790
neonatal crying *see* Neonatal crying behaviors
and neuroethics 16:664, 16:666–668
nonheterosexual British Asians 16:863
observation, and attitude measurement 2:181
optogenetic studies 17:268–269
outcomes, cardiac disease coping 3:114
of parents 5:377
patterns, autism spectrum disorders 2:276
and personality traits 17:814–818
and persuasion 17:943
political 18:459–464
during posttraumatic amnesia 18:638
and procrastination 19:77
pro-environmental behaviors 19:82–88
program evaluation 19:118
regulation, and self 21:441–443
regulatory agency 20:179
repertoire, neonatal crying as part of 16:476
retirement, modeling 20:603
role of context in 4:786–787
and self-regulation 21:553–555
sexual 21:786–787
sexual risk 21:797–801
of siblings during radical revolutions 21:925–926
suicidal, origins of 23:676–679
temporal organization of 18:813–814
and values 25:33–39, 25:45
voting, in referendums 25:296–298
work-related 17:824–825
youth physical activity motivation 15:916–918
Behaviors spreads
diffusion 6:412, 6:414
Behaviour 14:369
Behvarzes 6:43
Beijing 24:870–871
Beijing Consensus 3:505
Beijing Genomics Institute (BGI) 9:984, 11:318–320
Beijing Platform for Action 6:630, 10:193
Beijing Rules 12:956, 12:960–961
Beijing Seven Stars Group 5:598
Being, notions of 14:136
Beitz, C., works, *Political Theory and International Relations* 19:917
Bekhterev, Vladimir 19:413
Bekkering, E., on consumer–citizens 10:152
Bektashiya 24:603
Belgian Congo *see* Congo
Belgium 16:261
income inequality in 22:327
laicization in 13:218
tax system 23:341
and World War I 9:212
Belgrade Charter 7:733–734
Belief merge, phenomenon of 16:927
Belief-desire explanation 9:404–406
Belief–desire psychology *see* Theory of Mind (ToM)
Beliefs 4:163
about drugs 6:687–688
in ancestors 1:688–689
anthropology of 2:493–497
collective 4:163–168
culture as 9:51–52
definition 2:493, 25:28
in diverse anthropological traditions 2:494–496
anthropology of religion 2:496
Boasian tradition 2:495
cognitive anthropology 2:495–496
health psychology 2:496
intellectualist tradition 2:494–495
medical anthropology 2:496
public health 2:496
rationality debates 2:494–495
structural functionalism 2:495
and emotion regulation 7:454
environmental cognition 7:709–711
in Euro-American language use 2:493–494
eyewitness testimony 8:644, 8:647–648
formation of 17:780–781
freedom of 13:219
Glaube 18:436–437
identities, and 20:378–379
institutions 12:190–195
meritocracy
causes of 15:240–241
consequences 15:241
normativity of content (NCB) 14:844, 14:847–848
and opinion formation 17:243–244
parenting, across cultures 17:493
political 18:429
positive 4:2–3
versus practice
religious symbols 23:863–864
rationality of 12:335–336
and religion 20:721, 20:725
religious *see* Religious belief
representation 20:68–70
ritual 20:717, 20:719
and rituals 20:722–723, 20:725
rumors and 20:788
secularization 21:383–385
and self-concepts 21:469, 21:471
self-efficacy
development of 21:506–507
microgenesis of 21:507–508
social *see* Social beliefs
social, ideologies as 11:574–576
symbolic 20:724–725
and symbolism 23:861–864
tacit knowledge 13:77–78
and tradition 24:509
Belize 1:638
sociocultural aspects 1:621, 1:623–625
Bell, Alexander Graham 24:154
Bell, Catherine, works, *Ritual Theory, Ritual Practice* 20:309
Bell, Colin 4:362
Bell, Daniel A. 7:904, 12:85, 21:169
anomie 2:503
anticommunism 2:498–499, 2:502
capitalism 2:498–504
conception of knowledge society 13:106–108
early life 2:498–499
end-of-ideology debate 2:499–502
liberalism 2:499–500
mass culture 2:503
modernism 2:503
morality 2:498
neoconservatism 2:504
postindustrial society 2:502–504
postmodernism 2:503
the sacred 2:503
socialism 2:498–504
status politics 2:502
totalitarianism 2:499–501
works
The Coming of Post-Industrial Society 18:664, 18:667–668
The Cultural Contradictions of Capitalism 2:502–504, 18:664–665
The End of Ideology 14:745, 18:664
Bell, Daniel A., works, *Communitarianism and Its Critics* 4:320
Bell, David 21:808–809
on sexual citizenship 21:730–731
Bell, Derrick 5:266–267
works, *Civil Rights Chronicles* 5:267
Bell curve 23:418
intelligence 12:311
Bella Coola Indians 2:710
Bellagio-Harvard Guidelines, slavery 22:43
Bellah, Robert Neelly 5:499, 17:522
on civil religion 2:508–509, 3:680–684
contribution to social science 2:509
at Harvard 2:506–508
and public sociology 2:509–510
on religion and evolution 2:510–511
as social scientist 2:506
works
Apache Kinship Systems 2:506
Beyond Belief 2:509
The Broken Covenant: American Civil Religion in Time of Trial 2:508–509
Civil Religion in America 2:508
Confessions of a Former Establishment Fundamentalist 2:509
Durkheim and History 2:507
The Good Society 2:510
Habits of the Heart 2:509–510, 4:320
Heritage and Choice in American Religion 2:508
Law as an Intellectual Stepchild 2:508–509
The Modern Condition in the Light of Human Evolution 2:511
The New Religious Consciousness 2:509

Bellah, Robert Neelly (*continued*)
Religion in Human Evolution: From the Paleolithic to the Axial Age 2:510, 20:387
Religious Aspects of Modernization in Turkey and Japan 2:507
Religious Evolution 2:507
The Robert Bellah Reader 2:510
The Systematic Study of Religion 2:507
Tokugawa Religion: The Values of Pre-Industrial Japan 2:507
Transcendence in Contemporary Piety 2:509
Bellamy, Edward 4:524, 14:251, 24:862–863
works
Looking Backward: 2000–1887 2:150, 14:251
Uncle Tom's Cabin 14:251
Belle Epoque 17:517–518
Belle époque time period 16:462
Bellicose tradition 15:493
Bellman equation 21:333, 21:335–336
Bell-shaped curve, for per capita income and environmental degradation 13:145–146, 13:145f
Bellum omnium contra omnes 3:708
Belmont principles 20:514
Belmont Report 4:252, 8:38–40, 8:48, 13:559–560, 20:514
The Belmont Report 8:48
Belonging
and adoption 1:146
consciousness of 8:175–177
identity, nonheterosexual British Asians 16:864
politics, in migration 15:423–424
and territory 24:224–225
Below-replacement fertility 21:346–347
BEMI *see* Barts Explanatory Model Inventory (BEMI)
Benabib, Seyla 5:299
Benda-Beckmann, F. von 13:757
Bendix, Reinhard 2:512–514, 19:677
on authority 2:512–513
on bureaucracy 2:513
on intellectuals 2:512–514
on managerial ideology 2:512–513
on nation building 2:513
on power 2:513
on scholarly community 2:512
Benedict, Ruth Fulton 1:743, 1:747, 2:712, 7:145, 9:459, 9:461, 14:837–838, 16:237, 20:669–670
accomplishments of 2:517–518
comparative analyses by 2:518
contributions to anthropology 2:515–517
critique of Freudian theory 2:517
on humanism 2:518
inquiry into patterns of American culture 2:516
Japanese culture 2:517
lectures of 2:516–517
life of 2:515
on social science 2:517
works
The Chrysanthemum and the Sword 2:72, 2:515, 2:517, 12:794
Patterns of Culture 2:515–517, 6:590
Benedict, Stanley Rossiter 2:515
Benedict's Monte Cassino monastery 15:725
Benediktsson, Mike O., on cultural consumption 16:557
Beneficence 8:79, 12:607
biomedical research ethics 8:79–81
ethical aspects in surveys and polling 23:754
ethical principle of 20:518
Beneficia 9:111
Beneficiary right 19:173
Benefit adjustment model 12:517–518
Benefit finding 23:554
Benefit/Burden Scale for Children (BBSC), Revised 3:462
Benefit–cost analysis (BCA) 5:56, 7:728–729, 19:702, 25:19
Benefits 5:56
of environmental changes 7:708–709
research, assessment of 20:512
research funding 20:525–526
subjective well-being 23:649
physical health 23:649
social relationships 23:649
work 23:649
of voting 25:309–310
Benevolence 15:781
and intergenerational justice 12:396
in sexism 21:695–698
and values 25:28–30
Benevolent sexism 4:154
Benevolent societies 13:105
Benfey, Theodor, works, *Geschichte der Sprachwissenschaft* 14:215–216
Bengtson's solidarity paradigm 12:403
Benhabib, Seyla 2:519–524, 5:307
and Arendt 2:519, 2:522–523
childhood and youth 2:520
on citizenship 2:519, 2:523–524
on cosmopolitanism 2:519–520, 2:522–524
on critical theory 2:519, 2:522–524
on culture 2:519–523
on deliberative democracy 2:519, 2:522–523
on ethnicity 2:519, 2:522–523
on feminism 2:519, 2:522–523
on globalization 2:519, 2:522–524
and Habermas 2:519, 2:522–523
higher education abroad and academic positions of 2:520–521
on law 2:519, 2:522
multicultural family background of 2:520
political commitment of 2:521–522
on postmodernism 2:523
research interest and intellectual production of 2:522–524
on state 2:520–521, 2:523
on universalism 2:519, 2:521–524
Benign disinhibition (online dating) 17:213
Benign gerrymander, apportionment 7:343
Benign humor 11:417
Benjamin, Harry 24:550
Benjamin, Walter 1:747, 5:499, 14:200–201, 15:684
works
Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit 15:684
Das Passagen-Werk 15:684
Benjamini–Hochberg FDR-controlling method 16:84
Benji 11:49
Bennet, Craig 9:487–488
Bennett, Alice 25:625
Bennett, T., on Australian middle classes 2:242
Benney, M. 12:670
Benson, Paul 2:318
Benson, Peter 3:445–447
Bentham, Jeremy 2:525–530, 3:710, 6:255, 8:213–214, 10:537, 11:368, 12:508, 13:437–438, 13:570, 13:734–735, 20:681, 23:199, 24:980
contribution to social science 2:528–529
on individuals and collectives 11:818
influence of 2:529
as jurist 2:526–528
codification 2:527–528
fictions 2:527
pannomion 2:527–528
theory of law 2:527
utilitarian 2:526–527
legal positivism of 10:537
life of 2:525–526
and panapticon 18:922
science of morals and legislation 24:981–982
studies of 2:529
works
Collected Works (*CW*) 2:529
Constitutional Code 2:528
A Fragment on Government 2:526
An Introduction to the Principles of Morals and Legislation 2:526, 8:213–214, 12:508
Of Laws in General 2:527
Traité de legislation civile et penale 2:529
Utilitas 2:529
writings 2:525–526
Benthamic ambiguity 2:525–526
Benthamism 10:490–491
Bentley, A.F., works, *The Process of Government* 18:233
Benton Facial Recognition Test 8:678, 19:251
Benton Visual Form Discrimination 16:723
Benveniste, Émile 18:796
Ben-Yehuda, Nachman 15:798
Benzécri, Jean-Paul 5:1, 5:9
Benzodiazepines 6:669, 6:671, 6:685, 20:212–213
for somatoform disorders 23:13
Bercovitch, S., works, *Ideology and Classic American Literature* 1:620
Bereavement 2:531–536, 5:475, 5:862
across life span 2:537
and childhood cancer 3:461
complicated grief 2:533–534

diagnosis 2:534
intervention efficacy 2:533–534
coping 2:533–535
death/dying 5:868
depression 2:532, 2:534
DSM-5 2:533–534
grief and mourning 2:531
impact of 2:531–533
mortality 2:532
physical and psychological health outcomes 2:532–533
risk and protective factors 2:533–534
models 5:862
reactions to 2:531t
resilience 2:532–535
theoretical understanding 2:534–535
widow 2:531–533
Berelson, B., works, *The People's Choice* 7:563
Bergen Communiqué 2:742
Bergen school of meteorology 7:793
Berger, Peter 17:393–394, 19:677, 20:346
on cognition 5:577
on humor 11:418
on institutions 17:394
on omnivorousness 5:450
works
The Desecularization of the World 21:389
The Sacred Canopy 21:389
The Social Construction of Reality: A Treatise in the Sociology of Knowledge 17:391, 19:491, 24:135
Social Construction of Technological Systems 24:136
Ways of Seeing 12:780
Bergoglio, Argentinean Jorge 3:538
Bergson, Henri 21:146–147
on humor 11:416
Bering Strait 1:921–923
archaeological mounds 1:923
archaeological research 1:921–922
complex societies 1:923
Berkeley, George 7:567, 15:957, 19:410
on empiricism 7:567–568
Berkeley free speech movement 5:73
Berkeley School, cultural geography 5:409
Berkeley Village Law Project 13:517
Berkhoff, Karel C., works, *Harvest of Despair* 6:814
Berlant, Lauren 19:748–749
on affect 1:239
Berlin, homosexuals in 9:638
Berlin, Isaiah 2:542–547, 6:591–592, 10:536
biography of Marx 2:543
on Enlightenment 2:542–543, 2:546
on freedom 9:390
history of ideas 2:545–546
on identity 11:524–525
as intellectual 2:542–543, 2:545
on liberalism 2:542–546
on liberty 2:544–545
life and career of 2:542–543
on romanticism 2:542, 2:546
on totalitarianism 2:542, 2:545–546
on value pluralism 2:542, 2:546
works
Two Concepts of Liberty 9:398–399
Berlin Aging Study (BASE) 2:350, 6:378
Berlin Communiqué 2:742–743
Berlin Conference (1885) 3:313
Berlin model of intelligence structure (BIS) 12:950
Berlin Society for Empirical Philosophy 14:315
Berlin Treaty 6:773
Berlin Wall 25:838–839
Berlin wisdom model 25:591–592
Berlinische Monatsschrift 19:610
Berlo, David 4:281
Berlyne, D.E., on cultural consumption 22:331
Bermuda 1:638
Bernadou, John B. 2:71
Bernal, John B. 21:215
Bernard, Claude, works, *Introduction to the Study of Experimental Medicine* 8:541–542
Bernard, Russ 16:562
Bernard, Viola 25:625
Bernardelli, Leslie, and Lewis population model 18:585
Bernardino of Santa Lucia 15:580
Bernays, Martha 9:414
Berne Convention for Protection of Literary and Artistic Works in 1886 3:31, 12:263–264
Bernheim, Hippolyte 11:465–466
hypnotism studies 19:502
Bernoulli, Daniel 13:383–384
works, *Emolumentum* 7:88
Bernoulli, Jacob 23:424
Bernoulli, Nicholas 14:809
Bernoulli distributions 6:581, 8:521, 22:468, 23:418
logistic regression 2:587, 2:590–591, 2:594–595
Bernoulli trials 5:858, 6:581, 17:662
Bernstein, Eduard 3:740–741
works, *Probleme des Socialismus* 3:95
Bernstein's comparator model 15:958
Berr, Henri
works
Encyclopedia of the Social Sciences 11:72
Revue de synthèse historique 11:40
Berry, John 6:839
Berry, Lloyd, on sexual minorities of Russia 21:740
Berry–Esseen theorem 14:631–632
Bertalanffy, Ludwig von 14:383
Berthelot, Jean-Michel 13:782
Bertillon, Jacques 17:28
Bertrand competition 14:551–552
Besson, Jean 1:641
Best Evidence Encyclopedia (BEE) 4:883
Best friendships 9:443
Best interest principles 12:956, 12:958
and euthanasia 8:282
Best interests 3:474–475
Best linear unbiased estimates (BLUEs) 14:145, 14:148
Best linear unbiased predictions (BLUP) 14:166, 23:192
Best n-term selection 15:650–651
Best practice 8:330
Best Practices and Local Leadership program (BLP) 5:779
Best system interpretation of probability 19:27
Bestuzhev-Lada, Igor 20:838
Best–worst scaling (BWS) 2:548–552
discrete choice experiments 2:548–549
models 2:550–551
maximum random utility models of choice and response time 2:551
models of ranking by repeated choices 2:550–551
multinominal logit models 2:550
notation 2:550
multiprofile case 2:549
extensions of models to 2:551
object case 2:548
profile case 2:549
extensions of models to 2:551
properties and uses of scores for Maxdiff model 2:551–552
Beta activity, in electroencephalography 7:354, 7:357
Beta distribution 6:576
Beta learning model, in learning spaces 14:589–590
Beta testing 22:358
Beta-adrenergic antagonist, for fear-potentiated startle 8:865
Beta-binomial (negative hypergeometric) distribution 6:583
Beta-endorphin 20:212
Betamax 6:638
Bête noire 20:919
Bethlem hospital, in the West 6:47
Bethmann-Hollweg, Theobald Theodor Friedrich Alfred von 9:213
Betts' Questionnaire upon Mental Imagery 15:225
Between the Lions (BTL) 7:249–251
Between-group competition 10:416
Between-group variance, population heterogeneity 17:733
Betweenness
and human rights 11:377–378
and network centrality 16:535–539
Betweenness centrality 22:462
Between-network research 1:54
Between-person emotions, in workplace 7:508
and emotional intelligence 7:508
Between-school allocation 21:57
Between-series designs 21:986–987
combining within- and between-series designs 21:987
Between-subjects designs 8:537–538
completely randomized designs 8:537–538
matched pairs and block designs 8:538
Between-unit information, in crossover designs 5:352–353
Betz cells 15:965

Beuchat, Henri 14:826
BEV *see* Battery electric vehicle (BEV)
Beverages, alcoholic 1:518
Beveridge, Wiliam 23:4
  model of health insurance 10:643–644, 21:336, 10:644t
Beveridge curve 15:595–596, 15:595f
Beveridge Report on Social Security 13:15
Bevölkerungspolitik 8:220
Beyond gen-ethics 2:627
BFAS *see* Big Five Aspect Scales (BFAS)
BFI *see* Big Five Inventory (BFI)
BFLPE *see* Big-fish–little-pond effect (BFLPE)
BGB *see* German Civil Code (BGB)
BGI *see* Beijing Genomics Institute (BGI)
BGI Ark Biotechnology Co. LTD 11:319–320
Bhagavad Gita 21:173–174
Bhagidari initiative 24:898
Bharatiya Janata Party (BJP) 20:404
Bhaskar, Roy 10:787–788, 18:627, 22:490–491
  critical realism theory of 20:23
  works
    The Possibility of Naturalism 20:23, 22:491
Bhikkhunī movement 25:632–633
Bhikkhunīs 25:629–633
Bhindranwale, Sant Jarnail Singh 20:376
Bhopal gas tragedy 24:143–144
BHPS *see* British Household Panel Survey (BHPS)
Bhutan
  education in 7:195
  sustainable development 23:819
BI *see* Behavioral inhibition (BI)
Biafran Civil War 12:739
Bianfa 16:277
Biao 11:49
Bias 5:318, 5:322–323, 5:913, 14:354, 17:531, 21:169
  approach to probabilistic thinking problems 19:18–20
  attrition 19:738–739
  Chinese conceptualization 5:323
  construct 5:323
  evidence-based medicine 8:329
  extreme response style
    correction 23:762
    evaluation of effects of 23:761
  gender, in research 9:716
  implicit 19:859–860
  information 8:329
  interviewer 5:851–853, 5:855
  irrationality 12:719
  item 5:323
  judgments, Tversky 24:681–682
  method 5:323
  objectivity of research 17:93–94
  prospect theory 19:263, 19:266
  randomization 19:738–739
  removal, by experimental designs 8:536–537
  selection 8:329
  sources 5:323
  strategies for 5:319t
  strategies for identifying and dealing with 5:324t
  substitution 19:738–739
  toward purely recipient-oriented theorizing 12:944–946
  of traditional methods for time-varying treatments 14:343–344, 14:343f–344f
  Western conceptualization 5:323
Bias estimates 11:751
Bias hypothesis, for persuasion 17:945
Bias processes 24:130
Bias theory, of repetition priming 11:711
Biased Competition Model 2:165
Biased groups 4:727
Bias-motivated offenses 10:540–541
Bias–variance dilemma 5:913
  in heuristics 10:832–834
Bible 21:173–174
  adoption in 1:143
  traditions of history writing 11:67
Bible Lands 17:423
Biblical Flood 20:286–287
Biblical *nebiim* 19:196
Biblical prophecy 19:197–198
  Weber's analysis 19:197–198
Biblical Ten Commandments 20:333
Bibliography 19:621
Bibliometrics 12:865
BIC *see* Bayesian information criterion (BIC); Behavioral Indicators of Conscientiousness scale (BIC)
Bichat, Xavier 18:174
BICS *see* Basic Interpersonal Communicative Skills (BICS)
Bicultural identity integration 1:70
Bicultural integration, minority 8:145–146
Biculturalism 22:321
  and social identity 22:321–322
Bidding Strategies 2:207
BiDil (drug) 9:900–901, 16:760
Bidimensional stage of development 16:504
Bidirectional relations 3:445
Biehl, Janet, works, *Rethinking Ecofeminist Politics* 6:846–847
Bielby, D., on media audiences 25:82–83
Bielefeld School of history 2:553–558
  comparative methods 2:554
  formative moment 2:553–554
  historical sociology 2:553–555
  historicism 2:553–554, 2:557
  historiography 2:553–555, 2:557
  institutional backings 2:555–556
  international relations and perspectives 2:557
  macro theory 2:554–555
  Marxism 2:553–554, 2:556
  modernity 2:553–555
  modernization 2:553–555
  program 2:554–555
  research 2:556–557
  social history 2:553–557
  and Weber 2:553–555
Bien publique 1:9
Biernacki, Richard
  on labor 5:537, 25:676
Bifurcated consciousness, and intersectionality 12:658–659
Big capital countering 17:667
Big data 5:738–739, 5:822, 6:901, 12:93, 16:798, 22:602, 23:735, 3:285f
  for social and behavioral science research 22:617–618
  technologies impacting 23:712
Big Five Aspect Scales (BFAS) 4:660, 9:233
Big Five Inventory (BFI) 4:660, 9:250
  10-item version 9:232
  44-item version 9:232
Big Five personality traits 1:378, 2:259–260, 2:561, 7:446, 16:743, 17:797–798, 17:822, 17:857, 17:865–867, 17:871
  aggreableness 1:463–464
  assessment of 17:830–831
  beauty and labor market 2:386
  domains, gender differences in 9:751–752
  extraversion in 8:625
  genetic and environmental influences 17:884–885, 17:885f
  Honesty-Humility 11:170, 11:173
  and political orientation 18:429
  procrastination 19:75
  and self-concept 21:465–466
Big history 11:45
Big law 13:581–583
Big men 3:141, 11:895
  African tribal societies 1:288
  anthropology of 2:567–573
  crony capitalism 2:571
  democracy and 2:571
  early work and developing analyses 2:567–569
  egalitarianism and 2:568
  ethnographic Foundations 2:567–568
  political typologies and developmental puzzles 2:568–569
  rethinking 2:569–570
  in Southeast Asia 23:71–72
Big science 9:551–552, 11:315
Big Seven Factor Model 9:232
Big Six model 17:915
Big Society 12:240, 20:565
Big Three model 17:914
Big women 11:895
Big-fish–little-pond effect (BFLPE) 1:59–61, 12:429–430, 21:463–464, 21:473, 12:430f, 21:464f
  educational implications 1:60–61
  empirical evidence 1:59
  mechanism 1:60
  other outcomes and moderators 1:59–60
Biggart, Nicole, on cultural logics of action 5:539
Bijective functions 17:297
Biker clubs 13:887
Bikini nuclear test site 5:416–417
Bilateral damage
  of amygdalae 13:680
  to structures of medial circuit 13:680
Bilateral economic influences of repression 18:451

Bilateral Investment Treaties (BIT) 12:470
Bilateral parallel processing 12:417
Bilateral trade negotiations 12:571
Bilbao effect 3:665–666
Bilbao Metropoli-30 association 24:841
Bilbao Ria 2000 24:841
Bildgeschichte 2:24–25
Bildung
  and aesthetic education 1:229
  and cultural participation 5:447
  Humboldt on 25:287
Bildungsbürgerlicher Renaissancismus 20:425
Bildungsbürgertum 2:786–787
Bildwissenschaft 2:25
Bilharzia 25:429–430
Bilinear model 5:3
Bilingual abilities of interlocutors 3:918
Bilingual aphasia 2:574–579
  assessment 2:577
  cerebral lateralization 2:575–576
  inhibition 2:574, 2:576–578
  lexicon 2:575, 2:577–578
  memory 2:575–578
  morphology 2:574–575, 2:577–578
  neuroimaging studies of bilingual brain 2:576–577
  Paradis' neurofunctional model 2:575–576
  phonology 2:574–575, 2:577–578
  pragmatics 2:575–578
  recovery patterns 2:574–575
  rehabilitation 2:577–578
  representation of languages in brain 2:576
  semantic 2:574–575, 2:577
  syntax 2:574–575, 2:577–578
Bilingual Aphasia Test (BAT) 2:577
Bilingual brain, and neurolinguistics 16:687
Bilingual corpus 3:921
Bilingual education 2:580–586
  additive bilingualism 2:580–581
  assimilation 2:581, 2:584
  definitions 2:580–581
  dual language education 2:581–584
  findings and challenges 2:582–584
  heritage language 2:581, 2:583–584
  history 2:584
  immersion 2:580–582, 2:584
  language awareness 2:583–584
  language maintenance 2:581–584
  majority language/speaker 2:580–584
  minority language/speaker 2:580–585
  motivation and aims of 2:581
  multilingualism 2:580–581, 2:584
  prototypes 2:581–582
    strong candidates 2:581–582
    strong–weak continuum 2:582
    weak candidates 2:582
  resilience 2:584–585
  submersion 2:580, 2:582–584
  subtractive bilingualism 2:580–581
  transitional bilingualism 2:581–582
Bilingualism 9:486–487, 13:301, 16:57–62, 16:66, 21:349
  cognitive control and 9:487
  interaction of bilingual factors 16:57f
  and language contact 13:301–302, 13:304–305
  language modes 16:59f
  language processing 21:349–350
  and linguistic anthropology 14:173
  Roma people 20:751
  word-processing 21:352
Bilinguals 16:57, 16:59, 16:65
  language lateralization in 12:694
Bilip (Urapmin) 20:308
Bill & Melinda Gates Foundation 7:267–268, 8:797, 22:659–660
Bill of Rights (South Africa) 21:744, 21:746
Billable hours 13:630
Billiard ball theory 2:872
Billing, psychiatric ethics 8:66
Billing clients 13:630
Billion bootstraps 24:821
Bills of rights 4:705, 6:696, 11:362, 19:535
  and constitutions 4:710–714
  liberalism and 14:23
Bilum tree 9:806
Bimetallists 14:409
Bin Laden, Osama 3:356, 12:744, 20:403
Bin width construction 19:30–31
Binary, nonparametric regression for 16:874–875
Binary choice tasks 5:906
Binary code 13:554–555
Binary comparison 4:412
Binary data, optimal experimental design 17:256
Binary decision diagrams (BDDs) 5:937
Binary multinomial processing tree (BMPT) models 6:525
Binary preference data, in forced-choice personality assessment 17:840–841
Binary response models 2:587
  example, multilevel statistical analysis 23:384
Binary-valued undirected networks
  centrality measures for 16:533–535
    medial measures 16:535
    radial measures 16:533–535
Binding 1:676
Binding, memory *see* Memory binding
Binding of signals 16:634
Binding problem 9:483, 25:188
Binding theory (anaphora) 1:676–678, 1:676t–677t
Binet, Alfred 2:596–599, 5:354, 14:811–812, 19:412
  biography 2:596–597
  contributions to social and behavioral sciences 2:598
  and imageless thought controversy 12:703
  intellectual functioning studies by 12:249
  on intelligence 12:309–311
  metric intelligence scale of 2:597–598
  psychometrically sound assessment scales of 2:598–599
  study of abnormal children 2:597
  works
    Introduction to Experimental Psychology 2:596
Binet test 3:879
  impact on clinical psychology 3:878–879
Binet–Simon Intelligence Scale 2:597, 12:310
Bing Maps 3:162
Binge drinking 1:519–520
Binge eating disorder (BED) 6:822, 6:825–826
  psychotherapy for 19:376–377
  wanting/needing 25:354
Bininj Gun-wok (language) 14:222
Binnie, Jon
  on gay/lesbian tourism 9:639–640
  on LGBTI commercial sex 13:884
  on sexual citizenship 21:730–731
Binocular organization, and occipital lobe 17:128
Binocular rivalry 25:189
  paradigm 4:674
Binocular visual space 25:198
Binomial data, generalized linear models for 17:259
Binomial distribution 6:581, 19:419, 19:781
Binomial error models 23:424
Binswanger, L. 19:314
  Daseinsanalyse 19:316
Bio queens 13:875
Bioarchaeology 2:600–603
  biomechanics 2:601
  challenges 2:602
  estimating biological relatedness 2:602
  fertility 2:602
  infectious disease 2:600–601
  lifestyle and activity 2:601–602
  paleodemography 2:600
  population profiles and demographic history 2:602
  quality of life and health 2:600–601
    reconstructing diet and inferring nutrition 2:600–601
    violence and injury 2:601
  from racial classification to biological process 2:600
  trauma 2:601
Biobank ethics 2:604
  issues in 2:605–608
    commercialization 2:606–607
    commodification 2:606–607
    consent 2:608
    fairness and reciprocity 2:606
    good governance 2:605–606
    privacy 2:607
    public goods 2:605
    risk 2:607
    trust and trustworthiness 2:607
    unintended effects 2:607–608
  naming and framing of 2:604–605
Biobank Japan 11:318
Biobanking 8:81, 9:979, 9:1002, 9:1018–1019
Biobanks 1:758–760, 2:621–622, 2:680
  feedback 2:622
  right to know information 2:622
Biobehavioral perspective, and pain 17:454
Biocapacity 3:158–159

Biocapacity (*continued*)
relationship between population density and 3:159, 3:159f
Biocapital 2:678, 9:1016
Biochemical factors of mania 14:486
Biocitizenship 2:732
and biopolitics 2:610–611
biosociality 2:611
and digital life 2:611
genetic advocacy groups 2:610–611
health activism 2:610–611
health care 2:610–611
and identity 2:611
life sciences 2:610–611
patients' organizations 2:611
political economy of 2:611
Biocolonialism 11:308
Biocultural co-constructivism 2:351
Biocultural studies 19:363
Biodata 17:851–853
Biodemography 2:656, 8:413, 8:418
of fertility 9:21
and Hutterites 11:446
of mortality 15:844
warranties 17:173–175
see also Biomarkers in demographic research
Biodiversity 3:851, 11:774–775
in Africa 1:295–296
Anthropocene 1:726–727
and land use 13:238–241, 13:243
and social justice 9:972
Biodiversity ceiling 3:157
Bioecological system 7:243–244
Bioecological systems theory 8:732
Bioecological theory of intelligence 12:294
Bioeconomic cooperation 22:267–268
Bioeconomics 6:853
Bioeconomies 2:678–679, 9:977–978, 23:442
Bioethical brokers 2:873
Bioethical theory 2:626
Bioethics 1:763, 2:613–624, 2:680–681, 8:64, 13:557–564
abortion 2:613, 2:615
animal ethics 2:617
anthropology of 1:736–738
autonomy 13:557–560, 13:563
Belmont report 13:559–560
cases that shaped 13:560–563
conceptual analysis 2:633–634
concretizing medical bioethics and responding to biotechnology 13:559–560
contemporary, issues in 2:614–618
Declaration of Helsinki 13:559
definition 2:631
ethical, legal, and social implications (ELSI) program 8:43–45
European Conference of National Ethics Committees (COMETH) 13:560
euthanasia 2:615
genomics 10:32–41
and genomics, relationship between 2:619
government-sponsored research 13:561–562
cold war experiments 13:562
human plutonium experiments 13:562
U.S. sponsored syphilis studies in Tuskegee and Guatemala 13:561–562
history of 2:613
human nature 2:616–617
ethics of genome research and manipulation 2:616–617
neurosciences 2:617
informed consent 13:559–560, 13:562–563
medical ethics 2:615–616
allocation of resources 2:616
beginning of life 2:615
clinical practice 2:615–616
end of life 2:615
method of 2:613–614
international and intercultural aspects 2:614
morality and philosophical ethics 2:613–614
natural and social sciences 2:614
modern bioethics movement 13:558–560
Nazi experimentation 13:560–561
National Research Act 13:559–560
Project MKULTRA 13:562
Nuremberg Code 13:558–559
patient rights 13:558, 13:560–561, 13:563
patient–doctor relation 2:615–616
philosophical aspects 2:631–636
philosophical methods 2:633–635
philosophy and 2:631–632
philosophy and social problems in 2:632–633
in postgenomic age 2:625
beyond gen-ethics 2:627
changes in methodological approaches 2:625–627
dual-use problem 2:629
genomics 2:625, 2:627
human enhancement 2:629
international 2:628–629
neuroethics 2:627–628
population-level 2:628–629
public health ethics 2:628–629
vision for future 2:629
prisoner research 13:562–563
research on children and the mentally ill 13:561
Pfizer Meningitis Study in Nigeria 13:561
Willowbrook Hepatitis Study 13:561
research on the elderly 13:560–561
Jewish Chronic Disease Hospital Study 13:560–561
research on women 13:560
coercive contraception 13:560
scope of 2:613
society 13:557–558, 13:561
synthetic biology 2:617–618
traditions in 2:635–636
World War II 13:558–560
Bioethics and Biosafety Act 2005 23:443
Bioethics and population studies, microevolution 15:341
Biofeedback 2:408
Biofeedback trainings
for asthma 20:578–579
for chronic obstructive pulmonary disease 20:578–579
Biogeochemical cycles 9:567
Biogeographical ancestry 9:899
Biographical approach 3:25
Biographical factors, in offender rehabilitation 20:199
Biographical itinerary 1:324
Biographical methods, interviewing 12:673–674
Biographical studies 2:648
Biographical turn in history 2:646–648
Biographies 3:25, 6:494–495, 11:61, 21:34
collective identities 2:639
gender and intersectionality 2:640–641
history, memory and politics 2:638–639
leisure 13:855
methodology and methods
developments in 2:638
qualitative/interpretive 2:638–641
migration and transnational biographies 2:639–640
of Prophet Muhammad 11:63–64
psychoanalytic 19:327
research development 2:641
social exclusion and inclusion 2:641
Biographies, historical
2000 Years of 2:644–645
biographical turn 2:647–648
challenges 2:647–648
budding renaissance 2:646–647
defined 2:644–645
methodological approach 2:645–646
Biographization 11:552
Biohistories
and society–nature coevolution 22:259
Biolinguistics 16:678
Biological analogy, for industrial ecology 11:843–844
Biological and linguistic evolution, parallel between 8:367–369
Biological anthropology 19:313
Biological approaches
to extraversion 8:624
panic disorder 17:469–470
biological models 17:469
evidence regarding 17:469–470
to suicide 23:682–683
Biological aspects
of multilingualism 16:62
of suicide 23:678–679
Biological circadian rhythms 3:570–571
Biological correlates of negative symptoms in schizophrenia 19:292–293
Biological desiderata 2:792
Biological determinism 22:901–902
and gender 9:788–790, 9:793
see also Genetic deficit theory
Biological disposition of anxiety 1:806
Biological Economies (BE) research program 10:76–77, 10:80–81
Biological entities, autonomous agents 2:312
Biological evolution 16:332

evolutionary economics 8:430–435
functional explanation 9:522
Biological factors
of gender gap in life expectancy 9:704
measures, and infectious diseases 12:34
in post-traumatic stress disorder 18:704
Biological hazards 6:480–481
epidemics 6:480
Biological influences
on gender development 9:742–743
sensation seeking (SS) 21:610–612
biochemistry 21:611–612
genetics 21:610–611
psychophysiology 21:612
sexual orientation 21:773–778
Biological information processing 14:597
Biological kinship 1:143, 13:61
Biological maturation 17:886–887
Biological mechanisms, physical activity 18:88
Biological model
of panic disorder/agoraphobia 17:477
of personality 17:857–858
Biological models of personality 17:857–862
Biological mother, role in childcare 3:396–397
Biological origins, of number sense 17:49–51
Biological pathways
and health 22:369
and marital quality 14:514–515
socioeconomic status and health 22:934–935
Biological relatedness, estimation 2:602
Biological risk factors, obsessive-compulsive disorder 17:122
Biological sciences
commercialization of 17:584
patenting in 17:584–591
Biological sensitivity to context
evolutionary developmental psychology 8:424
Biological sex 9:698
Biological substrates, of personality 17:906–907
Biological theories
of intelligence 12:304–306
and prosocial behavior 19:229–230
of socialization and education 22:862
Biological warranty period 17:173, 17:176
Biological waters 25:437
Biological weapons 16:255
Biological/genetic factors on determining sex 21:678–679, 21:683t
Biologically linked markers, of transition to adulthood 24:578–579
Biologically primary abilities 8:394
Biologically productive area 3:159
Biologically secondary abilities 8:394
Biologization 5:475, 6:32
Biology 7:878, 8:463–465
of bipolar disorder 2:686–687
fishing 9:229
and functionalism 9:536
genetic factors (in organizational psychology) 9:919
and genetics 9:905
hybridization in 11:448–449
and institutionalism 12:183–184
and motivation 2:326–327, 2:329–330
and reproduction 20:450–454
research censorship 3:280, 3:282
role in life span development 14:102
Biomarkers 5:738, 9:1001, 14:513, 21:947
sleep disorders 22:82–83, 22:85–86
Biomarkers, in demographic research 2:656–662
methodological challenges 2:657–659
anthropometric measures 2:657–658
blood pressure 2:658
functional measures 2:657–658
involving biological specimens 2:658–659
in population-based surveys 2:659–661
reasons for 2:656–657
Biomass 13:241–242, 13:242f
energy flow analysis 7:628–631
Biome studies 7:792–793
Biomechanics 2:601
Biomedical data, archiving 1:918–919
Biomedical model of pain 3:556
Biomedical research, involving human subjects, ethics for 8:48, 8:50
Biomedical research ethics
involving humans 8:78–82
beneficence 8:79–81
confidentiality 8:80–81
guidelines content 8:79–81
informatics and biobanking 8:81
informed consent 8:79–81
justice 8:79–80
other guidelines 8:81–82
placebos 8:80–81
privacy 8:80–81
respect for persons 8:79–80
subjects exploitation in developing countries 8:81
universalism/pluralism 8:78–79
vulnerability 8:79–81
Biomedical sciences 2:663
drug 2:665–667
epistemology 2:663–664
experiment 2:663–667
future opportunities for scholarship 2:667
holism 2:663–664, 2:666
innovation 2:664, 2:667
models 2:664
molecular biology 2:664, 2:666–667
as practice 2:664–665
clinical trials 2:665
diagnosis and etiology 2:664–665
risk and enhancement 2:665
reductionism 2:663–664
and society 2:665–667
biomedicalization 2:666
patient/disease advocacy 2:666
philanthropy, and the state 2:665–666
and political economy 2:666–667
Biomedicalization 2:666, 6:507
Biomedicine 10:568–569, 21:183
anthropological perspectives 15:106, 15:108
definition 2:663–664
genealogy of term 2:663
grounded in laboratory science 2:664
philosophical themes 2:663–664
War on Cancer 2:666
Biomedicine culture 10:573–574
Biometric genetic theory 9:934
Biometric statistics 19:705
Biometrics 8:225–226
Biometrika 17:660–661
Biometry
Pearson on 17:660–661
Bion, Wilfred 19:504–505
Bio-objects 9:979–980
Biophilia 7:737
Biophysical carrying capacity 3:158
Biophysical specificity, brain stimulation reward 2:844–845
Biopic 2:221
Biopiracy 11:774–775
Biopolitical tattooing 5:246
Biopolitics 7:666–667, 8:242, 20:458
and biocitizenship 2:610–611
and governmentality 10:309–311
Biopower 1:736, 5:247–248, 20:452
and reproduction 20:452–453
Bioprospecting 11:774–775
Biopsychological perspective, of love 14:377–378
Biopsychosocial factors 2:656–660
Biopsychosocial model
for aging 6:379–382
behavior therapy 2:419–420
of behavioral medicine 2:463–464
of disability 6:467
hypertension 11:454–455
of illness 11:268
of International Classification of Functioning, Disability and Health 9:543–544
theories of sexuality 21:850
and UK psychiatric social work 19:302–305
Bioregionalists 23:814
Biosocial approaches to crime 2:670–675
Biosocial contract 22:397–398
Biosocial criminology 2:670–675
behavioral genetics 2:672–673
brain studies 2:673–674
criminality 2:670–674
environmental influences 2:670–674
evolution 2:670–672
evolutionary perspectives 2:672
evolutionary psychology 2:672
molecular genetics 2:673
moving forward 2:674
twin studies 2:670, 2:672–673
Biosocial effects 8:572–573
Biosocial surveys 2:660
Biosocial theories 11:165
Biosociality 2:611
concept of 9:977–978

Biosociology *see* Evolutionary sociology
Biospecimen handling 9:965–966
Biosphere, energy flows in 7:627–628
Biosphere eco-city 6:831–832
  see also Eco-cities
Bio-statistical theory (BST) 10:611
Biostatistics 8:701
Biotech patents 17:584
Biotechnologies 2:676–682, 3:374, 7:878, 10:40, 13:557–564, 20:287–288, 24:765
  agricultural 1:473–474
  commercialization of 2:678
  ethics in 10:32–33, 10:38–39
  expertise 2:680
  genetically modified organisms 2:677, 2:680
  genomics 2:677–679, 2:681
  governance 2:676, 2:679–680
  and health 10:574–575
  industry 2:678–679
  in intellectual property rights 12:262–263
  and medical bioethics 13:559–560
  molecular biology 2:676–677
  patents 12:270
  politics 2:679–681
    bioethics 2:680–681
    risk management 2:679–680
  reproductive medicine 13:40–41
  revolution 2:676–679
    institutional dimensions 2:678–679
    technological and epistemic space 2:677–678
  synthetic biology 2:678
  term 2:676
  transfer of 24:144–145
Biotechnology Revolution 1:566
Bioterrorism 3:280, 3:282
Biotic resources 23:37
Biotypology 8:248–249
BIP *see* Batterie Imagination et Perception (BIP)
Bipartite graphs 22:455
Bipedal transformation
  adaptation 2:723, 2:725–726
  bipedality 2:723–725
  body, evolution of 2:723
  body mass 2:723–726, 2:724f
  climate 2:724–725
  ecogeographic 2:724–725
  hominin 2:723–725, 2:724f
  human 2:723–725
  locomotion 2:723, 2:725
  sexual dimorphism 2:725
Bipedalism 15:472
Bipedality
  Body, evolution of 2:723–725
Biplots 5:3, 10:355, 10:355f
Bipolar adjectives, in semantic differential 21:563–565, 21:567–568
Bipolar affective disorder (BPD) 14:442–443, 15:213, 15:228
Bipolar cells 25:158
Bipolar disorder 2:683–690, 7:607, 12:420, 15:194, 15:213
  anticonvulsants 2:687–688
  antidepressants 2:685, 2:687–688
  antipsychotics 2:688
  biology 2:686–687
  clinical management 2:687
  cooccurring psychiatric illnesses 2:684–685
  course of illness 2:685–686
  depression 2:684
  diagnosis 2:685
  disclosures 2:689
  epidemiology 2:686
  history 2:683
  with hypomania 2:684
  lithium 2:687–688
  mania 2:683
  mental–physical comorbidity 4:390–391
  mortality 2:686
  prophylaxis 19:202
  psychodynamic characteristics 2:685
  psychopathological features 2:684
  suicide 2:684, 2:686–688
  treatment 2:687–688
  see also Mania
Bipolar I 14:485–486
Bipolar II 14:485
Bipolar illness 2:757–758
Bipolarity
  in Cold War 4:128, 4:132–133
  of maladaptive personality structure 9:246–247
  and realism 20:27–29
Biproportional adjustment 18:589
BIPs *see* Batterer intervention programs (BIPs)
Biran, Francois-Pierre Maine de 19:410
Birchmeier, Zachary, on networks and creativity 5:467
Bird song 14:367
Birds 14:369
Birkhoff's theorem 4:7
Birmingham Cultural Studies School 7:145
Birmingham Object Recognition Battery 19:251
Birmingham School 25:817
  theory of subcultures 25:816
  youth culture 25:808–811, 25:817
Birmingham University Cognitive Screen (BUCS) 10:767
Birth 16:867t
  anthropology of 2:697–699
  cultural anthropological perspectives 2:697
  fertility and 2:697–698
  fertility regulation 2:698
  human intelligence at 16:481–482
    controversial topic in philosophy and psychology 16:481
    developmental psychobiology of infant social mind 16:481–482
  infant mortality 2:699
  influences in adulthood
    parental and grandparental age at 6:755–756
    preterm births 6:753–754
    weight 6:753–754
  Lexis diagram 5:735
  life phase 5:509–510
  number, timing, and spacing of 8:788–789
  and parenting 8:457–458
  population register for 5:736
  premature birth 2:698–699
  prenatal care 2:698
  recurrent demographic events 5:740
  and religiosity of women 25:636
  and reproduction 20:450–454
  sexually transmitted disease 2:699
Birth cohorts 14:67
  generations 9:858
  social generation 9:859
  suicide risk in 23:686
Birth control 8:221, 8:247–248, 8:250–251, 9:3–4, 17:40–41
  and fertility transition 9:62
  and fertility transition, in China 9:47
  and motherhood 15:883–885
Birth matrix-mating rule (BMMR) model 18:589–590
Birth order
  influences in adulthood 6:754
  and sex selective abortion 21:687
  and sexual orientation 21:776–777
  see also Sibling order
Birth rate 2:699, 2:697t
Birth registration data 9:318–319
Birth squeeze 18:591
Birth weight 23:556–557, 23:559
  and infant mortality 12:19–20
  and twin studies 24:703–704, 24:707
Birthplace 5:474
Birthrates, Jewish 12:801, 12:803–804, 12:807–810
BIS *see* Behavioral inhibition system (BIS); Berlin model of intelligence structure (BIS)
Bisection function 24:309
Bisection point 24:308
Bisexuality 21:773
Bisexuals
  homophobia/heterosexism 10:790, 10:793
  Latinos in US gay culture
  see also Latinos
  and lesbians 13:894, 13:896–897
  LGBT movements 9:632, 9:635
  and queer anarchist politics 19:749
  queer theory
    and intersectionality 19:765–770
  sexual orientation 21:786–787, 21:789
  youth 9:649
  see also Lesbian, gay, bisexual, transgender (LGBT) individuals
Bishop of Caesarea 3:548
Bismarck, Otto von 4:340, 15:520, 19:569, 25:476, 25:529
Bismarck model 10:643–644, 10:644t
Bisphenol A 7:844
Bissenova, A. 3:318
Bistable modulation by opsins 17:271
BIT *see* Barriers to insulin treatment questionnaire (BIT); Bilateral Investment Treaties (BIT)

Bitter, and disgust 6:548
Biundi, G., works, *Vocabolaria Siciliano-Italiano* 14:423
Bivariate correspondence models 16:121–123
Bivariate model
growth with an exogenous outcome 14:331–333
for number of sex partners and drinking 14:331
Bivariate normal distribution 6:579
Bjerknes, Vilhelm 7:793
Bjorgo, Tore 22:431
BJP *see* Bharatiya Janata Party (BJP)
BJS *see* Bureau of Justice Statistics (BJS)
BK *see* Background knowledge (BK)
B&L politeness model 18:328–329
BLA *see* Basolateral amygdala (BLA)
Black, Donald 13:511
on social control 4:825
Black Act 1723 13:729
Black Action Defense Committee (BADC) 8:153–155
Black Africa and Malagasy law 13:619
Black and minority ethnic (BME) elders 15:440–441
Black Boomerang 25:396
Black box, of sciences and technologies 21:182
Black box causality 9:131
Black box generation
in stochastic optimization 23:474
efficient methods for 23:479
Black box of education 7:147
Black box of law 13:537
Black boxes (parable) 14:390
Black Catholics 20:294
Black children, protection of 3:420–421
Black Codes 5:875, 19:803
Black community work theories 4:380
Black Creoles 18:136
Black crime 5:183–184
arrest rates 5:184
imprisonment rates 5:184
property crime rate 5:184
ratio of violent crime rates 5:184
War on Drugs 5:184
see also Criminal justice system (CJS)
Black cultural capital 5:378
Black Death (bubonic plague) 6:480
Black empowerment 1:248
Black families 3:204
Black feminism 4:380
anthropology 8:899–900
criminology 8:915
feminists 7:118, 8:908
theory 11:2
Black freedom movement 1:305
Black helper 10:759–760
Black humor 11:417
Black isolation index 21:411–412
Black lesbian feminism 8:908
Black lesbians, representation in South Africa 21:747–748
Black metropolis 17:23
Black operations, war propaganda 25:396
Black Panthers 13:859
Black Power movement 5:84, 11:758–759
Black prisoners 18:927
Black Protestants (BPs) 20:294–295
Black racial identity 19:834–835
confrontation 19:834
dimensions of 19:834–835
integration 19:834
internalization 19:834
multidimensional model of 19:835
cultural dimension 19:835
physical dimension 19:835
psychological dimension 19:835
sociopolitical dimension 19:835
preconscious 19:834
Sanders-Thompson multidimensional model 19:835
Black radicalism 4:380
Black sheep effect 12:410
Black slaves, in South America 23:45
Black social work movement (BSWM) 1:305–306
The Black Stork 8:251
Black workers 23:158
Blackboard learning system 6:895
Blackboxing, in actor-network theory 5:524
Blacks 19:820–822
civil rights 3:692–694
discrimination of 5:154–156
and feminism 8:908
lesbian communities 13:871
oppression of 3:691
see also African Americans
Black–Scholes model 5:540, 14:632–633
Black–Scholes–Merton (BSM) theory 7:69
Blackstone, Sir William 6:634, 10:453, 12:508, 13:762, 19:169–170
Commentaries 4:249
Blackwell, Elizabeth 25:623
Blackwell, H. Richard 21:952–953
Black–white earnings differentials 1:256–257
Black-white relations 19:852
Black–white segregation 20:549–550
Blackwood, Evelyn, works, *Female Desire: Same-sex Relations and Transgender Practices across Cultures* 8:902
Blair, Tony 8:499, 22:842–843
Blair revolution 19:653–654
Blake, William 21:178, 22:836
Blame
and responsibility 20:581–582
Blameworthiness, and responsibility 20:587–590
Blaming 13:469, 13:594
in disputes 6:561–562
Blaming the victim 2:489
Blanché model 13:777
Blank slate, mind as 7:238
Blanket consent 9:965
Blanqueamiento 23:47
Blanqui, Auguste 22:837–838
Blasphemy, laws against 3:676
Blastogenesis 2:116
Blau, Peter 8:482
on social structure 22:724
theory of social exchange of 8:482–483
works
Exchange and Power in Social Life 8:482
Blau space 3:803
Blavatsky, Madame 2:891
Blend errors 23:230
Blended learning 6:567–569, 6:568t
Blended memory procedures, during childhood 3:465–467
Bleuler, Eugen 2:700–702
academic life of 2:700
basic concepts of association theory 2:700–701
biography of 2:700
on brain research 2:701
concept of 'degeneration' (Entartung) 2:702
on development of clinical psychiatry 2:701
Lehrbuch der Psychiatrie (*Textbook of Psychiatry*) 2:701
on schizophrenia
basic and accessory symptoms 2:701
works 2:700
writings on general psychology and philosophy 2:701–702
Blind patriotism 16:312
Blinding 24:538–539
in placebo-controlled double-blind studies 18:159–160, 18:162
Blindness 16:597–598
joint attention in children with 12:847–848
Blindsight 25:189
and unconscious mental activity 24:723
Bliss, William, works, *Encyclopedia of Social Reform: Including political economy, political science, sociology and statistics* 7:585, 7:591
Bliss point, reinforcement 20:209–210
Bloc Québécois 10:382
Bloch, Marc Léopold Benjamin, works 2:703–704, 4:407, 10:219–220, 11:40
and Annales school 1:708–712
anthropological approach of 2:703–704
on comparative historiography 2:703–704
contribution in feudalism 9:113–114
and economic history 2:703–704
on feudalism 2:703–704
on historiography 2:703–704
works
Annales d'histoire économique et sociale 11:40
Feudal Society 11:40
The Historian's Craft 2:704
La société féodale 9:113
Les caractères originaux de l'histoire rurale française (French Rural History) 17:670
Money and the Morality of Exchange 7:14
Royal Touch 11:40
Société féodale 22:311
Bloch, Maurice, on cognition 5:576
Block, Fred
on ideational embeddedness 5:536–537
on knowledge society 13:107–108

Block, Mitchel, works, *No Lies* 6:623
Block designs 8:538, 13:368
Block models 4:911–913, 4:915
structural balance theory 23:606–607
structure 4:911f
Block tests for outliers 23:389
Blockbusting, and residential segregation 20:542
Blocked random assignment 9:128
Blocking 15:156
in crossover designs 5:349, 5:353
nonequivalent group designs 16:852
Blocking effect, classical conditioning 3:765–766
Blockmodeling 5:557–558
and meaning 16:562
of social network data 19:433
structural equivalence 23:622–623, 23:625, 23:622f
Blockmodels 23:622
Blogosphere 12:587–588, 17:683–684, 21:251
Blogs 6:898
journalism 12:851, 12:855
Blood alcohol concentration (BAC) 1:508
Blood donation, and commodification 5:538
Blood feelings 9:810–811
Blood groups, Fisher's studies 9:223
Blood pressure
hypertension
reactivity 11:453–454
variability 11:455
and schizophrenia 21:24
Blood samples, repatriation of 9:966
Blood-Oxygen-Dependent ('BOLD' effect) 9:492
Blood-oxygen-level-dependent (BOLD) fMRI 9:526–527, 14:440
Bloom, Allan, works, *The Closing of the American Mind* 1:620
Bloom, Benjamin S. 14:753
on mastery learning 14:753–754
Bloomery iron smelting technology 24:131
Bloomfield, Leonard 2:705–709, 3:582
academic career of 2:705
on Algonquian languages 2:705
on anti-mentalistic program of linguistic analysis 2:708–709
on Austronesian language 2:705
biography of 2:705
on distributionalism 2:708–709
on standard behavioristic stimulus–response schema 2:707
structural analysis of language by 2:707–708
works
An Introduction to the Study of Language 2:705
Bloor, David 7:569–570, 24:136
positive support for reflexivity 20:94
strong program of 5:523
works
Knowledge and Social Imagery 20:94
BLP *see* Best Practices and Local Leadership program (BLP)
BLS *see* Bureau of Labor Statistics (BLS)
Blu, K.I., on race in North America 17:23
Blue and White 2:506
Blue Book of China's Cultural Industries 5:597
Blue Book Sociology 7:559
Blue Cards 11:665
Blue frits, usage in ancient Egypt 1:694–695
Blue space 23:294
Blue-collar industrial workforces 13:201
BLUEs *see* Best linear unbiased estimates (BLUEs)
Bluhdorn, Charles 9:157
Blüher, Hans 9:646
Blumenbach, Johann Friedrich 7:722
Blumenberg, H., works, *The Legitimacy of the Modern Age* 21:391
Blumer, Herbert 21:265, 22:267
on symbolic interactionism 23:856–857
works
The Polish Peasant in Europe and in America 7:560–561
Blumstein, Alfred 19:643
Blunting, coping styles, chronic illness 3:551–552
BLUP *see* Best linear unbiased predictions (BLUP)
Blut und Boden (German for 'blood and soil') ideology 13:228–229
BMDP (statistical package) 7:2, 23:377–378
BME *see* Black and minority ethnic (BME) elders
BMI *see* Body mass index (BMI)
BMIS *see* Brief Mood Introspection Scale (BMIS)
BMJ Group, Best Practice 8:330
BMMR *see* Birth matrix-mating rule (BMMR) model
BMO *see* Bounded mean oscillation (BMO)
BMOM *see* Bayesian method of moments (BMOM) method
BMPT *see* Binary multinomial processing tree (BMPT) models
BMTs *see* Bone marrow transplants (BMTs)
BN *see* Bulimia nervosa (BN)
BNC *see* British National Corpus (BNC)
BNM *see* Bare normativity of meaning (BNM)
BNP *see* British National Party (BNP)
BNST *see* Bed nucleus of the stria terminalis (BNST)
Board 14:265
Board games, in mathematics learning 4:71, 4:71t
Board interlocks 16:549
Board of directors 4:939–940
corporate governance 4:934
stockholders 23:491
voluntary organizations 25:261, 25:264–265
Board of Education of the Hendrick Hudson School District v. Rowley 23:222
Boas, Franz 1:746–747, 2:705–706, 2:710–713, 6:503, 7:145, 7:724, 8:138, 8:178–179, 14:837, 17:22–23, 19:677, 22:828, 23:1
anthropology, contributions to 1:746, 2:710–712
assumption of cultural heritage 2:710
biography 2:710–711
British functionalism 2:713
contribution to science 2:712
and cultural relativism 1:766
on culture of poverty 5:591–592
decisive period for 2:710
emic distinctions 2:711
on evolutionism 2:712
on explanations of symbolism 2:712
on functionalism 9:533
influences 2:712–713
neo-Boasian anthropology 1:749
on redistributive ceremonies 2:711
scientific explanation of unexplainable behavior by 2:713
study on American Indians 2:710
works
Handbook of American Indian Languages 25:576
Kwakiutl grammar 8:351
The Study of Geography 9:532
Boasian tradition, belief 2:495
Boasianism 1:747, 1:749
Bobbio, Norberto 2:714–718
biography 2:714
and democracy 2:714–718
and Italian Communist Party (PCI) 2:714
and jurisprudence 2:714
paradoxes and broken promises of democracy 2:716–718
role in Italian political life 2:714
works 2:714–718
The Absent Third 2:716
The Age of Rights 2:715–716
De senectute 2:716
Democracy and Dictatorship 2:715–716
The Future of Democracy: A Defense of the Rules of the Game 2:715–716
The Idea of the Republic 2:716
A Just War? 2:716
Left and Right 2:716
The Phenomenological Turn in Social and Legal Philosophy 2:715
The Philosophy of Decadence 2:715
Politics and Culture 2:715–716
Praise of Meekness 2:716
The Problem of War and the Roads to Peace 2:716
State, Government and Society 2:715–716
A Theory of Judicial Norms 2:715
A Theory of the Legal Order 2:715
The Use of Analogy in Legal Logic 2:715
Which Socialism? 2:716
The Absent Third 2:716
The Age of Rights 2:715–716
De senectute 2:716
Democracy and Dictatorship 2:715–716

The Future of Democracy: A Defense of the Rules of the Game 2:715–716
The Idea of the Republic 2:716
A Just War? 2:716
Left and Right 2:716
The Phenomenological Turn in Social and Legal Philosophy 2:715
The Philosophy of Decadence 2:715
Politics and Culture 2:715–716
Praise of Meekness 2:716
The Problem of War and the Roads to Peace 2:716
State, Government and Society 2:715–716
A Theory of Judicial Norms 2:715
A Theory of the Legal Order 2:715
The Use of Analogy in Legal Logic 2:715
Boccardo, Gerolamo, works, *Dizionario della economnia politica e del commercio: Cosi teorico come practico* 7:585
Bock, Philip K., works, *Handbook of Psychological Anthropology* 7:590
Bock–Aitkin EM (Expectation-Maximization) algorithm 19:438–439
Böckh, Richard 14:372
Bodhisattva 25:631
Bodies making population 18:602
Bodies of knowledge about emotions 7:531
Bodily adornment rituals, for transition to adulthood 24:579
Bodily defenses suppression, and infectious diseases 12:33
Bodily feedback, and embodied social cognition 7:410
Bodily movements 9:404
Bodily practice 8:196
Bodily symptoms, of hysteria 11:508
Bodily time 24:407–408
Bodin, Jean 7:721, 14:407
works, *Les Six Livres de la République* 1:6, 7:721
Bodinian absolutism 1:10
Bodleian Library at Oxford 5:146
Bodley, Rachel 25:623
Bodmer, Johann Jakob 15:11–12
Body 9:770–771, 10:43
anthropology of 2:719–720
asymmetry of 2:115–130
biographical significance of 2:638
cultural anthropology 2:720
cultural boundaries 2:720
and dance 5:667–668
disabled, normative assumptions of 6:454
early reflections on 2:719
eating disorder 6:822–828
effects of alcohol on 1:509
and embodied cognition 7:423
embodiment and culture 2:720–721
and emotions 7:473–475
expression of emotion 7:523
feminist analysis of 8:902
feminist theories 2:720
and gender archaeology 9:765–766
and health 10:571
history 9:771
history of 2:728–734
approaches 2:728–729
continuities, shift of paradigms, and new topics 2:729–730
culture 2:729–731, 2:733
emotions 2:728–729, 2:731–733
fields of research and types of analysis 2:730–731
functions 2:730–733
globalization 2:730–731
historiography 2:728, 2:730–731
image 2:729–731
mind 2:733
modernization 2:728–732
new trends and topics in twenty-first century 2:732–733
problem of corporeality 2:728–729
senses 2:728–730
western European history 2:728
imagery 2:719
interdisciplinary consolidation 2:721
and kinship 13:40
and mind 2:720–721
mind–body dualism 15:545–548
obesity *see* Obesity
of organism 4:415–416
phenomenology 18:20
religion 20:309–310
responsible for population census 3:290
and reproduction 20:450–454
and self 21:439–440
self-mutilation 24:32
and sexual orientation 21:776
as source in embodiment 7:410
tattoos 24:32–38
as theoretical problem 2:720
visual turn 25:211–212
Body, evolution of 2:723–727
bipedal transformation 2:723
body size as a baseline for other comparisons 2:726
variation in body shape 2:725–726
variation in body size 2:723–725
Body Action and Posture coding system (BAP) 7:523
Body art
tattoo as 24:36–37
use of term 24:37
Body dissatisfaction, prevalence of 2:735–736
Body dysmorphic disorder 17:113
Body function, and disability 6:467–469
Body image 2:735–740
across gender and life span 2:738–739
adolescents 2:735–740
adulthood 2:737–739
and cancer 3:74
childhood 2:737–738
children, media use and 15:39–40
coeducation schooling 3:929
concerns, preventing 2:739–740
determinants of 2:736–738
individual factors 2:737–738
sociocultural factors 2:736–737
in different cultural groups 2:739
family 2:736–737, 2:739
measurement of 2:736
media 2:736–740
peers 2:736–737, 2:739
prevalence of body dissatisfaction 2:735–736
self-esteem 2:735, 2:737–739
single-sex schooling 3:929
Body language, and communication 4:287
Body mass
evolution of body 2:723–726, 2:724f
Body mass index (BMI) 2:737, 4:747, 17:75
death and 11:319
and obesity 17:71–78
Body modification 18:849
Body movements, representations of 18:849
Body perception
interoception 12:614–616, 12:619
Body representations
in infancy 18:857
tactile processing contribution to 24:16–17
Body schema
like me theory 15:131–133
motor control 15:960–961
Body structure, and disability 6:467–469
Body techniques, sport 23:300–301
Body temperature, regulation by hypothalamus 11:489–490
Body weight perception 25:471–472
Body-awareness, and self 21:419–420
Body-centered frame of reference
versus allocentric frame of reference 23:119
versus retinocentric frame of reference 23:118–119
Body–mind–spirit model 23:288–289
of social work practice 10:532–533
Body-ownership
sense of 21:420
Body-part coding 1:875–876
Bodywork 21:650
Boeach-Gajano, S., on sanctity 15:577
Boehm, Gottfried 25:191
Boeing corporation 11:870–871
Boekaerts' model of SRL 21:536
Boer War (1899–1902) 8:239, 23:77
Bogen, Joe 16:686
Bogen, Joseph 23:294
Bohannan, Laura 1:743
Bohannon, Paul 13:469
Bohman, James 5:301
Böhm-Bawerk, Eugen von 3:91
Boileau, Nicholas 3:774
Boisguilbert, Pierre de 12:388
Boissevain, Jeremy, on brokers 2:871
Boko Haram 1:303
Bokros Package 18:692
Bolan, Richard 18:204
BOLD *see* Blood-oxygen-level-dependent (BOLD) fMRI
Bolívar, Simón 6:372–373, 23:46
Bolivarian Alliance for the Peoples of Our America *see* ALBA (Bolivarian Alliance for the Peoples of Our America)

Bolivarian alternative 5:370
Bolivarian Alternative for the Americas (ALBA) 1:624
Bolivarian republics 13:411–412
Bolivia 22:625
  craft production in 5:110
Bolla 15:578
Bolling, George M. 2:705
Bollywood 2:80
Bologna Declaration (1999) 2:741, 3:875
Bologna Process 2:741–746, 10:863, 10:868
  bachelor's degree 2:742–745
  competences 2:742–746
  definition 2:741–742
  doctoral studies 2:742–743
  effects of 2:744–746
  European Credits Transfer System 2:741
  harmonization 2:741–745
  impact on social work education 2:742–744
  master's degree 2:742–745
  new phase 2:744
  quality assurance 2:744
  research 2:742–743
  teaching methodology 2:743
BOLSA *see* Bonn Gerontologic Longitudinal Study (BOLSA)
Bolsa Familia 10:189
Bolshevik ideologists 16:292
Bolshevik Revolution 3:746, 8:227–228, 8:242, 20:859, 22:849
Bolshevik Social Democratic Party 18:400, 20:856–857
Bolsheviks 1:681, 4:303, 18:404–405
Bolshevism 4:302, 6:369, 6:373–374, 18:438–439, 22:834–835
  in Soviet Union 20:820
Bolshevization 18:405–406
Boltanski, Luc 2:747–754, 5:526
  biography of 2:747
  and Bourdieu 2:747–748, 2:751, 2:753
  on capitalism 2:747, 2:750, 2:753, 5:536
  on conspiracy 2:752–753
  critical sociology of 2:747–748, 2:751, 2:753
  on criticism 2:747–748
  on critique 2:747, 2:750–753
  on domination 2:750–752
  on emancipation 2:747, 2:751–752
  on justice 2:747–748
  on justification 2:748–752
  major works and contributions 2:747–753
  on power 2:748–753
  on pragmatism 2:747, 2:751–753
  on social class 2:747–748
  on social institutions 2:751–753
  on sociology of critique 2:747, 2:751–753
  works
    De la critique. Précis de sociologie de l'émancipation 2:751
    De la justification. Les économies de la grandeur 2:748
    Enigmes et complots. Une enquête à propos d'enquêtes 2:752
    La condition foetale. Une sociologie de l'engendrement et de l'avortement 2:750
    La souffrance à distance. Morale humanitaire, medias et politique 2:749
    L'amour et la justice comme compétences 2:748
    Le nouvel esprit du capitalisme 2:750
    Les cadres. La formation d'un groupe social 2:748
Bolton-Ockenfels model 19:903
Bolzano, Bernard, on heuristics 10:830
Bombay Provincial Municipal Corporation Act 1949 24:898
Bon festival 12:796
Bonaparte, Napoleon 6:372–373, 13:217, 17:423
  campaigns 25:363–364
  conscription 25:364
  fatherland, barrack and school 25:364
  French Revolution 9:409, 9:411
  westernization of world 25:364
Bonapartism 6:372
  and 'Caesarism' in Europe 6:372
Bond 17:334, 17:336–338
Bond markets 4:935
Bond model 16:531
Bondage, domination/discipline, submission/sadism, masochism (BDSM) communities 9:640–641, 19:754
Bondage, story of 20:331
Bondareva–Shapley lower bound 4:870
Bonded labor 22:44
Bondi, L. 7:166
Bonding social capital 16:520
Bonding theories 3:414
Bone
  chemistry 2:600–601
  death 5:871–872
  shapes 2:602
  structure, and activity 2:601
Bone marrow transplants (BMTs) 23:443–444
  model for 8:313, 8:313f
Bonerate 5:507
Bonham v. College of Physicians 12:897
Bonn Gerontologic Longitudinal Study (BOLSA) 6:378
Bonn Republik 2:21
Bono de Desarrollo Humano (BDH) cash transfer 3:408–409
Bonomo, Bruno 17:284
Bonyhady, Tim, works, *The Colonial Earth* 7:753
Book of Daniel 3:719–720
Book of Nature 14:243
Bookkeeping model, for prejudice reduction 18:837
Books
  and cultural capital 5:377
  and education 7:176
  and historiography 11:98–99
  university press 19:624
Booksellers 19:620–622
Boolean addition
  algebra 22:478–481
  of factors 4:442
Boolean connectives 14:295, 14:295t
Boolean distributivity laws in reasoning 16:925
Boolean Overlay 23:138
Boone, Christophe, genre categories 3:802
Boorman, S.A., on social networks 5:558
Boorse, Christopher, on diseases 17:3
Boosted modeling, in propensty scoring 19:142
Booth, Charles
  applied social research 1:851–853
  works
    Life and Labour of the People in London 7:559
Bootstrapping 16:119, 16:880–883, 20:492–494
  model selection by 15:656
  sample 20:492
BOP *see* Bottom of pyramid (BOP)
BoP/IB *see* Base of the pyramid/inclusive business (BoP/IB)
Bopp, Franz 20:324
Border ethnographies 2:763
Border/boundary theory
  of work–family conflict 8:760–761, 8:764, 8:761f
Borderlands 2:761
  Eastern Europe 6:811
Borderline personality disorder (BPD) 2:755–759, 9:244, 9:246, 10:501–502, 16:489, 16:491, 17:892–893, 24:278
  cognitive behavioral treatments 2:759
  conceptual shift 2:755–756
  diagnostic criteria 2:755t
  patients, fear-potentiated startle in 8:865
  psychodynamic psychotherapies 2:759
  structure 2:756–759
    differential diagnosis 2:757–758
    etiology 2:758–759
    treatment 2:759
Borderline personality organization 2:756–759
Borderology, in Mexico 1:630–631
Borders 18:377
  anthropology of 2:760–764
  cultural borderland 2:761
  of Eastern Europe 6:816–817
  and ecumene 7:111
  encountering, refugees 20:113
  frontiers 2:760–761
  frontiers and 9:464
  liminality 14:133–134
  pastoralism 17:582
  political borders 2:761–762
  securitizing 20:105
  social borders 2:762–763
  state borders 2:761–763
  statehood 23:348, 23:350
  symbolic borders 2:762–763
  territorial borders 2:761–762
  territorial conflict 24:214–220
  and territory 24:221–223

Borderscapes 2:763
Boredom
and common sense 4:2
entertainment 7:669, 7:671–672
Boredom susceptibility (BS) 21:608
Borgatta, E.F., works, *Encyclopedia of Sociology* 6:648
Borgatta, M.L., works, *Encyclopedia of Sociology* 6:648
Borges's imaginative and outrageous reduction forces 11:513
Borlaug, Norman 24:144
Born global firms 7:689–690
Bornstein, Kate, on gender outlaws 24:552
Borofsky, Robert 1:225
public anthropology 1:225
Bororo cosmology 13:72
Borrowing 11:24–25, 11:820–821
language contact 13:302–306
Borrowing strategies, job displacement insurance 12:829–830
Borrowing strength 23:185
Bortkiewicz, Ladislaus von 3:91
Bos indicus 16:840
Bosch, Hieronymus 3:146
Böschen, Stefan, on epistemic cultures 7:877
Boserup, Ester 2:765–767, 12:330–331, 25:641
on agricultural growth 2:766
Asian Drama research project 2:765
biography of 2:765
distinction between WID and GAD 2:766
gender and development (GAD) model 2:766
on marginalization 2:766
pioneering publications of 2:766–767
on population dynamics 18:568
theoretical contributions of 2:765–766
works
Persistent Inequalities volume 2:766
Woman's Role in Economic Development 2:766
Boserup hypothesis 11:408
Bosman Ruling 23:319–322
Bosnia and Herzegovina Household Survey Panel Series – Living in BiH 5:846
Boss, M., works, *Psychoanalysis and Daseinsanalysis* 19:492
Bossuet, Jacques-Bénigne 20:323
works, *Discours sur l'Histoire universelle* 20:323
Boston Diagnostic Aphasia Examination (BDAE) 3:792–794
Boston Gazette 6:567
Boston Gun Project 9:596
Boston Psychoanalytic Society 7:934
Boston Tea Party 22:506
Boswell, John 21:750
Botanical and zoological gardens 1:471
Botanical gardens 7:792
Bothered About Dungeons and Dragons (BADD) 15:590
Botkin, Daniel 7:792
Botswana, sociocultural aspects 23:75, 23:77, 23:79–80
Bott, Elizabeth 16:562
Bottle-feeding 2:848t
Bottom billion capitalism 24:821
Bottom of pyramid (BOP), and nongovernmental organizations 16:858–859
Bottomore, Tom 16:832
Bottom-up approach, demographic microsimulation 15:344
Bottom-up ascription, female managers 8:885
Bottom-up bias, visual selective attention 2:165
Bottom-up embodiment 7:409
Bottom-up methods, longitudinal data 14:348
Bottom-up modulation of right-ear advantage 6:360–363, 6:361f–362f
Bottom-up neural process 25:186
Bottom-up processing 25:153
Bottom-up view of mechanisms 14:878
Bouchard, Thomas 12:303–304
Boudjedra, Rachid
works
La pluie 21:753
La répudiation 21:751–752
Boudon, Raymond 1:668, 24:746
works,
The Logic of the Social 22:830
Tocqueville for Today 24:424
Bouglé, Célestin, works, *Essais Sur le Régime des Castes* 2:269–270
Bouhdiba, Abdelwahab 21:751–752
works, *La sexualité en Islam* 21:751
Bouillaud, J.-B.
on phonological disorders 18:46–47
Boulding, Kenneth 25:561
Boulé 6:61–62
Bounce back, resilience 4:354
Bounce forward, resilience 4:354–355
Bound Encoding of the Aggregate Language Environment *see* BEAGLE (Bound Encoding of the Aggregate Language Environment)
Bound serialities 16:265
Bound variables 16:6
Boundaries 3:202, 21:813–814
citizen science 3:632–633
democratic 6:93–94
ethnic 2:762, 8:149
of industrial ecology 11:849
Luhmann on 14:394
narrative networks 16:214–215
national 2:761–762
organizational 2:768–772
oversharing 17:439–441, 17:443
and personality disorders 17:441–443
political geography 18:377
postcolonialism 16:457
public sector organizations 19:599–600
redefining 19:600
resilience 20:557–558
social 2:762–763
of social systems and subsystems 21:529
state 9:464
of state 23:357–358
symbolic 23:850–855
territorial 9:464
territorial conflict 24:214–220
see also Borders
Boundary conditions
of cultural intelligence effects 5:437–438
genetic factors (in organizational psychology) 9:923
Boundary crossing 3:802
Boundary drawing 24:552–553
Boundary issues, postcolonial 18:661
Boundary making, and culture 5:618–619
Boundary marker, language as 13:252–253
Boundary objects 22:696
Boundary permeability, in meritocracy/tokenism 15:240, 15:243–244
Boundary problem, in differential diagnosis 6:387–388
Boundary strength, of classification systems 3:800
Boundary technique 8:632–633
Boundary tones 23:719
boundary vector cells (BVCs) 11:387
Boundary violations 21:815
psychiatric ethics 8:65
Boundary work, and identity 23:852–853
Boundaryless career 3:129–130
Bounded agency 15:941
Bounded confidence (BC) model of social influence 21:532–533
Bounded framing, and transfer of learning 21:99
Bounded mean oscillation (BMO) 14:633
Bounded observations 5:908
Bounded rationality (BR) 1:77–78, 2:444, 2:773–776, 4:488, 5:883–884, 5:911, 8:393–394, 17:353, 17:783, 18:290, 19:907–908, 22:704, 8:393f
applications 2:775–776
budgeting 2:775
electoral competition 2:775
evolution of cooperation 2:775
organizational reliability 2:775
prospect theory 2:775–776
subgovernments 2:775
asymptotic rationality 5:884
in behavioral economics 2:438–441
civil service 3:697–698
cognitive constraint 2:773–775
decision making 2:773–775
game theory 2:773, 2:775
and heuristics 10:830–832
hierarchies and markets 10:842–843
humans as information-processors 2:775
in learning 8:515
main branches in political science of 2:774–775
managerial and organizational cognition 14:481
minimax regret 5:883–884
and organization theory 1:104
and organizational theory 19:894
problem solving 2:773–775

Bounded rationality (BR) (*continued*)
rational choice theory 2:773–774, 19:902–903
satisficing theory 2:774–775
selective attention 2:773
Simon on 21:970–974
survival and evolution 5:884
working memory 2:773, 2:775
Bounded selfishness 2:439
Bounded solidarity 15:412
Bounded willpower 2:440–441
Bounding 20:883
Bounds approach 14:551–552
applications and extensions of 14:552
literature 14:554
structural methods 14:552–554
Bouraoui, Nina
works
Garçon manqué 21:753
La vie heureuse 21:753
Bourdieu, Pierre, works 2:777–783, 4:752, 5:557, 7:145–146, 7:415–417, 10:484, 12:121, 22:994, 23:309
on actor-network theory 5:525
on avant-garde art 2:322–324
and Boltanski 2:747–748, 2:751, 2:753
on capital 3:93
concept of cultural capital and theory of social classes 2:778–780
concept of distinction 12:588
concept of strategy 9:141
on conspicuous consumption/distinction 4:695, 4:697–698
on consumption 11:866, 16:555
contribution to cultural sociology 5:498, 5:501–502
on cultural capital 5:375–376, 5:378–380, 22:331–333
expansions and refinements of 5:376
on cultural participation 5:447
on cultural production 5:466
on economy of symbolic goods 2:778
elites 7:393–394
on fetishism 9:107
field as space of relations 5:466
on gender 2:777, 2:781
and health lifestyle research 14:115
on knowledge 13:72
and lifestyle 14:113–114
on meaning 16:561–562
on misrecognition and human sciences 11:384
multiple correspondence analysis method of 22:336
on new media 16:791
notion of habitus 2:777–778, 10:478, 10:486–488
notion of social capital 2:779–780
on public symbols and private discipline 23:863–864
reflexivity and engagement of 2:781–782
social inequality theory of 14:113–114
on social stratification 2:777–780
on social structure 2:778
on sociology 2:778–782
of culture 2:782
of education 2:779, 2:782
of science 2:782
on status 22:332
on structuralism 10:486–487
on taste and cultural capital 5:537–538
theory of class and cultural distinction 22:342
theory of domination by 2:778
theory of fields 2:780–781, 9:140–148
general theory 9:141–144
organizational fields and strategic action fields 9:146–147
origins and genesis of concept 9:140–141
uses and methodological issues 9:144–146
theory of practice 2:777–778, 5:553–554, 9:141, 18:783, 18:785
theory of state 2:781
theory of symbolic violence 2:781
use of correspondence analysis 5:8–9
works
Distinction: A Social Critique of the Judgement of Taste 7:16
Distinction and Rules of Art 5:500–501
Esquisse pour une autoanalyse 2:782
Homo Academicus 7:145–146
La Distinction 5:447, 5:451
Les Héritiers 2:779
Outline of a Theory of Practice 5:617–618, 10:486–487
Reproduction in Education, Society and Culture 7:291
The Rules of Art 9:142, 9:145
Sociologie de l'Algérie 2:777
Bourdieu and Passeron's theory 22:327
Bourgeois, Léon 9:466, 23:2–3
Bourgeois, Michael D., on knowledge and cultural consumption 16:557
Bourgeois class 4:686–687
Bourgeois democracies 6:369, 19:604
Bourgeois ideologies 10:743
liberalism and 14:24–25
Bourgeois law 22:844–848
Bourgeois liberties 9:399
Bourgeois pseudo-science, sociology as 3:249
Bourgeois Renaissancism 20:425
Bourgeois revolution 7:650–651, 7:654
French Revolution 9:409–411
Bourgeois sciences 22:141
Bourgeois society 4:339
Bourgeoisie 2:784–789, 3:739–740, 4:753, 7:379, 19:615, 19:697
capitalism 3:94–95
contraction and expansion 2:787–788
current research 2:788–789
definitions 2:784–786
formations of 2:785
functional equivalents 2:788
genesis of (eighteenth to nineteenth centuries) 2:786–787
as heterogeneous formation 2:785
historiography 2:788–789
history of 2:786–788
inclusive middle classes 2:785–786
limits of concept 2:788
and Marx, Karl 2:785
and Marxian theory 2:785
social formations in 2:784
Bourgeoisnationalist elitism 23:632
Bourgeois-Pichat, Jean 10:776
Bourguignon, F.J., works, Elsevier Handbook of Income Distribution 11:737
Bourricaud, François, works, *Critical Dictionary of Sociology* 7:589
Boutique medicine 2:623
Bovine spongiform encephalopathy (BSE) 20:706–707, 21:257
Bowen, Murray A., group psychotherapy 19:505
Bowen, W., on artists 2:58
Bower, Frank, works, *A Dictionary of Economic Terms for the Use of Newspaper Readers and Students* 7:590
Bower, Gordon, works, *Human Associative Memory* 4:105
Bowers, E.P., on adolescent self-regulation 21:550
Bowers v. Hardwick 21:770
Bowker, Geoffrey C. 21:188
Bowlby, Edward John Mostyn 2:790–794
see also Integrated attachment theory
Bowley, Sir Arthur Lyon 20:876
Bowling alone, Putnam's concept of 11:631
Bowman, Isaiah 18:374
works, *The New World: Problems in Political Geography* 18:374
Box plots 8:602, 10:353, 10:355–357, 8:602f, 10:353f
Boxenbaum, Eva, on translation of institutions 17:393
Box–Jenkins–Tiao model 12:652, 12:655
Boycott, divestment, and sanctions (BDS) movement 3:56
Boycotting 8:25–27
Boyd, N.A. 9:639
Boyd v. Albert Einstein Medical Center 10:595
Boyer, D. 2:873
Boyer, Pascal
works
The Naturalness of Religious Ideas 20:390–391
Religion Explained 20:390–391
Tradition as Truth and Communication 20:390–391
Boyle, Joseph 16:320
Boyle, Robert 8:577–579, 18:100–101, 21:213, 25:221
Boys
differential treatment of 9:744
education 7:119
educational performance 9:747
seclusion 5:512
BP *see* British Petroleum (BP)
BPD *see* Bipolar affective disorder (BPD); Borderline personality disorder (BPD)

BPN *see* Bandpassed noise (BPN)
BPO *see* Business process outsourcing (BPO)
BPRS *see* Brief Psychiatric Rating Scale (BPRS)
BPS *see* Business and professional services (BPS)
BPs *see* Black Protestants (BPs)
BPVS *see* British Picture Vocabulary Scale (BPVS)
BR *see* Bounded rationality (BR)
Brabant revolution 4:573
Brabantse revolutie 4:573
Bracero program 10:169–170
Bracken Basic Concept Scale 3:795
Brackenridge, C.H., on women's participation in sports 9:658
Bracketing, phenomenology 18:12
Bracton, Henry 20:781
Bradburn affect balance scale 7:490
Bradford, Judith 15:114–115
Bradford–Halifax School of model village builders 24:862–863
Bradley, David 25:561
Bradley, J.R., works, *After the Arab Spring* 20:640
Bradley–Terry–Luce (BTL) model 14:814
Brady Act 10:459
Brahimi Report 2000 17:633
Brahma Kumaris 16:806–807
Brahman groups 20:373–374
Brahmanism 25:628
Brahminical Hindu 11:78
Brahmins 3:223–225
Brahmo Samaj 20:374
Braid, James 11:465
  hypnotism studies 19:502
Braille (writing system) 16:597–598
Brain 2:847, 4:43, 4:46–48, 16:330–331, 21:932
  adolescent, and leisure 13:831
  aging *see* Aging brain
  asymmetry of 2:115–130
  bilingual, neuroimaging studies of 2:576–577
  correlates of right-ear advantage 6:363–365, 6:365f
  cortical areas in movement control, neuroimaging of 5:646
  and cybernetics 5:646
  dealing with motor redundancy 15:959–960
    equilibrium-point hypothesis 15:959
    optimal control theory 15:960–962
  defense system, and cardiac defense dynamics 23:544
  development 16:471
  differential aging
    successful cognitive aging 6:380–381
  and dreaming 6:651–657
  dysfunction 13:925–926
  educational neuroscience 16:728–735
  effects of alcohol on 1:508
  and empathy 7:543–545
  and false memories 8:713
  function, traditional tools for manipulating 17:268–269
  functional brain processes 16:601
  functional magnetic resonance imaging (fMRI) 9:528–529
  images 9:525
  and immune system 23:580–581
  and inferences 12:41–43
  and intelligence 12:290, 12:294
  interactive processes in 4:97–98
  investigation levels for studying 4:480f
  kindling
    areas, differential susceptibility of 13:32–33
    damage 13:33
  and language evolution 11:329–330
  lateralization 2:799
  learning process of 4:100
  localized regions, functional role of 16:632
  mapping of cognitive processes to 12:42
  mechanisms of emotion 7:479–481
    amygdala 7:479
    orbitofrontal cortex 7:479–481
  modeling technology 4:483
  multiple sensory representations in 16:94–95
  and neuroesthetics 16:657, 16:659–662
  organization of function
    interactive approaches to 4:97
    modular approaches to 4:97
  and pragmatic disorders 18:793–794, 18:793f
  regions
    essential for eyeblink conditioning 8:636–637
    implicated in phases of fear conditioning 1:811–813
    involved in moral functions 16:615f
    and memory 9:905, 9:907
    specialization of 4:96–97
  representation of languages in 2:576
  response to enriched environments 2:836–839
  and self 21:419–423
  and sexual orientation 21:775–776
  situated 16:600
  size 4:415–416
  social 16:601–604
  special mechanism, and flashbulb memory 9:255–256
  structural brain development 16:600–601
    in infancy 4:58–60
  studying LTP/LTD in 14:362–363
  systems
    computing transient spatial information 23:163–165, 23:164f–165f
    mediating route knowledge and cognitive mapping 23:166–167, 23:167f
  taste processing in 24:27–28
    effects of variety on food intake 24:27–28
    pathways 24:27
    pleasantness of taste of food 24:27–28
    primary taste cortex 24:27
    secondary taste cortex 24:27
    sensory-specific satiety 24:27–28
  theory 14:596–597
  and theory of mind 15:543
  topographic maps in 24:426–427
    sensory and motor topographic maps 24:427
  volume 2:830, 2:833, 2:829t
Brain, evolution of 2:827–834
  allometry 2:830, 2:832–833
  association cortex 2:827–828, 2:830
  brain volume 2:830, 2:833, 2:829t
  Broca's areas 2:831–832
  cerebral asymmetries 2:831–833
  cerebral cortex 2:827–828, 2:830–831
  direct evidence 2:827–828
  endocasts 2:827, 2:830–831
  human brain, characteristics of 2:828–831
    asymmetry 2:831
    brain organization and reorganization 2:830–831
    brain size, absolute and relative 2:828
    encephalization 2:828–830
  indirect evidence 2:828
  lunate sulcus 2:828, 2:830
  neurogenomics 2:827–828, 2:831
  paleoneurology 2:831–833
  reorganization 2:827, 2:830–832, 2:832t
  synthesis 2:831–833
  visual cortex 2:830
Brain areas 16:596f
Brain asymmetry 2:806–812
  absence of 2:806
  advantages of 2:809
  definition 2:806–807
  examples 2:806
  functional hemispheric 2:807–809
    left 2:807–808
    right 2:808–809
  and handedness 2:807
  individual-level 2:806
  ontogenesis of 2:809
  and pathology 2:810
  population-level 2:806
  and sex 2:809–810
  structural 2:809
Brain circuitry, obsessive-compulsive disorder 17:123
Brain connectivity
  and human language evolution 11:326
  within left hemisphere 13:402–403
Brain damage 23:335
  agnosia 1:440–441, 1:443
  neuropsychological rehabilitation 2:813–820
  perceptual deficits after 25:203–204
  recovery after 20:56
    age-dependent recovery 20:57–58
    behavior-specific recovery 20:57
    mechanisms underlying 20:58
    modulating functional recovery 20:59
    nature of 'recovery' after 20:56–57
    structural correlates of 20:58–59
  and visual mental imagery 25:163
  visuomotor deficits after 25:202–203
Brain death, electroencephalography for 7:361

Brain development 7:848, 20:56–57
epigenetics on 7:850
ontogenetic neurobiology of 2:821–826
apoptosis 2:824–825
cell migration 2:821–824
differentiation 2:823
embryonic development 2:821–822, 2:822f
epigenetic programming 2:821
and experience 2:825
fetal development 2:823–824
myelination 2:824–825
neural plasticity 2:823–825
neurogenesis 2:821–822, 2:824
neurotransmitters 2:823
phases 2:821
postnatal development 2:824–825
synaptic pruning 2:821, 2:824
synaptogenesis 2:823–824
prenatal/infant 18:853, 18:859
Williams syndrome 25:580, 25:582
Brain drain 3:132, 10:172, 10:252–253, 12:533–534, 12:536, 15:391–393
Brain Fag Syndrome (mental fatigue) 5:624–627
Brain function 23:280–281
and cognitive psychology 4:107
and mind 23:281
Sperry's contribution to 23:280
Brain gain 10:252–253
Brain imaging 1:399–400, 1:403–404, 9:476, 11:612–614
caveats and cautions of 13:490
developmental psychology 5:357–358
for dyslexia 6:729–730
techniques
CT 11:603
fMRI 11:603
fNIRS 11:603
PET 11:603
SPECT 11:603
Brain imaging procedures
for lateralization of language 13:398
Brain injury and pain 13:491–492
Brain lateralization, in animals 2:799–805
Brain lesions
effects on function 12:775
effects on serial order memory 21:639–640
Luria on 14:397–398
Brain mapping 9:476
Brain maturation 2:652–653
Brain mechanisms, evolutionary perspective 8:395
Brain networks, and semantic memory 21:576
Brain pathology, and confabulation 4:561
Brain plasticity 20:56
epigenetics 7:850–851
and language disorders 13:329, 13:333
and sleep 22:63–64, 22:67
Brain processes 20:357
Brain regions involved in speech production 23:250–252
arcuate fasciculus 23:251–252
basal ganglia 23:252
Broca's area 23:250–251
cerebellum 23:252
insular cortex 23:251
motor face 23:250
premotor cortex 23:250
supplementary motor area 23:252
Wernicke–Geschwind model 23:251
Brain reorganization, and language disorders 13:329–331
Brain repair, neuronal mechanisms of 16:620–621
Brain reserve 2:797
Brain reward circuits 16:640f
Brain scanning techniques, for aggression 1:377
Brain science, and anthropology 13:71–72
Brain size 4:415–416
and intelligence 12:305
Brain stem, and wake 22:88
Brain stimulation
and neurolinguistics 16:686
for obsessive-compulsive disorder 17:125
Brain stimulation methods 9:506–510
mental imagery 15:224
Brain stimulation reward (BSR) 2:841–846
addiction 2:842–843
anatomy and neurochemistry 2:842–843
discovery and context 2:841–842
dopamine 2:842–845
methodological issues and emerging directions 2:844–845
behavioral specificity 2:845
biophysical specificity 2:844–845
emerging directions 2:845
motivation 2:841–843
optogenetics 2:844–845
phenomenon 2:842
pleasure 2:843
schools of thought 2:843–844
hedonic 2:843
neuroeconomic 2:844
reinforcement learning 2:843–844
seeking and craving 2:843
Brain studies, about crime 2:673–674
Brain transplant devises 17:743
Brain Trust 22:698–699
Brain-based lie detection 13:492–493
Brain–behavior relationships
in infant development 18:858
Luria on 14:398
Brain–cognition–personality relations, aging of 2:796
Brain–computer analogy, and dualism 15:548
Brain–computer interaction (BCI) 2:816–817
cognitive enhancement 2:817
virtual reality 2:816–817
Brain–computer interfaces (BCI) 15:969, 16:666, 16:714
active 16:719
passive 16:719
reactive 16:719
Brain-death criterion 2:615
Brain-derived neurotropic factor (BDNF) 2:837–838, 2:848–849, 14:362, 14:364, 22:495, 24:533–534
Brain–gut axis 9:605
Brain–machine interface 15:969
Brainstem model of crying 16:477
Branch Davidians 16:807
Branch-and-bound algorithms 19:433–435
Branching processes
population dynamics 18:579–580
Brand, Adolf 9:644
works, *Gemeinschaft der Eigenen* (GdE) 9:644–647
Brand evaluation, in neuromarketing 16:701
Brand loyalty, Markov models for 14:577, 14:577t
Brandeis, Louis 19:555–556
Brandeis briefs 13:588, 13:767–768
creation of new law 13:768
determining facts 13:768–769
providing context 13:769–770
Brandenburg Gate, Berlin 3:771
Brandenburg-Prussian state 10:873
Brandes, Georg 16:830
Branding, in advertising 1:209, 1:213–217
Brandt, Vincent S.R. 2:72–73
Brandt Report 7:812
Brandtstädter, J., on adolescent self-regulation 21:549
Brans–Dicke field in gravitational studies 18:104
Brass, works, *Ethnicity and Nationalism: Theory and Comparison* 8:139–140
Brass, D.J., on negative social network ties 5:561
Brass, Freddy 15:111–112
Brass, William, on indirect estimation 6:143–146, 6:144t
Brass children surviving 6:143–144
Brass P/F ratio method 6:145
Bratman, Michael, joint action 12:840
Braudel, Fernand 1:708–712, 11:40, 25:671
and Mediterranean 11:40
structural theory 11:14
Braun-Blanquet, Josias 7:792
Bravery 10:522–523
Bray, Francesca 24:144
on agrarian political economy 1:457
works
A History of Agriculture in India 17:671
Science & Civilisation in China 17:671
Brazil 22:626–627
and Cardoso, Fernando Henrique 3:122, 3:125
carnival in 3:148
communities, pretend play in 18:871
gay/lesbian activism in 13:860–862
modern sociology in 3:123
Negroes in 3:123
Palmares 14:593
racial inequality 4:165–166
sociological tradition in 3:123
Brazil, Russia, India, and China *see* BRIC (Brazil, Russia, India, and China)

Brazil, Russia, India, China, and South Africa *see* BRICS (Brazil, Russia, India, China, and South Africa)
Brazilian Landless Farmers' Movement 1:684
Brazilian plantation economy 4:224–225
BRCA1 gene 9:971
BRCA2 gene 9:971
Breaching demonstrations 18:12–13
Breaching experiments, scientific controversies 21:282
Bread and circuses (entertainment) 7:670
Bread riots 22:509
Breadth citizen science projects 3:633, 3:635
Breadth-first search 2:32
Breadwinning, in dual-career families 14:606, 14:608
Break Index code 10:115
Breakdown point 20:731, 20:733
Breaking and entering 15:592
Breast cancer
  Bayesian experimental design for 8:524
  and physical activity 18:88
  and smoking 14:71–72
  test
    and social justice 9:971–972
Breast milk 2:847–848
Breast self-examination (BSE) 4:451
Breastfeeding 12:24, 2:848t
  brain 2:847
  child development 2:847–849, 2:851
  depression 2:847–848, 2:850–851
  effect on cognitive and neuronal development 2:847–851
    mechanism of action 2:848–849
    research findings 2:849–850, 2:850t
  intelligence 2:847, 2:849, 2:851
  and mortality rate 15:869
  perinatal psychiatry 17:725
  stimulation 2:847
  stress 2:847–849
  studies on 2:850t
  terminologies used to define 2:848, 2:848t
Breathing training
  for asthma 20:578–579
  for chronic obstructive pulmonary disease 20:578–579
  for panic disorder/agoraphobia 17:478
Breathlessness 20:576–579
Brecht, Arnold, works, *Political Theory: The Foundations of Twentieth-Century Political Thought* 21:169–170
Brecht, Bertolt 1:540
Breeding, genetic programming 2:42
Breiger, Ronald L.
  on music tastes 16:556
  on social network analysis 5:560
Breitinger, Johann Jakob 15:11–12
Brenner, Robert, on capitalist farming 1:457
Brentano, Franz 10:757, 19:413
Brentano, Lujo 3:21
Breslau, Daniel, on epistemic cultures 7:876
Bresnan, J., works, Lexical–Functional Grammar 4:830
Bretton Woods Agreement of 1944 13:16, 13:18
Breuer, Josef 9:414
  hysteria studies 19:502–503
  role in psychotherapy 9:414, 19:502–503
  works
    Jahrbuchfür psychopathologische und psychoanalytische Forschungen 12:903
    Studies on Hysteria 12:903
Breuilly, John 16:285
Breyer, Stephen 13:768
Brezhnev doctrine 4:305, 22:848
bRGCs *see* Basal radial glial cells (bRGCs)
BRIC (Brazil, Russia, India, and China) 1:941
Bricmont, Jean, works, *Chomsky Notebook* 25:344–345
Bricolage 17:393–394
Bricoleur analogy for structuralism 23:629
BRICS (Brazil, Russia, India, China, and South Africa) 3:100–101, 8:21, 10:182, 10:189, 14:281, 24:303
  law firms in 13:580–582
Bride-burning 6:644–645
Bridenbaugh, Carl 19:696–697
Brideprice
  determinants of 6:643
  prevalence of 6:642
  and property rights 6:643–644
Bridewealth 14:601
Bridge assumptions 1:75, 1:78
Bridge principles, of Albert 1:498
Bridge Schools 1:270
BRIDGES program 15:217
Bridging concept, governance as 10:298
Bridging inferences 12:42–43, 19:395
Bridging technology assessment 24:126
Bridging texts 10:332
Bridgman, P.W. 23:455
BRIEF *see* Behavior Rating Inventory of Executive Function (BRIEF)
Brief dynamic psychotherapy 19:344
Brief Infant-Toddler Social and Emotional Assessment Problems scale 22:946
Brief Mood Introspection Scale (BMIS) 7:521
Brief Psychiatric Rating Scale (BPRS) 19:290–291
Brief relational therapy (BRT) 19:346
Brief solution focused therapy 8:747
BRIEF-A *see* Behavior Rating Inventory of Executive Function for age from 18 to 90 (BRIEF-A)
BRIEF-P *see* Behavior Rating Inventory of Executive Function for ages of 2 and 5 (BRIEF-P)
BRIEF-SR *see* Behavior Rating Inventory of Executive Function self-report forms (BRIEF-SR)
BRIGHT *see* Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT)
Bringing Democracy Back In 1:527
Brint, Steven 12:278
  on social networks 5:558
Brinton, Mary C., on cultural capital 5:379–380
Briquet's syndrome 11:508
Brissett, D., works, *Life as Theater* 6:647
Britain *see* United Kingdom (UK)
Britannica 7:586
British Asians
  nonheterosexual 16:861–866
    forming and managing relationships 16:864–866
    identity, threats and coping 16:863–864
    identity process theory 16:862
    managing social representations 16:862–863
    social representations theory 16:862
    terminology for sexuality 16:861–862
  overview 16:861
British Association for the Advancement of Science 22:697
British birth cohort study (1946) 14:68, 14:73
British Broadcasting Corporation (BBC) 14:693, 14:708, 19:526–527, 19:529, 24:164
  Prison Study 4:155, 24:713–714
British Colonial Development and Welfare Act 1939 6:308
British colonialism 11:1
British Columbia Assembly of First Nations (BCAFN) 11:767
British contributors, historicism 11:94
British Council of Disabled People 6:462
British Crime Survey 5:719, 5:848
British cultural studies 2:852–855, 5:501
  analysis of popular culture 2:852–853
  media audiences 2:216–217
  politics of 2:854
  representation and identity 2:853–854
  texts and audiences 2:853
British Dependent Territory Citizens (BDTC) 10:310
British East India Company (EIC) 2:920
British Educational Reform Act 1988 21:136–137
British Empire 16:300–301
British empiricism 7:567–568
  and learning 13:701
British eugenics 8:225–226
British expansion 11:688
British Folklore Society 9:280
British historiography, welfare state 25:528–529
British Household Panel Survey (BHPS) 5:744–745, 5:846, 6:379, 16:556, 17:463, 17:465
  happiness and work 10:516
British Industrial Revolution 11:873–874
The British Journal of Sociology 10:280
British legal system 13:628–629
British model 24:274–276
  history 24:274–275
  theoretical foundations 24:276
British Monopolies 18:344
British Museum 16:142
British Muslims 12:238

British National Child Development Study 17:462
British National Corpus (BNC) 4:950
British National Party (BNP) 18:410–411
British National Survey of Health and Development 17:463–464
British Nationality Act 13:785
British North America Act 1867 7:298
British Petroleum (BP) 17:166
British Picture Vocabulary Scale (BPVS) 5:357
British Probation of Offenders Act 1907 19:43
British Psychological Association Code of Human Research Ethics 20:518
British Psychological Society 20:515
British regional planning 20:144–145
British Road Safety Act (BRSA) 6:251
British social anthropology 3:6–7, 12:794, 16:562
British Social Attitudes (BSA) Survey 12:550, 23:749
British Socio-Legal Studies Association 13:515
British structural–functional anthropology 22:722
British structuralism 23:630–631
Broad beauty appreciation (level of consciousness) 4:668–670
  initial steps of developing conscious control of thought and emotion 4:668–669
Broad personality traits, gender differences in 9:751–752
Broad Street cholera outbreak (1854) 3:248
Broadbent, Donald Eric 2:856–858
  behaviorism 2:856–857
  cognitive psychology 2:857
  empirical psychology 2:857
  stress 2:856–858
  works
    Decision and Stress 2:857
    Perception and Communication 2:856–857, 4:103–105
Broadcast communications
  and politics 14:713, 14:715–717
Broadcast news 16:819–820, 16:822
Broadcasting 12:492, 24:164
  distributive model of 24:167
  policies 24:164–165
  and political communication 18:345
  regulation 2:859–863
    adapting broadcasting regulation to age of digitization 2:863
    authoritarian and democratic regimes 2:859–860
    commercial 2:860
    communications 2:860–863
    convergence, globalization, and competition 2:861–862
    market regulation model 2:860
    media policy 2:860, 2:862
    media regulation 2:860, 2:862–863
    public service broadcasting 2:860–861
    radio 2:859–863
    research on broadcasting regulations 2:862–863
    television 2:859–863
    types 2:859–861
  systems in four different countries 24:164–165
  and telecommunications 14:735
Broadcasting institutions 24:165
Broadcasting medium 24:163–164
Broadcasting research unit 19:528
Broaden-and-build theory 2:864–865
Broaden-and-build theory of positive emotions 2:864–869
Broadened visual attention 2:865
Broadening a rule, in legal reasoning 13:788
Broca, Paul
  on phonological disorders 18:46–47
  view on handedness 10:504–505
    incidence of left hemisphere language representation 10:504–505
    'organic predisposition' to handedness 10:504
Broca, Pierre Paul 2:806–807, 16:682, 18:46–47, 19:412, 16:682f
  neuropsychological rehabilitation 2:813–814
Broca's aphasia 1:819–820, 4:91, 13:398, 14:176, 23:250–251, 1:820t
  adjectival passives 14:180
  with agrammatism 16:690
    chronic aphasic patient, case example 16:690–691
    linguistic theory-driven accounts 16:691
    neurological correlates of 16:691–692
    processing of sentences 16:691
    syntactic processing 16:692–694
    WM mechanisms 16:694
  sentence processing and 14:176
  syntactic processes in 14:181
  thematic roles 14:178
Broca's areas 2:806–807, 9:268, 13:398, 13:404, 16:694, 16:740, 18:47, 23:250–251, 10:512f
  brain, evolution of 2:831–832
  and handedness 10:506, 10:509–510
  and language 1:818, 10:506, 10:509–510, 15:583–585, 16:586, 16:588–590, 16:593
Brodie, Benjamin 11:506
Brodie, Bernard 6:247, 15:520
Brodmann's areas 18:817
Broken windows theory 1:790–791, 5:191, 6:71
Brokers/brokerage 22:439
  anthropology of 2:870–874
    colonialism 2:870–872
    current trajectories 2:872–873
    death of broker 2:871–872
    development 2:871–873
    ebb and flow 2:873–874
    ethics 2:873
    go-betweens 2:870
    history 2:870–871
    inequality 2:871
    intermediaries 2:870, 2:873–874
    manipulation 2:871
    market 2:872–873
    mediation 2:870–874
    migration 2:870, 2:873
    nongovernmental organizations 2:872–873
    patron–client relationships 2:870–872
    politics 2:870–874
    return of broker 2:872
    state 2:870–874
    tourism 2:872
    transactionalism 2:870–871
  definition 2:870
  in Mexico 1:627–628
  probabilities elicitation 7:382
  social capital 22:182–184
Brokmeyer, Hans Conrad 10:740
Bronfenbrenner and bioecological systems theory 8:732
Bronfenbrenner's ecological systems theory 3:445, 7:243, 12:543, 21:682–684, 21:683t
Bronze age 23:66
  Europe in 8:270–271
Bronzeville 10:123–125
Brook Farm Institute of Agriculture and Education 2:149
Brookberg 3:582
Brookes, Barbara, works, *Sites of Gender* 11:21
Brooking, Tom
  works
    Environmental Histories of New Zealand 11:21
    Seeds of Empire 11:21
Brookings (think tanks) 22:656
Brooklyn Heights 10:46–47
Brooks, Byron, works, *Earth Revisited* 14:255
Brooks Law 15:719
Brophy, K., on gay/lesbian tourism 9:641
Brother–sister matings 2:450–451, 11:727–728
Brought forward time, concept of 15:205
Brown, Donald
  on history 1:748
  works
    History, Hierarchy, and Human Nature 1:748
Brown, George Spencer 14:383
Brown, Jerry 18:367–368
Brown, Roger 20:905
Brown goods 9:725
Brown v. Board of Education 3:692, 5:267, 7:298, 12:899, 13:542–545, 13:568, 13:770, 18:735, 23:220
Browne-Wilkinson, Nicolas 13:782
Brownfields 24:866, 24:921–922
Brownian motion process 6:396–397
Brownlow Report of 1937 18:861
Brown-Peterson paradigm, proactive interference in 19:3–5, 19:4f, 19:5t
Brownstoning 10:46
  see also Gentrification
Brozgal, Lia Nicole, works, *Against Autobiography: Albert Memmi and the Production of Theory* 15:136

BRSA *see* British Road Safety Act (BRSA)
BRT *see* Brief relational therapy (BRT)
Brubaker, Rogers. on independent nation-states 16:309
Bruce effect 17:198
Brücke, Ernst 9:414
Bruening, Meg, on eating behaviors 16:558
Brugmann, Karl 2:705
Brundtland Commission 23:823
  Our Common Future 23:812
  see also World Commission on Environment and Development (WCED)
Brundtland Report 24:867
Brunei 23:63
  fertility transition in 9:85–87
  sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Bruner, Jerome 2:875–879, 4:104, 12:341, 19:415
  cognitive and psychological theories of 2:875–876
  and Cognitive Revolution 2:875–876
  on cultural psychology 2:875–876, 2:878
  and education 2:878
  on human development 2:875, 2:877
  life and career of 2:875
  on mind 2:875–876
  theoretical sources and inspirations 2:876–877
  works
    A Study of Thinking 4:17
    The Process of Education 22:727
Bruner's developmental theory 2:881–884
  language acquisition 2:882–883
  modes of representation 2:881–882
Brunschwig, Henri 11:688
  works, *Mythes et Réalités de l'Impérialisme Colonial Français* 11:688
Brunt, P., on gay/lesbian tourism 9:641
Brunvand, Jan Harold 20:791–792
Brushing, in cartographic visualization 3:164, 3:164f
Brussels I Regulation 13:616
Brussels Treaty Organization (BTO) 4:130
Brussels/Lugano Convention on Jurisdiction 13:618
Brute facts 23:523
Bruxism 17:281
Bryce, James 2:272–273
  works
    American Commonwealth 2:270
    The Hindrances to Good Citizenship 2:270
    Modern Democracies 2:270
Brym, Robert 12:279
Bryson, Norman 25:209–210
BS *see* Boredom susceptibility (BS)
BSA *see* British Social Attitudes (BSA) Survey
BSE *see* Bovine spongiform encephalopathy (BSE); Breast self-examination (BSE)
BSEE *see* Bureau of Safety and Environmental Enforcement (BSEE)
BSF-R *see* Bayley Short Form-Research Edition (BSF-R) mental scale
BSID *see* Bayley Scale of Infant Development (BSID) scores
BSM *see* Black–Scholes–Merton (BSM) theory
BSN *see* Bachelor of Science in Nursing (BSN)
BSR *see* Brain stimulation reward (BSR)
BST *see* Bio-statistical theory (BST)
BSWM *see* Black social work movement (BSWM)
BTL *see* Bradley–Terry–Luce (BTL) model
BTO *see* Brussels Treaty Organization (BTO)
BTOF *see* Behavioral theory of the firm (BTOF)
Bubba psychology 15:894
Bubble Act 1720 9:186
Bubbles
  in financial crises 9:185–186
  rational expectations 8:513, 8:515
Buchanan, Elizabeth 12:607
Buchanan, J. 19:535, 19:538
  contractarianism 4:805–807
  on division of labor 6:604
  works
    The Calculus of Consent 23:459
Bucharest, Museum of the Peasantry 1:755–756
Buchi, Eva 8:208
Buckingham, James Silk 24:862–863
Buckland, William 20:287
Buckle, Henry
  on historical explanation 11:11–12
  works
    History of Civilization in England 22:697
Buckley, William F., Jr. 1:604
  works, *God and Man at Yale* 4:690
BUCS *see* Birmingham University Cognitive Screen (BUCS)
Budapest City Identity Office 3:662–663
Buddha 24:726
  biography of 2:886–887
  Brahmanical thinking of 2:887
  teachings of 2:886–888
Buddhism 2:886–893, 3:533, 16:806–807, 20:331, 20:360, 20:366–367, 20:369, 20:398, 20:408, 20:411, 25:669
  anthropology of 20:308, 20:310–311
  beliefs about mental illness 19:308
  Buddhist laity 2:890
  in Burma 2:886, 2:889, 2:892
  versus Catholicism and Protestantism 2:891
  in China 2:888
  conflict between Shinto and 20:367–368
  construction of sense of national identity, role in 2:892
  egalitarianism in 25:628–629
  engaged 2:892–893
  humanistic 2:892–893
  and Internet 20:277
  in Japan 2:886, 2:888–892, 12:796
  later developments 2:889
  lay meditation movement and 2:892
  Mahayana 2:888
  Meiji restoration and 2:889
  Modernism 2:891
  modernist reformers and 2:887
  monasticism 2:891
  and morality 15:808
  in Nepal 2:886
  "no self" concept 19:402
  political authority and 2:889–890
  in pre-Meiji times 2:891
  present times 2:889
  radical individualism of the Buddha's message 2:890
  reformism and 2:890–893
  ritual forms 2:890–891
  salvation and afterlife 2:890
  Sangha/Monastic Community 2:887
  and science 21:173–174
  scriptures 2:888
  soteriology 2:886–887
  in South Asia 2:76, 2:78, 23:54–56
  in Southeast Asia 23:70–71
  spread of 2:887–889
  in Sri Lanka 2:891–892
  in Thailand 2:891–892
  Three Jewels of 2:887–889
  vajrayana (tantric) 2:888–889
  vinaya rules 2:887, 2:891–892
  western views 2:891
  women in 25:628
Buddhism, engaged 2:892–893
Buddhist catechism 2:891
Buddhist chronology 24:410
Buddhist legal traditions 23:60
Buddhist Modernism 2:891
Buddhist tantra 16:168
Buddhist *vipassana* meditation 16:168
Buddhist wedding ceremony 2:890
Budget Control Act 2011 2:902
Budget formulation 2:904
Budgeting
  action plan, components of 2:896–897
  alignment 2:898–899
  and anticipatory management 2:894–900
  and political science 2:904
  and principles 2:898–899
  as incrementalism 2:901–902
  bounded rationality for 2:775
  capital 4:927–928, 4:930
  commitment 2:895
  competitive turbulence 2:898
  coordination 2:894
  cycle 2:895–896
  definition 2:894
  delegation 2:895, 2:897
  divided government and 2:904
  forecasting 2:894
  foreign assistance and 2:904
  institutions and 2:904
  opportunity gap, and performance gap 2:897–898
  philosophy of anticipation 2:894–895
  political agendas 1:364–365
  political science aspects 2:901–905
  purposes of 2:895
  rational choice approaches to 2:903–904
  requirements 2:897
  resource allocation 2:894, 2:898
  resource consumption and creation 2:897

Budgeting (*continued*)
responsibility center 2:894
reward systems 2:894, 2:898–899
state building and 2:904
study of 2:901
continuously updated shared causal model 2:897
monitoring system 2:897
sense of direction 2:897
trade-off between goal achievement 2:897
trust between participants 2:897
Budget-maximizing bureaucrat 13:572
Buffer function 10:59
Buffer model 15:162–163
Buffering effects, adolescent health 1:121–125
Buffon, Comte de 7:722, 20:286–287
Buffon, G.L., works, *Histoire Naturelle* 20:768
Bühler, Karl 13:637, 13:639, 16:684, 18:795–796
works, *Sprachtheorie* 16:684
Build effect 2:866–867
Building a Harmonious Society slogan 5:600
Building models 23:233
Building permits 9:317, 24:331f
spectrum of 24:333, 24:333f
Buildings, and gender 9:711–712
Bulgaria
anthropology 1:754
population momentum values 18:573
Bulgarian Law for the Protection of the Nation 8:220–221
Bulimia nervosa (BN) 6:822, 6:825
antidepressants for 1:780
psychotherapy for 19:376
Bulletin of the Business Historical Society 3:21
Bullying 25:538
and adolescent friendship 9:435–438
and childhood stress 23:571
cyberbullying 5:638–644
of LGBTQ youth 21:87–88
prevention programs 10:671
and prosocial behavior 19:228–229
in workplace 25:723
Bulmer, M., on interviewing 12:671
Bumiputra 23:71
Bumiputras policy 24:897
Bund für Menschenrecht (BfM) 9:646–647
Bund für Muttereschutz und Sexualreform 9:644
Bundesamt für Verfassungsschutz 18:410
Bundesrechtsanwaltskammer 13:621
Bundle of rights 19:153
Bundling strategy 21:651–652
Bundy, McGeorge 1:951
Bunshu shakai 12:85
Bunzel, Ruth 14:837
Bunzl, Matti, neo-Boasian anthropology 1:749
Buonarroti, Michelangelo 21:742–743
Bupropion 1:779
Buraku Liberation League 12:795
Burakumin 16:19
Burawoy, Michael 19:677–678
Burch, B., on butch–femme subcultures 13:870
Burckhardt, Gottlieb 23:18
Burckhardt, Jacob 2:906–907, 20:425
biography of 2:906
and concept of Renaissance 20:425–426
on Greek form of religion 2:907
historical writing of 2:906
historiographical principles of 2:906
individualism hypothesis of 20:427
works
The Civilisation of the Renaissance 2:907
Die griechische Kulturgeschichte 2:906–907
Die Kultur der Renaissance in Italien (*The Civilization of the Renaissance in Italy*) 2:906–907, 11:75–76, 20:425
Die Zeit Constantins des Großen 2:906–907
Franz von Assisi und die Anfänge der Renaissance in Italien 20:425
Kunst der Renaissance (*Art of the Renaissance*) 2:906
Über das Studium der Geschichte 2:906
works of art 20:425
Burden
caregiver 3:135–138
of disease 18:178–179, 19:546–548
cancer 3:73–75
of mental disorder 19:304
Burden of Disease Project 17:147
Burdens of proof, and presumptions 13:790
Bureau of Applied Social Research (BASR), US 13:636, 13:638–639
Bureau of Census Geography Division, US 1:847
Bureau of Economic Analysis (BEA), US 10:304, 11:734
Bureau of Eugenics of the Soviet Academy of Sciences 8:228
Bureau of Justice Statistics (BJS), US 5:718, 10:304, 10:457, 18:945
Bureau of Labor Statistics (BLS), US 5:742–743, 5:799–800, 10:303–304, 16:91, 24:384
Green Goods and Services Survey (GGS) 10:374–375
on green jobs 10:372, 10:374–375
Bureau of Safety and Environmental Enforcement (BSEE), US 17:373
Bureau of the Census, US 10:304
Bureau of Transportation Statistics, US 10:304
Bureaucracies 2:270–271, 2:908–912, 2:919–923, 3:106, 15:677, 17:201, 22:842, 25:660
Arendt on 1:962–963
authority 2:913–914
Bendix on 2:513
bureaucratic contradictions 2:915–916
versus business administration 2:909–910
and civil service 3:695–700
critical approaches 2:914–916
and democracy 2:916
dysfunctions of 17:415–417
expertized 22:696
hierarchies and markets 2:911
Hilberg on 10:870–871
impact on societies 2:916
and implementation 11:692–693
and inequality 2:916
internal flaws 2:915–916
liabilities of 2:910–911
displacement of goals 2:910
spiraling growth 2:910–911
vicious circles 2:910
Mills on 15:533, 15:535–536
models of administrative behavior 1:104
and modernization 15:707, 15:709–711
negative consequences 2:914
organizational design 2:909
organizations 17:307, 17:310
organized crime as 5:172
political sociology 18:473
and professions 19:101–102
public *see* Public bureaucracies
random decision making 1:105
recent trend 2:916–917
regulatory agencies 20:182–183
research on 2:911
sociology of 2:913–918
systematic problems 2:916
versus traditional organizations 2:908–909
use of term 2:913
Weberian model 2:908–909, 2:911, 2:913–914
departures from 2:914–915
in work 25:687, 25:689–690, 25:692
Bureaucracies, redistributive
in socialist societies 22:849–850
Bureaucratic categories, and refugee labeling 20:107
Bureaucratic leadership 17:201
Bureaucratic model 23:513
Bureaucratic obstructionism, in Soviet Union 20:832
Bureaucratic organization theory 15:511
Bureaucratic politics game 1:557–558
Bureaucratic rationality 21:263
and age 1:334
Bureaucratic regulation *see* Regulation and administration
Bureaucratic simplification, land tenure 13:233
Bureaucratic sleepwalker (obedience) 17:64–65
Bureaucratic theory 17:314
Bureaucratic/administrative model 13:808–809
Bureaucratization 2:908–917, 2:919–923, 12:749–751, 17:416, 20:407
antiquity to the nineteenth century 2:919–920
of bar 13:630–631
defined 2:919
general phenomenon 2:922
of law firms 13:631
parliamentary politics 2:920–922
war and 25:403
Bureaucrats 6:58
and delegation 6:58

Bürger, P., works, Prosa der Moderne 15:683
Burgerkrieg 3:713
Bürgerliche Gesellschaft 3:708–709
Burgess, A., works, *A Clockwork Orange* 24:119
Burgess, Ernest W.
on city growth 22:265–266
works
Introduction to the Science of Society 7:561
Introduction to the Science of Sociology 2:176
Burgess, John W. 18:467
Burgess, Robert 6:70, 22:409–410
Burglary 19:165
prevalence of 19:160
Burials
rituals 20:724
warrior 4:610
Burke, Edmund 11:368, 12:741–742, 21:212, 24:744
on American/French Revolutions 1:598–599
on conservatism 4:689–690
works
Reflections on the Revolution in France 1:599, 4:689
Burke, Kenneth 14:383, 15:292
Burke, Peter 1:712, 2:924–928
Annales school 2:924, 2:927–928
on early modern Europe 2:924, 2:926–927
on European Renaissance 2:927
Freyre and 2:928
on prosopography 2:926
on public sphere 2:927
on social history 2:924, 2:926–927
on social theory 2:924–926
on sociology of knowledge 2:927
Burkholder–Davis–Gundy inequalities 14:633
Burkinabe Response to Improve Girls' Chances to Succeed (BRIGHT) 3:408
Burma
Buddhism in 2:886, 2:889, 2:892
legal system in 13:814
postcolonial law in 13:815
Burmese Muslims 12:728
Burned-Over District 2:147
Burnett, Graham, works, *Masters of All They Surveyed* 7:753
Burnham, Daniel 16:809, 24:863–864
Burnham, Mac 13:32
Burnout 2:929–932, 17:337–338
conceptualization 25:534
definition 2:930–931
engagement 2:931–932
health outcomes 2:930
interventions, research on 2:931–932
job performance 2:931
job satisfaction 2:929–930
job–person fit 2:931
mediation role of 2:931f
occupational health 2:929–931
organizational factors 2:929–932
research, history of 2:929–930
stress at work 2:929–931
teachers 24:72–74
in workplace, psychology of 25:534–536
Burnout syndrome 25:537
Burns, K., on gay/lesbian tourism 9:639–640
Burns, L., works, *Busy Bodies* 24:401
Burrow, John W. 12:256, 12:258–259
Bursars 21:66–67
Burst firing of hippocampal cells 20:215
Burt, Cyril 3:243
Burt, Ronald, on structural holes 5:467
Burt matrix 5:8
Burton, R. 9:571
Bush, George H.W. 1:827
Bush, George W. 16:451, 18:424
presidency 18:861
Bush, Robert R. 14:812
Bush, R.R., works, *Handbook of Mathematical Psychology* 14:814
Bush, Vannevar 24:764
Bush fallow systems 11:203
Bushfire 6:480
Bushman myth controversy 25:666
Bushmen, San 9:901–902
Business
application of social interdependence theory in 4:860
greening of 10:392
capitalism 10:392–393, 10:395–396
China 10:392–396
corporate social responsibility 10:394–395
drivers 10:394–395
elements 10:393–394
energy 10:393
environment 10:394
finance 10:393–394
global warming 10:392–393
industrial changes 10:393
industrialization 10:392, 10:394–395
nonsuitability of BAU 10:392–393
resources 10:393
transformation 10:392, 10:394
transition dynamics 10:395
as innovators and users of technology 3:22–23
problems of defining 5:759
regulatory agencies 20:181–183
role in global public policy 10:182–183
role in intimate partner abuse prevention 12:679
and society 3:1–3
environmental factors 3:2
legal factors 3:2
political factors 3:2–3
social accounting 3:3
social factors 3:1–2
stakeholders 3:1–3
sustainability 3:2
technological factors 3:3
Business administration *versus* bureaucracy 2:909–910
Business and professional services (BPS) 7:689
Business anthropology
application 3:4–7
areas of inquiry and practice 3:7–9
consumers, marketing, and advertising 3:8–9
ethnographically informed design 3:7–8
finance 3:9
organizational cultures 3:8
British social anthropology 3:6
early influences 3:4–5
following World War II 3:7
history 3:5–6
Human Relations School 3:5–6
institutions
business anthropology 3:4–9
perspectives of Hawthorne project 3:5
Rockefeller Foundation 3:5–6
Yankee City Study 3:7
Business as usual (BAU) 10:392, 10:395–396
nonsusustainability of 10:392–393
Business city 3:664–665
Business codes 8:21–24
codes of ethics 8:23
corporate culture 8:23
Corporate Governance Guidelines 8:22
corporate standards 8:22–23
functions and effect 8:23
Guidelines for Multinational Enterprises 8:21–22
history and actuality 8:21–22
ISO 26.000 Social Responsibility 8:22
minimum standards 8:22
problem standards 8:22
societal cooperation 8:21, 8:23
soft laws 8:22
values management system 8:22–23
Business community 24:798–799
Business corporation, and legal personalities 13:753, 13:755–756
Business cycles 5:651–653, 14:406–407, 14:410–411, 24:356, 24:358
and crime 5:123–124
Business Data Toolset (BDT) framework 12:637
domains of 12:637
behavior-dependent trust 12:637
behavior-enacting trust 12:637
cognitive/affect 12:637
Business development 17:364–367
Business education 3:10–14
brief history of 3:10
corporate curriculum 3:13–14
developing person for business 3:12–13
innovation and creativity 3:12
knowledge productivity 3:13–14
lifelong learning 3:13
personal skills 3:11–12
quality of leadership 3:12
strategy and 3:10–11
technical skills 3:11–12
Business entities
corporations 3:29–30
limited liability 3:30
Business establishments, databases 5:742–743

Business ethics 3:15–20
competition's role, in modern society 3:17–18
contractarian approaches 3:17
deliberative/discourse approach 3:16–17
and economic ethics 3:15
history of 3:15–16
implementation of 3:17
order ethics 3:17–18
topics 3:18–19
Business geography 1:846, 14:547
Business groups 12:372
diversification 6:585
Business history 3:21–26
methods and approaches 3:22–25
biographical approach 3:25
businesses as innovators and users of technology 3:22–23
companies and entrepreneurs 3:24–25
comparative advantage 3:23
corporate governance 3:23
corporations as cultural institutions 3:24
family business 3:23–24
finance capital and industrial production 3:22
firms as venue of social interaction 3:23
marketing history 3:24
organizational capability 3:23
scholarly discipline 3:21–22
Business History Review 3:21
Business law 3:21, 3:24, 3:27–32
affecting the legal environment 3:31–32
basic categories 3:27
remedies 3:29
Statutes of Fraud 3:28–29
Business model 3:33–36, 17:367, 23:524–526
activity 3:33–35
activity system 3:33–35
conceptualizations 3:33–34
and entrepreneurship 3:34–35
future research, opportunities for 3:36
innovation 3:33–36
and innovation 3:35
and strategic management 3:34
value appropriation 3:33, 3:36
value creation 3:33–36
value proposition 3:33–35
Business necessity, and discrimination 6:531–532
Business networks, in Southeast Asia 23:72
Business organizations, mobile communication in 15:631
Business platforms 3:37–42
double-sided markets 3:40
ecosystems 3:37–41
industry architecture 3:38
industry platforms 3:38–40
managerial implications and further research 3:40–41
modularity 3:38–41
network effects and multisided markets 3:40
product platforms 3:38
supply chain 3:38–40
types 3:37–38
Business practice, international advertising as 12:458–459
Business process outsourcing (BPO) 3:43
development of 3:43–44
and firm-level decisions 3:44
Indian BPO sector 3:44–46
Business records 5:759–760
Business registers 5:761
applications, economic statistics 5:761
Business sector, heavy crime burden of 19:163
Business statistics, and economic statistics 5:758
Business strategy 23:526
Business support, and entrepreneurship 7:687–688
Business tax 23:340–341, 23:341t
Business-related crimes 5:178, 5:235
Business-to-business transactions 23:709
Business-to-client (B2C) relations 8:487
Buss, A.H., temperament development approach 24:191–193
Buss, David, works, *The Evolution of Desire* 22:904
Busyness
academic approaches 3:50–51
acceleration of social change 3:52
defining 3:49–50
and family 3:50–54
household ethnographies 3:53
organizational implications 3:51–52
research and theory germane to 3:54–55
time allocation 3:52–53
Butch–femme subcultures 4:726, 13:869–874
communities, evolution of 13:869–870
discrimination and psychological research 13:872
identities 13:870–872
intersectionality 13:871
research on butch identity 13:871
research on femme identity 13:871
methodological issues related to research 13:872–873
relationship within 13:872
Butler, Judith 3:56–60, 11:548–549, 19:747–749
biography of 3:56–57
on desire 3:57–58
on ethics 3:58–59
impact 3:59–60
on international theory 3:58–59
on Jewish philosophy 3:59
on performativity of gender 3:57–58
on politics of cohabitation 3:59
on power 3:57–58
on psyche 3:58
on universality 3:58–59
works 3:57–59, 19:754
Antigone's Claim: Kinship between Life and Death 3:58
Bodies that Matter: On the Discursive Limits of "Sex&rdquo 2:730, 3:57–58
Frames of War: When is Life Grievable? 3:59
Gender Trouble: Feminism and the Subversion of Identity 3:56–57, 19:754
Parting Ways: Jewishness and the Critique of Zionism 3:59
Phenomenology of Spirit 3:58
Precarious Life: The Powers of Mourning and Violence 3:59
The Psychic Life of Power: Theories in Subjection 3:58
Subjects of Desire: Hegelian Reflections in Twentieth-Century France 3:56
Undoing Gender 3:58
Who sings the nation-state? Language, Politics, Belonging 3:59
Butler, Richard 16:251
Butler, Ruth, works, *Mind and Body Spaces: Geographies of Illness, Impairment and Disability* 6:459
Butler Act 1925 3:280
Butterfield, Herbert 20:284
works
The Englishman and His History 3:62
The Whig Interpretation of History 3:61–62
Butterworth, Brian 10:110
Buycotting 8:25–26
Buyer, consumer psychology 4:738–739
Buying fun
entertainment 7:671
Buzzard, Thomas 12:774
BVCs *see* boundary vector cells (BVCs)
BWL *see* Behavioral weight loss (BWL) treatment
BWOR *see* Bank Wiring Observation Room (BWOR)
BWS *see* Best–worst scaling (BWS)
Bygren, Lars O., on cultural participation 16:559
Bylaws 13:622
Bypassed shame 21:875
Byres, Terence J. 1:457
on agrarian political economy 1:457
Bystander intervention, and hazing 10:560
Bystanders, role in mass killings/genocide 14:704
Byun, Soo-yong, on cultural capital 5:378–379
Byzantine Empire 3:707–708, 3:720

C

C4 catchment 10:115
C scale 2:257–260
CA *see* Conjoint analysis (CA) applications; Conversation analysis (CA); Correspondence analysis (CA)
CA3 pyramidal cells 10:887–888
CAADP *see* Comprehensive African Agricultural Development Program (CAADP)
Caballero, Spanish 1:965
CABARET system 13:787
Cabet, Étienne 14:644
works
The Real Christianity according to Jesus Christ 2:149

Travel and Adventures of Lord William Carisdall in Icaria 14:255
Cabinas, Pierre Jean Georges 19:410
Cabinet systems 10:314
executive government 8:497–501
Cable news, impact on politics 14:716
Cable News Network International *see* CNNI (Cable News Network International)
Cabo, Luis, on forensic archaeology 9:333
CAC *see* Command and control (CAC) regulation
CACE *see* Complier average causal effect (CACE)
Caches, rituals 20:724
Caciques 23:47
Caciquismo 23:47
Cacosmia 17:199t
CADAM *see* Content Addressable Distributed Associative Memory (CADAM)
Caddell, Patrick 18:461
CADISOP *see* Complete ADISOP (CADISOP) model
Cádiz, carnival in 3:146
Cadmium, exposure to 7:844
Caenorhabditis elegans 7:835
CAEP *see* Council for the Accreditation of Educator Preparation (CAEP)
Caesarism 6:372
CAFA *see* Central Academy of Fine Arts (CAFA)
CAFLIS *see* Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS)
CAFOs *see* Concentrated animal feeding operations (CAFOs)
CAGE Questionnaire 14:5
CAH *see* Contrastive Analysis Hypothesis (CAH)
Cahokia 1:633–634
CAI *see* Computer Assisted Instruction (CAI)
Cairnes, John Elliot 7:95
Cairo Conference 20:459–460
Cairo Programme of Action 8:795–796
Cajal–Retzius (CR) cells 2:823
Cakravartin 23:67
CAL *see* Computer-assisted learning (CAL)
Calabresi, G. 19:170
Calcium (Ca^{2+})
activity and synaptic function 12:712–713
activity-dependent control of neuronal development 12:711–712
in aging and degeneration 12:713–714
neuronal, activity-dependent increases in 12:710–711
Calcium carbonate 1:726
CALCOPT algorithm 23:624
Calculability, McDonaldization 14:828
Calculations 19:700
authority, in quantification 19:702–703
commensuration 19:702
trust and impersonality 19:702–703
by computers 4:516–517
Calculations, mathematical
connections of 14:780, 14:780f
components 14:779, 14:779f
Calculative thinking 23:798–799
Calendar 11:55
Calendar methods, questionnaires 19:774
Calendric rituals
and holidays 17:35–36
and nostalgia 17:35–36
Calgary Depression Scale 19:291
Calibration
electroencephalography 7:355
eyewitness testimony 8:643–644
of microsimulations 15:344–345
self-efficacy 21:518
and spatial interaction 23:155
California Act 7:298
California Assembly 6:250
California Civil Rights Initiative 1:265
California effect 4:734
California F-scale 2:262–263
California law 20:154
California prisons 11:723
California Rural Legal Assistance (CRLA) 18:735
California Urban Futures (CUF) model 24:891
California Verbal Learning Test, Second Edition (CVLT-II) 16:724
California's Museum of Vertebrate Zoology 21:265
California's Street Terrorism and Prevention Act 9:596
Caligula's gambling 19:16–17
Caliper matching 19:143
Call to conquest 15:473
Callon, Michel 5:524, 16:562–563, 21:311
on ANT 5:524–525
on economic theory 5:540
Callosal agraphia 1:454
Callosal inhibitory mediation 12:417
Callosal syndromes 12:420–421
Callosotomies 12:418
Callous unemotional traits 19:451–452
and aggression 1:377
and behavior during infancy and early childhood 6:13
Calmodulin 16:753
Calmodulin kinase II, and memory 9:906–907
Calotype 20:420–421
CALP *see* Cognitive/Academic Language Proficiency (CALP)
Calthorpe, Peter 16:809–810
Caluinisme 12:762
Calvin, Jean
Reformation/confessionalization 20:98–103
Calvin, John 2:145
Calvinist doctrine of predestination 12:645
Calvinist movement 3:540
CAM *see* Complementary and alternative medicine (CAM); Confusion Assessment Method (CAM)
Cambodia 23:63
early states 23:67
fertility transition in 9:84–89
metal age 23:66
postcolonial law in 13:815
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Cambodian Muslims 12:729
Cambridge controversies 3:92
Cambridge Face Memory Test (CFMT) 8:678
Cambridge Face Perception Task (CFPT) 8:678
Cambridge School 12:256–259
Cambridge School of Intellectual History 3:61–65
Cambridge method 3:62
modern political thought, origins of 3:64–65
Pocock's historical writing 3:62–64
Skinner's historical writing 3:62–64
Cambridge University
Fisher at 9:222
Cambridge University Eugenics Society (UK) 9:222
Cambridge University Press 7:587, 18:916
companions series 7:588
Cameralism, and political economy 18:357
Cameralists 22:696–697
Cameras 20:419
in the courtroom 9:396
Cameron, Edwin, on sexual minorities in South Africa 21:744
Cameroonian Nso mothers 2:141
CAMHS *see* Child and adolescent mental health service (CAMHS)
cAMP *see* Cyclic adenosine monophosphate (cAMP)
Camp aesthetic 13:877
Camp in drag culture 13:877
cAMP response element-binding protein (CREB) 12:712, 16:471, 20:215
Campaign activity regulation, political money 18:396
Campaign finance 18:393–396
Campaign for Nuclear Disarmament (CND) 5:80
Campaigns
and direct democracy 25:294–295
elections 25:294
information sources 25:295
question wording 25:294–295
regulation 25:295
for physical activity 18:95–96
political *see* Political campaigns
and political efficacy 7:302–304
Campbell, Colin 4:753
Campbell, Donald Thomas 3:69–72, 8:330, 8:569, 14:383, 19:739
on biases affecting the generation of knowledge 3:71
biography of 3:69
contributions of 3:69–71
on discriminant validation 3:70
education 3:69
honors 3:69
legacy of 3:71–72
participation in interdisciplinary conference 3:70
works

Campbell, Donald Thomas (*continued*)
Experimental and Quasi Experimental Designs for Research 9:132
Unobtrusive Measures: Nonreactive Research in the Social Sciences 3:70
Campbell, John L., on bricolage 17:393–394
Campbell, Richard, on division of labor 6:602–603
Campbell Collaboration 8:330, 8:544
Campbellian conceptualization of validity 12:450, 12:454
Campbell's law 24:768
Camphill communities 24:275
CAMs *see* Cell adhesion molecules (CAMs)
CAMSIM (simulator) 8:728, 15:345
Camus, Albert 18:437–438
Canaanite polytheistic religions 12:869
Canada 8:152–156
anti-Black racism in social work education and practice 8:154–155
archaeology politics 1:889
censuses in 3:305
Christianity in 3:540, 3:542–543
Conservative Protestants 3:542–543
evangelicals 3:543
New Spirituality 3:543
Protestantism 3:542–543
shift toward adopting no religion 3:543
clinical psychology, history 3:881
confederation 4:574
critical race theory and anticolonialism 8:154
education policies and education equity in 7:267
efforts to advance antiracism in social work 8:153
ethnicity, race, and crime in 5:155
historical perspective of race and social work education 8:152–153
immigration policy 11:662–667
impact of family status on labor market in 8:758
indigenous child welfare practice in 11:762–768
patenting in biological sciences 17:588
population composition by race and ethnicity in 18:554–555, 18:554t
visible minority population 18:555t
silencing of antiracism by anti-oppression perspectives 8:153–154
sociocultural aspects *see* North America; sociocultural aspects
tax system 23:340
trade unions and social work in 24:491–496
wrongful convictions in 25:787
Canada Geographic Information System 10:58
Canada's General Social Survey 5:155
Canada's International Development Research Center 24:295
Canadian Aboriginals 11:762–763, 24:509
discrimination of 5:151, 5:155
Canadian Association of Schools of Social Work (CASSW) 8:152–154
Task Force on Multicultural and Multiracial Issues in Social Work Education 8:152
Canadian Charter of Rights and Freedoms 7:298
Canadian Election Study (CES) 22:601
Canadian First Nations communities 14:187
Canadian Institute for Genocide and Human Rights Studies 10:7
Canadian Longitudinal Study on Aging (CLSA) 22:601
Canadian Multiculturalism Act 1988 8:152
Canadian Psychological Association 20:515
Canadian Society for Brain, Behavior, and Cognitive Science, establishment 3:881–882
Canadian Study of Health and Aging 15:846–847
Canadian Survey of Income and Labour Dynamics 13:387
Canadian Union of Public Employees (CUPE) 24:493–494
Canberra Group Handbook on Household Income Statistics 11:733
Cancer 3:73–78, 18:176–177, 20:271
and birth weight 6:753–754
burden of disease 3:73–75
childhood cancer 3:459–464
and childhood stress 23:572–573
and mood disorder 7:499–500
and mortality 15:860–861
pain 3:73–75
personalized medicine 17:925–926, 17:928
and physical activity 18:88
activity among patients 18:88
breast cancer 18:88
colon cancer 18:88
psychological distress 3:73–75
body image 3:74
cognitive dysfunction 3:73–74
depression and anxiety 3:73–75
fear of recurrence 3:73
in partners of cancer patients and couple 3:75
positive aspects 3:74–75
sexual impairment 3:74
psychosocial interventions 3:75–76
for couples 3:76
for patients 3:76
quality of life 3:73–76
religion and 20:267
screening 3:78–82
challenges 3:80
examples of programs 3:79
health service 3:79–81
individual 3:78–81
programs as policies 3:79
provision and uptake, determinants of 3:79–80
provision and uptake, implications of 3:80–81
psychosocial factors 3:79–80
public health 3:78–79, 3:81
rationale for 3:78–79
research 3:80–81
social and psychological aspects 3:79–81
society 3:79–81, 3:80t
and testing, difference between 3:78
sexual functioning 3:74, 3:76
skin *see* Skin cancer
social support and recovery 22:736
spouse 3:75
Cancer therapy, evolutionary medicine 8:455–457
Cancun Agreements 7:771
Candan, Can, works, *My Child* 20:873
Candidate Countries Eurobarometer (CCEB) 5:848–849
Candidate decision strategy 21:950
Candidate genes 9:826, 9:1009, 9:1012–1013
Candidate genetic association studies 12:301
Candidates
choice, in political campaigns 3:66–67
decision to run 3:66
by parties 3:66–67
Candolle, Alphonse de 21:236
Canguilhem, Georges 9:368
Cannabinoid receptor activators 20:213
Cannabinols 6:669
Cannabis
addiction 1:97–98, 1:100
effect on prospective memory 19:275
Cannan, Edwin 14:372
on division o labor 6:603
Cannibalism 3:83–89
anthropology 3:85–86
theory 3:86–88
anthropophagy 3:83–88
antiquity 3:85
changing perceptions 3:84
current trends 3:88
famine-based 3:84
innocent 3:86
invisible 3:86
medicinal 3:85
missionaries 3:83–84
natural law 3:84–85
ritual 3:86–88
sacrificial 3:86
savage 3:83–85, 3:87–88
sorcery beliefs 3:87–88
taxonomy 3:85, 3:88
technological 3:85
ubiquity 3:84
Cannon, Walter 23:540
Cannon–Bard theory 8:868–869, 8:869f
Canon law 3:234, 13:801
Canonical correspondence analysis (CCA) 5:5
Canonical form for casual explanations 8:596t
Canonical models 16:122–123
Canonical neurons 16:736–737
and embodiment cognition 7:421
Canonical order of thematic roles 14:178
Canonization 2:780, 15:578–580

CANTAB (computerized test) 8:495
Cantonese-speaking children 24:270
Cantor, M., works, *The Hollywood TV Producer: His Work and His Audience* 24:165
Cantril, Hadley, works, *Invasion from Mars* 7:563
Cao Đài 16:805–806
CAP *see* Child and adolescent psychiatry (CAP); Common Agricultural Policy (CAP); Community Action Program (CAP)
Capabilities
  neo-Aristotelian freedom as 9:401
  welfare as 25:508
Capability-based poverty measure 18:743
Capacity 4:355, 4:358
  of executive government 8:500–501
  for resilience 18:708
  of visual working memory 25:176
Capacity building 7:579, 16:553
  social work 24:755
Capacity decisions
  supply chain management 23:710
Capacity investment
  congestion 24:632
CAPI *see* Computer-aided personal interviews (CAPI)
Capital 3:90–93, 3:105–106, 3:585–592, 13:107
  versus capital goods 3:91
  circulating 3:90
  constant 3:90–91
  cultural 3:93
  ecological 23:824
  economic growth 7:39–43
  and exchange 8:477
  fixed 3:90
  health as 10:617
  human 23:823
  marginalist approach 3:92
  natural 23:823
  profits 19:111–115
  social 3:93
  successivist approach 3:91
  symbolic 3:93
  theories
    Fisher's 9:218–219
  use of term 3:90
  and value 25:7–11
  variable 3:90–91
Capital & Class 14:649
Capital accumulation 7:46–47, 13:885
  and culture of economy 11:865
  and political ecology 18:350, 18:352
Capital asset pricing model (CAPM) 23:484
Capital budgeting 4:927–928, 4:930
Capital punishment 7:82, 18:880, 19:652–653, 19:655–656, 19:658, 24:120
  crime policy, in United States 5:180–181
Capital structure analysis 4:930–931
Capital switching 3:579–580
Capital volume 5:9
Capital-based approach, of environmental stress 7:806
Capitalism 3:105–110, 22:833–834
  accumulation 3:105–106, 3:108
  agrarian political economy transition to 1:457
  and anarchism 1:680–681, 1:685
  anthropology of 18:365
  and aristocracy 1:968
  and artisans 2:47, 2:49
  Bell on 2:498–504
  Boltanski on 2:747, 2:750, 2:753
  bourgeoisie 3:94–95
  and business anthropology 3:4–5, 3:7–8
  in China 2:67–68, 22:979
  class interests and functional explanation 6:242–243
  class structure of 3:746
  classical contributions 3:106–107
  commercial 4:684
  commodification 3:105–106, 3:108
  concepts 3:94–95
  and conservatism 4:684
  contradictory 14:652–654
  credit 3:106–108
  crisis 14:677
  crisis tendency 14:653–654
  criticism of 3:18
  critique and analysis 3:106
  definition 3:105, 3:108–109
  dependency theory 6:196–197
  deregulated 3:98
  development, early stages of 3:95–97
  and disability 6:453
  distinguished from market economy 3:107
  and division of labor 13:158–160, 13:162
  and economic anthropology 7:19
  and economic geography 7:33–34, 7:36
  elites 7:386–388
  environmental sociology 7:799–800
  and fetishism 9:107
  and finance 9:172–175
  and finance 9:177
  and financial crisis 3:579
  financialized 3:98
  French Revolution 9:409–411
  and geography 10:85
  global *see* Global capitalism
  and globalization 10:218–220
  globalized capitalist system 3:97–98
  globalized laissez-faire 3:97–98
  Gorz on 10:287–289
  and greening of business 10:392–393, 10:395–396
  and growth 6:24, 6:27–28
  Hirschman on 10:892–896
  and housing 11:237
  industrial 3:96–97, 4:687
  and imperialism 11:686
  and industrial societies 11:863–866
  inequality 3:105–108
  investment 3:105–108
  and labor unions 13:200
  and land tenure 13:232
  and macromethods 12:645
  Mann on 14:490–495
  market 3:105–108
  Marx on 3:105–107
  and mass media 14:733
  and materiality 14:766–768
  merchant capitalism 3:95–96
  and migration 11:648–649
  millennial 14:437–438
  mixed economy 3:97–98
  as mode of production 3:94–95
  and modernity 15:687–688
  and modernization 15:701
  and national liberation movements 16:242
  and nationalism 16:293
  in nineteenth century 3:94–95
  in North America 17:24
  and Oceania 17:163
  origins 3:105–106
  and planning during postsocialism 18:187
  political contexts and recent revival 3:107–108
  and political economy 18:352, 18:358, 18:360
  political sociology 18:472–473
  and postindustrial societies 18:663–668
  and postmodernism 18:676–677
  problematic reproduction of 14:652–654
  and production 15:714–717
  proletariat 3:94–95
  and property 19:153–158
  and punishment 19:641, 19:655–657
  regulated capitalism 10:220
  regulated market system 3:97
  renewal 3:97
  and reproduction 20:450–453
  and revolutions 20:644, 20:648
  role of affect in 1:238
  Schumpeter on 3:107
  Smith on 22:119, 22:122–123
  social work under 14:664–667
  socialism 3:105–106, 3:108
  as socioeconomic system 3:94–95
  in South America 23:44–45
  spirit of 9:670
  state 3:97
  theories 3:94–95
  and transnational corporations 10:221–222
  in twentieth century 3:95, 3:97–98
  and urban policy 24:900–901, 24:904–905
  wage labor 3:105, 3:108
  Wallerstein's contributions 25:341
  war and 25:402–403
  wart and 25:402–403
  Weber on 3:106–107, 25:456–458, 25:463–466
  welfare state 3:97–98
  world systems theory 25:748–752
Capitalism, Nature, Socialism 14:649
Capitalismo Amarillo 24:821–822
Capitalismo e escravidão no Brasil meridional (*Capitalism and slavery in Southern Brazil*) 3:123
Capitalist accumulation 3:105

Capitalist class system 22:745
Capitalist crisis, defense economics 6:5–6
Capitalist development 1:665
Capitalist dynamics, and urban conflicts 24:784–786
Capitalist economy, and social question 22:554–555
Capitalist ethos 12:645
Capitalist globalization 6:316, 20:673
Capitalist ideology, celebrity and 3:274–275
Capitalist mode of production 3:105–106
Capitalist modernization 4:686–687
Capitalist society 13:105
Capitalist spirit 25:463
Capitalist transformation, historical geography 11:18–19
Capitalist world system, role in resource exploitation 7:749
Capitalization 22:564
Capitalized critical psychology 5:262
Capitation 12:921
Capítulo Criminológico 5:240
Caplan, David 16:678–679, 16:678f
Caplan, Gerald 18:884
  works, *Principles of Preventive Psychiatry* 18:884
CAPM *see* Capital asset pricing model (CAPM)
Capo 14:422
CAPS *see* Cognitive-affective processing system (CAPS)
Captivity
  and liberation theology 14:33
  risk–captivity–harm thesis 5:246–247
Capture–recapture data 18:604
Capture–recapture methods 14:121, 18:603–608
  assumptions, relaxing 18:606–607
  heterogeneity 18:604–606
  incomplete contingency table 18:604–607
  log-linear model 18:604–606
  model averaging 18:606
  models 18:604–606
    Bayesian paradigm 18:605–606
    ecological capture–recapture studies 18:605
    model selection 18:606
    social/epidemiological capture–recapture studies 18:604–605
    statistical software 18:606
  multiple recaptures 18:603–604
CAQDAS *see* Computer-Assisted Qualitative Data Analysis Software (CAQDAS)
Car dependence, in cities, and sustainability 18:199–200
Car theft 19:162, 19:162f
  levels of 19:165
CARA *see* Cue-adaptive reasoning account (CARA)
Carbado, Devon 5:268
Carballo-Diéguez, A., on stereotyping of Latinos 9:629
Carbon, pricing
  and climate change 3:864
Carbon democracy 12:122
Carbon dioxide (CO_2) emissions 7:701–702, 7:728, 9:567
  climate change, economics of 3:861
Carbon fuel 7:623
Carbon intensity 10:373–375
Carbon isotopes 17:458
Carbon markets 7:717
Carbon monoxide (CO) emissions, reduction in 19:543
Carbon neutrality 7:775–776
Carbon stocks 13:241
Carceral feminism 19:281
Card stacking 25:397
Cárdenas, Lázaro 15:326, 15:328
Cardenismo 15:328
Cardiac Arrhythmia Suppression Trial (CAST) 8:328
Cardiac defense
  characteristics 23:541–544
    attentional *versus* motivational significance 23:542–543
    descriptive patterns of response 23:541
    habituation and dishabituation 23:541–542
    individual differences 23:543–544
    stimulus characteristics 23:541
    sympathetic *versus* parasympathetic mediation 23:542
  danger 23:539–540, 23:542–544
  dynamics, and brain's defense system 23:544
  emotion 23:540, 23:542–544, 23:543f
  fear 23:543–544
  heart rate 23:539–544, 23:541f–543f
  noise 23:540–544, 23:541f–543f
  and stress 23:539–545
    cognitive approach 23:540
    historical antecedents 23:540
    implications 23:544
    motivational approach 23:540
    natural defense 23:540
    protective reflexes 23:540–541
Cardiac diseases
  coping with 3:111–116
    actual coping strategies 3:112–114
    anxiety 3:114
    approach *versus* avoidance coping strategies 3:113
    behavioral outcomes 3:114
    congestive heart failure 3:111–114
    coronary heart disease 3:111–114
    definitions of stress and coping 3:111–112
    depression 3:112–113
    dispositional coping styles 3:112
    emotional well-being 3:112
    future directions 3:114–115
    medical outcomes 3:114
    medical outcomes, and mortality 3:112
    myocardial infarction 3:111–114
    quality of life 3:111–113
    rumination 3:113
    stressors 3:111, 3:113–115
    threat appraisal 3:113
  and depression 7:497
Cardiac events
  acute psychosocial triggers of 4:920–921
    implications for prevention 4:921
    physiological mechanisms underlying triggering 4:921
Cardiac Modulation of Startle (CME) effect 12:616
Cardiac rehabilitation (CR) 20:187–192
  background 20:187
  definition of 20:187
  development 20:187–188
  effectiveness of 20:189–190
    educational intervention 20:190
    exercise 20:190
    life style and behavioral risk 20:189
    morbidity and mortality 20:189
    psychosocial risk 20:189–190
  life style and behavioral risk 20:188
  psychosocial determinants and impact on CHD 20:188–189
    anger and hostility 20:188
    anxiety and depression 20:188
    misconceptions, illness beliefs and self-efficacy 20:188–189
  risk factor reduction 20:188
  treatment beliefs 20:188, 20:190–191
  understanding and improving uptake and adherence 20:190
Cardiac response 23:539
  cardiac–respiratory infections 23:539
    hyperventilation 23:539
    respiratory sinus arrhythmia 23:539
  humoral pathways 23:539
  neural pathways 23:539
Cardiac side effects of psychopharmacotherapy 19:465
Cardiac–respiratory infections 23:539
  hyperventilation 23:539
  respiratory sinus arrhythmia 23:539
Cardinal 22:879
  rational choice explanation 19:899
Cardinal principles, Spencer's 23:275
Cardinality principle 17:52–53
Cardiometabolic risk (CMR) 21:21
Cardiovascular disease (CVD) 17:77, 19:464–465, 20:187–188, 20:271
  and birth weight 6:753–754
  cardiovascular reactivity as cause of 3:117
  and depression 6:209–211
  implications 23:544
  and mortality 15:857, 15:859–861
  and physical activity 18:87–88
    biological mechanisms 18:88
    coronary heart disease 18:87–88
    stroke 18:88
  prevention, comprehensive community-based health education for 10:648–649
    early international diffusion 10:649
    first decade 10:648–649
    projects begun in 1980s and 1990s 10:649

side effects (Psychopharmacotherapy) 19:464–465
social support and recovery 22:735–736
Cardiovascular reactivity (CVR) 3:117–121, 11:453
as cause of cardiovascular disease 3:117
laboratory research on 3:118–119
individual tasks/challenges 3:118–119
milieu 3:118
poststress recovery phase 3:119
procedural timeline 3:118
stressful challenge, selection of 3:118
methodological issues 3:119–120
duration of stressor exposure 3:119
frequency of measurements 3:119
individual difference characteristics 3:119–120
probe measures 3:119
type 1 error 3:119
psychosocial stress 3:117
research on 11:454
Cardiovascular recovery 3:117–121
Cardiovascular revolution 1:163, 15:861–862
Cardiovascular system, psychosomatic disorders in 19:487
Cardon, Dominique, on social networks 16:557
Cardoso, Fernando Henrique 3:122–127, 25:749
debate and controversies on 3:125–126
education, traditions and training of 3:122–124
international actions 3:124–125
in Latin America 3:122, 3:124–125
political career of 3:125
theory of dependency 3:124–125
works
The Accidental President of Brazil 3:125
Capitalismo e escravidão no Brasil meridional (*Capitalism and slavery in Southern Brazil*) 3:123
Cór e mobilidade social em Florianópolis (*Race and social mobility in Florianopolis*) 3:123
Dependencia y Desarrollo en America Latina 3:124
CARDs *see* Child-Abuse-Related-Deaths (CARDs)
Care 6:25, 6:27–28
and feminist social work 8:939
gendered *see* Gendered care
health 10:631
health care 10:571–575
Care management, of older people 7:338–339
Care market 9:804
Care Programme Approach (CPA) 15:198
Care/caring work 11:147–149
Career Active System Triad (CAST) model 3:130–131, 3:131t
Career advancement 21:137
Career aspiration
coeducation schooling 3:929
single-sex schooling 3:929
Career criminals 1:312, 14:59–62
Career development and intervention, counseling psychology 5:71
Career goals, engagement and disengagement from 4:817
Career gypsies 18:665
Career making, teachers in United States 24:79–80
Career mobility, event history analysis for 8:308
Career patterns 10:314–315
Career success, and self-esteem 21:524
Career-less structure of teaching 24:79
Careers 3:128–134, 14:80, 25:662
in academic medicine 25:626
in applied geography 1:846–847
artists *see* Artists
of artists, and social networks 5:468–469
cultural production 5:609
factors influencing 3:130–131
future of 3:133
human resource management 3:131–133
individual and organizational perspectives 3:128
labor markets 3:128–132
models 3:129f
models, emerging 3:128–130
boundaryless career 3:129–130
kaleidoscope career 3:130
multidirectional career 3:130
postcorporate career 3:130
Protean career 3:130
outcomes 3:131–132
integration of practices into coherent system 3:132
national level 3:132
practices and systems 3:132
planning and management 3:130–131
-related issues 3:132–133
diversity 3:132–133
global careers 3:133
work–life balance 3:133
scientific disciplinary 21:288
and social mobility 22:426–429
studies 3:128, 3:133
systems, changing nature of 3:128–130
of visual sociology 25:193–194
women in 25:626
Caregiver 3:136
in game playing 18:219
strain 3:135, 3:137
Caregiver burden 3:135–138
aging population 3:135–137
Alzheimer's disease 1:587–589
caregiver strain 3:135, 3:137
definition 3:135
dementia caregiving 3:135–137
elderly caregiving 3:136–137
family caregiving 3:135–137
informal caregiving 3:135–137
mental health 3:135, 3:137
physical health 3:136–137
stress-health model 3:136–137
typical caregiver 3:136
Caregiver–infant synchrony 6:14–15
Caregiver–toddler play
preferences, differences in 18:222
similarities in gender differences and mutual specificity in 18:221
Caregiving 9:730–731
Alzheimer's disease 1:587–588
definition 3:135
future challenges 3:137
and health 10:575
multifaceted consequences 3:136
prevalence of 3:135–136
Carers (Recognition and Services) Act 1995 15:198
Care-seeking 11:596–602
determinants 11:597–599
automatic and deliberative processes in abnormal illness behavior 11:599
factors affecting during interpretation 11:598–599
symptom interpretation 11:597–598
symptoms 11:597
Caretaking, in childhood 3:454
Carey Infant Temperament Scale 22:944–946
Cargo cults 1:748, 15:531
anthropology 3:139–143
belief 3:140–141
birth of 3:139–140
desire for 3:140–142
futures 3:143
label 3:140
and Melanesian culture 3:142–143
organization 3:141–142
prophesies 3:141–142
Caribbean
carnival 3:145–149
education in *see* Latin America and Caribbean (LAC)
fertility transition in 9:68
condom 9:70–71
contraception 9:69–71, 9:70f, 9:71t
fertility trends 9:68–69, 9:68f–69f
first birth, in Latin America and Caribbean 9:69–70, 9:70t
marriage onset, in Latin America and Caribbean 9:70t
proximate determinates 9:69–71
reproductive rights 9:71
sexual debut, in Latin America and Caribbean 9:69–70, 9:70t
unintended pregnancy 9:71
sociocultural overviews of 1:638–643
anthropology of 1:639–640
class 1:641
colonialism 1:638–641
creolization 1:641
cultural reinterpretation 1:639
cultural values 1:640
culture of 1:640–641
family relations 1:642
geography of 1:638
history of 1:638–639
migration 1:638–639, 1:641–642
mobility and transnational relations 1:641–642

Caribbean (*continued*)
plantation societies 1:638–639
race 1:639, 1:642
slavery 1:638–639
social networks 1:641–642
sociocultural integration 1:640
transnational communities 1:642
Caribbean English creoles (CEC) 18:137
Caring
about students and respecting them 24:45
in close relationships 22:566
and empathy 7:542
partner's interest in 22:565–566
Caring conserver 18:666–667
Caring Dads 8:847
Caring for others 19:231
CARITAS (comprehensive, accessible, responsive, individualized, transdisciplinary, accountable, and systemic care) 10:105–106
Carlin, Jerome 13:516–517
works
Law and Society Review 13:516–517
Lawyers' Ethics 13:516–517
Lawyers on Their Own 13:516–517
Carlson, Gabrielle 25:625–626
Carlyle, Thomas 12:759
Carnality 4:267
Carne camp 3:145
Carnegie, Andrew 24:763
Carnegie Corporation 16:166
Carnegie Initiative on the Doctorate 6:610
Carnegie Institution of Washington (CIW) 1:104, 24:763
biomedical sciences 2:665–666
Carnelevare 3:145
Carniol, Ben, works, *Case Critical: The Dilemma of Social Work in Canada* 8:152
Carnival 3:145–149
approaches to 3:148–149
Caribbean 3:145–149
celebration 3:145–147
costume 3:145–148
dance 3:147–148
entertainment 7:670
festival 3:145–149
folklore 3:147–148
identity 3:145, 3:148–149
in Latin America 3:148
masquerade 3:145–149
in New World 3:146–148
northern Europe 3:146
nusic 3:145–148
performance 3:146–148
southern Europe 3:146
Carnival of Binche 3:146
Carnivalesque, liminality 14:135–136
Carolan, M.
works
Cheaponomics: The High Cost of Low Prices 10:80
Reclaiming Food Security 10:80
Caroline Islands 17:158, 17:160–161
Carolingian Renaissance 20:428–429
Caroll model, of school effectiveness 21:83
Carotid intima-media thickness 6:209
Carpenter, Patricia 12:304
Carpooling, transportation 24:623–624, 24:626
Carr, E.H. 11:111
Carr, Harvey 19:413
Carrera Pregnancy Prevention Program 21:825
Carrier, James, works, *Handbook of Economic Anthropology* 7:19
Carrier screening, ethics of 20:702–703
Carrier–Farrag ratio method 6:130–132
Carro camp 3:145
Carroll, John B. 14:752
Carroll's model for school learning 14:752–753
Carr-Saunders and Wilson analysis 19:107
Carrying capacity
abundance and distribution of limited resources, effect of 3:157
of biodiversity 3:157
biophysical 3:158
defined 3:155–156
of ecosystem 3:157
of Environment 3:155–160
human 3:157–158
negative density-dependence 3:155–156
population 3:155–157
relevance in anthropocene 3:159–160
Carryover effects 5:349–353, 12:452
Carson, Rachel
works
The Sea around us 7:747
Silent Spring and Blueprint For Survival 1:485, 3:374, 7:747, 7:781, 7:792, 21:179, 22:873, 23:809
CART *see* Classification and Regression Tree (CART) methodology
cART *see* Combined antiretroviral therapy (cART)
CART algorithm 23:393
Carta Mercatoria of 1303 13:915
Carte-de-visite 18:81–83
Cartels 14:431–432
Carter, J. 18:367–368, 18:461
works, *Palestine: Peace Not Apartheid* 17:430–431
Carter, Prudence, on cultural capital 5:378
Carter Administration, and area and international studies 1:946–947
Carter–McGoldrick model 3:385
Cartesian coordinates 14:284
graphics 10:358–359
Cartesian dualism 3:263, 17:709
Cartesian ego view 17:742
Cartesian linguistics 14:228–229
Chomsky on 3:526–528
Cartesian physics 18:107–108
Cartesian product of the sets 17:296
Cartesian space 21:533
Cartesianism 23:611
early criticisms and revisions of 15:546
recent criticisms 15:546
Cart@gène project 1:760–761
Cartier-Bresson, Henri 25:194–195
Cartographic animation 6:716
Cartographic Data Visualizer (CDV) 3:165
Cartographic film, as dynamic maps 6:716
Cartographic (map) scale 21:1
Cartographic strategies 24:263–266
Cartography 3:161–166, 3:620–621, 10:97, 10:174–176, 13:224–226, 13:229–230
cartographic visualization process 3:162f
categories of maps 3:171
democratization of 3:163
design 3:171–173
elevation data 3:171f
and exploration 3:163–164
exploratory cartographic functions 3:165f
generalization 3:170
geographic thought 10:70
history of maps 3:173–174
Humboldt on 25:279–284
Madagascar physiography 3:172f
and postcolonial geography 18:644–645
projection 3:169–170, 3:169f
resolution 3:168–169
scale 3:168, 3:168f
setting 3:161
spatial relationships 3:170
symbolization and data modeling 3:170
thematic map 3:172f
and visualization 3:162–163
Cartography and Geographic Information Science 10:60
Cartwright, N., on idealization 11:517, 11:520
Carver, Thomas Nixon, works, *Sociology and Social Progress: A Handbook for Students of Sociology* 7:588
CASA *see* Computational auditory scene analysis (CASA)
CASC *see* Clinical Assessment of Skills and Competencies (CASC)
Cascade review 12:863
Cascade screening 9:928
Case, definition 3:208–209
Case, R., staircase model 16:504
Case Con Manifesto 22:803
Case construction 18:631–632
Case law 3:671
value of databases in 5:803–807
critical services 5:805–807
organizational benefits 5:803–805
Case management (CM)
in psychiatric disorders 22:383–384
and social work 24:491–492
Case method 3:176–182
in classroom 3:176–177
aim 3:176–177
effectiveness 3:177
examples 3:178
levels of learning 3:178–179
teaching cases 3:177
definition of 3:177
design 3:177
The Case of the Mayagna (Sumo) Awas Tingni Community v. Nicaragua 11:799

Case processing 5:185–186
Case series 4:91–92
Case studies 3:194–201
collections 3:176–182
definition 3:208–209
differences in orientation 3:195
economic development 5:529–531
logic 3:202–207
programs 3:179
teaching cases 3:177
types of 3:209t
Case study methods 3:208–213
causal mechanisms 3:211
future directions 3:212
heuristic case studies 3:209, 3:211
history of 3:209–211
instances of class of events 3:209
interpretative case studies 3:209
new developments 3:211–212
philosophy of science 3:211
strengths and limits 3:211–212
typological theories and fuzzy logic 3:211
qualitative methods 3:208–212
structured focused comparison 3:208, 3:210–211
within-case method 3:208–210, 3:212
Case study protocol 3:198
Case study research 3:194, 3:200, 3:202
analysis 3:204–205
identifying tipping point 3:204
contemporary 3:196
credibility and generalizability, issues of 3:196–200
defining case in 3:195b
definition 3:202
deviant 3:203–204
early roots in sociology, psychology, and anthropology 3:195–196
exemplary 3:195–196
generalizing from a case 3:200
analytic 3:199
from another single-case study 3:200b
transferability 3:199
working hypothesis 3:199
interpretivist and realist orientations 3:197
interviewee's perspectives 3:199
plausible rival explanations 3:198, 3:198b
prototypical case 3:203
real-world context 3:194–195
reliability 3:198–199
selection of cases 3:202–204
challenges and pitfalls 3:205
in social sciences 3:202
teaching 3:196–197
theoretical generalization and refinement 3:205–206
origins of urban growth 3:205–206
thick description 3:197
trustworthiness 3:197
typical case 3:202–203
validity 3:197–198
working definition and its implications 3:194–195
Case-based reasoning (CBR) process 19:56
as an intelligent systems approach 19:56–57
applications of CBR 19:59–60
active CBR task domains 19:59
case-based 'retrieve and propose' systems 19:59
case-based support systems 19:59
in educational systems 19:59–60
in intelligent systems 19:56–57
role in human cognition for classification and interpretation 19:56
case adaptation 19:58
constraint satisfaction methods 19:58
evaluation and application 19:58
storage 19:58, 19:60
as a cognitive model 19:56
computer models 19:57
conversational case-based reasoning (CCBR) 19:57
efficiency benefits 19:57
interpretive 19:58
issues with 19:58–59
in developing case adaptation rules 19:58–59
harnessing Internet-based knowledge sources 19:59
integration with other methods 19:59
in long-lived 19:59
maintenance 19:59
retrieval and similarity assessment 19:58
learning done by CBR systems 19:60
quality benefits 19:57
relationship between artificial intelligence and 19:60
retrieval algorithms 19:57
rule-based methods 19:58
situation assessment 19:57
steps in 19:57–58
user acceptance 19:57
web-based sources *versus* textual sources 19:59
Case-based social work ethics 22:784
Case-control design, logistic regression 2:589
Case-control sampling *see* Outcome-based sampling
Case-control studies
epidemiology 7:829
evidence-based medicine 8:328–329
Casein 2:848–849
Case-level administrators 20:164
Case-oriented research 3:187–193, 3:190t
causation 3:191–192
of commonalities 3:191
goals of 3:187–188
logic of 3:188
on multiple instances 3:188–190
outcomes 3:191
correlation between a causal and outcome variable 3:192
versus variable-oriented research 3:190–192
bridging 3:192–193
Cases
Becker on 2:405
classification
assigning for classes 3:809
selection 3:806
facts 13:789–790
interpreting legal concepts with 13:786–789
Cash and Counseling Demonstration and Evaluation (CCDE) (Arkansas) program 21:501
Cash crop 17:677
Cash flow analysis 4:929
Cash flow decisions, supply chain management 23:710
Cash for care policies 22:775
Cash for care schemes, for older people 21:498, 21:500
Cash rental, land tenure 1:478
Cash transfers 3:214–222, 3:408–409, 25:479
conditional cash transfers 4:555–559
economic growth 3:217–218
future directions 3:220–221
human capital development 3:215–217
education 3:217
health 3:216–217
nutrition 3:215–216
humanitarian protection 3:218–219
and poor households 3:408–409
poverty 3:214, 3:216–218, 3:220–221
program governance modalities and challenges 3:220
social norm change 3:219–220
vulnerabilities 3:214, 3:216, 3:219–221
Cashinahua of Peru 21:617, 21:619
Cashmore, E., on gay athletes 9:659
Casing, narrative networks 16:214
CASM movement 20:881–882
CASS *see* Chinese Academy of Social Sciences (CASS)
Cassin, René 24:443
Cassiodorus 20:285
CASSW *see* Canadian Association of Schools of Social Work (CASSW)
CAST *see* Cardiac Arrhythmia Suppression Trial (CAST); Career Active System Triad (CAST) model
Castas 23:45
Caste 23:632
-based politics 3:224–226
colonial construction 3:224
contemporary context 3:224–225
Dalits 3:225–226
and hierarchy 3:223–224
in India 3:223
and inequality 11:896–897
local level 3:223–224
in South Asia 2:78–79
Caste society 21:137–138
Caste systems, in India 1:768
Castel, Robert 23:3–4
Castelike minority groups 8:132–133
Castellian binary 12:92–93
Castells, Manuel 3:579–580, 11:234–235, 11:283–284, 12:913–914, 18:202–203, 22:994, 24:866

Castells, Manuel (*continued*)
on collective consumption 3:229
on globalization 3:229–231
on information and communication technologies 3:229–231
on informational society 3:230
on interdisciplinarity 3:229
on Internet 3:230–231
lifestyle of 3:228–229
on power and change 3:230–231
on social movements 3:229–230
as social scientist 3:229
space of flows concept of 3:230
as urban sociologist 3:229–230
works
The Communication Power 16:798
The Information Age: Economy, Society and Culture 12:601–602, 18:668
The Power of Identity 4:176
Castiglione, Baldassar 14:244
Castoriadis, Cornelius 6:27, 19:332–333
Castro, Fidel 23:48
Casual explanations, canonical form for 8:596t
Casual sex 11:192–196, 17:486, 17:488, 17:490, 21:842
consequences of 21:842–843
trends in 11:192
see also Hookups
Casualties
of natural disasters 16:316
war 16:307–308
Casualty support, in disasters 22:757
Casuistic reasoning 13:776–779
Casuistry 2:634–635, 22:784
CASW *see* China Association of Social Workers (CASW)
CASWE *see* China Association of Social Work Education (CASWE)
CAT *see* Computerized adaptive testing (CAT); Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)
Catacomb 13:887
Catalexis 23:715
Catallaxy 6:92
Catalonia, carnival in 3:146
Catalyst, social work education as 22:792
Cataphora 1:674
Catastrophe, and stress 7:805–806
Catastrophic environmental stress 7:806–807
Exxon Valdez oil spill 7:806–807
hybrid forms of environmental stress 7:807
living in contaminated areas 7:806
living with nuclear power plants 7:806
Catastrophic errors 17:370–374
Catastrophic risks 24:118
Catatonia, schizophrenia 21:9, 21:13
CATCH *see* Child and Adolescent Trial for Cardiovascular Health (CATCH)
Catchments 10:113–115
Catechism 2:346
Catecholamine-opioid peptide reinforcement hypothesis 20:212–213
Catecholamines 16:749–750, 20:212
Catecholamines noradrenalin neuroendocrinology 16:651
Catechol-O-methyltransferase (COMT) 8:863–864
Categorial grammar 15:751–752
Categorical classification of disorders 6:388–389
Categorical covariate information 18:605
Categorical data 5:1
Categorical demography, and diversity 6:596
Categorical imperative (CI) 13:5–7
and eudaemonism 8:214
Kantian ethics
extracting and vindicating 13:5–7
living 13:7–8
Categorical inequality 11:909–912
Categorical observed variables, estimators for 23:617
Categorical perception 23:237–239
and language 13:294
Categorical self 21:469
Categorical thinking 11:909
Categorical variables
groups of 16:123
multivariate displays 10:354–355
regression modeling of 16:132–133
Categorical *versus* dimensional approaches 6:346
Categories
analytic, gender as 9:784
linguistics 19:288
degree of membership 19:287
representativity 19:287
structure 19:286–287
storing and using 22:187–188
Categories of maps 3:171
Categories of thought 13:70–71
Categorization 11:551–552, 23:241
by brain 2:800–801
concept learning 4:538–541
and racism 19:866–867
social causes of 6:244–245
and social identity 22:319–321, 22:323
Categorization errors 16:773
Categorization of faces 8:651–652
Category characteristic curve (CCC) 24:250
Category labels, survey questions 23:767–768
Category learning 4:654–655
and representation, neuropsychology of 4:540–541
Category markers, visibility of 22:320
Category object selectivity within ventral visual cortex 16:630–631
Category rating 21:604–605
Catharsis 10:430–431
early studies 19:502–503
Catholic Apostolic Church 15:531
Catholic aristocrats 11:7
Catholic Charismatic Renewal (CCR) movement 1:624–625
Catholic Christianity 3:233
Catholic Churches 3:233–235, 3:683, 13:820, 15:327, 18:100, 20:408, 24:602–603
and globalization 10:231
on homosexuality 13:858–859
and miracles
see also Miracles, cultural logic of
Catholic League 20:102–103
Catholic missionaries 15:616, 24:602–603
Catholic missionization, in South America 23:46
Catholic Reformation 20:98, 20:100–101
Catholic school advantage (CSA) 21:73
Catholic schools 20:302, 21:72–74
Catholic social justice tradition 22:386
Catholic theology 15:682
in era of Reformation and confessionalization 20:100–101
Catholicism 2:891, 3:233–236, 3:540, 8:258, 20:293–294, 24:604
accentuation of western church 3:234
alternatives to 3:538
apostles of 3:233
art cultures 3:234
authority of major churches 3:233–234
canon law and 3:234
Catholic creeds 3:233
as central form of Christianity 3:233
conservative trends 3:538
Conservatives *versus* Progressives 3:236
contraception law 3:235–236
Counter-Reformation 3:234
counter-reformation 3:240
early 3:233
and gender 3:237–242
Gregorian reform 3:234
growth of 3:236
in later Middle Ages 3:234
in Latin Americans' life 3:537
law of celibacy 3:234–235
and liberation theology 14:32–34, 14:36–37
martyrs 3:233
modern 3:240–242
mystical writings, influence of 3:234–235
orders of women 3:234–235
papacy and 3:233–236
post-Constantinian 3:233–234
reformation and 3:234, 3:236
religious orders 3:234, 3:236
Roman Catholics 3:234–235, 3:542
rulers of church 3:233
Second Vatican Council 3:235
shortage of priests and declericalization of the Church's ministry 3:236
sixteenth and seventeenth centuries 3:234–235
in South America 23:46
Tridentine 3:234
use of Latin 3:235
Vatican Council 3:234–235
virginity/celibacy, law of 3:233–235

CATI *see* Computer-assisted telephone interviewing (CATI)
Cato (Roman statesman) 11:37
CATO system of legal reasoning 13:787
Cats, intelligent movements of 23:292–294
Catt, Carrie Chapman 22:444
CATT study 24:282
Cattelino, J.R., on settler colonialism in North America 17:25
Cattell, James McKeen 12:307, 14:811–812, 19:472, 22:699
Cattell, Raymond Bernard 3:243–246
  biography 3:243
  bridging of psychology's past and future 3:245–246
  Child research program 3:243
  contributions of 3:243–246
  factor theory of 3:244
Cattell–Horn–Carroll (CHC) theory 12:250–251, 12:253, 12:285, 12:324–325
Cattle
  domestication 16:840
  raiding and pastoral conflict 17:582
Cattle, complex 17:578–579
Caucasian-American children 24:269–270
Caucasus, sex ratio at birth (SRB) 15:609
Cauchy distribution 6:578
Caucus for a New Political Science 18:469
Caudal intraparietal sulcus (cIPS) 5:26
Caudal premotor areas 18:849
Caudate 23:252
Caudillism 6:372, 6:375–376
Caudillismo 6:368, 20:909–911
Caudillos 20:907–911
Causal analysis 16:405
  confounding variables 3:247–248
  counterfactuals 3:247–248
  empirical social research (ESR) 7:558, 7:563
  field experiment 3:247–248
  and functional analysis 9:536
  hypothesis testing 3:248–249
  mechanism-based 14:873–879
  of mediating variables 15:66–68
  quasi-experiments 3:247–248
  randomization, experimentation, and social science research 3:247–248
  research in non-western societies 3:248–249
  selection bias 3:247
  in social sciences, history of 3:247–250
  statistical analysis 3:248
  structural equation modeling 3:248
Causal analytic epidemiology, of mental illness 15:183–184
Causal attribution
  conflict 4:583
  explanatory style 8:598
  and self-efficacy 21:504–505, 21:508
Causal chains 7:612–613, 14:873–874, 14:876, 14:878
Causal conclusion, use of term 3:255
Causal connections, generalizing 9:841–843
Causal counterfactuals, in social science research 3:251–254
Causal decision theory 5:97
Causal determinism, and responsibility 20:590
Causal effects 10:342, 15:66–67
  instrumental variables 17:111
  notation for 3:255–256
Causal explanations 11:12, 11:15–16
  of scientific realism 20:22
Causal functions 21:344
Causal hierarchy picture 3:261–266
Causal homogeneity 3:192
Causal identification, obesity 17:75–76
Causal indicators 11:751–752
  of SES 11:751
Causal inference 3:252–253, 3:255–260, 5:91–92, 6:868, 12:39, 12:42, 12:454, 17:109–110, 19:397
  for class size 3:751–752
  as comparison of potential outcomes 3:255
  compliance 4:452
  control variable 4:840–841
  definition 3:253
  design considerations 3:258–259
  epidemiology 7:826–827
    Hill criteria 7:827t
  and Neyman–Rubin model 3:252–253
  nonequivalent group designs 16:853
  notation for potential outcomes and causal effects 3:255–256
  observational studies 3:257–259
  overlap 3:257–258
  parametric assumptions 3:259
  principles to generalize 8:620
    causal explanation 8:620
    interpolation and extrapolation 8:620
    making discriminations 8:620
    relevant and irrelevant variation in *UTOS* 8:620
    surface similarity 8:620
  propensity scores 3:257–259
  and random assignment 19:883
  randomized experiments 3:257–258
  regression 3:258–259
  structural assumptions required for nonparametric identification 3:256–258
    common support 3:257–258
    ignorability 3:257
    stable unit treatment value assumption 3:258
    temporality 3:258
    unconditional ignorability 3:256–257
  for time-varying treatments 14:340–344
Causal influence identities 3:244
Causal judgment 2:202–203
Causal mechanisms 14:873–879, 22:415–417, 22:419
  case studies 3:211
  causal chains 14:873–874, 14:876, 14:878
  causation 14:873–874, 14:878
  conceptual core 14:877–878
  critical realism 14:874–875
  Elster on 14:875
  Hempelian view 14:874
  logic of inquiry 14:873–874
  meanings of mechanisms 14:874–876
  mechanisms and explanations in social sciences 14:873–874
  mechanistic explanation 14:873–874
  Merton and analytic sociology 14:875–876
  peripheral debates 14:877–878
    micro-foundations 14:877
    objects of inquiry 14:877
    observability 14:877
  in quantitative methodology 14:876
    Tilly's view 14:876
Causal mediating relations 15:64
Causal mediation analysis 15:75
  definitions 15:75–76
  goal 15:75
  linear counterfactual models (LCM) 15:77–78
  linear structural equation models (LSEMs) 15:76–77
  marginal structural models 15:78
  principal stratification 15:78–80
Causal modeling 8:687, 13:100, 22:548–550
Causal ordering
  of academic self-concept and achievement 1:56–57
    educational implications 1:57
    importance of multidimensionality and positive psychology 1:57
    self-enhancement and skill development models 1:56–57
  of self-concept 21:462–463
Causal pathways, Wright's contribution to 25:759
Causal reasoning 21:268–269, 21:271
Causal relatedness, and inferences 12:41–43
Causal relations, instrumental variables 12:205, 12:209
Causal relationships 21:268–269
  between the treatments 12:450
Causal responsibility 20:587
Causal salience 4:770
Causal statements 3:191
Causal step method 15:65–66
Causal theory of reference 20:69
Causal variables 13:394–397
Causality 9:129, 16:194, 16:196, 19:736–737, 23:592
  between culture and economic development 5:533
  and determinism 6:241–242
  Heider on 10:750–754
  ideological limitations on 23:593–594
  and interstate war 25:423–426
    causal-explanatory approaches in IR 25:424–426
    comprehensive accounts 25:426
    explanatory strategies 25:423–424
  and legal reasoning 13:776–777, 13:780, 13:782
  partial 23:425
  Popper on 18:524
  and social mechanisms 22:417
  social systems, Luhmann 14:383
  statistical analysis of 3:248

Causality (*continued*)
and statistical distributions and correlation 23:425–426
Causality analysis 16:196
Causality orientations 21:488–489
Causal/mechanical approach to scientific explanation 21:302–303
in cognitive neurosciences 21:303
compatibility of 21:303–306
Causation 3:261–266, 14:873–874, 14:878, 15:852–854
Cartesian dualism 3:261
causal hierarchy picture 3:261–262
and critical realism 5:277–279
Dowe's account 3:270
exclusion argument 3:262–263
Gassendi's objection 3:261
Lewis's counterfactual analysis 3:269–270
and mechanisms 22:417
'mechanistic' accounts of 3:270
in medicine 15:88–90
notion of 3:261
political power as 18:760–761
population-level 3:270–271
possible responses 3:263–264
rejecting the causal part of picture 3:263–264
rejecting the hierarchy part of picture 3:263
rejecting the metaphysical presuppositions of picture 3:264
problems with exclusion 3:264
reverse causality 15:851–852
scientific evidence and 21:297
self-efficacy 21:510
Causativization 25:3–4
Cause lawyers 18:390–391
and politics 18:390
Cause of death statistics 10:727
Cause-deleted life table 14:92
Cause–effect relationship 8:617
Cause-related marketing (CRM) programs 4:747
Causes 3:266–273
Causes and laws 3:272–273
causation 3:269–272
counterfactual analyses 3:269–270
naïve regularity theory 3:268
probabilistic theories 3:269–270
determinism *versus* indeterminism 3:266
laws of nature 3:267–268
Ceteris Paribus Laws 3:267–268
laws and dispositions 3:268
laws *versus* accidents 3:267
realist *versus* regularity accounts 3:267
stability and invariance 3:268
particular and general 3:266
preliminaries 3:266
process and mechanistic theories 3:270
Ramsey–Lewis view 3:267
Causes of death 15:857, 15:860–861
Causes of diseases 8:449
Cavalcanti, Alberto, works, *Pett and Pott* 6:621
Cavalry, warfare 25:415, 25:417
Cavan, Ruth, works, *Suicide* 7:561
Caxton, William 18:915
Cayenne (French Guiana) 1:638
Caylus, Comte de 3:761–762
Cayton, H.R., on race in North America 17:23
CBA *see* Cost–benefit analysis (CBA)
CBAR *see* Community-based action research (CBAR)
CBD *see* Central business district (CBD); Convention on Biological Diversity (CBD)
CBDRR *see* Community-based disaster risk reduction (CBDRR)
CBE *see* Council for Basic Education (CBE)
CBI *see* Combined Behavioral Intervention (CBI)
CBIM *see* Coaching Boys into Men (CBIM) program
CBITS *see* Cognitive Behavioral Intervention for Trauma in Schools (CBITS)
CBO *see* Congressional Budget Office (CBO)
CBPR *see* Community-based participatory research (CBPR)
CBR *see* Case-based reasoning (CBR) process; Crude birth rate (CBR)
CBS *see* Charles Bonnet syndrome (CBS); Columbia Broadcasting System (CBS)
CBS (Columbia Broadcasting Service) 14:693
CBS Evening News with Walter Cronkite 1:526
CBSAs *see* Core based statistical areas (CBSAs)
CBSM *see* Cognitive behavioral stress management (CBSM)
CBT *see* Cognitive behavioral therapy (CBT)
CC *see* Conscious control (CC)
C–C chemokine receptor type 5 (CCR5) 1:489–492, 1:489f
CCA *see* Canonical correspondence analysis (CCA); Council for Cultural Affairs (CCA)
CCAM *see* Coordinated community action models (CCAM)
CCAS *see* Committee of Concerned Asian Scholars (CCAS)
CCBE *see* Council of Bars and Law Societies of Europe (CCBE)
CCBR *see* Conversational case-based reasoning (CCBR)
CCBT *see* Computerized cognitive behavioral therapy (CCBT)
CC-BY *see* Creative Commons Attribution license (CC-BY)
CCC *see* Category characteristic curve (CCC)
CCCS *see* Chicago School and the Birmingham Centre for Contemporary Cultural Studies (CCCS)
CCDA *see* Confirmatory combinatorial data analysis (CCDA)
CCDE *see* Cash and Counseling Demonstration and Evaluation (CCDE) (Arkansas) program
CCF *see* Cohort-completed fertility (CCF)
CCG *see* Combinatory Categorical Grammar (CCG)
CCI *see* College Characteristics Index (CCI)
CCK *see* Cholecystokinin (CCK)
CCK-4 *see* Cholecystokinin tetrapeptide (CCK-4)
CCKAR 2:810
CCMD *see* Chinese Classification of Mental Disorders (CCMD)
C-command 3:919
CCP *see* Center for Community Progress (CCP); Chinese Communist Party (CCP); Cross-cultural psychology (CCP)
CCR *see* Catholic Charismatic Renewal (CCR) movement
CCR5 *see* C–C chemokine receptor type 5 (CCR5)
CCREM *see* Cross-classified random effects model (CCREM)
CCs *see* Constitutional courts (CCs)
CCSS *see* Childhood Cancer Survivor Study (CCSS)
CCT *see* Columbia Card Task (CCT); Conditional cash transfer (CCT); Consumer culture (CCT); Cross-cultural trading (CCT); Cross-cultural training (CCT)
CCTV 23:729
CCTV evidence 18:230
CD *see* Communicative dynamism (CD); Conduct disorder (CD)
CD4
and AIDS 1:488–492, 1:489f
CDA *see* Combinatorial data analysis (CDA); Critical discourse analysis (CDA)
CDBG *see* Community Development Block Grant (CDBG) program
CDC *see* Centers for Disease Control and Prevention (CDC)
CDCs *see* Community development corporations (CDCs)
CDE *see* Controlled direct effect (CDE)
CDF *see* Cumulative distribution function (CDF)
cDNA *see* Complementary DNA (cDNA)
CDPs *see* Census designated places (CDPs); Community development projects (CDPs)
CDR *see* Crude death rate (CDR)
CDS *see* Child Development Supplement (CDS); Child-directed speech (CDS); Credit default swap (CDS)
CDSs *see* City Development Strategies (CDSs)
CDV *see* Cartographic Data Visualizer (CDV)
CE *see* Consumer expenditure (CE) survey
CEA *see* Cost-effectiveness analysis (CEA)
CEACs *see* Cost-effectiveness acceptability curves (CEACs)
CEAFRD *see* Convention on Elimination of All Forms of Racial Discrimination (CEAFRD)
Ceausescu, Nicolae 18:700
CEB *see* Children ever born (CEB)

Ceballos-Lascuarian 7:105–106
CEBRAP *see Centro Brasileiro de análise e planejamento* (CEBRAP)
CEC *see* Caribbean English creoles (CEC); Council for Exceptional Children (CEC)
CED *see* Condition on Extraction Domains (CED)
CEDAW *see* Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW)
CEE *see* Central and Eastern Europe (CEE); Council for Economic Education (CEE) (United States)
CEER *see* Centers of Excellence in ELSI Research (CEER) Program
Ceja 23:43
Celebrations
  carnival 3:145–147
  of death 5:870
  and media events 15:48–52
  sport and 23:302–303
  of transition to adulthood 24:579–580
Celebrities 3:274–278, 20:484, 20:489
  charisma 3:275–276
  consumer culture 3:274–275
  and critiques of mass culture 3:274–275
    capitalist ideology 3:274–275
    versus heroism 3:275
  elites 3:274
  fandom and reception of 3:276–277
  future of 3:277
  media, uses of 3:277
  new media 3:274, 3:277
  ordinariness 3:275–276
  political economy of 3:276
    Hollywood star system and beyond 3:276
    production in twenty-first century 3:276
  role 3:275
  status 3:274–275
  as symbolic entities 3:275–276
    as democratic aristocracy 3:275
    search for 'Real' self 3:275–276
  television 3:275–277
Celebrity scientists 9:991–992
Celerity 6:255–256
Celestial mechanics 13:382
Celestial navigation system 16:346–347
Celibacy 3:237
Celibate Christian monastic life 3:238
Celine, Louis-Ferdinand 1:540
Cell adhesion molecules (CAMs) 12:711–712
Cell death, Ca^{2+} channels-mediated 12:714
Cell differentiation 20:57
  in brain development 2:823
Cell migration 20:57
  in brain development 2:821–824
Cell types, specific, optogenetic methods for targeting 17:269–270
Cell-based therapies 20:59
Cell-specific epigenetic effects 24:687
Cellular automata 1:368, 21:530
  agent-based models 1:367–368
Cellular effects tDCS 24:534
Cellular mechanisms
  sharing by memory 23:876–877
  for VOR motor learning 25:73
Cellular operant conditioning 20:214–215
Cellular pathology of Rudolf Virchow 15:119–120
CELS *see* Conference on Empirical Legal Studies (CELS)
Celtic Europe, state organization in 8:271–272
Celtic fringe 4:223
Celtic language 11:824
Celtic monasticism 15:725
Cemeteries 5:872–873
CEnR *see* community engaged research (CEnR)
Censal-ratio method 6:164
Censored data, survival analysis 23:774
Censorial jurisprudence 13:485
Censoring 14:54–55, 23:784–785
  in archiving 1:919
  compliance 4:450–451
  cultural mediators' role 5:442
  in event history analysis 8:306–307
  family reconstitution 6:141
Censorship
  administrative *versus* judicial censorship 9:393
  for Chinese cultural products 5:600
  and ignorance 11:579
  international law on prior censorship 9:393
  in research and scholarship 3:279–283
    biology 3:280, 3:282
    bioterrorism 3:280, 3:282
    Butler Act 1925 3:280
    Copernicus 3:279
    dual-use dilemma 3:280–281
    ethics of censoring dual-use research 3:281–282
    freedom of speech 3:281–282
    Galileo 3:279–280
    history, methods, and motivations 3:279–280
    human rights 3:281–282
    nuclear science 3:282
    religious censorship 3:279
    scientific research 3:282–283
    tobacco research 3:280, 3:282
  self-censorship 9:394
Censorship regimes 16:779
Censure, punishment 19:632–633
Census 3:290–301, 6:140, 6:163, 10:303, 11:6
  adjustment 3:306
  administrative record 3:306
  apportionment 3:302, 3:304–306
  biblical 20:875
  combined census 3:298
  in comparative contexts 3:294–295
  confidentiality 3:302, 3:304
  content 3:291–292
    core topics 3:291–292
    noncore topics 3:292
  content errors 3:293–294
  continuous measurement 3:305–306
  coverage errors 3:293–294
  data adjustments 6:126–135, 6:129f, 6:131f–133f
  data dissemination 3:294
  data processing 3:294
  definition 3:296, 3:302
  demography 6:176, 6:179
  economic statistics 5:761
  electoral district 3:303, 3:305
  features of national census 3:290–291
    geographical specificity 3:290
    household data 3:291
    responsible bodies 3:290
    simultaneity and periodicity 3:291
  functions and techniques 3:303–304
  future innovation 3:306
  future prospects 3:295
  government 20:875
  harmonization 3:294–295
  history 3:302–303
  household 3:302–305
  international census 3:290–295
  Internet 3:304
  inverse projection 6:149, 6:151–152
  machine tabulation 3:304–305
  methodologies 3:293–294
    accuracy assessment of data 3:293–294
    enumeration 3:293
  methodology 3:298
    in Austria 3:300–301
    in Canada 3:299
    in Estonia 3:300
    in France 3:301
    in Israel 3:299
    in Netherlands 3:300
    in Poland 3:300
    selected national census methodological practices 3:299–301
    in Turkey 3:299–300
    in United States 3:299
  microdata 3:306
  move to sample surveys 20:876
  National Statistics Office 3:290
  national traditions 3:304–306
    Canada 3:305
    France 3:305–306
    Great Britain 3:305
    issues 3:306
    United States 3:304–305
  population composition in North America 18:553–557
  population register 3:291–293
  postenumeration survey 3:306
  quantification 19:703–704
  recent shifts from traditional censuses to alternative models 3:292–293
  register-based 3:297
  rolling census 3:292, 3:298
  rolling sample survey 3:305
  for small areas 6:162–165
  and statistical surveys 23:424
  traditional census 3:296–297
  undercount 3:306
  urban poverty 24:915

Census Act 1920 3:305
Census Bureau, US 3:303–305
  Current Population Survey (CPS) 24:384
Census Bureau Data, US 5:695
Census data 5:735–736, 5:817–818, 5:820–822
  for demographic research 5:736
Census designated places (CDPs) 24:856
Census enumeration 6:171, 23:384–385
Census Method II X11 variant 24:351
Census microdata 3:284–289
  advantages 3:284
  big data 3:285f
  definition 3:284
  from genealogical sources 3:287–288
  Integrated Public Use Microdata Series (IPUMS) 3:286–288, 3:287f
  international 3:286–287
  limitations 3:285
  new applications of 3:288
  North Atlantic Population Project (NAPP) 3:287
  origins and development of US census microdata 3:285–286
    1990s 3:286
    1960s and 1970s 3:285–286
  population data 3:285–288
  Public Use Microdata Files (PUMF) 3:287
  restricted 3:288
  Samples of Anonymized Records (SARs) 3:287
Census Office 3:304–305
Census Research Data Centers (RDCs) 3:288
Census Transportation Planning Products 15:637
Census wars 8:92–93
Census X-12 seasonal adjustment filter 24:334, 24:334f
Censuses 1:156
  techniques of 3:303–304
Centenarians 17:175, 17:178–180
  oldest-old mortality 15:864–866
  and population aging 18:544
Center and periphery 20:746
  Cold War 4:123
  diffusion 3:309–310
  and ecumene 7:110–111
  end of 3:311
  levels and varieties 3:308–309
  mixed feelings 3:310–311
  Shils on 21:886–887
  soft power 3:310
Center for Alcohol Studies 1:501
Center for Community Progress (CCP) 21:917
Center for Education Policy (CEP) 1:232
Center for Global Education (CGE) 12:566
Center for Intercultural Documentation (CIDOC) 11:584
Center for International Development and Conflict Management (CIDCM) 5:836–837
Center for Law and Social Policy (CLASP) 19:556
Center for the Study of Higher Education (CESU) 7:183
Center for the Study of Law and Society 13:516
Center of a social system/domain 13:96
Center of mass (CoM) 3:779–780
Center of Systemic Peace (CSP) 5:836
Center on Media and Child Health (CMCH) 7:252
Center on Wrongful Convictions (CWC) 25:787
Centers for Disease Control and Prevention (CDC) 4:253, 5:723, 8:34–35, 22:658
Centers for Independent Living (CILs) 6:449, 6:462
Centers for Medicare and Medicaid Service (CMS) 20:526
Centers in modern world system 16:241–242
Centers of Excellence in ELSI Research (CEER) Program 8:44–45
Centers of pilgrimage 3:309
CENTO *see* Central Treaty Organization (CENTO)
Central Academy of Fine Arts (CAFA) 5:599
Central Africa
  sociocultural aspects 3:312–315
    challenges from Atlantic and colonizing Europe 3:313–314
    Congo 3:312–314
    endogenization 3:314–315
    patrilocal households 3:312
    Savannah 3:312
Central agraphias 6:735
Central alexias 6:732
Central America
  origins of food production in 9:303
  social science infrastructure 22:621–630
  sociocultural aspects 1:621–625
    Belize 1:621, 1:623–625
    Cold War legacies 1:621–622
    Costa Rica 1:621–625
    diaspora 1:622
    El Salvador 1:621–624
    gender politics 1:625
    Globalization 1:621, 1:623
    Guatemala 1:621–625
    Honduras 1:621–625
    Maya 1:621–622, 1:624
    Migration 1:622–623
    multiculturalism, religious change, and new forms of citizenship 1:624–625
    neoliberalism and its opponents 1:623–624
    Nicaragua 1:621–625
    Panama 1:621–625
    from revolution to remittances 1:622–623
    Roman Catholicism 1:624–625
  trafficking of women and children in 24:513–514
Central and Eastern Europe (CEE) 22:668
Central and Northern Europe 11:861
Central Asia
  anthropology of 3:316–322
    definitions 3:316–317
    environment 3:316–317, 3:319–320
    ethnicity 3:316–320
    hybridity, social life 3:320
    institutional framing for areas and topics 3:317
    Islam 3:317, 3:319–320
    migration 3:316–320
    nation-state 3:316–320
    non-Western accounts 3:317
    performativity, political 3:317
    performing politics 3:318–319
    postcolonialism 3:316–317
    postsocialist states 3:317–318
    property regime change 3:317–318
    religion 3:319–320
    Soviet Union, former 3:317–318
    spectacle, political 3:317
    statelessness 3:316–317
    tradition, social life 3:320
Central banks
  and financial crisis 9:183–184, 9:188–189
  monetary policy 15:727–734
Central business district (CBD) 11:236
Central coherence weaknesses 18:790
Central conceptual structures 16:504
Central dysgraphia 1:453–454
Central Europe 6:817
  child/infant mortality 12:15–20
  community development in 4:382
  conservative welfare states 25:481
  conservative welfare states in 25:481
  health crisis in 1:161–163, 1:162f
  holocene prehistory of 5:398–399
  postsocialist cities 3:605–607
Central explanatory problems, formulation of 1:563
Central Identification Laboratory (CIL) 9:335–336
Central language processes 6:732, 6:735
Central level
  of working memory 21:902–903
    WMG 21:904–906
Central limit theorem 14:631–632, 15:267, 19:36
  normal approximation 6:571–572
Central linguistic unit 15:827
Central nervous system (CNS) 7:606, 16:714
  and AIDS 1:488, 1:492
  asymmetry 2:125–127, 2:127f
  axis formation and positioning of 2:116–118
  development, anterior-posterior partitioning 2:118
  effects of alcohol on 1:508
Central nucleus (CeA) of the amygdala 8:870
Central order statistics 17:293
Central place theory 14:287–288, 20:149, 24:945–946
Central planning 18:188–189, 20:844
  economic transformation 7:74–75
  for environment in Soviet Union 20:831
  postsocialist cities 3:606–607
  Soviet politics 20:844

Central pontine myelinolysis, and alcohol 1:512
Central Product Classification (CPC) 5:760–761
Central reasoning mechanism, in semantic network 13:102
Central route processing 15:31
Central sensitization, phantom limbs 18:3–4
Central tendency, in spatial patterns
  lines 23:179
  points and areas 23:178–179
  spatial distribution 23:179–180
Central Treaty Organization (CENTO) 1:557
Central–eastern Europe 8:275
Centralities 7:636–637, 16:546, 22:458, 22:462
  network 16:532–539
  new forms of 3:589–590
Centrality of spirit
  of indigenous peoples 11:804
Centralization 16:535–536, 21:412, 22:265, 22:267
  and ergonomics 11:302
  in structural contingency theory 23:609–611
Centralized and decentralized office, and prosecution 19:211–212
Centralized decision-making mechanisms 6:602
Centralized nation-state, Japan 15:128–129
Centralized statism 4:689
Centralized systems 10:304–305
Centralized tracking 14:355–356
Centre de Sociologie de l'Innovation (CSI) 1:80
Centre for Applied Nonviolent Action and Strategies 22:510–511
Centre for Comparative Genocide Studies 10:7
Centre for Research on the Epidemiology of Disasters (CRED) 6:479
Centre National de la Recherche Scientifique (CNRS) 2:357–358, 22:669–670
Centrelink social workers 7:427
Centric systematic sampling 23:187f
Centripetalism, and conflict/consensus 4:591–592
Centro Brasileiro de análise e planejamento (CEBRAP) 3:122
Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS) 1:627
Centro Studi sulle Nuove Religioni *see* CESNUR (Centro Studi sulle Nuove Religioni)
Centurion Ministries of Princeton 25:786–787
Century dictionaries 24:306–307
Century of the gene 11:311–313
CEO *see* Chief executive officer (CEO)
CEOAEs *see* Click-evoked otoacoustic emissions (CEOAEs)
CEOs *see* Chief Executive Officers (CEOs)
CEP *see* Center for Education Policy (CEP)
CEPAL *see* Comisión Económica para América Latina (CEPAL)
CEPH *see* Fondation Jean Dausset Centre pour l'Etude du Polymorphisme Humaine (CEPH)
Cephas (Peter) 3:546
Ceques 23:39
CER *see* Comparative effectiveness research (CER)
Ceramics
  in Amazon 23:40
  in Andes 23:38
  archaeometry 1:907
  usage in ancient Egypt 1:694–695
Ceramics in archaeology 3:323–326
  and chronology 3:325
  classification of 3:324–325
  compositional studies 3:325
  ethnoarchaeology and 3:325–326
  folk art 3:324–325
  history of 3:323–324
  paste 3:325–326
  pottery 3:323–325
  potter's wheel 3:323–324
  style 3:323–324
  technology 3:323–324
  temper 3:325
CERCLA *see* US Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA)
CERD *see* Committee on the Elimination of Racial Discrimination (CERD)
Cerebellar atrophy, and alcohol 1:511
Cerebellar cognitive affective syndrome 3:328
Cerebellar damage 3:329
Cerebellar flocculus 25:71
Cerebellar neuronal processing 3:330
Cerebellar sequencing 3:330
Cerebellum 16:587, 23:252
  cognitive functions 3:327–331
  eyeblink conditioning
    feedback to CS and US pathways 8:638–639
    learning mechanisms 8:637
  and nonconditioned learning 2:96–97
  role in cognitive associative learning 2:97
  role in emotional associative learning 2:97–98
  role in motor associative learning 2:93–96
    eyeblink conditioning in animals 2:93–95, 2:93f–94f
    eyeblink conditioning in patients 2:95–96
  vestibular 25:63, 25:65–67
Cerebellum cognitive functions 3:328–329
  anatomophysiological cerebro-cerebellar interactions, principles of 3:327–328
  history 3:327
Cerebellum-related vWM impairments 3:328
Cerebral asymmetries
  brain, evolution of 2:831–833
Cerebral cortex 3:332–343, 16:737, 18:46–47, 3:332f
  brain, evolution of 2:827–828, 2:830–831
  cortical lobes and hemispheres 3:334–338
  cytoarchitecture of neocortex 3:340f
  development 3:335f
  differential lamination 3:338–340
    allocortex 3:339–340
    isocortex 3:338–339
  differential lamination of neocortex 3:341f
  frontal lobe 3:334, 3:336
  gyrus 3:336–337, 3:339–340
  human brain
    cholinergic projections of basal forebrain in 3:336f
    evolution in human phylogeny 3:336t
  and human language evolution 11:323
  interactions with thalamus 3:340–343, 3:342f
  interconnectivity and lateralization 3:334–338
  lobes 3:333f
  memory 3:336, 3:338
  microscopic anatomy of human hippocampal formation 3:342f
  neuronal density 3:335f
  neuronal migration during ontogeny of human neocortex 3:339f
  ontogeny and phylogeny of 3:333–334
  plasticity 3:338
  postcentral gyrus 3:336
  precentral gyrus 3:336–338
  projection of sensory pathways 3:337f
  pyramidal cell 3:339–340
  pyramidal tract 3:337–338
  regions 5:24f
  sulcus 3:333, 3:336
  synaptic plasticity 3:338
  thalamo-cortical and cortico-thalamo-cortical projections 3:343f
Cerebral cortical networks 13:681
Cerebral crowding 20:58–59
Cerebral dominance, substrates underlying 2:126–128
Cerebral dominance theory 16:686
Cerebral glucose metabolism 14:443
Cerebral hemispheres 23:293f
Cerebral injury, and posttraumatic amnesia 18:636
Cerebral lateralization, and bilingual aphasia 2:575–576
Cerebral nuclei 3:332f
Cerebral specialization 2:806–807
Cerebral structures involved in neuromarketing 16:701
Cerebral vascular diseases, and alcohol 1:512
Cerebrovascular accident (CVA), and language disorders 13:329–330
Cerebrovascular brain disease, electroencephalography for 7:362
Cerebrovascular disease 18:791–792
  religion and 20:266–267
Cerebrum 5:22–23
Ceremonial exchange 18:714

Ceremonial exchange debt 13:142
Ceremonial television 15:48, 15:51–52
Ceremonies 11:762–763
of indigenous peoples 11:805
media events 15:48–54
of naming 6:841–842
and transition to adulthood 24:577–581
Cérémonies et coutumes religieuses de tous les peuples du monde 7:662
CERN *see* European Organization for Nuclear Research (CERN)
Certainty 6:255–256
of legal punishment 6:256–257
Certainty, relativism 20:234
Certainty principle 6:260–261
Certificate for School Business Administrators (CSBA) 21:68
Certificate of School Business Management (CSBM) 21:67
Certificates of Confidentiality 18:943
Certiorari system 4:700
Cerulo, Karen A., on cognition 7:416
CES *see* Canadian Election Study (CES); Classroom Environment Scale (CES)
Césaire, Aimé, works, *Discours sur le colonialisme* 16:269
CESNUR (Centro Studi sulle Nuove Religioni) 16:808
CESP *see* Confederation of European Specialists in Paediatrics (CESP)
CESSDA *see* Consortium of European Social Science Data Archives (CESSDA)
CESU *see* Center for the Study of Higher Education (CESU)
CET *see* Cognitive evaluation theory (CET)
Ceteris paribus 3:268, 4:865, 25:335
apportionment 7:344
laws 3:267–268
Cetina, Karin Knorr 7:873–875, 13:205–206, 13:208
CEWG *see* Consultative Expert Working Group on Research and Development (CEWG)
CEWs *see* Conducted electronic weapons (CEWs)
CFA *see* Confirmatory factor analysis (CFA)
CFCs *see* Chlorofluorocarbon gases (CFCs)
CFCSA *see* Child Family and Community Service Act (CFCSA)
CFI *see* Comparative fit indices (CFI)
CFMT *see* Cambridge Face Memory Test (CFMT)
CFOs *see* Chief financial officers (CFOs)
CFPT *see* Cambridge Face Perception Task (CFPT)
CFR *see* Code of Federal Regulations (CFR)
CFS *see* Chronic fatigue syndrome (CFS); Current fact situation (CFS)
CFU *see* Community Forestry Unit (CFU)
CG *see* Complicated grief (CG)
CGD *see* Complicated grief disorder (CGD)
CGE *see* Center for Global Education (CGE); Computable general equilibrium (CGE) analysis
CGF *see* Continuous-gene-flow (CGF) model
CGIs *see* CpG islands (CGIs)
CGLIMs *see* Constrained generalized linear models (CGLIMs) approach
Cha, Meeyoung, works, *The Million Follower Fallacy* 22:423
Chadwick, Tabor 25:786
Chaebol 6:789
Chagas disease 18:177
Chagnon, Napoleon
on humor 1:735, 11:417
works
Human Social Behavior: An Anthropological Perspective 22:903
Chaikovskii, Petr 21:741–742
Chailley, Joseph
works, *Nouveau Dictionnaire d'Economie Politique* 7:585
Chain diffusion 11:622
Chain graph 10:341–342
Chain migration 12:438–439, 20:541
Chain of risk, life course epidemiology 14:71
Chained analysis, of geospatial data 10:101
Chains of opportunity 23:471
Chair holders 22:592
Chairs and councils 22:622
Chakrabarthy, Dipesh 18:648
Chalcolithic, South Asia 23:51–52
Chalk, F., works, *The History and Sociology of Genocide* 10:7
Challenge procedures 25:144–145
Challenge studies of anxiety 1:809
Challenger accident 24:137–138
and flashbulb memory 9:255–258
Challenging objects, and scientific revolution 21:320
Chalmers, Iain 8:327
Chalmers, Thomas C. 8:331
works, *Clinical Epidemiology* 8:327
Chamberlain, Neville 1:826–830
appeasement 1:826–830
Chambers, R., on global development, 10:153
Chambers, Whittaker, works, *Witness* 4:690
Chambless, Edgar 24:863
Chameleon socialization 15:82
Championship performance 21:463
Chan, Cheris Shun-ching, on life insurance 5:538
Chan, T.W.
on cultural capital 5:451
on occupation 22:333
Chance
objective probability 19:41
and probabilistic thinking 19:16–18, 19:21
probability 19:23–24, 19:34
statistics as taming 23:424
theory of 13:382–384
Chandler, Alfred D. 3:23
Chandler's analysis 23:522
Chang-an city 24:870
Changchun, Li 5:599
Change
and acculturation 1:69
Castells on 3:230–231
and division of labor 13:159–160
globalization and international division of labor 13:160
postindustrialist societies 13:159
work organization 13:159–160
institutional 12:184
Islamic social work 12:756
and marriage 14:601–602
and nostalgia 17:34–35
personality, measurement of 17:864–865
and public bureaucracies 19:531–532
stages of 19:511–512
Change, organizational
Charity Organization Society 22:803–804
concept of modernization 22:804
impact upon practitioners 22:803–804
poverty and deprivation 22:804
Settlement Movement 22:802–804
social inequality 22:804
social work in Britain 22:805–806
Change agent, organizational development 6:297–298
Change blindness 3:344–349, 24:15–16
consciousness 3:344, 3:348
eye movements 3:344–345
flicker paradigm 3:345–346, 3:345f
history and methods 3:344–346
masking 3:345–346
memory 3:344–348
mudsplashes paradigm 3:345–346
perception 3:344, 3:347
related work 3:347–348
theories 3:346–347
Change management, personalized medicine 17:929–930
Change map 6:715–716
Change numbness 24:15
Change processes 21:136–137
Change-directed improvement 16:551
Changes in business 17:851
Change-score analysis, nonequivalent group designs 16:852
Changing context (Frontiers in history) 9:464
expansion of scholarship in non-European areas 9:464
global geopolitics 9:464
spatial turn 9:464
Changing criterion design 21:986
Changing nature of science 17:13
'commercialization' of science 17:13
'public' and 'private' science 17:13
Channel, information theory 12:106–108, 12:106f
Channel decoding paradigm 23:261–262
Chantal, Jane Frances de 3:234–235
Chaos of Disciplines 11:29
Chaos theory 21:530
Chapel Hill study 1:351
Chappe, Claude 12:491
Character
ethics of 8:63
national character 16:233–236
skepticism about 25:151
Character contests 16:220
and homicide 11:166

Character-based social work ethics 22:784
Characteristics of textbooks 24:258
Characterization problem 5:901–902
Characterization studies, trade and exchange 24:481, 24:483
Characterologica and Kohut 19:342
Charcot, Jean Martin 2:596, 11:472
  hypnotism studies 19:502
Charge, security interests 21:402–403
Charisma 18:765
  artistic charisma 3:355
  authority 3:350–351
  authority and 2:271–272
  of celebrity 3:275–276
  charismatic movements 3:351–352
  Church 3:350–351
  Eisenstadt on 7:321–322
  as erotic capital 8:1, 8:4
  ethical prophecy 3:351
  evil leaders 3:356
  forms of religious association 3:351
  in information age 3:352
  leadership as 13:649
  legal-rationality 3:350–351
  origins of 3:353–354
  political charisma 3:355–356
  power 3:350
  and power 18:765
  recent notion of 3:356–357
  religious charisma 3:355–356
  revitalization movements 3:351
  routinization 3:350–351
  shamanism 3:350
  Shils on 21:885–887
  sociology of 3:350–351
  tradition 3:350–352
  transference and institutions 3:355
  use of term 3:353
  Weber's concept of 3:354–355, 25:458–461
Charismatic authority 25:464
Charismatic center 16:291
Charismatic ideology 5:441
Charismatic leadership 13:666, 25:463–464
Charismatic organizations 2:913–914
Charismatic phenomena 20:357
Charismatic religion 20:356
Charity 22:576–580, 25:475
  economics of 3:358–363
    economist's view 3:358–359
  and nongovernmental organizations 16:858–859
Charity Organization Societies (COS) 6:871, 19:478, 22:801–802, 24:728–729, 25:569
Charles Bonnet syndrome (CBS) 10:499
Charles I, King
  English Revolution 7:649–650, 7:653–654
Charm, as erotic capital 8:1
Charmed circle 19:755
Charreada 9:785
Charter, Leipzig 24:792–793
Charter 77 13:441–442
Charter of European Cities 23:848
Charter schools 7:154–155, 7:300, 21:71, 21:74–75
Charters, health 10:608–609
Chartier, Roger 18:918
CHAT *see* Cultural Historical Activity Theory (CHAT)
Chat rooms, terrorist 24:229–230
Chateaubriand, François-René de 4:686
Châtelet, Émilie du 18:101–102
Chattels, possession of 19:191
Chaucer, Geoffrey
  on personality 17:909
  works
    The Canterbury Tales 17:909
Chauvinism 16:312
  and anthropology 1:752
Chávez, Hugo 23:48
Chayanov, A.V. 1:459
  on peasant economy 1:457
  works
    The Theory of Peasant Economy 1:456–457
CHC *see* Cattell–Horn–Carroll (CHC) theory; Chicago Hellfire Club (CHC)
CHD *see* Coronary heart disease (CHD)
Che Guevara 20:621
Cheap labor 3:665
Chebel, Malek 21:752
  works
    Histoire de la circoncision: Des origines à nos jours 21:752
    La féminisation du monde: Essai sur "Les Mille et Une Nuits&rdquo 21:752
    Le corps dans la tradition au Maghreb 21:752
    Le livre des séductions 21:752
    L'encyclopédie de l'amour en Islam: Erotisme, beauté et sexualité dans le monde arabe, en Perse et en Turquie 21:752
    L'esprit de sérail: Perversions et marginalités sexuelles au Maghreb 21:751–752
Chechen struggle 3:717
Checking value 9:392–393
Checks and balances, republicanism 20:479
Cheka (secret police) 20:857–858
Chelkowski, P., works, *Staging a Revolution* 16:383
Chelsea 5:599–600
Chemical affected dating methods 3:567
Chemical energy 7:626
Chemical engineering 20:496
Chemical fertilizers 9:307–308, 17:677–678
Chemical sciences 3:371–375
  ancient technologies 3:371
  biotechnology 3:374
  Davy on 3:372–374
  electricity 3:372
  first chemical theories 3:371–372
  fragmentation 3:373
  Hoffman on 3:373–374
  Lavoisier's Revolution 3:372
  mature science 3:372–373
  metallurgy, alchemy, and pharmacy 3:371
  reduction 3:373–374
  status 3:374
  synthetic biology 3:374
Chemical thermodynamics 10:755
Chemoaffinity theory, Sperry's contribution 23:279–280
Chemosensory stimuli 11:607
Chen, Y., on cooperative marriage in China 21:735
CHER *see* Consortium of Higher Education Researchers (CHER)
Chernetsky, Vitaly 21:743
Chernobyl nuclear accident 20:833, 24:143–144
Chernoff, John 1:742
Chernyshenko Conscientiousness Scales 4:660
Cherry blossom 3:868
Cherubim 16:806
Chess 17:876–877
Chess, S., temperament development approach 24:191
Chevalier, French 1:965
Chevreuse 2:703
Chevron 17:166
CHF *see* Congestive heart failure (CHF)
Chiang, Mai
  works
    Pali *Jinakalamalipakaranam* 11:82
    Thai *Mulasasana* 11:82
Chiang Kai-shek 3:515
Chiapas-based Zapatista Liberation Army (EZLN) 11:758
Chiapello, Eve, on capitalism 5:536
Chiasm–callosum sectioned (split-brain) monkey 23:293
Chicago
  advocacy planning 1:221
  urban conflict 24:786
Chicago Board of Options Exchange 5:540
Chicago Conspiracy trial 18:499
Chicago Health and Social Life Survey (CHSLS) 12:686–688
Chicago Hellfire Club (CHC) 13:887
Chicago Law and Economics scholarship 13:794
Chicago Planning Commission 20:421
Chicago pragmatists 16:739
Chicago School 5:158–159, 5:191, 7:558, 7:561–562, 14:835, 17:24, 19:677, 21:131, 22:237–238, 22:996, 24:812, 25:678, 25:814
  antitrust policy 1:799
  community sociology studies 4:362–363
  and definition of situation 6:19
  Middletown and *Middletown in Transition* 3:195
  pragmatism 10:403
  qualitative case studies 22:993
  social control 22:237–239
  social ecology 22:263–268
  social work research 20:504
  studies of subcultures 23:638
  symbolic interaction 22:529
  symbolic interactionism 23:856–857
  and Thomas, William Isaac 6:19–21
  urban sociology 8:179–180, 21:238
  youth culture 25:814–815

Chicago School and the Birmingham Centre for Contemporary Cultural Studies (CCCS) 2:852–853, 23:638, 25:809
Chicago Teachers' Strike (CTU) 24:494
Chicanas 9:785
Chichewa, gender assignment in 9:796
Chief Bernard Ominayak and the Lubicon Lake Band v. Canada 11:798–799
Chief executive officer (CEO) 16:194, 17:614
  promotion and 19:136–137
Chief financial officers (CFOs) 7:70
Chiefdoms
  archaeology of 3:376–382
  and craft specialization 3:379
  critiques and refinements of 3:377–378
  and cultural evolutionary theory 3:376–377
  evolution of 3:381
  horticulture 11:204–205
  ideology of rulership 3:380
  network *versus* corporate strategies 3:378
  political and social features 3:378–379
  political economy 3:379
  and potlatch 18:715–717
  tribes 24:646
  warfare among 4:603
  warfare and militarism 3:380–381
Chiefly militias, warfare among 4:603
Chiefs 2:568, 11:895
  and big man 2:569
  transition to kings 11:895–896
Chieftaincies, African, external relations 1:288
Chieftains, in Europe 8:271–272
Chijun, Murayama 2:71–72
Chikafusa, Kitabatake, works, *Jinnô Shôtôki* 11:50–51
Child, I.L., works, *Six Cultures Series* 7:145
Child, use of term 12:956
Child abandonment, and adoption 1:137, 1:139
Child abuse 3:410, 3:412, 3:414, 3:418, 3:424, 3:473, 3:478–479
  in black 3:420–421
  defining 3:424–425
  in minority ethnic children 3:420–421
  and neglect 3:410, 3:412, 3:414
  and social work 3:473
Child amnesia 2:290–291
Child and adolescent mental health 8:745, 8:748
Child and adolescent mental health service (CAMHS) 3:386
Child and adolescent psychiatry (CAP) 3:383–390
  cultural considerations 3:384
  development 3:385
  as distinct medical specialty 3:383–384
  family environments 3:384–385
    authoritarian parenting 3:385
    authoritative parenting 3:385
    parenting 3:385
    permissive parenting 3:385
  features distinguishing 3:384
  information leaflet 3:389t
  most valued perspective 3:386
  principles of an effective approach to service delivery 3:386–388
    assessment 3:386–387
    assessment process 3:388
    context 3:386
    developmental assessment 3:387
    diagnosis and measurement 3:388
    family history 3:387
    high-quality services 3:386
    importance of development 3:387
    mental state assessment 3:388
    presenting concerns 3:387
    risk assessment 3:387–388
    setting the context 3:387
    social history 3:387
  school environments 3:385–386
  social contexts and experiences 3:384
  terminology 3:383
  training 3:383–384, 3:387
Child and Adolescent Trial for Cardiovascular Health (CATCH) 10:670
Child Behavior Checklist 16:726, 22:946
Child benefits, family policy 8:801
Child birth
  environmental factors of mental illness before and during 15:205
  and religiosity of women 25:636
Child care 3:391
  after-school care 3:392
  cognitive development 3:391–393
  effects based on family background 3:393
  extensive 3:391–392
  and fatherhood 8:840–841
  future care 3:393
  grandparent–grandchild relationships 10:336–339
  language development 3:392
  maternal deprivation 3:391
  maternal employment 3:391
  mother–child relationship 3:391
  peers 3:392–393
  public policy issues 3:393
    family leave *versus* infant care 3:393
    public *versus* private funding 3:393
  quality of 3:392
  residential 20:534
  single-mother families 3:391
  social skills 3:392–393
  socioemotional development 3:391–392
  and time use 24:379
  type of 3:392–393
Child care policies 16:162
Child characteristics, motivation 15:921–923
Child custody 15:566
  and forensic psychology 9:354
Child development 3:452, 3:480, 3:482–483, 11:276–277, 15:874, 15:876, 15:878
  across cultures 3:395–398
    human development 3:395–396
  and breast feeding 2:847–849, 2:851
  and childhood friendship 9:442–444
  computer games and 4:498
  cultural influences on 3:453–455
  definition 2:847
  developmental niche 3:395–397
  effects of prenatal stress on 23:556–561
  emotional interactions with caregivers 16:513–514
  families as educational settings 8:731–732
  impact of maternal mental illness on 17:724
  learning at school 16:515
  like me theory 15:131, 15:133
  and media 15:1–10
  and motherhood 15:883
  motivation factors 16:512–513
  motor skills 15:971–977
  and neonatal bonding 16:469–473
  Neo-Vygotskian Model of 16:513
    analysis of developmental periods 16:513–516
  object-centered joint activity 16:514
  and parenting 19:246
  paternal influences on 8:841–842
  pathways 3:395
  social situation of 25:332–333
  social-emotional development 22:942–948
  sociodramatic play 16:514–515
Child Development Supplement (CDS) 24:390
Child Family and Community Service Act (CFCSA) 11:765–766
Child fertility
  and mortality 8:822
  in old age 8:822
  and quality 8:821–822
Child friendly rights 12:958
Child guidance clinics, historical background 3:879
Child health, and socioeconomic status 22:924–925
Child Health and Human Development Study of Early Child Care 3:835
Child health psychology 10:679
  concept of health and illness 10:680–681
  coping with chronic illnesses 10:684
  developmental milestones 10:680t
  developmental stages of illness concepts 10:680t
  developmental tasks and competences 10:679–680
  dietary behavior 10:682–683
  emerging new interdisciplinary field 10:679
  family 10:679–684
  health status 10:681–682
  health-related behaviors in childhood and adolescence 10:681
  obesity 10:681–682, 10:684
  physical activity 10:683–684
  prevention and health promotion 10:682
  prevention programs 10:684
  psychosocial strain 10:684
  understanding of health 10:679

Child health revolution 10:199
Child health services, and family planning 8:788, 8:790–792
Child identity theft 5:634, 5:637
Child labor 3:405–409, 22:44
  cash transfers 3:408–409
  as cause for public concern 3:406
  definition 3:405–406
  policy implications 3:408–409
  and poverty 3:406–407
  regional estimates of 3:405t
  schooling 3:406–407, 3:409
  in South Asia 7:197
Child labor laws 3:481
Child language 13:245, 13:247, 25:325
Child language acquisition 12:782
Child maltreatment, and child protection 3:424–425
Child Maltreatment 2011 21:659
Child mental health 19:299
  psychiatric social work in 19:303–304
Child morality
  indirect estimation of 6:143–144, 6:144f, 6:144t
    consistency check 6:146
Child mortality 1:153
  decline in 15:868, 15:870–872
  and fertility 8:822
  and fertility transition 9:96
Child neglect 3:410–416, 3:418
  child abuse 3:410, 3:412, 3:414
  construction 3:411–412
    good enough and bad enough parenting 3:412
    measuring and proving 3:412
    preoccupation with dirt and disorder 3:411–412
  contextual factors 3:412–414
    attachment theory and mother-blaming 3:414
    drug and alcohol misuse 3:413
    family dysfunction 3:413–414
    gender 3:412–413
    intergenerational neglect 3:414
    poor parenting, characteristics of 3:413
    risk factors 3:413
  depression 3:413–414
  domestic violence 3:413–414
  fathers 3:412–414
  mothers 3:412–414
  neglect
    defining 3:410–411
    reemergence of 3:410
  nurturing 3:411–415
  physical and sexual abuse of 3:418
  poverty 3:414–415
    beyond simple causation 3:414
    social neglect 3:414–415
  threshold 3:410–412
Child physical abuse, and religion 20:320
Child prostitutes
  prosecution of 3:481
Child prostitution 3:427
Child protection 3:417–422, 13:722
  aboriginal children in 3:420
  critical issues in 3:418–420
    risk assessment 3:418–419
    skills 3:419–420
  defining 3:425
  disabled children 3:421
  and fathering 8:844–845, 8:847
  international issues in 3:423
    defining child abuse 3:424–425
    defining childhood 3:423
    defining children's rights 3:423–424
    in disasters and emergencies 3:425–426
    Internet abuse and online protection 3:427–428
    maltreatment 3:424–425
    protecting children from sexual abuse, exploitation, and sexual trafficking 3:426–427
    safeguarding children abused in armed conflicts 3:426
    social work 3:425–428
    war 3:426
    welfare 3:423, 3:427
  international policy context 3:417–418
  jurisdictions 3:417–418
  and motherhood 15:884–885
  understandings 3:420–421
    black children 3:420–421
    minority ethnic children 3:420–421
Child protection agencies 3:479
Child protection interface, psychiatric 19:304–305
Child protection services 3:418, 15:884
Child psychiatry, autism spectrum disorders 2:276–277
Child quality
  and fertility 8:821–822
    quantity–quality interaction 8:821–822
Child rearing, and religiosity of women 25:636
Child rights 12:956
Child Savers 25:814
Child schooling 3:406–407, 3:409
Child sexual abuse, law criminalizing 3:481
Child Sexual Behavior Inventory (CSBI) 21:841
Child soldiering 25:379
Child soldiers 3:430–438, 25:392–393
  adolescents 3:433–434
  and armed conflict 1:970–972
  depression 3:433–435
  experience in life-history context 3:431–434, 3:432f
  former, ethical conduct of research and interventions with 3:431
  future directions for research and intervention 3:435–436
  mental health of 3:435
  Nepal 3:431–435
  postconflict situations 3:431–434, 3:436
  post-traumatic stress disorder 3:433–435
  reintegration programs, traditional healing, and other interventions 3:434–435
  Sierra Leone 3:430–435
  trauma 3:430–436
  Uganda 3:430, 3:432–435
  as victims and actors 3:439–444
    context for engagement 3:440–441
    guerrilla fighters 3:439–443
    methodological considerations 3:439–443
    military regime 3:439–444
    political activism 3:441–443
    political engagement 3:440, 3:442–443
    revolution 3:440–444
Child spacing 8:823–824
Child survival 9:77
Child welfare 3:419–420, 20:537
  and child protection 3:423, 3:427
  fathers roles 3:420
  mothers role 3:420
Child welfare policies 20:689
Child well-being
  and cohabitation 4:120
  and family policy 8:803–804
Child-Abuse-Related-Deaths (CARDs) 19:305
Childbearing 9:19, 11:6–7
  within cohabitation, in Western societies 17:547–549, 17:549f
  economics of 8:821–825
    Becker on 8:821, 8:823
    child mortality and fertility 8:822
    child quality and fertility 8:821–822
    child spacing 8:823–824
    contraceptive 8:821–824
    cost of children 8:822–823
    demographic transition 8:822
    divorce risk 8:824–825
    dynamic fertility models 8:823–824
    Easterlin's relative income hypothesis 8:824
    fertility and support in old age 8:822
    human capital 8:821–823
    life-cycle dynamics 8:823–824
    marriage market 8:824–825
    nonmarital fertility 8:824–825
    quantity–quality interaction 8:821–822
  and employment among women 8:295–296
  and fertility 9:3–4, 9:12
  unmarried 8:768–769
Childbirth
  and motherhood 15:881, 15:883, 15:885
  perinatal psychiatry 17:723
Childcare
  across cultures 3:395–398
    accessibility of early childcare 3:397
    caretaking arrangements 3:396–397
    early childhood education and care by professional caretakers 3:397
  and culturally appropriate interventions 5:519
  and fertility recovery 9:777–778
Child-centered approach, early memory development 6:761
Child-centered research, street children 23:534, 23:537
Child-directed speech (CDS) 13:308

Child-directed speech (CDS) (*continued*)
and adult-directed speech (ADS) 3:400
influence on language development 3:399–404
associations between maternal linguistic input and language acquisition 3:399–400
in developmental disabilities 3:402
effects of maternal depression 3:402
extralinguistic features of 3:401
features 3:401
identifying features that impact language acquisition 3:400–401
pitch and prosody 3:400
social interaction, necessity of 3:401
universal nature 3:401–402
Childe, V. Gordon 4:607, 8:103, 14:670–671
CHILDES database 13:315
Child-focused activities 3:395–396
Childfree movement 20:839–840
Child-friendly environments 23:213
Child-friendly justice 12:957
Childhood 3:491, 6:842–843, 15:922–923, 20:298, 23:579–580, 24:176
and age policies 1:324
anthropological aspects 3:451
arts education 1:234
basic trust and attachment 3:454–455
body image 2:737–738
conceptions in anthropology 3:452–453
and conflict 4:634
conflict during 4:595–598
coping during 4:889
cultural acquisition 3:455–456
cultural scale and complexity 3:453
defining 3:423
developmental assets 3:445–450
development and health 11:267–269
developmental goals 3:455
early
development of attachment during
see also Attachment
empathy development during
see also Empathy
face perception during
see also Face perception
friendship during
see also Friendship
joint attention during 12:844–850
memory development during 15:147–154
play during 18:218–225
prosocial behavior during
see also Prosocial behavior
social referencing during
see also Social referencing
early adversity in *see* Early adversity
emotional development 3:454, 12:1–6
emotions in 12:4–5
environment 6:843
episodic memory in *see* Episodic memory
evolutionary developmental psychology 8:420–424, 8:426–427
foundational principles 3:492
friendship during
see also Friendship
gender differences 3:453–454
health-related behaviors in 10:681
HPA system 23:577, 23:580
identification 11:538–539
identity in 11:537–542
life course 14:76–78
life span development 14:109
logical and hypothetical reasoning in 14:306–307
long-term memory in *see* Long-term memory
mother–child communication in 15:876–878
autobiographical narratives 15:877
cross-cultural and quasi-experimental studies 15:877
discursive practice approach 15:877–878
nonverbal tests and developments in 21:896
and nostalgia 17:34–36
prosocial behavior during
see also Prosocial behavior
psychodynamic approaches and studies of 19:362
self-development in 21:492–497
developmental differences in domain-specific self-concepts 21:492–493
early childhood 21:492
later childhood 21:493
middle childhood 21:492–493
sexual development, longitudinal analyses of 14:328–339
sexuality 21:840–841
cultural and social differences 21:841
normative sexuality 21:840–841
short-term memory in *see* Short-term memory
social studies of 3:492
socialization in 22:866–872
stages of 3:451–452
stress and elevated cortisol levels in 23:581
study of, anthropological methods 3:456
in twenty-first century 3:456–457
temperament 24:185
trust across life span 12:638
verbal tests and developments in 21:896
Childhood amnesia 21:899
and early memory development 6:761
and memory development 15:147, 15:151–153
Childhood cancer 3:459–464
adjustment between dysfunction and resilience 3:462–463
bereavement 3:461
cognitive-behavioral therapy 3:462
development benefit 3:462
differential developmental paths and outcomes in survivors of 3:463f
effective coping by family counseling 3:459–460
ethics 3:461
intensive treatment phases 3:461
long-term adjustment outcome 3:462
long-term survival and reintegration 3:462
posttraumatic stress disorder 3:462
protective family resources for adaptation to 3:460t
from protective to open approach 3:461
psychological adaptation as balance of stress and resources 3:459
psychological counseling strategies and interventions 3:460–462
quality of life 3:459–460, 3:462
recurrent relapses, end-stage of disease, death, and dying 3:461–462
stressors, demands, and challenges 3:460t
supporting initial coping with diagnosis 3:460
trauma 3:462
Childhood Cancer Survivor Study (CCSS) 3:462
Childhood death attitudes 5:860–861
development stage 5:860–861
irreversibility 5:860
nonfunctionality 5:860
psychosocial stage 5:860–861
universality 5:860
Childhood deprivation 22:286–287
Childhood family processes 20:293
Childhood lead poisoning 7:759
Childhood memories 3:465–470
beyond preverbal-to-verbal transition 3:469
blended memory procedures 3:465–467
hide-and-seek task 3:465–466
magic shrinking machine 3:466–467
episodic memory 3:466, 3:469
evidence of verbal translation of preverbal memories 3:467–468
long-term fate of preverbal memories 3:467
long-term memory 3:465–466
nonverbal memory procedures 3:465
verbal memory procedures 3:465
verbal reminders 3:468–469
visual paired-comparison paradigm 3:465, 3:468–469
what-where-when memory 3:466
Childhood obesity/overweight 15:39–40
and media 15:5–6
Childhood physical illness 11:268
Childhood self-concepts 21:462
Childhood smoking prevention 8:300–301
Child-initiated social play 8:394
Childlessness 12:47–48
voluntary 12:46–48
Childrearing 3:395, 20:319–320
Children 3:491, 21:426
achievement in affordable private schools 1:270
adoptable, changing sources of 1:138–139
A-not-B task 8:491–493, 8:491f
attachment strategies 2:141
avoidant-dismissing style 21:426
as Bayesian statisticians 4:46
behavior problems 2:477
effects of socioeconomic status on 2:477–480
biomedical research on 13:561
from birth to 36 months of age

socioeconomic status and social-emotional development 22:942–948
capacities of 3:483
caring for 11:762–763
and ceremonial practices 11:762
changes in living arrangements of 8:841
characteristics, and parenting 22:869
cognitive development in 8:490–492
and cohabitation 4:120
common sleep disorders in 22:70
bruxism (tooth grinding) 22:73
difficulties with arousals 22:71–72
difficulty falling asleep 22:70–71
excessive daytime sleepiness 22:73
sleep-related breathing disorder 22:73
sleep-related movement disorders 22:72–73
and community 11:762
competencies and dispositions, development of 8:752–754
cultural capital 8:753–754
economic capital 8:753
social capital 8:754
variation in family structure 8:754
consequences of nonmarital childbearing for 9:16
as consumers 3:481
contributions to family relationships 22:158
cost of 9:63–64
cultural capital of 5:377
and culturally appropriate interventions 5:517–521
dependency of 3:483
as discipline 3:491
disciplining and nurturing, and religion 20:320
dominant research themes 3:493
and education 7:144–148
EF studies in 8:489, 8:495, 8:489t
methodological issues 8:494–495
effects of parenting and family structure on emotional development of 7:438–444
engagements 3:493–494
experiencing stress, working with 23:573–574
factors affecting spelling of 23:266–270
factual knowledge 15:284–285
and family 8:809–813
family size preferences 8:805–808
with focal injuries, neuroimaging studies of 16:588–589
geographies of 3:491–497
as global discipline 3:492
health and socioeconomic of family planning outcomes among 8:790
health status 10:681–682
human capital 9:60, 9:62
indigenous child caring 11:763
of labor intensity 9:63
and law 3:477–484
learning 4:41
academic learning 8:396
academic self-efficacy 8:396
explicit awareness of language sounds 8:395
motivation to learn 8:395–396
probabilistic learning 4:38
theory of mind 4:39
theory theory 4:41
learning process 3:451–456
lexical consolidation in 13:923
living with parents, percentage of 8:841t
and marriage 8:829–830, 8:832–833
mass trauma events in
psychological consequences 14:748
well-being and functioning in aftermath 14:748
as mathematicians 4:45–46
media effects on 15:35–41
mobilities 3:495
moral behavior 15:775
mortality 9:66, 12:20
in industrialized countries 12:14–20
in less developed countries 12:21–26
new directions for 3:495–496
nurturing of 11:762
and online pornography 17:219–220
opportunity cost of 9:62–63
parent–child tie, primacy of 8:782–783
participation 3:395–396
participation in civic life 3:481–482
as physicists 4:44–45
playful spaces 3:494
city center 3:494
schools 3:494
and political socialization 18:509–511
post-divorce adjustment of
child factors 14:523
interparental conflict 14:522
interventions for problems 14:521–522
parenting 14:522–523
stressful events 14:521–522
variability in outcomes 14:521–523
pragmatic disorders in 18:789–794
protection of 3:473–474
and psychosocial work 19:479–481
quantity and quality of 9:62–63
refugee 20:108
research 3:493
residential care *see* Residential child care
rights 3:456, 3:477–478, 3:480, 3:485–490, 3:492, 12:961
autonomy within family 3:488–489
basic human rights 3:486–487
conditional 12:961
contingent 12:961
education 3:486–487
equal opportunity 3:487
future 3:489
historical context 3:485–486
and juvenile justice 12:956–962
liberty and participation rights 3:488
meaning of term 3:486–489
protection rights 3:487–488
UN Convention on Rights of the Child 3:485
welfare rights 3:487
Roma 20:756
schema for 3:491–492
school learning, family socioeconomic status and 8:752
secure style 21:426
self-regulation 25:317
sleep pattern changes 22:70–71, 22:71t
as social actors 3:492
in social work
see also Social work
socialization 15:775
source monitoring in 23:33
spatial engagements 3:493–494
speech, language, reading and writing impairment in 3:794–795
sports 23:302–305
strains of dual-career families with 14:607
television viewing–aggression relationship 25:101
theory of mind 4:39
theory theory 4:41
time with, dual-career families 14:607
TOL task 8:493–494
trafficking in Latin America 24:512–516
universal child 3:492–493
value of 3:498–501, 5:538
child benefits 3:499
child costs 3:498–501
culture 3:500
decline of fertility 3:498–499
economic value 3:498–500
family 3:498–501
fertility 3:500–501
first- and later-born children 3:499
gender 3:499–500
measurement 3:499
parity 3:499, 3:501
psychological value 3:499
social capital 3:498, 3:500
socioeconomic variation 3:500
theoretical development 3:498–499
transition of fertility 3:501
variations 3:499–500
value of life for 25:18–19
war and political violence effects on 25:389–394
of working memory 8:494
young, direct assessment of memory of 6:761–763
Children, Young Persons and Their Families (CYPF) Act 1989 14:506, 14:508
Children Act 1908 5:209
Children Act 1989 1:146
Children and law
across children's lives 3:477–482
assessment of law's role in children's lives 3:482–483
caregiver–child relationships 3:477–479
educational experiences 3:480
juvenile justice and criminal law systems 3:480–481
market and civic life 3:481–482
parent–child relationships 3:477–478
peer relationships 3:482
procedural protection 3:480

Children and law (*continued*)
protection from maltreatment 3:478–479
terminating parental rights 3:479, 3:482
Children ever born (CEB) 15:398–399, 15:401–403, 15:401t
Children in care, and social work 3:474–475
Children of Divorce Intervention Program 14:523
Children of God/The Family International 16:806
Children understanding
of earth science concepts 21:275–276
of life science concepts 21:278
of physical science concepts 21:276–278
of space science concepts 21:275–276
Children's Assessment of Participation and Enjoyment 13:834
Children's behaviors
parents' causal attributions regarding 8:734
Children's Geographies 3:491
Children's learning, fostering 8:732–733
Children's Memory Scale (CMS) 16:724–725
Children's rights 3:492
defining 3:423–424
Children's Rights Index 3:417
Children's school learning
family socioeconomic status and 8:752
Children's Television Act of 1990 (CTA) 7:246–247
Child-specific beliefs, motivation 15:923–924
Child-to-child sexual aggression 21:840
Chile 22:626
gay/lesbian activism in 13:860–862
psychiatric care in 6:42–43
Chilean race 8:246
Chilembwe uprising 15:532
Chiliasm 15:529
Chimeras 10:38
Chimpanzees 13:123–125
Chin, E.M., on capitalism in North America 17:24
Ch'in and Han dynasties 2:919–920
China 2:65, 2:919–920, 3:865, 3:867, 5:454, 13:595, 22:640
achievements in technology and science 16:399
decimal system 16:399
dialectic materialism 16:398–399
Dujiangyan irrigation system 16:399
idea of evolution 16:398
in military field 16:399
movable-type printing 16:399
ancient and medieval China
ancient human settlement in 24:869–870
Feng-Shui 24:869–870
authoritarian approach 13:431
Beijing Consensus 3:505
birth control in 9:3–4
-centered view of neighbors 16:275–276
Christianity in 3:531–532
association between Imperialism and 3:531–532
during Communist regime 3:531–532
conversions to Christianity 3:531–532
radical indigenization of Christianity to Chinese culture 3:531–532
three-self churches 3:531–532
cities with foreign-occupied concessions 24:874
climatic variability 3:867
concessional loans 3:502–506
Confucianisms 6:778
corporate governance 4:936
cultural capital research in 5:379
cultural policies in 5:454
creative industries 5:457
cultural policy regimes in
see also Cultural policies
Cultural Revolution 25:816
death penalty in 5:876–877
development of sociology 24:584–585
diasporas 6:355
divorce in 14:518
economic cooperation 3:502
economic development 11:868–869, 11:871
economic dominance 11:858
economic growth 24:878–879
economic transformation 7:73–74, 7:76–79
elder care in 7:340
eugenics in 8:233–235
e-waste network 7:367f
family planning 15:611
fertility transition in 9:46–48
aging 9:47–48
birth control 9:47
consequences 9:48
demography 9:47–48
family 9:47–48
government programs 9:47–48
historical background 9:46–47
infanticide 9:47
Malthus on 9:46
marriage 9:46–48
One-Child policy 9:47–48
population control 9:46–47
process 9:46
sex ratio 9:48
foreign aid by 3:502–507
aid procedure 3:505
with Chinese characteristics 3:505–506
decision making 3:505
distribution and allocation 3:504–505
major impacts of 3:506
management 3:505
modalities and volumes 3:503
model 3:505–506
projects 3:503–504
terms and definitions 3:502–503
gender gap 15:612–613
genetics of alcohol metabolism 17:238–239
grants 3:502–503
greening of business in 10:392–396
Han Code 6:779
heterosexuality in 10:819–824
historiography development in 11:49
history of opium use 17:238
inheritance law in 19:175
intellectual property (IP) violations 13:431
interest-free loans 3:502–503
investment in West Africa 1:303
Islam in 12:728
law and courts 13:431
legal orders during middle period 6:779–780
Ming Dynasty 6:780
Ministry of Commerce (MOFCOM), China 3:505
modern, subimperialism, colonial modernity, and policies of gender revolution in 6:795
Mongol conquest of 12:728
multicultural education in 16:19
multinational corporations 16:79
NRMs 16:806–807
Official Development Assistance (ODA), China 3:502–503, 3:506–507
opium addiction in 17:237–238
opium smokers 17:237, 17:239–240
organized crime in 5:173
overseas aid 24:301
politics of remembrance in 4:184
professional urban planning 24:878–879
psychiatric care in 6:44–45
Qin Law 6:779
Qing Code 6:781–782
reform and rapid urbanization since 1980s 24:876–877
regional planning 20:134–140
religious forms in 20:369
sex ratio at birth (SRB) 15:609
sex selection in 15:610–611
sexual citizenship in 21:731
sexual minorities in 21:734–739
decriminalization, medicalization, and stigmatization 21:735–737
definitions and prevalence 21:734
family obligations and imperative of heterosexual marriage 21:734–735
quality citizens, internal divisions, and transgender people 21:738–739
sexual identity, cultural and social belonging 21:737–738
social work in 14:667–668, 22:789–794
abolition in 1950s 22:789–790
active agency of social work community 22:792
emergence in 1920s 22:789
future development 22:792–793
harmonious society 22:790–793
indigenization of social work 22:792
lack of strong knowledge base 22:793
rapid expansion and professionalization since 1999 22:790–791
re-emergence in late 1980s 22:790
regional disparities 22:792–793
social work education as catalyst 22:792
strong leadership role of government 22:791

tension between needs of masses and demand for professionalization 22:793
tension between Westernization and indigenization 22:793
in socialist societies 22:849–852
sociocultural aspects of 2:65–70
aging 2:69
anthropology 2:65
collectives 2:66–69
community building 2:68
consumerism 2:69
cosmopolitanism 2:67
education 2:65–69
environment 2:67, 2:69
ethnicity 2:66, 2:68
family 2:68–69
gender 2:69
health care 2:67
housing 2:66–68
inequality 2:68–69
Internet 2:68
land 2:66, 2:68
language 2:66, 2:68
life styles 2:69
Manchu 2:65–66
marriage 2:65, 2:68–69
media 2:68–69
migration 2:65–69
minorities 2:66–67
mobility 2:65–69
modernization 2:65, 2:69
nation 2:66
one-child-policy 2:69
popular culture 2:69
religion 2:65–66, 2:68–69
sexuality 2:69
social media 2:68
society 2:65–69
sociology 2:68–69
and Taiwan 2:65–66, 2:69
and Tibet 2:65, 2:67
tourism 2:68
transnationalism 2:69
unfolding modernities 2:67–69
unities and diversities 2:65–67
work units 2:66–69
Xinjiang 2:65, 2:67
South–South cooperation 3:505
Soviet model and the anti-city movement 24:874–876
Soviet–Chinese partnership/rivalry, in Cold War 4:128, 4:130
state organized enterprises (SOE) 3:659
state owned enterprises in 4:933, 4:936
Tang Code 6:780
technical cooperation 3:503, 3:505
techniques development in 5:530
traditional Chinese cities 24:872–874
transforming planned economy of 6:789–790
unregistered religions 16:806–807
urban definitions in 24:859–860
urban planning 24:869–881
urban planning education 24:879–881
urbanization in 24:973–979
housing 24:976
land politics 24:976
land-based accumulation 24:975
local citizenship and rights awareness 24:977
local state building and state entrepreneurialism 24:975–976
middle class 24:977–978
migrants 24:973–974, 24:976–978
remaking landscape 24:976–977
rise of industrial production and uneven development 24:974–975
socialist market economy 24:974
trend 24:973–974
uneven geography 24:977–978
urban as unit of accumulation 24:974–976
Yuan dynasty 6:780
China Artists Association 5:600
China Association of Social Work Education (CASWE) 22:790, 22:792
China Association of Social Workers (CASW) 22:790
China Children and Families Cohort Study 14:74
China First Coal Project 2:243
Chinese 20:1, 20:7–8
Cantonese 3:508
homophone and lexical compounding 3:509–510
literacy acquisition 3:510–512
Mandarin 3:508
orthography 3:510
syllable and tones 3:509
word reading and spelling 3:508–509
Chinese Academy of Contemporary Arts 5:600–601
Chinese Academy of Social Sciences (CASS) 22:637–639
Chinese capitalism 3:582
Chinese cities 24:814
Chinese Classification of Mental Disorders (CCMD) 10:820, 21:736
diagnostic category of homosexuality 3:884, 3:889
Chinese Communist Party (CCP) 2:66–68, 3:515, 5:596, 14:661, 22:790
Central Committee 5:596–597, 5:600–601
Chinese Cultural Revitalization Committee 5:454–455
Chinese cultural traditions 11:786
Chinese dynastic historians 11:120
Chinese eugenics 8:231
Chinese Exclusion Act 1882 15:463, 18:650
Chinese Filial Piety Awards Ceremony 21:735
Chinese folk religions 16:806–807
Chinese frontiers 9:466–467
Lattimore on 9:467
military defense 9:467
socioeconomic systems 9:467
Chinese intellectuals 12:278
Chinese International Economic and Trade Arbitration Commission 12:470
Chinese international migration 12:525
Chinese literacy acquisition 3:508–514
Chinese Muslims (Hui) 12:728–729
Chinese National Academy of Arts 5:600–601
Chinese negotiators 16:424
Chinese parenting 22:159
Chinese politics
in Cold War period 6:799–800
in post–Cold War period 6:800–801
Chinese Psychiatric Association 21:736
Chinese Revolutions 3:515–519
Chinese Revolutions, Twentieth Century 3:515–519
categories and definitions
Geming 3:516–517
standard categorization schemes 3:517
Chinese cases 3:518–519
cultural and intellectual change 3:517–518
diplomatic change 3:518
events 3:515–516
new Marriage Law 1950 3:515, 3:518
overlapping 3:517–518
politics 3:517
socioeconomic change 3:518
twenty-first-century perspective 3:519
unfinished revolutions 3:518
Chinese room 21:342
Chinese self 12:697
Chinese Tongzhi Conferences 21:737
Chinese Turkestan 3:316
Chinese World Order 16:275
Chinese writing 25:778
Chinese-speaking children 24:270
Chinmokhoe hoebo 24:585
Chint'ae, Son 2:72
Chipman, Alice 6:332
Chirographa 19:179–180
Chiropractic 10:562
Chishing'a 23:860–861
Chi-square distance 5:2
Chi-square test 5:2, 6:577–578, 11:494, 17:662, 23:616–618
in ethnography 8:190
loglinear models 16:128–129
Chladenius, Johann Martin 10:785, 11:121
Chlamydia trachomatis, and sexually transmitted infections 6:470
Chlorofluorocarbon gases (CFCs) 1:724
Chlorofluorocarbons 7:794
Chlorpromazine 1:778, 21:22–23
for schizophrenia 21:18
Choi, K., on men who have sex with men 21:736–737
Choice axiom 5:895–896, 14:804
Choice behavior 25:506–507
Choice probability, random utility model 5:901–905
Choice scholarships 21:71–72, 21:75–76
Choice theory
mathematical psychology 14:804–805
Choice tree model, random utility model 5:904–905

Choice-based sampling *see* Outcome-based sampling
Choice-of-law 13:616–617
Choices 3:520
as agency 3:520–521
anomalies of 16:642–643
axiomatic theories, Tversky 24:679–680
and collective action 3:524
as consumer 3:522–523
diffusion and random walk processes 6:395–401
economic psychology 5:886–891
as empowerment 3:521–522
and friendship 9:426
impacts on income distribution 11:735–737
international social work 3:521–522
in interpersonal relationships 5:427–428
mathematical psychology 14:813–814
as neoliberal policy 3:523–524
options, valuation of 19:440–441
predicting 19:441–442
random utility models 5:901–905
of resource measure 11:733
self-determination 3:520–521, 3:523–524
service-user involvement 3:520–524
sex work as 19:281–282
social work practice, engaged 3:520, 3:524
social worker 3:524–525
Cholecystokinin (CCK), and placebo effect 18:154
Cholecystokinin tetrapeptide (CCK-4) 11:607
and anxiety 1:804, 1:808–809
Cholera 18:176–177, 25:428–429, 25:433–434
in Haiti 25:431, 25:434
spread of 25:432, 25:432t
Cholera Treatment Centers (CTCs) 25:432–433
Cholesky components 23:401
Cholesky decomposition 23:401
Cholesterol Ester Transfer Protein 2:796
Cholesterol levels, and behavioral health 20:266
Cholinergic–aminergic imbalance hypothesis 22:83
Chomsky, Avram Noam 3:526–530, 6:504, 14:215–217, 16:678, 23:235
and anarchism 3:526
and behaviorism 3:527
binding conditions 1:675
on Cartesian linguistics 3:526–528
contribution to linguistics 3:526–530
generative grammar of 15:754
generative tradition of 23:890, 23:893–894
on grammatical agreement 10:317–318
and Harris 3:527
on imperialism 3:529–530
on Jewish intellectuals 3:526–527
lexical categories of 23:885
linguistic competence of 4:296
linguistics of 14:228–229
principles and parameter theory of 10:317–318
as public intellectual 3:530
and radicalism 3:527, 3:529
on structuralism 3:527
syntax of 15:753
on tacit knowledge of rules 13:74–76
faculty of language 13:75
Quine's challenge 13:75–76
theory of universal grammar 3:529, 9:206–210
parameters 9:206–207
parameters and cross-linguistic acquisition 9:207–209
universal principles and acquisition 9:209–210
transformational grammar of 15:754
works
Aspects of the Theory of Syntax 3:528, 9:873
Cartesian Linguistics: A Chapter in the History of Rationalist Thought 3:527–528, 14:217, 25:344
Language and Mind 3:528
The Logical Structure of Linguistic Theory 9:871–872
New York Review of Books (*NYRB*) 3:528
Reflections on Language 3:528–529
The Sound Pattern of English 3:528, 14:534
Syntactic Structures 4:104, 9:871, 23:885
Topics in the Theory of Generative Grammar 3:528
Chomsky Hierarchy 14:297
Chomsky–Herman collaboration 3:529
Chong-bok, works, *Tongsa kangmok* 11:50
Choo, Eugene 2:100, 2:104–106
Chopping/stretching transformation 19:79
Choreography 5:666–667
Chorography 13:224–229
Chorology 10:83, 13:228–229
and nature society 16:335
Chorus of prototypes 25:154
Choudary, Anjem 24:230
Choukri, Mohamed, works, *Le pain nu* 21:753
CHP *see* Republican People's Party (CHP) (Turkey)
Chraïbi, Driss, works, *Le passé simple* 21:752–753
Christ Cult 3:546–547
Christaller, Walter 23:199
Christchurch Tissue Bank, in New Zealand 9:964
Christendom 8:258
Christiaan Smuts, Jan, works, *Holism and Evolution* 17:911
Christian Bible 20:399
Christian blogs 20:276
Christian charity 4:338
Christian chronology 24:409–411
Christian church 13:801, 20:336
Christian church Fathers 16:168
Christian civilization 15:347
Christian conceptions of history 11:33
Christian education 15:614
Christian era system 24:411–412
Christian eschatology 15:531
Christian fundamentalism 20:409–410
Christian gospels 3:237
Christian grassroots communities 14:33
Christian historiography and historical thought
apocalyptic thought 11:33–34
basic models 11:33–34
current problems and questions 11:34
inside perspectives, problem of 11:31
interpretation of history 11:31–32
outside perspectives, problem of 11:31
selection mechanisms 11:32–33
Christian millenarianism
cargo cults 3:139–141, 3:143
Christian millennialism 15:531–532
Christian mission 3:237
Christian missionaries 24:602–603
Christian psychology 19:408
Christian radicalism 20:402, 20:404
Christian realism, and peacemaking 17:638
Christian Right movement 20:281
Christian theodicies, and functionalism 9:540–541
Christian theology 11:31, 20:284–285
Christian tradition 11:38
Christianity 2:144, 2:886, 2:891, 8:285–288, 10:485, 12:363, 15:229, 15:614, 20:285, 20:303–304, 20:322, 20:325, 20:340–342, 20:362, 20:368–369, 20:398, 21:173–174, 24:602
anomie in 1:718
anthropology of 20:307–311
in Arab nations 16:376
asceticism 2:60–61, 2:63
catholic *versus* Protestant 2:62
monastic 2:61–62
in Asia 3:531–533
among Karen tribal group 3:532
in China 3:531–532
Christian social militancy 3:531
converts of South India 3:531
in Indonesia 3:532–533
in Japan 3:532
in Korea 3:532
in Malaysia 3:533
modern style of 3:533
Pentecostal and Charismatic 3:533
in Singapore 3:533
in South India and Sri Lanka 3:531
in Taiwan 3:532
in Thailand 3:532
in Vietnam 3:532
in Australasia 3:543–544
beliefs about mental illness 19:307
in Canada 3:542–543
in Central Africa 3:314
in China during the missionary era 15:616
cultural logic of *see* Mission within the Christian tradition
Dalits and 3:226
in English-Speaking West 3:540–544
and family 8:809
historical realities 3:540–541
on homosexuality 3:885, 3:890–891
Illich's contribution 11:587

in Japan 12:796
in Latin America 3:534–539
conquest, resistance, accommodation 3:534–536
conversion 3:534–535
Liberation theology 3:537–538
Millenarian movements 3:537
popular religion 3:537
rebellions 3:535, 3:537
religion and everyday life 3:537
religion before the conquest 3:534
religion during the colonial period 3:536
religion in new republics 3:536–537
rituals 3:534–537
liberal 14:8–10
medieval 3:239–240
and miracles
see also Miracles
modern Catholicism 3:240–242
morality 15:807
New Testament era 3:237–238
in North Africa 16:376
origins 1:701
origins of 3:545–548
Patristics era 3:238–239
reformation 8:287
religious tolerance 18:238–239
in South Asia 2:76, 2:78
in Southeast Asia 23:70–71
strong reaction against Copernicanism 20:286
and syncretism 23:881
in United States 3:541–542
use of Internet 20:274
in West Africa 1:301–302
and world culture 10:227, 10:231
Christianity, origin of
apostolic succession 3:545, 3:547–548
Christ Cult and the Gospel 3:546–547
distinct visions of 3:548
early Jesus movements 3:545–546
Gospel tradition 3:547
heretical manifestations 3:548
historical thought and historiography 11:38
Jesus' teachings 3:546
miracle stories 3:546
'New Testament' canon 3:548
path to orthodoxy 3:548
pronouncement stories 3:546
theoretical considerations 3:545
Christianization 3:535, 15:350–351
and history 1:748
Christians, in the MENA region 15:357
Christman, John 2:316
Christus Medicus 15:119
Chromatin, structure and modifications 7:839–841
Chromatin-marking systems, epigenetic inheritance 7:833–834
Chromosome mapping 9:942
Chromosomes 9:826
epigenetic inheritance 7:833–834
Chronic diseases
diabetes 6:337–341
and environment 15:88
and gender 9:704, 9:706–707
prevalence, in old age 1:394, 1:396
Chronic disorder, managing experience of 6:475
Chronic fatigue syndrome (CFS) 17:134
Chronic illnesses
behavior 11:599–601
management 11:600
self and social conflict 11:600–601
children coping with 10:684
complementary and alternative medicine 10:564–565
coping with 3:549–554
accommodative coping and meaning-focused coping 3:550–551
behaviors 3:549
coping engagement and disengagement 3:550
distinctions in coping behaviors 3:550–551
emotion regulation 3:552
emotional suppression tendencies 3:551
experiences requiring coping efforts 3:549–550
health effects of 3:552
illness beliefs 3:552
individual differences in styles 3:551–552
interventions 3:552–553
monitoring and blunting coping styles 3:551–552
problem-focused and emotion-focused coping 3:550
relationship-focused coping 3:551
and human development 11:268–269
prevention and treatment of 10:675–676
social comparison 22:214
social support and recovery 22:736–739
Strauss on 23:528–530
Chronic mental disorder 8:234–235
Chronic mental illness, psychosocial interventions for 22:380–385
Chronic migraines 16:718
Chronic myelogenous leukemia (CML) 17:926
Chronic obstructive pulmonary disease (COPD)
comorbid psychological symptoms in 20:576–577
interventions in 20:578–579
social aspects in 20:577
Chronic offender 5:141, 5:143
Chronic pain 3:555–561, 3:556t
and mindfulness 15:551–552
role of learning in 17:452–453
treatments 3:558–560
based on cognitive behavioral model 3:559–560, 3:559t
based on operant model 3:559
comprehensive approaches 3:558–560
historical approaches 3:558
Chronic problems, operant techniques for 2:428–429
reinforcers which increase desirable behaviors 2:429
negative reinforcers 2:429
positive reinforcers 2:429
reinforcers which reduce undesirable behaviors 2:429
punishment 2:429
response cost 2:429
Chronic psychosocial adversity
and coronary heart disease risk 4:919–920
emotional responses 4:919
mediating mechanisms 4:920
other forms of chronic stress 4:919
protective psychosocial factors 4:919–920
work stress 4:919
Chronic stress, and interoception 12:618
Chronic stressors, work–family conflict as 8:761–762
Chronicity, pain 17:452–455
Chronicle 11:61
hijri calendar 11:64
Chronicles of Dissent (interview) 3:529–530
Chronological age 9:855
Chronological charts, longitudinal data 14:347
Chronological tables 24:413–414
Chronologization 14:77
Chronology 11:55, 11:64, 24:409–414
almanac 11:55
of ancient Egypt 1:691
in archaeology 3:562–567
age estimation 3:562
stratigraphic 3:564
Christian era system 24:411–412
chronological dating 11:55
constraints, settlement and landscape 21:656–657
cyclical 24:411
linear 24:410–411
Middle Ages 15:348
stratified 24:411
Chronometric dating methods 21:656–657
Chronometric function 19:474
Chronosystem, EST 6:841
Chronotype 3:568–573
and age 3:570
biological circadian rhythms 3:570–571
circadian system, organization of 3:569–570
cognitive performance rhythms 3:571
attention tasks 3:571
executive functions tasks 3:571
memory tasks 3:571
measurement of 3:568–569
mental disorders 3:572–573
attention deficit/hyperactivity disorder 3:573
drug use and dependency 3:572–573
eating disorders 3:572
mood disorders 3:572
personality characteristics 3:571–572
preventive and therapeutic approaches 3:573
and sex 3:570

Chrysi Avgi 18:412
Chrysostom, John 11:885
CHS *see* Comparative health systems (CHS)
CHSLS *see* Chicago Health and Social Life Survey (CHSLS)
Chu, Q.S., on sexual minorities in China 21:734
Chuang Tze 4:312
Chukotka 1:921, 1:923
  bronze in 1:923
  OBS culture in 1:924
  Thule culture in 1:924
Chung, Byung-Ho 2:74–75
Chunking, and memory binding 15:140, 15:143
Chuo University 12:795–796
Church 7:299
  and conservatism 4:689–690
  and state 3:574–578
    autonomy 3:575–576
    contemporary challenges 3:577–578
    cooperation 3:576
    freedom of religion or belief 3:574–578
    functions of religions 3:574–575
    human rights and institutional approaches 3:575–576
    identity 3:575, 3:578
    international instruments 3:575, 3:578
    migration 3:577
    pluralism 3:576, 3:578
    religious approaches 3:577
    separation of religions and state 3:574–577
    typology of systems 3:576–577
  terminology 3:574
Church attendance, religiosity and prejudice against minorities 20:382–383
Church Communities International 2:146
Church council of Chalcedon 17:435
Church history 11:31
Church of Cyprus 17:436
Church of Greece 17:435
Church of Jesus Christ of Latter-day Saints 16:805
Church of Scientology 16:806
Church of South India 17:438
Church of the Flying Spaghetti Monster 16:806
Church Universal and Triumphant 16:806
Churchill, Sir Winston Leonard Spencer 6:373, 9:214, 21:372
Church-related human rights networks 24:604
Church–state relationships in Europe 8:262–263
Churyu 2:788
Chvostizm 20:634
CI *see* Categorical imperative (CI); Cochlear implant (CI); Confidence intervals (CI)
Cialdini, Robert 22:349
  principles of influence 22:349–351
    authority 22:350
    commitment and consistency 22:349–350
    liking, attractiveness, similarity, ingratiation, and immediacy 22:350
    reciprocity 22:349
    scarcity 22:350–351
    social proof 22:350
Cicero 3:707, 4:312, 7:144–145, 13:799
  classical legacies of 3:771–772
  on mass society 14:744
  and republicanism 20:477–481
  works
    De Officiis 17:638
Cicourel, Aaron 23:437
CIDCM *see* Center for International Development and Conflict Management (CIDCM)
CIDI *see* Composite International Diagnostic Interview (CIDI)
CIDOC *see* Center for Intercultural Documentation (CIDOC)
CIDs *see* Common interest developments (CIDs)
CIES *see* Comparative and International Education Society (CIES)
CIEs *see* Cyber Islamic Environments (CIEs)
CIESAS *see* Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS)
Cigarette smoking
  religion and 20:266
  and social networks 16:558
CIHA *see* Comité; International d'Histoire de l'Art (CIHA)
CIIG *see* Continuous Instrument Improvement Group (CIIG)
CIIM *see* Common in-group identity model (CIIM)
CIL *see* Central Identification Laboratory (CIL)
CILs *see* Centers for Independent Living (CILs)
Cimarrones 23:45
CINAHL database 22:944
Cinchona bark (Peruvian bark) 18:176
Cincinnati Initiative to Reduce Violence (CIRV) 25:822
CINDA *see* Development University Center (CINDA)
Cinema Verité 6:622
Cinemas 18:530, 25:821–822
  early 9:166
  national 9:167–168
  sound 9:168
Cinematography 15:12–13, 15:15
Cingranelli and Richards (CIRI) physical integrity measure 18:446–447
Cingranelli–Richards (CIRI) Human Rights Dataset 5:837
Cingulate areas, connections with 18:847
Cingulate cortex, and emotions 7:477
Cingulotomy 23:18
  obsessive compulsive disorder 17:269
Cinnirella, M., nonheterosexual British Asian friendship 16:865
CIO *see* Congress of Industrial Organization (CIO)
CIOMS *see* Council for International Organizations of Medical Sciences (CIOMS)
cIPS *see* Caudal intraparietal sulcus (cIPS)
CIRC *see* Cooperative Integrated Reading and Composition (CIRC)
Circadian genes 22:494
Circadian rhythm sleep disorders 22:75–76
Circadian rhythms
  regulation by hypothalamus 11:488–489
  and sleep 22:90–91
Circadian typology 3:568
Circle of limited security, attachment 12:230
Circle of security model, for attachment 12:230, 12:233
Circuits (sexualized spatiality) 21:809–810
Circular causality 22:912–913
Circular characteristics, five- factor model (FFM) 2:563
Circular economy 10:393
Circular flows 3:91
Circular migration 11:649, 15:394, 18:598
Circular model
  marketing strategies in 14:556–557, 14:556f
    in game of imitation 14:557
    in game of movement 14:557
    in game of position 14:556–557, 14:557f
Circularity 11:649
Circulating capital 3:90
Circulation, globalization 10:233–235
Circulatory diseases, mortality from 1:163
Circumcision, Pharaonic 8:879
Circumfixes 15:821
Circumplex model, emotion 7:484
Circumplex theory 16:40
  and genocide 10:17
Circumpolar North 1:927–931
Circumstantial modality (deixis) 6:54–55
Circumstantial-focus (CF) construction 23:892
Circumventricular organs (CVOs) 11:487
CIRI *see* Cingranelli and Richards (CIRI) physical integrity measure; Cingranelli-Richards (CIRI) Human Rights Dataset
CIRV *see* Cincinnati Initiative to Reduce Violence (CIRV); Glasgow Community Initiative to Reduce Violence (CIRV)
CIS *see* Commonwealth of Independent States (CIS) countries; Community Innovation Surveys (CIS)
CISG *see* Convention on Contracts for the International Sale of Goods (CISG)
Cisgender 21:768, 24:549, 24:553
  and social work 21:802–803, 21:805–807
CISS *see* Coping inventory for stressful situations (CISS)
Citadel of dogmatism 20:828–829
Citalopram 1:779, 1:781
  for obsessive-compulsive disorder 17:125
Citation, theories of 21:325
Citation analysis 20:530

Citation classics 15:246
Citation index 5:809
Citations
journals 12:864–865
open access impacts on 12:867
patterns in academic publications, and degree of coherence of discipline 23:227–228
Citator 5:806, 5:808
CITES *see* Convention on International Trade in Endangered Species of Wild Fauna (CITES)
Cities 3:585–592, 4:363
built environment, in Africa 24:967–969
and citizenship 3:640–641
classical legacies 3:774
communication and transportation 4:274–275
culture types of 24:802
definition of 24:942–944
and development 6:282
factorial ecology of 24:801
and gender 9:710–711
growth of 22:265
healthy 24:832–833
intercultural comparison of 24:802
Jacobs on 12:777–779
and migration 15:424–425
and modernity 15:689–690
monocentricity and polycentricity of 24:823–824
under postsocialism 3:605
in ruins 18:670–672
slums 1:218
socialist 3:606–607
as a source of urban conflicts 24:786–787
and spatial mismatch 23:157–160
traditional Chinese cities 24:872–874
urban anthropology 24:774–782
and urban violence 24:955
urban violence in 24:949–950
and urbanization 24:960–964
WHO Healthy Cities project 24:833
Cities, internal structure of 3:593–604
accessibility and transport technology 3:593
of city centers 3:594f
distance decay 3:593–594
hierarchical structures 3:594
labor-oriented and leisure-oriented 3:604
land values gradient 3:593–594
political systems *see* Political systems
social ecological models of 3:597f
social gradients 3:593
Cities and financial crisis 3:579–584
capital switching 3:579–580
Castells on 3:579–580
crises around the world 3:581–582
crisis management 3:582–583
financialization 3:579–580
globalization 3:583
Harvey on 3:579–580
housing bubble 3:580–581, 3:583
mortgage market 3:581–583
occupy wall street (OWS) 3:582
overaccumulation 3:579–580
residential mortgage-backed securities (RMBS) 3:581–583
social movements 3:582–583
subprime lending in the United States 3:580–581
urbanization 3:579, 3:581
Citing Decisions tab 5:806
Citizen access to police 18:263
Citizen armies 16:306–307
Citizen competence, and environmental cognition 7:711
Citizen ethnographer 5:383
Citizen journalism 3:612–618, 9:392, 12:860
challenges of 3:612
critical discourse analysis 3:615–616
functionalist studies of 3:614
historical studies 3:616
postpositivist studies of 3:613–614
technological determinism and convergence in 3:614
theoretical approaches 3:613–614
Citizen mapping 3:619–624
cartography 3:620–621
geographic information systems (GIS) 3:619–621, 3:623
critical 3:620
geoweb 3:619, 3:621–623
methodologies for researching 3:621–623
neogeography 3:619
as social processes 3:620–621
spatial data and representation in 3:619–620
volunteered geographic information 3:619–620
Web 2.0 3:619–621
Citizen participation 3:625–630, 12:914
challenges in 3:627–630
empowering citizens 3:627–628
citizenship 3:625–626
enabling participation 3:629
normative and practical arguments for 3:626–627
representing citizens 3:628–629
representing communities 3:628–629
Citizen science (CS) 3:631–636, 21:253–254, 23:130
amateurs in science 3:631–633, 3:635–636
boundaries and expertise 3:632–633
ethical considerations 3:634–636
and knowledgeable publics 21:256
lay–experts 3:632–633
public engagement 3:631–635
science communication 3:631–632, 3:634
scientists' and public motivations 3:633–634
Citizen-consumer 4:753–754
Citizens 3:643
judgments of 13:749
knowledge about economy 7:28
personality profiles of 18:429
police–citizen encounters 18:258–259
of the state 19:604
sustainability lifestyles 23:828–829, 23:833
Citizenship 3:637–639, 3:643–644, 3:650–654, 5:824, 9:770, 11:641–647, 14:492, 17:516, 18:400, 20:673–674, 20:676–678
Addams on 1:95
alterity 3:639
anthropology 3:639–640
anthropology of 3:637–642
apolitical 3:651–652
Benhabib on 2:519, 2:523–524
biocitizenship 2:610
in Central America 1:624–625
in China 24:977
and cities 3:640–641
and community 4:388
concept of 15:196
conceptions 3:643–645
and criminology 5:246–249
cultural dimensions 24:588–589
data infrastructures for research into 5:826–828
destabilization of 3:638
development as core of modern community 9:539
ecofeminist 6:849
exclusion through disenfranchisement 11:642
flexibilization of 3:638
forms 3:637–638
genetic 9:1019
and globalization 10:235–236
historical development of 3:643–649, 3:655–656
and immigration 11:671–673
immigration and emigration 11:650–652
and immigration policy 11:662, 11:664
in interconnected world 11:651–652
internal *versus* external boundaries 3:657–658
legal dimensions 24:588–589
liberal democratic citizenship 3:647–648
liberalism and 14:24
margins 3:639
meanings of 3:650–651
and media 3:660
migrants 3:638–639
and military service 15:512
and modernity 15:689–690
moral aspect of 3:659
and nationalism 16:263–264, 16:291–292
nation-state sovereignty 3:638–639
new and/or disqualified persons 3:656–657
and new genetics 9:1018–1019
notion of 15:418
political 3:650–654
globalization 3:652–653
human rights 3:652–653
legal status 3:650–652
membership 3:650–653
nationality 3:651–652
nation-state 3:651, 3:653
participation 3:650–653
republican 3:650–652
self-governance 3:650–653
transnational 3:652–653

Citizenship (*continued*)
political economy of 20:895
politics 3:638–640
in migration 15:423–424
postnational, prospects of 3:652–653
as a process 3:659–660
and property 19:157–158
and public policy 3:657
refugees 22:769–770, 22:772
and republicanism 20:480
republicanism and liberalism 3:645–647
rights and obligations 3:655–656
of sex workers 19:279, 19:282–283
sexual *see* Sexual citizenship
sites of 3:637–638, 3:640
and social integration 12:245–246
sociological aspects 3:655–661
sociology, Marshall 14:626–627
current state 14:627–629
and statehood 23:348, 23:350
supranational 24:592–594
theories 3:655
transnational 24:590
universal/particular rights and individual/group rights 3:657
variations in regime types 3:658–659
world 24:588–589
Citizenship Act 1924 (US) 1:635
Citizenship education 16:19–20
conflicts in 22:730–731
Citizenship grant 7:575
Citizenship performance, and personnel selection 17:937
Citizenship without the vote *see Civitas sine suffragio*
Citizenville 3:660
Citoyen 13:438
Citrin, Jack, on national identity 16:312
Citron, Michelle, works, *Daughter-Rite* 6:623
CITs *see* Crisis intervention teams (CITs)
City Beautiful Movement 16:809, 24:863–865
City centers 3:494
City Connects 8:613–614
City Corporation Ordinance 24:898
City Development Strategies (CDSs) 23:516
City marketing 3:662–668, 24:920
business city 3:664–665
city branding 3:666
city images 3:665
cosmopolitan city 3:665
cultural city 3:665–666
global city 3:664
good city 3:665
green city 3:665
historical context 3:662–664
place wars 3:662
postindustrial city 3:664
selling the city 3:665
symbolic reconstruction of cities 3:665
tactics of 3:666–667
City of God 12:387
City of man 12:387
City planning 3:625–630
City population 24:859–860
City regionalism 20:127
City size, and sustainability 18:199
density and car dependence 18:199–200
City urban 24:859–860
City-building process 12:116
City-region plans 20:138
City-regions 20:138–140, 24:946–948
City-states
in East Africa 6:774
Greece 1:697–698
Civic competence of ordinary citizens 2:194
Civic culture 2:190–192
theory 2:190
Civic education 22:727, 22:730–731
Civic Education Survey (CivEd) 5:773
Civic engagement 16:794–795
and Skocpol, Theda 22:31, 22:34–35
in United States 11:631–632
Civic epistemologies 2:680, 21:204, 21:258
science and 21:258
Civic exchange theory 12:386
Civic humanism
Pocock's history of 3:63
Skinner's view of 3:63–64
Civic integration 11:672
Civic journalism 3:613, 15:46
Civic nationalism 8:92, 16:263, 16:310, 16:312, 22:889–890
Civic renewal 4:313, 4:316
Civic republicanism 6:91–93, 6:95, 9:402–403, 19:168, 20:895, 20:897
constitutive notion of 9:402–403
instrumental notion of 9:402
Pettit and 9:402–403
Sandel and 20:895, 20:897
Skinner and 9:402
Civic virtues, and communitarianism 4:308–309
Civic voluntarism model (CVM) 18:506, 22:178–179
Civic-respectable (C-R) style 21:875
Civil and political rights 12:956
Civil and Political Rights of the International Bill of Rights 2:671, 22:393
Civil and Procedural Law 13:410
Civil approach, to domestic violence 6:628–629
Civil contempt of court 15:593
Civil disobedience 18:424–426
effectiveness of 18:426
forms of 18:425
justification of 18:425–426
Civil disputes 6:557
Civil Gideon movement 18:738
Civil human rights 11:355
Civil law 3:669–673, 13:565, 16:357, 23:722–726
adjudication 5:104–106
categorization 3:669
culpa in contrahendo, principle of 3:671
developing of 3:671–672
difference between common law and 3:672
in European Union 3:672
and forensic psychiatry 9:348–349
history 3:670
Latin America 13:410–411
outlook 3:672–673
and punishment 19:632–634
Security Law
French civil code 21:403–404
German civil code 21:403–404
sources 3:670–671
structural aspects 3:669–670
theory of 24:983–984
trials 24:636–638
Civil law colonies, in sub-Saharan Africa 23:655–656
Civil law judges 12:880–881
Civil law jurisdictions, discovery in 6:520–521
Civil law model 12:881
Civil law systems 4:809–810, 8:582, 13:615, 13:628–629
Civil liabilities
and damages 12:911
Civil liberties 3:674–679, 9:389, 18:446–447, 18:449–451
basic issues 3:675–676
definitions 3:674–675
distinction between 'positive' and 'negative' rights 3:674
historical background 3:674
modernization 15:697
in practice 3:676–677
protection of 3:676
relationship to state 3:675
Civil Litigation Research Project (CLRP) 13:590, 13:595–596
Civil Operations and Rural Development Support Program (CORDS) 5:100
Civil partnership registration 21:731
Civil Procedure Rules 4:250, 6:516–520
Civil religion 2:508, 3:680–685, 7:665, 21:381, 22:888
American priestly 3:682
American prophetic 3:682
in Buddhist, Hindu, and Muslim contexts 3:682
as coercive 3:683
as a cultural container 3:682
definition 3:680–681
Durkheimian notions 3:680
and Enlightenment 7:665
future of 3:683–684
general religiosity and 3:683
key elements 3:682
poles in understanding of 3:680–681
in political mobilization 3:682–683
Rousseauian concept of 3:682–683
work on 3:683
Civil Religion in America 1:523, 2:508
Civil Renewal Unit 4:316
Civil rights 3:674–675, 3:686–690, 12:898–899, 12:956, 13:516, 13:542, 13:544–545, 13:547, 13:593, 18:387
abolishing racial discrimination 3:687
absolutist conceptions of 3:687–688
across national borders 3:688
and counterculture 5:84–85

and desegregation 6:218, 6:221–222
in fundamental sense 3:688
gay/lesbian 9:640
Marshall on 14:626–628
origins 3:686–687
prospects for 3:689
protection from government 3:687
and racism 19:855
role of court 3:688
scope 3:687
sources 3:686
universality 3:687–688
US tradition of formulation and enforcement 3:688–689
Civil Rights Act 1991 11:882
Civil Rights Act 1964 (CRA) 3:693, 4:741, 6:530–531, 6:598, 7:151, 11:882, 11:911, 18:431, 21:721, 24:551
Title VII 1:256–259, 8:915, 23:221
Civil rights activists 5:266–267
Civil rights laws 14:266–267, 20:884
procedure in 14:266
Civil Rights Movement 3:691–694, 7:811–812, 14:605, 16:18–19, 20:281, 21:107, 25:815
black oppression 3:691
challenges 3:693–694
goals 3:692–693
influence on stigma 23:463–464
international achievements 3:693
national achievements 3:693
origin of 3:691–693
political protests 18:423, 18:426
strategies 3:692–693
tactics 3:692–693
in the US 19:855
women in 9:738
Civil servants 3:695, 3:697, 19:97
Civil service 3:695–700
agency theory 3:697–700
blurred boundaries and continued relevance 3:700
bounded rationality 3:697–698
bureaucracy 3:695–700
civil servants 3:697
collaborative government 3:697
decision-making systems 3:698
delegation of power 3:697
executive branch 3:697–699
government 3:695–699
history 3:695–696
impact of privatization of 3:698–699
institution-building 3:700
modern 3:696
new states 3:696
nondemocratic regimes 3:696
new public management 3:698, 3:700
new version of 3:698
policy making 3:697–699
public administration 3:695–698, 3:700
rational model 3:696–697
and regulation 3:699
studying 3:697–698
Weber 3:696–698
Civil service ethics, public policy 19:583
Civil society 1:539–540, 3:366, 3:701–712, 9:674–675, 9:770, 10:738–740, 10:742, 13:441–442, 16:800–802, 18:386–387, 19:604, 19:606–607, 22:576–582, 22:750–751
in African countries 18:390
Alexander on 1:526–527
charity 22:576–580
and community 4:388
critique of 16:857
culture, Eastern European Studies 6:807
differentiation in 3:704–705
donating 22:580
donors 22:579–580
giving 22:576, 22:578–580
Hegel's analysis of 3:709–710, 10:738
justice-related associations 18:389
Koinonia Politike
and its historical modulations 3:707–708
transformed 3:708–709
late-twentieth century understandings 3:710–711
and liberal state 3:710
and liberalism 14:14
media 18:389
and modernization 15:707
NGO 22:576
and nongovernmental organizations 16:857
organizations 22:576–577
definition 22:576–577
practice 22:577
social policy 22:577
philanthropy 22:579–580
definition 22:579
political support 22:580
practice 22:579–580
political liberalism 18:389–390
political parties 18:390
political prerequisites of 23:344–345
political sociology 18:472–473, 18:477
religious groups 18:389–390
as result of exchange 12:387–388
role in global public policy 10:181–182
and self organization 3:701–702
as separate domain 3:709–710
and Skocpol, Theda 22:34–35
and social science 22:580–581
and state
liberalism and 14:23
third sector 22:576
voluntary sector 22:576–578
volunteering 22:577–579
definition 22:577–578
reasons for 22:578–579
social policy 22:579
volunteers 22:578
and Walzer, Michael 25:349
and world culture 10:227–228
Civil society activism 10:203
Civil society and ecological economics 6:860–861
Civil society groups 17:650–651
Civil Society Organization Index 3:368
Civil society organizations (CSOs) 12:371, 25:253, 25:259, 25:261
Civil sphere, Alexander on 1:526–527
Civil system, and social system 22:750–751
Civil union 3:708
Civil Union Act 2006 (South Africa) 21:744, 21:746
Civil unrest 5:367
Civil violence 24:786–787
Civil vital event registers 6:139
Civil wars 3:713–718, 8:97, 8:99
causes 3:716–717
concept 3:713–714
cross-national statistical studies 3:715–716
and crowds 5:364
English 7:649–653
literature 3:714–716
normative perspectives 3:715
North–South perspective 3:715
and territorial conflict 24:216–217, 24:219
violence 10:3–5
in West Africa 1:302–303
women's security 25:638
Civilians 25:372–373
legal reasoning/argumentation
early 13:777–778
later 13:778–779
and violence 25:122
Civility, and punishment 19:658
Civilizational analysis 7:323
Civilizational missions of monks 15:725
Civilizations 3:719–729, 23:371
archaeology of 23:371–375
Axial Age 3:726
expansion of 3:727
internal transformation 3:728
multiplicity 3:727–728
beyond totality 3:723–724
concept of 3:719
and cultures, clash of 20:403–404
and ecumene 7:110–111
and elites 3:726–727
and energy 7:616
global history 10:162–164, 10:166
and irrigation 12:724–725
and mental illness 15:576
modernity
cultural and political program of 3:728–729
multiple 3:729
prehistory 3:719–720
relativism 3:721–722
and relativism 20:230–231
Sarmiento on 20:908, 20:910–911
in South America 23:36
universalism 3:722–723
word and concept 3:720–723
Civilized monarchies 11:413
Civilized parenting 18:913
Civilized society 18:914
Civilizing mission 4:523
of women, in Middle East 16:367–368
Civilizing process 19:636–639
Civilizing programs 3:313

Civil–military relations 11:433, 15:491–493, 15:512
  emergence of 15:519
Civitas Dei 11:33
Civitas diaboli 11:33
Civitas sine suffragio 3:644–645
CIW *see* Carnegie Institution of Washington (CIW)
CJ *see* Criminal justice (CJ)
CJS *see* Criminal justice system (CJS)
CL *see* Computational linguistics (CL); Computational linguists (CL)
CLACSO *see* Latin American Council of Social Sciences (CLACSO)
Cladistics 11:24, 11:26
Cladocera 7:694
Claessens, Stijn, works, *Financial Crises: Causes, Consequences, and Policy Responses* 9:183
Claeys, Gregory 11:686
Claim repayment 2:353–356
Claimed authenticity, and nostalgia 17:36
Claiming 12:135–136, 13:469, 13:594
  in disputes 6:561–562
Clair, René, works, *Sous les toits de Paris* 9:168
Clairvoyance 16:170
Clan events 2:567–568
Clan politics 3:319
Clancy-Smith, Julia, works, *Rebel and Saint* 16:384
Clandestine grave, searching for 9:334–335
Clans 3:730–732
  ancestor 3:731
  in Central Asia 3:318–319
  definition 3:730
  descent 3:730–732
  ethnicity 3:732
  genealogy 3:731
  identity 3:732
  kinship 3:730–732
  lineage 3:731–732
  social evolution 3:730
  totemism 3:730–732, 24:453–457
  tribes 24:646
  and tribes 3:730
Clapham sect 10:492
Claptrap 19:157
CLARION 2:39
Clark, A., on superpositional storage 5:580
Clark, John Bates 3:91
  works, *Distribution of Wealth* 25:11
Clark, S.D., works, *The Social History of Canada* 22:601
Clark, Terry Nichols 4:364
Clark, W. 3:320
Clarke, A.E., works, *The Right Tools for the Job* 8:579
Clarke, Edith, works, *My Mother Who Fathered Me* 1:640
Clarke, G., on women's participation in sports 9:657
Clarke, Ronald 20:776
  works, *Modeling Offenders' Decisions: A Framework for Research and Policy* 20:776
Clarkson, Laurence 7:654
Clarkson, Thomas 22:442
Clash of civilizations 11:435–436, 16:225, 20:623
CLASP *see* Center for Law and Social Policy (CLASP)
CLASS *see* Classroom Assessment Scoring System (CLASS)
Class 1:965, 3:745–750, 14:649–650, 22:338–344
  analysis 22:339
  in artisans 2:47–48
  in Australia 2:242–244
  -based politics, diffusion 6:412
  and bear subculture 13:891
  and butch–femme subcultures 13:872
  capitalist 3:99, 3:101, 3:103
  in Caribbean 1:641
  categories as 11:896
  in commercial societies 14:647
  and community sociology 4:363–364
  consciousness 14:650
  conspicuous consumption/distinction 4:695–697
  constructing 22:343
  and cultural distinction 22:342
  and cultural lifestyles 22:331
  and cultural policy 5:461
  Dahrendorf on 5:662–665
  dealignments 25:303
  death of 3:749
  decline 3:747–749
  decomposition 3:747–749
  and development geography 6:315–317
  differentiation of peasantry 1:458
  division of labor 25:341–343
  and education 5:375–376
  experience from production to consumption 22:341–342
  and fashion 8:836–837
  and friendship 9:425–429
  Giddens on 10:135–137
  habitus 22:334
  historical sociology 11:27
  in historiography 22:339–341
    Germany 22:340
    United States 22:340–341
  and honor 11:182–183
  of horsemen 1:965–966
  and ideologies 11:574, 11:576–577
  and intersectionality 12:663–668
  and labor 25:677
  and leatherman subculture 13:889
  marriage as marker of 8:780–783
  in Marxism 3:745–747, 3:747t
  and nation 16:292–293
  noble 1:965
  and parental ties 8:783–784
  and peripheral nationalism 16:292
  and poverty 18:719–721
  prehistory of 3:745
  proliferation 3:747–749
  realignments 25:303–304
  and same-sex sexualities 20:872
  and sexuality 21:811
  social 3:755–759, 14:644–645
  as social action, experience, and culture 22:341
  and social status 22:331–332
  and social structure 22:714–715
  in South Asia 2:79–80
  in Southeast Asia 23:72–73
  stigma 23:465
  structure 14:650
  struggles 14:642–643, 14:645–646, 14:650
  symbolic boundaries 23:851–853
  theories of 22:338–339
  transition of kinship to 11:895–896
  in twentieth century 3:746–747
    occupational classes 3:747
    polar classes 3:746
    stratification 3:747
  and white-collar workers 25:566–567
  working classes 25:694
Class actions 3:733–738
  attorney fees 3:733, 3:737
  collective litigation 3:733–734, 3:736–737
  collective redress 3:733–734
  consumer policy 4:734–735
  goals 3:733–734
  history 3:733
  litigation funding 3:735
  procedures 3:734–735
  scholarship 3:737
  uses 3:735–737
    economic transactions 3:736
    mass accidents and environmental disasters 3:736
    product safety and marketing 3:736
    reparations 3:736–737
    social impact litigation against governments 3:735–736
    social policy litigation against industry 3:736
Class analysis 3:747
  Goldthorpe on 10:280–283
Class conditional likelihoods, in statistical pattern recognition 23:411–413
Class conflict 4:589
Class consciousness 3:739–744
  absence of 3:742
  academic debates 3:742–743
  and class position 3:740, 3:742–743
  experience and 3:741–743
  false consciousness 3:740–743
  and hegemony 10:748–749
  Marxist origins 3:739–742
  Marx's 4:175
  objectivism and 3:739–740, 3:743
  perspectives 3:740–742
  rise in America 25:55
  socialism and 3:740, 3:742
  use of term 3:739
  white-collar workers 25:567
Class culture 23:851–852
Class differences
  and intelligence 12:319–320
Class domination, political sociology 18:472–473
Class formation, and social mobility 15:640

Class hazing 10:557
Class ideologies 14:652
Class in itself/for itself 3:740–741, 3:743
Class inequalities 14:671
Class position 25:695
Class regime 3:756f
Class size 3:751–754
causal inference 3:751–752
education policy 3:751–753
as educational input 3:751
experimental research 3:752
long-term impacts 3:752–753
quasi-experimental research 3:752–753
reduction 3:751–753, 7:271
statewide policy-induced variation 3:753
and student outcomes 7:152–153, 7:153f
studying impact of 3:751–752
Class skeptics 3:749
Class society 14:675
Class stratification, in archaic states 23:373–374
Class structure 15:534–536
Class struggle 4:302, 14:650, 14:653, 14:663–664
and capitalist accumulation 6:242
causes of 6:242
in China's economic development 4:303
social justice 22:394
Class terms (classifiers) 3:812
Class theory, intellectuals 12:278
Class voting 25:299–301
dealignments 25:303
democracy 25:299–300
left-right continuum 25:299–301
manual/nonmanual 25:299–300
new directions 25:300–301
political participation 25:300
realignments 25:303–304
recent innovations 25:300
two class/two party era 25:299–300
Class-discrimination experiments 21:947
Classic cognitive therapy 4:542
second wave as 4:543
Classic language planning 13:353–354
Classic legal sociology 13:550–551
Classic public policy analysis 24:903
Classic social theory 15:687–688
Classic strain theories 23:495–496
Classic thesis, secularization 21:388
Classic Weber location triangle 23:173f
Classical additive error models 14:850–851
Classical anarchism 1:680–681
Classical antiquity 20:650
Classical aphasic syndromes 1:820t
Classical aphasiology, written language deficits in 1:450–451
Classical archaeology 3:760–763
antiquity 3:760–763
archaic epoch 3:760–761
definition of 3:760–761
Enlightenment 3:760–761
history of 3:761–762
orientalist research 3:760
paths of 3:762–763
prehistoric research 3:760
Renaissance 3:760–761
rules of 3:762
Classical conditioning 16:845–846
and attitude 2:169–170
aversion therapy 3:766–767
conditional response 3:764–766, 3:768
conditioned aversion 3:767
d-cycloserine 3:769
in development of psychopathology 3:765–766
avoidance 3:766
exemplars 3:766
individual differences 3:765–766
and early memory development 6:762
escape 3:767–768
fear 3:764–769
future directions 3:769
and learning 23:876
methods in psychotherapy 3:764–770
cognitive considerations 3:768–769
covert sensitization 3:767
sensitization 3:766–767
systematic desensitization 3:767–768
operant conditioning 3:766–767, 3:769
and persuasion 17:941
procedures and processes of 3:764–765
exposure to US or CS 3:765
generalization and discrimination 3:764–765
prolonged exposure 3:767–769
propranolol 3:769
reinforcement 20:207–208
and suicide 23:677
two-factor learning theory 3:766
unconditional response 3:764–766
see also Pavlovian classical conditioning
Classical constraint-satisfaction problems, local search methods for 4:718–719
Classical diglossia 6:417–418
Classical economics
innovation and employment 12:153
profits in 19:111–112
Classical elite theories 7:397
Classical empire 3:773–774
Classical Freudian theory *see* Freudian theory
Classical game theory 12:194
Classical genetics 16:13
Muller's contributions to 16:14
Classical geopolitics 10:91–93
Classical German historism 25:292
Classical Hellenistic biography 2:644
Classical Hollywood, genres 9:163
Classical hypothesis tests, model selection by 15:655
Classical interpretation of probability 19:24
Classical legacies 3:771–777
anthropology 3:775–776
Cicero on 3:771–772
city and satire 3:774
classical empire 3:773–774
classics and moderns 3:776
Epicureanism 3:774–775
Livy on 3:772–773
Plato on 3:774
Popper on 3:774
psychoanalysis 3:775–776
republican virtue 3:772
Skinner's view of 3:64
stoicism 3:774–775
Tacitus on 3:772–775
Classical mechanics 3:778–784
and motor control 3:778–784
dynamics 3:779–780
evidence of feed-forward, model-based control 3:781–782
evidence of intermittent feedback control 3:782–783
evidence of stiffness control 3:782
feed-forward, stiffness, and feedback control 3:780–783
internal models 3:778, 3:781–783
Jacobian matrix 3:779
Lagrangian dynamics 3:780
stiffness ellipse 3:782
synergy formations 3:778–779
and scientific revolution 21:318, 21:320
Classical multidimensional scaling 9:888–889
Classical narratology 16:195–196
Classical operationalization 8:132
Classical period
warfare in 25:414–415
organization 25:414–415
purposes 25:415
technology 25:414
Classical political economy 18:358–359
Classical pragmatism 18:803
Classical records, different generalizations of 16:902–905
Classical reliability theory 9:834
Classical republicanism *see* Civic republicanism
Classical school of criminology 5:232–233
Classical science 21:320
Classical scoring model of forced-choice personality assessment 17:842–843
Classical social theory 13:551
Classical sociology, anomie in 1:719
Classical test theory (CTT) 3:785–791, 4:510, 19:419, 19:423–430, 20:248–250
applications of 3:790
axioms of 3:786
basic equation of 20:248
classical true score model 19:423–424
correction for attenuation 19:425–426
effect of heterogeneity on reliability 19:425
estimated true score 19:426
mean of the observed score variable 19:424
reliability coefficient 19:424–425
standard error of measurement 19:425
variance of the error score variable 19:424
variance of the observed score variable 19:424
coefficient alpha 19:427–429
coefficient omega 19:429
computerized test construction 4:510
concepts 3:785–786

Classical test theory (CTT) (*continued*)
error variables 3:786–787
primitives 3:785–786
true score 3:786–787
discussion 3:789–790
equivalence models 19:427–428
congeneric measurements model 19:428
essentially parallel measurements model 19:427–428
essentially tau-equivalent measurements model 19:428
tau-equivalent measurements model 19:428
error variances 3:786–787
greatest lower bound 19:428–429
lower bounds to the reliability of a composite variable 19:428–429
models 3:787–789
assumptions 3:787, 3:787t
practical issues 3:789
parallel measurements model 19:426–427
correlation between parallel measurements 19:426
effect of test length 19:426–427
effect of test length on criterion-related evidence of validity 19:427
generalized Spearman–Brown formula 19:427
reliability from a single administration of a test 19:427
parallel tests 3:787–788
parameters of 3:790
principal statistical results used in 19:424
reliability 3:786–787
t-congeneric tests 3:788
t-equivalence tests 3:788
weaknesses of 20:249–250
Classical theories 10:850–851
of knowledge 22:229
Classical theory of value 25:7–9
objective
labor theory of value 25:7–9
Marx on 25:8–9
revival of classical theory of value 25:9
Ricardo on 25:7–8
Sraffa on 25:9
subjective (neoclassical theory)
equilibrium 25:10–11
marginal productivity 25:11
marginal utility 25:10–11
product distribution 25:11
utility analysis 25:11–12
utility and welfare 25:12
Classical transportation problem 23:174t
Classical true score model (CTSM) 19:423–424
correction for attenuation 19:425–426
effect of heterogeneity on reliability 19:425
estimated true score 19:426
mean of the observed score variable 19:424
reliability coefficient 19:424–425
standard error of measurement 19:425
variance of the error score variable 19:424
variance of the observed score variable 19:424
Classical true-score theory (CTST) 20:239–240
common factor model representation of 20:242–243
Classical typology 14:208–209
Classical world, history of 10:162–163
Classicism, transition of films to 9:166–167
Classification 4:531–533, 6:704–705
anthropology of 13:70–72
Bayes decision rule for 23:412–413
process 4:531
social, used by statisticians 23:426
social causes of 6:244–245
in social measurement 21:4–5
aggregate entities 21:5
taxonomies 21:4
typologies 21:4–5
in social sciences 3:805–810
clustering result, evaluation of 3:809
concepts of classes 3:808
determining similarity 3:807
evaluation 3:809
examples 3:806
extension 3:806, 3:808
intension 3:806, 3:808
limits of discussion 3:805–806
procedures to assign cases to classes 3:809
purposes 3:806
selection of cases 3:806
specifying properties 3:806–807
terminology 3:805
use of term 23:392
Classification and Regression Tree (CART) methodology 16:118
Classification for individual consumption according to purpose (COICOP) 5:707
Classification schemes, cultural production 5:609–610
Classification systems 3:799–800
symbolic boundaries 23:850–852
Classification task 21:950
Classification trees, in propensity scoring 19:142
Classification/regression trees 19:142
Classificatory types 10:774
Classifiers, linguistics of 3:811–818
Amazonian languages 3:812
Australian languages 3:814
diversity of subtypes 3:816t
Egyptian hieroglyphs 3:815t
genitival classifiers 3:812
grammaticalization 3:816–817, 3:817t
lumping and nonlumping 3:812
morphosyntax 3:817t
nominal classification 3:811–812
noun classes 3:812
noun classifiers 3:812–813, 3:813t
numeral classifiers 3:812–813, 3:813t
repeater classifier 3:817
sign languages 3:814
Southeast Asian languages 3:816
unique classifier 3:815–816
verb classifiers 3:814
verbal classifiers (verb-CL) 3:812
verbs in Jamingjung 3:814t
Classless society 1:965–966
Classlessness, intellectuals 12:278–279
Classroom
level in educational effectiveness research 3:833–838
Classroom
and childhood friendship 9:445
engagement and behavior in and gender 9:683
learning inside, implications 22:11
learning outside 22:10–11
implications 22:11
social studies education 22:731
Classroom assessment 3:819–824, 14:753–754
characteristics of high-quality 3:821–823
alignment 3:821–822
efficiency 3:823
fairness 3:822–823
practicality 3:823
reliability 3:822
validity 3:822
components of 3:819–820
interpretation 3:820
learning objectives or targets 3:819
use of the results 3:820
grading (marking) and reporting 3:823–824
summative and formative 3:820–821
Classroom Assessment Scoring System (CLASS) 3:835
Classroom characteristics 21:122
Classroom climate 3:825–832
instruments for assessing 3:825–828
teachers' attempts to improve 3:829–830
Classroom engagement 1:31
classroom management/structure 1:33
peer support 1:33
task characteristics 1:33
teacher support 1:33
Classroom environment 7:820
Classroom Environment Scale (CES) 3:825–826
Classroom experiences 15:894
Classroom learning
social-cognitive perspectives 13:674–675
Classroom management 24:56–57
physical arrangement of 24:47
positive classroom 24:47
routines and procedures 24:47
talk between teacher and student 24:47
Classroom Observation System 3:835
Classwide Peer Tutoring 4:884
Clastres, Pierre 1:682
Claude Reyes v. Chile (2006) 9:396
Clausal form, reasoning 5:935–936
Clauses 13:278, 16:4
knowledge space 4:7–8
Clausewitz, Carl von 15:519–520, 15:522, 23:523, 25:407
concept of the pause 25:409
on military history 15:509
on war 1:983, 25:380, 25:407–409

Clausius, Rudolf 12:107
Clawson, Mary Ann, on gender role in popular music 3:801
Clay tablet archives 5:700
Clay's model 10:46–47
Clayton, Daniel, works, *The Dictionary of Human Geography* 11:21
Clayton, Paula 25:625–626
Clayton Act 1914 1:797–798, 23:318
Clayton–Bulwer Treaty 1:827
CLE *see* Continuing Legal Education (CLE); Cultural Learning Environment (CLE)
Clean Air Act 1990 10:374
  National Ambient Air Quality Standards for key pollutants 7:773
Clean Air Act Amendments 1990 7:727, 24:619
Clean technologies 12:716
Clean water 25:429
Clear selection rule 1:77–78
Clear-cut process. organizational change as 6:299
Clearing Union Plan 13:18
Clearwater River Denesuliné protocol 14:187
Cleavages, political 4:692–694
Cleff, Ed van 16:228
Cleft habitus 9:142
Cleghorn, H.F., works, *Forests and Gardens of South India* 7:750
Cleland, J., on gay athletes 9:659
Clement of Alexandria 3:548
Clement of Rome 3:548
Clemmer, D., works, *The Prison Community* 18:920
Clergy
  malfeasance 20:256
  representational role of 20:281
  sexual misconduct 20:256–257
  syncretistic 23:881–882
Clericalism 13:216
CLES *see* Constructivist Learning Environment Survey (CLES)
Cleveland, Emmeline 25:623
Cleveland, Ohio, advocacy planning 1:221
Clever Hans phenomenon 8:573
Cleyre, Voltairine de 19:748
CLIA *see* Clinical Improvement Amendments 1988 (CLIA)
Click-evoked otoacoustic emissions (CEOAEs) 21:776
Client driven applied geography 1:845
Client populations, diversity, and social work stress/trauma 18:709
Client-based work groups 13:629–630
Client-centered psychotherapy 20:737–738
Client-centered therapy
  development 3:880
  hypothesis
    Rogers&rsquo 19:504
Clientelism 17:603–606
  in socialist societies 22:849–850
Clients
  and armed conflict 1:972–973
  narratives, understanding 16:208
Clifford, James 2:872
  works, *Writing Culture* 8:590
Clifford, W.K., works, *Common Sense of the Exact Sciences* 17:660
CLIL *see* Content and Language Integrated Learning (CLIL)
Climate 3:865–872, 7:720–722, 16:334–338
  body, evolution of 2:724–725
  definition 17:322
  of organizational diversity 4:740–741, 6:597–598
Climate, organizational *see* Organizational climate
Climate change 1:931, 3:839–847, 3:865–866, 6:479, 7:694–696, 7:769, 7:771–774, 7:776, 7:781, 8:943, 8:947, 18:104–105, 23:796–797, 24:612–613
  adaptation 3:848–849
  in Africa 1:294–295
  agriculture 3:849–850
  in Arctic 1:927–928, 1:931
  coastal areas and settlements 3:850–851
  and development 3:851
  disasters 3:851, 10:550–551
  and disasters 10:548–551
  and ecofeminism 6:849
  economic development 3:848, 3:850–851
  energy anthropology and 7:623–624
  food security 3:849–850
  frames and sponsors 3:855t
  and global poverty 3:848
  global warming 3:848–851
  and green social work 10:385, 10:387–388, 10:390
  and health 3:850, 7:742–743, 7:817
  international climate negotiations 3:842–843
  investment themes 10:372
  LTG study 14:141–142
  and media 3:853–859
  media reporting of 3:841–842
  at national level 3:843
  natural ecosystems and biodiversity 3:851
  nongovernment organization campaigns 3:840–841
  and personal behavior 3:844–846
  policies 10:377
  poverty alleviation 3:848, 3:851
  as relevant media issue worldwide 3:856
  and rural social work 20:808
  scientific reporting 3:839–840
  sustainable development 23:819–820
  sustainable energy 3:848
  and sustainable transportation 23:846
  transnational governance 3:843–844
  and urbanization 24:963
  and war 25:368
  water 3:848–849
  and water scarcity 25:445–446
Climate change, economics of 3:860
  adaptation 3:860, 3:863
  carbon, pricing 3:864
  caveats and open questions 3:862–864
  economic benefits 3:860
  economic costs 3:860–861
  economic damages, estimates of 3:860–861
  Intergovernmental Panel on Climate Change (IPCC) 3:860
  iterative risk management 3:863
  mitigation 3:861–862
  mitigation targets, framing 3:863–864
  risk aversion 3:863
  scenarios 3:860
  social cost of carbon (SCC) 3:860–861, 3:863, 3:863f
  Stern's review 3:860
  time preference 3:860–861
  United Nations Framework Convention on Climate Change (UNFCCC) 3:861
  'WRE' results 3:861–862
Climate Change Act (UK) 7:781
Climate history and human history 3:865
  climatic variability 3:867–869
    in Africa 3:868–869
    in China 3:867
    in Europe and the Mediterranean 3:867–868
    in Japan 3:868
    in South America 3:868
  documentary sources 3:868
  global climatic changes and forcing factors 3:865
  historical climatology 3:865–867
  impacts of extreme climate events on societies 3:870f
  weather, climate and history 3:869–871
Climate journalists 3:855–856
  work 3:855–856
Climate policy, and renewable energy 18:372
Climate politics 3:853, 3:856
Climate skepticism 7:793–794
Climate-neutral city *see* Eco-cities
Climatic determinism 3:866, 3:869
Climatic trends 3:868
Climatology
  Humboldt on 25:279–282
Clinical abortion 2:698
Clinical activities, for SES impact on childhood behavior problems 2:479–480
Clinical applications of stress inoculation training 8:608
Clinical Assessment of Skills and Competencies (CASC) 7:191
Clinical care, informed consent in 19:323–324
Clinical empathy 7:545–546
Clinical equipoise 18:166, 18:168, 20:519
Clinical ethics 2:613
Clinical Evaluation of Language Fundamentals 3:795
Clinical Evaluation of Language Fundamentals - Fourth Edition 16:725
Clinical genetics, issues in 2:619–620

Clinical Improvement Amendments 1988 (CLIA) 6:441
Clinical information, in combined pharmacotherapy and psychotherapy 19:496
Clinical interpretation, electroencephalography 7:357–358
Clinical intervention, motivation 15:907–913
Clinical interviews 6:342–343
Clinical issues, of antidepressants 1:781
Clinical judgment 6:344
Clinical management
  autism spectrum disorders 2:278–279
    diagnosis and assessment 2:278
    treatment 2:278–279
Clinical paradigm, to leadership in organizations 13:666–667
Clinical perspective, of neonatal crying 16:477
Clinical practice guidelines (CPG) 2:418
  behavior therapy 2:418
  psychological treatments 19:382–384
Clinical practices
  application of evidence-based medicine in 8:329–330
  ethics in 2:615–616
  proper names in 19:151
  of new genetics 9:1016–1018
    governance 9:1017–1018
Clinical pragmatics *see* Pragmatic disorders
Clinical Problem Solving Test (CPS), in psychiatric education 7:188
Clinical problems
  ABA for 2:410
  fear-potentiated startle for studying 8:864–865
Clinical psychology 1:854
  development 3:874
  early studies 3:874
  electrophysiological methods in 11:603
  field of study 3:873–874
  future trends, Europe 3:875–876
  of group psychotherapy 10:428–433
  Heider on 10:753
  history of
    in Canada 3:881
    in Europe 3:873–876
    in Mexico 3:881
    in North America 3:877–883
    post-World War II 3:880
  and homosexuality 3:884
  imaging methods in 11:603–611
  models 3:874
  neuroimaging techniques in 11:603–604
  organization, Europe 3:875
  personality assessment 17:828–830
    affective-motivational personality constructs and psychopathology 17:829
    dispositions of self-regulation 17:829–830
    personality disorders 17:829
  perspective of procrastination 19:76–77
  professionalization, Europe 3:875
  research 3:882
  scientist-practitioner model 3:874
  training in United States 24:517–527
Clinical psychology training
  in Europe 3:874–875
  sex differences 3:875
Clinical questions, formulation 8:328
Clinical research
  linking mental illness epidemiology to 15:183–184
  rationale for using placebo-controlled trials in 18:165–166
Clinical responsibility, in combined pharmacotherapy and psychotherapy 19:497
Clinical scientist model 24:517
Clinical significance 8:348
Clinical skills
  of forensic psychiatrist 9:349
  of medical students, assessment 7:190–191, 7:191f
    long case 7:190
    modified long case 7:190
    Objective Structured Clinical Examination 7:190–191
    workplace-based assessments 7:191
Clinical social work, and psychoanalysis 22:767
Clinical sociology 1:858–859
Clinical symptoms, personal intelligence 17:754–755
Clinical transport psychology 24:623–624
Clinical trial designs 19:387
  adaptive design 19:387
  stepped-care design 19:387
Clinical trials
  archiving 1:918–919
  biomedicine 2:665
  exclusion of women in 9:719–720
    political economy 9:721
    WCP and pregnant women 9:720
  placebo-controlled 18:164–173
  placebo-controlled double-blind studies 18:157, 18:159–160
  psychological treatments 19:384
  purposive sampling in 9:842
Clinical utility 9:927
  and health care integration 6:441–442
Clinician researchers, and personality psychology 17:911
Clinician–patient interaction 15:103
Clinicians, evidence-based practice 8:332–337
Clinicoanatomical method 16:682
Clinics, lesbian 15:112–114
Clinton, Bill 1:605–606, 12:760, 18:736
Cliometrica 7:58
Cliometrics 1:708–710, 7:51–53, 7:56, 19:696
Clique percolation method (CPM) 3:910
Cliques
  network analysis 16:519, 16:521
  in sociometry 22:455–456
CLIs *see* Cross-linguistic influences (CLIs)
Cliticization 23:890
Clitics 11:827
Clitorectomy 10:195
Clitoridectomy 8:878–880
Clock time 24:373–376
Clomipramine 1:778–780
  for obsessive–compulsive disorder 17:115, 17:125
Cloned plants of *Achillea*, growth of 10:44f
Clones 24:765
  in leatherman subculture 13:887
Clonidine 18:705
Cloning 9:1016
  stem cell research and 23:440–445
Close relationships 21:490, 22:563, 22:566
  abilities for 22:566–567
  commitment 22:566
  communal strength 22:563–564
  definition 22:563
  formation of 22:564–565
  fortitudes for 22:566–567
  high-quality close relationships 22:564
  initiation 22:565–566
  and interpersonal processes 22:564
  maintaining 22:565–566
Closed autopoietic systems 14:391
Closed circuit television *see* CCTV
Closed corporate community, peasants 17:677
Closed innovation 12:147
Closed list proportional representation 19:205, 19:207–209
Closed (authoritarian) polities 1:557
Closed population
  hypothesis of 6:149–150
  microsimulation models 15:344
Closed questions, survey 23:767
Closed systems
  organizations as 3:893
    adaptation 3:894–895
    design 3:895
    general systems theory 3:893–894
Closed testing procedures 16:84
Closed-loop Nash equilibrium 17:255
Closeness 17:563
  network centrality 16:534–539
Closeness centrality 16:534, 16:536, 22:462
Closet Era 9:632
Closing time effect 12:628
Closure 17:484, 22:479–480, 24:137
  social capital 22:182–184
Closure positive shift (CPS) 9:483, 9:487f
Closure theory 15:96
Clothing, of butch women 13:869
Cloud computing 4:493
Clozapine 19:201
  for schizophrenia 21:18–19
CLRP *see* Civil Litigation Research Project (CLRP)
CLRTAP *see* Convention on Long-Range Transboundary Air Pollution (CLRTAP)
CLS *see* Critical legal studies (CLS)
CLSA *see* Canadian Longitudinal Study on Aging (CLSA)

CLT *see* Communicative language teaching (CLT)
Club economy 9:615
Club of Rome
  model for global development 23:794–795
  report 23:794–795
Clubhouse Communities 10:689
Clubs, of bear subculture 13:891
Cluster analysis 16:116, 19:434, 19:731, 23:392, 24:825–826
  of classification 3:808–809
Cluster grouping, statistics 20:876
Cluster indicators 13:388
Cluster randomized trials (CRTs) 3:897–905, 24:282–283
  characteristics 3:897–898, 3:898t
  COMMIT study 3:900
  ethical implications of 3:899
  examples 3:899–900
    cluster–cluster trial 3:900
    individual–cluster trial 3:900
    professional–cluster trial 3:900
  general principles 3:897
  methodological implications 3:898–899
  Ottawa Statement 3:900–903, 3:901t
    feasibility of consent 3:902–903
    identification of research participants 3:901–902
    informed consent in organizations 3:903
    informed consent requirements 3:902
    justification for design 3:900
    need for research ethics review 3:900–901
    role of cluster gatekeepers 3:903
    timing of consent 3:902
  reasons for adopting cluster randomization 3:899
  research ethics committees 3:897, 3:900–904
Cluster sampling 13:385–386, 17:103, 23:187, 23:189
Cluster–cluster trial 3:900
Clustered data 19:734
  logistic regression for 2:594
  modeling 14:163–164
Clustered development 10:449
Clustered random assignment 9:128
Clustering 1:535, 16:4, 19:433–434, 21:411–412
  approximation methods 19:434
  blockmodeling of social network data 19:433
  clustering nonmetric data 19:433
  coefficient variance estimates adjusted for 23:382–383
  distance-based clustering of metric data 19:433
  exact methods 19:433–434
  metaheuristics 19:434
  structural equivalence 23:622–623
Clustering algorithms 23:623
Clustering coefficient 22:456–457
Clustering in social networks 3:906–912
  relative density measures 3:908–910
Clusters 7:63–64, 20:147, 20:149–150, 23:210–212
  in Asia 11:832
  in entrepreneurship 7:687–690
  evolution of 11:830–831
  geographical proximity and 11:828–830
  in Silicon Valley 11:828
  in time 11:831
  in varieties of capitalism 11:831–832
Clusters of organizations 12:623–625
  formation 12:624
  methods and approaches 12:624
  outcomes of local clusters 12:624
Cluttering (speech) 9:272
Clyde Valley Plan of 1946 20:131
CM *see* Case management (CM); Conventionality of meaning (CM)
CMA *see* Cost-minimization analysis (CMA)
CMC *see* Computer-mediated communication (CMC)
CME *see* Cardiac Modulation of Startle (CME) effect; Continuing medical education (CME); Coordinated market economies (CME)
CMHTs *see* Community Mental Health Teams (CMHTs)
CML *see* Chronic myelogenous leukemia (CML); Conditional maximum likelihood (CML) method
CMM *see* Congeneric measurements model (CMM)
CMOs *see* Collateralized mortgage obligations (CMOs)
CMR *see* Cardiometabolic risk (CMR)
CMRR *see* Common-mode rejection ratio (CMRR)
CMS *see* Centers for Medicare and Medicaid Service (CMS); Children's Memory Scale (CMS); Compact Muon Solenoid (CMS); Course Management System (CMS)
CMSA *see* Consolidated metropolitan statistical area (CMSA)
CMV *see* Concepts; method; and values (CMV)
CNA *see* Cognitive neoassociation theory (CNA)
CND *see* Campaign for Nuclear Disarmament (CND)
CNNI (Cable News Network International) 10:226
CNPC *see* Complex NP Constraint (CNPC)
CNPQ *see* Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPQ)
CNRC *see* Council of Title VI National Resource Center Directors (CNRC)
CNS *see* Central nervous system (CNS)
CNU *see* Congress for the New Urbanism (CNU)
CNVs *see* Copy number variations (CNVs)
Coaches, and youth physical activity motivation 15:917–918
Coaching, apprenticeship 1:864
Coaching Boys into Men (CBIM) program 12:681
Co-activation of elements 21:350
Coactive social flow 9:264
Coal power, subsidies for 18:367–368
Coal production, in Britain 15:555f
Coale, A.J.
  'I' indices 11:445–446
  'M' and 'm' indices 11:446
Coale–McNeil parameterization 6:134
Coalescent theory 11:292
Coalition Against Trafficking in Women 17:432
Coalition building, for physical activity 18:95
Coalition for Community Schools 8:613
Coalition for the Advancement of Foreign Languages and International Studies (CAFLIS) 1:947
Coalition governments
  and green parties 10:379–381, 10:383
  proportional representation 19:207–209
Coalition theories 1:556–557
Coalitional monotonicity 4:878
Coalitional strategic equivalence 4:873
Coalition-centric mapping 4:868
Coalitions 4:867, 9:579
  executive government 8:498–500
  and factionalism 8:683–685
  presidentialism 18:865–868
Coal-powered steam engines 7:617
Co-articulation (vocabulary) 25:226
Coase, R.H. 19:170
  on welfare economics 25:501–502
Coase theorem 5:253, 7:727, 13:455, 13:458, 20:866–867
Coast, West Africa 1:301–302
Coastal areas and settlements, and climate change 3:850–851
Coastal flooding 6:480
Coastal zones, in Africa 1:296–298
Coauthor model of network formation 16:575
Coauthors 7:868–870
  percentage 7:869t
Coauthorship networks 19:910–911
Cobb–Douglas production function 21:334
Coburg government style 17:511
Coburn amendment 22:664
Cobweb model 8:511f
Coca-colonization 6:404
Cocaine 6:669, 6:674–675, 19:463
  self-administration of 20:215
COCEI *see* Confederation of Workers, Peasants, and Students of the Isthmus (COCEI)
Cochinchina 25:95
Cochlear implant (CI) 16:714–715
  calibration sessions 16:714–715
  commercial 16:715
  components 16:714
  location 16:714f
  single-channel 16:715
Cochrane, Archibald 8:329–330
  works, *Effectiveness and Efficiency – Random Reflections on Health Services* 8:327
Cochrane Collaboration 8:327, 8:329–330, 8:544, 15:269

Cochrane Collaboration (*continued*)
logo of 8:330f
Cochrane Library 8:330
Cochrane's Q-test 15:268–269
Cochran–Mantel–Haenszel test 16:132
Cock, Jacklyn
on GASA 21:745
on gay rights movement in South Africa 21:745–746
Cocktail party problem 2:227
Cocktail–Party effect 3:342
Coconstructive perspective of aging 1:437
Coconstructivism in education 3:913–917
concept and process of coconstruction 3:913–915
context of linguistic discourse analysis 3:914–915
cooperative learning 3:914
learning communities 3:913
pedagogical concepts 3:915–917
collaborative learning and patterns of dialogues 3:916–917
instructional conversations and dialogic teaching 3:915–916
tutoring and scaffolding 3:915
peer interaction 3:913–914
perspective of situated and socially shared cognition 3:914
Piagetian perspectives, role of sociocognitive conflict 3:914
sociocultural theory 3:913–915, 3:917
Vygotskyan perspectives 3:914
zone of proximal development 3:914
Cocoyoc Declaration 7:770–771
Cocreated citizen science projects 3:632
Co-creation 6:585–588
Code, Chris 16:679
Code Civil security 21:403
Code d'instruction criminelle of 1808 5:215
Code mixing 16:60
Code of conduct 13:622–623
law as 17:400–401
Code of Federal Regulations (CFR), human subjects protection in research 8:50
Code of honor 14:432
Code of King Hammurabi 5:874–875
Code of Nuremberg 20:525–526
Code of Practice for Social Service Employers 20:584
Code of Practice for Social Service Workers 20:584–585
Code of professional conduct 18:194–196
Code of Professional Responsibility 13:620
Coded exclusion 23:731
Coded meaning 18:798
Codes of Civil and Criminal Procedure 23:61
Codes of conduct 10:255, 10:257, 17:13, 20:514
Codes of Corporate Social Responsibility (CSR) 4:736
Codes of ethics 8:40, 18:195–196
business codes 8:23
comparison of professional ethics 8:40–41
future research 8:41
in SBS 8:53–58, 8:54t
social work 22:782–783
Codes of ethics, psychiatry 8:65–66
boundary between ethics and law 8:65
confidentiality 8:65–66
double agent problem 8:66
false memory controversy 8:66
fees, billing, and reimbursement 8:66
informed consent 8:66
sexual misconduct and boundary violations 8:65
Code/space 12:93
CODESRIA *see* Council for the Development of Social Science Research in Africa (CODESRIA)
Code-switching (CS) 2:584–585, 3:918–925, 16:60, 16:66–67, 22:951
cherry-picking *versus* accountability 3:921
evaluating 3:921–923
versus borrowing 3:921
future directions 3:923–924
grammar of 3:919–920
and markedness 14:535–536
pre-eminence of counterexample 3:921
'rule-and-exception' paradigm 3:921
theory of 3:920
data for 3:921
variability, probability, and quantitative reasoning 3:920
Co-determination legislation 21:188
Codex Alimentarius 1:474
Codification 2:527–528, 13:567
of research ethics 21:315–317
in sub-Saharan Africa 23:655
Codification Projects 23:61
Codified knowledge 13:92–93
Coding 10:405, 22:545
of spatial distinctions in language forms 23:92
systematic review 23:908
CODIS *see* Combined DNA Index System (CODIS)
Coedès, George, works, *Histoire Ancienne des États Hindouisés d'Extrême-Orient* 11:86
Coeducation 3:926–932
Coeducation schooling 3:926
on academic- and career-related outcomes 3:927–929
on affective and social development 3:929–930
merits of 3:930t
prevalence of 3:926–927
research findings analysis 3:930–931
research on effects 3:927
Coefficient alpha
in classical test theory 19:427–429
Coefficient of generalizability 9:837
Coefficient of precision 20:241–242
Coefficient of relatedness 13:44–45
Coefficient of variation 11:900, 11:904, 11:906, 11:886t
Coefficient omega, in classical test theory 19:429
Coefficient variance estimates adjusted for clustering 23:382–383
Coefficients
latent variables 13:394–395
logistic regression, interpretation of 2:589–590
Coerced migration 11:649
Coercion 18:760
informed consent 12:112
and police use of force 18:258
in prison life 18:925, 18:927
and violence 25:117
Coercion model 2:472–473
Coercion–extraction cycle 23:362–363
Coercive Acts 1:610–611
Coercive control, definition 2:472
Coercive isomorphism 13:545–546, 17:391, 21:73, 22:981
Coercive language 18:614
Coercive power 17:68
and social exchange 8:485
Coercive violence 10:3–4
Coevolution 8:362, 9:567
Co-expressive speech and gesture 10:109–110
Cogito ergo sum 7:473
Cognate 21:350
definition of 4:424
evolution of language 8:366–367
Cognate set, definition of 4:424
Cognition 1:536–537, 2:349, 2:351, 2:876, 3:819, 4:542–543, 9:672, 12:285–286, 16:330–331, 17:766, 21:617, 21:848
and aggression 1:379
and action, in pragmatism 18:809–810
of animals 13:125
and anthropological comparison 4:413
approaches for religion 20:311
artificial intelligence 2:35, 2:39
brain–cognition–personality relations, aging of 2:796
and classical conditioning 3:768–769
cold and hot 7:416–417
and collective memory 4:195
in collectivism/individualism 4:208
comparative evolutionary psychology 4:395–396
components 5:113–115
computer modeling approach 4:103–104
and conflict 4:583, 4:594–597
and culture 5:576–581
beyond classical theory 5:576
classical theory of 5:577–578
problem of internalization 5:578–579
problem of meaning 5:579–580
defining intelligence 1:537
deliberative and automatic 7:416–418
of disputant 16:428
effects of media on children 15:35–37
comprehension 15:35–36
educational media use 15:36
infant development 15:35
relationships with academic outcomes 15:36–37

and electroencephalography 7:362
embodied social *see* Embodied social cognition
emotion in 7:461–466
environmental *see* Environmental cognition
experimental economics 8:550
load 8:550
and health 10:584
'historical' change 25:323
and identity movements 11:543
implicit social cognition 11:714
individual *see* Individual cognition
influences of emotion regulation on 7:454
information-processing approaches and communication research 4:103
and intellectual functioning 12:250, 12:252
knowledge spaces 4:6–9
and language 23:89, 23:92
language acquisition 1:536
macrostructure of 4:96–98
evidence of interactive processes in brain 4:97–98
modular *versus* interactive approaches to organization of function 4:97
specialization of brain regions 4:96–97
managerial and organizational 14:479–483
measuring intelligence 1:537
mechanisms underlying confabulation 4:566
associated features 4:566
constitutional processes 4:566
core processes 4:566
microstructure of 4:95–96
distributed representations 4:96
knowledge and learning in strengths of connections 4:96
lesion and behavior approaches 4:98
patterns of activity arising in ensembles of simple elements 4:95–96
modern human cognition 11:252–258
and motivation 2:326–327
of negotiator 16:426
neuroimaging 14:439–441, 14:443
operational mode of cerebellum in 3:329–330
outcome in transient global amnesia 24:564
and panic disorder 17:475
and perceptual learning 10:133
phenomenology 18:20–22
and play 18:214
in political belief system 11:557
and procrastination 19:76–77
and prosocial responding 7:318–319
public and private 15:546–547
and ritual 20:719
sequential decision making 1:536–537
social *see* Social cognition
and social beliefs 22:163
and stone tool fabrication 14:263
symbolic boundaries 23:850
and values 25:36–37
without control 18:819
Cognition, evolution of 4:1–5
behavior 4:1–2
case study 4:4
domains 4:2
evolutionary psychology 4:1, 4:4
genetics 4:3–4
neuroscience 4:2
psychological adaptations 4:1–2
psychological adaptations, methods for investigating 4:2–4
cross-cultural studies 4:3
cross-species comparisons 4:3
developmental studies 4:3
genes, fossils, and other sources of evidence 4:3–4
laboratory and naturalistic experiments 4:2
neuropsychological dissociations 4:2–3
Cognition building, action-oriented 1:106–107
Cognition in practice 19:361–362
Cognitive abilities 12:297–298, 15:372
and bounded rationality 2:437
explicit training in 7:141–142
general factor (*g*) 12:297–298
genetics of 12:297–302
and health 22:277
and income 7:132, 7:135
molecular genetic studies 12:300–302
candidate gene studies 12:301
genome-wide association studies (GWAS) 12:301
genome-wide complex trait analysis (GCTA) 12:301–302
personnel selection 17:937
quantitative behavioral genetic research 12:298–300
genetic influence on *g* across range of socioeconomic classes 12:299–300
genetic influence on *g* increases during development 12:299
Cognitive access 4:675–676
Cognitive aging 1:428–429, 4:10–15, 6:380–381, 1:429f
adult life development in 1:401–403
age-cohort-period model 4:10–11
data collection strategies 4:11
sequential strategies 4:11
age-comparative *versus* age change designs 4:10
cognitive training 4:13–14
compensation 4:12–13
definition 4:10
descriptive study of adult intellectual development 4:10
education and learning 7:140–141
experimental study findings 4:11–12
language 4:10, 4:12
memory 4:10, 4:12
findings from descriptive study on age changes in intellectual competence 4:12–13, 4:13f
methodological issues 4:10–11
role of response speed 4:11
slowing and reversing 4:12–14, 4:14f
source monitoring 23:28
Cognitive algebra 19:475
Cognitive and mental ability tests 17:853–854
Cognitive anthropology 1:748, 4:16–22, 4:28, 5:584–585, 7:145, 19:314, 19:360–362
cognition in practice/distributed cognition 19:361–362
cultural consensus/cultural domain analysis 19:361
cultural epidemiology 19:362
cultural models 19:360–361
future research 4:19–21
knowledge and knowing 4:20
language and cognition 4:20
universal/cultural variable 4:20–21
history (from ethnoscience to cognitive sciences) 4:17–18
revised premises 4:18
three premises 4:17–18
representation of knowledge 4:18–19
cultural model 4:19
mental images 4:19
paradigm 4:18–19, 4:19t
prototype 4:19
schema 4:19
script 4:19
taxonomy 4:18
and worldview 25:753–756
Cognitive appraisal 7:483, 23:566–567
Cognitive apprenticeship 1:863–864
and coconstructivism 3:916–917
instructional psychology 12:204
in school learning 1:871
Cognitive approaches
cardiac defense 23:540
to persuasion 17:943
to political decision-making 18:429–430
questionnaires 19:771–775
Cognitive archaeology 4:23–29
advancements in 4:28
associative approach 4:26–27
conditional approach 4:27
covering law model 4:24
current thinking about 4:26
direct historical approach 4:26
material engagement theory (MET) 4:26, 4:28
materiality approach 4:27–28
philosophical commitment 4:23–24
structural archaeological approach 4:27
trends 4:24–28
Cognitive aspects
of elicitation 7:382–383
of survey design 20:884
Cognitive Assessment System–2nd edition 12:287
Cognitive assessments 16:723–724
Cognitive associative learning, role of cerebellum in 2:97
Cognitive authority 5:577–578
Cognitive behavioral analysis system of psychotherapy 4:544
Cognitive Behavioral Intervention for Trauma in Schools (CBITS) 23:573–574

Cognitive behavioral model of pain 3:557–558
treatment for chronic pain 3:559–560, 3:559t
Cognitive behavioral stress management (CBSM) 3:552, 19:448, 23:553
Cognitive behavioral therapy (CBT) 1:502, 2:419–423, 2:431–434, 4:30–32, 4:542, 6:827, 8:748, 11:612, 11:615–616, 11:618–619, 14:501, 15:190, 15:194, 17:470, 17:594–595, 19:347, 19:460, 20:352, 24:288, 25:472, 11:616f
for anorexia nervosa 19:376
antidepressants 4:31
for anxiety disorders 25:138
applications 4:31–32
for asthma 20:578
for binge eating disorder 19:376–377
for bulimia nervosa 19:376
for childhood cancer 3:462
for chronic obstructive pulmonary disease 20:578
components of 4:545–546
assignment of homework 4:545–546
behavioral strategies 4:545
collaborative empiricism 4:545
psychoeducation 4:545
thought monitoring 4:545
criticisms 4:31–32
for depression 4:30–31, 19:374
emotions 4:30–31
in gastrointestinal disorders 9:609–610
for generalized anxiety disorder 19:374
groups 10:429
history and development 4:30
menstrual cycle disorders 10:463
for Obsessive–compulsive disorder 17:113, 17:115–117
for occupational stress management 17:149–150
for panic disorder 17:470–471, 19:374
in panic disorder 17:477
for perinatal mental illness 17:724
for personality disorders 17:896
in psychoneuroimmunology 19:448
for respiratory disorders 20:578–579
rheumatoid arthritis 20:655–656
for somatoform disorders 23:13
for weight reduction 25:472
Cognitive behaviorism 4:124
Cognitive biases 22:496
Cognitive capabilities 4:23–25, 4:27
Cognitive capitalism 13:106
Cognitive changes of acculturation 1:70
Cognitive classifications systems 3:803
Cognitive components, test anxiety 24:244
Cognitive conflict 21:271–272
Cognitive consistency models 22:346
of opinion formation 17:244
Cognitive consonance 4:76
Cognitive constraints, and bounded rationality 2:773–775
Cognitive controls
cognitive styles and learning styles 4:111
and hypnosis 11:467–470
and implicit social cognition 11:716–717
risk-taking 17:811
role of prefrontal cortex in 4:33
computational models 4:34–35
guided activation 4:35–36
outstanding questions 4:36
Stroop model 4:35–36, 4:35f
and selection 13:372
Cognitive cultural intelligence 5:433, 5:436
Cognitive deficits 6:76–78
Alzheimer's disease 1:585
Cognitive demands, contra-hedonic motivation 1:243–244
Cognitive development 2:847, 4:44, 12:327, 18:128–131, 19:71–72, 21:120–121, 21:125, 21:278, 18:130t
Bayesian statistics 4:38
Bruner's theory of 2:880
and child care 3:391–393
child education 4:38–42
exploration and experimentation 4:39
of children 22:327, 22:942
children's learning 4:41
and developmental psychology 5:354–355, 5:359
domain-general developmental mechanisms for learning 4:40–41
pretense 4:40–41
school readiness and success 4:40
self-regulation and executive function 4:40
domain-specific developmental mechanisms for learning 4:38–40
conceptual change 4:39–40
testimony 4:39
theory and probabilistic learning 4:38
theory of mind 4:39
theory theory and probabilistic learning 4:38
educational theories of 4:51–57
accommodation 4:52–53
assimilation 4:52–53
early development 4:51–52
emergence of cognitivism 4:51
equilibration 4:52–53, 4:55–56
implications 4:55–56
information processing 4:51
mechanism 4:52–53, 4:53f
Piaget on 4:52–54, 4:53f
role of culture 4:55
scaffolding 4:54–55
schema 4:53
social interaction 4:54–56
stages 4:53–54, 4:54t
Vygotsky on 4:54–55
zone of proximal development 4:54
effect of breastfeeding on 2:847–851
evolution and 8:394
and imitation by infants/young children 11:621–622
impact of schooling on 21:119
in infancy, neural mechanisms 4:58–65
and intelligence 12:319
and intentionality 12:347–348
lifespan, dual-process theory of 12:294
and mathematics learning 4:66–75
neo-Piagetian theories of 16:501–510
observation 4:41–42
overlapping waves theory 4:41
Piaget's theory of 18:120–124
play 4:38, 4:40–41
prenatal/infant 18:852–853
and prosocial behavior 19:229
role of pretend play in
see also Pretend play
self-development 21:493
and socioeconomic status 22:926–927
teaching and learning 4:39, 4:41–42
and temporal and object constitution 18:809
Williams syndrome 25:581–582
Cognitive developmental theories 15:788–789, 21:427–430, 21:680, 25:51
aging 21:430
competition and cooperation 4:856
crystallization 21:430
emotional transformation adulthood and later life 21:428–429
emotional transformation from childhood to adolescence 21:427–428
empirical research 15:788–789
vulnerability of later life regulation capacities 21:429–430
Cognitive dissonance 4:76–78, 22:528–530
aversion 4:76–78
cognitive consonance 4:76
Festinger on 4:76–78
foundations of dissonance theory 4:76
induced compliance effect 4:76
introduction of self 4:77
motivational states 4:76–78
new look model 4:77–78
and risk assessment 20:695
self-affirmation theory 4:77
self-consistency 4:77
self-perception theory 4:76–77
self-standards model of 4:78
and unconscious mental activity 24:723
Cognitive distortions 4:542
Cognitive dysfunctions
and cancer 3:73–74
in schizophrenia 21:13
Cognitive effects 19:465–466
on representations of smell, taste, and flavor 24:29–30
Cognitive elaboration strategies 13:714
Cognitive empathy 7:549, 17:776–777, 19:216–217, 19:219
Cognitive engagement 1:31
for direct voting 25:297
Cognitive enhancement 16:665–666
neuropsychological rehabilitation for 2:817
Cognitive epidemiology 22:277
education and learning 7:139
Cognitive epigenetics 22:283
Cognitive ergonomics 11:299–301

Cognitive errors 3:559t
Cognitive evaluation theory (CET) 17:612–613
Cognitive evolution 4:23, 4:25, 4:27–28
Cognitive explanations of grammar 9:470–471
Cognitive factors, and pain 17:453–454
Cognitive failure 12:131
Cognitive field 4:51–52
Cognitive film theory 9:162–163
Cognitive flexibility 8:489, 8:491–494
  and dysexecutive syndromes 6:726
Cognitive function
  development of 18:817–823
Cognitive functioning
  and aging 1:399–401
  and differential aging 6:381
  effects of alcohol on 1:508
Cognitive heuristics
  for persuasion 17:944
  taxonomies of 10:834
Cognitive immaturity 8:422
Cognitive impairment 23:559
  and behavior during PTA 18:638
Cognitive independence 17:97
Cognitive indicators 22:333–334
Cognitive interdependency, groups 10:434–438
Cognitive interviews
  eyewitness testimony 8:645–646
  in laboratory settings 20:884
  questionnaires 19:775
  scale development 7:932
Cognitive labeling, and aggression 1:374–375
Cognitive learning
  outcomes, relation of epistemological beliefs to 7:888
  strategies 13:712–713
  theories 7:240–241
Cognitive linguistics 9:473, 14:169
Cognitive loading 12:199, 22:188
  role in situational judgment tests 22:17
Cognitive mapping
  parahippocampal place area (PPA) 23:166–167
  retrosplenial cortex (RSC) 23:166–167
Cognitive maps 4:79–83, 23:112–113, 23:164–166, 4:81f
  brain systems mediating 23:166–167, 23:167f
  Lynch's classification 4:80t
  mapping 4:79
  naturalistic settings 4:82
  neurological approaches 4:82
  qualitative approaches 4:82
  virtual environments 4:82–83
Cognitive mechanics 4:12
  differential age-gradients of 1:430, 1:430f
Cognitive mechanisms 13:747
  evolutionary perspective 8:395
Cognitive misbehavior 5:203
Cognitive modeling, connectionist approaches 4:84–89
  adaptation and learning, mechanisms of 4:85–86
    backpropagation 4:86
    correlational learning 4:85–86
    delta rule 4:86
    reinforcement learning 4:86
  definitions 4:84
  example, learning past tenses 4:86–87
  history 4:84–85
    broader 4:85
    narrower 4:84–85
  issues
    computability 4:87–88
    connectionist representations 4:87
    dynamic systems approach 4:88–89
    relation to biological neural networks 4:88
    rules and constituent structure 4:87
    situated embodiment 4:88
Cognitive modeling, discrete-state models 6:524–529
Cognitive models
  computational neuroscience 4:482
  computational organization theory 4:485
  for obsessive-compulsive disorder 17:122–125
  of panic disorder/agoraphobia 17:476
  of spelling 1:453
Cognitive models, of personality
  information-processing models 17:873–875
    cognitive neuroscience perspectives 17:874–875
    cognitive-adaptive perspectives 17:875
  overview 17:870
  traits and self-beliefs 17:871–873
    cognitive content and self-schema 17:872
    cognitive process factors 17:872–873
  traits as cognitive constructs 17:871
Cognitive models of personality 17:870–875
Cognitive neoassociation theory (CNA), for aggression 1:375
Cognitive neuroprostheses 16:718
Cognitive neuropsychological model
  of spoken picture naming 3:793f
  written language impairments 1:451–454
    central dysgraphia 1:453–454
    peripheral dysgraphia 1:454
    processing models 1:452–453
Cognitive neuropsychological signs 6:76
Cognitive neuropsychological studies
  acquired and developmental 4:93
  approach to intervention 4:93
  associations 4:90–91
  case series 4:91–92
  computational 4:92
  dissociations 4:90–91
  individual difference 4:91
  single case studies 4:91
  theoretical impact of treatment studies 4:92–93
Cognitive neuropsychologists 4:90
Cognitive neuropsychology 4:98
  methods of 4:90–94
Cognitive neuroscience 4:95–102, 4:673, 11:711–712, 12:341–342, 13:461, 13:489, 16:680, 17:874–875, 22:485
  of art 16:656, 16:656f
  attention, neural basis of 2:163
  and brain aging 2:797
  and cognitive psychology 4:107–108
  future of 4:101
  macrostructure of cognition in brain 4:96–98
    evidence of interactive processes in brain 4:97–98
    modular *versus* interactive approaches to organization of function 4:97
    specialization of brain regions 4:96–97
  methods and approaches in 4:98–100
    behavioral neuroscience 4:98
    cognitive neuropsychology 4:98
    computational and mathematical modeling approaches 4:99–100
    functional and structural brain imaging 4:99
    lesion and behavior approaches 4:98
    neuronal recording studies 4:98–99
    transcranial magnetic stimulation 4:98
  microstructure of cognition 4:95–96
    distributed representations 4:96
    knowledge and learning in strengths of connections 4:96
    patterns of activity arising in ensembles of simple elements 4:95–96
  modularity 4:97
  neuroesthetics 16:656–663
  neuroimaging in 4:107–108
  neurophysiology 4:95, 4:99
  neuropsychology 4:95, 4:98–99
  open issues 4:100
    conscious experience 4:100
    learning process of brain 4:100
    unique cognitive capacities of human brain 4:100
  synapses 4:95–96, 4:100–101
  system-level organization 4:96–98
Cognitive overload, and attitude 2:171
Cognitive performance, in aging population 1:403–404
Cognitive performance rhythms 3:571
  attention tasks 3:571
  executive functions tasks 3:571
  memory tasks 3:571
Cognitive perspective of self 19:401–402
Cognitive phenomena 11:461
  partial amnesia 11:461
Cognitive plasticity, and brain aging 2:796–797
Cognitive polyphasia 20:445–446
Cognitive pragmatics 4:300
  differential age-gradients of 1:430, 1:430f
Cognitive praxis 21:177, 21:179
Cognitive predictors, of educational outcomes 17:787
Cognitive preferences, personality-religion links 17:803

Cognitive problems, and societal impact assessment (SIA) 22:874
Cognitive procedural skill learning, neural substrates of 16:848
Cognitive procedures, in ABA 2:410
Cognitive process
interhemispheric interaction in 12:416–424
Cognitive processes 20:50
and behavior
explanatory style 8:598–599
cognitive styles and learning styles 4:111
individualism 15:305–306
linking openness/intellect with creativity 17:771–772
decreased latent inhibition 17:771
divergent thinking 17:771
implicit learning 17:771
working memory 17:771–772
of motivation and executive function 21:554
in self-regulation 21:554–555
speech production 23:255–258
Cognitive processing therapy (CPT) 14:501
for mass trauma 14:749
for neuroticism 16:745
for posttraumatic stress disorder 19:374–375
Cognitive psychologists 12:672–673
Cognitive psychology 5:77, 12:197, 13:704, 17:314, 19:415–416, 20:881–882
and behaviorism 4:103
and brain function 4:107
Broadbent on 2:857
and cognitive neuroscience 4:107–108
and connectionism 4:107–108
generative linguistics approach 4:104
growth of 4:104–105
Heider on 10:753–754
history 4:103–109
and human–computer interaction 11:400–401
and learning 13:706–707
Marr on 14:597–598
memory models in 4:104–105
memory-for-language models 4:106–107
and neurobiology 4:107
pattern recognition in 4:104
and social cognition 22:204–205
Whorf on 25:575
Cognitive psychotherapy, for schizophrenia 21:12
Cognitive rationality 15:303
Cognitive reactivity, and mindfulness 15:550–551
Cognitive reappraisal, and emotion regulation 7:453
Cognitive rehabilitation, for hemispatial neglect 10:769
Cognitive relativism 20:227–232, 20:234–235, 20:792–793
about justification and warrant 20:235
about logic 20:235
about truth 20:234–235
epistemic relativism 20:235
Cognitive remediation therapy, for schizophrenia 21:12
Cognitive representation component, future orientation 9:562
Cognitive reserve 2:795, 2:797
education and learning 7:140
Cognitive restructuring (CR) 4:543–544, 20:436
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
grief 2:540
for panic disorder/agoraphobia 17:478
Cognitive revolution 2:408, 2:444, 2:875–876, 2:884–885, 4:17, 4:103, 15:162, 17:352, 21:558
and computers 4:516
and life management 14:85
of psychology 5:563, 12:197
Cognitive robotics 13:100
Cognitive scale 14:241
Cognitive schemas 4:543
nationhood as 16:294–295
Cognitive science (CS) 2:880–881, 4:25, 4:518, 5:934, 12:341–342, 14:320, 16:330–332, 22:913
Bruner on 2:875–876
cultural analysis 9:361–362
mental representations in 4:106–107
and psychology 4:106–107, 19:415–416
and realism 20:25
rise of 4:106–107
Cognitive science of religion (CSR) 20:356–357, 20:389–393
Cognitive semantics 13:933–935, 21:586
basic levels and onomasiological salience 13:934–935
conceptual metaphor and metonymy 13:933–934
frame semantics 13:934
prototype effects and radial sets 13:933
Cognitive skills 12:847
and beauty 2:385–386
and education 7:150–151
and educational gender gaps 9:683–684
Cognitive socialization 5:552, 5:578–579
Cognitive stage theory 1:436–437
Cognitive studies, of language learning impairment 13:346
Cognitive styles 2:264–265, 5:564–565
of coping, behavioral medicine 2:467
and knowledge 13:70, 13:72
Cognitive styles and learning styles 4:110–117
cognitive controls and cognitive processes 4:111
differential psychology 4:110
implications of styles for lifelong learning 4:114
individual differences 4:110, 4:112, 4:114
intellectual styles 4:110–111, 4:114
mental imagery, memory and thinking 4:112
metacognition 4:114
perceptual differentiation 4:111
personal differences in approach and performance 4:112–113
models based on orientation to study 4:113
models based upon learning skills and neuroscience 4:113
models focusing on experiential learning 4:112–113
models focusing upon instructional preference 4:113
personality styles 4:111
in psychology of education 4:110
research into executive functioning and differences in thinking 4:112
style construct, revisionist theories and problematizing 4:113–114
style differences and individuality 4:110
Cognitive styles assessment (CSA) 4:112
Cognitive synesthesia 14:193
Cognitive systems engineering, and ergonomics 11:298, 11:300–301
Cognitive theories 21:848
and coping 4:896
of intelligence 12:304
and learning 13:707–708
of personality 4:895–896
source monitoring 23:28–29
Cognitive Theory of Multimedia Learning (CTLM) 7:248–249
Cognitive therapies (CT) 2:426, 2:429, 2:431–433, 24:288
classic cognitive therapy 4:542
collaborative empiricism 4:542
common methods in 4:545–546
Barlow's unified protocol 4:546
components of CBT 4:545–546
concepts and methods of 4:542–547
for depression 19:375–376
future research directions 4:546–547
key concepts 4:542–543
assumptions 4:543
automatic thoughts 4:542–543
cognitive distortions 4:542–543
core beliefs 4:543
schemas 4:543
for obsessive–compulsive disorder 17:115–116
test anxiety 24:248
United Kingdom 3:874
waves of 4:543–545
first wave 4:543
second wave 4:543–544
third wave 4:544–545
Cognitive training 4:13–14
in psychosocial interventions for psychiatric disorders 22:382
Cognitive transformation 4:153
Cognitive trust 24:665–666
Cognitive tutors, in mathematics learning 4:71, 4:71t
Cognitive Work Analysis 11:300–301
Cognitive/Academic Language Proficiency (CALP) 2:583
Cognitive-adaptive theory 17:875
Cognitive/affect, BDT framework 12:637

Cognitive-affective processing system (CAPS) 16:745
obsessive-compulsive disorder 17:123
Cognitive-behavioral approach to organizational decision-making 24:720
Cognitive-behavioral interventions 6:683
Cognitive-behavioral psychotherapies (CBT) 19:336
for obsessive–compulsive disorder 17:124–125
Cognitive-behavioral skills training 6:682
Cognitive–behavioral therapy (CBT) 6:201, 20:194
Cognitive-contextual theories, of intelligence 12:306
Cognitive-emotional depression 8:341
Cognitive-emotional genre theory 9:162–163
Cognitive-emotional sphere 4:668, 4:671
Cognitivism 7:238
Computer Assisted Instruction (CAI) 4:494–495
in education 12:67–68
and embodiment cognition 7:421, 7:423
emergence of 4:51
and semantic differential 21:568–570
Cognitivist learning theory 23:214
Cognitivistic model of social action 1:78
Cogs and wheels, mechanism as 14:875
Cohabitations 6:116–119, 8:768, 14:606
family 8:766–770
and family therapy 8:815, 8:819–820
level 6:116–119
premarital 6:607
Second Demographic Transition 21:346–347
synthetic or hypothetical cohort measures 6:118–119
timing 6:119
in United States 4:118–121
in Western societies 17:545
childbearing within 17:547–549, 17:549f
incidence 17:545–548, 17:546t, 17:547f, 17:548t
people practicing cohabitation 17:548
Cohabiting families 8:768
Cohabiting women, fertility of 9:13
Cohen, A., on brokers 2:871
Cohen, D., works, *Handbook of Cultural Psychology* 5:312
Cohen, G.A. 1:665–666, 9:400–401
on political freedom 9:400–401
Cohen, Hermann 19:391
Cohen, Lawrence 20:772
on brokers 2:873
Coherence 5:857, 19:395, 19:397–398
and identity 11:521–523
and inferences 12:39–43
legal reasoning 13:778–779, 13:782–783
and relativism 20:235, 20:237–238
Coherence theory of justification 20:79–80
Coherence therapy 4:731
Coherent postsynaptic activity 9:477
Coherentism, and reflective equilibrium 20:79–80
Cohesion 12:243, 21:877, 22:462
and collective efficacy 4:170
in group therapy 10:431
of groups 10:423
in military 15:498–499
in military combat 15:512
organizational groups 10:408
social 12:243
structural equivalence 23:624–625
Cohesion, in social networks 3:906–912
group detection 3:910–911
communities 3:906, 3:908–911
deterministic group finding approaches 3:910
diffusion potential 3:907–908
modularity 3:909–910
search/optimization approaches 3:910–911
relative density measures 3:908–910
structural cohesion 3:906–908
Cohesive subgroups 16:564
Cohesiveness, self-categorization theory 21:457–458
Coh-Metrix 19:395
Cohn, Ludwig 5:35
Cohort, Boomer 20:594
Cohort analysis 11:229, 14:51
of age stratification 1:333–336
in demography 17:727–734
of population aging 18:542
Cohort changes, in obesity 17:78–79
Cohort effects, migration 15:436–437
Cohort life table 14:89–91
Cohort measures, marriage/divorce 6:118–119
Cohort mortality 15:870–871
crossover 15:849
Cohort studies 17:462
evidence-based medicine 8:328–329
Cohort-completed fertility (CCF) 6:156
Cohort-component method 6:165–166, 9:318
for population forecasts 18:593–594
advantages 18:594
curve fitting approach 18:593
incorporation of age structure 18:593–594
Cohort-component method, for small areas 6:166–167
Cohorts 11:5–6
and aging 1:328–329
comparisons 1:329
differences 1:328
changes in older cohorts 1:328–329
historical events 1:328
size and composition 1:328
social change 1:328
and fertility 9:10–12
and generations 9:866–867
leisure 13:855
life course 14:82
and political generations 9:861–862
cohort succession 9:862
COICOP *see* Classification for individual consumption according to purpose (COICOP)
CoIME *see* Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME)
COIN *see* Counterinsurgencies (COIN)
Coinage 19:177, 19:179–181
Coinsurance 10:640
Cointegrated vector autoregressive (CVAR) model 24:324
Cointegration 24:322–330
asymptotic analysis 24:328
asymptotic distribution of estimators 24:328
asymptotic distribution of rank test 24:328
autoregression formulation 24:323
cointegrated vector autoregressive model 24:324
Dickey–Fuller distributions 24:328
environmental Kuznets curves 13:147
error correction models 24:322, 24:325
further topics 24:328–329
integration and 24:325
interpretation of I(1) model 24:325–326
hypotheses on adjustment coefficients 24:326
hypotheses on long-run coefficients 24:326
normalization of parameters 24:326
relation between the models H(r) 24:325–326
likelihood analysis 24:326–327
checking for specification 24:327
likelihood analysis 24:327
reduced rank regression 24:327
likelihood inference 24:326
linear stationary processes 24:324–325
mixed Gaussian distribution 24:328
nonstationarity 24:322–323, 24:325–326, 24:328–329, 24:322f
of political data time series 18:517
regression formulation 24:323
relations 24:323, 24:325–326
unobserved components formulation 24:324
COINTELPRO *see* Counter-Intelligence Program (COINTELPRO)
Coital bleeding 21:798
Coke, Edward, Sir 6:696
Cokriging 23:194
Colander, D.C., works, *The Changing Face of Economics* 24:84–85
Colbertism, and political economy 18:357
Cold cognition 7:416–417
Cold deck imputation 16:923
Cold fusion 21:282
Cold memory 16:198–199, 24:434–435
Cold pressor, cardiovascular reactivity 3:118
Cold societies 1:748
Cold spells 3:868
Cold Spring Harbor Symposium (1950) 11:306–307

Cold War 4:128–134, 4:303, 4:764, 6:810–811, 6:813, 9:540, 11:685–686, 15:513, 17:169, 20:844, 21:371–372, 22:651
and area and international studies 1:946
and arms control 1:976
biomedical experiments 13:562
bipolar state system 23:349–351
bipolarity 4:128, 4:132–133
and communications research 4:278, 4:281
competitive cooperation between superpowers 4:132
components of term 4:128
concept and salience 4:128–130
confrontation 4:128–129
containment 4:128–129, 4:131
decline of European hegemony 25:419
and decolonization 18:659–661
democratic revolutions of post-Cold War era 20:623–625
détente 4:128–129, 4:131–133
diversification of area studies in Soviet Union during 1:950–951
duration 4:129–130
divisions of Europe and Germany 4:129–130
peculiarities of military–geostrategic balance 4:130
East Asian politics during 6:798–800
Chinese politics 6:799–800
Japanese politics 6:798–799
Korean politics 6:799
East–West conflict 4:128–133
European perspectives on 4:132
First World states 23:351
future research directions 4:133
Hilberg on 10:870
history of concept, major developments, and empirical results 4:131–133
and human development studies 6:307, 6:309–310
ideological antagonism 4:128–129
impact on social and behavioral sciences 4:122–127
in Americas 4:122–123
in Asia 4:125
center-periphery theories 4:123
conflict theories 4:122–124
cybernetics 4:124
developmental historicism 4:122
dialectical materialism 4:124
futures studies 4:122–123
game theory 4:122–123
heterodox economics 4:123
Keynesian macroeconomics 4:123–124
neoclassical microeconomics 4:123
newspeak 4:124
scientific positivism 4:122–123
in Soviet World 4:124–125
in Western Europe 4:123–124
international economy 4:129–130, 4:133
intersection of economic and strategic decisions 4:130–131
intra-West tensions 4:129–130
legacies in Central America 1:621–622
and Marxism–Leninism 14:656, 14:660–661
methodological problems 4:133
nuclearization 4:131
peace movements under 17:644
post-Cold War challenges 15:521–523
post-Cold War international system, Third World 24:302–303
and science 16:16
from second cold war to end of cold war 4:132–133
Soviet–Chinese partnership and rivalry 4:128, 4:130
state/society 23:345
Third World 24:300
winner of 4:130, 4:132–133
Colds 12:35
Cole, Johnnetta, works, *All American Women* 8:899–900
Cole, Michael 13:72
Cole, Simon A., on epistemic cultures 7:876
Coleman, James Samuel 4:135
achievements 4:137
contributions 4:136
appraisals of 4:137–138
educational research 4:136
empirical research 4:136
life 4:135
major studies of schools and educational processes 4:136
scholarship 4:135–137
use of mathematics 4:136
works
The Adolescent Society 7:291
Foundations of Social Theory 4:137
The Politics of the Developing Areas 16:371–372
Coleman, Robin Means, works, *African American Viewers and the Black Situation Comedy: Situating Racial Humor* 2:217
Coleman report 7:151–152, 7:268, 21:43, 21:119–120
Coleridge, Samuel Taylor 15:290
works, *Encyclopædia Metropolitana* 7:584
COLES *see* Constructivist-Orientated Learning Environment Survey (COLES)
Cole-Turner, Ronald 20:288
COLI *see* Cost of living index (COLI)
Colias butterflies, enzymes 1:87
Colias spp. 1:87
Collaboration 4:226, 4:489, 6:896–897, 7:869, 16:553, 17:540–542
by associations 25:257
in a broader context 21:113–114
community–labor alliances 24:491–492
during competition/cooperation 4:856–857
in education 21:116–117
and ethnography 20:89–91
extra-school examples 21:116
school–agency collaboration 21:116
school-to-school collaboration 21:116
feminist 8:902–903
foundational characteristics 21:112–113
intraschool examples of 21:114–116
consultation 21:114–115
co-teaching 21:115–116
team teaching 21:115
teams 21:114
logistics of 21:117
in multisited field studies 16:104–108
mutual goals in 21:113
parity 21:113
in schools 21:113–114
shared accountability 21:113
shared resources 21:113
shared responsibility for key decision-making 21:113
technical definition of 21:112
understanding of 21:112–114
voluntary nature of 21:112–113
Collaboration approach, for strategic planning 23:519–520
Collaboration network 22:461–465, 22:463f
Collaboration of Researchers for the Effective Development of Offender Supervision *see* CREDOS (Collaboration of Researchers for the Effective Development of Offender Supervision)
Collaborative circles 5:468, 22:105–106
in small group research 22:105–106
Collaborative citizen science projects 3:632
Collaborative consumption 4:139–144
background 4:139–140
differentiating from online collectives 4:140–142, 4:141t
effects of 4:142
entrepreneurship 4:140, 4:143
and evolution of consumer roles 4:140
exchange 4:139–140, 4:142–143
future research directions 4:142–143
marketing 4:139–140, 4:142–143
networks 4:139, 4:143
peers 4:139–140, 4:143
technology 4:139
Collaborative couples, and scholars, private lives of 21:34–36
Collaborative empiricism 4:545
Collaborative ethnography 1:223
Collaborative governance 10:296
Collaborative government
civil service 3:697
Collaborative inquiry 16:553
Collaborative learning 12:203
and coconstructivism 3:916–917
social-cognitive perspectives 13:675
Collaborative participation, in participatory action research 17:535–536
Collaborative planning 18:184–185, 18:193, 18:204–205
Collaborative practice 23:328–329
Collaborative remembering 4:191
Collaborative research 20:508
and epistemic community 7:867–872
interdisciplinary collaborations 7:870
international collaboration 7:869

in natural and social sciences 7:868–869
reasons for increasing collaboration 7:869–870
European Organization for Nuclear Research 7:864
industry 7:864–865
with laypersons 7:864
paradigms 7:864–865
professionals 7:864
science 7:862, 7:864
between scientists 7:864
with stakeholders 7:864–865
translational medicine 7:864–865
Collaborative social theory, and Frankfurt School 9:381–382
Collapse (media) 15:21–22
Collateral assets 21:401, 21:405
Collateral consequences
of criminal charges 15:592
and mass imprisonment 14:696–701
economic outcomes 14:697–699
family consequences 14:699–700
future research 14:696, 14:700
health 14:696–699
individual consequences 14:696–699
life course 14:696–697, 14:699
social exclusion 14:696–700
spillover effects 14:696, 14:698–699
well-being 14:697, 14:699
Collateral damage, war 25:377, 25:379
Collateralized debt obligations 9:171
Collateralized mortgage obligations (CMOs) 25:552
Collatio legum Mosaicarum et Romanarum 13:799
Collected memory 4:188–189
Collection of n data observable vectors 23:413
Collections
art 16:142
development of 16:146
knowledge of 16:146
materiality 14:766–767
and museums 16:143, 16:145
of royalty 16:142–143
of valued objects 16:142–143
Collections, case study 3:179–181
Case Centre, UK (formerly ECCH: European Case Clearing House), Cranfield University, UK 3:181
Case Consortium@Columbia, Columbia University, New York 3:179
Case Method Project, US Marine Corps University (MCU), Quantico 3:179
Case Program, ANZSOG, Melbourne 3:179
Case Program, Harvard Kennedy School, Cambridge 3:179
Case Study Program, The Institute of Public Administration of Canada (IPAC), Toronto 3:179
Case Study Unit, Lee Kuan Yew School of Public Policy, University of Singapore 3:179–180
Centre for Social Investment, Heidelberg University, Germany 3:180
China Case Center for Public Policy & Management (CCCPPM), Tsinghua University, Beijing 3:180
Electronic Hallway, University of Washington 3:180
Enduring Legacies Native Cases, Evergreen State College, Olympia, WA 3:180
Engineering Case Program, Rose-Hulman Institute of Technology, Terre Haute, IN 3:180
EnviroDx-Clinical Cases in Environmental Medicine, University of Utah 3:180
E-PARCC, Collaborative Governance Initiative, Program for the Advancement of Research on Conflict and Collaboration, Maxwell School, Syracuse University, New York 3:180
Epidemiologic Case Studies, Centers for Disease Control, Atlanta, GA 3:180–181
Global Health Delivery (*GHD*) *Project*, Harvard University, Cambridge, MA 3:181
Hubert Project, University of Minnesota, Minneapolis, MN 3:181
Innovations for Successful Societies (ISS), Woodrow Wilson School of Public & International Affairs and the Mamdouha S. Bobst Center for Peace and Justice, Princeton University, Princeton, NJ 3:181
INSEAD Humanitarian Research Group, Fontainebleau, France 3:181
National Center for Case Study Teaching in Science, University at Buffalo, Buffalo, NY 3:181
Pew Case Study Center, Institute for the Study of Diplomacy, Georgetown University, Washington, DC 3:181
School of Public Affairs and Administration Case Collection (SPAA), Rutgers University–Newark, NJ 3:181
Collective action 6:216–217, 15:253–254, 16:8–11, 19:534–535, 22:431, 22:435
choice and 3:524
crowds and 4:151–154
collective psychology, origins of 4:151–152
ESIM of 4:153–154
site of pathology, debating 4:152
social identity and 4:152–153
and group identity 4:175
and identity 11:526, 11:528–530
and identity movements 11:543, 11:545
macro–micro–macro theory of 10:17–18
Olson on 17:207–209
organizations as constructs of 17:415–417
and political protests 18:421–423
in political psychology 18:432–433
problem, institutional solutions to 19:541–542
and public opinion 19:562
rational choice explanation 19:896, 19:899
and social dominance orientation 22:252
social protests 22:512–514
classical approach 22:512–513
digital approach 22:513–514
importance of interdependent decisions 22:513
and solidarity 23:6
symbolic resources 4:177
use of tem 4:151
workers 20:866
Collective actors 15:19
Collective agency, cooperation as 4:857
Collective appropriation, hunters-gatherers 11:428–429
Collective association, sociology of professions 19:107
Collective autobiographies 6:495–496
Collective bargaining 4:771, 11:349, 11:860–861
trade unions 24:497t
working conditions 25:698–701
Collective bargaining legislation 23:591
Collective behavior 4:151–156, 5:362–363, 16:10–11
continuing exclusion, of collective action 4:154–155
collective resistance, neglect of 4:154–155
conformity bias 4:154–155
social change, decollectivization of 4:155
crowds and collective action 4:151–154
collective psychology, origins of 4:151–152
ESIM of 4:153–154
site of pathology, debating 4:152
social identity and 4:152–153
and deindividuation 6:38, 6:40
of groups 10:425–426
and leadership 13:652
and politics modeling 18:514
and rumors 20:788–789
and social identities 22:323
and social movements 22:447–448
sociology of 4:157–162
delineation of subject matter 4:157
emotions 4:157–159, 4:161
enduring dilemmas 4:159
framing theory 4:161
historical threats during past half-century 4:159
interest in variation 4:157
irrationalism 4:157, 4:160–161
Le Bon on 4:158
Mackay on 4:157
Marx on 4:158–159
opportunity structure 4:160
political contestation 4:160
resource mobilization 4:160
scientific neutrality 4:157
Smelser on 4:160
Tocqueville on 4:158
use of term 4:151
Collective beliefs 4:163–168, 12:411–414
category 4:164
difficulty in explaining 4:163–164
diversity 4:163

Collective beliefs (*continued*)
and home 11:156–157
and ideologies 11:573
ideology 4:164
knowledge 4:164
rationality 4:166–167
religion 4:165
and social mechanisms 4:164–166
sociological explanations 4:166–167
Collective biography 2:647
see also Prosopography
Collective capitalism, rise of 11:857–858
Collective change, cosmopolitanism and 5:41–42
Collective choice 6:92, 12:192
antecedents to 17:207–208
Collective choice theory 6:92
Collective conscience 4:200
Collective consciousness 5:42–43, 14:825–826
and knowledge 13:71
Collective consent, community engagement and 9:965
Collective consumption, Castells' concept of 3:229
Collective coping 4:903–904
Collective decision making 2:371–372
Collective efficacy 4:169–173, 4:324, 4:460, 4:836, 4:838, 5:159, 5:161, 5:191, 9:593
behaviors 4:169–172
communities 4:169–170, 4:172
consequences in neighborhoods 4:170–171
delinquency 6:71
distinguishing between cohesion and informal social control 4:170
feedback effect from neighborhood crime to 4:171
history of concept 4:169
measurement challenges 4:169–170
and social control 4:824
and social networks 4:171–172
theoretical extensions 4:172
Collective feelings, and collective beliefs 4:165
Collective goods, and rational behavior 19:909–910
Collective identities 5:538, 7:323, 11:529–530, 16:9–10, 21:176, 23:853
and biography 2:639
construction of 22:438
and ethnicity 8:148, 8:157–160
and language 13:251
memory 1:882–883
and social networks 16:563
in South Asia 2:77
Collective imagination, and nostalgia 17:34–35
Collective incentives 4:146
Collective intentionality 21:343–344, 22:491
Collective knowledge, Internet and 12:587–588
production 12:588
Collective landownership in China 5:598
Collective legal mobilization 13:593–594
Collective level studies, of social media 16:784–785
Collective litigation, class actions 3:733–734, 3:736–737
Collective memories 4:194–199, 4:524, 5:606, 11:129, 20:484, 20:486, 20:489
anthropology of 4:181–187
archaeology 1:881–884
and collected memories 4:188
collective rituals 4:196
and commemoration 4:235–236
complementary variants 4:194
cultural production 5:610–611
death/dying 5:868
in discipline-building 6:492
generations 9:866–867
group-based emotions 4:196–197
and groups 10:436–439
Halbwach's perspective on 9:866
history 4:194–198
and home 11:157
homogeneous variants 4:194
and individual memory 4:188
intergroup relations 4:194, 4:197–198
as a process 4:190–191
as a product 4:191–192
psychological antecedents of 4:195–196
cognition 4:195
conversation 4:195–196
emotions 4:196
narratives 4:196
psychological perspective 4:189–192
psychology of 4:188–193
social dimensions 4:189
social psychological functions of 4:196–198
reconciliation processes 4:197–198
social identity approaches 4:197
as social representations 4:198
social psychology 22:539
social representations of history 4:195, 4:197
sociology of 4:200–205
age of preservation 4:203–204
collective conscience 4:200
habitus 4:201
in Halbwachs 4:200
history 4:202–203
life history 4:203
modern society 4:203–204
monument 4:201, 4:203
museum 4:201, 4:203
negative heritage 4:203
order of recollection 4:202–203
past constructed by present 4:201–202
past is in present 4:200–201
politics of memory 4:204
postmodern time 4:203–204
social order and social memory 4:203
stock of knowledge 4:200
tacit knowledge 4:200–201
victimhood 4:198
Collective mentality 14:205
Collective mentoring, by affirmative action 9:619
Collective mind 6:38
Collective mobility, in collectivism/individualism 4:207–208
Collective naming 16:184–185
Collective organizing, informal sector 12:58
Collective ownership, principle of 9:966
Collective phenomena 21:877
in groups 10:436–438
dependency networks 10:438
groupishness 10:438
individual differences 10:438
motivation 10:437
norms 10:436–437
social embedding 10:436
status hierarchies 10:437–438
Collective political action 4:145–150
anger 4:145, 4:148–149
challenges 4:145
grievances 4:145–148
mobilization 4:145–149
participation of people in 4:145–149
emotions 4:148–149
identity 4:147–148
ideology 4:148
instrumentality 4:146–147
motivation 4:146–149
political protest 4:145
preparedness 4:147, 4:149
social embeddedness 4:149
transformation of potential into action 4:149
Collective psychology, origins of 4:151–152
Collective publics, and opinion formation 17:244–245
Collective redress, class actions 3:733–734
Collective representations 6:701, 20:443–444, 22:533, 22:991–992
Collective resistance, neglect of 4:154–155
Collective responsibility 20:583
Collective rights 3:674, 13:436
Collective rituals 4:196, 20:861
Collective selection 8:837
Collective self 7:497
elements of 25:29
Collective self-defense 10:17
Collective self-objectification 4:153
Collective sex 21:752–753
Collective socialization models 16:443
Collective solidarity, and social integration 12:246–247
Collective teacher self-efficacy 21:520
Collective unconscious, Jung on 12:904–905
Collective violence 20:338
and deprivation 6:216–217
instigators of 14:702–703
difficult life conditions 14:702
group conflict 14:702–703
self-interest 14:703
violent revolution 14:703
war 14:703

Collective wisdom production 12:588
Collective worlds, constitution of 1:83
Collectives 17:201–202
aggregative properties 11:816
in China 2:66–69
emergent properties 11:816
Latour on 13:414–418
and mass societies 11:818
problematic status of 11:818
reduction to individuals 11:818–819
as taxonomic groups 11:816
Collectivism 2:132–134, 2:148–149, 4:206–210, 5:567, 13:199, 16:398, 23:644
adolescent friendship in 9:433–434
conditions for formation 9:433
and adolescent prosocial behavior 19:223–224
argument in favor of 11:819
attributions 4:208
cognition, emotion, and motivation 4:208
collective mobility 4:207–208
common fate 4:207
and communication 4:209, 4:284
conflict resolution 4:209
conformity 4:209
describing self 4:208
distribution 4:206–207
and economic development 5:528–529, 5:531
economic status 4:207
emergence 11:811–812
emotions 4:209
exposure to other cultures 4:207
globalization and 4:209–210
hard, implausibility of 11:812–813
history 4:206
homogeneity 4:207
ideal of fitting-in rather than shining 4:208
versus individualism 5:531, 9:446
influence on homosexuality stigma 21:736–737
ingroups 4:207
and intercultural communication 12:358
interdependence 4:207
interpersonal relationships 5:426–427, 5:429–430
intrinsic motivation 12:696–697, 12:699–700
large families 4:208
little exposure to modern mass media 4:208
methodological 17:934
mobilizing for confrontations 4:208
ontological 11:819
philosophical aspects 11:811–815
population density 4:207
and prosocial behavior 19:239
recreation 4:209
resource distribution 4:209
skills in entering new groups 4:208
social behavior 4:208
social loafing 4:208–209
social movements 4:270–272
social stability 4:207
social support 4:208
socialization 4:208
soft, plausibility of 11:813–815
strict religious upbringing 4:208
transgressions of ingroup 4:209
utilitarian social thought 24:984–985
and utilitarianism 11:818
and values 25:28–31
well-being 4:208
Collectivist democracy 17:202
Collectivist theory 7:434–435
Collectivist values 17:825
Collectivistic cultures, divorce in 14:518
Collectivities 2:252
civilizational 3:726
Collectivity orientation (pattern variable) 1:382
Collectivization 4:304, 20:843
land 13:233
and property 19:155–157
socialist 13:234
College, Career, and Civic Life (C3) Framework for Social Studies 22:731–732
College Characteristics Index (CCI) 7:820
Collège de Sociologie 6:704–705
College grades, prediction, outliers in 23:387–388
College premium, and education 7:134–135
College publishing 19:625, 19:627–628
College sex 21:842
College students 25:815–816
psychosocial development research on 7:942–943
Colleges for Seniors 1:150
Collegial citizen science projects 3:632
Collegiality, and teacher leadership 24:60–62
Collegiate participation, in participatory action research 17:535–536
Colley, Linda 1:609
Collier, Jane 8:899
works, *Woman, Culture and Society* 8:898–899
Collinear facilitation 25:159–160
Collingwood, R.G. 8:289, 8:292, 11:13, 11:94
on historical explanation 11:13
works
The Idea of History 3:61
Collini, Stefan 12:256, 12:258–259
Collins, Harry
on disunity of sciences 7:875
doctrine of special relativism and ban on reflexivity 20:94
Empirical Programme of Relativism 24:136
Collins, H.M., works, *The Golem series* of books 21:310
Collins, Patricia Hill 12:659–660
Collins, Randall 4:211–217, 12:279, 16:563
conflict theory of 4:211, 4:214–216
early breakthroughs of 4:213–215
early years of 4:211
on historical sociology 4:211, 4:214–215
integrative and innovative sociological theory of 4:215
on interaction ritual theory 4:211, 4:213–216, 22:335
on microsociology 4:211, 4:213, 4:215–216
on network theory 4:215
on social change 4:214–215
on social stratification 4:212, 4:214–216
on sociology of education 4:211, 4:214
on sociology of emotions 4:214–216
on sociology of family 4:214–215
on sociology of gender 4:214–216
on sociology of love 4:214–216
on sociology of sex 4:214–216
studies in Harvard, Stanford, Berkeley 4:211–213
violence studies of 4:215–216
works
The Case of the Philosophers' Ring 4:214
Conflict Sociology 4:214
The Credential Society 4:214
The Discovery of Society 4:214
Does Capitalism Have a Future? 4:216
Four Sociological Traditions 4:214–215
Interaction Ritual Chains 4:215
Macro-History 4:215
Max Weber: A Skeleton 4:214–215
The New Economic Sociology 4:215
Sociological Insight 4:214–215
Sociology of Marriage and Family: Gender, Love and Property 4:214
The Sociology of Philosophies 4:213, 4:215
State and Society 4:212
Symbolic Interactionism: Perspective and Method 4:212–213
Theoretical Sociology 4:215
Three Sociological Traditions 4:214–215
Violence: a Micro-Sociological Theory 4:216
Weberian Sociological Theory 4:215
Collocational information 13:943
Collocations 13:932–933
Colluvium 7:693t
Cologne Carnival 3:146
Cologne War of 1583 20:102–103
Colombia 22:625
Colon cancer, and physical activity 18:88
Colonial construction, and caste system 3:224
Colonial empires 4:223
Colonial environmental history 7:747–750
ideas in twentieth century 7:751–752
and landscape imagery 7:752–754
Colonial era
British rule 11:79
historiography of 11:79–80
under colonial rule 11:79–80
modern India after decolonization 11:80
Colonial evangelists 15:615
Colonial expansion 4:223, 4:247–249
Colonial gender policy in modern Korea 6:794
Colonial historiography, Indo-Islamic period 11:79
Colonial history

Colonial history (*continued*)
Humboldt on 25:279, 25:284–285
of social work 16:457–458
Colonial labor market 22:57
Colonial law 4:264
in Southeast Asia 13:814–815
Colonial library 22:604
Colonial modernity
in modern China 6:795
in modern Japan 6:794–795
Colonial policing 18:246, 18:248
Colonial practices 11:763–764
Colonial protonationalism 16:268
Colonial railways 4:227
Colonial research societies, for social and behavioral science in Asia 22:616
Colonial scheme, gift as 10:142
Colonial science 7:750–751, 18:652–653
Colonial situation 8:192–193
Colonial Social Science Research Council 3:6
Colonial studies, ethnohistory 8:192
Colonial/institutional documents 8:195
Colonialism 3:536, 4:223–227, 11:685–686, 11:770, 11:795–796, 11:897–898, 18:647–650, 19:676–677, 19:804, 20:397, 22:57, 22:621, 23:633, 24:894
academic 11:898
and African cultural studies 1:277
and agrarian political economy 1:457–458
anthropology of 4:218–222
discipline 4:218–221
nationalism 4:220
problematization of 4:221
racism 4:218
self-consciousness 4:218
and anthropology of Korea 2:71–72
in Arctic 1:927–928
and British anthropology 1:768
and brokers 2:870–872
and cargo cults 3:139–141, 3:143
in Caribbean 1:638–641
collective memory 4:184
and communalism 4:263–264
and creolization 5:117–118, 5:120
and development 6:301–302
East Asian studies 6:785–786
elites 7:387
end of, and anthropology 1:768
fieldwork 9:149, 9:151
and genealogy 9:806–810
and homosexuality 13:858
Humboldt on 25:280, 25:282–284
Indigenous peoples 11:770–772
and indigenous peoples 11:806
and law 13:539–540
and materiality 14:766
Memmi on 15:136
and Middle East politics 16:372
minority-group status 15:571
and nationalism in East Asia 16:277
in North America 17:16–18
in Oceania 17:160–162
and peace/nonviolence 17:626
postcolonial geography 18:644
psychiatric care 6:46
and racism 19:857–858, 19:860–861
and sexual citizenship 21:731
and sexuality, in Africa 10:812
constructions of masculinities and feminities 10:813–814
in South Asia 16:279–281
Southern Africa 23:77–78
and spirit possession 23:283
and Third World 24:301
tribes and 24:645
and war 25:378–380
in West Africa
postcolonial social trends 1:303
social consequences 1:302–303
and whiteness/white supremacy 25:569–573
Colonialist elitism 23:632
Coloniality–modernity distinction 15:695
Coloniality/modernity paradigm 15:694–695
Colonic irrigation 10:563
Colonies of exploitation 4:226
Colonization 4:223–227, 9:466, 16:266
in archaeology 15:472–482
history 15:472
human dispersals 15:472, 15:477–479
neolithic-population movements 15:477–479
prehistory 15:473
Australia and New Guinea 2:229–231
civilizing mission and tradition, in Africa 1:286–287
colonized world, Memmi on 15:135–137
consequences of 4:226–227
decolonization 4:226
and de-Westernization 25:541–543
history and geography 4:224–225
Māori social work 14:505–506
and national liberation movements 16:243
Russian 9:466
semantic field 4:223–224
typology of 4:225–226
Colonization of law 13:588
Colonization of lifeworld, Habermas on 10:471
Colonizers, and ethnic groups 8:104
Color 4:231
visual mental imagery of 25:164–165
Color anomia 23:294
Color classification 4:228–230
cognitive and linguistic research on 4:228
color triad 4:229
color vision 4:229–230
linguistic characterization 4:228
'parallel' devices 4:229
symbolic meaning 4:229
three-value logic 4:229
Color perception, influence of color terms on 13:294–295
Color revolutions 20:624, 25:839
and youth movements 25:839–840
Color terms
influence on color perception and memory 13:294–295
linguistics of 4:231–234
B&K's definition of bct 4:232–233
categorization 4:231
color terms as linguistic objects 4:232
domain restriction 4:231
linguistic relativity 4:231–232
nonbasic color terms and noncolor meanings of color terms 4:234
practical problems with B&K definition of bct 4:233
theoretical issues concerning the B&K definition of bct 4:233
Color to grapheme 14:196
Color triad, and color classification 4:229
Color vision, and color classification 4:229–230
Colorblind racism 18:842, 19:859
Colorblindness 7:895–896, 18:835, 19:796, 19:799
inequality 7:894–896
and prejudice 18:835–837
ColorBrewer 3:172
Color-coded racism 19:852
Color-consciousness 18:835
and prejudice 18:835–837
Colored grapheme 14:196
Colored number synesthesia 14:196
Color-patch maps 24:265
Colossians 3:238
Colostrum 2:848–849
Colson, Elizabeth 1:768–769
Colten, C., works, *An Unnatural Metropolis* 11:19
Colugos
primates, evolution of 18:908
Columbia
mass media, Lazarsfeld on 13:638
tradition 15:248
voting model 25:302
Columbia Broadcasting Service *see* CBS (Columbia Broadcasting Service)
Columbia Broadcasting System (CBS) 24:169
Columbia Card Task (CCT) 17:811
Columbia Philosophy Department 21:453
Columbia Pictures 9:156–157
Columbia School 18:467
Columbine High School incident, aftermath and school safety 21:108–109
Columbus, Christopher 23:45
Coma, electroencephalography 7:356, 7:359–361
Comaroff, Jean
on magic 14:435, 14:437–438
Comaroff, John
on brokers 2:871
on magic 14:435, 14:437–438
Combat effectiveness 15:512
Combat stress 15:497
Combination treatments, for panic disorder 17:470–471
Combinatorial data analysis (CDA) 19:431–435
branch-and-bound algorithms 19:433–435
clustering 19:433–434

approximation methods 19:434
blockmodeling of social network data 19:433
clustering nonmetric data 19:433
distance-based clustering of metric data 19:433
exact methods 19:433–434
confirmatory 19:431–432
distributional tests 19:432
exact p-values 19:432
linear assignment 19:431
multinominal models 19:431–432
quadratic assignment 19:431
resampling 19:432
dynamic programming 19:432–434
metaheuristics 19:434–435
nonparametric statistics 19:431
permutation tests 19:431–432
randomization tests 19:431
seriation and scaling 19:432–433
approximation methods 19:433
exact methods 19:432–433
gradient-based seriation 19:432
multidimensional city-block scaling 19:432
paired comparison ranking 19:432
unidimensional scaling 19:432
variable selection 19:434–435
approximation methods 19:434–435
cluster analysis 19:434
discriminant analysis 19:434
exact methods 19:434
factor analysis and principal component analysis 19:434
regression 19:434
Combinatorial explosion 2:32
Combinatorial optimization
problem, computerized test construction as 4:510–511
model 4:511
objective functions 4:510–511
restrictions 4:511
Combinatory Categorical Grammar (CCG) 13:942, 13:946
COMBINE studies 1:501, 1:504
implications of 1:504
Combined antiretroviral therapy (cART) 1:491–493
Combined Behavioral Intervention (CBI) 1:504
Combined census 3:298
Combined DNA Index System (CODIS) 9:340, 9:958–959
Combined medicine, for post-traumatic stress disorder 18:705
Combined motor method, of Luria 14:396–398
Combined pharmacotherapy and psychotherapy 19:495
background 19:495–496
communication 19:497–499
ethical issues 19:498–499
legal issues 19:497–498
maximizing collaborative relationship 19:499
medication backup 19:495–496
negative aspects 19:497
clouded legal and clinical responsibility 19:497
inappropriate prescribing decisions without knowledge of therapy content 19:497
lack of reimbursement for collaboration 19:497
misconceptions of patient, therapist, and psychiatrist 19:497
potentially discrepant information given by patient to each clinician 19:497
splitting of clinicians 19:497
unclear confidentiality 19:497
positive aspects 19:496–497
cost-effective utilization of resources 19:496
enhancement of adherence 19:496–497
increased amount of clinical information 19:496
increased time and resources available for patients 19:496
professional/emotional support for both therapist and psychiatrist 19:497
selection of therapist by patient with similar ethnic/gender background 19:496
utilization of special talents of both therapist and psychiatrists 19:496
Combined statistical areas (CSAs) 24:857
Comfort women 2:74, 12:795–796
Comforting, prosocial behavior 19:233–235
Coming-of-age 24:577–578, 24:580
Coming-out models, and same-sex youth 9:651, 9:653–654
Comintern. see Communist or Third International (Comintern)
Comisión Económica para América Latina (CEPAL) 22:622
Comite d'Organisation des Recherches Appliquées sur le Développement Economique et Social (CORDES) 22:670
Comité International d'Histoire de l'Art (CIHA) 2:23
ComLabGames 24:85
Command, definition 13:484
Command and control (CAC) regulation
of business organizations 25:555
of environment 7:727, 7:773
Command theory of law 4:250, 13:764, 18:24, 18:26
Commemoration 4:235–242
aborted 4:239
anomalous 4:238
as cultural system 4:240–241
delayed 4:239
era of commemoration 4:237
history and collective memory 4:235–236
human rights movement 4:237
memory of 4:236–237
multiculturalism 4:239
networks 4:236
path dependency 4:236–237
politics of 4:239–240
politics of recognition 4:237
politics of regret 4:236
in post-heroic era 4:237–238
and sacrifice 20:860
surface 4:238–239
symbolism 4:235
theories 4:239
transitory 4:239
Commensurability of modes of influence 4:292
Commensuration
ecological economics against 6:855–857
quantification 19:702
Commentators 15:484–485
legal reasoning 13:778
Commerce
entertainment 7:671
Commercial arbitration 12:470–471
Commercial biobanking and research, direct-to-consumer (DTC) personal genetic testing 6:443
Commercial broadcasting
by radio 19:877–878, 19:880–881
regulation 2:860
Commercial capitalism 3:95–96
Commercial classifications 3:799
Commercial contracting, of organizations 17:398
Commercial data providers 5:710
Commercial disputes 6:557
Commercial impracticability 3:28
Commercial laws 3:30
Commercial newspapers 12:852
Commercial policy, international trade 12:568
trade negotiations for 12:568–569
unilateral trade policy 12:568
Commercial sex 14:542, 19:279–281
LGBTI 13:882–886
Commercial sex workers (CSW)
AIDS-related mortality 15:835
Commercial sexual exploitation (CSE) 24:512, 24:514
in Latin America 24:512–515
Commercial society, sociability in 12:387
Commercial sponsors, obligations to, surveys and polls 23:753
Commercial victimization 19:163
Commercialization 12:145, 20:422, 24:141, 24:765
of biotechnology 2:678
ethical issues with 2:606–607
and journalism 12:854
of media 19:605
of science 17:13
Commission (European Union) 8:265, 14:265
Commission for Racial Justice 7:757
Commission on Social Questions of the League of Nations 12:477
Commissives 18:797
Commissural fibers 12:416
cognitive functions and 12:418–419

Commissural fibers (*continued*)
interhemispheric excitation via 12:417
Commissures, contribution of Sperry to 23:280
Commissurotomy 23:294–295
anatomy 23:291f
color anomia 23:294
development of asymmetry of behaviors 23:295
neuropsychological investigation 23:295
total 23:294
COMMIT study (Community Intervention Trial for Smoking Cessation) 3:900
Commitment 4:267, 7:938, 17:334–340
antecedents of 17:336
budgeting 2:895
in close relationship 22:566
constructs, evolution of 17:334
cultural differences and influences 17:339
downsides of 17:339
general model of workplace commitment 17:335f
in groups 10:423
identity 21:434–436, 21:438
ideological 4:269
in long-term heterosexual relationships 20:912–917
and love 14:379
outcomes 17:335–336
participatory action research 17:536–537
process model of 17:335f
related constructs 17:336–338
burn-out 17:337–338
engagement 17:337
identification 17:337
motivation 17:337
satisfaction 17:336–337
stress 17:337–338
well-being 17:337–338
research and resolution
competing conceptual definitions 17:338
differences in commitment across targets 17:338–339
temporality 17:339
social influence 22:349–350
Commitment mechanisms, trade policy 12:577
Committee for a Sane Nuclear Policy 17:644
Committee for Legalizing Eugenic Sterilization 8:219–220
Committee for the Study of Population Policies 8:234
Committee of Concerned Asian Scholars (CCAS) 1:953
Committee of Training in Clinical Psychology 24:526
Committee on Lesbian Health Research Priorities 15:115
Committee on Publication Ethics (COPE) 12:864, 20:531
Committee on Science and Practice (CSP) 19:383–384
Committee on the Elimination of Racial Discrimination (CERD) 11:796–797
Committees of United States Congress 4:650
Commodification 5:538, 10:860, 11:234, 20:261, 24:839, 24:847
and capitalism 3:105–106, 3:108
of Cuban labor 5:370–371
and discourse 6:507
of emotion 11:146
and globalization 10:235
market 14:538–539, 14:541–542
and political ecology 18:352–353
of sexual activity 19:279
and social work 24:491
Commodified culture 6:551–554
Commodities 4:243, 7:13–14
and criminology 5:244–246
cultural meaning of 5:109–110
and exchange 8:476–480
exchange in kula ring 13:142
fetishism 9:105–107
in Melanesia 17:154–155
and money 15:747–748
Commoditization, collective memory 4:184–186
Commodity bundles 4:245
Commodity chains 4:243–246, 16:570
analysis, recent tendencies in 4:244–245
commodity bundles 4:245
consumption nodes 4:245
expanded version 4:244–245
production nodes 4:245
buyer-driven 4:243–244
conceptions 4:243–244
commodity circuits 4:244
global commodity chains 4:243–244
systems of provision 4:244
and consumption politics 4:245
dimensions of 4:243
governance 4:243–244
networks 4:243–244
producer-driven 4:243–244
semiotic analysis 4:245
supply chain 4:243–245
Commodity circuits 4:244
Commodity fetishism 3:740, 11:865–866, 20:203–205
Commodity production, and modernity 15:688
Commodity programs 1:480–481
Commodization
in craft production 5:108
definition 5:108
Common Agricultural Policy (CAP) 1:481
Common chronology 24:412
Common cold 12:35
psychoneuroimmunology 19:446–447
Common consciousness 13:550
Common couple violence 12:685–686
Common descent
evolution 8:385
Common descent, evolution 8:385
Common distribution, violations, t-test 8:10
Common environmental variance 12:298
Common factors 14:853
psychological therapies 19:388
adherence of participants 19:388
fidelity of treatment delivery 19:388–389
participant's expectancy of benefit 19:388
therapeutic relationship 19:388
therapist competence 19:388–389
and psychotherapy integration 19:511–513
Common fate, in collectivism/individualism 4:207
Common good 4:312, 4:316, 20:478
and public interest 19:552–553
republicanism 20:478
Common ground 12:95–96
Common in-group identity model (CIIM) 12:411
Common interest developments (CIDs) 9:612, 23:663
Common knowledge 12:840
Common law 3:669, 3:672, 4:247–251, 4:939, 7:925, 13:565, 13:619, 19:190, 19:194–195, 20:781–782, 20:784, 20:786, 21:402–403, 23:722–725
adjudication 5:104–106
in annexed or protected territories 4:248
classification 21:402
construction rules 4:811
direction of 4:250–251
in England 4:247–248
geographical diffusion of 4:248–249
historical development 4:247–248
procedure 4:247
theorization of 4:249–250
tort claims and 24:430
trials 24:636–638
use of term 4:247
working practices 4:247
Common law colonies, in sub-Saharan Africa 23:656–657
Common law contracts 4:809
judicial management of 4:811–812
tension between liberal and critical conceptions 4:812
Common law judges 12:880–881
Common law jurisdictions 13:585–587
discovery in 6:516–520
England 6:518–520
United States 6:516–518
Common law liability doctrine 10:595
Common law reaction, legal reasoning/argumentation 13:779–780
Common law systems 4:809–810, 8:582–583, 13:568, 13:615, 13:628–629, 14:268
procedure in 14:266
Common mental disorder 15:228
Common names, impairments of 19:147–152
Common pool problems 2:353–355
of property rights 24:529
Common pool resources (CPRs) 12:725, 16:416–417, 19:154
international public goods 19:540–543
in society 19:910
Common property 19:154–155
Common resource dilemma 22:245–246
Common Rule 4:252–257
activities covered by 4:253

assurances 4:252, 4:256
confidentiality 4:254, 4:256
criteria for approval of research 4:254–256
informed consent 4:254–256, 4:255t
risk-benefit assessment 4:254
enforcement of 4:256–257
equitable selection of subjects 4:256
institutional review boards 4:253–254
Office for Human Research Protections 4:255–256
privacy 4:253–254, 4:256
research ethics 4:255
surveys and polls 23:754–755
in United States 8:37–38
Common sense 2:632, 10:748
anthropology of 4:1–5
boredom 4:2
contingency 4:1–4
ethnography 4:2–4
evidence 4:2–4
fieldwork 4:2–3
hegemony 4:2–4
hybrid knowledge 4:3
power 4:2–4
rationality 4:4
reflexivity 4:2
relativism 4:1–2
religion 4:1, 4:4
self-evidence 4:2
self-referential 4:4
social improvement 4:4
universalism 4:1–5
ethnomethodology 8:203–205
Harold on 9:603
of networks 5:467
precursors 12:101
social representations 20:443–444, 20:446–448
Common Sense Media 7:252
Common support, causal inference 3:257–258
Common value auction model 2:207–208
general value model 2:208
independent information variables 2:207
strictly affiliated information variables 2:207–208
winner's curse 2:207
Commonalities 3:191
Common-mode rejection ratio (CMRR) 7:354
Common-object network 15:660–661
Commons, John R. 11:860
Common-sense model (CSM) 11:597
Commonsense naturalism 16:326
Commonsense psychology 2:201
Commonsense realism 20:22
of Reid 20:21
Commonsense reasoning 16:925
Commonwealth 3:707
republicanism 20:477–478
Commonwealth of Independent States (CIS) countries 7:76
Commonwealth Scheme for Mutual Assistance in Criminal Matters of 1990 (Harare Scheme) 24:599
Communal food production 14:255
Communal rationalism 22:164, 22:166
Communal relationships 22:563–566
Communal societies 2:148–150
Communal strength 22:563–564
Communal utopians 2:146–147
Communal violence 24:949
Communalism 4:263–265, 23:364
and adolescent prosocial behavior 19:223–224
colonialism 4:263–264
communitarianism 4:263
Hindu 4:263–265
Khalistan 4:264
law 4:264
Muslim 4:263–265
nationalism 4:263
in Pakistan 4:263–265
religious communities 4:263–265
Sikh 4:263–264
in South Asia 4:264
and Transnationalism 4:263–265
Communalism, universalism, disinterestedness, originality and skepticism *see* CUDOS (communalism, universalism, disinterestedness, originality and skepticism)
Communauté Européenne 4:341–342
Communes
anthropology of 4:266–272
alternative communities 4:266
anthropological and ethnographic studies 4:266–270
Brethren on 4:266–267
communalism and collective social movements 4:270–272
edited collections and general sociological study 4:269–270
ideological commitment 4:269
intentional communities 4:268–269
Israeli kibbutz 4:267–268
Japanese communes 4:268–269
ontological otherness 4:270–271
Russian Kolkhoz and Mao's communes 4:269
self-interest 4:267–268, 4:271
social orders 4:266
theory and practice 4:271–272
utopian social order 4:266
Communicable diseases
and natural disasters 10:617, 16:316
see also Infectious diseases
Communication 2:204, 5:704, 6:665, 8:745–746, 12:95–96, 14:733, 16:194–195, 20:881, 22:246, 24:416–418, 24:605
and autism spectrum disorders 2:276
and broadcasting regulation 2:860–863
building social structures through 14:392–393
in collectivism/individualism 4:209
in combined pharmacotherapy and psychotherapy 19:497–499
critical theory 5:305–306
death/dying 5:867
defined 4:285
deixis *see* Deixis
diffusion of innovation in 8:379
digital forms of 4:186
doctor–patient *see* Doctor–patient interaction
emotion and 7:478
gender differences in 9:754–755
language use and conversational style 9:754
nonverbal behavior 9:754–755
nonverbal communication 9:754–755
nonverbal decoding 9:755
verbal ability 9:754
generalized media
and system processing 9:541
governance by and as 10:294–295
and health literacy 10:655
and historiography 11:99
history of 4:273–277
cities 4:274–275
cultural history 4:274–275
cursus publicus 4:274–275
development phases in Europe 4:274–275
economic history 4:274–276
film 4:273–274
finance 4:273, 4:276
historiographical trends, further research, and new fields 4:275–276
media 4:273–276
television 4:273–275
transport networks 4:274
travel 4:276
intentionality in 12:347–351
interpersonal 4:274
intrapersonal 4:274
and knowledge 13:70, 13:73
by language 14:241
and language development 13:319–323
and language evolution 11:329, 11:331–332
in Latin America 22:628
and logic models 14:299, 14:302
mass 12:85
media *see* Media
and media 12:84, 12:86, 15:11–16
media events 15:48–54
media-bound 4:274
medium of 4:263–264
mobile communication 15:629–635
mother–child *see* Mother–child communication
negotiators and 16:424
newborn 16:482–483
one-step flow of 14:712–713
with organizations 12:73–74
politeness as 18:328
political 18:342–349, 18:463–464
populism 18:609
of possibility of relation 14:200–201
and postmodernism 18:677–679
principle of 19:670
to recognition 5:308–309

Communication (*continued*)
and ritual 20:717–719
Rogers on 20:737–739
Schütz on 21:146–148
in science 21:163–164, 21:219
self-categorization theory 21:458
and semiotics 21:594–595
SES differences in 13:325
social media 22:421–425
social media use 24:146–147
social representations 20:444–446, 20:448
in Southern Africa 23:80
symbolic 11:255–256
symbolically generalized media of 14:387
telecommunication 12:83–84
computer 12:86–87
two-step flow of 4:291–293
emergence of the concept 4:291–292
Lazarsfeld on 4:291, 13:638
use of term 4:273
Communication, international 12:490
ancient 12:490
broadcasting 12:492
data communication 12:493
international telephony 12:492–493
Internet 12:493
medieval 12:490
postal 12:490–491
telegraph 12:491
wireless telegraphy 12:492
Communication about science 21:163–164
Communication Activities of Daily Living 3:794
Communication and information technologies, in intellectual property rights 12:262–263
Communication and media 5:704, 12:84, 12:86, 15:11–16
methodological limitations of statistical sources on 5:710–711
insufficient data availability of official sources 5:710
insufficient data of nonofficial sources 5:710–711
national regulatory authorities (NRA) for 5:705
nonofficial statistical sources 5:706
data on internet-based communication 5:710
national and supranational sources 5:708–710
official statistical sources 5:705–708
official statistics 5:707–708
statistical environment 5:707
statistical programs and user needs 5:706
policy 5:706
cultural policy 5:706
industrial policy 5:706
research on 5:704–705
statistical data for 5:704–705
Communication and Symbolic Behavior Scales Developmental Profile 3:795
Communication apprehension 24:161
Communication ethics 15:42
Communication infrastructure theory 4:333
Communication networks, postal systems 18:640–643
Communication of climate science 3:839
nongovernment organization campaigns 3:840–841
scientific reporting 3:839–840
Communication phenomenon, modernization as 15:712
Communication process, international advertising as 12:457–458
Communication research
history of 4:278–283
academic associations 4:278, 4:281
Cold War 4:278, 4:281
as empirical social scientific discipline 4:280–281
geography 4:279–280
historiography 4:278–279
institutionalization 4:279–280
journalism education 4:280–281
and media studies on east–west battlefield 4:281–282
social science 4:278, 4:281–282
speech 4:279–281
Zeitungswissenschaft 4:279–280
and information processing approach to cognition 4:103
Communication skills
doctor–patient interaction 6:618
in social work 4:284–290
asking questions 4:289
collectivism 4:284
content and process 4:286–287
culture 4:284–285
digital age 4:287–288
forms 4:286–287
high- and low-context communication 4:284–285
identity and self-awareness 4:285–286
kinesics 4:287
microskills 4:288
nonverbal communication and body language 4:287
observation, attending, and listening 4:288
power distance 4:284
proxemics 4:287
responding and reflecting thoughts and feelings 4:288–289
skills in advocacy work 4:290
specialty areas 4:289–290
summarizing 4:289
uncertainty avoidance 4:285
using interpreters 4:290
using silence 4:289
verbal communication 4:287
working across languages 4:289
working with people who have hearing impairments 4:289
written communication 4:287
Communication sounds, in nonhuman primates 11:323–324
Communication structures, of groups 10:424
Communication studies 2:357
Communication systems 13:554
disciplines as 21:287–289
Communication technologies 17:714, 21:648
death/dying 5:868–869
and globalization 10:262–263, 10:266
Communication theory 6:266
and information processing approach to cognition 4:103
Luhmann on 14:384–385
elements as events and operations 14:385–386
Communications 2:361
Communications Act 1934 2:860, 19:878
Communications Act 1996 2:860
Communications infrastructure, and nongovernmental organizations 16:858
Communicative action 10:787
Habermas on 10:469–471, 10:473
and pragmatism 18:810
Communicative behavior, crying as 16:475
Communicative competence 6:505, 14:172, 22:864
adaptability 4:299
as a characteristic of human mind 4:295–296
Chomsky's 4:296
communication and interpersonal interaction 4:297
as complex range of linguistic abilities 4:296
context of debates 4:294
degree of consistency of 4:299–300
development of 4:300
dichotomy of competence/performance 4:295
diversity of speech and knowledge of language 4:294–295
effectiveness and appropriateness as core criteria 4:298–299
in electronic age 4:300
empirical research 4:299–300
linguistic aspects 4:294–301
mastering a second language 4:296–297
nature of 4:299
language and its relation to communication 4:295–296
social and linguistic perspectives 4:297–298
use of term 4:294
Communicative cues, goal-oriented 10:271–272
Communicative discourse 15:775–776
Communicative dynamism (CD) 10:116, 10:116f
Communicative ethics 12:137
Communicative gestures 1:876–877
Communicative interaction 14:203–204
Communicative language teaching (CLT) 9:330
Communicative leadership 10:295
Communicative memories 4:188
Communicative openness, and adoption 1:144–145

Communicative paradigm of Frankfurt School 9:384–385
Communicative performativity 17:708
Communicative planning 3:627–628, 18:204
Communicative processes, and collective memory 4:190–191
Communicative rationality 4:790–791, 4:794, 12:915–916
Communicative reason, Frankfurt School 9:383–384
Communicative skills 4:299–300
Communicative technologies, and tradition 24:506
Communism 3:518, 4:302–307, 4:762–764, 6:372–373, 6:810–813, 14:642, 14:644–647, 15:772–773, 16:399, 18:187, 18:434–435, 18:438–439, 20:634, 20:845, 22:833–834, 22:837–838, 22:844–847, 23:366, 24:448–452
  communalism 17:11
  in Eastern Europe 6:805–808, 6:817–819
  and Eastern European social work 22:795–797, 22:799
  history of 4:302
    bolshevism 4:302
    Brezhnev doctrine 4:305
    class struggle 4:302
    collectivization 4:304
    Communist ideas 4:302
    ideas of 4:302–303
    political movement 4:303–304
    political system 4:304–306
    socialism and 4:302
    socioeconomic system 4:304–306
  and journalism 12:851, 12:853–854
  overthrow of 20:624–625
  and production 15:714
  and socialism 22:837–838
  twentieth-century 14:656–662
Communist China 3:676
Communist community 4:339
Communist democracy 6:369
Communist ideas
  of communism 4:302–303
  quasireligious element of 4:302
  in Soviet bloc 4:302–303
  in USSR 4:302–303
  in Yugoslavia 4:302–303
Communist International (Comintern) 14:656, 14:659–660, 18:408
Communist Manifesto 4:302, 14:650
Communist parties 18:404–409
  Communist International 18:408
  control 18:407–408
  democratic centralism 18:404–405, 18:408
  electoral support 18:408
  and Leninism 18:404–405
  and Marxism 18:404–406, 18:408
  post-wall 18:408–409
  post-war fortunes 18:407
  revolution 18:404–408
  in socialist societies 22:849, 22:851–852
  strategy 18:406–407
  Third International 18:405–407
  totalitarianism 18:405, 18:408–409
  world Communism 18:406
Communist Party of China 22:790
Communist Party of Italy (PCI) *see* Italian Communist Party (PCI)
Communist Party of the Soviet Union (CPSU) 12:278, 20:825–827, 20:844
Communist Revolution 5:368–370
  economic crisis and revolutionary euphoria 5:369
  institutionalization and foreign military adventurism 5:369
  last decade of the twentieth century 5:369–370
  origins of the revolution and leadership legitimization 5:368–369
  revolutionary dynamics 5:369
Communist single-party states, in Europe 22:140–142
Communist states, former states of Eurasia, population change in 18:545–552
Communitarian Forum 4:313
Communitarian ideas 24:791
Communitarian liberalism 4:309
Communitarianism 2:634, 3:657, 4:308–310
  atomistic individualism 4:308
  civic virtues 4:308–309
  and communalism 4:263
  conservatism 4:308–309
  cooperative ethos and 4:312
  criticisms 4:312
  definition of 4:311
  education and community-based learning 4:315–316
  emergence 4:312
  good society 4:309
  identity 4:308–309
  and intergenerational justice 12:399–400
  key principles 4:314
    principle of citizen participation 4:314
    principle of cooperative enquiry 4:314
    principle of mutual responsibility 4:314
  liberalism 4:308–310
  and liberalism 4:312–313
  and mass communication 14:694
  modernity 4:313–314
  political ethics 4:309
  problems 4:313–314
  Sandel on 20:893–898
  secularization 4:310
  self-interpreting animals 4:308
  shared values 4:311, 4:314, 4:316
  and social change 4:314–315
  social criticism 4:309
  social integration 4:308–310
  sociology of 4:311–316
  solidarity 4:311–312, 4:314–315
  spheres of justice 4:309–310
  values 4:308–310
  and Walzer, Michael 25:349
Communitarians 2:145–146, 2:152, 4:317–321, 4:808
  versus liberals 4:317–320
  and responsive community platform 4:320–321
Communitas 18:146
Communities 2:144–145, 2:147, 2:510, 4:317–321, 4:337, 7:106, 14:256–257, 15:370–371, 17:541, 22:263–265, 22:267, 23:639–640
  alternative 4:266
  and anthropology 1:753–755
  art 5:598–600
  butch–femme, evolution of 13:869–870
  and collective efficacy 4:169–170, 4:172
  communitatianism 4:311–316
  and crime 23:495, 23:497, 23:498–499
  definition 4:378, 10:646
  development 10:719–721
  diverse, social work with 15:419–420
  economy and 22:267
  entertainment 7:669
  and environmental justice 7:758–759
  and ethnicity 8:102
  exclusivity of 4:315
  external factors to control development 22:267
  and fear of crime 8:856–858
  freedom as attaining autonomy through 9:401–402
  health workers 10:718–719
  human–environment relationships 11:406
  and indigenous peoples 11:805
  intentional 4:268–269
  joining (life phase) 5:510
  and law 13:538
  levels, working with men at 14:683
  LGBT
    drag in 13:876–877
    friendship outside 9:456
    friendship within 9:454–456
  and libraries 14:48–49
  Maori social work 14:506–510
  Maroons 14:591–592
  meaning 4:354
  and media 4:332–336
  and morality 15:807–809
  and nationalism 16:262–265
  network cohesion and clustering 3:906, 3:908–911
  in North America 17:22–23
  and nostalgia 17:35–36
  participatory action research 17:535–538
  peasants 17:674
  political protest 18:422
  prevalence and correlates of mental disorders in 15:180–181
  public interest 19:552–553
  queering 19:754–755
  and radio 19:877–878, 19:880
  rational-romantic version of 4:338
  and reentry 20:63–64
  for rehabilitation 10:719
  religion and new media 20:274
  resilience 4:355
  resources
    psychosocial work 19:479–481

Communities (*continued*)
transgender friends as 9:455
role in disaster relief 10:722
role in intimate partner abuse prevention 12:679–680
rural sociology 20:814–818
and segregation 21:408
shamanism 21:868
and social capital 22:174–175, 22:177–179
social context, in United States 4:385–389
social structure 22:720, 22:722
and social work 22:132–137
active agency of 22:792
response for disasters 7:427–431
and society 4:337–343
sociolinguistics 22:949–950
solidarity of 23:6
studying methods 22:266
sustainable transportation 23:847–848
symbolic boundaries 23:850–851, 23:853
teaching 16:730
transnational 6:355
village community 17:670, 17:673–674
volunteerism 25:268–272
and youth 4:632, 4:635–636
Communities, and crime 5:159, 5:161–162
areas of continued research 4:324–325
community differences 4:325
crime as independent variable 4:325
poverty 4:325
role of work as stabilizing force in communities 4:324–325
crime and types of communities 4:325–326
immigrant communities 4:326
public housing communities 4:326
rural communities 4:325–326
theoretical understandings 4:322–324
collective sense of community 4:324
institutional support mechanisms 4:323–324
role of collective efficacy 4:324
social networks within communities 4:323
Communities of affiliation 4:387–388
Communities of faith 20:254
Communities of Learners 10:418
Communities of practice (CoP) 1:867–868, 4:328–331, 13:258, 15:83, 16:552, 17:304, 19:361–362, 22:6
definition 4:328
development stages and elements as organizing structure 4:328–329
in digital world 4:330
discourses around 4:329–330
domain knowledge 4:328–329
establishment and emergence of 4:329
and gender 9:699
identity 4:328–329
and language 13:286–287
role of leadership of 4:330
in school learning 1:871–872
and social learning systems 4:328
tacit knowledge 24:5–6
types 4:329t
virtual 22:8–9
distance education 22:8–9
game-based learning 22:9
Communities of specialists 21:288
Community action, exemplifying 4:381–383
Community Action Program (CAP) 1:225, 3:625, 18:734
Community action research *see* Community-based action research (CBAR)
Community attachment model, of media 4:332–333
Community attitudes, advocacy 1:223, 1:227
Community Behavioral Health Organization 19:590
Community benefit agreements 1:221–222
Community building, in China 2:68
Community capacity
for physical activity 18:94–95
effective partnering and coalition building 18:95
essentials of community organization and community development 18:94–95
program evaluation tools 18:95
Community care 19:587–589
affordability of 6:50
in England 4:344–349
direct payment 4:345–346, 4:348
disabled people 4:344–345, 4:347–348
domiciliary care 4:345, 4:348
efficiency 4:346–347
family care 4:344–345, 4:347–348
funding and fairness 4:347–348
home care 4:345–347
implications for social care workers 4:346
individualization 4:345–346
intermediate care 4:345, 4:347
intersectoral collaboration 4:346–347
older people 4:344, 4:347–348
overall impacts of efficiency 4:347
personal budget 4:346
personalization 4:344
policies 4:345
quasi-markets 4:345
reablement 4:347
rise of consumerism 4:345–346
structure and scope 4:344–345
for learning disability 13:684, 13:686
single shared assessments 12:502–505
Community computing 21:188
Community Connection initiative of New Jersey Online 3:614
Community corrections 20:31–36
Community design 16:809–810, 16:812–814
Community detection 3:910
Community development 16:450–455
and The Agenda 10:144–145, 10:148
for physical activity 18:94–95
Community Development Block Grant (CDBG) program 16:452, 18:736
Community development corporations (CDCs) 16:450, 21:919–920
emergence 16:453
Community development projects (CDPs) 4:381
Community ecology, of pervasive community violence 25:105, 25:107, 25:109
Community efficacy 10:646, 10:651
Community engaged research (CEnR) 20:519
Community engagement 21:102
and collective consent 9:965
in disasters 22:757–758
and museums 16:145–146
Community equity 7:926
Community Forestry Unit (CFU) 6:302–303
Community genetics 4:350–353
activities 4:350–352
areas of interest 4:350–352
genetic literacy and education 4:351
genetic screening 4:350–351
definition 4:350
delineation 4:352
clinical genetics 4:352
population genetics/genomics and genetic epidemiology 4:352
public health genetics and genomics 4:352
in disadvantaged population 4:351
evolution 4:350
issues
economic 4:352
epidemiologic 4:351
ethical and legal 4:351–352
policy 4:352
psychosocial 4:351
public consultation about 4:351
in LMI countries 4:351
origin 4:350
in preconception care 4:351
in primary care 4:351
registries of genetic/congenital disorders 4:351
Community genetics journals 4:352
Community genetics network 4:352
Community genomics *see* Population genomics; Public health genomics
Community health 6:879
Community health workers, psychiatric care 6:44–45
Community Innovation Surveys (CIS) 12:154, 12:156–157
Community intervention projects, informed consent in 12:113
Community Intervention Trial for Smoking Cessation *see* COMMIT study (Community Intervention Trial for Smoking Cessation)
Community life, of Roma people 20:748–749, 20:751–752
Community Lodge program 18:884
Community members, and social research 22:270
Community Mental Health Teams (CMHTs) 19:303

Community networks 4:323
Community of historians 11:122
European 11:122
German 11:122
Community of learners 7:579
Community organization, for physical activity 18:94–95
Community outreach and engagement, Genographic Project 10:24
Community ownership, hospitals 11:210–211
Community Panel
Liberating our Children, Liberating Our Nations 11:765
Community participation 3:625
in health 4:401–403
Community perspective
of tacit knowledge 24:5
Community policing 18:253, 18:271, 18:879
Community politics, after Foucault 18:767–768
Community Power Debate 18:766, 18:768
Community property system 19:174
Community psychiatric nurses (CPNs) 19:302–303
Community psychiatry 22:380, 22:382–384
Community psychology 18:884
applied behavioral analysis for 2:410
Community punishment 19:46
Community reinvestment 16:452
Community Reinvestment Act 1977 (CRA) 16:452, 20:540–541, 24:930
Community relationships, early 9:451
Community research
ethics 20:509–512
Community resilience 4:355, 10:547–548, 20:565
and disaster 4:355–358
localization 4:358
and natural hazards 4:357
and nonviolent social crisis 4:358
and technological hazards 4:357–358
and violent social crisis 4:358
Community resilience to disasters 4:354–360
Community safety 5:148
Community schooling 8:612
Community Schools v. Seattle School 13:768
Community services 15:778–779, 19:46
psychiatric care 6:42, 6:44–45
Community sociology 4:361–368
Chicago School 4:362–363
class 4:363–364
community studies 4:363
cultural turn 4:365–366
early community studies 4:362
ethnicity 4:363–364
gender 4:363–364
globalization 4:366–367
market 4:364–365
neo-liberalism 4:364–365
new attitudes 4:364–365
power 4:363
spatial turn 4:366
tradition 4:363
Community studies 9:360–361, 14:203
anthropology 4:369–371
community in social sciences 4:369
critical anthropology 4:370–371
entry/reentry of community 4:369–370
globalization 4:370–371
migration studies 4:371
rural populations 4:370–371
solidarity 4:369–371
territorial units 4:369
transnationalism 4:370–371
and ethnography 8:179
in Mexico 1:629
peasants 17:676–677
Community supervision
definition 4:372
in United States 4:372–377
Community supported agriculture (CSA) 1:566–567, 1:570, 23:810
Community Toolbox, use of logic models in 14:303
Community treatment orders (CTOs) 19:303
Community union 7:574
Community violence 25:107
Community vitality 8:614
Community work 4:378–384, 22:522–523
definition 4:378
emancipation 4:381
empowerment 4:381–382
exemplifying 4:381–383
community development in Central Europe 4:382
community development projects 4:381
community youth development 4:382
globalization and community development 4:383
locally based community development 4:381–382
NGOs and community development 4:382
social pedagogy and community development 4:382–383
marginalized communities 4:378–381, 4:383
power relations 4:380–381
social change 4:378–380
social justice 4:378
social work education 22:775–776
theories and theorists 4:378–380
black and antiracist theories 4:380
feminist community work 4:379–380
pluralist community work 4:379
radical and socialist community work 4:379
thinkers 4:380
Alinsky 4:380–381
Freire 4:381
Gramsci 4:380–381
Community youth development (CYD) 4:382
Community-based action research (CBAR) 22:271–272
Community-based agencies 20:256
and religious congregations 20:256
Community-based audit 21:253
Community-based disaster risk reduction (CBDRR) 4:358–359
Community-based health interventions 10:646–652
barriers and challenges 10:650–651
community and state actions changing national policy 10:650
community efficacy 10:646, 10:651
community projects in other health topics 10:649–650
comprehensive community-based health education for CVD prevention 10:648–649
early international diffusion 10:649
first decade 10:648–649
projects begun in 1980s and 1990s 10:649
digital health 10:647–648, 10:651
master plan 10:650
multilevel 10:646–647, 10:651
new opportunities for greater cost-effectiveness 10:651
paradigm shift 10:646, 10:651
principles, methods, and resources 10:647–648
acceptance 10:647
comprehensive 10:647
dose 10:648
effective message 10:648
message design and audience segmentation 10:647–648
role of activism and advocacy 10:648
role of policy, laws, and regulations 10:648
variety, integration, and timing 10:647
public health policy 10:650
rationale 10:646–647
social marketing 10:646–651
theoretical formulations 10:647
Community-based interventions
social work 3:472–473
Community-based learning, and communitatianism 4:315–316
Community-based participatory research (CBPR) 17:543, 23:324, 23:329, 25:106–107
see also Community-based action research (CBAR)
Community-based programs, for skin cancer 23:698
Community-based research, in applied sociology 1:858–859
Community-defining consensus 5:504
Community–labor alliances 24:492
collaboration 24:491–492
Community-level theories 10:625
Community-oriented primary care (COPC) model 4:402
Community-scale technology 18:199
CommunityViz 24:891
Commutativity condition 21:606
Commuting
and spatial interaction 23:154–155

Commuting (*continued*)
and spatial mismatch 23:157–160
Commuting paradox 15:636–637
Commuting stress 24:627
Comorbid mental disorders, behavioral medicine for 2:467–468
Comorbidities
of disorders 6:389
and elder care 7:337
of gastrointestinal disorders 9:608–609
mental–physical 4:390–393
of panic disorder/agoraphobia 17:476
seizures 7:854–855
Compact cities 24:793
Compact Muon Solenoid (CMS) 18:104
Compagnonnages 2:49
Companies 3:24–25
role in industrial ecology 11:845–846
survey research 23:747–751
Companionate love 14:377–380
Companions 7:586–588
Company control, stock market 23:490
Company law 4:932–933
Comparable corpus 4:950–951
Comparable worth analysis, in discrimination 6:534
Comparable worth policy 9:731
Comparativ: Zeitschrift für Globalgeschichte (journal) 11:45
Comparative advantage 3:23
Comparative analysis
culture production 5:604–605
comparisons across symbol-producing realms 5:604
cultural mobility 5:605
high *versus* low forms 5:604–605
prestige, changing criteria of 5:605
higher education research 10:867
Comparative anatomy 14:367
Comparative and International Education Society (CIES) 7:146–147, 7:182
Comparative anthropology 2:567, 2:569–571
Comparative capitalisms 7:69
economic sociology 7:69
Comparative constitutional law 4:704
Comparative constitutionalism
Aristotelian notion of constitutionalism 4:706
diversity 4:704
global context 4:705–706
globalization and 4:704
interdisciplinary nature 4:704
issue of governance 4:707–708
normative context 4:704
politics of comparison 4:706–707
practice and theory of 4:708
sociolegal approach 4:706–707
Comparative contexts, population censuses in 3:294–295
Comparative criminology 12:462–469
Comparative cultural sociology 5:618–619
Comparative databases
availability of data 5:790
data archives 5:791
data harmonization 5:795
data infrastructure 5:791–792
definition 5:790
European Values Study (EVS) 5:791–792
creation of 5:791
creation of WVS 5:791
evolution of 5:791–792
international standardization 5:796
international survey programs 5:792
longitudinal data 5:791–792, 5:795
macro data 5:789
obtaining 5:790
methodological problems 5:795
micro data 5:789
obtaining 5:790–791
need for comparative research 5:789
organizational challenges 5:795–796
research infrastructures 5:790, 5:793, 5:796
selected individual- or household-level databases 5:792–793
electoral systems 5:793
Eurobarometer 5:792–793
European Social Survey 5:793
Gallup World Poll 5:792
International Social Survey Programme 5:792
other barometer-type surveys 5:793
World Health Organization surveys 5:793
supranational organizations and international official statistics data 5:794–795
topical macro databases from academic projects 5:793–794
Comparative Manifestoes Project 5:794
Correlates of War 5:793
European Data Center for Work and Welfare (EDAC) 5:794
IDEA project 5:793–794
Transparency International Index 5:793
World Database of Happiness 5:794
Comparative demography 8:418–419
across species 8:418
measures, methods, and data 8:418
widely occurring traits 8:418–419
Comparative economics, world systems theory as supplement to 25:749–750
Comparative education 4:430–431, 7:211–212
analytical research 4:430–431
model of worldwide activity in 4:428–429
Comparative effectiveness research (CER) 8:330
for therapies 24:281–283
Comparative efficiency test 8:265
Comparative evolutionary psychology 4:394–399
adaptations 4:394–395, 4:398
animals 4:395–397
anthropocentric 4:395, 4:397
biases 4:396–397
bridging gaps 4:398
cognition 4:395–396
continuity 4:394–398
evolved traits 4:394, 4:398
goals and limitations 4:394–395
comparative psychology 4:394–395
evolutionary psychology 4:395
human 4:394–396
uniqueness 4:395
need for evolutionary framework 4:395–396
'null' results, 'problem' of 4:397–398
Comparative fit 22:187–188
self-categorization theory 21:456–457
Comparative fit indices (CFI) 23:618
Comparative health systems (CHS) 4:400–404, 15:100–101
issues in 4:400–403
aging 4:401
changing demographics and emerging/re-emerging diseases 4:401
community participation in health 4:401–403
decentralization 4:400–401
health financing and expenditure 4:401
health policies 4:400–403
NGOs, role of 4:401
overview of 4:400
research, benefits and limitations of 4:401t
Comparative history 4:405–410
agenda and unsolved problems 4:409–410
disciplinary and national patterns 4:406–407
of emotions 7:513, 7:517
of European industrialization 11:104
functions of 4:407–408
of gender and material progress 10:166–167
methodology 4:408–409
and prosopography 19:258–259
and social mobility 22:428
Comparative interrupted time series designs 19:744–745
Comparative judgments, in forced-choice personality assessment 17:843–844, 17:847–848
Comparative law, and constitutions 4:712
Comparative legal change 13:568
Comparative linguistics 25:288–289
Comparative Manifestoes Project 5:794
Comparative methods, in evolutionary studies 4:415–422, 20:629, 20:631
adaptation in 4:417–418
in anthropology 4:411–414
binary comparison 4:412
cognition 4:413
controlled comparison 4:412–413
current debates 4:413–414
distant comparison 4:413
fluid comparison 4:413
functionalism 4:411–412
history 4:411–412
new pluralism of procedures in 4:412–413
particularism 4:411–412
regional comparison 4:412
temporal comparison 4:412–413
universalism 4:412

body, of organism 4:415–416
brain size 4:415–416
continuously varying traits 4:416–420
continuous-time Markov-transition process 4:420–421
evolution of political complexity 4:420–421
genotypic sex determination and adaptive radiations in the sea 4:421
dinosaurs genome size 4:417–418
evolution 4:416–419
generalized least squares approach 4:416–417
GLS regression 4:417
mammal brain and body size 4:417
in linguistics 4:423–426
ancestral language 4:423–426
application 4:424–425
assumptions and limitations 4:425
Indo-European cognates 4:424t
language families 4:423–424
reconstructed proto-languages 4:425–426
mammals 4:415–417
natural selection of 4:416
new developments 4:421–422
phylogenetic 4:416–417, 4:420
phylogenetic prediction
dinosaur genome size 4:417–418
feeding time and the evolution of food processing 4:418–419
phylogenetic signal 4:419–420
for revolutions 20:647
Comparative Party Manifestoes Project 5:833
Comparative personal risk perceptions 10:704–706
Comparative phenomenology of religion 20:346–347
Comparative philology 1:746
Comparative politics 12:181–182, 21:305, 24:40–41
Comparative prosopography 19:258–259
Comparative psychology 14:369
and perceptual learning 10:127–128
Comparative public policy, and agenda-setting 1:360
Comparative reconstruction 8:208, 11:824
Comparative religion 7:375, 20:322–323, 20:387–388
Comparative research
database 5:832
in education 4:436–437
ISSP 12:556
Comparative science of religion 20:322–323
Comparative social work 12:558–559
Comparative societies, Meiji Revolution 15:129
Comparative sociology 1:751
leisure and 13:855–856
Mann on 14:490–495
Comparative studies 4:439–444, 18:891–894
changes 4:440
counterfactual examples 4:444
definition 4:439
future directions 4:444
homogeneity 4:441
intellectual context 4:439–440
'internal' and 'external' validity of any scientific findings 4:440–441
methodological issues and problems 4:443–444
Mill's canons 4:440
Mill's indirect method of difference 4:441
opposite strategy 4:441
of time use 24:370–371
Comparative Study of Electoral Systems (CSES) 5:833
Comparative survey research 5:827–828
Comparative urbanism 10:50–51
Comparative/contextual field experiments 9:131
Comparative-historical conflict theories 22:957
Comparative-historical research on education 7:176–177
Comparative-historical sociology, contribution of Eisenstadt to 7:321
Comparativism, and postcolonial geography 18:645
Comparativists 7:374–375
Comparison 6:430
anthropology by 1:774–775
infancy as basis of 12:10
and reproduction 20:453–454
Comparison group experimental designs 22:293–294
Compartmental models, for social analysis 14:580–581
Compartmentalized time 24:410, 24:412–413
Com.pass model 17:571
Compassion, cognitive-anatomical basis of 16:612
Compassion fatigue 18:86, 22:758
in social work 18:710
Compassionate giving 3:365
Compassionate love 14:377–378, 14:380
Compassionate self-inqueery, queer anarchism 19:750
Compatibilism 9:405
Compatibilists 9:405
Compensation fund 12:216–217
Compensation law of mortality 6:756–757
Compensation of employees 12:519
Compensation systems, in law firms 13:578–579
Compensation theory, innovation and employment 12:152–157
Compensations 1:346, 1:349, 12:386, 13:448, 17:611–612, 20:56
advertising agencies 1:191–192
cognitive aging 4:12–13
and conflict 4:624–625
and employee performance 17:612–613
executive compensation 17:614–615
extrinsic compensation 17:612
goal setting 17:612
intrinsic compensation 17:611–612
for research participation 20:517
strategic compensation 17:614
workers 20:864–865
working conditions 25:701
Compensatory damages 20:414–415
award levels and trends of 20:414–415
awards reliability of 20:415
and general adequacy of damages 20:414
time trends of 20:415
Compensatory primary control 4:814–815
Compensatory secondary control 4:814–815
Compensatory stealing 19:162–163
Competence of juries 12:910
Competence stereotype 21:680–681
Competence trap 23:511
Competence-based deletion, idea of 19:59
Competence-Environment Press Model, of aging 6:887
Competence-Press Model of aging 6:885–886
Competence-related beliefs 1:21–22
change in mean level of 1:21–22
children's understanding of 1:21
Competencies 5:860, 7:215–216, 15:931, 17:767–768, 21:487, 21:489–490, 24:68
adequacy, in domains of importance 21:494
areas of 8:669
communication of emotion 8:669
emotion regulation 8:669
response preparation 8:669
beliefs, and epistemological beliefs 7:891
Bologna Process 2:742–746
and citizenship 3:656
control beliefs 4:819–820
cultural 22:822–823
equality of opportunity and 7:904–905
mathematical 14:818
personal intelligence 17:754
social work 5:516, 5:519
in sustainable practices 23:836–837
Competency to stand trial, of forensic psychologist 9:351–352
Competing concepts 23:728
Competing motivations 25:646–647
in functionalist grammar 9:474–475
Competing risks 14:92
Competing risks model 8:312, 17:174, 8:311f
Competing values framework (CVF) 17:328–329
Competing values model 17:323
Competition 4:856–861, 7:114–115, 12:146, 14:549–554, 18:770, 22:264, 22:266
achievement 4:857–859
and agreeableness 1:465–466
among experts, public bureaucracies 19:532
antitrust policy *see* Antitrust
artists 2:52–59
basic elements 4:859
behavioral learning theory 4:856–857

Competition (*continued*)
and broadcasting regulation 2:861–862
cognitive developmental theory 4:856
collaboration 4:856–857
competitive and individualistic efforts 4:859
cooperation 4:864–865
cooperative learning 4:859–860
definition 4:621
enhancing factors 4:859
in entrepreneurship 7:675–676, 7:679
and equilibrium 7:913, 7:915–916
and ethnic conflict 8:93
evolutionary economics 8:430–431
gender differences in 9:753–754
in aggression 9:753–754
in females 9:753
in males 9:753
in social play 9:754
of green parties 10:383
health care markets
effect on quality 10:600
managed competition 10:598–599
and innovation 12:165–166
interaction patterns 4:857–858
interpersonal relations 4:858–860
and market liberalization 13:159
multinational corporations 16:78–79
oppositional interaction 4:857–858
outcomes 4:858–859
in party systems 17:568
private health insurance 10:615–616
promotive interaction 4:857–859
psychological health 4:858, 4:860
psychological states 4:858
in public sector 16:770
scope of 16:774
use of 16:774
role in modern society 3:17–18
in schools 7:154–155
social cognitive theory 4:857
social interdependence theory 4:857
application of 4:859–860
breadth of 4:860
relationship among theory, research, and practice 4:860
in supply chain management 23:711
trade 12:576
in urban planning 24:882–885
for limited areas and special functions 24:885
major large-scale and city center competitions 24:883–885
Competition as a Cultural Phenomenon 14:497
Competition law 3:31
Competition models 13:313–314, 16:443, 21:363
Competition policy, communications and media 5:706
Competitive advantage
absorptive capacity 1:12–13, 1:15, 1:17
and interorganizational relationships 12:624
marketing 14:555–557
Competitive altruism
evolutionary group dynamics 8:443–444
Competitive analysis, for strategic planning 23:520
Competitive balance, sports economics 23:319–320
Competitive cities 24:792
Competitive cooperation between superpowers, during Cold War 4:132
Competitive corruption 5:14–15
Competitive elaboration, in realism 20:24
Competitive jungle worldview 22:250
Competitive particularism 5:14–15
Competitive price theory 12:100
Competitive search equilibrium 21:337
Competitive strategy analysis 23:522
Competitiveness, in urban megaprojects 24:838–841
Competitiveness and Security: Direction for Canada's International Relations 21:397
Complement control 4:832–833
Complementarity principle 12:404, 16:68, 18:501
Complementary and alternative medicine (CAM) 10:562–566
acute illness 10:565
chronic illness 10:564–565
controversy 10:562–563
cure 10:564–565
and medical profession 10:564
and orthodox medicine, consultation differences between 10:564–565
range 10:563–564
reasons for using 10:565
research in 10:564
rise of 10:562
types patients seeking 10:565–566
Complementary binding strategies 16:634–635
Complementary DNA (cDNA) 17:589
Complementary feeding 2:848t
Complementary healing practices *see* Complementary and alternative medicine (CAM)
Complementary learning systems 13:922
Complementing emotions 16:658–659
Complements 9:875
business platforms 3:37–42
Complete ADISOP (CADISOP) model 24:253
linear 24:254
Complete balanced structure 23:606
Complete cohort life table 14:91
Complete contracts 1:348
Complete independence, model of 23:618
Complete information 12:103
Complete linkage 23:393
Complete membership 17:532
Complete models, for urban systems 24:809
Complete spatial randomness (CSR) 23:96
Complete-case analysis 13:389, 15:603
Complete-case and weighting methods 15:603
Complete-data framework 4:470–471
Completely randomized designs 8:537–538
Completeness, assessment and adjustment of 6:128–129
Completeness theorem 16:924–925
Completing sentences test 19:671
Completion of orders 17:528
Complex adaptive systems 23:368–369
Complex domination 2:752
Complex emergencies 3:714–715
Complex equality 12:938
Complex genetic disorders 9:915, 9:917–918
Complex humanitarian emergency 20:112–113
Complex hunter–gatherers 17:19
Complex institutions, majoritarianism in 14:447
Complex integrative models, hysteria 11:509
Complex movement-derived sentences 16:693
Complex multistage sampling designs 13:386
Complex NP Constraint (CNPC) 16:4
Complex post-traumatic stress disorder (CPTSD) 24:436–437
Complex relationships, categorization of 23:117–118
Complex sampling designs 23:617
Complex societies
and dowries 6:642
Oceania 17:448–449
Complex sounds 23:243
hierarchical processing of 23:243–244
Complex systems, modularity and 15:718–720, 15:722–723
Complex traits
epigenome-wide association scans for 24:693–698, 24:694t–697t
power to detect epigenetic changes in 24:692–693
Complex trauma 16:200–201
Complexity 3:612, 6:612, 6:666–667, 22:459–460, 22:691–692, 22:914
of algorithm 1:530–531
disasters 10:548–549
and ergonomics 11:302
of extraversion 8:625–626
in high reliability organizations *see* high reliability organizations (HRO
in LAC educational system 7:178
mathematical models 14:794, 14:798
reasoning 5:937–938
social science of 22:915–916
social systems, Luhmann 14:384, 14:386
and systems theory 9:541
Complexity intervention units 14:6
Complexity theory 2:303, 2:305
cybernetics 5:647
organizational design and form 6:235, 6:237
Complexity thesis 20:284
Compliance 4:445–455, 8:527–530, 13:751, 13:823, 13:826
and adolescent personality 17:866
antisemitism 1:784–785

as-treated estimate 4:446, 4:450
causal inference 4:452
contract and 4:804–807
counterfactual model 8:528
modification of 8:529
extrapolating the CACE to full population 4:452–453
in-house counsel 12:126–128
IV approach 4:446–451
censoring 4:450–451
framework and notation 4:446–448
missing response data 4:451
PROSPECT study 4:450, 4:450t
randomization as 4:447f
relation between observed groups and latent compliance classes 4:447t
treatment effect on whole distribution of outcome 4:449–450
two-stage least squares (Wald) estimator 4:448–449
with legal authority
boundary conditions for procedural justice and legitimacy 4:458–459
deterrence and social regulation 4:456–457
future directions 4:458–461
integrating procedural justice theory with criminological accounts 4:460–461
legitimacy 4:460
normative modes of social regulation 4:457
procedural justice 4:457–458
and moral development 15:788–790
multiarm trials 4:453–454
noncompliance 4:445, 4:453–454, 4:445f
addressing in practice 8:530
difficulties with 8:527
as underdosing 8:529–530
organizational 17:397, 17:400–402
partial compliance 4:454
of patients 17:598
per-protocol analysis 4:446, 4:450
predictors of 4:453
and self 21:442
sensitivity analysis for violations of assumptions 4:452
for ER 4:452
for Violation of the Monotonicity Assumption 4:452
simple noncompliance 8:528–529
bounds for ACE using instrumental variable 8:529
systems 5:204
tax 15:738
see also Noncompliance
Compliance and legal authority 4:456–462
Compliance gaining, *versus* social influence 22:348
Compliance therapies, explanatory style 8:598
Compliance-oriented environment 11:697
Compliant helping 19:221–222
Complicated grief (CG) 2:532–534, 2:538
across life span 2:537
diagnosis 2:534
intervention efficacy 2:533–534
Complicated grief disorder (CGD)
anxiety 2:537–539
cognitive restructuring 2:540
criteria 2:540t
depression 2:537–541
diagnostic criteria for 2:539
emotion regulation 2:538, 2:540–541
life event 2:537
post-traumatic stress disorder 2:538–541
psychological intervention 2:539–540
psychotherapy 2:539–540
treatment of 2:539–540
Complication model 17:918
Complicit masculinities 14:681
Complicity
meaning of 14:640
in multisited field studies 16:107
Complier average causal effect (CACE) 9:131
extrapolating to full population 4:452–453
Component densities of mixture 15:624
Component literacy skills 3:508–509
Component methods 6:165
Componential analysis, of lexical semantics 13:931–932
Componential models, for facial behavior 7:522
Components in smooth testing
Cholesky components 23:401
diagnosis 23:400–401
Composite indicators 11:751
Composite International Diagnostic Interview (CIDI) 15:182–183
Composite measures, and cultural capital 5:377
Composite Memory Index 16:725
Composite method 6:165
Composite portrait 3:191
Composite positions, of personal identity 17:743–744
Composite scales, in social measurement 21:5–6
Composite variable, lower bounds to reliability of 19:428–429
Composition
household wealth 25:453–454
writing 25:768–770
Composition property 10:346
Compositional process maps 21:621
Compositional semantics 13:80–81, 16:674–675
Compositionality, principle of 15:750–751
Composure 17:286–287
Compound formation 23:716
Compound knowledge 3:511–512
Compounding 15:822
Comprehending consciousness 10:738
Comprehension 7:247–248, 7:250–251, 20:516–517, 20:881, 7:248f
question 19:771–772
sign language 21:935
Comprehension disorder 1:819–820, 1:820t
Comprehension disruption 21:622–623
Comprehension of media content, children 15:35–36
Comprehensive, accessible, responsive, individualized, transdisciplinary, accountable, and systemic care *see* CARITAS (comprehensive, accessible, responsive, individualized, transdisciplinary, accountable, and systemic care)
Comprehensive African Agricultural Development Program (CAADP) 9:309
Comprehensive Assessment of Spoken Language 3:795
Comprehensive community-based health interventions 10:647
Comprehensive Crime Control 5:236
Comprehensive Crime Control Act 1984, Title V 5:236
Comprehensive disability rights 13:501
Comprehensive education 22:328
Comprehensive egalitarianism 7:310–311
Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) (US) 12:213
Comprehensive examinations
for aphasia 3:793–794
for language disorders 3:795
Comprehensive School Reform (CSR) projects 21:81
Comprehensive sexuality education 21:825
Comprehensive theories of social phenomena 22:486
Compression 15:638
of morbidity 1:394, 4:463–468, 23:440
Compromise 6:558
in conflict/consensus 4:591
and long-term heterosexual relationships 20:912–914
Compromise kin group 3:732
Compromise option 19:444
Compromise value 4:878
Compromised learning 20:9
COMPSTAT *see* Computer Statistics (COMPSTAT)
Compulsions *see* Obsessive–compulsive disorder (OCD)
Compulsory education laws 3:480, 23:219
Compulsory heterosexuality 21:809
Compulsory insurance 25:476
Compulsory power 18:763–764
Compulsory retirement insurance 14:41
Compulsory schooling 7:168
Compulsory Sterilization Law 8:241
Compulsory voting 25:309–310
Computability 4:87–88
Computable general equilibrium (CGE) analysis 10:373–374, 10:377
Computation
approaches 2:30–31
to model evaluation 4:475–479
artificial/natural 2:27–31
by computers 4:516–517
evolutionary 2:30

Computation (*continued*)
expectation-maximization algorithm 4:469–474
managerial and organizational cognition 14:480
mathematical notion of 2:27
Computational approaches
to multisensory integration 16:101
Computational approaches to model evaluation 4:475–479
Bayesian model selection 4:476–477
good model 4:475–476
generalizability criterion 4:475–476
model complexity criterion 4:475
minimum description length 4:476
model evaluation at front end of experiment 4:478–479
adaptive design optimization 4:478
design optimization 4:478
other evaluation types 4:477–478
parameter space partitioning 4:477–478
prequential approach 4:477
statistics 4:476–477
Computational auditory scene analysis (CASA) 2:223–228
applications, evaluations, and future challenges 2:227–228
autocorrelation 2:223–226
glimpse decoder 2:227
grouping 2:224–227
pure model-driven grouping 2:226
signal-driven grouping 2:224–226
target source identification in unknown background 2:226–227
music 2:226–227
perceptual organization 2:223
signal decomposition 2:223–224
source separation 2:223, 2:225–226
speech 2:223–227
Computational cladistics 11:26
Computational cognitive neuropsychology 4:92
Computational complexity 21:583
geographic models 14:789
Computational historical linguistics 8:375
Computational humanities 9:365
Computational implications, neurons 16:706
Computational intractability, and heuristics 10:829, 10:831
Computational linguistics (CL) 4:949, 9:364–365
applications of corpus 4:951
corpus processing tools 4:951
probabilistic grammars 19:9, 19:11
trends and future 4:951
Computational linguists (CL) 21:581
Computational maps 4:88–89
see also Topographic maps
Computational models 4:676
cognition 4:99–100
of cognitive control 4:34–35
of lexicons 13:926–928
in neuropsychological rehabilitation 2:817
holistic approach 2:817
of scientific discovery 21:291–293
ARROWSMITH 21:294–295
BACON 21:294
data-driven and theory-driven approaches 21:292
future directions 21:295–296
general approaches 21:292
KEKADA 21:294
LAGRAMGE 21:295
limitations 21:292–293
MECHEM 21:295
PRET 21:295
representation 21:293
scope 21:292
tasks 21:293
tools for 21:293–295
in visual word recognition 25:216–217
Computational models of scientific discovery 21:291–296
Computational neuroscience 1:428, 1:431–433, 2:369, 4:480–484
algorithms 4:480–482
cognitive models 4:482
future directions 4:482–483
brain modeling technology 4:483
map/system 4:483
molecular/synaptic 4:482–483
neuron/network 4:483
investigation levels for studying brain 4:480f
learning and memory 4:482
levels of analysis 4:480f
realistic models 4:481
signal processing models 4:481–482
Computational neuroscience, Marr on 14:597–598
Computational organization theory (COT) 4:485–492
bounded rationality 4:488
cognitive models 4:485
current research 4:488–490
multilevel modeling 4:489
multimodeling and model interaction 4:488–489
rapid model development 4:489–490
defining organization 4:486
dynamic network analysis 4:488–489
model fidelity 4:490–491
model validation 4:490
multiagent stimulation 4:486–488
elements in environment 4:487–488
representations of space 4:487
representations of time 4:486–487
network-centric simulation 4:485
organizational models 4:485, 4:487, 4:487f
Computational phylogenetics 11:26
Computational procedures 9:626
Computational processes, in inferencing 19:396–397
Computational semantics 15:754–755
Computational social science 21:530
Computational sociology 9:365–366
Computed tomography 11:603
Computer assisted instruction (CAI) 4:493–497
behaviorism 4:495
cognitivism 4:494–495
constructivism 4:495–497
determining effects on students 4:496–497
information and communication technology (ICT) in education 4:495–496
current use of ICT in schools 4:496
ICT-enhanced learning environments in curriculum 4:496
integration in educational practice 4:496
IT in education, research on 4:497
need for new conceptualizations of use of 4:493
River City Project 4:494
software applications for educational use 4:493–495
augmented reality (AR) 4:495
data logging and modeling 4:495
drill and practice software 4:494
graphics software 4:493–494
multimedia cases 4:494
presentation software 4:493–494
simulations, haptics, games, and virtual worlds 4:494
spreadsheets 4:493–494
tutorials 4:494
Web 2.0 4:494–495
word processor 4:493–494
technological pedagogical content knowledge 4:496
Computer games 4:494
Computer games/gaming 20:9–10
and academic achievement 4:498
academic problems 4:499–500
attention problems 4:498–499, 4:501
mass media 4:498–500
pathological gaming 4:499–500
potential negative influences on 4:499
potential positive uses of 4:501
proposed mechanisms 4:498
Computer models, and cognitive psychology 4:103–104
Computer packages, time series analysis 24:345
Computer programs 21:291
Computer Scientists for Social Responsibility (CPSR) 21:188
Computer simulations 22:915
analytical sociology 1:672
Prisoner's Dilemma 18:935
Computer Statistics (COMPSTAT) 24:120
Computer technology 5:819
and the internet, globalization 22:994
role in self-regulated learning 21:544–545
Computer usage
learning and cognitive development 8:732
Computer-aided personal interviews (CAPI), time use research methods 24:381
Computer-assisted learning (CAL) 13:736, 20:9–10
Computer-Assisted Qualitative Data Analysis Software (CAQDAS) 19:674, 19:685–686

Computer-assisted telephone and personal interviews 10:729
Computer-assisted telephone interviewing (CATI) 18:461–462
  time use research methods 24:381
Computerization 23:731, 25:691
  impact of 21:187
    military roots and role of activism 21:187–188
Computerized adaptive testing (CAT) 19:436
Computerized assessments 6:346
Computerized cognitive behavioral therapy (CCBT) 19:297–298
  economics of psychiatric care 19:297–298
Computerized test construction 4:510–514
  adaptive tests 4:512
  algorithms for 4:511
  as combinatorial optimization problem 4:510–511
    model 4:511
    objective functions 4:510–511
    restrictions 4:511
  linear test forms 4:511–512
  multistage tests 4:512–513
  robust 4:513
Computerized tomography (CT) 11:615
Computerized training, and language learning impairment 13:349
Computer-mediated communication (CMC) 3:51–52, 4:504–509, 6:39–40
  communication perspective 4:504–505
  cues-filtered-out theories 4:505–506
  email 4:504–505
  Facebook 4:504–505, 4:508
  future directions 4:509
  hyperpersonal model 4:507
    channel 4:507
    feedback 4:507
    receiver effects 4:507
    sender effects 4:507
  media richness theory 4:505–506
    empirical findings 4:505–506
  online dating 17:213–215
  online information evaluation 4:507–509
  participatory websites 4:508
    aggregated user representations 4:508
    juxtaposition among sources 4:508–509
    proprietor content 4:508
    user-generated content 4:508
  social identification 4:506–507
  social identity model of deindividuation effects 4:506
  social information processing theory 4:506–507
  social media 4:504
  social presence theory 4:505
  theories of interpersonal adaptation 4:506–507
  Twitter 4:504
  warranting 4:507–508
Computer-mediated public health programs 10:688
Computers 4:515–518, 21:186–191, 24:889
  and 2:817
  from calculation to computation 4:516–517
  cognitive revolution 4:516
  contribution of Babbage to 4:515–517
  contribution of Simon to 4:516–518
  digital lexicography 13:940–941
  division of labor 4:515–516, 4:518
  in experimental laboratories 8:556
  as factory of workers 4:515
  hierarchical organization of mind 4:515–516
  in information society 12:84–86
  metaphor 4:515–518
  research and teaching assistance 4:518
  revolution 4:518
  telecommunication 12:86–87
  and telephone 24:158, 24:161
  from thought experiments to stimulation 4:517
  world of numbers and statistics 4:517–518
Computer-supported cooperative work 21:197
  infrastructure 12:120
Computing
  integrative visions 21:190
  pervasiveness of 21:186
  questionnaires judgment 19:772
  Scandinavian School 21:188
  in society 22:831
    social networks 22:831
    social Web 22:831
  standards and design 21:188
  Ubicomp 21:189
Computing Tabulating Recording Company 23:376–377
COMT *see* Catechol-O-methyltransferase (COMT)
Comte, Auguste 4:339, 4:519–524, 5:677, 6:241, 8:292, 13:112, 15:576, 18:33, 18:619, 18:625, 19:410–411, 19:782, 22:696–697, 22:996, 23:2
  assessment and influence 4:523–524
  and enlightenment 18:620
  founder of sociology 22:991
  on functionalism 9:536
  hierarchy of sciences of 21:241
  on history of science 4:520–521
  life and main works of 4:519–520
  on logical positivism 18:619
  positive politics and religion of humanity 4:522–523
  on positivism 4:520–521, 18:625–626
  positivistic conception of science by 18:625–626
  on social dynamics 4:522
  on sociology 4:521–522, 12:389, 18:620
  theory of divine right 4:521
  works
    Appel aux conservateurs 4:520
    Catéchisme positiviste 4:520
    Cours de philosophie positive 4:519–520, 21:241, 22:991, 24:583
    Discours sur l'ensemble du Positivisme 4:520
    The Positive Polity 4:361
    Synthèse subjective 4:523
Comtean positivism 18:620–621
Comunione e Liberazione 16:807
CONACYT *see* Consejo Nacional de Ciencia y Tecnología (CONACYT)
Conant, James Bryant 13:392, 24:762
Conative functions of language 12:783
Concatenated functions 17:297
Concatenated tables 5:4
Concatenative speech synthesis 23:260
Concealed ovulation 18:899
Concentrated animal feeding operations (CAFOs) 1:568–571
Concentrated disadvantage 4:322, 4:525–530, 4:322f
  and crime 4:322–324
  and social disorganization 4:322–323
  and urban violence 24:950–951
Concentrated ownership 4:943–944
Concentrated poverty 24:730–731
  and homicide 11:163
  neighborhood effects 16:443, 16:446–447
  urban sociology 24:931
Concentrated solar power (CSP) 7:636
Concentration 21:412
  deterioration of 25:170
  of disadvantage 4:525–526
    causes of 4:527–528
    consequences of 4:528–529
    examples around the world 4:526–527
    poverty in United States 4:527
  market concentration 1:798–799
  of political power 3:708
Concept analysis 17:527
Concept empiricism 7:568–569
Concept formation
  and cultural variations 4:534–535
  language role 4:533–535
  role of 4:531–535, 25:325–326
    action in 4:532
    mental operation 4:531–533
  and thought 4:534–535
Concept identification, discrete-state models 6:524–525
Concept languages 14:319
Concept learning 4:538–541
  categorization 4:538–541
  connectionist models of 4:654–657
    ALCOVE 4:655–656, 4:656f
    backpropagation 4:655–656
    category learning 4:654–655
    delta rule model 4:654–655
    feedback 4:654–655
    future directions 4:656
    multilayer networks 4:655–656
  decision-bound models 4:539
  exemplar theories 4:538–539
  hybrid models 4:540
  memory 4:538–541
  neuropsychology of category learning and representation 4:540–541
  perception 4:538–540
  prototype theories 4:538

Concept learning (*continued*)
rule 4:538–540
selective attention in 4:539–540
time course of classification 4:540
Concept mapping 24:101–102
Concept of learnability 8:358
Concept of matter 4:532
Concept of Potentiality 20:761
Concept of psychological tools 25:324
Concept of space 4:535
Concept of time 4:535–536
Concept of trajectory 14:51
Concept possession 13:81
Conception 20:451
Conception clans 24:456
Conception totemism 2:236
Conceptions of control 5:537
Conceptions of progress, Christian historiography 11:33
Concepts, method, and values (CMV) 12:258
Concepts
associative approach 4:26–27
cultural modernity 15:700
direct historical approach 4:26
and embodied cognition 7:422–423
in Empirical Implications of Theoretical Models 18:455
historical work 15:702–703
of history and their formation 11:54
media for preservation 11:54
proliferation of history 11:54
interdisciplinary approach 23:227
legal reasoning 13:777–783
of modernity 15:700–701
of modernization 15:700–701
in social sciences 15:701–702
of reflexive modernization 15:705
structural archaeological approach 4:27
time series analysis 24:341–344
Conceptual abstractions 20:203–204
dialectical critique of fetishism 20:204
humanist critique of naturalism 20:204
nominalist critique of realism 20:204
Conceptual analysis, legal philosophy 18:28–29
Conceptual artists 2:3
Conceptual change
instructional psychology 12:202
Conceptual change theory 21:278
Conceptual conservatism, Feyerabend on 9:117–118
Conceptual continuum 14:242t
Conceptual development 21:276
Vygotsky on 25:313–314
Conceptual frameworks, of economic statistics 5:760–761
Conceptual history 4:548–554, 18:486
before and after Koselleck 4:549–550
concepts in history of ideas 4:550
critique, transformation, and resumption of encyclopedia of basic concepts 4:551–552
discourses 4:548, 4:550–553
French approach 4:550–551
historical epistemology 4:550–551
historical semantics 4:548–553
historiography 4:548
history of mentalities and philosophy of science 4:550–551
international trends 4:552–553
Koebner on 4:549–550
lexicon *Geschichtliche Grundbegriffe* 4:551
linguistic turn 4:548
Luhmann on 4:552
philosophy of language 4:548
Schmitt on 4:551
Conceptual interface 9:877
Conceptual knowledge 12:326
semantic memory 21:572–578
Conceptual learning, policy change 3:843
Conceptual metaphor theory
of cognitive semantics 13:933–934
of language 13:297–298
Conceptual models
for emotions 7:462–464
social networks 22:366–367
Conceptual preparation 23:256
Conceptual relativism 20:235–236
about science 20:236
Conceptual relativity, in internal realism 19:665–666
Conceptual representation 16:675
Conceptual scale 14:241
Conceptual space 17:555–556, 17:556t
Conceptual/etiological theory of an intervention 15:65
Conceptualism 13:772
legal reasoning 13:779–780
Conceptuality, cardiovascular reactivity 3:118
Conceptualization 11:285
of a continuum for proficiency 16:66
in critical realism 5:280
and models 15:282–283
of neoliberalism
as ideology 16:462–463
neoliberal governmentality 16:463–464
as policy 16:464
of organizational role 6:597–598
of target object
in survey questions 23:766
of technology 24:113–114
Conceptualizations of role 4:741–742
Conceptually oral text 14:241
Conceptually written text 14:241
Concerted cultivation 5:378, 22:327
Concessional loans 3:503
Chinese foreign aid 3:502–506
Conciliation 15:71–72, 17:621
appeasement 1:826–828, 1:830
Conclusion (*explanandum*) 10:773–774
Concomitant of the rth order statistic 17:294
Concomitants of order statistics 17:293–294
CONCOR (The CONvergence of iterated CORrelations) 3:910, 23:622–623
Concordance
testing, nonparametrics 16:880–881
in twin studies 24:688–692
Concordance ordering 14:450–451
Concrete and abstract instantiations, in mathematics learning 4:68–69
Concrete interaction patterns, social structure as 22:716–717
Concrete operational child 18:128–129
Concrete operational thought 7:883
Concrete self-evaluations, self-development 21:492
Concubine 17:488–489
Concurrency, in urban growth management 10:448
Concurrent degree (undergraduate) 24:52
Concurrent expert testimony 8:585–586
Concurrent processes 16:541, 16:543
Concurrent sexual partnerships 21:798
Concurrent validity 6:345
Condillac, Etienne Bonnet de 19:410
Condition humaine 11:371
Condition on Extraction Domains (CED) 16:5
Conditional adaptations 8:422
evolutionary developmental psychology 8:422–423, 8:426
Conditional additivity 4:877
Conditional approximated likelihood 14:121–122
example 14:122
Conditional Bayes risk function for classification 23:412
Conditional Bayes risk function for learning 23:414
Conditional cash transfer (CCT) 6:274, 8:562, 25:479
impact on marriage and fertility 4:555–559
adolescent girls 4:558
adult female 4:556–557
developing countries 4:555
education 4:557–559
empirical evidence 4:556–558
empowerment 4:555
health 4:555, 4:557–558
other forms of transfers 4:558
school-age children 4:557–558
social protection 4:559
theoretical foundations 4:555–556
Conditional cash transfer (CCT) programs 18:751, 22:285
Conditional fallacy 5:95
Conditional fees 12:922, 14:267
Conditional heteroscedasticity, time series models with 24:312–313
Conditional independence 10:341, 10:345–346, 14:323–324
model of 16:132
loglinear models 16:125, 16:127
Conditional inference, model selection 15:650
Conditional inference model 15:272–273
Conditional knowledge 24:90
Conditional likelihood 14:120–121
example 14:121
Conditional likelihood analyses, for crossover designs 5:353
Conditional logistic regression 2:592–593

Conditional maximum likelihood (CML) method 2:592–593
in outcome-based sampling 17:105
Conditional monopolies 14:622
Conditional positive assortative matching 2:106
Conditional probabilities 14:632–633, 19:25
Bayes' theorem 2:378–379
Dutch Book theorem for 19:39
infant and child survival 16:52, 16:54–55
in outcome-based sampling 17:104
Conditional profile likelihood 14:121–122
Conditional reasoning 14:307–311
Conditional reference prior 14:122–124
Conditionalities
of norms 17:5, 17:7
Conditionals 5:95–96
counterpossible 5:96
probabilities of 5:96–97
subjunctive 5:97
Conditioned aversion 3:767
Conditioned emotional responses, and anxiety 1:808
Conditioned inhibition 3:765, 18:41–42
Conditioned reflex techniques 17:609–610
Conditioned reflexes 17:608–609, 20:211
Watson on 25:449
Conditioned response (CR) 3:764–766, 3:768, 8:849
Conditioned stimulus (CS) 3:765, 8:849, 15:156, 17:608–609
and fear 8:870
input circuitry
eyeblink conditioning 8:637–638
pathways, cerebellar feedback to 8:638–639
Conditioning 13:691, 13:693, 18:40–42
avoidance 13:692
behavior therapy 2:426
classical 1:374
fear-potentiated startle 8:860–864
operant 1:374
Pavlovian 13:692
Conditioning factors, and general strain theory 23:496–497
Conditioning therapy 2:408
Conditions, addiction 1:98
Condoms 21:793, 21:800
use, in Latin America and Caribbean 9:70–71
Condorcet, Marquis de 8:292, 18:917
Condorcet's jury theorem 11:889
Conduct
and anomie 1:720
and national character 16:233
regulation *see* Regulation of conduct
of research 20:510–512
Conduct disorder (CD) 2:470, 6:7, 6:13, 19:451–452, 23:557–560
psychiatric care for 19:299
Conduct of business, regulation and supervision 9:191
Conduct of conduct 5:257
Conduct parameter 14:553
Conducted electronic weapons (CEWs) 18:255, 18:258
Conduction aphasia 18:48, 18:57–58, 18:61, 1:820t
Conduit loans 20:16
Confabulation 4:560–569, 6:719–720
amnesia 4:560–562, 4:564
autobiographical memory 4:560, 4:566–567
brain pathology 4:561
classification 4:560–561
behaviorally spontaneous confabulation 4:560–561
cognitive mechanisms underlying 4:566
associated features 4:566
constitutional processes 4:566
core processes 4:566
definition and features 4:560
differentiation and etiology 4:561
disturbed temporal consciousness 4:562–563
executive dysfunction 4:562
memory reconstruction processes 4:565–566
control mechanisms underlying 4:565–566, 4:565f
monitoring 4:561, 4:564–567
motivational accounts 4:561–562
neomotivational accounts 4:567–568
and source monitoring 23:31–32
strategic models 4:561, 4:564–567
clinical evidence in support of 4:567
imaging evidence for 4:566
temporal context confusion 4:563
clinical evidence for 4:563–564, 4:563f
temporality models 4:561–564
imaging evidence for 4:564
Confactor rotation 3:243–244
Confederación Nacional del Trabajo 1:681
Confederacy 4:570–571, 4:573–575
Confederate Ireland 4:573
Confederation 4:570–571, 11:683
Africa 4:575
Pan-Africanism and postcolonial experiences 4:570–571
Arab world, during cold war 4:575–576
Asia and the Far East 4:574–575
Canada 4:574
characteristics 4:570–571
definition of 4:570–571
European integration 4:571–572
Germany 4:573
history of 4:570–576
Ireland 4:573
North America 4:573–574
postcolonial Latin America 4:574
postimperial and postcolonial transitory 4:572
Serbia and Montenegro 4:572
Switzerland 4:573
tribal 4:571
pre-Columbian Mexico 4:574
United States 4:573–574
Confederation of European Specialists in Paediatrics (CESP) 20:516
Confederation of Workers, Peasants, and Students of the Isthmus (COCEI) 11:756, 11:758
Conference of Regulatory Officers (CORO) 13:621
Conference on Empirical Legal Studies (CELS) 7:554–556
Conferences, professional associations 22:676–677
Confession building 20:99–100
Confession sexuelle d'un anonyme russe (The Sexual Confession of an Anonymous Russian) 21:740
Confessional age 20:98, 20:100, 20:103
Confessionalization 20:98–104
Calvin on 20:98–103
Catholic Reformation 20:98, 20:100–101
confession building 20:99–100
confessional age 20:98, 20:100, 20:103
cultural history 20:99–100
era, religious, social, and political structures and developments in 20:98
basic structures 20:98–100
Europe 20:101–102
Germany 20:101
late Middle Ages 20:100–103
Protestant and Catholic theology 20:100–101
historiography and periodization 20:98–100
Luther on 20:98–103
social history 20:99–100
state formation 20:100, 20:103
Zwingli on 20:100–101
Confidence 20:51–52
and risk 20:688
Confidence gap
and political trust 24:658
Confidence in government 24:658, 24:660
Confidence intervals (CI) 8:19, 9:124, 9:836, 11:496–497, 13:388, 15:275, 20:877, 23:423
Neyman on 16:824, 16:827
nonparametric statistics 16:889
for quantiles 17:292
Confidence ratings 12:81
Confidence–accuracy relationship 8:643–644
Confidential informants, in United States 16:250
Confidential Information Protection and Statistical Efficiency Act 18:943
Confidentiality 3:290, 8:80, 17:534, 20:514
biomedical research ethics 8:80–81
census 3:302, 3:304
in clinical genetics 2:622
in combined pharmacotherapy and psychotherapy 19:497
Common Rule 4:254, 4:256
fieldwork ethics 9:154
in health surveys 10:729
psychiatric ethics 8:65–66
of respondents, archiving 1:917–918
in social research 18:942–946

Confidentiality (*continued*)
data sharing and researchers' use of statistical databases 18:944–945
definitions 18:942
ethical and legal context 18:942–943
future 18:946
informed consent 18:946
professional codes 18:942
record linkage, and integrated data systems 18:945–946
research subjects 18:943
statistical disclosure limitation 18:944–945
surveys, randomized field trials, and other data collection efforts 18:943–944
use of state databases in United States 18:945
statistical disclosure limitation 4:577–581
in surveys and polls, protection 23:755–756
Configural face processing 8:652–654
Configurational epistemic culture 7:876
Configurational knowledge *see* Cognitive maps
Configuration-based processing (prosopagnosia) 19:252
Configurationism 23:611
Configuration-level issues
supply chain management 23:709–710
information support decisions 23:710
plant location decisions 23:710
product portfolio decisions 23:710
supply base decisions 23:709
Confirmation
atomism/holism 2:131–133
records with 16:904
Confirmation biases 16:419–420, 19:21, 22:496
Confirmation of evolution 8:469
Confirmatory combinatorial data analysis (CCDA) 19:431–432
distributional tests 19:432
exact *p*-values 19:432
linear assignment 19:431
multinominal models 19:431–432
quadratic assignment 19:431
resampling 19:432
Confirmatory data analysis 8:600
Confirmatory factor analysis (CFA) 3:788–789, 6:438, 8:686–690, 8:705, 8:707, 17:860
data-model fit assessment 8:688–689
identification 8:686–688
model modification 8:688–689
parameter estimation 8:686–688
of social dominance orientation 22:251
specification 8:686–688
Confirmatory mechanisms, conflicts 4:583–584
Confirmatory multidimensional scaling 16:45–46
Confirmatory stopping rule, of decision making 5:920
Conflict data sets 8:98t
Conflict embracing community 21:138
Conflict expansion, and political agendas 1:362–363
Conflict explanations, of ethnocentrism and prejudice 8:170–171
Conflict face negotiation, honor killing 11:188–189
Conflict face-negotiation theory (FNT) 11:185–187
Conflict group 4:623
Conflict Information and Analysis System (CONIAS) 5:836
Conflict interaction 4:582
Conflict management 17:631–634
disputes 6:560–561
social control 4:825
Conflict mediation 4:627–631
activities 4:628–629
definition 4:627–628
negotiations 4:627–628, 4:630
peace agreements 4:630
practice 4:628–630
research 4:628
strategies and ripeness 4:630–631, 4:630t
third parties 4:627–628, 4:630
timing 4:627, 4:630
Conflict of interest 1:344, 20:515
bargaining 16:415
human subjects protection in research 8:50–51
research 20:527
Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME) 20:515
Conflict of laws
definitions 13:614
interpersonal or intertemporal 13:615
legal diversity 13:615
parts of 13:615
sources of 13:615
in sub-Saharan Africa 23:655–656
Conflict resolution 16:435, 17:625, 17:628–629, 17:631–632
and agreeableness 1:467
in collectivism/individualism 4:209
organizational decision making 17:355–356
social psychology of 4:582–588
characteristics and goals 4:587
interactive conflict resolution 4:586–587
management styles, negotiation and mediation 4:586
problem-solving workshops 4:586
rationale in intergroup contact hypothesis 4:587–588
transfer and effectiveness 4:588
see also Ethnic conflicts
Conflict statistics 5:836–843
challenges 5:841–842
contemporary trends 5:840–841
disaggregation 5:840–841
extension 5:840
individualization 5:840
data 5:839–840
origins of 5:836–837
research, role in 5:841
Conflict Tactics Scale (CTS) 5:724, 8:914–915
Conflict theories 4:589–590
Cold War 4:122–124
Collins on 4:211, 4:214–216
of educational sociology 7:290, 7:293
and functional theory of stratification 9:538
of interviewing 5:854–855
and social structure 22:723
of socialization and education 22:864
view of schools 21:56–57
Conflict theorizing 22:956–957
analytical conflict theories 22:956
comparative-historical conflict theories 22:957
Marxian conflict theories 22:956–957
microlevel conflict theories 22:957
Conflict thesis 20:284
Conflict transformation 17:648–649
Conflict trap 3:715
Conflict-cohesion hypothesis 25:403
Conflict-related sexual violence 10:194
Conflict-related violence 10:194
Conflicts 4:589–593, 9:578, 10:2–3, 10:12, 13:510–511, 15:776–777, 16:433–435, 22:245, 25:371–373
and adolescence 4:634
and agonism 1:446–448, 4:591
anthropology of 4:620–626
attraction and displacement of emotions 4:622–623
changing identifications and alliances 4:625
as embedded social action 4:621
forms of actor conflict 4:621–622
goals 4:620
history 4:620–621
institutions of conflict regulation 4:624
methodological problems 4:621
motives 4:623
normative conflict 4:622
regulation and cohesion 4:623–624
retaliation, compensation, and punishment 4:624–625
archaeology of 4:600–613
archaic states and aristocratic armies 4:603–604
chiefdoms and chiefly militias 4:603
egalitarian hunter-gatherers and self-redress homicides 4:600–601
foragers 4:600–602
fortifications 4:612
iconographic representations in art and rituals 4:612–613
nation-states and industrial armed forces 4:604
skeletal trauma 4:612
tribes, feuds, and kin-militias 4:601–602
weapon combinations and technologies 4:611–612
weaponry 4:609–610
weaponry, in ritualized contexts 4:610–611
centripetalism 4:591–592
child soldiers 3:430–438

and childhood 4:634
and communities 4:636
compromise 4:591
consociationalism 4:591–592
Coser on 5:38
and crisis 4:592
definition 4:582, 4:594–595, 4:632–633
differentiations of 4:621f
and early friendships 9:450–452
and education, connections between 4:633–634
and emotions 7:502, 7:504–506
ethnic *see* Ethnic conflict
and factionalism 8:683–685
functionalism 4:590–591
groups against groups 22:246
injustice and 12:939
interdependence 4:590
interpersonal *see* Interpersonal conflict
interpersonal, face in 5:429
and intersectionality 12:661
intrapersonal *see* Intrapersonal conflict
and justice 12:950, 12:953
justice and 12:940
mediation 12:940–941
military. see Military conflict
Myrdal on 16:159, 16:162–163
nature of 4:589–590
Occupy movement 4:592, 4:592f
organizational 4:639–643
pastoral 17:582
and peace/nonviolence 17:625–626, 17:628–629
pluralism 4:590–591
post-1945 10:12–13
power-sharing 4:591
psychology of 16:423–429
cognition of disputant 16:428
emotion of disputant 16:428
social context 16:427–428
typology 16:427
settlement of
by just contracts 12:940–941
justice by 12:940
and shock, cultural psychiatry 5:476
social 12:243
and social problems 12:543–544
social psychology of 4:582–588
attitudes 4:582–584, 4:587–588
causal attribution 4:583
confirmatory mechanisms 4:583–584
escalation and intractability 4:585
ethnocentrism, social identity, and social dominance 4:585
images 4:583–585
prospect theory 4:584
self-fulfilling prophecy 4:584
social judgment and decision-making 4:584–585
social perception and cognition 4:583
stereotypes 4:583–584, 4:587
and socioemotional development 4:594–599
adulthood 4:596, 4:598
agreeableness 4:598
childhood 4:595–598
cognition 4:594–597
friendship 4:596–597
personality 4:598
social interaction 4:597–598
sociology of 5:662–664
and statehood 23:347–348, 23:350–351
and structural adjustment 23:598–599, 23:601–602
taxonomy of 22:805
territorial *see* Territorial conflict
and territory 24:223–225
and unintended consequences 17:420
and urban policies 24:903
and water scarcity 25:443–444
in world culture 10:228–230
and youth 4:634
see also Negotiations
Conflicts justice 13:617
Conflictual replicated game 4:871
Conformal projections 14:285
Conformist behavior 12:645
Conformist bias 24:130
Conformity 2:264, 2:266–267, 13:75–76
in collectivism/individualism 4:209
cultural transmission 5:403–404
models of 4:1–2
norms of culture 4:5
organizational groups 10:408–409
positive beliefs 4:2–3
and prejudice 18:831
psychology 4:1–4
rationality 4:1–3
reasons 4:1–5
social pressure 4:1–2
social rules in moral, political, and religious fields 4:3–5
and values 25:28–29
Conformity bias 4:154–155
Conformity studies, of obedience 17:61–63
Confounder effect 15:67
Confounding 17:108–109, 7:828b
evidence-based medicine 8:329
Confounding variables 5:325–326, 15:64
in causal analysis 3:247–248
control, by experimental designs 8:536
Confrontations
in Cold War 4:128–129
mobilizing, in collectivism/individualism 4:208
Confucian approach to learning 12:696–697, 12:699–700
Confucian class of state servants 2:919–920
Confucian dynamism 12:358
Confucian literati 25:95
Confucian meditation 16:168
Confucian monarchy 25:95
Confucianism 2:886, 2:889, 3:683, 16:806–807, 20:359, 20:367, 25:669
and harmony 10:529
in Japan 12:796
in Korea 2:72–73, 3:532
and morality 15:806, 15:808–809
Confucians 4:312
Confucius 4:312, 13:800, 14:253, 20:367
Confusion, and posttraumatic amnesia 18:636–639
Confusion Assessment Method (CAM) 14:5
Congeneric measurements model (CMM) 19:428
Congenital malformations 11:390
Congenital prosopagnosia 19:250
Congestion 7:726, 15:636–637
in transportation 24:629
airport delays 24:629, 24:631, 24:633–634
alleviation of 24:632–633
capacity investment 24:632
costs 24:630
future 24:633–634
information technology 24:633
intelligent transportation systems 24:633
investment 24:632
latent demand 24:632–633
modeling 24:630–632
policy 24:633
pricing 24:632–633
reasons 24:629–630
tolls 24:632–633
travel demand 24:629, 24:631–632
Congestive heart failure (CHF)
coping with 3:111–114
Conglomerate diversification 6:585
Conglomeration 16:266
Congo 3:312–314
Congregational form of governance 20:409
Congregations in religious organizations 20:410–411
Congrès International d'Architecture Moderne discourse on Urbanism 18:193
Congress for the New Urbanism (CNU) 16:810
Congress of Industrial Organization (CIO) 12:374
Congress of Racial Inequality (CORE) 3:693
Congress of Vienna 2:346, 18:398–399
Congressional Budget Office (CBO) 11:732, 11:736, 15:563–564
cash value 11:736
comprehensive income measure 11:735
Current Population Survey (CPS) 11:735
Congressional Office of Technology Assessment (OTA) 24:125
Congruence
motives
see also Explicit motives; Implicit motives
research 15:954–955
in positive development 15:954–955
Rogers on 20:737
and values 17:824, 25:28
in vocational interests 25:239
CONIAS *see* Conflict Information and Analysis System (CONIAS)
CONICET *see* Consejo Nacional de Investigaciones Científicas y Tecnicas (CONICET)
Conjoint analysis (CA) applications 14:865–867

Conjoint analysis (CA) applications (*continued*)
attributes 14:865–866
preference models 14:865–866
random utility theory 14:866
Conjoint choice designs 17:260–261
Conjoint measurement 14:862–864, 14:870
additive representations 14:863
specific theories 14:863
uniqueness 14:863
decision under risk and uncertainty 14:863–864
numerical values 14:862
objects and attributes 14:862
preferences and axioms 14:862–863
independence and cancellation conditions 14:862–863
ordering 14:862
solvability and archimean axioms 14:863
and social choice 14:863
Conjoint theory of social measurement 21:7
Conjugal relationships 8:725
Conjugation class 15:823–824
Conjunction fallacy 12:719–720
Conjunctional possessive 18:632
Conjunctions 13:276–278, 19:65
Conjunction-specific neurons 16:634
Conjunctive decision rule, of decision making 5:919
Conjunctural causation 3:191
Conjunctural coincidences 11:869
CONLEARN (face processing mechanism) 8:651
Connectedness
social work response for disasters 7:427, 7:431
and spirituality 23:285–286, 23:289
Connecticut 1:369
Connecting antisemitism 1:786–788
Connecting/networking, cultural mediators's role 5:441
Connection method
reasoning 5:935–936
systems based on 5:935
Connection thesis
law and morality 13:479
in legal philosophy 18:27
Connectionism 4:84–89, 18:57–61
artificial intelligence 2:37–38
knowledge representation 2:37–38
learning 2:37
and cognitive psychology 4:107–108
Thorndike on 24:306
Connectionist models 21:582–583
of concept learning 4:654–657
ALCOVE 4:655–656, 4:656f
backpropagation 4:655–656
category learning 4:654–655
delta rule model 4:654–655
feedback 4:654–655
future directions 4:656
multilayer networks 4:655–656
Connectionist networks, implicit rules in 13:81–82
Connections model of network formation 16:575
Connective action 22:586
Connectives 18:631
Connectivity 18:817–818
dynamic causal modeling 9:529
functional magnetic resonance imaging (fMRI) 9:529
Granger causality 9:529
mathematical model 9:529
neuroanatomical model 9:529
neuroimaging studies of 5:27
structural equation modeling 9:529
Connectivity studies, in ADHD 2:158–159
Connectome
and behavioral circuitry 9:884–885
and prefrontal cortex 18:814
Connerton, Paul 1:748
Connolly, William, works, *Terms of Political Discourse* 7:591
Connotation
and cultural sociology 5:499–500
in semiotics 21:596
Connubium 1:968
Conquest, and indigeneity 11:760
Conrad, Johannes, works, *Handwörterbuch der Staatswissenschaften* 7:585
Conring, Hermann 23:423
Consanguineal kinship 8:725, 8:813
terms 13:61
Consanguineous marriage 11:725, 14:600
comparison of incestuous matings with 11:728
in contemporary societies, attitude towards 11:726
detrimental health effects 11:727
history of 11:725–726
incestuous marriages 11:726–727
in Western societies, attitude towards 11:725–726
Conscience
freedom of
liberalism and 14:23–24
and moral development 15:788–790, 15:794
Conscience collective 4:175, 6:701–702, 23:6
Conscientiousness 4:658–665, 9:236, 16:493–494, 17:777, 17:787–788, 17:802–807, 17:914–915
with academic performance 17:788
and adolescent personality 17:864–866
at behavioral level 17:788
in Big Five model 16:743
development over life span 4:662–663
with effortful control 17:788
and grit 10:397, 10:399
and important life outcomes 4:661–662
with intrinsic motivation 17:788
lower-order structure of 4:658–659
measurement 4:659–660
and personality traits 17:920–923
and procrastination 19:75
self-control 4:658–660, 4:662
Conscientiousness Adjective Checklist 4:660
Conscientizaçao 4:381
Conscientization 16:916–917
Conscious control (CC) 4:666–672
and creative imagination 4:671
Conscious knowledge 13:85–87
in context of psychological theory 13:86–87
exhaustiveness and exclusiveness 13:86
judgment knowledge 13:87–88
structural knowledge 13:87
two criteria for test of 13:86
validating test of 13:86
Conscious procreation
ecological economics 6:860
Consciousness 10:738, 13:551, 13:554, 13:726, 16:326–327, 16:330–331, 16:636, 21:449–450, 21:452–453, 23:291–296
anthropology of 7:147
of belonging 8:175–177
and change blindness 3:344, 3:348
class consciousness 3:739–744
computer models of 12:704
discursive 7:416
and dreaming 6:652–653, 6:655–656
of a generation 9:866–867
generational 9:866–867
kernels of 9:867
in modern youth movements 9:868
higher-order representational theories of 4:677
historical *see* Historical consciousness
historical consciousness 1:748
and hypnosis 11:467–470
and implicit social cognition 11:714, 11:716
individual 17:932, 17:934–935
and intentionality 12:334
legal 13:538, 13:728–730
nature of 13:393
neural basis of 4:673–678
phenomenology 18:11–15, 18:18–21
practical 7:416
psychological development by changing interfunctional structure of 25:331
of Self 12:787–788
Skinner on 22:27
structuralism and 14:199
and tactile processing 24:15–16
visual, theory of 4:676–677
and visual mental imagery 25:167
Whorf on 25:576, 25:578
Consciousness-raising, for sexual harassment 21:721
Conscription 15:512–516, 15:519–522
demise of 16:307
in France 25:364
warfare 25:417–418
Consecration/editing 5:441
Consecutive degree (undergraduate) 24:52
Consecutive tests for outliers 23:389
Conseil d'Etat (French administrative court) 3:669
Conseil Francais du Culte Musulman 12:239

The Consejo Latinoamericano de Ciencias Sociales (CLACSO) 22:622
Consejo Nacional de Ciencia y Tecnología (CONACYT) 22:621, 22:624–625
Consejo Nacional de Investigaciones Científicas y Tecnicas (CONICET) 22:621, 22:627
Conselho Nacional de Desenvolvimento Cientifico e Tecnologico (CNPQ) 22:621, 22:626
Consensual agreement 15:70
Consensual crime 25:75–76, 25:79
Consensual solidarity 12:403–405
Consensualization 23:449
Consensus 13:510, 21:533, 21:877–879
  agonism 4:591
  centripetalism 4:591–592
  compromise 4:591
  concept of 10:742–743
  consociationalism 4:591–592
  and crisis 4:592
  Dahrendorf on 5:662–663, 5:665
  discrimination in markets 6:536–537
  functionalism 4:590–591
  interdependence 4:590
  Occupy movement 4:592, 4:592f
  peacebuilding with 4:592
  pluralism 4:590–591
  power-sharing 4:591
  Shils on 21:886–887
  and social order 9:539
  theory 4:589–593
Consensus ad idem (meeting of the minds) 4:810
Consensus committees 6:78
Consensus conferences 24:125
Consensus democracy 4:693–694
Consensus information 10:826
Consent 4:803
  biobank ethics and 2:608
  in common law system 4:809
  for data sharing, archiving 1:918
  informed *see* Informed consent
  marriage 14:601, 14:603
Consent-to-Search program (St. Louis) 10:460
Consequential damages 3:29
Consequentialism 4:679–683, 18:194, 19:581
  collective face 4:682–683
  defined 4:679–680
  global, and local 4:680–681
  in-house issues 4:683
  negative case 4:682
  positive case 4:681–682
  and practical reasoning 18:780–781
  role of 4:681
Consequentialist 8:74
Consequentiality, of flashbulb memory 9:256–257
Conservation 7:713, 18:127–129, 23:795, 23:798, 25:439–440
  alternative visions of 7:714
  cultural heritage 5:414–417
  of cultural heritage 5:488–490
  of energy and socioeconomic thought 10:756
  environmental 7:713, 7:716
  and environmental cognition 7:706, 7:708–710
  and environmental movements 7:764–766
  hunting/gathering societies 11:430
  in Soviet Union 20:831–833
Conservation citizen science projects 3:632
Conservation International 7:105
Conservation of force, principle of 10:756
Conservation of resources (COR) 25:734
  stress model 7:805
  theory 18:707, 19:369, 25:734
  of well-being and burnout 25:534–535
Conservatism 1:602–608, 2:257, 2:263, 4:308–309, 4:684–691, 9:410, 18:400, 20:397, 20:563
  in America 4:690
  as an intellectual movement 1:602–604
  and anticommunism 1:604
  approaches to problem 4:684–685
  and authoritarianism 2:263–264
  authority 4:689
  Buckley on 1:604
  Burke on 4:689–690
  centralized statism 4:689
  characteristics of conservative thought 4:685–686
  church 4:689–690
  deference 4:689
  definition 4:689
  definition and characterization 4:685
  European 4:689–690
  future of 4:691
  Goldwater on 1:605
  hierarchy 4:689
  historical overview 4:686–687
  individual rights 4:691
  and libertarianism 1:603–604
  Mannheim on 14:497
  and mass society 14:744–745
  Nisbet on 16:833–834
  nobility 4:689
  organizational 19:558
  political 19:557–558
  precursors 4:686
  Reagan on 1:605
  religious right 4:690
  Republican Revolution of 1994 1:605–606
  as a social movement 1:604–606
  in Soviet Union 20:820–821
  Tea Party 1:606
  tensions with 4:691
  tradition 4:689–690
  and traditionalism 1:603
Conservative 4:685
Conservative Christianity 14:10
Conservative construal of reflective equilibrium 20:78–79
Conservative enlightenment 7:660
Conservative movements 12:871
Conservative Protestants (CPs) 20:292–294
  disciplining and nurturing children 20:320
  equality within families 20:319
Conservative welfare state 24:573
Conservative–Liberal Democrat coalition 20:132
Conservatives (United States) 20:397
  The Fundamentals booklet 20:397
Considered judgments
  and reflective equilibrium 20:77–80
Consiglio Nazionale delle Ricerche (CNR) 22:669–670
Consilience of inductions, Darwinism 5:681
Consistency 20:5–7
  algorithms 8:727
  and embodiment 7:413
  of flashbulb memory 9:255–256
  of linguistic databases 5:814–815
  and personal identity 17:744–746, 17:748
  and persuasion 17:942
  social influence 22:349–350
Consistency conditions, choices 19:440–441
Consistency problem 25:215
Consistent correction, data adjustments 6:130, 6:134–135
Consociational democracy 6:94, 19:209
Consociationalism 4:692–694
  cleavages, political 4:692–694
  and conflict/consensus 4:591–592
  consensus democracy 4:693–694
  critiques of 4:693–694
  elites, political 4:692–694
  ethnic conflict 4:692
  explanation 4:692–693
  grand coalition 4:693–694
  Lijphart on 4:693, 4:692f
  minority veto 4:693
  political stability 4:692–693
  prescription 4:693–694
  proportional electoral system 4:693
Consolidated metropolitan statistical area (CMSA) 24:857
Consolidated Standards of Reporting Trials (CONSORT) guidelines 8:333
Consolidation 13:680–681
  of episodic memory 7:858–859
  fear conditioning 1:812
  interference 20:616–617
  in memory 24:204
  of memory at cellular level 23:874–876
  of procedural knowledge and sleep benefit 16:846–847
  and radio 19:880–881
  urban anthropology 24:776–778
Consonance 10:621
Consonant acculturation 2:85
Consonants 18:66–68, 23:715
Consonant–vowel (CV) syllables 6:358–359, 23:235
CONSORT *see* Consolidated Standards of Reporting Trials (CONSORT) guidelines
CONSORT Statement 8:544
Consortium of European Social Science Data Archives (CESSDA) 5:825, 5:832, 22:645–646

Consortium of Higher Education Researchers (CHER) 10:866–867
Consortium of Social Science Associations (COSSA) 22:675
CONSPEC (face processing mechanism) 8:651
Conspecificity 15:948
Conspicuous consumption 4:695–699, 4:757–758, 5:535, 8:28, 13:849, 16:555, 22:331–333, 25:192
  Bourdieu on 4:695, 4:697–698
  and class 4:695–697
  cultural capital 4:695, 4:697–698
  future 4:698
  habitus 4:695, 4:697–698
  inequality 4:695, 4:697–698
  status consumption 4:698
  Veblen on 4:695–698
Conspicuous leisure 22:332–333
  consumption 4:758
Conspicuous outrage 8:836
Conspiracy 7:392
  Boltanski on 2:752–753
Conspiracy doctrine 11:349
Conspiracy of Equals 22:836
Conspiracy theories of institutions 12:193
Constancy hypothesis 13:123
Constant, Benjamin 3:647
  on freedom 9:389–390
Constant capital 3:90–91
Constant causes, theory of 13:383–384
Constant risk function
  LEXIS diagram 6:157–159
    squares 6:158–159
    vertical-sided parallelograms 6:158
Constant transition intensities 8:314
Constantine (emperor) 3:233, 19:408
Constantinople 3:233–234, 3:548, 9:389
Constatives 18:797
Constituencies 23:903–904
  electoral geography 7:347
  proportional representation 19:205–207
  structure 7:349–350
Constituency (chunking) of phrasal units 23:886–889
Constituent Assembly 20:857
Constituent ordering 25:644–645, 25:652–655
  arbitrary patterning 25:655
  domain integrity 25:652–653
    Dutch 25:652–653
    Ngiti 25:652
  head proximity 25:653
  scope 25:653–655
Constituents, political representation 18:441–444
Constitution making 4:704, 4:706–707
Constitution of Athens 4:706
Constitution of Imperial Japan 12:795
Constitutional absolutism (royalism) 1:10
Constitutional Court of the Russian Federation 13:429
Constitutional courts (CCs) 4:700–703, 12:898–900, 23:723
  American system 4:700
  Austrian or concentrated model 4:700–701
  composition of judiciary 4:701–702
  decision making 4:703
  diffuse model 4:700
  as function of ordinary courts 4:700
  in historical and comparative perspective 4:700–701
  independence of judges 4:702
  international constitutional jurisprudence 4:701
  jurisdiction 4:702
  mixed systems 4:701
  mobilization 4:702–703
  political role of 4:702
  political system and 4:702
  sociology of 4:702–703
  use of term 4:700
Constitutional Democratic Party 20:856–857
Constitutional democrats (cadets) 20:858
Constitutional economics 11:568–569
Constitutional identity 4:705
Constitutional interpretation 4:704–705, 4:708
Constitutional order 18:381
Constitutional patriotism 10:740–741
Constitutional Political Economy 19:535
Constitutional reconstruction and amendment 4:704
Constitutional rules 12:192
Constitutional state, democratic 10:471–472
Constitutional theocracy 4:704–705
Constitutional theory 2:527
Constitutional Treaty 16:225
Constitutionalism 4:706–707, 13:916–917
  and adjudication 5:104–105
  comparative 4:704–709
  and incest 11:730–731
  Latin America 13:411–412
  legal form of 4:707
  republicanism 20:477–479
  Schmitt on 21:28–31
  theory and practice of 4:708
  traditional notions of 4:708
Constitutionalist tradition 11:730
Constitutionalization, of global IP Regime 12:269
Constitutions 4:710–715, 4:805–806, 4:808, 13:555, 19:535
  bill of rights 4:710–714
  common components 4:711
  comparative law 4:712
  conceptual approaches to written constitution 4:710
  constitutional law 4:710, 4:712, 4:714
  constitution-making 4:713–714
  courts 4:712–714
  design efficacy 4:713–714
  foreign law 4:713–714
  functions 4:711–712
    coordination 4:712
    international-signaling 4:713
    power-solidifying 4:712–713
    pre-commitment 4:712–714
    symbolic 4:713
  government 4:710–714
  liberalism and 14:23
  meanings 4:710
  norms 4:710–711, 4:714
  physicalism 18:106
  rights 4:710–714
  states 4:710, 4:713–714
  use of term 4:710
Constitutive approaches, of law 13:543
Constitutive conception of language 4:790–791
Constitutive dimensions, of law 13:604–605
Constitutive law
  as dependent variable 17:402
  as independent variable 17:401–402
Constitutive legal theory 13:810–811
Constitutive myths 16:293
Constitutive perspective 13:599–600
Constitutive rules 21:341, 21:343
  meaning 14:846–847
Constitutive system 14:202
Constitutive theories
  of legal consciousness 13:730–732
  of legal culture 13:730–732
Constrained choices 20:593
Constrained generalized linear models (CGLIMs) approach 17:729
Constrained optimization, microeconomics 15:334–336
Constrained theory of syntax 23:893–894
Constrained-induced therapy, for language disorders 13:332–333
Constraint satisfaction 4:721
  neurophysiology of 4:722
Constraint-based theories 18:79
Constraints 23:523–525
  human-demographic traits 8:414–416
  ideological *see* Ideological constraint
  institutions 12:190–192
  and speciation 1:87–88
  stochastic optimization 23:473, 23:475, 23:478–479
Constraints on agreement 10:321–322
Constraint-satisfaction algorithm 4:719
Constraint-satisfaction neural networks 4:720–722
Constraint-satisfaction problem (CSP) 2:33–34, 4:716–723
  assignment of values to variables 4:716
  attractors 4:720–722, 4:720f
  depth-first branch and bound search 4:719
  future trends in research 4:722
  hard and soft constraints 4:719, 4:722
  neural constraint-satisfaction 4:719–722
    algorithmic variations 4:720–721
    modeling with 4:721–722
  symbolic algorithms for 4:716–719
    backtracking search 4:718
    flexible CSP variants 4:719
    local consistency checking 4:716–718, 4:716f
    local search methods for classical CSPs 4:718–719

Constraints–facilitators framework, of adolescent leisure 13:832–833
Construal level, and embodiment 7:413
Construal level theory 4:741–742
Construct bias 5:318–319, 5:323, 7:929
Construct clarification 7:930–931
Construct inequivalence 5:319, 5:323
Construct relevance, content validity 4:774–775
Construct validation, in generalization 9:841
Construct validity 7:24, 8:618, 9:834, 12:450, 12:654–655, 17:542, 19:423, 19:739, 20:240, 24:640, 24:642
  in five-factor model of personality 9:246
  of social dominance orientation 22:250–251
Construct-driven approaches 17:851
Constructed-response assessments 3:819
Constructing beliefs about Knowledge 7:241
The Construction of Scientific Facts 8:577–578, 21:311
Construction–integration model
  of discourse 12:41
  of inferences 19:396
Constructionism 4:724–728, 13:116, 20:230
  approach to gender 9:788–793
  and authenticity 2:252–253
  delinquency, extensions 6:72
  Latour on 13:414–417
  methodology 4:724
    discourse analysis 4:726–727
    focus groups 4:727–728
    narrative analysis 4:725–726
    repertory grid technique 4:724–725
  phenomenology 18:11–12
  and relativism 20:230
  social 22:222
  theory of inferences 12:40–41
  see also Constructivism
Constructions 20:16–17
  of femininity 20:798
  of gender 20:797–798
  of masculinity 20:798
  of rural communities/places/societies 20:796–797
  of rurality 20:795–796
  social relations 20:796–797
Constructive controversy theory 4:856
Constructive data generation, in stochastic optimization 23:474
Constructive empiricism 7:569–570
Constructive liberty 11:126
Constructive mediation 4:627–628
Constructive operators theory 16:503–504
Constructive patriotism 16:312
Constructive proof 16:925
Constructive technology assessment (CTA) 22:874, 24:126
  alignment 22:874–875
  strategic niche management 22:874–875
  technology forcing strategy 22:874–875
Constructivism 2:883–884, 4:56, 4:724–729, 7:241, 7:882, 12:197, 13:704, 16:925, 19:491, 20:687, 22:817, 22:915, 23:63–64
  ANT 1:81–82
  approach, to political psychology 18:430–431
  approach for emotions 7:475
  and cognitive development 18:120–122
  Computer Assisted Instruction (CAI) 4:495–497
  in education 12:67–68
  ethnomethodology 21:195–196
  history of science 21:225–228
  methodology 4:724
    discourse analysis 4:726–727
    focus groups 4:727–728
    narrative analysis 4:725–726
    repertory grid technique 4:724–725
  Nisbet on 16:834–835
  and opinion formation 17:244–245
  phenomenology 18:11–12, 18:15
  philosophy 4:729–730
    Philosophy of As If 4:729–730
  and psychology 19:414–415, 19:417
  psychotherapies 4:730–732
  qualitative research 19:685–688
  versus realism 7:798–799
  and reflective equilibrium 20:79
  social 22:228–234
  sociological work 5:504
  sociology of science 21:263
  theories of inferences 19:395
  see also Constructionism
Constructivist antirealism 20:21
Constructivist institutionalism 12:188–189
Constructivist Learning Environment Survey (CLES) 3:826
Constructivist learning theory 23:214–215
Constructivist movement 21:231–232
Constructivist rationalism 25:275
Constructivist studies of science and technology 24:135
Constructivist-Orientated Learning Environment Survey (COLES) 3:828–830
Construct-related validity of SJT scores 22:15, 22:17–18
Constructs, relationship measurement using meta-analysis 23:910
Consultation 21:114–115
Consultation competency 24:520
Consultation psychiatry 19:484–485
Consultation–liaison psychiatry 14:1–7
  complexity intervention units 14:6
  future 14:4–6
    education 14:4–5
    guidelines and quality assurance 14:5
    integration of medical and psychiatric services 14:6
    research 14:5–6
  practice 14:3–4
    evaluation of interventions 14:4
    services in general hospital 14:3
    services in primary care and nursing homes 14:3–4
  psychiatric comorbidity in physically ill patients 14:1–3
    effects 14:2–3
    in general hospital 14:1–2
    in primary care sector 14:2
    psychiatric morbidity in nursing homes 14:2
  for psychosomatic disorders 19:484–485
  somatopsychic comorbidity 14:4–5
  training of staff 14:3–5
Consultative Expert Working Group on Research and Development (CEWG) 20:524
Consultative participation, in participatory action research 17:535–536
Consulting Council (Consejo Consultivo, Argentina) 4:403
Consumer action 8:25–26
Consumer activism 8:26–28
Consumer behaviors, wanting/needing 25:355
Consumer blindness 11:865
Consumer capitalism 1:216
Consumer culture (CCT) 3:8
  and celebrity 3:274–275
  consumption 4:758
  sport 23:313, 23:316
  youth culture 25:813–814
Consumer debt, and household wealth 25:453–454
Consumer expenditure (CE) survey 5:744, 5:746, 10:303, 16:91
  multiple imputation for 16:91–92
Consumer finances survey, multiple imputation for 16:92
Consumer Financial Protection Bureau 7:21
Consumer fraud 19:161–162
Consumer misleading and deception 1:201
Consumer movement 8:26–27
Consumer neuroscience 16:698–699
Consumer organizations, consumer policy 4:733, 4:735–736
Consumer pattern 4:751
Consumer policy 4:733–737
  adversarial legalism 4:735
  behavioral economics 4:734
  California effect 4:734
  challenges 4:736–737
  class action 4:734–735
  consumer organizations 4:733, 4:735–736
  corporate social responsibility 4:736–737
  food safety 4:734, 4:736–737
  free rider dilemma 4:733, 4:736
  governance in advanced economics 4:734–735
  ideas and ideologies 4:733–734
  information approach 4:734
  information model 4:734
  'loser pays' rule 4:735
  negotiation model 4:734
  new patterns of political consumerism 4:735–736
  precautionary principle 4:734, 4:736

Consumer policy (*continued*)
product safety 4:733–737
protection model 4:734–735
public interest litigation and 4:735
self-regulation 4:733
Consumer price indices (CPIs) 5:751, 5:753, 5:758, 5:760–762, 10:303
Consumer protection 3:31–32, 14:266
Consumer Protection Act 4:810
Consumer psychologists 5:887–888
Consumer psychology 4:738–743
acquisition 4:738
advertising 4:738, 4:740, 4:742
affect 4:738–740
behavior
complexity of 4:738
evolution of scholarly research on 4:739–740
introduction of psychology to 4:740
reasons for studying 4:738–739
behavioral decision theory 4:739
buyer 4:738–739
consumption 4:738–740
current content and foci on research 4:740–741
decision making 4:738–741
disposal 4:738
Economic Man 4:739
economics 4:738–739
emerging foci on research 4:741–742
emotions 4:740–741
framing 4:739
goods 4:738–739
marketing 4:738–739, 4:741
motivation 4:739
neuromarketing 16:698–699
purchase 4:738–740
rational choice theory 4:739
sensory perception 4:741
services 4:738–739
Consumer roles, evolution of 4:140
Consumer satisfaction 8:345
Consumer society 2:361, 18:671
advent of 4:758–759
criticism during postwar period 4:759
Consumer sovereignty 3:522
Consumer surplus 5:50–51
Consumer systems 4:675
Consumer values, in international advertising 12:455
Consumer welfare 20:166
Consumer-directed care, for older people 21:498–499
Consumer-directed services, for disablism 6:450
Consumerism 4:345–346, 4:752, 16:774, 23:794, 24:793
and applied criminology 1:835
capitalist 3:99, 3:101–104
in China 2:69
criticism during postwar period 4:759
culture–ideology of 3:101–102
in England 4:345–346
and working hours 13:851
Consumers
advertising 1:208–212
anthropology of 3:8–9
of area and international studies 1:944–945
awareness of food quality 10:151–152
choice as 3:522–523
consumer–citizen in Ecuador 10:152–153
demand 14:622
ethical 8:25–30
Internet 12:594–595
mental health 15:200
surplus notion 14:622
sustainability lifestyles 23:828–829
voice of, and Internet 12:591
Consumer-taxpayer (United States) 6:3
Consumption 4:751–761, 7:16–17, 11:733–734, 11:736, 11:736f
advent of consumer society and consumption 4:758–759
affluent society 4:759
alcohol 6:658–660, 6:662
and business anthropology 3:8
collaborative 4:139–144
conspicuous 4:695–699
conspicuous consumption 4:757–758
conspicuous leisure 4:758
consumer culture 4:758
consumer psychology 4:738–740
as creativity 8:325
criticism over consumer society and consumerism during postwar period 4:759
cultural 13:848
cultural capital 4:760
and cultural production 5:608, 5:610–611
and culture 5:537–538
of culture 5:732–733
downshifting, sustainable consumption, and green consumption 4:760–761
economic sociology 7:70–71
ethics of 3:19
food 9:286
geography of 4:744–750
body 4:747–749
connections 4:746–747
ethical trade 4:746–747
global middle class 4:744–746, 4:749
globalization 4:744–746
habitus 4:759–760
history of
changes in focus 4:754–755
definition 4:751
intellectual background and theories 4:751–752
knowledge and achievements 4:752–754
methodology 4:755
identity formation 4:758–759
and industrial societies 11:863–866
land use indicators 13:241–242
production-consumption interlinkages 13:242f
lifestyle 4:758–761
losses among involuntarily displaced workers 12:824t
mass consumption society 4:759
media 16:785
money 15:736–737
news 16:793–797
omnivore hypothesis 4:760
postmodern sociological approach 4:759
and postmodernism 18:676–679
rational choice explanation 19:899
remittances 12:521–522
risk tradeoffs in 25:17
significance of sociological approaches 4:757
Smith on 22:118–122
social changes and transition in focus of sociology 4:757–758
and social class 22:341–342
and social networks
see also Social networks
social patterning of cultural consumption and social class 4:759–760
in socialism 22:850, 22:852
sustainable 23:839–844
taste 4:757–760
trickle-down theory 4:758–759
in United States 1:618–619
urban settlements 24:836
use of term 16:555
Consumption nodes, commodity chains 4:245
Consumption politics, and commodity chains 4:245
Consumption smoothing, job displacement insurance 12:819–820, 12:823, 12:825
Consumption taxes 23:339
Consumption-based poverty measures 18:742–743
Contact
and adoption 1:144–145
and collective behavior 4:155
Contact experiences, of minorities 8:144
Contact hypothesis 6:219–220, 6:226–227, 12:410, 18:843–844, 19:219
Contact period 1:924–925
Contact theory 19:834
Contact zones 21:154
Contagion 23:686–687
in financial crises 9:185
Contagion processes, diffusion 6:411–413
Contagious crying, and empathy 7:551
Contagious distributions 23:468
Container models 1:66
of learning 7:415–416
Containerization 14:531
Containment, in Cold War 4:128–129, 4:131
Contaminated areas, living in 7:806
Contamination 25:107
Contemplation 21:172
Contemporary anthropologists 25:667
Contemporary British History 4:765–766
Contemporary concentrated urban poverty 24:929
Contemporary contractarianism and contractualism 4:807–808
Contemporary dictatorship 6:375–376

definitions and typologies 6:376
and dictator 6:375–376
dictatorship, tyranny, despotism 6:375
Contemporary economic methodology 7:102
Contemporary history 4:762–768
approaches, sources, and methods 4:766–767
defined 4:762–763
history of 4:763–764
nineteenth-century France 4:764
since the 1960s 4:764–766
Contemporary identity 11:535
Contemporary military geography 15:502–504
land use 15:502
military activities 15:502–503
military landscapes 15:503
military tactics 15:502
scope of 15:502–504
terrain 15:502
Contemporary models, time series 24:343–344
Contemporary neoliberal interrogation 5:257–259
Contemporary progressive political movements 19:765
Contemporary psychology of religion
developmental psychology 20:355–356
individual differences 20:355
mental health 20:355
mystical and religious experience 20:356
Contemporary publishing 7:588
Contemporary qualitative inquiry 19:676
Contemporary semiperiphery 6:196
Contemporary slavery 22:43
defining 22:43–44
estimating 22:44–45
forced marriage 22:44–45
policy measures to mitigate spread of 22:46–48
prevalence of 22:45
risk factors 22:45–46
scholarship on 22:45
sex trafficking 22:44
Contemporary social movements 15:19
Contemporary social systems 22:747
Contemporary society 13:108
Contemporary sociology 22:999
Contemporary systems theory 22:749
Contemporary tourism 9:293
Contemporary urban sociology 24:931–932
Contemporary warfare 25:378
Contemporary world 25:811
Contempt, and disgust 6:547–548
Contempt of court 15:593
Content
of communication 4:286–287
economic education 7:21–22
importance in future orientation 9:563
of interview 5:851–852
and organization 21:230
representational 20:67, 20:69–70
stereotypes 23:448
Content, census 3:291–292
core topics 3:291–292
noncore topics 3:292
Content Addressable Distributed Associative Memory (CADAM) 15:162
Content analysis 4:769–773, 9:362
classical or thematic 4:769–770
for embodiment 7:418
in ethnography 8:189
inferences to events *versus* perspectives 4:772
linguistic expressions 4:772
modal auxiliary verbs, use of 4:772
Osgood's contributions 4:769
qualitative health research 10:694
researcher as novice *versus* expert 4:771–772
of text populations 4:772–773
theme-relations within networks *versus* grammars 4:770–771
Content and Language Integrated Learning (CLIL) 2:580–582
Content and Service Registry (CSR) 23:134
Content biases 24:130
Content errors 6:127
census 3:293–294
Content internalization 5:578
Content of media coverage of climate change 3:858
Content producer 17:218
Content providers (CPs) 20:167–168
Content validity 4:774–777, 6:345, 7:24, 20:240
construct relevance 4:774–775
data analysis 4:775–776
evolving standards 4:776
expert judgment 4:774–775
history and terminology 4:774–775
representativeness 4:774–775
study design 4:775
domain definition 4:775
judgment process 4:775
selecting and training judges 4:775
support of test use 4:774
Content-involving explanation, of representations 13:77, 13:80, 13:82–83
Contentious jurisdiction 12:510
Contentious politics 22:435, 22:443
Contentment 2:791
Content-related evidence 3:822
Contestation
agonism 1:447
culture in 9:3–4
of Western intellectual hegemony 25:549
Contested commodities 5:538
Contested field 7:117
Contests, storytelling 16:221
Context biases 24:130
Context conditioning 8:849–850, 8:852–853
Context dependent memory 4:778–785
context-dependent recognition and context-dependent discrimination 4:780
factors affecting context effects 4:779–780
familiarity *versus* recollection 4:781–782
global *versus* local context 4:778, 4:780
ICE model 4:781
incidental *versus* intentional encoding of context 4:782–783
interference paradigms 4:779
item–context associations 4:781–782, 4:784
mental reinstatement 4:780
meta-analysis 4:780
nature of context and target stimuli 4:782
neuroscience studies of context effects 4:783–784
outshining 4:780
overshadowing 4:780
reinstatement paradigm 4:779, 4:779f
renewal effect 4:779
Context factors, offender rehabilitation 20:197–198
continuity of care 20:198
custody *versus* community 20:197
institutional climate 20:197–198
'natural' protective factors 20:198
staff factors 20:198
Context of evaluation, of logic models 14:300, 14:303
Context of food choice 9:298–299
Context of reception 2:85
Context reinstatement effects on recognition memory 20:41–42
Context setting
early childhood care and education (ECCE) 6:739–740
Context-based courses 21:230
Context-dependent discrimination 4:780
Context-dependent models
multiattribute linear ballistic accumulator model 5:904
of response time and choice 5:904–905
Context-dependent recognition 4:780
Context-free description of a language 14:296
Context-free grammars 19:10, 19:12–13
computation 19:13
inference 19:13
probabilities 19:13
relation of Gibbs distributions to 19:14
Context-free linear accumulator models 5:903–904
additive drift rate variability 5:904
multiplicative drift rate variability 5:903–904
relation of multiplicative LBA models with no start point variability and Fréchet drift rate variability to Luce (MNL) models 5:904
Context-free rules 14:296
Context-independent science system 21:152
Context–individual relations, adolescent self-regulation 21:547–552
Context-informed attachment research 2:141–142
Contexts
changing, linked lives in 14:232–233
culture as 17:615–616
in decision and choice 5:895–896

Contexts (*continued*)
and deixis 6:53
discourse as sociohistorical meaning-making practice in 6:510
effect on choice options 19:440, 19:444
effect on episodic encoding 7:857–858
effects of implicit attitudes measures 2:181
effects on segregation 21:407–409
effects on sentences 21:624
factors in lifestyle development 14:102
and flexible habitus 22:335
and life course 14:80, 14:82–83
and media systems 22:584
and neuroesthetics 16:659–661, 16:660f
-oriented discourse analysis 6:511
in qualitative research 19:686–687
question 19:771
and representation 20:70–71
role in diversity effects 4:740–741, 6:597
in social and behavioral sciences 15:323
of sustainable practices 23:834–836
temporal context models 24:201–202
youth culture 25:810–811
Context-sensitive rules 14:296
Contextual characteristics, as source in embodiment 7:410–411
Contextual factors 21:396–398
ecology and health 6:878–879
Contextual features 13:943
Contextual implication 5:88
Contextual integrity 12:612
Contextual interference effect 16:846
Contextual learning 15:156
Contextual performance (CP), organizational citizenship behavior 17:318
Contextual representation 16:198–199
Contextual social work 12:558
Contextual studies 4:786–789
contextual effects 4:788–789
ecological perspective 4:787
history of 4:786–787
methodological challenges 4:787–789
hierarchical nature of contextual influence 4:788–789
objective *versus* subject assessment 4:787–788
transactional nature of contextual influence 4:788
Contextual theory of meaning, Feyerabend on 9:122
Contextual variables 16:49
cross-cultural training 5:338
Contextualism 23:326
Contextuality, in critical social work 5:291–292
Contextualization, in ethnography 8:185
Contextualized learning
lifelong learning 14:96–97
Contextualized learning, lifelong learning 14:96–97
Contiguous conditioning 13:703
Continent valuation methods, valuation of benefits using 5:51–52
Continental ancestry 9:899
Continental Europe 13:792
Continental European Law 13:410, 13:412
Continental European perspective, of law and morality 13:479–482
Continental expansion 4:224–225
Continental philosophy 1:895–896, 2:359, 2:361, 20:762, 20:765
Continental realism 13:792
Contingencies
and common sense 4:1–4
leadership as 13:649–650
Pearson on 17:661–662
and self-efficacy 21:504–508
in social systems, Luhmann 14:385
strategic adaptation 23:501–502
structural contingency theory 23:609–614
Contingency analysis 4:769
Contingency management 2:409
Contingency principles
reinforcement 20:208
Contingency sentence 23:716
Contingency tables 5:1, 23:425
loglinear models 16:125–128
Contingency theory 2:915–917, 7:219
organizational design and form 6:236
Contingency thesis 22:230
Contingent fees 12:922
Contingent markets, stated preferences in 25:17–18
Contingent tolerance 13:34–35
Contingent valuation (CV) 5:51–52
cost–benefit analysis 5:51–53
ecological economics 6:856
of environmental cognition 7:709
Contingent work 18:665, 25:692
Continuing education 24:759
Continuing Legal Education (CLE) 13:738
Continuing medical education (CME) 18:10
Continuing professional development (CPD) 4:796–802, 13:738, 19:89–90, 22:779–780
challenges 4:801
leadership involvement 4:801
model 4:797
research base 4:797
shifts in thinking 4:797–800
social work education 22:779–780
in teacher education 24:54
Continuing professional education (CPE) 19:89–90, 19:92–93
time and focus of 19:89t
Continuing teacher professional learning 4:796
assessing impact 4:800–801
deepening professional knowledge 4:800
engaging students in new learning experiences 4:800
identifying students' knowledge and skills 4:799
identifying teachers' knowledge and skills 4:799–800
leadership 4:801
shifts in thinking 4:797–800
system challenges 4:801
Continuing Vocational Training Survey (CVTS) 22:22–23
Continuity 18:47
adolescent personality 17:864–865
in attachment 12:232
comparative evolutionary psychology 4:394–398
identity, nonheterosexual British Asians 16:864
and phonological disorders 18:47
Continuity hypothesis
evolutionary economics 8:434–435
Continuity thesis 16:329, 18:60
Continuous branching 11:820
Continuous core–periphery method 4:913–914
Continuous effect indicators 11:752
Continuous first-hand contact, in acculturation 1:68
Continuous Instrument Improvement Group (CIIG) 5:745–746
Continuous latent variable 11:752
Continuous light optical imaging 16:388–389
Continuous measurement, census 3:305–306
Continuous optimal designs 17:257–258
Continuous probability distributions 6:575
definitions and properties 6:575–576
individual continuous distributions 6:576–578
Beta distribution 6:576
Cauchy distribution 6:578
Chi-square distribution 6:577–578
exponential distribution 6:577
F distribution 6:578
Gamma distribution 6:577
normal distribution 6:576–577
Pareto distribution 6:578
Student t distribution 6:578
uniform distribution 6:576
Weibull distribution 6:577
multivariate distributions 6:578–579
bivariate normal distribution 6:579
probability density function 6:575
Continuous quantifiers 19:708–709
Continuous quantity, concept of 14:868–870
Continuous random variables, goodness-of-fit testing for 15:654
Continuous time, event history analysis in 8:310
Continuous time microsimulation 15:343
Continuous time models 23:470
for event–history analysis 14:353
for social analysis 14:580–582
Continuous traits, cultural evolution 5:404
Continuous univariate distributions 11:886
Continuous variables 11:751–752
multivariate displays 10:354
Continuous wave (CW) 16:388–389
Continuous-gene-flow (CGF) model 9:887
Continuously varying traits 4:416–420
Continuous-time, discrete-state processes, in event history analysis 8:305

Continuous-time Markov-transition process 4:420–421
evolution of political complexity 4:420–421
genotypic sex determination and adaptive radiations in the sea 4:421
Continuum model 23:449
Contour maps 24:266
Contour orientation 25:185
Contours, of Eastern Europe 6:816–817
Contra proferentem principle 12:216
Contraception 8:786–790, 9:3–4, 9:77, 11:194, 20:458
in affluent nations 20:469–470
in East Asia 9:56, 9:58
and fertility transition 9:51–52
in Latin America and Caribbean 9:69–71, 9:70f, 9:71t
prevalence, in Eurasia 18:547–548
of single/cohabiting women 9:14, 9:16–17
in South Asia 9:81–82
in Southeast Asia 9:87–88, 9:86t
Contraception law, in Catholicism 3:235–236
Contraceptive revolution 17:486
Contraceptive technologies 9:24
Contraceptive use, multilevel models 16:48–52
contextual variables 16:49
cross-level interactions 16:50
empirical Bayes 16:51
fixed coefficients 16:50
level-specific models 16:49
predicted probabilities 16:51–52
random intercepts 16:49
random slopes 16:49
variance parameters 16:50–51
between and within-group estimates 16:50
Contraceptives 21:709
and child fertility 8:821–824
and social networks 9:7–9
Contract approach, for strategic planning 23:519
Contract archaeology 10:53
Contract compliance program 1:257–259
Contract governance 10:295
Contract hypothesis, for persuasion 17:945
Contract law 1:204, 3:27, 4:942–943, 13:450, 13:741
Contract of employment 7:572
in United Kingdom 7:572
in United States 7:572
Contract research organizations (CROs) 18:7–10
Contract theory 4:942–943
Contract zone of settlements 6:558
Contractarianism 4:803–808
bargaining 4:805–806
business ethics 3:17
constitution 4:805–806, 4:808
contemporary contractarian positions 4:805–807
contract and compliance 4:804–805
and equality 7:909–910
Hobbesian 4:805–807
and intergenerational justice 12:396–400
Kantian 4:807
and morality 15:808
Contract-based property rights 4:338
Contracting 1:345, 1:347
hierarchies and markets 10:841, 10:843
Contracting out 10:258
arrangements 10:297t
Contraction independence 4:877
Contracts 12:103–104, 12:216
commercial impracticability 3:28
in common and civil law systems 4:809–810
consideration 3:27–28
cultural production 5:609
elements 3:27–28
employment 3:29
formation of 4:809
freedom to and freedom from 4:810–811
genuineness of assent 3:28
intend to bound 3:27–28
Japanese law 3:27–28
labor 25:676
legal perspectives 4:809–812
liberal construction of contract terms 4:811
living by 9:615
marriage 6:642
mistake of judgment 3:28
mistakes 4:809
mutual mistake of fact 3:28
neoliberal principles of 4:812
or status societies 1:766
oral 3:27–28
of organizations 17:398
parties to 4:809
principal-agent models 12:104
principle of freedom of contract 3:28
psychological contracts 19:366–371
short 3:27–28
unconscionability, doctrine of 3:28
US law 3:28
use of term 3:27, 4:809
voidable 3:28
Contractual citizen science projects 3:632
Contractual participation, in participatory action research 17:535–536
Contractualism 4:807, 7:926
and intergenerational justice 12:396–397
Contractualists 4:679
Contractualization of social services 24:849
Contracultures 23:638
delinquency 6:70–72
versus subcultures 23:638
Contradictions 16:410–411
and humor 11:418
and issues 19:52
of risk 20:693–694, 20:696
Contradictory consciousness 10:742–743, 13:471
Contradictory nature of capitalism 14:652–654
Contrariness 16:410–411
Contrast (linguistics) 6:432, 12:97–98, 22:212
in education 22:214
Contrast effects 6:430, 6:433, 20:883–884
in attitude measurement 2:178–179
of I/E model 12:428–429
Contrast gain control 25:159
Contrast model 21:580
Contrastive Analysis Hypothesis (CAH) 21:361
Contrat Social 7:657
Contrectationstrieb 21:846
Contribution biplot 5:3
Contribution limits, political money 18:395
Contribution monotonicity 4:874
Contributory citizen science projects 3:632
Contributory expertise 21:210
Control 7:483, 19:736–737, 24:247–248
death/dying 5:867
demand and 17:137
education and 1:546–547
McDonaldization 14:828
over interview 5:851–852
parental 2:471–472
versus perceived powerlessness 1:545
and self-efficacy 21:504–505, 21:507–508, 21:504f
separation of, management 14:471–473
social 4:822–827
sociocybernetics 22:913–914
stockholders 23:490–494
of work 25:689–690
see also Perceived control
Control, syntactic 4:828–833
arbitrary control 4:828
complement control 4:832–833
control shift 4:830
exhaustive control 4:828
explicit control 4:832–833
implicit control 4:832–833
language-specific differences 4:831, 4:833
causative *versus* con-causative coding 4:832
implicit control *versus* explicit control 4:832–833
modal *versus* nonmodal coding 4:831
minimal distance principle 4:829
obligatory control 4:828
partial control 4:828–829
pragmatic approach to control 4:831
raising and control 4:830
semantic approach to control 4:830
split control 4:829
Control behavior 4:813–818
action-phase model of developmental regulation 4:815–816, 4:815f
control beliefs 4:813
control striving 4:813–817
developmental regulation around developmental deadlines 4:816
engagement and disengagement from career goals 4:817
evolutionary origin and human characteristics of 4:813–814
further research and future directions 4:817
goal adjustment and goal engagement in transition to adulthood 4:816–817

Control behavior (*continued*)
motivation and life span development 4:814–815
perceived control behavior 4:813–814
primary control behavior 4:813–817
psychological perspectives of 4:813
role of individual agency 4:814–815
secondary control behavior 4:813–816
Control beliefs 4:813
health perspectives 4:819
alteration 4:820
assessment 4:820
control theory 4:820
control therapy 4:820
helplessness 4:820
locus of control 4:819
perceived behavioral control 4:820
perceived situational control 4:820
self-efficacy, mastery, and competence 4:819–820
self-regulation 4:820
Control conditions, for responsibility 20:588–590
Control constructs, self-efficacy distinguishing from 21:512–513
Control effects
impact of education environments 21:473
Control groups 22:291
creation of 8:568–569
design for workplace safety and health 25:731
Control of reproduction, and eugenics 8:219–220
Control shift 4:830
Control striving 4:813–817
Control systems 2:296
Control theories 4:820, 4:824
and crime 4:834–839
stochastic optimal control theory 6:709–710
Control theories of crime 4:834
empirical support for 4:834–836
sources of control 4:834–836
formal *versus* informal control 4:835–836
internal *versus* external control 4:834–835
varieties of 4:836–837
macro- *versus* micro-control theories 4:836–837
static *versus* dynamic control theories 4:836–837
Control therapy 4:820
Control variables 4:840–843, 17:253
observed variables 4:840–841
propensity score 4:841
regression adjustment in quasiexperimental and nonexperimental designs 4:841
regression adjustment in randomized experiments 4:840–841
regression and ANCOVA based approaches 4:840–841
validity and differential treatment effectiveness 4:841
optimal time path of 17:254
unobserved variables 4:841–843
causal inference 4:840–841
instrumental variable method 4:842
mediator variable method 4:842–843
observational research 4:841
Controllability 15:897
Controlled clinical trials developments 19:386–387
in medicine 19:386
in psychological treatments 19:387
in psychology 19:386–387
Controlled comparison 4:412–413
Controlled direct effect (CDE) 15:76
Controlled experiment, by Thorndike 24:305
Controlled motivation 15:931, 21:486
FTP and FO 9:563
Controller–controllee relation 10:317
Controlling shareholders 4:933
Control-value theory, I/E model 12:431
Controversies
new technologies 24:126–127
scientific 21:281–286
Convenience sampling 5:321
Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT) 12:956, 18:450–451, 24:442
Convention Concerning the Protection of World Cultural and Natural Heritage (1970) 5:488–489
Convention on Biological Diversity (CBD) 11:775, 11:802, 23:796
Convention on Contracts for the International Sale of Goods (CISG) 3:30
Convention on Elimination of All Forms of Racial Discrimination (CEAFRD) 11:357
Convention on International Trade in Endangered Species of Wild Fauna (CITES) 24:599
Convention on Long-Range Transboundary Air Pollution (CLRTAP) 1:724
Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) 5:495–496, 6:629–630, 6:793, 8:907–908, 10:192–193, 11:357, 11:363, 11:375–376, 12:956, 13:720, 20:469, 21:792, 24:752
Convention on the Elimination of All Forms of Racial Discrimination 11:796–797
Convention on the Elimination of All Forms of Violence against Women 20:459
Convention on the Prevention and Punishment of the Crime of Genocide 10:6
Convention on the Prevention of Major Industrial Accidents, and EU Directives 17:143
Convention on the Rights of Persons with Disabilities (CRPD) 5:495, 11:357, 13:497, 13:499–500, 23:213
committee 13:500
defining disability 13:500
introductory set of provisions 13:499
law and policy reforms 13:501
national level monitoring requirements 13:500
obligations 13:499–501
Optional Protocol 13:497
preambular paragraphs and articles 13:499
precondition of legislating for disability rights 13:501
rights of stakeholders 13:499
Convention on the Rights of the Child (CRC) 3:439, 3:485, 3:487, 3:489, 11:357
Convention refugee 20:105
Conventional approach
to rationality and emotionality 7:472
Conventional arbitration 6:558
Conventional game theory 9:587–588
Conventional monetary policy
down to zero lower bound 15:728
new ways of implementing 15:732–733
Conventional pluralism 1:447
Conventionalism 2:255, 2:257, 2:259, 8:576
and incest 11:730–731
Conventionalist approach to incest 11:730
Conventionalist view of scientific theories 8:576
Conventionality, and conscientiousness 4:659
Conventionality of meaning (CM) 14:844, 14:848–849
Conventions 1:610, 4:844–849, 12:191
as Agreement of action 4:844–845
defined 4:844–847
emergence 4:847
fashion as break with 8:837
institutions 12:190–194
rational choice explanation 19:899
solutions to recurrent coordination games 4:845–847
Lewis' definition 4:845–847
Convergence 4:152
and broadcasting regulation 2:861–862
consumer policy 4:734–735
corporate governance 4:937
economic growth 7:39–41
and epistemic cultures 7:878
increasing 12:880
language contact 13:305
Markov chain Monte Carlo methods 14:571
oculomotor control 8:630
between trade unions and social work practice 24:491
Convergence culture 12:588, 12:590
television 25:81–82
Convergence exchange theories 22:958
The CONvergence of iterated CORrelations *see* CONCOR (The CONvergence of iterated CORrelations)
Convergence of likelihood sequence 4:471

Convergence on minimal punishment 19:636, 19:642f
Convergence reasoning 19:594–595
Convergent thinking 17:770
Convergent validity 6:345, 21:461, 21:466
versus discriminant validity 3:70
Conversation
and collective memory 4:195–196
and linguistic anthropology 14:172–173
Conversation analysis (CA) 8:206, 9:599–600, 18:797–798, 20:881–882, 21:193–194, 21:362, 22:546
constructing turns and identifying actions 4:850–854
epistemics 4:852–853
knowledge 4:852–853
methods 4:852, 4:854–855
practical reasoning 4:850, 4:852
social constructionism 22:225
social interaction 4:850, 4:855
word searching and repair 4:854–855
Conversational abilities 4:300
Conversational case-based reasoning (CCBR) 19:57
Conversational implicature 18:796–797
Conversational interviewing 20:882–883
Conversational maxim approach 20:883–884
of politeness 18:327
Conversational style, gender differences in 9:754
Conversational/dyadic interactions, and collective memory 4:190–191
Converse, J.M., on interviewing 12:671
Converse, Phillip E.
demonstration of 11:557–558
formulation of political sophistication 11:555–557
Conversion disorder 23:11
explanatory style 8:599
Conversion hysteria 11:507–508
Conversion mixed methods designs 15:620–621
Conversion model, for prejudice reduction 18:837
Conversions
religious 20:356, 20:363–364, 21:382
Christianity 15:614–617
Judaism 12:801, 12:803, 12:808
Convex function, majorization/stochastic orders 14:450
Convex multistage problems, stochastic optimization of 23:478–479
Convict criminology 18:926–927
Conviction mode, disparities across 21:630–631
Convincing role models 2:583
Convolution 25:159
Fourier transform 9:376–377
Convulsive therapy 19:587–588
Co-occurrence data 16:43
Co-occurrence hypotheses, for persuasion 17:945
Co-occurring conditions, autism spectrum disorders 2:276–277
Cook, James 7:791, 17:161
Cook, Thomas, works, *Quasi-Experimentation: Design & Analysis Issues for Field Settings* 3:247
Cook Islands 17:159, 17:161–162
Cooke, Alistair 19:538
Cooke, B., works, *Participation: The New Tyranny?* 17:538
Cooking 9:291
and use of fire 9:197–199
Cool Japan 5:457
Cooley, Charles Horton 8:464
on shame 21:873–874
Coombs, Clyde H. 14:814–815, 19:474
Coon, Carleton, works, *Anthropology A to Z* 7:589–590
Cooper, Anne, works, *Sites of Gender* 11:21
Cooper, J.P. 7:651
Cooper, Robert, works, *Language Planning and Social Change* 13:354
Cooperating agents 2:310
Cooperation 4:460, 4:856–866, 16:332, 16:643–645, 18:798, 22:246–248, 22:264, 22:266, 22:702, 22:905–906
achievement 4:857–859
and agreeableness 1:465–466
altruistic reciprocity 4:865
among rational egoists 4:862–865
embedding 4:863–865
among socially rational individuals 4:864–865
framing effects 4:865–866
goals 4:865
basic elements 4:859
behavioral learning theory 4:856–857
cognitive developmental theory 4:856
collaboration 4:856–857
competition 4:864–865
competitive and individualistic efforts 4:859
in division of labor 13:157
enhancing factors 4:859
evolution, bounded rationality for 2:775
game theory 4:862–866
gender differences in 9:753–754
in aggression 9:753–754
in females 9:753
in males 9:753
in social play 9:754
and groups 4:863–866, 10:434
and Honesty-Humility 11:170
interaction patterns 4:857–858
interpersonal relations 4:858–860
irrigation 12:725
of joint action 12:839
kin selection 13:29–30
networks and 14:525
noncooperative games 4:865
oppositional interaction 4:857–858
outcomes 4:858–859
and politeness 18:326
prisoner's dilemma 4:863, 18:931–932, 18:934–936, 4:863f
promotive interaction 4:857–859
and prosocial behavior 1:580–582, 19:229–230, 19:233, 19:235–236
psychological health 4:858, 4:860
psychological states 4:858
rational choice theory 4:863
rationality 4:862–863, 4:865–866
reciprocity 4:863–865
signaling 4:862, 4:865–866
and social capital 22:177
social cognitive theory 4:857
social dilemma 4:862–863
social interdependence theory 4:857
application of 4:859–860
breadth of 4:860
relationship among theory, research, and practice 4:860
in social networks 19:910
social norms 4:864
social preferences 4:862, 4:865
social rationality 4:862, 4:864–866
solidarity 4:862
and trust 4:862–865, 24:668–670
and water wars 25:444–447
Cooperation agreements, criminal procedure 5:231
Cooperation games, experimental economics 8:549
Cooperation in Science and Technology (COST), European Union's research program 17:672
Cooperation of state and religions 3:576
Cooperative behavior 21:530
Cooperative breeding 2:140–141, 13:48
Cooperative competition, and artists 2:54
Cooperative decision-making 15:775
Cooperative ethos, and communitarianism 4:312
Cooperative game theory 4:867–880, 6:604
and power 15:669
Cooperative Integrated Reading and Composition (CIRC) 4:884
Cooperative learning 1:871, 3:914, 4:859–860, 4:881–886, 10:416
directions for further research and development 4:885
group goals and individual accountability 4:882–883
informal group learning methods 4:884–885
group investigation 4:885
jigsaw 4:884
learning together 4:884
peer tutoring 4:884
research on alternative pragmatic approaches to 4:883
Structured Team Learning methods 4:883–884
Cooperative Integrated Reading and Composition (CIRC) 4:884
IMPROVE 4:884
Peer-Assisted Learning Strategies (PALS) 4:884

Cooperative learning (*continued*)
Student Team Learning (STL) 4:883
Student Teams-Achievement Divisions (STAD) 4:883
Team Assisted Individualization (TAI) 4:884
Teams-Games-Tournament (TGT) 4:883
teachers' use of 10:420
theoretical perspectives on 4:881–882
Cooperative marriage 21:735
Cooperative *n*-person games 16:583
Cooperative problem-solving 4:316
Cooperative responding, in MMPI 15:566–567
Cooperative system, organization as 17:317
Cooperative work, computer-supported 21:197
Cooperativeness principle 20:882
Cooperatives 17:202–203
Cooperators
full 1:580
hypocritical 1:580
private 1:580
Cooptation 15:21–22
collective political action 4:149
and creative freedom 5:600–601
in organizations 1:104
Coordinate Structure Constraint (CSC) 16:4
Coordinate systems 14:285
motor control 3:778–779
Coordinated autobiographies 6:495–496
Coordinated community action models (CCAM) 12:682–683
Coordinated expectations 4:845
Coordinated market economies (CME) 11:831, 22:21, 22:708
Coordinates, graphics 10:368
Coordination 4:845, 10:313, 18:798
by constitutions 4:712
in division of labor 13:157
governance as 10:293
in moral decision-making 15:805
in organization theory 1:106–107
political protest 18:422
public sector organizations 19:601
Coordination games 4:845, 11:260
experimental economics 8:549
Coordination-level issues
supply chain management 23:710
capacity decisions 23:710
cash flow decisions 23:710
information flow decisions 23:710
material flow decisions 23:710
Coover, John E. 8:568
Coparenting 22:158
COPC *see* Community-oriented primary care (COPC) model
COPD *see* Chronic obstructive pulmonary disease (COPD)
COPE *see* Committee on Publication Ethics (COPE); Coping Orientation to Problem Experience (COPE)
Copeland, Morris A. 16:228
Copenhagen Consensus 12:547
Copenhagen Film Festival 3:666
Copenhagen Interpretation 17:89
Copenhagen model 21:396
Copenhagen talks on Green House Gas Reduction (2009) 7:771
Copernicus, Nicolaus
Copernican Revolution 13:135, 20:285–286
works
censorship of 3:279, 18:100
De revolutionibus orbium coelestium (*On the Revolution of the Heavenly Orbs*) 3:279, 20:285–286, 20:637
Coping 4:887–896, 17:872–873, 23:551–555, 23:566–567
adaptation 4:902
adversity 4:888–889, 4:891, 4:893
anticipatory 4:903
applications of research 4:900
appraisal 4:903
appraisals 4:887–890, 4:893
with bereavement 2:533–535, 5:862
with cardiac disease 3:111–116
challenge 4:887, 4:889
with childhood cancer 3:459–460
supporting with diagnosis 3:460
with chronic disease 1:586–589
with chronic illnesses 3:549–554
by children 10:684
collective 4:903–904
concept of 4:902
with death and dying 5:862
defense 4:902
definitions 4:887–888
depression 6:208–209
descriptive correlational studies 4:896–897
developmental perspectives on 4:891–893, 4:892f
as regulation 4:891
as strategies of action regulation 4:891–892
disability 6:467, 6:469–470
effectiveness 4:897–898
and emotion regulation 7:455
and emotional expressions 4:895
and emotional intelligence (EI) 4:899–900
conceptualizing coping as EI in action 4:899
empirical studies of EI and coping 4:899–900
emotion-focused 17:872–873
with family problems, and religion 20:320
divorce 20:320
and five-factor model (FFM) 4:895–897, 4:897t
gastrointestinal disorders 9:607
and gender 9:705–706
general coping theory 4:903
with grief model 5:862
and health 23:552–554
benefit finding 23:554
cognitive-behavioral stress management 23:553
emotional expression 23:553
mindfulness-based stress reduction 23:553
relaxation training 23:552–553
health psychology 10:676–677
Heider on 10:753
historical contexts 4:888–889
during childhood, adolescence, and aging 4:889
defense 4:888
interventions 4:889
stress 4:888–889
historical development of 4:902–903
human development 11:273
identity threats, nonheterosexual British Asians 16:863–864
with illness and stress
and personality 17:794–795
with job stress 4:902–904
individual and organizational perspective 4:903–904
methodological issues and future directions for research 4:904
methods 4:903t
research 4:904
life management 14:85–88
mastery 4:889–891
measurement models 4:896
accommodative and meaning-focused coping 4:896
engagement *versus* disengagement 4:896
proactive coping 4:896
with mental illness
and culture 5:571–572
and neuroticism 16:745–746
and optimism 17:264–265
with pain 17:453–454
and personality 4:895–901
as personality process 4:897
preventive 4:903
proactive coping 4:903
problem-focused 17:872
with psychological stress 23:585
with racial discrimination, academic motivation 19:792
resilience 4:891–892, 20:555–557
resources 4:902–904
risk 4:891–893
self-assessments of 17:873
skills, and relapse prevention 20:218
social comparison 22:211, 22:214
social support 4:887, 4:889–890
strain 4:902–904
stress 15:103
with stress 4:895–896
stress management 4:903–904
stressful life events 4:887–888
stressors 4:887–890
teacher stress/burnout 24:73–74
test anxiety 24:248
theoretical frameworks 4:895–896
threat 4:887–889, 4:892
transactional perspectives on 4:889–891, 4:890f
as process 4:890–891
research 4:889–890
to treatment, in behavioral medicine 2:467
vulnerability 4:887–888

work-related stress 23:547–548
Coping, relational 20:224–226
aim 20:225
facilitator/relational guide 20:225
graphic representation of 20:224–226
network 20:223
facilitation of 20:224
and its freedom of action 20:223
members of 20:224
network of motivated people 20:225
surrounding environment 20:225–226
time 20:225
Coping congruence between partners, dyadic coping as 11:593–594
Coping inventory for stressful situations (CISS) 4:896
Coping motives, and adolescent health 1:124
Coping Orientation to Problem Experience (COPE) 3:111–112
Coping responses, effects of media use on children 15:37
Coping skills therapies 4:543–544
Coping strategy, definition 16:862
Coping styles, hypertension 11:454
Coping system 15:948–949
Coping theory 20:351
Coppe, Abiezer 7:654
Copper Age, Europe in 8:270
Coproduction
of knowledge 21:218
of science and society 21:265–266
society and energy 7:615, 7:619–620
of solutions to problems, and green social work 10:385, 10:387
COPS *see* Criterion Orientated Personality Scales (COPS)
Coptic Orthodox Church 17:437
Coptic-Gnostic texts 3:546
Copy number variations (CNVs) 15:212, 15:338–339
Copy principle 7:568
Copyright Act 1710 12:263, 18:916
Copyright Act 1976 12:269
Copyright issues, journals, open access 12:867
Copyright law 12:266, 12:268–271
Copyrights 2:3, 3:31, 12:267–271
and Aboriginal art 11:773–774
of cultural content in Internet 12:589
law 13:524–525
publishing 19:620, 19:622–623
and secrecy 21:375
and trademark protections 12:270
Coquelin, Charles, works, *Dictionaire de L'economie Politique* 7:585
COR *see* Conservation of resources (COR)
Corbett, Elizabeth, works, *New Amazonia: A Foretaste of the Future* 14:256
Corcoran, Michael 13:32
Cordon sanitaire, populism 18:611–612
CORDS *see* Civil Operations and Rural Development Support Program (CORDS)
CORE *see* Congress of Racial Inequality (CORE)
Core 9:464, 9:468, 22:455–456
versus fringe imagery 18:247–248
powers 25:342
states 25:341–342
Core affect, in emotional expression 7:459
Core based statistical areas (CBSAs) 24:857
Core beliefs 4:543
Core center 4:871
Core competencies 10:720–721, 23:509
in medical social work 10:720–721
Core databases 5:844–850
barometer surveys 5:848–849
Afrobarometer 5:849
AmericasBarometer 5:849
Arab Democracy Barometer 5:849
Asian Barometer Survey 5:849
Eurobarometer 5:848–849
Latinobarometer 5:849
New Democracies Barometer 5:849
education surveys 5:847
international 5:847
national 5:847
general social surveys 5:845
international comparative research *see* Comparative databases
methodological characteristics of quantitative surveys 5:844
methodologies 5:844
sampling techniques 5:844
selected household panel studies of income and living conditions 5:845–847
defining panel studies 5:845
European-wide panel studies 5:846–847
examples of national panel studies 5:845–846
standardized survey 5:844, 5:846–849
thematic social science surveys 5:847–848
crime and victimization surveys 5:848
health-related surveys 5:848
labor-related surveys 5:847–848
national archives and statistical institutes 5:847
types of core national datasets 5:844–845
Core disgust 6:547
Core identities, in Europe 8:276
Core imagery 18:247–248
Core insights and predictions 1:345–347
Core knowledge 7:868, 12:325–326
Core model of social fluidity 15:643
Core self-evaluations (CSE) 4:905–910, 12:837
affective outcomes 4:907
broad in scope 4:905
cognitive outcomes 4:907–908
behavioral 4:907–908
mediating effects 4:908
moderating effects 4:908
evaluation focused 4:905
fundamental 4:905
issues 4:908–909
measuring 4:906–907
direct approach 4:907
indirect approaches 4:906
outcomes 4:907
trait indicators of 4:905–906
emotional stability 4:905
generalized self-efficacy 4:905
locus of control 4:905
self-esteem 4:905
Core Self-Evaluations Scale (CSES) 4:907
Core set, scientific controversies 21:283, 21:285
Core societies, world systems theory 25:748–749
Core subject knowledge, and media 15:7
Core symbols 23:861–862
Core theory 15:669
and power 15:669–670
Core topics of population censuses 3:291–292
Core values 19:564
death 5:871
Coreness 4:913–915
Core–periphery analysis, and development geography 6:314–316
Core-periphery hierarchy, dependency theory 6:196–197
Core–periphery patterns
in networks 4:911–916
block models 4:911–913, 4:915, 4:911f
continuous model 4:913–914
coreness 4:913–915
discrete model 4:912–913
extensions 4:914–915
social networks 4:911, 4:916
social structure 4:916
software, applications and scalability 4:915–916
Coresidence 12:405
households 11:228, 11:230–231
relationships 8:725
Co-resident domestic group 11:220
Coring archaeological sites 10:54
Corinthians 3:237
Coriolis forces 3:780
Corn laws 20:660
Corn model of agricultural production 25:7
Cornell Conferences on Therapy 18:157, 18:159–160
Cornell University 24:762–763
Cornette, J., works, *État Baroque* 1:9
Cornish, D., works, *Modeling Offenders' Decisions: A Framework for Research and Policy* 20:776
CORO *see* Conference of Regulatory Officers (CORO)
Corona Program 20:421–422
Coronado, Hernando 1:633–634
Coronary artery bypass grafting 20:190
Coronary atherosclerosis 4:917–921
Coronary heart disease (CHD) 4:917–922, 6:210, 9:695, 17:137, 22:125, 22:735
across the world 4:917–919
acute *versus* chronic psychosocial factors 4:918–919
methods of investigation 4:918
psychosocial factors 4:917–918
acute coronary syndrome 4:920

Coronary heart disease (CHD) (*continued*)
acute psychosocial triggers of cardiac events 4:920–921
implications for prevention 4:921
physiological mechanisms underlying triggering 4:921
coping with 3:111–114
coronary atherosclerosis 4:917–921
depression 4:917–921
and emotions 7:497–498
inflammation 4:920–921
lifestyle factors 4:920
myocardial infarction 4:917–919, 4:921
and physical activity 18:87–88
positive well-being 4:917–920
psychophysiological stress testing 4:918, 4:921
psychosocial determinants and impact on 20:188–189
and religion/spirituality (R/S) 20:266
risk, chronic psychosocial adversity and 4:919–920
emotional responses 4:919
mediating mechanisms 4:920
other forms of chronic stress 4:919
protective psychosocial factors 4:919–920
work stress 4:919
social isolation 4:917–921
social support 4:917–919
stress and recovery following acute cardiac illness 4:921
implications for patient care 4:921
stress management 4:921
Coronary thrombosis, and smoking 14:69–71
Corpora (linguistics) 4:949, 5:812–814
annotation types 5:813–814
components 5:812
data sources 5:813
metadata 5:814
uses 5:812–813
Corpora mystica 12:266
Corporal punishment 3:479, 5:232–233
Corporate consolidation, in publishing 19:627–628
Corporate control 23:490
Corporate crime 5:165, 25:551–552, 25:555–556
Corporate cultures 3:24, 4:923–926, 5:536
artifacts 4:924
business codes 8:23
changing 4:926
conceptual model for analyzing culture 4:924–925
creation 4:926
definition 4:923
description 4:925–926
espoused values 4:924–925
formal definition of culture 4:924
organizational culture analysis 4:923–924
subculture 4:923–924, 4:926
tacit assumptions 4:923–925
United States 17:348
Corporate curriculum 3:13–14
Corporate discipline, sociology of professions 19:107
Corporate diversity management 17:392
Corporate elite group 15:536
Corporate entrepreneurship, strategic adaptation 23:503–504
Corporate environmentalism 17:377, 19:82
Corporate finance 3:22, 4:927–931, 4:946
budgeting 4:927–928, 4:930
capital budgeting 4:927–928, 4:930
capital invested 4:928–931
capital structure analysis 4:930–931
decisions 4:928
dividends 4:927–930
financial accounting 4:929
financial planning 4:929–930
integrated system of financial analysis 4:928–929
investment decisions 4:927, 4:929–930
investment policies 4:927, 4:929
investment valuation 4:930
ratios 4:927–929
shareholders 4:927–931
Corporate geography 10:69
Corporate governance 1:347, 3:23, 3:29–30, 4:932–938, 13:667–668, 23:490
in China 4:936
convergence 4:937
corporate law 4:932–934, 4:936–937
goals 4:933–934
controlling shareholders *versus* minority shareholders 4:933
other stakeholders 4:934
shareholders *versus* creditors 4:933
shareholders *versus* employees 4:933–934
shareholders *versus* managers 4:933
international differences 4:935
Anglo-American 4:935–936
emerging economies 4:936–937
Europe and Japan 4:936
means of 4:934–935
board of directors 4:934
fiduciary duty, antifraud rules, and shareholder suits 4:935
gatekeepers 4:935
institutional shareholder activism 4:934
markets 4:935–936
voting 4:934–935
officers 4:932–935
problem 4:932–933
recent developments 4:936
federalization 4:937
gatekeepers 4:937
independent directors 4:936
institutional shareholder activism 4:936
OECD principles 4:937
rise and fall of good faith 4:936–937
Rule 14a-8 4:936
Corporate Governance Guidelines 8:22
Corporate groups 13:758
and ethnicity 8:148
Corporate hierarchies, promotions 19:135
Corporate income tax 4:948, 11:746
Corporate law 3:21, 4:932–934, 4:936–937, 5:269
for closely held corporations 4:939
corporate governance, role in 4:939
publicly traded corporations 4:939–945
shareholder protection and theories of the firm 4:942–943
Corporate lawyers 16:418–419
Corporate legal services 13:630
Corporate liberalism 18:299, 18:305
Corporate libraries 14:45, 14:49
Corporate pathos 16:181
Corporate philanthropy 22:170
Corporate postal systems 18:640–641
Corporate power and technology
agrarian political economy 1:460
Corporate restructuring 4:904
Corporate secrecy 21:377–378
Corporate social entrepreneurship 22:358
Corporate social responsibility (CSR) 4:942, 10:254, 10:257, 10:394–395, 16:78–79, 16:859, 19:82, 22:168–169, 24:475–477, 24:479
consumer policy 4:736–737
and corporate environmental performance 19:82
and greening of business 10:394–395
of in-house counsel 12:127
international development 8:61
Corporate societies 14:745–746
Corporate standards, business codes 8:22–23
Corporate structure 4:939–940
Corporate tax obligations 4:946
Corporate taxation 4:946–948
corporate income tax 4:948
economic behavior, impact on 4:946–947
multinational corporations 4:947–948
scope and history 4:946
Corporate welfare, and Corporate welfare 7:760
Corporation act (*Aktiengesetz*) 16:359
Corporation of University Promotion (CPU) 7:183
Corporations 3:29–30, 3:710, 5:769, 14:471–472, 17:144
capitalism 3:99–100, 3:103
controlling shareholders 4:941
as cultural institutions 3:24
de facto control of 4:940–941
dispersed *versus* concentrated ownership 4:943–944
financial 9:171–172, 9:175
globalization of 3:102
governing structure of 4:940
influence on 23:490–491
and law 17:397–404
publicly traded 4:940–941
as a separate legal entity from its owners 4:940
separation of ownership from control 4:941
share ownership 4:940–941
stakeholders 4:941–942
strategic investment planning 1:105
transferability of ownership 4:940

transnational 3:99
transnational corporations 10:221–223
use of term 4:939
Corporatism 12:374–375, 18:234
and policy networks 18:315–317
and welfare 25:477
Corporatist model, of vocational education 25:235
Corporatization 14:736
Corporeal property 13:107
Corporeal travel 5:412
Corporeality, and body history 2:728–729
Corpus
building 4:949–952
annotation and storage 4:951–952
data collection 4:949–951
design criteria 4:949
balanced *versus* specialized data collection 4:950
parallel *versus* comparable data collection 4:950–951
single language *versus* multilanguage data collection 4:950
spoken, written, or mixed data collection 4:950
synchronic *versus* diachronic data collection 4:950
defined 4:949
and exchange 8:477–478
lexicography 13:940–941
Corpus analysis 6:512, 6:513, 6:514
Corpus callosum 12:416, 16:737–738, 23:292–296
atrophy, and alcohol 1:512
congenital agenesis of 12:418
degree of interhemispheric connectivity 12:417
excitatory role of 12:418
hemispheric asymmetry and size of 12:417–418
of individuals with developmental dyslexia 12:417–418
interhemispheric inhibition and 12:417
Corpus Christi 23:47
Corpus Iuris Civilis of Justinian 3:669
Corpus Iuris Iustanianus 11:48
Corpus linguistics 4:949–953, 14:225–226
defined 4:949
Corpus planning, in language policy 13:353
Correa, Rafael 23:48
Correct index of residential segregation 21:412
Correct judgment 19:472
Corrected recognition scores, of recognition memory 20:37–38
Correction, data 6:126
Correction, in naturalism 16:329
Correction activity 4:771
Correction for attenuation 3:789
classical true score model 19:425–426
Correctional programs 6:682–683
Correctional psychiatry 9:349
Correctional strategies 18:940–941
Correctional treatment 20:193, 20:197–198
Correctional treatment program 11:721
Corrections, applied research (criminology and criminal justice) 5:129
Corrective feedback 13:308
Corrective justice 7:926, 12:935
Corrective rape 21:748
Correctives 14:755
Correctness 24:723
law and morality 13:480
and meaning 14:845–846
Correlated environments, influences in adulthood 6:754–755
Correlated equilibrium 5:97, 9:586–587
Correlated failure times, in multiple failure 23:780
Correlates, of subjective well-being 23:650
Correlates of personality traits 17:867
Correlates of War (COW) project 5:793, 5:836, 8:97–98, 8:98t
Correlation between parallel measurements 19:426
Correlation coefficient, mathematical psychology 14:811–812
Correlational learning 4:85–86
Correlations 17:662, 19:704
coefficients of 23:425
developmental psychology 5:354–355
ecological 6:868–869
multiple, Fisher's theories 9:223
Pearson's 17:661–662
word order 25:646
Correlogram 23:107–108
Correspondance analysis 5:6–11
Correspondence analysis (CA) 5:1–6, 5:560
active and passive variables 5:6, 5:9
Bourdieu's use of 5:8–9
contributions to inertia 5:3–4
extensions 5:4–5
canonical correspondence analysis 5:5
concatenated tables 5:4
fuzzy coding 5:5
multidimensional scaling on distance matrix 5:5
multiple correspondence analysis 5:4
ratings, rankings, and paired comparisons 5:4–5
subset correspondence analysis 5:5
future developments and related techniques 5:9–10
indicator matrix 5:7, 5:7t
matrix of input data 5:6–7, 5:7t
methodological background 5:8
principal investors and early work 5:8
principles of 5:6–8
relations between educational level and occupation groups 5:6
space of abilities 5:6
'symmetric map' for visualizing the solution 5:6–7
Correspondence matrix 5:1
Correspondence metaphor 25:177
Correspondence models 16:121–124
bivariate 16:121–123
canonical models 16:122–123
groups of categorical variables 16:123
inferential methods 16:121–122
multiple correspondence analysis 16:123
Correspondence studies 9:130
Correspondence theory of truth 5:272
Corridor location problem 23:174t
Corruption 3:19, 13:433
achieving control in contemporary times 5:18–19
definition 5:12–14
development of control of 5:16–18
etymology 5:12–14
evolution of control by global regions 5:18f
formula 5:18–19
control of 5:19
good governance as absence of 19:691
in Latin-America 13:411
and political trust 24:660–661
in postsocialist societies 18:699
public policy ethics 19:583
of youth
see also Youth
Corruptionem 5:12
Cortical areas 5:21–29, 16:586–587, 25:184–185, 25:187, 25:189
areas involved in arm actions 5:26–27
superior parieto-occipital cortex, precuneus, and mIPS 5:26–27
areas involved in eye movements 5:27
frontal eye fields 5:27
parietal eye fields 5:27
supplementary eye fields 5:27
areas involved in hand actions 5:25–26
anterior intraparietal sulcus 5:26
ventral premotor cortex 5:25–26
areas involved in head movements 5:27
ventral intraparietal sulcus 5:27
evolution 5:22–23
fMRI 5:23, 5:26–28
general organizational principles 5:27–28
grasping 5:23, 5:25–28
hierarchies of movement control 5:21–22, 5:22f
human cerebral cortex regions 5:24f
human cortical areas 5:24
parietal cortex 5:21–24, 5:27–28
primary motor cortex 5:24
reaching 5:25–28
secondary motor cortex 5:24–25
motor regions along medial wall of frontal lobe 5:25
premotor cortex 5:25
studies of connectivity 5:27
study techniques 5:23–24
Cortical columns, organization of 17:128–129
Cortical connectivity 5:31–33
Cortical development 5:30–34
auditory cortex 5:32–33
formation of areas 5:31
formation of layers 5:30–31
formation of networks 5:31–32
ocular dominance and orientation columns in visual cortex 5:31–32
patterned electrical activity and development of cortical networks 5:32–33

Cortical development (*continued*)
rewiring 5:32–33, 5:33f
Cortical dysfunction, in ADHD 2:157
Cortical field-like interference patterns, theory of 13:392
Cortical hierarchy 25:157
Cortical lamination 3:338–340
Cortical language circuit 13:403f
Cortical lobes, cerebral cortex 3:334–338
Cortical maps, and phantom limbs 18:3
Cortical networks, and inferences 19:397–398
Cortical neurocircuitry dysfunction, in ADHD 2:157–158
Cortical neurons, stimulus selectivity of 17:128
Cortical orientation selectivity 25:159–160
Cortical patterning, molecular pathways setting stage for 2:124–126, 2:125f–126f
Cortical region lateral magnocellular nucleus of the anterior neostriatum (lMAN) 2:692
Cortical semantic network, neuroanatomical distribution of 21:576f
Cortical specification 2:118–120, 2:119f–120f
Cortical spreading depression (CSD) 24:566
Cortico-cerebellar loops 3:330
Cortico-cortical communication 12:416
Cortico-hippocampal system 10:889–890
Cortico-parahippocampal–hippocampal circuits 10:888–889
Corticosteroids, and HPA axis 11:476–477, 11:479–480
Corticosterone 7:606, 16:490–491
Cortico-thalamo-cortical projections 3:343f
Corticotropin releasing factor (CRF) 16:651
Corticotropin-releasing hormone (CRH) 7:606, 10:463, 10:466, 11:476–477, 23:577
and anxiety 1:804, 1:806–809
-enhanced startle 8:862
and sleep 22:83–85
Cortisol 2:138, 7:606, 17:134, 17:138, 23:557–560, 23:578
measurement issues 23:578
and sleep 22:80, 22:82–85
and stress 23:571–572
COS *see* Charity Organization Societies (COS)
Cosa Nostra 14:422
Cosby Show 15:57
Cosche 14:422
Coser, Lewis A. 5:35–40, 7:380, 22:723
American socialism 5:38
Cohn on 5:35
conflict 5:38
dissent and intellectual success 5:37–38
on functionalism 9:539
functions and academic success 5:37
knowledge 5:38–40
legacy 5:40
Merton on 5:36–39
Parsons on 5:36–38
political sociology 5:37–38, 5:40
refugee 5:36
in Paris 5:35–36
Simmel on 5:36–37, 5:39
sociological theory 5:37, 5:39
as sociologist and socialist 5:39–40
structural functionalism 5:36–37
Cosgrove, D., works, *The Iconography of Landscape* 11:19–20
Cosmetic surgery, as erotic capital 8:2–3
Cosmides, Leda 8:463
Cosmogony 7:375–376
Cosmographer 1:734
Cosmological theories 20:288
Cosmology 13:70–72, 13:227
geomancy, Asian studies 6:786–787
play and 18:216
Cosmopolitan city 3:665
Cosmopolitan conceptions of racial identity 19:810
Cosmopolitan democracy 6:94
Cosmopolitan egalitarianism 7:312–313
Cosmopolitan employee 5:334
Cosmopolitan importance of philosophy 13:4
Cosmopolitan internationalism 4:339
Cosmopolitan regionalisms 8:278
Cosmopolitanism 3:664–665, 5:41–46, 12:237–241, 22:890–891, 24:588, 24:591, 12:236t
anthropology, Kant 5:42–43
applications of 5:45–46
Beck and 2:396–399
Benhabib and 2:519–520, 2:522–524
in China 2:67
and collective change 5:41–42
cosmopolitics 5:45
creolization 5:43–44
cultural differences 5:43, 5:45
Durkheim and 5:41–45
and multiculturalism 16:24–26
versus nationalism 5:43–44
new, debates around 5:44–45
and postcolonial geography 18:645–646
recent arguments, formalization of 5:44
Schmitt and 21:31–33
Simmel and 5:43–45
urbanization 5:45
Weber and 5:45
Weltanschauung 5:43
Weltkenntnis 5:42–43, 5:46
and world republic 25:740–742
Cosmopolitanization 17:652
Cosmopolites (community) 4:332
Cosmopolitics 5:45
COSSA *see* Consortium of Social Science Associations (COSSA)
COST *see* Cooperation in Science and Technology (COST)
Cost analysis 5:56
Cost belief, in vocational psychology 25:241
Cost benefit 8:345, 8:348–349
Cost containment practices 10:595–596
in hospitals 11:208, 11:211–213
Cost control, health care markets 10:598–601
Cost functions, and DoF problem 15:959–960
Cost internalization, torts in law and 24:432
Cost of living index (COLI) 5:753–754
Cost savings, sacrificing quality of managed care for 14:463–464
Cost tables 14:468
Cost value 1:37
Costa, Paul T.
work on five-factor model of personality 9:231
Costa Rica 22:624
sociocultural aspects 1:621–625
Costate variables 17:253
Cost–benefit analysis (CBA) 5:47–54, 5:56, 6:855, 7:80, 18:277–278, 19:297, 19:459, 24:130, 24:132
accounting for uncertainty 5:53
allocative efficiency 5:47–48, 5:50, 5:52–53
alternative projects and counterfactual 5:48
benefit analysis 5:56
benefit–cost analysis using nonmonetary units 5:58
computing net benefits and social surplus 5:50–51
contingent valuation 5:51–53
cost analysis 5:56
cost-effectiveness analysis (CEA) 5:56
cost-utility analysis (CUA) 5:57
of decriminalization of drugs 6:676–678
definition 5:55–57
discounting future costs and benefits 5:52–53
distributional 5:53–54
effectiveness 5:55–56
efficacy 5:55–56
efficiency 5:55–56
emerging directions in theory and research 5:58
utility assessment 5:58
in evaluation research 5:55–62
evidence-based practices 5:58
existence value 5:49
focus 5:57–58
foundations 5:47–48
goals 5:55–56
hedonic regression 5:51
identification and categorization of impacts 5:49
active use impacts 5:49
passive use impacts 5:49
issues of standing 5:48
leading advocates 5:57
of legal entitlements 5:253
LMST method 5:52
maximum expected net benefits rule 5:53
methodological issues and problems 5:59–61
biases in comprehensive measurement and monetization 5:59
itemization in reporting to allow generalization by readers 5:59

maximizing cost-benefit, cost-effectiveness, and cost-utility 5:60
monetization of lifetime earnings can exacerbate discrimination 5:59–60
monetization of volunteered and donated resources 5:59
relationships over range of values 5:60
social return on investment (SROI) 5:60–61
statistical models capture uncertainty, but cost-inclusive indices seldom do 5:60
opportunity cost 5:47, 5:50, 5:52–53
option value 5:49, 5:53
poorly-functioning markets, prediction and valuation of policies in 5:51–52
prediction of impacts 5:51
valuation of benefits using continent valuation methods 5:51–52
valuation of benefits using revealed preference methods 5:51
valuation of costs 5:52
program evaluation 5:55
public sector 5:50–51
risk 5:51, 5:53
of risk regulation 13:506
schools of thought 5:57
sensitivity analysis 5:53–54
shadow price 5:51–53
skills training 22:20
social discount rate 5:52
well-functioning markets, prediction and valuation of policies in 5:50
measuring benefits 5:50
measuring costs 5:50
willingness to pay 5:47–48, 5:50–53
Cost–benefit approach, risk 20:706–707
Cost-effective utilization of resources, in combined pharmacotherapy and psychotherapy 19:496
Cost-effectiveness 19:118
community-based health interventions 10:651
controlled trials 19:389
Cost-effectiveness acceptability curves (CEACs), economics of psychiatric care 19:297
Cost-effectiveness analysis (CEA)
benefit analysis 5:56
benefit–cost analysis using nonmonetary units 5:58
cost analysis 5:56
cost-benefit analysis 5:56
cost-utility analysis (CUA) 5:57
definition 5:55–57
effectiveness 5:55–56
efficacy 5:55–56
efficiency 5:55–56
emerging directions in theory and research 5:58
utility assessment 5:58
in evaluation research 5:55–62
evidence-based practices 5:58
focus 5:57–58
goals 5:55–56
leading advocates 5:57
methodological issues and problems 5:59–61
biases in comprehensive measurement and monetization 5:59
itemization in reporting to allow generalization by readers 5:59
maximizing cost-benefit, cost-effectiveness, and cost-utility 5:60
monetization of lifetime earnings can exacerbate discrimination 5:59–60
monetization of volunteered and donated resources 5:59
relationships over range of values 5:60
social return on investment (SROI) 5:60–61
statistical models capture uncertainty, but cost-inclusive indices seldom do 5:60
program evaluation 5:55
schools of thought 5:57
Cost-effectiveness evaluations, for quality of life 10:660
Cost-inclusive evaluation 5:56–60
Costly punishment 12:721
Cost-minimization analysis (CMA) 19:459
Cost–reward model of empathy 19:217–218
Costs 5:633
child 3:498–501
of children 8:822–823
in competition/cooperation 4:857
congestion 24:630
of criminalization of drugs 6:677–678
discovery
in England 6:520
in United States 6:518
expectancy-value-cost model of motivation 8:503–509
of fertility regulation 9:94–95
health care 15:102
healthcare, and personalized medicine 17:927
of identity theft 5:634–635
of illness 25:15
of interviews 5:853
pharmacoeconomics 19:458
of psychiatric care 19:296–297
of psychopharmacotherapy 19:457–461
of reproduction, and life history theory (LHT) 9:40–41
of sanctions 20:890–891
of voting 25:309
Costs of living socially 18:899–900
Costs of representation 14:267
Costs–benefits
political 12:387
Costume, carnival 3:145–148
Cost-utility analysis (CUA) 5:57, 10:617, 19:459
Co-surveillance 23:737
COT *see* Computational organization theory (COT)
Co-teaching 21:115–116, 21:116t
Cottage-craftsmen 5:108
CouchSurfing 8:487
Coulmas, Florian, works, *The Handbook of Sociolinguistics* 7:590
Council for Basic Education (CBE) 1:232
Council for Cultural Affairs (CCA) 5:456–457
Council for Economic Education (CEE) (United States) 24:83–84
Council for Environmental Education (UK) 7:733
Council for Exceptional Children (CEC) 7:160, 23:213
Council for International Organizations of Medical Sciences (CIOMS) 8:78–79, 8:81, 20:512, 20:514, 20:525–527, 21:316
Council for Mutual Economic Assistance (Comecon) 4:305
Council for the Accreditation of Educator Preparation (CAEP) 24:76–77
Council for the Accreditation of Educator Programs 23:213
Council for the Development of Social Science Research in Africa (CODESRIA) 22:604, 22:607, 22:675
Council of Area Studies Associations 1:947
Council of Bars and Law Societies of Europe (CCBE) 13:621
Declaration of Perugia 13:622
Council of Bars and Law Societies of the European Community 3:672
Council of Economic Advisors 22:698–699
Council of Europe 12:957
Council of Legal Education 13:734–735
Council of Title VI National Resource Center Directors (CNRC) 1:947
Council of Trent 3:234, 20:286
Council on Legal Education for Professional Responsibility 19:559
Council on Social Work Education (CSWE) 22:771
Counsel, in-house 12:126
Counseling
nondirective 9:995
offender rehabilitation 20:195
Counseling
childhood cancer
strategies and interventions 3:460–462
clinical psychology 24:525–526
ethnic and cultural differences in 5:63–68
cultural pluralism and tyranny of the shoulds 5:64
multicultural psychology 5:65–66
racial identity theory 5:66
racial–cultural identity development 5:65–66
psychology 5:69–72
career development and intervention 5:71
distinctiveness and diversity 5:69–70
fundamental features of field 5:70
healthy functioning and human effectiveness 5:71
methodological considerations 5:70

Counseling (*continued*)
multicultural counseling and research 5:71
personalities 5:69–71
roles and functions of counseling psychologists 5:70
scientific inquiry 5:70–71
theory and research 5:70–71
training in practice and science 5:71
unifying themes 5:69–70
Rogers on 20:737–742
Count data, nonparametric regression for 16:874–875
Count quantifiers 19:707
Counterargumentation 2:170
Counterattitudinal advocacy 22:352
Counterconcepts 4:337
Counterconditioning 18:41–42
Countercultural communities, and counterculture 5:85
Counterculture 5:73–86, 23:640
and youth culture 25:815–816
anti-Vietnam war 5:83
Beats 5:80
broader picture 5:84–85
civil rights 5:84–85
countercultural communities 5:85
cybernetics 5:646
emergence of youth cultures 5:80–81
feminism 5:85
festivals 5:83–84
gay rights movement 5:85
hippies 5:81–83
in history 5:77
legacy of 5:75–77
patriarchy 5:85
in postmovement period 5:74–75
protest 5:80–81, 5:84–85
psychedelia 5:83
psychedelic rock 5:82–83
1960s 5:81
women's liberation movement 5:85
Counterdiscourses
on sustainable development 23:813–815
nature-related counterdiscourses 23:813–814
Counterfactual conditionals 3:251, 5:87–88
Counterfactual interpretation of the mediation model 15:67
Counterfactual model 8:528
modification of 8:529
Counterfactual reasoning 14:307–309, 16:925
philosophical aspects
analysis of 5:95–96
background 5:95
in decision theory 5:97
in game theory 5:97
history 5:95
probabilities of 5:96–97
qualitative 5:87–94
causation 5:93
and evidence 5:91
Holland's model 5:91–92
interventionist counterfactuals 5:92–93
metalinguistic/support view 5:88–89
possible worlds view 5:89–91
powers 5:93
Rubin's model 5:91–92
quantitative 5:95–98
Counterfactual thinking, and grit 10:399–400
Counterfactuals 3:267, 3:269–270, 3:272–273, 22:292
account of dispositions 3:268
analyses 3:269–270
in causal analysis 3:247–248
in social science research
and causation 3:251
definition 3:251
empty counterfactuals 3:253–254
Neyman–Rubin model 3:252–253
prospective causal studies 3:251–252
Counterfinality 22:196–197, 24:746
Countering violent extremism (CVE) programs
in United States 16:250–251
Counterinsurgencies (COIN) 5:99–102, 11:392–396, 12:221, 13:534
anthropology 5:99–102
anti-communism 5:100–101
in Asia 5:100–101
'best practices' in 12:226
demographic studies 5:99, 5:101
direct approach 12:223
espionage 5:99, 5:101
in indirect approach 12:223–224
in Iraq 12:224
in Latin America 5:101–102
military anthropology 15:486
NATO intervention 12:226
primary challenges of effective 12:224
Project Camelot 5:101
psychology 5:99–102
social control 5:99
theory and practice 12:223–224
Third World 5:99
twentieth century 5:99, 5:101
United States 5:99–101
use of a 'continuum of force' by 12:224
Vietnam centrality 5:101
war 25:377
Counterinsurgency wars
in Latin America 25:122–123
media coverage 25:123
Counter-Intelligence Program (COINTELPRO) 11:395
Counter-Jihad movement 25:798–799
Countermajoritarian difficulty 12:897
Countermemory 4:184
Countermovements against medicalization 22:976–977
Counternarcissistic apparatus of power 1:445
Counter-objection 4:869
Counterproductive work behavior (CWB) 1:422–423, 7:447, 10:411, 12:835, 25:723–725
personnel selection 17:937
Counterproductive Work Behavior Checklist (CWB-C) 25:723
Counterproductivity, medicalization 11:585
Counterpublic 19:616–617
Counter-reformation 3:234, 20:98
and sainthood 15:578
Counter-Reformation Catholic Church 18:916
Counter-reformation Catholicism 3:240
Counterschools 21:39–41
Counterterrorism (CT) 13:530, 13:532, 24:237, 24:239–242
policing 18:264
training, in United States 16:250–251
Countertransference 19:340–341, 19:350–351, 24:547–548
Counterurbanization 2:307, 20:796
in Africa 24:970–971
Counting (quantification) 19:700
Counting process representation, for event history analysis 8:313
Countries
characteristics, and academic achievement 22:927–928
income, and academic achievement 1:43–44
influence on health outcomes 22:279
-level moral inclusiveness 15:785–786
-level patenting comparisons 11:835–837
role in nostalgia 17:35–36
Country/area studies 11:784–785
Country-club culture 20:396
Country-level studies 5:326
County Asylums Act 1845 19:587
Coup de grace 18:246–247
Coup d'état 5:837–840, 20:628
Couple therapy 2:428, 8:748, 19:515
for depression 19:376
Coupled slacks 16:542
Couples
dual-career
earning arrangements among 14:608
work schedules among 14:608–609
joint consumption of 8:119
partnerships in Western societies 17:545–552
psychological distress of cancer in 3:75
interventions 3:76
Coupling, in high reliability organizations (HRO) *see* high reliability organizations (HRO
Coupling constraints 15:638
Cour de Cassation 3:669
The Courage to Heal 8:710
Cournot competition 14:551–552
Cournot Oligopoly, example of Nash equilibria 9:577
Cournot's analysis of imperfect economic competition 15:304
Course Management System (CMS) 12:67
Court communities, in sentencing 21:627–628
Court community perspective 21:627
Court decisions, human resource management 11:351–352

Court judgments, university hazing, in United States 10:559
Court of Appeals for the Seventh Circuit 13:744–745
Court of Justice 8:265–266
Court on Human and Peoples' Rights 11:800
Court-appointed attorney 8:585
Court-awarded fees 12:922
Courtesan 17:489
Courtly diversions
  entertainment 7:670–671
Courts 13:604–606, 18:386–387, 18:389–390, 19:555–558, 19:560
  adjudication 5:103–107
  and administrative law 1:111–112
  in authoritarian states 13:427–434
  business and organization of 5:105–106
  and constitutions 4:712–714
  domestic violence 6:627–628
  equality principle 7:894–898, 7:900–901
  and executives 12:888
  law beyond the courts *see* Law beyond the courts
  law within the courts 13:598–599
  and legislatures 12:887–888
  and organizations 17:399–400
  public evaluation of 13:748
  role 13:609
  role in dispute transformation 6:564–565
Courtship 14:367
Coutin, S.B., on immigrants in North America 17:24
Covariance, spatial association as 23:102–103
Covariance adjustment 17:108–109
Covariance estimation, linear hypothesis 14:148–149
Covariance matrix
  latent variables 13:395
  for the random effects 14:167
  of test score variables 3:788
Covariance stationary process, spectral representation of 24:332
Covariance structure analysis 8:693, 8:695
Covariance structure modeling 8:686–688
Covariate adjustment 19:144
Covariate decompositions, generalized linear mixed models 9:851
Covariates 11:751
Covariation 21:268
Covenant at Sinai 12:870
Covenants on human rights 11:356–357
Coverage
  data 6:140–141
  health 10:640–644
Coverage errors 6:126–127
  census 3:293–294
Coverage of mental disorders 15:182
Coverbs 3:814
Covering law model 4:24, 10:773
Covert actions, simulation theory of 15:961–962
Covert behavioral problems 2:470
Covert conditioning *see* Covert sensitization
Covert ethnography 8:182
Covert language policy 13:355
Covert movement 16:6–7
Covert population experiments 9:130
Covert reading 6:733
Covert sensitization 3:767
Coverture, and feminist legal theory 8:922–923
COW *see* Correlates of War (COW) project
Cowan, Jane, works, *Culture and Rights* 11:363
Cowitness misinformation 8:645
Cowles Commission 12:207
Cox, B., on women's participation in sports 9:658
Cox process 23:195–196
Cox regression model 8:301, 14:56, 23:774, 23:777–779
  asymptotic distribution theory 23:779
  development 23:777–778
  in event history analysis 8:315–316
  special cases and generalizations 23:778–779
  survivor function estimation 23:779
Cox–Aalen model 8:303
Coy, M., works, *Apprenticeship* 5:110
  see also Contextual performance (CP)
CPA *see* Care Programme Approach (CPA)
CPAI-2 *see* Cross-Cultural Personality Assessment Inventory (CPAI-2)
CPAR *see* Critical participatory action research (CPAR)
CPC *see* Central Product Classification (CPC)
CPC Ver.2 (Central Product Classification) 5:707
CPD *see* Continuing professional development (CPD)
CPE *see* Continuing professional education (CPE)
CPG *see* Clinical practice guidelines (CPG)
CpG islands (CGIs) 24:683–685
CpG methylation 7:839
CPIs *see* Consumer price indices (CPIs)
CPM *see* Clique percolation method (CPM)
CPNs *see* Community psychiatric nurses (CPNs)
CPRs *see* Common pool resources (CPRs)
CPS *see* Clinical Problem Solving Test (CPS); Creative Problem Solving (CPS); Crown Prosecution Service (CPS); Current Population Survey (CPS)
CPs *see* Conservative Protestants (CPs); Content providers (CPs)
CPSR *see* Computer Scientists for Social Responsibility (CPSR)
CPSU *see* Communist Party of the Soviet Union (CPSU)
CPT *see* Cognitive processing therapy (CPT); Critical planning theory (CPT)
CPTED *see* Crime prevention through environmental design (CPTED)
CPTSD *see* Complex post-traumatic stress disorder (CPTSD)
CPU *see* Corporation of University Promotion (CPU)
CQ *see* Cultural intelligence (CQ)
CQS *see* Cultural Intelligence Scale (CQS)
CR *see* Cajal–Retzius (CR) cells; Cardiac rehabilitation (CR); Cognitive restructuring (CR); Conditioned response (CR)
CRA *see* Civil Rights Act (CRA); Community Reinvestment Act 1977 (CRA)
Crack cocaine 5:156
Crack epidemic 10:458
Crack-cocaine market, rise and fall of 5:199
Crack-cocaine thesis 5:199
Cracking, apportionment 7:343
Craft production 5:108–112, 12:329–331
  anthropology 5:109–110
    learning and apprenticeship 5:110
    Marxist studies 5:109
    material culture and folklore 5:109
    rebirth of material culture studies 5:109–110
    social, ethnographies in 5:109
  definition 5:108
  participant observation 5:108
  specialization, the division of labor, and commodization 5:108
  trajectories in global economy 5:110–112
    in developing societies 5:110–111
    in industrialized societies 5:111–112
Craft specialization
  in archaic states 23:373
  and chiefdoms 3:379
Craft workers, apprenticeship 1:867
Crafting 12:332
Crafts 2:15, 2:18, 2:46–47
  artisanal 2:47
  in European Middle Ages 2:46–47
  in Gallic and Germanic mythology 2:46–47
Craftwork, apprenticeship 1:865–869
Craig, Alexander, works, *Ionia: Land of Wise Men and Fair Women* 14:255
Craik, Kenneth H. 20:483
Cram schools in Japan 12:795–796
Cramer–Rao lower bound 21:976–977
Crane, D. 25:83–84
Cranial nerve nuclei, tension control in extraocular muscles 8:628–629
Cranston, Maurice, works, *Glossary of Political Terms* 7:591
Crater Mountain Wildlife Management Area 18:353
Cravath Model of law firms 13:578–579
Cravath system 13:630
Craven, Emily
  on GASA 21:745
  on sexual minority rights in South Africa 21:746
Craving 1:98–99
  brain stimulation reward 2:843
Crawfurd, John
  works
    Descriptive Dictionary of the Indian Islands and Adjacent Countries 11:84
    History of the Indian Archipelago 11:84

Crawling infants 22:558
motor development 15:975
CRC *see* Convention on the Rights of the Child (CRC); Criminology Research Council (CRC)
Cream skimming
economic regulation 20:167
health insurance 10:643
private health insurance 10:615–616
Creando, Mujeres 19:748–749
Creation
native theory 24:454
opposition of 24:453
Creationism 12:334–335
Creative behavior, Skinner on 22:26–27
Creative capitalism 24:821
Creative class 10:50, 12:90, 18:666
Creative Commons Attribution license (CC-BY) 12:867
Creative destruction 22:357–358
in entrepreneurship 7:676, 7:679, 7:686–687
and university press 19:629
Creative economy, existence and importance of 11:858
Creative freedom, and Chinese cultural policies 5:600–601
Creative industries 2:5, 5:727–728, 5:730–731
Chinese 5:597–598
and cultural policies 5:461–462
policy, in East Asia 5:456–457
Creative living 17:737
Creative management, of organization 22:804–805
Creative participation gap 22:921
Creative performances 14:204
Creative personality 17:770–773
Creative places, and social networks 5:469–470
Creative Problem Solving (CPS) 5:114–115
Creative Revolution 1:196–197
Creative Taiwan plan 5:456–457
Creative teaching methods, for economic education 24:84–85
Creative thinking, and mental imagery 15:223–224
Creative workers 5:609
Creativity 3:12, 17:232–233, 17:710, 17:770–773
and aesthetic education 1:231
cognitive processes linking openness/intellect with 17:771–772
decreased latent inhibition 17:771
divergent thinking 17:771
implicit learning 17:771
working memory 17:771–772
domain-specific 17:772–773
and gifted education 7:158, 7:160–164
intellectuals 12:278–280
and intelligence 12:314, 12:317, 12:320–321, 17:770–772
and learning 13:710–711
Lévi-Strauss on 13:908
and media 15:7
and organizational behavior 17:700–701
across situations 17:700
fostering 17:700–701
and organizational group diversity 10:410
and pragmatism 18:809–810
pretend play and 18:874
psychology 17:770, 17:772
and psychopathology 17:772
research history 17:770–771
social networks 5:467–468
and tradition 24:505–506
traits 17:770–773
Creativity complex 5:113
Creativity training 5:113–116
cognitive components 5:113–115
complex 5:113
as enhancement of potential 5:113
environmental and interpersonal factors 5:115–116
incubation 5:115–116
information processing 5:115
intrinsic motivation 5:115–116
motivation and attitude 5:115
overjustification effect 5:115
tactics 5:114t
CREB *see* cAMP response element-binding protein (CREB)
CRED *see* Centre for Research on the Epidemiology of Disasters (CRED)
Credentialing, and quality of care 14:463–464
Credentialism
and employment 7:127–128, 7:130–131
in medical profession 15:94
Credentials, of teachers in United States 24:76, 24:79
Credibility 10:302, 21:266
of Internet information 16:793–794
and reliability 17:287
in science 24:673–678
scientific knowledge 24:673–678
and social influence 22:351–352
plain folks appeal 22:352
source credibility 22:351
visualization 25:223–225
Credibility enhancing displays (CREDS) 20:392
Credible commitments 4:864
and democratic transition 23:370
and growth 23:370
Credit
and capitalism 3:106–108
and marriage 8:829
Credit card fraud 5:633–634
Credit default swap (CDS) 3:32, 3:582–583
Credit markets 1:478–479
job displacement insurance 12:829
Creditors 4:939–940
corporate governance 4:933
see also Bankruptcy
CREDOS (Collaboration of Researchers for the Effective Development of Offender Supervision) 19:47
CREDS *see* Credibility enhancing displays (CREDS)
Cree *see* Algonquian languages
Creese, Walter 24:862–863
CREFAL 7:183
Crenshaw, Kimberlé Williams 13:498
Creole 5:117–118, 5:120, 11:820, 13:307, 22:952
and Pidgin 18:133
extension of the term creole 18:135
lexifiers and 18:137–138
meaning 18:133–136
in settlement colonies 18:134
typological or genetic class of 18:138
and serial verb constructions 21:646–647
Creole continuum 18:134, 18:137
Creole genesis 18:138–142
Creole language 18:133–145
Creole Life-Cycle, historical perspective 18:136–137
Creolization 5:117–121, 14:594–595, 18:138, 25:195
and anthropology 5:118–119
in Caribbean 1:641
colonialism 5:117–118, 5:120
and cosmopolitanism 5:43–44
and diffusion 6:404
and ecumene 7:111
essentialism 5:119–120
and fast social change 5:120
hegemony 5:117–118, 5:120
and hybridity 5:118–119, 11:449–450
origin of concept 5:117–118
pidgin language 5:117–118
pidginization 5:118–119
postcolonialism 5:120
syncretism 5:118–119
and syncretism 23:883–884
in West Africa 1:302
and world culture 10:228
World System and globalization 5:120
CRESALC *see* Regional Center for Higher Education in Latin American and the Caribbean (CRESALC)
Cressey, Paul, works, *The Taxi-Dance Hall* 7:561, 7:564
Creutzfeldt-Jakob disease 18:177–178
Crèvecoeur, J Hector St. John de, works, *Letters from an American Farmer* 1:617
Crew resource management (CRM) 11:302
CRF *see* Corticotropin releasing factor (CRF)
CRH *see* Corticotropin-releasing hormone (CRH)
Crichton, Alexander 16:681, 18:178, 16:681f
Crick, Bernard, works, *The American Science of Politics* 18:468
Crick, Francis 14:243
Crime control 4:826, 4:834, 13:808–809
empirical support for 4:834–836
issues related to 5:192
sources of control 4:834–836
formal *versus* informal control 4:835–836
internal *versus* external control 4:834–835
varieties of 4:836–837

macro- *versus* micro-control theories 4:836–837
static *versus* dynamic control theories 4:836–837
Crime data
international 5:714
official data 5:712–713
police data 5:713–714
self-report studies 5:715–716
victimization surveys 5:714–715
Crime drop 5:153, 5:166
Crime elimination program 24:120
Crime events 20:772–774, 20:776–777
Crime hot spots 6:263–264
Crime mapping 5:145, 5:149
Crime monitoring/surveillance, applied research in 5:127–128
Crime patterns, offender 20:776
Crime policy
'Stand Your Ground' laws 5:177
in United States 5:176–181
AIG insurance company 5:178
capital punishment 5:180–181
death penalty 5:180–181
due process 5:176, 5:179
government branches that design and implement crime policy 5:180
law enforcement 5:177, 5:179–180
police 5:176–177, 5:180
policy defining crime 5:177–178
prosecution 5:176–178, 5:180–181
suppressing crime and protecting liberty 5:178–180
victimless crime 5:177–178
War on Drugs 5:178
white collar crime 5:176
Crime prevention
definition of 18:877
deterrence, forms of 18:877
of drugs and alcohol use 18:880–881
firearms and crime, relationship between 18:880
medical model 18:878
police and 18:878–879
principles 18:879
primary prevention programs 18:878
routine activity theory and situational 18:877
sentencing and punishment 18:879–880
situational 18:877–878
tertiary prevention programs 18:878
theory of 18:877–878
through environmental design (CPTED) 18:877–878
Crime prevention through environmental design (CPTED) 20:776
Crime protection effect, in United States 11:630–631
Crime rates 4:326
and guns 10:458, 23:142–143
Crime reduction, and social capital 22:178
Crime Survey for England and Wales 5:719
Crime surveys 5:848
Crime victimization, risks of 19:779
Crime victims, social justice for 22:388
Crime wave 15:23–28
Crime–gender gap 5:132
approaches to study 5:132–133
individual-level data 5:132–133
sociopsychological explanations 5:132–133
causal factors 5:134
connections between abuse during childhood, economic hardships, and criminal careers 5:134
explanations 5:135
National Crime Victimization Survey (NCVS) data over time 5:133
qualitative research on women's offending 5:133–134
research on narrowing 5:133
directions for future research 5:135
self-report surveys 5:134–135
sex-specific measures of economic marginalization on arrests 5:133
Uniform Crime Report (UCR) violent crime arrest rate 5:133
Crime-precipitated homicides 10:458
Crimes 5:151–157, 10:6–7
Aboriginals 5:151, 5:155
and age 1:310–317
and applied criminology 1:832–833
applied research on 5:127–131
by Asian and Pacific Island ancestry 5:185
biosocial approaches to 2:670–675
black 5:183–184
category 10:12
causes of 5:233
changes in science of 5:235–236
and collective efficacy 4:169–173
and communities types 4:325–326
immigrant communities 4:326
public housing communities 4:326
rural communities 4:325–326
correlates of 5:124–125
and criminal justice statistics 5:717
Bureau of Justice Statistics (BJS) 5:718
National Incident-Based Reporting System (NIBRS) 5:718
Sourcebook of Criminal Justice Statistics 5:718
Uniform Crime Reports (UCR) 5:717–718
definition 5:164–165
in legal discourse 5:164
definitions 22:411
desistance and persistence 5:137–144
deterrence, sanction perceptions of 6:250–254
deterrence theory 6:255–259
differential association 22:410
differential reinforcement in 22:411–412
discrimination and disparity 5:155–156
Canada 5:155
possible explanations for overrepresentations 5:155–156
United States 5:155
distribution of 5:166–169
and economics 7:80–86
environmental design and 20:775–776
ethnicity/race and 5:182
patterns 5:182–185
racial characteristics as causal factors 5:182
fear of 8:854–859
functional explanation 9:537
and gender 5:132–136
geography of 5:145–150
Gottfredson and Hirschi's general theory 21:481–482
and guns 10:458
Hispanic 5:184
and Honesty-Humility 11:170–171
honor 11:183–184
honor killing 11:185–191
against humanity 10:6, 11:369
genocide 10:18–20
imitation in 22:410–411
and immigration 5:152–155
contemporary empirical findings 5:153–154
generational differences 5:154–155
historical background 5:152–153
immigration–crime nexus 5:151–153
incidence of 5:182
interdisciplinary explanations 5:234–235
internationalization of 5:236
intersectionality and 12:663–669
knowledge about 5:164
levels of 15:591
life-course theories of 14:60
against mankind 10:6
measurement of 5:165–166, 5:235, 23:143–144
multidisciplinary approaches 5:234
at nodes, paths, edges 20:774–775
organized crime 5:170–175
patterns of 5:167
and personality 17:774–779
against persons 5:190
place research and 20:775
prevention of 18:877–882
and punishment 6:704
quantitative methods for 19:733–735
and race 19:837–844
and race/ethnicity 5:182–188
religion and 20:265
reported to police 5:190
reporting 5:190
and routine activities 20:772–773
minimal elements 20:772–773
in rural context 5:158
and self-conscious emotions 21:476–477
and self-control theory 12:462, 12:464, 21:481–485
self-esteem 21:524
self-report data 5:718
sensational 15:23
and sentencing 21:627
and SLT 22:409, 22:413
and social class 5:122–126
social control of 4:825–826, 22:235–241
social interaction and networks, impact of changing 5:164
and social learning theory 22:409

Crimes (*continued*)
sociological approaches 5:234
sociology of 5:189–193
and spatial dynamics 23:142–147
and strain theories 23:495–500
strain theories and *see* general strain theory (GST)
studies 5:151–152
transnational organized crime 24:596–601
trends and debates 5:194–202
age structure 5:196–197
age-specific homicide offending trends 5:194–196
decline 5:194–196
economic growth 5:199–200
explanations 5:196
gun control 5:196, 5:198–201
immigration 5:197
incarceration 5:197–198
legalized abortion 5:200
police force and policing strategies 5:198–199
race-specific homicide offending trends 5:196
rise and fall of crack-cocaine market 5:199
understanding 5:233–235
in United States 11:628–634
victimization surveys 5:718
victimless 25:75–79
whitecollar 5:203–207
Crimes, geography of
applied aspects of 5:149
approaches to 5:147–149
criminological theories 5:146
environmental cues 5:147–148
future directions 5:149
information bases 5:146
origins 5:145–146
recognition of patterns 5:145
subcultural theory 5:146
theoretical bases 5:146
Crimes of expediency 19:162–163
Criminal behaviors 5:233–235, 6:680–681, 13:571, 20:334
and forensic psychiatry 9:346
strength of 6:681
Criminal careers 1:310–312
Criminal changes 13:499
Criminal contempt of court 15:593
Criminal defamation 9:395
Criminal immigrant 5:152–153
Criminal intention 5:221–224
Criminal investigations, effect of genetics on 9:956–957
exculpation 9:957
inculpation 9:956–957
Criminal justice (CJ) 9:339–340, 9:343, 19:210, 19:214
applied criminology 1:832–838
applied research on 5:127–131
challenge to legitimacy 5:217
components of 5:215–216
courts and 5:216
debate on definition of deviance and criminality 5:216
discrimination in 6:533
domestic violence 6:627, 6:629
future perspectives 5:217
and genetics 9:971
goal of 5:215–216
incoherence 5:215
internal coherence of 5:215–216
medical therapeutic approach 5:217
in modern democracies 19:646
challenges of criminals 19:647–648
origins of 5:214–215
and police 18:252, 18:254
policies 5:200–201
practical impact 5:215
and probation 19:43–45, 19:47–48
punishment 5:216–217
and race 19:810
race–crime relationship 19:841, 19:843–844
radical perspectives 5:215
research 5:217
and social work practice 5:208–213
and antidiscriminatory practice 5:211–212
current trends 5:212
institutionalization 5:209
origins and development with offenders 5:208–209
rehabilitation 5:210
resettlement 5:210
restoration 5:211
socialization 5:210–211
typology 5:210–211
welfare and rights 5:211
what works 5:212
sociological perspectives 5:214
sociology of 5:214
study of 5:214
tribunals in Africa and Asia 5:216
and women 25:616–621
Criminal Justice Act 1948 5:128
Criminal justice orientations 13:808–809
Criminal justice research, social construction and transformation of disputes 6:561
Criminal justice system (CJS) 5:182, 5:712, 18:226–227, 18:229
and alcohol abuse 1:520
differential treatments in 5:185–187
arrest 5:185–186
case processing 5:185–186
charging decisions and plea negotiations 5:186
death penalty 5:187
incarceration 5:186
patterns of contextual effects 5:186
pretrial decisions 5:186
sentencing 5:186
social work within 9:803
Criminal law 5:219, 5:225, 6:678, 13:451
and forensic psychiatry 9:347–348
limits of 6:675–676
political trials and 18:500–501
systems 3:480–481
transnational organized crime 24:599–600
see also International criminal court
Criminal Law (China) 21:736
Criminal liability and criminal defenses 5:220
defenses negating *actus reus* 5:220–221
defenses negating *mens rea* 5:221–223
intoxication 5:222–223
mistake of fact as defense 5:222
mistake of law and color of right 5:222
exculpatory defenses that do not negate *actus reus* and *mens rea* 5:223–225
excuses 5:224–225
justifications 5:223–224
legal systems and criminal responsibility and 5:219
questionable assumptions about autonomy and criminal liability 5:219–220
Criminal mischief 15:593
Criminal networks 16:578–580
empirical test 16:579–580
model 16:578–579
policy implications 16:579
Criminal offending, and intersectionality 12:665–668
Criminal offending trajectories 5:141–142
Criminal opportunities
integrated perspectives 20:776–778
multilevel opportunity theory 20:777–778
situational crime prevention 20:777
Criminal organizations, aggressive and cynical scam 25:125
Criminal procedure 5:227–231
accusatorial procedure 5:227
adversarial and inquisitorial procedure, eclipse of 5:231
confrontation (right to) 5:227–228
consensual procedure 5:229, 5:231
cooperation agreements 5:231
customary law procedure 5:227
dossier (investigative) 5:227–228, 5:230
equality of arms 5:230
exclusionary rule 5:231
historical watersheds in 5:227–229
inquisitorial preliminary investigation, adversarialization of 5:229–230
inquisitorial trial judge, decline of 5:230–231
investigating magistrate 5:227–230
jury court (trial) 5:228–229
plea bargaining 5:231
presumption of innocence 5:228, 5:230
reasoned judgment 5:228–229
Criminal profiling, and forensic psychology 9:353–354
Criminal prohibition 6:677
Criminal prosecutions 13:530–531
avoiding 12:363
mafias 14:424
Criminal responsibility, of forensic psychologist 9:352
Criminal sanctions 5:204–205
Criminal trial verdicts 19:733

Criminal victimization, measurement in United States 5:720–726
Criminal/delinquent behavior 20:335
Criminalists, associations of 5:240
Criminality, and biosocial criminology 2:670–674
Criminalization 10:211
of children 12:961
of drugs 6:674–675
costs of 6:677–678
harms of drugs and benefits 6:677
and homeless 25:111–115
of homelessness 25:113–114
and intersectional queer theory 19:766
and sex work 19:279–283
Criminalogia Moderna 5:239
Criminals
discount rate 7:81
and guns 10:457–461
rational 7:81–82
Criminogenic effects of incarceration 18:940
Criminogenic needs, of women offenders 25:616
Criminological accounts of compliance 4:460–461
Criminological literature 13:571
Criminological stranger 5:248–249
Criminological theories 5:146, 5:205, 22:409–410
genocide 10:16–19
Criminology 5:182, 5:232, 18:254, 19:733, 20:772, 22:387
applied criminology 1:832–838
classical school of 5:232–233
contemporary issues 5:235–237
definition 5:239
development and current status of discipline 5:232
as a discipline 5:235
evidence-based policy 5:236
history of 5:238–243
Latin America 5:239–240
Scandinavia 5:240–241
United Kingdom 5:241–242
United States 5:238–239
importance of partnerships 5:236–237
international and comparative 12:462–469
comparing crime across countries 12:462–463
determinants of crime across countries 12:463–465
drug trafficking 12:465–466
firearm-related crime 12:466
homicide 12:462–467, 12:463t, 12:465t
income inequality 12:463–464, 12:466–467
macrolevel correlates of crime 12:463–464
microlevel correlates of crime 12:464–465
organized crime 12:465–467
regional characteristics of crime 12:465–467
self-control 12:462, 12:464
terrorism 12:466–467
internationalization of 5:236
link of delinquency to 6:68
school safety 21:107–111
social control in 4:824
transnational organized crime 24:596–601
violence 12:462–467
Criminology Research Act 1971 (Australia) 5:128
Criminology Research Council (CRC) 5:128
Criollos 23:46
Crise heureuse 17:420
Crises
capitalist 14:677
of psychological development 25:330, 25:332–333
in social psychology 22:222
The Crisis 6:691
Crisis intervention teams (CITs) 4:374
Crisis of trust 2:607, 21:257
Crisis states 3:715
Crisis theory 22:816
Crisis/exploration 7:938
CRISPR genomic editing method 10:39
Criss-cross theory of conflict 4:591
Criterion designs, changing 21:986
Criterion of meaning, verifiability 14:316–317
Criterion of relative abundance 24:483
Criterion Orientated Personality Scales (COPS) 17:830
Criterion validity 6:345, 20:240
Criterion-referenced grading 3:823
Criterion-related validity of SJT scores 22:16
Critical anthropological theory, human rights 11:360
Critical Appraisal Skills Program 8:329
Critical approaches
emotion, history of the concept 7:472
geographic thought 10:74
Critical awareness 16:197
Critical cartography 23:85–86
Critical cosmopolitanism 14:14
Critical criminology 5:244–250
citizenship 5:246–249
commodity 5:244–246
globalization 5:244
human capital 5:244–249
new developments and emerging trends 5:244
political economy 5:244–246, 5:248
postmodern societies 5:245–246
problem of reification 5:244–246
hyper-realists 5:245–246
Marx on 5:244–245
situationists 5:245
risk–captivity–harm thesis 5:246–247
of shadow 5:247–248
cultural realm and ultramodern shadow 5:248
linguistic realm and ultramodern shadow 5:247
material realm and ultramodern shadow 5:247–248
symbolic realm and ultramodern shadow 5:247
social justice 5:246–248
social person 5:244–245, 5:247–249
spectacle 5:245
of stranger 5:248–249
Critical discourse analysis (CDA) 3:615–616, 6:503, 6:506, 18:801, 19:673
in sociolinguistics 22:950
Critical discursive research, of racism 19:867
Critical elite theories 7:393–394, 7:397–398
Critical emancipatory psychology 5:262
Critical friend 7:579
Critical geopolitics 10:93–95
Critical health 10:628–629, 10:631
Critical Health Competence Test 10:654
Critical history 11:129
Critical incidents technique 4:775
Critical legal studies (CLS) 5:251–254, 5:267, 5:931, 13:475, 13:510, 13:741, 13:792, 13:794
antecedents 5:251–252
challenges 5:253
critiques 5:251–252
distinction between 'public' (coercive regulatory) and 'private' (freely-willed contractual consensual market) relations 5:253
doctrinal or policy arguments 5:252–253
early 5:252–253
European 5:254
ideologies 5:251
later 5:253
movements 5:251, 13:551–552
use of term 5:251
Critical management studies, power and authority 17:407–408
Critical masses 3:311
in social protests 22:513
Critical morality 13:483–488
Critical naturalism 20:23
Critical nostalgia 17:37
Critical ontology of ourselves 1:749
Critical participation, engineering 7:642–647
Critical participatory action research (CPAR) 17:540
Critical path network, task network models 16:540, 16:541
Critical pedagogy, of foreign languages 9:329, 9:331
Critical period model 14:68–69, 14:69f
with later modification 14:68–69, 14:69f
Critical periods 20:57–58
Critical pessimism 5:266
Critical planning theory (CPT) 5:255–260
contemporary neoliberal interrogation 5:257–259
critical critique 5:255–257
deconstruction 5:255
ideology 5:255, 5:257
importance of 5:259
liberal pragmatic interrogation 5:255–256

Critical planning theory (CPT) (*continued*)
neoliberalism 5:255
political economy 5:255
postpolitical 5:258
poststructural interrogation 5:256–257
poststructuralism 5:255–256
pragmatism 5:255
structuralist political economy interrogation 5:255
Critical political ecology, sustainable development 23:815
Critical political geography 23:83–84
Critical psychology 22:226
capitalized 5:262
critique of 5:264
Frankfurt School 5:261, 5:264
functional-historical analysis 5:263
Hegelian dialectics 5:263
Holzkamp School 5:262–264
Marxist–Leninist psychology 5:262–264
New Left 5:262, 5:264
organization of school of 5:262–263
origins and early history of 5:261–262
political psychology 18:430–431
themes and topics of 5:263–264
unemployment 24:735
Critical race 23:327
studies 19:805
Critical race theory (CRT) 1:787, 5:266–270, 5:518, 13:510, 13:792, 19:796, 19:805, 23:324–325
in American legal academy 5:267
articulation on race and racism 5:268
in Canada 8:154
future directions of 5:268–269
oppression, forms of 5:268
origin 5:266–267
reading of legal doctrine 5:267
themes of 5:266–268
traditional black/white paradigm 5:268
Critical rationalism 5:271–276, 14:317, 18:526
Albert on 1:495–496
correspondence theory of truth 5:272
criticisms 5:272–273
demarcation 5:271, 5:274
dogmatism 5:272, 5:274
fallibilism 5:271–273, 5:275
falsifiability 5:271–272
hypothetico-deductive method 5:273–274
methodology improvement 5:273–275
nonjustificationism 5:275–276
rationality 5:271–272, 5:274–276
realism 5:272
reason 5:275–276
truthlikeness 5:272, 5:275
Critical realism 1:896, 5:277–280, 10:787, 14:416, 14:419, 22:225, 22:490–491, 25:194–195
Albert on 1:495–496
assemblage theory 5:279–280
Bhaskar on 20:23
and causal mechanisms 14:874–875
causation 5:277–279
conceptualization 5:280
critical reception 5:279–280
definition 5:277
and fallibilism 5:277
humanistic geography 5:279
idiographic 5:278, 5:280
influence on geography 5:277–279
intellectual context 5:277
nomothetic approaches 5:278, 5:280
research design 5:278–280
and social mechanisms 22:415
social work-specific research 20:502–503
Critical reflection 5:289–291
and reflexivity 20:82–83, 20:85–86
social work education 22:779–780
Critical reflexivity
and social work practice 5:281–286
assessment 5:282–283, 5:285
challenges 5:285
and critical reflection 5:284–285
epistemology 5:281–285
neoliberalism 5:283, 5:285
objectivism 5:281–283
reflection 5:281
reflexivity 5:281–282
and social work 5:284
Critical research 7:415
participatory action research 17:535–538
Critical scholarship, race and law 19:807
Critical social enquiry (inquiry) (CSE) 5:294, 5:299–302, 22:269–270
and social work 22:273
critical social sciences 22:269, 22:273–274
and social justice 22:269–273
and social research 22:269, 22:272–273
Critical social sciences 5:294, 5:301, 18:626–627, 22:269, 22:273–274
Critical social theory 5:288
Critical social work
contemporary understandings 5:287–288
history of 5:287
practice 5:287–292
common ideas underpinning 5:288–290
contextuality 5:291–292
critical reflection and deconstruction/reconstruction 5:290–291
critical social theory 5:288
discourse 5:289–290
empowerment 5:288–289
feminism 5:287–288, 5:292
identity creation 5:290
identity politics 5:290
ideology 5:288
knowledge 5:289
language 5:289–290
narrativity 5:291
postmodernism 5:287, 5:292
power 5:289
problematization and research 5:291
radical social work 5:287
social difference 5:290
Critical sociology 22:996
and alienation 1:541
Boltanski 2:747–748, 2:751, 2:753
Critical Sociology 14:649
Critical studies, diffusion 6:404
Critical technology assessment 22:873
Critical theories 4:758, 5:293–310, 17:542, 17:633, 18:499, 19:873
Adorno's 5:295–296
of anarchism 1:683
Benhabib's 2:519, 2:522–524
Boltanski on 2:751
communication to recognition 5:308–309
contemporary 5:304–307
communicative turn 5:305–306
criticality 5:306–307
debate over modernity 5:307–308
democratization 5:306, 5:309
of educational sociology 7:290–291
environmental concern 5:299–300
explanation and interpretation/theoretical and practical reasons 5:301–302
and feminism 5:307
Frankfurt school 5:294
Fraser's 5:299
Fromm's 9:458–459, 9:462–463
gender issues reflected in feminist philosophy of science 5:299
globalization 5:300
Habermas's 5:296–298, 5:304–309, 10:469–473, 10:787
and German social theory 5:296–297
moral theory 5:302
and postmodernism 5:297
Hegelian concept 5:304
Honneth's 5:298, 5:304, 5:308–309, 11:176
Horkheimer's 5:295–296
early research program 5:294–295
ideal speech situation 5:305–307
Marxism, end of 5:300–301
methodological pluralism 5:301
positivism 18:627–628
pragmatism 5:301
procedural view of reason and justice 5:297–298
reason and linguistic turn 5:302
reification 20:203–205
resistance 5:308–309
of society, and Frankfurt School 9:381
Critical theorizing 22:961–962
American-style critical theories 22:962
Frankfurt school 22:961–962
postmodern critical theories 22:962
Critical thinking, and intelligence 12:320
Critical thought, and eudaemonism 8:215
Critical turns, tourism 24:471
Critical value 11:495–496
Critical visual methods 3:623
Critical white studies 25:571, 25:573
Critical whiteness 2:732
Critical window 7:848–849, 7:851
Critical/sensitive period 9:486
Critics, in cultural production/consumption 5:443
Critiques
Boltanski on 2:747, 2:750–753
engineering 7:643, 7:646
of human sciences 11:383–384

of human terrain system 11:395
ethical 11:395
financial 11:395
institutional 11:395
methodological 11:395
military strategic 11:395
organizational 11:395
political 11:395
professional 11:395
scholarly 11:395
of privacy 18:950
social science 22:692–693
Critiquing professional beliefs 4:799–800
CRLA *see* California Rural Legal Assistance (CRLA)
CRM *see* Cause-related marketing (CRM) programs; Crew resource management (CRM); Cultural resource management (CRM); Culturally reflexive model (CRM)
Croce, Benedetto 11:93–94, 11:96
Croft, Balliol 14:621
Croft, W.A. 17:555–556, 17:556t
Crohn's disease 8:457
Croly, David 4:524
Croly, Herbert 4:524
Cromwell, Oliver, English Revolution 7:649–650
Cromwell, Richard 7:649–650
Cronbach, Lee J. 12:304
Cronulla riots of 2005 2:241
Crony capitalism 2:571
in East Asia 6:791
Cronyism 5:345–346
Cropping intensity index 13:240
CROs *see* Contract research organizations (CROs)
Crosby, Alfred
works
The Columbian Exchange 25:343–344
Ecological Imperialism: The Biological Expansion of Europe, 900–1900 7:749, 25:343–344
Cross Bronx Expressway 24:955
Cross cousins 13:62
Cross cultural dimensions of research ethics 20:508–513
Cross fertilization 22:304
of cultures 5:628
Cross Media Annotation System (xMAS) 6:897
Cross modality matching 14:802–803
Cross referencing (publishing) 18:917
Cross-border movement 7:18
Cross-border terrorism 5:167
Cross-boundary regional plans 20:138
CrossCheck service 12:864
Cross-classified random effects model (CCREM) 17:731–733
Cross-country studies
on effect of beauty on labor market 2:388–389
qualitative studies, for economic development 5:531
Cross-country unemployment 12:104
Cross-cultural aesthetics and creativity 17:710
Cross-cultural application of performance 17:713
Cross-cultural aspects, of empathy development 7:552–553
Cross-cultural attachment research 2:140
Cross-cultural awareness, of mental health disorders 5:574
Cross-cultural communication 13:257
Cross-cultural comparisons
in ethnology 8:199–200
of teacher motivation 24:69
Theory of Mind 24:270–271
urban anthropology 24:774–775
Cross-cultural considerations, in self-development 21:495
Cross-cultural context, effects of parenting on prosocial behavior in 19:247–248
Cross-cultural data 8:461
Cross-cultural developmental research 6:326–331
Cross-cultural differences, in neonatal crying 16:476
Cross-cultural findings
emic approach 2:562–563
etic approach 2:562–563
five-factor model (FFM) 2:562–563
Cross-cultural frequency of war 4:615
Cross-cultural generalizability 21:462–463
of I/E model 12:428
Cross-cultural identity status model 21:436–437, 21:437f
Cross-cultural knowledge 11:790
Cross-cultural leadership 13:667
Cross-cultural management 11:785
organizational citizenship behavior 17:319–320
Cross-cultural perception and expression of mental illness 5:572
Cross-Cultural Personality Assessment Inventory (CPAI-2) 9:252
Cross-cultural psychology (CCP) 5:63–68, 5:311–317, 5:481, 5:570, 12:357–358, 14:397, 22:534–535
applications
biological revolution in psychology 5:316
cross-cultural variance 5:316
cultural anthropology 5:316
applied research 5:314–315
cultural pluralism and tyranny of the shoulds 5:64
definition 5:311
cultural psychology 5:311
culture-comparative research 5:311
indigenous psychologies 5:311
empirical research 5:313–314
future outlook 5:315–316
history 5:311–312
cognition 5:311
cultural anthropology, shift in 5:311–312
individualism–collectivism 5:311
perception 5:311
self-report questionnaire 5:311
social–psychological 5:311
and international advertising 12:457
and mental health disorders 5:570
methodology 5:313
experimental designs 5:313
grouping of humans 5:313
methodological issue 5:313
mixed methods 5:313
structural equivalence 5:313
multicultural psychology 5:65–66
racial identity theory 5:66
racial–cultural identity development 5:65–66
theory 5:312–313
behavior content 5:312
psychological process 5:312
sociocultural psychology 5:312
Cross-cultural research 5:318, 15:254–257, 21:462–463
on abusive supervision 23:706
power distance orientation 23:706
approaches, for early friendships 9:450–451
and bias 5:318
country-level studies 5:321
and equivalence 5:320–321
equivalence and transfer problems in 7:929–933
adaptation and assembly 7:930
construct clarification 7:930–931
documentation 7:932
emic 7:930–931
etic 7:930–931
item translation and formulation 7:931
scale development 7:932
test evaluation 7:932
individual-level studies 5:321
linguistic equivalence 5:320–321
minority acculturation 8:145
sampling cultures 5:321
sexual harassment 21:725
sources of bias 5:318–319
subjects 5:321
types of equivalence 5:319–321
Cross-cultural research methods
in psychology 5:318–321
in sociology 5:322–326
Cross-cultural sensitivity 5:574
Cross-cultural similarities and differences, in prosocial moral reasoning 7:318
Cross-cultural studies 2:792, 5:322
bias 5:322–323
sources 5:323
strategies for identifying and dealing with 5:324t
characteristics 5:322
of child development 19:362
on childhood 3:456–457
Chinese conceptualization 5:323
on eating disorders 6:826–827
equivalence 5:322–323
linguistic 5:325
types of 5:323–324

Cross-cultural studies (*continued*)
individual- and country-level studies 5:326
interpretation of 5:322
ambiguity of interpretation 5:322
mother–child communication 15:877
of psychosocial development 7:941
quantitative and qualitative approaches, use of 5:322
sampling cultures and subjects 5:325–326
Western conceptualization 5:323
Cross-cultural study of education 5:327–333
academic achievement 5:327–331
context 5:327, 5:329
culture 5:327, 5:329, 5:331
international assessments of achievement 5:327–328
International Association for the Evaluation of Educational Achievement (IEA) 5:327–329
literacy 5:327, 5:329–331
math education 5:327–329
motivation 5:327, 5:329, 5:331
Programme of International Student Achievement (PISA) 5:327, 5:329
schooling 5:327–329
science education 5:327–328
strategy use 5:329–331
support for learning
general trends and issues 5:330–331
home and family 5:329
individual attitudes and beliefs about learning 5:329–330
schools and teachers 5:328–329
universalism 5:331
Cross-cultural sustainability 7:106
Cross-cultural training (CCT) 16:75
in industrial and organizational psychology 5:334–341
cultural intelligence 5:336
focus on culture-specific and culture-general training 5:337–338
future research 5:340
inclusion of contextual and personality variables 5:338
intercultural training 5:334, 5:336, 5:339–340
leadership 5:338–339
mode and focus of training 5:335t
non-expatriate research 5:337
orientation programs 5:334–336, 5:340
postdeparture training 5:338
recent empirical research in 5:337–338
social learning theory 5:335–336
specific *versus* general training 5:337
teams 5:339–340
types of training 5:336–337
U-curve theory of adjustment 5:335, 5:335t
virtual team training 5:334–335, 5:339–340
Cross-cultural universality 21:679–680
Cross-cultural variations
action readiness 5:544
appraisals 5:544
and emotions 5:542–544
life phases 5:511
in personality 17:867–868
Cross-culture 6:748–750
of emotional competence 6:749–750
Cross-disability protests 6:474–475
Cross-disciplinary antecedents, of youth culture studies 25:809
Cross-disciplinary associations 22:677
Cross-disciplinary collaboration 21:36
Cross-disciplinary orientation, infrastructure 22:669
Cross-disciplinary research 20:881
Cross-disciplinary schools of thought 21:129
Cross-domain binding of memory 15:142–143
Cross-domain integration, of cognition and behavior 1:428
Cross-dressers 13:875
Cross-entropy goodness-of-fit criterion 23:414
Cross-Impact Matrix of PPA 17:761–762
Cross-indigenization 11:790
Crossing, in sociolinguistics 22:951
Cross-language
activation 21:352
interaction 21:349–350, 21:352–353
Crosslanguage mismatches in grammatical categories 3:918
Cross-level bias, in crossover designs 5:353
Cross-level fallacy 21:2
Cross-level integration, of cognition and behavior 1:428
Cross-level interactions 16:50
Crossley, Nick, on music tastes 16:556
Crosslinguistic comparison 25:645
Cross-linguistic comparisons 14:209–210, 19:356–357
of first language acquisition 9:201
Crosslinguistic influence 21:349–350
Cross-linguistic influences (CLIs) 16:57–58
in foreign language learning 9:329
Cross-modal binding of memory 15:142–143
Cross-modal cues, in multisensory integration 16:95–97, 16:99–101
Cross-modal matching 21:604–605
Stevens on 23:454
Cross-national barometer surveys 5:845
Cross-national comparisons 22:325, 22:327–328
culture 5:730, 5:733
social mobility 15:643
Cross-national databases 5:817, 5:819–821
Cross-national patterns in litigiousness 14:268–270
Cross-national perspective, education in 5:342–348
Cross-national research 22:633–635
quantitative methods 19:719–724
Cross-national studies 3:826–827
on political participation 18:505–506
Cross-national survey research, International Social Survey Program 12:556
Cross-national work 18:491–495
Cross-norm effect 10:437
Crossover analysis
genetic programming 2:41–43, 2:42f
narrative networks 16:215–216
Crossover designs 5:349–353, 8:538
baseline measurements 5:349, 5:351, 5:353
between-unit information 5:352–353
blocking 5:349, 5:353
carryover effect 5:349–353
cross-level bias 5:353
higher-order designs 5:351–352
intraclass correlation 5:349
Latin square design 5:352–353
linear mixed model 5:353
Mainland–Gart test 5:351
model-based analyses 5:352–353
period effect 5:349–350
repeated measures design 5:349–350, 5:352–353
two-period two-treatment design 5:349–351
unit effects 5:352–353
washout period 5:349–351
Williams square 5:352
within-unit correlation 5:349–353
within-unit information 5:351–353
Crossover effects 12:453
Crossover stress, of work–family conflict 8:763
Crossover trials 24:539
Cross-perspective audit strategy 15:621
Cross-population differences in prosocial behavior 19:238–243
Cross-racial helping 10:759–760
motivations for 10:760
in-group *versus* out-group members 10:760
strategic helping 10:760
Cross-sectional analysis, in political science 18:519–520
Cross-sectional design
on adult life development in cognitive aging 1:401–403
mental illness epidemiology 15:180, 15:183
Cross-sectional methodologies
applications in developmental psychology 5:354–360
Cross-sectional micro-level demographic surveys 5:820
Cross-sectional micro-level social surveys 5:820–821
Cross-sectional studies
aggression, social psychology of 1:376
epidemiology 7:828, 7:828b
simultaneous equations 21:982–983
Cross-situational consistency
in personality traits 17:920–921
Cross-subsidization 23:320–323
revenue sharing 23:322–323
sport labor market 23:320–322

Cross-validated error rate 16:119
Cross-validation
model selection 15:649, 15:656
Crosswords, and neuropsychological rehabilitation 2:817
Crow, Thomas 25:212
Crowd behavior
and identity 11:530
self-categorization theory 21:457–458
Crowd psychology 7:417
Crowding
and cognitive skills in children 8:732
Crowding hypothesis 12:694
Crowding in 12:404
Crowding out 1:348, 3:361, 12:404
Crowds
and collective action 4:151–154
collective psychology, origins of 4:151–152
ESIM of 4:153–154
site of pathology, debating 4:152
social identity and 4:152–153
during crisis and revolution 5:361
deindividuation 6:38–41
groups 10:425–426
protest 4:151
sporting 4:151
Crowds in history 5:361–366
civil war 5:364
community politics 5:363
conceptual constrictions 5:364–365
crowd riots 5:364
ethnic violence 5:361, 5:364
face-to-face public 5:365
mass protest 5:364–365
minority persecution 5:364
phases and schools 5:362
psychology 5:361–362
revolution 5:361–365
social history 5:361–362, 5:364
subsistence and liberties 5:362–363
Thompson on 5:361–363
Tilly on 5:361–362
urbanization 5:361–363
Crowd-sourced geographic information 23:129
Crowd-sourced organizations 6:239
Crowdsourcing 4:952–953, 12:588–589, 21:256, 22:423
Crown Prosecution Service (CPS) 19:210–211, 19:213
CRPD *see* Convention on the Rights of Persons with Disabilities (CRPD)
CRT *see* Critical race theory (CRT)
CRTs *see* Cluster randomized trials (CRTs)
Crucé, Emeric 4:572
Crucial experiment, Strong Program 23:594
Crude birth rate (CBR) 5:739, 6:111, 6:122
Crude death rate (CDR) 1:341, 6:119, 6:122
Crude divorce rates 14:518
Crude oil, peak oil 15:554–560
Crude rates, demography 6:110
Cruelty, torture and 24:442–446
Cruikshank, J., on settler colonialism in North America 17:25–26
Cruising 19:762
Crumbie Brown, John, works, *Hydrology of South Africa* 7:750
Crummey, Robert, on sexual minorities of Russia 21:740
Crusades 20:336
warfare 25:415–416
CryENGINE 3 25:133
Crying 16:493–494, 16:499
behaviors, neonatal 16:475–480
contagious 7:551
and empathy 7:543
gender differences in 9:755
Crystallization 21:402–403
Crystallized intelligence (Gc) 12:285, 12:317, 12:324–325, 14:101
Crystallized self-knowledge 17:816
CS *see* Citizen science (CS); Cognitive science (CS); Conditioned stimulus (CS)
CSA *see* Catholic school advantage (CSA); Cognitive styles assessment (CSA); Community supported agriculture (CSA)
CSAs *see* Combined statistical areas (CSAs)
CSBA *see* Certificate for School Business Administrators (CSBA)
CSBI *see* Child Sexual Behavior Inventory (CSBI)
CSBM *see* Certificate of School Business Management (CSBM)
CSC *see* Coordinate Structure Constraint (CSC)
CSD *see* Cortical spreading depression (CSD)
CSE *see* Commercial sexual exploitation (CSE); Core self-evaluations (CSE); Critical social enquiry (inquiry) (CSE)
CSES *see* Comparative Study of Electoral Systems (CSES); Core Self-Evaluations Scale (CSES)
CSI *see* Centre de Sociologie de l'Innovation (CSI)
CSI effect 9:958
CSM *see* Common-sense model (CSM)
CSOs *see* Civil society organizations (CSOs)
CSP *see* Center of Systemic Peace (CSP); Committee on Science and Practice (CSP); Concentrated solar power (CSP); Constraint-satisfaction problem (CSP)
CSR *see* Codes of Corporate Social Responsibility (CSR); Cognitive science of religion (CSR); Complete spatial randomness (CSR); Comprehensive School Reform (CSR) projects; Content and Service Registry (CSR); Corporate social responsibility (CSR)
CSW *see* Commercial sex workers (CSW)
CSWE *see* Council on Social Work Education (CSWE)
CT *see* Cognitive therapies (CT); Computerized tomography (CT); Counterterrorism (CT); C-Tactile (CT) afferents
CTA *see* Constructive technology assessment (CTA)
C-Tactile (CT) afferents 24:16
CTCs *see* Cholera Treatment Centers (CTCs)
CTOs *see* Community treatment orders (CTOs)
CTS *see* Conflict Tactics Scale (CTS)
CTSM *see* Classical true score model (CTSM)
CTST *see* Classical true-score theory (CTST)
CTT *see* Classical test theory (CTT)
CTU *see* Chicago Teachers' Strike (CTU)
CTUR *see* Oxford Centre for Time Use Research (CTUR)
CUA *see* Cost-utility analysis (CUA)
Cuba 1:638–639
carnival in 3:147
community participation in health 4:402
energy revolution 23:821
Humboldt on 25:280, 25:282, 25:284–285
and social work 14:664, 14:667–668
Cuban diaspora 5:367, 5:369
Cuban Missile Crisis (1962), United States presidential handling of 1:105
Cuban Revolution 1:621, 5:367–374
Communist Revolution 5:368–370
economic crisis and revolutionary euphoria 5:369
institutionalization and foreign military adventurism 5:369
last decade of the twentieth century 5:369–370
origins of the revolution and leadership legitimization 5:368–369
revolutionary dynamics 5:369
domestic politics 5:371–373
challenges 5:371
corruption 5:371
human rights 5:372–373
race relations 5:371–372
women 5:372
world wide web, cellphones, and bloggers 5:373
Spanish colony to republic 5:367–368
1933 Revolution 5:368
early republic 5:367–368
Republic's failure 5:368
in the twenty-first century 5:370–371
commodification of Cuban labor 5:370–371
foreign exchange 5:370–371
oil resales and joint ventures 5:370
old and new problems 5:370
tourism 5:370
US–Cuban relations 5:373
Cuban-Spanish American war 5:367
CUDOS (communalism, universalism, disinterestedness, originality and skepticism) 21:169
Cue conditioning 8:849–850, 8:852–853, 8:850f
Cue description, and retrieval 20:611–612
Cue ordering, in heuristics 10:831

Cue-adaptive reasoning account (CARA) 10:706–707
Cue-competition hypothesis, and symptom perception 23:867
Cue-overload account of proactive interference 19:5–6
Cues, recall 19:773
Cues-filtered-out theories, of computer-mediated communication 4:505–506
CUF *see* California Urban Futures (CUF) model
CUHR *see* Culturally unidentifiable human remains (CUHR)
Cuisines 9:313–314
  in anthropology 9:290–295
    authenticity 9:292–294
    culinary tourism 9:293–295
    culinary traditions 9:292–293
    cultures 9:290–292, 9:294–295
    ethnic foods 9:290–295
    future directions 9:294–295
    gastronomy 9:292–294
    globalization 9:292–293
    modernity 9:293
    national and regional cuisines and identities 9:293–294
    prevailing ethnographic and anthropological approaches 9:290–291
Culex quinquefasciatus 7:818
Culin, Stewart, works, *Korean Games* 2:71
Culinary tourism 9:293–295
Culinary traditions 9:292–293
Culinary triangle 9:291
Cullen, William 17:27–28
Culpa in contrahendo principle 3:671
Cult of personality, in Soviet Union 20:823
Cult of Yahweh 19:197
Cultic organization 3:143
Cultivation theory 15:30
Cultlike groupings 22:697
Cults 20:336, 20:360
  cargo cults 3:139
Cultural, production of 5:602–607
Cultural acquisition, in education 1:730
Cultural adaptation 15:472
Cultural advance 7:621
Cultural advocacy 20:758
Cultural affiliation, concept of 1:888
Cultural alienation 1:553–554
Cultural analysis
  ANT 5:526
  formal methods of 9:357–367
  of sport 23:315–316
Cultural and Creative Industries Development Act 2010 5:456–457
Cultural anthropology 1:774, 8:198, 8:478, 11:110, 15:687, 18:364–365, 18:384, 19:313, 24:590
  of body 2:720
  and cognitive psychology 4:105–106
  definition 1:765
  ethnohistory 8:192, 8:195
  fieldwork in 9:149–152
  and *Geist* 9:670
  Melanesia 17:154
  proto-anthropology 1:765–766
  and psychiatry 19:314
  and reproduction *see* Reproduction
  social control in 4:824–825
  and syncretism 23:882
Cultural approach
  geographic thought 10:72–73
  values 25:33–34
Cultural arbitrary 22:332
Cultural archive, spirit possession 23:283
Cultural areas
  of Europe 8:274
  of North Americans 1:634
Cultural aspects of migration 15:427–432
Cultural assimilation model 22:704
Cultural autonomy, language policy for 13:358
Cultural awareness 11:392, 11:785
Cultural behavior 17:952
  sexual masochism 17:952
Cultural beliefs and practices
  guiding child behavior and child socialization 3:455–456
  of indigenous people 9:964–965
Cultural belonging, of sexual minorities in China 21:737–738
Cultural bias 2:884
Cultural borderland 2:761
Cultural capital 3:93, 5:9, 5:375–381, 5:450, 5:463, 8:130, 16:791, 22:326–327, 22:331–332
  arts participation 2:7–9
  Bourdieu's concept of 2:778–780, 5:375–376, 8:132
  concept of 22:326
  conspicuous consumption/distinction 4:695, 4:697–698
  and consumption 4:760
  criticisms 5:379–380
  and cultural intelligence 5:438
  definitions 5:375–376
    Bourdieu's 5:375–376
  and education 7:145–146, 7:293
  empirical research 5:376–379
    findings 5:378–379
    operationalizations of cultural capital 5:376–377
    operationalizations of educational outcomes 5:377–378
  environment 15:453, 15:455
  family and schooling 8:753–754
  and field theory 9:144, 9:146
  future directions 5:379–380
  hypothesis 8:132
  remaking of 5:451
  and social inequality 22:328
  studies 9:362
  and taste 5:537–538
Cultural change, and reconceptualization of culture 1:770
Cultural characteristics
  mass killings/genocide 14:704
  autocratic *versus* democratic societies 14:704
  cultural devaluation 14:704
  monolithic *versus* pluralistic societies 14:704
  overly strong respect for authority in society 14:704
  past victimization or suffering 14:704
Cultural citizenship 3:638, 14:724, 16:19–20
  and migration 15:423
Cultural city 3:665–666
Cultural classification systems 5:442
Cultural clock 1:320
Cultural codes, mafias 14:425
Cultural community 22:264
Cultural competencies 5:433–439, 17:431–432
  arts socialization 2:9
  and intercultural communication 12:358–359
  and racialized groups 22:822–823
  of Roma 20:758
  and spirituality 23:288–289
Cultural conflict
  and delinquency 6:69
    extensions or reformulations 6:72
  models 16:443
Cultural consecration 5:443
Cultural consensus analysis 19:361
Cultural consequences, of genomics 10:40–41
Cultural considerations, in death and dying 5:861–862
Cultural consistency and punishment 19:646
Cultural construction of intergenerational ties 9:854–855
Cultural constructs, kinship 13:53, 13:55
Cultural consumption 5:442, 10:283
  European perspective 13:837–844
    cultural and recreational expenditures 13:837–839
    cultural participation trends 13:839–842
    international comparison 13:837–839, 13:842
  field, evolution affecting 5:449–450
    democratization of education 5:449–450
    increased rates of mobility 5:450
    rise of cultural omnivore 5:450–451
  and free time 13:847–849
    arts participation 13:847–849
    conspicuous consumption 13:849
  and leisure 13:845–851
  social inequality
  see also Social inequality
  social patterning of 4:759–760
Cultural contact 13:342–343
Cultural context
  of family of origin, fertility 9:21–22
  of fathering 8:844–845
  of international advertising 12:457–458
Cultural continuities
  in South America 23:46–47
    fertility 9:24
Cultural contradictions 2:503

Cultural creatives 18:666
Cultural creativity, and everyday life 8:324
Cultural criticism 1:766–767
role of Mumford of 16:137–141
Cultural critique
anthropology 5:382
in America 5:383–384
early twentieth-century ethnography 5:382–383
ethnicity, nation-states, and globalization 5:385–386
in Europe 5:384
feminism and Marxism 5:384–385
roots of critique in Enlightenment 5:386
and comparison 1:774–775
fetishism as 9:107
and industrial societies 11:863
Cultural defense 5:491–497, 13:811
anthropological jurisprudence 5:494–496
challenges for future 5:496
immigrants 5:493–494
multiculturalism 5:494
pluralism 5:492, 5:496
Cultural deficit theory 8:131
Cultural democracy 16:20
Cultural democratization 5:462–463
Cultural determinism, and self 21:446
Cultural development 4:752
psychological development as 25:331–332
Cultural developmentalism 5:585–586
Cultural deviance 15:25
and delinquency 6:69
Cultural difference model 8:132
Cultural difference theory 8:132
Cultural differences
in adolescent sexuality 21:841–842
in childhood sexuality 21:841
contra-hedonic motivation 1:245
and cosmopolitanism 5:43, 5:45
in counseling and psychotherapy 5:63–68
and ethnicity 8:148
and indigeneity 11:756, 11:758–759
in moral development 15:781–787
in morality 15:806–810
mother–child communication 15:874–880
in shame, guilt, and pride 21:478–479
in social referencing 22:560–561
spatial cognition development 23:123–124
cultural variation in FoR use 23:123–124
language and spatial categorization 23:123
student achievement 5:346
Cultural diffusion 5:403
Cultural dimension of science 21:180
Cultural dimensions system 12:358, 12:358t
Cultural discontinuities, in education 7:145
Cultural discrimination
and race 19:810
sexual minorities 13:880
Cultural dissonance 5:451
Cultural distance, and internal migration 12:514–515
Cultural diversity 16:22–23
anthropology, and advocacy 1:224
East Asian studies 6:785
Lévi-Strauss on 13:907–909
organizational culture 17:349–350
and political psychology 18:431–432
prevention science 18:888–889
relativism 20:234
Cultural domain 5:730
Cultural ecology 1:769, 6:838, 8:132–133, 10:547, 16:337, 17:22, 22:254–255
Cultural editing 5:416
Cultural efficacy, international social work student exchanges 12:563
Cultural entrepreneurs 5:606, 14:565
Cultural Environment Movement 15:56
Cultural epidemiology 19:362
Cultural equity 7:927
Cultural essentialism 19:855
Cultural evolution 5:388–389, 5:394–408, 8:357–360, 8:362, 8:389–390, 16:332, 25:274–276
archaeology 5:394–395, 5:397–399
versus biological evolution 1:85
continuous traits and linear Gaussian models 5:404
cultural macroevolution 5:390
cultural niche construction 5:406
cultural transmission 5:390–392
cultural transmission and consequences 5:402–404
group effects and conformity 5:403–404
horizontal and oblique transmission 5:403
innovation 5:404
vertical transmission 5:402–403
cumulative 5:391
Darwin's theory 5:388–389
and disgust 6:547
domain-specific *versus* content-free biases 5:390–391
and ecumene 7:110
and energy 7:621–622
Europe 5:395, 5:398–399
evolution of language 8:367–368, 8:370, 8:372, 8:374–375
evolutionary economics 8:431, 8:433–435
functional explanation 9:522
gene-culture coevolution 9:814, 9:816
gene–culture coevolution 5:391
gene–culture evolutionary theory 5:405–406
genetic change, culture-driven 5:391
genetics 5:394–395, 5:398–399
holocene prehistory of central Europe, perspective from 5:398–399
human biology 5:395–396
human cooperation, evolution of 5:391
of humans 11:254
and language development 13:322
language families 5:394–397
linguistics 5:394–399
metatheoretical considerations 5:405
methods 5:390
of morality 15:815–817
cultural group selection 15:816–817
gene-culture coevolution 15:817
moral memes 15:816
natural and cultural selection 5:404–405
neural basis of cultural transmission 5:391
New Guinea 5:398–399
phylogenetic arrays, importance of migration in spreading 5:396–397
phylogenetic studies of archeological and ethnographic cultures 5:397–398
phylogenetics 5:390
phylogeny and reticulation in prehistoric contexts 5:395–396
polynesia 5:397–399
processes of 24:129–130
reticulation 5:394–395
simulation approaches 5:406–407
social learning 5:388, 5:402, 5:407
and technology 24:133–134
theory, approaches to 5:402
Cultural evolutionary theory 24:132
chiefdoms and 3:376–377
Cultural evolutionism 8:386, 17:22
Cultural exchange 20:427–428
and Renaissance 20:427
Cultural expenditures, leisure and cultural consumption 13:837–839
Cultural factors
fertility transition 9:96
of mania 14:487
Culture
and health care 22:939
Cultural feminism 9:737
Cultural feminists 9:730–731
Cultural field 5:440–444
Cultural flows, and globalization 10:233–235
Cultural forms, economics and economic theory as 5:539–540
Cultural freedom 2:907
Cultural frontiers 2:761
Cultural gatekeepers 5:440–441
Cultural geography 5:409–413, 6:406–407, 22:301, 22:303, 22:304, 22:305, 23:83
Berkeley School 5:409
embodiment 5:409, 5:411–412
landscape 5:409–412
mobilities 5:412
new 5:409–411
performances and performativities 5:411–412
places 5:411–412
representations 5:409–412
tourism 5:410–412
urban 24:802
Cultural globalization 10:206, 25:83–84
Cultural goals and practices
differences, during preschool age 18:224
Cultural group selection, human cooperation 11:264–265
Cultural groups 1:901
Cultural health 10:628–631
Cultural heritage 5:414–419
anthropology 5:414–418
authenticity 5:414–416
conservation 5:414–417

Cultural heritage (*continued*)
conservation of 5:488–490
intangible 5:414, 5:417–418
and international organizations 12:540
invention of tradition approach 5:414–415
nostalgia 5:416–417
religion 5:417
tradition 5:414–417
UNESCO 5:414, 5:416–418
urban anthropology 5:417
world heritage 5:414–418
Cultural hierarchies, and cultural participation 5:449
Cultural Historical Activity Theory (CHAT) 6:611, 22:6
Cultural histories 5:420–425, 11:41, 11:102–106
anthropology 5:420, 5:422–423
Baroja on 3:150–151
Burke on 2:924–928
communication and transportation 4:274–275
cultural forms 5:420
definition 5:423–424
demography history 5:420–422
and family 8:738
Freyre on 9:420, 9:422
human science 5:422
intellectual history 5:421–423
and literacy 14:236, 14:238
memory 5:422
mentalities 5:420–423
founders of Annales 5:420–421
golden age 5:421–422
historical psychology 5:422
success and criticisms 5:422–423
plurality of practices 5:424
and prosopography 19:258
psychological history 5:420, 5:422
and Reformation/confessionalization 20:99–100
Roma people 20:752–753
social science 5:422
sociology 5:421–422
Cultural homogenization 25:367
Cultural human rights 11:355
Cultural hybridity, in South America 23:47
Cultural hybridization, and ethnicity 8:159
Cultural identity 9:434–435, 14:707, 21:432–433, 21:438, 22:313
globalization and 14:710
international debate on 14:709
and mass media 14:707–711
media, language and 14:710
and supernationalism 14:709–710
Cultural identity achievement, and adolescent social competence 22:218
Cultural idioms
of distress 5:571
nationalism as 16:294
Cultural impacts 7:108
Cultural imperialism 3:309–310, 14:709–710, 15:55, 22:824
East Asian studies 6:785–786
international debate on 14:709
and media audiences 2:219–220
and postcolonial geography 18:644–646
Cultural implications, in disasters 22:758
Cultural incompatibility, xenophobia 25:799–800
Cultural industries 5:730–731, 7:65, 14:735
advertising as 1:213–214
artistic/cultural markets 14:564–567
in China 5:597–598
and cultural objects 5:536
models of 5:728–729
Cultural inequality 22:331–332
Cultural influences
on children's development 3:453–455
basic trust and attachment 3:454–455
cultural scale and complexity 3:453
developmental goals 3:455
emotional development 3:454
gender differences 3:453–454
temperament development 24:194–195
Cultural information regimes, and Internet 12:590–591
Cultural integrationalism, in Caribbean 1:640
Cultural intelligence (CQ) 5:336, 5:433–439, 22:406
and cultural competencies 5:434
definition 5:433
future research 5:438
measurement 5:434–435
combining complementary measures 5:435
observer-reports 5:434–435
performance-based assessment 5:435
self-reports 5:434
as multidimensional intelligence 5:433
nomological network 5:435–438
antecedents 5:435–436
boundary conditions of effects 5:437–438
as mediator 5:437
as moderator 5:437
outcomes 5:436–437
and personality traits 5:434
Cultural Intelligence Scale (CQS) 5:434–435
Cultural invasion 15:55
Cultural issues
autism spectrum disorders 2:279
in genomics 2:624
issue evolution model 12:767
Cultural Jewish identity 12:871
Cultural knowledge 1:318
kinship 13:53
Cultural lag 13:524
Cultural learning, and mirror neurons 15:586
Cultural Learning Environment (CLE) 3:452
Cultural linkage 17:392
Cultural Logic: An Electronic Journal of Marxist Theory and Practice 14:649
Cultural macroevolution 5:390
Cultural markers, of transition to adulthood 24:577–581
Cultural market 14:564–568
art worlds 14:565
artistic reputation 14:565–566
cultural industry 14:564–567
gatekeeper 14:565, 14:567
high culture model 14:565
history 14:564–565
impact on culture 14:567
Internet 14:567
long-tail retailing strategy 14:567
marketization of art museums 14:566
neoliberalism 14:566
new technologies 14:567–568
privatization 14:566–567
production of culture 14:564–565
uncertainty 14:565
visual art markets 14:566–567
Cultural materialism 1:746, 7:622
Cultural matrix, and organizational citizenship behavior 17:319–320
Cultural mediators 5:440–446
cross-cultural exchange 5:445
developments impact on role and authority of 5:445
impact on symbolic production 5:442–444
commercial success 5:444
consumer behavior 5:444
cultural consecration 5:443
cultural evaluation (institutional context) 5:442–443
cultural legitimation 5:444
cultural tastes 5:444
reputation making 5:443
role in lifecycle of cultural products 5:440–442
censoring 5:442
classifying 5:442
cocreating/editing 5:441
connecting/networking 5:441
distributing 5:442
evaluating 5:442
meaning making 5:442
protecting 5:442
selecting from the supply (gatekeeping) 5:441
selling/marketing 5:441–442
supporting 5:442
Cultural memories 1:882, 4:188
and archaeology 1:881–884
Cultural microevolution 5:389
Cultural mistrust, of health care 19:816–817
Cultural mobility 5:376, 5:378–379, 5:605
arts socialization 2:8, 2:7f
and music industry 5:605
and religious realm 5:605
and sociopolitical 5:605
Cultural models 5:542–543, 5:546, 5:583–585
of cognitive anthropology 4:19
life phases 5:507, 5:512
Cultural models school 19:360–361
Cultural myths, and nostalgia 17:36
Cultural narratives, photos as 25:181–182
Cultural nationalism 19:855

Cultural neuroscience 5:566–567
Cultural niche construction 5:406
Cultural norms, and wealth flows 9:37
Cultural objects, creation and dissemination of 5:536
Cultural omnivores 9:362
  rise of 5:450–451
Cultural omnivorousness 5:451
Cultural participation 5:447–452, 22:333–334
  arts participation 2:10–12
  and cultural tastes 22:334–335
  current and emerging trends in theory and research 5:451
    remaking of cultural capital 5:451
    symbolic boundaries and social inequality 5:451
  distinction 5:450
  educational expansion 5:447, 5:449–450
  and educational outcomes 5:378
  facts and figures of trends 5:447–449, 5:448t
  fine arts 5:449
  Internet 12:590
  interpretation of trends in 5:449–451
    changes affecting field of cultural production and mediation 5:449
    evolution affecting field of cultural consumption 5:449–450
  omnivorousness 5:449–451
  and social networks 16:556–557
  studies 5:732–733, 5:732t
  trends in 13:839–842
Cultural patriotism 16:312
Cultural performances 17:708–709
cultural philistinism 16:829–830
Cultural pluralism 2:84, 5:64, 15:572–573
  in Caribbean 1:640
Cultural policies
  communications and media 5:706
  in East Asia 5:453–459
    countries in East Asia 5:454
    development 5:454–456
    flows of culture within East Asia region 5:457–458
    recent trends 5:456–457
  regimes in China 5:596–601
    art districts 5:598–600
    characteristics 5:601
    cultural industries 5:597–598
    political control *versus* creative freedom 5:600–601
    from subordinate politics to separate realm 5:596–597
  regimes in Western Europe 5:460–465
    defining 5:460–462
    funding 5:464–465
    from goals diversification to noncultural rationales 5:463
    historical legacies 5:460–461
    increasingly uncertain domain 5:461–462
    original trilogy 5:462–463
    rationales, specifying 5:462–463
    social, political, and institutional settings 5:463–464
    specifying policy domain 5:461
    structures and governance 5:463–465
Cultural populism 2:852–853
Cultural practices
  in South America 23:45
  and consumption 4:760
  for indigenous children 11:762
Cultural pragmatics 5:620
  Alexander on 1:527
Cultural principles, for physical activity 18:93
Cultural producers, sex workers as 19:282–283
Cultural production 5:602–603, 5:608–613
  artistic/cultural markets 14:564–565
  changes affecting field of 5:449
  dimensions of 5:608–611
    art 5:608–611
    authenticity 5:610
    careers, contracts, and status orders 5:609
    classification 5:609–610
    consumption 5:608, 5:610–611
    fan cultures, taste, and collective memory 5:610–611
    market and industry structures 5:608–609
    media 5:611
    music 5:608–611
    omnivorousness 5:611
    production factors 5:608
    prospects 5:611
  and education 7:147
  in education 1:730–732
  in networks 5:466–470
    art worlds 5:466–467
    Bourdieu's field as space of relations 5:466
    careers 5:468–469
    creative places 5:469–470
    markets and industries 5:469
    network structure and creativity 5:467–468
    relational thinking *versus* the lone genius 5:466
Cultural program 15:693–694
Cultural property 16:147, 19:155
  authenticity 2:253
Cultural provinces 7:110
Cultural psychiatry 5:471–477
  bereavement 5:475
  conflict and shock 5:476
  definition 5:475
  Gough's social contract 5:474
  identity 5:475
  illness *versus* disease 5:473–474
  interpretations 5:472–473
  Jung's archetype 5:471–472
  medically unexplained symptoms and somatization 5:472
  mode of culture and individual 5:472
  personalized approach 5:476
  prevention and management strategies 5:476
  refugee populations 5:474–475
  stigma and recruitment 5:472
  stress of migration 5:474
Cultural psychology 5:478–483, 5:563, 19:668, 25:756
  Bruner on 2:875–876, 2:878
  cognitive styles 5:564–565
  versus cross-cultural psychology 5:481
  cultural neuroscience 5:566–567
  custom complex 5:479–480
    genital surgeries in Africa 5:480
  definition of 5:478
  emotions 5:565–566
  indigenous psychology 5:482
  mental states 5:478–479
  and morality 15:807
  motivation 5:566
  versus national character studies 5:481–482
  perceptions 5:564–565
  psychic unity 5:479, 5:481
  recent research in 5:312
  self-processes 5:564
  self-reflexive social construction 5:478–479
  social attribution 5:564
  tenet of psychic pluralism 5:479
  theoretical position 5:312
  Vygotsky on 25:330–331
  and worldview 25:754–756
Cultural purification 11:451
Cultural racism 19:855, 19:864–865, 25:796–797
Cultural radicalism 9:633
Cultural realm, and ultramodern shadow 5:248
Cultural reinterpretation, in Caribbean 1:639
Cultural relativism 1:769, 3:420–421, 3:677, 5:382–384, 11:357, 11:362, 16:329, 20:234
  American school of anthropology 5:484
  anthropology of 5:484–487
  anti–anti relativism 5:485–486
  antifoundationalism 5:484
  Boasian 1:766–767
  and ethics 1:737
  versus ethnocentrism 1:765–766
  foundationalism 5:484
  and multiculturalism 16:29
  perspectivalism 5:484, 5:486
  positioned subject 5:485
  pragmatism 5:484
  relational understanding 5:485
  and social stratification 22:343
  social work 24:755
  types of 5:486
Cultural repatriation 2:602
Cultural repertoire, in cultural sociology 5:618
Cultural reproduction 5:376, 5:378–379, 18:721, 22:331
  arts socialization 2:8, 2:7f
Cultural resistance, and hegemony 10:743, 10:748
Cultural resource management (CRM) 5:488–490

Cultural resource management (CRM) (*continued*)
laws 5:489–490
legal protection 5:488–489
preservation 5:488–490
public support and education 5:490
repatriation 5:489–490
requirements 5:489–490
tourism and 5:488, 5:490
value of cultural resources 5:488
Cultural resources
tacit coordination using 18:422
Cultural Revolution (China) 3:516, 5:596, 25:816
Cultural revolutions 5:77
Cultural rights 3:674, 5:491–497, 12:956, 16:23–24
anthropological jurisprudence 5:494–496
challenges for future 5:496
formal guarantees in legal systems 5:494–496
hierarchy problem 5:495–496
immigrants 5:493–494
multiculturalism 5:494
pluralism 5:492, 5:496
scholarship on 5:493–494
Cultural rites of death, and sexual risk behaviors 21:799
Cultural safety 11:785
Cultural sciences 11:94
Cultural script 7:594
Cultural segregation 11:451
Cultural selection, for evolution 5:404–405
Cultural self-defense 25:797
Cultural sensitivity, and racialized groups 22:822–823
Cultural settings and gender role 21:684
Cultural shaping, of emotions 7:535
Cultural significance of age 1:334
Cultural socialization 8:130
Cultural sociology 1:523, 1:526, 5:498–503, 5:603, 5:617–621
Baroja on 3:150
classical precursors of 5:498–499
classical theory 5:552–553
contemporary variants of 5:500–501
cultural studies 5:501–502
cultural turn 5:500–502
debate in 5:552–554
embodied approach 5:552–553
externalization in 5:550–554
hermeneutics 5:500–501
and institutional logics 5:554
institutionalization of 5:501–502
internalization in 5:550, 5:553–554
interpretivism 5:498, 5:500
meaning 5:499–502
poststructuralism 5:498, 5:500–501
semiotics 5:499–502
strong program in 1:525–526
structuralism 5:498–501
symbolism 5:498–502
theoretical foundations 5:499–500
Cultural spaces
creation in China 5:598–600
modeling 22:336
Cultural specificity of emotions 7:459
Cultural stereotypes
families as educational settings 8:734
role in academic motivation 9:680
Cultural stimuli 22:331
Cultural structure and social structure, difference between 22:724–725
Cultural studies 1:933–937, 2:361, 23:640
African Studies 1:276–279
British 2:852–855
and cultural sociology 5:501–502
deconstructionism 1:935–936
and Foucault 14:200
Frankfurt School 9:385
and globalization 1:936–937
history of 1:933–934
linguistic turn 1:934–936
media audiences 2:216–217
new approaches to 1:934–936
postcolonial studies 1:935, 1:937
poststructuralism 1:935
of science 5:504–506
Soviet Union 20:820–824
conservatism 20:820–821
gulag 20:820–823
intellectuals 20:820
Marxism 20:820–821, 20:823
memory and mourning 20:823–824
nationalities 20:821–822
origins 20:820–821
radicalism 20:820–821
torture 20:822–823
violence 20:822–823
structuralism 1:935
Cultural studies of science 20:230–231
articulation of 5:505
community-defining consensus 5:504
emergence of 5:504
internalists and social constructivists, debates between 5:504
naturalistic (descriptive) approaches 5:505
normative and naturalistic (descriptive) approaches, debates between 5:505
place within science studies 5:504–505
science as historically evolving practices 5:504
science as justified knowledge 5:505–506
shared commitments and norms of scientific communities 5:504
themes 5:505–506
Cultural substance, tradition as 24:503–504
Cultural surround, reactive processes in 3:451–452
Cultural Survival (nonprofit organization) 1:223–224
Cultural symbols, and meaning 5:579–580
Cultural syndromes 5:571, 5:622
Ataque de Nervios ("Attack of Nerves") 5:624
bound to a specific culture 5:623
Brain Fag Syndrome ('Mental Fatigue') 5:624–625
Dhat or Semen-Loss Anxiety Syndrome 5:627–628
Dhat Syndrome ('Semen Loss') 5:625–626
in the Diagnostic Manuals (DSM and ICD) 5:623–624
DSM 5 5:623–624
DSM-IV-TR 5:623
ICD-10 5:623
future of 5:628
globalization and 5:628
historical and contemporary status of 5:622–623
impact of urbanization 5:628
Khyâl Cap ('Wind Attacks') 5:626
Nervios ('Nerves') 5:626
Shenjing shuairuo ('Weakness of the Nervous System') 5:626
short history of 5:622
society and 5:622–623
stability of 5:623, 5:626–627
Susto ('Fright') 5:626
Taijin Kyofusho ('Interpersonal Fear Disorder') 5:626
Windigo 5:627
Cultural systems 9:532
commemoration as 4:240–241
definition 1:737
religious symbols as 23:861
and social system 22:749–750
technological change and 24:105
Cultural taxonomies 4:285
Cultural terrain 11:392
Cultural themes of pilgrimage 18:146–148
Cultural theories
and hegemony 10:744
kinship 13:53
of Mary Douglas 20:709–710
tourism 24:471
Cultural theorizing (sociology) 22:960–961
strong program 22:961
weak program 22:960–961
Cultural therapy 5:616
Cultural time 24:407
Cultural toughness
definition of 5:335
Cultural tourism 24:466, 24:468
Cultural traditions 24:506
of regional geography 20:122
and social change 22:149
Cultural traits, and cultural evolution 5:401
in statistical taxonomies, operationalizing 9:2
Cultural transfer 4:409, 20:426
and Renaissance 20:426
Cultural translation 1:774–775
Cultural transmission 5:390–392, 5:401–404, 8:357
coevolution 8:362
conformist transmission 5:403
cultural evolution 8:357–360, 8:362
in education 1:730
in ethnic intermarriage 8:118, 8:121
group effects and conformity 5:403–404
horizontal and oblique transmission 5:403
innovation 5:404
iterated learning 8:357–360
and language development 13:322

language transmission and evolution of linguistic structure 8:357–361
language use, negotiation, and the formation of conventions 8:361–362
many-to-one transmission 5:403
neural basis of 5:391
one-to-many transmission 5:403
vertical transmission 5:402–403
Cultural trials, witchcraft 25:604–605
Cultural turn 2:884, 10:85, 10:87
in advertising 1:216–217
class consciousness 3:742–743
and community sociology 4:365–366
new orientations after 1945 11:76
of places 18:149–150
and unobtrusive measures 24:771–772
Cultural unity 5:64
Cultural universals 1:765–766
of kinship terminologies 13:62
problem of, and taxonomic groups 11:816
Cultural values 17:824, 25:27, 25:29–31
in Caribbean 1:640
definition 1:737
pay-for-performance 17:615
and work-related behaviors 17:824–825
Cultural variables 5:312
Cultural variations 1:765–766, 4:48, 4:534–535, 7:448–449
and adolescent social competence 22:218–220
in intentionality of children 12:351
in life phases 5:507–515
and symbolism 1:769
Theory of Mind 24:269–272
Cultural violence 17:622
Cultural wealth of nations 5:539
Cultural-contextual structures 12:404
Cultural-historical geography 11:19–20
Cultural-historical psychology, Luria on 14:397
Cultural-historical revival 20:325
Cultural-historical theory 25:319–320
Culturalism 12:237
Culturalist racism 5:518
Culturally appropriate interventions
in social work 5:516–522
antiracism 5:516–518
childcare 5:519
children 5:517–521
ethnicity 5:516–517, 5:519–520
future directions 5:520–521
historical backdrop 5:516–517
indigenous peoples 5:516–518, 5:520
interventions 5:519–520
professional application of cultural appropriateness 5:518–519
race, racism, and culture 5:517–518
Culturally diverse social and managerial systems 24:105–107
Culturally diverse societies, policies and norms in 11:679–680
Culturally reflexive model (CRM) 20:85–86
Culturally relevant pedagogies 8:756
Culturally responsive teaching 8:756
Culturally sensitive management practices 11:784
Culturally unidentifiable human remains (CUHR) 1:888
Culture, production of
boundary formation 5:606
in cognate areas 5:605–606
autoproduction 5:605–606
reception 5:605–606
collective memory 5:606
comparative analysis 5:604–605
comparisons across symbol-producing realms 5:604
cultural mobility 5:605
high *versus* low forms 5:604–605
prestige, changing criteria of 5:605
and field theory 9:140–141, 9:144–147
institutional levels of analysis 5:605
linking consumption with 22:336
micro levels of analysis 5:605
production perspective 5:602–603
production system in cultural diamond 5:603–604
prospects 5:606
and social structure 5:602–603
societal levels of analysis 5:605
tradition 5:606
Culture analysis, of corporate culture 4:923–924
Culture and personality School (United States) 7:145, 12:357, 16:233–235, 22:527–528
Culture area 11:436
Culture change, ethnohistory 8:192, 8:195
Culture contact
frontier 9:467
and parenting 17:496–497
Culture distance 5:334–335
Culture history, and ethnicity 8:103
Culture in nonhuman organisms 22:401–408
Culture industry 3:274
Frankfurt School 9:381–382
Culture jamming, and advertising 1:217
Culture learning theory 1:70
Culture maintenance, minority 8:145
Culture of circulation, and cultural studies 1:937
Culture of deprivation *see* cultural deficit theory
Culture of gangs 9:594–595
Culture of knowledge 22:229–230
Culture of narcissism 17:439
Culture of peace 25:843, 25:846
Culture of poverty 17:23, 22:873, 24:775
see also Cultural deficit theory
Culture of settlement 4:250
Culture of unity 20:323
Culture perspective, cultural analysis 9:363–364
Culture shock 5:334, 5:614–616
definition of 5:335t
epistemology of knowing other cultures 5:615–616
and mental illness 5:616
migration 5:616
multiculturalism 5:614
phases 5:615
reverse 5:615
Culture wars 3:542, 5:78, 20:281
gay/lesbian movements 9:634–635
Culture-bound idea, work as 25:688
Culture-bound psychogenic psychoses 5:622
Culture-Brain-Behavior Interaction model 16:600
Culture-formation, and public sphere 3:705
Culture-general training 5:337–338
Culture-historical approach, memory 1:883
Culture-historical archaeology 1:901, 1:903–904
Culture-ideology of consumerism 3:101–102
Culture-led regeneration 24:921–922
Culture/power/history orientation 15:688–689
Cultures 1:774, 2:293, 2:883, 3:719–720, 5:312, 5:388, 5:551, 5:563, 5:617–621, 6:748, 8:139–140, 9:357, 9:432, 12:3, 13:726, 16:29, 16:326–329, 17:231–233, 19:818, 20:328, 21:964–965, 22:881, 23:46, 23:435, 24:292–293, 25:322–323
acquisition in childhood 3:451–453, 3:455–456
and actor–network theory 5:523–527
adolescent prosocial behavior 19:223–224
and adoption 1:139–140
and aesthetic education 1:229–232
aesthetics 5:727–729
African Studies 1:276–279
anthropology 1:746, 5:617–618, 5:620
and art markets 14:567
artists and producers of 5:731–732
arts 5:618, 5:727–729, 5:732–733
and attachment research 2:140–142
early pitfalls 2:140
mother–infant relationships 2:140–141
outlook 2:141
and Australian aboriginal society
see also Australian aboriginals
authority and 2:273
and avant-garde art 2:320–324
-based differences
in justice reasoning 15:783–785, 15:784t
in transition from adolescence to adulthood 24:576–582
and bear subculture 13:891
behavioral inhibition across *see* Behavioral inhibition (BI)
belief 2:493
Benhabib on 2:519–523
and biology 6:10
and body 2:729–731, 2:733, 2:739
boundary making 5:618–619
Bourdieu's practice theory 5:553–554
and bureaucracy 2:914–915
of Caribbean 1:640–641
challenges in social stratification and class 22:341–342

Cultures (*continued*)
Bourdieu's class and cultural distinction 22:342
class as social action, experience, and culture 22:341
class experience from production to consumption 22:341–342
change 4:926
child and adolescent psychiatry 3:384
childcare and development across 3:395–398
and children's rights 3:485–490
cities types 24:802
and civilizations, clash of 20:403–404
claims of nurture 16:341
classification 5:728–729
cultural divisions by genre 5:729
models of cultural industries 5:728–729
process models 5:728
clothing 9:742
and cognition *see* Cognition; and culture
and collective beliefs 4:165–166
and commodities 5:109–110
as common (shared) knowledge 4:17
and communication 4:284–285
collectivism 4:284
high- and low-context communication 4:284–285
power distance 4:284
uncertainty avoidance 4:285
comparative method and social process 5:618–619
components 22:164–165
concept of 13:726–727
conflict 5:476
and conflicts 16:427–428
consumption and reception of 5:732–733
cultural participation studies 5:732–733, 5:732t
time-use and household expenditure surveys 5:733
contemporary views 5:582–589
cultural developmentalism and 'first-world's' burden 5:585–586
cultural models 5:583–585
deconstruction 5:584
ethnographic representation 5:584–585
fault lines 5:583–585
globalization 5:586–587
identity politics 5:584
Kroeber's and Kluckhohn's prediction 5:582–583
multiculturalism and the problem of difference 5:586
neo-positivism 5:584–585
popular objections and common misattributions 5:587–588
relativism 5:582, 5:584–585, 5:587
romantic pluralism 5:585
romanticism 5:585
skeptical postmodernism 5:584
standard view in North American anthropology 5:583
in contestation 9:3–4
as context 17:615–616
context to alcohol consumption 1:517
corporate *see* Corporate culture
creation 4:923, 4:926
creative industries 5:727–728, 5:730–731
cross-cultural similarities/differences in prosocial moral reasoning 7:318
cross-cultural study of education 5:327, 5:329, 5:331
cross-national comparison 5:730, 5:733
cultural industries 5:730–731
dancing 5:667–669
databases 5:727, 5:732
definitions 4:924, 5:471, 5:535, 5:582–583, 5:727–728, 9:450, 10:567, 10:747, 12:360, 23:637, 23:643, 23:645
Taylor's 1:766
and democracy 6:94
description 4:925–926
desirability of happiness 23:643–644
as determinant of mental health 5:570–575
conceptualization of culture 5:570
coping 5:571–572
cross-cultural awareness 5:574
cross-cultural perception and expression of mental illness 5:572
cross-cultural psychology 5:570
Diagnostic and Statistical Manual of Mental Disorders 5:571
etiological beliefs and folk approaches 5:572–574
impact of ignoring culture in therapy 5:574
nondominant ethnic minorities 5:570
non-Western healing practices 5:572–573
Western psychotherapy 5:572–573
worldviews 5:570, 5:572–574
development of self-regulation during early childhood across 21:553–556
developmental research across 6:326–331
and de-Westernization 25:547–549
differences in 23:644–646
gross domestic product (GDP) 23:644
happiness, definitions of 23:645
happiness, strategies in 23:645
individualism-collectivism 23:644
materialism 23:645
measurement 23:645–646
norms regarding happiness 23:644–645
positive emotions 23:644–645
differences in autobiographical memory 2:286
and diffusion 6:402–404
Dilthey on 6:425–428
and disgust 6:548
dissemination of 5:732
and doctor–patient interaction 6:615–616
and drinking 6:658–662
dual process theory 5:553
and early friendships 9:450–452
conflict 9:451–452
cross-cultural research approaches 9:450–451
versus familial/community relationships 9:451
intimacy 9:451
psychosocial outcomes 9:452
research gaps 9:452
shared activities 9:451
and early memory development 6:760
direct assessment method 6:761–763
indirect assessment method 6:760–761
East African studies 6:776
East Asian studies 6:784–787
Eastern European studies 6:805–809
and economic development 5:528–534
economic sociology 7:67–70
economics of 2:1–6
aesthetics 2:1, 2:4–5
art history 2:1, 2:4
of arts 5:731
copyright 2:1–3, 2:5
cultural and creative industries 2:5
of culture 5:728–731
heritage 2:4–5
and economy
see also Economy
effect on infant development 18:857
and embodiment *see* Embodiment and culture
embodiment and 2:720–721
and emotions 5:542–549, 7:513–517, 13:465
enmeshed relationship with health 5:622
and environmental movements 7:765
epistemic *see* Epistemic cultures
and essentialism 8:14
and ethnicity 8:102–104
and ethnocentricism 8:166–167
and ethnocentrism/prejudice 8:169–171, 8:173
ethnographic/clinical research method 4:925
and ethnography 8:185
evaluation 5:618–619
experience of happiness 23:644
of expertise, relations between 8:590–591
explanation for female genital mutation 8:880–881
exposure to other 4:207
expression of emotions 13:291
externalization in 5:550–554
factors in gender development 9:743–746
differential treatment of girls and boys 9:744
expectations of parents 9:743
gender socialization and role of parents 9:743
initiation rituals 9:745–746
role of mothers and parents 9:743–744
siblings and peers 9:745
task assignment 9:744–745
and fertility, anthropology of 9:1–5
and fertility transition 9:49–54
and field theory 9:142–143, 9:145–146
and finance 9:171, 9:173–175
five-factor model of personality across 9:249–253
in flux 9:3–4

and food preferences 9:299
food/cuisine 9:290–292, 9:294–295
foundations of early behavioral adjustment 6:15–16
gender 9:742
gender roles and stereotypes 9:747
and generation 9:853–855
global history 10:162–167
and globalization 10:262–263
globalization of 23:640
historical sociology 11:28–29
and historiography 11:100–101
and honor killing 11:185–190
households 11:228–231
human development 11:272
for human rights 11:363
and Hutterite fertility 11:445–446
as ideas 9:2–3
and imitation by infants/young children 11:622, 11:626
implications 5:567–568
  challenges 5:567–568
  contributions 5:568
and individuality 17:933–934
industrial geography 11:857–858
and inequality 23:851–852
influence of 5:622
influence on prosocial behavior 19:247
influences on interpersonal relationships 5:426–432
and institutional logics 5:550–556
and institution-building
  Eisenstadt's contribution 7:321
and intelligence 12:318–320
as interaction 9:50–51
and intercultural communication 12:357–360, 12:358t
internalization in 5:550, 5:552
and Internet
see also Internet
and interpretation 12:642–643
interpretation of term 5:491–492
issues, of Latinos 9:628–629
joint attention across
  during infancy and early childhood 12:844–850
key words 13:289
kinship and 13:49–51
knowledge system as grammar of 4:17
and language 13:288–292
and law 13:550–551
and leatherman subculture 13:889
legal 13:536–537, 13:727–728
Lévi-Strauss on 13:904–908
of LGBTI commercial sex 13:882–883
like me theory 15:133
and literacy 14:236–237, 14:239
Lizardo's procedural theory 5:553
at local/global confluence in Japan 12:795–797
logic of miracles *see* Miracles
and magic 14:438
and materiality 14:765
mathematical concepts development during early childhood across 14:772–784
mental representations of health and illness across 10:567–570
as mental schemes 9:2–3
microevolution 15:340–341
in Middle East 16:381–385
in Middle East and North Africa (MENA) 16:361
as moral demands system 2:273
motivational development across *see* Motives; negotiation of
multiculturalism 16:22–27
music as 16:149
narrative networks 16:212
national 16:262–263, 16:265
Native American tricksters 24:649–652
versus nature 1:765–766
negotiation of, and nostalgia 17:36
and negotiations 16:424
new concept of 7:175–176
notion of 7:875–876
of Oceania 17:160
oral 17:275–277
organizational *see* Organizational culture
and organizational conflict 4:639, 4:641
and organizational groups 10:411–412
and organizations
see also Organizations
paradox of free time 13:840–842
and parenting 17:493–494, 19:246–247, 22:159
pastoral 17:582
peasant 17:676–677
performance 5:618, 5:620
and personality 19:359–360, 19:362
and personality anthropologists 16:237
personality development and temperament 17:878
personality-religion links 17:804–805
perspectives of technology and economies 11:863–865
pluralism and tolerance 18:241
political culture 5:617, 5:619
politics of 9:283–284
and postwar sociology 5:617, 5:619–620
of poverty *see* Poverty
power and 18:337
as practice 5:618–619
and practices
  indigenous people 11:806
and pretend play 18:870–872
  in Brazilian communities 18:871
  in geographically distinct communities 18:871–872
in prison life 18:924–925, 18:927–928
problem of meaning 5:617–618
and production 15:716
professional 19:101–102, 19:104–105
profiling approaches 4:925
and prosocial behavior 19:238–240, 19:242
and prosocial development 19:230–231
  caring for others 19:231
  framing cultural issues 19:230
  modesty and politeness 19:231
  prosocialty at home 19:230–231
  race and ethnicity 19:231
  sharing 19:231
and psychosocial work 19:477–480
publishing 19:627
and punishment 19:659
and racialized groups 22:821–822
research and science as 20:509
as research specialty 5:557
resilience 20:558–559
right to *see* Right to culture
and risk 20:709–710
risk in gay/lesbian tourism 9:641–642
as ritual and interaction 5:619–620
role
  in cognitive development 4:55
  in disasters 10:548, 10:551–552
  in identity formation in adolescence 9:434–435
role of adolescence across 9:432–433
and role of gatekeeper 9:620–621
Rorty's contributions 20:761–762
rural geography 20:796
as schemas, attitudes, values, and beliefs 9:51–52
and secrecy 21:374–375
and self 5:563–569
self and identity development during adolescence across 21:432–438
self in
see also Self
and semantic differential 21:565, 21:568–569
and semiotics 21:594–595
and senses 21:617–619
and sexuality 21:819–824
as shared tradition 20:107
and small group research 22:105
and social and behavioral sciences 15:321–323
and social differentiation 6:393–394
and social history 22:308, 22:310–312
and social networks
see also Social networks
and social psychology 22:525–530
social science of 5:618
and social structure 22:714
social withdrawal across *see* Social withdrawal (SW)
sociology of 13:113, 13:116
sociology of arts 5:728
Southeast Asia 11:78–80
Soviet 20:853
-specific psychiatric disorders 5:622
statistics 5:727
  EU Statistical System of Culture 5:730
  transnational perspective 5:729–730
  UNESCO Framework for Cultural Statistics 5:729–730
and street children 23:533–538
'studies of culture at a distance' idea 16:235
and subjective wellbeing 23:643–647
and subjective well-being 23:652
symbols 5:552–553
taste 5:729
technology and environment 1:769

Cultures (*continued*)
touristic 24:467
and transference 24:545–548
transition to adulthood 24:573
transmission 9:747–748
and urban places 24:929
use of term 17:23
and value of children 3:500
and values 25:36–38, 25:41, 25:44–45
values development across
see also Values
variations
of defiance during adolescence 6:7–11
in prosocial responding 19:236
variations, and childhood friendship 9:444–446
classrooms and extracurricular activities 9:445
ideology 9:446
similarity 9:444–445
time with peers/family 9:445
working children 9:445–446
and vocabulary 13:288–289
in voting 25:305–306
world culture 10:225
youth 25:824
Cultures of evidence 7:579
Cultures of relatedness
kinship and difference in 9:427–431
Cultures research 5:542
definition 5:542
and emotion practices 5:545–547
emotional frequency 5:545–546
emotional intensity 5:545–546
rewards 5:547
role of concern salience 5:546–547
role of situations 5:546
Culture-sensitive theory of play 18:224
Culture-specific basic emotions 5:544
Culture-specific startle disorders 23:333
Culture-specific syndromes 6:826–827
Culture-specific training 5:337–338
Cult-watching groups 16:808
Cummins, R., on functional explanation 9:519–520
Cumulative advantage 1:331, 6:755
and artists 2:55–56
Cumulative causation 15:411
and migration 15:469
Cumulative continuity in marriage 17:798–799
Cumulative cultural evolution
gene-culture coevolution 9:816–817
Cumulative cultural evolution, origin of 5:391
Cumulative customer satisfaction 5:631–632
Cumulative disadvantage 1:331
Cumulative distribution function (CDF) 5:903, 6:575, 11:886, 19:29–30
Cumulative emissions 3:861–862
Cumulative evolution, of technology 24:132–133
Cumulative learning 21:48
Cumulative logit model 16:133
Cumulative prospect theory 2:446–447
Cumulative scale, in social measurement 21:6
Cumulative scaling, in ethnography 8:189–190
Cumulative voting 7:350, 19:206
Cunningham, G.B., on lesbian athletes 9:659
CUPE *see* Canadian Union of Public Employees (CUPE)
Curaca 23:47
Curation, social media 22:423–425
Curative factors 10:430
Curative rape 21:748
Curatorial consumer 5:418
Cure, complementary and alternative medicine 10:564–565
Curiositas 20:285
Curiosity 17:231
and health 7:498
Curious grapevine 11:363
Curle, James Herbert, works, *Our Testing Time: Will the White Race Win Through?* 19:803
Currency, anthropology of money 15:746–749
Currency crises 9:183–184, 9:189
Current counterterrorism (CT) policy 16:247
Current fact situation (CFS) 13:787
Current Impact Index 20:499
Current Population Survey (CPS) 5:692–693, 5:695, 5:747–748, 5:799, 8:119, 8:121, 10:303, 20:884
health and retirement study 5:747
living standards measurement study 5:747–748
Luxembourg income study 5:747
Current status data
in event–history analysis 14:353
Curricular validity 4:775
Curriculum 3:178–179, 21:101–102
behavioral/analytical 3:178
factual material 3:178
and law 13:723
legal rules 13:723
LGBTQ-inclusive, in schools 21:88
personal 3:178–179
social work 13:723
social work education 22:776–777
tangible user interfaces (TUI) in 24:24
Curriculum development 15:81
psychiatric education 7:185–187
Curriculum dimensions, gifted education 7:159–160
Curse of dimensionality 21:599–600
Cursus publicus 4:274–275, 18:640–642
Curtailed bureaucracies 2:922
Curtesy 19:173
Curtis, Edward 18:83
Curve fitting approach, for population forecasts 18:593
Curves
functional data analysis 9:514–516, 9:518
graphics 10:359–360
survival 23:777
Curvilinear contingency fits 23:611
Curvilinear relation, graphics 10:358
Cushing, Frank Hamilton 2:71
Cushing's syndrome 7:610, 17:134
Custodial role 19:527
Custody
alternatives to 19:44–45
and family law 8:780–784
Custom complex 5:479–480
genital surgeries in Africa 5:480
Customary international law 12:509
Customary law 3:670, 13:759, 13:915, 24:504
Africa 1:272
procedure 5:227
Customary legality 13:472
Customary practices 13:469
Customary rules of deliberation 9:620
Customer behavior, and customer satisfaction 5:632
Customer discrimination
and beauty 2:385
and labor market outcomes 17:74
Customer engagement behavior 4:140
Customer relationship, and customer satisfaction 5:632
Customer satisfaction 5:630
ACSI model 5:630–631, 5:631f
attribution theory 5:631
cumulative 5:631–632
customer behavior 5:632
customer relationship 5:632
disconfirmation of expectations model 5:630–631, 5:631f
marketing 5:630–632
stakeholder management 5:630–632
transaction-specific 5:630–631
value 5:630–631
Customers *see* Consumers
Customization challenges, in supply chain management 23:711
Customized craft production, persistence of 5:111
Customs 4:247–248
and authority 2:270–271
ethnopsychology 25:793
and taboo 24:1
Customs unions 12:574, 12:576, 12:578
Cutoff rule 5:907
Cuttica, C., works, *History of European Ideas* 12:255
Cutting-edge neuroscience 13:494
Cutting-edge research on legal mobilization 13:594–596
Cuvier, Georges 5:681–682, 8:384, 20:286–287, 21:241
works
The Animal Kingdom 12:175
Rapport historique sur les progrès des sciences naturelles depuis 21:241
CV *see* Consonant–vowel (CV) syllables; Contingent valuation (CV)
CVA *see* Cerebrovascular accident (CVA)
CVAR *see* Cointegrated vector autoregressive (CVAR) model

CVD *see* Cardiovascular disease (CVD)
CVE *see* Countering violent extremism (CVE) programs
CVF *see* Competing values framework (CVF)
CVLT-II *see* California Verbal Learning Test, Second Edition (CVLT-II)
CVM *see* Civic voluntarism model (CVM)
CVOs *see* Circumventricular organs (CVOs)
CVR *see* Cardiovascular reactivity (CVR)
CVTS *see* Continuing Vocational Training Survey (CVTS)
CW *see* Continuous wave (CW)
CWB *see* Counterproductive work behavior (CWB)
CWB-C *see* Counterproductive Work Behavior Checklist (CWB-C)
CWC *see* Center on Wrongful Convictions (CWC)
Cyber bulling 23:466
Cyber crime 5:633
  costs of 5:634–635
  definition 5:633
  example 5:636–637
  identity theft 5:633–637
  legislation 5:635–636
  prevalence of 5:634
  types of 5:633–634
  victims and offenders 5:635
Cyber Islamic Environments (CIEs) 16:778
Cyberactivism in Egypt 16:801–802
Cyberanthropology 12:600
Cyberbullying 5:638–644
  categories 5:639
  consequences of 5:642–643
  defining 5:638–640
    issues with 5:638–639
  males *versus* females 5:641
  media coverage 5:638, 5:642
  prevalence rates 5:640–642
  prevention and intervention 5:643
  research on 5:643
  taxonomies 5:639t
  and traditional bullying, comparison 5:639–640
  variations
    by age 5:641
    by gender 5:641
    by race/ethnicity 5:641
    by rural/urban samples 5:641–642
  victims and perpetrators 5:642
Cyberchase 15:7
Cybercitizens 9:726
Cybercrimes 5:192, 5:235, 12:595
Cyberculture 11:549–550, 12:600–601
  and religion, relationship between 20:276
Cybergogy
  tangible user interfaces (TUI) 24:23, 24:23f
Cyberhenge: Modern Pagans on the Internet 20:275
Cyberinfrastructure 22:602, 24:418–419
  in Canada 22:602
  challenge and promise of 22:602
  definition 22:602
  in USA 22:602
Cyber-infrastructure 21:189–190, 21:189f
Cybering *see* Cybersex
Cyberlove 21:813–814
Cybernetic feedback model of self-regulation 21:559–561
Cybernetic hierarchy, and AGIL scheme 1:385–387
Cybernetics 3:895, 5:645–650, 9:567–568, 18:103–104, 21:190, 22:912–913
  actor-network theory 5:645–646
  agency 5:645–648
  art 5:647–648
  brain 5:646
  and Cold War 4:124
  complexity theory 5:647
  counterculture 5:646
  cyborgs 5:645–646, 5:648
  enframing 5:649
  environment 5:647
  interdisciplinarity 5:645
  management 5:646–647
  mangle of practice 5:645–646
  modernity 5:649
  nonmodern ontology 5:645
  poiesis 5:649
  politics 5:648–649
  posthumanism 5:645–647
  of purposive actions 15:958–959
  robotics 5:646–648
  science and technology studies 5:645–646
  technocracy 5:648
  technologies of self 5:647–648
  viable system model (VSM) 5:646–647
Cybersecurity 12:595
Cybersex 21:813–814
Cyberspace 4:185–186, 4:812, 12:84, 12:88, 12:92, 12:600, 17:714, 21:813, 23:84–85
  and gender 9:709, 9:712–713
Cybersquatters 3:31
Cybersquatting TM 12:269–270
Cyberterrorism 24:230
Cybertime 24:408
Cyberwalk treadmill 25:133
Cyberwar *see* Cyberterrorism
Cyborg anthropology 21:182
Cyborg-optimism 2:731
Cyborgs 5:645–646, 5:648, 12:600
Cycles
  fashion 8:838
  population 18:559–561
  recognition by indigenous peoples 11:804–805
  of reflection 7:579
  time series 24:331–336
    band-pass filters 24:334–335
    census X-12 seasonal adjustment filter 24:334
    first difference filter 24:334
    lag filter 24:334
    relationship between spectral density and autocovariances 24:332
    spectra of commonly used stochastic processes 24:335
    spectral estimation 24:335–336
    spectral properties of moving average filters 24:333
    spectral representation of covariance stationary process 24:332
    spectrum of building permits 24:333, 24:333f
Cyclic adenosine monophosphate (cAMP) 9:881, 16:753, 24:533–534
  response element-binding (CREB) 9:881
Cyclical chronology 24:411
Cyclical industrial dynamics 5:651–654
  business cycles 5:651–653
  decomposition 5:651–652, 5:652f
  explanations 5:652–654
  Fourier analysis 5:652
  identification and measurement 5:651–652
  implications 5:654
  industry life cycles 5:651
  time series analysis in the frequency domain 5:652
  time series analysis in the time domain 5:652
Cyclical poor 18:729
Cyclical time 24:409
Cyclical unemployment 24:738, 24:741–742
Cyclicity-based approach to syntax 16:4
Cyclist culture 20:396
Cyclothymia 2:683–684, 14:484–485
CYD *see* Community youth development (CYD)
Cynicism 2:256, 18:419–420, 21:59
CYP19 *see* Aromatase cytochrome P450 (CYP19)
CYPF *see* Children, Young Persons and Their Families (CYPF) Act 1989
Cyprus, diglossia in 6:421
Cytoarchitecture of neocortex 3:340f
Cytopathic lentiviruses 1:488–489
Cytosine methylation 7:833–834, 7:839
Czarniawska, Barbara, on institutions 17:392
Czech Greens 10:380–381
Czechoslovakia 3:316, 16:259, 18:545–546

D

D measures, in implicit association test 11:704
D scale 2:257–258, 2:260, 15:567
Da Orta, Garcia 21:184–185
DAA *see* Disability Awareness in Action (DAA)
Dabashi, H., works, *Staging a Revolution* 16:383
DAC *see* Development Assistance Committee (DAC)
Dacin, M. Tina 17:393
Dacin, Peter A. 17:393
Dadaism 2:321–322
Dade County referendum in 1977 9:634
Dadrian, V.N., works, *A Typology of Genocide* 10:7
DADT *see* Don't Ask, Don't Tell (DADT)
Daedalus 2:508

Daejoseon Ilbon Yuhaksaeng Chinmokhoe 24:585
DAG *see* Directed acyclic graphs (DAG)
Daguerre, Louis Jacques Mandé 18:81, 20:420–421
al-Dahik, Abu Musab 24:233
Dahl, Robert A. 4:363, 5:655–660
  analyses 5:659
  behavioral approach to politics study 5:656–659
  behaviorism of 18:766
  on democracy 5:655–660
  on elites 7:392–393
  on influence 5:658–659
  on oppositions 5:658
  personal life 5:655–656
  pluralism of 5:656, 5:659
  on political economy 5:655–657
  on political inequality 5:658–659
  on political power 18:760–764
  on polyarchies 5:657–658
  on power 5:656–660
  works 5:656–659
    Politics, Economics and Welfare 18:233–234
    Who Governs? 24:943
Dahrendorf, Ralf 5:661–665, 9:539
  academia 5:661–662
  class 5:662–665
  consensus 5:662–663, 5:665
  on democracy 5:663–665
  on elites 7:392–393
  on entitlements 5:663–665
  on Erasmus intellectuals 5:664
  on functionalism 9:539–540
  on inequality 5:662
  liberalism of 5:665
  of liberty 5:661, 5:663–664
  life and career 5:661–662
  of life chances 5:661, 5:663–665
  pluralism of 5:661–662
  political career of 5:661–662
  on provisions 5:663–665
  on sociology of conflict 5:662–664
  on totalitarianism 5:662–664
  on work/labor 25:678
Dai Nihonshi 11:49–50
Dai Viet Su Ky Toan Thu 11:50
Daily diary data 7:507–508
Daily mobility 15:636–637
Daily stressors 22:159–160
DAIP *see* Duluth Domestic Abuse Intervention Project (DAIP)
Dalai Lama 2:889
Dalits, Indian cast 3:225–226, 22:44
Dalton, M., works, *Men who Manage* 2:914, 14:475
Daly, Mary 9:461
DALYs *see* Disability-adjusted life years (DALYs)
Damage caps 14:267
Damanhur 16:806
Damascus, cultural heritage in 5:416
Damasio, Antonio 1:741–742
  somatic marker hypothesis 12:617
Damaska, Mirjan 13:804
Dame comedians 13:875
Damnatio memoriae 11:63
Damocles syndrome 3:459
Damron's Men's Travel Guide 9:638
Dams 12:724
Dance
  anthropology of 5:666–670
    choreography 5:666–667
    dance event 5:666–667
    dancing culture 5:667–669
    ethnicity, nationalism, and transnationality 5:668
    mediation and preservation 5:668–669
    movement 5:667–669
    moving bodies, gender, and sexuality 5:667–668
    notation 5:668–669
    postcolonialism 5:667–668
  carnival 3:147–148
  evolutionary group dynamics 8:444
  movements 5:667–669
Dance event 5:666–667
Dancesport 5:668
Dandridge v. Williams (1970) 18:736
Dandy, Walter 16:686
Dangers
  and cardiac defense 23:539–540, 23:542–544
  involved in online dating 17:215
  liminality 14:133–136
  sex as 10:813–814, 10:816–817
  of unincorporated dead 5:872
Dani people, peace/nonviolence 17:626–627, 17:629
Daniel, Herbert 13:862
Daniel, Marc 21:750
Daniels, R.V. 20:857
Danish Psychological Association, establishment 3:874
Danish Radical Party 10:383
Danish Red–Green Alliance 10:383
Danish stød 15:828–829
D'Annunzio, Gabriele 16:257
Dansk Folkeparti (Danish People's Party) 18:415
Dantas-Saraiva-Cotegipe Law (Brazil) 22:41
Dante Alighieri, works, *Inferno* 17:429
DANVA *see* Diagnostic Analysis of Nonverbal Accuracy (DANVA)
Danwei system 21:734–735
Danziger, Kurt 13:913
Dao Duy Anh, works, *Short History of Vietnamese Culture* 11:85
Daoism 2:889
DAP *see* Deutsche Arbeiterpartei (DAP); Domestic Abuse Program (DAP); Draw-A-Person (DAP) test
Dapper, Olfert, works, *Description of Africa* 9:105
DAPs *see* Data analysis plans (DAPs)
DAP: PED *see* Draw a Person: Screening Procedure for Emotional Disturbance (DAP: PED)
DARE *see* Drug Abuse Resistance Education (DARE) program
Dark Ages, personality psychology in 17:908
Dark Alliance 3:613
Dark figure 5:713
  of crime 5:165
Dark Middle Ages 25:604–605
Dark play, ideas of 18:217
Dark tourism 4:185–186
Dark Triad of personality 5:671–675, 11:170–171, 17:830–831
  costs and benefits 5:671–672
    Machiavellianism 5:672
    narcissism 5:671–672
    psychopathy 5:672
  evolutionary psychology 5:673–674
  limitations 5:675
  measurement and theoretical issues 5:673–675
    life history theory 5:674
    niche-specialization hypothesis 5:674
    overall 5:674–675
    parasites and hosts 5:674
    social exchange theory 5:674
  overlap and distinctiveness 5:672–673
    resolution 5:673
    separationism 5:673
    unificationism 5:672–673
Darkness at El Dorado affair 25:34–35
Darnton, Robert 1:913, 11:110, 18:918
Darrah, C., works, *Busier Than Ever* 24:401
Darsan 21:618
Dartmouth College v. Woodward 13:755
Darwin, Charles Robert 1:704, 4:339, 5:676–682, 5:684, 5:686–687, 7:467–468, 7:723, 7:792, 8:253, 12:307, 13:70, 14:366–367, 14:642–643, 14:811, 19:124, 19:412, 20:287
  on adaptation 1:85
  on attitude 2:175
  behavioral research of 5:679
  on cultural evolution 5:388–389
  developmental theories of 8:384–385
    Malthus and population 8:384–385
    scientific influences 8:384
  doctrine of sexual selection 19:887
  on emotional expression 7:457–458, 7:467, 7:458f
  on evolutionary demography 8:406–407
  evolutionary models
    structure of 8:469
  evolutionary models of 14:621–622
    structure of 8:469
  and Fisher 9:222
  on instinct 12:175–178
  on irrationality 12:721–722
    ecological rationality 12:722
    informational format 12:722
    preference structure 12:722
  legacy 7:467
  life 5:676–677
  neo-Darwinian approach to modern human cognition 11:252–253
  on ontogeny of behavior 5:679

principle of differential natural selection 2:41
relevance of theories 5:679–680
scientific contribution of 5:677–679
series of books 5:676–677
theories of 8:233
evolution 4:721, 13:135–136, 21:173–174, 21:679
genealogical imagery 9:807
heredity 19:802
natural selection 17:17, 17:607, 19:887
works
Autobiography 3:155
The Descent of Man, and Selection in Relation to Sex 5:676–679, 7:457, 8:238–239, 12:176, 12:178, 20:287
The Expression of the Emotions in Man and Animals 1:237, 2:175, 7:457, 8:667, 18:83, 21:424
Inquiries into Human Faculty and Its Development 8:237
Natural Selection 5:676
On the Origin of Species 1:766, 5:676, 5:678, 5:681, 7:750, 7:792, 8:238–239, 8:385, 8:406, 12:176, 19:802, 20:122, 20:287, 23:274–275
Darwin, Erasmus 20:287
Darwin, John 11:685, 11:688–689
Darwin, Leonard
and Fisher 9:222–223
Darwin Awards 11:416–417
Darwinian Demon 8:451
Darwinian fitness 8:388, 17:173
use of term 1:86
Darwinian medicine *see* Evolutionary medicine (EM)
Darwinian revival 22:955–956
Darwinian selection 16:332
Darwinians 5:684
Darwinism 5:681–685, 14:366–367, 14:369, 21:238, 22:909
adaptation 5:683–684
in biology 8:385
history of life 8:385
natural selection 8:385
consilience of inductions 5:681
evolution 5:681–684
evolutionary economics 8:431–432
evolutionary medicine 5:683–684
Huxley's program 5:682
Islamic fundamentalism 12:742–744
and Lamarckism 8:383
meanings of 5:684
Mendelian genetics 5:682
modern 8:386–387
later developments 8:387
synthesis 8:386
natural selection 5:682
pre-Darwinian evolutionism 8:383–384
creation and design 8:383–384
evolution in the enlightenment 8:384
fossil record 8:384
revival of 22:955–956
social 5:684–691, 8:386
sociobiology 5:683
synthetic theory 5:682
Darwinism, generalized 8:431–432
Darwin's theorem 1:86–87
development of 8:384–385
Malthus and population 8:384–385
scientific influences 8:384
Darwin–Wallace papers 8:385
Das, J.P. 12:308
Das Gupta, J., works, *Language Problems of Developing Nations* 13:353
Dasgupta, Surendranath 7:374
Dashanzi International Art Festival 5:599
DASI *see* Dynamic approach to school improvement (DASI)
DAS-II *see* Differential Ability Scale–2nd edition (DAS-II)
Data 21:190
adoption 1:137–138
applications of quantitative memory models to 15:164
censuses
accuracy assessment 3:293–294
dissemination 3:294
processing 3:294
definition 5:830
environmental sociology 7:801–802
health policy 10:663
income 11:739–740
macroeconomic 5:751–757
narrative networks
coding 16:216
sources 16:216
time and 16:216
probability density estimation 19:29–32
urban violence 24:956
Data acquisition, functional magnetic resonance imaging (fMRI) 9:525–526
Data Act 1973 (Sweden) 18:943
Data adjustment and correction 6:126
age heaping 6:129
aggregate-level data 6:128–135
adjustments over time 6:134–135
age patterns 6:133–134
completeness 6:128–129
parity data 6:129
population distribution by age 6:129–133
census 6:126–135, 6:129f, 6:131f–133f
consistent correction 6:130, 6:134–135
digit preference 6:127–130, 6:133f
individual-level data 6:127–128
intercensal cohort analysis 6:128–129
item nonresponses 6:127–128
missing data 6:128
postenumeration survey 6:127–128
robust time series 6:134–135
types of errors 6:126–127
errors of content 6:127
errors of coverage 6:126–127
sampling errors 6:126
Data analysis 5:1
applications 23:417
content validity studies 4:775–776
evolving standards 4:776
functional data analysis 9:514–518
qualitative developmental psychology 19:672–674
discursive psychology and discourse analysis 19:673
grounded theory methodology 19:673
interpretive phenomenological analysis 19:673
metaphor analysis 19:673
narrative psychology and narrative analysis 19:673
performative social science 19:674
qualitative content analysis 19:672–673
video analysis 19:673–674
race and ethnicity 19:814–815
sequential experiments 21:635–636
Data analysis plans (DAPs) 16:82
Data analytics, behavioral analytics 24:107
Data archives 5:791
survey research 23:748–749
Data assessment
qualitative developmental psychology 19:670–674
autoethnography 19:672
group discussions and focus groups 19:671
interviews 19:671
participant observation and ethnography 19:672
participatory action research 19:672
qualitative experiments 19:671
visual data 19:671–672
written data 19:671
Data augmentation, in event–history analysis 14:353
Data centers, linguistic 5:816
Data collection and analysis (Europe) 7:558–560
administrative use 7:559
German historical school in economics 7:560
statistical use 7:559–560
Data collection methods, time use 24:392
diary supplements (IT) 24:392–393
internet and mobile apps 24:393
smartphone apps 24:393
web-based data collection 24:393
wearable cameras 24:393
Data collections 5:707, 21:634
and digital technology 12:604
functional approach 5:707
Genographic Project 10:24–25
growth curve analysis 10:441–443
institutional approach 5:707
interviewing 5:851–856
issues, in adoption 1:137–138
for job analysis 12:813–814
interview methods 12:813
observation methods 12:813–814
meta-analysis 23:909–910
privacy/confidentiality in social research 18:943–944
systems, required for health information systems 5:786
Data communication 12:493

Data compression, order statistics 17:294
Data curating 5:819–820
Data development, remote sensing 10:66
Data enclaves 18:945
Data evaluation, stochastic optimization 23:474
Data fabrication 21:313–314, 21:316
  and objective research 17:93–94
Data generating process (DGP) 7:1
Data generation, stochastic optimization 23:474
Data harmonization 3:294
  comparative databases 5:795
Data heterogeneity, meta-analysis 15:268–269
Data in multidimensional scaling 16:42–43
Data infrastructures 23:130
  comparative databases 5:791–792
  for electoral behavior 5:824–826
  for political behavior 5:824–826
Data integration 5:784, 7:793
Data issues, family and household demography 8:726
Data logging and modeling, Computer Assisted Instruction (CAI) 4:495
Data management 7:793, 14:187–188
  personalized medicine 17:927–928
Data matrix
  network text analysis 4:770t
  semantic text analysis 4:771, 4:771t
Data mining 23:730
  in EITMs 18:454–455
  journals 12:867
  technological change and 24:107
Data modeling, in cartography 3:170
Data protection, on Internet 12:594–597
Data protection agencies 12:596
  and privacy
    direct-to-consumer (DTC) personal genetic testing 6:444
Data providers, respect and trust for 10:303
Data quality 4:577, 6:140–141, 20:239, 20:246
  audit 4:577–578
  in citizen science projects 3:633
  and mortality crossover 15:849–850
  sample selection by 6:141
  surveys and polls 23:756
Data saturation 9:276
Data selection, archiving, protecting integrity of 1:917
Data sets
  multiply-imputed, analyzing 16:89
    inferences for scalar Q 16:89
    significance tests for multicomponent estimands 16:89
  statistical assumptions 8:10
Data sharing 5:790, 5:819–820
  about therapeutic effectiveness 24:283
  consent, archiving 1:918
  in social research 18:944–945
Data snooper, statistical disclosure limitation 4:577–579
Data stewardship 18:942
Data swapping
  probabilities elicitation 7:382
  statistical disclosure limitation 4:579
Data transformations 16:35–36
Data utility 4:578
Data warehouses 12:72–73
Databases 5:790, 5:817–823, 5:830–835, 5:837, 5:839, 5:841, 22:596, 22:601
  American National Election Studies 5:831
  applied social research 5:692–699
  archives 5:817–820, 5:831–832
    Inter-University Consortium for Political and Social Research 5:831–832
    necessity 5:831
    other European data archives 5:832
    Zentralarchiv für Empirische Sozialforschung 5:832
  archives and historical databases 5:700–703
  big data 5:822
  census data 5:817–818, 5:820–822
  classification 5:820–821
    cross-sectional micro-level demographic surveys 5:820
    cross-sectional micro-level social surveys 5:820–821
    longitudinal micro-level surveys 5:821
    macro-level social and economic data 5:820
  communications and media 5:704–711
  comparative research 5:832
  construction and usage, ongoing challenges in 5:821–822
  core 5:844–850
    international comparative research 5:789–797
  crime 5:712–716
  crime and delinquency 5:717–719
  cross-national data 5:817, 5:819–821
  culture 5:727, 5:732
  definition 5:830
  demography 5:735–741, 6:179
  development during and after Second World War 5:830
  economic 5:742–750
  education 5:763–774
  Genographic Project 10:25
  health 5:781–788
  influence on European social science 5:831
  Institute for Social Research 5:830–831
  legal 5:803–810
    citation index 5:809
    citator 5:806, 5:808
    commercial vendors 5:803
    dockets 5:808
    headnotes 5:803–805
    organizing statutes 5:807–808
    platforms providing legislative history 5:808
    secondary sources 5:808–809
    United States 5:803
    value of databases in case law 5:803–807
  linguistic *see* Linguistic databases
  longitudinal data, historiographic dimension of 5:832–833
  mechanisms behind change 5:818–820
    advances in computer technology 5:819
    gold-standard research 5:819
    increased ethos of data sharing and curating 5:819–820
  Miller on 5:831
  as modern phenomenon 5:817–818
  perspectives 5:834
  secondary analysis 5:817, 5:819–822, 5:831
  in social sciences 5:833–834
    barometer series 5:833–834
    European Social Survey 5:834
    European/World Values Surveys 5:834
    GSS and ISSP 5:833
    international comparisons of national election studies 5:833
    reasons for abundance 5:833
  and statistical systems (crime)
    international crime data 5:714
    national incident-based reporting system (NIBRS) 5:713
    official crime data 5:712–713
    outcome of data 5:716
    police data on crime 5:713–714
    self-report studies 5:715–716
    victimization surveys 5:714–715
  structural contingency theory 23:613
  studies of democratic quality 5:833
  survey research 5:831–834, 23:750–751
Databases, core (demography) 5:735
  administrative data 5:737
  census data 5:735–736
  demographic rates 5:739–740
    availability of 5:739
    comparability of 5:739
    period and cohort measures 5:739–740
    structure of 5:739
  microcensus data 5:736
  population register 5:736–737
  social science surveys 5:737–739
    big data 5:738–739
    prospective surveys 5:738
    retrospective surveys 5:737–738
Data-driven approach to scientific discovery 21:292
Data-driven testing, smooth testing 23:401–402
Data-gathering methods
  economic statistics 5:758–760
    business records 5:759–760
    directly measured statistics and derived statistics 5:758–759
    household surveys 5:760
    measurement 5:759–760
    problems of defining business 5:759
    public accounts and administrative records 5:760
Data-intensive science, journals 12:867
Dataism 22:987
Data-model fit 8:688–689
Data-oriented design 4:490
Datasets, land-use 13:241
DATATEXT 23:377–378
Date rape 19:889

Dating 11:61
  modern knowledge societies 13:107–108
Dating methods
  in archaeology 3:562–567
    chemical, temperature, and water affected dating methods 3:567
    isotopic 3:565–566
    radiation 3:566–567
    sidereal 3:564–565
Dating of arrival of humans 2:229–231
Dating recalled instances 19:773–774
Dating systems, in archaeology 23:744
Dativus sympaticus-constructions 18:632
Datum 14:285
Daubert v. Merrell Dow Pharmaceuticals, Inc. 8:583–584, 13:767
D'Augelli, A.R., on mental health of LGBTQ youth 21:87
Daughter of Bilitis (DOB) 9:633
Daughter relation 13:54
Davenport, Charles B. 8:224
David, Florence Nightingale 9:626
David, René 13:804
David Arnold, works, *The Tropics and the Traveling Gaze* 7:752–753
Davidoff, Paul 18:204
  advocacy planning 1:220–221
Davidson, D., on meaning 14:843
Davidson, Thomas 10:740
Davies, Norman, works, *God's Playground* 6:813–814
Davis, Allison 4:362
Davis, Gerald, on economic interest 5:537
Davis, Jefferson 4:574
Davis, Jerome, works, *An Introduction to Sociology* 7:588–589
Davis, Kenneth Culp 13:767–768
Davis, L.E., works, *American Economic Growth* 7:56
Davis, Mike 24:819
Davis, Natalie Zemon 4:152, 11:110
  works, *Martin Guerre* 2:645
Davis, Robert 16:809
Davis, William Morris 16:335
Davis-Delano, L., on apologetic behavior in sports 9:657–658
Davy, Georges 10:139
Davy, Humphry 21:226–227
  on chemistry 3:372–374
DAW *see* Division for the Advancement of Women (DAW)
Dawes Act 13:233
Dawkins, Richard 1:735
  works, *The Selfish Gene* 22:899, 22:901–902
DAWN *see* Diabetes Attitudes, Wishes and Needs (DAWN) study
Dawson, William 20:287
Dax, Marc 18:46–47
Day, Graham 4:363
Day care 2:793
  influence on infant development 18:856
Day reconstruction method (DRM) 10:589, 23:648
Day schools 11:764
Dayan, Daniel
  on media events 15:48–49, 15:51–52
  works
    Media Events: The Live Broadcasting of History 15:48, 15:51
    Owning the Olympics 15:52
Dayananda Saraswati 20:374
Daydreaming 15:7
Day-night task 8:492f
DAYTOP (Drug Addicts Treated On Probation) 24:276
D'Azeglio, Luigi Taparelli 22:386
DBH *see* Dopamine-ß-hydroxylase (DBH)
DBO (desires, beliefs, opportunities) theory 19:903
DBS *see* Deep brain stimulations (DBS)
DBT *see* Dialectical behavior therapy (DBT)
DC *see* Delivered costs (DC)
DCC *see* Deleted in colon cancer (DCC)
DCCS *see* Dimensional change card sort (DCCS)
DCD *see* Dominant Language Constellation (DCD)
DCE *see* Discrete choice experiments (DCE)
DCM *see* Dual Control Model (DCM) of sexuality
DCPR *see* Diagnostic Criteria for Psychosomatic Research (DCPR)
DCS *see* D-cycloserine (DCS)
DCT *see* Dimensional comparison theory (DCT)
D-cycloserine (DCS) 3:769, 17:125
  for post-traumatic stress disorder 8:609
DD *see* Delay discounting (DD)
DD-FRs *see* Duration at destination-specific fertility rates (DD-FRs)
d-dimensional feature vector 23:411
DDM *see* Death distribution method (DDM); Dynamic decision making (DDM)
DDR *see* Disarmament, demobilization, and reintegration (DDR) programs
DDT *see* Dichlorodiphenyltrichloroethane (DDT)
de Azpilcueta, Martin Navarro 14:407
de Bonald, Louis Gabriel 4:686
de Brosses, Charles 17:158
  works
    Grundrb der Geschichte aller Religionen 20:323–324
    Du Culte des Dieux Fétiches 9:106
de Coulanges, Fustel
  works
    De Tocqueville's L'Ancien régime et la Révolution Française (*The Old Regime and the French Revolution*) 11:75
    La Cité antique (*The Ancient City*) 11:75
De facto judicial independence 12:893
De Facto population estimates, applied demography 1:841
De facto power relationships, in production 15:715
De facto property rights 24:530
De facto racial subordination 19:808–809
De Finetti, Bruno 5:857
  adherent mass distributions 5:859
  coherency 5:857
  exchangeability theorem 5:857–859
  finite additivity 5:858–859
  scoring rule 20:928
De Gaulle, Charles 21:369, 21:371
De Genova, N., on immigrants in North America 17:24
De Graaf, Nan Dirk
  on cultural capital 5:378
  on cultural participation 22:333–334
  on parental reading 22:333
de Graff, J., works, *Take Back Your Time* 24:401
De jure desegregation 6:218–219
De jure judicial independence 12:892–893
De Jure population 1:841
De jure racial segregation 5:267
De jure racial subordination 19:808
De jure segregation 6:218–219
  racial segregation 5:267
  racial subordination 19:808
de la Blache, Paul Vidal 7:723
de las Casas, Bartolmé, on indigenous rights 11:795
de Maistre, Joseph 4:687
De Moivre, Abraham 23:424
de Morga, Antonio, works, *Sucesos de las Islas Filipinas* 11:83
De motu cordis 21:881
De Mundo 10:69
De natura locorum 7:721
de Nooy, Wouter, on artistic prestige 5:468
de Penafort, Raymond 15:579
de Prony, Gaspard, on division of labor 6:603–604
De Rienzi, Grégoire-Louis Domeny 17:158
De Sade, Marquis, works, *120 Days of Sodom* 17:948
De Saussure, Ferdinand 6:504
de Sauvages, Boissier 17:27–28
De Sica, Vittorio, works, *Bicycle Thieves* 9:168
de Soto, Hernando 1:633–634
de Tracy, Destutt 11:573
de Vattel, Emmerich
  on indigeneous rights 11:796
  works
    Droit de gens 3:771–772
de Victoria, Francisco, on indigenous rights 11:795–796
De Vigo, Giovanni, works, *Practica Copiosa in Arte Chirurgia* 16:681
De Vos, Pierre, on sexual minorities in South Africa 21:749
de Vries, Matthias, works, *Woordenboek der Nederlandsche Taal* 13:939
de Waal, Frans 19:887–888
  works, *Good-Natured* 22:905
Deacon, Terence W., works, *Incomplete Nature: How Mind Emerged from Matter* 25:344
Dead bodies, treatment of 9:948–949
Dead hand issue 13:742
Dead zones, irrigation 12:726
Deaf community 21:942

Deaf individuals, language lateralization in 12:694
Deafferentation 10:502
  phantom limbs 18:2, 18:5
Deafness 21:934–941
  joint attention in children with 12:848
Dealer–critic system 14:564
Dealignment, partisan 17:565–566
Deal-making 16:418–419
Deal-making lawyers 16:418–419
Dean, Howard 6:66
Deary, Ian 12:304–305
Death 5:739–740, 5:860, 18:174
  absorbing demographic events 5:740
  of ancestors 1:688–689
  anthropology of 5:870–873
    abortion 5:872
    bones 5:871–872
    celebrations 5:870
    cemeteries 5:872–873
    changes in mortuary ritual 5:872–873
    controversies over beginning and end of life 5:872
    danger of unincorporated dead 5:872
    emotional response of mourners 5:872
    as event *versus* process 5:871
    fieldwork dilemmas 5:870–871
    future directions 5:873
    ghosts 5:872
    good death *versus* bad death 5:872–873
    Hertz's analysis of first and second funerals 5:871
    as journey or rite of passage 5:871–872
    mourning 5:870–872
    parallels to other life cycle rituals 5:872
    ritual, centrality of 5:870–871
    rituals and core values 5:871
    social status 5:870
    spirits 5:870–872
  bereavement 2:531–536
  causes, in adult mortality 1:175–178
  causes of 12:21–24, 15:857, 15:860–861
  and childhood cancer 3:461–462
  and disgust 6:546–547
  by hazing 10:557–558
  medical causes 12:18–19
  and philosophy 18:210
  psychology of 5:860–864
  and religiosity of women 25:636–637
  sociology of 5:865–869
    afterlife 5:867–868
    ancestors 5:865
    awareness 5:866
    bereavement 5:868
    communication and control 5:867
    communication technology 5:868–869
    connection with living 5:865
    denial 5:866
    disenfranchisement 5:868
    funerals 5:867–868
    grief 5:865–866, 5:868
    internet 5:865, 5:868
    medicalization 5:865–866
    mourning 5:865–866, 5:868
    postmaterialism 5:867–868
    professionalization 5:866
    remembrance and collective memory 5:868
    ritual 5:865–866, 5:868
    secularization 5:866
    sequestration 5:866
    social death 5:867
    trajectories 5:867
  Strauss on 23:528
  university hazing 10:558–559
  and war 16:304–308
  wish for 12:755
Death and dying 7:595–597, 7:599
Death and dying, psychology of
  attitudes 5:860–861
    adolescent 5:861
    adulthood 5:861
    childhood 5:860–861
  coping with 5:862
  cultural considerations 5:861–862
  overview 5:860
Death attitudes 5:860–861
  adolescent 5:861
  adulthood 5:861
  childhood 5:860–861
Death distribution method (DDM) 1:154
Death intensity 14:89–90
Death of distance 12:91, 14:290
Death penalty 5:187, 5:874–878, 12:909, 13:768
  abolition of 5:875
  for atrocious crimes 5:875
  ban on 3:689
  in China 5:876–877
  countries retaining 5:874
  and crime policy 5:180–181
  historical context 5:874–875
  justifications for 5:874
  reach, retention, and rejection themes 5:874–875
  retaining 5:875–877
  in United States 5:876
  and wrongful convictions 25:786, 25:789
Death Penalty Information Center (DPIC) 25:787–788
Death rates 6:119–121
  Jewish 12:803, 12:807–810
DEB *see* Disordered eating behaviors (DEB)
Debasement, and financial crisis 9:183–185
Debasilectalization 18:136–137
Debo, A., on Native Americans 1:635
Debord, G., works, *The Society of the Spectacle* 5:245
Debriefing, in military/disaster psychiatry 15:498
Debt 2:353–356, 20:16
  economic psychology 5:888
  and household wealth 25:451, 25:453–454
  study of 5:888
Debt burden, and structural adjustment 23:600–601
Debt crises 9:184–185, 9:187–189
Debt enforcement 19:177–178
Debt trade 10:633–634
Debtor
  see also Bankruptcy
Debtor–creditor relationship 4:941
Debtors Act 1869 19:178
Debunking views, morality 8:71
Decalage, cognitive development 18:121–122
Decapacitation 9:808
Decarbonization 7:631
Decategorization model 12:411
Deceleration, multiple language 16:59
Decentered policy network analysis 12:772
Decentering 5:320
Decentralization 18:182, 18:677, 20:564, 22:265, 22:267, 24:823–824, 24:894–899
  comparative health systems 4:400–401
  of decision-making 21:103
  in educational planning 7:262–263
  of power 8:875–877
  public sector organizations 19:598–600
Decentralized and spontaneous social mechanisms 6:602
Decentralized policymaking 7:267
Decentralized systems 10:304–305
Deception 17:533–534
  experimental economics 8:548
  and online dating 17:215
  in research 20:517–518
De-Christianization 9:409
Decidability 14:319–320
Decimal system 16:399
Decimation of Native Americans 1:634–635
Decision analysis, DAG models for 2:365
Decision and choice 5:879
  bounded rationality 5:882–884
    asymptotic rationality 5:884
    minimax regret 5:883–884
    survival and evolution 5:884
  costly rationality 5:881–882
  heuristics 5:883
  Luce's choice axiom 5:895–900
  non-Bayesian learning 5:883
  satisficing 5:883
  Savage paradigm 5:880–881
    extended Savage paradigm 5:881–882
  Simon critique 5:881
  uncertainty 5:879–880
Decision behavior 17:782
  primitives 17:780
Decision boundaries 23:413
Decision domains, risk-taking 17:809–810
Decision field theory, mathematical psychology 14:805
Decision maker (DM) 5:906, 9:834, 10:620–621
  valuation of the terminal stock 17:253
Decision making 1:344, 2:371–372, 4:681, 5:911, 6:85–86, 13:461–462, 13:464, 18:288, 18:290, 19:320, 19:322–324, 19:584
  algorithms 1:531–533
  aspiration level theories 5:915
  and bounded rationality 2:773–775
  collective 6:86

concept of 20:927
and conflict 4:584–585
consumer psychology 4:738–741
in democracy 6:85
diffusion 6:413–414
dynamic *see* Dynamic decision making
ecological rationality *versus* internal consistency 5:912–913
robustness 5:913
structures of environments 5:912–913
economic, and genetics 9:952
and emotions 7:461–463, 7:473, 7:479
and environmental cognition 7:709, 7:711
in equality 6:87
and ergonomics 11:300
ethical 19:581
ethical, and public policy 19:580–581
in financial investments 15:738
in game theory 9:587–589
garbage can model 1:105
group 11:302
by groups 10:426
and health literacy 10:653–655
heuristics, classes of 5:913–915
one-reason decision-making 5:914–915
recognition-based decision-making 5:914
trade-off heuristics 5:915
historical background 5:911
impairment 19:452
influences of emotion regulation on 7:454
and informed consent 12:111–112
and interoception 12:617
Damasio's somatic marker hypothesis 12:617
irrationality in 12:719–721
framing 12:720
Public Goods Game and Costly Punishment 12:721
Ultimatum Game 12:720–721
judicial 12:885–889
managerial and organizational cognition 14:480–481
mathematical psychology 14:813–814
moral 13:465
moral *versus* selfish 16:612
muddling through view 1:105
neuroesthetics 16:661–662
nonrational theories 5:911–916
normative *versus* descriptive theories 5:912
offender 20:776
optimizing *versus* nonoptimizing theories 5:911–912
organizational 17:352–357
organizational process model 1:105
participatory 25:719–720
political 6:85–86, 18:429–430
political process model 1:105
practices, and ignorance 11:581–582
privacy 18:948
processes 17:515
and procrastination 19:77
psychology of 5:917–922
affect and emotion 5:921–922
descriptive approaches 5:918–919
heuristics 5:919–921
intuition *versus* deliberation and reasoning 5:921
normative theories 5:917–918
prospect theory 5:918–919, 5:919f
rational choice theory 5:917–919
rules 5:919–920
subjective expected utility theory 5:917–919
violations and anomalies 5:918
questionnaires 19:776, 19:779
and rational choice theory 19:893
rational model of choice 1:105
and risk 20:686, 20:689
social work ethics 22:785
systems, civil service 3:698
Tversky on 24:679–681
and uncertainty 24:718–720
and unintended consequences 17:420, 17:422
use of heuristics in 4:108
war 25:403
see also Judicial decision making
Decision Oriented Instructional Design (DO-ID) 12:198–199
see also 4C/ID model of instructional design
Decision rules, heuristics 5:913
Decision science 21:158
Decision strategy 21:949, 21:954–955
Decision studies (D studies) 9:835–837
Decision support systems 24:764–765
Decision theory 8:521, 16:430, 18:779–780
of actions 12:337
counterfactual reasoning in 5:97
of environmental decisions 7:710
prisoner's dilemma 18:933–934
and trust 24:669
Decision theory approaches
for quality of life measurement 10:659–660
EQ-5D 10:659
Health Utilities Index 10:659
Quality of Well-being scale 10:660
Decisional capacity 19:321–322
components of 19:321
Decisional roles, managers 14:473–474
Decision-bound models, concept learning 4:539
Decision-forcing cases 3:176–177, 3:180
examples 3:178
Decisionism 21:167
Decisions 23:239
corporate finance 4:928
diffusion and random walk processes 6:395–401
economic psychology 5:886–891
random utility models 5:901–905
types of 9:835
Decision-theoretic auction models 2:206
Decision-theory language 16:117
Decivilization 10:485
Declaration of Alma-Ata 4:401–402
Declaration of Barbados 1:224
Declaration of Helsinki 13:559, 20:514–515
Declaration of Human Rights 13:216–217, 24:442
Declaration of Independence 4:248
Declaration of Religious Freedom (*Dignitatis Humanae*) 24:604
Declaration on the Right to Development (DRD) 11:357
Declarations, Searle on 21:341, 21:344
Declarative knowledge 7:878, 11:556, 12:326, 13:708–709
Declarative memories 1:644–645, 5:923–926, 15:147, 22:63
consolidation, and sleep 22:66
and early memory development 6:759–760, 6:763
neural basis of 5:923–926
proactive interference in 19:1–3
in long-term episodic memory 19:3
in semantic memory 19:3
retroactive interference in 20:613–614
Declarative pointing 10:271–272
Declarative sentence 14:293–294
Declension 15:823–824
Decline
class 3:747–749
phenomena of 22:266–267
of religion thesis 21:385–386
of small states 22:111
Declivitous items, scaling of 21:6
deCODE genetics, Inc 1:760, 2:604, 9:808–809, 10:39
DeCODEme scans 1:761
Decoding 3:511–512
Decollectivization 13:234
Decolonial thought, and modernity 15:694
Decolonialization 16:268–269
Decolonization 4:223–226, 14:531–532, 16:301, 18:659–661
and human development 6:307–308, 6:310
of human sciences 25:548–549
and imperialism 11:684
and indigenous social work 11:807
modern India after 11:80
in Oceania 17:162
and postcoloniality 18:658–661
scientific 25:548
in Southeast Asia 23:72
Decommodification 11:234
welfare state 25:477
Decomposition analysis, for discrimination 18:830–831
Decomposition strategies, questionnaires estimation 19:774
Decompositions 5:651–652, 11:886, 11:888, 5:652f
class 3:747–749
inequality 11:906
object, survey questions 23:766
social differentiation as 6:391–392
Deconcentrating poverty 24:931
Deconnection phenomena 23:294
Deconstruction 10:93, 16:196

Deconstruction (*continued*)
beyond philosophy and literature 5:929–930
and critical hermeneutics 10:787
in critical social work 5:290–291
culture 5:584
deconstructive architecture 5:930
ethics 5:931–932
feminism 5:930
gender studies 5:930
law 5:931–932
and literary studies 5:928–929
and philosophy 5:927–928
politics 5:931–932
and postmodernism 18:680, 18:682
and psychoanalysis 5:930
queer theory 5:930
religion and theology 5:930–931
Deconstruction theory 21:530
Deconstructionism 18:12
and cultural studies 1:935–936
Deconstructionist critiques 12:647
Deconstructive approach
to gender 9:790–793
Deconstructive listening 4:726
Deconstructivism 22:668
Decorative art 14:566–567
Decorative morphology 11:514
Decoupling 16:764
see also Divorce
Decoupling of norms 22:982–983
of economic activities 14:761, 14:763
Decrementalism 2:902
Decrements 14:91
Decreolization 18:136–137
DECRG *see* Development Research Group (DECRG)
Decriminalization 13:818
of homosexuality in China 21:735–737
Decriminalization debate 6:674
cost–benefit analysis 6:676–678
current status of 6:675
harm principle 6:675–676
Décroissance 6:24–25
Dedekind completeness 17:300
Dedication 17:334, 17:336–337
school engagement 21:59–60
Dedoose.com, software package 22:545
Deduction and abstract theory 19:704
Deductive disclosure 18:945–946
Deductive nomological model 4:24
Deductive qualitative analysis (DQA) 20:502–506
Deductive reasoning 3:669, 5:933–940, 12:318, 12:646, 13:784–786, 14:306, 19:61–63
abstractions 5:933–935
formal languages 5:933–934
inferential relationship 5:934
syntactic inference 5:934–935
applications 5:938–939
hardware and software verification 5:938
knowledge systems 5:939
mathematical theorem proving 5:938
software and hardware synthesis 5:939
first-order logic 5:933–937
issues in automated deduction 5:935–938
artificial neural systems 5:938
clausal form, resolution, tableaux, and connection method 5:935–936
complexity, systems, and performance 5:937–938
first-order features 5:936
variants, specializations, and extensions 5:936–937
and legal reasoning 13:776–777, 13:780, 13:782
legal reasoning as deduction 13:784–785
logical calculus 5:934
mechanization 5:939
philosophy 5:939–940
program synthesis 5:939
proof search 5:935–936
propositional logic 5:933–935, 5:937–938
psychology 5:939–940
unification 5:936
Deductive-nomological explanations 8:594–595, 10:773
Deep agraphia 6:735
Deep alexia 6:732
Deep alphabetic systems 23:265
Deep Blue (IBM game machine) 2:33
Deep brain stimulations (DBS) 7:853, 16:665, 16:718, 23:20
for epilepsy 7:853, 7:855
obsessive-compulsive disorder 17:125, 17:269
Deep dysgraphia 1:453
Deep ecology 7:765, 23:813
Deep encoding, effects on recognition memory 20:41
Deep harmony 10:531, 10:533
Deep neural network (DNN) 23:262
Deep personality traits 17:801
Deep structures (generative grammar) 9:873
of history 1:708–710
Deepening commodity relations 1:459
Deeper-forces model 22:885–886
Deep-level diversity 4:739, 6:596
Deep-sea navigation 16:346
Deep-syntactic relations 6:186
Deepwater Horizon disaster 11:300, 11:302
Deepwater Horizon explosion 3:736
Deepwater Horizon oil spill 7:743–744
Deepwater oil and gas 17:167
Deese–Roediger–McDermott (DRM) list-learning task 8:711–712
Deethnocentrification 16:186
Defamation 9:394–395
criminal defamation 9:395
Internet defamation 9:395
transnational defamation litigation 9:395
Defamilialization 12:404
by crosscultural juxtaposition 11:866
Default logic 2:36
Default mode network (DMN) 11:459, 11:461, 11:468–470
Defeasible reasoning 13:785–786, 13:790
Defective translation of mental images 1:875
Defendant-rights state 19:645–646
Defended neighborhood hypothesis 10:543
Defended realism 7:883
Defense
cardiac 23:539–545
and coping 4:888, 4:902
of science in USSR 16:14–15
Defense (law)
and forensic psychologist 9:352
Defense (military)
analysts 19:576
versus deterrence 6:247
economics of 6:1–6
alternative interpretations 6:5–6
arms race 6:2, 6:5
capitalist crisis 6:5–6
externalities 6:3–5
historical contextualization of debates 6:2
imperialism 6:5–6
at mainstream margins 6:4–5
peace dividends 6:2
preliminary conceptual and methodological problems 6:1
privatization of army 6:3
problems related to statistical data sources 6:1–2
public good 6:2–3
research and development 6:3–4
spending within neoclassical mainstream 6:3–4
state 6:1–3, 6:5
strategy 6:3–5
technological spillovers 6:3–4
wars 6:1–6
Defense budget 6:1
Defense counsel 16:421
Defense mechanisms, in psychodynamic theories 22:815–816
Defense modernization 15:709–710
Defense of Marriage Act (DOMA) 6:563, 9:635, 13:618, 21:756
Defense of Marriage Act 1996 (DOMA) 21:770–771
Defenses *see* Criminal liability and Criminal defenses
Defensive hate crime 10:542
Defensive war 4:618
Defensiveness 11:454
Defensor Pacis 3:645
Deference, and conservatism 4:689
Deference democracies 2:270
Deferential review of factual determinations 1:824
Deferential wife 2:318
Deferred adaptations, evolutionary developmental psychology 8:422–423
Deferred gratification 9:198
Deferred imitation, and early memory development 6:763
Defiance
during adolescence 6:7–11
cross-national differences 6:8
as form of deviance 6:7
Defiant behavior

during infancy and early childhood 6:12–18
aggression 6:12
caregiver–infant synchrony 6:14–15
clinical implications 6:16–17
co-occurring developmental influences 6:14
cultural foundations of early behavioral adjustment 6:15–16
early deficits in social cognition 6:14
early disciplinary interactions 6:15
early verbal deficits 6:14
effortful control 6:13–17
emotion regulation 6:14, 6:16–17
interpersonal and transactional processes 6:14–15
noncompliance 6:15
parenting 6:14–16
patterns of temperament risk 6:13–14
self-regulation 6:12, 6:14–17
temper loss and oppositional defiant behavior 6:12–13
theory of mind 6:13–14
Defiant individualism 6:71–72
Defibulation 10:195
Deficient language 13:256–257
Deficit approach to educational achievement 1:731
Deficit hypothesis, gender 9:699
Deficit models 9:699, 21:255–256
of motivation in Australian Aboriginals 15:899
of public understanding of science 21:222
Defining Issues Test (DIT) 13:120, 15:783
Defining slavery 22:43–44
Definite descriptions 6:53
Definition of the situation 6:19–23
and American sociology 6:19
and Chicago School of Sociology 6:19
experimentalization 6:22
and Goffman 6:21–22
and Merton 6:21
microtheory after WWII 6:21–22
qualitative research 6:21–22
and social psychology 6:20
and Thomas 6:19–21
transdisciplinary adaptation 6:22
Definitional problems 8:95
Deflationism 18:806
Deforestation 12:717
De-Freudianization 19:329
Degeneration 6:75–76, 8:239, 8:241, 8:253, 8:255–256
Ca^{2+} in 12:713–714
Degeneration theories 8:225
Degeneration theories, Kraepelin's 13:128–129
Degenerative pathology, and activity 2:601
Deglobalization movements 11:451
Degradation 7:713
environmental 7:797–802
Degrading, torture and 24:442, 24:444–446
Degrammaticalization 10:333–334, 10:334t
Degree, measurement theory 14:868–870
Degree attainment, and educational outcomes 5:377
Degree centrality 22:462
measures 16:533–535
Degree of a node 16:573
Degree of physical incorporation 11:512
Degree of specificity, demography 6:110
Degrees and postdoctoral courses 22:628–629
Degrees of freedom (DoF) problem
cost-function-based solution to 15:959–960
force-field-based solution 15:959
Degrees of teachers 24:44
De-growth 6:24–30
alternatives strand 6:27–29
capitalism 6:27–28
degrowth alternatives 6:28–29
political 6:29
culturalist strand 6:26–27
consensus 6:26–27
limits liberate 6:27
ecological-economic strand 6:25–26
ecologically unsustainable 6:25–26
senseless 6:25
threshold 6:26
uneconomic and unjust 6:25
origins of 6:24–25
De-growth discourse, sustainable development 23:815
Dehumanization 17:68, 25:358–359
course 6:31–32
animalization 6:31
biologization 6:32
demonization 6:31–32
mechanization 6:32
objectification 6:32
dehumanizing metaphors 6:31–34
explicit, research on 6:32–33
functions 6:33–34
genocide 10:15–21
infrahumanization 6:34–35
intergroup violence 6:33–34
objectification and self-objectification 6:35–36
and prejudice 18:836
recent developments 6:36
subtle, research on 6:34–35
Dehydroepiandrosterone (DHEA) 21:761
neuroendocrinology 16:652
Deictic gesture 10:111
Deictoids 6:54–55
Deindividuation 4:152, 10:425–426
psychology of 6:38–41
aggression 6:39
collective behavior 6:38, 6:40
disinhibition 6:38–41
group norms 6:40–41
reconceptualization 6:40–41
research 6:39
self-awareness 6:38–40
social identity 6:38, 6:40–41
social influence 6:40–41
theoretical evolution 6:38–39
theoretical support 6:39–40
and tyranny 24:712
Deindustrialization 3:664, 11:873, 16:809–810, 20:566, 21:648, 24:792, 24:815–816, 24:921
of Beijing 5:598
and postindustrial societies 18:666
postsocialist cities 3:608
Deinstitutionalization 1:326, 17:393, 19:588
and Eastern European social work 22:797–798
of highbrow culture 5:449
psychiatry in low- and middle-income countries 6:42–46
abusive practices 6:42–43
Chile 6:42–43
China 6:44–45
colonialism 6:46
community health workers 6:44–45
community services 6:42, 6:44–45
Ethiopia and Egypt 6:42
family involvement 6:45–46
general hospital psychiatric admission units 6:45
India 6:45–46
institutionalization 6:42
Iran 6:43
Jamaica 6:43
mental health acts 6:42–43
mental illness 6:42–43, 6:45–46
Mexico 6:43
paucity of psychiatric resources 6:42, 6:46
primary health care 6:43–45
psychotropic medication 6:43
South Africa 6:43–44
traditional healers 6:44–46
psychiatry of 6:47–51
abuse 6:47–48
affordability of community care 6:50
alternatives to psychiatric hospital 6:49–50
antipsychotics 6:48–49
critiques 6:50–51
decline and fall of psychiatric hospital 6:48–49
rise of asylum 6:47–48
turning point 6:48
and residential child care 20:536
in residential child care 20:537–538
of traditional family 8:780
Deism 20:286
Deities 11:54
anthropomorphic 11:55
theriomorphic 11:55
Deixis 6:52–57, 8:351, 18:796
between deictic and anaphoric reference 6:53–54
deictoids 6:54–55
demonstratives 6:52–56
frames of reference 6:54
genesis of 6:56
honorifics, projective spatial relators, and modals 6:54–55
indexicality and context dependence 6:53
interaction 6:54

Deixis (*continued*)
meaning 6:52–53, 6:55
pragmatics of 6:56
semantics of 6:55–56
speech situation 6:52, 6:54
transposed deixis and text deixis 6:53–54
Dejerine, Jules Joseph 16:685, 16:685f
Dejonghe, T., works, *Sport en economie: samen in de spits* 23:318
Delaunay Triangulation 10:100
Delay discounting (DD) 16:641, 16:644
Delayed commemoration 4:239
Delayed entry, in event history analysis 8:317
Delayed repetition 18:58–59
Delayed start model 20:793
Delayed-matching-to-sample (DMTS) 15:155–156, 17:49
Delayed-non-matching-to-sample (DNMTS) 15:155–156
Delayed-return systems 11:895
Delayering 7:62
Delbrück, Berthold, works, *Einleitung* 14:216
Delegated child welfare 11:766–767
Delegation 1:344, 19:535–536
theory of 25:565–566
Delegation of power 6:58–60
bureaucrats 6:58
in civil service 3:697
future trends 6:59–60
lawmakers 6:58–59
politics 6:58–60
role of 6:58
study of 6:58–59
agency theory 6:59
early studies 6:58–59
Delegitimation 13:827–828
Deleted in colon cancer (DCC) 2:125
Deleuze, Gilles 1:237–238, 1:749, 19:750
on networks 16:567–568
Deliberate memory 21:896–897
Deliberate practice, and grit 10:399
Deliberation 1:446, 6:61–67, 12:908–909, 12:911
agonism 1:444, 1:446, 1:448
and decision making 5:921
democracy
legitimacy and 6:65
principles 6:64–65
institutions, kinds of 6:65
late-twentieth-century rediscovery 6:64–66
medieval 6:62
Middle Ages 6:66
political philosophy 6:64–65
political theory 6:61–64
polling 18:463–464
republicanism 20:479–481
risk-taking 17:811
wisdom of crowds 6:66
Deliberative approach, to business ethics 3:16–17
Deliberative behaviors 2:187–188
Deliberative cognition 7:416–418
Deliberative democracy 6:92–93, 20:479–481
Benhabib on 2:519, 2:522–523
political representation 18:444
Deliberative politics 20:479–480
Deliberative polling 6:65
Deliberative processing, in abnormal illness behavior 11:599
Deliberative test 19:63
Delincuencia 1:623
Delineation of subject matter 4:157
Delinquency 1:715, 4:836–838, 9:593, 22:410–411
anomie, strain, and subcultures 6:69–70
extensions or reformulations in 6:71–72
collective efficacy 6:71
contemporary developments 6:70–71
contraculture 6:70–72
and crime 1:310, 1:312–314, 23:496, 23:497, 23:498
cultural conflict 6:69
extensions or reformulations 6:72
cultural deviance 6:69
definition of field 6:68
differential association 6:72
and differential association 22:412–413
extensions in constructionist tradition 6:72
feminist theory 6:72
founding scholars 6:68–69
future recommendations 6:72–73
gangs 6:68–73
group delinquency 6:68–69, 6:72
labeling theory 6:69
extensions in 6:72
learning principles to 22:413
life course 6:71–72
link to criminology 6:68
Marxist theory 6:72
peer 22:412
prevention of 18:877–882
religion and 20:265
self report 6:70
self-esteem 21:524
social control theory 6:70
extensions or reformulations in 6:71
social disorganization 6:69
extensions or reformulations 6:71
social institutions 6:68, 6:71
social learning theory 6:70
social learning variables on 22:412
socialization 6:69, 6:71–73
subject matter 6:68
Delinquent subcultures 9:593, 25:819, 25:821
Delinquenti 14:424
Delirium 10:501–502, 13:360, 19:467
and alcohol 1:510
conditions causing 13:360
language processes in 13:360–367
psychomotor activity in 13:360
sleep–wake patterns 13:360
Delis–Kaplan Executive Function System (D-KEFS) 16:724
Delito y Sociedad 5:240
Delivered costs (DC) 23:511–512
Delivery
perinatal psychiatry 17:722–725
of punishment 19:634
Delivery of care, geriatrics 10:105–106
Delivery of health care, personalized medicine 17:928
Delivery of supports and services
reducing disablism in 6:450–451
disability activism 6:451
person-centered planning 6:450
self-directed and consumer-directed services 6:450
universal design 6:450–451
Delivery system 18:888
Dell, Gary
Connectionist Model 18:59–61
Interactive activation (IA) 18:45
Dell computers
business process outsourcing of 3:46
supply chain management of 23:711
Delluc, Louis, works, *Fièvre* 9:167
Deloria, Ella 1:743
Deloria, Vine, Jr., works, *Custer Died for Your Sins* 1:636
Delors Committee *see* Intergovernmental Panel on Climate Change
Delta activity, in electroencephalography 7:354, 7:357
Delta rule model 4:654–655
evaluation of 4:655
Delusional parasitosis 24:15
Delusions
of control 21:420–421
schizophrenia 21:9–10, 21:12–13
Demagicalization of relations among people 12:392
Demagoguery 14:744
Demand 1:479–480, 25:7, 25:10–12
and control 17:137
and family planning 8:787–788
housing 11:241
and innovation 12:162–163
job displacement insurance 12:825–826
Lacan on 13:214–215
social change 18:422–423
and supply, area and international studies 1:945–946
Demand for children 9:94–96, 9:98
Demand theory, ordinal revolution in 19:898
Demand–control models 17:135–136
for work-related stress 23:548
Demand-driven model 15:597–598
Demand-resource models 17:136, 21:60–61
of work–family conflict 8:760–762
Demarcation 5:271, 5:274
Dematerialization 23:799
and industrial ecology 11:846
Dembo, Tamara 19:414
Dembour, Marie-Bénédicte, works, *Culture and Rights* 11:363
De-McDonaldization 14:830
Dementia 1:437–438, 1:584–589, 1:591–592, 1:594–595, 6:75, 13:360, 18:791–792
and aging 1:398–399

and alcohol 1:511
characteristics of language disorders in 13:361t
definition 6:75–78
dementia syndrome of Alzheimer's type 6:75–77
dementia syndromes 6:75–77
frontotemporal dementia syndrome 6:77
diagnosis of Alzheimer's dementia 6:78
course of 6:78
mild dementia in old age 6:78
diseases 6:77–78
etiology 6:78
neurological diseases 6:77–78
electroencephalography for 7:361–362
geriatrics 10:105
language processes in 13:360–367
neuroimaging 14:443
neuropsychological rehabilitation for 2:817
placebo effect for 18:155
Psychiatric aspects 6:75–79
psychiatric care for 19:296–298
religion and 20:267
semantic *see* Semantic dementia
and transient global amnesia 24:563, 24:566–567
treatment, professional care, and preventions 6:78–79
Dementia care 20:690
Dementia caregiving 3:135–137
Dementia praecox 17:28, 21:9
Kraepelin on 13:129–130
Dementia syndromes 6:75–77
of Alzheimer's type 6:75–77
Dementia with Lewy bodies (DLB) 10:500–501, 13:361
language processes in 13:363
Demetriou, Andreas, analysis of developing architecture of mind 16:505
Demic diffusion 5:403
Demilitarization 15:492
military anthropology 15:487
Deming–Stephan procedure 18:589
Democide 10:6
Democracy 1:611, 2:191, 2:528, 3:106, 3:626, 3:644, 3:647, 3:674, 3:676, 4:704, 4:707, 6:25, 6:27, 6:61–67, 6:85, 7:351, 10:739–740, 11:369–370, 11:372–373, 12:368–369, 12:374, 12:915–916, 14:498, 14:678, 14:733, 17:202, 17:515, 17:620, 18:393–396, 18:399–400, 18:470, 19:603–604, 20:397, 20:633–635, 20:765, 20:768–770, 21:141–144, 22:833–834, 23:635, 25:276, 25:357, 20:635t
Addams on 1:93–95
and administrative law 1:109–110
and agonism 1:444–448
Alexander on 1:527
Arendt on 1:963
Athenian 1:698–699
workings of 6:61–62
autonomous in 6:85–86
and big man 2:571
Bobbio on 2:714–718
and citizenship 3:655–657, 3:659
and class voting 25:299–300
consensus *see* Consensus democracy
consociational *see* Consociationalism
control over military and war 15:520–521
and culture, relationship between 5:462–463
Dahl on 5:655–660
Dahrendorf on 5:663–665
and decision making 6:85
deliberative
legitimacy and 6:65
principles 6:64–65
direct *see* Direct democracy
e-democracy 6:66
effect to their opponents in wars 25:358–359
effects of wars on 25:361
and elites 7:390–393
and Enlightenment 7:664–665
and equality 7:909–910
French Revolution 9:409, 9:411
functions for 12:376
and global capitalism 3:103–104
and good governance 19:692
Hirschman on 10:892–893, 10:896–897
history of 6:80–84
age of democratic revolution 6:81
Aristotle 6:80
Athens 6:80
classical legacies 6:80–81
classical models 6:80–81
elections 6:82–84
electoral systems 6:82
forms of government 6:80–82
French revolution 6:81–82
mass politics 6:83
modernity 6:80–81
nineteenth century 6:81–82
popular politics 6:80–81
rhetoric 6:83
twentieth century 6:82–83
war 6:80, 6:82–83
and instrumental values 6:85
justification of 6:85
and law 13:435–442
versus legal rights 20:683
liberal 25:741, 25:743–744
and liberal culture 25:360–361
and liberalism 14:29–30
ambivalence 14:24
liberalism and modern representative democracy 6:63–64
and majority rule 14:446–447
market allocation 14:542
Masao on 14:635–640
media in 22:584–585
Mill on 15:524–525, 15:527–528
and modernization 15:697–698, 15:707, 15:711
Moore's contributions 15:769–772
and morality 6:85–86
and multiculturalism 16:22, 16:24, 16:26
and museums 16:146
Myrdal on 16:161–162
and nationalism 16:291, 16:294
normative theory 6:85–89
in North America 17:24–25
Olson on 17:210
parliamentary *see* Parliamentary democracy
Plato on 18:212
and pluralism 18:237
and political communication 18:347–348
political philosophy 6:64–65
and political power 6:85
and political repression 18:449
political sociology 18:472–475
political theory 6:61–64
and political trust 24:658, 24:661
and polling 18:463–464
and populism 18:611–612
and pragmatism 18:809
promotion 6:98
and public interest 19:553
as public justification 6:86–87
and public opinion 19:562, 19:564–566
and public reason 19:594–595
relative theory of 18:468
republicanism 20:478–480
republicanism and 6:62–63
and revolutions 20:644, 20:646
Sandel on 20:895–896
Schmitt on 21:30–31
science and 21:258
Skocpol on 22:34–35
social *see* Social democracy
and social justice 22:398–399
and social question 22:554–555
in South Asia 16:281–282
and telephone 24:157, 24:159
and Third World 10:188
Tocqueville on 24:421–422
Touraine's theory of 24:463–464
toward peace 25:359–361
training of aesthetic perception for 1:230
transitions 6:97–100
Arab Spring 6:98
authoritarianism 6:98
challenges 6:98–99
democratization 6:97–99
hybrid regimes 6:97
median voter model 6:97
modernization theory 6:98
political change 6:97–99
political regimes 6:97
tests 6:98
theories 6:97–98
in United States 1:619
and urban policy 24:902–903, 24:905–906
and Walzer, Michael 25:349
and war 25:357–362
war and 25:402–403
wart and 25:402–403
weak/strong 25:357–358
Weber on 25:459–460
and world republic 25:741, 25:745–746
Democracy and law

Democracy and law (*continued*)
areas of common interest and disagreement 13:436
in classic debates 13:436–438
contemporary 13:440
contribution of authoritarian and totalitarian regimes 13:440–442
freedom 13:437
individual rights and liberties 13:437–438
questions and arguments concerning 13:437
relationship 13:435–436
Democracy by decree 13:819
Democracy Index 24:451–452
Democrasubjection 4:4
Democratic accountability, of United States Congress 4:653
Democratic activism 18:494
Democratic aristocracy, and celebrity 3:275
Democratic capitalism 13:200
Democratic centralism 4:303, 14:657, 14:659–660, 18:404–405, 18:408
Democratic citizenship 6:87–88
Democratic coherence 19:536
Democratic consolidation 6:94
Democratic constitutional state, Habermas on 10:471–472
Democratic corporatism, and social democracy 22:243
Democratic Corporatist model 22:584
Democratic decision making 6:85
Democratic deficit 13:438
Democratic education 4:381, 22:730–731
Democratic faith 3:680–681
Democratic imaginary (demos) 22:890
Democratic instability 2:190–191
Democratic judge 12:879
Democratic justice 12:881
Democratic learning hypothesis 18:494
Democratic legitimacy, and presidentialism 18:864–865
Democratic narcissus 1:445
agon as 1:444–445
Democratic negotiating practices 18:401
Democratic parenting 8:734, 19:245–246
Democratic peace 7:664
Democratic peoples' republics 17:515
Democratic publics 5:525–526
Democratic quality, database 5:833
Democratic radicalism 4:687
Democratic regimes, broadcasting regulation 2:859–860
Democratic relativism, Feyerabend on 9:122–123
Democratic Republic of Congo 22:610–611, 25:384–385
education 22:610–611
funding 22:610
research 22:611
Democratic revolutions
and Enlightenment 7:665
of post-Cold War era 20:623–625
Democratic role of police 18:263–264
Democratic rule of law 13:435–436, 13:439
Democratic socialism 20:636
Democratic Socialist Republic of Sri Lanka *see* Sri Lanka
Democratic societies, and group violence 14:704
Democratic state, science and 21:214
transformation of modern democratic state 21:215–216
Democratic theory 6:90–96
Aristotle's 6:90–92
boundary problem 6:93–94
catallaxy 6:92
civic republicanism 6:91–93, 6:95
collective choice theory 6:92
contemporary variations 6:92
cosmopolitan democracy 6:94
deliberative democracy 6:92–93
democratic boundaries 6:94
liberal democracy 6:91, 6:93
Locke's 6:90–91
Mill on 6:91
participatory democracy 6:91–93
power-political pluralism 6:93
preconditions for democracy 6:94–95
consociationalism 6:94
culture 6:94
democratic consolidation 6:94
economic systems 6:95
radical-democratic pluralism 6:93
rational choice explanation 19:896
representative democracy 6:91–92, 6:94
Rousseau on 6:90–91
Schumpeter on 6:91
theocracy 6:95
Tocqueville on 6:91
Tocqueville's 6:91
value pluralism 6:95
Democratic therapeutic community (DTC) 24:274
Democratic totalitarianism 16:834
Democratic transition, East Asian politics 6:799–803
Democratic urbanism, anxiety for 24:789–790
Democratic value commitments 13:11–12
Democratic values
and administrative law 1:109
global governance 10:160
Democratization 2:190, 2:779, 3:612, 6:97–99, 13:567, 16:8, 16:798–799, 16:803, 17:511–514, 18:434, 20:745, 20:854, 22:633–635, 24:897, 25:512–513, 17:512t
and criminology 5:240
critical theory 5:306, 5:309
of education 5:449–450
and ethnic cleansing 8:84, 8:86–87
impact of 19:636
journalism 12:855
and majority rule 14:446–447
as mechanism 22:419
media imperialism 15:58
and modernization 15:697–698
and political communication 18:347–348
and social and behavioral sciences 22:143
social science practice with 22:633
Democratized peer community 17:683–684
Democratized shamanism 16:169
Democratizing policies 22:329–330
Democritus 14:243
Demographic analysis 6:109–110
Demographic and Health Surveys (DHS) 5:737, 8:787, 9:96–98, 12:434
AIDS-related mortality 15:836
Demographic aspects, of infertility 12:45–49
Demographic balance, and ethnic conflict 8:93–94
Demographic changes
geriatrics 10:105
welfare state 25:522
Demographic characteristics, of homeless people in United States 11:160
Demographic consequences of population aging 18:541–542
Demographic crisis 20:837
Demographic databases, family reconstitution 6:140
Demographic differences, in adolescent leisure activities 13:830
Demographic economics 11:224
Demographic engineering 8:92
Demographic factors, of fertility transition in East Asia 9:56–57
Demographic groups, suicide risk in 23:683
Demographic growth 8:220
Demographic history 2:602, 11:290–292, 22:309
Demographic measurement 6:109–115
crude rates and degree of specificity 6:110
cumulative experience 6:109
death rate 6:119–121
direct standardization 6:121
fertility measures 6:110–114
age-specific fertility rate 6:110, 6:112–113
demographic analysis 6:109–110
difficulties with TFR 6:113–114
hypothetical cohort 6:113
level 6:110–112, 6:111t
occurrence-exposure rate 6:110–113
order-specific fertility rate 6:111–112
parity-specific fertility rate 6:112, 6:114
period parity progression ratios 6:114
standardized fertility rate 6:110
synthetic cohort 6:113–114
tempo adjustment 6:114
TFR and associated indices 6:113
timing 6:112–113, 6:112t
indirect standardization 6:121
marriage, cohabitation, and union disruption/divorce 6:116–119
level 6:116–119
synthetic or hypothetical cohort measures 6:118–119
timing 6:119
migration 6:121–122
mortality 6:119–121, 6:120t
age standardization 6:121
life expectancy and life tables 6:121
overview 6:109–110

population analysis 6:109
population growth 6:122
rates, probabilities, and ratios 6:110
separation 6:118
standardized mortality ratio 6:121
Demographic microsimulation 15:343–346
agent-based models 15:344–345
applications 15:345
calibration 15:344–345
closed *versus* open population 15:344
continuous *versus* discrete time 15:343
heterogeneity 15:344–346
historical demography 15:345
HIV/AIDS 15:345
individual-based models 15:344, 15:346
kinship 15:344–346
modeling options 15:343–344
Monte Carlo methods 15:343
social interaction 15:344–346
starting population 15:344
strengths and limitations 15:345–346
top-down *versus* bottom-up 15:344
Demographic models 6:123–125
age patterns of demographic events 6:123
'one-sex' modeling 6:123
parametric models 6:124
semiparametric models 6:124–125
statistical demography *versus* mathematical demography 6:125
use of 6:123–124
Demographic rates 5:739–740
availability of 5:739
comparability of 5:739
LEXIS diagram 6:154–161
period and cohort measures 5:739–740
structure of 5:739
Demographic revolution 11:648
Demographic studies
counterinsurgencies 5:99, 5:101
of fertility, application of sociocultural anthropology to 9:1–2
Demographic techniques
data adjustment and correction 6:126
family reconstitution 6:138–142
indirect estimation 6:143–148
inverse projection 6:149–153
small area estimates and projections 6:162–169
Demographic theory 18:579
Demographic transformations 19:695
Demographic transition theory 15:869
Demographic transitions 11:6–7, 15:390–391, 15:394–395, 15:869, 17:40–41, 18:540, 15:391f
and age structure 1:338–340, 1:342
and child fertility 8:822
and elder care 7:335
in Eurasia 18:546–547
and fertility recovery 9:775–776
and life history evolution 9:28, 9:30, 9:33
life history theory (LHT) 9:43–44
and social networks 9:6, 9:8–9
stages 9:61–62
and wealth flows 9:35, 9:38
Demographic variables, and adolescent health 1:123
Demographics
and comparative health systems 4:401
of divorce 14:518
in military 15:513–514
and poverty 18:743–744
and property crime 19:163–164
Demography 1:839, 5:735, 7:817–818, 19:350
of adoption 1:137–138
data and collection issues 1:137–138
standardization of measures 1:138
and child/infant mortality 12:20
differences in procrastination 19:74
of disasters 6:101–108
educational research 7:281–282
effects of segregation 21:407–410
evolutionary *see* Evolutionary demography
and extreme longevity 17:176
families and households 8:725–730
and family history 8:738–742
family size preferences 8:805–808
and family therapy 8:815–820
and fertility transition 9:52
in China 9:47–48
Hutterite on 11:445–447
and labor history 13:169–171
life course 14:76
multilevel models in 16:48–56
multistate transition models in 16:109–115
and obesity 17:83–84
organizational 17:699–700
organizational ecology 6:890–892
of organizations 17:358–359
period and cohort analysis in 17:727–734
APC model-based demographic forecasting 17:732
cohort variation in associational studies 17:733–734
conceptualizations 17:727–728
descriptive analysis 17:728
estimation and interpretation challenges 17:728–730
hierarchical age-period-cohort models 17:731–732
intrinsic estimator for linear regression models 17:730
linear regression models for APC analysis 17:728–730
mixed effects models 17:731, 17:733
mortality 17:727–734
research designs 17:730–732
temporal dynamics in population heterogeneity 17:732–733
time change 17:728
of Southeast Asia 9:85–87, 9:85t–86t
statistics at origins of 23:423
twin studies in 24:703–709
and unintentional injuries 12:131–132
value of children 3:498–501
Demography, history of 5:420–422, 6:170–175
census enumeration 6:171
family reconstitution 6:173–174
Farr's views 6:171–172
generalized inverse projection 6:173–174
Graunt's views 6:171, 6:174
indirect estimation 6:173–174
Lexis diagram 6:171–172
life tables 6:171, 6:173–174
L'Institut National d'Etudes Demographiques 6:173
Lotka's views 6:173
Malthus's views 6:170, 6:172–173
nation-states 6:170–171
Princeton Office of Population Research 6:173
since 1900 6:176–181
age structure 6:177
censuses 6:176, 6:179
databases 6:179
eugenics 6:177–178
families 6:176–178
fertility 6:176–179
households 6:179
population dynamics 6:176–179
stable population theory 6:173–174
transition theory 6:173
Demokratisierung 18:489
Demokratizatsiia reform 20:845–846
Demonization 6:31–32
Demonstratives, in dexis 6:52–56
Demoralization of enemy 25:399
people as 22:887–888
DEMs *see* Digital elevation models (DEMs)
Demsetz, H. 19:169, 19:171
Demystifying Mentalities 5:423
Demythologization 2:907
Denationalization of classes 3:748
Denationalizing processes 10:256–259
Dendrites 16:703–707
axon 16:703–706
computational implications 16:706
electrical signaling results from conductance changes 16:704–705
and enriched environment 2:837–838
ion channels 16:704–706
membrane potential 16:704
motor reflexes 16:703
neurites 16:704
synaptic inputs 16:705
synaptic integration 16:705–706
synaptic plasticity 16:706
vertebrate phyla 16:703
Dendritic tree 16:703–704, 16:706
Dendrochronology 3:564–565
Dendrological analysis 17:458
Denenberg, Risa 15:111–112
Deneuve, Catherine 25:177
Deng Xiaoping 3:515
Denial 9:705–706
death/dying 5:866
self-defensive 10:706
Denkabstraktionen 20:203
Denmark
anthropology 1:754
migration to 15:383f
Dennis and Humphreys model 15:164

Denominational diversity 20:377–378
Dense social networks, populations with 9:9
Densely inhabited districts (DIDs) 24:859
Density
  city, and sustainability 18:199–200
  dynamic economic systems 15:298–299
  of social networks 16:546, 22:462
Density (statistics)
  fiducial/structural statistical invariance 9:125–126
  of graphs 22:454–455
  probability 19:34
  probability density estimation 19:29–32
Density dependence, organizational ecology 6:892–893, 6:893t
Density-dependent mechanism of population regulation 3:155
Density-weighted average derivative estimator (DWADE) 21:601
Dental anxiety 17:279–281
Dental fear 17:280–281
Dental microwear analysis 17:458
Dental phobia 17:280–281
Dentate gyrus (DG) 10:887–888
Dentate stamping technique 24:482
Dentures 17:281–282
Denver Developmental Screening Test II 3:795
Denver Developmental Test 7:929
Denzau, A.T., works, *Ideologies and Institutions* 11:563
Deontic powers 21:344
Deontological approach 4:679
Deontological ethics 19:581
  social work 22:783–784
Deontologists 4:679
Deontology 8:74, 18:194, 18:196, 20:893–894
  and morality 15:808
  and value of life 25:16
DeoxyHb 16:389, 16:392–393
Deoxyribonucleic acid (DNA) 9:1015–1016, 10:42–44, 11:289
  analysis 1:633, 9:343–344, 9:948–949, 9:955–959
  -based cosmetics 9:979
  databases 9:340–341, 9:958–960, 16:760
  in disaster victim identification 9:946–947, 9:949–950
  dragnets 9:955–956
  epigenetic inheritance 7:832–837
  evidence 10:455
  fingerprinting 9:955
  forensic *see* Genetic databases, forensic
  Genographic Project 10:22
  genomics, ethical issues in 10:32–41
  histone acetylation 7:840–841
  histone methylation 7:841
  histone monoubiquitination 7:841
  importance of 9:991
  methylation 1:37–38, 7:839, 22:486, 24:683–685, 24:687–690, 24:692–693, 24:698
    exercise 7:844
    genomic imprinting 7:850
    inheritance of modifications 7:842
  nucleosome, and histone tail modifications 7:841f
  packaging 7:839–841
  polymorphisms 2:452–453
  posttranslational modifications of core histones 7:839–840, 7:840t
  profiling 9:945, 9:947
    development of 9:955–956
  records, collection and retention of 10:33–34
  research, microevolution 15:338
  role of histones in transcriptional regulation 7:839–840
  signatures 1:759–760
  testing
    and wrongful convictions 25:785–790
  typing 21:161
  use in courtroom 8:46
  wars 9:339, 9:957–958
Deparliamentarization of democracy 17:513–514
Department for International Development (DFID) 1:270
Department of Defense 9:336
Department of Economic and Social Affairs (DESA), United Nations 24:753–754
Department of Health (UK)
  No Secrets guidance 7:329, 7:332
Department of Health and Human Services (DHHS) 4:252
Department of Housing and Urban Development (HUD), US 8:532–533, 11:159, 16:451
Department of Justice (DOJ)
  antitrust 1:796–797, 1:799–802
  Project Safe Neighborhoods 10:460
Department of Public Information (DPI), United Nations 24:752–753
Department store 4:753
Departmentalization, and in-house counsel 12:126–127
Departments in universities 22:592–594
Dependability 9:834
Dependence
  addiction 1:97, 1:100
  alcohol 1:508
  and drug use/abuse 6:687
  side effects (Psychopharmacotherapy) 19:464
Dependence ordering 14:451
Dependence theory, and power 15:667–669
Dependence-inducing distributions 10:342, 10:347
Dependencianistas 11:687
Dependencies in natural language 6:182–195
Dependency 3:483, 4:223–224, 15:136–137
  in adulthood 1:181–184
  Cardoso's ideas on 3:122, 3:124–125
  and developmental theories 3:126
  and human development 6:308–310
  in Memmi's thought 15:137
  and sociological changes 3:124
  and themes of disagreement 3:124
  theory 3:124–125, 3:308, 6:196–198, 10:187, 11:866–867, 18:362–364, 21:151–152, 24:300–301, 25:749
    capitalism 6:196–197
    core-periphery hierarchy 6:196–197
    and development studies 6:309
    division of labor 6:197
    exploitation 6:196–197
    and global capitalism 3:99–100
    in Latin America 22:144
    and liberation theology 14:33
    social and behavioral sciences 22:144
    and urbanism 6:283–284
    world systems 6:196–197
  and trust 24:668–669, 24:671–672
Dependency networks, and groups 10:438
Dependency ratio 1:327
  and elder care 7:335
  and labor market 1:413, 1:418
Dependency–support script 1:182
Dependent development theory 21:151
Dependent dimension 4:724
Dependent market economies 18:694
Dependent processes, life course 14:54
  at global level of society 14:54
  at macro level of society 14:54
Dependent variables 3:189–192, 7:613–614, 15:65, 19:725
  event history analysis 14:55
Dependent verb form 1:188
Dependent work 25:676
Dependent worker 7:573
Depersonalization 11:527, 21:455
Depictivist theory, mental imagery 15:220, 15:225–226
Depletion, oil 15:557–559
Depoliticization 6:26
  of constitutional law 13:742
  of megaprojects 24:840
Depo-Provera 24:119
  family planning programs 8:794
Deportation regime 1:622–623
Deposition, discovery 6:517–518, 6:520–521
Depreciation, and income tax 11:746
Depression 2:758, 3:112–113, 3:329, 6:199–213, 7:607, 8:349, 8:748–749, 11:605–607, 11:609, 15:194, 17:593, 17:595–596, 18:8, 21:523, 23:334, 23:556–560, 23:562–563, 8:348t
  behavior therapy 2:426
  bereavement 2:532, 2:534
  bipolar 2:684
  and breast feeding 2:847–848, 2:850–851
  and cancer 3:73–75
  and cardiac disease 20:188
  and child neglect 3:413–414
  in child soldiers 3:433–435
  cognitive behavioral therapy for 4:30–31
  coronary heart disease 4:917–921
  depressive syndrome 6:203

cultural 6:202–203
and diabetes 6:339–340
diagnosis (signs and symptoms) 6:200–201
dyadic coping 11:594
etiology 6:201–203
genetic 6:201–202
hormones 6:202
neurotransmitters 6:202
psychodynamic 6:201
psychological and behavioral 6:201
social 6:201
in functional GI disorders 9:608–609
and gender 9:705–706
geriatrics 10:104–107
and grief 2:537–541
health care 15:99–100
health risks of 7:497–498
history 6:199
and HIV infection 1:492–493
and hopelessness 11:199
in inflammatory bowel disease 9:608
interpersonal psychotherapy for 12:632
in lesbians 15:115
macroeconomics 14:406–409, 14:411
manic, management of 14:487
and marital quality 14:513–516
and mass trauma 14:747–749
mental–physical comorbidity 4:390–393
and negative symptoms 19:291
neurobiological model of 6:203f
optimism 6:207–211
and overgeneral memory 2:285–286
and patient adherence 17:599
perinatal 17:722–723
phenomenology 6:200–201
and physical activity 18:89
postnatal 17:723
and post-traumatic stress disorder 18:703–705
and procrastination 19:76–77
prognosis 6:204–205
psychiatric care for 19:296–297
psychological treatment 19:383
psychotherapy for 19:375–376
couple therapy approaches 19:376
individual/group approaches 19:375–376
religion and 20:264
and respiratory disorders 20:576–579
rheumatoid arthritis 20:657
and self-development 21:497
and self-efficacy 21:505–507
and serotonin 22:83–85
sex differences in 7:607
short-term psychodynamic therapy for 19:346
sleep abnormalities in 22:82–83
and social-emotional development of children 22:942, 22:944, 22:946
and somatoform disorders 23:11, 23:13
treatment and management 6:203–204
scales 6:204
suicide 6:204
Depressive anxiety 13:68
Depressive disorders 6:201t
social model of 6:202f
Depressive disorders, antidepressants for 1:780
Depressive episodes 6:201t
Depressive symptomatology 22:485–486
Depressive syndrome 6:203
cultural 6:202–203
Deprivation
bioarchaeology 2:600–601
culture of 5:591–594
experiment 7:458
relative 6:214–217
social *see* Social deprivation
Deprofessionalization 17:484
of judicial processes 15:71
Deproletarianization 24:818–819
Depth
of ethnography 16:106
and openness to experience 17:231
perception, and prenatal/infant development 18:853
Depth citizen science projects 3:635
Depth cues (vision) 25:156
Depth interview 12:671–672
Depth psychology 7:375
role of Jung in 12:902–906
Depth-first branch-and-bound (DFBB) search 4:719
Depth-first search 2:32
Der Eigene 9:644–647
Der sinnhafte Aufbau der sozialen Welt 10:435–436, 21:146–147
Der Witz 11:416
Deregulated, capitalism 3:98
Deregulation 10:258
and radio 19:880–881
Derge, David 18:461
Derivation 15:822
Derivational economy 9:877
Derivatives
Fourier transform 9:376
functional data analysis 9:516
modeling with 9:517–518
Derived demand 24:617
Derived joint commitment 12:841–842
Derived measurement 14:869–870
Derived statistics 5:758–759
Dermatology, psychosomatic disorders in 19:488
Dérogeance 1:965
Derrida, Jacques 4:772, 5:500, 6:503, 19:677–678
deconstruction 5:927–932
on frames 25:210
on linguistic turn 25:210
play of differences 15:291
on visual turn 25:209
DES *see* Diethylstilbestrol (DES); Differential Emotion Scale (DES); Dissociative Experiences Scale (DES)
DESA *see* Department of Economic and Social Affairs (DESA)
Desacralization 20:861
Desarguesian property 25:199
Descartes, René 3:261–263, 7:255–256, 18:101, 20:286, 20:443, 24:722–723
dualism proposal of 15:545–546
on heuristics 10:830
on mind-body separation 24:722
psychology 19:409–410
on unconscious 24:722
works
Discourse on Method 1:899
Meditations on first Philosophy 18:33
Descent 3:730–732, 8:290–291, 13:37–38
and genealogy 9:806–808, 9:810
as societal order 13:37
Descent system 16:840
Descent theory, and friendship 9:426–427
Deschooling, Illich's contribution 11:585, 11:588
Description
in dictionaries 13:938
role in ANT 5:526
Description logics (DL) 5:937, 14:319, 16:925
Description theory of reference 20:68
Descriptive analysis, demographic studies 17:728
Descriptive approaches
to decision making 5:918–919
to personality 17:827–828
Descriptive construal of reflective equilibrium 20:78
Descriptive epidemiology 15:179
progress and unresolved issues in 15:181–183
Descriptive ethnology 14:825
Descriptive grammars 23:885
Descriptive invariance 2:445–446
Descriptive linguistic work 20:901
Descriptive meritocracy beliefs 15:242–243
Descriptive methods
of philosophical analysis 18:28
time series 24:344
Descriptive models of mind, heuristics as 10:830–831
Descriptive orientation of linguistics 14:225–226
Descriptive phenomenology of religion 20:344–345
Descriptive psychology 6:426
Descriptive questions 19:53
Descriptive sociology 5:526, 13:586
Descriptive theories 11:15
Descriptivism 19:664
Desecuritization 21:398
Desegregation 6:218–228
apartheid 6:218, 6:220, 6:223–227
civil rights 6:218, 6:221–222
common identification 6:220
contact hypothesis 6:219–220, 6:226–227
controversies, debates, and future directions 6:225–226
de jure 6:218–219
discrimination 6:223–224, 6:226–227
and discrimination 6:530–531, 6:533–534
early social science case for 6:218–220

Desegregation (*continued*)
educational 6:218
poverty 6:221, 6:223–224, 6:224f
racism 6:218–227
and reduction of prejudice 6:220–221
schools 6:218, 6:221, 6:224–226
social change 6:219, 6:221, 6:223–227
and social justice 6:221–225
Desensitization 15:4
behavior therapy for 2:426–427
historical background 3:880–881
systematic 3:767–768
Desert asceticism 2:60–61
Desert bases 6:230
enumerate 6:232
locate 6:231–232
Desert claims
force of 6:232–233
institutional 6:229
structure of 6:230
Desertec scheme 7:636
Deserts
philosophical concept of 6:229–234
Deserts, philosophical concept of
bases 6:230–232
claims
force of 6:232–233
structure of 6:230
and entitlement 6:230
enumerating desert bases 6:232
achievement 6:232
effort/diligence 6:232
institutional 6:229–230
and justice 6:233
and merit 6:229
moral deserts 6:230
nonmoral 6:230
preinstitutional 6:229–230
rejection of 6:233–234
Deserts, West Africa 1:301–302
Deserving poor 18:730–731
Desiderata 19:728–731
knowledge representation 13:99
Design 20:15–16
in cartography 3:171–173
causal inference 3:258–259
of cluster randomized trials 3:900
constitutional 4:713–714
ethnographically informed, business anthropology 3:7–8
experimental 8:535–540
experimental laboratories 8:552
and integrative thinking 12:150
organizational 6:235–240
Design effect (deff) 13:387–388
Design efficiency, optimal experimental design 17:258
Design experiments
in instructional psychology 12:203–204
Design issues in outcome-based sampling 17:103
Design of an organization 4:486
Design optimization (DO) 4:475, 4:478
Design phase, ADDIE 12:198
Design principles, autonomous agents 2:309–310
Design review processes 10:449
Design sensitivity 3:258–259, 17:110–111
Design structure matrix (DSM) 15:719–720
Design technologies, engineering 7:645
Design thinking 12:145, 12:150
Designative function of language 4:791–794
Design-based approach, for sample surveys 20:886–889
Design-based sampling 23:186
Designer babies 20:474
Designer craftsmen 5:111
Designer drugs 6:686
Desirability 5:97
Desirability relation 4:873
Desirable personality, and interpersonal attraction 12:626–627
Desire
for cargo cults 3:140–142
for control 25:553–554
and homophobia/heterosexism 10:791–792
Lacan on 13:211–214
rationality of 12:336
sexual orientation 21:785–787, 21:789
Desire–belief–opportunity model 1:671–672
Desired family size 8:805
changes in desires 8:806
Desired total fertility rate (DTFR) 8:787
Desires, beliefs, opportunities *see* DBO (desires, beliefs, opportunities) theory
Desiring, and hope 11:199
Desiring subjects 8:902
Desistance 5:137–144, 20:196–197
adolescence-limited offender 5:139
chronic offender 5:141, 5:143
and crime 1:312–313
dual taxonomy theory 5:138
empirical findings 5:141–143
criminal offending trajectories 5:141–142
marriage, employment, and other adult life events 5:142
prospective prediction problem 5:143
group-based trajectory models 5:140–143
growth curve models 5:140–141
informal social control theory 5:138
intermittent offender 5:139–140
key emerging issues 5:143–144
key theories 5:138–139
life-course persistent offender 5:138–139, 5:141–142
methodological issues 5:139–141
definition problem 5:139–140
methods of analysis 5:140–141
operationalization complications 5:140
research 1:834
rise of interest 5:137–138
self-control theory 5:138
Deskaheh, Levi General 11:796
Deskilling 21:187, 25:691
of labor 6:604
Desmond Tutu model 18:837
Despair, and attachment 12:229
Despotic political order 18:388
Despotism 6:375
versus democracy, evolutionary group dynamics 8:444
of the eighteenth-century French ancient regime 3:708
Destination
choice in emigration from Europe 15:459
fertility at 15:399
Destination language skills of migrants 13:262–264
Destination-specific total fertility rate (D-TFR) 15:399
Destructive methods for forensic archaeology 9:335
Destructiveness 2:256
evolution of 14:703–704
Destructivity, and fire 9:199
Desubstantiation, and cultural heritage 5:417
Desynchronization, and time use 24:373–374
Detachment, in integrated attachment theory 12:229
Detainee, torture and 24:444–446
DETECT *see* Detection of simple structure multidimensionality (DETECT)
Detect-elect-connect model, of transfer in school learning 21:99
Detection of simple structure multidimensionality (DETECT) 6:438
Detection system 13:575
Detection theory 21:954
Detective stories 2:752–753
Detective style theories 21:625
Détente 24:450–451
during Cold War 4:128–129, 4:131–133
Determinate growth 9:41
Determinate sentences 18:880
Determinate theory of the profit rate 19:112
Determiner phrase (DP) 10:316
Determiners 14:295
Determinism 3:266, 9:405
and causality 6:241–242
early 6:241
feminist standpoint 6:241
Lévi-Strauss on 13:907
materialist 1:769
metaphysical implications 6:243
probability 19:23–24
of psychology 22:223–226
Strong Programme in the Sociology of Scientific Knowledge 6:245
technological 24:109–112
Deterministic axioms, making probabilistic structures 24:251–252
Deterministic cohesive group finding approaches 3:910
Deterministic geographic models 14:787
Deterministic methods 24:347–348

Deterministic model 23:470–471
Deterministic treatment strategy, in time-varying treatments 14:341
Deterrence 6:255, 13:572, 18:763, 18:877, 18:880
absolute 6:255–256
appeasement 1:826–828, 1:830
balance of power 6:247
complexities and 6:255–256
conditions for 6:256–257
additional 6:257
defense 6:247
escalation 6:248
game theory 6:248
general 6:255, 6:258
Grasmick–Bryjak study 6:256
in practice 6:248
preemptive war 6:248
restrictive 6:255–256
role of extralegal punishments in 6:257–258
and social control 22:235–236
specific 6:255
theory 6:255
central premises 6:256
link with other 6:258
in theory 6:247–248
torts in law and 24:432
Deterrence, sanction perceptions of 6:250–254
accuracy of 6:250–251
risk perceptions, on criminal behavior
Bayesian updating 6:253
deterrent effect of 6:251–252
experiential effect of 6:252–253
perceptual deterrence 6:250–253
punishment 6:250–253
Deterrence model of compliance 4:456–457
Deterrence policy 12:54–55
Deterrence strategies 25:819, 25:822
Deterrence theory 5:192
crime 6:255–259
of intimate partner abuse prevention 12:678
Deterrent effect 6:260–265
certainty principle 6:260–261
of imprisonment 6:261–262
of police 6:262–264
aggregate police presence and crime 6:262–263
police deployment and crime 6:263–264
Deterring crime 6:255
Deterritorialization 5:248
political geography 18:377
of world 10:310
Detrimental health effects of consanguineous marriage 11:727
Detroit Automation 2:296–297
Detumeszenztrieb 21:846
Deutsch, Karl Wolfgang 6:266–268, 16:291
contributions 6:266
intellectual progeny 6:267
on nationalism 16:291–292
professional career 6:268
works
Nationalism and Social Communication 6:266–267
The Nerves of Government 6:267
Quantitative Ecological Analysis in the Social Sciences 6:266
World Handbook of Political and Social Indicators 6:267
Deutsche Agrargeschichte 17:670–671
Deutsche Arbeiterpartei (DAP) 16:257
Deutsche Arbeitsfront 16:357
Deutsche Forschungsgemeinschaft (DFG) 22:670
Deutsche Gesellschaft für Soziologie
Ost- und Ostmittel-europa-Soziologie 22:676
Deutsche Post 18:642
Deutsche Volksunion (DVU) 18:410–411, 18:414–415
Deutscher bund *see* Germany; confederation
Deutscher Sonderweg 21:168
Deutsches Wörterbuch (German Dictionary) 13:939
Devaluation 9:730–731
and group violence 14:704
Devanagri script 25:777–779, 25:779t
DeVany, Arthur 24:530–531
Developed countries
comprehensive typology of welfare regimes in 25:486t
culture, and research ethics 20:510–512
and industrial societies 11:863–864
population momentum values 18:576
and science 21:150–152
welfare state regimes in 25:481t, 25:489t
Developed economies
sustainable transportation in 23:846, 23:848
Developers (real estate) 23:660, 23:662–664
Developing countries 17:145–146, 21:150–153, 24:302
adult mortality in 1:153
affordable private schools in 1:269–271
biomedical research subjects exploitation in 8:81
child/infant mortality in 12:21, 12:23–25
conditional cash transfers in 4:555
culture, and research ethics 20:509–512
economies of 7:123
technological catch-up for 7:125
education in 22:359
and global governance 10:160
and industrial societies 11:863–864
infertility in 12:45–46
informal work in 12:60–65
internal migration in 12:433–442
judicial systems in 12:880
population momentum values 18:576
relevance of educational policy in 25:485–488
application of welfare state typology to developing countries 25:487–488
familiarist regime 25:488
in productivist welfare regime 25:488
in protectionist regime 25:488
remittances 12:519–523, 12:519f
shadow economy labor force 21:864–865
technology in 13:522–525
urbanization 6:283–284
see also Third World
Developing economies, sustainable transportation in 23:846, 23:848
Developing regions 23:368
Developing societies
craft production in 5:110–111
development aid 5:111
for global market 5:110–111
for local use 5:110
Developing world, fertility transition in 9:92
Developing writers
knowledge for and about writing 25:769
skills, knowledge and strategies of 25:769
strategies for planning, drafting, revising, and editing 25:769–770
Development 1:244, 1:246, 3:124, 6:276, 7:12, 7:18, 7:809–815, 10:263–264, 22:254–256, 22:258, 24:846–852, 25:97, 25:342
adolescent 12:638
and leisure 13:831–832
aid 7:714–715
and anarchism 1:682–683, 1:686
and anthropology 1:775–776
attitudes toward 7:809–813
from 1800 to 1940s 7:810
after 1940s 7:811–812
after 1970s 7:812–813
of babbling 13:246
and brokers 2:871–873
Cardoso's contributions 3:124
in child and adolescent psychiatry 3:385
assessment 3:387
importance 3:387
childhood, and stress 23:572
children *see* Child development
cognitive *see* Cognitive development
critique of, and nongovernmental organizations 16:857–858
cultural variation in life phases 5:507–515
definition 20:12
academic and professional antecedents 20:12
good and bad development 20:12
dependency theory 21:151–152
dependent development theory 21:151
and disasters 10:547–549, 10:551
early, number sense in 17:49–50
ecological, external factors to control 22:267
effects of marital status and divorce on 14:518–524
effects on serial order memory 21:640–641
of EFs *see* Executive functions (EFs)
and employment 6:269–275
and evolution 7:810
factory system 20:259–260
of first words 13:246–247
and flow 9:261–262
genetic programming 2:44

Development (*continued*)
global *see* Global developments
and global south 10:186–190
goal-oriented behavior 10:268–274
growth of 7:714–715
human *see* Human development
in human resource management 11:348, 11:352
Illich's contribution 11:586–588
and industrial societies 11:866–867
in infancy 12:637
influences
on children's media use and preferences 15:1–3
during infancy and early childhood 6:14
and informal sector 12:56–57
and intelligence 12:317–318
issues and future directions 6:278–280
contested definition 6:279–280
crisis of developmental state 6:279
ecological crisis 6:280
economic restructuring 6:279
expanded embeddedness 6:279
globalization 6:279
role of state 6:279
Jews in 20:259
joint attention 12:844–850
LAC educational system 7:178–179, 7:181–183
language *see* Language development
language acquisition 13:245
of Latin America 3:123–124
and law 13:443–447
and learning 7:276
modernization
crises 15:698
human development sequence 15:697
political 15:697
modernization theory of 21:151
multinational corporations in 6:276–277
of multisensory integration 16:100–101
of nationalism 16:292–293
in occipital lobe 17:131
operations 20:17
perspectives on coping 4:891–893, 4:892f
as regulation 4:891
as strategies of action regulation 4:891–892
perspectives on motivation in youth sport and physical activity 15:914–920
postcolonial science studies 18:653
postcolonialism 21:153–154
process 20:14–17
psychosocial 12:639
reagency perspective views of 21:153–154
redefining state's developmental role 6:280
regional planning 20:134–135
rural economic 20:814–815, 20:817
of science concepts 21:278–279
of scientific concepts 21:275–280
site acquisition 20:15
social 10:264–266
social referencing of infants 22:556–558
emotion recognition 22:557–558
information-seeking 22:557
locomotion 22:558
referential understanding 22:558
social-anthropological aspects 6:301–306
antecedents 6:301
colonialism 6:301–302
early encounters of anthropologists 6:302
economics 6:303–304
era of sustainability, neoliberalism, and globalization 6:303–304
financial institutions 6:302
new directions 6:302–303
new millennium 6:304
return of anthropologists 6:302–303
structural adjustment programs 6:303–305
of sociology 3:124
in spatial cognition 23:112
stages of 20:14–17
and state 6:276–281
state's role in economic transformation (historical perspectives) 6:276–277
state's role in Third World 6:277–278
development planning 6:277
entrepreneurial initiative 6:277
import substitution 6:277
neoliberal critiques of state intervention 6:277–278
stereotypes 23:449
sustainable *see* Sustainable development
tasks and competences, in childhood 10:679–680
technology 20:18
technology and 13:521
temperament 24:176
and temperament 24:180–183
term 7:809
and Third World 24:300–303
time line of 20:15f
'transition to takeoff' stage of 21:150–151
of trust 12:637
United Nations Development Program (UNDP) 24:753
and urbanization 6:282–289
in age of economic globalization 6:285–287
circa 1950 6:283–284
circa 1975 6:282–283
circa 1990 6:285
of word combinations 13:247–249
Development, and migration 15:389–396
diaspora 15:393–394
integral part 15:389–390
international architecture of 15:394
managing 15:393
mobility transitions 15:390
origin and destination 15:390–393
destinations 15:390–391
origins 15:391–392
transnational turn in 15:392–393
remittances 15:393
Development, and transformation (health social work) 10:716–719, 10:717t
early twentieth century 10:716–718, 10:717t
environmental challenges 10:718, 10:717t
global scene 10:718
health disparity 10:718–719
nineteenth century 10:716–718, 10:717t
occupational social work 10:718
Development aid, craft production and 5:111
Development and law
capacity in the Global South 13:446
capitalism 13:443
change in economic and social relations 13:443
decision making 13:444
development economics 13:444–445
economic development 13:443–444
evaluation of experiences 13:445
evidence-based policies 13:446
financial development 13:444
idea of use of law 13:444
impact of global forces 13:445
innovations 13:445
legal reforms 13:445
mixed reforms 13:444
modernization theory 13:443
process of empowerment and capability enhancement 13:445–446
twentieth century 13:443–444
twenty-first century 13:444–446
Development Assistance Committee (DAC), China 3:502, 3:505–506
Development benefit, childhood cancer 3:462
Development cost and operating income 20:16f
Development economists 13:444
Development feasibility 20:15
Development geography 6:290–295, 6:314–319
agrarian change 6:314, 6:316–317
class 6:315–317
collaboration 6:292–293
contemporary themes 6:316–317
coproduction 6:292–293
emerging academic interfaces 6:293–294
engagement 6:291–292
feminist theory 6:315–317
gender 6:315–317
impact 6:291–292
imperialism 6:314–315
industrialization 6:314, 6:316–317
influence 6:291–292
Marxism 6:315
new development actors 6:293–294
participation 6:292–293
political ecology 6:317
political economy 6:317
postcolonial theory 6:315
post-World War II 6:314–315
race 6:315–317
urbanization 6:314–316
and youth 25:828–829
Development journalism 12:857
Development phase, ADDIE 12:198
Development planning 14:279, 20:801

Development Research Group (DECRG) 5:766
Development risks, and transfer of technology 24:143–144
Development settings, of econometric software 7:2
Development Trauma Disorder (DTD) 3:435
Development University Center (CINDA) 7:183
Developmental approaches, to morality 15:783t
Developmental assets 3:446–447
Developmental behavioral genetics and education 6:320–325
  concepts 6:320–322
    advances 6:322
    classical twin method 6:320–321
    heritability, shared environmental, and nonshared environmental effects 6:321
    quantitative behavior genetics 6:320–321
    twin method, requirements and assumptions of 6:321–322
  extended twin design 6:322
  findings 6:322–324
    environmental influences and intellectual development 6:323
    gene–environment correlation and interaction 6:323–324
    quantitative genetic findings in domains of ability and education 6:322–323
  Genotype–environment interaction (GE) 6:322
  implications for education 6:324
  individual differences 6:320–324
  motivation 6:322–323
  rGE 6:321
  school achievement 6:322–323
Developmental cognitive neuropsychology 4:93
Developmental cognitive neuroscience 16:391–392
Developmental contextualism, personality development 17:880–881
Developmental criminology 1:311–313
Developmental cycle, households 11:229
Developmental deadlines, Development regulation around 4:816
Developmental disabilities, CDS and language acquisition in 3:402
Developmental Disabilities Services and Facilities Construction Act 1970 6:474
Developmental discipline 15:776–777
Developmental disorders
  language *see* Language disorders
  of sign language 21:938–940
Developmental dyslexia 4:93, 16:685
  and language learning impairment 13:344
  see also Dyslexia
Developmental education 4:429, 13:712
Developmental effects, and retroactive interference 20:615
Developmental explanations 8:594
Developmental factors, and multiple language development 16:62
Developmental generalizability, of I/E model 12:428
Developmental historicism, and Cold War 4:122
Developmental inventory of black consciousness (DIBC) 19:834
Developmental issues, and acculturation 1:70–71
Developmental life-course (DLC) criminology 1:310–311
Developmental neurobiology 12:2–3
Developmental Neuropsychological Assessment Second Edition (NEPSY-II) 16:724
Developmental neuroscience, brain aging 2:795–798
Developmental niche
  children 3:395–397
  and education 22:864
Developmental niche framework 6:841
Developmental noise 10:44
Developmental pathways
  children 3:395
  of children, and intentionality 12:348–349
Developmental phonological dyslexia 4:93
Developmental phonology 18:76
Developmental plasticity 2:661, 8:451–452, 16:619
Developmental pragmatic disorders 18:790–791
Developmental privilege, and engineering 7:646–647
Developmental process, personality 17:882
Developmental prosopagnosia (DP) 19:250–251
Developmental psychobiology 12:2–3
  of infant social mind 16:481–482
Developmental psycholinguistics 9:201
Developmental psychology 2:349, 2:351, 15:775–776, 16:494, 19:414–415, 20:355–356, 21:558
  applications of cross-sectional methodologies in 5:354–360
    benefits and limitations 5:358–359
    brain imaging 5:357–358
    cognitive development 5:354–355, 5:359
    correlation 5:354–355
    IQ 5:358
    matching 5:356–357
    regression 5:355–356
    trajectory analyses 5:357
    Williams syndrome (WS) 5:357
  autism spectrum disorders 2:275
  cognitive development 18:120–124
  Harlow on 10:528
  of life management 14:85–88
  and morality 15:806–807
  and motives 15:953
  neonatal imitation 16:483–484
    love of shared knowledge 16:484
    Piaget's rejection of neonatal imitation 16:484
    psychoanalytic theory of neonatal identification with vital impulses of other 16:483–484
    role of emotions 16:484
  and perceptual learning 10:127–128, 10:132
  personality 17:879, 17:882
  qualitative *see* Qualitative Development psychology
  vocational interests 25:240
  Vygotsky on 25:332
Developmental psychopathology, and temperament 24:195–196
Developmental regulations 3:445, 22:152
  and adolescent self-regulation 21:547–550
  around Development deadlines 4:816
Developmental research 6:838
  across cultures and nations 6:326–331
    affordances of materials 6:329–330
    affordances of situations 6:327–329
    assessment settings 6:327–328
    challenges 6:326
    conceptions of knowledge 6:327
    conceptions of time 6:330
    infancy 6:326–329
    methodical challenges 6:326–328, 6:330
    recruitment of participants 6:327
    research methods 6:326
    sampling 6:327
    stimulus material 6:326, 6:329
Developmental schisms 12:404
Developmental science 3:445, 11:276–278
  personality 17:882
Developmental significance, of early friendships 9:448
Developmental stages 16:512
Developmental state model
  East Asian politics 6:798–799
  of vocational education 25:235
Developmental states 6:278
  analytical characteristics of 6:278
  East Asian 6:278
Developmental studies 6:307–313, 11:433–434, 18:180–185
  basic needs 6:310–311
  Cold War 6:307, 6:309–310
  decolonization 6:307–308, 6:310
  economic growth 6:307–311
  globalization 6:307–311
  historical origins 6:308
  inequality 6:307–312
  from modernization to dependency critiques 6:308–310
  neoliberalism 6:307, 6:309–310
  postdevelopment 6:310
  rights-based approaches 6:310–311
  source monitoring 23:33–34
    in Alzheimer's disease 23:34
    in children 23:33
    in healthy adult aging 23:33–34
  sustainable development 6:310–311
  Third World 6:307, 6:309
Developmental stutterers (DS) 9:266

Developmental systems
formulations 3:445
personality
see also Personality
Developmental tasks 14:86
gifted education 7:162
during life management 14:86
Developmental theories 13:67–68
engagement 1:32–33
learning 13:707
Developmental topographical disorientation (DTD) 11:390
Developmentalism 1:458–459
cultural *see* Cultural developmentalism
Developmentally appropriate curriculum 23:213
Deviance 13:150, 15:800, 22:235–241, 23:701, 24:123
and anomie 1:715
and antisocial behavior 1:790, 1:794
and crime 1:310, 1:312
disability as form of 6:473
emotional 7:536–537
and groups 10:425
legitimacy of 13:827
moral entrepreneur 15:796–797
and self-control 21:483–484
and youth culture 25:814–815
Deviance information criterion (DIC) 15:656–657
Deviance theory, workplace safety and health 25:728
Deviant behaviors 20:334–336
anomie and 1:719–721
Deviant case 3:203–204
Deviant nonprofit groups (DNGs) 25:255
Deviant voluntary associations (DVAs) 25:255
Deviants
emotions of 7:536–537
sexual acts 3:884
state dealing with 4:823–824
Devil worship 14:435, 14:437
Devil's mix 23:45
Devlin, Lord Patrick 25:75–76
Devotees 9:784
De-Westernization of social and behavioral sciences 25:541–550
(pre)colonial encounters 25:541–543
contestation of Western intellectual hegemony 25:549
epistemic pluralism 25:549
global diffusion of Western human sciences 25:541
institutional and material context of scientific de-Westernization 25:543–544
intellectual and argumentative strategies of programs 25:547–549
international and transnational networks 25:544–546
international scientific organizations 25:544–545
nation, socialism, and culture 25:547–549
national universities 25:541, 25:543–545, 25:548
politics 25:546–547
postcolonialism 25:546, 25:548–549
scientific decolonization 25:548
scientific migrations 25:541
Dewey, John 1:749, 6:332–336, 7:255–256, 7:568, 8:612, 9:460, 12:244–245, 14:694, 18:435, 19:413
on aesthetic education 1:230
biography of 6:332–333
in Chicago 6:332–333
contributions of 6:333–335
in field of aesthetics and fine arts 6:333
on democratic publics 5:525–526
distinction between religious and religions 6:335
life and work 6:333
Marxist–Leninist critique of Dewey's philosophy 6:335
on philosophy of education 6:333
on pragmatism 18:804–805, 18:808
relevance of contributions 6:335
on Soviet education 20:825–826
theory of democracy 6:335
works
Art as Experience 1:230, 6:334–335
Experience and Education 6:334
The Influence of Darwinism on Philosophy 6:334
The Public and Its Problems 6:334
The Quest for Certainty 6:334
Reconstruction in Philosophy 6:335
School and Society 6:332
DEX *see* Dexamethasone (DEX)
Dexamethasone (DEX) 7:607
Dexamethasone/corticotropin-releasing hormone (DEX/CRH) 22:85
d-exceedance record scheme 16:903
DEX/CRH *see* Dexamethasone/corticotropin-releasing hormone (DEX/CRH)
DF *see* Divergent force (DF) field; Dysregulated fear (DF)
DFBB *see* Depth-first branch-and-bound (DFBB) search
DFID *see* Department for International Development (DFID)
DG *see* Dentate gyrus (DG)
DGP *see* Data generating process (DGP)
DGPM *see* Dual Group Processes Model (DGPM)
Dhammasat 13:813–814
Dhammayutika Nikaya 2:891–892
Dharma 13:800–801, 25:630–631
Dharmapala, Anagarika 2:891
Dharmasastra 2:78, 23:58, 23:70
Dhat Syndrome *see* Semen-loss anxiety syndrome
Dhavan, Rajiv 13:515
DHEA *see* Dehydroepiandrosterone (DHEA)
DHHS *see* Department of Health and Human Services (DHHS)
Dhimmitude 15:136–137
Dhivehi 7:196
d'Holbach, Baron 20:286
works, *Social System* 22:827
DHS *see* Demographic and Health Surveys (DHS)
DHT *see* Double high-threshold (DHT) model
DI *see* Diffusion of innovations (DI)
Di Leonardo, Micaela 17:22–23
works, *Gender at the Crossroads of Knowledge* 8:899
Diabetes 6:337–338
adherence 6:338–340
angiopathy 6:337
diabetes mellitus 6:337–338
prognosis 6:337
self-management 6:338
further research 6:340–341
glycemic control 6:337–341
insulin 6:337–340
ketoacidosis 6:337, 6:339–340
lifestyle 6:337–338
mental disorders in 6:339–340
anxiety 6:340
depression 6:339–340
eating disorders and disordered eating behaviors 6:340
neuropathy 6:337, 6:340
and physical activity 18:87–88, 18:90
prognosis 6:337
psychosocial aspects 6:337–341
fears 6:339
hypoglycemia awareness 6:338–339
questionnaires 6:338t
self-management 6:338
and schizophrenia 21:23
self-management 6:338
type 1 diabetes 6:337–340
type 2 diabetes 6:337–341
Diabetes Attitudes, Wishes and Needs (DAWN) study 6:338
Diabetes treatment satisfaction questionnaire (DTSQ) 6:338t
Diabetes-specific problems 6:338
questionnaires 6:338
Diabetic ketoacidosis (DKA) 5:475, 6:337
Diabetic nephropathy 6:337
Diabetic retinopathy 6:337
DIAC *see* Directions in Advanced Computing (DIAC)
Diachronic corpus 4:950
Diachronic lexicology 8:207, 8:210
Diachronic linguistics 20:922–923
Diagnosis 6:342
of Alzheimer's disease 1:584–585
child and adolescent psychiatry 3:388
of depressive disorders 6:343–344
epidemiology 17:120–121
of intellectual disability and learning disorders 6:344
Kraepelin on 13:129–131
of mental health disorder 6:342–345
of obsessive-compulsive disorder 17:120
of personality styles and disorders 6:344

Diagnostic Analysis of Nonverbal Accuracy (DANVA) 8:678
Diagnostic and Statistical Manual of Mental Disorders (DSM) 5:571, 6:823, 17:592, 17:596, 17:949, 19:464, 21:523, 22:493–494
DSM-III 17:30–31, 24:549
personality disorder 17:893–894
somatoform disorders 23:10–11
DSM-IV 5:623, 6:200, 13:360, 15:182, 17:31–32
acute episode 6:200t
depressive disorder, diagnostic criteria 6:201t
depressive episodes, diagnostic criteria 6:201t
diabetes 6:338
diagnostic criteria for panic disorder and agoraphobia 17:475
language learning impairment 13:344–345
mania 14:484–485
somatoform disorders 23:10–11
on stressors 23:570
on substance dependence 19:464
DSM-IV-TR 2:755
DSM-V 2:755, 5:474, 5:623–624, 6:342, 15:182, 17:31, 17:949
bereavement 2:533–534
diagnostic criteria for panic disorder and agoraphobia 17:475
differential diagnosis 6:387–388
and five-factor model of personality 9:246–247
GD reclassification 17:592
on paraphilia 17:949
personality disorder 17:894
for personality disorder 9:246–247
post-traumatic stress disorder 18:702–703
psychosomatic disorders 19:486–487
on self-esteem 21:523
sexual motivation 21:763–766
social (pragmatic) communication disorder 18:790–791
somatoform disorders 23:10, 23:14
homosexuality 3:886
DSM-II 3:886–887
subsequent editions 3:888
schizophrenia 21:9, 21:11–12, 21:11t
sexual dysfunctions and paraphilic disorders 24:550
Diagnostic components, smooth testing 23:400–401
Diagnostic Criteria for Psychosomatic Research (DCPR) 19:488
Diagnostic test accuracy, meta-analysis 23:912
Diagnostic–prescriptive measures of students' current strengths and weaknesses 13:715
Diagnostics
generalized linear mixed models 9:850–851
between and within covariate decompositions 9:851
distribution of random effects 9:850–851
distribution of Y 9:850
missing data 9:851
Diagonal genealogy 9:809–810
Dialect, and social class 13:281–282, 13:285
Dialect geographers 6:351
Dialectic materialism 16:398–399
Dialectical behavior therapy (DBT) 2:431, 4:544, 6:827, 20:437–438
BPD 2:759
for personality disorders 17:896
Dialectical concepts 14:241
Dialectical constructivism 19:493
Dialectical critical realism 20:24
Dialectical critique of fetishism 20:204
Dialectical materialism 14:663
and Cold War 4:124
Dialectical method 10:490–491
Dialectical policy network analysis 12:771–772
Dialectical relationship between orality and literacy 14:244
Dialectics 14:642
claims 14:645–646
Geist 9:667–668
laws 14:642, 14:645
legal reasoning 13:776–779, 13:781–783
materialist 14:645, 14:647
Plato on 18:210
science 14:642–643, 14:646
Dialect-leveling 18:134
Dialectology 6:348–352
beginnings 6:348
computerization 6:350–351
decline and rebirth 6:350
methodology 6:349
motivation for developing 6:348–349
structuralist analyses 6:349–350
Dialectometry 6:351
Dialogic teaching, and coconstructivism 3:915–916
Dialogical approaches, autonomy 2:318
Dialogical self 4:731
Dialogicality 20:444, 20:448
Dialogue management (DM) 23:261
Dialogue patterns, and coconstructivism 3:916–917
Dialogues 23:798
neonatal imitation 16:486
Plato on 18:209
Ricoeur on 20:666–667
and science communication 21:222
and self 21:442–443
Vygotsky on 25:313
Diamond, Hugh Welch 18:83
Diamond, Jared
works
Collapse 7:724, 7:749
Guns, Germs & Steel: The Fates of Human Societies 7:724, 25:343–344
Diamond equilibrium 21:336–337
Diamond v. Chakrabarty (1980) 17:587, 24:765
Diaries 2:213
audience measurement 2:212–214
studying 19:671
Diarrhea 25:428–429
Diarrheal diseases, child mortality cause 12:23
Diary log techniques are 10:590
Diary methods 22:548
Diary time use methods 24:380–381
data collection 24:381
time diary research studies 24:382
Diaschisis, and stroke 13:330
Diaspora nationalism 16:292
Diaspora populations, in Eurasia 18:550
Diasporas 6:353–356, 10:213, 10:216, 11:451, 12:875–876, 15:354, 15:358
African Studies 1:278–279
Africans 6:353
in archaeology 15:472–482
history 15:472
human dispersals 15:472, 15:477–479
neolithic-population movements 15:477–479
prehistory 15:473
catastrophic meaning of 6:353
Central American diaspora 1:622
Chinese 12:527–528
decisive economic effects 12:528
interpolarity of relations 12:527–528
multipolarity of migration 12:527
as consciousness 6:354
dispersion 6:353–355
and ecumene 7:111
e-diasporas 6:354
ethnic 8:91–92
ethnicity 8:160
forms of 6:354
future of concept 6:355
ghetto 10:121
Greeks 6:353
Irish 6:353, 6:355
Jewish *see* Jewish Diaspora
Jews 6:353, 6:355
and liberalism 14:12–13
migration 6:354–355
and migration 15:423
as mode of cultural production 6:354
in North America 17:15, 17:17–18
origins and development 6:353
Palestinians 6:353
as political orientation 6:354
and public policy 6:354–355
refugee 20:108
and religion 20:311
Roma people 20:748–749
as social form 6:354
in Southeast Asia 23:72
sub-Saharan Africa 11:90–91
types of 6:353–354
Diasporic citizenship 15:423
Diasporic communities, East Asian studies 6:785

Diathesis-stress model 8:423–424, 9:819–820, 9:830, 24:196
Diatoms 7:694–695
Díaz, Porfirio 15:326
DIBC *see* Developmental inventory of black consciousness (DIBC)
DIC *see* Deviance information criterion (DIC)
Diccionario de autoridades publicado por la Real Academia Espanola (Dictionary of the Authorities published by the Royal Spanish Academy) 13:939
Dicey, A.V. 12:897, 13:819
Dichlorodiphenyltrichloroethane (DDT) 7:811, 18:177
Dichotic listening 6:357–367
  asymmetry of working memory 6:358, 6:358f
  auditory pathways and neuroanatomy of dichotic stimulus presentation 6:357f
  clinical and developmental applications 6:359–360
  consonant-vowel syllables paradigm 6:358–359
  effect of handedness 6:359
  effect of sex 6:359–360, 6:360f
  forced-attention paradigm 6:363–366, 6:364f
  Handbook of Dichotic Listening 6:359
  neuroimaging 6:364–365
  15O-PET method 6:364
  research history 6:357–358
  right-ear advantage
    bottom-up modulation 6:360–363, 6:361f–362f
    brain correlates 6:363–365, 6:365f
    top-down modulation 6:363
  and speech 6:357–360
Dichotic listening technique 2:807–808
Dichotomous conception 3:708
Dichotomous items, scaling of 21:5–7
Dichotomy of competence/performance 4:295
Dicken, P., works, *Global Shift* 23:84
Dickey–Fuller distributions 24:328
Dictator game 9:588–589, 11:259
Dictatorship 4:753, 6:368–377, 24:448, 24:450–451
  after end of second World War 6:374–375
  apologists for and critics 6:372
  Arendt on 1:961–962
  contemporary 6:375–376
    definitions and typologies 6:376
    and dictator 6:375–376
  in Latin America 6:372
  media in 22:584–585
  modern dictatorship 6:369
  modern period 6:368–369
  Moore's contributions 15:769–772
  in nineteenth century 6:372–373
    Bonapartism' and 'Caesarism' in Europe 6:372
  of proletariat 6:372–373
  Roman dictatorship 6:371
    dictatorship and revolutionary power 6:372
    to modern dictatorship 6:371–372
    origins and functions 6:371
    transformation and end 6:371
  Schmitt on 21:28–29
  social and behavioral sciences under 22:139–146
    authoritarianism 22:140, 22:144
    democratization 22:143
    dependency theory 22:144
    Eastern Europe 22:139–143, 22:145–146
    economics 22:141–146
    European communist single-party states 22:140–142
    Hitlerism 22:140
    in Latin America 22:143–145
    military junta 22:139
    modernity 22:139–140
    post-Stalinist social sciences 22:141–142
    socialist one-party state 22:143
    Soviet Union 22:139–142
    Stalinist model 22:141
  totalitarianism 6:374
  twentieth century
    fascist dictatorship 6:373
    totalitarian regimes 6:373–374
    between the two World Wars 6:373
Dictatorship of the proletariat (French Revolution) 9:410
Dictionaries 7:584–593
  growing availability of 7:584–585
  history of 7:584
  making *see* Lexicography
  systemization of 7:585
Dictionarium Britannicum (British Dictionary) 13:939
Dictionary of National Biography 15:248
Dictionary-attested loanwords 3:922
Dictionnaire de l'Académie Française (Dictionary of the French Academy) 13:939
Dictionnaire Etymologique Roman project 8:208
Didache 3:548
Didactic tradition, of military 15:508–509
Didactics 21:119–120
Diderot, Denis 20:767, 22:697
  works
    Encyclopédie (*Encyclopedia*) 7:584, 20:767
    The International Encyclopedia of the Social and Behavioral Sciences 7:585–586
DIDs *see* Densely inhabited districts (DIDs)
Die Aufklärung 7:656
Die Freundschaft (newspapaer) 9:646–647
Die Gesellschaft 14:497–498
Die Linke 18:610–611
Die Republikaner (REPs) 18:410–411, 18:414–415
Die Tat 14:497–498
Diencephalic amnesia 5:923–924
Dies natalis 15:579
Diet 10:583
  and adolescent health 1:125
  and bioarchaeology 2:600–601
    illness and infection 2:601
    stress and deprivation 2:600–601
  and epigenetics 7:844
  and gender 9:704–707
  and obesity 17:82–83, 17:85
  plant-based diets 1:568–569
  religion and 20:266
Dietary behavior, of children 10:682–683
Dietary restraint story 25:471
  body weight perception and weight control 25:471–472
  supporting behavior change 25:472–473
Dietetics 9:311
Diethylstilbestrol (DES) 21:774
Différance 18:680, 18:687–688
Difference approach to language and gender 13:257–258
Difference hypothesis, gender 9:699
Difference making 3:272
Difference measurement, representational theory of measurement 14:857–858
Difference principle 12:915, 19:168
Difference studies, *versus* intersectionality 12:660–661
Differences 12:235–238, 12:236t
  and equality 7:117
  in feminist legal theory 8:920
  and interpersonal attraction 12:627
  in Middle East and North Africa 15:357–358
  and postmodernism 18:687–688
  problem, and culture 5:586
Differend 18:687
Differential Ability Scale–2nd edition (DAS-II) 12:287
Differential aging 6:378
  application
    to cognitive functioning 6:381
    to personality and social functioning 6:381–382
  brain aging and successful cognitive aging 6:380–381
  definition 6:378
  perspectives from successful aging 6:379–380
  principle 6:378
    origins 6:378–382
Differential association 5:191, 5:239
  extensions or reformulations in 6:72
Differential association theory of criminal behavior 22:410
Differential association-reinforcement theory 22:409–410
Differential attrition, in outcome-based sampling 17:103
Differential decay hypothesis 15:33
Differential diagnosis
  borderline personality disorder (BPD) 2:757–758
  in psychiatry 6:384–390
    boundary problem 6:387–388
    categorical classification 6:388–389
    comorbidity 6:389

consequences of definition by operational criteria 6:386
consequences of failure to validate boundaries 6:388
DSM-5 6:387–388
hierarchy of 6:386–387
history of 6:385–386
ICD-10 6:386–388
logic of 6:384–385
practical issues in 6:389
reliability 6:385–386
single *versus* multiple diagnoses 6:386
use of operational criteria for 6:386
Differential effectiveness 7:219
Differential Emotion Scale (DES) 7:520
Differential emotions theory 21:424–425
emotions as core adaptive survival responses 21:424–425
stability and change of primary emotions 21:425
Differential evolution, genetic programming 2:41
Differential games 17:255
Differential impacts 22:295
Differential item functioning 5:319
Differential location in social structure 22:413
Differential offending 19:839–841
Differential parental investment 8:460
Differential parenting 2:475–476
Differential psychology 4:110
mathematical psychology 14:808, 14:812, 14:815
see also Personality psychology
perspective of procrastination 19:75–76
Differential reinforcement 6:258
and criminal behavior 22:411–412
Differential social organization 5:239, 22:413
Differential socialization 8:855
Differential stability, and adolescent personality 17:864–865
Differential susceptibility 9:829
to experience, evolutionary developmental psychology 8:425
Differential susceptibility model 9:830, 24:196
Differential susceptibility theory 8:425
Differential treatment effectiveness, control variables 4:841
Differentially methylated regions (DMRs) 24:689
Differentiation 6:539
AGIL 1:388–389
of classification systems 3:800
in gifted education 7:159–160
of human societies 22:991
and journalism 12:853
leader–member exchange (LMX) 13:644–645
and modernity 15:690
modernization 15:697
in moral reasoning 15:804
oral and literate expression 17:276–277
perceptual learning 10:130–131
social 6:391–394
of society 21:384–385
in structural contingency theory 23:611
Differentiation and integration (DnI) approach 23:611
Differentiation theory
functional 9:536
Luhmann on 14:387–388
Difficult heritage 5:416–417
Difficult unstable life conditions 16:73
Difficulty of replication 9:131
Difficulty parameter 8:340–341
Difficulty-modulated network 16:593
Diffuseness (pattern variable) 1:382
Diffusion 6:406–410, 11:695–696, 16:764, 16:768, 24:141
anthropological aspects 6:402–405
acculturation studies 6:403–404
creolization 6:404
critical studies 6:404
culture 6:402–404
diffusionism 6:402–403
globalization 6:404
historicism 6:403
hybridity 6:404
social life of things 6:404–405
studies and evolutionism 6:402
syncretism 6:404
world system 6:404
center–periphery 3:308–309
communication channels 8:379
cultural 5:403
cultural geography tradition 6:407–408
cultural traits 1:766
development as process of 6:409
dynamic economic systems 15:298–299
and ecumene 7:110–111
empirical regularities in 6:407
and fashion 8:835
and fertility transition 9:50–52
Hagerstrand
conceptualization of 6:407–408
methodological framework 6:408
in higher education management 10:845, 10:847–848
identity 21:434–436
infrastructure perspective 6:408–409
and innovation 6:409–410
of innovation 8:378–381, 12:169–170
compatibility 8:378
complexity 8:378
observability 8:379
relative advantage 8:378
trialability 8:378–379
market perspective 6:408–409
medical geography 15:88
new communication technologies 8:380–381
opinion leaders 8:380
and scaling impact 22:360
social protest 22:514–515
social system 8:380
in sociological analysis 6:411–416
actors, decision-making, and social interactions 6:413–414
agent-based models 6:415
behaviors spreads 6:412, 6:414
class and ethnic-based politics 6:412
contagion processes 6:411–413
endogenous feedback 6:414
fertility decline 6:412–413
formal models of social interactions 6:415
individual heterogeneity 6:414
microsimulation 6:415
modern diffusion models 6:413–415
persistence and reversibility 6:414
social movements 6:414
social network approaches 6:415
social organization 6:411–412, 6:414
social structure as context 6:414
structural *versus* diffusion explanations 6:412–413
traditional theories 6:411–412
weak identification of process 6:413
stereotypes 23:449
technological 6:409
time 8:379–380
early adopters 8:379
early majority 8:379
innovation-decision process 8:379
innovativeness 8:379–380
innovators 8:379
laggards 8:380
late majority 8:380
rate of adoption 8:380
Diffusion explanations 6:412–413
Diffusion of innovations (DI) 12:146, 16:788
Diffusion process
comparison of diffusion and accumulator models 6:400
and empirical data 6:399–400
error response times 6:399
quantile probability plot of RT distributions 6:399–400
scaling and distribution shape 6:399
speed *versus* accuracy 6:399
historical foundations 6:395–396
judgment 6:396, 6:398–399
mathematical methods for 6:397–398
models 6:396–397
neural studies of diffusion making 6:400–401
neuroeconomics 6:400
Ornstein–Uhlenbeck models 6:398–399
psychophysics 6:399
Diffusion spectrum imaging (DSI) 11:326
Diffusion tensor imaging (DTI) 9:492–494, 9:525, 13:490, 16:587, 18:61
for ADHD 2:158
advantages 9:494
for cortical area studies 5:24, 5:26–27
disadvantages 9:494
fMRI 14:440
measurements 9:492–494
Diffusion theories, and fertility 9:6–8
Diffusionism 6:402–403, 12:642
and anthropological comparison 4:411–412
Diffusionists 15:709

Diffusion-weighted imaging (DWI) 9:494
advantages 9:494
disadvantages 9:494
Digestive system 17:608
Digit preference
data adjustments 6:127–130, 6:133f
Digit Span Backwards, dysexecutive syndromes 6:726
Digital age
and dramaturgy 6:649
ethical considerations (internet research) 12:607–611
social work in 4:287–288
Digital anthropology
definition 12:601
and Internet 12:600–602, 12:604
Digital capitalism 24:589
Digital cartography 24:12
Digital cinema 9:165–166, 9:169–170
Digital communication 11:550–551, 11:870–871
Digital cultures 18:676
and postmodernism 18:676–677
Digital divide 12:603, 16:787–788, 21:188
diffusion of innovations theory to 16:788–789
inequality and 16:790–791
motivation 16:789
physical and material access 16:788–789
second-level 16:789
skill dimension 16:789–790
in terms of access to the Internet 16:787
usage dimension 16:790
Digital economies 7:36
Digital elevation models (DEMs) 10:99, 10:101
Digital environment, literacy in 14:247
Digital fiber-optic cables 24:155
Digital genealogy 9:808–809
Digital health 10:647–648, 10:651
Digital images 25:173–174
Digital immigrants 22:918
Digital inequality 12:587–588
Digital information economy, and multinational corporations 16:79–80
Digital journalism 12:858
Digital Learning Material (DLM) 12:66
Digital lexical system 13:941
Digital lexicography 13:940–941
Digital libraries 14:46, 21:189
Digital life, and biocitizenship 2:611
Digital manipulation 25:173
Digital media 12:600–604, 16:778, 18:528, 18:531
in Egypt 16:802
publishing 19:622–623
Digital media environment, audience measurement in 2:212
Digital natives, engagement of 22:918–923
creative participation gap 22:921
hanging out with peers 22:919–920
interest-driven engagement 22:920
Digital photography 25:173
as erotic capital 8:4, 8:6
Digital preservation, in libraries 14:47–48
Digital religion 20:277
Digital representations
in tangible interaction 24:20–21
Digital resources, for education 12:66–67
Digital revolution 12:568, 22:662–663, 24:104–105
Internet 22:662–663
Digital stratification, and Internet 12:603
Digital technologies
collective memory 4:183, 4:186
and management accounting 14:467–468
Digital utopia 25:173
Digital video devices 15:33
Digital video disk (DVD) 25:81
Digital video recorders (DVRs) 2:212, 25:82
Digital world, communities of practice in 4:330
Digital-based economies 7:65
Digitale Diathek 2:24
Digitale Wörterbuch der Deutschen Sprache (Digital Dictionary of the German Language) 13:941
Digital-information computers and policing 18:263
Digitalization 14:736–737, 24:163–165, 24:167
of discourse 6:507
Digitally-enabled social protests 22:514–515
first wave 22:514–515
second wave 22:515
Digitization 10:60, 19:605, 19:607
adapting broadcasting regulation to 2:863
of cash 15:748
of content 7:586
of labor 7:586
and university press 19:628–629
Diglossia 6:417–423, 13:253, 14:173, 22:951–952
in Arab-speaking world 6:421–422
classical diglossia 6:417–418
classificatory schemes 6:417–418
homogeneous and heterogeneous diglossia 6:418
partial *versus* total diglossia 6:418
Cyprus 6:421
extended diglossia 6:417–418
in Greek-speaking world 6:420–421
high variety 6:417–421
linguistic features of 6:418–419
Arabic 6:419
Greek 6:419
Haitian Creole 6:419
low variety 6:417–420, 6:422
power 6:420
prestige 6:420
social features of 6:419–420
in South Asia 7:195, 7:197
in Switzerland 6:422
morphology 6:422
phonetics/phonology 6:422
vocabulary 6:422
Dignity 5:476, 24:123
and elder care 7:338
and equality 7:906–907
and punishment 19:656–657
and worth of all human beings, respect for 22:782
3,4-Dihydroxy phenylalanine (DOPA) 16:749
Dijkstra's algorithm 2:32
Dilation orderings 14:450
Dilemmas
collective behavior 4:159
honor killing 11:185–186, 11:189–190
Dilthey, Wilhelm 1:747, 6:424–429, 11:12–13, 17:90, 18:625, 19:392, 20:88, 20:345, 21:167
criticisms and influence 6:428–429
on culture 6:425–428
on descriptive psychology 6:426
on external organization 6:425–426, 6:428
on feeling 6:425–428
and *Geist* 9:667, 9:669–670
on *Geisteswissenschaften* 10:785
hermeneutics 10:785
hermeneutics, expressions of life, and worldviews 6:426–428
historicity 6:424, 6:426–429
and human sciences 11:381–383
on human sciences, lived experience, and psychology 6:424–426
on interpretation 6:424–429
on nexus 6:425–429
on relativism 6:428–429, 20:233
on temporality 6:425–429
on understanding 6:424–429
Dilution effects 19:20
Diluvialism 20:287
DiMaggio, Paul 16:562, 17:375, 17:393
on boundary crossing 3:802
on classification system 3:799, 3:802–803
on cultural capital 5:376–379
on cultural consumption 22:336
on cultural participation 5:448–449
on culture 5:535
on genre 3:799
on organizations 17:391
on social network analysis 5:558, 5:561
Dimension reduction 5:5, 23:196–197, 23:421–422
Dimensional Assessment of Personality Pathology 17:915
Dimensional change card sort (DCCS) 8:493f
Dimensional characteristics, five-factor model (FFM) 2:563
Dimensional comparison theory (DCT) 6:430–436
hypotheses on 6:430t
hypothesis I 6:431
hypothesis II 6:431
hypothesis III 6:431–432
hypothesis IV 6:432
hypothesis IX 6:434
hypothesis V 6:432–433
hypothesis VI 6:433
hypothesis VII 6:433–434
hypothesis VIII 6:434

hypothesis X 6:434–435
I/E model 6:430, 12:425–427, 12:429, 12:431
Dimensional complexity, honor killing 11:185
Dimensional differentiation, of negative symptoms 19:290–291
Dimensional gender differences 9:751
Dimensional stage of development 16:504
Dimensional trait model of personality 9:242, 9:244–247
Dimensional weighting model 16:46
Dimensionality
of MDS solution 16:43–44
of tests 6:437
constructing multidimensional data 6:438
Stout's evaluation of essential unidimensionality 6:437–438
unidimensionality concept 6:438
Dimensions
orders 17:528
of organizational group diversity 10:410
survey questions 23:765–766
Diminished immunity 22:485–486
Diminishing marginal benefit of wealth 7:89
al-Din, Rashid, works, *Comprehensive Chronicles* (*Jami' al-tawarikh*) 11:65
Dingwall, W.O., works, *Neurolinguistics* 16:678
Dinosaurs genome size 4:417–418
feeding time and the evolution of food processing 4:418–419
Diocletian (AD 285) 11:37–38
Dionysian child 7:147
Diop, Assane 10:146
Diop, Cheikh Anta 1:277
Dioula 1:301
Dipesh, Chakrabarty, works, *Provincializing Europe* 18:660
Diphtheria 15:869
Diploma Directive, 1988 3:672
Diploma of School Business Management (DSBM) 21:67
Diplomacy
and balance of power 2:345, 2:347
and sanctions 20:891
Dirac-d-distribution 9:377
Fourier analysis 9:377–378
Dire helping 19:221–222
Direct access hypothesis 2:575
syntax–phonology interface 23:895–896
Direct action, and anarchism 1:684–686
Direct activation by opsins 17:270–271
Direct aggression 1:373
Direct assessment of early memory 6:761–763
Direct associative learning experiences 18:40
Direct Cinema 6:622
Direct confounding 10:348
Direct contact, and adoption 1:144–145
Direct costs, pharmacoeconomics 19:458
Direct democracy 25:293–298
and anarchism 1:681, 1:685–686
ballot propositions 25:293–296
and campaigns 25:294–295
elections 25:294
information sources 25:295
question wording 25:294–295
regulation 25:295
participation and engagement 25:295–296
opinion formation 25:296
voter turnout 25:295–296
structure and context 25:293–294
initiatives and referendums 25:294
varying forms and institutions 25:293–294
voting behavior in referendums 25:296–298
cognitive engagement 25:297
heuristics 25:297
mobilization 25:298
models of voting choice 25:297–298
rational choice 25:297
Direct discrimination 18:830
Direct effect 15:76–77
Direct effect model, for social networks 22:366–367
Direct employment regulation 7:571
Direct indexicality 13:258
Direct Inference principle 19:41
Direct inhibition by opsins 17:271
Direct material input (DMI) 14:761
Direct object 10:324
Direct observations, for measuring self-regulated learning 21:544
Direct optimization routines 3:910–911
Direct payments
community care in England 4:345–346, 4:348
older people care 21:498–499, 21:501
Direct personal experience in shaping public's views 13:748
Direct racial discrimination 6:542–544
Direct reference theory of names 20:69
Direct retrieval 20:608–610
Direct rewards, adolescent sexual risk 1:130–131
Direct sanctions, adolescent sexual risk 1:130–131
Direct scaling methods
achievements 21:605
exponents characteristics 21:605
intermodal comparisons 21:605
mathematical psychology 14:803
problems of 21:605–606
quantitative phenomenology 21:605
scales of sensation magnitude 21:605
Stevens's power law 21:605
Direct standardization, demographic measurement 6:121
Direct treatment effect 8:538
Direct violence 17:622
Directed acyclic graphs (DAG) 2:363, 22:279
for expert systems and decision analysis 2:365
learning from data 2:366
Directed evolution 10:39
Directed graph 22:478
Directed networks 22:461–462, 22:464f
centrality in 16:536–537
Directed path 16:536
Directed Patrol Project (Indianapolis) 10:460
Directed social learning 22:405
Direction, neural representations of 16:623–627
Direction of fit (DoF) 18:797, 21:341, 21:343–344
Directional relations 14:286
Directionality of succession 3:157
Directions in Advanced Computing (DIAC) 21:188
Directives 18:797
and self 21:442–443
Directly measured statistics 5:758–759
Directors, stock ownership and company control 23:493
Direct-to-consumer advertising (DTCA) 18:10
Direct-to-consumer (DTC) personal genetic testing 2:623, 6:440, 9:898
clinical utility and health care integration 6:441–442
commercial biobanking and research 6:443
companies 9:902, 9:980
data protection and privacy 6:444
genetic ancestry testing 6:442–443
genetic risk 6:440–442
geneticization and human identity 6:444
Human Genome Project 6:440
medicine 6:441
race and genetics 6:443
regulation of 6:441
risk screening 20:699
Dirichlet measures 16:878–880
Dirkmaat, Dennis, on forensic archaeology 9:333
Dirt, preoccupation, and child neglect 3:411–412
Dirty wars 25:380
Disabilities 1:328, 3:421, 6:467–471, 23:213, 23:215
ableism 6:453, 6:456
accessibility 6:467, 6:469
activity 6:467–469
assessment of 15:182
attitudes 6:467, 6:469
autism spectrum disorders 2:275, 2:279
biopsychosocial model 6:467
body function 6:467–469
body structure 6:467–469
causes 1:396–397
classification and measurement 6:468–469
social determinants and consequences 6:469
construction in Western history 6:453–454
coping 6:467, 6:469–470
disablement 6:452, 6:456
and elder care 7:336–337
environment 6:467–470
functioning 6:467–469
future challenges 6:460–461
gender and 9:784
geography of 6:458–461

Disabilities (*continued*)
criticality 6:458–459
research topics and directions 6:459–460
and health 10:573–574
impairment 6:452–456, 6:467–470
interventions 6:455
limitation 6:467, 6:469–470
living with 21:844
medical model of 6:454
medicalization 6:454–455
and natural disasters 16:316–317
need of treatment 15:182
obsessive-compulsive disorder 17:121
in old age 1:395–396
old age work participation 17:182–183, 17:186, 17:188
participation 6:467–470
restriction 6:467–469
people with
law and 13:497–503
personal factors and psychological determinants 6:469–470
prevalence 6:467–468
psychological consequences and adjustment 6:470
reproductive rights of women with 20:471
and sanism 15:201–202
social exclusion due to 22:287–288
social model 6:455
social work practice
anti-oppressive practice 6:456
not grounded in politicized disability activism 6:455–456
sociological aspects 15:103–104
sociology of 6:472–477
disability policy 6:476
environment 6:472–474, 6:476
as form of deviance 6:473
functional limitation 6:472
identity 6:472–476
illness experience 6:475
impairment 6:472–476
managing experience of chronic disorder 6:475
minority group model of 6:473–474
movements, emergence of 6:474–476
new conceptual directions 6:475–476
process of adjusting to 6:473
rehabilitation 6:472–474
sick role 6:472–473, 6:475
societal reaction theory 6:473
stigma 6:473, 6:475
stigma of 23:466
Strauss on 23:529–530
studies 6:456
challenging normative assumptions of disabled bodies 6:454
relevance to social work 6:453
Disability Awareness in Action (DAA) 6:464
Disability discrimination 13:497, 13:499–501
Disability Discrimination Act 1995 6:446
Disability factors, of multilingualism 16:62
Disability movement 6:462–466
beginnings of the independent living movement 6:462
CILs and other organizations of disabled people 6:462
current trends 6:463–464
European network of independent living 6:463–464
international 6:464
personal assistants 6:462–464
seven needs of independent living 6:463
social model 6:462–463
Disability-adjusted life years (DALYs) 5:57–59, 5:782, 10:634, 19:546
Disability-specific human rights treaty 13:499
Disabled bodies, normative assumptions of 6:454
Disabled children, child protection 3:421
abuse in 3:421
Disabled people, community care in England 4:344–345, 4:347–348
Disabled People against Cuts (DPAC) 6:464
Disabled Peoples' International (DPI) 6:464
Disabled peoples organizations (DPOs) 13:498–499
Disablement 6:452, 6:456
Disablement process 1:394
Disabling environments 6:459–461
Disabling practices 6:452
Disablism 6:446–451
and ableism 6:448
definitions 6:446–447
discrimination 6:446–450
models 6:447–448
independent living model 6:447
medical model 6:447
minority model 6:447–448
social model 6:447
reducing in delivery of supports and services 6:450–451
disability activism 6:451
person-centered planning 6:450
self-directed and consumer-directed services 6:450
universal design 6:450–451
reducing within social work 6:449–450
anti-oppressive practice 6:450
awareness among social workers 6:449–450
social work 6:448–449
and practice 6:449
Disadvantage indexes, and homicide 11:164
Disadvantaged groups
and advocacy planning 1:221
and meritocracy 15:242
and tokenism 15:244
Disadvantages, concentration of 4:525–526
Disaggregation 7:915, 16:772–773
of conflict 24:216–217
nonequivalent group designs 16:854
and performance management 16:773
to public sector organizations 16:773–774
Disagreement
in ethics 6:589–594
objectivity of research 17:93–94
Dis-alienation 1:542
Disappearances, political 18:446–447
Disappointment 11:695
Disarmament 1:976–977
Myrdal on 16:159, 16:162–163
and peace movements 17:643–645
Disarmament, demobilization, and reintegration (DDR) programs 3:434, 4:636
Disaster 6:479
causes of 10:547–550
community resilience to 4:354–360
defined 9:945
definition 10:546, 22:753–755
demography of 6:101–108
death and property discussion 6:103t
fertility, reproductive health, and family change 6:105–107
high impact disasters 6:103–104
impacts 6:102–103
migration and relocation 6:105
mortality and morbidity 6:104–105
processes and outcomes 6:104–107
due to climate change 3:851
and green social work 10:385, 10:387–388, 10:390
human impact of 22:754–755
demographics 22:754
economic implications 22:755
political implications 22:755
social implications 22:754–755
and human rights 6:481
human-induced 22:753
natural 22:753
natural (or environment) hazards 4:356–357
nonviolent social crisis 4:356–357
post-tsunami Japan 22:570
preparedness 6:481
protection of children in 3:425–426
and psychosocial work 19:479–482
recovery 6:482
relief 6:481–482
resilience 10:548
risk reduction 10:551
social work response to 7:427–428, 22:753–760
social work stress/trauma 18:707–710
in social, cultural and political context 6:478–484
and stress 7:804–805, 7:807, 7:805t
study of 6:101–102
definitions 6:101–102
disciplinary approaches 6:101
technological hazards 4:356–357
terrorism 4:356–357
violent crime 4:356–357
see also Catastrophic errors
Disaster cycle 22:753–754
of health action 16:317f
mitigation 22:753–754
preparedness 22:753–754
recovery 22:753–754
response 22:753–754
Disaster ethics 22:756

Disaster management 20:563–564
Disaster mental health 22:756–758
Disaster psychiatry 15:495–500
environment of disasters 15:495–496
help-seeking 15:499
management and care delivery 15:498–499
organizational management of traumatic stress 15:499
mental health consequences of disasters 15:496–497
psychiatric illnesses 15:497
possible technological advances 15:499
resilience 15:495–500
risk and protective factors 15:497–498
screening 15:498
social support 15:496, 15:498
Disaster relief 6:481–482
challenge 6:481–482
damage and needs assessments 6:481–482
provision of basic needs 6:481–482
Disaster risk reduction (DRR) 6:479
Disaster tourism 4:185–186
Disaster victim identification (DVI) 9:945
complexities of 9:946–947
DNA profiling in 9:946, 9:949–950
genetic methods in 9:945–946
management of dead 9:947
in treatment of human remains and dead bodies 9:948–949
use of genetic profiling in 9:947–948
Disavowal, and fetishism 9:105–110
DISC1 (disrupted in schizophrenia 1) 15:211
Discharge (bankruptcy) 2:353–356
Disciplinarity
of educational research 7:280
transitions of social scientists to 22:698–700
national differences 22:699–700
Disciplinary apparatuses 23:728
Disciplinary boundaries 20:359
Disciplinary continuum 1:792
Disciplinary cultures 6:491
Disciplinary differences, journals 12:865
Disciplinary differentiation 22:267
Disciplinary histories, in discipline-building 6:492
Disciplinary interactions, during infancy and early childhood 6:15
Disciplinary networks 22:661–662
Disciplinary power 18:767
Disciplinary practices, parental 2:472
Disciplinary revolution 20:315–316
Disciplinary welfare system 1:792
Discipline
and antisocial behavior 1:792
organizational 17:342, 17:344
of parents 22:157
school 21:108, 21:110–111
Discipline formation
in social sciences 6:485–490
age of international associations, public policies, and mass higher education 6:489
continuities and reassessments 6:489–490
history 6:485
institutionalization 6:487–488
internationalization 6:489
from moral and political philosophy to social science 6:485–487
research programs in interwar world 6:488
social question, the research university, and new nation-states 6:487–488
Discipline sociology 1:719
Discipline-building 6:491
differences in disciplinary histories 6:492–494
disciplinary collective histories in 6:492
infrastructure 6:491–492
personal documents 6:492, 6:494–496
personal experiences 6:491–492
in social sciences 6:491–497
Disciplines (academic) *see* Academic disciplines
Disciplining of children, and religion 20:320
Disclosure limiting masks 4:579
Disclosure regulation 20:170
Disclosures
in clinical genetics 2:622
in England 6:519
informed consent to research 12:111
and intimacy 14:375–377
same-sex youth 9:653–654
statistical disclosure limitation 4:577–581
in United States 6:517
Disconfirmation of expectations model, of customer satisfaction 5:630–631, 5:631f
Disconnecting antisemitism 1:786–788
from racism 1:787
Disconnection syndromes 1:655, 12:417
pure alexia 12:420
Discontinuous negation 16:412
Discontinuous quantifiers 19:708–709
Discordance, trait/disease-discordant monozygotic twin study 24:691
Discordancy testing for outliers 23:388–390
Discordianism 16:806
Discorsi 3:772
Discount, diversification 6:587–588
Discount rate 5:56
Discounted utility model (DUM) 2:447
Discounting 6:255, 7:729–730
Discounting future costs and benefits 5:52–53
Discours de réception 19:611
Discourse analysis 4:726–727, 14:202–207, 21:309–310, 22:530
communicative process 14:203–204
discourse of historians 14:203–204
environmental movements 7:763–766
Fouc on 14:203–204
historical a priori of thought, speech, and action 14:203
material world 14:205
others 14:204–205
pluralism and heterogeneity 14:204
problem of historical agency 14:203–204
qualitative developmental psychology 19:673
social constructionism 22:225
Discourse approach, to business ethics 3:16–17
Discourse comprehension 13:402
inferencing as component of 12:40–41
Discourse ethics 10:471
and pragmatism 18:810
Discourse explanations of grammar 9:471
Discourse framework 24:297
Discourse particles 13:256–257, 13:290
Discourse principle 5:296
Discourse prominence 18:632
Discourse relevance 18:634
Discourse space, popular music 3:803
Discourse studies 6:510–515
as analysis 6:512–513, 6:513f
corpus 6:512–514
as global field 6:513–514
interactionism 6:513–514
language-oriented 6:510–514
post-structuralism 6:511, 6:513–514
pragmatics 6:510, 6:512, 6:513, 6:514
social research 6:512, 6:514
structuralism 6:513
text 6:510–511, 6:513
as theory, 6:511–512, 6:513f
Discourse styles, and society 13:291–292
Discourse theory 10:471
Discourse variables, and social class 13:283–285, 13:284t
Discourse-analytical approach 4:195–196
Discourses 11:128–129, 21:396, 21:398, 22:546, 22:830, 23:83, 23:85
anthropology of 6:503–509
after structuralism 6:503–504
commodification 6:507
digitalization 6:507
dispersion in new millennium 6:506–507
globalization 6:507
ideologies of language 6:505–506
labor 6:503, 6:507
linguistic anthropology 6:506–507
linguistic approaches 6:504–505
mediatization 6:506
neoliberalism 6:503, 6:507
performativity 6:504, 6:506–507
politics 6:505–506
around communities of practice 4:329–330
concept of 6:503–504
conceptual history 4:548, 4:550–553
as constitutive of the social 6:510
discursive construction of subjectivity 6:510
feminist study of 9:701–702
and Foucault's énoncés 14:199–200
and homophobia/heterosexism 10:791–792
and identity 6:498–502
inferences in 12:37–44
and language 5:289–290
and linguistic anthropology 14:172–173

Discourses (*continued*)
music as 16:150
nationalism as 16:295
and opinion formation 17:244–245
organizational 17:409
on possession 23:282–283
sexual orientation 21:786, 21:789
and social constructionism 22:224–225
societal relevance of themes and topics 6:510
as sociohistorical meaning-making practice in context 6:510
Discoverability, of linguistic databases 5:815
Discovery 4:224
doctrine of 11:795–796
organizational decision making for 17:354–355
Discovery (legal aspects) 6:516–522
aims 6:516
attorney's fees 6:518, 6:520
in civil law jurisdictions 6:520–521
Civil Procedure Rules 6:516–520
in common law jurisdictions 6:516–520
deposition 6:517–518, 6:520–521
document request 6:520–521
electronically stored information 6:518, 6:520–521
in England 6:518–520
affirmative discovery 6:519–520
costs 6:520
disclosures 6:519
electronic discovery 6:520
sanctions 6:520
scope and limitations 6:518–519
Federal Rules of Civil Procedure 6:516–517
interjurisdictional discovery 6:521
interrogatory 6:517, 6:520–521
privilege 6:517–520
relevance 6:517, 6:519
specific disclosure 6:519–520
standard disclosure 6:519
in United States 6:516–518
affirmative discovery 6:517–518
costs 6:518
disclosures 6:517
electronic discovery 6:518
sanctions 6:518
scope and limitations 6:517
Discovery learning 13:710, 21:271
Discovery representation 11:402
Discreet Trial Training model 2:278–279
Discrepancy theory, job satisfaction 12:834, 12:836–837
Discrepancy-based methods
model selection by 15:655–656
bootstrap techniques 15:656
cross-validation techniques 15:656
Discrete choice 24:808–809
optimal experimental design 17:260–261
Discrete choice experiments (DCE) 2:548–549, 24:618
Discrete core–periphery method 4:912–913
Discrete distribution 6:575
Discrete emotions 7:459–460, 7:519–520
Discrete manifest variables, in personality psychology 8:703
Discrete multivariate analysis 16:131–136
alternative model types 16:134–135
experimental design 16:134
historical notes 16:136
loglinear models 16:133–134
regression modeling of categorical response 16:132–133
repeated measurement data 16:135–136
sampling assumptions 16:134
traditional analyses 16:131–132
Discrete probability distributions 6:580
Discrete random variables, goodness-of-fit testing for 15:653–654
Discrete time
event–history analysis in 14:351–353
microsimulation 15:343
Discrete time models 18:585, 23:470
for social analysis 14:578–580
Discrete time replicator dynamics 8:388
Discrete valued time series 24:314
Discrete variables 5:3
multivariate analysis 16:121–130
Discrete-flow architectures 12:78
Discrete-state models 6:523–524
of information processing 6:523–529
Bayesian statistics 6:528
cognitive modeling 6:524–529
concept identification 6:524–525
expectation maximization algorithm 6:528
full binary trees 6:525
hidden Markov chains 6:524
Markov learning models 6:524–525
mathematical learning theory 6:525
memory models 6:525
multinomial processing tree models 6:525–529
paired-associate learning 6:523–525
recognition memory 6:523, 6:527
signal detection models 6:523–524
source monitoring 6:527–528, 6:527f
threshold models 6:523–524
Discrete-time difference equations 14:795
Discrete-time hazard probability 23:787
Discrete-time hazards 14:351
Discrete-time survival analysis
sexual development 14:336–337
data preparation 14:336
female sexual debut and menarche 14:336–337
male and female sexual debut 14:336
Discretion 20:162
in adjudication 5:106
and ignorance 11:579
and rules 20:163–164
Discretionary powers, trust 24:668
Discretionary review 1:822
Discretization, mathematical models 14:792, 14:794, 14:796–798
Discriminability, and recognition memory 20:38
Discriminant analysis 16:116, 16:118, 19:434, 23:392
Discriminant functions 23:413
Discriminant validity 6:345
for construct validation 9:841
Discrimination 9:761, 18:828–833, 18:839–840, 19:555–556
antigay 21:769
on butch–femme women 13:872
compensation for past 1:250
decomposition analysis 18:830–831
and desegregation 6:223–224, 6:226–227
direct 18:830
direct *versus* indirect 18:830
disability 6:446–450
economics of 6:536–541
employment 18:830–831, 21:771
ethnicity/migration 8:143–146
ethnocentrism and prejudice 8:168–174
female managers 8:885
gender, in labor market outcomes 9:759–762
gender-nonconforming people 21:769
genetic 10:36–37
genocide, history of 10:10
group 18:831
in higher education 21:37
housing 21:771
and immigration 11:675–679
implications for social policy 18:832
indirect 18:830
ingroup favorability 18:831–832
legitimacy of 13:827
of lesbians 13:893–897
LGBT 21:769–770
and LGBT aging 13:866
and meritocracy/tokenism 15:240–244
minorities 15:569–570, 15:573
of minority ethnic elders 15:440–443
models of labor market discrimination 9:759–761
norms 18:828, 18:830–832
older workers 17:186–187
positive measures to avoid unlawful 1:248–249
and poverty 18:743–744
versus prejudice 18:839
preventing direct 1:250
racial 1:786, 19:857–862
regarding offspring 15:609–611
relation with prejudice 18:831–832
Roma 20:754
sanism 15:200–202
sexism 21:693–699
sexual minorities
consequences and responses 13:880–881
cultural, patterns of 13:880
sexual orientation 21:769–770
and social exclusion 22:287
stigma 23:462–463, 23:465
transgender 21:769
in United States 6:530–535
and urban poverty 24:915–916
wage 18:830
wage differentials 25:338–339
work 8:106–107, 8:109, 8:111

Discrimination, economics of 6:536
future directions 6:540–541
labor force, self-selection in 6:540
progress 6:540–541
theories of 6:536–537
market power 6:536–537
preferences 6:536
statistical discrimination 6:537
tastes 6:536
wage decompositions 6:537–539
wage inequities 6:539–540
Discrimination account of proactive interference 19:6–7, 19:6f–7f, 19:7t
Discrimination parameter 8:340–341
Discrimination-and-fairness perspective, of organizational diversity 6:597
Discriminative stimulus 20:211–212
Discriminatory abuse, of elders 7:331
Discriminatory auction 2:205
Discriminatory bullying of LGBTQ youth 21:87–88
Discriminatory intent, in United States 6:530–531, 6:533
Discursive agency 14:206
Discursive consciousness 7:416
Discursive constructions of poverty and place 18:726
Discursive frame analysis, in environmental movements 7:765
Discursive framing 16:564
Discursive institutionalism 12:189
Discursive methods (discourse analysis/discursive psychology) 22:546
Discursive power 17:408–409
Discursive practices
mother–child communication 15:877–878
in social sciences 14:200–201
Discursive psychology 22:224–225, 22:530
ethnic identity 8:113–114, 8:116
qualitative developmental psychology 19:673
Discursive turn 22:544
Discussion forums, and identity 6:500
Discussion-based learning 3:176
Discussione 9:452
Disdain 2:487, 2:489
Disease development, and emotions 7:497–498
Disease management
health promotion and support in 2:468
rheumatoid arthritis 20:655
Disease models
agent-based models 1:370
of alcoholism 1:501–502
Disease outcomes, and stress 23:553
Disease patterns, changes in 16:907
Disease susceptibility, and pessimism 6:209–210
Disease-resistant varieties 9:308
Diseases 7:815, 18:174
biomedical sciences 2:663
burden, in health statistics 5:782
as departure from statistical norm 17:1
ecology 7:817
emerging/re-emerging, and comparative health systems 4:401
environmental influences on 24:698
epigenome-wide association scans for 24:693–698, 24:694t–697t
evolutionary causes of 8:449
gender and 9:784
gene-culture coevolution 9:815–816
health, anthropology of 10:571–576
in history 18:174–179
as lesion 17:1–2
as loss of homeostatic equilibrium 17:2
measuring mental imagery abilities in 15:225
medicalization and 15:106
nosology 17:27–33
obesity 17:82
occupational 25:727
olfactory system 17:199–200, 17:199t
pathogenesis, LTP/LTD 14:362
political ecology of 15:89–90
power to detect epigenetic changes in 24:692–693
prenatal genetic selection for avoiding 10:35
prevalence, in old age 1:396
prevention, and health literacy 10:653, 10:655
and psychological stress 23:583–584, 23:586–587
as social deviance 17:2
social inequalities in 22:275–276
social support and recovery 22:735–740
cancer 22:736
cardiovascular disease 22:735–736
chronic illness 22:736–739
functional support 22:735–736
infectious and other diseases 22:736
interventions 22:736–737
roommate 22:738–739
social network 22:735–736, 22:738
spouse 22:735–739
structural support 22:736
spread of 10:199–200, 15:86–88
and urban violence 24:955
Disease-specific measures, of health outcomes 10:660–661
Disenchantment, and McDonaldization 14:829
Disenfranchised grief 13:865, 17:431
Disenfranchisement, exclusion through 11:642
Disengaged parenting, and socialization of children 22:867
Disengagement 4:896, 6:207–208
by Australian Aboriginal students 15:902
from career goals 4:817
coping, chronic illness 3:550
and health 6:209
and optimism/pessimism 17:263–267
and pessimism 6:207–208
theory of 17:180
Disequilibrium 14:653, 16:501
perspective of internal migration 12:444
structural unemployment 24:738
Disgust 6:546–549, 7:468–469
anger 6:547–548
animal reminder disgust 6:547
and arts 6:548
behavior 6:546, 6:548
bitter 6:548
contempt 6:547–548
core disgust 6:547
cultural evolution 6:547
death 6:546–547
development of 6:548
as emotion 6:546
fear 6:546–548
food 6:546–548
and food preferences 9:297–298
history and culture 6:548
individual differences 6:548
interpersonal and moral disgust 6:547
meaning of 6:546–547
neural correlates of 6:548
preadaptation 6:547
psychology 6:546
and related emotions 6:548
sympathetic magical law of contagion 6:547
Dishabituation 10:481–482
cardiac defense 23:541–542
Disincentives, work 25:492–493
welfare program features to reduce 25:493–494
Disinhibition 6:77, 21:608
cognitive function 6:725–726
and deindividuation 6:38–41
Disinhibition effect, in online dating 17:213
Disintegration, social 12:243
Disinterestedness 12:388, 17:11
Disjunction 9:209
Dismissing adults 2:139
Disney (film maker) 9:157
Disneyland
entertainment 7:673
Disordered eating behaviors (DEB), and diabetes 6:340
Disordered grief, diagnostic criteria for 2:539
Disorders
of consciousness, electroencephalography for 7:359–361
in DSM 17:915–916
and fear of crime 8:856–858
of judgment 6:76
and normality 17:2–4
preoccupation, and child neglect 3:411–412
and psychological stress 23:583–586
of reading 6:732–733
forms of 6:732–734
of writing 6:732
forms of 6:734–736
Disorganized attachment 12:230–231
Disorganized children 2:138
Disorganized inequality 3:758f
Disparate health services and treatment 19:815–817
changes in racial inequalities 19:817
cultural mistrust 19:816–817
physician–patient issues 19:816

Disparate health services and treatment (*continued*)
psychosocial causes of 19:816
Disparate impact proof structure, discrimination 6:531–534
Disparity, definition 6:597
Dispensability thesis 11:814–815
Dispersal policies 23:159
Dispersed ownership 4:943–944
Dispersion 23:181
diaspora 6:353–355
Displaced aggression 17:621
Displaced persons, and ethnic cleansing 8:85–88
Displacement effects 16:793
Displacement hypothesis 25:170
Displacement of crime 18:879
Displacements 7:715–716, 15:472–473, 24:838, 24:840
dynamics of 20:106
internal migration 12:433, 12:437–438
land 13:235
Display practices 16:142–145
Display rules 12:4
in emotional expression 7:458–459
Disposability 6:550–555
Disposable energy 7:617
Disposable income in affluent nations 11:740t
Disposable soma theory of aging 17:176
Disposable youth 6:550–551
commodified culture 6:551–554
domestic terrorism 6:554
governing-through-crime complex 6:551
modernity 6:550–551
neoliberalism 6:550–552, 6:554–555
politics of cruelty and politics of civic engagement 6:555
social abandonment 6:554–555
social death 6:550, 6:554–555
state violence 6:553–554
throwaway society 6:551
in United States 6:554–555
violence and corporate power 6:551–554
Disposal, consumer psychology 4:738
Disposition effect
behavioral economics 2:440
Dispositional coping styles, in cardiac disease 3:112
Dispositional intelligence 17:753
Dispositional selectionism, and functional explanation 9:523
Dispositions 3:268, 12:381–382, 20:650
to act/believe 22:334–335
just world as 2:487–488
mathematics 14:818
personality 2:488
temperament *see* Temperament
Disproportionate noun impairment 21:575
Disputatio 6:62
Dispute processes
and law 13:536
and legal pluralism 13:759
Dispute resolutions 13:604–608, 16:419
adjudication 5:103
definitions and types of processes 15:70–71
early neutral evaluation 15:70
in economics 6:557–559
arbitration 6:558–559
strikes 6:557–558
forums 13:591–592
hybrid or secondary processes 15:70
Med-Arb 15:70
mini trials 15:70
multiple forms of 15:72
primary processes 15:70
process 14:265
purpose of law 13:744
requirements based on setting 15:71
summary jury/judge trials 15:70
voluntary and consensual 15:70–71
Dispute transformation 6:562
Disputes
law's archaeology 6:564
and peace/nonviolence 17:625
social construction and transformation of 6:560–566
adjudication 6:560–561, 6:565
articulation of normative order, and production of law 6:565–566
blaming 6:561–562
claiming 6:561–562
conflict management 6:560–561
dialectic of narrowing and expansion 6:562
emergence 6:561
framing 6:562–564
grievance 6:560–562
language 6:562–564
law 6:560–561
legal consciousness 6:560–562
mediation 6:560–561, 6:564–565
naming 6:561–562
negotiation 6:561–562, 6:564–565
political mobilization 6:561
role of lawyers, courts, and other institutions 6:564–565
social change 6:563
third parties 6:560, 6:562–563
Disputing, organizational 17:399
Disputing process 13:590–591
Disqualified persons 3:656–657
Disraeli, Benjamin 3:773–774
Disrupted in schizophrenia 1 *see* DISC1 (disrupted in schizophrenia 1)
Disruption, fortification, and resilient design 23:174t
Disruptions, in law firms 13:583
Disruptive behavior 2:470
Disruptive change 21:105
Disruptive social innovation 22:355–356, 22:356t
Dissatisfaction of coalition 4:869
Dissemination 11:695–696
of culture 5:732
labor statistics 5:799–800
in Latin America 22:628
and logic models 14:299, 14:304
and manual-based treatment 14:500–501, 14:503
research 20:510–512
of utilitarian ideas 24:983
Dissent 4:318–319, 25:346
Dissent, organizational groups 10:408–409
Dissertation sur la musique moderne 20:767
Dissimilarities, multidimensional scaling for 16:41–42
Dissimilarity index 21:411
Dissociations 4:90–91
of hysteria 11:508
neuropsychological, anomia 19:148–149
Dissociative amnesia 1:651–652
cerebral glucose metabolism in 1:656f
sex ratio in 1:653
Dissociative disorders 1:651
Dissociative Disorders Interview Schedule 1:652
Dissociative Experiences Scale (DES) 1:652
Dissociative identity disorder, forensic psychology 9:352
Dissociative states 16:170
Dissolution
marital 14:515–516
of partnerships in Western societies 17:549–550
Dissolution of Muslim Marriages Act 23:61
Dissolution of the Third Reich 16:260
Dissonance, cognitive 4:76–78
Dissonant acculturation 2:85
Dissonant heritage 5:416–417
Distal cultural factors in gender development 9:746–747
division of labor 9:746
educational system and practices 9:746–747
status, religion, and political influence 9:746
Distal influences on gender development 9:742–746
Distance 10:70–72, 12:405, 14:286, 16:572, 22:462, 23:198–200
and market geographies 14:545–546
and migration 15:435
space as 23:198
Distance education 1:151, 6:567–570, 6:895–896, 7:919, 22:8–9, 24:759–760
defined 6:567
in higher education 6:569
history of 6:567
location 6:568
MOOCs 6:569
occurrence 6:568
in P-12 Schools 6:569
participation 6:567
reasons for 6:568
technology involved 6:568
Distance effect
numerical 17:46
priming distance effect 17:48
Distance from the mean principle 8:431
Distance models 16:34
Distance tasks 4:81–82
Distance-based centrality measures 16:533

Distance-based clustering of metric data 19:433
Distant comparison 4:413
Distant readings 21:190
Distant suffering 2:749
  and mass media 14:691
Distinction 4:695–699
  Bourdieu on 4:695, 4:697–698
  and class 4:695–697
  cultural capital 4:695, 4:697–698
  future 4:698
  habitus 4:695, 4:697–698
  inequality 4:695, 4:697–698
  legal reasoning 13:776, 13:778–780
  between omnivorousness 5:450
  social 4:759–760
  status consumption 4:698
  Veblen on 4:695–698
Distinctive assemblages 20:396
Distinctive (core) competence 23:509–512
Distinctive features 12:782–783
Distinctive voice of color 5:267
Distinctiveness
  of counseling psychology 5:69–70
  identity, nonheterosexual British Asians 16:864
  temporal 24:199–205
Distinctiveness account of proactive interference 19:6–7
Distinguishability tests, longitudinal data 14:347–348
Distinguishing language contact phenomena 3:922
  in terms of benchmarks 3:922–923
Distorted cognitive schemes, and suicide 23:677
Distortions, of economic evaluations during elections 7:28–29
Distraction 9:705–706
  and attitude 2:171
Distress
  of others, neural circuits associated with perception of 7:543–544
  psychological, of cancer 3:73–75
Distress/anger recall, cardiovascular reactivity 3:118–119
Distributed coding 16:635
Distributed cognition 4:189, 19:361–362
Distributed lag, Fisher's concept 9:219
Distributed leadership 16:553
  of teachers 24:60
Distributed representations 4:96
  and semantic memory 21:575
Distributed reproduction 20:454
Distributed signals 16:634–635
Distributed Situation Space model, of discourse 12:41
Distributed virtual teams 24:102
Distributing, cultural mediators's role 5:442
Distribution 7:14–15
  of Native Americans 1:634
    current status 1:636
  social justice 22:392
  television 25:80–82, 25:84–85
  welfare reforms 25:517–518
Distribution function 19:34
  majorization/stochastic orders 14:449–451
  see also specific distribution functions
Distribution markets 14:552
Distribution norms 8:443
Distribution of income 25:7–9
Distribution of linguistic databases 5:814–816
Distribution of wealth, and globalization 12:542–543
Distribution patterns, trade and exchange 24:483
Distribution theory, Fisher's 9:223
Distributional characteristics 15:827
Distributional cost–benefit analysis 5:53–54
Distributional semantics 13:932–933
Distributional tests 19:432
Distributional view, of embodiment cognition 7:423–424
Distributionalism 2:705–708
Distribution-free interval estimation 17:292
  confidence intervals for quantiles 17:292
  prediction intervals 17:292
  tolerance intervals 17:292
Distribution-free intervals 8:19
Distribution-free sequential theory of psychological judgment 19:473–474
Distribution-free statistic, rank-based nonparametric methods 16:891–895
Distributions, statistical *see* Statistical distributions
Distributions without a zero class 6:582
Distributive justice 7:538, 12:928, 12:930, 12:935, 12:938, 12:950, 13:749, 15:239, 15:242–243, 17:380, 17:718, 19:168, 20:519, 22:387
  egalitarianism 7:311
  political economy of 20:895
  in research funding 20:524
Distributivity 19:709–710
District magnitude, proportional representation 19:206
District Policing Partnership Boards (DPPBs) 18:269
Districting
  apportionment 7:342–344
  zonation, and region delineation problems 23:174t
Distrust
  and anomie 1:715–716
  of mass communication 14:690–692
    degradation of art 14:691–692
    fabrications and corruption of youth 14:691
    iconoclasm 14:690–691
    responsibility to distant suffering 14:691
Disturbance processes
  geoarchaeology
    human disturbance process 10:54–55
    nonhuman disturbance process 10:54
Disturbance smoothing 24:358–359
Disturbed sensation of limb ownership 21:420
Disturbed temporal consciousness, and confabulation 4:562–563
Disturbing social trend, and generational phenomenon 9:859
Disulfiram, for alcohol dependence 3:766–767
Disunity of sciences 7:875–876
DIT *see* Defining Issues Test (DIT)
Divergence, consumer policy 4:734–735
Divergent force (DF) field 3:782
Divergent thinking, and creativity 17:771
Divergent thinking loss, and prefrontal cortex injury 18:815
Divergent validity 21:466
Diverse experiences, of health 10:571–572
Diversification 6:585–588, 10:252, 12:210–211
  of area studies
    in Soviet Union during Cold War 1:950–951
    in United States 1:952–953
  business groups 6:585
  conglomerate 6:585
  discount 6:587–588
  of feminist anthropology 8:899–900
  intangible assets 6:586
  of livelihood 17:580–583
  measuring 6:586–587
  outsourcing 6:585, 6:588
  and performance 6:587–588
  profiting from innovation 6:586
  property 19:154–155
  and qualitative developmental psychology 19:670
  resource-based view 6:586–587
  scope economies 6:586
  strategic management 6:586
  understanding 6:585–586
  vertical integration 6:585
Diversion 15:31
  and probation 19:44–47
  in United States 4:372–377
    complete diversion 4:374
    enhanced probation diversion 4:375–376
    justice leverage diversion 4:374–375
    prevalence of diversion programs 4:374
    probation diversion 4:375
DIVERSITAS Program 7:791
Diversity 12:916, 16:28
  accommodation, in South Asia 2:77
  during adolescence
    in relationships 22:219
    in social contexts 22:219–220
    in social goals and values 22:220
  of bargaining situations 16:415–416
  and careers 3:132–133
  in China 2:65–67
  competency 24:521
  of counseling psychology 5:69–70
  cultural 1:773, 1:775, 1:777
  denominational 20:377–378

Diversity (*continued*)
early childhood care and education (ECCE) 6:740
and Eastern European social work 22:795, 22:799–800
enduring struggles, in South America 23:48
ethical
reduction of, by statistics 23:424
in ethics 6:589–594
ethnic, in Southeast Asia 23:71
and ethnicity 8:148
and gender 3:19
human 1:772–775
and immigration 11:642, 11:645
indigenous 11:806
institutional 12:193–194
Jacobs on 12:777–778, 12:780
and multiculturalism 16:29–32
organizational groups 10:409–410
creativity and innovation 10:410
dimensions 10:410
gains and losses 10:409
in organizational groups 10:409–410
in organizations 4:738–743, 6:595–600
affirmative action 6:598
age 4:739, 4:741, 6:596, 6:598
categorical and relational demography 4:739–740, 6:596
complex conceptualizations of 4:741–742, 6:597–598
diversity climate 4:740–741, 6:597–598
diversity training 4:741–742, 6:598–599
ethnicity 4:738–739, 4:741–742, 6:595–599
gender 4:738–742, 6:595–599
and goals 4:738, 6:595–598
management 6:598–599
management-led diversity initiatives 6:598–599
minorities 4:738, 4:741–742, 6:595, 6:598–599
outcomes 4:738–739, 6:595–596
race 4:738–739, 4:741–742, 6:595–596, 6:598–599
research trends 6:599
role of context and moderators in diversity effects 4:740–741, 6:597
social categorization 4:738–740, 6:595–597
typologies 4:740, 6:596–597
values 4:738–741, 6:595–598
perspectives/paradigms 6:597
post-heroic commemoration 4:237–238
practitioner-scholar model 24:521–522
religious, in Southeast Asia 23:70–71
of situational judgment test 22:16–17
technological change and 24:104–105
and Third World 24:302
Diversity cultural *see* Cultural diversity
Diversity of social Darwinism 5:687–690
Diversity perspectives/paradigms 4:740
Diversity Project *see* Human Genome Diversity Project
Diversity studies, *versus* intersectionality 12:660–661
Diversity training 17:392
Diversity visa participants 11:635–636
Divertissements, culture of 7:671
Divi 8:836–837
Divide and rule policy 16:456
Divided government 8:500, 18:865, 18:868
Divided societies, policing in 18:266–272
Dividends, corporate finance 4:927–930
Divine absolutism 1:10
Divinity, and morality 15:807–809
Division
legal reasoning 13:778–779
and Third World 24:302
Division for the Advancement of Women (DAW) 24:754
Division of Europe 6:811
Division of labor 4:338–339, 6:601–605, 6:701–702, 9:746, 10:187–188, 13:156–162, 22:721, 25:341–342, 25:344
capitalism 13:158–160, 13:162
class 25:341
and computers 4:515–516, 4:518
concepts, levels, and forms 13:156–157
in craft production 5:108
dependency theory 6:197
determinants 6:244
and dynamics of change 13:159–160
change to postindustrialist societies 13:159
change to work organization 13:159–160
globalization and international division of labor 13:160
ethnicity 25:341
foundational core/external commonplace 6:601–602
gender 25:341
and gender 25:690
gender discrimination 13:157, 13:160–161
geographic location 25:341
Hochschild on 11:146
households 11:230–231
industrialization 13:156, 13:158–160
inequality, social 13:158, 13:160–161
inside home 14:606–607
institutional frame 13:158–159
application of science, technological change, and automation 13:159
concentration in manufacture and factory 13:158–159
free labor 13:159
liberalization of markets and competition 13:159
private property 13:159
separation of household and workplace 13:158
intertemporal 6:601
Kaldor on 6:602
and marriage 8:829
in modern societies 13:158–161
nation 25:341
need for coordination and cooperation 13:157
in nineteenth century 6:603–604
and political economy 18:358
race 25:341
sexual 13:157–158, 13:161
shifting 24:791–792
and social differentiation 6:391–393
social differentiation, and cohesion 13:160–161
and social integration 12:246
social structure 22:721–722
in social thought 13:161–162
and socioeconomic development 13:157–158
and specialization, functional explanation 9:536–537
and time use 24:379
translation 24:583
in twentieth century 6:604–605
types of 6:602
from William Petty to Adam Smith 6:602–603
Young on 6:602
Division on Women and Crime (DWC) 8:916
Divisive feedback equations 25:158–159
Divorce 6:116–119, 6:606–609, 17:489–490
and behavioral problems 2:475
consequences 6:607–608
coping, and religion 20:320
correlates 6:606–607
additional correlates 6:607
premarital cohabitation 6:607
race–ethnic differences 6:607
socioeconomic status 6:606–607
creative potential, and marriage 14:602–603
demographics of 14:518
economics of 6:607–608, 8:829–834
consequences 8:831
efficiency 8:830, 8:833
policy issues 8:833
education 6:607–608
effects
on adolescents 21:842
on development 14:518–524
in Europe 6:606, 6:608
externalizing problems 14:519–523
family 8:767–770
and family history 8:742
and family law 8:780–784
internalizing problems 14:519–520, 14:522–523
level 6:116–119
levels and trends 6:606
and marital quality 14:512–516
post-divorce adjustment of children
child factors 14:523
interparental conflict 14:522
interventions for problems 14:521–522
parenting 14:522–523
stressful events 14:521–522
variability in outcomes 14:521–523
protective factors 14:521
responses of children to 14:520–521
risk, and child fertility 8:824–825

risks conferred by 14:518–520
and sexual activity of children 14:519
societal costs of 14:520
societal views of 14:518
stratification 6:608
substance use of children 14:519–520, 14:522–524
synthetic or hypothetical cohort measures 6:118–119
timing 6:119
trends, in Western societies 17:549–550
see also Marriage
Divorce Act (Canada) 20:468–469
Divorce rates 8:767, 8:782, 11:164
Divorce revolution 20:430
Divorced families, emotional development of children in 7:440
Divorced parents, academic achievement of children of 22:926
Diwa 11:791
Dix, Dorothea 25:625
Diyari, gender assignment in 9:795–796
DIY-research communities 23:329
Dizi, gender assignment in 9:796
Dizygotic (DZ) twins 15:204, 16:340, 21:779, 24:685, 24:690–691
Djaziri, Eyet-Chékib
works
Un "poisson" sur la balançoire 21:753
Une promesse de douleur et de sang 21:753
Djebar, Assia
works
Femmes d'Alger dans leur appartement 21:753
Ombre sultane 21:753
DJSGI *see* Dow Jones Sustainability Index (DJSGI)
DKA *see* Diabetic ketoacidosis (DKA)
D-KEFS *see* Delis–Kaplan Executive Function System (D-KEFS)
DL *see* Description logics (DL)
DLB *see* Dementia with Lewy bodies (DLB)
DLC *see* Developmental life-course (DLC) criminology
DLM *see* Digital Learning Material (DLM)
D-L-P formula 16:464
DLPFC *see* Dorsal lateral prefrontal cortex (DLPFC)
DM *see* Decision maker (DM); Dialogue management (DM)
DMC *see* Domestic material consumption (DMC)
D(H1,H2, A, chimpanzee) measure 11:293
DMI *see* Direct material input (DMI)
DMN *see* Default mode network (DMN)
DMNV *see* Dorsal motor nucleus of the vagus (DMNV)
DMRs *see* Differentially methylated regions (DMRs)
DMTS *see* Delayed-matching-to-sample (DMTS)
DNA *see* Deoxyribonucleic acid (DNA)
epigenetic inheritance 7:832–837
research, microevolution 15:338
DNA on loan 9:966
DNB Household Survey 5:846
DNGs *see* Deviant nonprofit groups (DNGs)
DNMTS *see* Delayed-non-matching-to-sample (DNMTS)
DNN *see* Deep neural network (DNN)
DNR *see* Do not resuscitate (DNR)
DO *see* Design optimization (DO)
Do no harm phase 24:838
Do not resuscitate (DNR) 7:597–598
Doable articulations 23:246
Doan, P., on LGBTI commercial sex 13:885
DOB *see* Daughter of Bilitis (DOB)
Dobbin, Frank
on industrial order 5:537
on theorization 17:392
Dobyns, H., on Native Americans 1:634
Docile bodies 8:900
Dockets 5:808
Docking 4:488–489, 4:491
DocStudy 7:281
Docta ignorantia 13:586–587
Doctor of physical therapy (DPT) 16:908–909
Doctoral education 6:610
changing policy and 6:611
current and emerging directions and issues in research 6:611
current knowledge 6:610–611
prevailing and emerging schools of thought 6:611
Doctoral studies, Bologna Process 2:742–743
Doctoral training 1:26–27
in Canada 1:27
examination 1:26
and habilitation 1:26–27
as interface between education and employment 1:26–27
in North America 1:27
in United States 1:27
Doctorate programs
educational research 7:280–281
degree production 7:281
preparation 7:280
Doctorates 22:681–682
for racial and ethnic minorities 22:682
in science and engineering 22:682–683, 22:683t
in social sciences 22:683
to women 22:682
Doctor–patient communication *see* Doctor–patient interaction
Doctor–patient interaction
adherence to medical advice 6:617–618
antecedents of 6:614–616
ethnicity, culture, and socioeconomic status 6:615–616
facilitation 6:614
gender 6:615
health care 6:613, 6:615–619
information giving 6:614
information seeking 6:614
Information-Motivation-Strategy model 6:617–618
interventions for 6:618
medically functional aspect 6:614
models 10:677–678
negative talk 6:614
nonverbal sensitivity 6:617–618
outcomes of 6:616–618
participatory aspect 6:614
partnership building 6:614
patient-centered care 6:613–614, 6:617
physical and psychosocial health 6:616
physician–patient communication 6:613–614, 6:618
positive talk 6:614
responsive aspect 6:614
satisfaction 6:617
social conversation 6:614
third parties and internet 6:616
in West 6:613–619
Doctor–patient relationship 15:122
impact on 8:68
Doctors' Trial 20:525–526, 21:316
bioethics 13:558
Doctrinal mode of religiosity 20:393
Doctrine of Discovery 11:795–796
Doctrine of Home Rule 24:910
Doctrine of multilevel analyses 22:486
Doctrines 5:252
Document Management and Electronic Archiving in IT-supported operations *see* DOMEA (Document Management and Electronic Archiving in IT-supported operations)
Document request, discovery 6:520–521
Documentary 9:164
defined 9:163–164
modes 9:164
Documentary film 6:620–624
affect 6:621, 6:623
audiences 6:620, 6:623
author as a fly in the soup – cinema verité 6:622
author as a fly on the wall – Direct Cinema 6:622
author as propagandist – Riefenstahl 6:622
author as reconstructing authentic other – Flaherty 6:621
author as reflexive roving eye – Vertov 6:621–622
author as social educator – Grierson 6:621
contemporary issues 6:623–624
expository mode of representation 6:620–621
versus fiction 6:620
genres 9:163–164
institutions 6:622
versus nonfiction 6:620
observational mode of representation 6:621
participatory mode of representation 6:621
performative mode of representation 6:621, 6:623
poetic mode of representation 6:621
politics of representation 6:621, 6:623
postcolonial 'other' as author 6:622–623
realism 6:621, 6:623
reflexive author 6:623

Documentary method, interpretation 13:116
Documentation 17:624–625
  cross-cultural research 7:932
Documents
  on human rights 11:356–357
  qualitative health research 10:694
  unobtrusive measures 24:768–769, 24:771–772
Dodd–Frank Wall Street Reform and Consumer Protection Act 4:936–937, 8:61, 15:733
Doers of program evaluations 19:117–118
Dogmatic method 11:31–32
Dogmatism 2:198, 2:257, 5:272, 5:274, 7:883–884, 21:262
Doha Development Agenda 12:569
DO-ID *see* Decision Oriented Instructional Design (DO-ID)
Doing gender concept 9:701–702
Doing identity, through crime, intersectionality in 12:666
DOJ *see* Department of Justice (DOJ)
Dokimasia 6:61–62
d'Olivet, Antoine Fabre 25:577
Dollar diplomacy 11:685
$1000 genomes 10:32
Dollard, John 12:11–12
  works, *Class and Caste in a Southern Town* 4:362
DOMA *see* Defense of Marriage Act 1996 (DOMA)
Domain convergence 23:896
Domain definition, in content validation 4:775
Domain integrity 25:650, 25:652–653
Domain knowledge, and communities of practice 4:328–329
Domain layering 23:896–897
Domain of international law and treaties 12:508
Domain representation 23:176
Domain restriction, color restriction 4:231
Domain-centered approach, of Sapir–Whorf hypothesis 20:905
Domain-general brain mechanisms 16:694
Domain-general maintenance mechanisms of working memory 21:904–906
Domains
  dualism of 15:546–547
    public and private cognition 15:546–547
Domain-specific competencies 21:120
Domain-specific creativity 17:772–773
Domain-specific knowledge, and intelligence 12:327
Domain-specific maintenance mechanisms, of working memory 21:903–904
Domain-specific models of passion 17:571–572
  entrepreneurial passion 17:571–572
  passion as part of the triangular theory of love 17:571
Domain-specific risk-taking (DOSPERT) scale 17:812
Domain-specific self-concepts 21:492
  developmental differences in 21:492–493
  individual differences in 21:494
Domain-specific *versus* content-free biases 5:390–391
Domain-specificity 6:431
  future orientation 9:563
Domain-unspecific models of passion 17:570–571
  com.pass model: passion as coincidence of commitment and desire 17:571
  dual model of passion 17:570–571
  harmonious versus obsessive passion 17:570–571
  passion as intrinsic, self-relevant absorption 17:571
Domashnii ochag (magazine) 20:839
DOMEA (Document Management and Electronic Archiving in IT-supported operations) 5:702
Domestic abuse, of elders 7:329
Domestic Abuse Program (DAP) 12:679
Domestic adoption 1:136–137, 1:142, 1:144
Domestic and intimate partner violence 10:194
Domestic and jural domains of kinship, distinction between 13:37
Domestic competition, multinational corporations 16:78–79
Domestic courts 13:821
Domestic extremists 16:247
  offending patterns and national security of in United States 16:251–252
Domestic family 11:228–231
Domestic household archaeology 11:220–221
Domestic law 10:258–259
Domestic legal commitments, and political repression 18:450
Domestic material consumption (DMC) 14:762
Domestic migration 11:649, 11:651
Domestic politics 5:371–373
  challenges 5:371
  corruption 5:371
  human rights 5:372–373
  race relations 5:371–372
  women 5:372
  world wide web, cellphones, and bloggers 5:373
Domestic processed output (DPO) 14:762
Domestic social problem 12:542
  family violence 12:542
Domestic space, safe 21:673
Domestic terrorism 6:554
  in United States 16:247–252
Domestic threats, and political repression 18:449–450
Domestic violence (DV) 6:633–636, 9:730, 9:733, 12:667–668, 20:253
  and child neglect 3:413–414
  and communities of faith 20:254
  definition 6:633–634
  and dowries 6:644
  and feminist legal theory 8:921, 8:923
  and feminist social work 8:938–939
  history 6:634
  homelessness 25:112
  honor killing 11:185–191
  institutional policies and practices 6:635
  intimate partner abuse *see* Intimate partner abuse (IPA)
  law 6:625–632
    arrest 6:626
    battered women's syndrome 6:625–627, 6:630
    civil actions 6:628
    civil approach 6:628–629
    Convention on the Elimination of All Forms of Discrimination against Women 6:629–630
    courts 6:627–628
    criminal aspects 6:626–628
    criminal justice 6:627, 6:629
    doctrinal changes 6:627
    gendered dimensions 6:630
    history 6:625–626
    and international human rights law 6:629–630
    international laws 6:630
    intimate partner violence 6:625, 6:627
    mandatory arrest 6:626–627
    prosecution 6:626–627
    protective order 6:628–629
    restraining orders 6:628–629
    tort 6:628
    US federal law 6:628
    vulnerability theory 6:625
  nature of problem, research methods, and findings 6:634–635
  overview 6:633
  and religion 20:319–320
  sociological explanations 6:633, 6:636
  theoretical explanations 6:635–636
  in United States 20:253
Domestic war, against terrorism 6:552
Domestic welfare in design of IP rules 13:521–522
Domestic work, and time use 24:369–371
Domestication
  Anthropocene 1:723
  of *Bos indicus* 16:840
  of camel 16:841–842
  cattle 16:840
  of fire 9:197–200
  and food production 9:300–305
  of herd animals 16:840
  of horse 16:841–842
  plant 16:840, 20:918
  of technology 9:725, 24:138–139
Domestic-based family firms 17:578
Domesticity 4:753
Domiciliary care, in England 4:345, 4:348
Dominance 22:264–265
  definition 15:92
  in established communities 4:315
  inequality 11:900–902
  language 16:58
  social dominance orientation 22:250–253
  sociology 22:897–898

Dominance approach to language and gender 13:256–257
Dominance feminism 9:730, 9:732
Dominance genetic variants 8:715
Dominance models, in forced-choice personality assessment 17:843–844, 17:847
Dominance principle 22:263
  and elections 18:417–418
Dominance reasoning 12:337
Dominance rule, of decision making 5:919
Dominant caste 3:223–224
Dominant category 6:640
Dominant culture 8:158
Dominant design 6:637–641
  boundary conditions 6:638–639
  emergence of 6:637–638
  entry timing advantages 6:639–640
  foundations 6:637
  limitations and critique 6:640
  organizational performance 6:640–641
  social, cognitive, and political factors in technology evolution 6:640
  technological paradigms 6:637, 6:640
  technology cycles 6:639
  window of opportunity 6:639–640
Dominant images, engineering 7:642, 7:646–647
Dominant Language Constellation (DCD) 16:68
Dominant model of science communication 21:222
Dominant narratives 4:726, 4:731
Dominant planning paradigm 24:867
Dominant-end conceptions 6:591–592
Dominated, relations 15:136–137
Domination 9:402–403, 17:203, 20:478
  and authority 2:271–272
  Boltanski on 2:750–752
  hegemony 10:748–749
  and ideologies 11:573–574, 11:576–577
  power and authority 17:407–408
  theory, Pierre Bourdieu 2:778
Dominica, carnival in 3:147
Dominican Republic 1:639, 22:624–625
Dominion period, potlatch in 18:715
Dominium 19:190, 19:192, 19:195
Donald Campbell
  works
    Experimental and Quasi-Experimental Designs for Research 3:247
    Quasi-Experimentation: Design & Analysis Issues for Field Settings 3:247
Donation 15:738, 22:580
Donauwörth conflict 20:102–103
Donders, Franciscus Cornelius 12:77, 14:810–811
Donnison, David 11:234
Donor nations, and structural adjustment 23:601–602
Donor value 5:49
Donor-language identity 3:919
Donors 22:579–580
  emerging 3:506
Don't Ask, Don't Tell (DADT) 15:514
Don't know survey responses, in Slovenian plebiscite, multiple imputation for 16:92
Donze, Patti, on gender role in popular music 3:801
Doob–Meyer decomposition 14:632
Doob, J.L
  maximal inequalities 14:631, 14:633
Dooley, Lucille 25:625
Doolittle, Myrick Hascall 9:625
DOPA *see* 3,4-Dihydroxy phenylalanine (DOPA)
Dopamine 9:1012, 16:749–752, 16:754, 20:212–213, 21:611–612, 25:351–354, 16:750f, 16:752f
  brain stimulation reward 2:842–845
  neuroendocrinology 16:651
  and placebo effect 18:154
Dopamine hypothesis 19:587–588
Dopamine receptor D4 gene (DRD4) 21:611, 22:495
Dopamine–norepinephrine–opioid peptide–cannabinoid theory of reinforcement 20:213
Dopaminergic system 9:953
  and ADHD 2:159
Dopamine-ß-hydroxylase (DBH) 16:749
Dopo lavoro 16:257
D-optimality 17:257, 17:259
  criteria of 17:257
Doreian, P., on negative social network ties 5:561
Dorn, Harold 21:236
Dorsal and ventral streams, integration of 25:205–206
Dorsal cortical streams 17:129–131
Dorsal lateral prefrontal cortex (DLPFC) 4:676, 16:640, 16:847, 18:817, 18:819–820, 21:421, 22:80–81
  serial order memory 21:639–643
Dorsal motor nucleus of the vagus (DMNV), interoception 12:614–615
Dorsal pathway 20:3
Dorsal premotor cortex (PMd) 5:25
Dorsal stream 23:245–247, 25:187, 23:247f
  activity 25:205–206
  and attention 2:163–166
  representation of action sounds 23:246
  for sensorimotor integration and control 11:324–326
  unified function o 23:246–247
Dorsal systems 18:61
Dorset
  art 1:923–924
  houses 1:923–924
  societies 1:923–924
Dorsolateral nucleus of the thalamus (DLM) 2:692
Dose, community-based health interventions 10:648
Dose–response relationship 16:199
DOSPERT *see* Domain-specific risk-taking (DOSPERT) scale
Dotson, Gary 10:455
Double agent problem, psychiatric ethics 8:66
Double auctions 2:208–209
Double circulation, and cultural studies 1:937
Double consciousness
  of DuBois 12:658–659
  and intersectionality 12:658–659
Double contingency, Luhmann on 14:390–391
Double dissociation model of developmental dyslexia 6:730
Double helix 9:990–991
Double high-threshold (DHT) model 6:524
Double jeopardy, women offenders 25:616–617
Double negation (grammar) 13:283
Double reflexivity 13:556
Double refutation 1:82
Double Relocation Policy 23:211
Double subject constructions 18:632
Double Trouble in Recovery (DTR) groups 10:689
Double-blind procedures 8:573
Double-blind randomized controlled trials 8:561
Double-blind review 12:863
Double-blind studies, placebo-controlled 18:157–163
Doubled vision of society 9:768
Double-difference designs *see* Comparison group
Double-loop learning 7:736
Doubleness 4:221
Double-sided markets 3:40
Doublets
  in spelling 23:268–269, 23:272
Doubling 5:4–5
  Doubling of grammatical relations 25:5
Doubly robust 15:604
Doubt
  and pragmatism 18:809
  and rumors 20:790–791
Douglas, Mary (née Tew) 5:499–500, 6:245, 13:71, 14:435–437
  on humor 11:416–417
  works
    Purity and Danger 1:769, 20:309–310
    The Two Bodies 2:729
    The World of Goods: Towards and Anthropology of Consumption 7:16
Doux commerce 7:664
Dow Jones Sustainability Index (DJSGI) 4:928
Dower 6:643–644, 19:173
Down syndrome (DS) 5:356–357
  joint attention in children with 12:846–847
Downer, Carol 15:111–112
Downey, Greg, works, *Learning Capoeira* 21:619
Downing, Lisa, on prejudice against women 21:748
Downs, Anthony, works, *An Economic Theory of Democracy* 18:484

Downshifting 4:760–761
Downsian mobilization 1:359
Downsizing 4:904, 7:62, 10:258, 18:664, 18:667
Downstream field experiments 9:130
Downtown 24:927, 24:932
Downward, P.
  works
    The Economics of Professional Team Sports 23:318
    Sports Economics: Theory, Evidence and Policy 23:318
Downward comparison principles in social psychology 22:212
Downward social comparison 10:430
Downward spirals 2:867
Dowries 6:642–645, 14:601
  arranged marriage 6:642
  brideprice 6:642
  complex societies 6:642
  determinants of 6:643
    brideprices 6:643
  domestic violence 6:644
  groomprice 6:643–644
  inheritance 6:643–645
  marriage contract 6:642
  marriage legislation 6:644
  marriage payments
    family characteristics as determinants of 6:643
    rise in 6:644
    social characteristics and occurrence of 6:642–643
    and women welfare 6:644–645
  population growth 6:642, 6:644
  prevalence of 6:642
  property rights 6:643–644
    brideprice 6:643–644
    dowry 6:644
  South Asia 6:642–645
  sub-Saharan Africa 6:642–643
Dowry deaths 6:644–645, 10:193, 10:195
Dowry murder 10:195
Dowty's treatment of lexical semantics 13:945
Doxa 18:783
Doxastic social connection 21:879
Doxepine 1:779
DP *see* Determiner phrase (DP); Developmental prosopagnosia (DP); Dynamic programming (DP)
DPAC *see* Disabled People against Cuts (DPAC)
DPI *see* Department of Public Information (DPI); Disabled Peoples; International (DPI)
DPIC *see* Death Penalty Information Center (DPIC)
DPMs *see* Dual-process models (DPMs)
DPO *see* Domestic processed output (DPO)
DPOs *see* Disabled peoples organizations (DPOs)
DPPBs *see* District Policing Partnership Boards (DPPBs)
DPSIR *see* Drivers-pressures-state-impact-response (DPSIR) scheme
DPSIR model of social ecology 22:259
DPT *see* Doctor of physical therapy (DPT)
DQA *see* Deductive qualitative analysis (DQA)
Draft system 23:320
Drag, and queer anthropology 19:755
Drag culture, LGBT 13:875–878
  community 13:876–877
  drag queens, drag kings, and faux queens 13:875
  early history 13:875–876
  political potential of 13:877–878
Drago, Luis M., works, *Los Hombres de Presa* 5:239
Dragonetti, Nicola C., on epistemic cultures 7:878
DragonflyIV 6:897
Drake, Durant, works, *Essays in Critical Realism* 20:21
Drake, S.C., on race in North America 17:23
Drama 17:710
  and alienation 1:540
Dramatism 6:646–647
Dramaturgical analysis 6:646–650
  actable idea 6:646–647
  contemporary applications 6:648
  contribution of academic fields to 6:647
  critiques and defenses 6:648–649
  digital age 6:649
  dramatism 6:646–647
  earlier conceptualizations 6:647–648
  future directions 6:649
  image 6:646–647
  impression management 6:648
  main concepts 6:646–647
  meaning 6:646–647, 6:649
  performance 6:646–649
  reality construction 6:646
  script 6:646–647
  social role 6:646
  stage 6:646
  symbolic interaction 6:647–648
Dramaturgical theories 22:958–959
Dramaturgy 10:276, 18:249, 22:239, 23:857–858, 24:636
  and metaphor 22:422
DRAM/EMPAL model 24:890–891
Draper, John William 20:284
  works
    History of the Conflict between Religion and Science 20:284
    Human Physiology 18:83
Draper–White thesis 20:284
Dravidian languages 13:340–341
Draw a Person: Screening Procedure for Emotional Disturbance (DAP:sPED) 16:725, 19:132–133
Drawing-as account of visualization 25:222
DRC *see* Dual route cascade (DRC) model
DRD *see* Declaration on the Right to Development (DRD)
DRD4 *see* Dopamine receptor D4 gene (DRD4)
DRD4 allele 15:340
Dream age 4:182–183
Dreaming 6:651–657
  definition and phenomenology 6:651–652
  generation and maintenance of dreams 6:655
  and idea generation 24:722
  neural correlates 6:652–655
  neuroanatomical studies 6:653–654
  neuroimaging studies 6:652–653
  neurophysiological and pharmacological studies 6:654–655
  and unconscious mental activity 24:723
Dreaming (Australian aboriginal worldview) 2:236
Drew Associates, *Primary* 6:622
Drexler, Anton 16:257
Drexler, K. Eric 16:188
Dreyer, Carl, works, *La Passion de Jeanne d'Arc* 9:167
Dreyfus, Alfred 2:380
Dreyfus affair 12:273–275
Drift load 24:133
Drift rate 5:903, 6:397
Drift rate variability
  additive 5:904
  Fréchet on 5:904
  multiplicative 5:903–904
Drill and practice software, Computer Assisted Instruction (CAI) 4:494
Drilling process 17:166
Drinking
  anthropology of 6:658–662
    consumption 6:658–660, 6:662
    culture 6:658–662
    current trends (1990–present) 6:660–662
    early approaches (1930–1960) 6:658
    ethnography 6:658–661
    everyday life 6:659, 6:661
    globalization 6:658, 6:660, 6:662
    health 6:658–662
    history of anthropology 6:658–659
    identity 6:658–662
    social construction 6:660
    social problems 6:658–659
    transformations (1960–1990) 6:658–660
  behavioral expressions 21:609–610
  bivariate model for 14:331
  four-factor model for 21:609
Drinking behavior
  multiple imputation for 16:91
  and sexual development 14:330
Drinking water, in Africa 1:297f
Drive, and instinct 12:177, 12:179
Driver alertness 24:625
Driver risk taking, attitudes, motivation, and emotion 24:625–626
Driver selection 24:626
Drivers of adult mortality 1:179–180
Drivers-pressures-state-impact-response (DPSIR) scheme 13:243f
Driving, behavioral expressions 21:609
Driving ability, and alcohol 1:510

Driving under the influence (DUI) 4:373, 15:593, 18:881
diversion 4:373, 4:375
Driving While Intoxicated (DWI) incidents 18:881
DRM *see* Day reconstruction method (DRM); Deese–Roediger–McDermott (DRM) list-learning task
Droit moral 12:263
Droits de l'homme 11:367–368
Drones 23:727–729, 25:379
in war 25:379–381
Drosophila, and memory 9:905–906
Drosophila melanogaster 9:879–881, 9:880f–881f, 9:883f–884f
Drought 3:865–868
Droysen, Johann Gustav 10:785–786, 10:873, 11:381, 18:625
DRR *see* Disaster risk reduction (DRR)
Drucker, Peter F. 22:699
on knowledge 13:108
on knowledge worker's productivity 6:664–665
on management as discipline 6:663–664
on practice of management 6:665–667
entrepreneurship 6:666
innovation 6:666
management by objectives 6:665
management ethics 6:666
management means communicating 6:665
managing 6:666
managing complexity 6:666–667
networks 6:666
strengths 6:665–666
systematic abandonment and change 6:665
weaknesses 6:665–666
works
The End of the Economic Man to Managing in the Next Society 6:663
The New Realities 6:663
Practice of Management 6:663
Druckman, J.N., works, *Cambridge Handbook of Experimental Political Science* 8:542
Drug Abuse Resistance Education (DARE) program 7:270–271, 18:881
Drug action, neurotransmitters 16:753–754
Drug addiction, wanting/needing 25:351–354
Drug Addicts Treated On Probation *see* DAYTOP (Drug Addicts Treated On Probation)
Drug control, international aspects 6:669–673
Drug court 4:374–376
Drug decriminalization or legalization 6:674
benefits 6:678
consensual nature of drug offenses and 6:677–678
definition of drug 6:674
of drugs 6:674–675
Drug dependence, for therapeutic community 24:278
Drug Enforcement Agency 21:107–108
Drug holidays 19:201
Drug misuse, and child neglect 3:413
Drug mules 25:618
Drug patents 10:618
Drug trafficking 12:465–466, 15:42–43
Drug use/abuse 1:517, 1:519, 1:521, 5:160–161, 5:177–178, 6:669
addiction 1:97–100
arm's-length analyses of the development 6:672
behavioral expressions 21:609–610
constructivist perspectives 6:672
and crime 23:495, 23:497
and homelessness 11:161
and homicide 11:166
intermittent, and impact 6:670
psychoactive substances 6:670–671
social patternings of 6:669–670
psychosocial aspects 6:685–689
attitudes and beliefs about 6:687–688
and dependence 6:687
distinguishing characteristics of drug-using behaviors 6:685
future directions 6:688
illegal drugs 6:686
illicit *versus* licit drug use 6:685–686
initial trajectory of drug use 6:686–687
intervention 6:686, 6:688
legal drugs 6:686
perceived risk 6:687–688
polydrug use 6:687
problem behaviors 6:686–687
psychoactive drugs 6:685
psychotherapeutic drugs 6:685
relapse 6:686
research traditions 6:688
social connotations of use 6:687
stages of involvement 6:686
treatment 6:685, 6:687–688
religion and 20:264
resistance education 10:671
social concerns about problems 6:670
social science analysis 6:671–672
Drug-related offenses 6:677–678
relation with other criminal activity 6:677
Drugs 5:160–161
Becker on 2:403–404
biomedicine 2:665–667
decriminalization of 6:674–678
illicit use and prevention 6:679–684
production 5:161
and school safety 21:110
that were legal in the United States 17:239–240
and transnational organized crime 24:597, 24:599
use of term 6:685
and victimless crime 25:75–79
Drummond, Leo 1:641
Drury, S., on lesbian athletes 9:659
Dry sex 21:799
Dryden text 13:266–267
Dryzek, John 5:300
DS *see* Developmental stutterers (DS); Down syndrome (DS); D(eep)-Structure (DS) representations
DSBM *see* Diploma of School Business Management (DSBM)
DSD *see* United Nations Division for Sustainable Development (DSD)
D-separation (directed acyclic graphical models) 2:363–364
DSGE *see* Dynamic; stochastic general equilibrium (DSGE) model
DSGE models, equilibrium 7:917
DSI *see* Diffusion spectrum imaging (DSI)
DSM *see* Design structure matrix (DSM); Diagnostic and Statistical Manual of Mental Disorders (DSM)
DST/CT *see* Dynamic systems theory/complexity theory (DST/CT)
D(eep)-Structure (DS) representations 16:1
DTC *see* Democratic therapeutic community (DTC); Direct-to-consumer (DTC) personal genetic testing
DTCA *see* Direct-to-consumer advertising (DTCA)
DTD *see* Development Trauma Disorder (DTD); Developmental topographical disorientation (DTD)
DTFR *see* Desired total fertility rate (DTFR)
D-TFR *see* Destination-specific total fertility rate (D-TFR)
DTI *see* Diffusion tensor imaging (DTI)
DTR *see* Double Trouble in Recovery (DTR) groups
DTSQ *see* Diabetes treatment satisfaction questionnaire (DTSQ)
Du Bois, William Edward Burghardt 1:287, 6:690–692, 8:137–138, 12:658–659, 17:541, 20:540, 25:571
applied social research 1:853
brief biography 6:690
as Director of Publications and Research at NAACP 6:691
education 6:690
on ethnic groups 8:137–138
interest in Marxism 6:691
legacy of 6:691–692
political disagreements and consequences 6:691
scientific contribution 6:691
understanding of African-American history 6:692
at Wilberforce College 6:690
works
Black Reconstruction In America 6:691
Dark Princess 6:691
In Battle for Peace: The Story of My eighty-third Birthday 6:691
The New Negro anthology 6:691
The Philadelphia Negro 4:362, 6:690–691, 7:560, 9:135–136
The Souls of Black Folk 6:690–691
The Suppression of the African Slave Trade to the United States of America 6:690
Dual actor failure

Dual actor failure (*continued*)
global public policy 10:179
evidence 10:181
Dual apprenticeship system, for vocational education and training 25:248
Dual center theory, for eating 11:420
Dual citizenships 3:657–658
Dual consciousness 22:852
Dual Control Model (DCM) of sexuality 21:850–851
Dual criminality 13:533
Dual diffusion model 6:400
Dual economy models 1:477
Dual explanandum strategy 15:174
Dual Group Processes Model (DGPM) 2:259
Dual identity 4:147, 12:411
Dual inheritance theory 11:247
Dual language education 2:581–584
Dual lattice analysis 5:560
Dual loyalty continuum 1:794
Dual mechanism theory 8:651
Dual mode model of decision making 19:908
Dual nature of law 13:481
Dual number 17:44
Dual offenders 20:438
Dual perspective, international social work student exchanges 12:563
Dual route cascade (DRC) model 6:736–737
Dual system 17:142
of vocational training 1:870–872
Dual taxonomy theory 5:138, 14:60
Dual track transitional system, in China 6:790
Dual Vocational Training System 1:862
Dual-aspect panpsychism, Wright's concept of 25:760
Dual-career couples 14:607, 14:609, 21:37–38
Dual-career families 14:605–610
breadwinning 14:606, 14:608
with children, strains of 14:607
couples
earning arrangements among 14:608
work schedules among 14:608–609
division of labor inside home 14:606–607
father involvement 14:609
housework 14:606–609
increases in women's employment 14:606
nonstandard work schedules 14:609
time pressure and total hours of family work 14:607
time with children 14:607
work–life balance 14:607–608
Dual-earner couples 14:608–609
Dual-earner families, and labor participation 8:758, 8:760
Dual-foci, and psychosocial work 19:478–482
Dual-frame sampling 16:920
Dualism 2:61, 6:334, 7:311–312, 9:430–431, 12:509, 13:142, 13:700–701, 19:409, 21:531
Cartesian, essential attribute of mind 24:722
and chemistry 3:372
mind–body 15:545–548
and morality 15:807–808
substance 18:107–108
Dualistic world order 13:821
Dualists 7:311
Dual-motives theory 15:951–952
Dual-process models (DPMs) 2:259–260, 2:535
of assimilation and accommodation 14:103–104
cognitive model 2:443
of information processing 22:207–208
of learning and memory 5:579
of recognition memory 20:39–40
and social network actors 5:559
Dual-process theories 1:122, 5:553, 7:416–418, 10:482, 13:464–465, 19:63
of human reasoning 12:722–723
implicit social cognition 11:717–718
of lifespan cognitive development 12:294
of memory 23:30
versus source monitoring framework 23:30
Dual-route model 6:732, 6:736–737
of reading 6:732
of spelling 6:736
Dual-task experiments 25:176–177
Dual-use problem 2:629
Dual-use research 3:280–281, 20:525
ethics of censoring 3:281–282
Duany, Andrés 16:809
Dubois, Pierre 4:571–572
Dubos, René 6:693–695, 15:104
antibiotics research of 6:694
standardization of BCG vaccines 6:694
vaccination against tuberculosis 6:694
biography of 6:693
biopsychosocial perspective on health 6:694
career of 6:693
culture techniques 6:693
microbial investigation 6:693
as French agricultural engineer 6:694–695
on human health 6:694
Ph.D. in soil microbiology 6:693–694
works 6:694
Mirage of Health: Utopias, Progress, and Biological Change 6:694
Only One Earth 6:694
So Human an Animal: How We are Shaped by Surroundings and Events 6:694
The White Plague: Tuberculosis, Man, and Society 6:694
Duby, George, works, *L'économie rurale et la vie des campagnes dans l'Occident medieval* 17:671
Duby, Georges, contribution to feudalism 9:113–114
Duchamp, Marcel, avant-garde art 2:320
Ducker, Peter
conception of knowledge society 13:106–107
works
The Age of Discontinuity 13:106
Duden, Barbara 2:732
works
Woman in the Body: A Doctor's Patients in Eighteenth Century Germany 9:771
The Woman Beneath the Skin: A Doctor's Patients in Eighteenth-Century Germany 2:729–730
Due credit assignment, archiving 1:919
Due diligence 12:910
Due process 6:696, 12:137, 13:764, 13:808–809, 17:718
in American legal system 6:696–697
basic elements of 6:697
and crime policy 5:176, 5:179
functionalism 6:698
German conceptions 6:698–699
in other national legal systems 6:698–699
post-Civil War amendments 6:697
procedural 6:697–698
protection of personality rights and physical integrity 6:698–699
substantive 6:697
Due process in law 6:696–699
Dueling, and honor 11:181–183
Duffy, S., works, *In Control* 4:346
Duguit, Léon 13:802, 18:467
Duhem, Pierre 7:569
Duhem–Quine thesis 8:576–577, 14:317, 21:308–309
Duhem's problem 5:273
Duhl, Leonard 24:867
DUI *see* Driving under the influence (DUI)
Dujiangyan irrigation system 16:399
Dukes v. Wal-Mart 13:594
Duloxetine 17:470
Duluth Domestic Abuse Intervention Project (DAIP) 12:682
DUM *see* Discounted utility model (DUM)
Dumais, Susan, on cultural capital 5:377–378
Dumb peasant theory 17:671
Dummy player axiom 4:872–873
Dumont, Etienne 2:529
Dumont, Louis Charles Jean 1:768, 8:290–291
works, *Homo Hierarchicus* 1:768, 6:703
Duncan, Otis Dudley 11:283, 20:816
Dunedin Multidisciplinary Health and Development Study 15:183
Dunham, Katherine 9:460
Dunlop, Charles E.M., works, *Glossary of Cognitive Science* 7:590
Dunlop, John T. 11:861
Dunn, John 3:62
works, *The Identity of the History of Ideas* 3:62
Dunner, Joseph, works, *Dictionary of Political Science* 7:591
DUP *see* Duration of untreated psychosis (DUP)
Duplex theory of love 14:379–380
Dupront, Alphonse 5:422
Duquenne, V., on social network analysis 5:560

Durable Inequality 16:787
Durand, R., on categories 15:576
Durand's classifications of visual tropes 20:651
Duras, Marguerite 1:540
Duration analysis 14:52, 14:57
in political science 18:515, 18:518–519
Duration at destination-specific fertility rates (DD-FRs) 15:399, 15:404–405
Duration of emotions 7:526
Duration of untreated psychosis (DUP) 19:201
Durbin–Watson test 24:341
Duress 3:28, 5:224–225
Durkheim, Émile 1:75–76, 1:540, 1:636, 4:338–341, 4:405–406, 4:524, 5:499, 6:241, 6:700–705, 10:214, 10:276, 10:485, 10:490, 11:865, 12:389, 13:112–113, 13:150, 13:551, 13:566, 13:586, 13:802–803, 14:414, 14:436, 14:825, 15:311, 18:472, 18:625, 18:939–940, 19:49, 19:125, 19:677, 20:269, 20:346, 20:359, 20:443–444, 21:261, 21:963, 22:828, 22:992, 22:996–997, 23:2–3, 23:683
on anomie 1:718–720
applied social research of 1:851–852
biography of 6:700–701
classic study of suicide 15:305
concept of anomie 1:714–715
contributions 6:701–703
religious sociology 6:704
social realism 6:701
sociology of education 6:703
sociology of law 6:704
and cosmopolitanism 5:41–45
on culture–cognition link 5:576–578
current significance of 6:703–705
definition of social facts 15:575
Durkheimian problem structure 19:49–50
on economy and culture 5:535
on Enlightenment 7:664–665
on functionalism 9:536–537
on habit 10:476–478
impact and influence of 6:703–705
on knowledge 13:71–73
on mechanical and organic solidarity 13:566
on nationalism 16:291
notion of the normal 13:586
on ontological collectivism 11:818
on political sociology 18:473
on religion 20:307, 20:309–310
on sacrifice 20:860–861
scientific ideas of 11:383
on social authority 2:269–270
on social causes 6:244–245
on social facts 22:991–992
on social structure 22:721
on sociology 22:996–997
on solidarity 23:6
on suicide
critiques and extensions 23:683–684
empirical tests 23:684–686
suicide statistics on 23:424–425
on symbolic boundaries 23:850
theorization of society 4:1–2
theory of suicide 23:676
and Weber 22:992
works
The Division of Labor in Society 2:82, 4:361–362, 5:498, 6:244, 6:601–602, 6:700, 13:162, 13:566, 13:802–803, 23:3
The Elementary Forms of the Religious Life 1:525, 1:748, 5:498, 6:700–702, 7:665, 11:818, 13:70, 16:291, 17:391, 20:307, 23:3
The Gift 6:703
The Normal and the Pathological 13:150
Primitive Classification 5:498, 13:70, 14:826
Professional Ethics and Civil Morals 5:535
The Rules of Sociological Method 6:244, 6:700–701
Suicide 1:714–715, 6:700, 6:702, 20:269, 22:192, 22:998
Durkheimian holism 15:303
Durkheimian individual 17:932–933
Durkheimian sociology 21:129
Durkheimian tradition of studying forms of classification 2:778
Durkheimians 6:706–708
D'Urville, Jules Dumont 17:158
Dutch auction 2:205–208
Dutch Book theorem 19:24–25, 19:39
for conditional probability 19:39
Dutch Carnival 3:146–147
Dutch colonialism 11:1
Dutch customary law 3:670
Dutch De Groenen 10:380
Dutch educational system 8:132
Dutch Golden Age 4:752–753
Dutch Potato Famine 1846–1847 6:753
Dutch public service system 2:861
Dutch University of Leiden 18:466
Duties 13:753, 17:144
definition 13:5–6
ethics of 13:5–8
and responsibility 20:581–582, 20:587–588
Dutta–Ray value, DR 4:869, 4:871
Dutthagamani, king 2:890
Duverger's law 7:351–352
DV *see* Domestic violence (DV)
DVAs *see* Deviant voluntary associations (DVAs)
DVD *see* Digital video disk (DVD)
DVI *see* Disaster victim identification (DVI)
Dvorak keyboard 17:8
DVRs *see* Digital video recorders (DVRs)
DVU *see* Deutsche Volksunion (DVU)
DWADE *see* Density-weighted average derivative estimator (DWADE)
Dwarfism 9:1019
DWC *see* Division on Women and Crime (DWC)
Dwelling, and motivation/actions 15:889–890
DWI *see* Driving While Intoxicated (DWI) incidents
Dworkin, A., on pornography 13:884–885
Dworkin, Gerald 2:316
Dworkin, Ronald 12:932, 13:764–765, 13:779–780
on legal philosophy 18:24, 18:26–28
theory of rights of 20:683–684
on welfare liberalism 9:399–400
Dyad 16:546
politeness from point of view of 18:328–329
Dyad social network 22:461–462, 22:462f
Dyadic conflict 4:639–640
Dyadic coping 11:593–595
as coping congruence between partners 11:593–594
depression 11:594
as interpersonal process 11:594–595
mental disorders 11:594
physical illness 11:594
we-disease 11:595
joint coping 11:594
as relationship-focused coping 11:593
social support 11:593
stress 11:593–594
theoretical concepts 11:593
Dyadic interfirm linkages 16:549
Dyadic intimacy process model 14:376–377
Dyadic leadership 13:641, 13:644
Dyadic structure in networks 22:468–469
Dyadic trust 24:665–666
Dyckman, John 18:204
Dying 5:860
and childhood cancer 3:461–462
psychology of 5:860–864
sociology of
see also Death
Dyirbal, gender assignment in 9:796
Dynamic, stochastic general equilibrium (DSGE) model 14:404
Dynamic approach to school improvement (DASI) 21:93–94
Dynamic assessment
of intellectual functioning 12:252–253
of team constructs 24:102
Dynamic Bayesian models 2:371
Dynamic capabilities 1:16, 23:509, 23:522
and absorptive capacity 1:15–17, 1:17f
integration into conceptualization 1:17, 1:13f, 1:16t
strategic adaptation 23:504
Dynamic constructivism theory 5:551
Dynamic control theories 4:836–837
Dynamic decision making (DDM) 6:709–714
alternative explanations for human performance 6:710–711
laboratory *versus* naturalistic decision research 6:713
learning to control dynamic systems 6:711–713
problem solving 6:710, 6:713

Dynamic decision making (DDM) (*continued*)
stochastic optimal control theory 6:709–710
Dynamic dimension of language 10:109–110
Dynamic economic systems 15:294–301
Dynamic equilibrium 7:912, 7:914–915, 17:886
Dynamic factor models 24:358
Dynamic fertility models 8:823–824
Easterlin's relative income hypothesis 8:824
life-cycle dynamics 8:823–824
Dynamic graphics, interactive displays 10:354
Dynamic linear model 24:354–358
Dynamic logics 13:103, 16:925, 16:927–928
of information flow 16:927
of natural language 16:927
Dynamic mapping
concept of 6:715
in geography 6:715–717
Dynamic model 22:471
of educational effectiveness 21:93
of endogenous interpersonal influences 15:659–666
family and household demography 8:728–729
of multilingualism 16:69
of school effectiveness 21:83
Dynamic Montague Grammar 15:754
Dynamic multiple social networks 22:478
algebraic structure of 22:481–483
Dynamic nature of education 7:218–219, 7:225–227
Dynamic network analysis 4:488–489
Dynamic person 17:764
Dynamic process 21:349
tradition as 24:503–504
Dynamic programming (DP) 6:710, 19:432–434
algorithms 23:262
Dynamic representation 18:442
Dynamic skill theory 16:505–508
Dynamic spatial phenomena 6:715
Dynamic strategies, in time-varying treatments 14:341
Dynamic systems approach 4:88–89, 12:344–345
developing through relationships 12:345
impact of 12:344–345
of socio-emotional development 7:521
Dynamic systems theory
personality development 17:880
prenatal/infant development 18:855
Dynamic systems theory/complexity theory (DST/CT) 16:69
Dynamic theories 14:192
Dynamic treatment strategy, in time-varying treatments 14:341
Dynamic TU games 4:879
Dynamic work 25:495–496
Dynamical systems 15:971, 21:530, 22:701
Dynamically inconsistent preferences, modeling 2:447
Dynamics
group 10:411–412
culture and 10:412
intragroup conflicts 10:412
stages 10:411–412
social network 22:475
structural balance theory 23:607–608
of time series in political science 18:517–518
Dynamism, of scientific disciplines 21:289
Dynastic chronicles
of Dai Viet 11:82
of Javanese 11:82
in Thai 11:82
Dynastic empires 16:297
Dynasties, Southeast Asia 11:78–79
Dynorphin neurons 20:212
Dys-appearing infrastructure 12:120–122
Dyscalculia 16:732–733
Dysexecutive amnesia 6:718–724
confabulation 6:719–720
disorders associated with 6:721–722
frontal amnesia 6:718–720
impaired emotion-based learning 6:721
impaired frequency judgments 6:721
impaired recall 6:719
impaired working memory 6:720
increased false alarms 6:720
Korsakoff's amnesia 6:722
memory impairments 6:719–721
metamemory impairment 6:720
perseveration 6:720–721
remedies 6:721
rupture of anterior communicating artery aneurysm 6:722
schizophrenic memory impairment 6:722
source amnesia 6:719
susceptibility to proactive interference 6:720
temporal discrimination impairment 6:719
vascular thalamic amnesia 6:722
Dysexecutive syndromes 4:33–34, 6:725–727
maintaining executive control 6:725–726
cognitive flexibility 6:726
disinhibition 6:725–726
ecological validity of assessment 6:726
frontal lobes 6:725–726
inhibition 6:726
initiation 6:725–726
neurologic disease 6:725–726
neuropsychological tests 6:725–726
planning 6:725
set-shifting 6:725
Dysfunction
childhood cancer 3:462–463
and normality 17:2, 17:4
olfactory system 17:199–200, 17:199t
Dysfunctional emotion regulation 7:496
Dysgraphia 1:450, 6:732, 12:420, 16:685
acquired 1:450–455
Dyshyponoia 2:576
Dyslexia 1:818–819, 2:810, 3:329, 4:90–91, 4:93, 6:732, 10:513, 12:420, 14:196, 16:685, 16:730–732, 20:7–8
acquired 6:730
in alphabetic languages 20:8–9
and asymmetry of body/brain 2:115–116, 2:125–126
developmental 6:728–731
abnormal development of phonological representations in 6:729
auditory processing problems 6:729
brain imaging studies 6:729–730
discrepancy definition 6:730
genetics of 6:730
normal development of lexical representations 6:728–729, 6:728f
phenotype of 6:729
phonological processing deficits in 6:729–730
subtypes 6:730
and language learning impairment 13:344–345, 13:347–348
visual word recognition for 25:217
Dyslexic individuals 9:483–484
Dyslipidemias
schizophrenia 21:23–24
Dyslogias 16:683
Dysosmia 17:199t
Dyspareunia 21:691–692, 21:716, 21:718
Dysphasias 16:683
Dysphoria 17:873
Dysphrasias 16:683
Dyspnea 20:577–578
Dysregulated fear (DF) 24:195
Dysregulation 21:553, 21:555
and adolescent health 1:123
Dyssomnia 22:79
Dysthymia 2:684
Dystonia 16:718
Dystopia 18:671–672
Dystopian 14:256–257
DZ *see* Dizygotic (DZ) twins

E

E professionalism 4:288
EA *see* Error analysis (EA)
EAB *see* Experimental analysis of behavior (EAB)
EAD *see* Encoded Archival Description (EAD)
EAEPE *see* European Association for Evolutionary Political Economy (EAEPE)
EAHCA *see* Education for All Handicapped Children Act (EAHCA)
EAI *see* East Asian Institute (EAI)
EAIR *see* European Association for Institutional Research (EAIR)
EAM *see* Early anatomically modern humans (EAM)
EAO *see* European Audiovisual Observatory (EAO)
EAR *see* Electronic activated recorder (EAR)
Ear knowledge 21:617

Earl of Hillsborough 1:610
Earle, Carville, works, *Geographical Inquiry and American Historical Problems* 11:18
Early abilities
numerical abilities 4:45–46
reasoning abilities 4:45–46
Early adulthood, structural brain development 16:600–601
Early adversity 23:581–582
lasting effects of 23:582
Early anatomically modern humans (EAM) 2:725–726
Early anthropogenic hypothesis 1:722–723
Early Boston Diagnostic Aphasia Exams 18:55
Early childhood 4:533, 21:279
defiant behavior during 6:12–18
development, effect of stress on 23:577–582
development of mathematical concepts during *see* Early mathematics
development of self-regulation during 21:553–556
education and care
family policy 8:801–802
imitation during 11:621–627
and media 15:2
motor development in 15:971–977
spatial cognition development during 23:116–125
Theory of Mind development during 24:268–273
Early Childhood Care and Development (ECCD) Law 6:743
Early childhood care and education (ECCE) 6:739–745
context setting 6:739–740
diversity 6:740
education for all (EFA) 6:740
exclusive breastfeeding as key communication actors, promoting 6:744
global development 6:739–741
mobile early childhood care and development programs 6:743–744
policy 6:739–741
quality 6:740–742
teacher education 6:741
in twenty-first century
Moscow Framework for Action and Cooperation 6:740–741
recent developments and global concerns 6:740
teacher preparation in vanuatu 6:741–742
UNESCO 6:740–741
UNICEF 6:743–744
Early childhood development (ECD) 6:739
Early childhood education and care (ECEC) 8:802
Early Childhood Longitudinal Studies (ECLS) Program 5:380, 5:696–697
Early children's socialization, in information society 4:667
Early Christianity 3:233
Early Developmental Stages of Psychopathology Study 15:183
Early education
effects on life span development 7:138–141
Early firm growth, alternative perspectives on 17:367–368
Early Head Start program 3:397
Early historic period, South Asia 23:54–56
Early industrial psychology, human resource management 11:350
Early Jesus groups 3:545
Early life conditions
influences in adulthood 6:752–758
aging 6:756–757
birth weight and preterm births 6:753–754
correlated environments 6:754–755
cumulative advantage 6:755
early parental death 6:756
fetal origins hypothesis 6:753
gestational age 6:752–754
life course epidemiology 6:752, 6:754–757
mortality selection 6:756–757
parental and grandparental age at birth 6:755–756
sibship size and birth order 6:754
Early life gene–environment interactions 7:844–845
Early mathematics 14:772–784
building and strengthening a culture of reasoning 14:776–783
capacities of mathematically talented children 14:777–778, 14:778f
diagonal length of square 14:772f
internal and external representation of numbers 14:775–776
fundamental number meaning 14:775
number symbols 14:775
number terminology–number words 14:775
object view–process view 14:776–778, 14:777f–778f
proper names–number names 14:775–776, 14:776f
mathematical playworlds boost reasoning 14:778–783
calculations are connected 14:780, 14:780f
components of calculations are connected 14:779, 14:779f
Event-Related Number Space Orientation 14:778–779, 14:779f
number range extension zero–four 14:779
place-value system 14:781, 14:781f
spiral stairs of calculation 14:779–780, 14:780f
Stellanian accounting system for their balance of Starthaler 14:781–783, 14:782f–783f
zero as number 14:780–781, 14:780f
M4-model 14:774–775, 14:774f
N4-model 14:775, 14:775f
number setting 14:772f
reasoning 14:773–774
mental constructions with accents 14:773–774, 14:773f
predicative- and functional-logical reasoning 14:773, 14:773f
understanding numbers 14:775–776
Early memory development 6:759–764
child-centered approach 6:761
childhood amnesia 6:761
classical conditioning 6:762
cultural approaches and 6:760
direct assessment method 6:761–763
indirect assessment method 6:760–761
cultural model 6:760
culture and cognitive development 6:760
declarative memory 6:759–760, 6:763
deferred imitation 6:763
East Asia cultural style 6:760–761
Euro-Caucasian-American cultural style 6:760
habituation paradigm 6:761–762
independent cultural model 6:760
interdependent cultural model 6:760
memory system model 6:759
mother-centered approach 6:761
neonatal and infant memory 6:759
nondeclarative memory 6:759
operant conditioning 6:762
research, future of 6:763
visual paired comparison paradigm 6:761–762
Early memory distortion work 8:711
Early modern, absolutism 1:6–7, 1:10
Early modern religion 2:507–508
Early modern times, genocide, history of 10:10–11
Early neutral evaluation 15:70
Early object permanence 4:44–45
Early organizational development 17:364–365
Early retirement 20:600
job displacement insurance 12:820, 12:830
Early sexual debut 21:797
Early short-term memory (eSTM) 15:157–158
Early speech
grammar learning 2:651–652
neural substrate for language development 2:652–653
sound discrimination 2:650–651
word learning 2:651
Early starters 14:60
Early warning systems (EWS) 5:841
Earmarking, and money 15:748–749
Earned Income Tax Credit (EITC) 13:195, 18:736
Earning subsidies 25:494–495, 25:494f
Earnings 25:335–336
Paretian distribution of 2:53
Earnings inequality 6:765–772
care of economists about 6:765
equality of opportunity 6:770–771
labor earnings 6:767–770

Earnings inequality (*continued*)
components 6:767–770
United States 6:767
labor supply 6:767–770
Earnings losses, job displacement insurance 12:819–821, 12:823, 12:825
Earnings potential 6:537
Earnings sharing regulation 20:168
Earth 10:83
Earth First! 23:814
Earth Liberation front (ELF) 16:249
Earth Observing System (EOS)-1 (Terra) 20:423
Earth Observing System (EOS)-2 (Aqua) 20:423
Earth Resources Survey (ERS) program 20:421–422
Earth Resources Technology Satellite (ERTS-1) 20:422
Earth science concepts 21:275–276
Earth.google.com 10:58
Earthquakes 6:480, 7:807
EAS *see* Emotionality; activity; and sociability (EAS) Temperament Survey
EASA *see* European Advertising Standards Alliance (EASA); European Association of Social Anthropologists (EASA)
Ease-of-learning (EOL) judgments 15:283–284
EAST *see* Extrinsic affective Simon task (EAST)
East
morality of 15:808–809
research ethics in 20:509
East, Gordon, works, *The Geography Behind History* 7:752
East Africa
agriculturalists 6:774–775
Arab influence 6:773
city-states 6:774
cultural concerns 6:776
foragers 6:775–776
globalization 6:776
history 6:773
kingdoms 6:774
modernity 6:776
pastoralists 6:774–775
prehistory 6:773
social forms 6:773–776
sociocultural aspects 6:773–777
East Asia
countries 5:454
population and economy size of 5:454t
cultural policies in
see also Cultural policies
cultural style, and early memory development 6:760–761
development of 6:278
fertility transition in 9:55–59
historical culture in 11:48–49
historiography, development of 11:49–50
China 11:49
Japan 11:49–50
Korea 11:50
Vietnam 11:50
historiology in 11:50–51
legal complex in postcolony 18:388
Meiji Revolution 15:126, 15:128
modern historiography in 11:51
nationalism in 16:275–278
advent of modern nationalism 16:276–277
Asianism 16:277
China-centered view of neighbors 16:275–276
Chinese World Order 16:275
colonialism 16:277
ethnic nationalism in periphery 16:277–278
globalization 16:277
imperialism 16:275, 16:277
modernization 16:276
nation states 16:276–277
national identity 16:276–277
Sino–Japanese War of 1894–1895 16:276–277
Tibet 16:277–278
Turkestan 16:277–278
Western threat and acceptance of international law 16:276
origins of food production in 9:301–302
as region 6:795–796
religions of 20:366–372
religious forms in
Buddhism 20:366–367
in China 20:369
Confucianism 20:367
folk religion 20:368–369
in Japan 20:370
Japanese scholarship for studies in 20:371
in Korea 20:369–370
millennial themes in 20:369–370
new religions 20:370–371
overview 20:366
religious integration in 20:367–368
scholastic traditions 20:371
Shinto 20:367–368
Taoism 20:367
social science academies in 22:637
urban policies in 24:895–896
East Asia and Pacific, research and teaching infrastructure in social and behavioral sciences in 22:631–636
financial setups
financial setups of research infrastructure development 22:632–633
funding 22:632
future research 22:635–636
Japan
development phases in 22:631
institutional structures of social sciences in 22:631–632
in twentieth and twenty-first centuries
Asian barometers and other types of scientific research 22:635
under Colonial Rule 22:633
fostering of social science practice with democratization 22:633
influence of globalization on research infrastructures 22:634
strengthening the Asian research base 22:634
East Asia Institute (EAI), Korea 22:642
East Asia legal traditions 6:778–783
China and East Asia legal orders during middle period 6:779–780
great tradition 6:778–779
Japan 6:778–782
Korea 6:778, 6:780–782
late imperial age 6:780–781
modern transformations 6:782
East Asian era names 24:409–410
East Asian Institute (EAI) 22:638, 22:640–641
East Asian politics 6:798–804
in Cold War period 6:798–800
Chinese politics 6:799–800
Japanese politics 6:798–799
Korean politics 6:799
democratic transition 6:799–803
developmental state model 6:798–799
influence of American political science 6:798, 6:800, 6:802
influence of media 6:798, 6:801–802
Liberal Democratic Party 6:798–799, 6:802
in post–Cold War period 6:800–803
Chinese politics 6:800–801
intraregional and extraregional comparative studies 6:803
Japanese politics 6:802–803
Korean politics 6:801–802
reform 6:799–803
Taiwan politics 6:800–801, 6:803
United States, East Asian students in 6:800–801
universities in East Asia 6:801–803
East Asian sociology 24:583
East Asian studies
economics 6:788–792
balance of power 6:791
China's transforming planned economy 6:789–790
industrialization 6:788–789, 6:791
Japan's lost decade and Asian financial crisis 6:790–791
modernization 6:788
newly industrialized economies 6:788–789
state 6:788–790
gender 6:793–797
colonial gender policy in modern Korea 6:794
colonial modernity, imperialist economy, national gender policy in modern Japan 6:794–795
economic liberalization, displacement of East Asia as region, gender scholarship, and transnational flexible economies 6:795–796
feminism 6:793–796
globalization 6:795–796

neoliberalism 6:793, 6:795
oppression 6:796
1840s to 1980s 6:793–795
subimperialism, colonial modernity, and policies of gender revolution in modern China 6:795
East Asian world system 1:924
East End study 4:363–364
East European Association of Schools of Social Work 12:475
East Timor 23:63
archaeological development 23:64
sociocultural aspects of *see* Southeast Asia
Eastenders (soap opera) 16:863
Easter Island, population dynamics 18:567
Easterlin, Richard 23:686
framework for fertility analysis 9:94–95
relative income hypothesis 8:824
Easterlin paradox 23:819
Eastern Air Lines Flight 401 crash 11:299
Eastern and Central Europe, Revolutions of 1989–1990 20:633–636
deviant type 20:633
peaceful revolutions 20:635–636
velvet revolutions 20:634–635
Eastern Asia, adult mortality in 1:171
Eastern Bloc 18:187–188, 18:191, 21:372
Eastern Europe 8:275–276, 12:15, 12:17–20, 18:187, 20:633, 22:993
adult mortality in 1:171
clinical psychology 3:874
development 3:874
dictatorship in 22:139–143, 22:145–146
fertility in former socialist countries in 9:99
gender in 9:782
Jewish Diaspora 12:802–806, 12:808, 12:810
postsocialist cities 3:605–607
psychotherapy 3:874
revolutions in 25:838–839
social work in 22:795–800
academization of social work and social work networks 22:799–800
communism 22:795–797, 22:799
deinstitutionalization 22:797–798
diversity 22:795, 22:799–800
ethnic conflicts 22:796–797
European union 22:795, 22:797–798
historical development before 1991 22:795–800
historical perspectives 22:799–800
international financial institutions 22:796, 22:798
people with disabilities 22:796–798
post-1991 practice 22:797–799
poverty 22:796–797
rereligization 22:796, 22:800
Roma people 22:796–798
schools of social work 22:795–796, 22:799
social assistance 22:795–800
social policy 22:795–796, 22:799–800
from state socialism toward neoliberalism 22:796–797
in socialist societies 22:849, 22:852
Eastern Europe borders 6:811
Eastern European Jewish life 12:875–876
Eastern European studies 6:810–815
in 1989 6:810
area studies 6:810–811, 6:813–814
and Cold War 6:813
communism 6:813, 6:817–819
contours and borders 6:816–817
of culture 6:805–809
area studies 6:805–808
communism 6:805–808
cultural formations after postcommunism 6:808
culture, civil society, and transition 6:807
Eastern Europe as cultural category 6:805–806
empire 6:806
European Union 6:805–806, 6:808
gender 6:807
knowledge 6:805, 6:808
nationalism 6:806–808
significance of nation 6:806–807
war 6:805–808
division of Europe 6:811
Eastern Europe borders 6:811
informal economy 6:818–819
as inorganic 6:810
Jewish histories 6:814
postcommunist transition 6:819–820
and regions 6:810
religious frameworks 6:810–811
society 6:816–821
one-sided development 6:817–818
turn to Europe 6:820
totalitarianism 6:818
and transitology 6:810
from where to where? 6:818–820
Eastern Navajo Dine Against Uranium Mining (ENDAUM) 7:760
Eastern North America, origins of food production in 9:303
Eastern totalitarian systems 24:143
Eastman, George 18:82
Easton, David, works, *The Political System* 18:469
East-West Center (EWC), Hawaii 22:638
East–West Cold War 4:128–133, 23:314
Eating
neural basis of 11:420–422
arcuate nucleus 11:421–422
dual center theory 11:420
economics 11:421
energy 11:420–422
environmental influences 11:420–421
food variety 11:421
history 11:420
homeostasis and allostasis 11:420
hunger 11:421
hypothalamus 11:420–422
neurobiology of hunger and satiety 11:421–422
nucleus of tractus solitarius (NTS) 11:421–422
reward systems 11:422
satiation and satiety 11:421
taste intensity 11:421
taste quality 11:420–421
Eating behaviors, and social networks 16:558
Eating Disorder Not Otherwise Specified (EDNOS) 6:822–823
Eating disorders 6:822–828, 23:334–335, 25:143
anorexia nervosa 6:823–825
binge eating disorder 6:825–826
bulimia nervosa 6:825
cross-cultural studies 6:826–827
and cultures in transition 6:826
and diabetes 6:340
diagnostic criteria as cultural models 6:822
eating disorder not otherwise specified 6:826
historical accounts of 6:822–823
motivating constructs in 6:826–827
psychotherapy for 19:376–377
anorexia nervosa 19:376
binge eating disorder 19:376–377
bulimia nervosa 19:376
treatments 6:827
EAUH *see* European Association of Urban Historians (EAUH)
EB *see* Eurobarometer (EB)
eBayization 14:829–830
Ebb and flow, brokers/brokerage 2:873–874
Ebbinghaus, Hermann 4:189, 13:691, 14:811, 19:412
quantitative memory models 15:160–162, 15:164–165
Ebbinghaus illusion 25:204f
EBBM *see* Evidence-based behavioral medicine (EBBM)
EBBP *see* Evidence-based behavioral practice (EBBP)
Ebert, Dominik, on epistemic cultures 7:878
EBF *see* Exclusive breastfeeding (EBF)
EBM *see* Evidence-based medicine (EBM)
Ebonics 18:136
E-books 6:899
EBP *see* Evidence-based practice (EBP); Evidence-based psychotherapy (EBP)
EBQ *see* Epistemological Beliefs Questionnaire (EBQ)
EBV *see* Epstein–Barr virus (EBV)
EC *see* Effortful control (EC); Embodiment cognition (EC); European Community (EC); Evaluative conditioning (EC)
EC Directive to Implement Minimum Health and Safety Requirements at Temporary or Mobile Construction Sites 17:143
EC Framework Directive of 1989 17:143
ECA *see* Economic Cooperation Administration (ECA)
ECB *see* European Central Bank (ECB)
Eccentricity, of vertex 16:534
Eccles, J.C., works, *The Self and its Brain* 18:525
Eccles' expectancy-value model 1:37
Ecclesia 18:435–436

Ecclesiastic grassroots communities *see* Foundational Christian communities
Ecclesiastical History 11:31
Ecclesiastical use, secularization 21:383, 21:386
Ecco database 21:190
ECDA *see* Exploratory combinatorial data analysis (ECDA)
ECDB *see* Extremist Crime Database (ECDB)
ECEC *see* Early childhood education and care (ECEC)
Echo intensity 15:163
ECHP *see* European Community Household Panel (ECHP)
ECHR *see* European Convention on Human Rights (ECHR); European Court of Human Rights (ECHR)
ECI *see* Employment Cost Index (ECI)
Ecker, Bogomir, works, *Dripstone Machine* 2:21f
Eckersley, Robyn 5:300
Eckstein, Harry, works, *Theory of Stable Democracy* 2:190
ECLA *see* United Nations Economic Commission for Latin America (ECLA)
Eclectic balance 1:562–563
Eclectic methodological pluralism 10:783
Eclecticism 9:421
  in psychotherapy 19:509–512
ECLS *see* Early Childhood Longitudinal Studies (ECLS) Program
ECM *see* Expectation–Conditional Maximization (ECM) algorithm
Ecocentric area of harmony 10:532–533
Ecocentric view of environmental policy making 7:780
Eco-cities 6:829–837
  challenges 6:835–836
  climate change 6:829, 6:835–836
  dimensions 6:829–830
  ecological modernization 6:829–830
  emerging cities 6:836
  global developments 6:832–835
    eco-city pioneers 6:832
    internationalization 6:833–834
    policy mainstreaming 6:832–833
    ubiquitous eco-city 6:834–835
  governance 6:829–831, 6:836
  low-carbon urban development 6:829, 6:831–833
  pioneers 6:832
  smart technology 6:834–836
  sustainability triple bottom line 6:830
  sustainable urban development 6:833, 6:835
  urban development, types of 6:830–831
  urbanization 6:829, 6:833–835
  variations on a theme 6:831–832
Ecocultural model 6:839
  behavioral outcomes 6:839
  ecological context 6:839
  population adaptations 6:839
  sociopolitical context 6:839
  transmission variables 6:839
Ecocultural theories of development 6:838–844
  applications 6:842–843
  conceptual thrust 6:839–843
  cultural ecology 6:838
  developmental niche 6:841
  ecological systems theory (EST) 6:840–841
  original ecocultural theory 6:839–840
  social ontogenesis 6:841–842
  theoretic thrust 6:839–843
Ecocultural theorization 6:839–843
Ecocultural thinking 6:838
Eco-efficiency 23:798
  and industrial ecology 11:846
Ecofeminism 6:845–850, 23:813–814
  activism 6:849
  citizenship 6:849
  climate change 6:849
  feminist political ecology 6:848
  future 6:848–849
  gender justice in environmental justice 6:848
  history of 6:845–847
    challenges 6:846–847
    essentialism 6:846
    social construction 6:846
  impacts 6:847
  lesbian, radicalized socialist 8:945–947
  in other environmentalisms 6:848
Eco-feminism 20:288
ECoG *see* Electrocorticography (ECoG)
Eco-garden city *see* Eco-cities
Eco-gastronomy, Slow Food movement 22:96
Ecogeographic clines in body size 2:724–725
Ecolabels 7:781
École des hautes etudes commerciales (school) 7:393–394
Ecole des Hautes Etudes en Sciences Sociales (EHESS) 11:40, 22:669–670
Ecole Libre model
  of political science 18:467
École normale supérieure (school) 7:393–394
Ecological approach, and perceptual learning 10:127–128
Ecological assets 3:446, 3:449
  indicators 3:447
Ecological budget 3:158–159
Ecological capital 23:824
Ecological capture–recapture studies 18:605
Ecological carrying capacity 3:156–157
Ecological conditions, differences, during preschool age 18:223
Ecological context
  of feminist psychotherapy 8:930, 8:932–933
  studying indigenous knowledge in 11:781–783
Ecological correlation 6:866
Ecological demography, and human ecology 11:284–287
Ecological determinants 10:623
Ecological economics 6:24, 6:851–864, 22:254, 22:258, 23:842
  civil society and 6:860–861
  against commensuration 6:855–857
  cost–benefit analysis (CBA) approach 6:855
  economic valuation 6:855, 6:857, 6:861
  economy embedded in society and nature 6:851–852
  entropy 6:853, 6:861
  environmental economics 6:852–853
  environmental services 6:854, 6:856–858
  financial debts compared to real wealth 6:854–855
  history and foundations 6:852–854
  human ecology 6:853, 6:855
  irreducibility of needs 6:855
  'isolated' systems 6:852
  new ecological macroeconomics without growth 6:854
  optimist paradox 6:857
  outlook 6:862
  Payment for Environmental Services (PES) 6:858
  on population 6:860
  property rights and resource management 6:857–858
  rebound effect 6:854
  social metabolism and indexes of (un) sustainability 6:858–860
  society for 6:861–862
  steady-state economy 6:853–854, 6:861
  strong sustainability 6:852–853, 6:856–857
Ecological enclaves 6:835–836
Ecological engineering, and food production 9:303–304
Ecological fallacy 21:2
  correlation 6:865–867
  example 6:865
  impact on research practices 6:866
  statistics of 6:865–867
    bounding individual-level parameters 6:867
    equivalence conditions for aggregate-level and individual-level parameters 6:866–867
    mitigation of problem 6:866–867
Ecological footprint (EF) 6:859–860, 12:717, 13:242, 18:352, 22:258, 22:258f
Ecological functioning, and sustainable development 23:825
Ecological Green Party of Mexico (PVEM) 10:380
Ecological hazards 6:480–481
  epidemics 6:480
Ecological inference 6:868–870
  correlation 6:868–869
  method of bounds 6:869
    causal inference 6:868
    statistical fallacies 6:868
  regression 6:869
    with random coefficients 6:869
Ecological integrity 13:241

Ecological Interface Design (EID) 11:300–301
Ecological knowledge, fishing 9:227–228
Ecological law of self-thinning 13:240
Ecological model of human development 22:943
Ecological modernization 6:829–830, 7:781
  and climate change 3:855
  of production 23:840–841
  and sustainable development 23:813
Ecological modernization theory (EMT) 7:798–802
Ecological momentary assessment (EMA) 7:507, 10:590, 10:592
Ecological perspective, of context 4:787
Ecological problems 7:796–797, 7:802
Ecological rationality 12:722
  of heuristics 10:832–834, 10:833f
  versus internal consistency 5:912–913
Ecological resilience 20:561
Ecological scarcity 23:824
Ecological science 10:547
Ecological social work 6:871–877
  alternative paradigm of nature 6:872–874
  conventional perspective 6:871–872
  core operational values 6:874–875
  ecosystems 6:872–875
  environmental social work 6:873–875
  future prospects 6:875
  green social work 10:387–388
  and green social work, commonalities between 10:386–387
  life-model 6:872
  major scholarly themes 6:874
  person-in-environment 6:871, 6:873–874
Ecological studies
  of infant development 18:855–857
  participatory action research in 17:535–539
Ecological sustainability 8:22
Ecological systems approach, to socialization and education 22:864
Ecological systems theory (EST) 6:840–841
  chronosystem 6:841
  exosystem 6:840
  macrosystem 6:840
  mesosystem 6:840
  microsystem 6:840
Ecological themes 20:288
Ecological theorizing 22:956
Ecological theory 3:803
  and property rights 19:169
  refugees 22:771–772
Ecological Utopias 14:256
Ecological validity 17:542
  of executive function assessment 6:726
Ecologically imposed monogamy 14:615
Ecologically oriented radical lesbian feminism 8:942–948
Ecologically unequal exchange 18:352
Ecologism 7:811
Ecology 7:791–793, 7:809, 18:138
  of aging 23:668
  agrarian political economy 1:460–461
  of Arctic 1:928
  and consumerism 3:102
  Darwin's evolutionary theory 7:792
  ecosystems 7:792–793
  evolutionary ecology 7:792
  Freyre on 9:420–421
  and health 6:878–883
    contextual factors 6:878–879
    developing high-leverage strategies for effective health promotion 6:881
    ecological perspectives on health and illness 6:878–879
    interdisciplinary scope 6:879
    macrosystem 6:880
    mesosystem 6:880
    microsystem 6:880
    policy 6:878–881
    reciprocal causation 6:879
    research findings from ecological analyses of health 6:880–881
    sociocultural environment 6:878
    systems theory and ecological paradigm 6:879–881
    targeting vulnerable subgroups in health promotion programs 6:881
  human 11:283–288
  of ignorance 1:749
  industrial *see* Industrial ecology
  and industrial societies 11:863–866
  of infrastructure 12:120
  and international advertising 12:455–456
  of language 13:354
  and leisure 13:847
  in Mexico 1:627–628
  in nineteenth century 7:791–792
  of North America 17:15–17, 17:20
  organizational 6:890–894
  and social change 22:148
  and social differentiation 6:394
  systems ecology 7:792–793
Ecology governing fertility and parental investment 9:31–33
Ecology groups, in Soviet Union 20:831–833
Ecology of aging 6:884–889
  aging in place 6:884–886
  field development 6:884
  goals and mission 6:884–885
  integrative views 6:887–888, 6:888f
  out-of-home environments 6:886–887
  person-environment relations
    approaching on macro-level of analysis 6:887
    objective approaches to 6:885–886
    subjective approaches to 6:886
  place identity 6:886
  role within aging research 6:885
  technological environments 6:887
Ecology of human 22:255
Ecology of leisure 13:832–833
E-commerce 12:594–595, 14:547–548, 23:713
E-communications platforms 6:895–903
Ecomuseum 1:755–756
Econeighbourhood 24:903–904
Eco-niche, human 11:256–257
Econometric model 7:122
Econometric Society (US) 9:220–221
Econometrica 24:345
Econometrics 7:1, 21:306, 24:337–338
  history of 7:8–11
    development era 7:8–9
    establishment of mature discipline 7:9–10
    estimation 7:9–10
    hypothesis testing 7:9–10
    identification 7:9
    measurement 7:8–10
    place in economics 7:10
    simultaneity 7:9–10
    vector autoregressions 7:9–10
  instrumental variables in 12:205–209
  for labor market outcomes 17:72, 17:74
  macroeconometric models 14:401–404
  quantification in 19:704
  software 7:1–7
    accuracy 7:5
    application studies 7:3
    common features 7:3–4
    development 7:5–6
    development settings 7:2
    estimation procedures 7:4–5
    evolution 7:1–3
    integrated packages 7:2
    interfaces 7:4
    Internet 7:3
    modern software 7:3–5
    packaged expertise 7:3
    program libraries 7:2
    programming languages 7:2–3
    user written programs 7:3
Economic action
  and community 4:386–387
  definition 5:535
Economic analysis 13:454–455
  of areas for research 13:451
  empirical work, importance of 13:455
  of law 13:448–449, 13:455–457
    criticism about 13:452
    foundations of 13:452
    sociological influences in 13:458
  neoclassical model 13:454–455
  of other areas of law 13:449–451
    contract law 13:450
    law enforcement and criminal law 13:451
    litigation 13:450–451
    property law 13:449–450
  positive economics 13:454
  traditional analysis contrasted with 13:449
Economic and Monetary Union (EMU) 2:902–903
Economic and Social Council (ECOSOC) 16:857
  United Nations 24:752
Economic and Social Research Council (ESRC) 16:191
Economic anthropology 7:12–20, 15:746–749
  contemporary futures 7:18–19
    capitalism 7:19
    cross-border movement 7:18

Economic anthropology (*continued*)
and exchange 8:476–480
finance 9:174–175
foundations and orientations of subdiscipline 7:13–17
agency 7:16–17
articulation 7:15–16
commodities 7:13–14
consumption 7:16–17
gifts 7:13–14
history 7:15–16
market 7:14–15
modern economy 7:16–17
modes of production (MOP) 7:15–16
nature of exchange 7:13–14
redistribution 7:14–15
neoclassical economics 7:12–13
Polanyi on 18:245
transitions to modern economy 7:17–18
Economic approach, to voting 25:302–303
Economic attitudes, and interests 12:392–393
Economic behavior 4:946–947
and genetics 9:951–954
impact on 4:946–947
and income tax
labor supply 11:744–745
saving and portfolio choice 11:745–746
liberal conception of 14:23
and networks 16:520
of trade unions 24:497–502
Economic bourgeoisie 2:785–787
Economic capital, family and schooling 8:753
Economic change 22:147–155
Economic classifications 5:760–761
standard 5:760–761
Economic collapse, LTG study 14:141
Economic Commission for Latin America 3:123, 4:123
Economic consequences
of immigration 15:383–384
of population aging 18:542–543
Economic considerations, urban growth models 24:807
Economic cooperation, China 3:502
Economic Cooperation Administration (ECA) 16:229
Economic cooperative networks 14:526
Economic cosmopolitanism 1:230
Economic costs
of climate change impacts 3:860–861
Economic crisis 6:28, 14:653, 16:260
in Africa 23:598–600
older age work participation 17:188
and voting 7:29–30
Economic damages from climate change, estimates of 3:860–861
Economic databases 5:742–750
business establishments 5:742–743
establishment data sets 5:743
future of 5:749
history of 5:742–743
households 5:742–749
macro databases 5:748–749
Eurostat's National Accounts 5:749
IMF's World Economic Outlook 5:749
National Accounts Main Aggregates Database 5:748
OECD Statistics Portal 5:748
OECD System of Composite Leading Indicators 5:748
UNCTAD statistics 5:749
World Bank databases on economic and growth 5:749
microeconomic databases 5:743–745
survey of consumer attitudes 5:744–745
survey of consumer finances 5:744
sample survey 5:742
wealth 5:744–747
Economic decision-making 5:879, 16:643–644
Economic decisions 5:896, 5:898–899
in Cold War 4:130–131
Economic decline 11:873
Economic determinism 6:242–243
Marx and 6:242–243
Economic development 5:749, 12:146, 20:802, 20:804, 24:795–797
agriculture's primacy in 1:477
and big man 2:568
and climate change 3:848, 3:850–851, 3:854–855
and culture 5:528–534
achievement motivation 5:528–529, 5:531, 5:533
causality 5:533
collectivism 5:528–529, 5:531
cross-country qualitative studies 5:531
dimensions 5:528, 5:531, 5:533
historical and in-depth case studies 5:529–531
Hofstede on 5:531, 5:533
individualism 5:530–531, 5:533
Islam 5:530–532
obedience 5:531
phases 5:528–529
politicized culture 5:532–533
power 5:530, 5:532
protestantism 5:528, 5:531, 5:533
religion 5:528, 5:531, 5:533
role of institutions 5:531–532
trust 5:530–533
World Values Survey 5:528, 5:531, 5:533
dual economy models 1:477
East Asian studies 6:786
and economic transformation 7:76–78
and family planning 8:786
and fertility 9:95–96
Hirschman on 10:892, 10:894–895
international social work student exchanges 12:562–563
population dynamics and 18:568
and property rights 24:530
and remittances 12:519, 12:523
and social work education 22:776
and sustainable transportation 23:846
and telephone 24:158–159
and urbanization 6:283, 24:962–963
Economic development policies 23:159
Economic Development Zones (EDZs) 24:877
Economic differences, and student achievement 5:343–344
Economic education 7:21–25, 24:83–88
assessment 7:23–24
assessment practices 24:86–87
classroom 7:22–23
content 7:21–22
course management software (CMS) 7:22–23
integration of teaching methods and media 24:84–86
applying creative and innovative teaching methods 24:84–85
service-learning approach 24:85
technology-integrated media 24:85–86
knowledge 7:21–24
literacy 7:21–22, 7:24
multiple-choice questions 24:86–87
and organizations 7:21
organizations promoting 24:83–84
Council for Economic Education 24:83–84
South African Foundation for Economic and Financial Education 24:84
student evaluations of teaching 7:24
student performance 7:24
teaching methods 7:22–23
Train-the-Trainers program 24:83–84
understanding 7:21–22
Economic effects
of advertising 1:208–209
of Chinese migrations 12:528
of minimum wages 15:561–563
Economic efficiency 13:793–794
and income tax 11:747–748
pharmacoeconomics 19:458
Economic equality 7:91
Economic equilibrium 17:504
Economic Espionage Act 1996 3:30
Economic ethics, and business ethics 3:15
Economic evaluations, during elections 7:26–32
Economic exchanges 16:643–645
Economic expansion 25:531
Economic explanations, for fertility transition 9:60–67
Economic fluctuations
as cycles 14:410
and structural unemployment 24:738
Economic forecasting 24:337–340
Economic gangs 9:595
Economic geography 7:33–37, 14:287–289, 14:291, 20:568, 20:571, 23:82
background 7:33–34
empirical challenges 7:36
and geographical economics 7:34–35
study objects 7:35–36
theory and methods 7:34
Economic globalization 10:206, 10:262, 17:249–250, 22:935
work-related stress 23:546

Economic growth 7:38–44, 13:443–444, 23:794
capital 7:39–43
and cash transfers 3:217–218
convergence 7:39–41
and crime 5:199–200
and culture 5:528–534
decline 7:41
definition and measurement 7:38–39
education 7:40–43
endogenous growth 7:40–41
and entrepreneurship 7:675–680
environmental Kuznets curves 13:144–147
fertility 7:39, 7:41–43
household economics 11:225
human capital 7:40–42
and human development 6:307–311
and inequality 7:43
institutions' role 7:40–41
macroeconomics 14:410–411
and macroeconomics 14:402
measurement 7:45–50
geospatial income measures 7:49
historical estimates 7:46
ideas 7:48
inequality 7:48
Kaldor's stylized facts 7:46
national accounting system 7:45–46, 7:48
national income 7:45
Penn world table 7:47–48
physical capital accumulation 7:46–47
structural drift 7:48–49
total factor productivity 7:46–49
mortality 7:39, 7:41–43
neoclassical growth 7:39–40
poverty traps 7:41
quality–quantity trade-off 7:42–43
technical progress 7:39–40, 7:42
theories of stagnation 7:39
transition to modern growth 7:41–42
exogenous models 7:42
unified growth theory 7:42
and urban poverty 24:916
Economic growth rate 7:45
Economic hazards, urban life health 24:831
Economic history 7:51–54, 11:103–105, 19:696–697, 22:307–308
cliometrics 7:51–53
communication and transportation 4:274–276
definition 7:51
financial crises 9:183–190
financial history 7:53
future of 7:53–54
global history 7:53
historical economics 7:53
labor 13:169
new, emergence of 7:51–52
origins and development 7:51
Polanyi on 18:243–245
quantitative
economic analysis of technological change 7:56–57
institutions 7:59–60
new directions in 7:58–60
Nobel Prizes 7:58
origins of 7:55
path dependence 7:58–59
returns 7:58–59
rise of 7:56–58
slavery debate 7:57–58
technology 7:58–59
in United States 7:55–60
search for balance 7:52–53
twentieth-century 7:51
Economic human rights 11:355
Economic impact of international immigration 12:511, 12:516–518
Economic imperialism 1:77, 12:393
Economic implications of disasters 22:755
direct losses 22:755
indirect losses 22:755
secondary losses 22:755
Economic independence, of women 24:572–573
Economic indicators 5:761–762, 9:358
Fisher's 9:219
Economic inequalities 7:903, 11:163
and family theory 8:819–820
in health 10:635
social change 22:194–195
Economic infrastructure 3:647
Economic institutionalism 12:186–187
Economic institutions 12:192
Economic integration 12:574–579
customs unions 12:574, 12:576, 12:578
economic effects on member countries 12:574–576
scale and competition effects 12:576
trade creation and diversion 12:574–576
winners and losers 12:576
free trade areas 12:574, 12:576–578
policy making 12:576–577
depth of integration 12:576–577
policy reform and commitment 12:577
regional integration agreements and world trading system 12:577–578
West African Economic and Monetary Union 12:574
Economic interests
and culture 5:537
sociology and 12:391–393
economic attitudes 12:392–393
scope 12:393
social order problem 12:391–392
Economic Journal 7:114
Economic laboratory 8:556–558
Economic landscape 23:84
Economic liberalism 25:676
and gay/lesbian activism in Latin America 13:861–862
Economic liberalization 24:815
in East Asia 6:795–796
Economic life
and religion 20:258–262
Economic Man model 4:739
Economic mobility 24:731
Economic modeling 14:266–267
Economic motivation, in United States 11:629–630
Economic nationalism, Central America 1:623
Economic neoinstitutionalism 12:187
Economic opportunities, intellectuals 12:279–280
Economic Opportunity Act 18:734–735
Economic optimism, and political trust 24:659–660
Economic organizations, and political economy 1:940
Economic outcomes, mass imprisonment 14:697–699
Economic perspectives, of entrepreneurship 7:683–684
Economic poverty, measurement of 18:741
Economic prosperity, and social capital 22:174–175, 22:178
Economic psychology 5:886–891
debt 5:888
economic behaviors of individuals 5:888–889
economic socialization 5:889
environmentalism 5:889
experimental economics 5:887, 5:890–891
impact of economy on individuals' psychology 5:889–890
investment 5:888–889
money 5:888–889
in new economic era 5:890–891
rationality 5:886–887, 5:890–891
recession 5:886, 5:888–890
relation to other research areas 5:886–888
saving 5:888
social dilemmas 5:887–888
taxation 5:888, 5:890
Economic rationalism 18:765
Economic realism, and social stratification 22:343
Economic reforms
Fisher's 9:219
post-socialist 18:690–696
acquis communautaire 18:693
agenda 18:690
communist legacies and challenges 18:690–691
Copenhagen criteria 18:693
dependent market economies 18:694
EU accession 18:693–694
foreign direct investment 18:694–695
gradualism 18:692
impact of financial crisis 18:694–695
liberalization 18:690–693, 18:695
neoliberals and critics 18:691–692
pathways 18:692–693
privatization 18:692–693
regulation 18:693
return to Europe 18:690
shock therapy 18:691–692
transition 18:690–691, 18:694–695
and property rights 24:530
Economic regulation 20:172, 20:178–179, 23:459
effects of 20:173–175

Economic regulation (*continued*)
monopoly regulation, empirical analysis of 20:173–174
regulation in multifirm, empirical analysis of 20:174–175
empirical methodologies 20:172–173
Economic Report of the President 11:732
Economic Research Service 10:304
Economic resources
and education 7:132–134
and family relations 15:736
Economic restructuring 7:61–66, 12:91–92, 22:194–195
geographic studies, themes of 7:61
postsocialist cities 3:608
Economic rights 3:674–676, 12:956
Economic Science Association 19:538
Economic segregation, and student achievement 5:346
Economic situation, of lesbians 13:895–896
Economic slavery 22:50
Economic socialization
economic psychology 5:889
study of 5:889
Economic sociology 7:67–72
comparative capitalisms 7:69
consumption 7:70–71
culture 7:67–70
embeddedness 7:67–68
finance 7:69–70
financialization and globalization 7:69–70
institutional theory 7:68–69
market conventions 7:69
market devices 7:68–69
markets 7:67–68
morality and 7:67, 7:70
network analysis 7:68–69
performativity theory 7:68–69
Polanyi on 18:244–245
political economy 7:68–69
Weber on 25:466
Economic special zones 24:849
Economic stabilization 14:409
Economic statistics 5:758–762
and business statistics 5:758
comparability and role of Eurostat 5:762
conceptual frameworks 5:761
consumer price index 5:758, 5:760–762
current economic indicators 5:761–762
data-gathering methods 5:758–760
business records 5:759–760
directly measured statistics and derived statistics 5:758–759
household surveys 5:760
measurement 5:759–760
problems of defining business 5:759
public accounts and administrative records 5:760
evolution as system 5:761
applications of business registers 5:761
censuses and registers 5:761
and official statistics 5:758
other classifications 5:761
prices 5:758, 5:760, 5:762
production 5:758–762
quantities 5:758
sample survey techniques 5:758–759, 5:761
standard economic classifications 5:760–761
technology 5:758
today's arrays of 5:762
types 5:759t
Economic status
acquired 11:896
in collectivism/individualism 4:207
intergenerational transmission of 17:465
Economic stratification, Veblen on 25:55
Economic surplus 25:7–9
Economic systems
and democracy 6:95
of knowledge society 13:109
and social system 22:746–748
Economic theories 1:574, 11:629, 12:104–105, 21:141, 21:143–144
cultural forms 5:539–540
defense economics 6:2
entrepreneur in 12:170
and equilibrium 7:912, 7:915–917
of firm 19:894
Fisher's 9:218
and industrial societies 11:864
interest and 12:386
margins of mainstream 6:4–5
and political economy 18:357–359
of recidivism 7:81
of regulation 20:166–171
and social science 12:388
of time allocation 24:370
of transaction costs 2:915
Economic theories, of criminal behavior 7:80–82
benefits and costs from crime 7:80–82
criminals, rational 7:81–82
criminal's discount rate 7:81
economic theory of recidivism 7:81
economics *versus* criminology 7:81
stigma of criminal activity 7:81
economics of law enforcement 7:83–85
certainty *versus* severity 7:83–84
extensions of the basic model 7:84–85
fines *versus* imprisonment 7:84
marginal deterrence 7:85
repeat offenders 7:85
risk preferences 7:84–85
empirical studies of economic model of crime 7:82–83
capital punishment 7:82
gun laws 7:82
imprisonment 7:82
income 7:82–83
income inequality 7:83
police 7:82
unemployment 7:83
Economic theories, of public enforcement
Beccaria's ideas 13:570–571
Becker's ideas 13:570–571
behavior of law-enforcement officials 13:572–573
behavioral theory 13:571–572
criminological approach *versus* 13:571
law as an incentive system 13:571–573
normative analysis 13:573–575
Economic thought
alternative schools of 1:573–578
American Institutionalism 1:573–574
Austrian school 1:574–576
feminism 1:576–577
heterodox economics 1:573, 1:577
market process 1:575–576
post-Keynesian economics 1:577
uncertainty 1:576–577
Economic transactions 3:736
Economic transformation 7:73–79
assessment and lessons 7:78–79
central planning and initial conditions 7:74–75
in China 7:73–74, 7:76–79
economic development 7:76–78
economic policies and reforms 7:75–76
enterprise restructuring 7:75–76, 7:78–79
historical perspectives, state's role 6:276–277
institutions and institutional change 7:76–77
market economy 7:73–75, 7:77–79
performance and outcomes 7:77–78
privatization 7:73–77, 7:79
in Russia 7:74–79
Russian and post-Soviet studies 20:847
theories and paradigms 7:73–74
transition economics 7:73
Economic trends, fertility 9:23
Economic uncertainty, fertility 9:21
Economic underdevelopment, African 1:289
Economic underpinnings, market geographies 14:545–546
Economic value
of children 3:498–500
of erotic capital 8:3–4
of prisoners 11:721
Economic voting 7:26, 7:29
EITMs in 18:456–457
behavioral and applied statistics concepts 18:456
devising formal and applied statistical analogs 18:456–457
linking formal and applied statistical analogs 18:457
Economic war 4:618
Economic world reversed 5:466
Economic worldviews 11:864
Economical alienation 1:552–553
Economics 11:102–103, 11:105–106, 17:780, 21:141–144, 23:6
and age stratification 1:336
of aging 20:599
applied social research 1:854–855
archaic states 23:371–373
of archaic states 23:371–373
of art and culture 2:1–6
of arts 5:731
of Australian aboriginals 2:235
bachelor's degrees 22:680t
behavioral *see* Behavioral economics

of childbearing 8:821–825
classical economics 7:99–100
and Cold War 4:124
consumer psychology 4:738–739
and crime 7:80–86
as cultural forms 5:539–540
of culture 5:728–731
of defense 6:1–6
and development 6:303–304
and dictatorship 22:141–146
discipline 7:93–96
'dismal science' designation of 15:332
and division of labor 6:602
and divorce 6:607–608
doctoral degrees 22:680t
East Asian studies 6:788–792
and eating 11:421
education 22:729
of education 21:119–120
empirical legal studies 7:554–556
equality 7:103
and ethics 7:87–92
equality in 7:91
and liberalism 7:87
Pareto principle 7:89–90
tools for ethics 7:90–91
utilitarianism 7:89–90
utility theory 7:88–89
value of life 7:91
of expectations 8:510–516
experimental 8:399
freedom 7:103
and freedom 9:400–402
Freyre on 9:420
health 10:614–619
heterodox 1:573, 1:577
heterodox economics 7:94–96
history of 7:93–98
engagement with 7:95–96
marginal revolution 7:96
in twenty-first century 7:96–97
household 11:224
housing 11:239
idealization in 11:517–518
of infant development 18:858
inference on expectations in 19:777
of information *see* Information; economics of
of innovation and technological change 12:160–168
of intergenerational relations 8:826–828
and journals 12:866
justice 7:103
and justice 7:83
and labor history 13:169–171
and law 12:935–936, 13:448–460
of law enforcement 7:83–85
certainty *versus* severity 7:83–84
extensions of the basic model 7:84–85
fines *versus* imprisonment 7:84
marginal deterrence 7:85
repeat offenders 7:85
risk preferences 7:84–85
mainstream 1:573
of marriage and divorce 8:829–834
master's degrees 22:680t
method a priori 7:99–100
methodology 7:99–103
of migration 15:433–439
of multinational corporations 16:78–79
natural experiments 8:546–547
old and new institutionalism in 17:190–195
peasant 17:678
performativity of 5:525
and personality 17:780–786
beliefs and information 17:783
incentives 17:784
preferences 17:781–783
technology 17:784
philosophy of 7:99–104
place of econometrics in 7:10
preference 7:102–103
primitives of 17:780
progress 19:125
prospect theory 19:263–265, 19:267
of psychiatric care 19:295–300
quantification in 19:704
rationality 7:102–103
of retirement 20:602
self-interest 7:102–103
skills training 22:20–24
Smith on 22:118–119
social dilemma 8:549
social networks in 16:572–581
choosing actions and links 16:575–576
criminal networks 16:578–580
games 16:572–574
key players 16:579–580
labor-market networks 16:577–578
network formation 16:574–575
and sustainable production and consumption 23:840–842
teaching 24:83–88
twentieth century philosophy 7:100–102
uncertainty in 24:718–719
of value of life 25:14–20
welfare *see* Welfare economics
Economics, of gender 9:757–763
time allocation
in family context 9:757–759
household production and 9:757
Economics, of safety 20:863–869
evidence on effects of OSHA standards 20:868
government mandate theory 20:868
law and 20:867
evidence on tort law stimulus to safety 20:867–868
theoretical stimulus of tort law to safety 20:867
L&E theory concerning government regulations 20:868
moral hazard 20:865
neoclassical economics 20:863–864
evidence consistent with 20:863–864
modified 20:864–866
qualifications concerning 20:864
new institutional economics 20:866
Coase theory and transaction costs economics 20:866–867
evidence concerning changes in liability rules 20:866–867
old institutional economics 20:864–866
collective action 20:866
effects of experience rating 20:865
evidence on effects of experience rating 20:866
workers' compensation 20:864–865
risk premiums 20:863–865
tort suits 20:863–865, 20:867–868
value of statistical life 20:863–864
Economics and Human Biology 19:697
Economies, Sociétés, Civilisations 11:40
The Economies of Scale 23:458
Economies of trust 24:676–677
Economism 2:778
Economist Intelligence Unit, *Democracy Index* 24:451–452
Economizing 23:511–512
transaction cost 10:843–844
Economy
of Arctic 1:928
and balance of power 2:347–348
and children/infant mortality 12:24–25
and communities 22:267
cultural perspectives of 11:863–865
culture and 5:535–541
connection, classics on 5:535
and consumption 5:537–538
cultural understandings of price and value 5:538–539
culture, development, and globalization 5:539
economics and economic theory as cultural forms 5:539–540
market culture 5:536–537
in markets 5:535–536
organization and production, cultural embeddedness of 5:537
definition 5:535
and emigration from Europe 15:460
enlightenment in 7:659
and fertility transition 9:96
in functionalist grammar 9:474
gender in 9:792–793
global history 10:167
globalization 10:206, 10:210–211
historical sociology 11:27–29
of India 2:80
informal 12:56, 12:58
and international advertising 12:456
Jacobs on 12:779
and Keynesianism 13:20–22
market *see* Market economy
and money 15:747–748
moral, in Southeast Asia 23:71
pastoral 17:580
political
in Australia 2:242–243
of socialist societies 22:849–851
of scope 6:586
shadow, and unemployment 21:863–865
of shortage, in socialist societies 22:850

Economy (*continued*)
of small states 22:108–112
and suicide 23:685
of symbolic goods 2:778
urban, in Africa 24:967–969
Economy labor force
shadow 21:862–863
developing and transition countries 21:864–865
OECD countries 21:863–864, 21:864t
worldwide aspects 21:863
Economy of symbolic goods 9:141–142, 9:145
Economy-based accounts 16:5–6
Econophysics 21:531
E-consumers 4:812
E-contracts 4:812
ECOSOC *see* Economic and Social Council (ECOSOC)
Ecosphere, African 1:294–298
Ecospirituality 23:289
Ecosystem evolution 17:367
Ecosystem level, land use intensity indicators at 13:240–241, 13:240f
Ecosystem services 13:239–240, 13:243
sustainable development 23:824
Ecosystems 3:128, 6:872–875, 7:809, 18:350–352
and business platforms 3:37–41
degradation, in Africa 1:295–296, 1:298
energy flow analysis 7:626–627
and sustainability 18:199
typology of 11:844f
Eco-tax reform 7:781
Ecotechnology 12:716
Ecotourism 7:105–109, 24:466–468
definition 7:105–106
misappropriation 7:108
in practice 7:107–108
wider socioeconomic and political context 7:107–108
principles 7:106–107
active engagement 7:106
educative 7:106–107
inclusivity 7:106
small scale 7:107
Quebec Declaration 7:105–106
as subset of sustainable tourism 7:106
sustainable 7:108
as western-centric construct 7:108–109
Ecotypes 9:420–421
E-courses and social sciences 22:629–630
Ecphory 13:681, 20:608, 20:610
ECPR *see* European Consortium for Political Research
ECSC *see* European Coal and Steel Community (ECSC)
Ecstasies 16:168
narrative 16:170
spirit possession 23:282
subdivided on psychological criteria 16:170
unconscious manifestations 16:170
unitive 16:170
world's ecstatic religious traditions 16:168
ECT *see* Electroconvulsive therapy (ECT)
ECtHR *see* European Court of Human Rights (ECtHR)
Ectodysplasin-A receptor gene (EDAR) 11:293–294
EDAR370A 11:293–294
Ectopic activity
phantom limbs 18:1
ECTS *see* European Credits Transfer System (ECTS)
Ecuador 22:625
consumer–citizen in 10:152–153
Ecuadorian cholera intervention project 25:429
Ecumene 7:110–112
borders 7:111
center-periphery relations 7:110–111
civilizations 7:110–111
creolization 7:111
cultural evolution 7:110
diaspora 7:111
diffusion 7:110–111
early perspectives 7:110–111
globalization 7:110–111
modern anthropology 7:111
multiculturalism 7:111
world history 7:111
Ecumenical anthropology 1:737
Ecumenical church history 11:31
Ecumenical council 3:233–234
ED *see Eindrucks* differential (ED); Erectile dysfunction (ED)
EDA *see* Electrodermal activity (EDA); Exploratory data analysis (EDA)
EDAC *see* European Data Center for Work and Welfare (EDAC)
EDAR *see* Ectodysplasin-A receptor gene (EDAR)
EDAR370A 11:293–294
EDC *see* European Defense Community (EDC)
Eddie Mabo and others v. the State of Queensland 2:237
Eddington, Sir Arthur Stanley 9:223
EDE *see* Equally distributed equivalent (EDE) income
E-democracy 6:66, 15:12
Eden, B. 9:646
EDF *see* Empirical distribution function (EDF) tests
Edge centrality 16:535
Edge connectivity 3:907
Edgerton, David 4:762–763
Edgeworth, Francis Beaufort 7:113
Edgeworth, Francis Ysidro 7:113–116
contributions to statistics 7:115–116
life of 7:113–114
on mathematical psychics 7:114–115
as member of Savile Club in central London 7:113–114
prolific writings in economics and statistics 7:114
teaching career 7:113
works
Economical Calculus 7:114–115
Edgeworth series 7:116
Mathematical Psychics 7:113–114
Mécanique Céleste 7:114
Mécanique Sociale 7:114
Methods of Statistics 7:116
New and Old Methods of Ethics 7:114
Utilitarian Calculus 7:114–115
Edgeworth, Maria 7:113
Edgley, C., works, *Life as Theater* 6:647
E-diasporas 6:354
Edible biomass
human–environment relationship 7:615–616
Edinburgh Depression Scale 15:190–191
Edinburgh Handedness Inventory (EHI) 2:807
Edinburgh Postnatal Depression Scale (EPDS) 7:932, 17:723
Edinburgh School 21:309
Edinburgh Strong Program 21:283
Edison, Thomas 24:154
Edited text 14:241
Editing of institutions 17:392–393
Editing of response 19:772
Editorial imagination 19:626–627
EDL *see* English Defence League (EDL)
Edmonton Narrative Norms Instrument 3:796
Edna McConnell Clark Foundation 19:556
EDNOS *see* Eating Disorder Not Otherwise Specified (EDNOS)
Educated bourgeoisie 2:786–787
Education 3:176, 7:121–123, 7:132–136, 7:299, 10:247, 12:196, 14:752, 15:596, 18:129–130, 22:606, 22:612–613, 22:861–865, 24:604–605
about Genographic Project 10:24
and adolescence 4:635
adult *see* Adult education
adult basic 7:122–123
of advanced practice nurses 16:911
aesthetic 1:229–235
affirmative action programs in 1:262–267
advent, growth, and persistence in higher education 1:264–265
partial retrenchment 1:265–266
agentic theory 22:863
American Indians 1:635
anthropology of 1:728–733, 7:298
cultural acquisition 1:730
cultural production 1:730–732
cultural transmission 1:730
directions and trends in United States 1:731–732
ethnic minorities 1:729, 1:731
ethnography 1:728–731
key concepts 1:730
learning 1:728–730, 1:732
multiple literacies 1:731–732
racial minorities 1:729, 1:731–732
research methods and related disciplines 1:730–731
social reproduction of inequality 1:731

teaching 1:728–729, 1:732
as tradition of inquiry 1:728–730
transculturation 1:732
transnationalism 1:731–732
application of Piaget's theory to 18:130–131
application of social interdependence theory in 4:859–860
applied behavior analysis for 2:410
and arts socialization 2:10–11
aspects of self-concepts 21:469
assimilation and contrast in 22:214
of attention 10:131
of Australian aboriginals 2:236–238
benefits 21:125
biological theories 22:862
behavior genetics 22:862
sociobiology and evolutionary psychology 22:862
Bruner on 2:878
and cash transfers 3:217
in China 2:65–69
cognitive ability 7:132, 7:135
college premium 7:134–135
and communitatianism 4:315–316
comparative research in 4:427
descriptive research 4:430
future 4:436–437
global education 4:431–432
globalization and leadership of United Nations 4:428
International Association for Evaluation of Educational Achievement (IEA) 4:427, 4:429
international education 4:429–430
key findings of IEA research studies 4:432–433
model of worldwide activity 4:428–429
organizational and developmental education 4:429
relationships, processes, and products 4:430–431
research program of IEA 4:432
research within organized developmental programs 4:429
systemic research and the use of models 4:434–436
UNESCO 4:427–429, 4:431
unfolded model of worldwide activity 4:428f
World Conference on Education for All 4:429
competency 24:520
computer games and academic achievement 4:498
and conditional cash transfers 4:557–559
and conflict, connections between 4:633–634
consultation–liaison psychiatry 14:4–5
contribution of Thorndike to 24:306
and control 1:546–547
cross-cultural study of 5:327–333
in cross-national perspective 5:342–348
analysis 5:347–348
cultural differences 5:346
data samples 5:346–347
ecology of student achievement 5:342–343
economic differences 5:343–344
economic segregation 5:346
family inequality 5:345–346
family socioeconomic status 5:344
future research 5:348
income inequality 5:345
inequalities and their mechanisms 5:345–346
methodological issues 5:346
microsystem effects in richer *versus* poorer countries 5:345
school inequality 5:346
school resources 5:344
schoolmates 5:344–345
teachers 5:343–346
tests and questionnaires 5:347
and cultural capital
see also Cultural capital, education and work
and cultural resource management 5:490
and culture of poverty 5:592–594
defined 7:256
definition 4:632–633
democratization of 5:449–450, 22:329
in developing countries 22:359
developmental behavioral genetics and 6:320–325
disadvantage, of Australian Aboriginals 15:900–901
and divorce 6:607–608
early childhood 3:397
ecological systems approach 22:864
economic 7:21–25
economic growth 7:40–43
and economic growth 7:121–126
economic growth and 7:123–124
economic resources 7:132–134
economics of 7:149
class size 7:152–153, 7:153f
cognitive skills 7:150–151
Coleman Report 7:151–152
competition and incentives 7:154–155
efficiency 7:153–155
finance of schools 7:155
importance of measured school inputs 7:151–152, 7:152t
outcomes 7:149–151
production 7:151–153
production function approach 7:149–152
school choice 7:154–155
secondary school completion rates 7:150, 7:150f
teachers and schools 7:152–153, 7:153f
educational attainment and levels of lifetime earnings 7:124
educational institutions and society 7:299
effects of income on 7:132–134
net effects 7:132–134
overall relationship 7:132
effects on earnings 7:134–135, 7:135t
sociological approaches 7:135
and employment 7:127–131
credentialism 7:127–128, 7:130–131
education attainment trends 7:129, 7:129t
educational expansion 7:127–131
imperfections and limits 7:128–129
income differences 7:129
knowledge society 7:127–128, 7:130–131
match between 7:129–130
overeducation 7:127, 7:130
professional relevance 7:127–128
transition 7:130–131
unemployment and employability 7:130
encounter between philosophy and 7:257–258
engineering 7:641–642, 7:644–646
entry to higher education 7:133
environmental 7:733
equality 5:766–767
equity in 7:918–923
access 7:918–922
employment 7:918–919, 7:921
graduates 7:919, 7:921–922
interventions to achieve 7:919–921
learning 7:918–922
measurements 7:921–922
mobility 7:919, 7:922
opportunities 7:918–922
performance 7:919, 7:921
policies 7:920–922
poverty 7:918, 7:920–922
quality 7:918–922
redistribution 7:922
resources 7:918–919, 7:921–922
social change 7:919, 7:921
expansion of 22:428
expenditures 5:767
fertility 9:20–21
and fertility transition 9:96
finance 5:772
formal system 7:122, 7:124
and gender 7:117–120, 7:200, 9:682–688
geography of 7:165–171
community development 7:167–169
early roots and development of 7:165–166
educational achievement 7:165–166, 7:168–169
educational attainment 7:165–169
enrollment 7:166–167
ethnic minority 7:165, 7:168–170
institutions of higher education 7:169–170
knowledge 7:165–166, 7:169–170
literacy 7:165–166, 7:168
main research topics 7:166–167
marriage register 7:165
mobility 7:165–170
organization, size, and location patterns of elementary schools 7:167–168
school 7:165–171
sociocultural milieu 7:168
teacher 7:165, 7:167–169
university 7:166, 7:169–170
for gifted and talented 7:158–164

Education (*continued*)
globalization in 10:247–253
higher education research 10:862–869
higher level 7:124
history of 7:172–177
books 7:176
comparative-historical research 7:176–177
in connection with systemic self-reflection 7:174–175
educational actors and the rediscovery of experience 7:175
as field of social science study 7:175–177
gender 7:175
from history of educational ideas to sociohistory of knowledge 7:176
human science 7:174–175
institutions 7:172–176
nation-states 7:173, 7:175
primary education 7:172–173
rise of modern educational system 7:172–173
school practices and a new concept of culture 7:175–176
secondary education 7:173
social structures 7:172–173
society 7:172–175
universities 7:173, 7:176
household economics 11:224–225
and imitation by infants/young children 11:624
inflation 11:591
and illiteracy 11:591
informal system 7:122
information and communication technology in 4:495–496, 12:66–71
intelligence changes by 12:320–321, 12:320f
and international advertising 12:456
and international organizations 12:538–541
of journalists 12:858–859
and knowledge 13:72–73
in Latin America and the Caribbean 7:178–184
leadership 13:655–663
and life expectancy 1:148
and life history evolution 9:32–33
and literacy 14:237, 14:239
major theoretical approaches 22:861–864
markets for 22:20–21
mathematics 14:816–821
mentoring 15:235–238
in Middle East and North Africa 7:210–214
educational quality 7:211–213
minority languages 7:213–214
quantitative expansion 7:211
religious education 7:213–214
structure 7:210–211
youth employment 7:213–214
and migration 15:435
and military schools 7:299
millennium development goals 4:632
and morbidity compression 4:464–466
multicultural 8:132–133
and names 17:225
of nonphysician health professionals 16:908–909
of nurses 16:910
object relations theory 22:862–863
and occupational opportunities 22:427–428
outcomes, operationalizations of 5:377–378
qualitative 5:377–378
quantitative 5:377
parents and teachers as partners in 17:499
and participatory action research 17:537
peace education 17:659
and poverty 18:744
in psychiatry *see* Psychiatric education
psychological theories 22:862–863
attachment theory 22:862–863
epigenetic theories 22:863
learning theories 22:863
psychoanalytic theories 22:862
qualifying function of 7:127–128
quality and equity in 7:227
quantification in 19:704
question of purpose in 7:256–257
for reducing prejudice in society 18:843
regression analysis 7:132–133
returns to 7:134–135
right to *see* Right to education
rights to children 3:486–487
Rogers on 20:737–738
Sarmiento on 20:907–909
schooling as modernizing function 7:124
secondary school 4:635
sexuality and social work 21:806
significance of motivation in 24:186–187
skills training 7:215–217, 22:20–24
and skills training
definition 7:215
generic and basic skills training 7:215–216
measurement 7:216
qualifications 7:216–217
skill acquisition and utilization 7:216
Skinner on 22:27–28
and social capital 22:174, 22:178–179
social cognitive learning theory 22:863
and social mobility 22:427
social studies *see* Social studies education
social work *see* Social work education
and social work 14:664, 14:666–668, 22:809–812
social work, as catalyst 22:792
social work, internationalization of 12:557
social work education 22:775–781
and socialization, integrative framework for studying 22:864–865
characteristics of individuals to be socialized 22:864
goals and values 22:865
proximal processes 22:864
self-socialization 22:865
units of environmental influence 22:864
and socioeconomic status 7:132–135
sociological theories 22:863–864
conflict theories 22:864
structural functionalism 22:863
symbolic interactionism 22:863–864
sociology, Randall Collins 4:211, 4:214
in South Asia 7:194–197
Soviet studies 20:825–830
standardized effects of income on 7:132–133, 7:135
status-distributive function of 7:127–128
structural inequality in 8:133–134
teacher expectation of students and labeling 8:134
tracking system and within-class ability grouping 8:133–134
unequal access to school and desegregation policy 8:133
in Sub-Saharan Africa 7:198–209
basic education 7:198, 7:202, 7:204
education and economic development 7:204
education and gender 7:200
education for all 7:198
enrollment and retention 7:199–205
grade repetition 7:200
learning achievement/educational quality 7:202–204
literacy 7:202–203
preprimary schooling 7:199
primary completion rates 7:200
primary education 7:199–200
SACMEQ assessments 7:202
secondary education 7:200–202
teacher education 7:204
teachers and factors affecting teaching process 7:204–205
tangible user interfaces (TUI) in 24:22–23
augmented reality (AR) 24:22
challenge of 24:23–24
constructive assembly 24:22
design of 24:23
embedded phenomenon 24:22–23
immersive environments 24:22
instructional context 24:24
integration with curriculum 24:24
interactive surface 24:22
malleability 24:24
participation 24:24
scalability 24:24
student adjustment 24:24
token + constraint system 24:22
teacher *see* Teacher education
and teacher leadership 24:60–61
teaching infrastructure 22:667–673
theory and practice, coconstructivism in 3:913–917
in United States 22:428
and values 25:28–29
and wealth flows 9:35–36
writing systems 25:780–784
and youth culture 25:810–811
Education Act 21:66
Education and learning 7:137–143
cognitive aging 7:140–141
cognitive epidemiology 7:139
cognitive reserve 7:140

education in adulthood 7:141–142
'full' tests 7:141
implications for education 7:138
innovating life span models of education 7:142
intelligence 7:138–139
life span cognitive development 7:137–138
holistic and multifunctional 7:137–138
multidirectional development 7:138
plasticity 7:138
life span development, effects of early education on 7:138–141
Reserve capacity 7:139–140
resilience 7:139–140, 7:142
structural lag 7:137
Education and vocational programs, offender rehabilitation 20:194
Education aspiration
coeducation schooling 3:929
single-sex schooling 3:929
Education citizen science projects 3:632–633
Education data
education finance 5:772
education-level graduates 5:772
labor force 5:772–773
staff 5:771–772
student enrollments 5:771
system-level 5:772
Education for a sustainable future (ESF) 7:734
Education for All (EFA) 4:427, 4:431–432, 7:198–199, 10:853
early childhood care and education (ECCE) 6:740
World Conference on 4:429
Education for All Handicapped Children Act (EAHCA) 6:474, 13:471–472, 23:221–222
amendments to 23:223–224
major principles of 23:222t
Education for sustainable development (ESD) 7:733
approaches to 7:735–736
definition 7:733–735
education for a sustainable future 7:734
future prospects 7:736–737
instrumental 7:735–736
interdisciplinarity 7:733–735
intrinsic 7:735–736
levels of learning 7:736
problems and contradictions 7:735
and resilient communities 7:736–737
social learning 7:735–736
social–ecological systems 7:736–737
unsustainability 7:733, 7:735, 7:737
Education innovation 7:267, 7:271
Education inputs, class size as 3:751
Education levels 20:293
Education market 7:118–119
Education of Mentally Retarded Children Act 1958 23:221
Education of the Handicapped Act 1970 (EHA) 23:221
Education policy 7:266–267
adoption of 7:270
on class size reduction 7:271
comparisons of policies and outcomes 7:268–269
design 7:269–270
development of 7:266
acors involved 7:267–268
in Canada 7:267
equity 7:266
future of 7:271
by policy entrepreneurs 7:268
problem definition and agenda setting 7:268–269
top-down foundation efforts 7:268
voluntary state applications 7:267–268
global context 7:266
implementation process 7:270
instruments or tools 7:269–270
multilayered governance and multiple education reform policies 7:271
research and 7:270–271
Education politics 7:266–267
Education production 7:218
Education production functions 7:225–226
Education reform activism 7:267, 7:271
Education research, and social network analysis (SNA) 22:461, 22:463–464
Éducation solitaire 20:769
Education statistics 5:763, 5:766
education equality 5:766–767
education expenditures 5:767
gender 5:766–767
social background 5:766–767
student demographics 5:767
Education surveys 5:847
international 5:847
national 5:847
Education tax credits 21:71–72, 21:76
Education transformation 12:66, 12:70
Education Welfare Officer (EWO) 19:302
Educational achievements 5:771, 5:773
comprehensive model of 4:436f
in different welfare regimes 25:484–485
degree of educational equality according to social background 25:485
educational level of general population 25:484–485
ethnicity and 8:130–135
Educational actors, history of 7:175
Educational agents, international organizations as 12:538–539
Educational attainment 20:294, 20:294t
and cultural capital 5:378
and cultural lifestyles 22:332–333
and gender 9:683
and personality traits 17:921–923
Educational change, and school improvement 21:91–92
Educational choices 22:326–327
educational content to children 7:246, 7:248
Educational continuum 13:736–738
continuing professional development 13:738
initial stage 13:737
vocational stage 13:737–738
Educational contribution 19:771
Educational credentials 7:122
Educational deprivation 22:286–287
Educational desegregation 6:218
Educational effectiveness 21:125
factors of 21:93
field of 7:224–228
research, classroom level in 3:833–838
theory and research 7:218–223
Educational effectiveness modeling 7:218–219
Educational effectiveness models 21:120, 21:123
Educational effectiveness research (EER) 7:218, 7:224, 21:80–81
analysis of the models 7:218–219
contribution of 7:227
dynamic model of 7:219–221, 7:220f
economic approach 7:218
effectiveness modeling 7:218
historical review of 7:224–226
first phase 7:225
fourth phase 7:225–226
second phase 7:225
third phase 7:225
input–output studies 7:218–219
methodological developments 7:226–227
testing the validity 7:221–222
Educational effectiveness theory 7:225
Educational environment instruments 3:828–830
associations between student outcomes and climate 3:829
evaluation of educational innovations 3:828–829
teachers' attempts to improve classroom climate 3:829–830
Educational environments 7:820–821
classroom environment 7:820
home environment 7:821
impact on self-concepts 21:472–473
improving 7:821–822
links between different 7:821
personal forms of educational environment instruments 7:822
qualitative data-collection methods 7:822
school environment 7:820–821
Educational evaluation 7:229–231
conceptualization 7:230–231
implementation across the LAs 7:231
intervention and implementation 7:230
methods 7:229–230
combined methods designs 7:229–230
micro and macro contexts of a program implementation 7:230
randomized controlled trials (RCTs) 7:229
Educational exhibitions 24:882, 24:886
Educational expansion, and cultural participation 5:447, 5:449–450
Educational goals, and school improvement 21:91–93
Educational governance 7:262–265
Educational ideas, history of 7:176

Educational implications, academic self-concept and achievement 1:56–57, 1:59–61
Educational inequalities 22:329
  Goldthorpe on 10:282–283
Educational initiatives, armed conflict 1:973–974
Educational institutions 7:232–237
  and area and international studies 1:943–947
  emergence of educational systems 7:236
  global model of schooling 7:236–237
  humanistic movement and Reformation 7:234–235
  Middle Ages 7:233–234
  origins 7:232–233
  reforming of universities 7:235
  role in intimate partner abuse prevention 12:680–681
  schooling in eighteenth and nineteenth centuries 7:235–236
Educational intervention, cardiac rehabilitation 20:190
Educational learning theory 7:238–245, 12:197
  changing paradigms 7:244
  cognitive perspective 7:240–242
    constructing beliefs 7:241
    controlling 7:242
    knowledge acquisition 7:241
    knowledge construction 7:241
  educational policies 7:244
  and environmental stimuli 7:238–240
  individual differences in learning 7:243–244
  involuntary learning 7:240
  learning as social interaction 7:242–243
    Bronfenbrenner's ecological system 7:243
    social cognitive theory 7:243
    sociocultural theory 7:242
  strengthening a behavior 7:240
  voluntary learning 7:240
  weakening a behavior 7:240
Educational linguistics 14:227–228
Educational marital matching 2:102t–103t
Educational media 7:246–254
  changes in focus/emphasis over time 7:251
  definitions 7:246–247
  knowledge base on learning from 7:249
  learning from 7:249–251
  measuring media exposure 7:252
  models and theories of learning from 7:247–249
  neuroscience of 7:251–252
  trends 7:251–252
Educational media use, effects on children 15:36
Educational neuroscience 16:728–735
  dyscalculia 16:732–733
  dyslexia 16:730–732
  executive function 16:733
  history 16:728–729
  implications 16:733–734
  individual variability 16:733–734
  literacy development 16:731–732
  need for 16:731–733
  number and mathematics 16:732
  pedagogy 16:729, 16:733–734
  prefrontal cortex 16:728, 16:733
  public health model 16:730
  science learning 16:732–733
Educational opportunity
  cases 23:220–221
  equality of 8:751
    family and schooling 8:751
Educational outcomes 17:787
  alternative personality models 17:789–790
  cognitive predictors of 17:787
  and five-factor model 17:787–789
    agreeableness 17:788
    conscientiousness 17:787–788
    emotional stability 17:788–789
    extraversion 17:789
    openness 17:789
  moderating factors 17:790
  noncognitive predictors of 17:787
  and personality 17:787–791
Educational paradigms 13:699–711
Educational philosophy 7:255–260, 20:768–770
Educational planning 7:261–265
  current practice 7:263–264
  evaluation 7:262, 7:264
  history 7:261–262
  human capital theory 7:261
  incentives 7:265
  inequality 7:261–262, 7:264
  long term planning 7:262–264
  monitoring 7:262–264
  participation 7:261, 7:264
  partners 7:262, 7:264
  projections 7:261
  quality education 7:262, 7:264–265
  simulations 7:263–264
  supply and demand 7:264
  from traditional planning to strategic planning 7:262–263
Educational policies 7:244, 25:481–490
  comparative perspective 7:298–299
  future development 25:489–490
  influence of welfare policy on educational outcomes 25:483–485
    degree of educational equality according to social background 25:485
    educational level of general population 25:484–485
  mechanisms for development of 7:266–271
  private schooling 25:482–483, 25:487–488
  public schooling 25:483, 25:485, 25:487–488
  relevance in developing countries 25:485–488
    application of welfare state typology to developing countries 25:487–488
    familiarist regime 25:488
    in productivist welfare regime 25:488
    in protectionist regime 25:488
  and school improvement 21:92–93
  and school safety 21:107
  and social policy
    combination in three regimes 25:482–483
    educational systems, structures of 25:483
    policy traditions in welfare states 25:481–482, 25:481t
    relationship between 25:481–483
  social welfare regimes 25:482
  tracking 25:485
  typology of welfare regimes in developed countries 25:486t
  typology of welfare state regimes in developed countries 25:489t
Educational Policy Analysis Archives (EPAA) journal 7:182–183
Educational production function 21:81
Educational productivity 7:272–278, 7:820–821
  in international context 7:273–274
  productive giftedness model 7:274–276, 7:274f
  significance of 7:272
  in terms of economic investments 7:272–273
Educational psychologist, conception interest 25:240
Educational psychology 7:240, 12:203–204
  personality assessment 17:831–832
    domains and methods of assessment 17:831
    motivation and volition 17:831–832
  twentieth-century, experimentation in 8:568
Educational quality 7:211–213
Educational reform, LAC educational system 7:179, 7:183–184
Educational research 7:279–288
  characteristics of profession 7:281–282
  demographic profile 7:281–282
  disciplinarity and interdisciplinary linkages 7:280
  diversity and doctoral production 7:281
  doctoral preparation 7:280
  doctorate degree production 7:281
  emergence of 7:279–280
  field 7:279
  higher education 7:279
  international footprint of 7:284–285, 7:285t
  journals and knowledge dissemination 7:283–284, 7:284f, 7:284t
  overall patterns 7:281, 7:281t
  professional associations 7:283
  professionals and workplace 7:282–283
  scientific fields 7:282, 7:284–286
  sectors of employment 7:282
  as social science 7:279–280
  structure and context of training 7:280–281
  taxonomy of doctorate programs 7:280–281
  in universities and research and development organizations 7:282–283

Educational resources, in home, and cultural capital 5:377
Educational results, inequality 22:327
Educational revolution 1:148
Educational settings
families as 8:731
Educational sociology 7:289–296
accountability movement and teacher burnout 7:294–295
conflict theory 7:290, 7:293
critical theory 7:290–291
formal education 7:289–290, 7:293
functionalist theory 7:290, 7:293
institutionalist perspective 7:290–291
interactionist theory 7:290
multicausality 7:292
network analysis 7:292–293
new sociology of education 7:290
or sociology of education 7:289–290
qualitative methods 7:291–292, 7:294
quantitative methods 7:291–293
research methods 7:291–293
school violence and bullying 7:294
social inequalities 7:293–294
standardized testing 7:293, 7:295
theories 7:290–291
Educational sovereignty 10:252
Educational staffing 5:771–772
Educational systems
emergence of 7:236
and gender development 9:746–747
in Japan 12:795–796
in North America 7:297–300
structures of 25:483
Educational theories, of cognitive development 4:51–57
Educational training of lawyers 13:629
Education-based meritocracy 22:325
Education-level graduates 5:772
Eductive learning 8:514
EDUG (software) 9:837–838
Edwardian legislation 25:528–529
EDZs *see* Economic Development Zones (EDZs)
EE *see* Environmental education (EE); Environmental enrichment (EE)
EEA *see* Environment of Evolutionary Adaptedness (EEA); European Economic Area (EEA); European Environment Agency (EEA)
EEC *see* European Economic Community (EEC)
EEG *see* Electroencephalogram (EEG); Evolutionary economic geography (EEG)
EEIC *see* English East India Company (EEIC)
EE-IOA *see* Environmentally extended input output analysis (EE-IOA)
EEO *see* Equal Employment Opportunity (EEO)
EEOC *see* Equal Employment Opportunity Commission (EEOC)
EER *see* Educational effectiveness research (EER); Effective end of reproduction (EER)
EES *see* European Election Study (EES)
EF *see* Ecological footprint (EF); Executive function (EF)
EFA *see* Education for All (EFA); European Free Alliance (EFA); Exploratory factor analysis (EFA)
Effect heterogeneity, optimal experimental design 17:256
Effect indicators 11:751
Effect sizes 8:345–346, 8:348
in gender differences 9:750–751, 9:750f
meta-analysis 15:272, 23:909–910
analyzing across a set of studies 23:910–912
Effective classroom
across prekindergarten (pre-K) to 12 environments 3:833
behavioral, emotional, and persistence oriented markers 3:833
goal of 3:833–834
multiple perspective 3:836
observer's perspective 3:835
professional development and 3:836–837
student's perspective 3:835–836
teacher's perspective 3:834–835
teaching through interactions framework 3:834t
Effective co-teachers 21:116
Effective demand, principle of 19:115
Effective end of reproduction (EER) 17:174
Effective performance appraisal
criteria for 17:716
leadership effect on 17:719
Effective program archives 1:919
Effective schools 1:46–47
Effective teacher paradigm 24:52
Effective teams 21:114
Effectiveness 1:9, 4:298–299
evidence-based psychotherapy (EBP) 8:345
global governance 10:159
of interventions
for learning 7:230–231
psychological treatments 19:382–384
in managerial and organizational cognition 14:479, 14:481
nonprofit organizations 16:915, 16:917
and probation 19:44–46
programs 19:117–118
of randomized controlled trials 8:542–544
of treatment, pharmacoeconomics 19:458
Effectuation 23:504
Effectuation theory, of entrepreneurship 7:682
Efferent neuronal connections of hypothalamus 11:486
Efferent neuroprostheses 16:718–719
invasive 16:718–719
noninvasive 16:719
Efferent–afferent firing contingencies 20:211
Effervescence 4:153
Efficacy 5:55–56, 7:242
and adolescent health 1:124–125
evidence-based psychotherapy (EBP) 8:345
of randomized controlled trials 8:542–544
of sanctions 20:891
Efficacy beliefs
and academic achievement 1:21–22
of teachers 24:45
Efficacy paradox, in placebo-controlled double-blind studies 18:161–162, 18:161f
Efficacy trials, pharmacoeconomics 19:458
Efficiency 2:920, 2:922, 3:823, 5:55–56, 22:840
and bureaucracy 2:915–916
in community care 4:346–347
overall impacts 4:347
economics 7:87
education 7:153–155
evidence-based psychotherapy (EBP) 8:345
of experimental designs 8:539
GMM estimation 15:296
for location procedures 16:896
and market allocation 14:539–540
McDonaldization 14:828
norms 17:8
and organization theory 1:104–105
regulatory agencies 20:181, 20:183–184
as unifying forces shaping organizational forms 17:417–419
in welfare economics 25:497–501
Efficiency analysis 18:277–278
cost–benefit analysis 18:277–278
cost-effectiveness analysis 18:278
Efficiency wage models 24:741
Efficiency-seeking, foreign direct investment 9:324–325
Efficiency-seeking behavior 12:580
Efficient bargaining model, trade unions 24:501–502
Efficient importance sampling (EIS) 15:760–761
Efficient market hypothesis (EMH) 23:487
Efficient market theory 23:482
fundamental analysis 23:482
technical analysis 23:482
Efficient utilization of labor 6:269–272
Effluent charges 7:727–728
Effort–distress model 17:135–136
Effortful control (EC) 6:13–17, 17:788, 24:182, 24:193, 24:195–197
and behavior during infancy and early childhood 6:13
Effort–reward imbalance (ERI) model 17:136, 17:148, 23:548, 25:538, 25:734
and demand–control 17:137
and justice 12:930
for work-related stress 23:548–550
Efforts
beliefs and academic motivation 19:792–794
as desert base 6:232
and reward balance 17:137
Efklide, A., model of self-regulated learning 21:536
EFPPA *see* European Federation of Professional Psychologists Associations (EFPPA)

EFPT *see* European Federation of Psychiatric Trainees (EFPT)
EFS *see* Excluded-from-School (EFS) children
EFs *see* Executive functions (EFs)
EFSA *see* European Food Safety Agency (EFSA)
EFT *see* Emotion-focused therapy (EFT)
EFTA *see* European Free Trade Area (EFTA)
Egalitarian 2:149, 2:153, 14:254–255
Egalitarian core 4:871
Egalitarian hunter-gatherers 4:600–601
Egalitarian societies 11:204, 11:220–221, 11:894–895
  assertively 11:894–895
  modern 11:898
  relatively 11:895
Egalitarianism 2:242, 7:91, 7:308–314, 12:915, 21:212
  and big man 2:568
  demands, site of 7:311–312
    dualism 7:311–312
    monism 7:311–312
  and feminist psychotherapy 8:931
  and indigenous peoples 11:805
  and majority rule 14:446–447
  modern thought 7:308–309
  and moral development 15:781–782, 15:784–786
  Pax Britannica 25:366
  post-Rawlsian egalitarianism 7:309–311
    comprehensive 7:310–311
    narrow 7:310
  and prejudice 18:834
  scope of demands 7:312–313
    cosmopolitan 7:312–313
    particularistic 7:312
  and sexism 21:695
  and teaching 24:77
  in United States 1:619
  and values 25:30
  and welfare 25:476
  in West Africa 1:301
Egalitarian-over-individual-worths value 4:869–870
Egalitarian-over-principal-contributions value 4:870
Egan, James P. 21:954
Egg donation, and commodification 5:538
Ego 10:484–485, 15:894, 19:338–339, 19:400–401
  aggression directed against 23:677
Ego alliance 24:287–288
Ego developmental stages 7:935
Ego identity 5:860, 17:745–746, 21:432–435, 22:313
Ego Identity Scale 7:941
Ego network 16:546, 22:462
Ego psychology
  conceptions of mind 22:761
  psychodynamic treatment 22:764
  psychopathology in 22:763
Ego Stage Development 7:941
Ego weakness 2:256–257
Ego–Alter–Object 20:444
Ego-centered networks 16:522
  network analysis 16:520, 16:522
Egocentric distance, and visual space 25:197
Egocentric frames 17:704
Egocentric network data 22:462
Egocentric representation 17:704
Egocentric spatial representations 11:386, 11:388–389
Egocentric spatial terms (language) 13:295
Egocentric speech, Vygotsky on 25:313
Egocentrism 15:3, 21:269
Ego-depletion theory 4:741–742
Egodystonic homosexuality 3:888
Egodystonic sexual orientation 3:888
Egoism 16:583–584
  and empathy 19:217–218
Egoistic giving 3:364
Egoistic suicide 1:719, 23:683
Egonets 22:474
E-government 3:698
Egyetemes néprajz 1:754
Egypt 16:801–802, 21:794
  ancient 2:919
  Civil Code 15:361
  legal system in 15:363
  literary activity in 16:362
  psychiatric care in 6:42
Egyptian blogging community 16:778
Egyptian hieroglyphics 9:266
Egyptian hieroglyphs 3:815t
Egyptian revolution 16:802
Egyptian writing 25:774–775
Egyptology 20:322
  of ancient Egypt 1:691, 1:693–694
EHA *see* Education of the Handicapped Act 1970 (EHA)
eHANPP *see* Embodied HANPP (eHANPP)
EHEA *see* European Higher Education Area (EHEA)
eHealth 5:785
E-health 10:624–625
  rheumatoid arthritis 20:658
EHH *see* Extended haplotype homozygosity (EHH)
EHI *see* Edinburgh Handedness Inventory (EHI)
EHPS *see* European Household Panel Survey (EHPS)
Ehrenberg, Richard 3:21
Ehrenburg, Ilia 20:823
Ehrlich, Anne, works, *The Population Bomb* 20:458
Ehrlich, Eugen 13:550, 13:802
  state law's constitutive power 13:550
Ehrlich, Paul R., works, *The Population Bomb* 6:311, 20:458
EI *see* Emotional intelligence (EI)
EIA *see* Environmental impact assessment (EIA)
Eiasler, Riane 9:461
EIC *see* British East India Company (EIC)
Eichhorn, Carl Friedrich 4:339–340
Eichmann trial 18:498, 18:500
EID *see* Ecological Interface Design (EID)
Eidetic reduction 19:315–316
EIF *see* External-internal fit (EIF) paradigm
Eigendynamik 8:837
Eigensinn 11:111
Eigenvalue decomposition (EVD) of covariance matrix 8:718
Eigenvalues-greater-than-one rule *see* Kaiser-Guttman rule
Eigenvector, network centrality 16:534–535, 16:537–539
Eigenvector measure 16:535
Eight-Queens problem 2:33f
Einbildungskraft 25:287–288
Eindrucks differential (ED) 21:567
Einhard, Eginhard 2:645
  works, *Vita Caroli Magni* 2:645
Einstein, Albert 6:396, 24:674
  general theory of relativity 19:64
  Special Theory 17:89
EIS *see* Efficient importance sampling (EIS)
Eisenberg, Nancy, prosocial responding theory 7:315–320
Eisenstadt, Shmuel Noah 2:922, 7:321–326, 20:745
  awards 7:321
  biography 7:321
  on Charisma 7:321–322
  civilizational analysis of 7:323
  on collective identities 7:323
  on comparative-historical sociology 7:321
  on culture and institution-building 7:321
  and Israeli society 7:323–324
  in Japanese civilization 7:323
  in Jewish civilization 7:323
  and macro-sociology 7:321–324
  on multiple modernities 7:321–322, 7:324–325
  on revolutions 7:322
  on social change 7:322
  on social exchange 7:322
  and sociological theory 7:324
  on tradition and modernization 7:322, 7:324
  works
    Civilta comparate 7:323
    The Early State in African Perspective 7:323
    Essays on Comparative Institutions 7:322
    The Form of Sociology: Paradigms and Crises 7:322
    From Generation to Generation 7:322
    Globality and Multiple Modernities: Comparative North and Latin American Perspectives 7:323
    Intellectuals and Tradition 7:322
    Japanese Civilization: A Comparative View 7:323
    Jewish Historical Experience: The Civilizational Dimension 7:323
    Knowledge and Society: Studies in the Sociology of Culture Past and Present 7:322
    Kulturen der Achsenzeit 7:323
    Max Weber on Charisma and Institution Building 3:355
    Modernization, Protest and Change 7:322

Multiple Modernities — A Paradigm of Cultural and Social Evolution 7:325
The Origins and Diversity of Axial Civilizations 7:323
Paradoxes of Democracy: Fragility, Continuity and Change 7:324
Patrons, Clients and Friends 7:322
The Political System of Empires 7:322
Protosociology 7:325
Revolution and the Transformation of Societies 7:322
Social Differentiation and Stratification 7:322
Society, Culture and Urbanization 7:322
Tradition, Change and Modernity 7:322
Eisenstein, Elizabeth 18:915, 18:917–918
Eisenstein, Sergei, works, *The Battleship Potemkin* 9:167
Eisner, Elliot, works, *Cognition and Curriculum, Reconsidered* 1:230
EISs *see* Epigenetic inheritance systems (EISs)
Eisuke, Zensho 2:71–72
EITC *see* Earned Income Tax Credit (EITC)
EITMs *see* Empirical Implications of Theoretical Models (EITMs)
EIT-system *see* System of the Methods of Emotional-Imaginative Teaching (EIT-system)
Ejaculatory disorders 21:715–716
inhibited ejaculation 21:716
premature (early) ejaculation 21:715–716
EJM *see* Environmental justice movement (EJM)
EJN *see* European Judicial Network (EJN)
EJSP *see European Journal of Social Psychology* (EJSP)
EJW *see* Equal Justice Works (EJW) Fellowship
EKC *see* Environmental Kuznets curve (EKC)
Ekklesia 6:61–62, 6:65
Ekman's law 19:474
EKS method, for microeconomic data 5:756
EL *see* Errorless learning (EL) for people with language disorders
El mestizaje 23:45–47
El Salvador 22:624
sociocultural aspects 1:621–624
E-lab 3:8
Elaborated social identity model (ESIM) 4:153–154
Elaboration, and senses 21:617–618
Elaboration likelihood model (ELM) 2:170, 15:31–32
of attitude-behavior consistency 2:186
of persuasion 17:943–944
Elaboration model 7:563
Elaboration strategies
for basic learning tasks 13:713
for complex learning tasks 13:713
Elaborative inferences *see* Forward inferences
Elastic Bunch Graph Matching 8:659–660
Elastic concept 1:790
Elastic tournament 13:630
Elasticity
of incentives 1:346
in population dynamics 18:581–582
Elder, Glen 22:150–151
Elder, T., works, *American Jewish Historical Society: Center for Jewish History* 10:6
Elder abuse 7:327–334
categories 7:329–330
discriminatory abuse 7:331
emotional/psychological abuse 7:330
financial/material abuse 7:330
neglect 7:330–331
physical abuse 7:330
sexual abuse 7:331
definitions 7:327–329
domestic abuse 7:329
harm 7:327–332
incidence and prevalence 7:331
institutional abuse 7:328, 7:332
mistreatment 7:327
practice issues and law 7:332
stranger abuse 7:329
victim's definition of 7:329
Elder Abuse Prevention and Caregiver Support Law (Japan) 7:332
Elder care 10:266
advocacy 7:336–339
ageism 7:336
comorbidity 7:337
context for social work with 7:335
policy context 7:336
population aging 7:335–336
critical gerontological social work practice 7:337–339
assessment 7:338
care management and therapeutic interventions 7:338–339
demographic transition 7:335
dependency ratio 7:335
dignity 7:338
and generation 9:853, 9:856
health and disability 7:336–337
health and social care 7:336–337
physical and mental health 7:336–337
and support 7:337
inequalities 7:336–337
life course 7:336–338, 7:340
marginalization 7:336
oldest old 7:335–336
social support 7:335
social work's capacity to respond to needs of aging population 7:339–340
transitions 7:335–337, 7:339–340
Elder care services 15:442–443
expectations of minority ethnic elders 15:441t
mainstream services 15:442–443
voluntary organizations 15:443
Elderhostel 1:150
Elderly
biomedical research on 13:560–561
Jewish Chronic Disease Hospital Study 13:560–561
see also Senior citizens
Elderly caregiving 3:136–137
Elders, importance by, indigenous peoples 11:805
E-learning 6:567, 6:895–903, 7:921, 12:66–67, 12:69, 12:197
apps 6:899
future trends 6:901
hybrids, collaborations, and partnerships 6:896–897
difference between informal and formal education 6:896–897
iCampus 6:897
Project Dragonfly 6:897
immersive environments 6:900
augmented reality 6:900
virtual environments 6:900
intelligent tutoring systems 6:899
learning games 6:900
mobile learning platforms 6:898–899
mobile computing 6:898
tablets and e-books 6:899
online learning communities 6:897–898
blogs 6:898
Internet forums 6:898
podcasts 6:897–898
social networks 6:897
video sharing websites 6:898
Wikis 6:898
open educational resources 6:896
smart phone 6:898, 6:900
traditional 6:895–896
distance learning 6:895–896
learning management system 6:895
video game 6:896–897, 6:900
Election change, linking manifestos to 18:419
Election studies, international comparisons of 5:833
Elections
campaigns 7:346–347
constituencies 7:347
and democracy 6:82–84
and direct democracy 25:294
economic evaluations during 7:26–32
distortions 7:28–29
economic voting theory 7:26
institutions and economic vote 7:29
knowledge of citizens about economy 7:28
observing economic crisis 7:29–30
other evaluations 7:30
pocketbook *versus* sociotropic economic evaluation 7:27
public opinion 7:26
responsibility for economy 7:29
restricted variance problem 7:30
retrospective *versus* prospective economic voting 7:27
survey research 7:27
gender quotas 8:888, 8:891
judicial 12:894–895
opinion polls 20:878
participation 25:309
and party identification 17:562–566
polling 18:459–465
presidentialism 18:864–869

Elections (*continued*)
United States Congress 4:649
voting 7:346–347
Elective share 19:175
Electoral behavior 5:824–829
data infrastructures 5:824–826
implications and challenges 5:828
Electoral cleavages 7:345–346
Electoral competition
bounded rationality for 2:775
and manifestos 18:417–418
Electoral district, census 3:303, 3:305
Electoral formulae 7:350
Electoral fortunes of green parties 10:381–382
Electoral fraud 5:367
Electoral geography 7:345–348
constituency 7:347
local influences on voting decisions 7:346–347
of votes 7:345–346
Electoral participation 18:504–505
Electoral performance of green parties 10:382–383, 10:382t
Electoral policies 17:518
Electoral politics, female quotas 8:890
Electoral realignment, issue evolution 12:765–767
Electoral revolutions 20:624
Electoral support, political parties 18:408
Electoral systems 7:349–352, 8:888–889, 17:517
classification 7:350–351
nonproportional electoral systems 7:351
party-list 7:351
proportional 7:351
STV 7:351
comparative study of 5:793
consequences 7:351–352
definition 7:349
and democracy 6:82
elements 7:349–350
ballot structure 7:349
constituency structure 7:349–350
electoral formulae 7:350
list proportional representation 19:205
and majority rule 14:447
mixed-member proportional 19:205–206
proportional representation 19:205–206
single-transferable vote 19:205
Electorate, effective role in politics 13:12, 13:12t
survey-research portrait of voters 13:12
Electric energy 7:626
Electric telegraph 12:490–491
Electric Telegraph Company 24:153
Electrical activity, and cortical development 5:32–33
Electrical signaling, neurons 16:704–705
Electricity
and chemistry 3:372
development of 7:617–618
early studies 19:502
Electroconvulsive therapy (ECT) 17:29, 19:587–588, 23:16–20
economics of psychiatric care 19:297
LTP/LTD 14:363
for mania 14:489
Electrocorticography (ECoG) 7:353, 7:356, 7:359
Electrode application, electroencephalography 7:354–355
Electrodermal activity (EDA) 16:700
Electrodermal measurement 2:181
Electroencephalogram (EEG) 2:181, 8:664–665, 9:476–479, 11:603, 11:616, 13:490, 16:699–701, 22:486–487
in ADHD 11:605
alpha rhythm 7:356–357
applications 7:359–362
brain death 7:361
cerebrovascular and structural brain disease 7:362
dementia 7:361–362
disorders of consciousness 7:359–361
epilepsy 7:359
psychiatry, psychology, and cognition 7:362
arousal response 10:480
beta activity 7:354, 7:357
brain oscillations investigated 9:483–484
coma 7:356, 7:359–361
conventional 9:478
delta activity 7:354, 7:357
electrical potential, size of 9:478–479
electrocorticography 7:353, 7:356, 7:359
for epilepsy 7:853–854
event-related potential (ERP) technique 9:479
gamma activity 7:354, 7:359
high-density 9:478
hypnosis 11:468–470
images of spontaneous brain activity 9:477f
infant cognitive development 4:60, 4:62
interpretative principles 7:356–359
abnormal classification 7:357, 7:357t
artifact discrimination 7:356
clinical interpretation 7:357–358
montages 7:358–359
normal and benign variant patterns 7:356–357
personnel 7:356
investigating language with 9:481–487
kindling 13:32
for language learning impairment 13:348
mu rhythm 7:357
neuromarketing 16:698
physiologic principles 7:353
recording principles 7:354–356
after recording 7:356
calibration 7:355
electrode application 7:354–355
head measurement 7:354
patient preparation 7:354
personnel 7:354
recording 7:355–356
seizures 7:354–355, 7:357, 7:359, 7:361–362
sharp waves 7:357–360, 7:360f
sleep 22:85, 22:87–90, 22:92
findings 22:82t
source localization techniques 9:478–479
spatial resolution 9:478
spikes 7:353–354, 7:357–361, 7:360f–361f
technical principles 7:353–354
theta activity 7:354, 7:357
Electromagnetic spectrum 20:419–420
Electromyography 2:181
Electron microscopy 24:481
Electron spin resonance (ESR) 3:566
Electronic activated recorder (EAR) 23:652
Electronic aggression 5:638, 5:640–641
see also Cyberbullying
Electronic articles, journals 12:867
Electronic commerce *see* E-commerce
Electronic computers 23:377
Electronic dictionaries 5:811–812
challenges 5:812
headwords and organization 5:812
types 5:811
Electronic discovery
in England 6:520
in United States 6:518
Electronic health records, psychiatric ethics 8:68
Electronic immobilizers 19:165
Electronic information, ethnography 8:188
Electronic information technology 1:913
cataloging 1:913
digital files 1:913
high-quality electronic scanners 1:913
hyperlink 1:913
reproduction of documents 1:913
Electronic libraries 7:586
Electronic man 18:918
Electronic monitoring (EM), diversion programs 4:375–376
Electronic navigation 16:347
Electronic payments 15:748
Electronic records, personalized medicine 17:927–928
Electronic resources, social work with refugees 22:770–771
Electronic trade, autonomous agents 2:311
Electronic waste (e-waste) 7:364–373
debates 7:370–371
geography 7:369–370
history 7:369–370
resource 7:369
trade 7:364–371, 7:365f
in China 7:367f
in Mexico 7:368f
in United States 7:366f
value 7:369
Electronic Word-of-Mouth (eWOM) 12:588
Electronically stored information (ESI), discovery 6:518, 6:520–521
Electrophysiological methods *see* Electroencephalogram (EEG); Magnetoencephalogram (MEG)
Electrophysiology 2:651–652
Humboldt on 25:279–280
Electroshock therapy 23:18–19

Elementary and Secondary Education Act 1965 (ESEA) 23:221
Elementary education 7:236
Elementary family 14:600
Elementary schools 7:167–168
Elementary societies 13:70–73
Elementary structures of kinship 23:627–628
Elementary theory of social relations 8:484
Elements in environment 4:487–488
Eleutheromania 8:214
Elevation data 10:61
ELF *see* Earth Liberation front (ELF)
Eli and Edythe Broad Foundation 7:267
Eliade, Mircea 7:374–377, 20:345
  characterization of anthropology by 7:376
  doctoral thesis on yoga 7:374
  life of 7:374–375
    teaching career 7:374–375
  on status of Jews 7:374–375
  theory of religion 7:375–376
    accounting for religion 7:376
    aspects of God 7:376
  works
    Encyclopedia of Religion 7:375
    History of Religions 7:375
    Maitreyi 7:374
    The Sacred and The Profane 20:345
Elias, Norbert 7:378–381, 7:514, 9:368, 16:239, 21:874, 24:746–747
  attempt at analyzing The American Civilizing Process 7:379
  brief biography 7:378
  on European history 21:874–875
  figurational sociology of 7:380
  idea of conceiving of society in terms of processes 7:378
  process of civilization 7:379
  on survival units 25:406–407
  'theory of human beings' (*Menschenwissenschaft*) 7:380
  works
    The Civilizing Process 11:27
    Gesammelte Schriften 7:380
    On the Process of Civilisation 16:239
    The Process of Civilization 2:729
    Studies on the Germans 16:239
    Über den Prozeß der Zivilisation (*The Civilizing Process*) 7:378
Elias–Lewis conjecture 21:872
Elicitation 7:382, 14:185
  of probabilities and probability distributions 7:382–385
    brokers 7:382
    cognitive aspects 7:382–383
    data swapping 7:382
    other issues 7:384–385
    prior distributions 7:383–384
    recoding 7:382
    subjective opinion within Bayesian framework 7:382
    topcoding 7:382
    utilities 7:384
  of startle 23:331–332
Elicitation procedures
  in kinship terminologies 13:63–64
Elimination procedures 9:625–626
Elimination-by-aspects-rule, of decision making 5:919
Eliminative materialism, and intentionality 12:334
Eliminativism 18:112
  in legal philosophy 18:29
Eliot, Charles 24:762
Elite clubs 1:28
Elite members of society 20:334
Elite movements 24:605
Elites 1:965, 1:968, 3:123–124, 3:126, 7:386, 23:632–633
  anthropology of 7:386–389
  authority of 2:270
  capitalism 7:386–388
  celebrity 3:274
  and civilization 3:726–727
  classic theorists 7:391–392
    functional, strategic elites 7:392–393
    governing éLite 7:391–392
    Michels 7:392
    Mosca 7:391
    Pareto 7:391–392
    pluralist elite theories 7:392–393
    political leaders and iron law of oligarchy 7:392
    ruling class 7:391
    theories of Mannheim, Dahrendorf, Dahl, Keller 7:392–393
  colonialism 7:387
  corporate 1:969
  critical elite theory 7:393–394
    Mills's 7:393–394
    power elite, field of power 7:393–394
  democracy 7:390–393
  emergence in Europe 8:270–271
  ethnographic studies of 7:387–389
  and experts 8:588–589
  history of 7:390–395
  and inequality 7:388
  interviewing of 5:854
  kinship 7:387
  Machiavelli on 7:390–392
  marriage market 1:968
  media 7:388
  medieval and modern vocabularies of 1:966
  methods 7:386–389
  neoliberalism 7:388
  partisan 12:766–767
  political 4:692–694
  politics 7:386
  revolutionary 20:646
  science 7:388
  and social mobility 22:426, 22:428–429
  and social stratification 22:709
  sociology of 7:396–401
    analysis, recent developments in 7:398–400
    classical theories 7:397
    critical theories 7:397–398
    functionalist theories 7:397
    habitus 7:398–399
    homogeneity 7:398–399
    internationalization 7:400
    power elite 7:397–398
    recruitment 7:397–400
    ruling class 7:397–398
    subelites 7:397, 7:399–400
    upper class 7:397–400
  studying up 7:387–388
  taste cultures 7:388
  theorization of 7:386
  use of term 7:390–391
Elites of Yahwism 12:869
Elitism 1:965–966, 6:690, 22:841
  and culture 5:462–463
  intellectual 12:280–281
  political agendas 1:362–363
Elitism of science 21:262–263
Elizabeth, Queen 23:733
Elizabeth Clare Prophet 16:807
Elizabethan Poor Law 15:884, 18:747–748
Elkhana, Yehuda, on intellectuals 13:72
Ellenberger, Henri 19:327
Ellickson, Robert 17:7–8
Elliot, G., works, *The Twentieth Century Book of the Dead* 10:7
Elliott, David 25:98
Ellis, Alexander 6:349
  works, *Research on Educational Innovations* 14:757
Ellis, Havelock 19:887
Ellsberg paradox 19:902–903
Ellul, Jacques 21:201
ELM *see* Elaboration likelihood model (ELM)
Elm (*Ulmus*) decline 7:696
Elmina Castle 5:416–417
ELRT *see* Expected likelihood ratio test statistic (ELRT)
ELS *see* Empirical legal studies (ELS)
ELSA *see* English Longitudinal Study on Aging (ELSA); Ethical; legal and social aspects (ELSA) program
Elsenhans, Theodor 19:391
Elsevier 21:322
ELSI *see* Ethical; legal; and social implications (ELSI) program
Elster, Jon 1:668, 7:402–408
  analytical Marxism 7:402
  analytical sociology 7:402
  behavioral economics 7:402
  causal mechanisms 14:875
  effects 7:406–407
  on functionalism 9:540–541
  influences 7:405
  institutional design 7:402, 7:404, 7:406
  intellectual trajectory 7:402–403
  Marxism 7:402–403
  philosophy of social science 7:405–406
  public choice theory 7:402
  rational choice theory 7:402–403, 7:406–407
  theoretical overview 7:406
  works 7:403–404
    Alexis de Tocqueville. The First Social Scientist 24:424
ELTE *see* Eotvos Lorand University (ELTE)

Elucidation of law 4:250
Elucutio 20:650–651
Ely, Richard T. 13:773
EM *see* Electronic monitoring (EM); Evolutionary medicine (EM); Expectation–maximization (EM) algorithms; Explanatory model (EM)
EMA *see* Ecological momentary assessment (EMA); European Medicines Agency (EMA)
E-mail 4:288, 4:504–505, 12:603, 16:522
Emancipation 14:19, 19:308, 19:854, 25:212
  Boltanski on 2:747, 2:751–752
  and community work 4:381
  and liberation theology 14:32, 14:35–37
  and slavery 22:40–41
  treatment research and access 8:34
Emancipation Act 1833 (UK) 22:40
Emancipation Proclamation 3:691
Emancipatory approaches, of social work 19:870–874
Emancipatory politics (Garrett) 22:803
Emancipatory pragmatics 18:801
EMAS *see* European Eco-Management and Audit System (EMAS)
Embarrassment 21:477–478
  causes 21:477
  functions 21:477
  individual differences in embarrassability 21:478
  and shame 21:477–478
Embase 8:328
Embassy of the Blessed Kingdom of God for all Nations 16:806
Embeddability, of orders 17:528
Embedded instruction
  metacognitive skills 24:93
Embedded liberalism 10:182, 13:200, 25:522
Embeddedness 17:302–304, 25:663
  economic sociology 7:67–68
  in entrepreneurship 7:688–689
  of infrastructure 12:119–120
  Polanyi on 18:244–245
  public sector organizations 19:600–601
  rational choice theory 19:901–902, 19:904
Embedding
  cooperation 4:863–865
    institutions 4:864
    social networks 4:864–865
    time 4:863–864
Embedding farming 17:666
Embedding theory of explanation 8:597
Emblematic CS 3:918
Emblems 10:110–111
Embodied action, Harold Garfinkel on 9:598–599
Embodied cognition 10:510–511, 12:40–41
  and mathematics education 14:818–819
Embodied cultural capital 5:375
Embodied human appropriation of net primary production (eHANPP) 13:242
Embodied human capital 7:48
Embodied land 18:352
Embodied language comprehension 13:374–375
Embodied person, concept of 8:944
Embodied phenomenon
  tangible user interfaces (TUI) 24:20–21
Embodied racism 19:867–868
Embodied simulation theory 7:421
Embodied social cognition 7:409–414
  bodily feedback 7:410
  bottom-up embodiment 7:409
  definition 7:409
  grounded social cognition 7:409, 7:412
  moderators 7:412–413
    consistency 7:413
    construal level 7:413
    familiarity 7:413
    modality match 7:413
  motor responses activation by higher cognitive processes 7:411
  sensory-motor activation relation to higher cognitive processes 7:409–411
    body as source 7:410
    contextual characteristics as source 7:410–411
  situated cognition 7:409, 7:411
  theoretical perspective 7:411–412
    soft *versus* hard approach of embodiment 7:412
  top-down embodiment 7:409, 7:413
Embodied technological change (ETC) 12:154, 12:156
Embodied-capital theory of life history evolution 9:28–34
  demographic transition 9:28, 9:30, 9:33
  education 9:32–33
  fertility and parental investment
    action of natural selection on 9:28–29
    among traditional humans 9:29–30
    ecology and proximate mechanisms governing 9:31–33
    fertility under modernizing conditions 9:32–33
    traditional foraging societies 9:31–32
  hunter-gatherers 9:28–30, 9:32
  implications for understanding current fertility transitions 9:33
  inclusive fitness 9:29–30
  intergenerational transfers 9:30
  life span 9:28–30
  menopause 9:28, 9:30
  senescence 9:29
  trade-offs 9:29, 9:32–33
Embodiment 4:88, 17:711–713, 21:616, 22:276–277, 24:33, 24:35–36, 25:663
  and body 2:720–721
  and cultural geography 5:409, 5:411–412
  and culture, cognition *see* Cognition
  and gender 8:900–902
  and mirror neurons 15:586–587
  phenomenology 18:20
  spirit possession 23:282
  sports 23:299–304
Embodiment and culture 7:415–419
  methodological strategies 7:417–418
    content analysis 7:418
    ethnography 7:418
    surveys 7:417–418
  theoretical strategies 7:416–417
    cold and hot cognition 7:416
    discursive and practical consciousness 7:416
    evaluation 7:417
    Maussian holism 7:416–417
Embodiment cognition (EC) 7:420–426
  affordances 7:422
  body and emotion 7:423
  canonical neurons 7:421
  cognitivism 7:421, 7:423
  concepts and words 7:422–423
  critiques and debates 7:423
  definition 7:420
  distributional approaches 7:423–424
  enactivism 7:423
  extended mind 7:424
  future challenges 7:423–424
  grounded cognition 7:421
  language comprehension 7:421–423
  mirror neurons 7:421–422
  notion of simulation 7:422
  perception *versus* action 7:420–422
  radical embodied cognition 7:420, 7:422
  representations 7:420–424
Embodiment theory 2:25
  and cultural sociology 5:552–553
Embourgeoisement 3:740, 3:748
Embree, John 23:71–72
Embryo Protection Act 23:441–442
Embryo selection 2:619–620
Embryology 5:677, 5:679
Embryonic development, formation of brain during 2:821–822, 2:822f
Embryonic life, bioethics, biopolitics, and uses of 23:441–442
Embryonic society 6:334
Embryonic stem cells (ESCs) 2:678, 2:681, 23:440–441
Emergence 4:485, 22:491, 23:368–369
  social differentiation as 6:391–392
Emergence hypothesis 4:233
Emergence in complex systems, principle of 22:704
Emergencies
  protection of children in 3:425–426
  and psychosocial work 19:477, 19:479–481
Emergency democracy 20:856
Emergency management cycle, for natural disasters 16:317–318
Emergency Medical Treatment and Labor Act 10:597
Emergency medicine, evolutionary medicine 8:457
Emergency relief, and social work responses, Australia 7:427–432
Emergency routine, social work stress/trauma 18:709
Emergenist sociology 14:414, 14:416

Emergent behavior, autonomous agents 2:310
Emergent grammar 10:333
Emergent motivation, and flow 9:261–262
Emergent norm theory (ENT) 4:152–153
Emergent properties 7:433–437
  collectivist theory 7:434–435
  ideas of in late-twentieth-century theoretical thought 7:435
  methodological issues 7:435–436
    impact 7:436
    measurement 7:436
  micro-macro link 7:434–435
  philosophical tradition of 7:433–434
  social systems 7:434–435
  society as 7:434–435
  sociological conceptualizations of social reality 7:434
  sociological tradition of 7:434–435
  system properties 7:433–436
  systems theory 7:434–435
Emergent (research) questions in mixed methods research 15:620–621
Emergent realism 20:21
Emergent (autonomous) strategy 23:509, 23:512–513
Emergent subgroups 22:882
Emergentism
  language development 13:313
  in legal philosophy 18:29
Emergentist coalition model, for word learning 25:229
Emerging adulthood 20:298, 20:302, 24:576, 24:580
Emerging Church 3:542
Emerging cities, in urban world 12:117
Emerging economies 12:482
  corporate governance 4:936–937
  law firms in 13:580–581
Emerging gateways, in United States 11:631
Emerging market multinational enterprises (EMMNEs) 12:485, 12:487
  internationalization strategies 12:488t
Emerging markets, and financial crisis 9:183–184, 9:186–188
Emergy 7:630
Emerson, Richard
  theory of exchange relations 8:484
  work on power-dependence relations 8:483–484
Emetics, for alcohol dependence 3:766–767
EMFACS *see* Emotion FACS (EMFACS)
EMH *see* Efficient market hypothesis (EMH)
EMIC *see* Explanatory Model Interview Catalogue (EMIC)
Emic, construct clarification 7:930–931
Emic aspects 7:931
Emic perspective, in ethnography 8:185
Emic–etic approach
  and ethnography 8:180
  of personality 9:251–252
Emigrants, history 15:446, 15:448
Emigration 11:654, 12:246, 15:470
  and citizenship 11:650–652
  from Europe 15:457f
    determinants 15:458
  in Europe 15:376–388
  functional explanation 9:537
  and population momentum 18:576–577
  Shklar on 21:893–894
Emiliani, F., works, *Paradigmi delle rappresentazioni sociali* 13:902
The Emin (new religious movement) 16:806
Eminence 7:276
Emissions trading 18:369–370, 18:372
Emma Goldman Clinic for Women 15:112
EMMNEs *see* Emerging market multinational enterprises (EMMNEs)
EMOs *see* Environmental movement organizations (EMOs)
Emotion control strategies, for teacher stress 24:74
Emotion development 6:746
  related domains 6:747–748
  underpinnings of 6:747–748
Emotion expression 7:484–486
  cross-cultural findings of 6:749–750
  cultural differences 7:487
  development trajectories 6:747
  functions 7:484–485
  individual differences 7:485
  modes 7:485
  regulation 7:485–486
Emotion FACS (EMFACS) 7:522
Emotion knowledge 2:293
Emotion management 7:534, 7:538–539
Emotion perception 7:486–487
  cultural differences 7:487
  functions of 7:486
  individual differences 7:486
  regulation 7:486–487
Emotion potential 5:543–545
  universal components 5:544–545
    emotion link 5:544–545
    indigenous emotions 5:545
  universal emotions 5:543–544
Emotion processing 8:676
Emotion processing specificity 8:680
Emotion recognition 6:750, 7:446, 7:448–449
  of infants 22:557–558
Emotion regulation 6:746–747, 6:749–750, 7:446, 7:448–449, 7:452–456, 11:605–607, 17:811, 24:248, 25:590, 25:595
  antecedents 7:453–454
    ability 7:454
    beliefs 7:454
    knowledge 7:454
  and behavior during infancy and early childhood 6:14, 6:16–17
  consequences 7:454–455
    cognition 7:454
    decision-making 7:454
    emotion 7:454
    mental health 7:455
    physical health 7:455
    social interaction 7:454–455
    well-being 7:455
  coping 7:455
  coping with chronic illness 3:552
  development of 12:4–5
  expressive suppression 7:453–454
  and grief 2:538, 2:540–541
  healthy 7:455
  and neuroticism 16:745–746
  reappraisal 7:453–455
  reasons 7:452
  strategies 7:452–453
  suppression 7:453–455
Emotion Review 8:668
Emotion space 7:530–531
  as a coin with two sides 7:531
Emotion theory 13:466
Emotion understanding 12:3
  cross-cultural findings of 6:750
  development trajectories 6:747
  trajectories 6:747–748
Emotion work 7:534, 7:536
  Hochschild on 11:146–149
Emotional abilities 15:372–373
Emotional abuse, of elders 7:330
Emotional associative learning, role of cerebellum in 2:97–98
Emotional behavior, of animals 2:801, 2:804
Emotional brain 11:909
Emotional communication 2:792–793, 6:746, 23:295
Emotional competence 6:749–750
Emotional contagion, and empathy 7:549, 7:551
Emotional context, source monitoring for 23:30
Emotional development 12:1–5
  in adulthood and later life 21:424–431
  attachment 2:137
  of children 3:454
    effects of parenting and family structure on 7:438–444
  displays 12:4
  emotion understanding 12:3
  future directions 12:5
  nature of 12:1–2
  neurobiological growth 12:2–3
  perceiving 12:3
  psychobiological growth 12:2–3
  relational influences on 12:5
  responses 12:3
  self-awareness 12:3–4
  self-understanding 12:3–4
  social influences on 12:5
Emotional deviance 7:536
Emotional dialect 7:508–509
Emotional disclosure, psychoneuroimmunology 19:448
Emotional disorders
  cognitive behavioral therapy for 4:30–31
  and neuroticism 16:744–745
Emotional distress, and adolescent health 1:124
Emotional empathy 7:549, 19:216, 19:219
Emotional energy 4:215–216, 23:370
Emotional engagement 1:31

Emotional experience 7:483–486, 8:517–518
age differences in 7:490–491
Emotional expressions 7:457–460, 7:496–497, 7:499–500, 8:663–664, 8:667, 23:553
affect program theory of 7:457–459
contemporary alternatives to 7:459
age differences in 7:491
core affect 7:459
Darwin on 7:457–458, 7:458f
in discourse style 13:291–292
discrete emotions 7:459–460
display rule 7:458–459
ethology 7:458
facial 8:650–652, 8:654
fixed action pattern 7:458
origins of research 7:457
psychological constructionism 7:459–460
Emotional factors, and pain 17:454
Emotional frequency 5:545–546
Emotional habituation 17:478
Emotional helping 19:221–222
Emotional inhibition 7:496–499
Emotional intelligence (EI) 4:668, 5:433, 7:445–451, 7:485–486, 7:508, 8:677, 8:679, 13:649, 17:752–753, 17:851, 17:871
approaches to 7:508
bad 7:447
counterproductive work behavior 7:447
and coping 4:899–900
discussion 7:448–449
future directions 7:448–449
cultural variability 7:448–449
differential impact for men and women 7:449
measurement considerations 7:449
methodological considerations 7:449
training 7:449
good 7:446–447
interpersonal relationships 7:446
job performance 7:446–447
leadership emergence 7:446
history 7:445
interrelated abilities 7:445
in leadership 7:509
leadership education 13:656, 13:661–662
and Mayer-Salovey four-branch model 7:508
personnel selection 17:938
streams of research 7:445
behavioral outcomes 7:446
performance tests 7:445
self- and peer-reports 7:445–446
ugly 7:447–448
leadership 7:448
negotiation 7:447–448
Emotional intensity 5:545–546
Emotional issues, gifted education 7:162–163
Emotional labor 7:509, 7:536, 21:648, 21:650
deep acting 7:509
Hochschild on 11:146–149
and leadership 7:509
surface acting 7:509
of well-being and burnout 25:535
Emotional life 2:793
Emotional marker 3:833
Emotional memory 14:127–128, 22:68
Emotional processing 13:679–680
and anxiety disorders 11:607
influence on phantom limb 18:4
Emotional processing model (EPM) 8:605–606
Emotional processing theory (EPT) 3:768–769
Emotional reactions, virtue ethics 25:151
Emotional regulations
age differences in 7:491–492
by temporal lobe 24:208
Emotional rehabilitation 10:716t
Emotional responses 7:484–486
coronary heart disease 4:919
of mourners 5:872
Emotional selectivity theory 21:426
Emotional social support 1:131
Emotional stability 4:905–906, 17:787, 17:822, 17:914–915
and academic performance 17:789
and adolescent personality 17:864, 17:866–867
in learning and education 17:789
for motivation 17:789
Emotional Stroop Test 8:679
Emotional support, for therapist and psychiatrist 19:497
Emotional suppression tendencies, coping with chronic illness 3:551
Emotional tactics, social influence 22:352–353
Emotional transformation 4:153
Emotional trust 12:637
Emotional valence, and flashbulb memory 9:256
Emotional well-being, in cardiac disease coping 3:112
Emotional/affective learning processes 3:451–452
Emotionality 7:472–473
Emotionality, activity, and sociability (EAS) Temperament Survey 24:191–193
Emotionalizing managerial and organizational cognition 14:481
Emotionally focused couple therapy 8:747
Emotionally unstable PD 24:279
Emotion-based learning, impairment 6:721
Emotion-focused coping 4:903–904, 23:551
chronic illness 3:550
Emotion-focused psychotherapy 20:740–741
Emotion-focused therapy (EFT) 4:731
for depression 19:375–376
Emotion-oriented therapy, test anxiety 24:248
Emotion-related physiology, and aging 7:491–492
Emotion-related prosocial behaviors 19:247
Emotions 1:241, 5:542, 5:565–566, 6:434, 6:746, 7:378–379, 7:445, 7:467–471, 7:483, 11:145–149, 12:1–4, 13:461, 15:802, 15:896, 16:738–740, 17:858–859, 17:871, 19:351, 21:424, 21:813, 23:291–292, 23:644–645, 24:173–175, 24:177
and abusive supervision 23:703
adult development, theoretical approaches 7:492–493
adulthood 21:428–429
affect program theory of 7:468
and aggression 1:379
and aging 7:490–495
and agreeableness 1:465–466
as analytical object 1:236
of anxiety 17:860
appraisal 7:526–531
arousal, and stress 24:567
assessment 7:519–525
arousal 7:520–523
autonomic measures 7:521
avoidance 7:523
body 7:523
discrete emotion approach 7:519–520
dynamic systems approach of socio-emotional development 7:521
expression 7:521–523
face 7:522
motivational component 7:523
from nonverbal behavior 7:521–523
physiology 7:521
subjective feeling 7:519–521
through motor actions 7:523–524
through questionnaires and self-report measures 7:521
valence 7:520–523
voice 7:522–523
attraction and displacement, in conflict 4:622–623
attributional approach 15:894–898
basic emotion theory 7:528–531
behavioral expressions, age differences in 7:491
and body 2:728–729, 2:731–733
and cardiac defense 23:540, 23:542–544, 23:543f
characterizing 7:526–528
affective phenomena 7:527
multicomponents 7:526–527
shaped by evolution and by sociocultural contexts 7:527–528
short term 7:526
triggering by events 7:526
in cognition 7:461–466
across lifespan 7:465
appraisal models 7:461–462
attention 7:461–465
basic emotion models 7:461
clinical implications 7:464–465
from conceptual models to research findings 7:462–464
decision-making 7:461–463
memory 7:461–462, 7:464–465

psychological construction model 7:462
in cognitive process 13:461
cognitive-anatomical models of 16:613–614
cognitive-developmental theory 21:427–430
aging 21:430
crystallization 21:430
emotional transformation adulthood and later life 21:428–429
emotional transformation from childhood to adolescence 21:427–428
vulnerability of later life regulation capacities 21:429–430
in collective behavior 4:157–159, 4:161
and collective memory 4:196
collective political action 4:148–149
in collectivism/individualism 4:208–209
commodification of 11:146
consequences of dimensional comparisons for 6:430–436
consumer psychology 4:740–741
contagion 7:487
and culture 5:542–549
cultures 13:465
Darwinian legacy 7:467–468
and decision making 5:921–922, 7:479
definition 5:542–543, 7:483
definitional issues 13:462
developmental neuroscience 16:601–602
specialized areas 16:603
developmental research on 12:2
differential emotions theory 21:424–425
emotions as core adaptive survival responses 21:424–425
stability and change of primary emotions 21:425
disgust 7:468–469
disgust as 6:546
of disputant 16:428
and dreaming 6:651–655
of drivers in transportation 24:625–626
educating
and channeling 13:464–465
strategies for 13:466
effects of media on children 15:37
coping responses 15:37
fear responses 15:37
and embodied cognition 7:423
emerging trends 5:547–548
as momentary constructions 5:547
as relational constructs 5:548
and emotional displays 12:4
emotional experience, age differences in 7:490–491
empathy *see* Empathy
and empathy 19:216–219
and ethnography 1:236–239
ethological tradition 7:467–468
events that give rise to 7:477
evolution 7:467–470
evolution of moral emotions 7:469–470
evolution of morality 7:469–471
evolutionary psychology and 7:469
experiential 8:517–519
corrective 10:431
experimental economics 8:550
expression 7:496–497, 7:499–500
expression of 7:484–486
facial 8:667–668
of fear 17:860
feeling 7:526–528, 7:530–531
feeling and display of 13:463
and film genres 9:162–163
functional neuroimaging of 16:609–613
functions of 7:477–479
future directions 7:493, 7:531–532
in group and institutional settings 13:465–466
group-level 13:466
and health 17:793
alexithymia 17:793
anger and hostility 17:793
constitutional vulnerability 7:499–500
and disease development 7:497–498
health behavior connection 7:498
as health risk 7:497–498
implications and future directions 7:500
mechanisms linking 7:498–500
physical health 7:496–500
physiological connection 7:498–499
as protector of health 7:498
reducing health risk 7:500
responses to illness 7:496–497
Type A behavior 17:793
Type D personality 17:793
hedonics of 21:425–426
attachment theory 21:426
emotional selectivity 21:425–426
optimization 21:425–426
Heider on 10:753
history of 7:472–476, 7:513–518
anthropology 7:513–514, 7:516–517
appraisal theory 7:474
body 7:473–475
comparative history 7:513, 7:517
culture 7:513–517
current issues and future directions 7:516–517
decision making 7:473
emotional energy 7:475–476
explicit field 7:513–514
gains and gaps 7:514
interaction 7:473, 7:475–476
limbic system 7:474
love 7:513–516
major themes 7:513
management 7:475
modernization 7:514–517
neurosciences 7:473–476
new understanding 7:513
private life 7:516
psychology 7:513–515, 7:517
public life 7:515
rationality and emotionality 7:472–473
social constructivism 7:475
sociology 7:475–476, 7:513–514, 7:516–517
turning points 7:515–516
Western European history 7:515
of hope 17:860
hope as 11:197–199
and HPA axis 11:477–480
and hypertension 11:453–454
identifying and interpreting 13:463–464
and identity movements 11:543–544
influences of emotion regulation on 7:454
informational role of 13:463
inhibition 7:496–499
and institutional structure 13:466
in intergroup relations 7:502–506
anger 7:505–506
conflict 7:502, 7:504–506
empathy 7:503–504
feeling 7:502–505
identity 7:503–505
motivation 7:502–505
pride and other positive emotions 7:504
reconciliation 7:504–505
shame and guilt 7:504–505
and interoception 12:616–617
current point of view 12:617
early concepts 12:616
leadership education 13:655, 13:659–662
and legislation 13:466
life span 21:424–425
managing 13:464
and meaning 21:563, 21:567
and mobilization of rights 13:466
modularity 7:469
moral 15:788–792, 15:794–795
motivational state 7:468–469
and nationalism 16:312–313
negative 7:496–500
of negotiator 16:426–427
neural basis of 7:477–482
amygdala 7:479
brain mechanisms of 7:479–481
cingulate cortex 7:477
explicit system 7:481
implicit system 7:481
orbitofrontal cortex 7:479–481
and neuroesthetics 16:658–660, 16:659f–660f
and neuromarketing 16:698, 16:700–701
and neuroticism 16:745
nonlinear dynamic systems 7:528–531
nonverbal expression of 13:463
and opinion formation 17:243
perception and expression of 7:483–489
perception and rehabilitation 2:816
perception of 7:486–487
and persuasion 13:466
phenomenology 18:19–20
physiological responses of 7:461–463
physiology, age differences in 7:491
positive 7:498–500
potential of 5:543–545
and prefrontal cortex injury 18:815–816
pride 7:469–470
process 7:483–484
psychological constructionism 7:528–531
psychological structure of 7:526–533
complementary bodies of knowledge 7:531

Emotions (*continued*)
emotion space 7:530–531
emotion space as a coin with two sides 7:531
emotional episodes distinguished from nonemotional episodes 7:528–530
influenced by evolution and sociocultural contexts 7:530
roles of and interrelations between emotion components 7:528
in reasoning process 13:463
reciprocal altruism 7:468
regulation 7:496
regulation, age differences in 7:491–492
related to disgust 6:548
role in doctrine 13:462
role in neonatal imitation 16:484
role in voting 25:306
shame 7:469–470, 21:871–876
and sleep-dependent memory consolidation 22:68–69
and social behavior 8:464
social exchange 8:486
and social norms 13:465
sociobiology and 7:468–469
sociology, Randall Collins 4:214–216
sociology of 7:534–540
affect control theory 7:537
cultural shaping 7:535
emotion management 7:534, 7:538–539
emotion work 7:534, 7:536
emotional deviance and emotions of deviance 7:536–537
emotional labor 7:536
equity, justice, and exchange 7:537–538
identity, emotion, and social structure 7:537
origins (1975–1985) 7:534–535
role taking 7:535
social change 7:538
social movements 7:534, 7:538
structural shaping 7:535–536
and time 7:538–539
solidarity 23:8
and startle response 23:331–332, 23:334–336
studies in four fields 1:236–237
and symptom perception 23:869
taste and smell 24:26–27
universality and cultural specificity of 7:459
valence 7:527–528, 7:530–531
for various legal actors 13:462
and work 7:507–512
in workplace 7:507
between-person (level 2) 7:508
interpersonal (level 3) 7:508–509
leadership (level 4) 7:509
multilevel model of 7:507
organizational level (level 5) 7:509–510
teams (level 4) 7:509
within-person (level 1) 7:507–508
Emotive functions of language 12:783
Emotivism, theory of 6:593
Empathic concern, and prosocial behavior 19:234
Empathic coping 11:593
Empathic understanding 16:585
Empathy 2:756, 7:541–548, 8:518, 15:790, 17:776–777, 19:216–220
as adaptive response 7:545
in adulthood 19:216
and aesthetic education 1:230–231
and agreeableness 1:466, 1:468–469
and altruism 19:218
aspects specific to humans 7:544–545
consequences 7:545–546
contact hypothesis 19:219
contribution of multisensory processing to experience of 7:544b
development during early childhood 7:549
contagious crying 7:551
cross-cultural aspects 7:552–553
defining 7:549–550
emotional contagion 7:549, 7:551
in first years of life 7:551–552
infancy 7:551–553
mechanisms of affect sharing 7:550
mental states 7:549–550
negative empathy 7:551
positive empathy 7:551–552
prosocial behavior 7:549–550, 7:552
self-awareness 7:551
sympathy 7:552
development of 20:435–436
emotional 19:216–219
erousal 19:216–218
evolution of 7:541–542
impairment 19:452
intergroup bias 19:218–219
and intergroup relations 19:218–219
power relations 19:218–219
in intergroup relations 7:503–504
in laws 13:463–464
and mirror neurons 15:585
and moral development 15:788–791, 15:794
and morality 15:814
motivation to care for others 7:542
neural circuits associated with perception of others' pain and distress 7:543–544
and parenting 19:244–245, 19:247–248
perspective taking 19:216–217, 19:219
prosocial behavior 19:233–236
and prosocial behavior 19:229
and prosocial motivation 19:217–218
and egoism 19:217–218
and prosocial responding 7:317–318
-related behaviors, neuroevolution of 7:542–543
research trends 19:216–217
Rogers on 20:737–738, 20:740–741
and self-conscious emotions 21:476–477, 21:479
volunteering 19:216–217
Empathy Learning Systems 1:270
Empathy–altruism hypothesis 19:218
Emperor system, in Japan 14:637, 14:639
EMPG *see* European Mathematical Psychology Group (EMPG)
Empire building
Pax Britannica 25:364–367
Empires 4:223, 4:753, 6:810–812, 11:685, 16:841–842, 23:63–64
Andes on 23:36
and anthropology 1:751–756
balance of power 2:345–347
of Chinggis 16:842
colonial, and Third World 24:301
in Eastern Europe 6:806
Eurasian 16:842
and imperialism 11:682–684
Islamic 15:354
Mann on 14:491–495
Meiji Revolution *see* Meiji Revolution
Ottoman on 15:354
premodern 23:364–365
small states 22:108–112
Southeast Asia 11:78–79
Empirical appearances 10:737–738
Empirical Bayes estimation 16:51
for social analysis 14:579–580
Empirical curve fitting 13:691
Empirical data
diffusion models and 6:399–400
error response times 6:399
quantile probability plot of RT distributions 6:399–400
scaling and distribution shape 6:399
speed *versus* accuracy 6:399
Empirical developments, teacher motivation 24:64
Empirical distribution function (EDF) tests 23:397–398
normality 23:398
Poisson assumption 23:398
Empirical ends 25:49
Empirical ethics 22:786
Empirical implications and connections 19:538
Empirical Implications of Theoretical Models (EITMs) 18:453–458
analogs 18:457–458
error-in-variables regression 18:458
expectations and uncertainty 18:457–458
relating behavioral analogs to empirical analog for measurement error 18:458
tools and components for behavioral analog 18:457–458
applied statistical procedures that fail to identify mechanisms 18:454–455
background 18:453
from concepts to analogs 18:455
data mining 18:454–455
developing formal and applied statistical analogs 18:455–456
economic voting 18:456–457
behavioral and applied statistics concepts 18:456
devising formal and applied statistical analogs 18:456–457
linking formal and applied statistical analogs 18:457
formal models and facts 18:454–455
garbage cans 18:454

identification of focal concepts 18:455–456
mechanism operationalization 18:455
noncumulative applied statistical practices 18:455f
omega matrices 18:454–455
prediction 18:454–455, 18:457
scientific accumulation 18:457
unification and evaluation of analogs 18:456
validation 18:454–455
Empirical legal pluralism 13:757
Empirical legal studies (ELS) 7:554–557, 13:457, 13:515
current status 7:555–556
economics 7:554–556
future of 7:556
law 7:554–556
methodology 7:554–556
origins 7:554–555
political science 7:554–556
psychology 7:554–556
sociology 7:554–556
statistics 7:554–555
Empirical methods 20:172–173
intensive and extensive design 11:818
Empirical program of relativism (EPOR) 21:202, 21:309, 24:136
Empirical psychology, Broadbent 2:857
Empirical research 14:112, 14:114–115, 22:992
adversarial expert testimony 8:586
cross-cultural training 5:337–338
expert testimony 8:586
infrastructure 22:667–670
nonadversarial expert testimony 8:586
normative criteria in 12:914
Empirical research topics, in cross-cultural psychology
between-culture variance 5:313–314
emotions 5:314
relationships between language and thinking 5:314
spatial orientation 5:314
Whorf's hypothesis 5:314
within-culture variance 5:313–314
Empirical science
concept formation in 10:774–775
methodological unity of 10:774
Empirical science studies 21:325–326
Empirical social research (ESR) 7:558–566
analysis of social data (Europe) 7:558–560
administrative use 7:559
German historical school in economics 7:560
statistical use 7:559–560
collection of social data (Europe) 7:558–560
community studies in United States 7:561–562
definition 7:558
development of 7:562–563
Lazarsfeld on 13:635–636, 13:638–640
questionnaire 7:563
in 1930–1940s 7:564
and sociological theory 7:564
stages 7:558
street corner society 7:562
survey analysis, principles of 7:563
trajectories 7:558
in United States 7:560–561
Empirical social sciences 13:587–588
Empirical social system 22:744–745
Empirical social workers 14:668
Empirical studies
of auctions 2:206
of scientific practice 8:577
Empirical tests
general strain theory 23:497
hierarchies and markets 10:843
of identification 23:616
Empirical validity, of psychoanalysis 22:765
Empirical work 13:456–457
behavioral law and economics 13:458–459
for multidisciplinary discourse on law 13:457–459
sociological influences in economic analysis of law 13:458
Empirically adequate model 14:797
Empirically supported intervention (ESI) guidelines 5:520
Empirically supported therapies (ESTs) 8:344, 19:382–385
controversy 19:384
identification 19:383–384
managed care 19:382
practice guidelines 19:382–383
Empiricism 7:567–570, 11:133–134, 18:100–101, 18:251, 18:253–254, 22:986, 25:292
basic thesis of 14:316
concept 7:568–569
and concept of introspection 12:702
constructive 7:569–570
Feyerabend on 9:117–118, 9:120–122
Hilberg on 10:871–872
history 7:567–568
knowledge 7:569
logical 14:313–318
and modern human cognition 11:252
and personality psychology 17:910
positivism 18:625–627
pragmatism 18:803–806
and psychology 19:410–411
and science communication 21:219
scientific knowledge and 7:569–570
twentieth-century 7:568
Empiricist antirealism 20:21
Empiricist approach to categories of thought 13:70–71
Empiricist epistemology 8:577
Empiriocriticism 14:313–314
Emplotment 16:205
Employability 7:130
Employee advancement, notion of 19:135
Employee compensation 17:611
Employee entrepreneurs 7:684
Employee green behaviors
determinants of 19:84–86
improving 19:86–87
Employee pay
pay-for-performance 17:612–613
Employee Protection Act 1994 21:725
Employee Retirement Income Security Act 10:596
Employee–customer interactions 17:342
Employees 7:573
corporate governance 4:933–934
workplace deviance 25:722–726
Employer attitudes
over older workers 17:186–187
negative attitudes 17:187
Employer discrimination
and beauty 2:384–385
and labor market outcomes 17:74
Employer provision, family policy 8:803
Employers' liability insurance funds 17:142
Employment 4:324–325, 11:120–121, 13:199–200, 25:662, 25:676
of adolescents 1:118
affirmative action programs in 1:262–267
advent, growth, and persistence 1:262–264
partial retrenchment 1:265–266
atypical 7:572–573
of Australian aboriginals 2:238
changing nature of 7:572–575
conditions 10:517
and desistance/persistence 5:142
and development 6:269–275
diversified arrangements 25:700–701
and education 7:127–131
and equity in education 7:918–919, 7:921
and ethnic identity 8:106, 8:109–112
green jobs 10:372
job design and evaluation 12:816–818
and knowledge 13:106
labor statistics 5:798–802
and macroeconomics 14:411
migrants 15:417
and obesity 17:70–74
promotions 10:518
public and private sector 10:519
and reentry after incarceration 20:62–65
regulation 7:571
direct employment 7:571
of standard contract of employment 7:572
regulation of 7:571–576
Roma people 20:750–751, 20:753
sectors 22:684
self-employment 10:518
skills training 22:21
sociology of, Myrdal's 16:162
state of the art 12:152–159
aggregate empirical evidence 12:155
classical economists 12:153
compensation theory 12:152–157
empirical evidence 12:154–157
high-tech 12:156–158
microeconometrics 12:154–155, 12:157–158
microeconomic empirical evidence 12:156–157
process innovation 12:152–157
product innovation 12:152–158

Employment (*continued*)
R&D 12:154–157
sectoral empirical evidence 12:156
technological unemployment 12:152–153
theory 12:152–154
status of social scientists 22:684
subsidies, and labor market policy 13:183–185
systems of 25:699
theory of 20:729
and unemployment rates 5:772–773
welfare reforms 25:517–518
white-collar workers 25:564–568
work anthropology 25:662, 25:664, 25:666
working conditions
policy 25:698–701
relationship 25:698, 25:700–701
Employment contracts 1:27–28
fixed-term 1:27–28
permanent 1:27
Employment Cost Index (ECI) 5:743
Employment deprivation 22:286
Employment discrimination 1:256, 1:259, 13:600, 18:830–831
black and Native-American employment 1:258
black–white earnings differentials 1:256–257
federal antidiscrimination law in United States for 6:531–532
quotas 1:259
and sexual harassment 21:721, 21:723
Employment equity 1:248
Employment Equity Act (Canada) 4:741, 6:598
Employment laws 3:29
Employment liability insurance 12:216–217
Employment loss, in shrinking cities 21:920
Employment Non-Discrimination Act (ENDA) 6:531, 9:635–636, 24:551
Employment opportunities, of women 9:64–66
Employment paradigm 17:143
Employment practices, impact of social media on 24:106
Employment protection legislation (EPL) 13:183–184
Employment rates 25:831–832, 25:832t
of age group 50–64 years 17:183–185, 17:184f–185f
Employment relationship, terms of 25:698
Employment Retirement Investment Security Act 1974 25:60
Employment sectors, educational research 7:282
Employment stability 6:273
Employment studies, and green jobs 10:376–377
Employment uncertainty, fertility 9:21
Employment White Paper of 1944 13:15
Employment–population ratios 13:190
Employment-to-population ratio 13:174, 13:176f
Empowering Dimension 13:422
Empowerment 3:522, 4:154, 5:288–289, 7:577–578, 8:34, 20:673, 20:677
and adult education 1:151
choice as 3:521–522
community 4:316
and community work 4:381–382
and conditional cash transfers 4:555
evaluation 7:577–583
and feminist psychotherapy 8:929–933
nonprofit organizations 16:916–917
outcomes 7:577
of patients 10:688–690
processes 7:577
relational 20:222–223
versus sick role 22:976–977
of women, and tourism 24:477–478
Empowerment evaluation 19:121
background 7:577
clarifying issues in 7:581
concepts 7:579
definition 7:577
principles 7:579
10-step 7:580–581
theories 7:577–578
empowerment theory 7:577
process use 7:578
role 7:578
self-determination 7:578
theories of action and use 7:578
three-step approach 7:579–580
future planning 7:580
mission 7:579–580
taking stock 7:580
Empowerment social work 11:644–646
Empowerment theory 7:577
Empowerment Zone and Enterprise Community program 16:450, 16:453
Empty category principle 16:5
Empty counterfactuals 3:253–254
Empty lands, Turner's concept 9:465
Empty subject 4:828
Empty-nest life 12:402
Emson, Daksha 5:471
EMT *see* Ecological modernization theory (EMT)
EMU *see* Economic and Monetary Union (EMU)
Emulators, agent-based models 1:371
Enablement, social conditions of 9:1018–1019
Enabling conditions 20:461, 20:464
Enabling contexts 11:700–701
Enabling environments 6:459
Enabling state 24:923
Enabling structures 13:163–165, 13:167
Enactivism 7:423
Enactment, negotiated orders 17:412
Encephalization 2:828–830
Encephalization Coefficient (EQ) 2:829
Encephalopathy, and alcohol 1:511
Enclave, and ethnicity 8:159
ENCODE *see* Encyclopaedia of DNA Elements (ENCODE) research
Encoded Archival Description (EAD) 5:702
Encoding 13:680–681
of episodic memory 7:856–858
context effect 7:857–858
neural correlates 7:858
linguistic presupposition 14:190–191
and memory development 15:148
quantitative memory models 15:162–163
Encoding specificity principle
of recognition memory 20:41
of retrieval 20:610–611
Encomienda 23:39
Encore adulthood 20:594
future research 20:595
in twenty-first century 20:594–595
Encounter groups 13:913, 20:738–739
Encounters of culturally diverse communities
in North America 17:17–19
Encouragement designs 9:130
Enculturation 5:578, 7:144, 13:544
learning as 1:871
Encultured brain 21:619
Enculturing feminism 8:905–909
Encyclical on Christian Marriage 8:219–220
Encyclopædia Britannica 7:586
Encyclopaedia of DNA Elements (ENCODE) research 9:980
Encyclopedia of Political Economy 7:590–591
Encyclopedia of the Social Sciences 6:647
The Encyclopedia of the Social Sciences 7:585, 7:587
Encyclopedia of Unified Science 25:89–92
Encyclopedias 7:584–593
of basic concepts in history 4:551–552
growing availability of 7:584–585
history of 7:584
systemization of 7:585
Encyclopediasts 22:697
Encyclopédie 7:656–657
Encyclopedism 25:88
End game strategies 14:525
End of life 2:615
controversies over 5:872
End of time 24:408
End organs, vestibular system 25:63–64
ENDA *see* Employment Non-Discrimination Act (ENDA)
Endangered languages 14:225–226
ENDAUM *see* Eastern Navajo Dine Against Uranium Mining (ENDAUM)
Endless frontier 8:574
Endocannabinoids, and placebo effect 18:154
Endocannibalism 3:85, 5:870
Endocasts
brain, evolution of 2:827, 2:830–831
Endocentric structures 23:888
Endocrine, side effects (psychopharmacotherapy) 19:464–465
Endocrine disorders 7:610
Endocrine functions, religion and 20:267
Endocrine system, hypothalamus 11:482–490

Endocrinology 7:606
and psychiatry 7:606–611
End-of-ideology debate, Daniel Bell 2:499–502
End-of-life care, personalized medicine 17:926–927
End-of-life choices 7:594–605
advance directives 7:597–598
alternative choices 7:601–603
culture in 7:594–596
denying death 7:595–596
future directions 7:601–603
medical behaviors that potentially shorten life (MBPSL) 7:599–601
arguments around 7:601
assisted dying 7:600
euthanasia 7:600
life-terminating acts 7:601
palliative sedation 7:601
terminal sedation 7:601
withdrawing support 7:599–600
withholding support 7:599–600
medicalizing, hospital death 7:596–601
nature in 7:594–596
refusal of treatment 7:599
Endogamy 8:118–122, 11:730, 15:570, 18:842–843, 8:118t
measuring 2:100–102
minorities 15:570, 15:573
rationalizing 2:105–106
in South Asia 2:78–79
Endogeneity 7:612–614
causal chains 7:612–613
endogenous *versus* exogenous variables 7:612
interdependent systems 7:613–614
of law 13:600
mutual dependencies 7:613–614
statistical *versus* mathematical analysis 7:613
Endogeneity problem, in cross-national research 19:723
Endogenization 3:314–315
Endogenous attention 4:675–676
Endogenous colonies 18:134
Endogenous deaths 12:17
Endogenous dynamism 3:92–93
Endogenous economic growth 7:40–41
Endogenous effects of law 13:552
Endogenous feedback, diffusion 6:414
Endogenous growth 20:149–150
sustainable development 23:825–826
implications of 23:826
Endogenous hybridization 11:451–452
Endogenous institutional change 22:982
Endogenous instruments 12:207–208
Endogenous languages 1:107
Endogenous noise generation, stochastic optimization for 23:474
Endogenous opioids, and placebo effect 18:154, 18:156
Endogenous sampling *see* Outcome-based sampling
Endogenous specialization 6:602, 6:604
Endogenous sunk costs 14:552
Endogenous technological change 7:48
Endogenous variable 21:980
Endometriosis 10:464
Endo-nostalgia, and tourism 5:416
Endophenotypes
obsessive-compulsive disorder 17:123–124
sleep disorders 22:82–83
Endophora 1:674
Endorois Welfare Council v. Kenya 11:800
Endorsement 2:316
Endostrategic mobilization 4:623
Endowment effects 17:782
experimental economics 8:548
Endowments, and twin studies 24:703–704, 24:706–709
End-stage of childhood cancer 3:461–462
End-stopped complex cells 25:160
E-neighbors study 4:336
Enemies, cultural meaning of 5:427
Enemy images 17:654–655
Enemy/attaching guilt 25:398
Enemyship 5:427
Energetic dogma 6:853
Energiewende (energy transition) 18:369
Energopolitics, and infrastructure 12:122
Energopower 7:623–624
Energy 16:508
anthropology of 7:621–625
and cultural evolution 7:621–622
energy development politics 7:622–623
in era of anthropogenic climate change 7:623–624
experts, culture of 7:622–623
fuel 7:621, 7:623–624
indigenous peoples 7:622–623
infrastructure 7:624
power 7:622–624
resource use 7:622–624
definition 7:626–627
and eating 11:420–422
emotional 7:475–476
and fire 9:197–200
geography of 7:633–640
and green jobs 10:372–375, 10:377
greening of business 10:393
landscapes 7:639
and power 7:627
regions 7:639
return on investment 7:630–631
school engagement 21:59–60
and society 7:615–620, 22:255–256
civilization 7:616
coproduction 7:615, 7:619–620
energy ladder model 7:618
exergy 7:615–616
fossil fuels 7:615, 7:618–620
industrial revolution 7:618–619
maximum power principle 7:615–616
renewable energy 7:618–620
social construction 7:618–619
social values 7:618
societal energetics 7:615–618
technology 7:615–619
transitions 7:620
systems
and Euclidian space 7:636–637
spatialities of 7:637–638, 7:638t
thermodynamics 7:626–627
transitions 7:633–634, 7:633t
units 7:626
Energy balances 7:628–629
Energy Conservation and Atmospheric Quality 7:774
Energy decade 7:623
Energy efficiency 7:772–776
measures of 7:630–631
Energy flow analysis 7:626–632
agrarian society 7:628–630
biomass 7:628–631
in biosphere 7:627–628
conventional energy statistics and balances 7:628–629
ecosystem 7:626–627
embodied energy and emergy 7:630
enthalpy 7:627
final energy 7:628–630
industrial society 7:629–631
land use 7:627
measures of energy efficiency and quality 7:630–631
decarbonization 7:631
energy return on investment 7:630–631
technoeconomic concepts of energy and exergy efficiency 7:630
net energy analysis 7:630–631
primary energy 7:626, 7:628–630
socioeconomic metabolism
energetic part of 7:629–630
energy in 7:628
sustainability 7:626, 7:630–631
Energy Information Administration 10:304
Energy mobilization at work 17:133–134
Energy quality, measures of 7:630–631
Energy return on energy investment (EROI/EROEI) 6:853, 6:859, 13:241, 15:554
Energy services 7:628–629
ENERGY STAR 7:774–775
Energy systems model 7:635f
Energy-saving behavior 7:700
Enforceable trust 15:412
Enforced circular migration 12:434
Enforcement
of Common Rule 4:256–257
and immigration 11:637–638
laws, political money 18:396
of morality 13:486–488
rules, and technology 24:117–118, 24:122–123
Enforcement of law 14:433–434
and decline of mafias 14:434
Enforcement powers 17:145
Enframing, cybernetics 5:649
Engaged anthropology 1:224, 1:226
human rights 11:365
Engaged Buddhism 2:892–893
Engaged citizenship 3:653
Engaged ethnography 8:182
Engaged followership (obedience) 17:64–65
Engaged observer, photographer as 18:84

Engagements 1:31, 1:831, 4:896, 12:380–381, 17:337, 21:52, 25:663
Bourdieu on 2:781–782
and burnout 2:931–932
career goals 4:817
children 3:493–494
of citizens, in direct voting 25:295–296
civic 20:594
in classroom 1:31
classroom context and 1:33
classroom management/structure 1:33
peer support 1:33
task characteristics 1:33
teacher support 1:33
in classrooms, and gender 9:683
context 3:440–441
coping, chronic illness 3:550
and crisis of trust 21:257
defined 1:31
family 20:595
family context and 1:32
in fatherhood 8:841
future directions 1:34–35
with learning activities 1:31
matter 1:32
measurement of 1:31–32
as modern form of appeasement 1:830–831
as multidimensional construct 1:33
in open education 17:228
person-centered approaches 1:34
with prosocial activities 1:31
public 20:594–595
reciprocal relations 1:34
with school 1:31
theoretical frameworks 1:32
motivational and developmental theories 1:32–33
psychological and motivational theory 1:32
school dropouts 1:32
trajectories of 1:34
Engaging
of patients in clinical interventions 15:910
and strategizing
engaging others 23:524–526
engaging the unique 23:523–524
Engaruka 6:774
Engel, David
educational experiences of children 13:471–472
works
Tort, Custom and Karma: Globalization and Legal Consciousness in Thailand 13:472
Engels, Friedrich 1:456, 3:740, 4:339, 7:650, 14:642, 16:223, 17:22, 22:697, 24:745
on division of labor 6:601
on feminism 1:636, 8:945–946
on historical materialism 6:242
works
Anti-Dühring 14:670
The Communist Manifesto 23:353
Condition of the Working Class in England 9:135–136, 14:644
Engel's biopsychosocial model
International Classification of Functioning, Disability and Health (ICF) and 9:545
Engel's law 4:751–752
Engerman, S.E.
works
The Reinterpretation of American Economic History 7:56
Time on the Cross: The Economics of American Negro Slavery 7:57
Engineering 7:642–643
expenditures in 9:550t
profession 7:643–644
Engineering resilience 20:561
Engineering studies 7:641–648
building 7:641–642
critical participation 7:642–647
critique 7:643, 7:646
cultural practices of 7:646
developmental privilege 7:646–647
dominant images 7:642, 7:646–647
education 7:641–642, 7:644–646
field 7:647
formation 7:641–647
gender 7:646–647
heterosexual privilege 7:646–647
nation, state, and country 7:646
normative holism 7:643–647
private industry 7:644–645
race 7:646–647
technical–social dualism 7:642–647
technology through design 7:645
work 7:641–645
Engineers 7:642–643
Engineer-scientists 21:318–320
England
Combination Acts 1799/1800 2:48–49
community care in 4:344–349
discovery in 6:518–520
affirmative discovery 6:519–520
costs 6:520
disclosures 6:519
electronic discovery 6:520
sanctions 6:520
scope and limitations 6:518–519
early foundations for archaeological survey in 21:654
enlightenment in 7:656–657
literacy in 14:238
mortality rate decline 15:869
probation in 19:45–47
English auction 2:205–208
English Civil War 7:649–653, 15:11–12
English common law 3:670, 19:190, 19:194–195
English Defence League (EDL) 18:414–415
English East India Company (EEIC) 23:59
English iambic pentameter 13:267
English intellectuals 12:274
English language 2:236–237, 14:209, 20:5–9, 22:994
empirical literature 13:747
as a lingua franca (ELF) 16:68
pidgins 18:134
scope preferences in first language learners of 13:314–315
surface-syntactic relations 6:187–189
as a third language 16:68
English law 4:248
English longbow 25:415
English Longitudinal Study on Aging (ELSA) 1:416
English Medium school in India 1:270
English natural theology 20:286–287
English number 17:43
English pluralism 18:232–233
English Poor Law of 1834 18:731
English postal system 12:491
English psychology, mathematical developments in 14:811–812
English Puritanism 18:101
English regular inflections 15:827–828
English regular plurals 18:77
English Revolution 5:74–75, 5:78, 7:649–655
bourgeois revolution 7:650–651, 7:654
Civil War 7:649–653
Cromwell on 7:649–650
gentry controversy 7:651
Glorious Revolution 7:650
King Charles I 7:649–650, 7:653–654
Long Parliament 7:649–650, 7:652
new British history 7:653
parliament 7:649–653
postrevisionism 7:653
Puritan Revolution 7:650–652
Puritanism 7:651–652
radicalism 7:651–654
revisionism 7:651–653
English Statute of Monopolies of 1624 12:263
English universities 22:593
English wh-imperative 13:290
English–Cantonese bilinguals 3:511
English-language skills 3:310
English-speaking countries 22:328
ENGOs *see* Environmental nongovernmental organizations (ENGOs)
Engram 20:608, 20:612
Enhance the QUAlity and Transparency Of health Research (EQUATOR) 8:329
Enhanced awareness of social agreements and social responsibility (level of consciousness) 4:670–671
Enhanced Thematic Mapper (ETM+) system 20:423
Enhancing Recovery in Coronary Heart Disease (ENRICHD) study 9:706
ENIL *see* European Network for Independent Living (ENIL)
Enlarged Jewish population 12:801
Enlarged thought–action repertoires 2:865
Enlightenment 2:27–31, 2:145–146, 4:312, 4:475–479, 4:684–685, 10:523–524, 11:547, 11:876, 12:363, 14:253–254, 18:660, 19:123–126, 20:769, 21:242–243, 22:836–837, 22:996

and academies relation 7:659
anticlericalism 7:664–665
applied social research 1:850, 1:854, 1:856
Bentham's theory 2:525
in Berlin 2:542–543, 2:546
circulation 7:659–660
civil religion 7:665
and classical archaeology 3:760–761
definition and meaning 7:656
democracy 7:664–665
democratic revolutions 7:665
discussions and themes (new) 7:660–662
Durkheim 7:664–665
in economy 7:659
emancipation 7:658–659
in England 7:656–657
epocal concept 7:660
in European 7:656–657
fetishism as countermodel of 9:106–107
fields 7:657–658
Foucault on 7:666–667
in France 7:656–657
Frankfurt School 9:382
freedom 7:664–665
French Revolution 9:412
in Germany 4:686, 7:656–657
Habermas on 7:667
habit 10:476
in human rights 7:660–662
humanist notion of subjective identity 11:548
Kant on 7:664, 7:666
liberty 7:664
and military history 15:506–507
Mill on 15:524–525
modernity 15:683
and modernity 7:656–658, 7:660, 7:662
modernization theories 15:700–701
monarchism 7:665
operative tools 7:657–658
in philosophy 7:656–658, 7:660–662
political 7:658–659
postmodernism 7:666–667
practices 7:659–660
rationality 7:666
reason 7:656–658, 7:664, 7:666–667
and reasoning 19:591
reform 7:658–659
religion 7:664–666
in religion 7:657–658
and revolution 7:660–662
romanticism 7:660
roots of critique in 5:386
routes 7:656–657
and science 7:656–661
and science communication 21:219–221, 21:223–224
and social constructionism 22:223
spaces 7:656–657
times 7:656–657
Tocqueville on 24:423
and tradition 24:504
utopia 7:658–659
Weber on 7:664–667
Enlightenment conceptions of history 11:128
Enlightenment Contested 7:660
Enlightenment era, in Oceania 17:158, 17:161
Enlightenment philosophy 4:339
Enlightenment project 18:619
Comte's 18:620
Enlightenment thinkers 22:697
Enloe, Cynthia 15:490–491
Ennett, Susan T., on cigarette smoking 16:558
Enquiry paradigm 13:776
ENRICHD *see* Enhancing Recovery in Coronary Heart Disease (ENRICHD) study
Enriched environments, 4:469–474
behavior 2:835–836
brain 2:836–839
age 2:838
applicability to primates 2:838–839
controls 2:836–837
hippocampus 2:837
sex 2:837–838
as therapy 2:838–839
visual cortex 2:837
dendrites 2:837–838
exercise 2:837
isolation 2:837, 2:837t
maze 2:835–836, 2:836f
social housing 2:835–837
Enrichment 14:755
gifted education 7:160
Enrichment Triad Model (ETM) 7:161
Enriquez, V.G.
theory of kapwa 11:792
works
From Colonial to Liberation Psychology 11:792
Enrollment, graduate 22:681, 22:681t
Enskillment, and experts 8:589, 8:591–592
Enslavement
historical process of 22:37–38
ENSO *see* Event-Related Number Space Orientation (ENSO)
ENT *see* Emergent norm theory (ENT)
Entailment 21:589–590
Entangled histories 11:44
Entäusserung 1:539–540
Enterprise liability, tort claims and 24:432
Enterprise provision, family policy 8:803
Enterprise restructuring, economic transformation 7:75–76, 7:78–79
Enterprise zones 24:866
see also Export processing factories (EPFs)
Enterprises, definition 17:308
Enterprising schools 10:250–251
Enterprising systems 10:250
Entertainment, 'agon, 'alea, 'only entertainment 7:669–674, 11:121
boredom 7:669, 7:671–672
bread and circuses 7:670
carnival 7:670
cities 7:673
commerce and buying fun 7:671
and community 7:669
courtly diversions 7:670–671
Disneyland 7:673
exhibition 7:672–673
festival 7:669–670
funscapes 7:672
globalization of 7:673
Hollywood 7:673
mapping 7:669–670
and mass communication 14:693–694
mass entertainment industry of nineteenth century 7:671–672
mimicry 7:670
play 7:669, 7:672
popular culture 7:670–671
postmodern art 7:672
recreation 7:670
sex entertainment 17:219
show business 7:669, 7:672–673
spectacle 7:670–672
variety 7:672
vertigo 7:670
world's fair 7:673
Entertainment industry, and cultural participation 5:449
Entertainment preferences, behavioral expressions 21:610
Entextualization 6:505
Entfremdung 1:539–540
Enthalpy, energy flow analysis 7:627
Enthusiasm, concept of 9:673–674
Entitativity, social categorization 22:189–190
Entities, social 22:500–505
Entitlement theory 7:926
Entitlements 6:230, 9:307, 20:15
Dahrendorf on 5:663–665
and deserts 6:230
in joint action 12:839–840
in joint commitment 12:842
Entity theories about intelligence 19:791–792
Entorhinal cortex, and spatial navigation 23:163, 23:167–168
Entorhinal layers II and III 10:888–889
Entorhinal/postrhinal cortices 16:625
Entrepreneurial farming 17:666–667
Entrepreneurial history 3:25
Entrepreneurial passion 17:571–572
Entrepreneurial system 22:746–747
Entrepreneurialism 24:847
urban *see* Urban entrepreneurialism
Entrepreneurialization of urban governance 24:847
Entrepreneurs 3:24–25, 13:72, 15:676–680, 21:141–143, 23:524
Entrepreneurship 6:666, 7:675–685, 12:170, 14:472–473, 14:475–476, 21:141–142
business models and 3:34–35
collaborative consumption 4:140, 4:143
competition 7:675–676, 7:679
concepts and definitions 7:675–676
creative destruction 7:676, 7:679
early scholarship 7:681–682

Entrepreneurship (*continued*)
economic perspectives 7:683–684
and ethics 3:19
failure and closure 7:679
field 7:682
firm creation 7:676–677
founders 7:682–684
geography of 7:686–691
agglomerations 7:687–688
born global firms 7:689–690
business support 7:687–688
clusters 7:687–690
creative destruction 7:686–687
embeddedness 7:688–689
ethnic entrepreneurship 7:688–689
Fordism and post-Fordism 7:687
French regulation theory 7:687
industrial linkage 7:688
networks 7:688–689
new firm formation 7:686–688, 7:690
new industrial districts 7:688
translocal/transnational relationships 7:688–689
global trends 7:684
importance of economy 7:678–679
industrial dynamics 7:683–684
innovation 7:675–679, 7:681–685
job creation 7:675, 7:679
management 7:684
market opportunity 7:682–684
personality 7:677–678
practice 7:684
process theories 7:682
productivity 7:675, 7:678–679
and profits 19:114
psychological and sociological approaches 7:683
risk 7:675–679
roles and networks 12:172–173
Schumpeter on 7:681–684
self-employment 7:676–677
social business 7:684
strategic adaptation 23:503–504
strategy 7:684
uncertainty 7:675–677
value creation 7:681, 7:684
venture capital 25:59–61
Entropy 12:107
information theory 12:107–108
Entropy index 21:412
Entscheidungsproblem, and algorithms 1:530
Entzauberung (disenchantment) 25:464
Enuma Elish 25:438
Enumeration 3:296, 3:299, 3:301, 8:875
field 3:298
mail 3:299
population censuses 3:293
and senses 21:617
traditional 3:296–297
Enumeration sample (E sample) 3:293
Enunciation (énonciation) 18:795
Environment 2:450–453, 4:755, 7:778, 7:809–819, 10:83, 14:674–676, 15:501, 20:146, 20:150, 25:373, 25:437
African 1:292–299
and aging *see* Ecology of aging
attitudes and behavior, measurement of 7:699–705
attitudes toward 7:809–813
from 1800 to 1940s 7:810
after 1940s 7:811–812
after 1970s 7:812–813
Beck on 2:397–399
and biosocial criminology 2:670–674
burden, and urbanization 24:963–964
and business 3:2
causes of psychopathy 19:452
in Central Asia 3:316–317, 3:319–320
in China 2:67, 2:69
and chronic diseases 15:88
and climate change 3:855
context dependent memory 4:778–785
and creativity training 5:115–116
and cultural production 5:469
and cybernetics 5:647
disability 6:472–474, 6:476
and disability 6:467–470
for education 7:820–823
effects on epigenetics 24:689–690
enriched 2:835–840
and evolution 7:810
and experts/expertise 21:206–207, 21:210
factors related to temperament 24:185
focus of psychoanalysis on 22:767
and gene *see* Gene–environment interactions
global
new, knowledge in 10:153
and greening of business 10:394
and health 6:878–883, 7:815–819
human ecology 11:286
hunting/gathering societies 11:430
of Iceman 17:460
influence on socialization and education 22:864
influences in adulthood 6:754–755
influences on diseases 24:698
limits to growth 10:203–204
linking to final functional protein levels 7:841
livability of 10:521–522
and mental illness
factors before and during birth 15:205
later causes 15:205–206
and multiple language development 16:61
in North America 17:19–20
open classroom learning 17:227–228
and pastoralism 17:579–580
and personality development 17:881
personality differences and development 17:884–890
population dynamics and 18:568
impact of population 18:568–569
influences, modeling of 18:569–570
problems, and political ecology 18:352
regulatory agencies 20:181–185
in relational social work 20:225–226
and remote sensing 10:64–66
resources, for psychosocial interventions 22:383–384
role in genetics 9:985–988
rural sociology 20:814–817
and sexual orientation 21:779–780, 21:783
and sleep 22:61
and social change 22:194
social sciences, application of GIS/remote sensing to 10:67–68
social system 22:742–744, 22:747
and social systems 9:541
and social systems, Luhmann 14:383–384
and society 23:201
Soviet studies 20:831–835
statistics on 5:775
European Environment Agency 5:777
framework for development 5:776
future of empirical data 5:779
history 5:775
Organisation for Economic Co-operation and Development 5:776
organization 5:775–776
Statistical Office of the European Union 5:776
sustainability 5:775, 5:777–779
UN Division for Sustainable Development 5:777
United Nations Environment Programme 5:777
United States Environmental Protection Agency 5:777
and systems 14:393–394
temperament–environment interactions 24:196
term 7:809
urban 24:801–802
urban growth models 24:807–808
war and disaster 15:496
Environment and History 17:672
Environment health 7:816
climate change and health 7:817
environmental justice 7:818–819
health within changing environments 7:818
political economy 7:817–818
urbanization 7:816
Environment of Evolutionary Adaptedness (EEA) 20:389, 22:895, 22:904
Environment social work, challenges in 10:718
Environmental activism, in Soviet Union 20:832–833
Environmental approaches, for physical activity 18:97
concepts and principles 18:94
Environmental archaeology 7:692–698
history 7:692
research themes 7:696–697
subdivisions 7:692–696
archaeobotany 7:694–695
geoarchaeology 7:693–694
geochronology 7:695–696
zooarchaeology 7:695–696
Environmental Archaeology 7:696
Environmental assessment 7:781

Environmental attitudes 7:801
Environmental awareness, Lorenz on 14:368
Environmental change 16:336–338
Environmental citizenship 23:829–830
Environmental cognition 7:706–712
  applying behavioral research 7:710–711
    ensuring citizen competence 7:711
    setting public policy 7:710–711
  conservation 7:706, 7:708–710
  decision theoretic measurement 7:710
  decision-making 7:709, 7:711
  eliciting beliefs and attitudes 7:709–711
  evaluating perceived risks and benefits 7:708–709
  heuristics 7:706–707
  holistic accounts 7:707–708
  judgment 7:709–711
  natural world 7:708
  probability 7:708, 7:710–711
  psychophysical measurement 7:709–710
  reductionist approaches 7:708
  regulation 7:710–711
  research strategies for describing decisions 7:706–707
  uncertainty 7:706–711
  values 7:706, 7:708–710
Environmental commitment 19:85
Environmental concern scale 7:703t
Environmental concern–behavior relation 7:703–704, 7:704f
Environmental concerns 7:702–703
  critical theory 5:299–300
Environmental conflict
  and ethnic cleansing 8:87–88
  and sustainable transportation 23:845
Environmental conservation and development 7:713–719
Environmental constraints
  in irrigation 12:725
  settlement and landscape 21:656
Environmental construal, and temperament 24:196
Environmental contingency, strategic adaptation 23:501–502
Environmental continuity 17:887
Environmental control of behavior, and prefrontal cortex injury 18:815
Environmental costs
  and tranporatation geography 24:612–613
Environmental crime 5:159–160
Environmental criminology 5:146, 20:774, 23:144
Environmental crisis, in Australia 2:243–244
Environmental degradation 7:713, 7:797–802, 7:809
  from Malthus to political ecology 7:713–714
  neo-Malthusianism 7:713–714
  and per capita income, bell-shaped relationship between 13:145–146, 13:145f
  political ecology 7:714
Environmental design 2:409
  and crime 20:775–776
  crime prevention through 20:776
Environmental determinants, of health 7:740
Environmental determinism 7:720, 7:797, 10:92, 16:334, 16:336, 16:338, 20:117–118, 20:122
  and academy 7:723–724
  biological metaphors 7:722–723
  evolutionary mechanisms in 7:722–723
  from Magnus to Montesquieu 7:721–722
  origins of 7:720–721
  Sauer on 20:918
Environmental determinists 7:810
Environmental development 17:887–888
Environmental disasters
  class actions 3:736
  and green parties 10:383
Environmental economic geographies 7:36
Environmental economics 6:852–853, 7:726–732
  amounts 7:728–729
  distributions 7:729–730
  growth and environment 7:730–731
  measurements 7:729
  policies 7:726–728
  problems 7:726
  quasi-experiments 7:729
  research design 7:729
  risk, perception and values 7:729
Environmental education (EE) 22:7
  approaches to 7:735–736
  definition 7:733–735
  future prospects 7:736–737
  interdisciplinarity 7:733–735
  levels of learning 7:736
  problems and contradictions 7:735
  social learning 7:735–736
  social–ecological systems 7:736–737
Environmental effects
  epigenetic inheritance 7:835–837
  of human activities 12:717
Environmental elicitation, and temperament 24:196
Environmental enrichment (EE) 7:851
Environmental explanations 8:594
Environmental externalities 24:617
Environmental extremists, in United States 16:249
Environmental factors
  of aggression 1:378–379
  and HPA axis activity 11:479
  in internal migration 12:438
  of mania 14:486
  middle adulthood development 15:370–372
    community 15:370–371
    family 15:370–371
    relationships 15:370–371
    work 15:370–371
  weight regulation 25:468–469
    appetitive behavior 25:470–471
Environmental feterminism 7:720–725
Environmental fluctuation 17:886
Environmental geography 25:559
  White's contributions to 25:559–562
Environmental harm, and transnational organized crime 24:598
Environmental hazards, urban life health 24:831
Environmental health 15:87–89
Environmental health and safety 7:740–746
  case studies 7:742–744
    Deepwater Horizon oil spill 7:743–744
    Fukushima nuclear power plant accident 7:744
    health impacts of climate change 7:742–743
  environmental determinants 7:740
  inequalities 7:741–742
  natural disasters 7:744
  social determinants 7:741
  sustainability and future challenges 7:745
  vulnerability and human health 7:741–742
  vulnerable populations 7:740–742
Environmental history 7:747–755
  colonial, themes and contexts in 7:747–750
  colonial environmental ideas in twentieth century 7:751–752
  early colonial scientists, environmentalism 7:750–751
  landscape imagery and colonial environmental history 7:752–754
Environmental History Review 17:672
Environmental impact assessment (EIA) 7:811, 22:873–874
Environmental impacts 1:484–486, 7:106, 22:254, 22:258–259, 22:257f
  of human activities 22:258–259
Environmental influences
  of anxiety 1:806
  developmental behavioral genetics and education 6:323
  of eating 11:420–421
    food variety 11:421
    taste intensity 11:421
    taste quality 11:420–421
Environmental injustice, green social work 10:387–388
Environmental justice 7:764–766, 7:799–800, 7:818–819, 22:386, 22:388–389
  gender justice in 6:848
  green social workers and 10:386–387
  risk 20:707–709
  sustainable development 23:820
  in United States 7:756–762
Environmental justice health symposium (United States) 7:757–758
Environmental justice movement (EJM) 7:764, 7:766
Environmental Kuznets curve (EKC) 6:857, 7:726–727, 13:144, 23:823, 23:825, 13:144f
  cointegration 13:147
  economic growth 13:144–147
  empirical analysis 13:146–147
  environmental thresholds 13:148
  implications of 23:825
  inverted-U curve 13:144, 13:146–147

Environmental Kuznets curve (EKC) (*continued*)
irreversibility of damage 13:148
literature 13:147–148
pollution Haven hypothesis 13:148
technological progress 13:145–147
time series 13:147
Environmental landmarks 11:389
Environmental laws 3:31
Environmental management 7:813
Environmental mastery 1:181
Environmental mismatch and genetic quirks
evolutionary medicine 8:452
Environmental modeling assumptions 23:413
Environmental modernization agenda 6:835
Environmental movement organizations (EMOs) 7:763–767
Environmental movements 11:781–782, 20:288
conservation 7:764–766
culture and forms of action 7:765
deep ecology 7:765
discourse analysis 7:763–766
environmental justice 7:764–766
foundation funding 7:766
green parties 7:765–766
impacts 7:767
institutionalization 7:765–767
preservation 7:763–767
professional social movement organizations 7:765
rise of 7:763–764
social movements 7:763–767
values and social bases of environmentalism 7:764–765
Environmental nongovernmental organizations (ENGOs) 7:364, 7:369, 20:573–574
Environmental perspective on psychological stress 23:584
Environmental planning 7:769–777, 20:129, 20:131–133
by business corporations 7:776
by cities 7:773–774
by governmental agencies 7:773–774
taxonomy of strategies for implementing 7:773
history of 7:770–772
intellectual context of 7:772
research issues and major categories of 7:772–776
research on the scientific basis of environmental planning 7:772–773
by universities and colleges 7:774–776
Environmental policy 1:482, 7:778
characteristics of 7:778–780
coping with complexity 7:779
coping with uncertainty 7:779
governing across political borders 7:780
governing through time and space 7:779–780
nonhuman entities, dealing with 7:780
policy areas 7:780
public bads 7:778–779
public goods 7:778–779
regulation, need for 7:779
definition 7:778
elements of 7:781–783
history 7:781
national policy styles 7:781–782
policy process 7:782
politics 7:782–783
green parties 10:380, 10:383
integration of 7:780
protection and regulation 7:778–783
Environmental preservation 23:201
Environmental press 17:876
Environmental problems
driving forces of 7:799
Environmental protection, welfare economics issues of 25:501–502
Environmental protection interventions 7:715–717
market regulation 7:716–717
reducing emissions from deforestation and forest degradation + 7:716–717
Environmental protection strategies 7:715–717
conservation 7:716
parks and protected areas 7:715–716
Environmental quality 13:145–146, 13:145f
and green jobs 10:372–375
Thorndike's contributions 24:306
Environmental racism 23:150
Environmental refugee 22:755
Environmental regulation 10:372, 10:374
Environmental risks 7:784–790
acceptability 7:784–785
analytic-deliberative process 7:786, 7:788–789
assessment 7:784, 7:788–789
compensatory principles and standards 7:785–786
knowledge 7:786
valuation of consequences 7:786
definition of 7:784
evaluation 7:784, 7:786–789
generic principles for human interventions 7:784–785
IRGC risk governance model 7:787, 7:789
moral systems 7:785, 7:788
public participation 7:788
risk governance approach 7:786–789
stakeholder involvement 7:787
Environmental Satellite (Envisat) 20:422–423
Environmental sciences 7:791–795
ecology 7:791–793
meteorology 7:793–794
Environmental selection, and temperament 24:196
Environmental services 6:854, 6:856–858
Environmental social work 6:873–875
and human rights 11:355
Environmental sociology 7:796–803
agnosticism *versus* pragmatism 7:798–799
American Sociological Association's Section on 7:796
capitalism 7:799–800
causes/driving forces of environmental problems 7:799
constructivism *versus* realism 7:798–799
constructivist excesses 7:798
early emphases and disciplinary context 7:796–797
ecological modernization theory 7:798–802
ecological problems 7:796–797, 7:802
environmental attitudes 7:801
environmental degradation 7:797–802
environmentalism 7:797
future projections 7:802
human ecology 7:799–800
impacts of environmental problems 7:799–800
improved data and increasing methodological sophistication 7:801–802
inequitable distribution and environmental justice 7:799–800
political economy 7:799–800
social construction of environmental problems 7:797–798
societal environmental relations 7:798
solving environmental problems 7:800–801
sustainability 7:799, 7:802
Environmental spillovers 12:717
Environmental stimuli 7:238–240
Environmental stress 7:804–808
case studies of catastrophic environmental stress 7:806–807
Exxon Valdez oil spill 7:806–807
hybrid forms 7:807
living in contaminated areas 7:806
living with nuclear power plants 7:806
catastrophe 7:805–806
definitions 7:805–806
capital-based approach 7:806
characteristics of stressors 7:805
variables intervening between stressors and their effects 7:806
disasters 7:804–805, 7:807, 7:805t
psychological stress 7:805
and risk 7:804–805
Environmental stressors, of aggression 1:378
Environmental studies
Humboldt on 25:284–285
participatory action research in 17:535–539
Environmental sustainability 1:460, 10:533, 18:203, 23:802
linkages with development 23:802
objectives 24:920–922
organizational 19:83
sustainable production and consumption 23:839–844
and tourism 24:475–476, 24:479
and transportation 23:846–847
Environmental tobacco smoke (ETS) 22:128–129
Environmental transformations, historical geography 11:19
Environmental uncertainty 22:247

Environmental unpredictability
evolutionary developmental psychology 8:426
Environmental variables, and adolescent health 1:123
Environmentalism 7:750–751, 7:782, 7:797, 7:811, 10:71, 20:359, 21:179, 23:199, 24:820–821
ecofeminism in 6:848
economic psychology 5:889
and intergenerational justice 12:397–398, 12:400–401
of poor 1:461
and Slow Food movement 22:94–96
sustainable agriculture 23:807, 23:809
and use of official statistics 23:426
values and social bases of 7:764–765
Environmentalist movements 3:711
Environmentally extended input output analysis (EE-IOA) 14:762–763
Environmentally relevant behavior 7:699–700
Environmental-themed explanans 4:24
Environment–growth debate, sustainable development 23:825
Environment–organism transaction perspective on psychological stress 23:584–585
Enzymes, that inactivate monoamines, blockade of 1:779
EOL *see* Ease-of-learning (EOL) judgments
EOS *see* Earth Observing System (EOS)-1 (Terra); Earth Observing System (EOS)-2 (Aqua)
Eoseewong, Nidhi, works, *Pen & Sail: Literature and History in Early Bangkok* 11:86
Eotvos Lorand University (ELTE) 12:564–565
EPA *see* United States Environmental Protection Agency (EPA)
EPA (evaluation, potency, activity) structure of meaning 21:565–570
EPAA *see* Educational Policy Analysis Archives (EPAA) journal
EPAM-VOC model, for vocabulary acquisition 25:229–230
EPDM *see* Ethiopian People's Democratic Movement (EPDM)
EPDS *see* Edinburgh Postnatal Depression Scale (EPDS)
EPFs *see* Export processing factories (EPFs)
EPH *see* Equilibrium-point hypothesis (EPH)
Ephesians 3:238
EPIC *see* European Prospective Investigation into Cancer and Nutrition (EPIC) questionnaires
Epic of Gilgamesh 21:172–173
EPIC technique 11:401
Epicureanism 3:774–775, 19:407–408
Epicureans, religious view of 20:285
Epicurus 19:407–408
Epidemic hysteria 11:508
Epidemic models 16:443
Epidemics 15:857–858
and children/infant mortality 12:24
Epidemiologic paradox 11:654–655
Epidemiological capture–recapture studies 18:604–605
Epidemiological Catchment Area 15:180
Epidemiological Clinicopathological Studies in Europe (EClipSE) 7:141
Epidemiological epigenetics 22:283
Epidemiological transitions 15:857–862, 18:174–175
versus health transition 15:859
theoretical concept 15:857
Epidemiology 7:816, 7:824–831, 15:86–90, 16:404, 21:712
of AIDS 15:835, 15:841
analytic 15:179
association measures 7:829
attributable risk 7:830, 7:830b
behavioral medicine 2:464–465
of bipolar disorder 2:686
causal inference 7:826–827
Hill criteria 7:827t
definition 7:824
descriptive 15:179
of epilepsy 7:853
evidence-based psychiatry 8:340–342
experimental and quasi-experimental approaches 7:827
HIV infection
adolescents and youth 11:138
heterosexual men 11:142
MSM 11:141
women 11:140
of hysteria 11:509
incidence 7:825–826, 7:826f
lethality 7:826, 7:826b
life course *see* Life course epidemiology
measures of morbidity 7:824
measures of potential for prevention 7:830
of men's sexual dysfunction 21:713–716
arousal disorders 21:714–715
aversion disorders 21:716
ejaculatory disorders 21:715–716
low sexual desire 21:714
sexual pain 21:716
see also Ejaculatory disorders; Sexual arousal disorders
of mental health and mental disorders 15:189
of mental illness 15:179–186
mortality 7:825–826
observational approach 7:827–829
case-control study 7:829
cross-sectional study 7:828, 7:828b
longitudinal study 7:828–829, 7:828b
odds ratio 7:829–830, 7:829b
panic disorder 17:469
of panic disorder/agoraphobia 17:475
population attributable risk 7:830, 7:830b
prevalence 7:824–825, 7:825f
public health 19:548–549
relative risk 7:829, 7:829b
risk factors 7:826–827
schizophrenia 21:10–11
scope of 7:824
smoking-related diseases 22:125–126, 22:129
social 22:275–281
study designs 7:827
of suicide 23:673–675
of women's sexual dysfunction 21:717–719
arousal disorders 21:717–718
aversion disorders 21:718–719
low sexual desire 21:717
orgasmic disorders 21:718
sexual pain 21:718–719
Epigenesis, definition 17:879
Epigenetic imprints 1:763
Epigenetic inheritance 7:832–838
acquired variations 7:834, 7:837
chromatin-marking systems 7:833–834
chromosomes 7:833–834
definitions and historical background 7:832–833
DNA 7:832–837
environmental effects 7:835–837
evolution 7:832, 7:836–837
gene activity 7:833–835
genetic and epigenetic inheritance compared 7:835
genomic imprinting 7:834–835
genotype 7:832
heredity 7:832–833, 7:836–837
implications 7:836–837
inheritance and evolution, broader concept of 7:836–837
long-term epigenetic inheritance 7:835
mechanisms of 7:833–834
phenotype 7:832–833, 7:835–837
practical implications 7:836
psychological stresses 7:835
RNA 7:833–834
small-RNA systems 7:834
steady-state systems 7:833
structural inheritance systems 7:833
transgenerational epigenetic inheritance 7:834–835
variation among individuals 7:832–833, 7:837
Epigenetic inheritance systems (EISs) 7:833
Epigenetic mechanisms 2:453
Epigenetic modifications 9:960–961
Epigenetic programming, and brain development 2:821
Epigenetic rules 22:903–905
Epigenetic theories, of socialization and education 22:863
Epigenetics 1:763, 2:478–479, 6:752, 7:832, 7:848, 9:1003, 16:498, 18:820, 22:282, 23:559, 24:683–702
in behavior and psychology 7:842–844
and behavior 7:843–844
DNA methylation 7:844
and exercise 7:844
stress response 7:842–843, 7:843f
and brain aging 2:795
and brain plasticity 7:851
of brain plasticity and behavior 7:848–851
change over time and site 24:687–688

Epigenetics (*continued*)
aging effects 24:688
cell-specific and tissue-specific effects 24:687
locus-specific effects 24:687
sex and gender effects 24:687
chromatin structure and modifications 7:839–841
concordance 24:688–692
dizygotic twins 24:685, 24:690–691
DNA methylation 7:839, 7:850, 24:683–685, 24:687–690, 24:692–693, 24:698
genomic imprinting 7:850
DNA packaging 7:839–841
histone acetylation 7:840–841
histone methylation 7:841
histone monoubiquitination 7:841
nucleosome, and histone tail modifications 7:841f
posttranslational modifications of core histones 7:839–840, 7:840t
role of histones in transcriptional regulation 7:839–840
environmental effects 24:689–690
epigenomic maps 24:684t, 24:685f
future prospects 24:698–699
gene–environment interactions 7:844–845
adolescence 7:845
early life 7:844–845
in utero 7:844
genetic effects 24:689
versus genetics 7:842
DNA methylation and miRNA 7:842
epigenetic modifications can be inherited 7:842
of glucocorticoid receptor 16:498
heritability of epigenetic modifications 24:688–689
and HPA axis 11:479
influences on brain development 7:850
influences on brain plasticity 7:850
large-scale national and international epigenomic consortia 24:683t
limitations 24:698
mechanisms 24:683–685
microRNAs 7:841
miRNAs link environment to final functional protein levels 7:841
synthesis and actions 7:841f
nature/nurture 16:343–344
notion of 1:763
obesity 17:83
and personality 17:912
and postnatal brain development 7:850–851
programming 7:849–850
posttranslational modification to histone proteins 7:849–850
RNA regulation 7:849
regulation, principal mechanisms of 7:839–841, 7:840f
social *see* social epigenetics
and social change 22:148–149
stochastic effects 24:690
and twin studies 24:690–692
classical twin study 24:690–691
combining different omic data 24:692
environmental influences on diseases 24:698
epigenome-wide association scans for complex traits and diseases 24:693–698, 24:694t–697t
longitudinal twin study 24:692
multivariate twin study 24:692
power to detect epigenetic changes in disease and complex trait 24:692–693
trait/disease-discordant monozygotic twin study 24:691
twins 24:685–686, 24:686f
Epigenome 2:450
Epigenome-wide association studies (EWAS), for complex traits and diseases 24:693–698, 24:694t–697t
Epigenomics 2:625
Epilepsy 10:500–501
antiseizure drug 7:853–855
deep brain stimulation 7:853, 7:855
definition 7:852–853
diagnosis and treatment 7:853–854
electroencephalogram 7:853–854
electroencephalography for 7:359
epidemiology 7:853
etiology 7:852–853
glia 7:853–855
hippocampal sclerosis 7:854
hyperexcitability 7:854
hypersynchrony 7:852–853
inflammation 7:853–854
intracarotid sodium amytal testing 12:689–695
status epilepticus 7:853
sudden unexpected death in epilepsy 7:853
Todd's paresis 7:854
transient epileptic amnesia *see* Transient epileptic amnesia (TEA)
and transient global amnesia 24:566
Epileptic discharges 12:775
Epileptic seizures 15:967
and alcohol 1:511
Epileptogenesis, kindling as model for 13:34–35
Epinephrine *see* Adrenalin
Epi-Olmec period 11:57
Epiphenomenalism 3:265, 15:175
mental causation and 15:174–175
Episcopal Council of Latin America 14:33
Episiotomy 5:493
Episodic buffer, and memory binding 15:141–143
Episodic change
organizations 6:299
area for specialists only 6:299
clear-cut process 6:299
exception in organizational life 6:299
Episodic enumeration, for health behaviors 10:588–589
Episodic histories 11:108
Episodic memory 1:400, 2:282, 2:286, 5:925–926, 7:856–861, 19:3, 22:63, 23:876, 24:565–566
and aging 1:400–403
autonoetic awareness 7:856, 7:858
of children 3:466, 3:469
consolidation 7:858–859
contribution to future thinking 9:556–557
debated role 9:557–558
in early childhood 21:897–898
emergence of 15:151–152
encoding 7:856–858
context effect 7:857–858
neural correlates 7:858
familiarity 7:859–860
frontal lobes 7:858–860
functional imaging studies 21:573–574
and hippocampal amnesia 10:876–877, 10:881, 10:883–884
incidental episodic memory 21:897–898
incidental learning 7:857
in Infancy 21:897
level of processing theory 7:857
long-term, proactive interference in 19:3
medial temporal lobe 7:858–860, 7:860f
multiple trace theory 7:858–859
recall 7:856–860
recollection 7:856, 7:859–860
relationship of semantic memory with 21:572, 21:577
retrieval 7:859–860
neural correlates 7:859–860
retroactive interference in 20:613–614
semantic-episodic memory distinction 21:572
visual representation in 25:177–178
Episodic-autobiographical memory 1:651, 13:678–679
brain regions relevant to 13:679–682
Epistatic genetic variants 8:715
Epistementalities 13:208
Epistemes
indigenous 11:804–805, 11:807–809
centrality of spirit 11:804
ceremonies 11:805
community, egalitarianism, and respectful individualism 11:805
holistic perspective 11:804
importance of elders 11:805
recognition of cycles and flux 11:804–805
values 11:805
Epistemic authority 21:266
and narratives 16:221
Epistemic beliefs 7:887, 7:890
Epistemic communities 3:839–840, 7:867–872, 7:878, 24:296–297
collaborative research
interdisciplinary collaborations 7:870
international collaboration 7:869
in natural and social sciences 7:868–869
reasons for increasing collaboration 7:869–870
concept of 24:297
criticisms 7:863

expanding analytical frameworks 7:863
genealogy and definition 7:862–863
general theories of conceptual innovation 7:870–871
Haas on 7:862–863
historical studies of
defining disciplines 7:871
local and national focus and consumption 7:871
methodologies 7:871
in international relations 7:867–868
paradigms 7:862–863
policy 7:862–863, 7:865
qualities 3:839–840
social sciences and natural sciences compared
consensus 7:868
publications 7:868
textbooks 7:868
structure of
communication networks 7:867
research communities 7:867
Epistemic competence, and responsibility 20:589
Epistemic contemporariness, and wisdom 25:602
Epistemic cultures 7:873–880
career of 7:876–877
analyses of research fields 7:876
further developments 7:877
nonacademic contexts 7:876–877
definition 7:873
description 7:873–874
epistemological styles 7:877
illustrations of 7:874–875
knowledge society 7:876–877
learning 7:877
lifeworld 7:874
macroepistemic/knowledge cultures 7:878–879
modeling dilemma 7:876
in organizations 7:878
policies and 7:877
practices 7:873–874
preceding notions of knowledge, culture, and disunity of sciences 7:875–876
and questions of interdisciplinary 7:877–878
risk society 7:879
Epistemic doubt 7:881–886
and adolescence development 7:882–884
Epistemic interpretation of probability 19:24–25
classical 19:24
logical 19:25
objective 19:25–27
subjective 19:24–25
Epistemic naturalism 16:326
Epistemic pluralism, de-Westernization 25:549
Epistemic prototype 18:138
Epistemic rationality 17:354
Epistemic realism 7:881
Epistemic relativism 20:235
Epistemic resources 23:329
Epistemic self-regulation 12:723
Epistemic space, biotechnology 2:677–678
Epistemic structuralism 14:793
Epistemic subject 18:121
Epistemic theories, of public reasons 19:592–593
Epistemics, conversation analysis 4:852–853
Epistemological anarchism, Feyerabend on 9:118, 9:121–122
Epistemological beliefs 7:887–893
development 7:888
measurement of 7:888–889
modeling 7:887–888
normative implications of 7:889
personal epistemology 7:887–892
relation to achievement motivation 7:889–891
achievement values 7:890–891
competency beliefs 7:891
goal orientation 7:889–890
relation to cognitive learning outcomes 7:888
self-efficacy 7:891
sophisticated 7:889
specificity of 7:888
Epistemological Beliefs Questionnaire (EBQ) 7:888, 7:890
Epistemological crisis 14:202
Epistemological objectivity 17:98–100
critiques 17:98–100
Epistemological quality 19:688
Epistemological questions 21:190
Epistemological styles 7:877
Epistemology 1:894–895, 2:251, 7:241, 7:881, 11:133, 13:205, 13:207, 16:197, 22:1–2, 22:988, 24:292–293
Addams on 1:93
and biomedicine 2:663–664
and critical reflexivity 5:281–285
definition 19:592–593
feminist 9:717–718
of feminist legal theory 8:918–919
first models of development 7:881
and gender 9:788–791, 9:793
human rights 11:364
hunting/gathering societies 11:430
implications of realism for 20:24
and indigenous knowledge 11:780–781
of knowing other cultures 5:615–616
Kraepelin on 13:127–128, 13:131
LAC educational system 7:182
legal reasoning 13:776–779, 13:781–783
liberalism 14:11, 14:15
logical empiricism/logical positivism 14:313–318
Lorenz on 14:368
organizational learning 13:697
participatory action research (PAR) 17:540–541
peer review 17:686, 17:688
Piaget on 18:120–121
Plato on 18:211
pragmatism 18:804–806
and psychology 19:406–407, 19:410
realist critique of traditional 10:788
science and technology studies 21:203–204
scientific evidence and 21:297, 21:300
of social constructionism 22:222–223
sociology of 22:984–988
Taylor on 24:41
uncertainty 24:716–717, 24:720
of wisdom 25:600–601
hermeneutics, poststructuralism, and analytic philosophy 25:601
impossibility of any meta-position 25:601
limits of justification 25:601–602
relevant elements 25:601–602
scientific pluralism 25:601
shifting from practical to theoretical philosophy 25:600–601
theoretical reason with characteristics of practical reason 25:601
uncertainty of fundaments 25:601
Wittgenstein on 25:601–602
Epistle to Diogentus (Christian apologetics) 3:548
Epistopics 20:96
EPL *see* Employment protection legislation (EPL)
EPM *see* Emotional processing model (EPM)
EPMM *see* Essentially parallel measurements model (EPMM)
EPO *see* European Patent Office (EPO)
Epoch, Anthropocene 1:722–727
Epochal concept of modernity 7:656, 7:660
Epochal shifts in the work force 17:313–314
Epoché 11:442
Epoptic
definition 25:703, 25:707, 25:705f
monitoring process 25:705
spatial and nonspatial information 25:705–706
working memory content 25:707
EPOR *see* Empirical program of relativism (EPOR)
EPQ *see* Eysenck Personality Questionnaire (EPQ)
EPR *see* Extended producer responsibility (EPR)
EPRDF *see* Ethiopian People Revolutionary Democratic Forces (EPRDF)
EPS *see* Extrapyramidal system (EPS)
EPs *see* Evoked potential (EPs)
EPSI *see* Erikson Psychosocial Stage Inventory (EPSI)
Epsilon productions 19:10
EPSP *see* Excitatory postsynaptic potential (EPSP)
Epstein, A.L. 1:768–769
Epstein, C., works, *Fighting for Time* 24:401
Epstein, J., on artificial societies 1:368–369
Epstein, Klaus, works, *The Genesis of German Conservatism* 4:684
Epstein, Steven, on epistemic cultures 7:876
Epstein–Barr virus (EBV) 12:35
EPT *see* Emotional processing theory (EPT)

EPTA *see* European Parliamentary Technology Assessment Network (EPTA)
EPZs *see* Export processing zones (EPZs)
EQ *see* Encephalization Coefficient (EQ)
EQ-5D *see* EuroQol (EQ-5D)
EQF *see* European qualification framework (EQF)
EQS 3:787–788
Equal access 9:729–730
Equal access theory 7:900
  critiques of 7:900–901
Equal but different 16:28–29
Equal compensation principle 1:346
Equal Credit Opportunity Act 1974 11:911
Equal division core 4:872
Equal Employment Act 1972 1:257
Equal Employment Opportunity (EEO) 11:351–352
Equal Employment Opportunity Commission (EEOC) 1:259, 4:741, 6:530–532, 6:598, 21:722
Equal environments assumption 2:451–452
Equal justice 22:393
  limitations of 22:395
Equal Justice Works (EJW) Fellowship 19:557
Equal Justice Works Guide to Law Schools 19:557
Equal participation and representation 13:436
Equal pay, and job evaluation 12:817
Equal Pay Act 1963 11:882
Equal protection 12:137
  of rights 13:436
Equal Protection Clause 13:768
Equal Rights Amendment (ERA) 13:594, 15:110
Equal treatment 13:436
Equal validity 20:237
Equality 1:965, 2:146–147, 2:149, 2:153, 3:643–644, 4:707–708, 4:869, 6:85, 6:87, 7:308, 7:310, 9:729–730, 12:235–237, 12:240–241, 13:438, 14:254–255, 17:619–620, 21:769–770, 21:772, 22:386, 23:365–366, 12:236t
  in decision-making 6:87
  and difference 7:117
  economic philosophy 7:103
  and feminist legal theory 8:919–920
  and freedom 9:398
  gender 13:722
  and justice 12:938, 12:950
  legal aspects 7:894–902
    antisubordination 7:896, 7:899
    color blindness 7:894–896
    courts, role of 7:894–898, 7:900–901
    critiques of equal access theory 7:900–901
    critiques of formal equality 7:896–897
    critiques of outcome theories of equality 7:899–900
    critiques of structural inequality 7:898–899
    formal inequality of treatment 7:894–897
    full participation 7:897–898, 7:901
    inequality of access 7:899–900
    inequality of outcome 7:899–900
    post-racialism 7:895
    racial polarization 7:895, 7:899–900
    second generation discrimination 7:896
    stigma 7:895, 7:898
    structural inequality of treatment 7:897–899
  and liberalism 14:26–28, 14:30–31
  in liberalism 14:22
  in Melanesia 17:155–156
  modernization 15:697
  and multiculturalism 16:22, 16:26
  of multivariate distributions 16:889
  philosophical aspects 7:906–911
    controversies about possibly legitimate sources of inequality 7:910–911
    democracy, and contractarian methodology 7:909–910
    dignity 7:906–907
    fairness 7:907–909, 7:911
    justice 7:906–911
    as moral and political principle 7:906–907
    natural equality and social treatment of individuals 7:908–909
    significant inequalities and ideal of equality 7:907–908
    social contract 7:909–911
  principle of 2:353
    implementation of 2:354–355
  promotion of 13:720–722
  and queer anarchism 19:749
  racial 6:219
  radical lesbian feminists 8:944
  of several populations 16:888
  of sexual minorities in South Africa 21:746–748
    politics of pride 21:747
    representation of black lesbians 21:747–748
  social dominance orientation 22:251
  social inequality 22:338–344
  social justice 22:392
  subordination and 9:730
  in welfare economics 25:498–500
Equality Act 2010 1:252
Equality of arms, in criminal procedures 5:230
Equality of opportunity 6:770–771, 7:903–905
  achievement 7:904
  ascription 7:904
  and competence 7:904–905
  economic inequality 7:903
  industrialization 7:904
  inequality of conditions 7:903
  inequality of educational opportunity 7:904–905
  meritocracy 7:904
  and mobility 7:903–904
  social inequality 7:903
  social justice 7:903–904
  and social justice 22:398
  social mobility 7:903–904
  and stratification 7:903
Equality rights 3:687
  of minorities 3:689
Equality within families 20:319
Equality–efficiency trade-off
  and income tax 11:747–748
Equally distributed equivalent (EDE) income 11:905
Equal-status contact 18:843–844
Equal-variance assumption 21:947
Equation-based approaches, hierarchy models 10:838–839
Equations
  graphics 10:359
  and structures, symmetry relationship between 14:795–798
EQUATOR *see* Enhance the QUAlity and Transparency Of health Research (EQUATOR)
Equidistant semantic scales 21:563, 21:563f
Equifinality 23:612
  of organizations 3:895
Equilibration 16:501, 18:126–127
  cognitive development 4:52–53, 4:55–56, 18:120–123
Equilibrium 4:844, 4:846–847, 7:912–917, 9:575, 9:577–578, 15:303–304, 25:10–11
  analogical uses of 7:913–914
  analysis in today's macroeconomics 7:917
  analytical concept of 7:914–915
  arguments, early forms of 7:912–913
  axiomatic analysis 7:915–916
  competition 7:913, 7:915–916
  DSGE models 7:917
  dynamic equilibrium 7:912, 7:914–915
  in economic search 21:336
  economic theory 7:912, 7:915–917
  existence of 7:915–917
  free market 7:915
  game theory 7:916–917
  general equilibrium theory 7:912, 7:914–917
  institutions 12:190–191, 12:194–195
  marriage 8:831–833
  models 7:916–917
  in organizational development 6:296, 6:299–300
  partial equilibrium theory 7:914, 7:917
  perspective of internal migration 12:444
  and politics modeling 18:512–514
  in population studies 8:563–564
  queues 19:786
  reflective 20:77
  stability of equilibrium 7:915–916
  static equilibrium 7:912, 7:914–915
  structural unemployment 24:738–742
    flow approaches 24:738
    PS–WS model 24:739–740
    stock approaches 24:738
  and time 7:915
Equilibrium analysis 14:624

Equilibrium distribution 19:786
Equilibrium firm 20:727
Equilibrium models, of functionalism 9:533
Equilibrium price dispersion 12:102
Equilibrium search models, effects of labor market policies in 12:103
Equilibrium theory, realism as 20:27
Equilibrium-point hypothesis (EPH) 15:959
  motor control models 15:959
Equipments
  ethnographic 8:188
  experimental laboratories 8:554
Equipotentiality 13:392
Equitable education 1:43, 1:45–46
Equitable remedies 4:809
Equitable selection of subjects, Common Rule 4:256
Equity 4:247–248, 12:916–917, 12:938, 17:380, 17:619, 18:275–276, 20:801–802, 20:805–806
  as an alternative to state law 7:924
  for business 4:941
  as a common expedient 7:924
  common law and 7:925
  contemporary theoretical views 7:927
  critiques 23:328–329
  in education 7:918–923, 22:330
    interventions to achieve 7:919–921
    measurements 7:921–922
  and emotions 7:537
  in England 7:925
  as a foundation of trading societies 7:925–926
  and health 10:618–619
  of health care 10:604–605
  imperfections of law and 7:924
  in personam 7:925
  and justice 12:949–953
  and law 12:934
  meritocracy as 15:239–240
  in neoliberal legal and political philosophy 7:926–927
  of opportunities 7:926
  as a referent 7:926–927
  as replacement to legal rules 7:924
  Roman legal regulation 7:924
  spatial *see* Spatial equity
  in sustainable transportation 23:848
  treatment research and access 8:32, 8:34–35
  in United States 7:926
  and urban poverty 24:917
  uses 7:927
  water and sanitation 25:435
  in Western legal tradition 7:924
Equity allocations 17:380
Equity financing 20:16
Equity funding 1:249–250, 20:13
Equity in law 7:924–928
Equity investors 20:16
Equity ownership 20:13
Equity planners 1:218
Equity planning 12:914, 19:823
  see also Advocacy planning
Equity premium puzzle 15:298
Equity returns, stock market 23:486–487
Equity Scorecards 7:922
Equity theory 7:537, 17:380, 19:367
  and justice 12:928–930
Equivalence 5:318, 5:322–323, 22:457–458
  in cross-cultural research 7:929–933
  structural 23:621
  types of 5:319–321, 5:323–324
Equivalence classes 17:297
Equivalence Constraint 3:919, 3:923
Equivalence factor 3:159
Equivalence models 19:427–428
  congeneric measurements model 19:428
  essentially parallel measurements model 19:427–428
  essentially tau-equivalent measurements model 19:428
  tau-equivalent measurements model 19:428
Equivalence of siblings, and kinship 13:65
Equivalence principal 14:859
Equivalence scale 11:733
Equivalence theorem, optimal experimental design 17:258
Equivariance, quantile regression 19:715
Equivocality, of computer-mediated communication 4:505
ER *see* Exclusion restriction (ER)
ERA *see* Equal Rights Amendment (ERA)
Era of commemoration 4:237
Era of the criminological silence 5:239–240
Erasmians 5:664
Erasmus
  on humor 11:416
  teaching and research infrastructure 22:672
Erasmus intellectuals 5:664
ERB *see* Ethical Review Board (ERB)
ERC *see* European Research Council (ERC)
ERCE *see* Regional and Comparative Regional Study (ERCE)
Erdmann, Benno 19:391
E-readers 6:899
Erectile dysfunction (ED) 21:690, 21:714
  and sexual motivation 21:763
Eretz Yisrael Shelanu (Our Land of Israel) 18:411
Erga-omnes effect 4:700–701
Ergative 11:826
Ergative (ERG)–absolutive (=nominative) patterns 10:317
Ergative languages 14:210, 14:212–213
Ergative predicate 10:327
Ergodic averages, Markov chain Monte Carlo methods 14:569, 14:571
Ergodicity 16:114
  inverse projection 6:150
ERGOMAS *see* European Research Group on Military and Society (ERGOMAS)
Ergonomics 11:297–305
  background 11:297–299
  cognitive systems engineering 11:298, 11:300–301
  group level analysis 11:301–302
  individual levels of analysis 11:299–301
    cognitive ergonomics 11:299–301
    motivational factors 11:301
    physical ergonomics 11:299
  macroergonomics 11:302–303
  organization level analysis 11:302–303
  scope of 11:298f
  situation awareness 11:299–302
Ergotropic regulation 23:294–295
ERI *see* Effort–reward imbalance (ERI) model
ERIC *see* European Research Infrastructure Consortium (ERIC)
Erickson, Bonnie H., on genre knowledge 16:556
Erickson, Fredrick 8:132
Ericksonian hypnosis 11:459, 11:462–463
Erigena, John Scotus 20:285
Erikson, Erik Homburger 2:350, 7:934–937, 11:532, 11:535, 11:537
  biography 7:934
  concept of identity 11:537–540
    classic psychoanalysis 11:537
    criticisms of 11:539–540
    object relations theory 11:537–538
  contributions of 7:936–937
  criticisms against 7:936
  ego identity *versus* identity confusion, crisis of 7:934
  on epigenetic scheme 11:539–540
  life span developmental schema of 7:934
  professional career of 7:934
  on psychoanalytic development 7:935
  see also Psychosocial developmental stages
  stage theoryof 1:435–436
  theory of psychosocial development 7:938–946
  validity of stages 7:937
  works
    Childhood and Society 7:934–935
    Dimensions of a New Identity 7:935
    Gandhi's Truth 7:934–935
    Identity: Youth and Crisis 7:934–935
    Identity and the Life Cycle in Psychological Issues 7:934
    Insight and Responsibility 7:934–935
    The Life Cycle Completed: A Review 7:935
    Life History and the Historical Moment 7:935
    Reflections on Dr Borg's Life Cycle 7:935
    Toys and Reasons 7:935
    Vital Involvement in Old Age 7:935
    Young Man Luther 7:934, 19:349
Erikson, Joan 7:935
Erikson Psychosocial Stage Inventory (EPSI) 7:940, 7:942
Eriksonian approach, to identity development 21:434–436
Erikson's developmental model 14:86
Erikson's psychosocial theory 11:538
Erikson's stage theory of adolescent development 6:681
Eritrean–Ethiopian War 25:421
eRm software 19:421
ERN *see* Error-related negativity (ERN)
Ernst Mach Society 14:314–315

EROI *see* Energy return on investment (EROI)
EROI/EROEI *see* Energy return on energy investment (EROI/EROEI)
Eroles, Rosa Florentina 7:113
Eros 9:647
Erotic capital 8:1–7, 17:490
  definition 8:1–2
  distribution of 8:3
  economic value of 8:3–4
  future research directions 8:6
  importance of 8:2–3
  in intimate relationships 8:5
  political value of 8:4–5
  research methods for measuring 8:5–6
  social value of 8:4–5
Erotic desire 18:613
ERP *see* Exposure and response prevention (ERP)
ERPs *see* Event-related brain potentials (ERPs); Event-related potentials (ERPs)
Errata 18:917
Erroneous questions 19:53–54
Error analysis (EA) 21:361
Error calibration models 14:850–851
Error correction models 7:9–10
  cointegration 24:322, 24:325
Error of judgment 19:471
Error of rejection 11:495
Error of sensory measurement 19:470
Error rates, hypothesis tests 11:500–501
Error response times, diffusion models 6:399
Error score variable, mean, classical true score model 19:424
Error variables 3:786–787
  latent variables 13:394
Error-in-variables regression, in EITMs 18:458
Errorless learning (EL) for people with language disorders 2:816
Error-rate estimation 16:119
Error-related negativity (ERN) 17:123–124
Errors 14:154, 20:50, 22:550, 24:642
  data 6:126–127
    errors of content 6:127
    errors of coverage 6:126–127
    sampling errors 6:126
  of measurement 20:239–240, 20:242, 20:248
  nonsampling 16:919–923
  objectivity of research 17:93–94
  in person perception 17:739–740
  polling 18:462–463
  population forecasts 18:595
  social measurement 21:7
  source monitoring 23:29
  in statistical assumptions 8:8–12
    data set 8:10
    linear regression 8:11
    sensitivity of Bayesian procedures 8:11
    theoretical results 8:9
    t-test 8:8–11
    violations of common distribution 8:10
    violations of independence 8:10–11
    violations of normality 8:9–10
  theory of 19:470
  see also Catastrophic errors
ERS *see* Earth Resources Survey (ERS) program; Extreme response style (ERS)
ERS-1 *see* European Radar Satellite (ERS-1)
Ersoy, Bulent, and homosexuality 20:870
Ertel, S., on semantic differential 21:567
ERTS-1 *see* Earth Resources Technology Satellite (ERTS-1)
E-rulemaking 1:110
Erythroxylon coca 19:463
ES *see* Experience seeking (ES)
ESA *see* European Space Agency (ESA); European System of Integrated Economic Accounts (ESA); Event-structure analysis (ESA)
Escalation 24:122–123
  and conflict 4:585
  deterrence 6:248
Escape training 2:408
Eschaton 21:388
Escitalopram 1:781
Escobar, Arturo, works, *Territories of Difference: Place, Movement, Life, Redes* 18:353–354
ESCs *see* Embryonic stem cells (ESCs)
ESD *see* Education for sustainable development (ESD)
ESDA *see* Exploratory spatial data analysis (ESDA)
ESEA *see* Elementary and Secondary Education Act 1965 (ESEA)
ESEM *see* Exploratory structural equation models (ESEM)
Esenin, Sergei 21:741–743
ESF *see* Education for a sustainable future (ESF); European Science Foundation (ESF)
ESHRE *see* European Society for Human Reproduction and Embryology (ESHRE)
ESI *see* Electronically stored information (ESI); Empirically supported intervention (ESI) guidelines
ESIM *see* Elaborated social identity model (ESIM)
ESM *see* Experience sampling method (ESM)
Esmeraldas, Maroons 14:593
E-sociology 17:444
Esotericism, and wisdom 25:598, 25:602
ESP *see* Extrasensory perception (ESP) ability
Esping-Andersen, Gøsta S. 3:659
  works, *The Three Worlds of Welfare Capitalism* 12:183
Espionage 11:393, 11:396, 11:398
  counterinsurgencies 5:99, 5:101
ESPON program 24:801
Espoused values, corporate culture 4:924–925
Esprit (magazine) 2:359, 20:665
Esquilleu Cave, Spain 17:459
ESR *see* Electron spin resonance (ESR); Empirical social research (ESR)
ESRC *see* Economic and Social Research Council (ESRC)
ESS *see* European Social Survey (ESS); European Standard Set (ESS)
Ess, C., works, *Digital Media Ethics* 12:612
Essen-Moller 17:30
Essential life span 17:173
Essential unidimensionality 6:437–438
Essentialism 8:13–15, 9:731–732, 14:14, 20:839
  and creolization 5:119–120
  definition 8:13
  and ecofeminism 6:846
  empirical investigations 8:14
  evolution of language 8:365, 8:368
  gender 8:14
  groups 8:13–15
  lay theory 8:14
  mental disorder 8:15
  personality 8:15
  prejudice 8:13–15
  of psychology 22:224
  race, ethnicity, and culture 8:14
  sexism 8:14
  sexual orientation 8:14
  in social psychology 8:13–14
  stereotyping 8:13–15
Essentialist tendency 4:532
Essentially contested concept, power as 18:765, 18:768
Essentially parallel measurements model (EPMM) 19:427–428
Essentially tau-equivalent measurements model (ETEMM) 19:428
Essig, Laurie, on Russian sexual life 21:740–741
ESSnet-Culture 5:730
ESSOL *see* European Standard Set of Loci (ESSOL)
EST *see* Ecological systems theory (EST)
Establishment data sets 5:743
Establishment Directive, 1998 3:672
Establishment economic data sets 5:743
Estate tax data 25:452–453
Estates 11:896, 19:174–176
  assemblies of 23:364
Esteem 4:849
  Honneth on 11:178
Estes, William K. 13:690–694, 14:812
Esthetic judgment 21:604
Esthetics 1:527
  neuroesthetics 16:656–663
Estimability, statistical 23:405
Estimated true score, classical true score model 19:426
Estimates, small area 6:162–169
Estimating slavery 22:44–45
Estimation
  fixed-effects statistical procedures 15:273
  heterogeneity evaluation 15:274
  of inequality 11:732
  inference models 15:272–273
  point and interval 8:16

random-effects statistical procedures 15:273–274
robust variance 15:275
of T2 15:274, 15:277
Estimation, of statistical parameters
in econometrics 7:9–10
instrumental variables 12:205–208
interval estimation 8:18–19
Bayesian intervals 8:19
distribution-free intervals 8:19
one-sided intervals 8:19
pivotal intervals 8:18–19
prediction intervals 8:19
test statistic 8:18
tolerance intervals 8:19
linear hypothesis 14:145–146
parameter 24:344
point estimators 8:16–17
Bayes's 8:17
evaluation 8:17–18
maximum likelihood 8:16
method of moments (MOM) 8:16
resampling methods for 20:492–495
statistical 23:407–408
of sufficient summary plots 14:159–160
theory of, in sampling 20:877
uncertainty 24:344
Estimation abilities 4:66, 4:69–70
Estimation error variances 15:273
Estimation procedures, econometric software 7:4–5
Estimation strategies
questionnaires 19:774–775
decomposition 19:774
response alternatives 19:774–775
subjective theories 19:774
Estimators
asymptotic distribution of 24:328
instrumental variables 12:206–207
latent variables 13:394
eSTM *see* Early short-term memory (eSTM)
Estonia, adult mortality in 1:162–165, 1:162f, 1:164f–165f
Estrogen 21:774
Estrous swelling 18:899
ESTs *see* Empirically supported therapies (ESTs)
ETC *see* Embodied technological change (ETC)
ETEMM *see* Essentially tau-equivalent measurements model (ETEMM)
Eternal peace, Kantian politics 13:9
Eternitas 21:388
ETFs *see* Exchange traded funds (ETFs)
ETHIC model 22:785
Ethic of care 12:936
Ethic of connection, religiosity of women 25:636
Ethical, legal, and social implications (ELSI) program 2:681, 8:43–47, 9:1017–1018
bioethics 8:43–45
of bioethics 2:627
DNA, use in courtroom 8:46
funding mechanisms and mechanics 8:44
genetics 8:43–46
of genetics 2:619, 10:32
of genome sequencing 11:313
of genomics in Asia 11:319–321
Human Genome Project 8:43
human subjects research 8:44
impact of 8:45–46
integration with genomics research and policy activities 8:44–45
National Institutes of Health 8:43–45
portfolio 8:44
privacy 8:45–46
program priorities 8:44
qualitative research 8:44
quantitative research 8:44
social sciences 8:44
Ethical, legal and social aspects (ELSA) program 2:619
Ethical accountability 2:648
Ethical behavior, autonomous agents 2:311–312
Ethical beliefs and practices, diversity and disagreement of 6:589
debates about implications of 6:590
fundamental disagreement 6:590–591
individual ethical uncertainty and 6:591–592
methodological concerns 6:590–591
nature and sources of 6:589–592
in practice 6:593–594
as a source of skepticism 6:593
variety of ethical objectivism and 6:592
variety of nonobjectivism and 6:592–593
Ethical codes, professional 8:21–24
Ethical conduct of research of former child soldiers 3:431
Ethical conflicts, in health care rationing 10:605
Ethical considerations, ethnomethodology 21:194–195
Ethical consumers 8:25–30
approaches to understand 8:27–28
boycotting 8:25–27
buycotting 8:25–26
characteristics 8:25–27
historical context 8:26–27
what do they do? 8:25–26
where are they? 8:26
who are they? 8:25
conspicuous consumption 8:28
consumer activism 8:26–28
consumer movement 8:26–27
critical remarks 8:28–29
ethical purchasing gap 8:27
ethical simplifier 8:25–26
future research 8:29
green consumers 8:25
green consumption 8:27
methodological issues 8:28
moral consumers 8:25, 8:27–28
sustainable consumerism 8:25–26
voluntary simplicity 8:25–26
Ethical consumption 8:27–28
Ethical cosmopolitanism 17:652
Ethical dilemmas
and armed conflict 1:973
social work ethics 22:784–786
Ethical guidelines
antisemitism 1:785–786
ethnic monitoring 1:786
Ethical hedonism 24:980–981
Ethical imagination, and sociality 22:858–859
Ethical implications of social control 24:121–123
automatic processes 24:121–122
displacement 24:122
escalation 24:122–123
goal conflicts 24:121
neutralization 24:122–123
second-order effects 24:122
unintended consequences 24:121–122
Ethical isolationism 6:593–594
Ethical issues
of antidepressants 1:781
in combined pharmacotherapy and psychotherapy 19:498–499
forensic genetic databases 9:343–344
genetic factors (in organizational psychology) 9:923
in genomics 10:32–41
and indigenous people 9:963–964
social work ethics 22:784–786
Ethical liberalism, and Christianity 14:8
Ethical objectivity 6:589, 6:592
Ethical plateaus, and knowledge translation 21:184
Ethical pluralism 6:594
Ethical poverty line 10:634
Ethical practice 22:806
Ethical principles 20:515–516
beneficence 20:518
fair distribution of research risks and benefits 20:519
informed consent 20:516
informing participants 20:516
and objectivity 17:93–95
conflicts of interest 17:94
error, bias, and disagreement 17:93–94
honesty 17:93
misrepresentation of data 17:94
openness 17:94–95
participant privacy, protection of 20:518
respect for persons 20:516–517
Ethical problems
euthanasia 8:282–283
social work ethics 22:783, 22:785
Ethical purchasing gap 8:27
Ethical questions 8:53–59
codes of ethics 8:53–58, 8:54t
early debates and institutionalization 8:53–56
anthropology 8:55–56
psychology 8:55
sociology 8:55
experiment 8:53, 8:57–58
freedom of inquiry 8:58
journal 8:53
leading cases 8:56–58
Alcoholics Anonymous 8:57

Ethical questions (*continued*)
Milgram Experiment 8:57–58
Project Camelot 8:58
Springdale case 8:56–57
participant observation 8:57
prehistory 8:53
professional association 8:59
relevant factors 8:58–59
research funding 8:53, 8:55–56, 8:58
research politics 8:58
responsibility 8:57–58
Ethical relativism 6:590–591
Ethical Review Board (ERB) 8:49
Ethical simplifier 8:25–26
Ethical trade, and consumption 4:746–747
Ethical turn 8:60
Ethical universalism 5:12–18
Ethical/societal concerns, forensic genetic databases 9:342–343
Ethics 2:631, 5:931–932, 6:666, 8:71–75, 12:136–137, 14:186–187, 15:813, 16:142, 17:92, 20:531–532, 20:667, 24:127–128
in age of genetic engineering 20:897
Albert on 1:498–499
in anthropological research 1:734–739
among emergent frontiers 1:738
anthropology of morality 1:737–738
bioethics 1:736–738
cultural relativism 1:737
ethics and anthropological science wars 1:734–735
ethnography 1:734–738
genealogies and normativity 1:735–737
history of science 1:734–735
human rights 1:737
human subjects protection 1:736
nation-state 1:735–737
normative structures 1:734, 1:736, 1:738
self-policing 1:734–736, 1:738
values 1:737–738
and antisocial behavior 1:792–794
of appeasement 1:828–829
in archaeology 23:745
archiving 1:916–920
aspects of intellectual property rights 12:262–265
aspects of research publication 20:529–533
biobank 2:604
bioethics *see* Bioethics
for biomedical research involving humans 8:78–82
and brokers 2:873
of care 2:614
feminist social work 8:936, 8:939
childhood cancer 3:461
of citizen science 3:634–636
and copyright 17:288–289
ethnography 8:190
eudaemonism 8:212–214, 8:216
fieldwork 9:153–155
and gendered care 9:801–804
genetic testing/screening and diagnosis 20:698–705
and human rights 11:353, 11:356, 11:358
human terrain system (HTS) 11:396–397
image 14:690–691
implications of cluster randomized trials 3:897–905
in intellectual property rights 12:264–265
and Internet 12:589
Internet research 12:606
in interviewing 5:853–855
issues, AIDS 1:493–494
journalistic 12:858–859
Kantian 13:5–10
and law, boundary between 8:65
Linguistic Society of America guidelines 14:186
media *see* Media ethics
mentoring 15:235, 15:237
metaethics 8:71–73
Milgram's obedience studies 17:65–66
neuroethics 16:664–669
normative ethics 8:73–75
participatory action research (PAR) 17:541–542
philosophy 18:32, 18:35
of placebo-controlled trials 18:164–173
planning 18:193–197
in codes of professional organizations 18:194–196
Plato on 18:211–212
pornography 17:219, 17:221
of privacy/confidentiality in social research 18:942–943
of program evaluation 5:55
psychiatric aspects of religion 20:349
and psychiatry 8:63–70
of public policy 19:580–585
administrative ethics 19:583
civil service ethics 19:583
corruption 19:583
decision making 19:580–581
evaluation 19:584
governance 19:580, 19:582–583
interest 19:581–582
policy circle 19:584
political ethics 19:583
principles 19:582–583
professional ethics 19:580
public administration 19:582–583
responsibility 19:583
stakeholder analysis 19:581–582
style 19:583–584
trust 19:584
values 19:582
publishing 12:864
Putnam on 19:664, 19:666
for randomized controlled trials 8:543–544
recognition politics 20:44, 20:46–48
relativism 20:233–237
research 20:514–522
research and access to treatment 8:31–36
research funding 20:523–528
social psychology 22:535–536
social representations 20:446
social work 2:109, 2:111–112, 22:782–788
social work with refugees 22:773
street children 23:537
in surveys and polling 23:752–757
Taylor on 24:42
turn in international development 8:60–62
of unobtrusive measures 24:770–771
and values 8:71–77, 25:33–35
virtue 25:147–152
visual methods 25:182
Ethics (internet research) 12:606–607
in digital age 12:607–611
dilemmas in conceptualization of 12:608–611
human subjects 12:608
informed consent 12:610
personally identifiable information 12:610–611
privacy concepts 12:608–610
in internet age 12:607–611
origin for scientific research 12:607
beneficence 12:607
justice 12:607
respect for persons 12:607
provocations for 12:611–612
Ethics committees, in science 8:49
Ethics tool 7:90–91
game theory 7:90
social choice theory 7:90
utility theory 7:90
Ethics work, social work ethics 22:787
Ethiopia 1:957, 22:611–614
child soldiers in 3:439–444
context for engagement 3:440–441
guerrilla fighters 3:439–443
methodological considerations 3:439–443
military regime 3:439–444
political activism 3:441–443
political engagement 3:440, 3:442–443
revolution 3:440–444
education 22:612–613
education in 7:199, 7:202
funding 22:612
psychiatric care in 6:42
research 22:613–614
Ethiopian Left 3:440
Ethiopian Orthodox Church 17:437
Ethiopian People Revolutionary Democratic Forces (EPRDF) 3:441
Ethiopian People's Democratic Movement (EPDM) 3:441
Ethnic, racial, and nationalist (E/R/N) movements 8:123
cognitive perspectives 8:127
cultural perspectives 8:127
evaluation of various perspectives 8:128–129
research questions 8:124–125
theoretical perspectives 8:125–127
colonialism and ethnicity 8:125
economic and instrument approaches 8:125–127
state- and nation-building 8:125
Ethnic and racial social movements 8:123
Ethnic arts 2:17
Ethnic Chinese, international migration by 12:525

Ethnic cleansing 8:83–90, 10:2–3, 10:7, 10:11, 15:571, 16:30–31, 17:436, 25:120
as classicide 8:85
comparisons, cases, and issues 8:87–88
definitions and properties 8:83–84
democratization 8:84, 8:86–87
displaced persons 8:85–88
as eliticide 8:85
environmental conflict 8:87–88
ethnic conflict 8:83
genocide 8:83–88
in historical contexts 8:84–85
human rights 8:86
humanitarian technologies 8:88
nationalism 8:83–86
nation-states 8:84–86
as politicide 8:85
population transfers 8:85–86
refugees 8:88
resource conflict 8:87–88
theorizing 8:85–86
war crimes 8:83–84, 8:86–88
Ethnic clustering, ghetto 10:123–125
Ethnic communities 8:136–137
Roma 20:755–757
Ethnic competition 8:126
Ethnic composition
and homicide 11:164
Ethnic conflicts 4:589, 8:83, 8:95–101
and consociationalism 4:692
datasets 8:98t
definition 8:95–96
description and classification 8:96–97
distribution of cases 8:98t
and Eastern European social work 22:796–797
explanation 8:97–99
geography of 8:91–94
competition 8:93
demographic balance 8:93–94
diaspora 8:91–92
ethnonationalism 8:92–94
identity 8:91–92
immigration 8:91–92
at interstate scale 8:91–92
at intrastate scale 8:92–93
irredentism 8:91–93
at microscale 8:93
nationalism 8:92, 8:94
polarization 8:91–93
population transfer 8:92, 8:94
regulation and resolution 8:94
secession 8:91–94
segregation 8:93
spatial scales 8:91
symbolic order 8:93
territory 8:91–94
outcomes 8:99–100
and war 25:367–369
Ethnic differences, in counseling and psychotherapy 5:63–68
Ethnic discrimination
acculturation and 1:73
and adolescent health 1:125
Ethnic disparities, in sentencing 21:628–630
Ethnic display events 3:147
Ethnic diversity, in Southeast Asia 23:71
Ethnic enclaves, job search and working in 8:110–111
Ethnic entrepreneurship 7:688–689
Chinese international migration 12:527
Ethnic exclusion 8:127
Ethnic foods 9:290–295
Ethnic ghettoes 18:600–601
Ethnic groups
approaches to 8:137–140
and nations 16:293
Ethnic hierarchies/bureaucracies 5:172
Ethnic homogenization, in Europe 8:277
Ethnic identities 8:106–107, 8:113–117, 8:123, 8:126–129, 8:137, 9:434–435, 11:676, 19:835
acculturation psychology 8:113, 8:115
and archaeology 8:102–105
affiliation 8:102, 8:104
anthropology 8:102–104
Childe on 8:103
community 8:102
culture 8:102–104
culture history 8:103
nationalism 8:104
new archaeology 8:103
of Australian aboriginals 2:236
defining 8:113
development of 8:114–115
discursive psychology 8:113–114, 8:116
ethnicity 8:113–116
and genomics 10:34
identification 8:113–116
identity development 8:114
identity negotiations 8:116
and immigration 11:676
intergroup relations 8:113–116
and intermarriage 8:120–122
measurement 8:107–109
empirical advancements 8:109
ethnosizer 8:108–109, 8:108f
theoretical advancements 8:107–108
multiple identities, complex structure of 8:115
narrative network analysis of 16:214
and national attachment 16:311
national identity 8:115–117
negotiating 8:116
Roma people 20:753
social identity theory 8:113–115
social psychology 8:113–114
sociocognitive approach 8:114
spirit possession 23:282
study, analysis levels and perspectives 8:113–114
well-being and consequences of perceived discrimination 8:115–116
and work 8:106–112
assimilation 8:107–112
discrimination 8:106–107, 8:109, 8:111
employment 8:106, 8:109–112
endogeneity issues 8:111
human capital 8:107, 8:109–111
informal networks 8:110
integration 8:106–107, 8:109–111
job search and working in ethnic enclaves or primary market 8:110–111
labor force participation 8:106, 8:110
labor market 8:109–111
marginalization 8:108–112
national identity 8:106–107, 8:109, 8:112
oppositional identity 8:106–107, 8:110–111
separation 8:108–109, 8:111
wages 8:111
Ethnic intelligentsia 4:382
Ethnic intermarriage 8:118–122
ancestry 8:118–119, 8:121, 8:118t
assimilation 8:118–119, 8:121–122
assortative matching 8:120
consequences of 8:121
cultural transmission 8:118, 8:121
determinants of 8:119–121
ethnic preferences 8:119–120
marriage market characteristics 8:120–121
preferences for related characteristics 8:120
endogamy 8:118–122
endogamy rates 8:118t
ethnic identity 8:120–122
exchange theory 8:119–120
immigrants 8:118–121
prevalence of 8:118–119
significance of 8:118
theories of 8:119
Ethnic jokes 11:418
Ethnic Judeans 12:870
Ethnic marital matching 2:100t, 2:103t
Ethnic markers 8:139
Ethnic minorities 1:68
and education 1:729, 1:731
in postsocialist societies 18:698
prisoners 18:927
Ethnic Minority British Election Study 25:305
Ethnic minority identity 9:435
Ethnic mobilization 8:126–129, 20:846, 20:852–853
Ethnic nationalism 8:92, 8:123, 8:125, 16:262, 16:265, 16:310, 16:312
in periphery 16:277–278
Ethnic neighborhoods, Ghettos and 10:124–125
Ethnic pride, of Australian aboriginals 2:236
Ethnic solidarity 8:123
Ethnic South and Southeast Asians, international migration by 12:530–537
Ethnic statistics, of Roma 20:754–755
Ethnic stratification 22:710
Ethnic subcultures, organized crime as 5:172
Ethnic violence
and crowds 5:361, 5:364
and globalization 10:235–236

Ethnicity 3:123, 3:687, 3:689, 5:151–157, 5:182, 6:8, 6:10, 6:351, 16:287, 16:297, 19:857, 20:363, 20:407, 20:409, 20:411, 24:603, 25:820
Aboriginals 5:151, 5:155
and academic achievement 1:44
and academic motivation 19:790–795
and adoption 1:146
ancient *versus* modern 8:136
in anthropology 1:769–770
anthropology of 8:148–151
basic model 8:151
corporate groups and cultural differences 8:148
ethnic groups and boundaries 8:149
personally meaningful ethnicity 8:150–151
social constructed ethnicity 8:149–150
approaches to 8:137–140
and archaeology 8:102–105
affiliation 8:102, 8:104
anthropology 8:102–104
Childe on 8:103
community 8:102
culture 8:102–104
culture history 8:103
nationalism 8:104
new archaeology 8:103
in archaic states 23:374
in Asian Societies 12:728–729
Barth on 8:149–151
-based politics, diffusion 6:412
Benhabib on 2:519, 2:522–523
in Canada 8:152–156
in Central Asia 3:316–320
in China 2:66, 2:68
of clan 3:732
and collective identities 2:639
collective identity 8:148
and community sociology 4:363–364
composition of North American population by 18:553–558
cultural critique and 5:385–386
and culturally appropriate interventions 5:516–517, 5:519–520
and dance 5:668
definition 19:812–813
discrimination and disparity 5:155–156
Canada 5:155
United States 5:155
diversity 8:148
division of labor 25:341–343
and divorce 6:607
docial identity 8:148–150
and doctor–patient interaction 6:615–616
and educational achievement 8:130–135
conflict view 8:134
critical pedagogy 8:133
critical race theory 8:133
cultural deficit theory 8:131
cultural difference theory 8:132
cultural ecological model 8:132–133
labeling 8:134
multicultural education 8:132–133
paradigm 1: genetic deficit theory 8:130–131
paradigm 2: culture-relevant explanations 8:131–133
paradigm 3: structural inequality in education 8:133–134
race 8:130–131, 8:134
self-fulfilling prophecy 8:134
theoretical explanations for achievement gap 8:130–134
tracking 8:133–134
and essentialism 8:14
ethnic boundaries 2:762
ethnic identity 8:113–116
in Europe 8:142–147
assimilation 8:143, 8:145–146
discrimination 8:143–146
ethnic diversity and inequality 8:143
majority attitudes and intergroup threat 8:143–144
media representations and contact experiences 8:144
migration and integration policies 8:143
minority acculturation 8:145
minority adjustment 8:145–146
multiculturalism 8:142–145
prejudice 8:142–144
stereotype threat 8:144–146
future research 1:44
Geertz on 8:150–151
and health 15:99–100
data analysis challenges 19:813
disparate health services and treatment 19:815–817
evidence-based interventions 19:815
health-enhancing resources 19:817
healthy immigrant phenomenon 19:815
institutional discrimination 19:812, 19:816
Latino health paradox 19:813–817
life expectancy 19:812–815, 19:817, 19:813t
Preston curve 19:813
prevalence of disease 19:812, 19:818
resources and acculturation 19:817–818
socioeconomic status 19:814–815
variations 19:813–814
and homicide 11:165
in immigrant groups 15:407, 15:409
and immigration 15:397, 15:399–401, 15:405, 15:400t–401t
and inequality 11:897
and intercultural communication 12:357, 12:359–360
and intimate partner violence 12:686–687
and labor history 13:171
and language 13:251–255, 14:173
and marital quality 14:513–514
metaanalytical studies 1:44
minority ethnic elders 15:440–445
minority groups 5:151–156
national identity 8:151
nationalism 8:148, 8:150–151
and organizational diversity 4:738–739, 4:741–742, 6:595–599
and prejudice 8:171
and prosocial behavior 19:231
racism 8:150
and religion 20:379
and residential segregation 20:540–541
and segregation 21:407–409
social boundaries 2:762–763
and social class 22:342
and socialization of children 22:870
and socioeconomic status (SES) 1:44
sociology of 8:157–161
acculturation 8:159
assimilation 8:159
being identified 8:159–160
collective identity 8:157–160
cultural hybridization 8:159
definition 8:157–158
diaspora and transnationality 8:160
enclave 8:159
identification 8:158–159
multiculturalism 8:158–159
nativism 8:159
studies 5:151–152
symbolic boundaries 23:853–854
tribe 8:148–150
tribes 24:645–647
in twenty-first century 19:796–799
urban anthropology 24:774–779
and urban planning 19:820–824
use of term 11:755
visual discourse of 25:207–208
vocabulary, brief history of 8:136–137
in voting 25:305
wage differentials 25:338–339
Ethnic-linguistic infrastructure 20:746
Ethnic/nationalistic movements 8:127–128
international governmental associations 8:127–128
transnational ethnic social movements 8:127
Ethnics 10:6–7, 10:12–13
Ethnoaesthetics, art 2:16–17
Ethnoarchaeology 7:694, 8:162–165
action archaeology 8:162
and ceramics 3:325–326
ethnography 8:162
middens 8:163–164
middle-range research 8:162
modern, development of 8:162–163
refuse deposits 8:163–164
tribes 24:646
uniformitarianism 8:163
at work 8:163–164
Ethnocentricism 8:166–167
culture 8:166–167
primitive society 8:166
Sumner on 8:166–167
Ethnocentrism 2:255–256, 6:590, 8:137, 11:118–119, 16:30, 18:411, 21:262
and conflict 4:585
versus cultural relativism 1:765–766
of disciplines 3:70
and gift 10:142
history of 8:168–174
culture 8:169–171, 8:173

ethnicity 8:171
evolutionary explanations 8:169–170
intergroup relations 8:168–172
personality 8:171–172
self-aggrandizement explanations 8:171–172
and social groups 8:169
social identity theory 8:171–173
social norms 8:172–173
socialization and normative explanations 8:172–173
threat and conflict explanations 8:170–171
and magic 14:435–436, 14:438
and morality 15:806
and Orientalism 17:423–424
and public opinion 19:565
Ethnocentrism theory 4:621
Ethnocide 1:223, 8:175–176
Ethnocultural differences, in health 22:277
Ethnodrama 11:758–759
Ethnogenesis
in anthropology 8:175–177
belonging 8:175–177
ethnocide 8:175–176
mechanisms 8:176
theoretical background 8:175
in twentieth century 8:176–177
Ethnographic arts 18:824
Ethnographic film 6:620–624
affect 6:621, 6:623
audiences 6:620, 6:623
author as a fly in the soup – Cinema Verité 6:622
author as a fly on the wall – Direct Cinema 6:622
author as propagandist – Leni Riefenstahl 6:622
author as reconstructing authentic other – Robert Flaherty 6:621
author as reflexive roving eye – Dziga Vertov 6:621–622
author as social educator – John Grierson 6:621
contemporary issues 6:623–624
expository mode of representation 6:620–621
institutions 6:622
observational mode of representation 6:621
participatory mode of representation 6:621
performative mode of representation 6:621, 6:623
poetic mode of representation 6:621
politics of representation 6:621, 6:623
postcolonial 'other' as author 6:622–623
realism 6:621, 6:623
reflexive author 6:623
Ethnographic immersion 17:532
Ethnographic methods
childhood studies 3:451, 3:456
see also Fieldwork; in social and cultural anthropology
Ethnographic observation 22:545
Ethnographic representation, culture 5:584–585
Ethnographic research 19:676
Ethnographic studies 21:264–265
Ethnographically informed design, business anthropology 3:7–8
Ethnography 1:728–731, 1:772–774, 8:162, 8:178–183, 8:199–201, 9:153–154, 11:393, 17:530, 21:195–197, 22:545, 24:291
African 1:286–287
of anarchist movements 1:683–686
alterglobalization movement 1:684–685
existing anarchism 1:686
global uprisings 1:685–686
Spanish anarchism 1:683–684
and anthropological comparison 4:414
and anthropological writing 1:740–743
and anthropology 1:751–756
in applied social research 8:184–191
analysis 8:188–190
content analysis 8:189
culture 8:185
emic and etic perspectives 8:185
equipment 8:188
ethics 8:190
fieldwork 8:186
holistic perspective and contextualization 8:185
interviewing 8:186–187
key actor or informant interviewing 8:187
key events 8:189
nonjudgmental orientation 8:185–186
outcroppings 8:188
participant observation 8:186
patterns 8:189
projective techniques 8:188
qualitative research 8:187, 8:190
questionnaires 8:187–188
selection, sampling, and entry 8:186
statistics and software 8:189–190
thinking 8:189
triangulation 8:189
unobtrusive measures 8:188
writing 8:190
written and electronic information 8:188
approaches to organizational conflict 4:640–641
art as 2:16–18
autoethnography 22:539
of body 2:720–721
Chicago School of Urban Sociology 8:179–180
and common sense 4:2–4
communes 4:266–270
community studies 8:179
of corporate cultures 5:536
covert 8:182
cuisines 9:290–291
and culture of poverty 5:591
and drinking 6:658–661
early twentieth-century 5:382–383
Eastern Europe 6:807–808
and education 7:144–148
and embodiment 7:418
emic–etic analysis 8:180
emotion and affect in 1:236–239
engaged 8:182
and ethics 1:734–738
ethnographic intelligence 8:182
ethnoscience 8:180
external corroboration of 24:291
of feminism 8:900–901
fiction 8:181–182
field observational research 9:135–138
fieldwork 1:765–766
and fieldwork 9:149–152
of finance 9:174–175
founding fathers 8:178–179
of gender 15:356
and genealogy 21:184
and global development 10:150–151
and health 10:571–575
holism 8:179
and interpretation 12:642–644
Islamism 15:357
of knowledge 13:71
and life phases 5:507, 5:509–510, 5:512
and linguistics 14:170
media audiences 2:217, 2:220
militarized 8:182
and morality 15:807
multisited 8:182, 13:540
and multisited field studies 16:103–107
narrative 8:181–182
netnography 8:182
new 8:180
oral and literate expression 17:276
virtues 17:277
participant observation 8:179, 8:182
prison life 18:925–928
property 19:154–155
qualitative developmental psychology 19:672
of reading 12:643
and reflexivity 20:84–86
in reflexivity 20:88–92
resource, infrastructure as 12:124
sensory 21:616–617
in social anthropology, craft production 5:109
for sociocultural aspects in North America 17:24–25
of speaking 6:505
sports 23:299–304
and STS 21:182–183
testimonio 8:181
thick description and reflexive ethnography 8:180–181
triangulation 8:180
trickster 24:648–657
Trobriand Islands 1:767
urban 24:774–777, 24:779–780, 24:929
village, Korea 2:72–73
and work anthropology 25:663–665
food collection 25:663–664
food production 25:664
manufacturing 25:664–665
Ethnography and Education International 7:146
Ethnohistory 1:636, 8:192–197
and anthropology 1:746–747

Ethnohistory (*continued*)
anthropology of 1:746–747
applied anthropology 8:192
colonial studies 8:192
contributions to knowledge 8:193–194
cultural anthropology 8:192, 8:195
culture change 8:192, 8:195
development of 8:192–193
emerging directions in research 8:196
feminist scholarship 8:192–193
gender studies 8:192–193
historical anthropology 8:192–194
historical archaeology 8:192
indigenous studies 8:192, 8:196
interdisciplinary scholarship 8:193, 8:196
methodological issues 8:194–196
settler colonialism 8:195
Ethnologia Balkanica 1:756
Ethnologia Europaea 1:756, 8:200–201
Ethnological ethnocentrism 19:677–678
Ethnologie Française 1:756
Ethnology 1:746, 1:772–773, 8:198–202, 15:290, 20:326–327
and anthropology 1:751–756
anthropology and ethnography 8:199–200
cross-cultural comparison 8:199–200
and folklore 9:280, 9:283
as formative period of anthropology 8:198–199
future 8:201–202
history 8:198, 8:200–201
Lévi-Strauss on 13:904, 13:908
multiplicity in Europe 8:200–201
terminology 8:198
Voelkerkunde on 8:200–201
Volkskunde on 8:200–201
and work anthropology 25:665
world anthropologies 8:198
Ethnomathematics, and worldview 25:755
Ethnomedicine 5:572–573, 10:568–569
Ethnomethodologists 22:997–998
Ethnomethodology 6:499, 8:203–206, 12:244–245, 18:11, 22:986
common sense 8:203–205
contemporary research initiatives 8:205–206
people-processing institutions 8:205
science and technology 8:205–206
talk and social interaction 8:206
Garfinkel's studies 8:204, 9:598–604, 10:786
intellectual origins 8:203–204
language 8:203–204
negotiated orders 17:412
Parsons on 8:203–204
phenomenology 18:11–13
and phenomenology 22:994
reasoning 8:203–206
relationship of symbolic interactionism to 12:353–354
of science and technology studies 21:192–199
society 8:203–204
theoretical implications 8:204–205
and use of metaphors 15:292–293
Ethnomethology 21:148
Ethnomusicology 16:156
Ethnonationalism 8:92–94, 8:96, 16:285–287
East Asian studies 6:785–786
Ethno-pluralism 11:574–575
Ethnopsychology 19:363
art 25:793
customs 25:793
language 25:793
myth 25:793
Wundt's contributions 25:793–794
Ethnoracial domination 10:122
Ethnoregionalisms, in Europe 8:278
Ethnoreligious groups 20:363
Ethnos, people as 22:887–888, 22:890
Ethnoscience 8:812, 19:360
and ethnography 8:180
and worldview 25:753–754
Ethnosemantics 14:170
Ethnosizer 8:108–109, 8:108f
Ethnotheories
parenting 17:493
Ethograms 14:367
Ethological models of anxiety 1:808
Ethological tradition 7:467–468
Ethology 2:791, 5:679, 14:367–369, 22:900
emotional expression 7:458
evolution of language 8:364–365
and instinct 12:179–180
Skinner's contribution to 22:26
Watson's contribution to 25:449
Ethos 8:71, 23:525
of science 17:685–686, 17:688, 21:314–315
Etic, construct clarification 7:930–931
Etic aspects 7:930–931
Etic perspective, in ethnography 8:185
Etic–emic analyses 21:4–5
Etiological beliefs, in mental health 5:572–574
Etiological heterogeneity 9:933
Etiological models, for obsessive-compulsive disorder 17:122–124
Etiological theory of functions 9:520–522
Etiology
of homosexuality 3:890
of mental illness 15:204–208
of revolutions 20:638–639
rheumatoid arthritis 20:655
of suicide 23:682–687
Etiquettes 7:378–380
Etkind, Aleksandr, on Russian sexual life 21:740
ETM *see* Enrichment Triad Model (ETM)
ETM+ *see* Enhanced Thematic Mapper (ETM+) system
Être-au-monde 19:315–316
Etruscan alphabet 25:777
ETS *see* Environmental tobacco smoke (ETS)
Etsi Deus non daretur 11:32
ETT *see* Excitation transfer theory (ETT)
Etudes Ricoeuriennes/Ricoeur Studies 20:667
Études rurales 17:669–670
Etymological dictionaries 8:209
Etymology 8:207–211
definition and scope 8:207
genetic affiliation and 8:208
health 10:607
loanwords and history 8:210
other linguistic fields and 8:208
in past 8:207–208
present times 8:210
reconstruction and 8:208–209
theory and methodology 8:209–210
Etzioni, Amitai 4:313, 4:317, 18:204
works
A Comparative Analysis of Complex Organizations 4:320
Genetic Fix 4:320
The Hard Way to Peace 4:320
An Immodest Agenda 4:320
The Moondoggle 4:320
The Moral Dimension 4:320
New Communitarian Thinking 4:321
The New Golden Rule 4:321
Responsive Communitarian Platform 4:320
The Responsive Community (*I & We*), 1987 4:320
The Spirit of Community 4:321
EU *see* European Union (EU)
EU Commission Consumer Survey 8:515
EU Emissions Trading System (EU ETS) 18:370
EU ETS *see* EU Emissions Trading System (EU ETS)
Euarchontoglires
primates, evolution of 18:908
Eucken, Rudolf 19:392
Euclid of statistics 20:928
Euclidean conceptualization, of number system 17:59
Euclidean coordinates in physical space 25:198
Euclidean distance model 14:285, 16:34, 16:37
Euclidean geometry 25:197
Euclidean metric 25:198
Euclidian space 7:636–637
Euclid's geometrical laws 2:23
Euclid's theory of perspective 25:197
Eudaemonism 8:212–217
categorical imperative 8:214
classical tradition 8:212–213
critical thought 8:215
ethics 8:212–214, 8:216
eviction from philosophy 8:214
good life 8:212–213, 8:215–216
Kant on 8:214–216
moral philosophy 8:212–214, 8:216
in philosophy of early modernity 8:213–214
redemption, happiness, and political order 8:214–216
self 8:212–215
utilitarianism 8:213–216
Eudaemonistic virtue ethics 25:149
Eudaimonia 10:607, 10:611
Eudaimonian ethics 13:5

Eudaimonic approach to well-being 15:187–188
Eudaimonic enjoyment 13:833–834
Eudaimonic wellbeing 21:489
Eudemonia 10:522
Eugenics 2:610–611, 8:224–232, 8:237, 8:407, 9:908–910, 9:990–991, 9:995–996, 9:1015, 10:32–33, 10:36
in 1933–1945 8:242–244
after World War II 8:256–257
in Asia 8:233–236
in China 8:233–235
in Japan 8:235
in Britain 8:240
on the continent 8:226–227
and control of reproduction 8:219–220
criticism of 8:221–223
cultural and political factors 8:237–238
in decline 8:244–245
demography 6:177–178
in Europe, 4:475–479
Fisher's views 9:220
in formation (1904–1933) 8:240–242
in France 8:241, 8:243
future of 8:245
in Germany 8:241–244
hereditarian 23:426
history of, 2:27–31
idiosyncratic science of 8:237–238
and intelligence 12:310–311
legacy of 8:245
measures 8:255–256
movement 16:14
negative 8:241
origins 8:238–239
Pearson on 17:660–661
popularity of 8:254–255
as population policy basis of, 4:469–474
positive 8:240
program 16:15
quantification in 19:704
and race 19:802–804
research, aims and problems 8:219
rise of 8:225
in Scandinavian nations 8:242
and scientism 8:225
and social Darwinism 5:688–690
and social justice 9:970
in Switzerland 8:242
in United States 8:240
Eugenics, in the Americas 8:246–247
biotypology 8:248–249
birth control 8:247–248, 8:250–251
euthanasia 8:251
gender 8:249–251
homiculture 8:248
immigration 8:249
institutions, organizations, and practitioners 8:247–248
Lamarckian inheritance 8:246–247
positive and negative eugenics in Latin America 8:249–251
race 8:246–247
sterilization 8:247–251
sterilizations 8:249–251
unfit, targeting 8:249
Eugenics Education Society of London (UK) 8:218–219, 9:222
Eugenics Laboratory 17:660–661
Eugenics movement 5:182, 17:40, 20:287–288
Eugenics Protection Law 8:235
Eulenburg, Count von 9:645
Euler, Leonhard 14:372
Euler equations 8:513
Euler's theorem 21:308–309
Eupatridai 6:61–62
Euphrates–Tigris river system 25:444–445
Eurabia 25:798
EURATOM *see* European Atomic Energy Community (EURATOM)
Eureka moment 12:84
EUREKA project 24:760
Euro-American modernity 15:687
Euroasia
population change in 18:545–552, 18:546t
aging 18:550–552
contraceptive prevalence 18:547–548
demographic transition 18:546–547
diaspora populations 18:550
fertility 18:547–548, 18:548t
international migration 18:549–551, 18:551t
life expectancy 18:549t
mortality 18:548–549
projections and implications of trends 18:551–552
Euro-Atlantic civilizations of modernity 15:710
Eurobarometer (EB) 5:792–793, 5:848–849
Eurobarometer Surveys 18:494–495
Euro-Caucasian-American cultural style, and early memory development 6:760
Eurocentricity 1:306–307
Eurocentrism 3:308–309, 14:435, 16:157, 16:459, 18:339, 25:342–343
and cultural heritage 5:417
and modernity 15:693–694
Eurocommunism 4:302–303, 14:661
Eurogang Project 25:820
Eurogentest 9:912
Europe 8:261–262, 16:259
anthropology at home in 1:753–755
archaeology of 8:269–272
Barbarian Europe and Roman Empire 8:272
Bronze Age 8:270–271
from chieftain to 'state' organization in Celtic Europe 8:271–272
emergence of elites 8:270–271
emergence of farming and social complexity 8:270
Iron Age 8:271–272
Neolithic and Copper Age 8:270
Paleolithic and Mesolithic 8:269–270
balance of power in 2:346
Beck on 2:396–399
Bologna Process 2:741–746
business ethics 3:16
capitalism, effects and implications of EU accession for 18:694
Catholic South 8:259
characterization 3:873
church–state relationships in 8:262–263
clinical psychology
future trends 3:875–876
groupings 3:873–874
history 3:873–876
organization 3:875
professionalization 3:875
training 3:874–875
Cold War 4:129–130
colonizing, challenges to Central Africa 3:313–314
communist single-party states 22:140–142
post-Stalinist social sciences 22:141–142
Stalinist model 22:141
concert of 2:346–347
corporate governance 4:936
critical legal studies 5:254
cultural critique in 5:384
cultural evolution 5:395, 5:398–399
cultural policy 5:455
development of renewable energy in 18:372–373
divorce in 6:606, 6:608
early modern, Burke on 2:924, 2:926–927
empirical social research (ESR) 7:558–560
enlightenment in 7:658
ethnicity and migration in *see* Ethnicity in Europe
eugenics in 8:237–245
fertility transition 9:64–65
fertility trends in formerly socialist countries of 9:99
Habermas on 10:470, 10:472–473
health psychology 3:874
history 10:162–167
history of family policy in 8:799–800
immigration in 11:669–673
immigration into Latin America 15:462–463, 15:465
impact of family status on labor market in 8:758–759
industrial relations in 11:860–861
Islam in 8:258
laicization in 13:218–219
leisure and cultural consumption 13:837–844
literacy in 14:238–240
locations of office quarters in 3:603–604
medical training in
length of training in 7:186f
table showing variations 7:185–187
mental health organizations 15:216–219
benefits of involvement with family organization 15:218
European Network 15:216–217
function of user and family organizations 15:217
policies and family 15:216

Europe (*continued*)
relationship between family organizations and statutory services 15:217–218
self-help groups 15:218–219
user and family organizations 15:216–217
World Health Organization 15:216–217
migrants in 16:30–31
and multiculturalism 16:30–31
migration and migrants in 15:376–388
asylum seeking 15:380
economic consequences of immigration 15:383–384
effects of policy 15:382–383
effects of population 15:384–386, 15:384t, 15:385f–386f
European Economic Area 15:379–380
family reunion 15:376, 15:378–379, 15:381, 15:379f
flows according to purpose of journey 15:378–380, 15:379f
foreign citizens and immigrants in EU countries 15:380–381
future 15:381–382
gross inflow to France 15:380t
labor market 15:381, 15:385–386
migration to Denmark 15:383f
migration to Germany 15:382f
pan-European policy development 15:385
population aging 15:381, 15:383–384, 15:386
population growth 15:382, 15:385f–386f
recent trends 15:376–378, 15:377f–378f
refugees 15:382
students 15:379, 15:381
migration out of 15:457–461
destination choice and return migration 15:459
determinants of emigration 15:458
end of age of mass migration 15:461
immigrant assimilation 15:459
immigration policy 15:460–461
income distribution 15:457, 15:460
labor market impact 15:459–460
patterns 15:457
persistence and volatility 15:458–459
real wages at home and economic convergence 15:460
skills 15:457–461
modern European history
peacemaking in 17:639–640
multinational corporations in 16:77
multiplicity of ethnologies in 8:200–201
nation 16:223–225
network of independent living 6:463–464
northern *see* Northern Europe
nostalgia in 17:35
partnerships in 17:545–552
patenting in biological sciences 17:588
patterns of religion in 8:259–262
perspectives on Cold War 4:132
policy knowledge
and United States, comparison 18:297–298
populism 18:610–612
Protestant North 8:259
provincializing 18:659–660
psychotherapy 3:874
racism in 19:848–849
during Reformation and confessionalization 20:101–102
religious boundaries of 8:258–259
Second Demographic Transition in 21:346–347
sociocultural overview 8:273–279
central–eastern Europe 8:275
core and peripheral identities 8:276
cultural area 8:274
eastern Europe 8:275–276
ethnic homogenization 8:277
globalization 8:278
historical regions and present socioeconomic gaps 8:276
imagined community 8:273
mass unemployment 8:277–278
Mediterranean Europe 8:275
migration 8:274, 8:277–278
misleading and dangerous ideas 8:274
North–South gap 8:276
northwestern Europe 8:275
patrimonialism 8:275–276
postnational Europe 8:277–278
rational capitalism 8:274–275
reference societies 8:276
regionalisms 8:278
serfdom 8:275
sociocultural order of nation-states 8:276–277
southeastern Europe 8:275
use of term 8:273–274
world-system 8:274–276
southern *see* Southern Europe
universities, science and technology at 24:758–761
urban policy in 24:900
western *see* Western Europe
western, racial discrimination in 6:543–544
working classes 25:694–695
Europe and the Mediterranean, climatic variability 3:867–868
European Advertising Standards Alliance (EASA) 1:204
European Association for Evolutionary Political Economy (EAEPE) 17:194–195
European Association for Institutional Research (EAIR) 10:864, 10:867
European Association for Research on Learning and Instruction 7:284–285
European Association of Geriatric Psychiatry 10:105
European Association of Social Anthropologists (EASA) 1:756, 12:600–601
European Association of Urban Historians (EAUH) 24:816
European associations
French Academy 2:89
Institut National des Sciences et Arts 2:89
Royal Academy 2:89
Royal Society of London 2:89
Société des observateurs de l'homme 2:89
European Atomic Energy Community (EURATOM) 8:264
European Audiovisual Observatory (EAO) 5:705
European Bologna Agreement 6:610
European Central Bank (ECB) 15:727–728
European Central Bank (ECB) Survey of Professional Forecasters 8:515
European Central Bank system 21:404
European Charter of Fundamental Rights 2000 21:770
European civil law countries 14:267
European Coal and Steel Community (ECSC) 8:264
European codes 13:585
European codification 3:670
European colonialism in Australia 11:2
European Commission 2:580, 22:643–646
Community Research and Development Information Service 18:942
for the Efficiency of Justice 14:269–270
funding in social research 22:644t
political economy of renewable energy 18:370
PROGRESS program 6:465
research ethical dilemma 20:525
European Community (EC) 4:341–342, 8:264
European Community Directive on Data Protection 10:303
European Community Household Panel (ECHP) 5:846
European conquest
impacts on Australian aboriginals 2:234–235
of South America 23:44–45
European conservatism 4:689–690
European Consortium for Political Research (ECPR) 18:470, 22:670–671
European Constitutional Treaty 25:293
European construction, prehistory of 4:571–572
European contact
impact on Native Americans 1:634–635
tribal decimation and relocation/concentration 1:634–635
tribal or individual Indian 1:635
urbanization 1:635
in Oceania 17:160–162
European Convention on Human Rights (ECHR) 3:417, 4:701, 12:934, 13:440–441, 13:595, 13:720, 15:197, 18:950
European Convention on Mutual Assistance in Criminal Matters 24:599
European Cooperation in Science and Technology 22:643–644
European Court of Human Rights (ECHR; ECtHR) 1:704, 5:229, 13:431–432, 18:396, 24:444

campaign activity, regulation of 9:393, 12:510, 18:396
European Court of Justice 12:510, 12:900, 13:623
European Credits Transfer System (ECTS) 2:741
European Data Center for Work and Welfare (EDAC) 5:794
European Defense Community (EDC) 4:131
European dimensions, migration history 15:448–450
European divorce rates 6:606
European Eco-Management and Audit System (EMAS) 7:813
European Economic Area (EEA) 15:379–380
European Economic Community (EEC) 2:741, 8:264, 12:574
European Economic Community (EEC) Treaty of 1958 8:267
European Educational Research Association 7:284–285
European Election Study (EES) 5:826–828
European Enlightenment 19:591, 19:676
European Environment Agency (EEA) 5:777
European exploration and expansion 6:589–590
European fascist movements 16:260–261
structural elements 16:261
European Federation of Professional Psychologists Associations (EFPPA)
activities 3:875
establishment 3:875
membership 3:874–875
European Federation of Psychiatric Trainees (EFPT) 7:185–186
European Fertility Project 9:6–7
European Food Safety Agency (EFSA) 21:253
European Free Alliance (EFA) 10:382
European Free Trade Area (EFTA) 6:465
European General Data Protection Regulation (GDPR) 22:875
European Group for the Study of Deviance and Social Control 5:241–242
European Health Examination Surveys 5:785
European Health Information Surveys 5:785
European Health Literacy Questionnaire 10:654
European Higher Education Area (EHEA) 2:741
European higher education systems
expansion of 24:761
funding 24:759
globalization and international competition, effect of 24:759–760
mechanisms of coordination and control 24:759
national policies 24:761
reforms 24:760–761
scientification of employment and work 24:759
trends 24:760
use of technology 24:759–760
European Household Panel Survey (EHPS) 5:744–745
European impact, on American regions 11:57–58
European industrialization 11:874–876
European integration 4:571–572
welfare state 25:522
European Journal of Social Psychology 13:899–900
European Judicial Network (EJN) 24:600
European laws, in Southeast Asia 13:814
European League Against Rheumatism 20:655
European legal systems 13:443
European lexifiers 18:141
European liberalism 4:687
European Mathematical Psychology Group (EMPG) 14:805, 14:808
European Medicines Agency (EMA) 24:284
European mental space 2:639
European million-cities, skylines of 3:598f
European modernity 15:687
psychoanalysis 19:326
European modernization 18:684–685
European Network for Independent Living (ENIL) 6:463–464
European Network of (ex-)Users and Survivors of Psychiatry 15:216–219
European Organization for Nuclear Research (CERN) 7:864, 11:315, 18:104
European Parliament 8:265
green parties 10:379–380, 10:382
European Parliamentary Technology Assessment Network (EPTA) 22:873
European Patent Office (EPO) 12:269, 17:588
European perspective, of law and morality 13:479–482
European Population Explosion 10:10
European Prospective Investigation into Cancer and Nutrition (EPIC) questionnaires 10:588–589
European Psychiatric Association 18:35
European qualification framework (EQF) 7:215
European Radar Satellite (ERS-1) 20:422–423
European regional planning 20:144–145
European religion from the Atlantic to the Black Sea 8:258–263
Britain 8:261
Catholicism 8:258
church–state relationships 8:262–263
cultural heritage 8:259–260
France 8:261–262
historical legacies 8:260
immigration 8:261
Islam 8:258
new arrivals 8:261
from obligation to consumption 8:260–261
orthodoxy 8:258, 8:262
protestantism 8:259
reformation 8:258–259
Russia, church, state, and religion in 8:262–263
secular reactions 8:261
European Renaissance, Burke on 2:927
European Research Council (ERC) 22:643–645
funding in social sciences and humanities 22:645t
Scientific Misconduct Strategy 20:515
European Research Group on Military and Society (ERGOMAS) 15:515
European Research Infrastructure Consortium (ERIC) 5:796, 5:828, 5:832, 22:645–646
European Research Network on Services and Space (RESER) 21:652
The European Review of Economic History 7:58
European Roman Catholic missions 15:614–615
European Rules for Juvenile Offenders Subject to Sanctions or Measures 12:957
European scholarship, on Ancient India 11:78
European schools 2:581
European Science Foundation (ESF) 5:828, 21:313, 22:643–646, 22:670–671
European self-awareness 8:273
European Social Charter (1961) 13:720
European social science, database influence on 5:831
European social science organizations 22:648–654
characteristics 22:650t
comparison 22:648–651
creation of 22:648–651, 22:649t
hegemony 22:648–651, 22:653–654
internationalization of sciences 22:648–649, 22:654
membership, structures and activities 22:652–654, 22:652f, 22:652t–653t
philanthropic foundations 22:648–649, 22:651
scientific organizations 22:648, 22:654
European Social Survey (ESS) 5:793, 5:834, 12:673, 22:645–646
European social theory 15:493
European socialist movement 22:834
European societies, Arendt on 1:961–962
European Society for Human Reproduction and Embryology (ESHRE) 12:47
European Society for Opinion and Market Research
Code of Marketing and Social Research Practice 18:943
European Society of Human Genetics 9:913
European Sociological Review 10:279
European Sourcebook of Crime and Criminal Justice Statistics 5:165
European Space Agency (ESA) 14:45, 20:422–423
European Standard Set (ESS) 9:342
European Standard Set of Loci (ESSOL) 9:958–959
European state formation 23:347–348

European statistical office (EUROSTAT) 5:762, 5:776, 5:779, 5:795, 5:801, 14:760–761, 22:116
National Accounts 5:749
European Survey on Income and Living Conditions (EU-SILC) 5:846
European Sustainable Development Strategy 7:771
European System of Integrated Economic Accounts (ESA) 5:753
European Time Use Survey 13:855
European unemployment system 12:827
European Union (EU) 4:341–342, 5:457, 10:854, 12:369–370, 13:504, 13:506, 13:616, 14:709, 17:633, 23:366, 24:164–165, 24:223, 24:588–589, 25:839
accession, and economic reform 18:693–694
effects and implications for capitalism in Europe 18:694
impact of financial crisis 18:694–695
basic freedoms 8:267
characterization 3:873
clinical psychology training, harmonization 3:875
compliance aspects 8:266
countries, foreign citizens and immigrants in 15:380–381
Directive on the Protection of Personal Data 12:596–597
Directives on lawyer mobility 13:620–621
and Eastern European Studies 6:805–806, 6:808
Equal Treatment Directive against sex discrimination 7:927
Federation of Family Associations of Mentally Ill People 15:216
formation of 8:264
fundamental rights 8:267
institutions 8:265
law and policy 1:248
legislative powers 8:265
membership
expansion 3:874
postsocialist societies 18:699–700
nation-state citizenship 3:658
policies and forms of intergovernmental cooperation 8:264
population momentum in 18:577
principles of the internal market 8:266–267
Racial Equality Directive, and the Constitution of South Africa 6:534
regulatory agencies 20:181, 20:184–185
research and teaching infrastructure 22:667–673
resourcing and funding, social and behavioral research 22:643–647
changing role of social sciences 22:644–645
Council of European Social Science Data Archives (CESSDA) 22:645–646
European Commission 22:643–646, 22:644t
European Cooperation in Science and Technology 22:643–644
European Research Council (ERC) 22:643–645, 22:645t
European Science Foundation (ESF) 22:643–646
European Social Survey (ESS) 22:645–646
Framework Programmes 22:643–646
funders, programs, and aims 22:645–646
history of 22:643–644
Horizon 2020 22:643–646, 22:643t
impact 22:646
Survey of Health, Ageing and Retirement in Europe (SHARE) 22:645–646
Vilnius Declaration 22:646
rule of law 8:264–265
and social work 22:795, 22:797–798
sources of law 8:265–266
statistical system of culture 5:730
transnational organized crime 24:597–600
Treaty of Maastricht 24:801
welfare retrenchment 25:522, 25:525
European Union law 8:264–268
European Values Study (EVS) Group 5:791–792, 5:834
creation of 5:791
evolution of 5:791–792
European Values Survey 18:494–495
European Zionist movement 16:363
European-American parenting 22:159
Europeanism, Sarmiento and 20:910
Europeanization 22:795
of historiography 11:65–66
parliamentary democracy in 17:514
European-wide panel studies 5:846–847
EuroQol (EQ-5D) 10:659, 19:298
EUROSTAT *see* European statistical office (EUROSTAT)
EU-SILC *see* European Survey on Income and Living Conditions (EU-SILC)
Eusocial species 13:48
Eusociality 13:26, 22:906
evolution of 13:48
kin selection and 13:30–31
Eustress 25:738
Eustress–distress model 17:135
EUT *see* Expected utility theory (EUT)
Euthanasia 7:594, 7:598, 7:600, 8:280–284
active euthanasia 8:281–284
aid in dying 8:282
ethical problems 8:282–283
ethics 2:615
eugenic 8:251
in Germany 8:282–284
good dying 8:280
Hippocratic oath 8:280
in the Netherlands 8:281, 8:283–284
physician-assisted suicide 8:281, 8:283–284
social science research on 8:283–284
in Switzerland 8:281, 8:283–284
types of 8:280–282
use of term 8:280
voluntary active euthanasia 8:281, 8:283–284
Euthanasia exterior 8:280
Euthanasia interior 8:280
Euthanasia operation 11:152
Euthanasia talk 7:600, 7:600b
Eutrophication, Anthropocene 1:724
Evaluability assessment, logic models 14:302–303
Evaluation 23:239
of analogs, in EITMs 18:456
axiological neutrality 16:757–758
cultural mediators's role 5:442
in educational planning 7:262, 7:264
Evaluation, potency, activity *see* EPA (evaluation, potency, activity) structure
Evaluation Accountability Standards 19:119
Evaluation apprehension 8:573
Evaluation phase, ADDIE 12:198
Evaluation research
cost-benefit and cost effectiveness analyses in 5:55–62
Evaluative attitudes (values) 25:42
Evaluative conditioning (EC), and attitude 2:169–170
Evangelical Christianity 8:285–288
Evangelical movement 10:492
Evangelical Protestantism 1:600
Evangelicalism 8:287–288, 10:491
Evangelicals 3:541, 20:280–281
Evans, David T., on sexual citizenship 21:730–731
Evans, G., works, *Mass Atrocity Crimes* 10:7
Evans, John H. 2:632
Evans-Pritchard, Sir Edward E. 1:681–682, 6:590, 7:376, 8:289–294, 13:71, 14:435–437
anthropology studies of 8:290–292
Azande kingdoms, study of 8:290–291
background 8:289
British academic life 8:289
career outline of 8:289
early influences of 8:289–290
education of 8:289
legacy of 8:293
in Oxford
British social anthropology 8:291
development of social anthropology 8:291–292
on relativism 20:235
review works 8:289
on style of anthropology 8:292
Sudan researches 8:289–291
works
Kinship and Marriage among the Nuer 8:290–291
The Nuer 1:767
The Nuer: A Description of the Modes of Livelihood and Political Institutions of a Nilotic People 8:290–291
Nuer Religion 8:291, 20:860
Principles of Art 8:290
Sanusi of Cyrenaica 8:291

Witchcraft, Oracles and Magic among the Azande 2:495, 8:290, 20:307
Evapotranspiration 25:437–438
Evasion, tax 15:738
Evasion of income tax 11:746–747
EVD *see* Eigenvalue decomposition (EVD) of covariance matrix
Event counts analysis, in political science 18:515
Event dating, questionnaires 19:773–774
Event history, family reconstitution 6:139–142
Event history analysis 8:295–309, 8:321, 14:51–58, 23:787–788
basic terminology 14:54
causal approach to interdependent systems 8:309
censoring 8:306–307, 14:54–55
childbearing and employment among women 8:295–296
notation and key functions 8:295–296
childhood smoking prevention 8:300–301
in continuous time 8:310
bone marrow transplantation 8:313, 8:313f
competing risks model 8:312, 8:311f
conditioning in multistate models 8:317
constant and piecewise constant transition intensities 8:314
counting process representation, likelihood 8:313
Cox regression model 8:315–316
delayed entry and right-censoring 8:317
freely varying (nonparametric) transition intensities 8:314–315
hypothetical calculations in multistate models 8:315–316
illness–death model 8:312, 8:312f
interaction between life–history events 8:312–313, 8:313f
interval censoring 8:317
Markov processes 8:313–315
multistate models 8:311–313
observational patterns 8:317
other parametric models for the transition intensities 8:314
partial model specification 8:316
semi- and nonparametric regression Markov models 8:315
state occupation probability 8:310–311, 8:315–316
statistical model specification 8:313–317
survival data 8:310–311, 8:310f–311f
time-dependent confounding 8:316–317, 8:316f
transition probability 8:310–316
two-state model for survival data 8:312
continuous-time, discrete-state processes 8:305
dependent variable 14:55
different observation plans 8:305–306
in discrete time 14:351–353
discrete-time hazard 14:351
embedding into continuous-time models 14:353
regression models and likelihood inference 14:351–352
structure of likelihood 14:352–353
effects of pregnancy on entry into marriage for couples living in consensual unions 8:309
hospital readmission among individuals with affective disorder 8:299–300
Markov process 8:296
multilevel analysis of life-course, cohort, and period effects in career mobility 8:308
multistate models 8:298–299
overview 8:295
parallel and interdependent processes 8:307–308
parallel processes at the level of the individual 8:308–309
recidivism among criminal offenders 8:295
recurrent events
analysis 8:296–297
multitype 8:297–298
remarks 8:301–303
semi-Markov process 8:298
statistical models 8:307, 14:55–56
terminology 8:306
time-varying covariates 8:307–309
transition rate 8:307
unobserved heterogeneity 8:308
women's rising educational and career investments, and changes in process of family formation 8:308–309
Event history calendars 5:737
questionnaires 19:774
Event nominalizations 23:891–892
Event occurrence in discrete-time data 23:785–787
Event questions 23:764–767, 23:765f
Event sequence analysis 8:320–322
event history methods 8:321
exploratory technique 8:320
future developments 8:322
methodological issues 8:321–322
minimal edit distance 8:321
origins 8:320–321
pattern-matching technique 8:321–322
standard causal methodologies 8:321
Event-based measurement scales 24:100
Event-driven media coverage of climate change 3:856
Event-history models, in political science 18:518–519
Event/interval calculus 13:103
Event-related brain potentials (ERPs) 8:681
Event-related designs 13:368–369
Event-related evoked potentials 11:470
Event-related fields (ERFs) 9:479
Event-Related Number Space Orientation (ENSO) 14:778–779, 14:779f
Event-related phenomena 9:479
Event-related potential (ERP) technique 9:479–481, 9:480f–481f
in bilingualism and second language research 9:486–487
in language development 9:484–486
components 9:488f
in pathological and special populations 9:487
nonliteral language addressing literal statements 9:484
responses reflecting aspects of language processing 9:481–483
study of figurative language 9:484
to target words illustrating N400 effects 9:484f–485f
Event-related potentials (ERPs) 2:164, 2:166–167, 10:703, 12:305, 16:673, 20:617–618
infant cognitive development 4:60–63
for inferences 12:38, 12:41
language development 16:708, 16:710
source monitoring 23:32–33
Events
death as 5:871
human–environment relationships 11:407
media 15:48–54
nationalism as 16:295
perceptual learning 10:129–130
TAME 24:211–212
triggering of emotions by 7:526
Event-structure analysis (ESA) 19:698
Every Child Matters agenda in England 8:615
Everyday concepts, Vygotsky on 25:312–314
Everyday interactions, qualitative developmental psychology 19:668, 19:673–674
Everyday knowledge 13:93
Everyday life
anthropology of 8:323–326
anthropological approaches 8:323–324
cultural creativity 8:324
problems and potentials 8:324–325
sociological approaches 8:323–324
study 8:323
traditions 8:323–324
undercurrents 8:325
and drinking 6:659, 6:661
and law 13:728, 13:730–731
legal scholarship for 13:730
sustainable practices in 23:834, 23:836
Everyday life and law 13:468, 13:542–545
cultural and symbolic implications of resorting to law 13:470
fallibility and vulnerability of law 13:472
law as action and experience 13:468
law as intervening in social life 13:469–470
law first orientation 13:468–469
law in everyday life; everyday life in law 13:470–472
legal realism 13:468–469
Everyday nationalism
interpretive studies 16:310–311
survey-based studies 16:311–312
Everyday practices, citizenship 3:641
Eviction rights *versus* security of tenure 11:244
Evidence (health information systems) 5:88–89, 5:781–782, 11:820–821

Evidence (health information systems) (*continued*)
common sense 4:2–4
counterfactuals and 5:91
scientific evidence and 21:298–300
torture and 24:442–444, 24:446
Evidence Act 23:61
Evidence-based behavioral medicine (EBBM) 8:333
Evidence-based behavioral practice (EBBP) 8:333
Evidence-based community-level strategies, for physical activity 18:95, 18:95t
Evidence-based crime policy 20:199–200
Evidence-based interventions, race and ethnicity 19:815
Evidence-based medicine (EBM) 8:327–331, 8:333, 9:719–720, 15:266, 19:705
application in clinical practice 8:329–330
critical appraisal of 8:329
critique, limitations, and debates 8:330–331
definition of 8:327
formulation of precise clinical questions 8:328
identification of best available evidence 8:328–329
key components 8:328f
levels of evidence 8:328–329, 8:329f
need for 8:327
outcomes 8:327–328
patient-oriented research 8:327–328
qualitative health research in era of 10:695
randomized controlled trials 8:328–331
resources and tools 8:330
Cochrane Library 8:330
other resources 8:330
systematic reviews and meta-analyses 8:330
Evidence-based policy 5:236, 24:295, 24:298
Evidence-based policy making 18:281–287, 19:570–572
challenges and criticisms 18:285–286
complex social problems 18:286
evaluation of 'welfare-to-work' programs 18:286
evidence-informed policy 18:286–287
factors enhancing use of evidence 18:284–285
roles and relationships 18:284–285
forms of evidence for policy 18:282–284
perspectives on evidence reliability 18:282–283
types and sources of evidence 18:283–284
historical development and overview 18:281–282
policy capacity 18:281–282, 18:284
policy decision making 18:281–282, 18:286–287
policy development 18:281–282
policy evaluation 18:282–284
policy knowledge 18:281–285
policy learning 18:285
policy skills 18:281–282, 18:284, 18:286
program evaluation 18:282–284
program management 18:284, 18:286
program outcomes 18:282–283, 18:286
randomized controlled trials (RCTs) 18:282–283
research evidence 18:286
Evidence-based policy movement 18:290
Evidence-based practice (EBP) 2:419, 5:58, 8:332–338, 8:346, 19:90, 19:93, 23:215, 25:767–771
in applied research (criminology and criminal justice) 5:129
best research evidence 8:333–335
patient characteristics 8:334
resources 8:334
role of health professionals 8:334–335
steps 8:335
clinicians 8:332–337
competencies 8:336
acquire 8:336
analyze and adjust 8:336
apply 8:336
appraise 8:336
ask 8:336
definition 8:332
EBBM to EBBP and EBP 8:332–333
EBM to EBBM 8:333
eclecticism 19:511
emerging directions 8:337
history 8:332–333
and integration 19:512–513
interprofessional model 8:333
methodological and practical challenges 8:336
randomized controlled trials 8:332–334, 8:336
reflexivity 20:82–87
risk and uncertainty 20:690
sexual dysfunction treatment 21:691–692
shared decision-making 8:333–334, 8:336
systematic review 8:333–337
writing 25:767–772
Evidence-based program movement 11:695
Evidence-based psychiatry 8:339–343
classification 8:339–341
diagnosis 8:339–341
epidemiology 8:340–342
evidence in
hierarchy 8:339t
improvement 8:341
experience of mental illness to different degrees 8:340–341
externalizing 8:340
internalizing 8:340
latent variable technique 8:340–342
mental disorders 8:340
experience in different ways 8:341
randomized controlled trials 8:339–342
Evidence-based psychotherapy (EBP) 8:344–350, 11:612–614, 11:618
methodological and conceptual issues 8:346–349
conflict of interest 8:348
evolution of EBP 8:348–349
interpretation of results 8:348
outcome criteria, type of 8:347
recruitment and selection into treatment 8:346–347
research designs 8:346
statistical analysis of results 8:348
psychotherapy research 8:344–345
consumer satisfaction 8:345
efficacy/effectiveness research 8:345
efficiency research 8:345
meta-analysis 8:345–346
process research 8:345
research on transfer into EBP 8:346
short history of 8:344–345
versus treatment-as-usual (TAU) 8:349
Evidence-based treatment
interpersonal psychotherapy 12:634–635
psychological treatments 19:384
for trauma 16:198
Evidence-driven jury 12:911
Evidence-informed policy 18:286–287
Evidence-informed policy making, for health statistics 5:785
Evidential decision theory 5:97
Evidentiality (grammar) 24:213
Evidentials in linguistics 8:351–356
deixis 8:351
evidentiality 8:351, 8:355
grammaticalization 8:354
lexical functional grammar 8:352
linguistic anthropology 8:351
linguistic typology 8:355
pragmatics 8:351
semantics 8:351, 8:353
tense and aspect 8:352
verbal category 8:351
E-Views 7:2, 7:4
Evil charismatic leaders 3:356
Evil cult 16:806–807
Evocation model of cultural symbols 5:579–580
Evo–devo (evolutionary–developmental biology) 8:473
Evoked potential (EPs), language development 16:708
Evoking 15:911
Évolué 3:313
Evolution 4:416–419, 7:467–470, 8:382–383, 8:385, 8:399–405, 8:469–470, 8:473, 13:227–229, 16:330–331, 20:313, 22:908–911, 22:991, 23:558–559
of adoption 1:143–144
of behavior 12:175–176, 14:367
approach 21:612
biological *versus* cultural 1:85
of body *see* Body; evolution of
concept of 7:836–837
of cooperation 11:259
and crime 2:672
cultural *see* cultural evolution
Darwinism 5:681–684
of economic statistics as system 5:761
applications of business registers 5:761

censuses and registers 5:761
ecosystem 17:367
and emotions 7:527–528, 7:530
of empathy 7:541–542
in the enlightenment 8:384
environmental determinism 7:722–723
and environment-development 7:810
epigenetic inheritance 7:832, 7:836–837
and eugenics 8:253–254
and food preferences 9:296, 9:298
genetic programming 2:44
human 17:458–459
and human ecology 11:283
in institutional economics 17:191, 17:194
and language *see* cultural transmission
of lateralization in vertebrates 2:801–802
life history theory (LHT) 9:40, 9:42–44
of marriage systems 14:611–619
microevolution 15:338
models of personality 17:899–905
of modern human cognition 11:252–258
of morality 7:469–471
Muller's contribution to 16:14
natural and social 8:388–391
adaptive social networks 8:390
cultural evolution 8:389–390
fitness 8:388–389
imitation 8:389
learning 8:389–390
local interaction 8:390
population structure 8:389
replicator 8:388–389
nonadaptive 8:383
within organization 12:72
organizational 17:367
perspectives of life span development 14:108
of political complexity 4:420–421
of primates 9:941, 18:891, 18:893
primitive society 18:911–914
religious *see* Religious evolution
and ritual 20:719
and social behavior 8:464–465
social Darwinism 5:686–687, 5:689–690
and social networks 22:475
societal 8:466
theories of sexuality 21:846–847
Wright's contribution 25:759
Evolution, history of 8:382–387
acquired characteristics, inheritance of 8:383–385
Darwinism in biology 8:385
history of life 8:385
natural selection 8:385
Darwin's theory, development of 8:384–385
Malthus and population 8:384–385
scientific influences 8:384
fossil record 8:384
human origins and social evolution 8:386
cultural evolutionism 8:386
social Darwinism 8:386
models of evolution 8:382–383
adaptation 8:383
Darwinism and Lamarckism 8:383
evolutionary trees 8:382–383
ladder of progress 8:382
nonadaptive evolution 8:383
modern Darwinism 8:386–387
later developments 8:387
modern Darwinian synthesis 8:386
natural selection 8:385
pre-Darwinian evolutionism 8:383–384
creation and design 8:383–384
evolution in the enlightenment 8:384
fossil record 8:384
Evolution origin of control behavior 4:813–814
Evolutional economic geography 7:64
Evolutionary algorithms, genetic programming 2:41–42
Evolutionary analogy, Skinner on 22:26
Evolutionary anthropology
gene-culture coevolution 9:814–815
of Korea 2:71
and wealth flows 9:35, 9:37
Evolutionary approaches, to cognition and development 8:392–394
academic development 8:394–395
brain mechanisms 8:395
cognitive mechanisms 8:395
implications for instructional activities 8:396–397
motivation to learn 8:395–396
folk biology and physics 8:393
folk psychology 8:392
heuristics and attributional biases 8:393–394
implications for education 8:394–397
social learning and 8:396
Evolutionary approaches, to prosocial behavior 1:580
Evolutionary basis of religion 20:360
Evolutionary biology 8:465
in early twentieth century 8:407
incorporation into demography 8:410–411
and institutionalism 12:183–184
Evolutionary computation 2:30
Evolutionary critical realism 20:24
Evolutionary demography 8:406–419
adaptation 8:417
biodemography 8:413, 8:418
comparative demography 8:418–419
across species 8:418
measures, methods, and data 8:418
widely occurring traits 8:418–419
demography beyond humans 8:417–418
application of demography to nonhumans 8:417
challenges in nonhuman demography 8:417
methods for nonhuman populations 8:417–418
divergence after World War II 8:407–408
in early twentieth century 8:407
fertility 8:407–411
fitness 8:413–417
human-demographic traits 8:414–417
optimality and constraints 8:414–416
shared by all human populations 8:416
variation among humans 8:416–417
incorporation of evolutionary biology into 8:410–411
mortality 8:407–410
mutation 8:413–417
natural selection 8:413–415
postreproductive life span 8:416, 8:418
renaissance 8:408
quantification 8:408–409
reason 8:409–410
reproduction 8:413–418
survival 8:413, 8:415, 8:417
theories of Darwin, Wallace, Malthus 8:406–407
twin population processes 8:413–414
evolution as demography 8:413
evolution of demography 8:413–414
shared concepts and focus 8:413
what evolutionists can learn from demographers 8:411
Evolutionary developmental psychology 8:420–429
adaptations 8:420–424, 8:427
and antisocial behavior 8:427
basic principles of 8:421–423
biological sensitivity to context 8:424
childhood 8:420–424, 8:426–427
conditional adaptation 8:422–423, 8:426
deferred adaptation 8:422–423
development and evolution 8:420–421
development constraining by both genetic and environmental factors 8:421–422
differential susceptibility to experience 8:425
environmental unpredictability 8:426
extended childhood 8:422
gene–environment interactions 8:421
life history theory 8:426
mental health
application to 8:423–427
evolutionary developmental psychological models of 8:424–425
model 8:423–424
mental health model 8:421, 8:423–426
natural selection 8:422–423
ontogenetic adaptation 8:422–423
risk-taking behaviors 8:425–426
and risky adolescent behavior 8:425–427
evolutionary life history framework 8:426–427
Evolutionary developments 20:357
and aging 1:435
Evolutionary differentiation 6:392–393
Evolutionary ecology 7:792, 8:411
and hunter-gatherers 11:425
Evolutionary economic geography (EEG) 7:35
goal of 7:35
Evolutionary economics 11:567
Evolutionary economics, evolutionary concepts application in 8:430–436
analogies to biological sphere and metaphorical use of 8:430–432
Darwinism, generalized 8:431–432
Metcalfes use of replicator dynamic 8:431

Evolutionary economics, evolutionary concepts application in (*continued*)
Nelson and Winters concept of organizational routines 8:430–431
biological evolution 8:430–435
competition 8:430–431
cultural evolution 8:431, 8:433–435
human behavior 8:430, 8:432, 8:434–435
naturalistic interpretation 8:432–435
American institutionalism 8:432–433
continuity hypothesis 8:434–435
evolution of institutions: Hayek and North 8:433
indirect evolutionary approach and game theory 8:434
thermodynamics 8:433–434
neo-schumpeterians 8:430
Evolutionary educational psychology 8:392
Evolutionary epistemology 3:71, 8:437–440, 14:367–368
gene-based evolutionary theory 8:437–438
individual learning 8:438
memetics 8:438–439
selection analysis 8:439–440
social learning 8:438
Evolutionary equality 23:394
Evolutionary explanations 8:594
of ethnocentrism and prejudice 8:169–170
Evolutionary game theory 5:405, 19:903–904
Evolutionary group dynamics 8:441–447
adaptation 8:442, 8:444–446
evolutionary psychology 8:441, 8:443
group living 8:441, 8:443
key adaptive problems in group life 8:441–445
competitive altruism 8:443–444
cooperation norms 8:443
despotism *versus* democracy 8:444
distribution norms 8:443
fear of strangers 8:445
following group norms 8:442
group cohesion 8:444
group coordination 8:441–442
group decision making 8:444–445
intergroup aggression and warfare 8:445
intergroup relations 8:445
leadership 8:442
religion, music, and dance 8:444
robust beauty of the majority rule 8:444–445
signaling and status emotions 8:443
social exchange 8:442–443
social identity 8:444
status 8:443–444
structural and social focal points 8:442
transactive memory 8:442
promise of 8:445–446
psychological processes 8:445–446
social brain 8:441
Evolutionary medicine (EM) 8:448–459
applications 8:454–459
antibiotic management in hospitals 8:454–455
birth and parenting 8:457–458
cancer therapy 8:455–457
emergency medicine and intensive care 8:457
helminths and Crohn's disease 8:457
public health and health behavior change 8:458–459
conceptual tools and medical relevance 8:449–453
environmental mismatch and genetic quirks 8:452
historical legacies and 'scars' of evolution 8:452–453
host–pathogen coevolution and genetic conflict 8:449–451
life-history trade-offs and aging 8:451–452
criticisms 8:453–454
Darwinism 5:683–684
and Darwinism 5:683–684
Evolutionary models 24:470
Evolutionary origin of language 8:364–365
Evolutionary paradigm, oral and literate expression 17:275
Evolutionary perspective 8:394–395
of neonatal crying 16:480
Evolutionary Prisoner's Dilemma 18:936
Evolutionary processes 11:289
Evolutionary programming 2:42
genetic programming 2:41–42
Evolutionary psychology 1:746, 8:151, 8:400–403, 8:463, 17:909, 18:912–913, 19:416, 21:679, 22:892, 22:895, 22:904, 22:985
and biosocial criminology 2:672
cognition, evolution of 4:1, 4:4
comparative evolutionary psychology 4:394–395, 4:398
Dark Triad 5:673–674
and emotions 7:469
and irrationality 12:721–723
and religion 20:389
of socialization and education 22:862
Evolutionary socialism 22:834
Evolutionary sociology 8:463–468, 18:628, 22:895–898
biology 8:463–465
dominance and social inequality 22:897–898
emotions and social behavior 8:464
evolution and social behavior 8:464–465
field background 8:463–464
gender and gender inequality 22:896–897
genes and social behavior 8:465–466
hormones and social behavior 8:466
mating and reproduction 22:896
neurosociology 8:464
social capital 22:897
societal evolution 8:466
Evolutionary strategies, genetic programming 2:41
Evolutionary synthesis 22:901
Evolutionary synthesis in biology 11:306
Evolutionary theories 1:237, 2:482–485, 7:621, 9:425–426, 13:44, 13:48–50, 15:794, 17:366–367, 19:887–888
Fisher's 9:223
of language development 13:319–323
in medicine 8:453
moral development 15:793–794
Muller's contribution to 16:14
of prosocial behavior 1:580
and religion 20:384–390, 20:393
on sexual strategies 21:683t
structure of, 'evo–devo 8:469–475
confirmation of evolution 8:469
confirming and testing models 8:473–475
Darwinian 8:469
evo–devo 8:473
evolutionary-developmental biology 8:473
expansions and alternative views of 8:472–473
falsification of evolution 8:473
genetics 8:470–471, 8:473–474
natural selection 8:469, 8:471–473
population genetic 8:470–472
quantitative genetics 8:471
random processes 8:471
selection theory 8:472
Evolutionary theorizing 22:955–956
Darwinian revival 22:955–956
stage models revival 22:955
Evolutionary trees 8:382–383
Evolutionary–developmental biology (evo–devo) 8:473
Evolutionism 5:677, 8:198–199, 11:93–94, 14:453, 14:671, 24:583
and anthropological comparison 4:411–412
cultural 8:386
and diffusion 6:402
in Korea 2:71
and magic 14:435–436
and morality 15:806
nineteenth century 8:810
pre-Darwinian 8:383–384
tribes 24:646
Victorian 1:766
Evolutionist political philosophy 13:231
Evolved capacities, of heuristics 10:832
Evolved traits
comparative evolutionary psychology 4:394, 4:398
EVS *see* European Values Study (EVS) Group
EVT *see* Expectancy value theory (EVT)
EWAS *see* Epigenome-wide association studies (EWAS)
E-waste *see* Electronic waste
EWC *see* East-West Center (EWC)
Ewe (language), fixed expressions compared with English 13:290
Ewick, P., works, *The Common Place of Law* 13:469
EWO *see* Education Welfare Officer (EWO)
eWOM *see* Electronic Word-of-Mouth (eWOM)

EWS *see* Early warning systems (EWS)
Ex ante modes of peer review 17:687
Ex ante moral hazard 12:212
Ex ante perspective, of value of life 25:15–16
Ex improviso speech 14:244
Exact logistic regression 2:593–594
Exact matching 19:143
Exact methods 19:432
  clustering 19:433–434
  seriation/scaling 19:432–433
  variable selection 19:434
Exact optimal designs 17:257–258
Exact *p*-values 19:432
Exact replication 20:441
Exaflood of big data 3:619
Ex-ante harmonization 5:795
Exaptations 9:522–523
Excavation
  in archaeology 23:745
  of human remains 9:335
Exception 1:823
  in organizational life 6:299
Exceptionalism, and regional geography 20:123–125
Exceptionalities 23:213–214, 23:217
Excessive demagogy 1:982
Excessive neonatal crying 16:476
Excessive oligarchy 1:982
Excessive police force, determining 18:255–256
Exchange 7:12
  in anthropology 8:476–481
    alienation 8:476–480
    barter 8:478, 8:480
    capital 8:477
    commodities 8:476–480
    concept 8:476–477
    economic anthropology 8:476–480
    examples from corpus 8:477–478
    gift economy 8:476, 8:479
    informal economy 8:480
    money 8:476, 8:478, 8:480
    new directions 8:478–480
    property 8:477, 8:480
    self-interest 8:476–477
    trade 8:477–480
    value 8:476–477, 8:479–480
  archaeology of 24:481–484
    central-place exchange 24:483
    characterization studies 24:481, 24:483
    current theory and research 24:482–483
    distribution patterns 24:483
    down-the-line exchange 24:483
    electron microscopy 24:481
    formalist approach 24:482
    future directions 24:483–484
    identification 24:482–483
    inductively coupled plasma-mass spectrometry 24:481–482
    neutron activation analysis 24:481
    obsidian 24:481–484
    substantivist school 24:482
    technological developments 24:481–482
    theoretical developments 24:482
    x-ray fluorescence spectrometry 24:481
  collaborative consumption 4:139–140, 4:142–143
  contours and consequences of 12:387–388
  and division o labor 6:603
  frontiers 9:464
  gift as 10:142
  international social work student exchanges 12:562–567
  justice of 12:928–931
  and marriage 14:599–601
  and money 15:746–749
  potlatch *see* Potlatch
  and production 12:331–332
  of products 3:376, 3:379
  and reproduction 20:450–451
  social 8:482–488
  strategies 17:604
Exchange network theories 8:484, 22:958
Exchange networks 15:668
  power in 15:667–671
Exchange relations, Emerson's theory of 8:484
Exchange scholars 18:467
Exchange systems
  in markets 14:559
Exchange theories 12:403, 22:993
  for ethnic intermarriage 8:119–120
  of generalized media 9:538
  groups 10:438
  of interviewing 5:853–854
  kula ring 13:139–140
  network analysis 16:520, 16:522
  and time use 24:370–371
Exchange theorizing 22:957–958
  convergence exchange theories 22:958
  exchange-network theories 22:958
  rational choice theories 22:957–958
Exchange tie 15:668
Exchange traded funds (ETFs) 23:483, 23:485
Exchangeability theorem 5:857–859
Exchange-resistance theory 8:484
Excision 10:195
Excitation 12:416–418
Excitation transfer theory (ETT) 1:374–375
  for aggression 1:374–375
Excitatory cortical neurons 25:161
Excitatory postsynaptic potential (EPSP) 14:358–359
Excitatory synaptic input 16:704
Excludability, international public goods 19:540–541
Excluded-from-School (EFS) children 19:303–304
Exclusion 7:118
  borders and 2:762–763
  gender and technology 9:723–724
  language policy for 13:357–358
  and power 15:668
  social *see* Social exclusion
Exclusion memory paradigm, task network models 16:543–544
Exclusion principle 3:262–263
Exclusion restriction (ER) 8:529
  assumption 4:446–447
  sensitivity analysis for 4:452
Exclusionary ideologies, xenonphobia 25:796–797, 25:800
Exclusive breastfeeding (EBF) 2:847–848, 2:848t
Exclusive Brethren 16:805–806
Exclusive disjunction 19:62
Exclusive inequality 1:9
Exclusive legal positivism 18:26–27
Exclusive nonpositivism 13:479–480
Exclusive positions, of personal identity 17:743
Exclusive positivism 13:479
Exclusiveness, conscious and unconscious knowledge 13:86
Exclusivity 21:815–816
  of communities 4:315
Excoriation disorder 17:113
Exculpation 9:957
Exculpatory defenses 5:220
  that do not negate *actus reus* and *mens rea* 5:223–225
    excuses 5:224–225
    justifications 5:223–224
Excuses 5:224–225
  for duress 5:224
  exculpatory defenses and 5:220
  for insanity 5:224
  for provocation (extreme emotional disturbance) 5:224–225
Executions 5:874–878
  in France 19:638f
Executive compensation 17:614–615
  agency theory 17:615
  social comparison theory 17:615
  tournament theory 17:615
Executive compensation, stock market 23:493
Executive control
  and aggression 1:377–378
  and intelligence 12:316–317
  processes 13:714–715
  structures 16:504
Executive dysfunctions 1:653–654
  and confabulation 4:562
  neuropsychological rehabilitation interventions for 2:816
Executive functioning
  cognitive styles and learning styles 4:112
Executive functions (EFs) 8:489, 9:486, 14:309, 17:873, 18:817, 25:715
  assessment of 16:724
  changes in concept of 8:489
  changes in studies of 8:489–490
    behavioral measures of 8:489–490
  cognitive development 4:40
  cognitive processes of 21:554
  cognitive styles and learning styles 4:112
  development of 8:490–494
    in early childhood 8:492–494
    in infants and toddlers 8:490–492
    in middle childhood 8:494
  and educational neuroscience 16:733
  and future thinking 9:555–556, 9:558
  role of prefrontal cortex in 4:33

Executive functions tasks 3:571
Executive government 8:497
  cabinet systems 8:497–501
  capacity, output, and performance of 8:500–501
  coalition 8:498–500
  divided government 8:500
  governance 8:498–500
  legislation 8:497, 8:499
  minister 8:497–501
  organization and functioning of 8:498–500
  origin of 8:497
  parliamentary systems 8:497, 8:499–500
  political parties 8:497–498
  presidential systems 8:497, 8:500
  prime minister 8:497–501
  semi-presidential systems 8:497–500
  separation of powers 8:497
  stability 8:500–501
  study of 8:498
  types of 8:497–498
Executive government, output of 8:500–501
Executive management, stock market 23:493
Executive Orders 13:505–506
  Executive Order 12898 7:758
  Executive Order 11 246 1:256
Executive oversight, and administrative law 1:112–113
Executive processes
  monitoring 25:705
  prefrontal cortical areas 25:707, 25:709
  working memory tasks, impairments 25:703
Executive stage of aging 1:436
Executives 14:471–473, 14:477
  and judicial behavior 12:888
Exegesis 16:806
Exegetical explanation, legal reasoning 13:777–778
Exemplar theories, concept learning 4:538–539
Exemplar-based model, of social categorization 22:187
Exemplars, in functionalist grammar 9:474
Exemplary exhibition 24:882, 24:887
Exemplary experiment 8:566
Exercise 10:583, 18:92–94, 18:97
  in cardiac rehabilitation 20:190
  and enriched environment 2:837
  and epigenetics 7:844
  and obesity 17:82–83
  religion and 20:266
  and sleep 22:60
  and time use 24:380
Exercise training
  for asthma 20:579
  for chronic obstructive pulmonary disease 20:579
Exergy 7:615–616, 7:627
Exergy efficiency, technoeconomic concepts of 7:630
Ex-feudi 14:423
Exhaustion 21:59, 25:537
Exhaustive control 4:828
Exhaustive processing 12:78
Exhaustively determinable 25:210
Exhaustiveness, conscious and unconscious knowledge 13:86
Exhibitionism 17:951–952
Exhibitionist society, and visual sociology 25:192–193
Exhibitions
  entertainment 7:672–673
  in urban planning 24:885–887
Exile
  in MENA 16:365
  Shklar on 21:893–894
ExIm Bank 3:505
Existence
  of equilibrium 7:915–917
  in social ontology 22:500, 22:502–504
Existence value, cost–benefit analysis 5:49
Existential copula constructions 18:631
Existential deterrence 6:247–248
Existential psychology 8:517
Existential quantifiers 14:295, 19:707–708
Existential self 21:469
Existentialism 10:85, 19:411, 19:491, 19:493, 22:817, 23:857–858
  and individuality 17:933–935
Existentialist feminism 9:738
Existentialist philosophy, Gorz on 10:286–288, 10:290
Exocannibalism 3:85
Exocentric structures 23:888
Exodus 5:77
Exogamy 11:730, 18:842
  minorities 15:570
Exogenous deaths 12:17
Exogenous languages 1:107
Exogenous noise generation, stochastic optimization for 23:474
Exogenous specialization 6:602
Exogenous variables 7:612, 21:980
Exonerations, and wrongful convictions 25:787–790
Exophora 1:674
Exosystem 6:840, 21:682–684
  families as educational settings 8:732
"Exotic becomes erotic" model 3:890
Exotic mortgages 3:580
Exotic 'other&rsquo 19:677
Expanded pidgins 18:133–134
Expanded version of commodity chains 4:244–245
Expansion 11:686
  dispute 6:562
  frontiers 9:465–466, 9:468
  Ottoman 9:468
  Qing and Romanov Empires 9:467
  Russia's territorial 9:466
Expansion consistency, systematic violations of 19:444
Expansion of morbidity 4:464
Expansive framing, and transfer of learning 21:99
Expansive goals 21:560
Expansive potential, and Internet 12:603
Expansive realization, and Internet 12:603
Expectancies 15:928
  constructs relating to
    attribution 21:50–51
    intelligence, implicit theories of 21:50
    self-concept 21:49–50
    self-efficacy 21:50
  and motivation 15:937, 21:49
  for success, and leisure motivation 13:833–834
Expectancy beliefs 8:504–505
Expectancy theory 17:314, 17:613
  job satisfaction 12:834
Expectancy value theory (EVT) 6:207, 15:894
  of achievement motivation 7:890–891, 9:677–678
  families as educational settings 8:734–735
  I/E model 12:430–431
  motivation 15:928–929
    implications for learning and instruction 15:929
  of motivation 15:937
  motivation factors 24:65–67
  motivation in Australian Aboriginals 15:900–901
  teacher motivation 24:64–67
  of vocational interests 25:240–241
Expectancy-value models 1:37–40, 6:207, 21:52
  motivation 15:921
  of test anxiety 24:247
Expectancy-value-cost model of motivation 8:503–509
  academic achievement 8:503–504, 8:506–508
  anxiety 8:505–506
  attainment 8:504–507
  cost component 8:505–508
    future 8:508
    past work 8:505–506
    present work 8:506–508
  effort costs 8:507
  expectancy component 8:504
  extrinsic motivation 8:505
  interest value 8:504, 8:506
  intrinsic motivation 8:505
  persistence 8:503–504
  utility value 8:504–507
  value component 8:504–505
Expectation damages 3:29
Expectation state theory 10:437–438
  groups 10:434–435, 10:437–438
Expectation–Conditional Maximization (ECM) algorithm 4:473
Expectation–maximization (EM) algorithms 4:451, 4:469–474, 6:528, 15:624–625
  advances since DLR 4:473
  discrete-state models 6:528
  example (estimation of mixing proportions) 4:472–473
  formulation 4:470–472
  history 4:469–470
  maximum likelihood estimation 4:470
Expectations
  adaptive expectations 8:510–512

adaptive learning approach 8:513–515
Bayesian learning 8:514
bounded rationality 8:515
bubbles 8:513, 8:515
economics of 8:510–516
eductive learning 8:514
in EITMs 18:456–458
expectational stability 8:514
experiments 8:516
forecasting 8:510–516
formation, empirical issues in 8:515–516
fitting survey data 8:515
testing for rationality 8:515
using economic frameworks to extract expectations 8:516
in history of thought 8:510
and learning 8:513–515
economic applications 8:515
other approaches and further issues 8:514–515
stability 8:514
statistical approach 8:514
least-squares learning 8:514–515
rational expectations 8:512–513
estimation and testing 8:513
forward-looking character 8:512
Lucas critique 8:512–513
multiple equilibria and 8:513
rational inattention 8:515
self-fulfilling prophecy 8:513
sunspots 8:513–514
temporary equilibrium approach 8:510–511, 8:514
traditional models 8:510–512
Expected family size 8:805
changes in expectations 8:806
Expected likelihood ratio test statistic (ELRT) 8:719–720
Expected Score 23:761
Expected survival analysis 23:773–774
Expected survival function 23:773
Expected theoretical RT 12:78–79
Expected utility theory (EUT) 2:443–444, 7:88–89, 16:640–642, 17:809–810
early criticisms of 2:444–445
and risk-taking 17:809–810
Expected value of risk 13:504
Expected value theory 15:669
and power 15:669–670
Expected *versus* desired consequences 22:196–197
Expenditure
comparative health systems 4:401
gift as 10:140
Expenditure limits, political money 18:395
Expenditure–income discrepancy 12:51
Expenses for investigation 14:267
Experience
anthropology, and liminality 14:136
art as 1:230
and brain development 2:825
class as 22:341
and class consciousness 3:741–743
consciousness of 4:100
for coping with chronic illness 3:549–550
of emotion 5:544
phenomenology 18:20
rediscovery of 7:175
role in face processing 8:654–655
role of culture 5:545–546
social 5:548
in social systems, Luhmann 14:384
Experience economy 24:792–793
Experience rating
effects, workers 20:865
evidence on 20:866
Experience sampling method (ESM) 1:242–244, 3:51, 7:507–508, 23:648, 25:242
for flow 9:261–263
for time use 24:369
Experience seeking (ES) 21:608
Experience-based sampling *see* Outcome-based sampling
Experienced spellers, factors affecting spelling of 23:270–271
Experiential canalization model 22:149
Experiential learning 6:892, 13:660
Experiential psychotherapy 8:517–520
goals 8:518
human nature and functioning 8:517–518
practice principles 8:518–519
process theory 8:518–519
Experiential theory 8:517
Experiential therapy 8:517–519
Experimental analysis of behavior (EAB) 2:407
Experimental approach, epidemiology 7:827
Experimental designs 8:535–540
Bayesian approach 8:521–526
between-subjects designs 8:537–538
completely randomized designs 8:537–538
matched pairs and block designs 8:538
compliance 8:527–530
features 8:536–537
comparison 8:536
control of extraneous and confounding variables 8:536
removal of bias 8:536–537
large-scale social experimentation 8:531–534
models and analysis 8:537
nested or hierarchical designs 8:539
optimality and efficiency of 8:539
planning 8:537
purpose 8:535–536
random assignment 19:883–886
split-plot designs 8:539
subject selection 8:537
terminology 8:535
within-subject designs 8:538–539
balanced crossover designs 8:538–539
crossover designs 8:538
Latin squares 8:539
Experimental economics 8:399, 9:582, 9:590
auction games 8:549–550
coordination and cooperation games 8:549
crossing disciplinary boundaries 8:550
cognitive load and heuristics 8:550
emotion and cognition 8:550
fairness, altruism, and inequality aversion 8:550
group dynamics 8:550
deception 8:548
disciplinary context and constituents 8:546–547
economic psychology 5:887, 5:890–891
endowment effects 8:548
framing effects 8:548
game theory 8:547–549
and game theory 9:582, 9:590
incentivize focal choices 8:548–549
individual decision-making games 8:549
institutional economic thought 8:550
institutions 8:546–548, 8:550
learning effects 8:548
market allocation 8:548
market design 8:549
market trading games 8:549
noncooperative games 8:549
nonmarket allocation 8:548
potential 8:546
theoretical backdrop 8:547–548
trust games 8:549
usefulness in progressing policy and science 8:550–551
Experimental graduate programs 19:351
Experimental laboratories 8:552–558
computer 8:556
management 8:558
in psychology 8:552–555
beyond participant pools 8:553
hardware, software, and equipments 8:554
human participants 8:552
lab design 8:552
lab location 8:552
online testing 8:554
participant payments 8:554
recruiting participants 8:552–554
undergraduate participant pools 8:553
research questions 8:556–558
context requirement 8:557
measurement devices 8:557–558
tools 8:557
Experimental methods 22:990
in population studies 8:559–565
Experimental philosophy 16:830
Experimental physiology 13:701
Experimental pragmatics 18:801
Experimental psychology 1:853, 4:212, 10:756–757, 13:391, 13:702, 17:609, 18:115, 19:413–415, 22:526, 22:528–530, 23:455
impact on 23:455
and introspection 12:702–703
laboratories *see* Experimental laboratories
of lexicons 13:925
mathematical psychology 14:808, 14:810–812, 14:815
and religion 20:389, 20:392
Wundt's studies 3:873
Experimental research, single-case designs 21:985–986

Experimental studies, infectious diseases 12:34–35
Experimentalization, definition of situation 6:22
Experimentation 13:443, 13:445
causal analysis 3:247–248
Experimentation, in psychology 8:566–570
guiding principle 8:569
origins of exemplary design 8:566–569
creating equal groups 8:568–569
nineteenth-century psychophysics 8:566–567
transfer of training phenomenon 8:567–568
twentieth-century educational psychology 8:568
Experimenter
artifacts 8:571–573
expectancy bias 8:573
expectancy effect 8:573
regress of 21:310
Experimenter artifacts 8:571–575
Experiments 9:128
architectural 2:151–152
biomedicine 2:663–667
class size 3:752
communal experiments of 1960s and 1970s 2:152–153
decline of interest in 8:580
development of new approaches 8:576
ethics of 8:53, 8:57–58
laboratory studies 8:577–579
Latour's naturalistic approach to scientific practice 8:579
material culture of 21:226
naturalistic approaches 8:577–579
and power 15:667–670
practices and regimes 8:579–580
randomized controlled trials *see* Randomized controlled trials (RCTs)
rational choice theory 19:903–905
in science and technology studies 8:576–581
sociological and historical exploitations of new theoretical arguments 8:576–577
Expert assistance 17:144
Expert evidence 21:208
Expert judgment, content validity 4:774–775
Expert occupations, and sociology of professions 19:107
Expert systems 21:187
DAG models for 2:365
Expert teacher 24:56–58
Expert testimony 8:582–587, 12:910
adversarial allegiance in 8:585–586
in civil law systems 8:582
in common law systems 8:582–583
empirical research on 8:586
laws governing the admissibility 8:583–584
legal safeguards against unreliable 8:584
opposing 8:584
as a potential remedy for adversarial allegiance 8:585–586
presiding judge 8:582
reliability of 8:584
during trial proceedings 8:585
in United States 8:583–584
Expert validity 17:542
Expert witnesses 10:594, 21:159–160
Expert-based medicine 8:327–328
Expertise 21:266, 21:311, 25:591–592, 25:595
biotechnology 2:680
citizen science 3:632–633
defined 24:56
econometric software 7:3
and ergonomics 11:300
and intelligence 12:319
I–O psychology
factors driving demand 11:883
and neuroesthetics 16:661
organizational information and knowledge 12:75
public bureaucracies 19:532
competition among experts 19:532
think-tanks 19:532
of teachers in United States 24:75–77, 24:79
Expertise reversal 4:73
Expertized bureaucracies 22:696
Experts 24:294, 24:296, 24:298
energy, culture of 7:622–623
scientific evidence and 21:298–301
social scientists as *see* Social scientists as experts/public intellectuals
Experts and expertise 21:206–211
admissibility of scientific evidence 21:208–209
anthropology of 8:588–592
elites 8:588–589
enskillment 8:589, 8:591–592
institutions 8:589–590
mainstreaming of 8:589–590
'paraethnography' and relations between cultures of expertise 8:590–591
professionals 8:590–592
reflexivity 8:590–591
roots in anthropology of knowledge 8:588–589
theory 8:589
environment 21:206–207, 21:210
expert evidence 21:208
interactional *versus* contributory expertise 21:210
lay expertise 21:209–210
as legal category 21:208–209
politics of science 21:207, 21:210
professions 21:207
public engagement with science 21:210
scientific authority 21:206–209
technocracy 21:207, 21:210
Expiation, and sacrifice 20:860–861
Expiration dates of warranties 17:173
Explanandum 8:595
Explanans 8:595
Explanation-based generalization 19:60
Explanation-based judgment process 12:908
Explanations 8:593–597, 11:126–128, 11:132, 11:134, 16:926, 22:437, 22:954–955
causal
in science 10:787–788
and understanding 10:786
and critical theory 5:301–302
deductive-nomological model of 8:594–595
developmental 8:594, 9:537
embedding theory of 8:597
environmental 8:594
evolutionary 8:594
functional explanation 9:519–524
functional or adaptationist 6:241–242
genetic 8:594
mechanism-based 22:415–420
mentalist 8:594
metaphorical 8:594
methodological individualism 15:313
mixed 8:594
model of aleatory explanations 8:596–597
myth as 16:173–175
myth as other than 16:176–177
objectivity 17:90
physiological 8:594
positivism 18:625–628
psychological explanations 8:594
in psychology 20:358
revolutions, theories of 20:628–630
-seeking questions 8:593–594
in social sciences 14:873–874
and sociological theories 22:954–955
statistical-relevance model of probabilistic explanation 8:595–596
tautological 8:594
teleological 8:594
Explanations, social
form of 15:302–304
giving primacy to individuals 15:302–303
individual motivation 15:303
rationality hypotheses and equilibrium 15:303–304
Explanatory model (EM) 10:567–568
for population forecasts 18:594
of reasoning 5:934
Explanatory Model Interview Catalogue (EMIC) 10:567–568
Explanatory personality models 17:828
Explanatory stories 21:230
Explanatory style
and health 8:598–599
attribution 8:598–599
causal attribution 8:598
cognitive processes and behavior 8:598–599
health anxiety 8:598–599
hypochondriasis 8:598
interpretation 8:598
somatic attribution 8:598
somatic symptom disorders 8:598–599
treatment 8:598–599
Explicator compound verbs 1:956
Explicit attitude 2:172–173

Explicit attitude measures 2:188
Explicit awareness of language sounds 8:395
Explicit dehumanization, research on 6:32–33
Explicit feedback, in mathematics learning 4:69–70, 4:70t
Explicit gender classifications 9:730
Explicit inferences 19:394–395
Explicit knowledge 13:74–80, 13:82–86
Explicit measures 11:714, 11:716–717
Explicit memory 1:644–645, 11:709–710, 23:876–877
  assessment across development 15:148
  assessment of 6:763
  development 15:148–151
  and medial temporal lobe 4:107
Explicit motives 15:951–952
  congruence of 15:952–954
  determinants of congruence of 15:954
Explicit rationing, of health care 10:603
Explicit representation 13:78–79
Explicit rewards 10:417
Explicit rules 13:78–79
Explicit self-concepts 21:470–471
Explicit self-reports of attitudes 2:178–180
  question answering process 2:178–179
Explicit storage of information 13:78–79, 13:82
Explicit syntactic control 4:832–833
Explicit theory construction 1:911
Explicit Theory of Mind 24:268–272
Explicit value premises 21:169–170
Exploitation 1:665–666, 10:206–207, 10:209–212, 16:548
  of biomedical research subjects 8:81
  dependency theory 6:196–197
  and inequality 11:910–911
  labor exploitation 10:209
  organ transplant 10:208–209
  protecting children from 3:426–427
  sex work as 19:281–282
  sexual exploitation 10:207–211
  and slavery 22:38–39
  treatment research and access 8:33–34
Exploitation theory 18:840
Exploration 3:163
  European in Africa 1:286
  identity development 21:434–437
  and visualization 3:163–164
Exploration journals 19:677
Exploration–exploitation dilemma 1:536–537, 16:641–642
Explorations in Economic History 7:58
Exploratory activity, perceptual learning 10:131–132
Exploratory combinatorial data analysis (ECDA) 19:431
Exploratory data analysis (EDA) 8:600, 10:351, 23:95–97
  of area data 23:96–97
  event sequence analysis 8:320
  everyday reexpressions 8:603
  Florida voting, example 8:601
  hotel occupancy, example 8:602–603
  literature 8:603
  teen drivers, example 8:602
  themes of 8:600–601
    displays 8:600
    reexpression 8:600
    residuals 8:600
    resistance 8:600
  univariate methods 8:600–604
Exploratory factor analysis (EFA) 8:686, 8:705, 8:707
Exploratory learning 1:15, 1:18
Exploratory methods in common use 19:730–731
Exploratory play 18:221–222
Exploratory spatial data analysis (ESDA) 23:95–96, 23:108–109
Exploratory structural equation models (ESEM) 8:707
Exploratory system 2:137
Exploratory techniques 23:96
Explosions, population dynamics 18:579
Exponence 3:185–186
Exponential distribution 6:577
Exponential families 23:420, 24:360
  log-linear models 16:125
Exponential formulas 14:351
Exponential growth, population dynamics 18:579–580
Exponential model 14:55, 23:192
Export processing factories (EPFs) 23:208
Export processing zones (EPZs) 10:200–201
Export-Import Bank of China 3:503
Exports, trade policy 24:485–488
Expository documentary 9:164
Expository mode of representation 6:620–621
Expository texts, and first language acquisition 9:204
Exposure, to toxic 10:546
Exposure and response prevention (ERP)
  obsessive-compulsive disorder 17:122, 17:124–125
Exposure in vivo
  for panic disorder/agoraphobia 17:478
Exposure population 5:740
Exposure rates
  LEXIS diagram for 6:156–157
    non-repeatable events 6:156–157
    repeatable events 6:156
Exposure techniques 18:41–42
Exposure therapies 8:605–611
  future directions 8:609
  imaginal exposure 8:606–607, 8:609
  in-virtuo exposure 8:606
  obsessive-compulsive disorder 8:606, 8:608–609
  obstacles to implementing 8:608–609
  panic disorder 8:606, 8:608
  pharmacotherapy 8:609
  post-traumatic stress disorder 8:606, 8:608–609
  theoretical models 8:605–606
    emotional processing model 8:605–606
    Mowrer's two-stage theory 8:605
  treatment characteristics of 8:606–607
    exposure 8:606
    generalization 8:606
    response prevention 8:606–607
  in vivo exposure 8:606–608
Exposure treatments 25:138
  behavior therapy 2:428
Expressionism
  and styles 16:565
Expressions 5:543
  affect programs (AP) 8:668
  of belief 8:73
  disgust 5:544
  emotion assessment from 7:521–523
  emotional 5:543
  of emotions 7:457–460
  facial 5:543–544, 8:667
  fear conditioning 1:812
  of focus 23:720
  of homophobia 10:790
  of life, Dilthey on 6:426–428
  linguistic meaning of 14:842
  LTP/LTD 14:361–362
  oral *see* Oral expression
  pull factors 8:667
  push factors 8:667
  vocal 5:543–544
  in writing 17:274–275
Expressive agon 1:446–447
Expressive communalism 20:300
Expressive forms
  African culture 1:276–279
  and group identity 4:174–180
Expressive movements, with neonates 16:482, 16:486–487
Expressive One-Word Picture Vocabulary Test Revised 3:796
Expressive suppression, and emotion regulation 7:453–454
Expressive Vocabulary Test 3:796
Expressives 18:797
Expressivism 8:72, 18:777
  and practical reasoning 18:780–781
Expropriation 22:845
Expulsion 18:843
  minority-group status 15:571
Extant Anabaptist utopian communities 2:147
Extended case method 3:200, 9:150–151
Extended Categorial Grammar 15:754
Extended deterrence 6:247–248
Extended diglossia 6:417–418
Extended families 8:810, 11:229–231
  emotional development of children in 7:441–442
Extended haplotype homozygosity (EHH) 11:291
Extended hippocampal system 10:880
Extended mind view, of embodiment cognition 7:424
Extended Parallel Processing Model 22:353
Extended peer community 17:683–684
  and science 21:250–251
Extended producer responsibility (EPR) 7:371
Extended schools 8:612–616
Extended self 15:14–15

Extended twin design 6:322
  family design 12:298–299
Extending working lives 1:415–417
Extension problem, in learning word meaning 25:227–228
Extensionist hypothesis, of kinship terminologies 13:63
Extensions (cooperative game theory) 4:871
  of certified knowledge 17:12
  of classification 3:806, 3:808
  core–periphery patterns 4:914–915
  of future time perspective 9:563
  Markov decision processes 14:574–575
Extensions of franchise 7:349
Extensive child care 3:391–392
Extensive game 9:583
Extensive magnitudes, measurement theory 14:868
Extensive measurement, representational theory of measurement 14:857
Extermination, minority-group status 15:571
External constituency surveillance 23:737
External control 4:834–835
External corroboration of ethnography 24:291
External criticism 21:451
External economies 14:622–623
External factors shaping policing 18:249
External frames, academic self-concept and achievement 1:58
External motivation 15:907–908
  for behavior maintenance 15:908
  clinical intervention 15:907–908
External nonconstituency surveillance 23:737
External organization, Dilthey on 6:425–426, 6:428
External political efficacy 7:301–306
External possession 18:632
External reasons, identity 11:525
External regulation 21:488, 24:68
External representation of numbers 14:775–776
External self-determination 11:797
External stability 4:869
External threats, and political repression 18:449–450
External unfolding 16:46
External validity 8:617–622, 9:129, 12:450, 14:354, 19:739
  complex interactions and 8:617–618
  definition and challenges 8:617–618
  evidence-based medicine 8:329–330
  in generalization 9:840
  inferences from samples 8:618–619
    causal 8:620
    generalizing from 8:619–621
  perceived trade-offs between internal and 8:617
  of placebo-controlled double-blind studies 18:162
  in population studies 8:562–565
  sampling frame for 8:618
  studies to reduce uncertainty about UTOS and *UTOS 8:619–621
  threats to 8:618
External validity, experimental estimates 22:297
External war 4:615
External womb (life phase) 5:509–510
External-internal fit (EIF) paradigm 23:522
Externalism 2:134
  Putnam on 19:665
Externalities 7:726, 19:168, 20:160
  defense economics 6:3–5
  norms 17:5, 17:8
  regulation 20:170
  regulatory agencies 20:181, 20:183–184
  and self-efficacy 21:505
  in social capital 22:174–175
  unintended consequences 24:747
Externalization 1:539–540
Externalizing disorders 17:916–917
Externalizing problems
  behavior problems 2:477–478
  and divorce 14:519–523
Externally driven population cycles 18:559–560
Externally sponsored research
  academic research and employment 1:49
Extinction 8:850–852
  in ABA 2:409
  classical conditioning 3:765, 3:768
  fear conditioning 1:812–813
  Oceania 17:449
  population dynamics 18:579
    reproduction 18:579–580
  reinforcement 20:207–208
Extinction learning 8:872
"Extirpation of idolatries" campaigns 6:441–442
Extortion 14:430
Extra question method 14:589
Extraction stage of pattern recognition process 23:411
Extracurricular activities
  of adolescents 13:829–830, 13:832–833
  and childhood friendship 9:445
Extracurricular time 7:276
Extrajudicial executions 18:446
Extralegal justice 12:932–933, 12:935
Extralinguistic features of child-directed speech 3:401
Extramarital affairs 17:486–487, 17:489–490
Extramarital sex 21:701, 21:821
Extraneous variables, control, by experimental designs 8:536
Extraocular muscles, cranial nerve nuclei tension control in 8:628–629
Extraordinary power, charisma 3:353
Extraordinary science, Kuhn on 13:134–135
Extraordinary writs 1:823–824
Extrapersonal space 17:703–704
Extrapolation 9:317–319
  to nonsampled entities 9:843–844
Extrapolation techniques 6:164
  for small areas 6:164, 6:166
Extrapontine myelinolysis, and alcohol 1:512
Extraposition 16:2
Extrapyramidal system (EPS) 15:966, 19:466
Extrasensory perception (ESP) ability 21:635
Extrastate war 25:420
Extrastriate areas, in visual cortex 17:129
Extrastriate form vision 25:160–161
Extrastriate visual areas, and visual mental imagery 25:164
Extraterritorial jurisdiction 10:449
Extra-university institutes
  infrastructure 22:669–670
Extraversion 3:244, 9:236, 17:777, 17:787, 17:802, 17:914–917
  and academic performance 17:789
  across lifespan 8:625
  and adolescent personality 17:864–866
  Big Five 8:625, 16:743
  biological approach 8:624
    reinforcement sensitivity theory 8:624
    revised reinforcement sensitivity theory 8:624
  as complex, dynamic system 8:625–626
  conceptual representation of facets of 8:625
  Eysenck on 8:624
  facets of 8:625t
  history 8:623–624
    ancient typological accounts 8:623
    early biological accounts 8:624
    early lexical accounts 8:623–624
    psychoanalytic accounts 8:623
  with impulsivity 17:789
  individual differences 8:623
  in interpersonal behavior 17:789
  with novelty-seeking 17:789
  and personality traits 17:920, 17:922–923
  and temperament 24:187
  underlying structure of 8:624–625
  and valued outcomes 8:625–626
Extreme evil case study
  for responsibility 20:590
Extreme order statistics 16:898, 17:293
Extreme peace negation 17:622–623
Extreme realism 5:89
Extreme response style (ERS) 23:758–763
  bias
    correction 23:762
    evaluation of effects of 23:761
  consequences of 23:758–759
  factors related to 23:761–762
  future directions 23:762–763
  index-based measurement of 23:759
  latent trait models 23:759
  measurement error 23:758–759
  psychometric modeling of 23:759–761
  rating scale 23:758–760, 23:762–763
  self-report measures 23:758–759, 23:762
  survey instrument 23:758–762
Extreme right 18:410–416
  anti-semitism 18:411–412, 18:414
  ethnocentrism 18:411
  fascism 18:410–413
  future of 18:415–416
  irredentism 18:411, 18:415–416

Islamophobia 18:412, 18:415
McCarthyism 18:410–412
movement, alternative forms of 18:414
Nazism 18:410–414
neofascism 18:411–412
neo-Nazism 18:410–412, 18:414
new wave 18:412–415
parties
definition 18:410–412
support, social bases of 18:413–414
tendency to schism 18:414–415
phenomena in twentieth and twenty-first centuries 18:412
populism 18:411–414
racism 18:411–414
radical right 18:410–414
right-wing populism 18:411–412
term 18:410
theoretical approaches 18:412–413
ultranationalism 18:411, 18:415–416
White Power 18:414
xenophobia 18:411–413
Extreme weather events, multilevel interventions during 10:389–390
Extremisms
macro/microsociologies 14:415
populism 18:611–612
in United States 16:249–251
Extremist Crime Database (ECDB) 16:247–248
Extremist factions 12:273–274
Extrinsic affective Simon task (EAST) 11:715
Extrinsic aspirations 21:490
Extrinsic compensation 17:612
Extrinsic contingencies 21:487–488
Extrinsic culture traits 2:85
Extrinsic mortality 6:757, 17:175
Extrinsic motivation 8:505, 21:51, 21:487, 24:187
in terms of degree of internalization 21:488
Extrinsic religiosity 20:382–383
Extroversion 16:493–494
Exurbia 23:663
Exxon (oil company) 17:166
Exxon Valdez tanker oil spill 3:736, 7:806–807
Eye contact, of children, and intentionality 12:348
Eye gaze 12:27–29
facial 8:650, 8:654
Eye knowledge 21:617
Eye movements 8:628
and change blindness 3:344–345
control 8:633–634
control of 8:628–630
cortical areas involved in 5:27
frontal eye fields 5:27
parietal eye fields 5:27
supplementary eye fields 5:27
developmental changes in 8:632
integration of information across eye fixations 8:633
perceptual span 8:632–633
during reading 8:631–632
in reading 8:631–634
regressions 8:631–632
saccades 8:631–632
Eyeblink, and startle response 23:331–333, 23:335
Eyeblink conditioning 8:635–641, 16:845–846
in animals, role of cerebellum 2:93–95, 2:93f–94f
anterior interpositus 8:636–637, 8:639
associative learning 8:635–637, 8:639–640
behavioral studies of 8:635–636
cerebellar feedback to CS and US pathways 8:638–639
conditioned and unconditioned stimulus input circuitry 8:637–638
future directions 8:639–640
identifying brain regions essential for 8:636–637
inferior olive 8:638–639
mechanisms of cerebellar learning in 8:637
paradigm 17:609
in patients, role of cerebellum 2:95–96
pontine nuclei 8:637–639
Purkinje cell 8:637–638
Eye-contingent paradigms 8:632–633
boundary technique 8:632–633
foveal mask technique 8:632–633
moving window technique 8:632–633
Eye-tracking 16:351–352
Eye-tracking devices 16:700
Eyewitness memory 8:710–711
Eyewitness testimony 8:642–649
beliefs 8:644, 8:647–648
calibration 8:643–644
cognitive interview 8:645–646
confidence–accuracy relationship 8:643–644
effects of misinformation 8:644–645
cowitness misinformation 8:645
factors 8:642–643
feedback 8:644
of forensic psychologist 9:353
gathering evidence from witnesses 8:645–647
interviewing techniques 8:645–646
lineups 8:646–647
juror's appraisal of witness reliability 8:647–648
memory 8:642–645, 8:647–648
postevent information 8:644, 8:647–648
weapon focus effect 8:642–643
Eyler, A.E., on LGBT aging 13:864
Eysenck, Hans Jürgen 3:243, 8:130–131
contributions to clinical psychology 3:874
criticisms 19:506
on extraversion 8:624
personality theory of 17:777, 17:802–803, 17:858
Eysenck Personality Questionnaire (EPQ) 9:250, 9:1010, 17:828
revised version (EPQ-R) 17:828
EZLN *see* Chiapas-based Zapatista Liberation Army (EZLN)
Ezrahi, Yaron 17:13
works, *The Descent of Icaru* 8:578–579

F

F2 area (premotor cortex) 18:849
F3 area (premotor cortex) 18:849
F4 area (premotor cortex) 18:849
F5 area (premotor cortex) 18:847–849
F6 area (premotor cortex) 18:849–850
F7 area (premotor cortex) 18:849–850
F distribution 6:578
F scale 1:660, 2:255–260, 9:459, 19:866
FA *see* Football Association (FA); Fractional anisotropy (FA)
FAB *see* Feminist Association for Bioethics (FAB)
Fabian, Johannes 1:740, 1:747
works, *Time and the Other: How Anthropology Makes Its Object* 21:751
Fabian society 22:698, 25:675
Face 5:428–429
expression of emotion 7:522
in interpersonal conflict 5:429
neural representation of 16:628–632
and politeness 18:328
and self 21:421–422
threats to face 5:428–429
visual scanning of 12:27
Face agnosia *see* Prosopagnosia
Face cognition 8:680–681
Face detection 8:658–659
Face Emotion Action Stimulus Test 19:251
Face identity 8:650, 8:676
individual differences in 8:676–678
measures of 8:678–679
Face identy 8:658
Face normalization 8:659
Face perception
during early childhood 8:650–657
categorization and recognition of faces 8:651–652
emotional expression 8:650–652, 8:654
eye gaze 8:650, 8:654
featural and configural processing 8:652–654
holistic processing 8:652–655
other race effect 8:655
perceptual narrowing 8:655
preference for faces 8:650–651
role of experience in face processing 8:654–655
role of social face information in face processing 8:654
visual memory 8:655
Face processing 8:663–664
brain network 8:664–665
infant cognitive development 4:59–63
Face recognition 4:90, 8:658
by animals 2:801
brain network 8:664–665
in brain-injured individuals 8:663–664
electroencephalography 8:664–665
emotional expressions 8:663–664
'face-specific' cortex 8:665–666
fusiform gyrus 8:664–665, 8:664f
magnetoencephalography 8:664–665
models 8:658–662

Face recognition (*continued*)
challenges 8:661–662
face detection and landmark localization 8:658–659
face identification 8:658
face normalization 8:659
face verification 8:658
feature extraction 8:659–661
feature matching 8:661
Gabor filter 8:659–660
holistic feature 8:659–660
illumination preprocessing 8:659–661
local binary pattern 8:659–660, 8:660f
local feature 8:660
sparse representation-based classifier (SRC) 8:661
and object recognition 8:663–666
prosopagnosia 8:663–666
and self 21:421–422
Face recognition units (FRUs) 8:663
Face specificity 8:679–680
of prosopagnosia 19:251–252
Face threatening process (FTP) 11:187, 11:189
Face validity 1:370–371, 4:775, 6:345
Face verification 8:658
Face-based emotion processing 8:677
Facebook 3:162, 4:288, 4:504–505, 4:508, 11:551, 12:600, 12:602–604, 17:215–216, 17:439–440, 21:808, 21:817, 24:147, 25:840
footing in 14:174
Internet privacy 12:593–594, 12:597
oversharing through 17:439–444
social protests 22:515
terrorists on 24:231–233
use in politics 14:717
Face-inversion effect 4:59
Face-negotiation theory (FNT), conflict 11:187
FACEs *see* Facie average causal effects (FACEs)
Face-threatening acts (FTAs) 18:327–328
Face-to-face communication/interaction 4:334, 4:505, 14:227, 16:791, 18:797–798
and folklore 9:281, 9:283
Facets 2:560–561, 2:563
Facet-specific organizational climate 17:323, 17:325
Facial Action Coding System (FACS) 7:459, 7:522, 8:668, 8:670
Facial characteristics 11:335
human mate choice based on 11:337–338
Facial electromyography, for assessing emotions 7:521
Facial emotion expressions 8:667
individual differences in 8:667–675
measurement of facial expression production 8:669–670
nomological net 8:673–674
structure of voluntary facial emotion expression (VfEE) 8:670–673
neural architecture of 8:668–669
theories on 8:667–668
Facial Expression Coding System 8:670
Facial expressions 5:543–544, 5:547, 7:457, 7:485
affect program theory of 7:457–459
contemporary alternatives to 7:459
Frijda's theory 5:545
origins of research 7:457
Facial information processing 8:676
individual differences in 8:676–678
measures of 8:678–679
Facial morphs 25:173
Facial perception 8:676–682
Facie average causal effects (FACEs) 3:253
Facilitation 10:622
of coping networks 20:224
Facilitative law
as dependent variable 17:399–400
as independent variable 17:399
Facilitator 20:224–225
Facilitator state 5:464
Facio de Combattimento 16:257
FACS *see* Facial Action Coding System (FACS)
Fact and law, distinction between 1:824
Factionalism 8:683–685
coalitions 8:683–685
conflict 8:683–685
groups 8:683–685
issues 8:684
leadership 8:683–684
organizations 8:683
political faction 8:683–684
status quo 8:684
structure, symmetry, and balance 8:685
study of 8:683–684
Factish 9:110
Factor analysis 8:686–691, 8:698–702, 12:357–358, 14:853, 19:79, 19:434, 20:248, 24:825–826
conceptual framework 8:692–693
future developments 8:696
health risk perception 10:703
historical background 8:695–696
idea 8:691
of intelligence 12:291–292, 12:294
mathematical psychology 14:812
members of family 8:693–694
and personality psychology 17:910
in personality psychology 8:705–707
determination of number of factors 8:706
exploratory and confirmatory approaches 8:705
extraction of factors 8:705–706
factor model 8:705
identification and rotational indeterminacy 8:706–707
for semantics 21:563, 21:565–568, 21:570, 21:565f
structure of family 8:693
Factor analyzers, mixture of 15:627
Factor endowments 7:57
Factor loading 19:79
Factor models, to big five theory 2:561
Factor specification equation 3:244
Factor V 17:232
Factories
concentration of labor in 13:158–159
emergence of 11:856–857
Factories Amendment Act 1844 17:142
Factorization theorem 23:419
Factors influencing teaching choice (FIT-Choice) 24:64–66
Factors of educational effectiveness 7:225
Factory 798 5:598–600
as Cultural Creative Industry Cluster 5:599
Factory era, organization of industry in 11:855–856
Factory Regulation Act 1833 17:141–142
FACTS *see* Faith, Community and Teachers for Students (FACTS)
Facts
and relativism 20:233
social 22:500–505
Factual knowledge 13:93, 14:313–314
Fact–value distinction 16:321, 16:755, 20:23
axiological neutrality
logical grounds 16:755
practical grounds 16:755–756
Facultad Latinoamericana de Ciencias Sociales (FLACSO) 22:622, 24:295
Facultative calibration model of personality 17:903
Faculties 22:588–589
employment 22:661–663
inferior 22:588
medieval 22:592–593
superior 22:590
Faculty of language 13:75
Faculty psychology 24:762
Fad, and fashion 8:835, 8:838
Fading, in amputees 18:2
Fagan Test of Infant Intelligence 18:854
Fagerstrom Test for Nicotine Dependence (FTND) 1:99
Fagging 10:557
Faience, usage in ancient Egypt 1:694–695
Failed states 23:348, 23:350, 23:354
Failing states 14:426
Failure
health care markets 10:598
implementation 11:692–693
design issues 11:693
and self-efficacy 21:504–508
Failure to treat 9:131
Failure-driven reminding events 19:56
Failure-time data and representation 23:776–777
Failure-time regression, survival analysis 23:777–779
Fair and Accurate Credit Transactions Act 5:635
Fair equality of opportunity 7:926
Fair game 14:630
Fair housing 24:931–932
Fair Housing Act 20:540–541, 22:702, 24:930
Fair Labor Association (FLA) 3:676, 10:257
Fair Labor Standards Act 17:612
Fair majority voting 19:207

Fair process effect 13:749
Fair ranking 4:874
Fair trade, Slow Food movement 22:94, 22:96
Fair trial 9:396
Fairchild, Henry, works, *Dictionary of Sociology* 7:589
Fairclough, Norman 6:506
Fairer systems of voting 17:649
Fairness 3:822–823, 7:483, 17:718–719
, conceptions of 8:485
and equality 7:907–909, 7:911
experimental economics 8:550
global governance 10:160
and justice 12:930–931
of leader 13:643, 13:645
of majority rule 14:446–447
media ethics 15:43
organizational decision making 17:355–356
promoting, global public policy 10:179–180
evidence 10:181–182
social justice 22:392, 22:394
standards, program evaluation 19:119
at work 23:548
Fairness equilibrium 9:589
Fairness instinct 15:814–815
Fairness principle 7:926
Fairness theory 23:702
Fairtrade and Fairmined Gold Standard 8:61–62
Fairy tales 9:281
Faith
and intimate partner abuse 12:682
in openness, mass communication 14:692–694
Faith, Community and Teachers for Students (FACTS) 24:494
Faith development, theory of 20:355–356
Faith Leader Training Program 12:682
Faith missions 15:616
Faith Trust Institute 20:256
Faith-based groups 3:683–684
Faits sociaux 6:701
Fake marriage 21:735
Faking 17:834–839
definition 17:834–835
definition of 17:834–835
as interaction between person and situation 17:835
as means to achieve personal goals 17:835
as response Set 17:835
existence 17:835
measurement 17:836–837
anchoring vignettes 17:837
Bayesian truth serum 17:837
decision trees 17:837
item response theory (IRT)-based approaches 17:837
overclaiming 17:837
social desirability scales 17:836–837
word use 17:837
and personality 17:834
prevention 17:837–838
fake resistant measures 17:837
reasoning with test-taker 17:837–838
and selection outcomes 17:836
and self-reports 17:834
and test criterion correlations 17:836
and test score's construct validity 17:836
Falkenhayn, Erich Georg Anton Sebastian von 9:214
Fallacies
of composition, rational choice explanation 19:896–897
linear hypothesis 14:150–152
Fallibilism 5:271–273, 5:275
Albert on 1:495–496
and critical realism 5:277
Fallowell, Duncan, on Russian sexual life 21:740–741
False alarms 6:720
False Belief Task paradigm 15:540–541
False beliefs
and pretend play 18:873
Theory of Mind 24:268–271
False confessions 25:788
False consciousness 3:740–743, 10:742, 13:750, 18:766, 22:721
of ruling class 22:526
False discovery rate (FDR) 9:529, 11:501–504, 16:82–83
defined 16:83
False friends 21:350
False legitimation 5:252
False memories 8:709–714, 20:50
autobiographical memory 8:709, 8:712
and brain 8:713
eyewitness memory 8:710–711
final remarks 8:713
imagination 8:710–713
memory distortion 8:711–712
misinformation effect 8:711–713
psychiatric ethics 8:66
in real world 8:709–711
recovered memory 8:710
source monitoring 8:713
suggestion 8:710–713
theories 8:712–713
False recognition (memory) 20:42
Falsifiability 5:271–272
Falsification
of data and methods 21:313–314, 21:316
of evolution 8:473
and objective research 17:93–94
principle of 4:440
Falsificationism 5:271, 5:274, 18:523–525
Feyerabend on 9:117–120
methodology 5:271–272
philosophy of 21:250–251
Popper on 18:522–527
Falun Gong (spiritual discipline), demonstrations of 2:69
Fama Sanctitatis 15:577
Fame 20:484–485, 20:488–490
Familial antecedents 11:540
Familial arrangements, and scholars 21:34–36
Familial relationships, early 9:451
Familial right to property 19:174
Familial searching 9:339, 9:960
Familial studies
eigensimplification approach 8:718–721
examples of 8:721–722
genetic inferences 8:715–724
genomics phase of heritability controversy 8:716–717
heritability concept 8:715–716
human statistical genetics 8:715–716
pregenomics phase of heritability controversy 8:716
sequence data, impact of 8:717–718
Familialism by default 12:404
Familiarist regime, educational policy in 25:488
Familiarity
context dependent memory 4:781–782
and embodiment 7:413
and episodic memory 7:859–860
and hippocampal amnesia 10:879–884
and interpersonal attraction 12:628
of recognition memory 20:37–40
Familiarity of being (*oikeiosis*) 16:482
Familiarity-based recognition memory 24:565–566
Familiarization period 15:148
Families 14:422, 18:597–600
and abusive supervision 23:702
and adoption 1:139–140
anthropology and history 8:809–814
children 8:809–813
Christianity 8:809
concepts of nature 8:812
household 8:809–813
incest 8:811
kinship 8:809–813
parents 8:809–812
private life 8:809
sexual relations 8:811
in Western society 8:809
background, child care effects based on 3:393
and body image 2:736–737, 2:739
characteristics 15:923
child health psychology 10:679–684
and childhood behavior problems 2:477–480
in China 2:68–69
climate 15:924
in collectivism/individualism 4:208
and demography 6:176–178
diversity 15:926
domestic 11:228–231
effects of income support policy 18:745
effects of military service on 15:515–516
extended 11:229–231
and fertility transition 9:49–52
and fertility transition, in China 9:47–48
functional basis 11:228
generation 9:853–856
genetics, and social justice 9:970–971
history of 8:738–744
cultural history 8:738

Families (*continued*)
demography 8:738–742
divorce 8:742
emergence of field 8:738–739
future directions 8:742–743
household structure 8:739–740
institutions 8:738–740, 8:743
interaction with process of industrialization 8:741–742
kinship 8:741
labor 8:741–742
life course 8:740–741
relationships 8:739, 8:742–743
social science 8:738
social ties 8:739, 8:741–743
states 8:743
strategies 8:742
work 8:738, 8:740–742
and home 11:156–157
honor 5:430–431
honor killing 11:185–190
household economics 11:224
impact of socioeconomic status on 22:159
influences on motivation 15:921–927
influences/paradox, and food preferences 9:298
as institution 8:766–771
alternative family forms 8:769
cohabitation 8:766–770
cohabiting families 8:768
decline of marriage 8:767–768
divorce 8:767–770
family structure 8:766
gay and lesbian families 8:769–770
increase in age at marriage 8:767
marriage 8:766–767
never married 8:767–769
sex 8:770
single adults and single parents 8:770
social organization 8:766
unmarried childbearing 8:768–769
and intergenerational justice 12:397, 12:399–400
involvement in psychiatric care 6:45–46
Jewish Diaspora 12:801, 12:803–804, 12:808–809
and labor history 13:169
of lesbians 13:894–895
LGBT friends as 9:454–455
transgender 9:455
LGBTQ 21:755–758
family of choice 21:755
individuals, parenting by 21:757
marriage and parenting legislation 21:756–757
in twenty-first century 21:755–756
limitation 11:6–7
and marriage payments 6:643
and marriage systems 14:611, 14:615, 14:618
matrifocal 14:822
and mental health policies 15:216
Middle Ages 15:349–350
and migration 15:428–429
obligations and imperative of heterosexual marriage in China 21:734–735
and partnerships, in Western societies 17:550–551
of people with learning disability 13:684, 13:686–689
policy, and poverty causes 14:823
privacy of 3:478, 3:482–483
psychosocial effects of Alzheimer's disease on 1:586–589
and psychosocial interventions for psychiatric disorders 22:382–383
reconstitution 11:5–7
and reentry 20:62–63
-related consequences of work–family conflict 8:760
relations, and wealth flows 9:37
relations, nonheterosexual British Asians 16:865
and relationships 15:370–371
and religion 20:317–321
and religiosity of women 25:635–637
role in fertility decision-making 9:37–38
role in war/political violence 25:391–392
Roma people 20:748–753
and Second Demographic Transition 21:346–347
size 11:5
social inequality and schooling 22:326–327
in social work*under* Social work
social work initiatives based on 3:473
and socialization, in infancy and childhood 22:866–872
socioeconomic status
and student achievement 5:344
sociology, Randall Collins 4:214–215
in South Asia 2:79
in Southeast Asia 23:73
and suicide 23:685
support for social work 3:473
system 11:6, 11:8–9
time of children with 9:445
toward aging 1:318–319
and value of children 3:498–501
Families as educational settings 8:731
child development 8:731–732
classic theories 8:731–732
Bandura and social cognitive theory 8:731–732
Bronfenbrenner and bioecological systems theory 8:732
Vygotsky's sociocultural theory 8:731
current research directions 8:736–737
expectancy-value theory 8:734–735
home environment 8:732–733
parent language 8:732–733
physical characteristics of 8:732
mesosystem 8:732
microsystem 8:732, 8:735–736
modeling 8:735
observational learning 8:735
parents' and siblings' behaviors 8:735–736
out-of-home activities 8:736
parents' school involvement 8:735–736
as role models and teachers 8:735
parents' beliefs and values 8:733–735
causal attributions regarding children's behaviors 8:734
cultural stereotypes 8:734
parents' expectations of the child and perceptions of children's competencies 8:733
values and interests 8:734–735
scaffolding 8:731, 8:733
Families Need Fathers 8:846
Families of choice 13:864–865
LGBTQ 21:755
Familism
interpersonal relationships 5:430
in United States 11:630
Familismo 9:628–629
Family and Fertility Surveys (FFS) 20:430
Family and Medical Leave Act (FMLA) 13:592–593, 13:609
Family and schooling 8:751–757
development of children's competencies and dispositions 8:752–754
cultural capital 8:753–754
economic capital 8:753
social capital 8:754
variation in family structure 8:754
educational opportunity, equality of 8:751
family socioeconomic status and children's school learning 8:752
home and school, discontinuities between 8:751
home support initiatives 8:754–756
health and nutrition 8:754–755
home–school partnerships 8:755–756
nonshared environment 8:756
research designs to study 8:751–752
response of schools 8:756
Family and work 8:758–765
border/boundary theory 8:760–761, 8:764, 8:761f
conflict 8:759–763, 8:759f
chronic stressors 8:761–762
crossover stress 8:763
documented influences 8:762–763
future directions 8:763–764
prevalence and consequences 8:760
survey measures 8:763
theories and perspectives 8:760–762
labor market patterns 8:758–759
in Asia 8:758–759
in Australia 8:758–759
in Canada 8:758
in Europe 8:758–759
in United States 8:758
Family Assistance Plan (FAP) 1969 18:736
Family business 3:23–24, 8:772–778
insights on next developments
drivers of development 8:776
methodological challenges 8:777
relevant topics 8:777
origins and development of 8:773
present to future
theoretical arena 8:775–776
theory building 8:776

theory testing 8:776
research areas and achieved results 8:773–775
Family capital, and education 7:919
Family capitalism 2:750
Family care, in England 4:344–345, 4:347–348
Family caregiving 3:135–137, 20:593–594
Family carers, for dementia 19:298
Family change
cohabitation *see* Cohabitation
and disasters 6:105–107
and parenting 17:495–496
Family characteristics
internal migration 12:443
motivation 15:923
and socioeconomic status 22:925–926
Family climate 15:924
Family consequences, and mass imprisonment 14:699–700
Family constitution 8:828
Family context
engagement 1:32
retirement in 20:606
Family counseling, for childhood cancer 3:459–460
Family demography 8:725–730
classification by subject matter 8:725–726
data issues 8:726
headship rate 8:728–729
kinship 8:725, 8:727–728
LIPRO household model 8:728
microsimulation 8:727–728
models 8:726–729
agent-based models 8:729
dynamic *versus* static models 8:728–729
macromodels *versus* micromodels 8:727–728
severing the link strategy 8:727
ProFamy model 8:728
relationships 8:725–729
SOCSIM model 8:728
unit of analysis 8:726
Family designs, for genetic studies of behavior 9:934–935
Family diversity 15:926
Family dysfunction, and child neglect 3:413–414
Family education programs 10:684
Family Educational Rights and Privacy Act 1974 (FERPA) 23:754–755
Family engagement 8:614
Family environments
child and adolescent psychiatry 3:384–385
authoritarian parenting 3:385
authoritative parenting 3:385
parenting 3:385
permissive parenting 3:385
Family firm *see* Family business
Family formation, event history analysis for 8:308–309
Family Group Conference (FGC) 14:506
Family history, child and adolescent psychiatry 3:387
Family homicides 11:166
Family identity, in genomics 10:34
Family income
and academic achievement 22:925–926
and education 7:132–136
Family inequality, and student achievement 5:345–346
Family instability, and cohabitation 4:120
Family interventions 8:745
Family investment model (FIM) 22:925
Family labor allocation 17:579
Family labor supply decisions 13:195–196
Family law 25:639
colonizing 15:362
and gender 9:785–786
in United States 8:779–785
class and parental ties 8:783–784
custody 8:780–784
deinstitutionalizing of traditional family 8:780
divorce 8:780–784
legal parentage as foundation of 8:781–782
marriage as marker of class 8:780–783
primacy of parent–child tie 8:782–783
redefinition of adult ties 8:781–782
traditional family of industrial era 8:779–780
Family Law Amendment (Shared Parental Responsibility) Act (Australia) 20:468–469
Family leave
and child care 3:393
psychosocial effects of Alzheimer's disease at 1:586–589
Family life cycle theory 8:746
Family Man program 8:847
Family migration 15:436
Family niches, and siblings 21:924–925, 21:927
Family obligation, and academic motivation 19:794
Family of inferences 16:82
Family organizations
mental health 15:216–217
benefits of involvement with 15:218
functions 15:217
Family planning 2:698, 8:786–793, 9:77–78, 9:909, 9:998
child health services 8:788, 8:790–792
contraception 8:786–790
education 21:827
estimates for fertility 8:791t
maternal health services 8:788, 8:790–792
mortality 8:789–790
number, timing, and spacing of births 8:788–789
population growth 8:786
Pritchett on 8:787–788
programs 8:788–789
benefits among children 8:790
benefits among mothers 8:789–790
health and socioeconomic outcomes among women and children 8:789–790
observational studies 8:788–789
randomized controlled trials 8:788
rconomic development 8:786
reproductive health 8:786–788
supply *versus* demand 8:787–788
Family planning, Female education, and Food supplementation *see* FFF (Family planning, Female education, and Food supplementation)
Family planning programs
in China 9:47
feminist perspectives 8:794–798
Cairo Programme of Action 8:795–796
ICPD 8:794–796
population paradigms 8:794
reproductive rights 8:795
sexual rights 8:796–797
women's rights 8:794, 8:796–797
youth 8:796–797
and social networks 9:6–9
in South Asia 9:80, 9:82–83
Family policy 8:799
child benefits 8:801
child well-being 8:803–804
constituents 8:801–803
child benefits 8:801
early childhood education and care 8:801–802
employer and enterprise provision 8:803
parental and maternal/paternal leaves 8:802–803
definition of field 8:799
expenditures 8:801
flexible work 8:803
gender equality 8:800–801, 8:803
general trends 8:803–804
history in Europe 8:799–800
models 8:800–801
reform 8:799, 8:801, 8:803–804
types of approaches 8:800–801
work–life balance 8:801, 8:804
Family problems
coping, and religion 20:320
divorce 20:320
Family reconstitution 6:138–142
civil vital event registers 6:139
data quality and coverage 6:140–141
demographic databases 6:140
event history 6:139–142
genealogies 6:139
historical demography 6:138–139
history and scope 6:138–139
household registers 6:140
life history 6:138–139
limitations 6:140–142
methods 6:140
parish registers 6:139–141
population registers 6:139–140
censuses 6:140
continuously updated 6:140
updated at fixed intervals 6:139–140
prospects 6:142
record linkage 6:139–140
statistical biases caused by sample selection 6:141–142
bias from delays in recording 6:141–142

Family reconstitution (*continued*)
selection by data quality 6:141
selection by migration 6:141
truncation and censoring 6:141
technique 6:138–139
Family relationships 22:156–159
and adolescent health 1:123
in Caribbean 1:642
forming 20:318–319
home and work balance 20:319
marriage 20:318
maternal fertility 20:318–319
paternal involvement 20:319
knowledge about 20:318
migrants 15:417–418
stressful relationship 22:158–159
sustaining 20:319–320
decreasing risk of domestic violence and sexual infidelity 20:319–320
disciplining and nurturing children 20:320
enhancing marital quality 20:319
risk of child physical abuse 20:320
Family resemblances 20:308
estimating sources of 9:934
Family resilience model 21:755–756
Family responsibility 20:582
Family reunion
and immigration policy 11:662, 11:664–665
and migration in Europe 15:376, 15:378–379, 15:381, 15:379f
Family sentiments 21:926–927
Family size 24:640
hypothesis tests 11:503
Family size preferences 8:805–808
births 8:807
changes 8:806
desired family size 8:805
expected family size 8:805
fertility 8:805–807
gender 8:807
ideal family size 8:805
intended family size 8:805
measurement 8:805–806
nonnumerical responses and uncertainty 8:806
parity 8:805
sex composition 8:806
Family socialization 2:10–11
Family solidarity, theory of 12:403–404
Family space, and kinship 13:55, 13:58
Family stress 22:159–160
Family stress model (FSM) 22:925
Family structures 12:404, 22:160
and behavior problems of children 2:474–475
age differences 2:475
gender differences 2:475
changes in 8:766
and cohabitation 4:120
diversity of 2:470–471
behavior problems, development of 2:470
effect on poverty 18:753–754
effects on emotional development of children 7:438–444
adoptive families 7:442
divorced families 7:440
extended family 7:441–442
kinship care 7:441–442
lesbian and gay families 7:442
protective factors 7:440–442
risk factors 7:439–441
single parent families 7:440
stepfamilies/remarried families 7:440–441
traditional families 7:439–440
and homicide 11:164
and kinship 13:55
and nonmarital childbirth 9:15–16
variation in, and schooling 8:754
Family studies, of sexual orientation 21:780
Family support 3:473
and fathering 8:847
Family support team (FST), in Australia for disasters 7:428–430
Family Survey Dutch Population and Cultural Capital and Social Exclusion survey 5:376, 5:380
Family systems 4:729, 4:731
Family theory 8:815–820
cohabitation 8:815, 8:819–820
economic inequality 8:819–820
economics of childbearing 8:821–825
gender 8:818
independent nuclear family system 8:816–817
role of relative cohort size 8:816–817
intergenerational relations 8:826–828
sex-role specialization 8:817–819
coping with risk and change 8:818–819
empirical status of independence hypothesis 8:818
mismatches 8:817–818
women 8:815–820
Family therapy 8:745, 19:515
defined 8:745–746
in emotions 8:747
systems approach 19:505
criticisms 19:505
theoretical overview 8:746–747
values and ethics in current practice 8:747–748
Family tree 11:820–821
language contact 13:300–301
Family values, and LGBTI commercial sex 13:884
Family violence 6:633–634
intimate partner abuse/violence *see* Intimate partner abuse (IPA)
Family welfare 3:418
Family work, total hours, in dual-career families 14:607
Family-based immigration 11:636–637
Family-based studies, life course epidemiology 14:73–74
Family-based therapy (FBT) 6:827
for anorexia nervosa 19:376
Family-oriented programs, offender rehabilitation 20:194
FamilyTreeDNA 9:988
Family-wise error rate (FWER) 9:529, 16:82–83
defined 16:83
Famine 11:7–8
and poverty 18:730
Famine-based cannibalism 3:84
Fandom
of celebrity 3:276–277
television 25:82–83
Fanon, Frantz
works
Peaux noires, masques blancs 15:136
The Wretched of the Earth 25:341
Fans 14:720, 14:723–724
cultures and taste, cultural production 5:610–611
Fantasies
and integrated attachment theory 12:229
and motivation/actions 15:889–891
Fantasizing, and hope 11:199
Fantasy proneness 11:458
Fantasy realization theory 15:889
Fante Confederacy 4:575
FAO *see* United Nations Food and Agricultural Organization (FAO)
FAP *see* Family Assistance Plan (FAP) 1969; Fixed action pattern (FAP); Functional analytic psychotherapy (FAP)
FAPE *see* Free appropriate public education (FAPE)
Faqīh 12:746t
Far East, confederation 4:574–575
Al-Farabi, Abu Nasr, works, *On the Perfect State* 2:144–145
Faraday, Michael 18:102, 24:153
FARC *see* Revolutionary Armed Forces of Colombia (FARC)
Faris, Robert E.L., works, *Handbook of Modern Sociology* 7:585
Farish, Matthew, works, *The Contours of America's Cold War* 11:20
Farkas, George, on cultural capital 5:378
Far-left, in United States 16:252
Farm ecology, peasant 17:677–678
Farm economics 1:485
Farm Security Administration (FSA) 18:84
Farmer, Paul 1:225
Farmers
affluent 11:895
peasants and 17:669
small 1:461
and social mobility 22:427
Farmers' markets (FMs) 1:566–567, 1:570–571
Farming 1:566–571, 16:838
clan 16:838–839
and colonialism 1:458
emergence in Europe 8:270
industrialization of 1:461
Farming communities 23:66
Neolithic 23:66

in Southeast Asia 23:66
in Thailand 23:66
Farming populations 11:7
Farming systems 1:485–487, 17:673–674
horticulture 11:205–207
research 1:472, 11:206
Farms 1:477–479
agricultural labor 1:478
credit markets 1:478–479
input markets 1:478–479
land tenure 1:478
marketing 1:479
organization of production 1:478
price determination 1:479
risk management 1:479
rural sociology 20:814–817
Faroqhi, Surayah, works, *Towns and Townsmen of Ottoman Anatolia* 16:382
Farr, William 6:171–172, 17:28, 22:697, 23:426
Farrell, Michael P., on collaborative circles 5:468
Far-right, in United States 16:248–249
Farrington, David 14:61
integrated cognitive antisocial potential theory 14:61–62
FARs *see* Floor-area ratios (FARs)
FAS *see* Fetal alcohol syndrome (FAS)
Fa-scheme (record statistics) 16:902
Fascism 4:753, 6:369, 6:373, 6:798, 15:771–773, 16:257–261, 18:410–413, 18:434–435, 18:438–439, 21:371–372, 24:448–452
Italian 22:444–445
political thought 18:483
and war 25:367–368
Fascist dictatorship 6:373, 16:257
Fascist Great Council 16:257
Fascist movements 16:261
European 16:260–261
structural elements 16:261
Fascists 22:445
Fashion 8:835–839
ambivalence toward 8:838
class 8:836–837
cycles 8:838
in different societies 8:835–836
diffusion 8:835
fad 8:835, 8:838
garment 8:835, 8:838
gender 8:836
identity 8:838
innovation 8:835, 8:837
mode 8:835–836
modernity 8:835
as process of change 8:836–838
as break with convention 8:837
inner dynamic 8:837
production 8:838
as psychological and social expression 8:837–838
social differentiation 8:836–837
system theories 8:837
sociology of 2:357, 2:360–361
style 8:835–838
trend 8:835–837
as unpredictable 8:838
Fashion industry, and cultural objects 5:536
Fashion race 8:835
Fashion victims 8:838
Fast algorithms 1:531–533
Fast and frugal algorithms 1:531–532
Fast optical imaging 16:389
Fast shadowing 21:623
Fast-and-frugal heuristics 5:911
Fastnacht carnival 3:146
Fat, and food texture 24:29
Fatal impact approach 17:161
Fatalism
in Russian grammar 13:289–290
in West Africa 1:302
Fatalistic suicide 23:683
Fatalities, of natural disasters 16:315–316
Fate restriction 2:823
Father involvement
in dual-career families 14:609
in intact families, quantitative and qualitative assessments of 22:866–867
motivation for and barriers to 8:841
Father relation 13:54
Father–child consensuses 12:405
Father–child relationships 8:803
and divorce 14:521
Father–daughter sexual relationships 11:727–728
Fatherhood 8:840–843
accessibility 8:841
and achievement 8:842
behavior problems 8:842
changes in 8:840
changes in living arrangements of children and parents 8:841
child care 8:840–841
engagement 8:841
gender division of labor 8:840
gender revolution 8:840
household work 8:840
motivation for and barriers to father involvement 8:841
nonresidential fathers 8:841–842
paternal influences on child development 8:841–842
responsibility 8:841
and second demographic transition 8:840
stepfathers 8:840–842
time use 8:840–842
transition to 8:841
Fathering 8:844
child protection 8:844–845, 8:847
family support 8:847
interventions 8:844–847
masculinity 8:847
programs 8:844–847
social work approaches to 8:845–847
critiques of current practice 8:845
evidence of effectiveness 8:846–847
initiatives to improve engagement of fathers 8:845–846
social/cultural context of 8:844–845
Fatherland, and French Revolution 25:364
Fathers
and child neglect 3:412–414
diaries of infants' early life 16:482–483
engagement, initiatives to improve 8:845–846
parenting of 2:475
role in gender development 9:743–744
Fathers Matter projects 8:846
Fatigue
and leisure 13:850
rheumatoid arthritis 20:656
Fatima (daughter of Muhammad) 12:731
Fatwá 12:746t
Fatwa-e-Alamgiri (sharia law) 23:59
Faujdar 23:59
Faulkner, Robert
on assortative matching 2:55
on careers of artists 5:468
Fault liability
torts in law and 24:431–433
Fault line wars 8:92
Fault-based approach to enforcement 13:573
Fauna, zooarchaeology 25:849–853
Faunalturbation 10:54
Faurisson Affair 3:529
Faux queens 13:875
Fava, Sylvia 23:664
Favelas 4:525–526, 4:528
Favoritism 5:17–18
Fawcett, H. 14:620–621
Fazah, Ziad 16:66
FBI *see* Federal Bureau of Investigation (FBI)
FBT *see* Family-based therapy (FBT)
FC *see* Fear conditioning (FC)
FCC *see* Federal Communications Commission (FCC)
FCP *see* Five-City Project (FCP)
FCTC *see* Framework Convention on Tobacco Control (FCTC)
FDA *see* Food and Drug Administration (FDA)
FDCs *see* Fixed-dose drug combinations (FDCs)
FDES *see* Framework for the Development of Environment Statistics (FDES)
FDI *see* Field Dependence-Independence (FDI); Foreign direct investment (FDI)
FDR *see* False discovery rate (FDR)
Fear 1:805, 16:494, 16:497–498, 18:40–42, 21:871, 24:436f
as an emotion 8:868
anthropomorphic (psychological) sense of 8:868
and cardiac defense 23:543–544
and classical conditioning 3:764–769
contemporary fear research 8:870–872
definition and background 8:868–869
diabetes-related 6:339
and disgust 6:546–548
and food preferences 9:299
and insecurity 22:369
liberalism of 21:891, 21:894–895

Fear (*continued*)
models of innate 8:869f
neuronal substrates 8:869–870
pathologies and clinical implications 8:872–873
psychological and neural aspects 8:868–874
Fear appeals 22:353
Fear arousal, and attitude 2:171
Fear circuitry disorders 25:106
Fear conditioning (FC) 2:97, 8:849–853, 8:870–871
amygdala 1:811–815
and anxiety disorders 8:852–853
basic terms 8:849
brain regions implicated in different phases of 1:811–813
acquisition 1:811–812
consolidation 1:812
expression 1:812
extinction 1:812–813
reconsolidation 1:813
definition 1:811
epigenetic factor 8:852
extinction memory 8:850–852
fear conditioning in animals 8:849
future research 8:852
genetic factors 8:852
hippocampus 1:811–815
in humans 8:849–850
importance 1:811
medial prefrontal cortex 1:812–815
neural basis of 1:811–817
neural mechanisms and modulators implicated in 1:813–814
understanding anxiety using 1:811
Fear learning 8:868
Fear of crime 4:326, 5:145–146, 5:148, 8:854–859, 13:750
changes in research focus 8:855–857
indirect victimization 8:856–857
neighborhood factors 8:855–856
personal characteristics 8:855
vulnerability 8:855–856
community 8:856–858
current and emerging trends in research and theory 8:857
disorder 8:856–858
importance and definitions 8:854
changes in measurement 8:854
patriarchy 8:855–856
perceived risk 8:854–855
researchers 8:856
shadow of sexual assault thesis 8:855
socialization 8:855
subcultural diversity 8:856–858
Fear of failure
cause for sexual inadequacy 21:689–690
and procrastination 19:76–77
Fear of isolation 22:346
Fear of legal sanctions 13:750–751
Fear of recurrence, of cancer 3:73
Fear of strangers 25:802–803
Fear of the masses 14:744–745
Fear on civil war 3:716
Fear responses, children, effects of media use on 15:37
Fear-potentiated startle 8:860–867
amygdala 8:861–865
anxiety disorders 8:860
bed nucleus of the stria terminalis (bnst) 8:862
behavioral genetics 8:863–864
conditioning 8:860–864
in humans 8:862–864, 8:863f
interoceptive threat 8:863–865
neural networks of 8:860–862, 8:861f
panic disorder 8:865
phobia 8:864–865
potentiation of mammalian startle response 8:860
using to study to clinical problems 8:864–865
Fear-related behaviors 8:871
Feasibility hypothesis 3:716–717
Feasibility of informed consent 3:902–903
Feasibility Standards 19:119
Feasting
and chiefdoms 3:378–379
and paleoethnobotany 17:460–461
Featural face processing 8:652–654
Featural models 21:581
Φ-feature 10:316–317
Feature comparison models of semantic processing 21:572–573
Feature detectors 25:185
Feature extraction, face 8:659–661
Feature geometry 18:78
Feature matching, face 8:661
Feature selection 16:119
in pattern recognition process 23:411
Feature space 23:411, 23:413
Features (linguistics), and prototypes 19:287–288
Febrile catatonia, and schizophrenia 21:20
February Regime ('dual power') 20:856–857
February Revolution 20:856
Febvre, Lucien 1:708–712, 5:421, 10:219–220, 11:40, 20:326
works
Ecole des Hautes Etudes en Sciences Sociales 11:40
Economies, Sociétés, Civilisations 11:40
Maison des Sciences de l'Homme 11:40
The Problem of Unbelief in the Sixteenth Century: The Religion of Rabelais 11:40
FEC *see* United States Federal Election Commission (FEC)
Fechner, Gustav Theodor 8:566–567, 10:756, 13:690–691, 14:810, 19:412, 19:470–472, 21:604, 23:453–454
on mathematical psychology 14:800–801
on theory of discrimination 19:473
on theory of mental judgment 19:470–472, 19:471f
works
Elemente der Psychophysik 19:470–472
Vorschule der Ästhetik 16:657
Fechner's Law 23:453–454
Fecondability 10:776–777
Fecundity of female maternal relatives of homosexual men 21:783
Federación Estudiantil Universitaria (FEU) 5:368
Federal and Medical Leave Act 1993 14:607–608
Federal antidiscrimination law, in United States 6:530–533
Federal Arbitration Act 1925 3:734–735
Federal Bureau of Investigation (FBI), US 5:127, 5:165, 5:194, 5:235, 19:733
Joint Terrorism Task Forces led by 16:250
National Instant Criminal Background Check System 10:459
Uniform Crime Reports (UCR) 19:837
Federal Child Abuse Prevention and Treatment Act 10:719–720
Federal Committee on Statistical Methodology 10:304
Federal Communications Commission (FCC), US 5:705, 7:252, 24:169
Federal constitution 1:612
Federal Contract Compliance Program Regulations 1:257
Federal convention 1:612–613
Federal Express 18:642
Federal funding, academic associations 22:664
Federal Housing Administration (FHA) 12:140, 16:451
Federal law shareholder suits 4:934–935
Federal Narcotics Bureau, US 15:590
Federal Policy for the Protection of Human Subjects *see* Common Rule
Federal recognition of neighborhoods 16:452
Federal Register 13:819
Federal Republic of Germany 4:764
Federal Reserve Bank
Philadelphia Survey of Professional Forecasters 8:515
Federal Reserve System 15:730
Federal Rules of Civil Procedure 6:516–517
Federal Rules of Evidence (FRE) 702 8:583
Federal Rules of Evidence (FRE) 706 8:582
Federal system 8:876
Federal Trade Commission (FTC) 5:633
Federal Trade Commission (FTC) Act 1:797
Federal welfare state 18:733–734
Federalism 4:570–571, 7:267, 8:875–877, 12:900, 24:908–913
defined 8:875
empirical study 8:875–877
causal approaches 8:876–877
descriptive approaches 8:875–876
normative study 8:875
proponents 8:875
and related terms 8:875
Federalization, corporate governance 4:937
Federal-State Cooperative Program for Local Population Estimates (FSCPE) 1:841
Federal-wide Assurance (FWA) 8:50

Fédération Internationale de Football Association (FIFA) 23:303
Federation of Cuban Women (FMC) 5:372
Federation of European Laboratory Animal Science Associations 1:706
Federation of International Human Rights Museums (FIHRM) 16:146
Federation of Latin American Therapeutic Communities 24:276–277
Fee systems in OECD countries 10:856t
Feeblemindedness 5:233–234, 8:249
Feedback 10:622, 14:755, 21:487, 22:913–914, 24:49
  centrality measures involving 16:533–534
  concept learning 4:654–655
  eyewitness testimony 8:644
  LTG study 14:139
  negative 22:913
  positive 22:913
  reinforcement 20:207–208
  from test administrators 7:932
  from test participants 7:932
Feedback control 3:781
  motor control 3:780–783
Feedback loops 15:946–948, 18:291, 25:158–159
Feedback mechanisms, in diffusion processes 6:414
Feedback of biobank research 2:622
Feedback systems 16:652
Feeder cells 16:634
Feed-forward consistency
  in spelling 23:267
Feed-forward control 3:781
  motor control 3:781–782
Feed-forward internal model 11:325
Feed-forward neural process 25:186
Feed-in tariffs, for renewable energy 18:369–370, 18:372–373
Feeding, regulation by hypothalamus 11:488
Fee-for-service delivery model 21:920
Feel-good activities 21:496
Feeling for the other 7:549–550
Feeling of knowing (FOK) 20:610, 24:89–90
Feeling of others 19:233–236
Feeling of rightness 20:611
Feeling with the other 7:549
Feelingful language 1:237
Feelings
  Dilthey on 6:425–428
  and emotions 7:502–505, 7:526–528, 7:530–531
  health risk perception 10:703
  as information 7:462
    heuristic 10:827
  same-sex youth 9:651
Fee-paying audiences 18:100
Fees 14:267
  in affordable private schools 1:269
  psychiatric ethics 8:66
Fee-shifting
  public interest 12:922
  statutory 12:922
FEF *see* Frontal eye fields (FEF)
Fehr-Schmidt model 19:903
Fehu 9:111
Feigned amnesia 1:655
Fein, H., works, *Genocide: A Sociological Perspective* 10:7
Feinberg, Leslie, on allyship 24:553
Feininger, Andreas 25:191
Feinstein, Alvan 8:327
Feitiçaria 9:105
Feitiço 9:105
Feld, Steven 1:236
Feldschers 25:624
Fellowship of Reconciliation 17:638
Felonies 15:591
  punishment 19:653
Felshin, J., on women's participation in sports 9:657–658
Felt rightness
  retrieval 20:611
Female alcohol drinkers 1:520–521
Female ascetic 3:238
Female bonded primate groups 18:900
Female choice, marriage as 14:615–616
Female circumcision (FGC) *see* Female genital mutilation (FGM)
Female condoms 21:797
Female defense polygyny 14:615
Female Education Society 15:616
Female genital cutting (FGC) *see* Female genital mutilation (FGM)
Female genital mutilation (FGM) 5:493, 8:878–882, 10:192–193, 10:195, 24:578
  clitoridectomy 8:878–880
  cultural explanation for 8:880–881
  extent and distribution 8:878
  as global concern 8:881
  history 8:879–880
  initiation rituals 8:878–879
  operation practices 8:878–879
  and transition to adulthood 24:578
Female honor killing 11:185
Female menstrual cycle 10:462–463
Female migration, by South and Southeast Asians 12:531–532, 12:535–536
Female (zenana) missionary societies 15:616
Female orgasmic disorder (FOD) 21:718
Female physicians 6:615
Female populations 15:608
Female power
  and sexuality, in Africa 10:816–817
  at work 8:883–887
    bottom-up ascription 8:885
    discrimination 8:885
    and gender inequality 8:885
    homophily 8:885
    rise of female managers 8:884–885
    sex segregation 8:885–886
    wage gap 8:885–886
Female primates, reproductive strategies of 18:897–898
Female prisoners 20:62
Female renunciation 25:633
Female sex offending/offenders 21:662
Female sexual debut 14:336
Female sexual orientation 21:779
Female sexual power 10:817
Female time 24:408
Female trickster 24:652
Female virginity 21:793
Female-headed families, and income inequality 11:742
Female–male life expectancy gap 15:865
Females
  contributions in leadership 25:626
  contributions in psychiatry 25:626
  gender differences in cooperation/competition in 9:753
  leadership, and use of force 25:639–640
  in medicine 25:626
  in psychiatry 25:626
  quotas in Parliament 8:888–892
    candidate selection 8:890–891
    numbers elected 8:889
    policy making 8:890
    public attitudes 8:890–891
Female-to-male arrest rates 5:133
Feminine 9:725
Feminine endings 13:275
Femininity 8:893–897, 20:798, 21:678–680, 21:684
  class 8:895–897
    hierarchies of 8:896
  consequences of upholding 8:894
  definition 25:30
  and education 7:118
  learning to be 8:893
  queering 8:896–897
  race 8:895–897
    hierarchies of 8:896
  scripts of 8:893
  and sexual bodies 8:895
  sexual identity 8:895–897
  and sexuality 8:894–895
  subordinated 8:893–894, 8:896–897
  theorizing 8:893–894
Feminism 3:57–59, 3:929, 4:765, 5:74, 5:76, 5:268, 5:930, 8:905–911, 8:926–927, 9:769, 9:781–783, 12:936, 21:179, 22:1–2
  African 8:909
  and anthropology 1:769
  in anthropology 8:898–904
    archeology 8:901
    collaboration and activism 8:902–903
    diversification of 8:899–900
    early frameworks 8:898–899
    embodiment 8:900–902
    emerging issues in twenty-first century 8:901–902
    ethnographies 8:900–901
    gender, history, and intersectionality 8:899
    gender violence 8:901–902
    globalization 8:899–903
    and immigration 8:901
    legacy 8:899

Feminism (*continued*)
new theoretical language, forging of 8:900
and political economy 8:899–900
postmodernism 8:900–901
sexuality 8:899–900, 8:902–903
transnationalism 8:901, 8:903
antipornography 17:220
in archaeological studies 9:764–767
audience research 2:218–219
and bear subculture 13:891
Benhabib on 2:519, 2:522–523
Black and 8:908
body, understanding of 2:720
Buddhism and 25:628–629
and counterculture 5:85
in criminology 8:912–917
approaches and methodologies 8:912–916
intersectionality 8:913
liberal feminism 8:912–914
masculinities 8:913–915
pathways theories 8:915
perspectives, development of 8:913–916
and professional organizations 8:916
radical feminism 8:912–914
standpoint feminism 8:913
critical theory and 5:307
cultural critique and 5:384–385
definition 8:929–930
and delinquency 6:72
and development geography 6:315–317
East Asian studies 6:793–796
ecofeminism *see* Ecofeminism
economics 1:576–577
enculturing 8:905–909
family planning programs 8:794–798
female quotas in Parliament 8:890–891
first-wave 8:906
Frankfurt School 9:385
and gender 9:788–793
and gendered care 9:800–801
and historiography 11:42
and homophobia/heterosexism 10:793
and households 11:231
and identity movements 11:545
and intersectionality 12:658–660
intimate partner violence 12:685–686
Islamic 12:744
and leatherman subculture 13:889
liberal 8:943–945
liberal Christianity 14:9
matrifocality and 14:823
in MENA 16:362, 16:365
and motherhood 15:882–885
Myrdal on 16:159–160
political ecology 6:848
and postmodernism 18:676, 18:678
post–Third-Wave 8:909–910
radical lesbian 8:942–948
religiosity of women 25:635–636
and reproduction 20:450–451, 20:454
science and technology studies 21:203
second-wave 8:907
and sex offenders 21:674–676
and sex work 19:280–282
and social work 5:287–288, 5:292, 22:135–137
social work research 8:935–941
aging 8:938
care 8:939
domestic violence 8:938–939
ethics of care 8:936, 8:939
focus and future of 8:938–939
gender 8:935–940
intersectionality 8:937
methodologies 8:936–938
objectives and methods 8:936
pluralisms 8:936
postmodern feminism 8:935–937
self-reflective research 8:935, 8:937
standpoint feminism 8:935, 8:937
structuralist feminism 8:936–937, 8:939
socialist/Marxist 8:906
and sport 23:314–316
and territory 24:225
Third World 8:909
third-wave 8:907–908
UN 8:906–907
Western 8:906
Feminist activism 23:324, 23:327–328
Feminist anthropology 15:425–426
Feminist Association for Bioethics (FAB) 9:720
Feminist bank teller experiment 12:721–722
Feminist community work 4:379–380
Feminist criminology, and intersectionality 12:663, 12:665–666
Feminist critical theory 9:739–740
Feminist critiques
and gender archaeology 9:764–767
of liberal law 13:439
of privacy 18:950
of sexism 23:326–327
Feminist empiricism 22:1–2
Feminist epistemology 22:1–2
and objectivity 17:100–101
Feminist geographers 21:809–810
Feminist historiography 21:34
Feminist identity 8:905
Feminist interviewing 12:673
Feminist jurisprudence 13:461–462
Feminist legal theory (FLT) 8:918–925, 13:510, 13:792
coverture 8:922–923
difference 8:920
domestic violence 8:921, 8:923
equality 8:919–920
freedom 8:921–922
human 8:919
and international human rights 8:923–924
law 8:920–921
liberal feminism 8:918–921
marriage and social reproduction 8:922–923
method and epistemology 8:918–919
queer theory 8:919–920, 8:922
radical feminism 8:919, 8:921–922
sex and sexuality 8:922
state 8:921
suffrage 8:918
Feminist methodology 19:683
Feminist movement, and de-Westernization 25:545–546
Feminist pedagogy 9:738
Feminist political science 8:926–928
Feminist political theory 8:926–928
and gender science
political science 8:927
and gender studies 8:927
political science 8:927–928
scholarship 8:926
Feminist psychoanalysis, Fromm on 9:460–462
Feminist psychotherapy 8:929–934
assessment and techniques 8:932
core assumptions and principles 8:930–932
definition 8:929–930
ecological context 8:930, 8:932–933
empowerment 8:929–933
future directions 8:933
as integrative approach 8:933
intersectionality 8:929–932
multiculturalism 8:929–933
origins and evolution 8:930
personal is political 8:930–931, 8:933
power analysis 8:931–932
social change 8:929, 8:931, 8:933
social identity analysis 8:931–932
social justice 8:929, 8:931, 8:933
Feminist research
methodology 9:717
politics and scientific rigor 9:718
scientific quality and ethics 9:717–718
validity in 9:718
Feminist revolution, and anthropological writing 1:741
Feminist roots of dilemma 22:2–3
Feminist scholars 8:928
Feminist scholarship 8:905
ethnohistory 8:192–193
Feminist standpoint theory 1:769, 12:658–659, 18:653, 23:324–325, 23:327
Feminist studies
in geography 9:689–693
concepts 9:690–691
methodologies 9:691–692
methods 9:691–692
in professional contexts 9:692
status and emerging directions 9:692–693
themes 9:690–691
see also Women's studies
Feminist theories 13:551, 21:851
in educational sociology 7:290
and gender archaeology 9:765–767
Feminist Therapy Institute Code of Ethics 8:932
Feminist Utopias 14:256–257
Feminist visual theory 25:208
Feminists 6:290, 6:293, 21:808–810
health 15:111

linguistics 9:700
in social work 14:679
standpoint epistemology 6:245–246
study of discourse 9:701–702
Feminities, colonial constructions in Africa 10:813–814
Feminization
and law 13:582
of migration 15:417
of poverty 18:722
of religious leadership 20:411
of work 18:668
Femme subculture *see* Butch–femme subcultures
Femocrats 21:725
Fénelon, François, on Jansenism 12:388
Feng-Shui 24:869–870
Feodalisme 9:111
Féodlité, French 9:111
Feral children 16:342
Feraoun, Mouloud, works, *Le fils du pauvre* 21:752–753
Ferguson, Adam 4:337, 7:722, 14:644, 19:123–124
on division of labor 6:603, 13:161–162
works
Essay on the History of Civil Society 3:701–702, 7:665
Ferguson, C., works, *Language Problems of Developing Nations* 13:353
Ferguson, Niall 11:689
Ferguson, Sir Alex 17:183
Ferm, Vergilius, works, *A Dictionary of Pastoral Psychology* 7:592
Fermi, Enrico 7:622
Fernea, Elizabeth Warnock 9:782
FERPA *see* Family Educational Rights and Privacy Act 1974 (FERPA)
Ferrero, Guglielmo, works, *Ancient Rome and Modern America* 2:270
Ferro, Vito Cascio 14:422
Ferrone, V., works, *Lezioni illuministiche* 7:659
Ferry, Jules 23:2–3, 25:365
Fertility 2:661, 2:697–698, 9:19, 9:40, 9:42–43, 11:5–6, 18:597–599, 21:705, 21:709
action of natural selection on 9:28–29
and adoption 1:137–139
age-specific fertility rate 15:397–399
among traditional humans 9:29–30
bioarchaeology 2:602
biodemography of 9:21
children ever born (CEB) 15:398–399, 15:401–403, 15:401t
in Chinese population
delayed marriage and 10:822
rate in aging population 10:821
of couples with children from previous relationships 20:432
and culture 9:1–5
abortion 9:2–4
aging 9:4
application of sociocultural anthropology to demographic studies 9:1–2
birth control 9:3–4
childbearing 9:3–4
in China 9:3–4
contraception 9:3–4
culture as ideas and mental schemes 9:2–3
culture in flux and contestation 9:3–4
in Islam 9:3
male fertility 9:3
migration and fertility 9:5
operationalizing cultural traits in statistical taxonomies 9:2
parenthood 9:3–4
reconciling local cultural knowledge and Western science 9:4
reproduction studies 9:5
decline in 9:44, 15:611
demography 6:176–179
and disasters 6:105–107
ecology and proximate mechanisms governing 9:31–33
economic determinants 9:60
and economic development 9:95–96
economic growth 7:39, 7:41–43
effects of income support policy 18:745
estimates for 8:791t
in Eurasia 18:547–548, 18:548t
and evolutionary demography 8:407–411
factors influencing pace of decline 9:97–98
family planning 8:786–793
and family size preferences 8:805–807
fertility regulation 2:698
household economics 11:224–225
Hutterite on 11:445
age patterns 11:446
Coale and Trussell's 'M' and 'm' indices 11:446
Coale's 'I' indices 11:445–446
control 11:446
and culture 11:445–446
limitations 11:446
micro- and macrosimulations 11:446
and immigration 15:397
2007 recession in United States 15:403–405, 15:404f
link between 15:397–399
nativity 15:400–401, 15:400t–401t
reproduction 15:397–399
time at destination 15:397–399
in United States 15:400–405, 15:400t–402t
impact of conditional cash transfers on 4:555–559
indirect estimation of 6:145
consistency check 6:145–147, 6:147f
intentions 9:20
macro-level determinants 9:22–24
maternal, and religion 20:318–319
measures 6:110–114
difficulties with TFR 6:113–114
gross reproduction rate (GRR) 6:113
level 6:110–112, 6:111t
net reproduction rate (NRR) 6:113
period parity progression ratios 6:114
tempo adjustment 6:114
TFR and associated indices 6:113
timing 6:112–113, 6:112t
meso-level determinants 9:22
micro-level determinants 9:19–22
under modernizing conditions 9:32–33
at origin and destination 15:399
population cycles 18:559–561
and population dynamics 18:563–564
postsocialist European countries 9:101–103
abortion 9:99–101, 9:103, 9:103f
age-specific birth rates 9:99, 9:102f
causes of decline 9:102
declining total fertility rates 9:101
former Soviet Union 9:99–102
nonmarital births 9:99, 9:102–104, 9:103f
other trends 9:102–103
postponement/permanent decline 9:101–102
preferences 9:21
preservation, in affluent nations 20:473–474
proximate determinants of 10:776
quantitiy–quality fertility model 24:708–709
in repartnerships 20:432
of single and cohabiting women 9:13
conceptual and theoretical challenges 9:16–17
consequences for women and children 9:16
contraception 9:14, 9:16–17
demographic trends 9:13–14
family structure 9:15–16
functionalist, historical, and policy literature 9:14–15
instability 9:16
measurement 9:13
poverty 9:16
pregnancy resolution 9:16–17
sexual activity 9:14–17
social disadvantage 9:14, 9:16
socioeconomic determinants 9:15–16
unmarried parents 9:13–16
and social networks 9:6–9
under socialism in Europe 9:99–101
1945–1989 9:99–100
age at first birth 9:100f
influence of socialist system 9:100–101
pattern 9:99–100
total fertility rate 9:100f
in sub-Saharan Africa 9:94t
total fertility rate (TFR) 15:397–399, 15:401–403, 15:401t–402t
destination-specific (D-TFR) 15:399
traditional foraging societies 9:31–32
trends in East Asia 9:55–56, 9:55f
value of children
decline 3:498–499
transition 3:501
and value of children 3:500–501

Fertility, determinants of
macro-level 9:22–24
attitude change 9:24
cultural continuities 9:24
economic trends 9:23
(un)employment trends 9:23
historical continuities 9:24
policy measures 9:23
value change 9:24
welfare regimes 9:23–24
meso-level 9:22
place of residence 9:22
social capital 9:22
social interaction 9:22
micro-level 9:19–22
biodemography of fertility 9:21
contraceptive technologies 9:24
cultural context of family of origin 9:21–22
economic uncertainty 9:21
education 9:20–21
employment uncertainty 9:21
fertility intentions 9:20
fertility preferences 9:21
gendered division of labor 9:20
human capital 9:20–21
income 9:20–21
intergenerational transmission of behavior 9:21
intergenerational transmission of values 9:21
partner 9:20
partnership 9:20
reproductive technologies 9:24
socioeconomic status 9:21–22
stepfamily fertility 9:20
Fertility and Family Surveys (FFS) 5:737
Fertility change
quantum and tempo 9:10–12
adjustment 9:12
childbearing 9:12
cohorts 9:10–12
distortion 9:10–12
tempo effect 9:12
translation equation 9:11–12
Fertility control 9:95–97
demography 6:178–179
Fertility culture 10:820
Fertility decision-making, role of family in 9:37–38
Fertility decline 9:775–776
diffusion 6:412–413
heterogeneous decline in advanced economies 9:776
Fertility desires 9:78
Fertility level 15:611
Fertility postponement 21:347
Fertility quantum 9:19
Fertility rate, in Euroasia 18:547–548
Fertility recovery, in industrialized countries 9:775–780
childcare 9:777–778
demographic transitions 9:775–776
evidence from time diary data 9:777–778
fertility decline, ideational change, and opportunity cost 9:775–776
gender attitudes 9:777–778
gender division of housework 9:775, 9:777
gender equity 9:777–779
gendered patterns of time use 9:777
heterogeneous fertility decline in advanced economies 9:776
housework 9:776–778
lowest-low fertility 9:775–778
research 9:778–779
second demographic transition 9:775–776
time allocation 9:776
Fertility tempo 9:19
Fertility theory 9:19–27, 9:40–45
embodied-capital theory of life history evolution 9:28–34
theory of intergenerational wealth flows 9:35–39
Fertility transitions 9:73–79
in China 9:46–48
contraception 9:56, 9:58
and culture 9:49–54
anthropology and demography 9:52
contraception 9:51–52
culture as interaction 9:50–51
culture as schemas, attitudes, values, and beliefs 9:51–52
diffusion 9:50–52
economics and culture in institutions 9:49–50
family 9:49–52
intentions 9:50
politics 9:49, 9:51–53
religion 9:51–53
social network 9:50–51, 9:53
social structure 9:49–50
decline to date 9:74–76
by country 9:75–76
decreasing marriage 9:56–57
demographic factors 9:56–57
in developing world 9:92
in East Asia 9:55–59
economic explanations for 9:60–67
approaches 9:60–61
child human capital 9:60, 9:62
child mortality 9:66
demographic transition stages 9:61–62
economic determinants 9:62–63
employment opportunities of women 9:64–66
Europe 9:64–65
global advancement in women's schooling 9:65–66
labor intensity of children and their opportunity cost 9:63
low income countries 9:64–65
value of women's time and cost of children 9:63–64
wages of women 9:63–66
explanation
demographic 9:76–77
in formerly socialist European countries 9:99, 9:101–102
implications for understanding 9:33
induced abortion 9:56, 9:58
in Latin America and Caribbean 9:68
macro economic changes 9:57
in Middle East and North Africa 9:73–79
population aging 9:55, 9:58
population decline 9:55, 9:58
prospects for trends 9:78–79
reduced form models 9:61
reduction in marital fertility 9:56
socioeconomic factors in fertility decline 9:57–58
in South Asia 9:80–83
in Southeast Asia 9:84–91
Brunei Darussalam 9:85–87
Cambodia 9:84–89
contraception 9:87–88, 9:86t
demographic overview 9:85–87, 9:85t–86t
demographic transition 9:88–89
explanations 9:87–88
Indonesia 9:85–89
Lao PDR 9:84–88
Malaysia 9:85, 9:87–89
marriage 9:87–89
Myanmar 9:85–89
Philippines 9:85–89
Singapore 9:84–89
Thailand 9:85–89
Timor-Leste 9:84–88
Vietnam 9:84–88
structural models 9:61
in sub-Saharan Africa 9:92–98
costs of fertility regulation 9:94–95
by country 9:93–94
cultural factors 9:96
delay in initiation 9:94–95
demand for children 9:94–96, 9:98
difference from other countries 9:97–98
Easterlin framework for fertility analysis 9:94–95
and economic conditions 9:96
education, infant and child mortality 9:96
fertility control 9:95–97
future prospects 9:98
onset, current state, and specificities 9:96
onset by place of residence 9:96–97
by region 9:93
stalling 9:97–98
supply of children 9:94–97
women's educational attainment 9:96
total fertility rate 9:55–56, 9:55f
women's educational attainment 9:57–58
women's employment 9:58
Fertility/mortality trade-off 15:846
Fest des Huhnes (Festival of the Chicken) 11:417
Festinger, Leon 19:414
on cognitive dissonance 4:76–78
Festival of cruelty, and punishment 19:655
Festivals
carnival 3:145–149
counterculture 5:83–84
entertainment 7:669–670
gay/lesbian tourism 9:640–641

related to fire 9:199–200
in South America 23:47
Fetal alcohol syndrome (FAS) 1:509, 2:698
Fetal development, formation of brain during 2:823–824
Fetal origins hypothesis 6:753, 11:269
Fetal personhood, in affluent nations 20:470
Fetal programming 7:850, 11:269, 14:67–68, 23:556
and prenatal stressing 23:557–558
Fetal viability 20:470
Fétiche 9:105
Fetish, use of term 9:105
Fetish foot 17:952
Fetish-gold 9:106
Fetishism 4:521, 9:105–110, 17:949, 17:951–952
in Africa 9:105–110
agency 9:109–110
animism 9:106–107
commodity 9:105–107
as countermodel of Enlightenment 9:106–107
as cultural critique 9:107
dialectical critique of 20:204
disavowal 9:105–110
historicity 9:107–110
idolatry 9:105–106
Latour on 9:109–110
misrecognition 9:105–110
and modernity 15:688
naturalism 9:106
object 9:105–110
origins of term 9:105–106
personification 9:107–108
Portugal 9:105–106
reification 9:107–110
as social creativity 9:108–109
sorcery 9:105, 9:107–108
territoriality 9:107–108, 9:110
untranscended materiality of 9:107–108
use of term 9:105
Fetishistic disorder 17:952
Fetish-oaths 9:106
Fetterman, D.M.
works
Empowerment Evaluation: Knowledge and Tools for Self-assessment and Accountability 7:581–582
Empowerment Evaluation Principles in Practice 7:581–582
Foundations of Empowerment Evaluation 7:581–582
Fetus 23:556–559
Fetzer, James H., works, *Glossary of Cognitive Science* 7:590
FEU *see* Federación Estudiantil Universitaria (FEU)
Feudal bureaucracy 16:400
Feudal economy 15:348
Feudal mutation 9:113
Feudal relationships 1:966
Feudal revolution 9:113
Feudal society 1:966–968, 4:339
Feudal Society 11:40
Feudal system 13:568
Feudalism 9:111–116, 11:518–519, 11:896, 19:193–194, 20:259, 22:839
and agrarian political economy 1:457
in artisans 2:47
Bloch's contribution 9:113–114
Duby's contribution 9:113–114
Ganshof's contribution 9:112
Heinrich's contribution 9:112
homage 9:111–112
investiture 9:111–112
legal model 9:111–112
Marxist model 9:112–113
Middle Ages 9:111–115, 15:348–349
mode of production 9:112–113
modern criticism 9:113–115
overview 9:115
peacemaking in 17:639
and production 15:714–715
Reynolds's contribution 9:114–115
sociological models 9:113
vassals 9:114–115
Feudalismus 9:111
Feuds 4:601–602, 4:614–615, 4:618
definition 4:621
and peace/nonviolence 17:625–626
Feudum 9:111
Feuer, Lewis 12:280
Feuerbach, Ludwig 14:643
works, *Essence of Christianity* 14:643
Feyerabend, Paul 9:117–123, 13:133–134, 18:626
on conceptual conservatism 9:117–118
on contextual theory of meaning 9:121
on democratic relativism 9:122–123
on empiricism 9:117–118, 9:120–122
on epistemological anarchism 9:121–122
on falsificationism 9:117–120
on incommensurability 9:120–121
on logical empiricism 9:117–118, 9:121
on philosophy of nature and rise of rationalism 9:122
on pragmatic theory of observation 9:119
on realism 9:120
on scientific philosophy and methodology 9:118
on theoretical pluralism 9:119–120
on Vienna Circle 9:119
Feyerabend, Paul K. 18:524–525
FFA *see* Fusiform face area (FFA)
FFBI *see* Five Factor Borderline Inventory (FFBI)
FFF (Family planning, Female education, and Food supplementation) 10:199
FFFS *see* Fight-flight-freeze system (FFFS)
FFM *see* Five-factor model (FFM)
Fforde, Arthur, Sir 19:526–527
FFPI *see* Five factor personality inventory (FFPI)
FFQ *see* Food frequency questionnaires (FFQ)
FFS *see* Family and Fertility Surveys (FFS); Fertility and Family Surveys (FFS)
FFT *see* Five factor theory (FFT)
FG *see* Fusiform gyrus (FG)
FGAs *see* First generation antipsychotics (FGAs)
FGC *see* Family Group Conference (FGC)
FGDs *see* Focus group discussions (FGDs)
FGF-2 *see* Fibroblast growth factor-2 (FGF-2)
FGFR3 gene 9:1019
FGIs *see* Focus group interviews (FGIs)
FGM *see* Female genital mutilation (FGM)
FHA *see* Federal Housing Administration (FHA)
FHSI *see* Food and Human Security Index (FHSI)
FI *see* Full interpretation (FI)
FIA *see* Fisher information approximation (FIA)
Fianna Fáil party (Ireland) 17:568
Fiber tracts
limbic system 14:127
fornix 14:127
medial forebrain bundle 14:127
stria terminalis and ventral amygdalofugal bundle 14:127
Fiber-optic cables 12:493
Fibers (cerebral cortex) 3:337–338
Fibroblast growth factor-2 (FGF-2) 20:59
Fichte, Johann Gottlieb 4:338
Fiction mode 9:163–164
Fictional ethnography 8:181–182
Fictions 2:527, 16:194, 16:196–197
Bentham's theory 2:527
versus documentary film 6:620
and historiography, conflation of 11:98–99
Fictitious commodity 19:156
Fictitious play 9:576
Fidelity 5:129, 22:13–14
in archiving 1:919
implementation 21:92
principle of 12:840
Fidesz (Fidesz–Magyar Polgári Szövetség, Hungarian Civic Union) 18:415, 18:611
Fiducia 21:401–402
Fiducial invariance 9:124–127
Bayesian analysis 9:124–127
density 9:125–126
example 9:124
parameter value 9:125–126
probability 9:124–126
scalar 9:124–125
Fiducial probability 14:117
Fiduciary duty, corporate governance 4:935
Fiduciary principle 8:65
Fiefs 9:111
Field data 23:95
Field Dependence-Independence (FDI) 4:111
Field education, and social work education 22:778–779
Field effect 9:140, 9:143
Field experiments 9:128–134, 12:908, 19:736–741
in causal analysis 3:247–248
causality 9:129
current and emerging directions 9:131–132

Field experiments (*continued*)
definitions 9:128–129
history 9:132
birth 9:132
late-twentieth century 9:132
second half of twentieth century 9:132
twenty-first century 9:132
for innovation and discovery 9:129
intellectual background 9:129
versus laboratory experiments 9:129
limitations 9:131
methodological problems and solutions 9:131
naturalism 9:128
and social networks 9:132
and social psychology 9:132
in social sciences 19:737–739
types of studies 9:129–131
Field linguist 14:183–184
Field notes 17:533
Field observational research
in anthropology and sociology 9:135–139
canons of validity 9:138–139
ethnographies 9:135–138
history and scope 9:135–136
interviewing 9:137–138
listening 9:136–137
notes and records 9:138
observing 9:137
selection, access, rapport 9:136
Field of feminist technology studies (FTS) 9:723
Field of inquiry 6:611
Field of power 7:393–394
Field of practice, religious 23:860, 23:864
Field of strategic possibilities 9:141
Field research 17:530
Field studies, multisited
see also Multisited field studies
Field theory 9:140–148, 13:913, 22:528
cultural analysis 9:362–363
cultural capital 9:144, 9:146
culture 9:142–143, 9:145–146
culture production 9:140–141, 9:144–147
economy of symbolic goods 9:141–142, 9:145
general theory 9:141–144
organizational fields and strategic action fields 9:146–147
origins and genesis of concept 9:140–141
social classes 9:142–144
social differentiation 9:140
in social network analysis 16:525
social structure 9:141
sociology of literature 9:141–142, 9:145–146
sociology of science 9:143
symbolic revolution 9:142–143, 9:145–146
theory of domination 9:141
theory of state 9:144–147
uses and methodological issues 9:144–146
Field trials, *versus* quasi-experiments 19:736
Field work, informed consent in 12:113
Field-data collection 3:298
Fieldnotes 9:150–151
Fields 21:809–810
and social representations 20:444–445
as space of relations 5:466
theory, of Bourdieu 2:780–781
Fieldwork 16:856, 20:505
African studies 1:278
and anthropology 1:773–776
apprenticeship 1:866–867, 1:869
and common sense 4:2–3
in education 7:147–148
ethics 9:153–155
ethnography 9:153–154
fieldworker intervention in ongoing social activities 9:155
informed consent 9:153–154
protection of confidentiality 9:154
recommendations for resolving dilemmas 9:155
relationships with research participants 9:154–155
ethnography 8:186
methodology 16:195
in multisited field studies 16:104, 16:106–107
from NGOs 16:856–857
in reflexivity 20:88–92
in social and cultural anthropology 9:149–152
colonialism 9:149, 9:151
ethnography 9:149–152
extended case method 9:150–151
fieldnotes 9:150–151
informants 9:149–151
origins of method 9:149–150
participant observations 9:149, 9:151–152
polymorphous engagements 9:151
practice 9:150–151
reflexivity 9:151
symbolic interaction 9:151
transformations in research 9:151–152
in United States 9:149–150
social work research 20:504
survey and excavation in archaeology 23:742–746
urban anthropology 24:780–781
Fieldwork dilemmas, death 5:870–871
Fieldworker 14:183–184
Fieldworks Language Explorer (FLEx) 14:187–188
FIFA *see* Fédération Internationale de Football Association (FIFA)
FIFO (first-in-first-out) 19:786
Fifteen Puzzle 2:32, 2:32f
Fifth Republic 18:867
Fight–flight response 1:804–805, 9:606, 19:445–446
and stress 23:571
Fight-flight-freeze system (FFFS) 8:624, 17:860
Fighting 12:839
Fighting skill 10:522–523
Fighting spirit 7:497
Figlio, Karl 2:730–731
Figurative art 14:566–567
Figurative language 13:402
and pragmatic disorders 18:790–794
purity thesis 15:289–290
Figurative language comprehension 13:372–373
Figurative schemes, social representations 20:446
Figure-ground asymmetries 11:706–707
Figure-ground relations 25:186–187
FIHRM *see* Federation of International Human Rights Museums (FIHRM)
Fiji 17:158–163
Fijian Social Justice Act 2001 1:249–250
Filangieri, G., works, *Scienza della legislazione* 7:661–662
Filibusters 4:651–652
Filipchenko, Yuri Alexandrovich 8:242
Filipino culture 5:615
Filkins, Dexter, works, *The Forever War* 25:361
Fillmore frame theory 13:934
Film genres 9:160–164
and auteur theory 9:160–161
basic film forms and universal formulas 9:163
classical Hollywood 9:163
cognition and emotion 9:162–163
documentary 9:163–164
film theory 9:160, 9:162–163
gender 9:161–162
history 9:162
nonfiction films 9:163–164
reception 9:162
ritual and ideological approaches 9:161
semiotics 9:161
Film history 9:165–170
aesthetics 9:165, 9:168
chronology 9:166–170
conversion to sound cinema 9:168
digital cinema 9:165–166, 9:169–170
early cinema 9:166
historiography since 1970s 9:169–170
Hollywood 9:165–169
motion pictures 9:166–167, 9:169
national cinemas 9:167–168
postwar internationalism 9:168–169
style 9:166–168
technology 9:165, 9:168–170
transition to classicism 9:166–167
Film industry 9:156–159, 15:15
Hollywood 9:156–157
industrial organization analysis 9:158–159
Marxist industry analysis 9:158
neoclassical industry 9:158
rest of the world 9:157–158
Film Optical Scanning Device for Input to Computers *see* FOSDIC (Film Optical Scanning Device for Input to Computers)
Film theory 9:160, 9:162–163
Filmer, Robert 9:388
works, *Patriarcha* 3:61
Filmmaking 9:161
Films

as art and industry 9:165–166
on bear subculture 13:891
communication and transportation 4:273–274
documentary and ethnographic 6:620–624
in Maghreb 21:753
MENA 16:361, 16:364–366
preserving dance 5:669
taxonomy 6:620–621
Filter bubble 22:423
Filter gain 24:334
Filter phase, in time series cycles 24:334
Filtering 24:358–359
Filter–rectify–filter sequence 25:160
FIM *see* Family investment model (FIM)
FIML *see* Full information maximum likelihood (FIML) method
Fin de siècle 12:743, 24:763
Final energy 7:628–630
Final Solution (World War II) 21:371
Final-offer arbitration 6:558–559
Finance anthropology 2:873
Finance capital and industrial production 3:22
Finances 20:16
anthropology of 3:9, 9:171–176
abstraction 9:172–175
capitalism 9:172–175
central assumptions and challenges 9:172
corporations 9:171–172, 9:175
critical approaches 9:172–174
culture 9:171, 9:173–175
economic anthropology 9:174–175
ethnography 9:174–175
financialization 9:171–175
inequality 9:171–175
markets 9:171–175
social studies of finance 9:173, 9:175
Wall Street 9:175
and behavioral economics 2:439–440
communication and transportation 4:273, 4:276
corporate 4:927–931
economic sociology 7:69–70
geography of 9:177–182
development 9:179–180
future research 9:181–182
global flows 9:178
hierarchy of financial centers 9:178–179
mapping financial literacy 9:180–181
pricing tradition 9:178
regional growth 9:179–180
greening of business 10:393–394
and libraries 14:46–47
party 18:393–397
psychology of 15:735–740
public 23:364
of schools 7:155
venture capital 25:59–60
Financial abuse, of elders 7:330
Financial accounting 4:929
Financial analysis, integrated system of 4:928–929
Financial capital, family and schooling 8:753
Financial centers hierarchy 9:178–179
Financial control 4:928
Financial crimes 25:552
Financial crises 9:183–190
banking crises 9:184–185, 9:188–189
bubbles, crises, and financial cycles 9:185–186
central banks 9:183–184, 9:188–189
contagion 9:185
currency crises 9:183–184, 9:189
debasement 9:183–185
debt crises 9:184–185, 9:187–189
emerging markets 9:183–184, 9:186–188
frequency 9:187–189, 9:189f
global financial crisis
causes of 9:194
regulatory response to 9:194–195
Great Depression 9:183–185, 9:187–189
Great Financial Crisis 10:201
impact on economic reform 18:694–695
inflation 9:183–184, 9:188
management of 3:582–583
monetary policy since 15:727–734
sudden stop 9:184–185
types 9:184–185
Financial cycles 9:185–186
Financial decision-making, remittances 12:521
Financial fair play 23:320–321
Financial history 7:53
Financial institutions
and development 6:302
special lending to 15:727–728
Financial instruments 19:572
Financial integration 9:191
Financial literacy mapping 9:180–181
Financial Market Stabilization Act 10:295
Financial markets
application of auctions to 2:209
ethics of 3:18
models of 15:298
Financial planning 4:929–930
Financial practices 12:117–118
Financial regulation 3:32, 9:191–196, 20:169–171
financial governance 9:192–193
financial markets 9:191–193
financial sector 9:191–192
financial services 9:191–193
financial supervision 9:191, 9:193–195
global financial crisis
causes of 9:194
regulatory response to 9:194–195
key concepts 9:191–192
macroprudential supervision 9:195
main open issues in 9:192–193
microprudential supervision 9:195
neoliberal paradigm 9:193
objectives and rationales 9:192
perspectives in academic literature 9:193–194
prudential regulation 9:191, 9:195
prudential supervision 9:191
reform of 9:191, 9:194–195
supervisory authorities 9:191, 9:193–194
Financial sector 9:191–192
Financial security, job displacement insurance 12:825
Financial services 9:191
The Financial Times 1:270, 16:770, 16:859
Financial Transaction Tax (FTT) 8:35
Financialization 3:579–580, 7:65–66, 9:171–175, 9:177, 20:817, 24:851–852
of capital flows 2:750
and culture 5:537
economic sociology 7:69–70
of the economy 24:851
Financialized capitalism 3:98
Financification 10:47–48
Financing
comparative health systems 4:401
health care, psychiatric ethics and 8:66–67
health policy 10:664
Financing decisions 4:927
Finch, Henry 25:465
Findhorn (new religious movement) 16:806
Findhorn Ecovillage project 2:152
Fine, Gary Alan 22:539
on meaning 16:561
Fine art photographers, careers and social networks 5:468
Fine arts, and cultural participation 5:449
Fine-branch niche
primates, evolution of 18:906
Fine-grained analyses 23:251
Fines 13:574–575
Finger Tapping Test (FTT) 16:723, 16:725
Fingerprints 9:955
Finite additivity 5:858–859
Finite frequency interpretations, objective probability 19:40–41
Finite relative frequency interpretation of probability 19:25–26
Finite semigroup, defined 22:479
Finite state grammars 13:83–84
Finite state Markov Chains
mathematical psychology 14:801–802
for social analysis 14:580
Finite state network (FSN) representations of knowledge sources 23:262
Finitely iterated Prisoner's Dilemma 18:934–935
Finitely unique structure 17:299
Fink, J., on lesbian athletes 9:659
Finkelnburg, Carl 16:683
Finland
anthropology 1:754
arts education in 1:231
income inequality in 22:327
relationship between culture and politics in 5:460–461
school burnout and engagement in 21:59–64
sex life in 21:665
declining frequency of intercourse 21:667–669
decreasing frequency of orgasms 21:669

Finland (*continued*)
discussions 21:669–671
gender differences in sexual desire 21:666–667
increasing lack of female sexual desire 21:666–667
method 21:665–666
sexual pleasure 21:669
sexual trends in 21:665
Finlay, Carlos 18:177
Finney, Charles Grandison 2:146
Finnis, John 12:933, 16:320
Finnish
acquisition of reading skills 20:5–9
Finnish Federation for Social Welfare and Health 12:500
Finnish Green League 10:383
Finnish Union of Psychologists, establishment 3:874
Finnish–English language 3:918
FINSEX study 21:665, 21:667
FIPC *see* Free informed prior consent (FIPC)
Fiqh 12:746, 12:752
Fire 9:197–200
and agrarianization 9:197–198
Anthropocene 1:722–723
in cooking 9:197–199
destructivity and 9:199
energy and 9:197–200
festivals 9:199–200
fighting 9:199
and fuel 9:197–200
in industrial society 9:198–199
original domestication 9:197
symbolic and religious aspects 9:199–200
Firearm-related crime 12:466
Firearms
acquisition of 10:457–459
and crime, relationship between 18:880
Firearms behaviors 10:458–459
Firearms laws, and reduction of gun violence 10:459
Firing costs, jobs 12:827–828
Firm creation, and entrepreneurship 7:676–677
Firm perspective 12:582
Firm theory, and property rights 24:529
Firm-as-governance structure 10:842
Firm-based vocational education and training 25:246
Firm-level patenting comparisons 11:838–840
Firms 4:946–947, 15:675–676, 17:302–303
characteristics, and innovation 12:160–162
economic theories of 19:894
in its selection environment 17:366–367
as open organizations 1:104
phenomenology of 17:390–391
regulatory agencies 20:181, 20:185
and trade unions 24:497–501
as venue of social interaction 3:23
First Amendment, US Constitution 6:534, 18:396
First birth, in Latin America and Caribbean 9:69–70, 9:70t
First chemical theories 3:371–372
First council of bishops 3:548
First cousin marriage 11:725–726
First difference filter 24:334
First European civil war 8:274
First fixation duration 8:632
First Forum on China–Africa Cooperation (FOCAC) Ministerial Conference 3:504
First funerals, Hertz's analysis of 5:871–873
First generation antipsychotics (FGAs) 21:22–23
First haircut, life phase 5:508
First Industrial Revolution 11:875
First intercourse 21:705–709, 21:707f–708f
First International Mathematics Study 4:433
First language (L1) 2:580–583, 21:360–361
First language (L1) acquisition 9:206–211
disjunction 9:209
field approaches 9:201–202
parameters 9:206
principles 9:209–210
semantic subset principle 9:209
target-language properties, impact of 9:203–204
Universal Grammar 9:206
universal trends in language development 9:202–203
usage-based theory 9:208
wh-movement 9:208
First language (L1) learners, scope preferences in 13:314–315
First law of thermodynamics 10:756
First marriage 8:767
First memory system 11:386
First moment of law and development 13:443
First National People of Color Environmental Leadership Summit 7:757
First Nations (Canada)
child raising 11:762
First person media 15:62–63
First philosophy 18:22
First recipients of assistance 18:709
First relationship, in Western societies 17:546–547
First responders, for intimate partner abuse 12:681–682
First Revolution of 1905–1906 20:856
First Vienna Circle 25:87, 25:90
First wave 6:459–460
First words of children 13:246–247
First World 6:276–277, 23:349, 23:351
and political economy 1:938–939
First years of life, empathy development during 7:551–552
First-born children, value of 3:499
First-generation antipsychotic medications, for schizophrenia 21:12
First-generation immigrants 5:154
First-generation rights 11:369–370, 20:680–681
First-hand contact, in acculturation 1:68
First-in-first-out *see* FIFO (first-in-first-out)
First-level personality 17:801
First-order asymptotic likelihood theory 14:124
First-order cybernetics 22:912
First-order features, reasoning 5:936
First-order logic (FOL) 5:936, 15:752
knowledge representation 13:100–101
language of 5:933–934
reasoning 5:933–937
First-order model, unit roots time series 24:362–363
First-order visual signal 4:677
First-past-the-post system 7:351
First-personal and social relations, analogies between 21:451–452
First-price auction 2:205
First-terminating or minimum time processing 12:78
First-wave feminism 8:906
First-world's burden 5:585–586
Firth, John Rupert 18:795
Firth, Sir Charles 7:650–651
Firth, Sir Raymond William 1:755, 1:767, 7:145
Firthian linguistics 14:228–229
Fiscal analysis 18:278
Fiscal austerity 16:462, 16:466
Fiscal equivalence, principle of 23:338
Fiscal policies 24:337
Fiscal-military model 16:304
Fischer, Eugen 8:224
Fisch's Capacity Model 7:248–249
Fish, Barbara 25:625–626
Fisher, Irving 9:218–221, 13:16, 14:406–407, 16:228, 17:504
biography of 9:218
causes supported by 9:219–220
distributed lag technique of 9:219
on economic indicators 9:219
on economic reforms 9:219
on economic theories 9:218–219
on Great Depression 9:220
on health reforms 9:219–220
and Keynes 9:220
legacies of 9:218–221
on monetary theories 9:219–220
on recognition (achievement) 9:220–221
Fisher, Margaret (née Hazard) 9:218
Fisher, Melissa, works, *Wall Street Women* 8:902
Fisher, Sir Ronald Aylmer 5:8, 9:222–225, 11:215, 11:494, 13:24, 14:812, 16:824–827, 17:660–662, 21:264, 23:376, 23:397, 23:425
biography of 9:222
choice of test statistic 11:494–495
on distribution theory 9:223
on evolutionary theories 9:223
on exact probability test 2:593
in ethnography 8:190
on fundamental theorem of natural selection 13:24–25
on genetics studies 9:222

on information number, unit roots time series 24:363–364
on kin selection 13:26
on linear discriminant function 16:119
on maximum likelihood statistics 9:223–224
on personality 9:224
on *p*-values 12:800
on research design theories 9:223
on statistical tables 9:224
on statistical theories 9:222, 9:224
works
The Design of Experiments 8:569, 9:223, 22:298
The Design of Field Experiments 9:132
The Genetical Theory of Natural Selection 9:223–224
Statistical Methods and Scientific Inference 9:224
Statistical Methods for Research Workers 9:223, 22:298
Statistical Tables for Biological, Agricultural, and Medical Research 9:224
Theory of Inbreeding 9:223
Fisher index 5:753–754, 5:756
Fisher information 4:510
Fisher information approximation (FIA) 4:476–477
Fisher information matrix 15:657
Fisherian experimental design model 19:714
Fisher–Kyburg interpretation 19:41
Fishing quotas 9:228
Fishing societies 9:226–229
biology 9:229
category of fishing 9:226
ecological knowledge 9:227–228
ethnography of 9:227
fishing quotas 9:228
gathering 9:226–227
gender roles and identities 9:227
hunting 9:226–227
maritime anthropology coming of age 9:228–229
market approach 9:228
natural and social models 9:226–227
practical skills 9:227
resource management 9:228
traditional knowledge 9:228
Fishlow, Albert, works, *American Railroads and the Transformation of the Ante-Bellum Economy* 7:56
Fishman, J.A., works, *Language Problems of Developing Nations* 13:353
Fission–fusion model 13:48
microevolution 15:339
Fit 12:836–837
social categorization 22:187–188
strategic fit 23:502
perspectives 23:501–502
FIT-Choice *see* Factors influencing teaching choice (FIT-Choice)
Fitness, evolutionary 8:388–389, 8:413–417
versus adaptation 1:85–86
concept of 1:85
as erotic capital 8:2
measures 1:86
and natural selection 1:86
and personality traits 17:899–903
relative 1:86
Fitness, of business environment 17:366–367
Fitness landscapes (metaphor) 24:131
Fitness proportional/roulette wheel selection 2:41
Fitting
of statistical models 23:787
of generalized linear mixed models 9:847–849
Fitting-in ideal, in collectivism 4:208
Fitzgerald, Frances, works, *Fire in the Lake: The Vietnamese and the Americans in Vietnam* 11:85
Fitzpatrick, Peter 13:810
Five Cs of PYD 3:447–449
Five Factor Borderline Inventory (FFBI) 9:244, 9:246
Five factor personality inventory (FFPI) 2:563
Five factor theory (FFT) 16:743, 17:886–887
temperament 24:187
Five Punishments 6:778
Five Stars movement 11:416–417
Five-City Project (FCP) 10:648–649
Five-factor model (FFM) of personality 2:559, 2:561, 4:895, 5:323, 7:225, 9:230–243, 9:1009, 17:766–767, 17:777, 17:787, 17:789, 17:797–798, 17:801–802, 17:821–822, 17:854, 17:871, 17:877, 17:884, 17:914–915, 17:937–938, 21:925, 24:182
agreeableness 17:787–788, 17:802
alternative representations of trait structure 9:232–233
Abridged Big Five Dimensional Circumplex 9:232
Big Seven Factor Model 9:232
HEXACO Personality Inventory – revised 9:232–233
Hogan Personality Inventory 9:233
Revised Interpersonal Adjective Scales: Big Five 9:233
alternatives 2:562
brief measures 9:231–232
Big Five Inventory – 10-Item Version 9:232
Big Five Inventory – 44-Item Version 9:232
Ten-Item Personality Inventory 9:232
circular characteristics 2:563
comparative validity of personality inventories 9:234
conscientiousness 2:559, 2:561, 17:787–788, 17:802
Costa and McCrae works on 9:231
cross-cultural findings 2:562–563
dimensional characteristics 2:563
and educational outcomes 17:787–789
emotional stability 2:559, 2:561, 17:787
extraversion 2:559, 2:561, 17:787, 17:802
on facets 9:236–237
on factors 9:236–237
future directions 2:564
Goldberg on 9:230–231
hierarchical characteristics 2:563
important facets
how many? 9:237–238
what facets? 9:237–238
individual differences 9:233
intellect/openness to experience 2:559–561
lexical approach 9:230–231
lexical hypothesis 2:559–560
and lower-level constructs 2:561
measures derived from International Personality Item Pool 9:233
NEO Personality Inventory, revised (NEO-PI-R) 9:231–234
neuroticism 17:802
nomological network 2:563–564
openness 17:787–788
openness to experience 17:802
paradigm 2:559
for personality disorder 9:242–248
absence of normal traits 9:246
advantages 9:244–246
diagnosis 9:244
dimensional trait model 9:242, 9:244–247
DSM-5 9:246–247
homogeneous trait constructs 9:245
improved construct validity 9:246
inclusion of normal, adaptive traits 9:245–246
individualized and precise description 9:245
limited number of traits 9:247
limited recognition of bipolarity of maladaptive personality structure 9:246–247
personality facets 9:238–239
heritability 9:238
personality development 9:238–239
self-observer agreement 9:238
personality psychology 2:559
of personality traits 17:922–923
practical usefulness 2:563–564
prediction of outcome criteria 9:239–240
contribution to behavior understanding 9:240
correspondence in 9:240
facets beyond five factor model 9:240
predictive validity of facets 9:239–240
specific variance in facets 9:239
psycholexical approach 2:562
questionnaire tradition 9:231
shared model of personality 2:560
structured interview 9:233–234
taxonomy 2:559
universality of 9:249–253
country profiles 9:250, 9:252
cross-cultural psychology 9:249
emic–etic approach 9:251–252
indigenous psychology 9:251–252
psycholexical studies 9:250–251

Five-factor model (FFM) of personality (*continued*)
questionnaire studies 9:250
see also Big-Five dimensions of personality
Fixations (vision) 3:344, 8:633
Fixed action pattern (FAP)
of emotional expression 7:458
Lorenz on 14:367
Fixed and random effects 19:734
Fixed capital 3:90
Fixed coefficients 16:49, 16:53
estimates of 16:50t
Fixed costs 14:550, 14:552–553
Fixed density function on processing time 12:79
Fixed effects 19:727
statistical procedures
meta-analysis 15:273
Fixed effects models 14:164, 23:204, 23:382
for class size 3:753
for divorce 14:518–519
hierarchy models 10:836–840
linear mixed models 14:162, 14:164, 14:166–167, 14:167t
Fixed fee schedules 13:630
Fixed internal sampling 23:188f
Fixed level testing, for significance 21:960
Fixed motor pattern 14:367
Fixed role therapy 4:730
Fixed unit effects model, for crossover designs 5:353
Fixed-dose drug combinations (FDCs), for AIDS 1:492
Fixed-effects estimates 16:49–50
of demographic variable impacts 24:705–708
Fixed-group measures of tolerance 18:491–494
Fixed-marginal models 19:431–432
Fixed-point account, of mutual beliefs 21:877–878
Fixed-price-ratio bimetallism 14:624
Fixed-term contracts 1:27–28
FK *see* Foreground knowledge (FK)
FL *see* Foreign languages (FL)
FLA *see* Fair Labor Association (FLA)
FLACSO *see* Facultad Latinoamericana de Ciencias Sociales (FLACSO)
Flaherty, Frances 18:83
Flaherty, Robert 6:621, 18:83
works
Moana 6:620
Nanook of the North 6:621, 25:181–182
Flâneur, concept of 24:780–781
Flash floods 6:480
Flash teams 24:106
Flashbulb memories 4:196, 9:254–259
accuracy, consistency, and challenges to special mechanism 9:255–256
and assassination of John F. Kennedy 9:254
Challenger disaster 9:255–258
consequentiality 9:256–257
emotional valence 9:256
first theory 9:254–255
narrative and rehearsal 9:257
9/11 attack 9:255, 9:257–258
novelty and surprise 9:256
recollection 9:255
and resignation of Margaret Thatcher 9:257
significance 9:257
social identity 9:258
sociocultural factors 9:257–258
vivid memories 9:254–256, 9:258
Flat social ontologies 14:416
Flattened hierarchy 24:274
Flattening of affect 19:291
Flattery of science 21:179
Flaubert, Gustave, on anti-Oriental bias 17:426
Flavell, John 18:123
Flavor
food 24:29
representation, cognitive and selective attentional effects 24:29–30
Flavor network 5:561
FLCs *see* Foreign legal consultants (FLCs)
FLD *see* Frontal lobe dysfunction (FLD)
Fleck, Ludwik 7:875
works
Genesis and Development of a Scientific Fact 21:309
The Structure of Scientific Revolutions 21:309
Flemish Interest 18:610–611
Fletcher, Garth 8:402
Fleury, Michel 10:776
FLEx *see* Fieldworks Language Explorer (FLEx)
Flexibility
and economic restructuring 7:62
of global capital 18:724–725
and perceptual learning 10:132–133
versus plasticity 7:138
at work 8:761
Flexibilization 24:848–849
Flexible accumulation 18:671
Flexible citizenship 6:796, 17:24
Flexible constitutions 4:700
Flexible constraint-satisfaction problem variants 4:719
Flexible discriminant rules 16:118
Flexible habitus 22:335
Flexible learning environments 23:213
Flexible lexemes 17:557–558
Flexible manufacturing technologies 14:466
Flexible space, in open education 17:227
Flexible specialization 11:869–870
and artists 2:54–55
Flexible work, family policy 8:803
Flexicurity 13:186
Flexion reflex 10:482–483
Flexner, Abraham 15:81, 25:624
Flexner Report 8:332, 15:81, 15:83–84, 16:910
Flight
of refugees 20:112–113
of talents 10:252–253
Fligstein, Neil
on conceptions of control 5:537
power theory 17:391
works
The Transformation of Corporate Control 14:475
Flings 17:489–490
Flip-flop model of regulation of sleep–wake 22:91–92, 22:91f
FLMP *see* Fuzzy Logical Model of Perception (FLMP)
Flocculus, cerebellar flocculus 25:71
Flood geology 20:287
Flooding
classical conditioning 3:767–768
coastal flooding 6:480
flash floods 6:480
ice jam flooding 6:480
overbank flooding 6:480
Flood-prone lands, special issues in 13:235–236
Floor-area ratios (FARs) 20:14
Floralturbation 10:54
Florida, Richard 24:867
Florida Affect Battery 8:678
Flourens, Jean Pierre 16:682
works, *Examen de la phrénologie* 16:682
Flourishers 2:867
Flournoy, Théodore, works, *Les Indes à la Planète Mars* 12:903
Flow (phenomenon) 9:260–265
autotelic personality 9:262
concept, applications of 9:264
emergent motivation, learning, and development 9:261–262
experience sampling method 9:261–263
intrinsic motivation 9:260–263
in leisure 9:263
measurement 9:262
model 9:261, 9:261f
optimal functioning 9:262
potential of overflow 9:263
psychophysiology of 9:264
social flow 9:263–264
state, essential components of 9:260t
in work 9:263
facilitation 9:263
Flow analysis 22:465–466
Flow equilibrium, of structural unemployment 24:740
Flow map 6:715–716
Flow theory, of vocational interests 25:241–242
Flow-betweenness centrality 16:536
FLT *see* Feminist legal theory (FLT)
FLTFLT *see* Feminist legal theory (FLT)
Fluency
disorders of 9:266–273
of recognition memory 20:39
Fluency heuristic, of decision making 5:920
Fluharty Preschool Speech and Language Screening Test II 3:795
Fluid comparison 4:413
Fluid historical geographies 11:20–21
Fluid intelligence (Gf) 12:285, 12:317, 12:324–325, 14:101
Fluid Reasoning Index (FRI) 12:284

Fluidity
in organizational development 6:299–300
Fluoxetine 1:779, 1:781
for obsessive-compulsive disorder 17:125
Fluvoxamine 1:779, 1:781
for obsessive-compulsive disorder 17:125
Flux
culture in 9:3–4
recognition by indigenous peoples 11:804–805
Fly in–fly out workers 2:243
Flying geese model 11:837–838
Flynn effect 12:311
FM *see* Frequency modulation (FM) radio broadcasting
FMC *see* Federation of Cuban Women (FMC)
FMLA *see* Family and Medical Leave Act (FMLA)
fMRA *see* Functional magnetic resonance imaging adaptation (fMRA)
fMRI *see* Functional magnetic resonance imaging (fMRI)
FMRP *see* Fragile X Mental Retardation Protein (FMRP)
FMs *see* Farmers’ markets (FMs)
FN *see* Front National (FN)
fNIRS *see* Functional near-infrared spectroscopy (fNIRS)
FNT *see* Face-negotiation theory (FNT)
FO *see* Future orientation (FO)
FOCAC *see* First Forum on China–Africa Cooperation (FOCAC) Ministerial Conference
Focal brain lesions, effects on serial order memory 21:639–640
Focal concepts identification, in EITMs 18:455–456
Focal concerns, in sentencing 21:627–628
Focal concerns perspective, race–crime relationship 19:843
Focal curve 19:781
Focal points 12:191
Focal retrograde amnesia 24:559
Focal seizures 7:852
occurrence of 13:34
Focal vision 23:293–294
Focus (linguistics) 12:95
and science 22:688–689
Focus group discussions (FGDs) 7:931
Focus group interviews (FGIs) 16:406
Focus groups 4:727–728, 9:274–279, 10:590, 12:672, 22:545
analysis 9:278–279
conducting session 9:277–278
criticisms 9:275
definitions 9:274
development 9:274
ethics 9:275
logistics 9:277
mixed-integrated methods 9:275
probing 9:278
qualitative developmental psychology 19:671
in qualitative health research 10:693–694
reporting 9:279
rigor 9:277
virtual focus groups 9:274–275
Focus markers 23:892
Focus of attention, proactive interference in 19:5
Focused information criterion 15:648–649
Focused interview 5:851–852, 12:672
Focused measures, spatial association as 23:103
Focused motivation 7:276
Focused organizational climate 17:328–329
Focusing, in clinical interventions 15:910
Focusing-oriented, experiential psychotherapy 20:740–741
FOD *see* Female orgasmic disorder (FOD)
Fodio, Shaykh Uthman dan 12:738
Foerster, Otfrid, neuropsychological rehabilitation 2:814
Fogel, Robert W.
works
Railroads and American Economic Growth 7:56
The Reinterpretation of American Economic History 7:56
Time on the Cross: The Economics of American Negro Slavery 7:57
FOI *see* Freedom of information (FOI)
FOIA *see* Freedom of Information Act (FOIA)
FOK *see* Feeling of knowing (FOK)
Fokolare (new religious movement) 16:807
FOL *see* First-order logic (FOL)
Folding space 17:304–305
Folk 9:281
Folk antisemitism 19:853
Folk approaches, for mental health 5:572–574
Folk art
ceramics in archaeology 3:324–325
Folk biology 8:392–393
Folk crafts production for national market 5:111
Folk devil 21:673
Folk knowledge 22:972
about science 21:230–231
Folk physical play 14:719–721
Folk physics 8:392–393
Folk psychology 2:201, 8:392–393, 8:393f
Folk religion 3:680–681
of shared customs 20:368–369
themes 20:369
Folk societies, peasants 17:676–677
Folk theory 21:269
of mind 2:203–204
Folkhem 22:889
Folk-life studies 17:670
Folklig 22:889
Folklore 9:280–284, 22:887–888
and anthropology 1:752–754
carnival 3:147–148
ethnology 9:280, 9:283
face-to-face interaction 9:281, 9:283
fairy tales 9:281
folkloristics 9:280–282, 9:284
history 9:280–281
approaches 9:281
discipline 9:280–281
terminology 9:280
Japanese 12:794
keywords 9:281–283
art 9:283
folk 9:281
genre 9:282
performance 9:282
text 9:282–283
tradition 9:281–282
nation building 9:280, 9:283–284
oral tradition 9:280
and politics of culture and heritage 9:283–284
ritual 9:280–281, 9:283
and rumors 20:788
textualization 9:283
Folkloristics 9:280–282, 9:284
Folklorists 20:791–792
Folk-urban continuum 24:508–509
Follicle-stimulating hormone (FSH) 10:462–463, 10:465–467
Followers, and leadership 13:648–652
Followership, authentic 2:248–249
Following, concept of 16:104, 16:106–107
Follow-up studies, psychotherapy 19:377–378
Folsom Europe 13:888
Folsom Street East 13:888
Folsom Street Fair 9:640–641, 13:888
Fon, of Benin 24:654
Fondation Jean Dausset Centre pour l'Etude du Polymorphisme Humaine (CEPH) 11:308–309
Fondation Nationale des Sciences Politiques 22:669–670
Fondation Thiers scholarship 2:703
Fontenelle, Bernard Le Bovier de, works, *Digression sur les Anciens et les Modernes* 15:682–683
Food
affordable, securing 10:389
in anthropology 9:290–295
authenticity 9:292–294
culinary tourism 9:293–295
culinary traditions 9:292–293
cultures 9:290–292, 9:294–295
ethnic foods 9:290–295
future directions 9:294–295
gastronomy 9:292–294
globalization 9:292–293
modernity 9:293
national and regional cuisines and identities 9:293–294
prevailing ethnographic and anthropological approaches 9:290–291
sharing 9:290
and disgust 6:546–548
flavor 24:29
future of 23:810
history and sociology of 9:311–315
cooking 9:313–314

Food (*continued*)
discourse 9:314–315
nutrition 9:312–313
representations 9:314–315
symbols 9:314–315
organic 22:94
pleasantness of taste of 24:27–28
Slow Food movement 22:94–99
variety 11:421
work anthropology
collection 25:663–664
production 25:664
Food access 9:307
Food accessibility 9:285–288
Food addiction, wanting/needing 25:354
Food and Drug Administration (FDA), US 2:623, 4:253, 20:179, 21:214
Food and Human Security Index (FHSI) 10:80
Food availability 3:869
Food balance sheet approach 9:307
Food consumption 9:313
and social networks 16:558
Food course 10:81
Food deserts 22:389
definition 9:286
in United States cities 9:285–289
Food environment, weight reduction 25:468–469
Food forensics 9:1005
Food frequency questionnaires (FFQ) 10:588–589
Food geographies 10:76–82
agri-food knowledge production 10:78
assemblage 10:76–77, 10:80–81
Australasian Agri-Food Research Network (AFRN) 10:76–77, 10:79, 10:81
Biological Economies (BE) 10:76–77, 10:80–81
conceptual reframings of food relations 10:77–78
development 10:79
enacting difference from knowing difference 10:79
enactive research 10:79–81
following as a methodological intervention 10:79–80
food course 10:81
food politics 10:76, 10:78, 10:80–81
food sovereignty and a radicalizing and expansionary initiative 10:79
healthy gendered foodscapes 10:80
hidden labor 10:78
institutional innovation for food security 10:79
La via Campesina (*LVC*) 10:79
ontological politics 10:80–81
reassembling how we know foodscapes 10:80–81
situatedness 10:79–81
Food insecurity 9:285–289, 9:307
Food intake, effects of variety on 24:27–28
Food justice 1:570
Food policy 1:482
Food politics 10:76, 10:78, 10:80–81
Food preferences
psychology/physiology of 9:296–299
acquisition 9:297–298
appetite 9:296
attitudes 9:297–298
basic relationship between humans and food 9:296
determinants of human food choice 9:297
disgust 9:297–298
evolution 9:296, 9:298
family influences and family paradox 9:298
fear and pleasure 9:299
hunger 9:296
intake 9:298–299
moralization and vegetarianism 9:298
nature and context of food choice 9:298–299
nutrients 9:296–297
Food production 17:459–460, 20:798
in Africa 9:301
agriculture 9:301–304
in Americas 9:302–303
Central and South America 9:303
Eastern North America 9:303
Mexico 9:303
southwestern United States 9:303–305
AMS radiocarbon dating 9:304
domestication 9:300–305
in East Asia 9:301–302
ecological engineering 9:303–304
goats 9:300–301
grasses 9:300–303
Holocene on 9:300–302, 9:304
legumes 9:300–301, 9:303
llama 9:303
PaleoIndian 9:303–304
pigs 9:300–303
sheep 9:300
in South Asia 9:301
in Southeast Asia and islands of Southwest Pacific 9:302
in Southwestern Asia 9:300–301
sustainable agriculture 23:807–810
upper Paleolithic 9:300–301, 9:303–304
zooarchaeology 9:304
Food quality, consumer awareness of 10:151–152
Food régime 9:312–313
Food rejections 9:297
Food relations, conceptual reframings of 10:77–78
Food responsiveness, weight reduction 25:470, 25:473
Food safety 1:484–485, 20:798
consumer policy 4:734, 4:736–737
Food security 9:307–310, 9:972, 13:241, 13:243
in Africa 1:293–294
and climate change 3:849–850
institutional innovation for 10:79
sustainable agriculture 23:808–809
Food sovereignty 1:485–486, 10:79, 17:667
Food systems 1:484–485, 1:487, 1:566, 1:568–571, 22:256
Foodscapes 10:80
reassembling 10:80–81
Foot gestures 1:875–876
Football, same-sex sexualities in 20:873
Football Association (FA) 23:303
Foote, Caleb 13:516–517
Footing (communication) 14:174
Foot-soldier jobs 2:52–53
Forager-Collector continuum model, of hunter-gatherers 11:425
Foragers 4:600–602, 11:428, 11:430
affluent 11:895
of East Africa 6:775–776
Foraging societies, traditional 9:31–32
Foraging theory 25:850–851
human behavioral ecology 11:248
Foraminifera 7:694
Foran, John, works, *Fragile Resistance* 16:382–383
Forbidden unallowed acts 23:23
Force fields-based solution to DoF problem 15:959
Force of mortality 14:89–90
Forced compliance theory 17:942
Forced draft urbanization 11:432
Forced heirship 19:174
Forced labor 22:44–45, 24:512–513
Forced migration 20:106
Forced succession 19:174
Forced-attention paradigm of dichotic listening 6:363–366, 6:364f
Forced-choice personality assessment 17:840–848
absolute judgments 17:841, 17:843, 17:848
advantages 17:841–842
binary preference 17:840–841
classical scoring model and ipsative scores 17:842–843
comparative judgments 17:843–844, 17:847–848
dominance models 17:843–844, 17:847
future research 17:848
graded preference 17:841
ideal-point models 17:843, 17:845–847
item response modeling 17:843
McCloy–Heggestad–Reeve unfolding model for multidimensional ranking blocks 17:846–847
Multi-Unidimensional Pairwise-Preference model 17:845–847, 17:847f
proportional preference 17:841
questionnaire data collection 17:840–841, 17:840t
scaling for responses 17:842–847
single-stimulus format 17:840, 17:842, 17:847–848
Thurstonian IRT Model 17:843–845, 17:844f–845f
Zinnes–Griggs Model for unidimensional pairwise preferences 17:844–846, 17:846f
Forced-choice surveys 5:851, 5:853, 7:418
Forces, police use of 18:255–259

Forces of production 14:651
Forcings, in climate change 3:865
Ford, Edmund Brisco, and Fisher 9:223
Ford, Henry 2:297
  on work process 13:160
Ford, J., works, *Social Class and the Comprehensive School* 7:291
Ford Foundation 1:933–934, 1:952, 9:620, 13:516, 19:556, 19:558, 22:651, 22:659, 22:993
Ford Model T 2:304
Ford Motor Company 19:536
Fordism 2:303, 7:63, 7:687, 10:220–221, 18:724–725, 20:673, 24:865, 25:677
  and postindustrial societies 18:664–666
  and transnational corporations 10:221
Fordist industrialism 2:750
Fordist method of production 15:685
Fordist movement, and political economy 1:938–939
Forecasting 9:316–323, 24:358–359
  budgeting 2:894
  expectations 8:510–516
  formal forecasting models 9:319–321
  leading indicator approach to 9:316–317
  limitations 9:322–323
  methods of 9:316
  perception routines for 23:425
  of political data time series 18:515–518
  population 18:593–596
  qualitative trend analysis methods 9:321–322
  quantitative trend analysis and extrapolation 9:317–319
  small areas 6:162
Forecasting geographic models 14:788
Forecasting models 24:337–338
  for forecasting higher moments 24:338
  multivariate models 24:338–339
  univariate models 24:337–339
Foreclosure
  ground zero 3:581
  identity development 21:434–437
Foregoing theory 23:469
Foreground knowledge (FK) 25:755
Foreign Affairs 11:433
Foreign citizens, in EU countries 15:380–381
Foreign crime groups 24:598
Foreign direct investment (FDI) 9:324–326, 12:480–482, 12:484, 12:486–487, 12:580–586, 20:565–566, 23:209, 25:96, 12:583f, 12:585f
  and economic reform 18:694–695
  efficiency-seeking 9:324–325
  explanations of 12:484–486
  foreign portfolio investment 9:324
  impact 9:325–326
  impact of 12:487–489
  location 9:325
  market-seeking 9:325
  motivations 9:324–325
  multinational corporations 16:74, 16:77–79
  multinational enterprises 9:324, 9:326
  and NGOs 9:325–326
  trends 9:324
Foreign economic policy, trade 24:486, 24:488
Foreign investors, in China 6:790
Foreign judgments, recognition and enforcement of 13:617–618
Foreign languages (FL) 21:350, 21:360
  in labor market 13:261–262
  teaching and learning 9:327–332
    communicative language teaching 9:330
    critical pedagogy 9:329, 9:331
    cross-linguistic influence 9:329
    definition 9:327
    emphases in second language research and teaching 9:328–329
    future directions 9:331
    innatist 9:328–329
    interactionist 9:327–331
    intercultural competence 9:331
    interlanguage 9:328–329
    learner-centered instruction 9:329–330
    and measurement 9:330–331
    multilingualism 9:329–330
    social media 9:331
    sociocultural theory 9:328, 9:330
    theories of language learning 9:327–328
Foreign law, and constitutions 4:713–714
Foreign legal consultants (FLCs) 13:623
Foreign markets, expansion into 6:585
Foreign operations, expansion of 12:584
Foreign policy
  appeasement 1:826, 1:830–831
  and area studies 1:949–953
  and health 10:203
  sanctions 20:891–892
    methodological problems 20:891–892
  and United States Congress 4:652
Foreign Policy and Global Health initiative 10:203
Foreign Policy Themes and Priorities 21:397
Foreign portfolio investment 9:324
Foreign rule of law *ex officio* 3:672
Foreign Service Institute 12:357
Foreign terrorist organizations (FTOs) 13:531
Foreign-born populations, in United States 11:628–630
Foreigner 22:54
Foreignness 10:171
Forel, Auguste, works, *The Sexual Question* 12:902
Forensic anthropology 9:333
Forensic applications, ABA for 2:410–411
Forensic archaeology 9:333–338, 11:3
  anthropology 9:333
  applications of 9:335–337
    human rights and mass graves 9:336–337
    joint POW/MIA accounting command 9:336
    mass fatality incident 9:336
  history 9:333
  methods 9:334–335
    recovering and excavating human remains 9:335
    searching for clandestine grave 9:334–335
    searching for skeletal remains deposited on ground surface 9:334
  research 9:337
  training 9:337
  working in medicolegal context 9:334
Forensic culture 7:876
Forensic DNA, and social justice 9:971
Forensic DNA profiling, development of 9:955–956
Forensic errors in sentencing 25:790
Forensic genomics 2:623
  identification in 10:33–34
Forensic offender rehabilitation (FORNET) 3:435, 24:439–440
Forensic psychiatry 9:346–350
  and civil law 9:348–349
  concept 9:346–347
  correctional psychiatry 9:349
  criminal behavior 9:346
  and criminal law 9:347–348
  definition of 9:347
  legal regulation 9:349
  organization and future of 9:349–350
  special clinical skills of psychiatrist 9:349
Forensic psychology 1:854
  assessment 9:354–355
    hypnosis 9:354–355
    polygraph 9:355
    psychological testing 9:354
  future 9:355
  historical background 9:351
  role of psychologist 9:351–354
    child custody 9:354
    competency to stand trial 9:351–352
    criminal responsibility 9:352
    eyewitness testimony 9:353
    other defenses 9:352
    personal injury 9:353
    prediction of violent behavior 9:352–353
    profiling 9:353–354
    sentencing 9:353
Forensic science 21:159–161, 24:120
Forensic testing, ethics of 20:703
Forensics 9:955–961
  and neuroscience 16:668
Foresight 24:126
Forest & Conservation History 17:672
Forest plot 15:268, 15:268f
Forest Town party, police raid at 21:744–745
Forestry 13:239–240, 13:242–243
"Forests are longing for the sea" movement 22:571–572
Forgetting 8:709
  and proactive interference 19:3–5
  and retroactive interference 20:613, 20:615, 20:617–618
  synaptic mechanisms and protein synthesis in 23:873–874, 23:877–878

Forgetting (*continued*)
and temporal distinctiveness 24:202–204
see also Memory; Source monitoring
Forkhead Box O3A (FOXO3A) gene 17:175–176
Forkhead box protein P2 *see* FOXP2 (Forkhead box protein P2)
Form (printing) 18:915
Formal adoption 1:136
Formal analogs
in EITMs 18:455–456
for economic voting 18:456–457
linking with applied statistical analogs 18:457
Formal and mandatory schooling 8:394
Formal assignment systems of gender 9:796–797
Formal categories (word order) 25:647
Formal competence, and personal identity 17:745–746
Formal control 4:835–836
Formal demography 1:839
Formal education system 3:13, 7:122, 7:124, 7:289–290, 7:293
and informal education, difference between 6:896–897
Formal empire 11:689
Formal equality, critiques 7:896–897
Formal forecasting models 9:319–321
Formal guarantees in legal systems 5:494–496
Formal industrial relations 25:687, 25:689
Formal inequality of treatment 7:894–897
Formal institutions 22:979–980
global governance 10:156
legal phiosophy 18:24
Formal knowledge 21:262
Formal language theory 14:296
Formal languages, reasoning 5:933–934
Formal law 13:471, 13:524
social change and 13:542
Formal logic 14:319
Formal (quantitative) memory models 15:161
Formal methods of cultural analysis 9:357–358
computational approach 9:364–366
humanities 9:365
information sciences 9:365
linguistics 9:364–365
sociology 9:365–366
contemporary approaches 9:361–364
content analysis 9:362
cultural capital studies 9:362
field theory 9:362–363
measuring meanings with cognitive science 9:361–362
networks and culture perspective 9:363–364
history 9:358–361
community studies 9:360–361
early uses of data to measure culture 9:358
Kroeber's analysis of fashion trends 9:358–359
Lévi-Strauss' contributions for interpreting myths 9:359–360
social work theory 9:360–361
surveying values, attitudes, and opinions 9:361
Formal models
of kinship 13:53–60
of politics 18:512–514
Formal names 16:184
Formal operational children 18:129–130
Formal operational thinking 5:481
Formal operations 7:881, 7:883, 14:306, 14:308
Formal organizations
and informal organizations, distinction between 17:412–413
Luhmann on 14:386–387
mobile communication in 15:631
negotiated orders 17:412–413
Formal phenomenology 20:345–346
Formal policing 18:248–249
Formal pragmatics 4:794
Formal rationality
and majority rule 14:447
race–crime relationship 19:841
Formal rationalization 20:203
Formal sampling theory 9:840
Formal science education 21:229–230, 21:234
Formal semantics 15:754
context-dependent phenomena and 15:755
debates on 15:755
new directions in 15:755
psycholinguistics and 15:755
Formal sociology 21:967
Formal systems 19:61
Formal theory of cycles 18:559–561
Formalism 13:439, 13:516, 14:319–320, 21:586
and decline of number 17:59–60
and kinship terminologies 13:63–65
legal reasoning 13:776, 13:779, 13:782
linguistic theory 9:471–472
Formality, and conscientiousness 4:659
Formalization 13:469, 14:295
of associations 25:254–255
of cosmopolitan arguments 5:44
hierarchies and markets 10:843
of land claims 13:235
microeconomics 15:332–334
of multidimensional scaling 16:41
organizational ergonomics 11:302
standard, theories with 2:341–343
Formal-logical thought 16:515
Forman, Paul 17:13
Formative assessment process 24:48–49
Formative classroom assessment 3:820–821
Formative evaluation 19:118
Formative indicators 11:751
Formative societies, Andes 23:38
Former prisoners, portrait of 20:62
Forms 18:915, 23:891–894
family 8:769
and meaning 17:934
organizational 2:768–772, 6:235–240, 17:417–419
Plato on 18:210
reality of 20:20
of survey questions 23:767
Forms of life 9:1016–1017
concept of 6:590
Forms of sociation 21:966
Formula Baetica 21:401–402
Formulas, electoral, proportional representation 19:206–207
Formulation of items 7:931
FORNET *see* Forensic offender rehabilitation (FORNET)
Fornication 9:729
Fornix 10:889, 14:127
For-sale property 20:14
Forsthoff, Ernst 15:520
Fort, R., works, *Sports Economics* 23:318
Fortes, Meyer 1:767, 9:868
Fortifications 10:499
war and conflict 4:612
Fortune 12:776
Fortune, Reo 14:837
The Fortune Encyclopedia of Economics 7:590–591
Fortune telling, in Korea 2:75
41st chair 2:333
Forum for the Empowerment of Women 21:747
Forums, terrorist 24:228
Forward citation rate 20:499
Forward genetics 9:905
Forward guidance
ongoing usefulness of 15:732
on policy interest rate 15:728–729
Forward inferences 12:39–40, 19:395–396
Forward inverse projection 6:150–152
Forward recurrence time 8:317
Forward–backward effect 15:157
FOSDIC (Film Optical Scanning Device for Input to Computers) 3:304–305
Fossil fuel incumbents, impact of RES-E deployment on 18:372
Fossil fuels 7:615, 7:618–620, 7:633–638, 9:199
Fossil primates
primates, evolution of 18:908
Fossil record 8:384
evolution 8:384
Fossilization
concept of 16:61
multiple language 16:61
Foster, John 10:733
Foster care 1:136–139
Fostering wisdom 25:595
Foto Marburg 2:24
Foucault, Jean Bernard 12:121
works, *History of Sexuality* 19:754
Foucault, Michel Paul 4:641, 4:726, 5:500, 6:503, 7:379, 9:368–373, 9:771, 11:129, 11:380, 11:548, 12:257–258, 12:900, 13:258–259, 18:938, 19:677–678, 19:747, 20:682, 21:237, 21:750, 22:994, 23:82, 23:309, 23:728

1979–1980 lectures 9:370–371
analysis, in critical planning theory 5:256
on anthropology 1:770
on authority 17:408–409
biography of 9:368
on Christianity rituals 9:370–371
community politics after 18:767–768
concept of governmentality 5:256–257
concept of scientific classification 11:548
discussion of parrhesia in Christianity 9:371
discussion of Plato's Alcibiades 9:370–371
on Enlightenment 7:666–667
Euripides's play *Ion*, analysis of 9:370–371
Foucauldian discourse analysis 19:673
on genealogy of the present 1:749
on governmentality 10:309–311
on human sciences 11:380, 11:383–384
on human studies 22:689–690, 22:692–693
intellectual trajectory 9:368–369
as a cultural diplomat 9:368
on knowledge 5:523
lecture at Collège de France 9:368
on linguistic turn 25:210–211
on mental illness 15:576
on narratives 16:218–221
notion of discourse 17:424–425
on Oedipus King by Sophocles 9:370–371
on Panopticon paradigm 23:728
on philosophy 2:305
on Plato's dialogues 9:370–371
on political power 18:762–763
on power 18:767
on prison life 18:922
on problematization 9:370
publication projects 9:372
qualifications 9:368
on queer anarchism 19:747–748
on reproduction 20:452
on sexuality 9:370
social sciences, discourses 9:370
on technologies of self and individualization 9:371
theories 22:997, 25:211
on visuality 25:208, 25:211
works 9:369–371
The Archaeology of Knowledge 9:368, 14:199–200
The Birth of the Clinic 2:729, 9:368–369
Discipline and Punish 2:729, 9:368–370, 9:372, 11:27
History of Madness 9:369
History of Sexuality 9:368–371, 13:860–861, 21:750, 21:768–769
La volonté de savoir 9:370
The Order of Things 9:368–369, 9:372
Foucault effect 9:372
Fouillée, Alfred 23:2–4
Foul-dealing problem 18:934
Foundation for Applied Psychiatric Anthropology 19:364
Foundation for the Establishment of the International Criminal Court 25:386
Foundation funding, for environmental movements 7:766
Foundational Christian communities 14:34
Foundational critical theory 17:432–433
Foundationalism, and cultural relativism 5:484
Foundations 16:772, 22:659–660, 22:663
Alfred P. Sloan Foundation 22:659
Andrew W. Mellon Foundation 22:659
Bill & Melinda Gates Foundation 22:659–660
Ford Foundation 22:659
funding 22:663–664
The Rockefeller Foundation 22:659
Founders (entrepreneurship) 7:682–684
Fountain Grove 2:147
Four Noble Truths 2:887
Fourastié, Jean 15:682
Fourcade, Marion, on economics 5:539–540
4C/ID model of instructional design 12:198–199
learning tasks 12:198–199
part-task practices 12:198–199
procedural information 12:198–199
supportive information 12:198–199
Four-factor model 21:51
Four-factor risk model 2:793
4-H Study of Positive Youth Development 3:447, 3:449, 21:549–551
Fourier, Charles 14:644, 22:836–837, 22:841
Fourier analysis 5:652, 9:374–379
definition 9:374
dirac-d-distribution 9:377–378
Fourier series 9:375, 9:378–379
Fourier transform 9:374–376
as spectral representation 9:376–377
generalized Fourier transform and some applications 9:377–378
impulse response 9:377
Shannon's sampling theorem 9:377–378
wavelets 9:378
higher dimensional Fourier transform 9:378–379
linear systems theory 9:376–379
power spectrum 9:375, 9:377
Fourier cosine transform 9:376
Fourier series 9:375, 9:378–379
Fourier transform 9:374–376
generalized 9:377–378
as spectral representation 9:376–377
convolution 9:376–377
derivatives 9:376
functions of differentiation 9:376
Fourier-Mellin transform 9:375
Four-level tonal model 18:71
Fournier, M., works, *Cultivating Differences* 5:606
Four-phase model of interest development 12:378–379, 12:379t
Four-point experiments, in visual space geometry 25:198–199, 25:199f
Four-step travel demand model 24:618
see also Urban transportation modeling system (UTMS)
Fourteenth Amendment, US Constitution 6:530–533
Fourth World 18:911, 24:302, 24:777
concept of 23:820
and sustainable development 23:820
Fouta Djallon (highland) 16:266–267
Fouts, J.T., works, *Research on Educational Innovations* 14:757
Foveal mask technique 8:632–633
Fox, George 7:654
Fox, Robin
Biosocial Anthropology 22:899
The Imperial Animal 22:903
Men in Groups 22:903
Foxe, John 3:720
works, *Acts and Monuments* 3:720
FOXO3A *see* Forkhead Box O3A (FOXO3A) gene
FOXP2 (Forkhead box protein P2) 13:311
FPC *see* Frontopolar cortex (FPC)
FPÖ *see* Austrian Freedom Party (FPÖ)
Fraassen, Bas Van 7:569–570
Fractals 21:2
Fractional anisotropy (FA), for brain asymmetries 2:121
Fractional factorial design 17:259
Fractional integration methods, for political data 18:517
Fractionation of visual working memory 25:176–177
Fractions 4:66–67, 4:69–70, 4:72
of 2m factorials 17:259
Fragile Families and Child Wellbeing Study 5:697, 14:699
Fragile X Mental Retardation Protein (FMRP) 14:361–362
Fragile X Mental Retardation Syndrome (FXS) 14:364
Fragility 4:633
Fragmentary languages 11:824
Fragmentation 13:446–447
of chemistry 3:373
of knowledge 6:601
within leisure 13:852
of mass media audience 14:713, 14:716–717
of online media audiences 16:794
of political science 18:470
public sector organizations 19:600–601
Frailty models 8:297, 15:844–845, 16:48, 16:52–53
for survival analysis 23:780–781
Frame error 16:919–920
dealing with 16:919–920
dual-frame sampling 16:920
random digit dialing 16:919–920
definition 16:919
Frame semantics 13:934
Frame-based systems, knowledge representation 13:102–103
Framed field experiments 9:129
Frames 10:277
in sampling 20:876
Frames of reference (FoR) 23:118

Frames of reference (FoR) (*continued*)
academic self-concept and achievement 1:57–59
educational implications 1:59
empirical evidence 1:58–59
simultaneous use of external and internal frames 1:58
dexis 6:54
internal/external frame of reference model 12:425–432
multiple 17:703–707
spatial 23:118–119
body-centered *versus* allocentric representations 23:119
retinocentric *versus* body-centered FoRs 23:118–119
use, cultural variation in 23:123–124
Framework Convention on Tobacco Control (FCTC) 22:128
Framework for the Development of Environment Statistics (FDES) 5:776
Framework Programs 22:643–646
Framing 12:720, 16:191
in agenda-setting 1:361
consumer psychology 4:739
disputes 6:562–564
effects, cooperation 4:865–866
and identity development 4:178
and inequality 11:910–911
in mass media 14:715
political agendas 1:364
Framing effects 2:445–446, 16:642–644, 17:782, 24:681
in decision making 5:918
experimental economics 8:548
Framing theory, and collective behavior 4:161
Framingham Heart Study 16:558
France 7:165, 8:261–262, 12:255, 12:257
adult mortality in 1:167–168, 1:167f
anthropology at home in 1:752–753
authority theories 2:269–270
censuses in 3:305–306
civil code 19:194, 21:403–404
clinical psychology, training 3:875
conceptual history in 4:550–551
economics in 5:539–540
enlightenment in 7:656–658
gross inflow to 15:380t
immigration policy 11:662–666
industrial policy in 5:537
infant mortality in 12:18–19
interviewing in 12:672–673
laicization in 13:216–217
second threshold 13:216
linguistic unification in 5:460–461
literacy in 14:238–239
psychoanalysis 3:875
public intellectuals in 22:699
racism in 19:848–849
revolution in 13:216–217
Roman Catholicism in 13:216–217
semipresidentialism in 18:867
sexual harassment in 21:725
small states 22:109–111
sociology 22:992
and World War I 9:212
France eugenics 8:226–227
Franchise 7:349, 7:351
Franchise bidding 20:168–169
Franchise monopolies 20:173
Franchising 12:921
and property rights 24:529
Franchising arrangements 17:365
Franco-French war 12:275
Francophone demography 6:110
Franco–Prussian war 6:172–173, 8:239, 15:174–175, 25:365
Franco–Soviet alliance of 1935 1:557
Franju, Georges, works, *Blood of the Beasts* 6:623
Frank, Jerome David 19:506–507
Frank, Josef 25:90
Frank, Philipp 25:90
Frank, T., works, *What's the Matter with Kansas?* 19:565–566
Frankel-Brunswick, E., works, *The Authoritarian Personality* 2:197–198
Frankenberg, Ronald 1:752–753
on race in North America 17:23
Frankfurt, Harry 2:316, 9:389
Frankfurt Institute for Social Research 19:125–126
Frankfurt School 1:496–497, 5:252, 5:261, 5:264, 9:380–386, 11:176, 13:794, 19:330–331, 22:696, 22:961–962
avant-garde art 2:321
critical theory 5:294, 19:617
critique of the culture industry 13:855
cultural studies 9:385
culture industry 9:381–382
history from 1923 to 1969 9:380–383
Adorno's rise 9:382–383
collaborative social theory 9:381–382
critical theory of society 9:381
institutional beginnings 9:380–381
philosophical turn and dialectic of enlightenment 9:382
rise to public prominence 9:382–383
history from 1970 to the present 9:383–385
communicative paradigm and historization 9:384–385
recognition, feminism, justice 9:385
role of Habermas 9:383
theoretical core of communicative reason 9:383–384
Marxism 9:380–381
modernity 9:383–385
role of Horkheimer in 9:380–384
role of Marcuse in 9:380–384
Frankfurter, Felix, works, *The Commerce Clause under Marshall, Taney, and Waite* 11:438
Frankl, V.E., works, *Man's Search for Meaning: An Introduction to Logotherapy* 19:492
Franks, David 8:464
Franks, Paul 19:666
Frank's theory of differential overqualification 15:598
Franz, Shepherd Ivory 13:392
neuropsychological rehabilitation 2:814
works
Nervous and Mental Re-education 2:814
Franz Ferdinand, Archduke, assassination 9:212, 9:215
Fraser, Nancy 5:299, 11:175, 20:48
critical theory 5:299
on politics of recognition 20:45–47
Fraternal birth order effect 21:783
and sexual orientation 21:776–777
Fraternal polyandrogynous marriage 14:618
Fraternal twins 24:703
Fraternities 12:236, 12:236t
and freedom 9:398
Fraternities, hazing in 10:554–561
Fratricidal war index 10:13
Fratricide 5:678
Fraud 3:28, 5:633, 21:262
and peer review 17:682
Frazer, Sir James George 9:672, 10:139, 14:453
and anthropology 1:746
works
The Belief in Immortality 2:494–495
The Golden Bough 1:746, 1:766, 7:589–590, 8:588, 20:307, 20:309
FRE *see* Federal Rules of Evidence (FRE) 702; Federal Rules of Evidence (FRE) 706
Freak radicals 5:74
Fréchet distribution 6:572
Fréchet drift rate variability, LBA model with 5:904
Free, M.D., works, *Race and Justice: Wrongful Convictions of African American Men* 25:789–790
Free actions 9:404–405
consequences of 9:406
and moral responsibility 9:406–407
Free agency 2:314–318, 23:320–322
Free agent 18:666
Free appropriate public education (FAPE) 3:480, 23:222
EAHCA principle 23:222t
Free association 24:545
Freud's view 19:503, 24:724–725
and meditation 24:726
and psychoanalysis 24:725–726
and repression 24:724–725
and Yoga 24:726
Free collective bargaining 11:349
Free enterprise system, and state provision of public goods economics and power 14:23
Free exercise rights 20:280
Free informed prior consent (FIPC), right to 11:801
Free labor 13:159, 25:697
Free landless workers 1:458
Free love 21:792
Free market 3:96, 14:733–734
and equilibrium 7:915
and neoliberalism 16:463
of vocational education 25:234–235
Free market ideology, and war 25:368

Free market islands 23:209–210
Free media coverage, of political campaigns 3:67
Free migration 11:649
Free mobility 25:199
Free movement, and immigration policy 11:662, 11:665–667
Free press 12:909–910, 19:603–604
Free professions 19:97
Free Recall Index 16:725
Free relatives 23:891
Free rider problem 4:733, 4:736, 19:534
Free riding behavior 4:940
  and Internet 12:588
Free speech movement (FSM) 21:340
Free testation and capitalism 19:175
Free time 13:840–842, 24:390
  cultural consumption 13:847–849
    arts participation 13:847–849
    conspicuous consumption 13:849
  demographic correlates of 13:846–847
    age 13:846
    ecological factors 13:847
    gender 13:846
    status 13:846–847
  international comparisons 13:847
  versus leisure 13:845
  measurement of work and 13:845–846
  organizations and volunteering 24:390
  social indicators 13:850–851
  social life 24:390
  social trends 13:847–850, 13:846f, 13:848f, 13:848t, 13:850f
  subjective aspects of 13:849–850
    appreciation of quality 13:849
    greater enjoyment 13:849
    less time pressure/fatigue 13:849
    time estimates 13:849–850
  television 13:846
  time diaries *versus* time estimates 13:849–850, 13:846f
  TV viewing 24:390
  US perspective 13:845–851
Free Time Motivation Scale for Adolescents 13:835
Free trade
  labor, division of 13:159
Free trade agreements 12:569, 24:485
Free trade areas 12:574, 12:576–578
Free trade policies 14:531
Free trade zones (FTZs) 23:208
Free Transit Agreement 24:514
Free translation 14:185–186
Free will 2:144–146, 2:316, 7:721, 9:407
  and action 9:404
  Kant on 13:7
  neurobiological correlates of 16:668
  and responsibility 20:590
Free word order 25:644
  languages 23:889
Free worker 25:696
FreeCycle 8:487
Freedom 3:686–688, 8:267, 9:405–406, 10:737–741
  of actions 9:404–406
  and American/French Revolutions 1:597–598
  of association and assembly, repression of 18:447–448, 18:448f
  and authority 2:270
  consequences of 9:407–408
  economic philosophy 7:103
  and Enlightenment 7:664–665
  and feminist legal theory 8:921–922
  historical origin 9:387–388
  history of 9:387–391
  Hobbes and Rousseau on 9:388–389
  and individuality 11:818
  and liberalism 14:13–14
  and long-term heterosexual relationships 20:913–914, 20:916–917
  market allocation 14:540–541
  Mill on 15:524
  national liberation movements 16:241–246
  political 9:398–403
  and postmodernism 18:686–687
  and public interest 19:553
  and reason 15:538
  republicanism 20:480–481
  and responsibility, media ethics 15:44–45
  right as compatibility of 13:8–9
  social justice 22:392–393
  and welfare 25:507–509
  of will 9:407
Freedom of belief 3:574–578, 13:219
Freedom of information (FOI) 9:396
Freedom of Information Act (FOIA) 23:755–756
Freedom of Information laws 21:378
Freedom of inquiry 8:58
Freedom of religion 3:574–578, 10:215–216
  repression of 18:447–448, 18:448f
Freedom of research 20:523–524
Freedom of speech 10:541
  and press, repression of 18:447–448, 18:448f
  and research censorship 3:281–282
Freedom of testation 13:865
Freedom Ride 11:758–759
Freedom to/from contract 4:810–812
Freelancers, and social networks 5:468
Freely varying transition intensities, Markov process 8:314–315
Freeman, Annis 9:461
Freeman, Caren 2:74
Freeman, Carla 1:640–641
Freeman, Cathy 2:237
Freeman, Derek 14:838–839
Freeman, J., on categories 15:576
Freeman, Richard
  future of higher education 15:596
  works
    The Overeducated American 15:596
Freenets 21:188
Freeports 23:208
Freestanding company 16:75
Free-to-air television 14:719, 14:722–723
Frege, Gottlob 2:631, 19:391, 23:596
  works, *Grundgesetze der Arithmetik* 25:611
Frege-Russell program 23:596–597
Frege's Puzzle 14:294
Freiheitliche Partei Österreichs (FPÖ) *see* Austrian Freedom Party (FPÖ)
Freire, Paulo 4:380–381, 17:540–541
  social work 22:134–135
  works
    Pedagogy of the Oppressed 3:521
French, gender assignment in 9:797
French Academy (*Académie française*) 2:89, 5:460–461
French case, intellectuals 12:273–274
French Civil Code 3:670, 7:925, 19:194, 21:403–404
French colonialism 11:1, 25:95
French Communist Party (PCF) 14:661
French concepts, of personality disorder 17:891–892
French cooking 9:292–293
French Declaration of the Rights of Man and the Citizen 3:647, 3:674
French Eugenics Society 8:241
French feminism 22:697
French Foundation for the Study of Human Problems 8:220–221
French gastronomy 15:576
French hegemony 25:290
French historical revolution 11:40
French Historiographe du Roi 11:120
French Institute for Central Asian Studies (Tashkent) 3:317
French Islam 24:603
French judgment 14:242
French jury model 5:228
French Labor Force Survey 10:303
French Law for the Protection of Mothers and Children 8:220–221
French philosophers 19:611
French Polynesia 17:158–159, 17:162–163
French regulation theory 7:687
French Revolution 1:597–601, 3:713–714, 5:17, 7:650, 7:657, 7:659–662, 9:409–412, 14:18–19, 16:224, 16:291, 17:516–517, 20:288, 20:621, 20:638–640, 20:645–647, 23:2–3, 23:364–365, 25:292, 25:363–364
  Arendt on 1:597, 1:599–600
  aristocracy 1:968
  bourgeois revolution 9:409–411
  Burke on 1:598–599
  capitalism 9:409–411
  crowds 5:362
  democracy 9:409, 9:411
  and democracy 6:81–82
  Enlightenment 9:412
  experience of 22:553–555
  foundations of modern social science 9:410–411
  freedom 1:597–598
  Furet on 1:600
  human rights 1:597, 1:599
  Marx on 1:598–599, 9:409–411
  Marxism 9:409–411

French Revolution (*continued*)
modern archives since 5:701
and modern public archive 1:909–910
modernity 9:409, 9:411
monarchy 9:409–411
Napoleonic wars 9:409
origins of ideology 9:409–410
and people welfare 25:475–476
religion 1:598–600
scholarly controversies concerning 9:411–412
stages 9:409
terror 9:409–411
Tocqueville on 1:597–600, 9:411, 24:421, 24:423
French Revolutionary wars 25:363–364
French School of Sociology 5:239, 14:825–826
French sociology
Boltanski's contribution 2:747–754
Durkheimians 6:706–708
French structuralism 22:723, 23:626–627
and anthropology 1:753
in Korea 2:73
French system of succession 19:174
French Système Probatoire de la Observation de la Terre (SPOT) satellite 20:422–423
French telegraph network 24:154
French tradition of statistics 5:6–7
French universities 22:590
French–Harary–DeGroot model 15:663
Frequency, of financial crises 9:187–189, 9:189f
Frequency domain
time series analysis in 5:652
Frequency histogram 19:29–30
Frequency judgments, impairment 6:721
Frequency measures, for health behaviors 10:588–589
Frequency modulation (FM) radio broadcasting 19:880
Frequency of exposure to pornography 18:614
Frequency questions 23:766–767
Frequency scales, questionnaires 19:771–772, 19:774–775
Frequency-dependent selection (balancing selection) 17:901
Frequency-resolved optical imaging 16:389
Frequentists 19:26, 20:928
sequential experiments 21:635
Fresh-start principle, of bankruptcy 2:353–354
implementation of 2:355–356
Freshwater 25:437
Freud, Anna 11:537, 13:67
ego defenses and ego functions 13:67
works
A Handbook of Child Psychology 7:591
Freud, Sigmund 1:636, 7:379, 8:590–591, 9:460–462, 9:638, 11:472, 18:47–48, 19:326–327, 19:330, 20:354–355, 20:357, 23:232–233, 23:682, 24:545–547
on adolescent personality 17:863
on applied social research 1:851, 1:853
biography 19:327
on clinical psychology 24:524
compromise–formation view of symptoms, on repression 24:724–725
criticisms 19:503
on determinism 24:726
on drive theory 22:814–815
early career and scientific outlook 9:413–414
admiration for Goethe 9:413–414
neurological work 9:413–414
education and early interests 9:413
exile to England 9:418
family background and early life 9:413
on fetishism 9:107
on free associations 19:503, 24:724–725
on homosexuality 9:656
on humor 11:416
hysteria studies of 11:506–507, 11:510, 19:502–503
impact on twentieth-century thought and culture 9:418–419
on infant memory 6:759
on introjection 5:578
legacy, controversies involving 22:764–767
on Leonardo da Vinci 24:725–726
model of group process 19:350
on neuroscience 9:414
dream analysis 9:415–416
model of the mind 9:416
neuropathology 9:414
recollected (manifest) dream content 9:415–416
and Nietzsche 15:290–291
on psychoanalysis 9:414–415, 19:326–327
ego 9:417, 19:338–339
growth of 9:416–418
id 9:417, 19:337–338
psychotic illness 9:417
superego 9:417, 19:339
psychoanalytic career of 9:415–416
on psychoanalytic movement and institutions and practices 19:327–328
on Russian sexual life 21:740
on sexuality concept 9:416
children's sexual theories 9:416
on sexuality studies 19:503
on theory of repression 12:903
on unconscious (psychology) 24:723–725
works
Analysis Terminable and Interminable 9:417–418
Beyond the Pleasure Principle 12:177, 19:330–331
Civilization and Its Discontents 2:729, 9:417–418, 19:330–331
Entwurf einer Psychologie 23:874
The Future of an Illusion 20:354–355
A General Introduction to Psychoanalysis 2:814
The Interpretation of Dreams 9:415, 12:903
Jahrbuchfür psychopathologische und psychoanalytische Forschungen 12:903
Jokes and their Relation to the Unconscious 9:416
The Psychology of the Unconscious 12:904
The Psychopathology of Everyday Life 9:416
Studies on Hysteria 9:415, 12:903
Thoughts for the Times on War and Death 9:417
Three Essays on the Theory of Sexuality 9:416
Totem and Taboo 9:417
Freudian researcher-as-expert 4:771
Freudian theory 24:454
conceptions of mind 22:761
psychodynamic treatment 22:763–764
psychopathology in 22:763
Freudianism, in Central Asia 3:320
Freundesliebe 9:644–645
Freyer, Hans 15:520
Freyre, Gilberto 2:928, 9:420–424
contributions
to anthropology 9:421
to cultural history 9:420, 9:422
to economics 9:420
to geography 9:420–421
to history 9:422
to psychology 9:421
to sociology 9:421–422
on miscegenation 9:423
on psychoanalysis 9:421
on race 9:420–422
reception of works of 9:424
on social ecology 9:420–421
on social history 9:422–423
theory of 9:422–423
works
The Masters and the Slaves 9:423
FRI *see* Fluid Reasoning Index (FRI)
Frictional unemployment 24:738–740, 24:742
Frictionless planes 11:516, 11:519
Frictions (economics) 21:333
Frictions (marriage) 8:832
Friedan, Betty 2:394
Friedkin–Johnsen model for fixed influence network 15:663–666
Friedland, Roger 17:393
Friedman, Lawrence 9:460, 13:803
works, *Contract Law in America* 13:517
Friedman, Marilyn 2:317
Friedman, Milton 1:604, 21:170
economic methodology 7:100–101
macroeconomics 14:401–403
works
Capitalism and Freedom 4:690, 14:41
Friedmann, Georges 22:699
Friedmann, John 18:182, 18:206
Friedrich, Carl J. 2:271, 20:669–670
Friendly compliance, and agreeableness 1:463
Friends
characteristics of 9:443–444
cultural meaning of 5:427
mafias 14:422
Friends of the Earth (FoE) 7:763–765
Friendship 23:1–2
during adolescence 9:432–441

attachment 9:432–434, 9:436
bullying 9:435–438
contextual conditions for social networks and friendship establishment 9:436–438
future directions 9:438
homophily 9:437
immigrants 9:432, 9:435, 9:437–438
impact of social change in societies 9:438
in individualistic *versus* collectivistic cultures 9:433–434
peer groups 9:432, 9:434–437
peer relationships 9:433, 9:438
prosociality 9:433
role of adolescence and peer experience across cultures 9:432–433
role of aggression in 9:435–436
role of culture in identity formation 9:434–435
and adolescent social competence 22:217–218
and agreeableness 1:465
anthropology of 9:425–431
altruism 9:425–428, 9:430–431
choice 9:426
class 9:425–429
descent theory 9:426–427
institutional relations 9:425
kinship and difference in cultures of relatedness 9:427–431
love (philia) 9:427
personal relations 9:425–427
terms 9:426
trust 9:425–426, 9:428, 9:430
Aristotle on 9:426–427
bisexual, gay, LGBT
see also Lesbian, transgender (LGBT individuals)
during childhood 9:442
characteristics negatively affecting 9:444
characteristics of friends 9:443–444
child development 9:442–444
classrooms and extracurricular activities 9:445
cultural ideology 9:446
cultural similarity 9:444–445
Hartup on 9:442
Piaget on 9:442–443
prevalence and assessment 9:442
psychosocial associations 9:443
quality 9:443–444
stability 9:444
Sullivan on 9:442–443
theories 9:442–443
time with peers/family 9:445
working children 9:445–446
and conflict 4:596–597
and culture 5:427–428
during infancy and early childhood 9:448–453
0–2 years 9:449
2–6 years 9:449–450
assessment 9:448–449
cultural variations 9:450–452
developmental significance 9:448
networks, nonheterosexual British Asians 16:865–866
provisions 9:450–451
structural balance theory 23:605
and trust 24:668–670
and youth physical activity motivation 15:918–919
Friendship homophily 9:437
Friendship ties, and social networks 16:564
Friere, Paulo 3:521
Friern Hospital 6:47
Frijda's theory, facial expressions 5:545
Fringe imagery 18:247–248
Fringes of a state 14:584–585, 14:588–589
Frisch, Karl von 14:367, 16:469
Frisch, Ragnar Anton Kittil 9:220
macroeconomics 14:400–401
Froebel, Friedrich 10:740
Frog pond effect 22:214
Fromm, Erich 20:669–671, 21:129
on authoritarianism 9:459–460, 9:462
on critical theory 9:458–459, 9:462–463
on gender, Freud, and psychoanalytic feminist sociology 9:460–462
on Marxism 9:458–460, 9:462
on Nazism 9:458–460
personal life and career of 9:458–459
on prejudice 9:460
as public intellectual and public sociologist 9:462
on social psychology 9:458–462
on sociology of race and racism 9:460
on women's studies 9:460
works
Anatomy of Human Destructiveness 9:458
The Art of Loving 9:458, 9:461
Beyond the Chains of Illusion 9:461
Escape from Freedom 9:458–462
The Heart of Man or The Anatomy of Human Destructiveness 9:460–461
Marx's Concept of Man 9:461
May Man Prevail? 9:461
The Revolution of Hope 9:461
The Sane Society 9:458, 9:460–461
Social Character in a Mexican Village 9:458
To Have or To Be? 9:458, 9:461
Front National (FN) 18:410–411, 18:413–415, 18:610–611
Front Populaire government 16:268
Frontal amnesia 6:718–720
Frontal cortex 8:849–851
and morality 16:607–608, 16:611–612, 16:615t
Frontal eye fields (FEF) 5:27, 18:847, 18:850
Frontal granular cortex 18:811
Frontal leucotomy 23:18t
Frontal lobe dysfunction (FLD) 13:492
Frontal lobes 1:644, 3:334, 3:336, 13:370–371, 18:814
disorders affecting 18:816
dysexecutive amnesia 6:718–724
dysexecutive syndromes 6:725–726
and episodic memory 7:858–860
mechanisms of spatial representation 17:706–707
motor regions along medial wall of 5:25
source monitoring 23:31–32
Frontal Systems Behavior Scale (FrSBe) 6:726
Frontal–temporal connections 13:371
Frontier migration 12:437
Frontiers 9:464–469, 24:821–822
borders 2:760–761
boundaries 2:760–761, 9:464, 9:466–467
changing context 9:464
Chinese frontiers 9:467
exchange 9:464
expansion 9:465
Ottoman frontiers 9:468
Russian frontiers 9:466–467
and social sciences 9:464
Turner thesis and critics 9:464–466
in history, and borders 9:464
land tenure 13:236
Frontopolar cortex (FPC) 16:641–642
Frontotemporal area 16:691–692
Frontotemporal dementia (FTD) 12:420
and asymmetry of body/brain 2:116
Frontotemporal dementia syndrome 6:77
Frontotemporal lobar degeneration 16:608
Frost, Nancy 12:304
Frotteurism 17:952
FrSBe *see* Frontal Systems Behavior Scale (FrSBe)
Frugal algorithms 1:531–533
Frühe Aufklärung 7:657
Fruit fly 9:879–886
A Fruitful and Pleasant Worke of the Best State of a Publique Weale, and of the Newe Yle Called Utopia 14:253
Fruitlands 2:149
Fruitless gene 9:881
FRUs *see* Face recognition units (FRUs)
Frustration
and behavior during infancy and early childhood 6:12–13
of contract, doctrine of 3:671
Frustration-aggression hypothesis 1:374
Frye test 21:299
Frye v. United States (1923) 8:583
FSA *see* Farm Security Administration (FSA)
F(B) scale 15:567
FSCPE *see* Federal-State Cooperative Program for Local Population Estimates (FSCPE)
FSH *see* Follicle-stimulating hormone (FSH)
FSIQ *see* Full Scale Intelligence (FSIQ)
FSM *see* Family stress model (FSM); Free speech movement (FSM)
FSN *see* Finite state network (FSN) representations of knowledge sources
FST *see* Family support team (FST)
FTAs *see* Face-threatening acts (FTAs)
FTC *see* Federal Trade Commission (FTC); Federal Trade Commission (FTC) Act
fTCD *see* Functional transcranial Doppler ultrasound (fTCD)
FTD *see* Frontotemporal dementia (FTD)

FTND *see* Fagerstrom Test for Nicotine Dependence (FTND)
FTOs *see* Foreign terrorist organizations (FTOs)
FTP *see* Face threatening process (FTP); Future time perspective (FTP)
FTS *see* Field of feminist technology studies (FTS)
FTT *see* Financial Transaction Tax (FTT); Finger Tapping Test (FTT)
FTZs *see* Free trade zones (FTZs)
Fuchs, Stephan, on cultural production networks 5:467
Fuel
  and energy 7:621, 7:623–624
  exploitation of 9:197, 9:200
  and fire 9:197–199
  in industrial society 9:198–199
Fuentes, Marlon, works, *Bontoc Eulogy* 6:621
Fuentes v. Shevin (1972) 18:735
Fuero real 13:858
Fuertière, Abbé, works, *Dictionaire universel* 22:887
Fugue 1:651
Führerstaat 25:357–358
Fujimori, Alberto 23:47
Fujimura, J.H., works, *The Right Tools for the Job* 8:579
Fujisawa Sustainable Smart Town 6:836
Fukushima nuclear power plant accident 7:744, 22:570, 22:572–573
Fukuyama, Francis 4:607–608
  works, *The End of History and the Last Man* 11:437
Fulani of Cameroon 3:325–326
Fulbright-Hays 1:946
Fulfillment 17:737–738
Full binary trees, discrete-state models 6:525
Full breastfeeding 2:848t
Full Faith and Credit Clause of the Constitution 13:618
Full information item factor analysis (TESTFACT) 6:438
Full information maximum likelihood (FIML) method 21:976–978, 21:982, 23:408
Full interpretation (FI) 9:877
Full matching 19:143–144
Full model, life course epidemiology 14:72
Full participation, and inequality 7:897–898, 7:901
Full response 15:21–22
Full Scale Intelligence (FSIQ) 12:284, 16:723–724
Full service agencies, advertising 1:190–191
Full writing systems 25:780–781
Fuller, Lon 12:933–934
Fuller, Margaret 2:515
Full-time craft specialists, in archaic states 23:373
Full-wave rectification 25:160
Fully explicit models 19:62
Fully functioning person 20:739
Fully integrated mixed methods design 15:621
Fully mediated, mediation model 22:549
Fulton, John 23:18
Function creep, forensics 9:959
Function specificity of prosopagnosia 19:251–252
Functional adaptation, of humans 18:126–127
Functional alternatives, in modes of production 15:716
Functional amnesia 1:651, 13:682
  cerebral blood flow 1:654–655
  changes in personality dimensions 1:653
  diagnostic features of 24:556t
  differential diagnosis of 1:655
  epidemiological data about 1:652–653
  magnetization transfer ratio measurement 1:655
  neural correlates of 1:655
  neuroimaging of 1:654–655
  pathogenetic models of 1:653–654
  prognosis 1:655
  risk factors and comorbidity 1:653
  treatment 1:655
  types of 1:652
  see also Psychogenic amnesia
Functional analysis 9:537–538
  and causal analysis 9:537
  of human control behavior 4:813–814
  management accounting 14:468
  Merton's 9:537–538
  requirements 9:537
Functional analytic psychotherapy (FAP) 2:433
Functional approach, data collection 5:707
Functional area of the brain 13:398
Functional Assessment of Communication Skills for Adults 3:794
Functional assimilation 16:484
Functional asymmetries
  language and handedness 10:504
Functional brain imaging 13:677–678, 16:390
  cognition 4:99
Functional brain processes 16:601
  recognition of others 16:601
Functional challenge, in legal reasoning/argument 13:780
Functional communication approach
  assessment of 3:794
  to language disorders 13:331–332
Functional connectivity, in ADHD 2:158–159
Functional corruption, in prisons 18:920
Functional data analysis 9:514–518
  curves 9:514–516, 9:518
  functional linear model and regularization 9:516–517
  human growth data 9:514
  mean function and registration problem 9:514–516
  modeling with derivatives 9:517–518
  nondurable goods index and more derivatives 9:516
  principal components analysis 9:516
  relationships with other branches of statistics 9:518
Functional deviance 25:725
Functional differentiation 6:391, 6:394, 13:555
  Luhmann's 14:388
Functional elements 23:232
Functional elites 7:392–393
Functional equation methods, mathematical psychology 14:803
Functional equivalence 5:319
  ISSP 12:550
Functional explanation
  philosophical aspects 9:519–524
    adaptation 9:520–522
    biological evolution 9:522
    cultural evolution 9:522
    Cummins&rsquo 9:519–520
    dispositional and virtual selection 9:523
    etiological theory 9:520–522
    exaptations 9:522–523
    explanandum 9:524
    malfunction 9:521
    Millikan's 9:521–522
    natural selection 9:520–522
    novelty 9:521–522
    Panglossianism 9:522
    proper and actual domains 9:522
    and selection 9:520–521
    and systems 9:519–520
    trait 9:522–524
Functional fixedness 18:819
Functional gastrointestinal disorders 9:605–606
  depression, anxiety, and stress in 9:608–609
Functional genomics 2:625, 9:1017–1019
Functional grammar 17:556–558
Functional illiteracy 11:590–591
Functional imaging 11:612–613, 11:618–619, 21:945, 23:251–252
  for attention-deficit/hyperactivity disorder 2:156–158
  findings of prosopagnosia 19:252–253, 19:252f–253f
  for lexicons 13:926, 13:928
  mental disorders
    analytical procedures 11:605–609
    experimental procedures 11:605–609
    symptom provocation 11:605–606
  of post-traumatic stress disorder 18:704
  for psychopathy 19:452–453
  and virtual spatial navigation 16:352–353
  of the visual brain 25:202
Functional interdependency 11:285
  groups 10:435–436
Functional knowledge 1:878
Functional level, internationalization at 12:584
Functional level processing
  grammatical encoding 23:256–257
Functional limitations
  of aging 1:395–396
  disability 6:472
Functional linguistics 14:210–211

markedness in 14:536
Functional literacy 14:236
Functional magnetic resonance imagery 9:525–531
Functional magnetic resonance imaging (fMRI) 4:47, 5:23, 5:26–28, 8:713, 8:849–850, 9:476, 9:491–494, 9:525, 10:703, 11:603, 11:612, 11:910, 12:690, 13:368, 13:489–490, 14:439–440, 16:673, 16:675, 17:506, 25:163
for ADHD 2:156–159
advantages 9:492
affective disorders 14:441
applications 9:525
attention, neural basis of 2:164, 2:166–167
blood oxygen-level-dependent (BOLD) response 9:526–527, 14:440
data acquisition 9:525–526
data analysis 9:528–530
connectivity 9:529
localizing brain activation 9:528–529
prediction 9:530
data structure 9:525
diffusion-tensor fMRI 14:440
disadvantages 9:492
of discourse comprehension 13:373–374
experimental designs 9:527
for eyeblink conditioning 2:96
face processing 8:664
future holds 9:530
general aspects 14:439
graphical representation 9:493f
hypnosis 11:466–470
infant cognitive development 4:60–62
intracarotid sodium amytal testing 12:690–691, 12:694
for language learning impairment 13:347–348
of language processing 13:369–374
embodied language comprehension 13:374–375
frontal and parietal lobes involved 13:369–370
phonological processing 13:369–370
temporal regions involved in 13:370
for lexicons 13:926
in lie detection 13:492–493
for long-term potentiation/long-term depression 14:363
machine 9:492f
magnetic resonance spectroscopy 14:440
mirror neuron system 11:331
neural representation 16:630–632
neuromarketing 16:698–701
perfusion fMRI 14:440
periods of the target activity 9:492
preprocessing 9:527–528
reconstruction 9:525–526
in school-aged children and adolescence 16:590–591
of semantic processing of words 13:370–371
angular gyrus involvement 13:371
frontal lobe involvement 13:370–371
frontal–temporal connections 13:371
fusiform gyrus (FG) involvement 13:371
middle and inferior temporal gyri involvement 13:371
parietal and temporal lobes involvement 13:371
of sentence processing 13:371–373
cognitive control and selection 13:372
figurative language comprehension 13:372–373
processing of structure and meaning 13:372
serial order memory 21:640
in spatial navigation 23:166, 23:168
taste and smell 24:29–30
three-dimensional representations 9:491
in toddlers and preschoolers 16:590
visual word recognition 25:215–216
Functional magnetic resonance imaging adaptation (fMRA)
for cortical area studies 5:23, 5:25
Functional models 14:850
Functional near-infrared spectroscopy (fNIRS) 9:476, 9:502–506, 11:603, 11:616
application to language studies 9:503–506
infant cognitive development 4:60–62
optical imaging results in infants and neonates 9:504f
Functional neuroimaging techniques
based on blood–brain interactions 9:487–502
based on electromagnetism 9:477–487
Functional neurological disorder, hysteria 11:507
Functional organization of premotor areas 18:847–850
caudal premotor areas (F2, F3, F4) 18:849
F5 18:847–849
rostral premotor areas (F6, F7) 18:849–850
Functional reduction 15:176
Functional reorganization approach, to language disorders 13:331
Functional solidarity 12:403, 12:405–406
Functional somatic symptoms, of psychosomatic disorders 19:485–486, 19:488
Functional specialization, in occipital lobe 17:130–131
Functional support, disease 22:735–736
Functional system, of brain 14:396, 14:398
Functional theories, organized crime 5:171
Functional theorizing 22:955
Functional transcranial Doppler ultrasound (fTCD) 10:505
Functional typology (linguistics) 14:210–211
Functional-historical analysis of categories 5:263
Functionalism 3:614, 10:567, 12:782–783, 12:787, 13:511, 13:702, 13:774, 14:453–455, 14:671, 21:129, 21:531, 21:586, 22:914–915
AGIL 1:382, 1:393
and anthropological comparison 4:411–412
in anthropology 9:532
approach for nationalism 16:310–311
approach to constitutions 4:711
and business anthropology 3:6
and Christian theodicies 9:540–541
conflict 9:539
and conflict/consensus 4:590–591
and consensus 9:540
counter-movements 9:541
criticism of 9:539–541
definition 9:536
as doctrine/theory 9:532–533
of educational sociology 7:290, 7:293
elite theories 7:397
equilibrium models 9:533
explanations in modes of production 15:715–716
functional prerequisites, and social stratification 9:538
functionalist movement 9:534
history of 9:536–542
hybrid cultural (Boasian/Parsonian) 9:533
ideas 9:532
institutions 12:190, 12:192–194
legal reasoning 13:776–777, 13:779–780, 13:782
linguistic theory 9:471–472
logical incorrectness of 9:540–541
and mental causation 15:173, 15:175–176
as method 9:533–534
methodological issues/problems in 9:534
multivocality 9:532–533
and nationalism 16:291
origins of 9:536
Parsonian 22:993
AGIL functions 9:537
and physicalism 18:110
and psychology 19:413
Putnam on 19:663–664
in social sciences 6:243
and sport 23:314–315
of stratification 15:641
structural 5:36–37
structural (Radcliffe-Brown) 9:533
studies of citizen journalism 3:614
synthesis in 9:538–539
and systems theory 9:541–542
theories of intentionality 12:335
utilitarian (Malinowskian) 9:532–533
view of crime and deviance 22:238
view on political institutions 12:182
Functionalist emotion theory 12:2
Functionalist historicism 9:533
Functionality of opsins in cells 17:270–271
Functional-logical reasoning 14:773
Functionally differentiated society 13:554
Functional–structural theory, Luhmann 14:383
Functioning
and disability 6:467–469
of executive government 8:498–500
Functions 9:532, 21:343
actual domains of 9:522

Functions (*continued*)
agentive 21:344
causal 21:344
of censuses 3:303–304
human body 2:730–733
nonagentive 21:344
ordered relational structures 17:296–297
recovery, and age 20:56
significance tests 21:959–960
status 21:344
urban 24:801–802
analysis of 24:801
Functions of humor 11:418–419
Functions of language 12:783
conative 12:783
emotive 12:783
metalingual 12:783
phatic 12:783
poetic 12:783
referential 12:783
Functions of Language 8:351
Functions of law 13:509–510
Functions redistribution, for access to justice 12:923–924
Fundamental cause theory 22:277
Fundamental democratization 2:906
Fundamental human rights 13:440–441
Fundamental measurement 14:869–870
Fundamental norm 12:509
Fundamental rights 4:704, 8:267
Fundamental science 7:791
Fundamental theorem
of factor analysis 8:705
of natural selection 22:900
of welfare economics 14:538
Fundamentalism 3:544, 10:214, 17:804, 20:342–343, 20:360, 20:381–383, 20:396, 20:409–410, 21:173–174
American Protestantism, case study 20:397–398
controversial features of 20:396
in culture 20:399–400
fundamentalist cultures 20:396
fundamentalist–modernist controversy 20:397
and hybridity 11:452
Islamic 12:741–745
market 5:536–537
phenomenological approach to 20:398
Protestant fundamentalist 20:397
reactions against modernity 9:539
and syncretism 23:883
Funding 9:550–554, 22:605–606, 22:608–609, 22:612
and academic achievement 1:46
big science 9:551–552
changing modes of legitimacy 9:553
for community care in England 4:347–348
controversies 9:552–553
cultural policy 5:464–465
definition and restrictions 9:550–551
for disciplinary associations 22:676
of ELSI program 8:44
in higher education
academic research and employment 1:51
history of social sciences 9:550–551, 9:553
for lesbian health 15:115
policy for science 9:552
regimes 9:551–552
research 9:552–553
research policy 9:552
science for policy 9:552
social/behavioral research 22:643–647
universities 18:308–312
Funding bias, in applied research (criminology and criminal justice) 5:128
Fundraising, philanthropy 3:362
Funerals
death/dying 5:867–868
Hertz's analysis of first and second funerals 5:871–873
of hunter–gatherers 5:870
Funerary archaeology, of ancient Egypt 1:692–693
Fungibility 15:745
Funk (music) 3:803
Funnel plot 15:267–268, 15:268f
Funnell, S.C., works, *Purposeful Program Theory* 14:299
Funscapes
entertainment 7:672
Fur trade, and potlatch 18:714–715
Furet, François 11:109
on American/French Revolutions 1:600
Furnivall, John S. 1:640
works, *Netherlands India: A Study of Plural Economy* 11:86–87
Furtado, Celso 25:749
Furu al-fiqh 12:752–753
Fusiform face area (FFA) 4:674, 8:676, 25:202
Fusiform gyrus (FG) 8:664–665, 13:371, 13:400–401, 14:197, 8:664f
Fusion model 17:621, 24:276–277
minority-group status 15:572
Fusionism 1:604
Future
reawrd for good deeds in 2:490
Future developments, psychological treatments 19:389–390
Future life domain
FTP and FO 9:561–565
Future orientation (FO) 9:562–563
autonomous motivation 9:563, 9:565
behavioral component 9:562
cognitive representation component 9:562
controlled motivation 9:563
domain specificity 9:563
future life domain 9:561–565
future thinking 9:561, 9:563
future time perspective and commonalties and differences 9:563
model expansion 9:562
motivational component 9:562
motivational consequences of 9:564–565
conditions for prompting 9:565
effect on academic achievement 9:564
effect on achieved identity 9:565
effect on resilience 9:564–565
perceived instrumentality 9:563–564
psychological distance 9:563
unidimensional approach 9:562
Future thinking 9:555–560
conceptual development and definitions 9:555–556
episodic contribution to 9:556–557
debated role 9:557–558
executive functions 9:555–556, 9:558
FTP and FO 9:561, 9:563
memory 9:555–559
mental time travel 9:556
non-episodic contribution to 9:558–559
prediction 9:555–557, 9:559
prospection 9:556–558
Future time perspective (FTP) 9:561–562, 19:368–369
autonomous motivation 9:563, 9:565
controlled motivation 9:563
future life domain 9:561–565
and future orientation
commonalties and differences 9:563
future thinking 9:561, 9:563
motivational consequences of 9:563–564
conditions for prompting 9:565
content 9:563–564
extension 9:563
perceived instrumentality 9:563–564
psychological distance 9:563
results from goal setting 9:561–562
Futures-focused school 21:105–106
Futurists, organizational design and form 6:239
Fuzziness in popular music market 3:802–803
Fuzzy coding, in correspondence analysis 5:5
Fuzzy logic
and artificial intelligence 2:36, 3:211
reasoning in 16:926–927
Fuzzy Logical Model of Perception (FLMP) 23:239–240
Fuzzy regression-discontinuity design 20:157
Fuzzy set analysis 4:443
Fuzzy Trace Theory 8:712–713
FWA *see* Federal-wide Assurance (FWA)
FWER *see* Family-wise error rate (FWER)
FXS *see* Fragile X Mental Retardation Syndrome (FXS)
Fynamic dimension of language 10:109

G

G-5 11:313
G-20 10:182–183
G-77 24:301–302
GA *see* Gamblers anonymous (GA); Genetic algorithm (GA)
GABA *see* Gamma-aminobutyric acid (GABA)
GABAergic interneurons 24:534
Gabelentz, Georg von der, works, *psychological subject* and *psychological predicate* 12:97

Gabor filter 8:659–660
face recognition 8:659–660
Gabor Fisher Classifier 8:659–660
GABs *see* Gender atypical behaviors (GABs)
GAD *see* Gender and Development (GAD) model; Generalized anxiety disorder (GAD)
Gadamer, Hans-Georg 1:473, 1:747, 4:794, 11:13, 20:88
on language philosophy 4:790–791, 4:794
on metaphor 15:291
works
Truth and Method 4:794, 10:787, 11:13
al-Gaddafi, Muammar 3:678, 24:451–452, 25:838
Gaddis, S. Michael, on cultural capital 5:380
Gadol, Joan 9:768
Gage 21:403
Gage immobilier 21:403
Gagné, R.M., works, *The Conditions of Learning* 12:197
GAI *see* General Ability Index (GAI)
Gaia 9:567–568
Gaia theory 9:567–568, 20:288
Gail score 10:704
Gain field responses 11:386
Gain–loss dynamics, and successful aging 23:667, 23:669, 23:671
Gains, of organizational group diversity 10:409
Gait, infant motor development after 15:976
Gaitskell, Hugh 22:840
GAL4 method 9:884
Galanter, E.G., works, *Handbook of Mathematical Psychology* 14:814
Galanter, Marc 13:766
Galatians 3:546
Galaxy Zoo 3:632, 3:634
Galbraith, John K. 13:106–107
works, *The Affluent Society* 1:215
Galen of Pergamum (AD 130-200) 15:119
Galenic humoralism 10:608
Galilean idealization 11:517–518
Galileo Galilei 3:778, 14:243, 18:101, 20:286, 21:173, 21:226–227
versus Congregation of the Holy Office 20:286
idealization 11:517–518
research censorship 3:279–280
works
Dialogo 25:344
Dialogue on the Two Principal World Systems 20:286
Letter to the Grand Duchess Christina 20:286
Letters on Sunspots 20:286
Sidereus Nuncius 3:279
Galison, Peter, works, *Image and Logic* 8:579
Gall, Franz Joseph
on cognitive functions 12:416
on neurolinguistics 16:682, 16:682f
Gallagher, Jack 11:687
works, *Africa and the Victorians* 11:687
Gallery owners, and price 5:538
Gallicanism 13:216–217
Gallows humor 11:417
Gallup, George 1:191, 12:672
Gallup, Washington D.C. 23:749–750
Gallup Leadership Summit 2:246
Gallup World Poll 4:402, 5:792, 10:169, 18:459–461, 19:567, 24:658
Galois connection 17:528
Galpin, Charles J. 20:816
Galton, Sir Francis 8:218–219, 8:233, 8:253, 9:569–572, 12:307, 14:811, 17:660–662, 18:174, 19:412, 19:418, 21:264
on anthropometry 9:571
on applied social research 1:851–853
on composite photography 9:571
on development of regression analysis 9:570
dual fascination with heredity and quantification 9:571
heredity studies of 23:425
on law of ancestral heredity 17:661
legacies 23:425
as medical student 9:569
method of word-association 9:571
on nature/nurture 16:340
practice of eugenics 9:569–570
research on human mental heredity 9:569–570
studies of heredity 9:569–570
works
Hereditary Genius: An Inquiry into Its Laws and Consequences 8:237, 12:298, 12:310, 14:811–812
Inquiries into Human Faculty 9:571
Galton Chair of Eugenics at University College London 8:240–241
Galton's box (measurement error model) 14:811
Galton's problem (cross-cultural measurements) 14:811
Galton–Watson branching processes 14:811, 18:579–580
Galvanic skin reflex (GSR) 10:481
neuromarketing 16:701
Galveston Orientation and Amnesia Test 18:637
GAM *see* General aggression model (GAM)
Gamaa i-Islamiya group 20:403
Gambetta, Léon 23:2–3
Gambian study 1:283
Gambier Islands 17:158–159
Gamblers anonymous (GA) 17:594
Gambling 5:719, 16:644
addiction 1:97–99
and victimless crime 25:75, 25:77–79
wanting/needing 25:354–355
Gambling disorder (GD) 17:592–597
Gambling strategy 14:630
Game of imitation, marketing strategies in 14:557
Game of movement, marketing strategies in 14:557
Game of position, marketing strategies in 14:556–557, 14:557f
Game playing, similarities, during infancy 18:219
Game theory 1:665, 4:844–848, 5:97, 6:248, 7:90, 9:573–591, 12:100, 12:194, 14:266–267, 14:549, 16:430, 19:902, 21:531–532
of actions 12:337
asymmetrically prescriptive/descriptive approach 16:431
backward induction 9:586–589
behavioral game theory 9:587–590, 16:430–431
bounds on human rationality 9:587–588
empirical regularities 9:588
learning 9:589–590
modeling player's thinking processes 9:589
players' motivations 9:588–589
and bounded rationality 2:773, 2:775
Cold War 4:122–123
cooperation 4:862–866
correlated equilibria 9:586–587
correlated equilibrium 9:586–587
and correlated equilibrium 5:97
decision making in 9:587–589
definition of 9:582
and deterrence 6:248
elements 16:430–431
of equilibrium 7:916–917
equilibrium refinements and stable equilibria 9:586–587
experimental economics 8:547–549, 9:582, 9:590
Fisher's studies 9:224
general-sum noncooperative games 9:576–578
Nash equilibria 9:577–578
n-Person games 9:579–580
and heuristics 10:830
human behavior 8:404
indirect evolutionary approach and 8:434
institutions 12:190–191, 12:194
and interactive choice 12:388
and interests 12:391–392
irrationality 12:720–721
iterated dominance 9:585–586, 9:588
mathematical psychology 14:813
microeconomics 15:335–336
modifications of game model 9:580–581
games with incomplete information 9:581
games with nontransferable utility 9:580
Nash equilibrium 9:584–585
and Nash equilibrium 5:97
noncooperative game models 9:582–586
perfect equilibrium 9:586–587, 9:590
and power 15:669
Prisoner's Dilemma 18:931–933
rational choice explanation 19:897
rationalizability 9:585–586, 9:588–589
relative deprivation 6:215
representation of games 9:573–575
extensive form 9:573–574
normal form 9:574–575
rules of negotiation 16:430–431

Game theory (*continued*)
for social capital 22:176–177
stable equilibrium 9:583, 9:586–587
standard assumptions 16:430–431
structures 16:430–431
symmetrically prescriptive approach of 16:431
two-person cooperative games 9:578–579
Nash's bargaining model 9:578
threats 9:578–579
two-person zero-sum games 9:575–576
infinite games 9:576
minimax theorem 9:576
mixed strategies 9:575–576
saddle points 9:575
and welfare 25:504
Game-based learning 22:9
Games 11:562
computer assisted instruction 4:494
definition of 9:582
differential 17:255
with a graph structure 4:879
learning 6:900
MC-value for monotonic NTU games 4:878
in partition function form 4:878–879
with a permission structure 4:879
play and 18:216–217
with precedence constraints 4:879
with restricted cooperation 4:879
with side-payments 4:867
without side-payments 4:867
on social networks 16:572–574
definitions 16:572–573
Katz–Bonacich network centrality 16:573
model with strategic complimentarities 16:573–574
Nash equilibrium 16:573–574
sport, anthropology of 23:299–306
of strategy 16:583
theory of 1:768–769
with upper bounds 4:879
with/without side-payments 4:867
Games against nature (heuristics) 10:834
Game-theoretic approach 2:206, 8:126, 15:306, 22:704
Gaming 16:773–774
Gamma activity, in electroencephalography 7:354, 7:359
Gamma data, generalized linear models for 17:259
Gamma distribution 6:577
Gamma oscillations 16:637
Gamma ROC 21:949
Gamma-aminobutyric acid (GABA) 2:823, 16:622, 16:750, 16:753
and sleep 22:89–92
Gamma-frequency band 16:635–636, 25:188
Gandhi, Mohandas Karamchand 16:280, 20:375, 21:178
Gang culture, causes and nature of 25:819–820
in different countries 25:821
emergence and growth of American gangs 25:820–821
media, role of 25:821–822
strategies for preventing gang violence 25:822
in the UK 25:821
Gang Resistance Education and Training (G.R.E.A.T.) 22:413
Ganges river system 25:444–445
Gang-labor system 22:56–57
Ganglion cells 16:634–635, 25:158, 25:184
Gangs
aggregations of gang nations 9:594
change and continuity 9:593–594
contemporary community research 9:593
culture of 9:594–595
definitional issues 9:592
delinquent 6:68–73
distribution and patterning of collective youth crime 9:594
ecological studies 9:593
ethnographic community studies 9:593
excessive behavior of bikers and sports hooligans 9:594
historical background of scholarly research 9:592–593
marginalization of 9:593–594
research trends and needs of 9:595–596
and school safety 21:110
sociology of 9:592–597
street 9:594
working or specialty 9:594
youth gang organization 9:593–594
Gannt chart 16:540, 16:540f
Gans, Herbert 4:364, 18:202–203, 24:730
Ganshof, François Louis, feudalism 9:112
Gantt, Horsley 17:609
Ganzeboom, H.
on cultural consumption 22:331
on cultural participation 22:333
on cultural socialization 22:334
on education attainment 22:332–333
on income 22:332–333
status-seeking theory 22:332
GAO *see* Government Accountability Office (GAO)
Gap studies 13:598–602
Garage biology 21:250
Garbage cans 23:513
in Empirical Implications of Theoretical Models 18:454
GARCH *see* Generalized autoregressive conditional heteroscedastic (GARCH) model
Garden city movement 1:219, 16:137–138, 16:809, 16:813–814, 24:863–865
Garden of Kings (*Bustan as-Salatin*) 11:83
Garden pathing 19:354–355
Garden slave (character) 3:147
Gardening, urban 1:567
Garden-path sentences 8:634, 21:623–625
Gardens of kingship 11:205
Gardiner, S.R. 7:650
Gardner, Carol Brooks 17:532
Garfield, Eugene 21:322
Garfinkel, Harold 6:505, 9:598–604, 10:786
biography 9:598–599
on common sense 9:603
on embodied action 9:598–599
on ethnomethodology 8:204, 10:786, 21:192–197
impact of work 9:603
local cohort of students and colleagues 9:599–601
major works and publications 9:601–603
on members' practices 9:598, 9:600, 9:602–603
on mundane reasoning 9:600
on social order 9:598–602
studies-of-work of 9:600–603
Garfinkel, Howard 17:393
Garifuna 1:624
gender assignment in 9:797–798
Garland, Diana 20:256–257
Garment, fashion 8:835, 8:838
Garofalo, James 20:774
works, *Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization* 20:774
Garofalo, Rafaele 5:233
Garrett, Merrill 18:50–56
Garrison, William Lloyd, works, *The Liberator* 22:40
Garsse, Y. v., works, *A Bibliography of Genocide Crimes against Humanity and War Crimes* 10:6
Garsten, Christina 17:349
works, *Apple World* 16:104
Garudhammas 25:629
Garvey, Marcus 6:691
Garveyism 19:798
GAS *see* Goal attainment scaling (GAS)
GASA *see* Gay Association of South Africa (GASA)
Gaskin, Ina May, works, *Spiritual Midwifery* 2:152
Gassendi, Pierre 15:546, 20:286
Gastrointestinal (GI) disorders
psychology 9:605–611
background 9:605
coping 9:607
depression, anxiety, and stress in functional GI disorders 9:608–609
depression, anxiety, and stress in IBD 9:608
GI tract and brain–gut axis 9:605
health-related quality of life 9:606–607
mental comorbidity 9:608–609
organic and functional GI disorders 9:605–606
psychological therapies in 9:609–610
psychosocial impacts 9:606–608
self-image and sexuality 9:608
social functioning and stigma 9:607–608
social support 9:607
stress 9:606
Gastrointestinal stromal tumors (GIST) 17:926
Gastrointestinal system, psychosomatic disorders in 19:487

Gastronationalism 22:97
Gastronomy
food/cuisine 9:292–294
Slow Food movement 22:96
Gastrulation 2:116–118, 2:821–822
Gate control theory of pain 3:557
Gate way drugs 6:686
Gated communities (GCs) 9:612–617, 24:224–225
club economy 9:615
diffusion of, in different contexts 9:612–613
efficiency 9:616
fear 9:612–613, 9:616
gates and walls for security 9:613–614
landscapes of walls 9:614–615
living by contract 9:615
neighborhood 9:612, 9:614–616
private urban governance 9:612–614
private urban governments, mass growth of 9:612
property prices 9:616–617
security 9:612–616
segregation 9:616–617
social homogeneity 9:616–617
urban morphology 9:612–614
Gatekeepers/gatekeeping 5:440–446, 9:618–622, 10:620–621
artistic/cultural markets 14:565, 14:567
artists as 5:469
cluster 3:903
collective mentoring by affirmative action 9:619
corporate governance 4:935, 4:937
definition 9:618–619
gender variant individuals 24:550
masters and apprentices in science 9:619
in publishing industry 9:619–620
in research funding 9:620
role, historical and cultural variation of 9:620–621
role in science 9:618–619
use of term 9:618–619
Gates, H.L., Jr., works, *Loose Canons: Notes on the Culture Wars* 1:620
Gates, R.T.L., works, *The City Reader* 24:940–941
Gates and walls for security 9:613–614
Gates Foundation 22:663
Gathering, fishing 9:226–227
see also Hunters–gatherers
Gathering societies, inequality 11:895
Gating mechanisms 15:968
GATS *see* General Agreement on Trade in Services (GATS)
GATT *see* General Agreement on Tariffs and Trade (GATT)
Gatterer, Johann Christoph 11:121
Gauchet, M., works, *The Disenchantment of the World* 21:391
Gaudet, Hazel
works
Invasion from Mars 7:563
The People's Choice 7:563
Gaughan, Monica, on alcohol consumption 16:558
Gauss, Carl Friedrich 9:623–627
asymptotic results of 9:625
Gaussian distribution 9:624
iterative fitting procedures 9:626
life of 9:623
on linear fitting problem 9:623–624
matrix algebra 9:626
on method of least squares 9:624
priority disputes with Legendre 9:624
set of $n-q$ recursive residuals 9:625
works
in astronomy 9:625
The Calculus of Observations 9:624
Theoria Motus Corporum Coelestium 9:623
A Treatise on Probability 9:624
Gauss–Doolittle procedure 9:625
Gauss–Hermite quadrature 9:848
Gaussian distribution 6:397, 9:378, 9:624, 19:471–472, 21:955
application in retinal function theories 2:369, 25:159
in Bayesian context 9:624
Gaussian errors, random coefficient models with 23:382
Gaussian example, multilevel statistical analysis 23:383–385
Gaussian likelihood function 24:323, 24:327
Gaussian maximum likelihood methods 24:327
Gaussian mixture model (GMM) 23:262
Gaussian quadrature 16:48
Gaussian random effects models 23:383
Gaussian receiver operating characteristic curve 21:949
Gaussian state space model 1:370
Gauss–Jordan procedure 9:625–626
Gauss–Markov theorem 9:625–626
Gauss–Newton procedure 9:626
Gauss–Seidel algorithm 9:626, 22:483
Gauthier, D., on contractarianism 4:805–807
Gautier, Etienne 10:776
Gautreaux Assisted Housing Program 4:326, 16:444
Gautreaux program 23:159
Gay 10:790–794, 18:533–534, 18:536–538, 21:769
bisexual, gay, gay and lesbian Latinos in US gay culture
see also Latinos
intersectional queer theory 19:765–767
minorities 15:573
movements in United States 9:632–636
neighborhood 19:763
number of youths 9:651
sexual orientation 21:786–789
and social work 21:802–806
tourism *see* Tourism; transgender (LGBT) individuals
in Turkey 20:871
youth 9:649
see also Lesbian
Gay, Lesbian and Straight Education Network (GLSEN) 21:87
Gay activism, in Latin America
see also Latin America
Gay Activists Alliance 9:633
Gay and bisexual men (GBM), Latino
see also Latinos
Gay and lesbian health movement 15:111
Gay and lesbian movement 9:633–634
Gay and Lesbian Organisation of the Witwatersrand (GLOW) 21:745–746
Gay Association of South Africa (GASA) 21:745
Gay cancer 13:861
Gay enclaves 21:755–756
Gay European Tourism Association 9:638–639
Gay families 7:442, 8:769–770, 21:756
Gay friendliness 9:639, 9:641
Gay gentrification 21:810
Gay legal studies 13:792
Gay liberation movement 9:633–634, 15:110–111
drag queens in 13:876
Gay male culture, in San Francisco 9:711
Gay Male S/M Activists (GMSMA) 13:887–888
Gay marriage 4:772
Gay men 9:711
rights of 3:689
Gay movement 10:793
Gay neigborhoods 21:810
Gay Nurses' Alliance 15:111
Gay Pride celebrations (Turkey) 20:873
Gay rights movement 5:76
and counterculture 5:85
Gay Roots 21:750
Gay sexuality 17:486–488, 17:490–491
sexual capital 8:1, 8:5–6
Gay space, nonheterosexual British Asians 16:865
Gay stepfamilies 21:757
Gay Times (magazine) 9:638
Gay wedding tourism 9:641
Gay-related racism 9:628
toward Latinos 9:629–630
GayRomeo 21:794–795
Gay-straight alliances (GSAs), in school 21:88–89
Gaze
like me theory 15:132
stabilization by vestibuloocular pathways 25:66–67
vestibulo-ocular reflex 25:70–74
Gaze cueing 12:27–28
Gaze direction 8:654
Gaze duration 8:632
Gaze following 10:270, 10:272, 12:28–29
Gaze-shifting movements 8:628–630
saccades 8:629–630
smooth pursuit 8:630
vergence 8:630
Gaze-stabilization movements 8:629
optokinetic response 8:629

Gaze-stabilization movements (*continued*)
vestibulo-ocular response 8:629
Gazetta 16:819
Gazetteers 10:61
Gazzaniga, Michael 16:686, 23:294–295
works
Handbook of Cognitive Neuroscience 7:590
Handbook of Psychobiology 7:591
GB *see* Global Barometer (GB)
GBD *see* Global burden of diseases (GBD)
GBTM *see* Group-based trajectory models (GBTM)
GCCs *see* Global commodity chains (GCCs)
GCI *see* GEOSS Common Infrastructure (GCI)
GCM *see* Growth curve models (GCM); Nosofsky's Generalized Context Model (GCM)
GCMs *see* General Circulation Models (GCMs)
GCP *see* Goal-corrected partnership (GCP); Good Clinical Practice (GCP); Gross cell product (GCP)
GCs *see* Gated communities (GCs)
GCSE *see* General Certificate of Postsecondary Education (GCSE)
GCTA *see* Genome-wide complex trait analysis (GCTA)
GCV *see* Generalized cross-validation (GCV); Gross calorific value (GCV)
GD *see* Gambling disorder (GD)
GDDS *see* General Data Dissemination System (GDDS)
GDI *see* Gross domestic income (GDI)
GDN *see* Global Development Network (GDN)
GDP *see* Global Detention Project (GDP); Gross domestic product (GDP)
GDPR *see* European General Data Protection Regulation (GDPR)
GDR *see* German Democratic Republic (GDR)
GE *see* General Electric (GE)
Geary, D.C. 8:394, 8:393f
Geary statistic 23:96, 23:107
Geary-Khamis method 5:756
Gebildete 12:274–275
GED *see* Georeferenced Event Dataset (GED)
Geddes, Sir Patrick 20:134, 21:178, 22:698, 24:863–864
urban planning 1:219
Geddesian approach 20:137
GEE *see* Generalized estimating equation (GEE)
Geek culture 12:591
Geertz, Clifford 1:747, 1:749, 2:74–75, 2:273, 5:500, 7:415, 8:150–151, 9:661–665, 14:435, 14:437, 22:690–692, 23:71–72, 24:291–293
on Bali 9:662–665
on brokers 2:870–871
on cognition 5:577
on culture 5:557, 5:560
on hermeneutics 1:769
on imagination 9:665
interpretation of 9:661, 9:663–664
on meaning 9:662, 9:664
Morocco 9:664
on politics 9:662, 9:664–665
on power 9:661, 9:664–665
on public symbols and private discipline 23:863–864
on symbolic action 9:662–663
on thick description method 9:662
on values 9:664
works
Agricultural Involution: The Process of Agricultural Change in Indonesia 11:86
The Interpretations of Cultures 5:617, 17:671
Islam Observed 16:377
Negara: The Theatre State in Nineteenth Century Bali 11:86
Gehl, Jan 24:866–867
Gehlen, A. 19:314
Gehry, Frank 2:22f
GEI *see* Genotype–environment interaction (GEI)
Geipel, Robert 7:166
Geist (spirit/mind) 9:666–670
cultural anthropology 9:670
dialectic 9:667–668
Dilthey on 9:667, 9:669–670
Hegel's concept 11:818
idealism 9:667, 9:669–670
legal science 9:666–669
linguistics 9:669
mind 9:666–667, 9:669
and nation in Romantic Germany 9:667–670
decline of *Geist* and rise of *Geisteswissenschaft* 9:669
German Historical School 9:668–669
Hegel on 9:667–668
lost legacy 9:669–670
spirit of capitalism 9:670
origins of term 9:666–667
Holy Ghost 9:666
Montesquieu's *Spirit of the Laws* 9:666–667
theology 9:666
Weber on 9:666, 9:669–670
Geistes 11:39
Geisteswissenschaft 9:669, 22:984
Geisteswissenschaften 1:747, 7:585, 10:736, 10:781–782, 10:785, 10:787, 11:75–76, 11:125, 11:380, 18:621–622, 18:625, 22:690
Geistige 12:274
Gelberg–Andersen behavioral model 19:548
Gellner, Ernest André 1:753, 7:147, 9:671–676, 16:291
on anthropology 9:671–672
on anti-relativism 9:675
assessment of 9:675
on elegant theory 16:292
fascination with Malinowski 9:671–672
ideal of civil society 9:674–675
impact as a public intellectual 9:675
life of 9:671
on mode of cognition 9:672
on nationalism 16:292
on nationalism and modernity 9:672–673
on penduum-swing model of Muslim societies 9:673–674
on philosophy 9:671–672
on socialist system 9:674–675
works
Conditions of Liberty 9:674
Muslim Society 16:378
Nations and Nationalism 9:672–673
Thought and Change 9:671
Words and Things 9:671
world history 9:671–672
Gemeinschaft 1:599, 4:338, 4:340–342, 4:361, 4:385, 5:416, 15:701–702, 22:105, 22:264
Gemeinschaft and Gesellschaft
in anthropology 1:766
influence of concepts 1:766
Tönnies' views 1:766
Gemeinschaft der Eigenen (GdE) 9:644–647
Gemeinschaft societies 24:505
Gemeinschaften 23:368
Gemeinschaftsverhältnis 16:357
Gemeinwesen 9:668–669
Geming 3:516–518
Gemonic stability theory 25:401–402
Gender 3:56–58, 4:751–753, 4:755, 5:766–767, 7:449, 8:926, 9:698–703, 9:757, 9:788–799, 10:796–800, 13:256, 15:329, 17:804–805, 18:599–600, 20:457–459, 20:836, 21:44–45, 21:678, 23:635
and academic motivation 9:677–681
acquisition of 21:681–682
expectations, from education system to work 21:682
family factors 21:681
media 21:682
social learning and observational learning theories 21:681
affirmative action 1:262–268
and aging 6:380–381
anthropology of 9:781–787
advances in language, law, and performance 9:784–786
in Africa 9:781–784
crossing threshold in new configuration 9:781
in Eastern Europe 9:782
emergent syntheses and twenty-first century configurations 9:783–784
erasing boundaries and fusing categories 9:781–782
global developments 9:782–783
in Middle East 9:781–782
in postmodern era 9:781
region 9:782–783
religion 9:782–783, 9:785–786
revolution 9:782–783
in archaic states 23:374

assignment 9:795–797
and authoritarianism 2:265–266
of bear subculture 13:890–891
bias, technology and fertility decline 15:611
and big man 2:570
and biographies 2:640–641
biological determinism 9:788–790, 9:793
biological factors 21:678–679
and body 2:720
body image across 2:738–739
Bourdieu on 2:777, 2:781
Butler's work on 3:57
and Catholicism 3:237–242
as central category 9:788
and child neglect 3:412–413
and childhood 3:453–454
in China 2:69
community of practice 9:699
and community sociology 4:363–364
comparative histories of 10:166–167
concept 6:793
and conflicts 16:427
construction of sexes 9:701
constructionist approach 9:788–793
conversational style 9:698
coverage in social sciences 9:789f
and crime 5:132–136
cultural beliefs 21:682–684
and cultural capital 5:378
cultural settings and 21:684
and culture, in Eastern Europe 6:807
and cyberbullying 5:641
and dance 5:667–668
deconstructive approach 9:790–793
deficit hypothesis 9:699
defined 21:678–679
definition 9:742, 13:869
description 20:836
and development geography 6:315–317
development of 21:680
differences in religion 20:380
dimensions of domestic violence law 6:630
and disasters 6:482
discrimination, regarding offspring 15:609–611
distribution in transient global amnesia 24:564
diversity and 3:19
division of labor 8:840, 9:758–759, 25:341–343
and division of labor 13:157, 13:160–161, 25:690
and doctor–patient interaction 6:615
doing gender concept 9:701–702
East Asian studies 6:793–797
eating disorders 6:827
economics of 9:757–763
in economy 9:792–793
and education 7:117–120, 7:175, 9:682–688
  academic achievement 9:682–686
  boys' education 7:119
  cognitive and socioemotional skills 9:683–684
  contested field 7:117
  education market 7:118–119
  educational attainment 9:683
  engagement and behavior in classrooms 9:683
  future directions 9:686
  gender-blind approaches 9:686
  gender-differentiated approaches 9:685–686
  historical perspectives 7:117
  identifying gender bias approaches 9:685
  interests 9:685
  intersecting inequalities 7:118
  neoliberalism 7:118
  new formations in gender theories 7:117–118
  potential explanations for gender gaps 9:683–685
  poverty, and globalization 7:119
  racialization, and social exclusion 7:118
  single-sex schools 9:686
  socialization 9:684
  societal responses to gender gaps 9:685–686
  stereotypes 9:684–685
  test scores and grades 9:682–683
education statistical systems 5:766–767
and engineering 7:646–647
and epigenetics 24:687
epistemology 9:788–791, 9:793
equality 20:836, 25:638–640
equality, and family policy 8:800–801, 8:803
and essentialism 8:14
family size preferences 8:807
and family therapy 8:818
and fashion 8:836
feminism 9:788–793
and feminism 8:899
feminist criminology 8:912–917
feminist linguistics 9:700
and feminist philosophy of science 5:299
and feminist social work 8:935–940
and fertility recovery
  attitudes 9:777–778
  division of housework 9:775, 9:777
  equity 9:777–779
and film genres 9:161–162
formal assignment systems 9:796–797
  morphological information 9:797, 9:797t
  phonological information 9:796–797
and free time 13:846
Fromm on 9:460–462
and general strain theory 23:497–498
and health 15:99
and health care 9:694–697
  medical research 9:694–696
  patient care 9:696–697
  physicians 9:696–697
  providers 9:696–697
  sex 9:694–696
  users 9:696
in history 9:791
and homicide 11:165
and honor 11:182–183
household archaeology 11:221–222
and households 11:228, 11:230–231
Illich's view 11:586
and immigration 8:901
in industrialized countries 9:775–780
  time use for fertility 9:777
and inequality 11:897
inequality, and social stratification 22:710–711
and intelligence 12:318–320
and intersectionality 9:698–699, 9:701, 12:657–667
and interviewing 5:853
and intimate partner violence 12:686–688
in Japan 12:795
and joking 11:417–418
and Judaism 12:875–878
and kinship 13:39–40
and labor history 13:171
and language 13:281–282, 13:285, 14:169, 14:171–172, 22:949–950, 22:952–953
and Law 9:729
and law scholarship 9:732
of leatherman subculture 13:888–889
and leisure 13:846, 13:854
  men and masculinity 13:854
  sexualities 13:854
  women and patriarchy 13:854
and lesbians 13:893, 13:895
lexicology 9:797–798
and liberalism 14:24
and life course 14:76–78
and marital quality 14:513, 14:515
and marriage 14:599–601, 14:603
and matrifocality 14:822–823
Members of the European Parliament (MEPs) 9:699, 9:702
and men's health 15:167–169
  societal messages around 15:169–170
methodology 9:790, 9:792
in Middle East and North Africa 15:356
in Middle East society 16:385
and migration 15:425–426
in military 15:514
modeling 9:798
and nationalism 16:265
Near Middle East/North African studies 16:367–370
and negotiations 16:424
and nostalgia 17:36
number of 9:795
in Oceania 17:160
and organizational diversity 4:738–742, 6:595–599
and parenting 2:475
and pathological gambling 17:593
performativity 9:701–702
performativity of 3:57–58
and physical health 9:704–708
  behavioral factors 9:704–705
  biological factors 9:704
  chronic diseases 9:704, 9:706–707
  coping 9:705–706
  depression 9:705–706

Gender (*continued*)
diet 9:704–707
gap in life expectancy 9:704
heart disease 9:704–707
implications for prevention and treatment 9:706–707
lifestyle 9:704, 9:706
physical inactivity 9:704–705, 9:707
psychosocial factors 9:705–706
smoking 9:704–707
stress 9:705–707
and place 9:709–714
architecture 9:711–712
buildings 9:711–712
cities 9:710–711
cyberspace 9:709, 9:712–713
future of 9:712–713
housing 9:709–712
neighborhoods 9:711
origins of field 9:709–710
private space 9:711
public space 9:709, 9:711
space 9:709–713
suburbs 9:710
urban design 9:711
urban planning 9:710
policies 6:793
contemporary policy, global governance, women's human rights 6:793
and political economy 8:899–900
in political science 9:791–792
and politics 8:888–889
post-Soviet situation 20:838–841
and powerlessness 1:547
practitioner-scholar model 24:522
predominantly semantic assignment systems 9:796, 9:796t
in psychology 9:790–791
queer 19:755–756
queer geography 19:759–761
quotas 8:888–889
racialization, and social exclusion 7:118
relations, in Southeast Asia 23:73
relationship to law
complexity of women's identities 9:731–732
equal access and formal equality 9:729–730
lesbian, gay, bisexual, and transgender (LGBT) issues 9:734
marriage 9:729
meanings of 9:729
portrayals of women 9:732
separate spheres ideology 9:729
stereotypes about women 9:729
subordination and equality 9:730
women in legal institutions 9:729
women's victimization 9:732–733
work/family conflict 9:730–731
-relevant differences in Eriksonian stages 7:941–942
and religiosity of women 25:635–636
and reproduction 20:450–453
reproductive technologies 9:723–724
responsiveness, women offenders 25:619
role in adolescent leisure activities 13:832
roles 20:837
roles and behaviors 21:680–681
roles and identities in fishing 9:227
in rural geography 20:797–798
Russian and Post-Soviet Studies 20:836–842
and same-sex youth 9:650–651, 9:653–654
schema 21:680–681
self-reinforcing nature of 21:682
through media 21:682
and Second Demographic Transition 21:346–347
and self-control 21:482
support 21:483
and sex offender 21:674
sex ratio at birth (SRB) 15:608–609, 15:612–613, 15:609t
sex selection, implications of 15:611–612
sex selective abortion 15:608–609, 15:611–612
sexism 21:693–699
and sexual harassment 21:722–726
and sexuality 21:821–822
and shamanism 21:869
and social class 22:342
and social dominance orientation 22:250, 22:252
social learning and observational learning theories of 21:683t
social media 24:146–152
in social mobility 22:428–429
and social research 9:715–722
bias 9:716
evidence-based medicine 9:719–720
exclusion of WCP and pregnant women in clinical trials 9:720
feminist 9:717–718
feminist methodology 9:717
feminist politics and scientific rigor 9:718
focus on women 9:716
integrity and validity 9:718–719
paradigm shift, methodological revolution 9:717
political economy in women's exclusion 9:721
politics of representation 9:719
positivism 9:717–718
reflexivity 9:717–718
reforms 9:720–721
research gaps 9:716–717
scientific soundness and ethics 9:717–718
shifts in usage of gender 9:715–716
validity in feminist research 9:718
women researchers 9:715
women's exclusion in clinical studies 9:719–720
and social vulnerability 10:549–550
socialization 21:678, 21:680
sociolinguistic studies of 9:700–701
sociology 22:896–897
in sociology 9:792
sociology, Randall Collins 4:214–216
son preference 15:610–611
Soviet era 20:836–838
specific socialization 24:371
spirit possession 23:282–283
stereotype 21:680–681
and street children 23:536–537
strict semantic assignment systems 9:795–796
subjectivity 9:792
and suicide 23:686
and symbolic boundaries 3:801
and technology 21:203
actor-network theory (ANT) 9:726–727
engineering field 9:724–725
exclusion 9:723–724
inclusion 9:726–727
material feminism 9:726–727
social shaping/coconstruction 9:724–726
and telephone 24:159–160
terms 9:795
theoretical developments 9:699
theories concerning 21:683t
and time use 24:369
transition to adulthood 24:572–573
use of term 9:781
and value of children 3:499–500
variationist approach 9:698, 9:700–701
violence 8:901–902
visual turn 25:207–211
wage differentials 25:338–339
and white-collar crimes 25:555
and white-collar workers 25:566–567
Gender and Development (GAD) model 2:766, 8:906
Gender archaeology 9:764–767
androcentric archaeology 9:765–766
approaches 9:765–766
basics 9:765
beginning 9:764
body 9:765–766
context of action 9:765
feminist critique 9:764–767
feminist theory 9:765–767
historic archaeology 9:766
masculinity 9:764, 9:766
material culture 9:764–765
practice 9:766
prehistoric archaeology 9:766
queer theory 9:766
rock art 9:765–766
social identity 9:764–765
Gender atypical behaviors (GABs) 21:678–679
Gender baiting 24:552
Gender bashing 24:550
Gender conflict 4:589
Gender development 9:742
concept of activity settings 9:744
culture and gender 9:742
development of culturally endorsed gender roles and stereotypes 9:747
distal cultural factors in 9:746–747
proximal and distal influences on 9:742–746

biological influences 9:742–743
cultural factors 9:743–746
Gender deviancy 9:733
Gender differences 9:699, 10:796–799
in adult mortality 1:178–179
in aggression 1:377
in caregiver–toddler play, similarities 18:221
childhood 3:453–454
in communication 9:754–755
language use and conversational style 9:754
nonverbal behavior 9:754–755
nonverbal communication 9:754–755
nonverbal decoding 9:755
verbal ability 9:754
cooking 9:291
in cooperation and competition 9:753–754
in aggression 9:753–754
in females 9:753
in males 9:753
in social play 9:754
dimensional *versus* taxonic differences 9:751
doctor–patient interaction 6:615
in effect of beauty on labor market 2:387–388
in fear of crime 8:855–856
in friendship 9:445
interpretation of effect sizes 9:750–751, 9:750f
in labor participation 8:758, 8:760, 8:762, 8:764
life phase 5:511
methodological issues 9:750–751
in oldest-old mortality 15:865
in personality 9:750–756
Big Five domains 9:751–752
broad personality traits 9:751–752
impulsivity, risk-taking, and sensation seeking 9:752
interpersonal circumplex model 9:752
narrow personality traits 9:752
temperament 9:752
and pornography 18:614
in prosocial moral reasoning 7:317
in social behavior 9:750–756
and social media 9:699
univariate *versus* multivariate differences 9:751
in variability 9:751
in work behavior 20:294–295
in wrongful convictions 25:788, 25:790
Gender discrimination 3:431–432
Gender disparities, in sentencing 21:630
Gender dysphoria 24:549
in China 21:738–739
Gender equality 20:836
tourism and MDG3 24:476–479
Gender Equality and Research Project 22:641–642
Gender equity 7:927
Gender expression
androgynous 13:870
butch–femme subcultures 13:869–870
definition 13:869
Gender gaps
in self-regulation skills 9:683–684
in voting 25:304–305
Gender history 9:768–774
entwined terms 9:770–772
body/bodies 9:770–771
outlook 9:772
sexuality 9:771–772
Middle Ages 15:349, 15:351
origins of 9:768–769
rewriting history 9:769–770
masculinities 9:770
Gender identities 3:929, 9:699, 9:788, 20:457, 20:460, 20:463–464, 21:768, 21:792, 21:794, 24:549, 24:551
butch–femme subcultures 13:870
definition 21:768
recognition of 21:771
scholarship in 21:771–772
and sexual motivation 21:760
Gender identity disorder (GID) 24:549
Gender ideology 11:576
Gender inequality 8:893–896, 25:513–514
and female managers 8:885
and fertility, in South Asia 9:80, 9:82
and hookups 11:193–194
sexual exploitation 10:208
sociology 22:896–897
Gender justice, in environmental justice 6:848
Gender mainstreaming, social work 24:755
Gender medicine, definition 9:694
Gender minority, and social work 21:802–807
Gender nonconformity, and same-sex youth 9:651, 9:653–654
Gender outlaws 24:552
Gender pay gap 9:761
Gender performance 8:781
Gender performativity, Butler's theory of 21:768–769
Gender Place and Culture (journal) 6:293
Gender politics 8:799
in Central America 1:625
Gender reassignment surgery (GRS) 24:549–550
Gender relations, and masculinity 14:681–684
Gender revolution
and fatherhood 8:840
policies in modern China 6:795
Gender role analysis, in feminist psychotherapy 8:932
Gender roles 9:788, 20:363, 23:437–438
gender, culturally endorsed 9:747
men 20:837
women 20:837
Gender scholarship, in East Asia 6:795–796
Gender similarities hypothesis 21:693–694
Gender socialization 9:743
fathers' role in 9:744
peers's role in 9:745
Gender stereotypes
families as educational settings 8:734
Gender stereotyping, effects of media 15:38
Gender studies 8:926–928
ethnohistory 8:192–193
in geography 9:689–693
concepts 9:690–691
methodologies 9:691–692
methods 9:691–692
status and emerging directions 9:692–693
themes 9:690–691
women *see* Women's studies
Gender theories 7:117–118
Gender transgression, homophobia/heterosexism 10:792
Gender variant individuals 24:549–553
community perspectives 24:551–553
modern transgender movement 24:551–552
trans* identity politics 24:552–553
diagnostic categories 24:549
GID and gender dysphoria 24:549
standards of care 24:550
medical perspectives 24:549–550
social and legal perspectives 24:550–551
future of trans* rights 24:551
legal recognition and protection 24:550–551
trans* case law 24:551
Gender-based violence 10:192–197, 25:124
in Latin America 24:513, 24:515
Gender-based wage inequality 13:600
Gendered care
and social work practice 9:800–805
care ethics 9:801–804
feminism 9:800–801
social care 9:801–802
social work theory 9:801–802, 9:804
Gendered division of labor 9:20
Gendered legal institution 9:729
Gendered life course 20:593–594, 20:593f
Gendered migration 11:649–650
Gendered Sexuality over the Life Course (GSLC) Model 21:840
Gendered time 24:408
Gender/gendered discourses 13:258–259
Gendering categories 9:790
Genderism 24:550
Gender-responsiveness position 25:619
Gender–technology hybrids 9:726–727
Gender-typed media 21:682
Gendlin, Eugene 20:740–741
Gene activity, epigenetic inheritance 7:833–835
Gene coding 1:37–38
Gene doctrine, Muller's contribution to 16:14
Gene expression 2:453, 23:580
behavior 9:933–934, 9:937
and microbiome analysis 9:960–961
by opsins 17:271
Gene function
measuring variation for behavior in 9:933
Gene hunting 11:311
Gene patenting 11:307
Gene selectionism 22:902

Gene sequencing methodology 9:975–976
Gene structure, measuring variation for behavior in 9:933
Gene therapy 2:620–621, 9:1017
Genealogical hypothesis, and kinship terminologies 13:63
Genealogical sources, census microdata from 3:287–288
Genealogical space (GS) 13:55
  formal representation of computations over 13:58–59
  and kinship 13:56, 13:58
Genealogies 6:139, 10:23–26, 10:28
  and ancestry 1:761–762
  in anthropology 9:806–812
    ancestry 9:806–810
    animal breeding 9:807
    assumptions 9:810–811
    bilum tree 9:806
    colonialism 9:806–810
    descent 9:806–808, 9:810
    diagonal genealogy 9:809–810
    digital genealogy 9:808–809
    imagination 9:810–811
    kinship 9:806–810
    marriage 9:807–808, 9:810
    method 9:807–808
    new reproductive technologies 9:806, 9:808–809
    pedigree 9:806–810
    photomontage 9:806, 9:810–811
    relatedness 9:806
    thinking 9:808
  of clan 3:731
  critique of religion 23:862–863
  definition 9:806–807
  of epistemic communities 7:862–863
  ethical 1:735–737
  and households 11:228–229, 11:231
  of modernity 21:391–392
  remaking of 9:808
  of secularization 21:388–389
  and STS 21:183–184
  tribes 24:646
Genealogy of (Melaka) Kings (*Sulalat'us-Salatin*) 11:83
Genealogy of the present 1:749
Gene-based evolutionary theory 8:437–438
Gene–context relations, of human development 11:279
Gene-culture coevolution 5:391, 5:405–406, 8:403–404, 9:813–818, 11:247, 11:336, 15:817, 16:332, 22:903–904
  AMY1 copy number 9:815
  cultural evolution 5:391, 9:814, 9:816
  cumulative cultural evolution 9:816–817
  disease 9:815–816
  evolutionary anthropology 9:814–815
  high-altitude hypoxia 9:817
  kuru 9:815
  lactase persistence 9:815
  malarial resistance 9:815–816
  microevolution 15:340
  nutrition 9:815
  prosociality 9:817
  sex ratio 9:813–814
  skin pigmentation 9:817
Gene-culture evolutionary theory 5:405–406
Gene-encoding regulator of G protein signaling 2 (RGS2) 22:495
Gene–environment correlations (rGE) 9:827
  active 9:827
  and behavioral genetics 6:321
  evocative 9:827
  and interaction 6:323–324
  passive 9:827
Gene–environment interactions 1:120, 7:844–845, 8:421, 8:465–466, 9:819, 9:827–831, 12:299–300, 12:325–326, 15:210, 15:213, 16:497, 22:156, 22:542
  adolescence 7:845
  allele by 9:828
  and behavior 9:936
  and behavioral genetics 6:323–324
  challenges in identifying 9:822–823
  definition of 9:819–820
  depression 9:819–822
  diathesis stress 9:819–820
  early life 7:844–845
  empirical evidence for 9:820–822
  and evolutionary developmental psychology 8:421
  genetic research, reception and interpretation of 9:823–824
  genome-wide (GWGEI) 9:828–829
  in health and well-being 9:819–824
  heritability by 9:828
  and HPA axis activity 11:480–481
  public opinion 9:823–824
  social science 9:821
  socioeconomic status 9:821, 9:823
  stress 9:819–820, 9:823
  twin studies 9:819
  types 9:819–820, 9:820f
  in utero 7:844
Gene–environment interplay 9:825
  additive genetic and environmental effects 9:826–827
  birth cohort differences 9:825f
  gene–environment correlations (rGE) 9:827
  genetic research terminology 9:826
  research limitations 9:831–832
Genentech 17:584
General Ability Index (GAI) 12:286, 16:723
General aggression model (GAM) 1:375–376, 1:375f
General Agreement on Tariffs and Trade (GATT) 10:156, 10:255, 12:568–569, 12:574, 14:709, 15:72, 16:465, 24:485
General Agreement on Trade in Services (GATS) 10:854, 13:620–621, 13:623
General Allotment Act 13:233
General Analysis of Selection 8:368
General Authoritarianism scale 2:259
General Certificate of Postsecondary Education (GCSE) 5:378–379
General child-rearing climate and style 15:924
General Circulation Models (GCMs) 7:793
General coalitional–worth monotonicity 4:874
General cognitive ability 12:297–298
General Comments on juvenile justice 12:958–959
General Component Latent Trait Model 19:438
General coping theory 4:903
General Data Dissemination System (GDDS) 5:800
General descriptive statistics 13:178
General deterrence 6:255
General duty provisions 17:143
General education, and arts socialization 2:12
General Electric (GE) 11:840, 24:763
General Electric Co. v. Joiner (1997) 8:583, 13:767
General Equal Treatment Act 6:544
General equilibrium theory 7:912, 7:914–917
  microeconomics 15:333–335
General equivalence theorem 17:258
  optimal experimental design 17:258
    quadratic model in one variable 17:258
General (g) factor of intelligence 12:284
General factor of psychosocial development (GFPD) 7:941
General fertility rate 5:739
General genetic law of development 25:317
General Growth Balance method 6:146
General Health Questionnaire (GHQ) 15:188
  happiness and work 10:515–517, 10:519
General hermeneutics 10:778–781
General hospital psychiatry, for psychosomatic disorders 19:484–485
General hospitals
  consultation–liaison psychiatric services in 14:3
  psychiatric admission units in 6:45
  psychiatric comorbidity in 14:1–2
General Household Survey 10:583
General individual–worth monotonicity 4:874
General intelligence 19:412
General intelligence hypothesis 22:406
General jurisdiction 13:616
General knowledge 22:16
General learning model (GLM) 15:6, 15:37–38
General linear hypothesis test 19:729
General linear model (GLM) 9:528–529
General linguistics 18:141–142
General partnership 3:30
General phenomenology of religion 20:346
General Plan East 11:152
General population
  educational level of 25:484–485

structure 18:581
studies, mental illness epidemiology 15:180, 15:182
General post-qualification training
psychiatric education 7:185
General principles
of law 3:670, 12:509
of the legal profession 13:623
General processing tree (GPT) models 6:526
General Register Office (GRO) (UK) 6:171–172, 23:426
General scales of personal intelligence 17:754
Omnibus scales 17:754
test of personal intelligence (TOPI) 17:754
General sensation seeking (GSS) 21:607
General Social Care Council (GSCC) 20:584
General Social Survey (GSS), US 1:267, 5:447, 5:826–827, 5:833, 5:844–845, 8:854, 8:884–885, 12:550, 13:387, 16:556, 17:462, 18:492, 18:494–495, 23:748, 24:658
General sociology 21:967
General strain theory (GST) 23:496–499
community 23:498
conditioning factors 23:496–497
crime 23:495–500
delinquency 23:496–498
drug use 23:495, 23:497
empirical tests of GST 23:497
extensions of 23:497–499
foundation for 23:496–497
gender and 23:497–498
life-course criminology and 23:498
looking back and looking forward 23:499
macrolevel 23:498–499
negative effect 23:496–498
of race–crime relationship 19:839
social control theory 23:496
social learning theory 23:496
strain 23:496–497
General symmetry, principle of 24:137–138
General systems theory (GST) 22:741, 22:744, 22:750
media 15:14
and organizational systems 3:893–894
General theory of fields 9:141–144
General theory of visual object recognition 25:157
General training, cross-cultural training 5:337
General value model, of auctions 2:208
General value type 2:205–206
General-based education, *versus* school-based VET 25:247–248
Generality
of psychological stress 23:585
of self-efficacy 21:512
Generalizability 4:91, 9:834, 17:531, 19:418–419
of evidence, in population studies 8:562–563
kindling 13:33
in qualitative health research 10:695
in sexuality surveys 15:316
Generalizability criterion 4:475–477
Generalizability question 12:665
Generalizability studies (G studies) 9:835
Generalizability theory (G theory) 9:834–838, 19:418–419, 20:248, 20:250
decision maker 9:834
decision studies (D studies) 9:835–836
decisions, types of 9:835
estimating a Person's universe score 9:835
extensions 9:837
generalizability studies (G studies) 9:835
illustrative generalizability design 9:836–837
for decisions 9:837
estimation of variance components 9:836
interpretation of variance components 9:836–837
numerical example 9:836
statistical model 9:836
overview 9:834
software for 9:837–838
universe and facets 9:835
universe of admissible observations 9:835
Generalizable knowledge 4:253
Generalizations 8:358–359, 9:839–844, 21:2–3
abstract constructs 9:840, 9:844
in cartography 3:170
causal connections 9:841–843
exposure therapies 8:606
extrapolating to nonsampled entities 9:843–844
formal sampling theory 9:840
homogeneous irrelevancies 9:841
in mathematics learning 4:66
meta-analysis 9:839, 9:842–844
purposive selection and construct validity 9:841
in qualitative research 19:686–688
random sampling 9:839–841, 9:844
task network models 16:543
Generalized additive model, nonparametric regression 16:876–877
Generalized additive modeling 19:729
Generalized agency 1:82
Generalized analysis of variance (GENOVA) 9:837–838
Generalized anxiety disorder (GAD) 1:805, 3:462, 11:617–618, 23:563
antidepressants for 1:780
psychotherapy for 19:374
Generalized autoregressive conditional heteroscedastic (GARCH) model 24:320
Generalized blockmodeling 19:433
Generalized Context Model (GCM) *see* Nosofsky's Generalized Context Model (GCM)
Generalized conversational implicatures 18:796–797
Generalized cross-validation (GCV) 15:649
model selection 15:649
Generalized Darwinism
evolutionary economics 8:431–432
Generalized entropy class 11:906
Generalized estimating equation (GEE) 16:135, 19:143–144, 23:383
Generalized exchange 8:483, 8:486–487
Generalized Fourier transform 9:377–378
Generalized Graded Unfolding Model (GGUM) 17:846
Generalized growth balance method (GGB) 1:154
Generalized increasing returns 6:604
Generalized inverse projection (GIP) 6:150–152
Generalized least squares (GLS) method 4:416–417, 21:981
mammal brain and body size 4:417
Generalized least squares (GLS) regression 4:416–417, 8:688
mammal brain and body size 4:417
Generalized linear mixed models (GLMMs) 9:845–852, 16:135
accommodating correlation in 9:845–847
model 9:847
best prediction 9:849–850
diagnostics 9:850–851
between and within covariate decompositions 9:851
distribution of random effects 9:850–851
distribution of Y 9:850
missing data 9:851
fitting and inferential methods 9:847–849
Bayesian methods 9:849
computing maximum likelihood estimates 9:848–849
maximum likelihood 9:848
numerical computations for maximum likelihood 9:848
other numerical approximations 9:848–849
software 9:849
nonlinear regression 9:847
types of inferences possible from 9:847
Generalized linear models (GLIMs) 1:659–664, 2:594, 16:135, 17:259, 19:727, 23:409, 23:420
for binomial data 17:259
for event–history analysis 14:351–352
family of models 17:259
for gamma data 17:259
logistic regression 2:594
for Poisson data 17:259
Generalized Lorenz curve (GLC) 11:901–902, 11:902f
Generalized media of interchange, and AGIL 1:389–390
Generalized method of moments (GMM) estimation 12:207, 15:294–296, 23:616
efficiency 15:296
semiparametric efficiency 15:296
time series central limit theory 15:295–296
Generalized model for latent variables 8:704–705
Generalized multinomial logit model 2:551

Generalized onomasiological salience 13:935
Generalized phrase structure grammar (GPSG) 15:751, 15:754
Generalized rank ordered logit model 2:551
Generalized reciprocity 7:15, 11:894
Generalized secretary problem (GSP) 5:907
Generalized seizures 7:852
Generalized self-efficacy 4:905–906
Generalized smooth tests 23:401
Generalized Spearman–Brown formula, parallel measurements model 19:427
Generalized stable survival functions, random utility models 5:902–903
Generalized symmetry principle 5:524
Generalized validity 19:18
Generalizing canonical correlation analysis 19:80–81
Generalizing from a case 3:200
  analytic 3:199
  from another single-case study 3:200b
  transferability 3:199
  working hypothesis 3:199
Generate-recognize strategy, of retrieval 20:610–611
Generating mechanisms, deprivation 6:215–216
Generation effect, source monitoring 23:29–30
Generation of €1000 movement 25:839
Generation Scotland study 14:74
Generation units 9:858
Generation X 9:867
Generational conflict 9:857–858
Generational consciousness 9:858
Generational differences, immigration and crime 5:154–155
Generational equity, and age stratification 1:336
Generational gap
  in attitudes 12:405
  in voting 25:305
Generational revolt 9:857, 9:859
Generations 9:857–860, 9:866–870, 19:9
  age groups in human society 9:857
  in anthropology 9:853–856
    age 9:853–855
    aging 9:854
    cultural construction of intergenerational ties 9:854–855
    culture 9:853–855
    elder care 9:853, 9:856
    habitus 9:853
    identity as facet of social difference 9:855–856
    importance in future work 9:856
    inequality 9:853, 9:855
    kinship 9:853–854
    meanings 9:853
    as principle for structuring society 9:855
    social change 9:853–854
    social organization 9:853, 9:855
    youth 9:854–855
  Arab Spring 9:870
  baby boomers 9:866–867
  birth cohort 9:858
  cohorts 9:866–867
  collective memory 9:866–867
  consciousness 9:866–867
    kernels of 9:867
    in modern youth movements 9:868
  disturbing social trend 9:859
  Generation of 1914 9:866–867
  Generation X 9:867
  generational idea, emergence of 9:858
  genetic programming 2:41, 2:43
  Halbwachs on 9:866
  leisure 13:855
  Mannheim on 9:866
  political *see* Political generations
  politics, and identity 9:870
  problems and potential analysis of 9:859–860
  social and cultural contexts of different patterns of age 9:867–868
  social generations, theory of 9:858–859
  symbolism and organization 9:867–868
  youth 9:867
    modern movements 9:868
    and student rebellion in 1960s 9:868–870
  and youth culture 25:810–811, 25:815–816
  youth in modern history 9:857–858
Generations and Gender Surveys (GGS) 5:737
Generation-set systems 9:855
Generation-skipping trusts 19:174–175
Generative basis of kinship terminologies 13:64–65
Generative entrenchment 24:131
Generative grammar 4:104, 9:871–878, 13:74–75, 13:83–84, 13:272, 13:274, 13:277–278
  adjunct 9:875
  complement 9:875
  conceptual interface 9:877
  deep structure 9:873
  derivational economy 9:877
  extensions and modifications 9:874–875
  grammatical transformation 9:872–873
  head 9:875
  lexical features 9:873
  markedness in 14:534–535
  merge 9:876
  minimalist program 9:876–878
  models of English meters 13:275
  move 9:875
  phonetic form 9:874–875
  phonetic interface 9:877
  phrase structure 9:872–873
  principles-and-parameters Theory 9:875–876
  representational economy 9:872
  standard theory 9:873–874
  surface structure 9:873
  transformational structure and levels of description 9:871–873
Generative Lexicon Theory 13:945
Generative linguistics, approach to cognitive psychology 4:104
Generative logic, in kinship terminologies 13:65
Generative phonology 18:76
Generative poetics 13:272
Generative probabilistic models 21:582
Generative problem, unobtrusive measures 24:770
Generative structures, kinship 13:58–59
Generative theories 13:273, 13:275
Generativity 1:182
  in adulthood 15:369
  and death/dying 5:861
  stage of aging 1:435
  versus stagnation 11:539, 11:541
Generic Conceptual Model 23:131–132
Generic effects
  placebo-controlled studies of 18:161
Generic human being 11:278
Generic model of psychotherapy 19:516
Generic skills 7:215–216
Generic statistical pattern recognition 23:411–413
  Bayes decision rule for classification 23:412–413
  feature selection and extraction 23:411
  flow chart of 23:412f
  probabilistic knowledge representation 23:411–412
Generic strategy 23:508
Genes 2:450
  definition 1:85
  and human mate choice 11:335–336
  interactive effect on behavior problems 2:474
  and longevity 1:394–395
  manipulation of 10:38–39
  and sexual orientation 21:779, 21:781–783
  and social behavior 8:465–466
Genes and behavior 9:879
  from component parts to networks 9:885–886
    networks of genes 9:885
    networks of neurons 9:885–886
  courtship 9:881–882, 9:884
  Drosophila 9:879–881, 9:880f–881f, 9:883f–884f
  from genes to circuits 9:882–885
    anatomical mutants 9:882–884
    'connectome' and behavioral circuitry 9:884–885
    genetic mosaics 9:882
    localization and genetic targeting 9:884
  memory 9:881
  phenotype 9:881–882, 9:885
  from single genes to many 9:879–882
    complex innate behavior 9:880–881
    plasticity and learning 9:881–882
    tropisms 9:879–880
Gene's-eye perspective 13:48–49
Genesis
  of deixis 6:56
  of legal pluralism 23:655

Genesis of discipline, modern psychology, Wundt and 25:793–794
Genetic admixture 9:887
genetic ancestry 9:887, 9:889f
genetic differentiation 9:887–888, 9:891–892
genetic isolation 9:887
genome fragment 9:887
at individual level 9:888–889
complexity of human demographic history 9:890–894
sex-biased admixture 9:889–890
linkage disequilibrium 9:887
migration 9:887, 9:890–892
at population level 9:887–888
Genetic advocacy groups 2:610–611
Genetic algorithm (GA) 1:531, 2:30, 2:41–44, 22:702–703, 23:139
Genetic analysis 9:956
of mortality 15:845
Genetic ancestry 9:887, 9:889f
Genetic ancestry testing 6:442–443, 9:898
ancestry informative markers (AIMs) 9:899–900
genetic admixture 9:901
large-scale projects based on 9:901–902
DTC testing companies 9:902
Human Genome Diversity Project (HGDP) 9:901
International HapMap Project 9:902
National Genographic Project (NGP) 9:901–902
limitations 9:902–903
ancestral populations derived from modern populations 9:902
conflation of race 9:903
contradictory results 9:903
databases 9:902
genetic nonrelatedness 9:903
statistical methods and ancestry estimates 9:903
mitochondrial DNA testing 9:899
single nucleotide polymorphism (SNP) 9:899–900, 9:902
types of 9:899–900
uses and ethical considerations 9:900–901
genealogical information 9:900
group affiliation/citizenship 9:900
human migration analysis 9:900
pharmacogenomic possibilities 9:900–901
Y-chromosome testing 9:899
Genetic and environmental factors, development constraining by 8:421–422
Genetic and epigenetic inheritance, compared 7:835
Genetic aspects, microevolution 15:338, 15:340
Genetic carbon dating 16:761
Genetic citizenship 9:1019
see also Biocitizenship
Genetic considerations, for language evolution 11:332–333
Genetic continuity 17:886–887
Genetic counseling 2:619–620, 9:908–914
concepts of 9:908–910
ethics of 2:614
indications of 9:910, 9:910t
paradigm development 9:908–910
paradigm shifts in 9:909t
process 9:910–911
in social context 9:912–913
Genetic counselors 9:909–910, 9:912–913
Genetic couplehood 9:998
Genetic databases, forensic
criminal justice 9:339–340, 9:343
ethical and societal concerns about 9:342–343
ethnic discrimination 9:343
international landscape 9:340–341
overarching societal and ethical issues 9:343–344
for police 9:341–342
typology of 9:341–342
Genetic deficit theory 8:130–131
Genetic determinism 8:421, 9:991
Genetic development 17:887–888
Genetic diagnosis 9:977
Genetic discrimination 10:36–37
Genetic diseases 9:908–909
abortion 9:994
disability 9:995
family planning 9:998
genetic testing 9:994–995, 9:998
newborn screening 9:997
preimplantation genetic diagnosis (PGD) 9:995–997
prenatal genetic diagnosis 9:994–995
reproductive decision making 9:998
risk 9:998
upon birth
testing for 9:997
Genetic disorders, Williams syndrome 25:579, 25:582
Genetic disposition of anxiety 1:806
Genetic diversity 11:309
in Asia 11:318–319
Genetic drift 11:289
Genetic engineering 9:915–918, 17:270
applications to behavior 9:917–918
biological background 9:915
cloning genes 9:915–916
ethics 20:897
gene therapy 9:917
modifying cells to produce proteins 9:916
modifying organisms 9:916–917
Genetic enhancement, prenatal genetic selection for 10:35–36
Genetic epistemology 18:116–118
cognitive development 18:120
Genetic etiology 15:210
Genetic explanations 8:594
Genetic factors
and HPA axis activity 11:479–480
in organizational psychology 9:919
affectivity (individual differences) 9:920–921
attitudes 9:921
behaviors 9:921–922
boundary conditions 9:923
in cognition (individual differences) 9:920–921
ethical issues 9:923
future research 9:923
legal issues 9:923
mediating effects of 9:922
as mediators 9:922–923
organizational research on 9:919–920
performance 9:921–922
in personality (individual differences) 9:920–921
values 9:921
Genetic inferences 8:715–724
Genetic influences 12:298–302
on religiousness 17:806–807
Genetic Information Non-Discrimination Act (GINA) 9:1005–1006, 10:37, 10:40
Genetic markers, Genographic Project 10:22–25
Genetic methods, in disaster victim identification 9:945–946
Genetic nurses 9:912
Genetic polymorphism 16:495
Genetic predispositions, and aggression 1:378
Genetic profile, Williams syndrome 25:579
Genetic profiling
in disaster victim identification 9:947–948
forensic 9:339
working of 9:339–340
Genetic programming (GP) 2:41–45
applications 2:43
automatic programming 2:41
breeding 2:42
crossover 2:41–43, 2:42f
differential evolution 2:41
evolution, development, and learning 2:44
evolutionary algorithms 2:41–42
evolutionary programming 2:41–42
evolutionary strategies 2:41
generation 2:41, 2:43
genetic algorithms 2:41–44
human-competitive results of 2:43–44
machine learning 2:41
mechanisms behind 2:41–42
methodological issues and future directions 2:44
mutation 2:41–43, 2:42f
natural selection 2:41
population 2:41–42, 2:44
progress and state-of-the-art 2:42–43
reproduction 2:41–42
search space 2:43–44
simulation of societies 2:44
Genetic proximity testing 9:342
Genetic recombination 9:898
Genetic record 11:292–293
Genetic relationship kernel or matrix (GRK) 8:716
Yang–Visscher approach 8:717–718
Genetic research 9:919–920
in genetic and environmental components 9:919–920

Genetic research (*continued*)
in individual differences 9:919
reception and interpretation of 9:823–824
on twin pairs 9:920
variance in trait 9:920
Genetic risks 9:911–912
Genetic salvage 11:309
Genetic screening
aim of 9:927
characteristics of 9:926–927
in community genetics 4:350–351
criteria 9:928
definition 9:926
for disease 9:926–931
ethics 20:698–705
and genetic diagnostics 9:926t
new developments in 9:930
psychosocial impact of 9:929–930
quality of 9:928–929
types 20:701–703
carrier screening 20:702–703
prenatal and neonatal screening 20:702
types of
cascade screening 9:928
at different stages of life 9:927–928
family-based screening 9:928
Genetic services 9:912
Genetic set-points theory 17:886, 17:886f
Genetic structuralism 14:416, 23:627
Genetic studies, and intelligence 12:305–306
Genetic testing 2:619–620
ethics 20:698–705
autonomy 20:699–700
confidentiality of genetic information 20:700
harms and benefits 20:699
justice and fair issues 20:700
privacy and right to know 20:700
public concerns 20:698–699
ethics, in workplace 20:703
guidelines 20:699, 20:701
and reproductive rights 20:473–474
types 20:701–703
forensic testing 20:703
preimplantation diagnosis 20:701
prenatal diagnosis 20:701–702
presymptomatic testing 20:702
testing in work place 20:703
The Genetical Evolution of Social Behavior 22:901
Genetical Society of Great Britain 9:222
Genetical theory of kin selection 13:26–28
The Genetical Theory of Natural Selection 13:26
Genetically modified foods (GMFs) 9:1004–1005
Genetically modified organisms (GMOs) 1:569–570, 2:677, 2:680, 7:811, 7:813, 9:972, 9:977, 9:1004–1005, 10:40, 21:180
Genetically modified plants and animals 1:566
Geneticization 9:976–977
of health/health care 2:620
and human identity 6:444
Geneticization thesis 9:976–977
Genetics 2:619–624, 8:218–219, 8:225, 8:228, 8:470–471, 8:473–474, 9:1001–1002, 10:43–44, 15:86, 15:88–90, 24:175
and aging 1:399, 1:401, 1:404
alcohol 1:507–508
of alcohol metabolism 17:238–239
in anthropology 9:939–944
adaptation 9:940
human variation 9:939–940
population genetics 9:940–941
and appetite behavior 25:470–471
approaches to memory 9:905–907
aspects, microevolution 15:338, 15:340
associations of schizophrenia 21:13
in asymmetry of body/brain 2:115–116, 2:126–127
and attachment 12:232–233
autonomy 9:1005–1006
biobanks 2:621–622
causes of psychopathy 19:451–452
change, culture-driven 5:391
and childhood behavior problems 2:479
children 2:620
clinical, issues in 2:619–620
cognition, evolution of 4:3–4
commercialization and cellular inventions 9:1003–1004
consent 9:1004–1005
counseling 2:619–620
and crime 2:672–673
cultural evolution 5:394–395, 5:398–399
data 9:1002–1003
data reports for participants of Genographic Project 10:25
in different legal fields 9:1006t
and disaster victim identification *see* Disaster victim identification
disposition 2:758
of dyslexia 6:730
and economic behavior 9:951–954
behavioral economics 9:951
behavioral genetics 9:952–953
decision-making 9:952
experiments 9:952–953
genome-wide association studies 9:953–954
heritability 9:951–954
heterogeneity 9:951
molecular genetics 9:953
risk 9:951–953
twins 9:952, 9:954
effects on epigenetics 24:689
ELSI program 8:43–46
versus epigenetics 7:842
DNA methylation and miRNA 7:842
inheritance of modifications 7:842
and eugenics 8:254
in Europe 9:1003–1004
evidence 9:1006
factors in organizational psychology 9:919–925
factors of mania 14:486
findings of OCD 17:115
Fisher's studies 9:222
and forensics
access to researchers 9:959
criminal investigations, effects on 9:956–957
CSI Effect 9:958
databases 9:958–960
DNA Profiling, development of 9:955–956
DNA wars 9:957–958
effects on forensic technologies 9:958
exculpation 9:957
familial searching 9:960
forensic identification 9:955
function creep 9:959
gene expression and microbiome analysis 9:960–961
inculpation 9:956–957
phenotypic profiling 9:960
short tandem repeats (STRs) 9:956–960
surreptitious sampling 9:959–960
use as evidence in court 9:957–958
forward 9:905
genetic exceptionalism 2:620
genetically modified foods (GMFs) 9:1004–1005
and genomics, difference between 9:977
and indigenous communities 9:962–968
influences on sexual orientation 21:779–784
affected sibling pair (ASP) method 21:781
association studies 21:781
behavior genetic studies 21:779
directions for future research 21:783
environment 21:779–780, 21:783
family studies 21:780
fraternal birth order effect 21:783
genes 21:779, 21:781–783
heritability 21:779–780
higher fecundity of female maternal relatives of homosexual men 21:783
linkage analysis 21:781–782
linkage studies 21:781–782
molecular genetic studies 21:780–783
twin studies 21:780
and intelligence 12:291, 12:294–295
of language learning impairment 13:347
and lateralization 13:406–407
of life span 15:845
and media 9:990–993
celebrity scientists, concerns, and publics 9:991–992
double helix 9:990–991
eugenics 9:990–991
future 9:992
genetic determinism 9:991
Genographic Project 9:992
hope 9:990–991
horror 9:990–991
hype 9:991
informatics 9:992
interactivity and mediatisation 9:992
Lifestory 9:990
meaning-making 9:990, 9:992

new genetics 9:990–991
personal genomics 9:991–992
public engagement with science 9:990–991
representation and publicity 9:990–991
scientific heroes 9:991
twenty-first century media 9:992
Mendel's studies 9:223
to mental deficiency 8:222
and mental illness 15:204–205
of mental illness 15:209
methods 9:941–943
biotechnological breakthroughs 9:942
chromosome mapping 9:942
mitochondrial DNA 9:942–943
repetitive DNA 9:943
stains 9:942
Y chromosomes 9:943
and neuroethics 16:664, 16:667–668
new 9:1015–1021
new scientific methods and approaches 9:1002–1003
Nobel Prize for 8:221–222
nondirectiveness 2:620
obesity 17:83
obsessive-compulsive disorder 17:121–122
and the operation of law 9:1005–1007
patents 9:1003–1004
of personality 17:861
gene-by-environment interactions 17:861
molecular genetics 17:861
statistical/behavioral genetics 17:861
personality differences and development 17:884–890
and post-traumatic stress disorder 18:704
and race 16:759–763
regulation 9:1004–1005
reverse 9:906
rights 9:1005–1006
risk 9:1004
role in biology 9:905
role on behavior problems 2:473–474
screening 2:620
and sexual orientation 21:775
side effects (Psychopharmacotherapy) 19:463
and social justice 9:969–974
BRCA1 9:971
BRCA2 9:971
breast cancer test and health disparities 9:971–972
eugenics and progressive spirit 9:970
forensic DNA and wrongful convictions 9:971
genetically modified organisms (GMOs) 9:972
golden rice and food security 9:972
governance 9:973–974
human rights 9:969–972
between individual and society 9:969
Innocence Project 9:971
kinship analysis and transnational justice 9:970–971
opposition 9:972–973
and society
biobanking 9:979
bioeconomies 9:977–978
bio-objects 9:979–980
current state of knowledge and theory 9:977–979
ethics of research 9:979
future directions 9:979–980
genetically modified organisms (GMO) 9:977
genomics 9:976–977, 9:979–980
intellectual antecedents 9:975–977
postgenomics 9:979
recombinant DNA 9:975
stem cells 9:977–980
and sociology 9:984–989
gene–environmental interactions 9:986
history 9:984–985
nature-nurture 9:985–986
role of environment 9:985–988
smoking 9:986–987
social behavior 9:984
twin studies 9:985
studies of behavior 9:932–938
biometric genetic theory 9:934
classical methods 9:932–933
estimating sources of family resemblance 9:934
experimental methods to study genetic variation 9:933–934
family and adoption designs 9:934–935
future challenges 9:937
gene expression 9:933–934, 9:937
gene–environment interplay 9:936
genome-wide association studies 9:933, 9:936–937
heredity 9:932
heritability and interpreting genetic estimates 9:936
linkage studies 9:933, 9:937
measuring variation in gene structure and function 9:933
Mendelian inheritance 9:932–933
methodological issues for future studies 9:936–937
phenotype 9:932–937
polygenic trait 9:934
population stratification 9:937
sample size and statistical power 9:936–937
sibling and twin designs 9:935–936
single nucleotide polymorphism 9:933, 9:936–937
statistical methods based on measured genotypes in humans 9:934–936
type I error and replication 9:937
in subjective well-being 23:649–650
of temperament 24:184–185
testing 2:619–620
and twin studies 24:703–707
uncertainty 9:1002–1003
values 9:1005–1006
and warranty period 17:176
Wright's contribution 25:759
Genetics education, in community genetics 4:351
Genetics services, access and quality of 4:351
Geneva, mortality decline in 15:871–872
Geneva Conventions 3:439, 25:382, 25:384–386
Geneva Emotion Recognition Test 8:678
Genital cutting 10:195
Genital knowledge 21:617
Genital manipulation and masturbatory behaviors in boys 21:840–841
Genital mutilation, female *see* Female genital mutilation (FGM)
Genital surgeries in Africa 5:480
Genitalia, and sexual motivation 21:759–760
Genitival classifiers 3:812
Genitive case in negative environments 16:413
Genitive possessive 18:632
Genocidal massacre 10:2
Genocidal violence 10:2
Genocide 1:224, 10:1–21, 11:369, 18:842
control 10:3–5
crimes against humanity 10:18–20
dehumanization 10:15–21
and ethnic cleansing 8:83–88
extermination 10:3–5
genocidal violence 10:2
history 10:11–12
ideology 10:15–16, 10:18–21
intergroup relations 10:16, 10:18–20
international law 10:15, 10:19
mass killing 10:15
minorities 15:571
minority-group relations 10:20
origins of 10:1–2
versus other types of genocidal violence in war 10:2–3
out-group hatred 10:16
political science theories 10:15–16
psychology of 14:702
autocratic *versus* democratic societies 14:704
cultural characteristics 14:704
cultural devaluation 14:704
difficult life conditions 14:702
evolution of destructiveness 14:703–704
future directions 14:705
group conflict 14:702–703
halting persecution and violence 14:705
instigators of collective violence 14:702–703
monolithic *versus* pluralistic societies 14:704
origins of violence between groups 14:703t
overly strong respect for authority in society 14:704
past victimization or suffering 14:704
prevention and reconciliation 14:705
role of bystanders 14:704
role of leaders 14:704–705
self-interest 14:703
turning against other 14:703

Genocide (*continued*)
violent revolution 14:703
war 14:703
racism 10:15, 10:18–21
recent criminological approaches 10:16–19
integrated theory 10:19
macro–micro–macro theory of collective action 10:17–18
theory of conformity 10:18–19
violentization and circumplex theory 10:17
social justice 22:394
socialization 10:18–19
state crime 10:15–16, 10:18–21
techniques of neutralization 10:16
as unilateral 10:4
war 25:377–380
in war 10:1–2, 25:367
war crimes 10:15, 10:20
and war crimes tribunals 25:382–387
xenophobia 10:19
see also Massive category killing
Genocide
future of 10:13
history 10:8–13
antiquity/city states (iron age) 10:9–10
antiquity/feudalism (bronze age) 10:8–9
early modern times 10:10–11
Middle Ages 10:10
tribal (primitive) society 10:8
twentieth century to end of World War II 10:11–12
world since 1945 10:12–13
International Criminal Court definition 10:6
meaning 10:6
study of 10:6–8
discipline 10:6–7
subjects of research 10:7–8
term 10:6
Genographic Legacy Fund (GLF) 10:24, 10:27
Genographic Project 9:992, 11:309
ancestry 10:22, 10:24–25, 10:28
autosomes 10:23
components 10:23–24
education (public participation) 10:24
Genographic Legacy Fund 10:24
research 10:23–24
genealogy 10:23–26, 10:28
genetic markers 10:22–25
inception 10:22–23
inheritance 10:23, 10:28
lineage 10:23, 10:25–27
mitochondrial DNA 10:23, 10:25–27
objectives 10:23
phase I 10:24–25
phase II, Geno 2.0 10:27–28
public participation 10:25
research component 10:24–25
community outreach and engagement 10:24
genetic data reports for participants 10:25
Genographic Project IDs 10:25
human subjects 10:25
laboratory analyses 10:25
project database 10:25
sample and data collection 10:24–25
research support 10:25
results, phase I 10:26–27
Genographic Legacy Fund 10:27
public participation 10:27
research component 10:26–27
y-chromosome 10:23, 10:25–28
Genographic Project IDs (GPIDs) 10:25
Genome biobank boom 1:760
Genome Institute of Singapore (GIS) 11:318
Genome research, ethics of 2:616–617
Genome revolution 19:827–829
Genomes 9:826, 10:39, 10:43
and bioethics, microevolution 15:339
Genome-wide association study (GWAS) 2:452–453, 7:842, 8:715–718, 9:826, 9:979–980, 10:32, 12:301, 15:212–214, 15:845
of behavior 9:933, 9:936–937
and economics 9:953–954
Genome-wide complex trait analysis (GCTA) 2:452, 12:301–302
Genomic imprinting 7:850, 8:450–451, 18:901, 22:910–911
epigenetic inheritance 7:834–835
and kinship 13:49
Genomic medicine, in Asia 11:318–321
Genomic sovereignty 16:759, 16:762
Genomics 2:619–624, 2:677–679, 2:681, 9:976–977, 9:979–980, 9:1015–1019, 11:312–313, 11:315–317, 16:762
artificial genes or genomes 10:39
and bioethics, relationship between 2:619
bioethics in 2:625, 2:627
biotechnology 10:32–33, 10:38–40
chimeras 10:38
cultural consequences 10:40–41
ethical issues in 10:32–41
ethnic identity 10:34
eugenics 10:32–33, 10:36
family identity 10:34
forensic identification 10:33–34
genetic discrimination 10:36–37
genetically modified organisms 10:40
germ line gene therapy 10:38
group histories 10:37–38
HGP and 2:621
human variation 1:758, 1:760–762
humans, genetics research 10:39
and identity 2:623–624
identity issues 10:33–34
individual ancestry 10:37
intellectual property 10:40
international perspectives and cultural issues 2:624
issues arising from medical successes 10:32–33
manipulation of genes 10:38–39
and media 9:990–993
patents 10:39–40
personal identity 10:34
personalized medicine 17:925, 17:927, 17:930
postnatal genetic testing 10:36
prediction of behavioral traits 10:37
prediction of future 10:34–37
prenatal selection for avoiding disease 10:35
prenatal selection for genetic enhancement 10:35–36
privacy 10:40
recreational and behavioral 2:623
research 10:39
integration of ELSI program with 8:44–45
somatic cell gene therapy 10:38
state-sponsored prenatal genetic selection 10:36
vision for 11:314f
Genopolitics 25:307
Genotype–environment interaction (GEI) 6:322, 8:715, 8:717
Genotypes 2:450, 9:826, 10:42–45, 11:289
description of 10:42–43
and environment 10:44
epigenetic inheritance 7:832
in humans 9:934–936
relationship with phenotype 10:43–44
GENOVA *see* Generalized analysis of variance (GENOVA)
Genpact 3:46
Genre analysis 19:673
Genres 13:290–291
anthropological writing 1:742–743
categories, consequences of 3:801–803
classifications 3:800–801
cultural divisions by 5:729
definition 3:799
diversity 3:803
film 9:160–164
folk 9:282
of humor 11:417
knowledge 16:556
media imperialism 15:57
writing instruction 25:768
GENSTAT (statistical package) 23:378
Gentile, Emilio 18:434–435
Gentlemanly capitalism 11:687–688, 14:528
Gentleman's war 25:380
Gentrification 10:46–52, 16:440, 20:541, 24:790, 24:793–794, 24:839–842, 24:931
comparative urbanism 10:50–51
definitions 10:46–48
Global South 10:50–51
globalization of 10:50–51
key theories and debates 10:48–49
neighborhood 10:46–50
new middle class 10:47–49
origins of the term 10:46
post-industrial society 10:48
postsocialist cities 3:607–610
and residential segregation 20:541
social class 10:50–51
stages/waves of 10:46–47

state gets involved 10:49–50
suburbanization 10:48
types 10:47–48
uneven development 10:46, 10:48
of urban neighborhoods 16:452–453
urban policy 10:50
urban redevelopment 10:47
urban renewal 10:49
Gentry 1:965
derivatives of 1:965
European history of 1:965–969
Gentry controversy, English Revolution 7:651
Genu 12:416
Genuine charisma 18:765
Genuine philosopher, attitude of 25:600
Genuine religion 16:805
Genuine slave societies 22:49–51
GEO *see* Group on Earth Observation (GEO)
Geo Web Portal (GWP) 23:134
Geoarchaeologist 7:693
Geoarchaeology 7:693–694, 10:53–57
artifact provenance 10:53, 10:56
grains and particle size analysis 10:55
micromorphology 10:53, 10:55–57
paleoenvironmental reconstruction and human impact 10:55–56
physical analysis of raw materials and archaeological artifacts 10:56
instrumental analytical techniques 10:56
thin-section petrography 10:56
x-ray fluorescence analysis, inductively coupled plasma-mass spectrometry and inductively coupled plasma-optical emission spectroscopy 10:56
principle of sedimentation 10:55
sampling soils and sediments 10:55
site discovery and definition 10:53–54
aerial photography and satellite imagery 10:53–54
coring archaeological sites 10:54
other techniques of remote sensing 10:54
remote sensing/archaeological prospecting 10:53
site formation process 10:54–55
human disturbance process 10:54–55
nonhuman disturbance process 10:54
site survey techniques 10:53
social 10:56–57
stratigraphy 10:53–56
Geoarchaeology 7:696
Geocentric theory 20:285–286
Geocentrism 8:168
Geochemistry 17:166
Geochronology 7:695–696
Geocoding 10:59, 24:261
Geocomputation 14:789
GeoDa program 23:203
Geodesic (mathematics) 16:532
distance 16:532
Geodetic datum 14:285
Geodetic surveys, in archaeology 23:744
Geoeconomics of regions 20:125
Geoengineering 7:794
GeoEye 20:423
Geographers: Biobibliographical Studies 10:69
Geographic ancestry 6:442
Geographic data 23:178
Geographic disparities, and children/infant mortality 12:21–22
Geographic distance, and internal migration 12:514–515
Geographic factor 16:335–336
Geographic features
data quality 23:128
describing 23:127
modeling time 23:127–128
terminology 23:126t
Geographic influence 16:335–336
alternatives to 16:336–337
Geographic information 10:174, 10:176–177, 23:130, 23:132
Geographic information science 1:847, 3:619
Geographic information system (GIS) 3:159, 3:162, 3:294, 3:619–621, 3:623, 6:162, 6:164–168, 7:800, 7:813, 10:58–68, 10:70, 10:86–88, 16:406, 19:735, 20:148, 20:420, 23:85–86, 23:114, 23:126, 23:130, 23:136, 23:731, 23:744–745, 24:265, 24:417, 24:470, 24:617–618, 24:803–804, 24:824, 24:890
analysis with 10:64
application to social sciences of environment 10:67–68
environment and society 10:68
land use and land cover 10:67
linking household-level data with remote sensing data 10:67
spatial modeling 10:67–68
spatial stratification/extrapolation 10:67
and archaeology 10:97–98
essentials 10:59
in future 10:62–63
hazards 10:68
natural resources 10:64
as research infrastructure 10:60
role in cartography 3:169
role of 23:137–138
satellite 10:64–66
visualizations based on 15:636
Geographic location 14:284
division of labor 25:341
Geographic mobility *see* Migration
Geographic Perspectives on Women (GPOW) Specialty Group 9:689
Geographic phenomena 23:126
Geographic profiling 23:144–145
Geographic reality, concept of layers 10:59f
Geographic referencing, absolute Location 14:285
Geographic Resources Analysis Support System *see* GRASS (Geographic Resources Analysis Support System)
Geographic scale, social movements 22:439
Geographic scholarship on poverty 18:726
Geographic space 4:79, 14:284–285
Geographic thought 10:69–75
cartography 10:70
coexistence of perspectives 10:73–74
corporate 10:69
critical approaches 10:74
cultural approach 10:72–73
experience of nature 10:72–73
experience of place 10:72–73
experience of society 10:72–73
history of 10:69–70
narratives 10:69–70
natural philosophies 10:71–72
natural sciences 10:71–72
postcolonial interpretations 10:74
postmodern interpretations 10:74
regional approach 10:70–71
scientific geography 10:69
social science approach 10:72
vernacular 10:69
Geographical access 1:64–67
definitions 1:64
measurement 1:65–67
factors affecting 1:66
types of 1:66–67
normative views 1:65
and spatial equity 1:64–65
use of term 1:64–65
Geographical determinism 2:710
Geographical distance 12:405
Geographical distribution
primates, evolution of 18:904
telephone 24:158
Geographical gerontology 1:407
Geographical history 11:21
Geographical materialism 14:674–675
Geographical mobility 22:426
Geographical position systems (GPS) 5:738
Geographical practice 10:83
regional 10:84
systematic 10:84
Geographical proximity 11:828–830
Geographical referencing 5:840–841
Geographical region, urbanization by 24:960t
Geographical specificity, population censuses 3:290
Geographically based economic data (G-Econ) 7:49
Geographically based online communities 4:335–336
Geographically distinct communities, pretend play in 18:871–872
Geographing 10:93–94
Geography 7:720, 7:722–723, 10:83–90, 15:504, 20:146–148, 25:825
of aging 1:407
emerging critical themes 1:409–410
emerging critical theory 1:410–411
places 1:409
spaces 1:407–409
of Caribbean 1:638
of children 3:491
of communication research 4:279–280
of consumption 4:744–750
contemporary discipline 10:85–88
geography in action 10:87–88
social theory 10:86–87

Geography (*continued*)
spatial science 10:86
critical realism in 5:277–280
of cyberspace 23:84–85
as description 10:70–71
development *see* Development geography
of disability 6:458–461
discipline creation 10:84–85
-isms 10:85
new geography 10:84–85
schisms 10:85
turns 10:85
education 22:728–729
of education 7:166
of energy 7:633–636, 7:638–639
of entrepreneurship 7:686–691
of ethnic conflict 8:91–94
as explanation 10:71–72
feminist studies in 9:689–693
food 10:76–82
Freyre on 9:420–421
gender studies in 9:689–693
and human ecology 22:266
Humboldt on 25:279–282
industrialization 11:869
intellectual origins 10:83–84
international business 12:484–489
market geographies 14:544–548
medical *see* Medical geography
of networks 16:567–571
of organization 17:301
and political ecology 18:350–354
postcolonial 18:644
and postmodernity 10:72–73
and poverty 18:729–730
qualitative methods in 19:682–684
collection of methods 19:682
continuing contributions 19:684
methodological issues 19:683–684
in philosophical and theoretical turns 19:682–683
refugees 20:111–116
regional 20:117–120
scale in 21:1–3
as science of terrestrial locations 10:70
of social movements 22:437
and state 7:721
of state 23:356–361
territory/territoriality 24:221
theoretical revolution 10:84–85
transitions 7:639–640
urban 24:800–806
of votes 7:345–346
Geohistory 1:710–711
Geoid 14:285
Geoinformatica 10:60
Geolinguistics 14:227
Geological archives 7:692–693
Geological metaphor 7:62–63
Geometric cues, use in infants 23:121–122
geometric cue use in reorientation by toddlers 23:121–122
nongeometric cue use by toddlers 23:122
processing of geometric information in young infants 23:121
Geometric distribution of cell 6:582, 21:529
Geometric figures 10:499
Geometric information, processing in young infants 23:121
Geometric mean 18:589
Geometry 13:70
axiomatic 2:339
graphics 10:360–368
one variable 10:360–362
three variables 10:365–368
two variables 10:362–365
legal reasoning 13:778–779
Geophysical hazards 6:480
earthquakes 6:480
erosion 6:480
landslides/rockslides 6:480
poisonous gases 6:480
sandstorms 6:480
soil loss and contamination 6:480
volcanoes 6:480
Geophysical surveys 9:335
Geophysiology 9:567
Geopolitical groups, ethnic groups as 8:136
Geopolitical traditions 10:91
classical 10:91–93
critical 10:93–95
Geopolitics 10:91–96, 25:372–373
in Arctic 1:927, 1:931
and infrastructure 12:122
political geography 18:374–377
of regions 20:125–126
in United States 1:617–618
and urban violence 24:954–955
Georeferenced Event Dataset (GED) 5:836
Georg Eckert Institute in Germany 24:259
George, Lloyd 13:14–15
George, Stefan 3:355
George C. Marshall Foundation 7:793–794
Georgian Order (historical archaeology) 11:1
Georgian word order 25:645
Georgics 14:252
Geospace 14:284
Geospatial analysis 3:621–623, 23:744–745
Geospatial applications, in archaeology 10:97–103
Geospatial data 23:126, 23:128
Geospatial income measures, economic growth 7:49
Geospatial techniques 20:419–420
GEOSS *see* Global Earth Observation System of Systems (GEOSS)
GEOSS Common Infrastructure (GCI) 23:133–134
Geostatistics 23:105, 23:191
Geovisualization 3:162
Geoweb 3:619, 3:621–623
Gerbner, George 4:279
on television 25:80–81
Geriatric Hopelessness Scale 11:198
Geriatric jurisprudence 13:424
Geriatric psychiatry 10:104–108
Alzheimer's disease 10:104–107
conceptual background 10:104
delivery of care 10:105–106
dementia risk 10:105
demographic change 10:105
depression 10:104–107
development of field 10:104–105
importance 10:105
medicine 10:104–106
mental disorders 10:106–107
professional organizations 10:105
research 10:107
scope 10:104
treatment 10:104–107
Geriatrics 13:420, 13:424
and warranties 17:173–175
GERM *see* Global Education Reform Movement (GERM)
Germ line gene therapy 10:38
Germ plasm 19:802–803
Weismann's theory of 19:802–803
Germ theory 8:248
as analogy for Dark Triad 5:674
German, and concept of introspection 12:703
German (language), variations of semantic differential 21:566–567
German academic community of historians 11:122
German Aging Survey 6:379
German ALLBUS 5:827
German barbarism 21:368
German Basic Law 1949 13:440–441
German *Bürgerliches Gesetzbuch* 3:670
German Civil Code (BGB) 3:671, 19:194, 21:403–404
German concepts, of personality disorder 17:892–894
German criminal law 12:364
German Democratic Republic (GDR) 4:281, 20:635
militarization of 15:485
German Ecological Democratic Party 10:380
German elites, homogeneity of 7:399
German Federal Cartel Office 18:344
German foundations, of modern geography 16:334–335
nature–society science 16:335–336
spatial–chorological science 16:334
ties between two views 16:335
German 'free law' movement 14:276
German Greens 10:379
German Historical School 9:668–669
German historicism 11:381
German historiography, welfare state 25:529–530
German history 25:292
German idealism 11:381
German League for National Regeneration and Heredity 8:218–219
German machinery 11:869
German National Educational Panel Study 5:847
German Nazism 7:514
German pharmaceutical companies 17:585–586

German psychology, mathematical developments in 14:810–811
German *Referendar* 3:672
German relative pronouns 23:892
German Research Council 22:655
German romanticism 15:684
German Social Democratic Party 12:280
German Societies for Population Policy 8:220
German Society for Racial Hygiene 8:218–219
German Socio-Economic Panel (GSOEP) 5:744–745, 6:379, 8:108–109, 10:516–519
German *Sonderweg* 2:553–554, 4:406–407
German Sterilization Law 1933 8:235
German sterilization model 8:220
German system of medical education 15:81
German universities 11:121, 22:591, 22:593, 24:763
German *Völkisch* Protection League 16:257
German working class, Moore's contributions 15:771
German youth movements 9:868
Germanic alliteration 13:276–277
Germanic languages 11:822, 23:718
German–Polish nonaggression treaty that 16:259
German–Soviet nonaggression pact 16:259
German–Spanish Treaty (1899) 17:161
German-speaking countries 22:328
Germany 16:259
  anthropology of 1:753–754
  apprenticeship and school learning in 1:861–864
    adolescence 1:861–863
    coaching 1:864
    cognitive apprenticeship 1:863–864
    exploration 1:863–864
    modeling 1:864
    scaffolding 1:864
    school instruction as apprenticeship 1:863–864
    school-to-work transition 1:862
    self-regulation 1:863
    situated learning 1:861, 1:864
    vocational orientation in school 1:862–863
    vocational training 1:861–862
  Beck on 2:396, 2:398
  childcare accessibility in 3:397
  civil code 19:194, 21:403–404
  clinical psychology 3:873–874
  Cold War 4:129–130
  confederation 4:573
  enlightenment in 7:656–657
  ethnic identity of immigrants in 8:110
  eugenics 8:226–227
  euthanasia in 8:282–284
  Grundschuld 21:403–404
  homosexual rights movement in 9:644
    divisions during Weimar Republic 9:646–648
    effects of World War I on 9:645–646
    in late nineteenth century 9:644–645
    rise of Hitler to power 9:646–648
  immigration policy 11:662–666
  industrialization 11:869
  laicization in 13:218
  literacy in 14:239
  migration to 15:382f
  modernization theory in historical work 15:702–703
  national character in 16:233–234
  origins of *Statistik* 23:423
  public intellectuals in 22:699
  during Reformation and confessionalization 20:101
  romantic, spirit in 9:667–670
    decline of *Geist* and rise of *Geisteswissenschaft* 9:669
    German Historical School 9:668–669
    Hegel on 9:667–668
    lost legacy 9:669–670
    spirit of capitalism 9:670
  social insurance in 10:663
  sociology 22:992
  stratification and social classes in 22:340
  tax system 23:340
  and World War I 9:212
Germinal choice 16:15
Germline gene therapy 2:617
Gernsbacher, Morton, works, *Handbook of Psycholinguistics* 7:590
Gerontechnology 1:151
Gerontocracies 1:320–321
  and youth culture 25:810
Gerontological geography 1:407, 1:409
Gerontological social work 7:335, 7:337–339
  assessment 7:338
  care management and therapeutic interventions 7:338–339
Gerontological Society in United States 1:325
Gerontology 2:351
Gerry, Elbridge 7:343
Gerrymandering
  apportionment 7:342–343
  purposes of 4:649
  United States Congress 4:649–650
    incumbency advantage 4:650
    primary elections 4:649–650
Gerschenkron, Alexander 6:276
Gerschenkron's typology 11:876
Gerteis, J., on narrative networks 16:214–215
Gerth, Hans 25:463, 25:465
Geschichte 11:73
Geschichte der romanischen und germanischen Völker von 1494–1514 25:290–291
Geschichte und Gesellschaft 11:45–46
Geschichten 11:73
Geschichtliche Grundbegriffe 4:551, 12:257, 18:489
Geschiedenis van Nederlandsch Indië (1930) 11:84
Geschlecht und Charakter, Weininger 25:611
Geschwind, Norman 16:685
Gesellschaft 1:599, 1:766, 4:338, 4:340, 4:361, 4:385, 15:701–702, 22:105, 22:264
Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS) 12:550
Gesellschaftsgeschichte 2:554–556, 11:104
GESIS *see* Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS)
GESIS: Leibniz-Institute for the Social Sciences, Mannheim 23:749
Gesner, Johann 16:681
  works, *Samlung von Beobachtungen aus der Arzneygelahrtheit* 16:681
Gestalt heuristic of proximity 10:830f
Gestalt psychology 4:52, 7:240, 12:79, 13:703, 19:414, 22:528, 25:157
  and cognitive psychology 4:104
  Heider and 10:750, 10:754
  and introspection 12:703
  Köhler and 13:123
  in social network analysis 16:524–525
Gestalt theory 10:429, 13:124–125, 19:493, 23:280
Gestation 2:698
Gestational age, influence in adulthood 6:752–754
Gesticulations 10:109–113
Gestsdottir, S., on adolescent self-regulation 21:548–549
Gesture Continuum 10:110–112
Gesture processing 13:375
Gestures
  beats as 10:111–112
  of blind 10:112
  of children, and intentionality 12:348–349
  cognition–visuospatial-actional 10:113
  context and psychological predicates of 10:112–115
  deictic 10:111
  and delayed auditory feedback (DAF) 10:112
  and discourse purposes 10:114
  dynamic dimension 10:109–110
  fluency and 10:112
  and goal-directed behavior 10:271–272
  and human communication 11:256
  in human–computer interface design 10:111
  iconic 10:111
  in information exchange 10:112
  inoculation against stuttering 10:112
  and joint attention 12:844–849
  and language 15:583–585
  and language evolution 11:331–332
  in linguistics 10:109–120
  metaphoric 10:111
  mimicry and 10:117–118
  neurogestures 10:111
  as a part of GP 10:116
  from practical actions 10:118–119
  psychological predicate of 10:113
  quotable 10:110–111
  as a referential symbol 10:111

Gestures (*continued*)
shared 10:117
signs 10:111
'social resonance' of 10:118
space 10:110
speech-framed 10:110–111
static dimension 10:109
temporal binding of speech and 10:112
transcription and measurement 10:111
Gesundes Volksempfinden 16:357
Get out the vote (GOTV) 25:310
Get tough on crime policies 5:194
Getting noticed (life phase) 5:511
Getting-to-Outcomes (GTO®) 7:580–581, 7:581t
Gettino, Octavio, works, *Hour of the Furnaces* 6:622–623
Getty Center 2:24f
Getty Foundation 5:417
Getty Research Portal 2:24
Geurts, Kathryn, works, *Culture and the Senses* 21:618
Gevisser, Mark
on Forest Town party police raid 21:744–745
on GASA 21:745
on gay rights movement in South Africa 21:745–746
GEW *see* Geneva Emotion Wheel (GEW)
Gewohnheitsverbrecher 16:357
GFC *see* Global financial crisis (GFC); Great Financial Crisis (GFC)
GFE *see* Goodness-of-fit ethics (GFE)
Gf-Gc theory of intelligence 12:291–292
GFPD *see* General factor of psychosocial development (GFPD)
GGB *see* Generalized growth balance method (GGB)
GGS *see* Generations and Gender Surveys (GGS); Green Goods and Services Survey (GGS)
GGUM *see* Generalized Graded Unfolding Model (GGUM)
GGvis software 16:46
GH *see* Growth hormone (GH)
Ghana, affordable private schools in 1:269
Ghazi thesis 9:468
Ghaziani, A., on social network analysis 5:560
Ghent unemployment insurance system 12:181
Ghettoisation 24:921
Ghettos 4:525–526, 4:528, 10:121–126
diaspora 10:121
disentangling poverty, segregation, and ethnic clustering 10:123–125
folk concept 10:121
Janus-faced institution of ethnic closure and control 10:122–123
shifting notion 10:121–122
tainted identity 10:125–126
and urban poverty 24:915–916
Venetian model 10:122
Wilson on 25:586–587
GHGs *see* Greenhouse gases (GHGs)
Ghiberti, Lorenzo, works, *Autobiography* 20:427
Ghitani, Gamal, works, *Zayni Barakat* 16:385–386
Ghonim, Wael 16:801–802
Ghost authorship 20:530
Ghost cities 21:919
Ghost Dance of 1870 15:531
Ghost towns 2:243
Ghosts, and death 5:872
GHQ *see* General Health Questionnaire (GHQ)
GHRH *see* Growth hormone-releasing hormone (GHRH)
GI *see* Gastrointestinal (GI) disorders
G.I. Bill *see* Servicemen's Readjustment Act 1944
Giant three (Extraversion, Neuroticism, and Psychoticism)
alternatives, five-factor model (FFM) 2:562
Giap, Vo Nguyen 20:621
Gibbon, Edward 19:676, 22:697
works, *Decline and Fall of the Roman Empire, Hume's History of England* 3:63–64, 11:73
Gibbs, Josiah Willard 9:218
Gibbs distributions
probabilistic grammars 19:13–14
inference 19:14
probabilities 19:14
relation to probabilistic context-free grammars 19:14
Gibbs sampling 14:569–570, 16:878, 16:880–883, 17:731, 14:570f
algorithm 15:764–765
Bayes' theorem 2:381
Gibney, Alex, works, *Enron: The Smartest Guys in the Room* 6:621
Gibson, Eleanor Jack
perceptual learning 10:127–134
works
Principles of Perceptual Learning and Development 10:127
Gibson-Graham, Julie, on rape 21:748
Gibson-Graham, Katherine, on rape 21:748
GID *see* Gender identity disorder (GID)
Giddens, Anthony 10:135–138, 22:994, 23:201, 25:83–84
on agency 10:135–136, 10:138
on class 10:135–137
on culture 5:557
on duality of structure 10:135–136
on globalization 15:705
on historical materialism 10:136
on institutional analysis of modernity 15:705
on knowledge 10:136–137
on Marxism 10:136
on modernity 10:135–137
on modernization theories 15:705
on political convergence 10:137–138
on practical consciousness 10:136
on pure relationships 10:137
on radicalized modernity 15:705
on reflexivity 10:136–137
on risk 10:137
on structuration 10:135–138
on structure 10:135–136
on 'third-way' politics 10:135, 10:137–138
Giddens–Blair's Third Way 24:848–849
Giddings, F.H., works, *The Principles of Sociology: An Analysis of the Phenomena of Association and of Social Organization* 2:175
Gide, André 9:637–638
on Russian sexual life 21:740–741
works
Return from the USSR 21:740–741
Gide, Charles 23:2–3
works, *Quatre Ecoles d'Economie Sociale* 23:3
Giere, Ronald N., on epistemic cultures 7:877
Gierke, Otto von 13:802
Gift economy 8:476, 8:479
and exchange 8:476, 8:479
Gift exchange, in big man system 2:571
Gifted children 7:158
Gifted education 7:158
and talent development 7:158–164
acceleration and enrichment 7:160
creativity 7:158, 7:160–164
curriculum dimensions 7:159–160
developmental tasks 7:162
differentiation in 7:159–160
intelligence 7:158, 7:160, 7:162–163
intervention strategies 7:163
learners, characteristics of 7:159
nurturing personal and social development 7:162–163
programs 7:160–162
schoolwide enrichment model 7:161–162
social and emotional issues 7:162–163
talent search model 7:161
teachers 7:160
trends and future directions 7:163–164
Giftedness
definition 7:158
and intelligence 12:314
Gift-giving, in Melanesia 17:154–155
Gifts 7:13–14, 23:1
as distinction mark in power relationships 10:141
as ethnocentric and colonial scheme 10:142
as expenditure and scapegoats 10:140
globalization 10:142
honor 10:140–142
as impossible exchange 10:142
kula ring 13:140–142
marital 14:601
against market economy 10:141–142
nobility 10:140–142
obligations of 10:140
potlatch 18:714–715
prestige 10:140–142
rationality 10:140–141
and reciprocity 10:140
as symbolic market 10:140–141
as total social fact 10:139–140
Gigerenzer 8:393–394

GIGO ('Garbage In, Garbage Out') 22:840
Gilbert, Felix 12:256
Gilbert, G.N., works, *Social Studies of Science* 21:323
Gilbert, Margaret, joint action 12:839–843
Gilbert, Milton 16:231
Gilbert, Richard 16:231
Gilbert Islands 17:158, 17:160
Gilded Age 25:55
Gillan and Quinton v. United Kingdom (2010) 18:950
Gillen, F.J., works, *Central Tribes of Native Australia* 9:135
Gillette, Camp, works, *The Human Drift* 14:255–256
Gilliéron, Jules 6:348
Gilligan, Carol 12:936, 21:875, 23:327
  works, *In a Different Voice* 9:730
Gilman, Daniel Coit 24:762–763
Gilroy, Paul, works, *Black Atlantic* 19:796–797
Gilson, E. 18:435
Gimi people 18:353
Gin epidemic 6:670
GINA *see* Genetic Information Non-Discrimination Act (GINA)
g-index (scientometrics) 21:322
Gingivitis 17:281
Gingrich, A. 1:753
Gingrich, Newt 1:605–606
Gini, Corrado 11:885, 22:698–699
Gini coefficient/Gini index 3:98, 7:918, 11:163, 11:732–736, 11:740–741, 11:886, 11:888, 11:900, 11:902, 11:906, 15:596, 23:820, 11:736f, 11:886t
Ginsburg, F.D., on democracy in North America 17:25
Ginsburg, M. 7:378
Ginzburg, Lev R. 3:156
Gioberti, Vincenzo 4:572
Giolitti, Giovanni 14:423, 16:257
Gion matsuri festival 5:417
Giovanni Ruccelai, works, *Zibaldone* 20:427
GIP *see* Generalized inverse projection (GIP)
Gippius, Zinaida 21:741–742
Giraffi, Alessandro, works, *Le Rivolutioni di Napoli* 20:637
Girl children 15:610
Girls
  differential treatment of 9:744
  educational performance 9:747
  rites for 5:509
  seclusion 5:511–512
Girod's analysis 11:592
GIS *see* Genome Institute of Singapore (GIS); Geographic information system (GIS)
GIScience 15:88
GIST *see* Gastrointestinal stromal tumors (GIST)
Gist 20:50
Gist memory 25:178
Gitlin, Todd, on television 25:80–81
Gittler, Joseph, works, *Review of Sociology* 7:588–589
Giudizio abbreviato 18:229–230
Giuffre, K., on negative social network ties 5:561
Give-some dilemma 22:245
Giving 22:576, 22:578–580
  compassionate 3:365
  egoistic 3:364
  in global perspective 3:367–368
  philanthropy 3:360–361
  present-day paradigms of 3:364–365
  scientific 3:365
  in United States 3:366–367
Giving environment, hunting/gathering societies 11:430
Giving up, and optimism/pessimism 17:264–265
Giza Plateau Mapping Project 1:694
Glacial sediments 7:693t
Glacken, Clarence, works, *Traces on the Rhodian shore* 7:747–748
Gladstone, William 3:351–352
Glaeser–Vigdor measure 20:550
Glamour girls 25:814
Glandular hormones 16:649t
  functions of 16:649–651
Glaser, Barney G.
  works
    The Discovery of Grounded Theory 10:403–404, 23:530–531
    Theoretical Sensitivity 10:404
Glasgow Community Initiative to Reduce Violence (CIRV) 25:822
Glasgow Rent Strike 11:234–235
Glasnost reform (Soviet Union) 20:845, 25:838–839
  and environment 20:833–834
Glass, Ruth 10:46
Glass, usage in ancient Egypt 1:694–695
Glazer, Nathan 20:670–672
  works, *Beyond the Melting Pot* 2:81
GLC *see* Generalized Lorenz curve (GLC)
Gleik, J., works, *Faster* 24:401
Glendon, Mary Ann, works, *Rights Talk* 4:320–321
GLF *see* Genographic Legacy Fund (GLF)
Glia, and epilepsy 7:853–855
Glick effect 6:607
GLIM (software) 23:378
Glimpse decoder, computational auditory scene analysis 2:227
Glimpse integration stage 2:226–227
GLIMs *see* Generalized linear models (GLIMs)
Gliogenesis 20:57
Glittering generalities 25:397–399
GLM *see* General learning model (GLM); General linear model (GLM); Good lives model (GLM)
GLMMs *see* Generalized linear mixed models (GLMMs)
Global 30 12:794
Global Action Plan, Eco-teams project 17:537
Global Agenda for Social Work and Social Development (The Agenda) 10:144–149
  building capacity and impact 10:148
  community development 10:144–145, 10:148
  global context 10:144–145
  history 10:145–147
  impact on social work profession 10:147
  influence on global policies 10:147–148
  organizational context 10:145
  service user organizations 10:145
  United Nations 10:144, 10:146–148
  values 10:144, 10:148
Global Agenda Observatory 12:499, 24:750
Global agraphia 6:735
Global alexia 6:734
Global Alliance for Justice Education 18:739
Global ancestry-based methods 9:888–890, 9:890f
Global aphasia 1:820t
Global audiences, media 2:219–220
Global awareness 17:432
Global Barometer (GB) 5:827
Global burden of diseases (GBD) 1:153, 15:182, 15:189
Global capitalism 3:99–104, 3:609–610
  alternatives to 3:104
  crises 3:102–103
  as culture–ideology system 3:102
  and democracy 3:103–104
  and diffusion 6:404
  as economic system 3:99–101
  and financial crisis 3:101
  and global south 10:186
  versus international capitalism 3:99
  and political economy 18:362–363
  as political system 3:101–102
  resistance to 3:103–104
  transnational corporations 3:99–103
Global careers 3:133
Global citizenship 10:251–252
Global city 3:585, 3:664
  model 3:587–589
  for new politico-cultural alignments 3:590
Global city networks 24:945–946
Global civil society 3:678, 22:514
Global climate changes
  and forcing factors 3:865
  and water scarcity 25:445–446
Global commodity chains (GCCs) 4:243–244, 7:33, 12:485
Global communications 5:769
Global competencies 10:251–252
Global consumerism 10:240
Global contest between the US and USSR 4:128–134
Global coordination, and telephone 24:157–158
Global cultural flows 24:465
Global cultural supermarket 14:562
Global culture 10:213–215, 17:633
Global delivery model 3:45–46
Global Detention Project (GDP) 2:109

Global Development Network (GDN) 24:296
Global developments 10:150
early childhood care and education (ECCE) 6:739–741
entities and actors' subjectivities 10:151–152
and gender 9:782–783
knowledge in new global environment 10:153
methodological encounters and tribulations 10:150–151
new subjectivities 10:152–153
position of international agencies 10:153
and tourism 24:474–475
Global diffusion of the Western human sciences 25:541
Global dimensions, migration history 15:447–448
Global distinctiveness model 24:199–200
Global Earth Observation System of Systems (GEOSS) 23:130
Global economics
influences of repression 18:451
and social work 24:756
technology and 13:521–522
Global economy 3:585–587, 3:589–591, 7:266, 10:240–243, 10:264–265
and brokers 2:873
craft production trajectories in 5:110–112
in developing societies 5:110–111
in industrialized societies 5:111–112
impact on urban poverty 24:916
labor, division of 13:160
Global ecumene 12:603
Global education 4:431–432
global monitoring of educational outcomes 4:431–432
systemic research and modeling 4:432
Global Education Digest 5:767
Global Education Reform Movement (GERM) 10:249
Global enterprising 10:250
Global environmental change, in Africa 1:294–295
Global environmental movement 7:767
Global ethics, and morality 15:809
Global financial crisis (GFC) 2:243, 2:245, 25:552
causes of 9:194
regulatory response to 9:194–195
Global financial integration 9:178
Global financial markets 14:562
Global financialization 15:746
Global flows of finance 9:178
Global forces
and liberalism 14:12–13
of oppression 17:249
Global gay 21:789
Global geographic models 14:787
Global governance 10:155–161, 10:293–294, 18:314
in context of rapid globalization 10:155–156
cooperative action 10:155
definition 10:155
elements 10:156–158
international organizations 10:157–158
norms 10:157
standards 10:156–157
states 10:157
and gender 6:793
for health 10:202–203
building up social protection floor 10:203
moving health higher up foreign policy agenda 10:203
putting more health in trade 10:202–203
supporting civil society activism 10:203
informal networks, formal institutions, and incentives 10:156
issues 10:159–160
democratic values and fairness 10:160
effectiveness 10:159
legitimacy and authority 10:160
law 10:156
legislative action 10:158
methods 10:158–159
global policy issue networks 10:158–159
harmonization of laws 10:158
hybrid institutions 10:159
international regimes 10:158
nongovernmental organizations 10:156, 10:159–160
and policy networks 18:317–318
definition 18:317
global policy 18:317–318
society 10:156
transnational networks 10:155–156, 10:158–159
United Nations and 10:156
Global Governance by Indicators 10:293–294
Global health actors in collecting and sharing of health data 5:786–787
Global Health Observatory 5:794
Global hegemonies and local struggle 4:184–185
Global history 4:409, 7:53, 10:162–168
civilization 10:162–164, 10:166
classical world 10:162–163
comparative histories of gender and material progress 10:166–167
connections and comparisons 10:165–166
cultural history 10:162–167
economic history 10:167
of Europe 10:162–167
historiographical traditions 10:162–164
modernity 10:165–167
nationalism 10:163, 10:166
nations 10:163–166
nation-states 10:162–164
religious history 10:162–163, 10:165–167
revival, and globalization 10:164
states 10:163–164
traditional, postmodern, and global histories 10:165
Global homogenization of education 10:252
Global human mobility, tourism 24:465–466
Global hunger 9:285
Global imagineering 3:664
Global income, and Keynesianism 13:20
Global industrialization 11:854, 11:856, 11:858
Global Islam 10:216
Global issue management
global public policy 10:180, 10:182–183
issue ambassadors 10:182
Global justice
Rawls' contributions 19:917
struggle for 22:395–396
Global justice movement 21:180
and global south 10:186–187
Global land grabbing 17:581–582
Global Land Project 13:220–221
Global law 13:555–556
Global law firms 13:631
Global Leadership and Organizational Behavior Effectiveness Research *see* GLOBE (Global Leadership and Organizational Behavior Effectiveness Research) project
Global leadership assignments, models of 5:338–339
Global market
in developing societies, craft production for 5:110–111
journals in 12:865–866
Global memory markets 4:185–186
Global memory models 20:38–39
Global metamodel optimization 23:476–477
Global middle class 4:744–746, 4:749
Global migration 10:169–173, 15:427–429
borders 10:169
deportation 10:170–171
development 10:172–173
globalization 10:169
governance 10:169
guest worker programs 10:169–170
immigration 10:169–170, 10:172–173
migrant rights 10:171–172
Ravenstein on 10:169, 10:172–173
of students 10:251
temporary migration 10:169–170, 10:170t
territorial sovereignty 10:169
xenophobia 10:171
Global Migration Group 10:172
Global Mindset Inventory 5:434
Global minimizer 23:414
Global model of schooling 7:236–237
Global modernization 14:202, 15:705
Global navigation 16:346
Global organizational climate 17:323–325
Global outposts 10:251
Global participatory citizen journalism 12:860
Global patterns of alcohol consumption 1:517–518
Global perspectives
giving in 3:367–368
indigenous peoples in 11:756–757

of nonphysician health professionals 16:912
sport in 23:303–304
Global policy, global networks 18:317–318
Global political economy 10:258
and health 10:633–634
Global politics, and governmentality 10:308–312
Global positioning system (GPS) 3:619, 10:58, 10:174–177, 15:636, 16:347, 23:114, 24:418–419, 24:470
general functioning 10:175–176
and information systems 10:176
and its social implications 10:176–177
origins and instances 10:174–175
Global poverty
climate change and 3:848
and development 18:723–724
Global private sector 10:255
Global processing 2:866
Global production networks (GPNs) 7:33, 7:64, 10:244, 12:485, 12:488, 13:165, 20:571–572, 23:209–210
Global public goods 10:179b
provision of 10:180f
Global public policy 10:178–185, 18:318
averting risks for dual actor failure 10:179
evidence 10:181
challenges 10:184
facilitating getting to the good 10:180
evidence 10:182–183
functions 10:178
G-20 10:182–183
G-77 24:301–302
global issue management 10:180, 10:182–183
globalization 10:179, 10:184
international cooperation 10:178–179, 10:181–184
promoting global fairness 10:179–180
evidence 10:181–182
public goods and global public goods 10:179b
relevant initiatives 10:183
role of business 10:182–183
role of civil society 10:181–182
sovereignty-compatible rationale for 10:183–184
supporting global sustainability 10:180–181
evidence 10:183
Global recession 24:851–852
Global revolutions, and migration 11:648–649
Global Scenario Group (GSG) 23:804
great transition of 23:804
Global scene, health social work 10:718, 10:717t
Global self-esteem 1:54–55, 21:492–494
Global sexual citizenship 9:639–640
Global shift 3:103
Global Slavery Index (2013) 22:43
Global social justice 22:398–400
Global social survey 22:634
Global social work 12:558–559
Global society 10:240
Global South 5:385, 6:276–277, 6:290–294, 7:36, 10:50–51, 13:446, 18:180–185, 20:676–678, 24:847–848, 24:850
anthropological reflections 10:186–191
antimonies of 10:188–190
Third World 10:186–188
antiglobalization movement 10:186–187
center–periphery conceptualization 3:308–309, 3:311
development 10:186–190
global capitalism 10:186
global justice movement 10:186–187
globalization 10:186–189
left politics 10:186–189
modernity 10:189–190
neoliberalism
and global south 10:187–190
postcolonialism 10:186, 10:189–190
postdevelopment 10:190
poverty policies in 18:751–752
right to the city in 20:677–678
social justice 10:188–189
Global spread, of modernization 15:710–712
Global standardization
East *versus* West 10:250
for economic competitiveness 10:248
of educational policies 10:247
of educational policy and practices 10:248–249
for educational quality 10:248
spread of 10:249–250
Global Study of Sexual Attitudes and Behaviors (GSSAB) 10:808
Global supervenience theses 18:110
Global supply chain 10:251
Global sustainability 10:180–181, 10:392
Global systems theory 21:153
Global Terrorism Database (GTD) 5:837, 12:466
Global theory of intelligence 2:598
Global umbrella campaign, in advertising 12:459
Global uprisings, anarchism 1:685–686
Global Urban Observatory (GUO) 5:779
Global value chains (GVCs) 12:485, 13:163, 13:165, 20:571–572
framework 4:745–746
impact of 12:487–489
locational determinants of 12:486–487
Global village 12:603, 16:105, 24:814
Global violence
against women 10:192–197
beyond legal interventions 10:193–194
domestic and intimate partner violence 10:194
dowry deaths 10:193, 10:195
gender-based violence 10:195–196
international institutional response 10:192–193
rape 10:193–196
sexual violence 10:194–195
traditional and cultural practices 10:195
women's agency 10:196
Global visual processing, in occipital lobe 17:130–131
Global voyaging 16:346
Global War on Terror 6:304
Global warming 3:848–851, 3:854–855, 6:479–480, 7:793–794, 17:165, 20:808
Anthropocene 1:726–727
and greening of business 10:392–393
in Oceania 17:163
and sustainable transportation 23:846
Global WASH (water, sanitation, and health) Campaign 25:428
Global welfare concerns 13:522–523
Global workers 5:334
Global workforce misbalance 15:84
Global workspace theory 4:676
Global world 10:242–243
cities 10:243–244
regions 10:243–244
GlobalEd 25:640
Globalism 10:263
Globalization 2:23–24, 3:158, 3:583, 3:719, 4:409–410, 4:704, 4:754–755, 4:944, 5:474–475, 6:197, 6:279, 6:291, 7:64–65, 7:927, 8:60, 8:285, 8:877, 10:206–207, 10:211, 10:239, 10:263–266, 10:867, 11:119, 11:688, 12:363, 12:372, 12:482, 13:472, 13:515, 13:519, 13:803, 14:54, 14:279, 14:677, 14:723, 14:736, 15:711, 16:230, 16:254–255, 16:458, 16:460, 17:633, 17:652, 19:607–608, 20:261, 20:411, 20:461, 20:801–802, 22:147, 22:149–152, 22:154, 22:890, 24:139, 24:794, 24:814–815, 24:838, 24:840, 24:842
and advertising 1:214–215, 1:217
agrarian political economy 1:459–460
and anarchism 1:684–685
of Anglo-American model of law firms 13:579–580
anthropology of 1:776, 10:233–238
circulation 10:233–235
commodification 10:235
cultural flows 10:233–235
disciplinary context 10:233
ethnic violence 10:235–236
future 10:236–237
heterogeneity or homogeneity 10:235
identification 10:233–235
locality 10:234–235
McDonaldization 10:235
minorities 10:236
sovereignty, citizenship, and violence 10:235–236
transnationalism 10:235–236
Beck on 2:396–399
Benhabib on 2:519, 2:522–524
and body 2:730–731
and broadcasting regulation 2:861–862
and capitalism 10:218–220

Globalization (*continued*)
in Central America 1:621, 1:623
challenges 10:254
characterization and meanings of 10:262–263
childhood study 3:456–457
and citizenship 3:652–653
and class inequality 10:220
collective memory 4:183–184, 4:186
and collectivism 4:209–210
and community 4:388
and community development 4:383
and community sociology 4:366–367
and community studies 4:370–371
comparative research in education 4:428
concept of 25:659, 25:661
of consumption 4:744–746
contemporary 10:198–201
accepting environmental limits to growth 10:203–204
austerity agenda 10:201
building up social protection floor 10:203
deepening neoliberalism 10:201–202
global governance 10:202–203
Great Financial Crisis 10:201
inequality 10:198, 10:201–202
loss of government policy space and capacity 10:200
moving health higher up foreign policy agenda 10:203
negative health 10:199–201
positive health 10:199
putting more health in trade 10:202–203
social determinants 10:199, 10:204
spread of disease 10:199–200
structural adjustment 10:198–203
supporting civil society activism 10:203
taxation 10:200–203
trade liberalization 10:198–201
turmoil in labor markets 10:200–201
and creolization 5:120
and criminology 5:244
and critical theory 5:300
critique of, and nongovernmental organizations 16:857–858
cultural critique and 5:385–386
and cultural studies 1:936–937
culture 10:262–263
and culture 5:586–587
culture and economy in 5:539
and debates 10:218–219
in the late twentieth century 10:219–220
definitions 10:218–219
denationalizing processes 10:256–259
depictions of 13:165
and development 6:303–305
and development geography 6:315–317
and diffusion 6:404
and discourse 6:507
and division of labor 13:160
domestic legal systems and 10:258–259
and drinking 6:658, 6:660, 6:662
East Asian studies 6:776, 6:795–796
economic sociology 7:69–70
and economies 10:206, 10:210–211
and ecumene 7:110–111
in education 10:247–253
of entertainment 7:673
ethics of 3:18
and ethnic conflict 8:128
in Europe 8:278
family violence 12:542
and feminism 8:899–903
finance 9:178, 9:180
of food system 17:82
food/cuisines 9:292–293
and gay/lesbian activism in Latin America 13:861–862
of gender identity 21:768
General Agreement of Tariffs and Trade 10:255
of gentrification 10:50–51
geographical approach to 10:239–242
geographical aspects of 10:239–246
and gift 10:142
global economy 10:264–265
and global history revival 10:164
global public policy 10:179, 10:184
and global south 10:186–189
glocalization 10:263–265
and governmentality 10:308–311
and health 10:574–575
in higher education 10:854–855
history of 25:279
and human development 6:307–311
and human trafficking 10:210–211
combating 10:211–212
and hybridity 11:450
ICT and 10:262–263, 10:266
impact of 5:628, 25:512
imperialism 16:460
and income inequality 11:741
and indigenous knowledge 11:780–783
and international organizations 12:538–539
international social work 10:264–266
and Internet 12:603
IP law 12:268–269
issues 16:461
Japan and 12:795–797
on job market 22:153
and labor history 13:172
of law 13:600–601, 13:608
legal aspects 10:254–261
and legal pluralism 13:539–540
and legal profession 13:631
Mann on 14:493–495
Manuel Castells 3:229–231
marketization of social services 10:258–259
and McDonaldization 14:828–829
and media imperialism 15:58
and modernity 15:690, 15:693–694
Montreal Protocol 10:255
and multiculturalism 16:31
multilateral agreements on trade 10:255
multilateral treaties and global environment 10:256–257
and multinational corporations 16:79–80
in multisited field studies 16:105–106
Mumford on 16:139
and music 16:151–152
and nation 16:225
national states and 10:254
and nationalism in East Asia 16:277
and networks 16:567–571
and nostalgia 17:34–35, 17:37–38
of opinion research 19:566–567
and oppression 17:248–249
organizations 17:310
origins of the term 10:219–220
parliamentary democracy in 17:514
place 10:242–244
of political science 18:470–471
and postcolonial geography 18:645–646
and postcolonial law 18:649
postcolonial science studies 18:652, 18:654–655
and postcoloniality 18:658–661
and postindustrial societies 18:668
postsocialist cities 3:608
postwar era 25:368
and poverty 7:119, 10:206–207
and poverty law 18:738–739
of production 25:701
and prosocial behavior 19:240
and public interest law 19:559–560
and publishing 19:622–623
and queer anthropology 19:756–757
rapid, governance in context of 10:155–156
and regulatory agencies 20:185
regulatory role nation-states 10:255
and religion 10:213–214, 20:311
functional conception 10:213–214
fundamentalism 10:214
global culture 10:213–215
glocalization 10:213
migration 10:216
new religious movements 10:214
postmodernity 10:215–216
religious authority, transnational organization, and diaspora 10:216
religious pluralism 10:215
theoretical perspectives 10:214
and religious identity 24:605
and religious nationalism 20:401–405
research, race/racism 19:799
research infrastructures 22:634
and revolutions 20:625–626, 20:648
and rural social work 20:808
of services 3:43–45
financial 3:45
knowledge management services 3:45
legal process outsourcing 3:45
voice-based 3:45
and sexual orientation 21:789
of Slow Food movement 22:97–98
and social justice 22:398
social space compression 10:218
social time acceleration 10:218
of societal modernity 10:214
and socioeconomic inequality 10:218, 10:222–223
in South Asia 16:282

in Southeast Asia 23:73
and South/Southeast Asian migration 12:530, 12:534, 12:536
space 10:242–244
and sport 23:313–316
of sport 23:304
and state formation 23:353–355
supply chains 23:712
and sustainable transportation 23:846
and syncretism 23:883
of television 25:83–84
and Third World 24:302
of tourism 24:472
trade policy 24:487, 24:489
and transition to adulthood 24:576–577, 24:579–581
and transnational corporations 10:221–223
transnational law (TL) 10:255–256
and unequal distribution of wealth 12:542–543
universities 18:308
and university press 19:628
and urban policy 24:802–803
urban sociology 24:928–929
and urbanization 6:285–287
and urbanization in Africa 24:966–967
of venture capital 25:61
and welfare 25:479
welfare measures 10:259
welfare state 25:522
in West Africa 1:301–302
and world culture 10:225
and world republic 25:740–741
world systems theory 25:749–751
and youth culture 25:808, 25:810–811
Globalized capitalist system 3:97–98
Globalized corporate spaces 10:242–243
Globalized gay consumer 9:639–640
Globalized *laissez-faire* 3:97–98
Globalizing identities 12:92
Global–local processing 2:865
GLOBE (Global Leadership and Organizational Behavior Effectiveness Research) project 17:824
Globus pallidus 23:252
Glocal 16:441
Glocalization 5:539, 10:213, 10:263–265, 17:37–38
and advertising 1:214–215
and world culture 10:228, 10:230
Glock, C.Y., works, *The New Religious Consciousness* 2:509
Gloeden, Wilhelm von 9:637–638
Glorious Revolution 1:611, 3:61, 7:650, 11:367, 17:516
Glory, and war 4:608, 4:610
Glottochronology 8:370–371, 11:23–26
GLOW *see* Gay and Lesbian Organisation of the Witwatersrand (GLOW)
GLS *see* Generalized least squares (GLS) method
GLSEN *see* Gay, Lesbian and Straight Education Network (GLSEN)
GluA1 16:489–491
GluA4 16:489–491
Glucagon 16:651
Gluckman, Max 1:752, 1:768–769, 6:245
on educational anthropology 7:145
Glucocorticoid hormones 7:606
Glucocorticoid receptors (GR) 11:477–480, 23:577–578
epigenetic regulation of 16:498
Glucocorticoid response elements (GREs) 22:485–486
Glucocorticoids 16:490–491
social isolation-induced disruption of AMPA synaptic delivery 16:490–491
GluR2 12:711
Glutamate (Glut) 16:750–751
transmission, and memory formation 15:158–159
Glutamic acid decarboxylase gene 22:495
Glycemic control, and diabetes 6:337–341
GM debacle 16:191
GMA *see* Growth Management Act (GMA)
GMD *see* Guomingdang (GMD)
GMFs *see* Genetically modified foods (GMFs)
GMM *see* Gaussian mixture model (GMM); Generalized method of moments (GMM) estimation; Groundwork of metaphysics of morals (GMM)
GMOs *see* Genetically modified organisms (GMOs)
GMSMA *see* Gay Male S/M Activists (GMSMA)
Gnaritas Nullius, intellectual property rights and 11:772–773
GNAT *see* Go/no-go association task (GNAT)
Gnostic Christians 3:548
Gnostic writings 3:548
Gnosticism 16:175, 18:437
GNP *see* Gross national product (GNP)
GnRH *see* Gonadotropin releasing-hormone (GnRH)
Go West policy 24:974–975
Goal adjustment in transition to adulthood 4:816–817
Goal attainment, in AGIL schema 1:381
Goal attainment scaling (GAS) 8:347
Goal directedness 12:342
Goal displacement 17:202–204
in bureaucracy 2:910
Goal engagement
and life span development 14:102, 14:104–106
in transition to adulthood 4:816–817
Goal explication, for user orientation in program evaluation 19:121
Goal focus 21:561
Goal frame, groups 10:437
Goal framing theory 1:78
prosocial behavior 1:581
Goal obstruction 7:483
Goal orientations 21:51, 25:684
and epistemological beliefs 7:889–890
self-regulated learning 21:542–543
Goal pursuit 10:713–714
Goal revision 25:683
Goal setting 25:682–683
FTP results from 9:561–562
and motivation 15:937
Goal striving 25:683
and motivation 15:937–938
Goal-corrected partnership (GCP) 12:231
Goal-directed actions 10:268, 10:271–272
parents' values and interests 8:734–735
Goal-directed behavior 15:946
Goal-oriented behavior
among human and nonhuman primates 10:268–274
communicative cues 10:271–272
methodological considerations 10:269–270
responses 10:271
theoretical orientations 10:268–269
visual behavior 10:270–271
definition 10:268
gaze-following 10:270, 10:272
goal-directed actions 10:268, 10:271–272
infants 10:268–272
intentions 10:268
joint attention 10:272
sociocognitive development 10:268–270, 10:272
visual habituation 10:269–270
Goal-oriented tableaux-based deductive systems 5:935
Goal-related operations 16:504
Goals 1:242–245, 10:710, 12:381, 15:928, 21:51, 21:490, 25:680–681
academic achievement 1:21f
development 1:23
concepts of 15:888
cooperation 4:865
disengagement 21:557
governing socialization and education 22:865
and hope 11:197–200
imitation by infants/young children on 11:622–623
of industrial ecology 11:846–849
for language policy 13:355–358
and motivation 15:937–938
pursuit 10:713–714, 21:557
pursuits
determinants of effective 15:888–889
hindrances 15:891–892
instigating determined 15:889–891
self-regulation of 15:889–892
and rationality 19:908–909
reengagement 21:557–558
and self-efficacy 21:504–507
self-regulation
impact of time perspective on 21:560
orientation 21:559, 21:561
setting 21:559–560
setting multiple goals 21:560–561
setting 10:712–713, 21:557
social, diversity during adolescence 22:220
and social control 22:238
and successful aging 23:670–671
of sustainability 23:803–804

Goals, Operators, Methods, and Selection *see* GOMS (Goals, Operators, Methods, and Selection) model
Goals diversification 5:463
Goals/alternatives matrix 18:277
Goal-setting theory 10:712–713, 25:681
Goats, and food production 9:300–301
Go-betweens, and brokers 2:870
GOBI (Growth monitoring, Oral rehydration, Breastfeeding, Immunization) 10:199
God image 20:351
Godard, Jean-Luc, works, *Breathless* 9:169
Godcasting 20:276
Goddard, Henry Herbert 3:879
Goddess, and motherhood 15:881
Gödel, Kurt 1:530
Godelier, Maurice 13:71
Godly prince 8:259
Godmilow, Jill, works, *Far from Poland* 6:623
God's acting in history 11:32
God's revelation 11:32
Godwin, William 24:980
Goebbels, Joseph 4:151–152, 16:258, 24:448
Goebbel's Lips (radio set) 19:878
Goethe, Johann Wolfgang von 17:426, 19:410–411, 21:178
  works, *Elective Affinities* 3:371–372
Goethe Institut 5:463
Goffman, Erving 2:401, 2:405, 5:525, 6:21–22, 6:505, 10:275–278, 11:550, 14:383, 17:532, 23:437
  biography 10:275
  and definition of situation 6:21–22
  on dramaturgy 10:276
  and Durkheim 10:276
  on frames and framing subjects 10:277
  importance for social sciences today 10:278
  and James 10:276
  methods 10:277–278
  on morals 10:276–277
  on narratives 16:218, 16:220
  on organizations 17:391
  on research program and intellectual context 10:275–276
  on ritualizations 10:276–277
  roles and identities 10:276
  on shame 21:873–874
  and Simmel 10:276
  on social identity 10:276
  on social interaction 10:275–278
  on social psychology 10:275
  on social roles 22:335
  on stigma and spoiled interaction 23:463
  on strategic interactions 10:277
  on symbolic interactionism 12:355
  on total institutions 10:276
  works 10:275–278
    Asylums 13:152, 18:920
    Forms of Walk 10:277
    Frame Analysis: An Essay on the Organization of Experience 6:21–22
    The Presentation of Self in Everyday Life 22:422
    Stigma 13:152
Gogol, Nikolai 21:741
Going home 20:107
Gold, Judith 25:625
Gold and Silver Commission of 1888–1889 14:621
Gold Open Access 12:866
Gold standard for intergenerational relationships assessment 12:403
Gold standard randomized controlled trials *see* Ideal randomized controlled trials
Gold standard research 5:819, 20:241
Goldberg, Lewis R., on five-factor model of personality 9:230–231
Goldberg v. Kelly (1970) 18:735
Golden Age (mythology) 14:251–252
Golden Age of Television 24:169
Golden Book of Venice 1:968
Golden Law 1888 (Brazil) 22:41
Golden rice 9:972–974
Golden Rule (bioethics) 2:613–614
Golden Rule (Christianity) 14:10
Goldenson, Robert, works, *Encyclopedia of Human Behavior: Psychology, Psychiatry, and Mental Health* 7:591–592
Goldenweiser, Alexander 2:515
Goldfrank, Esther 14:837
Goldman, Emma 19:748
  on queer anarchism 19:748
Goldman-Fristoe Test of Articulation 2 3:796
Goldschmidt, W., on capitalism in North America 17:24
Goldsmith, H.H., temperament development approach 24:193
Goldstein, Kurt 16:685
  neuropsychological rehabilitation 2:814–815
Goldstone, Jack 14:491
Goldthorpe, John H. 10:279–285
  biography of 10:279
  on class analysis 10:280–283
  on class scheme 3:747, 3:747t
  contribution to sociology 10:283–284
  on cultural capital 5:451
  on cultural consumption 10:283
  on educational inequalities 10:282–283
  on methodological issues 10:280
  on social mobility 10:281–282
  on social stratification 10:280
  on status analysis 10:280–281
  theoretical orientation of 10:279–280
    macrosociological perspective 10:279
    social action 10:279–280
  works
    Affluent Worker Study 10:279
    The British Journal of Sociology 10:280
    The Constant Flux 10:279
    European Sociological Review 10:279
    John H. Goldthorpe: Consensus and Controversy 10:283
    On Sociology 10:283
    Social Mobility and Class Structure in Modern Britain 10:279
Goldwater, Barry 1:605
Gomer, Nina 6:690
Gompertz double-exponential function 15:161
Gompertz–Makeham model of mortality 6:124, 6:159, 6:756–757, 14:56, 14:91, 15:864, 17:174, 17:727
GOMS (Goals, Operators, Methods, and Selection) model 11:400–402
Gonadal hormones 7:606
  with depression 7:607–608
  in schizophrenia 7:609–610
Gonadotropin releasing-hormone (GnRH) 10:462–464, 10:467
Gonads 11:484–486
Go/no-go association task (GNAT) 11:715
  and dysexecutive syndromes 6:726
Gonzo 17:218
Good, and public interest 19:552–553
Good city 3:665, 18:202–203
Good Clinical Practice (GCP) 8:78
Good companions 25:814
Good death 1:688, 5:866, 8:280, 19:310
  versus bad death 5:872–873
Good enough parenting 3:412
Good faith, corporate governance 4:936–937
Good Friday Agreement 4:693, 6:225
Good genes 11:343–344
Good governance 5:12, 10:293–294, 19:690–691
  as absence of corruption 19:691
  agenda 18:181–182
  definition 19:690
  and democracy 19:692
  equality 19:693
  ethical issues and 2:605–606
  objectivity 19:693
  procedural definitions 19:691
  as rule of law 19:691–692
  universalism 19:693
  Western vision of 5:16
Good government 18:203–205
Good health 1:280
Good judge, the 17:753
Good life
  concepts of 20:196–197
  and eudaemonism 8:212–213, 8:215–216
  notions of 10:521
    life-ability of person 10:521–522
    livability of environment 10:521
    subjective enjoyment of life 10:522
    utility of life 10:522
  philosophy of 10:521–523
Good Life, Good Death, Good Grief program 17:927
Good lives model (GLM) 20:437, 21:675
Good planning principles 16:810
Good Society 4:309
Good weather 6:229–230
Goode, Erich 15:798
Goodglass, Harold 16:685

Goodin, Robert E., works, *A New Handbook of Political Science* 7:591
Goodman, Bernice 15:114–115
Goodman, Paul 19:748–749
Goodman, Robert 24:865
Goodness of fit framework 13:691–692, 15:653–655, 16:132, 17:878, 21:947
  Bayesian model checks and criticisms 15:654–655
  for continuous random variables 15:654
  criterion 4:475–477
  for discrete random variables 15:653–654
  statistical distribution models 23:397–404
  for temperament 24:191, 24:196
  testing against specified alternatives 15:654
  using Pearson's chisquare 21:948
Goodness-of-fit ethics (GFE) 20:517
Goodness-of-fit tests 19:725–726
Goodnow, J.J. 4:104
Goods 4:317–318
  capital 3:91
    heterogeneous 3:92
  consumer psychology 4:738–739
  exchange, market allocation 14:542
  social 4:319
  types, and bargaining 16:415–416
Goodwin v. United Kingdom (1996) 9:396
Goody, Jack 9:672
  works, *The Domestication of Savage Mind* 1:746
Google 23:311
  Internet privacy 12:593–594, 12:596–597
Google Art Project 14:567–568
Google Books 7:586
Google Cultural Institute 14:567–568
Google Earth 9:229, 10:53–54, 10:58, 10:97–98
Google Maps 3:162, 10:97–98
Google Ngram 19:266f
Google Now 20:687
Google Scholar 21:322
Goolagong, Evonne 2:237
Gopinath, Suhas 10:247
Gorbachev, Mikhail 20:823–824, 20:845, 25:838–839
  New Political Thinking of 20:845
  reforms 20:845
    demokratizatsiia (democratization) 20:845–846
    glasnost (openness) 20:845
    perestroika (restructuring) 20:845
Gordley, James 13:777–778
Gordon, Thomas 3:773
Gore, Tim, on epistemic cultures 7:878
Gorer, Bateson 5:99
Gorer, Geoffrey 5:99
Göring, Hermann 16:259
Gorz, André 6:24–25, 6:27, 10:286–291
  on alienation 10:288
  on capitalism 10:287–289
  on existentialist philosophy 10:286–288, 10:290
  on immaterial economy 10:288
  influence and impact of 10:289–290
  life of 10:286–287
  on Marxism 10:287–290
  on political ecology 10:287–290
  Sartre and 10:286–289
  works 10:287–289
Goskompriroda (State Committee on Nature Protection) 20:834
The Gospel According to St. Luke 11:31
Gospel tradition 3:547
  Gospel of Luke 3:545–546, 3:548
  Gospel of Mark 3:546
  Gospel of Matthew 3:545–546
  Gospel of the Egyptians 3:237
  Gospel of Thomas 3:237, 3:545–546, 3:548
Gosset, William Sealy 11:495–496, 23:425
Gosudarstwo i rewoljucija 4:304
Gothic architecture 14:243
Gottdiener, Mark 4:364
Gottfredson, Michael 20:774
  general theory of crime 21:481–482
  self-control theory 21:481–485
  works
    Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization 20:774
Gottmann, J., works, *The Significance of Territory* 18:375
Gottsched, Johann Christoph 15:11–12
GOTV *see* Get out the vote (GOTV)
Gough's social contract 5:474
Gould, W.T.S. 7:166
Gouldner, Alvin W.
  on functionalism 9:540
  works
    The Coming Crisis of Western Sociology 9:540
    Patterns of Industrial Bureaucracy 2:914
Gould's analysis 7:538
Governance 4:707–708, 6:829–831, 6:836, 10:292–300, 12:218–219, 13:440, 13:760, 14:430, 14:472, 18:181–182, 20:162, 23:727–728, 23:730, 24:294, 24:296–298, 24:862, 24:864–866
  and administrative law 1:113
  biotechnology 2:676, 2:679–680
  character, Mancur Lloyd Olson, Jr. 17:209–210
  collaborative 10:296
  commodity chains 4:243–244
  concept, strength and weaknesses of 10:298–299
    as analytical tool for institutional architecture of modern statehood 10:298–299
    as interdisciplinary bridging concept 10:298
    as mirror of changes in statehood 10:298
  of consumer policy in advanced economics 4:734–735
  contract 10:295
  as coordination 10:293
  corporate 4:932–938
  and decline of mafias 14:434
  executive government 8:498–500
  financial 9:192–193
  global 10:155–161
  government to 10:292
  and governmentality 10:308, 10:310–311
  health data 5:781
  health policy 10:664
  hierarchies and markets 10:841–844
  hospitals 11:210–211
  multiple sites of 13:607–608
  network governance 10:298–299
  new genetics research/clinical practices 9:1017–1018
  as nonstate-centered concept 10:292
  organizational 17:397–399, 17:402
  by partnerships 10:293–294, 10:296, 10:294f
  Plato on 18:212
  and police 18:251–252
  public policy ethics 19:580, 19:582–583
  public sector organizations 19:600–601
  public–private partnerships 10:293–294, 10:296–298, 10:294f, 10:298f
  regional planning 20:129–130
  as rule making 10:293, 10:295–296
  of rules 13:486
  social work in Russia 22:810–812
  sustainability lifestyles 23:828–830
  theoretical considerations 10:292–293
  traditional *versus* modern 10:293, 10:293t
  transnational 10:295
  and urban policy 24:904
  of voluntary agencies 25:264–265
  water resource 25:446–447
  in world of politics and policies 10:293–297
    good governance and global governance by indicators 10:293–294
    governance at work 10:295–297
    governance by and as communication 10:294–295
Governance capacity building 18:205
  models 2:605
Governance of climate change 3:839, 3:842–844
  international climate negotiations 3:842–843
  at national level 3:843
  transnational 3:843–844
Governance regimes
  definition 5:14
  and their main features 5:15t
  typology of 5:14–16
Governance structures
  and markets 14:560
Governance system of a jurisdiction 7:267
Governing board, stockholders 23:492–493
Government Accountability Office (GAO) 1:232–233
Government action, linking manifestos to 18:419
Government by age 1:323
Government decision making 2:902
Government devolution, and urban policy 24:803
Government failure 18:274–276

Government funding, in applied research (criminology and criminal justice) 5:128
Government grants, on philanthropy 3:361–362
Government mandate theory 20:868
Government Performance and Results Act 1993 (GPRA) 23:516
Government Performance and Results Modernization Act 2010 23:516
Government policies
  for Australian aboriginals 2:235, 2:237–238
  space and capacity, loss of 10:200
Government regulations 13:505
  stock market 23:493
  for workplace safety 20:868
Government revenue 4:946
Government stability 8:500
Government statistics 10:301–307
  centralized *versus* decentralized systems 10:304–305
  collection and preparation of report 10:303–304
  national statistics, production of 10:301–302
  organization of 10:304–305
  principles of 10:302–303
  use of term 10:301
Government surplus 5:50–51
Governmental institutions, and public bureaucracies 19:531–532
Governmental resourcing in America 22:656–658
  Centers for Disease Control and Prevention (CDC) 22:658
  National Institutes of Health (NIH) 22:657–658
  National Science Foundation (NSF) 22:656–657
  Social Sciences and Humanities Research Council (SSHRC) 22:656
Governmentalities 1:736, 5:256–257, 10:308–312, 16:463–464, 21:396
  anthropology 10:308–311
  biopolitics 10:309–311
  globalization 10:308–311
  governance 10:308, 10:310–311
  institution 10:308–311
  and migration 15:424
  organization 10:310–311
  politics 10:308–312
  power 10:308–312
  and power 18:767
  and risk 20:687–688
  sovereignty 10:308–312
  state 10:308–312
  transnational 10:308–312
  and unemployment 24:735–736
Government-and-Binding framework 16:5–6
Government-backed insurance 13:505
Government-issued flood insurance 13:505
Government-linked research centers 22:619–620
Government-paid lawyers 12:920–921
  access to justice without supplying 12:923–926
Government-regulated public service broadcasting institutions 2:859–860
Governments 10:313–315, 22:604
  and area and international studies 1:943–947
  authoritarian 10:313–314
  branches that design and implement crime policy 5:180
  cabinet 10:314, 23:364
  and civil service 3:695–699
  and constitutions 4:710–714
  control of associations 25:255–256
  control over cultural industries 5:597–598
  definition 10:313
  duration of 10:314
    career patterns 10:314–315
  environmental justice in United States 7:757–758
  executive government 8:497
  forms of 6:80–82
  to governance 10:292–293, 10:293t
  green parties 10:383–384
  impact of 10:315
  instruments, implementation 11:693–694
  by law 20:781
  under law 20:781
  leadership role, in social work development 22:791
  liberal-democratic 10:313–314
  libraries 14:45–46, 14:48–49
  machinery of 19:521, 19:523–524
  media relations 3:178
  organization of 10:314
  and organized crime 5:171
  parliamentary *see* Parliamentary democracy
  presidential systems 10:314
  programs for fertility transition in China 9:47–48
  public bureaucracies 19:531
  quality of 19:690–694
  registration of associations with 25:254
  relations of Australian aboriginals with national and local 2:237
  republican form of 20:477–478
  revolutions, theories of 20:627–628
  and social capital 22:178
  social impact litigation against 3:735–736
  state/society 23:344
  and telephone 24:159
  tools, public policy 19:572–573
  trust in 24:658–659
    caring about 24:662
    corruption 24:660–661
    determinants 24:659–660
    incivility and gridlock 24:661
    measurement 24:659–662
    performance of government 24:660
    satisfaction with direction of country and economy 24:659–660
    social trust 24:661–662
    winners and losers 24:661
  types 10:313–314
Government-sponsored research 13:561–562
  cold war experiments 13:562
  human plutonium experiments 13:562
  U.S. sponsored syphilis studies in Tuskegee and Guatemala 13:561–562
GP *see* Genetic programming (GP); Grapheme-to-phoneme (GP) conversions; Growth points (GP)
GPA *see* Grade point average (GPA)
GPIDs *see* Genographic Project IDs (GPIDs)
GPNs *see* Global production networks (GPNs)
GPOW *see* Geographic Perspectives on Women (GPOW) Specialty Group
GPR *see* Ground-penetrating radar (GPR)
GPRA *see* Government Performance and Results Act 1993 (GPRA)
GPS *see* Geographical position systems (GPS); Global positioning system (GPS)
GPSG *see* Generalized Phrase Structure Grammar (GPSG)
GPT *see* General processing tree (GPT) models
Gqola, Pumla Dineo, on rainbow nation 21:747
GR *see* Glucocorticoid receptors (GR)
Graburn, N.H.H., works, *Ethnic and Tourist Arts* 5:110
GRACE *see* Gravity Recovery and Climate Experiment (GRACE)
Gradational regime
  social classes 3:756f
Grade level textbooks 24:257
Grade point average (GPA) 4:499, 17:787
GRADE (Grading of Recommendations Assessment, Development and Evaluation) system 8:330–331
Graded preference data, in forced-choice personality assessment 17:841
Graded response model (GRM) 8:700
Grades/grading (marking) 3:823–824, 17:787
  for educational gender gaps 9:682–683
  and educational outcomes 5:377
Gradient approximation, in stochastic optimization 23:475
Gradient-based seriation 19:432
Gradients, human–environment relationships 11:407
Grading of Recommendations Assessment, Development and Evaluation *see* GRADE (Grading of Recommendations Assessment, Development and Evaluation) system
Gradual Abolition of Slavery 22:443
Gradual retirement, older age work participation 17:187
Gradualism, and economic reform 18:692
Graduate education 24:762–763
Graduate School of Industrial Administration (GSIA) 21:971–973

Graduates, and equity in education 7:919, 7:921–922
Graeber, Daniel
on anarchism 1:683–686
and fetishism 9:108–109
works
Debt: The First 5000 Years 7:19
Graeco–Roman antiquity 18:795
Graff, Gerald, works, *Beyond the Culture Wars* 1:620
Grafik symbol system 13:58–59
Graham, B., works, *Modern Historical Geographies* 11:20
Grain size theory, and visual word recognition 25:215
Grains, geoarchaeology 10:55
Grameen Phone 15:632
Gramicidin 6:693–694
Grammars 2:651–652, 12:784, 13:272–273, 13:289–290, 13:942, 14:208, 14:210, 19:9–11
conceptualization of 13:272–273
context-free 19:12–13
of English 17:553
fatalism in Russian grammar 13:289–290
functional approaches to 9:470–475
autonomous functionalism 9:472
cognitive explanations 9:470–471
cognitive linguistics 9:473
competing motivations 9:474–475
directions and prospects 9:474–475
discourse explanations 9:471
dynamic approach 9:473–474
economy 9:474
exemplars 9:474
functionalism/formalism 9:471–472
iconicity 9:473–474
language change 9:473–474
mixed functionalism/formalism 9:472
optimality theory 9:474–475
processing explanations 9:470
prototypes 9:473–475
schools of analysis 9:471–474
sociolinguistics 9:474
typological functionalism 9:472–473
usage-based model 9:474
logic of 14:296–297
probabilistic 19:9–15
regular 19:11–12
SES differences in 13:325
theories on 21:625
universal 13:273
Grammatical agreement 10:316–323
constraints on agreement 10:321–322
long-distance agreement 10:319–321
theoretical accounts 10:317–319
Grammatical analyses, of serial verb constructions 21:646
Grammatical category
tense, aspect, mood, and evidentiality 24:210
Grammatical competence 4:296
Grammatical encoding 23:256–257
Grammatical incorporation 14:222–223
Grammatical nominalizations 23:891–892
Grammatical relations 10:324–329, 14:209–211, 23:889–891, 23:904, 25:1
direct object 10:324
ergative predicate 10:327
indirect object 10:324
null subject language 10:324
obligatory subject language 10:324
predicate 10:324
subject 10:324
thematic role 10:324
transitive predicate 10:326
unaccusative predicate 10:327
unergative predicate 10:326
Grammatical selection and description 13:942
Grammatical variables, and social class 13:283, 13:283t
Grammaticalization 3:816–817, 8:354, 10:330–335, 11:514, 16:412, 3:817t
aspects 10:330t
bridging texts 10:332
definition 10:330–331
degrammaticalization 10:333–334, 10:334t
emergent grammar 10:333
grammatical categories 10:330, 10:332
iconism 10:334
lexicalization 10:333–334
parameters 10:330–331
phonogenesis 10:334
phonological erosion 10:330
pragmatics 10:332–333
process
continuum of 10:332–333
dynamic of 10:333
unidirectionality 10:333–335
reanalysis 10:334–335
research 10:331–332
semantic bleaching 10:330, 10:332–333
symbolism 10:334
transcategorization 10:332–334
Gramsci, Antonio 3:710, 3:741, 4:380–381, 5:499, 11:41, 12:279, 13:730, 14:647, 18:473, 18:766, 22:697, 23:309
concept of hegemony 10:744, 10:746
on cultural hegemony 11:41
and industrial proletariat 11:41
on subaltern classes 11:41
writings in Mussolini's prisons 11:41
Gran, Peter, works, *Islamic Roots of Capitalism: Egypt* 16:382
Granadine Confederation 4:574
Grand Alliance 16:255
Grand Ayatollah Sayed Fadlullah 24:232
Grand coalition, and consociationalism 4:693–694
Grand narratives 16:196, 18:686
Grand strategy, appeasement 1:829
Grand synthesis 9:943
Grand theory 22:997
and institutionalism 12:182
of motivation 15:894
Grand-coalition-worth monotonicity 4:874
Grande encyclopedie 16:283
Grandes écoles 7:393–394
Grandparental age at birth, influences in adulthood 6:755–756
Grandparent–grandchild relationships 10:336–340
aging 10:336
changes over individual time 10:338
child care 10:336–339
historical context 10:336–337
importance 10:337–338
intergenerational relationships 10:337–339
intergenerational solidarity 10:337–338
third age 10:336
variation between countries 10:338–339
Grandparenting 12:405, 12:407
Granet, Marcel 10:139
Granger, Clive 7:9–10
Granger causality 9:529
Granger Representation Theorem 24:325
Granger–Schweder framework 8:301–303
Granjon, Fabien, on social networks 16:557
Granovetter, Mark, on social network analysis 16:529–530
Grant, J. 9:571
Grant, Madison 8:224
works
The Conquest of a Continent 19:803
The Passing of the Great Race 19:803
Grants 3:503
from China 3:502–503
Granularity problem 25:215
Graph representations, social network algorithms 22:454–455
bipartite graphs 22:455
directed graph 22:454
hypergraphs 22:454
multigraphs 22:454
undirected graph 22:454
Graph theory
network analysis 16:518–519, 16:521
and social networks 16:528–529
structural equivalence 23:621
Grapheme to color 14:194–195, 14:197
Grapheme to touch 14:196
Graphemes 14:193–195, 16:593–595
Grapheme-to-phoneme (GP) conversions 23:266–267, 23:269
Graphemic buffer agraphia 6:735
Graphemic buffer dysgraphia 1:452–454
Graphic representation of relational coping 20:224–226
Graphical displays, in exploratory data analysis 8:600
Graphical Markov models 10:341–350
applications and statistical research 10:344–348
implied dependences 10:346–347
implied independences 10:346
independence-predicting graphs 10:347–348
independence-preserving graphs 10:347–348
Markov equivalence of regression graphs 10:345–346

Graphical Markov models (*continued*)
distortions, graphical representations 10:348
regressions and data generating processes sequences 10:342–344
child development data 10:344
chronic pain treatment 10:343–344
regression graphs 10:342–343
types of model, general and historical remarks 10:341–342
Graphical models 14:324, 23:619
Bayesian 2:363–367
latent variables 13:395
Graphical representation of objects 16:34
Graphical user interfaces (GUIs) 7:4, 23:378–380
Graphicness (Anschaulichkeit) 25:208
Graphics 10:351–357
background 10:351–352
bar chart 10:360–364, 10:368
bar charts 10:351–353, 10:356, 10:351f–352f
biplots 10:355, 10:355f
box plots 10:353, 10:355–357, 10:353f
Cartesian coordinates 10:358–359
Computer Assisted Instruction (CAI) 4:493–494
coordinates 10:368
curves 10:359–360
curvilinear relation 10:358
equations 10:359
exploratory data analysis 10:351
future 10:357
geometry 10:360–368
one variable 10:360–362
three variables 10:365–368
two variables 10:362–365
histograms 10:351–353, 10:355–357, 10:352f–353f, 10:356f
history 10:359
interactive displays 10:352–354
dynamic graphics 10:354
features 10:352–353
interfaces 10:353
linking 10:353
mosaic plots 10:354–357, 10:355f
multivariate displays 10:354–355
categorical variables 10:354–355
continuous variables 10:354
projective displays 10:355
parallel coordinate plots 10:354, 10:354f
physical coordinates 10:359
pie chart 10:359–362
principles 10:359–360
regression analysis 14:157–161
scatterplots 10:352–356
software 10:357
specialist displays 10:355–356
missing values 10:355–356
weighted displays 10:356
statistical tools and plots 10:356–357
uncertainty 10:369–370
GraphoGame 20:9–10
Graphotactic cues 23:269, 23:271
Graph-theoretic depiction 23:605
Gras, Norman
works
Business and Capitalism 3:21
Casebook in American Business History 3:21–22
Guide to Business History 3:21
Graspable user interface (GUI) 24:20
Grasping, cortical areas in 5:23, 5:25–28
GRASS (Geographic Resources Analysis Support System) 10:97
Grasses, and food production 9:300–303
Grassland Biome Study 7:792–793
Grassroots democracy 10:381
Grassroots movements, theory of 12:913–914
Grass-roots sports fans 14:722–723
Grattan Institute 10:249
Graunt, John 6:171, 6:174, 20:875
works
Natural and Political Observations Made upon the Bills of Mortality 7:559, 23:423
Graves, Richard 14:457
Gravity models 24:611
of spatial interaction 23:152
of urban systems 24:808–809
Gravity potential method of geographical access 1:66
Gravity Recovery and Climate Experiment (GRACE) 20:423
Gray, Asa 5:678, 20:287
Gray, Elisha 24:154
Gray, J., on paratext 25:81
Gray, Mary L., works, *Out in the Country* 19:754–755
Gray economy 5:171
Gray Ghost Ventures 1:270
Gray market in guns 10:457
Gray matter 18:819–820
developmental neuroscience 16:600
Gray Oral Reading Tests - Fifth Edition 16:725
Grayson Clamp 16:715
GRD *see* Ground resolved distance (GRD)
Great Acceleration 1:724
Anthropocene 1:724
Great Ape Project 1:704
Great Britain *see* United Kingdom (UK)
Great Chronicle of Kings (*Mahazawingyi*) 11:83
Great circle 14:285
The Great Convergence 22:172
Great Crime Wave 19:646
Great Depression 2:255, 9:183–185, 9:187–189, 11:151, 16:450, 16:910, 24:764–765
egoistic giving during 3:364
Fisher's theories 9:220
public policy 19:569
Great divide theories
and anthropology 1:751
orality–literacy debate 17:276
Great Financial Crisis (GFC)
and health 10:201
Great Hanshin-Awaji Earthquake 4:203
Great Leap Forward 2:66, 24:973–974
Great power wars 25:420
Great Powers 2:346–347
role in Security Council 2:347
Great Proletarian Cultural Revolution 1:459
Great Rebellion 7:650–651
Great Recession 2:245, 25:830
evidence 25:830–832, 25:832t
poverty policy 18:749–750
Great Satan 20:403
The Great Satan of western imperialism 20:398
Great Smoky Mountain Study 15:183
Great Social Revolution 3:518–519
Great society 7:926
Great Spirit 5:573
Great Transformation 3:745–746
Great U-turn, and income inequality 11:740–741
Greater Antilles 1:638
Greater German Reich 11:152
Greater Romania Party (PRM) 18:610–611
Greatest Happiness Principle 2:526–527
Greatest lower bound, in classical test theory 19:428–429
GREBE system 13:787
Greco-Roman philosophy 20:284–285
Greco-Roman warfare 25:414
Greece
clinical psychology, development 3:874
mass society 14:743–744
psychology in 19:406–408
small states 22:110–111
Greek
and diglossia 6:419–421
language contact 13:300–301, 13:303–305
Greek alphabet 25:776–778, 25:778t
Greek astronomy 18:100
Greek Civil War 3:717
Greek concept of stasis 3:714
Greek geometry 2:338–339
Greek historiographical production, variety of 11:36–37
biographies 11:36
cultural histories 11:36
foundation histories 11:36
universal history 11:36
Greek historiography 11:35–36
Peloponnesian war 11:36
Persian wars 11:35
Thucydides' model of historiography 11:36
Greek life, and hazing 10:554–561
Greek meters 13:275
Greek natural science and medicine 20:285
Greek philosophy 7:255, 18:31–32
Greek republican citizenship 3:644
Greek script 14:242
Greek view of the world 20:285
Greeks, diaspora 6:353
Greekthink 10:554–555
Green, David M. 21:954
Green, Shane 11:774
Green building 7:775–776
Green capitalism 10:392, 23:815
Green cities 3:665, 6:834
Green consumerism 7:699, 8:25

Green consumption 4:760–761, 8:27
Green criminology 5:160
Green development 23:798
Green economy 14:763
Green Goods and Services Survey (GGS) 10:374–375
Green groups 7:745
Green growth strategies 6:836
Green initiatives 10:372
Green innovations 18:200
  and sustainability 18:200
Green jobs 19:84
  carbon intensity 10:373–375
  climate change policies 10:377
  conceptual framework of 10:373–374
  definitions and measurement of 10:374–376
  energy 10:372–375, 10:377
  environmental quality 10:372–375
  environmental regulation 10:372, 10:374
  existing employment studies 10:376–377
  green initiatives 10:372
  number and percentage of 10:375t
  program evaluation 10:377–378
  simulation models 10:372–373, 10:377
  in United States
    aggregate hours worked and total wage bill 10:377f
    characteristics of private sector green jobs 10:376t
    industrial distribution 10:376t
Green Jobs Act 2007 10:372
Green Marxism, sustainable development 23:815
Green movement 7:812
Green Open Access 12:866
Green Paradox 17:254
Green parties 7:765–766, 10:379–384
  antinuclear movements 10:381
  coalition government 10:379–381, 10:383
  electoral fortunes 10:381–382
  electoral performance 10:382–383, 10:382t
  emergence 10:379–380
  environmental policy 10:380, 10:383
  European Parliament 10:379–380, 10:382
  family 10:380–381
  government 10:383–384
  left-libertarian 10:380, 10:383
  new middle class 10:379, 10:381, 10:383
  new social movements 10:379–381
  nuclear power 10:379–381, 10:383–384
Green revolution 1:459, 1:566, 9:307–310, 12:726, 16:799, 24:144
  for Africa 9:309–310
  consequences of 9:308–309
  criticisms 9:309
  global food insecurity 9:309–310
  history of 9:307–308
  nature of high-yielding varieties in 9:308
  technologies 1:472–473, 1:484
Green social work 3:426, 10:385–391
  climate change 10:385, 10:387–388, 10:390
  coproduction of solutions to problems 10:385, 10:387, 10:390b
  defining 10:385–386
  disasters 10:385, 10:387–388, 10:390
  and ecological social work, commonalities between 10:386–387
  environmental degradation and environmental injustice 10:387–388
  examples 10:389–390
    multilevel interventions during extreme weather events 10:389–390
    securing affordable food 10:389
  and harmony 10:533
  and human rights 11:355
  interdisciplinarity 10:388
  neoliberal capitalism 10:387
  practice 10:388
  sustainable development 10:386–388
Green Street 25:822
Greenbeard, kin selection 13:29–31
Greenbeard effect
  kin selection 13:29–31
Greenberg, Joseph H. 14:209–210, 23:885
  works, *Language Universals* 14:533–534
Greenbergian functional-typological 23:885
Greene, Maxine, aesthetic education 1:230–231
Greenfeld, Liah 16:285
Greenfields 24:865–866
Greenham Common 19:748–749
Greenhouse effect 7:793–794
Greenhouse gases (GHGs) 7:769, 7:771–774
  emissions 1:723, 6:831–832, 6:854–855, 6:857, 7:798, 7:800–801, 20:571, 24:616
  and sustainable transportation 23:846
Greening 7:812
  of business 10:392
  of technology 12:716
  of theology 20:288
Greening jobs 19:84
Greenland 1:922
  enlightened paternalism 1:928
  Norse peoples in 1:925
  Thule people in 1:924–925
Greenleaf strategy 23:759
Greenpeace 3:854, 7:763–765, 23:814
Greentrification 10:47
Greenwashing 22:169
Greenwich Millennium Village 16:811
Gregg v. Georgia 13:768
Gregorian reforms 3:234
Gregory, Derek
  works
    The Colonial Present 11:20
    The Dictionary of Human Geography 10:88
    Ideology, Science and Human Geography 20:121
Gregory, S., on North American community 17:23
Gregory the Great 11:885
Greiffenhagen, Martin 4:684
Grele, Ronald 17:284–285
GREs *see* Glucocorticoid response elements (GREs)
Gresham's law 15:46, 18:356
Grey, Edward, 1st Viscount Grey of Fallodon 9:213
Griaule, Marcel 1:767, 7:376
  on intellectuals 13:72
Grice, Paul 18:796–797
  conversational implicature 21:589
  cooperativeness principle 20:883
Grid cells 11:387–388, 16:623
  and spatial navigation 23:163, 23:166–168
Grid organization development, organizational development 6:298
Gridlock, and political trust 24:661
Grief 2:531, 5:862, 7:515, 12:632–633
  across life span 2:537
    anxiety 2:537–539
    cognitive restructuring 2:540
    depression 2:537–541
    disordered grief, diagnostic criteria for 2:539
    emotion regulation 2:538, 2:540–541
    life event 2:537
    models 2:537–538
    post-traumatic stress disorder 2:538–541
    psychological intervention 2:539–540
    psychotherapy 2:539–540
    responses 2:537
  and bereavement 2:531
  and childhood stress 23:571
  death/dying 5:865–866, 5:868
  health risks of 7:497–498
  and integrated attachment theory 12:229
  and mass trauma 14:749–750
  nature of 5:870–871
Grief to Personal Growth Model 5:862
Grier, Colin, on anarchism 1:682–683
Grierson, John 6:621
Griesinger, Wilhelm 13:127
Grievance 3:716, 4:145–148, 12:223, 13:590, 13:592, 13:594–595
  disputes 6:560–562
  social protest 22:507, 22:509–510
Grievance procedures
  for sexual harassment in United States 21:723–725
Griffin, P.
  on women's participation in sports 9:657–658
  works
    Strong Women, Deep Closets 9:659
Griffith, D.W., works, *The Girl and Her Trust* 9:167
Griffith, Mary, works, *Three Hundred Years Hence* 14:255
Griffiths, J. 13:759
Grillo, Beppe 11:416–417
GRIM strategy (game theory) 18:935
Grimm, Jacob 22:887–888
GRIN2B gene 2:810
Grindr 21:794–795
Grindstaff, L., on television 25:80–81, 25:85
Grisez, Germain 16:320
Griswold v. Connecticut (1965) 18:948
Grit 12:381–382
  and achievement 10:397, 10:399–400
  and conscientiousness 10:397, 10:399

Grit (*continued*)
correlates and consequences of 10:399
and intelligence 10:397–399
measurement 10:398–399
mechanisms linking to outcomes 10:399–400
and motivation 10:400
and perseverance 10:397
prior research on persistence 10:397–398
psychological antecedents of 10:400
and retention 10:399–400
and success 10:397–398
Grit Scale 10:398–399
Grit-S 10:398–399, 10:398t
GRK *see* Genetic relationship kernel or matrix (GRK)
GRM *see* Graded response model (GRM)
GRO *see* General Register Office (GRO) (UK)
Grobalization 5:539
Grodal, T., works, *In his book Moving Pictures* 9:162
Groomprice 6:643–644
Großdeutschland 25:612
Großdeutschland (combat unit) 25:612
Gross, Jan Tomasz 6:814
works, *Neighbors: The Destruction of the Jewish Community in Jedwabne* 6:814
Gross, Neil 14:873
works, *Richard Rorty: The Making of an American Philosopher* 20:765–766
Gross calorific value (GCV) 7:629
Gross cell product (GCP) 7:49
Gross domestic income (GDI) 5:752
Gross domestic product (GDP) 3:100, 4:946, 5:751–752, 6:24–25, 7:178, 15:101, 16:227, 22:605, 23:203, 23:644, 24:894
of China 6:790
definition 5:752
economic cost of climate change impacts 3:860
and healthcare expenditure 10:666–667
per capita 25:641
R&D and modernization 20:497
and R & D expenditures 20:497–498
and welfare 5:754–755
Gross flows and nodes 9:177–178
Gross national product (GNP) 12:50, 15:782
per capita 5:323
Gross value added (GVA) 5:751–752, 5:754–755
Grossberg, Larry 6:551
Grosswissenschaft 3:762
Grotius, Hugo 3:771–772, 12:507, 13:802
consensual theory of the law of nations 12:507–508
definition of law of nation 12:508
jus gentium 12:507–508
works
De Iure Belli et Pacis 3:771–772
De Jure Belli Ac Pacis 12:507
Ground resolved distance (GRD) 20:420
Ground truthing 20:420
Grounded cognition 7:421
Grounded social cognition 7:409, 7:412
Grounded theory (GT) 1:911, 10:402–407, 10:436, 13:872, 20:502–506, 22:545–546, 22:987, 23:857–858
conducting 10:405–406
coding qualitative data 10:405
memo writing 10:405
theoretical sampling 10:406
constructivist 10:404
crucial characteristics 10:402
developments and changes 10:404
explanation of method 10:402–403
future directions 10:406–407
Glaser's imprint on 10:403
intellectual heritage of 10:403–404
objectivist 10:404
qualitative health research 10:692–693, 10:695–696
qualitative research 10:403
in qualitative research 19:685
Strauss's imprint on 10:403, 23:530–531
use of term 10:402
writings of pragmatist philosophers 10:403
Grounded theory methodology (GTM) 20:505–506
qualitative developmental psychology 19:673
Grounding
coconstructivism 3:914–915
physicalism 18:106, 18:111–112
Ground-penetrating radar (GPR) 9:335, 9:337
Groundwater 25:437
Groundwater management 25:437–438
Groundwork of metaphysics of morals (GMM) 13:5–7
Group analysis 24:275
Group Areas Act 6:225
Group average 19:80–81
Group boundaries of identity 11:528–529
Group care 20:536–538
Group coaching for leadership development 13:668–669
Group coherence 7:392, 24:711
Group cohesion
evolutionary group dynamics 8:444
Group conflict, and collective violence 14:702–703
Group consciousness 7:392
Group coordination
evolutionary group dynamics 8:441–442
Group decision making
evolutionary group dynamics 8:444–445
Group decision support systems 23:139–140
Group delinquency 6:68–69, 6:72
Group detection, cohesive 3:910–911
deterministic group finding approaches 3:910
search/optimization approaches 3:910–911
Group discrimination 18:831
Group discussions
qualitative developmental psychology 19:671
in qualitative health research 10:693–694
Group dynamics 8:441, 16:524–525, 22:703, 24:275
evolutionary 8:441–447
experimental economics 8:550
Group effects, cultural transmission 5:403–404
Group ergonomics 11:301–302
Group failure 24:714
Group formation
self-categorization theory 21:457–458
structural balance theory 23:606
Group goals
cooperative learning 4:882–883
Group Guidance Project in Los Angeles 9:596
Group harms 9:964
Group identities 11:526, 22:247, 23:464
boundaries 4:177
and boundaries 11:910
conceptualization 4:174–175
dimensions, layers and types of 4:179
emerging and evolving 4:175
empirical manifestations 4:175–176
and expressive forms 4:174–180
analytic approaches 4:176–177
identity work 4:177
theoretical and empirical issues 4:177–179
identity correspondence problems 4:177–179
intergenerational investments 4:178
legitimating 4:179
network appropriation thesis 4:178
in political psychology 18:431–432
selective incentives 4:178
situational relevance 4:176–177
and social movements 4:175–179
threshold model 4:178
types 4:179
Group influence 22:346
Group investigation
cooperative learning 4:885
Group leaders 10:437
Group learning techniques 10:417
Group living 22:854–855
Group matching 5:356–357
Group membership, reference group 20:72–74
Group on Earth Observation (GEO) 23:133–134
Group performance 17:612, 17:614–616
Group polarization, self-categorization theory 21:457–458
Group processes 10:422–427, 19:350
affiliation 10:423
in classroom 10:416–421
cohesion 10:423
collective behavior 10:425–426
and collective memory 4:191
commitment 10:423
in competition/cooperation 4:859
crowd 10:425–426
decision making 10:426
groupthink and polarization 10:426

intergroup decision making 10:426
remembering 10:426
deindividuation 10:425–426
development 10:423
direct influence and persuasion 10:423
formation 10:422–423
indirect influence and group norms 10:423
influence 10:423
intergroup dimension 10:426–427
leadership as 13:648, 13:652–653
performance 10:425
social loafing and social dilemmas 10:425
social identity 10:422–423, 10:425–426
social influence 10:423–424, 10:426
social psychology 22:541–542
structure
communication structures 10:424
deviance 10:425
and leadership 10:424
and positions 10:424
and roles 10:424
and status 10:424
studying 10:422
Group processes, in classroom 10:416
cognitive/elaboration perspectives 10:417
developmental perspectives 10:417–418
influencing factors 10:418–420
quality of students' discourse 10:418–419
task, role of 10:419–420
teacher, role of 10:419
research and practice 10:420–421
social cohesion approach 10:417
social–behavioral perspective 10:416–417
sociocultural and combined theoretical perspectives 10:418, 10:418t
teachers' use of cooperative learning 10:420
usable models of group process 10:420–421
Vygotskian perspective 10:418
Group processes, in organizations 10:408–415
change 10:413
change through non-western group processes 10:413
steps of organizational transformation 10:413
conformity and dissent 10:408–409
groupthink 10:408–409
minority influence 10:409
team mental models 10:409
definition and structural dimensions 10:408
degree of cohesion 10:408
norms and values 10:408
role and status 10:408
diversity 10:409–410
creativity and innovation 10:410
dimensions 10:410
gains and losses 10:409
dynamics 10:411–412
culture and 10:412
intragroup conflicts 10:412
stages 10:411–412
future research 10:414
in-group and out-group processes 10:412–413
integration of individual contributions 10:410–411
counterproductive work behavior 10:411
group performance and task type 10:411
synergy 10:410
top performance and organizational culture 10:411
and leadership 10:413
leadership in multinational corporations 10:413–414
shared leadership 10:414
social identity theory 10:412–413
Group Psychology and the Analysis of the Ego (*Massenpsychologie und Ichanalyse*) 19:350
Group psychotherapy 10:428–433, 24:275
approaches to 10:428–430
Bowen's 19:505
cognitive–behavioral therapy groups 10:429
cohesion and development 10:431
corrective emotional experiences 10:431
development 19:505
effectiveness of 10:431–432
existential and humanistic approaches 10:429–430
formal 19:505
interpersonal 10:429
interpersonal learning 10:430
Lewin's 19:505
for mass trauma 14:749
psychoanalytic 10:428–429
self processes in 10:431
social motivation in 10:430
Group randomized trials *see* Cluster randomized trials (CRTs)
Group rewards 10:416–417
Group selection 16:332, 22:900–901
evolution of morality 15:813
Group selves 21:452–453
Group size 9:276, 22:246
Group style, in cultural sociology 5:619
Group therapy 2:426–427, 10:429–430, 19:515
for depression 19:375–376
Group value model 19:367
Group work with men 14:683
Group-based analysis of crime 1:313–314
Group-based emotions 4:196–197
Group-based trajectory models (GBTM) 5:140, 14:62–63
considerations for 14:64
desistance/persistence 5:140–143
and life-course criminology 14:63–64
Group-centered leadership 20:738–739
Group-conditional densities 16:117
Group-covariance matrices 16:117–118
Groupe de Sociologie Politique et Morale (GSPM) 2:747
Grouping
computational auditory scene analysis 2:224–227
pure model-driven grouping 2:226
signal-driven grouping 2:224–226
target source identification in unknown background 2:226–227
Groupishness 10:434, 10:438
Group-level phenomenon
acculturation as 1:68–69
framework 1:69f
Group-level processes
political psychology 18:431–433
group identities and cultural diversity 18:431–432
international conflict 18:431
racism 18:431
social change and collective action 18:432–433
Group-membership of the entity 16:116
Group-posterior probabilities 16:117
Groups
with acquired economic status 11:896
cohesive, detection 3:910–911
deterministic group finding approaches 3:910
search/optimization approaches 3:910–911
in collectivism/individualism 4:208
cooperation 4:863–866
definition 10:422
and essentialism 8:13–15
ethnic 8:149
ethnocentrism and prejudice 8:168–174
and factionalism 8:683–685
forming impressions of 22:189
genomics 10:37–38
and identification, connection between 22:314–315
intergroup relations 7:502–506
interpersonal psychotherapy (IPT-G) 12:633
kinds of 11:817
legitimacy of status structures in 13:824–825
origins of violence between 14:703t
political representation 18:442
Rogers on 20:737–742
and social networks 5:560–561
social psychology 22:527–530
and social structure 22:720–722, 22:724–725
sociology of 10:434–440
cognitive interdependency 10:434–438
collective memory 10:436–439
cooperation 10:434
dependency networks, groupishness, and individual differences 10:438
evaluation 10:438–439
exchange theory 10:438
expectation state theory 10:434–435, 10:437–438
functional interdependency 10:435–436
goal frame 10:437
idioculture 10:436
joint production 10:437
motivation 10:437
norms 10:436–437

Groups (*continued*)
power-dependence 10:438
shared knowledge 10:436, 10:438–439
shared mental models 10:436
shared reality 10:436
small group research 10:435–436
social embedding 10:436
social identity 10:435–437
status hierarchies 10:437–438
structural interdependency 10:434–436, 10:438–439
teams 10:434, 10:436–438
transactive memory 10:436
transformational leadership 10:438
as sodalities 11:896
structured 11:816
structures 23:628–629
taxonomic 11:816
and terrorism 24:242
and workplace autonomy 25:720
Groups alive 10:438–439
Groups of transformations, invariance 12:706–707
Group-serving tendency 8:114
Groupthink 5:467–468, 5:470, 10:408–409, 10:426, 10:554–555
research, and tyranny 24:710–711
Groupwork 22:775–776
social work education 22:775–776
Grove, Richard, works, *Green Imperialism* 7:748
Growlr (smartphone app) 13:891
Growth 7:121–123, 9:40–41
animal's life history 9:40
of bureaucracy 2:910–911
and development 2:661
development continuum 7:123
economic policies and 7:125
of new jobs 24:916
in OECD countries 7:124
of population *see* Population growth
political economy of 20:895
relation to developing nations 7:123
Growth accounting 7:123–124
Growth Balance method 6:146
Growth control 10:447
Growth curve analysis 10:441–446
data collections 10:441–443
linear models for 10:443–444
multivariate growth curve analyses 10:445
nonlinear models for 10:444–445
recent issues in 10:445–446
Growth curve models (GCM) 5:140, 10:441, 19:734, 20:494, 23:619
desistance/persistence 5:140–141
Growth hormone (GH), and sleep 22:80, 22:82–83
Growth hormone-releasing hormone (GHRH)
and sleep 22:83–85
Growth Limitation Ordinance 10:451
Growth management 10:447–452
benefits 10:451–452
defined and early history 10:447
differentiated from related concepts 10:449–450
goals 10:447, 10:449–452
local 10:451
primary elements 10:447–449
coordination between governments 10:449
enhancing community life 10:449
growth guidance 10:448
infrastructure provision 10:448–449
overall form 10:447–448
protecting open spaces 10:449
regional 10:451
state 10:450–451
urban sprawl 24:937
Growth Management Act (GMA) 10:450
Growth modeling for sexual development
age-based time metric 14:333–334
menstrual age-based time metric 14:334–336
wave-based time metric 14:330–333
bivariate growth with an exogenous outcome 14:331–333
bivariate model for number of sex partners and drinking 14:331
ratings of days drunk 14:330
total number of sexual partners 14:330
Growth monitoring, Oral rehydration, Breastfeeding, Immunization *see* GOBI (Growth monitoring, Oral rehydration, Breastfeeding, Immunization)
Growth points (GP) 10:113, 10:116
contexts and 10:113
as a psychological predicate 10:113
Growth theories 7:38, 20:728
of law firms 13:580–581
unified 7:42
GRS *see* Gender reassignment surgery (GRS)
Grundbegriffe 4:548–550
Grundgesetz 23:441
Grundnorm 13:763–764, 13:792–793
Grundschuld 21:403–404
Grüning, Ferdinand 16:228
Grupo Gay da Bahia 13:861
GS *see* Genealogical Space (GS)
GSAs *see* Gay-straight alliances (GSAs)
GSCC *see* General Social Care Council (GSCC)
GSG *see* Global Scenario Group (GSG)
GSIA *see* Graduate School of Industrial Administration (GSIA)
GSLC *see* Gendered Sexuality over the Life Course (GSLC) Model
GSOEP *see* German Socio-Economic Panel (GSOEP)
GSP *see* generalized secretary problem (GSP)
GSR *see* Galvanic skin reflex (GSR)
GSS *see* General sensation seeking (GSS); General Social Survey (GSS)
GSSAB *see* Global Study of Sexual Attitudes and Behaviors (GSSAB)
GST *see* General strain theory (GST); General systems theory (GST)
gStudy (software) 21:544
GT *see* Grounded theory (GT)
GTD *see* Global Terrorism Database (GTD)
GTM *see* Grounded theory methodology (GTM)
GTO *see* Getting-to-Outcomes (GTO)
Gu, Ban 11:49
works, *Hanshu* 11:49
Gu, Jiegang 11:51
works, *Gushibian* 11:51
Guam 17:158–161
Guang, Sima, works, *Zizhi tongjian* 11:49–50
Guangdong Province 23:208, 23:210–212
specialized towns in (case) 23:211
Guanxi 11:786
Guardian ad litem 3:479
Guardian jobs 2:52–53
Guards 11:723
Guatemala 22:624
sociocultural aspects 1:621–625
syphilis studies in 13:561–562
Guattari, Félix 1:237–238, 19:750
on networks 16:567–568
Guerin, Daniel 19:748–749
Guerrilla fighters
and child soldiers 3:439–443
Guerrilla terrorism 17:620
Guerrilla warfare 12:221–222, 25:379
Guest authorship 20:530
Guest worker migrations 15:450
Guest worker programs 10:169–170, 11:664
Guha, Ranajit 23:632
works, *The Elementary Aspects of Peasant Insurgency* 11:42
GUI *see* Graspable user interface (GUI)
Guicciardini, Francesco 12:386–387
The Guide to Clinical Preventive Services 8:336
Guided activation model of prefrontal cortex 4:35–36
Guided discovery learning
instructional psychology 12:203
Guideline sentencing data 21:632
Guidelines for Multinational Enterprises 8:21–22
Guilds 2:46–51, 15:678–680
concepts 2:48–49
end of 2:48–49
history 2:48–49
in labor movements 2:49
in petit bourgeois 2:49
preindustrial 2:49
revisionism in 2:49–50
Guilford, J.P. 3:243
works, *Encyclopedia of Economics* 7:590–591
Guillen, Mauro, on cultural logics of action 5:539
Guilt 10:453–456, 21:475–477
adjudication outcomes for felony defendants 10:454t
cognitive-anatomical basis of 16:611–612
cultural differences in 21:478–479
definition 10:453–454
determination of 10:453

implications for deterring transgression and socially undesirably behavior 21:476–477
implications for interpersonal adjustment 21:476
implications for psychological adjustment 21:476
in intergroup relations 7:504–505
maladaptive 21:477
mistakes of 10:455
and moral development 15:788–790, 15:794
patterns of 10:454
in federal courts 10:454
in state courts 10:454
and religiosity of women 25:636
GUIs *see* Graphical user interfaces (GUIs)
Guizot, Francois, works, *Histoire de la révolution d'Angleterre* 7:650
Gulag 20:820–823, 20:853
Gülen movement 16:807, 20:406, 24:605
Gulf doctrine 21:169–170
Gulf War Propaganda 25:398–400
news media and 25:399–400
Gulf War syndrome 11:596
Gulf Wars 25:423
Gumbel distribution 6:572, 24:618
Gumplowicz, Ludwig 6:241, 6:243
works, *The Outlines of Sociology* 6:243
Gun control, and crime 5:196, 5:198–201
Gun laws 7:82
Gun shows 10:457
Gunboat policy 11:685
Gun-Free Schools Act 21:108
Gunkel, Henriette, on NCGLE 21:746
Gunman Writ Law 13:764
Gunpowder Age
warfare in 25:416–417
organization 25:416
purposes 25:416–417
technology 25:416
Gunpowder weapons 25:416
Guns 10:457–461
carrying 10:458–459
crack epidemic 10:458
crime, crime rates 10:458
crime-precipitated homicides 10:458
criminal gun market 10:457–458
firearms acquisition 10:457–459
firearms behaviors and criminological theory 10:458–459
firearms laws and the reduction of gun violence 10:459
gray market 10:457
gun shows 10:457
homicides 10:458–459
hot spots 10:459–460
law enforcement strategies and reduction of gun violence 10:460
military style weapons 10:457
retail sales 10:457, 10:459
routine activities 10:458–459
school shootings 10:457–458
shall issue permits 10:459
Guns carrying 10:458–459
GUO *see* Global Urban Observatory (GUO)
Guo, Moruo 11:52
works, *Zhongguo gudai shehui yanjiu* 11:52
Guomingdang (GMD) 3:515
Gurevich, Aaron, works, *Categories of Medieval Culture* 11:40
Gurian, Waldemar 18:434–435, 18:437–438
Gurland, Henny 9:461
Gurlesque (strip club) 13:884–885
Gürsel, Zeynep Devrim 2:873
Gurvitch, Georges 13:808
works
Traité de sociologie 7:588–589
Twentieth Century Sociology 7:588–589
Gustatory hallucinations 10:500
Gustatory input processing, by temporal lobe 24:208
Gut, axis formation and positioning of 2:116–118
Gut flora 9:605
Gut microbiota 9:605
Gut reasoning 2:194
Gutenberg, Johann 18:915, 18:917
Guthrie, Edwin 19:414
learning theory 13:706
Gutierrez, Gustavo 22:386
Guttenberg Galaxy 25:191
Guttentag, Marcia, works, *Too Many Women* 18:590–591
Guttentag–Secord theory of sex ratios 22:896
Guttman scales 2:179, 21:563
Guttman scaling 17:527
and ordered relational structures 17:297–298
Guttman scalogram analysis 21:6
Guugu Yimithirr language 10:110
Guy & Myriam Ullens Foundation 5:599–600
Guyana 1:638
Guyau, Jean-Marie 1:714, 1:719
Guzman, Patricio, works, *Battle of Chile* 6:622–623
GVA *see* Gross value added (GVA)
GVC *see* Global value chains (GVC) framework
GVCs *see* Global value chains (GVCs)
GWAS *see* Genome-wide association study (GWAS)
Gwei-Djen, Lu 16:397
Gwich'in protocol for traditional knowledge research 14:187
GWP *see* Geo Web Portal (GWP)
Gyémánt, Ladislau 4:571–572
Gynecology
psychosocial aspects 10:462–468
disorders not related to the menstrual cycle 10:465
disorders related to menstrual cycle 10:462–465
endometriosis 10:464
female menstrual cycle 10:462–463
infertility 10:465
menarche 10:463
menstrual cycle disorders 10:463
polycystic ovary syndrome 10:464
premenstrual syndrome and premenstrual dysphoric disorder 10:463–464
reproductive health 10:462
sexual dysfunctions 10:465
urinary incontinence 10:465
psychosomatic disorders in 19:488
Gypsies, Roma 20:754–755
Gypsy schools 20:756
Gyrus 3:336–337, 3:339–340

H

HA *see* Harm avoidance (HA)
Haar-like feature 8:658
Haarmann, H. 8:351–352
Haas, Peter
on epistemic communities 7:862–863
model 3:840
Habakkuk, H.J., works, *American and British Technology in the Nineteenth Century* 7:56
Habeas corpus 6:698
Habermas, Jürgen 1:473, 1:721, 2:519, 2:522–523, 3:711, 4:794, 5:255–256, 5:296–298, 6:506, 7:875, 10:469–474, 10:740–741, 11:175, 13:440, 13:555, 18:626–627, 19:616, 19:677–678, 20:203, 21:177, 22:994, 24:39
on colonization of lifeworld 10:471
on communicative action 10:469–471, 10:473
on critical theory 5:304–309, 10:469–473, 10:787
critiques and continuations 10:472–473
on democratic constitutional state 10:471–472
on discourse ethics 10:471
on Enlightenment 7:667
in Europe 10:470, 10:472–473
and Frankfurt School 9:383
on German social theory 5:296–297
hermeneutic claim to universality 10:469
increased complexity 10:471
on language philosophy 4:790–791, 4:794–795
on law 10:471–472
moral theoryof 5:302
on morality 10:471
on naturalism 10:472
on philosophy 10:469, 10:471–472
on placeholder and interpreter 10:469
on political public sphere 19:610
on politics 10:469, 10:471–472
on positivism 10:469
on postmodernism 5:297
on 'postnational' democracy 10:472
on pragmatism 18:810
on public reason 19:593–594
on public sphere 3:703
on rationalization 10:470–471
on reconstructive science 10:469–471

Habermas, Jürgen (*continued*)
on reification 20:203, 20:205
on religion 10:472
on republicanism 20:477
on social change 10:471
works
The Philosophical Discourse of Modernity 7:667
The Structural Transformation of the Public Sphere 5:619, 19:603–604, 19:616
Strukturwandel der Öffentlichkeit 19:603–604
Habilitation 1:26–27, 7:378
Habit learning 16:844
Habitat 22:264
Habitat II Conference (1996), Istanbul 7:771
Habits 2:882, 10:475–479
action 10:475–478
reflective model 10:478
addiction 1:98–99
Aristotle on 10:476
attitude 10:478
behavior 10:475–478
behaviorism 10:477–478
Bourdieu on 10:478
concept 10:475–476
Durkheim on 10:476–478
and Enlightenment 10:476
historical development 10:476–478
age of ferment 10:476–477
antiquity to early 1800s 10:476
demise 10:477–478
and instinct 12:175–177, 12:179
in institutional economics 17:191
Kant on 10:476
recent developments 10:478
in sustainable practices 23:835–836
Thomas on 10:476–478
and unconscious mental activity 24:723
Watson on 10:477
Weber on 10:476–478
Znaniecki on 10:477–478
Habits 2:509
Habitual arousal biases 23:294–295
Habitual cigarette smoker 6:670
Habitual knowledge 10:484–485
Habituation 10:480–483
brain processes of 10:480
cardiac defense 23:541–542
classical conditioning 3:765, 3:768
in learning 23:876
neural mechanisms 10:482–483
properties of 10:481
sensitization 10:480–481
stimulus intensity and 10:481
Habituation paradigm, and early memory development 6:761–762
Habituation–dishabituation task 17:199
Habitus 1:903, 2:719, 4:695, 4:697–698, 5:617–618, 7:145–146, 7:415, 8:900, 9:462–463, 10:475–478, 10:484–489, 16:238–239, 18:783, 18:785–786
Bourdieu on 2:777–778
cleft 10:488
collective memory 4:201
concept of 2:778
consumption 4:759–760
continuous and conscious religious 10:484
deducing from observed behavior 22:334
elites 7:398–399
flexible, context and 22:335
and generation 9:853
as generative principle of individual practices 10:487
as a heuristic tool 10:484–485
and language 13:286–287
and language policy 13:354
national 10:485
predatory 10:484–485
predicting observed behavior from 22:334–335
psychological 10:485
psychological individual 10:485
religious 10:484–485
sacred 10:484
social 10:485
in sociological theory of Pierre Bourdieu 10:486–488
theoretical 10:484–485
Weberian definition 10:484
Habsburg Empire 6:811, 16:299–300
Hacienda Ecocities 6:835–836
Häckel, Ernst 7:809
HAD *see* Human immunodeficiency virus (HIV)-associated dementia (HAD)
HADD *see* Hyperactive agency detection device (HADD)
Haddon, A.C. 14:453–454
Hadith 12:754–757
Hadot, Pierre 9:368
Hadrian 11:37–38
HADS *see* Hospital Anxiety and Depression Scale (HADS)
Hadza 6:775–776
Haeckel, Ernst 8:385, 22:263
Hage, G., on multiculturalism in Australia 2:241
Hägerstrand, Torstein 15:638
conceptualization of diffusion process 6:407–408
methodological framework 6:408
tradition 6:407–408
works
Time Geography 23:200
Hägerström, Axel 21:169
Haggett, Peter, works, *Locational Analysis in Human Geography* 20:123
Hagiography 2:644
Hague Convention for the Protection of Children 1:136–137, 1:139, 1:145–146, 17:640
Hahn, Hans 25:90
Haig–Simons income 11:733
Hair cells, vestibular 25:65
Hair cortisol analysis 23:578
Hair-pulling disorder *see* Trichotillomania
Haiti 1:639
water, health and inequality 25:431–433
epidemic 25:431–432
epidemic, dealing with 25:432–433
Haiti Carnival 3:147
Haitian Creole (HC) 18:136
diglossia in 6:419
Haitian political administration 25:432
Haitian revolution (St. Domingue) 22:40, 22:842, 25:282
Haji, Raja Ali
works
Silsilah Melayu dan Bugis dan sekalian raja-raja-nya (Genealogy of Malay and Bugis Princes) 11:84
Tuhfat al-Nafis (The Precious Gift) 11:84
Hajnal, John 11:8
HAL *see* Hyperspace analogue to language model (HAL)
Halberstam, Judith Jack 19:748–749, 25:211
Halbwachs, Maurice 4:188, 4:194, 5:499
on collective memory 4:200, 9:866
on generations 9:866
on memory 1:881–883
works
Collective Memory 1:748
Les Causes du suicide 1:715
HALCyon *see* Healthy aging across the life course (HALCyon) research
Haldane, J.B.S.
on kin selection 13:26
Haldane, John Burdon Sanderson 9:224, 13:24, 21:215
Haldiman Treaty of 1784 11:796
Hale, Ellery 24:763–764
Al-Halebi, Ibrahim b. Muhammed 12:750
Halévy, Élie 10:490–494
biography of 10:490
on dialectical approach 10:490–491
at École Libre des Sciences Politiques 10:490
impact as intellectual 10:492–494
intellectual contributions of 10:490–492
lectures on history of socialism by 10:494
on materialist theory of history 10:491–492
on methodism 10:491
parallels between Max Weber and 10:493
works
Histoire du socialisme européen 10:492
Halford, G.S., on relational complexity model 16:504–505
Half-split questioning rule 14:588
Half-structured interviews, for epistemological beliefs 7:888
Hall, Ann 23:315
Hall, D.G.E., works, *History of South-East Asia* 11:86–87
Hall, Edward T.
on intercultural communication 12:357, 12:359–360
works
The Silent Language 12:359
Hall, Granville Stanley 10:495–498, 19:414–415, 25:813
career 10:495

developmental psychology 10:497
five stages of human development 10:497
pedagogical ideas 10:497
work on adolescence 10:496
early life 10:495
as faculty at Johns Hopkins 10:495–496
intellectual pursuits 10:495
physiological relationships between muscles, nerves, and consciousness 10:495
pedagogical and psychological ideas 10:496
Presidency of Clark University 10:496
scholarly achievements during 10:496–497
study of education and psychology 10:495–496
Hall, John A. 9:672
Hall, S., works, *Resistance through Rituals* 23:638
Hallam, Henry, works, *Constitutional History of England* 3:61–62
Halle, M., works, *The Sound Pattern of English* 14:534
Halliday, M.A.K
notion of theme–rheme partition 12:97
Hallucinations 10:499, 10:513
auditory 10:500
mechanisms of 10:502–503
neural basis of 10:499–503
pharmacology of 10:502
psychiatric illnesses with 10:500–502
release theory of 10:502
schizophrenia 21:9–10, 21:12–14
sensory modalities 10:500
visual 10:499
Hallucinogens 6:686
Hallyu (Hanryu) 2:74, 5:458
Halo effect 17:841, 24:186
Halperin, M.H., works, *Strategy and Arms Control* 1:976
Hamada, Tomoko 17:349
Hamann, J.G., language philosophy 4:790–794
Hamann–Herder–Humboldt tradition 4:794
Hamartia 1:718
Hamburger Olympics 14:722
Hamdi v. Rumsfeld (2004) 6:698
Hamilton, Alexander, republicanism 20:479
Hamilton, Nigel 2:647–648
Hamilton Depression Rating Scale 1:781
Hamilton Pride 9:640
Hamiltonian functions 4:719, 17:253
Hamilton–Jacobi–Bellman equation 17:254–255
Hamilton's rule 13:45, 13:48–49, 15:812, 22:900
of kin selection 13:26–28, 13:31
Hammersley–Clifford theorem 22:470
Hamming distance 4:913
Hampden-Thompson, Gillian, on cultural capital 5:377, 5:379
Hampshire, Stuart 10:536
HAMS *see* Harm Reduction; Alcohol Abstinence and Management Support (HAMS)
Han Dynasty 2:66, 12:729, 13:800
Ḥanafī school 12:749
Hancock, John 1:610
Hancock, Keith 11:685
HAND *see* Human immunodeficiency virus (HIV)-associated neurocognitive disorders (HAND)
Hand actions
cortical areas involved in 5:25–26
anterior intraparietal sulcus 5:26
ventral premotor cortex 5:25–26
Hand knowledge 21:617
Hand movements 15:969
Hand technology 2:49–50
Handbooks 7:584–593
of Cambridge University Press 7:587–588
growing availability of 7:584–585
history of 7:584
of Oxford University Press publication 7:588
of Routledge 7:588
of Sage publication 7:587–588
systemization of 7:585
Handedness 2:807, 10:504
and asymmetry of body/brain 2:115, 2:121, 2:126–128
Broca's view 10:504–505
development of 10:507–508
and dichotic listening 6:359
embodied cognition and 10:510–511
evolutionary views 10:506
gender differences 10:508–512, 10:509f, 10:511f–512f
historical and geographical variation of 10:506
influential theory 10:506
and language, relationship between 10:504–514
language dominance and 10:505
and language lateralization 12:693–694
measuring 10:505
pathology 10:511–513
relationship between fetal lateralization and 10:507
'right shift gene' model 10:505–506
and sexual orientation 21:776
Handel, George Frideric, works, *Messiah* 17:908
Handicapped Children's Protection Act 1986 23:223
Handkerchief code 13:888
Hand-labeling 23:260
Handler, Joel 13:517
Handling, money 15:737–738
Handshake lemma 22:455
Hanh, Thich Nhat 2:892
Hanky code 13:888
Hankyoreh Shinmun (One Nation Newspaper) 9:393
Hannan, M., on categories 15:576
Hannerz, Ulf 2:873, 16:562
HANPP *see* Human appropriation of net primary production (HANPP)
Hansen, Alvin 20:146
Hansen, James 7:793–794
Hansen's J statistic 23:617
Hanshu 11:49
Hanson, Norwood Russell 8:576
Hansson, Albin 22:889
Hanukkah 12:872
Hanunoo language 4:232
Hanzi 20:1
HAPA *see* Health action process approach (HAPA)
Hapax legomenon 11:109
HAPC *see* Hierarchical age-period-cohort (HAPC) models
Haplodiploid mode of inheritance 13:30–31
Haplodiploidy 22:893, 22:906
Haplogroups 9:899
Haplorhines
primates, evolution of 18:907–908
Haplotype map (HapMap) 9:814
HapMap 11:309
Happiness 7:491, 10:521–525, 13:7, 22:878–879, 23:648–649, 24:401
collective, and utilitarianism 11:818
correlates of 23:650
in cultures 23:651
desirability 23:643–644
empirical research on quality of life 10:524–525
observed links with other qualities of life 10:524–525
rise of research on life satisfaction 10:524
eudaemonism 8:214–216
experience 23:644
and genetics 23:649–650
increase in 23:650–651
and longevity 23:649
and money 23:650
and motives 15:951, 15:955
norms 23:644–645
notions of good life 10:521
life-ability 10:521–523
livability of environment 10:521–522
subjective enjoyment of life 10:522–523
utility of life 10:522–523
philosophy of good life 10:521–523
and physical health 23:649
qualities of life 10:521t
and social relationships 23:650
strategies in 23:645
universal topics 10:522
classic focus on morality 10:523
historical variations 10:523–524
modern emphasis on subjective well-being 10:523–524
and utilitarian social thought 24:980–985
and welfare 25:504–506, 25:508
and welfare economics 25:497, 25:500–502
and work 10:515–520
anticipation effects 10:516–519
employment conditions 10:517

Happiness (*continued*)
individual and general unemployment 10:516
job satisfaction as driver of labor market outcomes 10:519
life satisfaction 10:515–520
no adaptation to unemployment 10:516–517
promotions 10:518
public and private sector 10:519
scaring effects of unemployment 10:517
self-employment 10:518
subjective well-being 10:515–520
unemployment and identity 10:517
unemployment and social norm 10:517
Happy person, self-efficacy as important element of 21:505–506
hapTEL 4:494
Haptic perception 24:14
Haptics
Computer Assisted Instruction (CAI) 4:494
and tactile processing 24:14
Hapu 14:505–506, 14:508–510
Harangues 15:62
Harappa 23:52–54, 23:52f
Harassment
discriminatory 6:531–532
and telephone 24:161
Haraway, Donna 22:1, 22:3
Hard collectivism, implausibility of 11:812–813
Hard constraints satisfaction 4:719, 4:722
Hard determinists 9:405
Hard embodiment 7:412
Hard positivism 13:485
Hard power 18:763
Hard problem of consciousness 4:673
Hardcore audiovisual pornography 17:220
Hardcore gonzo pornography 17:220–221
Harding, Sandra 19:677–678
Hardship, and poverty 18:741–742
Hard-to-fake signals 7:460
Hardware, experimental laboratories 8:554
Hardware synthesis, reasoning for 5:939
Hardware verification, reasoning for 5:938
Hardy–Weinberg principle 1:88, 11:290–291
Haredi community 12:873–874
Haredim 12:873–874
Harem defense polygyny *see* Female defense polygyny
Hargittai, Eszter, on cultural consumption 16:557
Hargreaves, David, works, *Social Relations in a Secondary School* 7:291
Hargreaves, Jennifer A. 23:315
on women's participation in sports 9:657
Haringey Council 20:758
Harlem Children's Zone 8:615
Harley, J.B., works, *The History of Cartography* 3:174
Harloe, Michael 12:913
Harloe's analysis 11:234
Harlow, Harry Frederick 10:526–528
on attachment 10:526
on developmental psychology 10:528
on isolation 10:526–527
on physiological psychology 10:527
on primate cognition 10:526–527
on primate research 10:527
Harm avoidance (HA) 17:794
Harm industries, businesses as 3:7
Harm principle 6:675–676
morality 13:486–487
Harm reduction 1:501–502, 22:394
approaches 6:675, 22:127
risk assessment 20:697
social justice 22:394
Harm Reduction, Alcohol Abstinence and Management Support (HAMS) 1:503–504
Harm to others, and victimless crime 25:76
Harmful activity 13:819
Harmless error 1:823
Harmonic analysis 14:633
Harmonic mean solution 18:588–589
Harmonious differentiation 18:836–837
Harmonious society 22:791
and Chinese social work 22:790–793
Harmonious versus obsessive passion 17:570–571
Harmonised European Time Use Surveys program 5:733
Harmonists 2:146
Harmonization
Bologna Process 2:741–745
census 3:294–295
concept of 2:605
of laws 10:158
process 16:229
Harmonized commodity description and coding system 5:707
Harmonized European time use surveys (HETUS) 5:746
Harmonized Indices of Consumer Prices (HICPs) 5:801
Harmony 12:243, 17:619
of elements 25:279–280
and social work practice 10:529–535
antioppressive practice 10:531
body–mind–spirit model 10:532–533
Confucianism 10:529
deep harmony 10:531, 10:533
ecocentric area 10:532–533
green social work 10:533
'othering' processes 10:530–531
personal area 10:531–532
philosophical tenets 10:529–530
relational area 10:530–531
social harmony 10:530–531, 10:533
Taoism 10:529, 10:532
therapeutic approach to social work 10:530
yin yang 10:532–533
Harmony Society 2:146
Harms
of alcohol abuse 1:518–519, 1:521
of drugs 6:677
of elders 7:327–332
risk 20:706–707
risk–captivity–harm thesis 5:246–247
Harnack, Adolf von 2:63
works, *Lehrbuch der Dogmengeschichte* 2:269
Harper, Frances E.W. 8:906
Harpy Eagle Project 18:353
Harré, Rom 10:787–788, 15:291, 18:627
works, *Varieties of Realism* 20:21
Harrer, Karl 16:257
Harries-Jones, Peter 8:291
Harriman, Philip, works, *Encyclopedia of Psychology* 7:591–592
Harrington, C.L., on media audiences 25:82–83
Harrington, James 11:413
Harrington, Michael, works, *The Other America* 18:719, 18:753
Harris, Cheryl 5:268
Harris, Cole
works
Historical Atlas of Canada 11:18
The Resettlement of British Columbia 11:18
Harris, Joseph, on division of labor 6:602–603
Harris, Louis 18:461
Harris, Marvin 1:769
Harris, William Torrey 10:740
Harris, Zellig 3:527
works, *Methods in Structural Linguistics* 3:527
Harris Matrix 3:564
Harris tweed 5:109
Harrison, Lawrence E., works, *Culture Matters* 5:617
Harrison, R., works, *Civil Code* 2:526
Harrison Narcotics Tax Act 1914 17:239
consequences 17:241
limitation on opiate and cocaine use 17:241
Harsanyi, John 9:582–583
Harsanyi payoff-configuration solution 4:878
Harsanyi–Shapley solution 4:878
Harsh parenting 2:472–474
Hart, Albert Bushnell 6:690
Hart, Herbert Lionel Adolphus 10:536–539, 12:899, 13:763, 16:323, 25:75–76
birth of 10:536
concept of law 18:26
conception of rights 10:538
contribution to legal, moral, and political philosophy 10:536–538
on distinction between habitual behavior and rule-governed behavior 10:537–538
importance for social sciences 10:538–539
on institutionalized system of social recognition 10:537
on 'internal attitude' to legal rules 10:537–538
interventions in normative and policy debates 10:537
on positivist notion of law 10:538
on primary and secondary rules 13:764
on recognition of legal rules 10:537–538
on reconstruction of legal positivism 13:764
work in analytical jurisprudence 10:537
works

The Concept of Law 2:527, 10:537, 13:763–764, 16:320–321
Hart, Keith 24:820
on commodization 5:108
on conceptualization of informality 24:818
Hartley, David 19:410
works, *Observations on Man, His Frame, His Duty, and His Expectations* 19:410
Hartshorne, Richard, works, *The Nature of Geography* 20:122–123
Hartup, Willard W., on childhood friendship 9:442
Hartz, Louis
on nationalism 16:310
works
The Liberal Tradition in America 1:597–598, 4:690
Harvard
case program 3:179
Global Health Delivery (GHD) project 3:181
Harvard Business School (HBS) 3:176
Harvard Case Histories in Experimental Science 8:577
Harvard Law Review 11:438
Harvard Law School 5:267, 13:734–735, 13:773
Harvard policy model, for strategic planning 23:517
Harvard Project Physics 3:825
Harvard tradition of small group research 22:102–103
Harvard University 24:762
social network analysis 16:529–530
Harvard University study 1:283–284
Harvest of Hate 10:871
Harvey, David 3:579–580, 12:913–917, 24:795, 24:846–847
on financial crises 3:579–580
works
The Limits to Capital 14:677
Social Justice and the City 24:943
Harvey, William 21:213
Hasan, Farhat 23:59
HASE *see* Housing Assistance Supply Experiment (HASE)
Hassard, John, on ANT 5:525
Hasse diagrams 17:526
Hastings, A.C.G., works, *Nigerian Days* 7:750–751
Hastings, Warren 23:59–60
Hatch Act 20:803, 20:815
Hate Crime Sentencing Enhancement Act 1994 10:540
Hate Crime Statistics Act 10:542
Hate crimes 10:540–545
characteristics of hate incidents 10:541
differences with other crimes 10:541
ethnicity and 10:543
examples 10:543
FBI report 10:542
incidence rate 10:543
legal distinctions 10:540–541
offenders, attitude of 10:543
police and 10:543
prevalence and distribution of 10:543
problems in data collection 10:541–542
reporting 10:542
responses 10:543–544
types of 10:542–543
use of term 10:540
victim groups 10:544
Hate Crimes Prevention Act 24:551
Hated stranger 25:796
Hateful joking 11:417
HathiTrust 14:47–48
Hauka spirits 23:283
Hau'ofa, E., on Oceania 17:158
Hausa 1:301
Hausdorff space 17:299
Haushofer, Karl 18:374
Haussmannization of Paris 10:46
Haute couture field 9:142–143
Havasupai 9:963b
Have-construction 18:632
Havel, Vaclav, works, *Kosovo and the End of the Nation-State* 3:678
Hawai'i 17:158–159, 17:161–162
Hawaiian Creole 18:138
Hawaiian cultural revival 4:185
Hawaiian sacrifice 20:861
Hawala 12:519
Hawkesworth, Mary
Encyclopedia of Government and Politics 7:591
Hawthorne, Nathaniel 2:149, 20:490
works, *The Blithedale Romance* 2:149
Hawthorne effect 8:572, 15:619, 19:738–739
Hawthorne experiments 2:915–916, 11:881–882, 16:527–528, 17:346, 22:101
perspectives of 3:5
Hayden, Cori 11:774
Hayek, Friedrich A. von 4:707, 7:926, 8:433, 10:494, 13:15, 16:167, 22:691–692, 25:273–278
biography of 25:273
on command-based organization 25:275
on constitutional conception of democracy 25:276
contributions to social sciences 25:277
on economics and Knowledge Problem 25:273–274
on equilibrium economics
formal 25:274
models of mathematical 25:274
impact and significance 25:276–277
on interplay between order of rules and order of actions 25:275
on limiting constitution 25:276
on market as a communication system, theory of 25:274
on monetary theory, capital theory, and business cycle theory 25:273
on political philosophy 25:276
on process of competitive selection 25:275–276
on spontaneous social order and cultural evolution 25:274–276
on theory of human mind 25:274
on theory of spontaneous social order 25:275
works
The Pure Theory of Capital 25:273
The Road to Serfdom 4:690, 25:276–277
The Sensory Order – An Inquiry into the Foundations of Theoretical Psychology 25:274
Haynie, Dana L., on alcohol consumption 16:558
Hazard, Geoffrey 13:516–517
Hazard function 23:471
Hazard negotiations 20:714
Hazard probabilities 23:785–786
estimated 23:788–789
Hazard rate function 8:310, 8:312
Hazard rate order, majorization/stochastic orders 14:450
Hazard rates 14:55
and mortality 15:845
random utility model 5:902
Hazard sequences 20:714
Hazard templates, and risk 20:714
Hazards 6:478, 10:698–699
characterizing 20:711
dimensions of 10:698–699
probabilities 10:699
and risk 20:706–711
and vulnerability 6:478
Hazards and disaster research 10:546–553
anthropological research in the twenty-first century 10:549–552
paradigm shift 10:547–548
Hazing
adolescence 10:556–557, 10:559
bystander intervention 10:560
definition 10:554–556
in fraternities and sororities 10:554–561
Greek life 10:554–561
initiations 10:555–558
prevention 10:559–561
rites of passage 10:556
social identity 10:556
social norms 10:560
in U.S. universities 10:556–559, 10:561b
alcohol abuse 10:557–558
court judgments 10:559
history 10:557
laws against 10:558
practices and deaths 10:558–559
prosecution and prevention 10:559
HazingPrevention.Org (HPO) 10:559–560
HBM *see* Health belief model (HBM)
HBS *see* Harvard Business School (HBS)
HbS allele 9:816
HC *see* Haitian Creole (HC)
HCA *see* Hierarchical cluster analysis (HCA)
HCG *see* Human chorionic gonadotropin (HCG)
HCI *see* Human–computer interaction (HCI)
HCL (IT company) 3:48
HCTB *see* High Casualty Terrorist Bombing Events (HCTB) dataset

HD *see* Head direction (HD) signal; Huntington's disease (HD); Hypothetico-deductive (HD) method
HDACs *see* Histone deacetylases (HDACs)
HDI *see* Human development index (HDI)
HDLs *see* High-density lipoproteins (HDLs)
11ß-HDS2 *see* 11ß-hydroxysteroid dehydrogenase type 2 (11ß-HDS2)
HE *see* Higher education (HE) market
HEA *see* Higher Education Act (HEA)
Head, Bessie, works, *When Rain Clouds Gather* 14:253
Head, Henry 16:684–685
  works, *Aphasia and Kindred Disorders of Speech* 16:684–685
Head direction cells 11:387, 16:623
  location of 16:623–624
  preferred firing direction 16:625
  relation with behavior 16:626
  responses in 3D 16:624–625
  and spatial navigation 23:163–164, 23:166–167, 23:164f
Head direction (HD) signal 16:623
  generation of 16:625–626
Head injury, and posttraumatic amnesia 18:636–638
Head measurement, electroencephalography 7:354
Head mounted display (HMD) 16:348
Head movements
  cortical areas involved in 5:27
    ventral intraparietal sulcus 5:27
Head proximity 25:653
Head shop products 6:686
Head Start program 18:720
Head-dependent relation 23:903–904
Head-Driven Phrase Structure Grammar (HPSG) 15:751, 15:754
Headedness-based accounts 16:5
Headless relative clauses 23:891
Headline indicators, in material flow analysis 14:761–762
Headnotes 5:803–805
Headship rate, family and household demography 8:728–729
Headwords, in dictionaries 5:812
Healing
  alternative and complementary *see* Complementary and alternative medicine (CAM)
  gender and 9:784
  and indigenous social work 11:804, 11:808–809
  witchcraft 25:605–608
Healing effect 8:571
Healing methods 15:118
  in indigenous cultures 15:118
Health 2:868–869, 6:208–211, 7:815–819, 11:654–661, 18:174, 20:594–595, 23:551–555
  ad normality 17:1–4
  African studies 1:280–285
  and aging 6:378–381
  anthropology of 10:571–576
    applied health 10:573
    biomedicine culture 10:573–574
    biotechnology 10:574–575
    body 10:571
    caregiving 10:575
    disability 10:573–574
    ethnography 10:571–575
    globalization 10:574–575
    health care 10:571–575
    images and diverse experiences 10:571–572
    local variations 10:572–573
    medical anthropology 10:571
    myriad social determinants 10:573
    risk 10:571–572, 10:574–575
    science and technology 10:575
  of Australian aboriginals 2:238
  and bioarchaeology 2:600–601
    reconstructing diet and inferring nutrition 2:600–601
    violence and injury 2:601
  as capital 10:617
  and cash transfers 3:216–217
  within changing environments 7:818
  charters 10:608–609
  children's concept of 10:680–681
  climate change 3:850
  conceptual components 10:611
  conceptualization and definitions 11:267
  and conditional cash transfers 4:555, 4:557–558
  consequences of bereavement 2:532–533
  contemporary philosophers in eighteenth and nineteenth century 10:608
  and coping 23:552–554
  definition of 10:567
  determinants of 7:741f
  dimensions of 1:394–395
  and disasters 6:105–107
  and drinking 6:658–662
  dynamic concept 10:612
  early life influences in adulthood 6:752–758
  and ecology 6:878–883
  economic inequalities 10:635
  education and work, lesbian *see* Lesbian
  effects, coping with chronic illness 3:552
  effects of alcohol on 1:509
  effects of media on children 15:39–40
    body image concerns 15:39–40
    weight 15:39
  and emotions 7:496
  employment, working conditions 25:698–701
  environmental 7:740–746
  etiologic and diagnostic correlates of 10:676–677
  etymology 10:607
  eudaimonia 10:607, 10:611
  and explanatory style 8:598–599
  future research 11:659–660
  gender and 9:784
  genetics, and social justice 9:971–972
  geography of 10:628–632
  global political economy and 10:633–634
  global trends since the 1940s 10:633
  and globalization 10:198
  historical definitions 10:607
    physicians and philosophers 10:607
  historical development 11:655
  household economics 11:224–225
  and human development 11:267–269
  human rights abuses 10:636–637
  inequalities 10:635, 22:286, 22:289
  and internal migration 12:440
  issues of lesbians 13:896–897
  and justice 12:928
  of Latino gays and bisexuals 9:630
  and LGBT aging 13:866–867
    effect of discrimination 13:866
    mental health 13:866
    physical health 13:866
  in low-income countries 10:633–639
  and marital quality 14:512
  and marital status 15:853–855
  and mass imprisonment 14:696–699
  measuring mental imagery abilities in 15:225
  medical geography *see* Medical geography
  of men 15:166–171
  mental representations in different cultures 10:567–570
    implications 10:568–569
  methodological issues 11:655–656
  migration and 15:407–410
  military anthropology 15:486
  of minority ethnic elders 15:441–442
  and mortality 15:845
  mortality rates 11:656–657
  mutual aid self-help (MASH) groups 10:688–690
  nonhierarchical social determinants of 22:276–277
  and nutrition
    family and schooling 8:754–755
    programs in schools 8:755–756
  in Oceania 17:163
  in old age 1:394–397
  online support groups 10:686–688
  operationalization of health by WHO 10:609–610
    classification systems 10:609–610
    quality of life and well-being and health-related quality of life 10:610
  outcomes of family planning among women and children 8:789–790
  and participatory action research 17:537
  and personality 17:792–796
  and personality traits 17:923
  perspectives of control beliefs 4:819
  and physical activity 18:87–90
  and physical changes 15:372
  physical health and health behaviors 11:657–658
  policy issues 10:636–637
  and population aging 18:541–543
  poverty 10:634–635
  race and ethnicity 19:812
  -related consequences of work–family conflict 8:760

and religion 20:269–272
and religious behavior 20:392–393
research, lesbians as topic of 15:114–115
resilience 20:558
and retirement 20:598–601, 20:605
right to *see* Right to health
in rural geography 20:799
salutogenesis 10:612
and self-efficacy 21:505–506, 21:508–514
self-esteem 21:523
from sixteenth century 10:608
view of physicians 10:608
and sleep 22:59–62
smoking and 22:125–131
and social behavior 22:283
and social capital 22:174, 22:178
social comparison 22:214–215
social determinants of 10:634–635
social exclusion, social deprivation and 22:285–290
social inequalities in 17:732–733, 22:275–276
social integration, social networks and 22:365–370
and social problems 12:545–546
social psychology of 15:102–104
and social relationships 15:853
social relationships pathways influencing 22:369
and socioeconomic status 15:852, 22:931–936
and startle reflex 23:331–337
and stress 23:551–552
and successful aging 23:667–671
twentieth and twenty-first century
concepts outside WHO in 10:611–612
contemporary philosophers of 10:610–611
criticism of WHO definition of health 10:612
twentieth century 10:608–609
constitution of WHO 10:608–609
idealistic intentions 10:609
and urban life 24:830–835
valuing 10:617–618
volunteerism 25:271
water and 25:428–436
and wealth 10:633–635
inequalities in 10:634–635
workplace safety and 25:727–732
Health access, universal 8:34–35
Health action process approach (HAPA) 10:578–579, 10:712, 10:714, 21:512
Health actions
disaster cycle of 16:317f
for natural disasters 16:317–318
Health activism, and biocitizenship 2:610–611
Health actors, in collecting and sharing of health data 5:786–787
Health and Retirement Study (HRS) 2:659–660, 5:698, 5:744, 5:747, 6:379, 20:602
Health and Safety at Work Act 1974 (UK) 17:144
Health and safety conditions in the workplace 25:698
Health anxiety, explanatory style 8:598–599
Health Assessment Questionnaire – Disability Index 20:656
Health behavior change
evolutionary medicine 8:458–459
personalized medicine 17:929
Health behaviors 10:582–587, 10:710–711, 10:713, 11:657–658
alcohol use 10:583–584
assessment of 10:588–593
behavior observation techniques 10:591
behavior record methods 10:590–591
ecological momentary assessment 10:590, 10:592
frequency measures 10:588–589
interviews 10:590
as multidimensional construct 10:588
nonreactive measures 10:591–592
physiological measures 10:591–592
quantity measures 10:588–589
questionnaires 10:588–590
reporting biases 10:589–590
self-report measures 10:588–591
attitudes 10:585–586
definition 10:582
diet 10:583
dimensions of 10:588t
distribution and prevalence of 10:584–586
health belief model 10:584–585
integrated models 10:585–586
social cognitive theory 10:585
theory of planned behavior 10:585
and emotions 7:498
health cognitions 10:584
health psychology 10:674–675, 10:678
health screening 10:583
intentions 10:585–586
life course epidemiology 14:71–72
measures, and infectious diseases 12:34
and obesity 17:78
oral 17:280–281
and personality 17:794
and physical activity 18:92–94, 18:96
physical activity/exercise 10:583
psychosocial theories of 10:577–581
public health 19:545, 19:548–549
relationship to sociodemographic factors 10:584
risk perceptions on 10:700
self-efficacy 10:585–586
sexual behaviors 10:583
smoking 10:582–583
and sun exposure 23:699
Health belief model (HBM) 1:121–122, 10:577, 10:584–585, 18:93
Health care 10:571–575, 22:937
assess to 10:637
and biocitizenship 2:610–611
changing context of 16:907
for children and infants 12:15, 12:17–19
in China 2:67
delivery, geriatrics 10:105–106
and doctor–patient interaction 6:613, 6:615–619
East Asian studies 6:786–787
expenditure and outcomes 10:637
and gender 9:694–697
and health literacy 10:653–655
Japanese studies 12:797
law and technology 13:525–526
legal issues, in United States 10:594–597
legal issues of
balancing access with coverage 10:596–597
cost containment practices, challenges to 10:595–596
historical view 10:594–595
liability doctrine 10:595
litigation context 10:595
medical liability for negligence 10:597
modern health care environment 10:595
physician autonomy 10:596
physician–patient relationship 10:596
public insurance mechanisms 10:595
social policy conflicts 10:597
in United States 10:594
and medical profession 15:93–97
organization and provision 15:100–102
changing role of physician 15:101–102
intractable cost crisis and managed care 15:102
technological imperative and public expectations 15:101
psychiatry 22:974–975, 22:977–978
rationing 10:603–606
concerns about 10:603–604
ethical and political issues 10:605
explicit *versus* implicit 10:603
policies and practices 10:604–605
priority setting in modern nations 10:605–606
of specific resources 10:604
role in explaining SES disparities in health status 22:940
and Skocpol, Theda 22:31, 22:33–34
social psychology of 15:102–104
clinician–patient interaction 15:103
disability and quality of life 15:103–104
neighborhood influences 15:103
social networks and social influence 15:103
stress and coping process 15:103
and social work 14:667–668
and socioeconomic status 22:937–941
systems 10:637
traditional healers 10:637
utilization 22:937–938, 22:938t
financial/affordability explanations 22:938–939
health care system explanations 22:939–940
knowledge, attitudinal, and cultural explanations 22:939
utilization in socioeconomic status (SES) 22:937–938
women in 25:622

Health Care Act 2005 7:599
Health care costs 10:605–606
Health care financing, psychiatric ethics and 8:66–67
Health care markets 10:598–602
  cost control 10:598–601
  empirical evidence 10:599–600
    effect of competition on quality 10:600
    structure 10:600
  failure, problems of 10:598
  health care research 10:599–601
  managed competition 10:598–599
  market-determined prices 10:600
  origins 10:598
  private insurance 10:599–600
  proliferation and variations 10:599
  public insurance 10:598–599
  regulated prices 10:600
  research 10:599–601
Health care occupations 15:96
Health care policy 15:94–95
Health care practitioners 5:477
Health care providers
  gender and 9:696–697
Health care regimes, adherence of patients to 17:598–602
Health care regulation 20:170
Health care services
  accessibility of 16:907–908
  exclusion from 22:287
Health care systems 10:666–668, 22:939–940
  patient behavior in 10:677–678
  personalized medicine implications for 17:927–928
    complexity of delivery of health care by nontraditional means 17:928
    consequences of predicting outcomes 17:928
    electronic records and data management 17:927–928
    need new systems for introducing innovations 17:928
    risk of increased costs 17:927
Health circles 17:151
Health cognitions 10:584
Health crisis, in central and eastern Europe 1:161–165, 1:162f, 1:164f–165f
Health data 5:781
Health data governance 5:781
Health differences
  social relations pertaining to 22:278
  social times in explaining 22:278
Health disparities 19:815
  and physical activity 18:96
  psychosocial causes 19:816
  and roles of health social work 10:718–719
  and social justice 9:971–972
Health economics 10:614–619
  adverse selection 10:615–616
  agency 10:615
  communicable disease 10:617
  competition 10:615–616
  cost-utility analysis 10:617
  cream skimming 10:615–616
  drug patents 10:618
  health as capital 10:617
  health maintenance organizations 10:616
  health policy and equity 10:618–619
  industrial organization of health insurance and health services industries 10:616–617
  information asymmetry 10:615
  insurance protection *versus* moral hazard effects 10:614–615
  managed care 10:615–616
  medical savings accounts 10:614–615
  noncommunicable disease 10:617
  pharmaceuticals and medical technology 10:618
  private health insurance 10:615–616
  quality-adjusted life years 10:617–618
  selective contracting 10:615–616
  state-contingent markets 10:614, 10:616
  supplier-induced demand 10:615
  valuing health 10:617–618
Health education 10:620–627, 21:825
  for CVD prevention 10:648–649
  early international diffusion 10:649
  first decade 10:648–649
  projects begun in 1980s and 1990s 10:649
  see also Sexuality education
Health education, and health promotion
  definitions 10:620
  determinants 10:623–624
    automatic 10:623–624
    ecological 10:623
    social cognitive 10:623
  Green's PRECEDE-PROCEED model 10:621
  intervention mapping 10:622–623
    long live love 10:622–623
  perspectives 10:620–621
  quality of planning 10:621–622
    reviews and meta-analyses 10:622
    theory-based interventions 10:621–622
  using theories in 10:624–626
    changing environmental conditions 10:625–626
    e-health 10:624–625
    learning, goals, and self-management 10:624
    stage models 10:624
Health expectancies 16:113–114
Health financing, comparative health systems 4:401
Health geography 1:407–409, 10:628, 15:86, 15:90
  air quality 10:630
  caring 10:631
  composition 10:630
  context 10:630
  HIV 10:629–630
  home 10:631
  inequalities 10:629–630
  place in 10:628–629
  self 10:631
  space in 10:628–629
Health goals, health psychology 10:674–677
Health information 5:781
Health information systems 5:781–788
  dimensions of 5:782–784
  dynamics between national and global health actors in collecting and sharing of health data 5:786–787
  health data governance 5:781
  health statistics 5:781–782
    historical origins of 5:782
    linking health and social statistics 5:782
    moving from mortality to disease burden 5:782
    sharing and classifying 5:782
    vital statistics 5:782
  information generated in other areas and sectors 5:783
  institutions providing health statistics 5:785
    analysis and interpretation 5:785
    collection 5:785
    ensuring evidence-informed policy-making 5:785
  international comparisons 5:783–784, 5:784t
  kind of information 5:784
  new developments and challenges 5:787
  public health 5:781–784
  purpose of 5:783
  requirements for countries 5:785–786
    analytical capacity 5:786
    data collection systems 5:786
    information technology 5:786
    institutional 5:785
    legal 5:785–786
    resources 5:786
    technical 5:786
  terminology 5:781–782
  use at international level 5:783
  World Health Organization and 5:781–787
Health insurance
  economic and risk aspects 10:640–645
    around the world 10:642–643
    asymmetric information 10:641–643
    Beveridge model 10:643–644, 10:644t
    Bismarck model 10:643–644, 10:644t
    copayment 10:640
    coverage 10:596–597, 10:640–644
    cream skimming 10:643
    deductibles 10:640
    market 10:640–643
    network providers 10:640
    preexisting conditions and risk selection 10:641
    premium 10:640–643
    supply side 10:641–642
  industrial organization of 10:616–617
  and managed care 14:462
  private 10:615–616
  and retirement 20:605
Health insurance experiment 8:532, 22:299
Health Insurance Portability and Accountability Act (HIPAA) 4:253, 8:46, 8:49, 23:754–755
Health interventions, community-based 10:646–652
Health literacy 10:653–656

communication 10:655
conceptualization 10:653–654
decision-making 10:653–655
definition 10:653
disease prevention 10:653, 10:655
field, characteristics of 10:655
health care 10:653–655
health promotion 10:653, 10:655
intervention 10:654
investment in 10:655
measurement 10:654
and patient adherence 17:599
people-centered care 10:653
as public health challenge 10:654
social determinants for health 10:653–654
see also Health education
Health Literacy Questionnaire 10:654
Health Maintenance Organization Act 1973 10:594–595
Health maintenance organizations (HMOs) 10:595–597, 10:599, 10:604–605, 10:616, 14:463–464, 15:101–102, 17:377
Health management task 6:709
Health outcomes 20:271
psychoneuroimmunology to improve 19:447–448
Health outcomes, assessment of 10:657–662
disease-specific measures 10:660–661
historical perspective 10:657–658
quality of life measurement 10:658–660
cost-effectiveness evaluations 10:660
decision theory approaches 10:659–660
EQ-5D 10:659
Health Utilities Index 10:659
Nottingham Health Profile 10:658–659
psychometric methods 10:658–659
Quality of Well-being scale 10:660
SF-36 10:658
Health policy 10:663–668
actors 10:663–664
comparison 10:663–666, 10:668
data 10:663
fee-for-service payment 10:667–668
financing 10:664
governance 10:664
and gross domestic product (GDP) 10:666–667
healthcare systems 10:666–668
National Health Service 10:663
politics 10:664, 10:667
private health insurance 10:663–664, 10:666–667
provision 10:664–665
regulation 10:665–666
self-regulation 10:666
social health insurance 10:663–667
state regulation 10:665–666
supply states 10:664
Health poor 18:741
Health problems
of alcoholic abuse 1:518–519
among poor 15:739
Health promotions 6:208, 10:620–627, 10:719, 18:885–886, 21:826, 21:830–831
in behavioral medicine 2:468
in children 10:682
concepts 10:669
effective, high-leverage strategies for 6:881
and health literacy 10:653, 10:655
and physical activity 18:91–99
programs 10:669–670
school-based 10:670–672
programs, targeting vulnerable subgroups in 6:881
Health psychology 2:496, 10:674–678
adherence 10:677–678
adjustment 10:675–676
in childhood 10:679
coping 10:676–677
development 3:874
etiologic and diagnostic correlates of health and illness 10:676–677
health behavior 10:674–675, 10:678
health goals 10:674–677
implementation intentions 10:675
intention-behavior gap 10:674–675
in oral care 17:280
of pain 17:451–455
patient behavior in health care system 10:677–678
patient-provider interaction 10:677–678
placebo effect 10:677–678
prevention and treatment of chronic illness 10:675–676
promotion and maintenance of health 10:674–675
risk perception 10:675–676
self-regulation 10:674
Health reforms, Fisher's 9:219–220
Health research
qualitative 10:692
and social network analysis (SNA) 22:461, 22:463–464, 22:467
Health risk appraisal 10:698–701
Health risk perception 10:698, 10:702–709
multidimensionality 10:703
intuition and feelings 10:703
psychometric paradigm 10:703
optimistic bias 10:704–705, 10:705t
personal risk perceptions 10:703–706
absolute 10:704
comparative 10:704–706
types 10:705t
pessimistic bias 10:704–705, 10:705t
reactions toward personalized risk feedback 10:706–707
cue-adaptive reasoning perspective 10:706–707
motivated reasoning perspective 10:706
quantity of processing perspective 10:706
risk assessment and numbers of risk 10:702–703
risk information 10:702–703, 10:706–708
studying 10:702
Health sciences, lifestyle in 14:114–115
Health selection 11:654
theoretical explanations 11:658–659
Health self-regulation 10:710–715
constructs and models 10:710–712
perceived self-efficacy 10:710–711
self-control 10:711–712
process-oriented model 10:712–714
motivation 10:712–713
volition 10:713–714
Health services, cancer screening 3:79–81
Health services industries, organization of 10:616–617
Health social work 10:716–724
areas of practice 10:719–720
medical social work 10:719
public health 10:719–720
public health social worker and disaster 10:720
challenges 10:721–723
in ethical issues 10:722
in health policy 10:722
in health social work practice 10:721–722
core competencies 10:720–721
development and transformation of 10:716–719
early twentieth century 10:716–718
environmental challenges 10:718
global scene 10:718
health disparity 10:718–719
nineteenth century 10:716–718
occupational social work 10:718
opportunities 10:721–723
in education 10:722–723
in knowledge development 10:722–723
overview of 10:716
definition 10:716
typology of 10:716t
Health statistics 5:781–782
historical origins of 5:782
institutions providing 5:785
linking health and social statistics 5:782
moving from mortality to disease burden 5:782
sharing and classifying 5:782
vital statistics 5:782
Health surveys 5:784, 10:725–730
definition 10:725
determinants of health 10:728
functioning 10:727–728
general design issues 10:725–726
health care utilization 10:728–729
health status 10:726
longitudinal designs 10:729
medical conditions 10:726–727
methodological research and analytic techniques 10:729
operationalization of health 10:725
privacy and confidentiality 10:729
summary measures 10:728
technological advances 10:729
Health system, comparative *see* Comparative health systems (CHS)
Health transition 10:571–572, 15:857–862
versus epidemiological transition 15:859
in more developed countries 15:859–861

Health transition (*continued*)
theoretical concept 15:857
Health Utilities Index (HUI) 10:659
Health-enhancing resources 19:817
Health-related quality of life (HRQL) 9:606–607, 9:609, 10:610, 25:108
and childhood cancer 3:462
Health-related surveys 5:848
Healthy adult aging, source monitoring in 23:33–34
Healthy aging, and retroactive interference 20:615
Healthy aging across the life course (HALCyon) research 14:73
Healthy attachment 5:481
Healthy Cities project 24:832–833
Healthy city 24:832–833
qualities of 24:833t
Healthy gendered foodscapes 10:80
Healthy immigrant phenomenon 19:815
Healthy life expectancy 4:463–464, 4:467, 17:179
Healthy life years 4:464
Healthy migrant effect 12:436
Healthy People 2020 program 15:110
Healthy-soul 20:354
Healy, Kieran, on commodification 5:538
Hearing
sign language disorders 21:934–941
Stevens on 23:455
Hearing impairments, working with people with 4:289
Hearst, William Randolph 15:12
Heart, axis formation and positioning of 2:116–118
Heart disease
and gender 9:704–707
psychoneuroimmunology 19:447
religion and 20:266
and stress 23:551–552
Heart rate
and cardiac defense 23:539–544, 23:541f–543f
and neuromarketing 16:700–701
Heart rate variability (HRV) 17:138
Heartbeat detection tasks 12:615
Heartbeat-evoked brain potentials (HEPs) 12:615–616
Heat energy 7:626
Heath, Melanie, on epistemic cultures 7:877
Heathershaw, J. 3:319–320
Heaton-Nicholls, Ioma 8:289
Heautoscopy 10:499
Heaven's Gate (religious group) 16:807
Heaviness 18:716
Heavy episodic drinking 1:519–520
Heavy oils 17:167
Hebb, Donald Olding 2:841, 10:731–735, 18:56, 19:415
on behavioral neuroscience 10:733
contributions to psychological theory 10:733–735
on cortex 10:734
early life and education of 10:731–732
on electronic circuitry 10:733
on learning changes at synapses 10:734
in McGill University 10:733
in Montreal Neurological Institute 10:732
on 'nature–nurture' debate 10:734
on neural model 10:734
on Pavlovian phase 10:733–734
on primate research with Lashley 10:732–733
on short-term expectancy and attention 10:734
works
Organization of Behavior 2:835
Hebbian Learning models 18:62
Hebb's rule 14:358–359
Hebdige, D., works, *Subculture: The Meaning of Style* 23:638
Heberer, Gerhard, works, *Anthropology A to Z* 7:589–590
Hebraic nationalism 16:295
Hebraism 11:67, 11:69–71
Hebrew 19:709
Hebrew Bible 12:876
Hebrew scriptures 3:233
and the Code of Hammurabi (1700 BC) 2:672, 22:393
Hebrew University 11:69
Hecht, Tobias 1:743
Hechter, Michael, works, *Containing Nationalism* 16:309
Heckman, James 17:783
Heckscher, Eli 7:124–125
Heckscher–Ohlin model of international trade 7:124–125
Hedegard, D., on racial identities 5:560
Hedonic adaptation
and long-term heterosexual relationships 20:916
in subjective well-being 23:650
Hedonic approach to well-being 15:187–188
Hedonic contingency hypothesis 2:171
Hedonic enjoyment 13:833–834
Hedonic regression 5:51
cost–benefit analysis 5:51
Hedonic school, brain stimulation reward 2:843
Hedonic well-being 21:489
Hedonism 10:523
ethical 2:326
motivation 2:326–327, 2:329–330
psychological 2:326
and values 25:28–29
and welfare 25:505–507
Hedonistic-oriented moral reasoning 7:317–319, 7:316t
Hegel, Denton Snider 10:740
Hegel, Georg Wilhelm Friedrich 1:636, 3:709, 3:774, 4:338, 8:779, 10:736–741, 11:175–176, 14:642–643, 16:834, 19:314, 19:411, 20:45, 21:890–891, 24:745
on alienation 1:551–552
analysis of civil society 10:738
biography of 10:736
on civil society 3:709–710, 10:738–740
on democracy 10:739–740
on dialectic 11:72–73
on freedom 10:737–741
and Habermas 10:740–741
on human rights 10:739
and Kant 10:736–738
and Marx 10:739–740
and Mead 10:740
on pedagogy 10:740
philosophical aspects 10:736–739
interests of young Hegel 10:736–737
phenomenology of spirit 10:737
philosophy of nature 10:738
philosophy of spirit 10:738–739
philosophy of the Real 10:738–739
science of logic 10:737–738
on psychology 10:738, 10:740
and Rousseau 10:736
on spirit 9:667–668
St Louis Hegelians 10:740
on state 10:738–740
and Stein 10:739
on struggle of recognition 25:409
on *Vernunft* and *Verstand* 10:785
on welfare 10:738–740
on work/labor 25:675
works
Doctrine of the Absolute Spirit 10:739
Doctrine of the Subjective Spirit 10:738
Elements on the Philosophy of Right 10:736
Encyclopedia of Philosophical Science 10:736
Phenomenology of Spirit 1:551–552, 3:58
Sittlichkeit 10:739
and Wundt 10:740
Hegel Craze 4:319
Hegelian dialectics 5:263
Hegelian roots 11:176–177
Hegelians, St Louis 10:740
Hegemonic culture 10:748, 15:716
Hegemonic femininity 8:893–897
Hegemonic masculinity 9:733–734, 14:679–681, 14:684
Hegemonic power of law 13:471
Hegemonic stability theory 24:486, 25:401–402
and political economy 1:940
Hegemonic sustainability discourse 23:812–813
Hegemonic view, of social work education 22:776–777
Hegemony 4:380–381, 5:251, 10:93–94, 10:747, 11:765, 13:471–472, 14:652, 14:694, 15:716, 17:425, 18:9, 18:761–762
anthropology of 10:747–749
civil society 10:747
class consciousness 10:748–749
cultural resistance 10:748
domination and subordination 10:748–749
future 10:749
Marxism 10:747–749
neo-Marxism 10:747–749
revisionists 10:747–748

state-colonial policies 10:748–749
audience, analysis of 10:745–746
and common sense 4:2–4
concept of 10:742–744
and creolization 5:117–118, 5:120
and cultural resistance 10:743
and cultural theories 10:744
within Europe 3:729
Gramsci's concept of 10:744, 10:746
journalism, analysis of 10:744–745
and political power 18:761–762
and power 18:766
and realism 20:27–28
and social change 10:743
social science organizations 22:648–651, 22:653–654
and sport 23:315
and state 23:360
subculture and popular culture studies 10:745
texts, analysis of 10:745
Hegghammer, Thomas 24:234
Heiberg, Astrid 25:625
Heidegger, Martin 1:539–540, 3:776, 4:793, 19:314
on language 14:199
on Nietzsche 15:291
works
On the Way to Language 4:793
The Origin of the Work of Art 4:793
Heidelberg Institute of International Conflict Research (HIIK) 5:836
Heidelbergensis 16:739
Heider, Franz 14:383
Heider, Fritz 2:201, 10:750–754
on aggression and prejudice 10:753
analysis of social perception 2:202
on attribution theory 10:752
on balance theory 10:751–752
on causal analysis 2:202
on clinical psychology 10:753
on cognitive psychology and psycholinguistics 10:753–754
on coping 10:753
on impersonal causality 2:202
interpretation of 10:752
on motivation and emotion 10:753
notion of invariances 2:202
on person perception 2:201–202
on personal causality 2:202
on political psychology 10:754
research method of 2:201
on unit relations 23:606–607
works
Psychology of Interpersonal Relations (*PIR*) 10:751
Height
as erotic capital 8:6
and sexual orientation 21:776
Heikkinen, M., on arts in Western Europe 5:461
Heilsgeschichte 11:32
Heim's File Change Semantics 15:754
Heinich, Nathalie 5:417
Heinroth, Johann Christian August 19:314
Heinsohn, G.
works
Genocide Prevention Now 10:7
Raphael- Lemkin-Institut für Xenophobie- und Genozidforschung 10:7
Heintz, Bettina, on epistemic cultures 7:876
HEIs *see* Higher education institutions (HEIs)
Heise, David R.J., semantic differential 21:568, 21:570
Heisenberg, Werner 20:287–288
Hekma, G., on gay athletes 9:658
Heliography 20:420–421
Hellebore 19:463
Hellenism 16:169
personality psychology in 17:907
Hellenistic Age 1:699
Hellenistic Empires 2:919
Hellenistic philosophy 25:599
Hellenists 3:548
Hellenized Christianity 20:285
Heller, Agnes 12:137
Hellmann, Siegmund 25:462
Helmholtz, Hermann Ludwig Ferdinand von 10:755–758, 14:810
achievements of 10:758
on conservation of energy and socioeconomic thought 10:756
on corollary discharge 15:958
efforts in science 10:756
on empiricist theory of perception 10:757
life of 10:755
German academic life 10:755
marriage of 10:755
on measuring time intervals in thought processes 10:756
on nativist–empiricist controversy 10:757
on nerve–impulse measurements 10:756
on ophthalmology 10:755
Handbook 10:757–758
On the Sensations of Tone 10:757–758
Scientific Papers 10:756
on physiological acoustics and optics 10:756–757
on resonance theory of hearing 10:757
scientific contributions of 10:756–757
on scientific foundations and aspects of music 10:756
work on psychology 10:757–758
Helminthic therapy 8:457, 8:457f
Helminths and Crohn's disease, evolutionary medicine for 8:457
Help giving 10:760, 10:763–764
Help seeking 10:763
mental illness 19:310–311
in military/disaster psychiatry 15:499
HelpAge India 7:327–328
Helper-therapy principle 19:381
Helping 19:217
and agreeableness 1:466
behavior, and intergroup relations 19:218–219
power relations 19:218–219
prosocial behavior 19:233–235, 19:238–239, 19:241
and prosocial behavior 19:228–231
Helping practices, within indigenous cultures 11:805–806
Helplessness, control beliefs 4:820
Helsinki Declaration 11:376, 13:440–441, 20:511
Helvering v. Davis (1937) 18:735
Hematopoietic stem cells 23:443–444
Hemeroby 13:240–241
Hemianesthesia 17:507–508
Hemianopic alexia 6:734
Hemispatial neglect 10:766–772, 17:509
assessment and diagnosis 10:767
Behavioral Inattention Test 10:767
cognitive rehabilitation 10:769
dorsal and ventral systems 10:767
future directions 10:770
high-resolution structural imaging 10:768
implicit visual information processing 10:768–769
lesion 10:767–768
lesion mapping in patients 10:768
sites 10:768
neuroimaging 10:766, 10:768, 10:770
orientation bias 10:769
perspectives on rehabilitation 10:769–770
phenomenon 10:766–767
right hemisphere 10:766, 10:768
spatial orientation 10:768
spatial representation 10:766–767, 10:770
stroke 10:766
Sunnybrook Neglect Assessment Procedure 10:767
term 10:766–767
underlying mechanisms 10:767
virtual reality lateralized attention test 10:767
visual exploration therapy 10:769
Hemispherectomy 13:405, 16:686
Hemispheres, of cerebral cortex 3:334–338
Hemispheric asymmetries 1:432, 2:806–807, 12:417–418, 13:398
and information processing 1:432
Hemispheric dominance 12:692, 12:694
Hemispheric lateralization 12:418–419
of language in children and adults 10:505–506
Hemispheric specialization theory 2:650, 2:653–654, 16:686
Hemispheric speech representation, patterns of 12:694
Hemodynamic response function (HRF) 9:526–527, 16:389
Hemoglobin 16:387
DeoxyHb 16:389, 16:392–393
OxyHb 16:389, 16:392–393
Hempel, Carl Gustav 1:902, 10:773–775, 11:12, 21:302
biography 10:773
of causal mechanisms 14:874
contributions 10:773–775
concept formation in empirical science 10:774–775
methodological unity of empirical science 10:774

Hempel, Carl Gustav (*continued*)
scientific explanations 10:773–774
covering-law model 1:892
on historical explanation 11:12–14
hypothetico-deductive (H-D) model 1:892
influence, impact, and current significance 10:775
Hempel–Oppenheim model 10:773
Hempel's dilemma 16:327–328
Henan eugenics law 1992 8:234–235
Henderson, H.M. 25:465
Hendry, David 7:9–10
Hengeveld, K. 17:556–558
Henisz, Witold, on world society theory 5:539
Hennemann size principle 3:779
Hennion, Antoine, on ANT 5:525
Henry, Joseph 24:153
Henry, Louis 6:149, 10:776–777, 11:5
biography of 10:776
Henry Molaison (HM) 16:847
Henry Sweet Society for the History of Linguistic Ideas (HSS) 14:218
Henry VIII, King 3:32
Hensen's node 2:121
HEP *see* High-energy physics (HEP)
Hepburn, Audrey 16:66
HEPs *see* Heartbeat-evoked brain potentials (HEPs)
Heraclitus 19:406
Herbal medicine 10:562
Herbart, Johann Friedrich 14:811
works
Psychology as a Science Based on Experience, Metaphysics, and Mathematics 19:411
Textbook in Psychology 19:411
Herberstein, Sigismund von, on Russian sexual life 21:740
Herbert, E. 20:323
Herbert, James 25:211
Herculaneum 3:761
Herder, G.F. 11:94–95
Herder, Johann Gottfried von 4:338, 4:790–794, 11:12–13, 12:741–742, 19:314, 19:676, 22:887–888, 23:199
on ethnos 15:290
on folk culture 1:765–766
on Islamic fundamentalism 12:741–742
on language and ethnicity 13:252–253
Herderian environmentalism 23:199
Hereditary disorders 9:998
Hereditary genius 12:176
Heredity 8:234–235, 9:932
changing views 8:254
and environment 8:221–223
epigenetic inheritance 7:832–833, 7:836–837
and instinct 12:176, 12:180
measurement of biological 23:425
Mendelian principles of 8:218
model of 8:218
scientific principles of 8:218
Heretical manifestations 3:548
Herfsttij der Middeleuwen 11:76
Heritability 2:452, 9:238, 9:821, 9:826, 9:951–954, 9:985, 9:1009–1011, 12:298, 17:889
of acquired characteristics 8:235
of behavior 12:175–176
concept of 8:715–716
developmental behavioral genetics and education 6:321
by environment conceptual models 9:829f
and environment interactions 9:829
of epigenetic modifications 24:688–689
of personality traits 17:899–903
and sexual orientation 21:779–780
and twin studies 24:704–705
Heritability controversy
genomics phase of 8:716–717
pregenomics phase of 8:716
Heritability estimation 8:716
Heritability ratio, and genetic studies of behavior 9:936
Heritability studies, and intelligence 12:305–306
Heritable genomic variation 8:715
Heritage 16:145–146
art 2:17
culture *see* Cultural heritage
memory 1:882–883
politics of 9:283–284
preservation, and cultural policy 5:462–463
Heritage Foundation in Washington DC 24:295
Heritage iconoclasm 5:416–417
Heritage language 2:581, 2:583–584
Heritage management studies 10:53
Heritage-baiting 5:415
Heritagization 5:414–417
Hermeneutic circle 22:691
Hermeneutic explanation theories 11:12–14
Hermeneutic phenomenology 19:491–492
Hermeneutic–dialectic tradition 1:473
Hermeneuticians 22:997
Hermeneuticists 19:677–678
Hermeneutics, 1:894–896, 2:508–509, 9:357–358, 9:360, 9:362, 9:364, 11:133–134, 20:345
and anthropology 1:747, 1:769
anthropology of 1:747
antinaturalism 10:778–779, 10:781–782
Boltanski on 2:752–753
concepts and tasks 10:778–779
critical 10:787
and cultural sociology 5:500–501
definition 10:785, 10:788
Dilthey on 6:426–428
general hermeneutics 10:778–781
historicist 10:787
history of 10:780–781, 10:785–789
early history 10:780–781
prehistory 10:780
and interpretation 12:641–642
interpretive social science 10:778–779
legal reasoning 13:776–777, 13:779–783
methodological dualism 10:781–782
methodological monism 10:781–782
methodological pluralism 10:783
methods and methodology 10:779
phenomenology 19:491–492
philosophical 10:787
philosophical hermeneutics 10:778
Plato on 18:209
critique of writing 18:209
dialogues 18:209
and positivism 18:626–627
and postmodernism 19:491–492
principles of benevolent interpretation 10:781
and psychotherapeutic approaches 19:492–493
and psychotherapy 19:491–494
recent developments 10:781–783
methods controversies 10:781–782
separation of disciplines 10:782–783
and reflexivity 20:84
Ricoeur on 20:666–668
spirit possession 23:283
systematization 10:785
Taylor on 24:39–41
understanding and interpretation 10:779–780
Verstehen on 10:778–779, 10:782
and wisdom 25:601
Hero myths 16:172, 24:650–651
Hero system 7:595
Herod, Andrew 13:163
Herodotus 7:144–145, 10:162, 11:35
and Greek historiography 11:35–36
history 1:765
history of the Persian Wars 18:380
Heroic entrepreneurs 24:821
Heroic suicide 15:511
Heroin 6:675
Heroin sniffers 17:240
Heroism, celebrity *versus* 3:275
Herophilus, on neurolinguistics 16:680–681
Herpes virus infections, recurrence 12:35
Herrenfall 9:111–112
Herring, Pendleton, works, *Group Representation Before Congress* 18:233
Herrnstein, R.J. 8:130–131
matching law 20:209
works
The Bell Curve 12:311
Herrschaft 3:353–355
legitimate 3:354
traditional 3:354
see also Authority
Herrschaftstypen 2:271
Hersen, Michel, works, *Handbook of Social Development* 7:590
Herskovits, Melville 1:639, 1:755, 2:712, 14:837
Hertz, Robert, analysis of first and second funerals 5:871–873
Hertzka, Theodor, works, *Freeland: A Social Anticipation* 14:256
Herzl, Theodore 2:147
works, *The Jewish State* 2:147

Herzog, Herta, works, *Invasion from Mars* 7:563
Herzog, Otthein 15:16
hESC *see* Human embryonic stem cell (hESC)
Heschl's gyrus 13:370, 13:404
Hesiod 14:252
Hesmondhalgh, David
on cultural industries 25:83
on cultural production 5:469
Hess, Moses 14:644
Heterodox economics 1:573, 1:577, 7:94–96
and Cold War 4:123
Heterogeneity 4:92
capture–recapture methods 18:604–606
demographic microsimulation 15:344–346
economics 9:951
effect on reliability, classical true score model 19:425
event history analysis 8:308
in event occurrence 8:297, 8:299–300
and globalization 10:235
of irrelevancies 3:70, 24:640
in LAC educational system 7:178
landscape 13:239, 13:243
meta-analysis 23:911
social representations 20:445–446
of treatment effects in microdata, instrumental variables 12:208
of value of life 25:18–19
public expenditures and social burden 25:19
values for children, working-age adults, and seniors 25:18–19
values for various types of individuals 25:18
values of others' lives – altruism 25:19
in world culture 10:228–230
Heterogeneity evaluation
meta-analysis 15:274
Heterogeneous capital goods 3:92
Heterogeneous diglossia 6:418
Heterogeneous fertility decline in advanced economies 9:776
Heterogeneous multimodule systems 2:38
Heterogeneous networks 21:265
Heterogeneous organizational forms, public sector organizations 19:599–600
Heterogeneous populations, mortality rates in 15:849
Heterogeneous preexisting subgroups 22:882–885
Heterogeneous space 4:441–442
Heterogeneous variances, time series with 24:320
Heterogeneous workgroups, demographically, behavior consistency across 17:699–700
Heteronomous hybridization 11:451–452
Heteronomous orientation 15:775
Heteronormative citizenship 21:728
Heteronormativity 8:907–908, 21:791
and sexuality 21:809–810
and social work 21:802–806
Heteropatriarchy, and lesbian feminism 8:942–947
Heteroperformance, in structural contingency theory 23:612–613
Heterophobia 15:136–137
Heteroscedasticity 14:147
Heterosexism 10:790–795, 21:809, 21:812
bisexuals 10:790, 10:793
desire 10:791–792
discourse 10:791–792
explanations for 10:791–792
feminism 10:793
gay 10:790–794
gay movement 10:793
gender transgression 10:792
heterosexuality 10:790–792
homosexuality 10:790–794
identity 10:791–792
institutional 10:791
containing sexuality 10:791
internalized homophobia 10:791
and intersectional queer theory 19:765–770
lesbian 10:790–794
oppression 10:794
power 10:794
queer theory 10:792
social movements 10:793
terminology 10:790
transgender 10:790–793
Heterosexual men
HIV infection 11:142–143
impact and epidemiology 11:142
improving future prevention efforts 11:143
intervention efficacy 11:142–143
selected predictors of risk 11:142
Heterosexual privilege
and engineering 7:646–647
and social work 21:803
Heterosexual sexuality 10:801–811
expression 10:803–805
Heterosexuality 2:393, 8:894–896, 8:942, 8:946–947, 10:790–792, 18:534–538, 21:768–769, 21:773, 21:809, 21:811
in Africa 10:812–818
African sexuality 10:814
colonial constructions of masculinities and feminities 10:813–814
colonialism 10:812
female power 10:816–817
mortality and HIV/AIDS 10:812–813
postcolonial critique and alternative analytical approaches 10:814
power of pleasure 10:816–817
sex as asset 10:816
sex as distinct from mortality 10:814–815
sex as risk and danger 10:813–814, 10:816–817
unlearning 10:814
women as active sexual beings 10:815–816
in China 10:819–824
delayed marriage and fertility 10:822
future of 10:823
gender imbalance and marriage squeeze 10:821–822
legendry Gods 10:819
love and couple relationships 10:822
low fertility rate in aging population 10:821
low sexual activity and satisfaction in couples 10:822
patriarchal heterosexuality 10:819
re-enlightenment in 10:820
sex aid industries 10:820
sex cultural festivals 10:821
sex law changes 10:820
sex shops 10:820
sexual permissiveness in youth 10:821
sexual restructuring 10:819–820
sexual studies in 10:819
definition 10:801–803
methodological issues in study of 15:316–317
orientation 21:785, 21:787–788
questioning, same-sex youth 9:652
sexual practices 10:805–808
see also Heterosexual sexuality
Heterosexuals 17:949, 17:951–952
friendship of LGBT individuals with 9:456
and long-term heterosexual relationships 20:915–916
Heterotopia 18:673
Heterotrophic organisms 7:627
Heterozygote advantage (balancing selection) 17:902
Heterozygotes, definition 1:85
Hettner, Alfred 16:334
HETUS *see* Harmonized European time use surveys (HETUS)
Heuristic approaches
mathematical competence 14:818
to strategy selection 6:711
Heuristic case studies 3:209, 3:211
Heuristic principles 5:274
Heuristic proof 17:254
Heuristic search 2:32
Heuristic study of technology 24:136–138
Heuristics 2:445, 4:46–48, 5:913–915, 10:829–835, 23:799
accuracy–effort trade-off 10:831–832
approach to probabilistic thinking problems 19:18–20
and biases 5:892, 8:393–394, 10:831
bias–variance dilemma 10:832–834
bounded rationality 10:830–832
and bounded rationality 19:908
building blocks 5:913
decision rules 5:913
search rules 5:913
stopping rules 5:913
cognitive, taxonomies of 10:834
cognitive illusions 10:830–831
cognitive limitations 10:830–831
computational intractability 10:829, 10:831
for decision and choice 5:892–894
evaluating performance of 5:892

Heuristics (*continued*)
historical overview 5:892
for inferences 5:893
for preferences 5:892–893
for social contexts 5:893–894
and decision making 5:919–921
definition 10:829
as descriptive models of mind 10:830–831
direct voting 25:297
ecological rationality 10:832–834, 10:833f
effectiveness of 10:831–832, 10:832f
and environmental cognition 7:706–707
evolved capacities 10:832
experimental economics 8:550
further directions 10:834
Gestalt 'heuristic' of proximity 10:830f
heuristics-and-biases program 10:831
irrationality 12:719
judgments, Tversky 24:681–682
one-reason decision-making 5:914–915
organizational decision making 17:353–356
person perception 17:740
for political behavior 11:556
as prescriptive procedures 10:829–830, 10:829f
problem solving 10:829–830
prospect theory 19:263–264, 19:266
recognition-based decision-making 5:914
reflexive practice 20:85–86
risk 10:830–831, 10:834
satisficing 14:479–480
Simon on 21:972–974
in social cognition 10:825
affect 10:826–827
anchoring and adjustment 10:826–827
attribution 10:826–828
availability heuristic 10:825–826
feelings as information 10:827
future directions 10:827–828
intuition 10:827
judgments under suboptimal conditions 10:825
processing fluency 10:827–828
rationality 10:825
representativeness heuristic 10:826
trade-off heuristics 5:915
uncertainty 10:830, 10:834
use in decision making 4:108
Heuristic-systematic model (HSM) 15:32
of attitude-behavior consistency 2:186
of persuasion 17:944–945
Hewlett–Packard 11:840
HEXA mutation 19:830
HEXACO model of personality 4:660, 5:671, 9:920, 11:169–173, 17:802
HEXACO-PI-R 9:232–233
Heydrich, Reinhard 16:260
Heymans, Gerardus 19:391
Heyneman–Loxley effect (HL effect) 1:43–44
Hezbollah 24:230–232
HFD *see* Human Fertility Database (HFD)
HFES *see* Human Factors and Ergonomics Society (HFES)
HFS *see* Hypoglycemia fear survey (HFS)
HGDP *see* Human Genome Diversity Project (HGDP)
HGP *see* Human Genome Project (HGP)
HGS *see* Hunting and gathering societies (HGS)
HHP *see* Hungarian Household Panel Study (HHP)
HI *see* Hybrid-isolation (HI) model
Hicks, John R. 3:92, 17:504
macroeconomics 14:401
works
Value and Capital 25:11
Hicks Paradox 23:588
HICPs *see* Harmonized Indices of Consumer Prices (HICPs)
Hidalgo, Miguel 3:536
Hidalgo (Spanish) 3:96
Hidden action 1:344–345
Hidden biases 17:108
choice of circumstances 17:110–111
design sensitivity, tactics to increase 17:111
elaborate theories 17:109
goals 17:110
instrumental variables 17:111
two control groups 17:109–110
Hidden characteristics problem 1:344–345
Hidden curricula, in education 7:147
Hidden Markov model (HMM) 19:11, 23:262, 23:416
discrete-state models 6:524
Hidden variables 8:696
Hide-and-seek task, during childhood 3:465–466
Hierarchical age-period-cohort (HAPC) models, demographic studies 17:731–732
Hierarchical analysis modules 17:761–762
Hierarchical Bayesian Models 17:731
Hierarchical characteristics, in five-factor model 2:563
Hierarchical cluster analysis (HCA) 19:730
Hierarchical clustering 23:393
Hierarchical communities, and social change 4:314
Hierarchical constructs 21:460, 21:465
Hierarchical data structures, multilevel statistical analysis 23:381
Hierarchical designs 8:539
Hierarchical governance, organization and the law 17:398
Hierarchical leagues 23:320
Hierarchical linear model (HLM) 14:62–63, 22:548, 22:550
meta-analyses 15:276–277
Hierarchical models 16:48
agent-based models 1:368–369
for multiple comparisons 16:82
for statistical clustering 23:394–395
of vision 25:153
Hierarchical nature of contextual influence 4:788–789
Hierarchical nested sampling 23:188f
Hierarchical networks
in modern societies 17:604
of semantic processing 21:572
Hierarchical organization of mind 4:515–516
Hierarchical participation 18:504–505
hierarchical predictive coding 25:157
Hierarchical processing, in visual cortex 16:629
Hierarchies 10:841–844, 18:435–436
in agrarian West Africa 1:301
and anarchism 1:682–684
of artist status 2:56
caste system 3:223–224
of classification systems 3:800
and conservatism 4:689
egalitarianism 7:310
of estates 19:156–157
foundations 10:841–842
and imperialism 11:682–684
job evaluation 12:816, 12:818
and markets 14:473–474
in Melanesia 17:155–156
naming 16:184
operationalization 10:842–843
asset specificity 10:842
bounded rationality 10:842–843
contracting 10:841, 10:843
economizing 10:843–844
empirical testing 10:843
full formalization 10:843
governance 10:841–844
human actors 10:842
interdisciplinarity 10:842
opportunism 10:842–843
organization 10:842–843
predictions 10:843
refutable implications 10:843
strategic behavior 10:842
transaction cost economics 10:843–844
transactions 10:842
in organizations 2:769
organized crime as 5:172
of sciences 21:241
of self-concepts 21:470
of society, in South Asia 2:77–78
of species 22:264–265
Hierarchy models
random and fixed effects 10:836–840
basic statistical model 10:836
Bayesian approach 10:836–839
equation-based approaches 10:838–839
inference 10:836–837
types 10:837–838
Hierarchy-attenuating ideologies 22:252
Hierarchy-enhancing ideologies 22:252
Hiestand, K.R., on butch–femme relationship 13:872
Higgins, H., on LGBTI commercial sex 13:885
High Casualty Terrorist Bombing Events (HCTB) dataset 5:837
High culture 3:800, 20:396
artistic/cultural markets 14:565
High density clusters 23:394
High impact disasters, demography 6:103–104

High Net Worth (HNW) philanthropy 3:367
High reliability theory (HRT) 17:371
High resistance, estimator 20:733
  median absolute deviation (MAD) 20:733
High risk–low return findings, stock market 23:485
High roads, of transfer of learning 21:97–98
High sainthood 15:577
High School Longitudinal Study of 2009 (HSLS:09) 5:696–697
High stakes testing, in education 7:295
High threshold theory (HTT) 19:473
high vocal center (HVC) 2:692–693, 2:692f
High-achieving students 7:277
High-altitude hypoxia
  gene-culture coevolution 9:817
High-asset neighborhoods 3:447–449
Highbrow cultural consumption 22:331, 22:333–336
High-context communication 4:284–285
High-context societies 12:357
High-density lipoproteins (HDLs) 1:509
high-dimensional context spaces 21:581–582
High-dimensional variable selection 15:650–651
  best n-term selection 15:650–651
  shrinkage 15:651
  sparse variable selection problem 15:651
High-energy physics (HEP) 7:873, 7:877
  versus microbiology 7:874–875
Higher cognitive processes
  in mammals 23:280
  motor responses activation by 7:411
  sensory-motor activation relation to 7:409–411
Higher dimensional Fourier transform 9:378–379
Higher education 7:279, 22:661–662, 22:680
  affirmative action programs in 1:264–265
  entry to 7:133
  institutions of 7:169–170
  and international organizations 12:538–540
  in Latin America and Carribean 7:178–184
  minority doctorates employed in 22:685
  professional development 19:92–94
  in social sciences 6:489
  social work education 22:775–776, 22:778
  universities 18:307–312
Higher Education Act (HEA) 1:943, 1:945–948
Higher education institutions (HEIs) 10:850–851
Higher education management 10:845–849
  academics 10:845–848
  diffusion 10:845, 10:847–848
  management roles and leadership effectiveness 10:847–848
  managerial revolution 10:846–847
  managerialism 10:846–847
  New Public Management 10:846–847
  university as special organization 10:845
Higher education (HE) market 10:850–861
  Australian 10:857
  conceptual framework 10:852
  emerging patterns of growth 10:859
  expansion of 10:853
  factors characterizing the biggest sending countries 10:857
  globalization and internationalization in 10:854–855
  international student mobility markets 10:855
  key features of dominant 10:857
  marketization 10:853–854
  north–south divide in 10:859–860
  notion of 10:853, 10:855
  population size and 10:857
  prevailing perception of low quality across the sector 10:857
  strong government commitment to economic growth 10:857
  study of destination countries 10:858t
  theorizing 10:850
    classical theory 10:850–851
    investment theory 10:851
    neoliberal theory 10:852
    postcolonial theory 10:852
    reproductive theory 10:851–852
  UK 10:856
  US 10:856
Higher education research 10:862–869
  academic profession 10:867
  comparative analysis 10:867
  growing supranational reference of 10:866–867
  higher education policy 10:862–867
  higher education system 10:862–863, 10:867–868
  international organizations 10:865–866
  internationalization 10:862, 10:866–867
  knowledge production by diverse experts 10:867–868
  learning 10:864–867
  lingua franca 10:867
  major findings 10:868
  roles of persons undertaking research 10:863–865
  teaching 10:864–865, 10:867
  thematic range 10:865–866
  theme-focused and practically relevant field 10:862–863
  variety of institutional bases 10:863–865
Higher Education Statistics Agency 5:736
Higher education system 10:862–863, 10:867–868
Higher mental functions 3:914
  Vygotsky on 25:312–313
Higher mental processes 16:511, 25:322
Higher order logic (HOL) 5:937
Higher order processing, of tactile information and tactile awareness 24:15–16
Higher psychological functions 25:331t
Higher-order analysis of variance 1:661–663
Higher-order association 10:888–889
Higher-order crossover designs 5:351–352
Higher-order efficiency, of system and single-equation estimators 21:978
Higher-order factor analysis 11:750
Higher-order influences 23:241
Higher-order representational theories of consciousness 4:677
Higher-order skills 3:819
Highest good theory, Kant's 13:7
High-income countries, institutionalization in 6:42
High-level vision 16:628–631, 25:153
  action and event understanding 25:155
  art appreciation 25:155–156
  background 25:153
  functional specialization and hierarchical organization 25:156–157
  object recognition 25:153–154
  scene interpretation 25:154–155
Highly reactive children 17:877
High-occupancy toll (HOT) lanes 24:616
High-occupancy vehicle (HOV) lanes 24:619
High-pass filter 24:335
High/political policing 18:247–248
High-precision storage of sensory information 25:179
High-quality classroom assessment
  efficiency 3:823
  fairness 3:822–823
  practicality 3:823
  reliability 3:822
  validity 3:822
High-quality close relationships 22:564
  and responsiveness 22:564
High-quality education 24:56
High-quality relationships
  consequences of 13:643–644
  development of 13:642–643
  importance of 13:644
High-quality services, child and adolescent psychiatry 3:386
High-quality teachers
  in United States 24:75–76
High-reliability organizations (HROs) 17:371, 24:748
  action research 17:372
  compressed time constraints 17:372
  extreme hierarchical differentiation 17:372
  high degree of accountability 17:372
  hypercomplexity 17:371
  immediate feedback, need for 17:372
  multiple decision makers 17:372
  reliability enhancing characteristics
    deference to expertise 17:372
    preoccupation with failure 17:372
    reluctance to simplify 17:372
    resilience 17:372
    sensitivity to operations 17:372
  of school effectiveness 21:83
  tight coupling 17:372
High-resolution structural imaging
  for hemispatial neglect 10:768
High-Scope Perry Preschool Program 6:740
High-skill policy 7:216

High-tech manufacturing sectors, innovation and employment 12:156–158
Highway Code 21:531
High-yielding varieties (HYV) 9:307–308
  principal characteristic 9:308
High-yielding varieties (HYV) seeds 1:459
HIIK *see* Heidelberg Institute of International Conflict Research (HIIK)
Hijabs incident 9:732
Hijras 12:738, 19:755
Hijri calendar 11:64
Hikayat Banjar 11:82–83
Hikkikamori 5:628
Hilberg, Raul 10:870–872
  on bureaucracy 10:870–871
  on Cold War 10:870
  on empiricism 10:871–872
  on Holocaust 10:870–872
  on national socialism 10:870–871
  on Neumann 10:870–872
Hilbert, David 1:530, 16:583
  axiomatic theories 2:339
HILDA *see* Household, Income and Labour Dynamics in Australia (HILDA) survey
Hildebrand, Bruno 4:340
Hilferding, Karl 11:686
Hilferding, Rudolf, works, *Das Finanzkapital* 3:95
Hill, K., works, *Ache Life History: The Ecology and Demography of a Foraging People* 8:411
Hill, Lewis 8:291
Hill, Rowland, works, *Post Office Reform: Its Importance and Practicality* 18:641
Hill, S.B.
  criteria of causal inference 7:827t
Hillary, Edmund 17:429
Hillbilly music 3:800
Hillman, M., works, *One false move: a study of children's independent mobility* 3:495
Hilton, James, works, *Lost Horizon* 14:255
Hilum, in plants 17:458
Himmler, Heinrich 9:647–648, 16:260
Hinde, R.A., works, *Growing Points in Ethology* 22:900
Hindelang, M., works, Victims of Personal Crime: An Empirical Foundation for a Theory of Personal Victimization 20:774
Hindelang, Michael 20:774
Hindenburg, Paul von 9:213
h-index 21:322
Hindi language 1:955–956, 7:195–196, 10:320
Hindi–Urdu controversy 4:263–264
Hindmarsh Bridge Island litigation (Australia) 5:495
Hindsight bias 13:748–749
Hindu cosmology and religion 18:216
Hindu fundamentalism 20:409–410
Hindu law 13:619, 13:800–801, 20:375, 23:58
Hindu Mahasabha 20:375
Hindu method of tribal absorption 2:77
Hindu nationalism 18:655, 20:375–376
Hindu yoga 16:168
Hinduism 3:533, 20:331, 20:360, 20:373–374, 20:398, 20:408, 20:410–411, 21:173–174
  anthropology of 20:308, 20:310–311
  beliefs about mental illness 19:308
  conceptualization of 20:374
  leadership in 20:410
  and morality 15:806, 15:808–809
  in South Asia 2:76, 2:78
  in Southeast Asia 13:813, 23:70
  Thugs 24:237–238
Hindu–Muslim relations 20:376
Hindus 20:373
  and communalism 4:263–265
Hine, Lewis 18:84, 25:181
Hine, Robert V., works, *California's Utopian Colonies* 2:144
Hinge groups 2:870
Hinshaw, Robert 25:559
  works, *Living with Nature's Extremes: The Life of Gilbert Fowler White* 25:559
Hintze, Otto 10:873–875, 25:410
  comparative approach to constitutional and governmental history 10:873
  contribution made to historiography by Comte and Spencer 10:874–875
  exploration of Prussian history 10:873–874
  extensive writings of 10:873
  implications of power interests 10:874
  on Lamprecht dispute 10:874–875
  principle of clear abstraction 10:874–875
  reception of Weber's writings 10:875
  on spatial-typological procedure 10:874
HIPAA *see* Health Insurance Portability and Accountability Act (HIPAA)
Hip-hop 25:843–847
Hippies 5:73, 5:75, 5:81–83
Hippocampal amnesia 10:876–885
  anterograde amnesia 10:876–879, 10:881
  associative memory 10:877, 10:879, 10:882–884
  episodic memory 10:876–877, 10:881, 10:883–884
  extended hippocampal system 10:880
  familiarity 10:879–884
  imaging 10:882–883
  lesion evidence 10:878–882
  object-related processing 10:877–878, 10:883–884
  pattern-separated memory 10:881, 10:883
  perirhinal cortex 10:877–879, 10:881–884
  recollection 10:879–884
  retrograde amnesia 10:876–879, 10:881, 10:883–884
  semantic memory 10:876–877, 10:883–884
  spatiotemporal context 10:878, 10:881–884
Hippocampal bursting 20:215
Hippocampal commissure 12:416
Hippocampal dependent system 11:389
Hippocampal formation 1:653–654, 13:679, 14:125–126
Hippocampal lesions 5:925
Hippocampal memory system 14:128–129
Hippocampal sclerosis 7:854
Hippocampal-neocortical dialogue 22:64, 22:64f
Hippocampus 1:644, 5:925, 8:849–851, 10:886–891, 11:388, 13:682, 16:626, 16:644, 19:274–276, 20:58, 25:713
  activation and eye movements 1:647
  clinical relevance 10:890
  cortico-parahippocampal–hippocampal circuits 10:888–889
  and dreaming 6:653–655
  and enriched environment 2:837
  and fear conditioning 1:811–815
  formation, microscopic anatomy of 3:342f
  fornix 10:889
  functions 10:877–878, 10:878f
  functions of 10:889–890
  general features of 10:886–887
  hippocampal formation 20:58
  implications in implicit or nondeclarative memory 1:647
  interconnectivity between deep and superficial layers 10:888
  local circuits 10:887–888
  LTP/LTD 14:358–364
  and memory binding 15:144–145
  parahippocampal region 10:886–887, 10:887f
  and process of consolidation 1:646
  recollection/familiarity distinction 1:647
  and remote memories 1:648
  role in memory formation 23:877
  serial order memory 21:642–643
  source monitoring 23:31–34
  and spatial navigation 23:163, 23:166, 23:168–169
  species differences 10:886–887
  structures 10:886
    anatomical organization of 10:889
  and transient global amnesia 24:565–568, 24:568f
  virtual spatial navigation 16:348–350, 16:352–354
  during working of working memory 1:646–647
Hippocampus remembers 15:547
Hippocampus/caudate nucleus dichotomy 11:389
Hippocampus–fornix connectivity 10:890
Hippocrates 17:27–28, 19:406–407, 22:485
  on mental disorders 3:877
  on personality 17:906–907
  philosophy 18:31–32
  psychology 19:406–407
Hippocrates of Cos 7:720, 15:119
Hippocratic oath 8:64–65, 8:280
  contemporary relevance of 8:64
  fiduciary principle 8:65
  honor in medicine 8:64–65

for psychiatrists 8:65
Hippocratic-Galenian medicine 15:119
Hirakana 20:1
Hiring, in law firms 13:578–582
Hiring standards 15:596
Hiroshima Peace Memorial 5:416–417
Hirsch, Paul, on cultural industries 5:536
Hirschfeld, Hermann Otto 5:8
Hirschfeld, Magnus 9:638, 9:644, 9:647
and Institute of Sexual Research 9:647–648
works
Wissenschaftlich humanitäres Komitee (WhK) 9:644–647
Hirschhorn, Larry, on knowledge society 13:107–108
Hirschi, T.
general theory of crime 21:481–482
self-control theory 21:481–485
works
Causes of Delinquency 4:836
Hirschl, Ran 4:704–705
Hirschman, Albert O. 20:660
on capitalism 10:892–896
on democracy 10:892–893, 10:896–897
on economic development 10:892, 10:894–895
on exit 10:895
on futility 10:896
on ideology 10:894–896
on industrialization 10:896
on intellectuals 10:892, 10:896–897
on jeopardy 10:896
Latin American influences 10:893–894
on perversity 10:896
on possibilism 10:894–895
on self-subversion 10:896–897
on unbalanced growth 10:894
on voice 10:895–897
Hisabetsu-burakumin 12:794
Hispalensis, Isidorus 7:721
Hispanic, use of term 9:628
Hispanic Catholics 20:294
Hispanic crime 5:184
age and sex distribution 5:185
illegal Hispanic immigrants and 5:185
imprisonment rates 5:185
neighborhood disadvantage and 5:184
within-city change and violent 5:185
see also Criminal justice system (CJS)
Hispanic ethnicity 19:813
Hispanic-owned businesses 11:630
Hispanics 8:160
discrimination of 5:155
ethnic composition in United States 18:557t
Hispaniola 1:638
Hispanoamerica 25:282
Histograms 10:351–353, 10:355–357, 10:352f–353f, 10:356f
probability density estimation 19:29–30
Histoire croisée 11:44
Histoire de la France rurale 17:670–671
Histoire et Sociétés Rurales 17:671–672
Histoire historisante 1:708–710
Histoire moderne 15:703
Histoire Sciences Sociales 11:45–46
Histoire totale 11:34
Histone deacetylases (HDACs) 7:840–841, 7:843
Histone modification 7:849–850, 22:282
social epigenetics 22:282–283
Histone proteins 7:834
Histones
acetylation 7:840–841
core, posttranslational modifications of 7:839–840, 7:840t
methylation 7:841
monoubiquitination 7:841
role of in transcriptional regulation 7:839–840
tail modifications 7:841f
Historia Agraria 17:671–672
Historia religionis 20:323–325
Historians 11:120–121
African 1:288
and *isms* 12:760–764
sport 23:307–312
Historic archaeology 9:766
Historic religion 2:507–508
Historical and comparative sociology 22:994
Historical and nonhistorical nations 16:287
Historical anthropology 1:708–709, 1:747
ethnohistory 8:192–194
Historical anthropometrics 19:697
Historical archaeology 11:1, 14:670–671
of contemporary material culture 11:3
ethnohistory 8:192
feminist and gendered approaches in 11:2
of free communities 11:2
as a global field 11:2
interdisciplinarity of 11:3
lack of diversity in 11:2
landscape-based approaches 11:2
of migrants 11:3
postcolonial theory 11:2
social, cultural, and economic issues 11:1
in studies of capitalism 11:1
as study of colonialism 11:1
study of the African Diaspora 11:2
Historical biography 2:644–649
Historical bloc, concept of 10:742–743
historical case studies of theoretical disputes 21:264
Historical change 1:708–709
in modern youth movements 9:868
Historical climatology 3:865–867
Historical comparative 22:31–32, 22:34–35
Historical consciousness 1:748, 11:114–117
modes of 4:182–183
Historical contexts
social work with refugees 22:770
studying indigenous knowledge in 11:781–783
Historical contextualization 4:341
Historical continuities, fertility 9:24
Historical culture 11:115–116
aesthetic dimension 11:115
cognitive dimension 11:115–116
comprehensive large-scale developments 11:118
East Asian 11:48–49
ethical dimension 11:116
political dimension 11:115
religious dimension 11:116
Historical databases 5:700–703
Historical demography 10:776, 11:5–9, 15:345, 19:696
of China 11:9
cross-cultural perspective 11:8–9
family reconstitution 6:138–139
inverse projection 6:149
of Japan 11:9
micro and macro 11:6–8
macrodemography 11:7–8
microdemography 11:6–7
social science 11:5
sources, methods, and impacts 11:5–6
Historical empathy 22:728
Historical epistemology 4:550–551
of objectivity 17:100–101
Historical epochs, mode of production 15:714
Historical ethnography 1:747
Historical events 1:328, 4:772
Historical explanation
causal explanations 11:12, 11:15–16
historical change, general interpretations 11:10–11
metahistories 11:11
theories of
descriptive theories 11:15
hermeneutic explanation 11:12–14
narrative explanation 11:14–15
philosophical aspects 11:10–16
postmodern theory 11:15–16
prescriptive theories 11:11–15
scientific explanation 11:11–12
structural explanation 11:14
Historical genres 9:162
Historical geography 11:17–22
capitalist transformation 11:18–19
cultural geography 11:19–20
environmental transformations 11:19
fluid 11:20–21
landscape change 11:19
national styles 11:20
place for 11:17–18
regional 11:18
review 11:21–22
Historical inquiry method 19:676
see also Inquiry-based learning
Historical institutionalism 11:562, 11:566–567, 12:183–184, 12:187–189
Historical legacies and 'scars' of evolution 8:452–453
Historical liberalism 14:19–20
Historical linguistics 1:959, 8:207–210, 14:225
evolution of language 8:365–367
numerical methods 11:23–26
Historical materialism 1:665, 6:242, 11:103–104, 12:121, 14:645,

14:670, 14:674–675, 14:766–768, 22:228, 24:109–110
concept of 22:339
Giddens on 10:136
and Marxism 14:649, 14:651–652
Historical Materialism 14:649
Historical memory 15:329–330
Historical methods 4:405
in sociology 22:991
Historical morphophonemics 12:447–448
Historical nonalienation, authenticity as 2:316–317
Historical objectivity 11:49
in East Asia 11:49
Historical phenomenology of religion 20:346–347
Historical positivism 1:709
Historical process, media imperialism 15:56–57
Historical professionalism 11:123–124
Historical progress 11:10–11
Historical psychology 5:422
Vygotsky on 25:330–331
Historical rationalists 5:504
Historical reasoning 6:486
Historical reconstruction 14:211
Historical regions, in Europe 8:274–276
central–eastern Europe 8:275
eastern Europe 8:275–276
Mediterranean Europe 8:275
northwestern Europe 8:275
present socioeconomic gaps 8:276
southeastern Europe 8:275
Historical Schools 11:93, 11:95–96
Historical semantics 4:548–553
Historical settlement patterns, in United States 11:629, 11:631
Historical social science 19:676, 19:695
Historical social system 22:744–745
Historical sociologists 16:293
Historical sociology 1:262, 11:27–30, 11:105, 15:700, 25:410
Bielefeld School 2:553–555
class 11:27
Collins's contributions 4:211, 4:214–215
culture 11:28–29
economy 11:27–29
identity 11:27
Mann's contributions 14:490–495
Marx's contributions 11:27
methodology 11:27–29
modernity 11:27
Moore's contributions 15:772
politics 11:28–29
revolutions 11:27–28
social movements 11:27
society 11:27–28
state 11:27–28
of state 25:420–422, 25:426
Weber's contributions 11:27
Historical sources
assessment of 11:413–414
records as 5:701–702
Historical theology 25:291–292
Historical thought 14:203
and historiography
Christian tradition 11:31–34
classical period 11:35–38
current trends 11:39–47
East Asia 11:48–53
indigenous cultures in Americas 11:54–60
Islamic tradition 11:61–66
modern history 11:72–77
Southeast Asia 11:82–88
Jewish tradition of 11:67–71
in South Asia 11:78–81
in sub-Saharan Africa 11:89–92
Historical thought and historiography (classical period)
antiquity 11:36
and Aristotle 11:35
Christianity, impact of 11:38
Greek historiographical production, variety of 11:36–37
Greek historiography 11:35–36
Roman imperial historiography 11:37–38
Roman republican historiography 11:37
Historical thought and historiography (current trends)
Annales 11:45–46
impact of 11:40–41
concept of nation, transformation of 11:44
cultural history 11:41
cultural turn impact on Marxist historiography 11:42–43
historical writing after 1990 (global academic community) 11:43–44
historiographical trends after 1990 11:44
historiography of politics after 1990 11:45–46
linguistic impact and cultural turn on writing of history 11:42
Marxism, impact of 11:40–41
Marxist historiography 11:45–46
reorientation of historical writing 11:44–45
social science, challenges to 11:41
social science models, dominance of 11:39–40
Historical thought and historiography (Islamic tradition)
Abbasids (caliphal dynasty, CE 749–1258) 11:63–64
Arabic (language, literature) 11:61–65
chronology 11:64
Europeanization 11:65–66
form of knowledge 11:61–62
mature tradition, 950–1800 11:64–65
and Persian culture 11:65
problem of origins, CE 622–825 11:63–64
producers of historical knowledge 11:62–63
as profession 11:61–62
prophet/prophethood 11:62–63
prosopography 11:64
Qur'an 11:63
transition (Islamic to European models, 1800–1950) 11:65–66
Umayyads (caliphal dynasty, CE 660-749) 11:63
Historical thought and historiography (in West, since eighteenth century)
cultural turn 11:76
emergence of a secular outlook 11:72–73
Hegel's dialectic 11:72–73
Marxism-Leninism 11:76
Marx's approach 11:73–74
new orientations after 1945 11:76
nineteenth-century conceptions of history and historical science 11:73–74
professionalization of historical studies 11:74–75
romantic historism 11:73–74
scientistic positivism 11:73–74
social Darwinism 11:75
social science turn 11:75–76
twentieth-century reorientations of historical thought and scholarship 11:75–76
Weber's approach 11:75–76
Western civilization 11:73, 11:75
Western tradition 11:72
Historical thought and historiography (indigenous cultures in Americas)
achievements 11:58–59
concepts of history and their formation 11:54
future problems 11:58–59
mesoamerica 11:54
mesoamerican historiography, components of 11:54–57
geographical horizon 11:55–56
historical chronology 11:55
origins of humans 11:55
polities 11:55
reflections on history 11:56–57
social history 11:56
sources 11:57
tutelary deities 11:55
regions of Americas 11:57–58
North America 11:58
South American Andes 11:58
Historical thought and historiography (Southeast Asia)
autonomous modernity 11:83
chronicle traditions 11:82–83
debates 11:85–88
dynastic chronicles 11:83–88
European enlightenment influences 11:83–84
in Indonesia 11:85
Marxist history 11:84–85
nationalist history 11:84–85
postnationalist history 11:85–88
regional history 11:85–88
in Thailand 11:85
thematic history 11:85–88
in Viet Nam 11:85
Historical time and place, and linked lives 14:231–232
Historical trends, time use research 24:391
mass observation 24:392–393
USDA 'Homemaker studies' 1923–1931 24:392

Historical turn, in philosophy of science 13:136
Historical wrongdoing, commemorating 4:237
Historical-Climatologic research 3:867
Historicism 4:521, 11:93–96, 11:132, 21:167–168
  Bielefeld School 2:553–554, 2:557
  British contributors 11:94
  crisis of 11:33–34
  defined 11:93
  and diffusion 6:403
  New Historicism 11:96
  nineteenth century German 10:786–787
  Popper's contributions 18:523, 18:525–526
Historicity 22:691, 24:460–461
  accumulation 24:460–461
  cultural model 24:460–461
  Dilthey on 6:424, 6:426–429
  and fetishism 9:107–110
  model of knowledge (image of society and nature) 24:460–461
Historicization of agency 14:206
Historie 11:35
Historikerstreit *see* Holocaust
Historiography 1:708–709, 4:405, 11:103–104, 11:129, 18:380, 18:382, 20:322, 22:307–311, 25:290–292
  absolutism and 1:6–7
  Bielefeld School 2:553–555, 2:557
  body 2:728, 2:730–731
  branch 2:644
  communication and transportation 4:275–276
  of communication research 4:278–279
  and conceptual history 4:548
  and fiction, conflation of 11:98–99
  film 9:169–170
  global history 10:162–164
  and historical thought
    development of modern historiography 11:52
    East Asian historiography, development of 11:49–50
    China 11:49
    Japan 11:49–50
    Korea 11:50
    Vietnam 11:50
    historical objectivity 11:49
    historiology in East Asia 11:50–51
    modern historiography in East Asia 11:51
    public history, metamorphoses of 11:52
    role and purpose 11:48–49
  of interviewing 12:670
  Jewish tradition 11:67–71
  labor history 13:169–171
  of linguistics 14:215, 14:217–218
  of Meiji Revolution 15:129
  migration history 15:447
  narration in 11:97–98
  of professions 19:99–100
  prosopography 19:256–257
  of psychoanalysis 19:328–329
  of racism 19:852–853
  Reformation and confessionalization 20:98–100
  of social and behavioral sciences 15:319–320, 15:323
  social stratification and class in 22:339–341
  of South Asia 11:78–81
  of sub-Saharan Africa 11:89–92
  working classes 25:695–696
Historiography and historical thought *see* Historical thought and historiography *entries*
Historiography and Historical Thought: Current Trends 11:76
Historische Sozialwissenschaft (historical social science) 2:554–556, 11:104
Historism 11:39
Historismus 11:122
Historization, of Frankfurt School 9:384–385
History 3:61–62, 6:176–181, 11:114–119
  African 1:278–279, 1:286–291
  of AGIL 1:381
  of alienation 1:539–543
  of ancient Egypt 1:691, 1:693
  of animal rights 1:703–704
  anthropological universals 11:114
  and anthropology 4:181–182
  of anthropology 6:658–659
    Mexico 1:627
    tribes 24:645–646
  anthropology and 1:746–750
  anthropology of 1:748–749
  of antidepressants 1:778–779
  of applied social research 1:850–857
  of Arab nationalism 16:271–274
  of archaeology 1:881–883, 1:899, 23:742
  of archives 5:700–701
  of area studies 1:933–934, 1:949–954
  Aristotle on 11:126–127
  of artistic/cultural markets 14:564–565
  of arts 2:17–19
  of associations 2:88–90, 25:253
  of attitude–behavior consistency 2:183–184
  balance of power 2:345
  of behavior therapy 2:413–416, 2:426–427
  of behavioral economics 2:443–449
  Bielefeld School 2:553–558
  of bilingual education 2:584
  of bioethics 2:613
  and biography 2:638–639
  of biomedical sciences 2:663
  of bipolar disorder 2:683
  of body 2:728–734
  of burnout research 2:929–930
  of business ethics 3:15–16
  of capitalism 3:105–110
  of case study methods 3:209–211
  of causal analysis in social sciences 3:247–250
  of censorship 3:279–280
  of change blindness 3:344
  of chemical sciences 3:371–375
  of Christianity 11:31–34
  of civic humanism 3:63
  of civil service 3:695–696
  of civil society/public sphere 3:701–706
  of clinical psychology
    Canada 3:881
    Europe 3:873–876
    Mexico 3:881
    North America 3:877–883
  of cognitive behavioral therapy 4:30
  of Cold War concept 4:131–133
  of collective efficacy concept 4:169
  collective memory 11:129
  and collective memory 4:194–198, 4:202–203
  and commemoration 4:235–236
  of communication and transportation 4:273–277
  of communication research and media studies 4:278–283
  of concept of attitude 2:175–177
  conceptions of time 11:114
  conceptual *see* Conceptual history
  Concert of Europe 2:346
    global successors 2:346–347
  of confederation 4:570–576
  Congress of Vienna 2:346
  of conspicuous consumption 4:695–699
  of content validity 4:774–775
  of contextual studies 4:786–787
  of coping 4:888–889
  of critical psychology 5:261–262
  crowds in 5:361–366
  cultural *see* Cultural history
  of cultural sociology 5:498–503
  of cultural studies 1:933–934
  definition 4:181
  of democracy 6:80–84
  of development studies 6:308
  developments 11:117–118
  of dichotic listening research 6:357–358
  of differential diagnosis in psychiatry 6:385–386
  disasters in 10:549–550
  discourse 11:128–129
  and disgust 6:548
  of distinction 4:695–699
  of domestic violence in law 6:625–626
  of East Africa 6:773
  of Eastern European Studies 6:810–815
  of eating 11:420
  of ecofeminism 6:845–847
  of econometrics 7:8–11
  economic *see* Economic history
  of economic databases 5:742–743
  economic growth 7:46
  of economics 7:93–98
    engagement with 7:95–96
    in twenty-first century 7:96–97
  of education 7:172–177
  of education for sustainable development 7:733–735
  of educational institutions 7:232–237
  of educational neuroscience 16:728–729
  of elites 7:390–395
  of emotions 7:472–476, 7:513–518

History (*continued*)
of empirical social research 7:558–566
encyclopedia of basic concepts in 4:551–552
of environmental education 7:733–735
epistemic communities and collaborative research 7:867–872
of equilibrium 7:912–917
of ESR 7:558
of ethical questions in social and behavioral sciences 8:53–59
of ethnic cleansing 8:83–90
of ethnocentrism and prejudice 8:168–174
ethnohistory 8:192–197
and ethnology 8:198, 8:200–201
event and narration 11:114–115
of everyday life 11:108–113
of evidence-based practice 8:332–333
of experimental methods in population studies 8:559–560
explanation 11:126–128
of extraversion 8:623–624
of family 8:809–814
of family and kinship 8:738–744
of family policy in Europe 8:799–800
of female genital mutation 8:879–880
of fertility transition in China 9:46–47
field observational research 9:135–136
film *see* Film history
of financial crises 9:183–190
of folklore 9:280–281
approaches 9:281
discipline 9:280–281
terminology 9:280
of food 9:311–315
Foucault on 11:129
of Frankfurt School 9:380–386
of freedom 9:387–391
Freyre on 9:422
functions of 11:125–130
of funding 9:550–554
of *Geist* 9:666–670
gender in 9:791
of genetics 9:984–985
genres 9:162
of gift 10:139
global 10:162–168
of graphics 10:359
of habit 10:475–479
of happiness 10:521–525
of health 15:119, 15:121
of health definitions 10:607
of health statistics 5:782
of hermeneutics 10:780–781
early history 10:780–781
prehistory 10:780
of heuristics 10:829–835
historical culture 11:115–116
historiography 11:129
of household economics 11:224
of human flourishing 3:63
of human intelligence 12:303–308
of human sciences 11:380–385
of hypnosis 11:465–466, 11:472–473
of IASSW 12:476–479
of ideas 2:545–546, 4:550, 12:255
African Studies 1:278
of informed consent to research 12:110–111
of instinct 12:175–180
of institutionalism 12:181–182
of institutionalism and neo-institutionalism 12:186–189
intellectual 3:62
of intelligence 12:309–312
intercultural communication 11:118–119
of International Council on Social Welfare 12:495–496
of interpretive research 12:648–650
reading 12:649
of intracarotid sodium amytal testing 12:689–690
isms and 12:759–764
of labeling theory 13:150–155
labor *see* Labor history
of laicization 13:216–219
of language families 13:337–343
of lexicography 13:940
and life course 14:80–83
life course in 14:76–79
of life management 14:85–86
of life span development 14:100
of linguistics (HoL) 14:215–219
Chomsky on 14:215–217
Kuhn on 14:215–216
language sciences 14:215, 14:217–218
proselytizing history 14:217
Saussure on 14:216–218
textbook history 14:215
Whig history 14:216–218
of literacy and illiteracy 14:236–240
and literature 11:97–101
at local/global confluence in Japan 12:795–797
of logical empiricism and logical positivism 14:313–314
LTP/LTD 14:358–359
of macroeconomics 14:406–413
management 15:675
of mania 14:484, 14:484f
of maps and cartography 3:173–174
Marxist geography 14:674
of mass society 14:743–746
master narratives 11:115
of material flow analysis 14:760–761
of mathematical psychology 14:808
of MDS 16:42
of measurement theory 14:868–872
media 11:116–117
of medical ethics 8:63–64
ethics of advertising 8:64
memory 11:115
of mental causation 15:172–173
of mentalities 5:420–423
of meta-analysis 15:266–271
of microeconomics 15:331–337
migration 15:446–451, 15:472
of migrations 4:409
military *see* Military history
of mobile communication 15:629–630
of modernity 15:692–696
modernization and modernity 15:702–703
of motivation 15:936–939
of multiculturalism 16:28–29
narrative networks 16:213–214
narrativity 11:125–128
of nature/nurture controversy 16:340–344
Near Middle East/North African studies 16:381–386
neoliberalism
crisis of 1890s 16:464
crisis of 1970s 16:464–465
Great Depression 16:464–465
of neuroendocrinology 16:648
of neuroesthetics 16:657
of neurolinguistics 16:678–689
of neuropsychological rehabilitation 2:813
Nietzsche on 11:128–129
of North America 17:15, 17:17, 17:19–20
of nosology 17:27–28
of obedience 17:61–66
of objectivity 17:88–91
occupational safety and health 25:727
of opsins 17:269
of organizational learning 13:695–696
particularism 1:766
of peace 17:625–626
peacemaking in 17:637–642
of personality disorder 17:891
personalized medicine 17:925
of philosophy 3:63–64
of Physical Education and Sport (HISPA) 23:307–308
of placebo-controlled studies 18:157, 18:159–160
of planning profession 24:883–884
of political psychology 18:427–428
of political thought 3:65, 18:480–490
of posttraumatic amnesia 18:636
poverty in 18:728–732
of proactive interference 19:1–2, 19:2f
process of enslavement 22:37–38
professionalization of 11:120–124
breadwinning 11:120–121
system of norms 11:121–124
proportional representation 19:208
of psychiatry 13:130
of psychoanalysis 19:326–329
of psychological stress 23:583
of psychologism 19:391–393
of psychology 19:406–417
of psychoneuroimmunology 19:445
of psychosomatic medicine 19:487
psychotherapy 19:517
of psychotherapy, in psychiatry 19:502–508
of publishing 19:620–621
of race and social work education in Canada 8:152–153
of racism 19:852–856
of radio 19:878
of randomized controlled trials 8:541–542
of realism 20:27
realistic view of 20:27–28
of reflective equilibrium 20:77–78

of regional planning 20:141–145
of reification 20:203–206
of relationship between organizations and law 17:397–398
of relationship between personality and health 17:792–793
of relativism 20:233
of religion 15:347, 15:349–350
of religions 20:384, 20:387
representation 11:125–126
representational theory of measurement 14:856–857
research, creativity and personality 17:770–771
of research on job satisfaction 12:834
of resource depletion 15:554–556
of retroactive interference 20:613
of revolutions 20:637–638
rhetorics 11:127–128
Roma people
  culture 20:752–753
  language 20:749–750
  occupation 20:750–751
  politics 20:751–752
of sanctions 20:890
of science 13:133–134, 13:136–137, 21:201
of science, constructivist perspectives 21:225–228
  laboratories 21:226–227
  local cultures 21:225–226
  material culture of experiment 21:226
  science as performance 21:226–227
  science at work 21:227–228
  scientific audiences 21:225, 21:227
  scientific practice 21:225–227
  skills 21:225–228
of science and media 21:163–164
of science and technology studies 21:200–205
of sciences 1:734–735, 13:116, 15:319–322, 17:99, 21:172, 21:241–246
  developments 1970–2010 21:244–246
  emergence and divisions of discipline 21:241–242
  narratives and periodization 21:242–244
of scientific disciplines 21:287–290
of scientific revolution 21:318–321
of self 21:445–448
self-efficacy 21:509
of self-efficacy 21:505
of semantic differential 21:563
of semantic memory 21:572–574
of sexism 21:693–694
of sleep 22:87–88
of small group research 22:100–107
of social and behavioral sciences, methodology 15:319–325
of social cognition 22:205–206
of social constructivism 22:228–229
of social control and deviance 22:235–241
of social democracy 22:242–243
of social inequality 22:338–344
of social movements 22:442–446
of social network analysis 16:524–531
of social psychological theory 22:525–531
social representations of 4:195, 4:197
of social science 9:984–985
and social sciences 11:102–107
  comparative history 11:104
  cultural history 11:102–106
  disciplinary divergence 11:102–103
  economic history 11:103–105
  economics 11:102–103, 11:105–106
  historical materialism 11:103–104
  historiography 11:103–104
  industrial society 11:104–105
  intellectual history 11:103–104
  interdisciplinary challenges 11:105–106
  interdisciplinary convergence 11:103–105
  interdisciplinary perspectives 11:106
  linguistic turn 11:105
  Marxism 11:103–104
  modernity 11:102–104
  modernization 11:103–104
  natural science 11:102–103
  technology 11:104
of social sciences 1:949, 7:588, 9:550–551, 9:553
of social sciences and disciplines 6:485
of social structure 22:719–726
of social work in Russia 22:807
of society 22:307, 22:311
of somatotherapy 23:16–21
of sport 23:307–312
of state 23:353–354, 23:362–367
of state dealing with deviants 4:823–824
of statistics
  econometrics 7:8–11
of street children 23:534
and structuralism 23:627–630
synthesis or recombination in 23:228
of teacher education 24:51–52
technological determinism 24:109–110
of technology 24:135–136
temporal distinctiveness models 24:199–201
of terminology of negative symptoms 19:290
of terrorism 24:237–238
theories and methods 11:131–137
theories of sexuality 21:846
and theory 3:62–64
of thought 19:16–17
of thought, Stigler, George Joseph 23:458
trickster as sign of 24:654–656
typologies of historical representation 11:116
of uncertainty 24:716–721
of unintended consequences 17:420, 24:744–749
of university hazing in United States 10:557
of university press 19:624–626
of unobtrusive measures 24:767–773
of urban violence 24:953–954
urbanization in Africa 24:965–966
of values 25:38–39
venture capital 25:59–60
of violence 25:116–121
of voluntary agencies 25:261–262
of warfare 25:414–419
of Weberian social thought 25:462–467
Western, construction of disability in 6:453–454
of Western penal development 19:655–657
of white-collar workers 25:564–568
witchcraft 25:604–605
women in 25:622
of workplace deviance 25:723
History and Sociology of the Biomedical Sciences (HSBS) 2:663, 2:667
History education 22:727–728
History effects (internal validity) 12:451
History from below (narrative) 11:108–113
  spirit possession 23:283
History of the Persian War 9:858
History of the Republic of Korea 11:52
History Workshop 11:45
History writing, *isms* in 12:763–764
Hitch, Graham 25:715
Hitchens, Christopher 24:451–452
Hitherto 3:719
Hitler, Adolf 1:826, 3:356, 4:151–152, 4:764–765, 16:257–258, 19:351, 21:369–370, 24:448–449
  appeasement 1:826–830
  and homosexuality 9:646–648
  and Nazi law 16:358
  penal policy 19:645
  works, *Mein Kampf* 16:358
Hitlerism 22:140
Hitoshi, Miyake 20:369
HITs *see* Human intelligence tasks (HITs)
Hitzb ut-Tahrir 16:807
HIV *see* Human immunodeficiency virus (HIV)
HIV/AIDS 10:634–635, 10:637, 20:271, 21:702, 21:704
  and child/infant mortality 12:24
  emergence of 20:447
  experiences of volunteers 25:270–271
  in immigrant groups 15:407–408
  and internal migration 12:440
  prevalence, in Southern Africa 23:75, 23:80
  prevention of 8:618–619, 21:825, 21:827–828
  and sexuality, in Africa 10:812–813
  see also Human immunodeficiency virus (HIV)
HIV/AIDS epidemic
  and mortality 15:834–843
    age patterns 15:839–841, 15:840f
    antiretroviral therapy 15:836–837, 15:839–843
    dynamics 15:835–836
    epidemiology 15:835, 15:841
    future 15:842–843
    global overview 15:836–839, 15:837f–838f
    impact on life expectancy at birth 15:841–842
    impact on population size and age structure 15:842

HIV/AIDS epidemic (*continued*)
regional trends 15:839, 15:839t
Hjelmslev, Louis 2:358
HKA *see* Hudson–Kreitman–Aguade (HKA) test
HKIAPS *see* Hong Kong Institute of Asia–Pacific Studies (HKIAPS)
HKIASP *see* Hong Kong Institute of Asia–Pacific Studies (HKIASP)
HL *see* Lexical pitch accent (HL)
HL effect *see* Heyneman–Loxley effect (HL effect)
HLM *see* Hierarchical linear model (HLM)
HM *see* Henry Molaison (HM)
HMD *see* Head mounted display (HMD); Human Mortality Database (HMD)
HMM *see* Hidden Markov model (HMM)
HMOs *see* Health maintenance organizations (HMOs)
HNW *see* High Net Worth (HNW) philanthropy
Ho, Karen
on corporate cultures 5:536
works, *Liquidated* 8:902
Ho, Peter, on landownership in China 5:598
Ho Chi Minh 24:449
HOA *see* Homeowners association (HOA)
Hoaglin, D.C., works, *Statistician's Guide to Exploratory Data Analysis* 8:603
Hoarding, and inequality 11:910–911
Hoarding disorder 17:113
Hobbes, Thomas 1:680–681, 3:674, 4:337–339, 6:696, 8:577, 13:762, 19:330–331, 19:410, 21:213, 22:696–697, 23:363–364, 24:118
conception of power 18:766
on contractarianism 4:804–808
on freedom 9:388–389
legacies 1:765–766
on mass society 14:744
on political thought 18:481
on war 4:607, 4:613
works
De Cive 2:715, 3:646
Leviathan 1:8, 3:646, 8:213, 11:817, 15:289–290, 20:768
Hobbesian absolutism (Leviathan) 1:10
Hobbesian moment 3:708
Hobbesian social contract 3:647
Hobsbawm, Eric J. 10:493
works
The Age of Extremes: A History of the World 18:663
Primitive Rebels 1:683–684
Hobson, J.A. 11:686
on imperialism 11:686
works
Imperialism: A Study 11:686
The War in South Africa. Its Causes and Effects 11:686
Hobson–Lenin thesis 11:686
Hochberg's method 16:83–84
Hoche, Alfred Erich 13:130
Hochschild, Arlie Russell 11:145–150
academic career, honors, and awards 11:145
on alienation 11:146
on care/caring work 11:147–149
on commodification of emotion 11:146
on division of labor 11:146
on emotion work 11:146–149
on emotional labor 11:146–149
on emotions 11:145–149
impact 11:147–149
major works and publications 11:146–147
on microsociology 11:147–148
on stalled revolution 11:147
theoretical influences of 11:146
works
The Second Shift 11:145–148
The Time Bind 11:145–148
Hodge, Charles 20:287
Hodgskin, T., on division of labor 6:601
Hodgson, Dennis 17:41
Hodiecentrism 2:925
Hodrick–Prescott (H–P) filter 5:651
Hodson Assessment of Phonological Patterns 3:796
Hoe agriculture, in West Africa 1:301
Hoebel, E. Adamson 13:469, 13:590–591
works, *The Cheyenne Way* 13:469, 13:590–591
Hoffman, Roald, on chemistry 3:373–374
Hoffmann-Martinot, Vincent 4:364
Hoffman's theory 19:244–245
Hofstadter, Richard 23:693
works, *Social Darwinism in American Thought* 23:693
Hofstätter, Peter Robert, on semantic differential 21:567, 21:569–570
Hofstede, Geert
on dimensions of culture 5:539
on economic development 5:531, 5:533
on intercultural communication 12:358–361
on organizational cultures 17:392
works
Culture's Consequences 12:358
Hog cycles 22:196–197
Hogan Personality Inventory (HPI) 4:660, 9:233
Hohfeld, W.N. 19:170
Hokiichi, Hanawa 11:50
works, *Gunsho* ruijû 11:50
HOL *see* Higher order logic (HOL)
Hölder's theorem 14:856
Holdouts 23:590
Holford, Castello, works, *Aristopia: A Romance-History of the New World* 14:256
Holism 1:768–769, 2:131–135, 5:677, 7:417, 15:187, 22:985, 24:862
and atomism
arguments 2:134–135
characterization 2:132–134
historical background 2:131–132
types of 2:131
and biomedicine 2:663–664, 2:666
and ethnography 8:179
and joint action 12:843
Maussian 7:416–417
in multisited field studies 16:104–105
of qualitative research 19:685, 19:687
and social ontology 22:502, 22:504
Holistic, developmental, systems-oriented perspective, of personality development 17:881–882
Holistic approaches, for environmental decisions 7:707–708
Holistic care 10:719, 10:722
for psychosomatic disorders 19:484–485, 19:489
Holistic feature, face recognition 8:659–660
Holistic Impact of Bereavement 5:862
Holistic language 8:358
Holistic models 4:443
of parenting 15:925–926
Holistic neuropsychological rehabilitation 2:817
holistic person–context interaction theory 3:445
Holistic perspective
in ethnography 8:185
of indigenous peoples 11:804
Holistic planning frameworks, sustainability 18:198–199
Holistic practice, spirituality 23:288
Holistic processing, face perception 8:652–655
Holistic systems 1:485
Holistic–interactionism approach, personality development 17:881
Holists 22:985
Holl, Karl 3:353–354
Holland, Dorothy 13:72
Holland, John L., model of vocational interests 25:239
Holland, Paul, counterfactuals model 5:91–92
Holland, Randstad 24:867–868
Hollerith cards 23:376–377
Hollerith machines 21:186, 25:565
Holling, C.S. 7:792
Hollinger, David, works, *Postethnic America* 1:620
Hollingshead, A.B., works, *Elmstown's Youth* 7:291
Hollingshead Index 2:477
Hollywood 9:165–169
artists' careers, and social networks 5:468
entertainment 7:673
film and video industry 9:156–157
genres 9:161, 9:163
Hollywood star system 3:276
Holmes, Oliver Wendell, Jr. 13:468, 13:772, 13:819
works, *The Common Law* 7:554, 13:819
Holmes, Stephen 4:705
Holmes and Rahe stress scale 6:201t
Holm's method 16:83–84
Holocaust 2:553–554, 5:619–620, 6:812, 8:237, 11:151–154, 16:358, 18:840, 20:381, 21:371
antisemitism 1:784–787
extermination 11:153–154
Hilberg on 10:870–872

immediate context of final solution 11:152–153
minorities 15:571
and North Americans 1:634
Holocaust and Genocide Studies 10:7
Holocene 7:713
continuities and changes, in Australia and New Guinea 2:232–233
food production 9:300–302, 9:304
greenhouse gas (GHG) 1:723
The Holocene 7:696
Holocene prehistory of central Europe 5:398–399
Holt, D., on habitus 22:334
Holton, Geoffrey 20:446
Holtorf, C., works, *Philosophy and Archaeological Practice* 1:896
Holy figure 18:146
Holy Ghost 9:666
Holy Grail approach 4:395
Holy places 18:147
Holy Spirit, possession of 23:282
Holzkamp School 5:262–264
Homage, feudalism 9:111–112
Homans, George C. 8:482
on norms 17:5
revolution against functionalism by 9:539–540
theory of social exchange 8:482–483
works
The Human Group 16:528, 22:102
Home
and school, discontinuities between 8:751
and work balance, and religion 20:319
Home atmosphere 1:45
Home care, in England 4:345–347
Home Direct Sales industry 7:539
Home environment 7:821
of children 22:947
effect on academic motivation 9:680
families as educational settings 8:732–733
parent language 8:732–733
physical characteristics 8:732
Home front phenomenon 15:520
Home Mortgage Disclosure Act 1975 20:540–541
Home ownership 12:216, 23:661–662, 23:664
Home Rule Act 1:930
Home Start 3:473
Home support initiatives 8:754–756
health and nutrition 8:754–755
home–school partnerships 8:755–756
Homecoming, Weberian social thought 25:466
Home-grown networks 20:488
Homeland nationalism 16:309
Homeless
and violence/criminalization 25:111–115
Homeless women 25:112
Homeless youth 25:112
Homelessness 11:244, 18:685, 25:111
criminalization of 25:113–114
definition 11:159
and domestic violence 25:112
future research 25:114
as perpetrators of violence 25:111–112
social exclusion due to 22:288
in United States 11:159
antecedents of/factors associated with 11:160–161
demographic characteristics 11:160
drug use/abuse 11:161
estimates 11:159–160
future directions for policy and research 11:161–162
growth in shelter/housing capacity 11:161
housing economics 11:161
mental health 11:161
patterns 11:160
policy changes 11:161
poverty 11:161
unemployment 11:160–161
victimization 25:112–113
as victims of violence 25:112
Homelessness Prevention and Rapid Rehousing Program 11:161
HomeNet Study 17:212–213
Homeomorphism 17:299
Homeopathy 10:562
Homeostasis 11:420
Ca^{2+} 12:713–714
and sleep 22:88
Homeostatic equilibrium, loss, disease as 17:2
Homeostatic maintenance, by hypothalamus 11:482, 11:486–487, 11:489
Homeowners association (HOA) 9:614
Homeownership, and residential segregation 20:540–541
Homer 3:771, 3:775–776
works, *Iliad* 4:609–610
Homes
in anthropology 11:155–158
collective beliefs 11:156–157
collective memory 11:157
definition and meaning 11:155
family 11:156–157
final thoughts 11:157
household 11:156–157
kinship 11:156
modern and postmodern homes 11:157
nationalism 11:157
tradition 11:155–156
ways of being at home 11:155–157
anthropology at 1:751–757
in America 1:755
associations and journals 1:756
community 1:753–755
empire 1:751–756
ethnography 1:751–756
ethnology 1:751–756
folklore studies 1:752–754
in France and United Kingdom 1:752–753
future 1:756
great divide 1:751
museums 1:755–756
nationalism 1:751–754
nuseum 1:752–756
peasantry 1:754–756
in rest of Europe 1:753–755
rural 1:751–753, 1:755
specific institutions 1:755
territory 1:751–752, 1:755
urban 1:751–752, 1:755–756
Volkskunde 1:751–754
prosocialty at 19:230–231
and schools, discontinuities between 8:132, 8:751
Home–school links, in children's education 17:500
Homeschooling 21:72, 21:76–77
Homesigns 10:111, 13:307
Homicide offending trends
age-specific 5:194–196
race-specific 5:196
Homicides 4:614–618, 5:161, 5:167, 5:194, 11:163–168, 12:462–467, 12:463t, 12:465t
aggregate-level studies 11:163–165
concentrated poverty 11:163
disadvantage indexes 11:164
family structure 11:164
inequality 11:163
population size and density 11:164–165
racial and ethnic composition 11:164
racial inequality 11:163–164
racial residential segregation 11:164
character contest 11:166
crime-precipitated 10:458
and guns 10:458–459
individual-level studies 11:165–167
age 11:165–166
alcohol and drug use 11:166
circumstances 11:166
gender 11:165
prior criminal record 11:166–167
race and ethnicity 11:165
relationship between victim and offender 11:166
population structure 11:164–165
prevention 11:167
social disorganization 11:163
strain 11:163
weapon 11:166–167
Homiculture, eugenics and 8:248
Homing patterns, Watson's contributions to 25:449
Homing spaces 24:131
Hominid, and language evolution 11:329
Hominidae, modern human cognition of 11:253
Hominin
body, evolution of 2:723–725, 2:724f
evolution of 11:254
occupation of SEA 23:64–65
tool tradition 23:65
Hominin occupation, of South East Asia 23:64–65
Homme civilisé (civil man) 20:768
Homme sauvage 20:768–769
Homo antecessor 8:269

Homo economicus 1:77, 2:437–438, 15:309
Homo educandus 11:586–587
Homo erectus 8:269, 16:739, 23:64
Homo floresiensis 23:65
Homo habilis 16:739
Homo heidelbergensis 8:269
Homo Hierarchicus 6:703
Homo hygienicus 15:121–122
Homo narrans, organizational members as 16:194–195
Homo neanderthalensis 8:269
Homo oeconomicus 3:93, 5:525, 17:9
Homo rudolphensis 2:830
Homo sapiens 2:615, 2:828, 15:338, 16:739
  demographic history of 9:890–894
  diet of 17:458
  and sacrifice 20:862
Homo sapiens sapiens 1:633
Homo sociologicus 1:76, 17:9
Homoerasure 9:657
Homoeroticism 9:651–653
Homoerotophobia 19:753–754
Homogamy 20:364
Homogeneity 4:441, 17:375–377
  in collectivism/individualism 4:207
  versus differences, hybridity 11:450–452
  for effective community life 4:315
  elites 7:398–399
  and globalization 10:235
  and national character 16:233, 16:235
  of Roma 20:754–755
  social context 4:386
Homogeneity analysis 5:8
  meta-analysis 23:911–912
Homogeneous association, model of 16:132
Homogeneous diglossia 6:418
Homogeneous preexisting subgroups 22:882
Homogeneous structure 17:299
Homogeneous trait constructs, in five-factor model 9:245
Homogenization
  of Maghreb 21:751
  nationalist 25:367–368
    fascism, war and 25:367–368
    postwar era and globalization 25:368
    war and genocide 25:367
  of online media audiences 16:794
  in world culture 10:228–230
  of the world's economies 12:584
Homogenizing effect 1:879
Homogenous cell populations, transgene expression in 17:271–272
Homohysteria, and sports 9:657–658
  navigation 9:658
Homology 7:459
Homomorphisms 13:58–59, 15:751, 17:298
  structural equivalence 23:622
Homonoia 3:644
Homonormativity 10:793–794, 21:809, 21:811
  and social work 21:805–806
Homonymy 16:185
Homophile movements 9:633, 9:636, 15:110–111, 24:551–552
Homophilia 21:751
Homophily 9:443, 9:449, 16:564, 22:702, 10:808t
  adolescent friendship 9:437
  female managers 8:885
Homophobia 10:790–795, 16:863, 16:865, 18:841, 21:812
  and alcoholism 15:112–113
  attitudes and expressions of 10:790
  bisexuals 10:790, 10:793
  desire 10:791–792
  discourse 10:791–792
  explanations for 10:791–792
  feminism 10:793
  gay 10:790–794
  gay movement 10:793
  gender transgression 10:792
  heterosexuality 10:790–792
  homosexuality 10:790–794
  identity 10:791–792
  institutional
    and permissible prejudice 10:792–793
  internalized 10:791
  and intersectional queer theory 19:766–768
  in Latinos 9:630
  lesbian 10:790–794
  of LGBTQ youth 21:87–88
  oppression 10:794
  power 10:794
  problematizing 10:793–794
    homonormativity 10:793–794
  queer theory 10:792
  social movements 10:793
  terminology 10:790
  transgender 10:790–793
  and youth, shifting intersections of 9:658–659
Homophone intrusions 23:270–271
Homophones 3:509–510, 4:92–93, 6:735
Homo-related fossil 11:293
Homosexual 17:949, 17:951–952, 21:768–769
Homosexual behavior 21:785–786
Homosexual Liberation Fronts 13:859–860
Homosexual men 15:169
  fecundity of female maternal relatives of 21:783
Homosexual paradises 9:637–638
Homosexuality 9:656, 9:770, 10:790–794, 11:545, 18:536–538, 19:753, 20:840, 21:779–780, 21:792
  and colonialism 21:731
  concept of 9:632
  contemporary issues in research 3:889–890
    homosexuality and mental illness 3:889–890
    sexual orientation change efforts 3:889
    theories on etiology of homosexuality 3:890
  early history 3:884
  early 1800s 3:885–886
  in Germany
  see also Germany
  late 1800s 3:885
  late 1900s 3:887–888
    declassification 3:887–888
    subsequent editions of DSM 3:888
  and mental disorder 3:888
  mid-1800s 3:885
  mid-1900s 3:886–887
    American psychoanalytic influence 3:886
    DSM 3:886
    DSM-II 3:886–887
  orientation 21:785–788
  in post-Soviet Russia 21:742–743
  in preevolutionary Russia 21:741–742
  psychiatric classification systems 3:888–889
  psychiatric conceptualization 3:885
  removal of 21:773
  and reproduction 20:453
  and social work 21:802, 21:804–806
  in Soviet Russia 21:742
  studies 3:882
  in Western culture 3:884
Homotopic inhibition theory 12:417
Homozygosity 11:727
Homozygotes, definition 1:85
Homuncularism, and personality 17:912
Homunculus man 24:15
Honduras 22:624
  sociocultural aspects 1:621–625
Hones, G. 7:166
Honest businessman, ideal of 3:16
Honesty 12:637
  behavior-dependent 12:637–638
  objectivity of research 17:93
Honesty-Humility 11:169
  agreeableness 11:169–170
  altruism 11:169–170
  Big Five 11:170, 11:173
  crime 11:170–171
  Dark Triad 11:170–171
  future directions 11:172–173
  HEXACO model of personality 11:169–173
  interaction with situations 11:171–172
  lexical studies 11:169–170
  and life outcomes 11:170–171
  methodological issues 11:172
  personality dimension explaining active cooperation 11:170
  psychology 11:173
  social dilemma 11:171, 11:173
  theoretical and empirical background 11:169–170
Honeybees
  memory and learning in 15:155–159
Honeymoon, definition of 5:335t
Hong Kong Cantonese–English bilinguals 3:511
Hong Kong flu 18:174
Hong Kong Institute of Asia–Pacific Studies (HKIAPS) 22:637–638, 22:640
Hong Kong International Arbitration Center 12:470

Honigmann, John J., works, *Handbook of Social and Cultural Anthropology* 7:590
Honneth, Axel 5:294, 11:175–180, 20:203, 20:205, 22:830–831
academic and intellectual achievements 11:176–178
critical theory 11:176
Hegelian roots 11:176–177
idea of recognition 11:177–178
struggles for recognition 11:178
on critical theory 5:298, 5:304, 5:308–309
output 11:175–176
on politics of recognition 20:45–47
on reification 20:203, 20:205
works 11:179
The Critique of Power 11:175
Disrespect 11:175
The Fragmented World of the Social 11:175
Freedom's Right 11:175
Geist 11:176
The I in We 11:175
Integrity and Respect: Principles of a Conception of Morality Based on the Theory of Recognition 11:175
Kampf um Anerkennung 11:175
The Pathologies of Individual Freedom 11:175
Pathologies of Reason 11:175–176
Phenomenology of Spirit 11:176
Redistribution or Recognition? A Political-Philosophical Exchange 11:175
Reification 11:175
Reification: A New Look at an Old Idea 11:175–176
Social Action and Human Nature 11:175
The Struggle for Recognition 11:175
Honor 11:181–184
in anthropology 11:181–182
ascribed 11:896–897
class and gender 11:182–183
comparative considerations 11:182
dueling 11:181–183
feminist studies 11:183
and gift 10:140–142
honor crimes 11:183–184
in interpersonal relationships 5:430–431
in medicine 8:64–65
Mediterranean 11:181–184
migration of 11:183
murder and multiculturalism 11:183–184
negative evaluation of 10:125
Pitt-Rivers on 11:181–183
sexuality 11:181, 11:183
violence 11:181–182
Honor crimes 16:863
Honor killings 10:195, 11:181, 11:183–191, 15:425, 21:793
conflict face negotiation 11:188–189
conflict face-negotiation theory 11:187
culture 11:185–190
defining and contextualizing 11:185–187
dimensional complexity 11:185
intercultural conflict dimension 11:186–187
moral dimension 11:186
sociocriminal approach 11:185–186
dilemma 11:185–186, 11:189–190
family 11:185–190
future directions 11:189–190
image 11:186–187, 11:190
izzat 11:185–187, 11:189
justice 11:185–186, 11:188–190
marital infidelity 11:189
morality 11:186
multilayered social–ecological framework 11:187–188
murder 11:185, 11:188–189
patriarchy 11:188–189
respect 11:185–187, 11:189–190
social ecological macrolevel and exolevel analysis 11:188–189
social ecological mesolevel and microlevel analysis 11:189
social–ecological integrative lens 11:188–189
violence 11:185–186
Honorary degrees 2:335–336
Honorifics 6:54–55
Honoris causa 2:335–336
Honor-related violence *see* Honor killing
Hont, Istvan, works, *Jealousy of Trade* 3:64–65
Ho'o pono pono 17:620
Hoogheemraadschap 15:678
Hooke, Robert 14:809
Hooker, Evelyn, homosexuality studies 3:882
Hooker, Joseph 9:571
Hooker, M.B., on legal pluralism 13:812
Hooking up 17:486–487, 17:490, 21:842–843
Hooks, Bell 15:493
Hookups 11:192–196
attention to 11:195
behaviors 11:192–193
and gender inequality 11:193–194
versus relationships 11:194
social class differences 11:194
see also Casual sex
Hooliganism 21:736
Hoover, Herbert 18:322
Hopcroft, Rosemary 22:896–897
Hope 11:197–201
adaptation function of 11:198
approaches and measures of 11:197–198
and change 12:756
as emotion 11:197–199
and goals 11:197–200
and health 7:498
and hopelessness, similarities and differences between 11:199
increasing 11:200
Islamic 12:754–756
magic as source of 14:436
for oppressed, liberation theology as 14:32–35
religion and 20:265
Hope Scale 11:197
HOPE VI *see* Housing Opportunities for People Everywhere (HOPE VI) Program
Hopedale Community 2:147
Hopelessness 6:199, 6:207–208, 11:197–201
decreasing 11:200
definitions and measures of 11:198–199
and hope, similarities and differences between 11:199
Hopelessness Scale for Children 11:198
Hopewell exchange system 4:602
Hopfield networks 4:719, 4:722
Höpfl, Harro 12:760–761
Hopi (language) 25:575–577
Hopkins, A.J. 11:687–688
Hopkins, Mark 10:495
Horizon 2020 22:643–646, 22:643t
Horizontal cells 25:158
Horizontal cultural transmission 5:403
horizontal equating 20:250
Horizontal expansion 12:481
Horizontal proliferation problem (arms control) 1:976
Horkheimer, Max 3:776, 4:339, 5:294, 5:499, 8:215, 9:380–384, 9:458–459, 14:735, 20:203
on critical philosophy and social theory 5:295
on critical theory 5:295–296
early research program 5:294–295
works
Dialiectic of Enlightenment 5:295
Eclipse of Reason 5:295
Hormonal treatments, benefits and risk of 13:866
Hormone systems, interaction of HPA axis with 11:478–479
Hormones
and aggression 1:378
biochemical properties of 16:649t
classification and biochemical properties of 16:648–649
as indicators of neurotransmitter and hypothalamic activity 16:654–655, 16:654t
interactions between 16:653
regulation by limbic–hypothalamo–pituitary processes 16:650f
and sexual motivation 21:762
and sexual orientation 21:774–775
and social behavior 8:466
and tactile processing 24:16
Hornberger, N., works, *Indigenous Literacies in the Americas: Language Planning from the Bottom Up* 13:355
Horney, Karen 9:459–461, 20:669–671, 21:129, 25:625
psychotherapy theories 19:503–504
Hornstein, Frances 15:112, 15:114
Horowitz, Daniel 8:139–140
works, *Ethnic Groups in Conflict* 8:139–140
Horowitz, Irving Louis, works, *Genocide: State Power and Mass Murder* 10:7

Horse race random utility models 5:901–905
Horseman 1:965–966
Horses, in warfare 25:415–416
Horticultural societies 11:202–207
  agriculture 11:202, 11:204–205
  agronomic regimes 11:202–204
  anthropology 11:202, 11:206–207
  chiefdoms 11:204–205
  farming systems 11:205–207
  future directions 11:206–207
  horticulture 11:202
  and inequality 11:894–895, 11:897
  livelihood strategies 11:206
  peasants 11:202, 11:204–206
  sociocultural orders 11:204–206
  sustainability 11:206
  tribes 11:204–205
Horticulture
  Australia and New Guinea 2:231
  terminology 11:202
Horton, Byrne J.
  Dictionary of Modern Economics 7:590–591
Horton, J.A., works, *West African Countries and People* 16:267–268
Horvat, Branko 16:231
Horvitz–Thompson estimator 20:886
Hoskins, W.G., works, *The Making of the English Landscape* 11:19
Hospitable environments 11:700–701
Hospital almoner 10:717t
Hospital Anxiety and Depression Scale (HADS) 6:204
Hospital research, negotiated orders in 17:411
Hospitalized dying 5:866–867
Hospitals 10:717t
  behavior in other nations 11:213
  community ownership 11:210–211
  cost containment 11:208, 11:211–213
  employment for social workers 10:716
  governance 11:210–211
  institutional environment of 11:208–209
  medical social work 10:719
  organization of work in 11:212
  organizational heterogeneity 11:210
  physician–organization structures 11:211–212
  professional norms 11:208–209
  psychiatric 6:42–46
  psychosocial aspects 11:208–214
  quality of care in 11:212–213
  readmission among individuals with affective disorder 8:299–300
  regulation 11:208–209
  reimbursement 11:208–209, 11:211
  responses to hybrid environments 11:210
  technical environment of 11:209–210
Host countries
  gains, migration 15:438
  refugees 20:113–114
Host national identity 9:435
Host society, tourism 24:467
Host–cleaner fish relationship 22:894
Hostile attributional bias 1:375
Hostile environment, and discrimination 6:531–532
Hostile noncompliance, and agreeableness 1:463
Hostile takeovers 23:491
Hostile working environment, and sexual harassment 21:722
Hostility
  and adolescent health 1:124
  and aggression 1:374–375
  and cardiac disease 20:188
  and health 17:793
  and health behaviors 7:498
  health risks of 7:497–498
  nonverbal behaviors 23:704–705
  in sexism 21:695–698
  verbal behaviors 23:704–705
Hostos, Eugenio María de 13:802
Host–pathogen coevolution and genetic conflict 8:449–451
HOT *see* High-occupancy toll (HOT) lanes
Hot cognition 7:416–417
Hot intelligences 17:750–751
Hot memories 16:198–199
  of traumatic events 24:434–435
Hot societies 1:748
Hot spots, and guns 10:459–460
Hot-and-cool framework 10:711
Hotelling, Harold 11:215
  book reviews 11:215
  career at Columbia University 11:216
  career at University of North Carolina 11:216–217
  contribution to multivariate statistical analysis 11:217–218
  early life of 11:215
  Hotelling's rule 14:138, 15:555
  mathematical focus of his research 11:218
  of originality and imagination 11:218
  papers in statistics and in mathematical economics 11:215–216
  statistical research of 11:216
  training as mathematician 11:218
Hotlines, for legal information 12:925
Hourly earnings 13:191
Hours 13:190–195, 13:197–198
House Committee on Patents 17:586
House names 16:184
House of Commons 17:516
House of Prayer 16:807
House of Representatives 4:649
  legislative procedure 4:651
Household, Income and Labour Dynamics in Australia (HILDA) survey 5:846, 15:599
Household activity ratio 20:773
Household archaeology 11:219–223
  domestic 11:220–221
  emergence of 11:219
  gender 11:221–222
  inequality 11:221
  microscale analysis 11:220
  modeling household activities in 11:220–221
  post-processual archaeology 11:221
  power 11:220
  residence 11:221
  sources of inspiration for household 11:220
  at small scale 11:219–220
  theory 11:221–222
  trends in 11:221–222
Household data, population censuses 3:291
Household demography 8:725–730
  classification by subject matter 8:725–726
  data issues 8:726
  headship rate 8:728–729
  kinship 8:725, 8:727–728
  LIPRO household model 8:728
  microsimulation 8:727–728
  models 8:726–729
    agent-based models 8:729
    dynamic *versus* static models 8:728–729
    macromodels *versus* micromodels 8:727–728
    severing the link strategy 8:727
  ProFamy model 8:728
  relationships 8:725–729
  SOCSIM model 8:728
  unit of analysis 8:726
Household economics 11:224
  applications 11:225
  Becker on 11:224–225
  demographic economics 11:224
  economic growth 11:225
  education 11:224–225
  family 11:224
  fertility 11:224–225
  health 11:224–225
  history 11:224
  household production 11:224–225
  integration 11:225–226
  labor supply 11:224–225
  modeling assumptions 11:224–225
  New Home Economics (NHE) 11:224–225
  policy, law, and technology 11:225
  time use 11:224–225
Household expenditure surveys 5:733
Household factors, and income inequality 11:742
Household finances 15:743–744
Household panel studies, of income and living conditions 5:845–847
Household panel surveys 5:844
Household production 11:224–225
Household production–consumption model 9:63
Household registers 6:140
Household slavery 22:57
Household surveys 5:760
Household wealth 25:450–452, 25:451t
Household work
  and fatherhood 8:840
  and time use 24:379–381
Household-level data 5:792–793
  linking with remote sensing data 10:67
Households
  in anthropology 11:228–232
    changes over time 11:230–231
    classification 11:229
    coresidence 11:228, 11:230–231

critical approaches 11:231
culture 11:228–231
definition 11:228–229
developmental cycle 11:229
division of labor 11:230–231
domestic family 11:228–231
extended family 11:229–231
feminist theory 11:231
gender 11:228, 11:230–231
genealogy 11:228–229, 11:231
kinship 11:228–231
locality 11:228
private sphere 11:231
public sphere 11:231
residence 11:228
rethinking 11:231
social stratification 11:230
transnational 11:228
variation 11:229–230
burglary 19:165
census 3:302–305
and cohabitation 4:119–120
databases 5:742–749
and demography 6:179
and family 8:809–813
and home 11:156–157
labor statistics 5:799, 5:801
and marriage 8:119
poor, rationality of 3:407
production, family theory 8:823
separation, and labor 13:158
structure of 8:739–740
House-hunting problem 12:102
Housekeeping functions 4:416
Housekeeping model 23:862
Housewifization 10:813
Housework 9:757–759
in dual-career families 14:606–609
and fertility recovery 9:776–778
Housing 24:908, 24:911–913
in China 2:66–68, 24:976
and cities 11:233–238
capitalism 11:237
commodification 11:234
connections 11:234–235
housing classes 11:236–237
informal housing 11:234
naming the parts 11:233–234
neighborhoods 11:233–235, 11:237–238
owning 11:233–234
privatization 11:234–235
renting 11:234, 11:237
residential mobility 11:236
segregation 11:233, 11:235–237
tenure 11:233–234, 11:237
understanding the distribution 11:235–237
urban managerialism 11:237
and gender 9:709–712
and LGBT aging 13:866–867
migrants 15:417
of minority ethnic elders 15:442
segregated 24:915–916
and segregation 21:408–409
statistics
history 5:778
and urbanization 5:778–779
in urban planning 19:822–823
urban sociology 24:927, 24:929–932
Housing Assistance Supply Experiment (HASE) 8:532–533
Housing bubble 3:580–581, 3:583
Housing capacity, of shelters in United States 11:161
Housing census 3:296, 3:299–300
Housing Choice Voucher Program (1983) 18:736
Housing classes 11:236–237
Housing economics 11:239
allocation 11:241–243
demand 11:241
markets 11:241–243
policy 11:243–244
real estate economics 11:243
supply 11:239–241
in United States 11:161
Housing Opportunities for People Everywhere (HOPE VI) Program 4:326, 16:451, 16:812
Housing poor 18:741
Housing prices in the Sunbelt 3:581
Housing satisfaction paradox 6:886
Housing unit method 6:164, 6:166–167
Houthakker, Hendrik, on division of labor 6:604
HOV *see* High-occupancy vehicle (HOV) lanes
Hovland, C.I., model of attitude change 2:170–173
Howard, Ebenzer, works, *Garden City of Tomorrow* 16:438–439
Howard, John 2:241
Howard, Sir Ebenezer 24:863
garden cities 1:219
Hoxha, Enver 18:700
Hoyle, Sir Fred 9:224, 20:288
HP *see* Harmonious passion (HP)
HPA *see* Hypothalamic-pituitary-adrenal (HPA) axis
HPG *see* Hypothalamic-pituitary-gonadal (HPG) axis
HPI *see* Hogan Personality Inventory (HPI)
HPL *see* Human placental lactogen (HPL)
HPO *see* HazingPrevention.Org (HPO); Hypothalamic-pituitary-ovarian (HPO) axis
HPSG *see* Head-Driven Phrase Structure Grammar (HPSG)
HPT *see* Hypothalamic-pituitary-thyroid (HPT) axis
HPV *see* Human papillomavirus (HPV)
HRAF *see* Human Relations Area Files (HRAF)
HRC *see* Human Rights Council (HRC)
HRF *see* Hemodynamic response function (HRF)
HRL *see* Human rights law (HRL)
HRM *see* Human resource management (HRM)
HROs *see* High-reliability organizations (HROs)
HRQL *see* Health-related quality of life (HRQL)
HRS *see* Health and Retirement Study (HRS); United States Health and Retirement Survey (HRS)
HRT *see* High reliability theory (HRT)
HRV *see* Heart rate variability (HRV)
HRW *see* Human Rights Watch (HRW)
Hs scale 15:567
HSBS *see* History and Sociology of the Biomedical Sciences (HSBS)
HSLS:09 *see* High School Longitudinal Study of 2009 (HSLS:09)
HSM *see* Heuristic-systematic model (HSM)
HSR *see* Human speech recognition (HSR)
HSS *see* Henry Sweet Society for the History of Linguistic Ideas (HSS)
5-HT *see* Serotonin (5-HT)
H-tone in morphemes 23:717
HTP *see* Human Terrain Program (HTP)
5-HTP *see* 5-Hydroxy-tryptophan (5-HTP)
HTS *see* Human terrain system (HTS)
HTT *see* High threshold theory (HTT)
5-HTT *see* Serotonin transporter (5-HTT) gene
5HTTLPR (polymorphic region) 22:495, 23:646
Hubbert, M. King 18:367
Hubbert model 14:139
peak oil 15:556–559, 15:556f–557f
Hubble, Edwin 20:288
Huber estimator 20:731
theoretical justification 20:734–735
Hubris, national 16:312
Hubristic pride 21:478
HUD *see* Department of Housing and Urban Development (HUD)
Hudson, Michael, works, *Arab Politics: The Search for Legitimacy* 16:383
Hudson, N., works, *Systema naturae* (The system of nature) 7:722
Hudson, William Henry 14:256
Hudson–Kreitman–Aguade (HKA) test 11:291
Huff, V., works, *Convicted but Innocent: Wrongful Conviction and Public Policy* 25:789
Huff model 1:848
Hughes, Everett C. 2:401, 2:404, 20:672, 21:265, 25:678
Hughes, H., on gay/lesbian tourism 9:639, 9:641–642
Hugh-Jones, Stephen 24:291
works, *Barter, Exchange, and Value* 7:14
Hughlings-Jackson, John 16:683, 18:47–48
Hugo, G., on multiculturalism in Australia 2:240
HUI *see* Health Utilities Index (HUI)
Huizinga, Johan 7:378
works, *The Waning of the Middle Ages* 20:426
Hukou system 12:437–438, 24:859–860, 24:973–974, 24:977

Hull, Clark L. 13:690–694, 14:813, 17:609, 19:414
Hull House Maps and Papers study 7:561
Hull House tradition 19:677
Hulu 25:81
Human action
  Bandura's view on 18:125–126
  unintended consequences of 24:744–745
Human actors, hierarchies and markets 10:842
Human agency 1:708–709, 1:711–712, 7:603
  life course theory 14:83
  and linked lives 14:232
Human appropriation of net primary production (HANPP) 6:859, 13:241–242
Human behavior 3:263, 13:571, 25:128
  adaptationism 8:400–402
  evolutionary economics 8:430, 8:432, 8:434–435
  evolutionary psychology 8:400–403
  gene–culture coevolution 8:403–404
  human behavioral ecology 8:400–401
  intellectual background 8:399–400
  and interests 12:392–394
  scientific descriptions of 21:194
  social change 22:148
  trends 8:404
Human Behavior and Evolution Society 8:463
Human behavioral ecology 8:400–401, 11:246–251
  areas of application 11:248–250
    foraging theory 11:248
    marital strategies 11:249
    reproduction 11:249–250
    resource transfers 11:248–249
  inclusive fitness 11:246–248
  life history theory 11:247–248
  optimal foraging 11:247–248
  optimality models 11:248
  origins and early history of field 11:246–248
    relationship to other fields using evolutionary approaches 11:246–247
    relationship to traditional social and behavioral sciences 11:247
    theoretical armamentarium 11:247–248
  and wealth flows 9:35
Human being 21:449–450, 21:452–453
Human biological variation, in new genetics 9:1018
Human biology 2:656–657, 2:661
  cultural evolution 5:395–396
Human body 15:119–121
  history of 2:728–734
Human bonding, and suicide 23:677
Human brain
  characteristics of 2:828–831
    asymmetry 2:831
    brain organization and reorganization 2:830–831
    brain size, absolute and relative 2:828
    encephalization 2:828–830
  cholinergic projections of basal forebrain in 3:336f
  evolution in human phylogeny 3:336t
  morphology, organization and localization of hypothalamus within 11:482–483
  natural computation in 2:28–29
  ontogeny of 3:334f
  studying LTP/LTD in 14:362–363
  unique cognitive capacities of 4:100
Human Brain Mapping conference in Boston 9:487–489, 9:489f
Human capacities, transforming 9:1017
Human capital 3:90, 7:123–125, 7:149, 7:261, 9:761, 13:260, 17:612, 20:294, 22:114–115, 23:823, 25:336–337, 25:495–496
  acquisition, effect of minimum wages on 15:564
  and cash transfers 3:215–217
    education 3:217
    health 3:216–217
    nutrition 3:215–216
  children 9:60, 9:62
  of children 8:821–823
  and criminology 5:244–249
  economic growth 7:40–42
  and ethnic identity 8:107, 8:109–111
  explanation 9:759–760
  fertility 9:20–21
  Fisher's theory 9:218
  and internal migration 12:436–438, 12:440
  and language 13:260, 13:262–263
  model
    for effect of beauty on labor market 2:386
    of internal migration 12:445
  and obesity 17:70–72
  and poverty 18:743–744
  promotions 19:138–139
  spillovers, and international immigration 12:517
  and time use 24:370
  universities 18:308
Human capital model 15:597–598
Human carrying capacity 3:157–158
  Earth's 3:158
Human cerebral cortex regions 5:24f
Human characteristics of control behavior 4:813–814
Human chorionic gonadotropin (HCG) 16:652–653
Human cognition 7:241
Human collective action 8:443, 14:253
Human colonization, Oceania 17:447–448
Human communication 16:194–195
  intentionality as foundational feature of 12:347–348
Human communities *see* Communities
Human condition (action systems) 1:390, 1:539–540
Human consciousness 14:675, 23:291–298
Human constructions, values 25:33, 25:35
Human cooperation, evolution of 5:391, 11:259–260
  as an evolutionary puzzle 11:260
  large-scale cooperation 11:262–265
    cultural group selection 11:264–265
    indirect reciprocity 11:262–263
    punishment 11:263–264
  small-scale cooperation 11:260–262
    direct reciprocity 11:261–262
    kin selection 11:260–261
Human cortical areas 5:24
Human cost of revolutions 20:640–641
Human culture, spoken expression in 17:274–275
Human development 2:192, 3:445, 3:451, 3:456, 10:634, 11:271, 13:70, 22:149–152
  in Arctic 1:927
  Bruner on 2:875, 2:877
  conceptualization and definitions 11:267
  criteria, problem of 11:271–272
    culture 11:272
    nature 11:272
    presumed gap bridging 11:271–272
  cultural matters in 3:395–396
  disabilities 11:267, 11:269
  and flow 9:261–262
  and health 11:267–269
  infancy 12:7–13
  and personality 17:909–910
  and psychodynamic thought 22:761–762, 22:767
  psychological conceptions 11:271–275
  statistics on 5:775
    future of empirical data 5:779
    Human Development Index 5:777–778
    Inequality-adjusted Human Development Index 5:778
    life expectancy 5:775, 5:777–778
    life satisfaction 5:775
    well-being 5:775, 5:777, 5:779
    World Population Prospects 5:778
  studies 6:307–313
  successful development 11:274
    coping 11:273
    efficacy 11:273
    resilience 11:273–274
  and temperament 24:180–183
  theories of individual agency in 14:103–104
  theory of 18:126–130
    child development 11:276–277
    development science 11:276–278
    individual–context relations 11:276–280
    life span 11:276–280
    plasticity 11:277–280
    positive development 11:280
    relational developmental systems theory 11:278–279
    optimization to developmental science within 11:279
    social justice 11:276, 11:280
  Vygotsky's theory of 25:316–321
Human development cycles 18:582–583
Human development index (HDI) 5:777–778, 5:795, 6:310, 7:918, 9:23

Human differences
in intelligence 12:309
theorization from North America 17:22
Human dignity 2:632, 3:689, 11:368, 11:370, 19:658, 24:443
Human Dimensions of Global Change project 20:422
Human dispersals, migration 15:472, 15:477–479
Human disturbance process, geoarchaeology 10:54–55
Human diversity
apportioning 19:826–827
comparison with other species 19:829
estimated fractions of 19:827t
Human ecology 6:853, 6:855, 11:283–288, 22:254–255, 22:263, 22:722
definition 11:283–284
environmental sociology 7:799–800
rubrics 11:284–286
adaptation 11:283, 11:285–286
ecological demography 11:284–287
environment 11:286
evolution 11:283
organization 11:284–285
population 11:285
space 11:283–284
sustenance organization 11:284–285
technology 11:285–286
sociological 11:284
urban 24:926–931
and urban life 24:831–832
Human economy 7:19
Human effectiveness, counseling psychology 5:71
Human embryology 2:116–120, 2:117f–118f
axis formation and positioning of gut, heart, and central nervous system 2:116–118
central nervous system development, anterior-posterior partitioning 2:118
cortical specification 2:118–120, 2:119f–120f
Human embryonic stem cell (hESC) 9:979, 23:440
Human engineering 11:880–881
Human–environment relationships 11:405–409
case study and comparative method 11:405
explanatory status of comparative case studies 11:405
qualitative and quantitative aspects 11:405
comparative case study method 11:408
comparative case study strategies 11:406–408
analogs 11:407–408
community 11:406
events 11:407
gradients 11:407
perspective 11:406
place 11:406
scale analysis 11:407
structure 11:406
system–subsystem 11:406–407
theory 11:408
time 11:407
transects 11:407
tools for analysis 11:408
Human epidemics 6:480
Human ethology 14:369
Human evolution 14:260–261, 14:263, 17:458–459
embodied-capital theory of 9:28–34
modern human cognition 11:252–258
prehistoric art 18:824–826
Human evolutionary genetics 11:289–296
adaptations and detection of positive selection 11:293–294
EDAR 11:293–294
genetic record 11:292–293
LCT 11:294
population genetics 11:289–290
population genetics inferences 11:290–292
Human extensibility 24:417–418
Human factors 11:297–305
background 11:297–299
cognitive systems engineering 11:298, 11:300–301
group level analysis 11:301–302
individual levels of analysis 11:299–301
cognitive ergonomics 11:299–301
motivational factors 11:301
physical ergonomics 11:299
macroergonomics 11:302–303
organization level analysis 11:302–303
scope of 11:298f
situation awareness 11:299–302
in transportation 24:625
Human Factors and Ergonomics Society (HFES) 11:298–299
psychology 11:880–881
Human Fertility Database (HFD) 5:739–740
Human flourishing, market allocation 14:541–542
Human freedom 18:687
Human genetics 6:443
Human Genome Diversity Project (HGDP) 9:901, 9:963–964, 10:23, 11:306, 16:759
antecedents 11:306–307
biocolonialism 11:308
bioethics 11:308
gene patenting 11:307
genetic diversity 11:309
genetic salvage 11:309
human rights 11:308
identity politics 11:308
indigenous rights 11:307–308
informed consent 11:308
initial controversy 11:307
legacies 11:309
old questions, new contexts, unexpected outcomes 11:307–309
population 11:306–309
population genetics 11:306
research 11:308–309
scientific racism 11:306–307
Human Genome Project (HGP) 2:619, 2:625, 2:627, 2:677–678, 5:742, 6:440, 7:864, 8:43, 9:969, 9:975–976, 9:991, 9:994, 9:1016–1019, 10:32, 11:311, 11:318–319, 17:590, 21:187–188
assessing 11:315–317
concise history 11:313
ELSI 11:313
and genomics 2:621
origins 9:223
prehistory 11:311–313
Human genomics, in Asia 11:318–322
Human geographers 10:85–86
social theorists 10:85–86
spatial scientists 10:85–86
Human geography 10:84–85, 10:244, 11:18, 11:76, 23:82
mathematical models in 14:785–790
Human growth data, functional data analysis 9:514
Human health 7:815
Human Heredity and Health in Africa initiative (H3Africa) 2:624
Human historicity 11:114
Human history, and civilization 3:726–727
Human immunodeficiency virus (HIV) 2:699, 7:815, 9:634, 9:642, 10:805–806, 11:138–144, 12:35–36, 15:111, 20:870, 20:872–873, 21:793, 21:797–801
adolescents and youth 11:138–140
impact and epidemiology 11:138
improving future prevention efforts 11:140
intervention efficacy 11:139
selected predictors of risk 11:139
clinical and immunological progression of illness 1:490–491
demographic microsimulation 15:345
health 10:629–630
heterosexual men 11:142–143
impact and epidemiology 11:142
improving future prevention efforts 11:143
intervention efficacy 11:142–143
selected predictors of risk 11:142
and homosexuality
in China 21:736–737
infection, treatment of 1:491–492
influence on gay/lesbian activism in Latin America 13:861
in Latinos 9:630
men who have sex with men 11:141–142
impact and epidemiology 11:141
improving future prevention efforts 11:142
intervention efficacy 11:141–142
selected predictors of risk 11:141
policy and ethical issues 1:493–494
and sexuality 21:804
in South Africa 21:748
testing issues 1:490
treatment of infections 1:491–492
women 11:140–141

Human immunodeficiency virus (HIV) (*continued*)
impact and epidemiology 11:140
improving future prevention efforts 11:140–141
intervention efficacy 11:140
selected predictors of risk 11:140
see also HIV/AIDS
Human immunodeficiency virus (HIV)-associated dementia (HAD) 1:492
Human immunodeficiency virus (HIV)-associated neurocognitive disorders (HAND) 1:491–492
Human impacts
of disasters
demographics 22:754
economic implications 22:755
political implications 22:755
social implications 22:754–755
on Oceania 17:449
anthropogenic landscapes 17:449
conversion of natural rainforests 17:449
declines in biodiversity 17:449
Human intelligence 12:303–308
at birth 16:481–482
controversial topic in philosophy and psychology 16:481
developmental psychobiology of infant social mind 16:481–482
development 12:306–307
Piaget theory 12:307
post-Piagetian theories 12:307
Vygotsky theory 12:307
measurement 12:307–308
distribution of IQ scores 12:308
historical background 12:307
IQ test 12:307
multiple intelligences 12:306
successful intelligence 12:306
theories 12:303–306
biological theories and research 12:304–306
brain size 12:305
cognitive theories 12:304
cognitive-contextual theories 12:306
genetic and heritability studies 12:305–306
implicit theories 12:303
neural efficiency 12:304–305
psychometric theories 12:303–304
Human intelligence tasks (HITs) 8:554
Human interaction 22:267
and sociality 22:856, 22:858
Human intimacy 17:948
Human judgment and memory, limits of 13:494
Human justness 22:396
Human language
and empathy 7:544
evolution of 11:323–328
brain connectivity 11:326
cerebral cortex 11:323
dorsal stream for sensorimotor integration and control 11:324–326
future studies 11:326
human auditory and language cortex 11:324
internal models 11:325
invariance and categorization of speech sounds 11:324
macaque monkey 11:323–326
premotor cortex 11:324
species-specific communication sounds in nonhuman primates 11:323–324
ventral stream 11:324, 11:326
Human language technology 5:811–813, 5:816
Human laws 12:138
Human learning, theories of 11:622
Human linkage, Fisher's studies 9:223
Human manipulation of nature 7:594
Human mate choice 11:335
based on facial characteristics 11:337–338
based on major histocompatibility complex 11:336–337
evolution of 11:340
mating strategies 11:340
men's mating strategies 11:340–342
long-term mating 11:341–343
short-term mating 11:340–341, 11:343–345
women's mating strategies 11:342–345
genes and 11:335–336
inbreeding 11:336, 11:338
inclusive fitness 11:338
kin selection 11:335
mating preference 11:335–337
phenotype matching 11:338
self-resemblance 11:338
sexual imprinting 11:337–338
twin study 11:335
Human memories 11:709
quantitative memory models 15:160–162
Human Microbiome Project 1:762–763
Human migration
analysis of 9:900
economics of 18:597–598
and language families 13:337, 13:340–341
Human mind
architecture of 12:314f
Lévi-Strauss on 13:905–907
and social psychology 22:526
Human mobility 15:638–639
Human models, of anxiety 1:808–809
challenge studies 1:809
state and trait anxiety 1:808
Human morality, Konrad on 14:368
Human Mortality Database (HMD) 5:739–740
Human motivation and the problem of institutional design 19:537–538
Human nature 2:131–132, 11:368, 11:370
bounds of 2:438–439
illness/wellness models of 17:911
and peace/nonviolence 17:624–629
prosocial behavior 1:580
and social justice 22:397–398
Human needs 22:396
and motivation 15:936–937
underlying self-concept development 21:472
Human origins and social evolution 8:386
cultural evolutionism 8:386
social Darwinism 8:386
Human papillomavirus (HPV) 3:78–79
Human participants
in experimental laboratories 8:552
in research, regulatory protections for 4:252–257
Human performance, alternative explanations for 6:710–711
Human placental lactogen (HPL) 16:652–653
Human plutonium experiments 13:562
Human populations
genomic estimates of Wright's Fst-statistics 19:828t
technological lineages and 24:130–131
Human poverty index 7:918
Human prehistory
evolution of language 8:375
Human primates, goal-oriented behavior among 10:268–274
Human Problem Solver (project) 14:814
Human productivity 7:274
Human psychology
instinct in 12:176–177
and profit 19:112–113
Human races, listing 19:825–826, 19:826t
Human rape-adaptation 19:887
Human rationality 12:721–723
Darwinian responses 12:721–722
ecological rationality 12:722
informational format 12:722
preference structure 12:722
dual process accounts of human reasoning 12:722–723
problems with specific experiments 12:721
Human reasoning 5:934
dual process accounts of 12:722–723
Human relations
crisis, in postsocialist societies 18:699
in organizational development 6:297
Human Relations Area Files (HRAF) 4:412, 7:145, 9:2, 19:752–753, 23:276
Human relations movement 11:351
I–O psychology 11:882
Human Relations School 3:5–7, 17:390
Human remains 2:600
recovering and excavating 9:335
treatment of 9:948–949
Human resource development cooperation 3:503
Human resource management (HRM) 3:131–133, 11:880, 12:817, 22:14
and performance appraisal 17:716
psychology of 11:348–352
adjustment 11:348, 11:352
behavioral sciences 11:351
conspiracy doctrine 11:349
development 11:348, 11:352
early industrial psychology 11:350
emergence of free collective bargaining 11:349

evolution of 11:348–352
human relations movement 11:351
industrial revolution 11:349
managing change 11:348, 11:352
private industry's approach to personnel management 11:350
retention 11:348, 11:352
scientific management 11:349–350
social legislation and court decisions 11:351–352
staffing 11:348, 11:352
US Civil Service Commission 11:350
virtual teams 11:352
responsibilities and objectives 11:348
and social context of PA 17:718
Human resources 13:203
for organizational environmental sustainability 19:83
Human rights 3:674–679, 3:686–689, 5:874–875, 10:206–207, 10:210, 10:225, 10:739, 11:367–375, 12:364–365, 13:420, 13:423, 13:721, 13:818, 13:821, 16:322–323, 17:620, 17:648–652, 18:389–391, 18:649, 20:457, 20:459–460, 20:464, 20:675, 22:396, 22:890–891, 25:123
abuse 25:639
and advocacy 1:223
and American/French Revolutions 1:597, 1:599
anthropology of 11:360–366
autism spectrum disorders 2:279–280
-based social work and gender mainstreaming 24:755
basic issues 3:675–676
betweenness 11:377–378
and business ethics 3:18
of children 3:486–487
and citizenship 3:652–653
covenants and documents on 11:356–357
International Covenant on Civil and Political Rights 11:356
International Covenant on Economic, Social, and Cultural Rights 11:356–357
cultural rights 5:491–497
defining 11:353–354
universality and indivisibility 11:353–354
definitions 3:674–675
and disaster 6:481
and disasters 22:756
and discipline 11:376–377
as a distinct discipline 11:377–378
enlightenment in 7:660–662
and ethics 1:737
and ethnic cleansing 8:86
and feminist legal theory 8:923–924
and forensic archaeology 9:336–337
and genetics 9:969–972
as God's rights 14:34
governments accountability 13:721–722
historical background 3:674
history and 11:376–377
history of 11:367–374
absolute *versus* relative rights 11:368
domestic *versus* international jurisdiction 11:368–369
first- *versus* second- *versus* third-generation 11:369–370
individuals *versus* groups as rights-holders 11:368
men *versus* all humans 11:368
moral *versus* positive 11:367–368
progress and failure 11:370–373
universality *versus* relativity 11:370
Human Genome Diversity Project 11:308
and inequality 11:898
instruments 11:375–376
international 11:375–377
international law 11:375–376
in international practice 3:677
international relations/politics and 11:377
international social work student exchanges 12:565
in Latin America 24:512, 24:514–515
and law 13:536–537, 13:539–540
legal perspective 11:375–376
and liberalism 14:14–15
Maroon communities 14:595
Marshall on 14:628–629
model of disability 13:498–499
and modernization 15:710–711
and morality 13:487–488
movement 11:376–377
museums 16:146
nation-state and 3:677–678
nature of 3:675
and nongovernmental organizations 16:858
norms 2:108
pluralism and tolerance 18:238, 18:241
and political repression 18:446, 18:448–451
in practice 3:676–677
and privacy 18:950–951
promotion of 13:720–722
and punishment 19:656–659
and radical social work 19:873
and recognition politics 20:44, 20:46
refugees 22:769–771, 22:773
relationship to state 3:675
and religion 3:575–576
research censorship 3:281–282
and sexuality 21:803–805
HIV 21:804
religion 21:804–805
socially enforced conformity 21:804
state-enforced conformity 21:804
slaves/slavery, history of 22:54
and social justice 13:721, 24:755
and social life 11:377
and social work 11:353–359, 13:720, 22:521–522
context 11:357–358
cultural relativism 11:357
environmental social work 11:355
ethics 11:353, 11:356, 11:358
green social work 11:355
Human Rights Watch 11:357
International Association of Schools of Social Work (IASSW) 11:353–354, 11:357–358
International Federation of Social Workers (IFSW) 11:353, 11:357–358
Millennium Development Goals 11:357
overview 11:353
positive human rights 11:353
social justice 11:353
sociolegal perspectives 11:376–377
sociology and 11:377
of street children 23:537
technology and 13:521
terrorist threats and 3:678
transnational activist network 11:377
treatment research and access 8:31, 8:33–35
Universal Declaration of Human Rights 11:354–356
challenges to 11:355–356
cultural, economic, and social 11:355
political and civil 11:355
provisions of 11:355
violations 2:110–111
and war crimes tribunals 25:385, 25:387
of women 6:793, 20:457
and world republic 25:743–744
Human rights, anthropology of 11:377
critique 11:365
culture 11:363
engagement 11:365
epistemology 11:364
future directions 11:365
history 11:360–363
1999–2007 11:362–363
2007–Present 11:363
1948–1980s 11:361–362
1980s–1999 11:362
UDHR, creation of 11:361
longue durée 11:360–361
overview 11:360
power 11:364
transnational advocacy 11:364–365
Human rights abuses, militarization and social construction of 10:636–637
Human Rights Act 1998 4:700, 15:197
Human Rights Campaign 9:630
Human Rights Council (HRC) 11:796–799
United Nations 24:753
Human rights education, and international organizations 12:539–540
Human rights law (HRL) 13:530
and domestic violence 6:629–630
torture and 24:442
Human rights litigation in domestic courts, transnational tort 18:502
Human rights movement, and commemoration 4:237
Human rights revolution 5:228
Human Rights Watch (HRW) 3:677, 11:357, 24:512
Human rights-based procedural safeguards 12:958
Human sciences 11:380, 18:625

Human sciences (*continued*)
critique 11:383–384
and cultural history 5:422
Dilthey and 6:424–426, 11:382–383
and education 7:174–175
history of 11:380–385
and labor history 13:171
methodological problems 11:383–384
neo-Kantianism 11:381–382
philosophical background 11:380–381
sociology of 11:384
Human security 17:619–620
Human services 11:695
risk assessment 20:693–694
and social work 14:665–666
Human settlement, changing sizes and locations 24:961
Human social behavior 8:460–461
Human sociality
and behavioral evolution 18:891, 18:893
character of 22:856–858
phylogeny of 22:894–895
Human socialization 25:670
Human society 3:160
Human sociobiology 22:899
Human spatial orientation, neural basis of 11:386–391
Human speech recognition (HSR) 23:262–263
Human sphere, African 1:292–294
Human studies, phenomenology 18:15
Human subjects 8:37–41, 12:608
definition of 4:253
Genographic Project 10:25
protection of
and ethics 1:736
in United States 8:48–52
research
in ELSI program 8:44
placebo-controlled trials 18:169
in social psychology experiments 22:536
Human Terrain Program (HTP) 25:34–35
Human terrain system (HTS) 8:182, 11:392, 15:486, 25:380–381
conjunctures 11:394–395
consequences 11:397–398
contexts 11:394–395
critics 11:395–396
diversity of critiques 11:395–396
from ethics to politics 11:396–397
impacts 11:397–398
practice of 11:392–394
recursion 11:394–395
theory of 11:392–394
in United States 11:392–399
Human territoriality 18:377
Human thinking, evolution of 25:576
Human trafficking 10:194, 10:207–210, 22:44–45, 22:48, 24:512
estimates of 10:209
exploitation 10:207–211
exploiters 10:210
globalization and 10:210–211
combating 10:211–212
and human rights 10:206–207, 10:210
labor trafficking 10:207–208, 10:210
sex trafficking 10:207–208, 10:210
transnational component of 10:207, 10:210–211
victims and survivors 10:209–210
of women and children in Latin America 24:512–516
Human unity 14:435
Human values, Wolfgang Köhler on 13:125
Human variation 9:939–940
history of 1:758–759
Human visual cortex 25:156
Human welfare, and intellectual property 13:523
Human well-being 16:322–323
Human–animal relations
and sociality 22:858
Human-centered automation 2:296
Human-competitive results, of genetic programming 2:43–44
Human–computer interaction (HCI) 11:400–404
and design process 11:403
development of usability evaluation methods 11:402
as discipline 11:403–404
influx of novel approaches 11:403
origins in human factors and cognitive psychology 11:400–401
Human–computer interface 25:133
Human-demographic traits 8:414–417
optimality and constraints 8:414–416
shared by all human populations 8:416
variation among humans 8:416–417
Humane warfare 11:397
Human-ecology integration, of cognition and behavior 1:428
Human–environment interactions 16:334–339, 22:728
Human–environment relations, in North America 17:19–20
Human-induced disasters 22:753
Humanism 2:645, 13:704, 19:409, 20:428, 22:589–590, 22:994–995
Lévi-Strauss on 13:904, 13:908
and use of rhetoric 15:289–290
Humanist critique of naturalism 20:204
Humanista 20:427
Humanistic anthropology 1:735
Humanistic Buddhism 2:892–893
Humanistic geography 5:279, 18:149, 22:303
place 18:149
placelessness 18:149
Humanistic movement 7:234–235
Humanistic philosophy, Veblen's 25:54
Humanistic psychology 5:76, 8:517, 20:739–740
early studies 19:504
Humanistic psychotherapy 4:725, 4:730–732
Humanistic spirituality 23:285–289
Humanistic theories 22:817
Humanistic universalism 17:158
Humanistic-existential psychotherapy 4:731
Humanistic-phenomenological approach, to regional geography 20:124
Humanitarian activities, relevance of social work to 1:972–973
Humanitarian disaster 3:715
Humanitarian equal treatment 12:363
Humanitarian interventions 10:310
Humanitarian law 11:371–372
and war crimes tribunals 25:382, 25:384–385
Humanitarian protection, and cash transfers 3:218–219
Humanitarian supply chains 23:712
Humanitarian technologies, and ethnic cleansing 8:88
Humanitarianism 11:397, 14:13, 25:124
and adoption 1:137
and human development 6:312
and nongovernmental organizations 16:858
Humanities 7:869, 7:871, 22:606, 25:343–344, 7:869t
crimes against 25:382
cultural analysis 9:365
hermeneutics in 10:788
unity of 13:907–909
Humanities discipline 22:255
in medicine 15:120
Humanized task approach, organizational ergonomics 11:303
Human–landscape relationships 10:55–56
Human-made capital 6:852–853
Human-made disasters 6:478, 6:481
capacities 6:479
economic impacts of 6:478–479
and poverty 6:478–479
resilience 6:479
vulnerability 6:478–479
Humans 4:394–396
arrival, in Australia and New Guinea
dating of 2:229–231
routes of entry 2:230
average man concept 23:424–425
body, evolution of 2:723–725
demographic history of 9:890–894
demography beyond 8:417–418
empathy aspects specific to 7:544–545
ethics for biomedical research involving 8:78–82
fear-potentiated startle in 8:862–864, 8:863f
fertility and parental investment among 9:29–30
and food, basic relationship between 9:296
food choice, determinants of 9:297
genetics research 10:39
as information-processors 2:775
neonatal bonding in 16:471–472
origins and social evolution 8:386
cultural evolutionism 8:386
social Darwinism 8:386
population history
evolution of language 8:367
serial order memory in
effects of development 21:640–641

effects of focal brain lesions 21:639–640
effects of neurological and psychiatric illness 21:641
neuroimaging 21:640
sources of anxiety in 1:805–806
genetic and biological disposition 1:806
life events 1:806
social and environmental influences 1:806
statistical methods based on measured genotypes in 9:934–936
uniqueness 4:395
Human-to-human network 4:486
Human-to-IT network 4:486
Humble novice 4:771
Humboldt, Alexander von 7:722, 7:791–792, 13:226–227, 25:279–285
on aesthetics 25:279, 25:282–284
American voyage of 25:280–284, 25:281f
on anthropology 25:279, 25:284
in Berlin 25:279, 25:284
on colonialism 25:279–280, 25:282–285
contributions to cartography 25:279–284
contributions to climatology 25:279–282
contributions to electrophysiology 25:279–280
contributions to geography 25:279–282
contributions to mineralogy 25:279
contributions to phytogeography 25:279
contributions to plant geography 25:279
contributions to political sciences 25:279
in Cuba 25:280, 25:282, 25:284–285
on environmental studies 25:284–285
on Hispanoamerica 25:282
on history of globalization 25:279
on history of science 25:284–285
on Mesoamerican cultures 25:282–284
in Mexico 25:280–282, 25:284–285
in Paris 25:280–284
publishing 25:280–282
on politics 25:282
in Prussia 25:279–280, 25:284
Russian–Siberian voyage 25:284
on scientific networks 25:284
studies, new editions, American travel diaries 25:284–285
travel literatures of 25:279, 25:284–285
on wars of independence 25:282
and Wilhelm von 25:279–280
works
Asie Centrale 25:284
Cosmos: A Sketch of a Physical Description of the Universe 7:722, 7:750, 13:227, 25:284
Essay on the Geography of Plants 25:280–282
Personal Narrative 7:748, 7:750
Political Essay on the Island of Cuba 25:282–284
Political Essay on the Kingdom of New Spain 25:282–284
Views of the Cordilleras and Monuments of the Indigenous Peoples of America 25:282–284
Voyage aux régions équinoxiales *du Nouveau Continent* 25:280
youth, education, and European travels 25:279–280
Humboldt, Wilhelm von 4:790–794, 14:241, 25:279–280, 25:286–289
contributions to anthropology 25:287–288
on imagination 25:287–288
on language 25:288
comparative study 25:288–289
on language and ethnicity 13:252–253
on liberalism 25:286–287
on linguistic project 25:287
on philosophy of language 25:286
on politics 25:286
on sexual difference 25:287–288
on typology 25:288–289
and University of Berlin and Bildung 25:287
works
The Diversity of the Human Language-Structure 25:288
On the Kavi Language on the Island of Java 25:288
Ueber den Nationalcharakter der Sprachen 4:791
on worldviews 25:288
Humboldtian typology 14:209, 25:288–289
Hume, David 3:90, 3:269, 4:337, 4:751, 4:844, 6:241, 6:590, 7:99, 7:567–569, 7:722, 8:74–75, 11:410–415, 18:524, 19:410, 19:666, 20:286, 22:697, 22:836–837
as *Chargé d'affaires* 11:411
on civilized monarchies 11:413
critic of religion 11:414–415
on economic behavior 11:413
on economics and politics 11:413
on experimental method of reasoning 22:989
general philosophical and psychological principles of 11:411–412
on historiography and history 11:413–414
legacies of 1:765–766
life and writings of 11:410–411
moral and political theory of 11:412–413
on necessity of allegiance 11:412–413
on origin of justice 11:412
on practical reasoning 18:779
on principle of uniformity of nature 11:411
on punishment and reward 11:411–412
on rules for possession and transfer of property 11:412
on small states 22:109
on virtue ethics 25:147, 25:150
works
Enquiry Concerning the Principles of Morals 11:410–412
History of England 11:414
Moral and Political 11:410
Philosophical Essays Concerning Human Understanding 11:410
Treatise of Human Nature 11:410–411
Humean problem of induction 7:569
Humes' Law 1:498
Humor 11:416–419
comparative perspective 11:416–417
and contradiction 11:418
faces of 11:417–418
domains of joking 11:417–418
genres 11:417
functions and meanings 11:418–419
and incongruity 11:416–417
and laughter 11:416–419
and ludicrous 11:416
nature of 11:416–417
anthropology 11:416–417
study 11:416
and relativism 11:416–417
tricksters 24:653–655
Humor: International Journal of Humor Research 11:416
Humoral pathways, cardiac response 23:539
Humoralism 15:119
Humorism 7:720–721
Humphrey, C., works, *Barter, Exchange, and Value* 7:14
Hundi 12:519
Hungarian 3:186
Hungarian Civil Liberties Union v. Hungary (2009) 9:396
Hungarian Goulash Communism 20:635
Hungarian Household Panel Study (HHP) 5:846
Hungarian Institute for National Biology 8:220–221
Hungarian revolution of 1956 18:700
Hungarian Society of Racial Hygiene and Population Policy 8:220
Hungary 3:316
anthropology 1:754
and property 19:156–157
Hunger 9:307, 9:309–310, 11:421
and food preferences 9:296
global 9:285
neural basis of 11:420–422
neurobiology of 11:421–422
in United States 9:285–286
Hunnicutt, B., works, *Work without End* 24:401
Hunt, L., works, *The Book That Changed Europe* 7:662
Hunter, Floyd, works, *The Community Power Structure: A Study of Decision Makers* 4:363
Hunter, Phillip, works, *Encyclopedia of Anthropology* 7:590
Hunter–gatherers 2:601, 11:202
archaeology 11:423–427
anthropology 11:424
behavioral ecology 11:425
biases 11:426
characteristics 11:423–424
current trends 11:425–426
evolutionary ecology 11:425
fishers 11:423
interpretive archaeology 11:426
mobility 11:424–425
postprocessualism 11:426

Hunter–gatherers (*continued*)
prehistory 11:426
processualism 11:425–426
in Australia and New Guinea 2:231
in East Africa 6:773–774
and life history evolution 9:28–30, 9:32
in North America 17:15–17, 17:19–20
warfare among 4:600–601
Hunting
and fishing 9:226–227
Hunting and gathering societies (HGS) 7:616, 11:428–430, 23:38
anthropology of 11:428–431
band society 11:429
collective appropriation 11:428–429
contested meanings 11:428–429
foragers 11:428, 11:430
giving environment, animism, relational epistemology, and conservation 11:430
immediacy 11:429–430
indigenous peoples 11:428–430
ontologies 11:430
relations with others 11:430
sharing of resources 11:429–430
in Arctic 1:928–929
specialization 6:244
Hunting hypothesis 11:423
Hunting societies, and inequality 11:895
Huntington, Ellsworth 7:724
works, *The Pulse of Asia* 7:751
Huntington, Samuel Phillips 11:432–437
biography of 11:432–433
influence of 11:437
works 11:433–437
American Military Strategy 11:433
American Politics: The Promise of Disharmony 11:435
Authoritarian Politics in Modern Society: The Dynamics of Established One-Party Systems 11:432
The Bases for Accommodation 11:432
Changing Patterns of Military Politics 11:432
The Clash of Civilizations and the Remaking of World Order 11:43, 11:435–436, 23:26
The Common Defense: Strategic Programs in National Politics 11:432
The Crisis of Democracy: Report on the Governability of Democracies to the Trilateral Commission 11:432, 11:434–435
Culture Matters 5:617
The End of History and the Last Man 11:437
No Easy Choice: Political Participation in Developing Countries 11:432
Political Order in Changing Societies 11:433–434
Political Power: USA–USSR 11:432
The Soldier and the State: The Theory and Politics of Civil–Military Relations 11:433
The Third Wave: Democratization in the Late Twentieth Century 11:435
Who Are We?: The Challenges to America's National Identity 11:436–437
Huntington controversy 17:298
Huntington's disease (HD) 2:450–451, 2:619, 6:444, 10:36–37, 16:847, 23:252
Hurbert, Jeanne 4:387
Hurricane Katrina 6:102, 7:807, 9:947, 9:949–950, 20:561
Hurst, James Willard 11:438–440, 13:517
biography of 11:438–439
on expansion of middle-class energy 11:438
impact of 11:439–440
scholarship 11:439
administrative law 11:439
man-made law 11:439
markets 11:439
monetary policy 11:439
social history of law 11:439
works
The Growth of American Law 11:439
Law and Economic Growth: The Legal History of the Lumber Industry in Wisconsin, 1836–1915 11:439
Law and the Conditions of Freedom 11:440
Law in Society 11:438
The Law of Treason in the United States 11:438
Hurston, Zora Neale 1:743
on race in North America 17:23
Hurtado, A.M., works, *Ache Life History: The Ecology and Demography of a Foraging People* 8:411
Hussein, Saddam 24:451–452, 25:398–399, 25:838
martyrdom of 20:331
Husserl, Edmund 2:631, 6:427, 10:484–487, 11:441–444, 14:382–384, 18:11, 19:314, 19:391, 19:413, 20:345, 21:146–147, 21:149
contexts of discussion/influence of 11:443
on crisis of European science and life-world 11:442–443
from descriptive to constitutive phenomenology 11:442
from early foundational work to antipsychologism 11:441–442
on epoché 11:442
on ideation 11:442
impact on social and behavioral sciences 11:443–444
on intentionality 11:441–442
life and academic career of 11:441
on phenomenology 18:11
on prepredicative experience 11:443
on transcendental ego 11:442
on *wesensschau* 11:442
works
Ideen I 20:665
Logische Untersuchungen 19:391
writings, subjects, approaches, and contexts 11:441–443
Hutcheson, Frances, on division of labor 6:602–603
Hutchins, Robert Maynard 21:971
Hutchins Commission 14:693
Hutchinson, Jonathan 12:774
Hutchison, Terence 7:100
Hutterian Brethren 4:266–267
Hutterite demography 11:445–447
and biodemography 11:446
fertility 11:445
control 11:446
and culture 11:445–446
as natural fertility standard 11:445–446
age patterns 11:446
Coale and Trussell's 'M' and 'm' indices 11:446
Coale's 'I' indices 11:445–446
limitations 11:446
micro- and macrosimulations 11:446
Hutterites 2:147
Hutton, James 9:567–568
Huxley, Aldous, works, *Brave New World* 2:153–154
Huxley, Julian 8:386
Huxley, Thomas Henry 5:683–684, 8:233, 9:571, 18:83
program 5:682
Huysmans, J.-K. 15:685
Hwang, K.K.
theoretical model of face and favor 11:792
works
Foundations of Chinese Psychology: Confucian Social Relations 11:792
Hy scale 15:567
Hyaluronic acid 16:476
Hybrid channels 2:861–862
Hybrid colonies, in sub-Saharan Africa 23:657
Hybrid courses 6:567
Hybrid cultural functionalism 9:533
Hybrid education 6:568
Hybrid environments, hospital responses to 11:210
Hybrid identities 22:7
Hybrid institutions, global governance 10:159
Hybrid knowledge, and common sense 4:3
Hybrid lab-field experiments 9:130
Hybrid legal systems 13:804
Hybrid models
artificial intelligence 2:38–39
architectures and representations 2:38–39
learning 2:39
concept learning 4:540
Hybrid national service programs 25:272
Hybrid null hypothesis significance test 12:655
Hybrid organizations 22:170–171, 22:173
Hybrid policing 18:249
Hybrid popular mass cultures 24:590–594
supranational citizenship 24:592–594
transnational citizenship 24:590
transnationalization of citizenship
via migration 24:590–591
via participation 24:591–592
via stakeholdership 24:591–592

Hybrid psychology, and mind–body dualism 15:547
Hybrid regimes, democratic transitions 6:97
Hybrid spaces 12:92
Hybrid subjects 14:227
Hybrid threats 25:419
Hybrid war crimes tribunals 25:385–386
Hybrid-isolation (HI) model 9:887
Hybridity 11:448–452, 11:534–535, 23:82
  and creolization 5:118–119
  and diffusion 6:404
  future 11:452
  globalization 11:450
  homogeneity *versus* differences 11:450–452
  hybridization 11:450
    terminology 11:448–449
  Latour on 13:414–418
  mestizaje, syncretism, transculturation, creolization 11:449–450
  modernity 11:450
  multiculturalism 11:448–452
  of new constitutions 13:601
  and postcolonial geography 18:644
  religion 11:449, 11:452
  of social life
    in Central Asia 3:320
  and syncretism 23:883–884
Hybridization 4:704–707, 11:450, 24:144
  terminology 11:448–449
  and world culture 10:228
Hybridization cycles 11:452
Hybrids 6:896–897
Hyde, Edward, works, *The History of the Rebellion and Civil Wars* 4:763
Hydra-headed monster 1:790
Hydro diplomacy 25:445–446
Hydrologic cycle 25:437
Hydrology 25:441
  archaeology 10:101
Hydrometeorological/climatological hazards 6:479–480
  flooding 6:480
  global warming 6:480
  wildfires 6:480
11ß-Hydroxysteroid dehydrogenase type 2 (11ß-HDS2) 10:466
5-Hydroxy-tryptamine (5-HT) 16:749, 16:751
5-Hydroxy-tryptophan (5-HTP) 16:495
Hygiene 18:174, 18:176
Hyman, Herbert 8:571, 19:677
Hymes, Dell, on communicative competence 4:296
Hyndman-Rizik, N., on migration in Australia 2:241
Hyogo Framework for Action 2005–2015: Building the Resilience of Nations and Communities to Disasters plan 6:479
Hyperactive agency detection device (HADD) 20:390
Hyperarticulation tendencies 23:715–716
Hyperautomobility 2:304, 2:306–307
Hyperbole, and Internet 12:601–602
Hyperbolic discounting utility theory 17:782
hyperbolic geometry 25:198
Hypercompetition, marketing 14:556–557
Hypercorrection (language) 13:285–286
Hypercorrectness (language) 9:698
Hyper-cortisolism 11:478t
Hyperdiversity, and migration 15:424–425
Hyperexcitability
  and epilepsy 7:854
  kindling 13:34
Hyperfactualism 18:469
Hypergeometric distribution 6:582–583
Hypermedia 19:59–60
  and learning 21:544
Hypermobility
  labor, division of 13:160
Hypernorms 3:17
Hyperobjectivist 20:95–96
Hyperosmia 17:199t
Hyperpersonal communication 17:214
Hyperpersonal model
  of computer-mediated communication 4:507
    channel 4:507
    feedback 4:507
    receiver effects 4:507
    sender effects 4:507
  of social media 22:422–423
Hyperplexia 23:332–333
Hyperpolarizing variable 25:161
Hyperprolactinemia 21:24
Hyperpunitiveness 6:553
Hyperrealism, and reification 5:245–246
Hyperreality 18:673–674
Hyperreligions 16:806
Hypersegregation 20:542, 21:412
  and residential segregation 20:540, 20:542
Hypersomnia 22:74–75, 22:79, 22:82
Hyperspace 18:673–675
Hyperspace analogue to language model (HAL) 21:581–582
Hypersynchrony, and epilepsy 7:852–853
Hypertension
  psychosocial aspects 11:453–457
    autonomic nervous system 11:454–455
    biopsychosocial models 11:454–455
    blood pressure reactivity 11:453–454
    blood pressure variability 11:455
    cardiovascular reactivity, research on 11:454
    coping styles 11:454
    diagnostic issues 11:455
    direct and indirect evidence 11:453
    emotion 11:453–454
    mood 11:453–455
    personality 11:453–455
    resilience 11:454
    stress 11:454–456
    treatment 11:455–456
  religion and 20:266
Hyperthreats 16:863
Hyper-unemployment 24:818–819
Hyperventilation 23:539
Hyperviolence 6:553
Hypnagogic hallucinations 10:499
Hypnoid states 9:415
Hypnopompic hallucinations 10:499
Hypnosis 11:458–464
  ability to hypnotize 11:461
  animal hypnosis 11:465, 11:468
  areas of application 11:462
    dentistry 11:462
    medicine 11:462
  context of 11:461
  definition 11:458
  forensic 9:354–355
  history 11:462–463, 11:465–466
  hypnotic phenomena 11:459–461
    cognitive 11:461
    motor-kinesthetic 11:459–460
    sensory-affective 11:460–461
  hypnotic trance 11:459
  hypnotizability 11:458
  neural basis of 11:465–471
    consciousness, and cognitive control 11:467–470
    fMRI 11:466–470
    future research 11:470–471
    role of suggestion 11:465–471
  psychiatry of 11:472–475
    definition 11:472
    essential elements and putative mechanisms 11:474–475
    history 11:472–473
    medicine 11:473–474
    psychotherapy 11:474
    therapeutic applications 11:473–474
  techniques and rituals 11:461–462
  therapeutic 11:466–467
  and unconscious mental activity 24:723
Hypnosis prone 11:458
Hypnotherapy 11:462
  in gastrointestinal disorders 9:609
Hypnotic training 11:473
Hypnotic trance 11:459
  benefit of 11:459
  in clinical practice 11:459
  ratification of 11:459
  techniques and rituals 11:461–462
Hypnotism 12:791–792, 22:526
  with psychopharmacotherapy 19:460
  for schizophrenia 21:19
  studies
    Bernheim's 19:502
    Braid's 19:502
    Charcot's 19:502
    Janet's 19:502
    Liebeault's 19:502
Hypnotizability 11:458
  and fantasy proneness 11:458
  and genetics 11:458
  and hypnosis prone 11:458
HYPO system of legal reasoning 13:787
Hypoactive sexual desire disorder 21:762–763, 21:712t
Hypochondriacal symptoms 5:628
Hypochondriasis 23:11, 23:11t
  explanatory style 8:598
Hypocoristics 16:184–185

Hypo-cortisolism 11:478t
Hypodermic needle model of media influence 2:216
Hypoglycemia
  awareness 6:338–339
    impaired 6:338–339
  diabetes and 6:337
  fear of 6:339
Hypoglycemia fear survey (HFS) 6:338t
Hypomania 19:467
  of bipolar disorders with 2:684
  treatment of 14:487
Hyposmia 17:199t
Hypostasis, and reification 20:203–204
Hypothalamic activity, hormones as indicators of 16:654–655, 16:654t
Hypothalamic dopamine pathway 17:953
Hypothalamic-pituitary-adrenal (HPA) axis 3:552, 7:606–607, 10:463, 10:466, 11:476–481, 11:485–486, 17:133–134, 17:138, 19:445–446, 23:565, 23:577, 16:653f
  basic concepts and regulation 11:476–477
  behaviors 11:478–479
  and bipolar disorder 14:486
  in bipolar disorder 7:607
  corticosteroids 11:476–477, 11:479–480
  determinants of individual differences in functioning 11:478–481
    effects of adverse environmental factors 11:479
    effects of genetic factors 11:479–480
    gene–environment interactions 11:480–481
    interactions with other hormone- and neurotransmitter systems 11:478–479
  development changes in 23:578–579
  diverse functions and clinical correlates 11:477–478
  dysfunction 7:607
  emotions 11:477–480
  epigenetics 11:479
  glucocorticoid receptor 11:477–480
  hyper- and hypo-cortisolism 11:478t
  mineralocorticoid receptor 11:476–477
  neuroendocrinology 16:652
  psychobiology of 23:577–578
  psychopathology 11:477–478, 11:480
  regulators in 23:579
  sleep 22:85
  and stress 23:571–572
  stress reaction 11:476–477
Hypothalamic-pituitary-gonadal (HPG) axis 11:486
Hypothalamic-pituitary-ovarian (HPO) axis 10:462–464, 10:467
Hypothalamic-pituitary-thyroid (HPT) axis 11:485
Hypothalamo-pituitary-gonadal (HPG) axis 7:606, 17:133–134
  in bipolar disorder 7:608
  neuroendocrinology 16:652
Hypothalamo-pituitary-thyroid (HPT) axis
  neuroendocrinology 16:652
Hypothalamus 7:606, 11:482–490, 14:126, 11:483f
  adrenal gland 11:484–486, 11:488–490
  circumventricular organs 11:487
  connections 11:483–486
    afferent and efferent neuronal connections 11:486
    hypothalamic–pituitary adrenocortical axis 11:485–486
    hypothalamic–pituitary gonadal axis 11:486
    hypothalamic–pituitary thyroid axis 11:485
    pituitary gland 11:483–485
    posterior pituitary 11:486
  and eating 11:420–422
  functions 11:486–490
    regulation of body temperature and immune system 11:489–490
    regulation of feeding 11:488
    regulation of sexual and reproductive behavior 11:489
    regulation of ultradian, circadian and infradian rhythms 11:488–489
    regulation of vegetative systems and adaptation to stress 11:486–487
    regulation of water balance 11:487–488
  gonads 11:484–486
  histamine, and wake 22:89
  homeostatic maintenance 11:482, 11:486–487, 11:489
  hormones related to pituitary gland 11:485t
  hypocretin, and wake 22:89
  leptin-action within 11:488f
  major hypothalamic nuclei 11:483t
  morphology, organization and localization within developed human brain 11:482–483
  neuroendocrine rhythmicity 11:482
  ontogeny 11:482
    pituitary gland 11:482
  sectioned in sagittal plane 11:484f
  sexual dimorphism 11:482, 11:489
  and sexual motivation 21:759, 21:761–762
  and sexual orientation 21:775
Hypothecation 19:190, 19:192–194
Hypothek 21:404
Hypothèque 21:403
Hypotheses 8:521
  involving moderation 15:672
Hypothesis of contamination 12:365
Hypothesis testing 11:491–499, 15:273–275, 19:725–730
  acceptance region 11:491
  a-level 11:493
  alternative hypothesis 11:491
  assumptions, role of 11:496
  asymptotic distribution 16:886–887
  asymptotics 11:493
  in causal analysis 3:248–249
  confidence intervals 11:496–497
  construction of tests 11:491–492
    Bayesian test 11:492
    intersection–union test 11:492
    likelihood ratio tests 11:491–492
    union–intersection test 11:492
  desiderata 19:728–730
  in econometrics 7:9–10
  empirical social research (ESR) 7:558
  errors of Type I and II 23:425–426
  evaluation of tests 11:492–493
  goodness-of-fit tests in common use 19:726
  history 11:494–495
  idea of 11:494–495
  invariant 12:708
  limitations of Neyman–Pearson approach 11:498
  loglinear models 16:128
  multiplicity of 11:500–505
    family size 11:503
    fixed and variable 11:503
    illustrative settings, error rates, and procedures 11:500–501
    mathematization *versus* realism 11:504
    multiple comparisons 11:501–502
    multiple determination 11:500–502
    NAEP Trial State Assessment 11:501–503
    plausible defaults for c 11:504
    prototype 11:503–504
    qualitative status 11:503
  Neyman on 16:826
  Neyman–Pearson formulation 11:495–496, 11:498
  null hypothesis 11:491–492
  power function 11:492
  problems in interpretation 11:497–498
    misinterpretations 11:497–498
  p-value 11:492–493
  rejection region 11:491–493
  role of null hypothesis 11:495
  score test 11:493
  for significance tests 21:960
  size 11:492–493
  in statistics 23:425–426
  student's *t*-test 11:495–496
  tests for independence/differences in common use 19:726–728
    monofactorial designs 19:726
    multifactorial designs 19:727–728
  type I error 11:491–492
  type II error 11:491–492
  Wald test 11:493
Hypothesis-generating methods 19:730–731
  desiderata 19:731
  exploratory methods in common use 19:730–731
Hypothetical cohort, demography 6:113
Hypothetical cohort measures, of marriage/divorce 6:118–119
Hypothetical limiting relative frequency 19:26–27
Hypothetical reasoning 14:308–310
Hypothetical survey 20:887
Hypothetical value function 19:264f
Hypothetical weighting function 19:264f
Hypothetico-deductive (HD) method 5:271, 5:273–274, 8:103, 10:782–783, 14:306–307, 14:310–311, 15:160, 23:424

hypothetico-deductive theory 11:890
Hypothyroidism 19:464–465
Hypoxic-ischemic events, and transient global amnesia 24:566–567
Hyppolite, Jean 9:368
Hysteria 11:506–511, 22:527
  abnormal illness behavior 11:509
  alternative terms and nosological status 11:507
    diagnosis and classification 11:507
  Breuer's studies 19:502–503
  clinical presentations 11:507–509
    Briquet's syndrome 11:508
    complex integrative models 11:509
    dissociative symptoms 11:508
    epidemic or mass hysteria 11:508
    hysterical personality 11:508
    neurobiology/pathophysiological mechanisms 11:508–509
    immunological mechanisms 11:508
    laterality 11:508
    neural circuitry 11:508
    neuroendocrine dysregulation 11:509
    sex hormone 11:508
    neurological conversion symptoms 11:507–508
    pain and other bodily symptoms 11:508
  epidemiology of 11:509
  Freud's studies 11:506–507, 11:510, 19:502–503
  functional neurological disorder 11:507
  initial concepts 11:506
  investigation and treatment 11:510
  psychoanalysis 11:506–507, 11:509–510
  relevance of behaviors 11:509
  somatoform disorder 11:507–509
  symptoms and outcome 11:509–510
  treatment, early 19:502–503
  use of term 23:10
HYV *see* High-yielding varieties (HYV)

I

I, detection of 21:440–441
I Can Do Program 10:672
I Mafiusi della Vicaria 14:430
I2 test 15:269
IACHR *see* Inter-American Commission on Human Rights (IACHR)
IADB *see* Inter-American Development Bank (IADB)
IADLs *see* Instrumental activities of daily living (IADLs)
IAIA *see* International Association for Impact Assessment (IAIA)
IALE *see* International Association of Legal Ethics (IALE)
Iambic pentameter, in language 13:266–267, 13:274
IAPT *see* Improving Access to Psychological Therapies (IAPT) program
IARR *see* International association of radical reflexivity (IARR)
IASC *see* Inter-Agency Standing Committee (IASC)
IASP *see* International Association for the Study of Pain (IASP)
IASR-B5 *see* Revised Interpersonal Adjective Scales: Big Five (IASR-B5)
IASSIST *see* International Association for Social Science Information Service &; Technology (IASSIST)
IASSW *see* International Association of Schools of Social Work (IASSW)
IAT *see* Implicit association test (IAT); Intracarotid amobarbital test (IAT); Intracarotid sodium amytal testing (IAT)
Iatrogenesis 11:585
Iatro-physics, man-machine model of 15:119–120
Iatro-technical concept 15:120
IB *see* Inclusive business (IB); International Baccalaureate (IB)
IBA *see* Independent Broadcasting Authority (IBA); International Bar Association (IBA)
Ibañez, G., on gay-related racism toward Latinos 9:629–630
IBD *see* Inflammatory bowel disease (IBD)
IBE *see* International Bureau of Education (IBE)
Iberian peninsula
  Baroja on 3:150–152
Ibero-Americana 20:918
IBLT *see* Instance-based learning theory (IBLT)
IBM *see* International Business Machines Corporation (IBM)
IBM punch card 3:304–305, 23:376–377
IBM SSP (Scientific Subroutine Package) 7:2
IBM System 360 23:377
Ibn al-Athir, works, *Complete History* (*al-Kamil fi al-ta'rikh*) 11:64–65
Ibn Khaldun 16:381
  works, *Examples* (*Kitab al-'Ibar*) 11:65
Ibn Sa'd, works, *Large Book of Generations* (*Kitab al-tabaqat al-kabir*) 11:64
Ibn-Hanbal 12:735
IBP *see* International Biological Program (IBP)
Ibrahim, Gamal 16:801–802
Ibrahim, Mahmood, works, *Merchant Capital and Islam* 16:382
IBS *see* Irritable bowel syndrome (IBS)
IBs *see* Individual budgets (IBs)
IC *see* Intercultural communication (IC); Intercultural competence (IC); Isolated condition (IC)
ICA *see* Intercountry adoption (ICA); International Coffee Agreement (ICA); International Communication Association (ICA); International Council of Archivists (ICA) (Paris)
ICAI *see* Intelligent computer assisted instruction (ICAI)
iCampus 6:897
ICAP *see* Integrated cognitive antisocial potential (ICAP) theory
Icarians 2:149
ICATUS *see* International classification of activities for time-use statistics (ICATUS)
ICC *see* International Chamber of Commerce (ICC); International Criminal Court (ICC); Item characteristic curve (ICC)
ICCPR *see* International Conference on Cultural Policy Research (ICCPR); International Covenant on Civil and Political Rights (ICCPR)
ICCs *see* Intraclass correlations (ICCs)
ICD *see* Index of cumulative deficits (ICD); International Classification of Diseases (ICD)
ICD-8 *see* International Classification of Diseases (ICD-8)
ICD-9 *see* International Classification of Diseases (ICD-9)
ICD-10 *see* International Classification of Diseases (ICD-10)
ICD-11 *see* International Classification of Diseases (ICD-11)
ICE *see* Immigration and Customs Enforcement (ICE); Internal combustion engine (ICE); Individual causal effect (ICE)
Ice jam flooding 6:480
ICE (item, context, ensemble information) model
  of context dependent memory 4:781
Iceland, anthropology 1:754
Iceman and His environment 17:460
ICEQ *see* Individualized Classroom Environment Questionnaire (ICEQ)
ICESCR *see* International Covenant on Economic, Social and Cultural Rights (ICESCR)
ICF *see* International Classification of Functioning; Disability and Health (ICF)
ICH *see* Intangible cultural heritage (ICH); International Conference on Harmonization (ICH); International Council on Harmonization (ICH)
Ichikawa Genzu 24:585
ICHoLS *see* International Conference on the History of the Language Sciences (ICHoLS)
ICIDH *see* International Classification of Impairments, Disabilities, and Handicaps (ICIDH)
ICIs *see* Inhibitory cortical interneurons (ICIs)
ICJ *see* International Court of Justice (ICJ)
ICLEI *see* International Council for Local Environmental Initiatives (ICLEI)
ICLEPS *see* International Center for Learning, Education, and Performance Systems (ICLEPS)
ICM *see* Idealized cognitive model (ICM); Integrated Causal Model (ICM); Intensive case management (ICM)

ICMJE *see* International Committee of Medical Journal Editors (ICMJE)
ICOMOS *see* International Council on Monuments and Sites (ICOMOS)
Iconic gesture 10:111
Iconic memory, and mental imagery 15:221
Iconic signs 25:171
Iconic turn 2:25
  and visual sociology 25:191
Iconicity 12:783, 14:210–211, 25:652
  Alexander on 1:527
  in functionalist grammar 9:473–474
  in linguistics 11:512–515, 14:534
  symbolism 20:722
Iconism 10:334
Iconoclasm 14:690–691, 20:859
Iconographic representations in art and rituals 4:612–613
Iconosphere, and visual sociology 25:191, 25:193–194, 25:196
Icons (semiotics) 4:235, 21:593
  and symbolism 23:860
ICOSH *see* International Committee for the History of Physical Education and Sport (ICOSH)
ICP *see* Independent Commission on Policing (ICP); Indochinese Communist Party (ICP); International Comparisons Project (ICP)
ICPD *see* International Conference on Population and Development (ICPD)
ICPHS *see* International Council for Philosophy and Humanistic Studies (ICPHS)
ICPSR *see* Inter-University Consortium for Political and Social Research (ICPSR)
ICR *see* Internal consistency reliability (ICR)
ICRC *see* International Committee of the Red Cross (ICRC)
ICRMW *see* International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICRMW)
ICRs *see* Imprinting control regions (ICRs)
ICRW *see* International Center for Research on Women (ICRW)
ICS *see* Interacting cognitive subsystems (ICS)
ICSERA *see* International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA)
ICSI *see* Intracytoplasmic sperm injection (ICSI)
ICSID *see* International Center for the Settlement of Investment Disputes (ICSID)
ICSS *see* Intracranial self-stimulation (ICSS)
ICSU *see* International Council for Science (ICSU)
ICSW *see* International Council on Social Welfare (ICSW)
ICT *see* Insulin coma therapy (ICT); Intentional change theory (ICT)
Ictal auditory hallucinations 10:500–501
Ictal period, seizures 7:854
ICTR *see* International Criminal Tribunal for Rwanda (ICTR)
ICTs *see* Information and communication technologies (ICTs)
ICTY *see* International Criminal Tribunal for the former Yugoslavia (ICTY)
ICVS *see* International Crime Victims Survey (ICVS)
ICWRN *see* Indigenous Child Welfare Research Network (ICWRN)
ID *see* Index of dissimilarity (ID); Instructional design (ID); Intellectual disability (ID); Intelligent design (ID)
Id 15:894, 19:337–338
IDEA *see* Individuals with Disabilities Education Act (IDEA); International Institute for Democracy and Electoral Assistance (IDEA)
Idea generation, and dreaming 24:722
Idea of progress 14:254
Idea of recognition 11:177–178
IDEA project 5:793–794
Ideal, knowledge representation 13:98–99
Ideal balance-of-power systems 1:556–557
Ideal contractarian theory of distributive shares 19:168
Ideal discourse situation 22:864
Ideal family size 8:805
  changes in ideals 8:806
Ideal observer 21:952, 21:954–955
Ideal of equality 7:907–908
Ideal points 16:37–38
Ideal randomized controlled trials 8:560
  double-blind 8:561
  key features of 8:560–561
Ideal self 4:725, 21:470
Ideal self-attraction 12:627–628
Ideal speech situation 5:305–307
Ideal types (psychiatric disorder type) 10:774–775, 11:518–519, 12:645, 19:317
  of rule and power 18:765–766
  Weber's 10:786
Ideal typology 18:31
Ideal utilitarianism 24:984
Idealism 10:85, 14:643, 14:646, 18:106–107, 18:488
  false charge of 23:592–593
  and *Geist* 9:667, 9:669–670
  Hegelian, and world spirit 11:818
  linguistic 4:794
Idealism–materialism dilemma 22:984–985
Idealization 11:516–517, 19:616
  Cartwright on 11:517, 11:520
  in economics 11:517–518
  Galileo on 11:517–518
  and liberalism 14:30–31
  Marx on 11:517–520
  Mill on 11:516–518
  ongoing disputes 11:518–519
  scientific explanation 11:517, 11:519–520
  Weber on 11:518–519
Idealized cognitive model (ICM) 19:287
Idealized norm, normality as 17:1
Ideal-object partitions 23:392
Ideal-point models
  in forced-choice personality assessment 17:843, 17:845–847
Ideals, neorepublican 20:480
Ideal-type Protestant ethic 12:645
Ideal-type sociology 15:48–49
Ideas
  association of 11:411
  cognition, and culture 5:577
  communism
    of Communist ideas 4:302–303
    of Marxism–Leninism 4:302–303
  culture as 9:2–3
  functionalism 9:532
  in history 22:108, 22:110, 22:112
  history of 2:545–546
    concepts in 4:550
  institutionalism 12:184
  intellectuals 12:278–280
  in social and behavioral sciences 15:321
  unmasking of 6:243
  utilitarian, dissemination of 24:983
Ideas of gurus (business writers) 17:851
Ideation 11:442
Ideational apraxia 1:877–879, 21:936
Ideational change, welfare state reform dynamics through 25:524–525
Ideational change approach, to fertility determination 9:61
Ideational culture mentalities 4:165–166
Ideational embeddedness 5:536–537
Ideational-themed explanans 4:24
IDEIA *see* Individuals with Disabilities Education Improvement Act (IDEIA)
Identical cognates 21:350
Identical monozygotic (MZ) twins 9:1010
Identical twins 24:703
Identifiability
  choices 19:442–443
  in flexible systems 17:557
  statistical 23:405–409
  in time-varying treatments 14:342–343
Identification of a structural equation model 23:616
Identification of genetic recombination, Muller's contribution to 16:14
Identification of outliers 23:388–389
Identification problems, in vocational education and training 25:246–247
Identifications 2:84–85, 2:316, 4:459, 11:521–531, 17:337
  childhood 11:538–539
  and conflict 4:625
  depersonalization 11:527
  in econometrics 7:9
  ethnic identity 8:113–116
  ethnicity 8:158–159
    being identified 8:159–160
  globalization 10:233–235

and groups, connection between 22:314–315
and language 13:253–254
social 11:527–528
and social change 11:529–530
social psychology of 11:526–527
statistical 23:405
background 23:405–406
Bayesian identification 23:408
definitions 23:406
identifiability 23:405–409
latent variable model 23:406–407
measurement model 23:406
trade and exchange 24:482–483
Identified regulation 21:488, 24:68
Identities 5:475, 5:928, 5:930, 9:690–691, 11:521–531, 11:543–544, 12:381, 14:284–285, 15:705, 16:195, 16:285–288, 16:777–778, 17:360–361, 20:338–343, 22:539, 22:878, 22:881, 23:82–85
and adoption 1:146
among immigrant minorities 11:676
in anthropology 11:532–536
agency 11:532–535
contemporary identity 11:535
contemporary influences 11:534–535
definitions 11:532–534
Erikson on 11:532, 11:535
identity politics 11:532, 11:534–535
person 11:532–535
self 11:532–535
in archaeology 1:901, 1:903
and biocitizenship 2:611
boundaries 11:528
and boundary work 23:852–853
British cultural studies 2:853–854
butch–femme subcultures 13:870–872
intersectionality 13:871
research on butch identity 13:871
research on femme identity 13:871
carnival 3:145, 3:148–149
in Central Asia 3:318–319
in childhood and adolescence 11:537–542
clan 3:732
coherence 11:521–523
collective action 11:528–530
collective political action 4:145, 4:147–148
and communication 4:285–286
and communitarianism 4:308–309
and communities of practice 4:328–329
concept of 7:938
construction 11:528–530
cultural 22:313
cultural, and authenticity 2:251
definition 16:563, 22:313–314
digital environments and 11:547
disability 6:472–476
and discourse 6:498–502
idiolect 6:499
inherited sources 6:498–499
interaction 6:498–501
on Internet 6:500–501
language 6:498–501
networks 6:499–501
optional sources 6:499
and self 6:499–500
and drinking 6:658–662
ego 22:313
emotions 7:503–505
and emotions 7:537
essential identity 2:252–253
ethnic 8:137, 8:140
see also Ethnic identity
and ethnic conflict 8:91–92
in Europe 8:276
external reasons 11:525
and fashion 8:838
formation in Latin America, gay/lesbian 13:860–861
futures online 11:552
gay, Latinos 9:628–629
and gay/lesbian tourism 9:639–640
and gender 9:783–784
generation, as facet of social difference 9:855–856
and generations 9:870
and genomics 2:623–624
ghetto 10:125–126
goals 15:889
Goffman's framework 11:550–551
historical sociology 11:27
in historiography 11:100–101
and homophobia/heterosexism 10:791–792
identity work 25:660–661
of individual 17:932–933
intentionality 11:522–523
internal reasons 11:524–525
Isaiah Berlin 11:524–525
Islam 11:523–524
issues, in genomics 10:33–34
ethnic identity 10:34
family identity 10:34
forensic identification 10:33–34
personal identity 10:34
journals 12:868
and language 13:286–287
and law 13:538
leadership 11:529–530
and linguistic anthropology 14:170–171
Marx on 11:523–525
in Middle East and North Africa 15:357–358
multiple 4:176–178
and music 16:150–151
narrative 16:204–206, 16:208–209
narrative networks 16:211, 16:213–216
national 16:311–312
national identity 11:526–528
nationalism 11:522–523
nonheterosexual British Asians 16:863
principles 16:862
threats and coping 16:863–864
normalization and stabilization of 11:552
and nostalgia 17:34–38
objectification of 11:548
objective 11:523–525
offline and online 11:547–553
online 11:551–552
performativity 11:548, 11:550–551
personal 4:152
philosophy 18:34–35
political 18:431–432
positive liberty 11:524–525
and postmodernism 18:676–678
problems in postsocialist societies 18:698–699
psychoanalysis 11:524
queer 19:766
queer anarchism 19:749–750
racialized 22:821–825
rationality 11:521–523
recognition 11:527–528
-related processes 11:527
relational 11:550–551
religion 11:522
and religion 3:575, 3:578
religion and new media 20:274
religious 20:338–341
religious, minority 8:145
representations of 11:528–529
representing 11:529
salience 11:526–528
and same-sex sexualities 20:871
and same-sex youth 9:649–654
second-order value 11:521, 11:523
self 11:522–525
self-identities 25:660–661
sexual 21:786–787
and queer anthropology 19:754–755
of sexual minorities in China 21:737–738
shared 4:154
social 4:152–153, 22:319–324
see also Social identity
social categories 11:526–530
and social change 11:529–530
social construction 11:524
social constructionism 22:225–226
social identities 25:660–661
social identity 11:526–530
and social movements 22:447, 22:449–452
and social networks 16:563
social ontology 22:500–501, 22:503–504
social work *see* Social work identity
solidarity 11:526, 11:530
subjective 11:521–523
taste information and usage 11:552
Taylor on 24:39–41
and territory 24:221, 24:224–225
theories of 11:547–549
and unemployment 10:517
use of term 22:313
virtual 11:549–550
volunteerism 25:268–272
well-being 11:527
women's studies 9:737–738, 9:740
and work 25:688
Identity achievement 22:218
Identity and agency 7:118
Identity control theory 7:537
Identity creation, in critical social work 5:290
Identity crisis, and group identity 4:176
Identity development 7:943, 11:537

Identity development (*continued*)
during adolescence 21:432–438
achievement 21:433–437
commitment 21:434–436, 21:438
cultural context 21:436–438
cultural identity 21:432–433, 21:438
diffusion 21:434–436
ego identity 21:432–435
exploration 21:434–437
foreclosure 21:434–437
identity statuses 21:434–437, 21:436t, 21:437f
independent development pathway 21:436–438
interdependent development pathway 21:436–438
moratorium 21:434–436
personality identity 21:434–436
adolescent, and leisure 13:831
in adulthood 11:538–539
in childhood 11:538–539
consolidation 4:178
construction processes 4:178
definition 21:432–433
Erikson's concept 11:537–540
ethnic identity 8:114
extension 4:178
formation process 11:539–541
developmental patterns of change 11:541
epigenetic consequences 11:541
familial antecedents 11:540
personality characteristics 11:540
future research 11:541–542
in infancy 11:538–539
versus Role Confusion 11:539–540
transformation 4:178
Identity diffusion 2:756–759
and motives 15:953
Identity disclosure 4:578
Identity does not matter view 17:743
Identity enhancer, homosexuality 16:866
Identity formation 7:883, 7:885, 16:515–516
in adolescence 9:434–435
consumption 4:758–759
Identity formation process 11:539
developmental patterns of change 11:541
epigenetic consequences 11:541
familial antecedents 11:540
personality characteristics 11:540
Identity issues following acquired brain injury 2:817
Identity lockbox 20:301
Identity maintenance needs 12:951
Identity movements 11:543–546
aboriginal 11:544–545
cognition 11:543
collective action 11:543, 11:545
emotion 11:543–544
feminism 11:545
Internet and 11:545
Islamist movements 11:545
multiculturalism 11:546
national 11:544
normative 11:545
particularistic 11:545
in Quebec 11:544
recognition 11:543–545
religious 11:544
in Scotland 11:544
social movements 11:543–545
universalism *versus* particularism 11:545–546
values 11:543–545
Identity negotiations, ethnic identity 8:116
Identity of sense 1:675
Identity over time 17:736
Identity performance theory 9:733–734
Identity personal *see* Personal identity
Identity politics 1:769–770, 5:290, 8:907, 9:769, 11:530, 20:46–47, 21:772, 22:319
and culture 5:584
in Mexico 1:628–629
and race identity 19:833
Identity process, and risk 20:715
Identity process theory (IPT) 16:862, 16:864, 20:714–715
Identity qua autonomous commitment 17:745–746
Identity realization, leadership as 13:652
Identity representation, leadership as 13:652
Identity status model 15:953
Identity statuses 7:938, 11:540–541, 21:434–437, 21:436t, 21:437f
Erikson's model 11:541
research on 7:943
Identity theft 5:633–637
costs of 5:634–635
definition 5:633
example 5:636–637
legislation 5:635–636
prevalence of 5:634
types of 5:633–634
victims and offenders 5:635
Identity theft legislation 5:635–636
Identity Theft Penalty Enhancement Act 5:636
Identity Theft Resource Center (ITRC) 5:634
Identity theories 20:300–301, 22:539
for ethical consumption 8:27–28
mental causation and 15:175–176
Identity threats
conflict, education, and youth 4:634
of xenophobia 25:804
Identity work, in narratives 16:221–222
Identity-based social movements 7:538
Identity-imparting 11:523
Identity-safe school environments 8:146
Ideographic script 14:242
Ideological, economic, military, and political *see* IEMP (ideological, economic, military, and political) model
Ideological antagonism, Cold War 4:128–129
Ideological approach, to film genres 9:161
Ideological attitudes 2:259–260
Ideological commitment, and communes 4:269
Ideological competitors 11:574–575
Ideological constraints 11:554–560
conceptual origins of 11:554–555
demonstration of Converse 11:557–558
formalization of 11:557
new conception of political sophistication 11:559–560
political knowledge 11:554, 11:559
reactions to Converse's formulation of political sophistication 11:555–557
Ideological dissensions 11:575–576
Ideological genocide 10:6
Ideological interest groups 12:372
Ideological level of analysis 13:899
Ideological pluralism 25:346
Ideological resources 21:135
Ideological secularism 12:872
Ideological self-descriptions 11:575–576
Ideological systems 22:750
Ideologies 6:245–246, 10:744, 11:94, 11:561–572, 13:111–115, 14:205, 14:652, 16:10, 23:324, 23:326–327, 24:296, 24:298
in administration 1:106
belief 2:494, 2:496
capitalist *see* Capitalist ideology
change 11:569–570
class 14:652
collective political action 4:145, 4:148
concepts of 11:562
cultural, childhood friendship 9:446
in economics 11:565
empirical relevance 11:570–571
qualitative analyses 11:571
quantitative analyses 11:571
end-of-ideology debate 2:499–502
genocide 10:15–16, 10:18–21
and hegemony 14:652
versus institutions 11:565–566
and intergroup relations 12:411–414
judicial behavior 12:886–887
Laissez-faire 16:462, 16:464
of language 6:505–506
legal 13:728
liberal 13:729
and linguistic anthropology 14:170–171
Mannheim's 14:497
Marxist critiques 10:787
in material social relations 14:652
of media sport texts 14:723
meritocracy as 15:240–243
motherhood as 15:882–883
nationalism as 16:293–294
neoliberalism as 16:462–463
origin 11:566–570
evolutionary approach 11:567
rational choice approach 11:567–569
in political science 11:564–565
populist 18:609
elements of 18:609–610
power of 17:655
and public opinion 19:562–564
racism 19:853–854
of rulership
chiefdoms 3:380

Smith's 22:119
social Darwinism 5:686, 5:689–690
social functions of 13:112
social science 22:693
in social sciences 11:564–565
and social work 5:288
in sociology 11:565
sociology of 4:164, 11:573–578
class 11:574, 11:576–577
collective beliefs 11:573
disassembling 11:575–576
domination 11:573–574, 11:576–577
explanatory schemes 11:576–577
and legitimate social order 11:574–575
liberalism 11:575–576
neoliberalism 11:576
origins 11:573–574
political values 11:574
radicalism 11:573
and scientific knowledge 11:576
as social beliefs 11:574–576
sociology of knowledge 11:574
terrorism 11:573, 11:577
stability 11:569–570
twentieth-century communism 14:656–662
unemployment 24:734–735
and worldview 25:755–756
Ideology, economy, military and political (IEMP) relationships 25:411
Ideology mandate, and manifestos 18:417, 18:419
Ideomotor apraxia 21:936
IDETF *see* Interdepartmental Emergency Taskforce (IDETF)
Idiocentrics 4:206
Idioculture 10:436
groups 10:436
Idiographic approach
in critical realism 5:278, 5:280
in geographic information systems 10:62–63
single-case designs 21:984–985
Idiographic science
Luria on 14:396, 14:398
regional geography as 20:122–124
Idiographic-nomothetic controversy 11:818
Idiolect 6:499
Idiomatic expression, lexicography 13:938
Idioms of distress 5:471
Idiosyncratic consciousness 11:644–646
Idiosyncratic gestural communication 11:256
Idola fori 25:288
Idolatry, and fetishism 9:105–106
IDP *see* Internally displaced people (IDP)
IDP Foundation 1:270
Idrisi (software) 23:138
IDS *see* Infant-directed speech (IDS)
IDUs *see* Injecting drug users (IDUs)
IDVS *see* International Dating Violence Study (IDVS)
IEA *see* Inclusive Education in Action (IEA); International Association for the Evaluation of Educational Achievement (IEA); International Energy Agency (IEA); International Ergonomics Association (IEA)
IEEE Transactions in Signal Processing 24:345
IEEPA *see* International Emergency Economic Powers Act (IEEPA)
IEMP *see* Ideology, economy, military and political (IEMP) relationships
IEMP (ideological, economic, military, and political) model 18:477
IEP *see* Individualized education program (IEP)
IES *see* Institute of Education Sciences (IES)
IEs *see* Intersession experiences (IEs)
IESALC *see* International Institute for Higher Education in Latin America and the Caribbean (IESALC)
IESBS *see* International Encyclopedia of Social and Behavioral Sciences (IESBS)
I-experience 11:532–533
If clause 1:187
IFG *see* Inferior frontal gyrus (IFG)
IFIs *see* International financial institutions (IFIs)
IFJ *see* International Federation of Journalists (IFJ)
IFRC *see* International Federation of Red Cross (IFRC)
IFRS *see* International Financial Reporting Standards (IFRS)
IFRs *see* International food regimes (IFRs)
IFSAR *see* Shuttle Interferometeric Synthetic Aperture Radar (IFSAR) mission
IFSW *see* International Federation of Social Work (IFSW)
Igbo 1:302
IGBP *see* International Geosphere-Biosphere Programme (IGBP)
IGLTA *see* International Gay and Lesbian Travel Association (IGLTA)
Ignatius of Antioch 3:548
Ignatow, Gabriel, on embodiment 7:418
Ignorability, causal inference 3:257
Ignorance 11:579–583, 12:744
asymmetric information 11:579
censorship 11:579
discretion 11:579
and kindred concepts 11:579
in knowledge theories 11:580–581
oblivion 11:579
practical consciousness 11:581
probability 11:579
risk 11:579–582, 20:708
secrecy 11:579–580
serendipity 11:580
in social action theories 11:579–580
sociology, agnotology and 11:582–583
strategic ignorance as organizational source 11:581–582
taboo 11:579
tacit knowledge 11:580
and uncertainty 24:718
vagueness 11:579
veil of ignorance 11:579
Ignorantia juris haud excusat 13:586–587
IGOs *see* Intergovernmental organizations (IGOs)
IGU *see* International Geographical Union (IGU)
IHA *see* Impaired hypoglycemia awareness (IHA)
IHDI *see* Inequality-adjusted Human Development Index (IHDI)
IHL *see* International humanitarian law (IHL)
IHSN *see* International Household Surveys Network (IHSN)
IHSR *see* Intergroup helping as status relations model (IHSR)
IIEP *see* International Institute for Educational Planning (IIEP)
IIF *see* Item information function (IIF)
IIPS *see* Intraparietal sulcus (lIPS)
IIRIRA *see* Illegal Immigration Reform and Immigrant Responsibility Act (IIRIRA)
IIS *see* Institut International de Sociologie (IIS)
IISUE *see* Institute for the Study of University and Education (IISUE)
IJCP see International Journal of Cultural Policy (*IJCP*)
Ijmāʿ 12:746t
Ijtihād 12:746t
IJzerbedevaart (Pilgrimage of the Yser) 18:414
IKC *see* Innovative knowledge community (IKC)
IKEAization 14:829–830
IKT *see* I-Kuan Tao (IKT)
I-Kuan Tao (IKT) 16:806–807
IL *see* Interlanguage (IL)
IL-6 *see* Interleukin-6 (IL-6)
ILEC *see* International Legal Ethics Conference (ILEC)
Ileri, Selim 20:870–871
Ileto, Reynaldo, works, *Pasyon and Revolution: Popular Movements in the Philippines* 11:86
ILI *see* Independent Living Institute (ILI)
The Iliad 16:155
Ill health, experience by minority ethnic elders 15:440–442
Illegal activities, fieldwork 9:154–155
Illegal corruption equilibrium 5:14–15
Illegal downloading 12:588–589
Illegal drugs 6:676, 6:686
Illegal immigrants 11:664–665, 20:114, 25:805
Illegal Immigration Reform and Immigrant Responsibility Act (IIRIRA) 11:637
Illegal migration 25:125
Illegal substances, possession of 15:593
Illegitimate children 8:780
Illegitimate hierarchies 5:251
Illegitimate power 21:138
Illegitimate private violence 25:116–117
Illich, Ivan 6:27, 10:611, 11:584–588
biography of 11:584

Illich, Ivan (*continued*)
impacts of 11:587–588
works 11:584–587
ABC: The Alphabetization of the Popular Mind 11:586–587
Celebration of Awareness 11:584
Deschooling Society 11:584
Energy and Equity 11:584
Gender 11:586
H_2O and the Water of Forgetfulness 11:586–587
In the Vineyard of the Text 11:587
Medical Nemesis 11:584
Tools for Conviviality 11:584
Illicit drug use 6:685–686
age of onset of 6:680
definition 6:679
epidemiology of 6:679–680
family socioeconomic status and 6:679
individual-level characteristics and 6:679
multicomponent interventions 6:682
peer group and 6:679
prevention and treatment of 6:680
psychosocial interventions 6:681–682
public policy 6:682
social environment, role of 6:679
social learning theory interventions 6:682–683
social marketing 6:681
theoretical foundation for prevention 6:680–681
Illinois tradition of small group research 22:103–104
Illiteracy 11:589–592, 14:236–240
LAC educational system 7:178–179
neural bases of reading 16:596–597
and social exclusion/deprivation 22:287
sociology of 11:589–592
education inflation 11:591
functional illiteracy 11:590–591
international survey of literacy skills 11:590–591
mass literacy campaigns 11:590–592
monolingual society 11:589
multilingual society 11:589
orality 11:589
social contexts and 11:591–592
statistics 11:590
and urban poverty 24:915
Illness
bioarchaeology 2:601
causes of 10:568
children's concept of 10:680–681
coping with
and personality 17:794–795
defined 10:657
definition of 10:567
versus disease 5:473–474
and dyadic coping 11:593–595
ecological perspectives on 6:878–879
etiologic and diagnostic correlates of 10:676–677
Kraepelin on 13:127, 13:129–130
living with 21:844
local variations in 10:572–573
from medicalization to pharmaceuticalization 18:8–9
mental representations in different cultures 10:567–570
implications 10:568–569
prevention and treatment of 10:675–676
psychoneuroimmunology 19:446–447
responses of emotions to 7:496–497
and self-efficacy 21:505–506
Strauss on 23:528
trajectories 23:529
types of 10:568
Illness anxiety disorder, explanatory style 8:599
Illness behavior 11:509, 11:596–602
chronic illness 11:599–601
management 11:600
self and social context 11:600–601
definition 11:596
determinants 11:597–599
importance of 11:596–597
schemata 11:597–598, 11:600–601
somatoform disorders 11:596, 11:599–601
Illness beliefs
and cardiac rehabilitation 20:188–189
chronic 3:552
and symptom perception 23:868–869
Illness experience, disability 6:475
Illness models of human nature 17:911
Illness–death model 8:312, 8:312f
Illocutionary speech act 14:172
Illocutions 18:797
Ill-Treatment and Torture (ITT) measures 18:446–447
Illumination preprocessing, face recognition 8:659–661
Illuminismo 7:656
Illusio, and fetishism 9:107
Illusion 11:458–459, 11:548
self-referenced 11:461
and sensory-affective phenomena 11:460
Illusory contour 25:186–187
Illusory correlation 12:410
Illusory flashing Marroquin circles 25:161
Illusory inference 19:62–63
ILM *see* Iterated learning model (ILM)
Ilness-related stress 23:572
ILO *see* International Labour Organization (ILO)
ILP *see* Institutional logics perspective (ILP)
ILPES *see Instituto Latinoamericano de Planificación Económica y Social* (ILPES)
ILS *see* Indirect least squares (ILS) estimators
Ilyas, Maulana Muhammad 20:375
IMaCh *see* Interpolated Markov Chains (IMaCh)
IMAGE *see* International Migration Around the GlobE (IMAGE) project
Image brokers 2:873
Image ethics, and mass communication 14:690–691
Image history 2:24–25
Image inspection 15:225
Image intensity 23:411
Image interpretation 20:420
Image manipulation 15:225
Image processing 21:292
Image transformation 15:225
Image-based constraints 25:156
Imagery 2:882
from aerial photography 10:61
autobiographical memory 2:285–286
mental 15:220
from satellites 10:61
and self 19:401–402
Imagery duration 15:225
Imagery–language dialectic 10:115–116
Images
concept, in photography 18:82–84
conflicts 4:583–585
coregistration 9:527
dramaturgy 6:646–647
of health 10:571–572
honor killing 11:186–187, 11:190
human body 2:729–731
of motherhood 15:881
scientific 25:221–225
Image-scoring 11:262
Imaginal exposure 8:606–607, 8:609
Imaginal neglect, visual mental imagery 25:165–166
Imaginary community 16:224
Imaginary interaction 21:813
Imagination 11:458–460, 20:248
emotional 21:815
and false memories 8:710–713
Geertz on 9:665
genealogical 9:810–811
Humboldt on 25:287–288
and media 15:7
online 21:815
and pretend play 18:870, 18:874
romantic 21:815
social representations 20:444, 20:446–447
sociological 15:537–538
Vygotsky on 25:332
Imagination inflation 8:710
Imaginative play 15:7
Imagined communities 8:273, 9:673, 11:526–527, 16:284–286, 17:37–38, 18:659–660, 20:363
in Europe 8:273
nation as 16:294
Imagined nations 20:401
Imagined (extra-sensory) perceptions 16:170
Imaging
in clinical psychology 11:603–611
in hippocampal amnesia 10:882–883
Imaging studies
in humans 24:29
odor 24:29
olfactory-taste convergence to represent flavor and influence of satiety 24:29
taste 24:29
of post-traumatic stress disorder 18:704
Imagining positive intergroup interactions 11:678–679
Imagistic mode of religiosity 20:393

IMD *see* Index of Multiple Deprivation (IMD)
IMF *see* International Monetary Fund (IMF)
Imipramine 1:778–780
Imitation 1:874–876, 4:164–165, 16:740–741, 1:876f
  Akers' SLT 22:410–411
  animal social learning 22:402–404, 22:406
  game of 14:557
  of goal-directed actions 10:269, 10:271
  during infancy and early childhood 11:621–627
    cognitive development 11:621–622
    culture 11:622, 11:626
    functional considerations 11:623–624
    imitating inferred goals 11:622–623
    imitating inferred strategies 11:623
    imitation of subgoals and means 11:622
    immediate and deferred imitation of acts on objects 11:621–622
    implications for parenting and education 11:624
    implications for theory 11:621–622
    intentions 11:622–624
    interim summary 11:623
    neonatal imitation 11:621
    practical implications 11:625
    preschoolers 11:621, 11:623–624
    regulating imitation 11:623–625
    social cognition 11:625
    social considerations 11:624–625
    social learning 11:621–623
    theories of human learning 11:622
    theory of mind 11:622–623, 11:625
  learning by 8:389
  like me theory 15:131–133, 15:132f
  and mirror neurons 15:586
  neonatal 11:621, 16:481–488
  and prenatal/infant development 18:853–854, 18:856–858
  principle of 4:758–759
Immaculate Conception of Mary 3:234–235
Immanence 21:391
Immanentization 18:436
Immaterial economy, Gorz 10:288
Immaterial labor, science communication as 21:223–224
Immaterialist theories, of personal identity 17:742–743
Immatriculation, in sub-Saharan Africa 23:656
Immature behavior 2:792–793
Immediacy 12:783, 21:623
  hunters-gatherers 11:429–430
  social influence 22:350
Immediacy hypothesis 21:623
Immediate memory 25:175
Immediate-return system 11:894–895
Immersion (bilingual education) 2:580–582, 2:584
Immersive environments 6:900
  augmented reality 6:900
  tangible user interfaces (TUI) 24:22
  virtual environments 6:900
Immersive virtual
  tangible user interfaces (TUI) 24:21
Immigrant adolescents
  identity for 9:435
  social interactions 9:437
Immigrant assimilation 15:459
Immigrant children, educational progress for 8:733
Immigrant congregations 20:295–296
Immigrant crime problem 5:153
Immigrant enclaves 15:414
Immigrant incorporation 11:655, 11:659
Immigrant integration 11:642–643, 11:665–666
Immigrant networks 12:515
Immigrant paradox 1:72–73
  and academic motivation 19:794
Immigrant policies 12:516
Immigrant religious organizations 20:295–296
Immigrant status, and socioeconomic status (SES) 1:44
Immigrant students, difficulties of 22:326
Immigrant victimization 5:151, 5:153, 5:155–156
Immigrant–host relations 8:142
Immigrants 15:462–464
  and adolescent friendship 9:432, 9:435, 9:437–438
  Americanization 1:218
  assimilation of 2:81–87
  community studies 4:371
  cultural and religious identities 5:518
  and cultural policy 5:461
  and cultural rights/defense 5:493–494
  ethnic and national identities 8:115
  and ethnic intermarriage 8:118–121
  in EU countries 15:380–381
  history 15:446–448
  illegal 25:805
  integration
  see also Integration
  'other' category, construction of 15:423–424
  versus receiving country citizen dichotomy 11:676
  and settlement houses 1:218
  and social work 15:416–421
  systemic barriers 11:642, 11:644
  undocumented 25:805
  and xenophobia 25:802–806
Immigration 2:86–87, 5:151–157, 10:169–170, 10:172–173, 11:638, 11:641–647, 11:654–661, 13:263, 15:462, 15:469–470, 20:295–296, 20:361, 21:396–397
  and adoption 1:138
  applied research on 11:635–640
  challenges 11:656
  and citizenship 11:650–652
  and crime 5:152–155, 5:197
    contemporary empirical findings 5:153–154
    generational differences 5:154–155
    historical background 5:152–153
    possible explanations for overrepresentations 5:155–156
  from Cuba 15:464
  from Dominican Republic 15:464
  economic consequences of 15:383–384
  and ethnic conflict 8:91–92
  ethnic identity
    and work 8:106–112
  in Europe 8:142–147
  European 15:462
  exclusion through disenfranchisement 11:642
  and family 15:428–429
  and fertility 15:397
    2007 recession in United States 15:403–405, 15:404f
    age at migration 15:397–399, 15:403, 15:405, 15:398f, 15:402t
    children ever born (CEB) 15:398–399, 15:401–403, 15:401t
    ethnicity 15:397, 15:399–401, 15:405, 15:400t–401t
    link between 15:397–399
    nativity 15:400–401, 15:400t–401t
    population mobility 15:397
    time at destination 15:397–399
    in United States 15:400–405, 15:400t–402t
  future research 11:659–660
  historical development 11:655
  immigrant victimization 5:151, 5:153, 5:155–156
  and inequality 15:429–431
  and integration 12:238
  Japanese 15:463
  labor market impact of 15:459–460
  and language 13:260–262
  methodological issues 11:655–656
  from Mexico 15:463
  and mortality crossover 15:849
  and nation 16:225
  during nineteenth century 15:462
  North American 17:24
  and parenting 17:496–497
  policy 15:460–461
  political aspects 11:669–674
    in Europe 11:669–673
    integration and citizenship 11:671–673
    liberalism 11:673
    multiculturalism 11:672
    in North America 11:669
    policies 11:669–671
    political science 11:669–670
    political sociology 11:669
    sovereignty 11:669–670
  and population composition in North America 18:553–556
  and population momentum 18:576–577
  and poverty 18:743–744
  and racism 19:858, 19:861–862
  and residential segregation 20:541
  and schooling 15:429
  second-generation 5:152–155
  social psychological aspects 11:675–681
    acculturation 11:676, 11:679–680

Immigration (*continued*)
beyond immigrant *versus* receiving country citizen dichotomy 11:676
discrimination 11:675–679
ethnic identity 11:676
identities among immigrant minorities 11:676
improving intergroup relations with intergroup contact 11:678–679
in-groups and the out-groups 11:675–676
intergroup conflicts 11:676–678
intergroup threat 11:676–679
multiculturalism 11:678–680
national majority 11:675–676
policies and norms in culturally diverse societies 11:679–680
prejudice 11:676–680
stereotypes 11:677–678, 11:680
in Southeast Asia 23:72
state and social work 11:641–642
studies 5:151–152
in United States 11:628–634
urban 24:929–930, 24:932
and xenophobia 25:804–805
Immigration and Customs Enforcement (ICE) 21:663
Immigration and Naturalization Service (INS) 1:631
Immigration controls 8:248–249
Immigration laws 5:268
Immigration policy 11:662–671
asylum 11:662, 11:664, 11:666
of Canada 11:662–667
citizenship 11:662, 11:664
family reunification 11:662, 11:664–665
of France 11:662–666
free movement 11:662, 11:665–667
of Germany 11:662–666
historical background 11:662–665
immigrant integration 11:665–666
implications for welfare states and labor markets 11:665–666
of Italy 11:662–665
of Japan 11:664–665
of Korea 11:664
Migration and Integration Policy Index (MIPEX) 11:666
multiculturalism 11:665–666
recent developments 11:664–665
securitization of 11:664–665
of Spain 11:665
types 11:662
of United Kingdom 11:665
of United States 11:662–665
work-based migration 11:662, 11:665
Immigration Reform and Control Act (IRCA) 1:622–623, 15:464
Immigration restriction 8:231
Immigration Restriction Act 8:226
Immigration surplus 15:438
Immigration–crime nexus 5:151–153
Immobilization of time 24:407
Immorality Amendment Act 2 1988 (South Africa) 21:744
Immorality Amendment Act 57 1969 (South Africa) 21:744
Immorality Amendment Bill (South Africa) 21:744–745
Immune Attack (game) 4:501
Immune function
religion and 20:267
and stress 23:552
Immune functioning, and emotions 7:499
Immune system 22:486
regulation by hypothalamus 11:489–490
Immunity, psychoneuroimmunology 19:445–449
Immunological détente 18:900
Immunological factors, obsessive-compulsive disorder 17:122
Immunological findings, of OCD 17:115
Immunological mechanisms, of hysteria 11:508
Immunology 12:32–35
psychoneuroimmunology 19:445–450
Immutable mobiles 5:524
I(1) model, interpretation of 24:325–326
hypotheses on adjustment coefficients 24:326
hypotheses on long-run coefficients 24:326
normalization of parameters 24:326
relation between the models H(r) 24:325–326
Impact, Population, Affluence, and Technology *see* IPAT (Impact, Population, Affluence, and Technology)
Impact evaluations 19:118
randomized controlled trials 8:541–545
Impact factors 20:530
of journals 21:322–324
Impact validity 17:542
Impact-oriented private-sphere environmentalism 7:701–702
Impaired confrontation naming 13:364
Impaired hypoglycemia awareness (IHA) 6:338–339
Impaired single-word comprehension 13:364
Impairing visual word recognition 4:90
Impairments
associated with spatial perception 17:704–705
disability 6:467–470
and disability 6:452–456, 6:472–476
reduction, in psychosocial interventions 22:382
Impartiality 4:804–805, 4:807, 23:592
quality of government as 19:693
Strong Program 23:593
Imperative, contingency 23:612
Imperative pointing 10:271–272
Imperfect competition 7:115
Imperfect competition theory 20:728
Imperial beginnings, indigeneity 11:755–756
Imperial postal systems 18:640
Imperial rule, restoration, in Japan 15:127
Imperial Telegraph Administration 24:154–155
Imperialism 3:95–96, 4:223, 4:225, 6:811–814, 11:395, 11:682–691, 14:658, 19:854–855, 21:368–369, 23:310, 24:894
Chomsky on 3:529–530
classic empire 3:773
decolonization 11:684
defense economics 6:5–6
and development geography 6:314–315
empire 11:682–684
end of 11:684
of free trade 11:687
geopolitics of 16:457–460
hierarchy 11:682–684
history of
American 11:689
Belgian 11:688–689
capitalist theory 11:687
collaborationist theory 11:687
concept 11:686–688
definition 11:685–686
Dutch 11:689
Hobson's contribution 11:686
Italian 11:688
Marxist theory 11:686
motives and means 11:689–690
national articulations 11:688–690
Portugal 11:688
Russian 11:689
institutional theory 11:682–684
and legal reform 15:360–361
media 15:55–59
multinational corporations 11:683–684
multinational states 11:684
and national liberation movements 16:243
and nationalism in East Asia 16:275, 16:277
neocolonialism 11:682–683
and postcolonial geography 18:644–646
protectorate 11:682–683
realism 11:682–683
relational contracting 11:683
in Rome 1:700
in social work ethics 22:786–787
in South Asia 16:281
spheres of influence 11:682
theories 11:683–684
use of term 11:682
and warfare 25:416–418
in West Africa 1:303
and whiteness/white supremacy 25:572–573
world systems theory 25:748–750
Imperialist economy in modern Japan 6:794–795
Impersonal causality 2:202
Impersonality
of bureaucracy 2:914–916
quantification 19:702–703
Implementation 10:313, 10:621, 10:623, 10:626
autonomous agents 2:310
and biological entities 2:312
bureaucracy 11:692–693

defined 11:695
design failures 11:693
explaining success and failure 11:692–693
failures 11:693
generic governmental instruments 11:693–694
regulation 11:693
service delivery 11:693–694
networks 11:694
of organizational research 17:372
policy 11:694
practitioner's view of 11:694
of randomized controlled trials 8:542–543
successful 11:692
urban policies 24:903
Implementation errors 11:262
Implementation intentions 19:271, 19:277
health psychology 10:675
and motivation 15:938
and motivation/actions 15:891–892
Implementation phase, of ADDIE model 12:198
Implementation research 5:129
Implementation science 11:695–702
active implementation frameworks 11:696
diffusion theory 11:695–696
effective implementation 11:697
drivers 11:698–699
implementation teams 11:700
improvement cycles 11:699–700
skilled implementation 11:698
stages 11:697–698
formula for success 11:696–697
rationale for 11:695
Implementation theory 19:579
Implicature 21:589–590
Implicit association test (IAT) 2:180–181, 7:415, 11:703–708, 11:715, 15:955, 17:817, 17:873, 19:859, 19:865, 21:523, 22:163, 23:447
for academic motivation 9:680
confounding influences on IAT effects 11:707
example 11:704, 11:703t
for prejudice 18:829
procedure 11:703–704
process models 11:705–707
figure-ground asymmetries 11:706–707
Quad model 11:707
random-walk processes 11:705–706
stimulus–response compatibilities 11:705
task-switching costs 11:706
reliability 11:704
for sexism 21:697
validity 11:704–705
group level 11:704
individual level 11:704–705
Implicit attitude 2:172–173
Implicit attitude measures 2:180–181, 2:188
context effects 2:181
low tech measures 2:181
response competition measures 2:180–181
response facilitation measures 2:180
response time measures 2:180–181
Implicit bias 19:859–860
and race 19:810
Implicit bias research, for discrimination 6:534
Implicit conception 13:81
Implicit control 4:832–833
Implicit feedback, in mathematics learning 4:69–70, 4:70t
Implicit inferences 19:394–395
Implicit learning
and creativity 17:771
and tacit knowledge of rules 13:84–85
AGL in patients with amnesia 13:85
skepticism about 13:85
Implicit linguistic competence 2:574, 2:576
Implicit meaning, case of 18:796–797
Implicit measures 11:703, 11:714, 11:716–717, 17:873
implicit association test (IAT) 17:873
Implicit memory 11:709–713, 15:147–148, 19:2, 23:876–877
assessment of 6:761–763
bias theory of repetition priming 11:711
cognitive neuroscience theory 11:711–712
initial resolution 11:710–711
multiple memory systems theory 11:710
neuroimaging effects reconceptualization 11:712
nondeclarative memory 16:844, 16:847
retroactive interference in 20:614
single-system processing theory 11:710
see also Nondeclarative memories
Implicit motives 2:488, 2:490–491, 15:951–952
congruence of 15:952–954
determinants of congruence of 15:954
development of 15:955
Implicit or nondeclarative memory 1:644–645
hippocampus implications of 1:647
Implicit person theory 25:684
Implicit rationing, of health care 10:603
Implicit religion 21:381
Implicit representation 13:78–79
and inference 13:78–79
Implicit rules 13:78–79
in connectionist networks 13:81–82
Implicit self-concepts 21:470–471
Implicit social cognition 11:714
applications 11:719
associative-propositional evaluation (APE) model 11:716–718
attitudes 11:714, 11:716–718
behavior prediction 11:717
cognitive control 11:716–717
consciousness 11:714, 11:716
dual-process theories 11:717–718
formation, change, and context effects 11:717–718
implicit 11:716
lack of process purity 11:718
measurement procedures 11:714–716
affect misattribution procedure 11:715–716
extrinsic affective Simon task 11:715
go/no-go association task 11:715
implicit association test 11:715
sequential priming tasks 11:714–715
mental association as core concept 11:714
mental representation 11:716–718
MODE (Motivation and Opportunity and DEterminants) model 11:716
prediction 11:717
prejudice 11:714, 11:716–717
quad model 11:718
relations between implicit and explicit measures 11:716–717
self-concept 11:714, 11:717
self-esteem 11:714, 11:717
stereotyping 11:714–715
Implicit theories 21:848
of intelligence 12:303, 21:50
Implicit Theory of Mind in infancy 24:268–269
Implicit visual information processing in hemispatial neglect 10:768–769
Implicit visual long-term memory 25:178
Implicit visual short-term memory 25:177
Import substitution 6:277
Import substitution industrialization (ISI) strategy 6:309
Importance sampling (IS) 15:758–762
construction of samplers 15:759–760
efficient importance sampling 15:760–761
posterior moments 15:759–760
smooth functional integration 15:761
Importation, in United States 11:629
Importation model 11:629
Imports, trade policy 24:485–487
Import-substitution industrialization 1:939
Import-substitution policies 25:749–750
Impossibility theorem 25:509
Impression formation model, of social categorization 22:187–190
Impression management 17:708, 17:834–835, 17:849, 18:771–772, 21:871, 22:539
dramaturgy 6:648
and social media 22:422–423
Impressionism, and styles 16:565
Imprinting 14:368
Imprinting control regions (ICRs) 24:687
Imprisonment 5:216, 7:82, 11:721–724
deterrent effect 6:261–262
goals of 11:721
and political repression 18:446–447
and poverty 18:754, 18:757
rates 11:721–723
sociological impacts 11:723
types 11:721
variation in imprisonment rates 11:721–723
of women 25:617–619
increase in 25:617–618
Imprisonment rates 5:184–185
IMPROVE 4:884
Improving Access to Psychological Therapies (IAPT) program 14:501, 19:590

Improving educational environments 7:821–822
IMPS *see* Inpatient Multidimensional Psychiatric Scale (IMPS)
ImpSS *see* Impulsive sensation seeking (ImpSS)
Impulse, and instinct 12:175–180
Impulse control 2:757
Impulse control disorder not elsewhere classified 17:592, 17:596
Impulse response, Fourier transform 9:377
Impulsive sensation seeking (ImpSS) 21:608
Impulsiveness 17:775–776
  and procrastination 19:73, 19:75–77
Impulsivity 17:593–594, 20:437
  and aggression 1:377–378
  gender differences in 9:752
  and motivation/actions 15:892
Impure altruism 1:579
Impure science 8:580
Imputation 4:868–869, 9:579, 13:389
  multiple 16:88–93
Imputation methods 16:923
  nonresponse error 16:922–923
IMR *see* Infant mortality rate (IMR)
IMRAD (introduction, methods, results, and discussion) format 20:529
IMS *see* Information-motivation-strategy (IMS) model
IMT *see* International Military Tribunal (IMT)
IMTFE *see* International Military Tribunal for the Far East (IMTFE)
In Defence of Empirical Psychology 2:857
In situ learning 22:5
In situ wh-questions 16:6
In utero gene–environment interactions 7:844
In vitro fertilization (IVF) 9:927, 9:995–996, 12:47
  risk screening 20:701
In vitro reinforcement paradigm 20:214–215
In vivo exposure 8:606–608
Inadequacy, school burnout as 21:59
INAH *see* Instituto Nacional de Antropología e Historia (INAH); Interstitial nuclei of the anterior hypothalamus (INAH)
Inalcik, Halil, works, *Economic and Social History of the Ottoman Empire: 1300–1914* 16:382
Inalienability 8:479
Inalienability rule 19:170
Inanimate energy systems 7:617
Inanimate objects 21:269
Inasmuch as education 7:198
Inattention blindness 3:348
Inauthentic 2:251–252
In-between spaces, liminality 14:131–132, 14:136
Inbreeding 11:336, 11:338, 11:725–729
  avoidance 18:900
  biologic outcomes of 11:727
  genetic effects of 11:726
  prevalence of deaths and defects in 11:728
  Wright's contribution 25:758–759
Inca Empire 23:44
Inca society, and culture of economy 11:865
Inca state 23:39
Incapacitation 6:256, 13:572, 13:574–575
Incarceration 5:186, 5:197–198, 11:721, 11:723, 13:575
  in Anglo-American countries 18:939
  in Continental European countries 18:939
  and crime 5:197–198
  criminogenic effects of 18:940
  in Europe 18:939
  explanations of 18:939–940
  in Finland 18:939
  rates 18:939
  reentry and reintegration after 20:61–66
    challenges 20:61–64
    emerging policy lessons 20:64–65
    employment 20:62–65
    interventions 20:61, 20:64–65
    needs 20:61–62, 20:65
    policy 20:61, 20:64–65
    program 20:65
    rehabilitation 20:61, 20:64
  in Scandinavian countries 18:939
  unanticipated consequences of 18:940
  in United States 18:939
Incentive intensity 1:345
Incentive pay 17:612
Incentive regulation 20:168–169
  earnings sharing regulation 20:168
  franchise bidding 20:168–169
  options 20:169
  price cap regulation 20:168
  regulatory lag 20:168
  yardstick regulation 20:168
Incentive salience 20:213–214
  drug wanting/needing 25:352–354
Incentive sensitization theory 25:351–352
Incentive theory 12:270
Incentive-based engagement strategies
  for patient adherence 17:600–601
Incentive-based regulation 20:176
Incentives 12:103–104, 14:267, 17:780–781
  and administration 1:105–106
  appeasement 1:828, 1:831
  educational planning 7:265
  experimental economics 8:548–549
  global governance 10:156
  job offer acceptance 12:826–827
  and motivation 15:937
  principal-agent models 12:104
  promotion-based 19:137
  to recruit and retain subjects 14:356
  in schools 7:154–155
  surveys and polls 23:756
  sustainable development 23:823, 23:827
  welfare reforms 25:516–517
Incentivization, international public goods 19:540, 19:542–543
Incest 11:725–729, 21:659–660
  and family 8:811
Incest aversion 15:815
Incest avoidance/taboos 11:726, 11:730–731
  constitutionalist 11:730–731
  conventionalist 11:730–731
  endogamy 11:730
  exogamy 11:730
  Lévi-Strauss on 13:905
Incestuous marriages 11:726–727
  biologic outcomes of 11:727
  direct investigation of progeny 11:727–728
  in pharaonic Egypt 11:726–727
  in Zoroastrian Persia 11:726–727
Incestuous union, types of 11:726–727
Incidence
  definition 15:179
  epidemiology 7:825–826, 7:826f
Incidence matrices, structural balance theory 23:605–607
Incidence of stress-inducing incidents 23:570–571
Incidence–prevalence rates, of voluntary agencies 25:262
Incident coronary heart disease (CHD) 2:420
Incidental encoding of context 4:782–783
Incidental learning, and episodic memory 7:857
Incidental parameters 14:123
Incivility
  and political trust 24:661
  in workplace 25:723
Inclusion
  borders and 2:762–763
  gender and technology 9:726–727
  and intersectionality 12:660
  language policy for 13:358
  social *see* Social inclusion
  and social dominance orientation 22:250–252
Inclusion memory paradigm, task network models 16:543–544
Inclusionary mechanisms of restoration 12:958
Inclusion/exclusion paradigm 16:543–544
  processing trees for 16:544, 16:544f
Inclusive business (IB) 22:168, 22:171–172
Inclusive cartography 24:9, 24:13
Inclusive club goods 16:416
Inclusive disjunction 19:62
Inclusive Education in Action (IEA) 23:213
Inclusive fitness 9:29–30, 13:26, 13:28–29, 22:900–901, 22:910–911, 13:28f
  human behavioral ecology 11:246–248
  kin selection and 13:44–46
  and life span development 14:108, 14:110–111
Inclusive fitness theory 11:338
Inclusive legal positivism 13:485–486, 18:26–27
Inclusive nonpositivism 13:479–480
Inclusive positivism 13:479
Inclusive practice, learning disability 13:687
Inclusiveness of moral universe 15:785–786
Inclusivity 22:874
Income 7:82–83, 7:132–136

and cognitive ability 7:132, 7:135
concept of 9:218
and cultural lifestyles 22:332–333
determinants of fertility 9:20–21
differences by education 7:129
distribution, effects of minimum wages on 15:563–564
and divorce 14:521–522
effects on education 7:132–134
net effects 7:132–134
overall relationship 7:132
household panel studies of 5:845–847
and inequality 11:900
differences 11:906
distribution 11:900
and internal migration
average income differentials 12:513
distribution in source and destination countries 12:513–514
and labor supply 13:191–195, 13:197
of minority ethnic elders 15:440–441
SES and social-emotional development of children 22:946–947
and social welfare 11:904
variation across countries 7:47–48
variation within country 7:48
Income capitalization techniques 25:453
Income deprivation 22:285
Income distribution 25:513
measurement issues 11:732–738
and migration from Europe 15:457, 15:460
Income distributional equity and law 13:452
Income dynamics, and poverty 18:744–745
Income inequality 6:537, 6:539, 7:83, 11:739–743, 22:327
and crime 12:463–464, 12:466–467
data 11:739–740
disposable income in affluent nations 11:740t
female-headed families 11:742
globalization 11:741
great U-turn 11:740–741
household factors 11:742
and justice 12:929
labor market forces 11:741–742
measurement 11:739–741
middle class 11:740
offshoring 11:741
pension policy 17:693, 17:695
and political trust 24:661
single mothers 11:742
skill biased technological change 11:741
and social change 22:194
social class 11:740
sources of income 11:739
and student achievement 5:345
top 1% 11:742
trends 11:740–741
and welfare reforms 25:519
Income redistribution 23:339
Income support, welfare state 25:476
Income support policy, labor supply, family, and fertility effects of 18:745
Income taxes 11:744–749, 23:339–342, 23:341t
corporate 11:746
depreciation 11:746
and economic behavior
labor supply 11:744–745
saving and portfolio choice 11:745–746
economic efficiency, and the equality–efficiency trade-off 11:747–748
evasion 11:746–747
nature of 11:744
progressivity 11:744, 11:747–748
proposals for tax reform 11:748
and redistribution of income 11:747
risk-taking 11:747
Income-producing property 20:13–14
Incoming reflection centrality scores 16:537
Incommensurability 14:868, 20:229
Feyerabend on 9:120–121
Kuhn on 13:133–137
of paradigms 21:308
value pluralism 25:22, 25:25
Incomparability, value pluralism 25:24–25
Incompatibilism 9:405
Incompatibilists 9:405
Incompatible texts 11:551
Incomplete contingency table, capture–recapture methods 18:604–607
Incomplete data 4:470–471
Incomplete information
and game theory 9:583, 9:589
inverse projection 6:149
Incomplete observation, survival analysis 23:774
Incompleteness theorem 16:924–925
Incongruence, and values 25:28
Incongruity, and humor 11:416–417
Inconsistency 19:64–65
Incorporated places 24:854–856
Incorporated structures 10:487
Incorporation
argument 14:221–222
defined 14:220
grammatical 14:222–223
noun 14:220–221
of outliers 23:388–389
Incorvio, Antonio 25:839
works, *Generation 1000* 25:839
INCOTERMS® 13:915–916
Increasing/decreasing quantifier 19:708
Incremental change, in industrial ecology 11:849
Incremental fit indices 8:688
Incremental innovation 12:147–148, 22:356, 22:356t
Incremental internationalization 12:584
Incremental theories about intelligence 19:791–792
Incremental validity 6:345
of SJT 22:16
Incrementalism 23:518, 19:354
budgeting as 2:901–902
challenges to 2:902–903
patterns of 12:583
Incubation, in creativity training 5:115–116
Inculpation 9:956–957
Inculturation 23:881
and syncretism 23:881
Incumbency advantage system, of United States Congress 4:650
Indecent exposure 15:593
Indefinitely repeated Prisoner's Dilemma 18:935–936
Indemnity 12:210–214
Indentured labor 15:447–448
Independence 1:9, 19:35
of axioms 2:340
extending 19:35
judicial 12:879, 12:881, 12:891–896
language policy for 13:356–357
in metaphysical objectivity 17:96–97
population dynamics 18:580
of primitive concepts 2:340
testing, nonparametrics 16:881
violations, t-test 8:10–11
Independence chains 14:569
Independence family model, and parenting 17:494–497
Independence problem, in cross-national research 19:721–723
Independence struggles
in MENA 16:363
Southern Africa 23:77–78
Independence–interdependence orientations 6:748–749
Independence-predicting graphs 10:347–348
Independence-preserving graphs 10:347–348
Independent Broadcasting Authority (IBA) 24:169–170
Independent Commission on Policing (ICP) 18:268–269
Independent contractors 7:573
Independent cultural model, and early memory development 6:760
Independent cultures, and play 18:218, 18:220–224
Independent development pathway, of identity 21:436–438
Independent directors, corporate governance 4:936
Independent identically distributed (iid) errors 24:327
Independent information variables, in auctions 2:207
Independent living 6:462–463
European network of 6:463–464
seven needs of 6:463
Independent Living Institute (ILI) 6:464
Independent living model of disability 6:447
Independent living movement, beginnings of the 6:462
Independent market, popular music 3:802
Independent nuclear family system 8:816–817
role of relative cohort size 8:816–817
Independent Payment Advisory Board (IPAD) 10:606

Independent practice association (IPA) 10:595
Independent private values, in auctions 2:206–207
Independent random variables, representation by 5:902
Independent reference 19:709
Independent selves 5:426
  motivational tendencies 5:426
Independent social work departments, development of 19:302–303
Independent specialists 3:379
Independent variables 3:189–192, 4:441–442, 15:64–66, 19:725
  loglinear models 16:125–127
Independent verb form 1:188
Independent walk M–H chain 15:764
Independent-observations decision strategy (IOS) 21:949–950
In-depth interviews 12:671–672, 17:532
INDEPTH Network 1:157, 6:139
In-der-Welt-sein 19:315–316
Indeterminacy 16:186, 22:691
  reasoning 19:591–592
  risk 20:708
Indeterminacy principle of Werner Heisenberg 20:287–288
Indeterminacy problem, in politeness 18:329
Indeterminate sentences 18:880
Indeterminism 3:266, 19:23
INDEX (rule-based impact assessment model) 24:891
INDEX (SDSS tool) 23:140
Index (semiotics) 21:593, 21:595–596
Index fossils 3:563, 14:260–261
Index matching 19:143–144
Index Number Institute (INI) (US) 9:219
Index numbers 5:753
  economic approach to 5:753–754
Index of consumer expectations 9:317
Index of cumulative deficits (ICD) 15:846–847
Index of dissimilarity (ID) 20:544, 21:411
Index of Global Philanthropy and Remittances 3:368
Index of Multiple Deprivation (IMD) 22:944
Index of Prohibited Books 18:916
Index of segregation (IS) 20:544–545
Index of sustainable economic welfare (ISEW) 6:853, 6:858
Index Visible (filing system) 9:220
Index wars 20:549
Index-based measurement of extreme response style 23:759
Indexes 18:917, 25:171
  associated with TFR 6:113
  inequality 11:900
Indexicality 6:53, 13:258
  symbolism 20:722
Indexing vocabularies 19:59–60
Index-shifting systems 15:720
India 1:955–956
  ancient *see* Ancient India
  archaeology of 23:50–57
  campaign funding 18:395–396
  caste system in 3:223
  communalist ideologies in 4:264
  diasporas 6:355
  education in 7:195–196
  fertility transition in 9:80–83
  gender gap 15:612
  Khadi cottage industry in 5:110
  modern, after decolonization 11:80
  nationalism in
    dominant strand in 16:279–280
    pan-Indian scene in nineteenth century 16:279
  partition of 12:433
  peace building and music in 25:844–845
  personal law in 4:264
  pharmaceutical industry 18:7
  psychiatric care in 6:45–46
  religions of 20:373–376
    Hindu law 20:375
    Hinduism 20:373–374
    internal warfare 20:375
    modern period 20:374
    nationalism 20:375–376
    non-Hindu religious communities 20:375
    public sphere 20:375
    revolution 20:375
    social movements 20:374–375
  sex ratio at birth (SRB) 15:609–610
  sex selection in 15:610–612
  Shils on 21:884
  sociocultural aspects of *see* South Asia; sociocultural aspects
  withholding/withdrawing supports (WH/WD) 7:599b
Indian Act 1876 (Canada) 1:635, 11:763–764, 11:796
Indian Ayurvedic texts 5:628
Indian BPO sector 3:44–46
  organizational structure 3:46–48
    Indian independents 3:47
    Indian IT industry subsidiaries 3:47–48
    Indian specialist 3:47
    MNC outsourcing service providers 3:47
    MNC specialist 3:47
    MNC subsidiaries 3:46–47
    US high-technology start-up 3:47
Indian Child Welfare Act 1:137
Indian Claims Act 1:746–747
Indian Claims Commission Act 1946 8:192
Indian diaspora 6:355
Indian Genome Variation Consortium 11:318–319
Indian Islam 20:375
Indian legal traditions
  ancient 23:58
  codification of procedural laws 23:61
  company period (1601–1833) 23:59–60
  postcolonial legal orders 23:61–62
  rights under colonial government 23:61
Indian management ethos 11:786
Indian Penal Code 4:248–249, 23:61
Indian Removal Act 1830 8:84
Indian Reorganization Act 1:635
Indian Sociological Association, Dalits and Backward Classes 22:676
Indian subcontinent 7:194
Indian syllabaries 25:777–778
Indian Traditional Knowledge Digital Library 21:188
Indianization, in Southeast Asia 23:70
Indians, in Britain *see* British Asians
Indic civilization 2:76
Indic studies, of Whorf 25:577–578
Indicative mood 24:212
Indicators 7:563, 11:750–754
  causal 11:751
  of depression 11:752
  development and selection 11:753
  effect 11:751
  measurement models 11:751–752
  measurement strategies 11:750–751
  for motivation 11:750–751
  multiple 11:751
  noncontinuous 11:752
  properties 11:752–753
  reliability of 11:752–753
  for strategic choices 21:325
  of sustainable development 23:802–803, 23:803t
  terms in social sciences 11:750
  use of term 11:750
  validity of 11:752–753
Indicators of Education Statistics (INES) program 5:763
Indifference, ethnomethodological 21:192–194
Indigeneity 11:755–761, 11:784
  conquest 11:760
  cultural difference 11:756, 11:758–759
  imperial beginnings 11:755–756
  indigenism and reindigenization 11:760
  inequality 11:759
  land 11:757–760
  mobilization and political process 11:757–759
  in national and global perspective 11:756–757
  political mobilization 11:757–759
  recognition 11:755, 11:757–759
  scholarly studies 11:760
  settler states 11:756
  in South America 23:48
  United Nations 11:755–758, 11:760
Indigenism 11:760
Indigenismo 1:626–627, 1:630, 15:328–329
Indigenista legacy, in Mexico 1:626–628
Indigenization 11:789–790, 16:458–459
  movement 16:458
  path 11:789
  of social work 22:792–793
Indigenous, meaning 11:795
Indigenous Australians *see* Australian Aborigines
Indigenous Bantu cultures 12:738
Indigenous capitalism 11:786
Indigenous child welfare practice (Canadian case) 11:762–768

affects of colonial practices 11:763–764
analysis of policy story 11:765
commonalities for 11:762–763
current and future strategies 11:767
delegation 11:766–767
first nations child raising 11:762
interim measures 11:765
interruptions for 11:763
provincial involvement 11:764–765
Spallumcheen experience 11:765–766
Indigenous Child Welfare Research Network (ICWRN) 11:767
Indigenous Chinese 12:697, 12:699–700
Indigenous corporations 11:786
Indigenous difference, definition 11:755
Indigenous emotions 5:545
Indigenous healing practices, sexual risk behaviors 21:799
Indigenous identity, in Mexico 1:628–629
Indigenous inhabitants, of Arctic 1:927
Indigenous knowledge 11:779–784, 21:177, 21:188
in Africa 1:293
archives 11:774
biodiversity and biopiracy 11:774–775
colonialism 11:770–772
copyright and Aboriginal art 11:773–774
definitions 11:779–780
definitions and terminology 11:771–772
environmental movement 11:781–782
and epistemology 11:780–781
future options and strategies 11:775–776
globalization 11:780–783
in historical and ecological context 11:781–783
Indigenous collections 11:774
and intellectual property rights 11:769–778
international legal instruments 11:776
law 11:770–772
libraries 11:774
museums 11:774
and politics 11:781
postcolonialism 11:779–782
research agreements and protocols 11:776
rights of indigenous peoples 11:781–782
sites of struggle 11:773–777
Sui generis law 11:776–777
traditional knowledge licenses and labels 11:777
tragedy of the commons 11:782
World Intellectual Property Organization 11:775
Indigenous land rights 17:577–578, 17:582–583
Indigenous management 11:784
capitalism 11:786
examples 11:784–786
notions of indigenous 11:784
Indigenous management styles 11:784–787
Indigenous methodologies, in interviewing 12:674
Indigenous minorities, of Australia 2:234
Indigenous movements 11:781–783
Indigenous nations 11:762
Indigenous peoples 9:962–963, 11:771
Aboriginals *see* Australian Aborigines
archaeology politics 1:885–890
benefit 9:965
bioethics in 9:966–967
biospecimen 9:963–966
community engagement and collective consent 9:965
cultural beliefs 9:964–965
and culturally appropriate interventions 5:516–518, 5:520
culture 9:964
decision-making structure 11:762
diversity of 9:963
and energy 7:622–623
ethical issues and 9:963–964
group harms 9:964
historiography and historical thought in 11:57–58
hunters-gatherers 11:428–430
and *Indian Act* 11:763
inherent knowledge 11:762
Maori, social work *see* Maori social work
methods 11:762
of North America 17:25
ownership 9:965–966
repatriation 5:489–490
'repatriation' of blood samples 9:966
of South America 23:36–37
and spirituality 23:285, 23:289
of sub-Saharan Africa 11:89–91
and whiteness/white supremacy 25:571, 25:573
Indigenous perspectives 11:784–786
Indigenous populations
advocacy 1:224
Australian *see* Australian Aborigines
Indigenous psychologies 5:311, 5:482, 11:788–794
achievements of 11:791
cross-indigenization 11:790
debates on 11:793
definition 11:788–789
development in various parts of world 11:790–791
in India 11:790
indigenization 11:789–790
indigenous theorizing 11:792–793
issues 11:793
in New Zealand 11:791
in Philippines 11:789–790
in Taiwan 11:790
Indigenous research 11:784
methods 11:791
Indigenous rights 11:362, 11:364, 11:795–803
authenticity 2:253
historical development of 11:795–796
colonialism and doctrine of discovery 11:795–796
formation of modern international legal system 11:796
instruments and institutions of 11:801–802
jurisprudence
international treaty 11:798–799
regional treaty 11:799–800
United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) 11:800–801
substantive provisions of 11:800–801
Indigenous scientific knowledge 19:155
Indigenous social work 11:804–810, 16:458–459
anticolonialism 11:807
based within indigenous epistemes 11:807–809
colonial impacts 11:806
decolonization 11:807
epistemes and ways of being 11:804–805
centrality of spirit 11:804
ceremonies 11:805
community, egalitarianism, and respectful individualism 11:805
holistic perspective 11:804
importance of elders 11:805
mino-pimatisiwin approach 11:804, 11:807–808
recognition of cycles and flux 11:804–805
values 11:805
healing 11:804, 11:808–809
helping practices within indigenous cultures 11:805–806
indigenous diversity 11:806
sharing circles 11:808–809
sweat lodge 11:807–809
Indigenous studies
ethnohistory 8:192, 8:196
of personality 9:251–252
Indigenous theorizing 11:792–793
Indigenous theory development 11:784
Indigenous women, violence against 10:195–196
Indignados movements 1:685–686, 25:839–840
Indirect aggression 1:373
Indirect assessment of early memory 6:760–761
Indirect confounding 10:348
Indirect connectivity 16:536
Indirect contact, and adoption 1:144–145
Indirect discrimination 18:830
in employment 6:534
Indirect effect 15:64, 15:67
mediation model 22:549
Indirect estimation 6:143–148, 19:782
of adult morality 6:145
Brass on 6:143–146, 6:144t
of child morality 6:143–144, 6:144f, 6:144t
consistency checks 6:145–146
adult morality 6:146–147, 6:147f
fertility 6:145–146
of fertility 6:145
of migration 6:145
quality of demographic data 6:146–147
Indirect euthanasia 8:281
Indirect evolutionary approach and game theory 8:434
Indirect indexicality 13:258
Indirect influence, of groups 10:423

Indirect least squares (ILS) estimators 21:981
Indirect measures 11:703
Indirect object (grammar) 10:324
Indirect racial discrimination 6:542–544
Indirect reciprocity 11:262–263
Indirect requests, and pragmatic disorders 18:788, 18:791–792
Indirect retrieval of memory 20:608, 20:610–611
Indirect standardization, demographic measurement 6:121
Indirect victimization, fear of crime 8:856–857
Indirect/future costs, pharmacoeconomics 19:458
Indirectness in language 18:326–328
Individual accountability 4:883
  in competition/cooperation 4:859
  cooperative learning 4:882–883
Individual actions
  analytical sociology 1:671–672
  and social facts 22:192
Individual addressability, mobile communication 15:630
Individual agency 3:521
  role in life span development 4:814–815
Individual and group, similarities and differences 23:114
Individual aptitude 7:276
Individual assets 3:446
Individual behavior
  and mortality 15:861–862
  Smithian model of 15:332
Individual budgets (IBs) 4:346
  for older people 21:498–503
Individual causal effect (ICE) 8:528
Individual characteristics, role similarities, during preschool age 18:223
Individual cognition 15:305–307
  and levels of description 15:306–307
Individual consequences, and mass imprisonment 14:696–699
Individual continuous distributions 6:576–578
Individual decision-making games, experimental economics 8:549
Individual differences 6:548, 8:669–674, 12:909, 15:894, 15:897–898, 18:854, 20:355, 24:173–177
  in adolescent leisure activities 13:832–834
  in agreeableness 1:463–466
  approaches to prosocial behavior 1:581
  assessment of 9:233
  cardiac defense 23:543–544
  cardiovascular reactivity 3:119–120
  cognitive development 18:122–123
  contra-hedonic motivation 1:244–246
  in coping styles with chronic illness 3:551–552
  in domain-specific self-concepts 21:494
  in embarrassability 21:478
  in extraversion 8:623
  in facial emotion expression 8:667–675
  in facial information processing 8:676–678
  genetic factors 9:919
    affectivity 9:920–921
    in cognition 9:920–921
    in personality 9:920–921
  and groups 10:438
  in HPA axis functioning 11:478–481
    effects of adverse environmental factors 11:479
    effects of genetic factors 11:479–480
    gene–environment interactions 11:480–481
    interactions with other hormone- and neurotransmitter systems 11:478–479
  in infancy 18:857–858
  in inferencing 19:398–399
  instructional psychology 12:203
  in intelligence 12:317
    reasons for 12:318
  in language capabilities 24:536–537
  in learning 7:243–244
  in life span development 14:103, 14:105
  in mathematics learning 4:72–73, 4:73t
  measurement of facial expression production 8:669–670
  modeling 19:443
  multidimensional scaling 16:37
  neural mechanisms in face cognition 8:680–681
  in neuroanatomy and neurophysiology 24:537
  nomological net 8:673–674
  in personality traits 17:899, 17:903
  prejudice as 18:834
  prosocial behavior 1:581
  of psychological stress 23:585
  psychology of 17:314
  in psychopathology 3:765–766
  and reinforcement 20:209
  in social referencing 22:560
    attachment style 22:560
    atypical development 22:560
    temperament 22:560
  structure of voluntary facial emotion expression (VfEE) 8:670–673
  temperament 24:180–182
  in values 25:41
  Western approaches to 24:173–174
  of white-collar crimes 25:553–554
Individual Differences Scaling 19:80–81
Individual disparate treatment, for employment discrimination 6:531, 6:533
Individual effects, of advertising 1:210–211
Individual factors
  of body image 2:737–738
  middle adulthood development 15:372–374
    cognitive abilities 15:372
    emotional abilities 15:372–373
    health and physical changes 15:372
    motivations 15:373
    personality development 15:373–374
  and multiple language development 16:62
Individual freedom principle 25:676
Individual heterogeneity, diffusion 6:414
Individual human rights 4:707
Individual identity, and race identity 19:833
Individual interest 12:378
Individual internalization of social constraints ('force to force oneself') 2:729
Individual learner factors 21:364–366
Individual learning 8:438
Individual level 9:22
  psychosocial effects of Alzheimer's disease at 1:586–589
  of stigma analysis 23:464
Individual life, secular chronicle of 2:645
Individual lifestyle 14:112
Individual mobility 23:464–465
  in meritocracy/tokenism 15:240, 15:242, 15:244
Individual motivation 15:303
Individual narratives 16:215
Individual negotiating behavior 16:434
Individual opinions 19:563
  factors that shape 19:564–566
Individual outcomes, and organizational citizenship behavior 17:319
Individual performance 17:612–614
Individual perspective of coping 4:903–904
Individual Placement and Support (IPS) Model 19:298, 19:303
Individual practice, sociology of professions 19:107
Individual pro-environmental behaviors 19:83–84
Individual psychology 25:793
  and social capital 22:177–178
Individual rationality 15:669
Individual responsibility 20:581–582
Individual rights 3:687–688, 9:730, 14:39–40, 14:42, 16:322
  and conservatism 4:691
Individual rights and liberties 13:437–438
Individual social mobility 15:239
Individual solitary flow 9:264
Individual therapy 19:515
  for depression 19:375–376
Individual variation, weight regulation 25:470
Individual work
  with men 14:683
  nature of 24:106–107
Individual-based models
  agent-based models 1:367
  demographic microsimulation 15:344, 15:346
Individual–cluster trial 3:900
Individual-collective relationship
  history of concept 11:816–819
  internal and external relations 11:817
  philosophical aspects of 11:817
Individual–context relations
  adolescent self-regulation 21:547–552
  of human development 11:276–280
Individual/group rights 3:657
Individualism 2:133–134, 2:510, 3:541, 4:206–210, 4:319–321, 5:567, 15:302–304, 15:807, 17:825,

17:932, 19:173–176, 21:174, 21:212, 21:872, 22:223, 22:225–226, 23:644
adolescent friendship in 9:433–434
conditions for formation 9:433
applied social research 1:850
attributions 4:208
and authenticity 2:251–252
cognition, emotion, and motivation 4:208
collective mobility 4:207–208
versus Collectivism 5:531
common fate 4:207
communication 4:209
conflict resolution 4:209
conformity 4:209
describing self 4:208
distribution 4:206–207
and economic development 5:530–531, 5:533
economic status 4:207
emotions 4:209
exposure to other cultures 4:207
history 4:206
homogeneity 4:207
ideal of fitting-in rather than shining 4:208
and individuality 17:934–935
influence on homosexuality stigma 21:736–737
ingroups 4:207
interdependence 4:207
interpersonal relationships 5:426–427, 5:429–430
intrinsic motivation 12:696–697, 12:699–700
and joint action 12:843
large families 4:208
liberal, and Pareto's rational actor model 14:25
and liberalism 14:24
little exposure to modern mass media 4:208
methodological *see* Methodological individualism
Mill on 15:524, 15:526–527
mobilizing for confrontations 4:208
modernity 15:682–684, 15:686
and personality psychology 17:908–909
philosophical aspects 11:811–815
population density 4:207
recreation 4:209
and reduction of collectives to individuals 11:818–819
resource distribution 4:209
skills in entering new groups 4:208
social behavior 4:208
social loafing 4:208–209
and social mechanisms 22:418–419
and social ontology 22:502, 22:504
and social psychology 22:525
social stability 4:207
social support 4:208
socialization 4:208
strict religious upbringing 4:208
and terrorism 24:239
Tocqueville on 24:422
transgressions of ingroup 4:209
and unemployment 24:733, 24:736
utilitarian social thought 24:984–985
and values 25:28–31
well-being 4:208
in West Africa 1:302
Individualism–collectivism 5:326, 6:748, 17:824
and parenting 17:494
Individualism–holism–systemism trilemma 22:984–985
Individualist action, legal reasoning 13:776
Individualist liberalism, and self 21:446
Individualistic collectivism 21:437
Individualistic competition 5:678
Individualistic descriptions 15:304–305
and cognition 15:306–307
of complex social facts 15:305
point of view and normative thinking 15:304–305
Individualistic efforts, during competition 4:859
Individualist-integration 12:237, 12:240, 12:236t
Individuality 1:9, 11:126, 17:932, 17:934–935
cognitive styles 4:110
respect for, and liberalism 14:24–25
Individualization 5:538, 10:622, 18:434, 24:571, 24:846–847
Beck on 2:397–398
in England 4:345–346
of events 13:588
Individualized Classroom Environment Questionnaire (ICEQ) 3:825–826
Individualized education program (IEP) 23:222
EAHCA principle 23:222t
Individual-level asymmetry 2:806
Individual-level comparative databases 5:792–793
Individual-level data
assessment and adjustment of 6:127–128
and ecological fallacy
bounding 6:867
equivalence conditions for 6:866–867
Individual-level interventions
for occupational stress management 17:148–150
effectiveness 17:149–150
implementation 17:150
Individual-level legal mobilization 13:590–593
Individual-level mechanisms 14:877
Individual-level moral inclusiveness 15:785–786
Individual-level phenomenon
acculturation as 1:68–69
framework 1:69f
Individual-level predicates 23:716
Individual-level processes
political psychology 18:429–430
cognitive approaches to political decision-making 18:429–430
personality profiles of citizens and political leaders 18:429
political beliefs and values 18:429
Individual-level racism 19:860
Individual-level stability, and adolescent personality 17:864–865
Individual-level studies 5:326
of homicide 11:165–167
Individual-level treatment, for mass trauma 14:749–750
Individual-oriented interventions, for skin cancer 23:698–699
Individuals
active role in shaping one's life 14:85, 21:558
to be socialized, characteristics of 22:864
benefits of skills training for 22:23
concept of 11:816
configurations
social structure as 22:717
conformity 4:190
Durkheimian 17:932–933
economic behaviors of 5:888–889
entrepreneur 7:682
and environment, interaction between 17:136
and freedom 11:818
genetics and social justice 9:969
giving primacy to 15:302–303
identification with groups 4:190
identity of 17:932–933
impact of economy on psychology 5:889–890
influence over organizational cultures 17:699
in institutional economics 17:191–194
and liberalism 14:13–14
and libraries 14:48
life courses of 22:263
non-Durkheimian 17:933–934
politeness from point of view of 18:328–329
and reproduction 20:450–454
role in life span development 14:102
screening and selection 21:328–332
self 21:449–453
and social structure 22:719–722
social treatment of 7:908–909
and time use 24:370–371
and unemployment 10:516
values for various types of 25:18
and workplace autonomy 25:718–720
Individuals in authority positions, legitimacy 13:825
Individuals with Disabilities Education Act (IDEA) 23:215, 23:223
Individuals with Disabilities Education Improvement Act (IDEIA) 23:223
Individual-worth monotonicity 4:874
Individuation 22:188
and body 2:719
land tenure 13:233–234
and Second Demographic Transition 21:346
Individuation process, Jung on 12:904–905
Indivisibility
of human rights 11:353–354

Indivisibility (*continued*)
and territorial conflict 24:215
Indo-Aryan languages 10:318
Indochinese Communist Party (ICP) 25:95
Indo-European comparative linguistics 8:208
Indo-European culture 13:585
Indo-European (IE) languages 3:186, 11:820–827, 13:337–343
cognates 4:424t
phonological system 11:825f
subgroups 11:822–824
Indo-European possessive pronouns 18:631–632
Indo-European trifunctionalism 3:723–724
Indo-Islamic period
historiography of 11:79
colonial and postcolonial historiography 11:79
medieval Persian chronicles 11:79
regional histories 11:79
Mughal period 11:79
Indology 20:322
after World War II 11:78–79
Indonesia 23:63–65
fertility transition in 9:85–89
historiography and historical thought 11:85
postcolonial law in the period of nationalism 13:815
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
urban policies in 24:897
Indoor Environmental Quality 7:774
INDSCAL model (individual differences scaling model) 16:46
Indubitably 24:844
Induced abortion 9:77, 9:79
in East Asia 9:56, 9:58
Induced innovation, theory of 1:472
Induced paraphasias 21:936
Induced strategy 23:512–513
Inducements, appeasement 1:826, 1:828, 1:831
Induction 1:536, 19:61, 19:63–64
legal reasoning 13:776–779, 13:782–783
LTP/LTD 14:359–360
Induction programs, for teachers in United States 24:78
Inductive inference 8:617
Inductive methods, empirical 23:424
Inductive reasoning 12:646
Inductive skill of reasoners 19:64
Inductive strategies, Strong Program 23:595–596
Inductively coupled plasma emission spectrometry 24:481
Inductively coupled plasma-mass spectrometry 10:56, 24:481–482
Inductively coupled plasma-optical emission spectroscopy 10:56
Inductive-statistical explanation 8:595
Indulgence, and values 25:30
Indulgent parenting style 19:245, 19:248
Indulging, and motivation/actions 15:889–890
Indus civilization 23:52–54
Indus River system 25:444–445
Industrial Accident Insurance Act 1884 17:142
Industrial agriculture
sustainable agriculture 23:809
Industrial and organizational (I–O) psychology 1:854, 11:879–884
ambidextrous leadership 11:883
cross-cultural training in 5:334–341
domain of 11:879
evolution of 11:881–882
Hawthrone experiments 11:881–882
1960s and 1970s 11:882
from 1980s forward 11:882
factors driving demand for research and expertise in 11:883
human engineering 11:880–881
human relations movement 11:882
organizational psychology 11:880
personnel psychology 11:879–880
role of research in guiding science and practice 11:882–883
scientific management 11:881
scientist–practitioner model 11:882–883
validity generalization 11:882–883
work-life fit 11:883
Industrial archaeology 11:1
Industrial armed forces 4:604
Industrial automation 2:296
Industrial behavioral management, applied behavioral analysis for 2:410
Industrial capitalism 2:750, 3:96–97, 4:338, 15:701
and disability 6:453
Industrial citizenship 13:203
Industrial cluster dynamics 11:828–833
Industrial cluster dynamics, special economic zones and 23:207–212
Industrial Commission 1901 5:153
Industrial concentration 2:54
Industrial deviance 1:794
Industrial districts 24:856–857
Industrial dynamics 11:834–842
country-level patenting comparisons 11:835–837
and entrepreneurship 7:683–684
firm-level patenting comparisons 11:838–840
global LED industry driven by technological appropriability (case) 11:840
industrial evolution 11:834, 11:838–840
industry cycle 11:841–842, 11:842f
industry-level patenting comparisons 11:837–838
innovation 11:835–838, 11:841
intellectual property 11:834, 11:837–842
knowledge diffusion 11:834, 11:837–838, 11:841
latecomers 11:834–842
light-emitting diode (LED) 11:837–838, 11:840
patent statistics as indicators of innovation and 11:835
worldwide 11:835
worldwide international patenting activity 11:834
Industrial ecology 11:843–853, 22:254, 22:256–258
boundaries 11:849
definition 11:843–846
biological analogy 11:843–844
dematerialization and eco-efficiency 11:846
forward-looking analysis 11:846
putting the elements together 11:846–847, 11:847f
role of companies 11:845–846
systems perspective 11:844–845
technological change 11:845
goals 11:846–849
positive and normative analysis 11:848
reducing risk *versus* optimizing resource use 11:847–848
transformative and incremental change 11:849
typology of ecosystems 11:844f
Industrial era, traditional family of 8:779–780
Industrial geography 11:854–859
agglomeration economies 11:858
emergence of managerial capitalism and factories 11:856–857
existence and importance of creative economy 11:858
global industrialization 11:854, 11:856, 11:858
industrial location 11:854–855, 11:858
knowledge economy 11:858
landscape 11:854–855
mechanization of manufacturing 11:855, 11:858
organization of industry in factory era 11:855–856
production systems 11:855–857
rise of collective capitalism 11:857–858
technological change 11:854–855, 11:857
Industrial linkage
in entrepreneurship 7:688
Industrial location 11:854–855, 11:858
Industrial metabolism, material flow analysis 14:760–761
Industrial organization 23:508–509
Industrial organization economics, film and video analysis 9:158–159
Industrial organization economists 14:549–551
Industrial organization of health insurance and health services industries 10:616–617
Industrial organization theory 8:547–548
Industrial organizational consultation 4:725
Industrial parks 23:208
Industrial policy 23:209–212, 25:691
communications and media 5:706
culture and 5:537
Industrial printing 18:916–917
Industrial production, in China 24:974–975
Industrial psychology
human resource management 11:350

and job evaluation 12:816–817
Industrial relations 7:571–572, 13:202–203, 25:689
Central and Northern Europe 11:861
concept of 11:860
Commons 11:860
Dunlop 11:860
critiques 11:862
end of an era 11:861–862
Europe, models 11:860–861
Japan 11:861
Latin Countries 11:861
new practices 11:862
North America 11:861
shift in 11:861–862
UK and the Republic of Ireland 11:861
Industrial research 24:763
Industrial restructuring, and regional geography 20:117–119
Industrial Revolution 11:648, 11:868–869, 12:363, 17:169, 18:663, 24:596, 25:670–671
in China 20:428–429
Renaissance and 20:428–429
and energy 7:618–619
family and schooling 8:751
human resource management 11:349
and industrial societies 11:863–867
slavery for 22:56–57
Smith and 22:119–120
Industrial societies 13:105, 13:107–108, 25:657–660
anthropology of 11:863–867
capitalism 11:863–866
consumption 11:863–866
cultural critique 11:863
cultural perspectives of technology and economies 11:863–865
development, underdevelopment, and sustainability 11:866–867
ecology 11:863–866
Industrial Revolution 11:863–867
industrialization and sociocultural condition of modernity 11:865–866
political economy 11:863
world, historical prerequisites 11:863
capitalist society 25:658–659
cult of the individual in 14:24
and economy 13:109
energy flow analysis 7:629–631
inequality 11:894
and land use indicators 13:239–240
tradition in 24:508
Industrial sociology 25:657–661
divisions of labor 25:658–659
major dimensions
conflicts 25:661
global social–political–economic change 25:658–659
Industrial therapy 6:47–48
Industrial transformation, and social question 22:554
Industrial Tribunals 14:265
Industrialism 25:512
urban life 24:837
Industrialization 2:906, 4:751–752, 11:233, 11:865–866, 11:868–878, 13:201, 14:20, 15:493, 16:286, 16:288, 18:202, 20:843, 24:812, 24:814–815, 25:671
bedrock universals 11:871
in China 11:868–869, 11:871
conflicts between diverse institutional logics 11:869–871
conjunctural coincidences 11:869
and development geography 6:314, 6:316–317
and division of labor 13:156, 13:158–160
and East Asian economic development 6:788–789, 6:791
and economic growth 5:528–529
effects on society 13:12
equality of opportunity 7:904
of Europe 23:794
economic growth and 23:794
of farming 1:461, 5:159–160
and fire 9:198–199
gay and lesbian culture 9:632
geography 11:869
in Germany 11:869
in Great Britain 11:868
and greening of business 10:392, 10:394–395
Hirschman on 10:896
and history 11:104–105
and hybridity 11:450
industrial revolution 11:868–869
knowledge economy 11:870
and life course 14:76–77
Mao Zedong on 11:869, 11:871
Marx on 11:869
of media 19:605
and modernization 15:712
and nationalism 16:290–293
organizations 11:869–871, 17:309–310
path dependency 11:870–871
positive legacies of layered histories 11:871
process, interaction of family with 8:741–742
Smith on 11:870–871
state planning 11:868
Taylor on 11:870
technological innovation 11:868–871
technology as composite 11:868–869
transatlantic migration 15:448
typologies 11:872–878
British industrial revolution 11:873–874
European industrialization 11:874–876
modern industry 11:873–874
process of 11:872–873
regions, towns and proto-industrialization 11:874
white-collar workers 25:565–566
Industrialized countries
gender patterned time-use for fertility recovery in 9:775–780
infant and child mortality in 12:14–20
components 12:16–19, 12:18t
factors 12:19–20
geographic perspective 12:14–16
historical perspective 12:14–16
medical causes 12:18–19
regional disparities 12:15–16, 12:16t
secular decline 12:14–15
in twentieth century 12:16t
infertility in 12:46–48
internal migration in 12:443–446
Industrialized societies
adult mortality in 1:170–180
craft production in 5:111–112
folk crafts production for national market 5:111
individual craft production for pleasure 5:112
outwork and workshop handcrafts 5:111
persistence of customized craft production 5:111
definition 1:170
Industrial/organizational psychology 17:327–329
Industries
biotechnology 2:678–679
characteristics, and innovation 12:162–166
appropriability 12:165–166
demand 12:162–163
technological opportunity 12:163–165
collaborative research 7:864–865
cultural and creative 2:5
engineering 7:644–645
film as 9:165–166
international advertising as 12:459–460
production 12:329–330
and social networks 5:469
social policy litigation against 3:736
tourism 24:465
venture capital 25:59
white-collar workers 25:565–566
work anthropology 25:663–666
Industrious revolution 3:96
Industriousness, and conscientiousness 4:658–659, 4:662
Industry association 5:706
Industry chains, visualization 25:223–225
Industry codes, comparability, multiple imputation for 16:91
Industry cycles 5:651
Industry life cycles 5:651
Industry platforms 3:38–40
architecture 3:38
Industry structure
cultural production 5:608–609
design and density 20:14
for-sale *versus* income-producing properties 20:14
ownership 20:13
property type 20:13–14
Industry-led training, in informal education 7:122
Industry-level financialization, impact of changes in 22:194
Industry-level patenting comparisons 11:837–838
Industry-transfer demonstration parks 23:211
Indymedia 22:514

INED *see* Institut National d'Études Démographiques (INED)
Inefficient institutions 12:193–194
  persistence of 12:194
Inegalitarian societies 11:895–896
  affluent foragers and farmers 11:895
  big men, big women 11:895
  chiefs 11:895
  chiefs to kings, kinship to class 11:895–896
Inequalities 3:755, 4:155, 5:662, 6:290, 7:388, 7:906–908, 16:465–466, 16:787, 17:648, 18:733, 18:736, 20:295, 20:851, 20:854, 22:325, 20:295t
  of access 7:899–900
  of bargaining power 18:228–229
  and brokers 2:871
  and capitalism 3:105–108
  categorical 11:909–912
  in China 2:68–69
  class 10:220
  colonialism 11:897–898
  in comparative perspective 22:708
  of conditions 7:903
  consequences 22:711
  conspicuous consumption/distinction 4:695, 4:697–698
  critiques 7:899–901
  cultural 22:331–332
  and culture 23:851–852
  definition 11:894
  earnings 6:765–772
  economic 7:48
  and economic growth 7:43
  in education 7:261–262, 7:264
  of educational opportunity 7:904–905
  egalitarian 11:894–895
    assertively 11:894–895
    relatively 11:895
  and elder care 7:336–337
  ethnic 8:143
  ethnicity-based 8:130
  evolution of 11:908–909
  exploitation 11:910–911
  finance 9:171–175
  framing 11:910–911
  functions and dysfunctions of 22:707–708
  gathering societies 11:895
  gender 22:710–711
  see also Gender inequality
  and generation 9:853, 9:855
  health 7:741–742, 10:629–630, 22:286, 22:289
  and health 10:198, 10:201–202
  in health 10:635
    economic inequalities 10:635
  hoarding 11:910–911
  and homicide 11:163
  horticultural societies 11:894–895, 11:897
  household wealth 25:454–455, 25:454t
  and human development 6:307–312
  human rights 11:898
  hunting societies 11:895
  income 11:739–743
  and indigeneity 11:759
  industrial societies 11:894
  inegalitarian 11:895–896
    acquired economic status 11:896
    affluent foragers and farmers 11:895
    ascribed honor 11:896–897
    big men, big women 11:895
    chiefs 11:895
    chiefs to kings, kinship to class 11:895–896
    dimensions 11:896
    feudalism and estates 11:896
    stratification 11:896–897
  information inequality 12:87–88
  legal aspects 7:894–902
    antisubordination 7:896, 7:899
    causes 7:894
    color blindness 7:894–896
    courts, role of 7:894–898, 7:900–901
    critiques of equal access theory 7:900–901
    critiques of formal equality 7:896–897
    critiques of outcome theories of equality 7:899–900
    critiques of structural inequality 7:898–899
    formal inequality of treatment 7:894–897
    full participation 7:897–898, 7:901
    inequality of access 7:899–900
    inequality of outcome 7:899–900
    post-racialism 7:895
    racial polarization 7:895, 7:899–900
    remedying 7:894–896
    second generation discrimination 7:896
    stigma 7:895, 7:898
    structural inequality of treatment 7:897–899
  legitimate sources of 7:910–911
  and long-term heterosexual relationships 20:914
  maintenance, and meritocracy 15:241–242
  Marshall on 14:626–627, 14:629
  measurement issues 11:732–738
  migration 11:898
  and migration 15:429–431
  modern egalitarian societies 11:898
  and mortality 15:852
  opportunity hoarding 11:910–911
  of outcome 7:899–900
  patron–client relationships 11:896
  peasant societies 11:894, 11:896
  political representation 18:443
  and poverty 18:719–721
  psychological foundations of 11:909–910
  queer anthropology 11:897
  and race–crime relationship 19:842
  racial 6:542–543, 11:163–164, 19:857–862
  racial, and health 19:817
  and racism 19:853, 19:855
  reaction to 23:150–151
  and recognition politics 20:45–48
  refugees 11:894, 11:898
  regional 23:149
  relative deprivation 6:215, 6:217
  and sexuality 21:819, 21:821–822
  slavery 11:897–898
  social 11:894, 11:908–913, 23:148–149
    in practice 11:910–912
  social, water and health 25:431–433
  social construction of 11:910
  social justice 22:392, 22:394
  and social justice 7:909
  and social problems 12:544
  socioeconomic 10:218, 10:222–223
  stereotype content model 11:909
  stratification 11:908–910, 11:912
  structural inequality 7:899
  structure 22:708–709
  and student achievement 5:345–346
    economic segregation 5:346
    family inequality 5:345–346
    school inequality 5:346
  and sustainable development 23:818, 23:820
  symbolic boundaries 23:850–851
  trickster as response to 24:654–656
  underclass 24:731
  urban 23:150
  in urban planning 19:823
  and urbanization 24:963
  and whiteness/white supremacy 25:569–572
  who, what, where, when, why, and how 11:732–735
    how, summary statistic and data source 11:734–735
    what, choice of resource measure 11:733
    when, choosing a time period 11:734
    where, measures adjusted for differences in geography 11:734
    who, standard of living of 11:733–734
    why, purpose of measure 11:736–737
Inequality analysis 11:885–893
  basic building blocks 11:885–886
  consequences of inequality 11:890–892
    conflict between subgroups 11:891
    husbands and wives wage inequality on marital cohesion 11:892, 11:887t
    public benefit of religious institutions 11:891–892
    sense of justice 11:891
  empirical 11:892
  framework for 11:885–886
  measures of 11:886, 11:886t
    probability distributions 11:886
    quantile functions 11:888f–889f
  preventing inequality effects among a family's children 11:892
  theoretical 11:886–892
    determinants 11:888–890
    properties and relations of inequality measures 11:888
Inequality aversion
  experimental economics 8:550
  rational choice theory 19:903–904
Inequality aversion model 2:447, 19:903
Inequality index 11:902
Inequality measurement 11:899
  axioms 11:902–904

components of problem 11:900–902
income and income distribution 11:900
indices 11:900
ranking and dominance 11:900–902
decomposition 11:906
implementation 11:906–907
income differences 11:906
social welfare and 11:904–906
income 11:904
ranking distributions 11:905–906
Inequality space (social class) 3:755–756
Inequality transfers 11:888
Inequality-adjusted Human Development Index (IHDI) 5:778
Inequitable distribution of environmental hazards 7:799–800
Inequities
and justice 12:950, 12:952
structure and process of 13:232
Inequity distress 12:929
Inert knowledge 12:326
Inertia 5:2
organizational path dependence 17:388
INES *see* Indicators of Education Statistics (INES) program; International Network for Engineering Studies (INES)
I-ness 5:622
Inevitability thesis 11:815
INF *see* Intermediate-Range Nuclear Forces (INF) Treaty
Infamy, Roma people 20:750–751
Infancy 6:747–750, 12:7–13, 24:174, 24:176
attachment assessment beyond 12:231
defiant behavior during 6:12–18
development, effect of stress on 23:577–582
development of attachment during
see also Attachment
developmental research 6:326–329
early, mother–child communication in 15:874–876
emotional development 12:1–6
emotions in 12:4
empathy during 7:551–553
friendship during
see also Friendship
identity in 11:538–539
imitation during 11:621–627
joint attention during 12:844–850
life span development 14:109
like me theory 15:131–134
memory development during 15:147–154
motor development in 15:971–977
neural mechanisms of cognitive development in 4:58–65
electroencephalography 4:60, 4:62
event-related potentials 4:60–63
face processing 4:59–63
functional bases of 4:60–62
functional magnetic resonance imaging 4:60–62
functional near-infrared spectroscopy 4:60–62
future research 4:62–64
language 4:58–59, 4:61–63
magnetoencephalography 4:60
memory 4:58–59, 4:61–62
numerical cognition 4:62
structural brain development 4:58–60
nonverbal tests and developments in 21:896
and perceptual learning 10:127, 10:129–133
play during 18:218–225
differences in appropriateness of play 18:219–220
differences in object-oriented play 18:220
differences in parental investment in play 18:219
differences in physical play 18:220–221
differences in social-oriented play 18:220
similarities in game playing 18:219
similarities in person- and object-oriented play 18:219
prediction 12:8
pretend play in cognitive development
see also Pretend play
prosocial behavior during
see also Prosocial behavior
psychological development in 16:506–507
and self 21:439–443
significance 12:11–12
social referencing during
see also Social referencing
socialization in 22:866–872
spatial cognition development during 23:116–125
stability 12:10–12
as stage of life 12:8
structural brain development 16:600–601
studies 12:8–11
agency 12:9–10
basis of comparison 12:10
challenges 12:8
life course 12:10–11
period of multiple adjustments 12:9
platform for theory testing 12:8–9
time of socialization 12:9
Theory of Mind development during 24:268–273
trust across life span 12:638
Williams syndrome 25:579–582
see also Childhood
Infancy defense 3:481
Infancy phase, breast-feeding 5:510
Infant and child survival, multilevel models 16:52–55
conditional probabilities 16:54–55
fixed coefficients 16:53
intraclass correlations 16:55
marginal probabilities 16:54–55
random effects 16:53–54
three-level hazard model 16:52–53
Infant bonding *see* Neonatal bonding
Infant care 3:393
Infant crying 16:475
Infant development 18:852–860
attachment 18:854, 18:856, 18:859
brain development 18:853, 18:859
childhood 18:852, 18:856–857
depth perception 18:853
dynamic systems theories 18:855
early origins 18:854–855
future directions 18:858–859
grammar learning 2:651–652
imitation 18:853–854, 18:856–858
learning 18:852–858
media use and 15:35
mileposts 18:853–854
from 0 to 6 months 18:853–854
from 6 to 12 months 18:854
from 12 to 24 months 18:854
neural substrate for language development 2:652–653
prefrontal cortex development 18:859
reflexes 18:852–853, 18:857–858
research 18:855–858
case of body representations 18:857
ecological studies 18:855–857
general developmental processes 18:857
individual differences 18:857–858
resilience 18:857–858
transdisciplinary studies 18:858
risk factors and individual differences 18:854
social development 18:856, 18:859
sound discrimination 2:650–651
sudden infant death syndrome 18:853, 18:857–858
transactionalism 18:855, 18:858
word learning 2:651
Infant face preference 8:651
Infant feeding *see* Breastfeeding
Infant gaze behaviors 12:27–31
change in focus 12:29
current and emerging direction in theory and research 12:29–31
autism spectrum disorders 12:30
language development 12:29–30
robotics 12:30
theory of mind 12:30–31
current knowledge 12:27–29
alternating gaze 12:28
gaze cueing 12:27–28
gaze following 12:28–29
visual scanning of face 12:27
eye gaze 12:27–29
intentions 12:27, 12:29–30
looking 12:28–29
mental states 12:27, 12:30
seeing 12:28
self-experience 12:30–31
social cognition 12:31
theoretical background 12:27
visual perception 12:30
Infant mental health, and social-emotional development 22:943, 22:947
Infant mortality 2:699, 15:857–860, 15:858f
decline in 15:868, 15:870–872
and fertility transition 9:96
in industrialized countries 12:14–20
in less developed countries 12:21–26

Infant mortality rate (IMR) 6:119–120, 12:14, 12:19–20
Infant–caretaker bonding 16:472–473
Infant-directed speech (IDS) 6:728, 8:733
  see also Child-directed speech (CDS)
Infanticide
  in China 9:47
  female 15:608, 15:611
  see also Abortion
  perinatal 17:723
Infantile amnesia, and memory development 15:151
Infantile neurosis 24:546
Infant–mother attachment
  see Mother–infant attachment
Infants 16:387–392
  action planning 19:69–70
  brain development 16:393
  control in 19:68–69
  development of EFs 8:490–492
  early life, fathers' diaries of 16:482–483
  goal-directed behavior 10:268–272
  and media 15:1–2
  memory 6:759
    direct assessment of 6:761–763
  and NIRS 16:393
  sequential actions 19:69
  social mind, developmental psychobiology of 16:481–482
Infatuation 9:650
Infections
  bioarchaeology 2:601
  psychoneuroimmunology 19:447
Infectious diseases 10:199–200
  bioarchaeology 2:600–601
  and children/infant mortality 12:23
  and mortality 15:857–860
  and natural disasters 16:316
  psychological aspects 12:32–36
    behavioral changes 12:32–33
    colds and upper respiratory illness 12:35
    experimental studies 12:34–35
    exposure to pathogens 12:32
    future directions 12:36
    human immunodeficiency virus 12:35–36
    immunology 12:32–35
    inflammation 12:34–36
    influenza 12:34–36
    longitudinal designs 12:34
    measures of health behavior and biological factors 12:34
    other infections 12:35–36
    psychological influences of symptoms 12:35
    recurrence of herpes virus infections 12:35
    standardized measures of symptoms 12:34
    stress 12:32–35
    suppression of bodily defenses 12:33
    symptom reporting 12:33
    verification of infection 12:33–34
  social support and recovery 22:736
  spatial analysis of 15:88
Inference methods, dynamic economic systems 15:298–300
Inference models
  meta-analysis 15:272–273
Inference statistics 16:926–927
Inference-decision spectrum 16:116
Inferences 12:37–44, 19:63–64, 19:394–399, 21:585
  agent-based models 1:371
  availability of information and reader's knowledge 12:40
  backward 12:39
  backward inferences 19:395
  Bayesian 2:373–376
  and brain 12:41–43
  bridging 12:42–43, 19:395
  causal 12:39, 12:42
  causal relatedness 12:41–43
  coherence 12:39–43, 19:395, 19:397–398
  coherence break 12:41–43
  as component of discourse comprehension 12:40–41
  computational processes 19:396–397
  constructionist theory of 12:40–41
  constructivist theory of 19:395
  correspondence models 16:121–122
  cortical networks 19:397–398
  differentiated view 12:39–40
  ecological 6:868–870
  elaborative inferences 12:39–40
  on expectations in economics 19:777
  explicit 19:394–395
  fallacy of composition 11:818
  forward 12:39–40
  forward inferences 19:395–396
  heuristics for 5:893
  in hierarchy models 10:836–837
  implicit inferences 19:394–395
  and implicit representation 13:78–79
  individual differences 19:398–399
  and intelligence 12:316
  invariance 12:705–707, 12:709
  joint action 12:839–840
  joint commitment 12:842
  knowledge 19:394–399
  language 19:394–395, 19:397–398
  legal reasoning 13:777, 13:782–783
  methodological issues 19:395–396
  minimalist theory of 12:40–41, 19:395
  neural bases 19:397–398
  off-line research methods 12:38
  online research methods 12:38–39
  possible from generalized linear mixed models 9:847
  pragmatic 12:39
  process character 12:39–40
  research, issues in 12:37–38
  research findings 12:39
  right hemisphere language 19:397
  in sample surveys 20:886–887
  types 19:395
  world knowledge 12:40–41, 12:43
Inferencing 13:268–269
Inferential I 8:353
Inferential II 8:353
Inferential methods
  generalized linear mixed models 9:847–849
Inferential relationship
  reasoning 5:934
Inferior births 8:234
Inferior frontal gyrus (IFG) 11:331, 13:404, 16:602–603, 16:673, 16:603f
Inferior frontal junction (IFJ) 16:673
Inferior lateral prefrontal cortex 13:681f
Inferior olive
  eyeblink conditioning 8:638–639
Inferior parietal lobule (IPL) 18:847
Inferior temporal cortex (ITC) 25:160–161, 25:201
Inferior temporal gyrus (ITG) 13:370
Inferotemporal cortex 24:206
Infertility 10:465
  and adoption 1:138
  demographic aspects 12:45–49
    assisted reproductive technologies 12:47–48
    childlessness 12:47–48
    definitions 12:45
    developing countries 12:45–46
    importance of fertility 12:46
    industrialized countries 12:46–48
    local definitions 12:45–46
    postponement 12:46–48
    primary infertility 12:45–47
    secondary infertility 12:45–46
    sexually transmitted diseases 12:46
    voluntary childlessness 12:46–48
  and sexuality 21:843
Infibulation 8:879–881
Infidelity, sexual, and religion 20:319–320
Infinite games 9:576
Infixes 15:821
Inflammation 12:34–36
  coronary heart disease 4:920–921
  and epilepsy 7:853–854
Inflammatory bowel disease (IBD) 9:605–607
  anxiety and 9:608
  coping with 9:607
  depression, anxiety, and stress in 9:608
  depression and 9:608
  psychoneuroimmunological pathways in 9:606f
  self-image and 9:608
  social support, effect of 9:607
  stigma and 9:607–608
  stress and 9:608
Inflation
  macroeconomic data 5:753–755
  macroeconomics 14:406–407, 14:409, 14:411
Inflation crisis 9:183–184, 9:188
Inflectional knowledge 3:511–512
Inflections 15:821–822
  TAME 24:210–212
Influence
  in action 18:772–773
  Dahl on 5:658–659
  history 18:770

literature and review 18:770
in organization 18:771
research on 18:774
nonlinear effects of 18:774
positive effects 18:774
and sanctions 20:890–892
social *see* Social influence
Influence functions 20:732–733
examples 20:732–733
Influence networks 15:660–662
Influenza 12:34–36, 18:174
Influx of Wealth 18:714–715
INFORM (Information Network Focus on Religious Movements) 16:808
Informal adoption 1:136
Informal associations 12:58
Informal caregiving 3:135–137
Informal control 4:223–224, 4:835–836
Informal curriculum 3:13
Informal economics 12:50–55
definitions 12:50–51
exclusions 12:51
denominations 12:50–51
direct approaches 12:51
discrepancy approaches 12:51–52
empirical estimates 12:52–54
indirect approaches 12:51–52
measurement approaches 12:51–52
multiple causes approach 12:52
multiple indicator 12:52
Informal economy
in Eastern Europe 6:818–819
and exchange 8:480
peasants 17:678
Informal education 7:919
and formal education, difference between 6:896–897
Informal education system 7:122
Informal empire 11:687
Informal employment 13:174, 13:180–181
categories 12:60–61
estimate of 12:61
minimum wage legislations and 12:63
in nonagricultural sectors 12:61
productivity 12:62
in total nonagricultural employment 21:865t
trade liberalization and 12:63–64
working conditions 25:701
Informal group learning methods 4:884–885
group investigation 4:885
jigsaw 4:884
learning together 4:884
Informal housing 11:234
Informal industrial relations 25:689
Informal instruction, in transition to adulthood 24:580–581
Informal interviews, in ethnography 8:187
Informal jobs, defined 12:61
Informal justice institutions 13:810
Informal networks, global governance 10:156
Informal oligarchs, rule of 17:202
Informal organization
negotiated orders 17:412–413
Informal organizational life 2:914
Informal policing 18:248
Informal sanctions, social control 4:822
Informal science education 21:230, 21:233–234
Informal sectors 12:56–59, 24:818–819
in Africa 24:967–969
collective organizing 12:58
informal associations 12:58
informal settlements 12:57–58
licensing 12:56–57
market liberalization 12:58
market meets informal economy 12:58
meanings of informal 12:56
microenterprise 12:57
neoliberalism 12:58
networking 12:57–58
participatory approaches 12:57–58
policy concerns 12:57–58
regulation 12:57
savings groups 12:57–58
sector/process 12:56–57
urbanization and development 12:56–57
and West 12:57
Informal settlements 12:57–58, 18:185
Informal social control, and collective efficacy 4:170
Informal social control theory 5:138
Informal social networks 16:549
Informal survivalism 24:819
Informal work
definition 12:60–62
in developing countries 12:60–65
macroeconomic correlates 12:60–62
measurement 12:60–62
Informal-based vocational education and training 25:249
Informality 13:174
Informality, urban 24:819–820
Informalization 24:820, 25:50–51
Informants
in ethnography 8:187
fieldwork 9:149–151
Informatics 9:992, 14:243
biomedical research ethics 8:81
Information 14:367, 19:320–321, 20:516
accessibility 22:207
archive, release timing to 1:919
availability, and inference 12:40
balanced processing of 2:247
and beliefs 17:780, 17:783
definition 12:72
economics of 12:100–105
common sense precursors 12:101
econometric methods 12:102–103
equilibrium price dispersion 12:102
future developments 12:104–105
intuitive precursors 12:101
two-sided search and bargaining 12:102
wage dispersion 12:102
episodic-autobiographical 13:680
explicit storage of 13:78–79, 13:82
integration in neurons and dendrites 16:703–707
knowledge and 13:93
and libraries 14:47–48
and media 15:11–14, 15:16
organizational 12:72–76
changes 12:74–75
communication 12:73–74
evolution 12:72
expertise 12:75
informal communication 12:73
innovation 12:73–74
messages 12:73–74
properties 12:72–73
sense making 12:72–73
perceptual 10:128
and postindustrial societies 18:666–668
presenting in dictionaries 13:938–939
quantification in bits 4:103
and representation 20:67–68, 20:70
retrieval of 13:681–682
and science communication 21:222–224
shortage 22:850
storage of 13:681
supply chain management
asymmetry 23:711
distortion 23:711
type, effect on social referencing 22:559
modality 22:559
valence 22:559
use of term
Fisher's 9:223
Information access
about randomized controlled trials 8:544
autonomous agents 2:311
Information acquisition 16:796
Information age 18:666
charisma in 3:352
economic restructuring 12:91–92
effects on transportation 24:625
geographies of urban restructuring 12:91–92
Information aggregation, and auction 2:208
Information and communication technologies (ICTs) 3:229, 4:667, 5:705, 6:66, 7:34, 10:263, 10:266, 11:838, 12:83, 12:87, 12:90, 12:587–588, 14:287, 14:289–291, 15:12, 16:787, 16:789, 16:791, 16:798, 21:150, 21:154–155, 21:648, 22:113–117, 24:621
Arab world in perspective 16:799–800
and biocitizenship 2:610–611
Castells on 3:229–231
definition 4:493
in education 4:495–496, 12:66–71
digital resources 12:66–67
education transformation 12:66, 12:70
e-learning 12:66–67, 12:69
learners 12:68–70
pedagogical issues 12:67–68
teacher knowledge 12:68
teachers 12:67
gender and technology 9:723, 9:726–727
and globalization 10:262–263, 10:266

Information and communication technologies (ICTs) (*continued*)
ICT-enhanced learning environments in curriculum 4:496
instructional psychology 12:204
literacy 21:120
organizations 17:311
planning methods 24:890–892
in schools 4:496
service activities across organizations and space 21:651–652
Information and computing technologies (ICT) 3:698
Information asymmetry 1:344–346
in health care 10:615
Information bias, evidence-based medicine 8:329
Information Bureau of Communist and Workers' Parties (Cominform) 4:303
Information content measures 21:581
Information costs, in voting 25:309
Information criteria, model selection 15:647–649
Akaike's information criterion 15:648
focused information criterion 15:648–649
Mallows' *Cp* 15:648
Schwarz Bayesian information criterion 15:648
value of R^2 in linear regression models 15:647–648
Information deficit model 3:844–845, 3:844f–845f
Information discrepancy, in combined pharmacotherapy and psychotherapy 19:497
Information economy 12:90
Information filtering, autonomous agents 2:311
Information flow decisions, supply chain management 23:710
Information flows
and mass media 14:712–713
in organizations 3:894–895
Information hiding, modularity and 15:718–719, 15:721–722
Information infrastructures 21:186
Information literacy, and libraries 14:46, 14:48
Information loss and statistical inference, in spatial autocorrelation 23:109
Information model, consumer policy 4:734
Information Network Focus on Religious Movements *see* INFORM (Information Network Focus on Religious Movements)
Information problems 19:536–537
Information processing
and authoritarianism 2:265
cognitive development 4:51
and creativity training 5:115
discrete-state models of 6:523–529
dual-process models of 22:207–208
impact of anxiety on 24:246
of intelligence 12:290–295
and short-term memory 21:902–905
and social cognition 22:206–207
Information processing capacity 6:75
Information processing system, in humans 21:904–905
Information propagation, and collective memory 4:191
Information regimes, cultural 12:590–591
Information retrieval
knowledge representation 13:100
Information revelation, and auctions 2:208
Information sciences
cultural analysis 9:365
ethnomethodology 21:197
mathematical psychology 14:808, 14:814–815
Information seeking 21:272
of infants 22:557
Information set 12:103
Information shortcuts 2:194
Information society 12:83–89, 13:107, 15:12
agenda 12:87–88
communication media 12:84, 12:86
cyberspace 12:84, 12:88
geography of 12:90–94
inequality 12:87–88
informatization 12:85–88
informatization and 12:83–84
mass communication 12:85
media 12:83–86, 12:88
power and 12:85, 12:87–88
situating 12:90
studies on 12:84–86
telecommunication 12:83–84, 12:86–87
thesis, critiques of 12:86–87
toward new geography of 12:92
Information society statistics 5:706, 5:708
Information society thesis 12:86
Information sources, direct democracy 25:295
Information structure
explanations of grammar 9:471
in linguistics 12:95
assertion 12:95–96
categories of information structure 12:96–97
common ground 12:95–96
communication 12:95–96
focus 12:95
knowledge 12:95
presupposition 12:95–96
question 12:95
research traditions 12:97
rheme 12:96–97
theme 12:96
theoretical background 12:95–96
topic 12:96
Information superhighway 12:87
Information support decisions, supply chain management 23:710
Information technology (IT) 21:186–191, 24:104–105, 24:613–614, 24:765
and archives 5:702
congestion 24:633
in education, research on 4:497
and genealogy 9:808–809
multinational corporations 16:79–80
and psychotherapy 19:379–381
required for health information systems 5:786
in society 22:831
social networks 22:831
social Web 22:831
transfer of 24:144–145
and university press 19:624
Information theories 12:106–109
applied in neuroscience 12:108
applied in psychology 12:108
arts socialization 2:7–9
channel 12:106–108, 12:106f
entropy 12:107–108
logarithm 12:107
noise 12:106
probability 12:107–108
quantifying information 12:106–108
Information technology (IT)-to-IT network 4:486
Information transfer 9:477, 12:418–419
callosal connections 12:419
Informational approaches, for physical activity 18:95–96, 18:96t
Informational capitalism 18:668
Informational cascades 23:369
Informational resources 21:135
Informational roles, managers 14:473–474
Informational safety campaign 25:729
Informational semantics, and intentionality 12:334
Informational social support 1:131–132
Informational society 12:85–86
Castells on 3:230
Informational-aesthetic conception of consciousness development 4:668
Informationalizing 5:246–247
Information-based policy instruments 19:572
Information/decision-making
and organizational diversity 4:738
Information/decision-making, and organizational diversity 6:595–596
Information-motivation-strategy (IMS) model 6:617–618
for patient adherence 17:600
Information-processing architectures 12:77–82
automatic processing or automaticity 12:79
challenges 12:81
issues 12:77–79
discrete flow 12:78
exhaustive processing 12:78
independence *versus* dependence 12:78
RT durations 12:78
serial *versus* parallel 12:78–80
seriality of operation 12:78
stopping rule 12:78, 12:80–81
workload capacity 12:78, 12:80–81
perceptual and cognitive holism 12:79
serial *versus* parallel processing 12:79–80
use of term 12:77
using factorial technology 12:80

Information-processing mechanisms
aging deficits in 1:430–432
attentional and inhibitory mechanisms 1:431
processing speed and variability 1:431
resource-reduction account 1:431–432
working memory 1:430–431, 1:431f
Information-processing models 17:873–875
cognitive neuroscience perspectives 17:874–875
cognitive-adaptive perspectives 17:875
and learning 13:708–709, 13:708f
Information-processing theories 22:331, 22:333–334, 7:239t
for aggression 1:375
of self-regulated learning 21:542
Information-processing view, quantitative memory models 15:161
Informativeness 1:346
Informatization 12:83, 12:85–88
and information society 12:83–84
and modernization 15:711–712
Informed choice 9:911
Informed consent 1:735, 2:616, 12:610, 17:534, 19:320, 20:516, 20:518
autonomy 20:510
and bioethics 13:559–560, 13:562–563
biomedical research ethics 8:79–81
clinical application 19:322–323
case example 1 19:322
case example 2 19:322
in clinical care 19:323–324
cluster randomized trials
feasibility of consent 3:902–903
in organizations 3:903
requirements 3:902
timing of consent 3:902
and cultural nuances 20:510–511
in developed and developing countries 20:510–511
documents, content and understanding of 20:511
elements of 19:320–322
decisional capacity 19:321–322
information 19:320–321
voluntarism 19:322
fieldwork ethics 9:153–154
in observations of group interactions 9:154
in genetics 2:621–622
human research participants 4:254–256
elements 4:255t
general requirements for 4:254–255
waivers of the general requirements for 4:256
of human subjects in research 8:48–51
and internet 12:610
in interviewing 5:853
McKee and Porter's dimensions for 12:609f
neuroethics 16:666–667
psychiatric ethics 8:66
in Psychiatry 19:320–325
of research subjects 12:110–115
autonomy 12:110–111, 12:113–114
coercion 12:112
in community intervention projects 12:113
decision-making capacity 12:111–112
disclosure 12:111
future 12:114
history 12:110–111
in internet-based research 12:113–114
in observational studies and field work 12:113
privacy 12:113
in studies based on review of records 12:112–113
undue influence 12:112
voluntariness 12:112
risk screening 20:700
in social psychology 22:536
in social research 18:946
in surveys and polls 23:755
Informed decision making 9:912
Informed instruction, metacognition 24:93
Infosys BPO Limited 3:48
Infractions 15:591–592
Infractions, conflict 4:622
Infradian rhythms, regulation by hypothalamus 11:488–489
Infrahumanization 6:34–35, 18:829
and prejudice 18:836
Infrahumanization theory 6:35, 12:413
Infra-individual determination 15:311–312
Infralimbic cortex 8:872
Infra-reflexivity 20:94
Infrastructural developments, time use research 24:383–384
American time use study 24:384
MTUS data archive 24:383–384
Infrastructure 4:273, 12:119–125, 24:911
agriculture 20:802, 20:804, 20:806
analytical uses 12:120–122
dys-appearing 12:120–122
ecologies 12:120
and statecraft 12:122
computer-supported cooperative work 12:120
cyberinfrastructure 22:602
discipline-building 6:491–492
embeddedness 12:119–120
energopolitics 12:122
and energy 7:624
epistemological issues 12:122–123
infrastructuring 12:119–124
interpellation 12:121
media as 22:583
methodological issues 12:123–124
as ethnographic resource 12:124
inversion and othering 12:123
relationality and modes of research 12:123–124
research 22:596
research and teaching 22:667–673
academic research system 22:669–670
applied research 22:670–671
cross-disciplinary orientation 22:669
development after 1945 22:667–668
empirical research 22:667–670
Erasmus's 22:672
extra-university institutes 22:669–670
institutionalization 22:667–670
international training 22:672
national environments and institutional structures 22:668–670
project funding 22:670
publicly financed research institutions 22:670
scientific evolution 22:667
strengthening European research base 22:670–672
scholarly 22:596
transportation 24:909
urban anthropology 12:119–122
utilities 12:120–122
Infrastructure for Spatial Information in Europe (INSPIRE) 23:130
key data themes 23:131t
Infrastructure investment 12:116–118
evolution from private to public function 12:116–117
public infrastructure investment 12:117
as public/private practice 12:117–118
access to private investment markets 12:118
emerging cites in urban world 12:117
financial practices 12:117–118
special purpose governments 12:117
and urban development 12:118
Infrastructure perspective, on diffusion 6:408–409
Infrastructuring 12:119–124
Ingelfinger rule 20:532
Ingenieros, Jose 5:239
Ingenuity 17:231
Ingestion, food, and social networks 16:557–559
Inglehart, Ronald, on political generations 9:862–864
Ingold, T., works, *The Perception of the Environment* 25:755
INGOs *see* International nongovernmental organizations (INGOs)
Ingram, Paul 8:710–711
Ingratiation, social influence 22:350, 22:352–353
fear appeals 22:353
In-group bias 12:410
Ingroup favorability, and discrimination 18:831–832
Ingroup favoritism bias 18:834
In-group identity 11:528–530
In-group process, organizational groups 10:412–413
In-group wrongdoing 7:505
Ingroup-outgroup 25:403
Ingroups
in collectivism/individualism 4:207
apology for transgressions 4:209
immigration 11:675–676
Inherent features 13:943
Inheritance 11:820, 13:758–759
of acquired characteristics 8:383–385
concept of 7:836–837

Inheritance (*continued*)
and dowries 6:643–645
of epigenetic modifications 7:842
and evolution, broader concept of 7:836–837
Genographic Project 10:23, 10:28
Mendelian *see* Mendelian inheritance
monogamous 14:617–618
of urban policies in Europe 24:902
Inheritance practices, by siblings 21:923, 21:926
Inheritance tax 23:341t
Inherited sources of sources 6:498–499
Inhibition 4:46–48, 12:416–418, 21:612
in bilingual aphasia 2:574, 2:576–578
cognitive function 6:726
by opsins 17:271
Inhibition control 8:492
Inhibitory cortical interneurons (ICIs) 2:823
Inhibitory gain 25:159
Inhibitory mechanisms, and aging 1:431
Inhibitory synaptic input 16:704
In-homogeneous structure 17:299
In-house counsel 12:126
compliance 12:126–128
corporate social responsibility 12:127
delimiting the role of 12:126–127
departmentalization 12:126–127
lawyers 12:126–128
legal profession 12:126–128
as organizational actors 12:127
as professional actors 12:127–129
professional autonomy 12:128
professional competition 12:127
professional independence 12:127–128
professionalization 12:126–128
staff function 12:127
In-house issues 4:683
In-house legal counsel 13:578–579
Inhuman treatment, torture and 24:442, 24:444–446
Inhumanity 11:370–371
Inhumanity thesis of human rights 11:370–371
INI *see* Index Number Institute (INI) (US)
Initial conditions, economic transformation 7:74–75
Initial discrete-time hazard model
fitting person-period dataset 23:788
substantive predictors in 23:789–790
Initial dividend yields, stock market prediction from 23:486
Initial explanatory frameworks 21:275
Initial Knowledge and Skill Standards for Gifted and Talented Education 7:160
Initial P/E multiples, stock market prediction from 23:486–487
Initial public offering (IPO) 4:940
Initiations
cognitive function 6:725–726
hazing 10:555–558
rites 1:767
rituals, and gender development 9:745–746
rituals, female genital mutation in 8:878–879
and sacrifice 20:860
and transition to adulthood 24:577
Initiatives, direct democracy 25:294
Injecting drug users (IDUs)
AIDS-related mortality 15:835
Injunctions, remedies and damages and 20:413
Injuries
bioarchaeology 2:601
and natural disasters 16:315–318
prefrontal cortex, symptoms of 18:815–816
disturbances of motor function 18:815
emotion and social behavior 18:815–816
environmental control of behavior 18:815
loss of divergent thinking 18:815
temporal memory 18:815
prevention of 12:132–133
in temporal lobe 24:208–209
unintentional *see* Accidents
workplace 20:863–869
Injurious experiences 12:136
Injustice
assimilation of 2:489
and conflicts 12:939
context and particularity 12:136
and health 12:928
law and 12:136–137
legal aspects 12:135–139
as a matter of malevolent human action 12:135
Merry's research 12:136
Moore's contributions 15:769–771
occupational stress management 17:148
reactions to 12:952–953
responses to 12:135–136
Shklar on 21:893–894
significance of procedure 12:137–138
study of 12:137
treatment of subject of 12:135
in urban planning 19:823
see also Justice
INL *see* Institut National du Logement Belge (INL)
Innate 14:367–368
Innate behavior, and neonatal bonding 16:472
Innate constraints, vocabulary acquisition theories based on 25:228
Innate ideas, rationalist doctrine of 24:722
Innate releasing mechanisms (IRM) 14:367
neonatal imitation 16:485
Innateness 12:343
Innatist position, in foreign language learning 9:328–329
Inner Asia 16:838, 16:841
Inner beast (soul) 18:211–212
Inner city 25:586–588
ghetto rise 12:140–141
labor market for unskilled work 12:141–142
and public policy 12:143–144
and residential segregation 20:541
Inner dialogue 4:730–731
Inner dynamic, of fashion 8:837
Inner man (soul) 18:211–212
Inner morality of law 12:933
Inner orientation 21:430
Inner senses 16:680, 16:680f
Inner speech, Vygotsky on 25:313
Inner suburbs 23:661, 23:663
Inner-directed network actors 5:558–559
Inner-directed social character 20:671
Innocence
and wrongful convictions
factual 25:785
legal 25:785
Innocence Project 25:787
Innocent cannibalism 3:86
Innocent victim 2:487
Innovation and diffusion, and law 13:601
Innovation ecosystems 22:360
Innovation life cycle 12:148
Innovation management, absorptive capacity 1:13
Innovation networks, network analysis 16:520–521
Innovation systems 12:147
Innovation-decision process 8:379
Innovations 3:12, 3:21–23, 6:666, 11:695, 12:170, 13:445, 20:360, 21:105–106, 21:141–144, 22:171, 22:355, 24:141, 24:763
actor network theory 12:171–172
animal social learning 22:405
and biomedicine 2:664, 2:667
and bridging TA 24:126
business models and 3:35
business platforms 3:37–42
classical theory 12:169–171
diffusion 12:169–170
entrepreneurship 12:170
social change 12:170–171
technological and behavioral changes 12:171
communication channels 8:379
in communication technology 12:73
contemporary theory 12:171–174
entrepreneurship roles and networks 12:172–173
social innovation 12:173–174
strategic reflexivity and innovation in organizations 12:173
technology-economic innovation 12:173
three perspectives 12:171–172
cultural transmission 5:404
and diffusion 6:409–410
diffusion of 8:378–381
compatibility 8:378
complexity 8:378
observability 8:379
relative advantage 8:378
trialability 8:378–379
economics of 12:160–168
appropriability 12:165–166
competition 12:165–166
demand 12:162–163

empirical studies in Schumpeterian tradition 12:160
firm characteristics 12:160–162
industry characteristics 12:162–166
market structure 12:160, 12:165–166
patents 12:161–162, 12:164–167
technological opportunity 12:163–165
effective 11:697
emerging directions and challenges 12:149–150
additive manufacturing 12:149–150
design and integrative thinking 12:150
management education 12:150
technological directions 12:149
and entrepreneurship 7:675–679, 7:681–685
and fashion 8:835, 8:837
future of theories 12:174
history 12:145–147
definition 12:146
diffusion of innovation 12:146
from linear to open 12:146
systemic nature of innovation 12:147
international migration effects on 12:517
and interorganizational relationships 12:622, 12:624
in law firms 13:583
management 12:148–149
knowledge management processes 12:148
leadership and organizational culture 12:148–149
measuring innovation 12:149
national systems of innovation 12:149
strategy 12:148
structure 12:148
and modernization 15:707–712
new communication technologies 8:380–381
opinion leaders 8:380
and organizational behavior 17:700–701
and organizational boundaries 2:769–771
organizational decision making for 17:354–355
and organizational group diversity 10:410
organizational knowledge 12:73–74
personalized medicine 17:928
in practice 11:696
and public bureaucracies 19:531–532
science and technology studies 12:171–172
and science communication 21:222–224
and social change 22:193
social entrepreneurship 12:173–174
social system 8:380
state of the art 12:152–159
aggregate empirical evidence 12:155
classical economists 12:153
compensation theory 12:152–157
empirical evidence 12:154–157
high-tech 12:156–158
microeconometrics 12:154–155, 12:157–158
microeconomic empirical evidence 12:156–157
process innovation 12:152–157
product innovation 12:152–158
R&D 12:154–157
sectoral empirical evidence 12:156
technological unemployment 12:152–153
theory 12:152–154
and telephone 24:162
time 8:379–380
early adopters 8:379
early majority 8:379
innovation-decision process 8:379
innovativeness 8:379–380
innovators 8:379
laggards 8:380
late majority 8:380
rate of adoption 8:380
types 12:147–148
incremental or radical innovation 12:147
innovation life cycle 12:148
open or closed innovation 12:147
product or process innovation 12:147
technological or nontechnological innovation 12:147
user-led innovation and service innovation 12:148
universities 18:308–309, 18:311–312
and venture capital 25:59, 25:61–62
waves of 22:701
Innovative economic behavior 12:645
Innovative entrepreneurs 7:676
Innovative knowledge community (IKC) 16:552
Innovative teaching methods, for economic education 24:84–85
Innovator's DNA, concept of 21:105
Inns of Court 4:248
Inoculation 15:869
Inositol triphosphate (IP_3) 16:753
Inpatient Multidimensional Psychiatric Scale (IMPS) 19:291
Input, role in language development 13:313–314
Input intensity indicators, land use 13:240
Input interaction model, of foreign language theory 9:329
Input markets 1:478–479
Input model, of foreign language theory 9:329
Input prices regulation 20:169
Input–output analysis (IOA) 11:850, 14:762–763
Input–output approach
to small group research 22:104
see also Production function approach
Inputs, team performance measurement 24:96
Inquiring teacher paradigm 24:53
Inquiry cycles 4:797, 4:799–800
Inquiry habit of mind 16:553
Inquiry in psycholinguistics 14:227
Inquiry-based learning 21:231
in social studies education 22:731–732
Inquiry-based teaching 21:231
Inquisitorial preliminary investigation, adversarialization of 5:229–230
Inquisitorial procedure 5:227–228
adversarial procedure and 5:231
Inquisitorial style of the European codified tradition 13:587
Inquisitorial systems, scientific evidence and 21:297–298, 21:300
Inquisitorial trial judge, decline of 5:230–231
INS *see* Immigration and Naturalization Service (INS)
Insanity 5:219–220, 5:224
categories of 17:28
Insanity defense *see* Criminal responsibility
Inscriptions, and actor-network theory 1:82–83
INSEAD (business school) 3:176
Insects group 7:694
Insecure avoidant attachment 12:230
Insecure-ambivalent attachment 12:230
Insecure-ambivalent children 2:138
Insecure-avoidant infants 22:560
Insecurely avoidant children 2:138
Insecure-resistant attachment 12:230
Insecurity
and attachment 12:231–232
attachment across life span 12:232
continuity and longitudinal correlates 12:232
maternal sensitivity 12:231–232
food 9:285–289
Insecurity stage, of Alzheimer's disease 1:586–587
Inshore littoral species 7:694
Inside perspectives problem, Christian historiography and historical thought 11:31
Inside Prime Time 24:165
Insideness, and person-environment relations 6:886
Inside/outside perspective on Christianity 11:31
Insiders
perspective of theology 20:357
in queer anarchist politics 19:749
Insights 19:66
Köhler on 13:123–124
Insistent query 10:110–111
Insomnia 22:60–61, 22:70–71, 22:73, 22:79–80, 22:82–83
side effects (Psychopharmacotherapy) 19:465
Inspection programs 17:145
INSPIRE *see* Infrastructure for Spatial Information in Europe (INSPIRE)
Instability, of children born to cohabiting parents 9:16
The Instability of Capitalism 3:97
Installation of hope 10:430
Instance-based learning theory (IBLT) 6:712
Instance-/exemplar-based model 6:711–712
Instant urbanism 20:678
Instantaneous photographs 12:593
Instantiation, physicalism 18:106, 18:110, 18:112

Instincts 12:175–180
behaviorism 12:179
criticisms and alternatives 12:179–180
Darwin on 12:175–178
drive 12:177, 12:179
ethology 12:179–180
evolution and heritability of behavior 12:175–176
gestalt theory of 12:177
habit 12:175–177, 12:179
heredity 12:176, 12:180
in human psychology 12:176–177
impulse 12:175–180
McDougall on 12:178–179
motivation 12:175, 12:178–179
reflex 12:175–179
in social sciences 12:177–179
vitalism 12:177
Instincts theory for social psychology 22:527–528
Instinktbewegung 14:367
Institut d'Etudes Politiques 18:468
Institut International de Sociologie (IIS) 22:675–676, 22:992
Institut National de la Statistique et des Etudes Economiques (INSEE) 3:305–306
Institut national d'études démographiques (INED) 8:244, 10:776
Institut National du Logement Belge (INL) 24:884
Institute for Earth Education 7:735
Institute for Experimental Psychology, Wundt's contributions 25:794
Institute for Social Research (ISR) 5:830–831, 23:748–749
Institute for the Study of Genocide 10:7
Institute for the Study of University and Education (IISUE) 7:183
Institute for Women's Policy Research 8:902–903
Institute of Criminology (UK) 5:241
Institute of Education Sciences (IES) 22:299, 23:215
Institute of Ethnology (Paris) 1:752
Institute of Medicine (IOM) 15:115, 16:910
definition of CER 24:282
The Future of Nursing 16:910
Lesbian Health: Current Assessment and Directions for the Future 15:115–116
Institute of Social Science (ISS), Japan 22:641–642
Institute of Social Studies in The Hague, the Netherlands
Initiatives in Critical Agrarian Studies program 1:457
Institute of Southeast Asian Studies (ISEAS) 22:638, 22:641
Institute on the Holocaust and Genocide 10:7
Instituted process, economy as 18:244
Institutes 22:606
Institutes of Divine and Secular Literature 7:584
Institution of motherhood 15:882
Institutional abuse, of elders 7:328, 7:332
Institutional analyses 5:608
of sport 23:314–315
Institutional anomie 1:716
Institutional applied geography 1:846–847
Institutional approaches 8:126
data collection 5:707
to prosocial behavior 1:581
United States presidency 18:861–863
Institutional aspects, of audience measurement 2:214
Institutional authoritarianism 2:199
Institutional autonomy 21:216
Institutional backings, Bielefeld School 2:555–556
Institutional care 20:534–537
for learning disability 13:684, 13:686–687
Institutional censorship 14:244
Institutional change 22:979–980
economic transformation 7:76–77
endogenous change 22:982
welfare state reform dynamics through 25:524–525
Institutional child welfare 15:884–885
Institutional constellations 15:693–694
Institutional constraints, public policy 19:573–574
Institutional context of scientific de-Westernization 25:543–544
Institutional decentralization 18:188–189
Institutional deserts 6:229–230
Institutional design, Elster on 7:402, 7:404, 7:406
Institutional dimensions, biotechnology 2:678–679
Institutional discrimination, race and ethnicity 19:812, 19:816
Institutional economic thought, experimental economics 8:550
Institutional economics 7:59–60, 21:305
and property rights 24:528
Institutional entrepreneurs 16:766, 16:768
Institutional environment of hospitals 11:208–209
Institutional facts 21:344
feature 21:344
Institutional families (comparative) of police 18:248
Institutional fees 14:267
Institutional framework, of labor statistics 5:799
Institutional framing, Central Asia 3:317
Institutional homophobia, and permissible prejudice 10:792–793
Institutional initiative 2:350–351
Institutional innovations 22:356, 22:356t
and advocacy planning 1:221
Institutional investors, stock market 23:490–493
Institutional isomorphism 17:400
Institutional logics 5:537, 16:766–767, 17:377
challenges 16:767
contributions 16:767
and cultural sociology 5:552–554
as culture 5:555
culture and 5:550–556
diverse, conflicts between 11:869–871
and entrepreneurs 5:551
institutional logics perspective (ILP) 5:550–552
versus institutional order 5:551
justification of 5:554–555
and markets 5:469
microfoundations of 5:551
morality 5:554
religious phenomena 5:550–551
structure and agency of 5:551–552
symbolic management of 5:554–555
value spheres 5:550–551
Institutional logics perspective (ILP) 5:550–552
Institutional mandate 17:11–12
Institutional misconduct 21:313–317
ethos of science 21:314–315
institutionalized self-regulation 21:314–315
regulation of research ethics after WWII 21:315–317
Institutional models 16:443
Institutional narratives 16:215
Institutional order 5:551
categorical elements 5:551
versus institutional logic 5:551
Institutional pluralism 16:767–768
Institutional power 18:763–764
Institutional quality hypothesis, and prosocial behavior 19:240
Institutional racism 19:847, 19:860–861, 19:864–865, 22:823
Institutional recognition, social psychology 22:535–536
Institutional reflexivity 20:96
Institutional relations, and friendship 9:425
Institutional requirements for health information systems 5:785
Institutional research 10:864
higher education research 10:863–865
Institutional Review Boards (IRBs) 1:736, 4:252–254, 8:38, 20:515
criticisms of 8:51
formulation of 8:49–50
structure and functioning 8:50
Institutional secrecy 21:377–378
Institutional settings
cultural policy 5:463–464
fertility 9:19, 9:24
Institutional shareholder activism 4:934, 4:936
Institutional solutions, to collective action problem 19:541–542
Institutional structures, teaching and research infrastructure 22:668–670
Institutional support 4:323–324
and community 4:324
Institutional theory 2:484–485, 7:68, 17:376
economic sociology 7:68–69
encore adulthood 20:593–594
and imperialism 11:682–684
organizations 17:308
retirement 20:593–594

Institutional transformations, privatization and 18:188–189
Institutional variation, in pension policy 17:693
Institutional viability, debates on 17:605–606
Institutional work 16:765–766, 17:377
Institutional-anomie theory 1:721, 5:191–192
Institutional-evolutionary paradigm 7:74
Institutionalism 12:181–189
  and agency 12:184
  American 1:573–574
  change 12:184
  constructivist 12:188–189
  and diffusion 6:404
  discursive 12:189
  economics 12:186–187
  in economics 17:190–195
    approaches that emphasize institutions 17:194–195
    demarcation 17:193–194
    evolution 17:191, 17:194
    habits 17:191
    individuals 17:191–194
    new 17:192–193
    organizations 17:190, 17:192–194
    original 17:191–192
    transaction costs 17:190, 17:192–193
  evolutionary biology 12:183–184
  and functionalism 12:182
  grand theory 12:182
  historical 12:183–184, 12:187–189
  history of 12:181–182
  ideas 12:184
  of law 13:600
  and Marxism 12:182
  new 12:183
  organizational 12:188
  and pension policy 17:694
  and political economy 18:359
  political science 12:187
  rational choice 12:182
  rational-choice 12:187, 12:189
  sociological 12:182
  sociology 12:187–188
  types of 12:182–183
Institutionalist perspective, educational sociology 7:290–291
Institutionalization 12:244–245, 12:375, 18:289, 22:621–622, 24:815–816
  across the region 22:621–622
  of Axial civilizations 3:727
  chairs and councils 22:622
  and communications research 4:279–280
  of cultural sociology 5:501–502
  and diffusion of values and norms 1:104
  of empirical social research 7:558
  of environmental movements 7:765–767
  and ethical questions 8:53–56
  of the industry 20:17–18
  infrastructure 22:667–670
  in low-, medium-, and high-income countries 6:42
  organizational
    and normative order 1:106
  of peer review 17:686–687
    ex ante modes 17:687
    rise of academic publishing 17:686–687
    role of royal academies 17:686
  of political psychology 18:428
  of political sociology after World War II 18:473–474
  and radical social work 19:872
  role of saga in 16:181–182
  and scientific revolution 21:318–319
  of slavery 22:50
  of social history 22:307–308
  of social sciences 6:487–488
  of social work in criminal justice 5:209
  sociological 1:107
  of urban policies in Europe 24:902
Institutionalized cultural capital 5:375–376
Institutionalized form, nationhood as 16:294–295
Institutionalized individualism 1:381–382
Institutionalized oppression, and racialized groups 22:822–823
Institutionalized slaveholding 22:49
Institutionalized values 5:583
Institutionalized wars 25:420
Institutional/occupational (I/O) model, in military sociology 15:514, 15:516
Institution-building, civil service 3:700
Institutions 7:59–60, 11:561–572, 12:181, 12:186, 12:190–195, 14:651, 17:6, 17:303, 20:160–161, 21:213, 21:216–218
  adolescent sexual risk 1:133
  and agency 17:393
    institutional logics 17:393
  anthropology 1:755
  architecture of modern statehood 10:298–299
  and area studies in United States 1:951
  beliefs 12:190–195
  Boltanski on 2:751–753
  and business anthropology 3:4–9
  capacity, for randomized controlled trials 8:544
  change 11:569–570
    dynamics of 12:194–195
    processes of 12:193
  community 4:388
  concepts of 11:562, 12:190–191
  of conflict regulation 4:624
  constraints 12:190–192
  conventions 12:190–194
  cooperation embedding in 4:864
  definition 11:562
  direct democracy 25:293–294
  diversity, explaining 12:193–194
  economic transformation 7:76–77
  and economic vote 7:29
  in economics 11:563–564
  and education 7:172–176
  effects of 12:192–194
  empirical relevance 11:570–571
    qualitative analyses 11:571
    quantitative analyses 11:571
  equilibrium 12:190–191, 12:194–195
  evolution of 8:433
  experimental economics 8:546–548, 8:550
  and experts 8:589–590
  family as 8:766–771
  and family history 8:738–740, 8:743
  and fertility transition 9:49–50
  film 6:622
  financial 14:473
  functionalism 12:190, 12:192–194
  game theory 12:190–191, 12:194
  and governmentality 10:308–311
  and health 22:279
  of higher education 7:169–170
  versus ideologies 11:565–566
  and individual intentions 11:817
  inefficient 12:193–194
  inefficient institutions, persistence of 12:194
  innovation for food security 10:79
  legal 13:536–537, 13:539
  legal reasoning 13:778–779
  lesbian health, establishing 15:113–114
  media as 22:583
  methodological individualism 15:311
  microfoundations of 17:393–394
    bricolage 17:393–394
    institutional work 17:394
  and multinational corporations 16:76, 16:79
  negotiated orders 17:413–414
  and neoinstitutionalism 17:393–394
  neorepublican 20:480
  norms 12:190–194
  origin 11:566–570
    evolutionary approach 11:567
  path dependence 12:194
  political, and political culture 20:162–164
  and political repression 18:450
  in political science 11:562
  population dynamics 18:567, 18:569
  and prosocial behavior 1:581–582
  providing health statistics 5:785
    analysis and interpretation 5:785
    collection 5:785
    ensuring evidence-informed policy-making 5:785
  psychology of 15:534
  rationality 12:191–192, 12:194
  and regulations 20:161–162
  religious 3:575–576
  role in bargaining 16:416–417
    to create institutions 16:417
    designs for different goods 16:416–417
    metaregimes, regimes, and international institutions 16:416
  role in dispute transformation 6:564–565
  role in economic development 5:531–532
  role in economic growth 7:40–41
  rules 12:190–194
  schools as 21:128–129
    characteristics 21:130–131
    historical relevance 21:130
  social and behavioral sciences 15:322–323
  and social dominance orientation 22:252

Institutions (*continued*)
in social science 11:562–564
in sociology 11:562–563
South Asia 2:79
stability 11:569–570
and time use 24:371
travel of 17:392–393
theorization 17:392
translation, editing, and adaptation 17:392–393
types 12:191–192
use of term 12:186
view of identity 16:215
Institutions change 11:569–570
Instituto di corrispondenza archeologica 3:762
Instituto Latinoamericano de Planificación Económica y Social (ILPES) 3:123
Instituto Nacional de Antropología e Historia (INAH), Mexico 1:626, 5:417, 22:623
INSTRAW *see* International Research and Training Institute for the Advancement of Women (INSTRAW)
Instructed actions 21:194
Instruction and training
of metacognitive skills 24:93
principles for 24:93–94
embedded instruction 24:93
informed instruction 24:93
prolonged training 24:93
Instructional alignment 14:755–756, 24:46–47
Instructional conversations, and coconstructivism 3:915–916
Instructional delivery 24:47–48
differentiated instruction 24:48
effective instruction design 24:48
relevant instructions 24:48
Instructional design (ID) 12:196–200
applications of 12:198–199
future theory 12:199
historical overview 12:196–198
and instructional technology 12:196
research and development 12:199
Instructional leader, teacher as 24:60
Instructional leadership 4:801
Instructional planning and teacher expectations 24:46
Instructional psychology 12:201–204
adaptive competence as desired end state 12:201–202
beliefs relating to learning 12:202
cognitive apprenticeship 12:204
conceptual change 12:202
constructive learning as lever for the transition from initial to the desired end state 12:202–203
self-regulated learning 12:202–203
situated learning and collaborative learning 12:203
design experiments 12:203–204
guided discovery learning 12:203
individual differences 12:203
information and communication technologies (ICT) 12:204
neuroscience and education 12:204
powerful learning environments to promote transition to desired end state 12:203–204
self-regulation skills 12:202–203
socio-constructivist conception of learning 12:201–202
transfer 12:202
Instructional sensitivity 3:822
Instructional sequencing 24:46
Instructional systems development (ISD) 12:196
Instructional technology 12:196
Instructional theory 12:197
and learning 13:699
Instructional time reduction for arts 1:232–234
Instructional validity 4:775
Instructions 12:196, 14:752–756, 15:928, 24:247, 25:325–327
and academic motivation 15:928–935
Instructive paradox 17:13
Instrumental activities of daily living (IADLs) 1:395, 4:464
Instrumental aid with needs 12:403
Instrumental analytical techniques 10:56
Instrumental approach, of law 13:543
Instrumental conception of language 4:791, 4:793–794
Instrumental conditioning 16:640–641
reinforcement 20:208
Instrumental education for sustainable development 7:735–736
Instrumental goals 18:275, 18:274t
Instrumental power 13:94
Instrumental psychology
Vygotsky on 25:330–331
Instrumental rationality 17:8, 18:778–780
Instrumental reason, and Enlightenment 7:666
Instrumental values 25:28, 25:49
Instrumental variable assumption 8:529
Instrumental variable methods 4:446, 4:842
compliance 4:446–451
censoring 4:450–451
framework and notation 4:446–448
missing response data 4:451
PROSPECT study 4:450, 4:450t
randomization as 4:447f
relation between observed groups and latent compliance classes 4:447t
treatment effect on whole distribution of outcome 4:449–450
two-stage least squares (Wald) estimator 4:448–449
Instrumental variables (IVs) 6:262–263, 12:207, 14:853, 17:111, 21:125–126
in statistics and econometrics 12:205–209
causal relations 12:205, 12:209
common problems with standard regression analysis of observational data 12:205
conditions for valid instruments 12:206
estimation 12:205–208
estimators and their distributions 12:206–207
heterogeneity of treatment effects in microdata 12:208
history of methods 12:207
instruments are endogenous 12:207–208
linear models 12:206–207
nonlinear models 12:207
potential pitfalls 12:207–208
regression 12:205–206
source of valid instruments 12:208–209
weak instruments 12:208
Instrumental variables estimation 23:383
Instrumental-affect, contra-hedonic motivation 1:242–243
Instrumentalism 6:333
and democracy 6:86
theories of intentionality 12:335
Instrumentality 4:145
collective political action 4:145–147
Instrumentality proposition 17:8
Instrumentation 12:451
Instruments
for research ethics 20:512–513
risk assessment 20:696–697
of school safety 21:109–110
in social and behavioral sciences 15:321–322
Instrumentum regni, mafia as 14:423, 14:427
Insula, and self 21:420, 21:422
Insular cortex 23:251
Insulin, and diabetes 6:337–340
Insulin coma therapy (ICT) 23:16–18
Insulin-like growth factor 2, gene coding for 8:450–451
Insulting 18:614
Insurance 12:210–214, 13:448, 13:505
adverse selection 12:211–212
commercial property and liability insurance 12:216
contemporary arrangements 12:215
as contract 12:216
definition 12:215–216
fraud 12:212–213
gatekeeping function 12:216
as governance 12:218–219
governing through risk 12:219–220
government regulation 12:216
health *see* Health insurance
health care markets 10:598–601
job displacement 12:819
and law 12:215–220
liability 12:213, 12:216–218
moral hazard 12:212
premiums 12:211
private 12:215–216
private health insurance 10:615–616
protection 10:614–615
as regulation 12:216–218
regulation of 12:213–214
risk sharing with complete markets 12:210–211

role of remittances 12:522
size and complexity of 12:215
social value 12:210
social value of 12:210
as a tool for increasing or decreasing social stratification 12:219
with transaction costs 12:211
welfare reforms 25:516–517
Insurance contract 12:210
Insurance coverage 10:599
Insurance fraud 12:212–213
Insurer 12:210
Insurgency 12:221–227, 22:431–432, 23:633
in 1900–2006 12:222f
contemporary 12:221–222
defined 12:221–222
distinguished from rebellion, guerrilla warfare, and civil war 12:222
future directions for policy and research 12:226
grievances 12:223
onset of 12:222–223
opportunity-based approaches 12:223
outcomes of 12:224–226
tactics and territory 12:222
Insurgent citizenship 3:639
Intake, and food preferences 9:298–299
Intangible assets, diversification 6:586
Intangible compensation 17:611
Intangible costs, pharmacoeconomics 19:458
Intangible cultural heritage (ICH) 5:414, 5:417–418
Integral liberation 14:34
Integralism 23:23–25
Integrated approach to social development 12:498
Integrated attachment theory 12:228
adult attachment interview 12:231
Ainsworth's 12:229–231
disorganized attachment 12:230–231
Mary Main 12:230–231
Strange Situation Procedure 12:230
antecedents of attachment security/insecurity 12:231–232
attachment across life span 12:232
continuity and longitudinal correlates 12:232
maternal sensitivity 12:231–232
assessment beyond infancy 12:231
-based interventions 12:233
Bowlby's trilogy of 12:228–229
attachment 12:228–229
fantasies *versus* real life events 12:229
loss and grief 12:229
primary vis-a`-vis secondary drive reduction 12:228–229
separation 12:229
circle of limited security 12:230
circle of security model 12:230, 12:233
and genetics 12:232–233
goal-corrected partnership 12:231
parental state of mind 12:231
psychoanalysis 12:228–229, 12:232–233
Integrated bidimenional stage of development 16:504
Integrated care 19:495
Integrated Causal Model (ICM) 20:389
Integrated cognitive antisocial potential (ICAP) theory 14:61–62
Integrated community schools 8:612
Integrated data systems
privacy in social research 18:945–946
Integrated delivery systems 14:463
Integrated disparity 19:676
Integrated factory 2:297
Integrated geographic models 14:787–788
Integrated indicators, land use 13:241
Integrated land use and transportation models 24:619–620
Integrated likelihood 14:122–123
example 14:122–123
Integrated marketing communication, advertising as part of 1:213
Integrated Model of Behavioral Prediction 15:39
Integrated model of school effectiveness research 21:81–82, 21:82f
Integrated packages, in econometric software 7:2
Integrated product policies 7:781
Integrated psychological therapy (IPT) 22:382
Integrated Public Use Microdata Series (IPUMS) 3:286–288, 3:287f
IPUMS International 5:794–795
Integrated regional resource management 20:141
Integrated regulation 21:488
Integrated rural development programs 1:458
Integrated science courses 21:230
Integrated systems 5:707
for access to justice 12:925–926
Integrated theories 21:850
and genocide 10:19
in police use of force 18:259
Integrated water resources management (IWRM) 25:441
Integration 1:71, 1:73, 2:84–85, 4:488, 8:264–265, 12:235–236, 12:418–419, 13:260, 13:262–263, 22:991, 23:239
and adult education 1:151
AGIL 1:381
callosal connections 12:419
and cointegration 24:325
community-based health interventions 10:647
defined 19:510
of devices 2:296
in educational practice 4:496
ethnogenesis 8:176–177
of findings in mixed methods research 15:621–622
immigrant 11:641–647, 11:665–666
diversity 11:642, 11:645
exclusion through disenfranchisement 11:642
framework 11:643–644
idiosyncratic consciousness and empowerment social work 11:644–646
and neoliberalism 11:642
outcomes 11:644
policy 11:641–642, 11:644–646
practical implications 11:644
social rights 11:641–642, 11:644
theoretical overview 11:642–643
immigration 11:671–673
of institutions, Polanyi on 18:244
in last 50 years 20:541
macro-symbolic conception of 12:235–236
and migration 15:418, 15:420
modes of 12:236–237, 12:236t
and multiculturalism 12:237–240
psychotherapy integration 19:509–514
social 12:243–248
and social differentiation 6:394
in structural contingency theory 23:609–611, 23:613
and Third World 24:302
and work 8:106–107, 8:109–111
Integration policies
in Europe 8:143
and xenophobia 25:805
Integration-and-learning perspective 4:740
of organizational diversity 6:597
Integrationism, racial 19:809
Integrative approaches 21:52
Eccles' expectancy-value model 21:52
engagement 21:52
Integrative conceptual framework 23:509–511
Integrative economic ethics 3:16–17
Integrative models
of negotiation 16:418
for obsessive-compulsive disorder 17:124
Integrative motivation, in foreign language learning 9:329–330
Integrative pedagogy 24:519–520
academic-scientific materials 24:519
attitudes 24:519
and National Council of Schools and Programs of Professional Psychology model 24:519
real examples 24:519
real experiences 24:519
role models 24:519
Integrative psychotherapy *see* Psychotherapy integration
Integrative social contracts theory (ISCT) 3:17
Integrative strategy for social action 1:77
Integrative test anxiety scale 24:245
Integrative thinking 12:150
Integrity
versus despair crisis 1:435–436, 11:541
in feminist research 9:718–719
Integrity factors 2:562
Intellect 17:231–233
and adolescent personality 17:864, 17:867
Intellect/openness to experience
five-factor model (FFM) 2:560–561

Intellectual, Berlin as 2:542–543, 2:545
Intellectual abilities 12:285, 12:287–288
Intellectual and moral leadership 10:747
Intellectual antecedents 9:975–977
Intellectual attainment, and intelligence 12:317
Intellectual capitalism 12:262–263
Intellectual competence, age changes in 4:12–13, 4:13f
Intellectual context
of modern virtue ethics 25:150
United States presidency 18:861–862
Intellectual core of absorptive capacity 1:12, 1:14
Intellectual craftsmanship, Charles Wright Mills 15:537
Intellectual dependence 11:91
Intellectual development 21:128
developmental behavioral genetics and education 6:323
Intellectual development theory, Piaget's 16:501–503
critiques 16:501–503
Intellectual disability (ID)
autism spectrum disorders 2:275–278, 2:280
Intellectual efficiency 17:231
Intellectual elitism 12:280–281
Intellectual environment 5:890
Intellectual functioning
assessment of 6:344
dynamic assessment 12:252–253
neuropsychologically based assessment 12:251–252
psychometric assessment 12:249–251
Intellectual groups 2:88–89
Intellectual history 3:62, 5:421–423, 11:103–104, 12:255–261, 12:762, 12:764
Burrow on 12:258–259
Cambridge School 12:256–257
Collini on 12:258–259
flapdoodle 3:62
Foucault on 12:257–258
French Dismissal 12:257–258
German *Begriffsgeschichte* 12:257–258
historical take 12:255–256
Japanese 14:635–636
in political thought 3:62
practitioners 12:259–260
Stigler's specialization 23:458
Sussex on 12:258–259
Intellectual History Review 12:255
Intellectual influences 18:181–184
cities and development 18:182
development studies 18:181–182
planning 18:182–184
Intellectual movements 4:317, 4:320
Intellectual orientations of science 21:216–217
Intellectual property (IP) 3:31, 7:48, 11:834, 11:837–842, 12:266, 17:365, 17:584, 17:587, 19:153, 19:155, 19:158
authenticity 2:253
frameworks, reorienting 13:524–525
in genomics 10:40
law in context 12:266–272
balance 12:268
biotechnology 12:270
constitutionalization 12:269
copyright 12:268–271
future 12:271
globalization 12:268–269
overlaps 12:270–271
patent 12:266–267, 12:270–271
propertization 12:267–268
software and internet 12:269–270
technology 12:268
technosocial 12:266
territoriality 12:268–269
theoretical approaches 12:266–267
trademark 12:266
rights, Ethical aspects 12:262–265
rules, domestic welfare in design of 13:521–522
Intellectual property (IP) regime 13:520
Intellectual property rights (IPRs) 12:154, 12:262, 20:499
in agriculture 1:472
economic importance 12:262–263
ethical aspects in 12:264–265
and Gnaritas Nullius 11:772–773
historic roots 12:263
indigenous knowledge and 11:769–778
international legal developments 12:263–264
notion of 12:262
rationale 12:263
Intellectual property (IP) system 13:521
Intellectual shock tactics 14:455
Intellectual society 13:106
Intellectual strategies of de-Westernization programs 25:547–549
Intellectual styles 4:110–111, 4:114
Intellectual technology 24:846–847
Intellectualism, and religion 20:309
Intellectuals 12:273–282, 13:114, 19:96, 24:295, 24:297
and anthropology 13:72
Bendix on 2:512–514
creativity 12:278–280
definition 12:277
economic opportunities 12:279–280
English intellectuals 12:274
Erasmus 5:664
French case 12:273–274
German case 12:274–275
Hirschman on 10:892, 10:896–897
ideas 12:278–280
intellectual elitism 12:280–281
intelligentsia 12:277, 12:280
MENA 16:361–366
origins 12:277
political opportunities 12:280
professions 12:277–279
Shils on 21:884–888
social origins 12:279
in Southern Europe 12:275
in Soviet Union 20:820
theories 12:277–279
class theory 12:278
classlessness 12:278–279
shifting social networks 12:279
twentieth-century 12:275–276
Intellectuels de gauche 12:275
Intellektuelle 12:274
Intelligence 3:246, 11:392–393, 12:283, 12:290–296, 12:309–323, 13:392, 15:776, 17:232, 17:787, 17:790, 21:120, 22:485
abstraction and inference 12:316
adult age differences in 1:428–430
and aggression 1:378
aging 12:294, 12:325
assessment measures 12:286–287
Cognitive Assessment System–2nd edition 12:287
Differential Ability Scale–2nd edition (DAS-II) 12:287
Kaufman Assessment Battery for Children–2nd edition 12:287
Stanford–Binet Intelligence Scales–5th edition (SB5) 12:287
Wechsler tests 12:286
Woodcock–Johnson test of cognitive abilities 12:286–287
assessment of 12:286–287
Cognitive Assessment System–2nd edition 12:287
Differential Ability Scale–2nd edition (DAS-II) 12:287
Kaufman Assessment Battery for Children–2nd edition 12:287
Stanford–Binet Intelligence Scales–5th edition (SB5) 12:287
Wechsler tests 12:286
Woodcock–Johnson test of cognitive abilities 12:286–287
attention and executive control 12:316–317
Bell Curve 12:311
Binet on 12:309–311
bioecological and lifespan theories of intellectual development 12:293–294
bioecological theory 12:294
dual-process theory of lifespan cognitive development 12:294
brain 12:290, 12:294
and breast feeding 2:847, 2:849, 2:851
challenges to theoretical status of ability factors 12:294–295
changes by education 12:320–321, 12:320f
cognitive development 12:319, 12:327, 18:120–121
comparability of human and computer 4:103–104
versus competencies 22:372
conceptual knowledge 12:326
core knowledge 12:325–326
core processes 12:315, 12:315t
creativity 12:314, 12:317, 12:320–321
and creativity 17:770–772
critical thinking 12:320

culture, gender, social class, and 12:318–320
culture-free/culture-fair tests 12:323–324
description 12:283
domains 12:315–316, 12:315t
education and learning 7:138–139
educational science 12:325–326
elusive concept of 12:290
entity *versus* incremental theories about 19:791–792
as erotic capital 8:4–6
ethnographic 8:182
eugenics 12:310–311
expertise 12:319, 12:327–328
factor analysis 12:291–292, 12:294
factor g, working memory, and specific abilities 12:324–325
Flynn effect 12:311
general and specific aspects about 12:317
genetic influence on 12:299
genetics 12:291, 12:294–295
genetics of 12:297–302
and gifted education 7:158, 7:160, 7:162–163
giftedness 12:314
and grit 10:397–399
history 12:283–284
human *see* Human intelligence
human differences 12:309
impact of social class 12:319–320
information processing 12:290–295
integrating psychometric intelligence with cognitive processes 12:292–293
mental speed theory 12:292–293
PPIK theory 12:293
triarchic theory 12:293
working memory theory 12:293
intelligence quotient 12:309–311
knowledge and beliefs 12:315, 12:315t
knowledge construction 12:326–327
lasting trends and an attempt of integration 12:290–291
leadership as 13:649
and learning 12:327–328
in lower socioeconomic status families 12:325
measuring 12:323–324
mental ability 12:309–311
mental operations 12:315–316, 12:315t
merit 12:309, 12:311
models 12:284–286
Cattell-Horn-Carroll (CHC) theory 12:285
PASS theory 12:285
in review 12:285–286
Sternberg's triarchic theory 12:285
Wechsler factor structure 12:284
motivation 12:292–295
nature of 12:311
nature-nurture 12:311
and prefrontal cortex 18:814
prior knowledge 12:327
probability 19:17–18
problem solving 12:313–314, 12:317, 12:319–321
products, measurement of 12:291–292
Gf-Gc theory 12:291–292
primary mental ability theory 12:291
structure of intellect model 12:292
three-stratum theory 12:292
two-factor theory 12:291
race 12:311
reasons for individual differences 12:318
representational capacity 12:316
as the result of gene–environment interactions 12:325–326
schooling 12:323–324
single shared *see* Single shared assessments (SSAs)
social and professional issues (in cognitive assessment) 12:287–288
species and groups 12:309–310
speed tests 12:323
stable and changing elements in 12:317–318
architecture and intellectual attainment 12:317
development 12:317–318
individual differences 12:317
Stanford–Binet scale 12:310–311
study of 2:598
talents 12:309
Terman on 12:310–311
test usefulness 12:287
tests 12:310–311
tests of 12:316
verbal and nonverbal intelligence tests 12:324
visual patterns, recognizing 12:319
WAIS 12:310–311
WISC 12:310–311
working memory 12:316, 12:324
Intelligence agencies, and area studies 1:949–952
Intelligence assessment 16:723–724
Vygotsky on 25:312
Intelligence construct requirements 22:372–373
Intelligence quotient (IQ) 12:309–311, 19:830
development, of Israeli children 21:126
and developmental psychology 5:358
and sibling order 21:924
scores 7:163
distribution of 12:308
tests 2:598, 8:130–131, 12:323, 17:853, 19:418–420
Intelligence tests 24:763–764
Intelligence-led policing 18:879
Intelligent agents 1:536
Intelligent behavior 4:23
Intelligent city 22:114
Intelligent Community Forum 22:116
Intelligent computer assisted instruction (ICAI) 4:494
Intelligent design (ID) 1:370, 20:288
Intelligent machines 18:666
Intelligent systems 19:56–57
Intelligent transportation systems (ITS) 24:633
Intelligent tutoring systems (ITSs) 4:494, 6:899, 21:187
Intelligent Vehicle Highway Systems 24:625
Intelligentsia 12:273, 12:277, 12:280
Intelligenz 12:273
Intended family size 8:805
changes in intentions 8:806
Intensification 12:329–331
agriculture 12:329–331
archaeology of 12:329–332
Australia and New Guinea 2:230–232
Boserup on 12:330–331
concept 12:329–330
consequences of 12:330
craft production 12:329–331
exchange 12:331–332
history of approaches 12:330
industry 12:329–330
landesque capital 12:329–330
prestige goods 12:331
social complexity 12:329, 12:331
standardization 12:331–332
studies
future of 12:330–331
methodological issues in 12:330
and time use 24:373, 24:375, 24:377–378
urbanization 12:329
workshops 12:329, 12:331
Intension of classification 3:806, 3:808
Intensionality 15:752
Intensity function 8:296–300
Intensity indicators, land use 13:239–241, 13:240f
at ecosystem level 13:240–241, 13:240f
input intensity 13:240
integrated indicators 13:241
output intensity 13:239–240
Intensive care, evolutionary medicine 8:457
Intensive case management (ICM), in psychiatric disorders 22:383–384
Intensive interpersonal relationships 9:428
Intensive interview 12:671–672
Intensive Language Program of the American Council of Learned Societies 2:705
Intensive magnitudes, measurement theory 14:868–870
Intensive oncological treatment
and childhood cancer 3:461
psychological counseling during 3:461
Intensive short-term dynamic psychotherapy (ISTDP) 19:345, 19:347
Intensive supervision, offender rehabilitation 20:196
Intensive supervision programs (ISPs) 4:373, 4:375–376
Intent to treat effect
in population studies 8:561
Intention reading 4:300
Intention superiority effects 19:271–272
Intention understanding, Theory of Mind development 24:269
Intentional action 1:671
Intentional change theory (ICT) 13:660
Intentional communities 2:144
communal societies 2:148–150

Intentional communities (*continued*)
and communes 4:268–269
Jewish intentional communities 2:148
organizing principles of 2:148
Intentional effects 8:572
Intentional encoding of context 4:782–783
Intentional families 9:456
Intentional fundamentalism 14:418, 22:418
Intentional psychology 11:814
Intentional self-regulation (ISR) 21:548–549
in adolescence 21:549
features 21:549–550
Intentionalism 12:932
Intentionality 11:441–442, 12:333–340, 15:539, 15:542–543, 18:33, 21:342–343
collective 21:343–344
and consciousness 12:334
family networks 21:755–756
forms of 21:343
in infants 12:340–341
as goal directedness 12:342
innate intentionality *see* Innateness
of knowledge 14:391
as organism–world relationship 12:343–344
phenomenology 18:11–13
reductive approaches to 12:334–335
social intentionality *see* Social intentionality
unintended consequences 17:420–422
world/intermediate objects 12:333
Intentionality in language and communication 12:347–351
cognitive development 12:347–348
developmental pathway 12:348–349
from eye contact to shared goals 12:348
motor control in vocal and gestural productions 12:348–349
intention reading 12:347
intentional agent 12:347–348, 12:350–351
joint attention 12:347–349
mental states 12:347–349
mind sharing 12:348
role of adult and input 12:349–351
adults' interpretations 12:350–351
affect and social interaction 12:349–350
cultural variations 12:351
scaffolding 12:347, 12:349, 12:351
theory of mind 12:347–348
definitions 12:347
foundational feature of human communication and language 12:347–348
mirror neurons 12:348
Intention-behavior gap 10:674–675
health psychology 10:674–675
Intentions 10:710
versus effects 22:196–197
fertility 9:50
goal-oriented behavior 10:268
health behavior 10:577–580
and health behaviors 10:585–586
Heider on 10:752–753
and imitation by infants/young children 11:622–624
infant gaze behaviors 12:27, 12:29–30
joint action in meshing 12:840
like me theory 15:131, 15:133–134
personal we-intentions 12:840–841
and practical reasoning 18:777–780
prospective memory 19:270–278
psychological concept 12:340
Intention-to-treat (ITT) analysis 8:346–347
Intention-to-treat (ITT) effect 4:445–446, 8:528
Intent-oriented private-sphere environmentalism 7:700–701
Interacteomics 2:625
Interacting cognitive subsystems (ICS) 11:402–403
Interaction engine 22:857
Interaction forms, social structure as 22:716
Interaction in organizational groups 10:411
Interaction order 8:206
in narratives 16:221–222
Interaction patterns, during competition and cooperation 4:857–858
Interaction process analysis 22:102–103, 22:103f
Interaction processes in organizations 21:135
Interaction ritual theory 22:335
Interaction rituals, Collins on 4:211, 4:213–216
Interaction theory
Luhmann on 14:386–387
phenomenology 18:19–21
Interactional artifacts 8:572–573
Interactional continuity in marriage 17:799
Interactional epistemology, social representations as 20:444–446
asymmetries and symmetries 20:445
Ego–Alter–Object 20:444
ethics 20:446
field 20:444–445
figurative schemes 20:446
heterogeneity and cognitive polyphasia 20:445–446
Interactional expertise 21:210
Interactional justice 12:930, 12:951, 17:381
Interactional sociolinguistics 18:798
Interactionism 25:678
and discourse studies 6:513–514
field theory 9:140
mental causation and 15:175
symbolic 12:352–356
Interactionist dualism 15:175
Interactionist position, in foreign language learning 9:327–331
Interactionist theorizing 22:958–959
dramaturgical theories 22:958–959
phenomenological theories 22:959
role theories 22:959
status theories 22:959
symbolic interactionist theories 22:958
Interactionist theory 6:72, 13:150–151
of educational sociology 7:290
Interactionist view of emotion 7:535
Interactions 14:322–323
culture as 5:619–620, 9:50–51
in dexis 6:54
and discourse 6:498–501
and emotions 7:473, 7:475–476
human 22:267
and intersectionality 12:657–662
language in 22:950–951
law as a facilitative arena for 17:398–400
methodological individualism 15:309
and moderation 15:672–674
organizational behavior 17:697–702
Schütz on 21:146–148
with selection 12:453
at sentence level 21:352–353
in social movement theories 22:451–452
at word level 21:350–352
Interactive activity 21:271
Interactive belief systems 10:436
Interactive case studies 3:180–181
Interactive conflict resolution 4:586–587
Interactive discussion-based learning 3:176
Interactive displays 10:352–354
dynamic graphics 10:354
features 10:352–353
interfaces 10:353
linking 10:353
Interactive documentary 9:164
Interactive ethnonationalism 8:92
Interactive governance 13:760
Interactive learning 7:23
Interactive processes in brain 4:97–98
approaches to cognition 4:97
Interactive revolution 21:814
Interactive social flow 9:264
Interactive text 14:241
Interactive voice response (IVR) 10:590
Interactive–developmental model of psychiatric disorders 22:381
Interactivity 16:146
media and genetics 9:992
see also Cybersex
Interagency Council on Statistical Policy 10:304
Inter-Agency Standing Committee (IASC) 19:480–481, 24:753–754
Inter-American Commission on Human Rights (IACHR) 11:801
Inter-American Court of Human Rights (Costa Rica) 11:799–800, 12:510
Inter-American Development Bank (IADB) 7:180
Intercalated (ITC) nuclei 8:870
Intercategorical complexity 12:661–662, 12:664–665
Intercensal cohort analysis, data adjustments 6:128–129
Interconnected world, citizenship in 11:651–652
Interconnectedness, and human trafficking 10:207
Interconnections 11:828
Interconnectivity, in cerebral cortex 3:334–338
Intercountry adoption (ICA) 1:142, 1:145

Intercourse *see* Sexual debut
Intercultural aspects of bioethics 2:614
Intercultural communication (IC) 1:776, 14:726
  anthropology of 12:357–361
    applications 12:359
    critique 12:360
    cultural competence 12:358–359
    culture 12:357–360, 12:358t
    ethnicity 12:357, 12:359–360
    Hall on 12:357, 12:359–360
    Hofstede's contribution 12:358–361
    key concepts 12:358–359
    origins 12:357–358
    recent developments 12:360–361
    rise and global spread 12:359–360
  and training 5:315
Intercultural comparison of cities 24:802
Intercultural competence (IC) 12:358–359
  foreign language learning 9:331
Intercultural conflict dimension, honor killing 11:186–187
Intercultural criminal law 12:362–367
Intercultural Development Inventory 12:358–359
Intercultural encounters, legal dilemma of 12:363
  conditions of agency 12:365
  presupposition of an interest 12:365
  in Roman Empire 12:364
Intercultural group processes, in organizations 10:412
Intercultural legal discourse 12:362–367
  basic idea 12:362
  levels 12:362
Intercultural relations 5:314–315
Intercultural situational judgment test (iSJT) 5:435
Intercultural training, cross-cultural training 5:334, 5:336, 5:339–340
Interculturalism 12:237, 16:31–32
Interdepartmental Emergency Taskforce (IDETF) 7:429
Interdependence 11:828, 22:220, 23:368–369
  autonomy and 1:182
  in collectivism/individualism 4:207
  in conflict/consensus 4:590
  language policy for 13:357
  in structural contingency theory 23:611
Interdependence family model
  and parenting 17:494–496
Interdependence hypothesis 2:583
Interdependent cultural model, and early memory development 6:760
Interdependent cultures, and play 18:220–223
Interdependent decisions, and collective action 22:513
Interdependent development pathway, of identity 21:436–438
Interdependent dyadic relationships 22:461
Interdependent global market 8:60
Interdependent marriage 8:782
Interdependent processes
  in event history analysis 8:307–308
    causal approach 8:309
  life course 14:53–54
    careers of spouses 14:53
    at intermediate level 14:53–54
    spouse selection (educational homogamy) 14:53
    trajectories at level of individual 14:53
Interdependent selves 5:426
  motivational tendencies 5:426
Interdependent systems, in variables 7:613–614
Interdisciplinarity 2:350, 12:257–259, 17:650, 24:813–814
  of area studies 1:949, 1:951, 1:953
  Castells on 3:229
  cybernetics 5:645
  of educational research 7:280
  and green social work 10:388
  hierarchies and markets 10:842
  legal reasoning 13:780, 13:782
  political psychology 18:427
  questions, and epistemic cultures 7:877–878
  of urban studies 24:941–943
Interdisciplinary approach
  to economics 23:228
  networks 23:226
  or recombination of specialties 23:226–228
  to sociology 23:228
Interdisciplinary collaborations 7:870
Interdisciplinary connections, in Mexico 1:627
Interdisciplinary genre 2:647–648
Interdisciplinary legal scholarship 13:515, 13:519
Interdisciplinary narrowness 3:70
Interdisciplinary research, Whorf on 25:578
Interdisciplinary scholarship, ethnohistory 8:193, 8:196
Interdisciplinary theory 5:233–235
Interest, psychology of 12:378, 12:380–382
  characteristics of 12:379
    affective and cognitive components 12:379
    content/object specific 12:379
    physiological/neurological basis 12:379
    relationship between learners and their environment 12:379
    unreflective 12:379
  impact 12:383–384
  measurement 12:382–383
  retrospective case study 12:379–382
    analysis 12:380–382
    overview 12:379–380
  working definitions 12:378–379
Interest (monetary)
  Fisher's theory 9:219
  Keynes's theory 9:218–219
Interest group liberalism 20:182
Interest groups 12:368–377, 18:398
  and administration 20:162
  definitions 12:368
  history 12:371–377
    constellations and trajectories 12:373
    functions for democracy 12:376
    patterns of organization 12:373–374
    secular changes 12:375–376
    systems of interest intermediation 12:374–375
    types 12:372–373
  and parties 12:369
  and pluralism 18:233–234
  and policy process 12:368–369
  politics 12:369–370
  regulatory agencies 20:182
  representation 12:369
  and welfare retrenchment 25:525
Interest on Lawyers' Trust Accounts (IOLTA) 18:737
Interest rate
  macroeconomics 14:406, 14:409
  monetary policy, forward guidance on 15:728–729
Interest rate spread 9:317
Interest theories 1:38
  for academic achievement 1:22
    developmental changes 1:22
Interest value, in expectancy-value-cost model 8:504, 8:506
Interest-based model of negotiation 16:418
Interest-driven engagement of digital natives 22:920
Interest-free loans 3:503
  from China 3:502–503
Interest-group politics 11:670
Interests 1:31, 15:921, 19:181, 21:51
  ambiguities in 12:389
  class, and functional explanation 6:242–243
  derivation of word 12:386
  element of negotiations 16:431
  gender-typed 9:685
  history of concept 12:386–390
  ownership interest 19:179–181
  patterns, autism spectrum disorders 2:276
  politics of 12:386–387
  psychology of 12:378–385
  public *see* Public interest
  public policy ethics 19:581–582
  rational choice explanation 19:899
  sociological analysis of 12:391–395
    altruism and selfishness 12:393–394
    axiological rationality 12:394–395
    economic attitudes 12:392–393
    game theory 12:391–392
    human behavior 12:392–394
    moral sentiment 12:393–394
    property rights 12:391–392
    rational choice 12:394
    rationality 12:394
    scope 12:393
    social action 12:391
    social norms 12:391–394
    social order problem 12:391–392
    utility 12:391, 12:393
    values 12:391–394
  utilitarian social thought 24:980–982, 24:984–985

Interests (*continued*)
vocational *see* Vocational interests
Weber's ideal 12:389
Inter-Ethnic Adoption Provisions (1996) 1:146
Interethnic communication 13:257
Interethnic relations, in postsocialist societies 18:699
Interethnic wars 3:714
Interfaces 11:549–550
econometric software 7:4
interactive displays 10:353
resilience 20:557–558
Interfaith organizations 20:407
Interference
language contact 13:301–302, 13:305
memory 24:200–201, 24:203–204
multiple language 16:60
retroactive interference 20:613–620
Interference paradigms, for context dependent memory 4:779
Interference theory 19:1
Intergenerational ambivalence 12:403, 12:406
Intergenerational class mobility 15:642–644
Intergenerational continuity, of attachment 2:139
Intergenerational dilemmas 22:246
Intergenerational hegemony, and inequality 11:897–898
Intergenerational justice 12:396–401
benevolence 12:396
communitarianism 12:399–400
contractarianism 12:396–400
contractualism and asymmetry 12:396–397
environmentalism 12:397–398, 12:400–401
family 12:397, 12:399–400
nonidentity problem 12:397–398
population 12:396, 12:398–399
prospects for 12:399–401
transgenerational community 12:399–400
universalism defended 12:400–401
sufficientarianism 12:401
sustainability 12:396, 12:401
utilitarianism and repugnant conclusions 12:398–399
Intergenerational neglect, children 3:414
Intergenerational property transmission 19:173
free testation and capitalism and 19:175
future of 19:176
and taxation policy implications 19:175–176
trusts and 19:174–175
wills and succession
among nations in twentieth century 19:175
throughout history and 19:173–174
Intergenerational relationships
economics of 8:826–828
altruism 8:826–827
family constitution 8:828
nonaltruism 8:827–828
grandparent–grandchild 10:337–339
Intergenerational school 1:151
Intergenerational sex 21:798
Intergenerational social mobility 22:934
Intergenerational solidarity 12:402–408
grandparent–grandchild relationships 10:337–338
Intergenerational ties, cultural construction of 9:854–855
Intergenerational transfers
and evolutionary demography 8:410
and life history evolution 9:30
wealth flows 9:35–39
Intergenerational transmission
of behavior, fertility 9:21
of divorce 14:520
panel surveys for 17:465
of values, fertility 9:21
Intergovernmental lobbying 12:373
Intergovernmental organizations (IGOs)
campaigns for climate science 3:841
and world culture 10:226, 10:228–229
Intergovernmental Panel on Climate Change (IPCC) 1:226, 1:294, 1:297–298, 3:860, 3:865, 7:791, 7:817, 7:862, 18:104–105, 20:422, 21:239
Second Assessment Report 7:794
Intergovernmental relations 8:875
Intergroup aggression and warfare 8:445
Intergroup behavior, social dominance orientation 22:252
Intergroup bias, and empathy 19:218–219
Intergroup competition, social dilemmas 22:246
Intergroup conflicts 4:582, 11:676–678
from majority perspective 11:676–677
from minority perspective 11:677–678
and racism 19:866–867
Intergroup contact 11:679
and community 4:387
improving intergroup relations with 11:678–679
Intergroup contact hypothesis 4:587–588
Intergroup coping strategy 16:862
Intergroup dynamics 4:153–154
Intergroup helping 10:759
cross-racial helping 10:759–760
relations between unequal groups 10:762–764
defensive helping 10:763–764
low-status group 10:763
self-categorization 10:761–762
social identity perspective on 10:761
Intergroup helping as status relations model (IHSR) 10:762
autonomy/dependency orientation 10:762–763
autonomy-oriented assistance 10:762–763
dependency oriented help 10:762–763
structural characteristics 10:762
Intergroup ideologies, social dominance orientation 22:252
Intergroup interaction 11:530
Intergroup level, of stigma analysis 23:464–465
Intergroup processes
and tyranny 24:711–713
Intergroup processses
of war/political violence 25:392
Intergroup relations 4:154–155, 8:445, 10:761–762, 12:409–415
cognitive perspective 12:409
and collective memory 4:194, 4:197–198
and emotions 7:502–506
ethnic identity 8:113–116
and ethnocentrism/prejudice 8:168–172
fundamental collective beliefs 12:411–414
genocide 10:16, 10:18–20
helping 10:759–765
and helping behavior 19:218–219
power relations 19:218–219
historical and comprehensive analysis of 12:409
ideologies and 12:411–414
implicit theories about nature of human beings 12:413–414
infra-humanization theory 12:413
stereotype content model 12:413–414
justice motive 12:412–413
belief in just world and intergroup relations 12:413
perceived justice 12:412
relative group deprivation 12:412
and legitimacy 12:412, 13:827
social dominance theory (SDT) 12:412
social identity theory (SIT) 12:412
system justification theory 12:412
political psychology 18:428, 18:431–432
social categorization and 12:410–411
decategorization 12:411
dual identity 12:411
mutual intergroup differentiation 12:411
recategorization 12:411
and social psychology 22:533–534
social psychology of 12:409
stereotypes 23:448–450
Intergroup threat
and immigration 11:676–679
majority 8:143–144
Intergroup violence, and dehumanization 6:33–34
Intergroups 10:426–427
decision making 10:426
INTERHEART study 9:704–707
Interhemispheric communication, Sperry's contribution 23:280
Interhemispheric cooperation 12:418–419
Interhemispheric excitation 12:417
Interhemispheric inhibition 12:417
Interhemispheric interaction and language 12:419–420
Interhemispheric interaction in language and cognitive processes 12:416–424
Interictal behavioral dysfunction 13:35
Interictal period, seizures 7:854–855
Interindividual level of analysis 13:899
Inter-Institutional Agreement 11:841
Interiorized 2:493

Interjections 13:290
Interjurisdictional discovery 6:521
Interlacing of activities, mobile communication 15:630
Interlacustrine kingdoms 6:774
Interlanguage (IL) 9:328–329, 21:361–363
Interlegality 13:539
Interleukin-6 (IL-6), and emotions 7:499
Interlibrary loans 14:47
Interlingual homographs 21:350
Interlingual homophones 21:350
Interlocking board 16:549
Interlocking directorates 12:623
Interlocks 14:473
Interlocutory appeal 1:823–824
Intermarriage 14:600
  ethnic 8:118–122
  Jewish 12:809–810, 12:808t
Intermediaries, and brokers 2:870, 2:873–874
Intermediate care, in England 4:345, 4:347
Intermediate endpoint research (mediating variable) 15:64
Intermediate institutions, Nisbet on 16:835
Intermediate objects 12:333
Intermediate order statistics 17:293
Intermediate progenitor cells (IPCs) 2:821–823
Intermediate variables 15:64
Intermediate-Range Nuclear Forces (INF) Treaty 1:977
Intermethod mixing 15:621
Interministerial Mission for Monitoring and Combatting Cultic Deviances (MIVILUDES) 16:808
Intermittent feedback control, motor control 3:782–783
Intermittent offender 5:139–140
Intermodal comparisons 21:605
Intermodal Surface Transportation Efficiency Act 1991 24:619
Internal and external critical positions 3:612–613
Internal and external validity 3:69
Internal articulation, of emergent system 7:173
Internal assets 3:446
Internal carotid artery 12:690–691, 12:691f
Internal coherence, definition 21:128–129
Internal colonialism 15:571, 19:847
Internal colonization 12:245
Internal combustion engine (ICE) 2:304, 17:169
Internal consistency 6:345
Internal consistency reliability (ICR) 20:242
Internal constituency surveillance 23:737
Internal control 4:834–835
Internal criticism 21:451
Internal democracy, in associations 25:256
Internal ecology model 23:513
Internal economies 14:622–623
Internal exile 21:894
Internal frames, academic self-concept and achievement 1:58
Internal frontier 19:156
Internal labor markets 17:311
  promotions 19:135
Internal locus of control, workplace autonomy 25:719
Internal market, principles of 8:266–267
Internal migration 15:389, 15:392
  in developing countries 12:433–442
    causes 12:436–439
    countries collecting internal migration data 12:434t
    displacement 12:433, 12:437–438
    environmental factors 12:438
    global estimates of number of internal migrants 12:435t
    human capital 12:436–438, 12:440
    impacts 12:439–440
    left-behind people 12:439–440
    levels and trends 12:434–436
    livelihoods 12:433, 12:436–440
    measurement and definitions 12:433–436
    remittances 12:433, 12:439–441
    rural–urban migration 12:436
    social capital 12:438–439
    state and other political actors 12:437–438
    urbanization rates 12:434–435, 12:434t
  and economics 15:435–436
    age 15:435
    distance 15:435
    education 15:435
    family migration 15:436
    other variables 15:435–436
    return and repeat migration 15:436
  in industrialized countries 12:443–446
    consequences of migration 12:445
    disequilibrium 12:444
    equilibrium 12:444
    family characteristics 12:443
    human capital model 12:445
    life cycle impacts on migration 12:445
    migration patterns 12:444
    personal characteristics 12:444–445
    place characteristics 12:443
    policies 12:445–446
    pressing questions 12:445–446
    rates of migration 12:443
    theoretical perspectives on migration 12:444–445
    utility maximization 12:444
Internal models, motor control 3:778, 3:781–783, 15:957–958
Internal motivation 15:907–908
  for behavior maintenance 15:908
  clinical intervention 15:907–908
Internal political efficacy 7:301–306
Internal rates of return (IRRs) 4:930, 20:16
Internal realism 18:806
  Putnam on 19:665–666, 20:22–23
Internal reasons, identity 11:524–525
Internal reconstruction (linguistics) 12:447–449
  historical morphophonemics 12:447–448
  regularization of systems 12:448
  universal and typological reconstruction 12:448–449
Internal representations 13:76, 13:79
  of numbers 14:775–776
Internal Revenue Service 19:556
Internal selection environment, role of 23:510
Internal self-determination 11:797
Internal source monitoring 23:28
Internal stability 4:869
Internal validity 12:450–454, 14:354, 19:739, 19:742, 19:744, 20:154–155
  evidence-based medicine 8:329–330
  experimental estimates 22:297
  of placebo-controlled double-blind studies 18:162
  threats, nonequivalent group designs 16:851
  threats to 12:450–453
    attrition 12:452
    carryover effects 12:452
    coping with threats 12:453–454
    crossover effects 12:453
    effect of a treatment 12:450
    'history' effects 12:451
    interactions with selection 12:453
    maturation 12:451
    measuring instrument 12:451
    regression toward mean 12:452
    'seasonality' effects 12:451–452
    selection differences 12:452
Internal war 3:714, 4:615
Internal working model (IWM) 2:137, 12:231
Internal/external frame of reference (I/E) model 6:430–435, 12:425–432, 21:464–465, 21:470, 21:465f
  academic achievement 12:425–426, 12:429
  academic self-concept 1:58
    and achievement 1:57–59
  assimilation and contrast effects 12:428–429
  control-value theory 12:431
  cross-cultural and developmental generalizability 12:428
  dimensional comparison research 12:425–427, 12:429, 12:431
  empirical support 12:427, 12:427f
  expectancy-value theory 12:430–431
  integration with other ASC models 12:429–430
    big-fish-little-pond effect 12:429–430, 12:430f
    reciprocal effects model 12:429–430, 12:430f
  longitudinal analysis 12:427, 12:429–431
  Marsh/Shavelson revision of 21:465, 12:426f
  meta-analysis 12:427–428
  multidimensionality 12:425
  policy and practice implications 12:431
  social comparison 12:425–426, 12:429–430
  structural equation modeling 12:428

Internal/external frame of reference (I/E) model (*continued*)
theoretical basis 12:425–427, 12:426f
theoretical extensions 12:430–431
Internalism 2:134, 13:782, 18:778
Internality, and self-efficacy 21:505
Internalization 4:459, 5:578, 25:324, 25:326–327
and business 12:484–486
of motivation 21:487–488
problem of 5:578–579
Internalization theory 12:268
Internalized conversation of gestures 14:834
Internalized homophobia 10:791
Internalized moral perspective, and authentic leadership 2:247
Internalized oppression, and racialized groups 22:823–824
Internalized-oriented moral reasoning 7:315–319, 7:316t
Internalizing behavior, and SES 22:944–947
Internalizing behavior problems 2:477–478
Internalizing disorders 17:916–917
Internalizing problems 2:460
and divorce 14:519–520, 14:522–523
Internally displaced people (IDP) 1:970, 12:437, 20:105, 20:107, 20:112–113, 25:431
and armed conflict 1:970
Internally driven population cycles 18:559–561
Internally generated code-switching 3:918
International adoption 1:136–137
International Adult Literacy Survey 5:847, 7:135, 7:135t
International advertising 12:455–461
as business practice 12:458–459
as communication process 12:457–458
consumer values 12:455
context 12:455–456
cross-cultural psychology 12:457
cultural context 12:457–458
definition 12:455
as industry 12:459–460
research for 12:456–457
as social force 12:460
social influence 12:460
International affairs journalism 14:715–716
International Affective Picture System 16:659
International agencies, position of 10:153
International aid agencies 1:270
International aid movement, and citizenship 3:658
International and regional human rights tribunals 11:801
International arbitration 12:470–473
advantages of 12:471
alternative dispute resolution and 12:471–472
application 12:470
arbitration centers 12:470
building an 12:471
commercial arbitration 12:470–471
features 12:471
investment arbitrations 12:470
sociolegal issues related to 12:471
in United States 12:472
use of term 12:470
International aspects of bioethics 2:614
International Association for Cross-Cultural Psychology 5:311
International Association for Impact Assessment (IAIA) 22:873
International Association for Semiotic Studies 21:594
International Association for Social Science Information Service & Technology (IASSIST) 1:917
International Association for the Evaluation of Educational Achievement (IEA) 4:427, 4:429, 5:327–329, 5:773, 7:919, 10:248, 21:119, 21:121
key findings of 4:432–433
relating to comparative education 4:433
relating to policy making 4:433
Language Education Study 9:330–331
phases of 4:433–434
research program of 4:432
International Association for the Study of Pain (IASP) 17:451
International Association of Academies 22:674
International Association of Bioethics 2:619
International Association of Genocide Scholars 10:7
International Association of Legal Ethics (IALE) 13:621
International Association of Legal Science (1950) 2:90
International association of radical reflexivity (IARR) 20:95
International Association of Schools of Social Work (IASSW) 1:785, 1:793–794, 10:145, 10:265, 10:530, 10:718, 11:353–354, 11:357–358, 12:474–479, 12:547, 12:567, 14:36, 19:478, 22:776, 24:750
current membership 12:475, 12:475f
global standards 12:474, 12:478, 22:521
history 12:476–479
crises and restoration 12:477
formative phase 12:476–477
internationalization of social work and education 12:479
years of expansion 12:477–478
language policy and publications 12:476
leadership and governing bodies 12:475–476
mission 12:474–475
presidents 12:475t
social work profession 12:475–476, 12:478
International associations, of social sciences 6:489
International Baccalaureate (IB) 10:249
IB Career-related Certificate Program 10:249
IB Diploma Programme (DP) 10:249
IB Middle Years Programme 10:249
IB Primary Years Programme 10:249
International Bar Association (IBA) 13:619–621, 13:623
International bioethics 2:628–629
International Biological Program (IBP) 7:792–793, 11:306–307, 11:309
International Biometric Society, establishment 9:222
International broadcasting, by radio 19:879
International Bureau of Education (IBE) 4:429–430
International business 12:480–483
geographic aspects 12:484–489
emerging market multinational enterprises 12:487
explanations of foreign direct investment 12:484–486
foreign direct investment 12:486–487
global production networks 12:485, 12:488
impact of foreign direct investment and global value chains 12:487–489
internalization 12:484–486
locational determinants of global value chains 12:486–487
multinational corporations 12:484
ownership 12:484–485, 12:487
transnational corporations 12:484
and multinational enterprise (MNE) 12:480t
International business law 12:507
International Business Machines Corporation (IBM) 5:539, 6:834–835, 17:615
microcomupters 23:378
International Campaign to Ban Landmines 22:514
International Cancer Genome Consortium 11:318
International capitalism, *versus* global capitalism 3:99
International census 3:290–295
International census microdata 3:286–287
International Center for Learning, Education, and Performance Systems (ICLEPS) 12:199
International Center for Research on Women (ICRW) 4:558
International Center for the Settlement of Investment Disputes (ICSID) 12:470, 15:72
International Centre for Prison Studies 11:721–722
International Chamber of Commerce (ICC) 12:470, 13:915–916, 15:72
International Court of Arbitration 13:916
International child protection systems 3:418, 3:421
International Christian University 12:795–796
International classification of activities for time-use statistics (ICATUS) 5:705
International Classification of Diseases (ICD) 6:78, 6:342
International Classification of Diseases (ICD-8) 3:886–887, 17:30

on homosexuality 3:886–887
International Classification of Diseases (ICD-9) 17:30
International Classification of Diseases (ICD-10) 5:623, 6:200, 17:31–32, 22:79, 24:549
acute episode 6:200t
criteria for diabetes 6:338
differential diagnosis 6:386–388
homosexuality 3:886–889
mania 14:484–485
schizophrenia 21:9, 21:11–12
somatoform disorders 23:10–11
International Classification of Diseases (ICD-11)
personality disorder 17:894
current proposal 17:894–895
psychosomatic disorders 19:486–487
International Classification of Functioning, Disability and Health (ICF) 9:543–549, 10:727–728, 20:656, 22:381–384
biopsychosocial integrative model 9:543–544
disability 9:543–545, 9:548
disability policy 9:548
and Engel's biopsychosocial model 9:545
environmental determinants 9:544, 9:547
health 9:543–544, 9:546–547
and health information 9:546–547
impairment 9:544
reduction of impairments 22:382
cognitive training 22:382
medication 22:382
remediation of persisting limitations 22:382–384
developing environmental resources 22:383–384
enhancing participation by vocational rehabilitation 22:383
family intervention 22:382–383
skills training 22:382
social and behavioral sciences, applications to 9:547–548
environmental factors, understanding 9:547–548
human rights and justice 9:548
universalism 9:548
understanding health in light of 9:545–546
International Classification of Impairments, Disabilities, and Handicaps (ICIDH) 6:472
International classifications, of labor statistics 5:798–799
International Code of Advertising Practice 1:204
International codes, for human subjects protection in research 8:48
International Coffee Agreement (ICA) 1:623–624
International collaborative research 20:508
International commercial arbitration 13:555
International commercial law 13:540
International Commission of Jurists 18:450
International Commission on Stratigraphy 1:722
International Committee for the History of Physical Education and Sport (ICOSH) 23:307–308
International Committee of Medical Journal Editors (ICMJE), recommendations 20:527
International Committee of the Red Cross (ICRC) 3:677, 25:382
International communication 12:490–494
ancient 12:490
broadcasting 12:492
data communication 12:493
international telephony 12:492–493
Internet 12:493
medieval 12:490
postal 12:490–491
telegraph 12:491
wireless telegraphy 12:492
International Communication Association (ICA) 4:276, 25:80–81
International communism, and Marxism–Leninism 14:660–661
International comparisons, macroeconomic data 5:755–756
purchasing power parities 5:755–756
real expenditures and real gross domestic product 5:755–756
International Comparisons Project (ICP) 5:755–756
International Conference on Cultural Policies of Development 5:493
International Conference on Cultural Policy Research (ICCPR) 5:453
International Conference on Emotions and Worklife 7:507
International Conference on Harmonization (ICH) 8:78–79, 8:81, 18:6–7
International Conference on Population and Development (ICPD) 8:794–796, 9:71, 20:457
International Conference on the History of the Language Sciences (ICHoLS) 14:218
International conflict, in political psychology 18:431
International Congress for School Effectiveness and Improvement 21:102
International connections, in Mexico 1:627
International context, obesity 17:78
International contracting 4:243–244
International Convention against Doping in Sport 6:669, 6:671
International Convention on the Elimination of All Forms of Racial Discrimination 6:534, 12:956, 13:720
International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICRMW) 10:171–172
International conversations, social problems 12:542
International cooperation
global public policy 10:178–179, 10:181–184
law and terrorism 13:533
Myrdal on 16:162
International Council for Local Environmental Initiatives (ICLEI) 6:832–833
International Council for Philosophy and Humanistic Studies (ICPHS) 22:677
International Council for Science (ICSU) 7:791, 22:675, 22:677
International Council of Archivists (ICA) (Paris) 5:702
International Council of Scientific Unions 20:422
International Council of Voluntarism, Civil Society, and Social Economy Researcher Associations (ICSERA) 25:252, 25:266–267
International Council on Monuments and Sites (ICOMOS) 5:488–490
International Council on Social Welfare (ICSW) 10:145, 10:147, 10:265, 10:718, 12:476–479, 12:495–500, 12:547, 12:567, 24:750
advocacy 12:497–498
expanding scope of work at global level 12:497–499
history of 12:495–496
integrated approach to social development 12:498
regional and national activities 12:500
social inclusion 12:498, 12:500
social justice 12:495–498
social protection 12:498–500
structure and organization of activities 12:496–497
vulnerability 12:496, 12:498–499
World Summit for Social Development (WSSD) 12:497–498
International Court of Justice (ICJ) 12:508–510, 13:820–821
International Covenant on Civil and Political Rights (ICCPR) 3:676–677, 5:495, 6:310, 11:356, 11:796–798, 12:956, 13:440–441, 18:450–451
Optional Protocol 3:676
International Covenant on Economic, Social and Cultural Rights (ICESCR) 3:676–677, 6:310, 7:573, 11:356–357, 11:796–797, 12:956, 13:440–441, 13:720
International crime data 5:714
Interpol 5:714
United Nations 5:714
International crime policy 5:177
International Crime Victims Survey (ICVS) 5:165, 5:848, 19:160
International Crimes Act 1973 25:386
International Criminal Court (ICC) 5:215–216, 11:369, 11:372, 13:595, 17:658, 18:500–502, 23:351, 24:445, 25:379
war crimes tribunals 25:386–387

International criminal law 12:507–508, 18:500
and war crimes tribunals 25:382–387
see also International Criminal Court (ICC)
International Criminal Tribunal for Rwanda (ICTR) 24:445, 25:383–385
International Criminal Tribunal for the former Yugoslavia (ICTY) 9:396, 24:445, 25:383–385
International criminal tribunals 12:510
International criminology 12:462–469
International cross-disciplinary associations 22:677
International Cultic Studies Association 16:808
International Data Systems 5:774
International Dating Violence Study (IDVS) 12:464
International Day Against Homophobia 21:737
International demography conferences 6:173
International development
ethical turn in 8:60–62
International development, ethical turn in 8:60
corporate social responsibility 8:61
ethical turn in 8:60
globalization 8:60
market integration 8:60
public campaigns 8:61–62
social movements 8:61–62
voluntary regulation 8:60–61
International differences, corporate governance 4:935
International diplomacy 1:982–983
contributions of Myrdal to 16:159
International disability movement 6:464
International disciplinary associations 22:676
International division of labor 13:160
International Economic Association (1949) 2:90
International economics 22:729
Smith on 22:119
International economy, in Cold War 4:129–130, 4:133
International education 4:429–430, 5:769
International education comparisons 5:769
International education industry 7:166–167
International education statistics 5:769–771
International education surveys 5:847
International Elder Law 13:422–423
comparative elder law 13:422
global and regional HR of older persons 13:422–423
growing interest in and development of the field 13:424
International Emergency Economic Powers Act (IEEPA) 13:531
International Encyclopedia of Social and Behavioral Sciences (IESBS) 20:549
International Encyclopedia of the Social Sciences 2:493, 25:341
The International Encyclopedia of the Social Sciences 7:585, 7:587
International Encyclopedia of Unified Science 18:623
International Energy Agency (IEA) 18:368, 18:370
International environmental law 12:507–508
International Ergonomics Association (IEA) 11:298–299
International ethical guidelines, for research involving human subjects 20:526–527
International Ethical Guidelines for Biomedical Research Involving Human Subjects 20:512, 20:514, 20:525–526
International Ethological Conferences 14:369
International Federation of Journalists (IFJ) 12:857, 12:859–861
International Federation of Red Cross (IFRC) 6:482
International Federation of Social Workers (IFSW) 1:785, 1:793–794, 7:336, 10:145, 10:147–148, 10:265, 10:530, 10:718, 11:353, 11:357–358, 12:474, 12:476, 12:478–479, 12:547, 12:567, 14:36, 19:478, 22:769, 24:750
ethical principles 3:521
International financial flows, and remittances 12:519–520, 12:522–523
International financial institutions (IFIs) 7:73–74
and Eastern European social work 22:796, 22:798
International financial law 12:508
International Financial Reporting Standards (IFRS) 13:916
International food regimes (IFRs) 1:460
International Gay and Lesbian Travel Association (IGLTA) 9:638–640
International Gay Games 9:640
International Geographical Union (IGU) 1:847, 22:677
International Geosphere-Biosphere Programme (IGBP) 7:791, 20:422
International Guidelines for Diagnostic Assessment 19:317
International Haplotype Map
see International HapMap Project
International HapMap Project 9:902, 9:1018, 11:309, 11:318
International Household Surveys Network (IHSN) 5:794
International Human Dimensions Programme 7:791
International Human Genome Project 2:616
International human rights 11:375–377
and feminist legal theory 8:923–924
law 1:248, 11:376, 12:507–508, 13:412–413
and domestic violence 6:629–630
International humanitarian and development agencies 19:479–480
International humanitarian law (IHL) 13:530
torture and 24:442
International Institute for Democracy and Electoral Assistance (IDEA) 18:394–395
International Institute for Educational Planning (IIEP) 4:429
International Institute for Higher Education in Latin America and the Caribbean (IESALC) 7:183
International Institute of African Languages and Cultures 3:6
International institutions, definition 16:416
International issues in child protection 3:423
International Journal for Sustainability in Higher Education 7:734
International Journal for the Psychology of Religion 19:311
International Journal of Cultural Policy (*IJCP*) 5:453
International Journal of Cultural Property 5:414
International Journal of Geographical Information Science 10:60
International Journal of Heritage Studies 5:414
International Journal of Illich Studies 11:588
International Journal of Psychology 11:793
International Journal of Social Psychiatry 19:311
International Labour Organization (ILO) 3:405–406, 5:369, 5:371, 5:800–801, 7:575, 10:257, 12:477, 23:208, 24:734, 24:818, 25:727
Convention No. 169 11:362
definition of basic needs 6:310
definition of minimum wages 15:561
establishment 9:215
Statistical Information and Monitoring Program on Child Labor 3:405
Treaty No. 169 11:796–798
International law and treaties 12:507–510
definition 12:507
domain 12:508
legislative and enforcement questions 12:508
nature and history 12:507–508
Roman *jus gentium* 12:507
International Law Commission 1993 25:386
International law firms 13:631
International law of Antarctica 12:508
International law of state responsibility 12:508
International laws 13:424–425
acceptance in East Asia 16:276
for domestic violence 6:630
genocide 10:15, 10:19
human rights 11:362, 11:365, 11:375–376
limitations of 22:395–396
terrorism and 13:530
International Lawyers Association 18:450
International Legal Ethics Conference (ILEC) 13:621

International legal process 12:509–510
International legal rules 12:508–509
International legal system, modern, formation of 11:796
International level regulation, of privacy on Internet 12:596–597
International macroeconomic imbalances 9:194
The International Marxist-Humanist 14:649
International Mathematics and Science Study 5:773
International mediation 4:627–628
International migration 6:271, 9:74, 9:78, 10:169–170, 11:649, 12:511–518, 15:389, 15:392–393, 15:412–413, 15:462, 15:469, 18:598, 20:361
  beginning 15:414
  changing geography 12:511–512
  determinants of 12:513–516
    average income differentials 12:513
    geographic and cultural distance 12:514–515
    immigrant networks 12:515
    immigrant policies 12:516
    income distribution in source and destination countries 12:513–514
  economic impact of 12:511, 12:516–518
  and economics 15:436–438
    assimilation and cohort effects 15:436–437
    gains to host country 15:438
    global gains from open borders 15:438
    labor market impact 15:437–438
  by ethnic Chinese 12:525
    current trends 12:525–526
    decisive economic effects 12:528
    diaspora 12:527–528
    ethnic entrepreneurship 12:527
    interpolarity of relations 12:527–528
    multipolarity of migration 12:527
    remittances 12:528
    revival of migrations from mainland China 12:526–527
    transnationalism 12:528
  by ethnic South and Southeast Asians 12:530–537
    brain drain 12:533–534, 12:536
    female migration 12:531–532, 12:535–536
    globalization 12:530, 12:534, 12:536
    issues 12:535–537
    migrant networks 12:534–536
    migrants 12:534
    migration industry 12:534
    patterns 12:530–534
    process 12:534–535
    refugees 12:533–534, 12:533t
    skilled migration 12:531–534, 12:536
    south–north migration 12:530–531, 12:533–534, 12:536
    south–south migration 12:530–531
    trafficking 12:534, 12:536
  in Eurasia 18:549–551, 18:551t
  first *versus* later trips 15:413–414
  impact flows on destination countries 12:516–518
    effects on output mix and on production techniques 12:516–517
    labor market effects 12:516
    productivity and innovation effects 12:517
    welfare-state effects 12:517–518
  Jewish 12:805–807, 12:806f–807f, 12:806t
  nonreciprocal exchange of favor 15:414
  and remittances 12:519–524
    activity in selected migrant origin-destination country pairs 12:520t
    consumption 12:521–522
    definition, magnitude and characteristics 12:519–521
    developing countries 12:519–523, 12:519f
    economic development 12:519, 12:523
    financial decision-making 12:521
    impacts 12:521–522
    insurance role 12:522
    international financial flows 12:519–520, 12:522–523
    investment 12:519–522
    motivations 12:521
    recipient countries 12:520t
  sanctioning 15:414
  social capital to 15:413
  technology adoption 12:516–517
International Migration Around the GlobE (IMAGE) project 12:434–435
International migration trends 15:416–419
International Military Tribunal (IMT) 25:383
International Military Tribunal for the Far East (IMTFE) 25:383
International mobility 10:240
International Monetary Fund (IMF) 1:724, 5:12, 7:204, 8:60, 10:182, 10:186, 10:198, 10:226, 10:601, 16:462, 18:724, 18:763, 23:598–603, 25:839
  structural adjustment programs 12:57
  World Economic Outlook 5:749
International Mr Bear Contest 13:890
International narcotics control system 6:671
International negotiations 12:572
  trade policy 24:488–489
International Network for Engineering Studies (INES) 7:641–642
International Network of Genocide Scholars 10:7
International networks, de-Westernization 25:544–546
International nongovernmental organizations (INGOs) 12:247, 18:450–451, 25:256, 25:259
  and world culture 10:225–229, 10:231–232
International official statistics, comparative databases 5:794–795
International Olympic Committee (IOC) 14:722, 23:303–304
International Organisation for the Support of Fighters for the Revolution (MOPR) 14:667
International Organization for Standardization (ISO) 10:225, 23:131
  ISO 9000 10:157
  ISO 26000 22:169, 22:173
  ISO 26.000 Social Responsibility 8:22
International Organization of Consumers Union 4:755
International Organization of Journalists (IOJ) 12:860–861
International organizations
  and education 12:538–541
    as agents 12:538–539
    cultural heritage 12:540
    globalization 12:538–539
    higher education 12:538–540
    human rights education 12:539–540
    knowledge economy 12:538–540
    technological education 12:539–540
    theoretical approaches 12:538
  global governance 10:157–158
  higher education research 10:865–866
International participation, in universal access to research products 8:34
International patent classifications (IPCs) 11:837
International Peace Research Association 22:677
International Personality Item Pool (IPIP) 11:172
  Abridged Big Five Circumplex scales (AB5C) 4:659–660, 9:232, 9:237
  measures derived from 9:233
International perspectives
  archaeology politics 1:889
  regulatory agencies 20:184–185
International Phonetic Alphabet (IPA) 18:65
International Physical Activity Questionnaires (IPAQ) 10:588–589
International Planned Parenthood Federation (IPPF) 8:786, 20:458
International Police Agency (Interpol) 5:714
International Police Organization (Interpol) 5:165
International policy
  ambivalence in 20:534–535
  child protection
    in Anglo-American countries 3:417
    in Belgian 3:418
    in Canada 3:417–418
    in Germany 3:418
    in New Zealand 3:417–418
    in Nordic and continental countries 3:418
    in United Kingdom 3:417–418
    United Nations convention 3:417
International political actors 17:650
International political economy 12:570
  trade policy 24:487
International Political Science Association (IPSA) 2:90, 18:468, 18:470–471
International politics

International politics (*continued*)
and balance of power 2:348
realism in 20:26
International prizes and awards 2:334–335
International Psychoanalytic Association 12:903–904
International Psychogeriatric Association 10:105
International psychosocial work, shifts in 19:479–481
International public goods 19:540–544
accessibility 19:540–541, 19:543
command-and-control 19:543
common-pool resources 19:540–543
context and institutional mode of provision 19:542–543
definition of public goods 19:540–541
excludability 19:540–541
implementation agreement 19:543
incentivization 19:540, 19:542–543
institutional answers to collective action problem 19:541–542
outcomes and impact of institutional provision 19:543–544
private provision 19:540
rival consumption 19:540–541
self-regulation 19:542–543
shadow of hierarchy 19:542
International Reading Literacy 7:212
International Red Cross 10:157
International refugee and immigration law 12:508
International refugee regime 20:111–112
International regimes
climate change 3:842–843
global governance 10:158
International Regional Science Review 20:147
International regulatory harmonization 9:191–193
International relations (IR) 3:64–65, 22:696, 22:699
Bielefeld School 2:557
causal-explanatory approaches to interstate war in IR 25:424–426
and constructivism 22:232
of Middle East 16:373–374
and political economy 1:939–940
and political power 18:763–764
Rawls' contributions 19:917
International relations modernization theory 11:433–434, 11:437
International relations theory 17:632–633
International relations/politics and human rights 11:377
International Research and Training Institute for the Advancement of Women (INSTRAW) 24:754
International research projects 24:759–760
International Residual Mechanism for Criminal Tribunals 25:385
International Risk Governance Council (IRGC) 7:787
risk governance model 7:787, 7:789
International School Improvement Project (ISIP) 21:91
International schools 2:581
International science 22:596
International scientific organizations, de-Westernization 25:544–545
International Self-Report Delinquency Study 5:165, 5:719
International service 10:854
International social problems 12:542–549
actors 12:547
clustering of 12:543
definitions 12:542
distribution of wealth 12:542–543
future 12:547
globalization 12:542–543
health 12:545–546
interpersonal relationships 12:545
key concepts 12:542
livelihoods 12:544
people and environment 12:545
people and knowledge 12:546
people and power 12:544–545
political–philosophical stances 12:543
re-envisioning 12:543
role of social workers 12:547
social problems 12:546
social solutions 12:546
sociological theories 12:543
strengths-based practice 12:543
well-being 12:545–546
International social science and humanities prizes 2:335
International Social Science Council (ISSC) 20:744–745, 22:675, 22:677
World Federation for Mental Health 22:677
International Social Survey Program (ISSP) 5:792, 5:820–821, 5:827, 5:833, 12:550–556, 18:494–495, 23:748, 12:551t–554t
Allgemeinen Bevolkerungsumfragen der Sozialwissenschaften (ALLBUS) 12:550
British Social Attitudes (BSA) Survey 12:550
comparative research 12:556
cross-national survey research 12:556
functional equivalence 12:550
General Social Survey (GSS) 12:550
and public opinion 19:563
International social work (ISW) 10:264–266, 12:557–561, 17:432, 24:512–516
in Asia 12:559–560
and choice 3:521–522
comparative social work 12:558–559
contextual social work 12:558
versus global social work 12:558–559
International Association of Schools of Social Work (IASSW) 12:474–479
internationalization of social work 12:557
profession and education 12:557
trafficking of women and children in Latin America 24:512–516
universal social work 12:557–558
western models 12:557
International social work organizations, at United Nations 24:750–751
International social work student exchanges 12:562–567
concepts and theoretical frameworks 12:563–564
cultural efficacy 12:563
dual perspective 12:563
economic development 12:562–563
future and assessment of 12:566
goals 12:564t
human rights 12:565
orientation 12:563–565
outcomes 12:564–565
planning 12:564–566
practical frameworks 12:564–566
social development 12:562–563, 12:567
types 12:564–566
International Socialism: A Quarterly Journal of Socialist Theory 14:649
International Socialist Review 14:649
International Society for Ecological Economics (ISEE) 6:853, 6:861–862
International Society for Gesture Studies 10:109
International Society for Humor Studies 11:416
International Society for Krishna Consciousness (ISKCON) 16:805–806
International Society for the History of Physical Education and Sport (ISHPES) 23:307–308
International Society for the New Institutional Economics (ISNIE) 17:192
International Society for the Prevention of Child Abuse and Neglect 3:418
International Society for the Study of New Religions 16:808
International Society of Medical Planning Professionals (ISMPP) 18:9–10
International Sociological Association (ISA) 2:90, 3:122, 13:515, 22:675, 22:912, 22:993, 25:193
International Standard Classification of Education (ISCED) 5:770–771, 5:770b
International Standard Industrial Classification of all economic activities *see* ISIC Rev.4 (International Standard Industrial Classification of all economic activities)
International Standard Industrial Classification of all Economic Activities (ISIC) 5:760, 5:798–799, 5:801
International standardization, comparative databases 5:796
International Statistical Congress 1:852, 3:303
International Statistical Institute 3:303, 9:222, 17:662–663, 20:876, 23:426

International student mobility markets 10:855
International studies
of culture 1:933–937
political economy 1:938–942
International Studies Association (ISA) 22:661
International survey of literacy skills 11:590–591
International survey programs, comparative databases 5:792
International survey research 5:827
International system of macroeconomic accounts 5:752–753
International system of nation-states 11:648–650
International systems 23:368
and political power 18:763
International tax law 12:507–508
International Telecommunications Union (ITU) 5:705, 18:346
International telephony 12:492–493
International terrorism 13:531
International Terrorism: Attributes of Terrorist Events (ITERATE) database 5:837
International trade 12:568
commercial policy and trade negotiations 12:568–573
economic integration 12:574–579
policy 24:485–487
International trade law 12:508
International traditions linkages 22:969
International trafficking in women 24:596
International training, infrastructure 22:672
International Training and Research Program 1:933
International treaties
and indigenous rights 11:796–798
jurisprudence 11:798–799
for political repression 18:450–451
International Typographers Union (ITU) 3:203–204
International Union for the Conservation of Nature and Natural Resources (IUCN) 7:810–811
International Union for the Scientific Study of Population (IUSSP) 1:281
International Union of Anthropological and Ethnographical Sciences 2:90
International Union of Anthropological and Ethnological Sciences (IUAES) 22:676–677
International Union of Psychological Science (IUPsyS) 2:90, 22:676–677
international victimization surveys 5:719
International Violence Against Women Survey 10:196
International Visual Sociology Association 25:193
International War Resister League 17:644
International Woman Suffrage Alliance 22:444
International Workers Day 24:492
International Workingmen's Association 1:680–681
International Year of the Woman 10:193
Internationalisms 17:641, 23:4
and peace movements 17:643–644, 17:646
in South America 23:48
Internationalist solidarity 22:847
Internationalist view
of social work education 22:776–777
Internationality 2:350
Internationalization 6:833–834, 10:264, 10:850, 10:852, 10:854, 10:867, 12:580, 13:614–615, 21:650, 22:622–623
of capital 3:93
elites 7:400
in higher education 10:854–855
higher education research 10:862, 10:866–867
of history 4:405
in Japan 12:795–796
journalism 12:855
of research 20:509–510
of sciences 22:648–649, 22:654
of sexual citizenship 21:732
of social sciences 6:489, 20:744–745
of survey research 12:672
Internationalization, patterns of 12:580–585
foreign direct investment 12:580–586, 12:583f, 12:585f
'how&rsquo 12:583–584
at functional level 12:584
new currents 12:584
investment modes 12:580–585
location 12:581, 12:583–585
MNCs 12:580–586
motives 12:580–581
old and new patterns of 12:585–586
process 12:584–585
sequence 12:580
'when&rsquo 12:581–582
firm perspective 12:582
'where&rsquo 12:582–583
patterns of incrementalism 12:583
'why&rsquo 12:580–581
emerging new motives 12:581
International-signaling
by constitutions 4:713
Internet 8:487, 10:61, 12:90–92, 12:490, 12:493, 12:606, 12:909–910, 13:521, 16:776, 16:791, 16:800, 18:678–679, 20:363, 20:793, 22:662–663, 23:640–641, 24:104–105, 24:891
and access to culture 5:461–462
and administrative law 1:110
advertising 1:214
anthropology of 12:600–605
cyberculture and digital anthropology 12:600–601
cyberspace 12:600
digital media 12:600–604
Facebook 12:600, 12:602–604
globalization 12:603
hyperbole and social change 12:601–602
network society 12:601–603
online knowledge production 12:604
social media 12:600, 12:602, 12:604
translocality 12:603
Twitter 12:600, 12:602
virtual 12:602
visual culture 12:604
Web 2.0 12:600
Wikipedia 12:600
YouTube 12:600, 12:602
and archives 5:702
artistic/cultural markets 14:567
Castells on 3:230–231
and census 3:304
and child development 15:3, 15:5
child pornography 20:438
in China 2:68
and climate change 3:853, 3:858
communities 4:333–335
computer-mediated communication (CMC) 4:504–509
conservative communities 20:276
and cultural consumption 16:557
and culture 12:587–592
audiences 12:590–591
material production and distribution 12:588–589
prospects 12:591–592
research perspectives 12:587–588
symbolic production 12:589–590
death/dying 5:865, 5:868
definition 12:600
and diasporas 6:355
and doctor–patient interaction 6:616
and econometric software 7:3
ethics of research 12:607–611
and gay/lesbian tourism 9:638–639
and global capitalism 3:102
and global terrorism 25:740–741
identity and discourse on 6:500–501
and identity movements 11:545
and informed consent 12:610
and journalism 12:851, 12:855
and McDonaldization 14:830
and media audiences 2:218
media ethics 15:46
and networks 16:567–570
oversharing through 17:439–444
and political communication 18:345–347
and politics 14:716–717
pornography *see* Online pornography
and postmodernism 18:676–679
and privacy 12:593–599
consumers 12:594–595
cybercrime and cybersecurity 12:595
data protection 12:594–597
development of Internet usage 12:594
Facebook 12:593–594, 12:597
Google 12:593–594, 12:596–597
international level regulation 12:596–597
national level regulation 12:596
national security 12:597
surveillance 12:595–596

Internet (*continued*)
technology 12:593–594
ubiquitous computing 12:598t
user demand 12:594–595
and religion 20:273, 20:275–276
Christianity 20:274
current approaches to study 20:276–277
Hindu rituals 20:275
OnIslam 20:273–274
Second Life 20:273–274, 20:277
smartphone apps 20:276–277
sale of Chinese cultural products in 5:600
and same-sex relationships 21:735
and same-sex sexualities 20:871
and secrecy 21:378
and sexual citizenship 21:732
sexual commerce in 13:884
and sexual minorities in China 21:737
and social capital 22:178
and social media 16:783, 16:785
and statistical analysis 23:380
surveys and polls on 23:756
and terrorism 24:227–236, 24:241–242
and therapeutic effectiveness assessment 24:283
Internet abuse
child protection and 3:427–428
Internet Archive 24:234
Internet dating 17:488
as erotic capital 8:6
Internet defamation 9:395
Internet experiments 9:132
Internet forums 6:898, 14:244
Internet fraud 5:235
Internet geographies 12:92
Internet interventions
and psychotherapy 19:380–381
Internet protocol 12:600
Internet protocol-based texting 15:630
Internet research 22:548
ethics 12:606
Internet service providers (ISPs) 20:167–168
Internet sex offenders 20:438–439
Internet support group (ISG) 10:687
Internet technology 5:704–705, 5:708, 5:710
Internet-based communications 5:769
Internet-based presentations 19:671
Internet-based research
informed consent in 12:113–114
Internet-based 'telewellness' programs 6:881
Internet-only offenders 20:438
Interoception 12:614–620, 23:866
assessment methods 12:615–616
heartbeat detection tasks 12:615
heartbeat-evoked potentials 12:615–616
visceral modulation of startle 12:616
body perception 12:614–616, 12:619
concepts 12:615
and decision making 12:617
Damasio's somatic marker hypothesis 12:617
and emotions 12:616–617
current point of view 12:617
early concepts 12:616
future directions 12:619–620
in learning and memory 12:618
in mental disorders 12:618–619
depression 12:618–619
eating behavior and eating disorders 12:619
panic disorder 12:618
somatoform disorders 12:619
neurophysiology and neuropsychology of 12:614–615
psychophysiology 12:616–617, 12:619–620
and self 21:419–420
and stress 12:617–618
accuracy during and after acute laboratory stressors 12:617
chronic stress 12:618
stress hormones 12:617–618
symptom perception 12:618–619
tasks, factors determining accuracy in 12:616
differential aspects of interoceptive accuracy 12:616
physiological determinants 12:616
Interoceptive accuracy 12:615
Interoceptive awareness 12:615
Interoceptive sensitivity 12:615
Interoceptive stimulation
and symptom perception 23:867
Interoceptive threat
and fear-potentiated startle 8:863–865
Interoperability 4:489
Interorganizational analyses
of issue networks 12:770
Interorganizational networks
clusters of organizations 12:623–625
social network analysis 12:622–623
Interorganizational relations 17:343–345, 17:375–376
network analysis 16:519
Interorganizational relationships 12:621–625
competitive advantage 12:624
consequences of 12:622
forms of 12:622
innovation 12:622, 12:624
learning 12:622–624
motives 12:621–622
networks
clusters of organizations 12:623–625
social network analysis 12:622–623
organization studies 12:622
social network analysis 12:622–623
interlocking directorates 12:623
principal methods 12:623
research subjects 12:623
social capital 12:623
theoretical approaches and methods 12:622
Inter-organizational relationships
social capital 22:184
Interparental conflict
and divorce 14:522
Interpellation
of the human subject, and language usage 14:200
and infrastructure 12:121
Interpenetration, Luhmann on 14:394–395
Interpersonal adaptation theories
of computer-mediated communication 4:506–507
Interpersonal adjustment
implications of shame and guilt for 21:476
Interpersonal attraction 12:626–630
desirable personality 12:626–627
differences and self-expansion 12:627
familiarity 12:628
ideal self-attraction 12:627–628
physical attractiveness 12:626
propinquity 12:628
reciprocal liking 12:628
similarity 12:627
situational influences on 12:628–629
Interpersonal circumplex (IPC) model
of gender differences 9:752
Interpersonal communication 4:274, 14:726
Interpersonal concepts and principles
for physical activity 18:92–93
Interpersonal conflict
adaptive correlates 4:597
developmental outcomes 4:597
mechanisms 4:597
definition 4:594–595
maladaptive correlates 4:597
developmental outcomes 4:597–598
mechanisms 4:597–598
Interpersonal conflict of laws 13:615
Interpersonal contexts
differences, during preschool age 18:223–224
Interpersonal coping strategy 16:862
Interpersonal deviance
in workplace 25:722–723
Interpersonal disgust 6:547
Interpersonal emotions, in workplace 7:508–509
and emotional dialect 7:508–509
Interpersonal factors
influencing creativity training 5:115–116
patient adherence 17:599
Interpersonal functioning assessment
empathy 2:756
intimacy 2:756
Interpersonal group psychotherapy 10:429
interpersonal learning 10:430
Interpersonal influences
in networks 15:660–662
mechanism, formalization of 15:662–664
Interpersonal intelligence 17:752
Interpersonal level
of stigma analysis 23:464–465
Interpersonal motivation 15:896
Interpersonal phenomenon
rejection of neonatal imitation 16:484
Interpersonal process
dyadic coping as 11:594–595
early defiant behavior 6:15–16

Interpersonal psychotherapeutic process 4:725
Interpersonal psychotherapy (IPT) 6:827, 11:615, 12:631–636, 19:460
for binge eating disorder 19:376–377
for bulimia nervosa 19:376
for depression 14:749, 19:375–376
development 12:631–632
evidence-based treatment 12:634–635
future directions 12:634–635
mediators and moderators 12:634
treatment efficacy 12:634
widespread dissemination 12:634–635
outcome research 12:633–634
adaptations 12:633–634
for perinatal mental illness 17:724
therapeutic stance and techniques 12:633
treatment phases, goals, and strategies 12:632–633
Interpersonal psychotherapy for group (IPT-G) 12:633
Interpersonal relatedness 17:867–868
Interpersonal relationships 3:929–930, 7:446, 12:545
during competition/cooperation 4:858–860
and constitution of personality structures 18:809
cultural influences on 5:426–432
choice and friendships 5:427–428
collectivism 5:426–427, 5:429–430
cultural meaning of friends and enemies 5:427
face 5:428–429
face in interpersonal conflict 5:429
familism 5:430
honor 5:430–431
individualism 5:426–427, 5:429–430
relational mobility 5:427–428
reputation 5:426–427, 5:430–431
social support 5:429–430
threats to face 5:428–429
value orientations and models of self 5:426–427
network analysis 16:518–520, 16:522
and self-esteem 21:524–525
Interpersonal roles, managers 14:473
Interpersonal skills 24:57–58
Interpersonal trust
and social capital 22:174–175, 22:178
Interpersonal trust across Lifespan 12:637–640
Interpersonal-oriented moral reasoning 7:316t
Interpol 24:597
and transnational organized crime 24:596–597, 24:600
Interpol DNA Gateway 9:340–341
Interpol Standard Set of Loci (ISSOL) 9:342
Interpolar Chinese migration 12:527–528
Interpolated Markov Chains (IMaCh) 16:113
Interpretation bias 22:496–497
Interpretation model
legal reasoning 13:781
Interpretations 2:495, 3:820, 5:472–473, 20:881
in anthropology 12:641–644
ethnography 12:642–644
hermeneutics 12:641–642
method 12:642–643
object of anthropology 12:642
performance 12:643
phenomenology 12:642–643
politics 12:643–644
of caches 20:724
choices 19:442–443
and critical theory 5:301–302
Dilthey on 6:424–429
of error phenomena
accommodations in error processes 23:230–231
building models 23:233
lexical selection processes 23:232
processing frames 23:231–232
sound exchanges and word exchanges 23:231
stranding errors and multiple levels of processing 23:231
of health 15:121
hermeneutics 10:779–780
and horizons of experience 10:787
in hypothesis testing 11:497–498
of ideational systems 20:722
managerial and organizational cognition 14:480
of meanings 10:693
and objectivity 17:89
postmodern methodology as 18:681
probabilistic grammars 19:9–12
of probability 19:23
epistemic 19:24–25
quantitative probability 19:38–41
Fisher–Kyburg interpretation 19:41
logical probability 19:39–40
objective probability 19:40–41
subjective probability 19:38–39
of symbols 20:721–722
and *Verstand* (understanding) 10:785
Interpretative case studies 3:209
Interpretative paradigm
and social action 1:76–77
Interpretative phenomenological analysis 22:546
Interpretative principles
electroencephalography 7:356–359
abnormal classification 7:357
artifact discrimination 7:356
clinical interpretation 7:357–358
montages 7:358–359
normal and benign variant patterns 7:356–357
personnel 7:356
Interpretative repertoire 22:225
Interpretative sociology, Schütz on 21:146
Interpretative understanding 22:992
Interpreter bias 8:572
Interpreters 4:290
Interpretive ambiguities in moderation 15:673–674, 15:673t
Interpretive analysis 18:204, 18:206–207
Interpretive anthropology 1:747, 1:749
Interpretive archaeology
and hunter-gatherers 11:426
Interpretive flexibility 24:136–137
Interpretive methods
macromethods 12:645–647
micromethods 12:648–651
Interpretive objectivity 12:647
Interpretive phenomenological analysis
qualitative developmental psychology 19:673
Interpretive problems
linear hypothesis 14:150–152
Interpretive research 12:648–650
inquiry as site of multiple methodological research practices 12:649–650
Interpretive social science 2:509, 10:778–779
Interpretive studies of everyday nationalism 16:310–311
Interpretivism 15:712, 22:690, 22:987
and cultural sociology 5:498, 5:500
and positivism 18:625–626
Interpretivist epistemology 4:726
Interpretivist normativism 2:751
Interprofessional model of evidence-based practice 8:333
Interracial contact 19:834
Interrater reliability 6:345, 8:585, 17:687
Interrelation
and senses 21:619
Interrelational stage of development 16:504
Interrogation
torture and 24:443–445
Interrogative utterance boundary tone 23:719
Interrogatory
discovery 6:517, 6:520–521
Interrole conflict 8:760
Interrupted time series quasi-experiment design 12:652
construct validity 12:654–655
external validity 12:655
internal validity 12:653–654
statistical conclusion validity 12:652–653
Interrupted time series quasi-experimental designs 12:652–656
Interrupted time-series (ITS) design 12:453
Interruptions
in negotiation 16:425
Intersecting inequalities 7:118
Intersection property 10:346
Intersectional anxieties 21:810
Intersectionality 7:118–119, 8:906, 8:908, 9:690, 9:698–699, 9:701, 12:657–662, 21:810–811, 22:321
advent of 12:659
approach 13:164
background and early roots 12:658–659
bifurcated consciousness 12:658–659
and biographies 2:640–641
and butch/femme identities 13:871–872
conflict 12:661
and crime 12:663–669

Intersectionality (*continued*)
background 12:663–664
class 12:663–668
classifications 12:664–665
criminal offending 12:665–668
in 'doing identity' through crime 12:666
feminist criminology 12:663, 12:665–666
gender 12:663–667
at macrolevel 12:666–667
methodological approaches 12:664–665
oppressions 12:663–668
race 12:663–667
theoretical links and developments 12:665
traditional theories of 12:665–666
to understand system involvement 12:667
victims of crime 12:667–668
and difference/diversity/multiculturalism studies 12:660–661
double consciousness 12:658–659
and drag culture 13:877
feminism 12:658–660
feminist 8:899–900
and feminist criminology 8:913
of feminist psychotherapy 8:929–932
and feminist social work 8:937
gender 12:657–662
inclusion 12:660
interaction 12:657–662
and masculinity 14:682
matrix of domination 12:660
methods and analysis 12:661–662
oppression 12:657–660, 12:662
and oppression 17:247–248
privilege 12:657–662
and queer theory 19:765–770
race 12:657–662
in race/racism 19:796–798
and sanism 15:200–202
sexuality 12:658–661
social class 12:657, 12:659–661
social identity 22:321
standpoint theory 12:658–659
stratification 12:657, 12:659–661
theory 9:731–732
women's studies 9:736–740
Intersection–union test
for hypothesis 11:492
Intersectoral collaboration
community care in England 4:346–347
Intersession experiences (IEs)
psychotherapy 19:518–519
study of 19:518–519
Intersex
and social work 21:802–803
Intersocietal level of analysis 13:900
Interspatial geographic models 14:787
Interstate Commerce Commission 20:161, 20:178–179
Interstate scale
ethnic conflict at 8:91–92
Interstate system
war 25:401–402
Interstate war 25:420–423
interstateness as reciprocity and similarity 25:423
state at war 25:421–422
war as organized violence 25:422–423
Interstitial nuclei of the anterior hypothalamus (INAH) 21:775
Interstitial spaces
liminality 14:131
Intersubjective perspective-taking 18:328
Intersubjective psychoanalysis 19:493
Intersubjective testability (publication) 20:441
Intersubjective theory
conceptions of mind 22:763
psychodynamic treatment 22:764
psychopathology in 22:763
Intersubjectivity 2:883, 16:183, 17:286–287
of humans 11:253–254
innate intentionality *see* Innateness
neonatal 16:483–484
primary intersubjectivity 12:343
secondary intersubjectivity 12:343–345
and self 21:443
and sociality 22:856–857
tertiary intersubjectivity 12:344
Intertemporal conflict of laws 13:615
Intertemporal decision behavior 17:782
Intertemporal division of labor 6:601
Intertemporal substitution 25:495
Intertextuality analysis 16:196
Interuniversity Consortium 2:190
Inter-University Consortium for Political and Social Research (ICPSR) 1:916, 5:693, 5:831–832, 23:748–749
Inter-University Seminar on Armed Forces and Society (IUS) 15:515–516
Inter-uterine device (IUD) 9:77
Interval censoring
in event history analysis 8:317
Interval estimation 8:18–19
distribution-free 17:292
Interval games 4:879
Interval orders 17:526–527
Interval scales
construction of 24:253
Interval-contingent assessments 10:590
Intervening variables 15:64
Intervention analysis, time series 24:319–320
Intervention cluster 21:104
Intervention competency 24:520
Intervention efficacy, for HIV infection
adolescents and youth 11:139
heterosexual men 11:142–143
MSM 11:141–142
women 11:140
Intervention implications 23:560
Intervention mapping 10:622–623
Intervention programs 25:539
Intervention strategies
gifted education 7:163
social representations 20:446–447
Interventionism 24:866
and mental causation 15:177
Mumford on 16:139
Interventions 10:620–621, 17:595, 24:174, 24:177–178, 25:539
and academic achievement 1:46–47
applied sociology 1:858–859
attachment-based 12:233
coping with chronic illness 3:552–553
and counterfactuals 5:92–93
for doctor–patient interaction 6:618
fathering 8:844–847
health literacy 10:654
measuring effectiveness using meta-analysis 23:909–910
to prevent accidents and injuries 12:132–133
psychological treatments, empirically supported 19:382–385
and reentry after incarceration 20:61, 20:64–65
social representations as 20:446–447
for social work stress/trauma
training models for 18:711
types of 18:709–710
for street children 23:537
subjective well-being 23:651
gratitude 23:651
kindness 23:651
visualizing selves 23:651
for youth physical activity motivation 15:916, 15:918–919
Interviewee 17:287–288
Interviewer bias 5:851–853, 5:855
Interviewers 12:673, 17:288
Interviewing 5:851–856, 12:670–676
control 5:851–852
in ethnography 8:186–187
exchange theory 5:853–854
eyewitness testimony 8:645–646
field observational research 9:137–138
formats and content 5:851–852
historiography of 12:670
history of
critical voices 12:673–674
interviewers 12:673
social acceptability of 12:674
special populations 12:673
survey interview I 12:670–671
survey interview II 12:672–673
unstructured 12:671–672
multimethod strategy 5:855–856
power theory 5:854–855
role theory 5:852–853
sexuality studies 15:314–317
situations 5:851
spring theory 5:852
symbolic interactionism 5:855
Interviews 17:284, 17:288, 17:851–852
formats of 5:851–852
for health behaviors 10:590
for job analysis 12:813
personnel selection 17:938–939
protocols, SRL 21:538
qualitative developmental psychology 19:671
in qualitative health research 10:693

in qualitative research 19:685, 19:687
qualitative research methods 22:544–545
in sampling 20:878
Interwar period
oriental studies in 1:949–950
Interwar world
social science research programs in 6:488
InterWell health promotion program 1:151
Inti Raymi 23:47
Intifida 16:384
Intimacy 2:756
in close relationships 22:565–566
and early friendships 9:451
versus isolation 11:539–541
and marriage 14:599, 14:601–603
and online dating 17:212–216
and radio 19:877–878, 19:881
as relationship process 14:375–377
disclosure, involvement, and responsiveness at work 14:376–377
dyadic intimacy process model 14:376–377
involvement behaviors 14:375–376
responsiveness 14:376
self-disclosure 14:375
sex 14:375
Intimacy crisis 1:435
Intimacy projects 17:764
Intimate citizenship 21:729–730
Intimate comradeship 9:645–646
Intimate labor
in sex work 19:279
Intimate partner abuse (IPA) 8:914–915, 12:685–686
applied research on 12:677–684
abusers 12:678–679
businesses 12:679
communities 12:679–680
coordinated community action models 12:682–683
current state of knowledge 12:678–683
definitions 12:677
educational institutions 12:680–681
first responders and medical personnel 12:681–682
methodological limitations and future directions 12:683
primary prevention 12:677–678, 12:682–683
program evaluation 12:677–679, 12:682–683
religious leaders 12:682
secondary prevention 12:677
socioecological model 12:678
tertiary prevention 12:677–678
theoretical perspectives 12:677–678
youth 12:680–683
Intimate partner terrorism 12:686
Intimate partner violence (IPV) 5:723, 6:625, 6:627, 6:633–634, 12:685–688
common couple violence 12:685–686
and gender 12:686–688
Johnson's categories of 12:685–686
and race/ethnicity 12:686–687
and sexual orientation 12:686
situational couple violence 12:686
in United States 5:723–725
see also Intimate partner abuse (IPA)
Intimate relationships
and adolescent social competence 22:217–218
fieldwork ethics 9:154
Intimate sciences
and infrastructure 12:124
Intimate terrorism 12:686
Intolerable Acts 1:610–611
Intolerance for Uncertainty scale 24:719
Intonation 1:188, 23:714–716
Intonation patterns 13:256–257
Intonational phrase 23:718–720
Intoxication 5:222–223
alcohol 1:510
effects of heroin 17:240
forensic psychology 9:352
Intracarotid amobarbital test (IAT) 12:689, 12:691
Intracarotid sodium amytal testing (IAT) 12:689–695
crowding hypothesis 12:694
current use 12:690
functional magnetic resonance imaging 12:690–691, 12:694
handedness 12:693–694
hemispheric dominance 12:692, 12:694
history 12:689–690
internal carotid artery 12:690–691, 12:691f
language lateralization
in bilingual and deaf individuals 12:694
early studies of 12:693
language study 12:692–693
alternative methodologies 12:693
magnetoencephalogram 12:690–691, 12:694
patterns of hemispheric speech representation 12:694
predictors of atypical language representation 12:693–694
problems and considerations 12:691–692
rationale and description 12:690–691
sign language 12:694
temporal lobe 12:690–691
Wada test 12:689–691, 12:690f
Intracellular signaling modulation by opsins 17:271
Intraclass correlations (ICCs) 14:63, 16:55
in crossover designs 5:349
Intracluster correlation
linear mixed models 14:162–163
Intracohort variability 1:329
Intracortical microstimulation 15:965
Intracranial self-stimulation (ICSS) 2:841–845
Intractability
and conflict 4:585
Intra-cultural change 2:192
Intracytoplasmic sperm injection (ICSI) 12:47
Intrafamilial allocations
and twin studies 24:707
Intragenerational mobility 15:640, 15:642
Intragenomic conflict 22:910–911
Intragroup conflicts
organizational groups 10:412
Intragroup processes
and tyranny 24:710–711
Intrahemispheric processing 12:417
Intraindividual aspect of sexual development 14:328–329, 14:339
Intraindividual level of analysis 13:899
Intraindustry spin-off processes 11:830
Intralocus sexual conflict 17:901
Intramethod mixing 15:621
Intransitivity
choices 19:440–442
Intra-organizational relationships
social capital 22:181
Intraparietal sulcus (IPS) 16:640–641, 16:673, 17:46, 17:49–50, 17:52
Intrapersonal communication 4:274
Intrapersonal concepts and principles
for physical activity 18:92–93
Intrapersonal conflict
adaptive correlates 4:595–596
developmental outcomes 4:595–596
mechanisms 4:595
definition 4:594
maladaptive correlates 4:596
developmental outcomes 4:596
mechanisms 4:596
Intrapersonal intelligence 17:752
Intrapersonal phenomenon
rejection of neonatal imitation 16:484
Intrapreneurs 7:676
Intrapsychic conflicts 19:339–341
Intrapsychic coping strategy 16:862
Intrasentential code-switching 3:918
Intrasentential mixing 3:918
Intrasituational switching, in linguistics 3:918
Intrastate scale
ethnic conflict at 8:92–93
Intrauterine device (IUD) 8:221
Intra-West tensions
in Cold War 4:129–130
Intreccio 14:424–425
Intrinsic case studies 3:197
Intrinsic compensation 17:611–612
Intrinsic education for sustainable development 7:735–736
Intrinsic goals 21:490
Intrinsic goods, and market allocation 14:542
Intrinsic interest, in vocational psychology 25:241
Intrinsic mortality 6:756
Intrinsic motivation 1:348, 8:505, 12:696–701, 15:931, 21:51, 21:487, 24:187
collectivism 12:696–697, 12:699–700
Confucian approach to learning culture 12:696–697, 12:699–700
and creativity training 5:115–116
culture-specificity *versus* universality 12:699–700

Intrinsic motivation (*continued*)
defined 12:696
environmental influences on 12:698–699
extrinsic rewards on, effects of 21:487
and flow 9:260–263
indigenous Chinese 12:697, 12:699–700
individualism 12:696–697, 12:699–700
in non-Western cultures 12:696–697
empirical evidence regarding intrinsic motivation 12:697–699
self-determination theory 12:699–700
Socratic approach to learning 12:696, 12:699–700
and vocational interests 25:241–242
in Western cultures 12:696
Intrinsic motivation theories
for academic achievement 1:23
developmental changes 1:23
of self-regulated learning 21:542–543
Intrinsic motor connections 18:847
Intrinsic regulation 24:68
Intrinsic reinforcement 20:207
Intrinsic religiosity 20:382–383
Intrinsic spatial terms (language) 13:295
Intrinsic value 1:37
task value 7:890–891
Introceptive exposure
for panic disorder/agoraphobia 17:478
Introduction, methods, results, and discussion *see* IMRAD (introduction, methods, results, and discussion) format
Introjected regulation 21:488, 24:68
Introjection 5:578
Introns 12:300
Introspection 17:232, 21:172
and behaviorism 12:703–704
classical controversy 12:702
and clearing out preconceptions 12:703
and experiment 12:702–703
history of concept 12:702–704
imageless thought controversy 12:703
phenomenal approach 12:703
philosophical background 12:702
as retrospection in memory 12:702
verbal reports 12:703
Introspectionism 17:532
Introspective psychology 12:787
Introversion
and temperament 24:187
Intruz 3:146
Intuition
and decision making 5:921
health risk perception 10:703
heuristics 10:827
managerial and organizational cognition 14:481
and reflective equilibrium 20:80
value pluralism 25:22
Intuitionistic logic 16:925
Intuitive precursors 12:101
Intuitive response 23:295
Intuitively appealing examples in prospect theory 19:266
Inuit (Arctic) 8:177
Inuit prehistory 1:921
Inuktitut (Eskimo) 14:221–223
Invariance 14:571
fiducial and structural 9:124–127
Jakobson on 12:783
of speech sounds 11:324
in statistics 12:705–709
estimation 12:707–708
groups of transformations 12:706–707
inference 12:705–707, 12:709
invariant hypothesis testing 12:708
models 12:706
Pitman on 12:705–706
principle of invariance 12:707
probability 12:705–706
Invariance principle 23:321–322
Invariance theorems for axiomatized theories 2:341
Invariant hypothesis testing 12:708
Invariant under translation 12:705
Invariants, in perceptual learning 10:129–130
Invasion from Mars 7:563
The Invasion from Mars (radio program) 22:345
Invented religions 16:806
Invented traditions 16:294, 17:35
nation as 16:294
Inventio 20:650
Invention 12:145–146, 24:141
concept of 24:129–130
and innovation, evolution of technology 24:131
of invention 24:141
and organizational boundaries 2:771
of tradition approach, cultural heritage 5:414–415
Inventories
factor structures, validation of 7:940–941
self-report methods 21:538–539
Inventory maps 3:171
Inventory of Personal Characteristics (IPC-7) 9:232
Inventory of Psychosocial Balance 7:941
Inventory of Psychosocial Development (IPD) 7:940
Inverse care law 22:286–287
Inverse extreme value maximum random utility model 2:551
Inverse Mill's ratio 21:415
Inverse probability 2:379–380
Bayes' theorem 2:379–381
Pearson's theorem 17:662
Inverse probability of treatment weighted (IPTW) estimator 15:78, 19:144
Inverse probability weighting (IPW) 15:603
Inverse projection (IP) 6:149–153
age distribution of death 6:151
aggregate data 6:149
back projection 6:150
backward projection 6:150–152
census 6:149, 6:151–152
developments 6:150–152
differentiated 6:151
ergodicity 6:150
forward projection 6:150–152
generalized 6:150–152
historical demography 6:149
incomplete information 6:149
life tables 6:149, 6:151–152
linear interpolation 6:149
method 6:149–150
optimization 6:150–152
origins 6:149
parish data 6:151
population reconstruction 6:150, 6:152
stable population 6:150–151
stochastic 6:151–152
trend projection 6:151
Inversion
infrastructural 12:123
tricksters 24:648–649, 24:651–653, 24:655–656
Inversion effect (prosopagnosia) 19:252–253
Inversion thesis 23:329
Invertebrates
lateralization in 2:802
development 2:802–803
function 2:803
Inverted-U curve 13:144, 13:146–147
Investigating constituent order across languages 25:645–646
Kiribati 25:645
Makwe 25:645–646
Nivkh 25:645
Tiwi 25:645
Investigation citizen science projects 3:632
Investigative journalism 12:859
Investigator obligations
placebo-controlled trials 18:166–169
Investiture
feudalism 9:111–112
Investment
and capitalism 3:105–108
economic psychology 5:888–889
in health literacy 10:655
infrastructure 12:116–118
and marriage 8:829
money 15:737–738
and profits 19:111–115
in R&D, LAC educational system 7:181
and regulation 20:178
remittances 12:519–522
in skills training 22:22–23
in transportation 24:632
Investment arbitrations 12:470
Investment climate 6:271
Investment decisions
corporate finance 4:927, 4:929–930
stock market 23:487–488
Investment modes 12:580–585
Investment policies
corporate finance 4:927, 4:929
Investment process
venture capital 25:60–61
Investment theory 10:851
Investment valuation 4:930
Investments in children
and twin studies 24:705–707

Investor urbanism 18:189–191
Investors
in area and international studies 1:946–948
Invidious comparisons 22:214–215
In-virtuo exposure 8:606
Invisible cannibalism 3:86
Invisible church 10:736
Invisible colleges 7:867, 9:619–620, 21:132
Invisible hand 11:864, 22:196
and political economy 18:358
unintended consequences 24:744–747
Invisible hand-spontaneous order interpretation (ISO) 24:744–747
Invisible religion 10:215
Invisible-hand processes 12:193
Involuntariness 11:458–460
Involuntary intoxication 5:222
Involuntary learning 7:240
Involuntary memory 2:286
Involuntary minorities 1:731, 8:132–133
Involuntary part-time employment 25:833t
Involution
in Southeast Asia 23:72
Involvement behaviors
and intimacy 14:375–377
Inward-oriented militarism 15:493
IOA *see* Input–output analysis (IOA)
IOC *see* International Olympic Committee (IOC)
IOJ *see* International Organization of Journalists (IOJ)
IOLTA *see* Interest on Lawyers' Trust Accounts (IOLTA)
IOM *see* Institute of Medicine (IOM)
IOM model 18:885
Ion channels 16:704–706
in neuronal activity 12:710–715
Ionescu, Nae 7:374
Ionia: Land of Wise Men and Fair Women 14:255
Ionotropic receptors 16:752–753, 16:753f
I-optimality 17:257
optimal experimental design 17:257
IOS *see* Independent-observations decision strategy (IOS)
Iowa School of Sociology 23:857
IP *see* Intellectual property (IP); Inverse projection (IP)
IP_3 *see* Inositol triphosphate (IP_3)
IP law 12:266–272
IP theory 12:266–267
IPA *see* Independent practice association (IPA); International Phonetic Alphabet (IPA); Intimate partner abuse (IPA)
IPAD *see* Independent Payment Advisory Board (IPAD)
IPAQ *see* International Physical Activity Questionnaires (IPAQ)
IPAT (Impact, Population, Affluence and Technology) 12:716–718
IPAT (Impact, Population, Affluence, and Technology)
as a casual relationship 12:716–717
environmental spillovers 12:717
nonlinearities in environmental systems 12:717
subjectivity of impacts 12:717–718
IPAT equation 7:799
I-PAT identity
population dynamics 18:569
IPC *see* Interpersonal circumplex (IPC) model
IPC-7 *see* Inventory of Personal Characteristics (IPC-7)
IPCC *see* Intergovernmental Panel on Climate Change (IPCC)
IPCs *see* Intermediate progenitor cells (IPCs); International patent classifications (IPCs)
IPD *see* Inventory of Psychosocial Development (IPD)
IPF *see* Iterative proportional fitting (IPF)
IP-final syllables 23:719
IPIP *see* International Personality Item Pool (IPIP)
IPL *see* Inferior parietal lobule (IPL)
IPO *see* Initial public offering (IPO)
IPPF *see* International Planned Parenthood Federation (IPPF)
Iproniazid 1:778
IPRs *see* Intellectual property rights (IPRs)
IPS *see* Intraparietal sulcus (IPS); Individual Placement and Support (IPS) Model
IPSA *see* International Political Science Association (IPSA)
Ipsative scores
in forced-choice personality assessment 17:842–843
Ipsative stability
and adolescent personality 17:864–865
Ipsos MORI, London 23:750
IPT *see* Identity process theory (IPT); Integrated psychological therapy (IPT); Interpersonal psychotherapy (IPT)
IPT-G *see* Interpersonal psychotherapy for group (IPT-G)
IPTW *see* Inverse probability of treatment weighted (IPTW) estimator
IPUMS *see* Integrated Public Use Microdata Series (IPUMS)
IPUMS-International 3:294–295
IPV *see* Intimate partner violence (IPV)
IPW *see* Inverse probability weighting (IPW)
IQ *see* Intelligence quotient (IQ)
IR *see* International relations (IR)
Iran
constitutional revolution 16:362
legal system in 15:363–364
MENA 16:361–365
psychiatric care in 6:43
religious politics 16:378–379
Iranian Revolution 18:476, 20:400, 20:645
and religious politics 16:378–379
Iran–Iraq War 24:604, 25:424
Iraq
Civil Code 15:361
legal system in 15:364
military operations in 13:534
post 9/11 16:364
IRB *see* United States Institutional Review Board (IRB)
IRBs *see* Institutional Review Boards (IRBs)
IRCA *see* Immigration Reform and Control Act (IRCA)
Ireland
confederation 4:573
locally based community development in 4:381–382
Irenaeus 3:548, 20:285
IRGC *see* International Risk Governance Council (IRGC)
Irhabi 007 24:230
Irish
diaspora 6:353, 6:355
Irish Greens 10:382–383
Irish Rebellion of 1641 7:653
IRM *see* Innate releasing mechanisms (IRM)
Iron Age
Europe during 8:271–272
South Asia during 23:54
Southeast Asia during 23:66–67
Iron continent 22:795
Iron Curtain 6:811, 6:814
Iron law of oligarchy 4:268, 17:201, 18:473
cultural factors 17:203–204
elites 7:392
methodological issues and emerging directions 17:204–205
Michels' original argument 17:201
in nonbureaucratic settings 17:202
organizational environment and 17:203
organization's structure and practices and 17:202–203
Iron law of wages 18:358
Iron Triangle (South Australia) 2:243
Iron triangles (interest groups) 12:375, 12:769–773
Irony 15:290
and pragmatic disorders 18:790–792
Iroquois Confederation 4:573–574
Irrational behavior
procrastination 19:73–78
Irrational commitments 1:559
Irrational criminals 13:575
Irrational systems 13:566
Irrationalism
and collective behavior 4:157, 4:160–161
Irrationality 18:777
Aristotelian tradition *versus* Heuristics and Biases program 12:719
and bureaucracy 2:914–916
in decision making 12:720–721
framing 12:720
Public Goods Game and Costly Punishment 12:721
Ultimatum Game 12:720–721
evolutionary psychology 12:721–723
implications for human rationality 12:721–723
Darwinian responses 12:721–722
ecological rationality 12:722
informational format 12:722

Irrationality (*continued*)
preference structure 12:722
dual process accounts of human reasoning 12:722–723
problems with specific experiments 12:721
judgment 12:722
methodological individualism 15:310–312
philosophical aspects 12:719–723
probability 12:719–722
in reasoning 12:719–720
base-rate neglect 12:720
conjunction fallacy 12:719–720
utility 12:722
Irrealis forms 24:212–213
Irredentism 18:411, 18:415–416
and ethnic conflict 8:91–93
Irregular adoptions 1:137
Irregular causes, theory of 13:383–384
Irregular immigrants 11:664–665
Irregular migration 11:675
Irreversible and Catastrophic Harm Precautionary Principle 13:507
Irrigated lands
special issues in 13:235–236
Irrigation 25:438–439
Irrigation communities 12:725
Irrigation societies 12:724–726
agroecology 12:726
archaic states 12:725
civilizations 12:724–725
common pool resources 12:725
cooperation 12:725
dams 12:724
dead zones 12:726
environmental constraints 12:725
green revolution 12:726
irrigation systems 12:724
small-scale 12:724–725
management 12:725
Marx on 12:724
maya 12:725
Mesopotamia 12:724, 12:726
prisoner's dilemma 12:725
tragedy of the commons 12:725–726
Wittfogel on 12:724–725
Irritable bowel syndrome (IBS) 9:605–607
anxiety and 9:608–609
coping with 9:607
depression and 9:608–609
sexuality and 9:608
stigma and 9:607–608
stress and 9:608–609
IRRs *see* Internal rates of return (IRRs)
IRT *see* Item response theory (IRT)
Irwin, John 20:61
works, *The Felon* 20:61
IS *see* Importance sampling (IS); Index of segregation (IS)
I-S explanations 10:774
ISA *see* International Sociological Association (ISA); International Studies Association (ISA)
Isarael, J., works, *Democratic Enlightenment* 7:660–661
Isard, Walter 20:146
Isarithmic maps 24:266
ISCED *see* International Standard Classification of Education (ISCED)
Ischophonia 16:680–681
ISCT *see* Integrative social contracts theory (ISCT)
ISD *see* Instructional systems development (ISD)
ISEAS *see* Institute of Southeast Asian Studies (ISEAS)
Isegoria 6:61–62
I-self 21:433
ISG *see* Internet support group (ISG)
Isherwood, Baron, works, *The World of Goods: Towards and Anthropology of Consumption* 7:16
Ishi 1:636
ISHPES *see* International Society for the History of Physical Education and Sport (ISHPES)
Ishwerwood, Christopher 9:638
ISI *see* Import substitution industrialization (ISI) strategy
ISIC *see* International Standard Industrial Classification of all Economic Activities (ISIC)
ISIC Rev.4 (International Standard Industrial Classification of all economic activities) 5:707
Isidore, bishop of Seville 20:285
ISIP *see* International School Improvement Project (ISIP)
Isip 11:791
Isis 15:248
iSJT *see* Intercultural situational judgment test (iSJT)
ISKCON *see* International Society for Krishna Consciousness (ISKCON)
Islam 2:144–145, 2:886, 4:523, 7:210, 7:213–214, 9:73–74, 9:77, 12:363, 20:331, 20:338–342, 20:362, 20:398, 20:408, 20:410–411, 21:173–174, 24:602
in Afghanistan 12:728
in Africa 1:287, 1:289
Almoravid movement 12:737
as an Asian religion 12:727–728
as an organizing principle 12:728–729
anthropology of 20:308, 20:310–311
against apartheid 12:739
and Arab nationalism 16:271–273
Arab nations 16:376
Arabic religious texts 12:729–730
in Asia 12:727–730
assassins 24:237–238
beliefs about mental illness 19:307
and birth control 9:3
in Central Asia 3:317, 3:319–320
in China 12:728–729
Colleges of Islamic Law 12:739
colonialism and 12:739
controversy on Islamic records 12:731
conversion of Mongol elite to 12:728
countries, Jewish Diaspora 12:803–804
in East African Coast 12:738–739
in East Coast 12:738
and economic development 5:530–532
and ethnicity in Asian Societies 12:728–729
in Europe 8:258, 17:425–426
in European languages 12:727
fatwas (legal opinions) 12:735–736
forums 20:276
gender in 9:782
German use of Muslims for administrative purposes 12:739
historical context 12:731–734
historical tradition 11:61
historiography 11:61
and homosexuality 16:864
identity 11:523–524
immutability of the Shari'a or divine law 15:614
independence struggle and 12:739
Islamic resurgence among Malaysians 3:533
Islamicate 12:727
'Islamist' critique of state 12:736
Islamist movements 12:730
jurists (faqih), influence of 12:735
in Kenya 12:739
leadership in 20:410
legal schools (*madhhab*) and madrasas 12:735
Maghreb sexual minorities *see* Maghreb, sexual minorities in
in Malaysia 12:728–729
marginalization of Arabia 12:731
meaning of submission 12:727
MENA 16:365
in Middle East 12:731–736, 17:423
in Middle East and North Africa 15:354–357
and Middle East politics 16:372–373
and Middle East society 16:381–385
militant reform 12:737
and modernity 16:376, 16:378
and morality 15:808–809
Muslim juristic assumption of 12:727
mystics (Sufis) definition 12:727
and national identity, relationship between 20:339–340
in Nigeria 12:739
in North Africa 16:376
organizational forms of 20:409
political systems, role in 12:729
'purification' of Muslim religious life 12:730
Quranic verses 12:731–733
rise of Shi' ism 12:733
rituals mandated by Islamic law 12:730
role in political systems 12:729
and same-sex sexualities 20:871–872
in South Africa 12:739
in South Asia 2:76–78, 12:728–729
in Southeast Asia 12:728–729, 13:813, 23:70–71
in Spain 17:423
in Sub-Saharan Africa 12:737–740
Sufism 12:733–734, 16:377

Sunni Moslems 16:376
in Tanzania 12:739
tension between textual and lived 12:727
theoretical concerns 12:727
in theory and practice 12:734–736
in tropical Africa 12:737
vernacular literatures of 12:729–730
in West Africa 1:301–302, 12:737
Islam influence 4:762
Islamaphobia 17:430
and anti-Arab prejudice 17:430
implications of 17:430
and Orientalism 17:429–431
Islamic chronology 24:410
Islamic Church 3:707–708
Islamic Development Bank 24:604
Islamic feminism 12:744
Islamic fundamentalism 12:741–745, 20:409–410, 21:792
al-Qa'ida 12:744
Darwinism 12:742–744
feminism 12:744
general features 12:741
Herder on 12:741–742
intellectual origins 12:741–743
incipient organization 12:743–745
Muslim reformism 12:741–742
radical Islamism and society 12:743
from reformism to fundamentalism 12:742–743
jihad 12:744–745
nationalism 12:741–744
Puritanism 12:742
revolution 12:741–744
social engineering 12:741–743
takfir 12:743
vanguard 12:744
Islamic Golden Age 21:173
Islamic historiography 11:61, 11:63
Islamic history 15:360
in MENA regions 15:360
Islamic law 12:746–753, 13:801, 15:360, 15:362–363, 15:365–367
in Africa 1:272
categories of legal acts 12:752
court 12:748–752
fiqh 12:746, 12:752
Ḥadīth 12:747–749
legal history 12:746–752
bureaucratization 12:749–751
colonialism and reformation, liberalization and authoritarianism 12:751–752
consolidation and formalization 12:748–749
constitutionalism 12:751
judges 12:748
jurist-professors 12:749
legal circles and networks 12:746–747
professionalization of legal schools 12:747–748
resistance 12:752
technocratization 12:749
in Malaysia nor Indonesia 13:815–816
in Middle East 12:751
in North Africa 12:751–752
orthodox Islamic jurisprudence, methodology of 12:752–753
Qur'an 12:746–747, 12:749, 12:752
Sharíah 12:750, 12:752
Sunnah 12:746–747, 12:752
uūl al-fiqh 12:752–753
Islamic legal theory 12:751
Islamic polities 23:58–60
Islamic radicalism 20:403–405
Islamic Republic of Afghanistan *see* Afghanistan
Islamic Republic of Pakistan *see* Pakistan
Islamic Revival 15:357
Islamic Revolution of 1978 16:363
Islamic slave systems 22:49–51
Islamic social work 12:754–758
Hadith on 12:754–757
hope 12:754–756
and change 12:756
implications for practice 12:757
keeping balance 12:757
Qur'an 12:754–757
religion 12:754–755, 12:757
spirituality 12:754, 12:756–757
strengthening relationship with Allah 12:756–757
Islamic transnationalism 24:604–605
Islamism 20:623
radical 12:743
Islamist movements
identity 11:545
and women, in Middle East 16:368–369
Islamization 15:119
in Southeast Asia 23:70
of western societies 22:506
Islamkonferenz 12:239
Islamofascism 24:451–452
Islamophobia 18:412, 18:415, 25:798–799
Islamophobic discourse of conservatives 21:791–792
Island ecosystems 17:449
ISMPP *see* International Society of Medical Planning Professionals (ISMPP)
Isms 22:162–167
absolutism 1:6, 1:10
correlates of 22:165–166
examples 22:164t
future directions 22:166
historical explorations 12:761–762
and history 12:759–764
in history writing 12:763–764
and personality 22:165
rethinking 12:762–763
scholars and 12:760–761
ISNIE *see* International Society for the New Institutional Economics (ISNIE)
ISO *see* International Organization for Standardization (ISO); Invisible hand-spontaneous order interpretation (ISO)
Isocarboxazid 1:778
Isolated condition (IC) 2:837, 2:837t
Isolated geographic models 14:787–788
Isolation 19:825, 19:829, 19:831, 21:411–412
Harlow on 10:526–527
of language zone 1:820t
and LGBT aging 13:866
Isomorphic model 14:793
Isomorphic organizations 13:545–546
Isomorphism 2:915, 16:764–765, 17:298, 17:375, 21:73, 22:981
coercive 21:73
defined 2:341
mimetic 21:73
of models 2:340–341
invariance theorems for axiomatized theories 2:341
quantum decision theory 2:341
normative 21:73
and world culture 10:229, 10:231–232
Isoniazid 1:778
ISOP *see* Isotonic probabilistic (ISOP) model
Isoperformance
in structural contingency theory 23:612–613
Isoprofit curve 24:500f
Iso-strain 17:136–137
Isotonic probabilistic (ISOP) model 24:252–253
linear 24:254
Isotopes 2:600–601
Isotopic dating methods 3:565–566
radiocarbon dating (14c) 3:565–566
AMS and 3:566
calibration 3:565–566
method 3:565
Isotropic scaling 19:80
Is/ought separation 16:755
ISPs *see* Intensive supervision programs (ISPs); Internet service providers (ISPs)
ISR *see* Institute for Social Research (ISR); Intentional self-regulation (ISR)
Israel 9:76
Jewish Diaspora 12:801, 12:803–810
MENA 16:361, 16:363, 16:365
peace building and music in 25:845
Israel, J.I., works, *A Revolution of the Mind* 7:662
Israel, Joachim, works, *The Context of Social Psychology. A Critical Assessment* 22:534
Israeli Center for Victims of Cults 16:808
Israeli Jews, religiosity of 12:873
Israeli kibbutz 4:267–268
Israeli *kibbutzim* 11:726
Israelite religion 12:869
Israelite society, early 12:869–870
Israel's Senior Citizens' Rights Act 13:425
ISS *see* Institute of Social Science (ISS)
Issac, B., works, *The Invention of Racism in Classical Antiquity* 19:796–797
ISSC *see* International Social Science Council (ISSC)
ISSOL *see* Interpol Standard Set of Loci (ISSOL)

ISSP *see* International Social Survey Program (ISSP)
Issue attention cycle 1:353–354, 3:841–842, 3:842f
Issue definition
  political agendas 1:364
Issue evolution 12:765–768
  critical elections 12:765
  electoral realignment 12:765–767
  illustrations 12:766–767
    abortion 12:767
    moral and cultural issues 12:767
    new religious divide 12:767
    racial issues 12:766–767
  mass public 12:766–767
  partisan activists 12:766–767
  partisan elites 12:766–767
  political issue 12:765–767
  sequence of 12:766
  standard model of partisan change 12:765
    criticisms 12:765
  structure of 12:766
Issue networks 12:769–773
  advocacy coalitions 12:772
  analysis and change 12:770–771
  causal mechanisms 12:771–772
  constant expansion 12:772
  decentered analysis 12:772
  dialectical analysis 12:771–772
  early developments 12:769
  increasing theoretical leverage of social network analysis 12:771–772
  interorganizational analyses 12:770
  policy outcomes 12:770–772
  typologies 12:769–770
Issue public hypothesis 14:713–714
Issue publics 2:195
ISTDP *see* Intensive short-term dynamic psychotherapy (ISTDP)
Istrians
  Italian *versus* Croatian 16:214–215
ISW *see* International social work (ISW)
IT *see* Information technology (IT) in education
Italian Code of Criminal Procedure (1988) 5:229
Italian Communist Party (PCI) 2:714, 3:710, 14:659
Italian fascism 16:257, 16:261, 16:284, 18:435, 22:444–445
  fascist dictatorship 16:257
  origins 16:257
Italian fascist movement 22:444
Italian Renaissance 4:752–753
Italy
  authority theories 2:270
  clinical psychology, development 3:874
  cultural physiognomies 20:428–429
  freedom 20:426
  heritage preservation in 5:462
  immigration policy 11:662–665
  laicization in 13:218
  modernity in fifteenth-century 20:425
  organized crime in 5:174
  psychoanalysis 3:874
  relationship between arts and politics in 5:460–461
  Renaissance culture 20:426
  Vitruvius' theory of architecture 20:427
  works of art 20:425
ITC *see* Inferior temporal cortex (ITC); Intercalated (ITC) nuclei
Item, context, ensemble information *see* ICE (item, context, ensemble information) model
Item analysis
  item response theory (IRT) 19:436
Item bias 5:319, 5:323, 7:929–930
Item characteristic curve (ICC) 19:420
Item development
  cross-cultural research 7:931
Item effect 9:836
Item equivalence 5:319
Item factor analysis models 19:438
Item information function (IIF) 19:420
Item nonresponses 13:385, 13:389, 15:602
  data adjustments 6:127–128
  in sexuality studies 15:315
Item response model 19:439
Item response modeling
  of forced-choice personality assessment 17:843
Item response theory (IRT) 11:752–753, 19:420, 19:423, 19:436–439, 20:248, 20:250, 24:250, 19:420f
  applications 8:701
  approaches based on 17:837
  Bayesian statistics 8:699–701
  computerized test construction 4:510
  framework 4:510, 4:513
  history 19:436
  item analysis 19:436
  models for dichotomous responses 19:436–437
  models for polytomous responses 19:437
  multidimensional item response theory 19:438
  noncompensatory multidimensional models 19:438
  nonparametric 8:701
  parameter estimation 19:438–439
  parametric 8:699–701
  personality psychology 8:704–705
  scoring tests 19:439
Item-and-arrangement 15:823
Item-and-process 15:823
ITERATE *see* International Terrorism: Attributes of Terrorist Events (ITERATE) database
Iterated dominance
  in game theory 9:585–586, 9:588
Iterated learning 8:357
  Bayesian analysis of 8:360
  cultural transmission 8:357–360
Iterated learning model (ILM) 8:359
Iterated Prisoner's Dilemma (two-person) 18:934–936
  finitely iterated 18:934–935
  indefinitely repeated 18:935–936
Iterative account, of mutual beliefs 21:877–880
Iterative fitting procedures 9:626
Iterative proportional fitting (IPF) 18:589
Iterative risk management
  climate change, economics of 3:863
Iterative simulation techniques
  for multiple imputation 16:90
Iteso 6:775
ITG *see* Inferior temporal gyrus (ITG)
I–thou dialogs 8:518
I–thou relationship 8:518
ITRC *see* Identity Theft Resource Center (ITRC)
ITS *see* Intelligent transportation systems (ITS)
ITSs *see* Intelligent tutoring systems (ITSs)
Itsue, Takamure 19:748–749
ITT *see* Ill-Treatment and Torture (ITT) measures; Intention-to-treat (ITT) analysis; Intention-to-treat (ITT) effect
ITU *see* International Telecommunications Union (ITU); International Typographers Union (ITU)
IUAES *see* International Union of Anthropological and Ethnological Sciences (IUAES)
IUCN *see* International Union for the Conservation of Nature and Natural Resources (IUCN)
IUD *see* Inter-uterine device (IUD); Intrauterine device (IUD)
Iudicanda 12:943, 12:945, 12:947–948
IUPsyS *see* International Union of Psychological Science (IUPsyS)
IUS *see* Inter-University Seminar on Armed Forces and Society (IUS)
Ius civile 7:924, 13:799
Ius commune 13:915
Ius gentium 12:364, 13:799, 13:915, 25:742
Ius honorarium 7:924
Ius naturale 13:779
Iusgentium 3:670
IUSSP *see* International Union for the Scientific Study of Population (IUSSP)
Ivanov, Aleksandr, works, *Two Models* (painting) 21:742–743
Ivanov, Viacheslav 21:741–742
  works, *Cors Ardens* 21:741
Ivens, Joris, works, *The Spanish Earth* 6:623
IVF *see* In vitro fertilization (IVF)
Ivison, Duncan
  Postcolonial Liberalism 18:650
Ivnev, Riurik 21:742–743
  works, *Self-Immolation* 21:741
Ivory Park EcoCity 6:835–836
IVR *see* Interactive voice response (IVR)
IVs *see* Instrumental variables (IVs)
Ivy League schools 1:635
Iwakura Mission (1871) 24:583
Iwi 14:505–506, 14:508–510
IWM *see* Internal working model (IWM)

IWRM *see* Integrated water resources management (IWRM)
Izinyangas 10:568
Izzat 1:191, 11:185–187, 11:189, 16:863

J

J. Walter Thompson (JWT) agency 1:191
Jackal, Robert, works, *Moral Mazes* 2:914
Jackknifing 23:416
Jacklin, C.N., works, *The Psychology of Sex Differences* 21:693
Jacks, G.V., works, *The Rape of the Earth* 7:752
Jackson, Andrew 8:84
Jackson, J.L., on race in North America 17:23
Jackson, John Hughlings 9:413–414, 12:774–775, 15:967
  biography of 12:774
  on clinical deficits associated with lesions 12:775
  concept of epileptogenic discharges 12:775
  concept of evolution of the nervous system 12:775
  concept of localized cerebral discharges 12:774–775
  on cortical neurophysiology 12:775
  on effect of lesions on function 12:775
  on epileptic discharges 12:775
  on negative and positive elements 12:775
  on observer of clinical phenomena 12:774
  on skill in precise clinical observation 12:775
  on topographical structure–function relationships 12:774–775
Jackson, Peter. A.
  on sexual citizenship 21:731
  works
    Maps of Meaning 5:409–410
Jacksonian principle, prenatal/infant development 18:852
Jacob, Ian 19:527
Jacob Wetterling Crimes Against Children and Sexually Violent Offender Registration Program 20:434
Jacobi, Mary Putnam 25:623
Jacobi algorithm 22:483
Jacobian elliptic sine functions 14:795
Jacobian matrices 3:779
  motor control 3:779
Jacobinism 6:372
Jacobs, Jane 12:776–781, 18:202–203
  assessment 12:779–780
  biography 12:776–777
  on cities 12:777
  on diversity 12:777
  on neighborhoods 12:777–778
  on social sciences 12:778
  works 12:777–779
    Cities and the Wealth of Nations 12:778
    Dark Age Ahead 12:778
    The Death and Life of Great American Cities 12:776, 24:943
    The Economy of Cities 12:778
    The Nature of Economies 12:778
    The Question of Separatism 12:778
    Systems of Survival 12:778
Jacobs, Jerry, on epistemic cultures 7:878
Jacobsen, Carlyle 23:18
Jacobsohn, Gary 4:705
Jacoby's process dissociation 18:836
Jahoda, M., works, *Die Arbeitslosen von Marienthal* 25:672
Jahoda, Marie 13:637–638
Jahrhundert 25:290–291
Jail diversion 4:373–375
Jail populations
  overlooked 20:32–33
    importance of research on jails 20:32–33
  as unexamined populations 20:31–36
  in United States 20:31–32
Jailbreaking 13:523
Jain legal traditions 23:60
Jainism 20:408
Jakobson, Roman 2:358, 6:504, 12:782–785, 16:680, 18:796, 20:651
  aphasia studies of 12:782, 15:291
  on child language 12:782
  on functional view of sound 12:782–783
  on functions of language 12:783
  on grammar 12:784
  on invariance 12:783
  legacy in linguistics 12:784–785
    influence on linguistics 12:784–785
    influence on other fields 12:785
  on metaphor 12:783–784
  on metonymy 12:783–784
  on phonological disorders` 18:48–49
  on relational structure 12:783
  on structural phonology 12:782
  on theory of distinctive features 12:782–783
  works
    Child Language, Aphasia and Phonological Universals 12:782
    Linguistics and Poetics 12:783
    On Language 12:782
    Poetry of Grammar and Grammar of Poetry 12:785
    Selected Writings 12:782
    Shifters 12:784
    Two Aspects of Language 12:783
Jama'at-i-Islami 4:264
Jamaica 1:638
  Maroons 14:593
  psychiatric care in 6:43
Jamaican Creole 18:137
Jamal, V., works, *Africa Misunderstood: Or Whatever Happened to the Rural-Urban Gap?* 24:969–970
James, Deborah 2:871–872
James, Paul 16:286
James, William 1:93–96, 3:878, 6:690, 7:568, 10:276, 12:8–9, 12:244–245, 12:340–341, 12:786–790, 18:435, 18:795, 19:64–65, 19:413, 20:354, 21:460, 22:319, 22:321, 25:560
  on clinical psychology 24:525
  on consciousness of self 12:787–788
  contributions to psychology and philosophy 12:787
  on family background and early life 12:786–787
  on functionalism 12:787
  on imperialism and War 12:789–790
  on introspective psychology 12:787
  as philosopher 12:788–789
  on pragmatism 12:788–789, 18:804, 18:808
  on principal psychological method 12:787–788
  on productive life 12:786–787
  on radical empiricism 12:789
  on religion and psychic research 12:789
  significance in modern times 12:790
  on theory of emotion 12:788
  on unconscious 24:723
  on voluntary movements 12:788
  works
    A Pluralistic Universe 18:232
    The Principles of Psychology 7:591, 12:787, 20:389
    Varieties of Religious Experience 20:354, 25:560
James (Gospel of Mark) 3:546
James–Lange theory 8:868–869, 12:788, 8:869f
Jameson, Fredric 11:549
Janapadas 23:54–55
Janelli, Dawnhee Yim 2:72–73
Janelli, Roger L. 2:72–73
Janet, Pierre 12:791–793
  on associations 12:791
  concept of psychological force 12:791
  on distinction between automatic and synthetic mental activity 12:791
  early and late period 12:791
  on hypnotism and post-hypnotic suggestions 12:791
  hypnotism studies of 19:502
  influence in current times 12:792
  on obsessive-compulsive rituals 2:427
  on psychological analysis 12:792
  on psychology in terms of conduct 12:792
  on synthesis 12:791
Janis, I.L., groupthink research 24:710–711
Janowitz, Morris 20:745
Jansenism 12:387–388
Janssen, Susanne, on popular music 3:800
Janus 16:235
Janus-faced institution, of ethnic closure and control 10:122–123
Janus-like brokers 2:870
Japan 5:454, 11:861, 14:270, 21:368–369
  Ainu people 6:785
  artisans 1:868
  Christianity in 3:532
  climatic variability 3:868
  colonialism in Korea 2:71–72
  communes 4:268–269
  corporate governance 4:936
  cultural capital research in 5:379
  cultural policies in 5:454–456
    creative industries 5:457

Japan (*continued*)
flow of culture 5:457–458
economy of 6:790–791, 11:857
education in 7:145–146
eugenics in 8:235
fertility transition in 9:55–59
historiography development in 11:49–50
immigration policy 11:664–665
legal history of 6:778–782
Maruyama Masao 14:635–641
Meiji Restoration, engagement of Western sociology 24:583–584
Meiji Revolution *see* Meiji Revolution
modern, colonial modernity, imperialist economy, national gender policy in 6:794–795
multicultural education in 16:19
multinational corporations in 16:76–78
narrative life story in 16:212
nationalism in 16:275–278
nostalgia in 17:35–36
post-tsunami 22:570–575
direction and ownership sense 22:574
discipline and security 22:572–573
diverse disaster experiences 22:570
economy of scale 22:573–574
Fukushima 22:570, 22:572–573
Kamiashi 22:572
local resilience 22:571–572
memories, resilience from 22:574–575
resilience development 22:573–574
speed of development 22:574
spontaneous resilience 22:573
Tohoku 22:570–571
religious forms in 20:370
Shinto 20:368
rice agriculture 6:785
sexual harassment in 21:725
sociocultural aspects of 12:794–797
Agrarian Japan 12:794–795
anthropology 12:794
culture/history at local and global confluence 12:795–797
educational system 12:795–796
gender 12:795
health care 12:797
majority and minorities 12:794–795
nonlocalized culture 12:795
religion 12:796
self-identities 12:794–795
sports 12:797
youth culture 12:797
value engineering in 14:468–469
Japan Commercial Arbitration Association 12:470
Japan Expo 5:457
Japan Society for the Promotion of Science 22:655
Japanese Association of Racial Hygiene 8:235
Japanese civilization
Eisenstadt's contribution 7:323
Japanese connection, organizational culture 17:347–348
Japanese eugenics 8:231
Japanese Immigration law (1952) 11:664
Japanese iMode system 15:630
Japanese language 20:1, 23:889
cognitive and neuropsychological studies
reading 20:3
writing 20:3
Kana scripts and reading 20:2–3
kanji script and reading 25:773–774
writing system 20:1, 25:778
Japanese militarism 15:490, 15:492–493
Japanese politics
in Cold War period 6:798–799
in post–Cold War period 6:802–803
Japanese syllables 23:715
Jardé law 24:285
Jargon aphasia 18:56–57
recovery from 18:56
Jarvie, Ian 9:671–672
Jarvik, Lissy 25:625–626
Jasanoff, Sheila, on epistemic cultures 7:879
Jaspal, R.
on nonheterosexual British Asians
friendship 16:865
social representations 16:862
Jasper, Herbert 10:732
Jaspers, Karl 25:463
on alienation 1:551, 1:554
works
General Psychopathology 19:492
Reason and Existence 1:554
Vom Ursprung und Ziel der Geschichte 20:388, 21:391
Jastrow, Joseph 8:567
Jataka (Rebirth) stories 2:888, 2:890
Jati 2:78–79
Jay, Martin 25:212
Al-Jazeera 16:776–777, 16:779
Jazz 3:803
Jazz Age 25:814
JCA *see* Jewish Colonization Association (JCA)
JCE *see* Joint criminal enterprise (JCE)
JDC *see* Job Demand-Control-Support (JDC) model
JDM *see* Judgment and decision making (JDM)
JD-R *see* Job demand–resource (JD-R) model
Jean, Piaget, on morality 15:806
Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act 21:662–663
Jeans, Sir James Hopwood, and Fisher 9:222
Jefferson, Gail 4:851
Jefferson, Thomas 12:897–898, 14:734
Jeffreys, Sir Harold 12:798–800, 17:662
on axiom system 12:799
on brilliant formulation 12:800
contributions to philosophy and practice of science 12:800
on mathematical physics and cosmogony 12:798
method of DNA analysis 9:955–956
on probability of a hypothesis 12:799
on reasonable degree of confidence, measure of 12:799
on scientific method 12:798–799
on statistical estimation 12:799
works
Cartesian Tensors and Operational Methods in Mathematical Physics 12:798
The Earth: Its Origin, History and Physical Constitution 12:798
Theory of Probability 12:798, 12:800
Jegnenet 3:442–443
Jehlen, M., works, *Ideology and Classic American Literature* 1:620
Jehovah's Witnesses 15:530
Jellinek, E. Morton 1:501
Jellinek, Georg 10:874–875
Jelloun, Tahar Ben
works
"""Défendre la diversité culturelle au Maghreb""&rdquo 21:753
L'enfant de sable 21:753
Les yeux baissés 21:753
Jena Model of Social Change and Human Development 22:151–152
Jenkins, Alan, works, *The Geography of Peace and War* 25:374
Jenkins, Henry
on online consumers 12:591
on television 25:81–83
works
Convergence Culture: Where Old and New Media Collide 2:218
Textual Poachers 2:217
Jenkins, Philip 15:798
Jennings, Herbert Spencer 8:221–222, 8:231–232
Jennings, Humphrey, works, *A Diary for Timothy* 6:621
Jensen, Arthur 8:130–131
Jensen, Joli, on media audiences 25:82–83
Jensen, Karen, on epistemic cultures 7:877
Jernudd, B., works, *Can Language Be Planned?* 13:354
Jerusalem, Wilhelm 19:391
Jespersen, Otto 18:796
Jespersen's cycle 16:412–413
Jessup, Philip 12:508
Jesuitismus 12:762
Jesus Army in England 16:806
Jesus movements 3:545–547
Jet lag, phenomenon of 15:638
Jeune Afrique (Young Africa) 15:135
Jevons, William Stanley 7:914, 14:644, 21:304, 25:7
works
Theory of Political Economy 25:9–10
Jew/Greek dichotomy 3:237
Jewish 15:135–136
Jewish American 25:346–347
Jewish Christianity 3:237
Jewish chronology 24:410
Jewish civilization, Eisenstadt's contribution 7:323
Jewish Colonization Association (JCA) 2:148

Jewish communities 24:602–603
genetic research on 16:761
life experiences of 12:876
future prospects 12:877
Jewish cultural production 12:875
Jewish Diaspora 12:801–810
assimilation 12:801
birthrates 12:801, 12:803–804, 12:807–810
conversions 12:801, 12:803, 12:808
database and conceptualization 12:801–802
death rates 12:803, 12:807–810
in Eastern Europe 12:802–806, 12:808, 12:810
family 12:801, 12:803–804, 12:808–809
intermarriage 12:809–810, 12:808t
international migration 12:805–807, 12:806f–807f, 12:806t
in Islamic countries 12:803–804
Israel and 12:801, 12:803–810
majority status 12:802
in Middle East 12:802–806
minority status 12:802
in Palestine 12:803–806
population growth (1700–1939) 12:802–803, 12:802f
population trends (1939–2013) 12:803–805, 12:804t–805t
in Shoah 12:803–806
in United States 12:803–810
urbanization 12:801
vital statistics and population composition 12:807–810, 12:808t–809t
Western Europe 12:803–806
Jewish feminists 12:875–876
communal life of 12:876
Orthodox Jewish feminist movement 12:876
Jewish histories, Eastern European studies 6:814
Jewish identity 20:398
patterns of 12:870–871
decline of antireligious or nonreligious 12:872
denationalized 12:871–872
dereligionized 12:871–872
difference between American Jews and Christians 12:872–873
historical identity 12:872
and levels of religious observance 12:873
rites of passage 12:872
secularist Zionism 12:871–872
Jewish immigrants 16:20
Jewish martyr 3:546
Jewish National Fund (JNF) 16:859
Jewish organizations 20:280
Jewish radicalism 20:404
Jewish Studies scholarship 12:875
Jewish tradition
historiography and historical thought 11:67–71
Biblical traditions of history writing 11:67
Hebraism 11:67, 11:69–71
Judaism 11:67–69, 11:71
major issues in modern Jewish historiography 11:70–71
modern Jewish scholarship and critical historiography 11:68
national identity 11:68
nations 11:69, 11:71
political agendas and writing of Jewish history 11:68–69
in Rabbinic literature and medieval sources 11:67–68
sixteenth-century Jewish historiography as response to catastrophe 11:68
Talmud 11:67–68
twentieth-century trends 11:70
Jewish Women International 20:256
Jewishness, Butler's work on 3:56–57, 3:59
Jews 20:328–329
and Arabs 15:137
diaspora 6:353, 6:355
Memmi on 15:135
and racism 19:852, 19:854–855
Jex-Blake, Sophia 25:623
Jhally, Sut, works, *Enlightened Racism* 2:217
Ji (China) 24:870–871
Jia jiehun 21:735
Jiang Zemin 3:518
JIC *see* Joint industry committee (JIC)
Jien, works, Gukanshô 11:50–51
JIF *see* Journal Impact Factor (JIF)
Jigsaw, cooperative learning 4:884
Jihads 9:468, 12:744–745, 16:266–267
Jim Crow laws 3:691–693, 5:267, 11:911, 17:23, 19:808–809
Jinruigaku Zasshi 2:71
JLD *see* Jyväskylä; Longitudinal study of Dyslexia (JLD)
JMP *see* Juhn–Murphy–Pierce (JMP) decomposition
JMP (computer program) 23:379
JNF *see* Jewish National Fund (JNF)
Joan Ganz Cooney Center 7:252
Joas, Hans 11:175, 12:244–245
Job analysis 11:880, 12:811–812, 17:936
data collection 12:813–814
interview methods 12:813
observation methods 12:813–814
intended applications 12:812–813
method users 12:813
methodological standards 12:814
outlook 12:814
person-oriented analysis 12:812–813
task analysis 12:811–812
theories 12:813
unit of analysis 12:812
Job assignment, promotions 19:137–138
Job assignment model 15:598
Job attitudes, and organizational citizenship behavior 17:317–318
Job availability 23:157
Job change and wage growth 19:135
Job characteristics model 12:834, 17:148, 25:337–338, 25:537, 25:718–719
Job competition model 15:597–598
Job control 25:537–539
Job creation 6:269, 21:334–335
and entrepreneurship 7:675, 7:679
and labor market policy 13:184, 13:187
policy changes for creating 6:271–272
Job creation condition 21:335
Job demand 17:147–148, 25:537–538
Job Demand-Control-Support (JDC) model 17:148
Job demand–job control model 25:538, 25:718–719
Job demand–resource (JD-R) model 25:734–735
of well-being and burnout 25:534
Job Descriptive Index 12:833–834
Job design 12:816–818
Job destruction 6:271, 21:337–338
idiosyncratic shocks 21:338
technological progress and obsolescence 21:338
Job displacement insurance 12:819
across economies 12:822–823
adequacy 12:823–825
aging and the long-tenured displaced worker 12:830
consumption losses among involuntarily displaced workers 12:824t
consumption smoothing 12:819–820, 12:823, 12:825
credit markets 12:829
design 12:828–829
early retirement 12:820, 12:830
earnings losses 12:819–821, 12:823, 12:825
financial security 12:825
firing cost issues 12:827–828
job search and offer acceptance incentives 12:826–827
active labor market policies 12:827
sanctions 12:827
losses 12:820–822, 12:822t
market 12:825–826
demand 12:825–826
supply 12:826
moral hazard 12:819, 12:825–827, 12:829
saving and borrowing strategies 12:829–830
service-conditioned severance benefits 12:828
severance pay 12:819–820, 12:822–823, 12:825–830, 12:828f
small-loss events 12:825
UI benefits over the duration of unemployment 12:828–829
wage insurance 12:819, 12:825–826
Job enrichment 11:301
Job evaluation 12:816–818
and equal pay 12:817
future directions 12:818
hierarchy 12:816, 12:818
human resources management 12:817
industrial psychology 12:816–817
labor market 12:816
and law 12:817–818
remuneration 12:816
research 12:816–817

Job evaluation (*continued*)
rise of 12:816
wages 12:816
Job finding 20:64
Job location 25:337
Job motivation, and organizational citizenship behavior 17:320
Job performance 1:420–421, 4:909, 7:446–447, 17:613–614
and burnout 2:931
and compensation 17:612–613
group performance 17:612, 17:614–616
individual performance 17:612–614
and organizational citizenship behavior 17:317–318, 17:320
and personality traits 17:921–923
reward systems 17:611–613, 17:615–616
Job polarization 22:194–195
Job resources 17:147–148, 25:538–539
Job satisfaction 1:423–425, 4:905–908, 7:508, 12:833–838, 25:660
absenteeism 12:835
affect 12:833–834, 12:836–837
attitude 12:833–837
and burnout 2:929–930
consequences of different levels of 12:834–835
counterproductive work behavior 12:835
definition 12:833
discrepancy theory 12:834, 12:836–837
as driver of labor market outcomes 10:519
expectancy theory 12:834
history of research on 12:834
job characteristics model 12:834
Job Descriptive Index 12:833–834
and job performance 7:508
measurement 12:833–834
Minnesota Satisfaction Questionnaire 12:833–834
organizational citizenship behavior 12:835
and organizational citizenship behavior 17:317–318, 17:320
performance 12:834–835, 12:837
personality 12:835–837
predictors of 12:835–836
promotions 19:136
reference group 20:72–73
social capital 22:183
social information processing 12:834, 12:837
theories of 12:836–837
turnover 12:835
Job search 12:826–827, 21:333–334
in ethnic enclaves or primary market 8:110–111
Job security 7:572–574
Job simplification 11:301
Job skills, and urban poverty 24:916
Job strain 22:933, 23:548
occupational stress management 17:148
Job stratification, artists and 2:52–53
Job stress
coping with
see also Coping
and disease 17:136–137
related outcomes 17:137–138
Job tenure 1:420–425
Jobbik Magyarországért Mozgalom (Hungarian political party) 18:410–413, 18:415, 18:610–611
Job-embedded education 7:122
Joblessness 12:141–142, 12:144
Job-oriented analysis 12:812
Job–person fit 2:931
Job-related training or education 7:122
within OECD nations 7:122
Job-related well-being, and burnout 25:537–540
bullying 25:538
consequences 25:539
exhaustion 25:537
intervention 25:539
intervention programs 25:539
job control 25:537–539
job demands 25:537–538
performance 25:539
personality 25:538–539
predictors 25:538–539
self-efficacy 25:539
strain 25:537–538
stress 25:539
stressors 25:537–539
theoretical models 25:537–538
work engagement 25:537–539
Jobs
green 10:372
and powerlessness 1:547
and pro-environmental behaviors 19:84
JOBS Act *see* Jumpstart our Business Startups Act (JOBS Act)
JOBS project 24:736
Job-security legislation 24:742
JOBSTART 8:532
Jochi University 12:795–796
Joel Beinin, works, *Workers on the Nile* 16:384
Joerges, Bernward, on institutions 17:392
Johannesburg Pride 21:747
Johannesburg (Earth) Summit (2002) 23:797
Johannesburg's spatial transition 24:947f
Johannsen, Wilhelm 10:42
John (Gospel of Mark) 3:546
John Paul II, Pope
works
Centesimus Annus 20:260
Laborem exercens (*On Human Work*) 20:260
John XXII, Pope 3:645
John-Fisk, Henna 12:563–564
Johns Hopkins University 18:467, 24:762–763
Department of Historical and Political Science 18:467
Johnson, Lyndon 11:911, 18:461, 20:540
Johnson, Marsha P. 24:552
Johnson, Michael P., on categories of intimate partner violence 12:685–686
Johnson, Philip 5:930
Johnson, Samuel 15:246–247
works, *A Dictionary of the English Language* 7:584, 22:887
Johnson v. McIntosh (1823) 11:796
Johnson–Reed Act 8:226
Johnson's law (psychophysics) 19:474
Johnston, Barbara Rose 1:225–226
Johnston, John Black 13:391
Johnston, L.D., works, *Monitoring the Future* (MTF) research 5:160–161
Join-count test 23:107
Joiner's theory 23:683
Joint actions 12:839–843, 18:798
Bratman on 12:840
cooperation in broad sense 12:839
individualism and holism 12:843
inferences from statements 12:839–840
joint commitment 12:841
derived 12:841–842
inferences 12:842
large populations 12:842
obligations and entitlements 12:842
nature and role off agreements 12:842–843
obligations and entitlements in 12:839
type 12:840
personal we-intentions 12:840–841
plural subject theory of 12:842
and related phenomena 12:842
Searle on 12:840–841
by selves 21:451–452
social groups 12:842
states of participants involved in 12:839
in terms of meshing personal intentions 12:840
Tuomela on 12:842–843
Joint angles 3:779
Joint attention 4:300, 10:272
of humans 11:254–255
during infancy and early childhood 12:844–850
autism spectrum disorder 12:846
blindness 12:847–848
in children with special needs 12:846–848
deafness 12:848
in diverse cultures 12:848–849
Down syndrome 12:846–847
gestures 12:844–849
language 12:844–849
parent–child interactions 12:845
play 12:845–849
social cognition 12:845–846, 12:849
social emotional understanding 12:845, 12:849
typically developing children 12:844–846
Williams syndrome 12:847
and intentionality 12:347–349
and Theory of Mind development 24:269
Joint commitment 12:841
derived 12:841–842
inferences 12:842
large populations 12:842
obligations and entitlements 12:842

Joint consumption technology, and artists 2:53
Joint coping 11:594
Joint Council on Economics Education
  Framework for Teaching Basic Concepts 22:729
Joint criminal enterprise (JCE) 25:384
Joint Cross-Impact Matrix 17:762
Joint design approach, organizational ergonomics 11:303
Joint distribution 16:117
Joint evaluation, in social networks 9:7
Joint industry committee (JIC) 2:214
Joint POW/MIA accounting command (JPAC) 9:336
Joint production, groups 10:437
Joint structure of choice probabilities and response times 5:901
Jokes
  and metaphors 15:292
  and pragmatic disorders 18:788, 18:790–792
Joking 11:416
  and age 11:418
  aggressive 11:417
  domains of 11:417–418
  and gender 11:417–418
  hateful 11:417
  militant 11:417
  psychological study of 11:416
JOL *see* Judgments of learning (JOL)
Jomon 12:794
Jones, Bryan D., on model of agenda-setting 1:359
Jones, Frank L., on national identity 16:312
Jones, Jim 3:356
Jones, Rufus 25:560
Jones v. City of Los Angeles 25:114
Jordan, legal system in 15:364
Jordan, Michael 23:315–316
Jordanowa, Ludmilla 2:729–730
Joseon Sahoe 24:586
Josephus, Flavius 11:67
Joule, James Prescott 21:226
Journal articles, and anthropological writing 1:740
Journal for Ethnopsychology and Linguistics 25:792–793
Journal Impact Factor (JIF) 12:864–865
Journal of Agrarian Change 1:456–457
Journal of Animal Behavior 25:449
Journal of Anthropological Research 11:362
Journal of Applied Social Science (*JASS*) 1:859
Journal of Archaeological Science 7:696
Journal of Contemporary History 4:766
Journal of Cross-Cultural Psychology 5:311
Journal of Economic and Business History 3:21
Journal of Economic Geography 7:35
Journal of Educational Psychology 15:896
Journal of Empirical Legal Studies (*JELS*) 7:554–556, 13:457
Journal of Environmental Education 7:733
Journal of Experimental Psychology 25:449
Journal of Genocide Research 10:7
Journal of Geographical Systems 10:60
Journal of Global History 11:45
Journal of Historical Geography 11:17–18
Journal of Legal Pluralism 13:758
Journal of Legal Studies (*JLS*) 13:456
Journal of Peasant Studies 1:457, 17:671
Journal of Policy Analysis and Management 19:576
Journal of Political and Military Sociology (*JPMS*) 15:515
Journal of Political Economy 11:438, 23:457, 23:459
Journal of Population Economics 11:226
Journal of Pragmatics 8:351
Journal of Regional Science 20:147
Journal of Religion and Health 19:311
Journal of Scandinavian Studies in Criminology and Crime Prevention 5:240
Journal of Sexology 21:846
Journal of Speculative Philosophy 10:740
Journal of Statistical Software 23:380
Journal of the American Society for Information Science and Technology (*JASIST*) 21:323
Journal of the American Statistical Association 24:345
Journal of the Plague Year 18:176
Journal of Theoretical Biology 22:900
Journal of Time Series Analysis 24:345
Journal of Transport Geography 3:495
Journal of Women's History 9:770
Journal of World History 11:45
Journal of World-Systems Research 14:649
Journal publishers as gatekeepers 20:511–512
Journalism 3:178–179, 11:120, 12:851–861, 16:795
  analysis, and hegemony 10:744–745
  blogging 12:851, 12:855
  case example 3:178
  citizen journalism 12:860
  climate change 3:855–856
  communism 12:851, 12:853–854
  current developments 12:854–855
    commercialization 12:854
    democratization 12:855
    economic crisis of news 12:855
    emergence of new forms of partisanship and decline of objectivity 12:855
    internationalization 12:855
    new technologies 12:855
  differentiation 12:853
  education of journalists 12:858–859
  ethics 15:42–43, 15:45–46
  historical evolution 12:851–853
  International Federation of Journalists 12:857, 12:859–861
  Internet 12:851, 12:855
  journalistic ethics 12:858–859
  market 12:851, 12:853–854
  mass media 12:857–861
  nationalism 12:851
  nonprofessionalism 12:860
  online journalism 12:857–860
  political parties 12:851–852, 12:854
  profession 12:858–860
  professionalism 12:852–854
  professionalization 12:852–854
  and the state 12:853
  UNI Global Union 12:860–861
  varieties of 12:853–854
  Weber on 12:852
Journalism education, and communications research 4:280–281
Journalistic culture 12:857
Journalistic objectivity 3:612–613
Journalistic professionalism, culture of 12:854
Journalists 12:851, 12:857–861
  privilege, to shield news sources 9:396
Journals 11:121–122, 12:862–868, 22:661–663, 22:665
  anthropology 1:756
  bibliometrics 12:865
  definition 12:862–863
    core functions 12:862
    critiques 12:863
    history 12:862
    new developments 12:863–864
    peer review 12:863–864
    publishing cycle 12:862–863
    types 12:863
  educational research 7:282–284, 7:284f, 7:284t
  and ethics 8:58
  in global marketplace 12:865–866
    access of researchers 12:866–867
    journal and articles numbers and trends 12:865
    market size and economics 12:866
    in scientific output 12:866
  mathematical psychology 14:805
  new developments in scholarly communication 12:867–868
    data-intensive science 12:867
    identity and trust 12:868
    mobile access and apps 12:867–868
    new forms of electronic articles 12:867
    publishing platforms 12:867
    semantic web and linked data 12:867
    social media 12:867
    text and data mining 12:867
  online publishing 12:862, 12:865–866
  open access 12:866–867
    copyright issues 12:867
    Gold Open Access 12:866
    Green Open Access 12:866
    impacts on use and citation 12:867
    system-wide perspectives 12:866–867
    variants 12:866
  peer review 12:863–864
    critiques 12:863
    new developments 12:863–864
    types 12:863
  professional associations 22:677
  publishing ethics 12:864
    Committee on Publication Ethics 12:864
    growth in article retractions 12:864
  in SBS 8:53
  and scholarly communication 12:862
  sport history 23:308

Journals (*continued*)
using 12:864–865
authors and readers 12:864
author's behavior, perceptions, and attitudes 12:865
citations and impact factor 12:864–865
disciplinary differences 12:865
long-term preservation 12:865
reading patterns 12:864
for women's studies 9:739
Journey, death as 5:871–872
Journey of Hope program 15:217
Journeyman 2:47
Journey-to-work 23:153–154
Jowett, B. 14:620–621
Joyriding 19:162
JPA *see* Juvenile Protection Association (JPA)
JPAC *see* Joint POW/MIA accounting command (JPAC)
JSM (John Stuart Mill) method of plausible reasoning 3:249
JSP *see* Jurisprudence and Social Policy (JSP) program
*j*th latent continuous indicator of latent variable 11:752
*j*th observation 5:908
*j*th pattern class 23:411–412
Judaism 11:67–69, 11:71, 11:153, 12:869–874, 15:614, 20:362, 20:398, 20:408, 21:173–174, 24:602
acculturated Jews 12:871
acculturated Western Jews 12:871
American and Israeli Jews attributes 12:872–873
anomie in 1:718
anthropology of 20:308, 20:310–311
beliefs about mental illness 19:308
boundaries of citizenship 12:870–871
centralization of sacrificial worship in Jerusalem 12:869–870
changes within synagogue services 12:871
in Eastern Europe 12:871
and gender 12:875–878
communal life 12:876
historical background 12:875–876
Jewish cultural production 12:877
Jewish feminist scholarship and activism 12:875
membership in Jewish congregations and communal organizations 12:876
research 12:876–877
women within Jewish life 12:875
historical identity of 12:872
influence on Western society 12:869
and Jewish identity 12:870–872
leadership in 20:410
and matrilineal principle 12:870
Max Weber's *Ancient Judaism* 12:869
in modern society 12:874
in North Africa and Arab nations 16:378
postexilic 12:869
Reform 12:871–872
reform 12:871
secularist Jewish movements 12:871
secularization and fundamentalism 12:872–874
Shabbat (Sabbath) and kashrut (dietary) observances 12:873
Six-Day War 12:872
transformation of Israelite religion into 12:870
Weber's analysis 12:869
Zealots 24:237–238
Zionist identity 12:870–871
Judaizers 3:548
Judeo-Christian tradition 15:807, 23:285, 25:669–670
Judges 8:582–584, 12:879–884, 18:387–388
adjudication 5:103–107
civil and common law judges 12:880–881
in civil law systems 8:582
in common law jurisdictions 8:582
evaluation of reliability 8:583
expanding power of 12:881
increasing convergence 12:880
judicial independence 12:879, 12:881
judicial systems in developing countries 12:880
judicialization of political life 12:881
and legal complex 18:388
for liberal politics 18:389–390
mediators and arbitrators 12:879
presence of women in judiciary 12:882
presiding 8:582
selecting and training, content validity 4:775
selection of 12:881–882
social origins of 12:882
social representativeness 12:882
types of 12:879–880
democratic 12:879
nondemocratic judge 12:880
traditional or predemocratic judge 12:879–880
Judgment and decision making (JDM) 5:917–922
Judgment conscious and unconscious knowledge 13:87–88
Judgment process, in content validation 4:775
Judgment sampling, in ethnography 8:186
Judgment under uncertainty 5:893
Judgments 10:559, 20:881
diffusion process 6:396, 6:398–399
and environmental cognition 7:709–711
and irrationality 12:722
Kantian ethics 13:7–8
moral 7:319, 15:788–789, 15:791–794
personality traits 17:815–816
recalling/computing, questionnaires 19:772
in risk 10:698–700
strategizing 23:523–526
under uncertainty, Tversky 24:681–682
Judgments of learning (JOL) 15:283–284, 24:89–90
Judgments under suboptimal conditions 10:825
Judicial activism 13:740
Judicial behavior 12:885–890
attitudinal model 12:886–889
courts and executives 12:888
courts and legislatures 12:887–888
ideology 12:886–887
judicial decision-making 12:885–889
judicial review 12:888
law 12:885
and legal formalism 13:744–745
organized interests 12:888
other justices 12:887
public opinion 12:888–889
solicitor general 12:888
Judicial decision making 12:885–889, 12:894–895, 13:793, 14:277
Judicial decisions, and administrative law 1:111–112
Judicial deference 12:900
Judicial elections, judicial independence and 12:894–895
Judicial errors 13:575
Judicial independence 5:104–105, 12:879, 12:881, 12:891–896, 18:450
definition 12:891–892
effects of 12:893–895
in comparative perspective 12:894
and judicial elections 12:894–895
and U.S. Supreme Court 12:894
judicial decision making 12:894–895
judicial review 12:891–895
measurement 12:892
de facto judicial independence 12:893
de jure judicial independence 12:892–893
relationship between *de jure* and *de facto* judicial independence 12:893
Judicial jurisdiction 13:616
in rem and quasi in rem 13:616
source of 13:616
Judicial management of contracts 4:811–812
Judicial notice doctrine 1:824
Judicial policy making 13:742
Judicial power and independence, models of 13:430–431
Judicial restraint 4:702, 13:740
Judicial review 4:700, 4:707, 12:888, 12:891–895, 12:897–901, 23:722–723
in ancient China 12:897
contemporary debate 12:899–900
democracy and 12:899
globalization of 12:898–899
historical origin 12:897
in Judeo-Christian tradition 12:897
in modern democratic polities 12:897–898
in United Kingdom 12:898–899
in United States 12:897, 12:899
Judicial systems 13:901
in developing countries 12:880
Judicialization of politics 12:881, 13:441
Judicialized American political system 12:898
Judiciary function 19:612
Judt, Tony 6:814

works, *Postwar* 6:814
Jugaad 11:786
Jugaad urbanism 24:820–821
Juhn–Murphy–Pierce (JMP) decomposition 6:539
Julian Stanley, works, *Experimental and Quasi-Experimental Designs for Research* 3:247
Jülich-Cleve-Berg conflict 20:102–103
Jump process 17:255
Jumpstart our Business Startups Act (JOBS Act) 3:32
Jun, Heejin, on careers and social networks 5:468
Jun, Pi, on transgenders 21:738
Jung, Carl Gustav 2:701, 5:471, 9:416, 12:902–906, 20:355
analytical psychology of 12:904–905
on collective unconscious 12:904–905
contemporary influence of 12:905–906
on individuation process 12:904–905
International Psychoanalytic Association 12:903–904
on linguistic relativity 25:576
on psychological types 12:904
on psychotherapy theories 19:503
works
Answer to Job 12:905
Mysterium Conjunctionis 12:905
On the Psychology of Dementia Praecox 12:903
Psychological Types 12:904
Synchronicity as a Principle of Acausal Relations 12:905
Zurich school of psychiatry 12:902
Jung's archetype 5:471–472
Junk science *see* Pseudoscience
Junk theory of aging 17:175
Junkspace 2:324, 18:675
Jural relationship 13:38
Jural-political domain 13:39
Juridical legal pluralism 13:757
Juridicalization of politics 12:899
Juridification 13:818
Juries 12:907–912
versus alternative triers 12:911
in civil cases 12:907
civil liability and damages 12:911
competence 12:910
composition 12:909
in criminal cases 12:907
decision-making process of 12:908
educative function of 12:908
evaluating 12:908–909
ideal 12:907
individual differences and 12:909
judges *versus* 12:907, 12:910–911
media and 12:909–910
medieval 12:907
methodological challenges in studying 12:908
modern American 12:907, 12:909
nullification and 12:910–911
peremptory challenge and 12:909
political roles of 12:907–908
prospective 12:909
reactions to jury trial 12:912
reactions to law 12:910–911
remedies and damages and 20:414–417, 20:416f
selection of 12:909
size and decision rule 12:909
and trials 24:637
unanimity *versus* majority decision rule 12:909
in the United States 12:907
Juris, Jeffrey S., works, *Networking Futures* 1:685
Jurisconsult 12:748
Jurisdictional disputes and professional monopoly 13:633
Jurisdictional political power 23:344
Jurisdictional question 13:615
Jurisdictions 13:476–477, 13:619
over child welfare 11:764
regional planning 20:145
and social work 22:810
Jurisdiction-specific noncompliance 12:959
Jurisprudence 2:526
Albert on 1:497–498
in America 13:755–756
Bentham on 2:527
Bobbio on 2:714
history 13:754–756
indigenous rights
international treaty 11:798–799
regional treaty 11:799–800
and utilitarian social thought 24:981–983
Jurisprudence and Social Policy (JSP) program 13:516–517
Jurisprudence constante 3:671
Jurisprudential gerontology 13:424
Jurisprudential history, legal personality 13:754–756
Juror, appraisal of witness reliability 8:647–648
Jury courts, in Europe 5:228–229
Jury law (1995), Spain 5:229
Jury trial 13:748
Jus ad bellum 12:508
Jus cogens 12:509
Jus commune 13:801
Just, Marcel 12:304
Just cause for war 25:398
Just city 12:913–918
Castells on 12:913–914
citizen participation 12:914
communicative rationality 12:915–916
democracy 12:915–916
development of concept within urban studies 12:913–914
diversity 12:916
equity 12:916–917
foundations for normative criteria in empirical research 12:914
Harvey on 12:913–917
Lefebvre on 12:913–915
making the implicit explicit 12:915
neoliberalism 12:915
Rawls on 12:915–916
right to the city 12:914, 12:916–917
spatial fix 12:913–914
spatial justice in the developing world 12:914
uneven development 12:914, 12:916–917
urban redevelopment 12:914–915
urban social movements 12:913–914
Young on 12:915–917
Just community school 15:776
Just deserts, doctrine of 5:217
Just noticeable difference (jnd) 19:470
Just price 19:111
Just War theory 17:637, 17:639–641, 21:33
Just world 2:487–492, 12:950, 12:952–953
assimilation of injustice 2:489
conceptualization of 2:487–488
consequences of 2:488–490
as disposition 2:487–488
hypothesis 2:487
justice motive *versus* justice motivation 2:488
personality dispositions and 2:488
social behavior and 2:488–489
subjective well-being and 2:490
trust in justice 2:489–490
Justice 3:644, 8:79, 12:607, 12:932–935, 12:937, 12:949–955, 17:620, 22:397, 25:347–348
access to *see* Access to justice
areas of interest and future challenges 12:953–954
biomedical research ethics 8:79–80
Boltanski on 2:747–748
conflict 12:950, 12:953
conflicts and 12:940
of constitutions and codified laws and ethics 12:939
criminal *see* Criminal justice
current interest and social applications 12:953
deserts 6:233
dilemmata 12:940
distributive 12:950, 17:718, 20:519
distributive justice 12:938, 17:620
divergent views of 12:937–938
due process model 17:718
economic philosophy 7:103
and emotions 7:537
equality 12:950
and equality 7:906–911
equitative justice 17:620
equity 12:949–953
ethical aspects in surveys and polling 23:754
forensic psychology 9:351
Frankfurt School 9:385
and health 12:928
background 12:928
concept of organizational justice 12:930
distributive justice 12:928, 12:930
effort–reward imbalance 12:930
equity theory 12:928–930
fairness 12:930–931
income inequality 12:929
interactional justice 12:930

Justice (*continued*)
justice of exchange 12:928–931
notions of justice 12:928–929
procedural justice 12:928, 12:930
social reciprocity 12:930
and honor killing 11:185–186, 11:188–190
interactional 12:951
intergenerational *see* Intergenerational justice
intergroup relations 12:412–413
belief in just world and intergroup relations 12:413
perceived justice 12:412
relative group deprivation 12:412
in international research 20:519–520
for Janitor 24:493
just and unjust contracts 12:940–941
just decision making 12:939
just world hypothesis 12:950, 12:952–953
and law 12:136–137, 12:932–936
and legitimacy 13:823, 13:827
and liberalism 14:29–31
and meritocracy/tokenism 15:239–240, 15:242–243
motivation 2:488
motive 2:488, 12:937
and multiculturalism 16:22, 16:24, 16:26
origins and early development of field 12:949–950
people caring about 12:951–952
philosophical aspects 12:943–948
agents of justice 12:947–948
bias toward purely recipient-oriented theorizing 12:944–946
causal pathways 12:947
first principle 12:944
iudicanda 12:943, 12:945, 12:947–948
natural inequalities 12:946
public criterion 12:943–944
Rawls theory 12:943–944
second principle 12:944
Sen's 12:946
toward a more adequate theory 12:946–948
and primordial sociobehavioral outcomes 22:879–881
procedural justice 5:297–298, 12:939, 12:950, 17:718
and public interest 19:552
punitive justice 17:620
of ratees 17:718–719
reactions to injustice 12:952–953
refugees 22:769–770, 22:772
regulatory agencies 20:183–184
relative deprivation 6:214–215, 6:217, 12:949
republicanism 20:477, 20:479–481
resource theory of social exchange 12:953
restorative 5:211
restorative justice 17:620
retributive justice 12:938–939
and rights 20:683–684
Sandel on 20:893–895
sensitivity 12:952–953
by settlement of conflicts 12:940
social *see* Social justice
in social exchange relationships 12:938
social psychological discussion of 12:949
social psychology 22:541
torts in law and 24:430–433
treatment research and access 8:31, 8:33, 8:35
trust in 2:489–490
and welfare 25:507, 25:509
and world republic 25:744
see also Injustice
Justice administration, socialist law and 22:847
Justice as fairness, Rawls' theory of 2:671, 19:914, 19:916–917, 22:392
Justice climate 17:383
Justice Denied (magazine) 25:787
Justice judgments 2:489
Justice leverage diversion 4:373
Justice perceptions
and abusive supervision 23:702
relational theories 23:702–703
Justice reasoning
culture-based differences in 15:783–785, 15:784t
moral values and 15:785
Justice system 14:254
Justice-enhancing devices 12:934
Justice-oriented enforcement 13:575
Justicia 12:137
Justification of moral and legal principles 12:362, 12:365
Justification problem, Albert on 1:496
Justifications
for affirmative action 1:250–251
combined reasons 1:251
compensation for past discrimination 1:250
creamy layer, resolution to 1:252
diversity-based rationales 1:250–251
egalitarian/redistributive argument 1:250
identity and recognition 1:250
issue of who loses and how much 1:252
maintain cultural values and institutions of community 1:250
for majority and minority community 1:251
membership in groups 1:252
preventing direct discrimination 1:250
rationale for protecting aboriginal groups 1:251–252
social cohesion argument 1:251
Boltanski on 2:748–752
and critical rationalism 5:272, 5:275
exculpatory defenses and 5:220
for lawful use of force 5:223
necessity 5:223–224
punishment 19:631–632
and reflective equilibrium 20:77–81
relativism about 20:235
self-defense and 5:223
and wisdom 25:601–602
Justified trusting 24:670–672
Justin (Martyr) of Samaria 3:548
Justinian, Emperor 13:801, 13:915
Corpus Iuris 13:915
Corpus juris civilis 13:801
Just-in-time manufacturing 18:665
Juvenile delinquencies 22:237, 25:814–815, 25:824
and youth culture 25:814–815
Juvenile Delinquents Act 8:249
Juvenile justice 3:480–481, 12:956–962
detention 12:960
race–crime relationship 19:838, 19:842–844
within UNCRC 12:960
Juvenile justice law 12:956
human rights instruments and 12:956–958
United Nations Standard Minimum Rules for administration of 12:956
Juvenile Protection Association (JPA) 25:814
Juvenile sex offending/offenders 20:438, 21:660, 21:662
Juvenile transfer 12:959
Juvenile waiver 12:959
Juvenilization of justice 19:640
Juxtaposition 23:343
of websites 4:508–509
JWT *see* J. Walter Thompson (JWT) agency
Jyväskylä Longitudinal study of Dyslexia (JLD) 20:9

K

K-12 curriculum 7:21, 14:80
K-1 groups 23:416
K 'test performance' classification 23:416
Kaan, Heinrich, works, *Psychopathia Sexualis* 21:846
Kabbalah 16:169
KABC-II *see* Kaufman Assessment Battery for Children version II (KABC-II)
Kaberry, Phyllis 7:145
Kabo, Viktor 9:672
Kabre of Togo 9:784
Kádár, János 24:449
Kadi justice 12:750
Kadt, E. de, works, *Tourism: Passport to Development?* 24:470–471
Kaernbach, C., works, *Psychophysics Beyond Sensation* 19:475
Kafir (non-Muslim) 12:727
Kafka, Franz 24:124
Kagan, J.
temperament development approach 24:193–194
works
Minangkabau Social Formations 5:109
Kahlbaum, Karl Ludwig 13:127
Kahneman, Daniel 8:393–394, 17:782, 24:679–682
prospect theory 19:262–263
Kainate receptors 12:711
Kain's hypothesis 23:157
Kaiser, M. 3:318
Kaiser-Guttman rule 8:706
Kaiser-Permanente (K-P) 15:102
Kaitiakitanga 11:785–786

Kakadu National Park 2:243
KAL (knowledge about language) 14:227–228
Kala, U., works, *Great Chronicle of Kings* (*Mahazawingyi*) 11:83
Kalahari Peoples Fund 1:224–225
Kalai–Samet payoff-configuration value 4:877–878
Kaldor, Mary, works, *Baroque Arsenal* 24:142–143
Kaldor, Nicholas
  on division of labor 6:602, 6:604
  stylized facts 7:46
  works
    Strategic Factors in Economic Development 6:604
Kaldor improvement 13:573
Kaldor–Hicks criterion 5:47–48
Kalecki, Michal 20:728
  on profits 19:115
  theory of employment 20:729
Kaleidoscope Career Model (KCM) 3:130
Kalenjin 8:177
Kalki tradition 15:531
Kalkutungu 3:184
Kalleberg, A., works, *Fighting for Time* 24:401
Kalman filter 23:196
  for prediction 24:358–359
Kalooban 11:791
Kaluluwa 11:791
Kalven, Harry 13:516
Kalyvas, Stathis 1:444, 21:305
  works, *The Logic of Violence in Civil War* 10:3–4, 21:305
Kam, L.Y.L., on cooperative marriage in China 21:735
Kamalayan 11:791
Kames, Henry Home, Lord 7:722, 8:292
  on division of labor 6:603
Kami 20:367–368
Kamiashi 22:572
Kamper, Dietmar 2:729
Kamprad, Ingvar 16:195
Kana scripts and reading 20:2–3
Kanaval 3:147
Kandel, Eric 9:886
Kandiyoti, D. 3:316, 3:318
Kane, Danielle
  on cultural consumption 16:557
  on cultural participation 16:556
Kang, W., on decriminalization of homosexuality in China 21:735–736
Kang Youwei 8:233
Kanji 20:1–3, 24:584
Kannada 3:186
  gender assignment in 9:795–796
Kansas Sexually Violent Predator Act 20:434–435
Kant, Immanuel 2:149, 2:315, 3:708–709, 4:337–338, 7:258, 7:722, 8:74–75, 8:779, 10:736–738, 11:93, 11:380–381, 12:388, 13:1–4, 13:437–438, 14:18, 16:334, 16:683, 19:314, 19:411, 22:890, 24:443
  on aesthetic education 1:229
  on aesthetic judgement 13:3
  on autonomy of ethical self-determination 13:3
  biography of 13:1
  on *Bürgerliche Gesellschaft* 3:708–709
  on 'cardinal propositions' of human knowledge 13:2
  on contractarianism 4:804, 4:807–808
  on cosmopolitan anthropology 5:42–43
  on 'cosmopolitan importance' of philosophy 13:4
  on critical refoundation of ethics 13:3
  education of 13:1
  on Enlightenment 7:664, 7:666
  on ethics 13:5–10
    of autonomy 13:5–8
    critique of practical reason 13:7
    doctrine of virtues 13:7–8
    of duty 13:5–8
    extracting and vindicating categorical imperative 13:5–7
    groundwork of metaphysics of morals 13:5–7
    judgment 13:7–8
    living the categorical imperative 13:7–8
    moral law, free will, and the highest good 13:7
  on eudaemonism 8:214–216
  final years of 13:4
  on habit 10:476
  on introspection 12:702
  legitimization for political rule, principle of 13:3
  on moral philosophy 4:338
  on organism, theory of 13:3
  on politics 13:5–10
    eternal peace 13:9
    perpetual peace 13:9
    and right 13:8–9
    right as compatibility of freedom 13:8–9
  on principle of the will 13:2–3
  on public use of reason 19:610–611
  on republicanism 20:478
  on scientific knowledge 13:2
  on self-organization, theory of 13:3
  on speculative thinking of pure reason 13:2
  on theory of critical justification 13:3
  on theory of world peace 13:3–4
  on things-in-themselves 13:2
  on transcendental philosophy 13:1–2
  on wheels 13:135
  works
    Critique of Judgement 13:3
    Critique of Pure Reason 13:2
    Dreams of a Spirit-seer 13:2
    Grounding for the Metaphysics of Morals 13:2
    Metaphysical Foundations of Natural Science 13:2
    Observations on the Feeling of the Beautiful and the Sublime 13:2
    Perpetual Peace 17:639, 25:740, 25:744–745
    Thoughts on the True Estimation of Living Forces 13:1–2
    Universal Natural History and Theory of the Heavens 13:1–2
Kantemir, Prince Antiokh 3:774
Kantian conception of autonomy 2:314–315
  and noumenal self 2:315
  and principles of justice 2:315
  and Rawls's notion 2:315
Kantian ethics 25:147
  see also Deontological ethics
Kantian heritage 11:380–381
Kantian idealism. 11:93
Kantianism 2:634–635, 11:381
  and human sciences 11:380–382
  objectivity 11:382
  value pluralism 25:23–24
Kantorowicz, Ernst H., works, *The King's Two Bodies* 2:729
Kanun 12:750
Kao 1:192
Kaplan, Robert, works, *The Revenge of Geography* 7:724
Kaplan–Meier estimator 23:771–772
Kapwa 11:792
Karadzic, Radovan 18:502
Karasek job demands–control (JDC) model 25:733–734
Karen tribal group 3:532
Kargil War 25:421
Kariel, Henry, works, *International Encyclopedia of the Social Sciences* 18:235
Karitiána and Havyaka Kannada negative sentences 16:413
Karl Friedrich Schinkel's museum of Berlin 2:22f
Karlinsky, Simon, on Russian sexual life 21:740–741
Karlsson, H., works, *Philosophy and Archaeological Practice* 1:896
Karma 13:800, 19:308
Karnataka State Farmers' Association 1:684
Kartagener syndrome 2:115
Karup–King–Newton formula 6:132
Karush–Kuhn–Tucker conditions 15:651
Kastenhofer, Karen, on epistemic cultures 7:878
Katakana 20:1
Katayama, Sen 22:839
Kathedersozialisten 4:340
Katib 11:62
Katona, George 5:887
Katz, Elihu 13:638
  on media events 15:48–49, 15:51–52
  works
    Broadcasting in the Third World 19:879–880
    The Export of Meaning 2:217
    Media Events: The Live Broadcasting of History 15:48, 15:51
Katz v. United States (1967) 18:948, 18:951
Katz–Bonacich network centrality 16:573–574, 16:579

Katz-Gerro, T., on cultural tastes 22:334–335
Kauffman, Stanley 1:368
  on adaptive radiation model 24:131
  and needs assessment 16:403
  random networks 1:368
Kaufman Assessment Battery for Children version II (KABC-II) 7:930–931, 12:287
Kaufman Pattern Reasoning scale 6:329
Kaupapa Māori 11:785
Kautsky, Karl 22:834
  works, *The Agrarian Question* 1:456
Kavafy, Constantine 21:742
Kayahara, Jennifer, on cultural consumption 16:557
Kazakhstan 3:317–320
k-balance (structural balance) 23:606
KBV *see* Knowledge-based view (KBV)
KCM *see* Kaleidoscope Career Model (KCM)
Keane, Michael, on Chinese cultural industries 5:597
Keeley, L., works, *War before Civilization. The Myth of the Peaceful Savage* 4:607
Keeping balance concept, in Islam 12:757
Keeping record of individual referents 10:318
Keeping the Rabble in Line (interview) 3:529–530
Keil, Frank C., works, *The MIT Encyclopedia of the Cognitive Sciences* 7:590
Keiō University 12:795–796
KEK *see* Kommission zur Ermittlung der Konzentration im Medienbereich (KEK)
KEKADA (scientific discovery system) 21:294
Kelbaugh, Douglas 16:809–810
Keller, H., works, *Cultures of Infancy* 3:396
Keller, Suzanne, on elites 7:392–393
Kelley, Donald 12:260
Kelley, H.H., works, *The Social Psychology of Groups* 8:483
Kelly, Erin, on theorization 17:392
Kelly, George 4:724
Kelly, Joan 9:768–769
Kelly, Raymond C., works, *Warless Societies and the Origin of War* 17:625
Kelpius, Johannes 2:146
Kelsen, Hans 12:881, 13:763, 18:435, 21:169
Kelty, Christopher, works, *Two Bits* 12:604
Kelvin, Lord 18:102–103, 21:226
Kendall, Katherine A. 12:475, 12:478
Kendall, Laurel 2:72–73
Kendon, Adam 10:113–114, 10:118–119
Kennedy, Anthony 24:444
Kennedy, Duncan 18:650
Kennedy, John F. 18:461
  assassination, and flashbulb memory 9:254
Kennedy, Paul 1:826
Kennedy Assassination 9:866–867
Kennell, J.H., works, *Maternal–Infant Bonding* 16:469
Kennewick man 1:633
Kensington Welfare Rights Union (KWRU) 17:250–251
Kenworthy, Jeff 24:866–867
Kenworthy, Marion 25:625
Kenya 6:773, 6:775–776
  gender in 9:786
  peace building and music in 25:845–846
Kenyan Muslims 12:739
Kenyon cells 15:157
Kept lawyers 12:127–128
Kerensky 20:857
Kern, R.M., on omnivorousness 22:336
Kernel estimation 21:599
Kernel methods 23:411
Kernel of neopositivism 14:315–316
Kernel smoothing, nonparametric regression 16:868–871, 16:869f–871f
Kernel width test 23:395
Kerr, Clark 7:904, 21:168
Kertcher, Zack, on epistemic cultures 7:878
Kerven, C. 3:318
Késenne, S., works, *The Economic Theory of Professional Team Sports: An Analytical Treatment* 23:318
Ketoacidosis, and diabetes 6:337, 6:339–340
Key, Valdimer Orlando, Jr. 13:11–13
Key, Valdimir Orlando, Jr.
  on American Progressive tradition 13:12
  assessment of professional impact and current importance 13:12–13
  biography of 13:11
  on democratic value commitments 13:11–12
  on effects of industrialization and urbanization on society 13:12
  on electorate's effective role in politics 13:12, 13:12t
    survey-research portrait of voters 13:12
  on gap between ideals and empirical reality 13:12
  scientific contribution of 13:11–12
  study of American politics 13:11
  works
    Public Opinion and American Democracy 13:12
    The Responsible Electorate 13:12
    Southern Politics in State and Nation 13:11
Key actors, in ethnography 8:187
Key events, ethnography 8:189
Key Indicators of the Labour Market (KILM) 5:798
Key opinion leaders (KOLs) 18:9, 20:526
Key player policy, social networks 16:579–580
Key stimulus 14:367–368
KeyCite 5:808
Keyfitz's formula 18:575, 18:586
Keyhole Markup Language 10:98
Key-Math3 16:725
Keynes, John Maynard 7:94–95, 7:114, 13:14–19, 14:411, 14:621, 16:228, 17:505, 18:359–360, 18:664, 19:39–40, 20:660, 20:728, 22:698–699, 25:273
  acceptance of theory and policies 13:18
  on aversion of banking crisis 13:14–15
  on balance-of-payments adjustment 13:16
  at Cambridge, London, and East Sussex 13:15
  as chief proprietor and chairman of weekly journals 13:15
  on Clearing Union Plan 13:18
  on disequilibrium between saving and investment 13:16
  on distinction between fixed and flexible prices 13:16
  on distinctive theory of probability 13:16
  on economic fluctuations, theory of 13:16
  on economic theories 13:109, 21:129
    interest theories 9:218–219
  and Fisher 9:220
  influence of 13:14
  Keynesian economics, debate of 13:18–19
  life of 13:14–16
  notation Y= C + *I* 13:17
  on profits 19:113–115
  on Quantity Theory of Money (QTM) 13:16
  on self-regulating economy 13:16
  on stabilization policy technique 13:17–18
  theoretical exercises of 13:15
  theoretical rationale for the New Deal 13:15
  on theory of effective demand 13:16–17
  on underemployment equilibrium 13:15
  works
    The Economic Consequences of the Peace 13:14–15
    The General Theory of Employment, Interest and Money 13:14–17
    Indian Currency and Finance 13:14
    The Monetary Theory of Production 17:955
    A Tract on Monetary Reform 13:15
    A Treatise on Money 13:15
    A Treatise on Probability 13:22, 24:718
Keynesian compact 23:345
Keynesian interventionism 2:750
Keynesian macroeconomics, and Cold War 4:123–124
Keynesian parameterization 9:320
Keynesianism 13:20–23, 14:653, 16:465, 18:304–305
  criticism of 13:21–22
  ideas, reception of 13:20–21
  and macroeconomics 14:401
  relevance of 13:22
  use of social statistics 23:426
Keyser, Leander, works, *A Handbook of Christian Psychology* 7:591
Key-word-in-context (KWIC) information 4:771–772
K-fold cross-validation, in statistical pattern recognition 23:416
Khadi justice 24:635–636
Khadija 12:731
Khaldun, Ibn 22:720
  works, *Muqaddimah: An Introduction to History* 9:858
Khaled, Leila 24:238
Khalid, A. 3:320

Khalidi, Rashid, works, *Palestinian Identity* 16:384
Khalistan, and communalism 4:264
Khalistani movement of Sikhs 20:401
Khan, Aisha 1:641
al-Khattabi, Abd al-Karim 24:603
Khazanov, Anatoly 3:318, 9:672
Khazraj tribe 12:732
Khmer Empire 13:812
Khozraschet 20:838
Khrushchev, Nikita 20:823
Al-Khwarizmi 21:173
Khyâl Cap ('Wind Attacks') 5:626
Kiaer, Anders 19:563
Kibbutz 2:148
  socialization 4:267–268
KIBS *see* Knowledge-intensive business services (KIBS)
Kidmapping, and political apportionment 7:343
Kierkegaard, Sören 1:539–540, 4:151, 19:314, 19:411
Kijaji 2:870–871
Kilchun, Yu 2:71
Killing, transformation to nonkilling 17:626
KILM *see* Key Indicators of the Labour Market (KILM)
Kim, Eleana J. 2:74
Kim, J.Y., works, *Dying for Growth: Global Inequality and the Health of the Poor* 7:19
Kim, Seong Nae 2:74
Kim, Seung-Kyung 2:73
Kim, U.
  works
    Indigenous and Cultural Psychology: Understanding People in Context 11:793
    Indigenous Psychologies: Research and Experience in Cultural Context 11:793
Kim Il Sung 24:449
Kimberley spear points 11:2
Kimchee Towns 10:125
Kimura, Motoo 13:24–25
  assessment of work 13:25
  biography of 13:24
  life in wartime Japan 13:24
  scientific contributions of 13:24–25
  on separation of genetics from environmental effects 13:25
  stochastic viewpoint on study of evolution 13:25
  study of mathematical problems 13:24
  study of natural selection and adaptation 13:25
Kin adoption 1:137
Kin altruism 22:703
Kin recognition
  kin selection 13:29–31, 13:29f
Kin selection 8:399–400, 13:26–31, 15:811–812, 22:900–901
  altruism 13:26–30
  cooperation 13:29–30
  and eusociality 13:30–31
  genetical theory of 13:26–28
  greenbeard 13:29–31
  Hamilton's rule 13:26–28, 13:31
  inclusive fitness 13:29
  kin recognition 13:29–31, 13:29f
  and kin recognition 13:29–31, 13:29f
  mechanisms of 13:29–30
  and morality 15:811–812
  mutual benefit 13:28t
  neighbor-modulated fitness 13:28
  origins of 13:26
  personal fitness 13:26, 13:28–29, 13:28f
  price equation 13:26
  relatedness 13:26–31
  selfishness 13:29, 13:28t
  spite 13:28, 13:28t
  theory 5:405
  viscosity 13:29, 13:31, 13:29f
Kin Term maps 13:56–57
Kin Term Space 13:55–58
Kin terminology 13:50–51
Kin work 10:696
Kindchenschema 12:9–10
Kinder, Katja 8:908
"Kinder and gentler" drug market 5:199
Kindertransport 21:353
Kindleberger, Charles, works, *Manias, Panics, and Crashes* 9:183
Kindling 13:32–35
  discovery 13:32
  electroencephalogram 13:32
  hyperexcitability 13:34
  as model for seizures and epileptogenesis 13:34–35
    consequences of multiple seizures 13:35
    creation of seizure focus 13:34
    interictal behavioral dysfunction 13:35
    multiple seizure generators 13:34
    occurrence of focal and secondarily generalized seizures 13:34
    pharmacoresistance and conditioned tolerance 13:34–35
    progression in severity of seizures 13:34
  plasticity 13:32
  properties 13:32–33
    brain damage 13:33
    differential susceptibility of brain areas 13:32–33
    generalizability 13:33
    persistence 13:33
    stimulation interval 13:33
Kindred concepts, ignorance and 11:579
Kinect controller 25:133
Kinematics, motor control 3:778–779
Kinesics 4:287
Kinesophobia 17:452–453
King, Martin Luther, Jr. 5:84, 16:324
  Civil Rights Movement 3:692–693
King, Mary Claire
  human rights genetics 9:971–972
King v. Smith (1968) 18:735
Kingdom of Bhutan *see* Bhutan
Kingdoms, East Africa 6:774
Kingdon, John 18:289
  model of agenda-setting 1:358
Kings, drag 13:875
  see also Drag culture
King's College (University of Cambridge) Sociobiology Group
  Current Problems in Sociobiology 22:902
Kings, transition of chiefs to 11:895–896
King's courts 4:247–248
Kingston, Paul W.
  on cultural capital 5:376, 5:379
Kin-militias 4:601–602
Kinney National Services 9:157
Kinsey, Alfred 10:801–802, 10:805–806
Kinship 7:387, 8:741, 13:41, 15:816
  ancestors 1:688–689
  in anthropology 13:36–43
  in Arctic 1:929
  of Australian aboriginals 2:235
  body and aesthetics 13:40
  clan 3:730–732
  culture 20:450–451
  in cultures of relatedness 9:427–431
  demographic microsimulation 15:344–346
  descent and marriage 13:37–38
  domain of 13:53
  domestic and jural, distinction between 13:37
  elementary structures of 18:912
  evolution of 13:44
    contextual mechanisms 13:46
    and culture 13:49–51
    kin recognition 13:46–47
    kin selection and inclusive fitness 13:45–46
    major transitions 13:48–49
    multilevel selection 13:48–49
    phenotypic matching 13:46–47
    relatedness 13:44–45
    and social structure 13:47–48
  and family 8:809–813
  family and household demography 8:725, 8:727–728
  formal models of 13:53–60
    algebraic structure 13:59
    cultural constructs 13:53, 13:55
    cultural knowledge 13:53
    cultural theory 13:53
    family space 13:55, 13:58
    family structure 13:55
    formal representation of computations over genealogical space 13:58–59
    genealogical relations 13:53, 13:58–59
    genealogical space 13:56, 13:58
    kin term maps 13:56–57
    Kin Term Space 13:56–58
    marriage 13:53–55, 13:58
    primary kinship 13:53–55
    terminology 13:57–58
  and gender 13:39–40
  and genealogy 9:806–810
  and generation 9:853–854
  and home 11:156
  hope for 13:38–41
  and households 11:228–231
  and joking 11:418
  in Korea 2:72–74
  law and order by another name 13:36–37
  Lévi-Strauss on 13:905–906, 13:909

Kinship (*continued*)
Lévi-Strauss structuralism 1:768, 23:628–629
LGBTQ 21:755
and marriage 14:599–600
and marriage systems 14:611, 14:616
and matrifocality 14:822–823
Melanesia 17:155–156
Middle Ages 15:349
naming 16:184
networks of relationships and personal kind terms 13:41
queer 19:754–755
and religion 20:317–321
reproductive medicine 13:40–41
in South Asia 2:79
and statehood 23:347–349
status of 13:41–42
structural analysis of 13:38
and symbolism 23:861–862
transition to class 11:895–896
tribes 24:645
Kinship analysis, and social justice 9:970–971
Kinship care, and emotional development of children 7:441–442
Kinship domain 13:55
Kinship relations 8:811–812
genealogy in remaking 9:809
Kinship structure, rural communities, in Africa 1:287–288
Kinship studies 1:766
Kinship terminologies 13:61–66
address terms 13:61
American 13:64f
biological kinship relations 13:61
classification of 13:62–63
consanguineal and affinal 13:61
cultural universals 13:62
elicitation procedures 13:63–64
formalism 13:63–65
and genealogical hypothesis 13:63
generative basis of 13:64–65
generative logic 13:65
kin term product 13:64–65
reference terms 13:62
relationship with form of social organization 13:63
structural logic of 13:63–64
structural properties of 13:65
systematic study of 13:62
Kintir 8:880
Kintirleey 8:880
Kinugasa, Teinosuke, works, *Page of Madness* 9:167
Kiobel v. Royal Dutch Petroleum (2013) 18:502
KIP *see* Knowledge-in-practice (KIP)
Kirchheimer, Otto 18:497
works, *Political Justice* 18:497
Kiribati 17:163, 25:645
Kirk, Russell, works, *The Conservative Mind* 4:690
Kissinger, Henry 10:93
Kitawala movement 15:531–532
Kitayama, S., works, *Handbook of Cultural Psychology* 5:312
Kitschelt, Herbert 25:305–306
Kivnick, Helen 7:935
Kjellén, Rudolf 18:374
KKK *see* Ku Klux Klan (KKK)
Klasse an sich (Marx) 3:745–746
Klasse für sich (Marx) 3:745–747
Klaus, M.H., works, *Maternal–Infant Bonding* 16:469
Kleiber's curve 13:148
Klein, Judy 7:96
Klein, Lawrence 20:147–148
macroeconomics 14:401, 14:403–404
Klein, Melanie Reizes 13:67–69
account of unconscious fantasy and transference 13:68–69
on anxiety-laden and intolerable unconscious fantasies 13:68
biography of 13:67
contributions 13:67
to child analysis and psychoanalytic clinical technique 13:67–68
controversial discussions about 13:67
on envy and death instinct 13:68–69
on envy of women's creative insides and psychic threat of images 13:69
on mechanisms as splitting, projection, and introjection 13:68–69
on paranoid-schizoid position 13:68
on practice and theory of child analysis 13:67
on psychic structure and psychic fantasy 13:68
on psychoanalytic theory 13:67
psychological developments studies 19:504
Kleinman, Sherryl, on meaning 16:561
Klein's model 17:469
Klesse, C., on gay/lesbian tourism 9:639–640
Kling, Rob 12:86
Klingemann, Hans-Dieter, works, *A New Handbook of Political Science* 7:591
Kliquefinder (software) 3:911
Kliuev, Nikolai 21:741–742
Klondike spaces 24:131
Kluckhohn, Clyde, prediction of culture 5:582–583
Klüver–Bucy syndrome 13:680
k-means algorithm 19:434, 23:183, 23:392–393, 23:395, 24:825–826
Knapped stone 14:260
K-Nect 6:898
Knez, Eugene 2:72
Knies, Karl 4:340
Knight, Frank H. 21:884–885, 21:887, 23:526, 25:273, 25:463
on entrepreneurship 19:114
on strategizing 23:526
works
Risk, Uncertainty, and Profit 21:970–971, 24:718
Knight, J.Z. 16:807
Knight, Robert 7:934
Knights 1:965–966
Knock on effect 5:49
Knockout animal method 9:934
Knock-out studies of learning and memory 9:906
interpretation of 9:907
Knorr Cetina, Karin 8:577–578
works, *The Manufacture of Knowledge* 8:577–578, 13:206–207
Knowers of the helping profession 25:573
Knowing
epistemological beliefs 7:887–893
organizational learning 13:697
Knowing how 15:282
Knowledge 2:495–496, 3:626–630, 5:781–782, 6:491–492, 12:95, 21:241, 22:913
about climate change 3:858
about family relationships 20:318
about semantic memory 21:572–574
and action 13:93, 6:238t
aesthetic 16:181
Afrocentric social work 1:305–306
agri-food knowledge production 10:78
anthropology of 20:227–228
roots of expertise in 8:588–589
application, managerial and organizational cognition 14:479
in archaeology 1:899–900
base on school effectiveness 21:81–83
categories, and semantic memory 21:575–576
categories of 13:91–94
categorizations of 13:93–94
center 13:92, 13:95
Chomsky on tacit knowledge of rules 13:74–76
faculty of language 13:75
Quine's challenge 13:75–76
codified 13:92–93
and cognitive relativism 20:227, 20:229–231
concept, developmental research 6:327
conceptual 12:326
conscious knowledge 13:85–87
in context of psychological theory 13:86–87
exhaustiveness and exclusiveness 13:86
judgment knowledge 13:87–88
structural knowledge 13:87
two criteria for test of 13:86
validating test of 13:86
construction 12:326–327
contributions of ethnohistory to 8:193–194
conversation analysis 4:852–853
core knowledge 12:325–326
Coser on 5:38–40
in critical social work 5:289
and cultural inequality 22:331
culture of 22:229–230
cultures 7:878–879
of developing writers 25:769
of developing writers for and about writing 25:769
directions, organizations 6:237

disciplines as unit divisions of 21:287
dissemination of 13:91, 13:93–94
domain-specific 12:327
and Eastern Europe Studies 6:805, 6:808
economic education 7:21–24
and emotion regulation 7:454
of environmental risks 7:786
epistemic communities/collaborative research 7:862
epistemological beliefs 7:887–893
everyday 13:93
factual 13:93
further perspectives 13:96
generation of 13:91–92, 13:94–96
generative grammar 13:74–75, 13:83–84
Giddens on 10:136–137
higher levels of 13:94
historical 1:748
implicit learning and tacit knowledge of rules 13:84–85
AGL in patients with amnesia 13:85
skepticism about 13:85
inert 12:326
and inferences 19:394–399
information 13:91–95
and information 13:93
infrastructure 16:858
infrastructures 21:189
-intensive jobs in the hierarchy of urban system 13:95–96
and international social problems 12:546
versus knowing 4:20
lack, and social work development 22:793
legal context of production of new 13:522–523
lexical 13:927
mechanisms 22:418
mental simulation 13:82–83
meritocracy 13:95–96
metacognitive 24:89–90
mobility of 13:92
modalities, and semantic memory 21:574–575
moral/orientational 13:93
and music 16:152–153
and neonatal imitation 16:484
and neuronal connections 4:96
new, creation of 15:575–576
in new global environment 10:153
and nongovernmental organizations 16:859–860
nonpropositional 13:93
notion of 7:875–876
objective 17:99
operative forms of 18:127
organization theory 13:95–96
organizational 12:72–76
changes 12:74–75
evolution 12:72
expertise 12:75
innovation 12:73–74
management 12:75
messages 12:73–74
properties 12:72–73
trends 12:74–75
periphery 13:94–96
personal levels 13:79, 13:81
Plato on 18:211
Plato's recollection (anamnesis) 24:722
policy
implications for appropriate forms of 18:308–309
transfer from universities into policy realm 18:310–311
policy knowledge 18:293–298
and postindustrial societies 18:663–665, 18:667–668
and power 13:91–92, 13:94–95
procedural knowledge 12:326
process of communicating different categories of 13:94
production
and peer review 17:685
production, in higher education research 10:867–868
production, new modes of 21:204
production, online 12:604
propositional 13:93
representation and rules: explicit and implicit 13:78–79
content-involving explanation 13:77, 13:80, 13:82–83
inference and implicit representation 13:78–79
rule-explicit and rule-implicit systems 13:79
research 13:88–89
restrictive concept of 13:108
science 21:236–239
and secrecy 21:374–378
of self and others 16:602
serial reaction time task 13:88
in simulation environments 4:488, 4:487f
social constitution of 21:201
and social constructionism 22:223
as social context of activity 21:98
social determinants of 21:261–262
sociohistory of 7:176
sociological 15:305–307
sociology of 4:164, 11:574, 13:111–118, 20:230
antecedents of 13:111–113
developments 13:115–117
as specialty area 13:113–115
sociology of, Burke 2:927
source of 7:890–891
in Southeast Asia 23:73
spatial disparities of knowledge 13:91–93, 13:96
spatiality of 13:91–92
learning processes 13:91
methodological problems of localizing knowledge 13:92
spatial differences of knowledge 13:91–92
stability of 7:890–891
and statehood 23:347–350
structure of 7:890–891
subpersonal levels 13:79, 13:82–83
superior 13:92
tacit 13:93
tacit knowledge 13:76–78
beliefs, and nonconceptual content 13:77–78
critiques 13:78
structure required for 13:76–77
tacit knowledge in philosophy 13:80–83
AGL task 13:83–84
compositional semantic theories 13:80–81
concept possession 13:81
finite state grammars 13:83–84
implicit rules in connectionist networks 13:81–82
learning and classification 13:84
theory and simulation in everyday psychological understanding 13:82–83
teachers 12:68
of teachers 24:51, 24:53–54
theoretical 13:93
and theory, current state of 9:977–979
translating legacy knowledges into public futures 21:183–184
types of 1:440
and uncertainty 24:716–718, 24:720
unconscious knowledge 13:85–87
exhaustiveness and exclusiveness 13:86
judgment knowledge 13:87–88
structural knowledge 13:87
and university press 19:625–627
verbal 13:93
in vocational education 25:236
of voters about economy 7:28
and wisdom 25:598–600
and writing 25:783
Knowledge about language *see* KAL (knowledge about language)
Knowledge acquisition 7:241, 13:712
ability for 7:890–891
absorptive capacity 1:13–16, 1:18
nature of 7:890–891
Knowledge assessment 4:6, 4:8
in learning spaces 14:586–588
Knowledge assimilation, absorptive capacity 1:14–16, 1:18
Knowledge base competency 24:521
Knowledge capital 19:109
Knowledge change 21:264
Knowledge claims 7:883–884, 24:673
Knowledge construction 7:241, 12:326–327
Knowledge consumption, and social networks 16:557
Knowledge creation 13:696, 16:552
Knowledge deficit 3:844–845, 3:844f–845f
Knowledge development, and scientific revolution 21:318–320
Knowledge diffusion 11:834, 11:837–838, 11:841
and innovation 12:173
Knowledge dissemination 7:283–284
Knowledge economy 3:13, 18:308, 24:759–760
industrial geography 11:858
industrialization 11:870

Knowledge economy (*continued*)
and international organizations 12:538–540
universities 18:307–308, 18:310
Knowledge empiricism 7:567, 7:569
Knowledge environment 7:170
Knowledge exploitation, absorptive capacity 1:14–16, 1:18
Knowledge gap 16:790
Knowledge generation, in Africa 1:298
Knowledge level of teachers 24:45
Knowledge management
absorptive capacity 1:12–13
processes 12:148–149
tacit 24:4–6
Knowledge producers, sex workers as 19:282–283
Knowledge production 21:150–152, 21:154, 24:763
Knowledge production cycle, in science 21:249–250
Knowledge productivity 3:13–14
Knowledge regime 24:297–298
Knowledge representation (KR) 2:37–38, 13:98–104
applications 13:99–100
robotics and planning 13:100
text interpretation and information retrieval 13:100
architectures 13:100–103
frame-based systems 13:102–103
large-scale systems 13:103
logic-based systems 13:100–101
semantic networks 13:101–102
artificial intelligence 13:98–99
desiderata 13:99
first-order logic 13:100–101
fundamental domains 13:103–104
other domains 13:103–104
time 13:103
ideal and reality 13:98–99
learning and 13:104
reasoning 13:98–100
temporal logic 13:101
Knowledge representation (KR), logics for 14:319–321
formalisms and semantics 14:319–320
in cognitive science 14:320
non-monotonic logics 14:320
qualitative spatial and temporal representation and reasoning 14:320
historical account 14:319
Knowledge retaining 13:696
Knowledge retrieval, levels of 1:440
Knowledge sharing 16:552
Knowledge society 7:876–877, 10:868, 12:88, 13:105–110, 15:12, 18:308, 18:665
conceptions of Peter Ducker and Daniel Bell 13:106–107
economic system of 13:109
and employment 7:127–128, 7:130–131
mature theories and antecedents 13:107
modern, dating 13:107–108
nature of 13:109–110
restrictive concept of knowledge 13:108
universities 18:307–309, 18:312
use of term 13:105–106
Knowledge spaces 4:6–9, 14:584, 14:589–590
Aleks on 4:8–9
clause 4:7–8
learning spaces 4:8
organizing body of knowledge 4:6–7
states 4:7
wellgradedness 4:8
Knowledge state, in learning spaces 14:584, 14:586–589
Knowledge structure, in learning spaces 14:584
Knowledge systems 13:678–679
reasoning for 5:939
Knowledge theories, ignorance in 11:580–581
Knowledge transfer 2:483, 13:696, 24:760
multinational corporations 16:79
Knowledge transformation, absorptive capacity 1:14–16, 1:18
Knowledge utilization 24:294
Knowledge workers 6:664–667, 18:665, 18:667
productivity 6:664–665
Knowledgeable publics 21:256
Knowledge-based economies 7:65, 17:341
Knowledge-based view (KBV) 23:522
Knowledge-building cycles 4:799–800
Knowledge–ignorance paradox 11:582
Knowledge-in-practice (KIP) 24:5–6
Knowledge-intensive business services (KIBS) 21:648–649
Knowledge-intensive goods 11:858
Knowledge-making, law as site of 21:160–161
Knowledge-poor inductive learning 19:60
Knowledge/power 11:392–394, 11:396–397
Knowledge-rich learning 19:60
Knox, Alexander, works, *The Climate of the Continent of Africa* 7:750–751
Koalitionfahig partisanship 2:191
Koch, Heinrich 17:1–2
Kockelman, Paul 1:237
Kocu, Resat Ekrem, works, *Encyclopaedia of Istanbul* 20:870
Kodaly, Zoltán 1:754
Kodi calendar 24:407
Koebner, Richard, contribution to conceptual history 4:549–550
Koffka, Kurt 10:750–751, 13:123, 19:414
works, *Principles of Gestalt Psychology* 22:533
Kogan, Maurice 7:591
Kohl, Helmut 18:393
Kohlberg, Lawrence 13:119–122
approach of 13:119–120
assumptions about moral judgment 7:315–317
biography of 13:119
early education of 13:119
influence of US psychology 13:119
on moral emotions 13:120
on moral judgment 13:119
on natural science epistemology 13:120
on normconforming behavior 13:120
on theory of moral development 15:782–783, 15:792, 15:806
on unity of moral judgment and action 13:119
on young children's moral development 13:120–121
claim of hierarchy 13:121
developmental sequence 13:121
mechanisms of development 13:120–121
'member of society' perspective 13:120
'prior to society' perspective 13:120
relationship between judgment and action 13:121
stage model 13:120
tests 13:120
theoretical revision 13:121
universal validity of stage model 13:121
Kohler, Robert 24:763–764
drosophila (fruit-fly) research of 8:580
works
Lords of the Fly 8:580
Köhler, Wolfgang 4:51, 19:414
accomplishments and impact on psychology 13:124–125
on animal cognition 13:125
on animal problem solving 13:123–125
on chimpanzees 13:123–125
chronology of 13:123t
death of 13:125
on human values 13:125
on insight 13:123–124
Koffka on 13:123
life and work of 13:123–124
on memory 13:124–125
on neurophysiology of perception 13:124–125
on perceptual organization 13:125
on psychoacoustics 13:123
on psychophysics 13:125
Wertheimer on 13:123–124
works
Die physischen Gestalten in Ruhe und im stationären Zustand 13:123–124
Kohn, Hans, on nationalism 16:310
Kohut, Heinz 19:504, 22:762
Koiné varieties of Creole 18:134, 18:138
Koinonia Politike
and its historical modulations 3:707–708
transformed 3:708–709
Kojiki 11:49
Kolakowski, Leszek 22:833
Kolami, gender assignment in 9:795–796
Kolb, J.H. 20:816
Kollontai, Alexandra 20:836
Kolmogorov, A. N. 13:24
axiomatic foundations 19:33–34
probability models 14:793
Kolmogorov backward equation 6:397
Kolmogorov complexity 1:534
Kolmogorov–Smirnov test 19:143
KOLs *see* Key opinion leaders (KOLs)

Kol'tsov, Konstantinovich 8:242
Kommission zur Ermittlung der Konzentration im Medienbereich (KEK) 5:705
Kommunist 20:838
Kompromat 20:847
Kondratieff waves 12:146
Konflikt-Simulations-Modell (KOSIMO) dataset 5:836
Kongo 3:313
König, René 4:341
  works, *Handbuch der Empirischen Sozialforschung* 7:588–589
Königlich Preubische Statistische Bureau 5:735
Kontradieff, Nikolai D. 16:229
Koolhaas, Rem 24:819–820
Koran 11:48
Korea
  Christianity in 3:532
  fertility transition in 9:55–59
  historiography development in 11:50
  immigration policy 11:664
  legal history of 6:778, 6:780–782
  modern, colonial gender policy in 6:794
  multicultural education in 16:19
  new religions 20:370
  religious forms in 20:369–370
  society and sociology in 24:585–586
  sociocultural overviews of 2:71–75
    anthropology, colonial origins 2:71–72
    conservative familialist ideology 2:73
    divided Korea and village ethnographies 2:72–73
    evolutionary anthropology 2:71
    history and memory 2:74
    Japan colonial rule in 2:71–72
    kinship 2:72–74
    neoliberalism 2:73
    North Korea 2:72, 2:74–75
    postcolonialism 2:72–73
    recent trends 2:73–75
    shamanism in 2:71–73
    social anthropology 2:71–72
    South Korea 2:72–75
    war 2:71–72, 2:74–75
Korean, scope preferences in first language learners of 13:314–315
Korean Anthropological Association 2:72
Korean argument nominalizations 23:893t
Korean Commercial Arbitration Board 12:470
Korean Labor & Income Panel Study (KLIPS) 5:846
Korean N-based nominalizations 23:893t
Korean politics
  in Cold War period 6:799
  in post–Cold War period 6:801–802
Korean protestant sects 20:368
Korean War Veterans Memorial 4:236–237
Korean writing systems 25:778
Koresh, David 3:356
Kornegger, Peggy 19:748–749
Korovin, Evgeny A. 22:847
Korsakoff syndrome 1:644, 6:722, 13:679, 19:275
Koselleck, Reinhart 14:383, 18:489
  conceptual history before and after 4:549–550
  works
    Geschichten 11:73
Kosher connection 24:598
Kôshôgaku 11:51
KOSIMO *see* Konflikt-Simulations-Modell (KOSIMO) dataset
Kossuth, Lajos 4:572
Kothari, U., works, *Participation: The New Tyranny?* 17:538
Koubaa, Khaled 16:801
Koyré, Alexandre 20:284
Kozin, Vadim 21:742
Kpaaîkpaapî 4:233
Kpaapî 4:233
Kpaapîkpaaî 4:233
KR *see* Knowledge representation (KR)
Kraaykamp, Gerbert
  on cultural capital 5:378, 5:380
  on social status and media consumption 16:557
Kracauer, Siegfried 11:111–112, 15:15
Kraepelin, Emil 13:127–132
  biography of 13:127
  on dementia praecox 13:129–130
  on diagnosis 13:129–131
  on epistemology 13:127–128, 13:131
  on history of psychiatry 13:130
  on illness 13:127, 13:129–130
  on mental illness 13:127, 13:129
  on natural disease entities 13:128–130
  on natural kind 13:129, 13:131
  neo-Kraepelinianism 13:130–131
  on neurobiology 13:128, 13:130–131
  on objectivity 13:131
  on psychiatric nosology 13:129–130
    basic aspects 13:129
    clinical aspects 13:129–130
  psychiatry as science 13:127–129
    degeneration theory 13:128–129
    experimental approach 13:128
    naturalism 13:128
    parallelism 13:128
    realism 13:128
  on psychopathology 13:127, 13:129, 13:131
  on psychosis 13:127, 13:129–130
  on schizophrenia 13:129–131
  on subjectivity 13:128
  twenty-first century psychiatry 13:130–131
  and Wundt 3:877–878
Krafft-Ebing, Richard von 13:888, 19:887
  works, *Psychopathia Sexualis* 3:885, 3:887–888
Krämer, Augustin 17:161–162
Kraus, Karl 18:435
Kreager, Derek A., on alcohol consumption 16:558
Krebs, Emil 16:66
  works
    Handbook of Evolutionary Psychology 7:591
    Sociobiology and Psychology 7:591
Kretschmer, Ernst 13:130
Krewes 3:147–148
Kriging 23:191, 23:193
Krio 16:268
Kripke, Saul 19:664
Kripkensteinian paradox 14:842–843
Kristensen, W. Brede 7:375
Kristeva, Julia 2:359, 19:333–334
  works, *Revolution in Poetic Language* 19:333
Kroeber, Alfred 1:747, 9:358–359
  analysis of fashion trends 9:358–359
  empirical tradition of 8:838
  on Native Americans 1:634, 1:636
  prediction of culture 5:582–583
Kropotkin, Peter 1:680, 19:748, 24:862–863
  on anarchocommunism 4:302
Krueger, Anne 6:277
Krugman, Paul 1:941, 6:602
  on division of labor 6:602, 6:604
Kruskal–Wallis statistic 16:893
k-sample problem, in nonparametric statistics 16:893–895
KSLA *see* Royal Swedish Academy of Agriculture and Forestry (KSLA)
Ku Klux Klan (KKK) 15:590–591
Kubary, John Stanislaw 17:161–162
Kuczynski, R.R. 14:372–373
Kuder Occupational Interest Survey 25:240
Kugler, Franz 2:23
  works, *Handbuch der Kunstgeschichte* (*Handbook of Art History*) 2:23
Kuhn, Thomas S. 5:683–684, 7:875, 8:576, 10:773, 13:133–138, 14:215–216
  and Feyerabend 13:133–134
  on history of science 13:133–134, 13:136–137
  impact on the behavioral and social sciences 13:137–138
  life of 13:133–134
  Neo-Kantianism 13:135
  on normal science 13:133–138
  on paradigms 13:133–138, 17:89
  on philosophy of science 13:133–134, 13:136–137
  on positivist conception of scientific progress 19:126–127
  on relativism 13:135–136
  on scientific revolution 13:134–135
  on sociology of scientific knowledge 13:133–134, 13:137–138
  on theory of scientific development 13:134–137
  works 13:133–134
    Black Body Theory and the Quantum Discontinuity, 1894–1912 13:134
    The Copernican Revolution: Planetary Astronomy in the Development of Western Thought 13:133–134
    The Essential Tension 13:134
    The Plurality of Worlds: An Evolutionary Theory of Scientific Development 13:134
    The Road Since Structure 13:134
    Sources for the History of Quantum Physics 13:133–134

Kuhn, Thomas S. (*continued*)
The Structure of Scientific Revolutions 12:146, 13:133–135, 14:215–216, 21:244, 21:308, 22:310, 22:688–689
Tour de force 1:81
Kuhnian paradigm shift 21:190
Kuhnian program 21:244
Kula 23:1
Kula, Witold, works, *Measures and Men* 11:40
Kula ring 13:139f
anthropology of 13:139–143
alienation 13:142
approaches 13:142
armshell 13:140–142
ceremonial exchange debt 13:142
commodity 13:142
exchange theory 13:139–140
gift 13:140–142
Malinowski on 13:139–142
Mauss on 13:142
necklaces 13:140–142
oversimplifications 13:140–142
present-day kula, features of 13:140
regional differentiation 13:142
Kula system 7:13
Kuláks 19:156
Kullback–Leibler discrepancy 15:656
Kulthum, Um 9:782
Kultur 1:753, 16:233–234
Kulturhistorische Methode 1:901
Kulturkreise 1:901
Kulturkreislehre 3:150
Kulturwissenschaft 11:39
Kulturwissenschaften 10:782
Kulturwissenschaftliche Bibliothek Warburg 2:25
Kumezo, Tsuboi 11:51
works, *Shigaku* kenkyûhô 11:51
Kumho Tire Company v. Carmichael (1999) 8:583–584, 13:767
Kunda, Gideon
on corporate cultures 5:536
on organizations and culture 17:390
Kunggari 3:186
Kunio, Yanagita 12:794
Kunitake, Kume 11:51
Künneth, Walter 18:437–438
Kun-reading 20:2
Kunsthalle, Hamburger 2:21
Kunstkammer 2:23
Kuoppamäki, Riitta, on epistemic cultures 7:876
Kuper, L., works, *Genocide: Its Political Use in the Twentieth Century* 10:7
Kurdish Peace and Democracy Party (BDP) (Turkey) 20:873
Kuru
gene-culture coevolution 9:815
Kussmaul, Adolf 16:683
Kuttab 11:62
Kuwait, legal system in 15:364
Kuzmin, Dmitrii 21:743
Kuzmin, Mikhail 21:741–743
works
The Trout Breaks the Ice 21:742
Wings 21:741
Kuznets, Simon, on national accounting system 16:230–231
Kuznets curves 13:144–149
bell-shaped relationship between per capita income and environmental degradation 13:145–146, 13:145f
environmental Kuznets curve 13:144, 13:144f
empirical analysis 13:146–147
literature 13:147–148
Kwakiutl potlatch 18:715
Kwan Yin (goddess) 2:888
Kwangju Uprising (1980) 2:74
KWIC *see* Key-word-in-context (KWIC) information
Kwon, Heonik 2:74–75
KWRU *see* Kensington Welfare Rights Union (KWRU)
Kyô-machiya houses 5:417
Kyoto Protocol 3:861, 6:835, 7:771, 23:796–797, 23:819–820, 24:620
Kyrgyzstan 10:189

L

LA *see* Lateral amygdala (LA); Linguistic anthropology (LA)
LA21 *see* Local Agenda 21 (LA21)
La belle indifference 11:509
La guerre des deux Frances 8:261
La Mettrie, Julien Offray de 19:410
La nation 16:283
La Poste 18:642
La RED 10:547
La Rochefoucauld, François de, works, *Maximes* 20:769
La Vía Campesina (LVC) 1:461, 10:79, 17:667
La voix du Dahomey (newspaper) 16:268
Label effect, generating 8:890
Labeled Synt-dependencies 6:183
Labeling 22:975, 22:997
antisocial behavior 1:791
Becker on 2:403–404
and delinquency 6:69
extensions in 6:72
and deviance theory 13:150
moral entrepreneur 15:796–797
and psychiatry 22:975
social control in 4:823
and societal reaction theory 13:150–154
Labeling theory, history of 13:150–155
Labia majora 8:879
Labia minora 8:879
Labianca, G., on negative social network ties 5:561
L'abigeato 14:423
Labinger, J.A.
on disunity of sciences 7:875
works
The One Culture 21:310
Labor 7:33–35, 13:156, 13:163–168, 17:664, 22:54
activism, Los Angeles immigrants 24:492–493
agency 13:163–167
agency of, to the forefront 13:163–164
and alienation 1:552–553
anomic division of 1:719
aristocracy 3:740–742
and capital 3:90–91
in chains and networks 13:165
as commodity 5:537
and constitution of social order 25:676–678
control 13:163, 13:165–166
cooperation, in West Africa 1:301
courts 7:571, 14:265
demand
and minimum wages 15:562
of older workers 17:186
and retirement 20:606
welfare reforms 25:517
and discourse 6:503, 6:507
division of *see* Division of labor
enabling structures 13:163–165, 13:167
and ethnicity 13:164
exploitation of 10:209
and family history 8:741–742
food 10:78
global production networks 13:165
global value chains 13:163, 13:165
and knowledge society 13:107
laws 7:572–573
market 7:574–575
mobilization 13:163
peasant agriculture 17:664
place 13:163–165
policy and practice 7:572
power 7:571–572
regimes 13:163, 13:165
regulation of 7:571–576
regulations 7:571
-related surveys 5:847–848
scale 13:163–166
sex work as 19:281–282
share of national income, decline in 22:194
shortage, in socialist societies 22:850–851
and slavery 22:55, 25:669
and social democracy 22:242–243
and social question 22:553–555
space 13:163–166
spatiality (of) struggles 13:165–166
as status 25:676–677
structural constraints 13:164–167
structural constraints and enabling structures 13:164–165
union 7:571, 13:163–166
work anthropology 25:662–667
Labor force 5:769–773, 11:861–862, 21:650–651, 22:683–685
employment sectors 22:684
employment status of social scientists 22:684
professional training 22:684
projections 1:414–415, 1:414t–415t
self-selection in 6:540
social science occupations 22:684

structural changes, and increase in leisure time 13:845
wages 22:684–685
Labor force participation (LFP) 9:23, 9:757–759, 13:190, 13:195, 13:197–198
and ethnic identity 8:106, 8:110
Labor force statistics 5:799–801
Labor geographers 13:164–167
Labor geography 13:163
Labor history 13:169–173, 22:309
Annales school 13:170
anthropology 13:170–171
beginnings 13:169
conceptual plurality 13:170–171
controversies 13:171–172
central categories 13:171
class/gender/ethnicity 13:171
great *versus* small stories 13:172
structure *versus* actor 13:171
demography 13:169–171
economic history 13:169
economics 13:169–171
family 13:169
globalization 13:172
historiography 13:169–171
human science 13:171
interpretation 13:169–170
Labor income/earnings 11:744
inequality 6:767–770
components 6:767–770
in United States 6:767
and value of life 25:15
Labor intensity of children 9:63
Labor leadership, in United States 15:535
Labor market institutions
in transition economies 13:174–182
Labor market outcomes
in transition economies 13:174–182
Labor market policy (LMP) 13:183–189, 21:863
activation 13:185
active 13:184–186
definition 13:183–184
development in time and space 13:185–187
employment subsidies 13:183–185
evaluative and explanatory perspectives on 13:187–188
expenditure 13:187f
flexicurity 13:186
job creation 13:184, 13:187
make work pay 13:185
passive 13:184–186
public employment services 13:184
training 13:184–185, 13:187
unemployment benefit replacement rates 13:186f
welfare reforms 25:518–519
Labor markets 4:935, 6:269, 7:574–575, 12:52, 13:174, 13:199, 13:201–203, 25:339
active policies 6:273–274
and aging 1:413–419
and artists' careers 2:3
and beauty 2:383–391
and career 3:128–132
in developing markets 6:270–271
in East and South Asia 6:270
and ethnic identity 8:109–111
flexibility, and older workers position, relation between 17:187
forces, and income inequality 11:741–742
foreign languages in 13:261–262
global developments and 6:269
and immigration policy 11:665–666
impact of immigration 15:459–460
institutions 13:174
evolution of 13:178
international migration effects on 12:516
and job evaluation 12:816
and language 13:260
in Middle East 6:270
migration impact on 15:437–438
and migration in Europe 15:381, 15:385–386
and minimum wages 15:561–564
mismatch in 15:595–601
networks 16:577–578
empirical test 16:578
model 16:577–578
in North Africa 6:270
outcomes, job satisfaction as driver of 10:519
patterns, and family statuses 8:758–759
in Asia 8:758–759
in Australia 8:758–759
in Canada 8:758
in Europe 8:758–759
in United States 8:758
performance 13:174
baseline estimation results 13:180t
data sources 13:177
determinants of 13:174
empirical results 13:180–181
labor market policies and 13:175
literature on 13:175
measures of 13:181t
methodology 13:178–180
in OECD countries 13:174–175
in transition economies 13:175–176
performance, and obesity 17:70–74
policy changes for 6:271–272
processes, and spatial mismatch 23:158–159
professions in 19:102–103
programs 25:834–835
search 21:334
segmentation 23:158
theories of 13:164
shadow 21:862–863
shortages 17:851
turmoil in 10:200–201
and unemployment 24:733, 24:735–736
and VET 25:232–237, 25:245–250
worker–firm relations 6:272–274
working conditions 25:698–700
Labor migration 17:664
Labor movement 3:742–743
guild traditions in 2:49
Labor parties, and social democracy 22:243
Labor policies, and work-related stress 23:549
Labor power 25:676
Labor process theory (LPT) 14:652–653, 17:342
Labor reallocation 6:270–271
Labor recommodification 18:749
Labor regulation 7:571
empower labor 7:571
wage bargain 7:571
Labor Relations or Employment Relations 'Commission' or Workers' Compensation Board 14:265
Labor scarcity paradox 7:56–57
Labor shedding strategy 13:185
Labor statistics 5:798–802, 22:697
administrative records 5:799
dissemination 5:799–800
households 5:799, 5:801
institutional framework 5:799
international comparisons 5:800–801
International Labor Office 5:800–801
Luxembourg Income Study 5:801
Organization for Economic Cooperation and Development 5:801
Statistical Office of the European Communities (Eurostat) 5:801
US Bureau of Labor Statistics 5:800
International Standard Industrial Classification of all Economic Activities (ISIC) 5:798–799, 5:801
labor force 5:799–801
sources 5:799
standard industrial classification (SIC) 5:798
standard occupational classification (SOC) 5:798–799
topics and classifications 5:798–799
international classifications 5:798–799
international concepts 5:799
national classifications 5:798
wage structure 5:798
Labor supply 13:190–198
collective model 13:191
and earnings inequality 6:767–770
economic development and trends in 9:758
effects of income support policy 18:745
empirical findings 13:194–197
differences in work–wage elasticities 13:197
family labor supply decisions 13:195–196
taxes 13:194–195
work behavior of older people 13:197
empirical regularities 13:190–191
empirical research 13:194
hourly earnings 13:191
hours–wage correlations 13:192, 13:197
household economics 11:224–225
income 13:191–195, 13:197
and income tax 11:744–745
life cycle models of 13:193
nature of available data 13:191–192
of older workers 17:185–186

Labor supply (*continued*)
economic and noneconomic factors 17:187
and retirement 13:197, 20:605
static models of 13:192–193
and taxes 13:193–194
variables classification 13:191
and welfare 25:496
welfare reforms 25:518
work hours 13:190–195, 13:197–198
work–age patterns 13:194
work–wage correlations 13:194
Labor theory 19:167
of exchange value 19:112
of value 3:318, 14:653, 20:662, 25:7–8
and liberalism 14:23
Labor trafficking 10:207–208, 10:210
Labor turnover, panel surveys for 17:465–466
Labor unions 7:571, 7:574, 11:349, 12:372, 13:199–204
as collective actors 13:202–203
in labor market 13:202
polity 13:203
prospects 13:203–204
in workplace 13:202–203
group rights 3:657
organization and union growth 13:200–202
rents due to 25:339
social sciences 13:199
unionism in industrial countries 13:199–200
Labor value 25:8–9
Laboratoire d'Anthropologie Sociale 1:753
Laboratories
analyses, Genographic Project 10:25
assessments, for aggression 1:376
decision research 6:713
definition 8:556
ethnographies 21:264–265
experimental *see* Experimental laboratories
experiments 22:633
organization of 7:874
practices 21:264–265
stressors, interoceptive accuracy during and after 12:617
Laboratory Life 8:577–578, 21:311
Laboratory research
on cardiovascular reactivity 3:118–119
cold pressor 3:118
conceptuality 3:118
distress/anger recall 3:118–119
individual tasks/challenges 3:118–119
mental arithmetic and serial subtraction 3:118
milieu 3:118
poststress recovery phase 3:119
procedural timeline 3:118
psychometric properties 3:118
social interaction tasks 3:118–119
stressful challenge, selection of 3:118
Stroop color-word test and mirror star-tracing procedure 3:118
Laboratory science 21:226–227, 21:237–239
and biomedicine 2:664
Laboratory studies 8:577–579, 13:205
developing agency 13:208–209
ethnomethodology 21:195–197
machines, conventions, culture 13:205–207
of science and technology 7:873, 21:202
Laboratory Temperament Assessment Battery (Lab-TAB) 24:193
Laborer–capitalist cycles 18:582
Labor-intensive agricultural systems 17:677
Labor–labor substitution, in minimum wages 15:562
Labour Force Survey (LFS) 5:847
Labriola, Antonio 22:834
Labrousse, Ernest 1:708–711
Lab-TAB *see* Laboratory Temperament Assessment Battery (Lab-TAB)
LAC *see* Latin America and Caribbean (LAC); Looked-After-Children (LAC)
Lacan, Jacques 13:211–215, 19:504
conceptions of mind 22:762
conept of mirror stage 13:211–213
on demand 13:214–215
on desire 13:211–214
lack theory of 13:213–214
on language 13:212–214
on need 13:214–215
the Other 13:211–215
on phallus 13:213–214
psychoanalysis 15:291
on psychodynamic treatment 22:764
on psychopathology 22:763
on real 13:213–215, 14:200
on signified 13:213
on signifier 13:213–214
on speech 13:212–214
on symbolic realm 13:212–215
on writing 13:214
Lack of expected reactions 6:77
Lack of insight and disorder of judgment 6:77
Lactase nonpersistence 11:294
Lactase persistence 9:815, 11:294
gene-culture coevolution 9:815
Lactase persistence in adulthood 15:340
L'Action 15:135
Lactobacillus lactis Dorner (LLD) 17:586–587
Lacunae and relapses 16:72
LAD *see* Language Acquisition Device (LAD)
Ladder of citizen participation 17:535
Laddering technique 17:762
Ladurie, Le Roy, works, *The Peasants of Languedoc* 1:712
Lady of Cofitachequi 1:633–634
Laesio enormis 3:28
L'Affaire des Foulards 5:494
Lafler v. Cooper (2012) 18:229
Lag filter 24:334
LAGRAMGE (scientific discovery system) 21:295
Lagrangian dynamics, motor control 3:780
Lagrangian function 3:780, 7:114
Lahire, Bernard
on dispositions 22:334–335
on habitus 22:335
on omnivorousness 5:450–451
Laicization 13:216–219
European construction 13:218–219
first threshold of 13:216–217
freedom of belief 13:219
modernity 13:216
problems 13:218–219
religious liberty 13:216–217, 13:219
religious persecution 13:216
religious toleration 13:217
revolution in France 13:216–217
second threshold of 13:217–218
secularization 13:216, 13:218–219
in various countries 13:218
wars of religion 13:216
Western European history 13:219
Laing, R.D., works, *The Divided Self* 19:492
Laissez faire economic system 2:525–526, 5:455, 18:299–301, 18:305, 20:581
capitalism 3:96
economic forces 20:399
ideology 16:462, 16:464
Smith's 22:119–122
liberal capitalism 20:260
and liberal capitalism 20:260
and liberalism 14:23
political philosophy of 11:349
politics 8:568
and sports 9:656
and urban design 24:790
Laissez-faire ideology 16:462, 16:464
Laissez-faire politics 8:568
Laissez-faire racism 18:842, 19:859
Laissez-faire style of leadership 25:639
Laitin civil war 3:716
Lakatos, Imre 18:524–525
on philosophy of economics 7:101
Lake La Yeguada, Panama 17:459
Lake Mohonk conference in 1967 4:436
Lakoff, George 20:652–653
Lala 21:738
Lalor, John Joseph, works, *Cyclopedia of Political Science* 7:591
Lamarck, Jean-Baptiste 7:723, 8:241, 20:287
works, *Zoological Philosophy* 12:175–176
Lamarckian biology 8:241
Lamarckian inheritance 8:246–247
Lamarckism 7:723, 19:802
Darwinism and 8:383
Lamba Legal Defense 10:792
Lambda (Grupo Lambda de Liberación Homosexual) 13:860
Lambda-abstraction 15:752–753
Lambek, Michael, on history 1:748
Lamination, cortical 3:338–340
Lamont, Michèle 4:760
on cultural capital 5:376, 5:379
on epistemic cultures 7:877
works
Cultivating Differences 5:606
Lampão da Esquina 13:860

Lamprecht, Karl, works, *Deutsche Geschichte* 11:75–76
Lamprecht dispute 10:874–875
Lancaster, Roger, works, *Life is Hard: Machismo, Danger and the, imacy of Power in Nicaragua* 8:902
Lancet Commission on the Future of Health Professions 16:908
Land change 13:220–223
  integrative study of 13:220
  land cover dynamics 13:221
  land use dynamics 13:221–222
  models 13:222
  remote sensing 13:221–222
  satellite instruments 13:221
  science, origins and motivations of 13:220–221
Land competition 13:241
Land Conservation and Development Act 1973 10:450
Land cover 13:220
  GIS/remote sensing for 10:67
Land degradation 18:352
  and political ecology 18:352
Land ethic 1:475
Land grabbing 1:460, 13:235, 17:581–582
  in West Africa 1:303
Land politics, in China 24:976
Land reform laws 12:751
Land reforms 1:458, 14:422–423
Land restitution 13:234–235
Land rights, of Australian aboriginals 2:237
Land system 13:239
Land tenure 1:478, 13:231–237
  advocacy and practice 13:236
  cash rental 1:478
  historical background 13:231–232
    evolutionist political philosophy 13:231
    persons, things, and systems 13:231–232
    structure and process of inequity 13:232
  land restitution 13:234–235
  metaphor and rhetoric 13:232
  minority land rights 13:234
  mobility 13:234
  new frontiers 13:236
  and property 19:153–157
  reform 13:232–236
    bureaucratic simplification 13:233
    collectivization and villagization 13:233
    contested formalization of land claims 13:235
    displacement, agrarian encroachments, and land grabbing 13:235
    land rights and tenure security in postsocialist settings 13:234–235
    special issues in irrigated and flood-prone lands 13:235–236
    titling and individuation 13:233–234
  sharecropping 1:478
  sovereignty 13:231–233
  struggles over land 13:231
  territoriality 13:236
  territory 13:231–232, 13:234, 13:236
Land use 13:220, 15:502, 16:337, 20:149, 22:256, 24:612–613
  of Africa 1:293
  energy flow analysis 7:627
  and food production 22:256
  GIS/remote sensing for 10:67
  and sustainable transportation 23:847–848
  in urban planning 19:822–823
  urban policy as 24:801, 24:909–910
Land Use Board of Appeals (LUBA) 10:450
Land use change
  Anthropocene 1:726–727
Land use datasets 13:241
Land use indicators 13:238–244
  agriculture 13:239–240, 13:242
  area demand 13:242
  area-related indicators 13:238–239
  biodiversity 13:238–241, 13:243
  biomass 13:241–242, 13:242f
  carbon stocks 13:241
  DPSIR scheme 13:242–243, 13:243f
  ecological footprint 13:242
  ecological integrity 13:241
  ecosystem services 13:239–240, 13:243
  energy return on investment 13:241
  food security 13:241, 13:243
  forestry 13:239–240, 13:242–243
  hemeroby 13:240–241
  human appropriation of net primary production 13:241–242
  industrial society 13:239–240
  intensity indicators 13:239–241, 13:240f
    at ecosystem level 13:240–241, 13:240f
    input intensity 13:240
    integrated indicators 13:241
    output intensity 13:239–240
  land competition 13:241
  land system 13:239
  landscape heterogeneity 13:239, 13:243
  metrics 13:238–241
  naturalness 13:240–241
  net primary production 13:240–242
  production and consumption indicators 13:241–242
  production-consumption interlinkages 13:242f
  spatial reference of land use information 13:241
  structural landscape elements 13:239
  sustainability 13:238, 13:242–243
  systems 13:242–243
Land use models 24:619–620, 24:792
Land use-transportation-emissions models 24:829
Landauer, Gustav, on queer anarchism 19:747–748
Land-based accumulation, in China 24:975
Land-based trading 18:714
Landes' thesis 19:613
Landesque capital, production 12:329–330
Land-Grant System 20:814–816, 20:818
Landholding 13:231–232
Land–industry cleavage 20:746
Landlordism 11:234
Landmark localization 8:658–659
Landmarking 20:883
Landownership, in China 5:598
Landry, Charles 24:867
Lands 13:224–226, 13:229
  of Africa 1:293
  in China 2:66, 2:68
  connection of indigenous peoples to 11:804
  conversions, types and quantity of 24:827
  and indigeneity 11:757–760
  of pre–Columbian South America 23:43–44
  rights *see* Right to land
Landsat Data Continuity Mission (LDCM) 20:423
Landsat satellite 24:827
  Landsat 2 20:422
  Landsat 4 20:422
  Landsat 7 20:423
  Landsat 8 20:423
Landscape 11:19
Landscape imagery, and colonial environmental history 7:752–754
Landscape model, of discourse 12:41
Landscape spatial pattern 10:65
Landscapes 10:70–73, 13:224–230, 15:503, 23:40, 25:373
  archaeology 21:654–658
    chronological constraints and considerations 21:656–657
    contemporary research strategies and ongoing debates 21:656–657
    early foundations for archaeological survey in Americas and England 21:654
    environmental constraints 21:656
    historical background 21:654–656
    interpretation of regional data 21:657
    population estimation 21:657
    studies, critical role of 21:657–658
    survey methodologies and sampling 21:656
  Berkeley school studies 13:229
  in China, remaking 24:976–977
  and cultural geography 5:409–412
  disciplinary hegemony, collapse of 13:228
  heterogeneity 13:239, 13:243
  historical geography 11:19
  and knowledge translation 21:184
  meanings of 13:225–227
  of North America 17:15–20
  perception 13:229–230
  in post-world war II geography 13:228–229
Landscaping 22:351
  defining and labeling issue 22:351
  limited choices/options 22:351
Lane, Allen 9:372
Lane, Robert E., on knowledge society 13:105–106
Lang, Fritz
  films
    M 9:168
    Metropolis 9:167
Lange, A. 9:647
Lange, O. 17:504
Langer, Susanne, works, *Philosophy in a New Key* 2:273
Langer, Suzanne 1:747

Langlois, Charles-Victor 1:909
Langscape (e-magazine) 1:224
Language Acquisition Device (LAD) 2:877
Language Acquisition Support System (LASS) 2:877
Language acquisitions 1:536, 13:245–250, 13:319, 13:321–323
  articulation 13:246
  audition 13:245–246
  CDS identifying features that impact 3:400–401
  emergentist accounts 13:249
  first words 13:246–247
  grammar learning 2:651–652
  and maternal linguistic input, associations between 3:399–400
  neural substrate for language development 2:652–653
  second/multiple language 16:57–58
  semantics 13:247
  sound discrimination 2:650–651
  and unconscious mental activity 24:723
  word combinations 13:247–249
    logical problem 13:248–249
    missing glue 13:248
    productivity 13:248
  word learning 2:651
Language analysis, by Jeremy 2:525, 2:527
Language attitudes 22:949–950
Language awareness 2:583–584
Language centers 16:682–684
Language change 1:959
  in functionalist grammar 9:473–474
  and serial verb constructions 21:647
Language characteristics after cerebellar injuries in children 16:589
Language competence 4:296–297
Language comprehension 6:76, 23:255
  embodied 13:374–375
  and embodiment cognition 7:421–423
  figurative 13:372–373
  nonverbal information in 13:375
Language contact 3:918–920, 3:922–923, 13:300–306
  Albanian 13:303–305
  Balkans 13:303–306
  bilingualism 13:301–302, 13:304–305
  borrowing 13:302–306
  convergence 13:305
  effects 13:305
  evolution of language 8:369
  family tree 13:300–301
  Greek 13:300–301, 13:303–305
  interference 13:301–302, 13:305
  language change 13:301, 13:305–306
  loanword 13:302–305
  manifestations 13:302–304
  mechanisms 13:301–302
  sprachbund 13:305
  substratum 13:301–302
  transfer 13:301–302, 13:304–305
  Turkish 13:304–305
  wave theory 13:300–301
Language death 14:841
Language description 14:184
Language development 2:290, 12:29–30, 13:307
  after injury to white matter pathways 16:589
  and child care 3:392
  cognitive approaches 13:310–311
  evolutionary theories of 13:319–323
    adaptationist account 13:321
    communication 13:319–323
    cultural evolution 13:322
    cultural transmission 13:322
    language acquisition 13:319, 13:321–323
    nativism and language learning evolution 13:319–321
    natural selection 13:319–320, 13:322–323
    social learning 13:321–323
    statistical learning and evolution of learnability 13:321–322
    Universal Grammar (UG) 13:319–322
  frequency 13:314–315
  genetic abnormalities and 16:589
  in humans 11:256
  influence of child-directed speech on
  see also Child-directed speech (CDS)
  internal factors 13:314–316
    reflexive pronouns 13:315–316
    scope preferences in first language learners of English and Korean 13:314–315
  lexical stress 13:309–310
  multilingual *see* Multilingual language development
  nativist approaches 13:307
  neural basis of 16:586–592
  neurophysiological markers of 16:708–713
    event-related potentials 16:708, 16:710
    evoked potential 16:708
    mismatch negativity 16:708–709
    N400 16:710–711
    P300 16:709–710
    tone burst 16:708–710
    verbal sound 16:708
  prelinguistic language processing in infants 16:589–590
  during preoperational stage 18:128
  processing cost 13:313–317
  relative clauses 13:314, 13:316
  role of child-directed speech 13:308
  role of input 13:313–314
  statistical learning 13:308–309
  uniformity of 13:316
  universal trends in 9:202–203
  Williams syndrome 25:580–581
Language disorders 1:818, 3:792–798
  assessment of aphasia 3:793
    bedside and screening tests 3:793
    cognitive neuropsychological model of spoken picture naming 3:793f
    comprehensive examinations and aphasia batteries 3:793–794
    specific aspects of language 3:794
  assessment of functional communication 3:794
  errorless learning for people with 2:816
  reference models 3:792–793
  speech, language, reading and writing impairment in children 3:794–795
    assessment of reading and writing 3:796
    comprehensive examinations and batteries 3:795
    screening tests 3:795
    specific aspects of language 3:795–796
  speech and language impairment in special populations 3:796–797
  treatment and remediation of 13:329–336
    approaches to aphasia therapy 13:330–333
    brain plasticity 13:329, 13:333
    brain reorganization 13:329–331
    cerebrovascular accident 13:329–330
    efficacy of aphasia therapy 13:330–333
    lesion studies 13:329–330, 13:332
    neuroimaging 13:329–330
    prognosis 13:329–330
    recovery after stroke 13:329–330
Language disturbance, aphasic signs 6:76
Language documentation 14:184–187
Language evolution 8:359–360
  experiment 8:360
Language families 6:506
  agriculture 13:337, 13:340–342
  archaeology and history of 13:337–343
  Austronesians 13:337–340, 13:342
  comparative method 11:821–822
  and comparative method 4:423–424
  correlation with archaeological prehistory 13:340
  dispersal histories with claimed archaeological correlations 13:340–342
  evidence 11:820–821
  genealogical relationship 11:821
  human migration 13:337, 13:340–341
  Indo-European languages 13:337–343
  inheritance 11:820
  origin 13:340–342
Language functions, recovery of 20:57
Language games 6:511, 18:795
  of power 18:768
  Wittgenstein's 10:786–787, 25:612–613
Language independent hypothesis 6:736
Language learning
  evolution 13:319–321
  theories of 9:327–328
Language learning impairment (LLI) 13:344–352
  auditory processing 13:346–347
  computerized training 13:349
  developmental dyslexia 13:344
  diagnosis 13:344–345
  electroencephalography 13:348
  etiology 13:345–348
    genetic studies 13:347
    linguistic studies 13:345–346
    neurobiological studies 13:347–348
    perceptual/cognitive studies 13:346
  goals for future research 13:349–350
  linguistic processing 13:346

magnetoencephalography 13:348
neuroimaging 13:347
neuroplasticity 13:349
phonological processing 13:345–346
remedial intervention 13:348–349
specific language impairment 13:344, 13:347–348
temporal processing 13:347–348
working memory 13:346, 13:348–349
Language loss 16:60
Language maintenance 2:581–584
Language of emotions 5:543
Language of rights 16:322
Language of thought 4:85, 4:87
and mental representation 20:70
Language planning 13:353–355
Language policies 22:952–953
acquisition planning 13:354
classic language planning 13:353–354
corpus planning 13:353
for cultural autonomy 13:358
for exclusion 13:357–358
goals and motivations for 13:355–358
ideologies 13:355
for inclusion 13:358
for independence 13:356–357
for interdependence 13:357
language planning 13:354–355
linguistic and social goals 13:356
management 13:355
practices 13:354–355
status planning 13:354
Language processes 13:369–374, 21:914
embodied language comprehension 13:374–375
frontal and parietal lobes involved 13:369–370
functional and neuroanatomic organization of 24:535–536
network organization 24:535–536
neuroanatomic organization 24:536
within-stage processing 24:536
functional brain imaging of 9:476–513
interhemispheric interaction in 12:416–424
nonverbal information in language comprehension 13:375
phonological processing 13:369–370
temporal regions involved in 13:370
Language processing
functional magnetic resonance imaging of 13:368–380
Language production 23:229, 23:233
Language replacement processes 13:341
Language resources 4:951
Language sciences 14:215, 14:217–218
Language studies 1:933–935
language subgroups 11:820
Language technology 4:949
Language training, migrants 15:417
Language transmission and linguistic structure 8:357–361
Language universals 14:209–210, 14:533–534, 14:536
in semantic parameters in subdomains 23:90–91
Whorf on 25:576, 25:578
Language use 18:799–800
gender differences in 9:754
and interpellation of the human subject 14:200
Language variation 6:349, 6:351
Language-brain theory-unification 14:177
Language-dominant hemisphere 2:807–808
Language-in-education planning (LiEP) 13:354
Language-oriented discourse studies 6:510–514
Languages, 2:882–883, 4:17–18, 4:844, 8:364–377, 9:698–703, 13:272, 14:202, 14:652, 16:224, 19:10–11, 21:932, 25:288, 25:322–323, 25:325–326
acquisition 21:352–353
as action 18:798
and administration 1:107
agent-based modeling 8:369
analysis of conversation and discourse 9:700–701
analytic philosophy of 10:786–787
applications of language phylogenies 8:373–375
investigating cultural evolution 8:374–375
population dispersal hypotheses, testing 8:373–374
quantifying processes of language change 8:374
as the archives of history 8:370
and asymmetry of body/brain 2:115–116, 2:120, 2:125–128
of Australian aboriginals 2:236–237
Austronesian 8:370, 8:373–375
-based nationality 20:339
biological and linguistic evolution, parallel between 8:367–369
change 8:366–369, 13:301, 13:305–306, 22:949–950
evolution of language 8:366–369
in China 2:66, 2:68
choice, social determinants of 22:949–950
cognate 8:366–367
and cognition 4:20
cognition determination 1:767
comparative study 25:288–289
comprehension 21:353
computational historical linguistics 8:375
constitutive conception of 4:790–791
contact 8:369, 22:952
contact, and cultural contact 13:342–343
continental philosophy of 4:790–795
Hamann–Herder–Humboldt tradition 4:790–793
in twentieth century 4:793–795
controversies 8:375
in critical social work 5:289–290
cultural evolution 8:367–368, 8:370, 8:372
investigating 8:374–375
deixis *see* Deixis
designative function of 4:791–794
development 6:728–729, 21:349–350
and dichotic listening 6:357–367
and discourse 6:498–501, 6:504–505
and dispute transformation 6:562–564
dominance 10:505–506
evolutionary views 10:506
dynamic dimension 10:109
ecology of 13:354
effects in literature 13:270
English *see* English language
essentialism 8:365, 8:368
and ethnicity 13:251–255
boundary marker 13:252–253
collective identity 13:251
determinism 13:252–253
diglossia 13:253
diversity 13:252–254
Herder on 13:252–253
Humboldt on 13:252–253
identification 13:253–254
minority languages 13:251–253
nationalism 13:252–254
switching 13:253
and thought 13:252–253
Whorf on 13:252–253
writing systems 13:253
ethnomethodology 8:203–204
ethnopsychology 25:793
ethology 8:364–365
and evolution 11:329–334
asymmetrical brain 11:329–330
communication 11:329, 11:331–332
genetic and physical considerations 11:332–333
gestural theory 11:331–332
hominid 11:329
lateralization 11:329–331
mirror neuron system 11:330–331
speech 11:329–333
vocalizations 11:329–333
evolutionary origin of 8:364–365
experimental study of 4:10, 4:12
faculty of 13:75
families
cultural evolution 5:394–397
feminist linguistics 9:700
and gender 9:785–786, 9:795–799, 13:256–259, 22:949–950, 22:952–953
historical background 8:370–372
Bayesian phylogenetic methods 8:372–373
evolutionary biology and phylogenetic revolution 8:371
historical linguistics and rise and fall of lexicostatistics 8:370–371
modern phylogenetics 8:371–372
historical linguistics 8:365–367
histories of Roma people 20:749–750
and history 11:105
human population history 8:367
human prehistory 8:375
hypercorrectness 9:698
idealistic conception of 4:794

Languages, (*continued*)
ideologies 13:355
and infant cognitive development 4:58–59, 4:61–63
and inferences 19:394–395, 19:397–398
inhabiting 10:109–110
instrumental conception of 4:791, 4:793–794
intentionality as foundational feature of 12:347–348
intentionality in 12:347–351
and intracarotid sodium amytal testing
alternative methodologies 12:693
lateralization 12:693–694
study 12:692–693
isms 12:759, 12:762–763
and joint attention 12:844–849
and labor market 13:260
and labor market success 13:260–265
Lacan on 13:212–214
lateralization
development of 10:507–508
pathology 10:511–513
lateralization of 13:398
and law 13:536–537
Lévi-Strauss on 13:905–907, 13:909
lexical consolidation 13:919–924
lineage 8:365–368
linguistics 8:370–371, 8:374–375
and literature 13:266–271
and logic 14:295–296
management 13:355
mathematical psychology 14:814
meme 8:368
memory and, in cognitive science 4:107–108
in MENA 16:361–362
mental state, ToM 24:269
in Mexico 1:629
and migration 15:429
and mirror neurons 15:583–585
gestural origins of language 15:583–585
motor theory of speech perception 15:583
models of semantic processing 21:572–573
mother–daughter interaction 9:700–701
motherese 9:698
national 16:262–264
and neurolinguistics 16:687
in Oceania 17:159–160
origin of 20:315
pacific settlement 8:373
paradigmatic and syntagmatic axes of 15:291
philosophy of 4:548
Searle's 21:340–341
phylogeny 8:365–366
phylogeny reconstruction 8:365–367
and play of differences 15:291
and poetic structure 13:272–280
politeness in 18:326–330
political 18:488–490
population thinking 8:368
practices 13:354–355
pretend play and 18:873
processes in delirium and dementia 13:360–367
processing 14:177, 14:181, 21:356
production 21:349–350
proficiency 21:351–353
Putnam on 19:664
queer 19:755–756
religious 20:310
replicator 8:367–368
representation in brain 2:576
of rights 13:536
Schütz on 21:148
and sexism 9:699–700
Skinner on 22:27
as social action 18:788–789
social and political organization of 22:952
and social class 13:281–287
social cognition 8:364–365
and social media use 24:149–150
social representations 20:444–448
in social sciences 6:486
and society 13:288–292
discourse particles and interjections 13:290
discourse styles 13:291–292
fatalism in Russian grammar 13:289–290
grammar 13:289–290
linguistic routines 13:290
personal autonomy and English wh-imperative 13:290
speech genres 13:290–291
speech styles 13:291
words and phraseology 13:288–289
sociolinguistics 8:368
as source of information on human prehistory 13:338–340
of South Asia 2:76–77
of Southern Africa 23:80
pre-Columbian 23:43–44
spatial *see* Spatial language
and spatial cognition development 23:123
spatial communication via 23:114
species 8:364–365, 8:367–368
-specific differences 4:831
speech community 8:368–369
static dimension 10:109
statistical analysis 23:379
study of land and tenure 13:236
subdivisions of
subversive nature of 14:200
theoretical developments 9:699
theories, Chomsky's 4:104
and thought 13:293–299, 20:903–906
categorical perception 13:294
color terms influence color perception and memory 13:294–295
conceptual metaphor 13:297–298
grouping effect 13:297
linguistic differences in temporal concepts 13:297–298
linguistic diversity 13:293–295
linguistic relativism 13:293
numerals 13:296–297
relational thinking 13:295
spatial terms influence nonverbal spatial tasks 13:295–296
time conceptions 13:294–295, 13:297–298
of time 24:407
transfer 21:349–350
use of biological methods for classification of 8:365–367
variation 22:949–950
as vehicle of social inequalities 22:327
working across 4:289
world languages 3:310
world-disclosing function of 4:790, 4:792–794
and worldviews 25:576
Language-specific features 3:511
Lanham Trademark Act 13:768–769
L'Année sociologique (journal) 1:767, 6:700–701, 6:706–707, 21:964–965
Lao Tze 4:312
Laos 23:63
fertility transition in 9:84–88
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
LAPACK (software) 7:2
Lapdancing 8:5
LaPiere study, prejudice 18:839–840
Lapita culture 17:446–447
Melanesia before 2:230
pottery 17:159–160
Laplace, Pierre Simon, Marquis de 13:381–384
Laplace, Pierre-Simon, Marquis de 7:116, 9:624, 13:381
on analytical theory of probability 13:382–383
on application of probability calculus to traditional fields 13:383
career of 13:383
on celestial mechanics 13:382
constant causes, theory of 13:383–384
decision-making aspect in work of 13:384
decision-theoretic approach of 9:625
and Gaussian law 23:424
influence in modern times 13:383–384
marriage and family of 13:383
on method of situation 13:382
program of studies 13:381–382
ratio estimation method of 20:875
on Rule of Succession 12:800
on theory of chance 13:382
on theory of chance and probability 13:384
works
Philosophical Essay on Probability 13:383
Principle of Insufficient Reason 19:39
Laplace approximation 9:849
Laplace integration 8:523–524
Laplace transforms 23:471
Laplace uncertainty principle 12:108, 19:40
Laplacian determinism 13:383–384
Laplanche, Jean 19:334
works
The Language of Psychoanalysis 19:334
Life and Death in Psychoanalysis 19:334

LAPOP *see* Latin American Public Opinion Project (LAPOP)
Laptop design 15:720
Laqueur, Thomas W., works, *Making Sex* 2:730
Larceny 15:592
L'Arche movement 24:275
Lareau, Annette, on cultural capital 5:376–380
Large Area Crop Inventory Experiment 20:422
Large networks 16:520
Large states 22:108–112
Large vocabulary continuous speech recognition (LVCSR) task 23:262
Large-scale asset purchases 15:729–732
  portfolio losses 15:731–732
  price inflation 15:731
Large-scale coding of objects 16:631
Large-scale cooperation 11:262–265
  cultural group selection 11:264–265
  indirect reciprocity 11:262–263
  punishment 11:263–264
Large-scale ID 12:196
Large-scale social experimentation 8:531–534
  alternatives to social experiments 8:533
  design issues 8:533–534
  health insurance experiment 8:532
  Housing Assistance Supply Experiment (HASE) 8:532–533
  JOBSTART 8:532
  New Jersey Negative Income Tax Experiment 8:531–532
  risks 8:533
Large-scale social surveys 13:385–390
  advantages 13:385
  background 13:385
  complex multistage sampling designs 13:386
  confidence intervals 13:388
  missing data 13:389
  recommendations 13:385
  standard errors 13:387–388
  statistical modeling with survey data 13:388–389
  survey weights 13:386–387
Large-scale systems, knowledge representation architectures for 13:103
Large-scale tests 3:819
Largest system of numbers 17:44–45
Larike trial 17:44
Larrey, Dominique 16:682
Larry Bird exemption 23:322
Larson, Magali Sarfatti, works, *The Rise of Professionalism* 18:386
Larson, Richard K., works, *The Evolution of Human Languages: Biolinguistic Perspectives* 25:344–345
Las Casas, Bartholome de 1:223
Las Vegas 3:662–663
LASA *see* Latin American Studies Association (LASA)
Lasch, Christopher, works, *The Culture of Narcissism* 1:620
Laser scanning, as geospatial data collection platform 10:98–99, 10:102
Lash, Scott 12:86–87
Lashley, Karl Spencer 10:732, 12:703, 13:391–393, 19:415
  biography of 13:391
  brain substrates of learning and memory 13:392
  college education of 13:391
  cortical field-like interference patterns, theory of 13:392
  honors of 13:393
  on localization of neocortex functions 13:392
  on mass action and equipotentiality 13:392
  on nature of consciousness 13:393
  positive contributions of 13:392
  on research on memory 13:392
  schooling of 13:391
  works
    Brain Mechanisms of Intelligence 13:392
Lasker, S., on LGBTI commercial sex 13:885
Laski, Harold 18:467
  works, *Authority in the Modern State* 2:270
LASS *see* Language Acquisition Support System (LASS)
Lassalle, Ferdinand 22:838
LASSI *see* Learning and Study Strategies Inventory (LASSI)
Lasswell, Harold 2:271, 3:614, 9:321–322, 18:288, 25:398
  works, *World Handbook of Political and Social Indicators* 6:267
Last Resort, principle of 16:5–6
Last-in-first-out *see* LIFO (last-in-first-out)
Latchkey children 14:607
LATE *see* Local average treatment estimates (LATE)
Late adult learners of second language 21:354–356
Late child learners of second language 21:354
Late imperial age, East Asia legal traditions during 6:780–781
Late insertion 15:824
Late Judaism 11:71
Late liberalism 14:15–16
Late life 11:267, 11:269–270
Late Middle Ages, Reformation and confessionalization during 20:100–103
Late Pliocene *Australopithecus* 18:891
Late positive potential (LPP) 10:703
Late posterior negativity effect 23:32–33
Late short-term memory (lSTM) 15:157–158
Late starters 14:60
Late-capitalism 18:668, 18:670, 18:673–674
Late-maturing oligodendrocytes myelinate 1:595
Latency
  additive factors method 4:105
  see also Process accidents; Response time (RT)
Latent categorical variable 11:753
Latent class analysis 3:801, 3:808, 8:341, 8:694, 16:135–136, 19:419, 19:436, 23:619
Latent continuum 19:436
Latent curves 10:444–445
Latent demand, and congestion 24:632–633
Latent Dirichlet allocation (LDA) 9:364–365
Latent factor 14:853
Latent growth curve models (LGCM) 14:62–63
Latent growth modeling (LGM) 14:328, 14:330, 14:337
Latent inhibition, and creativity 17:771
Latent mental network theory 12:80
Latent opinion 2:195
Latent pattern maintenance, AGIL schema 1:381
Latent profile analysis 8:694
Latent risk model 24:357
Latent semantic analysis (LSA) 21:581–582, 25:229
Latent structure 8:691, 8:698–702, 21:4–5
  conceptual framework 8:692–693
  estimation of parameters 8:695
  future developments 8:696
  historical background 8:695–696
  idea 8:691
  relationships between latent variables 8:694–695
  structure of family 8:693
Latent structure analysis 8:686–690
Latent trait models 8:694
  extreme response style 23:759
Latent trait theory 19:438
Latent variable analysis 8:686–687
Latent variables 3:788, 8:340–342, 8:691, 11:750–752, 13:394–397, 19:418–419, 19:436, 19:438, 20:239, 23:406–407, 24:642
  coefficients 13:394–395
  controlling for 4:841–843
    instrumental variable method 4:842
    mediator variable method 4:842–843
  covariance matrix 13:395
  error variables 13:394
  for estimators 13:394
  generalized models in personality psychology 8:704–705
  graphical models 13:395
  linear structural equation models 13:394–396, 13:394f
  measured variables 13:394–396
  measurement error 13:396
  methods, classification of 8:693, 8:693t
  models, construction of 13:396–397
  parameters 13:395–396
  parsimonious models 13:396, 13:396f
  preference models with
  see also Preference models

Latent variables (*continued*)
relationships 8:694–695
representing causal relationships between variables 13:396, 13:396f
Later life, sexuality in 21:844
Later life-course populations, and retirement 20:594
Lateral amygdala (LA) 8:870
Lateral geniculate nucleus (LGN) 25:159, 25:184
Lateral intraparietal (LIP) area 5:23, 5:27
Lateral lemiscus 10:482
Lateral occipital (LO) area 17:129–131, 25:202
Lateral occipital complex (LOC) 16:630–631
Lateral prefrontal cortex (LPFC) 18:817, 18:836
dorsal areas 18:818
ventral areas 18:818
Lateral premotor cortex (LPMC) 8:668
Laterality
and asymmetry of body/brain 2:115–116
and hysteria 11:508
neurolinguistics 16:686
Laterality quotient (LQ) 2:807
Lateralization 12:416, 13:398–401, 16:586–590
across life span 13:407–408
in animals 2:799–805
aggression 2:800–801
categorization 2:800–801
definition 2:799
development 2:802–803
emotion 2:801, 2:804
face recognition 2:801
function 2:803
functions of left and right hemispheres 2:801f
history 2:799
in invertebrates 2:802
limb preference 2:800
memory 2:802–804
response to predators 2:800, 2:802–803
revealing methods 2:799–800
spatial behavior 2:799
in vertebratesevolution 2:800–802
in cerebral cortex 3:334–338
of cognitive abilities, humans 23:280
genetics and 13:406–407
individual-level 2:799
of language 13:398
and language evolution 11:329–331
lexicosemantic processing and 13:400–401
in males and females 13:406
of multilingual populations 13:405
in pathological populations 13:405–406
perceptual, phonological and motor processing 13:399
at pragmatic level 13:401–402
in presurgical patients 13:406
at sentence level 13:401
of sign-language users 13:404–405
in special population 13:404–405
text and discourse comprehension, impact on 13:402
Later-born children, value of 3:499
Latifondi 14:423
Latifundia (large landed estates) 14:422–423
Latin America 7:36, 14:282
and Cardoso, Fernando Henrique 3:122
carnival in 3:148
child mortality in 12:21
Christianity in 3:534–539
counterinsurgencies in 5:101–102
criminology in 5:239–240
dependency 3:124–125
development of 3:123–124
dissemination and communication in 22:628
distribution of centers and universities 22:628
economic development 3:123–124
ethnic composition in United States 18:557t
eugenics in 8:246
fertility transition in 9:68
condom 9:70–71
contraception 9:69–71, 9:70f, 9:71t
fertility trends 9:68–69, 9:68f–69f
first birth, in Latin America and Caribbean 9:69–70, 9:70t
marriage onset, in Latin America and Caribbean 9:70t
proximate determinates 9:69–71
reproductive rights 9:71
sexual debut, in Latin America and Caribbean 9:69–70, 9:70t
unintended pregnancy 9:71
immigration control 8:248
legal complex 18:387–389
legal systems 13:410
Anglo-American law 13:412–413
Civil Law 13:410–411
constitutionalism 13:411–412
International Human Rights Law 13:412–413
legal actors 13:412
legal convergence 13:412–413
legal education 13:412
legal pluralism 13:411
Private Law 13:410–413
Public Law 13:410–411
lesbian/gay activism in 13:858–863
early activism 13:859–860
globalization and economic liberalism 13:861–862
identity formation 13:860–861
influence of HIV/AIDS 13:861
regional understandings of sexuality 13:858–859
state-directed activism 13:862
migration to and from 15:462–465
early migration patterns 15:462–463
European immigration into 15:462
Mexico–US migration 15:463–464
non-Mexican migration 15:464–465
political liberalism 18:387–389
populism 18:609–610
positive and negative eugenics in 8:249–251
postcolonial
confederation 4:574
social change in 3:125
social development 3:123–124
trafficking of women and children in 24:512–516
barriers to addressing 24:514
Central America 24:513–514
commercial sexual exploitation 24:512–515
consequences 24:514
gender violence 24:513, 24:515
human rights 24:512, 24:514–515
nongovernment organizations 24:512, 24:514–515
purpose 24:512–513
recruitment 24:512
risk factors 24:513
role of social work 24:514–515
Latin America and Caribbean (LAC)
educational systems 7:178–184
assessments 7:180–181
complexity and heterogeneity 7:178
development 7:178–179, 7:181–183
educational reform 7:179, 7:183–184
epistemological traditions 7:182
higher education 7:178–184
illiteracy 7:178–179
investment in R&D in region and educational research indicators 7:181
primary education 7:178–181
research 7:181–184
research by level 7:183–184
research traditions 7:181–183
science 7:180–181
secondary education 7:178–181, 7:183
trends and persistent issues 7:178–180
Latin American Council of Social Sciences (CLACSO) 22:143–144
Latin American eugenics 8:229–230
Latin American Faculty of Social Sciences (FLACSO) 22:143–144
Latin American Laboratory for Assessment of the Quality of Education (LLECE) 7:180–181
Latin American military dictatorships 22:143–145
politics and SBS before 1970s 22:143–144
1970s' dictatorships and their legacy for SBS 22:144–145
Latin American Public Opinion Project (LAPOP) 5:849
Latin American School of Social Sciences (FLACSO) 3:123, 7:183
Latin American Studies Association (LASA) 7:182
Latin Americans
gay and bisexuals, in US gay culture 9:628–631
health 9:630
Latin cultural issues and gay identity 9:628–629

Latino response 9:630–631
stereotyping and racism 9:629–630
use of term 9:628
Latin Countries 11:861
Latin language, morphological case system in 3:183–186, 3:184t
Latin squares 5:352–353, 8:539
LATINDEX 22:628
Latino paradox 10:572, 12:142–143, 19:813–817
Latino Threat Narrative 11:911
Latinobarometer/latinobarometro 5:793, 5:849
Latitude of acceptance (LOA) 22:352
Latitude of rejection (LOR) 22:352
Latitudes of noncommitment (LNC) 22:352
Latium 6:371
Latkin, Carl A., on alcohol consumption 16:557–558
Latour, Bruno 1:238, 5:524, 8:579, 13:205–206, 13:414–419, 16:562–563, 21:311, 22:439, 22:829–831
on actants 13:415–417
on actor-network-theory 13:415–417
alliances and impacts of 13:417
on collectives 13:414–418
on constructionism 13:414–417
on democratic publics 5:525–526
on ethnomethodology 22:830
on fetishism 9:106, 9:109–110
on hybridity 13:414–418
on institutions 17:392
on intermediaries 2:873
on laboratory work 5:524
on metaphysics and methods 13:415–417
on moderns 13:414–416
on ontology 13:414–417
on problem of constructivism 8:579
on realism 13:416–418
on semiotic analysis 8:579
translation 13:415–417
works
Knowledge and Reflexivity 20:94
Laboratory Life: The Construction of Scientific Facts 5:523, 8:577–579, 21:311, 24:136
On the Modern Cult of the Factish Gods 9:109
Reassembling the Social 5:525, 13:414
Science in Action 8:579
The Social Construction of Scientific Facts 8:577–578, 21:311
We Have Never Been Modern 5:524–525, 9:109, 11:779
Latter-Day Saints 20:302
Lattices 17:527
Latvian Green Party 10:380
Laudanum 17:239
Laughlin, Harry H. 8:224, 8:226
Laughter, and humor 11:416–419
L'Aumisme Religion Universelle de l'Unité des Visages de Dieu 16:806
Laura Spellman Rockefeller Memorial (LSRM) 3:4
Lauritsen, John, on Russian sexual life 21:740–741
Lauter, Paul 3:528
Lavaan (software) 19:421
Lave, J., works, *Situated Learning* 7:145–146
Lavisse, Ernest 1:709
Lavoisier, Antoine, hemical revolution 3:372
Law 3:266–273, 3:771, 13:483–488, 13:557–564
versus accident 3:267
and adoption 1:140
affect 13:463
Albert on 1:497–498
of ambassadors and diplomacy 12:507–508
anthropology of 13:535–541
colonialism 13:539–540
culture and language 13:536–537
dispute processes 13:536
historical developments 13:535–536
human rights 13:536–537, 13:539–540
legal consciousness and identity 13:538
legal pluralism 13:539–540
power 13:537–538
rule of law 13:537, 13:539
society 13:535–536, 13:539–540
transitional justice 13:539
archaeology politics 1:885–889
of association 13:701
Australian aboriginal worldview 2:236
autopoiesis in 13:554–556
Benhabib on 2:519, 2:522
of celibacy 3:234–235
change and evolution 13:565–569
children and 3:477–484
Cicero and 3:771–772
as code of conduct 17:400–401
cognition 13:464–465
and communalism 4:264
of comparative judgment 19:472, 19:474
mathematical psychology 14:801, 14:804
concept of, Bentham's 2:527
conflict of 13:614–618
constitution *see* Constitutions
as constitutive 13:550–553
limitations 13:552
modern critical perspectives 13:551–552
plurality of legal experience 13:552–553
protector of property 13:551
social power 13:550
continental European perspective 13:479–482
argument from correctness 13:480
claim 13:480–481
dual nature of law 13:481
necessity 13:480–481
outermost border 13:481–482
participant and observer 13:482
positivism and nonpositivism 13:479–480
Radbruch formula 13:480–482
reasonable disagreement 13:481
relativism 13:481
separation thesis and connection thesis 13:479
social facts 13:479–480, 13:482
and courts in authoritarian states 13:427–434
authoritarian ROL 13:432–433
authoritarianism 13:427
competition and informal practices 13:433
constitutions 13:429
corruption 13:433
functions of 13:428–430
hybrid regimes 13:427
legal positivism 13:428
legitimacy 13:427–430
leverage at work 13:431–432
models of judicial power and independence 13:430–431
political policing 13:428
rule by law 13:427, 13:432–433
critical and positive 13:483–488
of cultural differentiation 6:839
cultural resource management 5:489–490
and databases 5:803–810
decision-making 13:461–462, 13:464–465
definition 13:484, 13:535
deliberation 13:462, 13:465
of diminishing marginal utility (LDMU) 18:621–622
and dispositions 3:268
and disputes 6:560–561
production 6:565–566
domestic violence 6:625–632
due process in 6:696–699
economic analysis of, foundations and criticisms 13:452
basic assumptions 13:452
notions of fairness and law 13:452
economic explanation of 13:452
and economics 13:448–453
economics of its public enforcement 13:570–577
and elder abuse 7:332
emergence of the field 13:461–462
emotion 13:461–462
in group and institutional settings 13:465–466
identifying and interpreting 13:463–464
implementation and reform, implications for 13:464–465
in reasoning process 13:463
role in doctrine 13:462
for various legal actors 13:462
empathy 13:463–464
empirical legal studies 7:554–556
enforcement 13:486–488
equity in 7:924–928
and ethics, boundary between 8:65
facial expression 13:463
as a facilitative arena for interaction 17:398–400
and feminist legal theory 8:920–921
forensic psychology 9:351
future study 13:466
and gender 9:785–786
and global governance 10:156
governance of rules 13:486

Law (*continued*)
government by 20:781
government under 20:781
Habermas on 10:471–472
harm principle 13:486–487
harmonization of 10:158
against hazing in United States 10:558
household economics 11:225
human rights 13:487–488
identifying and investigating law's affective dimensions 13:462–463
images of 13:585
implicit bias 13:464
inclusive legal positivism 13:485–486
income distributional equity and 13:452
issues related to environment in Soviet Union 20:832
and job evaluation 12:817–818
judicial behavior 12:885
legal bases of social work 13:720–725
legal positivism and natural law 13:484–485
legal reasoning 13:461
legal validity 13:484–485
liberal 14:14–15
mobilization of 13:590–597
moral pluralism 13:483, 13:487
moral reasoning 13:463
morality of 20:783
neuroscience 13:461, 13:465
and organizations 17:397–404
philosophy *see* Legal philosophy
privacy laws 18:947–951
of privacy/confidentiality in social research 18:942–943
and race 19:807–811
and randomized controlled trials 8:543–544
relation to social sciences 13:585–589
and risk 13:504–508
role in community-based health interventions 10:648
as rule 13:585
rule of 23:362, 23:364–365
and science 21:156–162
decision science 21:158
disciplines and professions 21:158–159
DNA typing 21:161
forensic science 21:159–161
law as site of knowledge-making 21:160–161
natural law 21:156–157
posttraumatic stress disorder 21:159, 21:161
psychiatric expertise 21:159
psychology and law 21:158–159
recovered memory 21:159–160
regulatory science 21:157, 21:161
scientific evidence and expert witness 21:159–160
social sciences of risk 21:158
transition to modernity 21:156–157
U.S. Supreme Court 21:157–160
separability 13:485–486
separation thesis 13:485–486
against sexual harassment 9:730
and sexual harassment in United States 21:721–723
and sexual orientation 21:768–772
social norms 13:465
and society 13:509–514
field development 13:515–519
socio-legal perspective 13:604–613
sociology of 13:598
American Sociological Association Section on 13:598
law within the courts 13:598–599
value added by 13:601–602
and state 20:780–781
and statehood 23:347–351
and technology 13:520–527
copyright 13:524–525
developing countries 13:522–525
development 13:521
domestic welfare in design of IP rules 13:521–522
global economic considerations 13:521–522
global welfare concerns 13:522–523
health care 13:525–526
human rights 13:521
human welfare 13:523
interactions, complexity of 13:523–524
lagging formal law 13:524
liberty 13:521
need for legal coherence 13:523
patents 13:525–526
pervasive technologies 13:523
private law 13:525–526
public welfare 13:522–523, 13:525–526
reorienting IP frameworks 13:524–525
toward coherent welfare axis 13:520–521
TRIPS agreement 13:521–522
and terrorism 13:528–534
counterterrorism 13:530, 13:532
detention, interrogation, and prosecution 13:532–533
human rights law (HRL) 13:530
international cooperation 13:533
international humanitarian law (IHL) 13:530
international law 13:530
interrogation 13:532
investigation 13:532
law enforcement approaches to 13:530–531
military action in response to terrorism 13:533–534
motivations and operations 13:529
organization and tactics 13:529
pertinent law 13:529–530
rendition/extraordinary rendition 13:533
targeted killing 13:530, 13:534
United Nations Charter 13:530, 13:533–534
in United States and beyond 13:528
toleration 13:487
torture and 24:442–445
as typology of experience 17:401–402
use of term 13:585
utilitarianism 13:484–485, 13:487
and voting 25:309
Law, John 5:524, 14:407, 16:562–563
on actor-network theory 5:525
Law, Robin, works, *Sites of Gender* 11:21
Law, sociolegal perspective on 13:604–605
constitutive dimensions 13:604–605
early themes 13:605–608
gap 13:605–606
legal order, distributive tilt of 13:606–607
legal order (bargaining and negotiation) 13:606
multiple sites of governance 13:607–608
procedural dimension 13:604–605
recent themes 13:608–610
globalization of law 13:608
law and organizations, intersection of 13:609
legal consciousness 13:610
legal culture 13:610
legalization of society 13:608–609
new and developing themes 13:610
rights consciousness 13:610
social movements and law 13:609–610
regulatory dimension 13:604–605
Law and aging 13:420–426
history of 13:420–421
mapping the field of 13:421–422
tool for social change 13:425
Law and development 18:648
see also Postcolonial law
Law and development movement 13:515–516, 13:518–519
Law and economics (L&E) 13:454, 13:741, 13:793–794, 20:863
economic analysis 13:454–455
empirical work in 13:456–457
behavioral law and economics 13:458–459
for multidisciplinary discourse on law 13:457–459
sociological influences in economic analysis of law 13:458
foundational research 13:455–456
rational choice explanation 19:896, 19:900
of safety 20:867
evidence on tort law stimulus to safety 20:867–868
theoretical stimulus of tort law to safety 20:867
theory concerning government regulations 20:868
Law and geography
impact analysis and 13:474–475
legal geography and 13:475–478
regional studies and 13:474
Law and inequality 13:600
Law and neuroscience 13:489–496
adolescent justice 13:493–494
assessing dangerousness 13:493
brain imaging, caveats and cautions of 13:490
brain injury and pain 13:491–492

cognitive capacity and intent 13:492
cognitive neuroscience 13:489
future of 13:494
human judgment and memory, limits of 13:494
lie detection and recognition 13:492–493
neurolaw and groups 13:493–494
neurolaw and individuals 13:491–493
technological capacities 13:489–490
Law and society
areas of study 13:512–513
attitudes 13:513
functions of law 13:509–510
individual theorists 13:510–512
legal profession and actors in legal system 13:512–513
professional organizations 13:509
research 13:509
status differences 13:513
testing assumptions 13:513
Law and Society Association (LSA) 7:554, 13:509, 13:515, 13:518–519
Law and society movement 13:774–775
Law application 12:932
Law avoidance 13:470
Law beyond the courts 13:598–600
constitutive perspective 13:599–600
enforcement 13:600
globalization 13:600–601
governance 13:600–601
implementation 13:600
institutionalization 13:600
interpretation 13:600
lay engagement 13:599–600
legal consciousness 13:599–600
organizational practices 13:600
transnational legal transfers 13:600–601
Law Commission of New Zealand 1:143
Law Council of Australia, Legal Services Board (LSB) 13:621
Law enforcement
crime policy
in United States 5:177, 5:179–180
goal of 13:575
Law enforcement cooperation, and transnational organized crime 24:596, 24:599–600
Law enforcement strategies, and reduction of gun violence 10:460
Law enforcement/policing, and applied research (criminology and criminal justice) 5:129–130
Law firms 13:578–584
BRICS 13:580–582
compensation systems
in law firms 13:578–579
Cravath Model 13:578–579
death of big law 13:581–583
disruptions 13:583
diversity 13:580
emerging economies 13:580–581
feminization 13:582
future 13:583
globalization of Anglo-American model 13:579–580
growth theories 13:580–581
hiring 13:578–582
in-house legal counsel 13:578–579
innovation 13:583
magic circle 13:579
outsourcing 13:582
partnership 13:578–582
specialization 13:579–580
technology 13:582–583
in United Kingdom 13:578–579, 13:581–582
in United States 13:578–581, 13:583
Law first orientation 13:468–469
Law for the Prevention of Congenitally Diseased Progeny 9:908
Law for the Prevention of Progeny with Hereditary Diseases 8:220
Law for the Protection of German Blood and Honor 11:152
Law for the Protection of the Race 8:220–221
Law in action 13:598–599
Law in society 13:758
Law merchant 13:915
Law of effect (learning) 13:705, 19:908
Law of evolution 23:275
Law of exercise (learning) 13:705
Law of frequency (learning) 13:705
Law of gender discrimination 13:768
Law of international arbitration 12:508
Law of international institutions 12:508
Law of international organizations 12:507–508
Law of large numbers 19:35
Law of merchants 12:508
Law of nations 1:611, 11:796
Law of nature 3:266
Law of outer space 12:508
Law of popular affectivity 15:577
Law of recency (learning) 13:705
Law of recognition and succession of states 12:507
Law of retail gravitation 14:547
Law of small numbers 20:490
Law of substitution 14:623
Law of the Sea Tribunal in Hamburg 12:510
Law of treaties 12:507
Law on Protection of the Rights and Interests of the Elderly (China) 21:735
Law on the books 13:598–599
Law reform 13:567
Law reporting 4:247
Law schools 13:736
Law Society of New South Wales 19:103
Law within the courts 13:598–599
Law-enforcement officials, behavior of 13:572–573
LAWER *see* Life-terminating acts without an explicit request (LAWER)
Lawful revolutions 13:440
Law-in-space 13:476
Lawmakers, and delegation 6:58–59
Lawmaking 13:818
Law-related education (LRE) 13:736–737
Lawrence, G., on rural changes in Australia 2:240
Lawrence, Thomas B. 17:394
Lawrence III, Charles 5:268
Lawrence v. Texas (2003) 21:770
Laws of Manu, Indian philosophy 11:48
Laws of nature 5:87–88, 5:90–91, 5:93
Laws of sea 12:507
Laws of the Far East 13:619
Laws of thought 19:391
Laws of war 12:507–508
Lawsuit, torts in law and 24:429, 24:431, 24:433
Lawyer regulation
academic interest 13:620
binding regulations by authority 13:621
changing organizational forms and 13:622
definitional issue 13:619–620
globalization and technology, impact of 13:623
guidelines, guidance, comments, or commentary 13:622
journals 13:624
jurisdiction's constitution treaty 13:622
by law society or regulator 13:622
law society or regulator, by 13:622
lawyer conduct rules 13:621–622
Lawyer–client relationship 16:420
Lawyer-knows-best attitude 13:585
Lawyer-negotiators 16:418–420
Lawyers 13:598–601, 17:482–485
alternative approaches to 18:390–391
cause lawyering 18:390–391
political 18:390–391
in state power 18:391
definition of 13:619
government-paid *see* Government-paid lawyers
in-house counsel 12:126–128
life cycle 13:620
regulation of 13:619–627
role in dispute transformation 6:564–565
role in negotiation/bargaining 16:418–422
social organization of the profession 13:628–634
Lawyers Committee for Human Rights 3:676
Lay engagement 13:599–600
Lay expertise 21:209–210
Lay historian metaphor 4:195
Lay representations of illness and treatment 2:466–467
Lay theories
and essentialism 8:14
of wisdom 25:590
Laycock, Thomas 12:774
Layered social ontologies 14:416
Lay–experts, in citizen science 3:632–633
Laypersons, collaborative research with 7:864
Laywomen 25:630–631
Lazarsfeld, Paul Felix 4:135, 9:321–322, 13:635–640, 15:248–250, 15:800, 20:672
Adler and 13:639
administrative research of 13:640

Lazarsfeld, Paul Felix (*continued*)
Bühler and 13:637, 13:639
Bureau of Applied Social Research (BASR) 13:636, 13:638–639
Coleman and 13:639–640
and Columbia tradition 13:638
and Columbia University 13:636, 13:639–640
contributions 13:637–639
to audience research 13:638
to social research 13:636–637
on cross pressure 13:638
on empirical social research 13:635–636, 13:638–640
and Ford Foundation 13:639–640
on interviewing 12:671
Jahoda and 13:637–638
Katz and 13:638
Merton and 13:638–640
Mills and 13:638–639
narcotizing dysfunction of mass media 13:638
on opinion leaders 13:638
on personal equation 13:637–639
and Princeton Radio Research Project 13:636, 13:639
and Quételet 13:638
and Rockefeller Foundation 13:636, 13:639
Sills and 13:637
on sociology of social science 13:638
on two-step flow of communication 13:638
on vision 13:635–636
works
Communications Research 1948–1949 7:563
The People's Choice 7:563, 25:302
Radio & the Printed Page 7:563
Wirtschaftspsychologische Forschungsstelle 13:637–639
Zeisel and 13:635, 13:638
Lazarus and Folkman's stress model 1:70
Lazi 21:738
Lazy learning 19:60
LBA *see* Linear ballistic accumulator (LBA) model
LBK *see* Linearbandkeramik (LBK) period
LBPs *see* Local binary patterns (LBPs)
LBS *see* Location-based services (LBS)
LC *see* Locus coeruleus (LC)
LCA *see* Leaky competing accumulator model (LCA); Life cycle assessment (LCA); Luce's choice axiom (LCA)
LCDs *see* Liquid crystal displays (LCDs)
LCHD *see* Life Course Health Development (LCHD) model
LCM *see* Linear counterfactual models (LCM)
LCN *see* Low-copy-number (LCN) DNA profiling
LCPs *see* Life-course persistent offenders (LCPs)
LCT gene 11:294
LD *see* Linkage disequilibrium (LD)
LDA *see* Latent Dirichlet allocation (LDA); Linear discriminate analysis (LDA); Long-distance agreement (LDA)
LDC *see* Linguistic Data Consortium (LDC)
LDCM *see* Landsat Data Continuity Mission (LDCM)
LDCs *see* Less developed countries (LDCs)
LDP, Japan *see* Liberal Democratic Party (LDP, Japan)
Le, Van Huu, works, *Dai Viet su ky* 11:82
Le Bon, Gustave 4:152–153, 4:158, 22:323
crowd psychology 7:417
works
Crowd, The 4:151–152
The Psychology of Crowds 14:744
Le CD-Rom Jean-Jacques Rousseau 20:767
Le Center Intercantonal d'Information sur les Croyances 16:808
Le monde (newspaper) 21:753
Le Play, Frederic 11:576
works, *The Workers of Europe* 4:361
Le Quy Don, works, *Phu Bien Tap Luc* 11:83
LEA *see* Local Education Authority (LEA) for London
LEABRA 21:351–352
Leach, Sir Edmund Ronald 6:245, 13:71
works, *Political Systems of Highland Burma* 1:768–769, 2:72–73
Lead isotope analysis 24:481
Lead organization approach 23:519
Lead poisoning, childhood 7:759
Lead time, in supply chain management 23:711
Leader–follower relations 8:442
Leader–leader exchange (LLX) 13:644
Leader–member exchange (LMX) theory 13:641–647, 13:650, 17:325, 17:719
dyadic leadership 13:641, 13:644
exchange quality measurement 13:641–642
foundations of 13:641
high-quality relationships
consequences of 13:643–644
development of 13:642–643
importance of 13:644
LMX differentiation 13:644–645
and organizational citizenship behavior 17:319
relational theories 23:703
relative LMX 13:645
research opportunities 13:645–646
Leaders 10:314–315
charismatic
in administration 1:106
political sociology 18:473–474, 18:476
role in mass killings/genocide 14:704–705
women 25:625–626
Leadership 4:801, 6:663, 6:665, 6:667, 7:276, 7:448, 8:773–774, 8:777, 10:18–21, 13:648–654, 14:472–474, 14:476, 15:83, 16:553, 18:771, 18:774, 21:137, 22:248, 23:704–705, 25:592, 25:596
achieved *see* Big man
authentic *see* Authentic leadership
authoritarian 23:703–705
and authority 2:269–272
behavioral approach to 13:655–656
collective behavior 13:652
of communities of practice 4:330
contemporary psychology of 13:649–651
as contingency 13:649–650
as transaction 13:650
as transformation 13:650–651
contextual approach for 13:649–651
cross-cultural training 5:338–339
definitional issues 13:648
effectiveness, in higher education 10:847–848
evolutionary group dynamics 8:442
and factionalism 8:683–684
female leadership 25:639–640
followers 13:648–652
group processes 13:648, 13:652–653
and group processes 10:413
shared leadership 10:414
in groups 10:424
and identity 11:529–530
identity approach for 13:651–653
and military psychiatry 15:496, 15:498–499
in multinational corporations 10:413–414
negative 23:704–705
new psychology of 13:651–653
as identity realization 13:652
as identity representation 13:652
old psychology of 13:648–649
as charisma 13:649
as intelligence 13:649
and organizational behavior 17:700
changes 17:700
as function of strength of situation 17:700
and organizational climate 17:322–325, 17:327, 17:330
and organizational culture 12:148–149
on PA effectiveness 17:719
persuation 13:653
and political participation 18:507
political protest 18:422
politics 18:460–461, 18:463
populism 18:609
prototype 13:652–653
psychopathic 23:704–705
self-categorization theory 21:458
small group research on 22:103–106
social influence 13:648, 13:650–651
social work response for disasters 7:429–431
teacher *see* Teacher leadership
women 25:626
Leadership education 13:655–663
adult education 13:656–657
adult learning 13:660
coaching 13:661
developmental readiness 13:660
e-leadership 13:662
emerging trends in 13:661–662
emotional intelligence 13:656, 13:661–662

emotions 13:655, 13:659–662
evolving definitions and theories of leadership 13:655–657
experiential learning 13:660
intentional change theory (ICT) 13:660
leadership competencies 13:657–661
leadership development 13:655–662
leadership theories 13:658t
management development 13:656, 13:659, 13:662
MBA programs 13:659
mentoring programs 13:661
self-image 13:656
servant leadership 13:657
service-learning programs 13:657
teaching and developing 13:657–659
theory and practices related to developing leaders 13:659–661
transformational leadership 13:656
Ulysses Program 13:660–661
virtual teams 13:662
Leadership emergence 7:446
Leadership emotions, in workplace 7:509
emotional intelligence in 7:509
with emotional labor 7:509
Leadership for learning 4:801
Leadership in Energy and Environmental and Design (LEED) silver standard 7:774–775
Leadership in organizations 13:664–669
adaptive leadership 13:668
authentic leadership 13:664, 13:668
cross-cultural leadership 13:667
definition 13:664
group coaching for leadership development 13:668–669
moral leadership 13:667–668
proliferation of theories 13:664–667
attribution theory 13:665–666
charismatic leadership 13:666
clinical paradigm 13:666–667
order out of chaos 13:665
recent development 13:667–668
research areas 13:668–669
role of behaviors 13:665
shadow side of leadership 13:664
shared, collective leadership 13:667
situationists 13:664
traits theory 13:664–665
transformational leadership 13:664, 13:666, 13:668–669
Leadership in religious organizations 20:410–411
Catholic religious orders 20:410
congregations 20:410–411
future perspective 20:411
in local settings 20:410
recruitment and training 20:411
roles 20:410
Leadership role of government in social work development 22:791
Leading activity 16:513, 25:318
of adolescents 16:515–516
of infants 16:513–514
of middle childhood 16:515
of toddlers 16:514
Leading economic indicator (LEI) approach to forecasting 9:316–317
LEADMIX 9:888
League for the Protection of Motherhood and Sex Reform 9:644
League of Nations 9:215, 11:368, 11:371–372, 11:796, 17:640–641
Committee on Intellectual Cooperation 22:675
Slavery Convention of 1926 3:677
Leaking (word order) 25:653
Leaky competing accumulator model (LCA) 6:398
Lean management initiatives, workplace autonomy 25:720
Lean startup approach to new venture creation 7:684
Leaning, visual words 25:214–215
Leap, William, works, *Word's Out: Gay Men's English* 19:755–756
LEARN (lifestyle, exercise, attitudes, relationships, and nutrition) Program 19:377
Learnability 18:77
evolution of 13:321–322
Learned dependency 21:507
Learned helplessness 1:182, 1:545, 9:733, 21:507
and academic achievement 1:24
alleviating 1:24
Learned industriousness, and grit 10:398
Learned societies *see* Social science learned societies/professional associations
Learner-centered instruction, in foreign language learning 9:329–330
Learners 12:196, 12:198
gifted, characteristics of 7:159
and information and communication technology 12:68–70
Learners' native language 21:361
Learning 1:536–537, 2:482–483, 5:899, 7:848–851, 13:699–711, 18:125–126, 18:129, 18:131, 21:360, 23:111–112, 23:114, 23:292, 23:873–880, 25:322–327
adult *see* Adult education
algorithms 23:415
algorithms for 1:531
as apprenticeship 22:6
approach to persuasion 17:941–942
aptitudes 7:218
artificial grammar learning 13:83–84
artificial intelligence
connectionism 2:37
hybrid models 2:39
symbolism 2:37
assessment 14:753–754
Bayes risk function for 23:414
beliefs relating to 12:202
and bounded rationality 19:908
cellular mechanisms shared by memory 23:876–877
as a communal process 25:319
computational neuroscience 4:482
concept learning 4:538–541
consolidation and reconsolidation of memories 23:878f
container models of 7:415–416
to control dynamic systems 6:711–713
craft production 5:110
cultural *see* Cultural learning
DAG models from data 2:366
difficulties 20:9
spoken *versus* learning language 20:9
'direct transmission' view of 12:326
domain-general developmental mechanisms for 4:40–41
pretense 4:40–41
school readiness and success 4:40
self-regulation and executive function 4:40
domain-specific developmental mechanisms for 4:38–40
conceptual change 4:39–40
testimony 4:39
theory of mind 4:39
theory theory and probabilistic learning 4:38
and education 1:728–730, 1:732, 7:144–147
tangible user interfaces (TUI) 24:20–25
education and 7:137–143
environmental modeling assumptions 23:413
epistemic cultures and 7:877
and equity in education 7:918–922
estimative approach to 23:414
evaluation of generalization performance 23:415–416, 23:416f
and evolution 8:389–390
and expectations 8:513–515
economic applications 8:515
other approaches and further issues 8:514–515
stability 8:514
statistical approach 8:514
exploratory 1:15, 1:18
and flow 9:261–262
foreign language 9:327–332
genetic programming 2:44
goals 15:888
goals and self-management 10:624
higher education research 10:864–867
in honeybees 15:155–159
human, theories of 11:622
and instruction 15:928–935
intelligence and 12:327–328
interoception in 12:618
and interorganizational relationships 12:622–624
interpersonal 10:430
from interventions, cognitive development 4:38
and knowledge representation 13:104
knowledge representations 23:413
language learning impairment 13:344–352
lifetime and transformations of memory traces 23:879f
and limbic system 14:125, 14:127–129

Learning (*continued*)
long-term depression 23:873, 23:875–876, 23:875f
long-term potentiation 23:873, 23:875–879, 23:875f
LTP/LTD 14:358–360, 14:362–364
from mathematical models 14:794–795
mathematics 14:818–821
and memory
basic neural changes 13:677
functional and anatomical correlates 13:678–682
memory representation 13:677–678
memory consolidation at cellular level 23:874–876
memory recall and reconsolidation 23:877–879
and mental imagery 15:222–223
models in vestibulo-ocular reflex 25:71–72
motor control models 15:957–964
from multimedia 7:248–249
mutants
screens for 9:906
neural basis of 13:677–683
and neuronal connections 4:96
neuronal mechanisms of 16:620
nonparametric bootstrap simulation method 23:415
open education 17:227–228
environment 17:227–228
events 17:228
organizational 13:695–698
paradigms in neural networks and motor control 15:963–964
parametric bootstrap simulation method 23:415
perceptual *see* Perceptual learning
philosophy 13:699–704
behaviorism 13:702–703
British empiricism 13:701
cognitive psychology 13:704
constructivism 13:704
dualism 13:700–701
experimental physiology 13:701
experimental psychology 13:702
functionalism 13:702
Gestalt's 13:703
humanism 13:704
law of association 13:701
nature 13:699–700
neobehaviorism 13:703
nurture 13:699–700
positivism 13:701
psychoanalysis 13:703
reasoning 13:701
self 13:699–700
society 13:699–700
structuralism 13:701–702
and placebo effect 18:153–154
plasticity 23:873–875, 23:877–878, 23:875f
predictive approach to 23:414–415
and prenatal/infant development 18:852–858
prior knowledge 12:327
procedural 16:846
consolidation of procedural knowledge and sleep benefit 16:846–847
importance of practice for skill development 16:846
neural substrates of 16:847–848
stages 16:846
procedural models of 7:415–416
process of brain 4:100
program evaluation 19:118
and psychology 19:412–416
public policy 19:573
reading and writing 18:130–131
reinforcement 2:843–844, 23:413–414
role in chronic pain 17:452–453
school 1:861–864
for selfrealization 12:697
self-regulated 12:202–203
and self-regulation 21:554–555
signal transduction 23:874–875
situated 22:5–12
and sleep 22:63–69
as social interaction 7:242–243
Bronfenbrenner's ecological system 7:243
social cognitive theory 7:243
sociocultural theory 7:242
as sociocognitive activity 13:671
socio-constructivist conception of 12:201–202
sociocultural 16:603–604
style 12:69–70
supervised 23:413
support for
general trends and issues 5:330–331
home and family 5:329
individual attitudes and beliefs about learning 5:329–330
schools and teachers 5:328–329
synaptic bases of 23:873–874, 23:874f
synaptic learning rule 23:875, 23:875f
synaptic strength 23:873–878
synaptic transmission 23:873–875
and tacit knowledge 24:4–7
with a teacher 16:116
theories 13:704–711
advance organizers 13:710
assimilation theory 13:710
background 13:704–705
basic assumptions of behaviorism 13:704
behavioral psychology 13:704
cognitive psychology 13:706–707
cognitive theory 13:707–708
creativity 13:710–711
discovery learning 13:710
Guthrie's 13:706
information processing models 13:708–709, 13:708f
instructional theory 13:699
modern trends in behavioral theory 13:706
operant conditioning 13:705–706
respondent conditioning 13:705, 13:705t
schema theory 13:709–710
social learning theory 13:707
Thorndike's 13:705
Watson's 13:705
transformative 1:15, 1:18
transgenic/knock-out studies 9:906
and transition to adulthood 24:580–581
unsupervised 23:413
using writing to support 25:770
why and how of 13:699
Learning, mathematical learning theory 13:690–694
conceptual arguments 13:690
contexts 13:690
historical arguments 13:690
practice
mathematics in 13:690–693
psychology in 13:690–693
Learning activities engagement 1:31
Learning analytics, lifelong learning 14:93–94
Learning and instruction
argumentation 13:673–674
Bartlett way 13:670
classroom learning and teacher's role 13:674–675
contribution of experimental social psychology of cognitive instruments 13:671–672
from individual to collective levels 13:670–671
learning as sociocognitive activity 13:671
paradigms, shifting 13:671
levels of analysis 13:671
mastery goals 13:671
social interaction as dynamic and dialogical processes 13:672–673
social-cognitive perspectives 13:670
sociocognitive conflict 13:671, 13:675
sociocultural approach 13:670
Vygostkian way 13:670
Learning and Study Strategies Inventory (LASSI) 13:715–716
Anxiety Scale 13:716
application 13:716
Attitude Scale 13:716
Concentration Scale 13:716
Information Processing Scale 13:716
Motivation Scale 13:716
Selecting Main Ideas Scale 13:716
self-regulation component of strategic learning 13:716–717
Self-Testing Scale 13:716
skill component of strategic learning 13:716
Study Aids Scale 13:716
Test Strategies Scale 13:716
Time Management Scale 13:716
will component of strategic learning 13:716
Learning by doing 10:648, 14:553, 17:536
Learning communities, and coconstructivism 3:913
Learning curve 13:692–693
Learning disabilities 5:209, 15:212
and social work 13:684
community care 13:684, 13:686
definitions 13:684
families 13:684, 13:686–689
inclusive practice 13:687

institutional care 13:684, 13:686–687
normalization 13:684–685
philosophical and theoretical perspectives 13:684–686
policy frameworks and service provision 13:686
rights 13:685–688
role of social work 13:686–689
social model 13:684–686
special education 13:686
strengths-based approaches 13:687–689
Learning effects, experimental economics 8:548
Learning Environment Inventory (LEI) 3:825–826
Learning environments 7:277, 12:197–199, 21:119, 21:123, 24:47
guidelines for expected behavior 24:47
positive 24:47
routines and procedures 24:47
Learning games 6:900
Learning goals 21:535–537
Learning health care system (LHS) 24:283–285
Learning How to Learn Project 4:796
Learning Index 16:725
Learning management system (LMS) 6:895, 12:67
blackboard learning system 6:895
Sakai's 6:895
Learning materials
and educational progress 8:732
Learning networks, lifelong learning 14:94–96
Learning objectives or targets 3:819
Learning organizations 13:696–697
libraries as 14:47
social work education 22:779–780
Learning outcome 12:197–198
Learning outcome statistics 5:763–765
Learning potential, Vygotsky on 25:314
Learning processes 12:197–199, 13:91
and absorptive capacity 1:15, 1:17f
Learning resources, in open education 17:227
Learning spaces 4:8
Markov processes in 14:584–590
assessment engine 14:586–588, 14:586f
background 14:589–590
Bayesian operator 14:588
beta learning model 14:589–590
evaluating efficiency of 14:589
extra question method 14:589
fringes of a state 14:584–585, 14:588–589
half-split questioning rule 14:588
knowledge assessment 14:586–588
knowledge spaces 14:584, 14:589–590
knowledge state 14:584, 14:586–589
knowledge structure 14:584
learning engine 14:588–589
learning success probability 14:589
mathematical learning theory 14:589–590
miniature example 14:585–586
permutability 14:588
questioning rule 14:588
random walk tracking mechanism 14:589
response rule 14:588
standardized testing 14:589
uncovering knowledge state 14:584, 14:586–589
updating rule 14:587–588
Learning strategies 21:365
impact of anxiety on 24:246
Learning styles 21:365
cognitive styles and 4:110–117
personal differences in approach and performance 4:112–113
models based on orientation to study 4:113
models based upon Learning skills and neuroscience 4:113
models focusing on experiential learning 4:112–113
models focusing upon instructional preference 4:113
Learning success probability 14:589
Learning systems 12:196
Learning theories 2:482–483, 12:197, 15:788, 7:239t
of aggression 1:374
frameworks 23:214–215
mathematical 14:801–802
finite state Markov chains 14:801–802
linear operators models 14:802–803
mathematical learning theory 14:812–813
in parenting and behavioral problems 2:472
of socialization and education 22:863
Learning to learn 13:712–719
assessment of 13:715–717
checklists 13:715
questionnaire 13:715
talk-alouds 13:715
examples 13:712
historical overview 13:713
interventions 13:717–718
extended 13:717–718
tutoring and supplemental instruction (SI) 13:717
workshops and short courses 13:717
model of strategic learning 13:714–715, 13:714f
impact on a student's acquisition 13:714
variables impacting 13:714–715
strategies 13:712–713
for orchestrating cognitive and motivational resources 13:712–713
selection of 13:712–713
types 13:713–714
elaboration strategies 13:713
organization strategies 13:713–714
rehearsal strategies 13:713
Learning together
in classroom group processes 10:417
cooperative learning 4:884
Leasing, property 24:529
Least cost analysis, archaeology 10:101
Least preferred co-worker (LPC) model 13:649–650
Least preferred item 2:548
Least restrictive environment 23:222t
Least squares estimate (LSE) 9:624, 19:714, 24:362
simultaneous equations 21:981
Least-liked group 18:492–493, 18:495
Least-median-of-squares (LMS) estimator 20:735
Least-square prenucleolus 4:872
Least-squares estimate (LSE) 20:733–734
Least-squares learning 8:514–515
Leather bars 13:887
Leather Pride Night Committee 13:887
Leather square 13:887
Leatherman subculture 13:887–889
critiques of 13:889
current state of 13:888
gender and sexuality 13:888–889
history of 13:887–888
Leave-one-out cross-validation 23:416
Lebanese Broadcasting Corporation 16:776
Lebanon, legal system in 15:364
Lebensführung (life-conduct) 25:464
Lebensphilosophie 9:669, 10:289
Lebensraum (living space) 7:723, 10:91–92, 13:228–229, 22:41–42, 23:356, 24:449–450
Lebenswelt 20:666, 24:589
Lecky, William 10:491
Leclerc, Georges Louis 20:286–287
Leclère, Adhémard, works, *Histoire du Cambodge* 11:84
Leçon 13:72
LED *see* Light-emitting diode (LED); Local economic development (LED)
Lederach, John Paul 25:842
Lederman, Sully 10:776
Lee, Matthew 11:629
Lee, Sungmin, on narrative networks 16:216
Lee–Carter model 15:866
LEED *see* Leadership in Energy and Environmental and Design (LEED) silver standard
Leeds, M., works, *The Economics of Sport* 23:318
Leenhardt, Maurice 1:767
Leenstelsel 9:111
Lees-Haley Fake Bad Scale 16:723
Leeuw, Gerardus van der 7:375, 20:344–345
works, *Religion in Essence and Manifestation* 20:344
Leeward Antilles 1:638
Leeward Islands 1:638
Lefebvre, Henri 12:913–915, 22:438, 23:201
left anterior negativities (LANs) 9:482–483
Left cerebral hemisphere 13:398, 13:402–403
anatomical and cytoarchitectonic details 13:399f
brain connectivity within 13:402–403
Left dioid 22:483
algebra 22:483

Left Fascists 22:840–841
Left hemisphere 2:807–808, 16:690–694, 18:46–47, 18:61–62, 23:294–296
functions of 2:801f
Left hemisphere damage (LHD) 18:791–792, 21:930
Left hemisphere lesions 21:930–932
Left identity dynamic network 22:482
Left internationalism 10:188
Left politics, and global south 10:186–189
Left Socialist Revolutionary Party 20:856–857
Left-behind people, during internal migration 12:439–440
Left-handedness 9:269
Left-hemispheric strokes 2:810
Left-libertarian green parties 10:380, 10:383
Left-linear grammar 19:10
Left-right asymmetry of body 2:123–124
Left-right cleavage 25:303
Left-right continuum, and class voting 25:299–301
Left-wing libertarianism (Steiner) 14:43
Legacies
feminist 8:899
industrialization 11:871
socialist 3:609–610
Legacy-creating stage of aging 1:437
Legal academics/law professors 18:387–388
Legal actors 13:598–599
Latin America 13:412
Legal aid 12:919–923, 12:925–926
poverty law 18:734–735, 18:737–739
Legal analysis 13:450
Legal anthropology 13:802
Legal argumentation 13:776–783
common law reaction 13:779–780
early civilians 13:777–778
later civilians 13:778–779
methodological and theoretical approaches 13:776–777
Legal aspects
discovery 6:516–522
equality and inequality 7:894–902
Legal astuteness 3:27
Legal attitudes 13:729–730
Legal capacity 3:477
Legal category, experts as 21:208–209
Legal changes 13:566–568, 17:851, 18:389
Legal citizenship 3:638
Legal codification of racial categories 19:808
Legal coherence, and law 13:523
Legal complex 18:387–389
characteristics 18:387–388
international contexts for 18:389
in postcolony 18:388–389
toward political liberalism 18:388
Legal compliance, and diversity 6:595
Legal consciousness 13:471, 13:591–592, 13:599–600, 13:610, 13:726–733
constitutive theories of 13:730–732
disputes 6:560–562
as epiphenomena 13:728–729
as individual attitudes and opinions 13:729–730
research on 13:731–732
as structural residues 13:728–729
Legal constraints, on affirmative action 1:254
Legal convergence, Latin America 13:412–413
Legal corruption equilibrium 5:14–15
Legal costs 13:450–451
Legal Council of the League of Nations 10:6
Legal culture 13:470, 13:472, 13:610, 13:726–733
components of 13:727
concept of 13:727–728
constitutive theories of 13:730–732
scientific evidence and 21:297–299
Legal cynicism, and collective efficacy 4:172
Legal databases
see also Databases
Legal decision-making, Weberian typology of 24:635t
Legal diversity 13:615
Legal domination 2:921
Legal drugs 6:686
Legal duties 13:450
Legal eclecticism 13:758–759
Legal education 3:672, 13:445, 13:516–517, 13:734–739
aims of 13:735
educational continuum 13:736–738
establishment of organized 13:734–735
future of 13:738
globalization and 13:735
Islamic 12:747–749
specialization 13:738
trends in 13:735–736
Legal enforcement of racial hierarchy 19:808–809
de facto racial subordination 19:808–809
de jure racial subordination 19:808
Legal equality 7:907
Legal ethics 13:620, 13:622, 13:736
Legal evolution 13:565–566
Legal factors, of business 3:2
Legal families 3:669
Legal fees
conditional fees 12:922
contingent fees 12:922
court-awarded fees 12:922
Legal feminism 9:730
Legal fictions 1:143
Legal for lesbians 13:893
Legal formalism 13:468, 13:516, 13:565, 13:740–746, 13:792–793
conservative views 13:744
constitutional law and 4:707, 13:742–743
history 13:740
ideal type of 13:741
judicial behavior and 13:744–745
jurisprudential 13:741
nineteenth-century 13:741
as a normative system 13:744
precedent and 13:743
private law and 13:743–744
during Reagan Administration 13:741
Reagan years 13:744
revival of 13:741–744
rise and fall of classical formalism 13:741
rules *versus* standard 13:744–745
textualism and 13:743
use of term 13:740–741
Legal framework, of transnational organized crime 24:599–600
Legal freedoms 18:387
core legal rights 18:387
political rights 18:387
property rights 18:387
Legal globalization 13:555
Legal highs 6:686
Legal history of special education (in United States)
accomplishments 23:223–224
amendments to EAHCA 23:223–224
children with disabilities 23:219–220
early involvement by federal government in education 23:221
Education for All Handicapped Children Act 23:221–222
enactment of compulsory education laws 23:219
federal district court rulings in educational opportunity cases 23:220–221
formation of national advocacy groups 23:219–220
US Supreme Court's Ruling
in *Board of Education v. Rowley*) 23:222–223
in *Brown v. Board of Education* 23:220
Legal idealism 18:391
Legal indeterminacy 14:276
Legal institutions 1:481, 12:935, 13:598–599
Legal instructions 12:910–911
Legal insurance in Europe 12:217–218
Legal interpretation 12:932, 13:786
Legal issues
in combined pharmacotherapy and psychotherapy 19:497–498
genetic factors (in organizational psychology) 9:923
Legal issues: public opinion 13:747–752
Legal justice 12:137–138, 12:932
alternatives to 12:935–936
commonsense conception of 12:932–934
justice and 12:932–935
Legal language 12:934
Legal liability 13:448
Legal mobilization 13:428, 13:546–547
collective 13:593–594
contextual factors influencing 13:591–593
cutting-edge research on 13:594–596
individual-level 13:590–591
multidimensional process model (MPM) 13:591, 13:591f
'transformative' approach 13:593–594
Legal model, feudalism 9:111–112
Legal moralism 6:675–676
Legal moralism, and victimless crime 25:76
Legal motherhood 8:782
Legal occupations 13:619
Legal of linguistic databases 5:815

Legal order (bargaining and negotiation) 13:606
Legal orientalism and postcolonial law 18:649–650
Legal origins thesis 13:805
Legal orthodoxy 5:251
Legal parentage 3:477–478
  as foundation of family law 8:781–782
Legal paternalism 6:675–676
  and victimless crime 25:76
Legal personalities 13:753–756
  and business corporation 13:753, 13:755–756
  and classes of entities 13:753–754
  in jurisprudential history 13:754–756
  in Western culture 13:754
Legal philosophy 18:24–30
  Dworkin's 18:24, 18:26–28
  formal institutions 18:24
  legal positivism 18:25–27
    command theory of law 18:26
    Hart's concept of law 18:26
    inclusive *versus* exclusive legal positivism 18:26–27
    pure theory of law 18:26
    separation thesis 18:25–26
    validity in 18:26–27
  natural law theories 18:27–29
    conceptual and ontological analysis 18:28–29
    linguistic analysis 18:28
    naturalism in legal philosophy 18:29
    morally defined function of law 18:27–28
    moral realism and connection thesis 18:27
    reductionism, eliminativism, and emergentism 18:29
    social facts and the specification problem 18:28
    validity in natural law theories 18:27–28
  positive law, social facts, and morality 18:24–25
    legal philosophy, science, and political philosophy 18:25
    natural law theories *versus* legal positivism 18:25
    normativity of legal norms 18:24–25
    positive law and social practice 18:24
  social norms 18:24–26, 18:29
  and utilitarian social thought 24:981–983
  validity of law 18:25
Legal pluralism 13:552–553, 13:757–761, 13:809–810, 13:812
  dispute processes and 13:759
  empirical 13:757
  genesis, in sub-Saharan Africa 23:655
  history of 13:757–758
  juridical 13:757
  in Latin America 13:411
  legal orders, impact on 13:759
  and legal traditions 13:812–813
  local political authority, impact on 13:759–760
  property rights and 13:759
  role in modern sociology and anthropology of law 13:757
  rules of state law *versus* 13:759
  social working of legal rules and 13:759
  surveys on 13:760
  theoretical controversy 13:758
  traditional 13:810
  and trials 24:636–637
  use of term 13:757
Legal politics 13:425
Legal positivism 2:527, 6:487, 13:484–485, 13:762–766, 13:792–793, 16:324, 19:167
  core elements of 13:763–765
  debate between natural law and 13:762–763
  Dworkin's criticisms 13:765
  Hart's recharacterization of 13:793
  inclusive 13:765
  versus natural law theories 18:25
  social science and 13:765–766
Legal pressure, in diversion 4:373, 4:376
Legal Principles Dimension 13:421
Legal process and social science 13:767–771
Legal profession 3:672, 13:582, 13:820
  degree of specialization 3:669
  of in-house counsel 12:126–128
Legal professionals 13:619
Legal proof models 13:789–790
Legal protections
  for animal rights in research 1:703
  cultural resource management 5:488–489
Legal punishment 6:255–257
Legal realism 13:468–469, 13:516, 13:772–775
  contemporary 13:774–775
  definition 13:772–773
  difficulties in defining 13:772
  progressive reformers and 13:773
  realists' perspective 13:773–774
Legal realist tradition
  social change and 13:542–543
Legal realists 13:516, 24:432
Legal reasoning 13:439, 13:776–791, 14:276
  abduction 13:776–777
  action 13:776–780, 13:782
  analogy 13:776–783
  argumentation 13:787, 13:789–791
  and artificial intelligence 13:784, 13:786–787, 13:789–790
  authority 13:776–779, 13:781–782
  axiomatic reasoning 13:776–779
  burdens of proof and presumptions 13:790
  cases 13:786–789
    facts 13:789–790
  casuistic reasoning 13:776–779
  causality 13:776–777, 13:780, 13:782
  coherence 13:778–779, 13:782–783
  commentators 13:778
  common law reaction 13:779–780
  concepts 13:777–783
  conceptualism 13:779–780
  deduction 13:776–777, 13:780, 13:782
  as deduction 13:784–785
  deductive reasoning 13:784–786
  defeasible reasoning 13:785–786, 13:790
  dialectic 13:776–779, 13:781–783
  distinction 13:776, 13:778–780
  division 13:778–779
  early civilians 13:777–778
  epistemology 13:776–779, 13:781–783
  exceptions and rule hierarchies 13:785–786
  exegetical explanation 13:777–778
  formalism 13:776, 13:779, 13:782
  functionalism 13:776–777, 13:779–780, 13:782
  geometry 13:778–779
  hermeneutics 13:776–777, 13:779–783
  individualist action 13:776
  induction 13:776–779, 13:782–783
  inference 13:777, 13:782–783
  institution 13:778–779
  interdisciplinarity 13:780, 13:782
  later civilians 13:778–779
  and legal procedure 13:790–791
  legal proof models 13:789–790
  methodological and theoretical approaches 13:776–777
  models 13:781–782
    interpretation model 13:781
    policy model 13:781–782
    remedies model 13:782
    rights model 13:781
    rule model 13:781
  nominalism 13:779
  paradigm 13:776–779, 13:781–782
  person 13:778, 13:782
  philology 13:777
  positivism 13:776, 13:778–779
  principle 13:776–782
  realism 13:776, 13:780
  rule-based 13:786
  scholasticism 13:777, 13:779–781
  structuralism 13:776
  syllogism 13:780–781
  thing 13:778–779, 13:782–783
Legal recognition and protection
  gender variant individuals 24:550–551
Legal Reform Movement (LRM) 21:745
Legal regulation 13:555
  of forensic psychiatry 9:349
Legal requirements
  for health information systems 5:785–786
  punishment 19:631–632
Legal responses to disability 13:498
Legal responsibility, in combined pharmacotherapy and psychotherapy 19:497
Legal rights of handicapped children 13:471–472
Legal rules 13:448–452
  simplification of 12:924
Legal safeguards against unreliable expert testimony 8:584
Legal scholarship 5:269, 13:605, 13:792–798
  Chicago School model 13:794
  contemporary sociolegal studies and empirical legal studies 13:796–797

Legal scholarship (*continued*)
critical legal studies movement 13:794
critique of law (1970–1985) 13:793–794
doctrinal scholarship 13:793
early twenty-first century 13:794–797
jurisprudence and doctrinal studies between 1900 and 1970 13:792–793
law and economics movement 13:793–794
legal formalism 13:792–793
outsider scholarship 13:794–795
rehabilitation of law 13:793
standard legal scholarship 13:795–796
Legal schools
formalization of 12:748
professionalization of 12:747–748
Legal science
and *Geist* 9:666–669
use of term 13:585
Legal services 12:922–923
and rights revolution 18:734–735
Legal Services Corporation (LSC) 18:734, 19:558
Legal Services Corporation (LSC) Act 1974 18:737
Legal Services Initiative 13:625
Legal sociology 13:515, 13:551–552, 13:565–566
Legal status
of child 3:477
and citizenship 3:650–652
Legal studies (*fiqh*) 11:61
Legal styles 13:804
Legal systems 13:449–450, 13:452, 13:606–609, 13:753
African 1:272–275
classification of 13:799
Athens 13:799
China 13:799–800
classical legal traditions 13:801–802
India 13:800–801
Islamic law 13:801
modern, by European jurists 13:802
nineteenth century 13:802–803
Rome 13:799
Sunni schools 13:801
theory and critique of 13:803–805
tripartite classification 13:805
twentieth-century 13:803
formal guarantees in 5:494–496
Latino 13:410
private 13:807–811
religious 1:272
scientific evidence and 21:297–300
Legal theory 4:249–250
Bobbio's 2:715–716
Legal traditionalism 13:516
Legal traditions 13:804
of East Asia 6:778–783
of Southeast Asia 13:812–817
of sub-Saharan Africa 23:654–659
Legal transplants, in sub-Saharan Africa 23:655
Legal validity, and morality 13:484–485
Legalism, political trials and 18:497–502
Legalistic citizenship 3:652
Legalists 4:312
Legality 21:31–33
and punishment 19:658–659
Legalization 13:818–822
general overview 13:818–819
increasing legal regulation 13:819
of institutional behavior 13:820
in and of the international arena 13:820–821
narrower sense 13:818
of organizations 17:398
outside United States 13:821
private rights of action 13:819–820
of society 13:608–609
through private rights of action 13:820
trends 13:821–822
in US states 13:819–820
Legalized abortion 5:200
Legally authorized representative 4:254–255
Legally Mandated Organisations (LMOs) 23:132
LegalTrac 5:809
Lege fascissimi 16:257
Legend of Suns 11:55
Legendre, Adrien-Marie 9:624
Legibus solutus 1:7–8
Legislation 5:635–636, 11:762–764, 13:720–722
and administrators 1:112–113
adoption 1:143
emotion and 13:466
executive government 8:497, 8:499
implementation of 13:722–723
justifications
victimless crime 25:75–76
labor 25:676
for LGBTQ/sexual minority youth in schools 21:88
marriage 6:644
marriage and parenting, LGBTQ families 21:756–757
over Indigenous 11:765
regulatory agencies 20:181–184
science and art of
Bentham on 2:526
constitutional and democratic theory 2:528
Greatest Happiness Principle 2:526–527
pannomion 2:527–528
theory of fictions 2:527
social welfare 13:722
and social work 13:722
for therapeutic effectiveness assessment 24:284–285
in United States 16:250
Legislative action 3:671
Legislative apportionment, proportional representation 19:206–207
Legislative fact 13:767–768
Legislative history, platforms providing 5:808
Legislative institutions, for democracy 6:88–89
Legislative power 6:696–697
Legislative procedure, of United States Congress 4:651
Legislative quotas 8:888–889
Legislative redistricting 6:533
Legislative reform
mental health 15:196–197
for mental illness 15:197
Legislative response to opium addiction 17:240–241
Legislatures, and judicial behavior 12:887–888
Legitim 19:174
Legitimacy 2:697, 4:460, 10:295, 13:751, 13:823–828, 16:764, 16:766–767, 16:917, 18:262, 21:31–32
and authority 2:272
as authorization 4:457
boundary conditions for 4:458–459
consequences
of authority 13:825–827
of deviance, resource distributions, and discrimination 13:827
courts in authoritarian states 13:427–430
definition 13:823
delegitimation 13:827–828
and deliberative democracy 6:65–66
as dynamic and interactive 4:460
funding 9:553
global governance 10:160
of government 24:658
of individuals in authority positions 13:825
influence on law-related behavior 4:459
intergroup relations 13:827
justice 13:823, 13:827
and libertarianism 14:39–40, 14:43
and modern state 23:350–351
and nation 16:223–225
organizations 13:823–827
positive outcomes of 13:826
power 13:825–826
social exchange 13:825
as social process 13:823–824
social psychology 13:823, 22:541
of status structures in groups 13:824–825
as unifying forces shaping organizational forms 17:417–419
Weber on 25:459–460
Legitimacy crisis of planning 18:189
Legitimate authority 4:457
Legitimate institutions 4:457
Legitimate social order, ideologies and 11:574–575
Legitimate violence 25:116–117
Legitimation 5:444
intergroup relations 12:412
social dominance theory (SDT) 12:412
social identity theory (SIT) 12:412
system justification theory 12:412
of online cultural products 12:589
religious 3:534
sanctions 20:891
Legitimation crisis, micromethods 12:650–651
Legitimization 19:612
of evil 6:32–33

and meritocracy 15:241–242
Legrendre polynomials 23:400
LeGuin, Ursula 19:748–749
works, *The Dispossessed* 14:257
Legumes 9:300–301, 9:303
Lehmann, Walter, works, *Annals of Quauhtitlan* 11:57
Lehmann–Doksum quantile treatment effect 19:715
Lehnswesen 9:111
LEI *see* Leading economic indicator (LEI) approach to forecasting; Learning Environment Inventory (LEI)
Leibniz, Gottfried Wilhelm von 13:778–779, 15:173, 15:546, 19:411
on heuristics 10:829–830
Leipzig 21:921
Leistungsverwaltung 16:359
Leisure 3:50, 3:52–54
appreciation of quality 13:850
correlates of 13:846–847
criteria for, and quality of life 13:851
and cultural consumption 13:845–851
definition of 13:837
definitions in relation to work 13:845
European perspective 13:837–844
cultural and recreational expenditures 13:837–839
cultural paradox of free time 13:840–842
international comparison 13:837–839, 13:842
leisure society 13:839–840
reversal of the leisure/status gradient 13:840
scarcity 13:840–842
trends 13:839–842
flow in 9:263
versus free time 13:845
measurement of 13:845–846
and postindustrial societies 18:664–665
sociology of 2:361, 13:852–857
assessment, prospects, and future directions 13:856
challenges 13:852
comparative sociology 13:855–856
fragmentation 13:852
gender 13:854
history and status 13:852
life course 13:855
social class 13:853–856
societal role 13:855
status attainment 13:855
studies 13:852
well-being 13:855
work 13:852–854
subjective aspects of 13:850–851
time diaries 13:845, 13:847
and time use 24:379–380
Leisure activities of adolescents 13:829–836
assessing leisure preferences and participation 13:834–835
consequences 13:835
development 13:831–832
autonomy development 13:832
brain 13:831
changing relational concerns 13:832
gender role 13:832
identity 13:831
parents 13:831
peers and youth subcultures 13:831–832
eudaimonic enjoyment 13:833–834
extracurricular activities 13:829–830, 13:832–833
hedonic enjoyment 13:830, 13:833–834
individual differences 13:832–834
constraints–facilitators framework 13:832–833
ecology 13:832–833
expectancies for success and leisure motivation 13:833–834
personality 13:834
quality of experience and leisure motivation 13:833
intervention ways 13:835
time use 13:829–830
demographic differences 13:830
portfolios 13:830
prevalent activities 13:829–830
types and functions 13:830–831
Leisure class
of fashion 8:836
Veblen on 25:55
Leisure Interest Questionnaire 13:834
Leisure participation, assessment of 13:834–835
Leisure preferences, assessment of 13:834–835
Leisure science, lifestyle in 14:115
Leisure work 13:854
Leitmotifs 23:834, 23:836
Lemaître, Georges 20:288
Lemkin, Raphael 10:1
works
Axis Rule in Occupied Europe 10:1
History of Genocide 10:6
Series III: History of Genocide 10:6
Lemmas 15:827
Lemon v. Kurtzman (1971) 20:280
Len, coefficient for 14:156
LendingClub 4:139–140, 4:142
L'Engle, Madeleine, works, *A Wrinkle in Time* 2:151
Lenin, Vladimir Ilich 3:740–741, 11:686, 14:642, 14:657–658, 18:473, 20:857, 22:844, 22:846–847, 25:748–749
definition of classes 3:746
New Economic Policy 22:851
on peasantry 1:458
works
Development of Capitalism in Russia 1:456
Imperialism. The Highest Stage of Capitalism 11:686
State and Revolution 23:353
Leninism 25:749
political parties 18:404–405
political thought 18:483
Leninist world system 25:751
Lenneberg, Eric 16:678, 20:905
Lennon, John 25:844–845
Lenoir, Raymond 10:139
Lenski, Gerhard 8:466
Lenskyj, H.J., on women's participation in sports 9:657
Lent 3:145
Leone, Sierra 3:426
Leonhard, Euler 14:372
Leontief, Wassily 16:227, 20:147–148
Lepenies, Wolf, on disunity of sciences 7:875
Leptin 2:848–849
-action within hypothalamus 11:488f
Lermontov, Mikhail 21:741–743
Lerner, Daniel 6:267
works, *The Passing of Traditional Society: Modernizing the Middle East* 16:371–372, 16:382
Lerner, Jacqueline V. 3:447–449
Lerner, M.J. 2:487
Lerner, Richard M. 3:447–449
on adolescent self-regulation 21:548–549, 21:551
works
Handbook of Child Psychology 11:277
Lerner index of market power 14:549
Leroux, Pierre 23:1–2
Lersch, Laurenz, works, *Die Sprachphilosophie der Alten* 14:215
LES *see* Luxembourg Employment Study (LES)
Les économistes 18:357–358
Lesbian, gay, bisexual, and transgender (LGBT) 9:635–636, 9:734, 13:595, 13:879–881, 13:894, 15:314, 15:317, 21:769, 21:789
activism in Latin America
see also Latin America
aging 13:864
health, housing, and social care 13:866–867
health and well-being 13:866
and relationships 13:864–866
discrimination
consequences and responses 13:880–881
cultural, patterns of 13:880
drag culture *see* Drag culture; LGBT
friendship experience of 9:454–457
with LGBT community 9:454–456
outside LGBT community 9:456
future directions 13:881
future of 9:636
health of 15:110
issues in Russia 20:840
Lesbian, gay, bisexual, transgender, and intersex (LGBTI)
commercial sex 13:882–886
attitudes of 13:883–884
culture of 13:882–883
regulation of 13:884–885
in South Africa
see also South Africa
Lesbian, gay, bisexual, transgender/transsexual, queer/questioning

(LGBTQ) 21:792, 21:794–796, 21:826
bear and leatherman subcultures *see* Bear subculture; Leatherman subculture
Butch–femme subcultures *see* Butch–femme subcultures
family
see also Families
health challenges 21:795
issues in Jewish life 12:876
LGBTQ-friendly sexual health clinic 21:795
reproductive rights 20:471–472
school environment for
see also Schools, youth
and sexuality 21:819–820, 21:822
youth sexuality education 21:826
Lesbian activism
in Latin America
see also Latin America
in South Africa 21:744–746
Lesbian and Gay Equality Project (LGEP) 21:746
Lesbian androgyny 4:726
Lesbian ecofeminism, radicalized socialist 8:945–947
Lesbian families 8:769–770
emotional development of children in 7:442
Lesbian feminism 8:947, 9:633–634
Lesbian feminist movement 15:112–113
Lesbian health movement
building 15:115–116
Lesbian sex mafia (LSM) 13:887–888
Lesbian wages 13:896
Lesbians 8:942, 10:790–794, 13:893–898, 18:533–534, 18:536–538
bisexuals 13:894, 13:896–897
black, representation in South Africa 21:747–748
discrimination 13:893–897
economic situation 13:895–896
families 13:894–895
gay and lesbian, tourism *see* Tourism
gender 13:893, 13:895
health 15:110–117
creating 15:111–113
establishing institutions 15:113–114
movement, building 15:115–116
radical health movements 15:110–111
health and life course issues 13:896–897
intersectional queer theory 19:765–769
legal and normative challenges 13:893
methodological issues 13:893–894
minorities 15:573
movements in United States 9:632–636
radical lesbian feminism 8:942–948
rights of 3:689
and sexual orientation 21:785–789
and social work 21:802–806
stereotyping of 13:869
as topic of health research 15:114–115
in Turkey 20:871
workplace 13:894–895, 13:897
youth 9:649
see also Butch–femme subcultures
Lesbians-only self-help events 15:112
Leshner, Alan 6:688
Lesioning a model 4:92
Lesions
disease as 17:1–2
evidence of hippocampal amnesia 10:878–882
and hemispatial neglect 10:767–768
lesion mapping in patients 10:768
sites 10:768
and language disorders 13:329–330, 13:332
Leskien, August 2:705
Leslie, P.H. 14:372
Lesniewski, Stanislaw 19:666
Lesotho, sociocultural aspects 23:75, 23:77, 23:79
L'esprit des lois (The spirit of laws) 7:658, 7:721–722
L'Esquisse d'une théorie de la pratique 2:778
Less developed countries (LDCs)
infant and children mortality
and economy 12:24–25
epidemics 12:24
geographic disparities 12:21–22
infectious and parasitic diseases 12:23
malnutrition 12:23–24
medical causes 12:22–24
mortality measures 12:21
neonatal mortality 12:22–23
and politics 12:25–26
internal migration in 12:433–442
Lesser Antilles 1:638
Lesson planning 24:46–47
clear lesson and learning objectives 24:46
creating quality assignments 24:46
instructional alignment 24:46–47
instructional planning and teacher expectations 24:46
instructional sequencing 24:46
prioritising instructions 24:46
routines and automaticity 24:46
Lethality, epidemiology 7:826, 7:826b
Letter sound-symbol 7:229
Letter-by-letter reading 6:733
Lettere Remissoriali 15:578
Letter–sound correspondences 4:93
Leucippus 14:243
Levee effect 25:560
Level of fertility 6:110–112, 6:111t
Level of processing theory, of episodic memory 7:857
Levels, in social life 15:312
emergence 15:312
Levels of analysis
of organizational climate 17:329–330
social-cognitive perspectives 13:671
Levels of analysis (in social psychology) 13:899–903, 14:414–419
articulation 13:900–902
concept 13:900
illustrations 13:900–902
contemporary enrichments 13:900
intersocietal level 13:900
neurological level 13:900
four levels model 13:899–900
assessing explanations with four levels of analysis 13:899–900
ideological 13:899
interindividual 13:899
intraindividual 13:899
positional 13:899
Levels of consciousness (LOC) 4:666–667
development 4:667
expansion of 4:667
Levels of nature 3:264
Levels of processing 23:231
of recognition memory 20:41
of retrieval 20:609
Levels of reality 14:416–417
Level-specific models 16:49
Levelt, W. 16:678
Leverage 13:630
Leveraging, and antitrust 1:801
Leveson Inquiry 9:393
Levi, Giovanni 11:108
Levi, S. 3:317
Levinas, Emmanel 5:931
Levine, M.P. 21:755
Levins, Richard 3:156
Levinson, B.A., works, *The Cultural Production of the Educated Person* 7:145–146
Levinson, D.J., works, *The Authoritarian Personality* 2:197–198
Levirate 21:799
Lévi-Strauss, Claude Gustave 1:751, 1:753, 6:503, 9:359, 13:71, 13:904–910, 19:677
on Amerindian myths 13:906–907
analysis of myths 9:359–360
on anthropology 1:768
on bricolage 17:393–394
on creativity 13:908
on cultural diversity and unity of humanity 13:907–909
on culture 13:904–908
on ethnology 13:904, 13:908
on exoticism in Brazil and revelation in New York 13:904–905
on fundamentally symbolic taboo 13:905–906
on history 1:748
on human mind 13:905–907
on humanism 13:904, 13:908
on incest taboo 13:905
on kinship 13:905–906, 13:909
on language 13:905–907, 13:909
on linguistics 13:904–908
on model of exchange 20:451
on multiculturalism 13:908
on myths and structures 13:906–907
on racism 13:904, 13:908
on relativism 13:907–909
on religion 13:904–906
on social anthropology 13:908–909
on society 13:904–909
on structuralism 1:768, 13:907, 15:291, 23:627–630
kinship and group structures 23:628–629

savage thought 23:629
transformations and continuity 23:629–630
on symbolism 13:905–907
theory of mind 1:768
on universalism 13:905–907
works
The Elementary Structures of Kinship 23:627–628
La Pensée Sauvage 20:309, 25:753
The Raw and the Cooked 23:626–627
The Savage Mind 1:748, 23:626–629
Tristes Tropiques 2:359
Levit, Solomon G. 8:228
Levitt, H.M.
on butch–femme relationship 13:872
on femme identity 13:871
Levitt, Theodore 10:219
Levy, Elena 10:111
Lévy, Emmanuel 13:551
Levy, Jerre 23:294
Levy, Michelle, works, *Indigenous Psychology in Aotearoa: Realising Maori Aspirations* 11:791
Lévy-Bruhl, Lucien 1:767, 6:706, 8:292, 13:71, 14:436
works, *How Natives Think* 8:588
Lewin, Kurt 4:51–52, 13:911–914, 17:541, 19:414, 19:669
on applied psychology 13:911
assessment of works 13:913
biography of 13:911–912
concept of *Aufforderungscharaktere* 8:571–572
on creativity of dialogical thinking 13:912
on formal characteristics of life space 13:912
on group psychotherapy 19:505
Iowa period of 13:913
on marriages 13:911–912
research topics and fields of 13:912
on small group research 22:101–102, 22:104–105
on social interaction 13:912
on social psychological problems, research on 13:912
interaction between research and practice 13:912
on social psychology 22:528–530
use of research for policies of social change 13:912
works
Principles of Topological Psychology 10:435–436
Lewis, Arthur 18:182
Lewis, Bernard, on Said's Orientalism 17:423–425
Lewis, David 5:89–91
concept of convention 12:191
on extreme realism 5:89
theory of 5:89–91, 5:93
works
Convention 4:845
Lewis, Helen, on psychotherapy sessions 21:875
Lewis, Justin, works, *Enlightened Racism* 2:217
Lewis, Oscar 18:719–720
on culture of poverty 5:591–592
Lewis and Clark map 11:314f
LEWRG *see* London Edinburgh Weekend Return Group (LEWRG)
Lewy bodies 10:501
Lex animata 1:8–9
Lex injusta non est lex 12:933
Lex loci contractus 13:616–617
Lex loci damni 13:616–617
Lex loci delicti 13:616–617
Lex loquens 1:7–8
Lex mercatoria 12:471–472, 12:508
globalization and 13:917–918
historical development and current definitions 13:915–916
modern 13:915–916
theoretical analysis 13:916–917
as unwritten or, at least, not systematically collected principles and clauses 13:916
Lex rei sitae 13:616–617
Lex Salica 2:672, 22:393
Lexeme 3:183
Lexical access, in speech production 23:257
Lexical agraphia 1:453, 6:735
Lexical and frequency effects 21:624
Lexical and syntactic ambiguities of individual words 19:355
Lexical approach
to extraversion 8:623–624
of personality traits 17:922
Lexical borrowing 3:921
Lexical compounding 3:509–510
Lexical consolidation 13:919–924
behavioral evidence for 13:919–921
in children 13:923
complementary learning systems 13:922
memory consolidation 13:919, 13:922–924
neuroimaging evidence for 13:921–922
during sleep 13:922–923
speech 13:919, 13:921
during wake 13:923
Lexical decomposition structure 13:944–945
Lexical design
consequence of 13:946
Lexical dysgraphia 1:453
Lexical effect 18:56
Lexical field theory 13:931–932
Lexical functional grammar (LFG) 8:352, 15:751
Lexical hypothesis, five-factor model (FFM) 2:559–560, 9:230–231
Lexical items 13:256–257
Lexical Markup Framework 4:951
Lexical nonsemantic reading 6:733
Lexical phonology 18:79
Lexical pitch accent (HL) 23:716–717
Lexical productivity 14:222
Lexical representations
normal development of 6:728–729, 6:728f
Lexical retrieval 23:230, 23:233
Lexical selection processes 23:232
Lexical semantics 13:930–936
cognitive semantics 13:933–935
basic levels and onomasiological salience 13:934–935
conceptual metaphor and metonymy 13:933–934
frame semantics 13:934
prototype effects and radial sets 13:933
prestructuralist 13:930–931
semasiology 13:934–935
structuralist 13:931–933
distributional semantics 13:932–933
lexical fields and componential analysis 13:931–932
relational semantics 13:932
Lexical stress 13:307, 13:309–310
Lexical studies, of Honesty-Humility 11:169–170
Lexical tones 23:716–717, 23:719–720
Lexical variables, and social class 13:283–285, 13:284t
Lexical–Functional Grammar 4:830
Lexicalism 15:822–823
Lexicalist theories of sentence comprehension 21:625
Lexicalization 10:333–334
processes 14:222
Lexicographic heuristics 5:892–893
Lexicographic rule, of decision making 5:919
Lexicography 13:937–941
corpus 13:940–941
description 13:938
digital 13:940–941
history 13:939
idiomatic expression 13:938
lexical properties described 13:938
lexical units included 13:938
meaning 13:937–938, 13:940
presenting information 13:938–939
semantics 13:937–938, 13:940–941
Thorndike's contributions 24:306–307
vocabulary 13:939–940
Lexicology 13:937–941
and gender 9:797–798
Lexicon 5:812, 13:247–249, 13:925–929, 13:937–938, 13:942–947, 17:224
aphasiology 13:926
argument structure 13:943–944
and bilingual aphasia 2:575, 2:577–578
computational modelling 13:926–928
consequence of lexical design 13:946
decision 13:925
decomposition and event structure 13:944–945
experimental psychology 13:925
fMRI 13:926
functional neuroimaging 13:926, 13:928
future directions 13:928
knowledge, alternative conception of 13:927
neuropsychology 13:925–928
notion of 13:942–943
number of 13:926–927
parallel distributed processing 13:927–928

Lexicon (*continued*)
past tense verb generation 13:927–928
qualia structure 13:945–946
reading aloud 13:927–928
structure of lexical entry 13:943–946
traditional conception of 13:926
transcranial magnetic stimulation 13:925–926, 13:928
word classes and typing information 13:943
word knowledge and specific language activities 13:927–928
Lexicon/syntax mapping 17:558, 17:560
Lexicosemantic processing 13:400–401
Lexicostatistics 8:370–371, 11:23
Lexifiers 18:137–138
Leximin stable allocation 4:871
Lexis diagram 5:735, 6:154–161, 6:171–172
age 6:156
annual record of events for both time of/age at occurrence 6:157–161
constant risk function 6:157–159
squares 6:158–159
vertical-sided parallelograms 6:158
lifeline calculation 6:154f
mathematical framework of study of occurrence/exposure rates 6:156–157
non-repeatable events 6:156–157
repeatable events 6:156
rates 6:157
rates of first and second kind 6:156
regions 6:155f
special case of mortality at ages 0 and 1 6:160–161
variable risk function 6:159–160
squares 6:159–160
vertical-sided parallelograms 6:159
LexisAdvance database 5:803–807, 5:805f–807f
critical services 5:805–807
search capabilities on 5:807–808
Ley, David 10:48
Leydesdorff and Van den Besselaar, works, *Technology Analysis & Strategic Management* 21:322
LFG *see* Lexical functional grammar (LFG)
LFP *see* Labor force participation (LFP)
LFPs *see* Local field potentials (LFPs)
LG-ABN *see* Licking–grooming and arched-back nursing (LG-ABN)
LGBT *see* Lesbian, gay, bisexual, and transgender (LGBT)
LGBTI *see* Lesbian, gay, bisexual, transgender, and intersex (LGBTI)
LGBTQ *see* Lesbian, gay, bisexual, transgender/transsexual, queer/questioning (LGBTQ)
LGCM *see* Latent growth curve models (LGCM)
LGEP *see* Lesbian and Gay Equality Project (LGEP)
LGM *see* Latent growth modeling (LGM)
LGN *see* Lateral geniculate nucleus (LGN)
LH *see* Luteinizing hormone (LH)
LHD *see* Left hemisphere damage (LHD)
LHS *see* Learning health care system (LHS)
LHT *see* Life history theory (LHT)
Liabilities
of bureaucracy 2:910–911
displacement of goals 2:910
spiraling growth 2:910–911
vicious circles 2:910
remedies and damages and 20:413–417
and responsibility 20:589
torts in law and 24:429–433
Liability doctrine 10:595
Liability insurance 12:216–218
Liability litigation 10:594
Liability rules 19:170
workplace safety 20:866–867
Liaison psychiatry 14:1–7, 19:484–485
Liang, Qichao 8:233, 11:51, 24:585
works, *Zhongguo lishi yanjiufa* 11:51
Libby half-life 3:565
Libel 10:563
Liberal Christianity 14:8–10
development of modern liberalism 14:9
liberalism on the ground 14:9–10
liberation theology 14:9
modernity 14:8, 14:10
sexuality 14:10
sociological interpretations 14:10
varieties 14:8–9
ethical and social 14:8
feminist 14:9
rationalist 14:8
romantic 14:8
Liberal constitutionalism 4:704–705
Liberal continuity, and immigration 11:672–673
Liberal culture, democracy 25:360–361
Liberal democracy 6:91, 6:93, 13:437–438, 18:401, 20:676, 23:359–360, 25:346, 25:740–741, 25:743–744
law and 13:438–440
and neoliberalism 16:463
and pluralism 18:237
Liberal Democratic Party (LDP, Japan) 6:798–799, 6:802
Liberal democratic tradition 4:707
Liberal feminism 8:918–921, 8:943–945, 9:729–730
critique of 8:943–945
and feminist criminology 8:912–914
Liberal Jews 12:876
Liberal law 14:14–15
Liberal legalism 13:545
Liberal market economy (LME) 11:831–832, 22:708
Liberal model of media systems 22:584
Liberal modernization 15:710
Liberal parties, in Brazil 18:400
Liberal Party in Belgium 18:400
Liberal political theory 12:916, 18:339–340, 18:482–484, 24:790
Liberal pragmatic interrogation 5:255–256
Liberal religions 21:173–174
Liberal social Darwinism 5:687–688
Liberal societies 13:551
Liberal statism
containment of 18:304–305
limits of 18:303–304
Liberal theory of freedom 14:540–541
Liberal theory of industrialism 10:279
Liberal welfare state 24:573
Liberal-democratic government 10:313–314
Liberalism 2:715–717, 4:317–320, 4:342, 4:684, 4:687, 5:367, 7:87, 8:926, 10:164, 11:651, 14:18–21, 14:26–31, 15:326, 17:735, 18:398, 18:400, 20:397, 20:479, 20:763–764, 24:463–464, 25:276, 25:349, 25:405
and agonism 1:446, 1:448
anthropology of 14:11–17
civil society and state 14:14
diaspora, moral–ethical aspirations, and global force 14:12–13
epistemology 14:11, 14:15
individual, agency, and freedom 14:13–14
liberal law, human rights, universalism, and vernacularization 14:14–15
liberal multiculturalism 14:15
neoliberalism and late liberalism 14:15–16
precursors 14:12
relativism 14:12, 14:14–15
approach to peacemaking 17:637, 17:641
Bell on 2:499–500
in Berlin 2:542–546
and communitarianism 4:308–310, 4:312–313
contemporary 14:29–31
ideals in nonideal society 14:30–31
between social democracy and neo-liberalism 14:29–30
Dahrendorf on 5:665
definitions 14:11–12, 14:22–24, 14:26–28
and deliberative democracy 6:63–64
economic *see* Economic liberalism
equality 14:26–28, 14:30–31
of fear 21:891, 21:894–895
and global laissez-faire 16:462
Humboldt on 25:286–287
and ideologies 11:575–576
and immigration 11:673
impact on social science 14:22–25
justice 14:29–31
and libertarianism 14:25, 14:39, 14:43
Mannheim on 14:497–498
markets 14:26–30
and mass communication 14:692–694
Mill on 15:526–527
and modern society 14:24–25, 14:28–29
as ethos of modernity 14:28–29
as threatened by modernity 14:29
and the nation 14:20
neoliberalism in 16:462
new 18:299–306
organization and milieus 14:20–21
philosophy 14:22–23, 14:26–27
political 19:594
political order 14:18–19

politics 14:23–24, 14:27–28
Popper on 18:523
and populism 18:611
and public reason 19:594
rights 14:26–28, 14:30–31
Sandel on 20:893–895
Sarmiento on 20:908
and Schmitt, Carl 21:29–30
Smith on 22:122
social and economic order 14:19–20
social influences on formation 14:23
society 14:27
as threatened by modernity 14:25
and utilitarian social thought 24:984
Weber on 25:460
welfare 9:399–400
and welfare 25:477–479
Liberalism's triumphant 19:617
Liberalization 14:736, 18:369, 18:372, 20:635–636
and economic reform 18:690–693, 18:695
market 12:58
of markets 13:159
and radio 19:880–881
Liberal-legal political order 18:388
Liberal'no-Demokraticheskaya Partiya Rossii 18:410–411
Liberal-pluralist normative goal of journalism 3:613–614
Liberals 18:839
Liberation ecology 6:317
Liberation hypothesis, race–crime relationship 19:842–843
Liberation movements, national 16:241–246
Liberation theology 3:537–538, 13:859, 14:9, 20:623
and social work 14:32–38
anti-oppressive common grounds 14:32
captivity 14:33
Catholicism 14:32–34, 14:36–37
challenges 14:37
dependency theory 14:33
emancipation 14:32, 14:35–36
empowerment 14:32, 14:35–37
fighting against roots and effects of social oppression 14:35–37
foundational Christian communities 14:34
human rights as God's rights 14:34
integral liberation 14:34
participation 14:33, 14:35–36
poverty 14:33–34
as practical hope for oppressed 14:32–35
preferred action for poor people 14:34
preferred action for young people 14:34–35
research 14:37
social justice 14:32–33, 14:35–36
Libertarian anarchism (Rothbard) 14:42–43
Libertarian paternalism 23:829, 23:831
Libertarianism 9:400, 14:39–44, 19:167–168
and conservatism 1:603–604
functions of state 14:39
fundamentals 14:39–41
market 14:40–41
origins 14:39–40
principles 14:40
as ideology 14:43–44
individual rights 14:39–40, 14:42
legitimacy 14:39–40, 14:43
liberalism 14:39, 14:43
property rights 14:40–43
socialism 14:39–41
varieties 14:41–43
Nozick's minimal state libertarianism 14:41–42
Rand's objectivism 14:42
Rothbard's libertarian anarchism 14:42–43
Steiner's left-wing libertarianism 14:43
welfare 14:40–41
welfare state 14:41
Libertarians 9:405
Liberties 3:644–645, 7:308–309, 9:387–390, 9:398, 12:236t
Berlin 2:544–545
Dahrendorf on 5:661, 5:663–664
and democracy 6:86
and Enlightenment 7:664
in liberalism 14:22
and market allocation 14:542
national liberation movements 16:241–246
social justice 22:392
suppressing crime and protecting 5:178–180
technology and 13:521
Liberty as nondomination 20:478
Liberty Plaza 3:582
Liberty rights, of children 3:486, 3:488
Libidinal drives 21:183
Libidinal stage theory 13:68
Libido, Jung on 12:904
Lib–Lab coalition 14:21
Libraries 14:45–50, 21:322
academic 14:45, 14:49
collections 14:45–47, 14:49
community 14:48–49
corporate 14:45, 14:49
digital 14:46
digital preservation 14:47–48
finances 14:46–47
government 14:45–46, 14:48–49
HathiTrust 14:47–48
of Indigenous collection 11:774
individuals 14:48
information 14:47–48
information literacy 14:46, 14:48
as learning organizations 14:47
as memory organizations 14:47
organizational systems 14:46–47
place 14:49
public 14:45–49
school 14:46
types 14:45–46
Library of Congress 12:359
Libre Ecole des Sciences Politiques 18:467
Libri status animarum 6:138, 6:140–141
Libros del Saber 2:88
Libya, legal system in 15:364–365
LICAI technique 11:401
Licensing 13:629
informal sector 12:56–57
of teachers in United States 24:75–77
Licensing journalists 9:394
Licensure, of teachers in United States 24:76–77
Licit drug use 6:685–686
Licking–grooming and arched-back nursing (LG-ABN) 7:844
Licklider, J.C.R. 21:954
LIDAR *see* Light detection and ranging (LIDAR)
Lie detection
brain-based 13:492–493
fMRI-based 13:492–493
and recognition 13:492–493
Lie scales 17:823
Liebeault, Ambroise, hypnotism studies 19:502
Lieber, Francis 18:466
Lieberman, Victor, works, *Strange Parallels: Southeast Asia in Global Context, c.800–1830* 11:87
Liebes, Tamar, works, *The Export of Meaning* 2:217
Liebig, Justus 24:762–763
Liechtenstein 25:746
Lieh Mak, Felice 25:625
Lien 2:353, 2:355
Lienhardt, Ronald Godfrey 7:376, 13:71
LiEP *see* Language-in-education planning (LiEP)
Liepmann, Hugo 1:874, 18:48
Lieux de Mémoire 4:203, 11:46
Liezhuan 11:49
Life 21:967
connection of indigenous peoples to 11:804
expressions, Wilhelm Dilthey 6:426–428
qualities of 10:524–525, 10:521t
reimagining of 9:1016–1017
Simmel on 21:964
subjective enjoyment of 10:522–523
utility of 10:522–523
Life chances 14:113, 14:115
Dahrendorf on 5:661, 5:663–665
and social differentiation 6:393
Life choices 14:115
Life conditions, and collective violence 14:702
Life course analysis 14:51–54
of linked lives 14:233–234
Life course criminology 14:59–66
Life course epidemiology 14:67–75, 22:277
accumulation model
correlated exposures or chain of risk model 14:69–71, 14:70f
independent exposures and direct effects 14:69–70, 14:70f
single exposure, duration/dosage of 14:69–70, 14:70f
challenges and future directions 14:72–74
family-based studies 14:73–74

Life course epidemiology (*continued*)
multicohort studies 14:73
record linkage 14:73
critical period model 14:68–69, 14:69f
with later modification 14:68–69, 14:69f
historical background 14:67–68
influence in adulthood 6:752, 6:754–757
model selection, socioeconomic mobility and 14:72
models in practice 14:71–72
sensitive period model 14:68–69, 14:69f
with later modification 14:68–69, 14:69f
trajectories, health behaviors and 14:71–72, 14:72f
Life Course Health Development (LCHD) model 14:72–74
Life course sociology 22:540
Life course traditions 20:292
Life courses 1:318–319, 1:330–332, 2:656–657, 2:661, 3:445, 14:51–58, 14:80–84, 15:940, 20:592, 21:37–38
age policies 1:323–324
and age stratification 1:334–336
agency 14:80–83
agent in 15:941–942
in aging research 1:331
American 10:804–805
analysis 14:51–54
application example 14:56–58
basic concepts 14:80–81
challenges and new directions 14:83
cohort 14:82
components of 1:324
context 14:80, 14:82–83
and crime 1:310–314
dechronologization of 1:326
delinquency 6:71–72
dependencies processes on 14:54
at global level of society 14:54
at macro level of society 14:54
and elder care 7:336–338, 7:340
and family 8:740–741
and health 15:100
heterosexual expression over 10:803–805
history 14:80–83
in history 14:76–79
age 14:76–78
childhood 14:76–78
current state 14:78–79
demography 14:76
development 14:76–77
foundations 14:76
future directions of research 14:78–79
gender 14:76–78
industrialization 14:76–77
old age 14:76–77
positivism 14:76
premodern and modern 14:77–78
social science 14:76, 14:79
social structure 14:79
state formation 14:76
individualization of 1:324–325
and infants 12:10–11
interdependent processes on 14:53–54
careers of spouses 14:53
at intermediate level 14:53–54
spouse selection (educational homogamy) 14:53
trajectories at level of individual 14:53
issues of lesbians 13:896–897
and its paradigmatic principles 15:940
and leisure 13:855
age 13:855
biographies 13:855
cohort and generation 13:855
life cycle 14:81–82
life history 14:81
life span 14:81–82
life transitions 14:52
logic of trajectories 14:51–52
and mass imprisonment 14:696–697, 14:699
motivation 15:942–943
naming and 16:185–186
new directions 15:943
and obesity 17:79
parallel processes on 14:53–54
personality development 17:879, 17:881–882
perspective, linked lives in 14:230–231
roles 14:80–83
sequence of multiple life course transitions 14:52–53
sexuality over 21:840–845
and social mobility 22:427
stage, and marital quality 14:515
stratification 14:83
studying individual lives 14:80
temporalization of 1:324–325
theory
development 14:81–82
principles 14:82–83
theory, emergence of 15:940–941
timing 14:80–83
and trajectory of resilience 1:331
and trajectory of vulnerability 1:331
transitions 14:80–83
transitions in 14:52
tripartite social organization of 1:324
and youth culture 25:808–809, 25:811–812
Life cycle 7:594, 14:81–82, 14:230, 14:233
impacts on internal migration 12:445
Life cycle analyses 13:240–242
Life cycle assessment (LCA) 7:701, 11:845, 11:851
Life cycle models of labor supply 13:193
Life cycle rituals, parallels of death rituals to 5:872
Life cycle squeeze 13:855
Life cycle theory of regulation 20:182
Life domains, impact of procrastination in 19:77–78
Life events 14:230, 14:232
and adolescent health 1:124
antecedents 17:466
and anxiety 1:806
bereavement 2:537
and desistance/persistence 5:142
and panic disorder/agoraphobia 17:476
and psychological stress 23:584–586
Life expectancy 1:328, 1:395, 6:121, 9:319, 11:7, 14:91–92, 15:857–861, 16:112, 16:114, 17:178–179, 15:858f
and adult education 1:148
adult mortality in 1:171, 1:173–180, 1:173f, 1:176f–178f
in Africa 1:282t–283t
at birth 1:282t
and age stratification 1:335
at birth, impact of HIV/AIDS on 15:841–842
in developed and developing world 19:813t
in Eurasia 18:549t
gender gap in 9:704
behavioral factors 9:704–705
biological factors 9:704
psychosocial factors 9:705–706
healthy 4:463–464, 4:467
human development 5:775, 5:777–778
and obesity 17:75, 17:78
race and ethnicity 19:812–815, 19:817
see also Adult morality
Life experience 25:592–595
social work education 22:777
Life forms 9:1016–1017
Life histories 3:195, 14:81
collective memory 4:203
context, child soldier experience with 3:431–434, 3:432f
family reconstitution 6:138–139
Life history events, interaction between 8:312–313, 8:313f
Life history evolution, embodied-capital theory of 9:28–34
Life history evolution theory 9:40–45
Life history theory (LHT) 9:40, 14:108–111
age at first birth in 9:41
and cost of reproduction 9:40–41
of Dark Triad 5:674
demographic transition 9:43–44
evolutionary developmental psychology 8:426
and fertility 9:42–43
growth 9:41–42
environments 9:41–42
reaction norms 9:41–42
human behavioral ecology 11:247–248
investment 9:42
and longevity 9:42–43
quality 9:43–44
quantity 9:43–44
timing 9:41–42
trade-offs 9:42–43
Life history trade-offs and aging 8:451–452
Life imprisonment 12:961
Life insurance, and value 5:538
Life management
development psychology of 14:85–88
age-related expectations 14:86–87
behaviorism 14:85
cognitive revolution 14:85
coping 14:85–88

developmental tasks 14:86
efficacy 14:86, 14:88
life-goals 14:85, 14:88
maturation 14:85
personal goals 14:87–88
selection-optimization-compensation (SOC) model 14:87
self-regulation theories 14:86
history of 14:85–86
active role of individual in shaping one's life 14:85
definition 14:85–86
personal goals as building blocks for 21:558–559
Life narrative 2:646
Life orientation test (LOT) 17:264
Life orientation test-revised (LOT-R) 17:264
Life outcomes
and conscientiousness 4:661–662
and Honesty-Humility 11:170–171
and neuroticism 16:746
Life patents 9:1004
Life phases, cultural variation in 5:507–515
accomplishments 5:508–509
adolescent initiation 5:509
adulthood 5:512
birth and external womb 5:509–510
community, joining 5:510
cross-cultural variation 5:511
cultural models 5:507, 5:512
ethnography 5:507, 5:509–510, 5:512
first haircut 5:508
getting noticed 5:511
naming ceremony 5:508
rites of passage 5:508
separation 5:510–511
terminology 5:507–509
youth in limbo 5:511–512
Life politics (Ferguson) 22:803
Life satisfaction 1:329–330
and employment 10:515–520
human development 5:775
rise of research on 10:523
Life sciences 2:632, 11:311, 11:315, 21:278
and biocitizenship 2:610–611
in intellectual property rights 12:262–263
Life sciences industry 10:574–575
Life space 4:51–52
Life span 14:81–82, 14:230, 14:234
attachment across 12:232
body image across 2:738–739
change in neuroticism across 16:746–747
cognitive development, dual-process theory of 12:294
conscientiousness development over 4:662–663
distribution of memories across 2:284–285, 2:284f
emotion in cognition across 7:465
extraversion across 8:625
genetics of 15:845
and human development 11:276–280
and life history evolution 9:28–30
obsessive-compulsive disorder across 17:119–126
panic disorder and agoraphobia across 17:474–481
personal goals as building blocks for life-management throughout 21:558–559
personality development 17:879–881
phobias across 18:37–44
psychopathological effects of mass trauma across 14:747–751
self-concept development across 21:472
social phobia across 22:493–499
and warranties 17:172–173, 17:175–177
Life span cognitive development 7:137–138
holistic and multifunctional 7:137–138
implications for education 7:138
multidirectional development 7:138
plasticity 7:138
Life span development 2:349, 2:351, 11:274, 14:100–107, 15:369–370, 25:590
action-phase model 14:104, 14:106
adolescence 14:109–110
adulthood 14:110
Alzheimer's disease 1:585–586
areas for future research 14:105–106
childhood 14:109
dual-process model of assimilation and accommodation 14:103–104
effects of early education on 7:138–141
evolutionary assumptions 14:108–109
expectation of gains and losses across adulthood 14:101f
goal engagement/disengagement 14:102, 14:104–106
history 14:100
inclusive fitness 14:108, 14:110–111
infancy 14:109
influences on development 14:102–103
interindividual differences 14:103, 14:105
intraindividual change 14:103, 14:105
life history theory 14:108–111
metatheoretical propositions 14:100–102
methods 14:105
motivational approach to 4:814–815
motivational changes 15:373
motivational theory 14:104
multidimensional 14:101
multidirectional 14:101
old age 14:110
parental investment 14:108–110
phases 14:109–110
plasticity 14:101–102, 14:104–105
regulation 14:102–106
reproductive success 14:108–110
schema 7:934–935
secondary control 14:106
selection, optimization, and compensation 14:103
tasks 14:101–103, 14:105
testing-the-limits paradigm 14:102
theories of individual agency in human development 14:103–104
Life span psychology 14:86–88
and successful aging 23:667, 23:669–671
Life span theory of cognitive development 1:428, 1:432
Life stages 5:861
Life story 2:638, 2:641
narrative networks 16:212
Life stress, accidents due to 12:132
Life styles 14:112–115
cardiac rehabilitation 20:188–189
Life tables 6:121, 6:171, 6:173–174, 11:5–6, 14:89–92, 16:110–114
for Cohort 16:110–112
cohort life table 14:89–91
inverse projection 6:149, 6:151–152
multiple-decrement tables 14:91–92
in a period 16:112–113
procedures 14:91
statistical approaches to 16:113–114
Life transitions 14:230, 14:232
Life years, healthy 4:464
Life-ability of person 10:521–523
Life-course criminology
conceptual basis 14:59–60
Farrington's integrated cognitive antisocial potential theory 14:61–62
future directions 14:64–65
and GBTM 14:63–64
and general strain theory 23:498
historical background 14:62
and methodological advances 14:62
methodological techniques 14:62–63
Moffitt's dual taxonomy theory 14:60
Sampson and Laub's age-graded theory 14:60–61
theories of crime 14:60
Life-course persistent offenders (LCPs) 5:138–139, 5:141–142, 14:60
Life-course perspectives, of race–crime relationship 19:840–841
Life-course policies 1:326
Life-course transitions 1:324, 21:709
Life-cycle dynamics, and child fertility 8:823–824
Life-cycle impact of vocational education and training 25:249
Life-goals 14:85, 14:88
Lifeline 16:199–201, 16:200f
Lifelong education 1:149
Lifelong learning (LL) 3:13, 7:918, 7:920, 13:736–737, 15:82–83, 16:551, 24:759
and andragogy 1:148–149
barriers 14:97t
challenges 14:98
connecting dots of technologies for 14:97–98
contextualized learning 14:96–97
implications of 4:114
learning analytics 14:93–94
learning networks 14:94–96
learning paths 14:94–95
mobile learning 14:96–97
networked learning 14:94–96
and older age work participation 17:187
open educational practices 14:96
open educational resources (OER) and open education 14:96
open learner models 14:93–94
personalization approaches 14:97–98

Lifelong learning (LL) (*continued*)
in psychiatry 7:191–192
skills training 22:23
and social work education 22:779–780
styles for
implications of 4:114
support with new technologies 14:93–99
technologies for 14:93–99
technology-enhanced learning (TEL) 14:93–94
Lifelong Learning Accounts 1:150
Life-model, of social work 6:872
Life-project, of individual 17:935
Life-span models
of education, innovating 7:142
of motivation 21:63
Life-span theory of control 1:182
Lifestory (television drama) 9:990
Lifestory (television drama) 9:990
Life-story interviews 16:218
Lifestyle, exercise, attitudes, relationships, and nutrition *see* LEARN (lifestyle, exercise, attitudes, relationships, and nutrition) Program
Lifestyle factors, coronary heart disease 4:920
Lifestyle-exposure theory 20:774
and criminal victimization 20:774
of victimization 20:774
Lifestyle–routine activities theory (L-RAT) 20:774–775
Lifestyles 14:112, 18:665
on Adler 14:112–113
bioarchaeology 2:601–602
in China 2:69
consumption 4:758–761
cultural 22:331–334
definition 14:112
and diabetes 6:337–338
and gender 9:704, 9:706
and health 22:277
in health sciences 14:114–115
and *Lebensführung* 14:113
in leisure science 14:115
in market science 14:115
and morbidity compression 4:464–466
omnivorous, accounting for 22:335–336
and social stratification 22:342–343
and sociological theory 14:113–114
sport 23:316
sustainability 23:828–833
sustainable 23:834–836
by Weber 14:113
Life-support technology 5:866
Life-terminating acts without an explicit request (LAWER) 7:598
Lifetime frequency of opiate abuse 17:240
Lifetime prevalence 7:825
Life-work-and-times biography 2:647
Lifeworld 1:721, 18:20–21
colonization of 10:471
epistemic cultures and 7:874
Husserl on 11:442–443
LIFO (last-in-first-out) 19:786
Liga 20:286
Light, speed of 17:89
Light delivery, for opsins 17:271
Light detection and ranging (LIDAR) 10:65–66, 10:99, 23:744–745
archaeology 10:99
Light-emitting diode (LED) 11:837–838, 11:840
Light-emitting diode (LED) industry 11:837
Light-enhanced-startle effect 8:862
Lijphart, A. 4:693
typology of democratic regimes 4:692f
Lijst Pim Fortuyn 18:412
Like me theory 15:131–134
applications and implications 15:133
autism 15:133
culture 15:133
disorders 15:133
mirror neurons 15:133
body schema 15:131–133
child development 15:131, 15:133
counter theories 15:133–134
early experiments 15:131
gaze 15:132
imitation 15:131–133, 15:132f
intention 15:131, 15:133–134
related theoretical work 15:132–133
research prospects 15:134
self–other mapping 15:131, 15:133–134
social learning 15:131
testing 15:131–132
acts like me 15:132
experiences like me 15:132
perceives like me 15:132
theoretical background 15:131
Likelihood (probability) 14:116–119, 23:411
Bayes' theorem 2:379
in event history analysis 8:313
event–history analysis
inference 14:351–352
structure of 14:352–353
function 14:116–117
history and further reading 14:118–119
parameter estimation 14:117–118
principle 14:117
versus probability 14:116
relation to significance testing 14:118
significance tests
based on 21:959
statistical inference 14:116–118
statistical sufficiency 23:420–422
Likelihood analysis
-based methods, in event-history analysis 8:299
cointergration 24:326–327
checking for specification 24:327
reduced rank regression 24:327
estimation of litigation 14:267–268
function 16:132, 23:414
inference
cointegration 24:326
Likelihood observed data 23:414
of observing a feature vector 23:411–412
Likelihood principle 14:120–121
capture-recapture model 14:121
conditional likelihood 14:120–121
example 14:121
incidental parameters 14:123
integrated likelihood 14:122–123
example 14:122–123
marginal likelihood 14:120
example 14:120
profile likelihood 14:121–122
conditional approximated likelihood 14:121–122
example 14:120–121
modified profile likelihood 14:122
semiparametric model 14:124
example 14:124
Likelihood ratio 14:631
Likelihood ratio tests (LRT) 2:593, 8:716, 16:131, 21:952
for hypothesis 11:491–492
majorization/stochastic orders 14:450
Likelihoods
that eliminate nuisance parameters 14:120–124
Likert, Rensis, semantic differential 21:563
Likert ratings scales 16:404
Likert scales 2:179, 11:750, 21:563, 23:748–749
Likhachev, Ivan 21:742
Liking
social influence 22:350
and wanting/needing 25:353–355
Lilliputian figures 10:499
Lim, Alwyn, on world culture 5:539
Lima, Almeida 23:18
Limb apraxia 21:934–936
Limb preference, and lateralization 2:800
Limbic leucotomy 23:18t
Limbic system 13:679, 14:125–130, 16:625, 16:737
anatomy of 14:125–127
and emotions 7:474
fiber tracts 14:127
fornix 14:127
medial forebrain bundle 14:127
stria terminalis and ventral amygdalofugal bundle 14:127
functions 14:127–129
additional roles of structures 14:129
emotional memory 14:127–128
hippocampal memory system 14:128–129
learning 14:125, 14:127–129
limbic cortex 14:126
nuclear groups and major structures 14:125–126
amygdala 14:125
hippocampal formation and associated structures 14:125–126
hypothalamus and mamillary bodies 14:126
septal nuclei and nucleus accumbens 14:126
Limbic–hypothalamo–pituitary processes, hormone regulation by 16:650f
Lime (*Tilia*) decline 7:696
Liminality 14:131–137, 17:709, 17:711

and anthropology of experience 14:136
border 14:133–134
carnivalesque 14:135–136
danger 14:133–136
in-between spaces 14:131–132, 14:136
interstitial spaces 14:131
liminoid 14:133, 14:135
nonplaces 14:131, 14:134
pilgrimage 14:131, 14:133–135
place and space 14:133–135
and psychotherapy 14:135
ritual 14:131–136
threshold 14:131–133
and tourism 14:135–136
transition 14:131–134, 14:136
travel 14:131, 14:135–136
Turner on 14:132–133
van Gennep on 14:131–132
Limited dependent variables, in cross-national research 19:720
Limited information maximum likelihood (LIML) estimator 12:206, 21:976–978, 21:982, 23:408
Limited liability 3:30, 4:939–940, 12:213
Limited partnerships 3:30
venture capital 25:60
Limited quantity scarcity (LQS) 22:351
Limited security, and attachment 12:230
Limited time scarcity (LTS) 22:351
Limited vote (LV) 7:351, 19:206
Limiting frequency interpretations, objective probability 19:40–41
Limiting relative frequency interpretation of probability 19:26
Limiting vector 23:469
Limiting-pathway (LP) models of epistasis 8:717
Limits to growth (LTG) study 6:24–25
climate change 14:141–142
economic collapse 14:141
feedback 14:139
natural resources 14:138, 14:141–142
origins 14:138–139
overpopulation 14:138
political debate 14:141–142
pollution 14:138–139, 14:142
resource depletion 14:138, 14:142
results 14:139
scientific discussion 14:139–141
system dynamics 14:138–141
world modeling 14:138, 14:141
LIML *see* Limited information maximum likelihood (LIML) estimator
Limon, Jose 1:236
Lincoln, Abraham 3:691
commemoration 4:236
presidency 18:861
Lincoln–Peterson estimate 18:603
Lindblom, Charles E. 3:356, 18:204
political power 18:762–764
works
Politics, Economics and Welfare 18:233–234
Linder, R.M., works, *Handbook of Correctional Psychology* 7:591
Lindsay, Patrick, on division of labor 6:602–603
Line bisection task 2:808–809
Line Islands 17:158
Line symbols 24:265
Lineage
evolution of language 8:365–368
Lineage clan 3:731–732
Lineage functional analysis 9:537
Lineage Genographic Project 10:23, 10:25–27
Lineage society
of West Africa 1:300–301
Linear, functional data analysis modeling 9:516–517
Linear accumulator models, context-free 5:903–904
Linear ADISOP model 24:254
Linear assignment 19:431
Linear ballistic accumulator (LBA) model 5:903
with no start point variability and Fréchet drift rate variability 5:904
Linear CADISOP model 24:254
Linear chronology 24:410–411
Linear contingency fits 23:611
Linear counterfactual models (LCM) 15:77–78
causal mediation analysis 15:77–78
Linear damping 14:798
Linear discriminant function 16:117
Linear discriminate analysis (LDA) 8:659
Linear draft record model 16:902–903
Linear dynamic models, time series 24:317
Linear economy 10:393
Linear factor analysis, in forced-choice personality assessment 17:843
Linear filter 24:333
Linear Gaussian models, cultural evolution 5:404
Linear Gaussian state space models 24:354–358
ARIMA 24:356–357
ARIMA components models 24:356–357
dynamic factor models 24:358
latent risk model 24:357
multivariate models 24:357
regression and intervention effects 24:356
structural time series analysis 24:356
Linear growth models, sexual development 14:330
Linear hypothesis 14:144–149
Bayesian estimation 14:147
covariance matrix 14:145, 14:147
definition 14:144–145
estimation 14:145–146
fallacies and interpretive problems 14:150–152
model checking 14:147–148
regression 14:153–156
regression analysis 14:144, 14:157–161
relationship of quantities 14:150–152
testing 14:146–147
models 14:146
parameters 14:146–147
use of term 14:144
weighted models and covariance estimation 14:148–149
Linear interpolation, inverse projection 6:149
Linear ISOP model 24:254
Linear logistic model, optimal experimental design 17:256
Linear mixed models 5:353, 14:162–168
BLUP 14:166
fixed effect 14:162, 14:164, 14:166–167, 14:167t
intracluster correlation 14:162–163
longitudinal and clustered data 14:163–164
lowess 14:163–164, 14:166–167, 14:164f
MIT Growth and Development Study 14:166–167
prediction 14:166
random effect 14:162–168
repeated measures 14:162–168
restricted maximum likelihood 14:166–167
spline 14:166–167
Linear modeling 19:727
Linear models 1:348
dynamic economic systems 15:296–298
applications 15:297
efficiency 15:297–298
generalized 1:659–664
for growth curve analysis 10:443–444
instrumental variables 12:206–207
integration of 24:254
Linear operators models, mathematical psychology 14:802–803
Linear ordering errors 18:55
Linear prediction coefficients 23:260
Linear prediction filter 23:260
Linear predictive coding (LPC) of speech 23:260
Linear pricing
with multiple outputs 20:167
with single output 20:166
Linear probability model 16:132
Linear regression models 13:385, 14:157, 19:712
for APC analysis 17:728–730
conventional approaches and limitations 17:729–730
methodological process 17:728–729
theoretical problems 17:730
for crossover designs 5:352–353
errors 8:11
intrinsic estimator for 17:730
value of R2 in 15:647–648
Linear stationary processes, cointegration 24:324–325
Linear structural equation models (LSEMs) 13:394–396, 15:76–77, 13:394f
causal mediation analysis 15:76–77
effects 15:76–77
probability distributions 13:395–396, 13:395f–396f

Linear Structural Relationships *see* LISREL (Linear Structural Relationships) model
Linear systems theory, Fourier analysis 9:376–379
Linear test forms, computerized test construction of 4:511–512
Linear time 24:409
Linear *versus* chain-linked model of invention 6:409–410
Linearity, concept of 20:923
Linearbandkeramik (LBK) period 13:338
  colonial migrations 15:477–479, 15:478f–479f
Lineups, eyewitness testimony 8:646–647
Lingua franca
  higher education research 10:867
  of socialism 22:834
Linguicism 2:581
Linguist–community relations 14:186–187
Linguistic analysis, legal philosophy 18:28
Linguistic anthropology (LA) 1:237, 8:351, 14:169–175
  communicative competence 14:172
  diglossia 14:173
  and discourse 6:506–507
  ethnicity 14:173
  gender 14:169, 14:171–172
  in Mexico 1:629
  research topics 14:170–174
    conversation and discourse 14:172–173
    identity, socialization, and ideology 14:170–171
    new areas 14:174
    other traditional areas 14:173–174
    social markers 14:171–172
    text structure 14:171
  social class 14:171–174
  society 14:170–173
Linguistic approaches, for art 2:16
Linguistic area 1:955
Linguistic aspects of project content 17:759
Linguistic Atlas of New England (LANE) 6:349
Linguistic barriers 13:261
Linguistic characterization, and color classification 4:228
Linguistic choices 18:799
Linguistic codes, in education 7:147
Linguistic competence 14:172
Linguistic Data Consortium (LDC) 5:816
Linguistic databases 5:811–816
  corpora 5:812–814
    annotation types 5:813–814
    components 5:812
    data sources 5:813
    metadata 5:814
    uses 5:812–813
  data centers 5:816
  definition 5:811
  distribution and archiving 5:814–816
    discoverability 5:815
    legal and regulatory issues 5:815
    reuse 5:815–816
    standards and consistency 5:814–815
  electronic dictionaries 5:811–812
    challenges 5:812
    headwords and organization 5:812
    types 5:811
  examples 5:811
  human language technology 5:811–813, 5:816
  lexicon 5:812
  natural language processing 5:812–813
  types 5:811
Linguistic dependency 6:182–195
  as agreement 6:189–190
  as government 6:189
  notion 6:182
  semantic dependency 6:182–183
    properties 6:183
  syntactic dependency 6:183–194
    criteria for *Synt*-d 6:183–185
    deep-syntactic relations 6:186
    dependency tree as formalism for representing syntactic structures 6:185–186
    English surface-syntactic relations 6:187–189
    in first half of the twenty-first century 6:193–194
    labeled Synt-dependencies 6:183
    linguistic advantages of 6:191–192
    linguistic insufficiency of 6:192–193
    morphological dependency 6:189
    notion of Synt-d 6:185–186
    phrases and phrase-like units 6:193
    properties of syntactic dependency 6:185–186
    surface-syntactic dependency 6:185
    surface-syntactic relations 6:186–187
    and syntactic constituency 6:190–193
    Synt-connectedness 6:183
    Synt-dominance 6:183
  types 6:182–194
Linguistic determinism 13:252–253, 20:903–904
Linguistic disciples 15:826–827
Linguistic discourse analysis
  on coconstructivism 3:914–915
Linguistic distance 6:351, 13:261–262
Linguistic diversity 13:252–254, 13:293–295, 20:903–905
  in West Africa 1:300
Linguistic equivalence 5:320–321, 5:325
Linguistic features of diglossia 6:418–419
  Arabic 6:419
  Creole on 6:419
  Greek 6:419
Linguistic fieldwork 14:183
  contemporary 14:184
  data management, technology, and archiving in 14:187–188
  documentation of 14:184–185
  ethics and linguist–community relations 14:186–187
  history of 14:184–185
  research methods and activities 14:185–186
  use of term 14:183–184
Linguistic form in literature 13:266–267
  attenuated 13:267
  processing of 13:268
Linguistic historiography *see* Historiography of linguistics
Linguistic idealism 4:794
Linguistic impact and cultural turn on writing of history 11:42
  cultural Marxism 11:42
  ideas of postmodernism 11:42
Linguistic imperialism 22:951–952
Linguistic insufficiency of syntactic dependency 6:192–193
Linguistic markedness 14:533–537
  code switching 14:535–536
  cultural markedness 14:536
  functional linguistics 14:536
  generative grammar 14:534–535
  iconicity 14:534
  language universals 14:533–534, 14:536
  neutralization 14:533, 14:535
  phonological markedness in optimality theory 14:535
  Prague School 14:533
  second language acquisition 14:535
  semiotics 14:533, 14:536
  typological universals 14:533–534
Linguistic marketplace 13:286–287
Linguistic morphology 20:900
Linguistic nationalism 13:252–254
Linguistic paleontology 8:210
Linguistic phonetics 18:66
Linguistic possession 18:630–635
Linguistic pragmatics
  in Anglo-American tradition 18:799
  approaches and traditions 18:799–800
  case of implicit meaning 18:796–797
  in Continental European tradition 18:799
  defining field in 18:798–799
  deixis 18:796
  interdisciplinary interfaces 18:800–801
  notions in 18:795–798
  trends and perspectives 18:801
Linguistic presupposition 14:190–192
  dynamic theories 14:192
  encoding of 14:190–191
  pragmatic accounts 14:192
  projection 14:191
  semantic theories 14:191–192
Linguistic processing, and language learning impairment 13:346
Linguistic project, of Humboldt 25:287
Linguistic properties of psychological verbs 16:670–672
Linguistic proto-nationalism 13:254
Linguistic realm, and ultramodern shadow 5:247
Linguistic relativism 4:790–792, 13:293
Linguistic relativity 23:123
  Whorf on 25:575–578
  see also Whorfian Hypothesis (WH)
Linguistic relativity hypothesis *see* Sapir–Whorf hypothesis
Linguistic representation
  content and logical equivalence 20:69–70
  and language of thought 20:70

and reference 20:68–70
Linguistic routines 13:290
Linguistic semantics 21:585
Linguistic sign 15:826–827
Linguistic skills 9:150
Linguistic structure, evolution of 8:357–361
Linguistic studies, of language learning impairment 13:345–346
Linguistic synesthesia 14:193
Linguistic theory 14:177–178, 14:180–181
Linguistic turn 9:769–770, 14:199–202, 24:813, 25:210
in cultural studies 1:934–936
definition 14:199
and the post-modern epistemological crisis 14:202–203
Linguistic typology 8:355, 14:208
classical typology 14:208–209
functional typology 14:210–211
modern typology 14:209–210
typology and language change 14:211–212
Linguistic Typology 8:355
Linguistic unification in Europe 5:460–461
Linguistic universals 14:226
Linguistic variation theory 3:922
Linguistics 2:357–359, 2:361, 13:246, 14:224–229, 14:293–298
applications and extensions 14:228–229
approaches to agrammatism 14:176–182
archaeology 13:337–343
Chomsky on 3:526–530
classification
by interaction 14:226–228
by orientation 14:225–226
of color terms 4:231–234
comparative method in 4:423–426
cultural evolution 5:394–399
deixis *see* Deixis
evidentials in 8:351–356
evolution of language 8:370–371, 8:374–375
and *Geist* 9:669
generative grammar in 4:104
history of 14:215–219
iconicity in 11:512–515
incorporation 14:220–223
information structure in
assertion 12:95–96
categories of information structure 12:96–97
common ground 12:95–96
communication 12:95–96
focus 12:95
knowledge 12:95
presupposition 12:95–96
question 12:95
research traditions 12:97
rheme 12:96–97
theme 12:96
theoretical background 12:95–96
topic 12:96
internal reconstruction 12:447–449
Lévi-Strauss on 13:904–908
onomastics 17:223–226
prototype theory in 19:286–289
as a science 14:224–225
and semiotics 21:592–594
in social sciences 6:486
of tense, aspect, mood, and evidentiality 24:210
turn, in conceptual history 4:548
turn, in history 14:202–207
typology 14:208–214
Linguistics and Poetics 12:783
Linguistic-specific hypothesis 13:346
Linguistification of the sacred 5:305–306
Link function, logistic regression 2:594–595
Linkage, macro/microsociologies 14:415–416
Linkage analysis, sexual orientation 21:781–782
Linkage disequilibrium (LD) 9:887, 11:291
Linkage studies
of behavior 9:933, 9:937
of sexual orientation 21:781–782
Linked data, journals 12:867
Linked lives 14:53
in changing context 14:232–233
and historical time and place 14:231–232
and human agency 14:232
in life course perspective 14:230–231
life course theory 14:83
qualitative studies and life course analysis 14:233–234
quantitative studies 14:232–233
and timing 14:231
Linking, graphics 10:353
Links, social networks 16:575–576
Linnaean taxonomy 2:559
Linnaeus, Carolus 7:722
L'Institut National d'Etudes Demographiques 6:173
Linton, Ralph, idea of status and roles 23:428–429
Liou, K., works, *Handbook of Economic Development* 7:590–591
LIP *see* Lateral intraparietal (LIP) area
Lippe, Rudolf zur 2:729
Lippmann, Walter 1:351, 6:334, 12:859
geneticization thesis 9:976–977
Lipps, Theodor 19:391
LIPRO household model 8:728
Lipset, Seymour Martin 4:135, 18:474, 20:745, 21:169
works, *Continental Divide* 16:310
Liquid crystal displays (LCDs) 11:837–838
Liquidation 2:353–355
of a company 3:32
Liquidity 4:940
crises 19:181–182
LIS *see* Luxembourg Income Study (LIS)
Lisbon Recognition Convention 2:742
Lishman, William Alwyn
on neuropsychological rehabilitation 2:815
works
Organic Psychiatry: The Psychological Consequences of Cerebral Disorder 2:815
LISP-STAT (statistical analysis package) 23:379
LISREL (Linear Structural Relationships) model 3:787–788, 7:5, 23:405
List proportional representation 19:205
Listedness 15:823–824
Listeners, radio 19:877–878, 19:880–881
Listening
communication skill 4:288
field observational research 9:136–137
Literacy 5:771, 5:773, 7:165–166, 7:168, 7:194–195, 11:589, 13:260, 14:236–241
consequences of 17:275–276
cross-cultural study of education 5:327, 5:329–331
cultural history 14:236, 14:238
culture 14:236–237, 14:239
definition 13:260
definitions and mutual relations 14:236–237
economic education 7:21–22, 7:24
and economic well-being 13:261
education 14:237, 14:239
in Europe 14:238–240
and health 10:653–656, 22:287
measurement 14:237–238
oral history 14:236–238
and orality 14:241–245
orality–literacy debate 17:276
parental engagement in children's education 17:500
population studies 14:239–240
and productivity 13:260–261
reading 14:236–238
in South Asia 7:194–197
in Western Europe 14:237–238
and writing 25:782
see also Illiteracy
Literacy achievement, and socioeconomic status 22:927
Literacy acquisition, in Chinese 3:510–512
Literacy development, and educational neuroscience 16:731–732
Literacy education, in twenty-first century 14:246–250
digital environment 14:247
holistic potential of new literacy practices 14:248–249, 14:249f
multimodal literacy 14:246–250
new practices of writing 14:248
new ways of reading 14:247–248
Literal meaning
myth as 16:175
myth as other than 16:176–177
Literal movement 18:146
Literal paraphasia *see* Phonemic paraphasia
Literary conventions 18:530
Literary criticism 24:292
Literary Digest 18:459–460
Literary genres 2:752
Literary historical biography 2:645
Literary journalism 3:612–613
Literary onomastics 17:225
Literary Utopian thought 14:251–259
Literate expression 17:274–278
expression in writing 17:274–275

Literate expression (*continued*)
Great Divide theories 17:276
problems and approaches 17:275–276
consequences of orality and literacy 17:275–276
ethnographic and historical specificities 17:276
orality–literacy debate 17:276
romantic and evolutionary paradigms 17:275
recent insights and future directions 17:276–277
multiplicity and differentiation 17:276–277
practices and processes 17:277
senses and multimodality 17:277
virtues of ethnography 17:277
Literature
alienation in 1:540
and history 11:97–101
books 11:98–99
communication 11:99
concept of totality 11:100
concepts of truth 11:99
conflation of historiography and fiction 11:98–99
culture 11:100–101
meaning, memory, and identity 11:100–101
narration in historiography 11:97–98
narration in literature 11:98
works of art 11:100
in MENA 16:361–365
modernity 15:683
of Mumford 16:137–138
postmodern methodology as 18:681–682
for systematic review 23:907–908
Literature, Maghreb
in 1960–1990 21:752–753
in 1990–present 21:753
in 1950s 21:752–753
Literature reviews 8:618–619, 8:621
Lithics 14:260
and early humanity 14:260
economic perspectives 14:262–263
paleoethnological approach 14:261–262
and prehistory chronology 14:260
and prehistory discovery 14:260
processual approach 14:262
skill evaluation in 14:263
stone tools function, determination methods 14:263
technological analysis and classification 14:262
technological perspective 14:260–261
typotechnological approach 14:261
Lithium 19:202
for bipolar disorder 2:687–688
Litigants, nature of 14:270–271
Litigation 13:448, 13:450–451, 14:265–275, 19:555–556, 19:558–559
of civil disputes 6:557
consumer policy 4:735
cross-national patterns in litigiousness 14:268–270
definition 14:265
likelihood estimation that disputants will turn to 14:267–268
litigants and agents, role of 14:270
nature of litigants 14:270–271
organizational 17:397–401
process of 14:271
reforms 14:272
remedies and damages and 20:413–417
theoretical issues in understanding 14:265–267
uncertainty issue 14:267
winning and losing in 14:271–272
Litigation explosion 13:819
Litigation funding, class actions 3:735
Little fish-big fish phenomenon 19:213
Little Ice Age 3:865, 3:867–868
Little science 7:791
Little's law 19:786
Littlewood, John Edensor 9:224
Littlewood–Paley theory 14:633
Littré, Emile 23:2–3
works, *Dictionnaire de la langue française* 13:939
Liu, J.X., n MSM 21:736–737
Liu, M. 3:317, 3:319–320
Liu, Zhiji, works, *Shitong* 11:50
Liubov' bez granits 21:742–743
Livability 24:616, 24:620
of environment 10:521–522
Live TV sports 14:722
Lived civil religion 3:683–684
Lived experience, Dilthey on 6:424–426
Lived images 25:173
Lived-through experience 17:709
Livelihoods 7:33–36, 7:105, 7:108, 12:544, 25:662
in Africa 1:293, 24:967–969
and internal migration 12:433, 12:436–440
and primitive society 18:911–912
Liveliness, anthropological writing 1:741–742
LiveMocha 6:897
Liveness 24:167
Liver cirrhosis 1:509
Liver knowledge 21:617
Liverpool dock dispute 13:166
Livestocks
functions of 17:577–578
mobile 16:839–841
nonsedentary 16:838
ritual value of 17:579
sedentary 16:839–840
Living, and dead 5:865
Living apart together, in Western societies 17:548–549
Living arrangements of children and parents, changes in 8:841
Living automobility 2:305–306
Living conditions, household panel studies of 5:845–847
Living Constitution 13:742
Living law 13:439, 13:758, 13:809–810
Living primates 18:904–905
definition of 18:905–906
phylogenetic relationships among 18:906–908
Living space (Lebensraum) 16:259
Living Standards Measurement Study (LSMS) 5:744, 5:747–748, 5:766
Living wills 8:280
Living with Nature's Extremes: The Life of Gilbert Fowler White 25:559
Livingstone, David 20:288
Livingstone, Sonia 25:80–81
Livy 11:37
classical legacies 3:772–773
Lizardo, Omar 22:333
on consumption 16:555–556
on cultural capital 22:336
on cultural participation 16:556
on internalization 5:579
on practice 7:415
on procedure 5:553
on skills *versus* status seeking 22:332
on social network actors 5:559
LL *see* Lifelong learning (LL)
Llama, food production 9:303
LLECE *see* Latin American Laboratory for Assessment of the Quality of Education (LLECE)
Llewellyn, Karl Nickerson 13:469, 13:590–591, 14:276–278
American Legal Realism and 14:276–277
branches 14:277
'IdiosyncracyWing' of Realism 14:277
biography 14:276
impact and current importance 14:277
legal indeterminacy 14:276
philosophical defense 14:277
professional socialization and background of judges 14:277
Uniform Commercial Code 14:277
work on Native American law and dispute resolution practices 14:277
works
The Cheyenne Way 13:469
Llewellyn Park 9:614
LLI *see* Language learning impairment (LLI)
Llobera, Josep 16:285
LLX *see* Leader–leader exchange (LLX)
LME *see* Liberal market economy (LME)
LMICs *see* Low- and middle-income countries (LMICs)
LMOs *see* Legally Mandated Organisations (LMOs)
LMP *see* Labor market policy (LMP)
LMS *see* Learning management system (LMS); Least-median-of-squares (LMS) estimator; Local Management of Schools (LMS)
LMST method, in cost–benefit analysis 5:52
LMX *see* Leader–member exchange (LMX) theory
LNC *see* Latitudes of noncommitment (LNC)
LO *see* Lateral occipital (LO) area

LOA *see* Latitude of acceptance (LOA)
Loan Repayment Assistance Program (LRAP) 19:557
Loan security 21:405
Loanwords 3:919, 3:921, 8:210, 13:302–305
Lobban, Michael 13:779–780
Lobbies 12:371
Lobbying 3:24, 3:103, 12:370, 15:19
Lobotomy 19:587–588
LOC *see* Lateral occipital complex (LOC); Levels of consciousness (LOC)
Local Agenda 21 (LA21) 23:796
Local ancestry-based methods 9:888–889
Local Authority Social Services Act 1970 22:801–802
Local average treatment estimates (LATE) 24:707
Local binary patterns (LBPs) 8:659–660
  face recognition 8:659–660, 8:660f
Local buzz 11:829
Local citizenship, in China 24:977
Local clinical scientists 24:518–519
Local coherence 12:40
Local community, and media 4:332–333
Local consistency checking
  constraint-satisfaction problem 4:716–718, 4:716f
    arc consistency 4:717
    complexity and applicability 4:717–718
Local cultures, science as 21:225–226
Local economic development (LED) 14:279–283
  emergence of 14:279–280
  geographical spread of 14:281–282
  as place-based policy intervention 14:280–281
  reaffirming the case for 14:282–283
Local Education Authority (LEA) for London 21:65
Local enhancement
  animal social learning 22:403
Local feature, face recognition 8:660
Local field potentials (LFPs) 15:968
Local food movement 1:566–568
  community supported agriculture (CSA) 1:566–567, 1:570
  farmers' markets (FMs) 1:566–567, 1:570–571
Local geographic models 14:787
Local growth management 10:451
Local healthcare needs, international response to 20:513
Local log odds
  behavioral interpretation in marriage matching 2:104–106
  in education matching 2:102t
Local Management of Schools (LMS) 21:66
Local minima problem 16:36
Local molar validity 12:450
Local polynomial regression, nonparametric regression 16:870–872
Local scholars 3:867
Local sign 25:185
Local social justice 22:398–399
Local state building, in China 24:975–976
Local state corporatism, in China 24:975–976
Local state entrepreneurialism, in China 24:975–976
Local taxation 23:338–342
  business tax 23:340–341, 23:341t
  guidelines for reasonable tax assignment 23:339–340
  income tax 23:339–342, 23:341t
  inheritance tax 23:341t
  in OECD countries 23:340–341, 23:341t
  political economy of 23:341–342
  property tax 23:339–342, 23:341t
  rationale 23:338–339
  sales tax 23:340–342, 23:341t
  subfederal governments 23:338–341
  wealth tax 23:341, 23:341t
Local use in developing societies, craft production for 5:110
Locales 18:149, 22:438, 23:200–201
Localities 4:332
  globalization 10:234–235
  and households 11:228
  in Southeast Asia 23:73
  and territorial conflict 24:216
Localization 18:46–48, 18:58
  of international schools 10:249–250
  in neurolinguistics 16:678, 16:680–682, 16:684–686
Localized cerebral discharges 12:774–775
Localized counting, street children 23:535–536
Localized learning 11:829
Localizing knowledge, methodological problems of 13:92
Local-level politics, in Arctic 1:931
Locally based community development 4:381–382
Locally optimal designs 17:259
  for logistic regression in one variable 17:260
Locally weighted scatterplot smoothing (LOESS; LOWESS) 16:872, 16:876–877
  in linear mixed models 14:163–164, 14:166–167, 14:164f
Locarno Treaties of 1925 17:641
Location alternatives, rank-based 16:891–896
  efficiencies for location procedures 16:896
  k-sample problem 16:893–895
  multivariate two-sample location problem 16:895–896
  two-sample location problem 16:892–893
Location decisions of firms, conventional wisdom on 14:287–288
Location of place 18:149
Location optimizing algorithms 23:139
Location quotient 23:143–144
Location theories 5:278, 10:85, 14:287–292
  where
    people wish to live 14:289
    question 14:287
Locational attainment, and residential segregation 20:540
Locational possessive 18:632
Location-based services (LBS) 23:129
  mobile communication 15:632–633
Location-independent view cells 11:388
Locations 10:174–177
  absolute and relative 14:284–286
  experimental laboratories 8:552
  of firms, conventional wisdom on 14:287–288
  of foreign direct investment 9:325
  and growth 14:288–289
  psychotherapy 19:516–517
  of public facilities 14:290–291
Locative constructions, of spatial language 23:92
Locative expressions 13:311
Locative-focus (LF) construction 23:892
Loch Hyne (lake) 17:165
Lochner v. New York (1905) 6:697
Lock, Margaret, works, *Encounters with Aging* 20:454
Locke, John 1:475, 3:646, 3:674, 4:337, 7:255–256, 7:567, 7:912, 8:215–216, 11:367, 11:413–414, 12:8–9, 14:11, 14:18, 18:33, 19:167, 19:410, 20:443
  democratic theory of 6:90–91
  on empiricism 12:702
  on language 15:289–290
  on political thought 18:481
  on representational realism 20:21
  on sovereigns 11:817
  works
    An Essay Concerning Human Understanding 21:592
    Two Treatises of Government 3:61, 3:674, 11:367, 14:40, 17:516, 20:680
Lockean theory of labor and natural right 11:412, 12:267
Lock-in, organizational path dependence 17:385, 17:387–389
Lockman, Zachary, works, *Workers on the Nile* 16:384
LOCKSS *see* Lots of Copies Keep Stuff Safe (LOCKSS) Program
Lockwood, David 3:742
Locomotion
  body, evolution of 2:723, 2:725
  of infants 22:558
Locus coeruleus (LC) 16:471
  interoception 12:618
Locus of control 3:845, 4:905–906, 15:895–896, 17:343–344, 21:505
  health 4:819
Locus of Control Scale 15:895
Locus of dysfluency 9:266
Locus-specific epigenetic effects 24:687
Locutionary speech act 14:172
Locutions 18:797
LOESS *see* Locally weighted scatterplot smoothing (LOESS)
Lofland, Lyn 17:532
Logan, John 11:283–284
Logan–Stults calculation 20:550
Logarithm, information theory 12:107

Logic 14:293–298, 15:752, 19:391
of grammar 14:296–297
syntactic complexity 14:296–297
institutional 17:393
language and 14:295–296
meaning and 14:293–295
on natural language semantics 14:293–296
relativism about 20:235
science of 10:737–738
and semiotics 21:592–594
Logic games 16:927–928
Logic models 14:299–305
communication 14:299, 14:302
context of evaluation 14:300, 14:303
developing and using 14:303–304
dissemination 14:299, 14:304
evaluability assessment 14:302–303
generic 14:300f
methodologies for research and evaluation 14:302–303
nonlinear 14:302f
with feedback loops 14:303f
outcome evaluation 14:300–304
planning evaluation 14:303–304
for professional development program 14:300f
project management/monitoring 14:299
shared vision 14:303–304
strengths and weaknesses 14:304–305
for systematic review 23:907
theory of action 14:301f
theory of change 14:299, 14:301–304
vertical 14:302f
Logic of action, cultural 5:539
Logic of aggregation 1:75
Logic of exception 21:395–396
Logic of inquiry 14:873–874
Logic of routine 21:395–397
Logic of securitization 21:395–397
exception 21:395–396
routine 21:395–397
Logic of selection 1:75
Logic of the situation 1:75–78
Logic of transformation 1:75
Logic programming 14:320
Logical actions 17:505
Logical analysis, logical empiricism/logical positivism 14:314–316
Logical atomism 14:314, 25:612–613
Logical calculus, reasoning 5:934
Logical dialog games 16:927
Logical empiricism 10:774, 10:786, 14:313–318, 25:88–92
analytic philosophy 14:314, 14:317
analytic statement 14:317
critical rationalism 14:317
Duhem-Quine thesis 14:317
epistemology 14:313–318
Feyerabend on 9:117–118, 9:121
historical background 14:313–314
logical analysis 14:314–316
logical atomism 14:314
logicism 14:314
metaphysics 14:313–317
neoempiricist circles 14:314–315
Berlin Society for Empirical Philosophy 14:315
Lvov-Warsaw School 14:315
Vienna Circle 14:314–315
phenomenalism 14:313–318
philosophy 14:313–315, 14:317
of science 14:313–315, 14:317
prospects 14:317–318
scientific conception of world 14:315–317
unified science 14:315–317
Logical equivalence 20:69–70
Logical fallibilism, Albert on 1:495
Logical form 9:874–875
Logical formalism 14:295
Logical incrementalism, strategic planning 23:518
Logical inference 3:200
Logical interpretations of probability 19:25
Logical pluralism 16:928
Logical positivism 1:892, 1:894, 14:313–318, 18:619, 25:88, 25:91–92
analytic philosophy 14:314, 14:317
analytic statement 14:317
critical rationalism 14:317
Duhem-Quine thesis 14:317
epistemology 14:313–318
historical background 14:313–314
logical analysis 14:314–316
logical atomism 14:314
logicism 14:314
metaphysics 14:313–317
neoempiricist circles 14:314–315
Berlin Society for Empirical Philosophy 14:315
Lvov-Warsaw School 14:315
Vienna Circle 14:314–315
phenomenalism 14:313–318
philosophy 14:313–315, 14:317
of science 14:313–315, 14:317
Popper on 18:523
prospects 14:317–318
and psychology 19:414
Putnam on 19:662–663, 19:666
scientific conception of world 14:315–317
unified science 14:315–317
Logical positivists and logical empiricists 7:568
Logical probability 19:39–40
Logical problem, of language acquisition 13:248–249
Logical reasoning 14:308–310
Logical semantics of natural language 16:925
Logical truths, Wittgenstein on 25:612
Logic-based systems
knowledge representation 13:100–101
Logicism 14:314
Logicomathematical processes 18:129
Logics for knowledge representation 14:319–321
Logik der Forschung 18:523
Logische Untersuchungen 19:391
Logistic curve 19:782
Logistic equation 3:155–156
Logistic link function
in event–history analysis 14:351–352
Logistic model, oldest-old mortality 15:864
Logistic regressions 19:142
Bernoulli distribution 2:587, 2:590–591, 2:594–595
case-control design 2:589
coefficients, interpretation of 2:589–590
estimation 2:590–591
example 2:591–592
extensions of 2:592–594
for clustered data 2:594
conditional 2:592–593
exact 2:593–594
Fisher's exact test 2:593
generalized linear model 2:594
likelihood-ratio test 2:593
link function 2:594–595
logit 2:588–589, 2:594–595
maximum likelihood estimation 2:590–594
multiple 2:590
odds ratio 2:589–590, 2:593–594
in one variable
locally optimal designs for 17:260
probit function 2:588–589, 2:594–595
relative risk 2:587–590
Wald test 2:593
Logistic smooth transition autoregression (LSTAR) model 24:338
Logistics 14:531
focus groups 9:277
supply chain management 23:709, 23:712–713
Logit models 16:48, 16:51
logistic regression 2:588–589, 2:594–595
loglinear models 16:125–126, 16:129
Log-likelihood 14:117
Log-linear models 13:385, 14:322–327, 16:125–130, 18:604–605
applications and interpretations 14:323–324
capture–recapture methods 18:604–606
chi-square tests 16:128–129
conditional independence 16:125, 16:127
contingency table 16:125–128
discrete multivariate analysis 16:133–134
estimation and testing 14:324
examples 16:125–127
generalizations and applications 14:326
hypothesis tests 16:128
independent variables 16:125–127
logit models 16:125–126, 16:129
marginal models 14:324–325
applications of 14:325–326
directed Markov models 14:326
panel studies 14:325–326
maximum likelihood 16:127–128
model selection 16:129
multinomial distribution 16:125–126
Poisson distribution 16:125–129
regression for categorical variables 14:323
Markov models 14:323–324
residual analysis 16:128–129
Log-logistic model 14:56

Lognormal distribution 11:886, 11:888
Log-normal model 14:56
Logocentrism 5:247, 5:928, 5:930
Logograms 25:763
Logographic scripts 25:780
Logographic symbol 20:1–2
Logographs 20:7–8, 20:10, 25:780–781
Logography
  writing, evolution of 25:762–764
Logopandecteision 12:762
Logopenic variant of primary progressive aphasia (lvPPA) 13:363–364
  language processes in 13:364–366
  spontaneous speech in patients with 13:365–366
Logophoricity 1:678
Logos 23:525
Log-rank test 23:772
  survival analysis 23:772–773
Lokasangraha 11:786
Lombard, Denys, works, *Le Carrefour Javanais: Essai d'Histoire Globale* 11:87
Lombroso, Cesare 5:233–234, 8:239, 13:802
  works, *Criminal Man* 5:233
London Communiqué 2:742
London Court of Arbitration 15:72
London Court of International Commercial Arbitration 12:470
London Edinburgh Weekend Return Group (LEWRG)
  In and Against the State 22:803
London Metropolitan Police Act 1829 18:261
London Missionary Society 17:161
London Peace Society 17:643–644
London School of Economics (LSE) 3:6, 7:9–10, 17:520–521, 22:698
Lone genius, artists as 5:466
Lone Star Saloon 13:890
Loneliness
  and health 22:369–370
  and LGBT aging 13:866
Lone-parent families 2:471, 2:475
Long live love 10:622–623
Long memory, models with 24:313–314
Long Parliament, English Revolution 7:649–650, 7:652
Long Term Ecological Research Network (LTER) 7:792–793
Long term educational planning 7:262–264
Long-cycle theory 25:425
Long-distance agreement (LDA) 10:318–321
Long-distance nationalism 1:761
Long-distance questions 9:207
Longevity 17:175–176, 17:179–180
  definition 17:172
  oldest-old, determinants of 15:865–866
Longevity genes 17:175–176
Longevity of democracy 18:494
Longitude Act 1714 2:333
Longitudinal Adolescent Health Study 9:658–659
Longitudinal analysis
  of I/E model 12:427, 12:429–431
Longitudinal data 14:345–350, 19:734–735
  alternative bottom-up strategy 14:348
  analyses, of sexual development from late childhood through early adulthood 14:328–339
  classes of methods 14:345–346
  comparative databases 5:791–792, 5:795
  event–history analysis in discrete time 14:351–353
  historiographic dimension of 5:832–833
  micro-level surveys 5:821
  modeling 14:163–164
  motivation 14:345
  process dynamics 14:345
  sequential aggregation strategies 14:345
  top-down methods 14:348–349
  understanding psychological resilience 14:346–348
    abstract chronological charts 14:347
    aggregation of reduced response vectors 14:347
    data 14:346
    narratives 14:346–347
    pathway specification 14:347
    person-centered strategy 14:346
    problem 14:346
    reduced response vectors 14:347
    tests of distinguishability 14:347–348
Longitudinal designs 20:243–246
  on adult life development in cognitive aging 1:401–403
  infectious diseases 12:34
  structural equation modeling (SEM) 20:245
Longitudinal education surveys 5:844
Longitudinal Individual Database (LINDA) 5:846
Longitudinal Internet Studies for the Social Sciences (LISS) 5:846
Longitudinal research 3:447
  design features 14:354–356
    locating and tracking respondents 14:355–356
    recruitment and retention 14:354–355
    retaining respondents between interview waves 14:356
  factors associated with panel retention 14:354
  panel retention 14:354–357
  use of incentives to recruit and retain subjects 14:356
Longitudinal studies
  of attachment 12:232
  of crime 1:310–313
  epidemiology 7:828–829, 7:828b
  mental illness 15:183–184
Longitudinal Study of Australian Youth (LSAY) 15:901
Longitudinal twin study 24:692
Long-lasting effects of neonatal bonding 16:472–473
Long-period equilibrium 3:92
Long-range radio navigation (LORAN) 10:174
Long-run coefficients, hypotheses on 24:326
Long-tail retailing strategy, in artistic/cultural markets 14:567
Long-tenured displaced worker, job displacement insurance 12:830
Long-tenured job 12:825
  job displacement from 12:819, 12:822t
Long-term depression (LTD) 4:482, 11:613–614, 12:713, 16:620, 23:873, 23:875–876, 25:73, 23:875f
  Alzheimer's disease 14:363–364
  expression 14:361–362
  eyeblink conditioning 8:637
  hippocampus 14:358–364
  history 14:358–359
  induction 14:359–360
  learning 14:358–360, 14:362–364
  maintenance 14:362
  metaplasticity 14:359
  putative roles in disease and memory impairment 14:362
  studying in human brains 14:362–363
Long-term epigenetic inheritance 7:835
Long-term fate of preverbal memories, during childhood 3:467
Long-term heterosexual relationships
  commitment in 20:912–913
    influencing factors 20:914–916
  difficulties in 20:913
  profound satisfaction in 20:912
    influencing factors 20:914–916
    maintaining 20:916–917
  satisfaction and romantic compromises 20:913–914
  superficial pleasure in 20:912
Long-term learning 21:913–914
Long-term memory (LTM) 1:644–645, 2:289, 15:140–141, 15:158, 15:162, 20:51–53, 20:613, 21:896–898, 21:901, 23:874–879, 25:707, 13:678f
  attention and 4:104–105
  binding and association formation in 15:143–144
  of children 3:465–466
  episodic memory, proactive interference in 19:3
  mechanisms of change in 21:898–899
  mental images generated from 15:221
  primacy effect 4:104–105
  retroactive interference in 20:613, 20:618
  and sleep 22:63
  temporal distinctiveness 24:204
Long-term orientation, and values 25:30
Long-term physical disability, and natural disasters 16:316–317
Long-term potentiation (LTP) 4:482, 7:851, 11:613–614, 12:713, 14:128, 16:620, 20:617, 23:873, 23:875–879, 25:73, 23:875f
  Alzheimer's disease 14:363–364
  expression 14:361–362
  eyeblink conditioning 8:637
  hippocampus 14:358–364
  history 14:358–359

Long-term potentiation (LTP) (*continued*)
induction 14:359–360
learning 14:358–360, 14:362–364
maintenance 14:362
and memory 9:906–907
metaplasticity 14:359
putative roles in disease and memory impairment 14:362
and social isolation 16:489
studying in human brains 14:362–363
Long-term preservation, journals 12:865
Long-term retention, and memory development 15:149
Long-term sensitization effect 10:482
Long-term socioecological research 22:259
Long-term survival, childhood cancer 3:462
Long-term unemployment 13:175, 13:180–181
and social deprivation 22:286
Longue durée 11:360–361
Longwave wireless 12:492
Look-and-say method 23:266–267
Looked-After-Children (LAC) 19:303–304
Looking, infant gaze behaviors 12:28–29
Looking glass self 8:464, 21:871–872
Looking-time technique 4:46
Loop beliefs 21:877–880
Loose coupling 16:765
Loosely coupled systems, modularity and 15:718, 15:721
Lopreato, Joseph 8:463
LOR *see* Latitude of rejection (LOR)
LORAN *see* Long-range radio navigation (LORAN)
Lord North 1:609
Lordat, Jacques 18:45–46
articulatory dynamics 18:45–46, 18:46f
articulemes 18:46
gesturemes 18:46
on speech of adults with apoplexy 18:46–47
wickelphones 18:46
Lorde, Audre 8:908, 19:747
Lord's Resistance Army (LRA) 16:806, 20:342
Lorenz, Konrad 14:366–371, 16:469
biography 14:366–367
education 14:366
family 14:366
retirement 14:367
scientific career 14:366–367
contribution to knowledge 14:367–368
behavioral modules 14:367
environmental awareness 14:368
epistemology 14:368
ethograms 14:367
human morality 14:368
imprinting 14:368
innate forms of possible experience 14:367–368
moral-analogous behavior 14:368
nature–nurture 14:368
observation as research tool 14:367
Nobel Prize for medicine 14:367
relevance 14:368–370
politics 14:369–370
selected topics 14:369
works
On Aggression 14:369, 22:900
The Waning of Humaneness 14:369
Lorenz curve 11:732, 11:735, 11:888, 11:901–903, 11:905–906, 11:903f
Lorenz dominance 14:450
Lorenz ordering 14:451
Lorenz stable set 4:871
Los Angeles Games 14:722
Los Angeles immigrant labor activism 24:492–493
The Los Angeles Times 2:510
Los Hombres de Presa 5:239
Lösegeld 9:668–669
Loss aversion 2:438, 16:420, 16:642–644
Loss functions 8:17, 23:413
Bayesian experimental designs 8:523
Loss-based selection, in self-regulation 21:559
Losses 5:862
bereavement 2:531–536
experience across life span 2:538–539
and integrated attachment theory 12:229
multiple language 16:61
of ontological certainty 3:540–541
of organizational group diversity 10:409
value pluralism 25:23–24
LOT *see* Life orientation test (LOT)
Lot design and density measures 24:825–826
Lothian Birth Cohort (LBC) 7:139
Lotka, Alfred James 6:173, 14:372–374
biography of 14:372
on cycles of water, iron, chlorine, and other components of earth's crust 14:373
as editor of *Scientific American Supplement* 14:372
on law of productivity 20:530
on modern ecology 14:373
on possibility of man-made carbon dioxide leading to global warming 14:373
works
Elements of Physical Biology 14:373
Lotka–Leslie age-specific models 18:581
Lotka–Volterra differential equations 14:373, 14:792
LOT-R *see* Life orientation test-revised (LOT-R)
Lots of Copies Keep Stuff Safe (LOCKSS) Program 14:47–48
Lottery of the long ballot 13:12
Lotus Sutra (Saddharma-Pundarika) 2:888, 20:366–367
Lotz, A. 25:81–82
Lotze, Rudolph Hermann 19:411
Lou Movement 16:806
Loudness, Stevens on 23:453–455
Louis, Paul, works, *Le Colonialisme* 11:686
Louis Agassiz of Harvard 20:287
Louis XVIII, King 13:383
Louvre in Paris 16:142–143
Love 14:375–381, 21:813–815, 21:818
attachment 14:377–380
biopsychological perspective 14:377–378
commitment 14:379
companionate love 14:377–380
compassionate love 14:377–378, 14:380
and emotions 7:513–516
and friendship 9:427
for Honneth 11:176–178
and long-term heterosexual relationships 20:912–917
love stories 14:379–380
and marriage 14:603
and mutuality 5:308
passion 14:377–380
and personality 17:908
romantic love 14:377–380
sex 14:375, 14:377–380
sociology, Randall Collins 4:214–216
taxonomies 14:378–380
duplex theory 14:379–380
prototypes 14:380
quadrumvirate 14:378–379
styles 14:379
theory of love as story 14:379–380
triangular theory 14:379
virtuous 25:148
Love of God, and self-love 12:387–388
Love relationships 9:455
Love stories 14:379–380
Lovejoy, Arthur O. 12:255–258, 12:761
Lovelock, James 9:567, 20:288
Lover 17:486–489, 17:491
Lovers of wisdom 7:255
Low- and middle-income countries (LMICs) 22:125, 22:128
acceptable risk from withholding treatment in 18:168
Low context societies 12:357
Low external input agriculture, peasant agriculture 17:665
Low fertility 9:19–20, 9:22, 9:24
Low Income Housing Tax Credit Act 1986 18:324
Low intervention clinical trial 24:284–285
Low level motor disturbances 1:874
Low mood 6:202–203
Low roads, of transfer of learning 21:97–98
Low sainthood 15:577
Low self-control theory, applicability, and white-collar crimes 25:555–556
Low sexual desire, of men 21:714, 21:717
Low tech measures of implicit attitudes 2:181
Low threshold theory (LTT) 19:473–474
Low-carbon city 6:831–832
development of 6:829, 6:831–833
see also Eco-cities
Low-context communication 4:284–285
Low-copy-number (LCN) DNA profiling 9:957–958
Low-cost hypothesis 7:699–700, 8:27
Löwenheim–Skolem theorem 19:665
Lower Mississippi River Industrial Corridor 7:760
Lower-asset neighborhoods 3:447–449
Lower-class people 23:495

Lower-level constructs
and FFM dimensions 2:561
of personality 17:865–867
LOWESS *see* Locally weighted scatterplot smoothing (LOWESS)
Lowest-low fertility 9:775–778, 21:347
Lowey, Robert 2:712
Low-income citizens 3:629
Low-income countries 9:64–65
deinstitutionalization psychiatry in 6:42–46
health in 10:633–639
Low-income housing, provision 1:219
Low-Income Housing Tax Credit Program (1986) 18:736
Low-pass filter 24:335
Lowry Model 24:808
Low-structured therapy, offender rehabilitation 20:195
Low/very low birth weight 2:698–699
Loyalty, and administration 1:106
Loyalty swore 16:258–259
Loyalty-producing system-building, processes of 20:746
LP *see* Limiting-pathway (LP) models of epistasis
LPC *see* Least preferred co-worker (LPC) model; Linear predictive coding (LPC) of speech
LPFC *see* Lateral prefrontal cortex (LPFC)
LPMC *see* Lateral premotor cortex (LPMC)
LPP *see* Late positive potential (LPP)
LPT *see* Labor process theory (LPT)
LQ *see* Laterality quotient (LQ)
LQS *see* Limited quantity scarcity (LQS)
LRA *see* Lord's Resistance Army (LRA)
LRAP *see* Loan Repayment Assistance Program (LRAP)
L-RAT *see* Lifestyle–routine activities theory (L-RAT)
LRE *see* Law-related education (LRE)
LRM *see* Legal Reform Movement (LRM)
LRT *see* Likelihood ratio tests (LRT)
LSA *see* Latent semantic analysis (LSA); Law and Society Association (LSA)
LSAY *see* Longitudinal Study of Australian Youth (LSAY)
LSB *see* Law Council of Australia; Legal Services Board (LSB)
LSC *see* Legal Services Corporation (LSC); Legal Services Corporation (LSC) Act 1974
LSD *see* Lysergic acid diethylamide (LSD)
LSE *see* Least squares estimate (LSE); London School of Economics (LSE)
LSEMs *see* Linear structural equation models (LSEMs)
LSM *see* Lesbian Sex Mafia (LSM)
LSMS *see* Living Standards Measurement Study (LSMS)
LSRM *see* Laura Spellman Rockefeller Memorial (LSRM)
LSTAR *see* Logistic smooth transition autoregression (LSTAR) model
L-statistics 17:293
lSTM *see* Late short-term memory (lSTM)
LTD *see* Long-term depression (LTD)
LTER *see* Long Term Ecological Research Network (LTER)
LTG *see* Limits to growth (LTG) study
LTM *see* Long-term memory (LTM)
Ltm (software) 19:421
LTP *see* Long-term potentiation (LTP)
LTS *see* Limited time scarcity (LTS)
LTT *see* Low threshold theory (LTT)
L-type calcium channels 20:215
LUBA *see* Land Use Board of Appeals (LUBA)
Lubin, N., on Central Asia 3:317
Lubunya (Turkish) 20:872
Lucas, Robert E. 8:512–513
on macroeconomics 14:403–404
Luce, R.D.
beta learning model of 14:589–590
works
Handbook of Mathematical Psychology 14:814
Luce's choice axiom (LCA) 5:895–900, 5:902, 5:904, 19:475
applications in
learning 5:899
preferential decisions 5:898–899
future directions 5:900
historical connections 5:896–897
implications 5:896
relation with LBA model 5:904
validity of 5:897–898
Luce's ratio rule 14:813
Lucie-Smith, Edward, works, *The Bear Cult: Photographs by Chris Nelson* 13:890
Luckmann, Thomas 17:393–394, 19:677, 20:346
on cognition 5:577
on institutions 17:394
works
The Social Construction of Reality 17:391, 19:491
Ludendorff, Erich Friedrich Wilhelm 9:213
Ludicrous, and humor 11:416
Lüdtke, Alf 11:111
Ludwig, Carl 24:763
Lugbara 6:774–775, 13:71
Luhmann, Niklas 6:601–602, 13:439, 13:554–555, 14:382–389, 22:715–716, 22:913, 22:915
on autopoiesis 9:541
biography of 14:382
on conceptual history 4:552
on functionalism 9:541
on intellectual development 14:382–383
on meaning 16:561
on social systems 14:383–386, 14:390–395
action and experience 14:384
actions 14:394
autopoietic systems 14:391–395
boundaries 14:394
building social structures through communication 14:392–393
closed systems 14:391
communication and action 14:384–385
communication elements as events and operations 14:385–386
communication theory and double contingency 14:385
complexity 14:384, 14:386
double contingency 14:390–391
function and causality 14:383
functional differentiation 14:388
interpenetration 14:394–395
meaning 14:384, 14:391–392
meaning as selection 14:391–393
moment-to-moment choices 14:390–392
operation and observation 14:386
psychic systems 14:384, 14:390–391, 14:393–395
self-reference, autopoiesis, operational closure 14:386
system and environment 14:383–384
systems and environments 14:393–394
world with many centers 14:394
on society 14:386–388
differentiation theory 14:387–388
interaction and organization 14:386–387
theory of sociocultural evolution 14:387
theory of symbolically generalized media of communication 14:387
world society 14:386–389
on systems theory 9:541
on theory building and relevance for future social science 14:388–389
works
Weltgesellschaft 24:589
Luhrmann, Tanya, works, *Persuasions of the Witch's Craft* 20:310
Luirink, Bart, on GASA 21:745
Lukács, György 3:741, 5:244, 5:499, 21:171
Luke's Acts of the Apostles 3:547–548
Lumbreras, L.G., works, *La Arqueología de la América Andina* 23:36–37
Lumpa Church of Alice Lenshina 15:532
Lumpenproletariat 1:681, 22:888–889
Lumping position, classifiers from 3:817
Lumsden, Charles, works, *Promethean Fire* 22:903–904
Lunacy Act 1890 19:301, 19:308–309, 19:587
Lunar cycle 16:652
Lunate sulcus
brain, evolution of 2:828, 2:830
Lundberg, George 22:699
geometrical assumptions 25:199
works
Trends in American Sociology 7:588–589
Lung cancer 22:125–126
Fisher's theories 9:224
and mortality 15:860
Lunneborg, Clifford 12:304
Luo, Zhenyu 11:52
Luong, Jones 3:319
Lupercalia 3:145
Lurasidone 19:201
Lure 25:553

Luria, Alexander Romanovich 12:308, 14:396–399, 16:679–680, 18:57–58, 19:669–670, 16:679f
on brain lesions 14:397–398
on brain-behavior relationships 14:398
on combined motor method 14:396–398
on cross-cultural psychology 14:397
on cultural-historical psychology 14:397
on functional system 14:396, 14:398
on idiographic science 14:396, 14:398
on nature–nurture question 14:397
on neuroimaging 14:398
on neuropsychological rehabilitation 2:815
on nomothetic science 14:396, 14:398
on romantic science 14:398
on twins 14:397
works
Basic Problems of Neurolinguistics 16:679–680
The Man with a Shattered World: The History of a Brain Wound 2:815
The Nature of Human Conflicts 14:396–397
Luria Neuropsychological Battery 14:398
Lurker problem, in virtual communities 4:334
Luso (African community) 3:313
Luso-tropicalism 9:423
Luteinizing hormone (LH) 10:462–463, 10:465–467
Luther, Martin 2:145
Reformation/confessionalization 20:98–103
Lutheran church records 14:238
Lutheranism 18:917
Lutherans 3:542–543
Luxembourg Employment Study (LES) 5:801
Luxembourg Income Study (LIS) 5:744, 5:747, 5:801, 18:755
Cross-National Data Center 11:739
Luxemburg, Rosa 3:740–741, 11:686, 25:748–749
works, *The Accumulation of Capital* 3:95
Luxury 4:751
Luyia 8:177
LV *see* Limited vote (LV)
LVCSR *see* Large vocabulary continuous speech recognition (LVCSR) task
Lvov-Warsaw School 14:315
lvPPA *see* Logopenic variant of primary progressive aphasia (lvPPA)
Lwein, Kurt, works, *Field Theory in Social Science* 16:525
Lycées (*musées*) system 19:613, 25:364
Lyell, Charles 5:676, 20:287
works, *Principles of Geology* 5:676, 7:748, 20:287
Lyft 4:139–140
Lying, and prosocial behavior 19:228–229, 19:231
Lynch, Kevin 24:871
works, *The Image of the City* 4:79
Lynch, Michael 8:577–578
works, *Art and Artifact in Laboratory Science* 8:577–578
Lynchings 25:123
Lynd, Helen 20:669–670
works
Middletown: A Study in Modern American Culture 4:362, 17:22–23
Middletown in Transition 7:562
Lynd, Robert Staughton 1:755, 20:669–670
works
Middletown: A Study in Modern American Culture 4:362, 17:22–23
Middletown in Transition 7:562
Lyon Martin Women's Health Services 15:113–114
Lyotard, Jean-François 4:203–204, 11:10, 14:200
historical explanation 11:10–11
Lysenko, Trofim D. 8:229
Lysenkoism 8:234, 16:14–15
nature/nurture 16:341–342
Lysergic acid diethylamide (LSD) 5:81, 10:502, 16:170

M

M-15 movement 22:518
Ma Jun-Wu 24:585
Ma scale 15:567
Maas, Ineke, on cultural capital 5:377–378
Maasai tribe 6:773–775
Maastricht Treaty 2:741, 17:513–514, 24:599
MAB *see* Man and the Biosphere (MAB) program
Macadam, Elizabeth 12:477
MacAndrew Addiction scale (MAC-R) 15:567
Macaque monkey
brain of 17:127, 17:129–130
and human language evolution 11:323–326
MacArthur Communicative Development Inventory (MCDI) 13:325
MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR) measures 20:516–517
MacArthur Competence Assessment Tool for Treatment (MacCAT-T) 19:323–324
MacArthur Foundation 5:128
Macaulay, Stewart 13:470, 13:517
Macaulay, Thomas Babington 23:61
works, *The History of England from the Accession of James II* 3:61–62
MacBride Commission, *Many Voices, One World* 15:56
Maccabean martyrs 3:546
MacCAT-CR *see* MacArthur Competence Assessment Tool for Clinical Research (MacCAT-CR) measures
MacCAT-T *see* MacArthur Competence Assessment Tool for Treatment (MacCAT-T)
MacClancy, Jeremy 1:741
Maccoby, E.E., works, *The Psychology of Sex Differences* 21:693
MacDonald, Charles, on anarchism 1:682–683
Macdonald, S., on tourism 2:872
Mach, Ernst 14:453, 18:621, 18:625, 19:410–411, 23:425
Machian positivism 18:621–622
Machiavelli, Florentine Niccolò 3:645, 3:772, 11:413, 21:212–213, 22:696–697
elites 7:390–392
republicanism 20:477–479
works
Discorsi 3:772
The Prince 5:672, 23:353
Machiavellian absolutism (Tacitean) 1:10
Machiavellian power-dependence analysis 17:401
Machiavellianism 12:761–762
and personality 5:672
Machine learning 1:531
genetic programming 2:41
Machine tabulation, census 3:304–305
Machine translation (MT) module 23:261
Machine-based manufacturing 11:856
Machinery of government 19:521, 19:523–524
Machines
speech production by 23:259–260
speech recognition and production by 23:259–263
speech recognition by 23:261–263, 23:261f
Machismo 9:628–630
Machlup, Fritz 12:84
Machu Picchu in Peru 23:36
MacIntyre, Alasdair 4:341–342, 24:39
works
After Virtue 4:319
Marxism: An Introduction 4:319
Marxism and Christianity 4:319
A Short History of Ethics 4:319
Whose Justice? Which Rationality? 4:320
MacIver, Robert 2:271
Mackay, Charles 4:157
MacKenzie, Donald, on Black-Scholes model for pricing 5:540
Mackey, E., on democracy in Canada 17:25
Mackie, J.L.
on counterfatual reasoning 5:88
on division of labor 6:244
Mackinder, Halford 10:92, 18:374
MacKinnon, Catharine
on pornography 13:884–885
on sexual abuse of women 9:730
MacLean, P.D.
theory of triune brain 7:468
MacNeille completion 17:528
MAC-R *see* MacAndrew Addiction scale (MAC-R)
Macro databases 5:748–749
Eurostat's National Accounts 5:749
IMF's World Economic Outlook 5:749
National Accounts Main Aggregates Database 5:748

OECD Statistics Portal 5:748
OECD System of Composite Leading Indicators 5:748
UNCTAD statistics 5:749
World Bank databases on economic and growth 5:749
Macro evolutionary theory of law 13:565
Macro level
intersectionality and crime at 12:666–667
rational choice theory extensions at 19:901–902
Macro level ID 12:196
Macro practice, social work education 22:775–776
Macro theory, Bielefeld School 2:554–555
Macroanatomical brain lesions, moral behavior in patients with 16:607–609
Macro-autonomy 21:586
Macrobehavior 13:588
Macrocontextual views, on health 22:278–279
Macro-control theories 4:836–837
Macro-cosmos, migration history 15:446
Macro-data
comparative 5:789
obtaining 5:790
social and economic data 5:820
Macrodemography 11:7–8
Macroeconometric models, and macroeconomics 14:401–402
challenges 14:403–404
Macroeconomic data 5:751–757
cost of living index (COLI) 5:753–754
EKS method 5:756
Fisher index 5:753–754, 5:756
Geary-Khamis method 5:756
gross domestic product and gross value added 5:751–752
from gross domestic product to national income 5:752
international comparisons 5:755–756
purchasing power parities 5:755–756
real expenditures and real gross domestic product 5:755–756
International Comparisons Project (ICP) 5:755–756
international system of macroeconomic accounts 5:752–753
from national income to national accounts 5:752
real economic aggregates and inflation 5:753–755
economic approach to index numbers 5:753–754
gross domestic product and welfare 5:754–755
index numbers 5:753
sequence of account in SNA 5:753
Stone on 5:752
substitution effect 5:753–754, 5:756
Macroeconomic performance, and poverty 18:743–744
Macroeconomic theory 3:92
Macroeconomics 13:18, 13:21–22, 16:583, 20:661, 22:729
business-cycle model 14:404
Cowles Commission 14:401
equilibrium analysis in 7:917
and fertility transition, in East Asia 9:57
focus of 7:22
Harrod–Domar model 14:402
history of 14:406–413
aggregation 14:400–404
analysis of aggregate output and employment 14:411
business cycle 14:406–407, 14:410–411
challenge to macroeconometric modeling 14:403–404
contributions of Friedman 14:401–403
contributions of Frisch 14:400–401
contributions of Hicks 14:401
contributions of Klein 14:401, 14:403–404
contributions of Lucas 14:403–404
contributions of Tinbergen 14:400–403
crises 14:406, 14:408–410
depressions 14:406–409, 14:411
DSGE model 14:404
dynamics and growth 14:410–411
economic growth 14:402
fluctuations as cycles 14:410
inflation 14:406–407, 14:409, 14:411
interest rate 14:406, 14:409
Lucas critique 14:403–404
microfoundational programs and macroeconometric models 14:401–402
monetarist challenge 14:402–403
monetary theorists and economic stabilization 14:409
new classical macroeconomics 14:403–404
new neoclassical synthesis 14:404
origins of modern macroeconomics 14:400–401
Phillips curve 14:402–404
prehistory 14:406–407
quantity theory of money 14:406–409
rational expectations 14:403–404
representative-agent model 14:403–404
1886's 14:408–409
theory of money and prices 14:407–408
World War II and Keynesianism 14:401
IS-LM model 14:401
Klein–Goldberger model 14:401
Lucas critique 14:403
model-based extrapolation in 9:844
national accounting 23:426
representative-agent model 14:403–404
Solow–Swan model 14:402
Macroepistemic cultures 7:878–879
Macroergonomics 11:302–303
Macroextremism 14:415
Macrointerpretive explanation 12:646–647
Macro-justice, Kolm's theory of 7:907
Macrolevel correlates of crime 12:463–464
Macrolevel general strain theory 23:498–499
Macrolevel orientations 21:150
Macrolevel patterns of interaction 22:704
Macrolinguistics 14:228–229
Macromethods 12:645–647
Macro–micro–macro scheme 1:75
Macro–micro–macro theory of collective action 10:17–18
Macromodels, family and household demography 8:727–728
Macro-phylum 11:820
Macropractice, social work with refugees 22:770–771
Macroprudential stability 15:733
Macroprudential supervision 9:195
Macroremains, paleoethnobotany 17:456–457
Macroscale geographic models 14:786
Macroscopic asymmetries 2:809
Macrosimulations 8:728
Hutterite fertility 11:446
Macrosocial level, of stigma analysis 23:465
Macro-sociolinguistics 13:353
Macrosociology 14:414–421, 22:994
agent-based modeling 14:418–419
analytical sociology 14:417–419
balance 14:414–415
critical realism 14:416, 14:419
Eisenstadt's contribution 7:321–322, 7:324
emergenist sociology 14:414, 14:416
extremisms 14:415
ideology in 11:575–576
layered and 'flat' social ontologies 14:416
levels of analysis 14:414–419
levels of reality 14:416–417
linkage 14:415–416
mechanism-based thinking 14:417–419
modernization theories 15:701–702
rationality 14:418
scale-based thinking 14:417
social interactions 14:415–416
social network analysis 14:418–419
structural methodological individualism 14:416–418
theory of action 14:418
Macrostructure of cognition in brain 4:96–98
evidence of interactive processes in brain 4:97–98
modular *versus* interactive approaches to organization of function 4:97
specialization of brain regions 4:96–97
Macro-symbolic conception, of integration 12:235–236
Macrosystems 10:681, 21:682–684
and contextual studies 4:787
ecological systems theory 6:840
families as educational settings 8:732
and health 6:880
Macy v. Department of Justice, 2012 21:771
MAD *see* Median absolute deviation (MAD)
Mad, sanism 15:200–203
Madagascar, physiography 3:172f
Madame D 25:164–165
MADD *see* Mothers Against Drunk Drivers (MADD)

Maddox, Richard, on anarchism 1:684
Madhhab 12:746
Madison, James 3:675, 12:897–898, 19:538, 20:279
  works, *Federalist Number Ten* 1:617
Madrasahs 11:62
Madrid International Plan of Action on Ageing (MIPAA) 7:336
Madrigal v. Quilligan 8:251
Maeckelbergh, Marianne, on anarchism 1:685
Maffi, Luisa 1:224
Mafias 5:173, 14:422–427
  action-sets 14:424
  anthropology of 14:422–428
  aristocracy 14:423
  collusion 14:425
  criminal prosecution 14:424
  criminology of 14:429–434
    decline 14:433–434
    emergence 14:430–431
    enforcement of law 14:433–434
    expansion 14:432–433
    function 14:431–432
    governance 14:430
    hierarchy 14:432
  cultural codes 14:425
  friends 14:422
  as *instrumentum regni* 14:427
  and land reforms 14:422–423
  network structure 14:422–425
  origin 14:422
  protection 14:422, 14:424
  as social evil 14:425
  and transnational organized crime 24:596, 24:598
  violence 14:422–424, 14:426–427
Mafioso 14:430
Magdalenian migrations 15:474f, 15:476f
Magden, Perihan 20:873
Al-Maghili 12:737
Maghreb, sexual minorities in 21:750–754
  background 21:750
  film 21:753
  literature, 1960–1990 21:752–753
    1990–present 21:753
    1950s 21:752–753
  social sciences 21:751–752
  Western (mis)understandings of sexuality in 21:750–751
    ahistoricism 21:751
    homogenization 21:751
    orientalism 21:750–751
Magic
  anthropology of 14:435
    social evolutionism 14:435–436
    twentieth-century paradigms 14:436
  culture/ontology 14:438
  and ethnocentrism 14:435–436, 14:438
  and evolutionism 14:435–436
  and human unity 14:435
  in Japan 12:796
  and rationality 14:435–436, 14:438
  and religion 14:436–437
  and science 14:435–436
  and symbolism 14:436
  witchcraft 25:605–608
  and witchcraft 14:436–438
Magic circle 13:579
Magic of modernity 9:107
Magic shrinking machine
  during childhood 3:466–467
Magill, Frank
  works
    International Encyclopedia of Psychology 7:591–592
    International Encyclopedia of Sociology 7:588
    Survey of Social Science: Psychology Series 7:591–592
Magister 6:62
Magister vitae 11:61
Magna Carta 6:696–697
Magnet schools 6:218
Magnetic prospecting 10:54
Magnetic resonance imaging (MRI) 14:439, 17:861
  for ADHD 2:156–159
  cognition 14:439–441, 14:443
  dementia 14:443
  diagnosis 14:443
  findings 14:441–444
    affective disorders, studies 14:441–443
    Alzheimer's studies 14:443–444
    schizophrenia 14:441
  functional 14:439–440
    blood oxygen-level-dependent (BOLD) fMRI 14:440
    diffusion-tensor fMRI 14:440
    general aspects 14:439
    magnetic resonance spectroscopy 14:440
    perfusion fMRI 14:440
  for language learning impairment 13:347
  LTP/LTD 14:363
  neurobiology 14:439, 14:444
  pathophysiology 14:441, 14:443–444
  positron emission tomography 14:439–444
  in psychiatry 14:439–445
  psychosis 14:441–442, 14:442f
  single photon emission computed tomography 14:439–441
  structural 14:439
    principles 14:439
Magnetic resonance spectroscopy (MRS) 14:439–440
Magnetic sleep 11:465
Magnetic tape 23:378
Magnetism 11:472
  early studies 19:502
Magnetoencephalogram (MEG) 8:664–665, 9:476, 9:479–481, 11:603, 11:616
  benefits of 9:481
  brain oscillations investigated 9:483–484
  detecting changes in brain activity 9:479
  detectors of 9:479
  infant cognitive development 4:60
  intracarotid sodium amytal testing 12:690–691, 12:694
  investigating language with 9:481–487
  for language learning impairment 13:348
  neuromarketing 16:698–700
  visual word recognition 25:215–216
Magnificat Meal Movement 16:806
Magnitude estimation method 14:802–803, 21:605
Magnitude estimation method, Stevens on 23:454
Magnitude of marriage attraction 18:588
Magnitude scaling 21:604
Magnitudes
  and measurement theory
    extensive magnitudes 14:868
    intensive magnitudes 14:868–870
  and numbers 17:50–51
Magno cells 17:129
Magnus, Albertus 7:721
Magnus, Hugo 4:231–232
Maguire, Jennifer Smith, on culture and consumption 5:538
Magyar Igazság és Élet Pártja (Hungarian Justice and Life Party) 18:411
Maha Bodhi Society 2:891
Mahalle 3:319
Mahan, Alfred Thayer 18:374
Mahaprabhu, Chaitanya 16:805
Mahayana Buddhism 2:888, 2:891, 15:531, 20:328–329
  in Japan 12:796
  women in 25:631–632
Mahbubani, Kishore 22:637
Maher, V., on customized craft production 5:111
Mahikari (new religious movement) 16:805
Mahler, Margaret 11:537–538
Mahmood, Saba, works, *The Politics of Piety* 8:900, 20:311
Mahomo, Nana, works, *Last Grave at Dimbaza* 6:622–623
Mahoney, Jim 14:873
Maillart, Ella, on Russian sexual life 21:740–741
Main, Mary, attachment theory 12:230–231
Main clauses, ordering of 1:186
Main olfactory system (MOS) 17:196–198, 17:196f
Maine, Sir Henry James Sumner 13:565–566, 13:734–735, 13:802, 23:61
  on status and contract societies 1:766
  works
    Ancient Law 4:249, 13:565–566, 13:802, 19:153
Mainframes 23:377–378
Mainland–Gart test 5:351
Mainline Protestants (MPs) 20:292
Mainstream 23:637
Mainstream arts 2:17
Mainstream economics 1:573
Mainstream elder care services 15:442–443
Mainstream game theory 18:931
Mainstream labor market (segmentation) theory 13:164
Mainstream scholarship, race and law 19:807

Mainstream terrorism studies 22:431
Mainstreaming 15:30
Mainstreaming Environment and Sustainability in African Universities (MESA) program 7:734
Maintenance
  and adaptation 15:370
  animal's life history 9:40
  of health 10:674–675
  multiple language 16:58–59
Maintenance of Fair Trade 3:31
Maintenance (coping) self-efficacy 10:711
Mair, Johanna, on institutional work 17:394
Maison des Sciences de l'Homme 11:40, 11:76, 22:669–670
Maistre, Joseph de 12:741–742
Maitreyi 7:374
Maiwandi, Abdul Sattar 24:231
Maize, phytoliths 17:459–460
Majestas absoluta 1:7–8
Maji Maji Rebellion 15:532
Major continuous gateways, in United States 11:631
Major depression, prophylaxis 19:203
Major depressive disorder (MDD) 2:534, 6:824, 11:613, 11:616–619, 15:213, 16:614, 16:616, 18:703
  role of DNA methylation in 7:843
Major Episodes of Political Violence (MEPV) dataset 5:836, 8:97–98, 8:98t
Major histocompatibility complex (MHC) 11:335
  human mate choice based on 11:336–337
Major League Baseball (MLB) 23:302–303
Major League Baseball International (MLBI) 23:304
Major transitions, kinship and 13:48–49
Majoritarian 7:350–351
Majoritarianism 14:446–448
  in complex institutions 14:447
  democracy 14:446–447
  democratization 14:446–447
  egalitarianism 14:446–447
  electoral system 14:447
  fairness 14:446–447
  minority rule 14:446–447
  preference efficiency 14:447
  proportional representation 14:447
  voter preferences 14:446
Majority 15:569–573
Majority decision rule 12:909
Majority ethnonationalism 8:92
Majority groups, influence of 2:170–171
Majority identity 9:435
Majority language/speaker 2:580–584
Majority rule 14:446–448
  democracy 14:446–447
  democratization 14:446–447
  egalitarianism 14:446–447
  electoral system 14:447
  fairness 14:446–447
  and formal rationality 14:447
  minority rule 14:446–447
  preference efficiency 14:447
  proportional representation 14:447, 19:209
  voter preferences 14:446
Majority status, and Jewish Diaspora 12:802
Majority World 6:839
Majorization order 14:449–452
  applications and illustrative examples 14:451–452
  concordance ordering 14:450–451
  convex function 14:450
  distribution function 14:449–451
  hazard rate order 14:450
  likelihood ratio order 14:450
  probability 14:449, 14:451–452
  random variable 14:449–452
Makdisi, Ussama, works, *The Culture of Sectarianism* 16:384
Make work pay 13:185
Make-believe, myth as 16:177
Making history 4:181
Makwe language, word order in 25:645–646
Mala in se crimes 15:589
Mala prohibita crimes 15:589
Maladaptive anxiety 8:872
Maladaptive guilt 21:477
Maladaptive human behavior 7:702–703
Malapportionment 7:342–343
Malaria 1:283, 18:174–176
  child mortality cause 12:23
  Markov models for 14:578, 14:578t
Malarial resistance
  gene-culture coevolution 9:815–816
Malatesta, Errico, on queer anarchism 19:747–748
Malawi, sociocultural aspects 23:75, 23:77
Malay
  Annals 11:83
  and Arabic script 11:82–83
  customary law 13:813
  from Islamic writing 11:82
Malay Annals (*Sejarah Melayu*) 11:83
Malay Islamic Monarchy 12:729
Malay nationalism 12:728–729
Malayo-Polynesian languages 25:288–289
Malaysia 23:63, 23:67
  archaeologists 23:67–68
  economy of 24:897
  fertility transition in 9:85, 9:87–89
  New Economic Policy 4:741, 6:598
  sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Maldives
  education in 7:196
  sociocultural aspects *see* South Asia; sociocultural aspects of
Male bonding 10:554, 25:673
Male breadwinner–female homemaker model 14:606
Male coercion, polygynous marriage as 14:615–616
Male condoms 21:797
Male dominance, and matrifocality 14:822
Male fertility 9:3
Male orgasm 21:690
Male sexual debut 14:336
Male students, academic motivation in United States 9:677–678
Male time 24:408
Male warrior hypothesis 8:445
Male wives 19:752–753
Male-dominated female majority profession 14:679
Male-dominated settings, women working in 21:682
Male–female differences, in health 15:166–168, 15:170
Male/female primates, reproductive strategies of 18:897–898
Males, gender differences in cooperation/competition in 9:753
Males Courting Males 10:554
Malfattori (evildoers) 14:422
Malfunction 9:521
Malia, Martin, on sexual minorities of Russia 21:740
Malik b. Anas 12:735
Malingering 15:567, 16:722–723
Malinowski, Bronislaw Kasper 7:13–14, 7:145, 7:147–148, 8:178–179, 9:671–672, 10:139, 14:453–456, 18:795, 19:677, 21:128, 23:1
  on anarchism 1:681–682
  on anthropology 1:746, 14:453
  in development of British social anthropology 14:453
  ethnography studies of 1:767
  functionalism 9:532–533, 9:537, 9:672
  on functionalism 14:453–455
  functionalist holism 9:672
  on humor 11:416
  kula ring 13:139–142
  legacy of 14:455–456
  on magic 14:435–436
  Malinowskian fieldwork and ethnographic representation 14:456
  method of total immersion 14:454
  scientific theory of culture 14:455
  as teacher, pundit, and propagandist 14:455
  Trobriand Islands studies 1:767
  on Trobriand Man 14:454
  works
    Argonauts of the Western Pacific 9:135, 9:137
    The Family Among the Australian Aborigines 14:453
    Magic, Science and Religion 14:436
    The Natives of Mailu 14:454
Mallard, Grégoire, on epistemic cultures 7:876
Malleability 4:441
  tangible user interfaces (TUI) 24:24
Malleus Maleficarum of 1486 25:622
Mallows's *Cp* 15:648
Malmo Declaration 20:535
Malnutrition 3:870–871
  and children/infant mortality 12:23–24
  and schooling 8:753
Malone v. United Kingdom (1985) 18:950
Malouf, David 2:239
Maltese Cross model 2:858
Malthus, Thomas Robert 1:636, 6:170, 6:172–173, 11:5–6, 14:457–461,

16:228, 18:730, 19:124, 20:660–661, 20:663, 24:728
approach to economics 14:459
contribution to sociology and economy of populations 14:458–460
lowest stage of human society 14:459
reproductive power of human population 14:459–460
demographic science 14:459
'effective demand' for labor 14:459
evolutionary demography 8:406–407
fertility transition 9:46
impact of 14:460–461
life 14:457–458
model of population dynamics 2:765, 8:384–385, 11:6–7, 18:566–568, 18:581–583
polemic pamphlet 14:457
on population 8:384–385
principle of state assistance 14:460
as rector of Oakwood Chapel 14:457
theory of stagnation 7:39, 7:41
theory of stimulation 14:459
works
Definitions in Political Economy 14:459
Depreciation of Paper Money 14:459
An Essay on the Principle of Population 3:155, 5:676, 8:406
An Investigation of the Cause of the Present High Price of Provisions 14:459
Malthus Reconsidered 14:460
Observations on the Effects of the Corn Laws 14:459
Pamphlets on the Bullion Question 14:459
Principles of Political Economy Considered with a View to Their Practical Application 14:458–459
Maltreatment, protection from 3:478–479
Maltreatment by mentally ill mothers 17:724
MAM *see* Methylazoxymethanol (MAM)
MAMDC1 gene 22:495
Mamillary bodies 14:126
Mamluk Sultanate 12:749
Mammal brain and body size 4:417
Mammalian brain 16:634
and body size 4:417
Mammalian cloning 23:440–441
Mammalian long-term memory systems 5:925f
Mammalian startle response during fear, potentiation of 8:860
Mammals 4:415–416
brain and body size 4:417, 4:415f
Mammeri, Mouloud, works, *La colline oubliée* 21:752–753
Mammillary nuclei 5:923–924
Mammillothalamic tract 5:923–924, 10:889
Mamoulian, Rouben, works, *Applause* 9:168
Man and the Biosphere (MAB) program 6:831–832
Man the Hunter 11:428–429
Managed behavioral health organizations (MBHOs) 14:463
clinician-care managers 14:464
Managed care 10:594, 10:604–605, 10:615–616
ESTs 19:382
health care 15:102
litigation challenging cost 10:595
single-case designs 21:988
in United States 14:462–465
Managed care liability 10:595
Managed competition 10:616–617
Managed health care markets 10:598–599
Managed Professional Business (MPB) 19:104
Managed universities 10:847
Management 6:666, 14:471–478
and administration in organizations 1:103
of air quality 24:910
and business platforms 3:40–41
careers 3:130–131
cultural resource management 5:488–490
cybernetics 5:646–647
decision making in 1:104
and entrepreneurship 7:684
of experimental laboratories 8:558
ideologies 14:475–477
of irrigation 12:725
of knowledge within organizations 12:75
managerial behavior 14:473–474
managerial capitalism 14:471–473
organizational diversity 6:598–599
-led diversity initiatives 6:598–599
ownership and control, separation of 14:471–473
practice, and path dependence 17:389
relevance of technology to 24:114–115
role in organization 14:474–475
roles, in higher education 10:847–848
skills of teachers for open education 17:229
styles 14:475–477
of tasks with chronically-ill patients 23:529–530
value-based 17:348–349
of voluntary agencies 25:264–265
see also Organizational control
Management accounting 14:466–470
and digital technologies 14:467–468
future of 14:469–470
strides in 14:468–469
target costing 14:468–469
activity costing 14:469
cost tables 14:468
functional analysis 14:468
tools 14:468
value engineering collections 14:468–469
Management by objectives 6:665
Management capacity 24:760–761
Management competency 24:520–521
Management education 12:150
Management ethics 3:18–19
Management of dead 9:947
Management philosophy 6:667
Management practice 6:665–667
complexity 6:666–667
entrepreneurship 6:666
ethics 6:666
innovation 6:666
management by objectives 6:665
management means communicating 6:665
managing 6:666
networks 6:666
strengths 6:665–666
systematic abandonment and change 6:665
weaknesses 6:665–666
Management research, and tacit knowledge 24:4–5
Management theory 7:878
distinction between politics and administration 1:104
Managerial and organizational cognition (MOC) 14:479–483
computation *versus* interpretation 14:480
decision making 14:480–481
effectiveness 14:479, 14:481
emotionalizing 14:481
foundational concepts under scrutiny 14:480–481
knowledge application 14:479
mental models 14:480–481
refining methodological assumptions and techniques 14:481–482
rethinking bounded rationality and the role of intuition 14:481
upper echelons theory 14:480
Managerial behavior 14:473–474
Managerial capitalism 13:106, 14:471–473
emergence of 11:856–857
Managerial cultures 14:477
and modern globalization 14:477
Managerial domination 2:752
Managerial ideologies 14:475–477
of Bendix 2:512–513
Managerial re-segregation 8:884–886
Managerial revolution, higher education management 10:846–847
Managerial sciences 24:764–765
Managerial skepticism 13:592
Managerial styles 14:475–477
cross-cultural influence in 14:476
and cultures 14:476
and modern globalization 14:477
Managerial systems 24:105
culturally diverse, implications for 24:105–107
Managerialism 10:846–847, 16:463–464, 20:261, 22:303, 24:795
and ethical values 16:463–464
risk assessment 20:693
Managerialist approaches, authority *versus* power 17:405–407
Managerialization, risk of 13:592
Managers
corporate governance 4:933
female 8:884–885
Managing 15:675
Managing by numbers 16:771
intelligence systems 16:772
ranking systems 16:772
by targets 16:771–772
Managing change, in human resource management 11:348, 11:352

Manalansan, M.F.I., on immigrants in North America 17:24
Manchester Metropolitan University (MMU) 21:68
Manchester Scale 19:290–291
Manchester Statistical Society 1:852
Manchu 2:65–66
Mandal Commission 23:635
Mandamus 1:823–824
Mandarin Chinese 3:695, 16:19
Mandarin Chinese argument nominalizations 23:893t
Mandarin Chinese N-based nominalizations 23:893t
Mandate effect, triggering 8:890
Mandatory arrest, for domestic violence 6:626–627
Mandatory drug testing 6:682
Mandatory minimums 18:880
Mandatory *pro bono* legal services 12:922–923
Mandela, Nelson 25:838
Mandeville, Bernard
  on division of labor 6:602–603
  on self-interest 3:15–16
  works
    The Fable of Bees 4:751, 24:745
Mandrou, Robert 5:420–421
Manetho 1:691
Manga 5:457
Mangle of practice, cybernetics 5:645–646
Manhattan Project 3:280, 11:315
Manhood 9:783
Mania 2:683, 14:484–489, 19:318
  bipolar I 14:485–486
  bipolar II 14:485
  classification 14:485t
  course and prognosis 14:486
  cyclothymia 14:484–485
  diagnosis 14:484–485
  DSM-IV criteria 14:484–485
  epidemiology 14:485–486
  etiology 14:486–487
    biochemical factors 14:486
    cultural factors 14:487
    environmental factors 14:486
    genetic factors 14:486
    neurophysiological factors 14:486
    psychodynamic factors 14:486
  history 14:484, 14:484f
  ICD-10 14:484–485
  non-pharmacological treatment 14:487
  pharmacological interventions 14:487
    acute manic or hypomanic symptoms 14:487
    acute mixed episodes 14:487
    depressive symptoms 14:487
    electroconvulsive treatment in bipolar 14:489
    rapid cycling 14:487–488
Manic depression 15:194
Manic irritability 14:485
Manic–depressive illness, bipolar disorder *see* Bipolar disorder
Manic-melancholic syndrome 2:683
Manie sans délire 17:891
Manifest data 19:436
Manifest variables 20:239
  in personality psychology 8:703
Manifestations
  associated with bereavement 2:532–533
  of conflict 3:716
Manifeste des intellectuels 12:273
Manifesto of the Communist Party 22:834
Manifestos 18:417–420
  dominance principle 18:417–418
  and electoral competition 18:417–418
  ideology mandate 18:417, 18:419
  linking to election change 18:419
  linking to government action 18:419
  media 18:417–418
  nature of 18:417
  overall role and significance of 18:419–420
  party convergence 18:419
  policy 18:417–420
    equilibrium 18:417
  priorities 18:418–419
  quantitative analysis 18:417
  relative emphases 18:418
  using in political research 18:418–419
Maninka (language) 1:301
Manipulation 24:117
  and brokers 2:871
  cognitive 25:708f
  fronto-parietal circuit 25:707–709
  information 25:707
  mental 25:703
  monitoring and 25:705–706
  working memory 25:703
Manipulationist approach of scientific explanation 21:303
Manlinowski, Bronislaw 8:178–179
Man-made disasters 6:481, 17:370
Man-made diseases 15:860
Man-made law, Hurst's contribution 11:439
Mann, Heinrich, works, *Göttinnen* 20:425
Mann, Michael 6:241, 9:672, 14:490–495, 15:492, 25:410
  on capitalism 14:490–495
  on empires 14:491–495
  on globalizations 14:493–495
  on militarism 14:491
  on nationalism 14:492, 14:494
  on nation-states 14:490–495
  on networks 14:490–491
  on power 14:490–495
  on war 25:411–412
  works
    Incoherent Empire 14:490
    Sources of Social Power 11:27
Mann, Thomas
  works
    Buddenbrooks 1:968, 11:74–75
Manner of articulation 18:67–68
Männerbünde 9:646
Mannerist renaissance 20:428–429
Mannfall 9:111–112
Mannheim, Hermann 5:241
Mannheim, Karl 4:686–687, 6:241, 7:378, 7:875, 9:866, 11:94, 11:383, 11:574, 13:111, 13:113, 14:496, 14:635–636, 14:640, 21:261–262, 21:884–885, 21:887, 25:463
  biography of 14:496
  on conservatism 14:497
  contribution to human sciences 11:383–384
  contributions to democracy 14:498
  contributions to sociology of knowledge 14:497
  on elites 7:392–393
  on generations 9:866
  on ideologies 11:573–574, 14:497
  on liberalism 14:497
  on Marxism 6:244
  on mass society 14:745
  objectivity in social science 14:498
  on phenomenology 14:496
  on relativism 14:499
  on social planning 14:498
  on social utopias 14:497
  on socialism 14:497
  on sociology of knowledge 6:243–244
  works
    Competition as a Cultural Phenomenon 14:497
    Conservative Thought 4:686–687
    A Contribution to the Sociology of Knowledge 14:496
    Ideology and Utopia 6:243, 13:113, 14:496
    Man in Society in an Age of Reconstruction 14:498
    The Problem of Generations 14:497
    Science as a Vocation 14:497
    Structures of Thinking 14:496
  and youth movements 25:837–838
Mann–Whitney U statistic 21:948
Manoilescu, Mihail 25:749
MANOVA *see* Multivariate ANOVA (MANOVA)
Manpower systems 23:470
Manson, Charles 3:351, 3:356
Mantegazza, Paolo, works, *The Year 3000: A Dream* 14:256
MantiQitna 21:795
Manual class voting 25:299–300
Manual skills, infant motor development 15:974–975
Manual-based treatment 14:500–504
  cognitive and behavioral therapy (CBT) 14:501
  developmental history of 14:500–501
  dissemination 14:500–501, 14:503
  future prospects 14:503
  pros and cons of 14:501–503
  psychological treatment 14:500–503
  specific treatment manuals 14:501
  stepped care 14:503
  structured treatment 14:502
  system-wise adoption of 14:501
Manufacture
  concentration of labor in 13:158–159
  of consent 16:463
Manufacturers

Manufacturers (*continued*)
new orders for consumer goods/materials 9:317
new orders for non-defense capital goods 9:317
Manufacturing 7:33–34
supply chain management 23:710–713
work anthropology 25:664–665
Manumission 22:40
Manusmriti 23:58
Many-to-many mappings 22:486
MAO *see* Monoamine oxidase (MAO)
Mao Zedong 20:621, 24:449
communes 4:269
cultural revolution 9:866–867
industrialization 11:869, 11:871
role in Chinese revolutions 3:515–516
MAO-A *see* Monoamine oxidase-A (MAO-A)
MAOIs *see* Monoamine oxidase inhibitors (MAOIs)
Maoism 3:662, 4:302–303
and Marxism–Leninism 14:661
Maori
genetic research projects 9:964
people, indigenous rights of 11:799
spirituality (diacritical mark) 3:544
values 11:786
Maori social work 14:505–511
ANZASW Social Work Registration Board 14:507–508
colonization 14:505–506
community 14:506–510
hap&umacr 14:505–506, 14:508–510
Iwi 14:505–506, 14:508–510
professional social work 14:506
Treaty of Waitangi 14:505, 14:507
voluntary work 14:505–506, 14:508
Western worldview 14:505–507
whānau 14:505–506, 14:508–509
worldview 14:507
MAP *see* Maximum a posteriori (MAP) estimation
Map interpolation 23:108
Map of cognitive transformations (MCT) 4:668
for achieving enhanced awareness 4:670f
for appreciating value of thought 4:669f
for creating broad beauty appreciation 4:669f
to developing creativity 4:669f
Map projection 14:285
Map scale 14:285
MapInfo (MapInfo Corporation) 10:58
MAPKs *see* Mitogen-activated protein kinase (MAPKs)
Mapmaking 10:83
Mapping 16:506–507, 23:902
entertainment 7:669–670
scientific controversies 21:284
Mapping model 5:915
Mapping of science 21:325
Maprotiline 1:778
Maps 3:161, 10:174–177, 19:735, 23:126–127, 24:260
and animation 3:165
appropriate projection 24:260
cartography *see* Cartography
on-screen maps 3:161
population distribution 3:161f
principle cartographic strategies 24:263–266
representations of continually varying surfaces 24:266
representations using symbols 24:263–266
in social and behavioral science research 22:617
thematic 24:260
time by visual encodings in 6:715–716
understanding of data 24:261–262
visual or graphic variables in 24:262–263
Maps 2.0 3:619
maps.google.com (website) 10:58
Maquilas 1:623–624
MAR *see* Minorities at Risk (MAR) project; Missing at random (MAR)
Marat, E. 3:320
Marble Arch in London (1825) 3:771
Marburg Archive School 5:702
Marbury, William 12:897–898
Marbury v. Madison 4:700, 12:897–898
Marcé, Louis-Victor 16:686
Marcellinus, Ammianus 11:38
March, James, works, *Cognitive Limits on Rationality* 17:390
Marchall, Benjamin 24:414
works, *Chronological Tables* 24:414
Marchiafava–Bignami syndrome, and alcohol 1:512
Marcion of Sinope 3:548
Marconi, Guglielmo 12:492
Marcos, Ferdinand, works, *Tadhana: The History of the Filipino People* 11:85
Marcus, George E.
on multisited field studies 16:103–108
works
Annual Review of Anthropology 16:103
Ocasião: The Marquis and the Anthropologist, a Collaboration 16:104
Writing Culture 8:590
Marcuse, Herbert 7:379, 9:380–384, 19:331, 20:203
works, *One-dimensional Man* 14:745
Marcuse, Peter 12:913
Mardi Gras 3:145, 3:148
Marett, R.R. 8:289
Marfan's Syndrome 5:505
Marga 19:308
Marginal deterrence 13:574
Marginal effective tax rate (METR) 25:517
Marginal effectiveness 19:118
Marginal excess burden 5:50
Marginal likelihood 14:118, 14:120
example 14:120
Marginal log-linear model 14:325
Marginal man 2:870
Marginal mass 24:818
Marginal models 14:322–327, 23:383
Marginal probabilities, infant and child survival 16:52, 16:54–55
Marginal productivity 25:11
Marginal rate of subjective substitution (MRS) 25:498–499
Marginal rate of technical substitution (MRTS) 25:498–499
Marginal rate of transformation (MRT) 25:498–499
Marginal revolution, economics 7:96
Marginal social cost (MSC) curve 24:631
Marginal structural models 15:78
causal mediation analysis 15:78
Marginal utility 15:333, 25:10–11
mathematical psychology 14:809
Marginalism 19:125, 25:10–11
microeconomics 15:332–334
Marginalist approach 3:92
Marginalist authors 3:91
Marginalist economics
profits in 19:112–114
see also Neoclassical economics
Marginality 22:54
Marginalization 1:71, 1:73, 2:766, 12:960, 20:796–797
and elder care 7:336
in immigration 11:676
and intersectional queer theory 19:765–768
and political ecology 18:350, 18:352
stigma 23:463
and whiteness/white supremacy 25:569–573
and work 8:108–112
and worldview 25:754
Marginalized communities, and community work 4:378–381, 4:383
Marginalized groups 19:683
Marginalized masculinities 14:681
Margins, citizenship 3:639
Margulis, Lynn 9:567
Mariana Islands 17:158, 17:160–161
Marie, Pierre 16:684, 16:684f
Marie Skłodowska-Curie Actions (MSCA) 22:645
Marijuana 5:161, 6:674, 16:170, 21:610
Marijuana use 6:686–688
Marine Le Pen 25:305
Marine sediments 7:693t
Maritain, J. 18:435
Marital discord 15:736, 22:158
influence on socialization of children 22:869
Marital Distress scale 15:567
Marital duration 14:514
Marital examination 8:255
Marital fertility 8:769, 11:6–7
reduction in East Asia 9:56
Marital history, and repartnering 20:431
Marital inequality 11:890
Marital infidelity, honor killing 11:189
Marital quality 17:486–487, 17:490–491
and health 14:512
biological pathways 14:514–515
biomarkers 14:513
depression 14:513–516
divorce 14:512–516
future directions 14:516

gender 14:513, 14:515
life course stage 14:515
marital dissolution 14:515–516
marital duration 14:514
marital satisfaction 14:513–514, 14:516
psychological well-being 14:513
race 14:515
race and ethnicity 14:513–514
social selection 14:515
social support and integration 14:514
subgroup differences 14:513, 14:515
widowhood 14:512, 14:515–516
and religion 20:319
and sleep 22:60
Marital satisfaction, and marital quality 14:513–514, 14:516
Marital stability
and cohabitation 4:120
religion and 20:265
Marital status 10:798
effects on development 14:518–524
and mortality 15:853–855
and pathological gambling 17:593
Marital strategies
human behavioral ecology 11:249
Marital transfers 6:642–645
Maritime anthropology 9:228–229
Maritime networks 14:525–532
British 14:526–528
decline of 14:530–532
containerization 14:531
legal regulations 14:529–530
merchants 14:526–530
organizations 14:529–530
shipbuilders 14:527–530
technology changes 14:529–530
worldwide 14:528
Maritime trading systems 2:231
Mark, N., on cultural tastes 22:335–336
Marked fear 18:37, 18:42
Markedness 12:783, 12:785, 13:267, 13:269, 17:556–558
in linguistics *see* Linguistic markedness
Marker, Chris, works, *Sans Soleil* 6:623
Market activity, and Keynesianism 13:20–22
Market allocation 14:538–543
commodification 14:538–539, 14:541–542
criteria for assessing use of markets 14:539–542
democracy 14:542
efficiency 14:539–540
freedom 14:540–541
human flourishing 14:541–542
nature of goods exchanged 14:542
experimental economics 8:548
intrinsic goods 14:542
liberty 14:542
motivation 14:541–542
regulation *versus* alternatives 14:542–543
sexual relations 14:542
taxation 14:542–543
Market Bolshevism 18:692
Market capitalism 3:105–108
Market creation, issues for renewable energy 18:368–370
Market culture 5:536–537
Market demand and supply 14:622, 15:23–25
Market design, experimental economics 8:549
Market devices, economic sociology 7:68–69
Market discipline 22:590–591, 22:594
Market economics, and Keynesianism 13:21
Market economy 20:845, 24:894, 24:898–899
and capitalism, distinction between 3:107
economic transformation 7:73–75, 7:77–79
gift against 10:141–142
and welfare state 25:476–477
Market equilibrium 7:12–13
microeconomics 15:334–336
Market failure justification 19:538
Market failures 14:539–540, 18:274–275, 22:21
and new liberalism 18:299–300
regulatory agencies 20:181, 20:183
Market for corporate control 23:491
Market forces, and behavioral economics 2:437
Market fundamentalism 5:536–537
Market geographies 14:544–548
allocation 14:546
behavioral viewpoint 14:547
business geographics 14:547
distance 14:545–546
economic underpinnings 14:545–546
law of retail gravitation 14:547
location-based services 14:547–548
market area as territorial expression 14:544–545
prescriptive market areas 14:546
proximal area 14:545
public services 14:544
retail trade 14:544–545, 14:547
spatial analysis 14:546
spatial interaction 14:547
Thiessen polygons 14:545
time geography 14:547
transportation problem 14:546
Market information regimes 12:590–591
Market integration 8:60
Market integration hypothesis, and prosocial behavior 19:240
Market journalism 12:851, 12:853–854
Market liberalization 12:58
Market mechanism, microeconomics 15:331–332, 15:335
Market opportunity, and entrepreneurship 7:682–684
Market performance 14:549–554
Market perspective, on diffusion 6:408–409
Market portfolio 12:211
Market power 6:536–537
and antitrust 1:797–798, 1:800, 1:802
Market Power school 14:555
Market price, and industrial societies 11:865
Market process, economic thought 1:575–576
Market regulation, and radio 19:880–881
Market research
opinion polls 20:878
surveys 23:747–751
Market science, lifestyle in 14:115
Market size, journals 12:866
market socialism 22:834
Market society, and libertarianism 14:40–41
Market state terrorism 3:678
Market structure 14:549–554
cultural production 5:608–609
definition 14:549–550
game-theoretic approach to 14:550–551
and innovation 12:160, 12:165–166
Market structure regulation 20:169–170
Market trading, experimental economics games 8:549
Market triumphalism 20:896
Market work 9:757, 9:762
determinants of tastes for 9:757–758
Marketable permits 7:727
Market-based funding 24:760
Market-based incentives 7:773
Market-cum-technological environment 3:23
Marketing 1:479, 3:23–24, 4:139–140, 4:142–143
anthropology of 3:8–9
and audience measurement 2:211
class actions 3:736
consumer psychology 4:738–739, 4:741
and customer satisfaction 5:630–632
history 3:24
neuromarketing 16:698–702
pharmaceutical industry 18:9–10
university press 19:627
Marketing and advertising practices
ARPP system of professional regulation 1:206
British co-regulation 1:206–207
current issues 1:205
French co-regulation 1:205–206
laws and codes 1:205
technology and 1:204–205
Marketing and leasing 20:17
Marketing character 9:460
Marketing ethics 3:19
Marketing research, and semantic differential 21:563, 21:568, 21:570
Marketing strategies 3:664–666, 14:555–558
business city 3:664–665
in circular model 14:556–557, 14:556f
in game of imitation 14:557
in game of movement 14:557
in game of position 14:556–557, 14:557f
competitive advantage 14:555–557
competitive strategies 14:555
cosmopolitan city 3:665
cultural city 3:665–666
dual strategy 14:556

Marketing strategies (*continued*)
in external/internal perspective 14:555–556
final frontier of value 14:557
global city 3:664
good city 3:665
green city 3:665
hypercompetition 14:556–557
Market Power school 14:555
in new millennium 14:557–558
postindustrial city 3:664
in resource-based view perspective 14:556
win–lose scenario 14:556
Marketing studies 9:274
Marketization 5:538, 20:817, 24:826
of art museums 14:566
of social services 10:258–259
Marketization process 14:736
Marketized education 22:328–329
Markets 1:479–480, 7:14–15, 7:33–37, 7:926, 10:841–844, 22:980, 24:765
in archaic states 23:373
artistic/cultural 14:564–568
artists 2:56–57
autonomous adaptations 10:842
and brokers 2:872–873
causes of underinvestment in 22:21
and community sociology 4:364–365
corporate governance 4:935–936
culture in 5:535–536
data on participation rates and trends within and between countries 22:21–22
definition 14:538–539
development 10:150–152
in dual actor failure 10:179
economic exchanges 14:559
for education and training 22:20–21
embeddedness of 14:559–560
equilibrating relationships 14:560–561
finance 9:171–175
financial 9:191–193
foundations 10:841–842
operationalization 10:842–843
global cultural supermarket 14:562
governance structures 14:560
health care 10:598–602
health insurance 10:640–643
housing 11:241–243
Hurst's contribution 11:439
and informal economy 12:58
job displacement insurance 12:825–826
demand 12:825–826
supply 12:826
LGBT *see* Tourism; gay and lesbian
and liberalism 14:26–30
liberalization of 13:159
and money 15:746
moral limits of 20:896
and nation-state, relation between 22:121–122
operationalization 10:842–843
asset specificity 10:842
bounded rationality 10:842–843
contracting 10:841, 10:843
economizing 10:843–844
empirical testing 10:843
full formalization 10:843
governance 10:841–844
human actors 10:842
interdisciplinarity 10:842
opportunism 10:842–843
organization 10:842–843
predictions 10:843
refutable implications 10:843
strategic behavior 10:842
transaction cost economics 10:843–844
transactions 10:842
and organization theory 19:893–894
peasant farms 17:664
Polanyi on 18:244–245
provisioning relationships 14:561
regional analysis 14:561
Smith's invisible hand 14:24
and social networks 5:469
sociological model of 22:980
sociology of 7:67–68
conventions 7:69
transaction costs 14:560
Market-seeking foreign direct investment 9:325
Markham, Annette 12:607
Markov chain Monte Carlo (MCMC) method 2:375, 2:381, 8:373, 8:695, 10:839, 14:569–572, 15:763–765, 16:48, 19:439, 23:379–380, 23:382, 23:409
in agent-based models 1:370
asymptotic distribution 14:569
Bayesian statistics 14:569
convergence 14:571
ergodic averages 14:569, 14:571
in event–history analysis 14:352–353
Gibbs sampler 14:569–570, 14:570f
Gibbs sampling algorithm 15:764–765
limitations 14:571
for measuring judicial behavior 12:886
Metropolis–Hastings algorithm 14:570, 15:763–764, 14:570f
multiple-block M–H algorithm 15:764
parameter space partitioning 4:478
for parametric item response theory models 8:700
simulator 15:760
for social analysis 14:578–580
use in Bayesian statistics 2:381
inference 2:375–376
model selection 4:476
Markov chains 16:113
Markov decision processes (MDPs) 12:103, 14:573–575
algorithms for solving 14:574
animal-behavior example 14:573
definitions 14:573–574
related models and extensions 14:574–575
stochastic optimization of 23:478
Markov equivalence of regression graphs 10:345–346
Markov games 14:574
Markov learning models 6:524–525
Markov models 8:301, 14:323–324, 14:326, 23:469–470
choice of categories in 23:470
deterministic equivalent of 23:470–471
discrete time and continuous time systems 23:470
discrete time closed model 23:469
open Markov systems with fixed input 23:469–470
open Markov systems with fixed size 23:470
practical issues with 23:470–471
and social analysis 14:576–583
brand loyalty 14:577, 14:577t
compartmental models 14:580–581
continuous time models 14:580–582
discrete time models 14:578–580
empirical Bayes estimation of P 14:579–580
finite state Markov Chains 14:580
malaria 14:578, 14:578t
Markov chains 14:578–580
Markov processes 14:580–582
maximum likelihood 14:579
occupational mobility 14:577–578, 14:577t
stationary Markov chains 14:578–579
statistical inference-classical results 14:579
transition probabilities 14:578t
voter patterns 14:578, 14:578t
Markov processes 8:296, 14:801
event history analysis 8:313–315
constant and piecewise constant transition intensities 8:314
freely varying (nonparametric) transition intensities 8:314–315
other parametric models for transition intensities 8:314
semi- and nonparametric regression Markov models 8:315
in learning spaces
see also Learning spaces
for social analysis 14:580–582
Markov property 14:324
Markov random graph 22:470
Markov regime-switching model 5:652
Markov switching models 24:338
Markov time 14:632
Markov transition probability 12:102–103
Markov-perfect Nash equilibrium 17:255
Markov-transition process, continuous-time 4:420–421
evolution of political complexity 4:420–421
genotypic sex determination and adaptive radiations in the sea 4:421
Markwell, K., on gay/lesbian tourism 9:641
Maronite-Christians 2:241
Maroons
African origins 14:592–593
in anthropology 14:591–595
communities 14:591–592
in Esmeraldas 14:593
Jamaica Maroons 14:593
and nation-states 14:592

Ndyuka 14:593–594
Palenque de San Basilio 14:593
Palmares 14:593
retentions 14:593
Saamaka 14:593–594
San Lorenzo de los Negros 14:593
Marquesas 17:158–159
Marr, David 14:596–598
biological information processing 14:597
brain theory 14:596–597
cognitive psychology 14:597–598
computational theory 14:597–598
legacy 14:597–598
MIT period 14:597
neuroscience 14:596–598
vision 14:596, 14:598
works
Vietnamese Anti-Colonialism 1885–1925 11:85
Marres, Noortje, on public engagement 5:525–526
Marriage 2:65, 2:68–69, 6:116–119, 8:766–767, 9:73–74, 9:76, 11:5–6, 14:599–610, 21:792–794, 21:816–817
alliance 14:600
alliances 14:600
arrangement 14:602
changes in 14:605–610
cohabitation 4:118–121, 14:606
consanguinity 14:600
consent 14:601, 14:603
and creative potential of divorce 14:602–603
cultural institution of 13:54–55
decline of marriage 8:767–768
definition 14:599–600, 14:611–612
and desistance/persistence 5:142
in East Asia 9:56–57
economic dependency 14:602–603
economics of 8:829–834
children 8:829–830, 8:832–833
efficiency 8:830, 8:833
equilibrium 8:831–833
family decision-making 8:830
frictions 8:832
gains 8:829–830
market 8:831–832
matching 8:831–833
policy issues 8:833
power of partners in 8:830, 8:833
searches 8:832–833
traits 8:830–833
transfers 8:829–833
wages 8:829–830, 8:832–833
work 8:829–830, 8:833
ethnic intermarriage 8:118–122
event history analysis for 8:309
and exchange 14:599–601
and family structure 2:471
family theory 8:815–820
and feminist legal theory 8:922–923
and fertility, in South Asia 9:80–82
fertility of single and cohabiting women 9:13
and fertility transition, in China 9:46–48
and gender 14:599–601, 14:603
and genealogy 9:807–808, 9:810
gifts 14:601
heterosexual, Chinese family obligations and imperative of 21:734–735
impact of conditional cash transfers on 4:555–559
increase in age at 8:767
intermarriage 14:600
and intimacy 14:599, 14:601–603
kinship 13:53–55, 13:58
and kinship 13:37–38, 14:599–600
in Kutali (Sri lanka) 14:603
level 6:116–119
and love 14:603
as marker of class 8:780–783
measures 6:117t
network analysis 14:601–602
nonheterosexual British Asians 16:865
onset, in Latin America and Caribbean 9:70t
within panmictic circles 18:589
parenthood 14:599–600
and personality 17:797–800
cumulative continuity 17:798–799
interactional continuity 17:799
similarity and complementarity 17:798
traits 17:797–798
population register for 5:736
potlatch 18:715–716
recurrent demographic events 5:740
and religion 20:318
roles and change 14:601–602
same-sex 9:641
same-sex, in South Africa 21:746
and Second Demographic Transition 21:346–347
and sexuality 14:599–600, 14:603, 21:820–821
and social mobility 22:428–429
in South Asia 2:79
in Southeast Asia 9:87–89
and status 22:333–334
and structuralism 23:628–629
synthetic or hypothetical cohort measures 6:118–119
timing 6:119
and transition to adulthood 24:580
trends in Western societies 17:545
see also Divorce
Marriage contracts 6:642
liberalism and 14:23
Marriage legislation 6:644
LGBTQ families 21:756–757
Marriage markets 18:589
assortative mating in 2:100
characteristics of 8:120–121
and childbearing 8:824–825
as value of erotic capital 8:5
Marriage patterns, tribes 24:645–646
Marriage payments
family characteristics as determinants of 6:643
rise in 6:644
social characteristics and occurrence of 6:642–643
and women welfare 6:644–645
Marriage rate 18:588
Marriage register 7:165
Marriage squeeze index 18:591
Marriage squeezes 10:821–822, 15:612, 18:590–591
and sex selective abortion 21:688
Marriage systems 14:611–619
classification 14:611–613
distribution 14:613f
evolutionary perspectives 14:612
family 14:611, 14:615, 14:618
kinship 14:611, 14:616
mating and reproduction 14:611–612
monogamous marriage
as monogamous inheritance 14:617–618
as socially imposed monogamy 14:616–617
polyandrous marriage as response to ecological challenges 14:618
polygyny
as default mating/marriage system 14:614–615
polygyny threshold model, female choice, and male coercion 14:615–616
resource defense 14:615
subject definition 14:611
Married men 14:608
Marroquin illusion 25:161
MARS *see* Multivariate adaptive regression splines (MARS) model
Marschak, Jacob 5:879, 5:881–882
Marsden, George 20:398
Marsden, M. 3:317
Marsden, William 7:166
works, *History of Sumatra* 11:84
Marsh, Benjamin C. 1:219
Marsh, C., on interviewing 12:671
Marsh, J.P., works, *Man and Nature* 7:750
Marshall, Alfred 4:751–752, 13:14, 14:620–625, 16:228, 23:207
on administrative socialism 14:623–624
biography of 14:620–621
on consumers' demand 14:622
on consumer's surplus notion 14:622
on division of labor 6:604
on economic freedom and free trade analysis 14:623–624
economic studies of 14:622
on economic teaching 14:620–621
on economic thoughts 14:624
on economics 14:621–622
law of substitution of 14:623
on market demand and supply 14:622
on monetary arrangements 14:624
on policy views 14:623–624
on scale economies under competition 14:622–623
theoretical contributions 14:622–623
theory of value and distribution 14:622
works

Marshall, Alfred (*continued*)
The Economics of Industry 14:620
Elements of Economics of Industry 14:621
Industry and Trade 14:621
Money Credit and Commerce 14:621
The Pure Theory of Domestic Values 14:620
The Pure Theory of Foreign Trade 14:620–621
Marshall, John 16:685
Marshall, Leon C. 13:773
Marshall, P.J. 11:685
Marshall, Thomas Humphrey 3:647–648, 5:663, 12:245, 14:626–629, 25:529
as academic and citizen 14:626
on civil rights 14:626–628
on current state of citizenship 14:627–629
on human rights 14:628–629
on inequalities 14:626–627, 14:629
on political rights 14:627
on social rights 14:626–627
on sociology of citizenship 14:626–627
on sociology of rights 14:628
Marshall Field Foundation 22:656
Marshall Islands 17:158, 17:160–161, 17:163
Marshallian dynamics 15:335
Marshallian externalities 23:207
Marsh/Shavelson revised academic self-concept model 12:426f
Marti, Ignasi, on institutional work 17:394
Martial arts 4:612–613
Martial races 4:263
Martin, Henry, on division of labor 6:602–603
Martin, Robert F. 16:228
Martin, Terry 6:812
works, *Affirmative Action Empire* 6:812
Martineau, Harriet 1:540, 22:697
Martinez, Ramiro 11:629
Martinez-Alier, Joan, on agrarian ecology 1:461
Martingales 14:630–634
applications 14:632–633
definition 14:630–631
etymology 14:633–634
examples 14:631
historical notes 14:633
properties 14:631–632
Martin-Löf randomness 1:535
Martin–Quinn Score 12:886
Martyr, Justin 20:285
Martyrdom of Polycarp 3:548
Martyrs 3:233
Jewish 3:546
Maccabean 3:546
Maruyama Masao 14:635–641
after 1945 14:637–638
Cold War theology 14:638
on democracy 14:635–640
on emperor system 14:637, 14:639
on intellectual history 14:635–636
on Japanese thought 14:639–640
later works of 14:640
Mannheim on 14:635–636, 14:640
on modernism 14:635, 14:637, 14:639–640
on modernity 14:636–637, 14:639–640
on nationalism 14:635, 14:637, 14:640
on neo-Confucianism 14:636
Ogyu Sorai on (Confucian philosopher) 14:636
on political thought 14:635, 14:638
on self-possession 14:636
tennosei 14:637, 14:639
through 1945 14:635–637
Tokugawa era 14:635–636
on transfer of oppression 14:637
U.S.-Japan Security Treaty (ANPO) 14:638–639
Marx, Groucho 8:663
Marx, Karl Heinrich 1:456, 1:540, 1:636, 1:665, 1:680–681, 1:963, 3:90, 3:515, 3:710, 4:158–159, 4:339–340, 4:751–752, 5:244, 6:241, 6:372–373, 7:650, 9:112–113, 9:115, 10:218, 10:739–740, 11:103–104, 11:368, 11:869, 13:70, 13:112, 13:511, 13:551, 13:802, 14:642–648, 14:656–657, 15:13, 15:711, 16:223, 17:22, 18:382, 18:399, 18:472, 19:153, 19:168–169, 19:192, 19:330–331, 19:703–704, 20:359, 20:660, 20:682, 21:167, 21:261, 22:119, 22:583, 22:697, 22:834, 22:837–839, 22:844, 23:1–2, 23:727, 24:745, 25:8–9, 25:670
on alienation 1:551–555
on American/French Revolutions 1:598–599
applied social research 1:851–852
on apprenticeship 1:865–866
approach to history 11:73–74
biography of 2:543
on capitalism 3:105–107
on class struggles 14:642–643
on communism 14:642, 14:644–647
concept of artisans 2:47
conflict theories 22:956–957
contributions to historical sociology 11:27
on division of labor 6:601, 6:603, 13:162
on economic determinism 6:242
on economy and culture 5:535
and Engels 14:644–646
on fetishism 9:107
on feudalism 1:966–967
on French Revolution 9:409–411
on idealization 11:517–520
on identity 11:523–525
on ideologies 11:573–574
on individual-collective relationship 11:818
on industrialization 11:869
on industrialized societies 11:864–866
on irrigation 12:724
on labor 5:537
law of development of human history 14:642–643
on Marxism 14:656–657
mode of production 14:642–643
on modernization theories 15:700–701
on natural science 14:644
notion of labor 1:552–553
as orthodoxy 14:646–647
on peasantry 1:458
on philosophy 14:643
and political economy 14:642, 14:644–647
on political sociology 18:472–473
on practice theory 18:783–784
on praxis 18:784
on profits 19:111–112
on reification 5:244–245, 20:203–205
as revisionisms 14:646–647
on revolutions 20:621–623
on science 14:643
on social class 3:756
on social inequality 22:338–342
on social science 14:644
on social structure 22:721
on sociology 22:996–997
on technological determinism 24:110
on work/labor 25:675–676
works
Capital 3:90–91, 14:644, 14:647, 14:650, 14:677, 25:8
The Class Struggles in France 14:647
The Communist Manifesto 14:642–643, 23:353
A Contribution to the Critique of Political Economy 14:645
Economic and Philosophical Manuscripts 14:646
The Economic and Philosophical Manuscripts 5:498
The Eighteenth Brumaire of Louis Bonaparte 7:321, 14:647
Formen 14:670
German Ideology 5:498, 14:647
Gesamtausgabe 14:646
The Grundrisse 14:647, 14:650
Marx–Engels Gesamtausgabe 14:647
Theories of Surplus Value 7:590, 14:647
Vaterlandslosigkeit 24:589
Zur Kritik der politischen Ökonomie 13:802
Marxian historical dialectics 14:202
Marxian theories 21:809–810
Marxism 2:255, 2:716, 3:125, 4:380, 5:76, 5:931, 9:369, 9:372, 9:540–541, 9:769, 10:287–290, 11:103–104, 11:576–577, 11:688, 14:642, 14:645, 16:263, 16:459, 17:425–426, 18:382–383, 18:671–672, 20:260–261, 20:820–821, 20:823, 20:843, 20:851, 20:853, 21:143–144, 22:696–697, 22:833–834, 22:837–839, 22:841, 22:984–985, 22:994, 24:795, 25:99, 25:405
and advertising 1:215–216
analysis 12:913–914, 18:766
analytical 1:665–667
and anarchism 1:680–681, 1:683–684
Anthony on 10:136
approach 3:229, 13:163

approaches to social work 14:663–669
Britain 14:663–667
under capitalism 14:664–667
China 14:667–668
class struggle 14:663–664
Cuba 14:664, 14:667–668
dialectical materialism 14:663
education 14:664, 14:666–668
health care 14:667–668
human services 14:665–666
politics of social work 14:664
popular social work 14:667–668
radical social work 14:663–668
under socialism 14:667–668
state 14:664
understanding of society 14:663–664
United States 14:663–666
welfare 14:665–668
where state breaks down 14:667
Bielefeld School 2:553–554, 2:556
capital in 13:107
class consciousness 3:739–742
class in 3:745–747, 3:747t
and conflict 4:589
in contemporary sociology 14:649–655
contradictory nature of capitalism 14:652–654
cultural critique and 5:384–385
and delinquency 6:72
and development geography 6:315
on education 7:147
Elster on 7:402–403
end of 5:300–301
and Engels 14:645–646
and feminism 8:899–901
film and video industry analysis 9:158
Frankfurt School 9:380–381
French Revolution 9:409–411
Fromm on 9:458–460, 9:462
and hegemony 10:747–749
historical materialism 14:651–652
on knowledge society 13:108
labor theory 10:289
neo-Marxism 16:292–293
parties 18:404–406, 18:408
and political economy 18:362–365
as political movement 14:643
problematic reproduction of capitalism 14:652–654
production mode 15:714–716
on property 19:153
and prosopography 19:256–257
and radical social work 19:870–871
research program 14:650–654
ontological presupposition 14:650
social construction of reality 14:650–651
and revisionisms 14:646–647
and social history 22:308–311
as social science 14:643, 14:645–647
into sociology 14:649–650
studies of craft production 5:109
theory of imperialism 11:686
use within sociology 14:650
view of state formation 23:354
view on political institutions 12:182
working classes 25:696–697
world systems theory 25:748–751
and worldview 25:754–756
writing of history 11:40–41
critique of capitalism 11:40–41
dogmatic form of Marxist orthodoxy 11:40
economic analysis 11:41
Marxism–Leninism 14:656–662
Cold War 14:656, 14:660–661
Comintern 14:656, 14:659–660
ideas of 4:302
and international communism 14:660–661
Last Testament 14:659
and Lenin 14:657–658
Mao Zedong thought and 14:661
and Marx 14:656–657
national liberation movements 14:656, 14:660
as orthodoxy 14:659–660
Russian revolution 14:657–658, 14:660–661
of Soviet Union 11:76
and Stalin 14:658–659
Trotskyism 14:659–660
Marxist analysts 4:379
Marxist and socialist tradition 13:199
Marxist archaeology 14:670–673
contributions of Childe 14:670–671
southern 14:671–672
in Soviet Union 14:670
western 14:671
Marxist discourse 18:766
Marxist economics 18:360, 20:93
Marxist feminism 9:738
Marxist geography 14:674–678, 22:304
geographical materialism 14:674–675
historical materialism 14:674–675
history of 14:674
mode of production 14:675–676
politics of nature 14:677–678
politics of space 14:677–678
social relations with nature 14:675
space 14:675–676
spatial relations 14:676–677
Marxist historiography, cultural turn on 11:42–43
Marxist ideology, of communism 4:303
Marxist idiom 22:888–889
Marxist legal theory 13:551
Marxist model, feudalism 9:112–113
Marxist orthodoxy 22:834
Marxist science of the child 20:826
Marxist social thought 14:497
Marxist society 8:466
Marxist theories 3:94–95, 11:12, 12:913
of ideology 19:603
and property rights 19:168–169
of public opinion 19:565–566
Marxist-derived idea of law framing consciousness 13:551
Marxist-inspired cultural studies 14:694
Marxist-Leninist ideology 5:372
Marxist-Leninist psychology 5:262–264
Marxist-Leninist state 25:95
Marxist-Leninist traditions of journalism 4:279–280
Marxists 5:294, 11:237, 18:202–203, 20:718, 24:819
Mary, Immaculate Conception of 3:234–235
Maryanski, Alexandra, works, *On the Origin of Societies by Natural Selection* 20:386
Maryknoll Fathers and Brothers 15:614–615
MAs *see* Metropolitan areas (MAs)
Masanao, Nakamura 24:584
Masaniello Revolt 15:11–12
Masaryk, Thomas 21:169
MASC *see* Multidimensional Anxiety Scale for Children (MASC)
Maschler–Owen solution 4:878
Maschler–Owen value for hyperplane games 4:878
Masco, J., on democracy in United States 17:24
Masculine chauvinism 25:816–817
Masculine cultures 9:724
Masculinities 8:915, 9:690–691, 9:725, 9:733–734, 9:770, 11:344, 13:256, 14:685–689, 20:798, 21:678–680, 21:684
across life course 14:686–687
class-based 14:687–688
colonial constructions in Africa 10:813–814
and education 7:118
and fathering 8:847
and feminist criminology 8:913–915
and gender archaeology 9:764, 9:766
homosexuality and 9:656
and leisure 13:854
racialized 14:688
scripting theory 14:685–686
critiques 14:686
levels of 14:685–686
sexual script 14:687
and sexuality 21:819, 21:822
and social work practice 14:679–684
antioppressive practice 14:682
creating change at community and societal levels 14:683
diversifying men and masculinities 14:681–682
gender relations 14:681–684
group work with men 14:683
hegemonic masculinity 14:679–681, 14:684
individual work with men 14:683
intersectionality 14:682
issues facing men 14:680
men in women's profession 14:679–680
privilege 14:680–684
profeminist sensibility among men in social work 14:682
theorizing men and masculinities 14:680–681
working on one's own social location 14:682–683
and technology 9:725

Masculinity–femininity dimension 17:824
MASH *see* Mutual aid self-help (MASH) groups
Masking
in change blindness 3:345–346
statistical disclosure limitation 4:578–580
Masking effect, in outliers 23:389
Maskovsky, J., on capitalism in North America 17:24
Maskrosbarn 8:425
Masks, secrecy 21:374–375, 21:377
Maslach Burnout Inventory (MBI) 2:929–930
Maslow, Abraham H., works, *Motivation and Personality* 15:936
Masochism 13:888, 17:952
Mason, Otis 2:71
Masquerade, carnival 3:145–149
Mass (Catholic) 20:861
Mass accidents, class actions 3:736
Mass action 13:392
Mass action interpretation 9:584
Mass communications 6:680, 14:726–727
advertising 1:211–212
normative frameworks 14:690–695, 14:690t
communitarianism 14:694
degradation of art 14:691–692
distrust of 14:690–692
entertainment 14:693–694
fabrications and corruption of youth 14:691
faith in openness 14:692–694
iconoclasm 14:690–691
image ethics 14:690–691
media ethics 14:694
nineteenth-century criticisms and alternatives 14:692–693
other criticisms 14:693–694
public sphere 14:693–694
republicanism 14:691–692
responsibility to distant suffering 14:691
twentieth-century developments 14:693
political protest 18:422
Mass consumption society 4:759
Mass culture 3:274, 5:606, 20:396
Bell on 2:503
celebrities and 3:274–275
capitalist ideology 3:274–275
versus heroism 3:275
Mass data, ESR 7:558–560, 7:563–564
Mass entertainment industry of nineteenth century 7:671–672
Mass fatality incidents (MFIs) 9:336
Mass functions, probability 19:34
Mass graves, and forensic archaeology 9:336–337
Mass higher education 24:758–759
in social sciences 6:489
Mass hysteria 11:508
Mass idols 3:274–275
Mass imprisonment, and collateral consequences 14:696–701
family consequences 14:699–700
individual consequences 14:696–699
Mass incarceration 5:267
and punishment 19:656
Mass killings
genocide 10:15
psychology of 14:702
autocratic *versus* democratic societies 14:704
cultural characteristics 14:704
cultural devaluation 14:704
difficult life conditions 14:702
evolution of destructiveness 14:703–704
future directions 14:705
group conflict 14:702–703
halting persecution and violence 14:705
instigators of collective violence 14:702–703
monolithic *versus* pluralistic societies 14:704
origins of violence between groups 14:703t
overly strong respect for authority in society 14:704
past victimization or suffering 14:704
prevention and reconciliation 14:705
role of bystanders 14:704
role of leaders 14:704–705
self-interest 14:703
turning against other 14:703
violent revolution 14:703
war 14:703
Mass literacy campaigns 11:590–592
Mass magical movements 15:532
Mass media 4:291, 4:292, 5:704, 14:707, 14:726, 15:42, 18:528–529, 19:605–607, 22:583–587, 25:100
and advertising 1:217
anthropological perspectives 18:530
and capitalism 18:529
and computer games 4:498–500
and critical tradition 14:729–730
critique of ideology 14:729
political economy 14:729
public sphere theory 14:730
and cultural identity 14:707–711
audiovisual culture 14:707–708
and cultural imperialism 14:709–710
and national identities 14:707
culturalist approach 14:730–731
active subjects 14:730–731
centrality of meaning 14:730
cultural studies 14:730
field 14:726–727
basic orientations 14:726–727
nomenclature 14:726
organizing logics 14:727
tensions in 14:728–729
journalism 12:857–861
and knowledge about climate change 3:858
perspectives 14:727–729
agenda-setting research 14:728
content analysis 14:728
effects research 14:728
intervening variables 14:728
origins 14:727–728
policy horizon 14:731
media systems and norms 14:731
regulation and deregulation 14:731
for political advertising 18:331–334
political economy of 14:733–737
and politics 14:712–718
agenda-setting 14:714
anticipated agreement hypothesis 14:713–714
audience-centered effects 14:713–714
behavioral effects and institutions 14:712
broadcast communications 14:713, 14:715–717
changes in mass media ownership 14:715–716
flow of information 14:712–713
fragmentation 14:713, 14:716–717
framing 14:715
impacts of technology: television and cable news 14:716
Internet 14:716–717
issue public hypothesis 14:713–714
media-centered political behavior 14:714–715
minimal effects 14:713
news media bias 14:714–715
one-step flow 14:712–713
polarization 14:717
priming 14:715
professionalization of communications 14:716–717
propaganda 14:712, 14:716
public opinion 14:712–714
selective exposure 14:713–714, 14:717
spiral of silence theory 14:714
text analytics 14:715
theoretical origins 14:712–713
two-step flow 14:712–713
uses and gratification theory 14:713–714
and polling 18:459, 18:461, 18:463
publishing 19:621–622
for reducing prejudice in society 18:843–844
representations in 14:738–742
elite impacts 14:741–742
enhancement 14:740–741
entertaining 14:739
framing 14:739–740
and political and policy disputes 14:741
racial equality 14:740
status reinforcement 14:740–741
role in moral panics 15:801
school of thought 14:726–732
as sources for climate change information 3:858
and sports 14:719–725
Mass media effects on public evaluation 13:748
Mass media influence 22:345–346
Mass media sectors 14:735
marketizing of 14:736
Mass migration 11:648, 15:461
modern, features of 11:649–650
Mass mobilization 5:369
Mass pauperism, and social question 22:553
Mass politics 14:743–745

and democracy 6:83
structuring of 20:745–746
Mass production 24:104–105, 25:670
Mass production society 20:919
Mass protest 5:364–365
Mass public, issue evolution 12:766–767
Mass quantifiers 19:707
Mass schooling, and wealth flows 9:35–36
Mass secular schooling 7:291
Mass society 3:274, 14:743–746, 20:671
accommodation and persistence of concept 14:745–746
Cicero on 14:744
classical origins of concept 14:743–745
ancient roots 14:743–744
classical underpinnings from Hobbes to Tocqueville 14:744
conservatism and fear of the masses 14:744–745
and collectives 11:818
Greek 14:743–744
mass politics 14:743–745
and photography 18:84–85
Plato on 14:743
rise of modern theory of concept 14:745
Ortega and Mannheim 14:745
predominance and expansion 14:745
Roman 14:743
twentieth century origin of concept 14:743
Mass supervision 19:659
Mass supervision, and punishment 19:659
Mass trauma 14:747–751
children's psychological well-being and functioning in aftermath 14:748
depression 14:747–749
as global phenomenon 14:747
in adults 14:747–748
in children and adolescents 14:748
grief 14:749–750
group psychotherapy 14:749
individual-level treatment 14:749–750
interventions 14:748–749
narrative exposure therapy 14:749–750
natural disaster 14:747–748
post-traumatic stress disorder 14:747–750
resilience 14:748–749
school-based programs 14:749
torture 14:749–750
war 14:747–750
Mass violence 25:119
Massachusetts Institute of Technology (MIT) 24:764
Growth and Development Study 14:163–164, 14:166–167
MIT/Michigan tradition of small group research 22:102
OpenCourseWare 6:896
Massachusetts Peace Society 17:643–644
Massacres 10:2
Massad, Joseph A.
on Orientalism 21:751
translations 21:752
works
Desiring Arabs 21:751
Massell, G. 3:319–320
Masses, needs of 22:793
Massive category killing (MCK) 17:622–623
Massive modularity thesis 8:401–403
Massive multiplayer online games (MMOG) 4:494
Massive open online courses (MOOCs) 4:330, 6:567, 6:569, 10:251, 12:566, 14:96
Mass-mediated terrorism 24:227–228
Massumi, Brian 1:238
Master artisan 2:47
Master builder 24:864
Master degrees in social work and public health (MSW-MPA) 10:722
Master identity 16:216
Master narrative 11:115
Master of Social Work (MSW) Program, in China 22:790–791, 22:793
Master plan, community-based health interventions 10:650
Master Plan of Residential Land Use 20:421
Master Settlement Agreement 14:141–142
Master status 10:803, 10:807–808
Master–apprentice relation 1:866–867
Master–novice relations 1:866–867
Master's degree, Bologna Process 2:742–745
Mastery
control beliefs 4:819–820
coping 4:889–891
definition of 5:335t
experiences 22:540
Mastery goals
social-cognitive perspectives 13:671
Mastery learning 14:752
Bloom's learning for mastery 14:753–754
Carroll's model for school learning 14:752–753
elements of 14:755–756
correctives 14:755
enrichment 14:755
feedback 14:755
instructional alignment 14:755–756
implications 14:757
misinterpretations of 14:756–757
research results 14:757
Mastery motivation 24:188
Mastery orientation 13:643
Mastery-goal orientation, and epistemological beliefs 7:889–890
Masturbation 21:690
prevalence of 21:840–841, 21:843
Masturbatory behaviors in boys 21:840–841
Matamoros, Santiago de 23:46
Matched pair designs 8:538
Matched sampling 3:248
Matching 12:454, 17:108–109
algorithms 17:108–109
developmental psychology 5:356–357
goals 17:108
marriage 8:831–833
nonequivalent group designs 16:852
propensity score 17:108
sampling 17:108–109
selecting from reservoir of potential controls 17:108
Matching algorithms 17:108–109
Matching designs 19:745
Matching Familiar Figures Test 4:111
Matching hypothesis 4:111, 4:114, 22:366
Matching pennies (game theory example) 9:573, 9:584
Matching-to-sample 21:950
Mate choice *see* Human mate choice; evolution of
Mate preferences 11:341–342, 11:345
Material abuse, of elders 7:330
Material agency 13:206, 13:209
Material carriers 10:115–117
Material civilization 8:210
Material conditional 5:95
Material content, and personal identity 17:745–746
Material context of scientific de-Westernization 25:543–544
Material contradictions 14:649–653
Material culture 4:608–610, 4:612, 4:751–752, 21:184
African Studies 1:276, 1:278
and arts 2:18
craft production
and folklore 5:109
studies, rebirth of 5:109–110
of experiment 21:226
and gender archaeology 9:764–765
Material culture study, Ancient Egypt 1:693–695
Material deprivation
reference group 20:73–74
and religiosity of women 25:635
Material economy 13:109
Material efficiency (ME) 14:762
Material engagement theory (MET) 4:26, 4:28
Material feminism 9:723, 9:726–727
Material flow analysis (MFA) 11:850, 13:242, 14:760–764
accounting standards and headline indicators 14:761–762
applications 14:762–763
decoupling 14:761, 14:763
green economy 14:763
history 14:760–761
industrial metabolism 14:760–761
input–output analysis 14:762–763
material efficiency 14:762
physical accounting 14:760
policy use 14:763
satellite accounts 14:762–763
social metabolism 14:760, 14:763
stocks and flows 14:761, 14:763
sustainable consumption 14:763
sustainable production 14:763
Material flow decisions
supply chain management 23:710
Material footprint (MF) 14:762
material justice 13:617
Material practices 22:438

Material Product System (MPS) 16:229, 16:231
Material production and distribution, Internet 12:588–589
Material progress, comparative histories of 10:166–167
Material realities 10:151
Material realm, and ultramodern shadow 5:247–248
Material resources 21:135
Material semiotics 5:524
  actor-network theory 1:82–83
Material threat, and xenophobia 25:804
Material Transfer Agreement 11:841
Material welfare 25:505–506
Material world 14:765–770
Materialism 12:389, 14:642, 21:261, 22:526, 23:645
  children, effects of media use on 15:38–39
  and cultural anthropology 1:766
  dialectical 14:645, 14:647, 14:663
  in human history 14:642–643, 14:645
  Marxist geography 14:674–675
  and media 15:5–6
  and mind–body dualism 15:547
  in natural science 14:642, 14:645
  production mode 15:715
  in social science 14:642
Materialist theories
  of history 10:491–492
  of personal identity 17:743
Materialistic philosophy 20:286
Materialistic societies in new unified theory 22:881–882, 22:881t
Materiality 1:897, 17:305
  and body 2:721
  and culture 14:765
    actor network theory 14:765, 14:769–770
    artifacts 14:766–770
    capitalism 14:766–768
    collections 14:766–767
    colonialism 14:766
    historical materialism 14:766–768
    material world 14:765–770
    museums 14:766–767
    object-oriented ontology 14:769–770
    phenomenology 14:768–769
    ritual exchange 14:766
    sociality 14:765
  money 15:747
  untranscended, of fetish 9:105–108, 9:110
Materialization 10:115–116
  images and models as harbingers of 25:223
Materials, developmental research 6:329–330
Materials and Resource Conservation 7:774
Material-semiotic generative nodes 22:3
Maternal agency, in affluent nations 20:470
Maternal behaviors
  aspects of 13:327
  and neonatal bonding 16:470–472
  toward infants 2:141
Maternal bonding disorder 17:723–724
Maternal death 15:883
Maternal depletion syndrome 9:42
Maternal depression 22:157–158
  effects on child-directed speech 3:402
Maternal deprivation, and child care 3:391
Maternal education, and socioeconomic status 13:324
Maternal employment, and child care 3:391
Maternal fertility, and religion 20:318–319
Maternal health services, and family planning 8:788, 8:790–792
Maternal immune hypothesis 21:777
Maternal interactive behavior 2:138
Maternal linguistic input, and language acquisition, associations between 3:399–400
Maternal love 7:515
Maternal mortality 6:120–121
Maternal Protection Law 8:235
Maternal psychosocial vulnerability, adverse impact of 17:724
Maternal relatives
  marriage with 11:730
Maternal sensitivity
  attachment 2:138–139, 2:141
  and attachment theory 12:231–232
Maternal stress 19:363
Maternity, and population dynamics 18:581–582
Maternity leave, family policy 8:802–803
Math anxiety 8:505–506
Math education
  cross-cultural study of education 5:327–329
Math self-concept (MSC) 12:425–429
Mathematical analysis, of variables 7:613
Mathematical cartoons 18:566
Mathematical demography 6:125, 14:372
Mathematical economics 18:359
Mathematical expectation, probability 19:34–35
Mathematical expressions of age structure 1:341–342
Mathematical finance 14:632–633
Mathematical graduation 6:130–132
Mathematical ideograms 14:243
Mathematical imagination, constraints of 15:659–660
Mathematical language for spatial structure 25:197–198
Mathematical learning theory 14:589–590, 14:801–802, 14:812–813
  discrete-state models 6:525
  finite state Markov chains 14:801–802
  linear operators models 14:802–803
Mathematical lesioning 18:58
Mathematical linguistics 14:296
Mathematical literacy 5:763–764
Mathematical methods, for diffusion process 6:397–398
Mathematical modeling 4:475
  implicit association test (IAT) 11:705–707
  see also Computational approaches
Mathematical modeling approaches, cognition 4:99–100
Mathematical models
  central business district (CBD) 14:786
  for choice behavior 5:895–896, 5:900
  fourth paradigm of science 14:789
  in human geography 14:785–790
    agent-based models (ABMs) 14:786
    applications of models 14:788
    cellular automata (CA) models 14:786
    challenges 14:788–789
    computational complexity 14:789
    definitions 14:785
    deterministic *versus* stochastic models 14:787
    forecasting models 14:788
    geocomputation 14:789
    global or local models 14:787
    integrated *versus* isolated 14:787–788
    issues 14:786–787
    macroscale 14:786
    mathematical framework 14:787–788
    mesoscale 14:786
    microscale 14:786
    multi- or interspatial 14:787
    optimization models 14:788–789
    prospects 14:788
    simulation models 14:788
    space–time modeling 14:788–789
    spatial analysis 14:785–787
    spatial autocorrelation 14:788
    spatial scale 14:785–787
    statistical investigation 14:788
    typology 14:785–788
  input–output models 14:786
  location-allocation models 14:788
  methodologically integrated models 14:788
  optimization models 14:788
  in philosophy of science 14:791–799
    debates 14:792–795
    learning 14:794–795
    pluralism 14:794
    representations 14:793–794
    symmetry relationship between equations and structures 14:795–798
    validation 14:794
  simulation and forecasting models 14:788
  for social change 22:197
  spatially integrated models 14:787–788
  substantively integrated models 14:787
Mathematical playworlds boost reasoning 14:778–783
Mathematical programming 23:138–139
  computerized test construction 4:511
Mathematical psychology 4:103, 14:800–807, 14:809–811, 14:813, 14:815, 19:418
  Bayesian statistics 14:805
  Bayesian theory 14:809
  contributions of Fechner and Thurstone 14:800–801
  contributions of Tversky 24:679–682
  correlation coefficient 14:811–812
  decision field theory 14:805
  differential psychology 14:808, 14:812, 14:815
  European Mathematical Psychology Group 14:805

experimental psychology 14:808, 14:810–812, 14:815
factor analysis 14:812
game theory 14:813
information science 14:808, 14:814–815
journals, researchers, society 14:805
law of comparative judgment 14:801, 14:804
marginal utility 14:809
mathematical developments
in English psychology 14:811–812
in German psychology 14:810–811
mathematical learning theory 14:801–802
finite state Markov chains 14:801–802
linear operators models 14:802–803
measurement and choice 14:804–805
before nineteenth century 14:808–810
in nineteenth century 14:808–810
paired-comparison scaling 14:814
psychometrics 14:812, 14:814
psychophysics 14:802–803
direct scaling methods 14:803
functional equation methods 14:803
signal detection theory 14:803–804
response time 14:810–811, 14:814–815
response time models 14:805
roots of in twentieth century 14:812–815
choice and decision making 14:813–814
language and thinking 14:814
mathematical learning theory 14:812–813
measurement and scaling 14:814–815
perception and psychophysics 14:813
sequential sampling models 14:804–805
signal detection theory 14:813
Society for Mathematical Psychology 14:805
stimulus sampling theory 14:804
stochastic processes 14:814
theory of grammar 14:814
utility theory 14:804, 14:809, 14:813–814
Mathematical reasoning 14:773–774
Mathematical statistics, contribution of Neyman to 16:825, 16:827
Mathematical theorems, reasoning for proving 5:938
Mathematical theory of evolution 13:24
Mathematical thought 6:267
Mathematically gifted 7:158
Mathematically talented 7:158
Mathematically talented children, capacities of 14:777–778, 14:778f
Mathematico-deductive theory of learning 13:703
Mathematics 13:70, 14:243, 14:453, 18:100
concepts development during early childhood 14:772–784
contributions of Putnam 19:663, 19:665
developments
in English psychology 14:811–812
in German psychology 14:810–811
education 14:816–821
competence 14:818
disposition 14:818
embodied cognition 14:818–819
learning 14:818–821
problem solving 14:816–820
self-regulation 14:818
situated cognition 14:819
teaching 14:819–821
theories 14:817–818
and educational neuroscience 16:732
and mental imagery 15:223
before nineteenth century 14:808–810
in nineteenth century 14:808–810
philosophy of 14:870–871
social 22:990
Strong Program analysis 23:596–597
and structuralism 23:627
Mathematics learning
automaticity in 4:67–68, 4:68t
and cognitive development 4:66–75
concrete and abstract instantiations 4:68–69, 4:69t
individual differences in 4:72–73, 4:73t
in intervention effectiveness 4:73, 4:73t
interventions
explicit feedback 4:69–70, 4:70t
implicit feedback 4:69–70, 4:70t
that combat stereotype threat 4:71–72, 4:72t
learning
from cognitive tutors 4:71, 4:71t
from games 4:71, 4:71t
from worked examples 4:70, 4:70t
mindliness 4:67–68, 4:68t
misconceptions in learning 4:66–67, 4:67t
Mathematization 19:700
hypothesis tests 11:504
Mathews v. Eldridge 6:698
Mathieson, T., works, *Defences of the Weak* 18:922
Matilda effect 2:332
Mating
and marriage systems 14:611–612
sociology 22:896
Mating effort
human behavioral ecology 11:250
Mating patterns, genetic consequences of 18:898
Mating preference 11:335–337
Mating strategies 11:340
for men
long-term mating 11:341–342
short-term mating 11:340–341
for women
long-term mating 11:342–343
short-term mating 11:343–345
Mating system 18:897
Matiruku (periodic insanity) 14:487
Matlab Health and Socioeconomic Survey (MHSS) 8:564
Matrifocal family 1:640, 8:810
Matrifocality
analytic dimension of 14:823
definition 14:822
family policy and poverty causes 14:823
and feminism 14:823
term
and context 14:822–823
derivation of 14:822
Matrimandir 2:152
Matrix algebra 9:626
of least squares problem 9:626
network analysis 16:518–519
Matrix factorization (MF) with orthogonality constraints 16:119
Matrix of domination 12:660
Matrix structures, in structural contingency theory 23:613
Matsuda, Mari 5:267
Mattachine Society 9:633
Mattei, Ugo 13:805
Matteotti Crisis 16:257
Matter 16:508
Matthew effect 2:332, 9:618–620, 14:52
Matthew Shepard and James Byrd, Jr. Hate Crime Prevention Act 10:540–541
Matthews, M.H. 7:166
Matthiessen, F.O., works, *American Renaissance: Art and Expression in the Age of Emerson and Whitman* 1:619
Maturana, Humberto 14:383
Maturation 12:451, 12:454, 12:653–654
and crime 1:312–313
experience of 1:319–320
life management 14:85
and personality 17:865, 17:868
Mature adult language functioning 16:586–587
Mature behavior, parents' expectations for 8:733
Mature capitalism 10:151
Mature dependency 21:151
Mature theories, of knowledge society 13:107
Maturity gap 6:681
Mau Mau rebellion 16:269
Maududi, Maulana 4:264, 20:375
MAUP *see* Modifiable areal unit problem (MAUP)
Maurer, Bill 2:873, 9:174
Mauri 11:785–786
Mauriac, Francois 2:394
Mauryan Empire 23:54–55
Mauss, Marcel 5:499, 6:706–707, 7:13, 10:485, 14:825–827, 22:828, 23:1
biography of 14:825
collaboration with Henri Hubert 14:826
on conditions of collective solidarity/cohesion 14:827
and Durkheim 1:767, 6:244
on education, intellectual development, and academic options 14:825
on kula ring 13:142
legacies of 1:767
Maori concept of *hau* 7:13–14
Maussian holism 7:416–417
on nationalism 16:291
primitive economic systems, study of 14:827
on scholarly achievement 14:827
study of Eskimo life 14:826–827
study of total social facts 14:827
works

Mauss, Marcel (*continued*)
Barter, Exchange, and Value 7:14
The Gift: Forms and Functions of Exchange in Archaic Societies 6:703, 7:13, 17:154–155, 23:2
Money and the Morality of Exchange 7:14
Primitive Classification 13:70
The Social Life of Things 7:14
Techniques of the Body 2:729
MAVA *see* Multiple abstract variance analysis (MAVA) design
Max Hardcore 17:220–221
Max Planck institute for Behavioral Physiology 14:366–367, 14:370
Max Planck Institute for Demographic Research (Rostock, Germany) 5:739
Max Weber Gesamtausgabe project 25:461
Maxdiff model
for best–worst choice 2:550
properties and uses of scores for 2:551–552
Maximal covering location problem 23:174t
Maximal-contribution upper bound 4:873
Maximin designs 17:260
Maximin principle 12:940
Maximization 7:12–13, 12:388, 15:305–306
under constraint 12:388
Maximizes your expected utility (MEU) 20:927
Maximizing conception of rationality 18:778–780
Maxims, politeness as adherence to 18:326–327
Maximum a posteriori (MAP) estimation 2:369, 23:413–414
Maximum entropy principle 19:40
Maximum expected net benefits rule 5:53
Maximum likelihood 8:16, 8:688, 23:379–380
Pearson on 17:662
Maximum likelihood estimation (MLE) 2:368–369, 4:451, 4:470, 7:116, 8:695, 12:708, 14:116–117, 14:145, 14:148, 15:274, 15:624–626, 16:48, 17:292–293, 18:604, 20:733–734, 21:415–416, 23:203, 23:262, 23:414, 23:616
asymptotic theory of 8:695
Fisher's studies of 9:223–224
for generalized linear mixed models 9:848
computing estimates 9:848–849
numerical computations 9:848
other numerical approximations 9:848–849
for ignorable models 15:605
logistic regression 2:590–594
of loglinear models 16:127–128
for nonignorable models 15:605–606
in social analysis 14:579
statistical sufficiency 23:421
Maximum likelihood method 15:654, 15:656
in factor analysis 8:705–706
Maximum linkage 23:393
Maximum power principle 7:615–616
Maximum principle 17:253–254
conditions for optimal solution 17:253
for complicated optimal control problems 17:254
optimal time path of the control variables 17:254
Maximum random utility models, best–worst scaling 2:551
Maximum relative likelihood 14:118
Max-Planck Institute for Anthropology 3:317
Max-Planck-Gesellschaft 14:369
Maxwell, Henry, on division of labor 6:602–603
Maxwell, James Clerk 21:226, 21:293
Maxwell Jones model 24:274
Maxwell's Laws 9:479
May, Robert 7:792
Maya civilization
irrigation 12:725
sociocultural aspects 1:621–622, 1:624
Maya language
Whorf's studies of 25:575–576
Mayali, gender assignment in 9:798
Mayan arena 3:535
Mayan culture, childhood in 9:451
Mayan parents, prosocial behaviors 19:246, 19:248
Mayans, horticulture 11:205–206
Maybon, Charles, works, *Histoire Moderne du Pays d'Annam (1592–1820)* 11:84
Maybury-Lewis, David 1:224
Mayer, Tobias 9:625
Mayer–Salovey four-branch model 7:508
Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) 7:445, 7:508, 17:755
Mayflower Compact 1:616–617
Mayo, Elton 12:672
Mayr, Ernst 8:386
works, *Systematics and the Origin of Species* 19:825
Mayrinax Atayal argument nominalizations 23:892t
Maze 19:69
enriched environment 2:835–836, 2:836f
solving performance 19:71
tasks 19:71–72
Mazower, Mark, works, *Hitler's Empire* 6:814
Mazrui, Ali A., works, *The Trial of Christopher Okigbo* 14:253
Mazzarella, William, on affect 1:239
Mazzei, George 13:889
Mazzini, Giuseppe 4:339
MB *see* Mushroom body (MB)
MBCT *see* Mindfulness-based cognitive therapy (MBCT)
MBD *see* Multiple baseline design (MBD)
MBHOs *see* Managed behavioral health organizations (MBHOs)
MBI *see* Maslach Burnout Inventory (MBI)
MBPSL *see* Medical behaviors that potentially shorten life (MBPSL)
MBSR *see* Mindfulness-based stress reduction (MBSR)
MBT *see* Mentalization-based psychotherapy (MBT); Mentalization-based therapy (MBT)
MC *see* Monte Carlo (MC) methods; Multiple comparisons (MC)
Mc Kay, Henry D., works, *Delinquency Areas* 7:561
MCA *see* Multiple correspondence analysis (MCA)
McAdam, Doug 22:435
McAlister, John, works, *Viet Nam: The Origins of Revolution* 11:85
McAllister, Carlota, works, *War by Other Means: Aftermath in Post-Genocide* Guatemala 25:124
MCAR *see* Missing completely at random (MCAR)
McBride, Katherine 16:685
McCaa, Robert 3:287
McCabe, Janet, works, *Sex in the City* 2:219
McCall, Grant 11:417
McCall, William A. 8:568–569
works, *How to Experiment in Education* 22:298
McCarthy, Anna, on television 25:81
McCarthyism 18:103–104, 18:410–412
McClelland, Jay L. 12:304
McClintock effect 17:198
McCloskey, Deirdre 7:101
McCloskey, Donald 7:96
McCloy–Heggestad–Reeve unfolding model for multidimensional ranking blocks 17:846–847
McCosh, James 20:287
McCrae, R.R., five-factor model of personality 9:231
McCullagh, C.B.
works
The Logic of History 11:15
The Truth of History 11:15
McCullough, Karl-Erik 10:110
MCDA *see* Multicriteria decision analysis (MCDA)
MCDI *see* MacArthur Communicative Development Inventory (MCDI)
McDonaldization 5:536, 10:235, 14:828–830, 18:667, 22:94
advantages 14:829
critiquing and restudying 14:830
de-McDonaldization 14:830
disenchantment 14:829
globalization 14:828–829
and Internet 14:830
irrationality of rationality 14:828–829
neoliberalism 14:829
process 14:828
theorizing and studying 14:829–830
and Weber 14:828–829
McDonald's 10:235, 10:247
McDonough, Patricia, on cultural capital 5:377
McDougall, William 19:413–414
on instinct 12:178–179
works

An Introduction to Social Psychology 8:460, 12:178–179
McDowell, Linda 13:164
MCDs *see* Minor civil divisions (MCDs)
MCFA *see* Mixtures of common factor analyzers (MCFA)
MCFD *see* Ministry for Children and Family Development (MCFD)
McGee, Terry 23:663
McGilchrist, Iain 23:295
McGill Illness Narrative Interview (MINI) 10:567–568
McGrath, Joseph E., works, *Small Group Research* 22:103–104
McGurk effect 16:95
m-chlorophenylpiperazine (m-CPP), for obsessive–compulsive disorder 17:115
MCI *see* Mild cognitive impairment (MCI); Minimal cognitive impairment (MCI); More complete income (MCI); More comprehensive income (MCI); My Class Inventory (MCI)
MCII *see* Mental contrasting with implementation intentions (MCII)
McIntyre, Dan 13:32
MCK *see* Massive category killing (MCK)
McKay, H.D., on social class and crime 5:122–124
McKean, Samuel H., works *The Encyclopedia of Psychology: A series of graphic descriptions of human characteristics* 7:591–592
McKenzie, Lionel 7:915
McKenzie, Roderick
on social ecology 22:263–266
works
Behavioral Theory of Labor Negotiation 16:431
McLennan, John 13:802
McLeod, Poppy Lauretta, on networks and creativity 5:467
McLeod, Robert 10:733
McLuhan, Marshall 2:729, 12:603, 18:918, 22:583
works
The Gutenberg Galaxy 18:918
Understanding Media 18:918
MCM *see* Metacognitive model (MCM)
MCMC *see* Markov chain Monte Carlo (MCMC) method
McMichael, Philip 19:677–678
McNally, David 1:456
McNamara, Robert 5:100
McNaughton Test 9:347–348
McNeill, William H., works, *Plagues and Peoples* 11:45
McNeille completion 17:300
McNemar, Quinn 8:569
McNemar's test 19:143–144
McNollGast 19:536
MCP *see* Multiple comparisons procedures (MCPs)
m-CPP *see* m-chlorophenylpiperazine (m-CPP)
MCPs *see* Multiple comparisons procedures (MCPs)
MCT *see* Map of cognitive transformations (MCT); Metacognitive therapy (MCT)
MD *see* Medial dorsal nucleus (MD)
MDD *see* Major depressive disorder (MDD); Methyl donor deficiency (MDD)
MDE *see* Minimum detectable effect (MDE)
MDGs *see* Millennium Development Goals (MDGs)
MDL *see* Minimum description length (MDL)
MDMA *see* 3,4-Methylenedioxy methamphetamine (MDMA)
MDP *see* Minimal distance principle (MDP)
MDPs *see* Markov decision processes (MDPs)
MDS *see* Multidimensional scaling (MDS)
MDSO *see* Most different systems–same outcome (MDSO)
ME *see* Material efficiency (ME); Minority ethnic (ME) elders
MEA *see* Millennium Ecosystem Assessment (MEA)
Mead, George Herbert 1:93–94, 2:273, 10:740, 12:244–245, 14:831–836, 18:808, 22:267, 22:319, 22:321, 23:431
contributions to social psychology 14:835
early life and education of 14:831–832
on formation of the self 14:834–835
influence in social reform 14:833
influence of thought 14:835–836
professional career and social activism of 14:832–834
published materials 14:835–836
on symbolic interactionism 12:352, 12:354–355, 23:857
works 14:834–835
Mind, Self, and Society: From the Standpoint of a Social Behaviorist 14:835, 22:222, 22:533
Movements of Thought in the Nineteenth Century 14:835
The Philosophy of the Act 14:835
The Philosophy of the Present 14:835
Theory of Communicative Action 14:835
Mead, Lawrence 18:736
Mead, Margaret 1:742–743, 1:747, 2:712, 7:145, 9:459–461, 14:837–840, 16:237, 20:669–670, 25:561, 25:808–809
biography of 14:837
and Boas 1:766–767
cultural criticism of 1:766–767
education 14:837
study of socialization and education of Manus children 14:837
study to anthropology 14:837–838
anthropological insights and models 14:838
on women's rights and issues 14:838
works
Balinese Character: A Photographic Analysis 14:837–838
The Changing Culture of an Indian Tribe 14:837–838
Childhood in Contemporary Cultures 14:838
Coming of Age in Samoa 1:728–729, 14:837–839, 24:579
Culture and Commitment: A Study of the Generation Gap 14:838
The Fateful Hoaxing of Margaret Mead 14:839
The Feminine Mystique 14:838
An Inquiry into the Question of Cultural Stability in Polynesia 14:837
Male and Female 14:838
Margaret Mead and Samoa 14:839
The Mountain Arapesh 14:837–838
New Lives for Old: Cultural Transformation – Manus, 1928–1953 14:838
Sex and Temperament 14:837–838
Mead Report (American Anthropological Association) 8:56
Meade, James 16:228
Mead–Freeman controversy 14:838–839
MEADMIX (software) 9:888
Mean age of marriage 21:792, 21:791t
Mean function 14:153–154
functional data analysis 9:514–516
regression analysis 14:157–160
Mean interaction contrast (MIC) 12:80
Mean of the observed score variable, classical true score model 19:424
Mean reversion, in stock markets 23:483–484
Mean squared error (MSE) 1:659, 8:17, 19:426
Mean squared groups (MSGrps) 1:659
Mean squared prediction error, unit roots time series 24:366
Mean utterance length (MLU) 13:325
Mean world syndrome 15:30
Meaning 2:131, 5:577, 9:357–358, 9:361–362, 14:841–849, 23:901–902
as advocate 17:243–244
archaeology of symbolism and ritual 20:721–723, 20:726
in art 2:16–17
conceptualization of 16:561–562
constitutive rules 14:846–847
and context, relationship between 21:586
contextual theory of 9:121
and convention 14:848–849
and cultural sociology 5:499–502
and culture 5:617–618
Davidson on 14:843
in dexis 6:52–53, 6:55
discourse as sociohistorical meaning-making practice 6:510
of disgust 6:546–547
dramaturgy 6:646–647, 6:649
facts of use, correctness, and normativity 14:845–846
Geertz on 9:662, 9:664
grammatical and lexical 12:783

Meaning (*continued*)
historical background 14:841–842
in historiography 11:100–101
Kripkensteinian paradox 14:842–843
bare normativity of meaning 14:843–846
conventionality of meaning 14:844, 14:848–849
dimensions of semantic normativity 14:843–844
normativity of content: assertion 14:844, 14:847–848
normativity of content: belief 14:844, 14:847–848
prescriptive normativity of meaning 14:844, 14:846–847
rule-based normativity of meaning 14:844, 14:846–847
lexicography 13:937–938, 13:940
logic of 14:293–295
Luhmann on 14:391–392
meaning, rules, and prescriptions 14:846–847
naturalism 14:841, 14:843
normativity, truth, and meaning 14:844–845
normativity of content 14:847–848
of number 14:775
photographic 18:85–86
problem of 5:579–580
and purpose
and religion 20:265
in qualitative research 19:686–688
reductionism 14:843
regulative rules 14:846–847
as selection 14:391–393
and selfhood 21:446–447
and sexuality 21:819–821, 21:823
and social networks 16:561–566
analysis 16:562–563
forms 16:563–565
and social systems, Luhmann 14:384
in spirituality 23:285–286
and symbolic interactionism 16:561
of technology 24:113–114
verifiability criterion of 14:316–317
Wittgenstein on 14:841–843, 14:845, 14:847
Meaning construction, of visual words 25:216
Meaning finitism 23:594–595
and Strong Program 23:594–595
Meaning Reconstruction Theory 5:862
Meaning response
placebo as 18:158–159
placebo-controlled double-blind studies 18:157–159, 18:161
Meaning-constituted boundaries 14:394
Meaning-focused coping, chronic illness 3:550–551
Meaningful action 20:666
Meaningful gestures 1:875
Meaningfulness, representational theory of measurement 14:859–860
Meaninglessness 1:544, 14:392
Meaning-making 1:737, 4:729–732, 9:990, 9:992
cultural mediators's role 5:442
Meanings of humor 11:418–419
Meanings of mechanisms 14:874–876
Meaning-seeker approach, for social categorization 22:188
Mean-level change, personality stability 9:238
Mean-level stability, and adolescent personality 17:864–865
Means, imitation by infants/young children on 11:622
Means testing 25:477
welfare reforms 25:515, 25:518
Means–ends chains 25:49
Means–ends relationships 18:203, 24:118, 25:49, 21:504f
Mears, Ashley, on cultural industries 5:536
Measles 18:175–176
Measure terms (classifers) 3:812
Measured variables, latent variables 13:394–396
Measurement approaches
informal economics 12:51–52
underground economics 12:51–52
Measurement devices, in experimental laboratories 8:557–558
Measurement dimensions, of effectiveness factors 7:220–221
Measurement equation 24:354
Measurement error
in EITMs 18:456, 18:458
extreme response style 23:758–759
latent variables 13:396
scientific explanations 21:306
in sexuality surveys 15:315–316
Measurement error models 14:850–855
effect of 14:851
estimation in 14:852–853
multiple linear regression 14:852
nonlinear models and 14:854
related methods 14:853–854
factor analysis 14:853
regression calibration 14:853–854
regression with instrumental variables 14:853
simple linear regression 14:850–852
simple simulated example 14:851–852
Measurement invariance 19:420
Measurement issues, temperament development 24:195
Measurement model 19:419, 23:406
Measurement of engagement 1:31–32
observational measures 1:32
self-report measures 1:32
Measurement of work and free time 13:845–846
Measurement precision 19:419–420
Measurement scales, Stevens on 23:454–455
Measurement theories
conjoint 14:862–864
conjoint measurement 14:870
degree 14:868–870
derived measurement 14:869–870
and development of quantitative physics 14:868–870
extensive magnitudes and real numbers 14:868
intensive magnitudes and concept of continuous quantity 14:868–870
measurement of psychological attributes 14:870
fundamental measurement 14:869–870
history and philosophy 14:868–872
and issues external to quantitative science 14:870–871
philosophy of mathematics 14:870–871
philosophy of science 14:871
mathematical psychology 14:804–805, 14:814–815
operationism 14:871
ratio 14:868, 14:870–871
representationism 14:870–871
Measurement unit 5:323–324
Measurement unit equivalence 5:319–320
Measurements 3:785, 3:819, 20:239
child and adolescent psychiatry 3:388
in econometrics 7:8–10
of facial expression production 8:669–670
frequency, cardiovascular reactivity 3:119
income 11:739–741
income distribution 11:732–738
inequality 11:732–738
issues of single-case designs 21:985
on job satisfaction 12:833–834
performance 25:592–593
purpose 24:98
quantification 19:700–704
reliability 20:239–246
representational theory of 14:856–861
self-report scales 25:592
source 24:98–99
customers 24:98–99
objective sources 24:98–99
supervisors 24:98–99
team members 24:98–99
trained raters 24:98–99
of subjective well-being 23:648–649
teacher motivation 24:66, 24:69
timing 24:99
Tversky on 24:680
of uncertainty 23:424
validation 24:99
construct validity 24:99
content validity 24:99
criterion-related validity 24:99
validity 20:239–246
Measures of affirmative action 1:248–250
indirectly inclusionary measures 1:249
'outreach' programs 1:249
positive measures to avoid unlawful discrimination 1:248–249
preferential treatment 1:249–250
tie-break policies 1:249
training underrepresented group 1:249
Measures of potential for prevention, epidemiology 7:830
Measuring faking 17:836–837

anchoring vignettes 17:837
Bayesian truth serum 17:837
decision trees 17:837
item response theory (IRT)-based approaches 17:837
overclaiming 17:837
social desirability scales 17:836–837
word use 17:837
Measuring innovation 12:149
Measuring skills 7:215–216
Measuring slavery 22:46–48
Meatless Mondays campaign 1:568
Mecelle-i Ahkam-i Adliye (Ottoman civil code) 15:361
Mechanical computers 23:376–377
Mechanical energy 7:626
Mechanical Jurisprudence 13:773
Mechanical Marxism 18:363–364
Mechanical model of universe 20:286
Mechanical objectivity, tools of 19:702–703
Mechanical philosophies 20:286
Mechanical problem solving 1:878
Mechanical social solidarity 4:361–362
Mechanical solidarity 4:338–339, 6:701–702, 13:162, 18:668, 23:3, 23:6
Mechanical Turk, experimental laboratories 8:554
Mechanism of action, of antidepressants 1:779–780
Mechanism of missing data 15:602–603
Mechanism operationalization, Empirical Implications of Theoretical Models 18:455
Mechanism-based causal analysis 14:873–879
Mechanism-based explanation 1:670–671, 22:415–417
in analytical sociology 1:668–672
Mechanism-based thinking, macro/microsociologies 14:417–419
Mechanisms 3:269, 14:875–877
of kin selection 13:29–30
Mechanistic explanation 14:873–874
Mechanization 2:906, 6:32
of manufacturing 11:855, 11:858
of reasoning 5:939
MECHEM (scientific discovery model) 21:295
Medals 2:333–335
Med-arb 15:70
Medecins Sans Frontieres (MSF) 19:533, 25:433
Media 5:704, 11:121, 11:393, 11:396, 11:398, 17:711, 21:682
and aggression 1:378
audiences 2:216–222
and body image 2:736–740
and celebrity 3:277
and child development 15:1–10
ADHD-related behaviors 15:4–5
adolescence 15:3
aggressive behavior 15:4
core subject knowledge 15:7
creativity and imagination 15:7
early childhood 15:2
infants and toddlers 15:1–2
Internet 15:3, 15:5
materialism, parent–child conflict, and childhood overweight 15:5–6
middle childhood 15:2–3
negative effects 15:4–6
positive effects 15:6–7
prosocial behavior 15:6
school readiness 15:6–7
sexual socialization and sexual risk behavior 15:5
television 15:1–8
video games 15:1–8
in China 2:68–69
and climate change 3:853–859
agenda-setting approach 3:858
anthropogenic climate change as a global problem 3:854
behavioral effects 3:858
climate change as relevant media issue worldwide 3:856
climate journalists 3:855–856
climate politics 3:853, 3:856
content analysis 3:856–858
coverage from science of climate change to its societal impacts 3:856–858
difference in content of climate change coverage 3:858
ecological modernization 3:855
economic development 3:854–855
environment 3:855
event-driven media coverage 3:856
frames and sponsors 3:855t
global warming 3:854–855
Internet 3:853, 3:858
knowledge about climate change 3:858
media attention for climate change 3:857f
public sphere 3:854
relevance 3:853
representations 3:856–858
research field 3:853–854
scientific uncertainty 3:854
sources for climate change information 3:858
stakeholder communication 3:854–855
uses and effects 3:858
collective memory 4:183–186
communication and transportation 4:273–276
and community 4:332–336
community attachment model 4:332–333
geographically based online communities 4:335–336
local community 4:332–333
online newspapers 4:333
print newspapers 4:333
virtual communities 4:333–335
consumption 16:785
and crime 15:23–28
and cultural production 5:611
educational 15:36
effects on infant development 15:35
effects on political efficacy 7:303–304
and genetics 9:990–993
in Gulf War 25:399
influence on East Asian politics 6:798, 6:801–802
information society 12:83–86, 12:88
integration of economic education teaching methods and 24:84–86
and international advertising 12:460
and interviewing 12:672
in Korea 2:75
and manifestos 18:417–418
in Mexico 1:629–631
and music 16:152
nanotechnology in 16:191–192
narcotizing dysfunction 13:638
and political agendas 1:364
and political campaigns 3:67–68
and polling 18:459, 18:461, 18:463
and popular cultural 18:528–531
and pornography 17:219–221
and postmodernism 18:676–679
proliferation 18:529
psychiatric ethics 8:68–69
and public sphere 19:605–607
critical issues 19:607–608
mass media 19:605–607
principle of publicness 19:603–605
public 19:605–607
and religion 20:362–363
and science 21:163–166
and semiotics 21:594–595, 21:594f
and social movements 15:19–22
and social sciences 22:583–587
contexts 22:584
democracy, autocracy, dictatorship 22:584–585
effects 22:586
as infrastructure 22:583
as institutions 22:583
mass media 22:583–587
and modern state 22:584
new media 22:585–586
political communication 22:585–586
politics 22:583–586
power 22:583–584
studied by various social science disciplines 22:586–587
as systems 22:583–585
technological change 22:585
uses 22:585
in South Asia 2:80
technologies and social life 25:81–82
terrorism and 24:228
transnational 18:528–529
visual turn 25:207–209, 25:211–212
Media agendas 1:351–352
sources 1:354
Media anthropology 2:873
Media bias 14:714–715
Media centers 14:46
Media communication, and political psychology 18:430
Media content 5:705
Media effects models 15:29–34

Media effects models (*continued*)
agenda setting 15:29
contemporary 15:31–33
Media effects on children 15:35–41
advertising 15:38–39
childhood obesity 15:39–40
cognitive effects 15:35–37
comprehension 15:35–36
educational media use 15:36
infant development 15:35
relationships with academic outcomes 15:36–37
emotional effects 15:37
coping responses 15:37
fear responses 15:37
emotional responses 15:37
general learning model 15:37–38
health effects 15:39–40
body image concerns 15:39–40
weight 15:39
mitigating negative effects 15:40
parental mediation 15:40
social and moral development 15:37–39
materialism 15:38–39
prosocial effects of media exposure 15:39
racial and gender stereotyping 15:38
theorization 15:37–38
TV content effects on non-physical aggression 15:38
violence effects on physical aggression 15:38
youth 15:38
Media elites 7:388
Media environment 25:102–103
Media ethics 14:694, 15:42–47
fairness 15:43
freedom and responsibility 15:44–45
privacy 15:44
self-regulatory mechanisms 15:44–45
Internet 15:46
journalism 15:42–43, 15:45–46
media user ethics 15:46
quality of media content 15:45–46
truth and objectivity 15:42–44
Media events 15:48–54
audience 15:48–52
celebration 15:48–52
ceremonial television 15:48, 15:51–52
Dayan on 15:48–49, 15:51–52
Katz on 15:48–49, 15:51–52
in multichannel, hyperpartisan, and globalized world 15:52–53
paradigmatic conception 15:48–49
ritual 15:48–53
theoretical implications and conceptual tensions 15:50–52
Media exposure, measuring 7:252
Media history 15:11
audience 15:11, 15:15–16
cinematography 15:12–13, 15:15
communication 15:11–16
film industry 15:15
general systems theory 15:14
historical change 15:12–13
information 15:11–14, 15:16
major trends 15:11–12
print culture 15:14
propaganda 15:13–14, 15:16
public opinion 15:12
public sphere theory 15:13–14
reading 15:12–13, 15:15–16
research methods 15:15–16
television 15:12–16
theories of impact 15:13–15
theories of structure and content 15:15
writing 15:11, 15:13
Media imperialism 15:55–59
arguments
about audiences and effects 15:57
about genres 15:57
from and about historical process and effects 15:56–57
about media policy and regulation 15:57
about production and flow 15:57
about state and democratization 15:58
cultural imperialism paradigm 15:55
empirical research 15:55–56
political/policy struggles 15:56
rise of globalization 15:58
Media multiplexity hypothesis 22:425
Media opportunity structure 15:20
Media outlets for pornography 18:614–615
Media policies 2:860, 2:862, 24:164–165
Media politics and regulation 2:863
Media power 14:721–722
Media production 24:165
Media regulation 2:860, 2:862–863
Media reporting, of climate change 3:841–842
Media representations, of minorities 8:144
Media richness theory, of computer-mediated communication 4:505–506
empirical findings 4:505–506
Media slant 14:715
Media sociology 13:852
Media sports
analytical approaches to 14:721–724, 14:724t
forms, genres, and texts 14:720–721
Media sports cultural complex 14:719–720
Media sports political economists 14:722
Media standing 15:21
Media studies, history of 4:278–283
on east–west battlefield 4:281–282
Media talk shows 15:60–63
alleged 'effects' of 15:62
arguments in 15:62
audience participation 15:62
definition 15:60
evaluations and critical perspectives 15:62–63
history of 15:60
melodramatic and moralistic forms of 15:63
'on stage'/'backstage' arrangements 15:61
personal narratives in 15:61
phone-in program 15:61
Media usage, and social networks 16:557
Media use 5:705–706, 5:710
Media violence 25:100–104
effects type 25:101–102
consequences of 25:102
degree of realism of the violence 25:101–102
extent and graphicness of the violence 25:101
nature of the perpetrator 25:101
nature of the victim 25:101
presence of humor 25:102
presence of weapons 25:101
punished violence 25:102
reason for the violence 25:101
rewarded violence 25:102
and public policy 25:103
theories of effect 25:100–101
and young viewers 25:102
Media-bound communication 4:274
Media-centered political behavior 14:714–715
Mediacy 12:783
Media-designated leaders 15:21
Medial centrality measures 16:535
Medial dorsal nucleus (MD) 18:817
Medial forebrain bundle (MFB) 2:842, 2:844, 14:127
Medial intraparietal sulcus (mIPS) 5:26–27
Medial (MeA) of the amygdala 8:870
Medial prefrontal cortex (MPFC; mPFC) 16:602, 16:640, 18:817–818, 18:821, 18:836, 16:602f
dorsal areas 18:818
and fear conditioning 1:812–815
processing and integration of affective and social information 18:818
ventromedial portion of 18:818
Medial temporal cortex 18:817
Medial temporal lobes (MTLs) 1:644, 5:923–924, 19:274–275, 24:554, 5:924f
and episodic memory 7:858–860, 7:860f
and explicit memory 4:107
functions 10:877–878, 10:878f
serial order memory 21:639–640, 21:642
source monitoring 23:31
virtual spatial navigation 16:352
Medial thalamic nuclei 13:679
Median absolute deviation (MAD) 20:733
Median income 6:540
Median lifetime 23:785
Median polish (exploratory data analysis) 8:602–603, 8:603t
Median preoptic area, and sleep 22:89
Median test 16:886
Median voter model, democratic transitions 6:97
Media-oriented terrorism 24:235
Mediascapes 18:531
Media-sharing sites 24:146
Media–sports nexus 14:724–725
Mediated models
and genetic factors 9:922
and investigations of attitudes 9:922
and organizational research 9:922
and personality factors 9:922

Mediated moderation 22:549–550
Mediated sports scandals 14:724
Mediating mechanisms, coronary heart disease 4:920
Mediating variable 15:64–69
causal analysis of 15:66–68
in chronic disease research 15:64
complicated mediation models 15:68
effect 15:64
in regression models 15:65–66
tests of 15:65–68
examples 15:64–65
future directions 15:68
in prevention and treatment 15:65
in psychological theories 15:64
use of term 15:64
Mediation 15:70–74, 16:512, 25:80–81, 25:323–324
and adjudication 5:103, 5:105–106
and brokers 2:870–874
conflict 4:627–631
in conflict resolution 4:586
and cultural participation 5:449
and dance 5:668–669
disputes 6:560–561, 6:564–565
and moderation 15:672
peace/nonviolence 17:626
principle of 1:641
of social process 16:149
statistical 15:75–80
Mediation model 22:549
Mediation structure 13:55
Mediations: Journal of the Marxist Literary Group 14:649
Mediatization 14:719, 14:724, 15:50–51, 25:80–81
and discourse 6:506
Mediator variable method 4:842–843
Mediators
of aggression 1:379
controlled trials 19:389
cultural intelligence as 5:437
genetic factors as 9:922–923
judge *versus* 12:879
teachers as 24:60
team performance measurement 24:96–97
taskwork 24:96–97
teamwork 24:97
Mediatory myths 16:182
Medicaid 10:594–595, 10:599–601, 10:643–644, 11:209, 11:735, 12:219, 15:852, 16:907
Oregon rationing program 10:604–605
Medical advice, adherence, and doctor–patient interaction 6:617–618
Medical anthropology 2:496, 7:816, 10:571
and reproduction 20:452–454
Medical arms race 10:598
Medical behaviors that potentially shorten life (MBPSL) 7:599–601
arguments around 7:601
Medical causes, for children and infant mortality 12:22–24
Medical education 15:81–85
for future 15:83–84
as a professional field of inquiry 15:83
twentieth century 15:84
Medical ethics 2:615–616, 2:631–632
allocation of resources 2:616
beginning of life 2:615
bioethics 13:559–560
clinical practice 2:615–616
end of life 2:615
history of 8:63–64
Medical geography 15:86–91
causation 15:88–90
chronic disease and environment 15:88
environmental health 15:87–89
epidemiology 15:86–90
evolution of 15:86
exposure 15:89–90
genetics 15:86, 15:88–90
GIScience 15:88
health geography 15:86, 15:90
political ecology of disease 15:89–90
social environment 15:88–89
social epidemiology 15:89–90
spatial analysis of infectious disease 15:88
spatial data 15:88–90
spread of disease 15:86–88
toxicology 15:89
Medical identity theft 5:634, 5:637
Medical injury, tort claims and 24:432–433
Medical jurisprudence 10:594
Medical knowledge, moving to physicians and consumers 18:9–10
Medical liability insurance 12:216–217
Medical malpractice, remedies and damages and 20:414–415, 20:417
Medical model
behavior therapy 2:422
for complex and behavioral interventions 18:162
of crime prevention 18:878
of disability 6:447, 6:454
Medical outcomes, in cardiac disease coping 3:112, 3:114
Medical paternalism 2:615–616
Medical personnel, for intimate partner abuse 12:681–682
Medical perspectives, of gender variant individuals 24:549–550
Medical procedures 22:735–740
Medical profession 15:92–97, 15:784
and complementary and alternative medicine 10:564
dominance 15:94
changes in 15:96
heterogeneity 15:92–93
historical development 15:93–94
medical knowledge 15:94–95
medical practice 15:95–96
Medical research, gender and 9:694–696
Medical Research Council
National Survey of Health and Development 14:68
Medical risk, gay/lesbian tourism 9:641–642
Medical savings accounts (MSAs) 10:614–615
Medical Schools Council Assessment Alliance (MSCAA) (UK) 7:189
Medical social work 10:719, 10:716t
Medical sociology 15:98–104
aging 15:100, 15:104
comparative health systems 15:100–101
depression 15:99–100
disparities 15:99–100, 15:103
health maintenance organizations 15:101–102
morbidity 15:98
mortality 15:98–100
national institutes of health 15:98
population health 15:98, 15:100, 15:103
qualitative health research 10:696
rationing 15:100–102
social psychology of health and health care 15:102–104
clinician–patient interaction 15:103
disability and quality of life 15:103–104
neighborhood influences 15:103
social networks and social influence 15:103
stress and coping process 15:103
social stratification and health 15:98–100
age and life course 15:100
gender 15:99
race and ethnicity 15:99–100
socioeconomic status 15:98–99
Strauss on 23:527–529, 23:531
study of health care organization and provision 15:100–102
changing role of physician 15:101–102
intractable cost crisis and managed care 15:102
technological imperative and public expectations 15:101
Medical technology 10:618
Medical view of disability 13:498
Medical work, Strauss on 23:529
Medicalization 10:798, 15:105–109, 18:8
anthropological perspectives 15:106
biomedicine 15:106, 15:108
caveats about 15:106
characteristics of 15:105–106
countermovements against 22:976–977
death/dying 5:865–866
of deviance 6:672
of disability 6:454–455
disease 15:106
eating disorder 6:823
end of life 7:601–602
of homosexuality in China 21:735–737
medical social control 15:105
medical sociology 15:105, 15:107
and mental disorder 22:976
and mental health 15:106–107
and obesity 17:85
social consequences 15:107–108
sociological roots 15:105
spirit possession 23:283
use of term 15:105
uses of 15:108

Medicalization (*continued*)
and women's health 15:107
Medicalization theory 22:976
Medical–legal conflicts, evaluating 13:557–558
Medically unexplained physical symptoms (MUPS) 15:497
Medically unexplained symptoms (MUS) 8:599, 23:868–870
of psychosomatic disorders 19:485–488
Medically unexplained symptoms and somatization 5:472
Medical/social pathology model 13:808–809
Medicare 10:594–595, 10:643–644, 11:209, 11:735, 12:219, 15:848–849, 16:907
Medication backup 19:495–496
Medications 19:463–465
anticholinergic 19:465–466
antipsychotic 19:465–467
and delirium 19:467
metabolic side effects 19:464–465
for panic disorder 17:470–471
and placebo effects 19:467
for psychiatric disorders 22:382
SNRI 19:468
SSRI 19:468
Medici, Lorenzo di 21:212–213
Medici Palace in Florence 16:142
Medicinal cannibalism 3:85
Medicine 2:631–632, 18:174, 18:177
behavioral *see* Behavioral medicine
biomedicine 10:573–574
CAM *see* Complementary and alternative medicine (CAM)
evidence-based 8:327–331
geriatric psychiatry in 10:104–106
honor in 8:64–65
hypnosis 11:473–474
Illich's contribution 11:585–587
personalized medicine 17:925–931
postcolonial science studies 18:652–655
psychosomatic 19:484–490
quantification in 19:705
Strauss on 23:528
women in 25:626
Medicine (western), history of 15:118
institutional and organizational history 15:120–122
and molecular health 15:124–125
scientific and conceptual history 15:118–120
transition to molecular medicine 15:123–124
Medicine in National Socialism 15:122
Medicine-in-the-community 4:402
Medicines sans Frontieres 4:401
Medicolegal context, forensic archaeology 9:334
Medieval Christendom 18:239
Medieval cities in Europe 24:774–776
Medieval deliberation 6:62
Medieval economy, guilds 2:48
Medieval Europe, minstrelsy of 17:909
Medieval Jewish historiography 11:67–68
Medieval jurist 13:777–778
Medieval period, South Asia 23:56
Medieval Persian chronicles, Indo-Islamic period 11:79
Medieval rural history 17:672
Medieval science, scientific revolution 21:318–320
Medieval university 22:589–591
organization of knowledge 22:588–589
Medievalists 2:703–704
revolution of renaissance 20:426
Mediodorsal thalamic nucleus 18:811–812, 18:816
Meditation 2:886–887, 2:889, 2:891–892, 20:356
and free association 24:726
mental illness 19:308
Mediterranean 15:354, 25:34
and honor 11:181–184
Mediterranean Europe 8:275
Mediterranean syllabaries 25:775–776
Crete 25:775–776
Cyprus 25:776
Medium
photography as 18:81–86
publishing as 19:620–623
radio as 19:877
Medium-term expenditure frameworks (MTEF) 7:263
MEDLINE database 22:944
Medullary lamina 5:923–924
Meduna, Ladislas Joseph von 23:18
Medvedev, Dmitrii 1:941
Medvedyev v. France 2010 18:228
Meese, Edwin, III 10:455
Meet them and plead them lawyering process 15:593
MEG *see* Magnetoencephalography (MEG)
Megachurches 20:362
Megafauna, Australia and New Guinea 2:231–232
Megaliths, in South Asia 23:54
Megalomaniac 25:398
Megalopolitan structure, urban 24:801
Megamachine 21:179
Megan's Law 20:434
Megaprojects 24:838
urban *see* Urban megaprojects
Megoran, N. 3:319–320
Mehring, Franz 22:834
MEI *see* Migration Effectiveness Index (MEI)
Meier, Gebhard Theodor, works, *Historia Religionum* 20:322
Meiji Restoration 12:794, 15:129, 16:276, 17:162
engagement of Western sociology 24:583–584
Meiji Revolution 15:126–130
centralized nation-state 15:128–129
comparative societies 15:129
East Asia 15:126, 15:128
historiography 15:129
issues of debate on 15:129–130
modernity 15:129
modernization 15:127–130
nationalism 15:129–130
overthrow of *bakufu* 15:126–127
restoration of imperial rule 15:127
Meiji shrine 20:368
Meinecke, Friedrich 11:93–94
on nationalism 16:310
Meiners, Christoph, works, *Grundrb der Geschichte aller Religionen* 20:323–324
Meinig, D., works, *Shaping of America* 11:18–19
Meiroku Zasshi 24:584
Meirokusha publishing house 24:584
Meissner corpuscles 15:967
MEK *see* People's Mujahedin of Iran (MEK)
Mekong river system 25:444–445
Melamed, A.D. 19:170
Melancholia 14:484, 15:193, 19:318
Melanesia 2:567, 2:571–572, 17:153–160, 17:162–163, 17:446
anthropology 2:569–570
cargo cults 3:139
diversities 17:153
equality/hierarchy 17:155–156
ethnography 2:567
gift/commodity 17:154–155
kinship 17:155–156
before Lapita 2:230
missionization 17:153
modernities 17:156–157
political typologies 2:567–569
primitive capitalists 2:569
sexual antagonism 2:570
social and cultural anthropology in/of 17:154
South Pacific, islands of 17:154f
Melanesian contradiction 2:569
Melanoma, incidence 23:696
Melatonin 3:569–571, 3:573
neuroendocrinology 16:650–651
Mellin transform 9:375
Mellon, Andrew F. 2:24
Melodic intonation therapy, for language disorders 13:332
Melodrama 9:161–162, 15:62
Melody 13:273, 13:276–277
and neonatal crying 16:479
Melton, E.N., on lesbian athletes 9:659
Meltzoff, Andrew, like me theory 15:131–134
Melucci, Alberto 22:508
Member check, in feminist research 9:718
Members
in voluntary associations
types 25:253
Members of coping network 20:224
Members of family 8:693–694
factor analysis 8:693–694
latent class models 8:694
latent profile models 8:694
latent trait models 8:694
Members of the European Parliament (MEPs) 9:699, 9:702
Membership

and political citizenship 3:650–653
requirements for voluntary associations 25:258
social science learned societies/professional associations 22:674–675
tribal
Native Americans 1:635–636
Membership in public, and public reason 19:595
Membership in society 2:697
Membership roles 17:532
Membrane potential, neurons 16:704
Memes
evolution of language 8:368
Memetics 8:438–439
and cultural evolution 15:816
Memmi, Albert 15:135–139
biographical facts 15:135–136
impact 15:137–138
and Tunis 15:135
and Tunisian independence 15:135
woks
La statue de sel 21:752–753
works 15:136–137
Agar 15:136
Albert Memmi 15:135–136
The Colonizer and the Colonized 15:136–137
Decolonization and the Decolonized 15:136–137
Dictionnaire critique à *l'usage des incrédules* (*Critical Dictionary for the Use of Non-Believers* 15:136–137
Juifs et arabes (*Jews and Arabs*) 15:136–137
La Dépendance (*Dependence*) 15:136–137
La Statue de sel (*The Pillar of Salt*) 15:136
L'Aliénation 15:136
Le Buveur et l'amoureux (*The Drinker and the Lover*) 15:136–137
Le Désert ou la Vie et les aventures de Jubaïr Ouali El-Mammi (*The Desert or the Life and Adventures of Jubaïr Ouali El-Mammi*) 15:136
Le Mirliton du ciel (*The Reed Pipe of the Sky*) 15:136
Le Pharaon (*The Pharaoh*) 15:136
Le Portrait du colonisé précédé du Portrait du colonisateur (*Portrait of the Colonized Preceded by Portrait of the Colonizer*) 15:136
Le Racisme: Description, definition, traitement (*Racism: Description, Definition, Treatment*) 15:136
Le Scorpion ou la Confession imaginaire (*The Scorpion or the Imaginary Confession*) 15:136
L'Ecriture colorée, ou Je vous aime en rouge (*Writing in Colors, or I Love you in Red*) 15:136
Les Damnés de la terre (*The Wretched of the Earth*) 15:136
Les Français et le racisme (*The French and Racism*) 15:137
L'Homme Dominé (*Dominated Man*) 15:136–137
Libération du juif (*Liberation of the Jew*) 15:137
Nomade immobile (*The Immobile Nomad*) 15:136
Philological Quarterly 15:136
Portrait du colonisé précédé du Portrait du colonisateur 15:136–137
Portrait d'un Juif (*Portrait of a Jew*) 15:137
Térésa et autres femmes 15:136
Testament insolent (*Insolent Testament*) 15:136–137
Memo writing 10:405
Memoria 20:651
Memorials, war 16:306–307
Memories 17:286
and remembering 17:286
resilience from 22:574–575
Memory 6:812, 7:849, 7:851, 9:905–907, 14:809–811, 14:813, 14:815, 23:111–114, 23:873–880
and aging 1:398–406
animal models 9:905–907
and archaeology 1:881–884
'academic' meanings 1:883
archaeology and national agendas 1:883–884
archaeology as memory 1:883
archaeology *versus* memory 1:882–883
collective identity 1:882–883
and collective memories 1:882–884
contributions of Assmann 1:882
contributions of Halbwachs 1:881–883
contributions of Nora 1:882–883
cultural memory 1:882
culture-historical approach 1:883
heritage 1:882–883
history 1:881–883
Les Lieux des Mémoire 1:882
memory culture 1:882
memory studies 1:881–882
mnemonics 1:881–882
nation states 1:883–884
personal memory 1:881
remembrance as social construction 1:881–882
social memory 1:883
theoretical approaches 1:881–882
associative 4:105
auditory 4:104
autobiographical *see* Autobiographical memory
and bilingual aphasia 2:575–578
and biography 2:638–639
brain regions 9:905, 9:907
calmodulin kinase II 9:906–907
and cerebral cortex 3:336, 3:338
and change blindness 3:344–348
childhood memories 3:465–470
collective 4:235–236
collective memory 4:188–193
colonialism 4:184
of commemoration 4:236–237
commoditization 4:184–186
computational neuroscience 4:482
concept learning 4:538–541
confabulation *see* Confabulation
consolidation 13:919, 13:922–924
consolidation and reconsolidation of 23:878f
consolidation at cellular level 23:874–876
and countermemory 4:184
and cultural history 5:422
defining 15:147–148
digital technologies 4:183, 4:186
and dreaming 6:651–653, 6:655–656
drosophila 9:905–906
dual models 4:103–105
and emotions 7:461–462, 7:464–465
and ergonomics 11:300
experimental study of 4:10, 4:12
explicit memory 23:876–877
eyewitness testimony 8:642–645, 8:647–648
first screens for learning and memory mutants 9:906
flashbulb memories 4:196
future of genetic manipulations 9:907
and future thinking 9:555–559
global hegemonies and local struggle 4:184–185
globalization 4:183–184, 4:186
and historical consciousness 11:115
in historiography 11:100–101
and history 4:181
in honeybees 15:155–159
iconic 4:104
impairments related to dysexecutive amnesia 6:719–721
implicit memory 23:876–877
and infant cognitive development 4:58–59, 4:61–62
influence of color terms on 13:294–295
interoception in 12:618
Köhler on 13:124–125
and language
in cognitive science 4:105–106
and lateralization of animals 2:802–804
and limbic system
emotional memory 14:127–128
hippocampal memory system 14:128–129
localization of remembrance 4:183–184
long-term depression 23:873, 23:875–876, 23:875f
long-term memory 23:874–879
long-term potentiation 9:906–907, 23:873, 23:875–879, 23:875f
markets 4:185–186
media 4:183–186
and mental imagery 15:222–223
models in cognitive psychology 4:104–105
neural basis of 13:677–683
neural systems of 21:898–899
and neuromarketing 16:698, 16:701
nondeclarative 16:844–850
and nostalgia 17:34–35
and pain 17:454–455
place cells 23:876–877

Memory (*continued*)
plasticity 23:873–875, 23:877–878, 23:875f
politics of time 4:181–182
postcolony 4:182–184
and posttraumatic amnesia 18:636–638
prospective memory 19:270–278
recall and reconsolidation 23:877–879
recognition *see* Recognition memory
reconstructive nature 4:190
remembering 4:183, 4:185–186
sensory perception 4:104
serial order 21:639–644
shape of 4:182–183
sharing of similar cellular mechanisms 23:876–877
short-term memory 23:874–879, 23:878f–879f
signal transduction 23:874–875
and sleep 22:63–69
brain plasticity 22:63–64, 22:67
consolidation of declarative memories 22:66
consolidation of nondeclarative memories 22:66–67
contribution of within-sleep stimulations 22:67–68
emotion and sleep-dependent memory consolidation 22:68–69
hippocampo-neocortical dialogue 22:64f
interaction 22:65–67
learning 22:63–69
long-term memory 22:63
mechanisms of sleep-dependent memory consolidation 22:64
neuronal replay hypothesis 22:64
rapid eye movement sleep 22:63–69
sequence learning 22:67
sleep cycles and alternating sleep phases 22:63–64
slow wave sleep 22:64–68
synaptic downscaling 22:64
synaptic homeostasis theory 22:65f
social and cultural influences on 21:899
cultural factors 21:899
family factors 21:899
social contagion of 4:191
social dimensions 4:189
in Soviet Union 20:823–824
space 4:183–186
and symptom perception 23:869–870
synaptic bases of 23:873–874, 23:874f
synaptic plasticity 9:906–907
synaptic strength 23:873–878
synaptic transmission 23:873–875
temporal distinctiveness models of 24:199–205
and temporal lobe 24:208
temporality 4:182–183
traces, lifetime and transformations of 23:879f
transgenic/knock-out studies 9:906
interpretation of 9:907
trauma 4:181, 4:183, 4:185–186
violence, commercializing 4:185–186
and virtual spatial navigation 16:348–350
Memory assessments 16:724–725
Memory binding 15:140–146
Alzheimer's disease 15:144
association formation in LTM 15:143–144
attention 15:141–145
chunking 15:140, 15:143
convergence 15:144
episodic buffer 15:141–143
hippocampus 15:144–145
neural basis of 15:144–145
parietal cortex 15:145
prefrontal cortex 15:144–145
shape–color binding problems 15:141, 15:145
synchronization 15:144
in working memory 15:140–143
cross-domain and cross-modal binding 15:142–143
verbal binding 15:143
visuospatial binding 15:141–142, 15:141f
Memory consolidation, sleep-dependent 22:64
Memory conversations 2:292–293
American mother–child dyad 2:292
Chinese mother–child dyad 2:292
Memory cue, and retrieval 20:608–609, 20:611–612
Memory culture 1:882
Memory development
early *see* Early memory development
during infancy and early childhood 15:147–154
assessing explicit memory across development 15:148
autobiographical memory 15:147, 15:149, 15:151–153
autonoetic consciousness 15:151
childhood amnesia 15:147, 15:151–153
defining memory 15:147–148
emergence of episodic memory 15:151–152
encoding 15:148
explicit memory development 15:148–151
implicit memory 15:147–148
infantile amnesia 15:151
long-term retention 15:149
memory retrieval 15:149–151
mother–child talk 15:152–153
Ngā Pūmahara o Ngā Tamarikitanga 15:152–153
recognition memory 15:151
visual paired comparison 15:148
nonverbal tests and 21:896
social and cultural influences on 21:899
cultural factors 21:899
family factors 21:899
verbal tests and 21:896
Memory disorders 1:398
Memory distortion, and false memories 8:711–712
Memory impairments 1:879
LTP/LTD 14:362
and transient global amnesia 24:563
Memory loss 1:644
and Alzheimer's disease 1:584–585, 1:588
Memory measures, of inferences 19:395–396
Memory models 1:646–649
discrete-state models 6:525
quantitative 15:160–165
Memory of the griots 11:90
Memory organizations 19:56
libraries as 14:47
Memory recall, emotion and 7:479
Memory reconstruction processes 4:565–566
control mechanisms underlying 4:565–566, 4:565f
Memory representations
retrieval 20:608
Memory retrieval 15:149–151
Memory storage, emotion and 7:478–479
Memory studies 1:881–882
Memory system model 6:759
Memory systems
multiple memory systems theory 11:710
single-system processing theory 11:710
Memory tasks 3:571
Memory trace, and retrieval 20:608–612
Memory-based model
of discourse 12:41
of language comprehension 21:625
Memory-related neural substrates and future thinking 1:648–649
Memory-sharing
cultural practices of 2:291–293
styles 2:290
Memos, in qualitative research 19:685
Men
conversation style of 13:257
diversifying 14:681–682
fathering *see* Fathering
issues is social work 14:680
and leisure 13:854
mating strategies of 11:340–342
long-term mating 11:341–342
short-term mating 11:340–341
measurement of sexual motivation in 21:762–765, 21:764t–765t
profeminist sensibility in social work 14:682
sexual dysfunctions in
diagnostic procedures 21:690
psychological treatments of 21:690–691
validated treatments for 21:690–691
theorizing 14:680–681
and time use 24:369
in Women's Profession 14:679–680
working with 14:682–683
Men health 15:166–171
beyond gender 15:170
gender expectations 15:167–169
societal messages around 15:169–170
male–female differences 15:166–168, 15:170
morbidity 15:167, 15:170
mortality 15:166–167

social factors 15:169–170
Men who have sex with men (M-Coalition; MSM) 21:736–738, 21:795
AIDS-related mortality 15:835
HIV infection 11:141–142
impact and epidemiology 11:141
improving future prevention efforts 11:142
intervention efficacy 11:141–142
selected predictors of risk 11:141
MENA *see* Middle East and North Africa (MENA)
Menarche 10:463, 14:336–337
and transition to adulthood 24:578–579
Menchú, Rigoberta 1:621–622
Mendel, Gregor Johann 2:450, 8:218, 8:234–235, 8:254, 9:222, 11:311–312, 15:209
genetics studies
Fisher's analyses 9:223
Mendelian genetics 8:388–389
Darwinism 5:682
Mendelian inheritance 9:932–933
Mendelian patterns 9:915, 9:918
Mendelian transmissions 8:718
Mendelism 8:218–219, 8:221, 8:226, 8:230, 8:241
Mendel's laws 8:386
Menger, Carl 1:574, 4:340, 11:818–819, 21:168, 21:304, 24:745, 25:7
works
Grundsätze der Volkswirtschaftslehre 25:9–10
Menken, Jane 10:776–777
Mennell, Stephen, works, *Alexis de Tocqueville on Democracy, Revolution and Society* 16:237
Menninger, Karl 23:682
Mennonites 2:147
Meno 3:176
Menopause 7:608
and life history evolution 9:28, 9:30
psychosocial aspects 10:466–467
and sexuality 21:843
Mens rea 13:492
defenses negating 5:221–223
intoxication 5:222–223
mistake of fact as defense 5:222
mistake of law and color of right 5:222
Men's violence 6:633
Menshevik Social Democratic parties 20:856–857
Mensheviks 4:303, 20:857
Menstrual age-based time metric for sexual development 14:334–336
Menstrual cycle
disorders 10:463
disorders not related to 10:465
sexual dysfunctions 10:465
urinary incontinence 10:465
disorders related to 10:462–465
endometriosis 10:464
female menstrual cycle 10:462–463
infertility 10:465
menarche 10:463
menstrual cycle disorders 10:463
polycystic ovary syndrome 10:464
premenstrual syndrome and premenstrual dysphoric disorder 10:463–464
Mensural element 3:813–814
Mental ability, and intelligence 12:309–311
Mental ability tests 17:854
Mental affections 18:178
Mental arithmetic, cardiovascular reactivity 3:118
Mental associations 11:714
prediction 11:717
process purity of 11:718
Mental attributes, and mind–body dualism 15:545–546
Mental calculations 6:76
Mental capacity 19:321–322
Mental Capacity Act ` 7:599
Mental causation 3:261–265, 15:172–178
history 15:172–173
interventionism and 15:177
new compatibilism 15:176
problem of 15:173–174
canonical solutions 15:174–176
generalization 15:174
Mental chemistry 19:410
Mental comorbidity
of gastrointestinal disorders 9:608–609
depression, anxiety, and stress in functional GI disorders 9:608–609
depression, anxiety, and stress in IBD 9:608
Mental constructions with accents 14:773–774, 14:773f
Mental contrasting
and motivation 15:938
and motivation/actions 15:889–890, 15:892
Mental contrasting with implementation intentions (MCII) 15:892
Mental development, and meditation 16:512
Mental disorders 8:340, 15:188–189
in diabetes 6:339–340
anxiety 6:340
depression 6:339–340
eating disorders and disordered eating behaviors 6:340
and dyadic coping 11:594
epidemiology of 15:189
and essentialism 8:15
experience in different ways 8:341
interoception in 12:618–619
depression 12:618–619
eating behavior and eating disorders 12:619
panic disorder 12:618
somatoform disorders 12:619
as medicalized social construct 22:976
negative symptoms in 19:291
from neurodevelopmental perspective 15:188–189
and normality 17:3–4
in old age 10:106–107
physical disease following 4:391
religion and 20:264
and sleep 22:60–61
Mental disorders, genetics of 15:209
affective disorders 15:213–214
molecular genetics 15:213–214
twin studies 15:213
schizophrenia 15:211–212
molecular genetics 15:211–212
twin studies 15:211
twentieth century developments 15:209–211
Mental Disorders Act 1916 6:43–44
Mental energy 16:504, 17:873–874
Mental experience, components of 5:579
Mental fitness 8:249
Mental functions 3:329
Mental growth 25:575–576
Mental growth prediction, Vygotsky on 25:314
Mental handicap 13:684
Mental health 1:329–331, 10:522–523, 15:187–192, 19:586–587, 20:355, 22:386, 24:274
adolescent *see* Adolescent mental health
applying evolutionary developmental psychology to 8:423–427
antisocial behavior, evolutionary developmental model and 8:427
evolutionary developmental psychological models 8:424–425
evolutionary life history framework 8:426–427
mental health model 8:423–424
risky adolescent behavior, evolutionary developmental model and 8:425–427
of butch–femme women 13:872
caregiving 3:135, 3:137
child *see* Child mental health
of child soldiers 3:435
concepts 15:187
consequences of military operations and disasters 15:496–497
culture as determinant of 5:570–575
conceptualization of culture 5:570
coping 5:571–572
cross-cultural awareness 5:574
cross-cultural perception and expression of mental illness 5:572
cross-cultural psychology 5:570
Diagnostic and Statistical Manual of Mental Disorders 5:571
etiological beliefs and folk approaches 5:572–574
impact of ignoring culture in therapy 5:574
nondominant ethnic minorities 5:570
non-Western healing practices 5:572–573
Western psychotherapy 5:572–573
worldviews 5:570, 5:572–574
definition of 15:187
and divorce 14:519–520
effects of physical activity on 18:88–90
anxiety 18:89–90

Mental health (*continued*)
depression 18:89
epidemiology of 15:189
and homelessness 11:161
human subjects protection in research 8:49
of immigrants 15:407–408
influences of emotion regulation on 7:455
and law 21:158–159
and LGBT aging 13:866
of LGBTQ youth 21:87–88
mental disorders 15:188–189
epidemiology of 15:189
from neurodevelopmental perspective 15:188–189
migrants 15:417
Minnesota Multiphasic Personality Inventory (MMPI) 15:566–568
and motivation 15:907, 15:910
and natural disasters 16:316
and normality 17:1–4
personality 15:189
and physical health 15:188
positive functioning 15:188
prevention to 18:884
primary care 15:191
promotion and prevention 15:190–191
policy 15:190–191
selected, indicated and universal prevention 15:190
psychoneuroimmunology 19:448–449
psychosocial interventions for 22:380–385
relapse prevention, and mindfulness 15:550–551
relationships 15:189
religion and 20:263–265
resilience 15:189
resources 15:189
social capital 15:189
social structures 15:189
social support networks 15:189
and social work 15:193–199, 24:751, 24:753–755
in social work 15:200–203
and unemployment 24:733–737
use of term 24:734–736
and war/political violence 25:390–391
well-being 15:187–188
interventions to promote 15:190
Mental Health Act 1935 19:301
Mental Health Act 1959 15:196–197, 19:301–302
Mental Health Act 2007 15:196–197, 19:303
Mental health acts 6:42–43
mental health and social justice 22:389
Mental health disorder
definition 6:342
diagnosis 6:342–345
Patient Health Questionnaire (PHQ) 6:343
Mental health legislation 15:196
reform 15:196–197
Mental health problems 20:334–335
Mental health programs, mass trauma 14:749
Mental Health Religion and Culture 19:311
Mental health social work 15:193
influence of people with lived experiences 15:197–198
legislative reform 15:196–197
policy 15:196–197
psychiatric treatments, advances in 15:193–195
recovery 15:195–196
Mental health treatment 14:462–465
Mental health-care service providers 5:476
Mental illness 6:42–43, 6:45–46, 15:228, 15:230, 19:587–588
association with problematic forms of religiousness 20:350
attitude in Europe 15:229–230
attitude in North America 15:230
categories of 15:197
and civilization 15:576
clinical psychology 24:524
and culture shock 5:616
definition 15:228–229
attitude in Europe 15:229–230
attitudes 15:228
stigma 15:228–229
epidemiology of 15:179–186
assessment instruments and diagnostic classification 15:182–183
assessment of disability and need of treatment 15:182
burden 15:181
core health challenge of twenty-first century 15:184–185
coverage of disorders 15:182
cross-sectional design 15:180, 15:183
in first decade of twenty-first century 15:180t
general population 15:180, 15:182
linking epidemiology to clinical and basic neurobiological research 15:183–184
longitudinal studies and causal analytic epidemiology 15:183–184
prevalence and correlates in community 15:180–181
prevalence in primary care 15:181
progress and unresolved issues in descriptive epidemiology 15:181–183
questions related to comprehensive need assessment 15:184t
relevance of epidemiology of contemporary psychiatry 15:184t
representative sample 15:179, 15:183–184
risk factor 15:179, 15:183–184
subthreshold conditions and dimensional approaches 15:182
etiology of
early years of development 15:205
environmental factors before and during birth 15:205
genetic influences 15:204–205
later environmental causes 15:205–206
stress as a precipitant of mental disorder 15:206–207
experience in different degrees 8:340–341
family and patient organizations in Europe 15:216–219
genetics of 15:209
hierarchy of models 15:194f
homosexuality and 3:889–890
and human development 11:267–268
Kraepelin on 13:127, 13:129
legislative reform in 15:197
mania *see* Mania
melancholia 15:193
negative attitudes 15:231–232
age 15:231
educational status 15:232
gender 15:231
interventions 15:232
professions 15:232
religious beliefs 15:231
types of illness 15:231
obsessive-compulsive disorders 15:193
prophylaxis 19:202–203
recovery model of 15:196, 15:195f
religious beliefs about 19:307–308
Buddhism 19:308
Christianity 19:307
diagnosis 19:311
Hinduism 19:308
Islam 19:307
Judaism 19:308
religiously sensitive care for 20:350–351
social exclusion due to 22:288
and suicide 23:682–684
women in 25:626
Mental imagery 1:874, 23:111
analog-propositional debate 15:220–221
brain stimulation 15:224
definition 15:220
depictivist 15:220, 15:225–226
images generated from short-term memory *versus* long-term memory 15:221
measuring imagery abilities in health and disease 15:225
memory and thinking, cognitive styles and learning styles 4:112
in models of working memory 15:222
neuroimaging 15:221, 15:224–225
occipital cortex 15:222, 15:224
and perception 15:224–225
propositional theory 15:220–221, 15:225–226
representational format of images 15:220–221
and visual cognition 15:222–224
creative thinking 15:223–224
learning and memory 15:222–223
mathematics and numerical cognition 15:223
reasoning and mental models 15:223
Mental images 4:19
Mental logic theory 15:223
Mental mapping 20:652–653, 24:780–781
Mental model hypothesis 18:792–793
Mental model theory 15:223
Mental models 11:300, 12:199, 19:62–63
managerial and organizational cognition 14:480–481

and mental imagery 15:223
of readers knowledge 12:40
of risk 20:712–713
Mental mutilation 13:161
Mental operations 4:531–533
Mental processes
effects of alcohol on 1:508
social cognition 22:204–209
Mental reinstatement, and context dependent memory 4:780
Mental images, and knowledge 19:68–71, 20:67–70, 20:443
in cognitive science 4:106–107
and implicit social cognition 11:716–718
and language of thought 20:70
modality of 4:106–107
and processes 4:19
as prototypes of objects 4:106–107
of stereotypes 23:446–447
structure 4:106–107
visual images in 4:106–107
Mental representations of health and illness
across cultures 10:567–570
ayurveda 10:568–569
biomedicine 10:568–569
ethnomedicine 10:568–569
implications 10:568–569
pluralism 10:569
society 10:567–569
sociology 10:567
traditional medicine 10:568–569
Mental retardation 8:234–235
Mental rotation 21:269
Mental schemes, culture as 9:2–3
Mental simulation, for psychological understanding 13:82–83
Mental speed theory of intelligence 12:292–293
Mental state assessment, child and adolescent psychiatry 3:388
Mental states
in cultural psychology 5:478–479
empathy and 7:549–550
infant gaze behaviors 12:27, 12:30
and intentionality 12:347–349
language 24:269
reasoning in tasks 24:269–270
and theory of mind 15:541
Mental stress 23:540
Mental stress testing 4:918
Mental structure 18:127
Mental time travel 9:556
Mental Treatment Act 1930 15:196
Mental Treatment Act 1935 19:301
Mental Welfare Office (MWO) 15:196–197
Mental well-being 15:187–188
Mental workload, and ergonomics 11:300
Mentalism
Skinner on 22:27
in social work 15:200–203
and worldview 25:753–754
Mentalist explanations 8:594
Mentalistic concepts, autonomous agents 2:310
Mentalistic linguistics 14:228–229
Mentalities 1:708–710, 1:712
history of 4:550–551, 5:420–423
primitive 18:911, 18:913–914
Mentalization, attachment 2:139
Mentalization-based psychotherapy (MBT), BPD 2:759
Mentalization-based therapy (MBT) 19:346
for personality disorders 17:896
Mentalizing 15:539–540, 15:542
Mentally ill 15:228–234
biomedical research on 13:561
Mentally ill mothers 17:724
Mentally ill people, public attitudes towards 15:228–234
Mental–physical comorbidity 4:390–393
anxiety 4:390–393
bipolar disorder 4:390–391
consequences and implications 4:392–393
depression 4:390–393
epidemiology 4:390
mechanism 4:390–392
from mental disorder to physical disease 4:391
from physical disease to mental disorder 4:391–392
shared determinants 4:392
schizophrenia 4:390–392
substance use disorder 4:390–391
Mentoring 15:235–238
context 15:235
ethics 15:235, 15:237
relationships 15:236–237
supervision 15:235–236
checklist for students 15:236
checklist for supervisors and departments 15:236
Mentoring practices 7:277
Mentoring programs, offender rehabilitation 20:195
Mentors, role in nurturing productive giftedness 7:277
Mentors in Violence Prevention (MVP) program 12:681
MEP *see* Ministry of Environmental Protection (MEP)
MEPLAN model 24:890–891
MEPs *see* Members of the European Parliament (MEPs)
MEPSI *see* Modified Erikson Psychosocial Stage Inventory (MEPSI)
MEPV *see* Major Episodes of Political Violence (MEPV) dataset
M-equation (robustness) 20:734
Mercantile organization 11:874
Mercantilism
and political economy 18:356
and Third World 24:301
Mercantilist 3:96
Mercator projection 3:169
Mercenaries 15:518–519
and war 25:378
Merchant capitalism 3:95–96
Merchants, shipping 14:526–530
Mercier, Louis-Sébastien, works, *The Year 2440: A Dream If Ever There Was One* 14:254
Merck Sharp & Dohme 18:8
Mercy for Animals 1:705
Merezhkovskii, Dmitrii 21:741–742
Merge 9:876, 16:3
Merger Commission 18:344
Mergers and acquisitions specialists 16:418–419
Merging proofness (variable-population axioms) 4:875–876
Merit 1:250, 6:229
and intelligence 12:309, 12:311
Merit pay 17:612
Meritocracy 7:904, 13:95–96, 15:239–245, 19:96, 22:325, 22:329
boundary permeability 15:240, 15:243–244
discrimination 15:240–244
equality of opportunity 7:904
as equity 15:239–240
as ideology 15:240–243
advantaged groups 15:241–242
causes of meritocracy beliefs 15:240–241
consequences of meritocracy beliefs 15:241
disadvantaged groups 15:242
legitimization and maintenance of inequality 15:241–242
prescriptive or descriptive meritocracy beliefs 15:242–243
individual mobility 15:240, 15:242, 15:244
justice 15:239–240, 15:242–243
liberalism and 14:24
and prejudice 18:834
social identity theory 15:239–240
as social mobility 15:240
and well-being 15:241–242
Meritocratic principle 7:170, 13:95–96
Merleau-Ponty, Maurice 19:314, 19:316–317
works, *Phenomenology of Perception* 2:729, 18:11–12
Mernissi, Fatima 21:751–752
Merry, Sally, works, *Gender Violence: A Cultural Perspective* 8:902
Merryman, John 13:804
Mertha, Andrew C., on Chinese cultural policies 5:597
Merton, Robert K. 4:135, 5:36–39, 5:203, 6:21, 9:618–619, 12:280, 12:672, 13:638–640, 14:383, 14:875–876, 17:12, 17:685, 18:101, 18:940, 19:50–52, 20:284, 21:169, 21:200–201, 22:196, 25:465
on anomie 1:718, 1:720–721
on ANT 5:523
on changing nature of science 17:13
conception of anomie 25:815
contributions of 15:248–249
mass communication research 15:248
cumulative advantage model of 2:55
on ethos of science 21:314–317
on functional analysis 9:537–538

Merton, Robert K. (*continued*)
on improbable life trajectory 15:249–250
legacies of 15:250
on list of norms 17:11
Merton Thesis 21:262
on paradigmatic vision 15:247–248
on social class and crime 5:123
on social history 22:310
on specialization in social sciences 23:225
on strain theory 12:665, 23:495, 25:815
theoretical construction of norms 17:12
theory of anomie 1:715–716
typology, prejudice 18:839
works
Contemporary Social Problems 15:248–249
Dictionary of National Biography 15:248
Mass Persuasion 7:563, 15:248
The Negro in the Philadelphia Press 15:249–250
Science, Technology and Society in Seventeenth-Century England 15:248, 21:262
The Self-Fulfilling Prophecy 6:21
Social Structure and Anomie 1:715, 6:704
Social Theory and Social Structure 15:247
Thomas Theorem 6:21
Toward a Metric of Science: The Advent of Science Indicators 21:322
The Travels and Adventures of Serendipity 15:247
Mertonian norms 20:529–530
MES *see* Minimum efficient scale (MES)
MESA *see* Mainstreaming Environment and Sustainability in African Universities (MESA) program
Mescaline 10:502, 16:170
Me-self 21:433
Meshing hypothesis 4:114
Mesmer, Franz Anton 11:465, 11:472
animal magnetism theory 19:502
Mesmerism 11:472
Meso level ID 12:196
Mesoamerica 1:956, 11:54, 15:252
agriculture and craft specialization in 15:252
almanac in 11:55
archaeology of 15:252–265
background 15:252–253
chronological dating in 11:55
civilization 15:252–264
classic period 15:259–261
Monte Albán 15:260–261
Teotihuacan 15:259–260
Tikal 15:261
early formative years 15:257–258
Paso de la Amada 15:258
San José Mogote 15:258
San Lorenzo 15:257
Epiclassic Period 15:261–262
geography and environment of 15:253t
historiography 11:54–57
Indian historiography in 11:54
Late Postclassic 15:262–264
Aztec Triple Alliance 15:262–263
Mixteca Alta of Oaxaca 15:263
Tarascan Empire 15:263–264
logo-syllabic writing system 11:54
Mesoamerican cities 15:252
Mesoamerican civilizations 24:409
Mesoamerican cultures 25:282–284
Mesoamerican historiography, components of 11:54–57
geographical horizon 11:55–56
historical chronology 11:55
origins of humans 11:55
polities 11:55
reflections on history 11:56–57
social history 11:56
sources 11:57
tutelary deities 11:55
Mesocorporatist 15:95
Mesocorticolimbic dopaminergic pathway 16:639–641
Meso-historical approach, migration history 15:446
Mesolect 18:136–137, 18:141–142
Mesolevel analysis, television 25:84
Mesolimbic dopamine pathway 17:953
Mesolimbic dopaminergic system, and dreaming 6:654–655
Mesolithic Europe 8:269–270
Mesolithic South Asia 23:50–51
Mesopotamia, Egypt 15:118
Mesopotamiam irrigation 12:724, 12:726
Mesopotamian clay tablets 9:266
Mesopotamian empires of Babylon and Assur 2:919
Mesopotamian writing 25:773–774
Mesoscale geographic models 14:786
Mesosociology 4:215
Mesosystem 21:682–684
ecological systems theory 6:840
families as educational settings 8:732
and health 6:880
Message construction and perspective 19:356
Message tactics
social influence 22:352
message discrepancy 22:352
self-generated arguments 22:352
Message-learning model 2:170
Messages
and attitude 2:170
community-based health interventions
design 10:647–648
effectiveness 10:648
organizational communication 12:73–74
political campaigns 3:67
Messenger RNA (mRNA) sequences 7:834
Messiah 15:531
Messianic prophesies 20:331
Messinger, Sheldon 13:516–517
Messner, M., on homophobic disposition in sports 9:658
Mestiçagem 9:423
M-estimator 20:734
example 20:734
Mestizaje 11:758, 23:45
and hybridity 11:448–450
MET *see* Material engagement theory (MET); Metabolic Equivalent of Task (MET); Motivational enhancement theory (MET)
Meta-analyses 7:229, 8:618, 8:621, 15:266–281, 16:81, 17:777, 17:789, 17:804–805, 20:196, 21:42–46, 23:906–913, 24:639, 24:643
analyzing data heterogeneity 15:268–269
analyzing effect sizes across a set of studies 23:910–912
Bayesian meta-analysis 23:912–913
Bayesian statistical procedures 15:274–275
confidence intervals 15:275
robust standard errors 15:275
statistical power of tests 15:275
Central Limit Theorem 15:267
of context dependent memory 4:780
data collection, effect size 23:909–910
diagnostic test accuracy 23:912
effect sizes 15:272
evidence-based medicine 8:330
evidence-based psychotherapy (EBP) 8:345–346
examining heterogeneity 23:911
extensions of 15:270
future directions 23:912–913
in generalization 9:839, 9:842–844
of group psychotherapy 10:432
homogeneity analysis 23:911–912
I/E model 12:427–428
inverse variance weights 15:267
logic behind 15:267
mean and variance of effects, estimation of 15:272–274
estimating τ^2 15:274
fixed-effects statistical procedures 15:273
heterogeneity evaluation 15:274
inference models 15:272–273
random-effects statistical procedures 15:273–274
measures of effectiveness of intervention 23:909–910
measuring relationship between two constructs 23:910
modeling between-study variation in effect sizes 15:275–280
estimating τ^2 15:277
fixed-effects models 15:275–276
hierarchical linear models, relation to 15:276–277
mixed-effects models 15:276
multivariate meta-analysis 15:278–280
robust variance estimation 15:277–278
moderator analyses 23:911–912
and organizational diversity 4:739–740, 6:596–597
other measures 23:910
in psychotherapy effectiveness studies 19:506
random effects models 23:912
reporting, standardizing 15:269–270
of school effectiveness 21:81–83
for significance tests 21:960
social scientist experts and 22:695

systematic literature reviews 15:267–268
workplace values 17:824
Meta-analytic studies 15:68, 24:56
Metabolic Equivalent of Task (MET) 10:588–590
'Metabolic flows' in the economy 6:858–859
Metabolic monitoring
schizophrenia 21:25–26
Metabolic response 16:389
Metabolic side effects, of psychopharmacotherapy 19:464–465
Metabolic syndrome (MtS) 21:24
Metabolism, alcohol 1:507
Metabotropic glutamate receptors (mGluRs) 12:711
Metabotropic receptors 16:752–753
Metacognition 4:544, 7:242, 13:713, 15:282, 17:872–873, 19:71
availability deficiency 24:94
definition 15:282, 24:89
development of 24:91
domain-specific *versus* general metacognitive skills 24:91
and education 15:286–287
metamemory and educational processes 15:286
reading, and instruction 15:286–287
school performance 15:286
facets 15:282
instruction and training, principles for 24:93–94
embedded instruction 24:93
informed instruction 24:93
prolonged training 24:93
intellectual styles and 4:114
metacognitive conditional knowledge 24:90
metacognitive knowledge 24:89–90
development of 24:91
metacognitive skills 24:90–91
assessment of 24:92
development of 24:91
instruction and training of 24:93
as self-instructions 24:90–91
metacognitive training 24:94
metamemory 24:89–90
Nelson's model of 24:90
off-line assessments 24:92
online assessments 24:92–93
production deficiency 24:94
self-regulation 24:89, 24:92
teaching for 24:89–95
theory of mind (ToM) 24:91
'tip-of-the-iceberg' phenomenon 24:92
use of questionnaires 24:92
WWW&H Rule 24:93
Metacognitive conditional knowledge 24:90
Metacognitive cultural intelligence 5:433, 5:436
Metacognitive development
assessment 15:283–284
children's factual knowledge 15:284–285
conceptualizations and models 15:282–283
educational implications 15:282–288
metamemory–memory relations 15:285–286
precursors of 15:284–286
procedural metacognition 15:285
Metacognitive knowledge 24:89–90
development of 24:91
Metacognitive model (MCM) 2:172–173
Metacognitive skills 24:90–91
assessment of 24:92
development of 24:91
domain-specific *versus* general 24:91
instruction and training of 24:93
as self-instructions 24:90–91
Metacognitive therapy (MCT) 4:544
Metacognitive training 24:94
Meta-communication 21:595–596
Metacontrast paradigm 4:676
Metacontrol 12:417
Metadata 5:812, 23:128–129
of corpora 5:814
Meta-ethical relativism 6:590, 6:592–593
Metaethics 8:71–73
Metaevaluation 19:119, 21:325
Metafields (field theory) 9:146
Metaheuristics 19:434–435
Metahistories 11:11
Meta-ignorance 24:718
Metaknowledge 12:202
mathematical competence 14:818
Metal, usage in ancient Egypt 1:695
Metal age, Southeast Asian archaeology 23:66
bronze age 23:66
iron age 23:66–67
Metalanguages 5:929, 21:596, 21:596f
Metalingual functions of language 12:783
Metalinguistic awareness 3:510–512, 16:69
Metalinguistic knowledge 2:574, 2:576
Metalinguistic negation 21:589
Metalinguistic view, of counterfactuals 5:88–89
Metalinguistics 25:781–782
Metallurgy 3:371
archaeometry 1:906–907
Metalogical knowledge 14:308–311
Metalogues 1:743
Metamemory 15:282–283, 24:89–90
and educational processes 15:286
taxonomy 15:283
Metamemory impairment 6:720
Metamemory–memory relations 15:285–286
Metamodeling 12:81
Metamorphoses 14:252
Metanarratives 18:680–682
Metanorms 10:436–437
Metaphilosophy, Rorty's view 20:761
Metaphor analysis, qualitative developmental psychology 19:673
Metaphoric competencies 4:299
Metaphoric gesture 10:111
Metaphorical code-switching 22:951
Metaphorical explanations 8:594
Metaphors 10:111, 10:116–117, 13:272, 13:275, 13:278–279, 18:49, 20:651–653
and cognitive semantics 13:933–934
cognitive uses 15:289, 15:292
for computers 4:515–518
dehumanization 6:31–34
Freud on 15:290–291
history of concept 15:289–293
and jokes 15:292
of judgment 13:587
land tenure 13:232
language and culture 15:290
and metonymy 15:290–291
Nietzsche on 15:290–291
of organization 6:237–238
organizational learning 13:695–696
and paradigms 15:291
and philosophy of language 15:291–292
of policing 18:249
policy networks as 18:316
and pragmatic disorders 18:790–792
and rhetoric 15:289
development 15:289
and the social 15:292–293
theories
comparative 15:289
interactive 15:291
as tools 15:292
of written language 14:243–244
Metaphysical Club 18:804
Metaphysical determination 15:176
Metaphysical naturalism 16:326
Metaphysical necessity, physicalism 18:110–111
Metaphysical objectivity 17:96–98
alternative conceptions 17:97–98
independence 17:96–97
Metaphysics 18:436
Latour on 13:415–417
logical empiricism/logical positivism 14:313–317
Metaphysics 14:243
Metaphysics of Middle Ages 18:32
Metaplasticity 18:821
long-term potentiation/long-term depression 14:359
Meta-position, and wisdom 25:601
Metaregimes, definition 16:416
Meta-regression 15:267, 15:269–270, 23:912
Metarepresentations 4:300, 16:505
Metascientific inquiries 24:673
Meta-signs 21:595–596
Meta-stereotypes 17:803
Metatheoretical considerations, cultural evolution 5:405
Metatheory 13:227
MetaTutor 21:544
Metcalfes, J. Stanley
replicator dynamics 8:431
Metcalfes use of replicator dynamic 8:431
Meteorological climate 17:322
Meters 13:266–268, 13:273–276

Meters (*continued*)
for audience measurement 2:212–214
rhythmic structure 13:273
'unstressed' and 'stressed' context 13:273
use of 2:213
Methamphetamines 5:161
Method bias 5:319, 5:323, 7:929
Method of bounds 6:869
Method of decreasing abstraction 1:78
Method of fragments 15:752–753
Method of loci 15:222–223
Method of moments (MOM) 8:16, 14:852, 15:294–301
densities 15:298–299
diffusions 15:298–299
dynamic economic systems 15:294, 15:299–300
GMM estimation 15:295–296
efficiency 15:296
semiparametric efficiency 15:296
time series central limit theory 15:295–296
inference methods 15:298–300
linear models 15:296–298
applications 15:297
efficiency 15:297–298
minimum chi-square estimation 15:294–295
models of financial markets 15:298
moment matching 15:299–300
Pearson on 17:662
related approaches 15:299
Method variance 24:642
Methodenstreit 4:340, 7:875, 10:786, 11:75–76, 11:122–123, 19:678–679, 21:168, 22:992
Methodism 10:491–492
Methodists movement 3:540
Methodological advances, life-course criminology 14:62
Methodological approaches, for social change 22:151
Methodological collectivism 17:934
Methodological developments of educational effectiveness 7:226–227
Methodological dualism 10:781–782
Methodology
offender rehabilitation 20:199
demonstration studies versus routine practice 20:199
design quality 20:199
length of follow-up 20:199
outcome measures 20:199
researcher involvement 20:199
sample size 20:199
Methodological holism 22:997
Methodological individualism 4:188, 6:701, 7:406, 11:818–819, 15:308–313, 16:583–584, 16:787, 17:934, 19:53, 21:141–142, 21:303, 21:449–450, 21:453, 22:997
altruism 15:309
explanation 15:313
infra-individual determination 15:311–312
institutions 15:311
interaction 15:309
and liberalism 14:22–23
micro/macro transition and ontological issues 15:312–313
rationality of action 15:310–311
selfishness 15:309
and social mechanisms 22:418
social situation 15:311
in sociology 15:302
cognition and levels of description 15:306–307
giving primacy to individuals 15:302–303
individual cognition and sociological knowledge 15:305–307
individual motivation 15:303
individualism and form of explanations 15:302–304
individualism and levels of description of complex social facts 15:305
individualistic descriptions and uses of social theory 15:304–305
individualistic point of view and normative thinking 15:304–305
rationality hypotheses and equilibrium 15:303–304
reasons and cognitive processes 15:305–306
Methodological issues, of antidepressants 1:781
Methodological monism 10:781–782
Methodological nationalism 5:44, 16:103–104
Beck on 2:396–399
Methodological naturalism 16:326–328, 16:332
Methodological peace
first 21:411
second 21:411–412
Methodological pluralism 10:783
and critical theory 5:301
Methodological quarrel 22:997
Methodological rationalism, Albert on 1:495
Methodological reification 20:203–204
Methodological relativism 21:311
Methodological sieve 12:81
Methodological sophistication, environmental sociology 7:801–802
Methodological techniques, life-course criminology 14:62–63
Methodology 11:133–134
in EF studies 8:494–495
feminist research 9:717
Feyerabend on 9:118
historical sociology 11:27–29
of history of social and behavioral sciences 15:319–325
implications of realism for 20:24
and individual-collective relationship 11:818
individualism *versus* collectivism 11:818
issues in study of sexuality 15:314–318
participatory action research (PAR) 17:542
postmodernism 18:680–683
social mechanisms 22:415, 22:418–419
survey research 23:748–751
symbolic interactionism 23:856–859
Methods 12:606
elites 7:386–389
Latour on 13:415–417
needs assessment 16:403–405
Methods of science 6:245–246
Methyl donor deficiency (MDD) 7:844
Methylation 7:848–850, 22:282
histone 7:841
social epigenetics 22:282–283
Methylazoxymethanol (MAM) 4:58–59
3,4-Methylenedioxymethamphetamine (MDMA)
effect on prospective memory 19:275
Methylphenidate
cognitive enhancement 17:269–270
METI *see* Ministry of Economy, Trade and Industry (METI)
Metis 1:636
Metonymy 13:933–934, 15:289–290, 18:49, 20:651–653
and metaphors 15:290–291
Me-too medicines 20:524
METR *see* Marginal effective tax rate (METR)
Metraux, Rhoda 14:838
Metrazol 19:587–588
Metric axioms 21:580
Metric data, distance-based clustering of 19:433
Metric equivalence 5:323–324
Metric intelligence scale 2:597–598
Metric multidimensional scaling 16:36
Metric space 14:284–285
Metric tons of carbon dioxide equivalent (MTCDE) 7:775
Metric unfolding 16:46
Metro-Goldwyn-Mayer *see* MGM (Metro-Goldwyn-Mayer)
Metropolis 3:308, 24:856
Metropolis–Hastings (MH) algorithm 2:381, 14:570, 15:763–764, 14:570f
Bayes' theorem 2:381
multiple-block 15:764
Metropolis–Hastings within Gibbs sampling 19:439
Metropolitan areas (MAs) 23:663, 23:665, 24:856–858
Metropolitan art world 2:17–18
Metropolitan character 24:857
Metropolitan districts 24:855–857
Metropolitan divisions 24:857
Metropolitan habitus 10:49
Metropolitan Opera 14:567–568
Metropolitan plans 20:138
Metropolitan regional planning 20:143–144
Metropolitan statistical areas (MSAs) 24:856
recognizing and delineating 24:856–857
Metropolitan Urban Service Area (MUSA) 10:451
Mettrie, Julien Offray de La 20:286

MEU *see* Maximizes your expected utility (MEU)
Mexican Family Life Survey 5:846
Mexican Greens 10:380
Mexican Migration Project 11:632, 11:638
Mexican Revolution 10:188, 15:326–330, 20:621, 20:640
  anticlericalism 15:328
  armed phase 15:327
  Cardenismo 15:328
  historical memory 15:329–330
  Indigenismo 15:328–329
  international relations 15:328
  interpretations 15:326
  origins 15:326
  Pax Priísta (1940s–1960s) 15:329
  (Post)revolutionary reform and state building 15:327
  scholarship on 15:330
  women and gender history 15:329
Mexican Society of Puériculture 8:247–248
Mexican–American War 15:570
Mexico 15:326, 22:623
  carnival in 3:148
  clinical psychology, history 3:881
  educational policy 25:488
  e-waste network 7:368f
  gay/lesbian activism in 13:859–862
  Humboldt on 25:280–282, 25:284–285
  indigenous mobilization and political process in 11:757–759
  laicization in 13:218
  and migration 15:425
  origins of food production in 9:303
  peasants 17:676–677
  population composition by race and ethnicity in 18:556–557, 18:557t
  psychiatric care in 6:43
  sociocultural aspects *see* North America; sociocultural aspects
  sociocultural overviews 1:626–632
    community studies 1:629
    history of anthropology and nation building 1:627
    identity politics 1:628–629
    indigenismo 1:626–627, 1:630
    indigenista legacy 1:626–627
    indigenous identity and social movements 1:628–629
    interdisciplinary and international connections 1:627
    linguistic anthropology 1:629
    media 1:629–631
    migration and (post-)borderology 1:630–631
    modernization 1:627–628, 1:630
    nation-states 1:630
    neoliberalism 1:626–630
    political economy, ecology, and brokerage 1:627–628
    popular religiosity 1:629
    urban space 1:630
    violence 1:627–630
    visual anthropology 1:629–630
Mexico City 3:665
Mexico–US migration from Latin America 15:463–464
Meyer, Eduard 18:435
Meyer, Frank 1:604
Meyer, John W. 10:214, 17:393–394
  on institutions 17:392
  on organizational structures 5:537
  on organizations 17:391
  works
    Institutionalized Organizations: Formal Structure as Myth and Ceremony 17:390, 17:394
Meyer, M., on Native Americans 1:635
Meyer, Richard 2:24f
Meyer, Walter E. 13:516
Meyer Institute 13:516
Meyers, Diana Tietjens 2:317
Meynert, Theodor 9:414
Mezzofanti, Guiseppe 16:66
MF *see* Material footprint (MF); Matrix factorization (MF)
Mf scale 15:567
MFA *see* Material flow analysis (MFA)
M(ultilingualism)-factor 16:69–70
MFB *see* Medial forebrain bundle (MFB)
MFIs *see* Mass fatality incidents (MFIs)
Mgbeoji, Ikechi 18:650
mGENOVA (software) 9:837–838
MGM (Metro-Goldwyn-Mayer) 9:156–157
MH *see* Metropolis–Hastings (MH) algorithm
MHC *see* Major histocompatibility complex (MHC)
MHP *see* Model human processor (MHP)
MHSS *see* Matlab Health and Socioeconomic Survey (MHSS)
MI *see* Motivational interviewing (MI); Myocardial infarction (MI)
MIA *see* Missing in action (MIA)
MIBQ *see* Mother–Infant Bonding Questionnaire (MIBQ)
MIC *see* Mean interaction contrast (MIC); Military–industrial complex (MIC)
Miccolis, Leila 13:862
Michelet, Jules 20:425, 22:889
  works
    Histoire de France 20:425
    Le Peuple 22:889
Michels, Robert 17:201, 18:473, 24:744
  elites 7:392
  works
    Political Parties 2:269
Michigan Survey Research Center (SRC) 5:830–831
Michigan voting model 25:302
Michigan's Institute for Social Research (ISR) model 25:733
Mickey Mouse laws 1:893
Mickey Mouse problem (religion) 20:391
Micro and macro contexts of a program implementation 7:230
Micro- and macro-social contexts of crime 5:190
Micro data, comparative 5:789
  obtaining 5:790–791
Micro level, rational choice theory extensions at 19:902–904
Micro level ID 12:196
Micro simulations, Hutterite fertility 11:446
Microaffordances 7:422
Microaggressions 3:889–890, 15:200
Microanalysis
  for measuring self-regulated learning 21:544
  of self-regulated learning 21:544
Micro-autonomy 21:586
Microbial disease 6:694
Microbiology, *versus* high-energy physics 7:874–875
Microbiome analysis 9:960–961
Microbiome revolution 1:763
Microbiomes 1:762–763
Microbiomics 2:625
Microcensus data 5:736
Microchauvinisms 14:415
Microcomputers 23:378–379
Microconsciousness theory 4:676–677
Microcontextual views, on health 22:278–279
Micro-control theories 4:836–837
Microcoordination
  mobile communication 15:631
  and telephone 24:157–158
Micro-credit loans to women 25:641
Microdados 7:181
Microdata
  census 3:284–289, 3:306
  cross-sectional micro-level demographic surveys 5:820
  cross-sectional micro-level social surveys 5:820–821
  instrumental variables, heterogeneity of treatment effects in 12:208
  longitudinal micro-level surveys 5:821
Microdemography 11:6–7
Microeconometrics, innovation and employment of 12:154–155, 12:157–158
Microeconomic databases 5:743–745
  survey of consumer attitudes 5:744–745
  survey of consumer finances 5:744
Microeconomic foundations, of structural unemployment 24:740–741
Microeconomic theory 22:195
Microeconomics 11:818–819, 19:576–577, 20:661–662, 20:728, 22:729
  focus of 7:22
  history of 15:331–337
    Arrow–Debreu–McKenzie model 15:334–335
    behavioral economics 15:334
    constrained optimization 15:334–336
    formalization 15:332–334
    game theory 15:335–336
    general equilibrium model 15:334–335
    general equilibrium theory 15:333–335
    invisible hand in Adam Smith 15:332
    marginalism 15:332–334
    market equilibrium 15:334–336
    market mechanism 15:331–332, 15:335

Microeconomics (*continued*)
partial equilibrium theory 15:333, 15:335–336
profit maximization 15:332–333
revealed preference approach 15:334
self-interest 15:332
state of the art and future perspectives 15:336
supply-and-demand 15:332–334
'two–sided' markets 15:336
utility 15:333–334
Microenterprise, informal sector 12:57
Microethical judgments, of negotiators 16:426
Microevolution 15:338–342
assortative mating 15:340
bioethics and population studies 15:341
culture 15:340–341
DNA research 15:338
evolution 15:338
fission–fusion model 15:339
gene–culture coevolution 15:340
genetic aspects 15:338, 15:340
genomes and bioethics 15:339
migration 15:340
morphology and sexual selection 15:340
mutation 15:338–339
natural selection 15:339
population structure 15:338–340
problems 15:341
race and racism 15:341
Xavante Amerindians 15:340
Microevolutionary theories 13:565
Microfinance 24:821
Microfossils 17:458
Microfoundations 16:765, 17:367
and causal mechanisms 14:877
challenges 16:766
contributions 16:765–766
of institutions 17:393–394
and macroeconomics 14:401–402
Microfoundations movement 15:331
Microgenesis
of culture 5:548
of self-efficacy beliefs 21:507–508
Microhistorians 11:111–112
Microhistory 1:708–709, 4:408–409, 11:105, 11:108–113, 19:256–257
and intercultural history 4:408–409
Microlevel conflict theories 22:957
Microlevel correlates of crime 12:464–465
Microlevel opportunity 10:19
Microlinguistics 14:228–229
Micro–macro 21:531
Micro-macro integration 4:212
Micro-macro link 7:434–435
definition of situation 6:21–22
Micro–macro taxonomy 15:331
Micro/macro transition 15:312–313
Micromethods 12:648–651
crisis of interpretation 12:650–651
action research 12:651
legitimation crisis 12:650–651
representational crisis 12:650
resistances to qualitative studies 12:651
field of interpretive research and its history 12:648–650
inquiry as site of multiple methodological research practices 12:649–650
reading history 12:649
Micromodels, family and household demography 8:727–728
Micromorphology, in geoarchaeology 10:53, 10:55–57
Micromotives 13:588
Micronesia 17:446–447
as cultural areas 17:158–159
in global age 17:162–163
island types 17:159
see also Oceania
Micropolitan areas 24:857
Micropolitical literacy 21:135
Micropolitics 25:659
criticism, and areas for future development 21:138–139
in 1990s 21:138–139
elements 21:134–136
research 21:136–138
methods and strategies 21:136
topics 21:136–138
schools as organizations 21:137–138
Micropractice, social work with refugees 22:770–771
Microprudential supervision 9:195
Microreductionism approach 12:661
MicroRNAs (miRNAs) 7:841
inheritance of modifications 7:842
linking environment to final functional protein levels 7:841
synthesis and actions 7:841f
Microscale, ethnic conflict at 8:93
Microscale analysis
household archaeology 11:220
Microscale geographic models 14:786
Microscopic neuropathology, of transient epileptic amnesia 24:557
Microscopic observation 11:108
Microsimulation 8:727–728, 22:701–702
in demographic research 15:343–346
diffusion 6:415
family and household demography 8:727–728
Microsimulation models 6:166–167
Micro-simulations 8:315
Microskills, in communication 4:288
Microsocial order 8:486
Microsociology 14:414–421, 22:994
agent-based modeling 14:418–419
analytical sociology 14:417–419
balance 14:414–415
contributions of Collins 4:211, 4:213, 4:215–216
contributions of Hochschild 11:147–148
critical realism 14:416, 14:419
emergenist sociology 14:414, 14:416
extremisms 14:415
ideology in 11:575–576
layered and 'flat' social ontologies 14:416
levels of analysis 14:414–419
levels of reality 14:416–417
linkage 14:415–416
mechanism-based thinking 14:417–419
rationality 14:418
scale-based thinking 14:417
social interactions 14:415–416
social network analysis 14:418–419
structural methodological individualism 14:416–418
theory of action 14:418
Microstoria 11:42
movement 11:42
Microstoria analysis 16:196
Microstructure of cognition 4:95–96
distributed representations 4:96
knowledge and learning in strengths of connections 4:96
patterns of activity arising in ensembles of simple elements 4:95–96
Microsystem 21:682–684
ecological systems theory 6:840
families as educational settings 8:732, 8:735–736
and health 6:880
Microsystem effects in richer *versus* poorer countries
student achievement 5:345
Micro-theory, of organizational behavior 17:697–698
Microtubules 1:591
Mid *p*-value 21:960
Midas touch effect 24:16
Midcity Project 9:596
Middens 8:163–164
Middle adulthood, development in 15:369
environmental changes 15:370–372
community 15:370–371
family 15:370–371
relationships 15:370–371
work 15:370–371
future directions 15:374
individual, changes in 15:372–374
cognitive abilities 15:372
emotional abilities 15:372–373
health and physical changes 15:372
motivations 15:373
personality development 15:373–374
theoretical perspectives 15:369–370
adaptation function 15:370
maintenance function 15:370
variability in 15:370
Middle Ages 12:277, 15:347–353
Catholicism during 3:234
chronology 15:348
crisis 15:352
deliberations 6:66
educational institutions in 7:233–234
extensions and limits 15:348
family 15:349–350
feudalism 15:348–349
feudalism in 9:111–115
gender history 15:349, 15:351
genocide, history of 10:10
history of religion 15:347, 15:349–350
kinship 15:349

metaphysics of 18:32
migration 15:448–449
notion of 15:347–348
periodization 15:348
psychohistory 15:350
psychology in 19:408–409
role obligations in 13:846
structures 15:349–350
time 15:348
warfare in 25:415–416
organization 25:415–416
purposes 25:416
technology 25:415
wisdom and philosophy in 25:599–600
Middle category 23:764
Middle childhood, and media 15:2–3
Middle class omnivores 13:853
Middle classes 2:784–789
of Australia
and environmental crisis 2:243–244
and political economy 2:242–243
contraction and expansion 2:787–788
current research 2:788–789
definitions 2:784–786
formations of 2:785
functional equivalents 2:788
genesis of (eighteenth to nineteenth centuries) 2:786–787
historiography 2:788–789
history of 2:786–788
inclusive middle classes 2:785–786
and income inequality 11:740
limits of concept 2:788
and Marxian theory 2:785
social formations in 2:784
and social mobility 22:427–428
and urbanization, in China 24:977–978
white-collar workers 25:566–567
Middle East 9:73–79
Arab nationalism in 16:271, 16:273
gender in 9:781–782
Islam in 12:731–736, 17:423
Jewish Diaspora 12:802–806
Middle East and North Africa (MENA) 15:354–359, 16:798, 21:791, 21:793
culture in 16:361
Arab 16:361–366
art 16:361–362, 16:364–365
exile 16:365
feminism 16:362, 16:365
film 16:361, 16:364–366
independence struggles 16:363
intellectuals 16:361–366
Iran 16:361–365
Islam and revolution 16:365
Israel 16:361, 16:363, 16:365
literature 16:361–365
modernity 16:361
moral authority 16:361, 16:364
new public sphere 16:362–363
nineteenth-century Nahda 16:361
orientalism 16:361
political commitment and prison 16:364
question of language 16:361–362
theater 16:361, 16:364
translation and recognition 16:365–366
Turkey 16:361–362, 16:365–366
veil 16:362, 16:365
war stories 16:363–364
woman 16:362–365
Facebook penetration 16:800t
fertility transition in 9:73–79
gender 15:356
identities and differences 15:357–358
nation 15:355–356
religion 15:356–357
turning points in scholarship 15:354–355
Middle East and North Africa (MENA), education in 7:210
educational quality 7:211–213
minority languages 7:213–214
quantitative expansion 7:211
religious education 7:213–214
structure 7:210–211
youth employment 7:213–214
Middle East and North Africa (MENA), legal traditions 15:360–368
Algeria 15:362–363
Arab 15:360
Bahrain 15:363
brief legal histories and summaries of select states 15:362–366
civil law 15:361
colonialism, codification, and modernity 15:361–362
colonizing family law, liberalizing society 15:362
colonizing property law, liberalizing the economy 15:361–362
Islamicizing constitutions 15:362
constitutional law 15:362
criminal law 15:362
Egypt 15:363
imperialism and legal reform 15:360–361
Iran 15:363–364
Iraq 15:364
Islamic history 15:360
Islamic law 15:360, 15:362–363, 15:365–367
Jordan 15:364
Kuwait 15:364
Lebanon 15:364
Libya 15:364–365
modern nation-states 15:360–361
Morocco 15:365
Oman 15:365
Ottoman empire 15:361
postcolonial states 15:361
pre-Islamic history 15:360
Qatar 15:365
Saudi Arabia 15:365
secularization 15:361, 15:366–367
Sudan 15:365–366
Syria 15:365
Tunisia 15:366
Turkey 15:366
United Arab Emirates 15:366
Yemen 15:366
Middle frontal gyrus 16:701
Middle ground 8:193, 8:196
Middle Kingdom 16:275–276
Middle passage (slavery) 22:38
Middle period
Andes 23:38–39
China and East Asia legal orders during 6:779–780
Middle range theories (MRT) 1:670–671, 1:892, 1:901–902, 14:262, 15:247–248, 20:686, 20:690, 22:993, 22:997
social mechanisms 22:418
specialization and 23:227
Middle temporal (MT) area 25:202
Middle temporal gyrus (MTG) 13:370
Middle-class movements 2:49
Middle-class status 23:495
Middle-class youth movement 5:73–74, 5:76, 5:78
Middle-heavy age distributions 18:572
Middle-income countries, deinstitutionalization psychiatry in 6:42–46
Middleman minorities 15:573
Middle-range approaches, television 25:84
Middle-range research 11:219
in ethnoarchaeology 8:162
Middleton, John 13:71
Midlife
menopause, infertility, and miscarriage 21:843
sexualities in 21:841
sexuality in 21:844
Midlife crises 1:318, 14:77
Midlife in the United States (MIDUS) 14:513
Midlife psychological development 15:369–375
MIDM *see* Mutual intergroup differentiation model (MIDM)
Midrange quantifiers 19:708
Midterm memory (MTM) 15:157–158
MIDUS *see* Midlife in the United States (MIDUS)
Miedema, F., works, *Science* 3.0 21:254
Miège, Jean-Louis 11:688
Migraine, and transient global amnesia 24:566
Migraine aura 10:501
Migrant communities 24:589
Migrant foods 9:292
Migrant networks, South and Southeast Asians 12:534–536
Migrant rights 10:171–172
Migrant sovereignty 10:310
Migrants 3:638–639, 25:125
in China 24:973–974, 24:976–978
in Europe 15:376–388, 16:30–31
and social work 15:416–421
Migration Effectiveness Index (MEI) 12:435
Migration industry 12:534
Migration theory 6:402–403
material improvement 15:452
social connection 15:453
symbolic gratification 15:453

Migration to/from Latin America
early migration patterns 15:462–463
Mexico–US migration 15:463–464
non-Mexican migration 15:464–465
Migrations 5:474, 11:6, 11:8, 11:648–650, 11:654, 15:452–456, 15:636, 18:549–551, 19:825, 19:828, 20:105, 18:551t
of Africa 1:292–293
African 1:287
anthropology of 15:422–426
cities, hyperdiversity, and multiculturalism 15:424–425
diaspora 15:423
gender 15:425–426
governmentality 15:424
politics of citizenship and belonging 15:423–424
transnational dimensions 15:422–423
transnationalism 15:422–425
urban anthropology 15:422, 15:424
in archaeology 15:472–482
history 15:472
human dispersals 15:472, 15:477–479
neolithic-population movements 15:477–479
prehistory 15:473
in Australia 2:240–242
and biographies 2:639–640
and brokers 2:870, 2:873
and capitalism 11:648–649
in Caribbean 1:638–639, 1:641–642
in Central America 1:622–623
in Central Asia 3:316–320
in China 2:65–69
circular 11:649
civil violence 15:454–455
classical approaches 15:446–447
classification 12:433–434
climate change 15:454
coerced 11:649
cultural aspects 15:427–432
immigration, language, and schooling 15:429
immigration and family 15:428–429
immigration and inequality 15:429–431
definition 12:443
and demography 6:121–122
diaspora 6:354–355
and disasters 6:105
domestic 11:649, 11:651
in Europe 8:142–147, 8:274, 8:277–278, 15:376–388
assimilation 8:143, 8:145–146
discrimination 8:143–146
ethnic diversity and inequality 8:143
majority attitudes and intergroup threat 8:143–144
media representations and contact experiences 8:144
migration and integration policies 8:143
minority acculturation 8:145
minority adjustment 8:145–146
multiculturalism 8:142–145
prejudice 8:142–144
stereotype threat 8:144–146
and fertility 9:5
forced 20:106
free 11:649
frontiers 9:468
gendered 11:649–650
Genographic Project 10:22
global migration 15:427–429
global revolutions 11:648–649
and health 15:407–410
acculturation 15:408
chronic conditions 15:407
destination 15:408–409
historical development of the field 15:407
mental health 15:407–408
mortality 15:407
physical health 15:408
predeparture 15:407
psychological distress 15:407–409
return migration 15:409
travel 15:407–408
historiographical approaches 15:447
history 15:446–451
ascription 15:450
asylum seeker 15:446, 15:450
emigrant 15:446, 15:448
European dimensions 15:448–450
European expansion 15:447–448
global dimensions 15:447–448
historiography 15:447
immigrant 15:446–448
indentured labour 15:447–448
macro-cosmos 15:446
meso-historical approach 15:446
motivations, patterns, and typologies 15:446
research tasks 15:447
seasonal migration 15:449
settler migrations 15:447
spatial dimensions and research concepts 15:446–447
transatlantic African slave trade 15:447
of honor 11:183
immigrant assimilation 15:429, 15:431
and inequality 11:898
international 11:649, 12:511–518
international labor 17:678
and international social problems 12:544
international system of nation-states 11:648–650
and language families 13:337, 13:340–341
to and from Latin America 15:462–465
mass 11:648
modern, features of 11:649–650
in the MENA region 15:358
in Mexico 1:630–631
microevolution 15:340
minority ethnic elders 15:440–445
minority-group status 15:570
in multisited field studies 16:103–106
and nationalism 11:650
in North America 17:15, 17:18–19
in Oceania 17:159–160, 17:162–163
out of Europe 15:457–461
population register for 5:736
population stability with 18:585
poverty 15:429–431
and punishment 19:659
queer geography 19:760, 19:762
and race 11:651
racialization 15:427, 15:430–431
and religion 3:577
risk management 15:452–453
of Roma 20:754–755
rural sociology 20:817–818
rural–urban 24:962–963
sample selection by 6:141
scientific 25:541
seasonal 11:649
segregation 15:427, 15:430–431
and selectivity 11:650–651
and sex work 19:282
social connection 15:453
social mirroring 15:430–431
and social work 15:416–421
challenges to interventions 15:420
direct services 15:419
diverse communities 15:419–420
indirect services 15:419
integration 15:418, 15:420
integration policies 15:418
international migration trends 15:416–419
refugees 15:416–420
settlement 15:417–420
underutilization of service 15:419
and spatial interaction 23:153–154
theory 15:466–471
cumulative causation 15:469
emigration 15:470
immigration 15:469–470
international migration 15:469
neoclassical economics 15:466
new economics of migration 15:466–467
segmented labor-market theory 15:467–468
social-capital theory 15:469
state filter 15:469
theoretical synthesis 15:470
undocumented migration 15:469
world-systems theory 15:468–469
transnational 20:108
transnationalism 15:428, 15:431
urban 24:775–780
and urbanization 11:649
and urbanization in Africa 24:965–966, 24:969–972
xenophobia 25:803
Migrations, economics of 15:433–439
circular 18:598
cohort effects 15:437
competitive equilibrium 15:433
economic efficiency 15:433–434
gravity model 15:435
human 18:597–598
human capital approach 15:435–437
immigration surplus 15:438
internal migration 15:435–436

age 15:435
distance 15:435
education 15:435
family migration 15:436
other variables 15:435–436
return and repeat migration 15:436
international 18:598
international migration 15:436–438
assimilation and cohort effects 15:436–437
gains to host country 15:438
global gains from open borders 15:438
labor market impact 15:437–438
model 15:434–435
net gain 15:434
self-selection 15:435
wage differential 15:433, 15:435
Migration-trust networks 15:412
activation of 15:414
Miki, Nakayama 20:370
Mikloukho-Maclay, Nikolai 17:161–162
Milan systemic therapy 19:505
Milbank Memorial Fund (MMF) 17:39–41
Mild cognitive impairment (MCI) 1:591–592
Mild dementia in old age 6:78
Milgram, Stanley 2:273, 4:154–155, 4:256, 13:901
experiments 13:901
obedience 17:67–68
on small world structures 5:467–468
Milgram Experiment 8:57–58
Milgram's lost-letter paradigm 9:130–131
Milgram's obedience studies 17:61–66
Milieu therapy 24:275
offender rehabilitation 20:194
Milieus, social stratification 22:342–343
Militancy in Asian Christianity 3:531
Militant joking 11:417
Militants
anthropology 2:873
groups 22:432
nationalists 16:267
secular regimes 15:725–726
Militaries 15:518–523, 17:620, 25:371
action in response to terrorism 13:533–534
activities 15:502–503
adjunct 6:369
aggression and racial annihilation 16:259
anthropology of 15:483
bases 15:486–487
counterinsurgency 15:486
health and technology 15:486
lived experience 15:484
militarization, and demilitarization 15:487
soldiers 15:483–487
trauma 15:483, 15:486
veterans 15:485–486
violence 15:483–485, 15:487
war 15:483–487
buildup 15:491
child soldiers 3:430–438
civil–military relations, emergence of 15:519
conflict, changing nature of 15:513
conflicts 25:372
conscription 15:519–522
definition 15:484–485
dictatorships 6:368
economic geography 15:503–504
elite group 15:536
geography 15:501, 25:371
definition of 15:501
economy of 15:503–504
future directions in 15:504
history of 15:501–502
personnel 15:504
scope of 15:502–504
history 15:506–510
academic neglect of 15:507–508
arms 15:506
didactic tradition 15:508–509
emergence as separate history 15:506–507
Enlightenment 15:506–507
general staffs and official histories 15:508
influence of professional armies 15:506
modernity 15:507–509
nations 15:509
new 15:509
politics 15:506–509
popular 15:510
professionalization 15:506
rediscovery of Clausewitz 15:509
states 15:507
in twentieth century 15:507, 15:509
war as science 15:506–507
World War I 15:506–509
World War II 15:506–510
interventions, and peace movements 17:645
landscapes 15:503
reading and engaging with 15:503
mental health care 15:498–499
mercenaries 15:518–519
paradigms 24:143
patronage in research 24:764–765
personnel 15:504
post-traumatic stress disorder 18:703
professionalism 15:512
effects of military service on 15:515–516
regime, and child soldiers 3:439–444
relations on social relations, influence of 15:492
research 24:142–143
and rise of state 15:518–519
service, and same-sex sexualities 20:872
sociology 15:511–517
as academic discipline 15:515–516
changing demographics in force 15:513–514
changing nature of military conflict 15:513
civil–military relations and citizenship 15:512
conscription 15:512–516
effects of service on service members and families 15:515–516
foundations of interdisciplinary and applied approaches 15:511
military professionalism 15:512
motivation, cohesion, and effectiveness in combat 15:512
organizational changes 15:513
race, gender, and sexual orientation 15:514
reserve forces 15:514–515
Shils on 21:886
spending, within neoclassical mainstream 6:3–4
-style weapons 10:457
tactics 15:502
technology 4:225
arms transfers 24:143
transfer of 24:142–143
victory of democratic control over 15:520–521
and war 25:421–423, 25:425–426
Militarism 11:392–393, 15:490–494, 15:501, 15:504, 15:519–520, 15:724, 25:361, 25:371–374
academic study of 15:492
chiefdoms in 3:380–381
as excessive influence 15:491–492
as ideology 15:490–491
as the influence of military relations on social relations 15:492
inward-oriented 15:493
Japanese 15:492–493
landscapes of 15:503
Mann on 14:491
as military buildup 15:491
modernity and 15:492–493
outward-oriented 15:493
Prusso-German 15:492–493
scale of analysis 15:493
as use of force to resolve conflict 15:491
Militarization 15:487, 15:501
of human rights abuses 10:636–637
and urban violence 24:953–957
Militarized ethnography 8:182
Military geography 15:501–505
Military imaginary 15:484
military industrial complex 22:390
Military junta 22:139
Military operations in urban terrain (MOUT) 24:954–955
Military psychiatrists 6:48
Military psychiatry 15:495–500
cohesion 15:498–499
help-seeking 15:499
leadership 15:496, 15:498–499
management and care delivery 15:498–499
military mental health care 15:498–499
organizational management of traumatic stress 15:499
mental health consequences of operations 15:496–497
battle fatigue and combat stress 15:497
medically unexplained physical symptoms 15:497
psychiatric illnesses 15:497
possible technological advances 15:499

Military psychiatry (*continued*)
posttraumatic stress disorder 15:495–500
resilience 15:495–500
risk and protective factors 15:497–498
screening 15:498
social support 15:496, 15:498
war environment 15:496
Military schools 7:299
Military–geostrategic balance, in Cold War 4:130
Military–industrial complex (MIC) 15:490, 25:379
Mill, John Stuart 3:774, 7:95, 7:113, 7:567–568, 9:390, 9:414, 11:11–12, 11:885, 13:437–438, 14:19, 14:621–622, 15:524–528, 18:625, 19:410, 19:679, 20:663, 22:697, 25:75–77
on democracy 15:524–525, 15:527–528
democratic theory of 6:91
on economic methodology 7:99–102
on Enlightenment 15:524–525
on freedom 15:524
on idealization 11:516–518
on individualism 15:524, 15:526–527
on introspection 12:702
on labor 25:676
on liberalism 15:526–527
on political economy 15:525–526, 15:528
on utilitarian social thought 24:983–984
on utilitarianism 15:524–527
works
Autobiography 15:524, 15:527–528, 22:828
Civilisation 15:524
Considerations on Representative Government 15:527
Essays on Some Unsettled Questions in Political Economy 15:525–526
Logic of the Moral Sciences 13:112
On Liberty 9:390, 11:818, 15:526–528, 18:621–622, 25:75–76
On Representative Government 15:526–527
On the Logic of the Moral Sciences 15:525
Principles of Political Economy 15:525–528
System of Logic 18:621–622
Utilitarianism 8:213–214, 15:526–527
Millenarian movements 3:140, 3:143, 3:537
Millenarianism 15:529
Millenarism 15:529
Millennarian Christians 20:404
Millennial capitalism 14:437–438
Millennial Declaration of the United Nations 23:803
Millennial generation 3:542
Millennial movements 19:198
Millennial themes in East Asia 20:369–370
Millennialism 15:529–532
Christian 15:531–532
in Christian eschatology 15:531
communal purity and 15:531
demise of catholic apostolics 15:531
development in early and medieval Christianity 15:529
Hinduism and 15:531
historical provenance 15:529
introversionist position 15:531
modern millennial movements 15:530
in other traditions 15:531
Plymouth (Exclusive) Brethren response to hope 15:531
pre- and postmillennialism 15:529–530
prospect of the failure of prophecy 15:530–531
responses to failure of prophecy 15:530–531
Revolutionist movements and 15:532
setting dates 15:530
social constituency of 15:530
synonyms of 15:529
tribal peoples as millennialist 15:531
use of term 15:529
Millennium Development Goals (MDGs) 1:295, 4:632, 5:783, 6:304, 6:310, 7:919–920, 7:922, 8:939, 12:21, 12:498, 12:547–548, 13:520, 18:724, 23:797, 24:303, 24:474–476, 24:751, 24:848, 25:431, 12:548t
in Africa 1:295–298, 1:297f
and human rights 11:357
MDG3, tourism policy and discourse 24:476–479
urbanization 24:962
Millennium Ecosystem Assessment (MEA) 7:813, 12:717, 23:802
Miller, Daniel 1:641
works, *Material Culture and Mass Consumption* 7:16
Miller, David, works, *The Blackwell Encyclopedia of Political Thought* 7:591
Miller, Edward 3:890
Miller, George A. 19:415
Miller, James C. 13:566
Miller, Matthew 23:687
Miller, Neal 12:11–12
Miller, Warren E. 5:831
Millet system (Ottoman) 16:384
Millian methods 4:443
Millikan, Max 1:952
Millikan, R.G., functional explanation 9:521–522
Milliken, Frances J., on networks and creativity 5:467
Mills, Charles Wright 4:363, 9:540, 13:638–639, 15:533–538, 20:670, 25:463, 25:465
on bureaucracy 15:533, 15:535–536
on class structure 15:534–536
on elites 7:393–394
on freedom and reason 15:538
on intellectual craftsmanship 15:537
major works of 15:534
personal life of 15:533–534
on politics of responsibility 15:536–537
on politics of truth 15:536–537
on power elite 15:535–536
on power system 15:534–536
on pragmatism 15:533–534, 15:538
on psychology of institutions 15:534
on public intellectuals 15:537
on social psychology 15:534
on social stratification 15:534
on sociological imagination 15:537–538
on sociology of persons 15:534
on white-collar workers 15:535
works
The Power Elite 9:540, 14:745
White Collar 2:916, 14:745
Mills v. Board of Education of the District of Columbia 23:220–221
Mills–Muth model 11:240
Mill–Stephen conflict 25:75–76
Milnacipran 1:779
Milner, Brenda 10:732
Milo, Yuval, on Black-Scholes model for pricing 5:540
Milton, John 15:12
Milwaukee Parental Choice Program 21:72, 21:75
Mimesis 18:31
and human cognition 11:256
neonatal 16:482
Mimetic desire 18:34–35
Mimetic ethnographies, spirit possession 23:283
Mimetic isomorphism 13:545–546, 17:391, 21:73, 22:981
MIMIC *see* Multiple Indicator and Multiple Causes (MIMIC) approach
MIMIC method, for shadow work measurement 21:863
Mimicry 10:117–118
entertainment 7:670
Mimouni, Rachid, works, *Le fleuve détourné* 21:752
Mind 2:880–885, 4:152, 16:326–331, 21:341–342, 23:281
and body 2:720–721, 2:733
Bruner on 2:875–876
cognitive archaeology 4:23–29
computer models 12:704
conceptions of 22:761
in classical Freudian theory 22:761
in ego psychology 22:761
in Lacanian theory 22:762
in object relations theories 22:761–762
in relational and intersubjective theories 22:763
in self psychology 22:762
conscious awareness as essential attribute of 24:722
developing architecture of 16:505
and *Geist* 9:666–667, 9:669
heuristics as descriptive models of 10:830–831
hierarchical organization of 4:515–516
phenomenology 18:19–20
primitive 1:767
Putnam on 19:663–664
Searle on 21:341–343
and worldview 25:753–754
Mind and body differences 6:459
Mind frames 4:23–28

Mind reading, Thorndike on 24:305
Mind sharing, and intentionality 12:348
Mind the Gap project 22:922
Mind-altering substances 17:236
Mind–body dualism 15:545–548
  dualism of domains 15:546–547
    public and private cognition 15:546–547
  dualism of substances 15:545–546
    Cartesianism, early criticisms and revisions of 15:546
    proposal of Descartes 15:545–546
    recent criticisms 15:546
  materialism 15:547
  mental attributes 15:545–546
  person 15:545–548
  physicalism 15:548
  recent proposals of alternative distinctions 15:547–548
    dualism of Turing 15:548
    hybrid psychology 15:547
    supervenience 15:547–548
  soul 15:545–546, 15:548
  Wittgenstein on 15:547
Mind–body identity 15:172, 15:175
Mind–body problem 16:327–328, 19:409–410
  philosophy 18:32
Mind–body relations, psychosomatic disorders 19:484–485, 19:487–488
Mind/body split 21:190
Mindful abstraction, in transfer in school learning 21:97–98
Mindfulness 15:549–553
  automaticity 15:549–550
  -based interventions 15:551–552
  and behavior therapy 2:431, 2:433–434
  chronic pain 15:551–552
  cognitive reactivity 15:550–551
  mechanisms 15:550–551
  mood disorders 15:550, 15:552
  and physical health 15:551
  practices, and stress management 15:552–553
  relapse prevention in mental health 15:550–551
Mindfulness-based cognitive therapy (MBCT) 4:30, 4:544, 15:550–551
  for depression 19:375–376
Mindfulness-based educational programs 4:671–672
  goals and composition 4:671
Mindfulness-based stress reduction (MBSR) 15:551, 23:553
Minding the Baby (MTB) 12:233
Mindliness, in mathematics learning 4:67–68, 4:68t
Mindreading 15:539
Mind-set 12:381–382
Miner, H., on craft production 5:109
Mineral resources, limits to 15:554–560
Mineralocorticoid receptors (MR) 11:476–477, 23:577–578
Minerals Management Service (MMS) 17:373
MINERVA (memory model) 15:162–164
Mineshaft 13:887
Ming 3:519
Mingei 5:111
Mingione, Enzo 12:913
Minh, Hô Chì 25:96
Minhha, Trinh T.
  works
    Reassemblage 6:621
    Surname Viet Given Name Nam 6:623
MINI *see* McGill Illness Narrative Interview (MINI)
Mini computers 23:378
Mini Mental Status Exam (MMSE) 16:723
Mini trials 15:70
Miniaturization, and interviewing 12:674–675
Minimal cognitive impairment (MCI) 19:151
Minimal distance principle (MDP) 4:829
Minimal edit distance problem 8:321
Minimal group studies, tyranny 24:713
MINimal RESiduals *see* MINRES (MINimal RESiduals) procedure
Minimal state libertarianism (Nozick) 14:41–42
Minimal sufficiency, statistics 23:419–420
Minimal-contribution lower bound 4:873
Minimalist analysis 1:676
Minimalist Program 9:876–878, 16:5
Minimalist Syntax 10:317–318
Minimalist theories of truth 8:72
Minimalist theory of inferences 12:40–41, 19:395
Minimally punishing democratic state 19:636
  Blair revolution 19:653–654
  Netherlands 19:653
  US penal revolutions 19:654
  White March in Belgium (1996) 19:654
Minimax designs 17:260
Minimax regret 5:883–884
Minimax search 2:33, 2:33f
Minimax theorem 9:576, 16:583
Minimum chi-square estimation 15:294–295
Minimum description length (MDL) 4:475–476
  model selection following 15:657
Minimum detectable effect (MDE) 22:296
Minimum distance approach 1:66
Minimum efficient scale (MES) 17:366
Minimum norm quadratic unbiased estimation (MINQUE) 14:148
Minimum or single linkage 23:393
Minimum probability of error case 23:412
Minimum probability of error decision 23:413
Minimum qualifications 15:597–598
Minimum sentence of life imprisonment 12:961
Minimum spanning tree 23:393
Minimum standards, business codes 8:22
Minimum variance unbiased estimates (MVUEs) 14:145, 14:148
Minimum wage laws 24:740
Minimum wage legislations and informal employment 12:63
Minimum Wage Study Commission 15:562–563
Minimum wages 15:561–565, 25:339
  characteristics 15:561
  economic effects 15:561–563
    empirical evidence 15:562–563
    theoretical considerations 15:562
  effects 15:564
    on poverty and distribution of income 15:563–564
  labor demand 15:562
  labor markets 15:561–564
  textbook treatment of 15:562f
Ministering to others 19:233–234
Ministers, executive government 8:497–501
Ministry for Children and Family Development (MCFD) 11:765
Ministry for Family and Social Policies (Turkey) 20:872–873
Ministry of Commerce (MOFCOM) (China) 3:505
Ministry of Culture (China) 5:597
  Division of Cultural Industry 5:597
Ministry of Economy, Trade and Industry (METI) 5:457
Ministry of Environmental Protection (MEP) 6:833
Ministry of Finance (MOF) 3:505
Ministry of Foreign Affairs (MOFA) 3:505
Ministry of Housing and Urban-Rural Development (MOHURD) 6:833
Ministry of State for Antiquities (MSA) 1:691–692
MINITAB (statistical analysis package) 23:378
Minjung 2:73
Minnesota Longitudinal Study of Risk and Adaptation from Birth to Adulthood 2:793
Minnesota Multiphasic Personality Inventory, adolescent form (MMPI-A) 15:566
Minnesota Multiphasic Personality Inventory-2 (MMPI-2) 6:344, 9:1010, 16:723, 16:725–726
  adolescent version 16:725–726
  applications of 15:567–568
Minnesota Multiphasic Personality Inventory (MMPI) 15:566–568, 17:827
  child custody 15:566
  evaluating cooperative responding 15:566–567
    cannot say score 15:566
    infrequency scales 15:567
    K scale 15:566–567
    L scale 15:566
    TRIN and VRIN scales 15:567
  historical background 3:879
  malingering 15:567
  personnel 15:566–567
  profile interpretation 15:567
    content-based scales 15:567

Minnesota Multiphasic Personality Inventory (MMPI) (*continued*)
MMPI-2 clinical scales 15:567
special-problem or supplemental scales 15:567
validity 15:566, 15:568
Minnesota Population Center 3:288
Minnesota Satisfaction Questionnaire (MSQ) 12:833–834
Minor civil divisions (MCDs) 3:580, 24:854
Minorities 15:569–574, 19:813, 21:42–44, 21:46–47
acculturation 8:145
adjustment 8:145–146
broadening use of 15:573
characteristics 15:569–570
in China 2:66–67
consequences of minority-group status 15:571–573
assimilation 15:572
expulsion 15:571
extermination 15:571
fusion 15:572
pluralism 15:572–573
secession 15:571
segregation 15:571–572
discrimination 15:569–570, 15:573
emergence of minority-group status 15:570–571
annexation 15:570
colonialism 15:571
migration 15:570
endogamy 15:570, 15:573
exogamy 15:570
gay 15:573
genocide 15:571
globalization 10:236
group rights 3:657
Holocaust 15:571
immigrant, identities among 11:676
land rights 13:234
languages 13:251–253
in legal profession 13:632
lesbian 15:573
majority 15:569–573
middleman minorities 15:573
and multiculturalism 16:23–25
and organizational diversity 4:738, 4:741–742, 6:595, 6:598–599
prejudice 15:569–570, 15:572–573
proportional representation 19:206, 19:208
religiosity and prejudice against 20:381–383
authoritarianism 20:381–382
church attendance 20:382–383
dimensions of religiosity 20:381–383
extrinsic religiosity 20:382–383
fundamentalism 20:381–383
Holocaust 20:381
intolerance 20:381–382
intrinsic religiosity 20:382–383
psychological studies 20:381–382
sociological studies 20:382
synthesis 20:382–383
religious 15:357
rights of *see* Rights of minorities
transgender 15:573
utility, as descriptor 15:573–574
women as 15:573
Minorities at Risk (MAR) project 5:793
Minoritization of crime 19:640
Minority Critiques of the Critical Legal Studies Movement 5:267
Minority culture, identification with 22:218
Minority ethnic children, child protection 3:420–421
Minority ethnic (ME) elders
current and future developments 15:443–444
discrimination 15:440–443
elder care services 15:442–443
mainstream services 15:442–443
quality 15:443–444, 15:441t
voluntary organizations 15:443
expectations of service-delivery quality in care services 15:441t
experience of poverty and ill health 15:440–442
encouraging take-up of services 15:442
health 15:441–442
housing 15:442
income 15:440–441
social care 15:440, 15:442–444
unmet needs 15:444
wider recognition of specific needs of 15:443
Minority ethnic prisoners 18:927
Minority group model of disability 6:473–474
Minority group threat, and race–crime relationship 19:841–842
Minority groups 5:151–156
ethnic 8:114, 8:116
influence of 2:170–171
Minority influence, in organizations 10:409
Minority languages 7:213–214, 16:68–70
Minority language/speaker 2:580–585
Minority model of disability 6:447–448
Minority nationalism 16:309
in China 2:66
Minority persecution, crowds 5:364
Minority rights 14:446–447
Minority rule 14:446–447
Minority shareholders 4:933
Minority status 21:43–44
and Jewish Diaspora 12:802
Minority stress 3:889–890
and LGBT aging 13:866
Minority veto 4:693
Minority World 6:839
Minority-group relations, genocide 10:20
MINQUE *see* Minimum norm quadratic unbiased estimation (MINQUE)
MINRES (MINimal RESiduals) procedure 4:914
MINRES/SVD approach 4:914–915
Minstrelsy of Medieval Europe 17:909
Mintz, Jerome, works, *The Anarchists of Casas Viejas* 1:683–684
Mintz, Sidney W. 1:769
on Caribbean 1:639, 1:641
works
Sweetness and Power 1:746–747
Minuchin, Salvador, structural family therapy 19:505
Minwalla, O., nonheterosexual British Asian friendship 16:865
Minzokugaku 12:794
Miocene Epoch 22:894
MIP *see* Mixed integer programming (MIP) solvers
MIPAA *see* Madrid International Plan of Action on Ageing (MIPAA)
mIPS *see* Medial intraparietal sulcus (mIPS)
Miracles, cultural logic of 15:575–581
creation of new knowledge 15:575–576
mobilizing power 15:576–577, 15:579–580
relationship between science and religion 15:575–576
sainthood
modern 15:577–579
before Protestant schism 15:577–578
special commission for investigating miracles and saints 15:578–579
Miraculous absolutism 1:10
Miranda v. Arizona 12:899
Miranda warnings 13:573
Mirandola, Pico della 24:443
miRNAs *see* MicroRNAs (miRNAs)
Mirowski, Philip 7:96, 7:101
Mirror effect, and recognition memory 20:40–41
Mirror image assumption 16:585
Mirror movements, and asymmetry of body/brain 2:115
Mirror neuron circuit, motor control 15:959, 15:962
Mirror neuron system (MNS) 7:412, 15:582
and language evolution 11:330–331
Mirror neurons 1:875, 5:25, 7:416–417, 11:622, 12:343, 13:311, 15:582–588, 16:739–740, 18:848, 22:403–404, 23:236
action understanding 15:583
autism 15:582–583, 15:585
Broca's area 15:583–585
embodiment 15:586–587
and embodiment cognition 7:421–422
empathy 15:585
imitation and cultural learning 15:586
and intentionality 12:348
and language 15:583–585
gestural origins of language 15:583–585
motor theory of speech perception 15:583
like me theory 15:133
neonatal imitation 16:483–484
pantomime 15:584
and pragmatism 16:739–740
and speech 16:740
theory of mind 15:585–586
Wernicke's area 15:585
Mirror reading 16:845
Mirror stage concept, of Lacan 13:211–213

Mirror star-tracing procedure 3:118
Mirror tracing 16:845
Mirroring hypothesis 15:718, 15:720–722
Mirskij, Eduard Michajlovic 21:131
MIRT *see* Multidimensional item response theory (MIRT)
Mirtazapine 1:779
Mirzoeff, Nicholas 25:212
Misassignment 20:153–154
Miscarriages, and sexuality 21:843
Miscarriages of justice 15:71
Miscegenation, Freyre on 9:423
Mische, Ann 16:565
  three-mode lattice analysis 5:560
Miscommunication 13:257
Misconceptions 21:275
  cardiac 20:188–189
Misconduct
  and objective research 17:92, 17:94
  and peer review 17:682
  in voluntary organizations 25:263
Misdemeanors 15:589–594
  consequences 15:593–594
  felonies 15:591
  infractions 15:591–592
  mala in se crimes 15:589
  mala prohibita crimes 15:589
  minimum standards 15:589
  moral crusades 15:590–591
  moral entrepreneurs 15:589–591
  morality offenses 15:589
  petty offenses 15:589, 15:591, 15:593
  types of 15:592–593
  unclassified 15:592
Misdirection effect 2:172
Mises, Ludwig von 24:745
Misinformation
  in eyewitness testimony
    cowitness misinformation 8:645
    effects of 8:644–645
Misinformation effect 8:644, 8:711
  false memories 8:711–713
Misinformation paradigm 20:53
Mismatch hypothesis 8:452
Mismatch in labor market 15:595–601
  causes of 15:598–599
  consequences of 15:599–600
  job search 15:595–596
  policy implications 15:600
  relationship of mismatch models to existing theories 15:597–598
  skill 15:595–597, 15:596t
Mismatch negativity (MMN) 3:329, 9:482–486, 9:483f
  language development 16:708–709
Misnourishment, and labor market outcomes 17:70–71
Misrecognition 11:552
  and fetishism 9:105–110
  and human sciences 11:384
Misrepresentation of data 17:94
Misrepresentation problem (intentionality) 12:334
Missile crisis 5:369, 5:373
Missing at random (MAR) 15:603
Missing completely at random (MCAR) 15:603
Missing data 13:389, 15:602–607, 22:551
  adjustments 6:128
  Bayesian simulation methods 15:606
  complete-case and weighting methods 15:603
  generalized linear mixed models 9:851
  maximum likelihood
    for ignorable models 15:605
    for nonignorable models 15:605–606
  multiple imputation 15:604–605
  national health survey, multiple imputation for 16:91
  pattern and mechanism 15:602–603
  response data, compliance 4:451
  single imputation 15:604
Missing girls 15:608–613
  demographic records 15:608–609
Missing heritability 2:453, 8:717, 8:721–722, 16:499
Missing in action (MIA) 9:335–336
Missing value displays 10:355–356
Missing values 24:359
Missing women 15:611–612, 15:612t
Mission (Christianity) 15:614–617
Mission creep 8:51–52
Mission du patrimoine ethnologique (Mission for the ethnological heritage) 1:755
Mission hate crimes 10:542
Mission within the Christian tradition
  Christian foreign missionary 15:615
  colonialist attitudes among late Victorian missionaries 15:615
  concept 15:614
  and conversion to Christianity 15:615–616
  development of a vernacular literature 15:615
  historiography 15:615
  history of 15:614–615
  impact on Latin America 15:616
  modern missionary movement 15:614–615
  plural term missions 15:614
  regional studies 15:615–616
  Roman Catholic missionary 15:614–615
  sociological and gender studies 15:616
  in theological discourse 15:616–617
  women in 15:616
Missionaries 24:602–604
  accounts of Biblical prophets 16:169
  and African studies 1:277–278
  cannibalism 3:83–84
  influence in South Africa 1:289
Missionary activities of St Cyril and St Methodios 17:437
Missionary practice, syncretism 23:881–882
Missionary religions 2:886
Mission-driven companies 4:746
Missionization, in Melanesia 17:156
Mississippi Freedom Summer Project 3:693
Missouri v. Frye (2012) 18:229
Mistake of judgment 3:28
Mistakes, in organizations 17:422
Misthoforia 6:62
Mistreatment of elders 7:327
Mistress 17:488–489
Mistrust 1:544
  future research 1:549
  neighborhood disorder 1:548–549
  structural theory 1:548–549
MIT *see* Massachusetts Institute of Technology (MIT)
Mitchell, Arnold, works, *The Nine American Lifestyles* 1:840
Mitchell, Don, works, *The Lie of the Land: Migrant Workers and the California Landscape* 11:19–20
Mitchell, J. Clyde 1:768–769
  on meaning 16:562
Mitchell, Timothy 25:210–211
Mitchell, Wesley 1:574
Mitchell, W.J.T. 25:191
Mithraism 22:998
Mitigation, and natural disasters 16:317–318
Mitigation costs
  climate change 3:860–862
Mitigation targets, framing 3:863–864
Mitochondrial DNA (mtDNA) 9:942–943
  Genographic Project 10:23, 10:25–27
  profiling of 9:946t
Mitochondrial DNA (mtDNA) sequences 18:907
Mitochondrial DNA (mtDNA) testing 9:899
Mitochondrial transfer 10:38
Mitogen-activated protein kinase (MAPKs) 12:711–712
Mitsuminesama 4:202–203
Mitteis, Heinrich, contribution in feudalism 9:112
Mitteleuropa 6:811–812, 8:274
Mittelstand 3:24
Mitzvot (religious commandments) 12:872
MIVILUDES *see* Interministerial Mission for Monitoring and Combatting Cultic Deviances (MIVILUDES)
Mixed corpus 4:950
Mixed courts, in Europe 5:228–229
Mixed economy 3:97–98, 13:199, 16:771, 25:532
  competition in 16:771
Mixed effects models, demographic analysis 17:731, 17:733
Mixed episodes of mania, management of 14:487
Mixed explanations 8:594
Mixed functionalism/formalism 9:472
Mixed Gaussian distribution 24:328
Mixed integer programming (MIP) solvers 4:511
Mixed land use measures 24:825–826
Mixed legal system 13:619
Mixed methodology 15:618
  analysis strategies in 15:621
  challenges (in research) 15:622
  characteristics of 15:619–620
  data collection in 15:621
  designs 15:620–621
    conversion 15:620–621

Mixed methodology (*continued*)
integrated 15:621
parallel 15:620
sequential 15:620
movement 15:618
opportunities (in research) 15:622
postanalysis integration 15:621–622
research questions in 15:620–621
set of beliefs 15:618–619
Mixed methods 5:313, 15:618–623
in qualitative research 19:688
Mixed methods research 15:618
emergent (research) questions in 15:620–621
integration of findings in 15:621–622
preplanned (research) questions in 15:620
quality of inferences in 15:621–622
research question in 15:620–621
Mixed models 16:48
Mixed motive approach, for employment discrimination 6:531–532
Mixed personality disorders 17:895
Mixed political culture 2:191
Mixed race 19:835
Mixed Rasch models (MRMs) 17:837
Mixed strategies
in game theory 9:575–576, 9:584–585
mixed strategy equilibria 9:584–585
Mixed welfare state 18:324–325
Mixed-affect, contra-hedonic motivation 1:242
Mixed-effects modeling 19:728
Mixed-integrated methods 9:275
Mixed-member proportional (MMP) 19:205–206
Mixed-motive 22:245
Mixed-motive situations, and bargaining 16:415
Mixed-sex interaction 13:257–258
Mixed-sex schooling 3:926
Mixteca Alta of Oaxaca 15:263
Mixture distributions 15:624–625
Mixture models 15:624–628
factor analyzers 15:627
mixture of normal 15:626–627
ML estimation 15:625–626
model-based clustering 15:626
notation 15:624–625
number of components 15:626
for statistical clustering 23:393–394
t-distributions 15:626–627
Mixture of normal 15:626–627
Mixture sampling 16:117
Mixtures of common factor analyzers (MCFA) 15:627
Miyamairi rite 20:367–368
Miyazawa, Setsuo 13:515
MJI *see* Moral Judgment Interview (MJI)
ML *see* Maximum likelihood (ML) estimation
MLB *see* Major League Baseball (MLB)
MLBA *see* Multiattribute linear ballistic accumulator (MLBA)
MLBI *see* Major League Baseball International (MLBI)
MLE *see* Maximum likelihood estimation (MLE)
MLM *see* Multilevel random coefficient modeling (MLM)
MLQ *see* Multifactor Leadership Questionnaire (MLQ)
MLTM *see* Multicomponent latent trait model (MLTM)
MLU *see* Mean utterance length (MLU)
MMF *see* Milbank Memorial Fund (MMF)
MMI *see* Multiple Mini Interview (MMI)
MMN *see* Mismatch negativity (MMN)
M4-model 14:774–775, 14:774f
MMOG *see* Massive multiplayer online games (MMOG)
MMP *see* Mixed-member proportional (MMP)
MMPI *see* Minnesota Multiphasic Personality Inventory (MMPI)
MMPI-2 *see* Minnesota Multiphasic Personality Inventory-2 (MMPI-2)
MMPI-A *see* Minnesota Multiphasic Personality Inventory; adolescent form (MMPI-A)
MMS *see* Minerals Management Service (MMS)
MMSE *see* Mini Mental Status Exam (MMSE)
MMU *see* Manchester Metropolitan University (MMU)
M'Naghten rule 13:492
MNATP *see* National Museum of Popular Arts and Traditions (MNATP)
MNCs *see* Multinational corporations (MNCs)
Mnemonic window 21:914
Mnemonics 13:712
memory 1:881–882
for study of psychology 13:713
MNEs *see* Multinational enterprises (MNEs)
Mnestic block syndrome 1:651–654
MNL *see* Multinomial logit (MNL) model
MNP *see* Multinomial probit (MNP) model
MNS *see* Mirror neuron system (MNS)
Mo Tze 4:311
Moa 17:449
Mobil 17:166
Mobile access and apps, journals 12:867–868
Mobile applications, and psychotherapy 19:380–381
Mobile communication 15:629–635
among particular groups 15:631–632
business and formal organizations 15:631
teen use 15:631–632
women in developing countries 15:632
commercial use 15:631
definition 15:629–630
history of 15:629–630
individual addressability and interlaced activity 15:630
location-based services 15:632–633
methodological issues in studies 15:633
microcoordination 15:631
social cohesion 15:633
structural embedding in society 15:633
Wi-Fi-enabled devices 15:630
Mobile communication competence 4:300
Mobile conjugate reinforcement, and early memory development 6:762
Mobile early childhood care and development programs 6:743–744
Mobile learning, lifelong learning 14:96–97
Mobile learning platforms 6:898–899
mobile computing 6:898
tablets and e-books 6:899
Mobile livestock 16:839–841
Mobile money agent 2:873
Mobile phones 13:521, 24:157–161
Mobile technology 21:189–190
Mobile telephones 12:493
Mobile urbanism 24:948
Mobilities 7:165–170, 15:638, 24:616, 24:620
of Caribbean 1:641–642
children 3:495
in China 2:65–69
and community sociology 4:366–367
in cultural geography 5:412
equality of opportunity and 7:903–904
and equity in education 7:919, 7:922
and hunter-gatherer archaeology 11:424–425
and land tenure 13:234
population 15:397
of radio 19:877, 19:879
social 15:640
urban settlements 24:836–837
Mobilities 3:495
Mobility and travel–activity patterns 15:636–639
activity–travel patterns 15:637–639
automobility 15:637–638
daily mobility 15:636–637
human mobility 15:638–639
time geography 15:638
Mobility of people and national security 16:254–255
Mobility revolution 19:380
Mobility routes within profession 13:630
Mobility strategies 23:159–160
Mobility transition framework 12:433
Mobility transitions 15:390
Mobility-focused transportation 24:620
Mobilization 4:145–149
direct voting 25:298
indigenous people 11:757–759
populism 18:609
as predictor of political participation 18:506
religious 20:328–332
Mobilization theory, relative deprivation 6:216–217
Mobilizing power of miracles 15:576–577, 15:579–580
MOC *see* Managerial and organizational cognition (MOC)
Mock jury experiments 12:908
Moclobemide 1:778
Modafinil, cognitive enhancement 17:269–270

Modal age at death, oldest-old mortality 15:864
Modal logic 14:296
Modal model of memory 25:175, 25:713–714
Modal reasoning 5:95
Modality 5:91–92, 21:588, 21:932, 22:14, 22:688
Modality match, and embodiment 7:413
Modality of information 22:559
Modals 6:54–55
MODE *see* Motivation and Opportunity as DEterminants (MODE) model
Mode
  fashion 8:835–836
  multiple language 16:59, 16:59f
Mode 2 knowledge 18:309–310, 18:312
Mode 2 knowledge production 24:760
Mode of culture and individual 5:472
Mode of production 3:94–95, 14:671–672, 14:675–676
  in feudalism 9:112–113
Model assessment, Bayesian statistics 2:376
Model averaging 15:650
  capture–recapture methods 18:606
Model checking
  Bayesian statistics 2:376
  linear hypothesis 14:147–148
Model cities 16:452
Model comparison, Bayesian statistics 2:376
Model complexity criterion 4:475–477
Model ethical protocol 11:308
Model fidelity 4:490–491
Model human processor (MHP) 11:400–401
Model instability 24:339
Model interaction 4:488–489
Model learning, for suicide 23:677–678
Model ln-likelihood function 8:716
Model minority stereotype, and Asian-Americans 19:791
Model modification 8:688–689
λ-model of elastic properties of muscles 15:959
Model of frame selection 1:78
model of possibility 19:64
Model of Strategic Learning (MSL) 13:716–717
Model Penal Code (MPC) 13:751, 21:770
Model prison, Bentham, Jeremy 2:525
Model selection 15:647
  capture–recapture methods 18:606
  cross-validation and generalized cross-validation 15:649
  high-dimensional variable selection 15:650–651
    best n-term selection 15:650–651
    shrinkage 15:651
    sparse variable selection problem 15:651
  information criteria 15:647–649
    Akaike's information criterion 15:648
    focused information criterion 15:648–649
    Mallows' *Cp* 15:648
    Schwarz Bayesian information criterion 15:648
    value of R2 in linear regression models 15:647–648
  methods, asymptotic properties of 15:649–650
  pretesting and conditional inference 15:650
  smooth testing 23:402
Model selection tools 4:478
  see also Bayesian model selection
Model testing and selection 15:653–658
  Akaike's information criterion 15:653, 15:656–657
  Bayesian information criterion 15:653, 15:656–657
  Bayesian model selection 15:656–657
  classical hypothesis tests 15:655
  discrepancy-based methods 15:655–656
    bootstrap techniques 15:656
    cross-validation techniques 15:656
  following minimum description length principle 15:657
  goodness of fit 15:653–655
    Bayesian model checks and criticisms 15:654–655
    for continuous random variables 15:654
    for discrete random variables 15:653–654
    testing against specified alternatives 15:654
  maximum likelihood method 15:654, 15:656
  normalized maximum likelihood 15:653, 15:657
  posterior predictive model checks 15:655
Model validation 4:490, 14:794
Model Work Health and Safety Bill 17:144
Model-based adjustments 17:109
Model-based analyses, in crossover designs 5:352–353
Model-based ancestry estimation 9:888–889
Model-based clustering 15:626
Model-based data, statistical disclosure limitation 4:580
Model-based graphics, regression analysis 14:160
Model-based sampling 23:186
Model-based seasonal adjustment methods 24:351–352
  regression methods 24:351–352
Model-building activity 22:472
Model-driven spatial data analysis 23:97–98
Model-implied instrumental variable estimators 23:616–617
Model-implied instruments 23:616
Modeling
  in ABA 2:410
  families as educational settings 8:735
  multiple imputation 16:89–90
  with neural constraint-satisfaction 4:721–722
Modeling dilemma, and epistemic cultures 7:876
Modeling faking 17:834
  mixed Rasch models (MRMs) 17:837
  social desirable responding (SDR) 17:834
  structural equation modeling 17:836
Models
  biomedicine 2:664
  of bounded rationality 5:911
  computational 21:291
  see also computational models of scientific discovery
  of decadence 11:33
  for dichotomous responses 19:436–437
    two-parameter logistic (2PL) model 19:436
  of disordered personality 17:915
  of influence in networks 15:659–666
  of intelligence 12:284–286
    Cattell-Horn-Carroll (CHC) theory 12:285
    PASS theory 12:285
    in review 12:285–286
    Sternberg's triarchic theory 12:285
    Wechsler factor structure 12:284
  of labor market discrimination 9:759–761
  of lower-level openness structure 17:231–232
  needs assessment 16:402–403
  for polytomous responses 19:437
  of public enforcement
    choice between penalties and fines 13:574–575
    goals other than deterrence 13:575
    judicial errors 13:575
    marginal deterrence 13:574
    no deterrence of desirable acts 13:573
    normative framework 13:574
    optimality of maximum fines 13:573–574
    optimality of underdeterrence when detection is probabilistic 13:574
    probabilistic detection 13:573
    risk aversion 13:574
    strict *versus* fault-based approaches 13:573
  in science, homeomorphs and paramorphs 15:291
  as a scientific instrument 21:291
  social history 22:309–310
Model-theoretic interpretation of natural language 15:751–752
Model-theoretic semantics 15:750
Moderate state 18:386
Moderated mediation 22:549–550
Moderation 15:68, 22:549
  conceptual issues in testing 15:672–673
Moderator analyses, meta-analysis 23:911–912
Moderator variable 15:672–674
  conceptual and analytic issues in testing moderation 15:672–673
  examples of hypotheses involving moderation 15:672
  interpretive ambiguities 15:673–674, 15:673t
  main effect 15:672–673

Moderators 22:549
of attitude–behavior consistency 2:188–189
controlled trials 19:389
cultural intelligence as 5:437
embodied social cognition 7:412–413
consistency 7:413
construal level 7:413
familiarity 7:413
modality match 7:413
of psychological contact breach
person factors as 19:368–369
situational factors as 19:369
role in diversity effects 4:740–741, 6:597
Modern alphabets 25:764–765
Modern art, modernity 15:683–685
Modern behavior, emergence of 17:458–459
Modern bioethics movement 13:558–560
Modern bourgeoisie 2:785
genesis of 2:786–787
Modern capitalism 3:106–107, 13:566, 20:259–260, 20:313
Modern civil service 3:696
new states 3:696
nondemocratic regimes 3:696
Modern codes of law in Siam 13:814–815
Modern corporation, in property rights 24:531
Modern corporatism 12:374–375
Modern criticism, feudalism 9:113–115
Modern Darwinism 8:386–387
later developments 8:387
synthesis 8:386
Modern day slavery 25:125
Modern democracies 19:636
Modern dictatorship 6:369
and totalitarianism 6:369–370
Modern economic development 10:150
Modern economic growth 7:41–42
exogenous models for transition 7:42
Modern economy 7:16–18
artisans 2:46
Modern educational system 7:172–173
Modern egalitarian societies 11:898
Modern European university 24:758
Modern evolutionary synthesis 9:970, 12:176
Modern financial theory 9:177
Modern gender gap 25:304
Modern German history 25:292
Modern history 18:381
Modern homes 11:157
Modern human cognition 11:252–258
cultural evolution 11:254
Hominidae 11:253
hominin 11:254
human eco-niche 11:256–257
intersubjectivity 11:253–254
joint attention 11:254–255
language development 11:256
mimesis 11:256
pedagogy 11:254–257
symbolic cognition 11:255–256
toolmaking 11:254–255
Modern individualism 20:427–428
Modern industry 11:873–874
Modern intellectuals 12:278
Modern international legal system, formation of 11:796
Modern Jewish historiography 11:70–71
scholarship 11:68
Modern knowledge society 13:105, 13:107
Modern law 13:566
Modern legal sociology 13:550
Modern leisure 13:852–853
Modern *Lex mercatoria* 13:915
main characteristics and practical impact 13:916
Modern life story 2:645
Modern management 15:675–681
changing contexts of management practice 15:676–677
chronology 15:675–676
guilds 15:679–680
modern firm 15:678–679
theories of firm 15:677–678
Modern mass media, exposure, in collectivism/individualism 4:208
Modern mass migration 11:649–650
Modern medicine 2:663
Modern missionary movement 15:614–615
Modern money 15:746
Modern museum 16:142
Modern nation states 4:339
Modern nationalism 16:276–277
Modern organizations 16:179–180
Modern period, South Asia 23:56
Modern property rights 24:528–530
Modern public administration 19:522–523
Modern public life, golden age of 19:616
Modern racism 19:852
rise of 19:854–855
Modern reason, critique of 1:82
Modern religion 2:507–508
Modern residential location behavior, in network society 14:290
Modern scales of sexism 21:697
Modern science, formation of 17:12
Modern Sexism Scale (MSS) 21:697
Modern slavery 22:45–46
peaceful *versus* violent endings of 22:41
Modern social imaginary, Charles Taylor 24:41–42
Modern societies 4:338–339, 19:597, 25:670, 25:672–673
division of labor in 13:158–161
functional differentiation 9:541
leisure *versus* free time 13:845
and liberalism 14:28–29
as ethos of modernity 14:28–29
as threatened by modernity 14:29
and memory 4:203–204
role of competition in 3:17–18
social differentiation 6:391–394
and tradition 24:505
work in 25:689
Modern sport 23:300
Modern statehood 23:347–348
legitimacy and 23:350–351
Modern states 16:263–264, 17:515, 23:354
and media 22:584
political sociology 18:473
Modern synthesis of Darwinism 8:386, 22:901
Modern system of scientific disciplines 21:289
Modern terrorism 24:227
Modern theories, of personal identity 17:745–746
Modern transgender movement 24:551–552
Modern typology 14:209–210
Modern university press 19:624–626
Modern urban planning (Western), history of 24:862–868
achievements and legacy of 24:868
antecedents and beginnings 24:862–863
brownfields 24:866
city beautiful 24:863–865
first half of the twentieth century 24:863–865
garden city 24:863–865
governance 24:862, 24:864–866
greenfields 24:865–866
modernization 24:862–863
new millennium planning 24:866–868
New Urbanism 24:867
open space 24:862–865, 24:868
postconsensus urban planning 24:865–866
postwar master planning 24:865
reconstruction 24:865
sustainability 24:867
urban regeneration 24:866
urbanization 24:862
Modern war 17:620
Modern warfare 15:520
Modern western societies 18:434
Modern work 13:853
Modern world 3:720, 25:341–343
small states in 22:111–112
Modernism 8:225, 14:8, 18:684–685
Daniel Bell 2:503
and feminism 8:900
Japan 14:635, 14:637, 14:639–640
and nation 16:224–225
and sustainability 18:198
Modernities 3:720, 4:405–406, 7:380, 7:656–658, 7:660, 7:662, 11:102–104, 13:566, 15:682–686, 16:459, 18:383, 18:684–685, 21:388–392, 23:633–635, 24:166–168, 24:461–462, 24:813, 24:815, 25:340
alternative modernities 15:689
in anthropology 15:687
anthropology as nostalgia of 17:37
and Arab nationalism 16:271–272
attractiveness of 15:683–684
avant-garde 15:682–683, 15:685
Bielefeld School 2:553–555
and busyness 3:49–50
characteristics of 15:682–683
in China 2:67–69
cities and citizenship 15:689–690

civilization of 3:728–729
commodity production and fetishism 15:688
concepts (before 1945) 15:700–701
construction, tradition and 24:504–505
critical theory and 5:307–308
criticism of 15:684–686
critique and reappraisal (since 1980s) 15:703–705
contradictions of modernity 15:704
to postmodernism 15:704–705
cultural and political program of 3:728–729
culture as 21:390
cybernetics 5:649
and democracy 6:80–81
and dictatorship 22:139–140
differentiation 15:690
and disposable youth 6:550–551
East Africa 6:776
Enlightenment 15:683
eudaemonism in philosophy of 8:213–214
European, psychoanalysis 19:326
and fashion 8:835
food/cuisines 9:293
Frankfurt School 9:383–385
French Revolution 9:409, 9:411
Giddens on 10:135–137
and global history 10:165–167
and global south 10:189–190
historical sociology 11:27
in history 15:700
history of 15:692–696
alternative modernities 15:693–694
decolonial thought 15:694
eurocentrism 15:693–694
globalization 15:693–694
modernization theory 15:692–693
multiple modernities 15:693–694
postcolonial theory and coloniality/modernity paradigm 15:694–695
and social sciences 15:692
underdevelopment theory 15:693, 15:695
and hybridity 11:450
individualism 15:682–684, 15:686
and intellectuals 13:72
and its other 15:687–688
in Japan 15:129
and laicization 13:216
late, structural defects of 16:835–836
Latour on 13:414–416
and liberal Christianity 14:8, 14:10
liberalism as ethos of 14:24–25, 14:28–29
liberalism as threatened by 14:29
literature 15:683
Maruyama Masao on 14:636–637, 14:639–640
Marx on 15:700–701
Melanesian 17:156–157
MENA 16:361
and militarism 15:492–493
in military 15:507–509
modern art 15:683–685
modernization 15:682
and modernization 15:707–712
multiple 3:729
and music 16:151–152
of nationalism 16:262
new modernism 15:688–689
new theory of 15:705
and political thought 18:480–481
formation 18:482
politics of 16:459
postmodernism 15:685–686
progress 19:123, 19:125–126
of radio 19:877–878
and rationality 15:688
political and social 15:689–691
reformation 15:682
renaissance 15:682
and selfhood 21:446
Simmel on 15:684–685
Smith on 22:118–119, 22:121, 22:123
social political and economic relations 16:459–460
sociocultural condition of 11:865–866
in South America 23:47–48
in South Asia 2:80
in Southeast Asia 23:73
spirit possession 23:282
and sport 23:313–314
subject formation 15:690
and subjectivity 15:689
Taylor on 24:39, 24:42
theories in social sciences (after 1945) 15:701–702
theory in historical work 15:702–703
Germany 15:702–703
United States of America 15:702
threat to liberalism 14:25
and tourism 24:465, 24:468
and tradition 24:508–511
transition of law and science to 21:156–157
transnationalism 15:690
uncertainty 15:690–691
unintended consequences 24:747–748
urbanization 15:683
and war 25:363, 25:366
Weber on 15:701
Western 24:583
and witchcraft 25:605–607
Modernization 3:106, 3:122, 3:125, 3:314, 5:14–17, 11:103–104, 13:109, 15:697–699, 16:285, 18:181–184, 18:434, 21:242, 22:804, 23:369, 24:862–863
agricultural 1:458–459
Bielefeld School 2:553–555
and body 2:728–732
and brokers 2:871
center–periphery diffusion 3:309–310
in China 2:65, 2:69
civil liberties 15:697
concepts (before 1945) 15:700–701
critique and reappraisal (since 1980s) 15:703–705
contradictions of modernity 15:704
to postmodernism 15:704–705
definition 15:707–708
democracy 15:697–698
democratization 15:697–698
development crises 15:698
differentiation 15:697
in East Asia 6:793–795
and East Asian economic development 6:788
and emotions 7:514–517
equality 15:697
in history 15:700
and human development 6:308–310
human development sequence 15:697
and intellectuals 13:72
of Japan 15:127–130
in Korea 2:73–74
Marx on 15:700–701
in Mexico 1:627–628, 1:630
and Middle East society 16:381–385
modernity 15:682
and mortality decline 15:869
and nationalism 16:291–292
and nationalism in East Asia 16:276
new theory of 15:705
participation 15:697–698
and pension policy 17:693
political
determinants of successful paths to 15:698
routes to 15:697
simple and complex patterns of 15:697–698
political development 15:697
and political economy 1:938–939
progress 19:123, 19:125–126
of public sector management 7:263
R & D and 20:498
rationalization of authority 15:697
reflexive and nonreflexive 15:698
reflexivity 20:90–91
and revolutions 20:644
rights, political 15:697
risk society 15:698
self-expression values 15:697–698
sociological theories of 15:707–713
bureaucracy 15:707, 15:709–711
civil society 15:707
as communication phenomenon 15:712
democracy 15:707, 15:711
global spread 15:710–712
history of study 15:708–709
human rights 15:710–711
industrialization 15:712
informatization 15:711–712
innovation 15:707–712
modernity 15:707–712
non-Western societies 15:709–710
postindustrial society 15:707–708
social transformation 15:708–712
urbanization 15:707
in Southeast Asia 23:72
and Soviet education 20:825–830
of sport 23:307, 23:309–311
in sports 23:299–300, 23:303
survival values 15:697–698

Modernization (*continued*)
theories in social sciences (after 1945) 15:701–702
theory in historical work 15:702–703
Germany 15:702–703
United States of America 15:702
Wallerstein's critique 25:341
in West Africa 1:301
world systems theory 25:748, 25:750–751
Modernization theory 6:98, 13:443, 15:692–693, 21:150–151, 23:369, 24:300–301, 25:748, 25:751
and Africans colonies 1:288
German scientific culture and 15:702
in social sciences (after 1945) 15:701–702
Modernization thesis 25:512
Modernizing conditions, fertility under 9:32–33
Moderns, classics and 3:776
Modernus 15:700
Modes of incorporation 2:81
Modes of policing 18:248
Modes of production (MOP) 7:15–16, 18:362, 18:364
and class struggle 6:242
objective 7:16
work anthropology 25:662–663
Modes of religiosity 20:725–726
Modesty 19:231
Modica, M., on sanctity 15:577
Modiclus 4:872
Modifiable areal unit problem (MAUP) 23:95
Modification 23:886t
Modification-use of a Patient nominalization 23:892t
Modification-use of an Actor nominalization 23:892t
Modified Erikson Psychosocial Stage Inventory (MEPSI) 7:941
Modified Mini-Mental State Examination (3MSE) 9:845
Modified profile likelihood 14:122
example 14:122
Modifier–modified element 11:827
Modular intelligence 23:295–296
Modular Online Time Use Survey (MOTUS) 24:392
Modular partitioning, MDS 16:44f
Modular product architecture 15:721
Modularity 7:469, 19:355
and business platforms 3:38–41
cognitive neuroscience 4:97
cohesive groups 3:909–910
complex systems 15:718–720, 15:722–723
concept of 15:718–719
information hiding 15:718–719, 15:721–722
loosely coupled systems 15:718, 15:721
mirroring hypothesis 15:718, 15:720–722
modules 15:718–722
near-decomposable systems 15:718, 15:720
and organizations 15:720–722
process of 15:719–720
product architecture 15:718–719, 15:721
Modularity approaches to cognition 4:97
Modularity index 3:909
Modularity score 3:909–910
Modulation 19:64
Modulatory circuit, in fear-potentiated startle 8:861
Modus arguendi in law 13:778
Modus ponens 14:307
Modus tollens 14:307
Moetotlos 19:752–753
MOF *see* Ministry of Finance (MOF)
MOFA *see* Ministry of Foreign Affairs (MOFA)
MOFCOM *see* Ministry of Commerce (MOFCOM) (China)
Moffitt, Terrie 14:60
dual taxonomy theory 14:60
Mogutin, Iaroslav (Slava) 21:743
Mohan, G., works, *Beyond participation strategies for deeper empowerment* 6:292–293
Mohenjo Daro 23:53f
Mohists 4:311–312
Mohr, John
on networks 16:562
on social network analysis 5:560
MOHURD *see* Ministry of Housing and Urban-Rural Development (MOHURD)
Moieties 3:732
Moilanen, K.L., on adolescent self-regulation 21:549
Moisander, Johanna, on epistemic cultures 7:878
Moivre, Abraham de 9:624
Moknèche, Nadir 21:753
Moksha 14:253
Mol, Annemarie, works, *The body multiple* 5:525
MOLA *see* Multiobjective land allocation (MOLA)
Molar organizational climate 17:328–329
Molecular basis, of asymmetry of body/brain 2:121–128, 2:122f
Molecular biology 2:676–677, 18:103
and biomedicine 2:664, 2:666–667
Molecular clock 13:25
Molecular events, in neuronal activity 12:710–715
Molecular evidence
primates, evolution of 18:904, 18:906–908
Molecular evolution 13:24–25
Molecular genetics 2:450, 6:320, 12:297, 12:300–302, 21:611
and crime 2:673
and economics 9:953
and personality traits 17:921–922
studies of sexual orientation 21:780–783
Molecular health 15:124–125
Molecular hypothesis of reinforcement 20:215
Molecular imaging
affective disorders 14:441–443
molecular imaging studies 14:441–443
structural and fMRI studies 14:441
Alzheimer's studies 14:443–444
functional cerebral glucose and activation studies 14:443
molecular imaging studies 14:443–444
structural studies 14:443
cognition 14:439–441, 14:443
dementia 14:443
diagnosis 14:443
findings 14:441–444
schizophrenia 14:441
neurobiology 14:439, 14:444
pathophysiology 14:441, 14:443–444
positron emission tomography 14:439–444
in psychiatry 14:439–445
psychosis 14:441–442, 14:442f
with radiolabeled ligands 14:440
general aspects 14:440
principles 14:440
single photon emission computed tomography 14:439–441
Molecular interactions, computational neuroscience 4:482–483
Molecular mechanisms 22:416–417
Molecular medicine 15:123–124
Molecular memory 23:874–875
Molina, J.A., works, Household Economic Behaviors 11:226
Mollies 13:875
Mollusca 7:694
Molotch, Harvey 11:283–284
on Chinese art districts 5:599–600
Moltke, Helmuth von 15:520
MOM *see* Method of moments (MOM)
MoMA *see* New York Museum of Modern Art (MoMA)
Moment matching 15:299–300
Moment-to-moment choices 14:390–392
Moment-to-moment variation 21:947
Mommsen, Wolfgang J. 2:554
Monarchia absoluta 1:6
Monarchies 2:920, 23:363–364
French Revolution 9:409–411
and individual-collective relationship 11:817
Plato 18:212
Monarchism, and Enlightenment 7:665
Monasteries as learning centers 15:725
Monastic asceticism 2:61–62
Monastic medicine 15:119
Monastic medicine/presalernitarian medicine 15:119
Monasticism 2:891
within Anglo-Catholicism 15:725
becoming religious 15:724–725
Celtic 15:725
civilizational missions 15:725
'contemplative' and 'active' styles of religious life 15:725–726
contradiction in 15:726
detachment from responsibilities of marriage and family 15:724
life of the monk 15:724
oriental 15:724
and religious orders 15:725–726

'terminal' stage in 15:726
use of term 15:724
West and East 15:724–726
Monday Holiday Act 4:238
Mondialization, and nation 16:225
Monetarism, and macroeconomics 14:402–403
Monetary benefits, of psychiatric care 19:298
Monetary gain 5:911
Monetary market 12:52
Monetary policy
conventional, new ways of implementing 15:732–733
conventional monetary policy, down to the zero lower bound 15:728
forward guidance on the policy interest rate 15:728–729
Hurst's contribution 11:439–440
large-scale asset purchases 15:729–732
portfolio losses 15:731–732
price inflation 15:731
macroprudential concerns 15:733
ongoing usefulness of asset purchases and sales 15:732
ongoing usefulness of forward guidance 15:732
open market operations 15:730–731, 15:733
special lending to financial institutions 15:727–728
Monetary theorists 14:409
Monetary theory 15:748
Fisher's 9:219–220
and political economy 18:357–358
Monetary value, and moral value, relationship between 19:261f
Monetization 5:538
Money 21:405
anthropology of 15:746–749
commodity 15:747–748
earmarking 15:748–749
economy 15:747–748
exchange 15:746–749
market 15:746
materiality 15:747
payment 15:746–748
token 15:747–748
value 15:746–749
economic psychology 5:888–889
and exchange 8:476, 8:478, 8:480
political 18:393–397, 18:442
psychology of 5:889, 15:735–740
attitudes 15:736–737
consumption 15:736–737
donation 15:738
handling 15:737–738
personal and social significance 15:735–736
perspectives on poverty 15:738–739
priming 15:737
saving, spending, and investing 15:737–738
tax compliance and evasion 15:738
thinking 15:737
quantity theory of 14:406–409
sociology of 15:741–745
premodern to modern money 15:741–743
in theory 15:744–745
theory of 14:407–408
Money Attitude Scale 15:736
Money boys in China 21:738
Money fetishism 11:866
Money laundering 5:167, 5:216
Money priming 15:737
Money Supply (M2) 9:317
MoneyGram 12:519
Mongkut *see* Rama IV (Mongkut)
Mongo Big Men 3:312
Mongo patrilineal societies 3:312
Monin, C.C., on categories 15:576
Monism 7:311–312, 12:509
value pluralism 25:21–25
Monistic theory of state 18:232
Monitor 16:772
Monitor Model 21:362
Monitoring 1:346–349, 23:7, 23:369
coping styles, chronic illness 3:551–552
prospective memory 19:273–276
rheumatoid arthritis 20:658
workplace deviance 25:724–725
Monitoring groups 10:257
Monitoring intensity principle 1:346–347
Monitoring the Future (MTF) 5:160–161, 21:110
Monitory democracy 22:635
Moniz, Egas 23:18
Monkeys, intelligent movements of 23:292–294
"Monkish" virtues 15:817
Monks, ascetic 2:60–62
Monoamine hypothesis 19:587–588
Monoamine oxidase (MAO) 16:750, 21:611
Monoamine oxidase inhibitors (MAOIs) 17:470–471
Monoamine oxidase-A (MAO-A) 9:964, 14:442
nature/nurture 16:343
Monoamines 16:749–751
blockade of enzymes that inactivate 1:779
Monoaminooxidase inhibitors 1:778
Monocentricity of cities 24:823–824
Monochromatic near-infrared light 16:388
Monochronic time system 12:357
Monod, Jacques 21:215
Monofactorial designs 19:726
Monofactorial significance tests 19:726
Monogamous norm, spread of 18:902–903
Monogamy 15:816–817
as monogamous inheritance 14:617–618
socially imposed 14:616–617
Monograph, and anthropological writing 1:740, 1:742
Monographic method, empirical social research (ESR) 7:560
Monoid, algebra 22:479–481, 22:483, 22:480t–481t
Monolingual proficiency in language 16:66
Monolingual society, and illiteracy 11:589
Monolithic grammatical systems 18:141–142
Monolithic societies, and group violence 14:704
Monomania 17:891
Monomorphic associations 25:256–257
Mono-operation bias 24:639–640
Mono-operationalization 24:640
Monopolistic power 1:201
Monopolization
of media 19:605
and professionalization 19:98
Monopoly 21:141–144
in Chinese cultural industries 5:597–598
and mafia 14:430–431, 14:433
of organized crime groups 14:430
and price regulation 20:160–161
regulatory agencies 20:181
Monopoly regulation, empirical analysis of 20:173–174
restructuring, effects of 20:175–176
incentive-based regulation 20:176
performance 20:175–176
privatization, impact of 20:175
Monopoly union model, trade unions 24:500, 24:500f
Monopsony models, of minimum wages 15:562
Monopsony power 6:537
Monosyllabic signs 21:944
Monotheistic faith 13:801
Monotheistic idea 12:869
Monotheistic religions 3:727, 15:354, 20:362
Monotone regression 16:43
Monotonic logic 19:64
Monotonicity 4:447–448
Monotonicity assumption violation, sensitivity analysis for 4:452
Monotonicity axiom 4:873–874
Monoubiquitination, histone 7:841
Monozygotic twin study, trait/disease-discordant 24:691
Monozygotic (MZ) twins 15:204, 16:340, 16:344, 21:779, 24:703
fixed-effects estimates of impacts of demographic variables 24:705–708
Monro, Surya, on sexual citizenship 21:729–730, 21:732
Monroe Doctrine 1:617–618, 11:682
Monsiváis, Carlos 13:862
Montages, electroencephalography 7:358–359
Montagu, Ashley 2:712
peace studies of 17:624
Montague grammar 4:830, 15:750–757
dynamic 15:754
early work on linguistics 15:753
fragment of PTQ in 15:752–753
further developments and controversies 15:754–755
historical context of 15:750–751
linguistics–philosophy interactions 15:753–754
logic and intensionality 15:752

Montague grammar (*continued*)
naturalization of formal semantics 15:754
pragmatics 15:753
syntax–semantics interface 15:754
theory and the substance of 15:751–753
Montaigne, Michel Eyquem de 1:765–766
Monte Carlo integration
Bayesian experimental design 8:523–525, 8:525f
Monte Carlo (MC) methods 1:531, 15:763–767
additional uses in Bayesian inference 15:765–766
for demographic microsimulation 15:343
implementation issues 15:765
Markov Chain Monte Carlo (MCMC) methods 15:763–765
Gibbs sampling algorithm 15:764–765
Metropolis–Hastings algorithm 15:763–764
multiple-block M–H algorithm 15:764
simulation 5:53–54, 14:569, 15:758–762
Montenegro, confederation 4:572
Montesquieu,, Charles-Louis de Secondat, Baron de La Brede et de 3:708, 7:721–722, 8:292, 13:802, 21:890–892, 22:720
as philosopher 22:990
small states 22:110–111
works
De l'esprit des lois 7:658, 13:802
Lettres Persanes (Persian Letters) 1:765–766, 6:590, 13:802
The Spirit of the Laws 4:751, 9:666–667
Montgomery 4:574
Montgomery Assessment of Vocabulary Acquisition 3:796
Monthly Review: An Independent Socialist Magazine 14:649
Montreal Protocol 10:255–257
Monumenta Germaniae historica 11:74
Monumental architecture
in archaic states 23:374–375
in South America 23:36
Monumentality, in North America 17:15–16, 17:19
Monuments 4:235
collective memory 4:201, 4:203
and chiefdoms 3:378–381
MOOCs *see* Massive open online courses (MOOCs)
Mood 1:241
and attitude 2:171–172
and hypertension 11:453–455
and social referencing 22:558
TAME 24:212–213
Mood disorders 10:501–502, 15:194, 17:120–121
and cancer 7:497
and mindfulness 15:550, 15:552
and neuroticism 16:744–745
Mood modification 22:558
Mood stabilizers, with psychopharmacotherapy 19:460
Moody, J.
definition for walk 22:481–482
Moon, Sun Myung 16:807
Mooney, James, on Native Americans 1:634
Moore, Barrington, Jr. 18:474, 19:677
biography of 15:768–769
career 15:768–769
educational influence 15:768
family 15:768
intellectual influence 15:768
impact of 15:772–773
influence of Moore's methodology 15:772
responses to social origins of dictatorship and democracy 15:772–773
publications of 15:769–772
after *Origins of Dictatorship and Democracy* 15:770–772
before *Origins of Dictatorship and Democracy* 15:769
social *Origins of Dictatorship and Democracy* 15:769–770
works
Authority and Inequality under Capitalism and Socialism. United States of America, USSR and China 15:771
Injustice 15:768, 15:770–771
Moral Aspects of Economic Growth and Other Essays 15:771–772
Moral Purity and Persecution in History 15:771–772
Political Power and Social Theory 15:769
Privacy 15:771
Rational Discussion: Comparative Historical Notes on Its Origins, Enemies, and Prospects 15:772
Reflections on the Causes of Human Misery and on Certain Proposals to Eliminate Them 15:768
The Social Bases of Obedience and Revolt 15:768
Social Origins of Dictatorship and Democracy 1:456–457, 11:27, 15:768, 15:772–773, 22:31
Soviet Politics – the dilemma of power 15:769
Studies in Social and Cultural History 15:771
Terror and Progress USSR 15:769
Tour de force 15:769
Moore, Barrington, Jr 15:768–774
Moore, Grace E. 4:682–683
Heider and 10:751
Moore, Jason W. 1:460
Moore, Michael, works, *Roger and Me* 6:623
Moore, Robert 4:364, 11:236–237
Moore, Thomas, works, *Utopia* 7:810
MOP *see* Modes of production (MOP)
Mopping-up operations 14:216
MOPR *see* International Organisation for the Support of Fighters for the Revolution (MOPR)
Mora 20:1
Moral agency 2:314
Moral authority, MENA 16:361, 16:364
Moral autonomy 2:318–319
Moral behavior 15:775
Moral beliefs 20:335
moral calculus 5:232–233
Moral cognition
cognitive-anatomical models of 16:613–614
functional neuroimaging of 16:609–613
Moral communities' hypothesis 20:335–336
Moral community, and responsibility 20:587–588, 20:590
Moral constituents 18:443
Moral consumers 8:25, 8:27–28
Moral crusades 15:590–591
Moral decision-making 13:465
coordination in 15:805
Moral decisions, neural basis of 16:612–613, 16:613f
Moral demand systems 2:273
Moral deserts 6:230
Moral development 25:50–51
children, media effects on 15:37–39
cultural differences in 15:781–787
cognitive factors 15:786
culture-based differences in justice reasoning 15:783–785, 15:784t
developmental approaches to morality 15:783t
egalitarianism 15:781–782, 15:784–786
future directions 15:786–787
inclusiveness of moral universe 15:785–786
Kohlberg's theory 15:782–783
moral domain 15:781–782, 15:781t
moral foundations theory 15:781
moral inclusiveness, country-level, and individual-level 15:785–786
moral ingroup 15:785–786
moral values and development of justice-oriented reasoning 15:785
norms 15:782–786
other developmental dimensions 15:785
pathways 15:782
power distance 15:784–785, 15:784t
societal change 15:782
and education 15:775–780
Kohlberg on 15:775–776
Piaget on 15:775
and prosocial behavior 19:229
social cognitive domain theory and 15:777
Moral development theories
assumptions and early theoretical roots 15:788–789
cognitive-developmental theories 15:788–789
socialization theories 15:788
compliance 15:788–790
conscience 15:788–790, 15:794
emotions 15:788–792, 15:794–795
empathy 15:788–791, 15:794
empirical research 15:789–793
cognitive-developmental theories 15:788–789
socialization theories 15:789–791
guilt 15:788–790, 15:794

judgment 15:788–789, 15:791–794
moral behavior 15:788–790
neuroscience 15:794
prosocial behavior 15:790–791, 15:794
reasoning 15:789–791, 15:793–794
recent trends 15:793–794
social domain theory 15:792–793
Moral dilemma discussion 15:775–776
Moral dimension
honor killing 11:186
Moral disgust 6:547
Moral domain 15:781–782, 15:781t
Moral economy 25:662
peasants 17:677
in Southeast Asia 23:71
work anthropology 25:662
Moral education 7:294, 15:776–777
aims of 15:777–778
community service and moral self 15:778–779
educational applications 15:778
Moral emotions 13:120, 16:611, 21:476, 21:478
Moral emotions, evolution of 7:469–470
Moral entrepreneurs 15:589–591, 15:796–799
concept 15:796
construction of social problems and moral panics 15:797–798
deviance 15:796–797
labeling theory 15:796–797
for political correctness 15:798
quest of American puritans for dry society 15:796–797
Moral evaluation 8:73
Moral facts 6:700–702, 6:704
Moral field, social rules in 4:3–5
Moral foundations of society 2:748
Moral foundations theory 15:781, 15:813–814
Moral hazard 1:345, 10:614, 12:104
effects 10:614–615
health insurance 10:642–643
insurance 12:212
job displacement insurance 12:819, 12:825–827, 12:829
workplace safety 20:865
Moral hazard problem 12:101
Moral idealism 13:808
Moral inclusiveness, country-level and individual-level 15:785–786
Moral ingroup 15:785–786
Moral insanity 17:892
Moral issues
animal rights in research 1:703
issue evolution model 12:767
Moral Judgement Test (MJT) 13:120
Moral judgment
culture-based differences 15:783–785, 15:784t
impairment 19:452
and moral behavior, association between 7:319
moral values and 15:785
neuroethics 16:667–668
Moral Judgment Interview (MJI) 15:782, 15:784
Moral law, Kant on 13:7
Moral leadership 13:667–668
Moral limits of markets 20:896
Moral luck 6:232
Moral matrix 25:47–48
Moral maturation 23:327
Moral meanings and judgments 20:333–334
Moral motivation, cognitive-anatomical basis of 16:611–613
Moral obligation 15:775
Moral order of world culture 10:227
Moral orders 2:751, 11:186, 23:853–854, 24:675–676
symbolic boundaries 23:853–854
Moral panic theory 4:498
Moral panics 15:26, 15:797–798, 15:800–802
concept of 15:26–27, 15:800
controversies 15:802
in history 15:800–801
mass mobilization 15:802
media role in 15:801
variations in 15:801–802
Moral performance 18:926
Moral persons, and taboo 24:1–2
Moral philosophers, Scottish 22:990
Moral philosophy 2:631, 2:634, 6:485–487
eudaemonism 8:212–214, 8:216
Spencer's contributions 23:274–275
and Walzer, Michael 25:347, 25:349
Moral pioneers 9:998
Moral pluralism 13:483, 13:487
Moral pragmatism 16:426
Moral principles 13:764
equality as 7:906–907
Moral psychology 7:469, 15:812–814
Moral realism, in legal philosophy 18:27
Moral reasoning 13:463, 22:784
differentiation in 15:804
in psychology 15:803–805
coordination in moral decision-making 15:805
distinguishing morality from other domains of social judgment 15:804–805
social hierarchy and social justice 15:805
social interactions and moral judgments 15:803–804
stages in the development of moral judgments 15:804
two moralities and two types of social interactions 15:803–804
rights 15:803, 15:805
social convention 15:803–804
social inequalities 15:805
welfare 15:803–805
Moral reasoning, prosocial
approach to 7:315–317
age and gender differences 7:317
socio-cognitive skills 7:317
Moral reasons 18:780–781
Moral relativism 8:72, 20:236–237
Moral responsibility 20:584
Moral self 15:778–779, 15:790
Moral sense 5:687
Moral sentiments 5:678, 16:612
and interests 12:393–394
theory of 24:981
Moral solidarity, and society 24:586
Moral stage theory 15:788–789
Moral statistics 1:852, 7:560
Moral systems, environmental risks 7:785, 7:788
Moral theory, and utilitarian social thought 24:981
Moral treatment 19:586, 24:274
Moral uncertainty 7:90–91
Moral values
cognitive-anatomical basis of 16:613–614, 16:614f
and development of justice-oriented reasoning 15:785
and monetary value, relationship between 19:261f
Moral-analogous behavior 14:368
Morales, Evo 23:48
Moral–ethical aspirations, and liberalism 14:12–13
Moralism, legal 25:76
Moralities 5:554, 8:71, 8:212–213, 10:522–523, 12:264–265, 13:121, 13:483–488, 15:769–770, 15:781, 15:806–810, 16:606–618, 17:933–934, 20:333–334, 25:358–359
absolutism 15:807
and AIDS 1:493
altruism 16:606, 16:612
anthropological ethnography 15:807
anthropology of 1:737–738
autonomy 15:807–809
Bell on 2:498
and bioethics 2:613–614
brain regions involved in moral functions 16:615f
cognitive-anatomical basis of moral motivation 16:611–613
affiliative motivational states 16:612–613
guilt 16:611–612
moral values 16:613–614, 16:614f
moral *versus* selfish decisions 16:612–613, 16:613f
other-critical/blaming moral sentiments 16:612
pity, compassion, sympathy 16:612
cognitive-anatomical basis of social knowledge 16:610–611, 16:610f
cognitive-anatomical models of moral cognition and emotion 16:613–614
community 15:807–809
Confucianism 15:806, 15:808–809
content-based subdivisions 16:610–613
continental European perspective 13:479–482
argument from correctness 13:480
claim 13:480–481

Moralities (*continued*)
dual nature of law 13:481
necessity 13:480–481
outermost border 13:481–482
participant and observer 13:482
positivism and nonpositivism 13:479–480
Radbruch formula 13:480–482
reasonable disagreement 13:481
relativism 13:481
separation thesis and connection thesis 13:479
social facts 13:479–480, 13:482
and contractarianism 4:803–805, 4:807–808
contractrianism 15:808
critical and positive 13:483–488
cultural evolution of 15:815–817
cultural group selection 15:816–817
gene-culture coevolution 15:817
moral memes 15:816
cultural psychology 15:807
debunking views 8:71
and democracy 6:85–86
deontology 15:808
developmental approaches to 15:783t
developmental psychology 15:806–807
distinguishing from social judgment 15:804–805
divinity 15:807–809
dualism 15:807–808
in the East 15:808–809
enforcement 13:486–488
ethnocentrism 15:806
evolution of 15:811, 15:813–815
evolutionism 15:806
frontal cortex 16:607–608, 16:611–612, 16:615t
functional neuroimaging of moral cognition and emotion 16:609–613
global ethics and cultural change 15:809
governance of rules 13:486
Habermas on 10:471
and happiness 10:523
harm principle 13:486–487
Hinduism 15:806, 15:808–809
and honor killing 11:186
human cooperation 11:259
human rights 13:487–488
and ILP 5:554
implications for differential vulnerability in psychiatry 16:614–616
inclusive legal positivism 13:485–486
individualism *versus* collectivism 11:818
and legal philosophy 18:24–25
legal positivism and natural law 13:484–485
legal validity 13:484–485
and LGBTI commercial sex 13:884
major depressive disorder 16:614, 16:616
markets and 7:67, 7:70
mind made for 15:813–815
empathy 15:814
fairness instinct 15:814–815
incest aversion 15:815
trolleyology 15:814
moral emotions 16:611
moral pluralism 13:483, 13:487
Muslim 15:808–809
of nationalism 16:262–263
in patients with macroanatomical brain lesions 16:607–609
Piaget on 15:806
pluralism 15:807
and practical reasoning 18:780–781
problem of altruism 15:811–813
group selection 15:813
kin selection 15:811–812
reciprocal altruism 15:812
sexual selection 15:812–813
process-based subdivisions 16:613
psychopathy 16:614–616
public policy 19:580
relation of religion with 20:334–336
relativism 15:807
religion 15:807–809
semantics 16:610–611
separability 13:485–486
separation thesis 13:485–486
social cognition 16:607, 16:609–610
and social control 22:235, 22:238–239
and sports 23:302–303
symbolic boundaries 23:852–854
toleration 13:487
tricksters 24:650, 24:652–653, 24:656
universalism 15:806–807
utilitarianism 13:484–485, 13:487, 15:808
and victimless crime 25:75–76
vindicating views 8:71
of the West 15:807–808
and wisdom 25:600
witchcraft 25:606–608
Morality
and religion 20:333–337
Morality of care 15:776
Morality of law 20:783
Morality offenses 15:589
Morality-based honor 5:430
Moralization
and food preferences 9:298
of relations among people 12:392
Morally defined function of law 18:27–28
Moral/orientational knowledge 13:93
Morals
Goffman on 10:276–277
and politics, relation between 22:120–121
and values 25:36–37
Morals and legislation, Bentham's science of 24:981–982
Moran coefficient 23:107
Moran scatter plot 23:96–97
Moran's I test 23:96, 23:107, 23:109, 23:203
Moratorium, identity development 21:434–436
Moravians 2:146
Morbidity 2:656, 2:661, 20:269
and alcohol 1:509
in cardiac rehabilitation 20:189
development, alternative directions of 4:464
disasters 6:104–105
and health 22:279
measures, epidemiology 7:824
and men 15:167, 15:170
and social networks 22:367
and socioeconomic status 22:931–932
in twenty-first century 2:464–465
Morbidity compression 4:463–468
aging 4:463
alternative directions of morbidity development 4:464
assessment 4:464
education 4:464–466
evidence from studies 4:465–466
healthy life expectancy 4:463–464, 4:467
healthy life years 4:464
hypothesis 4:463–464
lifestyle 4:464–466
mortality 4:463–466
population health 4:464–465
prevention 4:463–464
social differences in 4:466
subjective health 4:464, 4:466
survey 4:464–466
More, Thomas 2:145, 22:836
works, *Utopia* 2:145
More complete income (MCI) 11:733
More comprehensive income (MCI) 11:735
Morel, B.A., works, *Traité des dégénérescences* 8:239
Moreno, Jacob L.
on social network analysis 16:524–525, 16:530–531
works
Who Shall Survive? 16:524–525, 22:101
Moreri, Louis, works, *Grand Dictionnaire historique* 7:584
Mores (ethics) 8:71
Moret Law (Spain) 22:41
Moretti, Franco, works, *Distant Reading* 9:365
Morgagni, Giovanni Battista 16:681
works, *De sedibus et causis morborum per anatomen indagatis* 16:681
Morgan, C.L., works, *Introduction to Comparative Psychology* 25:449
Morgan, John 25:623
Morgan, Lewis Henry 1:746, 7:622, 13:53, 13:802, 17:22
works
Ancient Society 1:636, 1:746, 1:766, 14:670, 19:153
Iroquois 1:636
League of the Ho-de-no-sau-nee, or Iroquois 9:135
Morgan, Mary 7:96
Morgan, Thomas Hunt 8:231–232, 8:234
works, *Evolution and Genetics* 8:221
Morgan's metaphors 15:677–678
Morgenson, Scott L., works, *Spaces between Us: Queer Settler Colonialism and Indigenous Decolonization* 19:754–755
Morgenstern, O., prospect theory 19:262–265

Mori, Arinori, and Japanese sociology 24:584
Morin, K.M., works, *Civic Geography: Geography in America, 1860–1890* 10:74
Mørk, Bjørn Erik, on epistemic cultures 7:877
Morley, David
  works
    Family Television 2:217
    The Nationwide Audience 2:217
Mormon Church 2:147, 16:805
Mormon polygyny 11:249
Mormons 16:805–806, 20:269, 20:293–294
Morningness–eveningness 3:568
Morocco
  gender in 9:785
  legal system in 15:365
  religious rituals and belief 16:377–378
Morphemes 15:819, 15:826, 23:230
  of a language 13:942
  morphosyntactic and semantic properties 15:823
  selection 15:820
  shape of 15:820
  types of 15:820–821
Morphew, C., on gay athletes 9:658
Morphogenesis 23:627
Morphological analysis, archaeology 10:101
Morphological arguments 14:221
Morphological awareness 3:510–512, 23:267–268
Morphological case
  linguistics 3:183–186
Morphological case system
  case, case form, and case marker 3:183
  case marking fused with number and gender 3:183–184
  classifying cases 3:185
  functions of the cases in Latin, and in Ancient Greek 3:184
  Georgian 3:186
  Hungarian 3:186
  Kalkutungu 3:184
  Kannada 3:186
  Kunggari 3:186
  Latin 3:183–186, 3:184t
  range of cases 3:184–185
  suffix-I 3:183
  Turkish sentence, case example of 3:183
  Yuwaalaray case marking 3:185t
Morphological dependency 6:189
Morphological effect, in spelling of children 23:267–268
Morphological errors 4:93
Morphological identification 14:212
Morphological information
  and gender assignment 9:797, 9:797t
  sensitivity of children to 23:267–268
Morphological principle 23:265
Morphological relatedness 15:831
Morphological typology 14:209
Morphology 3:512, 11:825–827, 14:229, 15:827–829, 17:560, 20:5, 20:9
  and bilingual aphasia 2:574–575, 2:577–578
  canta- 15:819
  comparing property 15:819–820
  correspondence between category and grammatical information 15:820
  debates in theories 15:822–824
  exponents and morphosyntactic features 15:824
  inflection 15:821–822
  interfaces of 15:821
  lexical listing *versus* rules 15:823–824
  in linguistics 15:819–825
  -mos 15:819
  -ré- 15:819
  of revolutions 20:639–640
  and sexual selection, microevolution 15:340
  stem change in German and Danish plurals 15:829
  TAME 24:211
  urban 24:800
    analysis 24:802
  use of term 15:819–821
  word formation 15:821–822
Morphophonemes 15:831
  as candidate lexical segments 15:831–832
Morphophonemic analysis 15:829–832
Morphophonemics 12:447–448, 15:826
Morphophonemics of Modern Hebrew (thesis) 3:527–528
Morphophonology 15:826–833
Morphosyntactic environments 17:554
Morpho-syntactic grounds 23:885–886
Morphosyntactic integration 3:923
Morphosyntax 3:817t
Morphosyntax of creole 18:141–142
Morrill, Calvin, on corporate cultures 5:536
Morrill, John 7:653
Morrill Act 1862 20:803
Morrill Land-Grant College Act 1862 20:815, 24:762–763
Morris, Brian, on anarchism 1:683
Morris, Charles 18:795
Morris, Clarence 12:137
Morris, Errol
  works
    Fog of War 6:621
    The Thin Blue Line 6:623
Morris, William 5:111, 21:178
Morris Water Maze (MWM) 23:166
Morrison, E., on LGBTQ families 21:756
Morse, Edward S. 2:71
Morse, Samuel 12:491
Mortality 1:158, 2:656, 2:658, 2:661, 5:860, 6:119–121, 9:41–42, 11:5–6, 11:654–661, 15:845, 18:87, 18:597–599, 20:269–271, 22:931
  adult *see* Adult morality
  age standardization 6:121
  and alcohol 1:509
  from Alzheimer's disease 1:395
  of bereavement 2:532
  biodemography of 15:844
  evolution of aging, health, and mortality 15:845
  fertility/mortality trade-off 15:846
  frailty models 15:844–845
  future 15:847
  genetics of life span 15:845
  hazard rates 15:845
  ideas in genetic analyses 15:845
  index of cumulative deficits 15:846–847
  multivariate survival function 15:845
  rectangularization of survival curve 15:845–846
  stochastic process model 15:846
  Strehler and Mildvan's model of mortality and aging 15:845–846
of bipolar disorder 2:686
and cardiac disease coping 3:112
in cardiac rehabilitation 20:189
child *see* Child morality
children 9:66
of children and infant
  in industrialized countries 12:14–20
crises 15:857–859
decline in poor countries 15:861–862
decline of
  cohort explanations 15:870–871
  explanations 15:868–871
  features 15:868
  period explanations 15:869–870
demographic analysis 17:727–734
differentials in 15:851–856
disasters 6:104–105
early life influences in adulthood 6:752–758
economic growth 7:39, 7:41–43
epidemiology 7:825–826
and evolutionary demography 8:407–410
and family planning 8:789–790
fluctuations 15:871
future research 11:659–660
and health 22:276–279
health care 15:98–100
and health statistics 5:782
historical development 11:655
and HIV/AIDS epidemic 15:834–843
of immigrants 15:407
and individual behavior 15:861–862
of infant and children
  in less developed countries 12:21–26
LEXIS diagram for 6:160–161
life expectancy and life tables 6:121
marital status and 15:853–855
measures 6:120t
and men 15:166–167
methodological issues 11:655–656
and morbidity compression 4:463–466
and obesity 17:75–81
in old age 1:395
of oldest-old 15:863–867
and population dynamics 18:562–564
predictors of 6:378–379
religion and 20:267
sex as distinct from 10:814–815
and sexuality, in Africa 10:812–813
social relationships and 15:853–855

Mortality (*continued*)
socioeconomic status 22:931–935, 22:932f
and socioeconomic status 15:871–872
socioeconomic status (SES) and 15:851–853
suicide 23:683–687
trends, in Africa 1:282
in twenty-first century 2:464–465
Mortality acceleration, and warranties 17:174–175
Mortality crossover 15:848–850
adult mortality 15:849
cohort mortality 15:849
empirical evidence 15:848
immigration 15:849
Medicare 15:848–849
period mortality 15:849
proposed explanations 15:849–850
poor quality data and age misreporting 15:849–850
selective survival and population heterogeneity 15:849
race 15:849–850
social security 15:848–849
United States 15:848–850
vital registration 15:849–850
Mortality data, on social integration/networks 22:367–368
Mortality rates 9:319
adult mortality 1:158–159, 1:158f–159f
in Eurasia 18:548–549
Mortality risk, value of change in 25:14
Mortality selection 6:756–757
Mortality transition 15:857–862
cardiovascular disease 15:857, 15:859–861
causes of death 15:857, 15:860–861
infectious diseases 15:857–860
theoretical concept 15:857
Mortgage debt, and household wealth 25:451, 25:454
Mortgage fraud 25:552
Mortgage lenders 3:581
Mortgage lending 3:580
Mortgage market 3:581–583
Mortgages 20:13–18, 21:402–404
Mortification of the self 18:920–921
Mortuary ritual, changes in 5:872–873
MOS *see* Main olfactory system (MOS)
Mos Italicus 13:778–779
Mosaic, genetic 9:882–884
Mosaic faith 12:870–871
Mosaic law 3:32, 3:548, 13:799
Mosaic maps 24:265
Mosaic plots 10:354–357, 10:355f
Mosca, Gaetano 7:386, 14:423, 18:472–473
elites 7:391
works
Elementi di scienza politica 2:270
Moscovici, Serge 20:444
on social representations 16:862
Moscow Centre for Gender Studies 20:836, 20:838
Moscow Framework for Action and Cooperation 6:740–741
Moseley, Michael, works, *Maritime Foundations of Andean Civilizations* 23:37
Moshav 2:148
Moshav Shitufi 2:148
Mosher, A.T., works, Thinking about Rural Development 20:805
Mosher, C.M., on leatherman subculture 13:888
MOSLER *see* Multiple Objective Structured Long Examination Record (MOSLER)
Mosquitoes, and diseases 18:177
Most different systems–same outcome (MDSO) 4:441–443
Mosteller, Frederick 13:690, 13:692–694, 14:812
Most-favored-nation treatment 12:568
Mostly straight 9:652
Mother earth, and motherhood 15:881
Mother ground 5:510–511
Mother involvement in intact families, quantitative and qualitative assessments of 22:866–867
Mother relation 13:54
Mother-blaming, and child neglect 3:414
Mother-centered approach, early memory development 6:761
Mother–child communication 15:874–880
Cameroonian Nso beliefs 15:875
child development 15:874, 15:876, 15:878
in childhood 15:876–878
autobiographical narratives 15:877
cross-cultural and quasi-experimental studies 15:877
discursive practice approach 15:877–878
as cradle of cultural meaning 15:876
in early infancy 15:874–876
methodological considerations and future directions 15:878
mundane communication 15:876
self-narratives 15:877
shaming/threatening strategies 15:876
storytelling 15:877–878
Mother–child consensus 12:405
Mother–child interactions 6:749
Mother–child relationship
and child care 3:391
and divorce 14:522
Mother–child talk
and memory development 15:152–153
Mother–daughter communication 21:681
Mother–daughter industrial system 10:221
Mother–daughter interaction 9:700–701
Motherese 9:698, 13:308
and neonatal crying 16:480
Mother-figure 2:790–791
Motherhood 8:782–783, 13:53–54, 15:881–886
advocacy 15:885
birth control 15:883–885
changes in 8:840
child development 15:883
child protection 15:884–885
childbirth 15:881, 15:883, 15:885
feminism 15:882–885
future of 15:885
goddess 15:881
as ideology 15:882–883
images and myths of 15:881
and matrifocality 14:822–823
mother earth 15:881
mothering 15:882–885
patriarchy 15:881–883
policing 15:883–885
power 15:881, 15:885
religion 15:881, 15:883
reproduction 15:881–882
state control 15:882
welfare state 15:884–885
see also Fatherhood
Motherhood wage penalty 14:608
Mother–infant attachment 16:471–472
from historical and anthropological perspective 2:140–141
Mother/infant bond 2:791–792
Mother–Infant Bonding Questionnaire (MIBQ) 17:723–724
Mother–infant games 18:219
Mother–infant interaction 17:723–724
and neonatal crying 16:475–476, 16:480
Mothers
and child neglect 3:412–414
health and socioeconomic benefits of family planning among 8:789–790
parenting of 2:475
role in gender development 9:743–744
Mothers Against Drunk Drivers (MADD) 15:590
Mother's face, recognition of 8:652
Mother–son communication 21:681
Motion capturing devices 25:133
Motion pictures 9:166–167, 9:169
Motion verbs 23:92
Motionese 11:624–625, 11:625f
Motivated reasoning 10:706
Motivated Strategies for Learning Questionnaire (MSLQ) 13:715, 24:92
Motivated tactician 22:206
Motivation 1:241–242, 2:482, 4:146–149, 4:908, 6:207, 7:218, 7:276, 10:712–714, 12:380–381, 15:452–456, 15:909–912, 15:928, 15:936–946, 16:789, 17:337, 17:849, 17:858–860, 17:871, 21:49, 21:365, 24:184–190
academic achievement *see* Academic achievement motivation
academic self-concept and achievement 1:60
accounts of confabulation 4:561–562
neomotivational accounts 4:567–568
achievement goal theory (AGT) 15:929–931
action control 15:937–938
approach–avoidance theory of 15:943
assessment of 15:909–912
attribution theory 15:936
attributional approach 15:894–898

in Australian Aboriginal populations 15:899–906
autonomous 21:486–488
avoidance and approach 2:326–331
basic human needs 15:936–937
behavior
change 15:909–912
maintenance 15:909–912
brain stimulation reward 2:841–843
to care for others 7:542
child characteristics 15:921–923
child-specific beliefs 15:923–924
in clinical intervention 15:907
in clinical interventions 15:907–913
cognition–emotion–action theory of 15:897
cognitive processes of 21:554
in collectivism/individualism 4:208
component, emotion assessment from 7:523
conceptualizations 15:907
consequences of dimensional comparisons for 6:430–436
consumer psychology 4:739
contemporary issues 15:907
and control behavior 15:894–898
controlled 21:486
and creativity training 5:115
cross-cultural study of education 5:327, 5:329, 5:331
definition 2:326
definitions 15:907
developmental behavioral genetics and education 6:322–323
and dimensional comparison theory (DCT) 6:432
of drivers in transportation 24:625–626
emotion and 7:478
emotions 7:502–505
and enthusiasm to teach 24:45–46
and epistemological beliefs 7:887–893
expectancies 15:937
expectancy-value model 15:921
expectancy-value theory (EVT) 15:928–929
expectancy-value-cost model of 8:503–509
and expectation 14:86–87
external 15:907–908
extrinsic 21:487, 24:187
familial influences on 15:921–927
family characteristics 15:923
for father involvement 8:841
future directions 15:925–926
family diversity 15:926
holistic models of parenting 15:925–926
influence 15:926
and future orientation 9:563–564
conditions for prompting 9:565
content 9:563–564
extension 9:563
and future time perspective 9:564–565
conditions for prompting 9:565
effect on academic achievement 9:564
effect on achieved identity 9:565
effect on resilience 9:564–565
goal setting 15:937
goal striving 15:937–938
goals 15:937–938
and grit 10:400
and groups 10:437
Heider on 10:753
implementation intentions 15:938
incentives 15:937
individual 15:303
and instinct 12:175, 12:178–179
and instruction 15:928–935
and intelligence 12:292–295
and interests 12:389
internal 15:907–908
internalization of 21:487–488
intrinsic 21:487, 24:187
for language policy 13:355–358
in late 1970s 15:896–897
life course 15:942–943
and market allocation 14:541–542
material improvement 15:452
mental contrasting 15:938
migration history 15:446
in military combat 15:512
and organizational citizenship behavior 17:320
outcome expectancies 10:712–713
parents' behaviors 15:924–925
activity-specific parenting 15:925
general child-rearing climate and style 15:924
modeling 15:924–925
parents' general beliefs 15:923–924
personal projects 15:937
and personality 17:831–832
and practical reasoning 18:778
predictors of achievement (1960s decade) 15:895
priming 15:938
prosocial 19:217–218
psychology of 15:887–893
attributions 15:887–890
behavior change interventions 15:892
concept of goals 15:888
current goal research 15:892
dealing with hindrances 15:891–892
determinants of effective goal pursuit 15:888–889
dwelling 15:889–890
fantasies 15:889–891
fantasy realization theory 15:889
implementation intentions 15:891–892
impulsivity 15:892
indulging 15:889–890
instigating determined goal pursuits 15:889–891
intention–behavior relation 15:891
mental contrasting 15:889–890, 15:892
self-completion theory 15:889
self-efficacy 15:887–888
self-regulation of goal pursuit 15:889–892
strategic automaticity 15:891–892
quasineeds 15:938
refinement 15:895–896
reflection 15:895–896
and reinforcement 20:209
replication 15:895–896
rewards 15:908–909
risk management 15:452–453
risk perception 10:712
role in language acquisition 2:574, 2:576
in 1980s 15:896–897
in 1990s 15:897
in 2000s 15:897–898
self-beliefs 10:713
and self-concepts 21:469–472
self-control 15:908
self-determination theory (SDT) 15:931–932
and self-efficacy 21:505
self-efficacy beliefs 15:937
self-enhancement 6:432
self-evaluation 6:432
self-improvement 6:432
self-maintenance 6:432
self-regulation 15:936–938
sexual *see* Sexual motivation
significance in education and well-being 24:186–187
and social beliefs 22:163
social cognitive theory (SCT) 15:932–933
social connection 15:453
Sorokin on 23:23
symbolic gratification 15:453
teacher 24:64–71
teaching as career 24:65
and temperament, interaction between 24:187–188
threat evasion 15:453–455
tradition in 15:894
and vocational interests 25:241–242
volunteerism 25:268–272
and wanting/needing 25:351–355
willpower 15:938
for writing 25:768
in youth sport and physical activity, developmental perspectives on 15:914–920
Motivation and Opportunity as DEterminants (MODE) model 11:716
of attitude-behavior consistency 2:187–188
Motivation factors 16:512–513
Motivation theory 12:813
Motivation to learn 21:51, 24:245–246
evolutionary perspective 8:395–396
Motivational approach, cardiac defense 23:540
Motivational behavior 15:914
Motivational changes 15:373
Motivational climate 15:918
Motivational components
future orientation 9:562
test anxiety 24:244
Motivational cultural intelligence 5:433, 5:436, 5:438
Motivational development 15:946, 21:121–122
impact of schooling on 21:119
systems theory of 15:946–950

Motivational enhancement theory (MET) 1:502–503
Motivational factors, human factors and ergonomics 11:301
Motivational hub 25:681–682
Motivational intervention, academic motivation 19:794–795
Motivational interviewing (MI) 15:909–912, 17:594–596
  definition 15:910–911
  engaging 15:910
  evoking 15:911
  focusing 15:910
  foundations 15:911
  limitations of 15:912
  planning 15:911
  principles of 15:911
  technical/micro-skills 15:911–912
  theoretical model of 15:909
Motivational model of pain 3:557
Motivational psychology, perspective of procrastination 19:76
Motivational significance, cardiac defense 23:542–543
Motivational states 7:468–469
Motivational states, and cognitive dissonance 4:76–78
Motivational theories, engagement 1:32–33
Motivational theory of life span development (MTD) 14:104
Motivation-behavior gap 1:246
Motivations for helping 10:760
  in-group *versus* out-group members 10:760
  strategic helping 10:760
Motives 15:951–956, 16:512–514, 16:516
  achievement 15:952–955
  affiliation-intimacy 15:952–953
  agreeableness 1:463–470
  collective political action 4:146–149
    emotions 4:148–149
    identity 4:147–148
    ideology 4:148
    instrumentality 4:146–147
  congruence research 15:954–955
    in positive development 15:954–955
  developmental psychology 15:953
  dual-motives theory 15:951–952
  explicit 15:951–952
    congruence of 15:952–954
    determinants of congruence of 15:954
  happiness 15:951, 15:955
  implicit 15:951–952
    congruence of 15:952–954
    determinants of congruence of 15:954
    development of 15:955
  of interorganizational relationships 12:621–622
  power 15:952–955
  utilitarian social thought 24:981–982
  well-being 15:951–955
Motor accidents, and alcohol 1:510
Motor actions, emotion assessment through 7:523–524
Motor associative learning, role of cerebellum in 2:93–96
Motor control
  premotor cortex 18:847, 18:850
  Sherrington on 21:882–883
Motor control models 15:957–964
  body schema 15:960–961
  brain dealing with motor redundancy 15:959–960
    cost-function-based solution to DoF problem 15:959–960
    equilibrium-point hypothesis 15:959
    force-field-based solution to DoF problem 15:959
    optimal control theory 15:960–962
  internal model 15:957–958
  learning paradigms in neural networks and motor control 15:963–964
  mirror neuron circuit 15:959, 15:962
  motor imagery 15:959, 15:961–962
  motor redundancy and cybernetics of purposive actions 15:958–959
  passive motion paradigm 15:960–962
    networks 15:960–961
    and simulation theory of covert actions 15:961–962
  reinforcement learning 15:964
  supervised learning 15:964
  unsupervised learning 15:964
Motor corollary discharge systems 16:625
Motor cortex 15:965–970, 24:427
  definition 15:965
  evolution of 15:966–967
  function of 15:967–968
    muscles *versus* movements 15:968
    population coding 15:968
    voluntary and reflex movement 15:968
  primary
    connections of 15:965–966
    identifying features of 15:965
  studies 15:966
Motor development
  in infancy and early childhood 15:971–977
    after gait 15:976
    manual skills 15:974–975
    motor behavior 15:971–972
    muscle control 15:971, 15:975
    patterns 15:972–976
    postural control 15:972, 15:974–976
    prone and neck control 15:973
    quadruped position and crawling 15:975
    sitting position 15:974
    stand and walk 15:975–976
    supine and roll 15:973–974
    theories 15:971–972
  and self 21:439–440
Motor effects 19:466
Motor execution of actions 1:874
  gestures 1:875
Motor face 23:250
Motor function disturbances
  and prefrontal cortex injury 18:815
Motor imagery, motor control 15:959, 15:961–962
Motor images, neonatal imitation 16:485
Motor intelligence, neonatal imitation 16:481
Motor learning 15:968–969
  vestibulo-ocular reflex
    cellular mechanisms for 25:73
    multisite hypothesis of 25:72–73
Motor map 15:967
Motor neurons 2:28
Motor procedural skill learning tasks 16:845
Motor processing 13:399
Motor redundancy 3:778–779, 15:958–959
  dealing of brain with 15:959–960
Motor reflexes 16:703
Motor responses activation by higher cognitive processes 7:411
Motor skill learning, neural substrates of 16:848
Motor skills 19:69
Motor speech disorder 23:251
Motor theory 23:235–236
  of speech perception 11:324, 15:583
Motor topographic maps, in brain 24:427
Motorized transportation 23:846–848
Motor-kinesthetic phenomena 11:459–460
Motorola 11:830
Mott, Lucretia 22:443
MOTUS *see* Modular Online Time Use Survey (MOTUS)
Moulin v. France (2010) 18:228
Mounds 1:923
Mount Hira 12:731
Mountz, Alison, works, Seeking Asylum 13:477
Mourners, emotional response of 5:872
Mourning 2:531
  and bereavement 2:531
  death 5:870–872
  death/dying 5:865–866, 5:868
  in Soviet Union 20:823–824
Mousepox 3:280
Mouse-to-elephant curve *see* Kleiber's curve
MOUT *see* Military operations in urban terrain (MOUT)
Mouth opening, neonatal imitation 11:621
Move to Opportunity (MTO) experiment 8:561, 8:564
Movement building of social justice 24:495
Movement circles 5:468–469
Movement for Gay Liberation 13:860
Movement for the Restoration of the Ten Commandments of God 16:806–807
Movement intellectuals 21:177–178
Movement litigation 13:593
Movement of Landless People (*Movimento dos Sem Terra*) 17:667
Movement societies 22:507
Movement theory 16:1
  barriers framework 16:4
  and constraints in syntax 16:1–7
  constraints on movement 16:4–6
  cyclicity-based accounts 16:4
  early 16:2
  economy-based accounts 16:5–6
  formulations of movement rules 16:2–3
  generalizing movement operations 16:3
  headedness-based accounts 16:5

motivations for 16:1–2
trace-based accounts 16:5
upward movement requirement 16:2–3
Movements
cortical areas engaged in 5:21–29
dance 5:667–669
disorders, and alcohol 1:512
prospective control of 23:292
MoveOn 22:514
MovieLens site 25:270
Movimentization of politics 4:145
Movimento Sociale Italiano (MSI) 18:411–412
Moving average seasonal adjustment methods 24:349–351
Moving families 6:149
Moving to Opportunity (MTO) for Fair Housing Program 2:660, 4:326, 16:444–446
Moving window technique 8:632–633
Mowrer, Ernest, works, *Family Disorganization* 7:561
Mowrer, O.H.
two-stage theory 8:605
works
Family Disorganization 7:561
Moyn, Samuel 11:372
Moynihan, Daniel Patrick
on North American community 17:23
works
Beyond the Melting Pot 2:81
Mozambique
education in 7:200
peace building and music in 25:846
sociocultural aspects 23:75–80
MPB *see* Managed Professional Business (MPB)
MPC *see* Model Penal Code (MPC)
MPFC *see* Medial prefrontal cortex (MPFC; mPFC)
Mplus (software) 3:787–788, 23:380
MPM *see* Multidimensional process model (MPM)
MPQ *see* Multidimensional Personality Questionnaire (MPQ)
MPS *see* Material Product System (MPS)
MPs *see* Mainline Protestants (MPs)
MPT *see* Multinomial processing tree (MPT) models
MR *see* Mineralocorticoid receptors (MR)
MRCPsych (Member of Royal College of Psychiatrists) 7:191
MRI *see* Magnetic resonance imaging (MRI)
MRMs *see* Mixed Rasch models (MRMs)
mRNA *see* Messenger RNA (mRNA) sequences
MRS *see* Magnetic resonance spectroscopy (MRS); Marginal rate of subjective substitution (MRS)
MRT *see* Marginal rate of transformation (MRT); Middle range theories (MRT)
MRTS *see* Marginal rate of technical substitution (MRTS)
Mrvar, A., on negative social network ties 5:561
MS-13 (gang) 1:622–623
MSA *see* Ministry of State for Antiquities (MSA)
MSAs *see* Medical savings accounts (MSAs); Metropolitan statistical areas (MSAs)
MSC *see* Marginal social cost (MSC) curve; Math self-concept (MSC)
MSCA *see* Marie Skłodowska-Curie Actions (MSCA)
MSCAA *see* Medical Schools Council Assessment Alliance (MSCAA) (UK)
MSCEIT *see* Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT)
MSE *see* Mean squared error (MSE); Multiple systems estimation (MSE)
MSF *see* Medecins Sans Frontieres (MSF)
MSGrps *see* Mean squared groups (MSGrps)
MSI *see* Movimento Sociale Italiano (MSI)
MSL *see* Model of Strategic Learning (MSL)
MSLQ *see* Motivated Strategies for Learning Questionnaire (MSLQ)
MSM *see* Men who have sex with men (MSM)
MS/MS *see* Tandem mass spectrometry (MS/MS)
Msp Foundation 15:114–115
MSQ *see* Minnesota Satisfaction Questionnaire (MSQ)
MSS *see* Modern Sexism Scale (MSS)
MST *see* Multisystemic therapy (MST)
MSW *see* Master of Social Work (MSW) Program, in China
MSW-MPA *see* Master degrees in social work and public health (MSW-MPA)
MT *see* Machine translation (MT) module; Middle temporal (MT) area
Mt Pelerin Society 22:699
MTB *see* Minding the Baby (MTB)
MTCDE *see* Metric tons of carbon dioxide equivalent (MTCDE)
MTD *see* Motivational theory of life span development (MTD)
mtDNA *see* Mitochondrial DNA (mtDNA); Mitochondrial DNA (mtDNA)
MTEF *see* Medium-term expenditure frameworks (MTEF)
M-test 5:347–348
MTF *see* Monitoring the Future (MTF)
MTG *see* Middle temporal gyrus (MTG)
MTLs *see* Medial temporal lobes (MTLs)
MTM *see* Midterm memory (MTM)
MTMM *see* Multitrait-multimethod (MTMM) models
MTO *see* Moving to Opportunity (MTO) for Fair Housing Program
MTP *see* MyTeachingPartner (MTP)
MTT *see* Multiple trace theory (MTT)
MTUS *see* Multinational Time Use Study (MTUS)
Mu rhythm, in electroencephalography 7:357
μ-Argus (software) 4:579
Mubarak, Ali, works, *Historical geography of Egypt* 11:65
Mubarak, Hosni 16:778, 18:424
Mubarak regime 16:779
Muchow, Martha 19:669
Muftī 12:746t
Mugabe, Robert 25:838
Al-Muhajiroun 16:807
Muhammad, Danish 25:798–799
Muhammad, Prophet 12:728, 12:731–732, 13:801
biography of 11:63–64
and Islamic historiography 11:63–64
and Qur'an 11:63
Mujeres Libres 19:748
Mukerjee, Radhakamal, on sustainability studies 22:267–268
Mukhia, Harbans, on agrarian political economy 1:457
Mukhtar, T., on cultural participation 5:448–449
Mulkay, M., works, *The Word and the World* 20:94
Mullaly, Bob, works, *Structural Social Work* 8:152
Muller, David 16:679
Muller, F. Max 4:405
works, *Introduction to the Science of Religion* 4:405
Müller, Friedrich Max 7:375
Müller, Georg Elias 10:757
Muller, Herbert J., works, *The Uses of the Past* 16:13
Muller, Hermann Joseph 8:220, 8:224, 16:13–17
concern for radiation protection 16:15–16
contribution to evolutionary theory 16:14
contribution to genetics 16:14
and defense of science in USSR 16:14–15
eugenics movement 16:14
field of radiation genetics 16:13–14
positive eugenics program 16:15
status after his death in 1967 16:17
views on world view of science 16:16–17
works
The Masses 16:14
Out of the Night 16:14
Psychophysics Beyond Sensation 19:475
Science 16:13–14
The Spark 16:14
Müller, Karl H., on epistemic cultures 7:876
Müller, Max 20:344
Muller v. Oregon 13:767
Muller's ratchet 16:14
Müller-Schumann Law of Associative Inhibition 19:1
Mullins, Nicholas 21:130, 21:132
Multi geographic models 14:787
Multiagent stimulation 4:486–488
elements in environment 4:487–488
representations of space 4:487
representations of time 4:486–487
Multiagent systems 2:310
autonomous agents 2:310
as models of the behavior of social systems 2:312
Multiarm trials 4:453–454

Multiattribute linear ballistic accumulator (MLBA) 5:904
Multicausality, educational sociology 7:292
Multi-choice game 4:879
Multicohort studies, life course epidemiology 14:73
Multicomponent drug use prevention program 6:682
Multicomponent latent trait model (MLTM) 8:700
Multicriteria decision analysis (MCDA) 23:138–139
Multicultural capital 5:451
Multicultural citizenship 6:594
Multicultural citizenship education 16:19–20
Multicultural counseling 5:71
Multicultural education 8:132–133
  academic achievement gap between ethnic minority and majority group students 16:19
  across nations 16:20
  balancing unity and diversity 16:20
  in East Asian Nations 16:19
  goals and scope of 16:18
  in Western Europe 16:18–19
Multicultural gourmets 11:452
Multicultural nation-states 16:19–20
Multicultural psychology 5:65–66
Multicultural social work 8:153
Multicultural society 13:550
Multicultural teams 5:339
Multiculturalism 1:628, 3:664–665, 6:594, 11:546, 11:678–679, 12:363, 16:22–27, 22:321, 24:779
  anthropology of 16:28–33
  in Australia 2:240–242
  in Central America 1:624–625
  and commemoration 4:239
  cosmopolitanism 16:24–26
  and cultural rights/defense 5:494
  and culture 5:586
  democracy 16:22, 16:24, 16:26
  and ecumene 7:111
  and egalitarianism 7:311
  equality 16:22, 16:26
  and ethnicity 8:158–159
  ethnicity/migration 8:142–145
  and feminist psychotherapy 8:929–933
  future of 16:26
  and gay/lesbian activism in Latin America 13:862
  and globalization 16:31
  history of 16:28–29
  of honor 11:183–184
  and hybridity 11:448–452
  and identity movements 11:546
  and immigration 11:672, 11:678–680
  and immigration policy 11:665–666
  and integration 12:237–240
  justice 16:22, 16:24, 16:26
  Lévi-Strauss on 13:908
  liberal 14:15
  liberalism and 14:24
  migrants in Europe 16:30–31
  migration 15:424–425
  minorities 16:23–25
  nationalism 16:22, 16:25
  New World countries 16:30
  and normativity 16:31–32
  in North America 17:24–25
  pluralism 16:24
  politics of 16:24–26
  postcolonial societies 16:30
  relevance of culture 16:22–24
    cultural rights 16:23–24
    diversity 16:22–23
  and social identity 22:321
  and sport 23:316
  in Taiwan 5:454–455
  and Walzer, Michael 25:349
Multiculturalism studies, *versus* intersectionality 12:660–661
Multiculture 12:238
  use of term 12:237
Multidimensional Anxiety Scale for Children (MASC) 2:478
Multidimensional citizenship education 22:730–731
Multidimensional city-block scaling 19:432
Multidimensional construct, engagement 1:33
Multidimensional continuum orientation 15:618
Multidimensional data, construction of 6:438
Multidimensional detection theory 21:950
Multidimensional intelligence, cultural intelligence as 5:433
Multidimensional item response theory (MIRT) 19:438
Multidimensional matching, trade-offs in 2:103–104
Multidimensional model 21:363
Multidimensional Personality Questionnaire (MPQ) 4:660, 9:1010, 16:743–744, 17:917
Multidimensional poverty 18:755
Multidimensional process model (MPM) 13:591, 13:591f
Multidimensional scaling (MDS) 4:82, 5:5, 9:892–893, 16:34–47, 19:79, 19:420–421, 21:579–580, 21:583, 24:386, 21:579f
  analysis 9:892–893
  classical 9:889, 9:892–893
  concept of 16:34
  confirmation of 16:45–46
  data in 16:42–43
  data transformations 16:35–36
  dimensionality 16:43–44
  dimensionality selection 16:36
  dissimilarities 16:41–42
  distance models 16:34
  fitting criteria 16:36
  formalization 16:41
  history of 16:42
  individual differences 16:37
  interpreting MDS solution 16:44
  optimal transformations 16:41–42
  practical details 16:44–45
  proximity data 16:35–36
  scale levels 16:36
  similarities 16:41–42
  software 16:46
  software for 16:38–39
  stress values 16:43
  three-way MDS 16:46
  unfolding 16:46
  unfolding analysis 16:37–38
Multidimensional self-concept 1:54–55
Multidimensional space 4:441–442
Multidimensionality 4:441
  of academic self-concept 1:57
  of health behaviors 10:588
  of prosocial behavior 19:235
  of self-concept 12:425
Multidirectional career 3:130
Multidirectionality
  in adult cognition 7:138
  in development 15:372–373
Multidisciplinary science 23:134–135
Multidisciplinary single shared assessments 12:503–505
Multidisciplinary theories 5:233
Multiepisode models 14:54
Multi-Ethnic Placement Act 1994 1:146
Multifaceted events 15:793
Multifaceted universe 9:835
Multifactor Leadership Questionnaire (MLQ) 13:651
Multifactorial designs, in quantitative linguistics 19:727–728
Multifactorial significance tests 19:727–728
Multifaith congregations 15:726
Multi-family therapy 8:745–746
Multifirm network organization 16:549
Multifunctionality, peasant agriculture 17:665–666
Multigenerational families 9:854
Multi-goal analysis 18:275–277
  goals identification 18:275–276
  goals justification 18:275–276
  policy comparison 18:277
  policy recommendation 18:277
  specification of impact categories 18:276
  specification of policy alternatives 18:276–277
Multigroup comparisons 5:324
  of confirmatory factor analytic models 5:323–324
Multilanguage corpus 4:950
Multilateral economic influences of repression 18:451
Multilateral trade negotiations 12:571, 12:569t
Multilayer backpropagation neural network 23:415
Multilayer feedforward artificial neural network 23:411
Multilayer networks, concept learning 4:655–656
Multilayered governance 10:308
Multilayered social–ecological framework, honor killing 11:187–188

Multilevel causal model of institutions 22:980–981
market mechanism 22:981
mimicking 22:981
mutual monitoring and enforcement 22:981
random mutations institutional innovations 22:981
Multilevel community-based health interventions 10:646–647, 10:651
Multilevel emotions model, in workplace 7:507
level 1 7:507–508
level 2 7:507–508
level 3 7:507–509
level 4 7:507, 7:509
level 5 7:507, 7:509–510
Multilevel interventions during extreme weather events 10:389–390
Multilevel methods
statistical analysis 23:381–386
binary response example 23:384
coefficient variance estimates adjusted for clustering 23:382–383
fixed effects models 23:382
Gaussian example 23:383–385
hierarchical data structures 23:381
marginal models 23:383
omitted variables 23:383
random coefficient models 23:381–382
robust estimator 23:382
Multilevel modeling 4:489, 7:226, 10:86, 19:727–728
health 22:278–279
social structure and personality 22:541
Multilevel models
in demography 16:48–56
Multilevel models in demography 16:48
of contraceptive use 16:48–52
contextual variables 16:49
cross-level interactions 16:50
empirical Bayes 16:51
fixed coefficients 16:50
level-specific models 16:49
predicted probabilities 16:51–52
random intercepts 16:49
random slopes 16:49
variance parameters 16:50–51
between and within-group estimates 16:50
of infant and child survival 16:52–55
conditional probabilities 16:54–55
fixed coefficients 16:53
intraclass correlations 16:55
marginal probabilities 16:54–55
random effects 16:53–54
three-level hazard model 16:52–53
Multilevel models of educational effectiveness 7:218–219
Multilevel opportunity theory 20:777–778
Multilevel organizational level analysis, negotiated orders 17:414
Multilevel random coefficient modeling (MLM) 22:550
Multilevel research, abusive supervision 23:705–706
Multilevel selection theory 15:813
Multilingual language development 16:57–64
bilingualism 16:57–62
biological and neurological aspects of 16:62
code switching and code mixing 16:60
disability factors 16:62
environmental factors 16:61
individual and developmental factors 16:62
interaction of bilingual factors 16:57f
language deceleration 16:59
language dominance and proficiency 16:58
language fossilization 16:61
language interference 16:60
language loss 16:61
language maintenance 16:58–59
language mode 16:59, 16:59f
language transference 16:59–60
second/multiple language acquisition 16:57–58
strategies for promoting 16:62
ultimate attainment 16:60–61
Multilingual society, and illiteracy 11:589
Multilingual speakers 16:58–59
Multilingualism 2:580–581, 2:584, 6:506, 16:65–71
assessment of multilingual proficiency 16:66
attitudes toward 16:65
challenges for future research 16:70
cognitive aspects of 16:67
connectivity between languages 16:67
definitions and terminology 16:65–66
dynamic systems theory/complexity theory (DST/CT) perspective 16:69
dynamics and complexity of 16:69–70
English as contact language in 16:68–69
and foreign language learning 9:329–330
functionality of language in 16:68
intelligence and/or age, linked to 16:67
and linguistic anthropology 14:173
prior language knowledge and awareness, effect of 16:69–70
social contexts 16:68–69
Multilocale ethnographies 21:182–183
Multimale groups, advertised ovulation in 18:899
Multimedia 3:181
learning materials 3:181
sources 3:181
Multimedia cases
Computer Assisted Instruction (CAI) 4:494
Multimedia Super Corridor project 24:897
Multimethod strategy, interviewing in 5:855–856
Multimodal Emotion Recognition Test 8:678
Multimodal emotional expressions 22:559
Multimodal information, perceptual learning 10:129–130
Multimodal literacy 14:246–250
Multimodal prosopagnosia 19:252–254
Multimodal semantics 21:576
Multimodal speech perception 23:240–241
Multimodal therapies
broad-spectrum approach 16:72
clients 16:72
cognitive restructuring 16:72
contraindications 16:72–73
seven-point perspective 16:72
test anxiety 24:248
theory and mechanisms 16:73
Multimodal transportation 24:617
Multimodality 18:800
oral and literate expression 17:277
Multimodeling 4:488–489, 4:489f
Multination country, China as 2:66
Multinational corporations (MNCs) 4:947–948, 10:219, 10:221, 10:254, 12:481–482, 12:484, 12:580–586, 16:74–80, 22:172
birth of 16:75–76
in development 6:276–277
diffusion of organizational knowledge 16:77–78
domestic competition 16:78–79
early history 16:74–75
economics and politics of 16:78–79
and foreign direct investment 16:74, 16:77–79
globalization and digital information economy 16:79–80
and imperialism 11:683–684
information technologies 16:79–80
institutions and 16:79
knowledge transfer 16:79
leadership in 10:413–414
national imprinting 16:79
national origins and institutions 16:76
organization of 16:76–77
outsourcing service providers 3:47
specialist 3:47
subsidiaries 3:46–47
transfer price 16:76
Multinational diversity 16:23
Multinational enterprises (MNEs) 12:484–487
foreign direct investment 9:324, 9:326
Multinational states, and imperialism 11:684
Multinational Time Use Study (MTUS) 5:733, 5:746, 24:375, 24:383–384
Multinational trade and treaty groups 15:72
Multinational troops 17:634
Multinomial distribution 6:583, 16:117
loglinear models 16:125–126
Multinomial logit (MNL) model 5:902, 14:866
best–worst scaling 2:550
Multinomial models 19:431–432
tests for 23:397–398
Multinomial probit (MNP) model 14:866
Multinomial processing tree (MPT) models 6:525–529
application areas and modeling issues 6:528–529
examples of 6:527–528
specification of 6:525–527

Multinomial processing tree (MPT) models (*continued*)
statistical inference for 6:528
task network models 16:543
Multiobjective decision making 23:138–139
Multiobjective land allocation (MOLA) 23:138
Multiparty system 17:567
Multiple abstract variance analysis (MAVA) design 3:244–245
Multiple access measures 1:66
Multiple adjustments, infancy as period of 12:9
Multiple agents, in supply chain management 23:710–711
Multiple assignment measures 20:154, 20:156–157
Multiple baseline design (MBD) 8:346, 21:985, 21:987
for workplace safety and health 25:731
Multiple category membership, in popular music 3:802
Multiple choice paradigm 14:804
Multiple communities 10:650
Multiple comparison groups, nonequivalent group designs 16:853
Multiple comparisons (MC) 9:529, 11:501, 16:81–87
adjusted p-values 16:84–85
Bayesian methods 16:85–86
Benjamini–Hochberg FDR-controlling method 16:84
false discovery rate (FDR) 16:82–83
family of inferences, prespecification 16:82
familywise error rate (FWER) 16:82–83
FWER-controlling methods 16:83–84
NAEP Trial State Assessment 11:502
Multiple comparisons procedures (MCPs) 16:82
Multiple components of emotions 7:526–527
Multiple contingency 23:610
Multiple correspondence analysis (MCA) 5:4, 5:6, 9:143, 16:123, 22:336
Bourdieu's use of 5:8–9
future developments and related techniques 5:9–10
indicator matrix 5:8
methodological background 5:8
principles of 5:6–8
Multiple cutoff points 20:154
Multiple determination 11:500
NAEP Trial State Assessment 11:501–502
Multiple determinism 22:486
Multiple diagnosis 6:386
Multiple distributions, Bayesian experimental designs 8:523
Multiple elitism, and political power 18:761
Multiple equilibria
and fertility 9:9
and rational expectations 8:513
Multiple ethnic identities, complex structure of 8:115
Multiple events, analysis of 8:295
Multiple failure, survival analysis 23:780–781
correlated failure times 23:780
frailty models 23:780–781
recurrent events 23:780
types and competing risks 23:780
Multiple frames of reference 17:703–707
encoding of spatial information by parietal neurons in 17:705–706
Multiple imputation 15:604–605, 16:88–93
applications in social-science research 16:90–92
comparability of industry and occupation codes over time 16:91
consumer expenditure survey 16:91–92
drinking behavior with follow-up for nonresponse 16:91
handling 'don't know' survey responses in Slovenian plebiscite 16:92
missing data in national health survey 16:91
survey of consumer finances 16:92
creating 16:89–90
choice of M 16:90
incorporating proper variability 16:90
iterative simulation techniques 16:90
modeling issues 16:89–90
data set, analyzing 16:89
inferences for scalar Q 16:89
significance tests for multicomponent estimands 16:89
secondary user 16:88
theoretical motivation for 16:88–89
Multiple Indicator and Multiple Causes (MIMIC) approach 12:52–53
Multiple indicators 11:751
Multiple intelligences 12:306
Multiple investigators 24:643
Multiple job holding 17:666
Multiple linear regression 14:852
Multiple literacies 1:731–732
Multiple logistic regression 2:590
Multiple memory systems 11:710, 25:712–713
and proactive interference 19:1–2
Multiple Mini Interview (MMI), in psychiatric education 7:188
Multiple modernities 15:693–694
Eisenstadt's contributions 7:321–322, 7:325
and gender, in Middle East 16:369–370
Multiple narratives 16:212
integration of 16:214
Multiple negation (grammar) 13:283
Multiple networks, in organizations 2:770–771
Multiple Objective Structured Long Examination Record (MOSLER) 7:190
Multiple observers of behavior 20:251
Multiple operationalizations 24:643
Multiple oppressions 6:448
Multiple outcome measures, nonequivalent group designs 16:853
Multiple outliers 23:389
Multiple paternity
benefits of 18:900
cost of 18:900–901
Multiple personality disorder
see Dissociative identity disorder
Multiple personas principle 17:910
Multiple perspectives, organizational design and form 6:237
Multiple populations 24:643
Multiple predictor variables, nonparametric regression with 16:874
Multiple raters 20:251
Multiple recaptures 18:603–604
Multiple regression 14:155–156
for moderation 15:673
Multiple sclerosis 12:420
Multiple seizures
consequences of 13:35
generators 13:34
Multiple selves 21:452–453
and self-development 21:493
Multiple sexual partnerships 21:798
Multiple sites 24:643
Multiple social network 22:478
algebraic structure of 22:478–481
Multiple sources of evidence 3:194–195, 3:199
Multiple stimuli varying along one dimension 21:950
Multiple systems estimation (MSE) 22:45
Multiple testing, for significance 21:961–962
Multiple time points, nonequivalent group designs 16:853–854
Multiple trace theory (MTT) 24:560
Multiple trace theory (MTT), of episodic memory 7:858–859
Multiple validities 15:621
Multiple views 6:716–717
Multiple visual areas 25:186
Multiple visual pathways 2:163
Multiple-block Metropolis–Hastings algorithm 15:764
Multiple-choice questions, in economic education 24:86–87
Multiple-decrement tables 14:91–92
Multiple-elitist 18:761
Multiple-factor analysis 14:801
Multiple-interval tasks 21:949–950
Multiple-item tests 20:248
Multiple-level games 24:488–489
Multiple-threshold theory 21:955
Multiplicative drift rate variability 5:903–904
Multiplicative model 8:297, 8:303
Multiplicativity condition 21:606
Multiplicity 17:305
of hypothesis tests 11:500–505
oral and literate expression 17:276–277
and postmodernism 18:687–688
of social identities 22:321–322
Multiplism 24:639
Multipolarity
of Chinese migration 12:527

and realism 20:27–29
Multiprocess theory 19:273
Multiprofile case 2:549
extensions of models to 2:551
Multiregional system, population momentum in 18:577
Multiscale governance fix 20:564–565
Multi-Scale Integrated Analysis of Societal and Ecological Metabolism (MuSIASEM) 22:258
Multisensory aspects of tactile processing 24:17
Multisensory integration 16:94–102
among different species and structures 16:101
computational approaches to 16:101
cross-modal cues 16:95–97, 16:99–101
development and plasticity in 16:100–101
multiple sensory representations in brain 16:94–95
in perception 16:95
during self-motion 25:68
sensor fusion 16:99
and superior colliculus 16:95–99, 16:96f
essential circuit in 16:99–100, 16:99f–100f
inverse effectiveness 16:99
spatial 16:96–97, 16:97f
temporal 16:97–99, 16:98f
Multisensory maps 24:11–12
Multisensory processing, contribution to experience of empathy 7:544b
Multisided markets 3:40
Multisite hypothesis of VOR motor learning 25:72–73
Multisite studies, controlled trials 19:387–388
Multisited archaeology 16:104
Multisited ethnography 13:540
Multisited field studies 16:103–108
anxieties and comforts 16:106–108
collaboration 16:104–108
complicity 16:107
depth 16:106
ethnography 16:103–107
fieldwork 16:104, 16:106–107
following 16:104, 16:106–107
globalization 16:105–106
holism 16:104–105
of Marcus 16:103–108
methodological and intellectual background 16:104–106
migration 16:103–106
scale 16:105, 16:107
siting 16:104, 16:106–107
social space 16:105–107
spatial turn 16:105
thick description 16:106
transnationalism 16:103–106
world systems 16:103, 16:105
Multisource appraisal technique 17:717–718
Multisource feedback 17:717–718
Multispectral scanners 20:419
Multistage designs 13:386
Multistage stochastic optimization 23:477–479
bandit problems 23:477
convex multistage problems 23:478–479
Markov decision processes 23:478
Multistage tests, computerized test construction of 4:512–513
Multistakeholder codes 10:257
Multistate models 8:298–299, 14:54
for event history analysis 8:311–313
conditioning 8:317
hypothetical calculations in 8:315–316
Multistate private-law disputes 13:615
Multistate transition models (in demography) 16:109–115
multistate life tables 16:110–114
for Cohort 16:110–112
in a period 16:112–113
statistical approaches to 16:113–114
multistate population models 16:114
multistate stable population 16:114
multistate stationary populations 16:114
without age 16:109–110
Multisystemic therapy (MST) 20:194
Multitasking 3:51
and time use 24:375
Multitrait-multimethod matrix 24:642
Multitrait-multimethod (MTMM) models 20:241, 20:243
Multi-Unidimensional Pairwise-Preference (MUPP) model 17:845–847, 17:847f
Multivariate adaptive regression splines (MARS) model 16:118, 16:874–875
Multivariate analysis 16:116–120, 17:662
discrete variables 16:121–136
Pearson's 17:662
Multivariate ANOVA (MANOVA) 13:690
Multivariate densities, probability density estimation 19:31–32
Multivariate discrete distributions 6:583
Multivariate displays 10:354–355
categorical variables 10:354–355
continuous variables 10:354
projective displays 10:355
Multivariate distributions 6:578–579
equality of 16:889
Multivariate gender differences 9:751
Multivariate growth curve analyses 10:445
Multivariate meta-analysis 15:278–280
combining correlated effect sizes within studies 15:278–279
multivariate mixed models 15:279–280
covariance components estimation 15:279–280
regression coefficients components estimation 15:279–280
robust variance estimator 15:280
Multivariate multiple-regression problem 19:79
Multivariate normal (MVN) distribution 7:116, 8:716
Multivariate pattern analysis (MVPA) 16:631
Multivariate statistical analysis 11:217–218
hierarchy of correlational measures 11:217
matrix calculations 11:218
Pearson's 17:661–662
principal components and canonical correlations 11:217–218
relations between variables 11:218
Multivariate survival function 15:845
Multivariate time-series models 24:338
Multivariate twin study 24:692
Multivariate two-sample location problem 16:895–896
Multivocality, and symbolism 23:860–861
Multivoxel pattern analysis (MVPA) 5:23, 9:530
for cortical area studies 5:23, 5:25
Mulvey, Laura, works, *Visual Pleasure and Narrative Cinema* 2:218
Mumbai, social movements in 22:854
Mumford, Lewis 16:137–141, 20:134, 21:201, 16:137f
on architecture 16:137–138, 16:140
on art 16:137–138
critical assessment 16:140
on garden cities 16:137–138
on globalization 16:139
on interventionism 16:139
life and work 16:137–140
background and intellectual development 16:137–138
early career 16:138
late career 16:139–140
mature career 16:138–139
literature 16:137–138
on pacifism 16:139
on politics 16:137, 16:139
on regionalism 16:137–138, 16:140
on technology 16:137–140
on urbanism 16:137–140
on utopias 16:137–139
works
The Brown Decades 16:138
City Development 16:138–139
The City in History 16:139
The Condition of Man 16:139
The Conduct of Life 16:139
The Culture of Cities 16:138–139
Faith for Living 16:139
From the Ground Up 16:138
The Golden Day 16:138
Green Memories 16:139
Herman Melville 16:138
The Highway and the City 16:138
In the Name of Sanity 16:139
Men Must Act 16:139
Mumford on Modern Art 16:138
Myth of the Machine 16:139
The Pentagon of Power 16:139
Sidewalk Critic 16:138

Mumford, Lewis (*continued*)
Sticks and Stones 16:138
The Story of Utopias 16:137
Technics and Civilization 16:138–139
Technics and Human Development 16:139
The Transformations of Man 16:139
The Urban Prospect 16:139
Münchhausen Trilemma 1:496, 1:498, 5:275
Muncie 3:202–203
Mundane reasoning, Harold Garfinkel on 9:600
Mungan, Murathan 20:870–871
Mungo 3 2:229
Munich, appeasement between Hitler and Chamberlain in 1:826–827
Munich agreement of autumn 1938 22:841
Munich vulnerability study, for sleep disorders 22:82–83
Municipal databases 10:61
Municipal Elder Law 13:423–424
Municipal legal systems 12:509
Municipal Research Bureaus 22:698
Municipalization of violence 24:783
Munif, Adbelrahman, works, *Cities of Salt* 16:385
Munn, Norman
works
Handbook of Psychological Research on the Rat 7:591
An Introduction to Animal Psychology 7:591
Munsell color chips 4:228
Munsterberg, Hugo
applied social research 1:854
Muntada Jensa 21:794
MUPP *see* Multi-Unidimensional Pairwise-Preference (MUPP) model
MUPS *see* Medically unexplained physical symptoms (MUPS)
Muqaddimah 11:62
Muratori, Ludovico, works, *Scriptores rerum italicarum* 11:72–73
Murder 6:633, 6:635, 10:6
of honor 11:183–184
honor killing 11:185, 11:188–189
Muren, Zeki, homosexuality of 20:870
Murik cultural intention 3:454–455
Murnau, F.W., works, *The Last Laugh* 9:167
Murray, Charles A. 8:130–131, 18:736
works, *The Bell Curve* 7:590, 12:311
Murray, Clark, works, *Handbook of Psychology* 7:591
Murray, Henry, works, *Explorations in Personality* 15:936
Murray, Martin 25:97
Murray, Stephen O., works, *Islamic Homosexualities: Culture, History, and Literature* 21:751
Murrow, Edgar R. 17:584
MUS *see* Medically unexplained symptoms (MUS)
MUSA *see* Metropolitan Urban Service Area (MUSA)
Muscarinic M3-receptor antagonism 21:22–23
Muscle control, in infants 15:971, 15:975
Muscles *versus* movements 15:968
Muscular fatigue 10:481
Musculature of eyes 8:628
Museon Arlaten 1:755
Museum of Ethnography (Kiev and Budapest) 1:755–756
Museums 2:21–22, 5:702, 14:766–767, 16:142–148
and anthropology 1:752–756
art 14:566
marketization 14:566
changes in twentieth century 16:144–146
emergence of universal survey museum 16:145
functions 16:144–145
funding 16:145
new museology 16:145–146
collecting for, and African Studies 1:277–278
collective memory 4:201, 4:203
contemporary museum 16:146
human rights 16:146
and social change 16:146
educational value 16:143–144
education as power 16:144
museum buildings 16:143
relationships with public 16:143–144
evolution of 16:142–143
in eighteenth century 16:142–143
in nineteenth century 16:143
future of 16:146–147
of Indigenous collection 11:774
model of 2:22f
origins of 16:142
and social and behavioral science research 22:617
sport 23:308
Mushroom body (MB) 15:156
MuSIASEM *see* Multi-Scale Integrated Analysis of Societal and Ecological Metabolism (MuSIASEM)
Music 4:235, 16:155–158, 23:639–640
African 1:278
anthropological aspects 16:149–154
archeology and iconography 16:156
carnival 3:145–148
characteristics of 16:155
computational auditory scene analysis 2:226–227
and cultural production 5:608–611
as culture 16:149
discipline and its methods 16:156–157
boundaries 16:156
developments 16:157
'relative' autonomy 16:156–157
as discourse 16:150
evolutionary group dynamics 8:444
and globalization 16:151–152
history and representation 16:155
and identity 16:150–151
identity exchanges and modern mythology 16:157–158
and knowledge 16:152–153
and media 16:152
as mediation of social process 16:149
and modernity 16:151–152
music history and the beginning 16:155–156
'other' musics 16:149
and peace building 25:842–848
performance 16:150
popular *see* Popular music
relative autonomy 16:156–157
Roma people 20:750–753
and social networks 16:556–557
as sonic practice 16:153
and technology 16:152
transitory nature 16:155
Music industry, new technologies 14:564, 14:567
Musical communication 23:296
Musical meaning 16:150
Musical notation 16:155
Musical notes 14:193
Musical taste 21:618
Musicology 16:155–157
Muskets, warfare 25:416–417
Muskogean peoples 8:176
Muslim Brotherhood 12:728, 12:751, 20:398, 20:400, 24:602–603
Muslim diaspora 6:355
Muslim girls 9:564
in Israel 9:564
Muslim law 4:264, 13:619
Muslim League 4:264, 20:375
Muslim legalities 12:746, 12:751
Muslim military traditions 12:727–728
Muslim minorities, in Europe
majority attitudes and intergroup threat 8:144
migration and integration policies 8:143
Muslim minority populations 12:729
Muslim movements 20:375
in India 20:375
Muslim nationalism 16:280–281
Muslim publics 20:311
Muslim reformism 12:741–742
Muslim tyranny 11:79
Muslim women 9:782
in western world 9:713
Muslim World League (*Rabitat al-'alam al Islami*) 24:605
Muslims 12:241
and communalism 4:263–265
equality within families 20:319
immigrants 20:339
and morality 15:808–809
online communities and spaces 20:273–274
OnIslam 20:273–274
social work *see* Islamic social work
in Western Europe 12:237–239
Mussolini, Benito 4:151–152, 16:257, 21:369–370, 22:444, 24:448–450
MUT *see* Moral Judgement Test (MUT)
Mut'a 21:793
Mutants, learning/memory 9:906
Mutation 9:915–918, 11:289
evolutionary demography 8:413–417
genetic programming 2:41–43, 2:42f

microevolution 15:338–339
Muller's contribution to 16:13–17
Mutation féodale 9:113
Mutation–selection balance (personality) 17:902–903
Mutatis mutandis 9:465, 12:152, 16:267
Mutatis mutandis 16:267
Mutual adaptation 4:299
Mutual advantage, and contractarianism 4:803–804, 4:806–807
Mutual affection 9:448
Mutual aid, and anarchism 1:680–682
Mutual aid self-help (MASH) groups 10:688–690
attraction and attrition 10:688–689
outcomes 10:689–690
process 10:689
Mutual beliefs 21:877–879
fixed-point account 21:877–878, 21:880
iterative account 21:877–880
and level problem 21:879–880
reflexive account 21:877–878, 21:880
as shared we-belief 21:880
Mutual benefit, kin selection 13:28t
Mutual dependencies, of variables 7:613–614
Mutual exclusivity assumption, in vocabulary acquisition 25:228
Mutual gaze 10:270
Mutual goals 21:113
Mutual help, and relational social work 20:223
Mutual intergroup differentiation model (MIDM) 12:411
Mutual mistake of fact 3:28
Mutual specificity in caregiver–toddler play similarities 18:221
Mutual support between generations 12:406
Mutual violent resistance 12:686
Mutual-face concern 11:186
Mutualism 1:579–580, 11:260
Mutuality 17:619–620, 19:154–155
Mutuality of influence, in acculturation 1:68
Mutuality principle 12:210–211
Mutually valorizing cultural symbols 9:595
Mutz, Diana, works, *Impersonal Influence* 14:713
Muxe of Oaxaca, Mexico 13:858
Muxes 13:858
MVN *see* Multivariate normal (MVN) distribution
MVP *see* Mentors in Violence Prevention (MVP) program
MVPA *see* Multivariate pattern analysis (MVPA); Multivoxel pattern analysis (MVPA)
MVUEs *see* Minimum variance unbiased estimates (MVUEs)
MWM *see* Morris Water Maze (MWM)
MWO *see* Mental Welfare Office (MWO)
My Class Inventory (MCI) 3:825–826
Myanmar 23:63
early states 23:67
fertility transition in 9:85–89
metal age 23:66
sociocultural aspects of *see* Southeast Asia, sociocultural aspects of
see also Burma
Myanmar Buddhists 3:532
Mycobacterium leprae 18:175
Mycobacterium ulcerans 7:815
Myelination 1:595, 4:59, 18:819
in brain development 2:824–825
Myers, Ronald 23:292
Myers' blended index 6:128–130
Myocardial infarction (MI) 6:209–210, 22:735
coping with 3:111–114
and coronary heart disease 4:917–919, 4:921
Myopathy, and alcohol 1:511
Myopia of monopoly, anthropological writing 1:740
Myopic behavior 17:783
MySpace profiles 24:149
Myrdal, Alva 16:159–164
as ambassador for peace and security 16:162–163
on child care policies 16:162
on conflict research 16:159, 16:162–163
on disarmament 16:159, 16:162–163
as educationalist for democracy 16:161–162
on feminism 16:159–160
on international diplomacy 16:159
preschool education of 16:160
on social justice 16:159–163
as social scientist for international cooperation 16:162
on sociology of employment 16:162
on sociology of women 16:162
as welfare theorist for families 16:160–161
works
Nation and Family 16:161
Myrdal, Gunnar 16:165–167, 20:540, 21:169
activities after World War II 16:166–167
aspects of black life 16:166
early life and career as an economist 16:165–166
expectations, notions of 16:165
on moral dilemma 16:166
on nationalism 16:310
political activities of 16:166
on values influencing social science 16:165
works
An American Dilemma 16:166, 20:540, 21:170
Asian Drama 16:166–167
Kris i befolkningsfrågan 16:166
Objectivity in Social Research 16:167
Myriad Genetics 17:588–590
Mysticism 16:168–171
and asceticism 2:62–63
and asceticism, distinction between 2:62–63
diversity in 16:168
experiences 16:170
Greek or Hellenistic Mystery 16:168
mystical experiences 16:168
mystical traditions 16:168
prophetism 16:169
psychological transformations 16:170
and religious experience 20:363
religious reveries 16:170
in scripture-based religions 16:169–170
shamanism 16:169
spirit medium 16:169
trance state 16:170
use of term 16:168
Whorf on 25:577–578
Mystification, ritual 20:718
MyTeachingPartner (MTP) 3:837
Mythical history 11:54–56
Mythologiques 23:627–628
Mythology, in South America 23:47
Native American tricksters 24:649–652
and structuralism 23:627–630
Mythopoesis 2:731
Myths 16:172
about Roma 20:755
Barthes on 2:358–359
ethnopsychology 25:793
function 16:172–177
Lévi-Strauss on 13:906–907
of motherhood 15:881
organizational 16:179–182
origin 16:172–173
in religion 16:172–178
as both other than explanatory and other than literal 16:176–177
bringing myth back to the world 16:177
make-believe 16:177
Malinowski science 16:174–175
as more than or other than explanatory in function 16:173–175
as other than literal in meaning 16:175
as 'primitive' counterpart to science 16:173
recent theories 16:177
self 16:176
of rights perspective, towards social change 13:545
ritual 20:717
and ritual 20:314
MZ *see* Monozygotic (MZ) twins

N

N100 9:481
N400 9:482
to target words illustrating effects of 9:484f–485f
N.W. Ayer & Son 1:189–190
N1 (event-related potential) 2:807–808
N170 (event-related potential) 8:664–665
N400 (event-related potential) 16:710–712, 16:712f
NA *see* Negative affectivity (NA); Trait negative affect (NA)
NAA *see* Neutron activation analysis (NAA)
NAACP *see* National Association for the Advancement of Colored People (NAACP)

NAAHoLS *see* North American Association for the History of the Language Sciences (NAAHoLS)
Nabokov, Vladimir Dmitrievich 21:741
NABSW *see* National Association of Black Social Workers, Inc. (NABSW)
NACC *see* National Aboriginal Consultative Committee (NACC)
NAcc *see* Nucleus accumbens (NAcc)
NACE (Nomenclature des Activités économiques dans les Communautés Européennes) 5:760
N-acetyltransferase 2 (NAT2) 15:339–340
NACJD *see* National Archive of Criminal Justice Data (NACJD)
Nadar *see* Tournachon; Gaspar Felix
Nadel, Siegfried 16:562
Nadelson, Carol 25:625
Na-Dene relationship between Athabaskan, Tlingit, and Haida 20:899
NADER *see* National Association of Directors of Educational Research (NADER)
Nader, Laura 7:387, 7:622–624, 13:517
Nader, Ralph 19:555–556
Naderism 8:26–27
NAEP *see* National Assessment of Educational Progress (NAEP)
NAFTA *see* North American Free Trade Agreement (NAFTA)
Nag Hammadi Library 3:546, 3:548
Nagaland, peace building and music in 25:845
Nagel, Ineke
  on cultural capital 5:378
  on cultural participation 22:333
Naglieri, Jack 12:308
Nagorno-Karabakh War 25:423
NAGPRA *see* Native American Graves Protection and Repatriation Act (NAGPRA)
Nahda 16:361, 16:381
Nahuatl 17:558, 17:560
NAICS *see* North American Industry Classification System (NAICS)
Nail household phenomenon 20:677–678
NAIP *see* National Agriculture Imagery Program (NAIP)
NAIRU *see* Nonaccelerating inflation rate of unemployment (NAIRU)
Naive Bayes 16:119
Naïve discriminative learning (NDL) 19:729
Naïve psychology, Heider and 10:754
Naïve realism 20:21, 22:223
Naïve regularity theory 3:268–269
Nakama 24:584
Naka–Rushton function 25:159
Naloxone, and placebo effect 18:154
Naltrexone 17:595
NAM *see* Nonaligned movement (NAM)
Nam regia maiestas non moritur 1:7–8
Name calling 18:614, 25:397–398
Name-Letter Task 21:523
NAMES Project *see* AIDS Memorial Quilt (NAMES Project)
Names/naming 12:135–136, 13:469, 13:594, 16:183–187
  and agnosia 1:440–441
  alterity 16:183
  clusters 17:224
  collective naming 16:184–185
  in disputes 6:561–562
  functional diversity 16:184
  hierarchy 16:184
  homonymy and seriality 16:185
  house names 16:184
  and life course 16:185–186
  nicknames 16:184–185
  onomastics 17:223–226
  patronyms 16:184
  personal naming 16:183, 16:185
  pronouns 16:183
  proper names 16:184–186
  reference for 20:68–69
  surnames 16:184–186
  teknonyms 16:184–185
  totemism 24:453–454
  transfer 17:225
NAMI *see* National Alliance for the Mentally Ill (NAMI)
Namibia
  education in 7:200, 7:204
  sociocultural aspects 23:76, 23:78–79
Namier, Lewis, works, *England in the Age of the American Revolution* 3:62
Naming ceremony, life phase 5:508
Naming patterns 4:235
Namson, Ch'oe 2:72
Nanjing 24:872–874
Nanking Massacre 12:796
Nanking Treaty of 1842 3:531–532
Nano art 16:189, 16:191
Nano-louse 16:190
Nanoscience 16:190
  attitudes toward 16:191
  public perceptions of 16:191
  sociocultural history of 16:188–189
Nanotechnologies 1:474, 16:188–189
  attitudes toward 16:191
  framing devices 16:192
  images and visions 16:189–191
  in media 16:191–192
  nano facts and nano fictions 16:189
  public perceptions of 16:191
  social and cultural framing of 16:192
  sociocultural history of 16:188–189
Nantissement 21:403
Nao, Deguchi 20:370
NAPM *see* National Association of Purchasing Managers (NAPM)
Napoleon, Louis 11:686
Napoleonic civil law countries 21:400
Napoleonic Code 13:628–629, 13:796, 19:174
Napoleonic (French) model 7:180
Napoleonic Reforms 3:696
Napoleonic wars, French Revolution 9:409
Napolitano, C.M., on adolescent self-regulation 21:551
NAPP *see* North Atlantic Population Project (NAPP)
NAPs *see* Nonaccidental properties (NAPs)
Al-Naqeeb, Khaldoun H., work, *State and Society in the Gulf and the Arab Peninsula* 16:383
NARA *see* National Archives and Records Administration (NARA)
Narayan, Kirin 1:743
Narbonne 2:703
Narcissism 17:439, 17:442–443, 21:524
  and personality 5:671–672
  and self-development 21:496–497
Narcissistic personality 17:439
Narcissistic theory of suicide 23:677
Narcissists 17:443
Narcoethnography 1:628
Narcolepsy 22:74–75
Narcotics Anonymous 1:503
Narcotizing dysfunction of mass media 13:638, 15:248
Nardi, P.M., on LGBT friendship 9:454
Narration theory 11:114–115
  biography 2:638–639, 2:641
  historiography 11:97–98
  literature 11:98
Narrative analysis 4:725–726, 22:546
  coding methods 4:726
  external narrative sequences 4:725
  internal narrative sequences 4:725
  Narrative Process Coding and System 4:725
  qualitative developmental psychology 19:673
Narrative approaches 3:866–867
  to organizations 16:194–197
    antenarratives 16:195
    anthropology 16:194–195
    causality 16:194, 16:196
    communication 16:194–195
    critical awareness 16:197
    enquiry 16:195–196
    epistemology 16:197
    fiction 16:194, 16:196–197
    fieldwork (methodology) 16:195
    identity 16:195
    leadership 16:196–197
    management 16:195–196
    narratology 16:195–196
    organizing 16:194–196
    plot 16:194
    sensemaking 16:195
    storytelling 16:194–197
    time 16:194
  of social comparison 22:211
Narrative attachment therapy 8:747
Narrative competence, and personal identity 17:745–748
Narrative devices 16:219–220
Narrative ecstasies 16:170
Narrative elements, networks of 5:561
Narrative ethnography 8:181–182
Narrative explanation theories 11:14–15
Narrative exposure therapy (NET) 16:198–203, 24:437–438, 16:200f
  character of 16:199–200

evidence for 16:200–201
for forensic offender rehabilitation (FORNET) 24:439–440
for mass trauma 14:749–750
theoretical underpinnings 16:198–199
therapeutic elements of 16:200
therapeutic procedure of 16:199
Narrative films (story films) 9:163
Narrative identity 17:746
theory of 6:429
Narrative interviews, qualitative health research 10:693
Narrative mapping 24:780–781
Narrative networks 16:211–217
aggregation and comparison 16:214–215
boundaries 16:214–215
casing 16:214
contributions and new directions 16:215–217
methodological challenges 16:216
theoretical/interpretive challenges 16:216–217
and culture 16:212
definition of narrative 16:211–212
and history 16:213–214
and identity 16:211, 16:213–216
institutional and individual narratives 16:215
and life story 16:212
multiple narratives, integration of 16:214
Nazi 16:212–216, 16:213f
network analysis 16:211, 16:213
types of narratives 16:212
Narrative psychology, qualitative developmental psychology 19:673
Narrative texts, and first language acquisition 9:204
Narrative-based social work ethics 22:784
Narratives 2:883, 4:729–732, 11:108–112, 11:121, 13:269–270, 16:218–222, 17:285–286, 19:682–683, 21:102
analysis 16:206–207
performance 16:207
structural 16:206–207
thematic 16:206
and collective memory 4:196
data, collecting 16:205–206
definition of 14:346–347
definitions 16:204–205
devices 16:219–220
epistemic authority 16:221
flashbulb memories as 9:257
Foucault on 16:218–221
future developments 16:209
Goffman on 16:218, 16:220
interaction order and identity work 16:221–222
Labovian model 16:207
longitudinal data 14:346–347
narrative identity 16:204–206, 16:208–209
nationalism as 16:295
participation frameworks 16:218, 16:220
reported speech 16:219, 16:221
representation of history 11:125–128
resistance and social categorizations 16:220–221
social accounts and character contests 16:220
social action and technologies for building stories 16:219
for social and behavioral science research 22:616–617
in social work 16:204–210
story versions 16:220–221
storytelling as storytelling contests 16:221
talk and interaction 16:204, 16:206, 16:209
technologies of selves 16:218–219
theory 16:205
understanding clients 16:208
understanding practices 16:208–209
Narrativity 12:646–647
in critical social work 5:291
definition 5:290
Narratology 16:195–196
Narrow egalitarianism 7:310
Narrow personality traits, gender differences in 9:752
Narrow reflective equilibrium 20:78
Narrowing, dispute 6:562
NAS *see* National Association of Scholars (NAS); Net additions to stock (NAS)
NASA *see* National Aeronautics and Space Administration (NASA)
Nasal epithelium 17:196
NAS-BSD *see* National Academy of Sciences-National Research Council's Board on Sustainable Development (NAS-BSD)
Nash, John 9:582, 9:584
on capitalism in North America 17:24
Nash bargaining problem 9:578, 21:335
example 9:580
theorem 9:578
trade unions 24:500–502
Nash bargaining solution 21:335–337
Nash equilibrium (NE) 5:97, 6:559, 9:577–578, 9:584–585, 12:100, 12:102–103, 16:432–433, 16:573–574, 16:576, 17:255, 18:512–513, 19:901–902
Battle of the Sexes 9:577
closed-loop 17:255
Cournot Oligopoly 9:577
definition 9:577
Markov-perfect 17:255
Prisoner's Dilemma 9:577, 18:931–933
theorem 9:577–578
Nash wage rule 21:337
NASS *see* National Agricultural Statistics Service Information (NASS)
Nasser, Gamal Abdel 25:838
NASSH *see* North American Society of Sport History (NASSH)
NASSPE *see* National Association for Single Sex Public Education (NASSPE)
NASW *see* National Association of Social Work (NASW)
NAT *see* Normal accidents theory (NAT)
NAT2 *see* N-acetyltransferase 2 (NAT2)
Nation building 25:512–513
Bendix on 2:513
concept of 1:305–306
in Europe 20:746
and folklore 9:280, 9:283–284
in Mexico 1:627
National Aboriginal Consultative Committee (NACC) 2:237
National Academy of Sciences (NAS) 5:235–236
National Research Council's Board on Sustainable Development (NAS-BSD)
sustainability transition of 23:803–804
National account systems 16:227–232
National accounting system 7:45–46, 7:48, 13:20
National accounts, macroeconomic data 5:752
National Accounts Main Aggregates Database 5:748
National advertising 1:209
National Advisory Commission on Civil Disorders 20:540
National advocacy groups, formation of 23:219–220
National Aeronautics and Space Administration (NASA) 14:45, 20:421–422, 21:214
National agendas, and archaeology 1:883–884
National Agricultural Statistics Service Information (NASS) 10:304, 20:423
National Agriculture Imagery Program (NAIP) 20:421
National Alliance for the Mentally Ill (NAMI) 15:217
National and regional demographic accounts 9:320
National Annenberg Election Study 2004 1:354–355
National Apprenticeship Scheme (2008) 22:6
National Archive of Criminal Justice Data (NACJD) 5:837
National archives 5:847
National Archives (Paris) 5:701
National Archives and Records Administration (NARA) 1:916, 5:702
National armies 16:305–306
National Assessment of Adult Literacy 7:194
National Assessment of Educational Achievement data 7:820
National Assessment of Educational Progress (NAEP) 7:820, 9:677
Trial State Assessment 11:501–503
multiple comparisons 11:502
multiple determination 11:501–502
National Association for Gifted Children 7:160
National Association for Single Sex Public Education (NASSPE) 3:927
National Association for the Advancement of Colored People (NAACP) 3:692–693, 6:691, 7:759

National Association for the Advancement of Colored People's Legal Defense and Education Fund 19:555
National Association for the Promotion of Social Science (UK) 22:990–991
National Association of Black Social Workers, Inc. (NABSW) 1:305
National Association of Directors of Educational Research (NADER) 7:279–280
National Association of Evangelicals 20:397–398
National Association of Purchasing Managers (NAPM) 9:317
National Association of Scholars (NAS) 2:112
National Association of Social Work (NASW) 6:874–875
National associations, for social sciences 22:990–991
National attachment 16:311
National Black Feminist Organization (NBFO) 8:908
National Board for Professional Teaching Standards (NBPTS) 24:79
National Broadcasting Company (NBC) 14:722, 24:169
National Broadcasting Corporation *see* NBC (National Broadcasting Corporation)
National Bureau of Economic Research (NBER) 24:808
National census
  in comparative contexts 3:294–295
  content 3:291–292
    core topics 3:291–292
    noncore topics 3:292
  data dissemination 3:294
  data processing 3:294
  features of 3:290–291
    geographical specificity 3:290
    household data 3:291
    responsible bodies 3:290
    simultaneity and periodicity 3:291
  methodologies 3:293–294
    accuracy assessment of data 3:293–294
    enumeration 3:293
  recent shifts from traditional censuses to alternative models 3:292–293
National Center for Early Development and Learning Multi-State Pre-K Study 3:835
National Center for Education Statistics (NCES) 1:233, 6:567, 21:43–44
National Center for Education Statistics Surveys Program 5:696–697
National Center for Health Statistics (NCHS) 10:303, 15:403, 16:91
National Center for Transgender Equality 24:553
National Centre for Social Research (NCSR, London) 12:550, 23:749
National Change of Address (NCOA) program 1:840, 14:355
National character 10:485, 16:233–240
  in Austria 16:239
  conduct 16:233
  in England 16:239
  and habitus 16:238–239
  homogeneity 16:233, 16:235
  and nation-states 16:233–234
  reframing 16:235
  self-perception 16:233, 16:235
  sensibility 16:233
  studies 16:234–235
  and war 16:234–235
National character studies 5:481–482, 7:145
National cinemas 9:167–168
National civic culture 16:20
National civil society 4:772
National classifications, of labor statistics 5:798
National Coalition for Gay and Lesbian Equality (NCGLE) 21:746
National Coalition for the Homeless (NCH) 25:113
National College for School Leadership (NCSL) 21:67
National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 4:252, 20:514
National Commission on Certification of Physical Assistants 16:911
National Commission on Excellence in Education, *A Nation at Risk* 7:298
National Commission on Law Observance and Enforcement 1931 5:153
National Committees for Research Ethics, Norway 20:518
National Comorbidity Survey Adolescent Supplement (NCS-A) 18:37
National Comorbidity Survey Replication (NCS-R) 1:100, 6:688, 15:180–181, 17:592
National Conference of Parent–Teacher Associations 25:100
National Council for Accreditation in Teacher Education (NCATE) 24:76–77
National Council for the Social Studies 22:727
National Council of Schools and Programs of Professional Psychology (NCSPP) 24:517
  competency areas 24:520
    consultation 24:520
    diversity 24:521
    education 24:520
    evaluation 24:520
    intervention 24:520
    knowledge base 24:521
    management 24:520–521
    psychological assessment 24:520
    relationship 24:520
    research 24:520
    supervision 24:520–521
  dissertations, types of 24:518
National Council of Teachers of Mathematics (NCTM) 14:755
National Council on Community Behavioral Healthcare 19:590
National Council on Economic Education (NCEE) *see* Council for Economic Education (CEE)
National Council on Measurement in Education 4:776
National councils 22:621
National Crime Victimization Survey (NCVS) 5:155, 5:165, 5:634, 5:714, 5:721–725, 5:848, 19:160, 19:733–734, 21:659–660, 5:722t
National cuisines and identities 9:293–294
National cultural repertoire, in cultural sociology 5:619
National culture
  as context 17:615–616
  organizational culture 17:346–349
  workplace values 17:823–824
National currency 15:743–744
National datasets, core 5:844–845
National Death Index 14:355–356
National Defense Authorization Act 6:554
National Defense Education Act 1958 (NDEA) 1:933, 1:946–947, 1:951, 25:815
  Title VI 1:951
National democracy, contradictions of 22:890–891
National development
  agrarian political economy
    1970s onwards 1:459–460
    1950s–1970s 1:458–459
National Development and Reform Commission (NDRC) 6:833
National Deviancy Conference 5:241–242
National differences 12:369
  in SWB 23:643–644
  white-collar workers 25:566
National disciplinary associations 22:675–676
National dividend 14:623, 16:228
National DNA Database (NDNAD) 9:341, 9:960
National economy, Albert on 1:497–498
National Education Association (NEA) 7:279–280
National Education Longitudinal Study (NELS) 5:376–378, 19:793
National education surveys 5:847
National election studies 2:195, 5:825–826, 5:828
National elites, tribes 24:645
National enthusiasm 4:238–239
National Environmental Justice Advisory Council (NEJAC) 7:757
National Environmental Policy Act (US) 7:781
National environments, teaching and research infrastructure 22:668–670
National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) 1:100, 1:501, 1:505, 6:688, 17:592
National Family Life and HIV Education Curriculum 21:825

National Family Violence Survey (NFVS) 5:724–725
National Federation of Independent Business v. Sebelius (2012) 10:596, 18:736–737
National Fund for the Protection of Family and Population in Hungary 8:220–221
National Gay and Lesbian Task Force 9:630
National Gay Health Coalition 15:111
National Gay Health Education Foundation 15:114–115
National gender policy in modern Japan 6:794–795
National Genographic Project (NGP) 9:901–902
National genomes, notion of 1:763
National 4-H Council 3:447
National health actors in collecting and sharing of health data 5:786–787
National Health and Nutrition Examination Survey (NHANES) 2:659, 15:846–847, 16:91, 17:77
National Health and Social Life Survey (NHSLS) 10:796, 10:801–802, 15:315
National Health Interview Survey 6:379, 10:303
National Health Service (NHS) (UK) 10:37, 10:599, 10:663, 15:852, 16:771–772
  Central Register 5:736
  and clinical psychology 3:874
National health survey, missing data, multiple imputation for 16:91
National Hearing Voices network 15:198
National historiographies 4:408
National Household Education Surveys Program 5:847
National Household Survey of Drug Abuse (NHSDA) 10:590
National Household Travel Survey data set 15:637
National Housing Association (US) 1:219
National hubris 16:312
National Human Genome Research Institute (NHGRI)
  Ethical, Legal, and Social Implications (ELSI) program 8:43–47
National identities 9:434–435, 11:435–437, 11:526–528, 16:283, 16:285–286, 16:311–312
  and Arab nationalism 16:271, 16:273
  ethnic identity 8:115–117
  and ethnicity 8:151
  frontier 9:467
  globalization and world culture 10:230–231
  Jewish tradition 11:68
  and mass media 14:707
  in Middle East and North Africa 15:355
  and migration 15:423–424
  and nationalism in East Asia 16:276–277
  and publishing 19:622–623
  in South Asia 16:279
  and work 8:106–107, 8:109, 8:112
National identity movements 11:544
National ideologies 11:564–565
National imprinting, multinational corporations 16:79
National Incident-Based Reporting System (NIBRS) 5:713, 5:718, 5:720–721, 5:725, 23:142
National income 7:45
  and labor earnings 25:15
  macroeconomic data 5:752
National Independence Movements 20:857–859
National Initiative for the Care of the Elderly (NICE) 7:339
National Institute for Health and Care Excellence (NICE) 6:203, 8:330, 10:605–606, 17:927, 19:590
National Innovation System 20:496
National Institute of Alcohol Abuse and Alcoholism (NIAAA) 1:501
National Institute of Child Health and Human Development (NICHD) 5:696
  Study of Early Child Care 3:397
National Institute of Mental Health (NIMH) 15:98, 17:32, 24:526
  RDoC initiative 21:14
National Institute on Aging (NIA) 5:698
National Institutes of Health (NIH) 4:252, 5:694, 8:43–45, 9:1019, 15:98, 20:515, 22:657–658
  Office of Behavioral and Social Science Research (OBSSR) 8:333
National intellectual cultures 22:970
National interest 16:285–288
National Internet Cafe Chains 5:600
National Intimate Partner and Sexual Violence Survey (NISVS) 5:721–722, 5:724–725, 12:685
National Labor Committee (NLC) 1:623
National Labor Relations (Wagner) Act 5:253, 11:349, 24:492
National Labor Relations Act 1935 (US) 25:699
National laws, for LGBTQ/sexual minority youth in schools 21:88
National Legal Services Authority v. Union of India (2014) 21:771
National Lesbian Health Care Survey (NLHCS) 15:114–115
National level regulation, of privacy on Internet 12:596
National liberation movements (NLMs) 16:241–246, 16:269, 20:621, 22:888, 22:890
  in Africa 16:241–245
  in Asia 16:241, 16:243–245
  centers and peripheries in modern world system 16:241–242
  colonization 16:243
  imperialism 16:243
  and Marxism–Leninism 14:656, 14:660
  nationalism 16:241–244
  new challenges, problems, and perspectives 16:245
  in South Asia 16:280–281
  stages of deployment of 16:242–245
National liberation narratives 14:202
National Long Term Care Survey 15:846–847
National Longitudinal Study of Adolescent Health (Add Health) 2:660, 5:696, 5:712, 8:465, 16:558
National Longitudinal Study of High School Class of 1972 (NLS-72) 5:693
National Longitudinal Survey of Youth 5:376, 5:738, 17:463–464
  NLSY79 5:693, 5:695–696
  NLSY97 5:694–696
National longitudinal surveys (NLS) 5:695–696, 5:744–746
  consumer expenditure survey 5:746
  panel study of income dynamics 5:745
  survey of income and program participation 5:745–746
  time use survey 5:746
National majority, and immigration 11:675–676
National memory, and archaeology 1:881–884
National minorities 16:23
National movements 16:288–289
National Museum of Natural History 1:636
National Museum of Popular Arts and Traditions (MNATP) 1:753, 1:755
National Museum of the American Indian Act (NMAIA) 1:636, 1:885–888
National musics 16:150
National Neighborhood Crime Study (NNCS) 5:184
National Oceanographic and Atmospheric Administration (NOAA) 20:421–422
National Office of Social Hygiene 8:241
National Opinion Research Center (NORC) 5:845, 12:550, 23:748
National Organisation for the Protection of Motherhood and Infancy 8:220
National Organization of Bar Counsel (NOBC) 13:621
National panel studies, examples of 5:845–846
National Penitentiary of Buenos Aires Institute of Criminology 5:239
National People's Congress (China) 5:600
National perspective, indigenous peoples in 11:756–757
National policy
  effect of community and state actions on 10:650
  styles 7:781–782
National postal systems 18:641–642
  origins 18:641
  postal reform 18:641–642
  recent developments 18:642
National pride 16:312
National Projections (2012) 9:319
National projects 1:758–759, 1:762
National Public Radio (NPR) 3:358, 19:880
National qualification systems 10:250

National Registry of Exonerations (NRE) 25:787–789
National regulatory authorities (NRA) 5:705
National Research Act 1974 4:252, 8:48
National Research Council (NRC) 22:597, 24:763–764
National Research Ethics Commission 8:33
National Resident Matching Program (NRMP) 7:188
National Resistance Army (NRA) 20:342
National Resources Planning Board (US) 1:219–220
National rollout 7:231
National roots of science of human variation and differences 1:758
  biobanks 1:759–760
  future 1:762–763
  genealogies and ancestry 1:761–762
  genetics 1:759–760, 1:762
  genomics 1:758, 1:760–762
  human variation, history of 1:758–759
  microbiomes 1:762–763
  national and the transnational 1:760–761
  personal genomics 1:761–762
National schools 21:128
National Science Advisory Board for Biosecurity (NSABB) 3:280–281
National Science Board
  Science and Engineering Indicators 21:322
National Science Foundation (NSF) 7:281–282, 7:286, 7:792–793, 8:507–508, 13:516, 20:524–525, 22:655–657, 22:664
  EarthCube 23:134
  Scientists and Engineers Statistical Data System 7:282
National security 3:678, 16:254–255, 24:765
  globalization and mobility 16:254–255
  implications 16:255
  and Internet 12:597
  meaning of 16:254
  size of the war machine and 16:255
  in United States 16:247–252
National security studies 16:253–256
National Service Framework for Mental Health (NSF-MH) 15:198, 19:590
National social integration 12:246–247
National Social Life, Health, and Aging Project (NSHAP) 10:796, 10:807–808, 14:513
National social welfare legislation 13:722
National socialism 8:256, 16:257–261, 18:434–435, 18:437–439, 21:369, 25:566–567
  antisemitism 1:784–785
  Dissolution of the Third Reich 16:260
  and fascism 14:498
  Hilberg on 10:870–871
  military aggression and racial annihilation 16:259
  Nazi State 16:258–259
  origins 16:257–258
  project of 15:710
  racial annihilation 16:260
  seizure of power 16:258
  social work 1:785
  World War II 16:259–260
National Socialist (Nazi) regime 3:24–25
National Society for the Study of Education (NSSE) 7:279–280
National sociological tradition 22:965
National sovereignty 16:262, 16:298
National Spelling Bee 10:399
National Standards for United States History 11:44
National Standards for World History 11:44
National Statement on Ethical Conduct in Research Involving Humans 20:515
National statistical institutes (NSI) 5:705, 5:845
National Statistical System 3:296
National statistics 10:301–302
National Statistics Office (NSO) 3:290
National styles, historical geography 11:20
National survey centers 23:748
National Survey of Adoptive Parents 1:138
National Survey of Youth and Religion 7:418
National Survey on Drug Use and Health (NSDUH) 6:251, 6:688
National systems of innovation 12:145, 12:149
National Theatre Movement 24:653–654
National traditions, of censuses 3:304–306
National Transgender Advocacy Coalition 24:553
National Transgender Discrimination Survey 24:550
National TV Violence Study (NTVS) review 25:102
National unity in Italy 16:297
National universities, de-Westernization 25:541, 25:543–545, 25:548
National University of Singapore
  Biopoleis Project 21:182
National urban systems 24:945–946
National Vacant Properties Campaign (NVPC) 21:917
National Violence Against Women and Men Surveys (NVAWMS) 5:721–725, 5:723t
National Vocational Qualifications 25:236, 25:247
National Women's Studies Association (NWSA) 9:739
National Women's Trade Union League 24:492
National Youth Anti-Drug Media Campaign 6:681
National Youth Survey (NYS) 5:712, 5:721, 5:723t
Nationaldemokratische Partei Deutschlands (NPD) 18:410–411
Nationalism 2:891–892, 4:339, 6:277–278, 6:690, 6:812, 10:485, 14:20, 15:327–328, 16:223–225, 16:262–265, 16:284–286, 16:297–299, 16:309, 18:376–377, 18:384, 19:854, 20:338–343, 20:363, 20:397, 23:366, 23:632, 23:853, 25:96–99
  in Africa 16:266–270
    colonial protonationalism 16:268
    evolution between World Wars until decolonialization 16:268–269
    genesis of modern 16:267–268
    National Liberation Movements 16:269
    older forms 16:266–267
    precolonial nation-states 16:267
    South Africa 16:268
    sub-Saharan Africa 16:267–268
  and anthropology 1:751–754
  in anthropology 1:769–770
  anticolonial 16:264
  Arab, historical aspects of 16:271–274
  and balance of power 2:346
  citizenship and 16:263–264
  civic 22:890
  and colonialism 4:220
  and communalism 4:263
  community and 16:262–265
  versus cosmopolitanism 5:43–44
  as cultural idiom 16:294
  and culture 16:262–263
  and dance 5:668
  defined by religion 20:402–403
  East Asian studies 6:785–786, 16:275–278
  Eastern European studies 6:806–808
  ethnic 22:890
  and ethnic cleansing 8:83–86
  and ethnic conflict 8:92, 8:94
  ethnicity 16:262, 16:265
  and ethnicity 8:104, 8:148, 8:150–151
  gender and 16:265
  and global history 10:163, 10:166
  global politics as 22:888
  historical aspects of Africa 16:266–270
  historical aspects of the West 16:283
  and home 11:157
  identity 11:522–523
  as ideology 16:293–294
  in India 20:375–376
  Islamic fundamentalism 12:741–744
  and journalism 12:851
  and language 16:262–264
  Mann on 14:492, 14:494
  Maruyama Masao 14:635, 14:637, 14:640
  Meiji Revolution 15:129–130
  methodological 5:44
  and Middle East politics 16:371–372
  and Middle East society 16:382, 16:384
  and migration 11:650
  and modernity 9:672–673
  modernity of 16:262
  morality of 16:262–263
  and multiculturalism 16:22, 16:25
  nation and culture 16:265
  nation and state 16:263–264
  and national liberation movements 16:241–244
  and nostalgia 17:36–37
  and popular culture 18:530–531
  postsocialist societies 18:698–699
  and radio 19:878–879

and religion 24:603
and resistance to conquest, in Africa 1:286
in settled times 16:310–312
interpretive studies of everyday nationalism 16:310–311
survey-based studies of everyday nationalism 16:311–312
small states 22:111–112
sociology of 16:263, 16:290–296
citizenship 16:291–292
class and nation 16:292–293
class and peripheral nationalism 16:292
classical sociology 16:290–291
comparative–historical sociology 16:293–294
democracy 16:291, 16:294
Deutsch's contribution 16:291–292
Durkheim's contribution 16:291
ethnic groups and nations 16:293
Gellner's contribution 16:292
industrialization 16:290–293
Mauss's contribution 16:291
modernization theory 16:291–292
nation as imagined community 16:294
nation as invented tradition 16:294
nationalism as cultural idiom 16:294
nationalism as discourse, narrative, and event: religion and nationalism 16:295
nationalism as ideology 16:293–294
nationhood as institutionalized form, practical category, and cognitive scheme 16:294–295
neo-Marxism 16:292–293
rational choice theory 16:295
social constructionist approaches 16:294–295
structural functionalism 16:291
uneven development 16:292–293
Weber's contribution 16:290–291
world system theory 16:292–293
in South Asia 16:279–282
Southeast Asian studies 11:78–80
sovereignty 16:262–265
in Soviet Union 20:821–822
and sport 23:314
and statehood 23:350–351
sub-Saharan Africa 11:90
traditional nationalism scholarship 16:309–310
tribes 24:646–647
and war 16:305, 16:308, 25:363–370
and warfare 25:417–419
in West 16:283
conflicts of interests 16:288
ethnic roots 16:287
evaluation 16:285–286
historical explanation 16:286–289
preconditions 16:286–287
as manifestations or consequences of modernization 16:287–288
problem of definition 16:283–286
in nineteenth century 16:283–284
in twentieth century 16:284–285
psychological, cultural, and emotional manipulation 16:288–289
world systems theory 25:749, 25:751
and xenophobia 25:804
Nationalist anthropology 1:753
Nationalist conflict *see* Ethnic conflicts
Nationalist historiography 16:297
Nationalist history 11:84–85
in Indonesia 11:85
in Southeast Asia 11:85
in Thailand 11:85
in Viet Nam 11:85
Nationalist homogenization 25:367–368
fascism, war and nationalism 25:367–368
nationalism, war and genocide 25:367
postwar era and globalization 25:368
Nationalist movements, social base of 16:292–293
Nationalist party 3:515
Nationality 6:8, 6:10
and citizenship 3:651–652
Nationality Act 1906 2:584
Nationalizations 4:305, 7:173, 12:582
Nationalizing nationalism 16:309
National-level survey of legal mobilization 13:595–596
Nationalökonomie 4:340, 22:697–698
Nationalsozialistische Deutsche Arbeiterpartei (NSDAP) 16:257, 18:412
Nationhood 16:294–295
Nations 6:276, 6:278, 10:313, 16:297–303, 18:380–382, 22:887–889
balance of power 2:346–347
birth of 22:121
in China 2:66
and class 16:292–293
and community 4:388
and culture 16:265, 25:29–30
and culture, Eastern European Studies 6:806–807
developmental research across 6:326–331
and de-Westernization 25:547–549
division of labor 25:341–343
ethnic groups and 16:293
global history 10:163–166
and health 22:279
as imagined community 16:294
as invented tradition 16:294
Jewish 11:69, 11:71
and labor 25:677
in Middle East and North Africa 15:355–356
and military 15:509
modernism 16:224–225
mondialization 16:225
nationalism 16:223–225
primordialism 16:224–225
state building 16:223
territory 16:223–225
before nationalism 16:293
sociology of 16:223–226
concepts 16:224–225
Europe 16:223–225
future 16:225
globalization 16:225
immigration 16:225
as late political meaning 16:223
legitimacy 16:223–225
and state 16:263–264
and territory 24:224
traditions 24:509
war potential of 16:253–256
Nation-state formation
beyond Europe 16:300–301
in Europe 16:299–300
political transformation 16:300
reform 16:299
separation 16:300
unification 16:299–300
Nation-state policy making 18:314–315
historical development 18:315
Nation-state sovereignty 3:638–639
Nation-states 4:604, 7:308, 10:218, 10:220, 15:519, 16:262–265, 16:297–303, 18:375, 19:607, 20:163, 22:887–888, 22:890–891, 23:347–348, 24:588–591
and capitalism 10:218–220
in Central Asia 3:316–320
centralized, Japan 15:128–129
cultural critique and 5:385–386
and education 7:173, 7:175
engineering 7:646
and ethics 1:735–737
and ethnic cleansing 8:84–86
global history 10:162–164
history of demography 6:170–171
and internal citizenship status 3:657–658
international system of 11:648–650
Mann on 14:490–495
and market, relation between 22:121–122
memory 1:883–884
in Mexico 1:630
and nationalism in East Asia 16:276–277
and political citizenship 3:651, 3:653
and punishment 19:655–657
and sex work 19:282
shared representations of *see* Shared representations of nation–state
social integration beyond 12:246–247
and social justice 22:398–399
and social sciences 6:487–488
sociocultural order in Europe 8:276–277
and territorial conflict 24:214, 24:216–219
transformation from colonies, in South America 23:46
and transnational corporations 10:221–223
and war 16:304–308
war before 16:304–305
wars among 25:420
Native America, tricksters in 24:649–652
Native American Graves Protection and Repatriation Act (NAGPRA) 1:636, 1:886–889, 9:966
Native Americans 1:633–637, 4:726
and adoption 1:137
anthropology and studies 1:636–637
archaeology politics 1:885–888, 1:890
contemporary status 1:635–636
distribution 1:636

Native Americans (*continued*)
tribal membership 1:635–636
distribution 1:634
dumping grounds 7:760
early 1:633–634
impact of European contact 1:634–635
tribal decimation and relocation/concentration 1:634–635
tribal or individual Indian 1:635
urbanization 1:635
origins 1:633
population size 1:634
for waste trading 7:760
Native mining 3:6
Native Title Act 1993 2:237
Nativelike proficiency 16:66
Native's wages, impacts of immigration on 11:638–639
Nativism 25:796–801
antifeminism 25:800
anti-immigration 25:797
antimulticulturalism 25:800
current usages 25:797–798
and ethnicity 8:159
exclusionary ideologies 25:796–797, 25:800
linguistic 13:321–322
and modern human cognition 11:252
neoracism 25:797, 25:800
new racism discourse 25:799–800
social phobias 25:797–798
Nativist–empiricist controversy 10:757
Nativity 11:654
and fertility 15:400–401, 15:400t–401t
Nativized pidgins 18:136
NATO *see* North Atlantic Treaty Organization (NATO)
Natorp, Paul 19:391
Natural areas 22:264, 24:831–832
Natural capital 6:852–853, 6:856–857, 23:823
sustainable development 23:823–827
Natural computation 2:27–31
dynamics of biological neurons 2:28–29
genetic algorithms 2:30
in human brain 2:28–29
mathematical notion of computation 2:27
neural networks 2:28–30
in neural systems 2:28
turing machine 2:27–28
Natural defense, cardiac defense 23:540
Natural disasters 3:425, 6:478, 22:753
astronomical hazards 6:481
biological or ecological hazards 6:480–481
and childhood stress 23:570
climate change 6:479
geophysical hazards 6:480
and health 7:744
health-related aspects 16:315–319
casualties 16:316
communicable disease 16:316
consequences 16:315–317
emergency management cycle and health actions 16:317–318
fatalities 16:315–316
infectious disease 16:316
injury 16:315–318
long-term physical disability 16:316–317
mental health 16:316
mitigation 16:317–318
preparedness 16:316, 16:318
recovery 16:317–318
rehabilitation 16:316–318
response 16:316–318
trauma 16:315–318
triage 16:316–318
vectorborne disease 16:316t
waterborne disease 16:316t
hydrometeorological/climatological hazards 6:479–480
and mass trauma 14:747–748
resilience 6:479
Natural disease entities 13:128–130
Natural disease entities, Kraepelin 13:128–130
Natural economy 25:662
work anthropology 25:662, 25:664, 25:666
Natural ecosystems 3:851
Natural environment measures 24:825–826
Natural equality 7:908–909
Natural evolution
philosophical aspects 8:388–391
adaptive social networks 8:390
cultural evolution 8:389–390
fitness 8:388–389
imitation 8:389
learning 8:389–390
local interaction 8:390
population structure 8:389
replicator 8:388–389
Natural experiments 9:129, 17:107
for social change 22:151
Natural fertility 10:776–777, 11:6–7
Natural hazards 4:357
community resilience and 4:357
research 11:407
Natural history 10:774
and anthropology 1:746
anthropology of 1:746
and culture, African Studies 1:277–278
Natural homo oeconomicus 16:293–294
Natural indirect effect 15:67
Natural inequalities, and justice 12:946
Natural kinds
Kraepelin on 13:129, 13:131
realism 20:20
Natural language metaphysics 15:754
Natural language processing 4:952, 5:812–813, 9:364–365
Natural language semantics 13:945
Natural languages 2:560–561, 2:564, 14:841, 16:925
dependencies in 6:182–195
dynamic logics of 16:927
logical semantics of 16:925
Natural law tradition 12:933–934, 22:990
Natural laws 11:367–368, 12:509, 13:428, 13:437–438, 13:484–487, 13:762–763, 13:796, 16:320–324, 18:27–28, 19:167, 21:156–157
anomie and 1:718–719
authority and 16:323
Bentham on 2:526–527
cannibalism 3:84–85
core of 16:320
human rights and 16:322–323
and indigenous rights 11:795
and individual-collective relationship 11:818
interest in 12:387
versus legal positivism 18:25
methods 18:28–29
conceptual and ontological analysis 18:28–29
linguistic analysis 18:28
naturalism in legal philosophy 18:29
reductionism, eliminativism, and emergentism 18:29
moral realism and connection thesis 18:27
origin 16:320
positive law and 16:323–324
reasoning 16:324
as a theory of right action 16:321–322
as a theory of value 16:320–321
validity in natural law theories 18:27–28
morally defined function of law 18:27–28
social facts and the specification problem 18:28
Natural liberty 9:389
Natural models, fishing 9:226–227
Natural monopolies 20:173
Natural orders 4:4
Natural parameter 14:854
Natural partial order 22:479, 22:482
Natural philosophers 21:173
Natural philosophy 20:284, 21:172–173
Natural psychological functions 25:331t
Natural reasoning 16:925
inferences 16:926–927
Natural religion 20:769
Natural resources 20:801–802, 20:806, 25:439–441
GIS/remote sensing for assessing 10:64
LTG study 14:138, 14:141–142
population dynamics and 18:566–568
transportation of 23:847
use of 22:257
Natural resource-seeking investment 9:325
Natural rights 4:521, 19:173
see also Human rights; history
Natural sciences 6:492, 11:102–103, 24:673, 25:343–344
anthropology of 1:735
and bioethics 2:614
epistemology 13:120
expenditures in 9:550t
geographic thought 10:71–72
Nobel Prizes 23:227
social science modeled on 22:990
and social sciences 7:868–869
consensus 7:868
publications 7:868
textbooks 7:868

and social sciences, relationship between 1:496–497
specialization in 23:225
Watson's contributions 25:448
Natural selection 3:86, 5:676–677, 5:679–680, 8:388, 8:400–401, 8:403, 8:469, 8:471–473, 11:289–290, 13:25, 20:284, 20:287, 22:900–901, 22:903, 22:908–910
of comparative method 4:416
for cultural evolution 5:404–405
and Darwinism 1:766, 5:682
evolution 8:385
evolution of social behavior by 22:892–893
evolutionary demography 8:413–415
on fertility and parental investment 9:28–29
Fisher's 9:223
and fitness 1:86
and functional explanation 9:520–522
genetic programming 2:41
and language development 13:319–320, 13:322–323
microevolution 15:339
and organizational conflict 4:641–642
Natural semantic metalanguage (NSM) approach 13:288
Natural surveillance 20:775–776
Natural theology 20:286
Natural unemployment rate 24:738–739
Natural Water Flow Approach 25:446
Natural world, and environmental cognition 7:708
Naturalism 16:325–332, 19:392, 22:491, 22:690
branches 16:325–327
and fetishism 9:106
Feyerabend on 9:122
field experiments 9:128
humanist critique of 20:204
Kraepelin on 13:128
in legal philosophy 18:29
meaning 14:841, 14:843
ontological *versus* methodological 16:326–328
and practical reasoning 18:777–778
psychoanalytical 4:268
and realism 10:788
Rorty's view 20:764–765
scientific 16:326
in social sphere 16:331–332
as a stance 16:325–327
Naturalist social science models, Taylor's 24:40–41
Naturalistic action impairment 1:877–878
Naturalistic approaches 8:577–579
Naturalistic decision making (NDM) 11:300, 14:481
Naturalistic decision research 6:713
Naturalistic metaphysics, Wittgenstein on 25:612
Naturalistic research 23:856–857
Naturalistic settings, in cognitive mapping 4:82
Naturalization process 20:295–296
Naturalization program 16:330–331
Naturalized epistemology 20:22
Naturalizing the mind 16:330–331
Naturally occurring data, qualitative research methods 22:545
Naturalness 13:240–241, 15:832
and land use indicators 13:240–241
Naturalness principle 15:827
Natural–technological synergy 7:807
Nature 13:229
and culture, in end-of-life choices 7:594–596
versus culture 1:765–766
human development 11:272
and learning 13:699–700
Marxist geography 14:677–678
philosophy of 10:738
and political ecology 18:350–354
and reproduction 20:450–454
social construction of 23:86
and war 25:373
Nature 21:323, 22:901
The Nature Conservancy 7:105
Nature of exchange 7:13–14
Nature of food choice 9:298–299
Nature of knowledge societies 13:109–110
Nature of prejudice 23:462–463
Nature of science 21:230–231
models and modeling 21:231
The nature of things 10:226
Nature-nurture 2:450, 9:825–826, 9:985–986, 16:331, 16:340–344
autism and Turner's syndrome 16:343
behaviorism 16:340–341
cultural claims of nurture 16:341
epigenetics 16:343–344
feral children 16:342
Galton on 16:340
and intelligence 12:311
Luria on 14:397
Lysenkoism in USSR and parallels in West 16:341–342
Monoamine oxidase A and antisocial behavior 16:343
origins 16:340
politics against nature 16:341–342
sex determination 16:342
twin studies 16:340, 16:342–344
X-linkage 16:343–344
Nature-related counterdiscourses
on sustainable development 23:813–814
limits to growth 23:814
radical ecology 23:813–814
Nature-society
contemporary, subfields in geography 16:337
in geography 16:334
science 16:335–336
entering twenty-first century 16:338–339
geography as 16:337–338
unified understanding 16:338
Natürpsychologie 17:910–911
Naturwissenschaften 18:621–622
Natyashastra 21:618
Nauru 17:158, 17:161–163
NAV *see* Net asset value (NAV)
Navigate terrains of war 25:380
Navigation
children's ability 11:389–390
in deep sea 16:346
electronic 16:347
global voyaging 16:346–347
improvements in 16:346–347
in Pacific 16:345
shores of Europe, Asia, and Africa 16:345–346
in spatial cognition 23:113–114
in spatial environments 11:386
spatial navigation 23:161–171
variability in large-scale neural network dynamics influences on 11:389
in virtual space, psychological and neural aspects 16:348
world system 16:346–347
Navigation maps 3:171
Navigational tools 16:347
Navigenics 9:980
NAVSTAR constellation 10:175
Nazi, and narrative networks 16:212–216, 16:213f
Nazi Enigma code 19:65
Nazi experiments 21:212
Nazi genocide 10:1–2
Nazi Germany 13:793
Arendt on 1:963
and obedience 17:63–65
relationship between arts and politics in 5:461
Universal Declaration of Human Rights 11:354
Nazi law 16:356–360
basic Nazi principles 16:356
civil law 16:357
continuity and discontinuity of 16:358–359
expression of völkische (racial–ethnic) 16:356
Hitler and 16:358
Holocaust and 16:358
justifications 16:359
leader principle (Führerprinzip) 16:356
legal culture 16:357–358
legal domains 16:359
legal staff 16:357
legal–philosophical point of view 16:359
legislation 16:356–357
modernity and regression 16:358–359
penal law 16:357
people's community (Volksgemeinschaft) 16:356
public law 16:356–357
terrorism and normality 16:358
Nazi movement 19:695–696
Nazi Party 16:258
Nazi Party Rally Grounds 5:416–417
Nazi psychology 18:437–438
Nazi regime 8:224, 8:226
Nazi State 16:258–259
Nazis atrocities 3:675

Nazism 4:753, 4:764, 6:812, 6:814, 7:514, 9:772, 16:284, 18:410–414, 21:370, 21:372, 22:993, 24:448–450
  Fromm on 9:458–460
Nazi–Soviet Pact of 1939 1:557
Nazpary, J. 3:317–318
NBC *see* National Broadcasting Company (NBC)
NBC (National Broadcasting Corporation) 14:693
NBER *see* National Bureau of Economic Research (NBER)
NBFO *see* National Black Feminist Organization (NBFO)
NBPTS *see* National Board for Professional Teaching Standards (NBPTS)
NCA *see* Assertion; normativity of content (NCA); Network of Concerned Anthropologists (NCA)
NCATE *see* National Council for Accreditation in Teacher Education (NCATE)
NCC *see* Neural correlate of consciousness (NCC)
NCDs *see* Noncommunicable diseases (NCDs)
NCES *see* National Center for Education Statistics (NCES)
NCGLE *see* National Coalition for Gay and Lesbian Equality (NCGLE)
NCH *see* National Coalition for the Homeless (NCH)
NCHS *see* National Center for Health Statistics (NCHS)
NCLB *see* No Child Left Behind (NCLB) Act
NCOA *see* National Change of Address (NCOA) program
NCS-A *see* National Comorbidity Survey Adolescent Supplement (NCS-A)
NCSL *see* National College for School Leadership (NCSL)
NCSPP *see* National Council of Schools and Programs of Professional Psychology (NCSPP)
NCS-R *see* National Comorbidity Survey Replication (NCS-R)
NCSR, London *see* National Centre for Social Research (NCSR, London)
NCTM *see* National Council of Teachers of Mathematics (NCTM)
NCVS *see* National Crime Victimization Survey (NCVS)
NDB *see* New Democracies Barometer (NDB)
NDEA *see* National Defense Education Act 1958 (NDEA)
Ndembu symbolism 13:72
NDL *see* Naïve discriminative learning (NDL)
NDM *see* Naturalistic decision making (NDM)
nDNA *see* Nuclear DNA (nDNA)
NDNAD *see* National DNA Database (NDNAD)
NDP, Canada *see* New Democratic Party (NDP, Canada)
Ndrangheta 14:429
NDRC *see* National Development and Reform Commission (NDRC)
Ndyuka Maroons 14:593–594
NE *see* Nash equilibrium (NE)
NEA *see* National Education Association (NEA)
Neanderthals 16:739
  and paleoethnobotany 17:458
Neapolitan emblem 10:110–111
Near East *see* Middle East
Near Eastern studies, on religion 16:376–380
Near Middle East studies
  on gender 16:367–370
    civilizing mission of women 16:367–368
    Islamist movements and women 16:368–369
    multiple modernities 16:369–370
  on politics 16:371–375
    Arab Spring 16:374–375
    colonialism 16:372
    international relations 16:373–374
    Islam 16:372–373
    nationalism 16:371–372
    orientalism 16:371–372
    post-World War II period 16:371–372
    social science 16:371–374
    and society 16:372–373
  on society and history 16:381–386
    contemporary approaches 16:382–384
    culture 16:381–385
    gender 16:385
    Islam 16:381–385
    modernization 16:381–385
    nationalism 16:382, 16:384
    orientalism 16:381–385
    social movements and social change 16:384–386
Near Oceania 17:446
  domestication of tree 17:446
  tropical horticulture 17:446
Near-balanced states, structural balance theory 23:606–607
Near-decomposable systems, modularity 15:718, 15:720
Nearest neighbor 23:393
Nearest neighbor/greedy matching 19:143
Near-infrared spectroscopy (NIRS) 16:387–396, 16:718
  future directions 16:394
  literature 16:390–392
    in clinical populations 16:392
    cognitive development 16:391–392
    functional specialization for language 16:390–391
    functional specialization of brain 16:390
    lateralization for language 16:390–391
  principles of 16:387–390
    continuous light optical imaging 16:388–389
    fast optical imaging 16:389
    frequency-resolved optical imaging 16:389
    versus other imaging techniques 16:389–390
    time-resolved optical imaging 16:389
  technical and methodological considerations 16:392–394
    brain development 16:393
    data analysis 16:393–394
    DeoxyHb 16:392–393
    experimental designs 16:393
    interpretation 16:393–394
    OxyHb 16:392–393
    probe design 16:392
    source 16:392
    TotalHb 16:392–393
  visual word recognition 25:216
Nebiim 19:197
NECAG *see* Northeast Community Action Group (NECAG)
Necessary inferences 12:39
Necessity 5:223–224, 23:362–363
Nechaev, Sergei 24:238–239
Neck, activity above 2:601–602
Neck control, infant motor development 15:973
Necklaces, in Kula ring 13:140–142
NECs *see* Neuroepithelial cells (NECs)
Nederlandse Omroep Stichting (NOS) 2:861
Nederveen, P.J., works, *Ethnicities and Global Multiculture: Pants for an Octopus* 8:140
Need for cognition (NFC) 7:942
Need for distinction 8:836
Need for orientation (NFO) 1:354
  agenda setting 1:354
Need for union 8:836
Need supportive 21:489–490
Needham, Joseph 16:397–401
  assessment and impact of 16:400
  contribution to knowledge and intellectual environment 16:398–399
  on distinction between modern science and medieval science 16:398
  ecumenical and universalistic view of history of science and technology 16:398
  fundamental criticism against 16:400
  honors 16:397–398
  life of 16:397–398
  works
    Chemical Embryology 16:397
    History is on Our Side 16:397
    The Nazi Attack on International Science 16:397
    The Sceptical Biologist 16:397
    Science, Religion, and Reality 16:397
    Science and Civilisation in China (SCC) 16:397
  worldview 16:398–399
Needham, Rodney 13:53
  works, *Belief, Language and Experience* 20:308
Needing 25:351–356
  binge eating and food addiction 25:354

consumer behaviors 25:355
dopamine 25:351–354
drug addiction, recovery, and relapse 25:351–354
incentive salience and utility 25:352–354
gambling 25:354–355
liking 25:353–355
motivation 25:351–355
pleasure 25:352–354
reward 25:351–355
sensitization 25:351–355
Needs
assessment methods 16:403–405
Lacan on 13:214–215
levels of 16:402
origins 16:402
of prisoners 20:61–62, 20:65
and risk 20:692–693, 20:695
types 16:402
and values 25:49
Needs assessment 12:197, 16:402–408
concepts 16:402
definition 16:402
future 16:406–407
guiding models 16:402–403
implementation concerns 16:405–406
and Kaufman's approach 16:403
methods 16:403–405
origins 16:402–403, 16:403t
phases of 16:404t
reasons for 16:405, 16:405t
Needs-oriented moral reasoning 7:315–319, 7:316t
Needy self 17:440–441
Neenghatú 23:46–47
Nefazodone 1:779
NEG *see* New economic geography (NEG)
Negation 16:409–414
contrariness and contradiction 16:410–411
form 16:409, 16:411–414
discontinuous 16:412
of existential, locative, or adjectival predicates 16:412
quadruple 16:413
zero 16:413
meaning 16:409
in relation to truth of sentence 16:410
scope and focus 16:409–410
Negative affectivity (NA) 7:497–498, 7:500, 17:788–789, 17:877
and symptom perception 23:867–869
Negative and negative-like contexts, uniting 16:413
Negative attitudes 15:231–232
age 15:231
educational status 15:232
gender 15:231
interventions 15:232
professions 15:232
religious beliefs 15:231
types of illness 15:231
Negative binomial distribution 6:582–583, 23:468
Negative buying *see* Boycotting
Negative carrying capacity 3:156
Negative civil liberties 3:674–676
Negative commemoration 4:235
Negative conception of freedom 9:388
Negative duties 20:680
Negative education 20:768
Negative effect, and general strain theory 23:496–497
Negative emotionality, and temperament 24:196
Negative emotions 6:747–748, 7:496–500
and neuroticism 16:743–747
regulation of 21:554
Negative empathy 7:551
Negative eugenics 8:233, 8:248
Negative feedback 21:487
Negative freedom 12:243
Negative health, of globalization 10:199–201
loss of government policy space and capacity 10:200
spread of disease 10:199–200
turmoil in labor markets 10:200–201
Negative heritage 4:203, 5:416–417
Negative hierarchical sorting 12:513
Negative income tax (NIT) 8:531, 22:297–298
Negative judgment 3:715
Negative knowledge 7:874–875
Negative liberty (NL) 9:398–399, 9:402
Negative matching 3:807
Negative medical consultation 6:614
Negative moment bounds, unit roots time series 24:365
Negative normalized log-likelihood 23:414
Negative peace 17:620–621
Negative politeness 18:326–328
Negative reciprocity 7:15
Negative reinforcement 1:374, 20:207, 20:213–214
behavior therapy 2:429
Negative reinforcers, behavior therapy 2:429
Negative sanctions 13:807–808
Negative self-perception 22:496
Negative social integration 12:243–244
Negative social interdependence 4:857–859
Negative social network ties 5:561
Negative symptoms
definition 19:290
etiology of 19:292
flattening of affect 19:291
history of term 19:290
poverty of speech 19:290–291
primary and secondary symptoms 19:292
psychiatric assessment of 19:290–294
in psychiatric/mental disorders 19:291
in schizophrenia 19:291
biological correlates in 19:292–293
course of symptoms in 19:292
relationship between positive and negative symptoms in 19:292
treatment in 19:293–294
social withdrawal 19:290
standardized assessment and dimensional differentiation 19:290–291
symptomatic overlap with depression and Parkinsonism 19:291
Negative taper rate 25:494–495
Negative valence 2:562
Negative violence 17:621
Negative-feedback regulation 7:606
Negative-state relief model of empathy 19:217–218
Neglect 3:410, 3:478–479, 17:705
of children *see* Child neglect
defining 3:410–411
depressed 3:413
disorganized 3:413
effects on neural function 16:489, 16:491
of elders 7:330–331
emotional 3:413
hemispatial neglect 10:766–772
reemergence of 3:410
Neglect alexia 6:734
Neglect of neglect 3:410
Neglected rural geographies 20:796
Neglective fictions 23:639
Negotiability 18:799–800
of predictions 23:594
Negotiable instruments 3:30, 19:180
Negotiated exchange 8:483, 8:486
Negotiated justice 18:226
Negotiated orders 25:659
and organizational conflict 4:641
in organizations 17:411–414
ambiguity, interpretation and enactment 17:412
bargaining 17:411–412
beyond classic formal/informal organization distinction 17:412–413
concept origins in hospital study 17:411
continuing and future power 17:413–414
ethnomethodology 17:412
institutions 17:413–414
multilevel organizational level analysis 17:414
official organization 17:411–413
organizational systems 17:413
processes and rules 17:411–412
sub-processes of negotiation 17:411–412
unofficial organization 17:412–413
Negotiated settlement 6:558
Negotiation analysis 16:430–431
Negotiations 6:557, 7:447–448, 16:430–431, 17:621
analytic response 16:431
barriers to 16:419–421
conflict mediation 4:627–628, 4:630
in conflict resolution 4:586
consumer policy 4:734
of culture
and nostalgia 17:36
disputes 6:561–562, 6:564–565
elements of analytic approach 16:431–435
alternatives to negotiated agreement 16:431–432
approach as whole 16:435
behavioral insight 16:434

Negotiations (*continued*)
changes in perceived zone of agreement 16:432–433
changing game 16:434–435
fundamental processes 16:433
interests 16:431
parties 16:431
structure representation 16:432
tension between creating and claiming value 16:434
further directions 16:435
lawyer–client relationship 16:420
in legal systems 16:418–419
models 16:418
organizational aspects
see also Bargaining
and organizational conflict 4:640–641
and power 15:667–668
psychological barriers to 16:419–420
psychology of 16:423–429
cognition of negotiator 16:426
emotion of negotiator 16:426–427
social context 16:423–424
culture 16:424
gender 16:424
power 16:424
social motives 16:423–424
social perception 16:425–426
microethical judgments 16:426
social process 16:424–425
phases and interruptions 16:425
strategies and tactics 16:425
trust and trustworthiness 16:425–426
role of lawyers in 16:418–422
self-categorization theory 21:458
statistical aspects 16:430–436
strategic barriers to 16:419
Strauss on 23:530
structural barriers to 16:420–421
see also Conflict
Negotiator's dilemma 16:419
Negros, Cardoso on 3:123
Negus, Keith, on musical categories 3:801
Nehru, Jawaharlal 20:375
Neier, Ayreh 3:677
Neighborhood 4:525–528, 9:612, 9:614–616, 10:46–50, 11:233–235, 11:237–238, 12:140–142, 16:437–442, 20:793
and adolescent sexual risk 1:132–133
architects' view 16:438–440
as community's social context 4:386, 4:388
concept in social sciences 16:437–438
consequences of collective efficacy in 4:170–171
effects 16:440–441
effects on academic achievement 22:926
feedback effect of crime to collective efficacy 4:171
and gender 9:711
and health 22:278–279
influences in health care 15:103
Jacobs on 12:777–778
planners' view 16:438–440
research, in the twenty-first century 16:440–441
and residential segregation 20:540–542
change 20:541
and responses, mortgage foreclosure crisis on 16:453–454
and segregation 21:407–409
unit 16:439
urban 24:927–928, 24:930–931
urban poverty 24:915–918
Neighborhood big men 2:571
Neighborhood disorder
and distress 1:549
and mistrust 1:548–549
and powerlessness 1:548
Neighborhood effects 5:147, 7:169, 16:443–449
concentrated poverty 16:443, 16:446–447
on poverty 18:753–756
research of 16:446–447
research on 4:325
segregation 16:443–444
stratification 16:443, 16:446–447
theory and mechanisms 16:443–444
from theory to evidence 16:444–446
use of term 16:443
on work–family conflict 8:764
Neighborhood factors, fear of crime 8:855–856
Neighborhood renewal 24:921
Neighborhood revitalization 16:450–455
in Canada 16:454
techniques 16:453
in Western Europe 16:454
see also Gentrification
Neighborhood school 8:613
Neighborhood sorting 24:916
Neighbor-modulated fitness, kin selection 13:28
Neighbors 21:351
to social history 22:309–310
Neighbors of a node 16:572
Neisser, Ulrich 15:162, 19:415
works, *Cognitive Psychology* 4:104
NEJAC *see* National Environmental Justice Advisory Council (NEJAC)
NELM *see* New economics of labor migration (NELM) model
NELS *see* National Education Longitudinal Study (NELS)
Nelson, Chris, works, *The Bear Cult: Photographs by Chris Nelson* 13:890
Nelson, Diane M.
works
Reckoning: The Ends of War in Guatemala 25:124
War by Other Means: Aftermath in Post-Genocide Guatemala 25:124
Nelson, Laura 2:73
Nelson, Reed E., on bricolage 17:393–394
Nelson concept of organizational routines 8:430–431
Nelson Mandela model 18:837
Nelson–Aalen estimator 8:297, 8:299–300, 23:772–774
Nemeth, Charlan Jeanne, on networks and creativity 5:467
Nemeth-Brown, Brendan, on networks and creativity 5:467
NEO Personality Inventory 9:231–234, 17:827
revised 4:659–660, 8:625, 16:743
NEO Personality Inventory - Revised (NEO-PI-R) 9:237, 9:250–251, 17:827
Neo-Aristotelian freedom as capability 9:401
Neoassociationist approach, to language disorders 13:332
Neobehaviorism 13:703
and semantic differential 21:563–566, 21:568–569
Neo-Boasianism 1:749
Neo-capitalist cybernetic order 18:668
Neo-cargoism 3:143
Neocatechumenal Way 16:807
Neoclassic approach, to language disorders 13:332
Neoclassic economic theory 18:352
Neoclassical economics 1:77, 7:12–13, 7:39–40, 10:189–190, 12:388–389, 15:334, 15:452–455, 15:466, 18:359–360, 19:901–902, 25:11, 25:55–56
film and video industry analysis 9:158
market equilibrium 7:12–13
maximization 7:12–13
and renewable energy 18:368
of safety 20:863–864
evidence consistent with 20:863–864
modified 20:864–866
qualifications concerning 20:864
stable preferences 7:12–13
see also Marginalist economics
Neoclassical microeconomics 21:304–305
and Cold War 4:123
Neoclassical models 13:454–455
Neoclassical political economy 1:939–940
Neoclassical revolution 14:624
Neoclassical synthesis, and macroeconomics 14:404
Neoclassical theory, defense spending within 6:3–4
Neoclassical utility maximization frameworks, for internal migration 12:436
Neocolonial movements 25:838
Neocolonialism 11:682–683, 11:685–686
and social work 16:456–461
colonial history of social work 16:457–458
imperial history of social work 16:457–458
indigenization 16:458–459
international social work of liberation 16:459–461
and structural adjustment 23:599, 23:601–603
sustainable development 23:815
Neo-Confucianism, Maruyama Masao on 14:636

Neoconservatism 19:872
Bell on 2:504
Neoconstructionism 15:822–823
Neocorporatism 12:369, 12:374
Neocortex
cytoarchitecture of 3:340f
differential lamination of 3:341f
neuronal migration during ontogeny of 3:339f
Neocortical connectivity 22:485
Neocortical information processing
Sperry's contribution 23:280
Neocortical memory circuits 13:682
Neo-Darwinism 22:899, 22:901, 22:904, 23:694
and Fisher 9:223
and LGBT tourism 9:637–638
to modern human cognition 11:252–253
and organizational conflict 4:641–642
Neo-Durkheimianism 6:707
Neoempiricism
circles 14:314–315
Berlin Society for Empirical Philosophy 14:315
Lvov-Warsaw School 14:315
Vienna Circle 14:314–315
Neoenvironmental determinism 7:724
Neo-Eskimo 1:921
Neoevolutionism 1:769
tribes 24:646
Neofascism 18:411–412
Neo-Freudianism 19:362, 21:129
Neofunctionalism 9:541, 11:425, 17:524
Alexander on 1:525
Neogeography 3:619, 10:87–88
Neogrammarian framework of linguistic research 14:216
Neogrammarian philology 6:348
Neogrammarianism 6:349
Neo-Gricean pragmatic theory of anaphora 1:677–679
Neo-Harringtonians 3:63
Neohumanist–idealist reception of Greece 2:907
Neoinstitutional school in organizational analysis 17:418
Neoinstitutional theory 21:152–153
of interorganizational relationships 12:622
Neoinstitutionalism 11:683, 12:186–189, 13:600, 17:393–394
constructivist institutionalism 12:188–189
discursive institutionalism 12:189
economics 12:186–187
historical institutionalism 12:187–189
and meaning 16:562
organizational institutionalism 12:188
political science 12:187
rational-choice institutionalism 12:187, 12:189
sociology 12:187–188
Neo-Kantian theories 2:317–318, 19:392
Neo-Kantianism 11:382, 11:384, 13:135
and human sciences 11:381–384
Neo-Kantians 19:392
Neo-Kraepelinianism 13:130–131
Neo-Lamarckian theory, of heredity 8:227
Neo-Lamarckism 8:228
Neoliberal art of governing cities 24:846–847
Neoliberal capitalism 8:60, 24:818–819
and green social work 10:387
and property 19:155
Neoliberal economics, and structural adjustment 23:598, 23:602
Neoliberal globalization 25:368
and queer anthropology 19:756
and sex work 19:279
Neoliberal governmentality 3:640–641, 16:463–464
Neoliberal ideology 5:257, 6:303, 18:724
Neoliberal institutional economics 11:876
Neoliberal policy, choice as 3:523–524
Neoliberal propensity 19:641
Neoliberal theory 10:852
Neoliberalism 2:750, 3:582, 5:255, 5:385–386, 6:197, 6:277–278, 6:291, 7:118, 7:716–717, 10:221, 10:254, 10:257–258, 12:915, 14:15–16, 14:29–30, 15:63, 16:459, 16:462–468, 16:857, 18:181, 18:184, 18:940, 19:804–805, 20:673, 20:816–817, 23:349–350, 24:735–736, 24:799, 25:123
and anarchism 1:684
and artistic/cultural markets 14:566
in Central America 1:623–624
and community sociology 4:364–365
conceptualization of 16:462–464
as ideology 16:462–463
neoliberal governmentality 16:463–464
as policy 16:464
consequences of 16:466–467
and critical reflexivity 5:283, 5:285
dangers of 16:463
definition 16:462
and development 6:303–305
and discourse 6:503, 6:507
disposable youth 6:550–552, 6:554–555
East Asian studies 6:793, 6:795
and Eastern Europe social work 22:796–797
and economic reform 18:691–692
elites 7:388
and global south 10:187–190
and governmentality 10:309
and health 10:201–202
history 16:464–466
and human development 6:307, 6:309–310
and ideologies 11:576
and immigrant integration 11:642
impact on social work 22:520–521
informal sector 12:58
in Korea 2:73
and liberal democracy 16:463
in liberalism 16:462
and McDonaldization 14:829
in Mexico 1:626–630
and political economy 1:939–941, 18:365
post-cold war 24:846–847
progress 19:125–127
and radical social work 19:870–872, 19:874
reproducing 24:850–851
resisting 25:839–840
risk assessment 20:693
roll-back 24:848
roll-out stage of 24:848
and social work 24:491
and state 23:358–360
transnational corporations 10:221–222
and unemployment 24:733, 24:735–736
and urban policy 24:900, 24:903, 24:905
and welfare 25:477–479
and work anthropology 25:663, 25:665–666
Neoliberalization 5:258
of urban economies 24:849–851
Neolithic communities 23:66
Neolithic Europe 8:270
Neolithic period 17:236
South Asia 23:51–52
Neolithic revolution 1:723, 11:317, 22:338, 25:670
in artisans 2:46
Neolithic/polished Stone Age 14:260
Neolithic-population movements
migration 15:477–479
Neologism 18:55–57
Neo-Machiavellian view of organizational conflict 4:642
Neo-Malthusianism 6:860, 7:713–714
Neo-Malthusians 7:811
Neomanagerialism 2:750
NeoMarxian approaches
to trade policy 24:486
Neo-Marxism 9:400–401, 16:292–293
in anthropology 1:769
class and nation 16:292–293
class and peripheral nationalism 16:292
on education 7:147
and hegemony 10:747–749
sport 23:314–315
Neo-Marxist approaches 16:293
Neo-Marxist deskilling 25:691
Neo-Marxists 12:913
Neomercantilism 10:164
Neonatal bonding 16:469–474
affiliation 16:469
attachment 16:469–472
child development 16:469–473
definitions 16:469–470
humans: mother–infant attachment 16:471–472
innateness of 16:472
long-lasting effects of 16:472–473
maternal behavior 16:470–472
neuroendocrine basis in 16:470
olfactory learning in rats 16:471
oxytocin and vasopressin 16:470–471
psychopathology 16:469, 16:472–473
social behaviors 16:471–473
Neonatal crying behaviors 16:475–480
anatomical basis and neurophysiological control 16:476–477

Neonatal crying behaviors (*continued*)
attachment-promoting signals 16:475
characteristics 16:477–479
from clinical perspective 16:477
cross-cultural differences 16:476
crying as communicative behavior 16:475
crying as part of typical behavioral repertoire 16:476
from evolutionary perspective 16:480
excessive crying 16:476
motherese 16:480
mother–infant interaction 16:475–476, 16:480
neural vocal pathway 16:476–477
nonhuman primates 16:475, 16:480
pain cries 16:479
parental care 16:475
properties, developmental changes of 16:479
separation call 16:475
Neonatal imitation 11:621, 16:481–488
affect attunement 16:487
alteroception 16:487
circular reactions 16:483–484, 16:487
dialogue 16:486
expressive movements 16:482, 16:486–487
features of 16:486–487
history of thinking about newborn awareness and communication 16:482–483
ancient dispute about neonatal mimesis 16:482
fathers' diaries of infants' early life 16:482–483
innate intersubjectivity 16:483–484
mirror neurons 16:483–484
modern developmental psychology 16:483–484
love of shared knowledge 16:484
Piaget's rejection of neonatal imitation 16:484
psychoanalytic theory of neonatal identification with vital impulses of other 16:483–484
role of emotions 16:484
motor images 16:485
nature of human intelligence at birth 16:481–482
controversial topic in philosophy and psychology 16:481
developmental psychobiology of infant social mind 16:481–482
neonatal self-awareness 16:487
neurobiology of 16:487
primary dialogue 16:487
proprioception 16:487
provocation 16:486
research on 16:484–487
shared experience 16:487
sympathetic engagement 16:481, 16:486–487
unconscious emotionality and schemas for objects 16:483–484
Neonatal memory 6:759
Neonatal mortality
in industrialized countries 12:17–20
in less developed countries 12:22–23
Neonatal screening 9:927–928
ethics of 20:702
Neonatal social isolation 16:489–492
AMPA receptors synaptic delivery, disruption of
glucocorticoids 16:490–491
whisker experience-dependent 16:490
attenuation of whisker–barrel function 16:491
barrel cortex 16:489–491
borderline personality disorder 16:489, 16:491
corticosterone 16:490–491
GluA1 16:489–491
GluA4 16:489–491
neglect 16:489, 16:491
plasticity 16:489, 16:491
Neonates 16:493–495, 16:498–499
temperament 16:493–494
understanding 16:493
Neonationalism 16:292, 22:890–891
Neo-nativism, and modern human cognition 11:252–253
Neo-Nazism 18:410–412, 18:414
Neopatriarchy 16:385
Neopatrimonial regimes 5:14, 23:369
Neo-Piagetian theories, of cognitive development 16:501–510, 18:122–123, 16:502t
Case's staircase model 16:504
Demetriou's analysis of developing architecture of mind 16:505
dynamic self-organization of integrative psychological structures 16:508
dynamic skill theory 16:505–508
Halford's relational complexity model 16:504–505
intellectual development theory 16:501–503
critiques 16:501–503
Pascual-Leone's theory of constructive operators 16:503–504
psychological development 16:505–508
developing through abstract tier 16:507–508
development through representational tier 16:507
in infancy 16:506–507
tracking development of psychological structures 16:506
web of development 16:506
psychological structures are dynamic constructions 16:508–509
scheme 16:501, 16:503–504
skills 16:505–508
stage theory 16:501–503
working memory 16:504–505, 16:508
NEO-PI-R *see* NEO Personality Inventory - Revised (NEO-PI-R)
Neoplatonism 18:208
and astrology 15:119
Neopluralism 18:761
and political power 18:761
Neopositivism 22:994
and culture 5:584–585
kernel of 14:315–316
Neopositivist empiricism 9:533
Neopragmatism 20:761, 20:763
Neoracism, xenonphobia 25:797, 25:800
Neorealism 3:842, 20:26–30
Neorepublican ideals 20:480
Neorepublican institutions 20:480
Neo-Romanism *see* Civic republicanism
Neo-Schumpeterians, evolutionary economics 8:430
Neosexism 21:697
Neosexism Scale (NS) 21:697
Neoshamanism 21:869
Neoteny, and personality 17:912
Neotraditional town planning 16:809–810
Neo-Vygotskian theories 16:511–517
accomplishments 16:516
psychological tools 16:511–512
shortcomings 16:516
Neo-Whorfian hypothesis 13:293–299
NEP *see* New economic policy (NEP)
Nepal 23:61
child soldiers in 3:431–435
education in 7:196
fertility transition in 9:80–83
sociocultural aspects of *see* South Asia
Nepalese 7:196
Nephrology, psychosomatic disorders in 19:488
Nepotism 2:920–921, 5:14, 11:261
Néprajz 1:754
NEPSY-II *see* Developmental Neuropsychological Assessment Second Edition (NEPSY-II)
Nerd girls 4:726
Neri, Filippo 15:580
Nerotti, Giovanni Giacomo 15:579
Nerve connections, Sperry's contribution 23:279
Nerve growth factor (NGF) 2:837
Nerve–impulse measurements 10:756
Nerves transplants, Sperry's contribution 23:279–280
Nervios ('Nerves') 5:626
Nervism, doctrine of 17:608
Nervous diarrhea 9:605
Nervous system 1:510–512
Sherrington on 21:881–883
NESARC *see* National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)
Nesselrath, Arnold, works, *Census of Antique Works of Art and Architecture Known to the Rena* 2:24
Nested designs 8:539
Nested social dilemmas 22:246
NET *see* Narrative exposure therapy (NET)
Net additions to stock (NAS) 14:762
Net asset value (NAV) 2:439–440
Net benefits, computation in cost–benefit analysis 5:50–51
Net energy analysis 7:630–631
Net impacts 22:295

Net migration 6:121–122
Net neutrality 20:167–168
Net nutrition, and mortality rate 15:870
Net operating income 20:14
Net present value (NPV) 4:930, 5:56
Net primary production (NPP) 7:616, 13:240–242
Net reproduction rate (NRR) 18:572–573
Net widening 19:652
Netflix 25:81
The Netherlands
  arts participation in 5:448t
  cultural capital research in 5:378
  euthanasia in 7:600, 8:281, 8:283–284, 7:600b
  rates of hospital deaths in 7:597
  WH/WD in 7:600
Netherlands Indies, town development in 24:776
Netherlands Institute for Social Research 12:590
Netnography 8:182, 19:671
Netville study 4:335–336
Netwar 24:228
Network analysis 5:557–558, 16:518–531, 21:182
  analytical concepts and techniques 16:521–522
  approach in social sciences 16:518–519
  block model 16:519
  clique 16:519, 16:521
  economic behavior 16:520
  economic sociology 7:68–69
  in educational sociology 7:292–293
  ego-centered networks 16:520, 16:522
  exchange theory 16:520, 16:522
  graph theory 16:518–519, 16:521
  innovation networks 16:520–521
  interorganizational relations 16:519
  interpersonal relations 16:518–520, 16:522
  large networks and structural features 16:520
  matrix algebra 16:518–519
  measuring social capital 16:520
  narratives 16:211, 16:213
  origins and development 16:519
  relational aspects of power 16:520
  relative merits, limits, and perspectives 16:522–523
  social inequalities 16:520
  spatial 10:100
  structural analysis 16:519
  structural equivalence 16:519, 16:521
  structure and actor behavior 16:520
Network centrality measures 16:532–539
  betweenness 16:535–539
  binary-valued undirected networks 16:533–535
    medial measures 16:535
    radial measures 16:533–535
    degree centrality 16:533–535
    distance-based centrality measures 16:533
    measures involving reflection/feedback 16:533–534
  centralization 16:535–536
  closeness 16:534–539
  directed networks 16:536–537
  edge centrality 16:535
  eigenvector 16:534–535, 16:537–539
  foundations 16:532–533
  social capital 22:182–184
  two-mode networks 16:538–539
  valued networks 16:536
  vertex 16:532–539
Network centralization statistics 16:521
Network connection types, and power 15:668
Network data 22:462
  egocentric 22:462
  sociocentric 22:462
Network density 5:467
Network effects 1:800
  and business platforms 3:40
  in monopoly 20:160–161
  tDCS 24:534–535
Network embeddedness 4:864
Network exchange theory 8:484, 15:669, 22:476
Network flow model 16:531
Network governance 10:298–299
  academic research and employment 1:51
Network journalism 3:614
Network man 2:750
Network meta-analyses 8:330
Network models of tasks 16:540–545
  additive factor method 16:541
  constructing of network 16:541–542
    generalization 16:543
    process durations estimation 16:543
    single process, prolonging 16:541
    task network inference 16:541–543
    two process, prolonging 16:541–542
  critical path network 16:540
  multinomial processing tree 16:543
  PERT (Program Evaluation and Review Technique) network 16:540–541, 16:543
  processing tree representations 16:543–544
    inclusion/exclusion paradigm 16:543–544
    tree inference 16:544
  psychological refractory period 16:540
  representation of processes 16:540–541
  selective influence 16:541–542, 16:544
Network of Concerned Anthropologists (NCA) 11:392
Network science, and social media 22:424
Network search problem 23:262
Network society 12:371, 12:375, 13:107, 16:791, 18:668
  and Internet 12:601–603
  science communication in 21:222–223
Network statistics, descriptive 16:521f
Network structure 14:422–425
Network text analysis 4:770
  data matrix from 4:770t
Network theory, Collins's 4:215
Network-based economies 7:65
Network-centric simulation 4:485
Networked learning, lifelong learning 14:94–96
Networked learning communities 16:551–552
  building capacity and support 16:553
  collaboration 16:553
  inquiry habit of mind 16:553
  key enablers of 16:552–553
  leadership 16:553
  purpose and focus 16:552
  relationships 16:552–553
  working of 16:551–552
Networked religion 20:277
Networked world, and community 4:388
Networking 16:915–917
  informal sector 12:57–58
  for school improvement 16:551
  transnational 24:295–296
Networking logic 18:668
Networkism *see* Television
Network-level analysis 22:462, 22:465
Networks 4:139, 4:143, 6:666, 10:61–62, 12:621–625, 17:302–305, 21:343, 22:981–982
  actor-network theory *see* Actor-network theory (ANT)
  appropriation thesis, group identity 4:178
  commemorative 4:236
  commodity chains 4:243–244
  communication 7:867
  cultural analysis 9:363–364
  culture 14:526, 14:531–532
  and discourse 6:499–501
  entrepreneurship 12:172–173
  in entrepreneurship 7:688–689
  environmental movements 7:763
  geography of 16:567–571
    actor-network theory 16:569–570
    commodity chains 16:570
    Deleuze and 16:567–568
    geography as rhizomes 16:568–569
    globalization 16:567–571
    Guattari and 16:567–568
    internet 16:567–570
    post-structuralism 16:567–570
    power-geometries 16:570–571
    space of flows 16:570
  immigrant 12:515
  implementation 11:694
  interorganizational *see* Interorganizational networks
  and knowledge translation 21:183–184
  Mann on 14:490–491
  maritime 14:525–532
  and migration 15:417
  of motivated people, in relational social work 20:225
  narrative *see* Narrative networks
  organizational path dependence 17:385–389
  in organizations 2:769
  organized crime as 5:172–173
  of people, in disaster resilience 22:570
  of perspectives 12:603

Networks (*continued*)
policy 18:314–319
properties of 16:546–547
random networks 1:368
rational choice theory 19:901–902, 19:904
social exchange 8:484–485
social movements 22:439
and social structure 22:720, 22:724
transportation 4:274
and university press 19:624, 19:626–627, 19:629
see also Social networks
Networks, functions of 16:547–550
exploiting 16:548
actor position 16:548
connectivity 16:548
network effect 16:548
tie strength 16:548
exploring 16:547–548
centrality 16:547
clustering 16:547
small worldliness 16:547–548
strong ties 16:547
structural hole 16:547
weak ties 16:547
network analysis 16:546
network organizations 16:549
network theory 16:550
organizational forms 16:548–549
organizational networks 16:546
organizing 16:548–550
community 16:549–550
dyadic interfirm ties 16:549
formal ties 16:548–549
informal ties 16:549
interlocking boards 16:549
multifirm ties 16:549
Networks (Soviet) 20:853
Neue Apostolische Gemeinde 15:531
Neuman, W. Russell
works, *The Future of the Mass Audience* 25:81
Neumann, Franz 10:870–872
Neural activity
basis of visual sensitivity and thresholds 25:188
development and plasticity of visual neural organization 25:189
encoding of visual scene 25:184
modulation, by attention and task demands 25:188–189
for visual awareness 25:189
Neural aspects
perception of action space 17:705–707
frontal lobe mechanisms of spatial representation 17:706–707
parietal neurons encode spatial information in multiple reference frames 17:705–706
virtual spatial navigation 16:352–353
and functional neuroimaging 16:352–353
in patient populations 16:352
and structural neuroimaging 16:352
Neural bases, of reading
of anxiety 1:806–808, 1:811–817
of attention 2:163
of attention-deficit/hyperactivity disorder 2:155–162
of cultural transmission 5:391
of dreaming 6:651–657
of emotion 7:477–482
of facial emotion 8:668–669
of fear conditioning 1:811–817
of hemispatial neglect 10:766–772
of hunger and eating 11:420–422
of hypnosis 11:465–471
of inferencing 19:397–398
of language development 16:586
behavioral studies of children with focal injuries 16:587–588
genetic abnormalities and 16:589
language characteristics after cerebellar injuries in children 16:589
of mature adult language functioning 16:586–587
cerebellum 16:587
cortical regions 16:586–587
white matter pathways 16:587
of memory binding 15:144–145
of reading 16:593
illiterate brain 16:596–597
learning to read 16:593–595
text-level reading 16:595
word-level reading 16:594–595
neural systems in adult reading 16:593
reading in blind 16:597–598
reading in different script systems 16:595–596
Neural circuits
associated with perception of others' pain and distress 7:543–544
Cloninger's model 16:494–495
dopamine system 16:497
Hariri's model 16:495
and hysteria 11:508
serotonin receptor 16:495–497
serotonin system 16:495
serotonin transmission 16:497
serotonin transporter 16:497
Neural constraint-satisfaction 4:719–722
algorithmic variations 4:720–721
modeling with 4:721–722
Neural correlates of consciousness (NCC) 4:673
competing views of 4:674–676
contents of 4:674
core 4:674
experimental approaches 4:674
global views of 4:675
phenomenal 4:675
prefrontal only view 4:676
theoretical implications 4:676–677
versus unconscious states 4:673–674
'visual activity only' view 4:675–676
of visual awareness 4:675
Neural correlates of disgust 6:548
Neural correlates of social development 16:600–605
challenges and opportunities 16:604
functional brain processes 16:601
amygdala 16:603
anterior cingulate cortex 16:601–602, 16:602f
anterior insula 16:603, 16:603f
emotion and response selectivity 16:601–602
inferior frontal gyrus 16:602–603, 16:603f
knowledge of self and others 16:602
medial prefrontal cortex 16:602, 16:602f
neural plasticity and sociocultural learning 16:603–604
recognition of others 16:601
self–other mapping 16:602–603
specialized emotion areas 16:603
superior temporal sulcus 16:601, 16:601f
gray matter 16:600
situated brain 16:600
social brain 16:601–604
structural brain development 16:600–601
white matter 16:600
Neural effects, of tDCS 24:533–535
Neural efficiency of intelligence 12:304–305
Neural function, effects of neonatal social isolation on 16:489–492
Neural impairment 19:452–453
functional neuroimaging 19:452–453
structural neuroimaging 19:452
Neural interfaces, for opsins 17:271
Neural mechanisms
of cognitive development in infancy 4:58–65
of habituation 10:482–483
of tactile processing 24:14–15
Neural networks 4:88, 16:117, 22:703, 23:98
algorithms 1:530–531
artificial intelligence 2:35, 2:37, 2:39
computation 2:28–30
of fear 18:42
of fear-potentiated startle 8:860–862, 8:861f
learning paradigms in 15:963–964
parallel distributed processing, in cognitive psychology 4:107
Neural pathways, cardiac response 23:539
Neural plasticity 5:567, 16:619–622
and brain aging 2:796–797
and brain development 2:823–825
during brain development 16:619–620
brain repair and 16:620–621
developmental neuroscience 16:603–604
learning and 16:620
lifecycle and 16:620
mechanisms of 16:621–622
and rehabilitation 15:969
use of term 16:619
ways of promoting 16:622
see also Plasticity
Neural prosthetics 15:969
Neural representations of direction 16:623–627
Neural representations of objects 16:628–633
agnosia 16:632

coding by large number of neurons 16:630–632
category selectivity within ventral visual cortex 16:630–631
large-scale coding of objects 16:631
object selectively in human visual cortex 16:630
organizing principles of object selectivity 16:631–632
coding by small number of neurons 16:628–630
building blocks of object selectivity 16:629–630
hierarchical processing in visual cortex 16:629
faces 16:628–632
functional magnetic resonance imaging 16:630–632
functional role of localized brain regions 16:632
high-level vision 16:628–631
lateral occipital complex 16:630–631
perception 16:628–631
prosopagnosia 16:632
Neural studies of diffusion making 6:400–401
Neural substrates
of motor skill learning 16:848
of perceptual/cognitive procedural skill learning 16:848
of procedural learning 16:847–848
Neural synchrony, as binding mechanism 16:634–638
Neural systems
artificial/natural computation in 2:28
of sleep 22:87–93
subserving visuospatial languages 21:930–932
Neural vocal pathway
and neonatal crying 16:476–477
Neurasthenia 12:791
Neurath, Otto 18:625, 25:90–92
works, *Empirische Soziologie* 6:488
Neurites 16:704
Neuritic plaques 1:592
Neuroactive steroids
neuroendocrinology 16:652
Neuroanatomy
and anxiety 1:806–807
of dreaming 6:653–654
of OCD 17:114, 17:114f
paraphilias 17:953
of writing 1:455
Neuroanthropology blog 19:364
Neuroarchaeology 4:28
Neurobiological factors 17:593–594
Neurobiology 13:368, 13:373, 14:369
and ADHD 2:160–161
basis of ADHD 2:155–159
and cognitive psychology 4:107
correlates of psychotherapy 19:519
of emotions 1:237
of hunger and satiety 11:421–422
of hysteria 11:508–509
Kraepelin on 13:128, 13:130–131
of language learning impairment 13:347–348
of neonatal imitation 16:487
neuroimaging 14:439, 14:444
perspective of procrastination 19:77
research
linking mental illness epidemiology to 15:183–184
post-traumatic stress disorder 18:704
Neurochemistry
brain stimulation reward 2:842–843
of OCD 17:115
paraphilias 17:953
Neurocognitive poetics, visual word recognition 25:217–218
Neurocognitive psychology, of visual word recognition 25:214–219
Neuroconstructivism, Williams syndrome 25:579
Neurocultural theory 7:484
Neurodegeneration
in AD 1:591–592
and prosopagnosia 19:253–254
Neurodegenerative diseases, and prospective memory 19:275–276
Neurodegenerative disorders 18:792
Alzheimer's disease 1:584–585, 1:587–588
Neurodevelopment 7:848, 23:556–559, 23:556t
Neuroeconomics 16:331–332, 16:639–647
brain stimulation reward 2:844
diffusion process 6:400
Neuroeducation 21:119–120
Neuroendocrine basis, in neonatal bonding 16:470
Neuroendocrine dysregulation, and hysteria 11:509
Neuroendocrine rhythmicity, and hypothalamus 11:482
Neuroendocrine system
in mood disorders 7:607–608
depression 7:607
gonadal hormone levels with depression 7:607–608
gonadal hormones in postpartum depression 7:608
HPA function 7:607
sex differences in depression 7:607
stress 7:607
stress hormones in postpartum depression 7:608
in schizophrenia 7:608–610
endocrine disorders 7:610
HPA function 7:609
risk for psychiatric illness 7:610
role of gonadal hormones 7:609–610
sex differences 7:609–610
stress 7:609
Neuroendocrinology 16:648–655
biochemical properties of major classes of hormones 16:649t
chemical structure 16:648–649
classification and biochemical properties of hormones 16:648–649
definitions 16:648
feedback systems 16:652
functions of glandular hormones 16:649–651
adrenal cortex 16:651
adrenal medulla 16:651
pancreas 16:651
parathyroid gland 16:651
pineal gland 16:650–651
pituitary 16:649–650
thyroid gland 16:651
glandular hormones according to sites of production and chemical classes 16:649t
history 16:648
hormone regulation by limbic–hypothalamo–pituitary processes 16:650f
hormones as indicators of neurotransmitter and hypothalamic activity 16:654–655, 16:654t
interactions between hormone systems 16:653
neuropeptides, tissue peptides, and neurosteroids 16:651–652
neurotransmitter challenge tests 16:654, 16:654t
of post-traumatic stress disorder 18:704
relationships between steroid hormones of HPA and GPA axes 16:653f
rhythms 16:652–653
sleep disorders 22:82–86
Neuroepithelial cells (NECs) 2:821–822
Neuroesthetics 16:656–663
aims and scope 16:656
approaches to 16:657–658
brain 16:657, 16:659–662
decision making 16:661–662
Fountain (Duchamp photograph) 16:660f
future avenues for research in 16:661–662
history 16:657
neuroimaging 16:657–659, 16:661–662
neuropsychology 16:657
preference 16:656–657, 16:660–661
relation with cognitive neuroscience of art 16:656f
research topics 16:658–661
context 16:659–661, 16:660f
emotion 16:658–660, 16:659f–660f
expertise 16:661
Neuroethics 2:627–628, 16:664–669
behavior 16:664, 16:666–668
brain–computer interfaces 16:666
cognitive enhancement 16:665–666
deep-brain stimulation 16:665
genetics 16:664, 16:667–668
informed consent 16:666–667
neuroimaging 16:664, 16:667–668
neuroprosthesis 16:666
neuroscience of ethics 16:667–668
moral judgment 16:667–668
neurobiological correlates of free-will and potential forensic implications 16:668
neurostimulation 16:665
privacy rights 16:666–667

Neuroethics (*continued*)
psychopharmacology 16:664
psychosurgery 16:665
Neuroevolution, of empathy-related behaviors 7:542–543
Neurofibrillary tangles (NFTs) 1:591–593, 1:592f–593f
Neurofibromatosis type II (NF2) 16:715
Neurofunctional models of short-term memory 21:915–916, 21:915f
Neurogenesis 2:795–796, 20:57–58
and brain development 2:821–822, 2:824
Neurogenetics 16:493
Neurogenomics 2:827–828, 2:831, 16:715
brain, evolution of 2:827
Neurogestures 10:111
Neurohistology, Sherrington on 21:881
Neurohypnotism 11:465
Neuroimaging 2:627–628, 2:650, 9:525, 13:461–462
of autobiographical memory 2:286
in clinical psychology 11:603
functional imaging of mental disorders 11:605–609
future challenges 11:609–610
imaging techniques 11:603–604
neuroimaging research of mental disorders 11:604–605
in cognitive neuroscience 4:107–108
of cortical areas engaged in movement 5:21–29
for dichotic listening 6:364–365
evidence for Lexical consolidation 13:921–922
for hemispatial neglect 10:766, 10:768, 10:770
hypnosis 11:470
for language disorders 13:329–330
for language learning impairment 13:347
Luria on 14:398
mental imagery 15:221, 15:224–225
of moral cognition and emotion 16:609–613
and neuroesthetics 16:657–659, 16:661–662
and neuroethics 16:664, 16:667–668
and neuromarketing 16:698–700
and personality 17:861–862
for pragmatic disorders 18:792–794
of psychogenic amnesia 1:654–655
and reconceptualization 11:712
serial order memory 21:640
and source monitoring 23:32
frontal lobes 23:32
in spatial navigation 23:163–164, 23:166–170
in transient global amnesia 24:567–568, 24:568f
visual word recognition 25:215–217
Neuroimaging studies
of bilingual brain 2:576–577
of children with focal injuries 16:588–589
of dreaming 6:652–653
functional imaging 16:590–591
language development after injury to white matter pathways 16:589
of mental disorders
aims 11:604–605
to functional alterations 11:605
on posttraumatic stress disorder (PTSD) 11:605
scope 11:604–605
to structural alterations 11:605
prelinguistic language processing in infants 16:589–590
Neurolaw 13:489
and groups 13:493–494
and individuals 13:491–493
Neuroleptic malignant syndrome 19:467
and schizophrenia 21:20
Neuroleptics
with psychopharmacotherapy 19:459–460
for somatoform disorders 23:13
see also Antipsychotics
Neurolinguistic processing
of psychological verbs 16:670–677
Neurolinguistics 13:249, 14:177, 16:678–689, 19:357
acquired reading and writing disorders 16:685–686
agrammatism 16:683–684
antiquity 16:680–681
aphasia 16:678–687
approach to language disorders 13:332
bilingual brain 16:687
birth 16:678–679
brain stimulation 16:686
cognitive neuroscience 16:680
dysgraphia 16:685
dyslexia 16:685
hemispherectomy 16:686
language and right hemisphere 16:687
laterality 16:686
localization 16:678, 16:680–682, 16:684–686
modern times 16:687
nineteenth century 16:682–684
psychological verbs 16:672–675
Renaissance 16:681
Russian roots of 16:679–680
as scientific discipline 16:679
twentieth century 16:684–686
Wada test 16:686
Neurological approaches, to agrammatism 16:690–697
Neurological aspects, of multilingualism 16:62
Neurological cognitive mapping 4:82
Neurological conversion symptoms, of hysteria 11:507–508
Neurological diseases 6:77–78
Neurological disorders
dysexecutive syndromes 6:725–726
effects on serial order memory 21:641
of memory and language, cognition research 4:107
Neurological level of analysis 13:900
Neurological underpinnings 23:241
Neuromarketing 4:741, 16:698–702
anterior cingulate cortex 16:701
attention 16:698, 16:700
consumer neuroscience 16:698–699
consumer psychology 16:698–699
design and launch of new product 16:700f
electrodermal activity 16:700
electroencephalogram 16:698–701
emotion 16:698, 16:700–701
functional magnetic resonance imaging 16:698–701
galvanic skin response 16:701
heart rate 16:700–701
magnetoencephalogram 16:698–700
main cerebral structures involved in advertising and brand evaluation 16:701
memory 16:698, 16:701
methodologies 16:699
neurophysiologic signals used by 16:699–700
middle frontal gyrus 16:701
neuroimaging 16:698–700
ventromedial prefrontal cortex 16:700–701
Neuromaturational theory (NT) 15:971
Neuromodulation, attenuated 1:432
Neuromotor control 9:266–267
Neuron projections, and opsins 17:271–272
Neuronal activity
ion channels and molecular events in 12:710–715
activity and synaptic function 12:712–713
activity-dependent control of neuronal development 12:711–712
activity-dependent increases in neuronal Ca^{2+} 12:710–711
Ca^{2+} in aging and degeneration 12:713–714
cell death 12:714
neurotrophins 12:712–713
plasticity 12:710, 12:712–714
transient receptor potential channels 12:711
voltage-dependent channels 12:710–712, 12:714
Neuronal density of cerebral cortex 3:335f
Neuronal development 2:847
Ca^{2+} activity-dependent control of 12:711–712
effect of breastfeeding on 2:847–851
Neuronal growth and death 16:621–622
Neuronal nitric oxide synthase (nNOS) 16:751
Neuronal recording studies, for cognition 4:98–99
Neuronal replay hypothesis 22:64
Neuronal substrates of fear 8:869–870
Neuronal synchronization 9:483
Neurons 16:703–707
axon 16:703–706
basic construction 16:703–704
computational implications 16:706
computational neuroscience 4:483
dendritic tree 16:703–704, 16:706

electrical signaling results from conductance changes 16:704–705
ion channels 16:704–706
membrane potential 16:704
motor reflexes 16:703
neurites 16:704
Sperry's contribution 23:279, 23:281
synaptic inputs 16:705
synaptic integration 16:705–706
synaptic plasticity 16:706
Thorndike's contribution 24:306
vertebrate phyla 16:703
Neuropathy, and diabetes 6:337, 6:340
Neuropeptides 16:651–652
and anxiety 1:804, 1:807–808
Neurophysiologic signals used in neuromarketing 16:699–700
Neurophysiological control, of neonatal crying 16:476–477
Neurophysiological factors of mania 14:486
Neurophysiological markers, of language development 16:708–713
Neurophysiology
cognitive neuroscience 4:95, 4:99
of dreaming 6:654–655
of interoception 12:614–615
paraphilias 17:953
of perception 13:124–125
Sherrington on 21:881
Neuropil threads (NTs) 1:591–593, 1:592f–593f
Neuroplasticity 7:848–849, 11:613–614, 11:619
remedies, for language learning impairment 13:349
Neuropragmatics 13:401–402
Neuroprosthesis 16:666
Neuroprosthetics 16:714–721
Neuropsychiatric issues, of AIDS 1:492–493
Neuropsychic structure 17:870
Neuropsychological approach, to adolescent sexual risk 1:129
Neuropsychological assessment
academic functioning 16:725
behavioral assessment 16:726
cognitive assessments 16:723–724
goals of 16:722
of intellectual functioning 12:251–252
malingering 16:722–723
memory assessments 16:724–725
Mini Mental Status Exam (MMSE) 16:723
motor functions 16:725
personality assessment 16:725–726
report 16:726
in research and forensic contexts 16:722
selecting an assessment battery 16:723
Neuropsychological deficits, obsessive-compulsive disorder 17:123
Neuropsychological dissociations, anomia 19:148–149
Neuropsychological impairments, of ADHD 2:155
Neuropsychological Insights 16:639–640
Neuropsychological rehabilitation 2:813–820
brain–computer interaction 2:816–817
cognitive enhancement 2:817
virtual reality 2:816–817
Broca on 2:813–814
computational modeling 2:817
holistic approach 2:817
definition 2:813
errorless learning for people with language disorders 2:816
Foerster on 2:814
Franz on 2:814
future directions 2:817–818
Goldstein on 2:814–815
history 2:813
identity issues following acquired brain injury 2:817
Lishman on 2:815
Luria on 2:815
need for evaluation 2:815–816
Newcombe on 2:815
perception of emotion and its rehabilitation 2:816
Poppelreuter on 2:814
restitution of working memory deficits 2:816
Russell on 2:815
support in the early stages of dementia 2:817
theories and models influencing 2:816
theory-driven interventions for executive dysfunctions 2:816
traumatic brain injury 2:814, 2:816–817
Zangwill on 2:815
Neuropsychological testing 16:722–727
Neuropsychological tests 13:361
dysexecutive syndromes 6:725–726
Neuropsychology 19:415–416
of category learning and representation 4:540–541
cognitive neuroscience 4:95, 4:98–99
of interoception 12:614–615
of lexicons 13:925–928
and neuroesthetics 16:657
of psychogenic amnesia 1:653
serial order memory 21:639–640, 21:642
of transient global amnesia 24:565–566
Neurorehabilitation 16:714
Neurosciences 5:566–567, 5:934, 17:871, 21:882, 22:635–636, 23:111
application 21:883
of information theory in 12:108
and autism spectrum disorders 2:275
Bayesian models in 2:368–372
cognition, evolution of 4:2
cognitive 4:95–102, 13:489
computational 4:480–484
and cross-cultural variation 5:567
and education 12:204
of education 16:728–735
emotions 7:473–476
ethics 2:617
of ethics 16:667–668
moral judgment 16:667–668
neurobiological correlates of free-will and potential forensic implications 16:668
law and 13:489–496
Marr on 14:596–598
moral 15:794
and neural plasticity 5:567
neuromarketing 16:698–702
of placebos 18:154–155
and psychology 19:416
psychoneuroimmunology 19:445–450
and realism 20:25
and retroactive interference 20:617–618
studies of context effects 4:783–784
Neurosis, transference 24:546
Neurosociology 8:464
basic unit of analysis 16:736–737
definition 16:736
determinacy 16:741
human brain
complexity 16:737
emotions and reality in 16:740
evolution of 16:738–739
imitation 16:740–741
limitations 16:737
projection of 16:737
relational 16:737
social nature of 16:739
two halves of 16:737–738
measures 16:738
mirror neurons 16:739–740
new unconscious 16:738
persistent inequality 16:741
unconscious voice signals of status 16:741
see also Mirror neurons
Neurosteroids 16:651–652
Neurostimulation 16:665
for cortical area studies 5:23–24
Neurotic depression 2:793
Neuroticism 9:236, 16:493–494, 16:743–748, 17:763–764, 17:787, 17:802, 17:814, 17:816, 17:914–917, 17:921–923
and adolescent personality 17:864, 17:866–867
and anxiety disorders 16:744–745
change across life span 16:746–747
and cognitive processing 16:745
and coping 16:745–746
definition 16:743–744
and emotion regulation 16:745–746
and emotional disorders 16:744–745
and emotions 16:745
life outcomes 16:746
and marital outcomes 17:797
and mood disorders 16:744–745
and negative emotions 16:743–747
trait perspective 16:743–744
universality of 16:743–744
Neurotransmitter activity, hormones as indicators of 16:654–655, 16:654t
Neurotransmitter agonists 22:486–487
Neurotransmitter challenge tests 16:654, 16:654t
Neurotransmitter release 10:482
Neurotransmitter systems
and ADHD 2:159–160
interaction of HPA axis with 11:478–479

Neurotransmitters 16:494–495, 16:497–498, 16:749–754
acetylcholine 16:749–752
adrenalin 16:749, 16:751
and anxiety 1:804, 1:806–807
and brain development 2:823
calmodulin 16:753
catecholamines 16:749–750
classes 16:749–750
cyclic adenosine monophosphate 16:753
definition 16:749
dopamine 16:749–752, 16:754, 16:750f, 16:752f
gamma-amino butyric acid 16:750, 16:753
glutamate 16:750–751
ionotropic receptors 16:752–753, 16:753f
mechanisms of action 16:751–752
mechanisms of drug action 16:753–754
metabotropic receptors 16:752–753
monoamines 16:749–751
noradrenalin 16:749, 16:751
panic disorder 17:469–470
receptor agonists and antagonists 16:753–754
receptors 16:752–753
second messenger 16:753
serotonin 16:749–750, 16:754
systems 16:750–751
Neurotrophins 12:712–713
Neustadt, Richard
United States presidency 18:861–862
works
Presidential Power 18:861–862
Neutral connotation 4:223
Neutral experts 8:582
Neutral inquisitor 13:747–748
Neutral monism 18:107
Neutral theory 11:290
of molecular evolution 13:24–25
Neutrality 1:87, 4:458, 22:987–988
axiological 16:755
of science 21:212, 21:217
Neutralization
and genocide 10:16
in linguistics 14:533, 14:535
and white-collar crimes 25:554
Neutralizing systems of control 24:122–123
Neutron activation analysis (NAA) 1:906, 24:481
NEV *see* Normal–normal equal-variance model (NEV)
Never married, family 8:767–769
Never-ending wars 3:714
New absolutism, Robert A. Nisbet 16:835–837
New American voter 2:194
New antisemitism theories 1:787
New archaeology 1:892–893, 1:901–902, 14:671, 23:63, 23:745
ethnicity 8:103
New Atheism movement 3:542
New behavioral economics 2:446
New British history 7:653
New Caledonia 17:159
New careers 3:129
New Chronicle of Myanma (*Myanma Yazawinthit*) 11:83
New Classical inframarginal theory of specialization 6:602
New classical macroeconomics 14:403–404
New compatibilism, and mental causation 15:176
New conceptual architecture 3:585–586
New Confucianism 5:600
New constitutionalism 13:441
New critical theory 9:534
New Critics, in United States 1:619
New cultural geography 5:409–411
of place 18:150–151
New cultural history 11:104, 11:109, 11:112
New cultural sociology 11:106
New Deal (US) 13:740–741, 13:772, 18:299–300, 18:302–305, 19:532
policies 1:220
New Deal Democrat 1:605
New Democracies Barometer (NDB) 5:849
New Democratic Party (NDP, Canada) 8:154–155
New economic geography (NEG) 7:34, 14:288–289, 14:291
New economic history 7:55–59, 11:5, 11:104–105
New economic mechanism 19:156
New economic policy (NEP) 20:843, 22:840
New economics of labor migration (NELM) model 12:436–437, 15:452, 15:455, 15:466–467
New economy 18:665
work in 25:692
New England towns 24:855
New European order 1:754
New firm formation 7:686–688, 7:690
and value creation 17:364–367
building a productive base 17:365–366
creating and sustaining returns 17:366
early organizational development 17:364–365
ecosystem evolution 17:367
firm in its selection environment 17:366–367
fitness, issue of 17:366–367
opportunity recognition 17:364
organizational evolution 17:367
organizing productive activity 17:366
New genetics 1:758, 9:990–991, 9:1015–1021
biobanking 9:1018–1019
biocapital 9:1016
cloning 9:1016
ELSI 9:1017–1018
forms of life 9:1016–1017
functional genomics 9:1017–1019
gene therapy 9:1017
genetic citizenship 9:1019
genomics 9:1015–1019
Human Genome Project 9:1016–1019
life forms 9:1016–1017
and race 16:759–760
crime and medicine 16:760
descent 16:760
history 16:760–761
identity 16:761–762
life sciences 16:759
origin 16:761–762
populations 16:759–760
race, ancestry, and human biological variation 9:1018
reproductive technologies 9:1016–1017
science, citizenship, and subjectivity 9:1018–1019
scientific governance 9:1016
scientific research and clinical practices 9:1016–1018
governance 9:1017–1018
reimagining of life 9:1016–1017
transforming human capacities 9:1017
social conditions of enablement 9:1018–1019
structural genomics 9:1017–1019
tissue culture 9:1016
New Green Revolution 9:309–310
New Guinea
archaeology of 2:229–233
colonization 2:229–231
dating of arrival of humans 2:229–231
highlands 2:230
Holocene continuities and changes 2:232–233
horticulture 2:231
hunter–gatherer 2:231
intensification 2:230–232
island Melanesia before Lapita 2:230
megafauna 2:231–232
Pleistocene on 2:229–233
rock art 2:232–233
routes of entry of humans 2:230
Sahul on 2:229–231
Sunda on 2:230
watercraft 2:230
cultural evolution 5:398–399
New historicism 1:935
New history 22:308–310
New home economics (NHE) 11:224–225
New immigrants 8:225
New immigration 15:448
New industrial districts 7:688
New institutional economics (NIE) 11:561, 11:563–564, 11:567, 17:190, 17:193, 19:154–155, 20:863
of safety 20:866
Coase theory and transaction costs economics 20:866–867
evidence concerning changes in liability rules 20:866–867
New institutionalisms 11:106, 11:561–572, 12:183, 17:7–8, 22:979–983, 23:725
in economics 17:190–195
and old institutionalism, demarcation between 17:193–194
elements of 22:980
institutional change 22:982–983
multilevel causal model 22:980–981
norms and formal rules 22:980–981
norms and networks 22:981–982

origin of norms 22:979–980
in social sciences 11:561
New International Economic Order (NIEO) 6:309, 10:186–187
New internationalisms, in South America 23:48
New Jersey Negative Income Tax Experiment 8:531–532
New Jersey Pinelands 10:451
New journalism 3:612–613
New Labour 19:524
New Left
and critical psychology 5:262, 5:264
political radicalism 25:816
and youth culture 25:813, 25:815–816
New Left Review 11:45, 14:649
New Legal Realism (NLR) 7:556
and law 13:797
New liberalism 18:299–306, 24:848–849
conflicting approaches to framing the social question 18:300–302
corporate liberalism 18:305
Keynesianism 18:304–305
laissez faire 18:299–301, 18:305
liberal statism
containment of 18:304–305
limits of 18:303–304
market failure and drive for new social knowledge 18:299–300
New Deal 18:299–300, 18:302–305
planning 18:303–305
progressivism 18:299–302, 18:305
turn to social insurance 18:302–303
welfare state 18:302–305
New managerialism (NM) 16:770, 20:693
critique of 16:773–775
competition 16:774–775
management 16:773–774
performance measurement 16:773–774
key concepts of 16:770–771
in public sector organizations 16:771–773
approaches to managing 16:772–773
extension of competition in 16:772
managing performance 16:771–772
measuring performance 16:771–772
New Marriage Law (1950) 3:518
New media 15:33, 16:776–777, 16:793–804, 18:668, 22:585–586
Arab Spring and 16:778–779
audiences 2:218
celebrity 3:274, 3:277
and democracy in Arab World 16:776–781
and digital divide 16:787–792
identity and 16:777–778
on political mobilization in Arab Spring 16:798
ICT 16:799–800
information communication technology 16:798–799
regimes response 16:779–780
regional authoritarianism and 16:776–777
and religion 20:273–278
authority 20:273–274
community and identity 20:274
current approaches to study religion 20:276–277
future areas of research 20:277
and Internet 20:275–276
online culture 20:273–274, 20:276–277
religious practice 20:273–277
ritual 20:274–275
social media 20:274, 20:276–277
use 20:273–275
and social capital 16:782–786
and social protest 22:512–517
see also Digital media
New media scholarship 16:793–795
New megacities 3:598–599
New Mexico Environmental Law Center 7:760
New middle class 2:786, 10:47–49
green parties 10:379, 10:381, 10:383
New millennium planning 24:866–868
New mobilities paradigm 15:639
New modernism 15:688–689
New motives, emerging 12:581
New museology 16:145–146
New Nations 21:884, 21:886
New new new religions 16:805
New new religions 16:805
New paradigm partnerships 22:358
New persons 3:656–657
New political culture (NPC) 4:364
New political history 19:695–696
New political institutionalism (NPI) 11:561–562
calculus approach in 11:562
cultural approach in 11:562
New product-development approach 12:146
New Proposals: Journal of Marxism and Interdisciplinary Inquiry 14:649
New public governance 13:796–797
New public health 6:879
New public management (NPM) 2:920, 2:922, 3:698, 3:700, 10:846–847, 16:770, 19:523, 21:652, 22:696, 22:779, 24:493, 24:921
academic research and employment 1:49
reforms, and public sector organizations 19:597–598
New racism 19:864–865, 25:799–800
New realism 18:83–84, 20:21
New regional geography 20:123–125
New regionalism 17:303, 20:125–126
New religions in East Asia 20:370–371
New religious divide, issue evolution 12:767
New religious movements (NRM) 16:805–808, 20:360–362
changes 16:807–808
characteristics 16:807
classification 16:807
dichotomous worldviews of 16:807
number of 16:805–806
opponents of 16:806
problems related to term 16:805
reactions to 16:808
use of term 16:805
The New Republic 25:346
New Right theories 14:30
New social history 11:104
New social movements (NSM) 3:521, 6:316–317, 7:766, 10:379–381, 17:644, 17:646, 21:178–179, 22:447
modernization theory in 15:705
research 3:103
New socialisms, in South America 23:48
New socialist man 22:851–852
New sociological institutionalism (NSI) 11:561, 11:563, 11:567
New sociology of education 7:290
New South Wales Aboriginal Education Consultative Group Incorporated (NSW AECG) 15:903
New spirit of capitalism 1:49, 2:750, 5:536
New spirituality 3:543
New states 23:354
civil service 3:696
New Testament 3:233, 3:237–238, 3:548, 20:335, 20:398
New theoretical framework 3:586–587
New unified theory (NUT) of sociobehavioral forces 22:878–885
elements 22:879
personal qualitative characteristics 22:879
personal quantitative characteristics 22:879
New urban politics 24:850
New urbanism 16:809–814, 18:193–194, 23:663–664, 24:867, 24:938, 16:810t
challenges and critiques 16:812–813
community design 16:809–810, 16:812–814
future of 16:813–814
neotraditional town planning 16:809–810
origins of 16:809–811
practice 16:811–812
smart growth 16:809–813
transect planning 16:812
transit-oriented development 16:809–810
New urbanist critics 24:909–910
New urbanists 2:307
New *versus* old destinations, in United States 11:631
New vitalism 5:646
New wars 3:714, 15:521–522, 24:954
New wave extreme right 18:412–415
New woman ideal 20:837
qualities 20:841
New World 1:639, 2:145–146, 11:449, 23:44–45
carnival in 3:146–148
migration from Europe to 15:457–461
and multiculturalism 16:30
New World Information and Communication Order (NWICO) 14:709, 15:55–56
New World Negro 1:639
New world of reference publishing 7:585–586
digitization of content 7:586
digitization of labor 7:586

New world of reference publishing (*continued*)
recent changes in reference publishing 7:585–586
shift toward electronic libraries 7:586
New world order 19:796, 20:402–403
New York City Teaching Fellows program 24:76
New York Longitudinal Study (NYLS) 24:181
New York Museum of Modern Art (MoMA) 24:886
New York Peace Society 17:643–644
New York Police Department (NYPD) 3:582
New York Review of Books 22:902
New York school of relational sociology 16:562
New York State Commission 17:28
The New York Times 2:82, 12:777, 23:459
The New York Times Book Review 2:510
New Yorker (magazine) 15:249–250
New Zealand 3:540, 17:158–159, 17:161–162
Anglican Church 3:543
Maori social work 14:505–511
New Age varieties 3:544
religious affiliation and congregational involvement, decline of 3:544
Roman Catholics in 3:544
New Zealand Greens 10:383
New Zealand Maori 11:785–786
Newborns 16:390–391
and NIRS 16:392
orbitofrontal brain activation in 16:390
screening of 9:997
speech perception in 16:389–390
see also Neonates
Newby, Howard 4:362
Newcomb, H., works, *The Producer's Medium* 24:165
Newcomb problem 5:97, 18:933–934
Newcombe, Freda Gladys 16:685
neuropsychological rehabilitation 2:815
works
Missile Wounds of the Brain: A Study of Psychological Deficits 2:815
Newell, Allen 12:304
origins of artificial intelligence 4:103–104
Newell, Sue, on epistemic cultures 7:876
Newell and Simon's theory 19:65
Newest Vital Sign 10:654
Newly arising religion, use of term 16:805
Newly industrialized countries (NICs) 5:528
Newly industrialized economies (NIEs) 6:788–789
Newly industrializing countries 3:100–101
Newman, Oscar 20:775
Newman. Oscar, works, *Defensible Space: Crime Prevention through Urban Design* 20:775
Newman, Peter 24:866–867
Newmeyer, Frederick, works, *Linguistics in America* 14:216–217
Newport Folk Festival 3:802
News 16:819–823
content and delivery 15:24–25
definition 16:819–820
diachronic and international perspectives 16:821–822
in electronic media 16:822
paradigm 16:820–821
and politics 14:712–718
production and consumption 16:793–797
News conferences and briefings 16:815, 16:817
News interview 16:815–818
broadcast 16:815
origins and institutionalization 16:815–816
practice of interviewing 16:815
roles of interviewer and interviewee
answering and resisting 16:817
participation and turn taking 16:816
question design and journalistic norms 16:816–817
question design and press–state relations 16:817–818
News journalism in Continental Europe 16:821
News media bias 14:714–715
News paradigm 16:820–821
News theme 15:26
Newsome, Gavin 3:660
Newspapers 11:574–575
commercial 12:852
and community 4:332–333
and journalism 12:855
and political communication 18:343–345
sport history 23:308
Newspeak 4:124
Newsweek 2:510
Newsweek International 11:437
Newswork 3:614–615
Newsworkers 12:857
Newton, Esther, works, *Mother Camp* 19:755
Newton, Isaac 7:567, 18:101, 20:286, 21:173
on mechanics 18:101–102
philosophical discussions and sermons of 18:101–102
on picture of a universe 20:286
works
Opticks 18:101–102
Principia Mathematica Naturalis Philosophiae 18:101–102, 20:286, 21:163
Newton, J., works, *Sex and Class in Women's History* 9:768–769
Newton, S. 3:316
Newton–Euler recursive algorithm 3:780
Newtonian paradigm 16:584–585
Newtonian synthesis 14:416
Newtonianism 18:101–103
Newton–Raphson algorithm 16:128
Newton's Laws 3:779, 10:44
Nexus
of contracts 1:347
Dilthey on 6:425–429
Neyman, Jerzy 9:626, 11:494–495, 19:563, 23:425–426
in Berkeley 16:825
on confidence intervals 16:824, 16:827
on fiducial inference 16:827
and Fisher 16:824–827
influence of 16:827
life of 16:824–825
on mathematical statistics 16:825, 16:827
Neyman optimal allocation 16:827
Neyman-Pearson lemma 16:826
and Pearson 16:824–826
works
The Grammar of Science 16:824
in Russia and Poland 16:824–825
scientific contributions of 16:826–827
hypothesis testing 16:826
sampling theory 16:827
statistical estimation 16:826–827
on smooth tests 23:398–400
on survey sampling 20:876–877
University College London 16:825
Neyman–Barton smooth tests 23:397–400
Neyman–Pearson lemma 3:248, 12:800, 15:654, 16:826
Neyman–Rubin model 3:252–253, 4:446
NF2 *see* Neurofibromatosis type II (NF2)
NFC *see* Need for cognition (NFC)
NFFO *see* Nonfossil fuel obligation (NFFO)
NFI *see* Normed fit index (NFI)
NFO *see* Need for orientation (NFO)
NFP-AN *see* Nonfat-phobic anorexia (NFP-AN)
NFTs *see* Neurofibrillary tangles (NFTs)
nfvPPA *see* Nonfluent/agrammatic variant of primary progressive aphasia (nfvPPA)
NFVS *see* National Family Violence Survey (NFVS)
Ng, Y.-K.
on division of labor 6:604
Nga Pumahara o Nga Tamarikitanga 15:152–153
Ngandu people, childcare of 3:396
NGF *see* Nerve growth factor (NGF)
NGO *see* Nongovernmental organizations (NGO)
NGOs *see* Nongovernmental organizations (NGOs)
NGP *see* National Genographic Project (NGP)
NHANES *see* National Health and Nutrition Examination Survey (NHANES)
NHANES III *see* Third National Health and Nutrition Examination Survey (NHANES III)
NHE *see* New home economics (NHE)
NHGRI *see* National Human Genome Research Institute (NHGRI)
NHP *see* Nottingham Health Profile (NHP)
NHPs *see* Nonhuman primates (NHPs)
NHS *see* National Health Service (NHS) (UK)
NHSDA *see* National Household Survey of Drug Abuse (NHSDA)
NHSLS *see* National Health and Social Life Survey (NHSLS)

NIA *see* National Institute on Aging (NIA)
NIAAA *see* National Institute of Alcohol Abuse and Alcoholism (NIAAA)
Nianershi kaoyi 11:49
Niarchos, Stavros 14:531
NIBRS *see* National Incident-Based Reporting System (NIBRS)
Nicaragua 22:623
  sociocultural aspects 1:621–625
NICE *see* National Initiative for the Care of the Elderly (NICE); National Institute for Health and Care Excellence (NICE)
Nice Carnival 3:146
NICHD *see* National Institute of Child Health and Human Development (NICHD)
Niche construction 1:88, 16:332
  by humans 11:256–257
  use of term 1:86
Niche construction theory 15:340
Niche journalism 14:716
Niche-specialization hypothesis, of Dark Triad 5:674
Niche-width dynamics, organizational ecology 6:892
Nichiren 2:889
Nicknames 16:184–185
Nicole, Pierre 12:387
Nicotiana 16:753–754
NICs *see* Newly industrialized countries (NICs)
Nidotherapy, for personality disorders 17:896
NIE *see* New institutional economics (NIE)
Nielsen, Arthur 2:213
Nielsen Corporation (A.C. Neilsen) 1:191
NIEO *see* New International Economic Order (NIEO)
Niepce, Joseph Nicéphore 20:420–421
Nierenberg, William 7:793–794
NIEs *see* Newly industrialized economies (NIEs)
Nietzsche, Friedrich Wilhelm 6:426–427, 11:128–129, 16:829–831, 19:411
  on alienation 1:554
  analysis and classification 11:129
  biography of 16:829
  of emotional state 16:830
  of experimental philosophy 16:830
  illness and its impact 16:830
  on metaphors 15:290–291
  philosophy 15:291
  of relativism 20:233
  sister's relationship to the sick man, effect of 16:830
  study of theology and philology in Bonn 16:829
  of virtue ethics 25:147, 25:149–150
  of wisdom 25:600–601
  works
    The Antichrist 16:830
    Beyond Good and Evil 1:554, 16:830
    The Birth of Tragedy 16:829
    Dawn 16:830
    Dionysian Dithyrambs 16:830
    Genealogy of Morals 1:554, 1:749
    Human, All-too-human 16:830
    Hymn for Life 16:830
    Mixed Opinions and Sayings 16:830
    Observations of World History 16:829
    On the Future of Our Educational Institutions 16:829
    On the Genealogy of Morals 16:830
    On the Pathos of Truth 16:829
    Thus Spoke Zarathustra 1:554, 16:830
    Twilight of the Idols 16:830
    Untimely Meditations 16:829–830
    The Wanderer and his Shadow 16:830
    Warning to all Germans 16:829
    With Thoughts Concerning Moral Prejudices 16:830
Nigeria
  affordable private schools in 1:269
  education in 7:199
  linguistic diversity in 1:300
  population momentum values 18:573
Nigerian Muslims 12:739
Nightingale, Florence 16:909
Nightmares, post-traumatic stress disorder 18:702, 18:704–705
Night-watch system 18:261
Nigrescence 19:834
  stages of 19:834
NIH *see* National Institutes of Health (NIH)
Nihilism 14:842–843
  in West Africa 1:302
NIJ *see* United States National Institute of Justice (NIJ)
Nile river system 25:444–445
Nile Valley 1:691
  settlements in 1:693
Nilotic 6:773, 6:775–776
NIMBY *see* Not-in-myback-yard (NIMBY)
NIMBYism (not-in-my-backyard syndrome) 16:813
NIMH *see* National Institute of Mental Health (NIMH)
Nine Beliefs of Hinduism 20:333
9/11 attacks 3:678, 25:379
  and flashbulb memory 9:255, 9:257–258
Nineteenth-century
  late, German homosexual rights movement in 9:644–645
  psychophysics, experimentation in 8:566–567
Nineteenth-century formalism 13:741
NIPT *see* Noninvasive prenatal testing (NIPT)
NIRS *see* Near-infrared spectroscopy (NIRS)
Nirvana 2:887, 19:308, 24:726
Nisbet, Robert A. 4:363, 15:709, 16:832–837, 22:829
  against constructivism 16:834–835
  identity, defining 16:832
  intermediate institutions 16:835
  as the loose individual 16:835
  new absolutism 16:835–837
  and notion 22:829
  organic solidarity 16:834
  and society 22:829
  structural defects of late modernity 16:835–836
  structural functionalism 16:834
  sui generis conservative 16:833–834
  works
    Contemporary Social Problems 15:248–249
    Present Age 16:832–833, 16:835–837
    Social Bond 16:832, 16:835, 16:837
    Sociological Tradition 16:832, 16:834–835
    The Twilight of Authority 16:833, 22:829
Nishi Amane 24:583–584
NISVS *see* National Intimate Partner and Sexual Violence Survey (NISVS)
NIT *see* Negative Income Tax (NIT)
Nitric oxide synthetase (NOS), and Ca^{2+} channels 12:714
Nitrogen oxides (NO_x) 7:773
Niue 17:158–159, 17:162
Nivelle, Robert-Georges (1856–1924) 9:214
Nivkh 25:645
Nixon, Richard 18:461
Niyoga 23:58
Nkoli, Simon 21:745–746
Nkonyane, Zoliswa 21:748
NL *see* Negative liberty (NL)
NLC *see* National Labor Committee (NLC)
NLHCS *see* National Lesbian Health Care Survey (NLHCS)
NLMs *see* National liberation movements (NLMs)
NLR *see* New Legal Realism (NLR)
NLS *see* National longitudinal surveys (NLS)
NLS-72 *see* National Longitudinal Study of High School Class of 1972 (NLS-72)
NM *see* New managerialism (NM)
NMAIA *see* National Museum of the American Indian Act (NMAIA)
NMDA *see* N-methyl-D-aspartate (NMDA) receptor
N-methyl-D-aspartate (NMDA) receptor 2:693–694, 8:850–851, 12:711–713, 14:359–364, 16:620
NML *see* Normalized maximum likelihood (NML)
N4-model of model 14:775, 14:775f
NMSC *see* Nonmelanoma skin cancers (NMSC)
NNCS *see* National Neighborhood Crime Study (NNCS)
nNOS *see* Neuronal nitric oxide synthase (nNOS)
No Child Left Behind Act (NCLB) 1:46, 1:232–233, 21:71, 23:215, 24:75, 24:78
No opinion filter 23:766–767
No Secrets guidance 7:329, 7:332
"No self" concept 19:402
NOAA *see* National Oceanographic and Atmospheric Administration (NOAA)
Noachian Flood 20:286–287, 21:172–173
Noah, Mordechai 2:148
NOBC *see* National Organization of Bar Counsel (NOBC)

Nobe, Shuichi 10:110
Nobel Prizes 1:91–92, 7:58
  for medicine 14:367
  for natural sciences 23:227
  for peace 17:640, 25:842
Nobilitas 1:966
Nobility 1:965
  and conservatism 4:689
  European history of 1:965–969
  and gift 10:140–142
  origins of 1:965
Nobiltà, Italian 1:965
Noble death 3:546
Noble man *versus* mass man 7:394
Noblesse 1:965
Noboru, Miyata 12:794
Nocebo effect 18:153–154, 18:156
Nocebo hyperalgesia 18:154
Nociception, and pain 17:451–454
No-corruption equilibrium 5:14–15
Nodal policing 18:249
Node connectivity 3:907–908
Node-level analysis 22:462–465
Nodes
  crime at 20:774–775
  in narrative networks 16:212–213
  social networks 22:461–462
Nodes of Ranvier 2:29
N-of-one-at-a-time designs 21:984
Nogo receptors 7:850–851
NOHARM *see* Nonlinear factor analysis (NOHARM)
Noise 22:247
  and cardiac defense 23:540–544, 23:541f–543f
  information theory 12:106
  and sleep 22:61
  tolerance 17:135
  and transportation 23:847
Noise generation, stochastic optimization for 23:474
Noise trials 21:946
Nomadism 16:838–843, 17:578
  in Altai 16:838
  and cattle domestication 16:840
  and clan farming 16:838–839
  and domestication of herd animals 16:840
  emergence of 16:838
  ethnoarcheological findings 16:839–841
  historical interpretation and evidence 16:841–842
  in Old World dry belt 16:838
  and pastoralism 17:577–578, 17:580–582
  and plant domestication 16:840
  Roma 20:754–755
  study of 11:156
Nomads 14:134, 16:838
Nomen facere 19:179–180
Nomenclatura/nomenklatura 2:788, 12:278, 20:844, 20:852, 22:845–846
Nomenclature des Activités économiques dans les Communautés Européennes *see* NACE (Nomenclature des Activités économiques dans les Communautés Européennes)
Nomina 19:179–181
Nominal approach 4:711
Nominal classification 3:811–812
  semantic profiles of 3:813t
  subtypes of 3:813t
Nominal number 17:43
Nominal quantifiers 19:708
Nominal scales, in ethnography 8:189–190
Nominalism, legal reasoning 13:779
Nominalist critique of realism 20:204
Nominalist thesis 22:230
NOMINATE scores 4:652–653
Nominated policing 18:248
Nominative (NOM)–accusative (ACC) patterns 10:317
Nominative-accusative case 11:825
Nomological explanations 21:303
Nomological network
  of cultural intelligence 5:435–438
  five-factor model (FFM) 2:563–564
Nomosphere 13:476–477
Nomothetic approach
  in critical realism 5:278, 5:280
  in geographic information systems 10:62–63
Nomothetic science
  Luria on 14:396, 14:398
  regional geography as 20:122–124
Nonacademic contexts, epistemic cultures in 7:876–877
Nonacademic intellectuals 12:279
Nonacademic self-concept 21:460–461
Nonaccelerating inflation rate of unemployment (NAIRU) 13:19, 24:739
Nonaccidental properties (NAPs) 25:154
Nonadaptive evolution 8:383
Nonadditive determinism 22:486
Nonadditive genetic effects 9:919–920, 12:298
Nonadherence
  patient 17:598–601
Non-adjacent dependencies 2:651–652
Nonagentive functions 21:344
Nonaligned movement (NAM) 10:186–188
Nonalphabetic scripts
  reading 20:1–4
Nonalphabetic scripts, reading
  cognitive and neuropsychological studies
    writing 20:3
    reading 20:3
  Japanese language 20:1
  Japanese writing system 20:1
  Kana scripts and reading 20:2–3
  Kanji script and reading 20:1–2
  writing, cognitive and neuropsychological studies 20:3
Nonalternating lemmas 15:827
Nonaltruism, and intergenerational transfers 8:827–828
Nonanthropocentric perspective 18:194, 18:196
Non-articulatory processes 23:251
Non-Austronesian languages 17:447
Nonbasic color terms 4:234
Nonbasic sciences 3:261
Noncapitalist 15:717
Nonce borrowing 3:921–922
Noncharismatic personalism 17:202
Non-Christians 3:535
Nonclinical applications of stress inoculation training 8:608
Noncognitive abilities, and health 22:277
Noncognitive predictors, of educational outcomes 17:787
Noncognitive skills 17:784
  and beauty 2:385–386
Noncognitive trust 24:665
Noncognitivism 8:72
  of Albert 1:498
Noncolor meanings of color terms 4:234
Noncombatant experience 3:431–432
Noncommunicable diseases (NCDs) 7:815–816, 10:151–152, 10:617, 18:91
  and public health 19:547–548
  UN review 24:752
Noncommunicative processes, and collective memory 4:190
Noncompensatory multidimensional models 19:438
Non-Compete Covenants 11:841
Noncompliance 4:445, 4:453–454, 4:445f
  addressing in practice 8:530
  and arms control 1:976
  difficulties with 8:527
  and early disciplinary interactions 6:15
  side effects (psychopharmacotherapy) 19:463
  as underdosing 8:529–530
  see also Nonadherence
Nonconceptual content, tacit knowledge 13:77–78
Nonconditioned learning, and cerebellum 2:96–97
Non-configurational languages 23:889
Nonconflict activities 15:502–503
Nonconjugal relationships recognition, and LGBT aging 13:865–866
Nonconsensual sterilizations 8:251
Nonconsequentialism 4:680
  collective problems 4:682–683
Noncontinuous effect indicators 11:752
Noncontinuous indicators 11:752
Noncontinuous latent variable 11:752
Nonconventional oil and gas 17:166–167
Noncooperative games 4:865, 9:582–586, 12:100
  experimental economics 8:549
  politics modeling 18:513
  relative deprivation 6:215–216
Noncore topics of population censuses 3:292
Noncorporate business forms 4:934
Noncultural rationales 5:463
Nondecision making, power 18:766
Nondeclarative memories 5:924, 15:147, 16:844–850

basal ganglia 16:847–849
consolidation, and sleep 22:66–67
and early memory development 6:759
habit learning 16:844
Huntington's disease 16:847
implicit memory 16:844, 16:847
motor procedural skill learning tasks 16:845
Parkinson's disease 16:847–848
and proactive interference 19:2–3
in perceptual representation system 19:2
in procedural memory 19:3
processes and mechanisms of procedural learning 16:846
consolidation of procedural knowledge and sleep benefit 16:846–847
importance of practice for development of skill 16:846
stages 16:846
question of awareness 16:848–849
retroactive and proactive interference 16:847
neural substrates of motor skill learning 16:848
neural substrates of perceptual/cognitive procedural skill learning 16:848
neural substrates of procedural learning 16:847–848
retroactive interference in 20:614–615
implicit memory 20:614
procedural memory 20:614–615
and sleep 22:63
tasks used for assessment 16:846–847
types 16:844–846
see also Implicit memory
Nondeclarative memory systems
proactive interference in 19:1–2
Nondemocratic judge 12:880
Nondemocratic regimes, civil service 3:696
Nondestructive methods for forensic archaeology 9:335
Nondeterministic polynomial (NP) 23:174
time-completeness, algorithms 1:532
Nondirective interviewing 12:672
Nondirective psychotherapy, Rogers on 20:737, 20:742
Nondirectiveness 9:909, 9:911, 9:913
Nondiscrimination 13:436, 13:439
Nondiscriminatory wage structures 6:538–539
Nondominant ethnic groups 16:288–289
Nondominant ethnic minorities, and mental health disorders 5:570
Nondurable goods index, functional data analysis 9:516
Non-Durkheimian individual 17:933–934
Nonecstatic forms of reverie 16:170
Nonelemental learning 15:156
Nonemotional episodes, *versus* emotional episodes 7:528–530
Nonemptiness 4:870–871
Nonenforcement 19:636
Non-episodic contribution to future thinking 9:558–559
Non-equilibrium systems 17:579–580
Nonequivalent comparison group (NECG) design 12:454, 19:745
Nonequivalent group designs 16:851–855
causal inference 16:853
elaborate designs 16:853–854
benefits of disaggregation 16:854
multiple comparison groups 16:853
multiple outcome measures 16:853
multiple time points 16:853–854
variations in amounts or types of treatments 16:854
minimizing selection differences 16:854–855
observational research 16:854
prototype 16:851–853
analysis of covariance 16:852–853
change-score analysis 16:852
matching/blocking 16:852
selection modeling 16:853
threats to internal validity 16:851
Nonet, P.
on evolution of law 13:566
Noneudaemonistic virtue ethics 25:149
Non-European civilizations 4:405
Non-European socialist countries 18:700–701
Nonexculpatory defense 5:220
Non-expatriate research, cross-cultural training 5:337
Nonexperimental designs, regression adjustment in 4:841
Nonexperimental research, for aggression 1:376
Nonfamily households 8:769
Nonfat-phobic anorexia (NFP-AN) 6:826–827
Nonfiction films
versus documentary films 6:620
genres 9:163–164
Nonfiction mode 9:163–164
Nonfluent/agrammatic variant of primary progressive aphasia (nfvPPA) 13:363–364
language processes in 13:364
Nonformal education *see* Informal education
Nonfossil fuel obligation (NFFO) 18:369
Nongeometric cue use by toddlers 23:122
Nongovernmental institutions, and public bureaucracies 19:531–533
Nongovernmental organizations (NGOs) 6:293, 7:713, 7:715–716, 12:371, 16:914–915, 17:302, 22:357, 22:576, 24:295–297, 24:512, 25:432
anthropological and historical aspects 16:856–860
and BOP 16:859
brief history of 16:857
and brokers 2:872–873
business codes 8:22
campaigns for climate science 3:840–841
in China 22:793
communications infrastructures 16:858
and community development 4:382
and comparative health systems 4:401–403
concepts of 16:856–857
development, critique of 16:857–858
in Eastern Europe 22:798
and foreign direct investment 9:325–326
global governance 10:156, 10:159–160
and global south 10:186–188, 10:190
globalization, critique of 16:857–858
human rights 16:858
humanitarianism 16:858
for indigenous peoples 11:756
informal sector 12:57
knowledge infrastructures 16:858
in Latin America 24:512, 24:514–515
left infrastructures 16:858
level of operation 16:914
orientation 16:914
as partners to MNCs 16:78–79
pragmatics of 16:859–860
religion of 16:858–859
rise of civil society 16:857
for social and behavioral science research 22:619
and social business 16:859
and social science 16:856
social work 24:750–754, 24:756
warfare 16:858
working with 16:856–857
and world culture 10:225–226
Nonhealth sector sources 5:784
Nonhermeneutic measures of culture 9:358
Nonheterosexuals
British Asians
see also British Asians
nonheterosexuality 21:822
encountering 16:863
and same-sex youth 9:650–654
use of term 16:861–862
Nonhierarchical social determinants of health 22:276–277
cognitive and noncognitive abilities 22:277
ethnocultural differences 22:277
lifestyle 22:277
Nonhuman agency 1:82
Nonhuman animals
serial order memory in 21:642–643
nonhuman primates 21:642
rodents 21:642–643
virtual spatial navigation in 16:353–354
Nonhuman demography
application 8:417
challenges 8:417
methods 8:417–418
Nonhuman disturbance process, geoarchaeology 10:54
Non-human embryonic stem cells 23:440
Nonhuman entities, and environmental policy 7:780
Nonhuman primates (NHPs)
goal-oriented behavior among 10:268–274
and neonatal crying 16:475, 16:480
neurophysiology of 5:23–24
serial order memory in 21:642
species-specific communication sounds in 11:323–324

Nonhumans
  and ANT 5:526
  in social referencing 22:561
Nonidentification 11:552
Nonidentity problem, and intergenerational justice 12:397–398
Nonindicative mood 24:212
Noninfectious diseases 12:36
Noninferiority designs, controlled trials 19:388
Noninstrumental values, and democracy 6:86
Noninteractional artifacts 8:572
Non-Intercourse Act 13:538
Noninvasive prenatal testing (NIPT) 10:35–36
Noninvasive research, experimental laboratories 8:552
Noninvasive tests 9:994–995
Nonissues, power 18:766
Nonjudgmental orientation, in ethnography 8:185–186
Nonjustificationism 5:275–276
Nonkilling 17:626–627
  transformation from killing to 17:626
Nonlawyers 12:924–925, 17:482
Nonlinear dynamical systems 4:47, 25:158–159
  and emotions 7:528–531
Nonlinear effects 18:774
Nonlinear factor analysis (NOHARM) 6:438
Nonlinear logic model 14:302f
  with feedback loops 14:303f
Nonlinear models 14:854
  for growth curve analysis 10:444–445
  instrumental variables 12:207
Nonlinear (diffusion) models 23:471–472
Nonlinear pricing 20:166–167
Nonlinear regression
  generalized linear mixed models 9:847
  optimal experimental design 17:261
Nonlinear time series 24:320, 24:360
  for the mean 24:311–312
    class of bilinear time series models 24:312
    class of threshold models 24:311–312
Nonlinearities in environmental systems 12:717
Nonlinguistic 'agent-first' heuristic 16:672
Nonliteral language, and pragmatic disorders 18:791–794
Nonlocal ethnographies, and citizenship 3:641
Nonlocalized culture, in Japan 12:795
Nonlumping position, classifiers from 3:817
Nonmanual class voting 25:299–300
Nonmarital births 8:783
  in formerly socialist European countries 9:99, 9:102–104, 9:103f
Nonmarital childbearing, and Second Demographic Transition 21:347
Nonmarital fertility 8:824–825
Nonmarket allocation 14:543
  experimental economics 8:548
Nonmarket valuation 7:728–729
Nonmelanoma skin cancers (NMSC) 23:696–697
  incidence 23:696
Non-Mendelian theory of heredity 8:227
Nonmetric data, clustering 19:433
Nonmetric multidimensional scaling 16:36
Nonmetropolitan countries, rural sociology in 20:817
Non-Mexican migration from Latin America 15:464–465
Nonmodern ontology, cybernetics 5:645
Nonmonetary economy 13:109
Nonmonetary relief, remedies and 20:413
Non-monotone function 16:886
Nonmonotonic logics 14:320, 16:926, 16:929, 19:64
  risky nature of 16:927
Nonmoral deserts 6:230
Nonnarrative films (nonstory films) 9:163
Nonnegativity 14:284–285
Nonofficial statistical sources 5:706
Nonoptimal stopping 5:908–909
Nonorganizational surveillance 23:737
Nonpainful phantom limb 18:2–5
Nonparametric additive models 21:601–602, 21:602f
Nonparametric Bayesian methods 23:197
Nonparametric bootstrapping 23:415–416
Nonparametric estimation and testing 16:117
  survival analysis 23:771–772
Nonparametric item response models 8:701, 19:419, 21:599–600
  applications 8:701
Nonparametric Parzen-window type learning algorithms 23:415
Nonparametric regression 8:315, 16:867–877
  for binary and count data 16:874–875
  generalized additive model 16:876–877
  local estimation 16:868
  LOESS 16:872, 16:876–877
  with multiple predictor variables 16:874
  penalized least squares 16:873–874, 16:873f
  regression trees and related methods 16:874–876
  response 16:874, 16:876
  with single predictor variable 16:867–873
    kernel smoothing 16:868–871, 16:869f–871f
    local polynomial regression 16:870–872
    with splines 16:872–873
  software implementation 16:876
Nonparametric statistics 16:885–890, 19:431, 20:157
  advanced computational methods 16:878–884
    Bayesian methods and Dirichlet measures 16:878–880
    bootstrapping and Gibbs sampling 16:880–883
    testing for concordance 16:880–881
    testing independence 16:881
  common tests 16:888–889
    equality of multivariate distributions 16:889
    equality of several populations 16:888
    independence of two variables 16:888–889
    location of symmetric distribution 16:888
  confidence intervals 16:889
  median test 16:886
  non-monotone function 16:886
  permutation tests 16:885–887
  rank-based methods 16:891–897
    efficiencies for location procedures 16:896
    k-sample problem 16:893–895
    location alternatives 16:891–896
    multivariate two-sample location problem 16:895–896
    two-sample location problem 16:892–893
  two-sample tests 16:885–888
    asymptotic distribution 16:886–887
    asymptotic power 16:887–888
    statistic 16:886
Nonparametric transition intensities, Markov process 8:314–315
Nonparticipation 3:627
Nonpartisanship 17:562–566
Nonpecuniary loss 24:429
Nonpersonality disorder psychopathology 17:916–917
Nonpharmacological treatment, rheumatoid arthritis 20:655–656
Non-physical aggression, of children, effects of TV content on 15:38
Nonphysician health professionals 16:907–913
  advanced practice nurses and physician assistants 16:910–912
    education and characteristics 16:911
    evolution of NP and PA 16:910–911
    roles and work settings 16:911–912
  allied health professionals 16:908
  changing context of health care 16:907
  disease patterns 16:907
  education and characteristics 16:908–909
  evolution of occupational and physical therapy 16:908
  global perspectives 16:912
  health care services, accessibility of 16:907–908
  nursing 16:909–910
    education 16:910
    historical perspectives 16:909
    vestiges of history in contemporary nursing roles 16:909–910
  organizational changes conducive to expansion of 16:908
  Patient Protection and Affordable Care Act (PPACA) 16:907–908, 16:912
  registered nurses 16:908–910
  roles and work settings 16:909
Nonplaces, liminality 14:131, 14:134
Nonpositivism 18:27
  law and morality 13:479–480, 13:482

Nonpreference responses, in crossover designs 5:351
Nonprocreative sexual acts 3:884
Nonprofessionalism, journalism 12:860
Nonprofit associations 25:252, 25:254–256, 25:260
Nonprofit institutions, in United States 18:320–325
Nonprofit management 22:170
Nonprofit organizations (NPOs) 3:683–684, 12:371, 22:168, 25:252, 25:259, 25:261, 25:264–265
  challenges 16:917
    accountability 16:917
    administrative costs 16:917
    evaluation 16:917
    impact of NGO operations 16:917
    legitimacy 16:917
    performance 16:917
    sustainability 16:917
  definition of 16:915
  diversity of 16:914–915
  functions of 16:915–917
    empowerment 16:916–917
    policy advocacy 16:915–916
    service delivery 16:915
  management of 22:170
  scale 16:914–915
  strategic planning 23:515–521
Nonprofit performing arts organizations 14:565
Nonprofit research, in criminology and criminal justice 5:129
Nonprofit sector 3:365–366
  geography of 16:914–918
  theories of 3:366
Nonproportional electoral systems 7:351
Nonproportional representation 7:350–351
Nonpropositional knowledge 13:93
Nonpsychometric evaluation, cross cultural research 7:932
Nonrapid eye movement (NREM) sleep 6:652–655, 22:63–65, 22:87
Nonrational theories 5:911–912
Nonreactive measures
  of health behaviors 10:591–592
    behavior observation techniques 10:591
    physiological measures 10:591–592
Nonreductive value pluralism 25:21, 25:25
Nonreflexive modernization 15:698
Nonreimbursable administrative time, in managed care 14:463
Nonreligious organizations, distinction between religious organizations and 20:406
Nonreligious people 20:296
Nonrenewable resources, peak oil 15:554, 15:556
Nonrepresentational theory, place 18:150
Nonresidential fathers 8:841–842
Nonresponse (data analysis) 20:878
Nonresponse error 16:921–923
  audience measurement 2:213–214
  dealing with 16:921–922
  definition 16:921
  imputation methods 16:922–923
  weight adjustments 16:922
Nonresponse weighting 15:603
Non-role relationship surveillance 23:737
Nonsampled entities, extrapolating to 9:843–844
Nonsampling errors 16:919–923, 20:881
  frame error 16:919–920
    dealing with 16:919–920
    definition 16:919
  nonresponse error 16:921–923
    dealing with 16:921–922
    definition 16:921
    imputation methods 16:922–923
    weight adjustments 16:922
  response error 16:920–921
    avoiding 16:920
    definition 16:920
    randomized response techniques for sensitive questions 16:920–921
  sources 16:923
Nonsensory scaling constant 19:474
Nonsequential design 21:634–637
Nonsexual experience seeking scale (NSES) 21:609
Nonshared environment, family and schooling 8:756
Nonshared Environment in Adolescent Development project 2:474–475
Nonshared environmental variance 12:298
Nonshared family 2:473
Nonsocial peer play, as maladaptive, during preschool age 18:223
Nonspatial feature-based selection, in visual selective attention 2:166
Nonstable population momentum 18:574
Nonstable populations 18:581–583
Nonstandard limiting distributions, time series 24:362–368
Nonstandard reasoning 16:924–929
  descriptive perspective 16:929
  directions and purposes in reasoning 16:926
  emergence of 16:924
  languages and 16:928
  logical semantics of natural language 16:925
  mechanisms and structures 16:928
  natural triangle of disciplines in 16:929
  social nature of 16:928
  structural rules for 16:928
  systematic theory of 16:928–929
  temporal strategic aspects of reasoning 16:928
  with uncertainty and probability 16:926–927
  valid consequence, notion of 16:925–926
  see also Reasoning; alternative views of
Nonstandard welfarism 25:509
Nonstate private legal systems 13:807–808
Nonstate war 25:420
Nonstate-centered concept, governance as 10:292
Nonstationarity, cointegration 24:322–323, 24:325–326, 24:328–329, 24:322f
Nonstationary time series 5:908, 24:316–317, 24:362–363, 24:367
Nonsteroidal anti-inflammatory drugs (NSAIDs) 8:328, 8:330
Nonstructural models 9:320–321
Nonsubordinate word order 1:188
Non-substance-related disorder (NSRD) 1:97, 1:99–101
Nonsymbolic magnitudes 17:46, 17:48–53
Nontechnological innovation 12:147
Nontraditional psychological treatments 19:379–381
Nontraditionalism 17:232
Nontransferable utility games 4:867, 4:877–878
Non-tuism 16:583–585
Nonulcer dyspepsia 9:606
Non-university-based applied geography 1:845
Nonutilitarian theories 4:682–683
Nonverbal behavior
  emotion assessment from 7:521–523
  gender differences in 9:754–755
Nonverbal communication 4:287
  gender differences in 9:754–755
  kinesics 4:287
  proxemics 4:287
Nonverbal decoding, gender differences in 9:755
Nonverbal Memory Index 16:725
Nonverbal memory procedures, during childhood 3:465
Nonverbal social cognition, goal-oriented behavior 10:268–274
Nonverbal spatial tasks, influence of spatial terms on 13:295–296
Nonverbal tests and developments in infancy 21:896
Nonviolence
  anthropological aspects 17:624–630
    anthropology as peace founder 17:624
    apologists for war 17:627–628
    colonialism 17:626
    conflict 17:625–626, 17:628–629
    conflict resolution 17:625, 17:628–629
    dispute 17:625
    documentation 17:624–625
    feud 17:625–626
    human nature 17:624–629
    mediation 17:626
    natural history 17:625–626
    nonkilling 17:626–627
    obstacles 17:628–629
    peace studies 17:624, 17:626–629
    positive peace 17:629
    reexamining tribal warfare 17:626
    Semai 17:625
    structural violence 17:629
    transformation from killing to nonkilling 17:626
  and political protests 18:425–426
  and revolutions 20:624–625
  and social justice 22:399
Nonviolent and Violent Campaigns and Outcomes Data Set 12:222

Nonviolent pornography 18:613
Nonviolent protest 18:425
Nonviolent resistance, effectiveness of 12:226
Nonviolent revolutions 20:644
Nonviolent social crisis, community resilience and 4:358
Nonvoting, implications of 25:310
Nonwarring societies 4:615, 4:617–618
Non-Western accounts of Central Asia 3:317
Non-Western civilization's period of modernization 15:711
Non-Western cultures 24:576–579, 24:581
Non-western group processes, organizational change through 10:413
Non-Western healing practices, for mental health disorders 5:572–573
Non-Western societies
causal research in 3:248–249
modernization of 15:709–710
Non-Western Utopianism 14:253
Non-Western views of social and behavioral sciences 25:541–550
Nonword repetition task 25:229
Non-work 25:676–677
Nootka 20:899–900
Nora, Pierre
on memory 1:882–883
works
Les Lieux des Mémoire 1:882
Noradrenalin 16:749, 16:751
and anxiety 1:804, 1:807, 1:809
Noradrenergic system, and ADHD 2:159–160
Norbert Elias Foundation 7:380
NORC *see* National Opinion Research Center (NORC)
Norcliffe, Glen, works, *The Ride to Modernity* 11:18–19
Nordau, Max, works, *Degeneration* 8:239
Nordenstreng, K., on television 25:84
Nordic and East Asian countries 22:328
Nordic Association of Schools of Social Work 12:475
Nordic countries 8:769
clinical psychology, development 3:874
mortality rate decline 15:869
tax system 23:341
Nordic criminology
see also Scandinavia
Nordic welfare
poverty policy 18:750
Nordics 19:801
Norepinephrine 21:611
and wake 22:89
Noriega, Guillermo Nuñez 13:861
Noritake, Koutaro 24:584
Norm of universalism 21:262–263
Norm referencing 12:249, 12:251
Normal accidents theory (NAT) 17:371
Normal aging 1:437–438
versus successful aging 23:668–669
Normal approximations 6:571–572
central limit theorem 6:571–572
Normal distribution 6:576–577, 23:424
Pearson's 17:661–662
see also Gaussian distribution
Normal exception 11:112
Normal mixture model 15:626–627
Normal ogive model for dichotomous data 19:436
Normal personality traits measures 17:754–755
Normal process, prejudice as 18:834
Normal reference rule 19:30–31
Normal science, Kuhn on 13:133–138
Normal sleep, regulation of 22:80–82
Normality 16:358, 17:1–4
critique of disease model in psychiatry 17:2–4
Boorse's critique 17:3
Sedgwick's critique 17:3–4
Szasz's critique 17:2–4
Wakefield and teleology 17:4
definitions 17:1–2
disease as departure from statistical norm 17:1
disease as lesion 17:1–2
disease as loss of homeostatic equilibrium 17:2
disease as social deviance 17:2
as idealized norm 17:1
disorder 17:2–4
dysfunction 17:2, 17:4
mental disorder 17:3–4
violations, t-test 8:9–10
Normality test
EDF tests 23:398
Neyman smooth tests 23:400
Normalization 9:733–734, 16:788–789, 21:529, 23:637
learning disability 13:684–685
Normalized maximum likelihood (NML) 15:653, 15:657
Normalizing constant, Bayes' theorem 2:379
Normal–normal equal-variance model (NEV) 21:947
Normal–normal unequal-variance (NUEV) model 21:948
Normal-range personality models 17:914–915
Norman, Ruth 16:807
Normative accounts of science 5:504
Normative assessments, regulatory agencies 20:183–184
Normative challenges, for lesbians 13:893
Normative conflict 4:622
Normative considerations, ethnomethodology 21:194–195
Normative construal of reflective equilibrium 20:78
Normative content of pluralism 18:236
Normative control 17:341–342, 17:344–345
Normative democratic theory
democracy as public justification 6:86–87
democratic citizenship 6:87–88
equality 6:87
ethical foundations 6:85
instrumental values 6:85–86
instrumentalism 6:86
justification of democracy 6:85
legislative institution 6:88–89
liberty 6:86
noninstrumental values 6:86
Normative development and parenting 2:793
Normative discourses 12:940
Normative ethical relativism 6:592
Normative ethics 2:634, 8:73–75
Normative explanations, of ethnocentrism and prejudice 8:172–173
Normative fit 22:187–188
self-categorization theory 21:456–457
Normative framework of public enforcement 13:574
Normative hetero-femininity 8:895–897
Normative holism, engineering 7:643–647
Normative identity movements 11:545
Normative industrial ecology 11:848
Normative isomorphism 13:545–546, 17:391, 21:73, 22:981
Normative methods 2:369–370
of political philosophy 18:28
Normative moral theory 2:634
elements 8:73–74
living traditions in 8:74
Normative order, articulation, and dispute transformation 6:565–566
Normative practice, writing as 25:782
Normative psychological contracts 19:369–370
Normative reference 20:72–73
Normative resources 21:135
Normative sexuality, in childhood 21:840–841
Normative solidarity 12:403, 12:406
Normative structures, and ethics 1:734, 1:736, 1:738
Normative theories
of decision making 5:917–918
and multiculturalism 16:31–32
Normative thinking 15:304–305
Normativism 14:841
Normativity
of content 14:847–848
assertion (NCA) 14:844, 14:847–848
belief (NCB) 14:844, 14:847–848
ethical 1:735–737
of legal norms 18:24–25
and practical reasoning 18:777–778
semantic *see* Semantic normativity
Normconforming behavior 13:120
Normed fit index (NFI) 23:618
Normlessness 1:544
and social control 22:236–237
Norm-referenced grading 3:823
Norms 4:247, 4:844–849, 7:483, 11:121–124, 11:259, 11:264–265, 17:5–10, 20:333–334, 23:6–8
basic facts 17:5
as behavioral regularities 17:6–7
conditionality 17:5, 17:7
constitutional 4:710–711, 4:714

in culturally diverse societies 11:679–680
defined 4:847–848
definition 17:5–7
effects 17:9–10
efficiency 17:8
externalities 17:5, 17:8
and formal rules 22:980–981
global governance 10:157
group 10:423
group, deindividuation 6:40–41
and groups 10:436–437
homo oeconomicus 17:9
homo sociologicus 17:9
human right 2:108
informal 4:848–849
institutions 12:190–194
legal 13:727
measurement 17:7
and moral development 15:782–786
and networks 22:981–982
organizational groups 10:408
origin of 22:979–980
origins 17:7–9
oughtness definition 17:5–6
patterns of approval and disapproval 4:848–849
political repression 18:450–451
and prejudice/discrimination 18:828, 18:830–832
rational choice explanation 19:899–900
regarding happiness 23:644–645
from regularities 4:848–849
in schools 21:56
in science 17:11–14
social 13:728
social exchange 8:482, 8:485
and social psychology 22:525–526, 22:528, 22:530
socialization 17:8–9
value-based 25:36, 25:38–39
values 17:6–9
and values 25:49–50
in virtual communities 4:335
Norms in science
Merton's list 17:11–12
scientific communication 17:11–13
scientific ethos and the changing nature of science 17:13
theoretical status and significance 17:12–13
Norms of culture 4:5
Norms of science 21:216, 21:262
Norm-sender 21:168
Norse, in Greenland 1:925
North, Douglass 6:276, 8:433
works, *The Economic Growth of the United States* 7:56
North Africa 9:73–79
see also Middle East and North Africa (MENA)
North African studies
gender 16:367–370
politics 16:371–375
religion 16:376–380
society and history 16:381–386
contemporary approaches 16:382–384
North America 11:861
archaeology of 17:15
Archaic period 17:16
colonialism 17:16–18
in context 17:20
diaspora 17:15, 17:17–18
ecology 17:15–17, 17:20
encounters and movement 17:17–19
history 17:15, 17:17, 17:19–20
human–environment relations 17:19–20
hunter–gatherers 17:15–17, 17:19–20
landscape 17:15–20
materializing social complexities through ritual 17:19
migration 17:15, 17:18–19
monumentality 17:15–16, 17:19
Paleoindian period 17:16
study 17:15–17
clinical psychology, history 3:877–883
confederation 4:573–574
educational systems in 7:297–300
historiography and historical thought in 11:58
immigration in 11:669
and Native Americans *see* Native Americans
native peoples, totemism 24:453
population composition by race and ethnicity in 18:553–558
in Canada 18:554–555, 18:554t–555t
census 18:553–557
ethnic composition of Asians and Latino/Hispanics 18:557t
historical perspective of ethnic composition 18:553–554
immigration 18:553–556
in Mexico 18:556–557, 18:557t
population momentum 18:553, 18:555–556
in United States 18:555–556, 18:556t
small states 22:110–111
social science infrastructure 22:661–666
research 22:596–603
resourcing and funding 22:596–603
social sciences funding in 22:655–656
sociocultural aspects 17:22–26
capitalist economies 17:24
communities of difference 17:22–23
contradictory democracies 17:24–25
immigration 17:24
racially stratified societies 17:23
settler colonial societies 17:25–26
theorizing human difference 17:22
upward social mobility in 12:279
urban policy in 24:908–914
North American anthropology, culture in 5:583
North American Arctic 1:921
archaeology of 1:921
ecological definitions of 1:921
North American Association for the History of the Language Sciences (NAAHoLS) 14:218
North American Black Feminism 8:908
North American Free Trade Agreement (NAFTA) 11:758, 12:470, 12:569, 12:574, 14:710, 24:485
and clinical psychologists 3:877
North American History of Science Society 21:241–242
North American Industry Classification System (NAICS) 5:798, 6:586–587
North American Society of Sport History (NASSH) 23:307–308
North Atlantic Population Project (NAPP) 3:287
North Atlantic Treaty Organization (NATO) 1:556–557, 3:678, 4:124, 10:156, 17:633, 24:170
North China Wireless Joint Equipment Factory 5:598
North Korea 2:72, 2:74–75
politics in *see* Korean politics
North Sea migrating system 15:449
North-American million-cities, skylines of 3:598f
Northcote–Trevelyan reforms in Britain 5:17–18
Northeast Bear Weekend 13:891
Northeast Community Action Group (NECAG) 7:756–757
Northeast Eurasia 1:921–922
Northern Andes 23:47
Northern Cooks 17:158–159
Northern Europe, carnival in 3:146
Northern Territory Emergency Response (Australia) 2:235
Northern Thai calendar 24:407
Northfield Military Hospital (UK) 19:505
North–south divide 6:290, 6:293–294
in higher education (HE) market 10:859–860
Northwest Coast of North America 1:956
Northwestern Europe 8:275
Northwestern University 13:516
Norwegian anthropology 1:754
Norwegian Program of Electoral Research 20:744
Norwegian Psychological Association, establishment 3:874
Norwegian Social Science Data Archive 5:793
Norwegian Social Science Data Services (NSD) 1:916, 5:832
NOS *see* Nederlandse Omroep Stichting (NOS); Nitric oxide synthetase (NOS)
Nosofsky, R.M., Generalized Context Model (GCM) 4:538–540, 4:655
Nosological status, of hysteria 11:507
Nosology 17:27–33
American Psychiatric Association's Classification
Fifth Edition (DSM-V) 17:31
Fourth Edition (DSM-IV) 17:31–32
Third Edition (DSM-III) 17:30–31
dimensional and hybrid alternative 17:32
future developments 17:33
history 17:27–28

Nosology (*continued*)
International Classification
eighth and ninth revisions (ICD-8 and ICD-9) 17:30
tenth revision (ICD-10) 17:31–32
origins of WHO's international classification 17:28–29
reasons for dissatisfaction with psychiatric classification 17:29
response to problems 17:29–30
uncertain validity 17:32
Nostalgia
anthropology of 17:34
in Asia 17:35–36
claimed authenticity, cultural myths, and negotiations of culture 17:36
country, community, childlike, calendric 17:35–36
in Europe 17:35
globalization 17:34–35, 17:37–38
identity 17:34–38
invented traditions 17:35
in Japan 17:35–36
memory 17:34–35
and nationalism 17:36–37
newer trends 17:37
past, reinvention of 17:35
salvage anthropology and critical nostalgia 17:37
tradition 17:35–36
transnationalism 17:37–38
uncertainty, change, and collective imagination 17:34–35
West 17:35–36
and cultural heritage 5:416–417
and primitive society 18:911, 18:914
Notary 17:482
Notation
best–worst scaling 2:550
dance 5:668–669
of tones 23:716
Notes
field observational research 9:138
publishing 18:917
Notestein, Frank Wallace 14:373, 17:39–42
achievements of 17:39–40
contribution to demography 17:40–41
on fertility control 17:40
Hodgson's criticism of 17:41
interest in Asia 17:40–41
life and career 17:39–40
and OPR's demographic discoveries and theories 17:39
revisionist movement of 17:41
on rural–urban fertility differentials 17:40
works
Medical History of Contraception 17:40
World Population 17:40
Not-for-profit hospitals 11:211
Not-for-profit organizations 14:567, 17:364
Notharctus 18:908
Nothing works
in applied research (criminology and criminal justice) 5:129
response to 19:44–45
Notice externalities 19:168
Noticiario de Historia Agraria 17:671–672
Notification systems 5:784
Not-in-myback-yard (NIMBY), taxation 23:339
Not-in-my-backyard syndrome
see NIMBYism (not-in-my-backyard syndrome)
Notions of indigenous 11:784
Notting Hill Carnival 3:146
Nottingham Health Profile (NHP) 10:658–659
Nouiba 21:752
Noujaim, Jehane, works, *Control Room* 6:621
Noun 11:825
Noun bias, in vocabulary acquisition 25:228
Noun classes 3:812
Noun classifiers 3:812–813, 3:813t
Noun incorporation 14:220–221, 25:2–3
Noun phrase (NP) 10:316, 10:324, 13:277–278, 14:178, 25:644
Noun/verb distinction at syntactic level 17:558
Nouvelle histoire 11:104
Novelistic Westerns 9:160
Novelties 7:483
and flashbulb memory 9:256
functional explanation 9:521–522
in organizations 2:770–771
in scientific disciplines 21:288
Novelty seeking (NS) 17:794, 21:607, 21:610–611
Novices 24:56–57
apprenticeship 1:868
Novick, Peter 23:327
Novus Ordo Seclorum 20:638
NO_x *see* Nitrogen oxides (NO_x)
Nozick, Robert 4:680, 24:745
on libertarianism 9:400
on minimal state libertarianism 14:41–42
works
Anarchy, State and Utopia 4:318, 14:41, 19:917, 25:348
NP *see* Nondeterministic polynomial (NP); Noun phrase (NP)
NPC *see* New political culture (NPC)
NPD *see* Nationaldemokratische Partei Deutschlands (NPD)
n-person games 9:579–580
characteristic function 9:579
example 9:579
core 9:579–580
Shapley value 9:580
theorem 9:580
solution concepts 9:580
N-persons Prisoner's Dilemma 18:934
NPI *see* New political institutionalism (NPI)
NPM *see* New public management (NPM)
NPOs *see* Nonprofit organizations (NPOs)
NPP *see* Net primary production (NPP)
NPR *see* National Public Radio (NPR)
NPs *see* Nurse practitioners (NPs)
NPT *see* Nuclear Non-proliferation Treaty (NPT)
NPV *see* Net present value (NPV)
NRA *see* National regulatory authorities (NRA); National Resistance Army (NRA)
NRC *see* National Research Council (NRC)
NRE *see* National Registry of Exonerations (NRE)
NREM *see* Nonrapid eye movement (NREM)
NRM *see* New religious movements (NRM)
NRMP *see* National Resident Matching Program (NRMP)
NRR *see* Net reproduction rate (NRR)
NS *see* Neosexism Scale (NS); Novelty seeking (NS)
NSABB *see* National Science Advisory Board for Biosecurity (NSABB)
NSAIDs *see* Nonsteroidal anti-inflammatory drugs (NSAIDs)
NSD *see* Norwegian Social Science Data Services (NSD)
NSDAP *see* Nationalsozialistische Deutsche Arbeiterpartei (NSDAP)
NSDUH *see* National Survey on Drug Use and Health (NSDUH)
NSES *see* Nonsexual experience seeking scale (NSES)
NSF *see* National Science Foundation (NSF)
NSF-MH *see* National Service Framework for Mental Health (NSF-MH)
NSHAP *see* National Social Life, Health, and Aging Project (NSHAP)
NSI *see* National statistical institutes (NSI); New sociological institutionalism (NSI)
NSM *see* Natural semantic metalanguage (NSM) approach; New social movements (NSM)
NSO *see* National Statistics Office (NSO)
NSRD *see* Non-substance-related disorder (NSRD)
NSSE *see* National Society for the Study of Education (NSSE)
NSW AECG *see* New South Wales Aboriginal Education Consultative Group Incorporated (NSW AECG)
NT *see* Neuromaturational theory (NT)
NTS *see* Nucleus of tractus solitarius (NTS)
NTs *see* Neuropil threads (NTs)
NTVS *see* National TV Violence Study (NTVS) review
Nubian identity 15:357–358
Nuclear accident 22:570
Nuclear club, and arms control 1:976
Nuclear DNA (nDNA), profiling of 9:946t
Nuclear electricity incumbents, impact of RES-E deployment on 18:372
Nuclear energy 7:626
Nuclear family 2:471, 8:809, 8:811, 8:816–817, 11:229
Nuclear micronesian 17:160
Nuclear Non-proliferation Treaty (NPT) 1:976–977
Nuclear power
and green parties 10:379–381, 10:383–384
subsidies for 18:367–368
Nuclear power plants

living with 7:806
when oil invades our water 7:806–807
Nuclear science, research censorship 3:282
Nuclear weapons 1:559–560, 11:298
Nuclear Weapons Freeze Campaign 17:645
Nuclearization, in Cold War 4:131
Nucleosomes 7:839, 7:841f
Nucleotides 9:826
Nucleus accumbens (NAcc) 14:126, 16:640, 16:640f
Nucleus of tractus solitarius (NTS) 11:421–422
interoception 12:618
Nudge theory 23:829, 23:831
Nuer 13:71
Nuestro Mundo 13:859
NUEV *see* Normal–normal unequal-variance (NUEV) model
Nueva Corónica y Buen Gobierno 11:58
Nugget effect 23:192
Nuisance, and antisocial behavior 1:790–793
parameters, statistical inference in presence of 23:421–422
Nuisance variables 8:571
Null hypothesis 11:491–492, 11:494–497, 19:704, 20:155, 20:927, 23:617–618
significance tests 21:958–959
Null models for networks 22:469–470
Null player axiom 4:872–873, 4:875
Null player out axiom 4:875
Null results, problem of 4:397–398
Null subject language 10:324, 10:326
Nullification 12:910–911, 12:934
Nullifying player axiom 4:873
Number, grammatical 17:43–45
Number development, Williams syndrome 25:582
Number meaning 14:775
Number names 14:775–776, 14:776f
Number of categories, in rating scale 23:768
Number processing 17:506, 17:509
Number range extension (zero–four) 14:779
Number sense 17:46–56
beyond 17:53
biological origins 17:49–51
across species 17:49
in early development 17:49–50
magnitude 17:50–51
subitizing 17:51
Number symbols 14:775
Number system
cultures without 17:57
Euclidean conceptualization of 17:59
evolution of 17:57–60
Number terminology 14:775
Number words 14:775
Numbers 4:45–46, 14:193, 14:775–776, 17:46–56
as an obligatory category 17:43
analysis of 23:423
approximate magnitude 17:46, 17:48–51
approximate number system 17:50–51
and computers 4:517–518
and educational neuroscience 16:732
internal and external representation of 14:775–776
journal and articles 12:865
access of researchers 12:866–867
market size and economics 12:866
in scientific output 12:866
nominals involved in 17:43
processing, signatures of 17:46–49
interpretations 17:47–49
numerical distance effect 17:46
numerical ratio effect 17:46–47
real numbers 14:868
semantics of 17:43
symbol grounding 17:52–53
symbolic numbers 17:51–53
tuning curves 17:48–49, 17:52, 17:48f
values 17:43–45
dual 17:44
largest system of 17:44–45
paucals 17:44
trial 17:44
Numeracy
achievement, and socioeconomic status 22:927
parental engagement in children's education 17:500
Numeral classifiers 3:812–813, 3:813t
Numerals 11:826
evolution of 25:762, 25:765
and language 13:296–297
Numeration 9:876
Numerical age 14:77
Numerical cognition 4:69–70
in infants 4:62
and mental imagery 15:223
Numerical constraints on time 13:103
Numerical distance effect 17:46
Numerical identity 17:742–744, 17:746, 17:748
Numerical methods 11:23
age–frequency relationship 11:26
early 11:23–24
future works 11:26
Ringe model 11:26
Sankoff model 11:25
Starostin model 11:25–26
Swadesh model 11:24–25
Numerical ratio effect 17:46–47
Numerical simulation method, family and household demography 8:728
Numerical symbols, systems of, in early civilizations 17:58
Numerical taxonomy 8:370
Numerical values, conjoint measurement 14:862
Numerical weather prediction 7:793
Nunavut 24:509
Nuns, Buddhist 25:629
NUPP *see* Public Policy Research Center (NUPP)
Nuptiality 21:792–793
Nuptiality patterns 11:8
Nuremberg Code of 1947 8:31, 8:48, 8:53, 13:558–559, 20:514–516, 21:316
Nuremberg Laws 16:356
Nuremberg trials 3:675, 3:677, 5:875, 13:820, 18:498, 18:500, 25:382–383
Nuremberg tribunal 20:514
Nurse Home-Visiting programs 18:887
Nurse practitioners (NPs) 16:910–911
education of 16:911
roles and work settings 16:911–912
Nursing 10:402, 16:909–910
education 16:910
historical perspectives 16:909
vestiges of history in contemporary nursing roles 16:909–910
Nursing homes
consultation–liaison psychiatric services in 14:3–4
psychiatric morbidity in 14:2
Nurture, and learning 13:699–700
Nurturing
and child neglect 3:411–415
of children, and religion 20:320
Nushu 9:785
NUT *see* New unified theory (NUT) of sociobehavioral forces
Nutrients, and food preferences 9:296–297
Nutrigenomics 2:623
Nutrition 9:312–313
and bioarchaeology 2:600–601
illness and infection 2:601
stress and deprivation 2:600–601
and cash transfers 3:215–216
and food deserts 9:285–287
gene-culture coevolution 9:815
and mortality decline 15:869
Nutritional effects, of breast feeding 2:848–849
Nutritional therapy 10:562
Nuwas, Abu 21:750
Nuwaubians 16:806
NVAWMS *see* National Violence Against Women and Men Surveys (NVAWMS)
NVivo, software package 22:545
NVPC *see* National Vacant Properties Campaign (NVPC)
NWICO *see* New World Information and Communication Order (NWICO)
NWSA *see* National Women's Studies Association (NWSA)
Nye, Joseph, works, *The Rise of China's Soft Power* 5:597
NYLS *see* New York Longitudinal Study (NYLS)
NYPD *see* New York Police Department (NYPD)
NYS *see* National Youth Survey (NYS)
Nzimbu shells 3:313

O

OA *see* Open access (OA)
OAC *see* Open Anthropology Cooperative (OAC)

Oak Ridge National Laboratory 23:142–143, 23:146
Oakeshott, M., works, *Rationalism in Politics and Other Essays* 4:685
Oakland Growth Study 17:462–463, 17:466
Oakley, Ann 9:792
OAS *see* Organization of American States (OAS)
Oasis civilizations 12:725
OASIS trial 20:189
OASys *see* Offender Assessment System (OASys)
OAU *see* Organization of African Unity (OAU)
OB *see* Organizational behavior (OB)
OBA *see* Online behavioral advertising (OBA)
ObaapaVitA trial 3:899
Obama, Barack 1:941, 7:895, 18:394
  presidential campaign of 14:716–717
OBE *see* Outcomes-based education (OBE)
Obedezco pero no cumplo 13:858–859
Obedience 4:154–155, 17:61–66
  agentic state 17:63–64
  and authority 2:272–273
  to authority 17:654–655
  banality of evil 17:63–64
  baseline condition 17:63
  from bureaucratic sleepwalker to engaged follower 17:64–65
  conformity studies 17:61–63
  and economic development 5:531
  explaining 17:63
  Nazi Germany 17:63–65
  science and ethics 17:65–66
  social psychological perspectives 17:67–69
    to authority, pervasiveness of 17:69
    distinctive feature 17:67
    Milgram's 17:67–68
    other approaches 17:68
    in workplace 17:68–69
  studies 17:62–63
  suggestion research 17:61
Oberschall, A.R., on interviewing 12:671
Obesity 4:401, 8:6, 15:408, 24:34
  and adolescent health 1:125
  child health psychology 10:681–682, 10:684
  childhood 15:39–40
  as complex problem 17:85
  critical perspectives 17:85
  cultural and biological factors 17:82–87
  and demography 17:83–84
  diet 17:82–83, 17:85
  disease 17:82
  ecological models of 17:82–83
  and exercise 17:82–83
  fat bodies, cultural meanings of 17:84
  fat studies 17:85
  genetic and epigenetic explanations of 17:83
  health 17:82–85
  history 17:77–78
  as identity 17:84
  and labor market performance 17:70–74
    absenteeism 17:70
    BMI 17:71–74
    customer discrimination 17:74
    econometrics 17:72, 17:74
    empirical methods and findings 17:72–74
    employer discrimination 17:74
    employment 17:70–74
    human capital 17:70–72
    misnourishment 17:70–71
    obesogenic environment 17:70
    relationship 17:70–72
    statistical discrimination 17:71, 17:74
  medicalization 17:85
  and mortality 17:75–81
    body mass index 17:75–78
    causal identification 17:75–76
    health behaviors 17:78
    in international context 17:78
    life course 17:79
    life expectancy 17:75, 17:78
    obesity paradoxes 17:77
    reverse causation 17:76
    role of obesity histories 17:77–78
    secular and cohort changes in 17:78–79
    statistical methods 17:77, 17:79
    variations in relative risks by age and smoking behavior 17:76–77
  psychotherapy for 19:376–377
    pediatric overweight 19:377
  and sleep 22:60
  and socioeconomics 17:83–84
  weight regulation 25:468
  weight stigma and discrimination 17:84–85
Obesity genes 25:470
Obesity paradox 17:76
Obesogenic environment 17:70
Object case 2:548
Object data 23:94–95
Object experiencers (OE) 16:670–672
Object knowledge, and semantic memory 21:573–577
Object nominalization 23:892
Object perception 2:201
Object permanence, in infant development 18:854
Object play, parental scaffolding of, during toddlerhood 18:222
Object recognition 8:663–666, 25:153–154
  Biederman's model 25:153–154
  Edelman's model 25:154
  and spatial localization, in visual perception 4:107
  two-stage view 4:104
Object relations 2:757–758
Object relations theories 11:537–538, 13:67–68
  conceptions of mind 22:761–762
  formulation 19:340
  psychodynamic treatment 22:764
  psychopathology in 22:763
  of socialization and education 22:862–863
Object representation 21:269
Object selectivity, building blocks of 16:629–630
Object substitution 18:873
Object trees 23:393
Object use, similarities, during preschool age 18:223
Object view of numbers 14:776–778, 14:777f–778f
Object-centric spatial terms (language) 13:295
Object-choice tasks 10:270–272
Objectification 6:32, 6:35–36
  homosexuality
    definition 16:862
  of identity 11:548
  and social representation 4:198
  and social representations 20:447
  totemism 24:453, 24:455–457
Objectified cultural capital 5:375, 5:379
Objection 1:823
Objectivation (*Entäusserung*) 20:203
Objective analysis 23:191
Objective assessments 3:820
  of context 4:787–788
Objective Bayesian inference 2:374
Objective behaviorism 13:703
Objective fault 5:221–222
Objective function, in stochastic optimization 23:473, 23:476, 23:478
Objective identity 11:523–525
Objective impact-oriented point of view 7:699
Objective interpretations of probability 19:25–27
  best system interpretation 19:27
  finite relative frequency interpretation 19:25–26
  limiting relative frequency interpretation 19:26
  need for 19:25
  propensity interpretation 19:26–27
  theoretical role interpretations 19:27
Objective justice 22:847
Objective knowledge 17:99, 22:1
Objective Measure of Ego Identity Status 21:434–435
Objective measures
  of personality 16:725
Objective physiological techniques 17:609
Objective possibility, principle of 12:647
Objective power, and perceived control 1:546
Objective prior 14:123
Objective probability 19:40–41
  chance and propensity interpretation 19:41
  limiting and finite frequency interpretations 19:40–41
Objective research 17:92
  data fabrication 17:93–94
  definition 17:92
  ethical aspects 17:92–95
  and ethical principles 17:93–95
    conflicts of interest 17:94
    error, bias, and disagreement 17:93–94
    honesty 17:93

misrepresentation of data 17:94
openness 17:94–95
ethics 17:92
evolving debate about 17:92–93
falsification 17:93–94
foundations of research ethics 17:93
misconduct 17:92, 17:94
moral limits 17:95
Objective spirit 6:427
Objective Structured Clinical Examination (OSCE) 7:190–191
Objective Structured Long Examination Record (OSLER) 7:190
Objective structures 10:487
Objective tests 3:823
Objective theories 17:92
of contracting 4:810
Objectivism 13:439, 14:42, 20:227–228, 22:986
authenticity 2:251
class consciousness 3:739–740, 3:743
and critical reflexivity 5:281–283
Objectivity 10:774, 12:647, 16:816, 19:676, 21:168, 21:170, 22:1–2, 23:325–326, 23:328
Albert on 1:496–497
challenges to 17:88
decline in journalism 12:855
epistemological 17:96, 17:98–100
critiques 17:98–100
explanation 17:90
feminist epistemology 17:100–101
historical epistemology 17:100–101
history of 17:88–91
history of science 17:99
and interpretation 17:89
limits 17:90–91
media ethics 15:42–44
metaphysical 17:96–98
alternative conceptions 17:97–98
independence 17:96–97
phenomenology 17:97–99
philosophical aspects 17:96–102
Pyrrhonist views 17:88
of qualitative research 19:687
and quantum mechanics 17:89–90
relativism 20:234, 20:236–237
and relativity 17:89–90
of research 20:523–524, 20:526
of science 21:212, 21:214, 21:216–217
scientific practice 25:221–225
historicizing and situating 25:222–223
social science 22:689
in social science 14:498
standards 17:89
and subjectivity compared 17:88
use of term 17:88
Object–location combinations 1:646
Object-oriented ontology 14:769–770
Object-oriented play
differences, during infancy 18:220
similarities, during infancy 18:219
Object-oriented politics 5:525–526
Object-related processing, and hippocampal amnesia 10:877–878, 10:883–884
Objects 1:352, 4:44–45
conjoint measurement 14:862
and fetishism 9:105–110
infants' imitation of acts on 11:621–622
implications for theory 11:621–622
neural representations of 16:628–633
scientific practice 25:221–225
Objects of environmental choice (OECs) 7:726–727
publicness of 7:728
Objects of inquiry, and causal mechanisms 14:877
Object-to-background contrast 20:420
Objet trouvé 9:108
OBLEQ *see* Outcomes-Based Learning Environment Questionnaire (OBLEQ)
Obligate eusociality
kin selection 13:30
Obligations
citizenship 3:655–656
of gift 10:140
in joint action 12:839–840
in joint commitment 12:842
and responsibility 20:587–588
Shklar on 21:893–894
surveys and polls
to commercial or private sponsors 23:753
to the public 23:753–754
to respondents 23:754–756
to sponsors 23:753
to U.S. federal agency sponsors 23:753
Obligatory control 4:828
Obligatory subject language 10:324
Oblique cultural transmission 5:403
Oblivion, and ignorance 11:579
OBOS *see* Our Bodies Ourselves (OBOS)
O'Brien, James Bronterre 22:838
OBS *see* Old Bering Sea (OBS) culture
Obscene humor 11:417
Obscenity 18:613
Observability, and causal mechanisms 14:877
Observable data distribution 23:414
Observable reality, and religion 20:307–308
Observances 4:235
Observation costs 5:906
Observation data
in personality assessment 17:850
in stochastic optimization 23:479
Observation method 12:814
Observation plans, in event history analysis 8:305–306
Observational approach
epidemiology 7:827–829
case-control study 7:829
cross-sectional study 7:828, 7:828b
longitudinal study 7:828–829, 7:828b
Observational documentary 9:164
Observational learning 1:374, 18:40, 21:681
animal social learning 22:403
families as educational settings 8:735
Observational methods, for time use 24:380
Observational mode of representation (film) 6:621
Observational notes during tests 7:932
Observational perspectives, psychotherapy 19:517
Observational Record of the Caregiving Environment 3:835
Observational research 3:252
control variable 4:841
experimental laboratories 8:552
nonequivalent group designs 16:854
selection bias 21:414–417
Observational studies 17:107–112, 19:742–743
adjusting for overt biases 17:108–109
matched sampling 17:108–109
model-based adjustments 17:109
appraising sensitivity to hidden bias 17:110
causal inference 3:257–259
central issues 17:107–108
definition 17:107
examples 17:107
effects on children of occupational exposures to lead 17:107
long-term psychological effects of the death of a close relative 17:107
of family planning 8:788–789
hidden bias detection 17:109–110
elaborate theories 17:109
two control groups 17:109–110
informed consent in 12:113
outcome-based sampling 17:103–106
population studies 8:559, 8:563–564
quasiexperimental designs (QEDs) 19:742–743
reducing sensitivity to hidden bias 17:110–111
choice of circumstances 17:110–111
design sensitivity, tactics to increase 17:111
goals 17:110
instrumental variables 17:111
Observation-execution situation 11:331
Observations 24:354
cognitive development 4:41–42
communication skill 4:288
limits of, in scientific laws 23:425
of procrastination 19:74–75
academic measures 19:74–75
other measures 19:75
in qualitative health research 10:694
qualitative research methods 22:545
ethnographic 22:545
participant 22:545
structured 22:545
systematic 22:545
in social systems, Luhmann 14:386
unobtrusive measures 24:768–771
Observed behavior
deducing habitus from 22:334
predicting from habitus 22:334–335
Observed group differences 5:323–324
Observed information 14:117–118
Observed outcomes 3:252
Observed score variable
classical true score model
mean 19:424

Observed score variable (*continued*)
variance 19:424
Observed variables 11:750, 20:239
controlling for 4:840–841
propensity score 4:841
regression and ANCOVA based approaches 4:840–841
regression adjustment in quasiexperimental and nonexperimental designs 4:841
regression adjustment in randomized experiments 4:840–841
validity and differential treatment effectiveness 4:841
Observer 17:631, 17:633
perspective of law and morality 13:482
Observer effects 8:572
Observer-reports, of cultural intelligence 5:434–435
Observing, field observational research 9:137
Obsessional neurosis 9:417
Obsessive versus harmonious passion 17:570–571
Organic agriculture
sustainable agriculture 23:809
Obsessive-compulsive disorder (OCD) 3:767–768, 9:243, 10:466, 15:193, 17:119–126, 23:563
antidepressants for 1:780
anxiety disorders 17:120–121, 17:124–125
body dysmorphic disorder 17:113
cingulotomy 17:269
clinical aspects 17:113–114
clinical symptomatology and course 17:113
differential diagnosis 17:114
epidemiology 17:114
cognitive models 17:122–125
compulsions 17:119–120, 17:122
cortico-striato-thalamo-cortical (CSTC) hypothesis of 17:123
deep-brain stimulation 17:269
diagnosis 17:120
disability and quality of life 17:121
early onset in 17:119–120
epidemiology 17:120–121
excoriation disorder 17:113
exposure therapies for 8:606, 8:608–609
hoarding disorder 17:113
life span aspects 17:120
mood disorders 17:120–121
obsessions 17:119–120, 17:122
outlook 17:125–126
pathogenesis 17:114–115
genetic findings 17:115
immunological findings 17:115
neuroanatomical hypothesis 17:114, 17:114f
neurochemical hypothesis 17:115
pathogenetic and etiological models of 17:122–124
brain circuitry 17:123
cognitive-affective processing 17:123
endophenotypes 17:123–124
need for integrative models 17:124
neuropsychological deficits 17:123
psychological models 17:122–123
prevalence 17:119–121, 17:123
prognosis 17:116
psychological treatments 19:383
psychosurgery 17:269
risk factors, etiological models, and pathogenetic mechanisms 17:121–122
biological risk factors 17:122
genetics 17:121–122
immunological factors 17:122
stressful life events and traumatization 17:121
selective serotonin reuptake inhibitors 17:113, 17:115–117
serotonin reuptake inhibitors 17:124–125
skin-picking disorder 17:113
symptoms 17:119–120
treatment 17:115–116, 17:124–125
brain stimulation and psychosurgery 17:125
care and guidelines 17:125
cognitive-behavioral psychotherapy 17:124–125
combination of psychological and pharmacological treatments 17:125
combined pharmacological and cognitive-behavioral treatment 17:116
drugs 17:116
natural course and placebo response 17:124
pharmacotherapy 17:125
psychological 17:115–116
in refractory cases 17:117f
trichotillomania 17:113, 17:115
Obshchaki 14:432
Obshchina 20:858
Obsidian 24:481–484
Obsidian hydration dating 3:567
OBSSR *see* Office of Behavioral and Social Sciences Research (OBSSR)
Obstetric practice 2:793
Obstetrics
psychosocial aspects 10:462–468
postpartum period 10:466
pregnancy 10:465–466
psychosomatic disorders in 19:488
Obstructive sleep apnea (OSA) 22:60
treatment in adults 22:73–74
Obtrusive measures 3:70
OC spray *see* Oleoresin capsicum (OC spray)
Ocasio, William, and markets 5:469
OCB *see* Organizational citizenship behavior (OCB)
Occam's razor 3:920, 19:727, 23:887–888
Occasion setting problem 15:156
Occident 17:428–429
Occidentalism 3:308–309
Occipital cortex, and mental imagery 15:222, 15:224
Occipital lobe 17:127–132, 25:163–164
binocular organization 17:128
columnar organization 17:128–129
development and plasticity in 17:131
dorsal cortical streams 17:129–131
interleaved pathways through V1 and V2 17:129
macaque, brain of 17:127, 17:129–130
organization, evolution of 17:129–130
orientation tuning 17:128
receptive fields 17:127, 17:129–131
recurrent and top-down influences 17:131
retinotopic projection to area V1 17:127–128
spatial generalization, functional specialization, and global processing 17:130–131
specialized extrastriate areas 17:129
stimulus selectivity of cortical neurons 17:128
ventral cortical streams 17:129–131
visual areas in 17:130f
Occult economies 14:435, 14:437–438
Occult Renaissance 20:426
Occupation 25:662
and cultural lifestyles 22:332–334
and gender economics 9:759–762
and skills 7:216–217
and social mobility 22:426–427
and sociology of work/industry 25:660
Strauss on 23:527, 23:529, 23:531
Occupation codes, comparability, multiple imputation for 16:91
Occupational classes 3:747
Occupational elites, in United States 1:618
Occupational Employment Statistics (OES) survey 5:743
Occupational hazards 23:549
Occupational health 17:133–140
adaptation to requirements at work 17:134
and burnout 2:929–931
effort–reward imbalance and demand–control 17:137
energy mobilization and anabolism at work 17:133–134
individual–environment interaction 17:136
job stress models and disease 17:136–137
job stress related outcomes 17:137–138
ongoing trends 17:138–139
physical and psychosocial adversities, combined 17:135
physiology 17:133
psychosocial interventions at work 17:138
theoretical models 17:135–136
unemployment and stress 17:134–135
Occupational health and safety 6:880
regulation of 17:141–146
see also Workplace safety and health
Occupational health and safety (OHS) regulation
criticisms 17:142–143
enforcement powers 17:145
in Europe 17:142
general duty provisions 17:143
inspection programs 17:145
process-based standards 17:143
provisions for expert assistance 17:144

regulation in developing countries 17:145–146
in Scandinavia 17:145
standard setting 17:143–144
traditional approach 17:141–142
worker involvement 17:144–145
Occupational history, of Roma people 20:750–751
Occupational mobility, Markov models for 14:577–578, 14:577t
Occupational Personality Questionnaire (OPQ32r) 17:844
Occupational Safety and Health Act 17:142, 25:699
Hazard Communication standard 20:864
Occupational Safety and Health Administration (OSHA)
enforcement activity 20:866
standards, effects of 20:868
Occupational secrets 18:253
Occupational segregation 9:760
Occupational social work 10:718
Occupational sorting, and beauty 2:385
Occupational Stress Indicator (OSI) 17:136
Occupational stress management 17:147–152
interventions 17:148–151
cognitive behavioral therapy 17:149–150
future challenges 17:151
health circles 17:151
individual-level 17:148–150
injustice 17:148
job strain 17:148
organization-level 17:150–151
resources 17:147–150
stressors 17:147–150
work design 17:149
work stress 17:147–150
working conditions 17:147–151
Occupational stress models 17:148
Effort–Reward Imbalance model 17:148
Job Characteristic model 17:148
Job Demand-Control-Support model 17:148
organizational justice 17:148
Occupational structures 25:660
Occupational therapy 16:908
Occupation–switch cycles 18:582
Occupy movements 3:104, 4:592, 17:250, 17:714–715, 18:424, 21:189–190, 4:592f
Occupy Wall Street (OWS) movement 3:582, 25:839–840, 25:844
Occurrence rates
LEXIS diagram for 6:156–157
annual record of events for both time of/age at occurrence 6:157–161
non-repeatable events 6:156–157
repeatable events 6:156
Occurrence-exposure rate 6:110–113
OCD *see* Obsessive-compulsive disorder (OCD)
OCDQ *see* Organizational Climate Description Questionnaire (OCDQ)
Ocean acidification, and Anthropocene 1:725–726
Oceania 3:139, 17:158–162, 17:446
adult mortality in 1:171, 1:179–180
austronesian expansion 17:446–447
complex societies, development of 17:448–449
early human settlement of 17:446
extinctions on 17:449
geographic regions
Melanesia 17:446
Micronesia 17:446
Polynesia 17:446
human colonization 17:447–448
human impacts 17:449
languages 17:447
lapita 17:446–447
sociocultural overviews of 17:158
Austronesian languages 17:159
decolonization 17:162
Enlightenment era 17:158, 17:161
European contact and colonialism 17:160–162
and gender 17:160
geographical nomenclature 17:158
and global warming 17:163
and health 17:163
initial human settlement 17:159–160
and Lapita pottery 17:159–160
and migrations 17:159–160, 17:162–163
and Oceanic languages 17:159–160
and Pacific Islands 17:158–162
and political organization 17:160
Polynesia and Micronesia as cultural areas 17:158–159
Polynesia and Micronesia in global age 17:162–163
society and culture 17:160
Oceanic chiefdom societies 17:448
Oceanic island 16:345
OCHA *see* Office for the Coordination of Humanitarian Affairs (OCHA)
Ochlocracy 14:743–744
O'Connor, Justin, on Chinese cultural industries 5:597–598
OCPD *see* Obsessive-compulsive personality disorder (OCPD)
Octopus pots 14:639
Ocular dominance columns in visual cortex 5:31–32
Oculomotor control 4:721, 8:628–630
abducens nerve 8:628–629
cranial nerve nuclei control tension in extraocular muscles 8:628–629
eye movements 8:628
gaze-shifting movements 8:629–630
saccades 8:629–630
smooth pursuit 8:630
vergence 8:630
gaze-stabilization movements 8:629
optokinetic response 8:629
vestibulo-ocular response 8:629
musculature of eyes 8:628
oculomotor nerve 8:628–629
trochlear nerve 8:628–629
OD *see* Organizational development (OD)
ODA *see* Official Development Assistance (ODA)
ODD *see* Oppositional defiant disorder (ODD)
Odds ratio
epidemiology 7:829–830, 7:829b
logistic regression 2:589–590, 2:593–594
Odegard, Peter, works, *Pressure Politics: The Story of the Anti-Saloon League* 18:233
ODM *see* Organizational decision making (ODM)
Odor discrimination 17:198–200
Odor identification 17:196, 17:198–200
Odum, Eugene 7:792
Odysseus 24:648
The Odyssey 16:155
OE *see* Object experiencers (OE); Organizational ecology (OE)
OECD countries *see* Organisation for Economic Co-operation and Development (OECD) countries
Oecological time 24:406
OECs *see* Objects of environmental choice (OECs)
Oedipal theory 24:454
Oedipus complexes 4:771, 19:400
Oedipus myth 9:359–360
OEO *see* Office of Economic Opportunity (OEO)
OER *see* Open educational resources (OER)
Oersted, Hans Christian 24:153
OES *see* Occupational Employment Statistics (OES) survey
OFC *see* Orbitofrontal cortex (OFC)
OFCCP *see* Office of Federal Contract Compliance programs (OFCCP)
Offender Assessment System (OASys) 19:46
Offender identification 24:120
Offender rehabilitation 20:193–202
context factors 20:197–198
continuity of care 20:198
custody *versus* community 20:197
institutional climate 20:197–198
'natural' protective factors 20:198
staff factors 20:198
desistance concepts 20:196–197
good lives concepts 20:196–197
methodological factors 20:199
demonstration studies *versus* routine practice 20:199
design quality 20:199
length of follow-up 20:199
outcome measures 20:199
researcher involvement 20:199
sample size 20:199
offender factors 20:198–199
age of offenders 20:199
biographical factors 20:199
personality factors 20:199
risk level 20:198
treatment drop out 20:199
treatment motivation 20:199
program 20:193–196

Offender rehabilitation (*continued*)
biological treatment 20:195
counseling 20:195
deterrence and punishment 20:195–196
education and vocational programs 20:194
family-oriented programs 20:194
intensive supervision 20:196
low-structured therapy 20:195
mentoring programs 20:195
milieu therapy 20:194
multisystemic therapy (MST) 20:194
restorative justice (RJ) programs 20:195
social-therapeutic prisons 20:194
sports and adventure programs 20:195
therapeutic communities (TCs) 20:194
program factors 20:197
control condition 20:197
individualization 20:197
intensity 20:197
program content 20:197
quality of program implementation 20:197
risk-need-responsivity (RNR) model 20:196
Offenders 5:633, 19:43–48
and applied criminology 1:833–834
decision making and crime patterns 20:776
development of social work with 5:208–209, 5:212
searches, routine activities and 20:774–775
sex *see* Sex offenders
and victim, relationship between homicide 11:166
weakening or incapacitation 24:119–120
women *see* Women; and criminal justice
Offending 5:190
rational choice 20:776
social organization of 5:203–204
Offense identification 24:120
Offense to others
and victimless crime 25:76
Offenses 25:821–822
misdemeanor 15:589–594
Offerings, rituals 20:724
Off-farm labor, peasants 17:678
Office automation 2:297
Office for Human Research Protections (OHRP) 4:255–256, 8:48–49
Office for the Coordination of Humanitarian Affairs (OCHA)
United Nations 24:753–754
Office of Behavioral and Social Sciences Research (OBSSR) 22:658
Office of Communications (Ofcom) 5:705
Office of Economic Opportunity (OEO) 18:734
Legal Services Program 18:734–735
Office of Federal Contract Compliance programs (OFCCP) 1:259
Office of Information and Regulatory Affairs (OIRA) 1:112–113
Office of Justice Programs 5:236
Office of Juvenile Justice and Delinquency Prevention (OJJDP) 19:838, 22:388
Office of Management and Budget (OMB), US 22:299, 24:856
Statistical and Science Policy Office 10:304
Office of National Statistics (ONS) 5:736
Office of Population Research (OPR) 17:39
Office of Price Administration and Civilian Supply (OPACS) 16:230–231
Office of Refugee Settlement, Program of Unaccompanied Alien Children 11:638
Office of Strategic Services (OSS), US 1:951, 5:99, 15:768–769
Office of the Special Adviser on Gender Issues and Advancement of Women (OSAGI) 24:754
Office of War Information (OWI) 1:951
Officers, corporate governance 4:932–935
Official crime data 5:712–713
Official Development Assistance (ODA), China 3:502–503, 3:506–507
Official economy, effects of underground economy on 12:54
Official histories
and Chinese historiography 11:48–49
and East Asian historiography 11:48, 11:52
of Japan 11:50
in Korea 11:48, 11:50
of military 15:508
and Qian 11:48
Official names 16:184
Official reality 2:752–753
Official statistics
and economic statistics 5:758
international official statistics data 5:794–795
Offline techniques 21:622
in inference research 12:38
Offloading containers 13:163
Offshore disasters, social work response, in Australia 7:430
Offshore planktonic species 7:694
Offshoring 3:43, 18:667, 25:692
and income inequality 11:741
OFS *see* Old-fashioned Sexism Scale (OFS)
OFT *see* Optimal foraging theory (OFT)
Ogbu, John 7:145
Ogburn, William F. 19:126, 22:698
Ogburn, William J. 13:774
OGC *see* Open Geospatial Consortium (OGC)
Ogre3D 25:133
Ogyu Sorai (Confucian philosopher) 14:636
Ohala's Frequency Code 23:714
Ohlin, Bertil 7:124–125
Ohm's law 5:92
OHP *see* Oregon rationing program (OHP)
OHRP *see* Office for Human Research Protections (OHRP)
OHS *see* Occupational health and safety (OHS) regulation
OIC *see* Organization of Islamic Cooperation (OIC); Organization of the Islamic Conference (OIC)
OIE *see* Old institutional economics (OIE)
Oil age 17:165–171
Oil Depletion Protocol 17:170
Oil industry 17:166–167
Oil resales and joint ventures 5:370
Oil shale 17:167
OIRA *see* Office of Information and Regulatory Affairs (OIRA)
O.J. Simpson trial 18:499–500
Ojibwa, gender assignment in 9:796
OJJDP *see* Office of Juvenile Justice and Delinquency Prevention (OJJDP)
Ok culture (New Guinea) 13:73
Okada, K. 3:318
Okeh Recording Company 3:800–801
Okiabeth 13:860
Okinawans 12:795
Okri, Ben, works, *Astonishing the Gods* 14:253
Oksapmin linguistic community 13:297
OLA *see* Optimal level of cortical arousal (OLA)
Olanzapine 19:201, 21:22–23
for schizophrenia 21:18–19
Olcott, Colonel 2:891
Old age 17:174–175, 17:178–181, 17:183
achieving 17:179
age policies 1:324
autonomy and dependency in 1:182–183
definitions of 17:183
disability 1:395–396
disease prevalence 1:396
fertility and support in 8:822
habits 17:180
health in 1:394–397
life course 14:76–77
life management
personal goals for 14:87–88
life phase 17:179
life span development 14:110
mortality 1:395
old age insurance 1:324
personal goals as building blocks for life-management into 21:558–559
and population aging 18:540–543
psychiatry *see* Geriatric psychiatry
psychology 17:179–180
retirement pensions 1:324
sociocultural meaning of 17:180
trust across life span 12:639
work participation 17:182–189
advantages and disadvantages of gradual retirement 17:187
background 17:182–183
consequences of working longer for well-being 17:187
demand for older workers 17:186
economic and noneconomic factors driving labor supply 17:187
economic crisis and policy recommendations 17:188
employer attitudes, age discrimination and regulation 17:186–187
employment rates of age group 50–64 years 17:183–185, 17:184f–185f
labor supply 17:185–186
lifelong learning 17:187
negative employer attitudes 17:187

on-the-job training 17:187
pensions 17:182–185, 17:187–188
policy-related research questions 17:187–188
productivity, wages, and substitution 17:186
relation between labor market flexibility and position of older workers 17:187
work disability 17:182–183, 17:186, 17:188
Old age insurance 1:324–325
Old behavioral economics 2:444–445
Old Bering Sea (OBS) culture 1:924
Old Diamond Community (Norco) 7:760
Old institutional economics (OIE) 17:190–191, 20:863
of safety 20:864–866
collective action 20:866
effects of experience rating 20:865–866
workers' compensation 20:864–865
Old institutionalism
in economics 17:190–195
and new institutionalism, demarcation between 17:193–194
Old middle class 2:786
Old Poor Laws 18:731
Old Swiss Confederacy 4:570–571, 4:573
Old Testament 9:266, 12:363, 20:335
Old World dry belt, Noadism in 16:838
Old world of reference publishing 7:584–585
availability of reference works 7:584–585
historic overview of reference works 7:584
systemization of reference works 7:585
Oldenburg Burnout Inventory 2:930
Older adults
common sleep disorders in 22:75
circadian rhythm sleep disorders 22:76
insomnia 22:75–76
REM sleep-behavior disorder 22:76
sleep-disordered breathing 22:76
sexuality in 21:844
sleep pattern changes 22:70–71, 22:71t
Older ages, well-being at 20:599
Older American Act 1965 1:325
Older displaced workers 12:830
Older driver safety 24:626
Older drivers (transportation) 24:623–624, 24:626–627
Older LGBT relationships 13:864–865
Older people
abuse *see* Elder abuse
care *see* Elder care
community care in England 4:344, 4:347–348
emerging critical themes 1:409–410
emerging critical theory 1:410–411
geography of 1:407–412
personal budget
banking services 21:501
benefits 21:500
making easy 21:501–502
skills maximization 21:501
stipulations 21:500–501
using 21:501
and places 1:409
self-directed care and individual budgets for 21:498–503
cash for care schemes 21:498, 21:500
consumer-directed care 21:498–499
direct payments 21:498–499, 21:501
evaluations 21:499
evidence 21:498–499
implementation 21:500
microlevel understandings 21:500
and space 1:407–409
work behavior of 13:197
Older workers
health of 20:599
and labor market 1:413–414, 1:417–418
Oldest-old 7:335–336, 15:863, 17:174–175, 17:178–180
mortality of 15:863–867
age pattern 15:864–865
centenarians 15:864–866
data and future 15:866
gender differences 15:865
Gompertz model 15:864
logistic model 15:864
longevity determinants 15:865–866
modal age at death 15:864
number of survivors 15:863–864
plateau 15:864
supercentenarians 15:864–865
theories 15:866
Old-fashioned Sexism Scale (OFS) 21:697
Olearius, Adam, on Russian sexual life 21:740
Oleoresin capsicum (OC spray) 18:255, 18:258
Olfactometer 17:199
Olfactory bulb 17:196–198, 17:197f
Olfactory cues 15:156
of autobiographical memory 2:285
Olfactory hallucinations 10:500
Olfactory learning in rats 16:471
Olfactory receptor neurons (ORNs) 17:196–198
Olfactory receptors 24:26
Olfactory system 17:196–200
accessory olfactory bulb 17:198
accessory olfactory system 17:198
anosmia 17:199–200, 17:199t
assessing olfactory function 17:198–199
dysfunction and disease 17:199–200, 17:199t
habituation–dishabituation task 17:199
main olfactory system 17:196–198, 17:196f
nasal epithelium 17:196
odor discrimination 17:198–200
odor identification 17:196, 17:198–200
olfactory bulb 17:196–198, 17:197f
olfactory receptor neurons 17:196–198
pheromonal effects 17:198
University of Pennsylvania Smell Identification Test 17:198–200
vomeronasal organ 17:198
Olfactory-ensheathing cells 17:196
OLI *see* Operational Land Imager (OLI); Organization-level interventions (OLI); Ownership-location-internalization (OLI) paradigm
Olick, Jeffey K. 4:200
Oligarchy
elites 7:392
iron law of 17:201–206
Plato on 18:212
and unintended consequences 17:421–422
Oligodendrocyte progenitor cells (OPCs) 2:824–825
Oligopolistic concentration 2:54–55
Oligopoly, Mancur Lloyd Olson, Jr. 17:207–208
Olivier de Sardan, J.P. 2:872–873
Ollantay 11:58
OLLI *see* Osher Lifelong Learning Institute (OLLI)
Olmstead v. United States (1928) 18:948
Olmsted, Frederick Law, Sr. 24:863
OLP *see* Ordinal linguistic personification (OLP)
OLS *see* Ordinary least squares (OLS) regression
Olson, Mancur Lloyd, Jr. 4:146, 17:207–211
antecedents to public/collective choice 17:207–208
oligopoly 17:207–208
public goods 17:208
biography of 17:207
on collective action 17:207–209
on democracy 17:210
on governance 17:209–210
impact of 17:210
on political economy 17:209
on public choice 17:209
on socialism 17:210
on voluntary provision model 17:208–209
works
Logic of Collective Action 11:564–565, 17:208–209, 22:512
Power and Prosperity 17:209–210
Rise and Decline of Nations 17:209
Olsson, J. 25:81–82
Olwig, Karen Fog 1:642
Olympic Games 2008 24:878
Olympic Movement 17:640
OM *see* Orthodox medicine (OM)
O'Malley, Pat 13:810
Oman, legal system in 15:365
Omani colonialism 11:1
OMB *see* Office of Management and Budget (OMB)
Ombudsman 14:265
Omega matrices, in EITMs 18:454–455
Omega Schools 1:270
Omertà 14:425–426
Omi, Michael, works, *Racial Formation in the Twentieth Century* 19:796
Omic data, twin study combining 24:692
Omitted variables, multilevel statistical analysis 23:383
Omnibus Diplomatic Security and Antiterrorism Act 13:531

Omnibus scales 17:754
Omnipredicativity 17:558–559
Omniscient informant 25:753–754
Omnisexual woman 21:751
Omnivore hypothesis, in consumption 4:760
Omnivorous cultural consumption 4:697–698
Omnivorousness 16:555–556
  cultural 22:335–336
  and cultural participation 5:449–451
  cultural production 5:611
OMR *see* Optical mark reading (OMR)
OMT *see* Organization and management theory (OMT)
On airs, waters, and places 7:720
On the Essence of Matter – Das Wesen der Materie nach Kant's Physischer Monadologie 21:963
On the Shoulders of Giants (OTSOG) 15:249
Onassis, Aristotle 14:531
Once-born 20:354
Oncology
  psychological aspects of cancer 3:73–77
  psychosomatic disorders in 19:488
1/N heuristic 5:912, 5:915
One Child Policy 1:145, 2:69, 8:231, 9:47–48, 10:821, 21:735, 24:977
One China Policy 3:506
One city, one product program 23:211
One in Nine Campaign 21:747–748
One person–one language strategy 2:582
One Planet Communities 6:834
One shot Prisoner's Dilemma (two-person) 18:930–934
  N-persons 18:934
  rationality considerations 18:933–934
  from social phenomena to game theory 18:931–933
One-clever-clue heuristic, of decision making 5:920
One-factorial analysis of variance (ANOVA) model 9:836
Oneida community 2:148
O'Neill, John, works, *Five Bodies* 2:730
One-mode networks 5:561
One-mode projection 22:455
Oneness 8:214–215
One-person psychology 24:546
One-reason decision making 5:914–915, 5:920, 19:908
  fast-and-frugal trees 5:914–915
  one-clever-cue heuristics 5:914
  take-the-best heuristic 5:914
One-sex life table 18:590
One-shot *versus* repeat player distinction 14:270–271
One-sided development, in Eastern Europe 6:817–818
One-sided intervals 8:19
One-sided violence 10:3
One-time meeting, in focus groups 9:274
One-variable-at-a-time rule positions 21:985
One-way analysis of variance 1:659–660
Ong, Aihwa
  on immigrants in North America 17:24
  works
    Spirits of Resistance and Capitalist Discipline 8:900
Online behavioral advertising (OBA) 1:204–205, 1:217
Online Christian identity 20:274
Online collectives 4:140–142, 4:141t
Online culture, and religion 20:273–274, 20:276–277
Online data evaluation, in stochastic optimization 23:474
Online dating 17:212
  computer-mediated communication 17:213–215
  dangers involved in 17:215
  deception 17:215
  disinhibition effect 17:213
  hyperpersonal communication 17:214
    contemporary theories about online relating 17:214
  intimacy 17:212–216
  less can mean more 17:213
  self-disclosure 17:215
  sites 17:214
  social information processing theory 17:213
  social networking sites 17:214–215
  social presence theory and social context cues theory 17:212–213
Online education 6:567
  prototypical course classifications 6:568t
Online exchange 8:487
Online *fatwas* 20:273–274
Online identity 11:551–552
Online information evaluation 4:507–509
Online journalism 12:857–860
Online learning communities 6:897–898
  blogs 6:898
  Internet forums 6:898
  podcasts 6:897–898
  social networks 6:897
  video sharing websites 6:898
  Wikis 6:898
  see also E-learning
Online lexical–semantic processes 21:914
Online listening techniques 21:622–623
Online methods, in inference research 12:38–39
Online networks
  and social protests 22:515–516
  terrorist 24:228–229
Online news consumption 16:793–794
  consequences of 16:794
  fragmentation of online media audiences 16:794
  homogenization of online media audiences 16:794
  and information acquisition 16:796
  limitations of previous research 16:795–796
  as a resource for civic engagement 16:794–795
  as a resource for civic participation 16:796
  to seek attitude-reinforcing content 16:796
  user as a content producer 16:795
Online newspapers, and community 4:333
Online petitions 22:514–515
Online pornography 17:217–222
  antipornography debates 17:220
  bad effects 17:219
  camera sites and pornography boundaries 17:219
  children and 17:219–220
  critical contextual analysis 17:220–221
  development of 17:217–218
  diversification 17:218–219
  pornification 17:220
  random estimate 17:218
  regulation 17:220
  social media 17:218
Online processing models, of opinion formation 17:244
Online protection, children 3:427–428
Online publishing, journals 12:862, 12:865–866
Online reading techniques 21:622
Online relationships *see* Sexuality and Internet
Online repositories 19:59–60
Online rituals 20:274–275
Online self-presentation 24:149
Online social networking 24:229
Online support groups (OSGs) 10:686–688
Online testing, experimental laboratories 8:554
Online trust 16:785
Onomasiology 13:934–935
Onomasticon 17:224
Onomastics 14:229, 17:223–226
  literary 17:225
  methodology 17:224–225
    name transfer 17:225
    toponymic research 17:224–225
  names and words 17:223–224
    lexicon and onomasticon 17:224
    name clusters and fields 17:224
    semantic distinctions 17:223
  socio-onomastics 17:225
    education 17:225
    personal names 17:225
On-reading 20:1–2
ONS *see* Office of National Statistics (ONS)
On-site sampling *see* Outcome-based sampling
Ontario Public Services Employees Union (OPSEU) 24:494
On-the-job learning of artists 2:56–57
On-the-job training 17:187
Ontogenesis 16:183–184
  of brain asymmetry 2:809
Ontogenetic adaptations 8:422
  evolutionary developmental psychology 8:422–423
Ontogenetic development 18:129
Ontogenetic neurobiology of brain development 2:821–826
Ontogeny 11:482, 14:369

of human cerebral cortex 3:333–334
pituitary gland 11:482
role of human econiche in 11:257
Ontological analysis, legal philosophy 18:28–29
Ontological certainty, loss of 3:540–541
Ontological dependence 2:132
Ontological naturalism 16:326–328
Ontological otherness, and communes 4:270–271
Ontological politics 10:80–81, 17:305
Ontological presupposition 14:650
Ontological questions 21:190
Ontological relativism 20:235–236
Ontological status
of nonhumans 21:190
of racial categories 19:808
Ontological theory, Wittgenstein's use of 25:612
Ontologies 1:894–897, 11:135, 17:424, 21:183–184
and ANT 5:526
of hunters-gatherers 11:430
implications of realism for 20:24
individualism *versus* collectivism 11:818–819
Latour on 13:414–417
and magic 14:438
nonmodern 5:645
participatory action research (PAR) 17:540–541
Plato on 18:210–211
science and technology studies 21:203–204
social *see* Social ontology
and strong social constructivism 22:231
Taylor on 24:39, 24:41–42
and uncertainty 24:716–717
OP *see* Obsessive passion (OP); Order-of-processing (OP) diagram
Opacity and arbitrariness 11:513
OPACS *see* Office of Price Administration and Civilian Supply (OPACS)
OPCs *see* Oligodendrocyte progenitor cells (OPCs)
OPEC *see* Organization of Petroleum Exporting Countries (OPEC)
Op-Ed Project 1:742–743
Open access (OA)
journals 12:866–867
copyright issues 12:867
Gold Open Access 12:866
Green Open Access 12:866
impacts on use and citation 12:867
system-wide perspectives 12:866–867
variants 12:866
in university press 19:629
Open and closed leagues 23:320–322
Open Anthropology Cooperative (OAC) 12:601
Open borders, global gains, migration 15:438
Open cities 24:784
Open classroom climate 22:731
Open Door Policy 20:137
Open education 14:96, 17:227–230
engagement 17:228
flexible space 17:227
learning 17:227–228
environment 17:227–228
events 17:228
learning resources 17:227
Piaget on 17:227
practices 14:96
teacher as facilitator 17:228–229
teaching 17:228–230
important functions 17:228–229
planning, relationships, and management skills 17:229
teacher characteristics, attitudes, and reactions 17:229–230
Open educational resources (OER) 6:896, 14:96
Open Geospatial Consortium (OGC) 10:62, 23:131
Open government, and public bureaucracies 19:532–533
Open innovation 12:147
Open Languages Archive Community 14:188
Open learner models, lifelong learning 14:93–94
Open list proportional representation 19:207, 19:209
Open market operations, of central banks 15:730–731, 15:733
Open Net Initiative 16:779
Open peer review 12:863
Open (democratic) policies 1:557
Open population microsimulation models 15:344
Open principals 21:137
Open questions 23:767
Open relationships 17:490–491
Open society 5:272–273, 7:926
The Open Society and its Enemies 18:525–526
Open Society Foundations 7:268
Think Tank Fund 24:295
Open Society Institute (OSI) 20:758
Open space 24:862–865, 24:868
Open Street Map 3:162
Open systems 2:915
innovation 12:173
organizations as 3:893
adaptation 3:894–895
boundaries 3:895
critical functions 3:894
design 3:895
equifinality 3:895
general systems theory 3:893–894
information and resource flows 3:894–895
self-regulation 3:895
personality development 17:880
Open-air museum 5:416
Open-architecture networking 12:600
Open-ended exploration 19:344
Open-ended interviews
for epistemological beliefs 7:888
Open-ended welfare programs 25:492–493, 25:492f
Open-mindedness 10:410
OpenMx 6:322
Openness 5:41, 5:45, 10:692, 16:493–494, 17:231–233, 17:777, 17:787–788
in adoption 1:144–145
with education 17:789
and effective learning strategies 17:789
mass communication 14:692–694
and objective research 17:94–95
principle of 19:670
and self 21:439, 21:443
subscales 17:232
Openness to experience 9:236, 16:743, 17:231–235, 17:802, 17:914–915, 25:593, 25:595
and adolescent personality 17:864, 17:867
background 17:231–232
as culture 17:231, 17:233
definition 17:231–232
as intellect 17:231, 17:233
measurement 17:231–232
models of lower-level openness structure 17:231–232
and personality traits 17:922–923
relationships with other constructs 17:232–234
academic achievement 17:232
creativity 17:232–233
intelligence 17:232
political and social attitudes 17:233
psychological well-being 17:233–234
religiosity 17:233
spirituality 17:233
Openness to sensation 17:232
Openness/intellect
changes in over time 17:771–772
cognitive processes linking with creativity 17:771–772
decreased latent inhibition 17:771
divergent thinking 17:771
implicit learning 17:771
working memory 17:771–772
of personality 17:771
Open-resource theory 22:50
OpenScene-Graph 25:133
OpenStreetMap 3:620
Operant conditioning 1:374, 2:415, 3:766–767, 3:769, 13:705–706, 22:409–411
and attitude 2:169
and early memory development 6:762
pain 17:452
and persuasion 17:941
Skinner's 21:847, 22:25–26
and suicide 23:677
Operant conditioning approach 21:847
to language disorders 13:331
Operant model of pain 3:557
treatment for chronic pain 3:559
Operant reinforcement 20:212
Operant theories, of self-regulated learning 21:542
Operating space 16:504
Operation Barbarossa 1:557
Operation Camelot 22:143–144

Operation Ceasefire (Boston) 6:264, 10:460
Operation Desert Storm 15:514, 17:645
Operation Murambatsvina 18:182, 24:968–969
Operation of social systems, Luhmann on 14:386
Operational capabilities 1:16
Operational closure, of social systems, Luhmann 14:386
Operational criteria, for differential diagnosis in psychiatry 6:386
  consequences of definition 6:386
Operational Land Imager (OLI) 20:423
Operational rules 12:192
Operational values, ecological social work 6:874–875
Operationalism 7:89
Operationalization 24:640
  of cultural capital 5:376–377
    qualitative 5:377
    quantitative 5:376–377
  of educational outcomes 5:377–378
    qualitative 5:377–378
    quantitative 5:377
  hierarchies and markets 10:842–843
  of target object, in survey questions 23:766
Operationism 14:871
  measurement theory 14:871
Operations 18:127, 18:130
  concrete 18:128–130, 18:130t
  formal 18:130t
  ordered relational structures 17:296–297
Operations management, supply chain management 23:709, 23:711–713
Ophelemité 17:504
Ophthalmology 10:755
Opiates 6:669, 6:674, 6:686
  use and abuse 17:236–242
Opinio juris 4:809, 12:509
Opinion
  and attitude 2:175–177
  of the court 1:824–825
Opinion assimilation 21:533
Opinion comparison 22:212
Opinion formation 17:243–245
  attitude change 17:244
  attitude formation 17:243–244
  attitude structure 17:243
  beliefs 17:243–244
  constructivism, discourse, and collective publics 17:244–245
  direct democracy 25:296
  emotions 17:243
  intent to act 17:243
  publics and meanings as advocates 17:243–244
Opinion leaders
  Lazarsfeld on 13:638
  in social influence 22:345–346
Opinion leadership 4:291, 4:292
Opinion polls 22:632–635
  market research of 20:878
Opinion Research Corporation 18:461
Opinion surveys 18:459
Opinion-based medicine 8:327–328
Opioid peptide 20:213
Opium 17:236, 19:463
  as a drug of abuse 17:237–239
  early written accounts of opiates 17:237
  and genetics of alcohol metabolism 17:238–239
  history 17:236–237
  use in China 17:237–238
Opium cake 17:236
Opium derivatives 6:685
Opium eaters 17:239
Opium poppy 17:236
Opium smoking 17:239–240
Opium wars 3:531–532, 12:525
Oportunidades 4:556–557, 18:751, 25:488
Oppenheim, Paul 21:302
Opponent process theory for agreeableness 1:468–469, 1:468f
Opportunism, hierarchies and markets 10:842–843
Opportunistic betrayal 16:426
Opportunistic offending 5:147–148
Opportunities
  access, and community social context 4:386
  and choice perspective 25:556
  equality of 6:770–771, 7:903–905
  and equity in education 7:918–922
  and sexual behavior 1:132
Opportunity cost 9:775–776
  cost–benefit analysis 5:47, 5:50, 5:52–53
Opportunity hoarding 16:564
  and inequality 11:910–911
Opportunity International 1:270
Opportunity perspective of crime 20:772
Opportunity recognition 17:364
Opportunity structure
  and collective behavior 4:160
  social mobility 15:642–644
Opportunity-based approaches in insurgency 12:223
Opposing expert testimony 8:584
Opposite-gender partners 10:802t
Opposition faction 8:684–685
Opposition norms 22:982
Oppositional consciousness, collective cultivation of 13:546
Oppositional defiant behavior, early developmental trends in 6:12–13
Oppositional defiant disorder (ODD) 6:7, 6:13, 9:444
Oppositional identity
  and academic motivation 19:793
  and work 8:106–107, 8:110–111
Oppositional interaction 4:857–858
Oppositions 11:575
  Dahl on 5:658
Oppressed 15:135–138
Oppression 15:136, 17:246–252, 17:428–434, 23:324–327
  antioppressive practice 17:248–249
    call to action 17:250–251
  antioppressive theory 17:248
  in Canada 8:152–156
  defining 17:246
  and diablism 6:450
  global forces of 17:249
  globalization 17:248–249
  and homophobia/heterosexism 10:794
  institutionalized 22:822–823
  internalized 22:823–824
  intersectional queer theory 19:766–768
  intersectionality 17:247–248
  and intersectionality 12:657–660, 12:662–668
  and lesbian feminism 8:942–947
  oppressed, liberation theology for 14:32–35
  and racialized groups 22:823–824
  restorative justice 17:251
  root causes 17:246–247
  sanism 15:200–202
  silencing by anti-oppression perspectives, Canada 8:153–154
  and social justice 22:397
  social work contributions 17:249–250
  social work profession 17:248, 17:250
  systemic 22:824
  theory 17:247
  types of 17:246
  and whiteness/white supremacy 25:569–570
  women's studies 9:736, 9:738
OPQ32r *see* Occupational Personality Questionnaire (OPQ32r)
OPR *see* Office of Population Research (OPR)
OPS *see* Optimization in primary and secondary control (OPS) model
OPSEU *see* Ontario Public Services Employees Union (OPSEU)
Opsins 17:269–271
  bistable modulation 17:271
  direct activation 17:270–271
  direct inhibition 17:271
  functionality 17:270–271
  gene expression 17:271
  historical development 17:269
  light delivery and neural interfaces 17:271
  methodology 17:269
  methods for targeting specific cell types 17:269–270
  modulation of intracellular signaling 17:271
Optic ataxia 17:509
Optic axons, Sperry's contribution 23:279–280
Optic nerve stimulation 16:716
Optic nerves, frogs 23:279–280
Optic tectum 23:279–280
Optical mark reading (OMR) 3:294
Optical telegraph 12:491
Optical tomography 16:388
Optical topography 16:388
Optimal case 20:248–249
Optimal control problem 17:253
  see also Maximum principle
Optimal control theory 17:253–255
  application of 17:254
    management of problems 17:254
    optimal level of asset holding 17:254

optimal management of natural resources 17:254
calculus of variations 17:254
discrete time version of 17:254
dynamic programming 17:254
extensions of 17:254–255
motor control models 15:960–962
stochastic 17:254–255
use of term 17:253
Optimal decision strategies 5:906
Optimal designs, Bayesian 8:521–522
Optimal development, resilience 20:557–558
Optimal experimental design 17:256–262
algorithms 17:256, 17:259, 17:261–262
binary data 17:256
conjoint choice designs 17:260–261
criteria of optimality 17:257
discrete choice 17:260–261
effect heterogeneity 17:256
exact and continuous designs 17:257–258
general equivalence theorem and design efficiency 17:258
quadratic model in one variable 17:258
generalized linear models 17:259
for binomial data 17:259
family of models 17:259
for gamma data 17:259
for Poisson data 17:259
I-optimality 17:257
linear logistic model 17:256
logistic regression in one variable, locally optimal designs for 17:260
2m factorial designs and fractions 17:259
nonlinear regression models 17:261
and parameter uncertainty 17:259–260
Bayesian designs 17:259–260
locally optimal designs 17:259
minimax and maximin designs 17:260
sequential designs 17:260
small effects 17:260
simple regression 17:256–257
Optimal foraging
human behavioral ecology 11:247–248
Optimal foraging theory (OFT) 11:248, 11:425
Optimal functioning, and flow 9:262
Optimal inductive inference 21:271
Optimal interpolation 23:191
Optimal interventions, psychological treatments 19:382
Optimal level of cortical arousal (OLA) 21:607–608
Optimal marginality 9:458
Optimal matching/net cost 19:143
Optimal policy 5:906
Optimal posterior point estimates 12:799
Optimal regulatory lag 20:168
Optimal search strategy 5:909
Optimal signal-processing procedure 21:954–955
Optimal spacing, order statistics 17:294
Optimal stopping problems 5:906
Optimal stopping rule 5:906–908
Optimal strategy 5:906–907, 5:909
Optimal tax 7:87
Optimal threat strategies 9:579
Optimal transformations, MDS 16:41–42
Optimality
criteria of 17:257
of experimental designs 8:539
in functionalist grammar 9:474–475
human-demographic traits 8:414–416
and phonological markedness 14:535
in politics modeling 18:512–513
and successful aging 23:667–671
Optimality models
human behavioral ecology 11:248
Optimism 1:536–537, 4:906–907, 6:207–211, 15:942, 17:263–267
climate change 3:841
and coping 17:264–265
and disengagement 17:263–267
downside to 17:266–267
as engine of resources 17:265–266
fostering and interfering with well-being 17:265
and health 7:498
and hope 11:199
magic as source of 14:436
and relational social work 20:224
religion and 20:265
and self-efficacy 21:504–505, 21:507–508
studying 17:264
Optimistic bias 10:698–701, 10:704–705, 10:705t
health risk perception 10:704
and risk perceptions 10:699–700
Optimistic explanatory style, and grit 10:400
Optimization 1:475, 5:911–912
decision-making 5:911–912
geographic models 14:788–789
inverse projection 6:150–152
of life span development 14:103
quantile regression 19:712–713
Optimization in primary and secondary control (OPS) model 14:87
Option value 5:49, 5:53
cost–benefit analysis 5:49, 5:53
and modularity 15:721
Optional sources of sources 6:499
Optional stopping 14:632
Optional stopping problems 5:906
Optogenetics 17:268–273
brain stimulation reward 2:844–845
genetic engineering 17:270
modifications to receptor function 17:272
opsins 17:269–271
bistable modulation 17:271
direct activation 17:270–271
direct inhibition 17:271
functionality 17:270–271
gene expression 17:271
historical development 17:269
light delivery and neural interfaces 17:271
methodology 17:269
methods for targeting specific cell types 17:269–270
modulation of intracellular signaling 17:271
restricted transgene expression in homogenous cell populations 17:271–272
rhodopsin 17:271
studying brain and behavior 17:268–269
targeting neuron projections 17:271–272
traditional tools for manipulating brain function 17:268–269
vectors 17:269–270
Optokinetic response 8:629
Optomotor-blind gene 9:883
Opus Dei (Roman Catholic Church institution) 3:534, 3:538, 16:807
OR gates 16:540–542
ORA *see* Organization for Refugees of America (ORA)
Oral and Written Language Scales 3:796
Oral archives 11:90
Oral argument 1:824–825
Oral expression 17:274–278
Great Divide theories 17:276
problems and approaches 17:275–276
consequences of orality and literacy 17:275–276
ethnographic and historical specificities 17:276
orality–literacy debate 17:276
romantic and evolutionary paradigms 17:275
recent insights and future directions 17:276–277
multiplicity and differentiation 17:276–277
practices and processes 17:277
senses and multimodality 17:277
virtues of ethnography 17:277
spoken expression in human culture 17:274–275
Oral formulaic theory 9:282
Oral health, psychology of 17:279–283
dental anxiety 17:279–281
dental fear and phobia 17:280–281
dentures and adaptation 17:281–282
periodontitis, bruxism, and orofacial pain 17:281
prevention of disease 17:280
psychoneuroimmunology 17:281
quality of life 17:279–280
stress 17:281–282
Oral historical interviews 11:2
Oral history 11:117
African 1:288
African Studies 1:276–278
archiving and disseminating 17:289
composure 17:286–287
credibility and reliability 17:287
defined 17:284
ethics and copyright 17:288–289
history 17:284–285
intersubjectivity 17:286–287
interview 17:288
interviewee 17:287–288
interviewer 17:288

Oral history (*continued*)
literacy 14:236–238
of Maroons 14:594
memory and remembering 17:286
in practice 17:287
problems of dealing with 1:288
recording and transcribing interviews 17:288
rehabilitation 17:285
subjectivity 17:286–287
sub-Saharan Africa 11:90
Oral language 1:450
Oral sex, unprotected 21:798
Oral traditions 8:195, 11:117
folklore 9:280
sub-Saharan Africa 11:89
Orality 14:241, 17:285
consequences of 17:275–276
and illiteracy 11:589
orality–literacy debate 17:276
of radio 19:877–878
Orang Asli 23:71
Orange Park project 10:733
Orature 17:274
Orbitofrontal cortex (OFC) 18:817
convergence of olfactory, taste, and visual inputs in 24:29
dysexecutive syndromes 6:725–726
and emotion 7:479–481
Orcutt, Guy 7:10
Order
form and function in constituent 25:650–652
different constituents, same function 25:651–652
functional modifier categories 25:651
same constituent, different functions 25:651
social *see* Social order
Order ethics 3:17–18
Order in Council 4:248
Order of recollection
collective memory 4:202–203
Order statistics 17:291–295
applications 17:293–294
data compression and optimal spacing 17:294
outliers and selection of the best 17:294
probability plotting 17:293
robust estimation 17:294
selection differentials and concomitants of order statistics 17:293–294
simultaneous inference 17:294
asymptotic theory 17:293
asymptotic independence results 17:293
central order statistics 17:293
extreme order statistics 17:293
intermediate order statistics 17:293
L-statistics 17:293
basic distribution theory 17:291–292
distribution-free interval estimation 17:292
confidence intervals for quantiles 17:292
prediction intervals 17:292
tolerance intervals 17:292
estimation of parameters 17:292–293
properties of ordered random variables 17:291
Order structures 23:629
Order topology 17:299
Ordered anarchy 1:681–682
Ordered assessments 21:7
Ordered relational structures 17:296–300
automorphism group 17:298–299
functions 17:296–297
and Guttman scaling 17:297–298
homomorphisms 17:298
isomorphisms 17:298
operations 17:296–297
order topology 17:299–300
relations 17:296–297
representational theory of measurement 14:857–859
several orders on one set 17:298
structures completion 17:300
and weights 17:297
Ordered weighted averaging (OWA) 23:138
Ordering axioms 14:862
Ordering principles 25:653
Ordering space 17:304
Ordering types 10:774
Orderliness, and conscientiousness 4:658–659
Order-of-processing (OP) diagram 16:543, 16:543f
as general arrangements of processes 16:543
Orders of creation 3:237
Order-specific fertility rate 6:111–112
Ordinal 22:879
rational choice explanation 19:899
Ordinal linguistic personification (OLP) 14:195
Ordinal multidimensional scaling 16:42
Ordinal scales
construction of 24:252–253
in ethnography 8:189–190
Ordinal sequence synesthesiae 14:193
Ordinalism 7:88
Ordinality 11:902
Ordinariness, celebrity 3:275–276
Ordinary intuition, value pluralism 25:22
Ordinary knowledge 2:749
Ordinary least squares (OLS) regression 12:206, 12:661–662, 21:598, 23:97, 23:384, 24:706–707
Ordinary legislative procedure 8:265
Ordinary naturalism 16:326–327
Ordinary naturalist 16:326
Ordinary people 13:587–588
Ordinary rationality, and collective beliefs 4:166
Ordnungspolitik 3:17
Ordonnance civile of 1667 7:925
Ordre public 12:264–265
Oregon rationing program (OHP) 10:604–605
Oregon v. Smith (1990) 20:280
Orexin, and wake 22:89
Organ donation, and commodification 5:538
Organic brain syndrome, and alcohol 1:511
Organic embodiments 22:2
Organic gastrointestinal disorders 9:605–606
Organic ideologies 11:564–565
Organic intellectuals 10:747, 17:541
Organic physics 10:756, 10:758
Organic social solidarity 4:361–362
Organic societies 1:719
Organic solidarity 6:701–702, 13:566, 18:668
in modern West 16:834
Organicism 22:991
Organisation for Economic Co-operation and Development (OECD) countries 1:148, 3:502, 4:744–745, 5:18, 5:763, 5:766, 5:769–770, 5:776, 5:790, 5:801, 5:874, 6:834, 7:180, 7:211, 7:267–268, 8:60, 9:317, 10:248, 10:257, 11:835–837, 16:465, 16:774, 21:119, 21:150, 21:232, 21:322, 21:649, 22:116, 22:325, 24:658, 24:662, 24:850
Adult Literacy and Life Skills Survey 5:847
countries, social assistance programs in 25:491t
Declaration on the Protection of Privacy on Global Networks 12:596–597
Declaration on Transborder Data Flows 12:596–597
family policy 8:801
family socioeconomic status and children's school learning 8:752
influence on global education policy development 7:268
jobs study 13:185
pension policy outside 17:694
principles, corporate governance 4:937
Principles of Corporate Governance 13:917
Program for International Student Assessment (OECD-PISA) 5:343, 9:679
Programme for International Student Assessment 5:847, 24:51–52
Reviews of National Policies for Education 7:236–237
shadow economy labor force 21:863–864, 21:864t
Statistics Portal 5:748
System of Composite Leading Indicators (OECD CLIs) 5:748
taxation in 23:340–341, 23:341t
teacher education 24:51–52, 24:54
trade union density 24:497
welfare retrenchment 25:521–526, 25:523t
women 's education 22:326
Organismic self-regulation 21:548–549
Organism-in-the-environment 18:354
Organisms
theory of 13:3
traits, evolutionary history 9:536
Organization and management theory (OMT) 17:405
critical views in 17:407–408
managerialist studies of power in 17:406–407

Organization for Refugees of America (ORA) 11:632
Organization for Social Science Research in Eastern and Southern Africa (OSSREA) 22:604, 22:608
Organization Man, (character) 18:665, 18:668
Organization of African Unity (OAU) 11:800
Organization of American States (OAS) 10:156, 11:799–800, 23:48
Organization of industry, in factory era 11:855–856
Organization of Islamic Cooperation (OIC) 21:792
Organization of Petroleum Exporting Countries (OPEC) 17:166, 20:571
Organization of the Islamic Conference (OIC) 24:604
Organization of work, in hospital 11:212
Organization principle of energy 7:634–636
Organization strategies
  for basic learning tasks 13:713
  for complex learning tasks 13:713–714
Organization structure, in popular music market 3:802
Organization theory (OT) 2:481, 16:764–765, 17:307, 19:892–895, 20:408, 21:137, 19:895t
  agency theory 19:892, 19:894
  bounded rationality 19:894
  classical 1:103
  divergence of 19:893
  and economic theories of firm, convergence and divergence of 19:894
  markets 19:893–894
  origins and premises 19:893
  and rational choice theory, link between 19:894
  rationality 19:893
  social-regulation processes 1:105
  transaction cost economics 19:892, 19:894
Organizational actors, in-house counsel as 12:127
Organizational aesthetics 16:180
Organizational age 1:420
Organizational ambidexterity 12:148
Organizational aspects, negotiation and bargaining 16:415–417
Organizational behavior (OB) 17:313, 17:327, 17:331–332, 17:697–702
  behavioral consistency across demographically heterogeneous workgroups 17:699–700
  creativity and innovation 17:700–701
    across situations 17:700
    fostering 17:700–701
  culture 17:698
    behavioral consistency across 17:698–699
    individuals' influence 17:699
    situation selection based on 17:699
  demography 17:699–700
  and genetic factors 9:921–922
  individual and collective goals 17:697
  influence in 18:771
  leadership 17:700
    as function of strength of situation 17:700
    some individuals can effect change more than others 17:700
  macro-organizational behavior 17:697
  micro-theory 17:697–698
  person–situation interactions 17:697–698
  politics in 18:771–772
  power in 18:770–771
  psychology of 17:313–316
    action theory 17:315
    bureaucratic theory 17:314
    cognitive psychology 17:314
    nature of field 17:313–314
    psychology of individual differences 17:314
    systems thinking 17:315
    unresolved issues and future directions 17:315–316
  research 17:697
  self-categorization theory 21:458
  white-collar workers 25:566
Organizational behavioral management, applied behavioral analysis for 2:410
Organizational benefits
  of legal databases 5:803–805
    LexisAdvance 5:803–804
    WestlawNext 5:804–805
Organizational boundaries 2:768–772, 3:895
  collaboration in research and development 2:769–770
  community principles 2:771
  flattening of hierarchies and the spread of networks 2:769
  learning by monitoring 2:768–769
  novelty and multiple networks 2:770–771
  ramifications 2:770
  user–firm arrangements 2:770
  workplace 2:768–769
Organizational capability 3:23
Organizational capitalism 2:750
Organizational challenges, comparative database research 5:795–796
Organizational changes 2:484, 21:135
  conducive to expansion of nonphysician health professionals 16:908
  and ecology 6:891–892
    benefits 6:891–892
    change process 6:891
  within military 15:513
Organizational characteristics, alternative schools 21:39–40
Organizational citizenship behavior (OCB) 1:422–423, 12:835, 17:937
  composition of 17:318
  construct and its boundaries 17:317
  contextual performances 17:318
  and cultural matrix 17:319–320
  current conceptions 17:317–318
  current trends and future tasks 17:320–321
  dark side of 17:320
  and individual outcomes 17:319
  interpersonal (OCB-I) 17:318
  job attitudes 17:317–318
  job motivation 17:320
  job performance 17:317–318, 17:320
  job satisfaction 17:317–318, 17:320
  leader–member exchange 17:319
  organizational (OCB-O) 17:318
  and organizational effectiveness 17:319
  performance evaluation 17:319
  social exchange 17:319–320
  theoretical roots 17:317
Organizational climate 17:327–333
  competing values model 17:323
  definition 17:322–323, 17:327
  facet-specific climate 17:323, 17:325
  future directions 17:331
  global climate 17:323–325
  history of 17:327
  leadership 17:322–325
  measurement of 17:329–330
  measuring 17:323–324
  molar climate *versus* focused climate 17:328–329
  organizational culture 17:322–323
  versus organizational culture 17:327–328
  organizational effectiveness 17:323
  psychological climate 17:322–324
  research on 17:330–331
  research on antecedents and outcomes 17:324–325
Organizational Climate Description Questionnaire (OCDQ) 7:820
Organizational cognition *see* Managerial and organizational cognition (MOC)
Organizational commitment, social capital 22:183
Organizational conflict 4:639–643
  culture 4:639, 4:641
  dyadic conflict 4:639–640
  ethnographic approaches to 4:640–641
  natural selection 4:641–642
  negotiated order 4:641
  negotiation 4:640–641
  and neo-Darwinism 4:641–642
  neo-Machiavellian view of 4:642
  perspectives on 4:639t
  political competence 4:642
  power 4:640–642
  and social psychology 4:639–640
Organizational context 21:627
  supervision in 19:91–92
Organizational control 17:341–345, 21:137
  ambiguities and absences 17:341–342
  break from command 17:341–342
  break from control 17:341–342
Organizational crime 25:554–556
Organizational culture analysis
  of corporate culture 4:923–924
Organizational cultures 5:536, 11:303, 11:351, 17:322–323, 17:327–328, 17:698
  anthropology of 17:346–351
    approaches 17:346–347
    cultural diversity 17:349–350
    Japanese connection 17:347–348

Organizational cultures (*continued*)
managing culture in transnational organizations 17:349–350
national culture 17:346–349
rise of 17:347–349
US corporate culture 17:348
value-based management 17:348–349
working with culture in transnational organizations 17:349–350
behavioral consistency across 17:698–699
business anthropology 3:8
center and periphery 1:106
individuals' influence 17:699
versus organizational climate 17:327–328
situation selection based on 17:699
variation in 1:106
Organizational decision making (ODM) 17:352–357
cybernetic model 17:354
fairness 17:355–356
garbage can model 17:354
heuristics 17:353–356
implications for organizational structures 17:356–357
information and communication technologies 17:356–357
logic 17:353–355
classic continuum 17:353–354
for innovation and discovery 17:354–355
rationality divide 17:353
organizational effectiveness 17:352–353, 17:355
politics 17:355–356
conflict resolution for efficiency and equity 17:355–356
power/efficiency controversy 17:355
uncertainty 17:353–356
value maximization 17:353
voting and gaming' conflict resolution mechanisms 17:355
Organizational democracy, and workplace autonomy 25:720–721
Organizational design 3:895
and bureaucracy 2:909
Organizational design and form 6:235–240
alignment 6:236–237
bridging theory and practice 6:237–238
complexity theory 6:235, 6:237
congruence model 6:236
contingency theory 6:236
futurists 6:239
items of information 6:239–240
knowledge directions 6:237
metaphors of organization 6:237–238
multiple perspectives 6:237
self-organizing networks 6:239–240
social relationships 6:236
systems theory 6:235
theories 6:236–237
trends and fads 6:238–239
Organizational development (OD) 6:296–300
change agent 6:297–298
conception 6:297
Confrontation Meeting 6:297–298
continuous change 6:299–300
as discipline 6:296–297
episodic change 6:299
area for specialists only 6:299
clear-cut process 6:299
exception in organizational life 6:299
equilibrium 6:296, 6:299–300
fluidity 6:299–300
historical overview 6:296
human relations 6:297
implementation 6:297–298
models 6:297–298
grid organization development 6:298
process consultation 6:298
survey-guided development 6:297–298
systemic interventions 6:298
paradoxical intervention 6:298
participation 6:296–299
resistance to change 6:296
shared power approach 6:297
Survey–Feedback approach 6:297
unfreezing 6:296–299
Organizational deviance 25:722–723
Organizational discourse 17:409
Organizational ecology (OE) 6:890–894, 17:358–363
demographic processes 6:890–892
age and size dependence 6:890–891
organizational change 6:891–892
organizational founding 6:890
organizational learning 6:892
density dependence model 6:893t
liabilities of newness and adolescence 6:891
organizational failure 6:892–893
processes 6:892–894
niche-width dynamics 6:892
population dynamics and density dependence 6:892–893, 6:893t
population learning 6:893–894
resource-partitioning theory 6:892
structural inertia theory 6:891–892
Organizational economics, approaches to interorganizational relationships 12:622
Organizational education 4:429
Organizational effectiveness 17:323, 17:327–332, 17:352–353, 17:355
organizational citizenship behavior 17:319
Organizational efficiency 21:628
Organizational emergence
business development 17:364–367
complexity 17:366
differentiation 17:364
early firm growth, alternative perspectives on 17:367–368
ecosystem 17:364, 17:367
new firm formation and value creation 17:364–367
building a productive base 17:365–366
creating and sustaining returns 17:366
early organizational development 17:364–365
ecosystem evolution 17:367
firm in its selection environment 17:366–367
fitness, issue of 17:366–367
opportunity recognition 17:364
organizational evolution 17:367
organizing productive activity 17:366
routines 17:366
stages of growth 17:365, 17:367
value creation 17:364–367
Organizational environmental sustainability, human resources for 19:83
Organizational ergonomics 11:302–303
Organizational evolution 17:367
Organizational failure, and organizational ecology 6:892–893
Organizational fields 9:146–147, 17:311, 17:375–378
adaptation 17:375, 17:377
effect of fields on institutions 17:377
effect of institutions on fields 17:376–377
empirical interests 17:376–377
foundations 17:375–376
homogeneity 17:375–377
institutional logics 17:377
institutional theory 17:376
institutional work 17:377
interorganizational relations 17:375–376
isomorphism 17:375
Organizational forms 2:768–772, 16:548–549
collaboration in research and development 2:769–770
flattening of hierarchies and the spread of networks 2:769
innovation 2:769–771
invention 2:771
novelty and multiple networks 2:770–771
public sector organizations 19:599–600
ramifications 2:770
workplace 2:768–769
Organizational founding 6:890
Organizational framework, of transnational organized crime 24:600
Organizational goals 2:481–482
Organizational heterogeneity, in hospitals 11:210
Organizational information and knowledge 12:72–76
Organizational institutionalism 12:188
Organizational interaction 21:135
Organizational justice 12:930, 17:135–136, 17:138–139, 17:148, 17:379–380
descriptive *versus* normative justice 17:379
dimensions 17:380–382
distributive justice 17:380
interactional justice 17:381
procedural justice 17:380–381
group-level justice 17:382
attraction–selection–attrition model 17:382
social information processing theory 17:382
multifoci justice 17:382
perceptions 17:379–380

rules in real-world settings 17:381
interactions between and among the types of justice 17:381–382
overall justice 17:382
Organizational justice model, for work-related stress 23:548–549
Organizational knowledge
diffusion of 16:77–78
and tacit knowledge 24:4–6
Organizational learning 12:148, 12:173, 13:695–698
absorptive capacity 1:12–13, 1:15–16
adaptation to the environment 13:695
behavioral change through trial and error 13:695
and ecology 6:892
economy of search 13:695
epistemology 13:697
field growth and divergences 13:696–697
history of concept 13:695–696
knowing in practice and becoming 13:697
learning curves 13:695
learning organization 13:696–697
metaphor 13:695
modification of mental maps 13:695
organizing 13:697
practice turn 13:697
and tacit knowledge 24:7
training and business management 4:114
Organizational level emotions 7:509–510
definition 7:509–510
Organizational management of traumatic stress 15:499
Organizational materialism 14:494
Organizational models, computational organization theory 4:485, 4:487, 4:487f
Organizational myths 16:179–182
mediatory myths 16:182
role of saga in institutionalization process 16:181–182
Organizational narratives 16:195
Organizational negotiation skills 22:805
Organizational networks 16:546
Organizational path dependence 17:385–389
complementarity effects 17:386
coordination effect 17:386
Formation Phase 17:385–386
implications for organizational research and managerial practice 17:389
inertia 17:388
learning effect 17:386–387
lock-in 17:385, 17:387–389
network 17:385–389
persistence 17:385, 17:388–389
Preformation Phase 17:385
related concepts 17:388
routines 17:385–386
self-reinforcing effects 17:386–387
self-reinforcing mechanism 17:385–389
theory 17:385–388
Organizational performance 17:611–612, 23:511
and diversity 4:738, 6:595
dominant design 6:640–641
structural contingency theory 23:609–614
Organizational perspective of coping 4:903–904
Organizational pluralism 18:234
Organizational processes, design of 11:303
Organizational pro-environmental initiatives 19:82–83
Organizational professionalism 13:630–631
Organizational psychology 11:880
genetic factors in 9:919–925
personality assessment 17:830–831
broad dimensions of personality 17:830
narrow facets of personality 17:830–831
Organizational reality 17:409
Organizational relationships, managing 21:137
diplomatic political perspective 21:137
Organizational reliability, bounded rationality for 2:775
Organizational research, and path dependence 17:389
Organizational resources 21:135
Organizational routines, Nelson and Winters concept of 8:430–431
Organizational size, contingency of 23:610
Organizational slack 17:366
Organizational sociology 18:252
Organizational structures, and decision making 17:356–357
Organizational studies, of uncertainty 24:719–720
Organizational symbolism 16:179–182
rites 16:179–181
sensory knowledge 16:181
types and functions 16:180–181
artifacts and aesthetic knowledge 16:181
symbolic actions 16:181
verbal symbols 16:180–181
Organizational systems
models 6:236
negotiated orders 17:413
Organizational tenure 1:420–423, 1:425
Organizational tradition 14:479
Organizational transformation 10:413
Organization-based instruments 19:572–573
Organization-level interventions (OLI)
for occupational stress management 17:150–151
effectiveness 17:150
implementation 17:151
participation of employees 17:151
Organizations 3:21, 3:24, 13:609, 16:8–9, 16:11, 16:194–197, 17:301, 17:307–312, 22:461, 22:463–464, 22:596–597
advertising agencies 1:191
advocacy 1:223–224
agency 17:311
authority and power 17:405–410
basic demography of 17:358–359
bureaucracy 2:913–918, 17:307, 17:310
catastrophic errors 17:370–374
action research 17:372
recent approaches 17:370–372
changes in 22:803–804
changing theories 17:311
civil society 22:576–577
definition 22:576–577
practice 22:577
social policy 22:577
closed and open systems 3:893
cognitive blindness in 1:106
commitment 17:334–340
complex, new institutionalism in the analysis of 16:764
context, The Agenda 10:145
creative management of 22:804–805
and culture 17:390–396
agency and institutional change 17:393
microfoundations of institutions 17:393–394
and neoinstitutionalism 17:393–394
phenomenology of firm 17:390–391
power and culture 17:391–392
travel of institutions 17:392–393
defining 4:486
density dependence 17:360–361
difference from social groupings 17:308–309
of disabled people 6:462
economic education 7:21
emergence of 17:309–310
emerging trends 17:361–362
epistemic cultures in 7:878
evolving perspectives on 17:307–308
of executive government 8:498–500
existence 17:308–309
and factions 8:683
factors related to burnout 2:929–932
generational 9:867–868
geography of 17:301–306
globalization 17:310
and governmentality 10:310–311
group processes in 10:408–415
hierarchies and markets 10:842–843
human ecology 11:284–285
industrial, role for state in 14:24
industrialization 11:869–871, 17:309–310
inertia 17:359–360
information and communication technologies 17:311
informed consent in 3:903
institutional economics 17:190, 17:192–194
institutional economists views 17:309–310
institutional theory 17:308
institutionalized categories 17:360
as interpretation systems 1:106
issues related to environment in Soviet Union 20:832
job design and evaluation 12:816–818
and law 13:543–547, 17:397–404
classical roots 17:397
compliance 17:397, 17:400–402
constitutive law as dependent variable 17:402

Organizations (*continued*)
constitutive law as independent variable 17:401–402
courts 17:399–400
current scholarship 17:398–402
disputing 17:399
facilitative law as dependent variable 17:399–400
facilitative law as independent variable 17:399
future directions 17:402
governance 17:397–399, 17:402
law as a facilitative arena for interaction 17:398–400
law as code of conduct 17:400–401
law as typology of experience 17:401–402
litigation 17:397–401
mid-century revival 17:397–398
regulative law as dependent variable 17:400–401
regulative law as independent variable 17:400
studies of commercial contracting 17:398
studies of legalization 17:398
studies of regulation 17:397–398
studies of transaction costs 17:398
torts 17:400–402
leadership in 13:664–669
and legitimacy 13:823–827
levels of analysis 17:310–311
and libraries 14:46–47
Luhmann on 14:386–387
members, as *homo narrans* 16:194–195
and milieus 14:20–21
modularity and 15:720–722
as moral communities 1:106
'natural system' model 17:307–308
negotiated orders in 17:411–414
network theorists views 17:309
niche 17:360
organization theory 17:307
organizing 17:309–311
patterns of 12:373–374
perspectives of strategic adaptation 23:502
population ecology 17:308
population ecology models 17:308
and production, cultural embeddedness of 5:537
professions in 19:101–106
promoting economic education 24:83–84
public administration 19:521–525
'rational system' model 17:307
rationality in 19:911
rationalization 17:307, 17:310
relational views of 17:309
resource dependence theory 17:308, 17:310
resource partitioning 17:361
selection 17:359–360
setup of universities 18:309–310
sexual harassment in 21:723–725
social capital in 22:181–185
and social change 22:193–195
social networks 17:309
social work in 22:801–806
sociology of 17:415–419
as constructs of collective action 17:415–417
dysfunctions of bureaucracy 17:415–417
efficiency and legitimacy as unifying forces 17:417–419
in space 17:302–303
organizations forming space 17:302–303
space-forming organizations 17:303
spatial representations production 17:303
special, universities as 10:845
storytelling in 16:196–197
strategic reflexivity and innovation in 12:173
as systems 3:893–894
levels 3:893–894
parts, relationships, and wholes 3:893
and technology 24:113–116
overview 24:113
theoretical perspectives 17:301–302, 17:301t
transaction cost economics 17:308–311
types of 17:302
ubiquity of 19:597
visual methods 25:182–183
volunteerism 25:268–272
ways of studying 16:195–196
of work 25:687–689, 25:691
see also High reliability organizations (HRO)
Organizations forming space 17:302–303
Organized activities, urban growth models 24:807
Organized attachment 12:230–231
Organized capitalism 3:106, 18:666, 25:565
Organized crime 5:170–175, 12:465–467, 14:425, 14:429
definition 5:170
as ethnic subcultures 5:172
expansion of 14:432
functional theories 5:171
gray economy 5:171
groups 14:430
as (ethnic) hierarchies/bureaucracies 5:172
as networks 5:172–173
as patron–client structures 5:172
scope and structure of 5:172–173
social disorganization 5:170–171
transnational 5:173–174, 14:433
in China 5:173
in Italy 5:174
in Russia 5:174
transnational organized crime 24:596–601
types of activities 5:170
UN definition 14:429
in United States 5:173
mafia 5:173
weak government 5:171
Organized hate group 10:542–543
Organized interests 12:371
judicial behavior 12:888
Organized medicine 13:472
Organized religious forms 3:540–541
Organized shortage, in socialist societies 22:850
Organized skepticism 9:620–621, 17:11, 21:214–215
peer review as 17:685
Organized social behavior, social influence as determinant of 21:457
Organized violence 17:649, 17:652
and territorial conflict 24:214–215, 24:219
war as 25:422–423
Organizing 16:194–196, 17:309–311
organizational learning 13:697
Organizing principles of object selectivity 16:631–632
Organizing space 17:305
Organogenesis 2:116
Organon model
of interpersonal communication 7:484, 7:486, 7:484t
signals 7:484
symbols 7:484–485
symptoms 7:484
of linguistic sign 18:795
Orgasms 21:669
decreasing frequency of 21:669
Oriental dance 5:668
Oriental depravity 17:429
Oriental identity 16:150
Oriental monasticism 15:724
Oriental Orthodox 17:435
diaspora 17:438
missionary outreach 17:438
Oriental philosophy 15:808
Oriental sensuality 17:429
Oriental splendor 17:429
Oriental studies in interwar period 1:949–950
Orientalism 3:308–309, 4:219, 10:87, 15:354–355, 15:693, 17:423–434, 18:647, 20:374, 25:799
in Ancient India 11:78
critiques 17:426, 17:431
in Eastern Europe 6:805
Islam 17:423, 17:425–426
key concepts in 17:428–429
meaning 17:428–429
orientalist stereotypes 17:429
Said's orientalism 17:428
legal 18:649–650
Maghreb on 21:750–751
in MENA 16:361
and Middle East politics 16:371–372
and Middle East society 16:381–385
'myth' of East *versus* West 17:426
and persistence of anti-Arab racism 17:429–431
and persistence of Islamaphobia 17:429–431
representations of Arab 17:429
representations of Muslim women post-9/11 17:429
resources for 17:426
Said on 17:423–425
critiques 17:425–426
for social work 17:428

teaching in social work 17:432–433
foundational critical theory 17:432–433
western gaze in social work 17:431–432
cultural competence 17:431–432
international social work 17:432
Orientalist reality 25:210–211
Orientalist research, classical archaeology 3:760
Orientalist stereotypes 17:429
Orientation 10:69–70, 10:73–74, 23:198–200
and posttraumatic amnesia 18:637–639
reference group 20:72–75
in spatial cognition 23:113–114
Orientation bias, and hemispatial neglect 10:769
Orientation channel 25:185
Orientation columns in visual cortex 5:31–32
Orientation programs, cross-cultural training 5:334–336, 5:340
Orientation skills 11:390
Orientation tuning 17:128
Orientation-Log 18:637
Orienting/regulation 17:877
Original ecocultural theory 6:839–840
Original position 2:315
Rawls on 19:914–915
Original society 4:339
Originalism 13:742–743
Origin–destination models 24:470
Origins of art 18:826–827
Origins of food production 23:67
Origins of war 4:616–617
Orinoco Delta of Venezuela 25:434
Orkidebarn 8:425
Ornithology 14:369–370
ORNs *see* Olfactory receptor neurons (ORNs)
Ornstein–Uhlenbeck (OU) models 6:398–399
Orobability models 24:344
Orofacial pain 17:281
Orosius 10:162
Orphanhood method, in indirect estimation 6:145
Orphans, adoption of 1:137–139
Ortega, José
on mass society 14:745
works
The Revolt of the Masses 14:743, 14:745
Ortelius, A., works, *Theatrum Orbis Terrarum* (*Theater of the Lands of the World*) 13:226
Orthodox Christianity 17:437, 20:331
Orthodox Church 20:408
in America 17:437
black Orthodox of Africa 17:437
of Byzantine empire 17:435
Ethiopian 17:437
Greek 17:436
Orthodox Diaspora 17:437
Orthodox of Uganda and Kenya 17:437
during Ottoman rule 17:435–436
Russian 17:436–437
Syrian 17:437
use of term 17:435
Orthodox churches 17:435–438
Orthodox economics 7:93
Orthodox Jewish feminist movement 12:876
Orthodox medicine (OM) 10:563
and CAM, consultation differences between 10:564–565
Orthodox of Bulgaria 17:435
Orthodoxy 2:780, 3:280
Marxism–Leninism as 14:659–660
Orthogenesis 5:677
Orthogonal matrix 19:80
Orthogonal Procrustes least-squares solution 19:79
Orthogonality criterion 19:80
Orthographic depth hypothesis 6:736
Orthographic material, visual mental imagery of 25:165
Orthographic-to-phonological recoding 21:913
Orthography 3:510, 3:512, 14:242–243, 20:5–10, 20:6f
Orthonormal functions 23:399–400
Ortner, Sherry
on capitalism in North America 17:24
on culture 5:557
works
Sexual Meanings 8:899
Ortssippenbücher 6:138–139
ORU model of skill mismatch 15:598
Orwell, George 18:434–435
works, *Nineteen Eighty-Four* 2:153–154
Oryza sativa japonica rice 23:66
OSA *see* Obstructive sleep apnea (OSA)
OSAGI *see* Office of the Special Adviser on Gender Issues and Advancement of Women (OSAGI)
Osborn, Frederick 17:39–40
OSCAR model 24:201
OSCE *see* Objective Structured Clinical Examination (OSCE)
Oscillations 16:635–636
Oscillatory synchrony 9:483
Osgood, Charles Egerton
semantic differential theory 21:563–570
works
The Measurement of Meaning 9:361
Semantic Differential Technique: A Sourcebook 9:361
Osgood, Cornelius 2:72
works, *The Koreans and Their Culture* 2:72
Osgood, W. 7:378
OSGs *see* Online support groups (OSGs)
OSH Act *see* Occupational Safety and Health Act (OSH Act)
OSHA *see* Occupational Safety and Health Administration (OSHA)
Oshana, Marina 2:319
Osher Lifelong Learning Institute (OLLI) 1:150, 7:142
Osho 16:807
Osho sannyasins 16:806–807
OSI *see* Occupational Stress Indicator (OSI); Open Society Institute (OSI)
OSIRIS (statistical analysis package) 23:378
OSLER *see* Objective Structured Long Examination Record (OSLER)
Osram 11:840
OSS *see* Office of Strategic Services (OSS)
OSSREA *see* Organization for Social Science Research in Eastern and Southern Africa (OSSREA)
OST *see* Out-of-school-time (OST) activity
Ostate variables 17:253–254
Ostensive definition 23:594–595
Osteoarthritis 2:601
Ostpolitik 4:129–130
Ostracization 16:864
Ostracoda 7:694
Ostrom, Elinor 19:538
Ostwald, Wilhelm, works, *Annalen der Naturphilosophie* 25:611
Oswald, R.F., family resilience model of 21:755–756
Oswin, Natalie, on NCGLE 21:746
OT *see* Organization theory (OT)
OTA *see* Congressional Office of Technology Assessment (OTA)
Other, the, Lacan on 13:211–215
Other (phenomenology) 18:21–22
Other-directed network actors 5:558–559
Other-directed social character 20:671
Other-face concern 11:186
Othering 22:821
and infrastructure 12:123
Otherization 16:864
Otherness 8:214–215
fear of 25:802–803
respecting 20:221–222
and selfhood 21:447
of Western 17:435
Other-race effect 8:655
Other-rated personality 17:790
Other-regulation, and self 21:442
Others
knowledge of 16:602
self–other mapping 16:602–603
Others' emotions, responses to 7:551
Otoacoustic emissions, and sexual orientation 21:776
Otoliths 25:63–64, 25:66, 25:69, 25:64f
Ottawa Charter of 1986 20:558
Ottawa Statement 3:900–903, 24:282–283, 3:901t
feasibility of consent 3:902–903
identification of research participants 3:901–902
informed consent in organizations 3:903
informed consent requirements 3:902
justification for design 3:900
need for research ethics review 3:900–901
role of cluster gatekeepers 3:903
timing of consent 3:902
Otto, Rudolf 18:437–438
Otto of Freising 7:721
Ottoman codes 15:360–361

Ottoman Empire 3:707–708, 15:354, 15:360–361, 16:297
  same-sex sexualities in 20:870
  slaves/slavery 22:54–55
Ottoman frontiers 9:468
  Byzantine imperial 9:468
  Islamic religious 9:468
  Köprülü 9:468
  nomadic warrior 9:468
  Wittek 9:468
Ottoman Land Code (1858) 15:361–362
Ottoman laws 12:749–750
Ottoman Turkish 16:362
Ötzi 17:460
OU *see* Ornstein–Uhlenbeck (OU) models
Ought self 21:470
Oughtness, norms as 17:5–6
Our Bodies Ourselves (OBOS) 15:112–113
Our Global Neighborhood 7:927
OUS *see* Overt, undifferentiated shame (OUS)
Out Now Consulting 9:637
Out of alignment 7:578
Outcome distribution, generalized linear mixed models 9:850
Outcome evaluation 19:118
  logic models 14:300–304
Outcome expectations
  health behavior 10:578, 10:580
  motivation 10:712–713
  self-efficacy 21:509, 21:512
Outcome satisfaction, in public evaluations 13:749
Outcome theories of equality, critiques 7:899–900
Outcome-based sampling 17:103–106
  design and robustness issues in 17:103
  extensions and related issues 17:105
  statistical inference in 17:104–105
Outcome-oriented environments 11:697
Outcomes
  pharmacoeconomics 19:458
  psychotherapy 19:382, 19:518–519
Outcomes prediction, personalized medicine 17:928
Outcomes-based education (OBE) 7:820
Outcomes-Based Learning Environment Questionnaire (OBLEQ) 3:827
Outcroppings, ethnography 8:188
Outdegree centrality 16:537
Outer suburbs 23:661, 23:663–665
Outgoing reflection centrality scores 16:537
Out-group hatred, genocide as 10:16
Out-group homogeneity
  effect of 12:410
  and social categorization 22:187, 22:189
Out-group identity 11:527–528
Out-group process, organizational groups 10:412–413
Out-groups, immigration 11:675–676
Outlier detection, time series 24:319–320
Outliers 11:496, 20:731, 23:387–391
  accommodation of 23:389–390
    trimming 23:390
    winsorizing 23:390
  block and consecutive tests 23:389
  college grades 23:387–388
  in complex data situations 23:390
  discordancy testing 23:388–390
  multiple 23:389
    masking and swamping 23:389
  order statistics 17:294
  paternity case 23:387
  in univariate samples 23:388–389
    causes 23:388
    meanings 23:389
    retention, identification, rejection, incorporation 23:388–389
Outlook Tower, in Edinburgh 22:698
Outlying observation *see* Outliers
Out-of-court assistance to unrepresented litigants 12:925
Out-of-home activities
  families as educational settings 8:736
Out-of-home environments, and aging 6:886–887
Out-of-school influences, in self-regulated learning 21:545
Out-of-school-time (OST) activity 3:449
Out-of-the-ordinary guarantees 24:673
Output harmonization 3:294–295
Output intensity indicators, land use 13:239–240
Output mix, international migration effects on 12:516–517
Outreach, in disasters 22:757
Outshining, and context dependent memory 4:780
Outside perspectives problem, Christian historiography and historical thought 11:31
Outsider, in queer anarchist politics 19:749
Outsider art, avant-garde 2:322
Outsider scholarship 13:794–795
Outsiders perspective of theology 20:357
Outsourcing 3:43, 10:258, 16:771, 18:248, 18:666–667, 21:650–651, 25:692
  civil service and 3:699
  diversification 6:585, 6:588
  in law firms 13:582
  supply chain management 23:709, 23:711
Outsourcing service providers, multinational corporations as 3:47
Outsports.com 9:659
Outward-oriented militarism 15:493
Outwork handcrafts 5:111
Overaccumulation 3:579–580
Overall justice 17:382
Overall life satisfaction 24:401
Overall survival probability 8:312
Overarching problems 19:49
Overarousal hypothesis 22:85
Overbank flooding 6:480
Overcoding, and queer anarchism 19:750
Overdeterminationism, mental causation and 15:176
Overdominance *see* Heterozygote advantage
Overeducation 7:127, 7:130, 15:596, 15:598
  notions of 10:868
Overfitting
  asymmetric effects on performance of 23:613
  in economic forecasting 24:339
Overflow, potential of 9:263
Overgeneral memory, and depression 2:285–286
Overidentification tests 23:617
Overimitation 11:622
Overindulgent parenting 3:385
Overjustification effect, in creativity training 5:115
Overlap, causal inference 3:257–258
Overlapping 24:58
Overlapping consensus 19:916–917
Overlapping generations model 18:568, 18:570
Overlapping waves theory 4:41
Overlooked jail populations 20:32–33
Overmedicalization 15:106, 17:82
Overpopulation 11:6–7
  LTG study 14:138
Overproduction, of artists 2:54, 2:56
Overreaching, property rights 24:531
Overrepresentations, in crime statistics 5:155–156
Overseas colonialism 19:849
Overseas contingency operations, war 25:379, 25:381
Overshadowing 15:156
  and context dependent memory 4:780
Oversharing 17:439–445
  boundaries and personality disorders 17:441–443
  definitions of 17:439–440
  genealogy of 17:439–440
  Narcissism 17:439, 17:442–443
  needy self 17:440–441
  performativity and slacking 17:443–444
  sociology of presence 17:444
  through social media and Internet 17:439–444
Oversight process 19:535
Oversimplification, property rights 24:530–531
Overskilling 15:598–600
Oversmoothed rules 19:31
Overspending, on urban megaprojects 24:842
Overt, undifferentiated shame (OUS) 21:875
Overt behavioral problems 2:470
Overt biases 17:108
  matched sampling 17:108–109
  model-based adjustments 17:109
Overt language policy 13:355
Overt linguistic categorization 3:812t
Overton, Leveller Richard, works, *An Arrow Against All Tyrants* 14:40
Overurbanization 6:316, 18:182, 24:819
Overweight, and addiction 1:97, 1:100–101
  see also Obesity
Ovid, Publius Ovidius, works, *Metamorphoses* 14:252, 16:155–156
Oviedo Convention 20:514

OWA *see* Ordered weighted averaging (OWA)
Owen, Richard 21:238
Owen, Robert 4:312, 21:178, 22:836
  works, *A New View of Society* 4:312
Owen, Roger, works, *The Middle East in World Economy: 1800–1914* 16:382–383
Owenites 4:311–312, 22:837–838
OWI *see* Office of War Information (OWI)
Own mental states, and theory of mind 15:541
Own-child method 5:736, 6:145
Ownership 8:772–775, 8:777, 19:192
  bailment 19:195
  of cultural content in Internet 12:589
  and international business 12:484–485, 12:487
  mass media 14:715–716
  property 19:177–183
  of research by-products, archiving 1:919–920
  security interests 21:400–405
  separation of, management 14:471–473
  stockholders 23:490–494
Ownership patterns 4:943–944
  history and politics explanation 4:943
  interest group explanation 4:944
  law matters 4:943–944
  private ordering explanation 4:944
Ownership-location-internalization (OLI) paradigm 12:485
Owning 11:233–234
OWS *see* Occupy Wall Street (OWS) movement
Oxfam International 16:916
Oxford Centre for Time Use Research (CTUR) 24:384
Oxford Movement 15:725
Oxford University Press (Oxford, UK and New York City) 7:587
  handbook series 7:588
OXT *see* Oxytocin (OXT)
OxyHb 16:389, 16:392–393
Oxymoron 13:697
Oxytocin (OXT) 14:514, 16:470–471, 16:650, 17:953
Ozpetek, Ferzan 20:870–871

P

P300 16:709–711, 16:709f–711f
P600 9:483
P* isolation index 21:411–412
P-12 Schools 6:569
PA *see* Performance appraisal (PA); Physical activity (PA)
Pa scale 15:567
PAA *see* Population Association of American (PAA)
PAC units *see* Personal action constructs (PAC units)
PACE *see* Promoting Aphasic's Communicative Effectiveness (PACE)
Pace Gallery 5:599–600
Pace of life 3:52, 3:54
  and time use 24:376
PACER *see* Public Access to Court Electronic Records (PACER)
Pachacamac in Peru 23:36
Pachakutik 23:47
Pachucki, Mark A., on food consumption 16:558
Pacific Islands
  archaeology of 17:446–450
  see also Oceania
Pacific region, social science academies in 22:637
Pacific Ring of Fire 6:480
Pacific settlement, evolution of language 8:373
Pacific social movements 3:140
Pacific War 3:139–140
Pacifism
  Mumford on 16:139
  and peace movements 17:643–647
  and peacemaking 17:637–638, 17:640–641
Packard, Vance, works, *The Hidden Persuaders* 1:215
Packing, apportionment 7:343
Pacta sunt servanda 12:509
PAD *see* Perception and attention deficit (PAD) model; Pleasure-Arousal-Dominance model (PAD)
Padala 12:519
Padilla v. Kentucky (2010) 18:229
Padoa's principle for proving independence of primitive concepts 2:340
Paedophilic disorder 17:951–952
PAF *see* Population attributable fraction (PAF)
PAG *see* Periaquaeductal gray (PAG)
Pagan areas 12:739
Paganism 9:807
Paget, James 11:506
Pahl, Ray 4:364, 11:237
PAHO *see* Pan American Health Organization (PAHO)
PAI *see* Personality Assessment Inventory (PAI); Private Attorney Involvement (PAI)
Pai Shan-ti Hui movement 20:369
PAID *see* Problem areas in diabetes survey (PAID)
Paid labors 25:670
Paid media advertising, for political campaigns 3:68
Paid work 24:387–389
  and time use 24:369–371, 24:379–380
Paige, Jeffrey 25:97
Pain 3:555–561
  and cancer 3:73–75
  classification 17:451–452
  definition of 3:555
  health psychology of 17:451–455
    biobehavioral perspective 17:454
    chronicity 17:452–455
    cognitive factors 17:453–454
    coping 17:453–454
    emotional factors 17:454
    future perspectives 17:455
    memory 17:454–455
    nociception 17:451–454
    operant conditioning 17:452
    respondent conditioning 17:452–453
    role of learning in chronic pain 17:452–453
  historical conceptualizations of 3:556–557
    biomedical model 3:556
    motivational model 3:557
    operant model 3:557
    psychogenic model 3:556
  and hysteria 11:508
  multidimensional conceptualization of 3:557–558
    cognitive behavioral model 3:557–558
    gate control theory 3:557
  orofacial 17:281
  of others, neural circuits associated with perception of 7:543–544
  phantom limb 18:1–2
    memories 18:1–2
    treatment 18:5
  and placebo effect 18:153–156
  rehabilitation 3:555, 3:559–560
  rheumatoid arthritis 20:655–656
  self-management 3:559–560
  terms 17:451
  treatments 3:558–560
    based on cognitive behavioral model 3:559–560, 3:559t
    based on operant model 3:559
    comprehensive approaches 3:558–560
    historical approaches 3:558
    multidisciplinary treatment 3:555–556, 3:559
  types of 3:556t
  utilitarian social thought 24:980–982
Pain cries, in neonates 16:479
Pain disorders, somatoform disorders 23:10–11
Paine, Robert 2:871
  works, *Patrons and Brokers in the East Arctic* 2:871
Paine, Thomas
  on republicanism 20:481
  works
    Common Sense 1:611
    The Rights of Man 7:664
Pain-prone families 17:453
Pair bonding 18:902
Pair-clustering model 6:527
Paired box homeotic gene 9 15:339
Paired comparison method 19:440–442
Paired comparison ranking 19:432
Paired courses 13:718
Paired-associate learning, discrete-state models 6:523–525
Paired-associate learning paradigm 6:524
Paired-comparison scaling, mathematical psychology 14:814
Pairwise-stable networks 16:574
Paisanos (countryman) 15:411, 15:413–414

Pakistan 12:729, 15:571, 20:375, 23:61
affordable private schools in 1:269–270
alcohol consumption in 1:517
archaeology of 23:50–57
and communalism 4:263–265
education in 7:196–197
fertility transition in 9:80–83
foundation of 16:280–281
historians of 11:80
relation between India and 20:376
sociocultural aspects of *see* South Asia; sociocultural aspects
Pakistanis, in Britain *see* British Asians
Pal review 21:252
Palace Museum in Beijing 2:21–22
Palace revolutions 20:638
Palau 17:158, 17:161
Palenque de San Basilio 14:593
Paleodemography, bioarchaeology 2:600
Paleodiets, archaeometry 1:907
Paleoeconomy 7:694
Paleoenvironment 7:694
Paleoenvironmental reconstruction and human impact 10:55–56
Paleoethnobotany 17:456–461
in archaeology 17:458–461
human evolution and emergence of modern behavior 17:458–459
Iceman and his environment 17:460
origins of food production 17:459–460
ritual and feasting 17:460–461
data 17:456–458
macroremains 17:456–457
phytoliths 17:457
pollen 17:457
starch 17:458
Ötzi on 17:460
Paleoethnological approach, to lithics 14:261–262
Paleoindian period
food production 9:303–304
North America during 17:16
Paleolithic Europe 8:269–270
Paleolithic populations 11:7
Paleolithic South Asia 23:50
Paleolithic/knapped Stone Age 14:260
Paleoneurology 2:827
brain, evolution of 2:831–833
Paleopathology 18:175
Paleopsychology 4:23–24
Palestine
Jewish Diaspora 12:803–806
Orientalism in 17:430–431
peace building and music in 25:845
Palestine Liberation Organization (PLO) 13:529
Palestinian community of Jesus 3:547
Palestinian Islamic Jihad (PIJ) 24:232
Palestinians, diaspora 6:353
Paley, William 9:390, 20:286
works, *Natural Theology or Evidences of the Existence and Attributes of the Deity* 12:175, 20:286
Palgrave, Robert H.I., works, *Dictionary of Political Economy* 7:585, 7:590
Pali language 2:888
Palinopsia 10:499
Paliperidone 19:201
Palliative care 5:867, 7:601–602
for the dying 8:281
personalized medicine 17:926–927
Palliative coping strategies, for teacher stress 24:74
Palliative function 12:412
Palliative sedation 7:601
Pallidotomy, for Parkinson disease 23:20
Palmares 14:593
Palmer, Craig T. 19:887
Palmié, Stephan 1:641
Palmonari, A., works, *Paradigmi delle rappresentazioni sociali* 13:902
Palo Alto Research Center (PARC) 3:6, 3:8
PALPA *see* Psycholinguistic Assessments of Language Processing in Aphasia (PALPA)
PALS *see* Peer-Assisted Learning Strategies (PALS)
Pálsson, Gisli, on embodiment 7:418
PalTalk chat 24:229
PaluTube 24:234
Paly, cognitive and brain development 8:394, 8:396–397
Palyi, Melchoir 25:462
Palynology 7:695
PAM *see* Positive assortative matching (PAM); Preparatory attentional and memory processes (PAM) theory
Pam Lynchner Sexual Offender Tracking and Identification Act 20:434
Pama-ngyngian language 23:889
Pampa 20:910
Pan, S., on marriage in China 21:735
Pan American Health Organization (PAHO) 4:401, 5:370–371
Panachage system 19:207
Pan-Africanism 1:277, 1:279, 1:287, 4:575, 6:690, 19:798
Panama 22:624
sociocultural aspects 1:621–625
Panama Pacific Exhibition (1915) 8:247
Panama Pacifico 6:833–834
Pan-Arab nationalism 16:777–778
Pan-Arabism 16:273
Panarchy 20:561–562, 20:565–566
PANAS *see* Positive and Negative Affect Schedule (PANAS)
Pan-Asia Population Genomics Initiative (PASNP 2.0) 11:319
Pan-Asia SNP consortium (PASNP 1.0) 11:319
Panasonic 6:836
Panca Sila (state ideology, Indonesia) 12:729
Pañcasila (five precepts) of Buddhism 20:333
Panckoucke, Charles-Joseph 15:11–12
Pancreas, functions of hormones 16:651
PANDAS *see* Pediatric autoimmune neuropsychiatric disorders associated with streptococcal infection (PANDAS)
Pandemic HIV-1 8:450
Pandering 15:592
Pan-diaspora 8:160
Pan-disability 6:462–463
Pandisciplinary approach to prevention 18:885
indicated preventive interventions 18:885
selective preventive interventions 18:885
universal preventive interventions 18:885
Panel attrition 14:354
Panel data, in skills mismatch 15:599, 23:202, 23:204
Panel fatigue 14:354
Panel retention 14:354
design features to maximize 14:354–356
locating and tracking respondents 14:355–356
retaining respondents between interview waves 14:356
Panel rotation error 16:923
Panel Socio-Economique "Liewen zu Letzeburg" (PSELL) 5:846
Panel studies
definition 5:845
European-wide 5:846–847
Panel Study of Belgian Households 5:846
Panel Study of Income Dynamics (PSID) 5:376, 5:697–698, 5:738, 5:744–745, 12:820, 17:462–463, 17:465–466, 24:390
Panel surveys 5:737, 17:462–467
antecedents of life events 17:466
changing populations 17:464–465
intergenerational transmission of economic status 17:465
poverty and welfare dynamics 17:465–466
problems with 17:463–464
unemployment and labor turnover 17:465–466
uses of 17:462–463
Pan-ethnic religious movements 8:128
Pan-European immigration policy development 15:385
Pangenesis 19:802
Panglossianism, functional explanation 9:522
Panhandling 15:592
Panic appraisal
and panic disorder/agoraphobia 17:477
Panic attacks 1:805, 17:468
experimentally-induced 1:809t
Panic disorders (PD) 1:805–806, 3:766, 3:768, 8:872, 17:468–481, 23:563
agoraphobia 17:469
antidepressants for 1:780
anxiety 17:469, 17:471
anxiety disorders 17:468–470
behavioral model 17:476–477
biological approaches 17:469–470
biological models 17:469
evidence regarding 17:469–470
biological model 17:477
clinical implications 17:479
cognitive model 17:476
comorbidity 17:476
definition 17:468–469, 17:474

development and avoidance behavior 17:475–476
diagnostic criteria 17:474
from DSM-IV and DSM-5 17:475
epidemiology 17:475
epidemiology, onset, and demographics 17:469
exposure therapies for 8:606, 8:608
and fear-potentiated startle 8:865
life events and stress 17:476
mechanisms of change 17:477
anxiety sensitivity 17:477
cognitive restructuring 17:478
exposure in vivo 17:478
guidelines 17:477–478
introceptive exposure 17:478
panic appraisal 17:477
pharmacological treatment 17:478
psychoeducation 17:478
relaxation and breathing training 17:478
medical conditions 17:476
neurotransmitter 17:469–470
panic appraisal 17:475
patients, fear-potentiated startle in 8:865
prevalence 17:475
psychological approaches 17:470
evidence regarding 17:470
psychological models 17:470
psychological treatments 19:383
psychotherapy for 19:374
treatment 17:470–471, 17:477–478
cognitive behavior therapy 17:470–471
combination treatments 17:470–471
medication 17:470–471
Paniclike symptoms 17:474–475
Pan-Indian civilization 2:76
Pan-Islamic movements 24:603
Pankov, Sergei 21:740
Panmictic circles, marriage within 18:589
Pannomion 2:527–528
Panofsky, Erwin 2:24
works, *Architecture gothique et art scolastique* 2:778
Panopticon 2:528–529, 6:43, 9:369–370
utilitarian social thought 24:982
Panorama 2:714–715
Panpsychism 18:107
Pan-Roman Catholicism, in South America 23:47–48
PANSS *see* Positive and Negative Syndrome Scale (PANSS)
Pansy craze 13:875–876
Pantagruel 7:584
Pantaleoni, Maffeo 17:503–504
Pantheismusstrei 7:660–661
Pantomime 10:111
and mirror neurons 15:584
Pantomime actions 1:877
Panunzio, Constantine, works, *A Student's Dictionary of Sociological Terms* 7:589
PAO *see* Partnership administrative organization (PAO)
Papacy 3:233–236
Papailias, Penelope, on history 1:748
Papal law 11:795
Papaver somniferum 17:236
Paper chase model 8:836
Papers in Regional Science 20:147
Paperwork Reduction Act 10:304
Papez circuit 8:869–870, 10:889
Papichristou v. City of Jacksonville 25:114
Papua New Guinea (PNG)
big man system 2:568–572
cargo cult 3:139
democracy 2:571
'stalled' economic development 2:571–572
Papuan languages 13:342
PAR *see* Participatory action research (PAR)
Par objects 13:68
Parabrachial nucleus (PBN) 16:896
interoception 12:614–615
Paradigm concept, as metaphor 15:291
Paradigm reason 13:776
Paradigm revolution 21:308
Paradigm shift
community-based health interventions 10:646, 10:651
qualitative developmental psychology 19:668, 19:674
Paradigmatic euthanasia conditions 8:281
Paradigmatic model, for urban systems 24:807–808
Paradigmatic principles 15:940
agency 15:940
life span development 15:940
linked lives 15:940
time and place 15:940
timing 15:940
Paradigmatic regulatory risk analysis 13:506–507
Paradigmatic systems 13:942
Paradigms 4:18–19, 12:648, 22:965–967, 24:137, 4:19t
diversity 4:740, 6:597
five-factor model (FFM) 2:559
Kuhn's concept 13:133–138, 17:89
legal reasoning 13:776–779, 13:781–782
in qualitative research 19:686, 19:688
in science and technology 24:141, 24:143
semiotics 21:593–594
witchcraft 25:607
Paradis, Michel 16:687
neurofunctional model of bilingual aphasia 2:575–576
Paradoxes
of assimilation 5:154
obesity 17:77
queer, histories of 19:752–753
secrecy 21:375
Paradoxical intervention, organizational development 6:298
Paraethnography 8:590–591
Para-government organizations 19:597
Paraguay 22:625
Parahippocampal place area (PPA) 11:388, 25:202
Parahippocampal region 10:886–887, 11:388, 10:887f
interconnectivity between deep and superficial layers 10:888
Parakkamabahu I 2:890
Para-language 14:229
Para-lawyers
definition of 17:482–483
empirical and theoretical issues regarding 17:483–485
deprofessionalization and 17:484
gendering of occupations and 17:484–485
law work structure and 17:483
legal services delivery effectiveness and 17:483–484
postprofessionalism and 17:484
profession and 17:484
future research and theory regarding 17:485
Paralinguistic interpretations of speaker intentions 23:714
Paralinguistics 23:714
Parallel analysis, in factor analysis 8:706
Parallel best–worst LBA model 5:905
Parallel coordinate plots 10:354, 10:354f
Parallel corpus 4:950–951
Parallel cousins 13:62
Parallel-distributed processing (PDP) 6:737, 23:295–296
of lexicons 13:927–928
Parallel electoral systems 19:206
Parallel institutions 16:417
Parallel measurements model (PMM) 19:426–427
correlation between parallel measurements 19:426
effect of test length 19:426–427
on criterion-related evidence of validity 19:427
generalized Spearman–Brown formula 19:427
reliability from single administration of test 19:427
Parallel mixed methods designs 15:620
Parallel process, in event history analysis 8:307–308
at level of individual 8:308–309
Parallel processing 12:79–80
streams 23:244–246
Parallel proportional profiles 3:243–244
Parallel relationships 17:486–492
flings 17:489–490
incidence and trends 17:486–488
kept mistress 17:488–489
'open' relationships 17:490–491
playfairs 17:489–490
sex parties 17:490–491
swinging 17:490–491
Parallel societies, creation of 16:29
Parallel tests 3:787–788
definition 3:787
identification 3:787
model 3:787t
testability 3:787
Parallelism 13:266–267, 23:891
Kraepelin on 13:128
mental causation and 15:174–175
Paramecium 7:833
Parameter estimation 23:616–617, 24:344

Parameter estimation (*continued*)
complex sampling designs and missing data 23:617
estimators for categorical observed variables 23:617
likelihood 14:117–118
maximum likelihood estimators 23:616
model-implied instrumental variable estimators 23:616–617
Parameter space 12:706
Parameter space partitioning (PSP) 4:475, 4:477–478
Parameter uncertainty 17:259–260
Bayesian designs 17:259–260
locally optimal designs 17:259
minimax and maximin designs 17:260
sequential designs 17:260
small effects 17:260
Parameterization 14:323–325
Parameters
estimation, order statistics 17:292–293
interpretation in regression 14:156
Parametric analysis 20:155–157
Parametric assumptions, causal inference 3:259
Parametric bootstrap simulation method 23:415
Parametric estimates 16:117
Parametric inference 23:419
survival analysis 23:773
Parametric item response theory 8:699–701
applications 8:701
Parametric learning algorithms 23:415
Parametric models 6:124
survival analysis 23:779–780
Paramount 9:157
Paramount Pictures 9:156, 9:158
Paranoia 2:753
Paranoid personality disorder 17:893
Paranoid-schizoid position 13:68
Paranormal phenomena 16:170
Paraphasia
misordering errors 18:51–52
Paraphilias 17:948–954
classifications 17:948–949
current classification 17:949
disorder 17:949
fetish foot 17:952
fetishistic disorder 17:952
gender 17:951
history 17:948
antiquity 17:948
nature of sexuality 17:949–950
neuroanatomy 17:953
neurochemistry 17:953
neurophysiology 17:953
nurture of sexuality 17:949–950
paedophilic disorder 17:951–952
prevalence of 17:951, 17:950t
procreation pleasure 17:952
sexual attitudes 17:950–951
history 17:950
religion 17:950–951
sexual behaviors 17:950–951
biology of 17:952–953
history 17:950
religion 17:950–951
shoes 17:952
Para-professions 17:483–484
Pararaton 11:82
Parareligious practice 20:408
Parascientific beliefs 4:164
Parasitic diseases, and children/infant mortality 12:23
Parasitic worms 8:457
Parasitism, Dark Triad 5:674
Paraskevidekatriaphobia 18:37–38
Parasomnias 22:75
Parasympathetic mediation, cardiac defense 23:542
Paraszt vármegye 9:468
Paratext 25:81
Parathyrin
neuroendocrinology 16:651
Parathyroid gland, functions of hormones 16:651
Paraventricular nucleus (PVN) 7:606, 11:476–477
PARC *see* Palo Alto Research Center (PARC)
Parens patriae suits 3:736
Parent language
and communication 8:732–733
Parent religious affiliation and youth 20:302–303
Parent to infant bonding *see* Neonatal bonding
Parent training programs 8:846–847
Parentage 3:477–478
Parental age at birth, influences in adulthood 6:755–756
Parental beaux arts participation, and education attainment 22:333
Parental beliefs, and play 18:223–224
Parental cancer
and childhood stress 23:572–573
and neonatal crying 16:475
Parental conversation styles 2:290
Parental death, early 6:756
Parental development, role in infant development 18:858
Parental education, effects on academic achievement 22:925–926
Parental emotion socialization 6:748
Parental employment
and children's educational progress 8:736
Parental engagement, in children's education 17:499
achievement 17:499–501
barriers 17:500
critique of research in field 17:500
future research 17:500–501
home–school links 17:500
literacy 17:500
numeracy 17:500
support and training 17:500
value 17:499–500
Parental ethnotheories 6:760, 19:246
Parental histories
and adult mortality 1:157
Parental investment 13:48–49
action of natural selection 9:28–29
among traditional humans 9:29–30
ecology and proximate mechanisms governing 9:31–33
fertility under modernizing conditions 9:32–33
human behavioral ecology 11:249–250
and life span development 14:108–110
in play during infancy, differences 18:219
and sibling order 21:924, 21:926
traditional foraging societies 9:31–32
Parental knowledge 7:438
Parental leave, family policy 8:802–803
Parental mediation, media effects on children 15:40
Parental mental health hypothesis 22:869
Parental personality 2:471
Parental practices 22:327
Parental reading, and education attainment 22:333
Parental rights 3:477–480, 3:482
and children's rights 3:489
Parental scaffolding of object play
differences, during toddlerhood 18:222
Parental socialization goals 2:141
Parental state of mind 12:231
Parental ties, and class 8:783–784
Parental values 3:452–454, 3:456, 21:42–46
Parent–child attachment 22:156–157
Parent–child conflict, and media 15:5–6
Parent–child interactions 19:245–246
impact of family socialization 22:867–868
social interaction approach 22:868
typological approach 22:867–868
and joint attention 12:845
Parent–child relationships 3:477–478, 22:156–158
and adolescent social competence 22:217
and emotional development of children 7:438–442
Parent–child structural solidarity 12:404
Parent–child tie, primacy of 8:782–783
Parent–child transmission 19:564–565
Parent–child–sibling interaction 12:845
Parenthood
and adoption 1:143
and fertility 9:3–4
and marriage 14:599–600
Parent–infant social-oriented 18:220
Parenting 2:458–459, 15:921, 15:923, 22:495–496
attachment 2:141
attitudes and beliefs across cultures 17:493
and culture 17:493–494
and culture contact 17:496–497
and family change 17:495–496
and socioeconomic context 17:494–495
autonomous-related self 17:495–497
and behavior during infancy and early childhood 6:14–16
behavior problems
associations between 2:472–473
development of 2:470
effects, direction of 2:473

interactive effect on 2:474
in siblings 2:475–476
child and adolescent psychiatry 3:385
authoritarian 3:385
authoritative 3:385
permissive 3:385
and childcare 3:396–397
and child's temperament/personality, interactions between 2:474
concept of 2:471–472
dimensions 2:471–472
control and discipline 22:157
democratic 8:734
differential 2:475–476
and divorce 14:522–523
effects on emotional development of children 7:438–444
protective factors 7:438
risk factors 7:438
styles 7:439
effects on prosocial behavior 19:245–246
altruism 19:246–248
authoritarian style 19:245
child development 19:246
in cross-cultural context 19:247–248
culture and 19:246–247
defining prosocial behavior 19:244–245
democratic parenting 19:245–246
empathy 19:244–245, 19:247–248
parent–child interaction 19:245–246
role 19:245–246
role modeling 19:245
socialization 19:244–246, 19:248
styles 19:245–248
evolutionary medicine 8:457–458
fathering *see* Fathering
good enough and bad enough 3:412
and health 11:269
and imitation by infants/young children 11:624
independence family model 17:494–497
interdependence family model 17:494–496
interparental relationship 22:158
LGBTQ
by individuals 21:757
legislation 21:756–757
poor, characteristics of 3:413
psychological functioning 22:157–158
psychological interdependence family model 17:495–497
queer *see* Queer parenting
and self-control 21:483
sensitivity and warmth 22:156–157
and social referencing 22:560
temperament 24:177
universalism 17:493
Parenting Early Intervention Programme (PEIP) 7:231
Parenting practices, effects on academic achievement 22:925–926
Parenting programs 7:230–231
Parenting/mating effort
human behavioral ecology 11:250
Parents 6:746–748
and adolescent leisure activities 13:831
arts participation 2:7–13
behaviors and attributes 5:377
changes in living arrangements of 8:841
and childhood behavior problems 2:478–479
cultural capital of 5:377
and family 8:809–812
implications, school environment for LGBTQ/sexual minority youth 21:89
interaction effect on infant development 18:855–856
involvement in schooling, and cultural capital 5:378–379
repartnering of 20:430–432
role in academic outcomes 1:40–41
role in gender development 9:743–744
expectations 9:743
role in nurturing productive giftedness 7:277
single 8:770
and socialization of children
instruction, advice giving, and consultation 22:868
as managers of children's opportunities 22:868
as source for children to learn gender schemas/stereotypes 21:681
and youth physical activity motivation 15:916–917
Parents and siblings, as role models and teachers 8:735
Parents behaviors 15:924–925
activity-specific parenting 15:925
general child-rearing climate and style 15:924
modeling 15:924–925
Parents' beliefs and values 15:923–924
educational settings 8:733–735
causal attributions regarding children's behaviors 8:734
cultural stereotypes 8:734
parents' expectations of child and perceptions of children's competencies 8:733
values and interests 8:734–735
Parents' school involvement
educational settings 8:735–736
Parent–teacher associations 8:755
Paretian distribution of earnings 2:53
Pareto, Vilfredo 1:967, 7:386, 17:503–505, 18:472–473, 20:663
economic actors, group of 17:505
economic equilibrium, analysis of 17:505
economic systems 17:504
elites 7:391–392
liberalism 14:25
life 17:503–504
logical actions 17:505
mathematics 17:504–505
professional career 17:503
works 17:504–505
Cours d' économie politique 17:504
Les systèmes socialistes 17:504
Manuale d'economia politica 17:504
Mind and Society (*Trattato Di Sociologia Generale*) 2:270, 17:505
Mythe vertuiste et la littérature immorale 17:505
Traité de sociologie générale 4:165
Pareto criterion 7:89
Pareto curve 2:52
Pareto distribution 6:571, 6:573, 6:578, 11:886
Pareto efficiency 5:47
Pareto improvement 13:573
Pareto improving 12:210
Pareto optimality 7:115, 14:538, 14:540, 14:623, 16:584–585, 25:497–500
Prisoner's Dilemma 18:931–934
Pareto principle 7:89–90
and utilitarianism 7:89–90
Pareto/NBD model 5:913
Pareto-optimal solutions 15:71
Parietal and frontal lobe maturity 11:390
Parietal areas, connections with 18:847
Parietal art 18:825–826
Parietal cerebral cortex 23:293–294
Parietal cortex 5:21–24, 5:27–28
and memory binding 15:145
and working memory 25:706–709
Parietal effect 23:32–33
Parietal eye fields (PEF) 5:27
Parietal lobe 13:371, 17:506–510, 25:184, 17:507f
gross anatomy and functional divisions 17:506
virtual spatial navigation 16:352–353
Parietal neurons 25:187–188
encoding of spatial information in multiple reference frames 17:705–706
Parietal reach region (PRR) 5:26–27
Paris, homosexuals in 9:638
Paris, Treaty of 1919 9:215
Paris Académie des Sciences 2:89
Paris Carnival 3:146
Paris Commune (1871) 22:834–835
Paris Convention for Protection of Industrial Property 12:263–264
Paris Peace Agreement 25:96
Paris Peace Conference 25:382
Paris Principles of 2007 3:439
Parish data, inverse projection 6:151
Parish registers 6:139–141
Parishes 20:410
Parisian Carnival 3:146–147
Parisian clubs in the Revolution of 1789 19:616
Parisian poststructuralism, and actor–network theory 1:81
Parité system 1:248, 1:253
Parity 11:6–7, 21:113
and family size preferences 8:805
and value of children 3:499, 3:501
Parity data, evaluation and correction of 6:129
Parity of participation, recognition politics 20:46–47
Parity progression ratios (PPRs) 6:112
Parity-specific fertility rate 6:112, 6:114

Park, Kwang-hyung, on careers and social networks 5:468
Park, Robert Ezra 4:362–363, 5:37, 8:138, 22:264, 22:266
works
Introduction to the Science of Society 7:561
Introduction to the Science of Sociology 2:176
Parke-Davis (pharmaceuticals) 17:585–586
Parker, Richard, works, *Beneath the Equator* 19:756
Parker, Theodore 20:697
Parkinsonism 19:466
negative symptoms 19:291
Parkinson's disease (PD) 6:443, 10:500–501, 16:675, 16:718, 16:847–848, 18:791–792, 23:252
effects on sign language use 21:937–938
placebo effect for 18:153–156
and prospective memory 19:276
Parks, for environmental protection 7:715–716
Parks, Rosa 3:693
Parliamentarian system 14:18–19
Parliamentarianism 18:864–868, 21:31
Parliamentary democracy 12:273–274, 17:517, 17:512t
democratization 17:511–514, 17:512t
development 17:511
functions and performance in 17:513–514
types and criteria 17:511–513
Parliamentary government 17:515
Parliamentary *Office d'évaluation des choix scientifiques et techniques* 24:127
Parliamentary systems 17:516, 20:163
executive government 8:497, 8:499–500
Parliamentary technology assessment (PTA) 24:125
Parliamentary technology assessment (PTA) bodies 22:873
Parliaments 17:515–519, 18:401
defined 17:515
English Revolution 7:649–653
evolution 17:516–518
female quotas in 8:888–892
functions and types 17:515–516
origins 17:516
political representation 18:441, 18:443
prospects 17:518–519
role of 17:516
types of 17:516
Parnas, David 15:718
Parnok, Sophia 21:742
Parochial conflicts 25:118
Parole 10:110
and punishment 19:656–659
in United States 20:31–32
Parole agencies 5:718
Parosmia 17:199t
Paroxetine 1:779, 1:781, 18:704, 22:85
for obsessive-compulsive disorder 17:125
Parr, Hester, works, *Mind and Body Spaces: Geographies of Illness, Impairment and Disability* 6:459
Parrington, Vernon, works, *Main Currents in American Thought* 1:619
Parry, J., works, *Money and the Morality of Exchange* 7:14
Parsimonious fit indices 8:688
Parsimonious latent variable models 13:396, 13:396f
Parsing 19:9
Parsonian concern for social order 18:11
Parsonian functionalism 13:855
Parsonian structural–functional strain theory 22:448
Parsonian theory 15:701–702
Parsons, Talcott 2:271, 2:506–507, 4:341, 5:36–38, 5:498–500, 6:601–602, 7:904, 14:382–385, 14:387–389, 14:415, 15:249–250, 15:708, 16:237, 17:2, 17:12, 17:520e524, 18:472, 21:146, 21:169, 21:885–886, 22:715, 22:968, 23:435, 25:463, 25:465
AGIL schema of 1:381, 17:522
on anomie 1:715–716, 1:720–721
on ethnomethodology 8:203–204
on capitalism 17:520–521
on functionalism 9:533
on ideas 5:578
on meaning 16:561
model of cognitive socialization 5:578
on neofunctionalism 17:524
on social structure 22:723
on structural functionalism 22:993
version of systems theory 8:483
voluntaristic action, theory of 1:76, 17:521–522
works
American Sociology 7:588–589
The Social System 1:382, 25:357
The Structure of Social Action 6:488, 6:703, 10:435–436
Parsonsian 21:885
PART *see* Program Assessment Rating Tool (PART)
Parti de l'intelligence 12:275
Partial breastfeeding 2:848t
Partial citizenship 10:310
Partial compliance 4:454
Partial control 4:828–829
Partial diglossia 6:418
Partial equilibrium 14:622–623
Partial equilibrium theory 7:914, 7:917
microeconomics 15:333, 15:335–336
Partial inclusion, and Internet 12:603
Partial least squares (PLS) 16:119
Partial mediation, mediation model 22:549
Partial model specification, in event history analysis 8:316
Partial order binary relation, defined 22:479
Partial orders 17:525–529
dimension 17:528
embeddability and completion of 17:528
extension of 17:527–528
related notions and diagrams 17:525–526
special classes 17:526–527
interval orders 17:526–527
lattices 17:527
semiorders 17:526
Partial registers 6:163
Partial reinforcement extinction effect (PREE) 10:398
Partial retrenchment, in employment and education 1:265–266
Partially linear models 21:601
Partially observable Markov decision processes 14:575
Partially ordered semigroup, defined 22:479
Partial-report paradigm 4:675
Participant observations 1:867–869, 17:530–534, 21:616–617, 22:545
advantages 17:530–531
economy 17:530–531
interpretive understanding 17:530
richness 17:530
validity 17:530
comparison with other field methods 17:532
in craft production 5:108
disadvantages 17:531
bias 17:531
generalizability 17:531
proofs 17:531
time 17:531
in empirical social research 7:561–562
ethical concerns 17:533–534
confidentiality 17:534
deception 17:533–534
informed consent 17:534
precision of description 17:534
and ethical questions 8:57
and ethnography 8:179, 8:182
in ethnography 8:186
fieldwork 9:149, 9:151–152
growth and development of 17:531–532
important works on 17:531–533
qualitative developmental psychology 19:672
qualitative health research 10:694
strategies of 17:532–533
access 17:533
developing relationships 17:533
exit strategies 17:533
field notes 17:533
socialization 17:533
use of term 17:530
Participant privacy, protection of 20:518
Participants
cross-cultural research, feedback from 7:932
experimental laboratories
beyond participant pools 8:553
payments 8:554
recruiting 8:552–554
undergraduate participant pools 8:553
joint action 12:839
peace movements 17:646
perspective of law and morality 13:482
Participants, Intervention, Comparison, Outcomes, and Study *see* PICOS (Participants, Intervention, Comparison, Outcomes, and Study)
Participation 1:31–32, 10:621

board 24:887
citizen 24:886–887
democratic 24:886
development geographies 6:292–293
disability 6:467–470
restriction 6:467–469
and liberation theology 14:33, 14:35–36
modernization 15:697–698
and self 21:442–443
skills training 22:21–22
and turn taking 16:816
in voting 25:306–307
Participation frameworks, and narratives 16:218, 16:220
Participation motivation, in youth physical activity 15:914–916, 15:918
Participation of citizens, in direct voting 25:295–296
Participation rights, of children 3:488
Participation–identification model 1:32, 1:35
Participationism 2:191
Participative leadership of teachers 24:60
Participative social action research and programming 24:755–756
Participatory action research (PAR) 17:535–539, 22:270–271, 23:324, 23:329
action research 17:536
applications 17:537–538
challenges and opportunities in undertaking 17:538
commitment 17:536–537
commitments 17:542
communities 17:535–538
critical research 17:535–538
critiques 17:537–538
definition 17:535–537, 17:540
education 17:537
epistemological orientation of 17:540–541
ethics 17:541–542
future of field 17:543
guiding principles and key features of 17:536–537
health 17:537
historical roots 17:541
learning by doing 17:536
methodological considerations 17:542
methods and outputs 17:537
ontological orientation of 17:540–541
participatory approaches 17:535–536
power 17:535–538
products 17:542–543
publics 17:542–543
purpose 17:542–543
qualitative developmental psychology 19:672
social change 17:536–538
in Social Research 17:540–544
Participatory and collaborative development and planning 18:184–185
Participatory approaches
informal sector 12:57–58
societal impact assessment 22:874
Participatory citizenship 3:652
Participatory decision-making, and workplace autonomy 25:719–720
Participatory democracy 6:91–93
Participatory Design Conference (PDC) 21:188
Participatory ergonomics 11:303
Participatory journalism 3:612–613
Participatory methodologies, street children 23:537
Participatory mode of representation (film) 6:621
Participatory plant breeding 1:472
Participatory research, street children 23:534–536
Participatory rights 3:688
Participatory rural appraisal 18:184
Participatory websites 4:508
aggregated user representations 4:508
proprietor content 4:508
user-generated content 4:508
Participatory-action researchers, in sociology 1:858–859
Particle filter 24:360
Particle physics 18:103–105
Particle size analysis, geoarchaeology 10:55
Particular rights 3:657
Particularism 5:12, 5:14–17, 7:312, 7:415
and anthropological comparison 4:411–412
historical 1:766
and identity movements 11:545–546
pattern variable 1:382
in Western societies 10:229
Particularistic egalitarianism 7:312
arguments 7:312
Particularistic identity movements 11:545
Partido Revolucionario Institutional 22:143
Partidul România Mare (Greater Romania Party) 18:411
Parties 10:313–315
element of negotiations 16:431
revolutionary 20:642–643, 20:645–646
Partij voor de Vrijheid (Party for Freedom, PvdV) 18:412, 18:415
Partisan activists, issue evolution 12:766–767
Partisan change, standard model of 12:765
criticisms 12:765
Partisan dealignment 17:565
Partisan elites, issue evolution 12:766–767
Partisan expert opinions 8:585–586
Partisan polyvalence 12:854
Partisanship
consequences 17:564–565
development 17:562–563
journalism 12:855
levels 17:563–564
Partiti 14:423
Partition of India 12:433
Partitioning 23:900–901
Partitions, statistical clustering 23:392–393
Partito Nazionale Fascista (PNF) 16:257
Partner bond formation, in heterosexual sexuality 10:801–811
Partnering, for physical activity 18:95
Partners
in education, parents and teachers as 17:499
in educational planning 7:262, 7:264
fertility of 9:20
Partners in Health (PIH) 1:225, 4:401
Partners of cancer patients
psychological distress in 3:75
psychological distress of cancer in interventions 3:76
Partnership administrative organization (PAO) 23:519–520
Partnership axiom 4:874
Partnership ethics, street children 23:537
Partnership recognition, and LGBT aging 13:865
Partnerships 6:896–897, 13:630, 10:297t
cohabitation 4:118–120
fertility 9:20
governance as 10:293–294, 10:296, 10:294f
in law firms 13:578–582
repartnering 20:430–433
and sleep 22:60
in teacher education 24:53–55
Partnerships formation, in Western societies 17:545–549
Partnerships in Western societies 17:545–552
dissolution 17:549–550
divorce trends 17:549–550
remarriage and repartnering 17:550
union dissolution 17:550, 17:550f
explanations for changes in partnership behavior 17:550–551
family 17:550–551
formation 17:545–549
childbearing within cohabitation 17:547–549, 17:549f
cohabitation 17:545
duration of cohabiting unions 17:547–548, 17:548t
incidence of cohabitation 17:545–547, 17:546t, 17:547f
living apart together 17:548–549
marriage trends 17:545
people practicing cohabitation 17:548
type of first relationship 17:546–547
union formation 17:545, 17:547–548, 17:550–551
Part-of-speech (POS) tags 4:951
Parts of speech 17:553–561
Part-time proletarians 25:697
Part–whole relation 11:750
Party convergence, and manifestos 18:419
Party elites 8:889
Party families, and populism 18:610–611
Party finance 18:393–397
Party identification (PI) 17:562–566
dealignment 17:565–566
elections 17:562–566
origin and evolution of concept 17:562
partisanship
consequences 17:564–565
development 17:562–563
levels 17:563–564

Party identification (PI) (*continued*)
political participation 17:562, 17:565
political socialization 17:562, 17:566
research controversies 17:565–566
voting 17:562–564
for voting choice 25:302–303, 25:306–307
Party lists 7:351
mixed systems 7:351
panachage 7:351
pure 7:351
voter-choice from 7:351
Party of hope 21:892–893
Party of memory 21:892–893
Party organization, of United States Congress 4:650–651
Party quotas 8:888
Party systems 17:567–569, 18:398–399, 18:402
classifying 17:567–568
Sartori on 17:567–568
and structures of competition 17:568
Party vanguard 3:741
Party-state 20:851–853
Parvo cells 17:129
PAs *see* Personal assistants (PAs); Physician assistants (PAs)
Pasa, Ahmet Cevdet, works, *History of the Ottoman Empire* 11:65
Pasai, Malay, works, *Hikayat Raja-raja Pasai* 11:82–83
Pascal, Blaise, works, *Pensées* 14:808–809, 20:769
Pasch, M., axiomatic theories 2:339
Pascual-Leone, Juan, theory of constructive operators 16:503–504
Pashukanis, Yevgeny 22:844
Pasivos 9:628–629
PASNP 1.0 *see* Pan-Asia SNP consortium (PASNP 1.0)
PASNP 2.0 *see* Pan-Asia Population Genomics Initiative (PASNP 2.0)
Paso de la Amada 15:258
PASS *see* Planning, Attention Simultaneous, and Successive (PASS) theory
Passeron, Jean-Claude 22:332
works
Les Héritiers 2:779
Reproduction in Education, Society and Culture 7:291
Passion
and love 14:377–380
and personality 17:908
Passion and motivation 17:570–576
defining and measuring 17:570–572
domain-specific models of passion 17:571–572
entrepreneurial passion 17:571–572
as part of triangular theory of love 17:571
domain-unspecific models of passion 17:570–571
com.pass model 17:571
dual model of passion 17:570–571
harmonious versus obsessive passion 17:570–571
as intrinsic, self-relevant absorption 17:571
future research, directions for 17:573–575
problems regarding conceptualization of passion 17:573–574
problems related to the aim to explain relevant criteria 17:574–575
state versus trait concepts of passion 17:574
general degree versus specific types of 17:573
relations to other constructs 17:572–573
harmonious and obsessive passion, correlates of 17:573
outcomes 17:573
predictors of passion 17:572–573
theoretical mappings of overlapping constructs 17:572
Passion Scale 13:835
Passionate problem 1:982
Passive articulators 18:66–67
Passive bystanders, role in mass killings/genocide 14:704
Passive control, in old age 1:182
Passive euthanasia 8:281
Passive genetic model of divorce 14:520
Passive labor market policy 13:184–186
Passive leisure, flow in 9:263
Passive motion paradigm (PMP) 15:960
as general approach to action generation 15:960–962
networks 15:960–961
and simulation theory of covert actions 15:961–962
Passive obedience 11:412
Passive pluralism 3:681–682
Passive processing of language sounds 8:395
Passive sensor systems 20:419
Passive tolerance 18:240
Passive use impacts, in cost–benefit analysis 5:49
Passivity 24:119
Passivity hypothesis 25:170
Passivization 16:2, 25:2
Passover meal 12:872
Passover narrative 20:331
Past 24:409
constructed by present 4:201–202
is in present 4:200–201
reinvention of 17:35
Past climate 3:865
Past tense verb generation 13:927–928
Past victimization/suffering, and group violence 14:704
Pasteur, Louis 8:579, 21:265
Pastoral development 17:581
Pastoral landscapes 20:797
Pastoral nomadism 16:838–839
Pastoral Psychology 11:793
Pastoral systems, in archaic states 23:372
Pastoralism 5:108
in anthropology 17:577–583
borders and conflicts 17:582
in Central Asia 3:318
definition of 17:577
and environment 17:579–580
future of 17:582–583
global land grabbing 17:581–582
and governmentality 10:311
herders *versus* animal needs 17:579
livestocks, functions of 17:577–578
pastoralists 17:579
perspectives on 17:579–580
politics of 17:580–581
rangeland enclosures 17:581–582
within regional economies 17:580
society *versus* continuum 17:579
theories 17:578
in West Africa 1:300–301
Pastoralists 11:202, 17:579
in East Africa 6:774–775
PAT *see* Preventing Alcohol Trauma (PAT)
Patent Act 1952 17:586–587
Patent analysis
research and development (R&D) in organizations 20:500
Patent law 12:268
territorial thinking in 12:268–269
Patent medicine 17:240
Patent statistics as indicators of innovation and industrial dynamics 11:835
Patenting, in biological sciences 17:584–585
American Civil Liberties Union *versus* Myriad Genetics 17:588–590
BRCA1 and 2 genes 17:588
Canada, Europe, and the United States 17:588
genes 17:587–588
Intellectual property 17:584, 17:587
products of nature 17:585–588, 17:590
1980s and 1990s 17:587
twentieth century 17:585–587
Patenting comparisons
country-level 11:835–837
firm-level 11:838–840
industry-level 11:837–838
Patenting system, research 20:524
Patents 3:31, 12:262–263, 12:266–267, 13:525–526, 24:144
biotechnology 12:270–271
drug 10:618
on genes 9:1003–1004
genomics 10:39–40
and innovation 12:161–162, 12:164–167
research and development (R&D) in organizations 20:497–498
Pater, Walter 20:425
works, *The Renaissance* 20:425
Paternal influences on child development 8:841–842
Paternal involvement, in religion 20:319
Paternalism
legal 25:76
women offenders 25:616
Paternalist 18:204
Paternity case, outliers in 23:387
Paternity leave, family policy 8:802–803
Path
in centrality network 16:532

in social network analysis 22:462
Path analysis 14:326
Sewall's method 25:759–760
Path dependency 7:55, 7:58–59, 8:479–480, 24:127
industrialization 11:870–871
institutions 12:194
organizational 17:385–389
public policy 19:573
welfare retrenchment 25:521
Path integration, in navigation 23:161–163
Path models, in educational sociology 7:292
Pathogen stress, and prosocial behavior 19:241
Pathogenesis, and personality 17:794
Pathogenetic models, of obsessive-compulsive disorder 17:122–124
Pathogenetic models of psychogenic amnesia 1:653–654
Pathogens, exposure, and infectious diseases 12:32
Pathological aging 1:437–438
versus successful aging 23:668–669
Pathological behaviors 17:859
Pathological gambling (PG) 17:592–597
behavioral treatment 17:594–595
cognitive and cognitive-behavioral therapies 17:594–595
gamblers anonymous 17:594
motivational enhancement and brief interventions 17:595
classification 17:592
epidemiology 17:592–593
etiology
demographic correlates 17:593
impulsivity and neurobiological factors 17:593–594
psychiatric comorbidity 17:593
medication treatment 17:596
treatments 17:595–596
antidepressants 17:595–596
glutamate 17:596
mood stabilizers 17:596
opioid antagonists 17:595
Pathological gaming 4:499–500
Pathological intoxication, alcohol 1:510
Pathological protein β-amyloid 1:591
Pathological responses, social work stress/trauma 18:707–708
Pathological tau 1:591–592, 1:594, 1:592f
Pathologization, medical, of homosexuality in China 21:735–737
Pathology
contributions of Sherrington to 21:881
and psychosocial work 19:479–482
Pathology of clients, and psychoanalysis 22:766
Pathophysiological mechanisms, of hysteria 11:508–509
Pathophysiology
neuroimaging 14:441, 14:443–444
of psychopathy 19:452
decision-making impairment 19:452
empathy impairment 19:452
moral judgment impairment 19:452
of schizophrenia 21:16
of transient global amnesia 24:566–567
Pathoplastic model 17:917–918
Pathos 23:525
Paths, crime 20:774–775
Pathways, moral development 15:782
Pathways theories, and feminist criminology 8:915
Patient, Intervention, Comparison and Outcome (PICO)
evidence-based medicine 8:328–329
Patient adherence
to cardiac rehabilitation 20:190
diabetes treatment 6:338–340
in respiratory disorders 20:577–578
Patient and public involvement (PPI) 4:400
Patient behavior in health care system 10:677–678
Patient care
and gender 9:696–697
implications, coronary heart disease 4:921
Patient education programs 10:684
Patient experiences in psychotherapy 19:518–519
Patient Health Questionnaire (PHQ) 6:343
Patient Health Questionnaire 4 (PHQ-4) 6:204
Patient populations
and retroactive interference 20:615
virtual spatial navigation in 16:352
Patient preparation, electroencephalography 7:354
Patient Protection and Affordable Care Act 2010 (PPACA) 7:602, 10:594, 10:606, 16:907–908, 16:912, 18:736–737, 20:170–171
Patient Register Data Service 5:736
Patient rights, and bioethics 13:558, 13:560–561, 13:563
Patient Self-Determination Act 1990 7:598
Patient stratification 2:623
Patient studies, and source monitoring 23:30–31
frontal lobes 23:31
medial temporal lobes 23:31
Patient variables, psychotherapy 19:373
Patient-centered care 18:35
and doctor–patient interaction 6:613–614, 6:617–618
Patient-centered therapy, experiential psychotherapy 8:517–518
Patient-centric behavioral care management 14:464–465
Patient/disease advocacy, biomedicine 2:666
Patient–doctor relation, ethics 2:615–616
Patient-driven competition 10:598–599
Patient-focus (PF) construction 23:892
Patient-led research (PLR), for therapeutic effectiveness 24:283
Patient-oriented research, evidence-based medicine 8:327–328
Patient–provider communication *see* Doctor–patient interaction
Patient-provider interaction, health psychology 10:677–678
Patients, delivering treatments unknown to 18:156
adherence to health care regimes 17:598–602
in combined pharmacotherapy and psychotherapy 19:497
adherence 19:496–497
information discrepancy 19:497
selection of therapist with similar ethnic/gender background 19:496
involvement in medical student training 7:187
rights of 7:599
role, in psychiatry 22:975, 22:977
role of cerebellum in eyeblink conditioning 2:95–96
seeking complementary and alternative medicine 10:565–566
Patient–therapist interaction, in psychotherapy 19:515–519
Patient–therapist relationship
concepts, history of 24:287–288
definition of 24:287
emphases 24:288–289
I–thou dialogs 8:518
research on 24:289
methodological issues 24:289
short-term psychodynamic therapy 19:346–347
Patriarchal power 13:257
Patriarchal terrorism 12:685–686
Patriarchalist absolutism 1:10
Patriarchy 5:161, 10:819
and counterculture 5:85
and fear of crime 8:855–856
honor killing 11:188–189
and leisure 13:854
and motherhood 15:881–883
and religiosity of women 25:635
Patrician Party 18:398
Patrilateral extended household 11:229
Patrilocal households 3:312
Patrimoine 5:414
Patrimonial rule 23:369
Patrimonialism, in Europe 8:275–276
Patrimonio 5:414
Patriot Act 24:117
Patriotic absolutism 1:10
Patriotism 16:285–286, 16:312
and war 25:363–369
Patristics era 3:238–239
Patrizzi, Francesco, on objectivity 17:89–90
Patron brokerage 17:604
Patron state 5:464
Patronage 5:14–15, 5:17–18, 14:425, 14:565
Patronage networks 23:369
Patronato Law (1880) (Cuba) 22:41
Patron–client networks 23:369
Patron–client relationships 9:426–427, 9:429, 17:603–606
and brokers 2:870–872
comparative analysis 17:605

Patron–client relationships (*continued*)
context of 17:604–605
and inequality 11:896
institutional viability, debates on 17:605–606
social exchange 17:604
strategies 17:604
study of 17:603–604
terminology 17:603
Patron–client structures, organized crime as 5:172
Patronyms 16:184
Patteggiamento 18:229–230
Pattern and practice cases, in employment discrimination 6:531–533
Pattern detection 23:95–97
Pattern formulation in SODS 21:529
Pattern of missing data 15:602–603
Pattern recognition 23:239
Pattern thinking 11:784–785
Pattern variables, AGIL schema 1:382–387
Pattern-matching, in realism 20:24
Pattern-matching technique 8:321–322
Patterns
of activity, cognition 4:95–96
ethnography 8:189
migration history 15:446
Pattern-separated memory, and hippocampal amnesia 10:881, 10:883
Patterson, C.J., on LGBTQ individual parenting 21:757
Patwardhan, Anand, works, *Father, Son and Holy War* 6:622–623
Paucals 17:44–45
Pauci potentes 1:966
Paucity of psychiatric resources 6:42, 6:46
Paul, G.A. 10:536
Paul, Hermann 2:705–706
psychological subject and psychological predicate 12:97
works
Prinzipien 14:216
Paul (New Testament) 3:545–547
Paul Bourke Award for Early Career Research 22:639
Paul II, John 3:235–236
works, *Laborem exercens* (*On Human Work*) 20:260
Paul VI, Pope 3:235
Pauperism 25:475–476, 25:479
Paupers 18:729
Pavlov, Ivan Petrovich 16:679–680, 17:607–610, 18:58–59, 19:413
on conditioned reflexes 17:608–609
on digestive system 17:608
early life of 17:607–608
influence on modern psychology 17:609–610
theory of classical conditioning 7:238–240
neutral auditory stimulus, experiments on 17:608
works
Lectures on the Work of the Digestive Glands 17:608
Reflexes of the Brain 17:607
Pavlov Pouch 17:608
Pavlovian classical conditioning 8:863–864, 15:156, 17:608–609
Pavlovian fear conditioning 8:870
Pavlovian theory-based model of temperament 2:562
Pax Americana 15:701
Pax Britannica 25:364–367
egalitarian impetus 25:366
Franco-Prussian militarism 25:365
industry 25:366
pupils into soldiers 25:365–366
World War I 25:366–367
Pax Priísta: 1940s–1960s 15:329
Pax Romana 17:638
Pay, for teachers in United States 24:78–79
Pay equity 7:927, 9:731
Pay equity movement 13:594
Pay-for-performance 1:349, 17:611–613
Payment, and money 15:746–748
Payment for care 9:804
Payment for Environmental Services (PES) 6:858
Payoff function 5:907–908
Payoff vector 4:868
Payoffs 4:869
PBC *see* Perceived behavioral control (PBC)
PBIS *see* positive behavioral interventions and supports (PBIS)
PBL *see* Problem-based learning (PBL)
PBN *see* Parabrachial nucleus (PBN)
PBS *see* Political belief system (PBS)
PBs *see* Personal budgets (PBs)
PBS Parents 7:252
P/BV *see* Price-to-book-value (P/BV) ratios, and stock market returns
PCA *see* Principal component analysis (PCA)
PCB *see* Psychological contract breach (PCB)
PCBD *see* Persistent complex bereavement disorder (PCBD)
PCBs *see* Polychlorinated biphenyls (PCBs)
PCE *see* Spanish Communist Party (PCE)
PCF *see* French Communist Party (PCF)
PCI *see* Italian Communist Party (PCI); Politicization of collective identity (PCI)
PCK *see* Pedagogical content knowledge (PCK)
PCOS *see* Polycystic ovary syndrome (PCOS)
PCR *see* Polymerase chain reaction (PCR)
PCSK6 gene 10:505, 10:513
PCT *see* Protocerebral calycal tract (PCT); Public choice theory (PCT)
PCV *see* Pervasive community violence (PCV)
PD *see* Panic disorders (PD); Parkinson's disease (PD); Prisoner's dilemma (PD); Professional development (PD)
Pd scale 15:567
PDAs *see* Personal Digital Assistants (PDAs)
PDC *see* Participatory Design Conference (PDC)
PDE *see* Pure direct effect (PDE)
PDI *see* Postmaterialism index (PDI); Principle of Domain Integrity (PDI)
PDMS-2 *see* Peabody Developmental Motor Scales, Second Edition (PDMS-2)
PDP *see* Parallel distributed processing (PDP)
PDS *see* Woodworth Personal Data Sheet (PDS)
PDSA *see* Plan, do, study, act (PDSA) improvement cycle
PE *see* Person–environment (PE) fit model; Physical education (PE)
PE1 *see* Pedunculus extrinsic neuron 1 (PE1) neuron
Peabody Developmental Motor Scales, Second Edition (PDMS-2) 16:725
Peabody Individual Achievement Test – Revised (PIAT-R) 3:795–796
Peabody Picture Vocabulary Test (PPVT) 3:796, 5:356
Peabody Picture Vocabulary Test – Revised (PPVT-R) 25:580
Peabody Social Museum 22:698
Peace 17:618–623
Addams on 1:91
anthropological aspects 17:624–630
anthropology as peace founder 17:624
apologists for war 17:627–628
colonialism 17:626
conflict 17:625–626, 17:628–629
conflict resolution 17:625, 17:628–629
dispute 17:625
documentation 17:624–625
feud 17:625–626
human nature 17:624–629
mediation 17:626
natural history 17:625–626
nonkilling 17:626–627
obstacles 17:628–629
peace studies 17:624, 17:626–629
positive peace 17:629
reexamining tribal warfare 17:626
Semai 17:625
structural violence 17:629
transformation from killing to nonkilling 17:626
and appeasement 1:826, 1:828
and balance of power 2:345–347
culture of peace 25:843, 25:846
definition 17:618–619
definition of 25:842–843
and democracy 25:359–361
global peace 25:744
Myrdal on 16:162–163
and related concepts 17:619–620
unconciliated traumas from the past 17:618
unresolved conflicts at present 17:618
and war 4:607, 4:610, 4:612–613, 25:374
Peace agreements, in conflict mediation 4:630
Peace building 8:99, 25:843, 25:847
in India 25:844–845
in Israel 25:845
in Kenya 25:845–846
in Mozambique 25:846

in Palestine 25:845
in Senegal 25:846–847
and youth and music 25:843–847
Peace dividends
conflict, education, and youth 4:634
defense economics 6:2
Peace movements 17:643–647
under Cold War 17:644
disarmament 17:643–645
early efforts 17:643–644
internationalism 17:643–644, 17:646
military intervention 17:645
new movements 17:644–645
new social movements 17:644, 17:646
Nuclear Weapons Freeze Campaign 17:645
outcomes 17:646–647
pacifism 17:643–647
participants 17:646
in post-Cold War era 17:645–646
postmaterialism 17:644–645
Peace of Augsburg 20:102–103
Peace of God 17:639
Peace of Westphalia (1648) 3:675
Peace Operations 2010 Reform Agenda 17:633
Peace Pledge Union 17:638
Peace politics 17:618
Peace processes 17:648–653
political and social 17:648
Peace promotion
advocacy for social justice and human rights 17:658–659
International Criminal Court (ICC) 17:658
peace education 17:659
postconflict 17:655–659
mental health and psychosocial support 17:655–656
peacebuilding 17:656–657
psychology of 17:654–659
social justice and prevention 17:657
transitional justice 17:657–658
Peace settlements 8:99
Peace support operations 17:631, 17:633
Peace systems 4:615, 4:617
Peacebuilding
with consensus 4:592
postconflict 17:656–657
Peace-building operations 17:631
Peace-enforcement operations 17:631
Peaceful abolitionism 22:40
Peaceful coexistence 22:847–848
Peaceful endings of modern slavery 22:41
Peaceful Families Project 20:256
Peacekeeping 17:631–636
cosmopolitan 17:633
differences in composition and conduct of 17:632
enlarged 17:633
established principles and practices of 17:632
historical development 17:634–635
operations 17:635–636
theories and concepts 17:631–633
UN and non-UN 17:633–634
in West Africa 1:303
Peacekeeping efforts 23:351
Peacemaking 17:637–642
in Ancient world 17:638
contemporary 17:641–642
internationalism 17:641
Just War theory 17:637, 17:639–641
in medieval world 17:638–639
in modern European history 17:639–640
in non-European history 17:640
pacifism 17:637–638, 17:640–641
reconciliation 17:637–638
secular and religious approaches to 17:637–638
theory underlying 17:637
treaties 17:639–641
twentieth century emergence 17:640–641
Peacemaking criminology 22:386–388
Peacemaking operations 17:631–632
The Peach Blossom Spring 14:253
Peak oil 14:142, 15:554–560
depletion claims 15:557–559
Hubbert model 15:556–559, 15:556f–557f
Limits to Growth model 15:557
nonrenewable resources 15:554, 15:556
perception and reactions to 15:557–559
resource depletion 15:554–556
Seneca Effect 15:557
sustainability 15:555
systematic view of 15:556–557
Peak travel 2:306–307
Peano, Guiseppe 23:596
Pearl, Raymond 8:220–222
Pearl River Delta 20:136–140
Pearson, Egon Sharpe 11:494–495, 16:824–826, 17:662, 23:425–426
Pearson, G., works, *The Deviant Imagination* 1:792
Pearson, Karl 5:8, 5:354, 8:240–241, 9:222, 9:571, 11:494, 14:811, 16:825, 17:660–663, 21:264
Biometrika 17:660–661
on biometry 17:660–661
on chi-squared test 17:662
contributions to statistics 17:661
on eugenics 17:660–661
Fisher and 17:660–662
Galton and 17:660–662
on law of succession 17:662
life of 17:660–661
on method of moments and inverse probability 17:662
on multivariate analysis 17:662
on multivariate statistics 17:661–662
on normal distribution 17:661–662
on partial causality 23:425
on statistics 23:426
on univariate statistics 17:661
use of Bayesian techniques 17:662
works
Ethic of Freethought 17:660
Grammar of Science 16:824, 17:660
and Yule 17:661–662
Pearson (education company) 1:270
Pearson Affordable Learning Fund 1:270
Pearson chi-squared test 23:397
Pearson curves 17:661
Pearson Foundations 10:249
Pearson statistic 16:131
Pearson's method of estimation 15:298
Peasant agriculture 17:664
big capital countering 17:667
ecological capital 17:664
embedding farming multiple job holding 17:666
enlarging ranks 17:667
entrepreneurial farming retreatment 17:666–667
labor 17:664
low external input agriculture 17:665
multifunctionality 17:665–666
Peasant elimination 1:459
Peasant movements 17:677
Peasant revolt 17:671
Peasant societies
inequality 11:894, 11:896
studies, anthropological 1:769
Peasant utopia 1:459
Peasant wars 17:677
Peasantries 11:205, 15:769–770
and anthropology 1:754–756
in 21st century 17:664–668
Peasants 1:461, 14:423, 14:425, 15:326, 15:328, 17:669–675, 23:632
after World War II 17:670–671
agrarian history 17:669–670
agricultural history 17:669–670
agricultural technology in 17:671
in anthropology 17:676
cash crop 17:677
closed corporate community 17:677
community studies 17:676–677
cultures 17:676–677
economics 17:678
farm ecology 17:677–678
folk societies 17:676–677
informal economy 17:678
labor-intensive agricultural systems 17:677
moral economy 17:677
off-farm labor 17:678
peasant movements 17:677
peasant wars 17:677
politics 17:677
profit motive 17:678
rebellion 17:677
remittances 17:678
repeasantization 17:678
resistance 17:677
rural social classes 17:676–678
subsistence orientation 17:677
via campesina 17:678
community 17:674
discussion about 17:672–674
economic networks 17:672
and farmers 17:669
horticulture 11:202, 11:204–206
in Mesopotamia 23:372
rural history network 17:672
in 1970s 17:671

Peasants (*continued*)
in 1980s 17:671
in 1990s 17:671–672
society 17:669
states and civilizations 23:372
study of European rural societies, network for 17:672
terminology 17:669–670
before World War II 17:670
PECDA *see* Professional engagement and career development aspirations (PECDA)
Pecheny, Mario, on Latin American homosexuals 13:858–859
Peck, Jamie 13:164
Peculiarities, white-collar workers 25:566
Pecuniary loss 24:429
Pedagogic differences, in lifelong learning 4:114
Pedagogical content knowledge (PCK) 12:67, 24:56, 24:75–76
Pedagogical leadership 4:801
Pedagogical optimism 7:182
Pedagogical pessimism 7:182
Pedagogical skills 24:56–57
Pedagogy 10:740
and coconstructivism 3:915–917
and educational neuroscience 16:729, 16:733–734
and information and communication technology 12:67–68
and modern human cognition 11:254–257
psychiatric education 7:185–193
public 6:552–553
Pediatric autoimmune neuropsychiatric disorders associated with streptococcal infection (PANDAS) 17:122
Pediatric overweight, psychotherapy for 19:377
Pedigree, and genealogy 9:806–810
Pedigree method 9:932, 13:45
Pedigree structures 8:720, 8:720f
Pedology, in Soviet Union 20:826
Pedophilia 21:660–661
Pedra Branca (Brazil) 6:833–834
PEDro *see* Physiotherapy Evidence Database (PEDro)
Peduncular hallucinosis 10:499
Pedunculus extrinsic neuron 1 (PE1) neuron 15:156–157
Peel paradigm of policing 18:246
Peer aggression 5:639t
Peer cultures, (self)socialization in 19:671
Peer delinquency 22:412–413
Peer effects 4:528–529
Peer experience, role across cultures 9:432–433
Peer groups
and adolescent friendship 9:432, 9:434–437
and self-control 21:482
support 21:483–484
Peer instruction 7:23
Peer interaction, and coconstructivism 3:913–914
Peer justice 17:383
Peer learning 10:416–417, 10:419–420
Peer networks, and identity 6:499–500
Peer play, during toddlerhood 18:221
Peer relationships 2:459–460, 3:482
and adolescent friendship 9:433, 9:438
and adolescent health 1:125
and adolescent social competence 22:217–218, 22:220
Peer review 13:506, 17:685–689, 20:530–531
accountability 17:685–686, 17:688
changing contexts of 17:687–688
disciplines 17:685–688
epistemology 17:686, 17:688
ethos of science 17:685–686, 17:688
institutionalization of 17:686–687
ex ante modes 17:687
rise of academic publishing 17:686–687
role of royal academies 17:686
journals 12:863–864
critiques 12:863
new developments 12:863–864
types 12:863
knowledge production 17:685
as organized skepticism 17:685
quality control 17:685–688
and quality control 17:680–684
background 17:680–681
Baltimore affair 17:682
blogosphere 17:683–684
extended and democratized community 17:683–684
extended peer community 17:683
fraud 17:682
misconduct 17:682
new forms of science 17:680, 17:682–683
pathologies of science in society 17:682–683
postnormal science 17:683–684
reproducibility 17:680, 17:682
retractions 17:682
science as endless frontier 17:681–682
self-correction 17:680–681
trust 17:681–682
role in research funding 9:620
science policy 17:685, 17:687
and university press 19:624
Peer Review Evaluation Process (PREP) 6:897
Peer support, in disasters 22:757–758
Peer tutoring 17:228
cooperative learning 4:884
Peer-Assisted Learning Strategies (PALS) 4:884
Peer-facilitated support groups 1:503
Peer-learning systems 7:122
Peer-led creativity 21:104
Peer-reports, of emotional intelligence 7:445–446
Peers 4:139–140, 4:143
and adolescent leisure activities 13:831–832
and body image 2:736–737, 2:739
and child care 3:392–393
and digital natives 22:919–920
and gender development 9:745
interaction effect on infant development 18:856
and social networks 9:6
time of children with 9:445
valued, characteristics of 19:793–794
and youth physical activity motivation 15:918–919
Peer-to-peer (P2P) connections 8:487
Peer-to-peer (P2P) networks 17:218
Peer-tutoring task 15:283
PEF *see* Parietal eye fields (PEF)
PEIP *see* Parenting Early Intervention Programme (PEIP)
Peirce, Charles Sanders 6:504, 7:568, 8:567, 10:111, 12:244–245, 18:795–796, 18:808, 19:472
pragmatism 18:803–804
Pejorative connotation 6:536
Peloponnesian War 17:638
Penal field 19:655
and punishment 19:655–656
Penal law 2:528, 16:357
Penal politics 19:636
Penal variation, and punishment 19:657–659
Penalized least squares, nonparametric regression 16:873–874, 16:873f
Penalties 13:574–575
and punishment 19:655, 19:657, 19:659–660
Penalty enhancement 10:541
Penan 20:308
Pendelton Act 1883 3:696
Penduum-swing model of Muslim societies 9:673–674
Penfield, Wilder 16:686, 16:686f
Peng, Fred 16:679
Penguin 1:270
Penicillin 10:609
Penis envy 9:461
Penn world table (PWT) 7:47–48
Pennsylvania Association for Retarded Citizens (PARC) v. Pennsylvania 23:220
Pennsylvanian prisons 18:938–939
Pen's parade 11:900
Pensées 14:808–809
Pension policy 17:690–696
administration 17:691–692
aging 17:690, 17:694–695
analysis of 17:691–692
benefits 17:691–692
Bismarckian–Beveridge classification system 17:692–693
coverage 17:691
defined contribution (DC) scheme 17:692
expanding institutional variation 17:693
financing 17:691
future 17:692–693
income inequality 17:693, 17:695

institutionalism 17:694
modernization 17:693
origins and development 17:690–691
outside OECD 17:694
pay-as-you-go (PAYGO) system 17:691
PAYG system 17:695
politics 17:693–694
social policy 17:692–694
typologies of 17:692–693
welfare state 17:692–694
Pension privatization 18:695
Pension systems 1:323–326
Pensions
and health 20:598–600
old age work participation 17:182–185, 17:187–188
public and private 20:603–605
PENTBOM investigation 13:532
Pentecostal Christianity 8:285–288, 10:216
Pentecostal congregations 21:174
Pentecostal movement 3:540
Pentecostal Protestantism 3:236
Pentecostalism 10:231
spirit possession 23:282–283
Pentecostals 3:542
witchcraft 25:608
Penturbia 18:666–667
Penty, Arthur, works, *Old Worlds for New: A Study of the Post-Industrial State* 18:664
Penumbra, and stroke 13:330
People 22:887–891
cosmopolitanism 22:890–891
democratic imaginaries 22:887–888
as demos 22:887–888
and environment, in international social problems 12:545
as ethnos 22:887–888, 22:890
forming impressions of 22:188–189
human rights 22:890–891
national democracy, contradictions of 22:890–891
national imaginaries 22:887–888
national variations 22:889–890
in organizations *see* Organizational behavior
photographs of 18:81–82
and power 12:544–545
republicanism 20:480
structural relations 22:888–889
of West Africa 1:300
People for the Ethical Treatment of Animals (PETA) 1:705
People of color, whiteness/white supremacy 25:569–573
People of color summit 7:757
People v. O.J. Simpson 9:957
People *versus* policy debate 16:453
People with disabilities
conceptualization of disability in law and policy 13:497–499
disability in law and policy, defining 13:500
disability inclusion in legal reform assistance 13:500–501
disability rights 13:501
and Eastern European social work 22:796–798
human rights model of disability 13:498–499
international human rights law standards on the rights of 13:499
law reforms and 13:501
legal responses to disability 13:498
medical view of disability 13:498
models of disability 13:498–499
persons with disabilities as victims of crime 13:499
social model of disability 13:498
UN disability convention structure and content 13:499–500
world population of 13:497
People-centered care, health literacy 10:653
Peoplemeters 2:213
People-processing institutions, ethnomethodology 8:205
People's assessor 22:845
The People's Choice 25:302
People's Choice (1948 American voting study) 4:291
Peoples Communication Charter 15:56
People's Front, democratic party 4:303
People's Global Action (PGA) 1:684–685
Global Action Days 1:684
People's history
agency, role of 11:109
future of 11:112
intellectual influences 11:109–110
origins 11:108–109
political sensibilities 11:109–110
People's Mujahedin of Iran (MEK) 16:807
People's participation 4:402
People's Republic of Bangladesh *see* Bangladesh
People's Republic of China (PRC) *see* China
People's Temple 16:807
People's Wireless (Germany) 19:878
Peopling of technologies 21:182
Pepper, David, works, *The Geography of Peace and War* 25:374
Peptidergic sleep regulation 22:84f
Per capita alcohol consumption 1:517–518
Per capita income, and environmental degradation, bell-shaped relationship between 13:145–146, 13:145f
Per capita nucleolus 4:872
Perceived barriers, health behavior 10:577
Perceived behavioral control (PBC) 4:820, 10:585
health behavior 10:578–580
Perceived benefits, health behavior 10:577
Perceived competence, and youth physical activity motivation 15:915–916
Perceived control 21:471–472, 21:516
versus powerlessness
and age 1:548
and gender 1:547
neighborhood disadvantage and disorder 1:548
objective power 1:546
self-concept of 21:471–472
self-estrangement 1:547
socioeconomic status 1:546–547
and work–family interactions 1:548
Perceived control behavior 4:813–814
Perceived discrimination, consequences of 8:115–116
Perceived endorsement of a product 15:32
Perceived instrumentality, FTP and FO 9:563–564
Perceived intelligence 13:649
Perceived powerlessness *see* Powerlessness
Perceived risk (PRA) 17:809–810
drug use/abuse 6:687–688
and fear of crime 8:854–855
as stable trait 17:809
Perceived severity, health behavior 10:577, 10:579–580
Perceived situational control 4:820
Perceived susceptibility, health behavior 10:577–579
Perceived value (PV) 23:511–512
Perceiver readiness
as determinants of social identity salience 21:456–457
self-categorization theory 21:456–457
and social categorization 22:187–188
Perceiving 20:50–51
Percent-variance-accounted-for 4:475
Perception and attention deficit (PAD) model 10:502
Perception and expression of mental illness, cross-cultural 5:572
Perception of action space 17:703–707
extrapersonal space 17:703–704
impairments associated with spatial perception 17:704–705
neural aspects 17:705–707
frontal lobe mechanisms of spatial representation 17:706–707
parietal neurons encode spatial information in multiple reference frames 17:705–706
peripersonal space 17:703–704
as represented in visual frames of reference 17:704
egocentric and allocentric frames 17:704
visual processing streams 17:704
spatial cognition 17:703
Perception–action model 7:550, 10:128
Perception-imagery equivalence hypothesis 25:163–164
Perceptions 23:111–112, 23:114, 23:235, 23:292
Batterie Imagination et Perception (BIP) 25:166–167
and change blindness 3:344, 3:347
concept learning 4:538–540
cultural psychology 5:564–565
direct scaling of 21:604–606
and embodied cognition 7:420–422
and ergonomics 11:299
of journal authors 12:865
like me theory 15:132
mathematical psychology 14:813

Perceptions (*continued*)
and mental imagery 15:224–225
multisensory integration in 16:95
neurophysiology of 13:124–125
object 2:201
of objects 16:628–631
person 2:201–202
of phenomenology 18:20
role in aesthetic experience 1:230
social 2:202–204
Stevens on 23:453, 23:455
and unconscious mental activity 24:723
visual 25:163–165
Perceptron 4:84–85
Perceptron-type learning algorithms 23:415
Percepts, materiality approach 4:28
Perceptual and cognitive holism 12:79
Perceptual consciousness 10:738
Perceptual deficits after brain damage 25:203–204
Perceptual deterrence 6:250–253
Perceptual differentiation 4:111
Perceptual grouping 16:635–636
Perceptual learning 10:127–134, 25:179, 25:189
affordances 10:129
animal–environment reciprocity 10:128
cognition 10:133
comparative psychology 10:127–128
in development 10:128–129
developmental psychology 10:127–128, 10:132
distinctive features 10:130
ecological approach 10:127–128
events, invariants, and multimodal information 10:129–130
hallmarks on human behavior 10:132–133
agency 10:132
flexibility 10:132–133
prospectivity 10:132
infancy 10:127, 10:129–133
perception–action reciprocity 10:128
perceptual information 10:128
process 10:130–132
differentiation and increasing specificity 10:130–131
exploratory activity 10:131–132
systems approach 10:127
Perceptual memory 13:678–679
brain regions relevant to 13:679–682
Perceptual narrowing, face 8:655
Perceptual objects 16:634
Perceptual organization, computational auditory scene analysis 2:223
Perceptual processing 13:399
Perceptual Reasoning Index (PRI) 12:284, 16:723
Perceptual recognition 23:240f
Perceptual representation system (PRS) 1:399, 11:710–712, 25:175
proactive interference in 19:2
Perceptual skill learning, neural substrates of 16:848
Perceptual span, eye movements 8:632–633
Perceptual studies, of language learning impairment 13:346
Perceptual symbols theory 5:579, 7:422
Perceptually contemporaneous offenses 8:855
Percutaneous transluminal coronary angioplasty 20:190
Perdomo, Rogelio Perez 13:515
Pereleshin, Valery, works, *Ariel* 21:742
Perestroika 2:922, 18:470, 18:691, 20:633, 20:845, 25:838–839
Perezhivanie, as unit of analysis 25:333
Perfect competition 7:114–115, 14:622
Perfect competition theory 20:728
Perfect equilibria 9:587
in game theory 9:586–587, 9:590
Perfect information 12:103
Perfect market assumptions 25:274
Perfection of Wisdom (Prajna-Paramita) 2:888, 25:632
Perfectionism, and procrastination 19:75
Perfects (grammar) 24:212
Perforant pathway 10:888–889
Performance 10:110, 17:334–336, 25:539
(pattern variable) 1:382
and abusive supervision 23:701–702
aesthetics and creativity 17:710
anthropological aspects 17:708–715
-based assessment, of cultural intelligence 5:435
-based emotional intelligence theories 8:673
carnival 3:146–148
central issues 17:713–714
concept of 17:708–709
cultural 5:618, 5:620
cultural geography 5:411–412
and cultural politics 17:711–712
and diversification 6:587–588
dramaturgy 6:646–649
of drivers in transportation 24:625
economic transformation 7:77–78
and equity in education 7:919, 7:921
of executive government 8:500–501
folk 9:282
genetic effects in 9:921–922
goals 15:888
of groups 10:411, 10:425
and interpretation 12:643
job 12:834–835, 12:837
measurement and management 16:771–772
motor control models 15:957–964
narrative 16:207
nonprofit organizations 16:917
in public sector services 16:770
queer 19:755–756
reasoning 5:937–938
ritual 20:719
ritual and media 17:711
science as 21:226–227
spirit possession 23:283
sports 23:299–302
sustainable development 23:818
of sustainable practices 23:834–836
theater and social action 17:710–711
tourism 17:712–713
Turner's legacy in theater 17:709–710
and unintended consequences 17:421
Performance anxiety, cause for sexual inadequacy 21:689–690
Performance appraisal (PA)
changes in focus of literature 17:716–717
definition 17:716
effectiveness criteria for 17:716
and evaluation 17:716–721
goal 17:716
leadership effect on 17:719
multisource appraisal 17:717–718
perceived justice of ratees 17:718–719
performance feedback 17:717
ratees' reactions 17:716
social context of 17:718
Performance assessments 3:821
Performance difficulty 4:676
Performance evaluation
autonomous agents 2:311
changes in focus of literature 17:716–717
effectiveness criteria for 17:716
goal 17:716
leadership effect on 17:719
multisource appraisal 17:717–718
and organizational citizenship behavior 17:319
perceived justice of ratees 17:718–719
performance feedback 17:717
social context of 17:718
Performance feedback 17:717
Performance intelligence quotient (PIQ) 12:284
Performance management 16:771–774, 17:716
Performance measurement 16:771–774, 24:99–102
automated performance measurement 24:100–101
behaviorally anchored ratings scales (BARS) 24:100
concept mapping 24:101–102
event-based measurement scales 24:100
rating scales 24:100
Performance measures of wisdom 25:592–593
advantages 25:593
disadvantages 25:593
Performance processes, self-regulation of 21:545
Performance tests, EI 7:445
Performance theory 9:733–734, 17:713
and gender 9:784–785
Performance-goal orientation, and epistemological beliefs 7:889–890
Performative elaboration 21:617
Performative mode of representation (film) 6:621, 6:623
Performative social science, qualitative developmental psychology 19:674
Performativities 13:258, 17:305
in cultural geography 5:411–412
and discourse 6:504, 6:506–507

of economics 5:525, 5:540
of gender 3:57–58, 9:701–702
political performativities in Central Asia 3:317
quantification 19:701–702
queer 19:755–756
and slacking 17:443–444
Performativity theory, economic sociology 7:68–69
Performing arts 2:4
Perfusion fMRI 14:440
Periaquaeductal gray (PAG) 23:866
interoception 12:614–615
Pericean logic 21:592–593
Perinatal injury 11:390
Perinatal psychiatry 17:722–726
adverse impact of maternal psychosocial vulnerability 17:724
breast-feeding 17:725
childbirth 17:723
delivery 17:722–725
disorders 17:722–723
anxiety disorders 17:722
depression 17:722–723
Edinburgh Postnatal Depression Scale 17:723
impact of mental illness on child development and outcome 17:724
maltreatment and poor mothering by mentally ill mothers 17:724
maternal bonding disorder 17:723–724
mother–infant interaction 17:723–724
psychotropic medications 17:725
puerperal psychosis 17:723
screening for postnatal depression 17:723
suicide and infanticide 17:723
treatment and service plan 17:724–725
pharmacological management 17:724–725
psychosocial interventions 17:724
Period analysis, in demography 17:727–734
Period effect, in crossover designs 5:349–350
Period incidence 7:825
Period life table 14:91
Period mortality 15:869–870
Period mortality crossover 15:849
Period of child development 25:324–325, 25:324f
Period of formal learning activity 25:325
Period of sociodramatic play 25:325
Period parity progression ratios (PPPR) 6:114
Period prevalence 7:825
Period total fertility rate (PTFR) 9:10–12
Periodgram 24:336
Periodic function, for time series cycles 24:331
Periodic insanity 14:487
Periodic limb movement disorder 22:74
Periodical indexes 5:809
Periodicity, population censuses 3:291
Periodization 4:762, 24:409–414
as historiological tool 24:412
Middle Ages 15:348
Reformation and confessionalization 20:98–100
types of terms 24:412
universality of the tripartite periodization 24:413
Periodontitis 17:281
Peripersonal space 17:703–704
Peripheral agraphias 6:735
Peripheral alexias 6:733
Peripheral dysgraphia 1:454
Peripheral language processes 6:733
Peripheral level, of working memory 21:902–903
working memory specific 21:903–904
Peripheral membership 17:532
Peripheral nationalism, and class 16:292
Peripheral neonationalism 16:292
Peripheral nervous system (PNS) 16:714
Peripheral process decisions 15:32
Peripheral route 15:31–32
Peripheral societies, world systems theory 25:748
Peripheral vestibular disorders 25:69
Peripheries 3:123–124, 3:126, 6:404, 13:96, 25:341–342
and center relationships 3:308–311
ethnic nationalism in 16:277–278
in Europe 8:276
in modern world system 16:241–242
working classes 25:695–696
Perirhinal cortex (PRC), and hippocampal amnesia 10:877–879, 10:881–884
Peri-urban 23:663–665
Perlocutionary speech act 14:172
Perlongher, Néstor 13:862
Permaculture 1:486, 1:566–567
Permanence, social work 3:474–475
Permanency movement 1:143–144
Permanent arms economy 25:379
Permanent associates 13:630
Permanent campaign in politics 14:716
Permanent contracts 1:27
Permanent Court of Arbitration 17:640
Permanent Court of International Justice 12:509–510
Permanent supportive housing (PSH) 11:161
Permanent visas 11:635–636
Permeability
and tyranny 24:713
at work 8:761
Permission culture 13:525
Permissive law 13:9
Permissive parenting 3:385
and socialization of children 22:867
Permissive parenting styles 19:248, 24:196
Permissive parents 7:439, 22:157
Permitted legal acts 23:23
Permutability, in learning spaces 14:588
Permutation matrix 19:80
Permutation tests 19:431–432
nonparametric statistics 16:885–887
Permutations 17:297
Perón, Eva 23:47
Perpetration 25:111–112
Perpetrators *see* Torture
youth as 25:393
Perpetual peace, Kantian politics 13:9
Perpetual subsidies 18:369
Perpetuity 1:9
Per-protocol analysis, compliance 4:446, 4:450
Perry, R.B., works, *Realms of Value* 8:76
Perry, Richard J., on anarchism 1:681–682
Perry Preschool Project 8:564
Persecution
halting of 14:705
and racism 19:852, 19:855
Roma people 20:748–749
Persecutory anxiety 13:68
Perseverance 10:522–523
and grit 10:397
and self-efficacy 21:504–505, 21:507–508
Perseveration 6:720–721, 18:56–58, 18:60
Persian
culture 11:65
historiography 11:65
language 11:65
Persian Wars
and warfare 25:414
Persiflage 11:417
Persistence 5:137–144, 17:794
adolescence-limited offender 5:139
chronic offender 5:141, 5:143
and crime 1:312–313
diffusion 6:414
dual taxonomy theory 5:138
empirical findings 5:141–143
criminal offending trajectories 5:141–142
marriage, employment, and other adult life events 5:142
prospective prediction problem 5:143
in expectancy-value-cost model 8:503–504
group-based trajectory models 5:140–143
growth curve models 5:140–141
informal social control theory 5:138
intermittent offender 5:139–140
key emerging issues 5:143–144
key theories 5:138–139
kindling 13:33
life-course persistent offender 5:138–139, 5:141–142
methodological issues 5:139–141
definition problem 5:139–140
methods of analysis 5:140–141
operationalization complications 5:140
in migration 15:458–459
and optimism/pessimism 17:263, 17:266–267
organizational path dependence 17:385, 17:388–389
of racial residential segregation 20:541–542
research on 10:397–398
rise of interest 5:137–138
self-control theory 5:138
unemployment 24:741–742
Persistence-oriented marker 3:833

Persistent complex bereavement disorder (PCBD) 2:534
Persistent fear 18:37, 18:42
Persistent somatoform pain disorder 23:11t
Persistent thoughts
see Obsessive–compulsive disorder (OCD)
Person (identity) 17:744
everyday understanding of sameness of persons 17:746–747
Person deixis 6:52
Person factors, of aggression 1:377–378
Person identity nodes (PINs) 8:663
Person in Environment (PIE) approach 18:708–709
PERSON model 17:741
Person of tomorrow 20:739
Person perception 2:201–202, 17:753
accuracy of 17:739–741
Personal action constructs (PAC units) 17:758
Personal aggression 25:722–723
Personal area of harmony 10:531–532
Personal assistants (PAs) 6:462–465
Personal authority 16:835
Personal autonomy 2:314, 13:290
and English wh-imperative 13:290
Personal behavior, and climate change 3:844–846
Personal budgets (PBs) 4:346
for older people
banking services 21:501
benefits 21:500
making easy 21:501–502
skills maximization 21:501
stipulations 21:500–501
using 21:501
Personal care, time use 24:389–390
eating 24:390
grooming 24:390
sleep 24:390
Personal causality 2:202
Personal characteristics
fear of crime 8:855
and internal migration 12:444–445
Personal choice 20:841
Personal computers 5:78
Personal construct psychology 21:232
Personal construct theory 4:724–726, 4:729
Personal contact/relationships in associations 25:257–258
Personal control 1:545, 15:942
versus powerlessness 1:545
Personal coping resources 4:902
Personal data
and internet data 12:594–595
ownership, autonomous agents 2:312
Personal Data Sheet, development 3:879
Personal development
gifted education 7:162–163
social work 24:755
Personal Digital Assistants (PDAs) 3:294, 10:590–591
Personal distress, and prosocial responding 7:317–319
Personal documents
in academic disciplines 6:494–495
construction methodology 6:495–496
in discipline-building 6:492
Personal epistemology 7:887–892
Personal equation, Lazarsfeld on 13:637–639
Personal factors, of disability 6:469–470
Personal fitness, kin selection 13:26, 13:28–29, 13:28f
Personal genomics 1:761–762, 9:991–992
Personal goals
as building blocks for life-management 21:558–559
for life management 14:87–88
Personal growth 17:233–234
Personal history/biography, personality assessment 17:850
Personal hygiene, and mortality rate 15:869
Personal identity 4:152, 9:434–435, 15:31, 16:215, 21:450, 22:882
comparison of theories 17:743–744
composite positions 17:743–744
exclusive positions 17:743
renouncing a univocal identity concept 17:744
consistency 17:744–746, 17:748
constraints 17:745–746, 17:748
criteria of 17:742, 17:744, 17:747–748
difference 17:745, 17:747–748
in genomics 10:34
narrative competence 17:745–748
narrative identity 17:746
numerical identity 17:742–744, 17:746, 17:748
philosophical aspects 17:742–749
philosophical problem 21:449–450
relational theories 17:743
identity does not matter 17:743
psychological continuity 17:743
sense of identity 17:746–748
everyday understanding of sameness of persons 17:746–747
facets of 17:747
implications 17:747–748
material content/formal competence 17:745–746
mechanisms for stabilizing identity 17:746
modern theories 17:745–746
postmodern approaches 17:745
stability of value commitments across a lifetime 17:747
theoretical-observing *versus* practical stance 17:746
theories 17:744–746
social identity 17:744–745
substance-based theories 17:742–743
immaterialist theories 17:742–743
materialist theories 17:743
thought experiments 17:743, 17:746
uniqueness 17:746–748
uses of basic terms 17:744
person 17:744
same 17:744
Personal idiosyncrasies 11:645
Personal influence 22:346
Personal information 23:727–728
of research participants 20:514
Personal injuries
of forensic psychologist 9:353
tort claims and 24:430
torts in law and 24:429–430, 24:433
Personal injury lawyers 24:429
Personal integrity rights, repression of 18:446–447, 18:450–451, 18:447f
Personal intelligence 17:750
ability measurement 17:752–753
ability scales 17:753–754
clinical symptoms 17:754–755
and competencies 17:750–756
competencies measurement 17:752–753
competencies of 17:754
contemporary view 17:750–751
hot/felt intelligences 17:750–751
intelligence theories of today 17:750
personality and personal intelligence 17:751
measurement 17:752–753
and mental abilities 17:755
normal personality traits 17:754–755
precursors 17:751–752
relation to self-evaluations 17:754–755
and socioemotional interests 17:754–755
and socioemotional traits 17:754–755
theoretical bases 17:751
theory of 17:752
work behavior 17:754–755
Personal intentions
joint action in meshing 12:840
we-intentions 12:840–841
Personal interest 12:378
phases of 12:378–379, 12:379t
in vocational psychology 25:240
Personal interview 12:671–672
Personal is political 17:248–249
Personal law 13:607
in India 4:264
Personal memory 1:881
Personal names 17:225
Personal naming 16:183, 16:185
Personal ownership 22:846
Personal probability, concept and uniqueness of 20:928
Personal projects 17:757
defined 17:757
and motivation 15:937
social ecology of 17:757–758, 17:757f
Personal projects analysis (PPA)
hierarchical statistical procedures 17:760–761
intimacy-related projects of lesbian, gay, and bisexual individuals 17:764
measurement criteria 17:758, 17:758t
methodology 17:758–763
psychometric qualities of 17:763
Personal qualitative characteristics, new unified theory 22:879
Personal quantitative characteristics, new unified theory 22:879

Personal references 17:852
Personal relationships 15:31
and friendship 9:425–427
Personal resources 21:135
and parenting 22:869–870
Personal Responsibility and Work Opportunity Reconciliation Act 1996 (PRWORA) 18:725, 18:736, 25:478
Personal risk perceptions 10:703–706
absolute 10:704
comparative 10:704–706
optimistic/pessimistic biases 10:705t
types 10:705t
Personal secrecy 21:377
Personal security 3:687
Personal significance of money 15:735–736
Personal skills 3:11–12
Personal spaces, and long-term heterosexual relationships 20:916
Personal strength, definition 23:572
Personal thefts 19:160–161
Personal tolerance 18:240–241
Personal transfers 12:519
Personal union 4:571
Personal values 25:27, 25:29–31
Personal variability, in risk representations 20:711–712
Personal well-being 24:623
Personalism 17:736
Personalist philosophical currents 17:735
Personality 1:245, 3:571–572, 4:895, 4:905, 7:218, 7:445–446, 16:493, 17:751, 17:766–773, 17:809–820, 17:857, 17:914, 21:365, 22:162–167, 22:247–248, 22:881, 24:173–174, 25:538–539
abilities 17:780–781
and accidents 12:130–132
and acculturation 1:71
across development 9:1010
and agreeableness 1:463–466
alternative personality models 17:789–790
approach-avoidance theories 17:859–860
personality traits as probabilistic constants 17:859–860
assessment 9:1010, 17:849–856
and faking 17:834–839
assessment, forced-choice 17:840–848
authoritarian *see* authoritarian personality
and authoritarianism 2:265
biological models of 17:857–862
biological substrates of 9:1011
biosocial perspective 17:858
brain–cognition–personality relations, aging of 2:796
changes during adolescence 17:863–869
Big-Five dimensions and lower-level constructs 17:865–867
correlates 17:867
cross-cultural variations 17:867–868
early views on 17:863
maturation 17:865, 17:868
measurement of personality, change, and stability 17:864–865
shared model of personality 17:863–864
changes in openness/intellect over time 17:771–772
children, interactions between parenting and 2:474
cognitive models of 17:870–875
cognitive theory of 4:895–896
common assumptions 17:766
competing arguments 17:766–767
and conflict 4:598
conscientiousness *see* Conscientiousness
and coping 4:895–901
counseling psychology 5:69–71
creative personality 17:770–773
and crime 17:774–779
culture and 19:359–360, 19:362
Dark Triad *see* Dark Triad of personality
definition 9:1009–1010, 17:906
descriptive correlational studies 4:896–897
development, systems theories 17:879–883
developmental contextualism 17:880–881
developmental process 17:882
developmental psychology 17:879, 17:882
developmental science 17:882
dynamic systems theory 17:880
future directions 17:882
holistic, developmental, systems-oriented perspective 17:881–882
holistic–interactionistic approach 17:881
life course 17:879, 17:881–882
life span 17:879–881
open system 17:880
probabilistic epigenesis 17:879–880
differences 17:803
differences and development: genetic and environmental contributions 17:884–890
and differential aging 6:381–382
disorders in DSM 17:915–916
and domain-specific creativity 17:772–773
and economics 17:780–784
and educational outcomes 17:787–791
educational performance 17:787
entrepreneurial 7:677–678
and essentialism 8:15
and ethnocentrism/prejudice 8:171–172
evolutionary models of 17:899–905
balancing selection 17:900–902
facultative calibration 17:903
fitness 17:899–903
frequency-dependent selection 17:901
heritability 17:899–903
heterozygote advantage 17:902
individual differences 17:899, 17:903
mutation–selection balance 17:902–903
selective neutrality 17:900
sex-dependent selection 17:901
spatiotemporal heterogeneity 17:900–901
experiential theory of 8:517
extraversion *see* Extraversion
factors, in offender rehabilitation 20:199
five-factor model *see* Five-factor model (FFM) of personality
functional perspective 17:858–859
gender differences in 9:750–756
and genetic factors 9:920
genetic influences on correlates of 9:1011
genetics of 9:1010–1011, 17:861
statistical/behavioral genetics 17:861
grit 10:397
and health 17:792–796
behavioral medicine 17:793
coping with illness and stress 17:794–795
emotions 17:793
health behavior 17:794
history 17:792–793
other explanations 17:795
pathogenesis 17:794
personality theories 17:793–794
temperament 17:792, 17:794–795
and Honesty-Humility 11:169
and hypertension 11:453–455
hysterical 11:508
influence on behavior 11:879
influences on religiousness 17:805–806
intelligence 17:770–772
and interpersonal attraction 12:626–627
and isms 22:165
and job satisfaction 12:835–837
and leisure 13:834
and marriage 17:797–800
cumulative continuity 17:798–799
interactional continuity 17:799
similarity and complementarity 17:798
traits 17:797–798
and mental health 15:189
Minnesota Multiphasic Personality Inventory 15:566–568
models 17:802
models of 17:787
models of disordered personality 17:915
models of normal-range personality 17:914–915
molecular genetics of 9:1012
motives and values 17:780–781
narratives 17:780–781
neuroimaging 17:861–862
neuroticism 16:743–748
nonpersonality disorder psychopathology 17:916–917
openness/intellect and five-factor model of 17:771
and organizational citizenship behavior 17:318
parental 2:471
personality disorder psychopathology 17:916–917
personality trait models 17:914
personality traits 17:915–916
personnel selection 17:937–938
and preferences 17:781–783
and procrastination 19:73–76
profiles of citizens and political leaders 18:429
psychology 17:770, 17:772

Personality (*continued*)
and psychopathology link 17:917–918
complication model 17:918
pathoplastic model 17:917–918
scar model 17:918
spectrum models 17:918
vulnerability model 17:917–918
ratings 17:787
reinforcement sensitivity theory 17:860–861
personality traits 17:860–861
RST 17:860–861
and religion 17:801–808
religion interaction 17:807
religious influences on 17:805
research history 17:770–771
and resilience 4:898–899
as self-regulatory system 17:767–768
self-theory of 20:738
and sibling order 21:924–925
social cognitive approach to 17:921
source processes 17:857–858
surface traits 17:857–858
and symptom perception 23:869
temperament 24:180–183
and temporal lobe 24:209
terrorist 24:239
theory and psychopathology 17:914–919
traits 17:770–773, 17:780–781, 17:802–803
assessment limitations 17:817
behavior 17:814–818
measures, self-knowledge assessed by 17:815, 17:815f
outcomes 17:816–819
reason for assessing 17:814–815
social cognitive views of self 17:817–818, 17:818f
stability of judgments 17:815–816
–state relations 17:816–817, 17:817f
traits model of 17:920
agreeableness 17:922–923
conscientiousness 17:920–923
cross-situational consistency 17:920–921
educational attainment 17:921–923
extraversion 17:920, 17:922–923
five-factor model 17:922–923
health 17:923
job performance 17:921–923
lexical approach 17:922
measuring traits 17:921
molecular genetics 17:921–922
neuroticism 17:921–923
openness to experience 17:922–923
quantitative genetics 17:921
rank-order stability 17:920, 17:922
selection of traits 17:922–923
socially relevant outcomes 17:923
sources of traits 17:921–922
subjective well-being 17:922–923
typological approaches 9:1009, 9:1011
and values 25:45
volunteerism 25:268–270, 25:272
at workplace 17:821–826
assessment 17:823
current concerns and future research 17:822–823
five-factor model 17:821–822
history 17:821–822
psychology 17:822
Personality and self-concept 21:465–466
multidimensional 21:465–466
self-esteem 21:466
Personality assessment 16:725–726, 17:827–834, 17:849
achievement motivation 17:831–832
assessment methods 17:851–855
biodata 17:852–853
cognitive and mental ability tests 17:853–854
interview 17:851–852
personal references 17:852
personality tests 17:854–855
backgrounds 17:827–828
descriptive approaches 17:827–828
explanatory personality models 17:828
clinical psychology 17:828–830
affective-motivational personality constructs and psychopathology 17:829
dispositions of self-regulation 17:829–830
personality disorders 17:829
definitions 17:827
educational psychology 17:831–832
domains and methods of assessment 17:831
motivation and volition 17:831–832
essential methods 17:849–850
impression management 17:849
observation data 17:850
personal history/biography 17:850
physiological evidence 17:850
self-deception 17:849–850
self-report 17:849
test performance 17:850
five-factor model 17:827
organizational psychology 17:830–831
broad dimensions of personality 17:830
narrow facets of personality 17:830–831
recent developments 17:851
applicant perceptions 17:851
changes in business 17:851
construct-driven approaches 17:851
ideas of gurus (business writers) 17:851
labor market shortages 17:851
legal changes 17:851
recommendations of consultants/academics 17:851
technological developments 17:851
reinforcement sensitivity theory 17:828
simple selection model 17:850–851
temperament 17:827–832
workplace 17:823
Personality Assessment Inventory (PAI) 16:725
Personality development 9:238–239, 11:537, 11:539, 11:541, 15:373–374, 17:884, 25:592
concept of 17:876
in context 17:876
continuous and categorical views of 17:876–877
future studies on 17:888–889
genetic and environmental influences 17:885–888
biological maturation 17:886–887
environmental development 17:887–888
environmental fluctuation 17:886
genetic and environmental continuity 17:887
genetic continuity 17:886–887
genetic development 17:887–888
genetic set-points 17:886
social maturation 17:887
manifestations of environmental and epigenetic factors 17:889
mean-level change 9:238
personality traits 17:884
quantitative and molecular genetic studies 17:889
rank-order stability 9:238
role of culture 17:878
temperamental contributions to 17:877–878
Personality differences 17:884
future studies on 17:888–889
genetic and environmental influences 17:884–885
manifestations of environmental and epigenetic factors 17:889
personality traits 17:884
quantitative and molecular genetic studies 17:889
Personality disorder psychopathology 17:916–917
Personality disorders 17:592–593, 17:596, 17:829, 17:891–898, 19:305, 24:278
Anglo-American concepts 17:892
antisocial personality disorder 17:892, 17:894, 17:896
changes in current system 17:895
classification 17:894t
clinical utility 17:895
diagnostic criteria 17:893t
DSM-5 17:894
five-factor model of personality 9:242–248
French concepts 17:891–892
future in diagnostic and clinical practice 17:897
German concepts 17:892–894
borderline personality disorder 17:892–893
DSM-III 17:893–894
history of concepts 17:891
ICD-11 17:894
current proposal 17:894–895
importance in general psychiatric management 17:897
mixed and unspecified personality disorders 17:895
and oversharing 17:440–443

psychopathy 17:891–892
short-term psychodynamic therapy for 19:346
treatment 17:895–896
outcomes 17:896
Personality disturbances, psychotherapy 19:373
Personality facets 9:237–239
heritability 9:238
personality development 9:238–239
self-observer agreement 9:238
Personality factors, adolescent health 1:121
Personality identity, development in adolescence 21:434–436
Personality integration model 3:452
Personality judgment *see* Accuracy of person perception
Personality psychology 8:703–708, 17:863–864, 17:867–868, 17:906–913
aggressiveness 17:910–911
Chaucer on 17:909
contemporary personology 17:909–911
Arthur Schopenhauer and Natürpsychologie 17:910–911
clinician researchers 17:911
nineteenth-century science 17:910
psychoanalysis 17:911
stages of human development 17:909–910
distinction between exploratory and confirmatory factor analysis 8:707
empiricism 17:910
factor analysis 8:705–707, 17:910
determination of number of factors 8:706
exploratory and confirmatory approaches 8:705
extraction of factors 8:705–706
factor model 8:705
identification and rotational indeterminacy 8:706–707
five-factor model (FFM) 2:559
generalized model for latent variables 8:704–705
from illness models to wellness models of human nature 17:911
individualism 17:908–909
minstrelsy of Medieval Europe 17:909
passions and individual in love 17:908
in Renaissance Literature 17:909
Kaiser-Guttman rule 8:706
parallel analysis 8:706
positive developments 17:911–912
epigenetics 17:912
homuncularism 17:912
neoteny 17:912
reduction of pre- and postnatal stressors 17:912
positive psychology 17:911–912
principles and disciplines 17:906–908
in Augustinianism 17:907–908
Dark Ages 17:908
early biological substrates 17:906–907
in Hellenism 17:907
romantic love 17:909
scree test 8:706
Shakespeare on 17:909
synopsis 8:707
tacit knowledge 17:910
unconscious 17:910
Personality science 17:763–764
Personality styles 4:110–111
Myers–Briggs Type Indicator 4:111
Personality tests 17:854–855
Personality theories
of racism 19:866
traits *versus* typologies 9:1009
Personality trait hierarchy 9:236–237
Personality trait models 17:914
Personality traits 1:563, 9:1010–1012, 16:498–499, 17:884, 17:915–916
and aggression 1:378
callous unemotional 1:377
and crime 17:774
causal relationship 17:774
common types of offenses 17:774
empathy 17:776–777
Eysenck theory 17:777
five-factor model (FFM) of personality 17:777
future directions 17:777–778
impulsiveness 17:775–776
temperament 17:774–775
and cultural intelligence 5:434
and religiosity 17:802–803
risk attitude as 17:809–810
risk-taking as 17:809–810
see also Big Five personality traits
Personality variables, cross-cultural training 5:338
Personalization 16:774, 20:582
community care 4:344
Personalized education 12:325
Personalized medicine 2:622–623, 17:925–931
behavior modification 17:926, 17:928
breadth of changes 17:925–927
diagnosis and screening 17:925–926
palliative care and end-of-life care 17:926–927
prevention 17:926
treatment 17:926
cancer 17:925–926, 17:928
direct-to-consumer testing 2:623
genomics 17:925, 17:927, 17:930
historical perspective 17:925
implications for health care systems 17:927–928
complexity of delivery of health care by nontraditional means 17:928
consequences of predicting outcomes 17:928
electronic records and data management 17:927–928
need new systems for introducing innovations 17:928
risk of increased costs 17:927
new advances enhance feasibility of improving 17:925
roles for behavioral scientists 17:928–930
change management 17:929–930
creating teams 17:930
interventions requiring behavior change 17:929
risk perception 17:930
treatment adherence 17:929–930
Personalized risk feedback 10:706–707
cue-adaptive reasoning perspective 10:706–707
motivated reasoning perspective 10:706
quantity of processing perspective 10:706
Personalizing education 4:112
and learning technologies 4:114
Personally meaningful ethnicity 8:150–151
Person-by-item interaction 9:836
Person-centered approaches, engagement 1:34
Person-centered expressive therapy 20:740–741
Person-centered methods 19:359–360
Person-centered planning, for disablism 6:450
Person-centered psychotherapy 20:737, 20:739–741
Person–context models, of human development 11:278
Personenverbandstaat 1:966
Person–environment (PE) fit 1:33, 1:35, 4:788, 17:137, 21:61, 23:547, 23:549, 25:537–538
at work 23:547–548
Person–environment relations
approaching on macro-level of analysis 6:887
objective approaches to 6:885–886
subjective approaches to 6:886
Person–Environment Stress Model of aging 6:886
Personhood 3:656, 16:183, 25:559
anthropology of 17:932–935
Durkheimian individual 17:932–933
existence and experience 17:934–935
form and meaning 17:934
Individualism and individuality 17:934–935
non-Durkheimian individual 17:933–934
and property rights 19:168
resistant, in socialist societies 22:851–852
and secrecy 21:374, 21:377
Personification, and fetishism 9:107–108
Person-in-environment
and social work 6:871, 6:873–874
Person-job fit model, of well-being and burnout 25:534–535
Personnel 15:504
electroencephalography 7:354, 7:356
Personnel administrators 20:164
Personnel psychology 11:879–880
workplace 17:822
Personnel selection 17:834–836, 22:15–17
psychology of 17:936
criterion development and measurement 17:936–937
job and work analyses 17:936
predictors 17:937–939

Personnel studies, applied research (criminology and criminal justice) 5:130
Person-oriented analysess 12:812–813
Person-oriented approach
  school burnout 21:61
  school engagement 21:61
Person-oriented play, similarities, during infancy 18:219
Person-period dataset 23:788
Person-related factors 25:539
Persons 17:735–738, 17:932–934
  challenges of conceiving 17:736–738
  continuity and radical discontinuity between animals or machines and 17:737
  fulfillment of 17:737–738
  and identity 11:532–535
  identity over time 17:736
  as a 'impersonal' or 'intrapersonal' forces 17:737
  land tenure 13:231–232
  legal reasoning 13:778, 13:782
  and mind–body dualism 15:545–548
  naming 16:183
  relation between self and 17:735–736
  and science 21:250
  and self 21:445
  and social networks 5:560–561
  sociology of 15:534
Person's universe score 9:835
Persons who stutter (PWS) 9:266
Person–situation debate 17:814
Person–situation interactions 11:171–172, 17:697–698
Person-to-person contact, and health 22:369
Perspectivalism, and cultural relativism 5:484, 5:486
Perspective taking 19:216–217, 19:219
  and empathy 7:544–545
  and prosocial responding 7:318–319
    association between moral judgment and moral behavior 7:319
    cross-cultural similarities and differences in prosocial moral reasoning 7:318
Perspective view 18:799
Perspectives 13:225–226, 13:228–229
  diversity 4:740
  in hermeneutics 10:785–786
  of organizational diversity 6:597
Persuasion
  and attitude 2:169–173
  in groups 10:423
  versus social influence 22:348
  in surveys and polls 23:756
Persuasion theories 17:941–947, 22:528
  attitudes and behavior 17:943
  cognitive approach 17:943
  consistency approach 17:942
  elaboration likelihood model 17:943–944
  heuristic-systematic processing 17:944–945
  learning approach 17:941–942
  self-perception approach 17:942–943
  unimodel 17:945
Persuasive communication 10:828
Persuation, leadership 13:653
PERT *see* Program Evaluation and Review Technique (PERT) network
Pertinent law 13:529–530
Peru 22:625
Peru–Bolivian Confederation 4:574
Peruvian archaeology 23:37
Pervasive community violence (PCV)
  health impact assessment 25:108–110
  measurement of 25:107–108
  related terminologies 25:107
Pervasiveness of computing 21:186
Pervasiveness of flexibility 17:558
Pervasiveness of technologies 13:523
Perverse effects 24:746
  unintended consequences 24:744, 24:746–747
Perverse subsidies 16:466–467
Perversions, sexual 17:948–954
Perversity thesis 20:660
PES *see* Postenumeration survey (PES); Public employment service (PES); Public engagement with science (PES)
Peschel, Oscar 7:723
Pessimism 6:207–213, 10:163–164, 17:263–267
  and anomie 1:715–716
  and disengagement 17:263–267
  giving up 17:264–265
  persistence 17:263, 17:266–267
  studying 17:264
Pessimistic bias 10:704–705, 10:705t
Pestalozzi, Johann Heinrich 17:955–958
  childhood, youth, and early career 17:955–956
  critical reception and research developments 17:957–958
  family origins and childhood 17:955
  journalistic advertising campaign 17:956
  livelihood 17:956
    as an independent writer 17:956–957
    in Burgdorf 17:957
    as institute director 17:957
    pedagogical work 17:957
    in Stans 17:957
  personal and literary life 17:955
  research controversy 17:957–958
  school and studies 17:955–956
  teaching method 17:957–958
  works
    Abendstunde eines Einsiedlers 17:956
    Bitte an Menschen-freunde 17:956
    Christoph und Else 17:956
    Freyheit meiner Vaterstatt 17:956
    Lienhard und Gertrud 17:956
    Meine Nachforschungen über den Gang der Natur in der Entwicklung des Menschengeschlechts 17:956
Pesticides 17:677–678
PET *see* Positron emission tomography (PET)
PETA *see* People for the Ethical Treatment of Animals (PETA)
Pétain, Philippe 9:214
Peter (New Testament) 3:545
Peter Principle 2:915–916
Peter the Great 1:754–755
Peters, Ted 20:288
Petersberg tasks 17:633–634
Petersburg paradox 14:630
Peterson, Richard
  on consumption 16:555
  on cultural preferences 5:449
  on omnivorousness 22:336
  works
    The Production of Culture 5:603
Peterson, Wes 21:952
Petit marronage 14:592
Petite bourgeoisie 12:278
Petitio principii 17:525
Petitions, online *see* Online petitions
Petric, B. 3:318
Petroglyph 18:825
Petroleum
  development of 17:165–166
    commercial 17:165–166
    drilling process 17:166
    onshore possibilities 17:166
  historical context 17:168–170
  origin of 17:165
  status of depletion 17:167–168
Petroleum disputes 12:471
Petroleum resource management system 17:166
Pettazzoni, Raffaele 7:375
Pettit, Philip 3:646
  on civic republicanism 9:402–403
  on republicanism 20:477–480
Petty, Sir William 16:227–228, 16:230, 23:423
  on division of labor 6:602–603
  works
    Political Arithmetick 6:602–603
Petty offenses 15:589, 15:591, 15:593
Petty theft 15:592
Petty tyranny 23:704t
Petty vandalism, and cultural heritage 5:416–417
Pew Research Center
  Hispanic Trends Project 9:628
PFA *see* Psychological first aid (PFA); Psychological First Aid (PFA) Mobile
Pfand 21:404
PFC *see* Prefrontal cortex (PFC)
Pfeiffer, G.P. 9:646
Pfister, Oskar 20:355
Pfister Indices 3:867
Pfizer Meningitis study in Nigeria 13:561
PFL *see* Preparation for future learning (PFL)
Pfungst, O. 8:572
PG *see* Pathological gambling (PG); Phoneme-to-grapheme (PG) conversions
PGA *see* People's Global Action (PGA)
PGD *see* Preimplantation genetic diagnosis (PGD)
PGM *see* Precision Guided Munitions (PGM)
PGS *see* Preimplantation genetic screening (PGS); Primary Genealogical Space (PGS)

Phalansteries 22:836–837
Phallogocentrism 15:291
Phallus 13:213–214
Phan Boi Châu 25:95
Phantom® device 25:133
Phantom heritability 8:717
Phantom limbs 18:1–5
  central sensitization 18:3–4
  deafferentation 18:2, 18:5
  ectopic activity 18:1
  nonpainful 18:2–5
    centrally triggered sympathetic-efferent somatic-afferent coupling mechanism 18:5
    cortical maps 18:3
    fading, telescoping, and shrinking 18:2–3, 18:3f
    psychological and emotional processes influences 18:4
    psychophysical correlates of sensations 18:3–4
  pain 18:1–2
    memories 18:1–2
    treatment 18:5
  sympathetically generated paresthesias 18:4f
Phantom-limb awareness 10:502
Phantom-limb hallucination 10:500
Phantosmia 17:199t
Pharaonic circumcision 8:879
Pharisees 3:546
Pharmaceutical industries 18:6–10
  clinical research 18:7–8
  clinical trials 18:7–10
  diseases 18:8
  drugs 18:6
  health 18:6–9
  hegemony 18:9
  illness 18:8–9
  as a key social actor 18:6
  marketing 18:9–10
  medical knowledge 18:9–10
  medicine 18:6, 18:9
  pharmaceuticals 18:6, 18:10
  physicians 18:9–10
  regulation 18:6–7
  research funding by 20:524
  trial income 18:7
Pharmaceuticalization 18:8–9
Pharmaceuticals 10:618
Pharmacia Corp. v. Alcon Laboratories 13:769–770
Pharmacoeconomics 19:458
  with anticraving drugs 19:461
  with antidepressants 19:460
  basic terms of 19:458–459
  with cognitive enhancers 19:461
  with hypnotics 19:460
  with mood stabilizers 19:460
  with neuroleptics 19:459–460
  study designs 19:459
  with tranquilizers 19:460
  types of analysis 19:459
Pharmacogenetics
  for ADHD 2:160
  personalized medicine 17:926
Pharmacogenomics 2:623, 10:33
Pharmacological approaches, for sexual disorder 21:691–692
Pharmacological reinforcers 20:213
Pharmacological studies, of dreaming 6:654–655
Pharmacological therapies 20:59
  for post-traumatic stress disorder 18:705
Pharmacological treatments
  for OCD 17:116
    combined with cognitive-behavioral treatment 17:116
  for panic disorder/agoraphobia 17:478
  for psychosomatic disorders 19:489
  rheumatoid arthritis 20:655–656, 20:658
Pharmacology 19:462
  alcohol 1:507
Pharmacoresistance, seizures 13:34–35
Pharmacotherapy 8:609
  combined with psychotherapy 19:495
  for obsessive-compulsive disorder 17:125
  for schizophrenia 21:12
Pharmacy 3:371
Phase theory 16:4
Phase-advance hypothesis 22:85
Phase-Impenetrability Condition (PIC) 16:4
Phases
  of negotiation 16:425
  therapeutic community 24:277–278
    assessment 24:277
    engagement 24:277
    preparation 24:277
    recovery 24:277–278
    treatment 24:277
Phasic prediction error signal 16:640–641
Phatic communion 18:795
Phatic functions of language 12:783
Phayre, Arthur, works, *History of Burma* 11:84
Phencyclidine 6:674
Phenelzine 1:778–779
Phenogenotype 9:813
Phenome 10:43
Phenomena
  associated with bereavement 2:532–533
  interpretation, objectivity 17:89
  social representations as 20:446–447
Phenomenal absence 23:296
Phenomenal absolutism 8:167
Phenomenal body 19:492
Phenomenal consciousness 4:675–676
Phenomenal neural correlates of consciousness 4:675
Phenomenal world 13:135
Phenomenalism 7:567–568
  logical empiricism/logical positivism 14:313–318
Phenomenological properties of autobiographical memory 2:285
Phenomenological reduction 21:617
Phenomenological sociology 21:148
Phenomenological theories 22:959
Phenomenology 1:896, 10:738, 12:658, 17:934, 19:491–492, 22:817, 22:994, 23:857–858
  of Alzheimer's disease 1:584–585
  of firm 17:390–391
  flow 9:260–265
  Husserl on 11:442
  and interpretation 12:642–643
  Mannheim on 14:496
  and materiality 14:768–769
  and objectivity 17:97–99
  philosophical aspects 18:18–23
    body 18:20
    cognition 18:20–22
    consciousness 18:18–21
    embodiment 18:20
    experience 18:20
    first philosophy 18:22
    interaction theory 18:19–21
    lifeworld 18:20–21
    minds and emotions 18:19–20
    the other 18:21–22
    perception 18:20
    social action 18:21
  places 18:149
  psychological 19:673
  and psychology 19:413
  of religious experiences 18:809
  Schütz on 21:146–149
  social, and symbolic interactionism 10:786
  in sociology 12:352–353, 12:355, 18:11–17
    background 18:11–12
    bracketing 18:12
    consciousness 18:11–15
    constructionism 18:11–12
    constructivism 18:11–12, 18:15
    ethnomethodology 18:11–13
    human studies 18:15
    Husserl on 18:11
    intentionality 18:11–13
    practice 18:14–15
    Schütz on 18:11
    social action 18:11
    subjectivism 18:11–15
    techniques 18:12–13
    theory 18:13–14
  of spirit 10:737
Phenomenon scale 21:2
Phenotype 2:450, 10:42–45, 11:289, 11:294
  description of 10:42–43
  of developmental dyslexia 6:729
  and environment 10:44
  epigenetic inheritance 7:832–833, 7:835–837
  genetic studies of behavior 9:932–937
  relationship with genotype 10:43–44
Phenotypic covariance matrix
  polygenic model for 8:716
  variance components (VCs) model for 8:716
Phenotypic gambit 8:400–401
Phenotypic profile, Williams syndrome 25:579–582
Phenotypic profiling 9:960
  forensics 9:960
Phenotypic variance 8:715
Phenylketonuria (PKU) 9:927–928, 9:997

Phenylketonuria (PKU) (*continued*)
genetic screening 2:620
prenatal and neonatal screening 20:702
Pheromonal effects 17:198
Philadelphia City Planning Commission 24:886
Philanthrocapitalism 22:168–169
Philanthropic foundations 22:648–649, 22:651
growth of 18:324
Philanthropic organizations 6:171
Philanthropy 4:338, 16:914–915, 22:168–170, 22:172–173, 22:579–580, 24:762–764
altruism 3:364
and biomedicine 2:665–666
definition 22:579
economics of 3:358–363
fundraising 3:362
future research 3:362
giving 3:360–361
and government grants 3:361–362
matching grants 3:362
taxes 3:360–361
giving in global perspective 3:367–368
giving in United States 3:366–367
historical meaning of 3:364
nonprofit sector 3:365–366
theories of 3:366
and policy knowledge 18:294–295
political support 22:580
practice 22:579–580
present-day paradigms of giving 3:364–365
prosocial behavior 1:581
research, emerging trends in 3:368–369
social capital 3:365
social impact investing 3:365
venture philanthropy 3:365
Philip II of Spain 15:579
Philippines 23:63–64
Americans in 23:64
cave sites 23:65
early childhood care and education (ECCE)
mobile early childhood care and development programs in 6:743–744
mobile playgroups in Davao City 6:743
fertility transition in 9:85–89
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Philippines law 13:815
Phillippino Muslims 12:729
Philips, Cathee Johnson 1:50
Phillips, David 23:687
Phillips curve 9:221
macroeconomics 14:402–404
Phillips Lumileds 11:840
Philological Quarterly 15:136
Philology 20:324
legal reasoning 13:777
metaphors in 15:290
theory of poetic logic 15:290
Philonic Judaism 3:237
Philosophers, view of health 10:607–608
contemporary philosophers 10:610–611
Philosophers' Committee 11:361
Philosopher's philosopher, Rorty as 20:761
Philosophes 7:656, 22:990
Philosophical alienation 1:551–552
Philosophical anthropology 9:671–672, 15:321, 19:313
and psychiatry 19:313
Philosophical aspects
equality 7:906–911
functional explanation 9:519–524
of individualism/collectivism 11:811–815
of irrationality 12:719–723
of justice 12:943–948
of natural and social evolution 8:388–391
objectivity 17:96–102
of personal identity 17:742–749
of phenomenology 18:18–23
postmodernism 18:684–689
rational choice explanation 19:896–900
reference and representation 20:67–71
relativism 20:233–238
republicanism 20:477–482
social properties 22:500–505
trust 24:668
welfare 25:504–510
Philosophical awareness of ethical diversity 6:589–590
Philosophical bioethics 2:631
Philosophical developments, of actor–network theory 1:83–84
Philosophical empiricists 5:504
vindication of acceptance of scientific claims 5:504
Philosophical ethics 2:613–614
Philosophical hermeneutics 10:778
Philosophical ideas and traditions 7:255–256
Philosophical linguistics studies 14:228
Philosophical ontology 22:490–491
Philosophical pluralism 1:895–896
Philosophical radicalism 21:311
Philosophical reasoning 16:926
Philosophical sociology 21:967
Philosophical sublimation 9:143
Philosophical tenets, of harmony 10:529–530
Philosophical themes, biomedicine 2:663–664
Philosophical theories of justice 12:935
Philosophical Transactions 18:917, 21:163
Philosophisches Wörterbuch der Philosophie (*Dictionary of Philosophical Terminology*) 4:548–549
Philosophy 3:63–64, 4:318–320, 9:671–672, 14:453, 18:31–36, 20:322–324, 23:63
archaeology of 1:891–898
Aristotle's 18:31–32
Bourdieu on 2:777–778
of culture 21:391
and education, encounter between 7:257–258
of education 7:255–256
contemporary 'field' of 7:258–259
of emergence 7:433–434
enlightenment 7:656–658, 7:660–662
of ethical consumption 8:27
ethics 18:32, 18:35
eviction of eudaemonism from 8:214
fire in 9:200
Frankfurt School 9:382
of good life 10:521–523
Greek 18:31–32
Habermas on 10:469, 10:471–472
as highest praxis 25:599
Hippocrates 18:31–32
of history 1:980
history of engagement 23:63
antiquarianism 23:63–64
early calls for science 23:63–64
of human sciences 20:668
of knowledge 13:74
tacit 13:80–83
of language 4:548
metaphors and 15:291–292
of language, Humboldt on 25:286
of law 18:24–30
in league with science 25:600
learning 13:699–704
liberal 14:26–27
of logic 19:391
logical empiricism/logical positivism 14:313–315, 14:317
of mathematics 14:870–871
of measurement theory 14:868–872
metaphysics of Middle Ages 18:32
of mind 3:263–264
mind–body problem 18:32
modern, practical intentions 25:600
moral–practical side 25:600
technical–practical side 25:600
of modern anonymity 21:149
modern use in psychiatry 18:35
of nature 10:738
Feyerabend on 9:122
neonatal imitation in 16:481
philosophical psychiatry 18:31
philosophical psychology 18:33–34
Plato on 18:31–32, 18:209–210
political thought 18:486–489
pre-Platonic understanding of 25:598
probability 19:23
professionalism 18:32, 18:35
and psychologism 19:392
psychopathology 18:31–33
of the Real 10:738–739
of reasoning 5:939–940
republican 20:481
Rorty's interest in 20:761, 20:764–765
of science 1:891–893, 4:550–551, 13:133–134, 13:136–137, 14:313–315, 14:317, 14:871, 15:319–320, 15:323, 21:201, 23:63
see also Scientific philosophy
mathematical models in 14:791–799
new archaeology and 1:892–893
Popper on 18:522–527
post-positivist philosophy 1:893
of science, and cases studies 3:211
of secularization 21:390–392
self and identity 18:34–35

of social science 22:489, 22:491
Socrates on 18:31–32
of spirit 10:738–739
of technology assessment (TA) 24:125
Utopian thought 14:251–259
of wisdom 25:598–603
as wisdom proper 25:599
Wright's contribution 25:759–760
Philosophy and Public Affairs 25:347
Philpott, Stuart 1:641–642
Phlogiston 3:372
Phobias 1:805, 8:872, 9:416–417, 18:37–44
acrophobia 25:141
aviophobia 25:140–141
and fear-potentiated startle 8:864–865
phobia of small animals 25:141–142
specific phobias 25:142
Phobic Origins Questionnaire 18:40
Phoenix 17:158
Phological disorders 18:45–64
Phonationals 18:65, 18:69–73
classification 18:69t
composition 18:69–70
four-level tonal model 18:71
multiregisters 18:71
system 18:69
tense *versus* lax vowels 18:73
three sets of voiceless stops 18:71–72
types 18:70
typology 18:71
Phonatory loop 9:268
Phone-in program 15:61
Phoneme ordering errors 18:51–52
Phonemes 4:92, 6:728–729, 15:827–828, 15:830–831, 16:593–595, 18:48
definition 25:226
to touch 14:196
to visual 14:196
Phoneme-to-grapheme (PG) conversions 1:453, 23:266–267, 23:269
Phonemic awareness 8:395, 20:8–9
Phonemic paraphasias 18:58, 21:935
Phonetic 23:229–231
Phonetic domain structure, nonsyntactic determinants of 23:897–898
Phonetic interface 9:877
Phonetic radicals 20:2
Phonetic sign 25:761–763, 25:765
Phonetically speaking 16:60
Phonetics 14:225, 15:826–827
articulatory 18:65–74
phonationals 18:69–73
classification 18:69t
composition 18:69–70
four-level tonal model 18:71
multiregisters 18:71
system 18:69
tense *versus* lax vowels 18:73
three sets of voiceless stops 18:71–72
types 18:70
typology 18:71
segments 18:66–68
consonants 18:66–68
secondary features 18:68
vowels 18:68
theoretical considerations 18:66
basic framework 18:66
linguistic phonetics being phonology 18:66
Phonogenesis 10:334
Phonographic recording, of interviews 12:674–675
Phonological agraphia 6:735
Phonological alexia 6:732
Phonological awareness 3:510–512, 23:267
and learning 16:731
Phonological Awareness Test 2 3:796
Phonological disorders 18:45, 18:47
from Jacques Lordat 18:45–46
Phonological domain structure, nonsyntactic determinants of 23:897–898
Phonological dysgraphia 1:453, 1:455
Phonological dyslexia 6:730
Phonological encoding 23:257
Phonological erosion 10:330
Phonological errors 4:92
Phonological information
and gender assignment 9:796–797
sensitivity of children to 23:266–267
Phonological mapping negativity (PMN) 9:482, 9:482f
Phonological markedness in optimality theory 14:535
Phonological paraphasic errors 18:51–52
Phonological processing 13:369–370, 13:399
and language learning impairment 13:345–346
Phonological processing deficits in dyslexia 6:729–730
Phonological recoding, visual word recognition 25:215
Phonological representations
abstractness of 18:77
in dyslexia 6:729
Phonological short-term memory 21:909–910
deficits 21:911–913
input-related dissociation 21:911
lexical–semantic long-term memory representations 21:911–912
pattern of impairment 21:911
verbal rehearsal 21:912–913
and long-term learning 21:912
neurofunctional architecture of 21:914
Phonological system 11:824–825
Phonological variables, and social class 13:282–283, 13:282t–283t
Phonological words 13:276
Phonological-articulatory loop 23:246
Phonology 12:782, 13:942, 14:229, 15:826–827, 18:47, 20:5–7
abstractness of phonological representation 18:77
American structuralist 18:76
and bilingual aphasia 2:574–575, 2:577–578
developmental 18:76
generative 18:76
goals of 18:75–77
history 18:75–76
Saussure and the Phonemic Principle 18:76
interaction with other areas of grammar 18:79
phonological regularities 18:77–78
rules *versus* constraints 18:79
stem change in German and Danish plurals 15:829
structure of representations in 18:78–79
as a system of rules and representations 18:76–77
theory, issues in 18:77–79
Phonotactic constraints 13:310
Phonotactics 15:827–828
Phosphenes 10:499
Phospholipase C (PLC) 16:753
Phosphorylation 12:712–713
Photo, preserving dance 5:669
Photo journalism 25:172–173
Photo-elicitation 25:182
Photogrammetric aerial surveys, in archaeology 23:744
Photographer, as engaged observer 18:84
Photographic montages 9:810–811
Photographs 19:671
as cultural narratives 25:181–182
in online profiles 24:148
in research process 25:182
Photography
aerial 10:53–54
as medium 18:81–86
and art 18:83–84
distant worlds 18:82
image concept 18:82–84
and mass society 18:84–85
photographer as engaged observer 18:84
photographs of people 18:81–82
and science 18:83
technology and revisions of photographic meaning 18:85–86
origins 18:81
and visual sociology 25:191–196
Photomontage, and genealogy 9:806, 9:810–811
Photoreceptors 25:158–159, 25:184
Photoshop 25:172
Photosynthesis 7:627
Phototherapy 23:20
Photovoice 16:406, 22:271
Photovoltaics 18:370
PHP *see* Principle of Head Proximity (PHP)
PHQ *see* Patient Health Questionnaire (PHQ)
PHQ-4 *see* Patient Health Questionnaire 4 (PHQ-4)
Phrase structure 9:872–873
Phraseology, and society 13:288–289
Phrases/phrase-like units, in syntactic dependency 6:193
Phrase-structure rules 23:886–889
Phthisis 17:29
p-hub location problem 23:174t

Phung Nguyen culture 23:66
Phylogenetic arrays, importance of migration in spreading 5:396–397
Phylogenetic comparative methods 4:416–417, 4:420, 8:374
Phylogenetic methods 8:370, 8:418
Phylogenetic prediction
  dinosaur genome size 4:417–418
  feeding time and evolution of food processing 4:418–419
Phylogenetic reconstruction
  primates, evolution of 18:904, 18:906
Phylogenetic revolution, evolutionary biology and 8:371
Phylogenetic signal 4:419–420
Phylogenetic studies of archeological and ethnographic cultures 5:397–398
Phylogenetic trees 8:449
Phylogenetics 8:370
  cultural evolution 5:390
  modern 8:371–372
Phylogeny 14:368
  evolution of language 8:365–366
  of human cerebral cortex 3:333–334
  of human sociality 22:894–895
  in prehistoric contexts 5:395–396
Phylogeny reconstruction
  evolution of language 8:365–367
Phylum 11:820
Physical abuse, of elders 7:330
Physical access, urban settlements 24:837
Physical accounting 14:760
Physical activity (PA) 2:601, 10:583, 18:87–99
  and arthritis 18:88, 18:90
  behavioral and social strategies 18:92–94
    health belief model 18:93
    intrapersonal and interpersonal concepts and principles 18:92–93
    policy and environmental concepts and principles 18:94
    social and cultural principles 18:93
    social support 18:94
    transtheoretical model 18:93
  building community capacity 18:94–95
    effective partnering and coalition building 18:95
    essentials of community organization and community development 18:94–95
    program evaluation tools 18:95
  of children 10:683–684
  and diabetes 18:87–88, 18:90
  effects on mental health 18:88–90
    anxiety 18:89–90
    depression 18:89
  effects on physical health 18:87
    activity among patients 18:88
    biological mechanisms 18:88
    breast cancer 18:88
    cancer 18:88
    cardiovascular disease 18:87–88
    colon cancer 18:88
    coronary heart disease 18:87–88
    further chronic diseases 18:88
    overall mortality 18:87
    stroke 18:88
  evidence-based community-level strategies 18:95, 18:95t
  exercise 18:92–94, 18:97
  foundations 18:91
  guidelines for adults and youth 18:95t
  health behavior 18:92–94, 18:96
  health disparities 18:96
  health promotion programs 10:671
  interventions 18:95–97
    behavioral and social approaches 18:96–97
    campaigns and informational approaches 18:95–96, 18:96t
    policy and environmental approaches 18:97
  physical education 18:96–97
  prevalence of 18:91t
  public health 18:91, 18:95
  and sleep 22:60
  transportation 18:94–95, 18:97
  urban design 18:97
  youth, developmental perspectives on motivation in 15:914–920
Physical adaptation for rape 19:887
Physical aggression 1:373, 18:614
  children, effects of media violence on 15:38
Physical and mental health 20:296
  of older people 7:336–337
Physical anthropology 1:746, 8:588
  quantification in 19:704
Physical attractiveness 12:626
  benefits of 8:4
  as erotic capital 8:1
  and labor market 2:383–391
Physical capital accumulation 7:46–47
Physical causation 3:261–265
Physical considerations, for language evolution 11:332–333
Physical coordinates, graphics 10:359
Physical dependency 19:464
  in adulthood 1:182
  effects of alcohol on 1:508
Physical determinism 16:438–440
Physical development, prenatal/infant 18:852–853
Physical disability, and natural disasters 16:316–317
Physical disease, mental disorder following 4:391–392
Physical disorders, behavioral medicine 2:465
Physical education (PE) 18:96–97
Physical education teaching 21:136
Physical environments, and cultural production 5:469
Physical ergonomics 11:299
Physical fitness, as leisure activity 13:847
Physical functioning, rheumatoid arthritis 20:656–657
Physical geography 10:83–86
Physical health 23:562
  and abusive supervision 23:702
  acculturation and 1:72–73
  caregiving 3:136–137
  and doctor–patient interaction 6:616
  effects of physical activity on 18:87
  and emotions 7:496–500
  and gender 9:704–708
  of immigrants 15:408
  influences of emotion regulation on 7:455
  and LGBT aging 13:866
  and mental health 15:188
  and mindfulness 15:551
  outcomes, bereavement 2:532–533
  -related consequences of stress 23:565–566
  religion and 20:266–268
  and sexuality 21:843–844
Physical health prophylaxis 19:203
Physical illness
  and dyadic coping 11:594
  and psychological stress 23:583–584, 23:586
Physical impedance, measuring 24:627
Physical inactivity (PIA) 18:91
  and coronary heart disease 4:920
  and gender 9:704–705, 9:707
Physical input output table (PIOT), in material flow analysis 14:762
Physical inputs 12:52
Physical insideness 6:886
Physical integrity, and political repression 18:446–447, 18:451
Physical minimality thesis 18:112
Physical play 14:719–720
  differences during infancy 18:220–221
Physical Quality of Life Index (PQLI) 6:310
Physical reasoning 13:103
Physical rehabilitational 10:716t
Physical representation
  in tangible interaction 24:20–21
Physical safety 22:206
Physical science concepts 21:276–278
Physical sciences 18:100–105
  Aristotelian understandings 18:100–101
  astronomy 18:100
  in Europe 18:100
  instruments 18:100–101
  military application 18:103–104
  Newtonianism 18:101–103
  origins 18:100–101
  social and cultural factors 18:100
  tensions between philosophical stances and methodological approaches 18:101
  twentieth-century 18:103–105
Physical symbol system hypothesis 4:516
Physical therapy 16:908
Physical trade balance (PTB) 14:761–762
Physical violence 25:116
Physical vulnerability 8:855
Physical withdrawal 1:508–509
Physicalism 16:327–328
  alternatives 18:112
  constitution 18:106
  elimination/location 18:112
  formulation of 18:108–112
  grounding 18:106, 18:111–112
  idealism 18:106–107
  instantiation 18:106, 18:110, 18:112

metaphysical necessity 18:110–111
and mind–body dualism 15:548
neutral monism 18:107
panpsychism 18:107
physical 18:106–108
possible worlds 18:109–111
a posteriori 18:108–109, 18:112
a priori 18:107–109, 18:112
property dualism 18:112
proto-panpsychism 18:106–107
qualia 18:109, 18:112
realization functionalism 18:110
substance dualism 18:107–108
supervenience 18:106, 18:110
token 18:109
type 18:109
Physically ill patients
psychiatric comorbidity in 14:1–3
effects 14:2–3
in general hospital 14:1–2
in primary care sector 14:2
psychiatric morbidity in nursing homes 14:2
Physician assistants (PAs) 16:910–911
education of 16:911
roles and work settings 16:911–912
Physician autonomy 10:596
Physician Payment Sunshine Act 20:526
Physician response to opiate use 17:240
Physician-as-gatekeeper 6:442
Physician-assisted suicide 8:281, 8:283–284
Physician–organization structures 11:211–212
Physician–patient communication 6:613–614
Physician–patient issues, health care 19:816
Physician–patient relationship 10:596, 17:598–600
Physicians
changing role of 15:101–102
and gender 9:696–697
moving medical knowledge to 18:9–10
and patient 15:121
view of health 10:607–608
women 25:622–623, 25:626
Physician's questions 6:614
Physico-theology 20:286–287
Physics 14:453
and social network analysis 16:528–530
Physics-based validation 1:370
Physiocracy, and political economy 18:357–358
Physiocrats 11:864, 18:357–358
Physiognomy 7:791
Physiologic principles, of electroencephalography 7:353
Physiological acoustics and optics 10:756–757
Physiological components, test anxiety 24:244
Physiological evidence, personality assessment 17:850
Physiological explanations 8:594
Physiological friendship 9:646
Physiological maturity 21:841
Physiological measures, for health behaviors 10:591–592
Physiological mechanisms, triggering cardiac events 4:921
physiological needs 21:487
Physiological process, in self-regulation 21:553–555
Physiological psychology 5:679, 10:527
Physiological reactivity, and pessimism 6:209–210
Physiological resources 2:867
Physiological responses
of emotions 7:461–463
universal emotions 5:544
Physiology
emotion assessment through 7:521
and emotions 7:498–499
of food preferences 9:296–299
occupational health 17:133
social 22:990
and tactile processing 24:16
of vestibular system 25:65–66
hair cells and afferents 25:65
vestibular nuclei 25:65–66
Physiotherapy Evidence Database (PEDro) 8:330
Phytogeography 25:279
Phytoliths 7:695
paleoethnobotany 17:457
PI *see* Party identification (PI); Principal investigator (PI); Proactive interference (PI)
PIA *see* Physical inactivity (PIA); Privacy impact assessment (PIA)
PIAAC *see* Programme for the International Assessment of Adult Competencies (PIAAC)
Piaget, Jean 2:349, 4:43–44, 12:11–12, 12:341, 14:306–308, 16:512, 18:114–120, 19:669, 25:324–326
administrative responsibilities, professorship and new directions of research 18:115–116
childhood and adolescence of 18:114
on childhood friendship 9:442–443
on circular reaction 15:958
cognitive development theory of 4:52–54, 18:115, 18:120–124, 22:814, 4:53f
assimilation 18:120, 18:122
beyond the theory 18:123–124
constructivism 18:120–122
decalage 18:121–122
developmental psychology 18:120–124
epistemology 18:120–121
equilibration 18:120–123
factors 18:121–122
genetic epistemology 18:120
heritage 18:122–123
individual differences 18:122–123
intelligence 18:120–121
mechanism 4:52–53, 4:53f
neo-Piagetian 18:122–123
psychology 18:121
schemes 18:120–122
stages 4:53–54, 18:121, 4:54t
unsolved issues 18:122
complementary studies of 18:114–115
on concrete operations 18:116
critique of behavioral explanations 18:125–126
early psychological research by 18:115
on evolution of mollusks 18:117
on experimental psychology 18:115
on factors of scientific creativity 18:118
on genetic epistemology 18:116–118
importance of work in social sciences 18:118
on infant memory 6:759
on intellectual activity 18:116
on intellectual development theory 16:501–503
on learning 25:317–318
on logical model of 'grouping' of mental operations 18:116
on neo-Piagetian theories of cognitive development 16:501–510
on open education 17:227
perspectives on coconstructivism 3:914
on pretend play 18:873
on psychological stages of children 4:43
on psychopathology 18:114–115
rejection of neonatal imitation 16:484
on reproduction of the 'sedum' plants 18:117
on schema 18:784–785
on structural analysis 18:116
studies of children's reasoning 18:115
theory of human development 18:125
theory of human development and education 18:126–130
application of 18:130–131
versus behavioral explanations 18:125–126
concrete operational stage 18:128–129
formal operational stage 18:129–130
preoperational stage 18:127–128
sensory-motor stage 18:127
Vygotsky's framework as a competing paradigm 18:131
theory of intelligence 12:307
post-Piagetian theories 12:307
works
Biology and Knowledge 18:117
The Child's Conception of Physical Causality 18:115
The Child's Conception of the World 18:115
The Child's Construction of Reality 18:115
The Child's Mental Imagery 18:117
The Equilibration of Cognitive Structures 18:117–118
Judgment and Reasoning 18:115
The Language and Thought of the Child 18:115
The Mechanisms of Perception 18:117
Memory and Intelligence 18:117
The Origins of Intelligence 18:115
Recherche 18:114
Treaty of Logic 18:116
PIAT-R *see* Peabody Individual Achievement Test – Revised (PIAT-R)

PIB *see* Pittsburgh compound B (PIB)
PIC *see* Phase-Impenetrability Condition (PIC)
Picasso, Pablo, drawings of 25:156
Pick, Arnold 16:684, 16:684f
Pickering, Andrew, works, *The Mangle of Practice* 8:579
Pickpocketing 19:161f
Pickvance, Chris 12:913
PICO *see* Patient; Intervention; Comparison and Outcome (PICO)
PICOS (Participants, Intervention, Comparison, Outcomes, and Study) 23:907
Pictographic symbols 24:263–264
Pictographs 25:780
Pictography
  writing, evolution of 25:762
Pictor, Fabius 11:37
Pictorial manuscripts 11:57
  sources of historical records 11:57
Pictorial (or depictivist) theory 15:220–221
Pictorial turn 2:25
Picture Act Theory 2:25
Picture Story Exercise (PSE) 15:951–952, 15:955
Picture writing 25:780
Picture-Based Value Survey 25:44
Pidgin-creole life-cycle 18:136
Pidginization 5:118–119
Pidgins 5:117–118, 13:307, 22:952, 23:46
  English 18:133–145
PIE *see* Person in Environment (PIE) approach
Pie chart 10:359–362
Piecewise constant exponential model 14:56
Piecewise constant transition intensities 8:314
Piercy, Marge 19:748–749
Piéron's law 19:472
Piesche, Peggy 8:908
Pietist intentional communities 2:145–146
Pietz, William, on fetishism 9:105, 9:107–108
Pignus 21:402
Pigou, Arthur C. 14:621, 16:228
Pigovian tax 25:501–502
Pigs, and food production 9:300–303
PIH *see* Partners in Health (PIH)
Piirainen, T., on Central Asia 3:317
PIJ *see* Palestinian Islamic Jihad (PIJ)
Piketty, Thomas, works, *Capital in the Twenty-First Century* 19:628
Pilgrimage 18:146–148
Pilgrimages 24:602
  association with images of death and rebirth 18:146
  for competing discourses and agendas 18:147
  cultural themes 18:146–148
  hajj 18:146
  holy figure 18:146
  holy places 18:147
  liminality 14:131, 14:133–135
  literal movement 18:146
  macro level of analysis 18:147
  methods of study of 18:147
  and miracles 15:576–577
  Muslim 16:377
  opportunities for social unification and moral regeneration 18:146
  research themes 18:147–148
  social scientific approach to 18:146–148
  themes in study of 18:146–147
  Turner's analysis 18:146–147
Pilkington Committee 19:527
Pillsbury, W.B., works, *Handbook of General Psychology* 7:591
Piloerection 7:467
Pim Fortuyn 18:610–611
PIMAP *see* Plug-in maximum a posteriori (PIMAP)
Pinard, Adolphe 8:241
Pinch, Trevor 8:577
  works, *The Golem series* of books 21:310
Pineal gland, functions of hormones 16:650–651
Pinel, John 13:32
Pinel, Philippe 18:178
Ping-jiang city 24:870
Pink, Daniel, works, *Free Agent Nation: The Future of Working for Yourself* 18:666
Pink Panther gang 24:598
Pink plague 13:861
Pinochet regime 3:683
Pinpricks experiments 8:567f
PINs *see* Person identity nodes (PINs)
Pintrich's model of self-regulated learning 21:536–537
Pinyin 20:7–8
Pioneering, cross-national research 6:266
Piore, Michaele 3:23
PIOT *see* Physical input output table (PIOT)
PIPA *see* Protect IP Act (PIPA)
PIQ *see* Performance IQ (PIQ)
PIR *see* Psychological insulin resistance (PIR); Psychology of Interpersonal Relations (PIR)
Pirate entrepreneurs 23:663
PIREDEU *see* Providing an Infrastructure for Research on Electoral Democracy in the European Union (PIREDEU)
PIRLS *see* Progress in International Reading Literacy Study (PIRLS)
PISA *see* Programme for International Student Assessment (PISA)
Pishtaco 23:46
Pissenlit 3:147
Pisum sativum 11:311–312
Pitcairn Island 17:158–159, 17:162
Pitch, child-directed speech 3:400
Pitch discrimination, and temporal lobe 24:208
Pitch-accent languages 23:716
PITF *see* Political Instability Task Force (PITF)
Pithecanthropus erectus 23:64
Pitman, Isaac 6:567
Pitman estimator 12:706
Pitt-Rivers, Julian A. 11:181–183
  works, *People of the Sierra* 1:683–684
Pittsburgh compound B (PIB) 14:443
Pituitary gland 11:482–485
  hormones, functions of 16:649–650
  hormones related to 11:485t
Pity, cognitive-anatomical basis of 16:612
Pius IX 3:234–235
Pius X 3:234–235
Pivotal intervals 8:18–19
Pivoting 22:456
PIXE *see* Proton induced X-ray emission (PIXE) analysis
Pizzo 14:424–425
PKC *see* Protein kinase C (PKC)
PKU *see* Phenylketonuria (PKU)
PL *see* Positive liberty (PL)
Place cells 11:387, 16:623, 16:625–626, 23:876–877
  and spatial navigation 23:163–164, 23:166, 23:168–170, 23:164f
Place characteristics, and internal migration 12:443
Place coding 25:185
Place deixis 6:52
Place identity, and aging 6:886
Place making, of tourism 24:469–471
Place memory 25:136
Place of residence, fertility 9:22
Place quality and governance capacity, coevolution of 18:205
Place randomized trials *see* Cluster randomized trials (CRTs)
Place research, crime and 20:775
Place wars 3:662
Place-based learning 22:7
Placebo effect 18:153–156, 19:467
  anxiety 18:153–154, 18:156
  cholecystokinin 18:154
  from clinical setting to physical performance 18:156
  definition 18:153–154
  delivering treatments unknown to patients 18:156
  dementia 18:155
  dopamine 18:154
  endocannabinoids 18:154
  endogenous opioids 18:154, 18:156
  expectations 18:153–156
  health psychology 10:677–678
  importance of prefrontal lobes in responsiveness 18:155
  learning 18:153–154
  neuroscience of 18:154–155
  nocebo effect 18:153–154, 18:156
  pain 18:153–156
  Parkinson's disease 18:153–156
  psychological mechanisms of response 18:153–154
  rewards 18:153–155
  sports 18:156
  therapeutic success as 24:723
Placebo response, for obsessive-compulsive disorder 17:124
Placebo treatment 20:154

Placebo-controlled double-blind studies 18:157–163
allocation concealment 18:159–160
antidepressants 18:161
blinding 18:159–160, 18:162
clinical trials 18:157, 18:159–160
definitions 18:157–159
efficacy paradox 18:161–162, 18:161f
external validity 18:162
history 18:157, 18:159–160
inadequacy of medical model for complex and behavioral interventions 18:162
internal validity 18:162
meaning response 18:157–159, 18:161
paradox of small specific and strong generic effects 18:161
practical issues 18:162
self-healing 18:157, 18:161
tacit presuppositions, conundrums, and critical aspects 18:160–162
Placebo-controlled trials, ethics of 18:164–173
acceptable risk from withholding treatment 18:168
in research in LMICs 18:168–169
clinical equipoise 18:166, 18:168
controversy 18:166–167
human subjects research 18:169
investigator obligations 18:166–169
key ethical guidance 18:169–171, 18:170t–171t
low risks from withholding treatment 18:168
no or negligible risk from withholding treatment 18:167–168
randomized clinical trials 18:164–167, 18:169
rationale for using in clinical research 18:165–166
therapeutic obligations 18:166, 18:168–169
Placebos, biomedical research ethics 8:80–81
Placelessness 18:149
Placenta 23:559–560
Placentophagy 3:86
Place-renewing leadership 20:565
Places 9:689–692, 18:149, 18:597–601, 19:802, 24:926
aging in 6:884–886
as assemblage 18:151
and crime 23:145–146
and cultural geography 5:411–412
cultural turn 18:149–150
gender and 9:709–714
in geography 18:149–152
in health 10:628–629
human–environment relationships 11:406
humanistic geography 18:149
and libraries 14:49
and liminality 14:133–135
and linked lives 14:231–232
new cultural geography (2000s–Present) 18:150–151
of nostalgia 17:35–36
and older people 1:409
political geography 18:375–376
radical geography 18:149–150
and regional geography 20:123–125, 20:127
social movements 22:438
spatial science 10:86–87
and territory 24:224
urban sociology 24:929–930, 24:932
youth 25:826
Place-value system 14:781, 14:781f
Plagiarism 20:530, 21:262, 21:313–317
Plagues 15:121, 18:174–179
ancient and medieval times 18:175–176
contemporary times 18:177–179
in history 18:174–179
modern times 18:176–177
Plain egalitarian value 4:869
Plain errors 1:823
Plain folks 4:16, 25:397
Plain folks appeal 22:352
Plain juxtaposition 18:631
Plainness penalty 2:384–386
Plaintiff's attorney 16:420–421
Plaintiff's lawyer 16:420–421
Plan, do, study, act (PDSA) improvement cycle 11:700
Plane surveys, in archaeology 23:744
Planetary temperature 9:567
Planetary urbanization 24:946
PlanIT Valley 6:836
Planned behavior, theory of 21:516, 21:518
Planned economy 22:844–845
Planned Lifetime Assistance Network 15:218
Planners 18:203
Planning 8:492, 10:713–714, 19:68–69, 22:840
action planning 19:69–71
careers 3:130–131
cognitive function 6:725
event planning 19:70–71
experimental designs 8:537
financial 4:929–930
focus groups 9:276–277
Jacobs' contributions 12:777–780
knowledge representation 13:100
land use 24:909–910
motivational interviewing (MI) 15:911
regional 20:141–145
skills of teachers for open education 17:229
in socialism 22:849–851
sustainability 18:198–201
and time use 24:373, 24:376
urban 12:776
Planning, Attention Simultaneous, and Successive (PASS) theory 12:285
Planning, Program, Budgeting (PPB) 2:902
Planning and development 18:180–186
collaborative planning 18:184–185
development studies 18:180–185
global south 18:180–185
governance 18:181–182
informal settlement 18:185
intellectual influences 18:181–184
cities and development 18:182
development studies 18:181–182
planning 18:182–184
modernization 18:181–184
participatory development and planning 18:184–185
policy 18:181
postdevelopment 18:182, 18:184
public participation 18:184
urban planning 18:180, 18:182–185
Planning evaluation, logic models 14:303–304
Planning history 24:816
Planning Institute of Australia' Code of Ethics and Professional Conduct 18:195–196
Planning methods *see* Competitions; Exhibitions
Planning methods and technologies 24:889–893
evolution of 24:889–890
planning as applied science 24:889–890
planning as communication 24:890
planning as design 24:889
future of 24:891–892
today's methods 24:891
today's technologies 24:890–891
Planning paradigms 18:193
Planning processes, sustainability 18:199
Planning support systems 24:890–891
agent-based models 24:891
CA models 24:891
Internet 24:891
large-scale urban models 24:890–891
rule-based models 24:891
Planning theory 19:821
cities, governance, and planning 18:202–205
coevolution of place quality and governance capacity 18:205
collaborative planning 18:204–205
communicative planning theory 18:204
in context 18:205–206
contribution of 18:206–207
good city 18:202–203
the good city 18:202
good government 18:203–205
governance capacity building 18:205
ideas and actions 18:205
interpretive analysis 18:204, 18:206–207
rational planning 18:204–205
spatial planning 18:202
substance and process 18:205–206
sustainable cities 18:203
system maintenance/transformation 18:206
theory and practice 18:203, 18:205
transformative initiatives 18:205
Plano, Jack, works, *Political Science Dictionary* 7:591
Plans for Progress 1:262–263
Plant domestication 16:840
Plant ecology 22:264
Plant geography, von Humboldt 25:279
Plant location decisions, supply chain management 23:710

Plantation societies, in Caribbean 1:638–639
Plantations 22:55
Plant-based diets 1:568–569
Planum polare 13:370
Planum temporale (PT) 13:370, 13:403
Plasmodia 18:175
Plasticity 2:351, 7:848
  and Ca^{2+} channels 12:710, 12:712–714
  and cerebral cortex 3:338
  versus flexibility 7:138
  human development 11:274
  and human development 11:277–280
  kindling 13:32
  and learning 9:881–882
  and life span development 14:101–102, 14:104–105
  and memory/learning 23:873–875, 23:877–878, 23:875f
  in multisensory integration 16:100–101
  and neonatal social isolation 16:489, 16:491
  neural, developmental neuroscience 16:603–604
  in occipital lobe 17:131
  and successful aging 23:669–671
  vestibulo-ocular reflex 25:72–73
Plate tectonics 15:557
Platen, August von 9:637–638
Plater-Zyberk, Elizabeth 16:809
Platform against Poverty and Social Exclusion 18:751
Platform programs 18:417–420
Platforms
  business *see* Business platforms
  providing legislative history 5:808
  remote sensing 10:65–66
Plato 3:707–708, 3:774, 4:269, 7:255, 8:237, 11:885, 11:890, 13:437, 14:252, 18:208–213, 19:407, 20:285, 25:598–599
  dialectic and forms of 18:210
  epistemology of 18:211
  on ethics and politics 18:211–212
  on governmentality 10:311
  on hermeneutical complications 18:209
    critique of writing 18:209
    dialogues 18:209
  on humor 11:416
  on mass society 14:743
  on meaning of philosophy 18:209–210
  Neoplatonism 18:208
  on neurolingustics 16:680–681
  on new society 9:861
  notion of asceticism 2:60
  ontology 18:210–211
  on personality 17:906–907
  on philosophy 18:31–32, 25:598
  pre-Platonic understanding of philosophy and wisdom 25:598
  on psychology 19:406–408
  on republicanism 20:477
  on sophism 18:208
  strategic implications 25:598–599
  on universals 11:816
  on unwritten doctrine 18:212
  works
    Academy 3:364
    Dialogues 6:589
    The Laws 12:932, 13:799, 17:907
    Phaidros 18:32
    The Republic 6:592, 6:601, 12:935, 13:246, 13:799, 14:252, 14:743, 22:878
    The Statesman 10:311
    Timaeus 17:907, 20:285
    Timaios 18:32
Platonism 3:774, 18:208
Platonists 17:97
Platt, Jennifer 12:670
  on interviewing 12:670
Platt, Tony 22:386–387
Plausibility 21:624
Plausible reasoning 3:249
Plausible rival hypotheses 3:69
Play (game)
  anthropology of 18:214–217
  attributes of 18:215–216
  behavior 18:215
  and childhood learning 3:455
  and cosmology 18:216
  during early childhood 18:218–225
  and early friendships 9:449–450
  entertainment 7:669, 7:672
  exploratory play 18:221–222
  and foundation-for-form 18:215
  framing 18:214–215
  game and 18:216–217
  gender differences in 3:453
  independent cultures 18:218, 18:220–224
  during infancy 18:219–221
    differences in appropriateness of play 18:219–220
    differences in object-oriented play 18:220
    differences in parental investment in play 18:219
    differences in physical play 18:220–221
    differences in social-oriented play 18:220
    similarities in game playing 18:219
    similarities in person- and object-oriented play 18:219
  interdependent cultures 18:220–223
  joint attention 12:845–849
  nonsocial play 18:223
  parental beliefs 18:223–224
  perception and practice of 18:215
  playful spaces 3:494
  during preschool age 18:222–224
    differences in cultural goals and practices 18:224
    differences in ecological conditions 18:223
    differences in interpersonal contexts 18:223–224
    similarities in nonsocial peer play as maladaptive 18:223
    similarities in object use, pretend play, and sex segregation 18:223
    similarities in role of individual characteristics 18:223
  and reality 18:214
  sociodramatic play 16:514–515
  sport and 23:302–303
  during toddlerhood 18:221–222
    differences in adult support 18:221–222
    differences in caregiver–toddler play preferences 18:222
    differences in parental scaffolding of object play 18:222
    differences in play partners 18:222
    differences in play themes 18:222
    similarities in effects of adult play scaffolding 18:221
    similarities in emergence of symbolic play and peer play 18:221
    similarities in gender differences and mutual specificity in caregiver–toddler play 18:221
  truth-values 18:216
Play partners, differences, during toddlerhood 18:222
Play themes, differences, during toddlerhood 18:222
Play-based learning 4:38, 4:40–41
Player-centric mapping 4:868
Playfairs 17:489–490
Playful 18:215
Playing with others 19:234–235
Play-mass 18:215
PLC *see* Phospholipase C (PLC); Professional learning community (PLC)
PLE *see* Public legal education (PLE)
Plea bargaining 5:216, 10:454, 18:226–231, 19:213
  adversarial procedure 18:230
  cases, early disposals 18:229–230
  comparative 18:226–227
  criminal procedure 5:231
  objections to 18:227–229
Plea bargains 21:630–631
Plea for rationality 5:272
Plea-bargaining process 12:667
Pleadings Game 13:790
Pleasantness 7:483
Pleasantness of taste of food 24:27–28
Pleasure 17:952
  brain stimulation reward 2:843
  and food preferences 9:299
  individual craft production for 5:112
  in leisure 13:847
  sexual 11:193–195
  and sexuality 10:816–817
  utilitarian social thought 24:980–982, 24:984
  value pluralism 25:21–25
  and wanting/needing 25:352–354
  welfare as 25:506
Pleasure-Arousal-Dominance model (PAD) 7:520–521
Plebeian Party 18:398
Plebiscite 16:284
Plebs 22:887
Plechanov, G.V., works, *The Role of the Personality in History* 2:645

Pledge 21:401–404
  security rights 19:193
Pleiotropy 9:885
Pleistocene, Australia and New Guinea 2:229–233
Pleistocene foragers 8:401–402
Pleistocene New Guinea 2:230
Plekhanov, Georgii V. 14:645, 22:834, 22:839
Plenary adoption 1:136
Plentitudo potestatis 1:9
Pleonexia 3:15
Plessner, Helmuth 2:729, 19:314
  on humor 11:416
Plessy v. Ferguson 3:691–692
Pliny the Elder, works, *Natural History* 7:584
PLM *see* Pseudologistic model (PLM)
PLO *see* Palestine Liberation Organization (PLO)
Ploetz, Alfred 8:241–242
Plomin, R., temperament development approach 24:191–193
Plot 16:194
Plot analysis 16:196
Plow, horticulture 11:204
PLR *see* Patient-led research (PLR)
PLS *see* Partial least squares (PLS)
Plugging in action 23:414
Plug-in maximum a posteriori (PIMAP) 23:262
Plural median 19:713
Plural societies
  multiculturalism 16:30
  reasoning in 19:595–596
  societal cohesion in 16:30
  in Southeast Asia 23:71
Plural society theory 1:640
Pluralism 3:540–541, 3:543, 11:556, 12:369, 12:374–375, 18:232–242, 18:467, 18:469, 20:399, 21:306–307, 25:346–349
  agonism 1:446–447
  American 18:233
  approach to American society 1:620
  associations 18:232–234
  biomedical research ethics 8:78–79
  and conflict/consensus 4:590–591
  corporatism 18:234
  critique of 18:237
  and cultural rights/defense 5:492, 5:496
  culture 18:241
  Dahl on 5:656, 5:659
  Dahrendorf on 5:661–662
  empirical and normative elements 18:236
  English 18:232–233
  and gender 9:783–784
  health and illness 10:569
  human rights 18:238, 18:241
  interest groups 18:233–234
  legal 13:539–540
  see also Legal pluralism
  and liberal democracy 18:237
  in mathematical models construction 14:794
  minority-group status 15:572–573
  and morality 15:807
  and multiculturalism 16:24
  and policy networks 18:315, 18:317
  political agendas 1:362–363
  and political power 18:761
  political sociology 18:473
  power 18:233–235
  of procedures in anthropological comparison 4:412–413
  psychic 5:479
  reasonable 19:592
  reasoning 19:591–592
  relative 18:237–238
  and religion 3:576, 3:578
  religious 10:215
    Arab nations 16:376
  romantic *see* Romantic pluralism
  in Southeast Asia 23:71
  sovereignty 18:232, 18:234–235
  state 18:232–235
  theoretical *see* Theoretical pluralism
  value pluralism 25:21–26
  and welfare 25:510
Pluralist and hybrid theories of property rights 19:169
Pluralist community work 4:379
Pluralist elite theories 7:392–393
Pluralist model 12:369
  of power 18:761
Pluralistic ignorance 21:877–879
pluralistic naturalism 16:327–328
Pluralistic protective belt 8:473
Pluralistic societies, and group violence 14:704
Pluralistic strategy for social action 1:77
Plurality 7:350–351, 21:533, 23:798
  and postmodernism 18:685–686
Plurality bloc voting, apportionment 7:344
Plurality rule 19:208
  in elections 18:865
Pluralization, agonism 1:446–447
Plurilatera trade negotiations 12:571
Plutarch 23:881
  works
    Bioi Parallelloi 2:644
    De musica 16:155–156
Plutonium experiments 13:562
Plymouth (Exclusive) Brethren response to hope 15:531
PMC *see* Premotor cortex (PMC)
PMd *see* Dorsal premotor cortex (PMd)
PMDD *see* Premenstrual dysphoric disorder (PMDD)
p-median location problem 23:174t
PML *see* Progressive multifocal leukoencephalopathy (PML)
PMM *see* Parallel measurements model (PMM)
PMP *see* Passive motion paradigm (PMP)
PMS *see* Premenstrual syndrome (PMS)
PMSAs *see* Primary metropolitan statistical areas’ (PMSAs)
PMT *see* Protection motivation theory (PMT)
PMv *see* Ventral premotor cortex (PMv)
PNAS *see* Proceedings of the National Academy of Science of the USA (PNAS)
Pneumonia, child mortality cause 12:23
PNG *see* Papua New Guinea (PNG)
PNI *see* Psychoneuroimmunology (PNI)
PNM *see* Prescriptive normativity of meaning (PNM)
PNS *see* Peripheral nervous system (PNS)
Poaching (hunting) 5:160
Poaching (media) 2:217
POAs *see* Property owners associations (POAs)
Pocketbook economic evaluation, during elections 7:27
Pocketbook on Cultural Statistics 5:708
Pocock, J.G.A. 3:62, 3:643, 7:653
  works
    The Ancient Constitution and the Feudal Law 3:62–63
    The History of Political Thought: A Methodological Enquiry 3:62
    The Machiavellian Moment 3:62–63
Pocomania cults of Jamaica 15:532
Podcasts 6:897–898
Podiebrad, George 4:572
POET (Population, Organization, Environment, and Technology) model 22:193–195, 22:722
Poetic functions 13:266–267
  of language 12:783
Poetic melody 13:276–277
Poetic mode of representation (film) 6:621
Poetic rendition, in qualitative research 19:687–688
Poetics 5:558
Poetry/poetics 12:783–784
Pogge, T., works, *Realising Rawls* 19:917
Pohang Iron and Steel Company (POSCO) Visiting Fellowship Program 22:638
Poiesis, cybernetics 5:649
Point estimators 8:16–17
  Bayes's 8:17
  evaluation 8:17–18
  maximum likelihood 8:16
  method of moments (MOM) 8:16
Point model of scientific discovery 21:281–282
Point of orientation in human life 2:646
Point of subjective equality (PSE) 24:308
Point of view, individualistic 15:304–305
Point pattern analysis 23:181
Point prevalence 7:825
Point process, times between events in 6:573
Point symbols, representations using 24:263–265
Point-based system, and international immigration 12:511
Pointillist approach 21:239–240
Point-in-polygon function 10:59
Poisson assumption test
  EDF tests 23:398
  Neyman smooth tests 23:400
Poisson counter model 6:399
Poisson counts 24:360

Poisson data, generalized linear models for 17:259
Poisson distributions 6:581, 21:949, 23:467–468
  loglinear models 16:125–129
Poisson model 23:467–468
Poisson processes 6:573, 21:334–335
Poisson regression 16:135, 23:774
Poker face 7:485
Poland, adult mortality in 1:162–165, 1:167–168, 1:754, 1:164f–165f, 1:167f
Polanyi, Karl 5:535, 6:276, 7:14, 22:849–850
  biography of 18:243–244
  on economic anthropology 18:245
  on economic history 18:243–245
  on economic sociology 18:244–245
  on embeddedness 18:244–245
  on exchange 18:244
  on forms of integration 18:244
  influence 18:244–245
  instituted process 18:244
  on markets 18:244–245
  on reciprocity 18:244
  on redistribution 18:244
  on societies 18:244–245
  Speenhamland law of 18:244
  substantive view of 18:244
  works
    Livelihood of Man 18:243–244
    Trade and Market in the Early Empires 18:245
Polanyi, Michael, on tacit knowledge 24:4–7
Polányi Károly *see* Polanyi, Karl
Polar classes 3:745–746
Polar coordinates 14:284
Polar partitioning, of multidimensional scaling 16:44f
Polarity profile, semantic differential of 21:563–564, 21:567, 21:564f
Polarization 21:533
  and ethnic conflict 8:91–93
  group 10:426
  political representation 18:441, 18:444
  and political trust 24:661
  politics 14:717
  structural balance theory 23:606–608
  of United States Congress 4:652–653
Polayk-Ruppert averaging 23:476
Police 18:246–250
  sociology of 18:251–254
Police agencies 18:263
Police data on crime 5:713–714
  problems with 5:713
Police force, and crime 5:198–199
Police imagery 18:247–248
Police reports, about criminal victimization 5:720–721
Police Service of Northern Ireland (PSNI) 18:269–270
Police–crime relationship 6:263
Polices 4:154, 5:717–718, 7:82
  apprehension risk 6:263
  certainty 6:260–261
  and crime policy 5:176–177, 5:180
  and crime prevention 18:878–879
  definition 18:247
  deterrence concept 6:260–261
  deterrent effect of 6:262–264
    aggregate police presence and crime 6:262–263
    police deployment and crime 6:263–264
  forensic genetic databases for 9:341–342
  hot spots 6:263–264
  incapacitation 6:260–262
  as organization 18:247
  problem-oriented policing 6:263–264
  and prosecutorial misconduct 25:788
  severity 6:260–261, 6:263–264
  use of force 18:255–259
    authority maintenance rituals 18:258–259
    coercion 18:258
    context 18:256–257
    determining excessive force 18:255–256
    integrated theory 18:259
    propositions 18:258
    public safety 18:255
    technological advancements 18:257–258
    violence 18:255–256
Polices, sociology of 18:251
  comparative research 18:253
  core *versus* fringe imagery 18:247–248
  determining subject 18:251–252
  emphasis 18:252
  focus 18:252
  future directions 18:254
  future of 18:250
  and metaphors 18:249
  methodological issues 18:253–254
  methods 18:249–250
  modes of 18:248
  as organization 18:247
  problematics 18:249–250
  role of 18:247
  role of external factors 18:249
  scholarship 18:246–247
  and sociological concern 18:251
  theory and research 18:252–253
  types 18:248–249
Policies 7:778, 10:864, 11:765, 18:181, 24:908
  adoption 1:140
  age stratification 1:334–336
  analysis, urban 24:802–803
    globalization and 24:802–803
    government devolution 24:803
  animal rights in research 1:704–705
  autism spectrum disorders 2:279
  cancer screening programs as 3:79
  choice probation 19:47
  class size 3:751–753
  community care
    in England 4:345
  congestion 24:633
  consumer 4:733–737
  cultural *see* Cultural policy
  in culturally diverse societies 11:679–680
  culture of poverty 5:592–593
  defining crime 5:177–178
  disability 6:476
  divorce-related 14:520
  early childhood care and education (ECCE) 6:739–741
  ecology and health 6:878–881
  economic integration 12:576–577
    depth of integration 12:576–577
    policy reform and commitment 12:577
  economic transformation 7:75–76
  educational 25:481–490
  effects on immigration 15:382–383
  environmental, in Africa 1:298
  epistemic communities 7:862–863, 7:865
  epistemic cultures and 7:877
  and equity in education 7:920–922
  evidence-based policy making
    capacity 18:281–282, 18:284
    decision making 18:281–282, 18:286–287
    development 18:281–282
    evaluation 18:282–284
    knowledge 18:281–285
    learning 18:285
    skills 18:281–282, 18:284, 18:286
  family 8:799
  fertility 9:23
  fertility transition, in South Asia 9:82–83
  gender 6:793
  and gender variant individuals 24:550–551
  global, influence The Agenda on 10:147–148
  governance in world of 10:293–297
  health 4:400–403, 10:618–619
  higher education 10:862–867
  homelessness in United States 11:161–162
  household economics 11:225
  housing 11:243–244
  immigrant 12:516
  immigration 11:641–642, 11:644–646, 11:662–671, 15:460–461
    and xenophobia 25:805
  implementation 11:694
  implications of poverty 18:720
  industrial *see* Industrial policy
  informality 12:57–58
  and information society 12:84–85, 12:87–88
  integration of ELSI program with 8:44–45
  of internal migration 12:445–446
  interventions, antipoverty effects of 18:745
  issues, AIDS 1:493–494
  labor market *see* Labor market policy (LMP)
  language 22:952–953
  learning disability 13:686
  against LGBTQ discrimination in schools 21:88–89
  manifestos 18:417–420
    equilibrium 18:417
  material flow analysis 14:763
  media 2:860, 2:862, 15:57
  media imperialism 15:56
  mental health 15:190–191, 15:196–197, 15:216

migration and integration, in Europe 8:143
nonprofit institutions in United States 18:320–325
mixed welfare state 18:324–325
social investigation in nineteenth century 18:320–321
think tanks 18:322–324
twentieth century 18:321–322
older age work participation
recommendations 17:188
research on 17:187–188
peer review 17:685, 17:687
pension 17:690–696
political science 18:460–461, 18:463–464
population aging 18:543
poverty policy *see* Poverty policy
and poverty 5:591
and principle, distinction between 20:683
for promoting sustainability lifestyles 23:831
public 19:569
public health 10:650
and reentry after incarceration 20:61, 20:64–65
refugee 20:105–107, 20:109
role in community-based health interventions 10:648
rural development 1:458–459
rural social work 20:808–809
sexual harassment 21:725
social exclusion/deprivation 22:289
sustainable development 23:826–827
sustainable production and consumption 23:840–841
telephone 24:159
tobacco control policies 22:128–129
trade 24:485–490
and UK psychiatric social work 19:301–302
usefulness of experimental economics in progressing 8:550–551
victim policy, developments in 1:836
vocational education and training 25:249–250
for WCP exclusion in clinical trials 9:720–721
women and criminal justice 25:619–620
Policies, of neoliberalism
consumerism 16:464
global capitalism 16:464
promotion of self-reliance 16:464
socioeconomic 16:464
Policing 18:251–252, 18:260–265, 19:641, 23:730
community and culture 18:261–262
cooperative efforts among police 18:264
and cultural aspects of police organizations 18:262
degree of professionalization 18:261
democratization processes 18:264
in developing nations 18:263
for fommunity 18:253
functions and organization 18:261
globalization of 18:263–264
history 18:260–261
modern 18:261
motherhood 15:883–885
multiple forms 18:264
research 18:252–253
sociology of 18:252–253
strategies for 18:262–263
technology and 18:262–263
and transnational organized crime 24:598–600
in United Kingdom 18:261
in United States 16:251, 18:261
urban violence 24:955–956
use of term 18:260
Policing, in divided societies 18:266–272
'alternative justice' in divided societies 18:268
and authoritarian regimes 18:267–268
features of 18:266–268, 18:267t
patten report 18:271, 18:269t
political conflict 18:266, 18:268, 18:271
political transition and police reform 18:268–271
dealing with past 18:271
police reform and internal agreement 18:268–270
security sector reform and external intervention 18:270–271
Policing, types of 18:248–249
formal 18:248–249
hybrid 18:249
informal 18:248
private 18:248
public 18:248–249
Policing Board 18:269
Policing borders 25:125–126
Policing of white-collar crime 5:204
Policing Ombudsman 18:269
Policing policies and practices 5:145–146
Policing strategies, and crime 5:198–199
Policing web 18:249
Policy agenda 7:268–269
Policy analysis 18:273–280, 19:570, 19:576–577, 24:294–297
client *versus* societal focus 18:274t
definition 18:273
evidence-based policy making 18:281–287
goals 18:273
instrumental goals 18:275, 18:274t
modes 18:275–279
efficiency analysis 18:277–278
incomplete analyses 18:278–279
multi-goal analysis 18:275–277
political feasibility 18:273, 18:276, 18:278
problems framing 18:274–275
public policy 19:570–572
substantive *versus* process focus 18:274t
use 18:279
Policy approaches
for physical activity 18:97
concepts and principles 18:94
Policy behavior styles in the policy cycle 18:288–291
agenda setting 18:288–289
decision making 18:290
policy evaluation styles 18:291
policy formulation 18:289–290
policy implementation styles 18:290–291
Policy change, policy networks and 18:316–317
Policy communities 12:769–773
and policy networks 18:315, 18:317, 18:319
and political agendas 1:363–364
Policy community concept 24:297
Policy curriculum 19:577–578
Policy cycle model of the policy process 18:288
Policy design 7:269–270
Policy development, SES and social-emotional development of children 22:946–947
Policy entrepreneurs 7:268
Policy evaluation 18:278–279, 18:288, 20:148
economics of expectations 8:512–513
Policy feedback, and Skocpol, Theda 22:33–35
Policy for science 9:552
Policy formulation 18:288
Policy implementation 7:270, 18:288, 19:584
Policy implications
criminal networks 16:579
for SES impact on childhood behavior problems 2:479–480
Policy initiation 10:314
Policy issue networks, global 10:158–159
Policy issues, child care 3:393
Policy knowledge, foundations 18:294
foundations 18:293–298
and American university system 18:295–296
debates 18:294
emergence of new philanthropy 18:294–295
Europe and United States, comparison 18:297–298
European Foundation Center (EFC) 18:297
and international development 18:296–297
public policy 18:294
Rockefeller Foundation 18:295
roles and impacts 18:293–294
roots 18:293
new liberalism 18:299–306
universities 18:307–313
venture philanthropy (VP) 18:295
Policy learning 24:296
Policy mainstreaming 6:832–833
Policy making
adjudication 5:103–104, 5:106
civil service 3:697–699
issue networks 12:769–773
Policy mobility 24:797–798, 24:850–851
Policy model, legal reasoning 13:781–782
Policy monopolies 1:359, 1:361
Policy networks 12:769–773, 18:314–319
advocacy coalition 18:315
corporatism 18:315–317
future challenges 18:318–319

Policy networks (*continued*)
global governance and 18:317–318
definition 18:317
global policy 18:317–318
nation-state policy making 18:314–315
historical development 18:315
pluralism 18:315, 18:317
policy communities 18:315, 18:317, 18:319
polycentric governance 18:316–317
private actors 18:315
theoretical status of concept at nation-state level 18:316–317
policy networks and policy change 18:316–317
policy networks as metaphors only 18:316
Policy planning 18:203–204
Policy process 7:782, 12:368–369, 23:508
agenda setting 7:782
consideration of policy options 7:782
decision-making 7:782
evaluation 7:782
implementation 7:782
problem emergence 7:782
public policy 19:569, 19:571–574
Policy Rainbow 7:741
Policy research 18:273
Policy responses, in Australia for disasters 7:428
Policy Review 1:842
Policy science, public policy 19:570–572
Policy windows 1:358–359
Policyholder 12:210
Policymakers 8:531–532
Policymaking
and administrative law 1:112–113
and political power 18:761–762
Policymaking discourse 7:215
Polis 13:436–437
Polish 13:290–291
Polish Law and Justice (PiS) 18:611
Politeness 19:231
accounting for 18:326–329
definition 18:326
Grice's Cooperative Principle 18:326–327
in language 18:326–330
accounting for 18:326–329
B&L politeness model 18:328–329
broad *versus* narrow scopes 18:328
conversational maxim approach 18:327
cooperation 18:326
face-threatening acts 18:327–328
implicature 18:328
indeterminacy problem 18:329
indirectness in language 18:326–328
negative politeness 18:326–328
new directions in research 18:329
point of view of individual, dyad, or social group 18:328–329
politeness as communicated or taken for granted 18:328
positive politeness 18:326–328
social relationship 18:326–330
speech act theory 18:327–329
universal principles 18:327–328
universality of face 18:328
Political action 25:346
Political Action project 18:505
Political activism, and child soldiers 3:441–443
Political actors 17:650
role in internal migration 12:437–438
Political advertisements 25:172
Political advertising 18:331–334
new directions 18:334
and vote production 18:331–333
voter autonomy 18:333–334
Political advocacy, mobilizing rights in 20:684–685
Political agendas 1:362–366
budgeting 1:364–365
conflict expansion 1:362–363
elitism 1:362–363
framing 1:364
issue definition 1:364
literature development 1:363
pluralism 1:362–363
policy communities 1:363–364
and political power 18:761–762
power 1:362–364
public policy 1:363–364
punctuated equilibrium 1:363
recent extensions 1:364–365
social movements and media 1:364
and writing of Jewish history 11:68–69
Political anthropology 1:681–682, 2:871–872, 8:290–291, 8:683–684, 10:309, 18:335–341
and history 18:337–338
and liberal political philosophy 18:339–340
making of 18:336–337
and nongovernmental organizations 16:856
processual approach of study 18:336–337
and state 18:338–340
statehood 23:347–348, 23:350–351
and violence 18:340
Political approach 8:126
Political arithmetic 7:559, 19:703, 20:875
Political artifacts/systems, technological determinism 24:111
Political aspects, of immigration 11:669–674
Political attitudes 2:262–263, 5:827, 18:459, 18:462–463
and political socialization 18:509
Political authority
political sociology 18:473
state/society 23:343
Political autonomy 2:314
Political behavior 5:824–829, 18:459–464
data infrastructures for research into 5:826–828
implications and challenges 5:828
mass media and 14:712
media-centered 14:714–715
Political being 23:363–364
Political belief system (PBS) 11:555, 11:557
Political beliefs 18:429
Political borders 2:761–762
Political campaigns 3:66–68
choice of candidates 3:66–67
decision to run 3:66
by parties 3:66–67
communicating with voters 3:67–68
free media coverage 3:67
paid media advertising 3:68
defining messages and setting strategies 3:67
organization 3:67
Political change
democratic transitions 6:97–99
psychological perspective of 22:147–148
Political channels, unconventional 22:508, 22:511
Political charisma 3:355–356
Political citizenship 3:650–654, 20:746
Political commitment, MENA 16:364
Political communication 16:815, 18:331–333, 18:342–349, 18:430–431, 18:463–464
broadcasting trends 18:345
consequences for democracy and democratization 18:347–348
constructivist, critical, and rhetorical approaches 18:430–431
contents 18:343
effects 18:343
and media 22:585–586
media communication 18:430
populism 18:609
postwar trends in newspapers 18:343–345
production processes 18:342
rise of Internet 18:345–347
Political community 3:707
political thought as constitution of 18:481
Political competence, and organizational conflict 4:642
"The political" concept
agonism 1:446–448
Political conflicts 5:837–838, 8:97–98, 8:98t
concepts of specific forms of political violence 5:838
and revolution 20:647
scope 5:838–839
Political conservatism, of cultural heritage 5:415
Political constitution and system of government 20:856–858
Political consumerism, new patterns of 4:735–736
Political contestation, and collective behavior 4:160
Political contexts
capitalism 3:107–108
informal sector 12:58
social work with refugees 22:770
for universities in twenty-first century 18:308
Political control, and Chinese cultural policies 5:600–601
Political controversies 21:283

Political correctness, moral entrepreneur for 15:798
Political cosmopolitanism 17:652
Political crimes 5:165
Political cultures 5:617, 5:619, 17:519
of opposition 20:645
Political cynicism, and political efficacy 7:302–303, 7:305
Political decision, in universal access to research products 8:34
Political decision-making, cognitive approaches to 18:429–430
Political destiny 16:290
Political development, and citizenship 3:655–656, 3:658–660
Political deviance 25:722–723
Political discourse 18:331, 18:333, 20:652–653
Political ecology 6:24–25, 6:290–292, 16:337, 18:600
anthropology of 18:350–355
actors and perspectives 18:352–353
agency 18:351, 18:353–354
capital accumulation 18:350, 18:352
commodification 18:352–353
definition 18:350
ecosystems and world systems 18:350–352
environmental problems 18:352
future prospects 18:354
geography 18:350–354
land degradation 18:352
marginalization 18:350, 18:352
nature 18:350–354
against political ecology 18:353–354
politics 18:354
power 18:350, 18:353–354
and development geography 6:317
of disease 15:89–90
environmental degradation 7:714
feminist 6:848
Gorz on 10:287–290
of health 7:817
Political economy 7:69, 7:93, 7:817–818, 13:161, 14:733, 23:82, 25:662
in anthropology 18:362–366
area and international studies 1:938–942
BRICs 1:941
economic organizations 1:940
Fordist movement 1:938–939
import-substitution industrialization 1:939
international relations 1:939–940
modernization 1:938–939
Neoliberal movement 1:939–941
post-2008 political economy 1:941
trade liberalization 1:938, 1:940
world systems theory 1:940
and articulation 18:364
in Australia 2:242–243
Beck on 2:397–398
of biocitizenship 2:611
and biomedicine 2:666–667
of celebrity 3:276
Hollywood star system and beyond 3:276
production in twenty-first century 3:276
chiefdoms 3:379
and chiefdoms 3:379
of citizenship 20:895
and criminology 5:244–246, 5:248
cultural analysis 14:735–736
Dahl on 5:655–657
declining popularity 18:364–365
and development geography 6:317
economics and 14:734
environmental sociology 7:799–800
French Marxist heritage 18:363–364
and global capitalism 18:362–363
of growth and distributive justice 20:895
historical and comparative 18:365–366
history of 18:356–361
Austrian economics 18:360
cameralism 18:357
capitalism 18:358, 18:360
classical political economy 18:358–359
division of labor 18:358
early modern economy 18:356–357
early neoclassical economics 18:359–360
economic theory 18:357–359
eighteenth century 18:357–358
institutionalism 18:359
invisible hand 18:358
Marxist economics 18:360
mathematical economics 18:359
mercantilism 18:356
monetary theory 18:357–358
nineteenth century 18:358–359
physiocracy 18:357–358
premodern economic thought 18:356
wealth 18:356–360
of hope 2:611
and industrial societies 11:863
institutionalizing public interest 14:734–735
of local and state taxation 23:341–342
Marx on 14:642, 14:644–647
in Mexico 1:627–628
Mill on 15:525–526, 15:528
and neoliberalism 18:365
Olson on 17:209
post-socialist 18:690–696
and radio 19:878
of regulation 20:178–180
of renewable energy 18:367–373
and reproduction 20:452
in rural geography 20:795–796
Smith on 22:118–119, 22:121
and social science 12:388
of socialism 22:141
of socialist societies 22:849–851
sociology 7:68–69
of tourism 24:470–471
and utilitarianism 12:388
in women's exclusion in clinical trials 9:721
work anthropology 25:662–665, 25:667
Political efficacy 7:301–307
antecedents of 7:303
campaigning 7:302–304
concept of 7:301–302
external 7:301–306
future research 7:305–306
internal 7:301–306
measurement 7:304–305
media effects on 7:303–304
political cynicism 7:302–303, 7:305
political participation 7:301–306
political trust 7:302–303
and reciprocal relationships 7:303–304
relevance of 7:302–303
subjective competence 7:301
survey items 7:301–302, 7:304–305
system responsiveness 7:301–303
voting 7:301–302, 7:304–305
Political elite group 15:536
Political elites, African, and social movements 1:287
Political engagement, and child soldiers 3:440, 3:442–443
Political ethics
and communitarianism 4:309
public policy 19:583
Political faction 8:683–684
Political field, social rules in 4:3–5
Political freedom 9:398–403
Arendt on 1:962–963
civic republicanism 9:402–403
constitutive notion of 9:402–403
instrumental notion of 9:402
Pettit on 9:402–403
Skinner on 9:402
Cohen on 9:400–401
economics and power 9:400–402
freedom as attaining autonomy through community 9:401–402
libertarianism 9:400
liberty, equality, fraternity 9:398
negative liberty 9:398–399, 9:402
neo-Aristotelian freedom as capability 9:401
neo-Marxism 9:400–401
Nozick on 9:400
positive liberty 9:398–399
post-Marxism 9:401
Raz on 9:401–402
right to own private property 9:400
Sen on 9:401
Two Concepts of Liberty 9:398–399
van Parjis 9:401
welfare liberalism 9:399–400
Dworkin on 9:399–400
Rawls on 9:399
Political generations
and cohort 9:861–862
cohort succession 9:862
Inglehart on 9:862–864
recent changes in postmaterialism 9:864–865
replacement, impact of 9:862
and youth movements 25:837–841
Political geography 15:502, 18:374–379, 23:83–84
apportionment 7:342–343

Political geography (*continued*)
approaches and themes 18:376–377
assumptions of political theory 18:376
assumptions of social theory 18:376
changing nature of sovereignty 18:377
deterritorialization 18:377
geopolitical conceptualizations 18:376–377
practice 18:376–377
reterritorialization 18:377
boundaries 18:377
citizenship 18:377
development in 1890s–1970 18:374–375
of environment 18:377–378
future 18:378
globalization and 23:83–84
identity 18:377
late-twentieth-century revival 18:375–376
political movements 18:378
rescaling of political practice 18:377
and youth 25:826–827
Political Geography 7:345, 18:375
Political history 18:380–385
of Southeast Asia 11:79–80
Political human rights 11:355
Political humor 11:418
Political ideal 20:675–676
Political ideas, Americaness of 21:893
Political identity 18:431–432
Political ideology 10:754, 18:494
Political implications
of disasters 22:755
legal rights 22:755
socioeconomic status and health 22:935
Political inequality 5:658–659
Political influence, on gender development 9:746
Political instability 1:557
Political Instability Task Force (PITF) 5:836
Political institutions 12:192, 18:513–514
fragmentation of 20:163
and mass media 14:712
public policy 19:569–570, 19:573–574
and welfare retrenchment 25:525
Political interventions, social work 3:471–472
Political issue, issue evolution 12:765–767
Political knowledge 11:554, 11:559, 16:794
Political languages 1:6–7, 1:9–10
Political lawyering 18:386–392
approaches to lawyers 18:390–391
approaches to politics 18:390–391
five propositions 18:387
and legal complex 18:387–389
political liberalism 18:386–387
Political lawyers 18:390–391
Political leaders
elites 7:392
personality profiles of 18:429
Political learning, and political socialization 18:509, 18:511
Political left, world systems theory 25:749, 25:751
Political liberalism 14:11, 14:19, 18:386–387, 20:894–895
civil society 18:386–387
and civil society 18:389–390
international contexts for 18:389
legal freedoms 18:387
moderate state 18:386
Popper on 18:523
and public reason 19:594
value pluralism 25:24
Political life 10:313
judicialization of 12:881
Political mobilization 16:257, 16:798–804
disputes 6:561
indigenous people 11:757–759
of religious resources 20:330
Political modernity 7:308–309
Political modernization 15:697–699, 18:474, 20:633
Political money 18:393–397
Australian High Court 18:396
campaign finance 18:393–396
democracy 18:393–396
enforcement and effectiveness 18:396
European Court of Human Rights 18:396
First Amendment, US constitution 18:396
India, campaign funding 18:395–396
institutional context 18:393–394
problems associated with 18:394
tools 18:394–396
contribution limits 18:395
expenditure limits 18:395
public disclosure 18:395–396
public subsidies 18:394–395
regulation of campaign activity 18:396
transparency 18:395–396
United Nations Convention against Corruption 18:395
United States
congressional elections 18:394, 18:396
elections 18:394–396
presidential elections 18:394–395
Political movement, Communist political movements 4:303–304
Political murder 18:446
Political norms, and political socialization 18:509
Political opportunities
intellectuals 12:280
and protest 18:423
Political order 11:433–434, 14:18–19
and eudaemonism 8:214–216
and industrial order 5:537
Political organizations
in Arctic 1:928–929
of language 22:952
in Oceania 17:160
political economy *see* Political economy
Political participation 18:504–508, 25:306–307
and citizenship 3:650–653
and class voting 25:300
conceptual issues in research on 18:506–507
leadership 18:507
outside advanced democratic countries 18:506–507
CVM 18:506
data infrastructures for research into 5:826–828
definitions of 18:504
as electoral and hierarchical 18:504–505
large-scale cross-national studies 18:505–506
mobilization as predictor of 18:506
and party identification 17:562, 17:565
and political efficacy 7:301–306
and social capital 22:178–179
see also Political behavior
Political parties 7:351, 12:368–369, 12:371, 12:373–374, 17:516–518, 18:334, 18:398–403
approaches and definitions 18:398–399
choice of candidates 3:66–67
class voting 25:299–301
communist *see* Communist parties
executive government 8:497–498
extreme right 18:410–416
female quotas 8:889–891
finance 18:393–397
issue evolution 12:765–768
and journalism 12:851–852, 12:854
manifestos 18:417–420
party systems 17:567–569
political participation 18:401–402
proportional representation 19:205, 19:208
stages of development (phases) 18:399–401
and welfare retrenchment 25:524–526
Political peace processes 17:648–649, 17:651
Political philosophy 6:485–487, 18:25, 18:486–488
and deliberative democracy 6:64–65
and public reason 19:592, 19:594
Rawls on 19:914, 19:917
Sandel on 20:893
transition to social science and history 18:482
Political philosophy, modern 17:735–736
Political pluralism 18:237
Political potential of drag 13:877–878
Political power 18:760–764
as causation 18:760–761
class system 22:339
Dahl on 18:760–764
and democracy 6:85
equality of 6:87
Foucault and 18:762–763
hegemony 18:761–762
international relations 18:763–764
international systems 18:763
Lindblom on 18:762–764
multiple elitism 18:761
neopluralism 18:761
pluralism 18:761
and policy making 18:761–762
political agendas 18:761–762
potlatch 18:716–717
power elite 18:761–763
regime theory 18:762–764

Political prerequisites of civil society 23:344–345
Political principle, equality as 7:906–907
Political process, of indigenous people 11:757–759
Political process theory 4:146
Political protests 4:145, 18:421–426, 22:507, 22:513–515
as assurance game 18:421–423
civil rights movement 18:423, 18:426
collective action 18:421–423
community 18:422
demand and supply 18:422–423
leadership 18:422
mass communications 18:422
nonviolence 18:425–426
political opportunities 18:423
power of social media 18:423–424
social media in Arab spring 18:423–424
social movements 18:424
civil disobedience 18:424–426
social networks 18:422–423
in social protest 22:507
tacit coordination using cultural resources 18:422
Political psychology 18:427–433, 18:463–464, 22:535
context 18:427–429
group-level processes 18:431–433
group identities and cultural diversity 18:431–432
international conflict 18:431
racism 18:431
social change and collective action 18:432–433
Heider on 10:754
history 18:427–428
individual-level processes 18:429–430
cognitive approaches to political decision-making 18:429–430
personality profiles of citizens and political leaders 18:429
political beliefs and values 18:429
institutionalization 18:428
interdisciplinarity 18:427
intergroup relations 18:428, 18:431–432
methods 18:428
political communication 18:430–431
constructivist, critical, and rhetorical approaches 18:430–431
media communication 18:430
political identity 18:431–432
political rhetoric 18:431
prejudice 18:427–428, 18:430–431
societal turn 18:428
Political radicalism 21:311
Political rationality, and modernity 15:689–691
Political realism 3:678
Political reform 14:251
Political regimes, democratic transitions 6:97
Political religions 18:434–435, 18:437–438, 21:380
Political representation 17:517, 25:220
accountability 18:442–443
constituents 18:441–444
deliberative democracy 18:444
group representation 18:442
identification of represented 18:442–443
inequality 18:443
origins and development 18:441–442
parliament 18:441, 18:443
performance of representative body 18:443–444
polarization 18:441, 18:444
political money 18:442
proportional representation 18:442–443
redistricting 18:443
role of representative 18:442
transnational citizenship 18:443
US Congress 18:441–442, 18:444
voting 18:442–443
Political repression 18:446–452
civil liberties 18:446–447, 18:449–451
disappearance 18:446–447
empirical studies 18:449–451
bilateral, multilateral, and global economic influences 18:451
democracy 18:449
domestic and external threats 18:449–450
domestic legal commitments and institutions 18:450
future 18:451
international treaties, norms, and INGOs 18:450–451
standard model 18:449
extrajudicial executions 18:446
of freedom of association and assembly 18:448f
of freedom of religion 18:448f
of freedom of speech and press 18:448f
human rights 18:446, 18:448–451
imprisonment 18:446–447
judicial independence 18:450
measurement 18:446–447
of personal integrity rights 18:446–447, 18:450–451, 18:447f
physical integrity 18:446–447, 18:451
political murder 18:446
theoretical approaches 18:447–449
torture 18:446–447, 18:450–451
Political research
surveys 23:748–751
using manifestos in 18:418–419
Political resocialization 18:510–511
Political revolution 11:648–649, 20:622, 20:628
Political rhetoric 18:431
Political rights
Marshall on 14:627
world systems theory 25:749
Political science 3:715–716, 5:655, 6:88, 6:487, 18:459–465, 22:679–681, 25:279
Addams on 1:95
applied social research 1:855
bachelor's degrees 22:680t
of bounded rationality 2:774–775
contributions of polling to 18:463–464
core infrastructure for 5:831–832
discipline
early years of 18:466–467
versus study of politics 18:466
doctoral degrees 22:680t
Empirical Implications of Theoretical Models 18:453–458
empirical legal studies in 7:554–556
gender in 9:791–792
Hilberg on 10:870–872
history of 18:466–471
after World War I 18:467–468
after World War II 18:468–469
American behavioralism 18:469
fragmentation 18:470
globalization 18:470–471
postbehavioral revolution 18:469–470
and immigration 11:669–670
institutionalism 12:187
master's degrees 22:680t
sanctions in 20:890
time series analysis in 18:515–521
Political science and theory 21:213
Political Science Quarterly 18:467
Political science theories, of genocide 10:15–16
Political scientists 21:119
Political sensibilities 11:109–110
Political settings, cultural policy 5:463–464
Political skill 18:771–774
Political socialization 18:509–511
adolescents 18:510–511
agents of socialization 18:509–510
children 18:509–511
initial development and ascendance of field 18:510
issue preferences 18:509
methodology 18:509–510
and party identification 17:562, 17:566
political attitudes 18:509
political learning 18:509, 18:511
political norms 18:509
political resocialization 18:510–511
resurgence and future prospects 18:510–511
topics and questions 18:509
waning interest 18:510
Political society 10:742
Political sociology 3:655–656, 18:472–479
civil society 18:472–473, 18:477
contemporary period 18:474–477
Coser on 5:37–38, 5:40
democracy 18:472–475
historical developments 18:472–473
Durkheimian tradition and pluralism 18:473
Marx and class domination, capitalism, and revolution 18:472–473
Weber and political authority, bureaucracy, and modern state 18:473
and immigration 11:669
institutionalization after World War II 18:473–474
leaders 18:473–474, 18:476

Political sociology (*continued*)
mid- and late twentieth-century period 18:474
power 18:472–477
and Skocpol, Theda 22:31
social class 18:472–473, 18:476
social movements 18:472, 18:475–476
Political sophistication
new conception of 11:559–560
reactions to Converse's formulation of 11:555–557
Political space for debate 11:575–576
Political stability 1:557, 4:692–693
Political strategies
of climate change 3:843
of communist parties 18:406–407
Political support
and philanthropy 22:580
for regulation 20:179
Political systems 8:876, 12:369, 16:820
big man system 2:567–569, 2:572
centrifugal and centripetal social gradients 3:595f
city growth and 3:596–597
concept of the city center 3:594–595
historical-political systems 3:594–595
impact of state socialism 3:595–596
impact on cities
social-ecological city models 3:594–597
land values gradient and zoning 3:596f
and social system 22:748–749
third dimension 3:596
urban fringes, development 3:596–597
Political technologies of body, and punishment 19:657–658
Political Terror Scale (PTS) 5:837, 18:446–447
Political theory 1:526, 2:716, 4:317–320, 5:665, 8:926–928, 18:381, 25:346–349
communitarianism 4:308–310
concept of interest in 12:386–387
and deliberative democracy 6:61–64
feminist 8:926–927
Shklar on 21:890–892
Political theory and ideology, Aron on 1:981–982
Political thought 3:61–63, 15:701–704, 18:381, 18:480–485
Cambridge School approach 18:487–489
conceptual history 18:489–490
as constitution of political community 18:481
definition 18:486
and early modernities 18:480–481
formation of modernity 18:482
in historical context 18:480
history 3:65, 18:486–490
Hobbes on 18:481
of Japan 14:635, 14:638
languages 18:488–490
liberal, triumph and crisis of 18:482–484
Locke on 18:481
origins of 3:64–65
Pufendorf on 18:481
renewal and critique in 18:484–485
Rousseau on 18:481
society 18:481–482, 18:484
Straussian approach 18:486–487
transition from political philosophy to social science and history 18:482
Political time 24:407
Political tolerance 18:240–241, 18:491–496
defined 18:491
democratic value of 18:491
explanatory frameworks for 18:494–495
individual-level explanations of 18:494
measures of 18:491–493
content-controlled measures 18:492–493
research on 18:493–494
future 18:495
themes 18:494–495
in United States 18:491
Political transformation 16:300
Political transition and police reform 18:268–271
Political trials 18:497–498
in established democracies 18:499–500
relevance of 18:502–503
in transition to democracy 18:498–499
in transnational legal order 18:500–502
criminal trials in international criminal courts 18:501–502
transnational tort human rights litigation in domestic courts 18:502
universal jurisdiction by domestic criminal courts 18:500–501
Political trust 24:658–663
caring about 24:662
confidence gap 24:658
confidence in government 24:658, 24:660
corruption 24:660–661
democracy 24:658, 24:661
determinants 24:659–660
economic optimism 24:659–660
incivility and gridlock 24:661
income inequality 24:661
legitimacy 24:658
measurement 24:659–662
performance of government 24:660
polarization 24:661
and political efficacy 7:302–303
satisfaction with direction of country and economy 24:659–660
social trust 24:661–662
Tea Party 24:661
winners and losers 24:661
Political values 18:429
and ideologies 11:574
Political violence 10:2–4, 22:430–436, 25:377
consequences of 22:434–435
biographical area 22:434–435
economic area 22:434
life-long consequences 22:435
political area 22:434
social change 22:434
and cyberterrorism 24:229
disengagement from 22:432–433
at macro level 22:433
at meso level 22:433
at micro level 22:433
effects on children 25:389–394
adolescents 25:389–392
aggression 25:391
armed conflict 25:389–391, 25:393
child soldiers 25:392–393
context 25:389–390
identifying risk and protective factors 25:391
intergroup process 25:392
intervention and prevention programs 25:393
mental health and psychopathology 25:390–391
resilience 25:389–390, 25:393
role of family 25:391–392
social ecology 25:390
social identity 25:392
youth as victims, perpetrators, and agents of social change 25:393
at macro level 22:431
major episodes of political violence (MEPV) 8:96–98, 8:98t
at meso level 22:431–432
at micro level 22:432
and territorial conflict 24:215–216
Political war 4:618
Political-economic approach, to regional geography 20:124
Political–philosophical stances, international social problems 12:543
Political–social instrumentalization of historical knowledge 2:907
Politicide 10:6
Politicization
of ignorance 11:582
of science 18:312
versus securitization 21:397
Politicization of collective identity (PCI) 4:147–148
Politicized culture, and economic development 5:532–533
Politicized professionalism 18:866
Politicized work 19:871–872
Politicizing of technological cultures 24:139
Politico-bureaucratic system 2:920
Politics 1:527, 6:814, 8:926, 15:518–523, 16:8–9, 18:375, 18:402, 20:667, 21:29–30
in action 18:772–773
and agonism 1:444–449
anthropological definitions of 18:335
and anthropology 1:227
appeasement 1:826–831
of archaeology 1:885–890
Canada 1:889
definition 1:885–886
future directions 1:889–890
historical background 1:886–888
indigenous peoples 1:885–890
international perspectives 1:889
key issues and current debate 1:888–889

law 1:885–889
NAGPRA (Native American Graves Protection and Repatriation Act) 1:886–889
Native Americans 1:885–888, 1:890
NMAIA (National Museum of the American Indian Act) 1:886–888
United States 1:885–886, 1:888–890
Arendt on 1:962
of art 20:706–710
and arts, relationship between 5:460–461
balance of power 2:348
behavioral approach to study 5:656–659
and biography 2:638–639
biotechnology 2:679–681
bioethics 2:680–681
risk management 2:679–680
of British cultural studies 2:854
and brokers 2:870–874
and business 3:2–3
caste-based 3:224–226
in Central Asia 3:318–319
and children/infant mortality 12:25–26
and citizenship 3:638–640
of citizenship and belonging in migration 15:423–424
and class 3:749
of commemoration 4:239–240
consumption, and commodity chains 4:245
of culture and heritage 9:283–284
and cybernetics 5:648–649
and delegation 6:58–60
deliberative 20:479–480
democracy *see* Democracy
of de-Westernization 25:546–547
diffusion 6:412
in discipline of geography 18:375
of discourse 6:505–506
displacement of 10:310–311
East Asian 6:798–804
efforts to slow down the shift to renewable energy 18:372
elites 7:386
energy development 7:622–623
enlightenment in 7:658–659
of ethnography 5:383
feminist research 9:718
and fertility transition 9:49, 9:51–53
formal modeling of 18:512–514
alternative approach 18:513–514
basic structure 18:512–513
chaos theorems 18:514
collective behavior 18:514
equilibrium 18:512–514
median voter theorem 18:513–514
Nash equilibrium 18:512–513
noncooperative game theory 18:513
optimality 18:512–513
political institutions 18:513–514
rational choice theory 18:512, 18:514
strategy 18:512–513
tragedy of the commons 18:514
voting 18:512–514
functional analysis 9:540
gay, in South Africa 21:744–746
Geertz on 9:662, 9:664–665
and gender variant individuals 24:550–553
and generations 9:870
governance in world of 10:293–297
and governmentality 10:308–312
Habermas on 10:469, 10:471–472
health care 10:664, 10:667
historical sociology 11:28–29
history 18:770
history of Roma people 20:751–752
human terrain system (HTS) 11:396–397
Humboldt on 25:282, 25:286
identity 11:532, 11:534–535
implementation 11:692
and indigenous knowledge 11:781
interest groups 12:369–370
and international advertising 12:456
and Internet 14:716–717
and interpretation 12:643–644
judicial decisions 13:605
Kantian 13:5–10
eternal peace 13:9
and right 13:8–9
right as compatibility of freedom 13:8–9
Latin American, before 1970s 22:143–144
liberal 14:27–28
liberalism 14:23–24
literature and review 18:770
and mass media 14:712–718
and media 22:583–586
media imperialism 15:56
and military 15:506–509
and morals, relation between 22:120–121
of multiculturalism 16:24–26
of multinational corporations 16:78–79
Mumford on 16:137, 16:139
against nature 16:341–342
Near Middle East/North African studies 16:371–375
object-oriented 5:525–526
in organization 18:771–772
measurement of 18:773
of organizational decision making 17:355–356
conflict resolution for efficiency and equity 17:355–356
power/efficiency controversy 17:355
peasant 17:677
and pension policy 17:693–694
of place 18:376
Plato on 18:211–212
and political ecology 18:354
political thought 18:486–490
public policy 19:573–574
queer anarchism 19:749–750
and race identity 19:833
of recognition 20:44–49
and religion 20:401, 20:404–405
republicanism 20:477–482
research on 18:774
positive effects 18:774
of responsibility 15:536–537
Russian and Post-Soviet Studies 20:843–850
science 21:239
and science 21:212–213
science communication as 21:221–222
sex work 19:281–283
and sibling order 21:923–924
in small states 22:108–112
and social and behavioral sciences 15:323
and social media 22:423–424
of social work 14:664
state 14:664
in sports 23:299, 23:302–304
study of 18:466
and telephone 24:159
Tocqueville on 24:423
of tradition 24:506–507
trans* identity 24:552–553
of truth 15:536–537
in the United States 20:279–283
as a vocation 3:351–352
and voting 25:309
Politics matter thesis 25:512
Politics of Authenticity 2:253
Politics of memory 4:204, 4:239–240
Politics of recognition, and commemoration 4:237
Politics of reform, East Asia 6:799–803
Politics of regret, and commemoration 4:236
Politics of representation, in feminist research 9:719
Politics of scale 24:946–948
Politics of science 21:207, 21:210
Wundt's contributions 25:794
Politics of time 4:181–182
Politics of truth 15:536–537
Polity 22:745, 22:747–751
Polity autonomy 15:698
Polity databases 5:790
Polity dependence 15:698
Polity dominance 15:698
Polizeistaat 20:781
Polizeiwissenschaft 22:696–697
Poll Tax 24:846
Pollen, paleoethnobotany 17:457
Pollination 23:824
Polling 18:459–465
AAPOR 23:752–753
accountability 18:463
Common Rule 23:754–755
contributions to political science 18:463–464
data quality 23:756
deliberation 18:463–464
efforts at persuasion 23:756
emerging uses of public opinion research 18:460–461
error sources 18:462–463
ethical aspects in 23:752–757
incentives 23:756
informed consent in surveys and polls 23:755
on Internet 23:756
leadership 18:460–461, 18:463
mass media 18:459, 18:461, 18:463
methodological and technical issues 18:461–462

Polling (*continued*)
obligations
to commercial or private sponsors 23:753
to the public 23:753–754
to respondents 23:754–756
to sponsors 23:753
to U.S. federal agency sponsors 23:753
policy 18:460–461, 18:463–464
political attitudes 18:459, 18:462–463
political behavior 18:459–464
political communication 18:463–464
political psychology 18:463–464
protection of respondents' confidentiality 23:755–756
public opinion 23:752–754
public opinion and democracy 18:463–464
sampling 18:459–463
scientific 18:459–460
standards for conduct of research 23:752–753
straw polls 18:459
subsequent expansion 18:461
survey research 18:459–464
U.S. federal agency protections for human research participants 23:754–755
voting 18:459–464
WAPOR 23:752–754
Pollock, Griselda 25:211
Pollock, Sheldon, works, *The Language of the Gods in the World of Men: Sanskrit, Culture and Power in Premodern India* 11:86
Polls, and public opinion 19:563–564, 19:566–567
Polluter pays 7:813
Pollution 7:778–779, 7:781
LTG study 14:138–139, 14:142
regulation 20:170
in Soviet Union 20:831–834
and sustainable transportation 23:846–847
and symbolism 23:860, 23:862
of water *see* Water pollution
Pollution control 7:772–773
Pollution Haven hypothesis 13:148
Polnische Wirtschaft 6:813
Polski Fiat Communism 20:635
Pólya, George, on heuristics 10:830
Polyamorists 7:535
Polyandry, as response to ecological challenges 14:618
Polyarchies 1:104, 5:657–658, 11:556
Polybius, Hellenistic historian 11:36–37
pragmatic history 11:36–37
Polycentric governance, and policy networks 18:316–317
Polycentric institutions 20:564
Polycentric mega-city 24:895
Polycentricity of cities 24:823–824
Polychlorinated biphenyls (PCBs) 7:757, 7:818
Polychoric correlations 23:617
Polychronic time system 12:357
Polycystic ovary syndrome (PCOS) 10:464
Polydrug use 6:687
Polyethnic diversity 16:23
Polygamy, in South Asia 2:79
Polygenic trait, genetic studies of behavior 9:934
Polygon overlay function 10:59
Polygraph 9:355, 13:492
Polygynous society 1:318–319, 1:319f
Polygyny 25:641
as default mating/marriage system 14:614–615
polygyny threshold model, female choice, and male coercion 14:615–616
resource defense 14:615
Polygyny threshold model 11:249
Polymerase chain reaction (PCR) 2:677, 9:916, 9:956, 9:975–976, 11:307–308, 18:898, 25:786
Polymorphic associations 25:256–257
Polymorphic lexicons 13:943
Polymorphisms 9:826, 12:300, 13:944, 19:463
Polymorphous engagements 9:151
Polynesia 17:446–448
as cultural areas 17:158–159
cultural evolution 5:397–399
in global age 17:162–163
island types 17:159
see also Oceania
Polynesian chief 2:569
Polynesian Outliers 17:158–159
Polynesian tattooing 21:374–375
Polyneuropathy, and alcohol 1:511
Polynomial, local polynomial regression 16:870–872
Polynomial analysis 20:155–156
Polynomial function 20:155
Polyopia 10:499
Polysaccharide coating of lobar pneumonia 6:693–694
Polysemy 16:410–411, 19:288
Polysomnographic recordings (PSG), for sleep 22:80
Polysyllabic signs 21:944
Polytechnical education, in Soviet Union 20:826, 20:828
Polytechnicalism, in Soviet education 20:826
Polytomous typologies 21:4
Polyunsaturated, long-chain fatty acid (PUFA) 2:848–849
POMC *see* Proopiomelanocortin (POMC)
Pomoan language 8:353
Pompeii 3:761
Pontine nuclei, eyeblink conditioning 8:637–639
Pool Cool program 23:698
Poole, G.S. 7:145–146
Pooled cross-sectional time series analysis 18:515–521
Poor 25:475–476, 25:478–479
Poor countries, decline of mortality in 15:861–862
Poor Law Amendment Act 1834 (UK) 16:464, 23:426
Poor Law Commissions 6:453
Poor laws 25:475
utilitarian social thought 24:982
Poor mothering by mentally ill mothers 17:724
Poor parenting, characteristics of 3:413
Poor people, preferred action, and liberation theology 14:34
Poor utilization of inventory assets, in supply chain management 23:711
Poorly-functioning markets, prediction and valuation of policies in 5:51–52
prediction of impacts 5:51
using continent valuation methods 5:51–52
using revealed preference methods 5:51
valuation of costs 5:52
POP *see* Principles for the Older Person (POP)
Pope Benedict XVI 23:46
Popkin's rational choice model 25:96–97
Poppelreuter, Walther, neuropsychological rehabilitation 2:814
Popper, Karl Raimund 3:774, 5:271, 7:569, 7:926, 11:12, 14:691, 18:522–527, 18:626, 19:52, 19:63
causality 18:524
economic methodology 5:271, 7:101–102
historicism 18:523, 18:525–526
liberalism 18:523
life 18:522–523
logical positivism 18:523
philosophy 18:526–527
political liberalism 18:523
positivism 18:523, 18:525–526
scientific reasoning 18:526
works 18:523–526
The Open Society and Its Enemies 2:715–716, 18:483
Popper, K.R., works, *The Self and its Brain* 18:525
Popperian contradictions 19:54
Popperian version of methodological individualism 15:303
Popper–Kuhn debate 21:308–309
Popular culture 14:425–427, 18:528–532, 20:650, 23:640–641
analysis of 2:852–853
and avant-gardes 2:323–324
in China 2:69
cinema 18:530
description 18:528–529
digital media 18:528, 18:531
entertainment 7:670–671
media 18:528–531
and nationalism 18:530–531
and other sociocultural processes 18:530–531
publishing 19:621–622
and rumors 20:792–793
scholarly perspectives 18:529–530
and sport 23:316
studies 2:357–358, 2:361
and hegemony 10:745
and television 18:528–531
and transnationality 18:529, 18:531

Popular Front strategy (communism) 18:406
Popular history 23:633–634
Popular music 3:799–804
  categories
    consequences of 3:801–803
    dynamics of 3:803
  classification systems 3:799–800
  legitimacy in comparative perspective 3:800
  social and symbolic boundaries 3:800–801
Popular politics 6:80–81
Popular religion, in Latin America 3:537
Popular religiosity, in Mexico 1:629
Popular science
  origins of 21:219–220
  science communication as 21:220–221
Popular shamanism 21:868–869
Popular social work 14:667–668
Popular sovereignty, populism 18:610–611
  republicanism 20:479
Popular writing *versus* scholarly writing 1:741–742
Popularity, and media ethics 15:46
Population, Organization, Environment, and Technology *see* POET (Population, Organization, Environment, and Technology) model
Population aging 1:338–339, 7:137, 7:335–336, 18:540–544
  centenarians 18:544
  cohort analysis 18:542
  demographic consequences 18:541–542
  in East Asia 9:55, 9:58
  economic consequences 18:542–543
    aging process and working-age population 18:542–543
    aging society 18:543
  health 18:541–543
  and migration in Europe 15:381, 15:383–384, 15:386
  modern demographic phenomenon 18:540–541, 18:541f, 18:541t
  old age 18:540–543
  policy 18:543
  retirement 18:540–543
  social consequences 18:543
  social security 18:542–543
  structure 18:542–543
Population Association of American (PAA) 1:841–842
Population attributable fraction (PAF) 14:520
Population attributable risk, epidemiology 7:830, 7:830b
Population biobanks 1:759
Population carrying capacity 3:155–157
  Ginzburg paradox 3:156
  Levins paradox 3:156
  paradoxes 3:156
Population census 3:296–301
  social and behavioral science research 22:616–617
Population change 11:6
Population coding 15:968
Population control, and fertility transition, in China 9:46–47
Population Council 8:786
Population cycles 18:559–561
  externally driven cycles 18:560
  fertility 18:559–561
  internally driven cycles 18:560–561
  limit cycles 18:560–561
  mechanisms driving 18:559
  new directions 18:561
  transient cycles 18:559–560
Population decline, in East Asia 9:55, 9:58
Population density
  in collectivism/individualism 4:207
  and homicide 11:164–165
  and specialization 6:244
Population dispersal hypotheses, testing 8:373–374
Population distribution 18:597, 3:161f
  by age, evaluation and adjustment of 6:129–133
Population dynamics 6:176–179, 7:792, 18:562–571, 18:579–587
  age distribution 18:580
  and age structure 1:338–340, 1:339f–340f
  age-structured cycles 18:581–582
  applications 18:586
  applications not requiring schedules of rates 18:562–563
  applications that assume schedules of rates 18:563
  applications that require schedules of rates 18:563
  Bernardelli, Leslie, and Lewis model (discrete time model) 18:585
  Boserup on 18:568
  branching processes 18:579–580
  carrying capacity 18:580
  demographic impact models 18:569
  Easter Island 18:567
  economic development and environment 18:568
  elasticity 18:581–582
  exponential growth 18:579–580
  extinction, stable size, and explosions 18:579
  fertility 18:563–564
  Galton–Watson 18:579–580
  general population structure 18:581
  generalization 18:586
  human development cycles 18:582–583
  impact of population on environment 18:568–569
  independence 18:580
  institutions 18:567, 18:569
  I-PAT identity 18:569
  kinship 18:563–564
  laborer–capitalist cycles 18:582
  Malthusian model 18:581–583
  Malthusian population dynamics and natural resources 18:566–568
  maternity 18:581–582
  modeling environmental influences on 18:569–570
  mortality 18:562–564
  occupation–switch cycles 18:582
  organizational ecology 6:892–893
  overlapping generations model 18:568, 18:570
  population growth 18:579
  predator–prey cycles 18:582
  prey-predator dynamics 18:567
  reproduction 18:562–564, 18:579–580
  Sharpe and Lotka model 18:584–585
  slow-fast dynamics 18:569
  small-scale mathematical models 18:570
  stability with migration 18:585
  stable population theory 18:580
  survivorship 18:564
  technology 18:566–569
  two-sex demographic models 18:588
Population ecology, organizations 17:308
Population ecology theories 2:482, 2:484
Population estimation
  settlement and landscape 21:657
  small area
    methods 6:164–165
    projection methods 6:165–167
    types of 6:163–164
Population forecasts 18:593–596
  cohort-component method 18:593–594
    advantages 18:594
    curve fitting approach 18:593
    incorporation of age structure 18:593–594
  population renewal 18:593–594
  risk analysis 18:596
  statistical uncertainty 18:595–596
    in events and rates 18:595
    expressing uncertainty via predictive distributions 18:596
    logical flaws in high and low variants 18:595–596
    probabilistic forecasts 18:596
  stochastic forecasts 18:596
  vital rates 18:594–595
    difficulty with explanatory models 18:594
    errors and debate on terminology 18:595
    self-fulfilling or self-negating prophecies 18:595
Population genetic evolutionary models, structure of 8:470–472
Population genetics 8:449, 11:289–290, 11:306
  in anthropology 9:940–941
  inferences 11:290–292
Population genetics theory 13:24–25
Population genomics 4:352
Population geography 18:597–602
  spatial demography 18:597–599
    components of change 18:597–598
    human migration, economics of 18:597–598
    modernization 18:599
    population distribution 18:597
    population systems 18:598–599
    population transitions 18:599
  studies 18:599–601
  transdisciplinarity 18:601–602

Population growth 18:579
and age structure 1:342
and class struggle 6:242
and demography 6:122
and dowries 6:642, 6:644
ecological economics 6:860
and family planning 8:786
and migration in Europe 15:382, 15:385f–386f
and urbanization in Africa 24:969–971
Population health 2:656–657, 2:659–661, 15:98, 15:100, 15:103
morbidity compression 4:464–465
Population heterogeneity
and crime 1:312
temporal dynamics in 17:732–733
between-group and within-group variance 17:733
origins and trends in social inequalities in health 17:732–733
Population history 22:309
Population learning, organizational ecology 6:893–894
Population methods 9:888
Population momentum
and age structure 1:342
analytic approach 18:573–574
empirical approach 18:572–573
and migration 18:576–577
nonstable momentum 18:574
North America 18:553, 18:555–556
over time 18:575–576
stable momentum 18:574–575
total momentum 18:575
Population movements
frontiers and 9:464
in North America 17:17–19
Population paradigms, family planning programs 8:794
Population policy 8:218
Population profiles 2:602
population projections 9:318
Population pyramid, and age structure 1:338–340, 1:339f
Population reconstruction, inverse projection 6:150, 6:152
Population registers 3:291–293, 5:736–737, 6:139–140, 6:163
censuses 6:140
continuously updated 6:140
updated at fixed intervals 6:139–140
Population renewal 18:593–594
Population Research 1:842
Population size
and higher education (HE) market 10:857
and homicide 11:164–165
impact of HIV/AIDS on 15:842
size estimation 18:603–608, 20:875
Population strategy 19:546
Population stratification 9:831–832
genetic studies of behavior 9:937
Population structure
and homicide 11:164–165
microevolution 15:338–340
and social change 22:193
Population studies 1:839, 18:599–601
development 18:600
difference 18:600–601
environment 18:600
experimental methods in 8:559–565
double-blind RCTs 8:561
equilibrium and longer-term effects 8:563–564
estimation of variability 8:562
external validity 8:562–565
generalizability of evidence 8:562–563
historical perspective 8:559–560
ideal RCTs 8:560–561
integrating theory and estimation with experimental designs 8:564
intent to treat and treatment on the treated estimates 8:561
levels of randomization 8:562
mechanisms 8:559–560, 8:562, 8:564
Move to Opportunity (MTO) experiment 8:561, 8:564
observational studies 8:559, 8:563–564
lifecourse 18:599–600
of literacy 14:239–240
sociospatial relations 18:601
Population thinking 8:368
evolution of language 8:368
Population transfers
and ethnic cleansing 8:85–86
and ethnic conflict 8:92, 8:94
Population vector 15:968
Population-average, infant and child survival 16:52, 16:54–55
Population-based programs, for skin cancer 23:698
Population-based reproductive carrier screening 9:927
Population-level asymmetry 2:806
Population-level bioethics 2:628–629
Populations 3:785
of Africa 1:292–293
age stratification as feature of 1:333–334, 1:333f
census *see* Censuses; population
census data 3:285–288
change in former communist states of Eurasia 18:545–552
of communities 22:266
composition by race and ethnicity in North America 18:553–558
demographic study of *see* Demography; history of
density and size
and prosocial behavior 19:239
and economic growth 7:42
effects on immigration 15:384–386, 15:384t, 15:385f–386f
and evolution 8:389
evolutionary demographics 8:413–419
genetic programming 2:41–42, 2:44
and global political economy 10:633
growth 3:158
human ecology 11:285
and intergenerational justice 12:396, 12:398–399
Jewish
composition 12:807–810, 12:808t–809t
growth (1700–1939) 12:802–803, 12:802f
trends (1939–2013) 12:803–805, 12:804t–805t
joint commitment 12:842
level policy responses for alcohol misuse 1:520
mobility 15:397
models 16:114
multistate stable population 16:114
multistate stationary populations 16:114
monotonicity 4:876
of Native Americans 1:634
panel surveys 17:464–465
proactive interference in different 19:5
projections 1:841
and prosopography 19:256–258
rural 20:814–818
social psychology 22:525–527
sociology and economy of 14:458–460
in South Asia 2:76
structure
microevolution 15:338–340
urban, past and future 24:958–961
Wright's contribution 25:759
Population-wide diffusion 16:788–789
Populicide 10:6
Populism 18:411–414, 18:609–612
agrarian 1:459
cordon sanitaire 18:611–612
and democracy 18:611–612
Europe 18:610–612
extremism 18:611–612
ideology, elements of 18:609–610
Latin America 18:609–610
liberalism 18:611
mobilization, leadership, style, or ideology 18:609
in party families 18:610–611
political communication 18:609
popular sovereignty 18:610–611
radical right 18:610
Populist model 8:837
Populist public history 23:308
Porat, Marc 12:85
Pornification 17:220
Pornography 18:533, 18:536–537, 18:613–618
consumption patterns 18:613–614
age of first exposure 18:614
frequency of exposure 18:614
gender differences 18:614
prevalence 18:613–614
rates of consumption 18:614
situational and interpersonal characteristics 18:615
content of 18:614
frequency of 'aggression' in adult films 18:614
defining of 18:613
effects of 18:615–618
on attitudes 18:615

attitudes toward pornography (ATP) 18:616
on behavior 18:617
beneficial 18:617–618
on cognition 18:616
sexist attitudes 18:615–616
sexually permissive attitudes 18:616
support on sexual aggression on women 18:615
impact on individual 18:613
online 17:217–222
opponents of 18:613
outlets for consumption 18:614–615
Internet and smartphone revolution 18:614
reasons for consumption 18:615
and victimless crime 25:75, 25:77–79
Port Royal Circle 14:169–170
Portable peoplemeters (PPMs) 2:213
Portable X-ray fluorescence spectrometry (PXRF) machines 24:481–482
Porter, Bernard 11:685
Porter, Edwin S., works, *The Great Train Robbery* 9:166
Porter, Michael E., works, *The Competitive Advantage* 24:866
competitive strategy analysis of 23:522
Portfolio choice, and income tax 11:745–746
Portfolio composition, household wealth 25:454
Portfolio losses 15:731–732
Portfolio management
in popular music market 3:802
stock market 23:482, 23:484–486
Portfolio models, for strategic planning 23:520
Portfolio workers 18:665
Portfolios 15:82–83, 18:291
leisure activity 13:830
Portrait Values Questionnaire (PVQ) 25:44, 25:52–53
Portraits, case studies 3:197
Portugal, fetishism in 9:105–106
Portuguese Civil Code of 1867 23:60
POS *see* Part-of-speech (POS) tags
POSCO *see* Pohang Iron and Steel Company (POSCO) Visiting Fellowship Program
Position 10:174–177, 23:435–436
structural equivalence 23:623
Positional constraints 23:265, 23:268–270, 23:272
Positional level of analysis 13:899
Positional level processing
grammatical encoding 23:256–257
Positional school 23:508
Positioned subject, and cultural relativism 5:485
Positioning
discourses 22:225
and identity 6:500
Positive action 1:248
Positive activities, and subjective well-being 23:651–652
Positive analysis, economics 13:454
Positive and Negative Affect Schedule (PANAS) 7:520, 23:648
Positive and Negative Syndrome Scale (PANSS) 19:290–291
Positive aspects, of cancer 3:74–75
Positive associations 25:398–399
Positive assortative matching (PAM)
conditional 2:106
measuring 2:102–103
rationalizing 2:106
Positive behavioral interventions and supports (PBIS) 23:216
Positive beliefs 4:2–3
mathematical competence 14:818
Positive buying *see* Buycotting
Positive check 11:5–7
Positive climate 3:833–834
Positive conception of freedom 9:387, 9:389
Positive coping 4:902–903
Positive core 4:871
Positive development 3:447
and human development 11:280
motive congruence in 15:954–955
Positive dialectic 10:490–491
Positive duties 20:680
Positive economics 7:114–115
Positive emotions 2:864, 6:746–747, 7:498–500
applications 2:867–869
broaden-and-build theory of 2:864–869
build effect and 2:866–867
health and 2:868–869
in intergroup relations 7:504
longer-term effects of 2:867
and religion 20:264–265
resilience and 2:868
Positive empathy 7:551–552
Positive eugenics 8:233
Positive feedback 21:487
Positive freedoms 25:508
Positive functioning, and mental health 15:188
Positive health 10:522
of globalization 10:199
Positive hierarchical sorting 12:513
Positive human rights 11:353
Positive illusions 10:700
Positive industrial ecology 11:848
Positive law 4:339–340, 11:369, 13:753–754, 16:323–324, 18:24
Positive liberty (PL) 9:398–399, 11:524
identity 11:524–525
Positive mental health 15:187–188
Positive morality 13:483–488
Positive nonviolence 17:621
Positive outcomes, armed conflict 1:971
Positive parenting 2:473
Positive peace 4:591, 17:621–622, 17:629
Positive politeness 18:326–328
Positive politics 4:522–523
Positive psychology 7:276, 15:187
academic self-concept in 1:57
approaches for motivation in Australian Aboriginals 15:899–901, 15:903, 15:905
and authentic leadership 2:245–246
personality 17:911–912
resilience 20:557
and vocational interests 25:239, 25:241
Positive psychology movement 4:666
Positive qualities 4:666
Positive reinforcement 1:374, 20:213–214, 25:729
behavior therapy 2:429
Positive reinforcers, behavior therapy 2:429
Positive rights 16:322–323
Positive sanctions 13:807–808
Positive science 7:665
Positive social integration 12:243–244
Positive social interdependence 4:857–859
Positive sociology 22:554–555
Positive symptoms in schizophrenia 19:292
Positive valence 2:562
Positive well-being, coronary heart disease 4:917–920
Positive youth development (PYD) 3:445
and adolescent self-regulation 21:549–551
conception of developmental processes 3:445
'Five Cs' of 3:447–449
4-H Study of 3:447, 3:449
Lerner and Lerner 3:447–449, 3:448f
model 3:445–446
out-of-school-time (OST) activity and 3:449
Positivism 2:715, 10:85, 11:134, 12:702, 13:112, 13:701, 13:916–917, 15:576, 18:254, 18:619–629, 19:410–411, 22:223, 22:226, 22:236, 22:301–302, 22:696, 22:986, 22:997–1000, 23:690
anthropology 1:748
Comte on 18:625–626
and concept of introspection 12:702
critical theory 18:627–628
criticisms 10:787–788
empiricism 18:625–627
explanation 18:625–628
hermeneutics 18:626–627
and history of science 4:520–521
interpretivism 18:625–626
law and morality 13:479–480
legal reasoning 13:776, 13:778–779
and life course 14:76
logical 14:313–318
versus methodological dualism 10:786
Popper on 18:523, 18:525–526
prospects of 18:627–628
realism 18:627
research 9:717–718
rise of 18:625
scientism 18:626
social theory 18:627
and sociology 18:626–627
transformation, Vienna Circle and 18:625–626
Positivism, history of 18:619
Comtean positivism 18:620–621
future 18:623
Machian positivism 18:621–622
Platonic political roots 18:619–620

Positivism, history of (*continued*)
Vienna Circle 18:622–623
Positivism, legal 18:25–27
separation thesis 18:25–26
validity in 18:26–27
command theory of law 18:26
Hart's concept of law 18:26
inclusive *versus* exclusive legal positivism 18:26–27
pure theory of law 18:26
Positivism, neo *see* Neo-positivism
Positivismusstreit 18:623
Positivist anthropology 1:748
Positivist criminology 5:239–240
Positivist dispute, Albert on 1:496
Positivist School 5:239
Positivist scientism 2:751
Positivist Society 4:520, 22:554
Positivist *versus* constitutive approaches, towards social change 13:544
Positivistic epistemology 22:431
Positron emission tomography (PET) 4:60, 9:490–491, 11:603, 11:612, 11:616, 13:490, 14:439–444, 19:274, 9:490f, 11:614t–615t
advantages 9:491
for cortical area studies 5:23, 5:25, 5:27
disadvantages 9:491
for eyeblink conditioning 2:96
hypnosis 11:466–467
investigating speech and language with 9:494–502
default language network 9:498
fiber tracts connecting brain regions associated with language 9:495–497
language rehabilitation and recovery 9:500–502
left hemisphere areas activated by spoken and written language tasks 9:496f
lexico-semantic processing 9:494
neural substrates underlying bilingualism 9:499–500
neural systems underlying linguistic pragmatic behavior 9:499
neurochemical bases of speech production 9:500
syntactic processing 9:494–495
pharmaceutical parameters 9:490
second modality 9:491
third modality 9:491
working of 9:491f
Posner, M.I., on adolescent self-regulation 21:547
Posner, Richard 12:935–936
Pospísil, Leopold 13:803
Possession 19:191–192
conceptualizing 23:282
discourses 23:282–283
Possession, linguistics 18:630
linguistic side of 18:632–634
literature on 18:630–631
typological research on 18:631–632
Possession of illegal substances 15:593
Possession paradox 3:675
Possessive adjectives 23:891
Possessive classifiers 18:631
Possessive pronouns 23:891
Possessive verbs 18:631
Possibilism 7:723, 20:122
Hirschman on 10:894–895
Possibility of progress (internal realism) 20:22–23
Possible realization 2:340–341
Possible selves 21:470
Possible selves theory of identity 6:500
Possible worlds 14:308, 14:310, 20:67, 20:69
physicalism 18:109–111
Possible worlds view, of counterfactuals 5:89–91
Post-academic science 21:316–317
Postadoption contact arrangements 1:144
Postal communication 12:490–491
Postal reform 18:641–642
Postal service 12:490–491
airplane transport 12:491
Postal systems 18:640–643
corporate 18:640–641
imperial 18:640
national 18:641–642
origins 18:641
postal reform 18:641–642
recent developments 18:642
and social science 18:642
Postanarchism 19:747
Postbehavioral revolution 18:469–470
Postbooking diversion 4:373
Postboomer generation 20:300
Post-borderology, Mexico 1:630–631
Postbureaucratic organizations 25:659
Postcapitalism 24:821–822
and postindustrial societies 18:665, 18:667
Postcentral gyrus 3:336
Post-Cold War era, peace movements in 17:645–646
Post-Cold War international system, Third World and 24:302–303
Post-Cold War neoliberalism 24:846–847
Post–Cold War period
East Asian politics in 6:800–803
Chinese politics 6:800–801
intraregional and extraregional comparative studies 6:803
Japanese politics 6:802–803
Korean politics 6:801–802
Postcolonial 23:633–634
modernities in 15:705
Postcolonial anthropology 4:221
Postcolonial computing 21:188
Postcolonial confederation attempts 4:572
in African states 4:575
in Latin America 4:574
Postcolonial geography 18:644
Anthropocene 18:645–646
cartography 18:644–645
comparativism 18:645
cosmopolitanism 18:645–646
cultural imperialism 18:644–646
globalization 18:645–646
hybridity 18:644
imperialism 18:644–646
legacies and directions 18:645–646
original scope of 18:644–645
race 18:644–645
revisions 18:645
self-reflexivity 18:645
subaltern 18:644
translation 18:645
whiteness 18:645
Postcolonial historiography
Indo-Islamic period 11:79
Postcolonial law 18:648–649
globalization and 18:649
legal orientalism and 18:649–650
in the period of nationalism 13:815
in twety-first century 18:647
perspective, defining 18:647–648
Postcolonial legal orders 23:61–62
Postcolonial science studies 18:652
anthropology 18:653
development 18:653
empire 18:652, 18:654
genealogies and antecedents 18:652–654
global claims 18:652
globalization 18:652, 18:654–655
history 18:652–653
hybridity 18:652–655
indigenous 18:653
medicine 18:652–655
national 18:652
postcolonial actors and scientific travels 18:654
technology 18:652–654
technoscience 18:654–655
Postcolonial societies, multiculturalism in 16:30
Postcolonial South Asia, nationalism in 16:281
Postcolonial studies 1:935, 1:937
Postcolonial technoscience 18:654
Postcolonial theory 10:852, 18:652–653, 21:154
and development geography 6:315
Postcolonialism 6:293, 11:42, 11:45, 16:456, 18:391, 21:153–154, 23:83
and anthropology 1:770
anthropology in Korea 2:72–73
in Central Asia 3:316–317
and creolization 5:120
and dance 5:667–668
de-Westernization 25:546, 25:548–549
domains of discourse 16:456
and global south 10:186, 10:189–190
and historical realities 16:456–457
and indigenous knowledge 11:779–782
and legal development, in Africa 1:273
and modernity 15:694–695
and sociopolitical challenges 16:456–457
studies 14:200
theoretical constraint of 16:457
Postcoloniality 18:658–662
decolonization 18:658–661
end of 18:660–661
globalization 18:658–661
provincializing Europe 18:659–660

statehood 23:348–349
subaltern studies 18:658–659
Post-Columbian admixture in the Americas 9:893
Postcommunism
cultural formations in Eastern Europe after 6:808
and youth movements 25:838–839
Postcommunist society 18:667
Postcommunist transition, in Eastern Europe 6:819–820
Postconflict situations, of child soldiers 3:431–434, 3:436
Postconscious 22:207
Postconsensus urban planning 24:865–866
Post-Constantinian Catholicism 3:233–234
Postcorporate career 3:130
Post-creole continuum 18:137
Postdemocratization era 23:657
Postdeparture training, cross-cultural training 5:338
Postdevelopment 6:29, 18:182, 18:184
and global south 10:190
Postdevelopment critique 18:182, 18:184
Postdisaster reconstruction 10:550
Post-divorce adjustment of children
child factors 14:523
interparental conflict 14:522
interventions for problems 14:521–522
parenting 14:522–523
stressful events 14:521–522
variability in outcomes 14:521–523
Postdoctoral experience
academic research and employment 1:50
Postdoctoral phase
academic research and employment 1:50
Postdoctoral scholars
academic research and employment 1:50
Postearnings announcement drift 23:483
Postempiricism 11:134–136
Postenumeration sample (P sample) 3:293
Postenumeration survey (PES) 3:293, 6:127–128
census 3:306
data adjustments 6:127–128
Posterior cingulate cortex 10:889, 18:817
Posterior cortical atrophy 19:253–254
Posterior density 12:799
Posterior distribution 2:375
Bayes' theorem 2:379–381, 8:17
calculation of 2:375–376
Posterior expected loss 8:17
Posterior parietal cortex (PPC) 5:21–22, 5:28, 17:506, 21:420–421, 25:201
anatomy 17:508
deficits associated with posterior parietal damage 17:509–510
acalculia 17:509–510
apraxia 17:509
hemispatial neglect 17:509
oculomotor apraxia 17:509
optic ataxia 17:509
working memory deficits 17:510
functions 17:508–509
Posterior pituitary 11:486
Posterior predictive model checks 15:655
Posterior probabilities
of the risk of arrest 6:253
significance tests 21:959–960
Posterior superior temporal sulcus (lPSTS) 16:673
Posteriori density 23:414
Postevent information, eyewitness testimony 8:644, 8:647–648
Postfeedback-perceived risk 10:706–707
Postfeminist movement 8:902
Post-Fordism 7:687
and postindustrial societies 18:663, 18:666–667
Post-Fordism regime 10:220
Post-Freudian concepts 19:349
Postfunctionalist conflict theory 5:578–579
Postgay world 13:881
Postgenome approaches 9:1002
Postgenomics 2:625, 9:979
Postgraduate legal studies 13:736
Post-heroic era, commemoration in 4:237–238
Posthumanism, cybernetics 5:645–647
Postictal period, seizures 7:854
Post-industrial 24:460–461, 24:463–464
Postindustrial age 18:663–664
Postindustrial city 3:664
Postindustrial economies, social rationality in 25:56
Postindustrial period, and economic growth 5:528–529
Postindustrial societies 10:48, 13:106–107, 18:663–669
Bell on 2:502–504
capitalism 18:663–668
deindustrialization 18:666
Fordism 18:664–666
globalization 18:668
history of term 18:663–664
information 18:666–668
knowledge 18:663–665, 18:667–668
leisure 18:664–665
and modernization 15:707–708
orienting theoretical developments 18:664–666
postcapitalist 18:665, 18:667
post-Fordism 18:663, 18:666–667
postmodernism 18:663–665, 18:668
tensions and competing visions 18:666–668
work in 18:663–668
Postindustrialism, and information society 12:86–87
Postindustrialist societies, division of labor changes 13:159
Post-Keynesian economics 1:577
Post-Marxism 9:401
Postmaterialism
death/dying 5:867–868
and environmental movements 7:764
and generational replacement 9:862–865
and peace movements 17:644–645
Postmaterialism index (PDI) 9:863–864
Postmaterialist values 25:304–306
Postmedieval archaeologists 11:2
Postmillennialism 15:529–530
Postmodern age 18:663–664
Postmodern art
entertainment 7:672
Postmodern condition 10:48
Postmodern critical theories 22:962
Postmodern era, gender anthropology and 9:781
Postmodern ethnography 12:644, 23:641
Postmodern family law 8:784
Postmodern feminism, social work 8:935–937
Postmodern history 10:165
Postmodern homes 11:157
Postmodern identity, and self 21:447
Postmodern legal feminism 9:733
Postmodern narratology 16:195–196
Postmodern pragmatism 18:810
Postmodern public administration 19:523–524
Postmodern reflexiveness, in advertising 1:217
Postmodern reflexivity 19:680–681
Postmodern societies, and criminology 5:245–246
Postmodern sociological approach, toward consumption 4:759
Postmodern theories 11:15–16, 18:688–689
of personal identity 17:745
Postmodern tourism 24:466
Postmodern urbanism 24:802
Post-modern wars 3:714
Postmodernism 7:259, 9:733–734, 10:87, 11:123, 14:202, 15:704–705, 16:456–457, 19:491–492, 19:697, 20:88, 21:309–310, 21:768, 22:689, 22:994–995
analyses 16:459–460
avant-garde art 2:321–322
Bell on 2:503
Benhabib on 2:523
and community sociology 4:365
criticism without foundations 18:688–689
and culture 5:584
deconstruction 18:680, 18:682
and development 8:906
and Enlightenment 7:666–667
and feminism 8:900–901
and indigenization 16:459
and interviewing 12:674
media audiences 2:217–218
and memory 4:203–204
metanarratives 18:680–682
methodology 18:680–683
as art and literature 18:681–682
for complex intertextural world 18:681
going beyond qualitative methods 18:680–681
as interpretation 18:681
linking substance and method 18:680
rejection 18:682
relativism and questioning postmodernists 18:681–682

Postmodernism (*continued*)
and military sociology 15:516
modernity 15:685–686
narrow definition 18:682
origins of 18:676
philosophical aspects 18:684–689
criticism without foundations 18:688–689
différance 18:687–688
difference and multiplicity 18:687–688
differend 18:687
grand narrative 18:686
modernity/modernism 18:684–685
postmodernity 18:685–686, 18:688
poststructuralism 18:686
pragmatism 18:685–686
representation 18:684, 18:686, 18:688
simulacra and simulation 18:688
truth 18:685–686
unity and plurality 18:685–686
universalism and freedom 18:686–687
and postindustrial societies 18:663–665, 18:668
progress 19:127
and psychology 19:417
and relativism 20:233, 20:236
social sciences 18:680–681
and social work 5:287, 5:292
of social work 16:459
in sociology 18:676–679
capitalism 18:676–677
communication 18:677–679
consumption 18:676–679
digital culture 18:676–677
feminism 18:676, 18:678
fields of research 18:678
identity 18:676–678
internet 18:676–679
media 18:676–679
methodology 18:677–678
new research directions 18:678–679
social interactions 18:677–678
state of research 18:678
technology 18:676–678
and sport 23:314
and symbolic interactionism 23:857–858
tenets of 18:677
analysis of postmodernity 18:677
critique of modernist sociology 18:677
Postmodernity 10:220, 15:704, 17:439–441, 18:685–686, 18:688, 19:126–127, 21:212, 21:217–218, 24:510
and arts 2:15
capitalism 17:441, 17:443
e-sociology 17:444
leisure in 13:845
and reflexivity 20:88–91
religion in 10:215–216
urbanism 18:670–675
Postnatal brain development, and epigenetics 7:850–851
Postnatal depression, screening for 17:723
Postnatal development, formation of brain during 2:824–825
Postnatal genetic testing 10:36
Postnatal stressors, and personality 17:912
Postnational citizenships, prospects of 3:652–653
Postnational Europe 8:277–278
Postneonatal mortality 12:16–18
Postnetwork, television 25:83
Postnormal science 17:683–684, 21:249, 21:253
Postpartum blues 17:723
Postpartum depression (PPD)
gonadal hormones in 7:608
stress hormones in 7:608
Postpartum period 10:466
Postpartum womb (life phase) 5:510
Post-PhD transitions 6:611–612
Post-photographic age 25:173
Postponement
of childbearing 9:19, 9:24
fertility 9:23–24
of motherhood 9:20–21
of parenthood 9:20–21
transition 9:19
and transition to adulthood 24:571–572
Postponement transition, in infertility 12:46–48
Postpositivism 11:134–136, 18:628
Alexander on 1:524
Postpositivist philosophy 1:893
Postprocessual archaeology 1:903
household archaeology 11:221
Postprocessualism 7:692
and hunter-gatherers 11:426
Postprofessionalism 17:484
Post-queering anthropology 19:757–758
Postracialism 18:842, 19:799
inequality 7:895
Postracism 19:804–805
Post-Rawlsian egalitarianism 7:309–311
comprehensive 7:310–311
narrow 7:310
Post-Renaissance rationalism 4:338–339
Postreproductive life span, evolutionary demography 8:416, 8:418
Postretrieval 20:611
Postrevisionism, English Revolution 7:653
Postrhinal/parahippocampal cortices 10:888–889
Postscarcity generation 25:837–838
Post-secular 21:392
Post-secular age 21:392
Post-socialism 20:847
political economy 18:690–696
transformations of real property 19:155–158
Postsocialism, planning during 18:187–188
central planning 18:188–189
communism 18:187
Eastern Bloc 18:187–188, 18:191
Eastern Europe 18:187
planning's legitimacy crisis 18:189
pressing challenges and new planning instruments 18:189–191
privatization and institutional transformations 18:188–189
socialism 18:187–189
Soviet Union 18:187
spatial planning 18:188–189, 18:191
state socialism 18:187, 18:189
transition period 18:188
transitional countries 18:191
urban planning 18:188–189, 18:191
Postsocialist cities 3:607–608
Central Europe 3:605–607
central planning 3:606–607
challenges of urbanization 3:609–610
deindustrialization 3:608
Eastern Europe 3:605–607
economic restructuring 3:608
gentrification 3:607–610
global capitalism 3:609–610
globalization 3:608
revitalization 3:605
segregation 3:605–608, 3:610
socialist city 3:606–607
socialist legacies 3:609–610
suburbanization 3:608–610
urban transformations 3:607–608
Postsocialist countries 21:918
Postsocialist Russia, social work in 22:808–809
Postsocialist settings, land rights and tenure security in 13:234–235
Postsocialist societies 18:697–701
in bourgeoisie/middle classes 2:787
corruption 18:699
crisis in human relations 18:699
definition 18:697
differences stemming from past 18:697–698
ethnic minority 18:698
EU membership 18:699–700
former 18:699–700
and former socialist societies/countries, difference between 18:700
interethnic relations 18:699
nationalism 18:698–699
non-European (former) socialist countries 18:700–701
poverty 18:698–699
present similarities/key problems of transition 18:698–699
problems of identity 18:698–699
problems of social interaction 18:699
restructuring of social life 18:697–698
unemployment 18:698–699
Postsocialist states, in Central Asia 3:317–318
Postsocialist transition, *versus* armed forces demobilization 25:123
Postsocialist urban development 18:190
Post-Soviet Russia
administrative justice in 13:431–432
homosexuality in 21:742–743
judicial power and independence in 13:430
see also Russia; Soviet Union
Post-Soviet society 20:853–854
Post-Soviet space 20:829–830
Poststress recovery phase, cardiovascular reactivity 3:119

Poststructural interrogation 5:256–257
Poststructuralism 5:255–256, 5:927, 5:929, 6:503, 7:259, 17:343, 18:686, 23:82
agonism 1:446
Barthes on 2:360–361
and cultural sociology 5:498, 5:500–501
and cultural studies 1:935
and discourse studies 6:511, 6:513–514
language and 14:199
and networks 16:567–570
Parisian 1:81
and queer theory 19:747
and sport 23:314, 23:316
and wisdom 25:601
Poststructuralist feminist theory 25:631
Postsubcultural theory 23:638
Postsubcultures, youth culture 25:817
Postsubiculum 16:625
Postsynaptic input 16:704
Postsynaptic potential 2:29
Post–Third-Wave feminism 8:909–910
Posttranslational modifications (PTMs)
of core histones 7:839–840, 7:840t
to histone proteins 7:849–850
Posttraumatic amnesia (PTA) 18:636–639
assessment 18:637
cerebral injury 18:636
cognitive impairment and behavior during 18:638
confusion 18:636–639
definition 18:636–637
head injury 18:636–638
history 18:636
memory 18:636–638
orientation 18:637–639
rehabilitation 18:638–639
reorientation 18:638–639
severity of TBI, recovery, and outcome 18:638
Posttraumatic growth (PTG) 18:708
social work 18:708
and stress 23:572–574
Posttraumatic stress
in social work 18:707–713
compassion fatigue 18:710
continuing with routine and resiliency 18:708
disasters 18:707–710
diversity of client populations 18:709
emergency routine 18:709
models of training for interventions 18:711
pathological responses 18:707–708
posttraumatic growth 18:708
ramifications of trauma 18:707
salutogenic responses 18:708
secondary traumatization 18:710
shared traumatic reality 18:710–711
terror 18:707–710
trauma 18:708–709
types of interventions 18:709–710
vicarious resilience 18:711
vicarious traumatization 18:710
war 18:707–709
Posttraumatic stress disorder (PTSD) 1:970–971, 2:414, 3:765–766, 3:768, 4:390, 8:850–851, 8:872–873, 10:720, 11:507, 15:194, 15:495–500, 15:552–553, 16:200–201, 17:82–83, 17:431, 17:655, 18:702–708, 21:159, 21:161, 23:563
associated brain regions 24:438f
behavioral therapy 18:702
biological factors in 18:704
functional neuroimaging studies 18:704
genetics 18:704
imaging studies 18:704
neurobiological research 18:704
neuroendocrine studies 18:704
in child soldiers 3:433–435
and childhood cancer 3:462
combined medicine and psychotherapy 18:705
depression 18:703–705
DSM-5 18:702–703
etiology 18:703–704
exposure therapies for 8:606, 8:608–609
forensic psychology 9:352
and grief 2:538–541
mass trauma 14:747–750
military 18:703
new and emerging research 18:705–706
nightmares 18:702, 18:704–705
and occupational health 17:134
patients, fear-potentiated startle in 8:864
pharmacological therapy for 18:705
prevalence, comorbidity, and prognosis 18:703
psychological treatment 18:705
and psychosocial work 19:479–482
psychotherapy for 19:374–375
related to torture 24:435–437, 24:439–440
sleep abnormalities in 22:83
and stress 23:572
symptoms 3:433–434
torture victims 18:703
veterans 18:702–703, 18:705
Posttrial surveys or interviews 12:908
Postulate of adequacy, Schütz 21:148
Postulate of foreigness 19:670
Postulatore 15:579
Postural control, in infants 15:972, 15:974–976
Postural Origins Theory 2:800
Posture, stabilization by vestibulospinal pathways 25:67–68
Postwar area studies, in United States 1:951
Postwar internationalism, and film history 9:168–169
Postwar master planning 24:865
Postwar period, rise of independent nation-states in 16:309
Postwar reforms, in Soviet education 20:828–829
Postwar Weberians 25:465–466
Post-Woodstock society 5:217
Post-World War II, development geography 6:314–315
Potential, training as enhancement of 5:113
Potential antecedents 25:723–724
person-based perspective 25:724
situation-based perspective 25:724
Potential life span 17:172–173
Potential observations 3:252
Potential outcomes 3:252–253, 15:64
causal inference as comparison of 3:255
notation for 3:255–256
Potential outcomes model
quasiexperimental designs (QEDs) 19:742–743
Potential Pareto improvement criterion 5:47–48
Potential *versus* behavioral performance 22:372
Potestas absoluta 1:8–9
Potestas legibus soluta 1:7–9
Potestas perpetua 1:7–8
Potlatch 23:1
as anthropological totem 18:717
Dominion period 18:715
and ecological variation 18:716
fur trade 18:714–715
invention of tradition 18:717
Kwakiutl 18:715
and marriage exchange 18:716
in northern region 18:715
precontact trade and ceremonial exchange 18:714
reifications 18:717
social divisions 18:716–717
in southern region 18:715
as total social phenomenon 18:715–716
use of term 18:714
Potlatch Law 18:715, 18:717
Potter, Beatrice 7:114
Potter, Daniel, on cultural capital 5:380
works, *History and American Society*. 1:615
Potter's wheel
ceramics in archaeology 3:323–324
Pottery
in ancient Egypt 1:694
ceramics in archaeology 3:323–325
Pottinger v. City of Miami 25:114
Pouchelle, Marie-Christine 2:729–730
Pound, Roscoe 13:511, 13:772
jurisprudence 13:773
law school 13:773
Pourashava Ordinance 24:898
Poutine mafia 14:429
Pouvoir absolu 1:6
Pouvoir constituant 4:700
Pouvoir constitué 4:700
Poverty 3:18, 3:757, 4:525–526, 6:290–292, 6:294, 10:265–266, 10:634, 12:142–143, 16:914–917, 18:722–727, 18:753–759, 19:555–556, 21:45–46, 24:775–776, 24:779, 24:818–821, 24:908, 24:912–913, 25:479, 25:513
absolute, relative, and anchored poverty 18:756–757
absolute poverty 18:723

Poverty (*continued*)
and academic achievement 22:924–926, 22:928–929
in Africa 1:292–294, 1:296–298
analysis 18:741
antipoverty effects of policy interventions 18:745
demographics 18:743–744
discrimination 18:743–744
family effects of income support policy 18:745
fertility effects of income support policy 18:745
human capital 18:743–744
immigration 18:743–744
income dynamics 18:744–745
labor supply effects of income support policy 18:745
macroeconomic performance 18:743–744
social mobility 18:744–745
status attainment 18:744–745
black families 3:420–421
cannibalism 3:83
capability measure 18:743
and cash transfers 3:214, 3:216–218, 3:220–221
and child labor 3:406–407
and child neglect 3:414–415
beyond simple causation 3:414
social neglect 3:414–415
child protection 3:418, 3:420
and childhood behavior problems 2:477–480
and child/infant mortality 12:24–25
in community crime research 4:325
concentrated 24:730–731
concepts 18:754–755, 18:754f
conceptual reason 18:743
culture of 5:590–595, 18:719–721, 18:753–755, 24:729–730
alternatives and follow-up 5:593–594
anthropology 5:590–594
from Boas to Lewis 5:591–592
class 18:719–721
critique of 5:593
definition 18:719–720
deprivation 5:591–594
education 5:592–594
ethnography 5:591
ideological backdrop of 18:719
inequality 18:719–721
Lewis's intellectual milieu 5:592
and policy 5:592–593
policy implications of 18:720
psychology 5:590–593
sociology 5:590, 5:592, 5:594
theoretical implications of 18:720–721
underclass 18:720–721
defining and measuring 18:722–723
definition 18:728
and desegregation 6:221, 6:223–224, 6:224f
development 18:723–724, 18:726
discursive constructions of 18:726
and Eastern European social work 22:796–797
education and training 18:744
effects of minimum wages on 15:563–564
and equity in education 7:918, 7:920–922
experience by minority ethnic elders 15:440–442
encouraging take-up of services 15:442
health 15:441–442
housing 15:442
income 15:440–441
explained in environmental terms 23:426
and family structure 18:753–754
feminization of 16:466, 25:478
geographic conceptualizations of 18:723–726
ghetto 10:123–125
global 18:723–724
and globalization 7:119
hardship 18:741–742
and health 10:634–635
health poor 18:741
in history 18:728–732
agriculture 18:730
artisans 18:729
causes 18:728–730
concept 18:728
famine 18:730
geography 18:729–730
paupers 18:729
size 18:728–729
society 18:728–730
welfare policies 18:730–731
welfare states 18:731
working classes 18:730
and homelessness 11:161
and homicide 11:163
housing poor 18:741
and human trafficking 10:206–208, 10:210
imprisonment 18:754, 18:757
inequality 18:723–726
law *see* Poverty law
and liberation theology 14:33–34
measurement 18:753–754
absolute income poverty measures 18:742
approaches 18:755–756
consumption-based poverty measures 18:742–743
relative income poverty measures 18:742
self-reliant inability 18:743
subjective measures 18:743
techniques 18:754–755, 18:754f
and migration 15:429–431
minority ethnic families 3:420–421
neighborhood effects 18:753–756
and neoliberalism 16:463
and nonmarital childbearing 9:16
penalization of 16:465–466
as personal failings 16:463
perspectives on 15:738–739
and policy 5:591
poor households, rationality of 3:407
in postsocialist societies 18:698–699
poverty rate 18:741–745
poverty ratio 18:742
reduction, and tourism 24:474–475, 24:478–479
relative poverty 18:723
research and social science methods 18:745–746
restructuring 18:723–726
and segregation 21:408–409
and SES 22:942–944
and social deprivation 22:285
social mobility 18:753–754
social problem 12:544
and social question 22:553–555
social stratification 18:753–754, 18:756–757
and social stratification 22:709
spatial inequality 18:757
and spatial mismatch 23:157–160
stigma of 23:465
street children 23:533–534
theory 18:753–754
threshold 18:722–723
and urban violence 24:949–951
War on Poverty 18:733–737
welfare reform 18:725–726
Wilson on 25:584–585, 25:587, 25:589
Poverty alleviation 3:848, 3:851
Poverty dynamics, panel surveys for 17:465–466
Poverty law 18:736–737
after the welfare state 18:735–738
Civil Gideon movement 18:738
federal welfare state 18:733–734
in global age 18:738–739
lawyers 18:737–738
legal aid 18:734–735, 18:737–739
legal services and rights revolution 18:734–735
Poverty line 18:754–757
Poverty neighborhoods, urban 24:915–918
The Poverty of Historicism 18:525–526
Poverty of speech 19:290–291
Poverty of stimulus argument 13:307
Poverty of stimulus situation 5:580
Poverty policy 18:747–752
active labor market policies (ALMP) 18:750
basic income guarantee (BIG) 18:749
conditional cash transfers (CCT) 18:751
deprivation 18:750–751
distinctions among poverty programs 18:748–749
entitlements versus conditions 18:749–750
European Social Funds 18:751
flex-security policies 18:750
growth, employment, and poverty 18:750–751
history and development of 18:747–748
in-kind benefits 18:747, 18:749
less eligibility 18:747–749
Mahatma Gandhi National Rural Employment Program 18:749
'make work pay&rsquo 18:747
means tests 18:747–749, 18:752
minimum incomes and activation 18:750
Nordic welfare 18:750

Platform against Poverty and Social Exclusion 18:751
Poor Law 18:747–748
poverty policies in the global south 18:751–752
social assistance 18:747, 18:749–751
social exclusion 18:750–751
social investment 18:750–751
tax credits 18:747, 18:749, 18:751
universalism 18:748
welfare reform 18:749–750
'welfare' programs 18:748–749
workfare 18:749
Poverty rate 18:741–745
Poverty ratio 18:742
Poverty reduction 10:199
Poverty reduction programs 22:285
Poverty traps 3:407–408, 7:41, 11:745
and child labor 3:407–408
Povinelli, Elizabeth
on affect 1:239
works
The Empire of Love 19:754–755
POW *see* Prisoners of war (POW)
Powell, Bingham 13:803
Powell, Walter W. 5:39, 17:375, 17:393
on organizations 17:391
Power 2:256, 3:57–58, 6:368–369, 8:875–876, 10:629–630, 11:548, 12:544–545, 16:83–84, 17:201, 21:134–135, 22:345
in action 18:772–773
Arendt on 1:963
of associations 25:258
and authority 2:272–273
distinction between 10:294–295
balance of 2:345
and deterrence 6:247
Bendix on 2:513
Boltanski on 2:748–753
Castells on 3:230–231
and collective memory 4:198
and common sense 4:2–4
and community sociology 4:363
and counterfactuals 5:93
in critical social work 5:289
and culture 17:391–392
anthropological studies 18:337
Dahl on 5:656–660
delegation of 6:58–60
and diglossia 6:420
and economic development 5:530, 5:532
and energy 7:622–624, 7:627
in established communities 4:315
and freedom 9:400–402
Geertz on 9:661, 9:664–665
and governmentality 10:308–312
history 18:770
and homophobia/heterosexism 10:794
household archaeology 11:220
human rights 11:364
and information society 12:85, 12:87–88
knowledge and 13:72–73, 13:94–95
labor 25:676
and language policy 13:354–355
and law 13:537–538
and legitimacy 13:825–826
literature and review 18:770
Mann on 14:490–495
and media 22:583–584
and motherhood 15:881, 15:885
and motives 15:952–955
and negotiations 16:424
one-dimensional view of 17:405–406
and organizational conflict 4:640–642
in organizations 17:405–410, 18:770–771
community studies 17:405–406
critical approaches 17:407–408
critical views 17:407–408
discursive power 17:408–409
and domination 17:407–408
managerialist approaches 17:405–407
in OMT 17:407–408
organizational discourse 17:409
third dimension 17:407
in participatory action research 17:535–538
of partners 8:830, 8:833
and pluralism 18:233–235
political 18:760–764
political agendas 1:362–364
and political ecology 18:350, 18:353–354
political sociology 18:472–477
primordial sociobehavioral outcomes 22:879–881
in prison life 18:924, 18:926
and psychiatry 22:975
in public bureaucracies 19:531–532
relational aspects, and networks 16:520
religious 20:328–332
research on 18:774
positive effects 18:774
and sanctions 20:890
second-dimensional view of 17:406
and secrecy 21:375–376
self-categorization theory 21:458
separation, executive government 8:497
and sexual orientation 21:786, 21:788–789
social constructionism 22:222–225
social psychology 22:541
and social sciences 21:213–214
in society 18:765–769
authority 18:765–766, 18:768
charisma 18:765
community politics after Foucault 18:767–768
Community Power Debate 18:766, 18:768
disciplinary power 18:767
discursive turn 18:766–767
essentially contested concept 18:765, 18:768
Foucault on 18:767
as generic concept 18:766
governmentality 18:767
ideal types 18:765–766
language games 18:768
nondecision making 18:766
nonissues 18:766
rational legal domination 18:765
real interests, and hegemony 18:766
tradition 18:765, 18:768
Weber on 18:765–766
in South America 23:45–47
and state 23:356–360
and syncretism 23:882–883
and territorial conflict 24:214–219
and territory 24:221–225
urban sociology 24:927–928, 24:930
and values 25:28–30
and violence 25:116–118, 25:120
women's studies 9:736–737, 9:740
Power analysis, in feminist psychotherapy 8:931–932
Power dependence theory 15:669
Power distance 5:428–429, 12:358
and communication 4:284
definition 25:30
and moral development 15:784–785, 15:784t
Power elites 7:393–394, 7:397–398
Mills on 15:535–536
political power 18:761–763
Power function 11:886
in hypothesis testing 11:492
Power in exchange networks 15:667–671
bargaining 15:667–669
cooperative game theory 15:669
core theory 15:669–670
dependence theory 15:667–669
exclusion 15:668
expected value theory 15:669–670
experiments 15:667–670
game-theory 15:669
negotiation 15:667–668
network connection types 15:668
profit 15:667–670
profit pools 15:668–670
Power inequalities, in social networks 8:485–486
coercive power 8:485
risk and uncertainty 8:485–486
Power law 14:802–803
Stevens on 23:453–454
Power mobilization 25:512
Power of judges 12:881
Power of social media 18:423–424
Power relations 3:625, 3:627–630, 19:683–684, 22:438
and community work 4:380–381
degrees of citizen power 3:627
treatment research and access 8:34
and welfare retrenchment 25:525–526
Power relationships, gift as distinction mark in 10:141
Power solidification, by constitutions 4:712–713
Power spectrum 9:375, 24:343
Fourier analysis 9:375, 9:377
Power system 15:534–536
Power theory, of interviewing 5:854–855
Power transfer function of filter 24:334
Power transformations 8:600
Power transition theory 25:401–402, 25:425
Power-conflict analysts 19:847

Power-conflict theory, racial relations 19:846–848
Power-dependence
  groups 10:438
  social exchange 8:484
Power-distance cultures 17:824
Power/efficiency controversy, organizational decision making 17:355
Powerful emotions 6:749
Powerful learning environments
  to promote transition to desired end state 12:203–204
Power-geometries 16:570–571
Powerless emotions 6:749
Powerless language 9:785
Powerlessness 1:544
  emotional consequences of 1:546
  future research 1:549
  heuristics in psychology and sociology 1:546
  learned helplessness 1:545
  and locus of control 1:545
  names and faces of 1:544–546
  versus perceived control 1:546–548
  versus personal control 1:545
  self-efficacy 1:545–546
PowerPoint 7:878
Power-political pluralism 6:93
Power-sharing 15:260–261
  and conflict/consensus 4:591
PPA *see* Parahippocampal place area (PPA); Personal projects analysis (PPA); Primary progressive aphasia (PPA)
PPACA *see* Patient Protection and Affordable Care Act 2010 (PPACA)
PPB *see* Planning; Program; Budgeting (PPB)
PPC *see* Posterior parietal cortex (PPC)
PPD *see* Postpartum depression (PPD)
PPGIS *see* Public participation GIS (PPGIS)
PPI *see* Patient and public involvement (PPI)
PPIK (process, personality, interests, and knowledge) theory of intelligence 12:293
PPMs *see* Portable peoplemeters (PPMs)
PPOs *see* Preferred provider organizations (PPOs)
PPP *see* Public–private partnerships (PPP)
PPPR *see* Period parity progression ratios (PPPR)
PPPs *see* Purchasing power parities (PPPs)
P-prims 4:532
PPRs *see* Parity progression ratios (PPRs)
PPS *see* Probabilities proportional to their size (PPS) sampling
PPVT *see* Peabody Picture Vocabulary Test (PPVT)
PPVT-R *see* Peabody Picture Vocabulary Test – Revised (PPVT-R)
PQLI *see* Physical Quality of Life Index (PQLI)
PR *see* Proportional representation (PR); Public relations (PR)
PRA *see* Perceived risk (PRA)
Practical category, nationhood as 16:294–295
Practical consciousness 7:416
  Giddens on 10:136
  and ignorance 11:581
Practical imperative, and Enlightenment 7:666
Practical knowledge, and scientific revolution 21:319
Practical network actors 5:558
Practical psychology 1:853
Practical rationality, and indigenous knowledge 11:780
Practical reason 13:7
Practical reasoning 18:776–781
  agency 18:776–781
  consequentialism 18:780–781
  conversation analysis 4:850, 4:852
  Hume on 18:779
  instrumental and maximizing rationality 18:778–780
  intention 18:777–780
  morality 18:780–781
  naturalism and normativity 18:777–778
  reasons and motivation 18:778
  reasons for action 18:776–779
  specification, expressive rationality, and moral reasons 18:780–781
  value 18:776–781
  Weber on 18:778
  will 18:778
Practical skills, fishing 9:227
Practical usefulness, five-factor model (FFM) 2:563–564
Practicality 3:823
Practice approaches, international social problems 12:543
Practice settings 13:629
Practice theorists 7:415
Practice turn
  in organizational learning 13:697
  in science and technology studies 21:192
Practice-oriented discourse studies 6:512–513
Practices 7:873–874, 10:864, 15:675–676, 21:241, 21:244, 21:395–396, 21:805–806
  activity forms 18:785–786
  applications, social work research 20:504
  apprenticeship 18:786
  axiomatic 2:339–340
    axioms as defining theories 2:340
    independence of axioms 2:340
    Padoa's principle for proving independence of primitive concepts 2:340
    primitive concepts 2:339
  biomedicine as 2:664–665
    clinical trials 2:665
    diagnosis and etiology 2:664–665
    risk and enhancement 2:665
  bird's-eye view 18:785
  Bourdieu's theory 18:783, 18:785
  business anthropology 3:4, 3:7–9
  careers 3:132
    integration into coherent system 3:132
  ceremonial 11:762–763
  child welfare 11:762
  cognition in 19:361–362
  colonial 11:763–764
  communes anthropology 4:271–272
  conflict mediation 4:628–630
  cultural 11:762
  culture as 5:618–619
  definition 18:782
  doxa 18:783
  entrepreneurial 7:684
  fieldwork 9:150–151
  gender archaeology 9:766
  guidelines, ESTs 19:382–383
  habitus 18:783, 18:785–786
  land tenure 13:236
  narratives, understanding 16:208–209
  organizational learning 13:697
  prosocial behavior 19:229
  psychotherapy 19:519
  reflexivity 20:84–85
    heuristic and process 20:85–86
  religious 11:764
  scaffolding approach 18:784
  schema 18:784–785
  scientific 21:225–227, 21:244–245
  social psychology 22:535–536
  social work education 22:778–779
    experience 22:777
  sport as 23:301–302
  storytelling 11:763
  sustainable 23:834–838
  theory, Pierre Bourdieu 2:777–778
  traditional 11:762
  university hazing 10:558–559
Practicing natural scientists 24:674
Practitioner, view of implementation 11:694
Practitioner model, historical background 3:881–882
Practitioner–client interactions 8:486–487
Practitioner-scholar model 24:517
  assessment 24:520
  competency areas 24:520
  consultation 24:520
  core curriculum 24:520
  diversity 24:521–522
  education 24:520
  education for professional psychology 24:517–518
  evaluation 24:520
  gender 24:522
  integrative pedagogy 24:519–520
  internship training 24:521
  intervention 24:520
  knowledge base 24:521
  management 24:520–521
  multiple and expanding roles 24:521
  practice, elements of 24:521
  practicum 24:521
  practitioner as local clinical scientist 24:518–519
  professional psychology 24:521
  of professional psychology 24:517–523
  psychological science 24:517–518
  reflective practice 24:521
  relationship 24:520

research 24:520
social responsibility 24:522
supervision 24:520–521
systematic evaluation 24:521–522
Praetor peregrinus 13:799, 13:801
Praetores 3:670
PRAF *see* Programa de Asignación Familiar (PRAF)
Pragmalinguistics 14:227
Pragmatic accounts 14:192
Pragmatic agon 1:447–448
Pragmatic anthropology 8:588
Pragmatic competence 2:574
Pragmatic disorders 18:788–794
acquired 18:791–794
neurodegenerative disorders 18:792
psychiatric disease 18:792
right hemisphere damage and left hemisphere damage 18:791–792
traumatic brain injury 18:792
Alzheimer's disease 18:792
and brain 18:793–794, 18:793f
in children and adults 18:789–794
dementia 18:791–792
developmental 18:790–791
autism spectrum disorders 18:790
DSM-5: social (pragmatic) communication disorder 18:790–791
specific language impairment 18:790
explaining 18:792–793
figurative language 18:790–794
indirect requests 18:788, 18:791–792
irony 18:790–792
joke 18:788, 18:790–792
language as social action 18:788–789
metaphor 18:790–792
neuroimaging 18:792–794
nonliteral language 18:791–794
notion of 18:789
Parkinson's disease 18:791–792
sarcasm 18:790, 18:792
schizophrenia 18:792
speech acts 18:788–792
Pragmatic inferences 12:39
Pragmatic language impairment (PLI) *see* Pragmatic disorders
Pragmatic method 12:788–789
diversity of interpretation 12:789
philosophical controversy about pragmatism 12:789
Pragmatic objectivity 15:43
Pragmatic reasonings 14:295
Pragmatic relevance 21:147
Pragmatic rules 14:841
Pragmatic theory of observation 9:119
Pragmatics 8:351, 10:332–333, 13:401–402, 14:172, 15:753, 15:826–827
approach to syntactic control 4:831
approach to language disorders 13:331–332
and bilingual aphasia 2:575–578
of deixis 6:56
and discourse studies 6:510, 6:512–514
linguistic 18:795–802
new directions in 15:755
Pragmatics–syntax interface 1:679
Pragmatism 1:893, 5:255, 5:484, 6:690, 7:259, 8:483, 12:936, 15:619, 18:204, 18:685–686, 18:803–807, 20:763, 22:994, 25:558–560, 25:678
Addams on 1:93
after Dewey 18:805–806
versus agnosticism 7:798–799
American 18:808
and ANT 5:525–526
anti-foundationalism 18:806
basic themes 18:808
Boltanski on 2:747, 2:751–753
classical 18:803
contemporary 18:810
and critical theory 5:301
definition 18:808–809
and definition of situation 6:19–20
deflationism 18:806
Dewey on 18:804–805
Dewey's 18:809
empiricism 18:803–806
epistemology 18:804–806
history of 18:808–810
influence of 18:809
James on 18:804
main variants of historical 18:809–810
Mills on 15:533–534, 15:538
Peirce on 18:803–804
and psychology 19:413, 19:417
and realism 20:25
Rorty's contributions 20:761, 20:763–764
and semiotics 21:593, 21:595
truth 18:803–806
Pragmatist social thought 14:497
Prague Circle 14:169
Prague Communiqué 2:742
Prague Linguistic Circle 12:782
Prague School 18:48–49
markedness concept in 14:533
Prague School of Typology 14:209
Prague Spring 18:422, 18:700, 25:838–839
Prasad (ritual) 21:618
Prasis epi lysei 21:401
Praxeological research, and sustainability 23:835–836
Praxis 1:874, 8:900, 18:783–785, 20:83
claim of philosophy to represent 25:599
philosophical theory as highest praxis 25:599
philosophy as wisdom proper 25:599
Prayer 20:356
Prayer, ritual 20:719
PRC *see* People's Republic of China (PRC); Perirhinal cortex (PRC)
Pre- and parasubiculum 10:889
Pre- and posttest, reliability of 3:789
Pre-Action Protocols 4:250
Preadaptation, and disgust 6:547
Preadolescence 7:882–883
Pre-Anthropocene impact 1:722–723
Preboundary lengthening 23:719
Prebsich, Raul 25:749
Precapitalist agrarian formations 1:457
Precariousness 3:56, 3:59, 14:244
Precautionary principle 12:717, 13:507
Precautionary principle, consumer policy 4:734, 4:736
Precedents 13:743
legal databases 5:803
principle of 4:247–250
PRECEDE–PROCEED model 10:621, 19:549
example 10:621
PRECEDE-PROCEED model of health promotion 10:621
Precentral gyrus 3:336–338
Precise utilitarianism 7:89–90
Precision
in describing location on Earth's surface 14:285
Precision Guided Munitions (PGM) 24:143
Precision measurement 21:226
Precision medicine 17:925
Precision weapons 24:955
Precocity 7:159
Precognition 16:170
Pre-Colombian civilizations 15:328–329
Precolonial law, in Southeast Asia 13:814–815
Precolonial nation-states 16:267
Pre-Columbian Mexico, confederation 4:574
Pre-Columbian people, in South America 23:43–44
Pre-commitment, by constitutions 4:712–714
Preconception carrier screening 9:927
Preconceptions 21:275
Preconditions, terrorism 24:240
Precontact times, in Oceania 17:160
Precontact trade, and potlatch 18:714
Precuneus 5:26–27
Pre-Darwinian evolutionism 8:383–384
creation and design 8:383–384
evolution in the enlightenment 8:384
fossil record 8:384
Predator–prey cycles 18:582
Predators
response of animals to 2:800, 2:802–803
Predatory habitus 10:484
Predemocratic judge 12:879–880
Predestination 12:646
Predicate logic 14:319
Predicate–object combination 23:716
Predicates 10:324, 23:901–902
Predication 23:886t
Predicative-logical reasoning 14:773
Predict and control behavior, Watson's contributions in 25:448–449
Predictability
and causality 23:425
McDonaldization 14:828
stock market 23:482–489
Prediction
functional magnetic resonance imaging (fMRI) 9:530

Prediction (*continued*)
and future thinking 9:555–557, 9:559
generalized linear mixed models 9:849–850
hierarchies and markets 10:843
and infancy 12:8
linear mixed models 14:166
of recidivism 20:34–35
sequence 1:536–537
situational judgment test 22:16
Prediction error 8:850
mean-squared prediction error 24:366
Prediction error theorem 1:536
Prediction intervals 8:19, 17:292
Predictive approach 23:414–415
Predictive coding account 25:157
Predictive distributions, expressing population forecast uncertainty via 18:596
Predictive inferences *see* Forward inferences
Predictive modeling, of geospatial data 10:101–102
Predictive validity 6:345
of facets 9:239–240
Predictive validity studies 17:860
Predictive value of a positive test (PVP) 21:328, 21:330
Predictors 25:538–539
abusive supervision 23:703–704
subordinate characteristics 23:704
supervisor characteristics 23:703–704
controlled trials 19:389
Predimensional stage of development 16:504
Predisciplinary period, expertize and public intellectuals in 22:696–698
Predominant breastfeeding 2:848t
Predominantly semantic assignment systems of gender 9:796, 9:796t
PREE *see* Partial reinforcement extinction effect (PREE)
Preemerging gateways, in United States 11:631
Preemption 15:21–22
Preemptive war, deterrence 6:248
Preening 1:852
Preevolutionary Russia, homosexuality in 21:741–742
Preexisting conditions, health insurance markets 10:641
Preexisting subgroups 22:882
Preexposure prophylaxis (PrEP), HIV infection 6:467–468
Prefabricated Extensive Green Roof Tray system 24:896–897
Preference data 16:37, 16:39
Preference efficiency, in majority rule 14:447
Preference models
conjoint analysis 14:865–866
with latent variables 19:440–444
consistency conditions 19:440–441
context 19:440, 19:444
identifiability and interpretation 19:442–443
individual differences 19:443
intransitivity 19:440–442
outlook 19:443–444
paired comparison 19:440–442
predicting choices 19:441–442
random utility maximization 19:441
ranking 19:440–443
revealed preferences 19:440
social dependencies 19:443
temporal differences 19:443
valuation of choice options 19:440–441
Preference responses, in crossover designs 5:351
Preference reversals 2:445
and procrastination 19:75–76, 19:76f
Preferences 16:924, 16:926, 16:928, 17:780
behavioral economics
construction 2:445
modeling dynamically inconsistent preferences 2:447
modeling prosocial preferences 2:447
heuristics for 5:892–893
media, children 15:1–3
and personality 17:781–783
welfare as 25:506–507
Preferences for Activities in Children scale 13:834
Preferencism 7:89
Preferential decisions 5:898–899
Preferential trade agreements (PTAs) 12:568–569, 12:571, 24:485
Preferential treatment 1:249–250
Preferred direction 15:968
Preferred group 1:251–252
Preferred orientation 25:159–160
Preferred provider organizations (PPOs) 15:102
Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA) 15:269–270
Prefigurative politics 22:133
Prefixes 15:821
Prefrontal areas, connections with 18:847
Prefrontal cortex (PFC) 1:653–654, 3:445–446, 4:675, 5:21–22, 8:489–491, 8:493–494, 11:613–614, 13:681, 16:494, 18:811–817, 23:866
academic achievement and life success 18:821
in adolescence 18:820
anterior 19:274
behaviors 18:812–814
in children 18:819–820
and connectome 18:814
definition 18:811–812
development 18:817–823
development, in infancy 18:859
disorders affecting frontal lobe 18:816
dysexecutive syndromes 6:725–726
educational neuroscience 16:728, 16:733
environmental influences in development 18:820
frontal granular cortex 18:811
functional development 18:819–820
in infants 18:819
injury, symptoms of 18:815–816
disturbances of motor function 18:815
emotion and social behavior 18:815–816
environmental control of behavior 18:815
loss of divergent thinking 18:815
temporal memory 18:815
intelligence and frontal lobes 18:814
interoception 12:614–615
mediodorsal thalamic nucleus 18:811–812, 18:816
and memory binding 15:144–145
metaplasticity and 18:821
neurodevelopmental disorders of
attention deficit hyperactivity disorder 18:820–821
autism 18:821
schizophrenia 18:821
ontogenetic development of 18:818–819
postnatal development 18:818–819
prenatal development 18:818
role in cognitive control 4:33
serial order memory 21:639–643
and sexual motivation 21:760–761
size of 4:33f
subregions of 18:817
function and connectivity of 18:817–818
temporal organization of behavior 18:813–814
and working memory 25:706–709
Prefrontal leucotomy 17:269
Prefrontal lobes
importance in placebo responsiveness 18:155
Prefrontal pedagogy 4:47
Prefrontals 16:736–737, 16:740–741
Prefscal (software) 16:46
Pregenomics phase of heritability controversy 8:716
Pregnancy 2:697–698, 10:465–466, 23:556–560
event history analysis for 8:309
nonmarital, resolution 9:16–17
of participants in sexual development analysis 14:330
perinatal psychiatry 17:722–726
unwanted 21:797, 21:799–800
Pregnancy discrimination 9:731
Pregnancy Discrimination Act 1978 6:531
Pregnant women, exclusion in clinical trials 9:720
Prehistoric archaeology 9:766, 14:671
Prehistoric art 18:824–827
birth 18:826–827
human evolution 18:824–826
interpretation 18:826–827
study 18:825–826
Prehistoric research, classical archaeology 3:760
Prehistory 1:746
of ancient Egypt 1:691
Anthropocene 1:722–723
hunter-gatherer archaeology 11:426
and lithics 14:260

migration 15:473
Prehistory of war 4:617
Prehomohysteria 9:657
Preictal period, seizures 7:854
Preimplantation diagnosis, ethics of 20:701
Preimplantation genetic diagnosis (PGD) 2:619–620, 9:927, 9:994–997, 10:35, 20:701
Preimplantation genetic screening (PGS) 9:927, 20:701
Pre-imputation solution 4:868–869
Preindustrial, guilds 2:49
Preindustrial societies 25:669–671
Preinstitutional deserts 6:229–230
Pre-Islamic history 15:360
Preiss, H.A. 9:646
Prejudice 1:563–564, 2:262–264, 2:266–267, 4:154–155, 9:460, 18:828–833
and agreeableness 1:467–468
ambivalent sexism 18:828–829
antisemitism 1:787
conformity 18:831
definition of 19:866
and essentialism 8:13–15
ethnicity/migration 8:142–144
Fromm on 9:460
Heider on 10:753
history of 8:168–174
culture 8:169–171, 8:173
ethnicity 8:171
evolutionary explanations 8:169–170
intergroup relations 8:168–172
personality 8:171–172
self-aggrandizement explanations 8:171–172
and social groups 8:169
social identity theory 8:171–173
social norms 8:172–173
socialization and normative explanations 8:172–173
threat and conflict explanations 8:170–171
and immigration 11:676–680
implications for social policy 18:832
implicit association test 18:829
and implicit social cognition 11:714, 11:716–717
individual 19:858
infrahumanization 18:829
intergroup relations 12:409
and LGBT aging 13:866–867
minorities 15:569–570, 15:572–573
against minorities 20:381–383
nature of 23:462–463
norms 18:828, 18:830–832
in political psychology 18:427–428, 18:430–431
racial 6:542–545, 19:858
versus racism 19:864–865
reduction, and desegregation 6:220–221
relation with discrimination 18:831–832
and residential segregation 20:542
role congruency model 18:828
sexism 18:828–829
and sexism 21:694–695
and social dominance orientation 22:250–251
and social psychology 22:528
toward LGBTQ youth in schools 21:86–87
and xenophobia 25:802–806
Prejudice, in society 18:834–845
as ambivalence 18:834–835
ambivalent sexism 18:835
automaticity and social control 18:835–836
aversive racism 18:834
color-blindness 18:835–837
color-consciousness 18:835–837
content of 18:841–842
corporate response 18:844
dehumanization 18:836
and discrimination 18:839–840
LaPiere study 18:839–840
Merton's typology 18:839
egalitarianism 18:834
elimination 18:843
education 18:843
extent of 18:842
color-blind racism 18:842
social distance scale 18:842
as individual differences 18:834
infrahumanization 18:836
mass media 18:843–844
equal-status contact 18:843–844
meritocracy 18:834
as normal process 18:834
overview 18:844
reduction 18:836–837
contact hypothesis 18:836–837
individualistic models 18:837
reconciliation 18:837
social identity theory 18:834–835
stereotype content model 18:835
stereotype threat 18:834–835
subordinate group, mood of 18:843
theories of 18:840–841
exploitation theory 18:840
normative approach 18:840–841
scapegoating theory 18:840
Preliminaries 23:902–903
Prelinguistic communication 4:300
Prelinguistic language processing in infants 16:589–590
Premack principle 2:409
Premarital cohabitation, and divorce 6:607
Premarital screening 9:927
Premarital sex 11:192–193, 21:791, 21:793, 21:820
Premature birth 2:698–699
Premediated acts, sex offences as 21:674
Premenstrual dysphoric disorder (PMDD) 10:463–464
Premenstrual syndrome (PMS) 10:463–464
antidepressants for 1:780
Premier presidentialism 18:867
Pre-millennialism 15:529–530
development 15:530
Premises (*explanans*) 10:773–774
Premiums
health insurance 10:640–643
insurance 12:211
ownership 19:181–182
Premodern professions 19:96
Premodern public administration 19:521–522
Pre-Modern Towns 24:812
Premonoclus 4:872
Premotor cortex (PMC) 5:25, 5:28, 18:846–851, 23:246, 23:250
classes of areas and cortical connections 18:846–847
connections with parietal, prefrontal, and cingulate areas 18:847
connections with spinal cord and intrinsic motor connections 18:847
functional organization of areas 18:847–850
caudal premotor areas (F2, F3, F4) 18:849
F5 18:847–849
rostral premotor areas (F6, F7) 18:849–850
and human language evolution 11:324
of humans 18:850
mirror neurons 18:848
motor control 18:847, 18:850
representations of movements 18:849
Prenatal care 2:698
Prenatal development 18:852–860
attachment 18:854, 18:856, 18:859
brain development 18:853, 18:859
childhood 18:852, 18:856–857
depth perception 18:853
dynamic systems theories 18:855
future directions 18:858–859
imitation 18:853–854, 18:856–858
Jacksonian principle 18:852
learning 18:852–858
physical and cognitive development 18:852–853
prefrontal cortex development 18:859
reflexes 18:852–853, 18:857–858
social development 18:856, 18:859
sudden infant death syndrome 18:853, 18:857–858
transactionalism 18:855, 18:858
Prenatal diagnosis 9:976–977
ethics of 20:701–702
Prenatal discrimination 15:609
Prenatal genetic diagnosis 9:994–995
Prenatal genetic selection
for avoiding disease 10:35
for genetic enhancement 10:35–36
state-sponsored 10:36
Prenatal genetic testing, and reproductive rights 20:473–474
Prenatal hormone levels 21:774
see also Hormones
Prenatal screening 9:927, 9:929
ethics of 20:702
Prenatal sex testing 21:686–687
Prenatal stress
altered child outcome 23:557

Prenatal stress (*continued*)
effect on fetal and child development 23:556–561
Prenatal stressors, and personality 17:912
Prenormal science 13:134–135
Prenucleolus, prenuc 4:869
Preobrazhenskii, Nikolai 21:741
Preoccupied adults 2:139
Preoperational children 18:127–128
Preoperational stage of development 16:501
PREP *see* Peer Review Evaluation Process (PREP)
PrEP *see* Preexposure prophylaxis (PrEP)
Preparation for future learning (PFL), transfer of learning as 21:98–99
Preparatory attentional and memory processes (PAM) theory 19:272–273
Preparedness
collective action 4:147, 4:149
for disaster 6:481
and natural disasters 16:316, 16:318
Prepilot interviews, scale development 7:932
Preplanned (research) questions in mixed methods research 15:620
Preposition 23:885
Pre–post designs 22:292–293
Prepotenza 14:425
Prepredicative experience of life-world 11:443
Prepression 5:246
Preprocessing, functional magnetic resonance imaging (fMRI) 9:527–528
Pre-prohibition rate 17:240
Prepulse inhibition, of startle 23:332
Pre-Raphaelite artwork 11:519
Preretrieval 20:610–611
Presage teacher characteristics 24:44–45
Presalernitarian medicine 15:119
Presbyterian form of governance 20:409
Presbyterians 3:542–543, 20:397
Preschool age
play during 18:222–224
differences in cultural goals and practices 18:224
differences in ecological conditions 18:223
differences in interpersonal contexts 18:223–224
similarities in nonsocial peer play as maladaptive 18:223
similarities in object use, pretend play, and sex segregation 18:223
similarities in role of individual characteristics 18:223
Theory of Mind development in 24:268–269, 24:272
Preschool education, Alva Myrdal 16:160
Preschool Language Scale 3:795
Preschoolers 21:276, 21:278–279
imitation by 11:621, 11:623–624
Prescribing decisions, inappropriate, in combined pharmacotherapy and psychotherapy 19:497
Prescribing Safety Assessment (PSA)
in psychiatric education 7:189
Prescription system 6:670
Prescriptive market areas, market geographies 14:546
Prescriptive meritocracy beliefs 15:242–243
Prescriptive normativity of meaning (PNM) 14:844, 14:846–847
Prescriptive procedures, heuristics as 10:829–830, 10:829f
Prescriptive theories
of historical explanation 11:10
of human explanation 11:11–15
Prescriptivism 14:841
Presentation 3:162–165
graphics 10:358
Presentation software
Computer Assisted Instruction (CAI) 4:493–494
Presenting time
in space–time cube 6:716
by time 6:716
by visual encodings in maps 6:715–716
Preservation
of archives 5:702
cultural resource management 5:488–490
and dance 5:668–669
Preservationists 7:810
Presidency
meaning 18:861
United States 18:861–863
behavioral revolution 18:862
current theory and research 18:862
future directions 18:863
institutional approaches 18:861–863
intellectual context 18:861–862
methodological and theoretical problems 18:863
Neustadt on 18:861–862
presidency-centered perspective 18:862
public law tradition 18:862
rational choice theory 18:862–863
scholarship, changing focus of 18:862
values of presidency scholars 18:862
Presidential Commission, *Challenge of Crime in a Free Society* 5:235
Presidential elections, United States 18:394–395
Presidential Executive Decree in 2007 13:498–499
Presidential systems 10:314
executive government 8:497, 8:500
Presidential task force 5:636
Presidentialism 18:864–869
coalitions 18:865–868
democratic legitimacy 18:864–865
Fifth Republic 18:867
parliamentarianism 18:864–868
plurality rule 18:865
semipresidentialism 18:867–868
winner-take-all 18:865
Presidents 10:314
United States, and Congress 4:652
Presoldier experience 3:431–432
Press, Andrea, works, *Women Watching Television* 2:219
Press freedom 9:392–397
access to information and free press *versus* fair trial 9:396
censorship of 9:393–394
administrative *versus* judicial censorship 9:393
international law on prior censorship 9:393
self-censorship 9:394
defamation law 9:394–395
criminal defamation 9:395
Internet defamation 9:395
transnational defamation litigation 9:395
definition 9:392
preventive counseling for 9:396–397
privacy 9:395–396
right to reply 9:394
and science communication 21:221–222
theories of 9:392–393
libertarian concept of 9:393
social responsibility theories 9:393
Press subsidies 9:394
Presser, S., on interviewing 12:671
Press–state relations 16:817–818
Pressure groups 12:371
Pressure-state-response (PSR) 13:242
Prestige 3:755–756
artistic 5:468
and diglossia 6:420
and gift 10:140–142
and language 13:282, 13:284–286
potlatch and 18:714–717
and public recognition 2:333–334
and social stratification 22:707, 22:709
Prestige bias 24:130
Prestige goods 15:257–260, 15:263–264
production 12:331
Prestige leader 2:269
Prestige of associations, of associations 25:258
Prestige structures 8:899
Preston curve 19:813
Prestructuralist semantics 13:930–931
Presumption of innocence 5:228, 5:230
Presuppositions 12:95–96, 16:410, 18:796, 21:589–590
Presurgical patients, lateralization in 13:406
Presymptomatic testing, ethics of 20:702
Presynaptic depression process 10:482–483
PRET 21:295
Preteceille, Edmond 12:913
Pretend play 18:870f
in cross-cultural context 18:870–872
in Brazilian communities 18:871
in geographically distinct communities 18:871–872
developmental correlates of 18:872–875
creativity 18:874
language 18:873
social competence 18:872–873
theory of mind 18:873–874
role in cognitive development during infancy and early childhood 18:870–876
imagination 18:870, 18:874

pretense 18:870–875
universality 18:870
similarities, during preschool age 18:223
Pretense
cognitive development 4:40–41
and pretend play 18:870–875
Preterm births 12:18–20
influences in adulthood 6:753–754
Pretest (pretreatment) assignment covariate 20:155
Pretesting
model selection 15:650
questionnaires 19:775
of survey questionnaires 20:881–882
Pretext approach, for employment discrimination 6:531–532
Pretransition fertility 11:7
Pretreatment group 20:152
Pretrial conference 8:585
Pretrial decisions 5:186
Pretrial publicity 13:748
Pretrial release 5:185–186
Prevalence 5:633
of childhood friendship 9:442
defined 15:179
definition of 15:179
of disease, race and ethnicity 19:812, 19:818
of elders 7:331
epidemiology 7:824–825, 7:825f
of friendship in early childhood 9:450
of identity theft 5:634
of obsessive-compulsive disorder 17:119–121, 17:123
Prevalence rates, of associations 25:253
Prevalence-specific phobias 18:38
Preventing Alcohol Trauma (PAT) 10:649–650
Preventing faking 17:837–838
fake resistant measures 17:837
reasoning with test-taker 17:837–838
Prevention 10:620–621
alcohol use 1:515
of disease 19:545–546
of diseases in children 10:682
programs 10:684
mass killings/genocide 14:705
and morbidity compression 4:463–464
of oral disease 17:280
personalized medicine 17:926
perspective on mass trauma 14:748
university hazing 10:559
Prevention and management strategies, cultural psychiatry 5:476
Prevention research 18:883–890
Prevention science 18:883
articulating strategies 18:884–885
primary prevention 18:884–885
secondary prevention 18:884
tertiary prevention 18:884
conceptual foundations 18:883–885
prevention to mental health 18:884
public health models 18:883–884
development of 18:885–886
prevention and promotion 18:885–886
resiliency 18:886
risk 18:886
emerging directions 18:888–889
cultural diversity 18:888–889
prevention and public policy 18:889
program fidelity *versus* program innovation 18:888
methodological challenges 18:886–888
assessing effectiveness 18:886–887
cost-effectiveness 18:887
models for prevention beyond mental health 18:885
pandisciplinary approach 18:885
program implementation and dissemination 18:887–888
delivery system 18:888
support system 18:888
synthesis 18:887
translation system 18:887
Preventive check 11:5–7
Preventive coping 4:903
Preventive dimension 13:422
Preventive diplomacy 17:632
Preventive patrol 18:879
Preverbal infants 2:136
Preverbal memories, during childhood
long-term fate 3:467
verbal translation 3:467–468
Preview benefit, eye fixations 8:633
Pre-vocabulary competencies, acquiring 25:226–227
Prey-predator dynamics 18:567
PRI *see* Perceptual Reasoning Index (PRI)
Price, D.H., works, *Counterinsurgency Field Manual* 11:396
Price, Derek J. de Solla 21:322
Price, Herwith, works, *Economic Dictionary* 7:590–591
Price, M., works, *Owning the Olympics* 15:52
Price cap regulation 20:168
Price cost margins 14:550
Price determination 1:479
Price earnings (P/E) multiples
and stock market returns 23:484–485
Price elasticity of demand 12:163
Price equation, kin selection 13:26
Price fixing, and antitrust 1:797–798, 1:801–802
Price inflation 15:731
Price Revolution 14:407
Price setting 7:69
Price theory, Stigler's contributions 23:458
Price Waterhouse v. Hopkins 24:551
Price–quantity diagrams, microeconomics 15:333
Prices 25:7–8, 25:10
Black-Scholes model for 5:540
cultural understandings of 5:538–539
economic statistics 5:758, 5:760, 5:762
of production 25:8–9
theory of 14:407–408
Prices design 20:166
Price-to-book-value (P/BV) ratios, and stock market returns 23:485
Pricing
and alcohol and consumption 1:520
congestion 24:632–633
airport runways 24:632–633
road pricing 24:632
microeconomics 15:336
neoclassical economics 18:359–360
nonlinear 20:166–167
with multiple outputs 20:167
and quality 20:167
and regulation 20:178–179
Pricing behavior, effect of minimum wages on 15:564
Pricing tradition, finance 9:178
Pride 7:469–470, 21:478
cultural differences in 21:478–479
in intergroup relations 7:504
national 16:312
and self 21:441
and shame 21:872–873
types 21:478
Pride parades 15:110–111
gay/lesbian 9:640–641
in South Africa 21:747
Priesthood 20:410
Priestley, Joseph 21:241
works, *The History and Present State of Electricity* 21:241
Prieto, Maria Rosario 3:868
Prigogine, Ilya 7:622
Primacy, to individuals 15:302–303
Primacy effect 25:712–713
Primary acoustic startle pathway 8:860–861
Primary appraisal, and coping 4:903
Primary care, prevalence of mental disorders in 15:181
Primary care mental health 15:191
Primary care sector
consultation–liaison psychiatric services in 14:3–4
psychiatric comorbidity in 14:2
Primary commodities, in Africa 24:967, 24:972
Primary control
compensatory 4:814–815
in life management 14:87
and self-efficacy 21:505
striving 4:813–814
Primary control behavior 4:813–817
Primary cultural undertaking, in East Asia 11:48
Primary data 5:692
Primary deviance 22:236
Primary drive reduction 12:228–229
Primary education 7:172–173
LAC educational system 7:178–181
Primary elections, United States Congress 4:649–650
Primary emotional development 24:276
Primary emotions 6:747, 21:425
Primary energy 7:626, 7:628–630
Primary Genealogical Space (PGS) 13:58–59
Primary groups 22:100–101
group process 22:541
Shils on 21:884–886

Primary health care
for child/infant 12:25
psychiatric care 6:43–45
Primary infertility 12:45–47
Primary key species 7:785
Primary kinship 13:53–55
Primary market, job search and working in 8:110–111
Primary mental ability theory of intelligence 12:291
Primary metropolitan statistical areas' (PMSAs) 24:857
Primary motor cortex 5:24, 8:668
Primary negative symptoms 19:292
Primary prevention (mental health) 18:884–885
Primary prevention programs 18:878
Primary progressive aphasia (PPA) 13:361
language processes in 13:363–366
nonfluent/agrammatic variant (nfvPPA) 13:364
Primary religions 20:388
Primary sampling units (PSUs) 13:386
Primary schools
and gender 7:117–120
Primary signed languages 21:928
Primary socialization 16:234
Primary taste cortex 24:27
Primary traits 9:236
Primary visual area (V1) 25:201
Primary visual cortex (PVC) 2:828, 2:830
Primate behavior 18:891–896
conceptual shifts in 18:892
Primate biology 18:891
Primate cognition 10:526–527
primate neurobehavioral studies 25:201
Primate research 10:527
Primate visual system 25:158
Primates
advertised male quality 18:899
advertised ovulation in multimale groups 18:899
benefit of multiple paternity to females 18:900
cost of multiple paternity for females 18:900–901
costs of living socially 18:899–900
determinants of sociality in 18:899
dispersal 18:900
diversity, social organization, and mating systems of 18:897
and enriched environment 2:838–839
factors determining primate mating systems 18:901
female bonded 18:900
female spacing 18:900
future directions 18:902–903
genetic consequences of mating patterns 18:898
genomic conflicts 18:901
incest avoidance 18:902
molecules and behavior 18:901
monogamy 18:897, 18:901–903
polyandry 18:897
polygamy 18:902
postmating female choice and sperm competition 18:901
reflections on human groups and families 18:901–902
reproductive strategies of 18:897–898
sex as a social tool 18:898
Primates, evolution of 18:904–910
colugos 9:941, 18:908
euarchontoglires 18:908
fine-branch niche 18:906
fossil primates 18:908
geographical distribution 18:904
haplorhines 18:907–908
living primates 18:904–905
definition of 18:905–906
phylogenetic relationships among 18:906–908
molecular evidence 18:904, 18:906–908
phylogenetic reconstruction 18:904, 18:906
primate fossil record 18:908–910
prosimians 18:904–908, 18:907f
simians 18:905, 18:908, 18:907f
SINEs strepsirrhines 18:907–908
tree-shrews 18:904–905, 18:908–909, 18:909f
Prime ministers 10:314
executive government 8:497–501
Prime of life 14:78
Primeval atom 20:288
Priming 13:678–679, 16:845, 25:175
in attitude measurement 2:180
for early memory development 6:762
in mass media 14:715
money 15:737
and motivation 15:938
process of 1:355
Priming distance effect (number representation) 17:48
Priming effects theory 25:100
Primitive accumulation 1:456, 1:458, 1:460
Primitive capitalists 2:569
Primitive communism 15:714, 19:153–154, 22:838
Primitive concepts 2:339
Padoa's principle for proving independence of 2:340
Primitive counterpart to science, myth as 16:173
Primitive cultures 24:646
Primitive currencies 15:746
Primitive domination 2:752
Primitive man 21:616
Primitive patriarchy, theory of 3:730
Primitive peoples 14:435–436
and religion 20:307
Primitive religion 2:507–508, 14:825–826
Primitive societies 13:37, 18:911–914, 23:371
and ethnocentricism 8:166
livelihood 18:912
mentality 18:913–914
modernity 15:687
nostalgia 18:914
social structure 18:912–913
Primitive sports 23:299–300
Primitive thought, and taboo 24:1
Primitives 3:785–786
and anarchism 1:681–682
of decision behavior 17:780–781
and magic 14:435–437
in Oceania 17:158
Primitivism 24:456
and anarchism 1:683–684
Primogeniture 21:923–924, 21:926
in sub-Saharan Africa 23:657–659
Primordial sociobehavioral outcomes (PSOs) 22:878–879
justice 22:879–881
power 22:879–881
rates of change 22:879–880, 22:880f
constant rate 22:879
decreasing rate 22:879
increasing rate 22:879
status 22:879–881
Primordialism 8:150–151, 8:157
and nation 16:224–225
Primordialists 16:224
Prince Stupid 3:147
Princeton Office of Population Research 6:173
Princeton Principles on Universal Jurisdiction 18:501
Princeton Radio Research Project 13:636, 13:639
Principal component analysis (PCA) 5:1, 8:659, 9:888–889, 9:937, 10:100–101, 11:292, 15:627, 16:119, 19:434, 19:730
functional data analysis 9:516
of traits 2:563
Principal Hessian Directions (pHd) 14:160
Principal inertias 5:2
Principal Investigator (PI)
academic research and employment 1:50
Principal stratification
causal mediation analysis 15:78–80
Principal-agent literature 19:536
Principal–agent (P–A) models 1:348–349, 12:104, 19:535–536
with asymmetric information 12:104
Principal–agent relationship 1:345–346
Principal–agent theory 1:344, 20:182–183
of rationality 19:911
Principal-axes method, in factor analysis 8:705–706
Principle of accountability, in linguistic code-switching 3:921
Principle of allocation 9:40
Principle of communication 19:670
Principle of compositionality 14:294
Principle of divergence 21:925
Principle of Domain Integrity (PDI) 25:652
Principle of extendibility, in vocabulary acquisition 25:228
Principle of Head Proximity (PHP) 25:652
Principle of hedonism 6:255
Principle of hermeneutic equity 10:781
Principle of homogamy 11:166–167
Principle of impartiality 5:103
Principle of Indifference 19:24

Principle of individualistic transmission 16:331–332
Principle of invariance 12:707
Principle of judicial independence 5:103
Principle of majority wins (linguistics) 4:424
Principle of nationalities 16:243
Principle of openness 19:670
Principle of party participation 5:103
Principle of patient autonomy, and euthanasia 8:282
Principle of provenance 5:701
Principle of solidarity, in clinical genetics 2:622
Principle of symmetry 4:167
Principle of the rule of law 5:103–105
Principle of transfers 11:903, 11:905–906
Principle-based social work ethics 22:783–784
Principled model of negotiation 16:418
Principles
  of benevolent interpretation 10:781
  empowerment evaluation 7:579
    accountability 7:579
    capacity building 7:579
    community knowledge 7:579
    community ownership 7:579
    democratic participation 7:579
    evidence-based strategies 7:579
    improvement 7:579
    inclusion 7:579
    organizational learning 7:579
    social justice 7:579
  of government statistics 10:302–303
  legal reasoning 13:776–782
  likelihood 14:117
Principles and parameters theory 9:875–876, 10:317–318, 16:1
Principles for the Older Person (POP) 11:357
Principles of justice
  application 19:915–916
  Rawls on 19:915
Principles of Knowledge Representation and Reasoning 14:319
Principlism ethics 8:31
Principlist approach 2:633
Print capitalism 16:294
Print culture 18:918, 19:621
  media 15:14
Print logic 18:918–919
Print newspapers, and community 4:333
Printing 14:242–243
  early practice 18:915–916
  industrial 18:916–917
  invention of 18:915
  as medium 18:915
  as a medium 18:918–919
  revolution in 18:917
  social–scientific interpretations of 18:915
  theories about 18:917–919
Printing house 18:915
Prior criminal record, and homicide 11:166–167
Prior distributions
  Bayes' theorem 2:379–380, 8:17, 8:523
  Bayesian inference 2:373–376
  elicitation of 7:383–384
Prior knowledge 12:327
Prior probability of the risk of arrest 6:253
Prioritarianism 7:91
Priority disputes 17:12, 21:281–282
  and de-Westernization 25:548
Priority setting, in modern nations 10:605–606
Prism adaptation 16:845
Prism biography 2:646
PRISMA *see* Preferred Reporting Items of Systematic reviews and Meta-Analyses (PRISMA)
Prison abolition 22:390
prison abolition movement 22:386
Prison code 17:442
Prison industrial complex 18:927–928, 22:390
Prison life 11:723
  authoritarian institutions 18:921
  California prisons 11:723
  in England and Wales 11:723
  Foucault on 18:922
  in Haiti 11:723
  regimes 18:921–922
  sociology of 18:924
    black and minority ethnic prisoners 18:927
    coercion 18:925, 18:927
    convict criminology 18:926–927
    culture 18:924–925, 18:927–928
    dominant theoretical models 18:925
    ethnography 18:925–928
    foundations 18:924–925
    mainstream studies in UK 18:925–926
    power 18:924, 18:926
    studies 18:927–928
    women's prisons 18:926
  supermax prisons 11:723
Prison works 19:45–47
Prison/crime ratio in France 19:639f
Prisoner biomedical research 13:562–563
Prisoner-per-robbery ratio 19:646
Prisoners
  characteristics 5:123–124
Prisoner's dilemma (PD) 4:682–683, 4:863, 11:171, 12:725, 16:584–585, 17:420, 18:930–937, 19:908, 22:245, 22:702, 4:863f
  backward induction 18:934
  computer simulations 18:935
  cooperation 18:931–932, 18:934–936
  decision theory 18:933–934
  evolutionary 18:936
  example of Nash equilibria 9:577
  matrix *versus* situation 18:931–932
  Nash equilibrium 18:931–933
  Newcomb's paradox 18:933–934
  Pareto optimality 18:931–934
  reciprocity 18:932
  and social capital 22:176–177
  Stag Hunt game 18:932
  Tit-for-Tat 18:935–936
  two-person, iterated 18:934–936
    finitely iterated 18:934–935
    indefinitely repeated 18:935–936
  two-person, one shot 18:930–934
    N-persons 18:934
    rationality considerations 18:933–934
    from social phenomena to game theory 18:931–933
Prisoners of war (POW) 1:612, 13:532
Prisons
  antiprison professionals and experts, role of 19:642–645
  birth of 19:655–656
  and budgets 19:641–642
  governance issues in 19:653
  and imprisonment 18:938–941
    comparative incarceration rates 18:939
    effective correctional strategies 18:940–941
    origins 18:938–939
    potential improvements 18:940–941
  in MENA 16:364
  minorities in 19:643f–644f
  opposition to 19:643–644
    in Finland 19:644
  robbery prison rates 19:647f
  torture and 24:443–446
  treatment of women in 25:618–619
  unemployment rates *versus* prison rates 19:645f
  utilitarian considerations 19:642
Pristine deformation 20:919
Pritchard, James, on ethnology 1:746
Pritchett, Lant, on family planning 8:787–788
Privacy 3:687, 8:80, 9:395–396, 12:608–610, 19:606, 19:608, 20:514, 23:662, 23:664, 23:727–728
  autonomous agents 2:312
  biomedical research ethics 8:80–81
  Common Rule 4:253–254, 4:256
  critiques of 18:950
  as developing field 18:947
  in ELSI research 8:45–46
  feminist critique 18:950
  in genomics 10:40
  geography of 23:727–732
  in health surveys 10:729
  human rights and 18:950–951
  of human subjects in research 8:49
  of individuals in social research 18:942–946
    data sharing and researchers' use of statistical databases 18:944–945
    definitions 18:942
    ethical and legal context 18:942–943
    future 18:946
    informed consent 18:946
    professional codes 18:942
    record linkage, and integrated data systems 18:945–946
    research subjects 18:943
    statistical disclosure limitation 18:944–945

Privacy (*continued*)
surveys, randomized field trials, and other data collection efforts 18:943–944
use of state databases in United States 18:945
informed consent 12:113
and Internet 12:593–599
legal protection of 18:947–951
media ethics 15:44
research 20:510, 20:512
scope of 18:947–949
and secrecy 21:377–378
in social media use 24:148–149, 24:151
and surveillance 23:736–737
Sveningsson's model 12:609, 12:609f
value of 18:949
Privacy commissioners 12:596
Privacy impact assessment (PIA) 22:875
Privacy laws 18:947–951
Privacy regulation 20:170
Privacy rights
neuroethics 16:666–667
Privacy theory 18:951
Private adoption 1:136
Private Attorney Involvement (PAI) 19:558
Private attorneys, and access to justice 12:921–923
Private bargains 4:811–812
Private car 2:303–304, 2:307
Private cognition 15:546–547
Private discipline 23:863–864
Private estates, in archaic states 23:372
Private family 9:768
Private funding, child care 3:393
Private health insurance 10:615–616, 10:663–664, 10:666–667
Private industry
approach to personnel management 11:350
engineering 7:644–645
Private insurance 22:362–363
health care markets 10:599–600
Private insurance industry 12:218–219
Private international laws 13:423, 13:614–615
see also Conflict of laws
Private investment markets, access to 12:118
Private irony, Rorty's concept of 20:765
Private justice 12:470–471
Private law 3:27, 3:669, 13:743–744, 19:184–185, 19:187–188
Latin America 13:410–413
of technology 13:525–526
Private legal systems
constitutive legal theory and 13:810–811
interrelationship between state law and 13:809–811
mechanisms of control 13:807
models of justice operative within 13:808–809, 13:809t
normative orders 13:807
semiautonomous quasi-legal set 13:807–808
significance 13:807–808
as a type of law 13:808
use of term 13:807
Private lives
and emotions 7:516
and family 8:809
of scholars 21:34
Private lives of scholars 21:34–38
Private miracles 15:576
Private moral features, and responsibility 20:590
Private orderings 4:864, 13:917
Private pensions 20:603–605
Private policing 13:807, 18:248
Private practice
infrastructure investment as 12:117–118
of lawyers 13:629–630
Private prisons 13:807
Private property 13:233
and division of labor 13:159
Private provision, international public goods 19:540
Private research organizations 24:762
for social and behavioral science research 22:619
Private rights *see* Property
Private rights of action 13:818–820
Private schooling, educational policy 25:482–483, 25:487–488
Private schools 21:73–74
in India 1:270
Private science 17:13
Private secrets 21:377
Private sector
applied demography in 1:840
components 1:840
household characteristics 1:840
personal demographic characteristics 1:840
psychographic component 1:840
employment in 10:519
Private Securities Litigation Reform Act 1995 3:735
Private security 18:270
Private space, and gender 9:711
Private speech, Vygotsky on 25:313
Private sphere
and households 11:231
and public sphere
and liberalism 14:23
sexual citizenship and 21:729
Private sponsors, obligations to, surveys and polls 23:753
Private urban governance 9:612–614
Private urban governments, mass growth of 9:612
Private value auction model 2:206–208
independent private values 2:206–207
strictly affiliated private values 2:206–207
Private vices, and public benefits 12:388
Private voluntary organization (PVO) 16:856
Private-sphere environmentalism
impact-oriented approach 7:701–702
intent-oriented approach 7:700–701
relation between intent- and impact-oriented measures 7:702
Privatization 6:270, 7:267, 10:258–259, 11:234–235, 14:736, 16:464, 19:157–158, 19:612, 20:175, 20:817, 24:790, 24:793, 24:846–848
of army 6:3
in artistic/cultural markets 14:566–567
of Chinese cultural industries 5:597–598
of civil service 3:698–699
and economic reform 18:692–693
economic transformation 7:73–77, 7:79
of enterprises 3:24
and institutional transformations 18:188–189
of judicial system 15:73
land rights and tenure security 13:234
and probation 19:47
and property 19:156–157
of religion thesis 21:386–387
in transition 7:75–76
of water 23:48
Privatization thesis 25:636
Privilège du prêteur de deniers 21:403
Privileged location thesis 6:245–246
Privileges 18:916
discovery 6:517–520
engineering studies 7:646–647
and intersectionality 12:657–662
of men 14:680–684
white-collar workers 25:565–566
Privy Council 4:248, 23:59
Prizes 2:333–335
international 2:334–335
social science and humanities prizes 2:335
PRM *see* Greater Romania Party (PRM)
Pro- and contra-hedonic affect-regulation motivation 1:241–242
age-related differences 1:244
cognitive demands 1:243–244
cultural differences 1:245
individual differences 1:244–246
instrumental-affect perspective 1:242–243
mixed-affect perspective 1:242
self-esteem 1:245
trait-related differences 1:245
Pro bono legal services 12:922–923, 19:558–559
and poverty law 18:737–738
Pro bono system 18:737–738
Pro forma court action 14:267
Proactive coping 4:903
Proactive interference (PI) 15:160, 16:847
AB-AD design to produce 19:1t
cue-overload account of 19:5–6
declarative memory 19:3
in long-term episodic memory 19:3
in semantic memory 19:3
in different populations 19:5
discrimination account of 19:6–7, 19:6f–7f, 19:7t
distinctiveness account of 19:6–7
forgetting 19:3–5
history 19:1–2, 19:2f
and multiple memory systems 19:1–2

and nondeclarative memory 19:2–3
in perceptual representation system 19:2
in procedural memory 19:3
in short-term/working memory 19:3–5
in Brown-Peterson paradigm 19:3–5, 19:4f, 19:5t
in focus of attention 19:5
susceptibility to 6:720
theoretical accounts of 19:5
Proactive mobilization of law 13:594–595
Proactive prosocial behavior 19:221–222
Proactive sexual desire 21:762
Proactivity, and successful aging 23:668–671
Probabilistic approach 2:36
Probabilistic epigenesis, personality development 17:879–880
Probabilistic expectations 19:778–779
elicitation of 19:778
inference in economics 19:777
Probabilistic explanation, statistical-relevance model of 8:595–596
Probabilistic grammars 19:9–15
computational linguistics 19:9, 19:11
context-free grammars 19:12–13
computation 19:13
inference 19:13
probabilities 19:13
Gibbs distributions 19:13–14
inference 19:14
probabilities 19:14
relation to probabilistic context-free grammars 19:14
interpretation 19:9–12
regular grammars 19:11–12
computation 19:12
inference 19:12
probabilities 19:11–12
speech recognition 19:9–12
statistics 19:11–14
stochastic grammar 19:9–10
syntactic coverage 19:9
transition probabilities 19:12
translation 19:9, 19:12–13
Probabilistic inference 2:365, 2:368, 2:370
Probabilistic knowledge representation 23:411–412
Probabilistic learning, cognitive development 4:38
Probabilistic population forecasts 18:596
Probabilistic projection methods 6:167
for small areas 6:167
Probabilistic reasoning 16:926–927
Probabilistic theories 3:269–270
Probabilistic thinking 19:16–22
assessments 19:17–18
chance 19:16–18, 19:21
example 19:18–20
heuristics and biases approach to problems of 19:17–18
history of thought 19:16–17
implications 19:20–21
risk 19:17
Probabilities 3:271, 10:698–699, 16:641, 19:33–42, 20:927, 23:468
Bayesian inference 2:373–376
central limit theorem 19:36
chance 19:34
conditional 19:25
of counterfactuals 5:96–97
De Finetti on 5:857
demographic 6:110
distribution function and related quantities 19:34
elicitation of 7:382–385
and environmental cognition 7:708, 7:710–711
epistemic probability 19:24–25
classical interpretation 19:24
logical interpretations 19:25
need for objective interpretations 19:25
subjective interpretations 19:24–25
fiducial 9:124–126
hypothetical limiting relative frequency 19:26–27
and ignorance 11:579
independence 19:35
extending 19:35
information theory 12:107–108
invariance 12:705–706
and irrationality 12:719–722
Kolmogorov's axiomatic foundations 19:33–34
law of large numbers 19:35
versus likelihood 14:116
majorization/stochastic orders 14:449, 14:451–452
mass functions and densities 19:34
mathematical expectation 19:34–35
objective interpretations 19:25–27
best system interpretation 19:27
finite relative frequency interpretation 19:25–26
limiting relative frequency interpretation 19:26
propensity interpretation 19:26–27
theoretical role interpretations 19:27
philosophical aspects 19:23
determinism and chance 19:23–24
interpretation 19:23
qualitative 19:37–38
formal theory 19:37
quantitative 19:37–38
formal theory 19:37
interpretations 19:38–41
random variables 19:34
randomness 19:33
sampling 20:877
stochastic processes 19:36
structural 9:126–127
uncertainty 19:33
Probabilities, of choice 5:895–899
Probabilities proportional to their size (PPS) sampling 13:386
Probability and My Life (article) 5:857
Probability approach 7:9
Probability density estimation 19:29–32
advanced algorithms 19:31
bin width construction 19:30–31
histograms and CDFs 19:29–30
multivariate densities 19:31–32
other materials 19:32
visualization 19:31
Probability density function (PDF; p.d.f.) 5:903, 6:571, 6:575, 11:886, 12:800, 23:262
Probability distributions 6:571, 6:573, 13:395–396, 13:395f–396f
elicitation of 7:382–385
Probability generating function (pgf) 6:581
Probability interpretation, of basic statistical model 10:836
Probability mass function (pmf) 6:580
Probability plotting, order statistics 17:293
Probability sample survey, census 3:302
Probability sampling 5:321, 13:385
for sexuality studies 15:314, 15:316–317
Probability theory 5:895, 13:382–384
Probability-and-penalty regimes 13:574
Probable error 13:691, 19:472
Probation 19:43–48
alternatives to custody 19:44–45
Anglo-American model 19:43
criminal justice 19:43–45, 19:47–48
diversion 19:44–47
effectiveness 19:44–46
origins and principles 19:43–44
policy choice and future of probation 19:47
prison works and case of England and Wales 19:45–47
privatization 19:47
and punishment 19:656, 19:658–659
punitiveness 19:46
rehabilitation 19:43–48
reintegration 19:44, 19:47–48
resettlement 19:44
response to nothing works 19:44–45
responsivity 19:45
risk–need–responsivity model 19:45
in United States 20:31–32
what works 19:45, 19:47
Probation diversion 4:373
Probation of Offenders Act 1907 (UK) 5:209, 5:241
PROBA-V (Probing Vegetation) mission 20:423
Probe measures, cardiovascular reactivity 3:119
Probing, focus groups 9:278
Probing Vegetation *see* PROBA-V (Probing Vegetation) mission
Probit function, logistic regression 2:588–589, 2:594–595
Probit model 16:132–133
Problem areas in diabetes survey (PAID) 6:338t
Problem articulation 19:52
Problem behavior model 1:122
Problem definition 7:268–269
Problem housing estates 5:147
Problem of aggregation 23:176
Problem of counterfactuals 5:95
Problem of description, and ethnomethodology 21:194–195
Problem of difference 5:586

Problem orientation 24:760
Problem selection
  in social sciences 19:49–55
Problem selection in sociology
  contradictions and issues 19:52
  Durkheimian problem structure 19:49–50
  overarching problems and subproblems 19:49
  problem articulation 19:52
  problem structures and their dynamics 19:49–52
    theories 19:51–52
  questions in
    descriptive questions 19:53
    descriptive to explanatory questions 19:52–53
    erroneous and well-stated questions 19:53–54
    superficial to deeper explanatory questions 19:53
    undertheorized questions 19:54
  Weberian problem structure 19:50–51, 19:50f
    revised 19:51
Problem solving 15:282, 15:286, 15:775, 19:61, 19:65–66, 22:588–589, 22:592, 22:594
  by animals 13:123–125
  and bounded rationality 2:773–775
  in children 19:70
  dynamic decision making 6:710, 6:713
  and ergonomics 11:300
  heuristics for 10:829–830
  higher nature of 19:70–71
  insights and 19:66
  and intelligence 12:313–314, 12:317, 12:319–321
  Köhler on 13:124
  and mathematics 14:816–820
  in school learning 1:871–872
Problem solving and reasoning
  case-based 19:56–60
  psychology of 19:61
Problem standards, business codes 8:22
Problematic reproduction of capitalism 14:652–654
Problématiques 19:334
Problematization, in critical social work 5:291
Problem-based learning (PBL) 15:81–82, 22:7–8
Problem-focused coping 4:903–904
  chronic illness 3:550
Problem-orientated policing 18:879
Problem-oriented policing 6:264
Problem-solving model of negotiation 16:418
Problem-solving techniques
  for asthma 20:578
  for chronic obstructive pulmonary disease 20:578
Problem-solving therapies 4:543–544
Problem-solving workshop 4:586
  characteristics and goals 4:587
  and intergroup contact hypothesis 4:587–588
  transfer and effectiveness 4:588
Proboscis 15:156
Procedural democracy 13:440
  Dahl on 5:658
Procedural dimension, of law 13:604–605
Procedural due process 6:697–698
Procedural fairness 6:696
Procedural instruments 18:290–291
Procedural invariance principle 2:445
Procedural irregularities 12:137
Procedural justice 4:456–458, 7:538, 12:939, 12:950, 13:749–750, 17:380–381, 17:718
  boundary conditions for 4:458–459
  effect on perceptions and experiences of power 4:459
  in fostering legitimacy and cooperation 4:460
  in health care rationing 10:605
  and justice 12:928, 12:930
Procedural knowledge 11:556, 11:558–559, 12:326, 13:708–709, 22:15–16, 22:728
Procedural liberalism 20:895–896
Procedural memory 1:399, 13:678–679, 22:63
  and aging 1:399
  proactive interference in 19:3
  retroactive interference in 20:614–615
  see also Nondeclarative memory
Procedural metacognition 15:283–285
Procedural models of learning 7:415–416
Procedural natural law 12:933–934
Procedural rationality 2:444, 5:911
Procedural regularity 6:696–697
Procedural theories of autonomy 2:314–317
  authenticity as historical nonalienation 2:316–317
  endorsement 2:316
  identification 2:316
  self-discovery 2:317
Procedural value 4:878
Procedure simplification, for access to justice 12:923–924
Procedures 14:266–267
  and science 21:250–251
Proceedings of the National Academy of Science of the USA (PNAS) 3:281, 21:323
Process, personality, interests, and knowledge *see* PPIK (process, personality, interests, and knowledge) theory of intelligence
Process accidents 17:370
Process and mechanistic theories 3:270
Process character of inference 12:39–40
Process consultation, organizational development 6:298
Process dissociation framework 13:88
Process dissociation procedure, of recognition memory 20:39–40
Process durations estimation, task network models 16:543
Process evaluation 19:118
Process innovation 12:147, 12:152–157
  versus product innovation 12:163
Process models
  of culture 5:728
  implicit association test (IAT) 11:705–707
    figure-ground asymmetries 11:706–707
    Quad model 11:707
    random-walk processes 11:705–706
    stimulus–response compatibilities 11:705
    task-switching costs 11:706
Process models of faking 17:834
  mixed Rasch models (MRMs) 17:837
  social desirable responding (SDR) 17:834
  structural equation modeling 17:836
Process of legal mobilization 13:590–591, 13:594
Process pluralism 15:70
Process research, psychotherapy 19:515–520
Process safety 17:370
Process theories, entrepreneurship 7:682
Process tracing, social mechanisms 22:415
Process use 7:578
Process view of numbers 14:776–778, 14:777f–778f
Process-based standards 17:143
Processes 17:871–872
  death as 5:871
  of negotiation 16:433
    claiming value 16:433
    creating value 16:433
  prolonging of
    single 16:541
    two 16:541–542
  revolution as 20:646–647
Process-experiential psychotherapy 20:740–741
Processing, visual words 25:215–216
Processing cost, language development 13:313–317
Processing explanations of grammar 9:470
Processing fluency, heuristic 10:827–828
Processing frames 23:231–232
Processing models, of written language impairments 1:452–453
Processing Speed Index (PSI) 12:284, 16:723
Processing tree representations 16:543–544
  inclusion/exclusion paradigm 16:543–544
  tree inference 16:544
Processo Apostolico 15:578
Processo Ordinario 15:578
Process–outcome findings, psychotherapy research 19:518–519
Process-produced data 5:737
Processual approach, to lithics 14:262
Processual archaeology 1:901–903, 3:378, 14:262, 23:745
Processual development, forces counteracting 22:267

Processual norm activation model 8:27
Processualism 7:692
and hunter-gatherers 11:425–426
Procrastination 19:73–78
behavioral economics and temporal motivation theory 19:76
conscientiousness 19:75
decision-making 19:77
definition 19:73
demographic differences 19:74
impact in different life domains 19:77–78
impulsiveness 19:73, 19:75–77
measurement of 19:74–75
motivational aspects 19:76
observation 19:74–75
academic measures 19:74–75
other measures 19:75
personality 19:73–76
perspective of clinical psychology 19:76–77
affective and cognitive aspects 19:76–77
attention deficit hyperactivity disorder 19:77
behavioral aspects 19:77
fear of failure, test anxiety, and depression 19:76–77
perspective of differential psychology 19:75–76
Big Five 19:75
perfectionism 19:75
self, self-esteem, self-handicapping 19:75–76
time, time orientation, and time preferences 19:76
perspective of neurobiology 19:77
preference reversal 19:75–76, 19:76f
research perspectives 19:75–77
self-deception 19:75
self-efficacy 19:73, 19:75–77
self-regulation 19:76–78
self-report 19:74
general measures 19:74
specific measures 19:74
situational perspective 19:77
and strategic delay, differences between 19:73
treatments 19:78
types 19:73
value 19:74, 19:76
volitional aspects 19:76
Procreation 17:952
Procrustes analysis 19:79–81
factor analysis 19:79
generalizations 19:80–81
variants and later developments 19:80
Procter & Gamble 1:192–193
Procurement, supply chain management 23:709
Producer surplus 5:50–51
Producers
of area and international studies 1:943–944
of culture 5:731–732
Producers cooperatives 5:253
Producers of historical knowledge (Islamic tradition)
kuttab 11:62
madrasahs 11:62
ulama 11:62
Product architecture, and modularity 15:718–719, 15:721
Product choice, regulation 20:178
Product development 3:663
Product distribution 25:11
Product Emotion Measuring Instrument (PrEmo) 7:520
Product innovation 12:147, 12:152–158
versus process innovation 12:163
Product liability 3:30
Product liability doctrine 12:216–217
Product liability litigation 24:433
Product platforms 3:38
Product portfolio decisions, supply chain management 23:710
Product safety
class actions 3:736
consumer policy 4:733–737
Production 7:15–16, 9:112, 15:714–717
agriculture 1:479
archaeology of 12:329
celebrity, in twenty-first century 3:276
cultural comparisons 5:603
in cultural diamond 5:603–604
of culture *see* Cultural production
culture studies 5:603
de facto power 15:716
economic statistics 5:758–762
educational 7:151–153
establishment in sociology 5:603
factors 5:603–604
of fashion 8:838
functional alternatives 15:716
land use indicators 13:241–242
production-consumption interlinkages 13:242f
media imperialism 15:57
in music industry 5:603
organizational process of 5:537
productive forces 15:714–715
regulation 20:178
inefficiencies 20:178
relations of 15:715
and forces 15:715
and reproduction 20:450, 20:452, 20:454
in shaping culture 5:603
and social class 22:341–342
superstructures
and bases 15:716
sustainable 23:839–844
urban settlements 24:836–837
Production cultures, television 25:83
Production deviance 25:722–723
Production function approach, for economics of education 7:149–152
Production mode, Marxist geography 14:675–676
Production nodes, commodity chains 4:245
Production of culture *see* Cultural production
Production of space 23:201
Production processes, political communication 18:342
Production rule model 6:711
Production systems, industrial geography 11:855–857
Production techniques, international migration effects on 12:516–517
Productionist ethic 1:475
Productionweise 9:112
Productive activity, organizing 17:366
Productive base, building 17:365–366
Productive exchange 8:486
Productive giftedness
defining 7:275
factors influencing
environmental 7:276
extracurricular time 7:276
home environment 7:276
individual aptitude 7:276
mentoring 7:276
peers 7:276
quantity and quality of instruction 7:276
school 7:276
time usage 7:273
model 7:274–276, 7:274f
nurturing 7:277
by altering home curriculum 7:277
by altering school and out-of-school curriculum 7:277
national reforms 7:277
role of parents 7:277
teachers and mentors, role of 7:277
productive outcomes 7:276
sustaining 7:275
ten-factor model 7:275–276
Walberg's model 7:275
Productive outcomes 7:276
Productive power 18:763–764
Productive-force determinism 24:110
Productivism 2:750
Productivist welfare regime, educational policy in 25:488
Productivity 3:155–157, 3:159, 5:769, 15:597, 17:673–674
and aging 1:417–418
and beauty 2:385
and entrepreneurship 7:675, 7:678–679
fire and 9:197–199
in informal employment 12:62
international migration effects on 12:517
language acquisition 13:248
and literacy 13:260–261
of older workers 17:186
Products of nature, patenting of 17:585–588, 17:590
Produsers (Internet) 12:591–592
Pro-environmental behaviors 7:699–700, 19:82–88
corporate social responsibility 19:82
employee green behaviors
determinants of 19:84–86
improving 19:86–87
environmental performance 19:82–84, 19:86

Pro-environmental behaviors (*continued*)
human resources for organizational environmental sustainability 19:83
individual pro-environmental behaviors 19:83–84
jobs and 19:84
organizational pro-environmental initiatives 19:82–83
Preisendörfer's scale 7:700t
Proenvironmental behavior 23:828
ProFamy model 8:728
Profane, ritual 20:717
Profeminist sensibility among men in social work 14:682
Professional actors, in-house counsel as 12:127–129
Professional anthropologists 24:291
Professional associations 12:372
American Anthropological Association 8:40
American Psychological Association 8:39
American Sociological Association 8:39–40
comparison of professional ethics 8:40–41
educational research 7:282–283
and ethics 8:59
International Sociological Association 8:40
other academic and professional organizations 8:40
Professional autonomy
of in-house counsel 12:128
in managed care 14:464
social work in Russia 22:810–812
Professional boundary violations, psychiatric ethics 8:65
Professional caretakers, childcare by 3:397
Professional class 12:274
Professional classifications 3:799
Professional codes, for privacy/confidentiality in social research 18:942
Professional competition, of in-house counsel 12:127
Professional constructs 4:730
Professional development (PD) 5:328–329, 19:89–94, 24:58
continuing professional development 19:89–90
continuing professional education 19:89–90, 19:92–93, 19:89t
in lifelong learning 4:114
pedagogic differences and 4:114
reflective practice 19:90–91
and supervision 19:92
supervision 19:91–92
higher education 19:92–94
in organizational context 19:91–92
research and scholarship in social work practice 19:93–94
of teachers 21:137
time and focus of 19:89t
Professional development program
logic model for 14:300f
Professional discipline 15:83
Professional engagement and career development aspirations (PECDA) 24:66
Professional ethics 19:821
for human subjects protection in research 8:48
public policy 19:580
and regulations 13:632–633
in social work 22:782, 22:786
Professional ethnographers 24:291
Professional habitus of studiousness 19:95–96
Professional identity, and whiteness/white supremacy 25:572–573
Professional independence, of in-house counsel 12:127–128
Professional inquiry 4:799
Professional judgment, social work 3:475
Professional knowledge 4:797–798, 4:800
Professional learning 4:797
policy 4:797
Professional learning community (PLC) 21:102, 21:114
Professional meetings 22:661–662
Professional names 16:184
Professional norms, in hospitals 11:208–209
Professional organizations
feminist criminology and 8:916
geriatrics 10:105
professions in 19:103–104
Professional power interpretation 19:108
Professional psychology 24:521
critical thinking in 24:519
education for 24:517–518
epistemology in 24:519
implications for 24:522–523
scientific training in 24:518
social nature of 24:522
Professional recognition
applied demography 1:841–842
in citizen science 3:635
Professional regulations 13:632–633
Professional responsibility 20:584–585
Professional roles 13:628–629
Professional service firm (PSF) 19:103
Professional shamanism 21:868–869
Professional social movement organizations 7:765
Professional social work 6:449, 14:506
Maori social work 14:506
Professional social worker, definition 14:506
Professional socialization 15:82–83
Professional supervision *see* Supervision
Professional support, for therapist and psychiatrist
in combined pharmacotherapy and psychotherapy 19:497
Professional team sports, economics of 23:318–319
Professional training, institutionalization of 19:96–97
Professional–cluster trial 3:900
Professionalism 11:120, 12:860, 15:82–83, 15:676, 15:680, 21:102
journalism 12:852–854
military 15:512
and philosophy 18:32, 18:35
and professional growth 24:46
psychiatric ethics 8:69–70
and social work 22:134–136
Professionalization 3:612, 19:95–96, 19:98–100, 22:679, 24:118
clinical psychology, Europe 3:875
of communications 14:716–717
conceptions and history 19:96–99
death/dying 5:866
definition 22:679
definitions 19:95–96
of historical studies 11:74–75
of history 11:51, 11:120
breadwinning 11:120–121
system of norms 11:121–124
in-house counsel 12:126–128
and institutionalization 19:96–97
journalism 12:852–854
and military history 15:506
in nineteenth century historical writing 11:75
process of 22:679–680
and radical social work 19:870–872
research on 22:679
of scientific disciplines 21:288–289
of social work
demand for 22:793
since 1999 22:790–791
variation and change in 22:679–680
Professionalized history, characteristic methods for 1:910
building national identities 1:910
focus on elites 1:910
geographically uneven attention 1:910
national specialization 1:910
Professionalized selection 12:881–882
Professionals
automation for 2:297
collaborative research 7:864
educational research 7:282–283
and experts 8:590–592
social workers as 24:493–494
Professional's ignorance 11:579
Professionhood 19:107
Profession-ready teachers 24:77
Professions 13:598–601
autonomy 19:95–98
cognitive and emotional orientations and strategies 19:97–98
conceptions and history 19:96–99
definitions 19:95–96
educational research 7:281–282
engineering 7:643–644
and expertise 21:207
'general theory' of 19:99–100
heteronomous and autonomous, distinguishing between 19:96
historiography 19:99–100
institutionalization of 19:97
intellectuals 12:277–279
as issue for paralawyers 17:484
journalism 12:858–860
and market and economic situation 19:98–99
medical 7:189, 7:191, 15:92–97

in organizations 19:101–106
bureaucracy 19:101–102
culture 19:101–102, 19:104–105
in labor market 19:102–103
in professional organizations 19:103–104
particularism *versus* universalism 19:99
place of ethics in definition of 8:64
social science 22:679–687
social work 20:581, 20:583, 20:585, 22:807, 22:810
social work, internationalization of 12:557
sociology of 19:107–110
collective association 19:107
corporate discipline 19:107
current problems 19:109
expert occupations 19:107
historical development of the field 19:107–109
individual practice 19:107
specialized education and examination 19:107
Strauss on 23:527–529, 23:531
teaching *see* Teaching
white-collar workers 25:564–566
Professor, women 25:623–624, 25:626
Proficiency, language 16:58
Profile case 2:549
extensions of models to 2:551
Profile likelihoods 14:118, 14:121–122
example 14:120–121
Profile of Nonverbal Sensitivity 8:678
Profiles 21:59, 21:61–63
Profiling, and forensic psychology 9:353–354
Profit maximization, microeconomics 15:332–333
Profit motive, peasants 17:678
Profit pools, and power 15:668–670
Profitability 14:549–550, 14:553
Profiting From Innovation framework 6:586
Profit-maximization leagues 23:319
Profit-oriented capitalism 19:846–847
Profits 25:11
absolute levels of 19:114–115
capital 19:111–115
and capital 3:90–91
classical economics 19:111–112
entrepreneurship 19:114
investment 19:111–115
Kalecki on 19:115
Keynes on 19:113–115
marginalist economics 19:112–114
Marx on 19:111–112
and power 15:667–670
rate of 3:91
rates, shares, and levels 19:114–115
risk 19:111–112, 19:114
saving 19:113–115
Smith on 19:111–112
Sraffa on 19:114–115
Progeny of incestuous matings, direct investigation 11:727–728
health status 11:728t
Program Assessment Rating Tool (PART) 22:299
Program budgeting 2:902
Program development
resilience 20:555–560
concepts and definitions 20:556–557
questions and observations 20:555–556
Program evaluation 18:278–279, 19:117–122
definition 19:117–118
distinction from other activities 19:117–118
empowerment evaluation 19:121
formative evaluation 19:118
for green jobs 10:377–378
impact evaluation 19:118
key concepts and assignment criteria 19:118
levels 19:118–119
behavior 19:118
learning 19:118
reactions 19:118
results 19:118–119
metaevaluation 19:119
outcome evaluation 19:118
process evaluation 19:118
program effectiveness 19:117–118
and selection bias 21:416–417
stakeholders 19:117, 19:119–121
standards for quality and fairness 19:119
user benefits 19:119–121
goal explication for user orientation 19:121
user-oriented evaluations 19:120–121
utilization types 19:119–120, 19:120t
users and doers of evaluations 19:117–118
utilization-focused evaluation 19:120–121, 19:120f
Program Evaluation and Review Technique (PERT) network 16:540–541, 16:543
Program evaluation tools, for physical activity 18:95
Program factors, offender rehabilitation 20:197
control condition 20:197
individualization 20:197
intensity 20:197
program content 20:197
quality of program implementation 20:197
Program fidelity 18:888
versus program innovation 18:888
Program for Detached Workers of the YMCA Metropolitan Chicago 9:596
Program for Health Education 21:825
Program implementation and dissemination (Prevention science) 18:887–888
delivery system 18:888
support system 18:888
synthesis 18:887
translation system 18:887
Program innovation, *versus* program fidelity 18:888
Program libraries 7:2
Program synthesis, reasoning 5:939
Programa de Asignación Familiar (PRAF) 4:556–557
Programme for International Student Assessment (PISA) 1:46, 4:431–432, 5:342–348, 5:376–377, 5:379, 5:763–764, 5:766, 5:771, 5:847, 7:180, 7:273, 7:919–920, 8:752, 9:679, 9:682, 10:248–249, 12:428, 14:247, 21:103, 21:119, 21:232, 21:464, 22:325–326, 22:927–928, 23:760, 25:248, 25:484
cross-cultural study of education 5:327, 5:329
Programme for the International Assessment of Adult Competencies (PIAAC) 5:327, 5:771, 5:774, 5:847
Programmed instruction program, for language disorders 13:331
Programming, television 24:172
cultural 24:169–170
entertainment 24:170–171
informational 24:170
news 24:170
regular television programming 24:169
advertisers in 24:169
Programming languages, econometric software 7:2–3
Programs
cancer screening
examples 3:79
as policies 3:79
and cash transfers 3:220
fathering 8:844–847
gifted education 7:160–162
acceleration and enrichment 7:160
schoolwide enrichment model 7:161–162
talent search model 7:161
PROGRESA (conditional cash transfer) 3:408, 4:556–559, 8:542, 8:562–564, 18:751
Progress 7:722–723
American exceptionalism 19:127
as Americanization (1950–1970) 19:126
dethroned (1914–1950) 19:125–126
economic growth 19:123–124
eighteenth century origins 19:123–124
Enlightenment 19:123–126
evolution 19:124–125
Frankfurt Institute for Social Research 19:125–126
historicism 19:123, 19:127
history of 19:123
Kuhn on 19:126–127
modernity/modernization 19:123, 19:125–126
neoliberalism 19:125–127
nineteenth-century triumph of 19:124
postmodernism 19:127
and postmodernity (1970–2010) 19:126–127
problem of (1880–1914) 19:125
romanticism 19:125
science, progress of 19:123–126
social differentiation 6:392–394

Progress (*continued*)
social science disciplines, ideology of 19:123, 19:125–126
theories of 19:123–124
Progress feedback, psychotherapy 19:519
Progress in International Reading Literacy Study (PIRLS) 4:434, 5:765–767, 5:773, 5:847
Progress monitoring 24:49
Progressive core values, of trade unions 24:494–495
Progressive education 20:770
Soviet 20:826
Progressive Era, pragmatism in 18:808
Progressive ideology 5:255
Progressive income tax 23:340
Progressive movement 13:773, 25:814
Progressive multifocal leukoencephalopathy (PML) 1:489
Progressive politics
and radical social work 19:870–871
and social work 22:521–522
Progressive reformers 13:773
Progressive sampling 23:189f
Progressive societies 13:565–566
Progressive syndromes of dementia 13:361
Progressives (grammar) 24:212
Progressivism 8:225, 18:299–302, 18:305, 20:397
Progressivity, of income tax 11:744, 11:747–748
Prohibition, arguments for/against 25:77–78
Prohibition (US), Fisher's views 9:220
Prohibition era, drag cultures during 13:875–876
Prohibition of Disguises Act 16 1969 (South Africa) 21:744
Prohibition on alcohol 6:670–671
Proinflammatory cytokine production 22:486
Proinflammatory markers, and emotions 7:499
Proinflammatory nuclear factor kB/Rel transcription factors 22:485–486
Project appraisal 17:759–761
ad hoc dimensions 17:759–761
cross-impact and hierarchical analysis modules 17:761–762, 17:761f
extrapolation of research findings 17:761
Joint Cross-Impact Matrix 17:762
matrix (affective dimensions) 17:760f
matrix (cognitive dimensions) 17:760f
standard dimensions 17:759
Project-based, external funding in academe 1:50–51
Project Camelot 1:952–953, 4:123, 8:58
Project categories 17:758
Project Courage 12:682–683
Project Cybersyn 5:647
Project DARE (Drug Abuse Resistance Education) 10:671
Project Dragonfly 6:897
Project Elicitation 17:758–759
Project funding, infrastructure 22:670
Project GLOBE 5:339
Project Head Start 2:877, 6:474
Project management/monitoring, logic models 14:299
Project Safe Neighborhoods (DOJ) 10:460
Project STAR 3:752–753
Project Troy 1:952
Project workers 18:665
Projection rules (generative grammar) 9:873
Projections
in cartography 3:169–170, 3:169f
linguistic presupposition 14:191
small area 6:162–169
Projective displays, multivariate 10:355
Projective drawings 19:132–133
Projective identification 13:68, 24:547
Projective measures 16:725
Projective methods
comprehensive systems 19:131
historical overview 19:129
projective drawings 19:132–133
psychometric considerations 19:129–130
reliability 19:130
Rorschach test 19:130–131
sentence completion tests 19:133
thematic apperception test 19:131–132
validity 19:130
Projective Movement 19:129
Projective spatial relators 6:54–55
Projective techniques
in ethnography 8:188
historical background 3:880
Projectivity 2:256
Project work
academic research and employment 1:51
Prokofiev's 'Classical' Symphony 20:662
Prokuratura 22:846–847
Prolactin levels
schizophrenia 21:24
Proletarian 6:369
Proletarian dictatorship 6:372–375
Proletarianization 5:109, 11:233, 25:566
of professionals 19:109
Proletariat 3:94–95, 4:339, 6:372–373, 14:642, 14:645, 14:647
Proletkult (cultural revolution) 20:859
Proliferation 24:143–144
of classes 3:747–749
health care markets 10:599
of survey modes 12:672
Prolific offender, sex offender as 21:673–674
PROLOG clauses 5:937
Prolonged grief disorder *see* Complicated grief disorder (CGD)
Prolonged training, for metacognition 24:93
Prometheus Image Archive 2:24
Prometheus method 17:589
Promise 12:637, 12:639
Promoting Aphasic's Communicative Effectiveness (PACE) 13:331–332
Promoting emotional and social competence 4:667
Promotion of health 10:674–675
Promotion schemes, and unintended consequences 17:421
Promotions 19:135–139
asymmetric learning and job assignment 19:137–138
corporate hierarchies 19:135
correlates and determinants 19:136
employment 10:518
fast track 19:136
human capital and symmetric learning 19:138–139
internal labor markets 19:135
job satisfaction 19:136
outcomes 19:135–136
signaling 19:138
tournament theory of 19:136–137
wages 19:135–139
Promotive interaction 4:857–859
Promotore Fidei 15:579
Prone position, infant motor development 15:973
Pronger, B., on gay athletes 9:658–659
Pronominal argument 14:221
Pronominal incorporation 14:221
Pronominal subjects 11:827
Pronouncement stories 3:546
Pronouns 11:826, 16:183
reflexive 13:315–316
Proof search, reasoning 5:935–936
Proof structures, for discrimination 6:531
Proof theory 13:100
Proofs 14:266, 16:924–928, 17:531
Proopiomelanocortin (POMC) 16:651
Propaganda
Chomsky on 3:529
cyberterrorism 24:227–235
and mass media 14:712, 14:716
media 15:13–14, 15:16
science of 25:395–396
Propaganda model 25:397–398
Propaganda of the deed 24:238
Propagandist, author as 6:622
Propagation 22:702–703
of immorality 21:178
Propensity 4:836–838
Propensity interpretation
objective probability 19:41
of probability 19:26–27
Propensity score 4:841, 8:346–347, 12:454, 17:108
adjustments 19:143–144
covariate adjustment 19:144
matching 19:143–144, 19:143t
stratification 19:144
weighting 19:144
assessment 19:143
causal inference 3:257–259
computer programs for computation of 19:145
computation 19:141–143
assignment prediction for multi-valued treatments 19:142–143
covariates selection 19:141–142
procedures 19:142
covariate balance assessment 19:144–145

description 19:141
implementation 19:141
limitations of PSAs 19:145
matching 19:143–144
methods for reducing selection bias 19:140–141
quasi-experiments 19:140
selection bias 19:140
Propensity score adjustments (PSA) 19:143–144
computer programs for 19:145
limitations 19:145
Propensity score matching (PSM) 2:849, 3:248
Propensity scoring 19:140–146
Propensity-driven behavior 25:128
Proper domains of function 9:522
Proper multiple imputation 15:605
Proper names 16:184–186
in clinical practice 19:151
impairments of 19:147–152
and numbers 14:775–776, 14:776f
Propertization 19:155
intellectual property law 12:267–268
Property 19:167
anthropology of 19:153–159
capitalism 19:153–158
citizenship 19:157–158
collectivization 19:155–157
common property 19:154–155
cultural property 19:155
ethnography and diversification 19:154–155
evolutionist approaches 19:153–154
Hungary 19:156–157
intellectual property 19:153, 19:155, 19:158
land tenure 19:153–157
privatization 19:156–158
real property, post-socialist transformations of 19:155–158
socialism, post-socialist transformations of real property 19:155–158
emergence of private rights of 19:169
evolution of jurisprudential and analytical conceptions of 19:169–170
and exchange 8:477, 8:480
and knowledge society 13:107
new frontiers of 19:171
ownership 19:177–183
philosophical justifications for and critiques of private rights of 19:167–169
civic republicanism 19:168
distributive justice 19:168
ecological theories 19:169
labour theory 19:167
libertarianism 19:167–168
Marxist theory 19:168–169
personhood 19:168
pluralist and hybrid theories of property rights 19:169
utilitarianism 19:168
right to *see* Right to property
and science 21:251
security *see* Security; commercial transactions
in socialism 22:849–851
as triadic relation 19:170–171
use of term 19:190
Property crimes 5:190, 19:160–166
burglary victimization, prevalence of 19:160–161, 19:161f
car theft and joyriding 19:162, 19:162f
consumer fraud 19:161–162
determinants 19:163–164
affluence 19:164
demographics 19:163–164
urbanization 19:163
drops in crime 19:164–166
heavy crime burden of business sector 19:163
opportunities of crime 19:161–162, 19:164–166
personal thefts 19:160–161, 19:161f
responsive securitization 19:165
routine activity theory 19:163
strain theory 19:164
trends 19:164–166, 19:164f
vehicle theft 19:162–163
Property deviance 25:722–723
Property dualism 18:112
Property insurance 12:217
Property law 13:449–450, 19:178
colonizing 15:361–362
Property owners associations (POAs) 9:612, 9:615
Property ownership 12:57, 19:177–183
banking 19:180
liquidity crises 19:181–182
nomina 19:179–181
premium and interest 19:181–182
property claims
as book accounts, debt instruments and money 19:179–181
and enforcement 19:177–178
quantification and settlement 19:178–179
settlement assets 19:177–179, 19:181–182
technical progress 19:182
Property prices 9:616–617
Property regime change, in Central Asia 3:317–318
Property regimes 16:840–841
Property rights 4:274–275, 4:707, 5:253, 13:759, 19:184–195, 24:528–532
antecedents 24:528
creation of 19:185
Dutch law 19:187
economics of 19:189
franchising 24:529
historical development 19:193–195
hypothecation 19:190, 19:192–194
institutional economics 24:528
and interests 12:391–392
leasing 24:529
and libertarianism 14:40–43
limitations 24:530–531
modern corporation 24:531
overreaching 24:531
oversimplification 24:530–531
Russian privatization 24:531
and marriage payments 6:643–644
brideprice 6:643–644
dowry 6:644
modern 24:528–530
common pool problems 24:529
de facto property rights 24:530
economic development and reform 24:530
externalities 24:528
rent seeking 24:529–530
theory of firm 24:529
variable access 24:529
nature of 19:184–185
ownership 19:192
politics of 19:187–189
possession 19:191–192
property 19:190–191
and resource management 6:857–858
resources 20:569–570
restrictions on the exercise of 19:187
security 19:192–193
social function of 19:185
tenant farming 24:529
for urban studies and planning 19:184
varieties of 19:185–187
Property rights movement 24:528
Property rules 19:167, 19:169–170
Property tax 23:339–342, 23:341t
Property tax abatements 24:796
Property types 20:13–14
Property-concept words (PCWs) 23:885–886
Prophecy, responses to failure of 15:530–531
Prophesy, cargo prophesies 3:141–142
Prophethood 11:63–64
Prophetic currents 19:197
Prophetic tradition (*hadith*) 11:61
Prophetism 16:169, 19:196–199
Biblical prophecy 19:197–198
emissary 19:196–197
ethical 19:196–197
exemplary 19:196–197
Lang's view of functionary *versus* autonomous charismatic 19:198
millennial movements 19:198
political element in 19:198
prophetic currents 19:197
prophetic revelations 19:197
religious movements 19:196
use of term 19:196
Weber's characterization 19:196–197
Prophets 11:63–64
in ancient Israelite society 19:197
and charisma 3:356
charisma and recognition of a personal calling 19:196
classical Biblical prophets 19:197
early 19:197
as founders of religious movements 19:196
in millennial movements 19:198–199
proof of gifts 19:196
and Qur'an 11:62–63

Prophets (*continued*)
role of 19:196
use of term 19:196
Weberian perspective 19:196
Prophylaxis, in psychiatry 19:200–204
adherence and importance 19:203
bipolar illness 19:202
future directions 19:204
major depression 19:203
mental illness 19:202–203
physical health prophylaxis 19:203
schizophrenia 19:200–202
Propinquity, and interpersonal attraction 12:628
Propoor tourism 24:474
Proportional electoral system 4:693, 7:351
Proportional hazards regression model 23:773
Proportional political representation 18:442–443
Proportional preference data, in forced-choice personality assessment 17:841
Proportional representation (PR) 7:350
closed lists 19:205, 19:207–209
coalition governments 19:207–209
constituencies 19:205–207
design elements 19:206–208
balloting choices 19:207–208
district magnitude 19:206
electoral formulas 19:206–207
electoral thresholds 19:207
tiers of representation 19:207
effects and evaluations 19:208–209
electoral systems, types of 19:205–206
list proportional representation 19:205
mixed-member proportional 19:205–206
single-transferable vote 19:205
fair majority voting 19:207
historical development 19:208
legislative apportionment 19:206–207
majority rule 19:209
and majority rule 14:447
minority representation 19:206, 19:208
open lists 19:207, 19:209
political parties 19:205, 19:208
semiproportional electoral systems 19:206
limited and cumulative voting 19:206
parallel systems 19:206
tiers of 19:207
Proportional representation (PR) systems, green parties in 10:382
Proportionality 4:869, 7:350, 8:265
Bayes' theorem 2:379
Proportionate universalism, social exclusion/deprivation 22:289
Propositional knowledge 13:93
Propositional learning 11:717–718
Propositional logic
language of 5:933
reasoning 5:933–935, 5:937–938
Propositional theory, and mental imagery 15:220–221, 15:225–226
Propositionality 2:493
Propositions, in police use of force 18:258
Propp, V. 1:754–755
Propranolol 3:769
Proprietary knowledge production 21:151
Proprietor content (online information) 4:508
Propriety Standards 19:119
Proprioception
neonatal imitation 16:487
vestibular system 25:63, 25:67–68
ProQuest Congressional 5:808
Prose recall 20:52–53
Prosecution 19:210
centralized and decentralized office and 19:211–212
and crime policy 5:176–178, 5:180–181
discretion at charging 19:213–214
for domestic violence 6:626–627
elements of private 19:211
independence and accountability to public and 19:212
new issues in twenty-first century 19:214
prosecutors as public officials and 19:210–211
and relationship to and with police 19:212–213
roles in court and 19:214
university hazing 10:559
Prosecution discretion 18:230
Prosecutorial behavior 13:573
Prosecutorial misconduct 25:788
Prosecutors 10:455
Proselytization 15:614–617, 20:378
in theological discourse 15:616–617
Proselytizing history 14:217
Proselytizing movements 16:805
Prosimians
primates, evolution of 18:904–908, 18:907f
Prosocial activities engagement 1:31
Prosocial behaviors 1:579–583, 7:550, 7:552, 19:216–220, 23:701–702
during adolescence 19:221–227
antecedents to 19:222–225
cultural orientations 19:223–224
methodological issues and future directions for research 19:224–225
origins and definitions 19:221–222
socialization 19:222–223
tendencies 19:221–223
in adulthood 19:216
cooperation 1:580–582
cross-population differences in 19:238–243
antisocial behavior 19:241
collectivism 19:239
culture 19:238–240, 19:242
explanations 19:239–241
globalization 19:240
helping 19:238–239, 19:241
institutional quality hypothesis 19:240
market integration 19:240
operationalizing 19:238
pathogen stress 19:241
population density and size 19:239
prosocial tradeoffs and social norms 19:241–242
religion 19:240
scale 19:238–239
sharing 19:238–239, 19:241–242
war experience 19:241
defining 19:244–245
definition 1:579
effects of parenting on *see under* Parenting
and empathy 7:549–550, 7:552
evolutionary theory 1:580
future directions 1:581–582
goal framing 1:581
helping 19:217
Hoffman's theory 19:244–245
human nature 1:580
individual difference 1:581
during infancy and early childhood 19:233–237
altruism 19:235
comforting 19:233–235
cooperation 19:233, 19:235–236
cultural variation in prosocial responding 19:236
empathy 19:233–236
feeling of others 19:233–236
future directions 19:236
helping 19:233–235
in infancy 19:233
ministering to others 19:233–234
multidimensionality 19:235
sharing 19:233–235
social development 19:233–234, 19:236
working and playing with others 19:234–235
influence of culture on 19:247
institutions 1:581–582
and intergroup relations 19:218–219
power relations 19:218–219
and media 15:6
and moral development 15:790–791, 15:794
philanthropy 1:581
punishment 1:580–581
rational choice 1:580
research trends 19:216–217
self-determination theory 19:245–246
and SES 22:944–946
social dilemmas 1:581
social networks 1:581
social preferences 1:580, 1:582
solidarity 1:579–580
theories 1:580–581
types 1:579–580
volunteerism 25:268–270, 25:272
Prosocial behaviors, during childhood 19:228–232
aggression 19:229
altruism 19:228–230
bullying 19:228–229
cooperation 19:229–230
cultural factors in prosocial development 19:230–231
caring for others 19:231
framing cultural issues 19:230

modesty and politeness 19:231
prosocialty at home 19:230–231
race and ethnicity 19:231
sharing 19:231
defining 19:228
future directions 19:231
helping 19:228–231
lying 19:228–229, 19:231
moral development 19:229
prosocial development 19:228–229
research and practice 19:229
theories of prosocial development 19:229–230
biology 19:229–230
cognitive development 19:229
empathy 19:229
social learning 19:230
Prosocial media, effects on children 15:39
Prosocial modeling, notion of 5:211
Prosocial motivation, and empathy 19:217–218
Prosocial preferences, modeling 2:447
Prosocial responding, Eisenberg's theory 7:315–320
approach to moral reasoning 7:315–317, 7:316t
approval-oriented moral reasoning 7:315–318, 7:316t
empathy 7:317–318
hedonistic-oriented moral reasoning 7:317–319, 7:316t
internalized-oriented moral reasoning 7:315–319, 7:316t
interpersonal-oriented moral reasoning 7:316t
needs-oriented moral reasoning 7:315–319, 7:316t
personal distress 7:317–319
perspective talking 7:318–319
self-focused-oriented moral reasoning *see* Hedonistic-oriented moral reasoning
self-reflective empathic orientation of moral reasoning 7:316t
situational factors 7:315–317, 7:319
stereotyped orientation of moral reasoning 7:317–318, 7:316t
sympathy 7:317–319
Prosocial tendencies 19:221–223
Prosocial tradeoffs 19:241–242
Prosociality 17:803–805
and adolescent friendship 9:433
gene-culture coevolution 9:817
Prosodic integration 23:230
Prosodic structure hypothesis 23:896–899
domain convergence 23:896
domain layering 23:896–897
interaction of prosodic markedness constraints and interface constraints 23:897–898
nonsyntactic determinants of phonological and phonetic domain structure 23:897–898
speech rate 23:898–899
Prosodic structure markedness constraints 23:895–896
Prosody 10:113–115, 15:827–829
child-directed speech 3:400
of language 23:714
Prosody effects 21:624–625
Prosopagnosia 1:441–442, 4:91, 8:663–666, 16:632, 19:250–255
acquired and developmental 19:250–251
amnestic 19:251–254
apperceptive 19:251–254
associative 19:251–254
configuration-based processing 19:252
congenital 19:250
face and function specificity of 19:251–252
functional imaging findings 19:252–253, 19:252f–253f
inversion effect 19:252–253
multimodal 19:252–254
and neurodegeneration 19:253–254
Prosopamnesia *see* Amnestic prosopagnosia
Prosopanomia 19:148
Prosopography 19:256–260
antiquity 19:256–257
Burke on 2:926
comparative 19:258–259
comparative history 19:258–259
cultural history 19:258
historiographic references 19:256–257
historiography 11:64
Marxism 19:256–257
merits and limitations 19:257–258
microhistory 19:256–257
population 19:256–258
statistics 19:256–257
PROSPECT study 4:450, 4:450t
Prospect theory (PT) 2:445–446, 2:775–776, 16:642–643, 17:782, 19:261–269
assessment 19:265–266
friendly criticism and normative–descriptive shuffle 19:265
intuitively appealing examples 19:266
role of new type of experiments 19:265–266
biases 19:263, 19:266
citations of text search 19:267f, 19:267t
of decision making 5:918–919, 5:919f
economics 19:263–265, 19:267
Google Ngram 19:266f
heuristics 19:263–264, 19:266
history and context 19:261–262
hypothetical value function 19:264f
hypothetical weighting function 19:264f
influence in behavioral and social sciences 19:266–268
Kahneman on 19:262–263
Morgenstern on 19:262–265
Neumann on 19:262–265
psychology 19:262–265, 19:268
rational choice theory 19:261, 19:264
relationship between monetary value and moral value 19:261f
risk averse behavior 19:262–263
risk seeking behavior 19:262–264
and sanctions 20:890
Savage on 19:262–264
Tversky on 19:262–263, 24:680–681
uncertainty 24:718–719
Prospection, and future thinking 9:556–558
Prospective control 19:68–69, 19:71
Prospective economic voting 7:27
Prospective memory 1:882, 19:270–278
anterior prefrontal cortex 19:274
avoiding interruptions 19:277
distinguishing characteristics of 19:271
empirical study of 19:270–271
encoding 19:271
impaired 19:275–277
implementation intention encoding 19:277
improving 19:277
laboratory approaches 19:270–271
medial temporal lobe 19:274–275
multiprocess theory 19:273
neural correlates of 19:273–275
neurodegenerative diseases 19:275–276
normal aging 19:276–277
preparatory attentional and memory processes theory 19:272–273
reminders and smartphone technology 19:277
retrieval 19:272–273
storage 19:271–272
substance abuse 19:275
tasks, categories of 19:271
Prospective memory functioning 19:276
Prospective observational cohorts 24:282
Prospective prediction problem, and desistance/persistence 5:143
Prospective surveys 5:737
Prospectivity, and perceptual learning 10:132
Prosperity property 4:871
Prostate cancer 15:170
among males 3:79
Prostate-specific antigen (PSA) test 3:78–79, 17:925
Prosthesis, vestibular disorders 25:74
Prostituted science 1:734
Prostitution 5:719, 18:533–535, 19:279–285, 21:792
carceral feminism 19:281
commercial sex 19:279–281
and sex trafficking 19:280–283
sexual activity, commodification of 19:279
and victimless crime 25:75–79
Protagonists 18:202
Protean career 3:130
Protect IP Act (PIPA) 13:524
Protected areas 7:715–716
Protection 11:763–764, 14:429–431
child 11:765, 11:767
of children, in disasters 22:756
cultural mediators's role 5:442
of individual 4:707
mafias 14:422, 14:424–425
private 14:431
of vulnerable people, in disasters 22:756
Protection in evaluation, EAHCA principle 23:222t
Protection model, consumer policy 4:734–735

Protection motivation theory (PMT) 10:577–578
Protection rights, of children 3:487–488
Protectionism 12:570
  in Central America 1:623
  trade policy 24:486–488
Protectionist regime, educational policy in 25:488
Protective approach to childhood cancer 3:461
  side effects of 3:461
Protective belt 8:473
Protective Dimension 13:421–422
Protective factors
  adolescent health 1:121–126
  bereavement 2:533–534
  resilience 20:555–559
Protective order, for domestic violence 6:628–629
Protective psychosocial factors, coronary heart disease 4:919–920
Protective reflexes, cardiac defense 23:540–541
Protective regulation 20:161
Protectorates
  and imperialism 11:682–683
Protein kinase C (PKC) 16:753
Protein synthesis
  and Ca^{2+} neuronal activity 12:712
  in learning, memory, and forgetting 23:873–880
Protest movements
  and globalizing identities 12:92
  spirit possession 23:283
Protestant asceticism 2:62
Protestant awakening 15:614–615
Protestant Buddhism 2:891
Protestant Christianity 16:169
Protestant civic piety 3:680–681
Protestant Ethic 5:75, 5:77–78, 10:523, 20:313, 20:345, 25:463
  Weber on 25:463
The Protestant Ethics and The Spirit of Capitalism 14:414–415
Protestant foreign missionaries 15:615
Protestant fundamentalism 24:603
Protestant missionaries 24:602–603
Protestant missionary society for women 15:616
Protestant nationalization of territorial culture 20:745
Protestant Reformation 3:540, 20:314–316, 21:383–385
Protestant religions 21:173–174
Protestant schism, sainthood before 15:577–578
Protestant schools 20:302
Protestant theology, in era of Reformation and confessionalization 20:100–101
Protestant transnationalism 24:604
Protestant Union 20:102–103
Protestantism 2:891, 3:538, 19:49, 20:280, 21:385
  in Canada 3:542–543
  and Christianity 8:286
  and economic development 5:528, 5:531, 5:533
  in Japan 12:796
  in Korea 3:532
  Protestant trajectory 3:541
  in United States 3:541
Protestantization 3:541
Protesters 22:507, 22:509–510
Protests 12:229, 16:8, 16:10–11, 16:799–800
  of Australian aboriginals 2:234–235, 2:237
  and counterculture 5:80–81, 5:84–85
  environmental, in Soviet Union 20:833
  and urban policies 24:903
Protoaddressees 15:874–875
Proto-Algonquian language 2:708
Proto-arithmetic 17:57–58
  to systems of numerical symbols 17:58
Protobureaucratic system 4:224
Proto-Canaanite alphabet 25:764
Protocerebral calycal tract (PCT) 15:157
Protocol sentences 14:317
Protocols, and knowledge translation 21:183–184
Protoconversation 15:874–875
Proto-fantasy 19:333
Proto-Germanic language 4:423
Proto-Humanism 20:428–429
Proto-Indo-European languages 17:51–52
Proto-industrialization 5:108, 11:874
Protolanguages 11:824, 13:337–338
  as a 'be' language 11:826
  clitics 11:827
  definition 4:424
  pronominal subjects 11:827
  reconstructed 4:425–426
  structure of 11:824
  subject–object alignment 11:826
  subordinate system 11:827
  word order 11:826–827
Proton induced gamma ray emission 24:481
Proton induced X-ray emission (PIXE) analysis 1:906, 24:481
Protonationalism 2:892
Proto-panpsychism 18:106–107
Protoracism 19:796–797
Proto-Renaissance 20:428–429
Proto-Romance language 4:423, 4:426
Proto-Sinaitic alphabet 25:764, 25:764f
Prototype causal-ordering model 21:463f
Prototype heuristics 5:920
Prototype model 4:531–532
  of social categorization 22:187
Prototype theories
  of cognitive semantics 13:933
  concept learning 4:538
Prototype theory in linguistics 19:286–289
  background knowledge 19:287
  categories 19:288
    degree of membership 19:287
    representativity 19:287
    structure 19:286–287
  categorization 19:286–287
  features 19:287–288
  polysemy 19:288
Prototypes 4:19
  of bilingual education 2:581–582
  in functionalist grammar 9:473–475
  leadership 13:652–653
  love 14:380
Prototypic emotional expressions 8:667
Prototypic nonequivalent group designs 16:851–853
  analysis of covariance 16:852–853
  change-score analysis 16:852
  matching/blocking 16:852
  selection modeling 16:853
Prototypical case 3:203
Proudhon, Pierre-Joseph 1:680–681, 22:837
Proust, Marcel 21:742–743
  on autobiographical memory 2:282–286
Proved producing reserves 17:166
Proved undeveloped reserves 17:166
Provenance, of museums 16:142
Provenzano, Bernardo 14:422
Providencia 1:640
Provident Knowledge Society 13:105
Providing an Infrastructure for Research on Electoral Democracy in the European Union (PIREDEU) 5:826
Province, and metropolis 3:308
Provincialism 3:662
Provision systems, commodity chains 4:244
Provisional government 20:856–859
Provisional orientation 17:202
Provisions
  cancer screening
    determinants 3:79–80
    implications 3:80–81
  Dahrendorf on 5:663–665
  friendship 9:450–451
  health policy 10:664–665
Proviso, in libertarianism 14:41–42
Provocation 5:219, 5:224–225
  and aggression 1:378
  neonatal imitation 16:486
Proxemics 4:287
Proxies 6:138–139
Proximal area, and market geographies 14:545
Proximal development zone 4:54
Proximal influences on gender development 9:742–746
  biological influences 9:742–743
  cultural factors 9:743–746
Proximal processes, for studying socialization and education 22:864
Proximal similarity 9:841
Proximate mechanisms governing fertility and parental investment 9:31–33
Proximate–ultimate dichotomy 1:88
Proximity, Gestalt heuristic of 10:830f
Proximity data 16:35–36
Proximity errors, in spelling 23:270
Proximity/distance matrices 19:80
Proxscal 16:46
Proxy, social comparison 22:211–212, 22:214

Proxy variables 11:750
Prozac 18:8
PRR *see* Parietal reach region (PRR)
PRS *see* Perceptual representation system (PRS)
Prudence 2:350–351, 8:212–213
Prudential regulation 9:191, 9:195
Prudential supervision 9:191
Pruitt–Igoe Housing Project 18:877–878
Prüm Decision 9:340
Prüm regime 9:341–342
Prussia
  Franco-Prussian militarism 25:365
  Humboldt on 25:279–280, 25:284
Prussian Secret State Archives 5:701
Prussian state conservatism 25:290–292
Prussian state economy 10:873
Prussian War Ministry, on homosexuality in Germany 9:645
Prusso-German militarism 15:492–493
PRWORA *see* Personal Responsibility and Work Opportunity Reconciliation Act 1996 (PRWORA)
PSA *see* Prescribing Safety Assessment (PSA); Propensity score adjustments (PSA); Prostate-specific antigen (PSA) test
P(p) scale 15:567
PSE *see* Picture Story Exercise (PSE); Point of subjective equality (PSE)
Psellotes 16:680–681
Pseudodiagnosticity 19:19–21
Pseudo-events
  in media 15:49
  and tourism 24:466
Pseudo-expressions 23:903
Pseudo-hallucinations 10:499
Pseudohomosexuality 9:645–646
Pseudologistic model (PLM) 24:309
Pseudo-maximum likelihood method, in outcome-based sampling 17:105
Pseudomonas aeruginosa bacteria 13:29–30
Pseudo-neurotic schizophrenia 17:892–893
Pseudo-panels, for political data 18:519
Pseudoscience 21:208
  magic as 14:435–436
Pseudoscientific discourse 20:565
Pseudosmia 17:199t
Pseudoword-related network 16:593
Pseudo-words, spelling of 23:269, 23:271–272
PSF *see* Professional service firm (PSF)
PSG *see* Polysomnographic recordings (PSG)
PSH *see* Permanent supportive housing (PSH)
PSI *see* Processing Speed Index (PSI)
Psicologia 13:902
PSID *see* Panel Study of Income Dynamics (PSID)
Psilocybin 10:502, 16:170
PSM *see* Propensity score matching (PSM)
PSNI *see* Police Service of Northern Ireland (PSNI)
PSOs *see* Primordial sociobehavioral outcomes (PSOs); Public sector organizations (PSOs)
PSP *see* Parameter space partitioning (PSP)
PSR *see* Pressure-state-response (PSR)
P-STAT (statistical analysis package) 23:378
PSUs *see* Primary sampling units (PSUs)
Psychasthenia 12:791
Psychedelia, and counterculture 5:83
Psychedelic rock, and counterculture 5:82–83
Psychedelics 16:170, 21:610
Psychiatric anthropology 19:363
Psychiatric aspects of sleep disorders 22:79–86
Psychiatric assessment, of negative symptoms 19:290–294
Psychiatric care
  economics of 19:295–300
    child and adolescent mental health 19:299
    computerized cognitive behavioral therapy 19:297–298
    conduct disorder 19:299
    costs 19:296–297
    dementia 19:296–298
    depression 19:296–297
    effectiveness 19:297
    evaluation 19:295–296
    family carers 19:298
    issues and challenges 19:299
    making trade-offs 19:297
    modeling 19:298–299
    monetary benefits 19:298
    utility measurement 19:298
Psychiatric classification of diseases, reasons for dissatisfaction with 17:29
Psychiatric community *see* Therapeutic community
Psychiatric comorbidity 17:593
Psychiatric complications 1:510
Psychiatric deluge 22:802
Psychiatric discourse 14:204
Psychiatric disease 18:792
Psychiatric disorders 3:572–573, 19:457–458
  attention deficit/hyperactivity disorder 3:573
  culture-specific 5:622
  drug use and dependency 3:572–573
  eating disorders 3:572
  mood disorders 3:572
  negative symptoms in 19:291
  and sleep 22:82
  social intervention in 22:380–385
Psychiatric drugs 19:458
  anticraving drugs 19:458
  antidepressants 19:458
  cognition enhancers 19:458
  hypnotics 19:458
  mood stabilizers 19:458
  neuroleptics 19:458
  tranquilizers 19:458
Psychiatric education 7:185–193
  assessment 7:189–190
  clinical skills assessment 7:190–191, 7:191f
    long case 7:190
    modified long case 7:190
    Objective Structured Clinical Examination 7:190–191
    workplace-based assessments 7:191
  curriculum development 7:185–187
    general post-qualification training 7:185
    length of training in Europe 7:186f
    specialty training 7:185–187
    table showing variations 7:185–187
    undergraduate 7:185
  life-long learning 7:191–192
  patient involvement 7:187
  profession 7:189, 7:191
  recruitment 7:188–189
  selection 7:187–188
Psychiatric ethics 8:63–70
  bioethics 8:64
  contemporary codes of ethics (AMA and APA) 8:65–66
    boundary between ethics and law 8:65
    confidentiality 8:65–66
    double agent problem 8:66
    false memory controversy 8:66
    fees, billing, and reimbursement 8:66
    informed consent 8:66
    sexual misconduct and boundary violations 8:65
  ethics and health care financing 8:66–67
  Hippocratic oath and tradition 8:64–65
    contemporary relevance of 8:64
    fiduciary principle 8:65
    honor in medicine 8:64–65
    for psychiatrists 8:65
  history of medical ethics 8:63–64
    ethics of advertising 8:64
  novel ethical challenges in twenty-first century 8:67–70
    electronic health records 8:68
    safeguard measures 8:69–70
    social networking/media 8:68–69
    telepsychiatry 8:67–68
  place of ethics in definition of profession 8:64
  professional definition 8:64
  professionalism 8:69–70
  types of theories 8:63
  of virtue and character 8:63
Psychiatric expertise 21:159
Psychiatric genetics 2:452
Psychiatric hospitals
  alternatives to 6:49–50
  decline and fall of 6:48–49
  development 3:874
Psychiatric illness
  family therapy for 8:745
Psychiatric illnesses 7:610, 8:745–748
  cooccurring bipolar disorder 2:684–685
  effects on serial order memory 21:641
  with hallucinations 10:500–502
  military/disaster psychiatry 15:497
Psychiatric institutions 23:528
Psychiatric nosology 13:129–130
Psychiatric practice, contemporary, religion in 19:309–310
  improving practice 19:310–311

Psychiatric social work, in United Kingdom 19:301–306
Psychiatric-child protection interface 19:304–305
Psychiatrists 6:48
  beliefs about religion among 19:308–309
  in combined pharmacotherapy and psychotherapy
    misconceptions of 19:497
    professional/emotional support 19:497
    utilization of special talents 19:496
  Hippocratic oath for 8:65
Psychiatry 2:752, 6:75, 7:606, 18:31–36, 19:313, 19:586
  anthropological aspects 19:313–319
  aspects of behavior therapy 2:426–430
  child and adolescent psychiatry 3:383–390
  classification systems 3:888–889
  cultural anthropology 19:314
  cultural *see* Cultural psychiatry
  definition of 19:307
  deinstitutionalization 6:42–46
  of deinstitutionalization 6:47–51
  differential diagnosis in 6:384–390
  disease model in 17:2–4
    Boorse's model 17:3
    Sedgwick's model 17:3–4
    Szasz's model 17:2–4
    Wakefield's model 17:4
  eating disorders 6:823
  and electroencephalography 7:362
  and endocrinology 7:606–611
  evidence-based 8:339–343
  forensic *see* Forensic psychiatry
  geriatric 10:104–108
  hermeneutical 19:313
  historical background 3:877
  of hypnosis 11:472–475
  importance of personality disorder in 17:897
  informed consent 19:320–325
  liaison and consultation 14:1–7
  magnetic resonance and molecular imaging in 14:439–445
  military and disaster 15:495–500
  modern use of philosophy in 18:35
  and moral neuroscience 16:614–616
  natural–scientific 19:313
  perinatal 17:722–726
  phenomenological–anthropological diagnostics 19:314–318
  philosophical 18:31
  and philosophical anthropology 19:313–314
  prophylaxis in 19:200–204
  psychotherapy in, history 19:502–508
  rehabilitation model 19:589b
  relevance of mental illness epidemiology of 15:184t
  and religion 20:348–349
    rationale for rapprochement between 20:349–351
  as science, Kraepelin 13:127–129
  and sociology 22:974–978
    countermovements against medicalization 22:976–977
    diagnosis 22:976
    empowerment and recovery *versus* sick role 22:976–977
    health care 22:974–975, 22:977–978
    mental disorder as medicalized social construct 22:976
    patient role 22:975
    post-World War II theory 22:975
    power and labeling 22:975
    social exclusion and stigma 22:977–978
  symptoms and concepts 19:316
  and women 25:622–627
  women in 25:624–626
Psychic autopoietic systems 14:390–391, 14:393–395
Psychic cost 15:434
The Psychic development of the white rat correlated with the growth of its nervous system 25:449
Psychic distance 12:583
Psychic life 13:68
Psychic pluralism, tenet of 5:479
Psychic secretions 17:609
Psychic systems 14:384
Psychic unity 5:479, 5:481
Psychoacoustics 13:123
  Stevens on 23:455
Psychoactive drugs 6:685, 19:462–463
  anticholinergic effects 19:464
  autonomic side effects of 19:464
  characteristics of 19:463–464
  and dependence 19:464
  use of term 6:685
Psychoactive substances 6:669
  use of 6:670–671
Psychoaesthetics 25:155
Psychoanalysis 2:701, 9:369, 13:703, 17:911, 19:336–343, 19:345, 19:350, 19:492–493, 22:527, 22:696–697, 24:524
  antifeminist 22:766
  for attachment 12:228–229, 12:232–233
  in Axial Age 24:726
  classical legacies 3:775–776
  and clinical social work 22:767
  as critical social science 10:787
  criticisms 19:506
  deconstruction and 5:930
  development 3:874
  elitist and irrelevant 22:765–766
  empirical validity 22:765
  feminist sociology, Erich Fromm 9:460–462
  focus on environment 22:767
  and free association 24:725–726
  Freud's theory of mental apparatus 19:336–339
    ego 19:338–339
    id 19:337–338
    superego 19:339
  Freyre on 9:421
  group psychotherapy 10:428–429
  history of 19:326–329
    European modernity 19:326
    Freud biographies 19:327
    from Freud to present 19:326–327
    historiography of psychoanalysis 19:328–329
    methodology problems 19:328–329
    psychoanalytic movement and institutions and practices 19:327–328
    schools of psychoanalysis 19:326–328
    theory of the unconscious 19:326
  hysteria 11:506–507, 11:509–510
  identity 11:524
  interpretation in, assent by the subject as criterion of correctness 24:723
  and metaphors 15:290–291
  metaphors and metonymy 15:290–291
  pathology of clients 22:766
  perspectives of self 19:400–401
  psychoanalytic treatment 19:339–342
    contemporary trends of 19:341–342
    countertransference 19:340–341
    object relations theory formulation of 19:340
    process 19:340
    psychoanalytic theory of psychopathology 19:339
    structural formulation of 19:339–340
    transference 19:340–341
  Riesman on 20:669–670
  role of Jung in 12:902–906
  STPD *see* Short-term psychodynamic therapy (STPD)
  transference 24:545–548
  treatment results 19:337
  Yoga 24:726
Psychoanalysis, in sociology 19:330–335
Psychoanalysis in sociology 19:330
  and Castoriadis 19:332–333
  and Freud 19:330–331
  integration of 19:331–334
    current trends in European psychoanalysis 19:332–334
    disciplinary articulations 19:331
    feminism 19:332
    Laplanche, Jean 19:334
    postmodernism 19:332
    poststructuralist psychoanalytic sociology 19:331–332
    structuralist psychoanalytic sociology 19:331–332
  and Kristeva 19:333–334
Psychoanalytic anthropology 19:362–363
Psychoanalytic approach to extraversion 8:623
Psychoanalytic interview 5:851–852
Psychoanalytic movement 19:327–328
Psychoanalytic psychotherapies 19:336, 19:345
Psychoanalytic theories 20:355, 22:2
  of neonatal identification with vital impulses of other 16:483–484
  in parenting and behavioral problems 2:472
  of psychopathology 19:339
  of socialization and education 22:862

Psychoanalytic thought, contemporary critiques of 22:764–767
Psychoanalytic treatment 19:339–342
contemporary trends of 19:341–342
countertransference 19:340–341
object relations theory formulation of 19:340
psychoanalytic theory of psychopathology 19:339
structural formulation of 19:339–340
transference 19:340–341
Psychoanalytic treatment process 19:340
Psychoanalytical naturalism 4:268
Psychobiological Model of Temperament and Character 8:625–626, 24:184, 24:187
Psychobiological process, in self-regulation 21:554
Psychobiology
of attachment 2:138–139
of HPA axis 11:476–481
of stress 23:571–572
Psychocardiology 19:487
Psychocultural model 3:452
Psychodrama 10:429–430
Psychodynamic characteristics, of bipolar disorder 2:685
Psychodynamic factors of mania 14:486
Psychodynamic psychotherapy 19:344
Psychodynamic theories 22:816
on therapeutic relationships 24:287–288
Psychodynamic therapy 11:612, 11:615
in gastrointestinal disorders 9:609
short-term 19:344–348
for somatoform disorders 23:10
Psychodynamic thought and social work 22:761–768
attachment 22:762, 22:765
conceptions of mind 22:761
in classical Freudian theory 22:761
in ego psychology 22:761
in Lacanian theory 22:762
in object relations theories 22:761–762
in relational and intersubjective theories 22:763
in self psychology 22:762
contemporary critiques of psychoanalytic thought 22:764–767
controversies involving Freud's legacy 22:764–767
human development 22:761–762, 22:767
psychoanalysis
antifeminist 22:766
and clinical social work 22:767
elitist and irrelevant 22:765–766
empirical validity 22:765
focus on environment 22:767
pathology of clients 22:766
psychopathology
in classical Freudian theory 22:763
in ego psychology theory 22:763
in Lacanian theory 22:763
in object relations theories 22:763
in relational and intersubjective theories 22:763
in self psychology theory 22:763
psychopathology and treatment 22:761
self-object 22:762
treatment
brief psychodynamic treatment 22:764
classical Freudian 22:763–764
ego psychology 22:764
Lacanian 22:764
object relations 22:764
relational and intersubjective 22:764
self psychology 22:764
unconscious 22:761, 22:763–764
Psychoeducation 4:545
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
for panic disorder/agoraphobia 17:478
Psychogenic amnesia 1:651–658, 24:563
age at diagnosis 1:652–653
cerebral blood flow 1:654–655
changes in personality dimensions 1:653
differential diagnosis of 1:655
epidemiological data about 1:652–653
neuroimaging of 1:654–655
neuropsychology of 1:653
pathogenetic models of 1:653–654
prognosis 1:655
recurrence of 1:653
risk factors and comorbidity 1:653
treatment 1:655
types of 1:652
see also Functional amnesia
Psychogenic model of pain 3:556
Psycho-historical biography 2:646
Psychohistorical research 19:350–351
Psychohistorical thesis 19:351
Psychohistory 7:935, 19:349–352
crisis years 19:351
development 19:349–350
meaning 19:349
Middle Ages 15:350
problems 19:350–351
Psycholexical approach, five-factor model (FFM) 2:562
Psycholexical studies of personality 9:250–251
Psycholinguistic Assessments of Language Processing in Aphasia (PALPA) 3:793–794
Psycholinguistic variables 4:92
Psycholinguistics 14:228, 15:755, 19:353, 21:350, 21:355–357, 25:578
crosslinguistic comparisons 19:356–357
current-day 23:264–265
Heider on 10:753–754
message construction and perspective 19:356
neurolinguistics 19:357
and semantics 21:565, 21:569
sentence comprehension 19:354–355
sentence production 19:355–356
word recognition 19:353–354
Psycho-linguistics 15:826–827
Psychologica 13:902
Psychological abuse, of elders 7:330
Psychological acculturation 1:68–72
framework 1:69f
Psychological adaptations 4:1–2
methods for investigating 4:2–4
cross-cultural studies 4:3
cross-species comparisons 4:3
developmental studies 4:3
genes, fossils, and other sources of evidence 4:3–4
laboratory and naturalistic experiments 4:2
neuropsychological dissociations 4:2–3
Psychological adjustment, implications of shame and guilt for 21:476
Psychological and Ethnological Studies on the Origins of Music 21:963
Psychological androgyny 21:695
Psychological anthropology 7:145, 19:359–365
childhood studies 19:362
cognitive anthropology 19:360–362
cognition in practice/distributed cognition 19:361–362
cultural consensus/cultural domain analysis 19:361
cultural epidemiology 19:362
cultural models 19:360–361
current and emerging directions 19:362–363
electronic resources 19:363–364
intellectual background 19:359–360
methodological and theoretical issues 19:363
psychodynamic approaches 19:362
Psychological approaches
for art 2:16
to entrepreneurship 7:683
panic disorder 17:470
evidence regarding 17:470
psychological models 17:470
to social media 22:422–423
spirit possession 23:282–283
to suicide 23:682–683
Psychological aspects
of cancer screening 3:79–81
virtual spatial navigation 16:350–351
attention 16:351
strategies, behavior, and cognitive processes 16:350–351
Psychological assessment 6:342, 24:520
Psychological attributes, measurement of 14:870
Psychological barriers to negotiation 16:419–420
Psychological behaviorism 5:77
Psychological benefit 5:58
Psychological capital (PsyCap) 2:246, 2:248
Psychological climate, of organizations 17:322–324
Psychological coherence, identity, nonheterosexual British Asians 16:864
Psychological concept of schools 21:128
Psychological consequences of mass trauma events

Psychological consequences of mass trauma events (*continued*)
in adults 14:747–748
in children and adolescents 14:748
Psychological construction model, for emotions 7:462
Psychological constructionism 7:459–460
and emotional expression 7:459–460
and emotions 7:528–531
Psychological constructivism 19:491
Psychological consumption, and social networks 16:557–559
Psychological continuity, and personal identity 17:743
Psychological contract breach (PCB) 19:366
and affective reactions 19:368
antecedents 19:368
and behavioral reactions 19:368
consequences of 19:368
and employee attitudes 19:368
moderators of 19:368–369
versus related constructs (theoretical distinctions) 19:367–368
Psychological contract violation 19:368–369
Psychological contracts 19:366–371
antecedents and 19:368
characteristics
idiosyncratic 19:366
perceptual 19:366
typology 19:366
violation 19:366
consequences of 19:368
future research opportunities 19:369–370
moderators of 19:368–369
person factors as 19:368–369
situational factors as 19:369
psychological contract breach *versus* constructs 19:367–368
theoretical perspectives 19:367
equity theory 19:367
group value model 19:367
social exchange theory 19:367
Psychological correlates of wisdom 25:593–595
Psychological cost 5:58
Psychological desiderata 2:792
Psychological development 16:505–508
by changing interfunctional structure of consciousness 25:331
as cultural development 25:331–332
in infancy 16:506–507
Klein's studies 19:504
through abstract tier 16:507–508
through representational tier 16:507
tracking development of psychological structures 16:506
web 16:506
Psychological diagnostic testing 25:794
Psychological disorders 25:438
and psychological stress 23:586
Psychological distance, FTP and FO 9:563
Psychological distress
among immigrants 15:407–409
of cancer 3:73–75
Psychological effects of racial discrimination 6:544
Psychological empiricism 7:569–570
Psychological empowerment, workplace autonomy 25:719–720
Psychological essentialism 23:446
Psychological events 3:261
Psychological explanations 8:594
Psychological expression, of fashion 8:837–838
Psychological factors, behavioral medicine 2:465
Psychological first aid (PFA) 22:756–757
goals 22:756
Psychological First Aid (PFA) Mobile 19:380
Psychological force 12:791–792
Psychological (dys)functioning, rheumatoid arthritis 20:657
Psychological game theory 2:447
Psychological games 9:589
Psychological health
and abusive supervision 23:702
during competition/cooperation 4:858, 4:860
Psychological health outcomes, bereavement 2:532–533
Psychological history 5:420, 5:422
Psychological holism 12:79
Psychological insulin resistance (PIR) 6:338
Psychological intelligence 17:751
Psychological interdependence family model, and parenting 17:495–497
Psychological interventions
for grief 2:539–540
for sexual dysfunction
in men 21:690–691
in women 21:691–692
Psychological job analysis 12:811–812
Psychological mindedness 17:751–752, 17:754
Psychological models, for obsessive-compulsive disorder 17:122–123
Psychological needs 21:489
Psychological perspectives, on sex offenders 21:674–675
Psychological political science 11:556
Psychological predicate 10:113
Psychological process, influence on phantom limb 18:4
Psychological racism 19:835
Psychological reality
of grammatical structures 21:623
of linguistic concepts 19:353
Psychological refractory period, task network models 16:540
Psychological resilience, longitudinal data 14:346–348
abstract chronological charts 14:347
aggregation of reduced response vectors 14:347
data 14:346
narratives 14:346–347
person-centered strategy 14:346
problem 14:346
reduced response vectors 14:347
tests of distinguishability 14:347–348
Psychological resources 2:867
Psychological Review 15:897
The Psychological Review 25:449
Psychological science 24:517–518
Psychological stages 4:43
Psychological states
during competition and cooperation 4:858
Wittgenstein on 25:613
Psychological statistics 15:304
Psychological stresses 7:805, 23:583–587
conceptual challenges and problems 23:585
coping 23:585
diseases 23:583–584, 23:586–587
disorders 23:583–586
environment–organism transaction perspective 23:584–585
epigenetic inheritance 7:835
generality *versus* specificity 23:585
historical considerations and concerns 23:583
individual differences 23:585
life events 23:584–586
methodological considerations 23:585–586
and occupational health 17:133–135
physical illness 23:583–584, 23:586
psychological disorders 23:586
recent advances and future directions 23:586–587
research traditions and major definitional perspectives 23:583–585
response perspective 23:584
stimulus or environmental perspective 23:584
Psychological structures
as dynamic constructions 16:508–509
integrative, dynamic self-organization of 16:508
tracking development of 16:506
Psychological testing, forensics 9:354
Psychological theories
engagement 1:32
of socialization and education 22:862–863
Psychological therapies
evidence for 8:748–749
anxiety 8:748–749
depression 8:748–749
Psychological tools 16:511–512, 25:324
internalization of 16:512
Psychological transformations 16:170
Psychological treatments
combined with pharmacological treatments, for obsessive-compulsive disorder 17:125
controlled clinical trials developments 19:386–387
effectiveness 19:389
effectiveness of 19:372–378
empirically supported 19:382–385
accountability 19:382
clinical practice guidelines 19:382–384
clinical trials 19:384
controversy 19:384
depression 19:383

effectiveness of interventions 19:382–384
evidence-based treatment 19:384
identification 19:383–384
managed care 19:382
obsessive-compulsive disorder 19:383
optimal interventions 19:382
panic disorder 19:383
practice guidelines 19:382–383
psychotherapy 19:382–384
treatment manuals 19:383–384
treatment outcomes 19:382
future developments 19:389–390
hypnosis 11:472–475
manual-based 14:500–503
nontraditional 19:379–381
evolution 19:379–380
information technology 19:379–381
Internet interventions 19:380–381
mobile applications 19:380–381
self-help 19:379, 19:381
social climate and rise of new forms 19:380–381
virtual reality therapy 19:379–380
wireless devices 19:380–381
for OCD 17:115–116
outcomes of controlled trials 19:389
cost-effectiveness 19:389
mediators 19:389
moderators 19:389
predictors 19:389
for post-traumatic stress disorder 18:705
randomized controlled trials 19:386–390
therapy process 19:388–389
common factors 19:388
therapist competence 19:389
treatment fidelity 19:389
treatment manual 19:388–389
types of controlled trials 19:387–388
clinical trial designs 19:387
multisite studies 19:387–388
noninferiority designs 19:388
Psychological types, Jung on 12:904
Psychological understanding, tacit theory and simulation in everyday 13:82–83
Psychological verbs
linguistic properties of 16:670–672
neurolinguistic processing of 16:670–677
neurolinguistic research using 16:672–675
argument structure 16:672–673
compositional semantics 16:674–675
conceptual representation 16:675
syntactic argument structure of 16:670
Psychological violence 25:116
Psychological well-being 2:792, 17:233–234
and authentic leadership 2:247–248
and marital quality 14:513
Psychologism
history of 19:391–393
Psychologism, history of 19:391
controversy in early-twentieth- century Germany 19:391–392
and philosophy 19:392
and sociologism 19:392
Psychologists
counseling 5:70
forensic, role of 9:351–354
Psychologization 20:48
Psychology 2:880–885, 4:394, 10:738, 10:740, 17:780, 18:31–36, 19:391, 19:406–411, 22:679–680
academic 1:853
antecedents of grit 10:400
and anthropology 13:71–72
application of information theory in 12:108
applied, technological change and sociotechnical system 24:104–108
applied social research 1:853–854
approach to accident and injuries study 12:130–131
Aristotle's 19:407–409
attributional processes 2:201–204
avoidance and approach motivation 2:326–331
in bachelor's degrees 22:680–681, 22:680t
behavior therapy in 2:418–419
behavioral economics 2:437–442
and behavioral economics 2:443–448
Bruner on 2:875–876
of burnout 2:929–932
of cancer 3:73–77
childhood cancer 3:459–464
clinical 1:854
and Cold War 4:124
of collective memory 4:188–193
comparative 4:394–395
of competition and cooperation 4:856–861
conformity 4:1–4
consumer 4:738–743
and contextual studies 4:786
counseling 5:69–72
counterinsurgencies 5:99–102
creativity 17:770, 17:772
cross-cultural research methods in 5:318–321
crowds 5:361–362
cultural *see* Cultural psychology
and culture of poverty 5:590–593
of decision making 5:917–922
of deindividuation 6:38–41
Descartes's 19:409–410
developmental 10:132
see also Developmental psychology
Dilthey on 6:424–426
disability 6:467–471
consequences and adjustment 6:470
determinants 6:469–470
disgust in 6:546
doctoral degrees 22:680t
economic *see* Economic psychology
and education 7:145, 7:147
of education
style construct in 4:110
and electroencephalography 7:362
and emotions 5:548, 7:513–515, 7:517
of emotions 7:526–533
empirical legal studies in 7:554–556
empiricism 19:410–411
epigenetics in 7:842–844
epistemology 19:406–407, 19:410
ethical questions in 8:55
evolutionary 4:395
experimental 1:853
of food preferences 9:296–299
forensic 1:854
Freyre on 9:421
gastrointestinal disorders 9:605–611
gender in 9:790–791
Greece 19:406–408
Harlow on 10:526–528
health 10:674–678
Hippocrates 19:406–407
historical background 3:877
history of 19:412–417
beginnings of experimental psychology 19:413–415
behaviorism 19:413–414
challenges to twentieth-century conventions 19:417
constructivism 19:414–415, 19:417
developmental psychology 19:414–415
at end of twentieth century 19:416–417
functionalism 19:413
gestalt psychology 19:414
influence of logical positivism 19:414
late-twentieth-century developments 19:415–416
learning 19:412–416
in midcentury 19:415
neuropsychology 19:415
neuropsychology, neuroscience, and evolutionary psychology 19:416
nineteenth-century foundations 19:412
postmodernism 19:417
pragmatism 19:413, 19:417
resurgence of cognitive psychology and cognitive science 19:415–416
social psychology 19:414
subfields and subspecialties 19:414–415
voluntarism, structuralism, and phenomenology 19:413
and Honesty-Humility 11:173
of human resource management 11:348–352
humanistic psychology 20:739–740
in indigenous country 11:789
of indigenous people 11:789
of individual differences 17:314
industrial and organizational 1:854, 11:879–884
infectious diseases 12:32–36
of inferences 19:394–399
in discourse 12:37–44
of institutions 15:534
of interest 12:378–385
of interpersonal attraction 12:626–630
James's functionalist 18:809
and law 21:158–159
of love and intimacy 14:375–381
of mass killings and genocide 14:702
master's degrees 22:680t
mathematical *see* Mathematical psychology

Psychology (*continued*)
mechanisms of placebo response 18:153–154
Middle Ages 19:408–409
mind–body problem 19:409–410
of money and finances 15:735–740
moral reasoning in 15:803–805
of motivation and actions 15:887–893
of negotiation and conflict 16:423–429
neonatal imitation in 16:481
before nineteenth century 14:808–810
in nineteenth century 14:808–810
of oral health 17:279–283
of organizational behavior 17:313–316
of peace promotion 17:654–659
of personnel selection 17:936
perspectives of values 25:36–40
philosophical 18:33–34
Piaget's theory of cognitive development 18:121
Plato on 19:406–408
political 18:427–433, 18:463–464
practical 1:853
prejudice in society 18:834–838
pre-Socratic philosophy 19:406
projective methods in 19:129–134
prospect theory 19:262–265, 19:268
psychoneuroimmunology 19:445–450
quantification in 19:704
of racism 19:858–860
rationalism 19:411
of real self 19:400–405
of reasoning 5:939–940
of religion 20:384–385, 20:389
Renaissance 19:409
research traditions in 12:314
Rome 19:408
of rumors 20:788–790
of self-conscious emotions 21:475–480
sensory 21:616–617, 21:619
social cognition 22:204–209
of social comparison 22:210–215
social constructionism in 22:226
of social dilemmas 22:245–249
of social exchange 8:483
social network analysis in 16:525–527
of speech production 23:255–258
of spelling 23:264–273
subfields of 19:359
of subjective wellbeing 23:648–653
of suicide 23:676–678
aggression directed against ego 23:677
appeal to human bonding 23:677
classical and operant conditioning 23:677
distorted cognitive schemes 23:677
model learning 23:677–678
narcissistic crisis 23:677
of sustainability lifestyles 23:830–831
of taste and smell 24:26–31
tools 25:330–333
of transportation 24:623–628
uncertainty in 24:719
of vocational interests 25:239
of volunteerism 25:268–272
Watson's contributions in 25:448–449
of well-being and burnout in workplace 25:534–536
of white-collar crimes 25:553–554
of wisdom 25:590–597
Wundt's contributions 25:793–794
Psychology critical *see* Critical psychology
Psychology experiment 5:895–897, 5:899
Psychology of Interpersonal Relations (PIR) 10:751
The Psychology of Money 21:963
Psychology of religion
attitudes 20:356
brain processes 20:357
charismatic religion 20:356
cognitive science of religion (CSR) 20:356–357
conversion 20:356
developmental psychology 20:355–356
evolutionary developments 20:357
explanations in psychology 20:358
general issues 20:357–358
history 20:354–355
individual differences 20:355
meditation 20:356
mental health 20:355
mystical and religious experience 20:356
prayer 20:356
relation to general psychology 20:357
relationship with theology 20:357
separation between sociology and psychology 20:357
Psychometric constructs 6:345
Psychometric evaluation, cross cultural research 7:932
Psychometric function 14:800, 19:472
Psychometric intelligence, integrating with cognitive processes 12:292–293
Psychometric methods, for quality of life measurement 10:658–659
Nottingham Health Profile 10:658–659
SF-36 10:658
Psychometric modeling of extreme response style 23:759–761
Psychometric models, for genetic studies of behavior 9:934
Psychometric paradigm
health risk perception 10:703
of risk 20:711
Psychometric properties, cardiovascular reactivity 3:118
Psychometric Society 19:418
Psychometric tests 17:851
historical background 3:879
Psychometric theories of intelligence 12:303–304
Psychometrically sound assessment scales 2:598–599
Psychometrics 14:805, 14:812, 19:418–422, 19:436
analysis of choice data 19:440
classical test theory 19:423–430
combinatorial data analysis *see* Combinatorial data analysis (CDA)
concepts 19:419–421
developments in 19:421
historical background 3:879
history 19:418–419, 19:436
of human abilities 1:428–432
intellectual functioning assessment 12:249–251
item analysis 19:436
item response theory 19:436–439
mathematical psychology 14:812, 14:814
models for dichotomous responses 19:436–437
models for polytomous responses 19:437
multidimensional item response theory 19:438
noncompensatory multidimensional models 19:438
parameter estimation 19:438–439
projective methods 19:129–134
in R 19:418
use of term 19:418
Psychometrika 14:812
Psychometry, for age differences in intelligence 1:428–430
Psychoneuroimmunology (PNI) 2:464, 19:380, 19:445–450
cognitive behavior therapy and psychotherapy 19:448
common cold studies 19:446–447
emotional disclosure 19:448
heart disease 19:447
history 19:445
illness 19:446–447
immunity 19:445–449
to improve health outcomes 19:447–448
infections and vaccinations 19:447
mental health 19:448–449
oral health 17:281
relaxation techniques 19:448–449
and stress 19:445–446
wound healing 19:446
Psychopathological features, of bipolar disorder 2:684
Psychopathology 2:793, 11:603–605, 11:613, 17:914, 18:114–115
classical conditioning in development of 3:765–766
avoidance 3:766
exemplars 3:766
individual differences 3:765–766
in classical Freudian theory 22:763
and creativity 17:772
in ego psychology theory 22:763
historical background 3:877
and HPA axis 11:477–478, 11:480
Kraepelin on 13:127, 13:129, 13:131
in Lacanian theory 22:763
and neonatal bonding 16:469, 16:472–473
nonpersonality disorder 17:916–917
in object relations theories 22:763
and personality 17:829, 17:917–918
of personality 17:914–919
personality disorder 17:916–917
and philosophy 18:31–33
in relational and intersubjective theories 22:763

in self psychology theory 22:763
and self-esteem 21:523
and war/political violence 25:390–391
Psychopathy 13:681, 19:451–456
aggression 19:451–453
antisocial behavior 19:451–453
antisocial personality disorder 19:451
callous-unemotional traits 19:451–452
causes 19:451–452
environmental 19:452
genetic 19:451–452
conduct disorder 19:451–452
development 19:453
and morality 16:614–616
neural impairment 19:452–453
functional neuroimaging 19:452–453
structural neuroimaging 19:452
pathophysiology 19:452
decision-making impairment 19:452
empathy impairment 19:452
moral judgment impairment 19:452
and personality 5:672
and personality disorder 17:891–892
treatment 19:453
violence 19:451–453
Psychopharmacology
and neuroethics 16:664
Psychopharmacotherapy 19:457–458, 19:462
costs of 19:457–461
precautions 19:461
side effects 19:462–469
study designs 19:459
Psychophysical correlates of phantom limb sensations 18:3–4
Psychophysical experimentation 8:566–567
Psychophysical parallelism 13:702
Psychophysical theory and laws 19:470–476
Psychophysics 13:125, 14:802–803, 14:810
approaches to environmental decisions 7:709–710
category rating 21:604
cross-modal matching 21:604–605
diffusion process 6:399
direct scaling methods 14:803
achievements 21:605
exponents characteristics 21:605
intermodal comparisons 21:605
problems of 21:605–606
quantitative phenomenology 21:605
scales of sensation magnitude 21:605
Stevens's power law 21:605
Fechner's solution 21:604
functional equation methods 14:803
local *versus* global 21:605–606
magnitude estimation 21:604
magnitude production 21:604
magnitude scaling 21:604
mathematical psychology 14:813
nineteenth-century 19:470–472
nineteenth-century experimentation in 8:566–567
numerals 21:606
power law (Stevens) 23:453–454
scaling methods 21:604–605
signal detection theory 14:803–804
of speech perception 23:236–237
twentieth-century 19:472–475
use of term 19:470
Psychophysiological measures of attitudes 2:181
Psychophysiological problems, Wundt's contributions 25:792
Psychophysiological stress testing, coronary heart disease 4:918, 4:921
Psychophysiological symptoms, of psychosomatic disorders 19:485–486
Psychophysiology 2:464, 21:612, 25:138, 25:144–145
of flow 9:264
interoception 12:616–617, 12:619–620
Psychosis 8:746–749, 10:513, 21:9–13, 21:22–23
alcohol 1:510–511
Kraepelin on 13:127, 13:129–130
neuroimaging 14:441–442, 14:442f
religion and 20:264
transference 24:546
Psychosocial age 1:420
Psychosocial aspects
cardiac rehabilitation 20:187–192
diabetes 6:337–341
drug use and abuse 6:685–689
gynecological and obstetric health 10:462–468
of hospitals 11:208–214
hypertension 11:453–457
respiratory disorders 20:576–580
rheumatoid arthritis 20:655–659
Psychosocial associations, of childhood friendship 9:443
Psychosocial capacity building 10:720
Psychosocial care 10:719
and medical social work 10:719
Psychosocial causes, of health disparities 19:816
Psychosocial crisis 14:86
Psychosocial development
Erikson's theory of 7:938–946
across times 7:941
age and attributes associated with adulthood 7:941
'anatomy is destiny' interpretation 7:939
criticisms 7:938–939
gender-relevant differences in Eriksonian stages 7:941–942
identity status 7:938, 7:943–944
inventories' factor structures, validating 7:940–941
limitations 7:944
modifications 7:939–940
practical applications 7:944
research themes 7:940
scope and methodology 7:940
self 7:939, 7:943–944
society 7:943–944
stages, validation of 7:940–941
stages and correlates 7:942–944
trajectory of each stage 7:941
validation of eighth and ninth stages 7:941
individual freedom, liberalism and 14:24
Psychosocial development theory 22:943
Psychosocial developmental stages 7:934–935
adolescence years 7:935–937
adult years 7:935–937
'configurational' approach to children's play therapy 7:935
cross-cultural applicability, context of 7:935
crucial periods ('crises') of development 7:936
ego developmental stages 7:935
individual's senses of oneself 7:935
life span developmental schema of 7:935
middle adulthood years 7:935
optimal development in childhood 7:935
provision of 'better-than-average expectable' conditions 7:935–936
social institutional rituals, context of 7:935
Psychosocial educational environment 7:820
Psychosocial effects 8:572–573
of Alzheimer's disease 1:585–586
disease progression and the necessity of care 1:587–588
at individual and family level 1:586–589
last stage 1:588–589
onset of disease 1:586–587
stage of insecurity and emerging losses 1:586–587
of breast feeding 2:849
Psychosocial factors
cancer screening 3:79–80
of coronary heart disease 4:917–918, 4:917t, 4:918f
acute *versus* chronic 4:918–919
of gender gap in life expectancy 9:705–706
Psychosocial foundations, focus groups 9:275
Psychosocial functioning 12:637–639
Psychosocial health, and doctor–patient interaction 6:616
Psychosocial impacts, of gastrointestinal disorders 9:606–608
coping 9:607
health-related quality of life 9:606–607
self-image and sexuality 9:608
social functioning and stigma 9:607–608
social support 9:607
Psychosocial interventions 6:681–682, 22:380–385
for cancer 3:75–76
interventions for couples 3:76
interventions for patients 3:76
community psychiatry 22:380, 22:382–384
conceptual framework 22:381
International Classification of Functioning, Disability and Health 22:381–384
cognitive training 22:382
developing environmental resources 22:383–384

Psychosocial interventions (*continued*)
enhancing participation by vocational rehabilitation 22:383
family intervention 22:382–383
medication 22:382
reduction of impairments 22:382
remediation of persisting limitations 22:382–384
skills training 22:382
in perinatal psychiatry 17:724
target populations 22:380–381
vulnerability models 22:381
at work 17:138
Psychosocial issues, of AIDS 1:492–493
Psychosocial models 6:680
Psychosocial outcomes, and early friendships 9:452
Psychosocial perspective, social work 24:755
Psychosocial risk, of cardiac rehabilitation 20:189–190
Psychosocial strain, child health psychology 10:684
Psychosocial stress, cardiovascular reactivity 3:117
Psychosocial studies, in social constructionism 22:226
Psychosocial support, in disasters 22:756
Psychosocial theories
of health behavior 10:577–581
action planning 10:578–579
attitude 10:578, 10:580
comparison of theories 10:579–580
health action process approach 10:578–579
health belief model 10:577
intention 10:577–580
outcome expectations 10:578, 10:580
perceived barriers 10:577
perceived behavioral control 10:578–580
perceived benefits 10:577
perceived severity 10:577, 10:579–580
perceived susceptibility 10:577–579
pros and cons 10:579–580
protection motivation theory 10:577–578
response efficacy 10:577–578
self-efficacy 10:577–580
social cognitive theory 10:578
stages of change 10:579
subjective norm 10:578, 10:580
theory of planned behavior 10:578
transtheoretical model 10:579
Psychosocial work 19:477–483
children 19:479–481
critical issues 19:481–482
culture 19:477–480
current applications 19:479
definitions 19:477–478
disasters 19:479–482
dual-foci 19:478–482
emergencies 19:477, 19:479–481
emerging trends and directions 19:481
history, theoretical underpinnings and leading figures 19:478–479
international, shifts in 19:479–481
pathology 19:479–482
post-traumatic stress disorder 19:479–482
strengths 19:479, 19:481
Psychosomatic depression 8:341
Psychosomatic medicine 19:484–490
alternative terminology 19:484–485
concepts 19:484
consultation liaison psychiatry 19:484–485
criticisms of term psychomatic 19:485–486
disorders 19:486–488
assessment 19:488
behavior therapy 19:489
cardiovascular system 19:487
classification of 19:486–487
dermatology 19:488
diagnosing 19:486
gastrointestinal system 19:487
management 19:488–489
nephrology 19:488
obstetrics and gynecology 19:488
oncology 19:488
pharmacological treatments 19:489
psychotherapies 19:489
respiratory system 19:488
treatments 19:488–489
functional somatic symptoms 19:485–486, 19:488
general hospital psychiatry 19:484–485
historical aspects 19:487
holistic care 19:484–485, 19:489
medically unexplained symptoms 19:485–488
mind–body relations 19:484–485, 19:487–488
psychophysiological symptoms 19:485–486
somatization 19:486–488
somatizing patient in medical clinics 19:488
stress 19:485–489
Psychosurgery 16:665, 23:18
obsessive compulsive disorder 17:269
for obsessive-compulsive disorder 17:125
Psychotherapeutic drugs 6:685
Psychotherapy 4:724–726, 8:347, 11:609, 14:500, 19:317–318, 19:382–384, 19:515
alliance 19:517–519
classical conditioning methods in 3:764–770
combined with pharmacotherapy 19:495
constructivist approach to 19:493
definition 19:372, 19:502
development 19:504–506
effectiveness of 19:372–378, 19:506–507
common factors 19:373–374
follow-up studies 19:377–378
negative effects of 19:377
patient variables 19:373
treatment during and lasting effects 19:372–373
effects 11:615t
in anterior cingulate cortex (ACC) and insula 11:618f
ethnic and cultural differences in 5:63–68
cultural pluralism and tyranny of the shoulds 5:64
multicultural psychology 5:65–66
racial identity theory 5:66
racial–cultural identity development 5:65–66
in Europe 3:874
evidence-based psychotherapy 8:344–350
experiential 8:517–520
experimental designs 11:617t
feminist 8:929–934
feminist approaches 19:504
generic model of psychotherapy 19:516
Gestalt therapy 19:493
for grief 2:539–540
group *see* Group psychotherapy
and hermeneutic phenomenology 19:491–492
and hermeneutics 19:491–494
hermeneutics and 19:492–493
history 19:517
in psychiatry 19:502–508
humanistic schools 19:504
hypnosis 11:474
interpersonal *see* Interpersonal psychotherapy
intersession experiences 19:518–519
and liminality 14:135
mechanisms of action 11:613–614
for negative emotions 7:500
neural correlates of 11:615–616
neurobiological correlates of psychotherapy 19:519
neurobiological mechanisms 11:612–613
neuroimaging research on 11:613t
nontraditional 19:379–381
in orbitofrontal cortex (OFC) and striatum 11:617–618
patient and therapist experiences in psychotherapy 19:518–519
patient–therapist interaction 19:515–519
and patient–therapist relationship 24:288–290
and phenomenology 19:491–492
and postmodernism 19:491–492
for post-traumatic stress disorder 18:705
process research 19:515–520
established findings 19:518
focus 19:517–518
methods 19:518
new research areas 19:518–519
and practice 19:519
variables 19:517–518
progress feedback 19:519
in psychoneuroimmunology 19:448
for psychosomatic disorders 19:489
relationships 19:517–518
research 19:515–516
location 19:516–517
observational perspectives 19:517
temporal units 19:517
Rogers on 20:737–743
for schizophrenia 21:12
sessions 19:516–519

for somatoform disorders 23:12–14
for specific interventions 19:374–377
anxiety 19:374–375
depression 19:375–376
eating disorders and obesity 19:376–377
STPD *see* Short-term psychodynamic therapy (STPD)
and systems theory 19:505
theories
Adler's 19:503
Horney's 19:503–504
Jung's 19:503
theory 19:515, 19:517–518
and traditional medicine 19:502
and treatment planning 6:344–345
methodological issues 6:345–346
in United States 14:462–465
virtual reality in 25:138–146
Psychotherapy integration
assimilative integration 19:511
effectiveness of 19:512
historical context 19:509–510
pathways to 19:511–512
personality variables 19:511
recent growth reasons 19:510–511
research support 19:512
theoretical integration 19:511–512
Psychotic disorders 10:511–513, 21:9–13
Psychoticism 17:802–803, 17:914, 20:355
Psychotropic medications
in perinatal psychiatry 17:725
in psychiatric care 6:43
Psychotypology 16:66–67
PsycINFO database 22:944
PT *see* Planum temporale (PT); Prospect theory (PT)
Pt scale 15:567
PTA *see* Parliamentary technology assessment (PTA); Parliamentary technology assessment (PTA) bodies; Posttraumatic amnesia (PTA)
PTAs *see* Preferential trade agreements (PTAs)
PTB *see* Physical trade balance (PTB)
PTFR *see* Period total fertility rate (PTFR)
PTG *see* Posttraumatic growth (PTG)
PTMs *see* Posttranslational modifications (PTMs)
Ptolemaic astronomy 18:100
Ptolemaic cartography 13:225–226, 13:229
Ptolemaic dynasty 1:691
Ptolemaic theory 20:286
Ptolemy of Alexandria 20:285–286
Geographia 3:173–174
PTS *see* Political Terror Scale (PTS)
PTSD *see* Posttraumatic stress disorder (PTSD)
Puberty, and transition to adulthood 24:578–579
Puberty rites 5:509
Public
as advocate 17:243–244
genetic 9:991–992
and LGBTI commercial sex 13:884
and museums 16:143–145
obligations to, surveys and polls 23:753–754
Public Access to Court Electronic Records (PACER) 5:808
Public accountability 17:605–606
Public accounts 5:760
Public administration 2:901, 2:922, 19:521–525
and administrative law 1:113
civil service 3:695–698, 3:700
definition of 1:579
machinery of government 19:521, 19:523–524
modern 19:522–523
postmodern 19:523–524
premodern 19:521–522
public management 19:521, 19:523
new 19:523
public policy ethics 19:582–583
virtual and supranational 19:524
Public adoption 1:136
Public agendas 1:351–352
Public anthropology 1:225–226
Public attitudes 15:228–234
on female political quota 8:890–891
towards mentally ill people 15:228–234
Public awareness 1:848
Public benefits, and private vices 12:388
Public broadcasting 19:526–530
Public broadcasting, general concept and background 19:526–528
Public bureaucracies 2:901–902, 19:531–533
alternatives to 19:533
and change 19:531–532
as agents 19:531–532
expertise 19:532
competition among experts 19:532
think-tanks 19:532
governmental institutions 19:531–532
innovation 19:531–532
nongovernmental institutions 19:531–533
permanent government 19:531
power 19:531–532
regulatory bureaucracy 19:531, 19:533
secrecy and open government 19:532–533
Public campaigns, in international development 8:61–62
Public capital facilities 11:259
Public choice 17:209, 19:534–539
antecedents to 17:207–208
collective action 19:534–535
constitutions 19:535
delegation 19:535–536
democratic coherence 19:536
empirical implications and connections 19:538
human motivation and the problem of institutional design 19:537–538
information problems 19:536–537
key problems 19:534–537
Leviathan, controlling 19:535
principal-agent models 19:535–536
rent-seeking 19:537
transactions costs 19:534
Public choice approach, antitrust 1:800
Public choice theory (PCT) 5:253, 7:402, 19:534–535, 19:538
Public cognition 15:546–547
Public concerns
child labor as cause for 3:406
genetic testing 20:698–699
Public consultation 21:255, 21:258–259
frameworks 21:257–258
Public criterion of justice 12:943–944
Public disclosures
political money 18:395–396
Public discourses 15:19, 19:615
Public education 21:103
Public education system (in North America) 7:298–299
churches 7:299
cultural emphasis on 7:299
experimentation 7:299–300
innovation 7:299–300
legal history 7:298
military schools and education 7:299
reforms 7:300
religious schools 7:299
tertiary sector 7:299
Public employment service (PES), and labor market policy 13:184–185
Public engagement 5:525–526
and anthropological writing 1:740–742
in citizen science 3:631–635
and science communication 21:222
Public engagement with science (PES) 9:990–991, 18:6, 21:210, 21:255
citizen science and knowledgeable publics 21:256
deficit model 21:255–256
engagement and crisis of trust 21:257
engaging beyond engagement 21:259
frameworks of consultation 21:257–258
science and civic epistemologies 21:258
science and democracy 21:258
science communication as engagement 21:258–259
Public enlightenment 13:441
Public evaluations 2:332–337
Public evaluations of legal system
attorneys, of 13:747–748
courts, of 13:748
direct personal experience 13:748
distributive justice in 13:749
fear of crime 13:750
fear of legal sanctions 13:750–751
hindsight bias 13:748–749
legitimacy and compliance 13:751
mass media effects 13:748
outcome satisfaction in 13:749
police, of 13:747
procedural justice in 13:749–750
relative deprivation in 13:749
retributive justice in 13:750
Public expectations, health care organizations 15:101
Public expenditures
and value of life 25:19

Public expenditures (*continued*)
and welfare retrenchment 25:522–523, 25:525
Public funding 22:663–664
child care 3:393
Public funds, for higher education 24:759
Public futures, translating legacy knowledges into 21:183–184
Public goods 7:726, 10:179b, 15:23–24
biobank ethics and 2:605
and collective action 22:512–513
defense as 6:2–3
definition 19:540–541
international 19:540–544
Olson on 17:208
for philanthropy 3:359
republicanism 20:477–478, 20:481
state provision, liberalism and 14:23
Public goods dilemma 11:171, 22:245–246
Public goods game 11:259–260, 11:262–264, 12:721–722
Public goods theory 2:447, 17:8
Public governance, public sector organizations 19:600–601
Public health 2:496, 3:178, 10:728, 15:122, 15:124, 18:176–177, 22:386
addiction 1:100–101
and cancer screening 3:78–79, 3:81
challenge, health literacy as 10:654
definition 19:545–546
and educational neuroscience 16:730
evolutionary medicine 8:458–459
health information systems 5:781–784
and physical activity 18:91, 18:95
and sex offenders 21:676–677
and social justice 22:389
as social science 19:545–551
behavioral theory 19:545–546, 19:550
epidemiology 19:548–549
evidence base 19:549–550
future 19:550
health behavior 19:545, 19:548–549
noncommunicable disease 19:547–548
theories and models 19:548–549
trends in global burden of disease 19:546–548
Public health access 25:123–126
Public health consequences 20:269
Public health disaster response plans 16:318
Public health ethics 2:628–629
Public health genomics 4:352
Public health mental systems 19:586–590
Public health models 1:501–502, 18:883–884
Public health nurse 16:910–911
Public health policy, community-based health interventions 10:650
Public health social work 10:719–720
and disaster 10:720
disease prevention 10:716t
global health 10:716t
Public helping 19:221–222
Public history
metamorphoses of 11:52
sport history 23:308
Public housing 16:450–451
Public hygiene 15:118, 15:121
Public ideology 11:565
Public imagination 9:567–568
Public incivilities 5:191
Public information society 12:87–88
Public infrastructure investment 12:117
Public insurance, health care markets 10:598–599
Public intellectuals 25:349–350
Chomsky on 3:530
and cultural policy 5:463–464
Mills on 15:537
social scientists as *see* Social scientists; as experts/public intellectuals
Public interest 12:389
common good 19:552–553
community 19:552–553
concept 19:552
democracy 19:553
follow 19:553
freedom 19:553
good 19:552–553
implementation 19:553
justice 19:552
responsibility 19:552–553
self-interest 19:552
substantive manner 19:552–553
Public interest groups 12:372
Public interest law 19:555
antecedents 19:555–556
evolution
funding 19:558–559
globalization 19:559–560
growth 19:556–557
ideology 19:557–558
firm 19:556
Public interest litigation 23:62
and consumer policy 4:735
Public interest theory 19:535
Public journalism 3:613
Public journalism movement 14:694
Public justification 6:86
democracy as 6:86–87
and public reason 19:594–595
Public landownership in China 5:598
Public law 3:27, 3:669, 16:356–357, 19:185, 19:188
Latin America 13:410–411
Public law tradition
United States presidency 18:862
Public legal education (PLE) 13:736–737
Public libraries 14:45–49
Public life 24:794
and emotions 7:515
reemergence of religion in 24:603–604
Public management 19:521, 19:523
Public mental health systems 19:586
decline of asylums 19:587–589
deinstitutionalization movement 19:588
post-deinstitutionalization 19:588
psychotherapy, rise of 19:587
rise of public psychiatry 19:587–588
treatments, advances in 19:587–588
future 19:590
latest initiatives in 19:589–590
England 19:590
United States 19:590
and psychiatric rehabilitation 19:588–589
reforms 19:589–590
recovery movement 19:589
rise of asylums 19:586–587
middle ages 19:586
renaissance 19:586–587
Public miracles 15:576–577
Public misperception of crime rates 13:750
Public monopoly broadcasting, of radio 19:877–880
Public moralists 12:274
Public motivations, and citizen science 3:633–634
Public office 5:12–13
Public opinion 2:193–196, 13:568, 13:747–752, 19:562–568, 19:603–604, 19:611–612, 21:791, 25:395–396, 25:398
'activated' opinions 2:195
on affirmative action 1:267
from audience to public 19:612
collective action 19:562
contemporary studies of 2:193
democracy 19:562, 19:564–566
ethnocentrism 19:565
factors that shape aggregate-level opinions 19:566
factors that shape individual opinions 19:564–566
formation *see* Opinion formation
'fully informed' choices 2:194
ideology 19:562–564
impact of constituents' opinions 2:194
international social survey program 19:563
judicial behavior 12:888–889
and mass media 14:712–714
media 15:12
Neyman on 19:563
normative concerns about the quality of 2:194
pioneering analyses of 2:193–194
political consequences of 2:194–195
poll 19:563–564, 19:566–567
and polling 18:463–464
research 18:460–461
racism 19:564–566
research, globalization of 19:566–567
and rumors 20:788
and science communication 21:221, 21:223–224
scientific study of 19:563–564
scope and methods 2:193
as social force 19:566
social movement 19:562
stereotype 19:562–563, 19:565
survey 19:563–567
surveys and polls 23:752–754
from theater to tribunal 19:611–612
Tocqueville on 24:423–424
uncomplimentary characterizations of 2:194
world values survey 19:566–567

Public opinion polling 23:749–750
Public Order Administrative Penalty Regulation 21:736
Public order offenses, and victimless crime 25:77–78
Public organizations 2:902, 2:904
  strategic planning 23:515–521
Public Outreach Assessment 1:225
Public participation 18:184
  constructive technology assessment (CTA) 24:126
  environmental risks 7:788
  in Genographic Project 10:24–25, 10:27
  in health 4:402
Public participation GIS (PPGIS) 3:620
Public participation in scientific research (PPSR) *see* Citizen science (CS)
Public pensions 20:603–605
Public philosophy 12:87
  America in search of 20:895–896
Public police, evaluation of 13:747
Public policies 1:323–326, 2:901, 2:903–904, 4:946
  anti-crime 5:176–177
  case example 3:178
  case programs 3:179
  case study 3:179–180
  and diaspora 6:354–355
  for environment 7:710–711
  ethics of 19:580–585
  evidence-based policy making 19:570–572
  global *see* Global public policy
  government tools 19:572–573
  instruments 19:572t
  knowledge 18:294
  learning 19:573
  origins 19:569–570
  outlook 19:574
  path dependency 19:573
  policy analysis 19:570–572
  policy science 19:570–572
  political agendas 1:363–364
  political institutions 19:569–570, 19:573–574
  politics and institutional constraints 19:573–574
  prevention science 18:889
  process 19:569, 19:571–574
  public spending 19:569–570, 19:573
  RAPID outcome mapping approach 19:571f
  RAPID outcome mapping approach (ROMA) 19:571f
  and religion 20:364–365
  science and technology studies 21:204
  on skills training 22:22
  for social sciences 6:489
  urban *see* Urban policy
  and welfare retrenchment 25:525
Public policing 18:248–249
Public policy analysis 18:273
Public policy analysts 18:273
Public policy experiments 8:531, 8:533
Public policy issues
  child care 3:393
Public Policy Research Center (NUPP) 7:183
Public policy schools 19:576–579
  challenges to 19:578–579
    difficulty of management 19:579
    intellectual challenge 19:579
    leadership education 19:579
    in problem solving 19:578
  changes to 19:578–579
    tradeoff between breadth and depth 19:579
  concentration on analytics and professionalism 19:577–578
  and other programs 19:577–578
  similarities in 19:577
Public practice, infrastructure investment as 12:117–118
Public psychiatry 19:586–590
  decline of asylums 19:587–589
    deinstitutionalization movement 19:588
    post-deinstitutionalization 19:588
    psychotherapy, rise of 19:587
    rise of public psychiatry 19:587–588
    treatments, advances in 19:587–588
  future 19:590
  latest initiatives in 19:589–590
    England 19:590
    United States 19:590
  and psychiatric rehabilitation 19:588–589
  reforms 19:589–590
    recovery movement 19:589
  rise of asylums 19:586–587
    middle ages 19:586
    renaissance 19:586–587
Public rationality, and public reasons 19:592
Public realm 12:777–778
Public reason 19:591–596
  conceptions 19:592–595
    convergence accounts 19:594–595
    epistemic theories 19:592–593
    membership in public 19:595
    public rationality 19:592
    public reason qua political reason 19:594
    shared practice theories 19:593–594
  democracy 19:594–595
  diversity of reason 19:591–592
    enlightenment 19:591
    pluralism and indeterminancy 19:591–592
    reasonable pluralism 19:592
    relativism and social science 19:591
  Habermas on 19:593–594
  liberalism 19:594
  in plural society 19:595–596
  political liberalism 19:594
  political philosophy 19:592, 19:594
  public justification 19:594–595
  Rawls on 19:592, 19:594–595
Public reason qua political reason 19:594
Public recognition 2:332–337
  academic distinctions 2:333
  awards, medals, and prizes 2:333–335
  honorary degrees 2:335–336
  international awards/prices 2:334–335
  social science and humanities prizes 2:335
  prestige 2:333–334
  ranking research 2:336
  renown 2:334–335
  and reputation 2:332
  scientific capital 2:333
Public relations (PR) 15:43
  and science communication 21:221–222, 21:224
Public rules of legislation and administration 5:251–252
Public safety, and police use of force 18:255
Public schooling, educational policy 25:483, 25:485, 25:487–488
Public science 17:13
Public secrecy 21:374–377
Public sector 3:695
  applied demography in 1:840–841
  cost–benefit analysis 5:50–51
  employment 10:519
Public sector consumerism 16:774
Public sector employment, and social work 24:491, 24:493
Public sector management 16:770
Public sector management, education as 7:263
Public sector markets 16:770–771
Public sector organizations (PSOs) 19:597
  agencification 19:597
  autonomy 19:599–601
  classifications 19:599
  consequences for governing 19:600–601
  coordination 19:601
  decentralization 19:598–600
  definition 19:599–600
  embeddedness 19:600–601
  fragmentation 19:600–601
  functions 19:598–599
  groups of organizations 19:600
  heterogeneous organizational forms and blurring boundaries 19:599–600
  new managerialism in
    approaches 16:772–773
    extension of competition 16:772
    managing performance 16:771–772
    measuring performance 16:771–772
  and new public management reforms 19:597–598
  new types of activities 19:598
  para-government organizations 19:597
  privatization 19:599
  public governance 19:600–601
  quangos 19:599, 19:601
  reaggregation 19:601
  redefining organizational boundaries 19:600
  state-owned enterprises (SOEs) 19:598
  whole-of-government approach 19:601
Public service broadcasting 2:212, 2:860–861, 14:693, 19:527
  principles of
    commitment to education of the public 19:528

Public service broadcasting (*continued*)
encouraging competition in good programming 19:529
independence from vested interests 19:529
provision for minorities 19:528
rules of broadcasting 19:529
serving the public sphere 19:528
universality of appeal 19:528
universality of availability 19:528
in twenty-first century 19:529
Public services 19:604, 24:789–790, 24:793
market geographies 14:544
Public sociology 1:858–860, 20:815, 22:696
Public space
dangerous 21:673
and gender 9:709, 9:711
privatization of 24:790, 24:793
urban design and 24:789–794
Public spending, public policy 19:569–570, 19:573
Public sphere 3:701–706, 9:770, 14:735, 19:610–614, 20:375, 24:164–165, 24:167
Arendt on 3:704
Burke on 2:927
and climate change 3:854
culture-formation 3:705
definition 19:610–611
French philosophers 19:611
Habermas's political public sphere 19:610
Kant's public use of reason 19:610–611
differentiation in 3:704–705
eighteenth-century history 19:610–614
Habermas on 3:703
and households 11:231
idea 3:702–703
mass communication 14:693–694
and media 15:23–24, 19:603–609
critical issues 19:607–608
mass media 19:605–607
principle of publicness 19:603–605
public 19:605–607
nineteenth- and twentieth-century 19:615–619
cities and media 19:617–618
history as decline 19:616
idealization 19:616
participation and exclusion 19:616–617
publicness 19:615
nineteenth and twentieth-century history 19:615–619
and private sphere
and liberalism 14:23
prohibited books, subversive words 19:612–613
public opinion 19:611–612
from audience to public 19:612
from theater to tribunal 19:611–612
and science communication 21:219, 21:221–224
sexual citizenship and 21:729
state/society 23:345
women and 19:613
Public sphere theory 14:730
media 15:13–14
Public subsidies, political money 18:394–395
Public support
of arts 2:2–3
and cultural resource management 5:490
Public transit 24:624
Public understanding of science 9:976–977
Public Use Microdata Files (PUMF) 3:287
Public use of reason 19:603
Public Utilities Regulatory Policies Act (PURPA) 18:367–368
Public value 19:529
Public violence 25:128–132
coordinated destruction 25:129–131
ritual harm 25:131
crude typology of 25:129f
episodic social interaction 25:128–129
fragmented resistance 25:129, 25:131
opportunism 25:129, 25:131
organizational outcomes 25:129–130
violent games 25:129, 25:131
visible social processes 25:128, 25:130
Public visual communication 3:162
Public Voices Fellowship Program 1:742–743
Public welfare 8:531, 25:528
and technology 13:522–523, 13:525–526
Publication 7:868
of research 20:511–512
slant of 14:715
Publication bias 20:441–442
meta-analysis 15:267–268
in systematic review 23:907–908
Publicity 6:64–65, 19:603
of genetics 9:990–991
science 21:251–252
and science communication 21:220
Publicity of movements 15:21
Public-law disputes 13:615
Publicly financed research institutions, infrastructure 22:670
Publicly traded corporations 4:939–945
Publicmarkets 4:940
Publicness 19:603–605
Public–private development 20:18
Public–private dichotomy 9:768–769
Public–private hybrids 2:770
Public–private partnerships (PPP) 3:843–844, 13:445, 24:795–796, 24:841–842
and governance 10:293–294, 10:296–297, 10:294f
and urban planning 24:842
at work, phase model of 10:298f
Publishers 19:620–623
Publishing 19:620–630
bibliography 19:621
booksellers 19:620–622
copyright 19:620, 19:622–623
culture 19:627
digital media 19:622–623
as field of study 19:621–622
historical perspectives 19:621
social science perspectives 19:621–622
history 19:620–621
mass media 19:621–622
popular culture 19:621–622
print culture 19:621
research themes 19:622–623
dynamics of concentration 19:622
globalization and national identities 19:622–623
technological challenges 19:623
university press *see* University press
Publishing cycle 12:862–863
Publishing ethics 12:864
Committee on Publication Ethics 12:864
growth in article retractions 12:864
Publishing industry, gatekeeping in 9:619–620
Publishing platforms, journals 12:867
Publizistikwissenschaft 4:280
PubMed 8:328–329, 20:530
Pudeur 21:871
Puériculture 8:241, 8:246–247, 8:249
Puerperal psychosis 17:723
Puerto, Felipe Carillo 15:329
Puerto Madero megaproject 24:839–840
Puerto Rico 1:638
PUFA *see* Polyunsaturated, long-chain fatty acid (PUFA)
Pufendorf, Samuel von 13:802, 21:167
on political thought 18:481
Puissance absolue 1:6–7
Puja (online Hindu worship) 20:276
Pular language 16:266–267
Pull factors, Organon model 7:484
Pulling levers 11:167
Pulling levers strategy 10:460
Pull-out classes 2:582
Pulmonary rehabilitation 20:576, 20:578–580
PUMF *see* Public Use Microdata Files (PUMF)
Punch cards 21:186, 23:376–377
Punctual axiom 4:872
Punctuality, and conscientiousness 4:659
Punctuated equilibrium 2:903
political agendas 1:363
Punctuated equilibrium model, of agenda-setting 1:359
Punditry model of news 14:715
Punic Wars 17:638
Punishment 2:526–527, 5:216–217, 5:232–233, 5:874–875, 5:877, 6:255–256, 6:258, 7:80–81, 7:240, 13:570, 17:859–860, 18:879–880, 19:631–635
allocation 19:634–635
alternative 19:651–652
in applied behavioral analysis 2:408
in the Australian states 19:640f
behavior therapy 2:429
beyond penal state 19:634–635
capitalism and 19:641
censure and sanction 19:632–633
certainty of 6:260–263, 13:571

civil law 19:632–634
comparative politics of 19:636–654
and conflict 4:624–625
cultural consistency and 19:646
delivery 19:634
and deterrence 6:250–253
elite control of crime and punishment policy 19:653
historical rarefaction 19:636–654
Church endorsement of hellish 19:639
penal sentimentalism and elite indifference 19:639
state elites' conversion to leniency 19:636–639
and human cooperation 11:263–264
internationalization of rules and norms 19:648–651
legal requirements and justification 19:631–632
in Netherlands 19:638f
and offender rehabilitation 20:195–196
and parenting 2:472
premodern forms of 18:940
pretrial detentions as 19:651
prosocial behavior 1:580–581
punitiveness 19:632–635
as sentencing 19:633–634
severity of 13:571
since 1880 19:641f
social and legal aspects 19:655–661
birth of prison 19:655–656
capital punishment 19:655–656, 19:658
civility 19:658
culture 19:659
dignity 19:656–657
festival of cruelty 19:655
human rights 19:656–659
law and society 19:655
legality 19:658–659
mass incarceration 19:656
mass supervision/new banishments 19:659
migration 19:659
parole 19:656–659
penal field 19:655–656
penal variation and change 19:657–659
penality 19:655, 19:657, 19:659–660
political technologies of body 19:657–658
probation 19:656, 19:658–659
risk and prevention 19:659
social structure 19:657
solidarity 19:657
surveillance societies 19:656
social control 4:823–826, 19:634
and social control 22:235–236
state monopolization of 19:636
torture and 24:442–446
utilitarian social thought 24:982
women offenders 25:616–621
Punishment power 8:485
Punitive damages 20:415–417
levels of awards of 20:416
rate of awards of 20:415–416
relationship with compensatory awards 20:416–417
time trends of 20:417
Punitive populism 19:642–643
Punitive upsurge 12:960
Punitiveness 19:632–635
and probation 19:46
Punjab Female School Stipend Program 4:557
Punnett, Reginald Crundall 9:222
Punning 11:417
Puranas 23:70
Purchase, consumer psychology 4:738–740
Purchase of Development Rights programs 10:449
Purchasing, online 12:590
Purchasing obligation, for renewable energy 18:369
Purchasing power of gay/lesbian tourists 9:640
Purchasing power parities (PPPs) 5:755–756, 7:47
of China 6:790
Pure alexia 6:733
Pure altruism 1:579
Pure direct effect (PDE) 15:76
Pure economics 6:602
Pure Food and Drug Act 17:240
Pure Land Buddhist ideas 20:369
Pure model-driven grouping, computational auditory scene analysis 2:226
Pure motor speech disorder 1:820t
Pure reference, anomia 19:147–148
Pure relationships 10:137
Pure science 8:580, 21:263
Pure structure 23:634
Pure test 19:436
Pure theory of law 13:763, 18:26
Pure word deafness 1:820t
Purely recipient-oriented theorizing about justice 12:944–946
Purification of religion from science 9:106–107
Puritan Revolution 7:650–651
English Revolution 7:650–652
Puritanism 15:577, 17:440, 25:816–817
English Revolution 7:651–652
Islamic fundamentalism 12:742
Purity and pollution 3:223
Purity of blood 20:446
Purkinje cell, eyeblink conditioning 8:637–638
PURPA *see* Public Utilities Regulatory Policies Act (PURPA)
Purpose, in spirituality 23:285–286
Purposeful sampling, in qualitative research 19:687
Purposive behaviorism 13:703
Purposive rationality (Zweckrationalität) 25:466
Purposive sampling, in generalization 9:842
Purposive selection, in generalization 9:841
Purposive social action, unintended consequences of 24:744–747
alternative and emerging directions in theory and research 24:746–747
current refinements and changes in focus over time 24:745–746
intellectual backgrounds 24:744–745
Purposive social intervention, unintended consequences of 24:747–748
Pursuit of self-esteem 21:525
Pursuit rotor 16:845
Push factors, Organon model 7:484
Pushkin, Alexander 21:741
Push–pull theory of migration 15:416
Pu-sik, Kim 11:50
works, *Samguk sagi* 11:50
Putamen 23:252
Putative fear-conditioning circuit 8:871f
Putin, Vladimir 1:940–941, 13:430, 13:432, 20:847–848
Putnam, Hilary 18:810, 19:662–667
concept of bowling alone 11:631
on ethics 19:664, 19:666
on externalism 19:665
on functionalism 19:663–664
on internal realism 19:665–666, 20:22–23
conceptual relativity 19:665–666
on language 19:664
life of 19:662
on logical positivism 19:662–663, 19:666
on mathematics 19:663, 19:665
on mind 19:663–664
Quine and 19:662–663
on science 19:663
on values 19:666
works 19:662–666
The Analytic and the Synthetic 19:662–663
Brain and Behavior 19:663
Hilbert's Tenth Problem 19:663
The Many Faces of Realism 20:21
The Meaning of Meaning 19:664
Realism and Reason 19:664–665
Putnam–Boyd argument 19:663
Puusepp, Ludvig 23:18
Puzzle box studies 24:305
PV *see* Perceived value (PV)
p-values 11:494–495, 19:432
hypothesis testing 11:492–493
in significance tests 21:960
PVC *see* Primary visual cortex (PVC); Primary visual striate cortex (PVC)
PVEM *see* Ecological Green Party of Mexico (PVEM)
PVN *see* Paraventricular nucleus (PVN)
PVO *see* Private voluntary organization (PVO)
PVP *see* Predictive value of a positive test (PVP)
PVQ *see* Portrait Values Questionnaire (PVQ)
PWS *see* Persons who stutter (PWS)
PWT *see* Penn world table (PWT)
PXRF *see* Portable X-ray fluorescence spectrometry (PXRF) machines
PYD *see* Positive youth development (PYD)
Pygmies 3:312
Pyramidal cells 3:339–340

Pyramidal motor system 15:966
Pyramidal tract 3:337–338
Pyrrhonists, objectivity views 17:88
Pythagoras 19:406

Q

Q document 3:545–546
Qafar (language), gender assignment in 9:796–797
Qāḍī 12:746t
QALYs *see* Quality-adjusted life years (QALYs)
QAP *see* Quadratic assignment procedure (QAP)
Qatar, legal system in 15:365
Qazi 23:59–60
QB6 *see* Questionnaire Big Six Scales (QB6)
QCA *see* Qualitative comparative analysis (QCA); Quality control analysis (QCA)
QCAmap (software) 22:545
QEDs *see* Quasiexperimental designs (QEDs)
QGIS (geospatial application) 10:97
qgraph (software) 19:421
Qian, Sima
  on official history 11:48
  works
    Shiji 11:48
Qibla for Salah prayers 21:173
Qing dynasty 2:65, 24:814–815
Qingshigao 11:52
Qiyās 12:746t
QOC (questions, options, and criteria) technique 11:402
QOL *see* Quality-of-life (QOL)
QOP *see* Quantity of processing (QOP)
Qowmiyya nationalism 16:777–778
QR *see* Quantifier Raising (QR)
QRP *see* Questionable research practices (QRP)
Q-sort, in personality assessment 17:841
QTI *see* Questionnaire on Teacher Interaction (QTI)
QT-interval prolongation
  schizophrenia 21:25
QTL *see* Quantitative trait loci (QTL)
QTM *see* Quantity theory of money (QTM)
QTN *see* Quantitative trait nucleotide (QTN)
Quad model 11:707, 11:718
Quaderni Storici 11:42
Quadrals 17:44–45
Quadratic assignment 19:431
Quadratic assignment procedure (QAP) 22:469–470
Quadratic loss, Bayesian experimental designs 8:523
Quadratic model in one variable 17:258
Quadrilateral 13:329
Quadrumvirate of love 14:378–379
Quadruped position, infant motor development 15:975
Quadruple Condition 14:800
Quadruple negation 16:412–413
Quaestio 6:62
Quaker community 25:561
Quakerism 25:559–560
Qualia 18:109, 18:112
Qualia structure 13:945–946
Qualification agreement 1:27
Qualifications 7:215–217, 7:257
  and unemployment 22:286–287
Qualifications mismatch, in labor market 15:597
Qualifying association 22:674
Qualitative analysis
  Strauss on 23:527, 23:530–531
Qualitative approaches, in cognitive mapping 4:82
Qualitative assessments of mother and father involvement in intact families 22:866–867
Qualitative attitudinal research 19:777–778
Qualitative case studies 22:992–993
Qualitative comparative analysis (QCA) 3:210–212, 5:560, 12:647
Qualitative content analysis 22:545
  qualitative developmental psychology 19:672–673
Qualitative data 19:676
Qualitative dependent variables, in cross-national research 19:720
Qualitative developmental psychology 19:668–675
  everyday interactions 19:668, 19:673–674
  future directions 19:674
  methodological premises of 19:670
    postulate of foreigness 19:670
    principle of communication 19:670
    principle of openness 19:670
  new interest in 19:668–669
  paradigm shift 19:668, 19:674
  pioneers of 19:669–670
  procedures for data analysis 19:672–674
    discursive psychology and discourse analysis 19:673
    grounded theory methodology 19:673
    interpretive phenomenological analysis 19:673
    metaphor analysis 19:673
    narrative psychology and narrative analysis 19:673
    performative social science 19:674
    qualitative content analysis 19:672–673
    video analysis 19:673–674
  procedures for data assessment 19:670–674
    autoethnography 19:672
    group discussions and focus groups 19:671
    interviews 19:671
    participant observation and ethnography 19:672
    participatory action research 19:672
    qualitative experiments 19:671
    visual data 19:671–672
    written data 19:671
  sociocultural turn 19:674
  subjective experience 19:668–669, 19:672–673
Qualitative expectations 19:778–779
Qualitative field methods 17:530
Qualitative health research 10:692
  characteristics 10:692–693
  content analysis 10:694
  contributions 10:696
  in era of evidence-based medicine 10:695
  grounded theory 10:692–693, 10:695–696
  medical sociology 10:696
  methods of analysis 10:694–695
  methods of inquiry 10:693–694
    documents 10:694
    group discussions/focus groups 10:693–694
    interviews 10:693
    observations 10:694
  narrative interviews 10:693
  participant observation 10:694
  sampling, sample size, and generalizability 10:695
  semi-structured interviews 10:693
Qualitative methods 7:229, 19:676–684
  of case studies 3:208–212
  childhood studies 3:456
  developments in logics and methodologies 19:679–680
  for educational sociology 7:291–292, 7:294
  institutional emergences of 19:676–678
  Kantian problem of relation between object and selection 19:678
  object of inquiry 19:678–679
  postmodern reflexivity in the shared domain of 19:680–681
  and postmodernism 18:680–681
  qualitative developmental psychology 19:671
  and visual sociology 25:193, 25:196
Qualitative probability 19:37–38
  formal theory 19:37
Qualitative research 7:415, 9:131, 9:274–276, 9:278–279, 10:402–403, 15:618, 19:493, 19:685–689, 22:544–547
  authenticity 19:688
  constructivism 19:685–688
  context 19:686–687
  of cultural capital 5:377
  on definition of situation 6:21–22
  educational outcomes 5:377–378
  in ELSI program 8:44
  ethnography 8:187, 8:190
  expansion of 22:544
  generalization 19:686–688
  holism 19:685, 19:687
  interview 19:685, 19:687
  meaning 19:686–688
  media audiences 2:217–219
  methods of collecting/generating qualitative data 22:544
    focus groups 22:545
    interviews 22:544–545
    naturally occurring data 22:545

observation 22:545
structured methods of data collection 22:545
methods of data analysis 22:545–547
content analysis 22:545
conversation analysis (CA) 22:546
discursive methods (discourse analysis/discursive psychology) 22:546
grounded theory 22:545–546
interpretative phenomenological analysis 22:546
narrative analysis 22:546
reliability 22:547
rhetorical analysis 22:546–547
thematic analysis 22:546
validity 22:547
mixed methods 19:688
objectivity 19:687
online research 19:671
paradigm 19:686, 19:688
reflexivity 19:687
researchers 20:252
sampling 19:687–688
selection 19:687
social work-specific 20:502, 20:504–506
transferability 19:688
validity 19:687–688
on women's offending 5:133–134
Qualitative science studies 21:323
Qualitative sociological research, Schütz on 21:148
Qualitative status, hypothesis tests 11:503
Qualitative studies
of linked lives 14:233–234
micromethods 12:651
Qualitative thought, artistic thinking as 1:230
Qualitative trend analysis methods 9:321–322
Quality
and actors 10:151
appreciation, free time 13:849
of childhood friendship 9:443–444
of Chinese population 2:68
data 6:140–141
of early childhood care and education 6:740–742
of education 7:918–922
of elder care services 15:443–444, 15:441t
of indirect estimation data 6:146–147
of Internet content 12:588
life history theory (LHT) 9:43–44
of media content 15:45–46
of planning 10:626
in residential child care 20:537–538
standards, in urbanization 5:775
standards, program evaluation 19:119
of systematic review 23:908
teacher, in United States 24:78–79
of teaching 7:219, 7:221, 7:227, 21:103, 24:54
of treatment, health care 10:600
uncertainty among artists 2:53–55
Quality (pattern variable) 1:382
Quality assurance
Bologna Process 2:744
consultation–liaison psychiatric services 14:5
in peer review 17:682–684
Quality citizens, in China 21:738–739
Quality control
peer review 17:685–688
and peer review 17:680–684
science 21:252
cycle 21:249–250, 21:252
and statistical probability 23:425
Quality control analysis (QCA) 4:442–443
Quality deterioration, economic regulation 20:167
Quality education 7:262, 7:264–265
Quality of care, in hospitals 11:212–213
Quality of child care 3:392
Quality of experience, and leisure motivation 13:833
Quality of government 19:690–694
different conceptions of 19:690–691
good governance 19:690–691
as absence of corruption 19:691
and democracy 19:692
as rule of law 19:691–692
as government efficiency 19:692–693
as impartiality 19:693
intellectual background to debate 19:690
state capacity 19:692
Quality of inferences, in mixed methods research 15:621–622
Quality of life (QoL) 10:521, 10:524–525, 15:188, 17:279, 24:383, 10:521t
Alzheimer's disease 1:589
and bioarchaeology 2:600–601
reconstructing diet and inferring nutrition 2:600–601
violence and injury 2:601
and cancer 3:73–76
and cardiac disease 3:111–113
and childhood cancer 3:459–460, 3:462
implications 24:403
in Japan and Singapore 5:455
leisure criteria for 13:850–851
measurement 10:658–660
cost-effectiveness evaluations 10:660
decision theory approaches 10:659–660
EQ-5D 10:659
Health Utilities Index 10:659
Nottingham Health Profile 10:658–659
psychometric methods 10:658–659
Quality of Well-being scale 10:660
SF-36 10:658
obsessive-compulsive disorder 17:121
oral health-related 17:279–280
and side effects, of psychopharmacotherapy 19:468
and sleep 22:61
sociological aspects 15:103–104
WHO 10:610
see also Health-related quality of life (HRQOL)
The Quality of Reporting of Meta-analyses *see* QUOROM (The Quality of Reporting of Meta-analyses)
Quality of Well-being (QWB) scale 10:660
Quality teaching, in United States 24:76–77
Quality-adjusted life years (QALYs) 5:57, 10:617–618, 10:659, 17:927, 19:296–298, 19:459, 19:546, 25:16–17
Quality–quantity trade-off, economic growth 7:42–43
QUAL-quan 15:618
Quangos, public sector organizations 19:599, 19:601
Quannei 21:738
QUAN-qual 15:618
Quantification 19:695–699, 19:707, 24:812
delimitation of 19:710–711
in history
culturalist challenge 19:696–697
defined 19:695
developments 19:697
importance of 19:697–698
research contributions 19:695–696
rise 19:695
of information 12:106–108
internal classifications 19:707–708
social sciences 19:700
anthropometry 19:704
authority of calculation 19:702–703
censuses 19:703–704
commensuration 19:702
evidence-based medicine 19:705
historical perspective 19:703–705
measurement 19:700–704
in modern social and behavioral sciences 19:704–705
nineteenth century and rise of statistics 19:703–704
performativity 19:701–702
plethora of units 19:701
producing uniformity 19:701–702
quantities and standards 19:700–702
sampling 19:703–705
standardization 19:701
theories and measures 19:700
trust and impersonality 19:702–703
Thorndike's contributions in 24:306
of warranty period 17:173–174
Quantified Self social movement 21:190
Quantifier lowering rules 16:6
Quantifier Raising (QR) 16:6
Quantifiers 19:707–711
classification of 19:707
continuous 19:708–709
count 19:707
decreasing 19:708
discontinuous 19:708–709
existential 19:707–708
as grammatical category 19:710
increasing 19:708
mass 19:707
midrange 19:708
nominal 19:708
scopal 19:709–710
scope 19:709–710
strong 19:708
universal 19:707–708

Quantifiers (*continued*)
use of term 19:707
verbal 19:708
weak 19:708
Quantile function (QF) 11:886
Quantile probability plot of RT distributions 6:399–400
Quantile regression 19:712–718
computational aspects of 19:716
example 19:713–714
extensions 19:717
future developments 19:717
interpretation of 19:714–716
quantile treatment effects 19:714–715
robustness 19:716
transformation equivariance 19:715
optimization 19:712–713
quantiles 19:712–713
ranks 19:712–713
statistical inference for 19:716–717
uses 19:712
Quantiles 19:712–713
confidence intervals for 17:292
treatment effects 19:714–715
Quantitative analysis
of manifestos 18:417
Quantitative assessments of mother and father involvement in intact families 22:866–867
Quantitative behavioral genetics 6:320–321, 12:297
Quantitative cohort analysis, and social mobility 22:427
Quantitative crime data
analysis 19:734
history 19:733
longitudinal and spatial analysis 19:734–735
sources of 19:733–734
Quantitative cross-national research methods 19:719–724
challenges 19:719–720
endogeneity problem 19:723
problem of independence 19:721–723
problem of selection bias 19:720–721
qualitative and limited dependent variables 19:720
Quantitative expansion 7:211
Quantitative experimental methods 22:548
Quantitative genetics 2:450–452, 6:320, 8:471
personality traits 17:921
Quantitative linguistics 19:725–732
hypothesis-generating methods 19:730–731
desiderata 19:731
exploratory methods in common use 19:730–731
hypothesis-testing methods 19:725–730
desiderata 19:728–730
goodness-of-fit tests in common use 19:726
tests for independence/differences in common use 19:726–728
Quantitative memory models 15:160–165
associations 15:163–164
buffer model 15:162–163
encoding 15:162–163
H. Ebbinghaus 15:160–162, 15:164–165
human memory 15:160–162
information-processing view 15:161
recent developments 15:161–164
applications to data 15:164
MINERVA 15:163
REM 15:163–164
TODAM 15:164
retrieval 15:162–164
storage 15:162–164
Quantitative methods 6:266, 19:695–697
causal mechanisms in 14:876
Tilly on 14:876
and crime 19:733–735
for educational sociology 7:291–293
Quantitative models
for social change 22:197
Quantitative moral psychology 14:869
Quantitative nonexperimental methods 22:548
Quantitative phenomenology 21:605
Quantitative physics
development and measurement theory 14:868–870
extensive magnitudes and real numbers 14:868
intensive magnitudes and concept of continuous quantity 14:868–870
measurement of psychological attributes 14:870
Quantitative probability 19:37–38
formal theory 19:37
interpretations 19:38–41
Fisher–Kyburg interpretation 19:41
logical probability 19:39–40
objective probability 19:40–41
subjective probability 19:38–39
Quantitative reasoning 3:920
Quantitative research 5:836, 5:840–841, 7:415, 20:252, 22:547–551
additional comments 22:550–551
error 22:550
missing data 22:551
variable coding 22:550–551
of cultural capital 5:376–377
of educational outcomes 5:377
in ELSI program 8:44
methods of collecting/generating data 22:547–548
ambulatory assessment 22:548
diary methods 22:548
experimental 22:547
internet research 22:548
nonexperimental 22:548
survey 22:548
trace measures 22:548
methods of data analysis 22:548–551
causal modeling 22:548–550
levels of analysis 22:550
mediated moderation 22:549–550
mediation model 22:549
moderated mediation 22:549–550
moderation 22:549
Quantitative revolution 6:408, 15:501–502
geography 10:84–85
Quantitative science 21:322
measurement theory and issues external to 14:870–871
philosophy of mathematics 14:870–871
philosophy of science 14:871
Quantitative studies, of linked lives 14:232–233
Quantitative surveys 5:844, 22:993
methodologies 5:844
sampling techniques 5:844
Quantitative trait loci (QTL) 12:300–301, 12:305
Quantitative trait nucleotide (QTN) 8:719
association testing 8:721, 8:721f
DNA sequence variant at 8:719
-specific heritability 8:719
Quantitative trend analysis 9:317–319
Quantities
economic statistics 5:758
life history theory (LHT) 9:43–44
linear hypothesis 14:150–152
quantification 19:700–702
Quantity measures, for health behaviors 10:588–589
Quantity of capital 3:92
Quantity of processing (QOP) 10:706–707
Quantity theory of money (QTM) 13:16, 14:406–409, 18:357
Quantity–quality (Q–Q) fertility model 24:708–709
Quantum (fertility) 9:10–12, 9:19, 15:397–398
QUANTUM (quantifier organization) 19:695
Quantum chemistry 18:103
Quantum decision theory 2:341
Quantum logics 16:925
Quantum mechanics 19:24–25, 19:27
Copenhagen Interpretation 17:89–90
and objectivity 17:89–90
Quantum theory 18:103–105
Quarantelli, E.L. 17:370
Quarreling 12:839
Quarterly Journal of Biology 22:903
Quasi rent 14:622
Quasi-actors 15:312
Quasi-analysis 14:316
Quasi-control subjects 8:574
Quasi-erga-omnes effect 4:700
Quasiexperimental designs (QEDs) 19:742–746
observational studies 19:742–743
potential outcomes model 19:742–743
regression adjustment in 4:841
types 19:743–745
comparative interrupted time series designs 19:744–745
nonequivalent comparison group 19:745
regression discontinuity designs 19:743–744
validity of 19:742–743

Quasi-experiments 3:69, 7:729, 8:345–347, 9:129, 19:140, 19:736–741
in causal analysis 3:247–248
in epidemiology 7:827
versus field trials 19:736
regression-discontinuity design 20:152, 20:154–155
research/studies
class size 3:752–753
mother–child communication 15:877
Quasi-markets 10:598–599
in England 4:345
Quasi-Markov simplex model 20:245–246, 20:246f
Quasineeds, and motivation 15:938
Quasi–Newton algorithms 9:848
Quasireligion 21:382
Quasireligious practice 20:408
Quasi-states 23:366
Quasi-stationarity, population 18:579–580
QUASSH (quantitative social science history) 19:697
Quaternary circuit 3:580
Quayle, Oliver 18:461
Quebec
educational system 12:280
identity movements in 11:544
Quebec Act 1:610–611
Quebec Declaration 7:105–106
Québec genome 1:760–761
Quechua 23:44, 23:46–47
Queens, drag/faux 13:875
Queer 13:875, 21:808–811
and lesbian feminism 8:946
sexual orientation 21:788–789
Queer anarchism 19:747–751
history 19:747–749
identity/politics 19:749–750
compassionate self-inquiry 19:750
equality 19:749
insider/outsider 19:749
representation 19:749–750
state of sexuality 19:750
Queer anthropology 19:752–758
future and post-queering anthropology 19:757–758
histories of sexualities and queer paradoxes 19:752–753
and inequality 11:897
kinship, sociality, and queering community 19:754–755
language and gender 19:755–756
reflexivities and methodologies 19:753–754
transnationalities and new sexual subjectivities 19:756–757
Queer geography 19:759–764, 21:809–810
body 19:760
city 19:761
gay neighborhood 19:763
gender 19:759–761
home 19:760–761
migration 19:760, 19:762
mobilities 19:759, 19:762
nation-state 19:761–762
place-making 19:762–763
scale 19:760–762
sexuality 19:759–762
space 19:759–761, 19:763
Queer identity 19:766
Queer legal theory 9:734
Queer Nation movement 19:754, 24:552
Queer parenting 19:754
Queer politics 24:552
Queer repositioning 19:766–767
Queer spaces 21:809–810
Queer theories 19:747, 21:730, 21:769, 21:809–810, 21:851, 24:552, 25:629
and feminist legal theory 8:919–920, 8:922
and gay/lesbian activism in Latin America 13:862
and gender archaeology 9:766
and homophobia/heterosexism 10:792
and intersectionality 19:765–770
concepts development 19:765–766
criminalization 19:766
gay 19:765–767
heterosexism 19:765–770
homophobia 19:766–768
intersectional queerness 19:767–768
lesbian 19:765–769
marginalization 19:765–768
oppression 19:766–768
queer identity 19:766
queer repositioning 19:766–767
racialization 19:765–768
sexual orientation 19:765–766
transgender 19:765
transnational 19:768–769
white privilege 19:768
Queer turn 19:754
Queer Undocumented Immigrant Project (QUIP) 9:630
Queerball Strikes Back! event 9:640–641
Queering 21:752
Querelles des anciens et des modernes 15:700
Quesnay, François 3:90, 16:228
Quest orientation 20:381
Question answering process, for attitude measurement 2:178–179
attitude scale formats 2:179
evaluating attitude objects 2:178–179
implications 2:179
reducing socially desirable responding 2:179–180
Question asking (communication skill) 4:289, 21:272
Questionable research practices (QRP) 21:314
Questionnaire Big Six Scales (QB6) 11:172
Questionnaire of Sex-Role Stereotypes 21:695
Questionnaire on Teacher Interaction (QTI) 3:825–826
Questionnaires
for census 3:296–300
cognitive approaches 19:771–775
asking questions about behavior 19:772–774
autobiographical memory 19:773
behavioral reports 19:775
calendar methods 19:774
cognitive interviewing 19:775
estimation strategies 19:774–775
event dating 19:773–774
event history calendars 19:774
frequency scales 19:771–772, 19:774–775
pretesting 19:775
question comprehension 19:771–772
question order effects 19:772
rating scales 19:772
reconstructive memory 19:773–774
respondents' tasks 19:771–772
response order effects 19:772
self-report 19:773–774
social desirability 19:771–772
collection of subjective expectations data through 19:776–780
decision making 19:776, 19:779
elicitation of probabilistic expectations 19:778
frontiers of research 19:779
inference on expectations in economics 19:777
qualitative attitudinal research 19:777–778
qualitative *versus* probabilistic expectations 19:778–779
subjective concepts and measures 19:776–777
subjective probabilities 19:776–779
survey research 19:777
data collection, forced-choice personality assessment 17:840–841, 17:840t
designers 23:764
emotion assessment through 7:521
empirical social research (ESR) 7:563
ethnography 8:187–188
five-factor model of personality 9:231
for health behaviors 10:588–590
personality studies 9:250
student achievement 5:347
Questions 23:764
design 16:817–818
explanation-seeking 8:593–594
in problem selection
descriptive 19:53
descriptive to explanatory questions 19:52–53
erroneous and well-stated 19:53–54
superficial to deeper explanatory questions 19:53
undertheorized 19:54
in questionnaires
comprehension 19:771–772
order effects 19:772
social 22:553–555
survey
absolute frequencies 23:767, 23:769
category labels 23:767–768
characteristics 23:765–768
closed questions 23:767
conceptualization and operationalization of target object 23:766

Questions (*continued*)
decomposition, object 23:766
event questions 23:764–767, 23:765f
form 23:767
frequency questions 23:766–767
guidelines for writing 23:768–769
middle category 23:764
no opinion filter 23:766–767
number of categories 23:768
open questions 23:767
parts of 23:764–765
rating scales 23:764–768, 23:766f
relative frequencies 23:767, 23:769
response categories 23:767–768
standardized interviews 23:767
structure 23:766–767
topic, type, and response dimension 23:765–766
understanding 19:771–772
context 19:771
response alternatives 19:771–772
wordings
direct democracy 25:294–295
for sexuality 15:317
Questions, options, and criteria *see* QOC (questions, options, and criteria)
Quetelet, Adolphe 6:171–172, 19:781–783
on applied social research 1:851–852
average man concept of 19:781–782, 23:423–424
on constant mortality 19:782
distinction between *causes constantes* and *causes accidentelles* 19:782–783
human-ontogenetic development as a lifelong process 19:783
life and career of 19:781
notion of *l'homme moyen* 19:781–782
on probability theory and astronomy 19:781
psychology, contributions to 19:783
ratio estimation approach of 20:875–876
sociology, contributions to 19:782–783
statistics and demography, contributions to 19:781–782
work on demographic subjects 19:781
Quételet, Adolphe 13:638
Quetiapine 19:201
Queues 19:784–789
Bayesian approach of 19:786–788
discipline 19:786
equilibrium 19:786
notations 19:786
statistical analysis 19:786–788
systems 19:785–786
Quid pro quo harassment 21:721–723
Quiet violence 4:358
Quijo 23:45
Quinceañera 24:579
Quine, Willard Van Orman 3:263–264, 7:568–569, 8:576, 18:626, 19:662–663
on metaphors 15:291
on scientific pluralism 25:601
on tacit knowledge 13:75–76
on wisdom 25:601
Quine, W.V.O. 6:593
Quinlan, A., on narrative networks 16:215–216
Quinlan, E., on narrative networks 16:215–216
Quinn, Naomi 13:72
Quinn and Rohrbaugh Competing Values Framework, of school effectiveness 21:83
Quinney, Richard 13:511, 22:386–387
QUIP *see* Queer Undocumented Immigrant Project (QUIP)
Quis custodiet 21:249–251
Qunxue 24:584–585
Quod principi placuit legis habet vigorem 1:7–9
QUOROM (The Quality of Reporting of Meta-analyses) 15:269
Quota-referential system *see* Single-transferable vote
Quotas 1:248
Qur'an 9:266, 11:48, 11:62–63, 12:746–747, 12:749, 12:752, 12:754–757, 20:231, 20:335, 20:398–399, 21:173–174
Qur'anic exegesis (*tafsir*) 11:61
Quraysh clan 12:731
QWB *see* Quality of Well-being (QWB) scale
Qwelane, John 21:746

R

R (statistical analysis language) 23:379
R UV UGLY campaign 8:458–459
RA *see* Retrograde amnesia (RA)
Rabbinic Judaism 11:67, 12:870
Rabbinic literature 11:67–68, 12:876
Rabbinics 12:876
Rabelais, on humor 11:416
Rabies in raccoons 1:369–370
Rabinow, Paul 1:749
works
French Modern 1:749
Marking Time 1:749
Race 5:151–152, 5:182, 6:8, 6:10, 7:720–724, 8:137, 8:139, 8:233–236, 18:597, 18:599–600, 24:912
and academic motivation 19:790–795
and adoption 1:137, 1:146
affirmative action 1:262–268
and bear subculture 13:891
Blacks 5:154–156
see also Blacks
blending ideologies in South America 23:45
in Canada 8:152–156
historical perspective of 8:152–153
in Caribbean 1:639, 1:642
composition of North American population by 18:553–558
conflation of 9:903
and crime 5:182–188, 19:837–844
attribution theory and racial stereotyping 19:843
criminal justice 19:841, 19:843–844
differential offending 19:839–841
focal concerns perspective 19:843
general strain theory 19:839
juvenile justice 19:838, 19:842–844
liberation hypothesis 19:842–843
lifecourse perspectives 19:840–841
minority/racial group power threat thesis 19:841–842
official statistics 19:837–839, 19:837t–838t
selection bias 19:841–843
social and self-control theories 19:839
social disorganization theories 19:839–840
subcultural theories 19:840
symbolic threat hypothesis 19:842
theory of inequality and social control 19:842
urbanism and formal rationality 19:841
and cultural capital 5:378
as cultural construction 8:138
culturally appropriate interventions 5:517–518
definition of 19:812–813, 19:857
and development geography 6:315–317
differences in intelligence 12:311
discrimination and disparity 5:155–156
in Canada 5:155
possible explanations for overrepresentations 5:155–156
in United States 5:155
division of labor 25:341–343
and divorce 6:607
and drag culture 13:877–878
and engineering 7:646–647
and essentialism 8:14
and ethnicity 5:182–188, 8:130–131, 8:134, 8:137–140
eugenics in the Americas 8:246–247
face perception 8:655
Freyre on 9:420–422
future aspects 19:831
genetic aspects 19:825–832
and genetics 6:443
geography of 19:801–806
and eugenics 19:802–804
late twentieth century 19:804–805
medicalization of 19:801
twenty-first century 19:804–805
and health 15:99–100
data analysis challenges 19:813
disparate health services and treatment 19:815–817
evidence-based interventions 19:815
health-enhancing resources 19:817
healthy immigrant phenomenon 19:815
institutional discrimination 19:812, 19:816
Latino health paradox 19:813–817
life expectancy 19:812–815, 19:817, 19:813t
Preston curve 19:813
prevalence of disease 19:812, 19:818
resources and acculturation 19:817–818
socioeconomic status 19:814–815
variations 19:813–814

Hispanics 5:155
and homicide 11:165
and inequality 11:896–897
and intersectionality 12:657–667
and interviewing 5:852–853
and intimate partner violence 12:686–687
and law 19:807–811
affirmative action 19:808
antiessentialist and cosmopolitan conceptions of racial identity 19:810
antiracist legal reform 19:809–810
antidiscrimination 19:809–810
antisubordination 19:810
cultural discrimination 19:810
implicit bias 19:810
integrationism and separatism 19:809
criminal justice 19:810
definition 19:807
Jim Crow laws 19:808–809
law as cause of racism 19:807
mainstream *versus* critical scholarship 19:807
racial categorization 19:807–809
legal codification of racial categories 19:808
legal enforcement of racial hierarchy 19:808–809
de facto racial subordination 19:808–809
de jure racial subordination 19:808
ontological status of racial categories 19:808
and marital quality 14:513–515
in medicine 19:830–831
microevolution 15:341
and migration 11:651
in military 15:514
and mortality crossover 15:849–850
and new genetics 9:1018
'new' genetics and 16:759–763
in North America 17:23
and organizational diversity 4:738–739, 4:741–742, 6:595–596, 6:598–599
and postcolonial geography 18:644–645
and prosocial behavior 19:231
and residential segregation 20:540–542
and segregation 21:407–409
and social class 22:342
as social construct 22:821
and social networks 5:560
sociology 9:460
and spatial mismatch 23:157–160
-specific homicide offending trends 5:196
studies 5:151–152
and suicide 23:686
symbolic boundaries 23:850–851, 23:853
theory, and stigma 23:461–462
and urban planning 19:820–824
urban sociology 24:926–927, 24:930–931
use of term 8:137, 8:139
variation in intelligence 19:830–831
visual turn 25:207–208
in voting 25:305
Wilson on 25:585–586
Race and ethnic politics, competition theories 8:125–126
Race and ethnicity 11:657, 12:144
and cyberbullying 5:641
Race- and gender-based stereotypes 15:25
Race and racism, microevolution 15:341
Race betterment 8:247
Race defilement 11:152
Race discrimination 9:731–732
Race equity 7:927
Race hygiene 16:14–15
Race identity 19:833–836
Race jurisprudence 5:268
Race purity 19:855
Race records 3:800
Race relations 5:371–372
Race Relations Act 1976 5:517
Race Relations Amendment Act 2000 5:517
Race relations training 17:392
Race riots 24:949–950
Race suicide 8:247
Race to the Top competitive grant program 7:267–268
Race-conscious actions 5:266
Rachid O., works, *L'enfant* ébloui 21:753
Rachmanism 10:48
Racial and ethnic difference
and adopted children 7:442
Racial and ethnic identity
East Asian studies 6:785–786
Racial and ethnic relations 19:845–846
Racial annihilation 16:260
Racial catalogs 19:826
Racial categories 20:884
legal codification of 19:808
ontological status of 19:808
Racial centrality 19:835
Racial colonialism and colonial capitalism 25:571
Racial composition, and homicide 11:164
Racial conflict 4:589, 19:846–848
Racial covenants 16:452
Racial differences in intelligence test scores 5:323–324
Racial discrimination 4:155, 19:833–834, 23:159
acculturation and 1:73
and adolescent health 1:125
antisemitism 1:786
conceptual complexities 6:542–543
coping 19:792
direct discrimination 6:542–544
direct/indirect distinction 6:542–543
empirical ambiguities 6:543
indirect discrimination 6:542–544
inequalities 6:542–543
job promotion 6:545
prejudice 6:542–545
psychological effects of 6:544
remedies for 6:544–545
and spatial mismatch 23:157, 23:159–160
stereotype threat 6:544
in United States and western Europe 6:543–544
Racial disparities
and poverty 24:730–731
in sentencing 21:628–630
in wrongful convictions 25:788–790
Racial empire 16:300
Racial equality 6:219
Racial formation 19:847
Racial formation theory 19:801, 19:804
Racial group consciousness 19:833
Racial group identification 19:833
Racial hierarchy, legal enforcement of 19:808–809
Racial identities attitudes scale 19:834
Racial identity theory 5:66, 19:833
antiessentialist conceptions of 19:810
cosmopolitan conceptions of 19:810
development of 19:833–835
future of 19:835
group consciousness 19:833
group identification 19:833
identity politics 19:833
and individual identity 19:833
politics 19:833
white identity 19:835
Racial ideology 19:835, 19:846–847
Racial inequality, and homicide 11:163–164
Racial issues, issue evolution 12:766–767
Racial minorities, and education 1:729, 1:731–732
Racial nullification 13:749
Racial polarization, and inequality 7:895, 7:899–900
Racial recursion 9:1018
Racial regard 19:835
Racial relations 19:845–851
assimilation approaches 19:846
continuing racial oppression 19:848
global racial order 19:849–850
power-conflict approaches 19:846–848
racial and ethnic groups 19:845–846
racism
in Europe 19:848–849
in former European empires 19:849
Racial residential segregation, and homicide 11:164
Racial salience 19:835
Racial segregation 20:540
African 1:288
Racial segregation laws 13:820
Racial stereotyping, media, effects of 15:38
Racial stratification 19:847, 22:710
Racial unconscious 4:152
Racial violence 19:804–805
Racial–cultural identity development 5:65–66
Racialization 2:85, 7:118, 19:860, 25:570
intersectional queer theory 19:765–768
and migration 15:427, 15:430–431
and whiteness/white supremacy 25:570–571, 25:573
Racialized groups
social work with 22:821–826
cultural competence and cultural sensitivity 22:822–823
culture 22:821–822
implications for social work practice 22:824–825

Racialized groups (*continued*)
institutionalized oppression 22:822–823
internalized oppression 22:823–824
marginalization 22:823–824
oppression 22:823–824
othering 22:821
race as social construct 22:821
racialized identities 22:821–825
social work ethics 22:823
systemic oppression 22:824
Racialized identities 22:821–825
Racialized (in)justice 12:959–960
Racine, Ron 13:32
Racism 1:786–787, 5:266, 7:724, 8:240–245, 11:151, 18:411–414, 19:864–869, 20:546–547, 22:55–57, 23:366, 25:797
ambivalent 19:865
antisemitism 19:852–854
in Australia 2:239
authoritarian personality 19:866
aversive 19:865
categorization 19:866–867
challenges 19:862
civil rights 19:855
and colonialism 4:218
contemporary approaches to 19:858–862
critical discursive research 19:867
cultural 19:864–865, 25:796–797
culturally appropriate interventions 5:517–518
professional application of cultural appropriateness 5:518–519
definition 19:864–865
definition of 19:857–858
and desegregation 6:218–227
disconnection of 1:787
domination 19:852–855
embodied racism 19:867–868
environmental 10:387–388
and ethnicity 8:150
in Europe 19:848–849
in France 19:848–849
genocide 10:15, 10:18–21
genocide, history of 10:10
geography of 19:801–806
late twentieth century 19:804–805
modern racial discourse 19:801–802
twenty-first century 19:804–805
historiography of 19:852–853
history of 19:852–856
definition 19:853
emergence 19:853–854
end 19:855
rise of modern racism 19:854–855
in twentieth century 19:855
implicit bias 19:859–860
inequality 19:853, 19:855
institutional 19:864–865
institutional racism 19:860–861
integration of multiple perspectives 19:868
intellectual history 19:858
Jews 19:852, 19:854–855
Lévi-Strauss on 13:904, 13:908
Memmi on 15:136–137
microevolution 15:341
and migration 15:423–424
new 19:865
new racisms 19:859
persecution 19:852, 19:855
personality theories 19:866
in political psychology 18:431
versus prejudice 19:864–865
and public opinion 19:564–566
racial discourse 19:868
realistic group conflict theory 19:866–867
religion 19:852–854
religious intolerance 19:852–853
responses to 19:861–862
right-wing authoritarianism 19:866
in Roma 20:757
slavery 19:852–854
social cognitive theories 19:866
social identity theory 19:866–867
as social process 19:860
in social work 1:786–787
sociology 9:460
sociology of 19:857–863
in South Africa 19:849
and sport 23:314–316
stereotyping 19:865–866
stigma 23:462–463
toward Latino gays and bisexuals 9:629–630
in twenty-first century 19:796–800
beyond borders of United States 19:798
dismantling structural racism 19:799
ethnicity 19:796–799
new engagements of 19:798–799
new globalization research 19:799
racial formation 19:797–798
racial formation paradigm 19:796–797
three hundred years of 19:798
in United States 19:848
as white privilege 19:868
and whiteness/white supremacy 25:569–573
xenophobia 19:852–853
Racist Nazi laws 12:933
RADAR console (signal detection) 21:952
RADAR program (health care) 12:681–682
Radbruch, Gustav 12:933, 13:762–763
Radbruch formula 13:480–482
Radcliffe-Brown, Alfred Reginald 1:680–682, 7:145, 7:376, 8:289–291, 9:672
on anthropology 1:746
on functionalism 9:533
on humor 11:416–417
on lineage systems 9:537
structural functionalism 1:767, 8:290–291, 9:537
Radial arm maze (RAM) 16:349–352, 16:350f
Radial centrality measures 16:533–535
degree centrality 16:533–535
distance-based centrality measures 16:533
measures involving reflection/feedback 16:533–534
Radial glial cells (RGCs) 2:821–823
Radial set model, of cognitive semantics 13:933
Radiant energy 7:626
Radiation dating methods 3:566–567
archaeomagnetic dating 3:566–567
Radiation protection, Muller's concern for 16:15–16
Radical approach
emotion, history of the concept 7:472
Radical behaviorism, Skinner 22:26–27
Radical community work 4:379
Radical constitutionalism 13:411–412
in Latin America 13:411–412
Radical construal of reflective equilibrium 20:78–79
Radical constructionism 22:2
and affect 1:237
Radical constructivist studies of science and technology 24:135
Radical ecology, sustainable development 23:813–814
Radical embodied cognition (REC) 7:420, 7:422
Radical enlightenment 7:660–661
Radical evil, Arendt on 1:963
Radical feminism 8:919, 8:921–922
and feminist criminology 8:912–914
Radical geography 10:85
of places 18:149–150
Radical health movements, lesbian 15:110–111
Radical individualism, and social stratification 22:342–343
Radical innovation 12:147, 23:512
Radical Islamism 12:743
Radical legal autonomy 13:555
Radical lesbian feminism
ecologically oriented 8:942–948
critique of liberal feminism 8:943–945
future 8:943
heteropatriarchy 8:942–947
oppression 8:942–947
queer 8:946
radicalized socialist lesbian ecofeminism 8:945–947
revolution 8:942–943, 8:945–947
sexual identity 8:942–947
sexuality 8:942–944, 8:946–947
for twenty-first century world 8:942
Radical liberalizations 24:792
Radical microsociology 22:335
Radical monopoly 11:585
Radical movement in geography 14:674
transformation 14:674
Radical nationalist movements 16:309
Radical organizational theory
control 17:342–343
centrality 17:344–345
critiques 17:343–345
debates and disagreements 17:342–343
external to internal 17:343–344
new challenges 17:343–345
responses 17:343–345
scope of workplace 17:344–345
Radical psychology 5:264

Radical reflexivity 20:94–95
Radical revolutions, behavior of siblings during 21:925–926
Radical right 18:410–414, 25:304–306
  populist 18:610
Radical right gender gap 25:304–305
Radical skepticism 20:227, 22:988
Radical social constructivism 22:231–232
  and ontology 22:231
  and the social 22:231–232
Radical social work (RSW) 5:287, 14:35, 14:663–668, 19:870–876, 20:83
  emancipatory approaches 19:870–874
  exploring social work practice 19:873–874
  human rights 19:873
  institutionalization 19:872
  lost in 1990s 19:872
  Marxism 19:870–871
  neoliberalism 19:870–872, 19:874
  professionalization 19:870–872
  progressive politics 19:870–871
  reclaiming twenty-first century 19:873
  social justice 19:870, 19:873–874
  social movements 19:871, 19:874
  solidarity 19:871–872, 19:874–875
  theoretical evolution 19:873
  tradition from 1970s 19:871–872
Radical-democratic pluralism 6:93
Radicalism 2:525–526, 12:761, 20:96, 20:397
  Chomsky on 3:527, 3:529
  English Revolution 7:651–654
  and ideologies 11:573
  in Soviet Union 20:820–821
Radical/'red' Tacitism 3:772
Radin, Paul, on intellectuals 13:72
Radio 19:877
  broadcasting regulation 2:859–863
  commercial broadcasting 19:877–878, 19:880–881
  community 19:877–878, 19:880–881
  in first decades of twenty-first century 19:881
  history and political economy 19:878
  international broadcasting 19:879
  intimacy 19:877–878, 19:881
  liberalization, deregulation, and consolidation 19:880–881
  listeners 19:877–878, 19:880–881
  market regulation 19:880–881
  mobility 19:877, 19:879
  modernity 19:877–878
  nationalism 19:878–879
  orality 19:877–878
  public monopoly broadcasting 19:877–880
  1920s and 1930s 19:878
  1940s and 1950s 19:879
  1960s and 1970s 19:879–880
  1980s and 1990s 19:880–881
  talk shows 19:881
  technology and cultural form 19:877–878
  and television 19:877, 19:879–881
Radio Act 1927 (US) 19:878
Radio Act 1996 (Sweden) 19:880
Radio audience measurement 2:212
Radio broadcasting 24:164
Radio casting 12:492
Radio communication 15:631
Radio Corporation of America (RCA) 24:169
Radio-based wireless devices 15:629
Radiocarbon dating 1:905, 3:565–566, 7:695–696
  AMS and 3:566
  calibration 3:565–566
  method 3:565
Radio-Keith-Orpheum (RKO) 9:156–157
Radiolabeled ligands
  molecular imaging with 14:440
    general aspects 14:440
    principles 14:440
Radszuweit, Friedrich 9:646–648
Radway, Janice, works, *Reading the Romance: Women, Patriarchy, and Popular Literature* 2:218–219
Radzinowicz, Leon 5:241
Rae, John, on division of labor 6:603–604
Raelians 16:805
RAF *see* Rote Armee Fraktion (RAF)
Raffles, Thomas Stamford
  works
    Genealog of (*Melaka*) *Kings* (*Sulalat'us-Salatin*) 11:83
    History of Java 11:84
RAFI *see* Rural Advancement Foundation International (RAFI)
Ragin, Charles, on qualitative comparative analysis 5:560
Rahe, Paul A., works, *Against Throne and Altar* 18:487
Rahman, Ziaur 7:194
Raiffa, Howard, works, *The Art and Science of Negotiation* 16:431
Railway industry 21:186
Rainbow nation 21:747
Rainiero (sainthood candidate) 15:579
Raising Boys' Achievement Project 9:685
Rajneeshee movement *see* Osho sannyasins
Rakhi 9:138
RAM *see* Radial arm maze (RAM); Realistic accuracy model (RAM)
Rama IV (Mongkut), King 2:891–892
Ramadan 12:730
Ramadan War 25:423
Ramakrishna (Hindu saint) 20:374
Ramifications, organizations 2:770
Ramsey–Neumann (RN) cardinal 19:899
RAN *see* Rapid automatized naming (RAN)
Rancière, Jacques 25:212
Rand, Ayn
  objectivism 14:42
  works
    Atlas Shrugged 14:42
    The Fountainhead 14:42
RAND Corporation, USA 19:576, 21:971–973, 22:656, 24:295
RAND Database of Worldwide Terrorism Incidents (RDWTI) 5:837
RAND Health Insurance Experiment 8:563–564, 10:642
Randkulturen 3:760
Random assignments 5:322, 9:128, 19:883–886, 20:153, 22:295
  acceptance of 19:883
  causal inference 19:883
  conditions for 19:884
  objections to 19:883–884
  purpose of 19:883
  social programs evaluation 19:884
  to treatments 12:453
  working toward 19:884–885
Random coefficient models 16:50, 23:381–382, 23:385
  ecological regression with 6:869
  with Gaussian errors 23:382
  period and cohort analysis in demography 17:732
Random digit dialing (RDD) 16:919–920, 18:461–462
Random effects 16:49–50, 23:204, 23:383
  generalized linear mixed models 9:850–851
  hierarchy models 10:836–840
  infant and child survival 16:53–54
  linear mixed models 14:162–168
  meta-analysis 23:912
  in quantitative linguistics 19:727
  variance component 15:269
Random error 24:639–640, 24:642
Random experiments 3:785
Random field theory 9:529
Random intercepts model 16:49
  period and cohort analysis in demography 17:732
Random match probability (RMP) 9:956–957
Random measurement errors 11:750–751
Random networks 1:368
Random processes 8:471
Random sampling 5:321, 9:839–841, 9:844, 20:876, 23:423, 23:426, 23:187f
  development, United States 20:877–878
  stratified 20:877
Random slopes model 16:49
Random treatment strategy 14:341
  in time-varying treatments 14:341
Random unit effects model, for crossover designs 5:353
Random utility maximization 19:441
Random utility models 5:901–905
  choice probability 5:901–905
  choice tree model 5:904–905
  context-dependent models of response time and choice 5:904–905
  context-free linear accumulator models 5:903–904
    additive drift rate variability 5:904
    multiplicative drift rate variability 5:903–904
    relation of multiplicative LBA models with no start point variability and Fréchet drift rate variability to Luce (MNL) models 5:904

Random utility models (*continued*)
generalized stable survival functions 5:902–903
hazard rate 5:902
independence of option chosen from time of choice 5:902
2N-ary choice tree model 5:904–905
representation by independent random variables 5:902
Random utility representation 5:901
Random utility theory 1:66, 24:618
conjoint analysis 14:866
Random variables
majorization/stochastic orders 14:449–452
Monte Carlo methods 15:758–762
probability 19:34
Random walk M–H chain 15:764
Random walk processes 6:395–401, 11:705–706
historical foundations 6:395–396
Random walk tracking mechanism 14:589
Random-digit-dial telephone surveys 13:386
Random-effects statistical procedures
meta-analysis 15:273–274
Randomization 3:258, 8:567, 8:569, 23:425
causal analysis 3:247–248
levels in population studies 8:562
Randomization bias 19:738–739
Randomization tests 19:431
Randomized block design 8:538
Randomized clinical trial (RCT) 19:736, 20:519
placebo-controlled trials 18:164–167, 18:169
Randomized controlled trials (RCTs) 2:533–534, 4:459, 7:229, 8:345–347, 8:541–545, 9:129–130, 11:462, 17:281, 19:140, 19:742, 21:499, 22:298
access to information about 8:544
biomedical sciences 2:665
definition and rationale 8:541
effectiveness 8:542–544
efficacy 8:542–544
for efficacy trials 7:230
ethics and law 8:543–544
evidence-based medicine 8:328–331
evidence-based practice 8:332–334, 8:336
evidence-based psychiatry 8:339–342
examples 8:542
of family planning 8:788
history 8:541–542
ideal 8:560
implementation of 8:542–543
institutional capacity for 8:544
mental health 15:190–191
psychological treatments 19:386–390
random assignment 8:541–544
unbiased estimates of effect 8:541
uses 7:229
Randomized designs, completely randomized designs 8:537–538
Randomized experiments 12:453–454, 17:108, 17:110, 19:736–737
causal inference 3:257–258
regression adjustment in 4:840–841
Randomized field trials, privacy/confidentiality in social research 18:943–944
Randomized phase start-point designs 21:988
Randomized phase-order designs 21:988
Randomized response techniques for sensitive questions 16:920–921
Randomized rollout designs 9:130
Randomized steepest ascent 23:395
Randomized trials, compliance 4:445–455
Randomness
probability 19:33
in stochastic optimization 23:473, 23:479
Random-walk/drift diffusion models 14:804
Range of the good 14:545
Rangeland enclosures 17:581–582
Rank
of gentry 1:965
system of 1:967
Rank, Otto 9:416, 12:672
Rank ordering 5:610
Rank societies 11:895–896
Rank test, asymptotic distribution of 24:328
Rank-based nonparametric methods 16:891–897
location alternatives 16:891–896
distribution-free statistic 16:891–895
efficiencies for location procedures 16:896
k-sample problem 16:893–895
multivariate two-sample location problem 16:895–896
two-sample location problem 16:892–893
Ranke, Leopold von 1:910, 11:39, 11:51, 11:94–95, 11:121–122, 11:126, 19:676, 25:290–292
on classical German historism 25:292
on empiricism 25:292
on French hegemony 25:290
on historical theology 25:291–292
on historiography 25:290–292
on modern German history 25:292
on Prussian state conservatism 25:290–292
on Thukydides 25:290
works
Geschichte der romanischen und germanischen Völker von 25:290–291
Historischpolitische Zeitschrift 25:290
Preußische Geschichte 25:290–291
Zur Kritik neuerer Geschichtsschreiber 25:290–291
on world history 25:290–291
Rankean paradigm 11:39, 11:74
Ranked lineage 3:731
Ranking
auction 2:207
choices 19:440–443
inequality 11:900–902
distributions 11:905–906
models by repeated choices 2:550–551
quantile regression 19:712–713
Ranking research, and public evaluation 2:336
Ranking selection 2:42
Ranking systems 16:772
Rank-order paradigm 22:210–211
Rank-order stability 9:238
of personality traits 17:920, 17:922
Rao, H., on categories 15:576
Rapanui (Easter Island) 17:158–159, 17:162
Rapaport, David 7:934
on projective tests 3:880
Rape 10:193–196, 21:659–661
of black lesbians in South Africa 21:747–748
and sexual coercion 19:887–891
in United States 5:721–723
Rape and sexual coercion
alcohol, role of 19:888–890
biological explanation for 19:887–888
cultural explanation for 19:888
evolutionary theory 19:887–888
legal and historical contexts of 19:888–889
mechanisms involved in 19:887
prevention strategies 19:890–891
in United States
social correlates 19:890
statistics 19:889–890
Rape law reform 9:730
Rape-free societies 21:676
Rape-prone societies 21:676
Raphael, Beverly 25:625
Rapid automatized naming (RAN), and dyslexia 6:729
Rapid cycling disorder 14:485
management of 14:487–488
Rapid Estimate of Adult Literacy in Medicine 10:654
Rapid eye movement (REM) sleep 4:673–674, 6:652–655, 22:87
density 22:82–83, 22:85–86
functional neuroanatomy of 6:654f
and memory 22:63–69
neural system of 22:89–90
sleep hypothesis of depression 22:85
Rapid model development 4:489–490
RAPID outcome mapping approach (ROMA) 19:571f
RAPO *see* Resource-activity-processoutcome (RAPO) model
Rapp, R., on democracy in North America 17:25
Rappaport, Judith 25:625
Rappaport, Roy
works
Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People 1:769, 18:351
Ritual and Religion in the Making of Humanity 20:386
Rapprochement 20:920
Rara 3:147
Rare events 19:733
Rare sequence variants 8:717, 8:722
Rareté 25:10
Rarotonga 17:158–159

Rasanyagam, J. 3:317–318
Rasch models 8:698–699, 16:135–136, 18:605
estimation in 8:699
Rasch-scaled test 17:525
Rashtriya Swayamsevak Sangh (RSS) 20:375–376
Rasputin, Grigorii 21:740
Rassenhygiene movement 8:241–242
Rastafari movement 16:806
Raster model 10:59–61, 10:59f–60f
Rastriya Swayamsevak Sangh (RSS) 4:264
Rasuly-Paleczak, G. 3:317, 3:319
RAT *see* Rational action theory (RAT)
Rat reality show 21:253
Ratchet, Muller's contribution to 16:14
Ratcliffe, Francis, works, *Flying Fox and Drifting Sand* 7:751–752
Ratcliff's diffusion model 6:396–397, 6:399, 15:164
Rate function 8:296–298
Rate of interest 13:20, 19:113–114
Rate of profit 25:7–9
Rate processing constraint hypothesis 13:346
Ratees' reactions, to performance appraisal 17:716
Raters' biases 17:716, 17:718
Rates, demographic 6:110
Rates of change, primordial sociobehavioral outcomes 22:879–880, 22:880f
constant rate 22:879
decreasing rate 22:879
increasing rate 22:879
Ratification 1:612–613, 13:497
Ratified hearers 16:205
Rating scales 6:204, 23:758, 23:764–768, 24:250, 23:766f
of attitude measurement 2:179
extreme response style 23:758–760, 23:762–763
questionnaires 19:772
team constructs 24:100
Ratings 2:211
for audience measurement 2:211, 2:213–214
semantic differential 21:563–571
Ratio effect, numerical 17:46–47
Ratio extrapolation methods 6:164
Ratio matching, in propensity scoring 19:143–144
Ratio methods 6:166
Ratio of likelihoods 21:949
Ratio-correlation method 6:164–165
Rational action theory (RAT) 4:136, 10:279–280
Rational actor model 23:513
Rational addiction 6:672
Rational behavior 24:807
Rational bourgeois capitalism 19:893
Rational brain 11:909
Rational capacity 9:408
Rational capitalism 22:748–749
in Europe 8:274–275
Rational choice approach 4:848, 8:126, 11:561, 13:572, 22:986, 22:997–999
to budgeting 2:903–904
and interests 12:394
prosocial behavior 1:580
and value pluralism 25:21, 25:24–25
Rational choice explanation 19:896–900
collective action 19:896, 19:899
convention 19:899
democratic theory 19:896
fallacies of composition 19:896–897
game theory 19:897
law and economics 19:896, 19:900
norms 19:899–900
Scottish Enlightenment 19:896
social order 19:900
value theory 19:897–898
cardinal, ordinal, Ramsey–Neumann cardinal 19:899
interests, consumptions, and welfare 19:898–899
Rational choice institutionalism 11:562, 12:182, 12:187, 12:189
Rational choice model, behavior voting 25:297
Rational choice theory (RCT) 1:77, 1:669, 1:671–672, 4:825–826, 6:258, 17:781, 18:470, 19:261, 19:264, 19:892–895, 20:359, 22:957–958, 22:997–998, 23:7–8, 25:96–97, 19:895t
and bounded rationality 2:773–774
consumer psychology 4:739
cooperation 4:863
and crime 23:145
of decision making 5:917–919
decision-making 19:893
and definition of situation 6:22
divergence of 19:893
Elster's 7:402–403, 7:406–407
and interest 12:389
nationalism 16:295
of offending 20:776
and organization theory, link between 19:894
origins and premises 19:892–893
of politics modeling 18:512, 18:514
scientific approach 22:994–995
and social mechanisms 22:418–419
in sociology 19:901–906
baseline model 19:901
bounded rationality 19:902–903
embeddedness 19:901–902, 19:904
experiments 19:903–905
extensions at macro level 19:901–902
extensions at micro level 19:902–904
inequality aversion 19:903–904
links to empirical research 19:904
looking ahead 19:904–905
networks 19:901–902, 19:904
social dilemmas 19:902, 19:904
surveys 19:904
trust 19:904–905
and United States presidency 18:862–863
wide versions of 1:77–78
Rational choice value theories 19:900
Rational constructivism 25:275
Rational decision-making 5:911
Rational egoists
cooperation among 4:862–865
embedding 4:863–865
Rational expectations 8:512–513
estimation and testing 8:513
forward-looking character 8:512
Lucas critique 8:512–513
macroeconomics 14:403–404
multiple equilibria and 8:513
Rational expectations hypothesis 12:102–103
Rational ignorance 2:194, 19:537
Rational inattention, expectations 8:515
Rational legal domination, power 18:765
Rational legal proof 13:789–791
Rational model, civil service 3:696–697
Rational network actors 5:558
Rational planning 18:204–205, 24:617, 24:621
Rational planning model 24:889–890
Rational reconstruction 10:773
Rational relations 21:450–451
Rational selves 4:726
Rational theories 5:911–912
Rational transportation 24:617
Rational-critical debate 19:615
Rationalism 7:567–568, 7:883–885, 15:700–701, 15:703, 15:705, 15:807, 20:770
climate change 3:841
critical rationalism 5:271–276
Feyerabend on 9:122
liberal Christianity 14:8
and liberalism 14:22
and psychology 19:411
turn in social movement theory 22:448
Rationalist approaches to prosocial behavior 1:580
Rationality 1:536–537, 2:131–133, 2:494–495, 3:106–107, 10:522–523, 12:333–339, 13:566, 14:418, 15:676, 21:344–345, 21:452, 22:704, 24:508, 24:510
of action 12:337
are we rational? 12:337
asymptotic 5:884
of belief 12:335–336
and bureaucracy 2:913–916
and collective beliefs 4:166–167
and common sense 4:4
conformity 4:1–3
and contractarianism 4:807–808
cooperation 4:862–863, 4:865–866
critical rationalism 5:271–272, 5:274–276
of desire 12:336
ecological 12:722
economic philosophy 7:102–103
economic psychology 5:886–887, 5:890–891
Elster on 7:403, 7:406
and emotions 7:472–473

Rationality (*continued*)
and Enlightenment 7:666
in game theory 9:584–589
and gift 10:140–142
heuristics 10:825
human 12:721–723
identity 11:521–523
institutions 12:191–192, 12:194
instrumental 18:778–780
and interests 12:394
axiological 12:394–395
and magic 14:435–436, 14:438
maximizing conception of 18:778–780
methodological individualism 15:310–311
and modernity 15:688
notion of 4:2
occlusion in social movement theory 22:447–448
in organization theory 19:893
of organizational action 21:138
organizational decision making 17:353
ought we be rational? 12:338
plea for 5:272
and politeness 18:327
Prisoner's Dilemma 18:933–934
and public reasons 19:592
and reasoning 12:338
and relativism 20:234–235
relativism of 19:592–593
and religion 20:309
in society 19:907–912
collective goods 19:909–910
cooperation 19:910
in organizations 19:911
principal-agent theory 19:911
in social networks 19:910–911
social preferences 19:908–910
strategic formation 19:910–911
transaction cost theory 19:911
theory 19:907–909
bounded rationality 19:907–908
goals 19:908–909
heuristics 19:908
learning 19:908
as utility maximization 19:907
and unintended consequences 17:420–421
Rationality hypotheses 15:303–304
Rationality of science 21:212
Rationality paradigm 16:584
Rationality requirements 15:306
Rationalizability 4:874
in game theory 9:585–586, 9:588–589
Rationalization 13:109, 13:114–115, 18:434, 20:407, 22:148, 24:118
of authority 15:697
and bureaucracy 2:913–917
Habermas on 10:470–471
and McDonaldization 14:828–829
organizations 17:307, 17:310
and reification 20:203–205
of religion 15:577
universities 18:308, 18:312
Rationalization processes 19:49
Rationing 10:603
health care 15:100–102
of health care 10:603–606
Ratio-rule models 24:200
Ratios
corporate finance 4:927–929
demographic 6:110
estimation 20:875–876
measurement theory 14:868, 14:870–871
Rats, olfactory learning in 16:471
RATS (Regression Analysis of Time Series) 7:4, 7:6
Ratzel, Friedrich 7:723, 10:91–92, 18:374, 23:199
concept of *Lebensraum* 7:723
works
Politische Geographie 18:374
Raub-wirtschaft, Ernst Friedrich's concept 20:918–919
Raumer, Rudolf von, works, *Geschichte der germanischen Philologie* 14:216
Rauwolfia serpentina 19:463
Rauwolfia 19:463
Raven progressive matrices test 12:316
Raven's Colored Progressive Matrices (RCPM) 5:356, 5:359
Raven's Standard Progressive Matrices (RSPM) 5:323, 5:359
Ravenstein, Ernst 10:169, 10:172–173
Raw data 5:812
Raw material equivalent (RME) 14:762
Raw materials
industrialization 11:869
physical analysis of 10:56
Raw stress 16:41
Rawlings, Craig M., on knowledge and cultural consumption 16:557
Rawls, John 2:671, 4:341–342, 6:591–592, 7:926, 12:137, 12:915–916, 16:323, 19:913–918, 20:893–895, 20:897, 22:392, 22:397–398
conception of good 20:894
conception of justice 22:325–326
on contractarianism 4:807
on difference principle 12:915
education 19:913
family origins 19:913
impact/influence 19:917
on justice 12:943–944
first principle 12:944
second principle 12:944
on political liberalism 20:894–895
post-war academic life of 19:913–914
on public reason 19:592, 19:594–595
on self 20:894
theory of reflective equilibrium 2:633, 12:945–946
wartime experiences of 19:913
on welfare liberalism 9:399
works 19:914–917
achievements 19:914
application of principles of justice 19:915–916
A Brief Inquiry into the Meaning of Sin and Faith 19:913
Collected Papers 19:913
Fifty Years After Hiroshima 19:913
global justice 19:917
institutions 19:914
international relations 19:917
Justice as Fairness: A Restatement 19:914
Justice as Fairness: Political not Metaphysical 19:916
Kantian Constructivism in Moral Theory 19:916
Law of Peoples, The 19:914
original position 19:914–915
Outline of a Decision Procedure for Ethics 19:914
Political Liberalism 17:735, 19:916–917, 20:897
principles of justice 19:915
reflective equilibrium 19:916
social justice 19:914
Social Unity and Primary Goods 19:916
A Theory of Justice 3:675–676, 4:318, 7:87, 7:309, 11:568, 12:137, 16:323, 18:484, 19:914, 24:40
Two Concepts of Rules 19:913
veil of ignorance 19:914–915
Rayleigh distributions 21:949
Raz, J.
freedom as attaining autonomy through community 9:401–402
Razan, Hayashi, works, *Honchô tsugan* 11:49–50
RBC *see* Recognition-by-components (RBC) theory
RBV *see* Resource-based view (RBV)
RCA *see* Radio Corporation of America (RCA)
RCAP *see* Royal Commission on Aboriginal Peoples (RCAP)
rCBF *see* Regional cerebral blood flow (rCBF)
R-class estimators 23:408
RCO *see* Relative concentration index (RCO)
RCPM *see* Raven's Colored Progressive Matrices (RCPM)
RCPysch *see* Royal College of Psychiatrists (RCPysch)
RCs *see* Roman Catholics (RCs)
RCT *see* Randomized clinical trial (RCT); Rational choice theory (RCT)
RCTs *see* Randomized controlled trials (RCTs)
RD *see* Reward dependence (RD)
R&D *see* Research and development (R&D)
RDCs *see* Research Data Centers (RDCs)
RDD *see* Random digit dialing (RDD); Regression discontinuity designs (RDD)
RDIM *see* Revolving Door Identification Model (RDIM)
rDNA *see* Recombinant DNA (rDNA)
RDoC *see* Research Domain Criteria (RDoC)
RDS *see* Relational developmental systems (RDS) theory
RDWTI *see* RAND Database of Worldwide Terrorism Incidents (RDWTI)
REA *see* Right-ear advantage (REA)
Reablement, community care in England 4:347

Reachability 22:462
Reachability index 16:533
Reaching, cortical areas in 5:25–28
Reaction approach, of social comparison 22:211
Reaction norms, life history theory (LHT) 9:41–42
Reaction time 19:472
  see also Response time (RT)
Reactions, program evaluation 19:118
Reactions to receiving help 10:763
Reactive attitudes, as susceptibility to 20:589
Reactive heritability 17:903
Reactive oxygen species (ROS), and Ca^{2+} channels 12:714
Reactive religiosity 8:146
Reactive strategies of policing 18:263
Reactive system, autonomous agents 2:309
Reactive violence 25:118
Reactivity 2:457–458, 2:461, 17:876
  blood pressure 11:453–454
  cardiovascular 3:117–121
Reactivity–reflexivity pair 21:531
Readers 12:641
Readers, journal 12:864
Reader's Digest 10:92
Reader's knowledge, and inference 12:40
Readerships, anthropological writing 1:740–741
Reading
  accuracy of word and pseudoword 20:6f
  acquisition 16:594
  in blind 16:597–598
  character, number and proportion of 20:7t
  in different script systems 16:595–596
  eye movements during 8:631, 8:633
  eye movements in 8:631–634
  integration of information across eye fixations 8:633
  and literacy 14:236–238
  media 15:12–13, 15:15–16
  neural basis of 16:593
  neural systems in adult reading 16:593
  new ways of 14:247–248
  nonalphabetic scripts 20:1–4
  perceptual span 8:632–633
  text-level 16:595
  using writing to support 25:770
  word-level 16:594–595
Reading, dyslexia *see* Dyslexia
Reading acquisition 16:594, 20:5–11
  external influences 20:5
  internal influences 20:8
  writing skills 20:5
  writing systems *see* Writing systems
Reading aloud 13:927–928
Reading and instruction 15:286–287
Reading behavior
  and cultural capital 5:377–378
  journal 12:864
Reading disorders 16:685–686
Reading impairment in children, assessment of 3:794–795
Reading literacy 5:763–765
Ready to Learn Initiative 7:251
Reagan, Ronald 1:605, 18:725, 19:158, 24:846–847
  presidency 18:863
Reagan Administration
  and area and international studies 1:947
Reagan-era deregulatory reforms 10:258
Reaganism 24:866
Reaganomics 16:465, 24:846
Reagan–Thatcher revolution 7:18
Reagency 21:153–154
Reaggregation, public sector organizations 19:601
REAL *see* Resumed Education for Adult Learners (REAL) program
Real
  Lacan on 13:213–215
  philosophy of 10:738–739
  structure of 22:762
Real abstractions, reification 20:204–205
Real economic aggregates
  macroeconomic data 5:753–755
Real economy 12:56
Real estate 3:580, 19:155–157
Real estate development 20:12–19
Real estate economics 11:243
Real estate economies, urbanization 5:775, 5:778
Real estate finance 20:12
Real estate industry
  players 20:13f
  understanding 20:14f
Real estate investment 20:13
Real expenditures, international comparisons of 5:755–756
Real gross domestic product, international comparisons of 5:755–756
Real interests, and power 18:766
Real life events, and integrated attachment theory 12:229
ReAL model, in implicit association test 11:706–707
Real numbers, measurement theory 14:868
Real property, post-socialist transformations of 19:155–158
Real self 21:470
  search, and celebrity discourse 3:275–276
Real socialism 4:303, 4:305
Real wages, and emigration from Europe 15:460
Real world, false memories in 8:709–711
Realabstraktionen 20:203
Real-business-cycle models 5:652–653, 14:404
Realis forms 24:212–213
Realism 5:272, 5:504–505, 7:881, 13:780, 18:210, 20:20–30, 22:225
  versus adaptationism 22:904
  American critical realism 20:21
  analytical 1:668–669
  anarchy 20:26–29
  approach to peacemaking 17:637, 17:639–640
  balance of power 20:26–29
  Bhaskar's critical realism 20:23
    critical naturalism 20:23
    transcendental realism 20:23
  bipolarity 20:27–29
  cognitive and neurosciences 20:25
  commonsense realism of Thomas Reid 20:21
  competitive elaboration: testing presumed mechanisms 20:24
  versus constructivism 7:798–799
  constructivist antirealism 20:21
  critical 22:490–491
  critical rationalism 5:272
  current perspectives 20:21–23
  definition 20:20–21
  disagreements and debates 20:28–29
  and documentary/ethnographic film 6:621, 6:623
  empiricist antirealism 20:21
  Feyerabend on 9:120
  future directions 20:24–25
  hegemony 20:27–28
  hypothesis tests 11:504
  and imperialism 11:682–683
  implications for methodology 20:24
  implications for ontology and epistemology 20:24
  implications for social and behavioral sciences 20:23–24
  internal realism 19:665–666
  Kraepelin on 13:128
  Latour on 13:416–418
  legal reasoning 13:776, 13:780
  Locke's representational realism 20:21
  modifications and varieties 20:29–30
  multipolarity 20:27–29
  and naturalism 10:788
  new realism 20:21
  nominalist critique of 20:204
  pattern-matching 20:24
  and positivism 18:627
  pragmatism 20:25
  problems 20:29
  Putnam's internal realism 20:22–23
    commonsense realism 20:22
    fact-value dichotomy 20:23
    possibility of progress 20:22–23
  reality and scientific theories 20:20–21
  reality of forms 20:20
  reality of things in world 20:20
  retroduction: inferring underlying mechanisms 20:24
  scientific 22:489
  scientific realism 20:22
    causal explanation 20:22
    naturalized epistemology 20:22
  and their opponents 20:227–228
  transcendental 5:277
  varieties in history 20:21
  varieties in international politics 20:26
  vision 20:26–28
    assumptions 20:26
    as equilibrium theory 20:27
    historical roots 20:27
    view of history 20:27–28
  Weber on 25:460–461
Realissimum 18:437

Realistic accuracy model (RAM)
of person perception 17:741
Realistic conflict theory 12:411, 25:803–804
Realistic group conflict theory
of racism 19:866–867
Realistic models, computational neuroscience 4:481
Realistic theories 1:556, 1:558–559
Realistic threats, of xenophobia 25:803–804
Realist–relativist continuum 20:94
Realists 13:773–774
on formalist understandings of judicial decision-making 13:774
scholarship of 13:773–774
versus regularity accounts 3:267
Realities 7:874
and global development 10:151
knowledge representation 13:98–99
native's perspective of 8:185
and performance 17:708, 17:711, 17:713
and play 18:214
in pragmatism 18:809
principle of 5:578
social construction of 22:723–724
social science perspective of 8:185
Reality construction, dramaturgy 6:646
Reality constructionism 18:11
Reality monitoring 23:28
Reality television 3:276, 14:691, 15:62–63
audiences 2:219–221
Realization functionalism 18:110
Realized absorptive capacity 1:14–15
Realm of ideology 13:112
Realpolitik ambitions 16:290
Realschule 25:611
Real-time processing of language 15:755
Real-time technology assessment (RRTA) 24:126
Real-valued utility function 20:927
Real-world currency 8:547
Reappraisal, and emotion regulation 7:453–455
Reappropriation (*Aneignung*) 20:203
Reappropriation of tricksters 24:656
Reason of State absolutism 1:10
Reasonable disagreement, law and morality 13:481
Reasonable pluralism 19:592
Rawls on 19:916
Reasoned judgment, criminal procedure 5:228–229
Reasoning 1:878, 4:45–46, 5:933, 15:776, 23:111–112
alternative views of 16:925–926
Boolean distributivity laws in reasoning 16:925
constructivism 16:925
counterfactual reasoning 16:925
dynamic logics 16:925
intuitionistic logic 16:925
natural reasoning 16:925
paradoxes of material implication 16:925
quantum logics 16:925
scientific reasoning 16:925–926
and decision making 5:921
deduction 19:61–63
dual systems of 19:63
ethnomethodology 8:203–206
individualism 15:305–306
induction 19:63–64
irrationality in 12:719–720
base-rate neglect 12:720
conjunction fallacy 12:719–720
knowledge representation 13:98–100
and learning 13:701
legal 13:776–783
mathematical *see* Mathematical reasoning
mechanism-based 22:417
and mental imagery 15:223
mental state, in ToM tasks 24:269–270
moral 15:789–791, 15:793–794
practical 18:776–781
and rationality 12:338
Reasoning abilities 18:114–115
Reasoning process, emotion's role in 13:463
Reason-of-state literature 12:386–387
Reasons 10:738
for action 18:776–779
conformity 4:1–5
critical rationalism 5:275–276
diversity of 19:591–592
enlightenment 19:591
pluralism and indeterminancy 19:591–592
reasonable pluralism 19:592
relativism and social science 19:591
economic methodology 7:103
enlightenment 7:656–658
and Enlightenment 7:664, 7:666–667
fragmentation of 19:591–592
and freedom 15:538
and linguistic turn 5:302
procedural view of 5:297–298
public 19:591–596
theoretical and practical 5:301–302
and trust 24:670–671
for war 4:617–618
Reay, Diane, on cultural capital 5:377–379
Rebellion
peasants 17:677
and slavery 22:39–40
Rebinding in retrieval 20:611–612
Rebirth 19:308
REBs *see* Research Ethics Boards (REBs)
Reburial 1:885, 1:888–889
Rebus principle 25:781
REC *see* Radical embodied cognition (REC)
Recall
of episodic memory 7:856–860
facilitating 19:773–774
cues and reference periods 19:773
dating recalled instances 19:773–774
impairment 6:719
memory 23:877–879
Recall-and-count model 19:773
Recalling of judgment, in questionnaires 19:772
Recategorization 12:411
Received view, of science 8:576–578
Receiver operating characteristic (ROC) curve 14:803, 19:473–474, 21:329, 21:947–949, 21:954, 21:947t, 21:949f
approaches to fitting 21:948
for binary-response data 21:948f
Gamma 21:949
Gaussian 21:949
index of sensitivity 21:948
NUEV 21:948–949
proportion correct p(c) 21:948
for rating-response data 21:948f
same–different 21:950
source of 21:949–950
Receiving country (adoption) 1:139
Recency effect 25:713
Reception
of celebrity 3:276–277
of culture 5:605–606, 5:732–733
film 9:162
Freyre on 9:424
Searle on 21:345
Receptive fields (RFs) 17:127, 17:129–131, 25:185
and attention 2:163–167
Receptivity, sexual 21:761–762
Receptor adaptation 10:481
Receptor subtypes, stimulation and/or blockade by antidepressants of 1:779–780
Receptors
neurotransmitters 16:752–753
agonists and antagonists 16:753–754
taste and olfactory 24:26
Receptors function, modifications to 17:272
Recession 24:846–847, 24:851–852
of 2007, and fertility in United States 15:403–405, 15:404f
economic psychology 5:886, 5:888–890
Rechtsradikal 18:410
Rechtsstaat ('rule of law' state) 3:710
Recidivism 20:31–36, 20:193–197, 20:199
accurately assessing and predicting 20:34–35
among criminal offenders 8:295
within community context 20:32
overlooked jail populations 20:32–33
importance of research on jails 20:32–33
probation, parole, and jail populations in United States 20:31–32
reentry programming 20:33
risk assessment tools 20:33–34
risk–need–responsivity (RNR) model 20:34
Recipient-oriented theorizing, about justice 12:944–946
Recipients' posttest scores 12:452
Reciprocal actions and effects 21:964, 21:966–967
Reciprocal altruism 7:468, 15:812
in nonhuman animals 15:812
Reciprocal causation, and health 6:879
Reciprocal determinism 15:932, 22:486
self-efficacy 21:510–511

Reciprocal effects model (REM) 12:429–430, 21:462–463, 12:430f
challenges to 21:463
of motivation in Australian Aboriginals 15:903
Reciprocal exchanges, in socialist societies 22:849–850
Reciprocal fairness 2:447
Reciprocal imitation 16:486–487
Reciprocal impact axiom 4:874
Reciprocal liking, in interpersonal attraction 12:628
Reciprocal recognition 21:450–451
Reciprocal relations, engagement 1:34
Reciprocal relationships, political efficacy 7:303–304
Reciprocal surveillance 23:737
Reciprocal Trade Agreements Act 1934 (RTAA) 12:570, 24:488
Reciprocity 1:581, 6:64, 7:14–15, 12:568–569, 20:222–223, 22:247, 7:15f
animal–environment reciprocity 10:128
balanced 7:15
cooperation 4:863–865
direct 8:400, 11:261–262
generalized 7:15
gift and 10:140
indirect 8:400, 11:262–263
interstateness as 25:423
in mutual attitudes 1:73
negative 7:15
perception–action reciprocity 10:128
of perspectives 9:601
and potlatch 18:714, 18:717
principle of 12:363
Prisoner's Dilemma 18:932
social exchange 8:483
social influence 22:349
symmetrical 9:445
at work 23:548
Reciprocity models 11:261–263
Reciprocity principle 22:541
Reciprocity theory 11:261–262
and real-world pairwise cooperation 11:261–262
Recitals of events 11:61
Reckless, Walter C., works, *Vice in Chicago* 7:561
Reckless behaviors, and accidents 12:132
Recodings
probabilities elicitation 7:382
statistical disclosure limitation 4:580
Recognition
agnosia 1:440–443
Butler on 3:56–58
communication to 5:308–309
identity movements 11:543–545
and indigeneity 11:755, 11:757–759
in MENA 16:365–366
politics of 20:44–49
ethics 20:44, 20:46–48
Fraser–Honneth debate 20:45–47
human rights 20:44, 20:46
identity politics 20:46–47
inequality 20:45–48
as matter of concern for social work 20:47–48
parity of participation 20:46–47
social justice 20:44–48
public *see* Public recognition
struggles for 11:178
Recognition and Enforcement of Judgments 13:618
Recognition heuristic, of decision making 5:920
Recognition memory 5:925, 20:37–43
assessment of performance of 20:37–38, 20:38f
corrected recognition scores 20:37–38
discriminability and response bias 20:38
Deese–Roediger–McDermott (DRM) effect 20:42
discrete-state models 6:523, 6:527
dual-process signal detection (DPSD) model 20:40
encoding specificity principle 20:41
factors affecting performance of 20:40–42
context reinstatement effects 20:41–42
facilitative effects of deep encoding 20:41
false recognition 20:42
mirror effect 20:40–41
familiarity 20:37–40
fluency 20:39
global matching models 20:38–39
levels of processing model 20:41
and memory development 15:151
misses and false alarms 20:37–38
models 20:38–40
dual-process models 20:39–40
single-process models 20:38–39
process dissociation procedure 20:39–40
recollection 20:39–42
Recognition of faces 8:651–652
Recognition of others 16:601
Recognition of Prior Learning (RPL) 7:920–921
Recognition primed decision (RPD) model 11:300, 14:481
Recognition theory, Frankfurt School 9:385
Recognition-based decision-making 5:914
fluency heuristic 5:914
recognition heuristic 5:914
Recognition-by-components (RBC) theory 25:154
Recollection
context dependent memory 4:781–782
of episodic memory 7:856, 7:859–860
flashbulb memory 9:255
and hippocampal amnesia 10:879–884
of recognition memory 20:39–42
Recombinant DNA (rDNA) 2:677, 2:679–680, 9:975, 9:1015–1016
Recombinant inbred (RI) strains 9:933–934
Recombination 9:915, 11:289
Recommendations of consultants/academics 17:851
Recommended acts 23:23
Reconceptualizations
Merton on 15:246–249
of social structure 22:724–725
Reconciliation 12:958, 17:619, 23:812–813
emotions 7:504–505
mass killings/genocide 14:705
and peacemaking 17:637–638
Reconciliation procedures 17:648
Reconciliation processes, and collective memories 4:197–198
Reconfiguration of gender 9:781
Reconsolidation 8:851–852, 8:872
fear conditioning 1:813
memory 23:877–879
of reminder memory 15:159
Reconstruction 11:824, 24:865
after disasters 10:549–551
in critical social work 5:290–291
of a life 2:644
in retrieval 20:611–612
Reconstructive doing, artistic creation as 1:230
Reconstructive memory 20:50
background 20:50
long-term memory 20:51–53
misinformation paradigm 20:53
pictures and scenes 20:53
prose recall 20:52–53
psychology of 20:50–55
questionnaires 19:773–774
sensory and short-term memory 20:50–51
word lists 20:51–52
Reconstructive science 10:787
Habermas's contribution 10:469–471
Record indicators 16:898–900
Record linkage
family reconstitution 6:139–140
life course epidemiology 14:73
privacy in social research 18:945–946
Record methods, for health behaviors 10:590–591
Record on appeal 1:824
Record statistics 16:898–906
different generalizations of classical records 16:902–905
distribution of record times 16:900
distribution of record values 16:900–901
distributions of kth record values 16:902
δ-exceedance record scheme 16:903
extreme order statistics 16:898
F^{α}-scheme 16:902
linear draft record model 16:902–903
records with confirmation 16:904
records with restrictions 16:903–904
sequential ranks and record indicators 16:898–900
strong records 16:900
weak records 16:898, 16:900
Record times, distribution of 16:900
Record values, distribution of 16:900–901
Recording
demographic events 6:141–142
electroencephalography 7:355–356
and transcribing interviews 17:288
Recording therapy 20:738
Recordings of language 14:183, 14:185
Records

Records (*continued*)
definition 16:898
field observational research 9:138
as historical sources 5:701–702
informed consent in studies based on review of 12:112–113
Recovered memory 8:710, 21:159–160
Recovery 4:355, 4:357
alcohol abuse 1:521
cardiovascular 3:117–121
from disaster 6:482
evacuations challenges 6:482
gender perspective 6:482
psychosocial impacts 6:482
social capital in 6:482
from diseases 22:735–740
drug addiction 25:351–354
of fear 8:850–851
from mental illness 15:195–196
and natural disasters 16:317–318
versus sick role 22:976–977
Recovery Movement 19:202
Recovery of human remains 9:335
Recovery patterns 2:574–575
Recovery self-efficacy 10:711
Recreation
in collectivism/individualism 4:209
entertainment 7:670
Recreational drug use and rights 6:676
Recreational expenditures, leisure and cultural consumption 13:837–839
Recreational genetics 6:442
Recreational genomics 2:623
Recreational travel 9:639
Recruitment 14:354–355
in administration 1:106
elites 7:397–400
of participants, in developmental research 6:327
in political participation 18:506
into psychiatry 7:188–189
volunteerism 25:269, 25:271–272
RECs *see* Research Ethics Committees (RECs)
Rectangularization of survival curve 15:845–846
Recurrence fear, of cancer 3:73
Recurrent demographic events 5:740
Recurrent events
analysis 8:296–297
multitype 8:297–298
Recurrent loops, in occipital lobe 17:131
Recurrent pain 3:556t
Recursion, of shame 21:871
Recursive feature elimination (RFE) 16:119
Recursive publics 12:602
Recursive single-response regressions 10:342
Recursivity 13:601, 20:91–92
Recycling 7:699–700, 7:769, 7:775–776, 7:700t
and sustainable transportation 23:845–847
RED *see* Research on Bono's Product (RED) campaign
Red Crescent Societies 6:482
Red Cross movement 11:372, 17:640
Red de Protección (RPS) 4:556–557
Red October 20:857
Red Purge 14:638
Red Queen principle 8:449–450
Red relief services 14:667
Red terror 20:857
Redakteur 16:821
Reddit 22:422–423
Redefinition, family networks 21:755–756
Redemption, eudaemonism 8:214–216
Redescending M-estimators 20:735
Redfield, Robert 1:769
on brokers 2:870
Redistribution 18:714
of income 11:747
politics of 20:45–47
social justice 22:392
Redistributive bureaucracies, in socialist societies 22:849–850
Redistributive justice 22:387
Redistricting, political representation 18:443
Redlining 16:452
and residential segregation 20:542
Redress 12:958, 14:266
Reduced competition 6:537
Reduced form equations 21:980
Reduced rank regression 24:326–327
likelihood analysis 24:327
Reduced response vectors
longitudinal data 14:347
aggregation 14:347
Reducing emissions from deforestation and forest degradation 7:716–717
Reduction
approaches to intentionality 12:334–335
of chemistry 3:373–374
Reduction hypothesis 25:170
Reduction of Hazardous Substances (RoHS) directive 7:371
The Reduction of Stigma in Schools program 21:89
Reductionism 13:136, 14:315, 16:583, 22:224
and biomedicine 2:663–664
in legal philosophy 18:29
meaning 14:843
and mental causation 15:175
Muller's view 16:16
Reductionism approaches, for environmental decisions 7:708
Reductive value pluralism 25:21
akrasia 25:22–23
arguments of 25:21–24
implications of 25:24–25
ordinary intuition 25:22
unavoidable loss 25:23–24
Redundant community management 3:454
Reed, Adolph 24:730
Reemployment wage losses 12:819
Reentry
after incarceration 20:61–66
challenges 20:61–64
emerging policy lessons 20:64–65
employment 20:62–65
interventions 20:61, 20:64–65
needs 20:61–62, 20:65
policy 20:61, 20:64–65
program 20:65
rehabilitation 20:61, 20:64
Reentry programming, recidivism 20:33
Re-envisioning international social problems 12:543
Reeves, Steve 13:890
Reexpression 3:164, 8:600
Referees 17:852
Reference (syntax) 23:886t
Reference books 7:588–591
Reference groups 20:72
American soldier study 20:73–75
anticipatory socialization 20:73–74
eligibility 20:74
group membership 20:72–74
group process 22:541
job satisfaction 20:72–73
in methodology of social research 20:74–75
normative reference and development of self 20:72–73
orientation 20:72–75
relative deprivation 20:73–74
role-taking 20:72–73
social comparisons and material deprivation 20:73–74
in social influence 22:346
social status 20:72–75
Reference machine 1:534
Reference maps 3:171
Reference materials, abundance of 7:586–587
Reference memory 24:308–309
Reference models, for speech and language impairment 3:792–793
Reference periods 19:773
Reference points, in framing effects 2:445–446
Reference populations 9:899
Reference societies 8:276
in Europe 8:276
Reference terms, and kinship terminologies 13:62
References 20:67–71
causal theory of 20:69
description theory of 20:68
direct reference theory of names 20:69
and linguistic representation 20:68–70
content and logical equivalence 20:69–70
and language of thought 20:70
theories for reference for names 20:68–69
in personality assessment 17:850, 17:852
References problem, in learning word meaning 25:227–228
Reference-tracking system 1:674–676
Referendums
direct democracy 25:294
voting behavior in 25:296–298
Referent known exactly (RKE) model 24:309
Referential coherence 12:39
Referential expressions 14:222
Referential functions of language 12:783

Referential understanding, of infants 22:558
Referentially opaque conditions 12:333–334
Reflected glory 21:473
  academic self-concept and achievement 1:60
Reflecting thoughts and feelings 4:288–289
Reflection 5:281
  centrality measures involving 16:533–534
  critical 5:284–285, 5:289–291
  and reflexivity 20:82–86
Reflections on history 11:56–57
  among Mesoamerican Indians 11:56
  and human life 11:56–57
Reflections on the Revolution in France 20:628
Reflective behavior 10:624
Reflective equilibrium 2:633
  applications 20:81
  concept 20:77
  considered judgments 20:77–80
  construals 20:78–79
    general *versus* individual 20:79
    normative *versus* descriptive 20:78
    radical *versus* conservative 20:78–79
    wide *versus* narrow 20:78
  constructivism 20:79
  criticisms 20:79–81
    objections 20:80–81
    worries about coherentism 20:79–80
  intuition 20:80
  justification 20:77–81
  method and history 20:77–78
  Rawlsian model 19:916, 20:80–81
Reflective Judgment Interview (RJI) 7:888
Reflective judgment model, of epistemological beliefs 7:887–888
Reflective listening 15:911–912
Reflective model of action 10:478
Reflective practice 24:521
  professional development 19:90–91
    and supervision 19:92
Reflective teacher paradigm 24:52
Reflex (language), definition 4:424
Reflexes 14:367, 16:506–507
  and instinct 12:175–179
  and prenatal/infant development 18:852–853, 18:857–858
  startle 23:331–337
Reflexive abstraction 18:117–118
Reflexive account, of mutual beliefs 21:877–878, 21:880
Reflexive author 6:623
Reflexive capabilities 21:531
Reflexive designs *see* Pre–post designs
Reflexive documentary 9:164
Reflexive environmental law 13:555
Reflexive ethnography 8:180–181
Reflexive inquiry 20:90
Reflexive law 13:439, 13:555
Reflexive modernization 2:397, 15:698, 15:705, 20:95–96
  Beck on 2:397
Reflexive narrative sequences 4:725
Reflexive practice 20:83–84
  heuristic and process 20:85–86
Reflexive programmatics 20:94
Reflexive pronouns 13:315–316
Reflexive sociology 22:995
Reflexive tourism 24:468
Reflexivists 20:94
Reflexivity 4:727, 5:281, 8:913, 17:525, 18:799, 19:679, 20:82–87, 20:90, 20:93, 22:691, 23:592
  in anthropology 20:88–92
    critique 20:89–90
  assemblages 20:91
  ban on 20:94
  Bloor's 20:94
  Bourdieu on 2:781–782
  Collins' 'management strategy' for 20:94
  and common sense 4:2
  critical 5:281–282
  critical reflection 20:82–83, 20:85–86
  criticisms 20:94–95
  discourse of 15:712
  epistemic 21:672
  ethnography 20:84–86
  ethnography in 20:88–92
  ethnomethodological 21:193
  evidence-based practice 20:82–87
  and experts 8:590–591
  feminist research 9:717–718
  fieldwork 9:151
  fieldwork in 20:88–92
  Giddens on 10:136–137
  hermeneutic 20:84
  infra-reflexivity 20:94
  institutional 20:96
  modernization 20:90–91
  new literary forms 20:95
  and postcolonial geography 18:645
  postmodern 19:680–681
  postmodernity and 20:88–91
  practicing 20:84–85
  present situation 20:95–96
  as a problem 20:93–94
  of qualitative research 19:687
  radical 20:94–95
  reflection 20:82–86
  and reflexive practice 20:83–84
    heuristic and process 20:85–86
  and research 20:84
  of risk 20:707–709
  Rolfe's model of 5:283
  in 2000s 20:91–92
  in science and technology studies 20:93–97
  social science's ubiquitous 20:95–96
  and Strong Program 23:593
  symbolic interactionism 12:353–354
  and work anthropology 25:667
Reflexivity analysis 1:677
Reflexology 10:562
Reform Judaism 12:871–872
Reform movement 7:559, 7:562, 12:871, 12:912
Reform of financial regulation 9:191, 9:194–195
Reformasi 13:815–816
Reformation 2:145, 3:62–64, 3:234, 3:236, 8:258–259, 12:363, 14:254, 16:224, 16:298, 20:98–104
  Calvin on 20:98–103
  Catholic Reformation 20:98, 20:100–101
  Christianity 8:287
  confession building 20:99–100
  confessional age 20:98, 20:100, 20:103
  cultural history 20:99–100
  era, religious, social, and political structures and developments in 20:98
    basic structures 20:98–100
    Europe 20:101–102
    Germany 20:101
    late Middle Ages 20:100–103
    Protestant and Catholic theology 20:100–101
  historiography and periodization 20:98–100
    counter-reformation 20:98
  Luther on 20:98–103
  modernity 15:682
  religious movement, and educational development 7:234–235
  social history 20:99–100
  state formation 20:100, 20:103
  Zwingli on 20:100–101
Reformed Churches 3:543
Reformed Judaism 21:174
Reformer's science, anthropology as 1:737
Reformism 2:890–893
  Muslim 12:741–742
Reformists 20:634
Reformpädagogik 1:870–871
Reforms 16:299
  antiracist legal 19:809–810
  economic transformation 7:75–76
  family policy 8:799, 8:801, 8:803–804
  income tax 11:748
  LAC educational system 7:179, 7:183–184
  land tenure 13:232–236
    bureaucratic simplification 13:233
    collectivization and villagization 13:233
    contested formalization of land claims 13:235
    displacement, agrarian encroachments, and land grabbing 13:235
    land rights and tenure security in postsocialist settings 13:234–235
    special issues in irrigated and flood-prone lands 13:235–236
    titling and individuation 13:233–234
  parliamentary 2:528
  social 2:526
  use of term 25:515
  for WCP exclusion in clinical trials 9:720–721
Refraction effect 9:140
Reframing 6:563–564
Refrigerator mothers 16:343
Refugee Convention 2:108, 2:111
Refugee populations 5:474–475
Refugee status, social exclusion due to 22:288
Refugees 2:108–112, 25:125

Refugees (*continued*)
academic concepts 20:105–106
in anthropology 20:105–110
and armed conflict 1:970–973
asylum seekers 20:105, 20:107–108
children 20:108
definition 20:111
diasporas 20:108
dynamics of displacement 20:106
and ethnic cleansing 8:88
ethnographic challenges 20:108
in Europe 15:382
geographical aspects 20:111–116
encountering borders 20:113
flight 20:112–113
future directions 20:114–115
host countries 20:113–114
internally displaced peoples 20:112–113
international refugee regime 20:111–112
territorialization 20:112–115
UNHCR 20:111–115
and inequality 11:894, 11:898
legal definitions 20:105–106
and migration 15:416–420
research on 20:112
situations and research themes 20:106–107
and social work 22:769–774
advocacy 22:771, 22:773
citizenship 22:769–770, 22:772
core therapies and concepts 22:771–772
definitions 22:769–770
ecological theory 22:771–772
electronic resources informing macro- and micropractice 22:770–771
ethical challenges 22:773
historical and political contexts 22:770
human rights 22:769–771, 22:773
justice 22:769–770, 22:772
literature review 22:770
resilience 22:770–771
roles and practices 22:772–773
settlement 22:769–773
welfare regimes 22:770
women 22:771–773
South and Southeast Asians 12:533–534, 12:533t
studies on 20:105
future of 20:108–109
warehousing 20:107
Refunctioning ethnography 16:103
Refurbishment of professional ideals 13:630–631
Refuse deposits 8:163–164
Refutable implications, hierarchies and markets 10:843
Refutation 16:926
Regeneration games 24:866
Regev, Motti, on popular music 3:800
Reggie program 19:557
Reggio, Godfrey, works, *Koyanisqatsi* 6:621
Regime instability 8:126–127
Regime of love 2:748
Regime of peace 2:748
Regime theory 18:763
and political power 18:762–764
Regimes
definition 16:416
prison 18:921–922
revolutions, theories of 20:627–629
Regimes of justice 2:748
Region connection calculus, RCC-8 14:320
Regional analysis
of markets 14:561
Regional and Comparative Regional Study (ERCE) 7:180–181
Regional approach, geographic thought 10:70–71
Regional authoritarianism 16:776–777
Regional Center for Higher Education in Latin American and the Caribbean (CRESALC) 7:183
Regional cerebral blood flow (rCBF) 11:615
Regional characteristics of crime 12:465–467
Regional chauvinism, and anthropology 1:752
Regional cities 20:134
Regional comparison 4:412
Regional conflicts, warfare 25:418–419
Regional cuisines and identities 9:293–294
Regional data, interpretation, settlement and landscape 21:657
Regional development 14:288
Regional development planning 20:134–137
in Africa 1:298
Regional differentiation, Kula ring 13:142
Regional disparities
and infant and child mortality 12:15–16
in social work development 22:792–793
Regional economic development and policy 20:150
Regional economic growth, and urban poverty 24:916
Regional Economic Studies (RES) 22:641
Regional economics 14:287–289, 14:291
Regional economies, pastoralism within 17:580
Regional Educational Laboratories (RELs) 7:283
Regional geography 10:84, 20:117–128
areal differentiation 20:122–123
birth of 20:122
critique 20:126
from environmental determinism to regional synthesis 20:117–118
exceptionalism 20:123–125
future 20:126–127
geoeconomics of regions 20:125
geopolitics of regions 20:125–126
idiographic science 20:122–124
industrial restructuring 20:117–119
new 20:118–119, 20:123–125
new regionalism 20:125–126
nomothetic science 20:122–124
place 20:123–125, 20:127
regional worlds 20:125–126
relational approach 20:127
relational/territorial 20:119
scale 20:125–127
scales 20:119
schools of thought 20:121f
as science 20:122–123
socially constructed 20:124
space 20:122–125, 20:127
territory 20:124–127
traditional 20:122
Regional growth 14:287–288
Regional growth management 10:451
Regional historical geographies 11:18
Regional histories, Indo-Islamic period 11:79
Regional institutionalization 20:124
Regional integration, and state formation 23:353
Regional integration agreements (RIAs) 12:574, 12:577–578
Regional items, scaling of 21:6
Regional migration 18:598
Regional Plan of New York and Its Environs 20:142
Regional planning 20:129–134
academic interests 20:130–131
in China 20:136–139
historic context 20:136–137
urbanization challenges 20:137
city regions 20:138–140
development planning 20:134–135
doctrine of 20:135–136
and economic development 20:131–132
environmental planning 20:129, 20:131–133
future of 20:132–133
governance 20:129–130
in historical focus 20:129–130
history of 20:141–145
Appalachian Regional Commission 20:143
British 20:144–145
European 20:144–145
future prospects 20:145
jurisdictions 20:145
metropolitan 20:143–144
origins and practice 20:142–145
Regional Plan of New York and Its Environs 20:142
Regional Planning Association of America 20:142
settlement 20:141
Tennessee Valley Authority 20:142–143
terminology 20:141–142
regional analysis 20:141
regional-cities 20:134
strategic planning 20:129–133
transport planning 20:129–131
types 20:139t
urban regeneration 20:131
urbanization challenges 20:134–135, 20:137
waves of 20:137–139, 20:137t
Regional Planning Association of America (RPAA) 20:142
activities 1:219
garden cities 1:219
membership 1:219

Regional science 20:146–151, 24:945
analysis, methods 20:149–150
intellectual progress 20:148–150
location and 20:149
subject matter 20:147–148
Regional Science and Urban Economics 20:147
Regional Science Association International (RSAI) 20:147
Regional Science Policy and Practice 20:147
Regional Social and Cultural Studies (RSCS) 22:641
Regional specialization 21:650
Regional Strategic and Political Studies (RSPS) 22:641
Regional subarea analysis 24:803–804
Regional survey programs, large-scale, settlement and landscape 21:655
Regional treaty, jurisprudence 11:799–800
Regional worlds 20:125–126
Regionalisms 20:125–126
and anthropology 1:752
in Europe 8:278
Mumford on 16:137–138, 16:140
trade policy 24:485, 24:489
Regionalization
of media productions 2:220
parliamentary democracy in 17:514
Regional/local subcultures, and cultural policy 5:461
Regions 22:438
description 20:129–130
and economic development 20:131–132
and gender 9:782–783
Register-based census 3:297
Registered nurses (RNs) 16:908–910
Registers 3:298
economic statistics 5:761
in linguistics 14:171–172
Registration problem, functional data analysis 9:514–516
Registries 5:784
Reglobalization 10:231
Regnal years 24:409–411
Regression 12:205–206, 14:153–156, 16:81–82, 19:434
causal inference 3:258–259
developmental psychology 5:355–356
ecological 6:869
with random coefficients 6:869
interpreting parameters 14:156
logistic *see* Logistic regression
multiple regression 14:155–156
nonparametric 16:867–877
optimal design
logistic regression in one variable 17:260
nonlinear regression models 17:261
simple regression 17:256–257
simple linear regression 14:154–155
diagnostics 14:154–155
fitting 14:154
Regression analyses 14:144, 17:662, 19:727, 20:153
for class size 3:751
definition of 14:157
of education 7:132–133
graphics 14:157–161
estimation of sufficient summary plots 14:159–160
mean function 14:157–160
model-based graphics 14:160
special purpose graphics 14:160
sufficient summary plots 14:158–159
variance function 14:157–160
Regression Analysis of Time Series *see* RATS (Regression Analysis of Time Series)
Regression approaches
observed variables 4:840–841
quasiexperimental and nonexperimental designs 4:841
in randomized experiments 4:840–841
Regression artifacts 12:653–654
Regression calibration 14:853–854
models 14:851
Regression coefficient 15:66
Regression decomposition approach
discrimination 18:831
Regression discontinuity 20:153
Regression discontinuity designs (RDD) 12:453–454, 19:743–744, 20:152–159, 21:125–126
advantage of 20:154
analysis 20:155–156
cutoff points 20:156
multiple assignment measures 20:156–157
steps in 20:155–156
applications 20:152
assumptions of 20:155
comparison group pretest variance 20:155
continuous pretest distribution 20:155
cutoff criterion 20:155
pre–post distribution 20:155
treatment implementation 20:155
attraction of 20:152
for class size 3:752–753
for a compensatory treatment showing a positive effect 20:153f
considerations and variations 20:153–155
assignment variations 20:153–154
multiple assignment measures 20:154
multiple cutoff points 20:154
selection of cutoff 20:153
cutoff point 20:152
definition 20:152
history 20:152
internal validity 20:154–155
interpretation of 20:153
methodological developments 20:157–158
design variations 20:157–158
fuzzy RD design 20:157
nonparametric statistical analysis 20:157
model for basic 20:155
specification 20:155
model specification 20:155
parametric analysis 20:155–157
pre and post measures 20:154
pretreatment cutoff value 20:153
statistical power 20:155
structure 20:152–153
treatment comparisons 20:154
posttreatment measurement variations 20:154
Regression effect 21:605
Regression equations, estimates, distribution of 21:975, 21:978–979
Regression for categorical variables 14:323
Regression formulation 24:323
Regression graphs 10:341–343
Regression imputation 16:922–923
Regression methods 24:347–348
seasonal adjustment 24:351–352
Regression modeling, of categorical response 16:132–133
Regression models
event–history analysis in discrete time 14:351–352
survival analysis 23:774
Regression techniques 6:164–165
Regression toward mean 12:452
Regression trees 23:393
nonparametric regression 16:874–876
in propensity scoring 19:142
Regression with instrumental variables 14:853
Regressions, eye movements 8:631–632
Regressive dialectic 10:490–491
Regressive transference neurosis 24:546
Regret 14:307, 21:817
value pluralism 25:23
Regret theory, of decision making 5:921
Regular conventional gas 17:167–168
Regular conventional oil 17:167–170
Regular equivalence, structural equivalence 23:621, 23:623
Regular grammars 19:10–12
computation 19:12
inference 19:12
probabilities 19:11–12
Regularization
functional data analysis 9:516–517
Regulated market system 3:97
Regulation and administration 20:160–165
cost 20:162
emergence and justification 20:160–161
institutions 20:161–162
political institutions and political culture 20:162–163
Regulation S (Securities and Exchange Commission) 3:32
Regulation theory, entrepreneurship 7:687
Regulations, of lawyers
with accurate information 20:166–168
linear pricing with multiple outputs 20:167
linear pricing with single output 20:166
nonlinear pricing 20:166–167
quality deterioration and cream skinning 20:167
two-sided markets 20:167–168
broadcasting 2:859–863
campaign, direct democracy 25:295
and civil service 3:699
competing for benefits 20:171

Regulations, of lawyers (*continued*)
of conduct, social work ethics 22:785–786
content providers (CPs) 20:167–168
coping as 4:891
definition 20:166
disclosure regulation 20:170
economic 20:172
economic theory of 20:166–171
empirical analysis of 20:172–177
environmental 7:710–711
environmental policy 7:779
externalities regulation 20:170
financial 9:191–196
financial regulation 20:169–171
health care regulation 20:170
health policy 10:665–666
hospitals 11:208–209
of imitation by infants/young children 11:623–625
implementation in 11:693
incentive regulation 20:168–169
earnings sharing regulation 20:168
franchise bidding 20:168–169
options 20:169
price cap regulation 20:168
regulatory lag 20:168
yardstick regulation 20:168
informal sector 12:57
insurance sector 12:213–214
internet service providers (ISPs) 20:167–168
of lawyers
changes in focus and emphasis over time 13:622–623
current state of knowledge 13:620–621
databases and software resources 13:624–625
definitions 13:619–620
emerging trend 13:623–624
methodological problems 13:624
sources and methods of 13:621–622
of LGBTI commercial sex 13:884–885
life span development 14:102–106
market allocation 14:542–543
media 2:860, 2:862–863, 15:57
political economy of, 20:178–180
economic regulation 20:178–179
estimating political support 20:179
intervention 20:178
investment 20:178
measuring distributional effects 20:178–179
modeling regulatory agency behavior 20:179
pricing 20:178–179
product choice 20:178
production 20:178
production inefficiencies 20:178
of pollution 20:170
of privacy on Internet 12:596–597
privacy regulation 20:170
of research 20:509–510
role in community-based health interventions 10:648
sector regulation 20:170–171
service quality regulation 20:170
social work 13:723
of therapeutic effectiveness assessment 24:284–285
vertical divestiture 20:169–170
in vertical related industries 20:169–170
input prices regulation 20:169
limiting the disadvantaging of retail rivals 20:169
market structure regulation 20:169–170
working conditions 25:698–702
Regulative law
and organizations 17:397–398
as dependent variable 17:400–401
as independent variable 17:400
Regulative rules, meaning 14:846–847
Regulators, of risk 20:706–707
Regulatory agencies 20:181–186
bureaucracy 20:182–183
business 20:181–183
efficiency 20:181, 20:183–184
empirical inquiry 20:182–183
environment 20:181–185
of European Union 20:181, 20:184–185
externality 20:181, 20:183–184
firm 20:181, 20:185
interest group 20:182
interest group liberalism 20:182
international and comparative perspectives 20:184–185
justice 20:183–184
legislation 20:181–184
life cycle theory of regulation 20:182
market failure 20:181, 20:183
monopoly 20:181
normative assessments 20:183–184
principal-agent theory 20:182–183
regulatory reform 20:183
rent seeking 20:182
zero-risk mind-set 20:183
Regulatory bioethics 2:636
Regulatory bureaucracy 19:531, 19:533
Regulatory dimension, of law 13:604–605
Regulatory focus theory 25:684–685
Regulatory issues, of linguistic databases 5:815
Regulatory lag 20:168
Regulatory protections for human research participants 4:252–257
Regulatory reform 20:183
Regulatory regimes 14:736
Regulatory science 21:157, 21:161
Rehabilitation 3:480, 13:575, 15:969
bilingual aphasia 2:577–578
cardiac 20:187–192
of chronic pain 3:555, 3:559–560
disability 6:472–474
of drivers in transportation 24:626
for hemispatial neglect 10:769–770
and natural disasters 16:316–318
of offender 20:193–202
of offenders 5:210
posttraumatic amnesia 18:638–639
of prisoners 20:61, 20:64
and probation 19:43–48
and sex offenders 21:676–677
vocational rehabilitation 22:383
Rehabilitation Act 1973 6:473–474, 23:221
Rehabilitation of law 13:793
Rehabilitation social work
behavioral 10:716t
emotional 10:716t
physical 10:716t
social 10:716t
Rehabilitative treatment program 11:721
Rehearsal, and flashbulb memory 9:257
Rehearsal strategies
for basic learning tasks 13:713
for complex learning tasks 13:713
Rehn-Meidner model 13:185
Reich, William, on Russian sexual life 21:740–741
Reich Chancellery 16:258–259
Reichard, Elias Caspar, works, *Versuch einer Historie der deutschen Sprachkunst* 14:215
Reichian analysis 22:696
Reichsanstalt, Physikalisch-Technische 18:102–103
Reichskriegsgericht 16:357
Reichsverwaltungsgericht 16:357
Reichswehr, in Bavaria 16:257–258
Reid, Anthony, works, *Southeast Asia in the Age of Commerce, c.1450–1680* 11:87
Reid, Thomas 19:411
commonsense realism 20:21
Reification 5:244–246, 13:551–552
and fetishism 9:107–110
history 20:203–206
alienation 20:203–205
critical theory 20:203–205
critique of conceptual abstractions 20:203–204
critique of real abstractions 20:204–205
dialectical critique of fetishism 20:204
Habermas's contribution 20:203, 20:205
Honneth's contribution 20:203, 20:205
humanist critique of naturalism 20:204
hypostasis 20:203–204
Marx's contribution 20:203–205
methodological reification 20:203–204
nominalist critique of realism 20:204
rationalization 20:203–205
Simmel's contribution 20:203–205
social reification 20:204–205
Weber's contribution 20:203–205
Reiki 10:562
Reimagining of life 9:1016–1017
Reimbursement
for combined pharmacotherapy and psychotherapy 19:497
hospital 11:208–209, 11:211
psychiatric ethics 8:66
Reindigenization 11:760
Reinforcement 1:536–537, 1:655, 7:240, 10:622, 20:207–210
antecedents 20:208
applied behavior analysis 2:407–411
A-R-D model 20:209
as behavioral regulation 20:209–210

bliss point 20:209–210
classical conditioning 20:207–208
concept history 20:207–208
consequences 20:207–210
constraints in generalizing from animal laboratory 20:208–209
contingencies, Skinner 22:25–26
contingency principles 20:208
extinction 20:207–208
feedback 20:207–208
instrumental conditioning 20:208
intrinsic 20:207
motivation, reward, and individual differences 20:209
negative 20:207
neurochemical substrates 20:211–216
reward dominant individuals 20:209
schedules 20:208
secondary 20:207
shaping 20:208, 20:210
Reinforcement learning 8:389–390, 14:575, 15:964, 19:908, 23:413–414
brain stimulation reward 2:843–844
motor control 15:964
Reinforcement sensitivity theory (RST) 8:624, 17:789–790
and personality 17:828
personality traits 17:860–861
revised 8:624
Reinforcement systems 20:211
in cellular operant conditioning 20:215
molecular hypothesis of 20:215
Reinforcers, behavior therapy
which increase desirable behaviors 2:429
negative reinforcers 2:429
positive reinforcers 2:429
which reduce undesirable behaviors 2:429
punishment 2:429
response cost 2:429
Reinforcing neurons
brain self-stimulation evidence 20:212–213
drug self-administration evidence 20:213
electrophysiological evidence 20:214
methodological caveats and terminology 20:213–214
Reinhart, Carmen, works, *This Time Is Different* 9:183
Reinstatement 8:850–851
Reinstatement paradigm, for context dependent memory 4:779, 4:779f
Reinsurance 7:785
Reintegration
after incarceration 20:61–66
challenges 20:61–64
emerging policy lessons 20:64–65
employment 20:62–65
interventions 20:61, 20:64–65
needs 20:61–62, 20:65
policy 20:61, 20:64–65
program 20:65
rehabilitation 20:61, 20:64
childhood cancer 3:462
and probation 19:44, 19:47–48
programs for child soldiers 3:434–435
Reintegrative shaming 1:835
Reintegrative stage of aging 1:436–437
Reiter, D., works, *Democracies at War* 25:357–358
Reiyukai(diacritical mark) 20:409
Reject/don't reject decision 11:495
Rejecting parenting style 24:196
Rejection of outliers 23:388–389
Rejection region 11:491–493
Rejection sensitivity 22:564–567
Rejectionist integration 11:644
Reklos and Davourlis v. Greece (2009) 18:950–951
Relapse prevention (RP) 20:217–220, 20:435
addictive disorders 20:217–219
conceptual basis of 20:217
coping skills 20:218
efficacy of 20:218–219
methodological issues 20:219
strategies 20:218
Relapses
childhood cancer 3:461–462
drug addiction 25:351–354
Related attributes, social comparison 22:211–213
Related attributes hypothesis 22:211
Relatedness 13:44–45, 15:931, 21:487–490, 24:68
culture, kinship and difference in 9:427–431
and genealogy 9:806
kin selection 13:26–31
and spirituality 23:285–286
Relation
ordered relational structures 17:296–297
and regional geography 20:119
and self 21:439, 21:441–443
Relational accounting 15:743
Relational aggression 1:373
Relational analysis 9:359–360
Relational approach, to regional geography 20:127
Relational area of harmony 10:530–531
Relational axiom 4:872
Relational biography 2:647
Relational class analysis 5:560
Relational coherence 12:39
Relational cohesion, and social exchange 8:486
Relational complexity model 16:504–505
Relational composition
algebra 22:478–480, 22:483
binary operation of 22:479
Relational concerns, and adolescent leisure activities 13:832
Relational contracting 11:683
and imperialism 11:683
Relational demography, and diversity 6:596
Relational developmental systems (RDS) theory 3:445, 11:278–279
adolescence within 3:446–447
optimization to developmental science within 11:279
Relational diffusion 22:439
Relational empowerment 20:222–223
Relational epistemology, hunting/gathering societies 11:430
Relational feminists 9:730
Relational Gompertz model of fertility 6:146
Relational guide *see* Facilitator
Relational identity 11:550–551
Relational leadership
leader–member exchange (LMX) theory 13:641–647
Relational models 6:134
Relational reasoning, and mental imagery 15:223
Relational regulation theory (RRT) 22:367
Relational semantics 13:932
Relational social work 20:221–226
coping network and its freedom of action 20:223
methodological and technical principles 20:223–226
facilitation of coping networks 20:224
facilitator 20:224
graphic representation of relational coping 20:224–226
members of coping network 20:224
optimistic look forward 20:224
reciprocity and relational empowerment 20:222–223
respecting otherness 20:221–222
self-help and mutual help 20:223
social capital 20:223
social relationship 20:221
theoretical bases and key concepts 20:221–223
Relational sociology, and meaning 16:562
Relational space (geography) 16:567
Relational spirituality, religion 20:317–321
Relational structure 12:783
Relational text analysts 4:770
Relational theories 17:743
abusive supervision 23:702–703
fairness theory 23:702
social exchange theory 23:702
uncertainty management theory 23:702
conceptions of mind 22:763
of personal identity 17:743
identity does not matter 17:743
psychological continuity 17:743
psychodynamic treatment 22:764
psychopathology in 22:763
Relational thinking
and cultural production 5:466
and language 13:295
Relational transformation 4:153
Relational transparency, and authentic leadership 2:247
Relationalism 11:792
Relationality 11:551
and infrastructure 12:123–124
and sociality 22:855–859
Relationality principle
of identity 6:499
relational mobility 5:427–428
Relationism 21:964

Relations of production 14:650–651, 14:653, 15:715
Relationship competency 24:520
Relationship conflict in groups 10:412
Relationship psychotherapy 20:742
Relationship purchasing 8:25–26
Relationship-based social work ethics 22:784
Relationship-centered care 6:614, 6:618–619
Relationship-focused coping
  chronic illness 3:551
  dyadic coping as 11:593
Relationships 8:745–747, 16:552–553, 23:577–579
  within butch–femme subcultures 13:872
  cohabitation 4:118–121
  customer *see* Customer relationship
  between cyberbullying victims and perpetrators 5:642
  diversity, during adolescence 22:219
  familial/community 9:451
  family and household demography 8:725–729
  and family history 8:739, 8:742–743
  forming and managing, nonheterosexual British Asians 16:864–866
    family relations 16:865
    friendship networks 16:865–866
  grandparent–grandchild 10:336–340
  of hunters-gatherers with others 11:430
  LGBT aging and 13:864–866
    nonconjugal relationships recognition 13:865–866
    older relationships 13:864–865
    partnership recognition 13:865
  in markets 14:560–561
  mentoring 15:236–237
  process, intimacy as 14:375–377
  psychotherapy 19:517–518
  romantic 9:654
  and self-esteem 21:523–524
  social networks 22:461–462
  in Western societies 17:545–552
Relative adoptions 1:137
Relative amount 11:886, 11:888
Relative clauses 13:314, 13:316, 23:891
Relative concentration index (RCO) 21:412
Relative constancy, principle of 16:793
Relative dating (archaeology) 23:744
Relative density measures 3:908–910
Relative deprivation 1:716, 4:165, 6:214–217, 7:502–503, 13:749, 15:467, 18:731
  action 6:216–217
  and collective beliefs 4:165
  and collective violence 6:216–217
  game theory 6:215
  generating mechanisms 6:215–216
  inequality 6:215, 6:217
  justice 6:214–215, 6:217
  and justice 12:949
  mobilization theory 6:216–217
  noncooperative games 6:215–216
  reference group 20:73–74
  revolution 6:214, 6:216–217
  social movement 6:216–217
  status 6:216
  themes and theory 6:214–215
Relative deprivation models 16:443
Relative deprivation theory 4:146, 25:804
Relative effectiveness, evaluation of 19:118
Relative frequency categories 23:767, 23:769
Relative income hypothesis, of child fertility 8:824
Relative income poverty measures 18:742
Relative judgment theory (RJT) 6:396, 19:473–474
Relative language proficiency 21:351
Relative leader–member exchange 13:645
Relative location 14:284–286
Relative mobility 7:903–904
Relative noncentrality index (RNI) 23:618
Relative pluralism 18:237–238
Relative poverty 18:723, 18:756–757
Relative ranking of scores 9:837
Relative rates 15:642–643
Relative risk
  epidemiology 7:829, 7:829b
  logistic regression 2:587–590
Relative spatial terms (language) 13:295
Relative theory 18:103–105
  of democracy 18:468
Relatively egalitarian societies 11:895
Relativism 5:312, 5:484, 7:881–882, 8:72–73, 18:681–682, 19:676, 20:227, 20:233–238, 20:399, 21:308, 21:311, 22:2, 22:225
  arguments about 20:237–238
  cognitive 20:227–232
  cognitive relativism 20:234–235
    about justification and warrant 20:235
    about logic 20:235
    about truth 20:234–235
    epistemic relativism 20:235
  coherence 20:235, 20:237–238
  and common sense 4:1–2
  conceptual/ontological relativism 20:235–236
    about science 20:236
  and constructionism 20:230
  cultural 11:357
  see also Cultural relativism
  cultural relativism 20:234
  and culture 5:582, 5:584–585, 5:587
  definition 20:233
  democratic *see* Democratic relativism
  Dilthey on 6:428–429, 20:233
  early 1:765–768
  and epistemological skepticism 20:227–228
  ethics 20:233–237
  Evans-Pritchard on 20:235
  history 20:233
  and humor 11:416–417
  incommensurability 20:228–230
  Kuhn on 13:135–136
  and law 13:481
  Lévi-Strauss on 13:907–909
  liberalism 14:12, 14:14–15
  linguistic 4:790–792
  Mannheim on 14:499
  methodological 21:311
  moral relativism 20:236–237
  and morality 15:807
  motivations 20:233–234
    cultural diversity 20:234
    disappearance of old certainties 20:234
    situatedness of all judgments 20:233–234
  Nietzsche on 20:233
  objectivity 20:234, 20:236–237
  postmodernism 20:233, 20:236
  rationality 20:234–235
  reasoning 19:591
  semantic relativism 20:237
  and Strong Program 23:595
  time-bounded 21:311
  universalism 20:233–234
  versus universalism 1:765–766
  value 20:233–234, 20:236–237
  value pluralism 25:24
  varieties 20:234–237
  in West *versus* other countries 20:230–232
  Wittgenstein on 20:235
  and world culture 10:229
Relativists 20:230
Relativity, Einstein's
  linguistic, and color terms 4:232
  and objectivity 17:89–90
  Special Theory 17:89
Relativization 16:2
Relativized minimality 16:5–6
Relaxation techniques
  behavior therapy 2:426–427
  in psychoneuroimmunology 19:448–449
Relaxation training 23:552–553
  for asthma 20:579
  for chronic obstructive pulmonary disease 20:579
  for panic disorder/agoraphobia 17:478
RelayRides 8:487
Releasing mechanism 14:367–369
Relevance theory 18:798–799
Relevant social groups 24:136–138
Reliability 3:244, 3:786–787, 3:822, 5:318, 6:345, 8:689, 9:834, 12:250, 12:637, 19:418, 20:248
  assessment 20:239–247
  in criterion-referenced measurement 20:250–251
  diagnostic 6:385–386
  estimation of 20:249
  in ethnographic research 20:251–252
  of evidence, policy making 18:282–283
  implications of 20:249
  implicit association test (IAT) 11:704
  issues for observational measures 20:251–252
  measurement 20:248–252
  projective methods 19:130
  relationship with validity 20:252
  as a signal-to-noise ratio 19:418–419

theory 19:418–419
see also High reliability organizations (HRO)
Reliability assessment 20:239
concepts of 20:239–240
designs for assessing 20:240–246
common factor model representation of CTST 20:242–243
internal consistency reliability 20:242
longitudinal methods 20:243–246
MTMM models 20:243
reliability models for single measures 20:243
simple reliability models 20:241–242
studies of reliability 20:241
test–retest reliability 20:243
validation studies 20:241
Reliability coefficient, classical true score model 19:424–425
Reliability from single administration of test, parallel measurements model 19:427
Reliability of expert testimony 8:584
Reliability of indicators 11:752–753
Reliability of scores, SJT 22:15
Reliability ratio 14:851
Reliability seeking organizations (RSOs) 17:373
Reliance, and trust 24:668–670, 24:672
Reliance on science 7:797–798
Relics of sainthood candidate 15:580
Religion and Cyberspace 20:275–276
Religion and Violence e-Learning Project 20:256
Religion of humanity 4:522–523
Religion on the Internet: Research Prospects and Promises 20:275
Religion Online: Finding Faith on the Internet 20:275
Religion thesis
decline of 21:385–386
privatization of 21:386–387
Religione primordial 18:437–438
Religions 2:507–508, 2:887–888, 3:719–720, 15:412, 16:224, 18:389–390, 20:322–327, 20:338–343, 20:396–399, 21:388–389, 21:391, 22:149
and abuse 20:253–257
and adolescent prosocial behavior 19:224
Albert on 1:499
in America 2:508
and American/French Revolutions 1:598–600
and ancestors 1:688–689
anthropology of 2:496, 20:307–312
belief 20:308
body 20:309–310
cognitive approaches 20:311
definitions 20:307–308
emergence of subfield 20:307
globalization and diaspora 20:311
language 20:310
rationality 20:309
ritual 20:308–309
secularism 20:308–311
witchcraft 20:310
world religions 20:310–311
approach to peacemaking 17:637–638
and Arab nationalism 16:271–273
of archaic states 23:374
in Arctic 1:929–930
of Australian aboriginals 2:236
autonomy of 3:575–576
behavioral health 20:265–266
cigarette smoking 20:266
diet 20:266
exercise 20:266
sexual behavior 20:266
weight 20:266
belief 2:496
Buddhism 18:389–390
cargo cults 3:142
causal processes 17:805–807
environmental influence on religiousness 17:805–806
genetic influence on religiousness 17:805–806
personality influences on religiousness 17:805–806
personality-religion interaction 17:807
religious influences on personality 17:805
in Central Asia 3:319–320
in China 2:65–66, 2:68–69
Christianity 20:340–342
clinical application 20:268
collective beliefs 4:165
collective nature of 20:328
collective violence 20:338
during colonial period 3:536
and common sense 4:1, 4:4
conflict, influence on 20:281
before the conquest 3:534
Constitution's religion clauses 20:280
for control over collective resources 20:328
and cultural heritage 5:417
as cultural system 14:437
defining 20:263
definitions 20:279
democratic roles of 20:279–280
disconnection of 20:331
diversity and revivalism, in Southeast Asia 23:70–71
of East Asia 20:366–372
East Asian studies 6:786–787
and economic development 5:528, 5:531, 5:533
and economic life 20:258–262
capitalism 20:261
modern capitalism and its critique 20:259–260
precapitalist configurations of 20:258–259
reflective religious critiques of 20:260–261
and Enlightenment 7:664–666
enlightenment in 7:657–658
and ethnogenesis 8:176
ethnographic study of 8:289–293
of Europe 8:263
evolutionary group dynamics 8:444
and family 20:317–321
child-rearing 20:319–320
coping with family problems 20:320
divorce 20:320
forming family relationships 20:318–319
home and work balance 20:319
marriage 20:318
maternal fertility 20:318–319
paternal involvement 20:319
knowledge about family relationships 20:318
sustaining family relationships 20:319–320
decreasing risk of domestic violence and sexual infidelity 20:319–320
disciplining and nurturing children 20:320
enhancing marital quality 20:319
risk of child physical abuse 20:320
understanding of prevention and intervention 20:320–321
and fertility transition 9:51–53
foreign 2:892
functions of 3:574–575
fundamentalism 20:342–343
and gender 9:782–783, 9:785–786
and gender development 9:746
genocide, history of 10:10
global history 10:162–163, 10:165–167
globalization and 10:213–217
in globalized cultural context 10:231–232
government control of 2:892
Habermas on 10:472
and health 20:263–272
end of life care 20:271
enhancement of social support networks 20:271
public health consequences 20:269
reduction of behavioral risk factors 20:270
religious coping 20:271
religiousness and health outcomes 20:271
research in 20:271–272
Hindu 18:389–390
history 20:322–323, 20:348–349
historiography 20:323–325
new directions 20:352
rationale for rapprochement between 20:349–351
religious struggles 20:350
spirituality 20:350–351
and hybridity 11:449, 11:452
identity 20:338–341
and identity 11:522
as immorality and deviance 20:336
intellectual reflection 20:323
and international advertising 12:456
Islam 18:389–390, 20:338–342
and Islamic social work 12:754–755, 12:757
in Japan 12:796

Religions (*continued*)
and joking 11:418
Judaism 11:67–69, 11:71
and justice 12:938
Lévi-Strauss on 13:904–906
liberal Christianity 14:8–10
and magic 14:436–437
and mental health 20:263–265
and metal disorder 20:264
alcohol and drug use/abuse 20:264
anxiety 20:264
depression 20:264
psychosis 20:264
suicide 20:264
in Mexico 1:629
in Middle East and North Africa 15:356–357
missionary 2:886
mobilization and power 20:328–332
modernization of 2:892
and morality 15:807–809
morality and 20:334–336
morality and social control 20:333–337
and motherhood 15:881, 15:883
myth in 16:172–178
and nationalism 16:295
and nationalism in South Asia 16:280–281
Near Eastern/North African studies 16:376–380
Neo-Marxist deprivation compensation theories 25:635
and new media 20:273–278
authority 20:273–274
community and identity 20:274
current approaches to study religion 20:276–277
future areas of research 20:277
and Internet 20:275–276
online culture 20:273–274, 20:276–277
religious practice 20:273–277
ritual 20:274–275
social media 20:274, 20:276–277
use 20:273–275
and personality 17:801–808
personality-religion links 17:801–805
age 17:804–805
behavioral implications 17:803–804
cognitive preferences 17:803
culture 17:804–805
fundamentalism 17:804
gender 17:804–805
implications of 17:803–804
intense religious change 17:804
meta-stereotypes 17:803
personality differences 17:803
personality models 17:802
personality traits 17:802–803
as predicting psychological outcomes 17:807
religiosity 17:802–803
robustness 17:803–804
spirituality 17:804
stereotypes 17:803
traits 17:802
phenomenology of 20:344–347
analytical approaches 20:344–345
descriptive approach 20:344–345
social accounts 20:344, 20:346–347
physical health 20:266–268
Alzheimer's disease 20:267
cancer 20:267
cerebrovascular disease 20:266–267
dementia 20:267
endocrine function 20:267
heart disease 20:266
hypertension 20:266
immune function 20:267
mortality 20:267
political behavior and 20:281–282
political mobilization of religious resources 20:330
and politics 20:401, 20:404–405
and positive emotions 20:264–265
hope 20:265
meaning and purpose 20:265
optimism 20:265
self-esteem 20:265
well-being 20:265
in precapitalist societies 20:328
and prosocial behavior 19:240
and psychiatry 19:307
beliefs about mental illness 19:307–308
beliefs among psychiatrists 19:308–309
Buddhism 19:308
Christianity 19:307
contemporary psychiatric practice 19:309–310
diagnosis 19:311
help seeking 19:310–311
Hinduism 19:308
improving practice 19:310–311
Islam 19:307
Judaism 19:308
meditation 19:308
sin 19:307–308
spirituality 19:309, 19:311
stigma 19:308, 19:310
psychiatry and 20:348–353
attachment theory 20:351
attachment to god 20:351
coping 20:349–351
ethics 20:349
new directions 20:352
psychology of 20:354–358
purification from science 9:106–107
and racism 19:852–854
rationalization of 15:577
religion and everyday life 3:537
religion in new republics 3:536–537
religious beliefs and environmental attitudes 20:282
religious characteristics and actual behaviors 20:335
religious institutions and politics 20:280–281
religious movements 20:281
religious political movements 20:281
religious symbols 20:279
research trends 17:801–802
right to 3:687, 3:689
rituals *see* Rituals
and sacrifice 20:860–862
salvation 2:886
versus science 15:575–576
and sciences 20:284–291
biotechnology 20:287–288
Copernican Revolution 20:285–286
Draper–White thesis 20:284
ecology and environmental movement 20:288
eugenics movement 20:287–288
Galileo's theory 20:285–286
geology and theological considerations 20:286–287
in Greece 20:284–285
historiography of 20:284
mechanical philosophy and natural theology 20:286
during middle ages 20:285
modernist–fundamentalist controversy 20:287
Muslim contribution 20:285
in Rome 20:284–285
twentieth-century physics and cosmology 20:287–288
secular religions 21:380–382
separation of church and state 20:280
and sexuality 21:804–805
'shamanic' form of 2:888–889
shamanism 21:868–869
social control and 20:334–336
social health 20:265
altruistic activities 20:265
delinquency/crime 20:265
marital stability 20:265
social capital 20:265
social support 20:265
as a social institution 20:328
and social movements in modern societies 20:329–330
and social stratification 20:292
adult family processes 20:293–294
childhood family processes 20:293
contraception, use of 20:293–294
divorce 20:293
education levels 20:293
fertility behavior 20:293–294
gender differences in work behavior 20:294–295
human capital 20:294
immigration 20:295–296
income level 20:295
marriage and cohabitation 20:293–294
nonreligious people 20:296
physical and mental health 20:296
religion as acculturation resource centers 20:295–296
sexual activity, approaches to 20:293–294
status attainment and life course traditions 20:292
theoretical approaches 20:292
wealth inequality 20:295, 20:295t
work behavior 20:294–295

sociology of 20:359–365, 25:463, 25:465–466
in South America 23:46
in South Asia 2:76–77
and spirituality 23:285–289
spiritually integrated treatment 20:351–352
theoretical approaches 20:351
therapeutic directions 20:351–352
transformation 20:351
treatment effectiveness 20:349–352
and state 3:577, 13:219
and state, separation of 3:574–577
and stratification 20:292–297
and suicide 23:684–685
suicides 20:269
unintended health consequences of religious practices 20:271
well-being 20:271
symbolic boundaries 23:854
as symbolic system, genealogical critique of 23:862–863
as system and ideology, Arab nations 16:379
and taboo 24:1
Taoism 18:389–390
theories of monotheistic religion 1:766
totemism 24:455–457
in the United States 20:279–283
use of term 20:298, 20:359
and values 25:30–31
versus religiosity 20:361
Western-derived exclusivist definition of 2:890
witchcraft 25:604–605, 25:608–609
and worldview 25:755–756
and youth 20:298–306
cognitive development 20:301
conversion 20:301
ethnicity, class and gender 20:303–304
genetic and birth order influence 20:303
multiple identities 20:300–301
parents, influence of 20:302–303
pilgrimages 20:300
religion, effects of 20:303
religiosity 20:298
religious institutions, role of 20:302
religious service attendance 20:301–303, 20:299f
religious socialization, agents of 20:301–302
schools, influence of 20:302
spiritual development, trajectories 20:300–302
spirituality 20:298
trajectories of religious involvement 20:300–302
Zionism 20:341–342
Religion/spirituality (R/S) 20:263–264
and coronary heart disease (CHD) 20:266
Religionswissenschaft 20:344
Religiosity 17:233, 17:802–803, 20:298, 20:361, 20:377
across generations in Britain 20:300
and agreeableness 17:805
beliefs, identities, and personal religiosity 20:378–379
and conscientiousness 17:805
dimensions of 20:298, 20:381–383
ethnicity and 20:379
as explanatory factors 20:364
extrinsic religiosity 20:382–383
gender and 20:380
intrinsic 20:382–383
macrolevel dimensions of 20:379
microlevel processes and 20:379–380
against minorities 20:381–383
personal characteristics and 20:302–303
as religious behavior 20:378
religious traditions and denominations 20:377–378
research, future directions in 20:380
risk-aversive people and 20:303
of societies 20:360
trends in youth 20:298–300
of women 25:635–637
Religious, forms, studies of 21:381–382
Religious actors 17:651
Religious apps (smartphone) 20:276–277
Religious association, forms of 3:351
Religious behavior 20:391–393
Religious beliefs 4:164, 20:292–293, 20:364, 20:384–395
anthropology 20:385–386, 20:385f
cognitive science of religion 20:389–393
comparative religion 20:387–388
early theories 20:384–385
evolutionary psychology 20:389
evolutionary theory 20:384–390, 20:393
experimental psychology 20:389, 20:392
history of religions 20:384, 20:387
psychology of religion 20:384–385, 20:389
and religious delusion, differentiation between 20:350
sociology 20:386–387
Religious boundaries 18:146–147
Religious broadcasting 20:362–363
Religious censorship 3:279
Religious change, in Central America 1:624–625
Religious charisma 3:355–356
Religious classification 20:380
Religious cleavage resilience, and voting 25:304
Religious commitments 20:295
Religious communities 19:616–617
and communalism 4:263–265
Religious conflicts 20:336
Religious control 20:328–330
against British colonial control 20:329–330
by religious organizations 20:330
Religious diversity 20:361
impact on religious involvement 20:364
Religious dogmatism 20:465
Religious education 7:213–214
Religious evolution 2:507–508, 20:313
archaic religion 2:507–508
during Axial Age 20:313–314
causes of 20:315–316
directional changes 20:313
early modern religion 2:507–508
historic religion 2:507–508
modern religion 2:507–508
primitive religion 2:507–508
Protestant ethic 20:313
Protestant Reformation 20:314–316
shape of 20:314–315
Religious experiences
defended against scientific skepticism 18:809
phenomenology of 18:809
Religious experts 8:588–589
Religious families 20:256
conflict 20:254
Religious field
of practice 23:864
social rules in 4:3–5
Religious freedom 18:239
rationalism and liberalism 14:23
Religious fundamentalism 20:396–400
Religious habitus 10:484
Religious history, laicization 13:216–219
Religious homogamy 20:293
Religious homogeneity 20:335–336
Religious identity 11:528–529, 20:338–341
minority 8:145
transnational *see* Transnational religious identities
Religious identity movements 11:544
Religious ideologies 20:328
framing of political issues and 20:330–331
resources 20:328
Religious institutions 20:302, 20:328
agents of 20:328
alignment with political and economic elites 20:328–329
positive effects of 20:303
religious schools and temples 20:330
Religious international nongovernmental organizations (RINGOs) 20:411
Religious intolerance, and racism 19:852–853
Religious leaders
abusive 20:256–257
response to abuse 20:254, 20:256–257
role in intimate partner abuse prevention 12:682
Religious liberty 13:216–217, 13:219, 18:240
Religious life 15:724
'contemplative' and 'active' styles of 15:725–726
Religious men, abusive 20:255
Religious mobilization
in precapitalist states 20:328–329
religious schools and temples, role of 20:330
trajectories in 20:331
Religious movements 19:196
Religious narratives 20:330–331
Religious nationalism 3:680–681, 20:401–405
Christian radicalism 20:402, 20:404
future of 20:404–405
in global era 20:403–404

Religious nationalism (*continued*)
globalization 20:401–405
Islamic radicalism 20:403–405
Jewish radicalism 20:404
and politics 20:401, 20:404–405
religious violence 20:402–404
secularism 20:403
terrorism 20:403–405
Religious neutrality 20:374
Religious nongovernmental organizations 16:858–859
Christian form 16:859
Islamist forms 16:859
and secular civil society 16:858–859
social and development services of 16:858–859
Religious orders 3:234, 3:236
Religious organizations 20:362, 20:406–412
authority 20:408–410
Buddhist monastic institutions (sangha) 20:408
Christian organizations 20:408
church, basic concept of 20:407
concepts of sect and cult 20:407
distinction between nonreligious organizations and 20:406
diversity and distinctiveness 20:406–407
dynamics of 20:408
economic underpinning of 20:408
forms of governance 20:409
fundamentalism 20:409–410
history 20:407–408
leadership 20:408, 20:410–411
local 20:362
mosque committees 20:408
participation in ecumenical or interfaith activities 20:407
products, and production processes 20:406
sects and cults 20:407
types and subtypes 20:407–408
Religious participation 20:377–380
Religious persecution 13:216
Religious phenomena 5:550–551, 20:322–323
and cognition 5:577–578
Religious pluralism 3:540–541, 3:543, 3:681–682, 10:215, 15:616–617, 20:407, 24:589–590, 24:604
Religious power 20:328
in precapitalist societies 20:328–329
of religious institutions 20:329–331
of religious movements 20:330
trajectories 20:331
Religious reveries 16:170
Religious right
and conservatism 4:690
and gay/lesbian movements 9:634–635
Religious ritual practices 20:271
Religious rituals, and beliefs in Morocco 16:377–378
Religious saliency 20:382–383
Religious schemata 20:328, 20:331
Religious schools 7:299, 20:380
Religious self-identification 16:295
Religious service attendance 20:299, 20:301, 20:299f
Religious socialization
agents of 20:301–302
parental influence 20:302
peer influence 20:302
Religious sociology 14:825
Religious specialists 20:362
Religious struggles 20:350
Religious studies, of syncretism 23:881–882
Religious survivors 20:253, 20:255
Religious symbols 23:861, 23:863
belief *versus* practice 23:863
Religious tolerance 8:143, 13:217, 18:238–240
Religious traditions 20:313
Religious upbringing, in collectivism/individualism 4:208
Religious violence 20:402–404
Religious wars *see* Wars of religion
Religious women, unique vulnerabilities facing 20:254–255
Religiously sensitive care 20:350–351
Religiousness 17:801, 20:298, 20:356
relationship to health and well-being 20:349–350
Relish, Michael, on genre knowledge 16:556
Relocation 14:289–290
and disasters 6:105
Relocation/concentration of Native Americans 1:634–635
RELs *see* Regional Educational Laboratories (RELs)
REM *see* Rapid eye movement (REM) sleep; Reciprocal effects model (REM); Retrieving effectively from memory (REM)
Remaining reserves 17:166
Remapping control, resources 20:573–574
Remarriage
and divorce 14:522
in Western societies 17:550
Remarried families, emotional development of children in 7:440–441
Remediation, of language learning impairment 13:348–349
Remedies 3:29, 14:266
adjudicator and 20:417
assessment and comparison of damage awards and 20:417
compensatory damages social scientific study and 20:414–415
award levels and trends 20:414–415
compensatory awards reliability 20:415
general adequacy of damages 20:414
time trends 20:415
damage rules across countries and 20:413
nonmonetary relief and 20:413
punitive damages and 20:415–417
levels of awards 20:416
rate of awards 20:415–416
relationship with compensatory awards 20:416–417
time trends 20:417
Remedies model, legal reasoning 13:782
Remembering 8:710–711
Remembering, group 10:426
Remembrance
death/dying 5:868
localization of 4:183–184
as social construction 1:881–882
Remilitarization of Rhineland 16:259
Reminders, and prospective memory 19:277
Reminders of money 15:737
Reminiscence bumps 2:284–286, 4:191–192
Remissoriali 15:578
Remittances 6:271
in Central America 1:622–623
in Chinese international migration 12:528
internal migration 12:433, 12:439–441
and international migration 12:519–524
peasants 17:678
REML *see* Restricted maximum likelihood (REML)
REML-EB *see* Restricted maximum likelihood-empirical Bayes estimation (REML-EB)
Remote management&rsquo 24:107
Remote memories 1:648
Remote memory impairment 24:558–560
autobiographical memory loss and theoretical models of memory 24:560
biology 24:559
clinical features 24:559–560, 24:560f
cognitive architecture 24:559
Remote Oceania 17:446
Remote sensing 10:53, 10:64–68, 20:419–424
analysis with 10:64
application to social sciences of environment 10:67–68
environment and society 10:68
land use and land cover 10:67
linking household-level data with remote sensing data 10:67
spatial modeling 10:67–68
spatial stratification/extrapolation 10:67
in archaeology 10:98
and environment 10:64–66
data development 10:66
platforms 10:65–66
spectral, spatial, and temporal domains 10:65
geographic information systems 20:420
hazards 10:68
historical development 20:420–423
image interpretation 20:420
and land cover 13:221–222
natural resources 10:64
other techniques of 10:54
resolution 20:420
satellite 10:64–66
sensor systems 20:419–420
Remuneration
and job evaluation 12:816
Ren, Xuefei, on Chinese cultural industries 5:597–598
Renaissance 1:765, 2:646–647, 7:656–658, 9:388, 22:589–590

and classical archaeology 3:760–761
cultural physiognomy 20:426
and cultural production 5:466
development of concept
medievalists, revolution of 20:426
Renaissance 20:425–426
genesis, research problem of 20:428
mannerist 20:428–429
and modern individualism 20:427–428
art historical findings 20:427
renaissance humanism 20:427–428
revolution in discourse 20:427–428
modernity 15:682
political theories 3:62–64
psychology during 19:409
and scientific revolution 21:319–320
sociohistorical dimension 20:426–427
Renaissance art 20:426
Renaissance Florence 16:564
Renaissance humanism 20:427–428
Renaissance literature, personality in 17:909
Renaissance people 20:427
Renaissance Utopian literature 14:253–254
Renaissance women 20:427
Renan, Ernest 22:889, 24:603
works, *What is a Nation* 22:889
Rendition/extraordinary rendition 13:533
Renewable energy 7:618–620, 7:636–638
political economy of 18:367–373
climate policy 18:372
development economics 18:368
development in Europe and world 18:372–373
emissions trading 18:369–370, 18:372
feed-in tariffs 18:369–370, 18:372–373
global surge of photovoltaics 18:370
idea 18:367
impact of RES-E deployment on fossil and nuclear electricity incumbents 18:372
issue of market creation 18:368–370
liberalization 18:369, 18:372
neoclassical economics 18:368
political efforts to slow down the shift 18:372
purchasing obligation 18:369
R&D funding 18:367–368
renewable portfolio standard 18:370
renewables obligation 18:369
subsidies for nuclear and coal power 18:367–368
technological innovation system 18:370
technological learning 18:369–370
Renewable Energy Act 18:372
Renewable energy sources electricity (RES-E) 18:370
impact on fossil and nuclear electricity incumbents 18:372
Renewable portfolio standard (RPS) 18:370
Renewable resources 18:566–567
Renewables obligation (RO) 18:369
Renewal 8:850–851
of capitalism 3:95
strategic adaptation 23:504
between trade unions and social work practice 24:491
Weberian social thought 25:466
Renewal effect, in context dependent memory 4:779
Renewal models 23:471
Rent control, housing 11:244
Rent gap theory 10:48
Rent seeking 19:537
and property rights 24:529–530
regulatory agencies 20:182
Rentiers 17:505
Renting 11:234, 11:237
Rents
defined 18:358–359
wage differentials 25:339
Rent-seeking behavior 20:161
Renunciation
categories of 25:633
Illich's contribution 11:584–587
Renyao 21:738
Renzulli Learning System (RLS) 7:162
Reorganization 2:354–355, 20:561–562
brain, evolution of 2:827, 2:830–832, 2:832t
Reorganizational stage of aging 1:436–437
Reorientations 23:634
and posttraumatic amnesia 18:638–639
Repair
of speech errors 23:257–258
word 4:854–855
Reparations 13:68
class actions 3:736–737
Repartnering 20:430–433
fertility of couples with children from previous relationships 20:432
and marital history 20:431
of parents 20:430–432
stability 20:430–432
in Western societies 17:550
Repatriation 9:966, 16:145–147
archaeology 1:885–890
cultural resource management 5:489–490
Repatriation paradigm 9:966
Repeasantization 17:678
peasant agriculture 17:665
Repeat migration 15:436
Repeated best and/or worst choice model 2:550–551
Repeated cross-sectional analysis, for political data 18:515, 18:520
Repeated measurement data, discrete multivariate analysis 16:135–136
Repeated measures, linear mixed models 14:162–168
Repeated measures design 5:349–350, 5:352–353
Repeated phoneme effect 18:56
Repeater classifier 3:817
Repertoires of contention, notion of 22:508
Repertory grid technique 4:724–725
Repetition 13:273, 13:276, 13:278
Repetition priming 11:710–712
bias theory of 11:711
long-term 11:709
Repetitive and restricted behaviors (RRBs) 2:276
Repetitive behaviors
see Obsessive–compulsive disorder (OCD)
Repetitive drumming and/or dancing 21:868
Repetitive publications 20:530
Repetitive sex offenders
treatment of 20:434–440
Repetitive transcranial magnetic stimulation (rTMS) 13:490, 23:19
LTP/LTD 9:507, 14:363
Repgrids 4:724
Replaceability, and war 25:366, 25:369
Replacement fertility 9:775, 18:572–573, 18:575–577
Replacement rates 11:25
Replacement workers 23:591
Replicability 16:81–82
Replication 2:484, 9:915, 20:441–442
exact and approximate 20:441
in genetic studies of behavior 9:937
intersubjective testability 20:441
publication bias 20:441–442
reproducibility 20:441
in social and behavioral sciences 20:441–442
type-I error 20:441–442
Replicative entrepreneurs 7:676
Replicator
evolution of language 8:367–368
Replicator dynamics 8:388–389
Replicator dynamics, Metcalfes use of 8:431
Reported speech, and narratives 16:219, 16:221
Reporting 3:823–824
meta-analysis 15:269–270
Reporting biases, in health behavior assessment 10:589–590
Reportive markers 8:353–354
Repositioning, queer 19:766–767
Representational crisis, micromethods 12:650
Representational economy 9:872
Representational mapping 16:507–508
Representational measurement theory 17:296, 17:298
Representational realism (Locke) 20:21
Representational schemes 16:504
Representational set 16:506–507
Representational system 16:507–508
Representational theory of measurement (RTM) 14:856–861
applications of 14:860–861
automorphisms of measurement structures 14:859
formulation of measurement within 14:856–857
meaningfulness 14:859–860
ordered relational structure 14:857–859
problems of 14:860
scale type 14:858–859
theorems 14:857–858
comments on the axioms 14:858

Representational theory of measurement (RTM) (*continued*)
difference measurement 14:857–858
extensive measurement 14:857
Representational theory of mind, in primates 10:268–269
Representationalism, Rorty's contributions 20:762
Representationism, measurement theory 14:870–871
Representations 3:628–629, 8:927, 12:342, 12:369, 16:146, 16:506–507, 19:332–334, 20:67–71, 20:797, 22:438, 25:207–212
artificial intelligence
hybrid models 2:38–39
symbolism 2:35–36
belief 20:68–70
British cultural studies 2:853–854
and change 12:344
concept learning 4:538–541
category learning 4:540–541
connectionist 4:87–88
content 20:67
and context 20:70–71
cultural geography 5:409–412
explicit 13:78–79
in feminist research 9:719
format of mental images 15:220–221
general notion 20:67–68
implicit 13:78–79
information 4:87–88, 20:67–68, 20:70
and intelligence 12:316
knowledge 4:85
knowledge representation 13:98–104
of a life 2:644
linguistic 20:68–69
content and logical equivalence 20:69–70
and language of thought 20:70
map theory 20:70
mathematical models and 14:793–794
mental 20:67–70
neural 4:88
no need for 12:344
political 18:441
possible worlds 20:67, 20:69
and postmodernism 18:684, 18:686, 18:688
proportional 19:205
of queer anarchists 19:749–750
and resources 12:344
in science 25:220–221
of space 4:487
of time 4:486–487
Twin Earth 20:68–69, 20:71
see also Social representations
Représentations collectives 6:701
Representative, political 18:442
Representative democracy 2:717–718, 6:63–64, 6:91–92, 6:94, 22:635–636
Representative firm 20:727
Representative fraction (RF) 21:1
Representative government 19:536
Representative publicity of monarchs 19:615
Representative research 17:763–764
Representative sample, mental illness epidemiology 15:179, 15:183–184
Representative sampling, for sexuality 15:316
Representative survey 7:564
Representative-agent model 14:403–404
Representativeness 11:112
content validity 4:774–775
of samples 9:275
Representativeness heuristic 10:826
of decision making 5:919–920
Represented, in political representation 18:442–443
Represented emotions 16:658–659
Repressed traumas, transference 24:545
Repression 3:688, 19:330–331, 19:334, 24:524
and free association 24:724–725
of legal mobilization 13:594–595
political 18:446–452
Repressive hypothesis 21:740
Repressive liberalism 11:673
Repressive regimes 16:779
Reproducibility
peer review 17:680, 17:682
publication 20:441
Reproduction 8:810–812, 9:40, 11:7, 20:457–458, 20:461, 20:464
animal's life history 9:40
cost of 9:40–41
and cultural anthropology 20:450–456
1980–2000 20:451–454
abortion 20:450–453
adoption 20:451–453
biology 20:450–454
biopower 20:452–453
birth 20:450–454
body 20:450–454
capitalism 20:450–453
comparison 20:453–454
concept, construction of 20:454
exchange 20:450–451
feminism 20:450–451, 20:454
gender 20:450–453
individual 20:450–454
medical anthropology 20:452–454
nature 20:450–454
nature of reproduction and kinship culture 20:450–451
political economy
and reproduction 20:452
production 20:450, 20:452, 20:454
science technology studies 20:453–454
sex 20:450–453
vital statistics 20:451–452
evolutionary demography 8:413–418
and fertility 15:397–399
genetic programming 2:41–42
human behavioral ecology 11:249–250
life history 11:250
parental investment 11:249–250
parenting/mating effort 11:250
and marriage systems 14:611–612
and motherhood 15:881–882
and population dynamics 18:562–564, 18:579–580
reproductive medicine 13:40–41
sociology 22:896
Reproduction studies, anthropology and demography in 9:5
Reproductive and sexual rights 20:457
historical trajectories 20:457–460
political and conceptual inquiries 20:464–465
human rights 20:464
indispensability and insufficiency of rights 20:464–465
religious dogmatism 20:465
reproduction 20:464
science 20:465
secularity 20:465
sexuality 20:464
2000s policy and political landscape 20:460–464
United Nations policy arenas 20:459–460
Reproductive assimilation 16:484
Reproductive behavior, regulation by hypothalamus 11:489
Reproductive biology, and warranties 17:173
Reproductive choice, in affluent nations 20:473–474
Reproductive decision making 9:998
Reproductive health 10:462
and disasters 6:105–107
and family planning 8:786–788
Reproductive health services 21:828–830
Reproductive justice, in affluent nations 20:470–471
Reproductive memory 20:50
Reproductive rights 8:233, 19:825, 20:457–467, 21:792
in affluent nations 20:468–476
assisted human reproduction 20:472–473
assisted reproductive technologies 20:468–469, 20:471–474
contraception and abortion 20:469–470
fertility preservation 20:473–474
fetal personhood and maternal agency 20:470
LGBTQ reproductive rights 20:471–472
preimplantation genetic diagnosis 20:472, 20:474
prenatal diagnosis 20:474
reproductive choice and prenatal genetic testing 20:473–474
reproductive justice 20:470–471
women with disabilities 20:471
ecological economics 6:860
family planning programs 8:795
in Latin America and Caribbean 9:71
Reproductive strategies of male and female primates 18:897–898
Reproductive success, and life span development 14:108–110
Reproductive technologies
and fertility 9:24

and gender 9:723–724
genealogy 9:806, 9:808–809
and new genetics 9:1016–1017
Reproductive theory 10:851–852
REPs *see* Die Republikaner (REPs)
Republic 18:380
and solidarity 23:2–4
Republic of China (ROC) 3:515
The Republic of India *see* India
Republic of Maldives *see* Maldives
Republican 3:643
Republican freedom 9:403
Republican People's Party (CHP) (Turkey) 20:873
Republican political reeducation 9:409
Republican Revolution of 1994 1:605–606
Republican virtue 3:772
Republicanism 1:699–700, 4:317–320, 8:158, 18:380, 18:487–488, 20:397, 20:477–482
Adams on 20:477
Aristotle on 20:477
checks and balances 20:479
Cicero on 20:477–481
and citizenship 3:650–652
civic 9:402–403
common good 20:478
commonwealth 20:477–478
constitutionalism 20:477–479
deliberation 20:479–481
deliberative democracy 20:479–481
deliberative politics 20:479–480
democracy 20:478–480
and democracy 6:62–63
freedom 20:480–481
Habermas on 20:477
Hamilton on 20:479
Harrington on 20:477
justice 20:477, 20:479–481
Kant on 20:478
liberty as nondomination 20:478
Machiavelli on 20:477–479
and mass communication 14:691–692
neorepublican ideals 20:480
neorepublican institutions 20:480
Paine on 20:481
people 20:480
Pettit on 20:477–480
philosophy 20:481
Plato on 20:477
popular sovereignty 20:479
public good 20:477–478, 20:481
republican form of government 20:477–478
revival 20:478
rule of law 20:479–480
Sandel on 20:895, 20:897
tradition 20:477–478
République des lettres 20:529
Repugnancy clause, in sub-Saharan Africa 23:656–658
Repugnant conclusions, and intergenerational justice 12:398–399
Reputational cascades theory 8:107–108
Reputational networks 20:483
Reputations 5:443, 14:431–433
of artists 2:56–57
artists, and social networks 5:468–469
ascertaining 20:484–486
indicators 20:484–485
Craik model 20:483
dimensions of 20:483–484
disparity between two groups 20:490
effects of 20:486–488
factors affecting survival of 20:490
in interpersonal relationships 5:426–427, 5:430–431
level of eminence 20:484
overview 20:483–484
production of 5:611
and public recognition 2:332
trajectories 20:488–491
'What?' question 20:484
Rereligization, and Eastern European social work 22:796, 22:800
Rerum Novarum 20:260
RES *see* Regional Economic Studies (RES)
Res cogitans 2:728
Res extensa 2:728
Res judicata 13:618
Res Militaris 15:515
Resacralization 18:434
Resampling methods 19:432, 20:492–495
bootstrap method 20:492–494
growth curves 20:494
hypothesis testing 20:492
linear regression 20:492
related methods 20:494–495
Re-scaling of territoriality 10:241
Rescorla–Wagner model 3:768
Rescue Takeover Act 10:295
RES-E *see* Renewable energy sources electricity (RES-E)
Research 10:864, 22:588–594, 22:604, 22:606–610, 22:613–614, 24:763
accuracy 17:739
adolescent prosocial behavior 19:224–225
aesthetic education 1:231, 1:233
agenda-setting 1:357
aging 6:885
animal rights in 1:703–707
anthropology of education 1:730–731
armed conflict 1:972
attachment
methods and problems 2:139–140
audience measurement 2:211–214
autism spectrum disorders 2:279–280
autonomy and dependency in adulthood 1:183–184
bargaining and negotiation 16:417
Bielefeld School 2:556–557
body history 2:730–731
Bologna Process 2:742–743
on broadcasting regulations 2:862–863
on bureaucracy 2:911
burnout 2:929–930
interventions 2:931–932
business platforms 3:40–41
on butch identity 13:871
butch–femme subcultures
methodological issues 13:872–873
cancer screening 3:80–81
causal, in non-western societies 3:248–249
censorship in 3:279–283
child soldiers 3:435–436
class size 3:752
quasi-experimental research 3:752–753
climate change 3:853–854
Cold War 4:133
collaborative consumption 4:142–143
communication and transportation 4:275–276
comparative health systems, benefits and limitations of 4:401t
in complementary and alternative medicine 10:564
computational organization theory 4:488–490
multilevel modeling 4:489
multimodeling and model interaction 4:488–489
rapid model development 4:489–490
computers 4:518
conflict mediation 4:628
constraint satisfaction 4:722
consultation–liaison psychiatry 14:5–6
integration of medical and psychiatric services 14:6
consumer 4:740–742
on consumer behavior 4:739–740
control behavior 4:817
control variable in 4:840–843
counseling psychology 5:70–71
creativity and personality 17:770–771
in critical social work 5:291
cross-cultural training 5:337–338
focus on culture-specific and culture-general training 5:337–338
future 5:340
inclusion of contextual and personality variables 5:338
non-expatriate research 5:337
postdeparture training 5:338
cultural intelligence 5:438
definition of 20:509
definition of sexuality as object of 15:314–315
dehumanization
explicit 6:32–33
subtle 6:34–35
dichotic listening 6:357–358
dramaturgical analysis 6:649
drug use/abuse 6:688
education 5:348
see also Educational research
emotions 7:462–464
ethnohistory 8:196
on euthanasia 8:283–284
evidence-based policy making
evidence 18:286
experimental laboratories 8:556–558
context requirement 8:557
measurement devices 8:557–558
tools 8:557

Research (*continued*)
feminist *see* Feminist research
on femme identity 13:871
fertility recovery 9:778–779
fields analyses, and epistemic cultures 7:876
fieldwork, transformations in 9:151–152
findings from ecological analyses of health 6:880–881
forced-choice personality assessment 17:848
forensic archaeology 9:337
funding 9:552–553, 22:658
gaps in early friendships 9:450, 9:452
Genographic Project 10:22
genomics 10:39
geriatric psychiatry 10:107
goal pursuit 15:892
grammaticalization 10:331–332
group 10:435–436
health, Africa 1:280–281, 1:284–285
health, lesbians as topic of 15:114–115
health care 10:599–601
higher education 10:862–869
homelessness in United States 11:161–162
homosexuality 3:889–890
homosexuality and mental illness 3:889–890
sexual orientation change efforts 3:889
theories on etiology of homosexuality 3:890
implications, school environment for LGBTQ/sexual minority youth 21:89
infant development 18:855–858
infrastructure 22:667–673
for international advertising 12:456–457
I–O psychology
factors driving demand 11:883
role in guiding science and practice 11:882–883
job evaluation 12:816–817
knowledge 13:88–89
labor supply 13:194
LAC educational system 7:181–184
on leadership in organizations 13:668–669
legal
databases 5:803–810
liberation theology 14:37
life courses 14:78–79
life span development 14:105–106
linguistic anthropology 14:170–174
mass imprisonment 14:696, 14:700
media audience 2:220–221
feminist 2:218–219
third generation 2:217–218
media history 15:15–16
media imperialism 15:55–56
methods, educational sociology 7:291–293
methods, social psychology 22:544–552
methods, sustainable practices 23:834, 23:837
migration history
concepts 15:446–447
tasks 15:447
modes, infrastructure 12:123–124
of neighborhood effects 16:446–447
on neonatal imitation 16:484–487
neuroesthetics 16:658–661
of new genetics 9:1016–1018
governance 9:1017–1018
of obedience 17:61–66
older age work participation 17:187–188
opinion, globalization of 19:566–567
organizational climate 17:324–325
organizational diversity 6:599
party identification 17:565–566
on persistence 10:397–398
perspectives, procrastination 19:75–77
perspectives on Internet and culture 12:587–588
philanthropy, emerging trends in 3:368–369
policy 9:552
politeness in language 18:329
political efficacy 7:305–306
political participation 18:506–507
post-traumatic stress disorder 18:705–706
neurobiological research 18:704
practices, impact of ecological fallacy on 6:866
prevention research 18:883–890
process, photographs in 25:182
projects, academic research and employment 1:49–52
prosocial behavior 19:229
prosocial behavior and empathy 19:216–217
psychological *see* Psychological research
of psychological stress 23:583–587
psychotherapy 19:515–516
location 19:516–517
new areas 19:518–519
observational perspectives 19:517
temporal units 19:517
public opinion 18:460–461
publishing, themes in 19:622–623
dynamics of concentration 19:622
globalization and national identities 19:622–623
technological challenges 19:623
qualitative *see* Qualitative research
qualitative health research 10:692
questionnaires 19:779
rational choice theory 19:904
reflexivity and 20:84
revolutions 20:644
in Roma 20:758–759
schizophrenia 21:13–14
school effectiveness 21:80–85
scientific controversies 21:282, 21:283–284
second language acquisition 9:328–329
on self-conscious emotions 21:479
self-regulated learning 21:542, 21:544–545
semantic memory 21:577–578
for SES impact on childhood behavior problems 2:479–480
on sexuality 21:846
sexuality and social work 21:806
shamanism 21:869
dead ends and detours 21:869
positive directions 21:869
social constructionism 22:225
social exchange 8:483–485
Emerson's theory of exchange relations 8:484
key assumptions 8:485
networks 8:484–485
power and dependence 8:484
social influence, influence of 22:348–349
on social justice 22:398–399
social protests 22:516
social science
programs 6:488
social science, urban geography 24:804
in social work practice 19:93–94
social work-specific 20:502–507
on socialist societies 22:849
sociality 22:858–859
sociocultural 21:849
stigma 23:461–464
Allport's theory of nature of prejudice 23:462–463
early race theory 23:461–462
Goffman's stigma and spoiled interaction 23:463
influence of civil rights movement 23:463–464
social gospel 23:462
strategies, settlement and landscape 21:656–657
strategies for describing environmental decisions 7:706–707
supply chain management 23:712–713
surveys and polls 23:752–753
sustainable production and consumption 23:840–843
technological change 24:107
temperament, history of 24:180–181
time use research methods 24:379–382
traditions in psychology 12:314
transactional coping 4:889–890
treatment, ethics 8:31–36
unemployment 24:735
and United States presidency 18:862
universities 18:307–312
of violence 25:117–118
voluntary agencies 25:266–267
women and criminal justice 25:619–620
in women's studies 9:739–740
work-related stress 23:549–550
on youth physical activity motivation 15:919
Research and development (R&D) 22:656–657
absorptive capacity 1:14–16
collaboration in organizations 2:769–770
defense economics 6:3–4
for environment in Africa 1:298
funding for renewable energy 18:367–368
innovation 12:160–168
innovation and employment 12:154–157
international business 12:486
in organizations
contemporary debates in R & D management 20:500

economic determinants and impact of R & D activities 20:498–499
educational research 7:282–283
innovation 20:496
limitations of R & D statistics 20:499
modernization 20:497–498
patent information versus R & D activities 20:499–500
modernization 20:498
specialization by corporate function 20:497
specialization by discipline 20:496–497
specialization by institution 20:497
technology commercialization 20:496–497
social and behavioral science
Asian centers 22:618–619
Asian government-linked research centers 22:619–620
Asian private research organizations 22:619
drivers 22:616–618
in teacher education 24:54–55
Research Assessment Exercise 9:552
Research bias, gender 9:716
Research by-products, ownership, archiving 1:919–920
Research centers 22:604
Research Committee on Conceptual and Terminological Analysis of the International Social Science Council 8:95
Research Committee on the Sociology of Law of the International Sociological Association 13:598
Research conduct 8:50
Research Data Centers (RDCs) 3:288
Research decision-making competence 20:516–517
Research designs 7:729, 9:223
for critical realism in geography 5:278–280
demographic studies 17:730–732
Research Domain Criteria (RDoC) 17:32
Research ethics 8:41, 11:397, 20:508–522
autonomy 20:510
bioethics
on children and the mentally ill 13:561
on the elderly 13:560–561
prisoner research 13:562–563
U.S. government-sponsored research 13:561–562
on women 13:560
biomedical research involving humans 8:78–82
codification and institutional regulation after WWII 21:315–317
Common Rule 4:255
community research 20:509–512
conduct and dissemination 20:510–512
benefits assessment 20:512
content and understanding of consent documents 20:511
journal publishers as gatekeepers 20:511–512
publication and utilization 20:511–512
contextual concerns 20:508–510
East and West 20:509
growth of international collaborative research 20:508
internationalization and regulation 20:509–510
research and science as culture 20:509
developed country culture 20:510–512
developing country culture 20:509–512
feminist research 9:717–718
foundations of 17:93
history of 20:514–515
human subjects protection in United States 8:48–52
informed consent 12:110–115
international initiatives 20:512–513
instruments 20:512–513
responding to local healthcare needs 20:513
privacy 20:510, 20:512
research subjects 20:508–512
review, cluster randomized trials 3:900–901
secrecy 21:378
Research Ethics Boards (REBs) 8:37, 8:49
Research Ethics Committees (RECs) 3:897, 3:900–904, 8:49, 8:51, 20:515
Research ethics review 20:515
Research Evaluation 21:322
Research funding
academic research and employment 1:51–52
ethics 20:523–528
broader impact 20:524–525
conflict of interest 20:527
distributive justice 20:524
dual use 20:525
freedom of research 20:523–524
funding by companies which sell addiction products 20:524
funding by pharmaceutical industries 20:524
ICMJE's Recommendations 20:527
influence 20:526–527
International Ethical Guidelines 20:526–527
material and immaterial outcomes 20:525–526
objectivity of research 20:523–524, 20:526
patenting system 20:524
Physician Payment Sunshine Act 20:526
responsibility of scientists and science 20:524–526
Scientists' Oath 20:524–525
transparency 20:526–527
and ethics 8:53, 8:55–56, 8:58
gatekeeping in 9:620
Research gaps
gender research 9:716–717
street children 23:534–535
Research infrastructures 22:596
comparative databases 5:790, 5:793, 5:796
definition 22:596
foundational period of 22:599–601
in Canada 22:600–601
in United States 22:600
period of punctuated evolution 22:601
in Canada 22:601
in United States 22:601
Research integrity 8:37–38, 8:41
Research Library Group (RLG) (San Francisco) 5:702
Research management 21:322–323
Research methods 7:562–563
and ethics 8:58
for measuring erotic capital 8:5–6
Research misconduct 8:38, 8:41
Research on Bono's Product (RED) campaign 4:747
Research on circulation of law 13:599–600
Research on consumption of law 13:599–600
Research on formal law 13:598–599
Research on reception of law 13:599–600
Research participants
capacity to consent 20:516–517
fieldwork ethics 9:154–155
illegal activities 9:154–155
intimate relationships 9:154
human, regulatory protections for 4:252–257
identification, cluster randomized trials 3:901–902
Research Policy 21:322
Research politics, and ethics 8:58
Research products, universal access to 8:34
Research program, Marxism 14:650–654
contradictory nature of capitalism 14:652–654
historical materialism 14:651–652
problematic reproduction of capitalism 14:652–654
social construction of reality 14:650–651
Research publication, ethical aspects 20:529–533
authorship 20:530
citation analysis 20:530
ethics 20:531–532
impact factor 20:530
Mertonian norms 20:529–530
peer review 20:530–531
plagiarism 20:530
and science policy 20:529–530
scientific journals 20:530
scientific misconduct 20:530
Research question
formulation for systematic review 23:906–907
in mixed methods research 15:620–621
Research space, experimental laboratories 8:552
Research subjects
informed consent of 12:110–115
in social research 18:943
Research-based practice 23:215
Researchers
access to journals 12:866–867
mathematical psychology 14:805

Researchers (*continued*)
use of statistical databases in social research 18:944–945
Researcher–subject dichotomy 4:726–727
RESER *see* European Research Network on Services and Space (RESER)
Reservation wage equation 21:334
Reservations 1:248
Reservatum ecclesiasticum 20:102
Reserve clause 23:320–321
Reserve forces, in military 15:514–515
Reserve growth 17:166
Reserved seats 8:888
Resettlement
of offenders 5:210
and probation 19:44
Resettlement Administration (US) 1:220
Residence, and households 11:228
household archaeology 11:221
Resident registration (Houkou) system 24:876
Residential child care 20:534
ambivalence in international policy 20:534–535
Boys Town approach 20:537
child welfare 20:537
deinstitutionalization 20:536
effectiveness, quality and standards in 20:537–538
group care 20:536–538
institutional care 20:534–537
move away from 20:536
UN Guidelines on Alternative Care for Children 20:535
use of 20:536–537
Residential cultures 3:757
Residential design and density 20:14f
Residential distribution of social groups in urban space 21:411
Residential housing 24:224–225
Residential instability 11:628–629
Residential location 14:289–290
Residential mobility 11:236, 14:354, 18:597–598
Residential models, for urban systems 24:808–809
Residential mortgage-backed securities (RMBS) 3:581–583
Residential moves 5:736, 5:738
Residential schools 11:764
Residential segregation 20:540–543, 20:549–554, 21:407, 21:409, 21:411
assimilation 20:542
blockbusting 20:542
class 20:549
consequences of racial isolation 20:541
dissimilarity 20:549–550, 20:552–553
diversity 20:551
ethnicity 20:540–541, 20:551–552
exposure 20:549–550, 20:552–553
gentrification 20:541
geographic aspects 20:544–548
choice-constraint debate, interpretation 20:545
division of residential space 20:544
index wars, measurement 20:544–545
ghetto 10:124
homeownership 20:540–541
and homicide 11:164
hypersegregation 20:540, 20:542
immigration 20:541, 20:549–551
inequality 20:549, 20:551–553
and inequality 20:545–547
and racial inequality 20:546–547
inner city 20:541
and integration in last 50 years 20:541
locational attainment 20:540
measurement and future study 20:542
measuring segregation patterns 20:549
mixed-race 20:551
multiethnic households 20:551
neighborhoods 20:540–542, 20:549–552
change 20:541
patterns and change in 20:549–550
poverty 20:549
preferences 20:549, 20:551–553
prejudice 20:542
race 20:540–542, 20:550, 20:552
race, class, and the changing urban mosaic 20:552–553
racial
persistence of 20:541–542
in United States 20:540–541
redlining 20:542
segregation, continuing 20:551–552
spatial measures of 20:552
steering 20:542
suburbs 20:541
Residential spaces, urban growth models 24:807
Residential treatment 24:277
Residual analysis, loglinear models 16:128–129
Residual decomposition 6:539
Residual (restricted) maximum likelihood (REML) 14:148
Residual Mechanism 25:385
Residual spatial autocorrelation 23:203
Residual time density 12:78–79
Residuals 8:600
Resilience 2:868, 4:354, 4:633, 4:898–899, 6:479, 20:555–567, 21:398, 21:529–530, 22:358–359, 22:757, 23:172–174
allostasis 20:555–556
bereavement 2:532–535
bilingual 2:584–585
bounce back 4:354
bounce forward 4:354
boundaries and interfaces 20:557–558
health 20:558
optimal development 20:557–558
positive psychology 20:557
childhood cancer 3:462–463
and cities 20:561–563
climate change 20:563–564
community 4:355
community resilience 20:565
coping 4:891–892, 20:555–557
culture 20:558–559
definition 4:355, 6:479
diagnostics 20:558
disaster 10:547–548
disaster management 20:563–564
disasters 7:428, 7:431
education and learning 7:139–140, 7:142
effect of future orientation on 9:564–565
human development 11:273–274
hypertension 11:454
in infancy 18:857–858
and mass trauma 14:748–749
meaning 4:355
and mental health 15:189
in military/disaster psychiatry 15:495–500
outlook 20:559
panarchy 20:561–562, 20:565–566
polycentric institutions 20:564
principles and practice of promoting 20:558–559
assessment 20:558
evaluation 20:559
prevention and intervention 20:558–559
program development 20:555–560
concepts and definitions 20:556–557
questions and observations 20:555–556
promoting, psychosocial work 19:479–481
protective factors 20:555–559
psychological 14:346–348
refugees 22:770–771
resources 20:556–559
risk factors 20:555–559
strengths 20:555–559
stress 20:555–559
sustainable development 23:824
implications of 23:825
and sustainable development 23:817, 23:819–820
systems thinking 20:561, 20:566
and transformation 4:359–360
urban ecology 20:563
urban economic resilience 20:565
of urban economies 20:565–566
and urban governance 20:561–565
vicarious 18:711
and war/political violence 25:389–390, 25:393
of welfare state through policy feedback 25:524
Resilience theory 22:359, 25:728
Resiliency 18:886
and stress 23:572–573
Resilient city 22:114–115
see also Eco-cities
Resilient communities, and education for sustainable development 7:736–737
Resist group 3:528
Resistance 8:600, 13:471–473, 16:465–466
of Australian aboriginals 2:234–235, 2:237
collective, neglect of 4:154–155
critical theory 5:308–309
market 16:466–467
in narratives 16:220–221
peasant agriculture 17:666
peasants 17:677

and slavery 22:39–40
studies of 18:337–338
and territory 24:225
and youth culture 25:816–817
Resistance to change, organizational development 6:296
Resistant ambivalent attachment 16:470
Resistant tuberculosis 18:177–178
Resnais, Alain, works, *Hiroshima mon amour* 9:169
Resocialization 5:615, 23:799
Resolution 17:619, 20:420
cartography 3:168–169
reasoning 5:935–936
Resonance model, of discourse 12:41
Resonance theory of hearing 10:757
Resort enclaves 24:470
Resource allocation, urban growth models 24:807
Resource attraction, by voluntary agencies 25:265–266
Resource commodities 20:569
Resource conflict, and ethnic cleansing 8:87–88
Resource conservation 7:774
Resource consumption 24:373–378
and time use, connections between 24:373–375
Resource cycle 20:570
Resource defense polygyny 14:615
Resource demands, of risk assessment 20:693–696
Resource dependence theory 2:915, 17:308
of interorganizational relationships 12:622
organizations 17:308, 17:310
Resource dependency
sustainable development 23:825–826
implications of 23:826
Resource depletion 15:554–556
LTG study 14:138, 14:142
Resource depression, zooarchaeology 25:851
Resource dilution model 6:754
Resource distributions
in collectivism/individualism 4:209
legitimacy of 13:827
Resource extraction, functional model of 15:555
Resource flows, in organizations 3:894–895
Resource frontiers 2:760
Resource limitation 1:879
Resource management, fishing 9:228
Resource markets 20:570–572
urban markets 20:572
value chain coordination 20:571–572
Resource mobilization
and collective behavior 4:160
Resource mobilization (RM) theory 4:146, 22:447–449
Resource peripheries 20:572–574
remapping control 20:573–574
Resource policy, agriculture 1:482
Resource regimes 20:569
Resource theory of social exchange 12:953
Resource transfers
human behavioral ecology 11:248–249
Resource use
in Africa 1:293
and energy 7:622–624
optimizing, and industrial ecology 11:847–848
Resource-activity-processoutcome (RAPO) model 5:60
Resource-based view (RBV) 8:775, 23:508–509
diversification 6:586–587
marketing strategies in 14:555–556
Resource-oriented proof-theoretic logics 16:928
Resource-partitioning theory, organizational ecology 6:892
Resource-reduction theory, and aging 1:431–432
Resources 2:616, 18:770–771, 18:774
classification 20:568–569
in combined pharmacotherapy and psychotherapy
availability for patients 19:496
cost-effective utilization 19:496
control of 20:569–570
coping 4:902–904
and cultural inequality 22:331
cycles 20:570
definitions 20:568
and equity in education 7:918–919, 7:921–922
evidence-based medicine 8:330
exploitation 20:570–571
and global economy 20:568
greening of business 10:393
health-enhancing 19:817
markets 20:570–572
mental health 15:189
occupational stress management 17:147–150
ownership 20:569–570
peripheries 20:572–574
remapping control 20:573–574
psychological adaptation to childhood cancer 3:459–460, 3:460t
required for health information systems 5:786
resilience 20:556–559
and revolutions 20:645
self-regulation 21:557–562
themes of 20:568
welfare as 25:507–508
at work 23:548
Resources for Bears web site 13:891
Resourcing, social/behavioral research 22:643–647
Respect 11:808
for authority in society, and group violence 14:704
definition 13:5–6
and honor killing 11:185–187, 11:189–190
and long-term heterosexual relationships 20:915
for persons 8:79, 20:516–517
biomedical research ethics 8:79–80
Respectability 1:640–641
Respectful individualism, and indigenous peoples 11:805
Respiratory disorders 20:576–580
anxiety 20:576–579
asthma and COPD
comorbid psychological symptoms in 20:576–577
interventions including psychosocial aspects in 20:578–579
social aspects in 20:577
behavioral medicine 20:578
breathlessness 20:576–579
cognitive-behavioral therapy 20:578–579
depression 20:576–579
dyspnea 20:577–578
links between psychosocial aspects and disease processes 20:577–578
patient adherence 20:577–578
pulmonary rehabilitation 20:576, 20:578–580
self-management 20:577–580
Respiratory muscle weakness 20:579
Respiratory sinus arrhythmia 23:539
Respiratory system, psychosomatic disorders in 19:488
Respondeat superior, doctrine of 3:29
Respondent conditioning
and learning 13:705, 13:705t
pain 17:452–453
see also Classical conditioning
Respondent conditioning model 17:452–453
Respondents
confidentiality of, archiving 1:917–918
obligations to, surveys and polls 23:754–756
of questionnaires, tasks 19:771–772
age-related differences 19:772
judgment recalling/computing 19:772
question context 19:771
response alternatives 19:771–772
response editing 19:772
response formatting 19:772
understanding question 19:771–772
Responding, communication skill 4:288–289
Responding, prosocial, cultural variation in 19:236
Responding emotions 16:658–659
Response alternatives, of attitude measurement 2:178–180
Response bias 19:473–474, 24:640, 19:474f
and recognition memory 20:38
Response competition effect 2:180
Response competition measures of implicit attitudes 2:180–181
Response cost, behavior therapy 2:429
Response dimensions 23:765
Response efficacy, health behavior 10:577–578
Response error 16:920–921
audience measurement 2:213–214
avoiding 16:920

Response error (*continued*)
definition 16:920
randomized response techniques for sensitive questions 16:920–921
in sexuality surveys 15:315–316
Response facilitation effect 2:180
Response facilitation measures of implicit attitudes 2:180
Response formatting 19:772, 22:17–18
Response instructions of situational judgement test 22:15
Response order effects, questionnaires 19:772
Response perspective on psychological stress 23:584
Response prevention, exposure therapies 8:606–607
Response selectivity, developmental neuroscience 16:601–602
Response sensitization 10:481
Response speed 4:11
Response styles 11:454, 23:758
extreme response style 23:758–763
latent class representations of 23:760
Response surface methodology (RSM) 23:476
Response time (RT) 19:474, 24:640
distributions 6:399–400
quantile probability plot of 6:399–400
shape 6:399
error response times 6:399
mathematical psychology 14:810–811, 14:814–815
random utility models 5:901–905
Response time (RT) measures of implicit attitudes 2:180–181
Response time (RT) models
mathematical psychology 14:805
Response to intervention 21:114, 23:216
Response variable, regression analysis 14:157–161
Response-based sampling *see* Outcome-based sampling
Responses 6:479, 14:367
cardiac defense 23:541
of children to divorce 14:520–521
goal-oriented behavioral 10:271
and natural disasters 16:316–318
questionnaires
alternatives 19:771–772, 19:774–775
editing 19:772
formatting 19:772
survey questions
categories 23:767–768
dimension 23:765–766
Responsibilities redistribution, for access to justice 12:923–924
Responsibility 10:628, 10:630–631
Butler on 3:56–59
concept 20:587–588
and conscientiousness 4:659
definition 20:581
and deserts 6:231–232
in fatherhood 8:841
liberalism and 14:23–24
moral 9:406–408
philosophical aspects 20:587–591
as accountability 20:589–590
as attribution 20:589
blameworthiness 20:587–590
case studies 20:590
causal determinism and moral responsibility 20:590
causal responsibility 20:587
conditions for responsible agency 20:588–589
control conditions 20:588–590
duty and obligation 20:587–588
epistemic competence 20:589
extreme evil 20:590
and free will 20:590
holding responsible and being responsible 20:589
liability 20:589
moral community 20:587–588, 20:590
private moral features 20:590
as susceptibility to reactive attitudes 20:589
politics of 15:536–537
and public interest 19:552–553
of scientists and science 20:524–526
in social work 20:581–586
blame 20:581–582
collective 20:583
duty 20:581–582
family 20:582
individual 20:581–582
moral 20:584
professional 20:584–585
rights 20:581–583, 20:585
risk 20:581–582
social 20:582–583
Responsibilization 21:502, 24:848
Responsibilization strategy 1:791–792
Responsible conduct of research 8:37, 8:41
Responsible research and innovation (RRI) 24:127
Responsible stage of aging 1:436
Responsible tourism 24:474
The Responsive Communitarian Platform: Rights and Responsibilities 4:313
Responsive communitarians 4:320–321
Responsive securitization 19:165
Responsive strategy
conduct 23:504–505
making 23:505–506
Responsiveness 22:564
and intimacy 14:376–377
Responsivity, and probation 19:45
Respublica litterarum 20:529
Restatements 3:669
Resting state connectivity studies (RSC), for cortical area studies 5:24, 5:27
Resting-state data 9:529
Restless legs syndrome 22:74
Restoration 12:958
Restorative justice 1:835–836, 2:672–673, 5:192, 5:211, 12:939, 12:958, 17:657, 20:437, 22:386, 22:388, 22:393–394
and oppression 17:251
Restorative justice procedures 17:650
Restorative justice (RJ) programs, offender rehabilitation 20:195
Restorative niches 17:764
Restraining orders 6:628–629
Restraint, principle of 1:823
Restricted census microdata 3:288
Restricted maximum likelihood (REML) 16:48, 24:691
linear mixed models 14:166–167
Restricted maximum likelihood-empirical Bayes estimation (REML-EB) 17:731–732
Restriction endonuclease 9:915–916
Restriction enzymes 9:942
Restrictions, records with 16:903–904
Restrictive concept of knowledge 13:108
Restrictive continuity, and immigration 11:672–673
Restrictive deterrence 6:255–256
Restructuring 21:103
Results, program evaluation 19:118–119
Results bias 24:130
Resumed Education for Adult Learners (REAL) program 1:149
Retail rivals, disadvantaging of 20:169
Retail sales, of guns 10:457, 10:459
Retail trade
areas 23:154
and market geographies 14:544–545, 14:547
Retaliation, and conflict 4:624–625
Retaliatory hate crime 10:542
Retardation hypothesis 25:170
Retention 14:354–355
and grit 10:399–400
in human resource management 11:348, 11:352
long-term 15:149
of outliers 23:388–389
teacher, in United States 24:78
Retention rate 8:371
Reterritorialization, political geography 18:377
Rethinking flows 12:92–93
Rethinking Marxism: A Journal of Economics, Culture & Society 14:649
Rethinking places 12:92–93
Rethinking political 23:635
Reticence 2:459
Reticularis pontis caudalis (RPC) 8:860–861
Reticulation
cultural evolution 5:394–395
in prehistoric contexts 5:395–396
Retinal function, theories of 25:158–159
light intensities in natural scenes 25:158–159
retinal neurons, role of 25:158
Retinal implants 16:715–716
Retinal neurons 25:158
Retinal receptive fields 25:159

Retinocentric frame of reference, *versus* body-centered frame of reference 23:118–119
Retinol axons 23:279–280
Retinotopic maps 25:184
Retinotopic projection to area V1 17:127–128
Retired Senior Volunteer Program 1:150
Retirement 1:324, 20:592
  and aging 1:415–417, 1:415t
  behavior, modeling 20:603
  definition 20:602
  economics of 20:602
    in family context 20:606
    health and health insurance 20:605
    labor demand 20:606
    public and private pensions 20:603–605
    wealth and savings 20:605
  future research 20:595
  gradual retirement 17:187
  and health 20:598–601
    economics of aging 20:599
    empirical evidence 20:599–600
    older workers 20:599
    pensions 20:598–600
    theoretical mechanisms 20:598–599
    well-being at older ages 20:599
  historical overview 20:592–593
  institutional gendered life-course approach 20:593–594, 20:593f
  job displacement insurance 12:820, 12:830
  labor supply 20:605
  and labor supply 13:197
  new later life-course stage 20:594
  pensions 1:324
  and population aging 18:540–543
  Social Security 20:602–606
  systems 1:324
  trends and patterns 20:602–603
  in twenty-first century 20:594–595
  in United States 1:325
  from work 25:692
Retirement behavior 1:424
Retirement History Study (RHS) 5:698
Retracted paper 20:530–531
Retractions, and peer review 17:682
Retrenchment
  welfare 25:521–527
  welfare state 25:477–479
Retributive justice 12:938–939, 13:750
Retributive punitivism 12:958
Retrieval 1:648, 20:608–612, 20:881
  cue description 20:611–612
  direct 20:608–610
  ecphory 20:608, 20:610
  encoding specificity principle 20:610–611
  engram 20:608, 20:612
  of episodic memory 7:859–860
    neural correlates 7:859–860
  feeling of knowing 20:610
  feeling of rightness 20:611
  felt rightness 20:611
  generate-recognize strategy 20:610–611
  indirect 20:610–611
    postretrieval 20:611
    preretrieval 20:610–611
  levels of processing principle 20:609
  memory 15:149–151
  memory cue 20:608–609, 20:611–612
  memory representations 20:608
  memory trace 20:608–612
  monitoring and verification processes 20:611
  processes invoked during 20:609f
  quantitative memory models 15:162–164
  reconstruction and rebinding in 20:611–612
  routes 20:608–611
  source monitoring framework 20:611
  source-constrained retrieval 20:610
  working with memory procedure 20:610–612
Retrieval algorithms 19:57–58
Retrieval induced forgetting 4:191
Retrieval interference 20:615–616
Retrieval mode 20:610
Retrieval of information 13:681–682
Retrieving effectively from memory (REM) 15:163–164
Retroactive inhibition (RI) 15:162
Retroactive interference (RI) 15:160, 16:847, 19:1, 20:613–620
  in declarative memory 20:613–614
  developmental effects 20:615
  episodic memory 20:613–614
  forgetting 20:613, 20:615, 20:617–618
  healthy aging 20:615
  historical overview 20:613
  long-term memory 20:613, 20:618
  mechanisms underlying 20:615–617
    consolidation interference 20:616–617
    retrieval interference 20:615–616
  and neuroscience 20:617–618
  in nondeclarative memory 20:614–615
    implicit memory 20:614
    procedural memory 20:614–615
  patient populations 20:615
  semantic memory 20:613–614
Retroduction, in realism 20:24
Retrofitting 6:829
Retrograde amnesia (RA) 1:644, 1:651, 10:876–879, 10:881, 10:883–884, 13:682, 24:554
Retrograde episodic-autobiographical memory impairments 1:655
Retrograde memory distinctions 1:645–646
  autobiographical recall of recent and remote memories 1:646
  semantic facts and episodic experiences 1:645–646
Retrospective data 5:737
Retrospective economic voting 7:27
Retrospective interviews, in ethnography 8:187
Retrospective life history data 14:51–52
Retrospective sampling *see* Outcome-based sampling
Retrospective surveys 5:737–738
  advantages 5:737–738
Return migration 15:409, 15:436
  in Europe 15:459
Return of religion 25:608–609
Return on investments (ROI) 4:928, 5:56, 15:554
  energy 7:630–631
Returns 7:58–59, 13:262–263
  estimation, of vocational education and training 25:248–249
Returns to education 7:134–135, 21:119–120
Reunification 2:21
Reuse, of linguistic databases 5:815–816
Reuter, Julia, on epistemic cultures 7:876
Revealed preference methods, valuation of benefits using 5:51
Revealed preference theory, of desires 12:337
Revealed preferences 19:440
Revelation, periodic 21:374
Revelation of Arès 16:806
Revenge 17:621
Revenge homicide 4:605
Revenue Act 1978 24:765
Revenue sharing 23:322–323
Reversal design 21:986, 21:988
  for workplace safety and health 25:731
Reverse causality 15:851–852
Reverse causation, obesity 17:76
Reverse correlation 5:543
Reverse culture shock 5:615
Reverse engineering 24:122
Reverse genetics 9:906
Reverse hierarchy theory 25:156–157
Reverse inference (brain mechanism) 17:52
Reverse Record Check (RRC) 3:293–294
Reverse sexism 21:694
Reverse survival techniques, for fertility estimation 6:145
Reverse transcription 1:488–489
Reversibility, diffusion 6:414
Reversible error, principle of 1:823
Review of Agrarian Studies 1:457
Review of Economics of the Household 11:226
Reviewers 5:440, 5:442–444
Reviews, law 5:808–809
Revisability thesis 11:814
Revised hierarchical model (RHM) 21:351–352, 21:354
Revised Interpersonal Adjective Scales: Big Five (IASR-B5) 9:233
Revised Reinforcement Sensitivity Theory (r-RST) 17:828
Revisionisms 1:626
  English Revolution 7:651–653
  Marx on 14:646–647
Revisionists 10:747–748, 18:383–384
Revitalization 11:630
  postsocialist cities 3:605
  see also Neighborhood revitalization
Revitalization movements 3:351
Revival, republican 20:478
Revival of Chinese migrations 12:526–527
Revivalism 8:287–288
  religious, in Southeast Asia 23:70–71

Revivalist Christianity 8:285–288
Revivification 20:333
Revolt and war 25:403–404
Revolution in Military Affairs (RMA) 15:522
Revolution theory 6:216, 20:627
  definition 20:627–628
  explanation 20:628–630
  validation 20:630–631
Revolutionaries 6:369
Revolutionary Armed Forces of Colombia (FARC) 11:364–365
Revolutionary consciousness 22:721
Revolutionary Homosexual Front (Frente Homosexual de Acción Revolucionaria, FHAR) 13:860
Revolutionary movement 20:627–629
  and functional theories 20:629–630
Revolutionary orientations, in modern youth movements 9:868
Revolutionary outcome 20:627–629
  and structural theories 20:630
Revolutionary power, of charisma 3:353–355
Revolutionary situation 20:627–629
Revolutionary socialism 22:834
Revolutionary United Front (RUF) 25:385
Revolutions 3:715, 10:85, 10:314–315, 14:650, 16:8–10, 20:621–626, 20:637, 20:642–649, 22:32, 23:23–24
  biotechnology 2:676–679
    institutional dimensions 2:678–679
    technological and epistemic space 2:677–678
  capitalism 20:644, 20:648
  in Central America 1:622–623
  and child soldiers 3:440–444
  and collective violence 14:703
  and communist parties 18:404–408
  comparative methods 20:647
  computers 4:518
  conceptualizing 20:621–622
    common usage 20:621–622
    political revolution 20:622
    social revolution 20:622
  crowds 5:361–365
  democracy 20:644, 20:646
  democratic revolutions of post-Cold War era 20:623–625
    Arab spring 20:625
    electoral (color) revolutions 20:624
    nonviolence 20:624–625
    overthrow of communism 20:624
  from description to explanation 20:622
  in eastern Europe 25:838–839
  Eisenstadt's contribution 7:322
  elusive definition of 20:643–644
  emerging directions 20:645–648
  English Revolution, The 7:649–655
  and enlightenment 7:660–662
  etiology 20:638–639
  evolution of 20:637–638
  French Revolution 9:409–412
  and gender 9:782–783
  globalization 20:648
  and globalization 20:625–626
  historical sociology 11:27–28
  history of 20:637–641
  in human lives 20:640–641
  Islamic fundamentalism 12:741–744
  Islamism 20:623
  and lesbian feminism 8:942–943, 8:945–947
  major research traditions 20:644
  marginalist 14:646
  Marx on 20:621–623
  in MENA 16:365
  morphology 20:639–640
  national liberation movements 20:621
  new meaning of 20:642–643
  parties 20:642–643, 20:645–646
  political sociology 18:472–473
  post-world war II era 20:644–645
  and production mode 15:714–715
  proletarian 14:647
  relative deprivation 6:214, 6:216–217
  and slavery 22:39–40
  social 14:642–643
  social movements 20:645, 20:647–648
  states 20:643–646, 20:648
  structure *versus* agency 20:622–623
  taxonomy 20:638
  theories of 20:627–632
  use of term 7:650
  of values 20:623–625
  violence 20:643–644, 20:646–648
  and war 25:403–404
  working classes 25:695
  world systems theory 25:748–751
Revolutions of 1989–1990
  deviant type 20:633
  Eastern and Central Europe 20:633–636
  peaceful revolutions 20:635–636
  velvet revolutions 20:634–635
Revolving Door Identification Model (RDIM) 7:161–162
Revue de MAUSS (*Mouvement antiutilitariste en sciences socials*) 6:703
Revue de synthèse historique 6:703, 11:75–76
Revue internationale de sociologie 22:992
Revue Tocqueville/Tocqueville Review 24:424
Reward dependence (RD) 17:794
Reward prediction error (RPE) 16:640–641
Reward systems 16:640–642, 17:611–613, 17:615–616
  amount of reward 16:641
  budgeting 2:894, 2:898–899
  and dreaming 6:655
  eating 11:422
  of science 21:262–263
  timing of reward 16:641
Reward-detecting sensory neurons 20:211
Rewards 5:232–233, 6:255, 6:258, 15:908–909, 17:859–860
  addiction 1:98–99
  for artists 2:58
  brain stimulation reward 2:841–846
  in competition/cooperation 4:857
  and effort balance 17:137
  and motivation 15:908–909
  and placebo effect 18:153–155
  and reinforcement 20:209
  taste and smell 24:26–27
  and wanting/needing 25:351–355
Reward-signaling stimuli 20:211–212
Rewiring, of cortex 5:32–33, 5:33f
Rewrite grammar 19:10
Rewriting relation 19:10
Rex, John 4:364, 11:236–237, 22:723
Rex solutus legibus est 1:8–9
Reynolds, Bertha 14:664–665
Reynolds, Susan, contribution in feudalism 9:114–115
RF *see* Rockefeller Foundation (RF)
RFE *see* Recursive feature elimination (RFE)
RFI *see* Rod-and-frame illusion (RFI)
RFs *see* Receptive fields (RFs)
RGCs *see* Radial glial cells (RGCs)
rGE *see* Gene–environment correlation (rGE); Gene–environment correlations (rGE)
  developmental behavioral genetics and education 6:321
RGS2 *see* Gene-encoding regulator of G protein signaling 2 (RGS2)
RHD *see* Right hemisphere damage (RHD)
Rheinberger, Hans-Jörg, works, *Toward a History of Epistemic Things* 8:580
Rheinish Carnival 3:146
Rheme 12:96–97
Rhesus monkey
  serial order memory 21:642
Rheticus, Georg 20:285–286
  works, *University of Wittenberg* 20:285–286
Rhetoric and semiotics 20:651
Rhetoric reduction 20:651
Rhetoric of reaction 20:660
Rhetorical analysis 5:927–930, 20:650–654, 22:546–547
  argumentation and thinking 20:653
  classical antiquity 20:650
  mental mapping 20:652–653
  metaphor 20:652–653
  metonymy 20:652–653
  modern rhetoric
    rhetoric and semiotics 20:651
    visual rhetoric 20:651–652
  rhetoric reduction 20:651
  rhetorical system 20:650–651
Rhetorical approach, to political psychology 18:430–431
Rhetorical meanings, of media conventions 21:165
Rhetorical system 20:650–651
  actio 20:651
  dispositio 20:650
  elucutio 20:650
  inventio 20:650
  memoria 20:651
Rhetorics 6:62, 11:127–128, 23:525
  development 15:289
  land tenure 13:232
  and metaphors 15:289
  political 18:431
  and science communication 21:220–223

strategizing 23:525
tropes 15:289–290
Rheumatoid arthritis 20:655–659
adaptation 20:657
depression and anxiety 20:657
e-health 20:658
etiology, prevalence, course and disease management 20:655
nonpharmacological treatment 20:655–656
pharmacological treatment 20:655–656, 20:658
psychological (dys)functioning 20:657
psychophysical aspects 20:655–657
fatigue 20:656
pain 20:655–656
physical functioning 20:656–657
regular monitoring and timely support 20:658
self-efficacy 20:655–658
self-management 20:658
social participation and work participation 20:657–658
symptoms 20:656–658
and well-being 20:655
RHI *see* Rubber hand illusion (RHI)
Rhizomes, geography as 16:568–569
Rhode Island 1:610
Rhodes, Geoffrey, works, *The Mind at Work: A Handbook of Applied Psychology* 7:591
Rhodesia 23:77–78
Rhodopsin 17:271
RHS *see* Retirement History Study (RHS)
Rhymes 13:267–268, 13:276–277
Rhythms 13:273–275, 16:652–653
RI *see* Recombinant inbred (RI) strains; Retroactive inhibition (RI); Retroactive interference (RI)
RIAs *see* Regional integration agreements (RIAs)
Ribbentrop, Berthold von, works, *Forestry in the British Empire* 7:750
Ribonucleic acid (RNA) 11:289
epigenetic inheritance 7:833–834
-induced silencing complex (RISC) 7:841
Ribonucleic acid (RNA) regulation 7:849
Ribot, Théodule-Armand 2:596
Ricardian equivalence 8:826–827
Ricardo, David 3:90, 4:751–752, 6:241, 14:407, 14:644, 23:199, 25:7–8
and Corn laws 20:660
bullionist view of 20:661
on classical economics 20:660
on de-skilling 20:663
on diminishing returns 20:660–662
on free trade 20:660
on gold price 20:661
on international trade 20:661–662
on labor theory of value 20:662
on laissez-faire 20:661, 20:663
at London Stock Exchange 20:660
on macroeconomics 20:661
and Malthus 20:660–661, 20:663
on microeconomics 20:661–662
on Old Corruption 20:663
on perversity thesis 20:660
on population increase and class conflict 6:242
on profit 20:663
on rent 20:661–662
Ricardian vice 20:661
and Smith 20:660, 20:662
on sociology 20:663
theories of 19:124
comparative advantage 20:661–662
on wages 20:660–663
works
On the Principles of Political Economy and Taxation 3:90, 14:644
Ricardo–Viner model 12:570
Ricci, Matteo 3:531–532
Ricci, Michelangelo 15:11
Rice, Condoleezza 24:451–452
Rice production, in India 9:308
Richard Laws 21:676–677
Richards, Audrey 7:145
Richards, E., on multiculturalism in Australia 2:240
Richards, Ivor Armstrong 15:290–291
Richardson, Diane
on sexual citizenship 21:729–730, 21:732
on sexual rights 21:728
Richardson, Tim
discourse analysis 5:256
Riches, Colin 21:65
Rich-get-richer hypothesis 16:783, 22:424
Richta, Radovan 13:106–108, 21:131
Richter, Melvin 18:490
Richthofen, Ferdinand von 7:723, 16:334
Rickert, Heinrich 19:391, 21:167
Ricketts, M.L. 7:374–375
Ricklefs, Merle, works, *Jogjakarta under Sultan Mangkubumi, 1749–1792: A History of the Division of Java* 11:87
Ricoeur, Paul 19:677–678, 20:88, 20:665–668, 24:292
career of 20:665
legacy of 20:667–668
life of 20:665
on metaphor 15:291–292
works
The Course of Recognition 11:179, 20:667
as dialogue and translation 20:666–667
Fallible Man 20:666
From Text to Action 20:665
Memory, History, Forgetting 20:665
On Translation 20:666
Oneself as Another 20:665–667
The Rule of Metaphor 20:665
Symbolism of Evil 20:666
Time and Narrative 20:665–668
The Voluntary and the Involuntary 20:666
Rida, Muhammad Rashid 12:742
Riddare, Swedish 1:965
Ridgeway, Sally, on movement circles 5:468–469
Ridley, Clarence 21:971
Riefenstahl, Leni 6:622
works, *The Triumph of Will* 6:621–622
Rieff, Philip 2:273
Riemannian spaces 25:198–199
Riesman, David 7:934–935, 12:670, 20:669–672
academic life of 20:671–672
apprenticeship of 20:669–670
and Benedict 20:669–670
and Friedrich 20:669–670
and Fromm 20:669–671
and Glazer 20:670–672
in Harvard College 20:669–670, 20:672
and Horney 20:669–671
and Hughes 20:672
on inner-directed social character 20:671
on law schools 20:669–670
and Lazarsfeld 20:672
and Lynd 20:669–670
mass society 20:671
and Mead 20:669–670
and Mills 20:670
on other–directed social character 20:671
on psychoanalysis 20:669–670
on social network actors 5:558–559
sociological bestseller of 20:670–671
on tradition-directed social character 20:671
in University of Chicago 20:670, 20:672
works
The Lonely Crowd 9:462, 18:664–665
Riess, Ludwig 11:51
Rifkin, Jeremy, works, *The End of Work* 18:666
Riggs, Marlon, works, *Tongues Untied* 6:621
Right action, virtue ethics 25:147–148
Right frontal effect 23:32–33
Right hemiplegia 18:46–47
Right hemisphere 2:808–809, 23:294–296
functions of 2:801f
hemispatial neglect 10:766, 10:768
and neurolinguistics 16:687
Right hemisphere damage (RHD) 18:791–792, 21:930
Right hemisphere dysfunction 1:653
Right hemisphere language, and inferences 19:397
Right hemisphere lesions 21:932
Right of reply 9:394
Right shift gene model 10:505–506
Right Socialist Revolutionaries 20:857
Right temporal pole (RTP) 19:150
Right temporoparietal junction (RTPJ) 15:794
and body-ownership 21:420
Right to a hearing 6:697–698
Right to confrontation 5:227–228
Right to counsel 12:920–922
Right to culture 11:801
Right to free informed prior consent 11:801
Right to have rights approach 20:683
Right to health, education and work 11:801
Right to know genetic information 2:622
Right to land 11:798
Right to life 16:322
Right to own private property 9:400
Right to property 11:799–801

Right to remembrance 13:441
Right to self-determination 11:797, 11:800–801
Right to the city 12:914, 12:916–917, 20:673–679
bridging political ideal and working slogan 20:675–676
capitalist globalization 20:673
citizenship 20:673–674, 20:676–678
critical urban theories 20:673–674
empowerment 20:673, 20:677
evolution of the concept 20:673–675
global south 20:676–678
human rights 20:675
inclusion 20:673–676
international initiatives 20:673, 20:675–676, 20:678
liberal democracy 20:676
right to the city in global south and beyond 20:677–678
urban politics 20:673–674
urban social movements 20:673, 20:675, 20:677–678
Right to the City Alliance 20:676
Right to the truth 11:368, 11:372, 13:441
Right-censoring 8:310–311, 8:313
in event history analysis 8:317
Right-ear advantage (REA)
bottom-up modulation 6:360–363, 6:361f–362f
brain correlates 6:363–365, 6:365f
top-down modulation 6:363
Right-handedness 9:269, 23:296
Right-hemisphere consciousness 23:291–292
Right-hemisphere conservatism theory 2:121
Right-hemispheric dominance for spatial processing 2:808–809
Righting reflex 15:910–911
Right-posterior lesions 23:296
Rights 3:643–645, 3:647–648, 4:317–318, 13:610, 13:753–754, 14:266
autism spectrum disorders 2:279
children 3:492
of children 3:485–490
citizenship 3:655–656
civil 13:609
claim, expanding variety of 20:680–681
first-generation rights 20:680–681
rights beyond liberal western tradition 20:681
second-generation rights 20:681
third-generation rights 20:681
consciousness 13:610
and constitutions 4:710–714
and consumer policy 4:735
contemporary philosophical debates about liberal legal 20:681–682
atomism, egoism, and selfishness, and 20:682
qualifications for bearing 20:681–682
and criminal justice 5:211
gay/lesbian 9:640
human 13:604
jurisprudence and legal theory and 20:683–684
democracy *versus* legal rights 20:683
rights and justice 20:683–684
and Kantian politics 13:8–9
as compatibility of freedom 13:8–9
land 13:234–235
and learning disability 13:685–688
legal 13:607–609
and liberalism 14:26–28, 14:30–31
logic of 20:680
and moral reasoning 15:803, 15:805
and mutuality 5:308
origin and evolution of 20:680
political 15:697
practice, empirical social science studies of 20:684–685
politics of rights 20:684–685
rights and everyday disputing 20:684
and responsibility 20:581–583, 20:585
and sanism 15:200–202
sexual citizenship *see* Sexual citizenship
social 4:319, 13:721
sociology of, Marshall 14:628
of soldiers 16:306–307
Supreme Courts and 23:722, 23:725–726
trans* 24:551
virtual ethics conceptions of 25:147–148
welfare 25:507, 25:509
women 14:14–15
of women 9:786
Rights awareness, in China 24:977
Rights debate 13:544–545
Rights in people, in socialist societies 22:850–851
Rights model, legal reasoning 13:781
Rights of indigenous peoples 3:687, 11:781–782
Rights of man 9:412
Rights of minorities 11:797
Rights revolution, poverty law 18:734–735
Rights to the city 20:674
Rights-based approaches, to human development 6:310–311
Rights-to-the-state 12:900
Right-to-manage model
trade unions 24:501–502
Right-wing authoritarianism (RWA) 2:198, 2:257–260, 11:677, 19:866
Right-wing authoritarianism (RWA) scale 2:263
versus social dominance orientation 2:264
Right-wing populism 18:411–412
Rigidification 8:72–73
Rigidity-of-the-extreme perspective 2:265
Rigidity-of-the-right perspective 2:265
Rigoberta Menchú Controversy 1:621–622
Rigor, focus groups 9:277
Riina, Salvatore 14:422
Riis, Jacob 18:84, 24:863, 25:181
RIKEN Center for Genomic Medicine 11:318
Rimassa, A., works, *Generation 1000* 25:839
Rimassa, Alessandro 25:839
Rimbaud, Arthur 1:540
R-index 21:948
Ringe model 11:26
RINGOs *see* Religious international nongovernmental organizations (RINGOs)
Rio + 10 Conference 23:797
Rio + 20 Conference 7:772, 18:194, 23:797
see also United Nations Conference on Sustainable Development
Rio Branco Law (1871) (Brazil) 22:41
Rio Declaration on Environment and Development 23:796
Rio Earth Summit (1992) 23:796, 23:813
Riot, Pussy 20:840
Riot hypothesis, and affirmative action 1:264
Riot prone individuals 4:152
Ripeness, conflict mediation 4:630–631, 4:630t
Ripeness for review 1:823–824
RISB *see* Rotter incomplete sentence blank (RISB)
Risk analysis 18:596
Risk appraisal 10:698
Risk assessments 12:501
child and adolescent psychiatry 3:387–388
child protection 3:418–419
tools, recidivism 20:33–34
Risk attitude 17:809–810
Risk averse behavior 19:262–263
Risk aversion 13:574, 17:809, 25:296
Risk behavior 1:115, 1:117–119
Risk communication 10:698
Risk compensation theory 25:728
Risk culture, and uncertainty 24:720
Risk factor reduction, cardiac rehabilitation 20:188
Risk factors
adolescent health 1:121–126
of aggression 1:379–380
bereavement 2:533–534
of crime 12:463
of emotional development of children 7:438–441
epidemiology 7:826–827
mental illness epidemiology 15:179, 15:183–184
obsessive-compulsive disorder 17:121–122
biological risk factors 17:122
genetics 17:121–122
immunological factors 17:122
stressful life events and traumatization 17:121
resilience 20:555–559
of suicide 23:675–676
urban life health 24:831
Risk feedback, personalized risk feedback 10:706–707
Risk function 8:17
Risk homeostasis theory 25:728
on LGBTQ individual parenting 21:757
Risk information, health risk perception 10:702–703, 10:706–708
Risk management 1:479
biotechnology 2:679–680
Risk perceptions 2:489–490, 17:810–811

on criminal behavior
deterrent effect of 6:251–252
experiential effect of 6:252–253
defining 10:698
dimensions of hazards 10:698–699
goals of 10:699
hazard probabilities 10:699
health psychology 10:675–676
health risk 10:702–709
on health-related behaviors 10:700
importance of health risk perceptions 10:698
judgments 10:698–700
and optimistic bias 10:699–700
personalized medicine 17:930
Risk pooling 15:412
and marriage 8:829
Risk predictors
HIV infection
adolescents and youth 11:139
heterosexual men 11:142
MSM 11:141
women 11:140
Risk preferences 13:571, 17:809
Risk premiums
workplace safety 20:863–865
Risk ratio 7:829, 7:829b
Risk reduction
and industrial ecology 11:847–848
insurance 12:210
Risk reduction strategies (RRS) 10:388
Risk regulation 13:504
government and 13:505
Risk screening
ethics 20:698–705
Risk seeking behavior 19:262–264
Risk selection, health insurance markets 10:641
Risk society 1:749, 5:246, 7:879, 13:107, 15:93, 15:521–522, 20:686–687, 20:709, 24:127, 24:144, 24:747
Beck on 2:397
modernization 15:698
and tourism 24:468
and uncertainty 24:720
Risk taking 17:809–813
caveats 17:812
domain 17:809–810
by drivers in transportation 24:625–626
framing effects on 17:809–810
future directions 17:812
gender differences in 9:752
and income tax 11:747
multiple processes in 17:811
occupational 25:728–729
and perceived-risk attitude 17:810
as personality trait 17:809–810
and risk perception 17:810
and sibling order 21:924
Risk tradeoffs in consumption 25:17
Risk-averse, insurance 12:210
Risk-based principles 12:219–220
Risk-benefit assessment, Common Rule 4:254
Risk–captivity–harm thesis 5:246–247
Risk-compensating wages 25:17
Risk-managed production, pharmaceutical industry 18:7–8
Risk-need-responsivity (RNR) model 5:129, 20:196, 20:437, 21:675
and probation 19:45
recidivism 20:34
Risk-reward ratio 12:219
Risk–risk analysis 13:506–507
Risks 10:547–551, 10:698, 17:370–371, 19:43–48
activities giving rise to 13:504
analysis and management 13:506–507
assessment 13:506
assessment, in social work 20:692–697
assessment by societies 13:504
Beck on 2:396–399
in biomedicine 2:665
cognitive dissonance 20:695
complexities of 20:692–694
conferred by divorce 14:518–520
contradictions of risk 20:693–694, 20:696
coping 4:891–893
cost–benefit analysis 5:51, 5:53
decision under 14:863–864
decision-making 5:911
defined 2:397
defining 10:698
definition of 10:702
disaster 6:481
disclosure, measures of 4:579–580
disclosure of information 13:505
in drinking 21:609–610
in drugs 21:609–610
and entrepreneurship 7:675–679
environmental 7:784–790
of environmental changes 7:708–709
and environmental stress 7:804–805
expected value of 13:504
factors, of child neglect 3:413
future 20:696–697
gay/lesbian tourism 9:640–641
Giddens on 10:137
harm reduction 20:697
versus hazard 20:711
and health 10:571–572, 10:574–575
health insurance 10:640–645
health risk perception
assessment 10:702–703
numbers of 10:702–703
and heuristics 10:830–831, 10:834
human services 20:693–694
insurance and 22:362–363
'intuitive understanding' of 20:713
and law 13:504–508
law and 13:505
management 13:506–507
managerialism 20:693
need 20:692–693, 20:695
neoliberalism 20:693
offender rehabilitation 20:198
perspective on mass trauma 14:748
placebo-controlled trials
acceptable risk 18:168–169
low risks 18:168
no or negligible risk 18:167–168
preoccupation with failure 17:372
and probabilistic thinking 19:17
profits 19:111–112, 19:114
and punishment 19:659
reliability and validity of instruments 20:696–697
representation 20:711
resource demands of risk assessment 20:693–696
response to 13:505–506
and responsibility 20:581–582
risk management 20:695–696
sex as 10:813–814, 10:816–817
in sexual behavior 21:609
in smoking 21:609–610
and social exchange 8:485–486
social justice 20:696–697
social psychological perspectives 20:711–716
affect 20:713
hazard templates 20:714
identity process 20:715
mental models 20:712–713
personal variability 20:711–712
psychometric paradigm 20:711
social amplification 20:713–714
social representations 20:714–715
vulnerabilities 20:712
in social sciences 20:686–691
audit 20:686, 20:688
confidence 20:688
decision making 20:686, 20:689
evidence-based practice 20:690
managing abuse and violence 20:688–690
as new external condition 20:686–687
as new internal condition 20:687–688
sociological approaches 20:686
subjectivity 20:686–687, 20:690
trust 20:688
social sciences of 21:158
societal actors and 13:505
sociology and politics of 20:706–710
assessment 20:706–707
averaging 20:707
cost–benefit approach 20:706–707
cultural theory 20:709–710
and cultures 20:709–710
environmental justice 20:707–709
harm 20:706–707
hazard 20:706–710
hierarchy of uncertainty 20:708
ignorance 20:708
indeterminacy 20:708
normal accidents 20:708
reflexive modernization 20:709
reflexivity of 20:707–709
uncertainty 20:706–708
unknowns 20:708
sociology and politics of cost–benefit calculations 20:706–707
true 13:504
and trust 24:668–669, 24:671
and uncertainty 24:717

Risks (*continued*)
unintended consequences 24:747–748
use of term 13:504
in vocations 21:610
Risk-taking behaviors 18:820
and accidents 12:132
evolutionary developmental psychology 8:425–426
Risk-transfer war 16:307
Risk–value tradeoff 17:809–810
Risky adolescent behavior, evolutionary developmental model and 8:425–427
Risky behavior 21:858
Risky families 22:158–159
Risperidone 19:201
for schizophrenia 21:18–19
Rissho Kosei-kai 16:806–807, 20:409
Rites
ancestral 1:688–689
organizational symbolism 16:179–181
Rites (nonreligious), anthropology of 1:767, 11:55
Rites of passage 7:145, 14:131–133, 20:309, 20:311
and anthropology 1:752
death as 5:871–872
and hazing 10:556
initiation 1:767
life phase 5:507–508
in schools 21:57
in transition to adulthood 24:577–578, 24:577t
Ritter 1:965
Ritter, Karl 7:722–723, 16:334
Ritterschaft 1:965
Ritual approach, to film genres 9:161
Ritual classifications 21:73
Ritual commitment 20:719
Ritual dance 5:668
Ritual exchange, and materiality 14:766
Ritual experts 8:588–589
Ritual forms 2:886–888, 2:892
Buddhist 2:890–891
Tantric 2:889
Ritual healing 15:580
Ritual leadership 14:135
Ritual of consensus (American studies) 1:616–617
Ritual time 24:407
Ritualism, and anarchism 1:683–684
Ritualistic practices, and nationalism 16:310–311
Ritualizations 7:467–468, 10:276–277, 11:512–513, 20:309
Ritualized contexts, weaponry in 4:610–611
Ritualized homosexuality 19:753
Ritual/pro forma marriage 21:735
Rituals 3:534–537, 6:701, 6:704, 20:314, 20:717–720
action 20:717, 20:719
agent 20:719
ancestral 1:688–690
archaeology of 20:722–724
issues 20:725–726
ladder of inference' and processual approaches 20:722–724
material basis for 20:724–725
ritual as practice and structured deposition 20:724
beliefs 20:717, 20:719
cannibalism 3:86–88
chiefdoms 3:376, 3:379
cognition 20:719
communication 20:717–719
culture as 5:619–620
death 5:871
centrality of 5:870–871
mortuary ritual, changes in 5:872–873
death/dying 5:865–866, 5:868
direct historic approach 20:723
evolution 20:719
folklore 9:280–281, 9:283
iconographic representations in 4:612–613
of kingship 20:373–374
liminality 14:131–136
materializing social complexities in North America 17:19
media events as 15:48–53
and miracles 15:577
mystification 20:718
myth 20:717
and paleoethnobotany 17:460–461
performance 17:711, 20:719
prayer 20:719
profane 20:717
religion and new media 20:274–275
religious 20:308–309
sacred 20:717
sacrifice 20:717
and sacrifice 20:860–861
of secular religions 21:380–382
in South America 23:46
spatialilty 20:719
and symbolism 23:860–864
transformation 20:718–719
of transition to adulthood 24:579–580
tricksters 24:648–649, 24:651–656
of truth 23:348
Ritzer, George
on globalization 10:240
on glocalization and grobalization 5:539
on McDonaldization 5:536
works
The McDonaldization of Society 18:667
Rival consumption, international public goods 19:540–541
Rival socialism, and worldwide diplomacy 22:840–841
Rival therapy systems 19:513
Rivalries 10:92–93
River water sharing agreements 25:444–445
Rivera, Sylvia 24:552
Rivers, W.H.R. 14:453–454
Rivista di Filosofia 2:714
Rizzolatti, Giacomo 22:985
RJ *see* Restorative justice (RJ) programs
RJI *see* Reflective Judgment Interview (RJI)
RJT *see* Relative judgment theory (RJT)
RKE *see* Referent known exactly (RKE) model
RKO *see* Radio-Keith-Orpheum (RKO)
RLG *see* Research Library Group (RLG) (San Francisco)
RLS *see* Renzulli Learning System (RLS)
RM *see* Resource mobilization (RM) theory
RMA *see* Revolution in Military Affairs (RMA)
RMBS *see* Residential mortgage-backed securities (RMBS)
RME *see* Raw material equivalent (RME)
RMP *see* Random match probability (RMP)
RMS *see* Rostral migratory stream (RMS)
RMSEA *see* Root mean square error of approximation (RMSEA)
RN *see* Ramsey–Neumann (RN) cardinal
RNA *see* Ribonucleic acid (RNA)
epigenetic inheritance 7:833–834
RNA interference (RNAi) 7:834
RNAi *see* RNA interference (RNAi)
RNI *see* Relative noncentrality index (RNI)
RNM *see* Rule-based normativity of meaning (RNM)
RNR *see* Risk–needs–responsivity (RNR) model
RNs *see* Registered nurses (RNs)
RO *see* Renewables obligation (RO)
Road network 23:175f
Road pricing 24:632
Road Scholar program 1:150
Robbe-Grillet, Alain 1:540
Robberies 19:637f
Robbins problem 5:938
Roberts, John 13:768
Robertson, William 24:811
Robinson, Andrew, on anarchism 1:683
Robinson, Chase F., works, *Encyclopaedia of Islam* 11:61
Robinson, Douglas 18:648
Robinson, G. 9:624
Robinson, Joan 20:727–730
biography of 20:727
career of 20:727
contributions to economic theory 20:727
on distribution of income between wages and profits 20:729
dynamic analysis of capitalism by 20:728–729
on growth theory 20:728
on imperfect competition theory 20:728
interpretation of Marx 20:728
marriage and family of 20:727
on microeconomics 20:728
on neoclassical method of long-period equilibrium 20:728–729
on perfect competition theory 20:728
on 'representative' firm as an 'equilibrium' firm 20:727
works
The Accumulation of Capital 20:729
The Economics of Imperfect Competition 20:727–728
An Essay on Marxian Economics 20:728
Essays in the Theory of Employment 20:728
Work to Live 24:401
Robinson, Ronald 11:687

works, *Africa and the Victorians* 11:687
Robinson–Patman Act 1:799
Robo-Polls 18:462
Robot behavior 2:309
Robotic farmhands 1:474
Robotic technology 1:474
Robotics 5:646–648, 12:30, 21:187, 21:292
knowledge representation 13:100
technological change and 24:104, 24:106
Robotlike therapy 14:502
Robots, and autonomous agents, distinction between 2:309
Robust computerized test construction 4:513
Robust estimation
multilevel statistical analysis 23:382
order statistics 17:294
Robust nucleus of the archistriatum (RA) 2:692–693, 2:692f
Robust regression 20:735
Robust standard errors, in meta-analyses 15:275
Robust statistics, in qualitative linguistics 19:730
Robust time series, data adjustments 6:134–135
Robust variance estimation, in meta-analyses 15:275, 15:277–278
Robustness 8:600, 20:731–736, 21:529–530
for deviations 11:496
estimators 20:733–736
finite samples 20:735
highly resistant regression estimators 20:735
Huber's estimator, theoretical justification 20:734–735
least-squares 20:733–734
maximum likelihood 20:733–734
M-estimator 20:733–734
redescending M-estimators 20:735
weighted least-squares 20:735–736
indicators 20:732–733
breakdown 20:733
high resistance 20:733
influence functions 20:732–733
sensitivity functions 20:732
inference 20:736
personality-religion links 17:803–804
quantile regression 19:716
Robustness inference 20:736
Robustness issues in outcome-based sampling 17:103
ROC *see* Receiver operating characteristic (ROC) curve; Republic of China (ROC)
Rochester Child Resilience Project 10:672
Rock art 18:825–826
Australia and New Guinea 2:232–233
and gender archaeology 9:765–766
Rock music 25:816
Rockefeller, John, Jr. 3:5–6
Rockefeller, John D., III 17:39
Rockefeller Brothers Fund 19:556
Rockefeller Foundation (RF) 1:933, 3:5–6, 9:620, 13:636, 13:639, 18:295, 22:659, 22:993, 24:763–764
biomedical sciences 2:665–666
Rockefeller Institution for medical research (Rockefeller University) 24:763
Rockefeller philanthropies 22:696, 22:698
Rocker, Rudolf, on queer anarchism 19:747–748
RocznikiDziejowSpolecznych i Gospodarczych 11:76
Rod-and-frame illusion (RFI) 25:204
Rodents, serial order memory in 21:642–643
Rodinson, Maxime, works, *Islam and Capitalism* 16:382
Rodney King trial 18:499–500
Rodwin, Lloyd 20:147
Roe v. Wade (1973) 12:899, 18:948
Roeh 19:197
Roentgen, Conrad Wilhelm 18:83
Rogers, Carl Ransom 12:672, 12:674–675, 18:32, 20:737–743
applications to other fields 20:738–739
client-centered therapy 3:880, 19:504, 20:737–738
communication 20:737–739
counseling 20:737–742
criticism 20:741
education 20:737–738
education and influences 20:741–742
empathy 20:737–738, 20:740–741
groups 20:737–742
humanistic psychology 20:739–740
influence 20:740–741
initial impact on professions 20:738
nondirective approach 20:737, 20:742
personal life 20:742
person-centered psychotherapy 20:737, 20:739–741
political implications 20:739
positions held by 20:742
recording therapy 20:738
scientific research and theory 20:738
student-centered psychotherapy 20:738–739
works
Becoming Partners: Marriage and Its Alternatives 20:739
Carl Rogers on Personal Power 20:739
The Carl Rogers Reader 20:740
Client-Centered Therapy 20:738
Clinical Treatment of the Problem Child 20:742
Counseling and Psychotherapy 20:738
Freedom to Learn 20:738
On Becoming a Person 20:740
A Way of Being 20:739
Rogers, Natalie 20:740–741
Rogers, P.J., works, *Purposeful Program Theory* 14:299
Rogoff, Kenneth, works, *This Time Is Different* 9:183
Rohan, Henri de 12:387
Röhm, Ernst 9:647–648
RoHS *see* Reduction of Hazardous Substances (RoHS) directive
ROI *see* Return on investments (ROI)
Roi-thaumaturge 1:6–7
Rokeach, M.
theory of values 25:42–43
works
The Open and Closed Mind 2:198
Rokeach Value Survey (RVS) 25:42–43
Rokkan, Stein 3:647–648, 6:267, 20:744–747
biography of 20:744
on cleavages 20:745–746
comparative approach of 20:746–747
on democratization 20:745
educational career of 20:744
on internationalization of social sciences 20:744–745
on macro-contexts 20:747
on model-building 20:747
on state formation and nation-building in Europe 20:746
on structuring of mass politics 20:745–746
support of UNESCO for 20:745
works
Scandinavian Political Studies 20:744
Roksa, Josipa, on cultural capital 5:380
ROL *see* Rule of law (ROL)
Role conflict theory 23:429
of work–family conflict 8:760, 8:764
Role congruency model 18:828
of prejudice 18:828
sexism 21:698
Role differentiation 23:429–430
organizational groups 10:408
Role enactment of individuals 8:573
Role identity 23:433
Role identity model 25:270
Role modeling, and parenting 19:245
Role models, and cigarette smoking 22:126, 22:129
Role of associations 2:90–91
Role of celebrity 3:275
Role of in-house counsel 12:126–127
Role of managers 14:474–475
corporate control 14:475
decisional 14:473–474
informational 14:473–474
interpersonal 14:473
as negotiator 14:474–475
as power holders 14:475
Role of police 18:247
Role relationship surveillance 23:737
Role relationships 11:533
Role responsiveness 24:547
Role set 23:429
Role strain 23:429
Role theories 22:722–723, 22:959
of interviewing 5:852–853
and tyranny 24:712–713
Role-blurring hypothesis 8:761
Role-making 23:437
Role–person merger 23:432
Roles
of advanced practice nurses 16:911–912

Roles (*continued*)
individual and 17:933–935
life course 14:80–83
and marriage 14:601–602
of nonphysician health professionals 16:909
as process 23:431–433
in social anthropology and sociology 23:435–436
gender 23:437–438
role-making 23:437
role-taking 23:437
sex 23:437–438
in social networks 16:565
social structure as 22:714
stable expectations 23:430–431
and status, social psychology of 23:428
structural aspects 23:435–439
Role-taking 23:437
and emotions 7:535
reference group 20:72–73
Role-taking shifts 23:432
Roll, Barbara von 25:624–625
Roll, infant motor development 15:973–974
Roll-back neoliberalism 24:848
Roll-call analysis 19:695–696
Rolling back state 7:18
Rolling census 3:292, 3:298
Rolling gardens 11:203
Rolling sample survey, census 3:305
Rolling surveys 13:386
Roll-out neoliberalism 24:848
ROMA *see* RAPID outcome mapping approach (ROMA)
Roma 20:754
children 20:756
common myths about 20:755
cultural competence 20:758
and Eastern European social work 22:796–798
ethnic statistics of 20:754–755
future practice and research 20:758–759
homogeneity 20:754–755
marginalized ethnic community 20:755–757
migration 20:754–755
nomadism 20:754–755
origin of 20:754–755
Romani language 20:755
service users, social workers' engagement with 20:757–758
antiracist responses 20:757–758
racism 20:757
women 20:756–757
Roma people 20:748–753
bilingualism 20:751
community life 20:748–749, 20:751–752
cultural histories 20:752–753
diaspora 20:748–749
employment 20:750–751, 20:753
ethnic identity 20:753
families 20:748–753
language histories 20:749–750
music 20:750–753
occupational histories 20:750–751
overview 20:748–749
persecution 20:748–749
political histories 20:751–752
socialist wage labor 20:751
travelers 20:748–749, 20:751
unemployment 20:750
Roman alphabet 25:777
Roman Catholic canon law 5:227
Roman Catholic Church 3:233, 3:545, 3:548, 10:216, 13:758, 20:406, 21:173, 24:603
Roman Catholic missionary 15:614–615
Roman Catholic schools 4:137
Roman Catholicism 3:234–235, 8:286, 15:725, 20:288
in France 13:216–217
in Italy 13:218
in Mexico 13:218
in Spain 13:218
Roman Catholics (RCs) 20:293, 20:295
missions in Vietnam 3:532
Roman citizenship 1:700
Roman criminal law 12:364
case example 12:364
Roman dictatorship 6:371
dictatorship and revolutionary power 6:372
to modern dictatorship 6:371–372
origins and functions 6:371
transformation and end 6:371
Roman Empire 3:233, 3:548, 6:810–811, 16:297
in Europe 8:272
Roman imperial historiography 11:37–38
Tacitus 11:37
Roman *ius civile* (civil law) 13:915
Roman *jus gentium* 12:507, 12:509
Roman law 3:670, 4:248, 13:801, 14:243
compensation in 12:386
property rights 19:190–194
Roman Law of the Twelve Tribes (449 BC) 2:672, 22:393
Roman Mercury 24:648
Roman model - Cicero 3:645
Roman philosophy 25:599
Roman private law 12:507
Roman Republic 4:337
Roman republican historiography 11:37
Roman Society of Anthropology 8:218–219
Roman texts 13:777–778
Romance 9:650–651
LGBT friendship and 9:455–456
Romance etymology 8:208
Roman-Dutch law 3:669, 23:60
Romani language 20:749–750, 20:755
Roman–Jewish war (66–70 CE) 3:547
Romano-Germanic law 13:619
Romanov, Konstantin 21:741–742
Romans, Sarah 25:625
Romantic historism
nineteenth-century conceptions of history and historical science 11:73–74
Ranke's conception 11:73
Romantic liberalism, Christianity 14:8
Romantic love 7:515, 14:377–380
and personality psychology 17:909
Romantic orientation 9:650
Romantic paradigm, oral and literate expression 17:275
Romantic partners, trust in 12:638–639
Romantic pluralism, and culture 5:585
Romantic primitivism 17:158
Romantic relationships
and long-term heterosexual relationships 20:913–914
same-sex youth 9:654
Romantic science, Luria on 14:398
Romantic ties, and social networks 16:564
Romanticism 21:178, 24:504, 24:821–822
and authenticity 2:252
in Berlin 2:542, 2:546
and culture 5:585
enlightenment in 7:660
progress 19:125
Sarmiento on 20:907–910
Romanticization of rural life 17:35
Romdromas 4:382
Rome 3:233–234
early
class in 3:745
early empire 1:700–701
late republic 1:700
later empire 1:701–702
legacy of antiquity 1:702
mass society 14:743
and middle republic 1:699–700
psychology in 19:408
small states 22:108–111
Rome prize 24:883
Rome Statute 13:821, 18:501–502
Romer v. Evans (1996) 21:770
Rommel, Peter 16:681
Romney, Mitt 9:172
Roniger, Luis, works, *Globality and Multiple Modernities: Comparative North and Latin American Perspectives* 7:323
Rooke, A., on lesbians 13:882
Rookie draft 23:320
Roommate, hospital 22:738–739
Roos, Charles F. 9:220
Roos, J.P. 8:465
Roosevelt, Franklin Delano 14:18–19, 16:450–451, 22:698–699, 25:96
New Deal policies of 1:220
presidency of 18:861–862
urban planning by 1:220
Roosevelt, Theodore 1:827, 20:815
Root mean square error of approximation (RMSEA) 23:618
Rootless cosmopolitan 5:41
Root-mean-squares-error 4:475
Roots (linguistics) 15:820–821
Rorschach Inkblot Test 16:725
Rorschach Performance Assessment System (R-PAS) 19:131
Rorschach Research Foundation 19:131
Rorschach test 19:130–131
historical background 3:879
Rorty, Richard 6:335, 20:761–766
on antirepresentationalism 20:762–763

legacy of 20:765–766
life of 20:761
pragmatism of 20:763–764
scope and role of philosophy 20:764–765
works 20:761
Concept of Potentiality 20:761
Contingency, Irony and Solidarity 20:761, 20:763–765
The Linguistic Turn 20:761, 20:764–765
Method, Social Science and Hope 20:765
The Philosopher as Expert 20:764–765
Philosopher as Expert, The 20:764–765
Philosophy and the Mirror of Nature 20:761–762, 20:764
Philosophy as Cultural Politics 20:761
Philosophy as Transitional Genre 20:761
Poetry 20:761
Richard Rorty: The Making of an American Philosopher 20:765–766
ROS *see* Reactive oxygen species (ROS)
Rosaldo, Michelle Z. 1:236–237
works, *Woman, Culture and Society* 8:898–899
Rosaldo, Renato 1:236–237, 1:743
on democracy in North America 17:25
Roscher, Wilhelm 4:340, 10:874
Roscigno, Vincent J., on educational outcomes 5:378
Roscoe, Will, works, *Islamic Homosexualities: Culture, History, and Literature* 21:751
Rose, Nicolas 2:732
Rosen, Sherwin, on division of labor 6:604
Rosenbaum, J., on reputation of artists 2:57
Rosenbaum, P.S., work, *Minimal Distance Principle (MDP)* 4:829
Rosenberg, Alexander 7:102
Rosenquist, J. Niels, on alcohol consumption 16:558
Ross, E.A., works, *Social Control* 2:271
Rossby, Carl-Gustaf 7:793
Rossby waves 7:793
Rossellini, Roberto, works, *Rome, Open City* 9:168
Rossetti Infant-Toddler Language Scale 3:795
Rossi, Peter H., works, *Armed and Considered Dangerous: a Survey of Felons and Their Firearms* (*ACD*) 10:457–459
Ross's Island constraints 16:4
Rostovtzeff, Mikhail, works, *Social and Economic History of the Roman Empire* 22:311
Rostow, Walt Whitman 1:952
on planning and development 18:181
on scientific methods 21:150–151
works
The Stages of Economic Growth: A Non-Communist Manifesto 6:309, 18:724, 25:340
Rostral migratory stream (RMS) 2:824
Rostral prefrontal cortex 18:818–819
Rostral premotor areas 18:849–850
Rostral ventromedial medulla (RVLM), interoception 12:614–615
Rostrum 12:416
Rotating panel design 5:714
Rotation 8:693–694
Rotation length 13:240
Rotational indeterminacy, in factor analysis 8:706–707
Rote Armee Fraktion (RAF) 15:797–798
Rote learning 16:773
Roth, Guenther 25:465
Rothamsted Experimental Station (UK), Fisher at 9:222
Rothbard, M.N.
on libertarian anarchism 14:42–43
works
America's Great Depression 14:42
Man, Economy, and State 14:42
Rothbart, M.K.
on adolescent self-regulation 21:547
temperament development approach 24:193
Rothman, Lorraine 15:112
Rothman, N. 2:872
Rotikov, Kostya 21:742–743
works, *Drugoi Peterburg* 21:742–743
Rotter incomplete sentence blank (RISB) 19:133
Rotuma 17:158–159
Rouanet, H., on correspondence analysis 5:560
Rouch, Jean
works
Chronique d'une Été 6:622
Moi, un Noir 6:622
Rousseau, Jean-Jacques 1:540, 3:16, 3:647, 8:215–216, 11:885, 13:438, 20:767–771, 21:890–891, 22:697, 22:888
on democratic theory 6:90–91
education of 20:768–770
on freedom 9:388–389
on general will 11:817
life of 20:767–768
on political thought 18:481
reception and enduring influence of 20:770–771
on society 20:768–770
on theory of human nature 1:765–766
works 20:767–768
Confessions 20:767
Contrat social 20:767
Discours sur les sciences et les art 20:767
Dissertation sur la musique moderne 20:767
Emile ou de l'éducation 20:767
Manuscrit Favre 20:768
Mémoire à *Mr de Mably* 20:767
Nouvelle Héloise 20:767
The Social Contract and Discourses 7:665, 22:827, 23:353
Rousseau, L 4:751, 10:736
Route knowledge 23:164–166
brain systems mediating 23:166–167, 23:167f
Routine activities 4:836–838
and crime 20:772–773
minimal elements 20:772–773
and guns 10:458–459
and offender searches 20:774–775
Routine activity theory 19:163
and crime 23:145
of property crime 19:163
Routines
organizational path dependence 17:385–386
in sustainable practices 23:834, 23:836
Routing 10:174–177, 23:173–174
Routinization 17:68
of charisma 3:350–351
Routledge (London and New York City) 7:587
handbooks 7:588
reference works 7:588
Rowan, Brian 17:393–394
on organizational structures 5:537
Rowe and Kahn aging model 6:379
Rowland, Henry 24:763
Rowntree, Benjamin Seebohm, works, *Poverty, A Study in Town Life* 7:559
Row-principal biplot 5:3
Roy, Bill, on musical categories 3:800–801
Roy, O. 3:316, 3:318–319
Roy, Rammohan 20:374
Roy, William
on power in organizations 17:391–392
works
Socializing Capital: The Rise of the Large Industrial Corporation in America 17:391
Roy Très Chrétien 1:6–7
Royal academies, role in peer review 17:686
Royal Academy (Sweden) 2:89
Royal Academy of Science (Paris) 18:100–101
Royal Asiatic Society 22:616
Royal College of Psychiatrists (RCPysch) 5:473, 17:28
Royal Commission on Aboriginal Peoples (RCAP) 11:765
Royal Medico-Psychological Association 17:28
Royal postal service 12:491
Royal Proclamation of 1763 11:763
Royal Prussian Statistical Bureau 22:697
Royal Society of London 2:89, 9:222, 16:144
Royal Statistical Society (UK) 1:852, 9:222, 17:662–663
Royal Swedish Academy of Agriculture and Forestry (KSLA) 17:671–672
Rozanov, Vasilii, works, *People of the Moonlight* 21:740–742
Rozin, Paul, on food consumption 16:558
RP *see* Relapse prevention (RP)
RPAA *see* Regional Planning Association of America (RPAA)
R-PAS *see* Rorschach Performance Assessment System (R-PAS)
RPC *see* Reticularis pontis caudalis (RPC)
RPD *see* Recognition primed decision (RPD) model
RPE *see* Reward prediction error (RPE)
RPF *see* Rwandan Patriotic Front (RPF)
RPL *see* Recognition of Prior Learning (RPL)

RPS *see* Red de Protección (RPS); Renewable portfolio standard (RPS)
RRBs *see* Repetitive and restricted behaviors (RRBs)
RRC *see* Reverse Record Check (RRC)
RRI *see* Responsible research and innovation (RRI)
RRS *see* Risk reduction strategies (RRS)
r-RST *see* Revised Reinforcement Sensitivity Theory (r-RST)
RRT *see* Relational regulation theory (RRT)
RRTA *see* Real-time TA (RRTA)
RSAI *see* Regional Science Association International (RSAI)
RSC *see* Resting state connectivity studies (RSC)
RSCS *see* Regional Social and Cultural Studies (RSCS)
RSM *see* Response surface methodology (RSM)
RSOs *see* Reliability seeking organizations (RSOs)
RSPM *see* Raven's Standard Progressive Matrices (RSPM)
RSPS *see* Regional Strategic and Political Studies (RSPS)
RSS *see* Rastriya Swayamsevak Sangh (RSS); Rural Sociological Society (RSS)
RST *see* Reinforcement sensitivity theory (RST)
RSW *see* Radical social work (RSW)
RT *see* Response time (RT)
RTAA *see* Reciprocal Trade Agreements Act 1934 (RTAA)
RTM *see* Representational theory of measurement (RTM)
rTMS *see* Repetitive transcranial magnetic stimulation (rTMS)
RTP *see* Right temporal pole (RTP)
RTPJ *see* Right temporoparietal junction (RTPJ)
Rubber hand illusion (RHI) 21:420, 24:17
Rubbing mirror party 21:737–738
Rubin, Donald 5:91
  model of counterfactuals 5:91–92
Rubin, J., works, *Can Language Be Planned?* 13:354
Rubin causal model 19:742–743
Ruccelai, Giovanni, works, *Zibaldone* 20:427
Rucellai, Giovanni 3:15
Rudd, Kevin 2:237
Rudimentary Poker, game theory example 9:573–574
Ruesink, M., works, *Race and Justice: Wrongful Convictions of African American Men* 25:789–790
RUF *see* Revolutionary United Front (RUF)
Ruiz, Raul, works, *Of Great Events and Ordinary People* 6:622–623
Rule by records, concept of 8:193–194
Rule creators *see* Moral entrepreneurs
Rule enforcers *see* Moral entrepreneurs
Rule following 2:134, 14:841–849
  see also Meaning
Rule formalism 13:740
Rule learning 2:652, 15:155–156
Rule making, governance as 10:293, 10:295–296
Rule model, legal reasoning 13:781
Rule of critical testing 5:273
Rule of inheritance 1:968
Rule of law (ROL) 2:715–716, 3:676, 3:678, 4:250, 4:705, 12:138, 12:897, 12:900, 12:934, 13:427, 13:439, 13:445–446, 13:537, 13:539, 13:595, 13:721–722, 13:917, 17:620, 23:59
  anatomy and teleology 20:784–785
  arbitrary power 20:781–783
  authoritarian 13:432–433
  development of 13:722
  good governance as 19:691–692
  law and state 20:780–781
  legal and socio-legal 20:785–786
  republicanism 20:479–480
  Supreme Courts and 23:722
  thin and thick 20:783–784
Rule of reason analyses, antitrust 1:798
Rule of subtyping 13:944
Rule of the Sun 16:346
Rule of the wise, aristocracy 1:965–966
Rule of tit for tat 22:702
Rule ordering 18:77–79
Rule-based normativity of meaning (RNM) 14:844, 14:846–847
Rule-based reasonings 19:59
  of legal interpretation 13:786
Rule-explicit systems 13:79
Rule-governed behavior 2:432, 20:209
Rule-implicit systems 13:79
Rule-oriented 17:68
Rulership, ideology of 3:380
Rules 20:161–163
  algorithms 1:529, 1:531–533
  and bureaucracy 2:913–914
  of classical archaeology 3:762
  concept learning 4:538–540
  decision making 5:919–920
  and discretion 20:163–164
  institutions 12:190–194
  negotiated orders 17:411–412
  of recognition 13:764
  versus standard 13:744
  of state law 13:759
  in sustainable practices 23:834, 23:836
  tacit knowledge, Chomsky on 13:74–76
    faculty of language 13:75
    Quine's challenge 13:75–76
  of war 1:611
Rules Committee 4:651
Rules of Conciliation and Arbitration 13:917
Rules of thumb, for decision making 5:892
Ruling class 7:391, 7:393–394, 7:397
  elites 7:397–398
  and social stratification 22:709
Ruling through signals 10:294
Rumanian Church 17:435
Rumelhart, David E. 12:304
Rumination, for cardiac disease 3:113
Ruminative exploration 21:435
Rummel, R.J., works, *Statistics of Democide* 10:7
Rumors 20:788–794
  changing information ecology and new media 20:792–793
  characteristics 20:789–791
  definition of 20:788
  modern mediated societies and 20:791–792
  panic 20:790
  research history 20:788–789
  research on 20:789
Rural, and anthropology 1:751–753, 1:755
Rural Advancement Foundation International (RAFI) 11:307–308
Rural change, in Australia 2:239–240
Rural communities, kinship structure, in Africa 1:287–288
Rural community characteristics and crime 5:161–162
Rural contexts, community in 4:385
Rural crime 5:158–162
Rural criminology 5:160
Rural culture, Sarmiento on 20:907, 20:910
Rural development 1:484
Rural development policy 1:481–482
  agricultural education 1:481
  agricultural extension 1:481
  agricultural research 1:481
  legal institutions 1:481
  rural infrastructure 1:481–482
Rural gentrification 10:47
Rural geography 20:795–800
  cultural geography 20:796
  gender in 20:797–798
  legacy of political economy 20:795–796
  neglected rural others 20:796
  sexuality in 20:797–799
  as social construct 20:796–797
Rural guerrilla revolution 4:302–303
Rural History 17:670–672
Rural housing 20:806
Rural idyll 20:797
Rural industries 20:801–805
Rural infrastructure 1:481–482
Rural livelihoods, peasant agriculture 17:664–665
Rural manufacturing 11:854
Rural material civilization, and anthropology 1:755
Rural mythologies 5:162–163
Rural planning 20:801–806
  current priorities and approaches in 20:805–806
  defining 20:802
  in the development decades (1940s–1980s) 20:802–803
  Indian Model 20:804–805
  services-oriented 20:804
  Third World approach to 20:804
  US model of 20:803–804
Rural populations 5:158
  and community studies 4:370–371
  and cyberbullying 5:641–642
Rural social classes, peasants 17:676–678

Rural social work 20:807–813
climate uncertainty and globalization 20:808
defining rural 20:807
opportunities 20:810–811
policy 20:808–809
practice 20:809–810
improved model of 20:811
services 20:808–811
setting and context 20:807–808
Rural societies 17:669–675
after World War II 17:670–671
discussion about 17:672–674
1970s 17:671
1980s 17:671
1990s 17:671–672
terminology 17:669–670
before World War II 17:670
Rural Sociological Society (RSS) 20:814
Rural sociology 20:814–819
new 20:818–819
origins of 20:815–816
research themes 20:816–818
Ruralia 17:672
Rurality 20:795–796
Rural–urban continuum 6:284
Rural–urban dichotomy 24:961
Rural–urban fertility differentials 17:40
Rural–Urban interface 20:814–815, 20:818–819
Rural–urban linkages 24:961–962
Rural–urban migration 12:436, 24:962–963
Ruritanian allegory 9:673
Rush, Benjamin 16:908
Ruskola, Teemu 18:650
Russell, Bertrand 4:682–683, 16:583, 23:596
works, *Principles of Mathematics* 25:611
Russell, William Ritchie
on neuropsychological rehabilitation 2:815
works
The Traumatic Amnesias 2:815
Russell Sage centers 13:516
Russell Sage Foundation 2:446, 13:516, 13:518, 22:698
Russett, B.M., works, *World Handbook of Political and Social Indicators* 6:267
Russia 25:322, 25:324, 25:326–327
adult mortality in 1:162–168, 1:162f, 1:164f–167f
anthropology 1:754–755
church, state, and religion in 8:262–263
civil war 9:216
economic transformation 7:74–79
environment in 20:832–835
gender assignment in 9:797–798
modern, sexual minorities in 21:740–743
post-Soviet, homosexuality in 21:742–743
preevolutionary, homosexuality in 21:741–742
Soviet, homosexuality in 21:742
Western gaze 21:740–741
organized crime in 5:174
Orthodoxy 8:262
sample surveys 20:876
social work in 22:807–812
education 22:809–812
governance and autonomy 22:810–812
history 22:807
jurisdiction 22:810
postsocialist Russia 22:808–809
social policy 22:807–809, 22:812
social welfare under transition from socialism 22:807–808
training 22:809–810
and World War I 9:212, 9:214
see also Soviet Union
Russia Longitudinal Monitoring Survey (RLMS) 5:846
Russia–Georgia War 25:423
Russian and post-Soviet studies
on gender 20:836–842
on politics 20:843–850
on society 20:851–855
Russian and post-Soviet studies (politics)
crisis in Ukraine 20:849–850
economic transformation 20:847
post-2000 20:847–848
Soviet politics 20:843–845
transition to democracy and capitalism 20:846–847
from USSR to Russian federation 20:845–846
variation in rest of former USSR 20:848–849
Russian and post-Soviet studies (society)
culture 20:853
deviance 20:853
networks 20:853
research on Soviet society 20:851
Soviet ethnic and gender relations 20:852–853
soviet society, characteristics of 20:851
Soviet stratification 20:851–852
class 20:852
status 20:852
Russian Civil War 20:856
Russian colonialism 11:1
Russian Communist Party (Bolshevik) 14:658
Russian Constitution of 1906 21:741
Russian Cultural Revolution of 1860s 17:607
Russian eugenics 8:227–229
Russian Eugenics Society 8:228
Russian frontiers 9:466–467
anthropology 9:466
colonization 9:466
cultural studies 9:466
Turner thesis in 9:466
Russian grammar, fatalism in 13:289–290
Russian million-cities, skylines of 3:598f
Russian Orthodox Church 20:331
Russian privatization, in property rights 24:531
Russian Revolution 9:411, 20:639–640, 20:642–643, 20:856–859
and class consciousness 3:741
disintegration of empire 20:858–859
Iconoclasm 20:859
and Marxism–Leninism 14:657–658, 14:660–661
National Independence Movements 20:858–859
political constitution and system of government 20:856–858
social upheaval 20:858
Russo, Giuseppe Genco 14:422
Russo–Japanese War 16:276
Rustling livestock 5:159
Ruthven, M. 3:320
RVLM *see* Rostral ventromedial medulla (RVLM)
RVS *see* Rokeach Value Survey (RVS)
RWA *see* Right-wing authoritarianism (RWA)
Rwanda, education in 7:199
Rwandan genocide (1994) 17:635
Rwandan Patriotic Front (RPF) 25:384–385
Ryan, Caitlin 15:114–115
Ryang, Sonia 2:74–75
Ryba, R. 7:166
Rybczynski theorem 7:124–125
Ryle, Gilbert 1:747, 10:536

S

S and Marper v. United Kingdom (2009) 18:950
S language 23:379
SA *see* School achievement (SA); Selective availability (SA); *Stormtroopers* (SA)
SAA *see* Sample average approximation (SAA); Society for American Archaeology (SAA)
Saamaka, Maroons 14:592–595
Saami, territory of 1:927–929
SAARC *see* South Asian Association for Regional Cooperation (SAARC)
Sabel, Charles 3:23
Saber 12:757
Sabine, George, works, *A History of Political Theory* 18:483
SABRE (airline reservation system) 24:764–765
Sabsay, Leticia, on sexual citizenship 21:731
Saccades 3:344, 8:631–632
and attention 2:166
oculomotor control 8:629–630
Saccadic suppression 3:345
SACF *see* Sample autocorrelation function (SACF)
Sachs, Ignacy 23:798
Sack, Robert, works, *Human Territoriality* 18:375
Sackett, David 8:327
Sacks, Harvey 18:797–798
Sacks, K.B., on capitalism in North America 17:24
SACPA *see* Substance Abuse and Crime Prevention Act (SACPA)
Sacralization 20:861
Sacramento Area Latino Study of Aging (SALSA) 9:845

Sacred 23:594
  and religion 21:380–382
  ritual 20:717
Sacred and the profane 6:702, 23:594
Sacred habitus 10:484
Sacred History 14:252
Sacred landscapes, rituals 20:724–725
Sacred texts 20:331
Sacred works 3:684
Sacredness 7:375–376
Sacrifice 20:717, 20:860–862
  as catharsis 20:861–862
  as causality 20:861
  commemoration 20:860
  as communion 20:861
  expiation 20:860–861
  as gift 20:860–861
  initiation 20:860
  and long-term heterosexual relationships 20:915
  origins of 20:862
  religion 20:860–862
  ritual 20:717, 20:860–861
  as symbol system 20:861
  terminology of 20:860
Sacrifice zones 7:758–761
Sacrificial cannibalism 3:86
SAD *see* Separation anxiety disorder (SAD); Sexual aversion disorder (SAD); Social anxiety disorder (SAD)
SADC *see* South African Development Communities (SADC) region
Saddle points 9:575
  definition 9:575
  matrix game 9:575
  theorem 9:575
Sadism 13:888, 17:952
Sadness *see* Depression
Sadomasochistic (S/M) culture 13:887–889
SAE *see* Sigma Alpha Epsilon (SAE)
Saecularizatio 21:388
Saeculum 21:388
Saeed, Khaled 18:424
Saez measures of income 11:736–737
Safe Conduct Pass 25:399
Safe Dates program 12:681
Safe houses 24:119
Safe School Initiative (SSI) 21:108–109
Safe Schools Improvement Act 21:88
SAFEFE *see* South African Foundation for Economic and Financial Education (SAFEFE)
Safeguard measures, psychiatric ethics 8:69–70
Safer sex campaigns 9:642
Safer sex practice 21:799–800
Safety
  behavioral modifications for 12:133
  economics of 20:863–869
  employment, working conditions 25:698–701
  environmental 7:740–746
  environmental modifications for 12:133
  health and safety education 12:133
  physical 22:206
  risk of zero tolerance 21:107–108
  of school *see* School safety
  in sustainable transportation 23:848
  transportation 24:623–628
  workplace, and health 25:727–732
Safety auditing 17:370
Safety behavior 1:423
SAFHS *see* San Antonio Family Heart Study (SAFHS)
Saga, role in institutionalization 16:181–182
Sagar, Vidya 6:45
SAGE *see* Semiautomatic ground environment (SAGE); Student Achievement Guarantee in Education (SAGE) program; Study on global AGEing and adult health; (SAGE)
Sage (London) 7:587
  handbook series 7:587–588
Saguenay, mortality decline in 15:871–872
Sahara 16:840, 16:842
Sahlin-Andersson, Kerstin 17:393
Sahlins, M. 1:769, 5:557, 7:622, 24:293
  on history 1:748–749
  standard of goodness 1:749
  works
    Stone Age Economics 7:15
Sahoe 24:585–586
Sahul 2:229–231
Said, Edward Wadi 17:428, 18:649–650, 23:70
  legacies 1:770
  on orientalism 17:423–425, 17:428, 21:750
    critiques 17:425–426
  works
    Orientalism 1:770, 7:667, 11:20, 11:78, 16:371–372, 16:382, 16:457, 17:428, 18:647, 18:658–660
    The Question of Palestine 17:430
Saint Bartholomew's Day massacre of 1572 4:152
Saint Benedict 20:288
  rules of 2:61–62
Saint Francis of Assisi 3:355–356, 15:577, 20:288
Saint Guinefort 15:577
Saint Ireneus of Lyons Informational Consultative Center 16:808
Saint Louis 15:577
Saint Paul
  against celibacy 3:237
  issue of women covering their heads 3:237
  letter to the Corinthians 3:237
  women's speech 3:237
  works
    Letter of the Galatians 3:237
Saint Thomas Augustine 12:933, 18:32
  works, *City of God* (*De Civitate Dei*) 2:144–145, 3:708, 11:32
Sainthood
  modern 15:577–579
  before Protestant schism 15:577–578
  special commission for investigating miracles and saints 15:578–579
Saint-Simon, Henri de 7:390, 14:644
  visoion of technocratic society 15:576
Saint-Simonianism 22:697
Sakai (learning management system) 6:895
Salami slicing 20:530
Salamisation 21:314
Salary cap 23:321–322
Salduz v. Turkey (2008) 18:230
Sale of goods 3:30
Sales 20:17
  university press 19:627
  without advertising 1:209
Sales tax 23:340–342, 23:341t
Salganik, Matthew J.
  on market information regimes 12:591
  on music tastes 16:556
Salience 13:568, 20:790
  identity 11:526–528
  of religion 20:303
  sexual 21:761
  and social categorization 22:187–189
  social identity 21:456–457
Salient discrimination 16:199
Salinization, and irrigation societies 12:726
Sallo, Denis de 15:11
Sallust 11:37
Salomon, Albert 25:463
Salomon, Alice 9:800, 12:476–477
Salomonsen, Karla 7:934
SALSA *see* Sacramento Area Latino Study of Aging (SALSA)
SALT *see* Strategic Arms Limitation Talks (SALT)
Salutogenesis 10:612
Salutogenic responses, social work stress/trauma 18:708
Salvage anthropology 15:687
  and nostalgia 17:37
Salvation 12:645–646, 19:307, 20:366–367, 25:743
Salvation Army 22:388
Salvation history 11:32
Salvation religions 2:886
  Weber on 25:458–459
Salverda, W., works, *Oxford Handbook of Economic Inequality* 11:737
Salvo, Carl di 21:188
SAM *see* Search of Associative Memory (SAM) model; Selective accessibility model (SAM); Self-Assessment Manikin (SAM); Sympathetic adrenomedullary (SAM) system
Same (identity) 17:744
  everyday understanding of sameness of persons 17:746–747
Same–different experiment 21:950
Same-gender partners 10:802t
Samejima's model 19:437
Same-sex attraction, youth 9:652
Same-sex desire 19:769
Same-sex domestic violence 12:687
Same-sex marriage 8:769–770, 8:781, 9:635–636, 9:641

ban 21:770–771
legal status in Latin America 9:629
Same-sex orientation, biological influences 21:775–776
Same-sex sexualities 21:769
in Turkey 20:870–874
current situation 20:871–873
history 20:870–871
issues, visibilities, and absences 20:873
Samguk yusa 11:50
SAMHSA *see* Substance Abuse and Mental Health Services Administration (SAMHSA)
Sammaniyya Sufi order 12:738
Samoa 17:158–163
Samois 13:887
Sample autocorrelation function (SACF) 24:318–319
Sample average approximation (SAA) 23:474
theoretical properties of estimates 23:474–475
Sample average treatment effect on the control (SATC) 3:256
Sample average treatment effect on the treated (SATT) 3:256
Sample average treatments effect (SATE) 3:255–256
Sample extended autocorrelation function (SEACF) 24:318
Sample partial autocorrelation function (SPACF) 24:318–319
Sample Registration System (SRS) 9:80
Sample selection
by data quality 6:141
by migration 6:141
Sample sizes 13:385
genetic studies of behavior 9:936–937
in qualitative health research 10:695
for significance tests 21:961
Sample space 14:793
Sample surveys 5:742
accuracy of factual reports 20:883
background and history 20:881
cognitive aspects of 20:884
cognitive interviews in laboratory settings 20:884
impact of theories 20:884
cognitive aspects of survey design 20:881–885
context effects 20:883–884
design-based approach 20:886–889
economic statistics 5:758–759, 5:761
factual questions *versus* questions about attitudes or opinions 20:882–884
hypothetical survey 20:887
inference in 20:886–887
interviewer behavior 20:882
methodological transfers 20:881–882
model-based approaches 20:886–889
analysis 20:888–889
descriptive and analytic uses of surveys 20:887–888
small area estimation 20:888
stochastic models 20:886–888
survey design 20:889
Sample-based allocation rules 16:117–118
Sample-based surveys, for sexuality studies 15:316–317
Samples of Anonymized Records (SARs) 3:287
Sampling 3:299–300, 5:321, 9:276
analysis of variance 20:878
antecedents of modern survey methods 20:875–876
convenience 5:321
developmental research 6:327
in ethnography 8:186
history of sample surveys 20:875–880
Neyman's paper on 20:876–877
polling 18:459–463
probability 5:321
probability theory 20:878
purposive and random compared 20:876–877
in qualitative health research 10:695
in qualitative research 19:687–688
quantification 19:703–705
quota controls in 20:878
settlement and landscape 21:656
soils and sediments, geoarchaeology 10:55
statistical disclosure limitation 4:580
study of nonsampling error 20:878–879
systematic 5:321
theory of proportionate stratified 20:876
Sampling assumptions, discrete multivariate analysis 16:134
Sampling cultures 5:321, 5:325–326
Sampling errors 6:126, 20:878–879, 23:185–186
audience measurement 2:212–213
Sampling of bonds model 8:695
Sampling soils and sediments 10:55
Sampling subjects 5:325–326
Sampling techniques, empirical social research 7:563–564
Sampling theory
empirical social research 7:559, 7:563
Neyman's 16:827
Shannon's 9:377–378
Sampson and Laub's age-graded theory 14:60–61
SAMs *see* Social Accounting Models (SAMs)
Samuel, Ralph, works, *Explanation of Behaviour* 4:318
Samuelson, Paul Anthony 9:625
Samurai 24:583
San (Kalahari) 8:177
San Antonio Family Heart Study (SAFHS) 8:720, 8:720f
San Antonio Independent School District v. Rodriguez (1973) 18:736
San Francisco, gay male culture 9:711
San Francisco Bear weekend 13:891
San José Mogote 15:258
San Lorenzo (MesoAmerica) 15:257
San Lorenzo de los Negros (Veracruz) 14:593
Sancta Gallia 22:502
Sanctified place 18:146
Sanction threats 6:262
Sanctioning 17:6–7
Sanctions 13:807–808, 22:248–249
A's burden of response 20:890
concepts 20:890
conflict 4:622
diplomacy 20:891
efficacy 20:891
foreign policy 20:891–892
methodological problems 20:891–892
historical context 20:890
indicators of success 20:891
influence 20:890–892
jobs 12:827
legitimation 20:891
perceptions of deterrence 6:250–254
power 20:890
prospect theory 20:890
punishment 19:632–633
role of costs 20:890–891
in schools 21:56
and social dilemmas 22:512
Sanctions, discovery
in England 6:520
in United States 6:518
Sanction-service 4:622
Sanctity 15:577
Sandel, Michael 20:893–898
on American Revolution 20:895
biography of 20:893
on civic republicanism 20:895, 20:897
on democracy 20:895–896
on genetic engineering 20:897
influence 20:897
on liberalism and limits of justice 20:893–895
on moral limits of markets 20:896
Rawls and 20:893–895, 20:897
on self 20:893–895
Taylor and 20:893, 20:897
works
Case against Perfection: Ethics in the Age of Genetic Engineering 20:893
Democracy's Discontent: America in Search of a Public Philosophy 4:320, 20:893, 20:897
Justice: A Reader 20:893
Justice: What's the Right Thing to Do? 20:893
Liberalism and Its Critics 4:318
Liberalism and the Limits of Justice 4:317–318, 4:320, 20:893
The Procedural Republic and the Unencumbered Self 4:320
Public Philosophy: Essays on Morality in Politics 20:893
Sanders-Thompson multidimensional model, black racial identity 19:835
racial centrality 19:835
racial ideology 19:835
racial regard 19:835
racial salience 19:835
Sandfort, T.G.M., on mental health of LGBTQ youth 21:88
Sandpilot 14:135

Sandra, Morgen, works, *Wall Street Women* 8:902
Sanford, R.N., works, *The Authoritarian Personality* 2:197–198
Sanger Centre 11:313
Sanger sequencing 11:316–317
Sangha Act 1902 2:891–892
Sangomas 10:568
al-Sanhuri, Abd al-Razzaq 15:361
Sanism, in social work 15:200–203
  advocacy 15:202
  antisanist practice 15:202–203
  consumer/survivor 15:200
  current conceptualizations 15:200–201
  disability 15:201–202
  discrimination 15:200–202
  intersectionality 15:200–202
  mad 15:200–203
  mad studies 15:201, 15:203
  oppression 15:200–202
  responses 15:202
  rights 15:200–202
  social model 15:201–202
  stigma 15:201
  terms 15:200
Sanitation 25:428–429, 25:431–435
Sanitation programs 25:434
Sankoff model 11:25
SANS *see* Scale for the Assessment of Negative Symptoms (SANS)
Sanskrit language 2:888
Sanskritization 3:223–224, 11:896
Santa Clara Center for Occupational Health (SCCOSH) 7:370
Santa Cruz Women's Health Collective
  Lesbian Health Matters! 15:114
Santayana, George 6:690
Sao Paolo pride parade (2011) 9:640
SAP *see* Single assessment process (SAP); Société; d'anthropologie de Paris (SAP)
Sapir, Edward 1:743, 2:705–706, 6:245, 7:145, 9:459, 20:899–903, 25:575–576
  on American Indian linguistics 20:900
  Athabaskan, linguistic work on 20:899
  on cultural relativism 1:767
  descriptive linguistic work of 20:901
  dissertation on Takelma language 20:899
  general career of 20:899–900
  impact and current importance of 20:901
  scientific contributions of 20:900
  theoretical insights of 20:901
  views on history of language 20:901
  works
    Language 20:899
Sapir–Whorf hypothesis 5:314, 13:252–253, 13:288, 14:170, 14:209, 20:901, 20:903–906, 25:575
  behavior-centered approach 20:905
  domain-centered approach 20:905
  empirical research 20:904–905
  future prospects 20:905
  historical development 20:904
  language and thought 1:767
  linguistic determinism 20:903–904
  linguistic diversity 20:903–905
  nature and scope 20:903–904
  structure-centered approach 20:905
SAPROs *see* Social aspects/public relation organizations (SAPROs)
SAPS *see* Scale for the Assessment of Positive Symptoms (SAPS)
SAPs *see* Structural adjustment programs (SAPs)
SAR *see* Synthetic Aperture Radar (SAR) sensor
Sarariiman (salary-man) 2:788
Sarayaku v. Ecuador 11:801
Sarbanes–Oxley Act 2002 (SOX) 4:936–937, 13:917, 21:652, 23:493
Sarcasm, and pragmatic disorders 18:790, 18:792
SARF *see* Social amplification of risk framework (SARF)
Sargan's chi-squared statistic 23:617
Sargut, Gökçe 5:469
Sarkozy, Nicolas 12:239, 18:393
Sarmiento, Domingo F. 20:907–911
  on barbarism 20:908, 20:910
  biography 20:907–908
  caudillos 20:907–911
  on civilization 20:908, 20:910–911
  on education 20:907–909
  on Europeanism 20:910
  interpretation of *caudillismo* 20:909–911
  on liberalism 20:908
  literary legacy of 20:908–909
  on Pampa 20:910
  on progress 20:907–910
  on romanticism 20:907–910
  on rural culture 20:907, 20:910
  on travel literature 20:909
  on urban culture 20:907, 20:910
Sarraute, Nathalie 1:540
SARS *see* Severe acute respiratory syndrome (SARS)
SARs *see* Samples of Anonymized Records (SARs)
Sarton, George 15:248
Sartori, G., on party systems 17:567–568
Sartre, Jean-Paul 10:286–289, 19:314
  philosophy of 10:487
  works
    Critique of Dialectical Reason 1:748
SAS *see* Standard American Sociology (SAS)
SAS software *see* Statistical Analysis System (SAS) (software)
SAT *see* Scholastic Aptitude Test (SAT); Situational action theory (SAT)
Sataani 3:314–315
SATC *see* Sample average treatment effect on the control (SATC)
SATE *see* Sample average treatments effect (SATE)
Satellite accounts, in material flow analysis 14:762–763
Satellite broadcasters 16:776
Satellite imagery 10:53–54
Satellite systems
  in land cover research 13:221
Satellite television 16:776–777
Satellites 9:943
  GIS/remote sensing 10:64–66
Satiation, and eating 11:421
Satiation principle 22:541
Satiety
  and eating 11:421
  neurobiology of 11:421–422
Satiety sensitivity, weight reduction 25:470
Satire 11:417
  classical legacies 3:774
Satisfaction 17:336–337, 23:595
  of customer *see* Customer satisfaction
  in doctor–patient interaction 6:617
  in long-term heterosexual relationships 20:912–917
  volunteerism 25:268–271
Satisfaction account 14:192
Satisfaction with life scale (SWLS) 23:648
Satisfiability 14:319–320
Satisficing 5:911, 14:479–480
  for decision making 5:892–893
  organizational theory 1:103–105, 1:107
Satisficing behavior 18:290
Satisficing theory, and bounded rationality 2:774–775
SATT *see* Sample average treatment effect on the treated (SATT)
Saturation process 5:935
Saturnalia 3:145
SAT-V *see* Verbal section of the SAT (SAT-V)
Satyagraha (truth force) 18:425
Satyam Computers 3:47–48
Satyr 13:889
Saudi Arabia 16:802–803
  legal system in 15:365
Sauer, Carl Ortwin 1:846, 7:723, 9:420–421, 20:918–920
  bête noire 20:919
  on colonial destructive impacts 20:919
  on cultural diversity 20:919
  on cultural geography 10:85, 20:918
  on environmental determinism 20:918
  horizons of 20:918–919
  on modern mass production society 20:919
  natural science 20:919
  on plant domestication 20:918
  social sciences, distrust of 20:919–920
  works
    The Agency of Man on Earth 20:919
    Agricultural Origins and Dispersals 20:919
    Aztatlan 20:918
    Encyclopedia 20:919
    Folkways of the Social Sciences 20:919
    Ibero-Americana 20:918
    Man's Role in Changing the Face of the Earth 20:919
    The Morphology of Landscape 13:227, 20:918
    The Road to Cibola 20:918
Saunders, Peter 4:362
Saussaye, P.D. Chantepie de las 7:375

Saussure, Ferdinand de 2:705–706, 5:499, 5:501–502, 10:109–110
Saussure, Ferdinand de 20:921–925
on antireferentialism 20:923
on associative relations 20:924
on concept of linearity 20:923
Cours 20:922, 20:924
on course in general linguistics 20:922–923
on diachronic linguistics 20:922–923
on distinction between paradigmatic and syntngmatic axes 15:291
education and career of 20:921–922
on interdisciplinary legacy 20:921
on langue 20:922–924
on linguistics 11:512, 14:228–229, 20:921–922, 21:593–594
main tenets of course 20:923–924
on meaning 20:923
on model of communication 20:924
on parole 20:922–924
on reception and adaptation 20:924–925
on semiology 20:924–925
on semiotics 20:922, 20:924–925
on sign 20:922–924
on structuralism 20:924–925
summa cum laude 20:922
on synchronic linguistics 20:923–924
on syntagmatic relations 20:923
on systems 20:922
works
Cours de Linguistique Générale (*Course in General Linguistics*) 7:590, 14:169, 14:216–218, 20:921
Mémoire sur le systéme primitif des voyelles dans les langues Indo-Européennes 20:921–922
Savage, cannibalism 3:83–85, 3:87–88
Savage, Leonard Jimmie 5:857, 20:926–928
contribution to social and behavioral sciences 20:928
contribution to statistics 20:926–927
decision-making, concept of 20:927
Euclid of statistics 20:928
personal probability, concept and uniqueness of 20:928
poor eyesight of 20:926
on probability and utility 20:927
on prospect theory 19:262–264
scientific methods and 20:926–927
works
The Foundations of Statistics 20:926–927
Savage Minds (blog) 12:601
Savage neoliberalism 25:368
Savage paradigm 5:880–881
Savage thought, structuralism 23:629
Savagery 5:182
Savages 20:333
Savannah 3:312
Savannas, in West Africa 1:300–302
SAVE *see* Sliced Average Variance Estimation (SAVE)
Save the Children (non-governmental organization, UK) 12:961, 19:481
SAVI *see* Strength and vulnerability integration (SAVI) model
Savigny, Friedrich Carl von 4:339–340, 11:381
Saving strategies, job displacement insurance 12:829–830
Savings
and behavioral economics 2:440–441
economic psychology 5:888
and household wealth 25:450–452, 25:454–455
and income tax 11:745–746
money 15:737–738
and profits 19:113–115
and retirement 20:605
Savings groups, informal sector 12:57–58
Savior siblings 9:995–996
Savoir 7:215
Say, Jean-Baptiste 12:170
Sayigh, Rosemary, works, *Palestinians: From Peasants to Revolutionaries* 16:384
SBA *see* Small Business Association (SBA)
SBICs *see* Small Business Investment Corporations (SBICs)
Sblizhenie 9:466
SBM *see* School Business Manager (SBM)
SBMDP *see* School Business Manager Development Programme (SBMDP)
SBS *see* Social and behavioral sciences (SBS)
SBTC *see* Skill biased technological change (SBTC)
SC *see* Subject Condition (SC)
Sc scale 15:567
SCA *see* Survey of consumer attitudes (SCA)
SCA6 *see* Spinocerebellar ataxia type 6 (SCA6) cerebellar degeneration
Scaffolding 4:54–55, 16:505–506
apprenticeship 1:864
and coconstructivism 3:915
families as educational settings 8:731, 8:733
human development 25:318
and intentionality 12:347, 12:349, 12:351
Scalability
core–periphery patterns 4:915–916
tangible user interfaces (TUI) 24:24
Scalar equivalence 5:320, 5:323–324
Scalar property 24:308
Scalar timing 24:308–309
Scale analysis, human–environment relationships 11:407
Scale development, cross-cultural research 7:932
prepilot and cognitive interviews 7:932
Scale economies under competition 14:622–623
Scale for the Assessment of Negative Symptoms (SANS) 19:290–291
Scale for the Assessment of Positive Symptoms (SAPS) 19:290
Scale of analysis of research 7:415
Scale type, representational theory of measurement 14:858–859
Scale-based thinking, macro/microsociologies 14:417
Scale-Invariant Memory, Perception, and Learning *see* SIMPLE (Scale-Invariant Memory, Perception, and Learning)
Scales
cartography 3:168, 3:168f
construction 22:439
meaning of 21:1–2
in multisited field studies 16:105, 16:107
regional geography 20:119, 20:125–127
and social movements 22:439
spatial science 10:86–87
sustainable development 23:818–819
youth 25:826
Scales of measurement (Stevens) 23:454–455
Scaling 3:244, 3:621–623, 19:432–433
approximation methods 19:433
exact methods 19:432–433
mathematical psychology 14:814–815
multidimensional city-block scaling 19:432
RT and accuracy 6:399
in social measurement 21:5–7
dichotomous items 21:5–7
ordered assessments 21:7
unidimensional scaling 19:432
Scalogram analysis, in ethnography 8:189–190
SCAN *see* Schedules for Clinical Assessment in Neuropsychiatry (SCAN)
Scandals, and telephone 24:160
Scandinavia 2:921–922
clinical psychology, scientist-practitioner model 3:873–874
criminology in 5:240–241
Scandinavia eugenics 8:226–227
Scandinavian anthropology 1:754
Scandinavian Research Council for Criminology (SRCC) 5:240
Nordic Draftee Research (NDR) project 5:240–241
Scandinavian Social Democratic welfare model 16:465
Scanlon, T.M., contractarianism 4:807
Scanning 10:60
Scanning tunneling microscope (STM) 25:222–223
Scapegoating theory 18:840
Scapegoats 18:840
and gifts 10:140
Scar model of low self-esteem 17:918, 21:523
Scarcity
of leisure time 13:840–842
social influence 22:350–351
of water *see* Water scarcity
SCAs *see* Standard consolidated areas (SCAs)
Scatterplots 10:352–356
SCC *see* Supreme Constitutional Court (SCC)
SCCOSH *see* Santa Clara Center for Occupational Health (SCCOSH)
SCD *see* Social communication disorder (SCD)
SCED *see* Single-case experimental design (SCED)
Scenarios 24:127
of sustainable development 23:802–803

Scene documentation, in human remains excavation 9:335
Scene interpretation 25:154–155
Scenic resources 20:806
ß-centrality 16:535
SCF *see* Survey of Consumer Finances (SCF)
Schaap, Andrew 1:446
Schaefer, Fred K., works, *Exceptionalism in Geography* 20:123
Schaefer, William 25:786–787
Schaffer, Simon 8:577
  works, *Leviathan and the Air-Pump* 8:577–579
Schaie, K.W., stage theory of cognition 1:436–437
Schapire, Anna 25:90
Schattschneider mobilization 1:359
Schatz, E. 3:319
Schatz, Robert T., on patriotism 16:312
Schedule for Nonadaptive and Adaptive Personality 17:915
Scheduled castes 3:225–226
Scheduled tribes 3:226, 22:44
Schedules, reinforcement 20:208
Schedules for Clinical Assessment in Neuropsychiatry (SCAN) 15:182–183
Scheff, Thomas J. 7:380, 13:150, 13:152–153
Scheffler, Israel 7:259
Schein, Edgar, on organizational culture 5:536
Scheler, Max 4:341, 10:485, 18:435, 19:314, 21:261–262
  work
    The Place of Man in the Universe 19:314
Schelling, Thomas C. 1:668, 15:520, 22:702, 24:745
  model of residential segregation 20:551–552, 21:532–533, 22:702–704, 21:533f
  works
    Micromotives and Macrobehavior 22:513
    Strategy and Arms Control 1:976
    The Strategy of Conflict 16:431, 24:216
Schelsky, Helmut 13:106–107
  on scientific knowledge 13:105–106
Schema models, of semantic memory 21:577–578
Schema theory 13:709–710
  cognitive view 13:709
  contemporary view 13:709–710
  early 13:709
Schema therapy 4:544
Schema-focused psychotherapy (SFT), BPD 2:759
Schemas 4:19, 18:632, 18:784–785
  culture as 9:51–52
  for objects 16:483–484
  of semantic memory 21:573
  use of term 13:72
Schemata 4:189
Schematization 5:579
Scheper-Hughes, Nancy 1:734
  on brokers 2:873
Scherer's theory on emotional competence 8:673
Schicksalsgemeinschaft 8:274
Schiebinger, Londa 23:327–328
Schildkraut, Deborah, on national identity 16:312
Schiller, Friedrich, aesthetic education 1:229
Schiller, Glick 1:642
SCHIP *see* State Children's Health Insurance Program (SCHIP)
Schipperges, Heinrich 2:729
Schism, extreme-right parties tendency 18:414–415
Schismogenesis 16:151
Schistosomiasis 25:429–430
Schizoid personality disorder 2:757
Schizophrenia 2:453, 2:701, 2:810, 3:329–330, 7:608–610, 10:500, 10:513, 12:420, 14:441, 15:194–195, 15:211–212, 15:228, 17:737, 18:792, 18:821, 19:587–588, 20:271, 21:9–15, 23:332–333
  attenuated psychosis syndrome 21:12–13
  catatonia 21:9, 21:13
  causes and pathophysiology of 21:16
  changes in focus 21:13
  classification 21:11–12
  and creativity 17:772
  current and emerging directions in theory and research 21:13–14
  current knowledge 21:10
  definitions 21:9
  delusions 21:9–10, 21:12–13
  diagnosis 21:12
    criteria 21:11t
  disease course characteristics of 21:10f
  endocrine disorders 7:610
  epidemiology, illness-associated factors and burden of disease 21:10–11
  first-rank symptoms 21:9t
  functional studies 14:441
  hallucinations 21:9–10, 21:12–14
  HPA function 7:609
  intellectual background 21:9–10
  Kraepelin 13:129–131
  memory impairment 6:722
  mental–physical comorbidity 4:390–392
  metabolic consequences 21:21–27
    antipsychotic agents 21:22–23
    blood pressure 21:24
    cardiometabolic risk (CMR) in 21:21–22, 21:21f
    diabetes 21:23
    dyslipidemias 21:23–24
    metabolic syndrome (MtS) 21:24
    monitoring 21:25–26
    mortality in 21:21–22
    prolactin levels 21:24
    QT-interval prolongation 21:25
    screening 21:25–26
    vulnerability 21:22
    weight gain 21:23
  molecular genetics 15:211–212
  negative symptoms in 19:291
    biological correlates in 19:292–293
    course of symptoms in 19:292
    relationship between positive and negative symptoms in 19:292
    treatment in 19:293–294
  pathophysiology of 21:13–14
  prevention 21:12–13
  prophylaxis 19:200–202
  psychosis 21:9–13
  psychotic disorders 21:9–13
  risk for psychiatric illness 7:610
  role of gonadal hormones 7:609–610
  schools of thought and leading advocates 21:9–10
  sex differences 7:609–610
  stress 7:609
  structural studies 14:441
  subtypes of 21:16t
  symptoms of 21:17, 21:17t
  treatment 21:12
  treatment of 21:16–20
    antipsychotics 21:17–20, 21:18t–19t
    emergency treatment 21:19–20
    future perspectives 21:20
    pharmacotherapy 21:18–19, 21:18t–19t
    principles 21:17–18
    psychosocial treatment 21:19
  twin studies 15:211
Schizophrenic autism 2:701
Schlegel, Friedrich 20:324
Schleicher, August 14:209
Schleicher's model 13:300–301
Schleiermacher, Friedrich Ernst Daniel 10:785–786, 11:12–13
Schlesinger, A.M., Jr., works, *The Disuniting of America* 1:620
Schlick, Moritz 18:622
Schlick Circle *see* Vienna Circle
Schlieffen plan 9:212
Schmalenbach, Herman 2:271
Schmidt, Emerson, works, *Man and Society* 7:588–589
Schmidt, Johann (linguist)
  model of language contact 13:300–301
Schmidt, Johann 16:681
Schmidt, Raymund, works, *Die Wissenschaft der Gegenwart in Selbstdarstellungen* 6:496
Schmitt, Arno 21:751
  works, *Sexuality and Eroticism among Males in Moslem Societies* 21:750
Schmitt, Carl 15:520, 16:359, 21:28–33, 21:167, 25:409
  biography 21:28
  conception of democratic sovereignty 21:29
  on conceptual history 4:551
  on constitutional theory 21:30–31
  critique of liberalism 21:29–30
  on dictatorship 21:28–29
  on international legal order 21:31–33
  on liberal cosmopolitanism 21:31–33
  on sovereignty 21:28–29
  theory of democracy 21:30–31
  on war 25:409
  works

Concept of the Political 21:29, 22:828
Constitutional Theory 21:31
Nomos of the Earth 21:32
Political Theology 21:28
Schmoller, Gustav von 3:21, 4:340, 6:690, 10:873, 13:802, 21:168
Schneider, Arnd 2:18
Schneider, Arnd, works, *The Oxford History of Historical Writing* 11:46
Schneider, David 1:755, 24:293
Schneider, Jeffrey 9:645
on homosexuality in Germany 9:645
Schoenberg, Gustav, works, *Handbuch der politischen oekonomie* 7:585
Scholarly associations 2:88–91
current developments 2:90–91
early history 2:88–89
internationalization 2:90–91
modern history 2:89–90
social role 2:91
Scholarly book publishing 19:624–625, 19:627–629
Scholarly communication
and journals 12:862
new developments in 12:867–868
data-intensive science 12:867
identity and trust 12:868
mobile access and apps 12:867–868
new forms of electronic articles 12:867
publishing platforms 12:867
semantic web and linked data 12:867
social media 12:867
text and data mining 12:867
Scholarly community 2:512
Scholarly discipline 3:21–22
Scholarly infrastructure 22:596
historical development of 22:597–599
Canada 22:599, 22:598t
United States 22:597–599, 22:598t
Scholarly journals 12:862–868
Scholars
American studies 1:615, 1:619–620
presidency, values of 18:862
private lives of 21:34–38
antinepotism rules and their consequences 21:37–38
collaborative couples 21:34–36
familial arrangements 21:34–36
and students 7:170
Scholarship
censorship in 3:279–283
class actions 3:737
entrepreneurship 7:681–682
legal 19:807, 19:810
presidency 18:862
on public police 18:246–247
in social work practice 19:93–94
university press 19:624, 19:628–629
Scholarship granting organizations (SGOs) 21:71–72
Scholarship on cultural rights 5:493–494
Scholastic Aptitude Test (SAT) 12:723, 21:43–44
mathematics (SAT-M) 7:161
Scholastic texts 14:242
Scholasticism 3:234, 14:243, 19:408
legal reasoning 13:777, 13:779–781
Scholte, Bob 1:747
School achievement (SA) 21:42–54
complex relations 21:52–53
constructs relating to expectancy 21:49–51
constructs relating to values 21:51–52
determinants 21:48–49, 21:49f
developmental behavioral genetics and education 6:322–323
integrative approaches 21:52
and motivation 21:49
School and teacher effects, educational effectiveness research 7:225
School as workplace, in United States 24:77–79
School burnout 21:59–64
cynicism toward school 21:59
demands-resources model 21:60–61
developmental changes (during educational transitions) 21:61–62
exhaustion 21:59
future directions 21:63
person-oriented approach 21:61
prevention 21:62
sense of inadequacy 21:59
in social contexts 21:62
School business management 21:65–70
in Australia 21:68–69
brief history 21:65
current research 21:69
evolution of 21:67
future of 21:69–70
as a growing profession 21:67–68
international perspective 21:68–69
School Business Manager (SBM) 21:67
School Business Manager Development Programme (SBMDP) 21:67
School choice 21:71–79
catholic schools 21:73–74
charter schools 21:74–75
education tax credits 21:76
future research 21:77
homeschooling 21:76–77
private schools 21:73–74
scholarships 21:75–76
theoretical perspectives on 21:72–73
in United States 21:71–72
voucher programs 21:75–76
School collaboration
culture supportive of 21:117
evidence base 21:117
professional preparation for collaborative responsibilities 21:117
School composition effect 7:169
School culture, and school improvement 21:92
School curriculum, context and components of 4:435f
School dropouts 1:32
School education 8:130
School effectiveness 8:614, 21:80
and school improvement, link between 21:92–94
School effectiveness research 21:80–85
conceptual models and theory in 21:83–84
educational effectiveness research 21:80–81
educational production function 21:81
effects size 21:82–83
future of 21:84
integrated model 21:81–82, 21:82f
knowledge base on school effectiveness 21:81–83
meta-analyses 21:81–83
school improvement 21:80–81, 21:83–84
types of 21:81t
School engagement 21:59–64
demands-resources model 21:60–61
developmental changes (during educational transitions) 21:61–62
dimensions 21:59–60
absorption 21:59–60
dedication 21:59–60
energy 21:59–60
future directions 21:63
person-oriented approach 21:61
promotion 21:62
in social contexts 21:62
School environments 7:820–821, 24:67
child and adolescent psychiatry 3:385–386
dimensions of 7:820
and socioeconomic status 22:926
School evaluation 7:221, 7:220f
School functioning 21:80f
School funding, and academic achievement 1:46
School Health Education Evaluation 10:672
School improvement 7:218, 7:222, 21:80–81, 21:83–84, 21:91–95
advisory and research team 21:93–94
changes at classroom level 21:92
changes at school level 21:91, 21:93
dynamic approach to school improvement 21:93–94
dynamic model of educational effectiveness 21:93
educational change 21:91–92
educational goals 21:91–93
educational policy 21:92–93
factors of educational effectiveness 21:93
International School Improvement Project 21:91
practice and theory in 21:91–92
school culture 21:92
and school effectiveness, link between 21:92–94
school self-evaluation 21:93–94
school stakeholders 21:91, 21:93–94
student learning 21:92–94
teaching–learning processes 21:92
School inequality, and student achievement 5:346
School instruction 16:512–513, 16:515
School intake mix 22:328
School leaders 21:103, 21:137
School learning 1:870–873
apprenticeship models as solutions for problems with 1:870
cognitive apprenticeship 1:871

School learning (*continued*)
communities of practice 1:871–872
problem-solving 1:871–872
Reformpädagogik, situated learning 1:870–871
articulation and reflection 1:871
cooperation 1:871
as enculturation 1:871
multiple perspectives 1:871
by solving authentic problems 1:871
traditional 1:872
for transfer 21:96–100
School learning environment (SLE) 7:219
School libraries 14:46
School location 21:46
School management 21:101–106
curriculum and pedagogy 21:106
definition of 21:101
as a field of practice 21:101
as a field of study 21:101–102
innovation and future perspectives 21:105–106
relationship between change, reform, and innovation 21:105f
McKinsey & Company report 21:103–104
meta-analyses in different settings 21:102–105
OECD report 21:102–103
preparation and ongoing development 21:106
prognosis for 21:105
'techno-structural' interventions for 21:105
School of the Americas (SOA) 15:485–487
School performance 15:286
of minority students 8:145–146
School practices, history of 7:175–176
School psychology 24:525
School quality effects 3:930
School readiness
and media 15:6–7
and success 4:40
School resources
and schooling effects 21:121
and student achievement 5:344
School safety 21:107–111
Columbine aftermath 21:108–109
defining 21:107
discipline 21:108, 21:110–111
drugs and gangs 21:110
educational policy 21:107
future of 21:111
Gun-Free Schools Act 21:108
instruments of 21:109–110
measures, side effects of 21:110–111
risk of zero tolerance 21:107–108
Safe School Initiative (SSI) 21:108–109
security 21:107–111
violence 21:108
School shootings 10:457–458
School socialization, and arts 2:10–11
School teams 21:114
School to school collaboration 21:112–118
School violence and bullying 7:294
School-age children, marriage and fertility of 4:557–558
School–agency collaboration 21:116
School-average achievement 1:59–60
School-based programs, for mass trauma 14:749
School-based vocational education and training 25:245–246
dual apprenticeship system *versus* 25:248
versus general-based education 25:247–248
School-centered innovation 21:136
School-going child, concrete operational satge 18:128–129
Schooling 3:455, 9:74, 9:77–78
assessing total effect of 21:125–126
regression-discontinuity design 21:125–126
seasonality of learning 21:126
child, and poverty 3:406–407, 3:409
cross-cultural study of education 5:327–329
in eighteenth and nineteenth centuries 7:235–236
and employment 25:834t
and family 8:751–757
and fertility, in South Asia 9:80, 9:82
global model of 7:236–237
impact on cognitive and motivational development 21:119
importance of 21:125
intelligence 12:323–324
in Latin America and Caribbean 9:69
and migration 15:429
purposes of 21:55–56
total impact of 21:125–127
and twin studies 24:704–708
and youth culture 25:813–815
Schooling, effects of 21:119
on cognitive development 21:120–121
context-specific cognitive dispositions for achievement 21:120
disciplinary composition of research on 21:119–120
early findings and current wisdom 21:120–122
future directions 21:123–124
integrative frameworks on 21:120
as interdisciplinary field 21:120
methodological challenges in studying 21:123
measurement issues 21:123
on motivational development 21:121–122
practical implications 21:123
problems with differential intake or selection 21:123
school resources and 21:121
Student Characteristics 21:121
teaching quality and 21:121
tracking/composition and 21:121
Schooling (in North America)
conflict of ideals 7:298
colonization 7:298
curriculum 7:298
federal role in US public schools 7:298
language 7:298
native peoples 7:298
religion 7:298
effect of language 7:297
effect of religion 7:297
Schooling, and social inequality
basic mechanisms 22:326
democratizing policies 22:329–330
impact of organizational characteristics 22:328–329
large social gaps 22:325–326
role of family 22:326–327
school factors 22:327–330
story told by international comparisons 22:327–328
School-Level Environment Questionnaire (SLEQ) 7:820–821
Schoolmates, and student achievement 5:344–345
Schools 3:494, 7:165–171, 21:128–133, 16:867t
alternative 21:39–41
bullying prevention programs 10:671
choice 7:154–155
and communitarianism 4:315
completion rates 7:150, 7:150f
desegregation in 6:218, 6:221, 6:224–226
as differential learning environments 21:119–120
disciplinary backgrounds 21:119–120
drug abuse resistance education 10:671
and education 7:145, 7:148
environment for LGBTQ/sexual minority youth 21:86–90
contemporary LGBTQ youth 21:86–87
gay-straight alliances 21:88–89
homophobia and bullying 21:87–88
implications for research 21:89
implications for schools, parents, and students 21:89
inclusive curriculum 21:88
national/state-level laws 21:88
finance 7:155
full service or community 8:612
evidence of outcomes 8:614–615
implication of 8:615
rationales 8:613
scope 8:613–614
health programs and personnel, implications for 10:672
impact of teacher leadership on 24:61
influence on academic motivation 9:680
inputs, impact of 7:151–152, 7:152t
as institutions 21:129
characteristics 21:130–131
historical relevance 21:130
learning 1:861–864
adolescence 1:861–863
coaching 1:864
cognitive apprenticeship 1:863–864
exploration 1:863–864
instruction as apprenticeship 1:863–864
modeling 1:864
scaffolding 1:864
school-to-work transition 1:862
self-regulation 1:863
situated learning 1:861, 1:864
vocational orientation 1:862–863

methodological considerations 21:132
micropolitics of 21:134–140
Napoleonic 25:364
as organizations 21:137–138
characteristics 21:137–138
Iannaconne's analysis 21:138
physical activity health promotion programs 10:671
religious 20:380
for religious professionals 20:330
self-esteem within 21:495–497
shared knowledge base 8:612
smoking prevention programs 10:670–671
as social systems 21:55–58
between-school allocation 21:57
interpretations 21:57
norms 21:56
rites of passage 21:57
sanctions 21:56
socialization 21:56–57
theoretical perspectives 21:56
within-school allocation 21:57
stress prevention programs 10:672
and student outcomes 7:152–153, 7:153f
teacher stress/burnout 24:72–74
of thought 21:128–129
characteristics 21:128–129
historical relevance 21:129
use of term 21:128
urban/rural setting, of Australian Aboriginal students 15:904
use of term 21:128
violence and aggression prevention programs 10:671
Schools and Staffing Survey and the Teacher Follow-up Survey 24:78
Schools of psychoanalysis 19:326–328
Schools of social work, and Eastern European social work 22:795–796, 22:799
Schools of thought
brain stimulation reward 2:843–844
hedonic 2:843
neuroeconomic 2:844
reinforcement learning 2:843–844
seeking and craving 2:843
of regional geography 20:121f
schizophrenia 21:9–10
tacit knowledge 24:5–6
community perspective 24:5
knowledge-in-practice 24:5–6
treatment research and access 8:31–32
School-to-Work Opportunities Act 1994 1:861
School-to-work transition 1:862
Schoolwide enrichment model 7:160–161
for gifted education 7:161–162
Schopenhauer, Arthur 19:411
on personality 17:910–911
works
The World as Will and Idea 17:910
Schor, J., works, *The Overworked American* 24:401
Schreiner, Olive 2:515
Schröder, Gerhard 8:499
Schroeder, Paul 1:826
Schroedinger, Erwin 20:287–288
Schröger, E., works, *Psychophysics Beyond Sensation* 19:475
Schuetz, Alfred 18:797–798
Schultz, Henry 9:626
Schultz, Jennifer, on music tastes 16:556
Schultz, Wolfram 16:640–641
Schumpeter, Joseph A. 2:190, 7:114, 7:686–687, 12:147, 12:169, 16:227, 17:504, 21:141–145
on capitalism 3:107
on democracy 21:141–144
on democratic theory 6:91
on division o labor 6:603
early life and career of 21:141–143
difficult mid years 21:142–143
works 21:141–142
on economic theory 21:141, 21:143–144, 22:358
on economics 21:141–144
on entrepreneur 21:141–143
on entrepreneurship 7:681–684, 19:114, 21:141–142
on innovation 12:160, 21:141–144
on Marxism 21:143–144
on monopoly 21:141–144
on socialism 21:141–144
theory of growth 21:338
in United States 21:143–144
works 21:143–144
Capitalism, Socialism and Democracy 3:95, 12:146, 18:483, 21:141, 21:143–144
Das Wesen 21:141–142
The Fiscal Crisis of the Tax State 21:143
Gustav vs. Schmoller und die Probleme von heute 21:143
History of Economic Analysis 20:661, 21:141
The Instability of Capitalism 21:143
The Sociology of Imperialisms 21:143
Theorie der wirtschaftlichen Entwicklung 21:141–142
Schur order 14:450
Schütz, Alfred 10:435–436, 18:11, 20:345–346, 21:146–149, 24:135
aims of 21:146
approach, intellectual sources of 21:146–147
Bergson and 21:146–147
on communication 21:146–148
Husserl and 21:146–147, 21:149
on interaction 21:146–148
on interpretative sociology 21:146
on language 21:148
Parsons and 21:146
on phenomenological sociology 21:148
on phenomenology 18:11, 21:146–149
on philosophy of modern anonymity 21:149
on postulate of adequacy 21:148
on pragmatic relevance 21:147
on qualitative sociological research 21:148
significance of approach for social sciences 21:148–149
on sociology of everyday life 21:147–148
theory of life world 21:147–148
consequences 21:148
on vitae 21:146
Weber and 21:146–148
works
Der sinnhafte Aufbau der sozialen Welt 10:435–436
The Meaningful Constitution of the Social World 10:786
The Phenomenology of the Social World 17:391
on worldly phenomenology 21:149
Schutzstaffel (SS) 16:356
Economic and Administrative Main Office 11:153
Schwartz, James, works, *Pursuit of the Gene* 16:17
Schwartz, Richard 13:566
Schwartz, S.H.
theory of values 25:43–44
circular continuum of 25:43–44, 25:44t
common characteristics in different theories 25:43
empirical verification of circular model of 25:44
motivational continuum of 25:43f
Schwarz, James, works, *In Pursuit of the Gene* 16:17
Schwarz Bayesian information criterion 15:648
Schweickert, R., latent mental network theory of 12:80
Schweikard, Wolfgang 8:208
Schweizer, Thomas 16:562
Schweizerische Volkspartei (SVP) 18:410–411, 18:415
Schwendinger, Herman 22:386–387
Schwendinger, Julia 22:386–387
SCI *see* Science Citation Index (SCI)
SCIEM2000 Project 1:695
Science 11:380, 21:241–247, 21:297–301, 22:954
anthropological *versus* ethics 1:734–735
archaeology of 1:891–898
collaborative research 7:862, 7:864
as culture 20:509
and democracy 21:258
and development 21:150–155
disunity of 7:875–876
and division of labor 13:159
educational system in Latin America and Caribbean 7:180–181
elites 7:388
as endless frontier 17:681–682
enlightenment in 7:656–661
ethnomethodology 8:205–206
history and philosophy of 21:201
history of 13:133–134, 13:136–137, 17:99, 25:284–285
ideology 13:111–112
impact on society 21:282–283
interdependence with law 21:300
land change, origins and motivations of 13:220–221

Science (*continued*)
and law 21:156–162
decision science 21:158
disciplines and professions 21:158–159
DNA typing 21:161
forensic science 21:159–161
law as site of knowledge-making 21:160–161
natural law 21:156–157
posttraumatic stress disorder 21:159, 21:161
psychiatric expertise 21:159
psychology and law 21:158–159
recovered memory 21:159–160
regulatory science 21:157, 21:161
scientific evidence and expert witness 21:159–160
social sciences of risk 21:158
transition to modernity 21:156–157
U.S. Supreme Court 21:157–160
law's requirement from 21:297
versus legal inquiry 21:298–299
and legal philosophy 18:25
and magic 14:435–436
Marx on 14:642–643
masters and apprentices in 9:619
myth as primitive counterpart to 16:173
natural and human distinguished 10:785–786
new forms of 17:680, 17:682–683, 21:248–254
Academic Spring 21:248
blogosphere 21:251
citizen science 21:253–254
consequences 21:252–253
extended peer community 21:250–251
fetters 21:251
historical background 21:248–249
knowledge production cycle 21:249–250
pal review 21:252
persons 21:250
postnormal science 21:249, 21:253
problematic case 21:253
procedures 21:250–251
property 21:251
publicity 21:251–252
quality control 21:252
quality control cycle 21:249–250
quis custodiet 21:249–251
rat reality show 21:253
Science 3.0 21:254
science of bads 21:253
secrecy 21:251, 21:253
structures of scientific activity 21:249–250
and new genetics 9:1018–1019
pathologies in society 17:682–683
as performance 21:226–227
and philosophy 14:643
philosophy in league with 25:600
philosophy of 4:550–551, 13:133–134, 13:136–137, 14:313–315, 14:317, 14:871
see also Philosophy of science
and photography 18:83
and politics 21:167
in social science 21:167–171
progress of 19:123–126
purification of religion from 9:106–107
Putnam on 19:663
realist metatheory of 10:787–788
regional geography as 20:122–123
and religion 20:284–291, 21:172–175
relativism about 20:236
versus religion 15:575–576
responsibility of 20:524–526
role of gatekeeper in 9:618–619
as self-validating enterprise 17:12
separate, behavioral economics as 2:443
as social institution 21:200–201
and social movements 21:176–181
of social order 22:554
social studies as 22:727–729
sociology of 13:115–116, 20:230
of state 21:212–218
sociotechnical approach to 5:523–524
tension in relationship with law 21:297–299
as topic 24:674–675
usefulness of experimental economics in progressing 8:550–551
visualization in 25:220–225
war as 15:506–507
as *Wissenschaft* 14:643
at work 21:227–228
Science 21:323
Science, geography of 21:236–240
fieldwork 21:237–239
global 21:237–239
histories, theories, and concepts 21:236–237
knowledge 21:236–239
laboratory 21:237–239
local 21:236–238
politics 21:239
scale 21:238
situating science 21:237–238
sociology 21:236
spatiality 21:236–237, 21:239–240
tracking science 21:239
Science, sociology of
classical anticipations 21:261
knowledge, social determinants of 21:261–262
production and consumption of scientific knowledge 21:261
science in/as society 21:265–266
of scientific knowledge 21:263–265
social organization of the scientific community 21:262–263
structural contexts of scientists' behavior 21:261
Science, technology, and society (STS) 24:135
anthropology of 21:182–185
mapping of field with nearest neighbors 21:182–183
multilocale ethnographies 21:182–183
open futures 21:184–185
peopling of technologies 21:182
translating legacy knowledges into public futures 21:183–184
Science, technology, engineering, and mathematics *see* STEM (science, technology, engineering, and mathematics)
Science 3.0 21:254
Science and engineering, growth of 20:496–497
Science and media 21:163–166
approaches to studying 21:165
audience responses to media content 21:165
political and economic interests of media producers 21:165
rhetorical meanings of media conventions 21:165
history 21:163–164
understanding 21:164–165
Science and religion
distinction between natural and supernatural 21:172–173
distinguishing between Eastern and Western religions 21:173
epistemologies of 21:173
individuality 21:174
in prehistoric times 21:172
religion of interest 21:173–174
scientific study of religion 21:174–175
speculation, contemplation, and introspection 21:172
time period of evolution 21:172–173
Science and Society: A Journal of Marxist Thought and Analysis 14:649
Science and state
in authoritarian states 21:214
authorities of 21:212–213
in democratic societies 21:214
institutional and intellectual orientations of science 21:216–217
interpenetration 21:216
postmodern condition and recognition 21:217–218
power 21:213–214
transformation of modern democratic state 21:215–216
during wars 21:214–215
Science and technology 21:212, 22:1
academic 24:758–759
challenges in higher education systems 24:759–760
European universities 24:758–759
at European universities 24:758–761
and health 10:575
social study of 21:186–191
at United States universities 24:762–766
Science and Technology (S&T) statistics systems 7:181
Science and technology studies (STS) 1:471, 1:896–897, 6:836, 12:171, 13:137–138, 13:205, 13:207, 18:6, 18:354, 20:93–96, 21:282, 23:324, 23:326–328
and cybernetics 5:645–646
ethnomethodology of 21:192–199

actor-network theory 21:196–197
computer-supported cooperative work 21:197
constructivism 21:195–196
conversation analysis 21:193–194
ethnography 21:195–197
Garfinkel on 21:192–197
indifference 21:192–194
information science 21:197
instructed actions 21:194
laboratory studies 21:195–197
normative and ethical considerations 21:194–195
possibility of scientific descriptions of human behavior 21:194
practice turn in science and technology studies 21:192
problem of description 21:194–195
reflexivity 21:193
scientific practice 21:193, 21:195
social practices 21:192
sociology of science 21:194
Strong Program 21:192–193
symmetry postulate 21:192–194
technical practice 21:197–198
topic and resource 21:193–194
transcendence 21:196
workplace studies 21:197
experts and expertise 21:206–211
history of 21:200–205
actor-network theory 21:203
emergence of new field 21:202
from epistemology to ontology 21:203–204
feminism 21:203
and gender 21:203
history and philosophy of science 21:201
laboratory studies 21:202
new modes of knowledge production 21:204
precursors 21:200–202
and public policy 21:204
science, technology and society 21:202
science as social institution 21:200–201
scientific controversies 21:202
scientific expertise 21:204
social constitution of knowledge 21:201
social construction of technology 21:202–203
social impact of technology 21:201–202
social shaping of technology 21:203
sociology of scientific knowledge 21:202
sociotechnical systems 21:200, 21:202–203
structure of scientific revolutions 21:201
technological turn in science studies 21:202–203
indigenous knowledge 11:779–783
innovation 12:171–172
scientific controversy in 21:284–285
Science as Culture 20:96
Science Citation Index (SCI) 21:262, 21:322
Science communication 21:219–224
in citizen science 3:631–632, 3:634
climate change 3:839–840
communication in science 21:219
deficit model of public understanding of science 21:222
dialogue 21:222
dominant model of 21:222
empiricism 21:219
as engagement 21:258–259
Enlightenment 21:219–221, 21:223–224
as immaterial labor 21:223–224
information 21:222–224
innovation 21:222–224
in network society 21:222–223
origins of popular science 21:219–220
as politics 21:221–222
as popular science 21:220–221
press office 21:221–222
public engagement 21:222
public opinion 21:221, 21:223–224
public relations 21:221–222, 21:224
public sphere 21:219, 21:221–224
publicity 21:220
rhetoric 21:220–223
science journalism 21:221
scientific community 21:219–221
theoretical models of 21:222
Science education 21:229–235, 21:278
assessment 21:232
assessment of 21:232
assumptions about learning and implications for teaching 21:231–232
attitudes towards 21:229–230
content and organization 21:230
context-based courses 21:230
conventional media and 21:233
cross-cultural study of education 5:327–328
defined 21:229
electronic media and 21:233–234
evolution of 21:229–233
changing purposes 21:229
formal 21:229–230, 21:234
informal 21:230, 21:233–234
integrated science courses 21:230
on nature of science 21:230–231
relationship with technology 21:230
research 21:233
design-based 21:233
mixed method 21:233
qualitative 21:233
school-age learners and 21:233
teacher education 21:232–233
Science fiction 16:188–190
Science for policy 9:552
Science in/as society 21:265–266
Science Indicators 21:322
Science journalism, and science communication 21:221
Science Laboratory Environment Inventory (SLEI) 3:826–827
frequency response alternatives 3:826
Korean version of 3:826–827
Science learning, and educational neuroscience 16:732–733
Science Linkage Indicators
research and development (R&D) in organizations 20:499–500
Science of anthropology 1:734–736
Science of bads 21:253
Science of climate change 3:856–858
Science of ideas 13:111–112
Science of logic 10:737–738
Science policy 20:529–530, 21:322, 21:325
Science studies 21:322–323, 21:325–326
archaeological history 1:895
technological turn in 21:202–203
Science Teachers' School Environment Questionnaire (STSEQ) 7:820
Science technology studies, and reproduction 20:453–454
Science wars 7:873, 7:875, 17:92–93, 18:311, 20:236, 21:310
Sciences Humaine 11:380
Sciences Morales et Politiques 2:89
Sciences of man 11:40
Scientia sexualis 21:750
Scientific accumulation, in EITMs 18:457
Scientific activity, structures of 21:249–250
Scientific American 18:719–720
Scientific archaeology 1:901–902
Scientific artifacts 25:221–224
Scientific audiences 21:225, 21:227
Scientific authority 21:206–209, 21:261–262, 21:266
Scientific capital, and public recognition 2:333
Scientific collaboration 21:34–35
Scientific Committee on Problems of Environment (SCOPE) 25:561–562
Scientific communications 21:322
Scientific communities 13:136–137, 21:288
and science communication 21:219–221
Scientific conception of world 14:315–317
unified science 14:315–317
basic thesis of empiricism 14:316
debate on protocol sentences 14:317
kernel of neopositivism 14:315–316
verifiability criterion of meaning 14:316–317
Scientific concepts 16:511, 16:515, 21:278
biological concepts 21:269–270
development in children 21:268–274
early understanding 21:268
physical concepts 21:268–269
processes skills 21:270–272
Vygotsky on 25:312–314
Scientific controversies 21:202, 21:281–286
actor-network theory 21:284
Bath School 21:283
breaching experiment 21:282
core set 21:283, 21:285
'distance lends enchantment' phenomenon 21:283
Edinburgh Strong Program 21:283
mapping 21:284
as method 21:282
over impact of science on society 21:282–283
priority disputes 21:281–282
research, strands of 21:282

Scientific controversies (*continued*)
at research frontiers 21:283–284
in STS 21:284–285
Scientific curiosity 21:271–272
Scientific cyber-infrastructure 21:189
Scientific decolonization 25:548
Scientific development, theory of Thomas S. Kuhn 13:134–137
Scientific de-Westernization, institutional and material context of 25:543–544
Scientific disciplines 21:287–290
disciplines as communication systems 21:287–289
future of 21:289–290
modern system of 21:289
unit divisions of knowledge 21:287
Scientific discourse 21:264–265
Scientific discovery 21:150, 21:264, 21:291
see also computational models of scientific discovery
Scientific empirical method 19:326
Scientific ethos 17:13
Scientific evidence 21:159–160, 21:297
law–science interdependence and 21:300
legal control of 21:299
scientific enterprise legal understanding and 21:299–300
tensions in law–science relationship and 21:297–299
culture 21:297–298
scientific *versus* legal inquiry 21:298–299
Scientific evolution, teaching and research infrastructure 22:667
Scientific excellence 9:618–619
Scientific expertise 21:204, 21:214, 21:216–217
Scientific explanation 21:302–307
causal/mechanical approach 21:302–303
in cognitive neurosciences 21:303
compatibility of 21:303–306
claims 21:303
current philosophical theories 21:302–303
compatibility of 21:303–306
Hempel–Oppenheim model 21:302–303
and idealization 11:517, 11:519–520
manipulationist approach 21:303
nomological explanations 21:303
theories 11:11–12
unification approach 21:302
compatibility of 21:303–306
Scientific facts 5:524
Scientific fictions 4:520–521
Scientific fields 7:282, 7:284–286, 22:679–680
Scientific foundation for relativism 6:590
Scientific freedom 16:16–17
Scientific geography 10:69
Scientific giving 3:365
Scientific governance, in new genetics 9:1016
Scientific heroes 9:991
Scientific humanism 16:398–399
Scientific images 25:221–225
Scientific inquiry
in counseling psychology 5:70–71
value-freedom of, Weber's 22:992
Scientific institutions 21:216
Scientific instruments 24:141
Scientific journals 20:530
Scientific knowledge 7:569–570, 13:2, 13:105–106, 13:109, 18:686, 21:213–214, 21:216–217, 21:308–312, 22:229, 24:141, 24:673–674
and business anthropology 3:5
ideologies and 11:576
theories
and pragmatism 18:809
Scientific knowledge, sociology of 13:133–134, 13:137–138, 20:228, 20:230, 21:202, 21:261–265
actor networks and social worlds 21:265
interests and knowledge change 21:264
laboratory practices and scientific discourse 21:264–265
scientific discovery 21:264
Scientific leadership research 13:655
Scientific literacy 21:222
Scientific management 15:675, 15:677, 18:664
I–O psychology 11:881
Scientific management movement 11:349–350, 25:659
Scientific merit 20:518
Scientific methods 17:660, 21:308
Savage's contributions 20:926–927
Scientific migrations 25:541
Scientific milieu 7:170
Scientific misconduct 20:530, 21:313–317
ethos of science 21:314–315
fabrication of data 21:313–314, 21:316
falsification of data and methods 21:313–314, 21:316
Merton on 21:314–317
social norms 21:314
sociological imagination 21:314–315
Scientific naturalism 12:858, 16:326–330
Scientific naturalist 16:326
Scientific neutrality 4:157
Scientific ontology 22:489–490
Scientific organizations 22:648, 22:654
Scientific output, and journals 12:866
Scientific paradigm 25:576
Scientific personality assessment 17:827
Scientific philosophy 9:118
Scientific pluralism 25:601
Scientific polling 18:459–460
Scientific positivism, and Cold War 4:122–123
Scientific practice 21:225–227
ethnomethodology 21:193, 21:195
visualization 25:220–225
Scientific progress 13:135–136
Scientific publications 21:288
Scientific quality of feminist research 9:717–718
Scientific race theory 18:436
Scientific racism 11:306–307, 17:22
Scientific realism 20:22, 22:986
Scientific reasoning 16:925–926
inferences 16:926–927
Popper on 18:526
Scientific reporting, climate change 3:839–840
Scientific research
funding, in United States 22:993
research censorship 3:282–283
Scientific revolutions 13:112–113, 13:116, 19:700, 21:177, 21:212, 21:242–244, 21:318–321, 22:591, 21:177t
antiquity 21:319–320
challenging objects 21:320
classical mechanics 21:318, 21:320
early modern 21:318–320
engineer-scientists 21:318–320
historical and sociological context 21:318–319
institutionalization 21:318–319
knowledge development 21:318–320
Kuhn and 13:134–135
medieval science 21:318–320
of modern age 20:428–429
Renaissance and 20:428–429
practical knowledge 21:319
progress of sciences 21:243
renaissance 21:319–320
structure of 21:201
structures and achievements 21:319–320
term 21:318
theoretical knowledge 21:319
Scientific rigor, in feminist research 9:718
Scientific risk assessment 20:709
Scientific role, institutionalization of 21:263
Scientific social justice 9:972
Scientific socialism 18:436, 21:178, 22:834
Scientific socialization-by-filiation 9:619
Scientific Society of Agroecology (SOCLA) 1:486
Scientific study of religion 21:174–175
Scientific theories, and reality 20:20–21
Scientific trust, and ethics 8:58–59
Scientific truth 13:587
accountability of 24:675–676
credibility of 24:674–675
economies of trust 24:676–677
evaluation and resources 24:673–674
Scientific uncertainty, in climate change 3:854
Scientific validity of expert testimony 8:584
Scientific world conception 25:91
Scientifically engineer negligible senescence (SENS) 17:175
Scientification of policy 18:312
Scientific-practical sense 15:124
Scientific-technological revolution 13:106–108
Scientism 22:987, 22:994–995
and positivism 18:626
Scientistic positivism 11:73–74
Scientist-practitioner model 11:882–883, 24:517, 24:524–527
Committee of Training in Clinical Psychology 24:526
counseling 24:525–526

Freud on 24:524
historical background 3:880
James on 24:525
mental illness 24:524
psychoanalysis 24:524
Scandinavia 3:873–874
school psychology 24:525
in United Kingdom 3:874
Scientists
collaborative research between 7:864
motivations, and citizen science 3:633–634
responsibility of 20:524–526
Scientists' Oath 20:524–525
Scientometrics 21:322–327
delineation of reference sets 21:324–325
disciplinary organization of 21:323
future perspective 21:325–326
impact factors of journals 21:322–324
indicators for strategic choices 21:325
mapping of science 21:325
Science Citation Index (SCI) 21:322
theories of citation 21:325
use in policy and management contexts 21:325
use of term 21:322
Scientometrics 21:322
SCLC *see* Southern Christian Leadership Conference (SCLC)
SCL-90R *see* Symptom Checklist-90R (SCL-90R)
SCM *see* Stereotype content model (SCM)
SCMs *see* Social cognition models (SCMs)
SCN *see* Suprachiasmatic nucleus (SCN)
SCNT *see* Somatic cell nuclear transfer (SCNT)
Scopal quantifiers 19:709–710
SCOPE *see* Scientific Committee on Problems of Environment (SCOPE)
Scope economies, diversification 6:586
Scope principle 25:651
Scope quantifiers 19:709–710
Scopes trial 20:284, 20:287
Scopus 21:322
Score test, in hypothesis testing 11:493
Scores, for Maxdiff model, properties and uses 2:551–552
Scoring keys, in situational judgment tests 22:15
SCOT *see* Social construction of technology (SCOT)
Scotland
group identity 4:176–177
identity movements 11:544
Scott, David, on Caribbean anthropology 1:639
Scott, James C. 23:71, 25:98
moral economy model of 25:98
works
The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia 1:682–683, 11:86
Gender and the Politics of History 9:769
The Moral Economy of the Peasant 1:456–457
Scott, Joan 12:875
Scottish bourgeoisie 16:292
Scottish Centre for Social Research 23:749
Scottish Enlightenment 11:410–411
rational choice explanation 19:896
Scottish Hebridean Island, weavers in 5:109
Scottish national identity 11:544
Scottish National Party 10:382
Scottish universities 22:591
Scotus, Duns 14:868–869
Scoups database results 8:489t, 8:490f
SCP *see* Situational crime prevention (SCP); Social and Cultural Planning Office (SCP) (Netherlands, the); Structure-conduct-performance (SCP) paradigm
SCR *see* Skin conductance response (SCR)
Scree test 8:706
Screen theory 9:162
Screening 21:328–332
applications and future concerns 21:331
basic concepts 21:328–330
cancer 3:78–82
health 10:583
in military/disaster psychiatry 15:498
personalized medicine 17:925–926
systematic review 23:908
tests
estimating the accuracy of 21:330–331
importance of context in interpreting results of 21:330
Screening hypothesis 7:124
Screening of criminal cases 10:455
Screening questions 6:343
Screening tools 6:346
Script theory 21:849–850
for aggression 1:375
Scripts 4:19, 16:595–596, 16:594f
dramaturgy 6:646–647
reading 16:595–596
of semantic memory 21:573
Scriptural primitivism 15:725
Scruff 13:891
Scrutiny, managerial and organizational cognition 14:480–481
Scruton, Roger, works, *A Dictionary of Political Thought* 7:591
SCSL *see* Special Court for Sierra Leone (SCSL)
SCT *see* Self-categorization theory (SCT); Social cognitive theory (SCT); Social comparison theory (SCT); Sociocultural theory (SCT); Standard consolidation theory (SCT) of memory
S-curve measures 16:788
Scylla of blind historical forces 11:111
SD *see* Semantic differential (SD); Shortest derivation (SD); Sustainable development (SD)
SDA *see* Southern Demographic Association (SDA)
SDB *see* Sleep-disordered breathing (SDB)
SDF *see* Social Democratic Federation (SDF)
SDGs *see* Sustainable development goals (SDGs)
SDI *see* Strategic Defense Initiative (SDI); Survey of Dictionary-based Isms (SDI)
SDICs *see* Spatial Data Interest Communities (SDICs)
SDIs *see* Spatial data infrastructures (SDIs)
SDK *see* Software Developers Kit (SDK)
SDO *see* Social-dominance orientation (SDO)
SDP *see* Social Democratic Party (SDP)
SDQ *see* Self Description Questionnaires (SDQ)
SDR *see* Social desirable responding (SDR); Social discount rate (SDR)
SDS *see* Self-directed support (SDS)
SDSS *see* Spatial decision support systems (SDSS)
SDSS tool *see* INDEX (SDSS tool)
SDT *see* Second demographic transition (SDT); Self-determination theory (SDT); Signal detection theory (SDT); Social dominance theory (SDT)
SDU *see* Situation–decision–utility (SDU) triplets
SE *see* Social entrepreneurship (SE); Subject experiencers (SE)
Sea Shepherd Conservation Society 23:814
SEAbank *see* Social Ecological Assessment Databank (SEAbank)
SEACC *see* Southeast Asian Cultural Collection (SEACC)
SEACF *see* Sample extended autocorrelation function (SEACF)
Seafarers 16:345
Seafaring navigation 18:100
Seale, Clive, works, *Constructing Death* 7:596
Sealed-bid auctions 2:205–209
SEAMEO ARCAFA (Southeast Asian Ministers of Education Organization Applied Research Centre for Archaeology and Fine Arts) in 1971 23:64
Seamless web and symmetry 24:137–138
Search
artificial intelligence
symbolism 2:35–36
economics of 12:101, 21:333–339
heuristic search 2:32
minimax search 2:33, 2:33f
Search bias, meta-analysis 15:267–268
Search externality 21:335
Search for truth 5:230–231
Search Institute 3:446–447
asset-building framework 3:446
Search movement 1:144
Search of Associative Memory (SAM) model 15:162–163
Search rules, heuristics 5:913
Search space, genetic programming 2:43–44
Search theory 12:100–101
formal development of 12:101–102
Searches (marriage) 8:832–833
Search/optimization cohesive group finding approaches 3:910–911

Searle, John Rogers 19:415–416, 21:340–345, 22:490–491
background 21:340
on joint action 12:840–841
philosophy of language 21:340–341
philosophy of mind 21:341–343
philosophy of society 21:343–345
reception 21:345
works
Intentionality 21:345
Seasonal adjustment 24:344, 24:347–353, 24:355
ARIMA models 24:347–348, 24:351–353
X11/X12 ARIMA 24:351–353
Census Method II X11 variant 24:351
characteristics of seasonality 24:348–349
model-based methods 24:351–352
regression methods 24:351–352
moving average methods 24:349–351
STAMP (Structural Time Series Analyser, Modeller and Predictor) 24:352–353
stochastic model based methods 24:352–353
TRAMO-SEATS 24:352
Seasonal adjustment by signal extraction (SIGEX) 24:352
Seasonal adjustment methods 24:347–348
Seasonal dummy model 24:355
Seasonal migrations 11:649, 15:449, 15:472, 15:636
Seasonal patterns, in stock markets 23:484
Seasonality, characteristics of 24:348–349
Seasonality effects 12:451–452
Seasonality of learning 21:126
Seattle Longitudinal Study (SLS) 6:381
Sebenius, James K.
works
3D Negotiation 16:431
The Manager as Negotiator 16:431
Negotiation Analysis 16:431
SEC *see* Securities and Exchange Commission (SEC)
Secession
and ethnic conflict 8:91–94
minority-group status 15:571
and territorial conflict 24:214, 24:216, 24:219
Sechenov, Ivan 19:413
Second demographic transition (SDT) 9:24, 9:775–776, 12:404, 21:346–348, 24:571, 24:575
attitudes 21:346–347
below-replacement fertility 21:346–347
cohabitation 21:346–347
europe 21:346–347
family 21:346–347
and fatherhood 8:840
fertility postponement 21:347
gender 21:346–347
individuation 21:346
inequality 21:347
lowest-low fertility 21:347
marriage 21:346–347
nonmarital childbearing 21:347
values 21:346–347
Second economies 18:691, 20:852–854
in socialist societies 22:851–852
Second funerals, Hertz's analysis of 5:871–873
Second generation antipsychotics (SGAs) 21:22–23
for schizophrenia 21:12
side effects 21:23t
Second generation discrimination, inequality 7:896
Second generation models 24:314
Second language (L2) 2:580–583, 20:9–10
mastering 4:296–297
Second language (L2) acquisition (SLA) 21:349–359
Second language (L2) learning 21:360–367
for adult 13:301–302
Second language (L2) research 9:486–487
Second law of psychophysics 19:471
Second law of thermodynamics 7:621
Second memory system 11:386
Second messenger molecules 16:753
Second modernity 2:396–399
Second recipients of assistance 18:709
Second shift 14:607
Second slavery 22:39
Second stage teachers in United States 24:79
Second Vatican Council 3:235, 24:604
Second wave 6:459–460
Second wave cognitive behavioral therapy 4:543–544
Second Wave feminist movement 9:729–730
Second World 6:276–277, 23:349
and political economy 1:938, 1:940
Secondarily generalized seizures, occurrence of 13:34
Secondary analysis of databases 5:817, 5:819–822, 5:831
Secondary appraisal, and coping 4:903
Secondary control
compensatory 4:814–815
in life management 14:87
selective 4:814–815
and self-efficacy 21:505, 21:507
striving 4:814
Secondary control behavior 4:813–816
Secondary cultural characteristics 1:731
Secondary data 5:692
Secondary data sets 5:692, 5:695–698
for applied research 5:694–695
availability of 5:693–694
Fragile Families and Child Wellbeing Study 5:697
Health and Retirement Study (HRS) 5:698
history of 5:692–693
National Center for Education Statistics Surveys Program 5:696–697
National Longitudinal Study of Adolescent Health 5:696
National Longitudinal Surveys (NLS) 5:695–696
Panel Study of Income Dynamics (PSID) 5:697–698
US Census Bureau Data 5:695
Wisconsin Longitudinal Study (WLS) 5:698
Secondary deviance 13:152, 22:236
Secondary drive reduction 12:228–229
Secondary education 7:173
LAC educational system 7:178–181, 7:183
Secondary emotional development 24:276
Secondary features, segments 18:68
Secondary gain model of pain 3:557
Secondary groups, group process 22:541
Secondary infertility 12:45–46
Secondary institutions, political function of 4:339
Secondary key species 7:785
Secondary motor cortex 5:24–25
motor regions along medial wall of frontal lobe 5:25
premotor cortex 5:25
Secondary negative symptoms 19:292
Secondary prevention (mental health) 18:884
Secondary prisonization 14:699
Secondary reinforcement 20:207
Secondary religions 20:388
Secondary rules 13:793
Secondary schools
completion rates 7:150, 7:150f
and education 4:635
and gender 7:117–120
Secondary sources, for legal databases 5:808–809
law reviews 5:808–809
periodical indexes 5:809
Secondary taste cortex 24:27
Secondary transfer effect 11:678–679
Secondary traumatization, in social work 18:710
Second-best solution 1:345–346
Second-best welfare policies 25:498–499
Second-generation immigrants 5:152–155
Second-generation rights 11:369–370, 20:681
Second-language acquisition (SLA) 14:535, 16:57–58, 16:62, 16:65, 21:360
core notions 21:350
definitions 21:360–361
dynamics of language learning across life span 21:353–356
late adult learners of second language 21:354–356
late child learners of second language 21:354
factors influencing 21:364–366
affect 21:365
age 21:365
aptitude 21:365
attitudes 21:365
environmental/setting factors 21:364
individual learner factors 21:364–366
learning strategies 21:365
learning styles 21:365
motivation 21:365
naturalistic *versus* foreign nature 21:364
personality 21:365
sex 21:365
social distance 21:366

technology 21:364
versus foreign language acquisition 21:360
historical context 21:349–350
interaction of bilingual factors 16:57f
interaction of languages 21:350–353
interactions at the sentence level 21:352–353
interactions at the word level 21:350–352
versus learning 21:360
research and teaching 9:328–329
theories 21:361–364
behaviorist approaches 21:361
cognitive approaches 21:361–362
interractionist approaches 21:362–363
sociocultural approaches 21:363–364
versus third language acquisition 21:360
in twenty-first century 21:366
Second-level personality 17:801
Second-order cybernetics 22:912
Second-order values 11:521, 11:523
Second-parent adoption 21:756
Second-wave behavior therapies 2:431–432
Second-wave behavioral economics 2:446–448
modeling dynamically inconsistent preferences 2:447
modeling prosocial preferences 2:447
Second-wave feminism 8:907, 8:913–914
Secord, Paul F., works, *Too Many Women* 18:590–591
SECPT *see* Socially Evaluated Cold Pressor Test (SECPT)
Secrecy 19:608, 21:262
in adoption 1:144–145
anthropology of 21:374–379
culture 21:374–375
paradox 21:375
personal secrecy 21:377
power 21:375–376
public secrecy 21:375–377
secret societies 21:376–377
state and corporate secrecy 21:377–378
study 21:378
and ignorance 11:579–580
impact on environmental policies in Soviet Union 20:833–834
Internet 21:378
knowledge 21:374–378
masks 21:374–375, 21:377
personhood 21:374, 21:377
practice 21:374
privacy 21:377–378
public bureaucratic 19:532–533
reading 21:375
research ethics 21:378
and science 21:251, 21:253
Simmel on 21:374–377
Secrecy Act 2009 (Sweden) 18:943
Secret societies 21:376–377
brotherhoods 14:424
in West Africa 1:301
Secret speech 22:851
Secretary problem 5:907
Secrete shame, and silence/violence 21:875–876
Secretion 17:608
Section of Geriatric Psychiatry of the European Psychiatric Association 10:105
Section of Old Age Psychiatry of the WPA 10:104–105, 10:107
Sections of disciplinary associations 22:676
Sector regulation 20:170–171
Sectoral transformation of labor force 18:663
Sector/process, informal 12:56–57
urbanization and development 12:56–57
and West 12:57
Sects 20:360
political 21:380, 21:382
Secular changes 12:375–376
in obesity 17:78–79
Secular ethics 11:186
Secular nationalism 16:291, 16:295
Secular outlook 11:72–73
historical thought in eighteenth century 11:72–73
historiographical practice in eighteenth century 11:72–73
modern conception of history 11:72
Western civilization 11:73
Secular religions 18:438, 21:380–382
civil religion 21:381
conversion 21:382
defining 21:381
examples 21:380–381
implicit religion 21:381
political religions 21:380
quasireligion approach 21:382
religious forms, studies of 21:381–382
rituals 21:380–382
sacred 21:380–382
sect 21:380, 21:382
theoretical significance 21:382
therapeutic religions 21:380
Secular state 20:401–405
religious challenges to 20:401–402
Secularism 20:403
approach to peacemaking 17:637–638
and religion 20:308–311
Secularist Jewishness 12:871–872
Secularist movements 12:871
Secularity 20:465
and secularization 20:465
Taylor on 24:42
Secularization 3:540–541, 8:261–262, 13:216, 13:218–219, 18:434, 20:331, 20:360–361, 20:407, 21:382–394
age 21:392
beliefs 21:383–385
classic thesis 21:388
and communitarianism 4:310
death/dying 5:866
definition 21:388
differentiation and secularization of society 21:384–385
ecclesiastical use 21:383, 21:386
Enlightenment 21:386
etymology 21:388
as historical process 21:383–384
of law 12:751
linear-progressive process of 18:434
paradigm 21:384
philosophy of 21:390–392
post-secular 21:392
religion thesis
decline of 21:385–386
privatization of 21:386–387
sociology of 21:389–390
term 21:383–384
Weber on 25:458
west *versus* rest 21:392–393
Secularization theories 20:379
Secularization thesis 7:665, 21:388–389
Secularized Eastern European Jews 12:871
Secure attachment 12:230, 16:470
Secure knowledge 11:122–123
Secure–autonomous adults 2:139
Secured lending transactions 3:30
Securely attached children 2:137–138
Securities and Exchange Commission (SEC) 3:32, 13:457
Securities lawyers 16:418–419
Securitization 3:581, 3:583, 20:564–565, 21:395–399, 24:928
definitions of 21:395
environment 21:398–399
future avenues of research 21:397
global health 21:398–399
logics of 21:395–397
methods 21:397–398
migration 21:398–399
versus politicization 21:397
strategies of contesting securitization 21:398
Security 2:526, 2:791–793, 9:612–616, 22:572
commercial transactions 21:400
economic importance 21:405
English Common Law 21:402–403
French civil code 21:403–404
functional approach 21:404–405
German civil code 21:403–404
Graeco–Roman origins 21:401–402
defense economics 6:1–6
land tenure 13:234–235
Myrdal on 16:162–163
nomadism 16:842
rights 19:192–193
security interests 21:400–401
of status relations 10:763
studies 25:638
and values 25:27–28
women, national and international security 25:638–643
Security, attachment 12:231–232
attachment across life span 12:232
continuity and longitudinal correlates 12:232
maternal sensitivity 12:231–232
Security apparatuses 23:728
Security dilemma 3:716, 15:519

Security dilemma (*continued*)
and war 25:423–425
Security law
French civil code 21:403–404
German civil code 21:403–404
Security rights 19:193
Security sector reform and external intervention 18:270–271
Security statistics 5:836–843
challenges 5:841–842
contemporary trends 5:840–841
disaggregation 5:840–841
extension 5:840
individualization 5:840
data 5:839–840
research, role in 5:841
Security system 15:948
Security tools and technologies 18:263
SED *see* Survey of Earned Doctorates (SED); Survey of English Dialects (SED)
Sedation 19:466
Sedentarization 17:580–582
Sedgwick, Eve 19:747
works, *Epistemology of the Closet* 19:754
Sedgwick, Peter, on diseases 17:3–4
Sedimentation
geoarchaeology
principle 10:55
sampling 10:55
Sediments 7:692–695
categories 7:694
SEE *see* Stanford Engineering Everywhere (SEE) initiative.
See something, say something campaigns 24:120
Seed companies 1:474
Seed money 3:361
Seeing, infant gaze behaviors 12:28
Seeker-minded evangelicals 3:542
Seeking, brain stimulation reward 2:843
SEF *see* Social–ecological framework (SEF); Supplementary eye fields (SEF)
SEFA *see* Stepwise exploratory factor analysis (SEFA)
SEG *see* Synthetic extinct generation method (SEG)
Segmental units 15:827
Segmentation 16:843, 19:354
Segmented assimilation 2:84, 2:86, 8:159, 12:236–237
Segmented Discourse Representation Theory 15:754
Segmented labor-market theory, and migration 15:467–468
Segmenting 9:276
Segments 18:66–68, 18:75, 18:77–78
consonants 18:66–68
secondary features 18:68
vowels 18:68
Segregation 2:83, 3:204, 6:536, 9:616–617, 11:233, 11:235–237, 12:141–142, 16:635, 18:597–598, 18:600–601, 20:549–554, 22:266
African-American households from white households 19:820
continuing 20:551–552
demographic effects of 21:407–410
community 21:408
contextual effects 21:407–409
ethnicity 21:407–409
evolution of empirical investigation 21:409
housing 21:408–409
mechanisms 21:408
methodological issues 21:408–409
neighborhood 21:407–409
poverty concentration 21:408–409
race 21:407–409
spatial models 21:407–409
urban areas 21:407–409
desegregation 6:218–228
era of racial segregation in housing 19:820
and ethnic conflict 8:93
ghetto 10:123–125
job 6:537, 6:540
and migration 15:427, 15:430–431
minority-group status 15:571–572
and neighborhood effects 16:443–444
postsocialist cities 3:605–608, 3:610
by race 19:820–821
racial segregation in American cities 19:823
residential 20:540–543
residential segregation analysis 19:821–822
role in urban poverty 24:915–916
and spatial mismatch 23:157–158
urban 24:801
urban sociology 24:930
and urban violence 24:949–951
Segregation indices 21:411–413
Seignobos, Charles 1:909
works, *Introduction des Etudes Historiques* 11:122
Seinsgebundenheit of knowledge 7:875
SEIS *see* Shared Environmental Information System (SEIS)
Seisin 19:193–195
Seitz, Frederick 7:794
SEIU *see* Service Employees International Union (SEIU)
Seizures 19:467
and alcohol 1:511
consequences 7:854–855
ictal period 7:854
interictal period and comorbidities 7:854–855
postictal period 7:854
preictal period 7:854
definition 7:852–853
electroencephalography for 7:354–355, 7:357, 7:359, 7:361–362
etiology 7:852–853
focal *versus* generalized 7:852
kindling as model for 13:34–35
consequences of multiple seizures 13:35
creation of seizure focus 13:34
interictal behavioral dysfunction 13:35
multiple seizure generators 13:34
occurrence of focal and secondarily generalized seizures 13:34
pharmacoresistance and conditioned tolerance 13:34–35
progression in severity of seizures 13:34
and transient epileptic amnesia 24:554–556, 24:558
Seken 22:346
Sekentei 22:346
Seki, Shinpachi 24:584
SEL *see* Social-emotional learning (SEL) programs
Sela, R. 3:317
Selassie, Haile 16:806
Selbsthilfe 14:425
Selected-response tests 3:819
Selecting from the supply (gatekeeping) 5:441
Selection 4:871, 15:851, 15:853–855
in ethnography 8:186
and functions 9:520–521
indirect selection 15:852
of individuals 21:328–332
marriage selection 15:854
models 15:653–658
in qualitative research 19:687
social selection 15:851
in structural contingency theory 23:613
Selection, optimization, and compensation (SOC) model of life span development 1:435, 2:351, 14:87, 14:103
childhood and adolescence development 3:447
self-regulation
during adolescence 21:549
self-regulation during adulthood 21:558–559
successful aging 23:669–671, 23:670t
Selection approach, of social comparison 22:210–211
Selection bias 3:205, 5:203, 14:354, 19:140, 20:152, 23:158
in causal analysis 3:247
evidence-based medicine 8:329
observational research 21:414–417
and polling 18:459
problem 21:414–415
program evaluation and 21:416–417
race–crime relationship 19:841–843
selection on observables 21:414–416
self-selection 21:415
social experiment 22:294
solutions 21:415–416
statistics 21:414–418
Selection bias problem, in cross-national research 19:720–721
Selection differences 12:452
nonequivalent group designs 16:854–855
Selection differentials, order statistics 17:293–294
Selection effects 3:930
Selection modeling, nonequivalent group designs 16:853
Selection of judges 12:881–882
Selection of the best, order statistics 17:294
Selection on observables 21:414–416

Selection process, social exclusion 22:286, 22:288
Selection theory 8:472
Selection/development considerations, team performance measurement 24:97–99
construct selection 24:98
measure referent 24:98
measurement purpose 24:98
measurement source 24:98–99
measurement timing 24:99
Selectionism 22:26
Selectionist models 8:431
Selectionist social Darwinism 5:688–690
Selective accessibility model (SAM) 10:827, 22:213–214
Selective acculturation 2:85
Selective aphasia 2:574–575
Selective association 8:871
Selective attention 12:419
anatomical locus of 2:163–164
and bounded rationality 2:773
in concept learning 4:539–540
effects on representations of smell, taste, and flavor 24:29–30
Selective availability (SA) 10:175
Selective contracting 10:615–616
Selective exposure, in politics 14:713–714, 14:717
Selective immigration 11:630
in United States 11:630
Selective incentives 1:580–581, 4:146, 19:535
and social dilemmas 22:512
Selective influence
of experimental factors 12:80
task network models 16:541–542, 16:544
Selective neutrality (personality) 17:900
Selective optimization with compensation (SOC) 6:378, 6:380
Selective secondary control 4:814–815
Selective survival, and mortality crossover 15:849
Selectively influencing processes 16:541–542, 16:544
Selective-serotonin reuptake inhibitors (SSRIs) 1:778–779, 17:470, 18:8, 18:704, 19:203, 19:460, 19:464
obsessive-compulsive disorder 17:124–125
for obsessive–compulsive disorder 17:115
for for obsessive–compulsive disorder 17:113, 17:115–117
for perinatal mental illness 17:724–725
in sex therapy 21:691
Selectivity, and migration 11:650–651
SELF *see* Society for Ethical and Legal Philosophy (SELF)
Self 4:77, 4:152, 17:735–738, 19:400–405, 21:445–448
aging 21:427
analogies between, and brain 21:419–423
body and 2:720
and brain
action-awareness 21:419–421
affective self 21:419–420
body-awareness 21:419–420
face 21:421–422
face-recognition 21:421–422
frontal areas 21:421
future directions 21:423
insula 21:420, 21:422
interoception 21:419–420
parietal areas 21:420–421
self-recognition 21:421–422
self-representing self 21:422–423
sense of agency 21:420–421
sense of body-ownership 21:420
sensorimotor self 21:420–421
categorical 21:469
-centered conception of human interactions 4:311
challenges of conceiving 17:736–738
cultural psychology in 5:563–567
culture and 5:563–569
definition 5:563
describing, in collectivism/individualism 4:208
development in adulthood and later life 21:424–431
development of 20:72–73
direction 2:756
early development in culture 21:439–444
awareness of 21:440–441
behavior, regulation of 21:441–443
body 21:439–440
compliance 21:442
dialogue 21:442–443
directives 21:442–443
infancy 21:439–443
intersubjectivity 21:443
motor development 21:439–440
openness 21:439, 21:443
other-regulation 21:442
participation 21:442–443
pride 21:441
relation 21:439, 21:441–443
self-conscious affects 21:441
self-regulation 21:442
shame 21:441
shyness 21:441
as entity
cognitive perspective 19:401–402
psychoanalytic perspectives 19:400–401
and eudaemonism 8:212–215
existential 21:469
experiential theory 8:517–518
first-personal and social relations, analogies between
critical evaluation 21:451
joint activities 21:451–452
rationality over time 21:452
fulfillment of 17:737–738
group selves and multiple selves 21:452–453
ideal 21:470
identity 2:756
and identity 11:522–525, 11:532–535
identity over time 17:736
and identity/discourse 6:499–500
and illness behavior 11:599–601
imagery 19:401–402
implications 5:567–568
challenges 5:567–568
contributions 5:568
individual as 17:932–933
knowledge of 16:602
lack of integration 2:756
and learning 13:699–700
life span 21:425
and meaning 21:446–447
methodological individualism, implications for 21:453
models, and interpersonal relationships 5:426–427
and modernity 21:446
and multiple identities 4:178
needy self 17:440–441
nonatomistic view of 17:736
notion of 4:726–727, 19:363
and otherness 21:447
ought 21:470
person 21:445
personal identity, philosophical problem 21:449–450
philosophy 18:34–35
possible 21:470
and procrastination 19:75–76
and psychosocial development 7:939, 7:943–944
Rawls on 20:894
real 21:470
reasoning 4:152
-related brain-generated awareness 23:296
relation between person and 17:735–736
rise of 15:807
Sandel on 20:893–895
self motivation 19:402
self-affirmation and self-compassion 19:403
self-awareness 19:403–404
self-evaluation 19:403–404
self–other mapping 16:602–603
self-regulation and self-monitoring 19:403
Smith on 22:122–123
social cognitive views of 17:817–818, 17:818f
social comparison theory 19:404
social psychology 22:529
in social sciences 21:445–446
sociality of selves 21:450–451
and sociohistorical transformations 21:447–448
stability over time 19:402–403
symbolic interactionism 12:352–353, 12:355
and taboo 24:2
technologies of 5:647–648
theoretical premises 5:563–564
through interaction with others 19:404
Self Description Questionnaires (SDQ) 1:54–55, 21:461–462, 21:469
Self memory 2:289–290
Self psychology
conceptions of mind 22:762

Self psychology (*continued*)
psychodynamic treatment 22:764
psychopathology in 22:763
Self-acceptance, same-sex youth 9:654
Self-accounting 20:838
Self-actualization 10:522, 16:170
Self-actualizing tendency, in psychotherapy 20:740
Self-affirmation 19:403
and academic motivation 19:794
Self-affirmation theory, of cognitive dissonance 4:77
Self-aggrandizement explanations, of ethnocentrism and prejudice 8:171–172
Self-appointed policing 18:248
Self-ascription 8:103–104
Self-as-knower (the I) 21:460
Self-as-known (the Me) 21:460
Self-assessed health 20:600
Self-Assessment Manikin (SAM) 7:521
Self-attributions, and self-regulated learning 21:543
Self-awareness 4:285, 5:563, 8:392, 19:328, 19:403–404
and authentic leadership 2:247
and communication 4:285–286
and deindividuation 6:38–40
and empathy 7:551
neonatal 16:487
Self-beliefs 17:871–873
Self-building 16:183
Self-categorization 10:761–762, 11:527–528, 11:530
and defensive helping 10:763–764
Self-categorization theory (SCT) 4:147, 11:526, 12:410, 21:455–459, 22:186–189
accentuation 21:456, 21:458
crowd behavior 21:457–458
depersonalization and self-stereotyping 21:455
group polarization 21:457–458
leadership 21:458
organizational behavior 21:458
perceiver readiness and fit as determinants of social identity salience 21:456–457
power 21:458
process 21:455–456
and social identity 22:314–316
social influence as a determinant of organized social behavior 21:457
tests and applications of 21:457–458
and tyranny 24:713
Self-characterization 4:730
Self-commodification, and social media 24:148
Self-compassion 19:403
Self-competence scale 21:522
Self-completion theory, of motivation/actions 15:889
Self-complexity 6:432
Self-concept theories, of self-regulated learning 21:542
Self-concepts 6:430–434, 17:831–832, 19:401–404, 21:49–52, 21:433, 21:438, 21:460–468, 21:516, 21:522, 24:244–246
of ability 15:923–924
academic 1:54–63, 21:470–472
see also Academic self-concept
and academic achievement 21:471–472
effects 21:471–472
empirical evidence 21:471
achievement and 21:462–463
belief 21:469, 21:471
causal ordering of 21:462–463
championship performance 21:463
children's 21:462
conceptual issues 21:469
consequences of dimensional comparisons for 6:430–436
cultural views of 2:291
development 21:472–473
developmental trajectories across the life span 21:472
human needs underlying 21:472
impact of educational environments 21:472–473
developmental perspective of 21:461–462
different perspectives on attributes 21:470
domains, low correlations among 1:55–56
domain-specific
developmental differences in 21:492–493
individual differences in 21:494
explicit *versus* implicit 21:470–471
frames of references 21:463–465
big-fish–little-pond effect 21:464f
dimensional comparisons 21:464–465
gender differences in 21:462
general 21:470
and group identity 4:174
hierarchies and attributes 21:470
history and significance 21:460
and identity 6:501
implications 21:473
and implicit social cognition 11:714, 11:717
interdependent 5:426
independent 5:426
internal/external frame of reference (I/E) model 12:425–432
motivation 21:469–472
multidimensional perspective of 21:460–461
and personality 17:814–815
personality and 21:465–466
relationship between age and 21:461–462
research with Australian Aboriginals 15:903–905
self-efficacy 21:469, 21:471
self-enhancement model 21:462
skill development model 21:462
temporal comparisons 21:464–465
types and structures 21:470–471
Self-confidence 21:516
and self-efficacy 21:505–506, 21:508
Self-confrontation method 4:731
Self-conscious affects, and self 21:441
Self-conscious emotions 21:475–480
aggression 21:476, 21:478–479
authentic pride 21:478
behavioral pride 21:478
crime 21:476–477
cultural differences 21:478–479
embarrassment 21:477–478
causes 21:477
functions 21:477
individual differences in embarrassability 21:478
and shame 21:477–478
empathy 21:476–477, 21:479
hubris 21:478
moral emotions 21:476, 21:478
new directions in research 21:479
pride 21:478
types 21:478
shame and guilt 21:475–477
adaptive shame 21:477
implications for deterring transgression and socially undesirably behavior 21:476–477
implications for interpersonal adjustment 21:476
implications for psychological adjustment 21:476
maladaptive guilt 21:477
Self-consciousness 10:738, 17:736, 20:45, 21:449–450, 24:651–653
Self-consistency 15:942, 21:471–472
and cognitive dissonance 4:77
Self-constituting 11:523
Self-control 4:834–838, 9:407–408, 10:711–712, 15:908, 18:126, 20:334, 21:519
and adolescent health 1:123
and aggression 1:377–378
behavioral 2:410
and conscientiousness 4:658–660, 4:662
development of 21:481–482
and motivation 15:908
and self-regulation 21:557–558, 21:561–562
Self-control theory
and crime 12:462, 12:464, 21:481–485
gender 21:482
Gottfredson and Hirschi's general theory of crime 21:481–482
peer groups 21:482
stability 21:483
support 21:482–484
support for gender differences 21:483
support for parenting 21:483
support for peer groups and deviant behavior 21:483–484
desistance/persistence 5:138
of race–crime relationship 19:839
Self-corrective methods, in feminist research 9:718
Self-deception 17:93, 17:849–850
and procrastination 19:75
Self-defense 5:219, 5:223, 5:225
and powerlessness 1:545

Self-defensive denial 10:706
Self-delusion, and self-development 21:496
Self-denial, asceticism 2:60–64
Self-denigration 22:852
Self-destructive behaviors 20:334
Self-determination 3:520, 7:578, 9:909, 10:215, 11:369, 18:237–238, 18:240, 20:514, 22:890
  and advocacy 1:223
  in Arctic 1:930
  and choice 3:520–521, 3:523–524
  defined 7:578
  of peoples 20:480
  right to *see* Right to self–determination
  social work 24:755
Self-determination research 21:486
Self-determination theory (SDT) 1:32, 9:561–564, 12:697, 12:699–700, 13:833–835, 15:776–777, 15:931–932, 21:486–491, 21:525, 25:684
  implications for learning and instruction 15:931–932
  meta-theory of 21:486–487
  prosocial behavior 19:245–246
  teacher motivation 24:68
    consequences of 24:68
  workplace autonomy 25:719–720
Self-determined motivation 15:931
Self-development 11:537–539, 15:373–374
  during adolescence 21:432–438
    cultural context 21:436–438
    self-perceptions and self-esteem 21:433–434
  in childhood and adolescence 21:492–497
    abstract self-evaluations 21:492
    children's level of self-esteem 21:494–495
    cognitive development 21:493
    concrete self-evaluations 21:492
    cross-cultural considerations 21:495
    depression 21:497
    developmental differences in domain-specific self-concepts 21:492–493
    early adolescence 21:493
    early childhood 21:492
    global self-esteem 21:493–494
    individual differences in domain-specific self-concepts 21:494
    later adolescence 21:493
    later childhood 21:493
    mid adolescence 21:493
    middle childhood 21:492–493
    multiple selves 21:493
    narcissism 21:496–497
    self-delusion 21:496
    self-enhancement 21:496–497
    self-esteem movement 21:495
    self-esteem within both society and our schools 21:495–497
    social comparison 21:493, 21:495
  definition 21:432–433
Self-directed care, for older people 21:498–503
Self-directed search
  for disablism 6:450
  for vocational interests 25:239
Self-directed support (SDS) 21:498
Self-direction 2:756
  and values 25:28–29
Self-directive 17:767–768
Self-discipline, asceticism 2:60–64
Self-disclosure 17:444
  and intimacy 14:375–377
  and online dating 17:215
Self-discovery 2:317, 10:431
Self-distanced reasoning 25:592
Self-dysfunction 8:518
Self-efficacy 1:181, 1:329–330, 1:545–546, 4:820, 8:396, 10:585, 10:710–711, 10:714, 12:381, 15:942, 21:50, 21:504–508, 21:535–536, 21:538–539, 24:247, 25:539, 25:681–682
  action 21:504–508
  agency 21:504f
  assessment 21:517–518
  behavior determined by 21:511–512
  beliefs
    development of 21:506–507
    microgenesis of 21:507–508
  calibration 21:518
  and cardiac rehabilitation 20:188–189
  causal attribution 21:504–505, 21:508
  causes 21:504, 21:506
  concepts 21:504–505
  conceptual issues 21:512–513
    clarification 21:513
    distinguishing from other control constructs 21:512–513
    specificity *versus* generality of self-efficacy 21:512
  contingency 21:504–508
  control 21:504–505, 21:507–508, 21:504f
  control beliefs 4:819–820
  definition 21:509
  depression 21:505–507
  education aspects 21:508, 21:515–521
  educational and therapeutic aspects 21:508
  effects on achievement 21:517f
  and emotion regulation 7:454
  and epistemological beliefs 7:891
  externality 21:505
  factors affecting 21:517
  failure 21:504–508
  future directions 21:520
  goals 21:504–507
  health 21:505–506, 21:508
  and health behavior 10:577–580
  and health behaviors 10:585–586
  and health-related behavior 21:509–514
    determinants 21:510
    evidence for causation 21:510
    reciprocal determinism 21:510–511
  historical development 21:509
  history 21:505
  and hope 11:199
  illness 21:505–506
  as important element of happy and successful person 21:505–506
  internality 21:505
  learned dependency 21:507
  learned helplessness 21:507
  and learning 13:707
  means-ends-relations 21:504f
  method for increasing 21:511
    Bandura's four sources of information 21:511
    behavior change techniques 21:511
  motivation 21:505
  and motivation/actions 15:887–888
  optimism 21:504–505, 21:507–508
  outcome expectancies 21:509, 21:512
  perseverance 21:504–505, 21:507–508
  prediction of education outcomes 21:517
  primary control 21:505
  and procrastination 19:73, 19:75–77
  rheumatoid arthritis 20:655–658
  risk estimation 20:712
  role in multiple theories of behavior 21:509
  role in social cognitive theory 21:509
  secondary control 21:505, 21:507
  and self-concepts 21:469, 21:471
  self-confidence 21:505–506, 21:508
  versus self-confidence 21:516
  self-regulated learning model 21:518–519
  for self-regulation 21:518–519
  social cognitive theory 21:515–516
  social learning theory 21:509
  social models 21:519
  sources of information 21:515f
  stress 21:506, 21:508
  and sun protection 23:697–698
  teacher 21:520
  well-being 21:505, 21:508
Self-efficacy beliefs 15:933, 23:697–698
Self-efficacy improvement 10:624
Self-efficacy theory, and motivation 15:937
Self-employment 10:518, 12:61, 20:839, 25:696
  and entrepreneurship 7:676–677
Self-enhancement 5:564, 21:472, 21:526
  academic self-concept and achievement 1:56–57
  personality trait 17:817
  and self-development 21:496–497
Self-enhancement biases 5:426
Self-enhancement model 21:462
Self-enumeration 3:296, 3:299–300
Self-equilibrium hypothesis 21:434
Self-esteem 1:329–330, 3:929, 4:905–906, 10:431, 16:24, 19:401–402, 21:466, 21:469, 21:516, 21:522–528
  and body image 2:735, 2:737–739
  children's level of 21:494–495
    within both society and our schools 21:495–497
    competence-adequacy in domains of importance 21:494
    cross-cultural considerations 21:495
    incorporation of opinions of significant others 21:494–495
  conceptualizations 21:522

Self-esteem (*continued*)
contra-hedonic motivation 1:245
development in adolescence 21:433–434
developmental changes 21:526
enhancement 21:526
etiology 21:526
global 1:54–55, 21:493–494
and group belonging 10:423
identity, nonheterosexual British Asians 16:863
and implicit social cognition 11:714, 11:717
in interpersonal relationships 21:524–525
malleability of 17:816
measurment 21:522–523
implicit self-esteem 21:523
negative consequences 21:525–526
origins 21:526
and procrastination 19:75–76
protective function of 21:524
pursuit of 21:525
relation to important life outcomes 21:523–524
academics 21:524
aggression 21:524
career success 21:524
crime 21:524
delinquency 21:524
physical health 21:523
psychopathology 21:523
well-being 21:523
and religion 20:265
worldview verification model of 15:241
Self-esteem inventory 21:522
Self-Esteem Movement, and self-development 21:495
Self-esteem scale 21:522
Self-estimated personal intelligence 17:754
Self-estrangement 1:544, 1:547
Self-evaluations 6:431–432, 6:434, 7:577, 19:403–404, 21:522–523
of driving habits 24:625
of schools 21:93–94
social comparison 22:211–214
Self-evidence 4:2, 19:391
Self-expansion, and interpersonal attraction 12:627
Self-experience, infant gaze behaviors 12:30–31
Self-expression values, modernization 15:697–698
Self-face concern 11:186
Self-focus 6:39
Self-focused-oriented moral reasoning *see* Hedonistic-oriented moral reasoning
Self–friend similarity 5:427
Self-fulfilling prophecies 1:669–671, 6:21, 8:513, 14:877, 23:449, 24:186
population forecasts 18:595
Self-generated arguments 22:352
Self-governance 19:582
and citizenship 3:650–653
principle of 19:583
Self-government rights 16:26
Self-handicapping
by Australian Aboriginal students 15:902
and procrastination 19:75–76
Self-harm 12:755
Self-hatred 23:463–464
Self-healing, placebo-controlled double-blind studies 18:157, 18:161
Self-help 4:622, 18:730
principle of 19:583
in psychotherapy 19:379, 19:381
and relational social work 20:223
social control 4:825
Self-help clinics 15:111–113
Self-help groups 10:686–691
for mental illness 15:218–219
mutual aid self-help (MASH) groups 10:688–690
Self-help gynecology movement 15:111–112
Self-help legal assistance 12:925–926
Selfhood 17:933
Self-identification 12:761
butch–femme subcultures 13:872–873
same-sex youth 9:653
Self-identities 2:756, 23:433, 24:36
in Japan 12:794–795
Self-identity development
and adolescent social competence 22:218
Self-image, gastrointestinal disorders 9:608
Self-imposed limitations 14:206
Self-improvement 6:431–432
Self-inflicted harm 23:682–683
Self-insight 17:850
Self-instructional training
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
Self-instructions
metacognitive skills as 24:90–91
Self-insurance 6:274
Self-interest 12:386–387
and collective violence 14:703
and communes 4:267–268, 4:271
economic philosophy 7:102–103
and exchange 8:476–477
justice motive 12:937
microeconomics 15:332
Mill's views 12:388
and opportunism 1:348
and public interest 19:552
rational choice theory *see* Rational choice theory (RCT)
and social order 12:387
and virtue 12:387–388
Selfish decisions, neural basis of 16:612–613, 16:613f
Selfish genes 8:387, 22:910–911
Selfish herd 18:899
Selfishness 8:213, 12:391, 12:393, 15:309, 19:902–903
and interests 12:393–394
kin selection 13:29, 13:28t
Self-knowledge 11:533, 21:522
assessed by personality trait measures 17:815, 17:815f
Self-legitimization 24:848–849
Self-liking 21:522
Self-limiting institutionalization, of environmental movements 7:766
Self-limiting radicalism, of Green parties 7:766
Self-locomotion, and perceptual learning 10:130
Self-love 8:213
enlightened 12:387
and love of God 12:387
Self-made man 20:896–897
Self-management 20:675
of chronic pain 3:559–560
diabetes 6:338
respiratory disorders 20:577–580
rheumatoid arthritis 20:658
Self-mockery 11:417
Self-monitoring 15:284, 19:403
speech 23:257–258
weight reduction 25:472–473
Self-motion, multisensory integration during 25:68
Self-motivation 19:402
role in self-regulated learning 21:543
Self-mutilation 24:32
Self-myopia 11:524
Self-negating prophecies, population forecasts 18:595
Self-object, and psychodynamic thought 22:762
Self-objectification 6:35–36
Self-observation 21:519
Self-observer agreement 9:238
Self-ordered pointing task (SOPT) 21:639
Self-organization 8:517–518, 20:564, 21:529–531
and personality development 17:880
sociocybernetics 22:914–915
theory of 13:3
Self-organizing dynamical systems 21:529–534
Self-organizing dynamical systems (SODS)
autonomy 21:529
basic features 21:529–530
dynamics 21:530
examples of 21:532–533
future prospects 21:533
historical development 21:530–531
methodological considerations 21:531–532
pattern formulation 21:529
robustness and resilience 21:529–530
simulation models 21:532
theoretical considerations 21:531
Self-organizing networks 6:239–240
Self-organizing social systems 21:531
Self-organizing teams, workplace autonomy 25:720
Self-orientation (pattern variable) 1:382
Self–other distinctions 7:503–504
Self–other mapping, like me theory 15:131, 15:133–134
Self-other similarity, and social identity 22:316–317

Self-ownership 9:400
Self-perceived effects of pornography 18:618
Self-perception approach to persuasion 17:942–943
Self-perception theory 22:528–529
of cognitive dissonance 4:76–77
Self-perceptions 1:329–330, 15:915, 22:496
development in adolescence 21:433–434
and national character 16:233, 16:235
sense of competence 1:329–330
sense of self-worth 1:329–330
Self-policing, and ethics 1:734–736, 1:738
Self-presentation
and social media 22:422, 24:148–149
Self-processes 5:564
Self-prophecy effect 17:942
Self-quantifying diagnostic tools 10:574
Self-realization 1:868, 14:28, 18:237–238, 18:240
Self-recognition 21:421–422
Self-redress homicides 4:600–601
Self-reference 21:530
in social systems 14:386
Self-referenced grading 3:823
Self-referent executive function (S-REF) theory 17:873
Self-referential emotions 12:3–4
Self-reflection 3:821, 17:767–768, 25:592–593, 25:595
Self-reflective empathic orientation of moral reasoning 7:316t
Self-reflective research, feminist social work 8:935, 8:937
Self-reflectiveness 17:736
Self-reflexive, controlled macro-comparison *see* Distant comparison
Self-reflexive social construction 5:478–479
Self-reflexivity, and postcolonial geography 18:645
Self-regarding attitudes 2:318
Self-regarding motives 24:981–982
Self-regulated learning (SRL) 4:800–801, 6:434, 12:202–203, 21:518, 21:535–546
changes in emphasis 21:543
characteristics of 21:535–536
definition 21:541–542
future directions for research and theory 21:544–545
goal orientation theories 21:542–543
information processing theories 21:542
intellectual context for research 21:542
intrinsic motivation theories 21:542–543
methodological issues 21:544
methodologies and issues 21:538–539
self-report methods 21:538–539
microanalysis 21:544
models 21:536–538
Boekaerts' model 21:536
Efklides' model 21:536
Pintrich's model 21:536–537
Winne and Hadwin's model 21:537
Zimmerman's model 21:537–538
operant theories 21:542
out-of-school influences 21:545
role of computer technology 21:544–545
role of self-motivation 21:543
role of teachers 21:545
self-attributions 21:543
self-concept theories 21:542
versus self-regulation of performance processes 21:545
social cognitive theories 21:542
structured diaries 21:544
training, effects of 21:543–544
volition theories 21:542–543, 21:545
Vygotsky on 21:542
Self-regulated Strategy Development (SRSD) model 21:543
Self-regulating economy 12:388, 13:16
Self-regulating market 7:14
Self-regulation 1:35, 10:710, 10:712, 12:381, 13:622, 16:514, 17:872, 19:403, 24:176–177, 25:682–683
and abusive supervision 23:703
during adolescence 21:547–552
developmental regulations 21:547–550
future directions 21:550–551
intentional 21:548–550
organismic 21:548–549
positive youth development 21:549–551
adolescent health 1:122
in adulthood, 21:557–562
active role of individual in shaping one's life 21:558
aging 21:560
gains 21:559, 21:561
goal focus 21:561
goal orientation 21:559, 21:561
goal setting 21:559–560
impact of time perspective on goals 21:560
losses 21:559, 21:561
personal goals as building blocks for life-management 21:558–559
resources 21:557–562
self-control 21:557–558, 21:561–562
setting multiple goals 21:560–561
apprenticeship 1:863
and behavior during infancy and early childhood 6:12, 6:14–17
cognitive development 4:40
cognitive theories of 17:873
consumer policy 4:733
control beliefs 4:820
defined 17:876
definitions 21:547–548, 21:553–554
development during early childhood 21:553–556
behavior 21:553–555
cognitive processes of motivation and executive function 21:554
concomitant developmental factors 21:555
dysregulation 21:553, 21:555
early years 21:555
historical and clinical contexts 21:553
interrelationship between social, physiological, and cognitive processes 21:554–555
learning 21:555
physiological and psychobiological processes 21:554
socialization 21:554
trajectory 21:555
dispositions, and personality 17:829–830
of function systems 13:556
of goal pursuit 15:889–892
health psychology 10:674
institutionalized 21:314–315
international public goods 19:542–543
in learning 4:114
for life management 14:86
and mathematics education 14:818
metacognition 24:89, 24:92
and motivation 15:936–938
of organizations 3:895
of performance processes 21:545
personality trait 17:818
phases of 25:682–683
and procrastination 19:76–78
and self 21:442
self-distortion 17:873
for self-efficacy 21:518–519
skills, instructional psychology 12:202–203
structure of 25:682
and temperament 24:185–186, 24:188, 24:193
temperamental 17:877
Vygotsky on 25:313
weight reduction 25:473
Self-regulation model (SRM) 6:470
Self-regulation skills, gender gaps in 9:683–684
Self-regulatory mechanisms, media ethics 15:44–45
Self-regulatory processes 10:257–258
Self-regulatory skills, mathematical competence 14:818
Self-regulatory systems 18:126, 18:131
personality as 17:767–768
Self-reinforcing learning effects 17:386–387
Self-reinforcing mechanism, organizational path dependence 17:385–389
Self-reinforcing processes 21:530
Self-related judgments 6:76
Self-reliance 14:667–668
versus mutuality 4:148
Self-reliance of poverty measures 18:743
Self-reliant dimension 4:724
Self-report inventories 6:343–344
Self-report measures
emotion assessment through 7:521
extreme response style 23:758–759, 23:762
of health behaviors 10:588–591
behavior record methods 10:590–591
interviews 10:590
questionnaires 10:588–590
of procrastination 19:74
general measures 19:74
specific measures 19:74
Self-report methods 21:538–539
interview protocols 21:538
inventories 21:538–539

Self-report methods (*continued*)
methodological issues about self-reports 21:539
teacher ratings of learner's SRL 21:539
think-aloud protocols 21:538
traces 21:539
Self-report offending surveys 5:165
Self-report questionnaire 5:311
Self-report scales 8:669, 21:522, 25:592
advantages 25:592
disadvantages 25:592
Self-report studies of crime 5:715–716
Self-report surveys 5:125, 5:134–135, 5:190, 5:718, 19:734
Self-reported gang membership 9:595–596
Self-reported health 22:931
Self-reported insight 10:432
Self-reported offenses 17:776–777
Self-reported teaching style (TSS) 24:66
Self-reports 17:849
of attitudes 2:178–180
of cultural intelligence 5:434
and delinquency 6:70
of emotional intelligence 7:445–446
questionnaires 19:773–774
of values 25:36, 25:39
Self-representation 11:533, 19:340–341
Self-representing 13:386
Self-representing self 21:422–423
Self-reproducing social structures 22:980
Self-resemblance, human mate choice 11:338
Self-respect 16:24
Self-restraint
asceticism 2:60–64
and values 25:30
Self-satisfaction 21:519
Self-schemas 17:871–872, 19:401
Self-selection (statistics) 21:415
Self-serving assessments 16:419–420
Self-socialization 22:865
Self-standards model of cognitive dissonance 4:78
Self-stereotyping 21:455
and social identity 22:314–316
Self-stimulation 20:212–213
Self-stimulation behavior 20:212
Self-surveillance 23:737
Self-theory of personality 20:738
Self-transcendence 17:817–818, 25:50–51, 25:591
Self-validating enterprise 17:12
Self-verification (personality trait) 17:817
Self-worth 21:522, 21:526
Seliger, R.V., works, *Handbook of Correctional Psychology* 7:591
Seligman, C.G. 8:290, 10:139, 14:453–454
Sellars, Roy Wood, works, *Critical Realism* 20:21
Selling spirituality 20:408
Selling/marketing, cultural mediators's role 5:441–442
SELS *see* Society for Empirical Legal Studies (SELS)
Selten, Reinhard 9:582
Selves, technologies of 16:218–219
Selznick, Philip 13:516, 13:808, 17:376, 20:783
on evolution of law 13:566
works
Law, Society, and Industrial Justice 13:516, 13:518
TVA and the Grass Roots 17:302–303
SEM *see* Sexually explicit material (SEM); Social Ecological Model (SEM); Structural equation modeling/models (SEM)
Sem software 19:421
Semai people, and peace/nonviolence 17:625
Semantic agraphia 6:735
Semantic algebra 15:751
Semantic analyses, of serial verb constructions 21:646
Semantic approach to control 4:830
Semantic barriers 20:445
Semantic bleaching 10:330, 10:332–333
Semantic categories 25:645, 25:649
Semantic competence 15:755
Semantic (or 'conceptual') criteria 25:649
Semantic dementia 21:573
Semantic dependency 6:182–183
properties 6:183
Semantic differential (SD) 21:563–571, 21:581
bipolar adjectives 21:563–565, 21:567–568
cognitivism 21:568–570
consolidation and internationalization 21:568–570
culture 21:565, 21:568–569
data on political judgments 21:564f
development and early career 21:566
differentiation and pluralization of applications 21:570
EPA (evaluation, potency, activity) structure 21:565–570
equidistance 21:563, 21:563f
Ertel on 21:567
factor analysis 21:563, 21:565–568, 21:570, 21:565f
on fairness of tornado 21:563f
Heise on 21:568, 21:570
history 21:563
Hofstätter, semantic differential 21:567, 21:569–570
Likert, semantic differential 21:563
marketing research 21:563, 21:568, 21:570
neobehaviorism 21:563–566, 21:568–569
Osgood on 21:563–570
other types of scaling 21:563–565
polarity profiles 21:563–564, 21:567, 21:564f
prosperity and critique 21:567–568
psycholinguistics 21:565, 21:569
reliability 21:566, 21:568–569
SSCI article on 21:567f
stereotype 21:566–567
theoretical and institutional context 21:565–566
Thurstone on 21:563
validity 21:566, 21:568–569
variations in German-speaking countries 21:566–567
Semantic Differential Scale 2:179
Semantic domain, space as 23:89–90
Semantic facts and episodic experiences 1:645–646
distinction between 1:647–648
Semantic grammars 19:698
Semantic invariance 14:533
Semantic knowledge 5:926
Semantic loss, impact of 21:573
Semantic memory 1:399–400, 1:875, 5:925–926, 19:3, 21:572–578, 23:876, 24:565–566
and aging 1:399–402
amodal semantics 21:574–577
anterior temporal lobe 21:574, 21:576–577
brain networks 21:576
categories of knowledge 21:575–576
attribute knowledge 21:575–576
current status of knowledge and areas of controversy 21:576–577
definition 21:572
development of knowledge about 21:572–574
distributed representations 21:575
early language models of semantic processing 21:572–573
feature comparison models 21:572–573
hierarchical network model 21:572
spreading activation model 21:572
future directions in research 21:577–578
and hippocampal amnesia 10:876–877, 10:883–884
impact of semantic loss 21:573
modalities of knowledge 21:574–575
multimodal semantics 21:576
neuroanatomical distribution of cortical semantic network 21:576f
object knowledge 21:573–577
proactive interference in 19:3
relationship with episodic memory 21:577
retroactive interference in 20:613–614
schema models 21:577–578
schemas and scripts 21:573
semantic dementia 21:573
semantic-episodic memory distinction 21:572
superordinate category knowledge and subordinate knowledge 21:574
technology provides a window on to the neural basis of 21:573–574
visual representation in 25:177–178
word knowledge 21:572–573, 21:575, 21:577–578
Semantic memory impairment 1:879
Semantic model theory 14:792
Semantic networks 2:36, 21:581
central reasoning mechanism in 13:102
knowledge representation 13:101–102
Semantic normativity
bare normativity of meaning 14:843–846
conventionality of meaning 14:844

dimensions of 14:843–844
normativity of content: assertion 14:844, 14:847–848
normativity of content: belief 14:844, 14:847–848
prescriptive normativity of meaning 14:844, 14:846–847
rule-based normativity of meaning 14:844, 14:846–847
Semantic P600 (sP600) 16:674
Semantic processing of words 13:370–371
angular gyrus involvement 13:371
frontal lobe involvement 13:370–371
frontal–temporal connections 13:371
fusiform gyrus (FG) involvement 13:371
middle and inferior temporal gyri involvement 13:371
parietal and temporal lobes involvement 13:371
Semantic radical 20:2–3
Semantic reading errors 4:93
Semantic reconstruction 8:208–209
Semantic relativism 20:237
Semantic roles 21:588–589
Semantic similarity 21:579
cognitive psychology of 21:579–584
connectionist modeling 21:582–583
general models 21:579–581
generative probabilistic models 21:582
high-dimensional context spaces 21:581–582
model evaluation 21:583
semantic differentials 21:581
semantic networks and featural models 21:581
structured representations 21:580
transformational accounts of similarity 21:580–581
Semantic subset principle 9:209
Semantic system for sentential logic 19:61
Semantic text analysis 4:770–771
data matrix 4:771t
Semantic theories 14:191–192
Semantic types of adverbial clauses 1:186–187
Semantic unity or homogeneity 17:557–558
Semantic variant of primary progressive aphasia (svPPA) 13:363–364
language processes in 13:364
spontaneous speech in patients with 13:364
Semantic Web 12:867, 14:320, 21:293
Semanticists 21:585
Semantic/lexical rules 14:841
Semantic-pragmatic deficit syndrome 18:789
Semantics 2:882, 8:351, 8:353, 13:100, 13:942, 14:229, 15:826–827, 17:554–558, 21:585–591, 23:901, 25:205
and bilingual aphasia 2:574–575, 2:577
of deixis 6:55–56
distinctions between names and words 17:223
future developments 21:590
language acquisition 13:247
lexical 13:930–936
lexicography 13:937–938, 13:940–941
modality 21:588
and morality 16:610–611
of negation 21:585
prototype theory 19:286–289
sentence 21:586–590
situation type and aspect 21:587–588
and syntax 21:586
tense 21:587
Semantics of number 17:43
Semasiographic systems 25:780
Semasiology 13:934–935
Semen Retention Syndrome 5:628
Semen-loss anxiety syndrome 5:625–628
Semester-long learning to learn 13:717–718
SEMI *see* Short Explanatory Model Interview (SEMI)
Semiautomatic ground environment (SAGE) 24:764–765
Semiautonomous social fields 13:758
Semicircular canals 25:63–69, 25:64f
Semi-dependent worker 7:573
Semidwarf wheat 9:307–308
Semigroup, algebra 22:479–481, 22:483
Semi-Markov model 8:298–300
Semiological structuralism, and advertising 1:216
Semiology 2:358–359, 23:626–627
Semiorders 17:526
Semiosis 21:593, 21:593f
Semiotics 2:357–362, 2:853, 12:783–785, 20:651, 20:721, 20:726, 21:592–597, 24:292–293, 25:207–211
abduction 21:595
communication 21:594–595
connotation 21:596
and cultural sociology 5:499–502
culture 21:594–595
deixis *see* Deixis
and film genres 9:161
icon 21:593
ideologies 6:506
index 21:593, 21:595–596
logic and linguistics 21:592–594
classical heritage 21:592
Pericean logic 21:592–593
Saussurean linguistics 21:593–594
and markedness 14:533, 14:536
media 21:594–595
meta-signs and meta-communication 21:595–596, 21:596f
paradigm 21:593–594
pragmatism 21:593, 21:595
structuralism 21:594
symbol 21:592–593
syntagm 21:593–594
theory of science 21:595
visual methods 25:181
Semiparametric efficiency, generalized method of moments estimation 15:296
Semiparametric models 6:124–125, 14:124, 21:598–603, 23:773, 21:602f
example 14:124
nonparametric additive models 21:601–602, 21:602f
nonparametric models 21:599–600
partially linear models 21:601
single-index models 21:600–601
survival analysis 23:779
Semi-parametric regression Markov models 8:315
Semiperiphery 25:341–342, 25:751
world systems theory 25:748–749, 25:751
Semi-presidential systems 17:513
executive government 8:497–500
Semipresidentialism 18:867–868
Semiprofessions 19:107
Semiproportional electoral systems 19:206
limited and cumulative voting 19:206
parallel systems 19:206
Semi-proportional single transferable vote 7:351
Semiring, defined 22:480
Semisocial motives 24:981–982
Semistructured interviews 12:671–672
qualitative health research 10:693
Semitic religious practices and institutions 24:602
Semitic sacrifice 20:861
Semivariograms 23:107–108
Semmel, Bernard 10:493
Semmel thesis 10:493
Semple, Ellen Churchill 7:723
Sen, Amartya 20:114
justice 12:946
neo-Aristotelian freedom as capability 9:401
works
Collective Choice and Social Welfare 7:90
Sénac, Jean 21:753
Senate 4:649
legislative procedure 4:651–652
Senate culture mentalities 4:165–166
Sending country (adoption) 1:136–137
Senegal, peace building and music in 25:846–847
Senegambia 4:575
Senescence
evolution of 8:451
and life history evolution 9:29
Senior adults, psychosocial development research on 7:942
Senior citizens 13:425
Seniority
and inequality 11:897
in teaching 24:77
Seniors, value of life for 25:18–19
Sennett, Richard, works, *The Craftsman* 25:233
SENS *see* Scientifically engineer negligible senescence (SENS)
Sensation 23:111
direct scaling of 21:604–606
and ergonomics 11:299
Stevens on 23:453–454

Sensation magnitude, scales of 21:605–606
Sensation seeking (SS) 21:607–614
  behavioral expressions 21:608–610
  biological influences 21:610–612
  definition 21:607
  evolution 21:612
  gender differences in 9:752
Sensation seeking scales (SSS), development of 21:607–608
  cross-cultural shortened form of ZKPQ (ZKPQ-50-cc) 21:608
  form II 21:607
  form IV 21:607–608
  form V 21:607–608
  impulsive sensation seeking (ImpSS) 21:608
  Zuckerman–Kuhlman–Aluja personality questionnaire 21:608
Sensationalism 12:86
  media ethics 15:45
Sensay 3:147
Sense enumerative lexicons 13:943
Sense generation 11:116
Sense of agency 21:420–421
Sense of body-ownership 21:420
Sense of Coherence (SOC) 10:612
Sense of community 4:324
Sense of fairness 22:895
Sense of identity 17:746–748
  everyday understanding of sameness of persons 17:746–747
  facets of 17:747
  implications 17:747–748
  material content/formal competence 17:745–746
  mechanisms for stabilizing identity 17:746
  modern theories 17:745–746
  postmodern approaches 17:745
  stability of value commitments across a lifetime 17:747
  theoretical-observing *versus* practical stance 17:746
  theories 17:744–746
Sense of place 18:149, 18:151
SENSEable City Lab 23:146
Sensemaking 16:195
  organizational information 12:72–73
  and realism 20:25
Sensemaking theory 17:390
Senses
  anthropology of 21:615
    cognition 21:617
    cultural approach 21:617–619
    embodiment 21:616
    ethnography 21:616–617
    hierarchy 21:615–616
    from measurement to meaning 21:616–617
    psychologies 21:616–617, 21:619
    sensory approach 21:619
  and body 2:728–730
  oral and literate expression 17:277
Sensibility, and national character 16:233
Sensitive period model 14:68–69, 14:69f
  with later modification 14:68–69, 14:69f
Sensitive periods 18:819
Sensitive questions, randomized response techniques for 16:920–921
Sensitivity 6:345–346, 9:928–929, 15:282–283, 20:732
  cultural 22:822–823
  justice 12:952–953
  maternal 2:138, 2:141
Sensitivity analysis 5:53–54, 17:110, 19:459
  Bayesian statistics 2:376
  population forecasts 18:594
  for violations of assumptions 4:452
    for ER 4:452
    for violation of the monotonicity assumption 4:452
Sensitivity function 14:800
Sensitivity training 13:913
Sensitization 3:766–767, 10:480–481
  dishabituation *versus* 10:481–482
  in learning 23:876
  and wanting/needing 25:351–355
Sensor fusion, and multisensory integration 16:99
Sensor neurons 2:28
Sensor systems 20:419–420
Sensorial synesthesia 14:193
Sensorimotor actions 16:506–507
Sensorimotor function 21:759–760
Sensorimotor integration 2:371
Sensorimotor integration and control, dorsal stream for 11:324–326
Sensorimotor maps 11:386
Sensorimotor self 21:420–421
  sense of agency 21:420–421
  sense of body-ownership 21:420
Sensorimotor stage of development 16:501, 16:504
Sensorineural hearing loss 16:714
Sensory adaptation and recovery 21:605
Sensory aphasia 18:47
Sensory consciousness 10:738
Sensory cortex 24:427–428
Sensory deficits 23:335
Sensory deprivation 21:868
Sensory discrimination 21:955
Sensory ethnography 21:616–617
Sensory experiments 21:953
Sensory integration 2:370
Sensory knowledge, organizational symbolism 16:181
Sensory memory 20:50–51
Sensory modalities, hallucinations in 10:500
Sensory motor 18:127, 18:130t
Sensory neuroprostheses 16:714–718
  cortical interfaces 16:717
  peripheral interfaces 16:717
Sensory noise 21:946
Sensory pathways, cortical projection of 3:337f
Sensory perception, consumer psychology 4:741
Sensory perceptual representation 16:198–199
Sensory perfection 21:616
Sensory psychologies 21:616–617, 21:619
Sensory representation of stimuli 19:470–471, 19:471f
Sensory representations 24:427
Sensory stimulation, and cognition 5:577–578
Sensory topographic maps, in brain 24:427
Sensory variability 21:946
Sensory-affective phenomena 11:460–461
  auditory deafness 11:460
  negative visual 11:460
  olfactory hallucinations 11:460
  positive auditory hallucinations 11:460
  positive visual 11:460
  in psychotherapeutic practice 11:460
  taste hallucinations 11:460
Sensory-motor activation
  relation to higher cognitive processes 7:409–411
    body as source 7:410
    contextual characteristics as source 7:410–411
Sensory-specific satiety, in food intake 24:27–28
Sentence bargaining, objections to 18:227–229
Sentence completion tests 19:133
Sentence comprehension 19:354–355
  interpretation 21:625–626
  phenomena 21:623–625
    ambiguity resolution 21:623–624
    context effects 21:624
    immediacy 21:623
    lexical and frequency effects 21:624
    prosody effects 21:624–625
    structure sensitivity 21:623
  psychological models of 21:625
  psychology of 21:621–626
  task of 21:621–622
  ways of studying
    offline techniques 21:622
    online listening techniques 21:622–623
    online reading techniques 21:622
Sentence meaning 21:340, 21:585–586
Sentence processing 13:371–373
  cognitive control and selection 13:372
  figurative language comprehension 13:372–373
  processing of structure and meaning 13:372
Sentence production 19:355–356
Sentence prosody 23:718–720
  prosodic hierarchy 23:718–719
  tone structure 23:719–720
Sentence semantics 21:586–590
Sentences, language evolution 11:324–326
Sentencing 5:186, 18:879–880, 19:212–214
  court communities 21:627–628
  disparities across mode of conviction 21:630–631
  focal concerns 21:627–628
  limitations 21:631–632
  punishment as 19:633–634
  racial/ethnic disparities 21:628–630
  role of forensic psychologist in 9:353
  sex/gender disparities 21:630

theoretical perspectives of 21:627–628
trial penalty 21:631
Sentential logic 19:61
Sentential Subject Condition (SSC) 16:4
Sentience, animal rights in research 1:705
SEP *see* Socioeconomic position (SEP)
Separability thesis 13:485–486
Separate spheres ideology 9:729
Separation 1:71, 1:73, 16:300
Bowlby's studies 19:504
definition 4:740–741, 6:597
demographic measurement 6:118
in immigration 11:676
and integrated attachment theory 12:229
and morality 13:485–486
theme of 2:790–791
Separation (ethnogenesis) 8:177
Separation (life phase) 5:510–511
Separation anxiety 2:791
Separation anxiety disorder (SAD) 23:563
Separation anxiety test 2:139
Separation assault 9:733
Separation call, and neonatal crying 16:475
Separation criteria 10:346
Separation thesis 13:764
law and morality 13:479
legal validity 18:25–26
Separation-individuation process of infancy 11:538
Separationism, Dark Triad 5:673
Separatism 8:946–947
racial 19:809
Separatist movements 13:529
Separatrix 14:797
Sephardic saints (*zaddiqim*) 24:602
SEPs *see* Somatosensory evoked responses (SEPs)
Septal nuclei 14:126
Sequence alignment methods 24:470
Sequence analysis, event *see* Event sequence analysis
Sequence of issue evolution 12:766
Sequence of tenses 24:211
Sequence processing, language evolution 11:324–326
Sequential actions 19:69
Sequential aggregation strategies 14:345
Sequential contribution value 4:870
Sequential decision making 1:536–537, 5:906–910
in binary choice tasks 5:906
for different sample distributions 5:908–909
for known sample distribution 5:907–908
number of observations bounded 5:908
number of observations unbounded 5:908
with multiple information source 5:909
observations 5:906
with one decision criterion 5:906–909
role in applied research 5:906
secretary-type of problems 5:907
with two decision criterion 5:909–910
for unknown sample distribution 5:907
use of term 5:906
Sequential designs 17:260
Sequential equilibria 9:586–587
in game theory 9:583, 9:586–587
Sequential exchangeability, in time-varying treatments 14:342–343
Sequential Memory Index 16:725
Sequential memory processing 15:157–158
Sequential mixed methods designs 15:620
Sequential MultiPhased Adaptive Randomized Trial (SMART) 8:337
Sequential positivity, in time-varying treatments 14:342–343
Sequential priming tasks 11:714–715
Sequential probability ratio test (SPRT) 6:396, 21:637
Sequential processes 16:541–543
Sequential publication 20:530
Sequential ranks, in records 16:898–900
Sequential sampling 5:906
Sequential sampling models 14:804
mathematical psychology 14:804–805
Sequential search problems 5:906
Sequential statistical methods 21:634–638
Bandit problems 21:636–637
Bayesian case 21:636
dcollection 21:634
Frequentist case 21:635
nonsequential design 21:634–637
Sequential stopping 21:634
Sequestered problem solving (SPS) 21:98
Sequestration, death/dying 5:866
Serano, Julia
on allyship 24:553
on antiidentity movement 24:552
Seraphim 16:806
Serbia
confederation 4:572
and World War I 9:212
Serebrovsby, Aleksandr 8:224, 8:228
Serendipity, and ignorance 11:580
Serendipity effects 16:584
Serfdom, in Europe 8:275
Serhane, Abdelhak 21:752
works
L'amour circoncis 21:752
Les enfants des rues 21:752
Messaouda 21:752
Serial monogamy 17:486, 17:489, 17:491
Serial order memory 21:639–644
amnesia 21:639–640
dlPFC 21:639–643
fMRI 21:640
future directions 21:643
hippocampus 21:642–643
in humans
effects of development 21:640–641
effects of focal brain lesions 21:639–640
effects of neurological and psychiatric illness 21:641
neuroimaging 21:640
medial temporal lobes 21:639–640, 21:642
neuropsychology 21:639–640, 21:642
nonhuman animal studies 21:642–643
nonhuman primates 21:642
rodents 21:642–643
prefrontal cortex 21:639–643
rhesus monkey 21:642
temporal order 21:639, 21:641
working memory 21:639, 21:641–642
Serial position curves 24:202–204
Serial processing 12:79–80
Serial reaction time task 13:88, 16:845
Serial subtraction, cardiovascular reactivity 3:118
Serial verb constructions (SVCs) 21:645–647
Creole languages 21:646–647
definition and importance 21:645
distribution in world's languages 21:646
grammatical analyses 21:646
and language change 21:647
related constructions 21:647
semantic analyses 21:646
syntactic analyses 21:646
West African languages 21:645–647
Serial verbs, types of 21:645–646
Serialism 18:57–61
Seriality 16:185
Seriation 19:432–433
approximation methods 19:433
exact methods 19:432–433
gradient-based seriation 19:432
paired comparison ranking 19:432
Seriation (archaeology) 23:744
Seriation problems 19:432
Sering, Max 22:698
Serious games 4:501
Serious mental illness (SMI), religions and 20:349
Seroconversion, in AIDS 1:491
Serotonergic synapse 16:496f
Serotonin (5-HT) 9:1012, 16:749–750, 16:754, 21:611–612
and anxiety 1:804, 1:806–808
and depression 22:83–85
and neonatal temperament 16:495
transporter 9:820–821, 16:495
and wake 22:88–89
Serotonin and norepinephrine reuptake inhibitors (SNRIs) 17:470, 19:467
Serotonin (5-HT) reuptake inhibitors (SRIs)
for obsessive–compulsive disorder 17:115, 17:124–125
Serotonin transporter (5-HTT) gene 8:863–864
Serotonin1A receptor levels 14:441–442
Serson, Joan 7:934
Sertillanges, G. 18:435
Sertraline 1:779, 1:781
for obsessive-compulsive disorder 17:125
Servant leadership 13:657
Servers, audience measurement 2:212–214
Service, Elman 7:622
Service, volunteerism 25:268–272
Service class 2:788
Service climate 17:330
Service delivery
child and adolescent psychiatry 3:386–388
implementation challenges in 11:693–694
Service economy 21:649–650

Service Employees International Union (SEIU) 24:493
Service industry 21:649–650
  regional specialization 21:650
  uneven development 21:650
Service provision 24:278–279
  learning disability 13:686
Service quality regulation 20:170
Service user organizations 10:145
Service Users Research Group England (SURGE) 15:198
Service workers 18:667
Service-conditioned severance benefits, job displacement insurance 12:828
Service-learning (SL) 15:778–779
  approach, for economic education 24:85
  programs 13:657
Servicemen's Readjustment Act 1944 16:306, 18:734
Services 3:43–45, 7:33–35, 7:65
  consumer psychology 4:738–739
  definition 21:648
  financial 3:45, 9:191–193
  knowledge management services 3:45
  legal process outsourcing 3:45
  rural social work 20:808–811
  supply chain management 23:709–711, 23:713
  voice-based 3:45
Services automation 2:297
Services economy, geography of 21:648–653
  conceptual approach 21:648–649
Service-user 3:523
Service-user involvement 15:194–195
  and choice 3:520–524
  and mental health 15:196
SES *see* Socioeconomic status (SES)
Sesame Street 7:249–250
  educational program 15:6
  exposure to 15:36
Seselelame 21:618
Sessions, psychotherapy 19:516–519
SET *see* Student evaluations of teaching (SET); Supportive expressive therapy (SET)
Set, several orders on 17:298
Set estimator 8:16
Setai 24:584
Set-asides 1:260
Set-shifting (cognitive function) 6:725
Set-theoretical models, axiomatic theories 2:338, 2:340–342
Settlement Act 1992 11:799
Settlement archaeology 21:654–658
  of ancient Egypt 1:693
  chronological constraints and considerations 21:656–657
  contemporary research strategies and ongoing debates 21:656–657
  development 21:654–655
  early foundations for archaeological survey in Americas and England 21:654
  environmental constraints 21:656
  historical background 21:654–656
  interpretation of regional data 21:657
  large-scale regional survey programs 21:655
  pattern research at smaller scales of analysis 21:655–656
  population estimation 21:657
  studies, critical role of 21:657–658
  survey methodologies and sampling 21:656
Settlement assets 19:177–179, 19:181–182
Settlement house movement, by Addams 1:91, 1:93–95, 1:218
Settlement houses 16:450
Settlement Movement 22:801–802
Settlement performance, geographical access and 1:65
Settlements 4:223–224, 7:115
  informal 12:57–58
  market equilibrium allocation and 7:115
  and migration 15:417–420
  refugees 22:769–773
  regional planning 20:141
  in trials 24:636, 24:638
  urban 24:836–837
Settler colonialism
  ethnohistory 8:195
  in North America 17:25–26
Settler migrations 15:447
Settler states, indigenous peoples in 11:756
Set-top-boxes (STBs) 2:213
SEU *see* Subjective expected utilities (SEU)
Seuse, Heinrich 2:63
718 Joint Factory 5:598
Seven-point multimodal perspective 16:72
Seventh-day Adventists 15:530–531, 20:269
Severance pay, job displacement insurance 12:819–820, 12:822–823, 12:825–830, 12:828f
Severe acute respiratory syndrome (SARS) 11:318, 12:566
Severe mental illness 24:278–279
Severing the link strategy, family and household demography 8:727
Severity 6:255–256
  of addiction 1:100
  and certainity 7:83–84
  of seizures 13:34
Sewell, William 20:816
Sex 2:256, 3:570, 17:948, 21:365
  as asset 10:816
  construction of gender 9:701
  defined 15:315
  definition 9:742, 13:869
  as demographic variable 1:401
  and dichotic listening 6:359–360, 6:360f
  as distinct from mortality 10:814–815
  and epigenetics 24:687
  and family 8:770
  and feminist legal theory 8:922
  and gender 9:788–794
  and health care 9:694–696
  and inequality 11:897
  and the law 18:534–536, 18:538
  and love 14:375, 14:377–380
  and reproduction 20:450–453
  as risk and danger 10:813–814, 10:816–817
  as a social tool 18:898
  sociolinguistic studies of 9:700–701
  sociology, Randall Collins 4:214–216
Sex aids 10:820
Sex appeal, as erotic capital 8:1, 8:4–5
Sex Civilization Survey 21:734
Sex composition, and family size preferences 8:806
Sex crimes 21:659–664
Sex crimes
  against adults 21:661
  against children 21:659–661
    child sex trafficking 21:661
    incest 21:660
    pedophilia 21:660–661
  emerging issues in 21:662–663
    female sex offenders 21:662
    in institutional settings 21:662–663
    juvenile sex offenders 21:662
  explanations of sex offending 21:661–662
Sex cultural festival 10:821
Sex Description and Sex Status Act 2003 (South Africa) 21:771
Sex determination, nature/nurture 16:342
Sex differences 11:340
  in adult mortality 1:178–179
  clinical psychology training 3:875
  and enriched environment 2:837–838
  in temperament 24:176
  virtual spatial navigation 16:351–352
Sex discrimination 9:731–732
  and harassment 21:721–722
Sex disparities, in sentencing 21:630
Sex education 21:825
Sex hormones 7:608
  and hysteria 11:508
  and sexual motivation 21:762
Sex in the media 21:833
Sex industry 13:882
Sex life
  challenges 21:665–671
Sex life in Finland 21:665–666, 21:669–671
  declining frequency of intercourse 21:667–669
  decreasing frequency of orgasms 21:669
  gender differences in sexual desire 21:666–667
  increasing lack of female sexual desire 21:666–667
  sexual pleasure 21:669
Sex markets
  and affiliation networks 21:702
  and partner selection 21:701–702
Sex offender legislation 20:434
Sex offenders 21:659, 21:672–677
  challenging problems in offender treatment 20:438
  cognitive-behavioral therapy (CBT) framework
    adult relationship, improving 20:435
    anger management 20:436
    antisocial personality/lifestyle impulsivity, control of 20:437
    cognitive restructuring 20:436

empathy, developing 20:435–436
targeting sexual fantasy and deviant sexual arousal 20:436–437
definitions and terminologies 21:672–673
Internet 20:438–439
Kansas versus Hendricks 20:434–435
management in community 20:438
moral panics and research findings 21:673–674
feminism 21:674–676
gender issues 21:674
good lives model 21:675
premediated acts 21:674
prolific offender 21:673–674
psychological perspectives 21:674–675
rehabilitation, public health, and social work 21:676–677
risk need responsivity 21:675
sociogical perspectives 21:675–676
stranger danger 21:673
theoretical and therapeutic approaches 21:674
relapse prevention (RP) 20:435
as 'repetitive' derives 20:434
treatment models for 20:435
dialectical behavior therapy (DBT) 20:437–438
good lives model 20:437
for juvenile sex offenders 20:438
restorative justice model 20:437
risk-needs-responsivity (RNR) model 20:437
Sex offending 21:659
Sex parties 17:490–491
Sex partners, bivariate model for 14:331
Sex ratio at birth (SRB) 15:608–609, 15:612–613, 15:609t
Sex ratios
areas 21:687
and ethnic intermarriage 8:120–121
and fertility transition, in China 9:48
gene-culture coevolution 9:813–814
variation at birth 21:686–687
explanations for 21:688
Sex reassignment surgery, in China 21:738–739
Sex roles 23:437–438
biological factors 21:678–679
Bronfenbrenner's ecological theory and 21:682–684, 21:683t
development and education 21:678–685
evolution and selection of 21:679–680
theories concerning 21:683t
Sex segregation 3:926
female managers 8:885–886
similarities, during preschool age 18:223
Sex selection 15:609–610
implications of 15:611–612
prenatal 10:36
prohibition of 15:613
Sex status 22:542
Sex stereotypes 3:929
Sex surveys 17:486–490, 21:665
in Finland 21:665–666
Sex therapy
desire, excitement, and orgasm 21:689–690
efficacy 21:691
sexual dysfunctions in men
diagnostic procedures 21:690
psychological treatments of 21:690–691
validated treatments for 21:691
sexual dysfunctions in women
diagnostic methods 21:691
psychological treatments of 21:691–692
validated treatments for 21:691–692
Sex tourism 17:487
Sex trafficking 10:193–194, 10:207–208, 10:210, 19:280–283, 21:661, 22:44
in Canada 17:432
protecting children from 3:426–427
in United States 17:432
Sex work 19:279–285, 21:799, 21:837
criminalization 19:279–283
feminism 19:280–282
intimate labor 19:279
legal, spatial, and sociotechnical dimensions of 19:279–281
migration 19:282
and nation state 19:282
neoliberal globalization 19:279
politics 19:281–283
technology 19:280
tourism 19:282
as violence, exploitation, labor, and choice 19:281–282
Sex workers
activism 19:281, 19:283
citizenship of 19:279, 19:282–283
as knowledge and cultural producers 19:282–283
Sexagesimal cycle 24:411–414
Sex-biased admixture 9:889–890
Sex-dependent selection (balancing selection) 17:901
Sex–education interaction 23:385
Sexism 3:928–929, 18:835, 18:841, 21:693–699
benevolence 21:695–698
benevolent 4:154
consequences 21:698
discovery of 21:693–694
and essentialism 8:14
hostility 21:695–698
and language 9:699–700
measurement 21:696–697
ambivalence 21:698
first phase 21:696
modern scales 21:697
old-fashioned attitude scales 21:697
second phase 21:697
and prejudice 18:828–829
role congruity theory 21:698
social status 21:694
stereotype threat 21:698
stereotypes 21:693–698
stigma 23:464
stigmatization 21:693
types 21:694–696
ambivalence 21:695–696
prejudice 21:694–695
SEXOFSPEAKER 19:727
Sexology
Arabic 21:752
sexual orientation 21:785–786, 21:789
Sexophobia 21:741–742
Sex-positive feminism 9:732–733
Sex-Role Egalitarianism Scale (SRES) 21:697
Sex-role socialization 13:256–257
Sex-role specialization, family theory 8:817–819
coping with risk and change 8:818–819
empirical status of independence hypothesis 8:818
mismatches 8:817–818
Sex-role spillover theories 21:723
Sex-role theory 14:681
Sex-selective abortions 20:463, 21:686–688
bachelors 21:687–688
birth order 21:687
factors contributing to sex selection 21:686
marriage squeeze 21:688
prenatal sex testing 21:686–687
sex ratios areas 21:687
son preference 21:686–687
substitution hypothesis and role of technology 21:687
variation in sex ratios at birth 21:686–687
explanations for 21:688
Sex-specific measures of economic marginalization on arrests 5:133
Sex-typed behavior 21:678–679
Sexual abuse 21:659–661
and childhood stress 23:570
of children 3:424, 3:426–427
of elders 7:331
protecting children from 3:426–427
Sexual activity
of children, and divorce 14:519
commodification of 19:279
future research 10:799–800
and nonmarital childbearing 9:14–17
at old age, defining 10:796–797
at old age, determinants *see* Sexual activity determinants
Sexual activity determinants
attitudes 10:798–799
believes 10:798–799
cohorts 10:799
dyadic characteristics 10:797
marital status 10:797
physical health 10:797–798
psychological well-being 10:797–798
Sexual agency 21:840
Sexual aggression 6:633
Sexual antagonism, Melanesia 2:570
Sexual arousal disorders
of men 21:714–715
pathophysiological problems 21:714–715
psychosocial factors 21:715
treatment of 21:691
of women 21:717–718
Sexual assault 21:659
on college campuses 21:662–663
in correctional facilities 21:663

Sexual assault (*continued*)
in military 21:660–661
in United States 5:721–723
Sexual attitudes 17:950–951, 21:665, 21:833–834
of heterosexual sexuality 10:801–811
history 17:950
religion 17:950–951
Sexual attraction 10:802t
and sexual orientation 21:773, 21:775
Sexual aversion disorder (SAD) 21:716
Sexual behaviors 1:128–129, 10:583, 18:617, 21:665, 21:707–709, 21:825–826, 21:829–831, 21:833, 21:835
and age 1:129
biology of 17:952–953
by gender 1:129
history 17:950
and opportunities 1:132
regulation by hypothalamus 11:489
religion 17:950–951
religion and 20:266
same-sex youth 9:650, 9:652–653
and social networks 21:700–704
and socialization 1:130
sociocultural scripts of 17:950
surveys 15:314–317
and telephone 24:160
in urban adolescents 21:706f
Sexual capital 8:1, 8:5
Sexual characteristics 21:786–787
Sexual citizenship, sexual minorities and 21:728–733
application of Western conceptions to diverse societies 21:731–732
critiques 21:730–731
Sexual coercion 11:193–194, 21:722–723
and rape 19:887–891
Sexual competence, as erotic capital 8:2
Sexual conflict 22:910
Sexual continuum, and same-sex youth 9:650
Sexual culture 21:665
in Western Europe 18:533–539
Sexual debut 14:336, 17:487–488, 17:490, 21:841
age of 14:330
concomitant aspects of 21:709
consequences of 21:709
correlates and contexts of 21:706–709
determinants of 21:706–708
cross-national variation and trend 21:705–706
emerging directions in 21:709
as erotic capital 8:5
future research in 21:709
in Latin America and Caribbean 9:69–70, 9:70t
Sexual desire 10:791, 21:665
and arousal 18:537–538
female 21:666
gender differences in 21:666–667
Sexual development 21:825
from late childhood through early adulthood, longitudinal analyses of 14:328–339
age at time of interview 14:330
age of menarche 14:330
age of sexual debut 14:330
alcohol use 14:328
behavioral change 14:328–329
bivariate growth with an exogenous outcome 14:331–333
bivariate model for number of sex partners and drinking 14:331
data preparation 14:336
discrete-time survival analysis 14:336–337
drinking behavior 14:330
empirical examples 14:329
event occurrence/onset 14:329
female sexual debut and menarche 14:336–337
growth modeling: age-based time metric 14:333–334
growth modeling: menstrual age-based time metric 14:334–336
growth modeling: wave-based time metric 14:330–333
intraindividual 14:328–329, 14:339
linear growth models 14:330
male and female sexual debut 14:336
materials and procedure 14:329–330
methodological limitations 14:338–339
number of participant pregnancies 14:330
panel data 14:333
participants 14:329
ratings of days drunk 14:330
substantive implications 14:337–338
summary of findings 14:337–339
total number of sexual partners 14:329–330
Sexual deviance 3:885, 20:334
Sexual deviation 3:888
Sexual difference
and gender 9:788–792
Humboldt on 25:287–288
Sexual dimorphism
body, evolution of 2:725
and hypothalamus 11:482, 11:489
Sexual diversity 13:861, 21:753
Sexual division of labor 13:157–158, 13:161
Sexual double standard 8:895, 8:897
Sexual dysfunctions 10:465, 10:808–809
classification 21:712t
common risk factors for 21:713f
epidemiology 21:711–720
of men 21:713–716
response understanding 21:711–712
of women 21:717–719
Sexual education 18:534
Sexual exploitation 10:207–211
Sexual expression, at older age 10:796–800
Sexual fantasy and deviant sexual arousal 20:436–437
Sexual functioning, and cancer 3:74, 3:76
Sexual harassment 6:562–563, 13:592, 13:820
policies 21:725
in United States 21:721–727
Sexual health 21:825–826, 21:829–831
of median age 21:708f
of young people 21:706f–707f
Sexual identities 9:650, 12:658, 21:678, 10:802t
and gay/lesbian tourism 9:639–640
and lesbian feminism 8:942–947
and queer anthropology 19:754–755
Sexual impairment, and cancer 3:74
Sexual imprinting 11:337–338
Sexual infidelity, and religion 20:319–320
Sexual initiation 21:700
see also Sexual debut
Sexual intention 21:700
Sexual intercourse, declining frequency of 21:665
Sexual invert 9:632
Sexual maturity 9:41
Sexual minorities 13:879–881, 21:769
categories, social construction of 13:880
in China *see* China
discrimination
consequences and responses 13:880–881
cultural, patterns of 13:880
family *see* Families
future directions 13:881
in Maghreb *see* Maghreb
and same-sex youth 9:652–654
and sexual citizenship 21:728–733
application of Western conceptions to diverse societies 21:731–732
critiques 21:730–731
and social work 21:802–807
in South Africa *see* South Africa; sexual minorities in
Sexual misconduct, psychiatric ethics 8:65
Sexual morality 15:815
Sexual motivation 21:759–767
attractiveness 21:760
DHEA 21:761
DSM 5 21:763–766
erectile dysfunction 21:763
gender identity 21:760
genitalia 21:759–760
hypoactive sexual desire disorder 21:762–763
hypothalamus 21:759, 21:761–762
in men and women, measurement scales 21:762–765, 21:764t–765t
prefrontal cortex 21:760–761
proactive sexual desire and solicitation 21:762
receptivity 21:761–762
sensorimotor function 21:759–760
sex hormones 21:762
sexual orientation 21:760–761, 21:763
sexuality 21:759–762
valence and salience 21:761
Sexual network, hypothetical 21:854f
sexual objectification of women 9:730

Sexual Offences Act 21:770
Sexual orientation 3:888–889, 10:801–802, 20:457, 20:460, 20:463, 21:678–679, 21:684, 21:768
antisodomy laws 21:770
ban on same-sex marriage 21:770–771
biological influences 21:773–778
body 21:776
brain 21:775–776
2D:4D finger ratio 21:776
fraternal birth order effect 21:776–777
future directions 21:777
genetics 21:775
handedness 21:776
height 21:776
hormones 21:774–775
hypothalamus 21:775
otoacoustic emissions 21:776
sexual attraction 21:773, 21:775
x chromosome 21:775
concept 21:785
defining 21:773
definition 21:768
and essentialism 8:14
genetic influences on 21:779–784
heterosexuality as historically contingent 21:768–769
historical and social construction 21:785–790
behavior, characteristics, and identities 21:786–787
beyond constructionism 21:789
bisexuality 21:786–787, 21:789
comparative perspectives 21:788–789
desire 21:785–787, 21:789
discourse 21:786, 21:789
gay 21:786–789
heterosexuality and homosexuality 21:787–788
lesbian 21:785–789
power 21:786, 21:788–789
queer 21:788–789
rise of historical approaches 21:785–786
sexology 21:785–786, 21:789
transgender 21:787, 21:789
homosexuality as historically contingent 21:768–769
and intersectional queer theory 19:765–766
and intimate partner violence 12:686
and law 21:768–772
laws 21:769
theoretical approaches to 21:771–772
in military 15:514
recognition of gender identity 21:771
and same-sex youth 9:649–652, 9:654
scholarship in 21:771–772
and sexual motivation 21:760–761, 21:763
Sexual orientation change efforts (SOCE) 3:889
Sexual orientation law 21:771–772
ideology 21:771
Sexual paradise 9:637–638, 21:740–741
Sexual partners 14:329–330
use of term 10:589
Sexual permissiveness 13:567
Sexual perversions 17:948–954
Sexual pleasure 21:669
Sexual politics, in Arab world 21:791–796
Sexual pollution 14:437
Sexual practices 21:835–836
Sexual problems 21:701
Sexual psychopaths 20:434
Sexual relations
with close relatives 11:730
and family 8:811
Sexual relationships between children 3:482
Sexual reproduction 18:897
Sexual response cycle 21:711–712
Sexual revolution 5:75
Sexual rights 20:457–467, 21:792
family planning programs 8:796–797
Sexual risk 1:128
and adolescent sexual behavior 1:128–129
Sexual risk behaviors 21:797–801
adolescent health 1:121–123, 1:125
coital bleeding 21:798
cultural rites of death 21:799
dry sex or vaginal douching 21:799
early sexual debut 21:797
indigenous healing practices 21:799
intergenerational sex 21:798
and media 15:5
multiple and concurrent sexual partnerships 21:798
safer sex practice 21:799–800
sex work 21:799
substance abuse 21:799
unprotected anal sex 21:798
unprotected oral sex 21:798
unprotected vaginal sex 21:797
unwanted pregnancy 21:797, 21:799–800
Sexual satisfaction 10:822, 10:809t
Sexual scripts 10:802–805, 10:807–808, 21:840–842
Sexual selection 5:677–678, 22:892, 22:910
sociobiology 22:892, 22:896
Sexual selection theory 15:812–813
Sexual self-definition 16:861
Sexual sensation seeking scale (SSSS) 21:609
Sexual services 17:486–487, 17:491
as erotic capital 8:5
Sexual side effects (psychopharmacotherapy) 19:467
Sexual socialization, and media 15:5
Sexual Sterilization Act (Canada) 20:471
Sexual Strategies Theory (SST) 21:847
Sexual trends 21:665
Sexual victimization 21:659
Sexual violence 1:128, 10:194–195, 21:794, 25:117
Sexual well-being 10:808–809
Sexualities 2:393, 3:56–58, 9:770–772, 11:2, 11:193–195, 19:330–332, 20:459–460, 20:463–464, 21:791
in Africa 10:812–818
in Arab world 21:791–792
criminalization of homosexual acts 21:795
public indecency 21:795
reproductive rights 21:792
restrictions on girls' physical activity and socialmobility 21:793
in archaeological studies 9:764–767
in Australia 21:833–839
of bear subculture 13:890–891
behavior/identity 16:863
in China 2:69
future of 10:823
studies 10:819
surveys 10:821–822
containing 10:791
cultural aspects 21:819–824
AIDS 21:822–823
extramarital sex 21:821
and gender 21:821–822
inequality 21:819, 21:821–822
LGBTQ 21:819–820, 21:822
marriage 21:820–821
masculinity 21:819, 21:822
meaning 21:819–821, 21:823
non-heterosexuality 21:822
premarital sex 21:820
social reproduction 21:819–823
society 21:819–823
and dance 5:667–668
definition
as object of research 15:314–315
in surveys 15:315
domains of 9:650–651
and feminism 8:899–900, 8:902–903
and feminist legal theory 8:922
Freud's studies 19:503
gastrointestinal disorders 9:608
and gender 9:783–784, 18:536–537
geographies of 19:759–761, 19:763
and harassment 21:722
historical approaches to 21:785–786
and honor 11:181, 11:183
infantile 19:503
and intersectionality 12:658–661
of leatherman subculture 13:888–889
and leisure 13:854
and lesbian feminism 8:942–944, 8:946–947
LGBT friendship and 9:455–456
and liberal Christianity 14:10
and marriage 14:599–600, 14:603
marriage and 21:792–794
nature of 17:949–950
nurture of 17:949–950
over life course 21:840–845
adolescence 21:841–842
childhood 21:840–841
hooking up 21:842–843
in later life 21:844
from young adulthood through midlife 21:841
and pornography 17:219–221
queer, histories of 19:752–753
queer geography 19:759–762
regional understandings, Latin America 13:858–859
in rural geography 20:797–798
and sexual motivation 21:759–762

Sexualities (*continued*)
and social media 24:150
and social psychology 22:527
and social work 21:802–807
challenging binaries 21:802–803
cisgender 21:802–803, 21:805–807
education 21:806
gay 21:802–806
global context 21:803
heteronormativity 21:802–806
heterosexual privilege 21:803
HIV 21:804
homonormativity 21:805–806
homosexuality 21:802, 21:804–806
and human rights 21:803–805
intersex 21:802–803
lesbian 21:802–806
practice 21:805–806
religion 21:804–805
research 21:806
response 21:805–806
socially enforced conformity 21:804
state-enforced conformity 21:804
taxonomy 21:802–803
transgender 21:802–805
transsexual 21:802–804
and space 21:808–812
difference 21:808
feminist 21:808–810
geographies of 21:808–809
heteronormativity 21:809–810
homonormativity 21:809, 21:811
homophobia 21:812
identities 21:808–811
intersectionality 21:810–811
justice 21:809
and other identities 21:810–811
place 21:808–811
queer 21:808–811
queer spaces, circuits and fields 21:809–810
subjectivities 21:812
transphobia 21:811
state, and queer anarchism 19:750
study, methodological issues in 15:314–318
heterosexuality 15:316–317
interviewing 15:314–317
item nonresponse 15:315
LGBT 15:314, 15:317
measurement or response error in surveys 15:315–316
NHSLS 15:315
probability sampling 15:314, 15:316–317
question wordings 15:317
representative sampling and generalizability 15:316
sample-based surveys 15:316–317
sexual behavior 15:314–317
sexually transmitted infections STIs 15:315
social desirability 15:315–316
terminology for 16:861–862
theories of 21:846–852
and visibility 25:208, 25:211–212
youth and 21:792–794
Sexualities, theories of
behaviorism 21:847–848
biopsychosocial approach 21:850
cognitive theories 21:848
Dual Control Model (DCM) of 21:850–851
evolutionary psychology 21:846–847
feminist theory 21:851
importance 21:846
integrated theories 21:850
queer theory 21:851
research 21:846
social learning theory 21:848
social theories 21:848–850
script theory 21:849–850
social constructionism 21:849
social exchange theory 21:848–849
sociocultural research 21:849
sociobiology 21:846–847
symbolic interactionism 21:851
Sexuality and Internet 21:813–818
alternative romantic environment 21:816–817
cyberlove 21:813–814
cybersex 21:813–814
future of 21:817–818
greater satisfaction 21:815
greater self-disclosure 21:814–815
increasing boundaries violations 21:815
increasing opportunities 21:814
lesser commitment 21:815
lesser vulnerability 21:814–815
reducing exclusivity 21:815–816
Sexuality education 21:794, 21:825–832
abstinence-only sex education 21:826
abstinence-plus sex education 21:826
addressing learners 21:826–827
disabled 21:826–827
LGBTQ youth 21:826
comprehensive 21:826
effectiveness 21:830–831
frameworks and standards 21:828–830
abstinence-only sex education framework 21:828–829
abstinence-plus sex education framework 21:829
International Planned Parenthood Federation (IPPF) framework 21:829–830
United Nations Population Fund report 21:829
World Health Organization (WHO) frameworks 21:829
public policies and legal frames 21:827–828
terminology and definitions 21:825–826
in United States 21:828
comprehensive sexuality education frameworks 21:830
Sexuality over the life course 18:533, 18:535–536, 18:538
Sexualized language 24:150
Sexually active marriages 21:844
Sexually aggressive behavioral outcomes 18:617
Sexually dangerous persons 20:434
Sexually explicit material (SEM) 15:5
Sexually transmissible infections 21:836
Sexually transmitted diseases (STDs) 1:128, 2:699, 18:534–535, 18:537–538, 19:779, 21:826, 21:855–856, 21:858
and infertility 12:46
prevention of 21:826, 21:829–831
Sexually transmitted infections (STIs) 9:642, 10:805–808, 15:315, 21:705, 21:709, 21:797–801, 21:853–861
Seymour-Smith, C.
works
The Macmillan Dictionary of Anthropology 7:589–590
Wealth of Nations 7:590
Seyyid, Mohammed Ahmed ibn 15:531
SEZs *see* Special economic zones (SEZs)
SF-12 *see* Short Form 12 (SF-12)
SF-36 *see* Short Form 36 (SF-36)
SfAA *see* Society for Applied Anthropology (SfAA)
SFS *see* Site-frequency spectrum (SFS)
SFT *see* Schema-focused psychotherapy (SFT); Structural family therapy (SFT)
SGAs *see* Second generation antipsychotics (SGAs)
SGLMG *see* Sydney Gay and Lesbian Mardi Gras (SGLMG)
SGOs *see* Scholarship granting organizations (SGOs)
Shack Dwellers International 24:969
Shadow, criminology of 5:247–248
ultramodern shadow
cultural realm and 5:248
linguistic realm and 5:247
material realm and 5:247–248
symbolic realm and 5:247
Shadow economy 12:50–54
definition of 12:50–51
policy consequences of 12:54–55
see also Informal economics
Shadow economy estimations
direct approaches 12:51
discrepancy approaches 12:51–52
indirect approaches 12:51–52
measurement approaches 12:51–52
Shadow of sexual assault thesis 8:855
and fear of crime 8:855
Shadow price, cost–benefit analysis 5:51–53
Shadow side of leadership 13:664
Shadow work 21:862–867
economy and unemployment 21:863–865
economy labor force 21:862–863
developing and transition countries 21:864–865
OECD countries 21:863–864, 21:864t
worldwide aspects 21:863
informal employment in total nonagricultural employment 21:865t

labor market 21:862–863
state regulation 21:862, 21:864
tax pressure 21:862
theoretical considerations 21:862–863
undeclared work 21:862, 21:866
Shafak, Elif 20:873
al-Shahi, Ahmed 8:291, 12:735
Shakai 24:584–585
Shakai-gaku 24:584
Shakers 2:146–147
Shakespeare, William 20:68–69
and personality psychology 17:909
works
As You Like It 17:909
Hamlet 9:365, 20:637
Shaleka Melaku 3:442
Shall issue permits, for guns 10:459
Shallow alphabetic systems 23:265
Sham 24:538
Shamanic practices 20:300
Shamanism 3:350, 16:168–169, 21:868–870
in agricultural communities 16:169
altered states of consciousness 21:868–869
in Arctic 1:929–930
communities 21:868
democratized 16:169
East Asian studies 6:786
Korean 2:71–73
popular and professional 21:868–869
and psychotherapy 19:502
religions 21:868–869
research 21:869
dead ends and detours 21:869
positive directions 21:869
Siberian origin and common features 21:868
spirits 21:868–870
terminological precision 21:869–870
Shame 7:469–470, 11:181–184, 21:475–477
acknowledgment of 21:875
adaptive 21:477
class and gender 11:182–183
comparative considerations 11:182
Cooley on 21:873–874
definitions of 21:871
Elias on European history 21:874–875
and embarrassment 21:477–478
Goffman on 21:873–874
implications for deterring transgression and socially undesirably behavior 21:476–477
implications for interpersonal adjustment 21:476
implications for psychological adjustment 21:476
in intergroup relations 7:504–505
Lewis on psychotherapy sessions 21:875
looking glass self 21:871–872
pride and alienation 21:872–873
secrete, and silence/violence 21:875–876
and self 21:441
Shandong RuYi 2:243
Shanghai, creative industries in 5:597–598
Shangshan xiaxiang 24:973–974
Shanin, Teodor 1:456–457
works, *Introduction to Shanin* 17:669
Shannon, Claude E. 12:106, 14:243, 23:261–262
Shannon information, Bayesian experimental designs 8:522
Shannon's sampling theorem 9:377–378
Fourier transform 9:377–378
Shanty town 4:526
Shape space 25:154
Shapes of time 2:21
Shapin, Steven 8:577
works, *Leviathan and the Air-Pump* 8:577–579
Shaping
in ABA 2:409
reinforcement 20:208, 20:210
Shapiro, Martin, works, *Courts* 4:250
Shapiro v. Thompson (1969) 18:735
Shapley value 4:870, 9:580
Sharabi, Hisham, works, *Neopatriarchy: A Theory of Distorted Change in Arab Society* 16:385
SHARE *see* Survey of Health, Ageing and Retirement in Europe (SHARE)
Share ownership 4:940–941
Sharecropping, land tenure 1:478
Shared, collective leadership 13:667
Shared accountability 21:113
Shared activities, of young children 9:451
Shared assessment, single *see* Single shared assessments (SSAs)
Shared attention mechanisms, and collective memory 4:190
Shared beliefs
cohesion 21:877
collective phenomena 21:877
concept of 21:877–878
consensus 21:877–879
mutual beliefs 21:878–879
and level problem 21:879–880
as shared we-belief 21:880
Shared decision-making, in evidence-based practice 8:333–334, 8:336
Shared environment 9:826
Shared Environmental Information System (SEIS) 5:777
Shared environmental variance 12:298–299
Shared experience, neonatal imitation 16:487
Shared family 2:473
Shared goals, of children, and intentionality 12:348
Shared governance approach 23:519–520
Shared identity 4:154
Shared knowledge groups 10:436, 10:438–439
Shared leadership, in organizations 10:414
Shared mental models 10:436
groups 10:436
Shared model of personality 17:863–864
five-factor model (FFM) 2:560
Shared norms 10:157
Shared power approach, organizational development 6:297
Shared practice theories, of public reasons 19:593–594
Shared reality
and collective memory 4:190
groups 10:436
Shared representations of nation-state 16:309–314
limitations and future directions 16:312–313
nationalism in settled times 16:310–312
interpretive studies 16:310–311
survey-based studies 16:311–312
traditional nationalism scholarship 16:309–310
Shared resources 21:113
Shared responsibility for key decision-making 21:113
Shared situation awareness 11:301–302
Shared social identification 11:527
Shared traumatic reality (STR) 18:710–711
in social work 18:710–711
Shared values 4:311, 4:314, 4:316
of trade unions and social work practice 24:491–492
Shared vision, and logic models 14:303–304
Shared water systems 25:444
Shared/nonshared environmental effects 6:321
Shareholders 4:933, 4:939, 12:210–211
corporate finance 4:927–931
versus creditors 4:933
versus employees 4:933–934
minority 4:933
protection of 4:942–943
suits 4:935
see also Stockholders
Shareholder-value view 7:70
Share's theory 4:93
Shari'a law 3:708, 12:750, 12:752, 13:801, 21:731, 21:792, 21:795, 25:798–799
Sharing 11:808
in children 19:231
and marriage 8:829
prosocial behavior 19:233–235
Sharing (prosocial behavior) 19:238–239, 19:241–242
Sharing circles, and indigenous social work 11:808–809
Sharing economy 8:487
Sharing of resources
in Arctic 1:928–929
by hunters-gatherers 11:429–430
Sharing organizations 6:239
Sharon, Ariel 18:500
Sharp waves, in electroencephalography 7:357–360, 7:360f
Sharpe, F.R. 14:372
Sharpe and Lotka population model 18:584–585
Shatalov, Aleksandr 21:743
Shaw, Clifford 12:304

Shaw, Clifford (*continued*)
on social class and crime 5:122–124
works
Delinquency Areas 7:561
The Jack-Roller 7:561
The Natural History of a Delinquent Career and Brothers in Crime 7:561
Shaw, George Bernard 8:220, 8:241
Shaw, Martin 15:492
Shawnee prophet 15:532
Sheep, and food production 9:300
Shehuixue 24:584–585
Shelby County v. Holder (2013) 6:533
Sheldon, Charles, works, *In His Steps: What Would Jesus Do?* 14:256
Shelley, Percy Bysshe 15:290
Shelters for homeless people, in United States 11:161
Shenjing shuairuo ('Weakness of the Nervous System') 5:626
Shenzhen Special Economic Zone 24:896
Shepard diagram 16:42, 16:42f
Shepardize 5:805–806
functions 5:808
LexisAdvance 5:809
Shepherd of Hermes 3:548
Sheps, Mendel 10:776–777
Sherif, M., works, *The Psychology of Social Norms* 10:435–436
Sherman Antitrust Act 1890 1:740, 1:796–798, 3:31, 13:820, 23:318
Sherpa, Nepalese 17:429
Sherrington, Sir Charles Scott 21:881–883
attraction to poetry of 21:881
concepts in current world 21:883
life of 21:881–882
on mind 21:882
on motor control 21:882–883
on nervous system 21:881–883
on neurohistology 21:881
on neurophysiology 21:882
on neuroscience 21:882–883
on pathology 21:881
as philosopher 21:882–883
as physiologist 21:882
on relation between mind and energy 21:882
on spinal reflex behavior 21:882
works
The Brain and its Mechanism 21:882
The Integrative Action of the Nervous System 21:881
Man on his Nature 21:882–883
Sherzer, Joel 6:505
Shichigosan festival 20:367–368
Shift work 3:570
Shifting balance theory, Wright's contribution 25:759
Shifting cultivation 11:202–203
Shifting time–space comparison *see* Fluid comparison
Shi'i Islam 12:747, 21:793
Shi'ism 12:733, 20:331, 20:404
Shiji (*The Records of the Grand Historian*) 11:48
Shi-jia 11:49
Shils, Edward 21:169, 21:884–888, 25:463, 25:465
on atomic bomb 21:884
on authoritarian personality 21:885
on center and periphery 21:886–887
on charisma 21:885–887
consensus 21:886–887
early writings 21:885–888
on India 21:884
on intellectuals 13:72, 21:884–888
and Knight 21:884–885, 21:887
and Mannheim 21:884–885, 21:887
on military sociology 21:886
New Nations 21:884, 21:886
and Parsons 21:885–886
on primary groups 21:884–886
on tradition 21:885–887
and Wehrmacht 21:885
Shilubana case 13:760
Shining ideal, in individualism 4:208
Shinnyoen 16:806–807
Shinobu, Orikuchi 12:794
Shinran 2:889
Shinreiky&omacr 16:805
Shinto 20:367–368
conflict between Buddhism and 20:367–368
festivals and ritual calendar 20:368
kami concept 20:367–368
miyamairi rite 20:367–368
in modern Japan 20:368
and national identity 20:368
shrines 20:368
state, association with 20:368
Shintoism 20:408
in Japan 12:796
Shipbuilders 14:527–530
Shipping
British 14:526–528
tramping 14:529–531
Shire, J., on women's participation in sports 9:658
SHIs *see* Social health insurances (SHIs)
Shiv Sena (Shiva's Army) 18:410–411
Shiva, Vandana 21:180
Shizhen, Li 21:184–185
Shklar, Judith N. 12:135, 12:138, 18:498, 18:502–503, 21:889–895
on emigration 21:893–894
formative years of 21:889
Hegel and 21:890–891
on injustice 21:893–894
on legacy 21:894–895
on liberalism of fear 21:891, 21:894–895
Montesquieu and 21:890–892
on optimal marginality 21:892
on political theory 21:890–892
reflections on obligation and exile 21:893–894
Rousseau and 21:890–891
on totalitarianism 21:889–890
transatlantic encounters of 21:892–893
works
After Utopia 21:889–890
American Citizenship: The Quest for Inclusion 21:893
The Faces of Injustice 12:135, 21:893
Freedom and Independence 21:890–891
Legalism 21:890
Men and Citizens 21:890
Ordinary Vices 21:891
Shmakov, Gennadii 21:742
Shneidman, Edwin 23:682
Shneyer, Aaron 25:845
Shoah (Holocaust) 12:803–806
Shock generator (Milgram) 17:62
Shock therapy 12:281
and economic reform 6:789–790
Shock uncertainty 24:338–339
Shockley Semiconductor (company) 11:830
Shoes, role in sexual fantasy 17:952
Shogoro, Tsuboi 2:71
Shôkô, Asahara 20:370
Shopfloor culture 25:817
Shopping 10:152
Shopping behavior 7:700, 7:700t
Short Explanatory Model Interview (SEMI) 10:567–568
Short Form 12 (SF-12) 15:188
Short Form 36 (SF-36) 4:464, 10:658, 15:188
Short neural precursors (SNPs) 2:821–822
Short sleep 22:59
Short tandem repeats (STRs) 9:956–959, 10:25, 19:828
Shortest derivation (SD) 9:877
Shortest-path betweenness centrality 16:535
Shortest-path closeness centrality 16:534
Short-run momentum, in stock markets 23:482–483
Short-term memory (STM) 1:594, 1:644–645, 13:679, 15:157, 15:162, 19:3, 20:50–51, 21:896, 21:901–909, 23:874–879, 25:175, 23:878f–879f
Atkinson and Shiffrin model 21:901–902
attention 21:904–906
in auditory 25:706–707
behavioral work 25:703
of brain injured patients 21:910–911, 21:911f
central level and peripheral level in working memory 21:902–903
central level in WMG 21:904–906
documented impairments 25:706
domain-general maintenance mechanisms 21:904–906
domain-specific maintenance mechanisms 21:903–904
functional architecture of 21:909–910
lexical–semantic representations 21:910
phonological short-term memory 21:909–910
information maintenance 25:705
information processing 21:902–905
maintaining information in working memory 21:906
mechanisms of change in 21:898–899
mental images generated from 15:221

neurofunctional architecture of 21:914
neurofunctional models of 21:915–916
olfactory 21:910
origin 21:909
peripheral level in WMS 21:903–904
proactive interference in 19:3–5
psychological and neural aspects 21:909–916
spatial 21:910, 25:705–706
deficits 21:913
tactile 21:910
temporal distinctiveness 24:204
tests 25:705
and transient global amnesia 24:566
uses 21:913
visual 21:910
deficits 21:913
visual–verbal material 25:706
working memory 21:901–902
Short-term plasticity 12:713
Short-term psychodynamic therapy (STPD) 19:344–348
cognitive behavior therapy *see* Cognitive behavioral therapy (CBT)
for depression 19:375–376
efficacy of 19:346–347
future of 19:347
historical background 19:345–346
individualized psychotherapies 19:347
monitoring patient outcomes 19:347
process research 19:347
specific training in diagnostic-specific treatment 19:347
Short-term storage space 16:504
Shoshonius 18:908
SHOT *see* Society for History of Technology (SHOT)
Shovel test pits 9:335
Show business
entertainment 7:669, 7:672–673
SHP *see* Swiss Household Panel (SHP)
Shreeves, R. 3:318
Shrines 4:235, 20:368
Shrinkage
Bayes' theorem 2:380
variable selection 15:651
Shrinkage estimate 16:82, 16:86
Shrinking 21:917
in amputees 18:2
Shrinking cities 21:917–922
banks 21:919
decline 21:918–919
demolitions 21:919–921
depopulation 21:917–919
detroit 21:920–921
employment loss 21:920
industrial collapse 21:918, 21:921
land 21:919–921
Leipzig 21:921
origins 21:917
policy response 21:919–920
redevelopment 21:919, 21:921
shrinkage 21:918–920
suburbanization 21:917–919, 21:921
vacant properties 21:919–920
Shrinking cities syndrome 24:866
Shriving 3:146
Shrovetide 3:146
SHS *see* Subjective happiness scale (SHS)
SHSS *see* Stanford Hypnotic Susceptibility Scale (SHSS)
SHT *see* Single high-threshold (SHT) model
Shu 11:49
Shubin, Neil
works
The Universe Within: The Deep History of the Human Body 25:344
Your Inner Fish: A Journey into the 3.5 Billion Year History of the Human Body 25:344
Shugendo 12:796
Shuras 23:60
Shutter speed 25:194–195
Shuttle Interferometeric Synthetic Aperture Radar (IFSAR) mission 20:423
Shutz, Alfred 4:200
Shyness 2:456, 22:495, 24:176–177
and self 21:441
SI *see* Symbolic interactionism (SI)
Si scale 15:567
SIA *see* Societal impact assessment (SIA)
Siam 13:814–815
Siberia, Arctic 1:927–930
Siberia, shamanism 21:868
Sibling deidentification 22:158
Sibling designs, for genetic studies of behavior 9:935–936
Sibling histories 1:156–157
Sibling order 21:923
behavior during radical revolutions 21:925–926
competition 21:924–925
family niches 21:924–925, 21:927
family sentiments 21:926–927
inheritance practices 21:923, 21:926
IQ 21:924
parental investment 21:924, 21:926
personality 21:924–925
primogeniture 21:923–924, 21:926
risk-taking 21:924
social, economic, and political consequences 21:923–924
social behavior 21:923
Sibling relationships 3:482, 13:55
organ donation between siblings 3:482
Sibling system, and socialization of children 22:869
Siblings 22:158
behavioral problems in 2:475–476
and gender development 9:745
interaction effect on infant development 18:856
as role models and teachers 8:735
Sibling–toddler play 18:222
SIBs *see* Social Impact Bonds (SIBs)
Sibship size, influences in adulthood 6:754
SIC *see* Standard Industrial Classification (SIC)
Sicherman and Galor's theory of career mobility 15:598
Sicherungsabtretung 21:404
Sicherungsuebertragung 21:404
Sicilian mafia 14:422, 14:425
Sicily (mafia) 14:430–433
Sick humor 11:417
Sick role 6:472–473, 6:475, 11:509
empowerment and recovery *versus* 22:976–977
Sickle model 14:56
Sick-soul 20:354
SID *see* Supplier-induced demand (SID)
SIDCER *see* Strategic Initiative for Developing Capacity in Ethical Review (SIDCER)
SIDE *see* Situational deindividuation (SIDE) theory; Social identity model of deindividuation effects (SIDE)
Side effects, of psychopharmacotherapy 19:462
adolescents 19:467–468
assessment 19:463
children 19:467–468
classification 19:463–467
autonomic 19:464
cardiac 19:465
cardiovascular 19:464–465
dependence 19:464
endocrine 19:464–465
insomnia 19:465
metabolic 19:464–465
elderly 19:468
genetic factors 19:463
historical perspective 19:462–463
management of 19:463
neurological/behavioral 19:465–467
cognitive effects 19:465–466
delirium 19:467
hypomania/mania 19:467
motor effects 19:466
neuroleptic malignant syndrome 19:467
sedation 19:466
seizure 19:467
noncompliance 19:463
placebo effects 19:467
and quality of life 19:468
sexual 19:467
vulnerability 19:463
Sidereal dating methods 3:564–565
dendrochronology 3:564–565
Sideways glance 21:619
Sidgwick, Henry 4:681
works, *The Methods of Ethics* 8:75
Sidney, Algernon 3:772
SIDS *see* Sudden infant death syndrome (SIDS)
Siècle des Lumières 15:683
Siemens and Halske Company 2:921
Siemens Brothers 3:21
SIENA (statistical analysis package) 22:471
Sierra Club 7:763–764
Sierra Leone, child soldiers in 3:430–435
SIESCA *see* Sistema de Internalización de la Educación Superior Centroamericana (SIESCA)
Sieyès, Abbé 22:887

Sieyès, Emmanuel 22:827
SIFFM *see* Structured Interview for the Five-Factor Model of Personality (SIFFM)
SIG *see* Special Interest Group (SIG)
Siger of Brabant 20:285
SIGEX *see* Seasonal adjustment by signal extraction (SIGEX)
Sigma Alpha Epsilon (SAE) 10:558–559
Sign languages 3:814, 21:928–933
  acquisition 13:307, 21:943
  acquisition of 21:929
  comprehension 21:935
  disorders of 21:934–941
    aphasia and limb apraxia 21:934–936
    apraxia 21:936–937
    autism spectrum disorder 21:938–939
    developmental disorders 21:938–940
    induced paraphasias 21:936
    Parkinson's disease 21:937–938
    sign paraphasias 21:935–936
    specific language impairment 21:934, 21:939–940
    stuttering 21:939
  grammar 21:930–932
  and language lateralization 12:694
  left hemisphere lesions 21:930–932
  linguistic principles of American sign language (ASL) 21:942–943
    deaf culture 21:942
    language 21:942
    sign language acquisition 21:943
    structure of ASL 21:942–943
  modalities 21:942
  neural organization of 21:944–945
    aphasia 21:944
    functional imaging 21:945
    hemispheric specialization 21:944
    right hemisphere in ASL, role of 21:945
  neural systems subserving visuospatial languages 21:930–932
  nonlanguage spatial processing 21:932
  production 21:935
  psycholinguistic aspects of 21:943–944
    attention 21:944
    memory for signs 21:943–944
    perception 21:944
    sign recognition 21:943
  psychological and neural aspects 21:942–945
  right hemisphere lesions 21:932
  -specific deficits 21:939–940
  structure of 21:928–929
  users, lateralization of 13:404–405
Sign paraphasias 21:935–936
Sign recognition 21:943
Signal decomposition, computational auditory scene analysis 2:223–224
Signal detecting tasks 5:906
Signal detection models 6:523–524
Signal detection theory (SDT) 19:473, 21:946–951
  analysis of multidimensional stimulus 21:955
  background 21:952
  books and conferences 21:955
  detection paradigm 21:946
  detractors conceptualized under 21:955
  discrimination paradigm 21:946
  history of 21:952–956
  importance for psychology 21:955
  mathematical psychology 14:803–804, 14:813
  multidimensional detection theory 21:950
  multiple stimuli varying along one dimension 21:950
  multiple-interval tasks 21:949–950
  receiver operating characteristic (ROC) curve 21:947–948, 21:947t
  of recognition memory 20:37–38, 20:38f
  response bias 21:947
  sensitivity 21:947
  single-interval task 21:946–947, 21:946t
  standard 21:947
  Thurstone's model 21:953
  use of term 21:946
  Vision Research at Michigan 21:952–953
Signal extraction 24:356
Signal processing algorithms 23:260
Signal processing models, computational neuroscience 4:481–482
Signal transduction, and memory/learning 23:874–875
Signal-driven grouping, computational auditory scene analysis 2:224–226
Signaling 1:581, 12:104
  cooperation 4:862, 4:865–866
  promotions 19:138
Signaling pathways and gradients, in asymmetry of body/brain 2:121–128, 2:122f
Signal-plus-noise trials 21:946
Signals 21:946
Signal-to-noise ratio 16:119
Signans 12:783
Signatum 12:783–784
Signed graphs, structural balance theory 23:605–606
Signed rank sum test 16:888
Significance, role in flashbulb memory 9:257
Significance, tests of 21:957–962
  achievable *p*-values 21:960
  alternative hypothesis 21:959–961
  approximate inference 21:959
  combining 21:960
  difficulties 21:961–962
    multiple testing 21:961–962
    sample size 21:961
  fixed level testing 21:960
  metaanalysis 21:960
  mid p-value 21:960
  model-based inference 21:958–960
    based on likelihood 21:959
    hypothesis testing 21:960
    models and null hypothesis 21:958–959
    significance functions and posterior probabilities 21:959–960
  relation of likelihood to 14:118
  two-sided testing 21:960–961
Significant anomalies 13:134–135
Significant others, incorporation of opinions of 21:494–495
Signified, Lacan on 13:213
Signifier, Lacan on 13:213–214
Signs 10:111
  and symbolism 23:860–861, 23:864
    Peirce's theories 18:809
Sigwart, Christoph 19:391
Sikh, and communalism 4:263–264
Sikh separatism in the Punjab 20:376
Sikhism 20:408, 20:410–411
  in South Asia 2:77
Sikolohiyang Pilipino 11:788, 11:792
Silbey, S., works, *The Common Place of Law* 13:469
Sildenafil, in sex therapy 21:691–692
Silence
  and oral history 17:286, 17:288
Silence, using (communication skill) 4:289
Silence hypothesis 9:658
Silencing of homosexuality 16:863
Silent identities 11:524
Silent language, culture as 12:357
Silent Revolution 5:790–791
Silent Spring and pesticide industry 15:558
Silent trade of goods 16:75
Silicon Valley 25:59
  information and communication technology industry 11:828, 11:830, 11:832
Silicon Valley Toxics Coalition (SVTC) 7:370
Silk Road 21:184–185
Sills, David, works, *International Encyclopedia of Social Sciences* 2:349
Sills, Davis 13:637
Siloviki 20:847
Silver, Jacob 1:847
Silver Economy 17:182
Silver workers 17:183
Silverman, D., on interviews 12:674
Silverman, Max 25:211
Silverman, Sydel, on brokers 2:871
Silvio Berlusconi 18:610–611
SIM *see* Synthetic informant model (SIM)
SimCon 25:640
Simiand, François 12:389
Simians
  primates, evolution of 18:905, 18:908, 18:907f
Similarities 1:535–536
  changing face of 22:211–212
  and interpersonal attraction 12:627
  multidimensional scaling 16:41–42
  and social influence 22:350
  and structural equivalence 23:624
Similarity coefficients 3:807
Similarity judgments 24:679, 24:682
Similarity/contiguity 12:783–784
Simkhovitch, Mary Kingsbury 1:218–219
Simmel, Georg 2:269, 4:341, 4:362, 5:36–37, 5:39, 5:498, 6:690, 8:138, 10:276, 11:865, 15:684, 18:625, 20:359, 21:128, 21:374–377, 21:963–969, 22:828, 24:744, 25:193

on alienation 1:551, 1:553
analysis of fashion 8:836
biography of 21:963
concept of alter ego 21:966–967
concept of form 21:966
and contemporary sociology 21:967–968
and cosmopolitanism 5:43–45
life of 21:964
on modernity 15:684–685
on object of sociology 21:965–966
on reification 20:203–205
and social sciences 21:963–965
on social structure 22:722
sociologies, proposal for 21:967
works
Goethe 21:965
The Fundamental Questions of Sociology (*Grundfragen der Soziologie*) 21:965
Introduction to the Moral Sciences (*Einleitung in die Moralwissenschaft*) 21:964
Kant und Goethe 21:965
Lebensanschauung – View of Life 21:965
On the Essence of Matter – Das Wesen der Materie nach Kant's Physischer Monadologie 21:963
The Philosophical Culture (*Philosophische Kultur*) 21:965
The Philosophy of Fashion (*Philosophie der Mode*) 21:965
The Philosophy of Money (*Philosophie des Geldes*) 1:553, 21:965
The Principal Problems of Philosophy (*Hauptprobleme der Philosophie*) 21:965
The Problem of Sociology (*Das Problem der Sociologie*) 21:964
The Problems of the Philosophy of History (*Probleme der Geschichtsphilosophie*) 21:965
Psychological and Ethnological Studies on the Origins of Music 21:963
The Psychology of Money 21:963
Religion (*Die Religion*) 21:965
Rembrandt 21:963–964
Sociology (*Sociologie*) 21:964
Über Sociale Differenzierung 21:963–964
Simon, Herbert
critique on decision and choice 5:881
Simon, Herbert Alexander 12:304, 15:718, 17:783, 18:290, 21:970–974, 23:522–523
Bach and 21:971–973
Barnard and 21:971, 21:973–974
on bounded rationality 21:970–974
in Carnegie Mellon University 21:971, 21:973–974
on computers 4:516–518
Graduate School of Industrial Administration (GSIA) 21:971–973
on heuristics 21:972–974
and Knight 21:970–971
Nobel Prize 21:970
on organizational theory 1:103–104
RAND 21:971–973
and Ridley 21:971
on uncertainty 21:970–971, 21:973–974
in University of Chicago 21:971
works
Cognitive Limits on Rationality 17:390
Simon, Richard 20:323
Simon, Theodore 12:307
Simons, Henry, works, *A Positive Program for Laissez Faire* 23:457
SIMPLE (Scale-Invariant Memory, Perception, and Learning) 24:200–201
Simple A–B design 21:986
Simple adoption 1:136
Simple and capacitated plant location problem 23:174t
Simple arithmetic mean 18:589
Simple domination 2:752
Simple linear regression 14:154–155, 14:850–852
diagnostics 14:154–155
fitting 14:154
Simple noncompliance 8:528–529
bounds for ACE using instrumental variable 8:529
Simple random sampling 23:617
Simple regression, optimal design for 17:256–257
Simple relationships, categorization of 23:116–117
Simple rules, for decision making 5:893–894
Simple selection model 17:850–851
Simple societies 13:36–37, 16:234–235
Simplicity 1:536–537
Simplicity Postulate 12:800
Simplification, bureaucratic 13:233
SIMPOC *see* Statistical Information and Monitoring Program on Child Labor (SIMPOC)
Simpson, A., on settler colonialism in North America 17:25
Simpson–Angus extrapyramidal system scale 19:463
Simpson's paradox 14:150–152
Sim-pua marriage 11:726
Simulacra, and postmodernism 18:688
Simulated annealing 23:395
Simulation approaches, for cultural evolution 5:406–407
Simulation environment, elements in 4:487–488
agents 4:487–488
knowledge 4:488
tasks 4:488
Simulation geographic models 14:788
Simulation of societies, genetic programming 2:44
Simulation theory 18:19
Simulations 19:64, 22:13–14, 24:360
agent-based models 1:367–372
Bayesian inference 2:375–376
Computer Assisted Instruction (CAI) 4:494
in educational planning 7:263–264
embodied 5:579
and embodied cognition 7:422
for green jobs 10:372–373, 10:377
and postmodernism 18:688
virtual spatial navigation 16:348–350
Simulators, autonomous agents 2:311
Simultaneity
in econometrics 7:9–10
population censuses 3:291
Simultaneity problem 6:261–262
Simultaneous (spatial) autoregressive model 23:105–106
Simultaneous Equation Estimates 23:408
Simultaneous equations 21:975–979
asymptotic distribution
single equation 21:976–978
whole system 21:976
estimation 21:980–983
cross-section *versus* time series studies 21:982–983
decline of simultaneous equations models 21:983
estimating a complete system 21:982
estimating one structural equation at a time 21:981–982
identifiability of structural parameters 21:981
least squares estimators 21:981
methods 21:981
need for estimation of parameters 21:981
problem of choosing and testing a model 21:983
structural and reduced form equations 21:980
supply-demand example 21:980
types of variables 21:980
higher-order efficiency 21:978
regression equations 21:975, 21:978–979
single equation estimators 21:976–977
exact distribution of 21:977
system, identification of 21:975
whole system estimation 21:976
Simultaneous inference, order statistics 17:294
Simultaneous tilt illusion (STI) 25:204, 25:204f
Simultaneous treatment design 21:986–987
Sin, and mental illness 19:307–308
Sinclair, K.O., on discriminatory bullying of LGBTQ youth 21:87–88
Sinclair, Raven, works, *Wicihitowin: Aboriginal Approaches to Social Work* 8:154
Sine qua non 20:240
SINEs strepsirrhines
primates, evolution of 18:907–908
Singapore 23:63
arts education in 1:231
cultural policies in 5:453–456
creative industries 5:456–457
fertility transition in 9:84–89
judicial power and independence in 13:430–431
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Singapore International Arbitration Center 12:470

Singer, André 8:292
Singer–Prebisch thesis 6:309
Singh Sabha movement 20:375
Singing training
for asthma 20:579
for chronic obstructive pulmonary disease 20:579
Single adults 8:770
Single assessment process (SAP) 12:501, 12:503, 12:505–506
Single case studies 4:91
Single contingency 23:610
Single diagnosis 6:386
Single factor reductionism 5:233–234
Single factories *see* Export processing factories (EPFs)
Single fixation duration 8:632
Single high-threshold (SHT) model 6:524
Single imputation 15:604
Single interest groups 12:372
Single Language corpus 4:950
Single mediator methods 15:68
Single mothers, and income inequality 11:742
Single nucleotide polymorphisms (SNPs) 2:452, 7:842, 8:718–719, 9:815, 9:826, 9:899–900, 9:902, 9:959, 10:25–27, 11:291, 11:318, 12:300, 12:305, 15:210, 15:212, 17:889
behavior 9:933, 9:936–937
Single parenthood 21:45–46
Single parents 8:770
Single predictor variable
nonparametric regression with 16:867–873
kernel smoothing 16:868–871, 16:869f–871f
local polynomial regression 16:870–872
with splines 16:872–873
Single process (prolonging), task network models 16:541
Single shared assessments (SSAs)
in social work 12:501–506
community care 12:502–505
context 12:501–503
multidisciplinary 12:503–505
process 12:503–505
reflections 12:505–506
requirement of service users 12:503
single assessment process 12:501, 12:503
Single transferable vote (STV) 7:349, 7:351, 19:205
Single women, fertility of 9:13
Single-case designs 8:346, 8:348–349, 21:984–989
alternating treatment design 21:986–987
analytic aims 21:985
analytic aims of 21:985
application of 21:987–988
accountability and managed care 21:988
analysis of outcome data 21:987–988
validity, improving 21:988
in applied behavior analysis 2:407–408
applied research 21:985
changing criterion design 21:986
experimental research 21:985–986
idiographic 21:984–985
measurement issues of 21:985
multiple-baseline design 21:985, 21:987
reversal design 21:986, 21:988
simultaneous treatment design 21:986–987
subject matter of 21:984–985
varieties of 21:985–987
between-series designs 21:986–987
combining within- and between-series designs 21:987
within-series designs 21:986
withdrawal design 21:986
Single-case experimental design (SCED) 2:815–816
Single-cell recordings in nonhuman primates 4:674
Single-cell studies in monkeys 15:968–969
Single-decrement life table 14:92
Single-equation estimators 21:976–977
asymptotic expansions of distribution of 21:977–978
exact distribution of 21:977
Single-index models 21:600–601
Single-interval task 21:946–947, 21:946t
Single-issue terrorists 16:247
Single-loop learning 7:736
Single-member districts (SMDs) 19:205, 19:208
Single-mother families 11:742
child care in 3:391
Single-nontransferable vote (SNTV) 19:206
Single-parent families
academic achievement in 22:926
emotional development of children in 7:440
and labor participation 8:758–759, 8:763
Single-photon emission computed tomography (SPECT) 11:603, 11:615, 14:439–441
Single-process models, of recognition memory 20:38–39
Single-sex education 9:686
single-sex interaction 13:257
Single-sex schooling 3:926–932
on academic- and career-related outcomes 3:927–929
on affective and social development 3:929–930
merits of 3:930t
prevalence of 3:926–927
research findings analysis 3:930–931
research on effects 3:927
Single-sex schools 9:686
Single-stage stochastic optimization 23:473–477
data generation: constructive *versus* black box 23:474
data treatment: batch *versus* online 23:474
global metamodel optimization 23:476–477
noise generation: exogenous *versus* endogenous 23:474
response surfaces 23:476
sample average approximation 23:474
theoretical properties of estimates 23:474–475
stochastic approximation 23:474–475
gradient approximation 23:475
step size selection 23:475
theoretical properties 23:475–476
Single-stimulus format, of personality assessment 17:840, 17:842, 17:847–848
Single-subject designs *see* Single-case designs
Single-system processing theory 11:710
Singular actors 15:312
Singular political system 6:369
Singular value decomposition (SVD) 4:914, 5:2, 16:119
Singular–plural distinction 14:211
Sinhala 7:197
Sinhala nationalism 2:891
Sink habitat 3:156
Sinn Fein 18:610–611
Sino–Japanese War of 1894–1895 16:276–277
Sinologists 6:799
Sinology 20:322
Sino-Soviet relationship 20:137
Sino-Soviet Security Treaty (1950) 1:561
Sino-Tibetan Burman 8:208
Siow, Aloysius 2:100, 2:104–106
SIP *see* Social information processing (SIP) theory; Stochastic inverse projections (SIP)
SIPP *see* Survey of Income and Program Participation (SIPP)
SIPRI *see* Stockholm International Peace Research Institute (SIPRI)
SIR *see* Sliced Inverse Regression (SIR)
Sira 11:61
Siraj, A., on social representations 16:862
SIR/DBMS (software) 23:378
SISCF *see* Surveys of Inmates in State and Federal Correctional Facilities (SISCF)
Sistema de Internalización de la Educación Superior Centroamericana (SIESCA) 22:624
Sister language, definition 4:424
Sisterhood 18:215
Sisupalgarh eastern rampart 23:55f
SIT *see* Social identity theory (SIT); Stress inoculation training (SIT)
SITC *see* Standard International Trade Classification (SITC)
Site configuration/design problem 23:174t
Site of pathology, debating 4:152
Site selection 20:15
Site survey techniques, in geoarchaeology 10:53
SITEAR *see* Society of Intercultural Education, Training and Research (SITEAR)
Site-frequency spectrum (SFS)-based methods 11:291
Sites, archaeology 23:743
Sites (geoarchaeology)
discovery and definition 10:53–54

aerial photography and satellite imagery 10:53–54
coring archaeological sites 10:54
other techniques of remote sensing 10:54
remote sensing/archaeological prospecting 10:53
formation process 10:54–55
human disturbance process 10:54–55
nonhuman disturbance process 10:54
Sites of learning, in teacher education 24:52–53
Site's situation 10:62
Site-specific dance 5:666
Siting, in multisited field studies 16:104, 16:106–107
Sitte, Camillo 24:863
Sitting Bull 1:636–637
Sitting position, infant motor development 15:974
Sittlichkeit 4:338
Situated cognition 7:409, 7:411, 25:316
and mathematics education 14:819
Situated identities 6:499
Situated knowledges 8:901, 22:1–2, 23:324–326, 23:328
and object 22:1–2
feminist roots of dilemma 22:2–3
subjects, objects, and agency 22:3
uptake and critique 22:3
Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective 22:1
Situated learning 1:870–871, 12:203, 22:5–12
and apprenticeship 1:861, 1:864
articulation and reflection 1:871
cooperation 1:871
critiques of 22:9–10
as enculturation 1:871
environmental education 22:7
implications of learning inside/outside classroom 22:11
in situ 22:5
learning as apprenticeship 22:6
learning outside classroom 22:10–11
multiple perspectives 1:871
place-based learning 22:7
problem-based learning 22:7–8
as situated 22:5–6
by solving authentic problems 1:871
virtual CoP 22:8–9
distance education 22:8–9
game-based learning 22:9
Situated learning theory, on coconstructivism 3:914
Situatedness 4:88, 22:8
in food geography 10:79–81
Situating science 21:237–238, 21:266
Situation awareness
definition 11:299–300
and ergonomics 11:299–302
shared 11:301–302
Situation calculus 13:103
Situation selection, based on organizational culture 17:699
Situation type and aspect 21:587–588
Situational action theory (SAT) 4:460
Situational analysis 10:404
and brokers 2:870
Situational approaches to prosocial behavior 1:581
Situational awareness 11:392
Situational clues in social actions 1:78
Situational code-switching 22:951
Situational control 4:820
Situational couple violence 12:686
Situational crime prevention (SCP) 18:877–878, 20:777
Situational deindividuation (SIDE) theory 22:422
Situational demands 17:834–835
Situational effects 8:572–573
Situational factors, and prosocial responding 7:315–317, 7:319
Situational influences on interpersonal attraction 12:628–629
Situational interest 1:38, 12:378–379, 12:379t
in vocational psychology 25:240
Situational judgment tests (SJTs) 22:13–19
applicants views 22:17
definition 22:13–14
design considerations 22:14–15
development 22:14
evidence-based evaluation of 22:15–17
diversity 22:16–17
measure 22:15–16
popularity 22:16–17
prediction 22:16
reliability of scores 22:15
history 22:13–14
in psychiatric education 7:188
Situational meanings 6:591
Situational mechanisms 22:151–152
Situational model, of readers knowledge 12:40
Situational perspective, of procrastination 19:77
Situationally accessible memory 16:198–199
Situation–decision–utility (SDU) triplets 6:712
Situationist psychology, virtue ethics 25:147, 25:151
Situationists, and reification 5:245
Situations, developmental research 6:327–329
Six Sigma management methodology 10:225
Six-Day War 12:872
686 Program 6:44–45
60s scoop 11:764
Sixtus V, Pope 15:578
Sixty Minutes (magazine) 24:171
64-share system (marine vessel ownership) 14:527
Size dependence, organizational ecology 6:890–891
Size effect, in stock markets 23:484
Size of poverty 18:728–729
Size principle 1:556–557
SJTs *see* Situational judgment tests (SJTs)
Skadden Fellowship 19:557
Skalicka, Vladimir 14:209
Skaneateles Community 2:149
SKE *see* Stimulus known exactly (SKE) model
Skeggs, B., on LGBTI commercial sex 13:884
Skeletal remains 9:334
Skeletal trauma 4:612
war and conflict 4:612
Skeptical postmodernism, and culture 5:584
Skepticism 5:485–486, 6:301–302, 6:593, 7:89, 7:883, 10:220, 12:386, 20:227, 21:920, 22:172, 22:690–691, 22:986, 22:988
about character 25:151
about implicit learning 13:85
Skepticism–dogmatism adaptation 7:883–884
Sketch maps 4:80
SKF 82958 20:213
Skhodki 14:432
Skill biased technological change (SBTC) 11:741
and income inequality 11:741
Skill building
exercises
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
and transition to adulthood 24:580–581
Skill development models 21:462
academic self-concept and achievement 1:56–57
Skill differentials 25:336–337
Skill dimension of digital divide 16:789–790
Skill evaluation, in lithics 14:263
Skill learning, neural substrates of 16:848
Skill levels, and promotion 19:138
Skill mismatch 15:595–597, 15:596t
Skill utilization 17:135
Skill-based learning 1:865, 1:868–869
Skilled migrants 18:598, 18:601
Skilled migration 15:393
South and Southeast Asians 12:531–534, 12:536
Skilled reading 8:395
Skilled witnesses 8:583
Skills
artisans 2:46
child protection 3:419–420
communication with children 3:419–420
eyes wide open approach 3:419
working relationships with parents 3:419
of developing writers 25:769
dynamic skill theory 16:505–508
guilds 2:49
and migration from Europe 15:457–461
scientific practice 21:225–228
versus status seeking 22:332
teaching, behavior therapy 2:428
theory of mind 15:542–543

Skills mismatch 22:23, 23:158
Skills obsolescence 15:597
Skills training 22:20–24
adult education 22:20
benefits for individual and society 22:23
causes of underinvestment in 22:21
Continuing Vocational Training Survey (CVTS) 22:22
cost–benefit analysis 22:20
data on participation rates and trends within and between countries 22:21–22
economics 22:20–24
employment 22:21
estimates of amount invested in 22:22–23
life-long learning 22:23
markets for education and training 22:20–21
in psychosocial interventions for psychiatric disorders 22:382
public policies 22:22
public subsidization 22:20
skill availability 22:23
vocational skills training 22:20–23
Skin 24:36
Skin cancer 23:696–700
basal cell carcinoma 23:696–697
community-based programs 23:698
future directions 23:699
incidence and causation 23:696
individual-oriented interventions 23:698–699
melanoma and NMSC incidence 23:696
population-based programs 23:698
predictors of sun exposure 23:697
predictors of sun protection behavior 23:697–698
prevention 23:698
risk factors 23:696–697
squamous cell carcinoma 23:696–697
tanning 23:696–699
ultraviolet radiation 23:696–699
Skin conductance response (SCR) 1:442
Skin knowledge 21:617
Skin pigmentation
gene-culture coevolution 9:817
Skinner, Burrhus Frederic 12:11–12, 13:691, 13:693–694, 17:609, 19:414, 22:25–30
on behavioral therapy 22:28–29
biography of 22:25
on contingencies of reinforcement 22:25–26
contribution to evolutionary analogy 22:26
on creative behavior 22:26–27
education 22:27–28
on ethology 22:26
on introspection 12:703
on language and consciousness 22:27
on mentalism 22:27
operant conditioning 7:238–240
on operant conditioning 22:25–26
other contributors to psychology of his time 22:29
principles of behavioral analysis 8:483
on radical behaviorism 22:26–27
on selectionism 22:26
Skinner box 22:25–26
on social philosophy 22:29
on teaching machines 22:27–29
on utopia 22:25, 22:28–29
on verbal behavior 22:25, 22:27
works
Verbal Behavior 3:527, 9:871
Skinner, Quentin 3:62, 3:646, 12:256
civic republicanism 9:402
works
The Foundations of Modern Political Thought 3:62–63, 18:484–485
Liberty before Liberalism 3:64
Meaning and Understanding in the History of Ideas 3:62
Skinnerian behaviorism 1:574
Skin-picking disorder 17:113
Skinship 21:619
Skipped-generation households 10:336
Sklar, K.K., on interviewing 12:671
Skocpol, Theda 14:491, 22:31–36
on American civic life 22:34–35
on American democracy 22:34–35
analysis of social revolutions 3:205
on Chinese revolutions 3:517–519
goals of 22:35–36
social revolutions, causes of 22:32
United States welfare state 22:33–34
works
Boomerang 22:32
Bringing the State Back In 22:32
Civic Engagement in American Democracy 22:34
Diminished Democracy: From Membership to Management in American Civic Life 22:31
Health Care Reform and American Politics 22:33–34
Inequality and American Democracy: What We Know and What We Need to Learn 22:35
Obama and America's Political Future 22:33–34
The Politics of Social Policy in the United States 22:33
Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United Sates 22:31
Social Revolutions in the Modern World 22:32
States and Social Revolutions: A Comparative Analysis of France, Russia and China 11:27, 22:31
The Tea Party and the Remaking of Republican Conservatism 22:35
Vision and Method in Historical Sociology 22:32
Skog, Ole-Jørgen, on alcohol consumption 16:557–558
Skolemization 5:936
Skolnick, Jerome 13:516–517
Skrinkage 16:51
Sky Kingdom sect 16:806
Skyscrapers 9:712
Skywalks 24:119
SL *see* Service-learning (SL) approach
SLA *see* Second-language acquisition (SLA)
Slack-maximizing bureaucrat 13:572
Slacktivism 22:423
SLANG 21:356
Slave labor system 3:304
Slave societies 4:339, 22:38–41, 22:51–55
genuine, approaches to study 22:50–51
Slave trade 22:37–38, 22:40–41
in Africa 1:286
moral and political significance for Africa 1:287–288
Slaveholding societies, comparative approaches to 22:49–50
Slavery Convention of the League of Nations (1926) 22:43
Slaves/slavery 14:591–592, 19:801–803, 20:258–259, 22:49–58
abolition of 22:442
abolition of in Great Britain 22:442–443
in Africa 22:54–57
agency 22:54, 22:56
Aristotle on 22:55
Atlantic world 22:55–57
black, in South America 23:45
in Biblical societies and times 22:54–55
capitalist plantation mode 22:52–53
in Caribbean 1:638–639
Civil Rights Movement 3:691–692
comparative approaches to slaveholding societies 22:49–50
debate 7:57–58
defining 22:43–44
dependence on 22:51
description 22:54
embedded demesne mode 22:52
estimating 22:44–45
Finley's contribution 22:54–55
history of 22:54–58
human rights 22:54
for industrial revolution 22:56–57
and inequality 11:897–898
institutionalized slaveholding 22:49
lineage mode of articulation 22:52
Maroons 14:591–595
Ottoman Empire 22:54–55
policy measures to mitigate spread of 22:46–48
predatory circulation mode 22:52
profitability of 7:56–57
and racism 19:852–854
risk factors 22:45–46
Roman or urban–latifundic mode 22:52
Shakespeare's view 22:54
slave societies 22:51–53
genuine, approaches to study of 22:50–51
as social institution 22:43
as socially dead 22:54
societal organization 22:53
societies 22:54–55
urban–industrial modes of articulation 22:52

and violence 22:37–42
emancipation 22:40–41
exploitation, and control from antiquity to modern period 22:38–39
historical process of enslavement 22:37–38
manumission 22:40
middle passage 22:38
peaceful *versus* violent abolitionism 22:40
peaceful *versus* violent endings of modern slavery 22:41
resistance, rebellion, and revolution 22:39–40
resurgence of violent slavery in twentieth century 22:41–42
second slavery 22:39
slave society 22:38–41
slave trade 22:37–38, 22:40–41
societies with slaves 22:38
Slavic languages 11:824
SLC6A4 gene 22:494
SLE *see* School learning environment (SLE)
Sleep 22:59–62
alternating phases 22:63–64
and cortisol 22:80, 22:82–85
cycles 22:63–64
definitions 22:87
non-REM sleep 22:87
rapid eye movement sleep 22:87
demographic factors 22:59–60
age 22:59–60
partnership 22:60
dreaming *see* Dreaming
duration 22:59–61
environmental factors 22:61
and growth hormone 22:80
health conditions 22:60
obesity 22:60
physical activity 22:60
history 22:87–88
homeostasis 22:88
insomnia 22:60–61
lexical consolidation during 13:922–923
and memory 22:63–69
mental disorders 22:60–61
suicidal tendency 22:61
neural systems 22:87–93
acetylcholine 22:89
arousal 22:88–90
basal forebrain 22:89
brain stem 22:88
circadian 22:90–91
GABA 22:89–92
hypothalamus – histamine 22:89
hypothalamus – hypocretin (orexin) 22:89
median preoptic area 22:89
norepinephrine 22:89
REM sleep 22:89–90
serotonin 22:88–89
sleep 22:89
state regulation 22:90–92
transitions between states 22:91–92
ventrolateral preoptic area 22:89
wake 22:88–89
neurobiology of 22:83–85
quality of life 22:61
slow-wave sleep 22:80
stages of 22:80f
stress 22:60
variants 22:59
short sleep 22:59
Sleep benefit, and procedural knowledge, consolidation of 16:846–847
Sleep disorders 22:79–86
and alcohol 1:512
anxiety and PTSD 22:83
ascending arousal system 22:81f
VLPO projections to 22:81f
biomarker 22:82–83, 22:85–86
bruxism (tooth grinding) 22:73
depression 22:82–83
difficulties with arousals 22:71–72
difficulty falling asleep 22:70–71
dyssomnia 22:79
EEG and HPA system 22:85
EEG findings 22:82t
endophenotype 22:82–83
excessive daytime sleepiness 22:73
frequency and risk factors 22:79–80
future perspectives 22:85–86
hypersomnia 22:79
insomnia 22:73, 22:75–76, 22:79–80, 22:82–83
markers of antidepressant treatment outcome 22:85
model of normal and impaired peptidergic sleep regulation 22:84f
neuroendocrinology 22:82–86
psychiatric disorders 22:82
recurrence 22:80, 22:85
regulation of normal sleep 22:80–82
REM density 22:82–83, 22:85–86
sleep-disordered breathing (SDB) 22:73–74, 22:76
sleep-related breathing disorder 22:73
sleep-related movement disorders 22:72–73
vulnerability 22:82–83
Sleep for memory hypothesis 22:63
Sleep splindles 6:653
Sleep-dependent memory consolidation 22:64, 22:68–69
Sleep-disordered breathing (SDB) 22:73–74, 22:76
Sleeper effect 15:32–33
Sleep-phase advance protocol 22:85
Sleep-related breathing disorder 22:73
Sleep-related movement disorders
hypersomnia 22:74–75
narcolepsy 22:74–75
periodic limb movement disorder 22:74, 22:76
restless legs syndrome 22:74, 22:76
Sleep–wake cycle 3:570, 3:573
SLEI *see* Science Laboratory Environment Inventory (SLEI)
SLEQ *see* School-Level Environment Questionnaire (SLEQ)
SLEs *see* Stressful life events (SLEs)
SLI *see* Specific language impairment (SLI)
Sliced Average Variance Estimation (SAVE) 14:159–160
Sliced Inverse Regression (SIR) 14:159–160
Slicher, B.H., works, *The Agrarian History of Western Europe A.D. 500–1850* 17:671
SLID *see* Survey of Labour and Income Dynamics (SLID)
Slim city *see* Eco-cities
Slip-of-the-tongue 18:47, 18:49–50
Slippery slope argument 7:601
Sloan, Alfred 24:143
Sloan Foundation 2:446
Slocum, Karla 1:640
Slovenian Public Opinion (SPO) Survey 16:92
Slovenská národná strana (Slovak National Party, SNS) 18:411, 18:610–611
Slow Food movement 22:94–99, 24:377
authenticity 22:96–98
early years 22:94–95
and environmentalism 22:94–96
fair trade 22:94, 22:96
gastronomy to eco-gastronomy 22:96
against global standardization 22:96–97
globalization 22:97–98
growth and change 22:95–96
market creation 22:96–97
organic movement 22:94
sustainability 22:95–96
Slow Food philosophy 22:97
Slow wave activity (SWA) 22:63–64
Slow-fast dynamics 18:569
Slow-wave sleep (SWS) 19:271–272, 22:80
and memory 22:63–68
SLS *see* Seattle Longitudinal Study (SLS)
SLSA *see* Socio-Legal Studies Association (SLSA)
SLST *see* Social Learning Strategies Tournament (SLST)
SLT *see* Social learning theories (SLT)
SLU *see* Spoken language understanding (SLU)
Slum clearance 16:440
public housing 16:450–451
urban renewal 16:451–452
Slum Dwellers International 18:184
Slum-free movements 20:674
Slums 1:218, 4:525–527, 24:775–778, 24:819–821
and urban poverty 24:915–916
Slyvester, J.J. 24:763
SMA *see* Supplementary motor area (SMA)
Smacof package 16:46
Smacof software 19:421
Small, Mario, on narrative networks 16:215–216
Small, Medium, and Micro-Enterprises (SMMEs) 24:968
Small areas 6:162–169
administrative records 6:163, 6:165
administrative units and user-defined areas 6:167–168

Small areas (*continued*)
census 6:162–165
cohort component models 6:166–167
data and sources 6:163
estimates, projections and forecasts 6:162
extrapolation techniques 6:164, 6:166
geographic information system 6:162, 6:164–168
housing unit method 6:164, 6:166–167
population estimates
methods 6:164–165
projection methods 6:165–167
types of 6:163–164
probabilistic projection methods 6:167
smallness 6:162–163
surveys 6:163, 6:165
Small Business Association (SBA) 1:264
Small Business Investment Corporations (SBICs) 25:60
Small effects, optimal experimental design 17:260
Small farmers 1:461
Small group research 22:100–107
Bales's contribution in 22:101–103
collaborative circles 22:105–106
contemporary attempts to revive 22:105–106
decline 22:104–105
cultural explanations 22:105
sociological explanations 22:104–105
theoretical explanations 22:104
Harvard tradition 22:102–103
Hawthorne experiments 22:101
Illinois tradition 22:103–104
interaction process analysis 22:102–103, 22:103f
leadership 22:103–106
Lewin's contribution in 22:101–102, 22:104–105
MIT/Michigan tradition 22:102
social groups 22:102
social psychology 22:100, 22:102–105
sociometry and other early research on 22:101, 22:101f
theorizing primary group 22:100–101
Small multiples 6:716
Small states 22:108
in Britain 22:109–111
crisis of eighteenth century 22:111
decline 22:111
economy 22:108–112
empire 22:108–112
in France 22:109–111
in Greece 22:110–111
historical overview 22:108–109
Hume on 22:109
ideas in history 22:108, 22:110, 22:112
large states 22:108–112
in modern world 22:111–112
Montesquieu on 22:110–111
nationalism 22:111–112
in North America 22:110–111
political thought 22:109
politics 22:108–112
in Rome 22:108–111
survival by alliance 22:109–110
survival by union 22:110–111
survival by virtue 22:110
union of 22:110–111
war 22:108–112
Small world structures 5:467–468
Small worlds, and social media 22:424–425
Small-area estimation 13:388–389
models for 20:888
Small-commodity production 2:47
Smallest space analysis (SSA) 24:386
Smallness 6:162
Smallpox 15:869, 18:175–176
Small-RNA systems, epigenetic inheritance 7:834
Small-scale cooperation 11:260–262
direct reciprocity 11:261–262
kin selection 11:260–261
Small-scale industry, artisans 2:46
Small-scale mathematical models, population dynamics 18:570
Small-scale societies 11:261, 11:263–264
Small-world networks 16:527, 16:530, 16:530f
Small-world phenomenon 22:424, 22:471
SMART *see* Sequential MultiPhased Adaptive Randomized Trial (SMART)
Smart, J.J.C. 4:681
Smart cities 6:831–832, 22:113–117, 23:730
in an open world 22:113–114
definitions and conceptualization of 22:114–115
human capital 22:114–115
information and communication technologies (ICTs) 22:113–117
intelligent city 22:114
policies on 22:116–117
ranking procedures for 22:115–116
resilient city 22:114–115
sustainability 22:113–116
urban development 22:115–116
urban governance 22:113–114
urban policies 22:114
urbanization trends 22:113
see also Eco-cities
Smart grid digital technology 6:834–835
Smart growth 10:449–450, 16:809–813, 20:18, 24:792, 24:935, 24:938
Smart Growth and Neighborhood Conservation Act 10:450–451
Smart phones 6:898, 6:900
social media and 12:602
SMART Recovery (non-profit organization) 1:503
Smart regulation 18:291
Smart technology 6:834–836
Smartphone technology, and prospective memory 19:277
Smartphone/table applications (app) 16:406
SMAs *see* Standard metropolitan areas (SMAs)
SMC *see* Somatomotor cortex (SMC); Supplementary motor complex (SMC)
SMDs *see* Single-member districts (SMDs)
Smell
and appetite 24:26–27, 24:30
olfactory receptors 24:26
psychology of 24:26–31
cognitive and selective attentional effects 24:29–30
flavor representation 24:29
imaging studies in humans 24:29
olfactory–taste convergence to represent flavor and influence of satiety 24:29
texture of food 24:29
reward and emotions 24:26–27
Smelser, Neil J.
approach of collective behavior 4:160
works
Handbook of Sociology 7:585, 7:588–589
International Encyclopedia of the Social and Behavioral Sciences 2:349
SMEP *see* Society of Multivariate Experimental Psychology (SMEP)
SMG *see* Supramarginal gyrus (SMG)
SMI *see* Serious mental illness (SMI)
Smiling, gender differences in 9:755
SMIs *see* Stress management interventions (SMIs)
Smith, Adam 1:540, 3:647, 3:709, 7:99, 11:413, 13:161, 13:802, 14:11, 14:459, 14:644, 16:228, 19:123–124, 19:233–234, 20:483, 20:660, 20:662, 22:196, 22:697, 24:811, 25:670, 25:675
applied social research of 1:850, 1:854
and birth of nations 22:121
on capitalism 22:119, 22:122–123
on consumption 22:118–122
on division of labor 6:602
as founder of economics 22:118–119
ideological twists of 22:119
on industrial revolution 22:119–120
on industrialization 11:870–871
on international economics 22:119
on invisible hand 11:818–819, 15:332, 21:530
on labor 5:537, 25:676
on laissez-faire approach 22:119–122
on liberalism 22:122
model of division of labor 11:870–871
on modernity 22:118–119, 22:121, 22:123
new conception of society 22:120
on new relation between morals and politics 22:120–121
on political economy 13:161, 22:118–119, 22:121
on profits 19:111–112
on relation between nation-state and market 22:121–122
scientific point of view of 22:122
on self, other, and social 22:122–123
on social change 22:123
on sociology 22:119, 22:123
works
Theory of Moral Sentiments 1:579, 2:443

The Wealth of Nations 2:47–48, 3:90, 3:96, 6:603, 7:87, 7:590, 7:665, 20:259–260, 20:660, 25:7
Smith, Anthony 9:673, 16:286–287
Smith, Barbara 8:908
Smith, Bernard, works, *The European Vision and the South Pacific* 7:753
Smith, C.A., on butch–femme relationship 13:872
Smith, Charles W., on value 5:538–539
Smith, Christian
on social network actors 5:558–559
on social network analysis 5:559
Smith, Dorothy 19:677–678, 23:329
Smith, G. 3:318–319
Smith, John Maynard 22:901
Smith, Linda Tuhiwai 11:771
Smith, Llevellyn, works, *New Survey of London Life and Labour* 7:559
Smith, Michael G. 1:640
Smith, Neil 10:48
Smith, Philip, on national identity 16:312
Smith, Raymond T. 1:640
Smith, Robertson 13:71
Smith, Rogers
on nationalism 16:310
works
Civic Ideals 3:655–656
Smith, T.
on narrative networks 16:212, 16:214–216
on social network analysis 5:559, 5:561
works
Guide for the Modern Bear 13:891
Smith, Vernon 19:538
Smith, William Robertson 1:766, 24:602
on social groups 13:70
works
Lectures on the Religion of the Semites 24:602
Smith v. City of Salem 24:551
Smith–Lever Act 20:815
Smith's Friends 16:806
Smithsonian Institution 18:320–321
Smithsonian repatriation 1:887
SMMEs *see* Small, Medium, and Micro-Enterprises (SMMEs)
SMO *see* Social movement organizations (SMO)
Smoke-free laws 22:128–129
Smokestack chasing 3:663, 18:666
Smoking 10:582–583, 20:270
addiction 22:126–127
and adult mortality 1:179–180
behavior, and obesity 17:76–77
behavioral expressions 21:609–610
and breast cancer 14:71–72
causes of 22:126–127
cessation 22:127
for asthma 20:579
for chronic obstructive pulmonary disease 20:579
and coronary heart disease 4:920
and coronary thrombosis 14:69–71
and disease 22:125–126
and gender 9:704–707
prevention 4:463–464
prevention programs 10:670–671
public place legislation 22:128–129
advertising and promotion 22:129
interaction of policies 22:129
smoke-free workplace 22:128–129
taxation 22:128
tobacco control policies 22:128–129
Smoot Hawley Act (US) 12:568, 12:570
Smooth functional integration 15:761
Smooth muscle relaxants 21:691–692
Smooth nomadic space (geography) 16:567–568
Smooth pursuit, oculomotor control 8:630
Smooth testing
modern developments in 23:400–402
components diagnosis 23:400–401
data-driven testing 23:401–402
generalized smooth tests and Cholesky components 23:401
model selection 23:402
Neyman smooth tests 23:398–400
Smoothing
disturbance 24:359
exponential 24:355
kernel *see* Kernel smoothing
state 24:359
SMP *see* Society for Mathematical Psychology (SMP)
SMPY *see* Study of Mathematically Precocious Youth (SMPY)
SMRs *see* Standardized mortality ratios (SMRs)
SMSAs *see* Standard metropolitan statistical areas (SMSAs)
Smuggling 1:610
Smuts, J.C., works, *Holism and Evolution* 7:751
SNA *see* Social network analysis (SNA); System of national accounts (SNA)
Snail mail 12:603
SNAP-25 *see* Synaptosomal-associated protein 25 (SNAP-25)
Snapshot-matching' mechanism of place recognition 25:136
SNC *see* Substantia nigra (SNC)
SNCC *see* Student Nonviolent Coordinating Committee (SNCC)
Snider, James G., works, *Semantic Differential Technique: A Sourcebook* 9:361
SNM *see* Structural nested models (SNM)
Snob effect 4:696
Snow, C.P. 24:674
works, *The Two ultures* 25:343
Snow, Charles P., on disunity of sciences 7:875
Snow is white (bioconditional statement) 20:68
Snowflake analogy 4:91
SNPs *see* Short neural precursors (SNPs); Single nucleotide polymorphisms (SNPs)
SNRIs *see* Serotonin and norepinephrine reuptake inhibitors (SNRIs)
SNS *see* Slovenská; národná; strana (Slovak National Party, SNS); Social networking service (SNS)
SNSs *see* Social network sites (SNSs)
SNTV *see* Single-nontransferable vote (SNTV)
Snyder, Louis 16:285
Snyder, Timothy, works, *Bloodlands* 6:813–814
So Ko-jong, works, *Tongguk t'onggam* 11:50
SOA *see* School of the Americas (SOA); Stimulus onset asynchrony (SOA)
SOAEs *see* Spontaneous otoacoustic emissions (SOAEs)
Soap opera 18:528
Soar 11:401–402
SOC *see* Selection, optimization, and compensation (SOC) model of life span development; Selective optimization with compensation (SOC); Sense of Coherence (SOC); Standard occupational classification (SOC); Standards of Care (SOC) for Health of Transsexual, Transgender, and Gender Nonconforming People
SOCE *see* Sexual orientation change efforts (SOCE)
Sociability 3:638–641, 22:827–828
Social
concept of 22:827–832
definition 22:305
discourse as constitutive of 6:510
and sociality 22:854–856
Social (history of concept) 22:827
after World War II 22:828–829
Arendt on 22:828–829
computing and information technology 22:831
Durkheim on 22:828
in early social science 22:828
in early twenty-first century 22:829–831
in eighteenth century 22:827
in Germany 22:828
in late twentieth century 22:829–831
Latour on 22:829–831
Mauss on 22:828
philosophy 22:829–831
in social history 22:829–831
Weber on 22:828
Social abandonment 6:554–555
Social abilities 22:371, 22:378
Social acceleration 3:52
Social accounting 3:3
Social Accounting Models (SAMs) 3:3
Social accounts, and narratives 16:220
Social action 12:843, 18:798, 22:132–138, 22:522–523, 22:985, 22:987, 25:464
class as 22:341
conceptualizing 22:134–135
embedded, conflict as 4:621
Goldthorpe on 10:279–280
and interests 12:391
language as 18:788–789
for narration 16:219

Social action (*continued*)
phenomenology 18:11, 18:21
and theater 17:710–711
Weber on 25:459
Social action theories
and education 7:147
ignorance in 11:579–580
Social activities
fieldworker intervention in 9:155
transformational model of 22:491
Social actors 2:749, 22:914, 24:461–464
children as 3:492
Social agencies 25:261
contributions to social work education 22:775
Social agenda, United Nations 24:750
Social agreements awareness 4:670–671
steps of method 4:670–671
Social amplification of risk framework (SARF) 20:713–715
Social analysis 9:770
and Markov models 14:576–583
Social and behavioral science institutions, in Asia 22:615–620
Social and behavioral sciences (SBS) 2:88, 15:319–325
actors 15:320–321
analytical perspectives 15:319–320
contexts 15:323
culture 15:321–323
under dictatorship 22:139–146
empirical methodology 15:320
historiography 15:319–320, 15:323
history of ethical questions in 8:53–59
history of science 15:319–322
ideas 15:321
impact of Cold War on 4:122–127
implications of realism for 20:23–24
institutions 15:322–323
instruments 15:321–322
philosophy of science 15:319–320, 15:323
politics 15:323
replication in 20:441–442
sociology of science 15:319–320, 15:322–323
Western and non-Western views of 25:541–550
Social and biological transitions 15:369–371, 15:373
Social and cognitive connections among scientists 21:263
Social and cognitive self-efficacy, model of 8:395
Social and Cultural Planning Office (SCP) (Netherlands, the) 5:447
Social and cultural representations, and old age 17:180
Social and economic inequalities 12:944
Social and economic order 14:19–20
Social anomalies 24:746
Social anthropology 18:364–365, 22:992
ethnographies in, craft production 5:109
fieldwork in 9:149–152
in Korea 2:71–72
Lévi-Strauss on 13:908–909
in Melanesia 17:154
social network analysis in 16:528–529
Social anti-individualism
Wittgenstein on 25:613–614
Social anxiety 8:239, 22:494–495
Social anxiety disorder (SAD) 22:493, 24:194
cognitive factors 22:496
developmental course of 22:494
age-dependent presentation 22:494
prevalence throughout lifespan 22:494
treatment consideration 22:494
genetic factors 22:494–495
parental influence 22:495–496
sensitizing events 22:496
temperamental factors
anxiety sensitivity (AS) 22:495
behavioral inhibition (BI) 22:495–497
Social appraisal 7:487
Social apprenticing 6:842
Social approaches, for physical activity 18:96–97
Social archaeology 14:671–672
Social area analysis 22:266, 22:302
and crime 5:124–125
Social arrangements 4:809
Social ascent
and migration 22:426
opportunities for 22:427
and social mobility 22:426
in United States 22:428
Social aspects
of Alzheimer's disease 1:584
of cancer screening 3:79–81
of tactile processing 24:16
Social aspects/public relation organizations (SAPROs) 20:524
Social assistance, and Eastern European social work 22:795–800
Social assistance programs, in OECD countries 25:491t
Social attitudes 2:265–266, 19:562–568, 22:162–167
factors that shape aggregate-level opinions 19:566
factors that shape individual opinions 19:564–566
isms 22:163–165
Social attribution 5:564
Social authority 13:768
Social avoidance 2:456–457
Social axioms, definition of 25:28
Social background
degree of educational equality according to 25:485
effects, mapping 22:332–333
Social balance theory 16:525, 16:530–531
Social behaviorist 14:835–836
Social behaviors 2:488–489, 5:690
of children 3:452, 3:454
in collectivism/individualism 4:208
current primate studies in 18:892–893
emotions and 8:464
evolution and 8:464–465
evolution by natural selection 22:892–893
gender differences in 9:750–756
genes and 8:465–466
and genetics 9:984
hormones and 8:466
linking agreeableness to 1:465–468
in negotiation 16:434
and neonatal bonding 16:471–473
and prefrontal cortex injury 18:815–816
and sibling order 21:923
and social cognition 22:204
Sorokin's work 23:23
and temperament 24:180
Social belief system 4:152
Social beliefs 22:162–167
ideologies as 11:574–576
isms 22:163–165
values 22:164–166
Social belonging, of sexual minorities in China 21:737–738
Social boding, emotion and 7:478
Social borders 2:762–763
Social boundaries, popular music categories 3:800–801
Social brain 10:269, 13:900, 15:540, 16:601–604
Social brain hypothesis 8:441, 22:485
Social burden, and value of life 25:19
Social business 22:168–173
base of the pyramid 22:171–172
bottom of the pyramid 22:169
corporate social responsibility 22:168–169, 22:172
and entrepreneurship 7:684
hybrid organizations 22:170–171, 22:173
inclusive business 22:168, 22:171–172
ISO 26000 22:169, 22:173
nonprofit management 22:170
nonprofit organizations, managing 22:170
philanthropy 22:168–170, 22:172–173
social entrepreneurship 22:170–171
sustainability 22:168–169, 22:173
Social business management 22:173
Social capital 3:93, 3:627, 4:323–324, 6:482, 14:52, 14:113–114, 15:411–415, 15:453, 15:455, 16:782–786, 18:181–182, 18:783, 21:42, 22:174–185, 22:933, 23:369
among organizations 22:184
benefits 22:182–183
Bourdieu on 2:779–780
brokerage 22:182–184
causes 22:183–184
closure 22:182–184
communities 22:174–175, 22:177–179
and community 4:385–386, 4:388
concept 22:181–182
crime reduction 22:178
economic prosperity 22:174–175, 22:178
education 22:174, 22:178–179
effects of 22:178–179
experimental evidence 22:176–177
family and schooling 8:754
and fertility 9:22
forms 22:182
game theory 22:176–177

health 22:174, 22:178
and health 22:369
individual psychology 22:177–178
international migrants 15:413
inter-organizational relationships 22:184
and interorganizational relationships 12:623
interpersonal trust 22:174–175, 22:178
intra-organizational relationships 22:181
leisure activities 13:846
measurement 22:175–176, 22:176f
mental health 15:189
network centrality 22:182–184
and networks 16:520
origins of 22:175–178
outcomes 22:183
and philanthropy 3:365
and relational social work 20:223
religion and 20:265
role in internal migration 12:438–439
social control 22:182–183
social influence 22:181, 22:183–184
social information 22:181–184
and social media 22:424–425
and social media, linking 16:783–784
social networks 22:181–184
sociology 22:897
sources of 15:412–413
subjective well-being 22:174, 22:178
tie strength 22:184
trust 22:183–184
and trust 24:668–669
and value of children 3:498, 3:500
voluntary activities 22:174–175, 22:178–179
welfare 22:176
Social capital theory 3:131, 15:411
and migration 15:469
Social capitalism 22:173
Social capture–recapture studies 18:604–605
Social care 9:801–802
and LGBT aging 13:866–867
minority ethnic elders 15:440, 15:442–444
Social Care Institute for Excellence 22:771
Social care workers, implications, in England 4:346
Social catalyst 11:646
Social catchment field, of liberal parties 14:21
Social categories
and identity 11:526–530
and social networks 16:564–565
and social structure 22:719
Social categorization 8:150, 22:186–191
accessibility 22:187–188
availability 22:187–189
controversies and enduring issues 22:190
entitativity continuum 22:189–190
fit 22:187–188
comparative fit 22:187–188
normative fit 22:187–188
forming impressions
of groups 22:189
of people 22:188–189
impression formation 22:187–190
and individuation 22:188
intergroup relations 12:410–411
decategorization 12:411
dual identity 12:411
mutual intergroup differentiation 12:411
recategorization 12:411
meaning-seeker approach 22:188
and narratives 16:220–221
and organizational diversity 4:738–740, 6:595–597
outgroup homogeneity 22:187, 22:189
perceiver readiness 22:187–188
salience 22:187–189
self-categorization theory 22:186–189
stereotypes 23:446
stereotyping 22:188–189
storing and using categories 22:187–188
Social causation 3:261–265
Social causation, and life course epidemiology 14:72
Social change 1:328, 13:472, 13:542, 13:550, 13:567, 13:593–594, 22:147–155, 22:171
actors and structures 22:195
specifying assumptions 22:195–196
agents, youth as 25:393
analysis, Tocqueville 24:422–423
and biography 2:637
Cardoso's ideas on 3:122–125
cargo cults 3:139–140
Collins on 4:214–215
and communitarianism 4:314–315
and community work 4:378–380
and consumption 4:757–758
and creolization 5:120
decollectivization of 4:155
definition of 13:543–544
direct *versus* indirect effects 13:543–544
positivist *versus* constitutive approaches and 13:544
definition of law and 13:542–543
demands of active aging 22:154
and desegregation 6:219, 6:221, 6:223–227
disputes 6:563
economic inequalities 22:194–195
Eisenstadt's contribution 7:322
and emotions 7:538
and entrepreneurship 7:684
and equity in education 7:919, 7:921
as evolutionary 9:538
expected *versus* desired consequences 22:196–197
and feminist psychotherapy 8:929, 8:931, 8:933
and fertility decline 9:775–776
and generation 9:853–854
Habermas on 10:471
and hegemony 10:743
and human behavior 22:148
and identity/identification 11:529–530
impact in societies 9:438
impact on family socialization 22:870–871
individual actions and social facts 22:192
and innovation 12:170–171
intentions *versus* effects 22:196–197
and Internet 12:601–602
and law 19:559–560
linking macro with micro 22:150–151
loci of social change and POET model 22:193–195
mechanisms through which law brings about 13:544–547
organizational and institutional resistance and 13:545–546
rights debate and 13:545
social movements and law and 13:546–547
methodological approaches 22:151
Middle East 16:384–386
models 22:197
and participatory action research 17:536–538
political change 22:147–148
political psychology 18:432–433
psychological research 22:148–149
cascading effects 22:149
comprehensive effects 22:149–150
cultural traditions 22:149
polyvalent effects 22:150
syntax of change 22:149
recent research 22:151–153
benefits of social change 22:153
demands 22:152
moderating effects of context 22:152–153
negotiating demands of social change 22:152
well-being 22:152
requirements of theories 22:192–193
role of national contexts 22:153
Smith on 22:123
social differentiation 6:391
and social justice 22:135–136
and social movements 22:447–449, 22:518–523
by tradition 24:508, 24:510
and war 4:607–608, 4:611–612, 25:409–412
work anthropology 25:663–664
see also Erotic capital
Social Change and Crime Rate Trends: A Routine Activity Approach (Review) 20:772
Social change approaches 22:358
Social character 9:460, 9:462–463
of movements 22:443
Riesman on 20:671
Social characteristics of marriage payments 6:642–643
Social choice
conjoint measurement and 14:863
and welfare economics 25:500–501
Social choice theory 7:90
Social citizenship 3:638–639, 3:652, 20:746, 25:476–477, 25:511
and human rights regimes 3:639
Social classes 3:755–759, 10:50–51, 13:199, 21:842, 22:275
see also Class

Social classes (*continued*)
Boltanski on 2:747–748
Bourdieu on 2:778–780
and childhood behavior problems 2:477, 2:479
children's school learning 8:752
'class effect' rationale 3:758
class voting 25:299–301
and crime 5:122–126
business cycle, unemployment, and prisoner characteristics 5:123–124
correlates of crime 5:124–125
McKay on 5:122–124
Merton on 5:123
research tradition 5:123–125
Shaw on 5:122–124
social area analysis 5:124–125
Sutherland on 5:123
theoretical tradition 5:122–123
victim and self-report surveys 5:125
culturalist phase (c.1985–1995) 3:756–757
definition 15:640
and field theory 9:142–144
gradational regime 3:756f
and health 22:276
and income inequality 11:740
and intelligence 12:318–320
intergenerational mobility 15:642–644
and intersectionality 12:657, 12:659–661
and labor history 13:171
and language 13:281–287, 14:171–174
alternatives 13:286–287
communities of practice, habitus, and linguistic marketplace 13:286–287
dialect 13:281–282, 13:285
early studies 13:281
gender 13:281–282, 13:285
grammatical variables 13:283, 13:283t
identity 13:286–287
language change 13:285–286
lexical and discourse variables 13:283–285, 13:284t
methods 13:281–282
other sociolinguistic patterns 13:285, 13:285t
phonological variables 13:282–283, 13:282t–283t
prestige 13:282, 13:284–286
results 13:282–285
social networks 13:286
social status 13:281–282, 13:285–286
style 13:281–282, 13:285–287, 13:282t
urbanization 13:281
and leisure 13:853–856
and media audiences 2:217, 2:221
political sociology 18:472–473, 18:476
postmodernist phase (c.1995–present day) 3:757–758
regime 3:756f
and sexual behavior 11:194
social patterning of 4:759–760
stratification 12:802
structuralist phase (c.1945–1985) 3:755–756
transition to adulthood 24:572–573
Wilson on 25:585–586
Social classification 22:199–203
biological perspectives 22:200
common themes 22:202–203
categories 22:202
groups 22:202
membership 22:202
stratification 22:203
definition 22:199
identity 22:199, 22:202
pragmatic forms 22:201–202
age 22:201
caste 22:201
class 22:201
ethnicity 22:201
gender 22:201–202
nationalism 22:201
race 22:201
sexuality 22:201–202
psychological perspectives 22:200
sociological perspectives 22:200–201
and taboo 24:1–2
Social climate and psychological treatments 19:380–381
Social cognition 2:460–461, 4:300, 16:426, 17:314, 22:204–209, 22:534
and autism spectrum disorders 2:277
characteristics 22:204
and cognitive psychology 22:204–205
definitions and features 22:204
early deficits in 6:14
evolution of language 8:364–365
goal-oriented behavior 10:268–274
heuristics in 10:825
history of 22:205–206
and imitation by infants/young children 11:625
implicit association test 11:705
infant gaze behaviors 12:31
information processing 22:206–207
dual-process models 22:207–208
and joint attention 12:845–846, 12:849
and morality 16:607, 16:609–610
and small group research 22:104, 22:104f
social perception 22:204–208
and social referencing 22:560
thinking about social issues 22:206–208
reliance on mode of thinking 22:206–207
views of self 17:817–818, 17:818f
Social cognition models (SCMs)
and health behaviors 10:584
Social cognitive ability tasks 22:374t
Social cognitive approach to personality 17:921
Social cognitive determinants 10:623
Social cognitive domain theory 15:777
Social cognitive learning theory 22:863
Social cognitive theory (SCT) 1:374, 7:238–239, 7:243, 8:731–732, 10:578, 10:585, 15:932–933, 18:93, 19:549, 21:515–516, 22:530, 7:239t
basic principles of 15:932
observational learning 15:932
self-efficacy 15:932
triadic reciprocity 15:932
competition and cooperation 4:857
implications for learning and instruction 15:932–933
interactions between individuals, within social contexts 15:932
learning enhancement through observation 15:933
self-efficacy beliefs 15:933
of racism 19:866
role of self-efficacy in 21:509
of self-regulated learning 21:542
Social cohesion 10:417, 14:825–826, 15:412–414
mobile communication 15:633
and social structure 22:721
Social collective, and values 25:29
Social communication 16:286
and autism spectrum disorders 2:276
Social communication disorder (SCD) 18:790–791
Social community of others 10:533
Social compact, and social work 24:756
Social Comparison Record 22:211
Social comparison theory (SCT) 6:430, 15:660, 19:404, 23:624
Social comparisons 6:430–431, 6:433–434, 21:463–464
academic self-concept and achievement 1:59–61
and group psychotherapy 10:430
I/E model 12:425–426, 12:429–430
psychology of 22:210–215
ability 22:210–214
affect–cognition priming 22:211
assimilation and contrast 22:212
changing face of similarity 22:211–212
chronic illness 22:214
coping 22:211, 22:214
education, assimilation and contrast in 22:214
frog pond effect 22:214
narrative approach 22:211
opinion comparison 22:212
original theory 22:210
proxy 22:211–212, 22:214
reaction approach 22:211
related attributes 22:211–213
selection approach 22:210–211
selective accessibility model 22:213–214
self-evaluation 22:211–214
social inequality, invidious comparisons, and health 22:214–215
social reality 22:210
unidirectional drive upward 22:210–213
upward assimilation theory 22:212–213
reference group 20:73–74
and self-development 21:493, 21:495
Social competence 22:371, 22:375–378
during adolescence 22:216–221
cultural variations 22:218–220
diversity in relationships 22:219
diversity in social contexts 22:219–220

diversity in social goals and values 22:220
friendships and intimate relationships 22:217–218
parent–child relationship 22:217
peer relationship 22:217–218, 22:220
self and cultural identity achievement 22:218
social-developmental milestones 22:217, 22:220
sociocultural systems 22:217
concepts 22:375–377
definitions 22:375t
design issues 22:377–378
framework of 22:376f
measurement issues 22:377–378
pretend play and 18:872–873
Social competition 23:464–465
Social complexity 23:66
in Europe 8:270
production 12:329, 12:331
and warfare 4:616–617
Social conditions
of autonomy 2:314, 2:318–319
of enablement 9:1018–1019
Social conflict 9:539, 12:941, 25:48–50
injustice and 12:939
political trials and 18:500
views of justice and 12:937–938
Social connectedness 2:867
Social connotations, of drug use/abuse 6:687
Social consequences of population aging 18:543
Social conservatism 2:255
Social constancy 22:192
Social constitution of knowledge 21:201
Social constitutionalism 13:411
in Latin America 13:411
Social constructed ethnicity 8:149–150
Social construction 4:730–732, 5:542–543, 5:545–546, 5:548, 8:577–578, 10:786, 13:115
of adoption 1:142–143
of disputes 6:560–566
and drinking 6:660
of environmental problems 7:797–798
of identity 11:524
of mental illness 24:135
of nature 23:86
of ordinary knowledge 24:135
of race 19:833
race as 22:821
of reality 14:650–651, 22:538, 22:723–724
remembrance as 1:881–882
scientific evidence and 21:299
of scientific knowledge 21:309
of sexual minority categories 13:880
and social movements 22:450–451
of technology 21:202–203
theories 14:200
Social construction of technology (SCOT) 9:724, 24:111, 24:135
artifacts 24:136–138
bodies of work 24:135
constructivist studies of 24:135
developments in 24:139
heuristic study of 24:136–138
in nanotechnologies 24:139
normative 24:139
within organizational sociology 24:139
origin and development of 24:135–136
political dimensions 24:139
proponents of 24:136
theoretical perspective 24:138–139
Social constructionism 5:523–524, 9:699, 10:404, 10:786, 13:153, 14:651, 16:329, 17:390, 19:491, 21:849, 22:817
conversation analysis 22:225
critical psychology 22:226
determinism 22:223–226
discourse analysis 22:225
epistemology 22:222–223
essentialism 22:224
individualism 22:223, 22:225–226
issues and debates 22:225–226
identity and subjectivity 22:225–226
realism *versus* relativism 22:225
relationship between individual and social 22:226
knowledge in context 22:223
origins 22:222
positivism 22:223, 22:226
power 22:222–225
in psychology 22:226
psychology's model of person 22:223–225
antiessentialism 22:223–224
centrality of social interaction and discourse 22:224–225
rejecting reductionism 22:224
research 22:225
psychosocial studies 22:226
time and place 22:223
Social constructionist approaches 4:724, 4:726, 16:294–295
nation as imagined community 16:294
nation as invented tradition 16:294
nationalism as discourse, narrative, and event 16:295
nationhood as institutionalized form, practical category, and cognitive scheme 16:294–295
Social constructivism 3:614–615, 18:101, 20:230, 22:228–234, 23:82–86
actor network theory 22:230–232
approach for emotions 7:475
culture of knowledge 22:229–230
and ecofeminism 6:846
facets 22:228
historical background 22:228–229
history of science 21:225–228
international relations 22:232
methodological principle 22:230
and moral entrepreneur 15:798
radical 22:231–232
and ontology 22:231
and the social 22:231–232
social epistemology 22:229–231
nonveritistic approach 22:230–231
veritistic approach 22:229–230
social ontology 22:232–233
sociological analysis of science 22:229–230
strong program of sociology of science 22:230
STS and 12:602
Social constructivist theories 22:817–818
Social constructivists 21:244
Social context cues theory, of online dating 17:212–213
Social context of activity, knowledge as 21:98
Social context of fathering 8:844–845
Social contexts
child and adolescent psychiatry 3:384
conflict 16:427–428
culture 16:427–428
gender 16:427
diversity, during adolescence 22:219–220
and illiteracy 11:591–592
illness behavior 11:600–601
negotiation 16:423–424
culture 16:424
gender 16:424
power 16:424
social motives 16:423–424
PA process 17:718
school burout in 21:62
school engagement in 21:62
for universities in twenty-first century 18:308
Social contract 3:680, 6:90, 11:817, 13:437–439
and equality 7:909–911
and liberalism 14:23
Social contract theories, Rawls' contributions 19:914–915
Social control 4:822–827, 4:834–838, 13:550, 13:808, 15:105, 20:333–334, 22:235–241, 23:727–728, 23:734–735
anomie 22:238
anomie and 1:720–721
and antisocial behavior 1:790, 1:792–793
Black on 4:825
Chicago School of Sociology 22:237–239
collective efficacy and 4:824
concentration of poverty, association between 20:334
conflict management 4:825
contemporary 24:117–118
control theory 4:824
and counterinsurgencies 5:99
of crime 4:825–826
in criminology 4:824
cultural anthropology 4:824–825
deterrence 22:235–236
engineered 24:118
engineering of 24:117
ethical implications 24:121–123
forbidden under formal systems of 20:335
goals 22:238
history and sociology of state dealing with deviants 4:823–824
informal sanctions 4:822

Social control (*continued*)
labeling 4:823
mass executions and ethnic cleansing as form of 20:336
medical 15:105
morality 22:235, 22:238–239
neutralizing 24:122–123
normlessness 22:236–237
original meaning of 4:822–823
and police 18:251
and prejudice 18:835–836
punishment 4:823–826, 22:235–236
and punishment 19:634
and race–crime relationship 19:842
rational choice theory 4:825–826
relation of religion with 20:334–336
and religion 20:333–337
religious adherence and participation as 20:336
self-help 4:825
and sexual behavior 21:700
six strategies of 24:118–121
skeptic reaction 4:823
and social capital 22:182–183
social order 4:822–824, 22:235–236
strain theory 22:238
and technology 24:117–124
third parties 4:825
values 22:235–240
witchcraft 25:605, 25:607
of work 25:689–690
employee participation 25:689–690
technological progress control 25:690
Social control hypothesis 5:147–148
Social control model 9:829
Social control theory 5:146, 13:150–153, 23:496
and crime 23:496
and delinquency 6:70
extensions or reformulations 6:71
of race–crime relationship 19:839
Social convention 15:777–778
Social cost of carbon (SCC) 3:860
calculations 3:860–861, 3:863
Social creativity 23:464–465
fetishism as 9:108–109
Social crisis, and social question 22:553
Social criticism, and communitarianism 4:309
Social Darwinism 1:680, 1:682, 5:677–678, 5:684–691, 7:723, 8:239, 8:386–387, 8:407, 10:92, 11:75
Ammon on 5:689–690
anthropology 5:689
current trend 5:690
de Lapouge on 5:689–690
definition 5:686
eugenics 5:688–690
evolution 5:686–687, 5:689–690
ideology 5:686, 5:689–690
in Korea 2:71
liberal 5:687–688
selectionist 5:688–690
socialism 5:688–689
sociobiology 5:686, 5:690
and Spencer 22:991
Spencer on 5:686–687
unity and diversity of 5:687–690
Veblen on 25:55
Social death 5:867, 6:550, 6:554–555
form of slaves/slavery 22:54
Social decision-making 16:643–645
Social deixis 6:54
Social democracy 22:242–244, 25:529
in Brazil 3:125
challenges 22:243–244
democratic corporatism 22:243
history and significance of 22:242–243
labor 22:242–243
labor parties 22:243
and liberalism 14:29–30
socialism 22:242–244
state socialism 22:242–243
and Walzer, Michael 25:349
welfare state 22:242–243
Social Democratic Federation (SDF) 2:498
Social Democratic Party (SDP) 6:690, 16:160–161, 22:838–839, 22:842–843
Social democratic welfare state 24:573
Social Democrats 21:142
Social demography, and family therapy 8:815–820
Social dependencies, modeling 19:443
Social deprivation
area deprivation 22:285–286
concepts and definitions related to 22:285
educational deprivation/lack of qualifications 22:286–287
employment deprivation/long-term unemployment 22:286
and health 22:285–290
health inequalities 22:286, 22:289
illiteracy 22:287
income deprivation and poverty 22:285
policy responses 22:289
proportionate universalism 22:289
and religiosity of women 25:635
selection process 22:286, 22:288
social gradient 22:289
Social desirability
and agreeableness 1:464
questionnaires 19:771–772
and sexuality 15:315–316
Social desirability bias 8:573, 19:564
Social desirability scales 17:836–837
Social desirable responding (SDR) 17:834
Social determinants
of health 7:741, 10:199, 10:204, 10:634–635, 10:653–654, 22:931, 22:935
nonhierarchical 22:276–277
treatment research and access 8:34
Social determinants and consequences of disability 6:469
Social determinism 2:751, 11:413
Social development 10:264–266
and aggression 1:377
children, media effects on 15:37–39
materialism 15:38–39
prosocial effects of media exposure 15:39
racial and gender stereotyping 15:38
theorization 15:37–38
TV content effects on non-physical aggression 15:38
violence effects on physical aggression 15:38
gifted education 7:162–163
global agenda for 10:144–149
The Global Agenda for Social Work and Social Development (The Agenda) 10:144–149
international social work student exchanges 12:562–563, 12:567
neural correlates of 16:600–605
prenatal/infant 18:856, 18:859
prosocial 19:233–234, 19:236
social work 24:755
and social work education 22:776–777
Williams syndrome 25:581
and writing 25:782
Social deviance 22:703
disease as 17:2
Social dialectology 13:281
Social differences
in adolescent sexuality 21:841–842
in childhood sexuality 21:841
identities and 5:290
in morbidity compression 4:466
Social differentiation 6:391–394
and cohesion 13:160–161
consequences of 6:393–394
culture 6:393–394
as decomposition or emergence 6:391–392
division of labor 6:391–393
driving forces of 6:392–393
ecology 6:394
and fashion 8:836–837
and field theory 9:140
integration 6:394
life chance 6:393
ligatures 6:393
Luhmann's conceptualization 6:392
modern society 6:391–394
polycontexturality 6:392
progress 6:392–394
social change 6:391
social evolution 6:394
social integration 6:394
systems integration 6:394
value-spheres 6:391–392
Social dilemmas
ad groups 10:425
cooperation 4:862–863
economic psychology 5:887–888
and Honesty-Humility 11:171, 11:173
prosocial behavior 1:581
psychology of 22:245–249
in public goods provision 22:512–513
rational choice theory 19:902, 19:904
Social dilemmas, psychology of 22:245
observations and explanations 22:246–248
communication 22:246
environmental uncertainty 22:247
group identity 22:247

group size 22:246
noise 22:247
personality 22:247–248
reciprocity 22:247
tacit coordination 22:246–247
research trends in 22:249
solving dilemmas 22:248–249
leadership 22:248
sanctions 22:248–249
types 22:245–246
common resource dilemma 22:245–246
intergenerational dilemmas 22:246
intergroup competition 22:246
nested social dilemmas 22:246
prisoner's dilemma 22:245
public good dilemma 22:245–246
Social disadvantage, and nonmarital childbearing 9:14, 9:16
Social discount rate (SDR) 5:52
Social disorganization 4:322–323, 4:325, 4:836, 4:838
and collective efficacy 4:324
and delinquency 6:69
extensions or reformulations 6:71
heuristic model of 4:323, 4:322f
and homicide 11:163
and organized crime 5:170–171
in United States 11:628–629, 11:631
Social disorganization theories 5:161, 5:191, 5:197, 5:239, 11:628–629, 22:237–238
of race–crime relationship 19:839–840
Social displacement hypothesis 16:783
Social distance 15:230, 21:366
Social distance scale 18:842
definition 18:842
Social distinction, and consumption 4:759–760
Social distinction model 9:829
Social diversity 24:794
Social division 19:612
Social domain theory, and moral development 15:792–793
Social dominance, and conflict 4:585
Social dominance theory (SDT) 12:412
Social drama 17:709
Social dynamics 4:521, 22:990
Social Ecological Assessment Databank (SEAbank) 17:763–764
Social ecological exolevel analysis, honor killing 11:188–189
Social ecological macrolevel analysis, honor killing 11:188–189
Social ecological mesolevel analysis, honor killing 11:189
Social ecological microlevel analysis, honor killing 11:189
Social Ecological Model (SEM) 18:94
Social ecology 6:843, 8:180, 22:156, 22:159–160, 22:254–262
academic traditions 22:254–255
biohistory and society–nature coevolution 22:259
Chicago School 22:263–268
concepts and their transfer to social and behavioral sciences 22:263
concepts used in social and behavioral sciences 22:264–265
methods 22:266
origins of approach 22:263
reasons for loss of influence 22:266–267
reasons for using 22:265–266
successors 22:267–268
environmental impacts of human activities 22:258–259
of personal projects 17:757–758
regulation, governance and sustainability transitions 22:259–260
society's biophysical structures 22:255–258
energy and society 22:255–256
land use and food production 22:256
social metabolism/industrial ecology 22:256–258
socioeconomic status 22:159
and war/political violence 25:390
Social economy 3:366
Social education
author as 6:621
social studies as 22:730–731
Social effects of automation 2:298–300
Social embeddedness 4:149, 7:64
Social embedding 4:621, 4:863–865
and groups 10:436
Social engagement, and health 22:369
Social engineering 24:118–119
Islamic fundamentalism 12:741–743
of socialist societies 22:851–852
Social enquiry, for social work 22:269–274
critical social enquiry 22:269–270
Social enterprise 22:170–173
and nongovernmental organizations 16:858–859
Social entrepreneurs 7:676
Social entrepreneurship (SE) 22:168, 22:170–171
defined 22:170–171
innovation 12:173–174
Social environment 9:827, 9:829, 9:831, 15:88–89
and social cognition 22:204
Social epidemiology 2:656, 2:659, 15:89–90, 22:275–281, 22:283
age-period-cohort analysis 22:275, 22:278
directed acyclic graphs 22:279
field 22:275
fundamental cause theory 22:277
micro- and macrocontextual views 22:278–279
country, nations, institutions, and welfare regimes 22:279
neighborhood and area 22:278–279
morbidity 22:279
mortality 22:276–279
multilevel modeling 22:278–279
nonhierarchical social determinants of health 22:276–277
cognitive and noncognitive abilities 22:277
ethnocultural differences 22:277
lifestyle 22:277
perspectives on methodology, social policy, and translation to practice 22:279–280
social inequalities in health and disease 22:275–276
empirical evidence 22:276
social class 22:276
socioeconomic position 22:276
social relations pertaining to health differences 22:278
social stratification 22:276–277
social times in explaining health differences 22:278
socioeconomic gradient 22:275–277, 22:279
Social epigenetics 22:282
cognitive epigenetics and 22:283
and developmental disorders 22:283
disease and 22:283
environmental health 22:282–283
epidemiological epigenetics 22:283
histone modification 22:282–283
inheritance 22:282
methylation 22:282–283
research
foundational and animal model 22:282–283
limitations and future of 22:283–284
Social epistemology 22:229–231
nonveritistic approach 22:230–231
veritistic approach 22:229–230
Social equality 7:308, 7:313, 10:123
Social equity 7:927, 24:920
Social evolution 1:900
AGIL schema 1:388–389
clan 3:730
philosophical aspects 8:388–391
adaptive social networks 8:390
cultural evolution 8:389–390
fitness 8:388–389
imitation 8:389
learning 8:389–390
local interaction 8:390
population structure 8:389
replicator 8:388–389
social differentiation 6:394
Social evolutionism 9:536
and magic 14:435–436
Social exchange 13:641
and assortative mating 2:103–104, 2:106
Eisenstadt's contributions 7:322
and emotions 7:537–538
evolutionary group dynamics 8:442–443
and legitimacy 13:825
and organizational citizenship behavior 17:319–320
patron–client relations 17:604
resource theory of 12:953
Social exchange relationships, justice in 12:938
Social exchange theory 8:482–488, 19:367, 21:848–849, 22:541, 23:702
applications 8:486–487
Blau and Homans 8:482–483

Social exchange theory (*continued*)
of Dark Triad 5:674
emotions 8:486
empirical research 8:483–485
Emerson's theory of exchange relations 8:484
key assumptions 8:485
networks 8:484–485
power and dependence 8:484
generalized exchange 8:483, 8:486–487
negotiated exchange 8:483, 8:486
norms 8:482, 8:485
online exchange 8:487
power inequalities in networks 8:485–486
coercive power 8:485
risk and uncertainty 8:485–486
productive exchange 8:486
reciprocity 8:483
relational cohesion and 8:486
roots of 8:483
anthropological 8:483
psychological 8:483
solidarity 8:482–483, 8:486
trust 8:482, 8:486–487
Social exclusion 2:641, 10:265–266, 24:790, 24:792
and child protection 3:420
and health 22:285–290
concepts and definitions related to 22:285
criticisms of 22:288–289
discrimination 22:287
exclusion due to asylum seeker or refugee status 22:288
exclusion due to disability 22:287–288
exclusion due to homelessness 22:288
exclusion due to mental illness 22:288
exclusion from services and health/inverse care law 22:287
health inequalities 22:286, 22:289
illiteracy 22:287
policy responses 22:289
proportionate universalism 22:289
selection process 22:286, 22:288
social gradient 22:289
mass imprisonment 14:696–700
in psychiatry 22:977–978
and urbanization 24:963
Social experiences, child and adolescent psychiatry 3:384
Social experiments 8:531, 8:541–545, 9:132, 14:251, 22:291–300
application (1960–2000) 22:298–299
application (2001–2014) 22:299
comparison group designs 22:293–294
data analysis 22:296–297
description 22:291
designing 22:295–296
ethics 22:297–298
external validity of experimental estimates 22:297
history of 22:298
importance 22:291–292
internal validity of experimental estimates 22:297
population studies *see* Population studies
pre–post designs 22:292–293
rationale for 22:294–295
selection bias 22:294
Social explanada, analytical sociology 1:671–672
Social expression, of fashion 8:837–838
Social externalism 2:134
Social externalities 24:617
Social fabric 5:287
Social face information, role in face processing 8:654
Social facilitation effect 22:532
Social facilitation theory 22:103
Social factors
of business 3:1–2
and men's health 15:169–170
Social facts 6:703, 10:110, 12:192, 13:768–769, 21:344, 22:721
definition of 15:575
description, individualism and 15:305
Durkheim's 6:244, 22:991–992
and groups 11:818
and individual actions 22:192
and law 13:479–480, 13:482
in legal philosophy 18:28
Social features of diglossia 6:419–420
Social fields 5:9
Social flexibility 22:373
Social flow 9:263–264
Social fluidity 15:642–644
degree of 15:642
Social force
international advertising as 12:460
public opinion as 19:566
Social formations 16:285
in bourgeoisie/middle classes 2:784
Social forms 22:192
of East Africa 6:773–776
and social structure 22:720
Social framework evidence, for employment discrimination 6:531–533
Social functional, emotional expression 7:484
Social functioning
and differential aging 6:381–382
gastrointestinal disorders 9:607–608
Social games (heuristics) 10:834
Social generalization 25:48
Social generational ordering 9:855
Social generations 9:857–858
birth cohort 9:859
theory of 9:858–859
Social genetics 23:23
Social geoarchaeology 10:56–57
Social geography 1:407–410, 22:301–306
behavioral challenge 22:302
new directions 22:304–306
future for 22:305–306
social process and spatial pattern *versus* spatiality 22:304–305
politicized approaches 22:302–304
urban 24:801
Social gerontology 1:407
Social goals and values, diversity, during adolescence 22:220
Social gospel, stigma 23:462
Social gradient
in health 15:851
social deprivation 22:289
Social groups 22:982
ethnocentrism and prejudice 8:169
joint action 12:842
and knowledge 13:70
politeness from point of view of 18:328–329
small group research of 22:102
Social harmony 10:530–531, 10:533
Social hazards, urban life health 24:831
Social health, and religion 20:265
Social health insurances (SHIs) 10:644, 10:663–667
Social heterogeneity 23:373
Social hierarchy 1:965, 15:805
Social history 4:406–407, 4:765, 19:695–696, 22:830
Bielefeld School 2:553–557
Burke on 2:924, 2:926–927
child and adolescent psychiatry 3:387
crowds 5:361–362, 5:364
culture 22:308, 22:310–312
demographic history 22:309
early institutionalization 22:307–308
economic history 22:307–308
Freyre on 9:422–423
historiography 22:307–311
history of society 22:307, 22:311
labor history 22:309
and Marxism 22:308–311
Merton on 22:310
new history 22:308–310
population history 22:309
postwar expansion 22:308–311
national variations and international circulation of ideas 22:308–309
topics, neighbors, models 22:309–310
will to synthesis 22:310–311
and Reformation/confessionalization 20:99–100
revenge of success 22:311–312
and social inequality 22:338–343
Social history of law, Hurst's contribution 11:439
Social homogeneity 9:616–617
structural equivalence 23:623–624
Social housing, enriched environment 2:835–837
Social human rights 11:355
Social hygiene movements 8:241–244
Social identification 11:527–528
and computer-mediated communication 4:506–507
Social identities 4:152–153, 9:434–435, 11:526–530, 17:744–745, 22:313–324
actions, reactions, and consequences 22:322–323
collective behavior 22:323
stereotype threat 22:323

armed conflicts 17:655
biculturalism 22:321–322
and biocitizenship 2:611
categorization 22:319–321, 22:323
changes in 22:322
childhood 3:452
and conflict 4:585
connection between groups and identification 22:314–315
definition 22:313–314
and deindividuation 6:38, 6:40–41
and ethnicity 8:148–150
evolutionary group dynamics 8:444
and flashbulb memory 9:258
and gender archaeology 9:764–765
Goffman on 10:276
and groups 10:422–423, 10:425–426, 10:435–437
and hazing 10:556
identity importance 22:320–323
importance in social and behavioral sciences 22:315
insider and outsider perspectives on 22:320–321
intersectionality 22:321
measurement 22:315–317
multiculturalism 22:321
multiplicity of 22:321–322
role and categorical 4:174
self-categorization theory 22:314–316
self-other similarity 22:316–317
self-stereotyping 22:314–316
and social media 22:422–423
identity performance and impression management 22:422–423
psychological approaches 22:422–423
social psychology 22:313–315, 22:317
socialization 22:313–314
stereotypes 23:447, 23:450–451
symbolic interactionism 22:313–314
and values 25:31
Social identity analysis, in feminist psychotherapy 8:931–932
Social identity approaches, to collective memory 4:197
Social identity model of deindividuation effects (SIDE) 6:40
of computer-mediated communication 4:506
Social identity salience, perceiver readiness and fit as determinants of 21:456–457
Social identity theory (SIT) 10:435–436, 10:761–762, 12:412, 20:382–383, 22:539, 25:804–805
ethnic identity 8:113–115
and ethnocentrism/prejudice 8:171–173
meritocracy 15:239–240
and organizational groups 10:412–413
and prejudice 18:834–835
of racism 19:866–867
and tyranny 24:713
of war/political violence 25:392
Social imaginary, Charles Taylor 24:41–42
Social imagination, and aesthetic education 1:230–231
Social Impact Bonds (SIBs) 22:360
Social impact investing 3:365
Social impact litigation against governments 3:735–736
Social impact of computing 21:187
Social implications, of disasters 22:754–755
Social inclusion 2:641
in Eastern Europe 22:797–798
ICSW 12:498, 12:500
and social integration 12:245–246
values and ethics 22:756
Social increasing returns 6:604
Social indicators research 10:524
Social inequalities 5:451, 11:890, 11:894, 11:908–913, 16:791, 22:275, 22:338–344, 22:939–940
challenge of culture in history 22:341–342
Bourdieu's class and cultural distinction 22:342
class as social action, experience, and culture 22:341
class experience from production to consumption 22:341–342
in child and infant health 12:15, 12:19–20
from classes to lifestyles, milieus, and radical individualism 22:342–343
constructing class and inequality 22:343
in cultural consumption patterns 22:331–337
criticisms and controversies 22:334–335
cultural inequality, theoretical accounts 22:331–332
empirical tests of theories 22:332–334
future directions 22:335–336
and division of labor 13:158, 13:160–161
from economic realism to cultural relativism 22:343
and education 7:293–294
and ethnicity 8:137–138
gender, race, and ethnicity 22:342
in health 17:732–733, 22:933–934
in health and disease 22:275–276
empirical evidence 22:276
social class 22:276
socioeconomic position 22:276
Marx on 22:338–342
and media sports representation 14:724
and moral reasoning 15:805
and networks 16:520
opportunity 7:903
in practice 11:910–912
and production mode 15:714
and schooling 22:325–327
basic mechanisms 22:326
democratizing policies 22:329–330
impact of organizational characteristics 22:328–329
large social gaps 22:325–326
role of family 22:326–327
school factors 22:327–330
story told by international comparisons 22:327–328
social comparison 22:214–215
social history 22:338–343
and social mobility 22:427–428
social structure 22:338–340, 22:343
sociology 22:338–339, 22:341–343, 22:897–898
stratification and social classes in historiography 22:339–341
in Germany 22:340
in United States 22:340–341
theories of social stratification and class formation 22:338–339
Thompson on 22:338, 22:341
urban conflicts 24:785
Veblen on 25:55
Weber on 22:338–340, 22:342
working class 22:339–343
Social influence model 1:122
Social influence-based drug prevention programs 15:64–65
Social influences 15:659–660, 22:345–354, 22:532, 22:534
of anxiety 1:806
Cialdini's principles of influence 22:349–351
authority 22:350
commitment and consistency 22:349–350
liking, attractiveness, similarity, ingratiation, and immediacy 22:350
reciprocity 22:349
scarcity 22:350–351
social proof 22:350
versus compliance gaining 22:348
credibility and social relationships 22:351–352
plain folks appeal 22:352
source credibility 22:351
definition 22:348
deindividuation 6:40–41
as determinant of organized social behavior 21:457
emotional tactics 22:352–353
examination of 22:346
and fertility 9:7–8
group influence 22:346
and groups 10:423–424, 10:426
and health care 15:103
ingratiation 22:352–353
fear appeals 22:353
of international advertising 12:460
landscaping 22:351
defining and labeling issue 22:351
limited choices/options 22:351
and leadership 13:648, 13:650–651
mass media influence 22:345–346
message tactics 22:352
message discrepancy 22:352
self-generated arguments 22:352
personal influence 22:346
versus persuasion 22:348
research, influence of 22:348–349
and sexual behavior 21:700, 21:703
and social capital 22:181, 22:183–184
theoretical frameworks 22:349
Social informatics 12:86

Social information, and social capital 22:181–184
Social information dynamics 16:927
Social information processing (SIP) theory 17:213, 17:382
for aggression 1:375
of computer-mediated communication 4:506–507
and job satisfaction 12:834, 12:837
Social innovation 12:173–174, 22:355–361
contexts 22:357
definitions 22:356
diffusion and scaling impact 22:360
disruptive 22:355–356, 22:356t
distinctive challenges of 22:359–361
evaluation 22:360
failure, and unintended consequences 22:361
inception 22:359
innovation ecosystems 22:360
innovation in social relations 22:358–359
innovation to address social market failures 22:359
resilience 22:358–359
resilience theory 22:359
social enterprise 22:357
social entrepreneurship 22:358, 22:360
theoretical traditions 22:357–359
theory 22:358–359
wicked problems 22:355, 22:357
Social insideness 6:886
Social institutions 4:834–838, 12:191, 13:592, 22:540
Boltanski on 2:751–753
delinquency 6:68, 6:71
science as 21:200–201
slavery as 22:43
Wilson on 25:586–587
Social insurance 22:362, 23:4, 25:529
benefits
guarantee of 22:363
public–private mix of 22:363
range of 22:363
characteristic items of
financing 22:362
organization 22:363
future developments in the field of 22:363–364
international standards of
higher levels of social standards 22:363
minimum standards 22:363
and new liberalism 18:302–303
and private insurance 22:362
and social security 22:363
and working conditions 25:699
Social integration 12:243–248, 25:47
action integration 12:243–244
actor integration 12:243
beyond nation-state 12:246–247
citizenship 12:245–246
collective solidarity 12:246–247
and communitarianism 4:308–310
division of labor 12:246
and health 22:365–370
historical and theoretical conceptualizations 22:365–366
mortality data 22:367–368
social support 22:366
and marital quality 14:514
national integration 12:246–247
negative effects of 12:245
social inclusion 12:245–246
and social order 12:244–245
and social question 22:553, 22:555
and social structure 22:721, 22:724–725
and suicide 23:683–686
and systems integration 12:245–246
transnational integration 12:246–247
world society 12:243, 12:245–246
Social integration mechanism 20:336
Social intelligence 5:433, 17:750–751, 22:371–379
animal social learning 22:406
and autism spectrum disorders 2:275
as erotic capital 8:5
measurement approaches 22:373–375
theoretical models 22:373
Social intelligence hypothesis 4:395–396
Social intentionality 12:342–343
mirror neurons 12:343
theory of mind 12:342–343
Social interaction tasks, cardiovascular reactivity 3:118–119
Social interactional theory, of emotions 7:534–535
Social interactionist approach, to collective memory 4:189
audience tuning/shared reality 4:190
communicative processes 4:190–191
conversational/dyadic interactions 4:190–191
group processes 4:191
information propagation and social networks 4:191
noncommunicative processes 4:190
shared attention mechanisms 4:190
Social interactions 11:108, 13:671–673, 14:415–416, 18:772, 25:128–129
and autism spectrum disorders 2:276
and CDS 3:401
cognitive development 4:54–56
and collective beliefs 4:164–165
Collins on 4:211–217
and conflict 4:597–598
conversation analysis 4:850, 4:855
demographic microsimulation 15:344–346
diffusion 6:413–414
as dynamic and dialogical processes 13:672–673
ethnomethodology 8:206
and fertility 9:6–9, 9:22
formal models, diffusion 6:415
Goffman on 10:275–278
in infancy 12:349–350
influences of emotion regulation on 7:454–455
interviewing 5:851–856
and moral judgments 15:803–804
development stages 15:804
moralities and types of social interactions 15:803–804
and postmodernism 18:677–678
problems in postsocialist societies 18:699
and social constructionism 22:224–225
and socialization of children 22:868
Social interdependence theory
competition and cooperation 4:857
application of 4:859–860
breadth of 4:860
relationship among theory, research, and practice 4:860
Social interests of scientists 21:264
Social intimacy 19:754–755
Social investigation, in nineteenth century 18:320–321
Social investment, poverty policy 18:750–751
Social irresponsibility of advertising 1:201–202
Social isolation 1:544
coronary heart disease 4:917–921
neonatal 16:489–492
Social issues
gifted education 7:162–163
and old age 17:180
stigma 23:462
thinking about 22:206–208
Social judgment
and conflict 4:584–585
distinguishing morality from other domains of 15:804–805
of person perception 17:740
Social justice 2:716, 7:308–310, 7:312, 7:926, 10:385, 12:138, 13:720–721, 15:805, 16:461, 16:464, 17:540, 17:542, 22:132–138, 22:303, 22:304, 22:305, 22:386, 22:392–400
academic analysis of 22:387
Alinsky on 22:135
and anthropological writing 1:740
background 22:392
Catholic tradition 22:386
class struggle 22:394
community 22:132–137
and community work 4:378
conceptualizing 22:134–135
contemporary challenges 22:136–137
and criminology 5:246–248
CSE and AR 22:269–273
defining 22:386–387
democracy 22:398–399
and desegregation 6:221–225
different uses of 22:386
criminological applications of 22:386
distribution 22:392
and diversity 4:738, 6:595
empirical research on 22:398–399
equal justice 22:393
equal justice, limitations of 22:395
equality 22:392
equality of opportunity 7:903–904, 22:398
fairness 22:392, 22:394
feminism 22:135–137

and feminist psychotherapy 8:929, 8:931, 8:933
first generation of rights 22:393
freedom 22:392–393
Freire on 22:134–135
future possibilities for concept and practice 22:399–400
future prospects 22:396
and genetics 9:969–974
genocide 22:394
global justice, struggle for 22:395–396
and global south 10:188–189
globalization 22:398
harm reduction 22:394
and human development 11:276, 11:280
human justness 22:396
human nature 22:397–398
human needs 22:396
human rights 22:392, 22:396
and human rights 11:353
ICSW 12:495–498
inequality 22:392, 22:394
international law, limitations of 22:395–396
justice 22:393
and liberation theology 14:32–33, 14:35–36
liberties 22:392
local and global 22:398–399
mapping social work's commitment to 22:133–134
mental health and 22:389
mobilization, and trade unions 24:492
movement building of 24:495
Myrdal on 16:159–163
nation-state 22:398–399
nonviolence 22:399
operationalizing 22:135–136
oppression 22:397
peacemaking and 22:387–388
professionalism 22:134–136
promotion of 22:782
psychological identification of 22:387
public health and 22:389
and radical social work 19:870, 19:873–874
Rawls's theory of 7:909–910, 19:914
Rawls's thoery of 22:392
and recognition politics 20:44–48
redistribution 22:392
research 22:386–387
restorative justice 22:393–394
and risk assessment 20:696–697
second generation of rights 22:393
social justice 22:394–395
social movements 22:132–137
social structure 22:397–399
and social work 22:521–522, 24:755
sovereign justice, struggle for 22:395
spirituality 23:288–289
structural injustice 22:398–399
structural violence 22:398–399
Taylor on 24:40
theories 22:397–398
third generation of rights 22:393
Universal Declaration of Human Rights (UDHR) 22:392
in urban planning 19:821–822
war on drugs 22:395
Social justice theory 4:146
Social knowledge
cognitive-anatomical basis of 16:610–611, 16:610f
new liberalism 18:299–300
Social laboratory 7:561, 8:556–558
Social landscapes, urban 24:804
Social laws 11:813
Social learning 3:451–452, 6:258, 8:438, 16:643–645, 18:291, 22:410, 22:485, 25:102
in animals 22:401–408
applied behavior analysis 2:407–408, 2:410
cognitive development 4:39
cultural evolution 5:388, 5:402, 5:407
and environmental education 7:735–736
and evolutionary perspectives 8:396
and fertility 9:7–8
and group psychotherapy 10:430
and imitation by infants/young children 11:621–623
and language development 13:321–323
like me theory 15:131
and pain 17:452
policy change 3:843
and prosocial behavior 19:230
and social referencing 22:558
Social learning principles 6:682–683
Social learning process 23:799
Social Learning Strategies Tournament (SLST) 5:407
Social learning systems, and communities of practice 4:328
Social learning theories (SLT) 1:374, 2:472, 5:239, 5:335–336, 13:707, 16:73, 18:126, 21:681, 21:848, 22:409, 23:496, 25:100–101
Akers&rsquo 22:410–412
definitions 22:411
differential association 22:410
differential reinforcement 22:411–412
imitation 22:410–411
and crime 23:496
and delinquency 6:70
empirical tests of 22:412–413
interventions 6:682–683
practical applications of 22:413
roots of 22:409–410
self-efficacy 21:509
social structure and social learning model (SSSL) 22:413–414
Social learning/sociotherapy model 24:276
Social legislation, human resource management 11:351–352
Social liberalism, Christianity 14:8
Social life 25:194–195
in Central Asia 3:320
human rights 11:377
restructurion, in postsocialist societies 18:697–698
and values 25:27–28
Social life of the honeybee colony 15:155
early short-term memory (eSTM) 15:158
glutamate transmission and memory formation 15:158–159
late short-term memory (lSTM) 15:158
learning
associative 15:158
contextual 15:156
under natural conditions 15:155–156
neural correlates of 15:156–157
nonelemental 15:156
rule 15:155–156
sequential memory processing 15:157–158
visual categorization 15:155
long-term memory (LTM) 15:158
memory formation 15:156–157
midterm memory (MTM) 15:158
navigation and communication 15:155
reconsolidation of reminder memory 15:159
Social life of things 6:404–405
Social lives 17:801
Social loafing 10:425
in collectivism/individualism 4:208–209
Social location 9:858–859
own, working of men on 14:682–683
Social markers, and linguistic anthropology 14:171–172
Social Market Economy 11:568
Social marketing 6:681
community-based health interventions 10:646–651
sustainability lifestyles 23:829, 23:831
Social marking
learning and instruction 13:671
Social mathematics 22:990
Social maturation 17:887
Social measurement 21:4–8
classification in 21:4–5
aggregate entities 21:5
taxonomies 21:4
typologies 21:4–5
error 21:7
relations 21:5
scaling in 21:5–7
dichotomous items 21:5–7
ordered assessments 21:7
theory 21:7
Social mechanics 23:23
Social mechanisms 14:875, 22:415–420
analytical sociology 22:415, 22:419
-based theorizing 22:417–418
causality 22:417
and causation 22:417
critical realism 22:415
future challenges 22:419
and individualism 22:418–419
mechanism-based explanation 22:415–417
methodology 22:415, 22:418–419
middle-range theory 22:418
in political science 21:305–306
process tracing 22:415
rational choice theory 22:418–419

Social mechanisms (*continued*)
rise of 22:415
theory of action 22:419
Social media 2:860, 3:627, 3:629, 5:704–705, 13:521, 16:782, 16:800, 22:421–425, 22:634
and advertising 1:213, 1:217
in Arab spring 18:423–424
automated language analysis 24:150
in China 2:68
collective level studies 16:784–785
and computer-mediated communication 4:504
consumption 16:785
curation 22:423–425
definition 22:421–422
effects 16:783–785
and foreign language learning 9:331
future directions for research 22:425
hyperpersonal model 22:422–423
impact on employment practices 24:106
and Internet 12:600, 12:602, 12:604
and journals 12:867
network science 22:424
online dating 17:212
online pornography 17:218
oversharing through 17:439–444
politics 22:423–424
power of 18:423–424
and religion 20:274, 20:276–277
self-presentation 22:422
small worlds 22:424–425
social capital 22:424–425
and social capital, linking 16:783–784
and social identities 22:422–423
identity performance and impression management 22:422–423
psychological approaches 22:422–423
social network analysis 22:421–422
social network sites 22:421–422, 22:424–425
strong ties 16:784
teens, gender, and self-presentation in 24:146–152
adolescents 24:146–150
communication 24:146–147
implications 24:150–151
language 24:149–150
privacy 24:148–149, 24:151
profile content 24:147–148
profile visibility and perceived audience 24:148–149
sexuality 24:150
social media use 24:146–147
textual communication and interaction 24:149–150
truthfulness of self-presentation 24:149
visual self-presentation 24:148
and trust 16:785
user-generated content 16:782
weak ties 16:784
Social media communication 4:288
Social memory 1:883, 4:203, 22:373
Social metabolism 6:24–25
industrial ecology 22:254, 22:256–258
material flow analysis 14:760, 14:763
Social metabolism and indexes of (un) sustainability 6:858–860
Social mirroring, and migration 15:430–431
Social mix 22:328
Social mobility 3:757, 22:333–334, 22:426–429, 24:731
alternative approaches 15:644–645
and aristocracy 22:427–428
and careers 22:426–429
class formation 15:640
in Communist countries 22:428
cross-national comparisons 15:643
debates on 22:427–429
definitions 22:426–427
and education 22:426–428
equality of opportunity 7:903–904
and equity in education 7:919, 7:922
of ethnic groups 22:426
gender in 22:428–429
Goldthorpe's contributions 10:281–282
historical research, future 22:429
of immigrants 22:426
intergenerational class mobility and social fluidity 15:642–644
intragenerational mobility 15:640, 15:642
meritocracy as 15:240
in middle class 22:427–428
opportunity structure 15:642–644
and poverty 18:744–745, 18:753–754
quantitative research on 22:427–428
rates 15:640–644
rates of 5:450
research questions and traditions of study 15:640–641
in social history research 22:426–427
social opportunities 15:640
social origin 15:640–642, 15:644
social stratification 15:640–642, 15:644
and social stratification 22:709–710
sources 22:426–427
status attainment 15:641–642
status group formation 15:640
in United States 22:428
and urban poverty 24:916
in Western societies 22:426
Wilson on 25:586–587
of women 22:426
Social mobilization 6:267, 24:842–844
Social models
of disability 6:447, 6:455, 6:462–463, 13:498, 15:201–202
fishing 9:226–227
for learning disability 13:684–686
Social monitoring, urban 24:803–804
Social morphology 18:668, 22:721
Social motivation, in group psychotherapy 10:430
Social motives, negotiations 16:423–424
Social movement groups 3:683–684
Social movement organizations (SMO) 16:8–11
Social movements 3:202, 3:582–583, 3:711, 4:148, 5:74, 6:29, 8:123–129, 12:371–372, 13:542–547, 13:609–610, 15:19, 15:801–802, 16:8–12, 17:359, 17:361, 18:424, 21:266, 22:32, 22:437, 22:447–453, 22:509, 22:513–516, 23:4–5, 24:459–463
active participation and involvement in 21:177
affirmative action 1:264, 1:267
analysis of 22:437–439
contemporary conceptions of space 22:438
networks and 22:439
place and 22:438
scale and 22:439
spatial-analytic analysis 22:437–438
and anarchism 1:680, 1:683–684, 1:686
Castells on 3:229–230
civil disobedience 18:424–426
cognitive approach to 21:177
collective 4:270–272
as a collective form of social behavior 21:176
as communicative action 21:177
and community 4:388
conceptualization 16:8–9
conservatism as 1:604–606
contemporary 21:179–180
control over followers 15:21
definition 22:437
diffusion 6:414
and emotions 7:534, 7:538
environmental movements 7:763–767
focal analytic issues 16:10–11
future directions 22:435
geographic concepts 22:437–439
geographic strategies of 22:439–440
geographic tactics 22:439–440
geography of 22:437–441
for global justice 21:180
and group identity 4:175–179
historical perspectives 21:177–178
historical sociology 11:27
history of 22:442–446
abolition of slavery in Great Britain 22:442–443
from eighteenth to twentieth century 22:445–446
social character of movements 22:443
transnational 22:445
twentieth-century 22:444–445
woman's suffrage in United States 22:443–444
homophobia/heterosexism 10:793
identity 22:447, 22:449–452
identity movements 11:543–545
impact on development of political ideology 21:178
in India 20:374–375, 22:854
indigenous knowledge by 21:177
interaction in 22:451–452
in international development 8:61–62
mass media, role of 15:19–20
measuring success of 15:21–22
and media 15:20

media opportunity structure and 15:20
in Mexico 1:628–629
in Middle East 16:384–386
for mobilization of human and material resources 21:176
movement constraints in 15:20
movement opportunities in 15:20
new 22:449–450
new social movements 17:644, 17:646
occlusion of rationality in 22:447–448
and political agendas 1:364
political process approach 22:430
political sociology 18:472, 18:475–476
and political violence 22:430–436
political violence and 22:430
politics and 21:176–177
prominence of 15:21–22
protests against biotechnology and genetically modified organisms (GMOs) 21:180
and public opinion 19:562
publicity as validation 15:21
questions 16:10–11
and radical social work 19:871, 19:874
relations between science and 21:176
relative deprivation 6:216–217
resource mobilization theory 22:448–449
revolutions 20:645, 20:647–648
science and new 21:178–179
Slow Food movement 22:94–99
social change 22:447–449
social construction 22:450–451
social media, role of 15:19
social protest 22:507, 22:509
and social work 22:132–137
and social work practice 22:518–524
strategic dilemmas of 15:20–21
subject matters of 22:430
terrorism 24:240–241
turn to rationalism in 22:448
use of term 21:176–177
varieties of 16:9–10
working definition of 21:176
and youth movements 25:837
Social movements Melanesian *see* Cargo cults
Social multiplier effects, and fertility 9:9
Social museums 22:698
Social neglect, children 3:414–415
Social network algorithms 22:454
cliques 22:455–456
clustering coefficient 22:456–457
cores 22:455–456
graph representations 22:454–455
role equivalence 22:457–458
shortest-path centralities 22:458
triad census 22:456–457
Social network analysis (SNA) 3:378, 10:100, 14:418–419, 16:524–531, 19:698, 21:853–861, 22:421–422
application across disciplines 16:530
for applied social and behavioral sciences 22:461
assumption 22:461
background 22:461–463
concepts 22:461–463
measures 22:461–463
balance theory 16:525, 16:530–531
current applications 22:463–466
flow analysis 22:465–466
network-level analysis 22:465
node-level analysis 22:463–465
disciplinary trajectories
social anthropology 16:528–529
social psychology 16:525–527
sociology 16:527–528
and education research 22:461, 22:463
field theory 16:525
gestalt psychology 16:524–525
Granovetter on 16:529–530
group dynamics 16:524–525
Harvard University 16:529–530
Hawthorne studies 16:527–528
and health research 22:461, 22:463–464, 22:467
increasing theoretical leverage of 12:771–772
of interorganizational networks 12:622–623
interlocking directorates 12:623
principal methods 12:623
research subjects 12:623
social capital 12:623
Moreno on 16:524–525, 16:530–531
networks of two families 16:529f
origins of ideas 16:524–525
small worlds 16:527, 16:530, 16:530f
social physics 16:528–530
social structure 16:524–525, 16:527, 16:529–531
sociograms 16:526f
sociometry 16:524–525, 16:527–528, 16:530–531
structural balance 16:527f
for structures of relationships 22:461
Warner on 16:527–528
White on 16:529–530
Social network approaches, diffusion 6:415
Social network data, blockmodeling of 19:433
Social network measures 16:521
Social network models
dynamic models 22:471
models bridging analytic level 22:468
statistical models 22:468–472
Bernoulli graphs 22:468–469
dyadic structure in networks 22:468
extra-dyadic local structure in networks 22:470–471
notation 22:468
null models for networks 22:468
prospects 22:471
Social network sites (SNSs) 24:229, 17:214–215, 22:421–422, 22:424–425, 24:146
Social network theory 13:256
and organized crime 5:172–173
use in military combat 15:516
Social networking 18:668
location-based 15:633
psychiatric ethics 8:68–69
Social networking service (SNS) 21:858
Social networks 1:581, 4:323–324, 6:897, 7:492, 11:550–552, 15:411–413, 20:486–488, 21:154, 22:454–456, 22:458, 22:461, 22:473–477, 22:831, 23:368, 22:463f
and adolescent sexual behavior 1:128–129
algebra of 22:478–484
algorithms and software 22:454–460
analysis
concepts and techniques of 22:473–475
statistical models for 22:475
approach 21:853–854
bear subculture 13:890
in Caribbean 1:641–642
challenges 21:858–859
clustering and cohesion in 3:906–912
and collective efficacy 4:171–172
and collective memory 4:191
and community 4:386–387
community and 24:930
and consumption 16:555–560
cultural participation and tastes 16:556–557
definition and intellectual context 16:555
future directions 16:559
history 16:555–556
media usage and knowledge consumption 16:557
psychological consumption 16:557–559
contextual conditions for 9:436–438
cooperation embedding in 4:864–865
core–periphery patterns 4:911, 4:916
cultural production in *see* Cultural production in networks
and culture 5:557
beyond persons and groups 5:560–561
kinds of actors inhabiting 5:558–559
narrative elements 5:561
negative ties 5:561
as research specialty 5:557–558
directed network 22:461–462
disease recovery 22:735–736, 22:738
dyad 22:461–462, 22:462f
dynamic models of influences in 15:659–666
constraints of mathematical imagination 15:659–660
formalization of influence mechanism 15:662–664
illustration 15:664–665
influence networks 15:660–662
in economics 16:572–581
embodied orientations 15:660
evolution and network dynamics 22:475
and fertility 9:6–9
demographic transition 9:6, 9:8–9
diffusion models 9:6–8
empirical evidence 9:7–8
family planning programs 9:6–9
implications 9:8–9, 9:8f
multiple equilibria 9:9
peers 9:6

Social networks (*continued*)
populations with dense *versus* sparse networks 9:9
social influence 9:7–8
social interactions 9:6–9
social learning 9:7–8
social multiplier effects 9:9
of fertility 9:19
and fertility transition 9:50–51, 9:53
formation of 16:574–575
coauthor model 16:575
connections model 16:575
static models 16:574
and health 22:365–370
assessment 22:367
biological pathways 22:369
conceptual models 22:366–367
direct effect model 22:366–367
historical and theoretical conceptualizations 22:365–366
loneliness 22:369–370
morbidity 22:367
mortality data 22:367–368
person-to-person contact 22:369
social capital 22:369
social participation and engagement 22:369
social relationship pathways 22:369
social support 22:366, 22:369
stress-related models 22:366
and health care 15:103
and language 13:286
and LGBT aging 13:864–866
and meaning *see* Meaning
method/theory 22:475–476
node 22:461–462
organizations 17:309
in political protests 18:422–423
prosocial behavior 1:581
rationality in 19:910–911
cooperation 19:910
strategic formation 19:910–911
relationships 22:461–462
and sexual behavior 21:700–704
dysfunction 21:701
extramarital sex 21:701
frequency 21:701
identification 21:700–701
initiation 21:700
intention 21:700
multiple sexual partners 21:703
online networks 21:702
orientation 21:700–701
partner selection 21:701–702
risky behavior 21:702–703
satisfaction 21:701
sex markets 21:701–702
use of drugs 21:703
shifting, intellectuals 12:279
and social capital 22:177, 22:181–184
and social media 16:782–785
social relations 22:473
social structure 22:473, 22:475
and social structure 22:717
in STI transmission 21:857–858
effective preventive programs 21:857–858
information provision 21:857
normative pressure *versus* social support 21:857
online networks as socially (un) embedded places 21:858
structural characteristics of 21:854–856
across-time network property 21:856
dyadic network property 21:855
ego-level network property 21:854–855
triadic network property 21:855–856
structural equivalence 23:621
tie 22:461–462
transnational organized crime 24:598–599
triad 22:461–462, 22:462f
undirected network 22:461–462
and unintended consequences 17:420
Web 2.0 4:494–495
Social Networks 21:323
Social neuroscience 2:369–372, 13:900, 22:485–486
as an interdisciplinary scientific field 22:485
doctrine of multilevel analyses 22:486
of empathy 13:463
prospects 22:487
techniques used in 22:486–487
of twenty-first century 22:485–488
Social norms
and cash transfers 3:219–220
cooperation 4:864
and ethnocentrism/prejudice 8:172–173
and ethos of science 21:314
and hazing 10:560
and interests 12:391–394
legal philosophy 18:24–26, 18:29
of prosociality 19:241–242
intergroup relations 12:409–410
and unemployment 10:517
Social observation 18:40
Social ontogenesis 6:841–842
ceremony of naming 6:841–842
social apprenticing 6:842
social priming 6:841–842
Social ontology 15:312–313
collective intentionality 22:491
and constructivism 22:232–233
critical realism 22:490–491
defined 22:489
emergence 22:491
naturalism 22:491
philosophical aspects 22:500–505
cases of elimination 22:501
cases of replacement 22:501
commitments 22:501
existence 22:500, 22:502–504
in France 22:503
holism 22:502, 22:504
identity 22:500–501, 22:503–504
individualism 22:502, 22:504
issues 22:501–504
multiple realizability 22:504
reduction 22:501, 22:503
reference 22:500, 22:502–503
supervenience 22:502, 22:504
philosophical ontology 22:490–491
philosophy of social science 22:489, 22:491
realism 22:489–491
scientific ontology 22:489–490
social theory 22:489–490
transformational model of social activity 22:491
Social opportunities, and social mobility 15:640
Social oppression, fighting against roots and effects of 14:35–37
Social orders 2:751, 4:212–213, 13:901–902, 24:138, 25:677
African tricksters 24:652–654
alternative 4:266
and collective memory 4:203
concepts of
emergence 12:386
and consensus 9:539
constitution of 25:676–678
labor, class, and nation 25:677
labor as status 25:676–677
transformations in work and reconfiguration of social sciences 25:677–678
of early modern Europe 18:916
Garfinkel on 9:598–602
and individual-collective relationship 11:817
and interests 12:391–392
rational choice explanation 19:900
and religion 20:307–310
science of 22:554
social control 4:822–824
and social control 22:235–236
and social integration 12:244–245
spirit possession 23:283
and taboo 24:1–2
Social organism 22:716
Social organization 4:323
and apprenticeship 1:866–867
of archaic states 23:373–374
in Arctic 1:928–929
of Australian aboriginals 2:234
diffusion 6:411–412, 6:414
family mode of 8:766
forms, and kinship terms 13:63
and generation 9:853, 9:855
of language 22:952
of legal profession 13:628
contemporary changes 13:630–631
context of work 13:631–632
educational training 13:629
gender and racial minorities 13:632
globalization and 13:631
jurisdictional disputes and professional monopoly 13:633
licensing 13:629
mobility routes within profession 13:630
practice settings 13:629
private practice of lawyers 13:629–630
professional ethics and regulations 13:632–633
professional roles 13:628–629
titles 13:628–629

of the scientific community 21:262–263
disciplines and specialities 21:263
norms of science 21:262
scientific role 21:263
scientists' autonomy and cognitive authority 21:262
stratification and scientific careers 21:262–263
underrepresentation of women 21:262–263
and social question 22:553
and social structure 22:719
of socialist societies 22:849–851
of tradition 24:505
tribes 24:646
Social origins
and intellectuals 12:279
of judges 12:882
Moore's contributions 15:770
and social mobility 15:640–642, 15:644
Social participation 4:402
and health 22:369
rheumatoid arthritis 20:657–658
Social pathologies, in underclass 24:729
Social patterning of cultural consumption and social class 4:759–760
Social peace processes 17:648, 17:650–651
Social pedagogy
and community development 4:382–383
social work education 22:775–776
Social pentagram 20:224–226, 20:225f
Social perception 2:202–204, 22:373
of conflict 4:583
negotiation 16:425–426
microethical judgments 16:426
trust and trustworthiness 16:425–426
and social cognition 22:204–208
Social performance 1:523, 1:526–527
Social person, and criminology 5:244–245, 5:247–249
Social phenomena
Prisoner's Dilemma 18:931–933
and religion 20:344, 20:346–347
Social philosophy 22:829–831
and alienation 1:540
Skinner's 22:29
Social phobias 22:493–499, 23:563, 25:797–798
antidepressants for 1:780
exposure therapies for 8:606
and fear-potentiated startle 8:864
Social physics 16:528–530
Social physiology 22:990
Social planning 14:498
Social play, gender differences in 9:754
Social policies 3:261, 8:800, 22:33–35, 25:512–513
and civil society organizations 22:577
and Eastern European social work 22:795–796, 22:799–800
and educational policy
combination in three regimes 25:482–483
educational systems, structures of 25:483
policy traditions in welfare states 25:481–482, 25:481t
relationship between 25:481–483
and health 22:279–280
pension 17:692–694
and prejudice/discrimination 18:832
and social work 22:807–809, 22:812
and volunteering 22:579
and welfare retrenchment 25:521–522, 25:524–526
Wilson on 25:587–588
and work-related stress 23:549
Social policy conflicts 10:597
Social policy issues 2:793
Social policy litigation against industry 3:736
Social position
and collective beliefs 4:164
and health 22:931–933
Social power 7:534–535
Social practices
ethnomethodology 21:192
and legal philosophy 18:24
and performance 1:866–869
sport as 23:301–302
and sustainability lifestyles 23:831–832
tradition as 24:503–504
Social pragmatic theories, of semantics 25:228
Social pragmatism 2:751
Social preferences 7:492
and behavioral economics 2:447
cooperation 4:862, 4:865
of eugenicists 8:237
prosocial behavior 1:580, 1:582
and rationality 19:908–910
Social presence theory
of computer-mediated communication 4:505
of online dating 17:212–213
Social pressure, conformity 4:1–2
Social prestige, and cultural inequality 22:331
Social priming 6:841–842
Social principles, for physical activity 18:93
Social problems 12:546
of alcoholic abuse 1:519
definition 12:542
domestic 12:542
drinking 6:658–659
globalization of 12:542
and health problems 10:572
societal impact assessment (SIA) 22:874
and United Nation 12:547
Social problems construction, and moral entrepreneur 15:798
Social processes 23:733
citizen mapping as 3:620–621
culture in 5:618–619
legitimacy as 13:823–824
negotiation 16:424–425
phases and interruptions 16:425
strategies and tactics 16:425
in self-regulation 21:554–555
terrorism 24:240–241
Social programs evaluation, and random assignment 19:884
Social proof, social influence 22:350
Social properties, philosophical aspects 22:500–505
Social protection 3:214–222, 7:573–574
conditional cash transfers 4:559
and health 10:198–201, 10:203
ICSW 12:498–500
Social Protection Floor (SPF) Initiative 10:146
Social protection policies 6:272, 6:275
Social protest 22:506–511
bread riots 22:509
collective action 22:507, 22:509, 22:511–514
classical approach to 22:512–513
digital approach to 22:513–514
importance of interdependent decisions 22:513
contemporary transnational 22:509–510
conventional and unconventional protest 22:508, 22:511
critical mass 22:513
defined 22:507
demonstrations 22:506, 22:508–510
evolution of digitally-enabled protests 22:514–515
first wave 22:514–515
second wave 22:515
Facebook 22:515
forms of 22:508–510
grievances 22:507, 22:509–510
historical transformation of the forms of 22:508–509
modular collective action 22:509
and new media 22:512–517
online networks and protest diffusion 22:515–516
political in 22:507
political protest 22:507
political protests 22:513–515
protest tactics 22:508
protesters 22:507
reasons for 22:507–508
repertoires of contention 22:508
research 22:516
and revolutions 22:510–511
social in 22:508
strikes 22:506, 22:508–509
threshold models 22:513
transnational environmental movement 22:510
Twitter 22:515–516
unconventional political channels 22:508, 22:511
Social psychological grievance theories 4:146
Social psychological perspectives
justice 12:949–955
of risk *see* Risk
Social psychological theory 22:525–531
American growth after World War II 22:528–529
attitudes 22:527–530

Social psychological theory (*continued*)
cognitive dissonance 22:528–530
crisis and rise of attribution theory 22:529–530
culture 22:525–530
discourse analysis 22:530
discursive psychology 22:530
experimental psychology 22:526, 22:528–530
field theory 22:528
groups 22:527–530
instincts 22:527–528
Lewin's 22:528–530
long past of 22:525
norms 22:525–526, 22:528, 22:530
origins of disciplines in nineteenth century 22:525–527
populations 22:525–527
psychoanalysis 22:527
symbolic interactionism 22:529
twentieth century 22:527
between world wars 22:527–528
Social psychology 6:20, 14:835, 16:11, 17:814, 19:414, 22:205, 22:532–543, 23:527, 23:531
affect control theory 22:539
agency 22:538–540
of aggression 1:373–380
collective memory 22:539
of collective memory 4:194–199
of conflict and conflict resolution 4:582–588
cross-cultural psychology 22:534–535
defining area of theoretical and empirical interest 22:532
domain 22:534–535, 22:535f
essentialism in 8:13–14
ethnic identity 8:113–114
evolutionary 8:460–462
experimental data 8:461
future directions 8:461
experimental laboratories 8:552
and field experiments 9:132
Fromm on 9:458–462
Goffman on 10:275
group process 22:541–542
of group processes 10:422–427
historical roots and development of new scientific endeavor 22:532–534
of identification 11:526–527
identity theory 22:539
intergroup relations 22:533–534
and legitimacy 13:823
levels of analysis 22:534, 22:536
levels of analysis in 13:899–903
and liberalism 14:25
Mills on 15:534
and organizational conflict 4:639–640
philosophy for 10:787–788
political psychology 22:535
practices, ethics, and institutional recognition 22:535–536
in qualitative research methods 22:545
in quantitative research methods 22:546
racism 19:864–869
research methods 22:544–552
of rumors 20:788
and small group research 22:100, 22:102–105
social cognition 22:204–209, 22:534
and social identity 22:313–315, 22:317
social influence 22:532, 22:534
social network analysis in 16:525–527
social representations 22:533, 22:535
social structure and personality 22:539–541
sociobiology 22:542
sociological 22:538
symbolic interaction 22:538–539
symbolic interactionism 12:352
of values 25:41–46
and vocational interests 25:240–243
Social psychology journals 13:902
Social push model 9:829, 9:831
Social question 6:487–488, 8:246–247, 22:553–555, 22:836–837
autonomy of society 22:553–554
capitalist economy 22:554–555
democracy and socialism 22:554–555
experience of French Revolution 22:553–555
industrial transformation 22:554
labor 22:553–555
mass pauperism 22:553
of new liberalism 18:300–302
poverty 22:553–555
science of social order 22:554
social crisis 22:553
social integration 22:553, 22:555
social organization 22:553
social stability 22:553–554
solidarity 22:554–555
Social ranking 2:79
of chiefdoms 3:376, 3:378–379
Social rationality
cooperation 4:862, 4:865–866
and modernity 15:689–691
in postindustrial economies 25:56
Social reaction theory, of deviance 22:239
Social Readjustment Rating Scale (SRRS) 23:584
Social realism 6:700–701
Social reality
social comparison 22:210
sociological conceptualizations of 7:434
Social reasoning 16:927
Social reciprocity, and justice 12:930
Social recovery, disasters 10:550
Social referencing 12:3, 12:5
autism spectrum disorder 22:560
during infancy and early childhood 22:556–562
changes over development 22:557–558
classic studies in 22:556
cultural differences 22:560–561
developmental course of 22:556–558
earliest evidence of 22:557
effect of information type 22:559
future directions 22:561
individual differences 22:560
mood 22:558
in nonhumans 22:561
preference of infants 22:558–559
parenting 22:560
social cognition 22:560
social learning 22:558
Social Reflection Measure (SRM) 13:120
Social reform movements 22:698
expertize associated with 22:696
Social reforms 14:255–256
applied research 1:851–852
and trade unions 24:492
Social regulation 4:456–457, 11:812–813
and anomie 1:714–715
normative modes of 4:457
and suicide 23:683, 23:685–686
Social rehabilitation 10:716t
Social reification 20:204–205
Social reinforcement model of pain 3:557
Social rejection, and aggression 1:379
Social relations 14:675, 22:854–855, 22:857–859
and health differences 22:278
and language 14:200
like logical relations between propositions 10:786–787
with nature 14:675
and selves 21:449–452
and social networks 22:473
Social relations model, of person perception 17:740
Social relationships 22:461
in adulthood 22:563–569
in adulthood, close relationship 22:563
measurement approaches used to assess 22:367t
mental health 15:189
and mortality 15:853–855
organizational design and form 6:236
pathways influencing health 22:369
and politeness 18:326–330
and relational social work 20:221
and social influence 22:351–352
and successful aging 23:670
and youth physical activity motivation 15:916–919
Social releaser 14:368–369
Social representations 13:899–902, 14:825–826, 16:866, 20:443–449, 22:529, 22:533–534
anchoring 20:447
collective memories as 4:198
collective representations 20:443–444
common sense 20:443–444, 20:446–448
communication 20:444–446, 20:448
defining 20:443–444
dialogicality 20:444, 20:448
formation and transformation of 20:447
formation of 20:447
of history 4:195, 4:197
imagination 20:444, 20:446–447
as interactional epistemology 20:444–446
asymmetries and symmetries 20:445
Ego–Alter–Object 20:444
ethics 20:446

field 20:444–445
figurative schemes 20:446
Heterogeneity and Cognitive Polyphasia 20:445–446
intervention strategies 20:446–447
language 20:444–448
managing 16:862–863
encountering nonheterosexuality 16:863
sexuality behavior/identity 16:863
sources 16:862
mental representations 20:443
objectification 20:447
as phenomena and interventions 20:446–447
of risk 20:711, 20:714–715
and social psychology 22:533, 22:535
study approaches 20:448
study methods 20:448
themata 20:446–447
theory 16:862
Social representations theory (SRT) 4:198, 20:714–715
Social representativeness of judges 12:882
Social reproduction 13:885, 22:329
and feminist legal theory 8:922–923
and gender 9:792–793
of inequality 1:731
and sexuality 21:819–823
Social reproduction theory 18:721
Social research 22:269, 22:272–273
on gender 9:715–722
and discourse studies 6:512, 6:514
methodology, reference group in 20:74–75
participatory action research in 17:540–544
privacy/confidentiality in 18:942–946
surveys 23:747–751
Social Research 24:401
Social Research Council (SRC) of the United States 1:68, 22:597
Social research strategy 22:986–987
Social resilience 10:547
Social resources 2:867, 21:135
Social responsibility 3:16, 11:805, 20:18, 20:582–583
practitioner-scholar model 24:522
Social responsibility awareness 4:670–671
steps of method 4:670–671
Social responsibility theories 9:393
Social return on investment (SROI) 5:60–61
Social revolutions 20:622, 20:628, 25:404
causes of 22:32
Social rights 12:956, 25:476
and immigration 11:641–642, 11:644
Marshall on 14:626–627
studying changes in 25:523–524
Social robots 24:106
Social role theory, and sexism 21:694
Social roles 15:369–370
of archaeology 1:903
dramaturgy 6:646
in middle adulthood 15:371
structural equivalence 23:623
Social rules
in moral, political, and religious fields 4:3–5
norms, and conformity 4:5
politeness as 18:326
Social science academies
in East Asia and Pacific 22:637
Academia Sinica, Taiwan 22:639–640
Academy of the Social Sciences in Australia (ASSA) 22:638–639
Chinese Academy of Social Sciences (CASS) 22:639
East Asia Institute (EAI), Korea 22:642
East Asian Institute (EAI), Singapore 22:640–641
East-West Center (EWC), Hawaii 22:638
excellence 22:637–638
Hong Kong Institute of Asia–Pacific Studies (HKIASP) 22:640
Institute of Social Science (ISS), Japan 22:641–642
Institute of Southeast Asian Studies (ISEAS) 22:641
Universities Service Centre for China Studies (USCCS) 22:640
in European Union 22:648–654
Social Science and Medicine 19:311
Social science application 21:35–36
Social science approaches
geographic thought 10:72
to history 11:41
and historiography 11:41
Social science associations, in United States 22:661–662
Social Science Citation Index (SSCI) 3:202, 21:567–569
Social science degrees 22:680–682
Social science disciplines, ideology of 19:123, 19:125–126
Social Science Electronic Data Library (SSEDL) 5:693–694
Social science history 11:109, 19:695, 19:697
Social science infrastructure 22:604–605, 22:621
Africa 22:604–614
case studies 22:608–614
Democratic Republic of Congo 22:610–611
education 22:609
Ethiopia 22:611–614
funding 22:608–609
research 22:609–610
South Africa 22:608–610
Central and South America 22:621–630
degrees and postdoctoral courses 22:628–629
dissemination and communication in Latin America 22:628
e-courses and social sciences 22:629–630
historical background 22:621
colonialism 22:621
developments after independence 22:621
in individual countries 22:623–627
Argentina 22:627
Bolivia 22:625
Brazil 22:626–627
Chile 22:626
Colombia 22:625
Costa Rica 22:624
Dominican Republic 22:624–625
Ecuador 22:625
El Salvador 22:624
Guatemala 22:624
Honduras 22:624
Mexico 22:623
Nicaragua 22:623
Panama 22:624
Paraguay 22:625
Peru 22:625
Uruguay 22:625–626
Venezuela 22:624
institutionalization 22:621–622
across the region 22:621–622
chairs and councils 22:622
internationalization 22:622–623
Latin America
dissemination and communication in 22:628
distribution of centers and universities 22:628
North America 22:661–666
research 22:596–603
resourcing and funding 22:655–660
in SSA 22:605–608
education 22:606
funding 22:605–606
research 22:606–608
Social science inquiry 20:93
Social Science Japan Journal (*SSJJ*) 22:641–642
Social science learned societies/professional associations 22:674–678
functions 22:674
membership 22:674–675
system of associations 22:675–677
conferences 22:676–677
international cross-disciplinary associations 22:677
international disciplinary associations 22:676
journals 22:677
national disciplinary associations 22:675–676
sections 22:676
Social science method, for economic analysis 13:454–455
Social science methodology, on diffusion 6:408
Social Science Methods in Legal Education 13:516
Social science models, dominance of 11:39–40
Social science professions 22:679
Social science professions and professionalization 22:679–687
Social science research
ethical standards for 20:515
organizational and methodological aspects of 22:993
urban geography 24:804
contribution 24:804

Social Science Research Council (SSRC) 1:933–934, 5:831, 22:597, 22:658–659
Social science surveys 5:847–848
crime and victimization surveys 5:848
health-related surveys 5:848
labor-related surveys 5:847–848
national archives and statistical institutes 5:847
Social science theories 19:126
Social science turn 11:75–76
social Darwinism 11:75
Weber's historical sociology 11:75–76
Social sciences 2:131–132, 2:554, 4:337, 8:182, 9:825, 9:828–831, 11:5, 11:123–124, 13:199, 22:688, 22:828, 23:3, 23:200–201, 24:673
alienation in 1:540–541
alternative conceptions of human studies 22:692–693
critique 22:692–693
disciplines/ideologies 22:693
literature 22:692
Baltes's contributions to 2:351
and bioethics 2:614
Castells's contributions to 3:229
causal analysis 3:247–250
changing role in EU research funding 22:644–645
and civil society 22:580–581
classification in 3:805–810
and communications research 4:278, 4:281–282
community concept in 4:369
counterfactuals in
empty counterfactuals 3:253–254
Neyman–Rubin model 3:252–253
prospective causal studies 3:251–252
and counterinsurgencies 5:99–102
critical 10:787
and cultural history 5:422
of culture 5:618
databases in 5:833–834
barometer series 5:833–834
European Social Survey 5:834
European/World Values Surveys 5:834
GSS and ISSP 5:833
international comparisons of national election studies 5:833
reasons for abundance 5:833
discipline formation in 6:485–490
as disciplines 22:679
distrust of 20:919–920
diversity 22:682–683
gender 22:682
non-US recipients 22:682
race/ethnicity 22:682
doctoral degree 22:682–683
international comparisons in 22:683
educational research as 7:279–280
empirical social sciences 13:587–588
enrollment and degrees in 22:680–682
bachelor's degrees 22:680–681
doctoral degree completion 22:681–682
doctoral degrees 22:681
graduate enrollment 22:681
master's degrees 22:681
time-to-degree 22:681–682
environment, application of GIS/remote sensing to 10:67–68
expenditures in 9:550t
and family history 8:738
field experiments in 19:737–739
foundations 22:659–660
Alfred P. Sloan Foundation 22:659
Andrew W. Mellon Foundation 22:659
Bill & Melinda Gates Foundation 22:659–660
Ford Foundation 22:659
The Rockefeller Foundation 22:659
French Revolution 9:410–411
and frontiers 9:464
funding in North America 22:655–656
future 22:685–686
governmental resourcing in America 22:656–658
Centers for Disease Control and Prevention (CDC) 22:658
National Institutes of Health (NIH) 22:657–658
National Science Foundation (NSF) 22:656–657
Social Sciences and Humanities Research Council (SSHRC) 22:656
hermeneutics in 10:786, 10:788
and history 11:102–107
history of 1:949
and human interrelations 22:828
human studies 22:689–691
heuristically 22:690
institutionally 22:689–690
morally 22:690
positions 22:690–691
hybridization in 11:448–449
idea of 22:688–694
idealization in 11:517–518
impacts
of computers on 4:515–518
of Enlightenment on 7:664–668
of liberalism on 14:22–25
impediments to human studies 22:691–692
complexity 22:691–692
instability 22:691
interim conclusion 22:692
pre-interpretation 22:691
instinct in 12:177–179
interdisciplinarity 12:386
Jacobs on 12:778
labor force 22:683–685
employment sectors 22:684
employment status of social scientists 22:684
professional training 22:684
social science occupations 22:684
wages 22:684–685
and law, aspects of relationship 13:585–587
and legal positivism 13:765–766
and life course 14:76, 14:79
mechanisms and explanations in 14:873–874
and media 22:583–587
Middle East politics 16:371–375
and modernity 15:692
modernization theory (after 1945) in 15:701–702
and natural sciences, relationship between 1:496–497
new doctorate recipients, current patterns for 22:683
and nongovernmental organizations 16:856, 16:860
philosophy of 22:489, 22:491
and postal systems 18:642
postmodern 18:680–681
as professions 22:679
public health as 19:545–551
quantification 19:700
reasoning 19:591
reconfiguration of 25:677–678
reference works publishers 7:587–588
Ashgate (Surrey, UK and Vermont) 7:587–588
Cambridge University Press (Cambridge, UK and New York City) 7:587
Elsevier (Amsterdam and Oxford) 7:587
Oxford University Press (Oxford, UK and New York City) 7:587
Routledge (London and New York City) 7:587
Sage (California) 7:587
Sage (London) 7:587
Wiley-Blackwell (Oxford, UK and Massachusetts) 7:587
of risk 21:158
rival claims of explanation and understanding 10:787
role in American antidiscrimination legal practice 6:533–534
Rorty's contributions 20:761
Rorty's impact on 20:765
scientistic and humanistic approaches 22:994–995
selfhood in 21:445–446
on sexuality, Maghreb 21:751–752
and Simmel, Georg 21:963–965
Social Science Research Council (SSRC) 22:658–659
study, history of education as field of 7:175–177
taxonomic distinctions 22:688–689
focus 22:688–689
modality 22:688
objectivity 22:689
scope 22:689
transition from political philosophy to 18:482
uncertainty in 24:718–720
visual methods in 25:181–183
Wallerstein's contributions 25:343
on water and health 25:433–434
Western views of 25:541–550
women and minorities labor force 22:685
gender 22:685

race/ethnicity 22:685
Wright's contribution 25:759–760
Social Sciences and Humanities Research Council (SSHRC) 22:655–656
Social sciences disciplines 6:492
Social scientific approach to pilgrimage 18:146–148
Social scientific theories 16:172
Social scientific thinking 24:40–42
Social scientists 9:420, 13:454, 22:684
as experts/public intellectuals 22:695–700
in predisciplinary period 22:696–698
transitions to disciplinarity 22:698–700
typology 22:695–696
Social security 4:338–339, 25:677
and mortality crossover 15:848–849
and population aging 18:542–543
and retirement 20:602–606
social insurance and 22:363
and welfare 25:477–478
Social Security Act 1935 18:725, 18:734, 24:492
Social Security Administration 17:175
Social Security Administrative 5:698
Social security fraud 5:633, 5:636–637
Social security Number (SSN) 5:634
Social security wealth (SSW) 1:416
Social selection
and life course epidemiology 14:72
and marital quality 14:515
Social settings, cultural policy 5:463–464
Social shaping of technology 21:203
Social sharedness 8:445
Social sharing of emotion 7:487
emotional interpretation 7:487
empathy 7:487
social comparison 7:487
Social shifter 25:809
Social shyness 1:564
Social significance of money 15:735–736
Social Silicon Valleys 22:357
Social simulation 22:701–705
computational models 22:701–705
Social situation
of child development 25:332–333
methodological individualism 15:311
Social skills 7:485, 24:56–57, 22:375t
and child care 3:392–393
in competition/cooperation 4:859
as erotic capital 8:1
evaluation 22:377
measurement approaches 22:377
Social skills training, in psychosocial interventions for psychiatric disorders 22:382
Social slavery 22:50
Social solidarity 8:483, 13:902, 16:834
agrarian, West Africa 1:301
Social solutions 12:546
Social sorting 5:246, 23:728, 23:731
Social Sources Thesis 13:764
Social space 1:408, 5:6, 5:9
in multisited field studies 16:105–107
Social space compression 10:218
Social stability
in collectivism/individualism 4:207
and social question 22:553–554
Social stasis 18:783
Social statics 4:521–522
Social statistics, linking health statistics and 5:782
Social status
attainment 15:641–642
Becker on 2:402–403
and death 5:870
as erotic capital 8:2
labor as 25:676–677
and language 13:281–282, 13:285–286
and leisure 13:846–847
reference group 20:72–75
and sexism 21:694
Social status differentiation, urban 24:801
Social stratification 3:742, 3:755–758, 22:338–344, 22:706–712
assets and examples of advantaged and disadvantaged groups 22:707t
basic concepts 22:706–707
Bourdieu on 2:777–780
Collins on 4:212, 4:214–216
and cultural lifestyles 22:333
and education 5:375
and functional prerequisites 9:538
gender inequality 22:710–711
Goldthorpe on 10:280
and health 15:98–100, 22:276–277
age and life course 15:100
gender 15:99
race and ethnicity 15:99–100
socioeconomic status 15:98–99
in historiography 22:339–341
in Germany 22:340
in United States 22:340–341
inequality
in comparative perspective 22:708
consequences 22:711
functions and dysfunctions of 22:707–708
structure 22:708–709
interests and power 9:538
legitimacy, and consensus 9:538
Mills on 15:534
and poverty 18:753–754, 18:756–757
poverty and underclass 22:709
prestige 22:707, 22:709
racial and ethnic stratification 22:710
ruling class, elites, and upper class 22:709
social mobility 22:709–710
and social mobility 15:640–642, 15:644
Sorokin's work 23:24
status 22:707, 22:709, 22:711
and suicide 23:686
theories
conflict 9:538
theories of 22:338–339
work 22:707–711
Social stratification function of insurance 12:219
Social structure and social learning model (SSSL) 6:72, 22:413–414
Social structures 11:109, 13:728, 13:730–731, 15:940, 16:562, 22:714–715, 22:723, 23:437, 23:733
agency 22:724–725
Bourdieu on 2:778
building through communication 14:392–393
and class 22:714–715
classical analysis 22:720–722
Durkheim's 22:721
Marx's 22:721
Simmel's 22:722
Weber's 22:721
communities 22:720, 22:722
as concrete interaction patterns 22:716–717
as configurations of individuals 22:717
as context, diffusion 6:414
core–periphery patterns 4:916
and culture 5:602–603, 22:714
distinguished from cultural structure 22:724–725
division of labor 22:721–722
and education 7:172–173
and emotions 7:537
and fertility transition 9:49–50
and field theory 9:141
as forms of interaction 22:716
and function 22:713–714
groups 22:720–722, 22:724–725
history 22:719–726
individuals 22:719–722
kinship in 13:36, 13:39
language 22:719
and life course 14:79
mental health 15:189
networks 22:720, 22:724
and personality 22:539–541
prehistory of concepts 22:719–720
and primitive society 18:911–913
process reconceptualizations of 22:724–725
and punishment 19:657
as role 22:714
social categories 22:719
social cohesion 22:721
social forms 22:720
and social inequality 22:338–340, 22:343
social integration 22:721, 22:724–725
and social justice 22:397–399
social network analysis 16:524–525, 16:527, 16:529–531
and social networks 22:473, 22:475, 22:717
social organization 22:719
Standard American Sociology 22:723–724
Blau's attempted synopsis and program 22:724
conflict theory 22:723
social construction of reality 22:723–724
theoretical reflections 22:724
as stories 22:717
structuralism 22:716
structuralist approaches 22:722–723

Social structures (*continued*)
American structural–functional anthropology 22:722–723
British structural–functional anthropology 22:722
French structuralism 22:723
Parsons 22:723
systems theory 22:715–716
urban 24:801–802
use of term 22:719
Wilson on 25:585
Social studies education 22:727–734
civic education 22:727, 22:730–731
democratic education 22:730–731
future of 22:732
as science 22:727–729
economics education 22:729
geography education 22:728–729
history education 22:727–728
as social education 22:730–731
conflicts in citizenship education 22:730–731
international perspectives 22:731
teaching practices for 22:731–732
inquiry-based learning 22:731–732
socratic seminars 22:732
structured academic controversy 22:732
Social studies of finance 9:173, 9:175
Social support 1:544, 4:324, 8:392, 23:566–567, 25:733–734
caregiving 3:137
in collectivism/individualism 4:208
and community 4:387–388
and coping 4:887, 4:889–890
coronary heart disease 4:917–919
cultural meanings of 5:429–430
for diseases 22:735–740
dyadic coping 11:593
and elder care 7:335
gastrointestinal disorders 9:607
and group psychotherapy 10:431
and health 22:366, 22:369
within immigrant communities 15:409
and marital quality 14:514
and migration 15:417
in military/disaster psychiatry 15:496, 15:498
mortality data on 22:367–368
for older LGBT people 13:865
for physical activity 18:94
religion and 20:265
and sexual behavior 21:700–701
and socialization of children 22:870
and trust 1:548–549
Social support networks 20:271
mental health 15:189
Social support theory 6:71
Social surplus 19:112
and cost–benefit analysis 5:50–51
Social survey movement 7:558, 7:560–561, 7:563
Social sustainability 23:798
Social systems 7:434–435, 9:532, 22:914, 23:198, 24:105
autopoiesis 9:541
civil system and 22:750–751
cultural system and 22:749–750
culturally diverse, implications for 24:105–107
defining properties of 22:741–743
diffusion of innovation in 8:380
economic system and 22:746–748
and environment 9:541
historical and empirical, classification of 22:744–745
levels of 9:541
loosely structured, in Southeast Asia 23:72
Luhmann on 14:383–386, 14:390–395
action and experience 14:384
communication and action 14:384–385
communication elements as events and operations 14:385–386
communication theory and double contingency 14:385
function and causality 14:383
meaning 14:384
operation and observation 14:386
psychic systems 14:384
self-reference, autopoiesis, operational closure 14:386
system and environment 14:383–384
political system and 22:748–749
school as 21:55–58
structural elements 22:745–751
types 22:743–744
Social television 25:81
Social tensions, mid-nineteenth century 18:400
Social theories 2:357–358, 2:361, 5:665, 10:86–87, 10:239–240, 13:586, 15:700–702, 19:330–332, 22:489–490, 22:966–967, 23:82–88, 24:461–464, 25:96–97
advertising in 1:216–217
and age stratification 1:334–336
economics and generational equity 1:336
effects of stratum size 1:335
interstratum inequality 1:335–336
interstratum variation in intrastratum inequality 1:336
paradigm 1:334–335
Alexander's contributions 1:523–524
applied research 1:850–851
Burke on 2:924–926
collaborative 9:381–382
and Giddens' duality of structure 10:788
Mead's pragmatic 18:809
and positivism 18:627
uses of 15:304–305
Weber on 25:465–467
white-collar workers 25:566–567
Social Theory and Social Structure (STSS) 15:247
Social therapy, for schizophrenia 21:12
Social thinker 22:205–206
as a cognitive miser 22:205–206
as a consistency seeker 22:205–206
Social thinking 2:865–866
Social thought 4:337, 15:703
division of labor in 13:161–162
impact of social question on 22:553–555
utilitarian 24:980–985
Social ties 15:413
community 4:385–387
and family history 8:739, 8:741–743
Social time acceleration 10:218
Social times 24:406
in explaining health differences 22:278
Social tolerance 18:240–241
Social trajectories 14:231–234
Social transformation 13:550, 17:648–649
and modernization 15:708–712
processes, institutions of conflict regulation as central determinants for 4:624
and sustainability lifestyles 23:832
Social trends 9:859
transportation 24:624, 24:626
Social trust, and political trust 24:661–662
Social typing 13:152–153
Social understanding 22:373
Social unrest, and global development 10:153
Social upheaval 20:858
Social utopias 14:497
Social values
of character 15:775
and energy 7:618
insurance 12:210
Social vulnerability 7:741–742
Social war 1:966, 4:618
Social waste 16:231
Social Web 22:831
Social welfare 16:231
and inequality 11:904–906
International Council on Social Welfare 12:495–500
Social Welfare Action Alliance (SWAA) 17:250–251
Social welfare function (SWF) 11:904–905, 25:500–501, 25:508–510
Social welfare management 18:725–726
Social welfare policy, use of labor statistics 23:426
Social welfare programs 18:733
Social welfare regimes, educational policy 25:482
Social welfare under transition from socialism 22:807–808
Social well-being ratings in ATUS 2010 24:398–401
Social withdrawal (SW) 2:456–462, 19:290
biological correlates of 2:458
development of 2:457–458
emotional and cognitive outcomes of 2:460–461
internalizing problems 2:460
self-concept and coping 2:460
social cognition 2:460–461
measures of 2:457
in social contexts
cultural influence of 2:458
in parent–child relationships 2:458–459
in peer relationships 2:459–460
subtypes 2:456–457

Social work 1:95, 1:785, 6:479, 9:801–802, 10:264–266, 16:458, 22:386, 22:389–390
Afrocentricity 1:305–309
anti-Arab racism 17:429–431
antioppressive 5:287–288
in antisemitism 1:784–785
approaches to fathering 8:845–847
armed conflict and 1:970–975
authentization of 16:459
in Canada 8:152–156
and child protection 3:425–428
children and families in 3:471–476
adoption 3:474–475
alternative care 3:474
child abuse 3:473
children in care 3:474–475
community-based interventions 3:472–473
family-based initiatives 3:473
forms of practice 3:473–475
future directions 3:475–476
permanence 3:474–475
professional judgment 3:475
protection of children 3:473–474
state 3:471–476
structural and political interventions 3:471–472
supporting families 3:473
underpinning levels and approaches to 3:471
in China 22:789–794
and choice *see* Choice
in climate change 6:479
colonial 16:458
commitment to social justice 22:133–134
communication skills in 4:284–290
contributions to oppression 17:249–250
critical *see* Critical social work
culturally appropriate interventions in 5:516–522
and disablism 6:448–449
and practice 6:449
in disaster recovery 6:480
and disasters 22:753–760
advocacy 22:757
assessments 22:757
casualty support 22:757
community engagement 22:757–758
crisis intervention 22:757
cultural implications 22:758
human rights concern 22:756
observations 22:758–759
outreach 22:757
populations at risk 22:756
psychological first aid (PFA) 22:756–757
psychosocial interventions 22:756
social inclusion 22:756
staff and peer support 22:758
in earthquake relief 6:480
in Eastern Europe 22:795–800
ecological *see* Ecological social work
efforts to advance antiracism in, Canada 8:153
and emergency relief, Australia 7:427–432
feminist research in 8:935–941
field, contribution of Addams to 1:95
in geophysical hazards 6:480
global agenda for 10:144–149
The Global Agenda for Social Work and Social Development (The Agenda) 10:144–149
green 10:385–391
health social work 10:716–724
human rights and 11:353–359
and immigration 11:641–642
imperial history of 16:457–458
international *see* International social work
international definition 1:785
international social work student exchanges 12:562–567
internationalization of 12:557
and interview 12:671–672
and islamaphobia 17:429–431
Islamic 12:754–758
and law 13:720
curriculum 13:723
governments accountability 13:721–722
legal provisions implementation 13:722–723
national social welfare legislation 13:722
regulating social work 13:723
transnational declarations 13:720–721
and learning disability 13:684
legal bases of 13:720–725
of liberation 16:459–461
liberation theology and 14:32–38
Maori's contribution 14:505–511
Marxist approaches to 14:663–669
and mental health 15:193–199
mental health, mentalism, and sanism in 15:200–203
migration and 15:416–421
narrative in 16:204–210
and national socialism 1:785
and neocolonialism 16:456–461
by numbers 12:504–505
with older people 7:335–336
policy context 7:336
population aging 7:335–336
in organizations 22:801–806
in Britain 22:802
definition 22:802
managing organization 22:804–805
nature of 22:802–803
organizational change 22:803–804
structure development for 22:801–802
orientalism in 17:428
politics of recognition 20:44–49
posttraumatic stress in 18:707–713
practice, ASB implications for 1:790–795
practice, engaged in 3:520, 3:524
in preventing human rights violations 6:481
and promotion of policies 16:460
psychiatric *see* Psychiatric social work
and psychodynamic thought 22:761–768
psychopathology and treatment 22:761
with racialized groups 22:821–826
radical *see* Radical social work
reducing disablism within 6:449–450
anti-oppressive practice 6:450
awareness among social workers 6:449–450
and refugees 22:769–774
relational *see* Relational social work
relevance of disability studies to 6:453
research and theory building 20:502–507
analytic induction 20:504–506
applications to practice research 20:504
critical realism 20:502–503
deductive qualitative analysis 20:504–505
definitions 20:502–504
fieldwork and Chicago tradition 20:504
grounded theory 20:504
qualitative research 20:502, 20:504–506
response to disasters 7:427–428
responsibility in 20:581–586
risk and uncertainty 20:686–691
risk assessment in 20:692–697
rural 20:807–813
in Russia 22:807–812
and sex offenders 21:676–677
and sexuality 21:802–807
single shared assessments in 12:501–506
social enquiry and action research for 22:269–274
and spirituality 23:288–289
structural 5:288
sustainable development and 23:817–822
teaching orientalism in 17:432–433
and trade unions, in Canada 24:491–496
and western gaze 17:431–432
whiteness/white supremacy and 25:569–574
Social work education 22:775–781
academic learning 22:777–778
anti-Black racism in, Canada 8:154–155
in Canada
historical perspective of 8:152–153
community work 22:775–776
continuing professional development 22:779–780
in Eastern Europe 22:796
economic development 22:776
global standards for curricula 22:776–777
groupwork 22:775–776
higher education 22:775–776, 22:778
impact of Bologna Process on 2:742–744
International Association of Schools of Social Work (IASSW) 12:474–479
internationalist and hegemonic view 22:776–777
life and practice experience 22:777
lifelong learning, critical reflection and the learning organization 22:779–780
macro practice 22:775–776
models 22:775–776
practice and field education and supervision 22:778–779
social development 22:776–777
social pedagogy 22:775–776
and social work identity 22:775
structures, history of 22:776
transformations in identity 22:777

Social work education (*continued*)
university and social agency contributions 22:775
Social work ethics 2:109, 2:111–112, 22:782–788
challenges 22:784–785
code of ethics 22:782–783
decision making 22:785
deontological ethics 22:783–784
ethical dilemmas 22:784–786
ethical issues 22:784–786
ethical problems 22:783, 22:785
ethics work 22:787
imperialism in 22:786–787
professional ethics 22:782
from professional ethics to ethics in professional life 22:786
and racialized groups 22:823
regulation of conduct 22:785–786
themes in 22:783t
theories 22:783–784
character- and relationship-based approaches 22:784
narrative- and case-based approaches 22:784
principle-based ethics 22:783–784
utilitarianism 22:783–784
values 22:782
virtue ethics 22:784
Social work identity
and social work education 22:775
transformations in 22:777
Social work practice 13:720
and alcohol misuse 1:521
anti-Black racism in, Canada 8:154–155
criminal justice and 5:208–213
critical *see* Critical social work practice
critical reflexivity and 5:281–286
disability 6:456
anti-oppressive practice 6:456
gendered care and 9:800–805
harmony and 10:529–535
implications, in Africa 23:602
masculinity and 14:679–684
minority ethnic elders 15:440–445
radical social work 19:873–874
in Roma 20:754
rural 20:809–810
improved model of 20:811
social movements and 22:518–524
activism 22:518–520
community work and social action 22:522–523
impact of neoliberalism 22:520–521
social change 22:518–523
social justice, human rights, and progressive politics 22:521–522
whiteness/white supremacy and 25:569–574
Social work process, and reflexive practice 20:85–86
Social work profession 12:475–476, 12:478
impact of The Agenda on 10:147
and oppression 17:248, 17:250
whiteness/white supremacy and 25:569–574
Social work regulation 13:723
Social work theories 9:801, 9:804, 22:813–820
application 22:819
cultural analysis 9:360–361
importance and use 22:813–814
relationship between perspective, theory, method, and model 22:814
social constructionism 22:817
types 22:814–819, 22:815t
behavioral, cognitive, and social learning theories 22:816
critical theories 22:818–819
developmental theories 22:814
humanistic theories 22:816–817
psychodynamic theories 22:814–816
social constructivist theories 22:817–818
systems theories 22:818
Social Worker Action Network (SWAN) 24:492
Social workers 2:112, 3:472, 3:474, 3:520–521, 3:524–525, 4:345
in child protection 3:425–427
disability awareness among 6:449–450
engagement with Roma service users 20:757–758
impact of armed conflict on 1:973
as professionals and union members 24:493–494
role in international social problems 12:547
Social Workers Registration Act 2003 14:505
Social working of legal rules 13:759
Social work-specific research 20:502–503
Social worlds 21:265
Strauss on 23:530
Social/behavioral research, ethical guidelines and codes of conduct in 8:37–42
future research 8:41
major American social and behavioral professional associations
American Anthropological Association 8:40
American Psychological Association 8:39
American Sociological Association 8:39–40
comparison of professional ethics 8:40–41
International Sociological Association 8:40
other academic and professional organizations 8:40
regulatory contexts
Australia 8:37
Canada 8:37
comparison of 8:38–39
United Kingdom 8:38
United States 8:38
responsible conduct of research 8:41
Social-cultural heterogeneity 8:158
Social-cultural psychology, in Soviet Union 20:828–829
Social-developmental milestones, and adolescent social competence 22:217, 22:220
Social-dominance orientation (SDO) 2:258–260, 2:264, 11:677, 12:412, 19:866, 22:250–253
collective action 22:252
construct validity of 22:250–251
gender 22:250, 22:252
inclusion 22:250–252
institutional discrimination 22:252
intergroup ideologies and intergroup behavior 22:252
low, psychology of 22:251–252
origins of 22:250
prejudice 22:250–251
versus RWA 2:264
system justification theory for 22:251
Social–ecological framework (SEF) 11:187–188, 11:190
Social–ecological integrative lens, honor killing 11:188–189
Social-ecological systems 22:260
education for sustainable development 7:736–737
framework 22:259
Social-emotional development
of children from birth to 36 months of age 22:942–948
anxiety 22:944, 22:946–947
criteria 22:944
depression 22:942, 22:944, 22:946
early intervention and income subsidy 22:946–947
home environment 22:947
importance 22:942–943
infant mental health 22:943, 22:947
internalizing behavior 22:944–947
methods 22:943–944
policy development 22:946–947
poverty 22:942–944
prosocial behavior 22:944–946
results 22:944–946, 22:945t
search strategy 22:944
theoretical underpinnings 22:943–944
withdrawn behaviors 22:946
definition 22:942–943
Social-emotional learning (SEL) programs 4:900
Social-emotional selectivity theory 21:425
Social-emotional world 21:872
Social-intuitionist theory 13:465
Socialism 2:716–717, 4:380, 4:684, 6:690, 7:308–309, 7:311, 13:199–200, 14:642–643, 14:650–651, 18:187–189, 18:398–399, 20:633, 20:838, 20:847, 21:141–144, 22:833–835, 22:839, 23:1, 23:3–4, 25:122
agrarian political economy transition to 1:459
Bell on 2:498–504
and capitalism 3:105–106, 3:108
city 3:606–607
and class consciousness 3:740, 3:742

cold war 22:842–843
and communism 22:837–838
consumption in 22:850, 22:852
and de-Westernization 25:547–549
dissenters and mass murder 22:841–842
and environment, Soviet Union 20:831–832
fertility trends in formerly socialist countries of Europe 9:99
forerunners and pioneers 22:836–838
historical aspects 22:836–843
legacies 3:609–610
and libertarianism 14:39–41
Mannheim on 14:497
Olson on 17:210
planning in 22:849–851
as political doctrine 14:644
and production 15:714, 15:716–717
and property 19:156–157
property in 22:849–851
real property, post-socialist transformations of 19:155–158
rival, and worldwide diplomacy 22:840–841
social Darwinism 5:688–689
and social democracy 22:242–244
Social Democratic Party 22:838–839
and social question 22:554–555
social welfare under transition from 22:807–808
social work under 14:667–668
in South America 23:48
and Soviet education 20:825–827, 20:829–830
in Soviet Union 20:820–821
Wallerstein's contributions 25:341
Socialism and Democracy 14:649
Socialism in one country, concept of 24:449–450
Socialist 18:187
Socialist calculation debate 19:536–537
Socialist city, model of 3:598f
Socialist collectivization 13:234
Socialist community work 4:379
Socialist intelligentsia 2:788
Socialist internationalism 22:834
Socialist land reforms 17:581–582
Socialist landmasters, in China 24:975
Socialist law 13:619, 22:844
international law perception and 22:847–848
justice administration and 22:847
stages in development of 22:844–845
structure of 22:845–846
substantive and procedural institutions and 22:846
Socialist market economy, in China 24:974
Socialist one-party state 22:143
Socialist realism, in Eastern Europe 6:807
Socialist reform, in socialist societies 22:851–852
Socialist Register 14:649
Socialist Revolutionary 20:856–857
Socialist societies 4:339, 20:843–846
anthropology of 22:849–853
clientelism 22:849–850
communist party 22:849, 22:851–852
consumption, in socialism 22:850, 22:852
eastern Europe 22:849, 22:852
economies of shortage 22:850
People's Republic of China 22:849–852
planning, in socialism 22:849–851
political economy and social organization 22:849–851
property, in socialism 22:849–851
research 22:849
social engineering and resistant personhood 22:851–852
soft budget constraints 22:850
Soviet Union 22:849, 22:851
and postsocialist societies, difference between 18:700
statehood 23:348–350
Socialist wage labor, Roma people 20:751
Sociality 13:38–41
in anthropology 22:854–860
actor-network theory 22:855–856
affect 22:856–859
ethical imagination 22:858–859
human interaction 22:856
human sociality character of 22:856–858
human–animal relations 22:858
intersubjectivity 22:856–857
relationality 22:855–859
research 22:858–859
scope 22:854–856
the social 22:854–856
subjectivity 22:854, 22:856, 22:858–859
costs to 18:899–900
and materiality 14:765
and personal naming 16:183, 16:185
queer 19:754–755
Socialization 2:264–265, 6:748–749, 7:257, 9:859, 15:82–83, 15:775, 17:533, 19:360–362, 21:42, 22:156, 22:159, 22:828, 22:861–865
in administration 1:106
adolescent prosocial behavior 19:222–223
agentic theory 22:863
and art 2:7–14
biological theories 22:862
behavior genetics 22:862
sociobiology and evolutionary psychology 22:862
of children 20:379
cognitive *see* Cognitive socialization
in collectivism/individualism 4:208
and cultural tastes 22:334–335
delinquency 6:69, 6:71–73
ecological systems approach 22:864
and education 7:144–148
and education, integrative framework for studying 22:864–865
characteristics of individuals to be socialized 22:864
goals and values 22:865
proximal processes 22:864
self-socialization 22:865
units of environmental influence 22:864
and educational gender gaps 9:684
family, in infancy and childhood 22:866–872
assessments of mother and father involvement in intact families 22:866–867
child characteristics 22:869
contemporary perspectives 22:866
ethnicity 22:870
impact of social change 22:870–871
influence of marital discord 22:869
parental instruction, advice giving, and consultation 22:868
parent–child interaction 22:867–868
parents as managers of children's opportunities 22:868
personal resources 22:869–870
sibling system 22:869
social interaction approach 22:868
social support 22:870
socioeconomic status 22:870
typological approach 22:867–868
and fear of crime 8:855
gender 9:743
genocide 10:18–19
infancy as time of 12:9
of law 13:438–439
limitations of behaviorist and social learning accounts of 15:775
and linguistic anthropology 14:170–171
norms 17:8–9
object relations theory 22:862–863
of offenders 5:210–211
and parenting 19:244–246, 19:248
perspectives in studying 22:861
political *see* Political socialization
psychological theories 22:862–863
attachment theory 22:862–863
epigenetic theories 22:863
learning theories 22:863
psychoanalytic theories 22:862
religious 20:379
in schools 21:56–57
and self-regulation 21:553–554
and senses 21:618–619
in sexual behavior 1:130
social cognitive learning theory 22:863
and social identity 22:313–314
sociological theories 22:863–864
conflict theories 22:864
structural functionalism 22:863
symbolic interactionism 22:863–864
of teachers 21:137
and temperament 24:180–183
and youth culture 25:809–810
Socialization explanations, of ethnocentrism and prejudice 8:172–173
Socialization goals 2:141
of mothers 17:496
Socialization theories 15:788
empirical research 15:789–791
Socialized health care programs 12:219
Socialized labor 25:676
Socializing effect 7:257
Socially constructed 10:628–629

Socially constructed regional geography 20:124
Socially deviant behavior 3:888
Socially disinterested children 2:456–457
Socially enforced conformity, sexuality 21:804
Socially Evaluated Cold Pressor Test (SECPT) 12:617
Socially imposed monogamy 14:616–617
Socially rational individuals
  cooperation among 4:865–866
    framing effects 4:865–866
    goals 4:865
Socially relevant outcomes of personality traits 17:923
Socially responsible investments 22:168–169
Socially responsive courts 4:809
Socially shared cognition, on coconstructivism 3:914
Socially shared retrieval-induced forgetting 4:191
Socially undesirably behavior, implications of shame and guilt for 21:476–477
Social-oriented play, differences, during infancy 18:220
Social-republican property models of landholding 5:253
Social-scientific history 11:104–105
Social-scientific knowledge 2:749
Social/socioeconomic inequalities 22:275
Social/software sorting 23:731–732
Social-structural regularities 11:812
Social-structural trends 18:477
Social-therapeutic prisons, offender rehabilitation 20:194
Sociedade Brasileira Contra a Escravidão (Brazilian Anti-Slavery Society) 22:41
Societal aggression 1:379–380
Societal and cultural difference 3:720
Societal approach, to vocational education 25:233–234
Societal aspects, of telephone 24:159–161
Societal breakdown, and social movements 22:447–448
Societal change, and moral development 15:782
Societal community, AGIL schema 1:388–389
Societal constitutionalism 12:269
Societal constitutionalization 13:556
Societal cooperation, business codes 8:21, 8:23
Societal corporatism 12:374
Societal costs of divorce 14:520
Societal cultures
  diversity of 16:23
  protection 16:25
Societal dynamics, and administration 1:107–108
Societal energetics 7:615–618
Societal environmental relations 7:798
Societal evolution 8:466
Societal impact assessment (SIA) 22:873–877
  approaches 22:874–875
  cognitive problems 22:874
  context 22:873–874
  focus 22:873–874
  future perspective 22:876
  methodologies 22:874–875
  privacy 22:874–875
  social problems 22:874
  surveillance impact assessment 22:875–876
Societal impacts of climate change 3:856–858
Societal interest 2:350
Societal mobilization 15:490–491
Societal multilingualism 22:951–952
Societal narratives 17:655
Societal organization, and slavery 22:53
Societal psychology, political psychology 18:428
Societal reaction theory 13:150–154
  disability 6:473
  and labeling theory 13:151–154
Societal reconciliation 17:648
Société d'anthropologie de Paris (SAP) 21:616
Société des observateurs de l'homme 2:89
Société d'Ethnologie Française 1:756
Société d'Histoire et d'Épistémologie des Sciences du Langage (S.H.E.S.L.) 14:218
Societies 4:337, 8:289, 13:585, 13:604–606, 20:627–629, 21:964–967, 22:878, 22:887–891
  advanced 13:72
  after World War II 22:828–829
  aging 18:543
  American studies 1:615–620
  Arendt on 1:962
  autonomy of 22:553–554
  based on status or contract 1:766
  benefits of skills training for 22:23
  and bioethics 13:557–558, 13:561
  and biomedicine 2:665–667
    biomedicalization 2:666
    patient/disease advocacy 2:666
    philanthropy, and the state 2:665–666
    and political economy 2:666–667
  and business 3:1–3
  cancer screening in 3:79–81, 3:80t
  in China 2:65–69
  and civilization 3:726–729
  and community 4:337–343
  and concept of individual 11:816
  conceptualization of 16:524
  in deeper-forces model 22:885–886
  definition of 22:828
  different, fashion in 8:835–836
  differentiation and secularization of 21:384–385
  diverse, application of Western sexual citizenship conceptions to 21:731–732
  in early social science 22:828
  in early twenty-first century 22:829–831
  Eastern European studies 6:816–821
  and education 7:172–175
  in eighteenth century 22:827
  elementary 13:70–73
  as emergent property 7:434–435
  and energy 7:615–620
  and environment 10:68
  ethnomethodology 8:203–204
  fishing 9:226–229
  genetics and social justice 9:969
  and global governance 10:156
  historical sociology 11:27–28
  horticultural 11:202–207
  hot/cold 1:748
  human, analysis of 22:265
  identity and intersubgroup relations 22:882–885
    heterogeneous preexisting subgroups 22:882–885
    homogeneous preexisting subgroups 22:882
  and illness 10:567–569
  impact of social change in 9:438
  and individualism 17:932–933
  information society 12:83–89
  irrigation 12:724–726
  and Islamism 12:743
  Japanese 24:583
  Korean 24:585–586
  and language 13:288–292, 14:170–173
  language in 22:951–952
  in late twentieth century 22:829–831
  and law 13:509–514, 13:535–536, 13:539–540
  and learning 13:699–700
  legalization of 13:608–609
  leisure 13:839–840
  Lévi-Strauss on 13:904–909
  liberal 14:27
  Luhmann on 14:386–388
    differentiation theory 14:387–388
    interaction and organization 14:386–387
    theory of sociocultural evolution 14:387
    theory of symbolically generalized media of communication 14:387
  Marxist understanding of 14:663–664
  mathematical psychology 14:805
  Middle East 16:372–373
  moral foundations of 2:748
  Near Middle East/North African studies 16:381–386
  new conception of 22:120
  in new unified theory of sociobehavioral forces 22:878–885
    elements 22:879
    overview 22:878–879
    rates of change 22:879–880, 22:880f
    subgroups 22:882
    subtypes of materialistic societies 22:881–882
    types of societies 22:881
  of Oceania 17:160
  philosophy of 21:343–345

Polanyi on 18:244–245
and political thought 18:481–482, 18:484
postindustrial *see* Postindustrial societies
postsocialist 18:697–701
and poverty 18:728–730
power in 18:765–769
prejudice in 18:834–845
and psychosocial development 7:943–944
rationality in 19:907–912
relevance of 22:829–831
respect for authority, and group violence 14:704
responses to gender gaps 9:685–686
gender bias approaches 9:685
gender-blind approaches 9:686
gender-differentiated approaches 9:685–686
role in life span development 14:102
role of leisure in 13:855
Russian and Post-Soviet studies 20:851–855
science of 21:965
self-esteem within 21:495–497
settler *see* Settler colonialism
and sexuality 21:819–823
with slaves 22:38
social differentiation 6:391–394
with social divisions 18:266
social inequality 22:338–344
social system *see* social system
sociological social psychology 22:538–543
sport history 23:307–308
South American 23:43–44
and state 23:343–346
Cold War 23:345
crisis of Staendische Gesellschaft 23:344
government 23:344
historical roots of distinction 23:343–344
jurisdictional political power 23:344
Keynesian compact 23:345
political authority 23:343
political prerequisites of civil society 23:344–345
public sphere 23:345
statism 23:345
totalitarian state 23:345–346
in twentieth century 23:345–346
unsettled distinction 23:343
structural embedding of mobile communication in 15:633
and territory 24:224–225
types of 22:878–886
usage 22:827
views of divorce 14:518
Society and law 13:475–477, 13:515
legal education 13:516
legal realists 13:516
1960s and early 1970s 13:516
relevance of themes and topics in discourse studies 6:510
studies 13:515
in United States 13:515–516
Society for American Archaeology (SAA) 1:886
Society for Applied Anthropology (SfAA) 3:6
Society for ecological economics 6:861–862
Society for Empirical Legal Studies (SELS) 7:554–556
Society for Ethical and Legal Philosophy (SELF), discussion group 25:347
Society for History of Technology (SHOT) 7:641
Society for Intercultural Education, Training, and Research 12:357
Society for Juridical Anthropology 5:239
Society for Mathematical Psychology (SMP) 14:805
Society for Psychological Anthropology (SPA) 19:363
Society for Research in Child Development (SRCD) 11:276–277
Society for schools, metaphor of 21:137–138
Society for Social Studies of Science (4S) 7:641
Society for the Amelioration 22:443
Society for the Study of Social Problems 22:675–676
Society for Universal Inquiry and Reform 2:149
Society for Visual Anthropology 25:182–183
Society for Worldwide Interbank Financial Telecommunications (SWIFT) 10:156–157
Society Islands 17:158–159
The Society of Brothers 2:147
Society of gazing 25:193
Society of Gerontology and Geriatrics in France 1:325
Society of Intercultural Education, Training and Research (SITEAR) 12:359–360
The Society of Lesbian and Gay Anthropologists (SOLGA) 19:753–754
Society of Multivariate Experimental Psychology (SMEP) 3:245
Society of surveillance 25:192
Society of the design 25:192
Society of the display 25:192
Society of the icons 25:192
Society of the spectacle 25:192
Society of the Woman in the Wilderness 2:146
Society of utilitarians 4:682
Society-centered theories of trade policy 24:486–487
Society–nature coevolution 22:254, 22:259
Society-related ideologies 11:564
Socio economic status, and arts education 1:234
Socio-behaviorist theory 22:528–529
Sociobiology 8:399–400, 8:404, 8:463, 22:542, 22:899–911
and anthropology 22:903
broader role for 22:905
collective effort 22:899–900
conflict about cooperation 22:905–906
controversy 22:901–903
centers for education and research 22:902–903
early attacks 22:901–902
political critique of bad science 22:902
Darwinism 5:683
and emotions 7:468–469
evolutionary psychology 22:904
inclusive fitness 22:900–901
kin selection 22:900–901
paradigm shift 22:899
realism *versus* adaptationism 22:904
scope and history of 22:908–909
social Darwinism 5:686, 5:690
of socialization and education 22:862
and sociology 22:892–898
adapted mind 22:893–894
current inquiry in evolutionary sociology 22:895–898
dominance and social inequality 22:897–898
evolution of social behavior by natural selection 22:892–893
evolutionary psychology 22:892, 22:895
gender and gender inequality 22:896–897
mating and reproduction 22:896
phylogeny of human sociality 22:894–895
sexual selection 22:892, 22:896
social capital 22:897
transforms 22:903–904
value-laden science 22:904–905
Sociobiology challenge to social sciences 22:994–995
Sociobiology controversy 22:901–903
Sociocentric network data 22:462
Socio-centric societies 5:622
Sociocentrism 14:826, 25:317–318
Sociocide 17:623
Sociocognition, skills in prosocial moral reasoning 7:317
Sociocognitive activity
learning as 13:671
Sociocognitive approach, ethnic identity 8:114
Sociocognitive conflict 13:671, 13:675
and coconstructivism 3:914
Sociocognitive development, and goal-directed behavior 10:268–270, 10:272
Socio-constructivist conception of learning 12:201–202
Sociocriminal approach, honor killing 11:185–186
Sociocultural acculturation 1:72
Sociocultural anthropology 1:772, 12:794
application to demographic studies of fertility 9:1–2
Sociocultural applications, applied behavioral analysis for 2:411
Sociocultural aspects
of creolization 5:117–121
of Melanesia 17:153–157
Sociocultural condition of modernity 11:865–866

Sociocultural contexts
and emotions 7:527–528, 7:530
of rape 19:888
Sociocultural correlates of values 25:28–29
Sociocultural dynamics 23:24–25
Vietnamese revolution 25:98
Sociocultural effects, of advertising 1:209–210
Sociocultural environment, and health 6:878
Sociocultural evolution
Luhmann's 14:387
Parsons&rsquo 9:539
Sociocultural factors
and aging 6:379, 6:382
of body image 2:736–737
Sociocultural inequality, and professionalization 19:95
Sociocultural integration, in Caribbean 1:640
Sociocultural learning, developmental neuroscience 16:603–604
Sociocultural milieu 7:168
in the family 7:168
Sociocultural ontology 5:472
Sociocultural order of nation-states, in Europe 8:276–277
Sociocultural orders, and horticulture 11:204–206
Sociocultural orientation 25:322–323
Sociocultural overview, Europe 8:273–279
Sociocultural patterns, professions 19:95
Sociocultural processes, popular culture 18:530–531
Sociocultural psychology 5:312
Sociocultural research 21:849
Sociocultural styles, professions 19:95
Sociocultural systems 22:914
and adolescent social competence 22:217
Sociocultural theory (SCT) 7:242, 8:731, 9:328, 9:330, 9:432, 21:363
of coconstructivism 3:913–915, 3:917
Vygotsky's 25:329–334
Sociocultural turn, qualitative developmental psychology 19:674
Sociocybernetics 22:912–917
actors 22:913–914
autopoiesis 22:914–915
circular causality 22:912–913
concept 22:912
control 22:913–914
cybernetics 22:912–913
defined 22:912
epistemological constructivism 22:915
methodology 22:915
organization 22:914–915
origin 22:912
self-organization 22:914–915
social science of complexity 22:915–916
steering 22:913–914
systems 22:912–914
Sociodemographic factors, relationship of health behaviors to 10:584
Sociodemography, and work-family conflict 8:762
Sociodigital technology, digital natives *see* Digital natives
Sociodramatic play 9:449
Socioecological model
for Alzheimer's disease 1:585–586
for intimate partner abuse 12:678
Socioecology 18:893
Socioeconomic changes, welfare state 25:522
Socioeconomic context, and parenting 17:494–495
Socioeconomic development
and division of labor 13:157–158
and globalization 10:218–223
Socioeconomic factors in fertility decline, in East Asia 9:57–58
Socioeconomic gaps, in Europe 8:276
Socioeconomic gradient 22:214–215
of health 22:275–277, 22:279
Socioeconomic metabolism
energetic part of 7:629–630
energy in 7:628
Socioeconomic mobility, life course epidemiology 14:72
Socio-Economic Panel Study (SOEP) 5:846
Socioeconomic position (SEP) 22:275
and health 22:276
and life course epidemiology 14:69, 14:72
Socioeconomic revolution 3:518
Socioeconomic sciences and humanities (SSH) 22:644, 22:671
Socioeconomic segregation 1:45–46
Socioeconomic status (SES) 5:354, 8:426, 11:654–656, 11:659, 21:42–46, 22:937, 24:33
and academic achievement 1:43
child health 22:924–925
cognitive development 22:926–927
country characteristics and student achievement 22:927–928
family characteristics 22:925–926
family investment model 22:925
family stress model 22:925
literacy achievement 22:927
neighborhood effects 22:926
numeracy achievement 22:927
parental education 22:925–926
parenting practices 22:925–926
poverty 22:924–926, 22:928–929
school environment 22:926
in analyses of covariance 1:43
behavioral medicine 2:465
Chapin's definition 1:43
of children from birth to 36 months of age 22:942–948
anxiety 22:944, 22:946–947
criteria 22:944
depression 22:942, 22:944, 22:946
early intervention and income subsidy 22:946–947
home environment 22:947
infant mental health 22:943, 22:947
internalizing behavior 22:944–947
methods 22:943–944
policy development 22:946–947
poverty 22:942–944
prosocial behavior 22:944–946
results 22:944–946, 22:945t
search strategy 22:944
theoretical underpinnings 22:943–944
withdrawn behaviors 22:946
and children's school learning 8:752
and cultural capital 5:378–379
definition 2:477, 22:943
definitions of 13:324
in demographic information 1:43
and divorce 6:606–607
and doctor–patient interaction 6:615–616
and education 7:132–135
in educational literature 1:43
effect on infant development 18:856
effects on academic performance 1:43–48
effects on childhood behavioral problems 2:477–480
epigenetics 2:478–479
externalizing behavior 2:477–478
family 2:477–480
genetics 2:479
importance of problems in children 2:478
importance of relation 2:477–478
internalizing behavior 2:477–478
mechanisms 2:478–479
parents 2:478–479
poverty 2:477–480
research, clinical activities, and policy implications 2:479–480
social class 2:477, 2:479
social class gradient 2:477
and fertility 9:21–22
and health 15:98–99, 19:814–815, 22:931–936
biological pathways 22:934–935
morbidity 22:931–932
mortality 22:931–935, 22:932f
political implications 22:935
racial differences 19:814
relationship with social position 22:931–933
social determinants of health 22:931, 22:935
social inequalities 22:933–934
and health care 22:937–941
and health care utilization 22:937–938
financial/affordability explanations 22:938–939
health care system explanations 22:939–940
knowledge, attitudinal, and cultural explanations 22:939
indicators of 1:43
of individuals 11:750
influence on language development 13:324–328
differences in communication 13:325
differences in grammar 13:325
differences in vocabulary 13:324–325
other SES-related differences 13:325–326
sources of differences 13:326–327

theoretical implications 13:327
in interaction 1:43
maternal education and 13:324
media use 15:36
and mortality 15:851–853, 15:871–872
causal pathways 15:852–854
and motivation in Australian Aboriginal students 15:903–904
and pathological gambling 17:593
and political participation 18:505
and poverty 18:744–745
and powerlessness 1:546–547
as predictor variable 1:43
and pretend play 18:871
in quasiexperimental studies 1:43
and socialization of children 22:870
and student achievement 5:344
and work-family conflict 8:763
Socioeconomic status model 25:306
Socioeconomic system 3:94–95
Socioeconomic variation, and value of children 3:500
Socioeconomically underprivileged ethnics 8:159
Socioeconomics, and obesity 17:83–84
Socioemotional abilities 8:669–670, 8:677–678
Socioemotional adjustment, and temperament 24:191, 24:194, 24:196–197
Socioemotional development
and child care 3:391–392
and conflict 4:594–599
Socioemotional leaders 23:430
Socioemotional selectivity theory 21:560
Socioemotional skills, and educational gender gaps 9:683–684
Socioemotional wealth 8:774
Socioenvironment systems 23:805–806
Sociogenesis 6:841
Sociogenetic law, Vygotsky on 25:313
Socio-genomic theory of personality 17:887, 17:887f
Sociogical perspectives, sex offenders 21:675–676
Sociograms 16:527, 22:473–474, 16:526f
Sociohistorical approach, worldview 25:754
Sociohistorical circumstances, attachment 2:136
Sociohistorical humanity 6:425
Sociohistorical inquiry 19:676, 19:679
Sociohistorical methodologies 19:676
Socio-historical perspective, of UK psychiatric social work 19:301–302
Sociohistorical transformations, and selfhood 21:447–448
Sociohistory of knowledge 7:176
Sociolegal environment, and reentry 20:64
Sociolegal perspective 13:604–605
constitutive dimensions 13:604–605
early themes 13:605–608
gap 13:605–606
legal order, distributive tilt of 13:606–607
legal order (bargaining and negotiation) 13:606
multiple sites of governance 13:607–608
procedural dimension 13:604–605
recent themes 13:608–610
globalization of law 13:608
legal consciousness 13:610
legal culture 13:610
legalization of society 13:608–609
new and developing themes 13:610
organizations, intersection of 13:609
rights consciousness 13:610
social movements and law 13:609–610
regulatory dimension 13:604–605
Sociolegal studies 13:475, 13:792
Socio-Legal Studies Association (SLSA) 13:515
Sociolinguistic competence 4:296, 4:299
Sociolinguistic dialectology 6:351–352
Sociolinguistic patterns, social class differentiation in 13:285, 13:285t
Sociolinguistic variationist studies 13:256
Sociolinguistics 3:455–456, 3:923, 6:505, 22:949–953
change, variation, community 22:949–950
code switching 22:951
creoles 22:952
diglossia 22:951–952
evolution of language 8:368
and functionalist grammar 9:474
interaction and audience 22:950–951
language and gender 22:949–950, 22:952–953
language attitudes 22:949–950
language policy 22:952–953
pidgins 22:952
society 22:951–952
language contact 22:952
social and political organization of language 22:952
societal multilingualism 22:951–952
for sociocultural aspects in North America 17:25
status and solidarity 22:950
Socio-linguistics 15:826–827
Sociologica 2007 10:283
Sociological alienation 1:553–554
Sociological analysis of science 22:229–230
Sociological approaches
to entrepreneurship 7:683
to everyday life 8:323–324
spirit possession 23:282–283
to suicide 23:683–684
values 25:33–34
Sociological conceptualizations of social reality 7:434
sociological constructivists 5:504
Sociological criminology 5:240
Sociological exchange theory, of time allocation 24:370–371
Sociological hazards 6:481
Sociological human ecology 11:284
Sociological imagination 15:537–538, 25:657–658, 25:660–661
and ethos of science 21:314–315
Sociological imperialism 14:826
Sociological implications on motor cortex 15:969
Sociological institutionalism 1:107, 11:562, 11:566–567, 12:182, 12:187–188, 16:764–765
in political science 11:562
Sociological jurisprudence 13:773
Sociological knowledge 13:516, 15:305–307
cognitive processes 15:305–306
Sociological laws, deterministic 6:241
Sociological models, feudalism 9:113
Sociological neoinstitutionalism 5:537, 17:418
Sociological practice 1:858–860
Sociological practitioners 1:858–859
Sociological semantics 15:249
Sociological social psychology 22:538–543
Sociological studies, of avant-gardes 2:322–323
Sociological system 22:747
Sociological systems theory 22:741, 22:746, 22:748–750
Sociological theories 11:165, 12:543, 15:247, 22:914, 22:954–964
Alexander's 1:523
Collins's 4:215
Coser's 5:37, 5:39
Eisenstadt's contribution 7:324–325
and empirical social research (ESR) 7:564
fundamental disagreement points 22:954
level of analysis 22:954
sociological theorizing as normative 22:954
sociology as science 22:954
and industrial societies 11:865
sex offending 21:674
of socialization and education 22:863–864
Sorokin's work 23:24
state of 22:962
theoretical perspectives 22:955–962
American-style critical theories 22:962
analytical conflict theories 22:956
comparative-historical conflict theories 22:957
conflict theorizing 22:956–957
convergence conflict theories 22:958
convergence exchange theories 22:958
critical theorizing 22:961–962
cultural theorizing 22:960–961
Darwinian revival 22:955–956
dramaturgical theories 22:958–959
ecological theorizing 22:956
exchange theorizing 22:957–958
exchange-network theories 22:958
Frankfurt school 22:961–962
functional theorizing 22:955
interactionist theorizing 22:958–959
Marxian conflict theories 22:956–957
microlevel conflict theories 22:957
phenomenological theories 22:959
postmodern critical theories 22:962
rational choice theories 22:957–958
role theories 22:959

Sociological theories (*continued*)
stage models revival 22:955
status theories 22:959
strong program 22:961
structural and structuralist theorizing 22:959–960
symbolic interactionist theories 22:958
weak program 22:960–961
Sociological traditions 22:965–973
contemporary 22:968–969
midcentury US 22:968
national subtraditions 22:969
Sociological Wing of Realism 14:277
Sociologism 19:392
Sociologists
classical 4:361–362
and Spencer, Herbert 23:274
and traditions 22:965–966, 22:968–969
Sociologists for Women in Society 22:675
Sociologization 5:577
of criminology 5:240
Sociology 4:137–138, 4:320, 4:521–522, 10:280, 15:576, 15:943, 16:764, 21:236, 22:679–680, 22:912
achievements 22:998–999
of action 23:527–530
Touraine's view 24:461, 24:463–464
Addams's contributions 1:94
of affirmative action 1:262–268
of aging 1:327
age structure and its implications 1:327–328
and life course 1:330–332
research strategies for 1:330
significance of cohorts 1:328–329
and well-being 1:329–330
alternative spirituality 20:360–361
analytical 1:668–673
anomie before 1:718–719
and anthropology 1:766
applied social research 1:851–853
apriorism 22:986
of art 20:706–710
of arts 5:728
in bachelor's degrees 22:680–681
biographical influences 23:274
biographical research 2:637–643
of biomedical sciences 2:663
Bourdieu's 2:778–782
of culture 2:782
of education 2:779, 2:782
of science 2:782
in Brazil 3:123
of bureaucracy 2:913–918
of chemical sciences 3:371–375
in China 2:68–69
of citizenship 3:655–661
of citizenship, Marshall 14:626–627
classical *see* Classical sociology
of collective behavior 4:157–162
of collective memory 4:200–205
of communitarianism 4:311–316
complex identity 22:999–1000
and Comte, Auguste 18:620, 18:623
of conflict 5:662–664
'conversion' and 'switching' of religion 20:363–364
of crime
challenges of change and continuity 5:192
early studies 5:189
history 5:189
issues of method, measurement, and levels of explanation 5:189–190
issues related to crime control 5:192
positivist view 5:189
social distribution and explanation of crime 5:190–192
social factors 5:189
crisis (from 1970) 22:993–995
critical sociology 22:996
of critique, Boltanski on 2:747, 2:751–753
cultural analysis 9:365–366
and cultural history 5:421–422
and culture of poverty 5:590, 5:592, 5:594
dataism 22:987
definitions, Weber's 10:786
of delinquency 6:68
development of 3:124
disciplinary boundaries 20:359
discipline 22:996–998
of divorce 6:606–609
doctoral degrees 22:680t
in economic analysis of law 13:458
of education 7:289–290
Collins on 4:211, 4:214
educational attainment 20:364
of emergence 7:434–435
of emotion 7:475–476
Collins on 4:214–216
and emotions 7:513–514, 7:516–517
of emotions 7:534–540
empirical legal studies in 7:554–556
empiricism 22:986
environmental 7:796–803
of epistemology 22:988
epistemology of 22:984–988
ethical questions in 8:55
ethnicity 20:363
of everyday life 21:147–148
evolutionary 8:463–468
explanations of ethical diversity 6:589–590
of family
Collins on 4:214–215
of fashion 2:357, 2:360–361
field observational research in 9:135–139
formation of academic discipline 22:991–992
for/of police 18:246
founders of sociology 20:359
French *see* French sociology
Freyre's contributions 9:421–422
fundamentalism 20:360
of gender
Collins on 4:214–216
gender in 9:792
gender roles 20:363
and genetics 9:984–989
Goldthorpe's contributions 10:280, 10:283–284
Habermas's contributions 10:469–474
history of 22:989
writing and teaching of 22:989
and human rights 11:377
of human sciences 11:384
idealism–materialism dilemma 22:984–985
and illness 10:567
individualism–holism–systemism trilemma 22:985
industrial 3:122
influence of religion on fertility 20:364
international migration 20:361
interpretation *versus* explanation 22:999
interpretivism 22:987
of knowledge 4:164, 22:2–3
antecedents of 13:111–113
Burke on 2:927
developments 13:115–117
Mannheim's contributions 14:497
and social causes of categories of thought 6:244
as specialty area 13:113–115
labeling theory 22:997
laicization 13:216–219
of law 13:808, 18:252
of leadership in organizations 13:664–669
of leisure 2:361, 13:852–857
levels of analysis 20:361
life course, influence of 20:363
of literature 9:141–142, 9:145–146
of love 4:214–216
Mannheim conception of 11:565
Marx conception of 11:565
master's degrees 22:680t
media and forms of communication 20:362–363
medicalization in 15:105, 15:107
of medicine 15:120
military 15:511–517
modernization theory 15:700–702, 15:705
of money 15:741–745
mysticism and religious experience 20:363
of narratives 16:218–222
national traditions 22:991
nationalism 20:363
of nationalism 16:263
nature of, as science 11:818–819
new religious movements (NRM) 20:360
objectives 22:996–997
and old age 17:179–180
orientations 22:997–998
originated in moral statistics 23:228
of persons 15:534
of police 18:251
of poverty 18:753–759
of presence 17:444
of prison life 18:920, 18:924
of professions 17:484
and psychiatry 22:974–978
and psychology, separation between 20:357
public policy and religion 20:364–365
qualitative styles of 10:280
of racism 19:857–863
rational choice model 22:999

relationship of symbolic interactionism to 12:353–354
of religion 20:378, 20:386–387, 21:390, 25:463, 25:465–466
religion, defining 20:359
religious affiliation 20:364
religious diversity 20:361
religious organizations 20:362
religious specialists 20:362
of revolutions 20:642–649
of rights, Marshall 14:628
ritual and religious activity 20:362
role of action in 1:75–76
of rural 20:814–819
of science 9:143, 15:248, 15:319–320, 15:322–323, 21:194, 21:322
scientific realism 22:986
of scientific revolution 21:318–321
scientism 22:987
sects 20:360
secularization 20:360
of secularization 21:389–390
of sex 4:214–216
and Simmel, Georg 21:967–968
skepticism 22:986
and small group research 22:104–105
Smith's contributions to 22:119, 22:123
and social inequality 22:338–339, 22:341–343
social network analysis in 16:527–528
social or evolutionary basis of religion 20:360
social research strategy 22:986–987
of social science 13:638
socialization 20:363
and sociobiology 22:892–898
and sociological theory 1:980–981
of solidarity 23:6–9
Spencer's contributions to 23:274, 23:276–277
of sport 23:313–317
stages in development 22:989–990
of state 18:252, 23:368–370
of Strong Program 23:592–597
subdisciplines proliferation 22:994
suicide 23:676
Durkheim's theory 23:676
of suicide 23:682–688
teaching of 3:122–124
of technology 24:109, 24:111
terminology 11:812
theory and methods 20:361–362
translation 24:583–587
Chinese development of sociology 24:584–585
concepts 24:584
east Asian sociology 24:583
society and sociology in Korea 24:585–586
Western sociology, Meiji Japan's engagement of 24:583–584
of translation 1:80–82
uncertainty in 24:719–720
use of term 22:991
of values 25:47–53
visual *see* Visual sociology
of voluntary associations 25:252–260
of voting 25:302–308
of war 25:405–413
Weberian perspective 22:994
of work 25:687–693
Sociology as social science (SSS) 10:280
Sociology of scientific knowledge (SSK) 8:576, 13:133–134, 13:137–138, 13:205, 17:12, 21:282–284, 24:135–136
attacks on 21:310
concepts 21:310
discourse analysis 21:309–310
Edinburgh School 21:309
Edinburgh–Bath approach 21:309
Fleck's work 21:309
intellectual and methodological paradigm 21:309
laboratory studies 21:309
Lakatos's work 21:308–309
origin 21:308–310
Popper–Kuhn debate 21:308–309
post-1970s 21:308
scientists' choices of research topics and social factors 21:310
significance 21:310–311
Strong Programme 21:309
theoretical resource in 21:309
Sociomaterial context, of sustainable practices 23:834–836
Socio-meter model 21:526
Sociometrics Corporation 1:916–917
Sociometry 16:524–525, 16:527–528, 16:530–531, 22:101, 22:101f
Socio-onomastics 17:225
education 17:225
personal names 17:225
Sociophysics 21:531
Sociopolitical approaches, for art 2:16
Sociopsychological explanations to crime–gender gap 5:132–133
Sociospatial ordering 13:884–885
Sociotechnical systems 2:304, 2:307, 21:200, 21:202–203
and sustainable production and consumption 23:839, 23:843
technological change, applied psychology of 24:104–108
technological determinism 24:111
Sociotechnical systems approach (STSA) 17:315
organizational ergonomics 11:303
Sociotechnical systems design 2:300
Sociotechnical systems theory (STS) 17:308, 25:720
Sociotropic economic evaluation, during elections 7:27
Socius 3:638–639, 22:827
SOCLA *see* Scientific Society of Agroecology (SOCLA)
Socrates 3:546, 7:255, 19:407
philosophy of 18:31–32
pre-Socratic philosophy 19:406
works
Phaedrus (Plato) 18:209, 18:211–212
Socratic approach to learning 12:696, 12:699–700
Socratic method 3:176
Socratic seminars, in social studies education 22:732
Socratic teaching methods 13:712
Socratic *versus* Confucian approaches to learning 12:699–700
SOCSIM model 8:728, 15:344–345
Sodalities, groups as 11:896
Söderblom, Nathan 7:375
Sodium amytal test 6:359
intracarotid sodium amytal testing 12:689–695
Sodomy 9:641, 9:729, 21:769–770
in China 21:736
in Latin America 13:858
laws 21:770
in South Africa 21:744
Sodomy laws 10:792–793
SODS *see* Self-organizing dynamical systems (SODS)
SOEs *see* State-owned enterprises (SOEs)
Sofer, Jehoeda 21:751
works, *Sexuality and Eroticism among Males in Moslem Societies* 21:750
Soft assembly 12:344
Soft budget constraints
in East Asia 6:791
in socialist societies 22:850
Soft collectivism, plausibility of 11:813–815
Soft constraints satisfaction 4:719, 4:722
Soft determinism 24:109
Soft embodiment 7:412
Soft heredity 8:227
Soft indicators 5:738
Soft law 13:917
Soft legislation 25:744
Soft money 22:607
Soft power 2:75, 5:597
center–periphery relationships 3:310
Soft-law regulators 13:619–620
Softsystems methodology 15:305
Software 12:93, 22:459–460
analysis 22:459–460
for Bayesian experimental design 8:524
core–periphery patterns 4:915–916
data 22:459
econometric 7:1–7
for ethnography 8:189–190
experimental laboratories 8:554
generalized linear mixed models 9:849
graphics 10:357
interaction 22:460
and internet, intellectual property law 12:269–270
social networks 22:454–460
structural contingency theory 23:613
visualization 22:460
Software applications for educational use 4:493–495
augmented reality 4:495
data logging and modeling 4:495

Software applications for educational use (*continued*)
drill and practice software 4:494
graphics software 4:493–494
multimedia cases 4:494
presentation software 4:493–494
simulations, haptics, games, and virtual worlds 4:494
spreadsheets 4:493–494
tutorials 4:494
Web 2.0 4:494–495
word processor 4:493–494
Software Developers Kit (SDK) 7:2–3
Software implementation, for nonparametric regression 16:876
Software programs 24:118
Software sorting 23:731–732
Software synthesis, reasoning for 5:939
Software verification, reasoning for 5:938
Soh, C. Sarah 2:74
Sohm, Rudolph 3:353–354
works, *Kirchenrecht* 2:269
SoHo (Manhattan) 5:599–600
SOI *see* Structure of intellect (SOI) model of intelligence
Soils 7:692–694
geoarchaeology, sampling 10:55
Sointula 2:150
Soka Gakkai 12:796, 16:806–807, 20:409
Soka Gakkai International 20:406
Sokoto Caliphate 1:301
Sokoto jihad 12:738
Sola fide 20:101
Sola gratia 20:101
Sola scriptura 20:101
Solanas, Fernando, works, *Hour of the Furnaces* 6:622–623
Solar city 6:831–832
see also Eco-cities
Solar Temple 16:807
Soldiers 15:504
and military anthropology 15:483–487
rights 16:306–307
students as 25:365–366
see also Child soldiers
SOLGA *see* The Society of Lesbian and Gay Anthropologists (SOLGA)
Solicitation 15:592
and sexual motivation 21:762
Solicitor general, judicial behavior 12:888
Solicitors Regulation Authority (SRA) 13:621
Solid waste, transportation of 23:847
Solid Waste Disposal Act 1965 7:369
Solidarism 23:1–5
Solidarity 4:154, 4:311–312, 4:314–315, 12:236–237, 12:239, 15:412, 20:763–764, 23:1–5, 23:369–370
affectual 12:403–405
from archaism to modernity 23:1–2
associational 12:403, 12:405
Bengtson's solidarity paradigm 12:403
collective *see* Collective solidarity
collective action and 23:6
and community studies 4:369–371
conflict and ambivalence 12:402–403
consensual 12:403–405
cooperation 4:862
functional 12:403, 12:405–406
grandparents, role of 12:407
and identity 11:526, 11:530
intergenerational 12:402
and language 22:950
macro-level analysis 12:406–407
mechanical 6:701–702
and mutuality 5:308
normative 12:403, 12:406
normative perspective 23:7
organic 6:701–702
principle of 19:583
prosocial behavior 1:579–580
and punishment 19:657
and radical social work 19:871–872, 19:874–875
rational choice perspective 23:7–8
and republic 23:2–4
social exchange 8:482–483, 8:486
and social question 22:554–555
and society 23:1
sociology of 23:6–9
structural 12:403, 12:405
structural perspective 23:7
theory of family 12:403–404
three-generation perspective 12:407
throughout world 23:4–5
and trade unions 24:491–494
and values 23:6–8
Solidarity (trade union) 25:838–839
Solidarity City Campaign 24:492–493
Solidarity-conflict paradigm 12:403
Solid-diaspora model 8:160
Solitary-active play 18:223
Solitary-passive play 18:223
Solitude 1:9, 2:460
Solly, Martin, works, *Xenophobe's Guide to the Italians* 16:237
Sologub, Fedor, works, *Petty Demon* 21:741
Solomon, Rebecca 25:625
Solomonic judgements 7:403
Solow, Robert M. 7:122
Solsona's Carnival 3:146
Solutions
multidimensional scaling 16:44
clustering of points 16:44
directions 16:44
regions 16:44
Solutions and values of games 4:868–870
accessibility 4:871
extensions 4:871
nonemptiness 4:870–871
selections 4:871
solution-specific structural issues 4:871–872
Solvability axioms 14:863
Solving dilemmas 22:248–249
leadership 22:248
sanctions 22:248–249
Solzhenitsyn, Alexander, works, *The Gulag Archipelago* 20:640
Somatic attribution, explanatory style 8:598
Somatic cell gene therapy 10:38
Somatic cell nuclear transfer (SCNT) 2:678, 10:34, 12:746–747, 12:749, 23:440, 23:443
Somatic difficulties, in work 25:534
Somatic gene therapy 2:616
Somatic marker hypothesis 12:617
Somatic revolution 7:415–416
Somatic sensations, explanatory style 8:598–599
Somatic symptom disorders, explanatory style 8:598–599
Somatization 19:486–488
Somatization disorder 23:10–12, 23:11t
risk model for 23:12f
Somatoform autonomic dysfunction 23:11t
Somatoform disorders 11:596, 11:599–601, 23:10–15
anxiety disorders 23:11–14
characteristics 23:11t
classification and epidemiology 23:10–11
cognitive-behavior therapy 23:13
depression 23:11, 23:13
diagnosis 23:12–13
DSM-V 23:14
etiology and course 23:11–12
hypochondriasis 23:11
hysteria 11:507–509
pain disorders 23:10–11
psychodynamic therapy 23:10
psychotherapy 23:12–14
risk model for 23:12f
stepped-care models 23:14
treatment 23:13–14
Somatomotor cortex (SMC), interoception 12:614–615
Somatoparaphrenia 21:420
Somatopsychic comorbidity, consultation–liaison psychiatry 14:4–5
Somatosensory cortices 16:620–621, 17:506–507
anatomy 17:506–507
deficits associated with damages 17:507–508
anosognosia 17:507
astereognosis 17:508
hemianesthesia 17:507–508
functions 17:507
interoception 12:614–615
Somatosensory evoked responses (SEPs) 3:327
Somatostatin
neuroendocrinology 16:651
Somatotherapy 23:16–21
deep brain stimulations 23:20
electroconvulsive therapy 23:16–20
electroshock therapy 23:18–19
phototherapy 23:20
psychosurgery 23:18
transcranial direct current stimulations 23:20
transcranial magnetic stimulation 23:19
repetitive 23:19
vagal nerve stimulation 23:19

Sombart, Werner 2:47, 3:21, 4:752
  works, *Why Is There No Socialism in the United States?* 1:598
Somers, Margaret
  on ideational embeddedness 5:536–537
  on social network analysis 5:558
Somnambulism 11:465, 11:472
Somos (we are) 13:860
Son preference 15:610–611
  and sex selective abortion 21:686–687
Son relation 13:54
SONAR (SOund Navigation And Ranging) 19:472–473
Song, Jesook 2:73
Song behavior 2:692
Song Sokha 2:72
Songbirds 2:691
Song-control nuclei 2:693
Songzhuang Association for the Promotion of Arts 5:600
Songzhuang Town (Beijing, China), cultural industries in 5:597–598
Sonic practice, music as 16:153
Sonno joi 16:276
Sony 9:157
SOPA *see* Stop Online Piracy Act (SOPA)
Sophism 18:208
Sophisticated epistemological beliefs 7:889
Sophists 1:765
SOPT *see* Self-ordered pointing task (SOPT)
Sorcery
  and fetishism 9:105, 9:107–108
  witchcraft 25:604, 25:608
Sorensen, Clark, on Koran villages 2:73
Sorenson Molecular Genealogy Foundation 1:762
SORKC scheme *see* Stimulus Organism Reaction Kontingency Consequence Scheme (SORKC scheme)
Sorokin, Pitirim Alexandrovich 2:271, 15:249–250, 16:284
  in America 23:23
  biography of 23:22–23
  impact 23:25–26
  in Russia 23:23
  on social genetics 23:23
  on social mechanics 23:23
  works 23:23–25
    Altruistic Love 23:25
    American Sex Revolution 23:25
    Contemporary Sociological Theories 23:22–24
    The Crisis of Our Age 23:25
    Golod kak faktor 23:23
    Man and Society in Calamity 23:25
    Power and Morality: Who Shall Guard the Guardians? 23:25
    Prestuplenie i kara, podvig i nagrada 23:22–23
    Russia and the United States 23:25
    Sistema sotsiologii (*System of Sociology*) 23:22–23
    Social and Cultural Dynamics 4:165–166, 23:22–25
    Social Mobility 23:22–24
    Society, Culture and Personality 23:25
    The Sociology of Revolution 23:22–24
    S.O.S. The Meaning of Our Crisis 23:25
    Sotsiologiya 23:26
    System of Sociology 23:23
    The Ways and Power of Love 23:25
Sororate 21:799
Sororities, hazing in 10:554–561
Sortal element 3:813–814
Sorting, in naturalism 16:329
Soteria model 24:278
Soteriology 2:886–887
Sotsiologicheskie issledovaniya 23:26
Soul
  and mind–body dualism 15:545–546, 15:548
  Plato on 18:211–212
Soule, Sarah, on theorization 17:392
Sound cinema 9:168
Sound correspondence 4:424, 11:822
Sound exchanges and word exchanges 23:231
SOund Navigation And Ranging *see* SONAR (SOund Navigation And Ranging)
Sounds
  functional view of 12:782–783
  segmentation of 25:764
Soundscapes 16:152–153
Source amnesia 6:719, 23:30–31
Source credibility
  and attitude 2:170
  social influence 22:351
Source monitoring 23:28–35
  amnesia 23:28, 23:30–31, 23:33
  assessment 23:29
  cognitive aging 23:28
  cognitive theories 23:28–29
  and confabulation 23:31–32
  developmental studies 23:33–34
    in Alzheimer's disease 23:34
    in children 23:33
    in healthy adult aging 23:33–34
  discrete-state models 6:527–528, 6:527f
  for emotional context 23:30
  errors 23:29
  event-related potentials 23:32–33
  false memories 8:713
  framework 23:28–29
    versus dual-process theories 23:30
  generation effect 23:29–30
  hippocampus 23:31–34
  and neuroimaging 23:32
    frontal lobes 23:32
  and patient studies 23:30–31
    frontal lobes 23:31
    medial temporal lobes 23:31
Source monitoring framework, of retrieval 20:611
Source rocks 17:165
Source separation, computational auditory scene analysis 2:223, 2:225–226
Sourcebook of Criminal Justice Statistics 5:718
Source-constrained retrieval 20:610
Source-filter model 23:260
Sources 21:38
  mesoamerican historiography 11:57
Sources' perspectives 4:772
Sourcing, supply chain management 23:712–713
Sousveillance 23:735
South Africa 14:281–282, 16:268, 22:608–610
  Apartheid labor policy 12:437–438
  arts education in 1:231
  education in 7:200
  influence of missionaries 1:289
  internal migration in 12:434
  legal traditions in 23:658–659
  post-apartheid, brokers in 2:872
  psychiatric care in 6:43–44
  racism in 19:849
  sexual minorities in 21:744–749
    current status 21:749
    equality 21:746–748
    history of lesbian and gay activism 21:744–746
South African Constitution of 1996 21:744
South African Development Communities (SADC) region 24:84
South African Foundation for Economic and Financial Education (SAFEFE) 24:84
South African management practices 11:786
South African Truth and Reconciliation Commission 18:271
South America
  archaeology of 23:36–37
    Amazon River 23:36, 23:39–40
    Andes 23:36–39
    biotic resources 23:37
    civilization 23:36
    ecological diversity 23:36–37
    geography of 23:36
    indigenous peoples 23:36–37
    monumental architecture 23:36
  climatic variability 3:868
  historiography and historical thought in 11:58
  native peoples, totemism 24:453
  origins of food production in 9:303
  social science infrastructure 22:621–630
  sociocultural aspects 23:43–49
    ALBA (Bolivarian Alliance for the Peoples of Our America) 23:48
    Amazon 23:43–44
    Andes 23:43–48
    blackness 23:45
    cultural continuities 23:46–47
    European conquest 23:44–45
    fusion of cultures 23:46
    human and topographic diversity 23:43
    indigeneity 23:48
    mestizaje, cultural practices, and ideologies of racial blending 23:45
    modernities and alternative modernities 23:47–48
    new socialisms, new internationalisms, and enduring struggles for diversity 23:48
    power 23:45–47

South America (*continued*)
pre-Columbian people 23:43–44
religion 23:46
rituals 23:46
southern cone 23:43–44, 23:47
trade 23:43–47
transformation from colonies to nation-states 23:46
UNASUR 23:48
South Asia 1:955–956
archaeology of 23:50–57
Asokan edict 23:55f
Buddhism 23:54–56
early historic period 23:54–56
Harappa 23:52–54, 23:52f
Indus civilization 23:52–54
Iron Age 23:54
Mauryan Empire 23:54–55
Medieval and Modern Periods 23:56
megaliths 23:54
Mesolithic 23:50–51
Mohenjo Daro 23:53f
Neolithic and Chalcolithic 23:51–52
Paleolithic 23:50
Sisupalgarh eastern rampart 23:55f
Vijayanagara 23:56, 23:56f
Vitthala temple 23:56f
area of 2:76
and communalism 4:264
dowries 6:642–645
education 7:194–197
in Afghanistan 7:195
in Bangladesh 7:195
in Bhutan 7:195
and child labor 7:197
and diglossia 7:195, 7:197
in India 7:195–196
in Maldives 7:196
in Nepal 7:196
in Pakistan 7:196–197
on South Asian Association for Regional Cooperation (SAARC) 7:194
in Sri Lanka 7:197
fertility transition 9:80–83
abortion 9:81–82
in Bangladesh 9:80–83
contraception 9:81–82
data sources 9:80
direct causes 9:81–82
family planning programs 9:80, 9:82–83
fertility trends 9:80–81
future prospects 9:83
gender inequality 9:80, 9:82
in India 9:80–83
marriage 9:80–82
in Nepal 9:80–83
in Pakistan 9:80–83
schooling 9:80, 9:82
socioeconomic, cultural, and policy influences 9:82–83
in Sri Lanka 9:80–83
total fertility rate 9:82, 9:81f
historiography and historical thought 11:78–81
Ancient India 11:78–79
colonial era 11:79–80
culture 11:78–80
dynasties 11:78–79
empire 11:78–79
Indo-Islamic period 11:79
nationalism 11:78–80
political history 11:79–80
Islam in 12:729
legal traditions 23:58–62
movements 2:80
nationalism in 16:279–282
colonialism 16:279–281
democracy 16:281–282
dominant strand in Indian nationalism 16:279–280
empire 16:279–280, 16:282
future 16:281–282
Gandhi and 16:280
globalization 16:282
imperialism 16:281
interpretations 16:281
Muslim nationalism and foundation of Pakistan 16:280–281
national identity 16:279
national liberation movements 16:280–281
pan-Indian scene in nineteenth century 16:279
in postcolonial South Asia 16:281
religion 16:280–281
origins of food production in 9:301
political liberalism in 18:388
sociocultural aspects of 2:76–80
accommodation of diversity 2:77
area and population 2:76
caste 2:78–79
civilization 2:76–78
collective identities 2:77
family, kinship, and marriage 2:79
hierarchical organization 2:77–78
media, movements, and modernity 2:80
new institutions and classes 2:79–80
religion and language 2:76–77
village 2:78
urban policies in 24:898
use of term 2:76
South Asian Association for Regional Cooperation (SAARC) 7:194
South Asians, international migration by 12:530–537
South Korea 2:72–75, 5:454, 24:922
cultural policies in 5:455–456
creative industries 5:457
flow of culture 5:458
gender gap 15:613
sex ratio at birth (SRB) 15:609
sex selection 15:609, 15:613
urban policies in 24:895
South Korean politics *see* Korean politics
South Pacific islands 17:154f
South Sea Islanders 11:516
South Sudan–Sudan border war 25:421
Southeast Asia
adult mortality in 1:171
archaeology of 23:63
Asian laws, Asian values 13:815–816
British and Dutch legal policy 13:814
colonial law 13:814–815
colonialism 13:815
customary law 13:812
development 13:815
European laws in 13:814
fertility transition in 9:84–91
future of law in 13:816
indigenous 13:812
Islam in 12:728
legal pluralism and 13:812–813
legal traditions of 13:812–817
nationalist history in 11:85
origins of food production in 9:302
postcolonial law in the period of nationalism 13:815
precolonial law 13:813–814
sociocultural aspects of 23:70–74
big man and theater states 23:71–72
conceptualization 23:70
diasporas and business networks 23:72
ethnic diversity 23:71
family and gender relations 23:73
globalization, urbanization, and locality 23:73
Indianization and Islamization 23:70
loosely structured social systems and involution 23:72
moral economy 23:71
plural societies 23:71
religious diversity and religious revivalism 23:70–71
social class and strategic group formation 23:72–73
values and knowledge 23:73
urban policies in 24:896–898
Southeast Asia Treaty Organization 1:561
Southeast Asian archaeology 23:63
complex polities 23:66–67
development 23:63–64
early hominin occupation 23:64–65
early states 23:67
farming communities 23:66
future research directions for 23:68
iron age 23:66–67
late Pleistocene/early Holocene hunter–gatherers 23:65
metal age 23:66
political concerns 23:67–68
theoretical tensions 23:67–68
Southeast Asian Cultural Collection (SEACC) 22:641
Southeast Asian languages 3:816
Southeast Asian Ministers of Education Organization Applied Research Centre for Archaeology and Fine Arts *see* SEAMEO ARCAFA (Southeast Asian Ministers of Education Organization Applied Research Centre for Archaeology and Fine Arts) in 1971
Southeast Asians, international migration by 12:530–537

Southeast Louisiana Flood Protection Authority 7:760
Southeastern Europe 8:275
Southern Africa, sociocultural aspects 23:75–81
  Angola 23:77–79
  Botswana 23:75, 23:77, 23:79–80
  colonialism and independence struggle 23:77–78
  history 23:75–77
  language and communication 23:80
  Lesotho 23:75, 23:77, 23:79
  livelihood 23:79–80
  Malawi 23:75, 23:77
  Mozambique 23:75–80
  states and politics 23:78–79
  Swaziland 23:75, 23:77, 23:79
  Zambia 23:75, 23:77, 23:79
  Zimbabwe 23:75, 23:78–80
Southern Baptists 21:174
Southern Christian Leadership Conference (SCLC) 3:693
Southern Cooks 17:158–159
Southern Demographic Association (SDA) 1:842
Southern Europe, carnival in 3:146
Southern Marxist archaeology 14:671–672
Southern Paiute 20:899–900
Southern Sociological Association 22:675–676
Southern theory 2:604–605, 18:653
  and global south 10:190
Southern urbanism 20:677–678
South–north migration 12:530–531, 12:533–534, 12:536
South–South cooperation 3:505
South–south migration 12:530–531
South–South remittance flows 15:746
Southwest Pacific islands 9:302
Southwestern Asia, origins of food production in 9:300–301
Southwestern United States, origins of food production in 9:303–305
SOV *see* Subject–object–verb (SOV)
Sovereign 13:762
Sovereign justice, struggle for 22:395
sovereign states 12:507–508
Sovereignist attitude 16:225
Sovereignty 2:527–528, 4:521, 4:705, 11:670, 17:516
  -compatible rationale for global public policy 10:183–184
  food 10:79
  and globalization 10:235–236
  and governmentality 10:308–312
  and immigration 11:669–670
  land tenure 13:231–233
  national 16:262–265
  and pluralism 18:232, 18:234–235
  Schmitt on 21:28–29
  of state 23:358–359
  and territory 24:223–224
Sovereignty issues, of Australian aboriginals 2:237
Soviet bloc 4:302–303
  in Czechoslovakia 4:304
  in European communist countries 4:305
Soviet collective 4:269
Soviet Communist dictatorship 20:856
Soviet deviance 20:853
Soviet ethnic 20:852–853
  and gender relations 20:852–853
Soviet journalism 16:822
Soviet Marxist archaeology 14:670
Soviet political system 4:304
Soviet politics 20:843–845
  collectivization 20:843
  fusion of polity and economy 20:844
  industrialization 20:843
  new economic policy (NEP) 20:843
  Soviet institutions 20:845
  structure of power 20:844
Soviet Russia, homosexuality in 21:742
Soviet studies
  education 20:825–830
    Communist Party of the Soviet Union 20:825–827
    crisis and ambiguous contemporary legacies 20:829
    Dewey on 20:825–826
    educational experiments and radical progressive education 20:826
    modernization 20:825–830
    polytechnical education 20:826, 20:828
    potential and limits of postwar reforms 20:828–829
    socialism 20:825–827, 20:829–830
    Sputnik on 20:827
    Stalinist traditionalism and system building 20:826–828
    Vygotskii on 20:826, 20:828–829
  environment 20:831–835
    activism 20:832–833
    authoritarianism 20:832, 20:834
    central planning 20:831
    conservation 20:831–833
    ecology groups 20:831–833
    glasnost 20:833–834
    legacies 20:834–835
    legal and organizational issues 20:832
    pollution 20:831–834
    protests 20:833
    Russia 20:832–835
    secrecy 20:833–834
    socialism 20:831–832
Soviet Union (USSR) 1:570, 5:367, 6:810, 6:812, 18:187, 21:368, 22:841–842
  anthropology of 3:317–318
  area studies in 1:949–951
    diversification during Cold War 1:950–951
    oriental studies in interwar period 1:949–950
  arms control 1:976–978
  authoritarianism and environment in 20:832, 20:834
  Bolshevism in 20:820
  bureaucratic obstructionism in 20:832
  Cold War 4:128–134
  communist ideas in 4:302–303
  conservation in 20:831–833
  conservatism in 20:820–821
  cult of personality in 20:823
  cultural studies 20:820–824
    conservatism 20:820–821
    gulag 20:820–823
    intellectuals 20:820
    Marxism 20:820–821, 20:823
    memory and mourning 20:823–824
    nationalities 20:821–822
    origins 20:820–821
    radicalism 20:820–821
    torture 20:822–823
    violence 20:822–823
  culture studies 20:820–824
  defense of science in 16:14–15
  dictatorship in 22:139–142
  ecology groups in 20:831–833
  environmental activism in 20:832–833
  environmental protests in 20:833
  and environmental law 20:832
  fertility in 9:99–102
  Glasnost reform 20:845, 25:838–839
  homosexuality in 21:742
  in socialist societies 22:849, 22:851
  industrialization 22:840
  intellectuals in 20:820
  Marxism–Leninism of 11:76
  Marxist archaeology in 14:670
  memory 20:823–824
  mourning 20:823–824
  nationalism in 20:821–822
  pedology in 20:826
  pollution in 20:831–834
  polytechnical education in 20:826, 20:828
  and postsocialism 18:187
  population change in 18:545–548, 18:550–551
  radicalism in 20:820–821
  social work in 22:807–810
  socialism in 20:820–821
  socialist societies anthropology of 22:849, 22:851
  Soviet–Chinese partnership/rivalry, in Cold War 4:128, 4:130
  Stalinism in 20:820–823
  torture in 20:822–823
  violence in 20:822–823
Soviet Union law (1918) 19:175
Soviet women 20:853
Soviet World, impact of Cold War in 4:124–125
Sovietization of North Korea 2:72
Soweto Pride 21:747
SOX *see* Sarbanes–Oxley Act 2002 (SOX)
SO_x *see* Sulfur oxides (SO_x)
Soyinka, Wole, works, *Seasons of Anomy* 14:253
Sozialistische Reichspartei (SRP) 18:410
Sozialökonomik 21:143–144
Sozialstaat 3:710, 25:528
Soziokultur programs 5:461
SP *see* Supportive psychotherapy (SP)
SPA *see* Society for Psychological Anthropology (SPA)

Space 6:458–460, 9:689–692, 10:84–85, 10:242–244, 13:228, 17:303–305, 18:376–377, 18:599–601, 19:801–802, 19:804–805, 23:82–88, 23:199–200, 24:813
as distance 23:199–200
as environment 23:199
environmental psychology 23:200
as experience 23:200
folding space 17:304–305
and gender 9:709–713
in health geography 10:628–629
and human ecology 11:283–284
and liminality 14:133–135
Marxist geography 14:677–678
and older people 1:407–409
ordering space 17:304
production and regulation of 24:820–821
psychological, moral, and religious significance 23:200
queer geography 19:759–761, 19:763
and regional geography 20:122–125, 20:127
relative 17:89
representations of 4:487, 4:487f
and rurality 20:795–796
and scientific explanation 10:62
sexuality and 21:808–812
social experience as a localized experience 23:200
and social movements 22:438
and territory 24:224–225
urban 24:801
urban sociology 24:926–930
and youth 25:826
Space of flows 12:91, 16:570
Castells on 3:230
The space of flows 24:866
Space of places 12:91
Space of relations, field as 5:466
Space scholarship, sexuality and 21:811
Space science concepts 21:275–276
Space-adjusting technologies 24:609–611
Space-based selection, in visual selective attention 2:165–166
Space-filling objects 23:106
Space-forming organizations 17:303
Space-in-law 13:476
Spaceship simulator 15:949–950
Space–time convergence 15:638
Space–time cube 6:716
Space–time modeling, geographic models 14:788–789
Space–time path 24:416–417
Space–time prism 24:416–417, 24:416f
Space–time visualization 6:715–716
SPACF *see* Sample partial autocorrelation function (SPACF)
Spade, Dean, on trans* politics 24:550–551
Spain
clinical psychology in
development 3:874
training 3:874
cultural policies in 5:464
immigration policy 11:665
laicization in 13:218
Spandrels of San Marco 8:453
Spanish anarchism
anthropology of 1:683–684
violent demise of classical anarchism 1:681
Spanish anthropology, Baroja on 3:150
Spanish colonialism 11:1
Spanish colony to republic 5:367–368
1933 Revolution 5:368
early republic 5:367–368
Republic's failure 5:368
Spanish Communist Party (PCE) 14:661
Spanish flu 18:174
Spanish system of succession 19:174
Spanish words 15:819
Spanish–American War of 1898 17:161
Spanish–English communities 3:918
Spanish-English-speaking adults 16:60
Spann, Othmar 18:435
Spanning items, scaling of 21:5
Spanos, Nicholas 11:466
The Spark 16:14
Spark technology investment zones 23:211
Sparse population code 25:187
Sparse representation-based classifier (SRC) 8:661
face recognition 8:661
Sparse social networks, populations with 9:9
Sparse variable selection problem 15:651
Sparta 1:698
Spartacus International Gay Guide 9:638
Spatial agglomeration 23:209
Spatial allocation 7:636
Spatial allocation problem 23:174t
Spatial analysis 5:278, 14:785–787
in archaeology 10:100–101
in geography 23:94
of infectious disease 15:88
market geographies 14:546
Spatial and temporal representation 14:320
Spatial association
as clusters 23:103
as covariance 23:102–103
as difference 23:103
focused association 23:103
as interaction 23:103
maps showing 23:100–101
in matrix terms 23:101–102
measures of 23:100–104
Spatial attention 12:419
Spatial autocorrelation 23:95–96, 23:105–110, 23:181–183, 23:187–189, 23:203
geographic models 14:788
measuring and testing for
autocorrelation tests 23:107
local autocorrelation, test for 23:108
semivariograms 23:107–108
spatial correlograms 23:107–108
and spatial data analysis 23:108–109
exploratory spatial data analysis 23:108–109
information loss in statistical tests 23:109
map interpolation 23:108
modeling spatial variation 23:109
spatial sampling 23:108
Spatial behavior, and lateralization 2:799
Spatial categorization, in infancy 23:116–118
Spatial choice, and spatial interaction 23:155–156
Spatial clusters, spatial association as 23:103
Spatial cognition 4:79, 4:82, 14:320, 17:703, 23:111–115, 25:133–137
geographic study of 23:111–112
topics in 23:112–114
history of, in geography 23:112
navigation 23:113–114
and nonverbal spatial tasks 13:295–296
orientation 23:113–114
spatial knowledge structures and processes 23:112–113
virtual reality in 25:135–136
Spatial cognition development
during infancy and early childhood 23:116–125
body-centered *versus* allocentric representations 23:119
categorization of more complex relationships 23:117–118
categorization of simple relationships 23:116–117
cultural differences 23:123–124
geometric cue use in reorientation by toddlers 23:121–122
geometric module 23:121–122
nongeometric cue use by toddlers 23:122
processing of geometric information in young infants 23:121
retinocentric *versus* body-centered FoRs 23:118–119
spatial frames of reference 23:118–119
Spatial communication, via language 23:114
Spatial correlograms 23:107–108
Spatial data 23:126–129
analysis, model-driven 23:97–98
data models 23:126–127
feature data structures 23:126–127
modeling solids 23:127
raster data structures 23:127
and data types 23:94–95
for medical geography 15:88–90
metadata 23:128–129
and representation in citizen mapping 3:619–620
and tyranny of data 23:94–95
ubiquitous 23:129
Spatial data infrastructures (SDIs) 3:619, 23:130–135
multidisciplinary interoperability 23:133–134
as social networks 23:130–133
Spatial Data Interest Communities (SDICs) 23:132
Spatial data mining 23:98
Spatial decision support systems (SDSS) 23:136–141
components of 23:137, 23:137f
implementations of 23:140
Spatial derivative transformations 23:411

Spatial differences of knowledge 13:91–92
Spatial dimensions, migration history 15:446–447
Spatial disparities of educational achievement and attainment 7:168–169
Spatial distribution 23:198–199
  and primitive society 18:911
Spatial divisions of labor 7:61–63, 13:164
Spatial domains of remote sensing 10:65
Spatial Durbin model 23:98, 23:94t
Spatial econometrics 20:148, 20:150, 23:105
Spatial economy, and location 20:149
Spatial equilibrium 14:288, 14:291
Spatial equity 23:148–151
  access to public facilities and resources 23:149–150
  analysis of 23:150–151
  and geographical access 1:64–65
  location of undesirable facilities 23:150
  meaning of 23:148–150
  overcoming the 'unfairness' of nature 23:149
  recent developments 23:150–151
  regional inequality 23:149
  and social inequality 23:148–149
  voting and representation 23:149
Spatial error models 23:203, 23:94t
Spatial extrapolation
  GIS/remote sensing for 10:67
Spatial fetishism
  and development geography 6:316
Spatial fix 12:913–914
Spatial Fourier transforms 23:411
Spatial frames of reference 23:90–91, 23:118–119
  body-centered *versus* allocentric representations 23:119
  retinocentric *versus* body-centered FoRs 23:118–119
Spatial frequencies 25:186
Spatial generalization, in occipital lobe 17:130–131
Spatial heterogeneity 23:187–188
Spatial imagery 15:223, 15:225
Spatial inequality, and poverty 18:757
Spatial inhibition 21:605
Spatial interaction 23:152–156, 24:611
  accessibility 23:155–156
  calibration 23:155
  commuting 23:154–155
  and market geographies 14:547
  migration 23:153–154
    typical empirical effects on 23:154t
  modeling 23:152–153
  models for 23:153t
  spatial association as 23:103
Spatial interaction data 23:95
Spatial job search 23:158–159
Spatial justice in developing world 12:914
Spatial knowledge, navigation based on enduring 23:164–167
Spatial lag models 23:205
Spatial language 23:89–93
  adpositions 23:90, 23:92
  coding of distinctions in language forms 23:92
  and cognition 23:89, 23:92
  locative constructions 23:92
  motion verbs 23:92
  as semantic domain 23:89–90
  universals and variation in semantic parameters in subdomains 23:90–91
    frames of reference 23:90–91
    topology 23:90
Spatial learning, virtual spatial navigation 16:348–349, 16:352
Spatial memory 24:566
  virtual spatial navigation 16:348–349
Spatial mismatch 23:155, 23:157–160
  cities 23:157–160
  commuting 23:157–160
  evidence of 23:157–158
  impacts of 23:159
  and labor market processes 23:158–159
  poverty 23:157–160
  race 23:157–160
  racial discrimination 23:157, 23:159–160
  segregation 23:157–158
  suburbanization 23:157–158
  unemployment 23:157, 23:159
  and urban policy 23:159–160
Spatial mismatch phenomena 24:916
Spatial models 10:67–68
  evolutionary models 24:470
  origin–destination models 24:470
  of segregation 21:407–409
  structural models 24:470
  tourism geography 24:470
Spatial navigation 23:161–171
  aging of navigational abilities 23:167–169, 23:169f
  based on transient spatial knowledge 23:161–164
    brain systems computing transient spatial information 23:163–165, 23:164f–165f
    path integration and spatial updating 23:161–163
  complexity of 23:162f
  entorhinal cortex 23:163, 23:167–168
  fMRI 23:166, 23:168
  grid cells 23:163, 23:166–168
  head direction cells 23:163–164, 23:166–167, 23:164f
  hippocampus 23:163, 23:166, 23:168–169
  navigation based on enduring spatial knowledge 23:164–167
    brain systems mediating route knowledge and cognitive mapping 23:166–167, 23:167f
    route knowledge and cognitive maps 23:164–166
  neuroimaging 23:163–164, 23:166–170
  place cells 23:163–164, 23:166, 23:168–170, 23:164f
  posterior parietal cortex (PPC) 23:163–164
  self-motion 23:163
  striatum 23:166, 23:168
  triangle completion task 23:161–162, 23:162f
  virtual reality 23:163, 23:166–168, 23:167f
Spatial optimization
  examples of 23:173–174
  historical roots of 23:173
  model 23:172–177
  problem 23:174t
    detailed view of 23:174–176
  research needs in 23:176–177
Spatial organization theory 14:546
Spatial orientation 5:314, 16:624–626
  and hemispatial neglect 10:768
  and vestibular system 25:67
Spatial pattern and shape 25:186–187
  grouping processes and figure/ground 25:186–187
  receptive field structure and specificity 25:186
  spatial location and layout 25:187–188
  spatial primitives 25:186
  visual recognition of objects, faces, and scenes 25:187
Spatial patterns
  analysis of 23:178–184
  central tendency
    lines 23:179
    points and areas 23:178–179
    spatial distribution 23:179–180
  describing 23:178–180
  future directions 23:183–184
  mapping 23:181–183
  quantifying 23:180–181
  spatial features 23:179t
Spatial planning 5:258, 18:188–189, 18:191, 18:202
Spatial point patterns 23:96
Spatial point platform 23:97
Spatial point processes 23:195–196
Spatial prediction 23:193–195
  basics of kriging 23:193
  multivariate methods 23:194–195
  trend models and their estimation 23:193–194
Spatial random fields 23:192–193
Spatial reasoning 13:103, 21:269
Spatial reference of land use information 13:241
Spatial regression models 23:97–98, 23:94t
Spatial relations 25:154
  absolute Location 14:285–286
  in cartography 3:170
  Marxist geography 14:676–677
Spatial representations 21:579–580
  frontal lobe mechanisms of 17:706–707
  and hemispatial neglect 10:766–767, 10:770
Spatial representations production 17:303
Spatial resolution 20:420, 20:424
Spatial sampling 23:108
Spatial scale geographic models 14:785–787
  issues 14:786–787
  macroscale 14:786
  mesoscale 14:786

Spatial scale geographic models (*continued*)
microscale 14:786
Spatial scales 21:1–3
ethnic conflict at 8:91
Spatial science 10:85–86
Spatial selectivity 23:245
Spatial short-term memory 21:910
deficits 21:913
neurofunctional architecture of 21:914–915
Spatial statistical methods 23:191–197
Bayesian spatial statistics 23:195
geostatistics 23:195
hierarchical models 23:195
comparisons 23:191–192
extensions 23:197
perspectives 23:191–192
recent advances 23:196–197
dimension reduction 23:196–197
nonparametric Bayesian methods 23:197
spatial–temporal statistics 23:196
spatial point processes 23:195–196
spatial prediction 23:193–195
basics of kriging 23:193
multivariate methods 23:194–195
trend models and their estimation 23:193–194
spatial random fields 23:192–193
Spatial statistics 20:148, 20:150, 23:105, 23:178, 23:185–190
population attribute 23:185
sampling
design-based approaches 23:186
model-based approaches 23:186
sampling plans 23:186–187
and spatial variations 23:187–190
Spatial stratification, geographic information systems/remote sensing for 10:67
Spatial thinking in social sciences 23:198–201
production of space and postmodernity 23:201
environment and society 23:201
space 23:199–200
as distance 23:199–200
as environment 23:199
environmental psychology 23:200
as experience 23:200
psychological, moral, and religious significance 23:200
social experience as a localized experience 23:200
spatial distributions 23:198–199
spatial turn 23:200–201
structurationist approach 23:200–201
Spatial time 24:407–408
Spatial tools 21:269
Spatial turn 23:198
and community sociology 4:366
frontiers 9:464
in multisited field studies 16:105
social sciences 23:200–201
Spatial updating, in navigation 23:161–163
Spatial-analytic analysis, social movements 22:437–438
Spatiality 7:637, 10:240–243, 25:824–825, 25:827
ritual 20:719
science 21:236–237, 21:239–240
of struggles 13:165–166
Spatial–temporal models (STMs) 23:202–206
atypical STMs 23:202–203
constructing 23:205–206
cross-sectional STMs 23:203
with spatial effects 23:203–204
panel STMs 23:204
with spatial effects 23:204–205
Spatial–temporal statistics 23:196
Spatiotemporal analysis 24:418–419
Spatiotemporal data 9:529
Spatiotemporal heterogeneity 17:900–901
Spatiotemporal reasoning 13:103
SPC *see* Sustainable production and consumption (SPC)
Speaker's meaning 21:340
Speaking 10:109
thinking for 25:782–783
Spearman, Charles E. 3:243, 8:691, 12:303, 14:812, 19:412
factor analysis 8:695
general intelligence 8:691
Spearman rank correlation coefficient 11:218
Spearman–Brown formula 19:427–428
for lengthened tests 3:788
parallel measurements model 19:427
Special audiences, of advertising 1:211
Special commission for investigating miracles and saints 15:578–579
Special Court for Sierra Leone (SCSL) 25:385
Special courts (*Sondergerichte*) 16:357
Special Drama 9:784
Special economic zones (SEZs) 23:207–208, 24:895–896
in China 23:210
clusters and 23:210–212
defined 23:208
from experimentation to growth and beyond 23:210–211
Guangdong 23:208, 23:210–212
industrial policy 23:209–212
objectives of 23:209
rationale and organization 23:208–209
role assigned to 23:209
spatial agglomeration 23:209
specialized towns 23:207, 23:211–212
success/failure of 23:209–210
supra-firm organizations 23:207, 23:211
Special education 21:114, 23:213–218
autism spectrum disorders 2:278–279
evidence-based practice for 23:215
frameworks 23:215–216
future directions and issues in 23:217
for learning disability 13:686
training in 23:214
in United States 23:219–224
accomplishments 23:223–224
amendments to EAHCA 23:223–224
children with disabilities 23:219–220
early involvement by federal government in education 23:221
Education for All Handicapped Children Act 23:221–222
enactment of compulsory education laws 23:219
federal district court rulings in educational opportunity cases 23:220–221
formation of national advocacy groups 23:219–220
US Supreme Court's Ruling in *Board of Education v. Rowley* 23:222–223
US Supreme Court's Ruling in *Brown v. Board of Education* 23:220
Special educational needs 21:114
Special educators 23:214
Special events, gay/lesbian tourism 9:640–641
Special Interest Group (SIG) 3:825
Special Juvenile Delinquency Project 25:814–815
Special measures 1:248
Special needs, joint attention in children with 12:846–848
Special Period in Peacetime 5:369
Special populations
interviews 12:673
speech and language impairment in 3:796–797
Special Programme for Research and Training in Tropical Diseases (TDR) 20:514
Special purpose graphics, regression analysis 14:160
Special Rapporteur on Violence against Women 6:630
Special relativism 20:94
Special rights 1:248
Special rule, U.S. legislative procedure 4:651
Special sciences 3:261
Special territories 23:212
Special Tribunal for Lebanon (STL) 25:385
Special Weapons and Tactics (SWAT) 24:953, 24:955–956
Special Weapons and Tactics (SWAT) vehicles 6:552
Special writing needs, students with 25:770
Specialist, multinational corporations 3:47
Specialist area, organizational change as 6:299
Specialist displays 10:355–356
missing values 10:355–356
weighted displays 10:356
Specialist training 1:866
Specialization 12:329–332, 13:629–630, 21:529, 22:588, 24:118
as adaptation to social density 6:244
agriculture 12:329–331
for APRNs 16:910
archaeology of 12:329–332
Boserup on 12:330–331
of brain regions 4:96–97

concept 12:331
consequences of 12:331
craft production 12:329–331
in craft production 5:108
and division of labor
functional explanation 9:536–537
exchange 12:331–332
history of approaches 12:331
industry 12:329–330
labor 6:601–604
landesque capital 12:329–330
in law firms 13:579–580
prestige goods 12:331
social complexity 12:329, 12:331
standardization 12:331–332
in structural contingency theory 23:609–610, 23:612
studies
future of 12:332
methodological issues in 12:331–332
urban 24:801
urbanization 12:329
workshops 12:329, 12:331
Specialized agencies, of United Nations 24:750–757
Specialized associations 22:662
Specialized corpus 4:950
Specialized education and examination, sociology of professions 19:107
Specialized sector 23:211
Specialized towns (STs) 23:207, 23:211–212
Specialties 6:491, 7:867, 7:870
Specialty gangs 9:594
Specialty publishing 19:628
Specialty Selection Test (SST), in psychiatric education 7:188
Specialty training, psychiatric education 7:185–187
Speciation, and constraint 1:87–88
Species
comparative demography across 8:418
general principles applying across 8:460–461
hierarchy of 22:264–265
multisensory integration among 16:101
number sense across 17:49
research, and animal rights 1:706
Species (Gattung) 19:317
evolution of language 8:364–365, 8:367–368
Species-energy hypothesis 13:241
Speciesism 2:615
Species-specific communication sounds in nonhuman primates 11:323–324
Species-specific defensive reactions (SSDR) 8:868
Specific Affect Coding System 8:670
Specific animal-type phobia 18:37–38
Specific area scales 17:753–754
good judge 17:753
measurement scales of good judges 17:753
trait understanding 17:753
specific deterrence 6:255
Specific disclosure, discovery 6:519–520
Specific environmental variance 12:298
Specific language impairment (SLI) 10:513, 13:344, 13:347–348, 18:790, 21:934, 21:939–940
Specific performance, and nonmonetary relief 20:413
Specific phobias 23:563
across the lifespan 22:495
defining 18:37
diagnostic features of 18:37–38
exposure therapies for 8:606
neuroscience perspective 18:38–40
prevalence of 18:38
treatment
exposure in virtual reality 18:42
exposure techniques 18:41–42
exposure-based treatment 18:40–41
Specific training, cross-cultural training 5:337
Specification problem, law 18:28
Specificity 6:345–346, 9:928–929
pattern variable 1:382
and perceptual learning 10:130–131
of psychological stress 23:585
of self-efficacy 21:512
Specified ignorance 24:718
SPECT *see* Single-photon emission computed tomography (SPECT)
Spectacle 4:752, 24:793
and criminology 5:245
entertainment 7:670–672
Spectacle, political
in Central Asia 3:317
Spectator theory of knowledge 18:805
Spectral density function 24:331
Spectral domains of remote sensing 10:65
Spectral imaging 10:98
Spectral representation theorem 24:332
of covariance stationary process 24:332
Spectral resolution 20:420, 20:424
Spectrum, Fourier analysis 9:375, 9:377
Spectrum models 17:918
Speculation 21:172
Speculative idealism 10:736–737
Speculative realism 5:646
Speculative searching 9:958–959
Speculators 17:505
Speculum majus (*The Greater Mirror*) 7:584
Speech 14:203, 15:584–585, 20:6–7, 24:533, 24:535, 25:325
as action 18:797
in action 9:136
codes 14:173–174
and communications research 4:279–281
computational auditory scene analysis 2:223–227
and dichotic listening 6:357–360
impairment in special populations 3:796–797
and knowledge of language, diversity of 4:294–295
Lacan on 13:212–214
and language evolution 11:329–333
and lexical consolidation 13:919, 13:921
recognition/production by machines 23:259–263
relation with writing 25:780–781
Vygotsky on 25:313
Speech, parts of 17:553–554, 23:885–886
approaches to 17:555–558
Croft's 17:555–556, 17:556t
Hengeveld's Functional Grammar 17:556–558
Tagalog and parts-of-speech hierarchy 17:559
criteria for distinguishing 17:554–555
cross-linguistic labeling of lexical categories 23:886
flexible 17:557
flexible lexemes 17:557
identifiability in flexible systems 17:557
morphology, role of 17:560
noun/verb distinction at syntactic level 17:558
pervasiveness of flexibility 17:558
semantic unity or homogeneity 17:557–558
short case studies and problems 17:558–560
Speech acquisition 23:236
Speech act adverbial clauses 1:187
Speech act of identifying reference 17:223
Speech act theory, of politeness 18:327–329
Speech acts 8:594, 14:241, 21:340–341, 21:344, 21:396
and pragmatic disorders 18:788–792
Speech analysis
concatenative speech synthesis 23:260
Hidden Markov Model 23:262
linear prediction model 23:260
Speech chain 23:259
Speech community 22:949
evolution of language 8:368–369
Speech errors, psychology of 23:229–234
interpretations of error phenomena 23:230–232
accommodations in error processes 23:230–231
building models 23:233
lexical selection processes 23:232
processing frames 23:231–232
sound exchanges and word exchanges 23:231
stranding errors and multiple levels of processing 23:231
methodological note 23:229
types of error 23:229–230
Speech event 12:783–784
primary factors 12:783
Speech impairment
in children, assessment of 3:794–795
reference models for 3:792–793
Speech perception 2:650–651, 23:235–242, 25:226
demise of categorical perception 23:237–239
Fuzzy logical model 23:239–240
higher-order or top-down influences 23:241
mirror neurons 23:236

Speech perception (*continued*)
motor theory of 15:583
multimodal speech perception 23:240–241
neurological underpinnings 23:241
as pattern recognition 23:239
psychophysics 23:236–237
Speech processing 23:243
cortical organization of 23:243–249
dorsal pathway, role of 23:245–246
Speech production 25:227
by machines 23:259–260
neural basis of 23:250–254
psychology of 23:255–258
conceptual preparation 23:256
errors 23:255–258
grammatical encoding 23:256–257
language comprehension 23:255
levels and processes 23:256–258, 23:255f
lexical access 23:257
methodological issues 23:255
phonological encoding 23:257
self-monitoring and repair 23:257–258
Speech rate 23:898–899
Speech recognition
and grammars 19:9–12
human *versus* machine 23:262–263
by machines 23:261–263
system design 23:262
and temporal lobe 24:207–208
Speech rights 3:686–689
Speech science 14:225
Speech situation, dexis 6:52, 6:54
Speech sounds, invariance and categorization of 11:324
Speech-framed gestures 10:110
Speech-motor control system 9:266–269, 9:272
Speed, diffusion models 6:399
Speed and accuracy of performance 8:677, 8:679
Speed of information processing, and aging 1:431
Speed tests, intelligence 12:323
Speenhamland law 18:244
Speke, J. 9:571
Spelling 3:508–509, 3:511–512, 20:6–9, 23:264–273
abstract rules or statistical learning 23:271–272
of children, factors affecting 23:266–270
sensitivity to morphological information 23:267–268
sensitivity to phonological information 23:266–267
sensitivity to spelling constraints 23:268–270
stage models 23:266
cognitive models of 1:453
current-day psycholinguistics 23:264–265
doublets 23:268–269, 23:272
dyslexia *see* Dyslexia
of experienced spellers, factors affecting 23:270–271
graphotactic cues 23:269, 23:271
homophone intrusions 23:270–271
information encoded in written words 23:265–266
morphological awareness 23:267–268
morphological principle 23:265
phonological awareness 23:267
positional constraints 23:265, 23:268–270, 23:272
Spelling constraints, sensitivity of children to 23:268–270
Spell-Out model 16:5
Spence, Thomas 22:837
Spencer, Baldwin, works, *Central Tribes of Native Australia* 9:135
Spencer, Herbert 4:339, 5:677, 6:241, 7:723, 8:233, 9:570, 12:774, 14:622, 15:709, 18:625, 19:124, 20:287, 21:167, 22:697, 23:274–278, 24:583
cardinal principles of 23:275
contributions to sociology 23:274, 23:276–277
on functionalism 9:536, 23:275–276
on geopolitics 23:275–276
influence in biology 23:275
influence in education and philosophy 23:275
law of evolution of 23:275
on moral philosophy 23:274–275
on organismic analogy 23:275
on political sociology 23:275–277
on social Darwinism 5:686–687
on sociology 22:991
on synthetic philosophy 23:275
works
Descriptive Sociology 7:589–590, 23:274, 23:276
Education: Intellectual, Moral and Physical 24:584
Encyclopedia of Social and Cultural Anthropology 7:590
First Principles 23:275
The Introduction of Sociology 24:585
Principle of Sociology 24:584
Principles of Biology 12:178, 23:275
The Principles of Ethics 23:274–275
The Principles of Psychology 12:176
Principles of Sociology 12:178, 15:709, 23:276
Social Statics 1:92, 23:274–275
The Study of Sociology 24:585
Spencer, Jonathan 2:871–872
works, *Encyclopedia of Social and Cultural Anthropology* 7:590
Spending, money 15:737–738
Spengler, Oswald 18:435
works, *Decline of the West* 14:744
Sperber, Dan, on anthropology of knowledge 13:71
Sperm competition 18:901
Sperm donation 9:809
and commodification 5:538
Sperry, Roger Walcott 13:393, 16:686, 23:279–281, 23:291–292
on brain function 23:280–281
causal potency of motor intentions 23:292
chemoaffinity hypothesis 23:279–280
on electrical fields in neocortical function 23:280
on higher cognitive function in mammals 23:280
on interhemispheric communication 23:280
on lateralization of cognitive abilities in humans 23:280
on neocortical information processing 23:280
on nerve connections 23:279
on nerves transplantation 23:279–280
on neurons 23:279, 23:281
scientific contributions of 23:279
'split brain' studies 23:292
works
Merging Mind, Brain, and Human Values 23:280–281
SPF *see* Social Protection Floor (SPF) Initiative
Sphere, and worldview 25:755
Spheres of influence 12:226
and imperialism 11:682
Spheres of justice 4:309–310
Spider phobia 18:39–40
brain responses 18:39f
Spigel, Lynn 25:81–82
on television 25:81
Spikes, in electroencephalography 7:353–354, 7:357–361, 7:360f–361f
Spike-triggered averaging 15:968
Spillman, Lyn, on financialization 5:537
Spillover 9:131, 22:159–160
Spillover effects, mass imprisonment 14:696, 14:698–699
Spinal cord, connections with 18:847
Spinal reflex behavior 21:882
Spindler, G., on educational anthropology 7:145
Spine development 18:818–819
Spinocerebellar ataxia type 6 (SCA6) cerebellar degeneration 3:328
Spinoza, Baruch 2:132, 4:337, 19:411
Spiral of silence theory 14:714, 15:30, 25:399
Spiral stairs of calculation (SSC) 14:779–780, 14:780f
Spirit medium 16:169
deference to family elders 16:169
in democratized forms 16:169
Spirit mediumship 23:282–283
Spirit possession
anthropology of 23:282
altered state of consciousness 23:282
colonialism 23:283
conceptualizing possession 23:282
cultural archive 23:283
discourses 23:282–283
ecstasy 23:282
embodiment 23:282
ethnic identities 23:282
gender 23:282–283

hermeneutical approaches 23:283
history from below 23:283
Holy Spirit 23:282
medicalization 23:283
mimetic ethnographies 23:283
modernity 23:282
new developments 23:283
pentecostalism 23:282–283
performance 23:283
protest movements 23:283
psychological and sociological approaches 23:282–283
and spirit mediumship 23:282–283
symbolism 23:283
trance 23:282
transcending of class boundaries 23:283
trauma 23:283
traveling spirits 23:283
upholding moral order 23:283
war 23:283
witchcraft 25:608
Spiritist movements 20:410
Spirits 9:666–670
ancestral 1:689
death 5:870–872
phenomenology of 10:737
philosophy of 10:738–739
shamanism 21:868–870
Spiritual abuse 20:253
Spiritual assessment 23:287–288
Spiritual care 23:287
Spiritual development, trajectories 20:300–302
Spiritual gifts 3:353
Spiritual healing 10:562
Spiritual homosexuals 21:741–742
Spiritual identity 20:298
Spiritual movement, Bellah's contribution 2:507
Spiritual salvation (*Seelenheil*) 8:214–215
Spirituality 3:542, 3:544, 4:267, 17:233, 17:804, 20:263, 20:298, 20:406, 22:149, 23:285–290, 24:642
and adolescent prosocial behavior 19:224
concept and definitions 23:285–287
connectedness 23:285–286, 23:289
contributions from social work 23:288–289
cultural competence 23:288–289
ecospirituality 23:289
holistic 23:288
humanistic spirituality 23:285–289
indigenous peoples 23:285, 23:289
and Islamic social work 12:754, 12:756–757
meaning 23:285–286
and mental illness 19:309, 19:311
origins and development 23:285
a personal relationship with Jesus Christ 20:298
practice models 23:287–288
psychiatric aspects of 20:350–351
purpose 23:285–286
relatedness 23:285–286
relational 20:317–321
religion 23:285–289
social justice 23:288–289
spiritual care 23:287
world view 23:286, 23:289
youth perception in United States 20:298
Spiritually integrated treatment 20:351–352
Spiro, Jarrett, on small world structures 5:467–468
Spite, and kin selection 13:28, 13:28t
Spivak, Gayatri 2:732, 18:658
SPL *see* Superior parietal lobule (SPL)
Splash queer theory 13:861–862
Splenium 12:416
Splice 13:476–477
Spline interpolation 6:133
Splines
linear mixed models 14:166–167
nonparametric regression 16:872–873
Splintering urbanism 20:678
Split account system 1:192
Split brain research 23:291–298
commissurotomy patients 23:294–295
defect of awareness 23:296
experiments 23:292–294
intelligent movements of cats and monkeys 23:292–294
history of 23:291–292
perception, learning, and prospective control of movements 23:292
Split control 4:829
Split treatment, in combined pharmacotherapy and psychotherapy 19:497
Split-brain patients 12:417–419
Split-plot designs 8:539
Splitting proofness 4:875–876
S-PLUS (statistical analysis package) 23:379
SPO *see* Slovenian Public Opinion (SPO) Survey
SPOC *see* Superior parieto-occipital cortex (SPOC)
Spoiled interaction, stigma 23:463
Spoken corpus 4:950
Spoken expression in human culture 17:274–275
Spoken language understanding (SLU) 23:261
Sponsored mobility 9:618–619
Sponsored projects, academic research and employment 1:51
Sponsoring 3:19
Sponsors, surveys and polls 23:753
Spontaneous behaviors 2:187–188
Spontaneous expressions 8:667, 8:673
Spontaneous order, unintended consequences 24:744–745
Spontaneous otoacoustic emissions (SOAEs) 21:776
Spontaneous retrieval 19:273–276
Spontaneous retrieval processes 19:273
Spontaneous social order 25:274–276
Spoon test 15:151–152, 15:152f
Spoonerisms 10:112, 23:229
Sport history 23:307–312
associativity concept 23:309
definition of 23:307
evidence and theory 23:308–309
future of 23:311–312
journals 23:308
public history 23:308
research agenda 23:310–311
societies 23:307–308
subject development 23:307
Sport labor market 23:320–322
Sport psychology, applied behavioral analysis for 2:411
Sporting ideology 21:136
Sports
anthropology of 23:299–306
allure 23:304–305
athletics 23:299–305
body techniques 23:300–301
children 23:302–305
comparison 23:301
embodiment 23:299–304
ethnographic attention 23:302–303
ethnography 23:299–304
global perspective 23:303–304
modernization 23:299–300, 23:303
moralities 23:302–303
performance 23:299–302
play and celebration 23:302–303
politics 23:299, 23:302–304
social practice 23:301–302
transnationalism 23:303–304
of Australian aboriginals 2:237
behavioral expressions 21:609
economics of 23:318–323
future directions 9:659–660
gay/lesbian 9:656–660, 20:873
as heteromasculine preserve 9:656–657
homohysteria and 9:657–658
navigation 9:658
in Japan 12:797
placebo effect in 18:156
shifting intersections of youth and homophobia 9:658–659
sociology of 23:313–317
consumer culture 23:313, 23:316
cultural analysis 23:315–316
feminism 23:314–316
functionalism 23:314–315
globalization 23:313–316
hegemony 23:315
institutional analysis 23:314–315
lifestyles 23:316
modernity 23:313–314
multiculturalism 23:316
nationalism 23:314
neo-Marxism 23:314–315
popular culture 23:316
postmodernism 23:314
poststructuralism 23:314, 23:316
racism 23:314–316
in twenty-first century 23:316
youth, developmental perspectives on motivation in 15:914–920
Sports, economics of 23:318
competitive balance 23:319–320

Sports, economics of (*continued*)
methods of cross-subsidization 23:320–323
revenue sharing 23:322–323
sport labor market 23:320–322
professional team sports 23:318–319
uncertainty of outcome 23:319–320
US profit-maximizing objectives *versus* European win-maximizing objectives 23:319
Sports and adventure programs, offender rehabilitation 20:195
Sports economy 14:720, 14:724
Sports Illustrated 9:659
SPOT *see* French Système Probatoire de la Observation de la Terre (SPOT) satellite
Spousal bereavement 25:636–637
Spouse abuse 6:633–634
Spouse relation 13:54–55
Spouses 8:766
disease recovery 22:735–739
personality, and marriage 17:797–800
psychological distress of cancer in 3:75
SPPA *see* Survey of Public Participation in the Arts (SPPA)
Sprachbund 1:955, 1:959–960
language contact 13:305
Sprachtheory 6:52
Sprawl 23:660–661, 23:665
Spread of disease 10:199–200, 15:86–88
Spreading activation model of semantic processing 21:572
Spreadsheets, computer assisted instruction 4:493–494
Spring theory, of interviewing 5:852
Springdale case 8:56–57
SPRT *see* Sequential probability ratio test (SPRT)
SPS *see* Sequestered problem solving (SPS)
SPSS (Statistical Package for the Social Sciences) program 4:517, 9:837–838, 16:34, 23:377–378
Spuhler, J.N., works, *The Evolution of Man's Capacity for Culture* 20:386
Spurious correlation 8:691
Spurious mechanisms 15:851
Sputnik 1:933, 20:827
Sputnik Shock 2:877
Spy novels 2:752–753
Spykman, John 18:374
Squamous cell carcinoma 23:696–697
Squash, phytoliths 17:459
SRA *see* Solicitors Regulation Authority (SRA)
Sraffa, Piero 25:9, 25:273
on profits 19:114–115
works
Production of Commodities by Means of Commodities 25:9
SRB *see* Sex ratio at birth (SRB)
SRC *see* Michigan Survey Research Center (SRC); Social Research Council (SRC) of the United States; Sparse representation-based classifier (SRC)
SRCC *see* Scandinavian Research Council for Criminology (SRCC)
SRCD *see* Society for Research in Child Development (SRCD)
SRD *see* Substance-related addictive disorder (SRD)
Srebrenica and Bosnia-Herzegovina massacre (1995) 17:635
S-REF *see* Self-referent executive function (S-REF) theory
SRES *see* Sex-Role Egalitarianism Scale (SRES)
SRH *see* Studies of adult self-rated health (SRH)
Sri Lanka
archaeology of 23:50–57
Buddhism in 2:891–892
education in 7:197
fertility transition in 9:80–83
history of 11:80
and migration 15:425
sociocultural aspects of *see* South Asia; sociocultural aspects
Sri Lanka Christians 3:531
Sri Vijaya Empire 13:812
Srivastava, Sameer B.
on cognition 7:416, 7:418
on social network actors 5:559
SRL *see* Self-regulated learning (SRL)
SRM *see* Self-regulation model (SRM); Social Reflection Measure (SRM)
SROI *see* Social return on investment (SROI)
Srole, Leo 1:720
works, *The Social Systems of American Ethnic Groups* 2:83–84
SRP *see* Sozialistische Reichspartei (SRP)
Srpska radikalna stranka (Serbian Radical Party) 18:411
SRRS *see* Social Readjustment Rating Scale (SRRS)
SRS *see* Sample Registration System (SRS)
SRSD *see* Self-regulated Strategy Development (SRSD) model
SRT *see* Social representations theory (SRT)
Srull, Thomas K., works, *Handbook of Social Cognition* 7:590
SS *see* Schutzstaffel (SS); *Schutzstaffel* (SS); Sensation seeking (SS); S(urface)-Structure (SS) representations
SSA *see* Smallest space analysis (SSA); Sub-Saharan Africa (SSA)
SSAs *see* Single shared assessments (SSAs)
SSC *see* Sentential Subject Condition (SSC); Spiral stairs of calculation (SSC)
SSCI *see* Social Science Citation Index (SSCI)
SSDR *see* Species-specific defensive reactions (SSDR)
SSEDL *see* Social Science Electronic Data Library (SSEDL)
SSH *see* Socioeconomic sciences and humanities (SSH)
SSHRC *see* Social Sciences and Humanities Research Council (SSHRC)
SSI *see* Safe School Initiative (SSI); Statewide Systemic Initiative (SSI)
SSK *see* Sociology of scientific knowledge (SSK)
SSN *see* Social security Number (SSN)
SSP *see* Strange Situation procedure (SSP)
SSRC *see* Social Science Research Council (SSRC)
SSRIs *see* Selective-serotonin reuptake inhibitors (SSRIs)
SSS *see* Sensation seeking scales (SSS); Sociology as social science (SSS)
SSSL *see* Social structure and social learning model (SSSL)
SSSM *see* Standard Social Science Model (SSSM)
SSSS *see* Sexual sensation seeking scale (SSSS)
SST *see* Sexual Strategies Theory (SST); Specialty Selection Test (SST)
S(urface)-Structure (SS) representations 16:1
SSW *see* Social security wealth (SSW)
St Louis Philosophical Society 10:740
Staatslehre 4:708
Staats-Lexikon oder Encyclopädie der Staatswissenschaften 7:585
Staatswissenschaft 18:466, 22:697
Staatswissenschaften 22:992
Stability 9:580, 15:895–897, 17:781–782
of aggression 1:377
of childhood friendship 9:444
of executive government 8:500–501
expectations 8:514
and infancy 12:10–12
personality, measurement of 17:864–865
repartnering 20:430–432
of self-control 21:483
Stability and change (in personality) 17:887
Stability and invariance 3:268
Stability of equilibrium 7:915–916
Stability of writing systems 25:765
Stability thesis 22:230
Stability under learning 8:514
Stabilization
and technology 24:137
Stabilization, economic 14:409
Stabilization technique 13:17–18
Stable age distribution 18:584–585
Stable contextual features 17:764
Stable democratic political cult 2:191
Stable distribution 6:572
characteristic function of symmetric stable law 6:572
definition 6:572
theorem 6:572
Stable equilibrium, in game theory 9:583, 9:586–587
Stable equivalent, and age structure 1:342
Stable isotopes 2:600–601
Stable laws, symmetric, characteristic function of 6:572
Stable population momentum 18:574–575
Stable population theory 6:173–174, 18:580
classical applications of 18:562–565
applications not requiring schedules of rates 18:562–563

applications that assume schedules of rates 18:563
applications that require schedules of rates 18:563
fertility 18:563–564
kinship 18:563–564
mortality 18:562–564
reproduction 18:562–564
survivorship 18:564
Stable populations 18:584–587
age structure in 1:342
inverse projection 6:150–151
multiside 16:114
Stable preferences 7:12
Stable size, population dynamics 18:579
Stable unit treatment value assumption (SUTVA) 3:258, 4:446–447, 8:528, 15:75–76
Stack, C., on North American community 17:23
Stacked tables 5:8, 5:7t
Stacking, apportionment 7:343
STAD *see* Student Teams-Achievement Divisions (STAD)
Staendische Gesellschaft crisis 23:344
Staff, military 15:508
Staff and volunteers, in disasters 22:758–759
Staff function, of in-house counsel 12:127
Staffing, in human resource management 11:348, 11:352
Stag Hunt game 18:932
Stage, dramaturgy 6:646
Stage models 10:624
revival of 22:955
of spelling 23:266
Stage theories 15:700–701, 16:501–503
Stage-based coming-out models 9:651
Staged or controlled speech events 14:185
Stages, in individual development 21:232
Stages of change 15:909
health behavior 10:579
transtheoretical model of 15:909
Stagflation 18:736
Stagnation
stage of aging 1:435
theories 7:39
Staircase model 16:504
Stakeholder analysis 18:273, 18:278, 22:874–875
public policy ethics 19:581–582
Stakeholder citizenship 24:591–592
Stakeholder communication, about climate change 3:854–855
anthropogenic climate change as global problem 3:854
ecological modernization 3:855
economic development 3:854–855
scientific uncertainty 3:854
Stakeholder involvement approaches 7:578
Stakeholder management, and customer satisfaction 5:630–632
Stakeholder management approaches, strategic planning 23:517–518
Stakeholders 3:1–3, 19:117–118, 22:169
area and international studies, in United States 1:943–948
collaborative research with 7:864–865
corporate governance 4:933–934
environmental risks 7:787
program evaluation 19:117, 19:119–121
of school improvement 21:91, 21:93–94
Stalin, Joseph V. 6:243, 13:428, 20:634–635, 21:368, 21:370–372, 24:448–451
era, homosexuality during 21:742
and Marxism–Leninism 14:658–659
Stalin Constitution of 1936 13:429
Stalin Revolution of 1929–1930 20:856
Stalinism 5:301, 6:812–813, 18:434–435, 22:841
education 20:826–828
model of social sciences 22:141
post-Stalinist social sciences 22:141–142
in Soviet Union 20:820–823
Stalinist economic planning 18:626
Stalinization 18:406
Stalking 15:593
in United States 5:725
Stalling, of fertility transition 9:97–98
STAMP *see* Structural Time Series Analyser, Modeller and Predictor (STAMP)
Stamp Act 1765 1:609
Stamped scrip, Fisher's 9:220
Stampede 22:900
Stamping-in reinforcement process 20:211–212
Standard & Poor's 500 (S&P 500) stock index 9:317
Standard American English 14:173
Standard American Sociology (SAS) 22:722–724
Blau's attempted synopsis and program 22:724
conflict theory 22:723
social construction of reality 22:723–724
theoretical reflections 22:724
Standard commercial insurance 12:819
Standard consolidated areas (SCAs) 24:857
Standard consolidation theory (SCT) of memory 1:646
Standard coordinates 5:2
Standard deviation 1:852
Standard diffusion model 17:392
Standard disclosure
discovery 6:519
Standard disclosure, discovery 6:519
Standard error 20:492
Standard error estimates 13:385, 13:388
Standard error of measurement, classical true score model 19:425
Standard Industrial Classification (SIC) 5:798, 6:586–587
Standard International Trade Classification (SITC) 5:760
Standard legal scholarship 13:795–796
Standard logic 16:924–925
Standard metropolitan areas (SMAs) 24:857
Standard metropolitan statistical areas (SMSAs) 24:857
Standard modernization theory 8:285
Standard mortality ratio (SMR) weighted estimator 19:144
Standard occupational classification (SOC) 5:798–799
Standard Oil Company 17:166
Standard possessive constructions 18:632
Standard regression analysis 12:208–209
of observational data 12:205
Standard same–different task 21:949
Standard serial model 12:78–79
Standard Social Science Model (SSSM) 20:389, 22:893–894
Standard theory of sequential analysis 5:909
Standard tree for ordered processes 16:544
Standard tree for unordered processes 16:544
Standard welfarist philosophies 25:508–509
Standard-form insurance policy 12:216
Standardization
behavior therapy 2:418, 2:421
empirical social research 7:558–560, 7:562–563
global governance 10:156–157
production 12:331–332
quantification 19:701
of vocational education 25:234–237
Standardized assessment, of negative symptoms 19:290–291
Standardized bias estimate 19:144–145
Standardized fertility rate 6:110
Standardized interviews 23:767
Standardized mortality ratios (SMRs) 6:121, 21:21–22, 23:108–109, 23:773
Standardized testing
for educational sociology 7:293, 7:295
in learning spaces 14:589
Standardized tests 3:819
Standardized World Income Inequality Database (SWIID) 5:795
Standards 1:474, 21:101
of linguistic databases 5:814–815
quantification 19:700–702
in residential child care 20:537–538
Standards Movement 9:330–331
Standards of Americans with Disabilities Act 23:466
Standards of behavior, business codes 8:21–24
Standards of Care (SOC) for Health of Transsexual, Transgender, and Gender Nonconforming People 24:549, 24:551
Standards-based grading 3:823
Standing, infant motor development 15:975–976
Standing Committee of the National People's Congress of Gansu Province 8:234–235
Standing issues, cost–benefit analysis 5:48
Standing to appeal 1:823
Standpoint epistemology 6:245–246
Standpoint feminism
and feminist criminology 8:913

Standpoint feminism (*continued*)
social work 8:935, 8:937
Standpoint of social (*Gesichtspunkt des Sozialen*) 22:828
Standpoint theory 23:324–330
central arguments 23:324
collaborative practice 23:328–329
contemporary 23:328
content critiques 23:326–328
criticisms and elaborations 23:324–326
domains of standpoint thinking 23:326–329
feminist 23:324–325, 23:327
and intersectionality 12:658–659
Stanford Binet test of intelligence 13:347
Stanford Engineering Everywhere (SEE) initiative 6:896
Stanford Hypnotic Susceptibility Scale (SHSS) 11:467
Stanford Prison Experiment 4:154–155, 13:901–902, 24:711–713
Stanford Three Community Study 19:549–550
Stanford–Binet Intelligence Scales 10:732, 12:310–311, 17:770, 17:787
5th edition (SB5) 12:287
Stanford–Binet test 12:307, 12:950
Stanhope's metal press 18:916
Stanko, Betsy 8:855–856
Stanley, Julian C. 7:161, 8:569
works, *Experimental and Quasi-Experimental Designs for Research on Teaching* 9:132, 22:298
Stanton, Elizabeth Cady 22:443
Stanton, Frank, works, *Communications Research 1948–1949* 7:563
Stanton, Gregory, works, *The International Alliance to End Genocide* 10:7
Stapel, F.W., works, *Geschiedenis van Nederlandsch Indië* 11:84
Star, Susan Leigh 21:188
STAR 3 24:312
Star jobs 2:52–53
Starbuckization 14:829–830
Starch, paleoethnobotany 17:458
Stare decisis 12:886
legal databases 5:803
Stare decisis principle (doctrine of precedent) 3:669, 4:249–250, 4:700, 24:430
Starhawk 19:748–749
Starostin model 11:25–26
Starr, Frederick 2:71
START *see* Strategic Arms Reduction Talks (START); Study of Terrorism and Responses to Terrorism (START)
Starting hypothesis 6:594
Startle
fear-potentiated 8:860–867
visceral modulation of 12:616
Startle reflex 23:331–337
addiction 23:335
anxiety disorders 23:333
brain damage 23:335
clinical conditions 23:332–335
depression 23:334
eating disorders 23:334–335
emotion 23:331–332, 23:334–336
eyeblink 23:331–333, 23:335
nonclinicial individual differences 23:335
response and modification 23:331–332
attentional modulation 23:332
elicitation and measurement 23:331–332
modulation of startle by affectively valenced stimuli 23:332
prepulse inhibition 23:332
ubiquity 23:331
schizophrenia 23:332–333
sensory deficits 23:335
startle syndromes 23:332–333
culture-specific startle disorders 23:333
hyperplexia 23:332–333
startle stimulus-induced disorders 23:333
stress 23:333
and well-being 23:335–336
Startle stimulus-induced disorders 23:333
STATA (statistical software) 23:379
State anxiety 1:808
State autonomy 22:31, 22:33, 22:35
State breakdown 20:627–628, 20:631
State building 15:327
nation 16:223
Ottoman 9:468
theory of 23:362–365
and war 25:366, 25:369
State capitalism 3:97, 5:294
State capture 5:14–15, 5:15t
State Children's Health Insurance Program (SCHIP) 14:699
State control, and motherhood 15:882
State crime, genocide 10:15–16, 10:18–21
State databases, in United States 18:945
State dependence
and crime 1:312–313
of response synchronization 16:637
State elites' conversion to leniency 19:636–639
State entrepreneurialism, in China 24:975–976
State equation 24:354
State fiduciary duty cases 4:934
State formation 23:353–355, 25:403
in Africa 1:287, 1:289
in Europe 20:746
future of state 23:354–355
globalization 23:353–355
history of state 23:353–354
and life course 14:76
and military 15:518–519
modern states 23:354
problem of state 23:353
and Reformation/confessionalization 20:100, 20:103
regional integration 23:353
and warfare 25:405, 25:410
witchcraft 25:604–606
State fragility 5:841
State growth management 10:450–451
State hypothesis, of hypnosis 11:467–468
State interest 12:386–387
State intervention, use of statistics in determining 23:426
State interventionism 13:20–23
State monopoly, in Chinese cultural industries 5:597–598
State occupation probabilities 8:315
in event history analysis 8:310–311, 8:315–316
State of emergency 21:28
State of sexuality, queer anarchism 19:750
State of the art innovation and employment 12:152–159
State of the Mankind study 23:750–751
State planning, industrialization 11:868
State power 2:919, 10:91
State probabilities 19:786
State reconstruction 4:707–708
State regulation, shadow work 21:862, 21:864
State rituals 23:351
State sciences 6:487
State secrecy 21:377–378
State self-esteem assessment 21:522–523
State self-esteem scale 21:522–523
State smoothing 24:358–359
State socialism 18:187, 18:189, 20:847, 20:851, 22:834
Eastern Europe social work under 22:796–797
and social democracy 22:242–243
State socialist 18:187
State societies, Andes 23:38
State space 14:54
State space methods (time series) 24:354
analysis 24:358–360
diagnostic checking 24:359
disturbance smoothing 24:359
filtering 24:358–359
forecasting 24:358–359
Kalman filter for prediction 24:358–359
parameter estimation 24:360
state smoothing 24:359
computing 24:360
linear Gaussian state space models 24:354–358
ARIMA components models 24:356–357
dynamic factor models 24:358
latent risk model 24:357
multivariate models 24:357
regression and intervention effects 24:356
structural time series analysis 24:356
local level model 24:355–356
non-Gaussian time series 24:360
nonlinear time series 24:360
recent developments 24:360
unobserved component models 24:355–356
State taxation 23:338–342
business tax 23:340–341, 23:341t
guidelines for reasonable tax assignment 23:339–340
income tax 23:339–342, 23:341t
inheritance tax 23:341t
in OECD countries 23:340–341, 23:341t

political economy of 23:341–342
property tax 23:339–342, 23:341t
rationale 23:338–339
sales tax 23:340–342, 23:341t
subfederal governments 23:338–341
wealth tax 23:341, 23:341t
State terror 10:4
State v. Chapple 13:769
State variables 17:253
State violence 6:553–554
State-appointed professions, members of 19:97
State-centered approaches of trade policy 24:487–488
State-centered thesis 25:513
State-centrism 15:493
State–church cleavage 20:746
State–city interface 24:784–785
State-colonial policies, hegemony 10:748–749
State-contingent markets 10:614, 10:616
Statecraft, and infrastructure 12:122
Statecraft theory 22:696–697
Stated preferences
in contingent markets 25:17–18
see also Revealed preferences
State-directed activism, and gay/lesbian activism in Latin America 13:862
State-enforced conformity, sexuality 21:804
Statehood
anthropology of 23:347–352
advocacy 23:350
borders 23:348, 23:350
citizenship 23:348, 23:350
conflict 23:347–348, 23:350–351
criteria for statehood 23:348
discipline of anthropology and state 23:349–350
kinship 23:347–349
knowledge 23:347–350
law 23:347–351
legitimacy and modern state 23:350–351
modern statehood 23:347–348
nationalism 23:350–351
political anthropology 23:347–348, 23:350–351
postcoloniality 23:348–349
socialist societies 23:348–350
strategy of self-representation 23:350
subject formation through linking territory to people 23:348–349
violence 23:350–351
war 23:347–351
Statehood changes
governance as 10:298
modern, institutional architecture of 10:298–299
State-led gentrification 10:50
Stateless societies 18:336
and anarchist theory 1:681–683
Statelessness, in Central Asia 3:316–317
State-level laws, for LGBTQ/sexual minority youth in schools 21:88
Stateliness 23:366–367
Statel-led training, in informal education 7:122
Statements, Foucault's énoncés 14:199–200
State-of-the-art face recognition algorithms 8:658
State-owned broadcasting institutions 2:859–860
State-owned enterprises (SOEs) 19:598, 23:210
in China 4:933, 4:936, 5:598–599
and firms 7:75
reform 23:210
State-recorded social statistics 19:733
State-run schools, practices of 3:480
States 4:339, 10:738–740, 14:650, 14:664, 18:374–375, 20:160–163, 22:438–439, 23:353, 23:371, 24:343–344, 24:765, 25:349
administrative needs of 7:559
age of equality 23:365–366
age of monarchy 23:363–364
anthropology and 18:338–340
archaeology of 23:371–375
assemblies of estates 23:364
balance of power between 2:345
Benhabib on 2:520–521, 2:523
and biomedicine 2:665–666
borders 2:761–763
and brokers 2:870–874
capacity 19:692
capacity, and revolutions 20:645
and church 3:574–578
and civil society
liberalism and 14:23
as complex systems 23:368–369
concept of 3:314
and constitutions 4:710, 4:713–714
as corporation of citizens 11:817
credible commitments
and democratic transition 23:370
and growth 23:370
dealing with deviants 4:823–824
defense economics 6:1–3, 6:5
definition 2:246
definition, problems of 23:362
and development 6:276–281
disintegration 23:366–367
in dual actor failure 10:179
economics 23:371–373
agricultural and pastoral systems 23:372
full-time craft specialists 23:373
markets 23:373
tribute, taxes, and private estates 23:372
urbanism 23:372–373
and family history 8:743
and feminist legal theory 8:921
formation, time consistency and 23:370
future of 23:354–355
geography of 23:356–361
boundaries 23:357–358
hegemony 23:360
neoliberalism 23:358–360
origins and features 23:356–359
power 23:356–360
sovereignty 23:358–359
territory 23:356–357
global governance 10:157
global history 10:163–164
and governmentality 10:308–312
historical sociology 11:27–28
history of 23:353–354, 23:362–367
immigration 11:669–674
and immigration 11:641–642
and journalism 12:853
and law 20:780–781
law, rule of 23:362, 23:364–365
and liberalism 14:14
media imperialism 15:58
and migration 15:469
and military history 15:507
modern 23:362–366
modernization 3:122
nationalism 16:290–296
and nonstate communities 23:369
and observations in large numbers 23:424
and pluralism 18:232–235
and political institutions 12:182
political liberalism 18:386–387
power 18:391
premodern alternatives 23:364–365
public bureaucracies 19:531–533
redefining developmental role 6:280
religion and world views 23:374
monumental architecture 23:374–375
writing 23:375
and revolutions 20:643–646, 20:648
rise and fall of 23:375
role in East Asian economic development 6:788–790
role in economic transformation 6:276–277
role in internal migration 12:437–438
role in Third World development 6:277–278
development planning 6:277
entrepreneurial initiative 6:277
import substitution 6:277
neoliberal critiques of state intervention 6:277–278
roles
in liberalism 14:24
social organization 23:373–374
class stratification 23:373–374
ethnicity 23:374
gender 23:374
and social work 3:471–476
social work and 14:667
and society 23:343–346
Cold War 23:345
crisis of Staendische Gesellschaft 23:344
government 23:344
historical roots of distinction 23:343–344
jurisdictional political power 23:344
Keynesian compact 23:345
political authority 23:343
political prerequisites of civil society 23:344–345
public sphere 23:345
statism 23:345

States (*continued*)
totalitarian state 23:345–346
in twentieth century 23:345–346
unsettled distinction 23:343
state building, theory of 23:362–365
and statistics 23:423, 23:426
terminology 3:574
territorial 24:223–224
territorial conflict 24:214–220
theorizing 23:359–360
theory 23:362–363
theory, Pierre Bourdieu 2:781
and urban policy 24:900–902, 24:904–905
at war 25:421–422
States *versus* traits 17:920
State-sponsored prenatal genetic selection 10:36
Statewide policy-induced variation
class size 3:753
Statewide Systemic Initiative (SSI) 3:830
Static control theories 4:836–837
Static dimension of language 10:109
Static equilibrium 7:912, 7:914–915
Static models
family and household demography 8:728–729
of labor supply 13:192–193
of network formation 16:574
Static treatment strategy, in time-varying treatments 14:341
Static-group comparison designs *see* Comparison group
Stationarity, of political data time series 18:517
Stationary Markov chains, for social analysis 14:578–579
Stationary population 18:572–577
Stationary populations 16:114
Stationary time series process 24:316–317
Stationary vector 23:469
Stationers' Company 18:916
STATIS method 19:81
Statism 23:345, 25:741
Statist liberalism 18:299, 18:303–305
Statistical agencies 10:302–303
analysis and interpretation of data 10:306
challenges faced by 10:306–307
continually developing more useful data 10:305
cooperation with data users 10:305
coordination and collaboration 10:306
dissemination of data 10:305
evaluations 10:306
fair treatment of data providers 10:305–306
independence 10:305
mission 10:305
openness 10:305
practices of effective 10:305–306
quality and professional standards 10:306
research program 10:306
use of statistics 10:306
Statistical analysis 19:696
of causality and social science research 3:248
multilevel methods 23:381–386
outliers 23:387–391
of variables 7:613
Statistical analysis languages 23:379
Statistical analysis packages 23:376–380
current choices 23:379–380
history 23:376–379
mainframes 23:377–378
microcomputer 23:378–379
statistical analysis languages 23:379
Statistical Analysis System (SAS) (software) 7:3, 9:837–838, 13:388, 23:378, 23:380
Statistical and Science Policy Office 10:304
Statistical approach to learning 8:514
Statistical assumptions, errors in 8:8–12
Statistical biases
caused by sample selection, family reconstitution 6:141–142
bias from delays in recording 6:141–142
selection by data quality 6:141
selection by migration 6:141
truncation and censoring 6:141
Statistical classification 23:392
Statistical clustering 23:392–396
algorithms 23:392–393
hierarchical clustering 23:393
object trees 23:393
partitions 23:392–393
probabilistic classification 23:393–395
hierarchical models 23:394–395
mixture models 23:393–394
probabilistic search 23:395
validity of clusters 23:395
Statistical conclusion validity 12:652–653, 19:739
Statistical correlation 23:425–426
Statistical data sources, defense economics 6:1–2
Statistical databases, in social research 18:944–945
Statistical decision theory 16:430
Statistical demography 6:125
Statistical disclosure limitation 4:577–581
data quality 4:577
audit 4:577–578
data utility 4:578
data snooper 4:577–579
data swapping 4:579
masking 4:578–580
measures of disclosure risk 4:579–580
restricted data through 4:579
recodings 4:580
samplings 4:580
suppression 4:579–580
in social research 18:944–945
synthetic or model-based data 4:580
topcoding 4:579
types of disclosure 4:578–579
Statistical discrimination 6:537
and labor market outcomes 17:71, 17:74
Statistical distribution models
goodness of fit tests 23:397–404
EDF tests 23:398
examples 23:402–404
modern developments in smooth testing 23:400–402
Neyman Smooth tests 23:398–400
orthonormal functions 23:399–400
tests for multinomial models 23:397–398
Statistical distributions 23:425–426
normal 23:424
special and continuous 6:575
approximations 6:571–574
extreme value distributions 6:572–573
special and discrete 6:580–584
stable distribution 6:572
Statistical estimability 23:405
Statistical estimation 16:826–827
Statistical fallacies 6:868
Statistical fictions 14:64
Statistical flukes 23:415
Statistical generalization 3:200
Statistical genetics 8:715–716
eigensimplification of likelihood function 8:718–721
impact of sequence data 8:717–718
Statistical hypothesis testing 19:471
Statistical identification 23:405
Statistical inference 3:200
likelihood 14:116–118
for MPT models 6:528
Neyman on 16:827
in outcome-based sampling 17:104–105
in presence of nuisance parameters 23:421–422
from survey data 20:878
Statistical inference testing 8:566
Statistical information
on enterprises and industries 5:704
on market concentration 5:704
on media content 5:705
on media use 5:705
on size and structure of market demand 5:704
on technical infrastructure 5:704–705
on use of telecommunications service and Internet 5:705
Statistical Information and Monitoring Program on Child Labor (SIMPOC) 3:405
Statistical institutes 5:847
Statistical invariance, fiducial and structural 9:124–127
Statistical learning 2:651–652, 13:307–309
language development 13:321–322
spelling as 23:271–272
Statistical methods
of ancestry testing 9:903
for obesity 17:77, 17:79
Statistical Methods in Medical Research 4:470
Statistical modeling with survey data 13:388–389
Statistical models
event history analysis 14:55–56
for event history analysis 8:307
of language comprehension 21:625

for significance tests 21:958–960
of social change 22:197
for social network analysis 22:475
Statistical norm, disease as departure from 17:1
Statistical Office of the European Communities (Eurostat) 5:801
Statistical Office of the European Union 5:776
Statistical Package for the Social Sciences *see* SPSS (Statistical Package for the Social Sciences) program
Statistical pattern recognition 23:411
applications of 23:416–417
future prospects 23:417
generic 23:411–413
Statistical power 20:155, 20:158
genetic studies of behavior 9:936–937
Statistical power of tests 15:275
Statistical probability 23:424–425
Statistical purposes 10:301
Statistical regression 23:425
Statistical software, capture–recapture methods 18:606
Statistical sufficiency 23:418–422
ancillary statistics 23:420–422
approximate sufficiency 23:421–422
asymptomatic sufficiency 23:421
basic notation 23:419
complex models 23:422
dimension reduction 23:421–422
exponential families 23:420
inference in presence of nuisance parameters 23:421–422
likelihood 23:420–422
maximum likelihood estimate 23:421
statistical inference 23:418–420, 23:422
sufficiency and minimal sufficiency 23:419–420
summary statistics 23:418
Statistical surveys, history of sampling 20:875–880
Statistical systems 5:742–750, 22:600
applied social research 5:692–699
archives 5:700–703
communications and media 5:704–711
crime 5:712–716
crime and delinquency 5:717–719
current population survey 5:747–748
health and retirement study 5:747
living standards measurement study 5:747–748
Luxembourg income study 5:747
demography 5:735–741
economic 5:758–762
education 5:763–774
national longitudinal surveys 5:745–746
consumer expenditure survey 5:746
panel study of income dynamics 5:745
survey of income and program participation 5:745–746
time use survey 5:746
Statistical systems (in education) 5:763
education equality 5:766–767
education expenditures 5:767
gender 5:766–767
learning outcome statistics 5:763–765
criticism 5:765
Statistical tables, Fisher's 9:224
Statistical uncertainty
population forecasts 18:595–596
in events and rates 18:595
expressing uncertainty via predictive distributions 18:596
logical flaws in high and low variants 18:595–596
Statistical-relevance model of probabilistic explanation 8:595–596
Statistics 19:700, 20:926, 24:341, 24:343–344
Bayesian 2:373–377
Bayes's theorem 20:928
and behavioral sciences 20:928
branches, and functional data analysis 9:518
bureaus of 23:423–424, 23:426
computer data processing in 23:426
and computers 4:517–518
concept of average 23:424
contribution of Karl Pearson to 17:661
contributions of De Finetti to 5:857
culture 5:727
databases 5:817–823
of ecological fallacy 6:865–867
empirical legal studies 7:554–555
energy 7:628–629
on environment, urbanization, and human development 5:775
in ethnography 8:189–190
Gaussian law 23:424
history of 23:423–427
econometrics 7:8–11
hypothesis testing in 11:491–493
illiteracy 11:590
instrumental variables in 12:205–209
integration of inferential into probabilistic models 23:425
invariance in 12:705–709
Jewish population 12:807–810, 12:808t
on labor, and social welfare policy 23:426
labor and work 5:798–802
law of large numbers 23:424
convergence conditions 23:424
likelihood in 14:116–119
mathematical 23:425
meanings of word 23:423
methods based on measured genotypes in humans 9:934–936
moral 23:424–425
nineteenth century 23:423, 23:426
normal law 23:424
official 23:423–424, 23:426
for prosopography 19:256–257
queues 19:786–788
regularity of series of 23:424–425
and reproduction 20:451–452
rise of 19:703–704
Savage's contributions 20:926–927
selection bias 21:414–418
as social science 22:697
and social sciences 20:928
street children 23:534–536
improving localized counting 23:535–536
research gaps 23:534–535
system of notation 23:425
territorial connotation 23:423
theories
Fisher's 9:222–223
tools, graphics 10:356–357
Wright's contribution 25:759
Status 3:757–758
and authority 2:269–272
of celebrity 3:274–275
and cultural consumption 16:555–556
and cultural inequality 22:331–332
indicators 22:333–334
and free time 13:846–847
and gender development 9:746
inconsistency 22:333–334
labor as 25:676–677
and language 22:950
and leisure 13:846–847
and media consumption 16:557
or contract societies 1:766
organizational groups 10:408
and potlatch 18:715
primordial sociobehavioral outcomes 22:879–881
and religiosity of women 25:635
and role, social psychology of 23:428
as social position 23:428–429
as social prestige 23:429–430
as social process 23:430–431
social psychology 22:538–542
and social stratification 22:707, 22:709, 22:711
social structure 22:714
as stable extensions of position 23:431–433
structural aspects 23:435–439
symbolic boundaries 23:850–854
white-collar workers 25:565–566
see also Social status
Status (social position) 23:435
conformity 23:436
mapping 23:436
social structure 23:437
socialization 23:436
as structural position 23:435–436
Status analysis, Goldthorpe on 10:280–281
Status attainment 20:292
and leisure 13:853
and social mobility 15:641–642
Status attainment theory 22:333–334
Status characteristics theory 22:541–542
Status civil francais (civil status)
civil law colonies 23:655
Status consumption 4:696–698
Status cultures 5:376
Status epilepticus 7:853
Status frustration, delinquency 6:71–72
Status functions 21:344
Status group formation, and social mobility 15:640

Status hierarchies, and groups 10:437–438
Status incongruity hypothesis, sexism 21:698
Status leveling hypothesis 14:52
Status maintenance hypothesis 14:52
Status markets 5:536
Status motives, arts socialization 2:9
Status of groups 10:424
Status orders, cultural production 5:609
Status passage/constructionist model 13:808–809
Status planning, in language policy 13:354
Status politics 2:502
Status preservation, and social networks 5:468
Status quo 8:684
Status relations 10:762–763
  model 10:762–763
Status theories 22:959
  arts socialization 2:7–9
Status-distributive function, of education 7:127–128
Status-seeking theory 22:332
  versus skills 22:332
Statut coutumier status
  civil law colonies 23:655
Statute law 4:248
Statutes 4:810–811
  organizing with legal databases 5:807–808
    annotations 5:808
    challenge 5:807
    search capabilities on LexisAdvance and WestlawNext 5:807–808
    Shepardize and KeyCite functions 5:808
Statutes of Fraud 3:28–29
  in England 3:28–29
  in France 3:28–29
  in Japan 3:28–29
  in United States 3:28
Statutory insurance scheme 12:217
Statutory services, relationship between family organizations and 15:217–218
StatXact 16:134
STBs *see* Set-top-boxes (STBs)
STDs *see* Sexually transmitted diseases (STDs)
Steady-state distribution 19:786
Steady-state economy 6:853–854, 6:861
Steady-state systems, epigenetic inheritance 7:833
Stebbing, E.P.
  works
    The Creeping Desert in the Sudan and Elsewhere in Africa 7:751–752
    Forestry in West Africa and the Sahara 7:751–752
    The Forests of India 7:751
Steel, Carolyn 24:867
Steel Pan 3:147
Steelband clashes 3:147
Steele, Richard 15:11–12
Steering
  and residential segregation 20:542
  sociocybernetics 22:913–914
Steglich, Christian, on social networks 16:556–557
Steiger, Anatolii 21:742
Stein, Lorenz von 3:710, 10:739
Steinberg, Joel, works, *In Search of Southeast Asia* 11:87
Steiner, Hillel
  left-wing libertarianism 14:43
  works
    An Essay on Rights 14:43
Steinkuehler, Constance 4:501
Steinthal, Heymann 16:683, 16:683f
Stellanian accounting system 14:781–783, 14:782f–783f
Stem 15:820–821
STEM (science, technology, engineering, and mathematics) 24:765
  careers 1:40
  course-taking in high school 1:40–41
  disciplines 22:608
  fields 9:684–685
  jobs 1:40
Stem cells 9:977–980
  banking 9:978–979
  research, and cloning 23:440–445
    bioethics, biopolitics, and the uses of embryonic life 23:441–442
    clinical ethnography 23:443
    embryonic stem cells 23:440–441
    feminist critiques 23:443
    hematopoietic stem cells 23:443–444
    mammalian cloning 23:440–441
    oocytes 23:443
    science and technology studies 23:443
    state and regulation 23:442–443
    translational medicine, experimentation, and medical tourism 23:443–444
Stem change as a morphophonological mechanism 15:829
STEM disciplines 12:865
Stem-and-leaf displays 8:601
Stem-family household 11:229
Stemmer, B., works, *Handbook of Neurolinguistics* 16:679
Stenfors, Sari, on epistemic cultures 7:878
Stenographers, in interviewing 12:674–675
Step size selection, in stochastic optimization 23:475
Stepchildren 20:430–433
  and divorce revolution 20:430
  and union dissolution 20:432
Stepfamilies
  and cohabitation 4:118
  emotional development of children in 7:440–441
  and fertility 9:20
Stepfathers, fatherhood of 8:840–842
Stepfather–stepchildren relationships 20:431
Stephen, James Fitzjames 23:61, 25:75–76
Stepped wedge designs 9:130
Stepped-care models 6:203t
  and manual-based treatment 14:503
  for somatoform disorders 23:14
STEPPS *see* Systems Training for Emotional Predictability and Problem Solving (STEPPS)
Step-relations 8:813
Stepwise exploratory factor analysis (SEFA) 8:707
Stereotactic subcaudate tractotomy 23:18t
Stereotype content model (SCM) 12:413–414, 17:271–272, 18:841
  and inequality 11:909
  and prejudice 18:835
  of sexism 21:695
Stereotype threat
  and African-Americans 19:791
  ethnicity/migration 8:144–146
  in mathematics learning 4:71–72, 4:72t
  and prejudice 18:834–835
  and sexism 21:698
  and social identities 22:323
Stereotyped orientation of moral reasoning 7:317–318, 7:316t
Stereotypes 2:256, 17:803, 18:835, 18:841–842
  accuracy 23:448–449
  assessment of 23:447
  of beauty 2:384
  change 23:450
  conflicts 4:583–584, 4:587
  content 23:448
  defining 23:446
  development and diffusion 23:449
  functions of 23:447–448
  gender 9:684–685
  gender, culturally endorsed 9:747
  and immigration 11:677–678, 11:680
  influence on target group 23:450–451
  intergroup relations 12:413–414, 23:448–450
  mental representation of 23:446–447
  and public opinion 19:562–563, 19:565
  semantic differential 21:566–567
  in sexism 21:693–698
  for sexual harassment 21:723, 21:725
  social categorization 23:446
  social identity 23:447, 23:450–451
  stages involved in 23:449–450
  stigmatization 23:447, 23:450–451
  threat 23:450–451
Stereotyping 4:154, 19:859–860
  in advertising 1:215
  and essentialism 8:13–15
  and implicit social cognition 11:714–715
  of Latino gays and bisexuals 9:629–630
  race–crime relationship 19:843
  and racism 19:865–866
  self-categorization theory 21:457–458
  and social categorization 22:188–189
  stigma 23:462–463, 23:465–466
Sterilizations 8:219–220, 8:225–228, 8:231, 8:234–235, 8:240–245, 8:255–256, 9:77
  eugenic 8:249–251
  eugenics in the Americas 8:247–251
  nonconsensual 8:251
Stern, William 19:669

Sternberg, Robert 12:285
triarchic theory 12:285
Steroid hormones and vocal development 2:692–693
Steroids, neurosteroids 16:651–652
Steuart, James 14:644
works, *Philosophy of Right* 14:644
Stevens, Stanley Smith 21:605, 23:453–456
beginnings of experimental psyhology 23:453
on cross-modal matching 23:454
Fechner and 23:453–454
on hearing 23:455
impact on experimental psychology 23:455
on loudness 23:453–455
on magnitude estimation 23:454
on perception 23:453, 23:455
on power law 21:605–606
on psychoacoustics 23:455
psychophysical power lawof 23:453–454
on scales of measurement 23:454–455
on sensation 23:453–454
Stevenson, Robert B., works, *Handbook on Research in Environmental Education* 7:735
Steward, Julian H. 1:769, 7:622, 17:22
works
Handbook of South American Indians 20:386
The People of Puerto Rico 1:639
Steward Machine Company v. Davis (1937) 18:735
Stewardship theory 8:775
Stewart, K., on North American community 17:23
Stewart, Maria 8:906
Stewart B. McKinney Homeless Assistance Act 1987 11:159
STG *see* Superior temporal gyrus (STG)
STI *see* Simultaneous tilt illusion (STI)
Stiffness control 3:781
motor control 3:782
Stiffness ellipse, motor control 3:782
Stiftung, Bertelsmann 22:400
Stigler, George Joseph 20:172, 23:457–460
on analytical rigor 23:459
governmental activities of 23:457
on history of thought 23:458
personal history of 23:457
scientific contributions of 23:457–459
economic regulation 23:459
economics of information 23:458
essays, editorship, and reviews 23:459
intellectual history 23:458
organization of industry 23:458–459
price theory 23:458
works
The Calculus of Consent 23:459
The Citizen and the State 23:459
Economics of Information 23:458
The Economies of Scale 23:458
The Economist as Preacher 23:459
Essays in the History of Economics 23:458
The Intellectuals and the Marketplace 23:459
Memoirs of an Unregulated Economist 23:459
The Organization of Industry 23:457
Production and Distribution Theories 23:458
The Theory of Competitive Price 23:458
The Theory of Price 23:458
Stigma 15:228–229, 23:461–466, 25:493
addressing subtlety 23:465
class 23:465
contemporary levels of analysis 23:464–465
at individual level 23:464
at interpersonal/intergroup level 23:464–465
at macrosocial level 23:465
of criminal activity 7:81
disability 6:473, 6:475
discrimination 23:462–463, 23:465
in divorce 14:518
future frontiers 23:466
gastrointestinal disorders 9:607–608
groups in focus in the early twenty-first century 23:465–466
aging 23:465–466
disability 23:466
poverty 23:465
and inequality 7:895, 7:898
marginalization 23:463
mental illness 19:308, 19:310
racism 23:462–463
and recruitment 5:472
research 23:461–464
Allport's theory of nature of prejudice 23:462–463
early race theory 23:461–462
Goffman's stigma and spoiled interaction 23:463
influence of civil rights movement 23:463–464
social gospel 23:462
and sanism 15:201
sexism 23:464
social issues 23:462
stereotyping 23:462–463, 23:465–466
target's perspective 23:464–465
weight stigma 17:84–85
Stigmatization 13:592
of homosexuality in China 21:735–737
in psychiatry 22:977–978
in schizophrenia 21:14
and sexism 21:693
stereotypes 23:447, 23:450–451
Stillman, S., on butch–femme relationship 13:872
Stimulants 6:686, 19:464–468, 21:610
and attention-deficit hyperactivity disorder 3:882
Stimulation 24:533
and breast feeding 2:847
by computers 4:517
hypothesis 25:170
of insular cortex 23:251
Stimulation interval, kindling 13:33
Stimulation-facilitation approach, to language disorders 13:331
Stimuli 2:326, 14:367–368
addiction 1:97–98
agnosia 1:440–443
context dependent memory 4:782
Stimuli in a block 13:368
Stimulus characteristics, cardiac defense 23:541
Stimulus chimeras 23:294–295
Stimulus discrimination, and classical conditioning 3:764–765
Stimulus enhancement
animal social learning 22:403
Stimulus format 22:15–18
Stimulus generalization, and classical conditioning 3:764–765
Stimulus known exactly (SKE) model 24:309
Stimulus materials, developmental research 6:326, 6:329
Stimulus onset asynchrony (SOA) 16:540–542
coupled slack from 16:542
Stimulus Organism Reaction Kontingency Consequence Scheme (SORKC scheme)
suicidal behavior 23:677–678, 23:678f
Stimulus perspective on psychological stress 23:584
Stimulus sampling theory, mathematical psychology 14:804
Stimulus selectivity of cortical neurons 17:128
Stimulus-induced oscillations 16:635–636
Stimulus-induced response synchronization 16:637
Stimulus-model comparator theory 10:482
Stimulus-Organism-Response (S-O-R) model 4:739
Stimulus–response compatibilities 11:705
Stimulus–response framework 18:125
Stinchcombe, A., on assortative matching 2:56
STIRPAT *see* Stochastic version of the I-PAT identity (STIRPAT)
STIs *see* Sexually transmitted infections (STIs)
STL *see* Special Tribunal for Lebanon (STL); Student Team Learning (STL)
STM *see* Scanning tunneling microscope (STM); Short-term memory (STM); Systemic transactional model (STM)
STMs *see* Spatial–temporal models (STMs)
Stochastic analysis 14:632
Stochastic approximation 23:474–475
gradient approximation 23:475
step size selection 23:475
theoretical properties 23:475–476
Stochastic complexity 4:476
Stochastic effects on epigenetics 24:690
Stochastic geographic models 14:787
Stochastic grammar 19:9–10
Stochastic (Itô) integral 14:632
Stochastic inverse projections (SIP) 6:151–152
Stochastic largeness order 14:449–450

Stochastic modeling 13:692
Stochastic models 23:467–472, 24:347–348
accounting equations 23:468–469
characteristic feature of 23:467
individual unpredictability 23:467
Markov models 23:469–470
choice of categories in 23:470
deterministic equivalent of 23:470–471
discrete time and continuous time systems 23:470
discrete time closed model 23:469
open Markov systems with fixed input 23:469–470
open Markov systems with fixed size 23:470
practical issues with 23:470–471
nonlinear (diffusion) models 23:471–472
Poisson model 23:467–468
renewal models 23:471
sample surveys 20:886–888
seasonal adjustment methods based on 24:352–353
of social change 22:197
terminology and background 23:467
vacancy chains 23:471
Stochastic optimal control theory 6:709–710, 17:254–255
Stochastic optimization 23:473–481
constraints 23:473, 23:475, 23:478–479
future challenges 23:479
efficient methods for black box, observational data 23:479
efficient methods for large multistage problems 23:479
solution evaluation methods 23:479
multistage 23:477–479
bandit problems 23:477
convex multistage problems 23:478–479
Markov decision processes 23:478
objective function 23:473, 23:476, 23:478
randomness 23:473, 23:479
single-stage 23:473–477
data generation: constructive *versus* black box 23:474
data treatment: batch *versus* online 23:474
global metamodel optimization 23:476–477
gradient approximation 23:475
noise generation: exogenous *versus* endogenous 23:474
response surfaces 23:476
sample average approximation 23:474
step size selection 23:475
stochastic approximation 23:474–475
Stochastic orders 14:449–452
applications and illustrative examples 14:451–452
concordance ordering 14:450–451
convex function 14:450
distribution function 14:449–451
hazard rate order 14:450
likelihood ratio order 14:450
probability 14:449, 14:451–452
random variable 14:449–452
Stochastic pendulum equation 14:798–799
Stochastic population forecasts 18:596
Stochastic process model of mortality 15:846
Stochastic processes 14:631, 19:36, 24:316
mathematical psychology 14:814
time series cycles 24:335
Stochastic Processes and their Applications 24:345
Stochastic programming 23:478–479
Stochastic structure of time series observations 24:316
Stochastic version of the I-PAT identity (STIRPAT) 18:569
Stochastic volatility (SV) models 24:338, 24:360
Stock equilibrium, of structural unemployment 24:738–740
Stock market predictability 23:482–489
adaptive market hypothesis 23:487
asset pricing 23:483
efficient market theory and its critics 23:482
equity returns 23:486–487
evidence of short-run momentum 23:482–483
high risk–low return findings 23:485
from initial dividend yields 23:486
from initial P/E multiples 23:486–487
investment decisions 23:487–488
mean reversion 23:483–484
P/BV ratios and returns 23:485
P/E multiples and returns 23:484–485
portfolio management 23:482, 23:484–486
seasonal patterns 23:484
size effect 23:484
stock valuation 23:482, 23:485, 23:487
Stock markets 4:939–940, 4:944
venture capital 25:59
Stock of knowledge 4:200
Stock ownership 23:490
Stock valuation 23:482, 23:485, 23:487
Stöcker, Helene, League for the Protection of Motherhood and Sex Reform 9:644–645
Stockholders
ownership and control 23:490–494
board of directors 23:491
corporate control 23:490–491
corporate governance 23:490
executive compensation 23:493
governing board 23:492–493
government regulation 23:493
influence on top management and corporation 23:490–491
institutional investors 23:490–493
resolving separation of stock ownership from company control 23:490
shareholders, directors, and executives 23:493
top management 23:491–492
see also Shareholders
Stockholm Chamber of Commerce 12:470
Stockholm Declaration of 2003 20:535
Stockholm International Peace Research Institute (SIPRI) 8:34–35, 16:163
Stockholm Peace Appeal of 1950 17:641
Stock-outs, in supply chain management 23:711
Stocks 3:90
and flows 14:761, 14:763
Stoddard, Lothrop, works, *The Rising Tide of Color Against White World-Supremacy* 19:803
Stoicism 3:774–775
Stokes, Gale 16:287–288
Stoler, Ann, works, *Carnal Knowledge and Imperial Power* 8:900
Stoller, Paul 1:743
Stolper–Samuelson theorem 12:570
Stone, Clarence 4:365
Stone, John, works, *Alexis de Tocqueville on Democracy, Revolution and Society* 16:237
Stone, Richard 16:227–231
macroeconomic data 5:752
Stone, usage in ancient Egypt 1:694
Stone Age 14:260
Stone age economics 20:258
Stone knapping 11:255
Stone tool 14:260–261
function of 14:263
Stonewall riots 24:551–552
drag culture during 13:877
gay/lesbian movements during 9:632–633
Stoneworking, in ancient Egypt 1:694
Stop Online Piracy Act (SOPA) 13:524
Stopping rules 5:906
heuristics 5:913
Storage
quantitative memory models 15:162–164
short-term memory 21:901–908
Storage corpus 4:951
Storage of information 13:681
Storch, Johann 2:729–730
Storehouse metaphor 25:177
Stories
love stories 14:379–380
and social networks 16:563–564
social structure as 22:717
Storm and stress model 6:10
Storming 22:850–851
Stormtroopers (SA) 16:258
Story network analysis 16:196
Story Resistance Task 10:397–398
Storytelling 10:111, 16:194–197, 16:204–205, 16:207, 16:221, 18:681
and communities 4:333
in organizations 16:196–197
source of educating children 11:763
Stour, F., works, *The City Reader* 24:940–941
Stout, W., evaluation of essential unidimensionality 6:437–438
Stovel, K. 16:211
on narrative networks 16:212–216
STPD *see* Short-term psychodynamic therapy (STPD)
STR *see* Shared traumatic reality (STR)

Strabismus 5:32
Strabo of Amaseia 7:720
Strafprozessordnung 18:229–230
Straight-line theory 2:83–84
Strain 14:475–476, 25:537–538, 25:733–738
  caregiver 3:135, 3:137
  child health psychology 10:684
  and coping 4:902–904
  and crime 19:839
  definition 23:547
  and delinquency 6:69–70
    extensions or reformulations in 6:71–72
  due to job *see* Job strain
  family 23:702
  and general strain theory 23:496–499
  and homicide 11:163
  physical health 23:702
  psychological health 23:702
  types 14:475–476
Strain theories 1:715
  and crime 23:495–500
    classic theories 23:495–496
    community 23:495, 23:497, 23:498–499
    delinquency 23:496, 23:497, 23:498
    drug use 23:495, 23:497
    general strain theory 23:496–499
    looking back and looking forward 23:499
    social control theory 23:496
    social learning theory 23:496
  of crime 22:238
  of property crime 19:164
Strain-based conflict, and work–family conflict 8:759
Strand, Michael, on practice theory 7:415
Stranding errors and multiple levels of processing 23:231
Strang, David, on institutions 17:392
Strange Situation procedure (SSP) 2:137–139, 2:792, 3:396, 12:230, 16:470
Stranger abuse, of elders 7:329
Stranger adoption 1:137
Stranger homicides 11:166
Strangers
  criminology of 5:248–249
  danger 21:673
  fear of 8:445, 25:802–803
Strasser, Garold, on networks and creativity 5:467
Strasser, Gregor 16:258
Strata software 13:386, 13:388
Strategic action fields 9:146–147
Strategic adaptation 23:501–507
  adaptive organizations 23:502–503
  corporate entrepreneurship 23:503–504
  environmental contingency 23:501–502
  fit 23:502
    perspectives 23:501–502
  organizational perspectives 23:502
  renewal and dynamic capabilities 23:504
  responsive strategy
    conduct 23:504–505
    making 23:505–506
Strategic Arms Limitation Talks (SALT) 1:976–977
Strategic Arms Reduction Talks (START) 1:977
Strategic asset-seeking investment 9:325
Strategic automaticity, and motivation/actions 15:891–892
Strategic behavior, hierarchies and markets 10:842
Strategic choice of structural contingency theory 23:612
Strategic climates 17:329
Strategic compensation 17:614
Strategic competence 4:296
Strategic complimentarities, social network model with 16:573–574
Strategic context determination 23:512
Strategic decisions, in Cold War 4:130–131
Strategic Defense Initiative (SDI) 1:977
Strategic delay, and procrastination, differences between 19:73
Strategic dissonance 23:510
Strategic educational planning 7:262–263
Strategic elites 7:392–393
Strategic ellipse 7:636
Strategic Entrepreneurship Journal 7:684
Strategic essentialism 2:732, 8:155
Strategic formation, in social networks 19:910–911
Strategic game 9:583
Strategic group formation, in Southeast Asia 23:72–73
Strategic helping 10:760
Strategic ignorance, as organizational source 11:581–582
Strategic Initiative for Developing Capacity in Ethical Review (SIDCER) 20:514
Strategic innovation theory perspective 12:173
Strategic intelligence, technology assessment for 24:127–128
Strategic interaction model 10:277
Strategic issues management approaches 23:518
Strategic learner 13:714–715
Strategic learning model 13:714–715
  impact on a student's acquisition 13:714
  variables impacting 13:714–715
Strategic management 23:508–514
  business models and 3:34
  distinctive (core) competence 23:509–512
  and diversification 6:586
  economizing 23:511–512
  emergent (autonomous) strategy 23:509, 23:512–513
  genesis and development 23:508
  implications 23:511
  induced strategy 23:512–513
  industrial organization 23:508–509
  integrative conceptual framework 23:509–511
  organizational performance 23:511
  resource-based view 23:508–509
  role of internal selection environment 23:510
  strategic dynamics 23:509–510
  strategic position view 23:508–509
  strategizing 23:511–512
  strategizing, economizing, and performance 23:511–512
  strategy diamond framework 23:509, 23:511
  strategy implementation 23:513
  strategy-making process 23:508, 23:511–513
  theoretical foundations 23:508–509
Strategic Management Journal (*SMJ*) 23:508
Strategic marketing 14:555–558
Strategic memory 21:897–898
Strategic models, of confabulation 4:561, 4:564–567
  clinical evidence in support of 4:567
  imaging evidence for 4:566
Strategic niche management 22:875
Strategic planning 24:842
  collaboration approach 23:519–520
  competitive analysis 23:520
  contract approach 23:519
  as framework for innovation 23:518–519
  Harvard policy model 23:517
  history and development 23:516
  logical incrementalism 23:518
  portfolio models 23:520
  for public and nonprofit organizations 23:515–521
  regional 20:129–133
  stakeholder management approaches 23:517–518
  strategic issues management approaches 23:518
  as system 23:517
Strategic position 23:508–509
Strategic power 17:737
Strategic recognition 23:510
Strategic reflexivity, in organizations 12:173
Strategies 4:47–48, 9:573, 12:148, 15:282, 15:285–286
  contingency of 23:610
  defense economics 6:3–5
  of developing writers 25:769
  entrepreneurial 7:684
  imitation by infants/young children on 11:623
  in mathematics learning 4:66–67, 4:69–70, 4:72–73
  mixed 9:575–576
  in negotiation 16:425
  optimal 9:576
  persons/environments control 24:118–121
    advanced technology 24:123
    capital punishment 24:120
    crime elimination program 24:120
    dimensions 24:120–121
    offender weakening or incapacitation 24:119–120
    offense/offender/target identification 24:120
    place or activity exclusions 24:120
    primary direct preventive efforts 24:118–119

Strategies (*continued*)
secondary engineering 24:119
social intervention 24:119
stand-alone mechanical solutions 24:123
target devaluation 24:119
target insulation 24:119
target/facility removal 24:119
technical solutions 24:123
political campaigns 3:67
pure 9:576
of social movements 22:439–440
threat 9:578
Strategies of control 17:342–343
Strategist 23:522–524
Strategizing 23:522–526
Clausewitz on 23:523
constraints 23:523–525
engaging others 23:524–526
engaging the unique 23:523–524
entrepreneur 23:524
judgment 23:523–526
Knight on 23:526
practice 23:523–524
profit 23:525–526
rhetoric 23:525
Simon on 23:522–523
strategy field's history 23:522–523
uncertainty 23:523–526
Strategizing arcs 23:525
Strategy instruction, in writing 25:769–770
Strategy use
cross-cultural study of education 5:329–331
Strategy-making process 23:512–513
Strathern, Marilyn, works, *After Nature* 7:594
Stratification 5:325–326, 11:896–897, 19:144, 21:529, 22:338
acquired economic status 11:896
ascribed honor 11:896–897
class 3:747
dimensions 11:896
and divorce 6:608
equality of opportunity and 7:903
feudalism and estates 11:896
and inequality 11:908–910, 11:912
and intersectionality 12:657, 12:659–661
and life course 14:83
and neighborhood effects 16:443, 16:446–447
and religion 20:292–297
social 22:338–344
in vocational education 25:234, 25:236–237
Stratification (Soviet) 20:851–852
class 20:852
status 20:852
Stratified chronology 24:411
Stratified Cox model 23:774
Stratified medicine 17:925
Stratified random sampling 23:187f
Stratified reproduction 20:452
Stratified sampling 5:321
Stratigraphy
in archaeology 3:562–564
constructing chronologies 3:564
in classical archaeology 3:762
in geoarchaeology 10:53–56
Stratton, Frederick John Marrian, and Fisher 9:222
Stratum indicators 13:388
Strauss, Anselm Leonard 21:265, 23:527–532, 25:678
on chronic illness 23:528–530
management work 23:529–530
trajectories of illness and medical work 23:529
web of negotiation and social worlds 23:530
on disability 23:529–530
on grounded theory and thinking through fieldwork 23:530–531
influence 23:531
intellectual genesis of 23:527–528
on medical sociology 23:527–529, 23:531
on medicine and illness 23:528
death 23:528
psychiatric institutions 23:528
on occupation 23:527, 23:529, 23:531
on profession 23:527–529, 23:531
on qualitative analysis 23:527, 23:530–531
on social psychology 23:527, 23:531
on sociology of action 23:527–530
on symbolic interactionism 23:527
works
Continual Permutation of Action 23:527, 23:530
The Discovery of Grounded Theory 10:403–404, 23:530–531
Mirrors and Masks: The Search for Identity 23:527
Negotiations: Varieties, Contexts, Processes and Social Order 23:530
Qualitative Analysis for Social Scientists 10:404, 23:531
The Social Organization of Medical Work 23:529
Strauss, David Friedrich 10:736, 14:643
works, *Life of Jesus* 14:643
Strauss, Leo 18:469, 18:486–487
works, *Natural Right and History* 4:690
Straussians 18:486–487
Straw polls 18:459
Strawson, P.F. 15:546
Stream of consciousness theory (James) 13:702
Streams, emotional intelligence 7:508
Street centerline data 10:61
Street children 3:456, 23:533–538
agency 23:536
child-centered research 23:534, 23:537
definitions 23:533–534
gender 23:536–537
history 23:534
participatory research 23:534–536
poverty 23:533–534
responses and solutions 23:537
human rights discourse 23:537
interventions 23:537
participatory methodologies and partnership ethics 23:537
statistics 23:534–536
improving localized counting 23:535–536
research gaps 23:534–535
street girls 23:536–537
youth homelessness 23:534, 23:536
Street connectedness 23:535
Street corner society, and empirical social research (ESR) 7:562
Street crimes 5:182
Street culture 16:327
Street design measures 24:825–826
Street gangs 9:594
Street girls *see* street children
Street Transvestite Action Revolutionaries 24:552
Street View technology 14:567–568
Street-level bureaucrats 1:794
and antisocial behavior 1:794
Streets 12:777–778
Strehler-Mildvan's model of mortality and aging 15:845–846
Streitberg, Wilhelm, works, *Geschichte der indogermanischen Sprachwissenschaft seit ihrer Begründung durch Franz Bopp* 14:216
Strength analysis, in feminist psychotherapy 8:932
Strength and vulnerability integration (SAVI) model 7:492–493
Strengthening a behavior 7:240
Strengths and Difficulties Questionnaire 22:944–946
Strengths-based approaches, for learning disability 13:687–689
Strengths-based interventions, for disasters 7:428, 7:431
Strengths-based practice, holistic approaches 12:546
and international social problems 12:543
Stress 1:401, 1:805, 7:607, 7:609, 9:819–820, 9:823, 15:103, 16:40, 17:337–338, 22:148–149, 22:758, 22:805, 23:333, 23:551–555, 23:557, 24:403, 25:539
acculturative 1:70
adaptation by hypothalamus 11:486–487
bioarchaeology 2:600–601
and breast feeding 2:847–849
Broadbent on 2:856–858
busyness and 3:51–54
and cardiac defense 23:539–545
in cardiac disease 3:111–112
commuting 24:627
and coping 4:888–889
coping with 4:895–896
and personality 17:794–795
coronary heart disease 4:919
work stress 4:919
dyadic coping 11:593–594
effect in newborn 23:557
risk of autism 23:557
effects on childhood development 23:570–576
abuse 23:570, 23:572–574

and development 23:572
incidence of stress-inducing incidents 23:570–571
psychobiology of stress 23:571–572
resiliency and posttraumatic growth 23:572–573
trauma 23:570–574
working with children experiencing stress 23:573–574
environmental 7:804–808
in functional GI disorders 9:608–609
and gastrointestinal disorders 9:606
and gender 9:705–707
and health 23:551–552
disease outcomes 23:553
heart disease 23:551–552
immune function 23:552
wound healing 23:552
and health behaviors 7:498
and hypertension 11:454–456
and infectious diseases 12:32–35
in inflammatory bowel disease 9:608
and interoception 12:617–618
accuracy during and after acute laboratory stressors 12:617
chronic stress 12:618
stress hormones 12:617–618
and mental disorder 15:206–207
negative effects of 22:805
and oral health 17:281–282
within organizations 22:805
and panic disorder/agoraphobia 17:476
prevention programs 10:672
psychological *see* Psychological stress
psychological adaptation to childhood cancer 3:459
and psychoneuroimmunology 19:445–446
and psychosomatic disorders 19:485–489
and recovery following acute cardiac illness 4:921
implications for patient care 4:921
resilience 20:555–559
and self-efficacy 21:506, 21:508
and sleep 22:60
teachers 24:72–74
and transient global amnesia 24:567
at work 23:546–550
and burnout 2:929–931
concepts 23:547–549
coping 23:547–548
demand–control model 23:548
economic globalization 23:546
effort–reward imbalance model 23:548–550
empirical evidence 23:548–549
fairness 23:548
future challenges for research 23:549–550
labor policies 23:549
occupational hazards 23:549
organizational justice model 23:548–549
person–environment fit 23:547–548
practical implications 23:549–550
promotion 23:546–547, 23:549
reciprocity 23:548
resources 23:548
role 23:546–547
social policies 23:549
theoretical models 23:547–549
work-related 12:132
see also Coping; with job stress; Work stress
Stress-1 16:41, 16:43
Stress hormones, and interoception 12:617–618
Stress inoculation training (SIT) 8:605–611
conceptual model of 8:607
implementation of intervention 8:607–608
nonclinical and clinical applications of 8:608
Stress management
and coping 4:903–904
coronary heart disease 4:921
and mindfulness pracatices 15:552–553
occupational 17:147–152
Stress management interventions (SMIs) 17:148
CBT-based 17:150
occupational stress 17:149
Stress management programs, in behavioral medicine 2:468
Stress management techniques
for asthma 20:578
for chronic obstructive pulmonary disease 20:578
Stress of migration 5:474
Stress process 22:540–541, 22:541f
Stress psychobiology 23:577–578
Stress reaction, and HPA axis 11:476–477
Stress response, epigenetics of 7:842–843, 7:843f
Stress Scale 22:389
Stress shift 23:718
Stress system reactivity 23:580
Stress theory 18:707
families as educational settings 8:736
Stress values, multidimensional scaling (MDS) 16:43
Stress–diathesis model 10:676–677
Stressful experiences at work 23:547
Stressful life events (SLEs) 2:535
coping 4:887–888
during divorce 14:521–522
obsessive-compulsive disorder 17:121
Stress-health model, caregiving 3:136–137
Stress–illness relationship 10:676–677
Stress-induced transient myocardial ischemia 4:921
Stressor exposure duration, cardiovascular reactivity 3:119
Stressors 4:896–899, 25:537–539, 25:733–738
cardiac disease 3:111, 3:113–115
childhood cancer 3:460t
coping 4:887–890
definition 23:547
environmental
characteristics of 7:805
and effects, variables intervening between 7:806
occupational stress management 17:147–150
Stress-related mental health 22:389
Stress-related models, for social networks 22:366
Stress–response disorders 2:538
Stria terminalis 14:127
Striate cortex, and visual mental imagery 25:163–164
Striate cortical function 25:159–160
Striated space (geography) 16:567–568
Striatum, and spatial navigation 23:166, 23:168
Strict approach to enforcement 13:573
Strict conditional 5:95–96
Strict cyclicity 16:4
Strict liability, torts in law and 24:431–432
Strict liability regimes 7:781
Strict order (mathematics) 17:525
Strict scrutiny, of discrimination 6:530–531, 6:533
Strict semantic assignment systems, of gender 9:795–796
Stricter penalties, in United States 16:250
Strictly affiliated information variables, in auctions 2:207–208
Strictly affiliated private values, in auctions 2:207
Strictly monotone transform of a generalized stable survival function 5:902
Stridhan 6:644
"Strike Hard" Campaigns 5:876–877
Strikes 6:557–558, 23:588–591
collective bargaining legislation and 23:591
condition 23:589
economists' perspective 23:588
empirical analysis of strike activity 23:590
equilibrium model 6:557
fluctuations in strike activity 23:590–591
incidence 23:591
'joint cost' model of 23:590
nonequilibrium model 6:557
private information theories of 23:588–589
seasonal patterns in 23:591
social protest 22:506, 22:508–509
Strip for Jacky campaign 21:794
Strogatz, Stephen H., on small world structures 5:467–468
Stroke
and hemispatial neglect 10:766
and physical activity 18:88
recovery after 13:329–330
Strong Interest Inventory, for vocational interests 25:239
Strong practice theory 7:415
Strong Program 5:504, 6:245, 18:101, 21:192–193, 21:264, 21:309, 22:230, 22:961, 24:136, 24:675
causality 23:592
ideological limitations on 23:593–594
converging evidence 23:595
criticisms 6:245
crucial experiments 23:594
in cultural sociology 5:618–620
defined 23:592

Strong Program (*continued*)
false charge of idealism 23:592–593
impartiality and symmetry 23:593
inductive strategies 23:595–596
mathematics analysis 23:596–597
meaning finitism 23:594–595
reflexivity 23:593
relativism 23:595
social, role of 23:593
Strong records 16:900
Strong situation, organizational culture as 17:698
Strong social constructivism 22:231–232
Strong stress meter 13:266–267
Strong sustainability 23:824
Strongest Families 2:479–480
Strongly efficient network 16:574
Strong/weak quantifier 19:708
Stroop color-word test 3:118
Stroop effect 23:264
Stroop model of cognitive control 4:35–36, 4:35f
Stroop task 4:35
Stroop test, dysexecutive syndromes 6:726
STRs *see* Short tandem repeats (STRs)
Structural adjustment
in Africa 23:598
and health 10:198–203
Structural adjustment policies 1:282, 1:284–285, 18:181
Structural adjustment programs (SAPs) 7:198, 10:186–188, 18:724, 24:967, 24:969–970
in Africa 23:598–603
Structural alignment 21:580
Structural analysis, network analysis 16:519
Structural and structuralist theorizing 22:959–960
Structural approaches to prosocial behavior 1:581
Structural assumptions
for nonparametric identification of causal inference 3:256–258
common support 3:257–258
ignorability 3:257
stable unit treatment value assumption 3:258
temporality 3:258
unconditional ignorability 3:256–257
Structural balance theory 23:605–608
balance theory 23:605–606
balanced and near-balanced states 23:606–607
block models 23:606–607
empirical support 23:608
friendship 23:605
group formation 23:606
incidence matrices 23:605–607
origins and development of 23:605
polarization 23:606–608
signed graphs 23:605–606
structural theorems 23:606
studying adjustment toward balance 23:607–608
Structural barriers to negotiation 16:420–421
Structural brain asymmetry 2:809
Structural brain development 16:600–601
in infancy 4:58–60
Structural brain disease, electroencephalography for 7:362
Structural brain imaging, cognition 4:99
Structural business statistics 5:707
Structural causal effects 18:454
Structural centrality and information diffusion processes 16:525–527
Structural characteristics, alternative schools 21:39–40
Structural cohesion 3:906–908
Structural connectivity, in ADHD 2:158
Structural conscious and unconscious knowledge 13:87
Structural constraints 13:164–167
and enabling structures 13:164–165
Structural contingency theory 17:417–418, 23:609–614
centralization 23:609–611
configurationism or Cartesianism 23:611
contingency imperative or strategic choice 23:612
current and emerging directions in theory and research 23:612–613
asymmetric effects on performance of overfit and underfit 23:613
fits 23:613
isoperformance and heteroperformance 23:612–613
links to databases and software 23:613
matrix structures 23:613
new variables 23:613
selection 23:613
current knowledge 23:609–610
organizational size 23:610
strategy 23:610
uncertainty 23:609–610
from curvilinear to linear fits 23:611
definitions 23:609
from differentiation to interdependence 23:611
equifinality 23:612
integration 23:609–611, 23:613
intellectual background 23:609
from single to multiple contingency fits 23:610
specialization 23:609–610, 23:612
structures fitting contingencies 23:611–612
technology 23:610–611
Structural couplings 13:555
Structural determinist approach 22:723
Structural drift, in economic growth 7:48–49
Structural elements, of social system 22:745–751
Structural embedding, of mobile communication in society 15:633
Structural equation modeling/models (SEM) 6:322, 7:221, 7:226, 8:686, 10:444, 14:812, 19:729–730, 20:245, 22:197, 22:548, 23:615–620, 24:642, 24:691
in causal analysis 3:248
developments and extensions 23:619
in forced-choice personality assessment 17:844
general 23:615–616
for genetic studies of behavior 9:934
graphical models 23:619
growth curve models 23:619
identification 23:616
I/E model 12:428
latent class models 23:619
linear 13:394–396, 13:394f
parameter estimation 23:616–617
complex sampling designs and missing data 23:617
estimators for categorical observed variables 23:617
maximum likelihood estimators 23:616
model-implied instrumental variable estimators 23:616–617
testing 23:617–619
alternative methods of assessing model fit 23:618–619
global model fit 23:617
hypotheses regarding individual parameters 23:617–618
measures based on comparative fit to a baseline model 23:618
measures based on errors of approximation 23:618
use of term 23:615
without latent variables 13:396f
Structural equations 21:980
estimating a complete system 21:982
estimating one equation at a time 21:981–982
Structural equivalence 5:319, 16:564–565, 22:457, 23:623–624
blockmodeling 23:622–623, 23:625, 23:622f
clustering 23:622–623
cohesion 23:624–625
definition 23:621–622
empirical applications 23:624–625
graph theory 23:621
history and significance 23:623–624
homomorphism 23:622
methods 23:622–623
network analysis 16:519, 16:521
position 23:623
regular equivalence 23:621, 23:623
roles 23:623
similarity 23:624
social homogeneity 23:623–624
and social networks 16:557
Structural explanation theories 11:14
Structural explanations 6:412–413
Braudel's structural theory 11:14
structures of meaning 11:14
Structural family therapy (SFT), Minuchin's 19:505
Structural formulation of psychoanalytic treatment 19:339–340

Structural functionalism 3:6, 9:533, 9:538–539, 14:415, 20:307–308, 23:428–431
anthropological traditions 2:495
Coser on 5:36–37
and nationalism 16:291
Deutsch 16:291–292
Gellner 16:292
Nisbet on 16:834
Radcliffe-Brown's 1:767
reactions to 1:768–769
of socialization and education 22:863
view of schools 21:56–57
Structural genomics 2:625, 9:1017–1019
Structural history 11:104
Structural holes 5:467
Structural hybridism 11:451–452
Structural identity of explanation and prediction 10:774
Structural imaging studies, for ADHD 2:155–156
Structural individualism 15:311
and analytical sociology 1:671
Structural inequality
critiques of 7:898–899
of treatment 7:897–899
Structural inertia theory 6:891–892
Structural inheritance systems 7:833
Structural injustice 22:398–399
and social justice 22:398–399
Structural integration 2:85
Structural interdependency, and groups 10:434–436, 10:438–439
Structural interventions, social work 3:471–472
Structural invariance 9:124–127
Bayesian analysis 9:124–127
density 9:125–126
example 9:124
parameter value 9:125–126
probability 9:126–127
scalar 9:124–125
Structural lag
education and learning 7:137
Structural linguistics 14:199, 14:226
Structural logic of kinship terminologies 13:63–64
Structural macroeconomic models 9:320–321
Structural magnetic resonance imaging 14:439
principles 14:439
Structural methodological individualism 14:416–418
Structural modeling 14:850
Structural models 6:166, 22:471
spatial models 24:470
Structural nested models (SNM) 15:78
Structural neuroimaging
for psychopathy 19:452
and virtual spatial navigation 16:352
Structural phonology 12:782
Structural poor 18:729
Structural power 18:763–764
Structural properties of kinship terminologies 13:65
Structural racism 19:799
Structural rules for nonstandard reasoning 16:928
Structural shaping, of emotions 7:535–536
Structural social work 14:35–36
Structural solidarity 12:403, 12:405
Structural stability, and adolescent personality 17:864–865
Structural support, disease recovery 22:736
Structural theories 14:113–115, 20:630, 23:7, 23:606
of trade policy 24:486
hegemonic stability theory 24:486
(neo)Marxian approaches 24:486
Structural Time Series Analyser, Modeller and Predictor (STAMP) 24:352–353
Structural time series model 24:355–356
Structural unemployment 24:738
cyclical unemployment 24:738, 24:741–742
definitions 24:738
disequilibrium phenomena 24:738
economic fluctuations 24:738
flow equilibrium 24:740
frictional unemployment 24:738–740, 24:742
microeconomic foundations 24:740–741
persistence 24:741–742
policies influencing 24:742–743
stock equilibrium 24:738–740
Walrasian system 24:738–739
Structural variables 22:413
Structural violence 10:633–634, 17:622, 17:629, 22:398–399, 25:117, 25:122
and social justice 22:398–399
Structural-constraints strand 13:164
Structuralfunctionalist, and political anthropology 18:335–336, 18:340
Structuralism 1:748–749, 4:201, 12:121, 12:782, 12:784, 13:36, 13:38, 13:701–702, 14:793, 20:651, 22:716, 22:997–998, 23:201, 23:626–631
and advertising 1:216
and anthropological comparison 4:411–412
anthropology 23:626–627, 23:630
in anthropology 1:768
Barthes on 2:359–360
British 23:630–631
Chomsky on 3:527
classifications 23:627–629
and cultural sociology 5:498–501
and cultural studies 1:935
and deconstruction 5:929–930
discourse after 6:503–504
and discourse studies 6:513
French 23:626–627
and functional analysis 9:537
genetic 23:627
history 23:627–630
legacy of 23:630–631
legal reasoning 13:776
of Lévi-Strauss 13:907, 23:627–630
kinship and group structures 23:628–629
order structures 23:629
savage thought 23:629
transformations and continuity 23:629–630
and mathematics 23:627
metaphors and metonymy 15:291
mythology 23:627–630
Native American tricksters 24:649–652
and psychology 19:413
and revolution 20:622–623
Saussure on 20:924–925
semantics 13:931–933
distributional semantics 13:932–933
lexical fields and componential analysis 13:931–932
relational semantics 13:932
and semiotics 21:594
social structure 22:722–723
in wider sense 23:627
working classes 25:695
Structuralist analysis of myths 9:359–360
Structuralist emotion theory 12:2
Structuralist feminism, social work 8:936–937, 8:939
Structuralist linguistics 15:826
Structuralist phenomenology 20:345
Structuralist political economy interrogation 5:255
Structural-phenomenological method 19:492
Structuration 10:86–87, 10:150
Giddens on 10:135–138
and individual-collective relationship 11:817
theories 10:787
Structuration process 17:377
Structuration theory 14:416, 20:84
of regional geography 20:124
Structure 12:148
and actors, in social change 22:195
assumptions 22:195–196
of education 7:147
and equations, symmetry relationship between 14:795–798
and factions 8:685
family 8:766
fitting contingencies 23:611–612
Giddens on 10:135–136
and labor history 13:171
Lévi-Strauss on 13:906–907
literature 14:554
multisensory integration among 16:101
narrative 16:206–207
of sign language 21:928–929
social networks 5:467–468
of survey questions 23:766–767
STRUCTURE (algorithm) 19:828
Structure and agency
of institutional logics 5:551–552
peasants 17:671
Structure of groups 10:424–425

Structure of intellect (SOI) model of intelligence 12:292
Structure of issue evolution 12:766
Structure of knowledge 25:341–344
Structure sensitivity 21:623
Structure theorem 16:525
Structure-centered approach, of Sapir–Whorf hypothesis 20:905
Structure-conduct interventionist period, antitrust 1:798–799
Structure-conduct-performance (SCP) paradigm 14:550
Structured academic controversy, in social studies education 22:732
Structured anamnestic questionnaire, hypoglycemia awareness 6:338t
Structured Clinical Interview for Dissociative Disorders 1:652
Structured Clinical Interview for DSM-IV 15:182–183
Structured dependency, in adulthood 1:182
Structured deposition, archaeological concept of 20:724
Structured diaries 21:544
  for measuring self-regulated learning 21:544
Structured focused comparison, for case studies 3:208, 3:210–211
Structured immersion (bilingual education) 2:582
Structured Interview for the Five-Factor Model of Personality (SIFFM) 9:233–234
Structured interviews
  for five-factor model of personality 9:233–234
  for measuring self-regulated learning 21:544
Structured methods of data collection 22:545
Structured movement systems 5:666
Structured observation 22:545
Structured Photographic Articulation Test II 3:796
Structured Photographic Expressive Language Test 3 3:796
Structured representations 21:580
Structured structures 10:487
Structured Team Learning methods 4:883–884
  Cooperative Integrated Reading and Composition (CIRC) 4:884
  IMPROVE 4:884
  Peer-Assisted Learning Strategies (PALS) 4:884
  Student Team Learning (STL) 4:883
  Student Teams-Achievement Divisions (STAD) 4:883
  Team Assisted Individualization (TAI) 4:884
  Teams-Games-Tournament (TGT) 4:883
Structured treatment, manual-based 14:502
Structure–function relationships 2:692
Structures (linguistic units or constructions) 23:891–894
Structuring society, generation as principle for 9:855
Struggles 11:575
  Honneth on 11:178
  over land 13:231
  spatiality of 13:165–166
Stryker, Roy E. 18:84
Strzelecki, P.E., works, *Physical Description of New South Wales* 7:750
STS *see* Science; technology; and society (STS); Science and technology studies (STS); Sociotechnical systems theory (STS); Superior temporal sulcus (STS)
STs *see* Specialized towns (STs)
STSA *see* Sociotechnical systems approach (STSA)
STSEQ *see* Science Teachers' School Environment Questionnaire (STSEQ)
Stuart, Charles 7:649–650
Stubben, Joseph 24:863
Stuchka, Peteris 22:844
Student achievement 21:103
  ecology of 5:342–343
  see also Academic achievement
Student Achievement Guarantee in Education (SAGE) program 3:753
Student assessment data 5:773–774
Student characteristics 21:121
Student engagement 1:31
student enrollments 5:771
Student evaluations of teaching (SET) 7:24
Student exchanges, international social work 12:562–567
Student learning 8:614, 21:92–94
Student motivation 13:714
Student movements, in 1960s 25:837–838
Student Non-Discrimination Act 21:88
Student Nonviolent Coordinating Committee (SNCC) 3:693, 5:84, 25:815–816
Student outcomes
  and teacher behaviors 24:44–50
Student participation, and education 7:146–147
Student progress
  assessment of 24:48–49
    feedback 24:49
    formative assessment 24:48
    progress monitoring 24:49
  and critical reflexivity 5:282–283, 5:285
Student rebellions, youth 9:866
Student self-assessment 3:821
Student sit-in movement 3:692
Student Team Learning (STL) 4:883
Student Teams-Achievement Divisions (STAD) 4:882–883, 10:416–417
Student-centered psychotherapy 20:738–739
Studentification 10:47
Students 18:599
  American studies 1:615
  of Australian Aboriginals 15:901–903
  economic education
    evaluations of teaching 7:24
    performance 7:24
  impact of teacher leadership on 24:61–62
  implications, school environment for LGBTQ/sexual minority youth 21:89
  individual rights 3:480
  and migration in Europe 15:379, 15:381
  reading and learning, using writing to support 25:770
  as soldiers 25:365–366
  with special writing needs 25:770
  supervision checklist for 15:236
  VET schools, postponing tracking 25:248
Students Against Drunk Drivers 15:590
Students' age 21:45
Studia humanitatis 20:427
Studies of adult self-rated health (SRH) 12:440
Studies of reliability 20:241
Studies-of-work 21:195
Studies-of-work, Garfinkel on 9:600–603
Study of Assets and Health Dynamics among the Oldest Old (AHEAD) 5:698
Study of genocide 10:6–8
  discipline 10:6–7
  subjects of research 10:7–8
Study of Mathematically Precocious Youth (SMPY) 7:161
Study of societies, and AGIL schema 1:388–389
Study of Statewide Education Programs 3:835
Study of Terrorism and Responses to Terrorism (START) 5:837
A Study of Values (SoV) questionnaire 25:42
Study on global AGEing and adult health, (SAGE) 2:660–661
Studying sideways 8:591
Studying subjective well-being (SWB) 23:643
  around world 23:643–644
    desirability of happiness across cultures 23:643–644
    experience of happiness across cultures 23:644
  cultures differences in 23:644–646
    gross domestic product (GDP) 23:644
    happiness, definitions of 23:645
    happiness, strategies in 23:645
    individualism-collectivism 23:644
    materialism 23:645
    measurement 23:645–646
    norms regarding happiness 23:644–645
    positive emotions 23:644–645
  definition 23:643
  future research 23:646
Studying up, elites 7:387–388
Study-skills training, test anxiety 24:248
Stuff sociotechnology 24:137–138
Stumpf, Carl 10:757, 19:413
Sturtevant, Edgar H. 2:705
Sturzo, Luigi 18:437–438
Stuttering 12:420, 21:939
  aberrant cerebral laterality of language 9:269

acquired (AS) 9:272
animal modeling of 9:269–272
brain imaging 9:267–268
cerebral volumetric differences with nonstutterers 9:268
diffusion tensor imaging (DTI) 9:268
magnetoencephalography (MEG) 9:268
PET and fMRI experiments 9:268
clinical features of 9:266–267
definition 9:266
difficulties 9:266
emotional stress and 9:266–267
genetic predisposition 9:267
gesture inoculation against 10:112
locus of dysfluency 9:266
'secondary characteristics' of 9:266
strategies to overcome dysfluencies 9:266
speech-related maneuvers 9:267
theories of 9:268–269
cerebral hemisphere laterality 9:269
STV *see* Single transferable vote (STV)
Stychin, Carl F., on sexual citizenship 21:730–731
Style and self-regulation in learning 4:114
Style construct, revisionist theories and problematizing 4:113–114
Style differences, cognitive styles 4:110
Styles 13:291
coping with chronic illness
individual differences in 3:551–552
monitoring and blunting 3:551–552
fashion 8:835–838
film 9:166–168
and language 13:281–282, 13:285–287, 13:282t
parenting 19:245–248
populism 18:609
public policy 19:583–584
and social networks 16:565
Stylized estimates, for time use 24:379–380
Suba 8:177
Subaltern 16:457, 24:819–820
Subaltern classes 11:41
Subaltern consciousness 23:634
Subaltern counterpublics 3:705
Subaltern historiography 23:632
Subaltern history 23:632–636
against elitist historiography 23:632–633
gramsci in India 23:632
imbrication of elite and subaltern politics 23:633–634
modernities 23:634–635
reorientations 23:634
rethinking political 23:635
Subaltern internationalism 10:189
Subaltern studies 11:42–44, 18:647–648, 18:658–659, 23:632, 23:634
Subalternity 16:184, 16:186
Subarctic 1:927–929
Subcategorization information 13:943
Subcategory of a lexical item 13:943
Subclasses 15:603
Subcortical motor-control structures 15:966
Subcortical neurocircuitry dysfunction in ADHD 2:157–158
Subcortical structures 1:644
Subcultural diversity, and fear of crime 8:856–858
Subcultural theories 5:146
of race–crime relationship 19:840
Subculture studies, and hegemony 10:745
Subcultures 23:637–642
Subcultures
corporate 4:923–924, 4:926
and delinquency 6:69–70
homosexual 9:638
sociology of 23:637–638
definition of 23:637–640
ecosystem of 23:640
methodology 23:640–641
youth culture 25:816–817
Subelites 7:397, 7:399–400
Subfederal governments, and taxation 23:338–341
Subfossil 7:692
Subgame perfect equilibria 9:586, 12:103–104
Subgame perfectness 17:255
Subgenual anterior cingulate (ACC) 18:39–40
Subgoals, imitation by infants/young children on 11:622
Subgovernments 12:769–773
bounded rationality for 2:775
Subgroup differences 22:16–17
Subgroup identity 22:882
Subgroups 22:462, 22:882
Subimperialism in modern China 6:795
Subincision 24:455
Subintentional determination 15:311–312
Subitizing 17:51
Subjacency 16:4
Subject artifacts 8:571–575
Subject assessment, of context 4:787–788
Subject choice
coeducation schooling 3:928
single-sex schooling 3:928
Subject Condition (SC) 16:4
Subject experiencers (SE) 16:670–672
Subject formation
and modernity 15:690
statehood 23:348–349
Subject knowledge 24:56–57
Subject matter
delineation of 4:157
of phonetics 14:224
of single-case designs 21:984–985
Subject positions (discourses) 22:226
Subject Raising 16:2
Subject selection, experimental designs 8:537
Subject–action–object triplets 4:771
Subject–auxiliary inversion 16:2
Subjectification 7:257, 11:548
Subject-initiated surveillance 23:737
Subjective alienation 1:544
meaninglessness 1:544
mistrust *see* Mistrust
normlessness 1:544
powerlessness *see* Powerlessness
self-estrangement 1:544
social isolation 1:544
Subjective competence, and political efficacy 7:301
Subjective concepts and measures 19:776–777
Subjective enjoyment of life 10:522–523
Subjective expectations, data collection through questionnaires 19:776–780
Subjective expected utilities (SEU) 17:314
and decision making 5:917–919
Subjective expected utility theory 17:783
Subjective experience
flow 9:260–265
qualitative developmental psychology 19:668–669, 19:672–673
Subjective fault 5:221
Subjective feeling, emotion assessment through 7:519–521
Subjective happiness scale (SHS) 23:648
Subjective health, and morbidity compression 4:464, 4:466
Subjective identity 11:521–523
Subjective interpretations of probability 19:24–25
Subjective magnitude 21:604
Subjective measures of poverty 18:743
Subjective mental states 23:296
Subjective norm, health behavior 10:578, 10:580
Subjective opinion, within Bayesian framework, elicitation 7:382
Subjective phenomena, questionnaires 19:776–777
Subjective prior distribution, Bayesian inference 2:374
Subjective probabilities 5:882, 19:38–39
axiomatic theories 2:338–339, 2:341–342
definitions 19:38–39
Dutch Book theorem 19:39
for conditional probability 19:39
questionnaires 19:776–779
Subjective spirituality 22:164
Subjective task values 1:37
and academic achievement 1:22
development 1:22
Subjective theories, questionnaires estimation 19:774
Subjective time 24:393–394
Subjective time use 24:398
affect 24:398–401
happiness 24:401
overall life satisfaction 24:401
social well-being ratings in ATUS 2010 24:398–401
time pressure 24:401–404
outcomes from time inputs 24:404
QOL implications 24:403
questions 24:403–404
stress 24:403
time crunch scale 24:403
Subjective utility, axiomatic theories 2:338–339, 2:341–342
Subjective well-being (SWB) 2:487–488, 2:490, 22:150–151, 22:153

Subjective well-being (SWB) (*continued*)
benefits 23:649
physical health 23:649
social relationships 23:649
work 23:649
correlates of 23:650
and culture 23:643–647
current and future directions 23:651–652
definition 23:648–649
and employment 10:515–520
genetics, role of 23:649–650
and happiness 10:523–524
hedonic adaptation 23:650
historical background 23:648
increase in 23:650–651
interventions 23:651
gratitude 23:651
kindness 23:651
visualizing selves 23:651
measurement 23:648–649
and personality traits 17:922–923
psychology of 23:648–653
self-report measures of 23:648–649
and social capital 22:174, 22:178
Subjectivism 8:72
phenomenology 18:11–15
Subjectivism–objectivism 21:531
Subjectivities 2:305, 2:307, 10:692–693, 12:647, 17:285–287, 17:736, 19:326, 19:330–331, 19:333, 21:812
Bayes' theorem 2:380
and discourse studies 6:510
and gender 9:783–784, 9:792
global developments 10:151–153
and modernity 15:689
new genetics 9:1018–1019
and objectivity compared 17:88
in qualitative research 19:687
and risk 20:686–687, 20:690
sexual 19:756–757
social constructionism 22:225–226
and sociality 22:854, 22:856, 22:858–859
unemployment 24:735–736
and values 25:33, 25:35
Subject–object 18:784–785
Subject–object alignment 11:826
Subject–object relations, linguistic turn 14:199
Subject–object–verb (SOV) 1:955
Subjects 10:324, 24:460–462
sociology of 24:459
Subjects of significance 17:737
Subject-specific, infant and child survival 16:52, 16:54–55
Subject-specific knowledge 21:120
Subject–verb agreement 10:317–318, 10:322
Subject–verb–object (SVO) 4:770, 14:209–210
Subjunctive conditionals 5:87
Subliminal perception, and unconscious mental activity 24:723
Submarine cables 12:490–492
Submarine telegraphy 24:155
Submersion (bilingual education) 2:580, 2:582–584
Suboptimal conditions, judgments under 10:825
Suboptimal search behavior 5:909
Suboptimality 22:196
Subordinate attributions 23:704
Subordinate knowledge, resistance to disruption 21:574
Subordinate masculinities 14:681
Subordinate normative order 13:810
Subordinate precipitating factors 23:704
Subordinate system 11:827
Subordination 9:730
hegemony 10:748–749
Subpolitics 21:176–177
Subprime lending 3:580–581
defined 3:580
Subprime loans 3:581
Subprime mortgage crisis of 2008 3:32
Subproblems 19:49
Subretinal implants 16:716
Sub-Saharan Africa (SSA) 14:282, 16:267–268, 22:604
dowries 6:642–643
education in 7:198–209
fertility transition in 9:92–98
historiography and historical thought 11:89–92
Anglo-American debate 11:90
and anthropology 11:89–90
dependence 11:89–91
diaspora 11:90–91
indigenous peoples 11:89–91
nationalism 11:90
nations 11:89–90
oral history 11:90
oral traditions 11:89
origins 11:89
written history 11:89–91
Islam in 12:737–740
legal traditions of 23:654–659
African customary law 23:654–655
case of South Africa 23:658–659
civil law colonies 23:655–656
codification 23:655
common law colonies 23:656–657
conflict of laws 23:655–656
genesis of legal pluralism 23:655
hybrid colonies 23:657
immatriculation 23:656
legal transplants 23:655
postdemocratization era 23:657
primogeniture 23:657–659
repugnancy clause 23:656–658
structural adjustment in 23:598
Subset correspondence analysis 5:5
Subsets
applied demography 1:839
basic/academic demography 1:839
Subsidiaries, multinational corporations 3:46–47
Subsidiarity, budgeting 2:895
Subsidiarity principle 8:265, 19:583
Subsidiary theory 8:576
Subsidies
earning and wage 25:494–495, 25:494f
for nuclear and coal power 18:367–368
Subsistence
and anarchism 1:682
work anthropology 25:662–665
Subsistence cultures, caregivers in 18:223–224
Subsistence orientation, peasants 17:677
Substance Abuse and Crime Prevention Act (SACPA) 4:375
Substance Abuse and Mental Health Services Administration (SAMHSA) 6:680
Substance dependence, definition of 6:687
Substance dualism 18:107–108
Substance use disorders (SUD) 1:99–100, 17:592–594, 17:596
mental–physical comorbidity 4:390–391
Substance use/abuse 6:685–689, 20:271, 23:564
and adolescent health 1:121–125
of children, and divorce 14:519–520, 14:522–524
and panic disorder 17:476
and prospective memory 19:275
and sexual risk behaviors 21:799
Substance-based theories
of personal identity 17:742–743
immaterialist theories 17:742–743
materialist theories 17:743
Substance-related addictive disorder (SRD) 1:97–100
Substance-related disorders 25:143
VR applications in 25:143
Substances
and aggression 1:379
dualism of 15:545–546
Cartesianism, early criticisms and revisions of 15:546
proposal of Descartes 15:545–546
recent criticisms 15:546
Substances ingestion with psychoactive properties 21:868
Substantia nigra (SNC) 16:751
Substantialism, and field theory 9:140
Substantive autonomy 2:314–315
Substantive causes of women's inequality 9:730
Substantive due process 6:697
Substantive instruments 18:290–291
Substantive knowledge 22:728
Substantive rationality 2:444, 13:566
Substantive reason, and Enlightenment 7:666
Substantive variables 8:571
Substitution bias 19:738–739
Substitution effect 21:865–866
macroeconomic data 5:753–754, 5:756
in minimum wages 15:562
Substitution errors 18:49, 18:58
Substitution hypothesis 21:687
Substitution principle 14:294
Substitutions
for older workers 17:186
in sacrifice 20:860

Substratum, language contact 13:301–302
Subsumption architecture 2:310
  autonomous agents 2:310
Subsymbolic level 4:85, 4:87
Subsystems 22:745–751
Subsystems hypothesis 2:575
Subtle dehumanization, research on 6:34–35
Subtlety of stigma 23:465
Subtractive bilingualism 2:580–581
Subtractive feedback 25:158–159
Subtractive loop 25:158–159
Subtractive/substitutive acculturation 2:85
Subtyping model, for prejudice reduction 18:837
SUBUD 16:806
Suburbanism 23:660, 23:664–665
Suburbanization 10:48, 18:190, 21:917–919, 21:921, 24:860, 24:864, 24:866, 24:919, 24:921, 24:927
  causes 23:661–662
  consequences 23:665
  density 23:660–661, 23:665
  developers 23:660, 23:662–664
  exurbia 23:663
  home ownership 23:661–662, 23:664
  inner suburbs 23:661, 23:663
  metropolitan areas 23:663, 23:665
  outer suburbs 23:661, 23:663–665
  peri-urban 23:663–665
  postsocialist cities 3:608–610
  privacy 23:662, 23:664
  and spatial mismatch 23:157–158
  sprawl 23:660–661, 23:665
  suburbia 23:664
  transportation 23:660–662, 23:664–665
  urban fringe 23:660–661, 23:663, 23:665
Suburbia 23:664
Suburbio 23:660
Suburbs 24:935–936
  development of 23:662–663
  and gender 9:710
  and residential segregation 20:541
  types of 23:663–664
Sub-UTOS 8:619
Subversive sociology 4:137
Success
  and academic achievement 1:20
  definition 23:667
  and grit 10:397–398
  implementation 11:692–693
Successful aging 1:437–438, 6:378, 23:667–672
  adaptation 23:667–669, 23:671
  cognitive aging 6:380–381
  definition 23:667–668
  difference from normal and pathological aging 23:668–669
  differential aging as 6:379–380
  gain–loss dynamics 23:667, 23:669, 23:671
  health 23:667–671
  life span psychology 23:667, 23:669–671
  optimality 23:667–671
  plasticity 23:669–671
  proactivity 23:668–671
  processes 23:669–671
  SOC model 23:669–671, 23:670t
  well-being 23:669–671
Successful aging models 17:179–180
Successful intelligence, theory of 12:306
Successful person, self-efficacy as important element of 21:505–506
Successful tokens 15:243–244
Succession 8:773–775, 8:777, 22:265–266
  see also Wills and succession
Successive monogamy 14:612
Successive polygamy 14:612
Successivist approach 3:91
SUD *see* Substance use disorders (SUD)
Sudan, legal system in 15:365–366
Suddaby, Roy 17:394
Sudden death syndrome 5:573
Sudden infant death syndrome (SIDS) 12:17–19, 18:853, 18:857–858
Sudden stop (financial crisis) 9:184–185
Sudden unexpected death in epilepsy (SUDEP) 7:853
SUDEP *see* Sudden unexpected death in epilepsy (SUDEP)
Sudhoff, Karl 21:241–242
Suetonius 11:37–38
Suez Canal 14:531
Sufficiency, sustainable development 23:815
Sufficiency principle (heuristic-systematic processing model) 17:945
Sufficient summary plots 14:158–159
  estimation 14:159–160
Sufficientarianism, and intergenerational justice 12:401
Suffixes 15:821
Suffocation alarm system 17:469
Suffrage 17:515–518, 18:398–400, 23:365
  and feminist legal theory 8:918
Suffrage movement 22:443
  of women 9:737
Sufi orders 24:603
Sufi religious brotherhoods 24:602–603
Sufism 12:730, 12:733–734, 16:168–169, 20:328–329
Sugar Act 1:609
Sugar daddies 17:489
Sugarscape 1:368–370, 1:370f
Suggestion 11:458, 11:461–463
  false memories 8:710–713
  in hypnotizability scales 11:459
  for ideomotor movements 11:459
  posthypnotic 11:461
  role in hypnosis 11:465–471
Suggestion research 17:61
Sui generis 12:741, 12:794–795
  assumption 20:384
  assumption, religious belief 20:384
  Chinese policing 18:248
  Islamic fundamentalism 12:741
  Japanese cosmology 12:794–795
  law 11:776–777
Sui generis law 11:776–777
Sui generis@formation 18:248
Suicidal tendency, and sleep 22:61
Suicide 1:719, 1:851, 6:204, 6:702, 12:755, 20:269, 20:335–336, 23:563–564, 23:673–681
  and adolescent health 1:121–122, 1:124, 1:126
  anomic 1:714–716, 1:719
  attempts *see* Suicide attempts
  as basic human existential problem 23:676
  behavior, origins of 23:676–679
  biological aspects 23:678–679
  and bipolar disorder 2:684, 2:686–688
  birth cohorts 23:686
  definition 23:673, 23:682
  Durkheim on 23:683
    critiques and extensions 23:683–684
    empirical tests 23:684–686
  and early parental death 6:756
  and economy 23:685
  epidemiology 23:673–675
  etiology of 23:682–687
  and family structure 23:685
  and gender 23:686
  and hopelessness 11:199
  ideas 23:673–674
  importance of 23:673
  measurement 23:682
  perinatal 17:723
  prevention 23:679
  prevention of 23:687
  psychological and biological approaches 23:682–683
  psychological aspects 23:676–678
    aggression directed against ego 23:677
    appeal to human bonding 23:677
    classical and operant conditioning 23:677
    distorted cognitive schemes 23:677
    model learning 23:677–678
    narcissistic crisis 23:677
  and race 23:686
  rank order 23:674f
  rates 6:205t
  and religion 23:684–685
  religion and 20:264
  risk factors 23:675–676
    gender 23:675–676
  and social stratification 23:686
  sociological approaches 23:683–684
  sociological aspects 23:676
    Durkheim's theory 23:676
    general sociological trends 23:676
  sociology 22:996–998
  statistics 23:674f–675f
    WHO 23:673t
  terrorist tactic 24:238
  theory of contagion 23:686–687
  therapy 23:679–680
Suicide attempts 23:675
  risk factors 23:675–676
    gender 23:675–676
  and suicide, differing and common aspects between 23:675
Suicide problem 22:697
Suiming, Pan, on AIDS prevention 21:736
Suitable enemies 6:672

Suitable target 25:552–553
Sukyo Mahikari 16:806–807
Sulcus 3:333, 3:336
Sulfur 14:373
Sulfur oxides (SO_x) 7:773
Sullivan, Alice, on cultural capital 5:378–379, 22:333
Sullivan, Harry Stack 9:461, 19:505, 20:901, 21:129
  and childhood friendship 9:442–443
  children's friendships 9:448
  Sullivan-Piaget thesis 9:443
Sullivan, R.M. 16:470
Sullivan method 16:114
Sullivan-Piaget thesis 9:443
Sully, James, works, *Teacher's Handbook of Psychology* 7:591
Sulphonamides 15:869
Sumatraliths 23:65
Sumerian Code of UrNammu 2:672, 22:393
Summa potestas 1:7–8
Summarizing (communication skill) 4:289
Summary jury/judge trials 15:70
Summation theory 4:93
Summative classroom assessment 3:820–821
Summer bridge programs 13:718
Sumner, William Graham 8:137, 13:802, 23:689–695
  anti-imperialist to sociologist 23:692–693
  Baconianism 23:690–691, 23:693
  early life 23:689
  education 23:689–691
  and ethnocentricism 8:166–167
  in-group 23:689
  mores 23:689, 23:692–694
  neo-Darwinism 23:694
  politics 23:690–691
  positivism 23:690
  preacher to scholar in politics 23:690–691
  religious views 23:690
  social Darwinism 23:693–695
  social theorist 23:691–692
  sociological legacy 23:693–695
  works
    The Challenge of Facts 23:694
    Folkways 6:590, 23:692–693
    The Science of Society 23:693
Sumo 12:797
Sun, Meng, on Chinese cultural industries 5:597–598
Sun exposure 23:696–700
  appearance motivation 23:696–699
  health behavior 23:699
  predictors 23:697
Sun protection behavior
  predictors of 23:697–698
  self-efficacy 23:697–698
Sun protection self-efficacy beliefs 23:697–698
Sun Stone 11:55
Sun Tzu 24:239
Sun Yat-sen 3:515
Sunda shelf of Southeast Asia 2:230
Sundsvall, mortality decline in 15:871–872
Sung, Kim Il 2:74–75
Sunk costs 5:918
Sunnah (Islamic law) 8:878–879, 12:746–747, 12:752
Sunni Islam 21:791
Sunni legal schools 12:748
Sunnybrook Neglect Assessment Procedure 10:767
Sunspots, in expectations 8:513–514
Sunstein, Cass 4:705
Super villages, in China 2:68
Super-additive cover 4:873
Super-additivity 4:874
SuperAgers 6:380
Superalterns 14:15
Supercentenarians, oldest-old mortality 15:864–865
Superdiversity 16:32
Superego 15:894, 19:339, 19:400
Supergentrification 10:46–47
Superinclusive nonpositivism 13:479–480
Superior colliculus
  and multisensory integration 16:95–99, 16:96f
    essential circuit in 16:99–100, 16:99f–100f
    inverse effectiveness 16:99
    spatial 16:96–97
    temporal 16:97–99
Superior parietal lobule (SPL) 18:847
Superior parieto-occipital cortex (SPOC) 5:26–27
Superior temporal gyrus (STG) 13:370
Superior temporal sulcus (STS) 13:370, 15:582, 16:601, 18:848, 24:206, 8:664f, 16:601f
Superkinship trait 22:901
Supermarket revolution 1:460
Supermax 24:446
Supermax prisons 11:723
Supernatural 21:172–173
Supernatural entities 16:328–329
Superordinate category knowledge, resistance to disruption 21:574
Superorganic systems 23:275
  social selection in 23:275
Superorganism 5:677
Superpositional storage 5:580
Superpowers 4:766–767
  during Cold War 4:132
Superreferences 1:913
Superstitions 2:256
  in Japan 12:796
Superstructures 10:744, 22:844–845
Supervenience
  and mind–body dualism 15:547–548
  physicalism 18:106, 18:110
  social ontology 22:502, 22:504
Supervised learning 16:116, 23:413
  motor control 15:964
Supervision
  abusive 23:701–708
  financial 9:191, 9:193–195
  professional development 19:91–92
    higher education 19:92–94
    in organizational context 19:91–92
    research and scholarship in social work practice 19:93–94
  social work education 22:778–779
  of young researchers 15:235–238
Supervision competency 24:520–521
Supervisor bullying 23:704t
Supervisor incivility 23:704t
Supervisor Machiavellianism, abusive supervision 23:703
Supervisor social undermining 23:704t
Supervisor traits 23:703
Supervisor verbal aggression 23:704t
Supervisors checklist 15:236
Supervisors' experience, abusive supervision 23:703–704
Superwesterns 9:160
Supine position, infant motor development 15:973–974
Supplemental housing 10:449
Supplementary eye fields (SEF) 5:27, 18:847, 18:850
Supplementary motor area (SMA) 5:25, 6:725–726, 16:848, 21:421, 23:252
Supplementary motor complex (SMC) 5:25
Supplier-induced demand (SID) 10:615
Supply
  of artists, excess 2:54, 2:56
  and family planning 8:787–788
  housing 11:239–241
  job displacement insurance 12:826
  social change 18:422–423
  transportation 24:629
Supply and demand
  of education 7:264
  simultaneous equations 21:980
Supply base decisions, supply chain management 23:709
Supply chain
  and business platforms 3:38–40
  and commodity chains 4:243–245
Supply chain management 23:709–713
  complexities associated with 23:710–711
    competition 23:711
    information asymmetry 23:711
    lead time 23:711
    multiple agents 23:710–711
    uncertainty 23:711
  configuration-level issues 23:709–710
    information support decisions 23:710
    plant location decisions 23:710
    product portfolio decisions 23:710
    supply base decisions 23:709
  coordination-level issues 23:710
    capacity decisions 23:710
    cash flow decisions 23:710
    information flow decisions 23:710
    material flow decisions 23:710
  definition 23:709–710
  impact of technologies on 23:712
  inefficiencies 23:711
    customization challenges 23:711
    distortion of information 23:711
    poor utilization of inventory assets 23:711

stock-outs 23:711
logistics 23:709, 23:712–713
manufacturing 23:710–713
new developments 23:712
global supply chains 23:712
humanitarian supply chains 23:712
new technologies impacting supply chains and big data 23:712
sustainable supply chains 23:712
operations management 23:709, 23:711–713
outsourcing 23:709, 23:711
procurement 23:709
research 23:712–713
service 23:709–711, 23:713
sourcing 23:712–713
Supply of children, and fertility transition 9:94–97
Supply side, health insurance markets 10:641–642
Supply-and-demand model, microeconomics 15:332–334
Supply–demand diagrams 24:631–632
Support groups, online 10:686–688
Support system 18:888
Support vector machine (SVM) 8:661, 23:411
Support view, of counterfactuals 5:88–89
Supported familialism 12:404
Supportive behavior 15:412
Supportive Dimension 13:422
Supportive emotional colored atunements 12:350
Supportive environment, urban life 24:833–834
Supportive expressive therapy (SET) 19:345–346
Supportive psychotherapy (SP), BPD 2:759
Supportive writing context, creating 25:768–769
Support–learning–action sequence 19:373
Suppressants 21:610
Suppression
of emotions 7:453–455
statistical disclosure limitation 4:579–580
Suprachiasmatic nucleus (SCN) 11:476–477
Suprachoroidal implant 16:716
Suprachoroidal–transretinal stimulation 16:716
Supra-firm organizations 23:207, 23:211
Supramarginal gyrus (SMG) 17:506
Supranational courts 12:900
Supranational law 13:616
Supranational organizations, comparative databases 5:794–795
Supranational public administration 19:524
Supranational reference 10:866–867
Supranationalism 24:592–594
Suprasegmental units 15:827
Suprasegmentals 23:714–721
Suprastate law 13:616
Supremacism 14:841
Supreme Administrative Law Court of Prussia 13:429
Supreme Constitutional Court (SCC) 15:363
Supreme courts 23:722
German 3:669, 3:671–672
growth in policy intervention, and limits of power 23:725
institutional structure
agenda setting, workload, and internal structure 23:723–724
main types of 23:722–723
in political process
political influences on Supreme Courts and legitimacy 23:724–725
social scientific research on 23:725–726
Supreme Soviet's Conservation Commission 20:833
Sure thing principle, of violations 5:918
Suresha, R., works, *Bears on Bears* 13:891
Surething principle 20:927
Surface agraphia 6:735
Surface alexia 6:733
Surface commemoration 4:238–239
Surface dysgraphia 1:453, 1:455
Surface dyslexia 4:91, 6:730
Surface structure 9:873
Surface-level diversity 4:739, 6:596
Surface-syntactic dependency 6:185
Surface-syntactic relations 6:186–187
SURGE *see* Service Users Research Group England (SURGE)
Surgency/extraversion 17:877
Surgency/sociability 17:877
Surgent engagement 17:867
Suriname 1:638–639
Maroons 14:592–595
Surnames 16:184–186
Surplus
production 12:329
trade unions 24:498
Surplus product, and profits 3:91–92
Surprise, and flashbulb memory 9:256
Surreptitious sampling 9:959–960
Surrogacy 20:473
Surrogate endpoint research (mediating variable) 15:64
Surrogate endpoints 8:328
Surrogate motherhood 8:812
Sursurunga 17:44–45
Surveillance 11:552, 15:31, 17:341–345, 23:727–732, 25:122–123
agent-initiated 23:737
border 25:124
commodification of 23:738
communication and 23:738
defining 23:734–736
essays 23:733t
external constituency 23:737
external nonconstituency 23:737
internal constituency 23:737
on Internet 12:595–596
monetarization of personal data 23:738
myth of 23:738
nonorganizational 23:737
non-role relationship 23:737
nonstrategic and strategic 23:735
privacy and 23:736–737
reciprocal 23:737
role relationship 23:737
self-surveillance 23:737
slack 23:738
society 23:733–734
softening of 23:738
strips 23:738
structures and processes 23:737–738
studies 23:734
subject-initiated 23:737
and telephone 24:160
urban violence 24:955–957
use of term 23:733–734
value conflicts and 23:738–740
Surveillance constriction 23:738
Surveillance creep or gallop 23:738
Surveillance impact assessment 22:875–876
Surveillance societies 24:117
and punishment 19:656
Surveillance studies 12:595–596, 23:733–741
Survey analysis 7:558, 7:563
Survey data 5:737–739
big data 5:738–739
fitting on expectations 8:515
household 25:453
prospective 5:738
retrospective 5:737–738
Survey design, cognitive aspects 20:881–885
Survey instrument, extreme response style 23:758–762
Survey knowledge *see* Cognitive maps
Survey methods
settlement and landscape 21:656
urban geography 24:804
Survey nonresponse 14:354, 14:356
Survey of consumer attitudes (SCA) 5:744–745
Survey of Consumer Finances (SCF) 5:744, 11:733, 11:735, 16:92
Survey of Dictionary-based Isms (SDI) 22:164
Survey of Earned Doctorates (SED) 7:281
Survey of English Dialects (SED) 6:349
Survey of Family, Income, and Employment (SoFIE) 5:846
Survey of Health, Ageing and Retirement in Europe (SHARE) 5:793, 6:379, 12:405, 20:602, 22:645–646
Survey of Income and Program Participation (SIPP) 5:744–746, 10:303, 13:388
Survey of Labour and Income Dynamics (SLID) 5:845
Survey of Public Participation in the Arts (SPPA) 1:234, 5:447–449
Survey of Social Science: Psychology Series 7:591–592
Survey questions 23:765–766
Survey research 5:325–326, 23:747–751
academic research 23:748–749, 23:751
centers 23:748–749
GESIS: Leibniz-Institute for the Social Sciences, Mannheim 23:749

Survey research (*continued*)
Institute for Social Research, Ann Arbor 23:748–749
National Centre for Social Research, London 23:749
National Opinion Research Center, Chicago 23:748
companies 23:749–751
Gallup, Washington D.C. 23:749–750
Ipsos MORI, London 23:750
TNS Emnid, Bielefeld 23:750–751
Westat, Rockville 23:750
data archives 23:748–749
database 5:831–834
databases 23:750–751
definition 23:747
distinguishing survey centers from companies 23:747–748
on economic evaluations for elections 7:27
market research 23:747–751
methodology 23:748–751
political research 23:748–751
politics 18:459–464
public opinion polling 23:749–750
questionnaires 19:777
social research 23:747–751
superpopulation and modeling 20:877
Survey weights 13:386–387
Survey-based studies of everyday nationalism 16:311–312
Survey-guided development, organizational development 6:297–298
Surveys 12:51, 22:548
AAPOR 23:752–753
in archaeology 23:744–745
for census 3:296–300
Common Rule 23:754–755
criminal victimization 5:721–725
cultural participation surveys 5:732–733, 5:732t
data quality 23:756
efforts at persuasion 23:756
for embodiment 7:417–418
ethical aspects in 23:752–757
extreme response 23:758–763
history of 12:670–673
history of sample surveys 20:875–880
household expenditure surveys 5:733
incentives 23:756
informed consent in surveys and polls 23:755
on Internet 23:756
morbidity compression 4:464–466
obligations
to commercial or private sponsors 23:753
to the public 23:753–754
to respondents 23:754–756
to sponsors 23:753
to U.S. federal agency sponsors 23:753
political efficacy 7:301–302, 7:304–305
privacy/confidentiality in social research 18:943–944
protection of respondents' confidentiality 23:755–756
public opinion 23:752–754
and public opinion 19:563–567
quantification in 19:705
rational choice theory 19:904
representative method 20:876
representative survey, first tested in Norway 23:426
sampling techniques 23:426
sexuality
definition 15:315
measurement or response error in 15:315–316
for small areas 6:163, 6:165
special populations 12:673
standards for conduct of research 23:752–753
statistical 20:875–880
time-use expenditure surveys 5:733
units of observation 10:303
U.S. federal agency protections for human research participants 23:754–755
of values, cross-cultural 25:30–31
WAPOR 23:752–754
Surveys and survey series 1:156–157
adult mortality 1:156–157
Surveys of Inmates in State and Federal Correctional Facilities (SISCF) 10:457
Survival
evolutionary demography 8:413, 8:415, 8:417
and values 25:30
Survival analysis 8:295, 23:771–775, 23:783–793
Aalen additive hazards model 23:774
accelerated failure time model 23:774
additional topics 23:781
censored data 23:774
comparison with expected survival 23:773–774
competing semiparametric model 23:779
Cox model 23:774, 23:777–779
asymptotic distribution theory 23:779
development 23:777–778
special cases and generalizations 23:778–779
survivor function estimation 23:779
design decisions 23:783–784
estimated hazard probability 23:788–789
estimated survivor function over time 23:787
estimation and comparison of survival curves 23:777
event-history data in a person-period dataset 23:787–788
extending 23:792
failure-time data and representation 23:776–777
failure-time regression 23:777–779
history 23:771
incomplete observation 23:774
initial discrete-time hazard model 23:788
Kaplan–Meier estimator 23:771–772
log-rank test 23:772–773
multiple failure times and types 23:780–781
correlated failure times 23:780
frailty models 23:780–781
multiple failure types and competing risks 23:780
recurrent events 23:780
Nelson–Aalen estimator 23:772–774
nonparametric estimation and testing 23:771–772
parametric inference 23:773
parametric models 23:779–780
Poisson regression 23:774
regression models 23:774
time-to-event data 23:771
Survival data
event history analysis of 8:310–311, 8:310f–311f
two-state model for 8:312
Survival functions 14:91
random utility models 5:902–903
Survival of the fittest, use of term 1:86
Survival probability 23:785
Survival units 16:238–239, 25:406–407
Survival values, modernization 15:697–698
Survivalism 23:814
Survivor function estimation, survival analysis 23:779
Survivor function over time, estimated 23:787
Survivor movement 15:197
Survivorship, and population dynamics 18:564
Suspect softening or weakening 24:118, 24:122–123
Susser, I., on capitalism in North America 17:24
Sussex 12:256, 12:258–259
Süssmilch, Pastor 13:383–384
Sustainability 3:18, 6:28, 7:771, 7:774–776, 8:81, 15:638–639, 16:334, 16:337–339, 18:198–201, 18:203, 20:18, 20:805–806, 22:113–116, 22:168–169, 22:173, 23:794–807, 23:809–810, 24:616, 24:620–621, 24:867, 25:437–438
and alternative food movements (AFMs) 1:568, 1:570–571
city size 18:199–200
density and car dependence 18:199–200
concept 18:198
concept of 7:770–771
definition 7:771
development 23:801–802
and development 6:303–304
ecosystem 18:199
and energy 7:626, 7:630–631
environment 5:775, 5:777–779
environmental 23:802
environmental health and safety 7:745
environmental sociology 7:799, 7:802
global public policy 10:180–181
evidence 10:183
goals and targets 23:803–804
great transition of GSG 23:804
Millennial Declaration of the United Nations 23:803

sustainability transition of NAS-BSD 23:803–804
green innovations 18:200
holistic frameworks 18:198–199
horticulture 11:206
indicators and scenarios 23:802–803
and industrial societies 11:866–867
and intergenerational justice 12:396, 12:401
and land use indicators 13:238, 13:242–243
linkages between development and environmental sustainability 23:802
modernism 18:198
organizational environmental 19:83
papers published 23:801f
peak oil 15:555
planning processes 18:199
of Slow Food movement 22:95–96
'Small is Beautiful' approach 18:199
socioenvironment systems 23:805–806
and tourism 24:467–468
transportation 18:199
in United States 23:803t
urban form 18:199
urban sprawl 24:935–936
well-being 23:802–805
Sustainability competence 7:737
Sustainability lifestyles 23:828–833
behavior change 23:828–829
behavioral economics 23:831
citizens 23:828–829, 23:833
consumers 23:828–829
environmental citizenship 23:829–830
governing 23:828–830
libertarian paternalism 23:829, 23:831
nudge theory 23:829, 23:831
policy approaches for promoting 23:831
psychology of 23:830–831
social marketing 23:829, 23:831
social practices and 23:831–832
social practices approach 23:832
social transformation and 23:832
Sustainability science 23:804–806, 23:805f–806f
Sustainability studies, social ecology 22:267–268
Sustainability transition 22:254
Sustainability triple bottom line 6:830
Sustainable agriculture 1:486, 23:807–811
agribusiness 23:808
development 23:807, 23:809
ecological 23:809
economic 23:809–810
environmentalism 23:807, 23:809
food production 23:807–810
food security 23:808–809
industrial agriculture 23:809
in practice 23:810
organic agriculture 23:809
social 23:810
Sustainable business 10:392
Sustainable cities 18:203
see also Eco-cities
Sustainable consumerism 8:25–26
Sustainable consumption 4:760–761, 14:763, 23:834–836
Sustainable development (SD) 3:158–159, 5:258, 6:26, 6:481, 6:829, 7:105, 7:771, 7:778–781, 7:812, 7:815, 20:141, 23:795, 23:801–802, 23:812–816, 23:823–827, 24:792–793
counterdiscourses on 23:813–815
criticism of the conventional concepts 23:815
green Marxism/critical political ecology 23:815
as neocolonial strategy 23:815
sufficiency and degrowth discourse 23:815
defined 24:867
definition 23:823
as a dialogue of values 23:798–799
and disaster risk 6:481
ecological functioning, implications of 23:825
ecological modernization 23:813
ecosystem services 23:824
education for 7:733
EKC 23:825
implications of 23:825
endogenous growth, technological innovation, and resource dependency 23:825–826
implications of 23:826
environment–growth debate 23:825
and green social work 10:386–388
hegemonic sustainability discourse 23:812–813
incentives 23:823, 23:827
linkages with environmental sustainability 23:802
natural capital 23:823–827
nature-related counterdiscourses 23:813–814
limits to growth 23:814
radical ecology 23:813–814
perspectives of 23:815–816
policy issues and options 23:826–827
principle of 19:583
resilience 23:824
implications of 23:825
Rio + 10 Conference 23:797
Rio + 20 Conference 23:797
Rio Declaration on Environment and Development 23:796
Rio Earth Summit 23:796, 23:813
and social work 23:817–822
climate change 23:819–820
definition 23:817
environmental justice 23:820
Fourth World 23:820
inequality 23:820
Millennium Goals 23:818
performance 23:818
resilience 23:817, 23:819–820
scale 23:818–819
sustainable transitions 23:820–821
well-being 23:819
sustainable growth 23:812–813, 23:815
total capital stock 23:823–824
and urbanization 24:962–963
economic development 24:962–963
weak and strong sustainability 23:824
World Commission on Environment and Development (WCED) 23:812–813
Sustainable Development Goals (SDGs) 23:797, 24:751–752
Sustainable ecotourism 7:109
Sustainable energy 3:848
Sustainable growth 23:812–813, 23:815
Sustainable human development 6:310–311
Sustainable practices 23:834–838
behavioral change 23:834
competences 23:836–837
context 23:834–836
everyday life 23:834, 23:836
habits 23:835–836
leitmotifs 23:834, 23:836
lifestyles 23:834–836
methodology and methods 23:836–837
performance 23:834–836
practical turn and consequences 23:834–835
praxeological research 23:835–836
research methods 23:834, 23:837
routines 23:834, 23:836
rules 23:834, 23:836
sociomaterial context of 23:834–836
Sustainable production 14:763
Sustainable production and consumption (SPC) 23:839–844
ecological economics 23:842
ecological modernization 23:840–841
economics 23:840–842
policy 23:840–841
research 23:840–841
varieties 23:842–843
sociotechnical systems 23:839, 23:843
theoretical developments 23:842–843
transitions 23:839, 23:841, 23:843
Sustainable Sites and Responsible Land Development 7:774
Sustainable society 10:380
Sustainable supply chains 23:712
Sustainable tourism 7:105–106, 24:467–468
Sustainable transportation 23:845–849
alternative energy 23:847
anthropogenic activity 23:846
climate change 23:846
developed economies 23:846, 23:848
developing economies 23:846, 23:848
and economic development 23:846
environmental conflict 23:845
environmental sustainability 23:846–847
equity 23:848
global warming 23:846
globalization 23:846
greenhouse gases 23:846
land use 23:847–848
measuring sustainability 23:848
motorization 23:846–848

Sustainable transportation (*continued*)
natural resources 23:847
noise 23:847
pollution 23:846–847
recycling 23:845–847
safety 23:848
solid waste 23:847
sustainable communities 23:847–848
Sustainable urban development 6:833, 6:835
Sustainable urbanism 6:829–830, 6:832–836
Sustainable water management 25:437–442
Sustenance acquisition 11:286
Sustenance differentiation 11:285
Sustenance organization, human ecology 11:284–285
Susto ('Fright') 5:626
Sutherland, Edwin
on social class and crime 5:123
theory of differential association 22:409
Sutta Pitaka 2:888
SUTVA *see* Stable unit treatment value assumption (SUTVA)
Suzerain des suzerains 1:6–7
SV *see* Stochastic volatility (SV) models
SVCs *see* Serial verb constructions (SVCs)
SVD *see* Singular value decomposition (SVD)
Sveningsson, M.
model of Internet privacy 12:609, 12:609f
SVM *see* Support vector machine (SVM)
SVO *see* Subject–verb–object (SVO)
SVP *see* Schweizerische Volkspartei (SVP)
svPPA *see* Semantic variant of primary progressive aphasia (svPPA)
SVTC *see* Silicon Valley Toxics Coalition (SVTC)
SW *see* Social withdrawal (SW)
SWA *see* Slow wave activity (SWA)
SWAA *see* Social Welfare Action Alliance (SWAA)
Swadesh model 11:24–25
Swadesh-style lexicostatistics 11:25
Swahili 6:773–774, 6:776, 9:786
Swamping effect, in outliers 23:389
SWAN *see* Social Worker Action Network (SWAN)
Swanton, John 10:139
Swarms 2:310
SWAT *see* Special Weapons and Tactics (SWAT)
Swaziland, sociocultural aspects 23:75, 23:77, 23:79
SWB *see* Studying subjective well-being (SWB); Subjective well-being (SWB)
Sweat lodge, and indigenous social work 11:807–809
Sweating sickness (sudor anglicus) 18:174
Swedberg, R., works, *The Handbook of Economic Sociology* 7:588
Sweden
adoption in 1:146
anthropology 1:754
literacy in 11:591
mortality rate decline 15:869
Swedish Center Party 10:383
Swedish Left Party 10:383
Swedish model
of literacy 14:238
of welfare state 21:178, 25:530
Swedish Psychological Association, establishment 3:874
Sweet, Henry 18:48
Swell guys 25:814
SWF *see* Social welfare function (SWF)
Swidler, Ann 5:501
on culture 5:557
SWIFT *see* Society for Worldwide Interbank Financial Telecommunications (SWIFT)
SWIID *see* Standardized World Income Inequality Database (SWIID)
Swinging (sex) 17:490–491
Swiss Civil Code 3:671
Swiss Green Liberals 10:383–384
Swiss Household Panel (SHP) 5:846
Swiss People's Party 18:610–611
Switch design 8:653
Switch-face 8:653
Switching, language 13:253
Switching of religions 20:363–364, 20:408
Switzerland
clinical psychology 3:873
confederation 4:573
diglossia in 6:422
morphology 6:422
phonetics/phonology 6:422
vocabulary 6:422
euthanasia in 8:281, 8:283–284
tax system 23:341
Vocational Education and Training (VET) system 1:872
SWLS *see* Satisfaction with life scale (SWLS)
SWS *see* Slow-wave sleep (SWS)
Sydenham, Thomas 11:506
Sydenham's chorea 17:115
Sydenstricker, Edgar 17:39
Sydney Gay and Lesbian Mardi Gras (SGLMG) 9:640
Sykes, G., works, *The Society of Captives* 18:920
Syllabaries 25:780
Syllabic constraints 18:57
Syllable 18:65–66, 18:68–69, 18:71–73
Syllabus of Errors 11:93
Syllogisms 19:63
legal reasoning 13:780–781
Symbiosis 17:621
Symbiotic relationship 15:21
Symbol grounding 17:52–53
Symbol systems 7:247
in religion 20:387
Symbolic action 9:662–663
Symbolic algorithms
for constraint-satisfaction problem 4:716–719
backtracking search 4:718
flexible CSP variants 4:719
local consistency checking 4:716–718, 4:716f
local search methods for classical CSPs 4:718–719
Symbolic analysts 12:90
Symbolic anthropology 6:659–660, 23:860, 23:862–864, 24:293
Symbolic borders 2:762–763
Symbolic boundaries 5:451, 23:850–855
art 23:852
challenges and future directions 23:854
class 23:851–853
cognition 23:850
community 23:850–851, 23:853
current theory and research 23:851–854
culture and inequality 23:851–852
identity and boundary work 23:852–853
moral order 23:853–854
religion 23:854
symbolic politics 23:853–854
definition 23:850
ethnicity 23:853–854
history 23:850–851
inequality 23:850–851
intellectual context 23:850
morality 23:852–854
nationalism 23:853
popular music categories 3:800–801
race 23:850–851, 23:853
status 23:850–854
Symbolic capital 3:93, 18:783
Symbolic cognition, of humans 11:255–256
Symbolic communication 18:40
Symbolic domination 6:505, 9:785
Symbolic economy 13:109
Symbolic entities
celebrities as 3:275–276
as democratic aristocracy 3:275
search for 'Real' self 3:275–276
Symbolic fields 16:180
Symbolic function, of constitutions 4:713
Symbolic individuation 16:505
Symbolic interaction 22:538–539
dramaturgy 6:647–648
fieldwork 9:151
Symbolic interactionism (SI) 1:76–77, 6:21, 9:539–540, 10:556, 10:786, 12:352–356, 14:831, 14:835, 19:491, 19:677, 21:851, 22:222, 22:267, 22:529, 22:993, 23:431–433, 23:527
and actions 12:352–354
Becker on 12:352–354
Collins on 4:213
criticisms 12:354
developments 12:354–356
Goffman on 12:355
and interviewing 5:855
and meaning 16:561
methodology 23:856–859
analytic induction 23:857
Blumer's 23:856–857
Chicago School 23:856–857
current trends 23:858
dramaturgy 23:857–858

existentialism 23:857–858
grounded theory 23:857–858
labeling 23:857
Mead's 12:352, 12:354–355, 23:857
naturalistic research 23:856–857
phenomenology 23:857–858
postmodernism 23:857–858
Twenty Statements Test 23:857
and organizations 17:391
origins and development 12:352
phenomenology in sociology 12:352–353, 12:355
and pragmatism 18:809–810
reflexivity 12:353–354
relationship to ethnomethodology and mainstream sociology 12:353–354
self 12:352–353, 12:355
and social identity 22:313–314
social psychology 12:352
of socialization and education 22:863–864
theoretical and methodological implications 12:353
traditions 12:354–356
Symbolic interactionist theories 22:958
Symbolic laws 15:591
Symbolic market, gift as 10:140–141
Symbolic meaning, and color classification 4:229
Symbolic numbers 17:51–53
Symbolic order, and ethnic conflict 8:93
Symbolic phenomenon, magic as 14:436–438
Symbolic play, during toddlerhood 18:221
Symbolic politics 23:853–854
Symbolic production, and Internet cultural products 12:589–590
Symbolic racism 19:859
Symbolic realm 13:212–215
and ultramodern shadow 5:247
Symbolic revolution, and field theory 9:142–143, 9:145–146
Symbolic signs 25:171
Symbolic system 2:882
Symbolic thought 16:514–515
Symbolic threat hypothesis, race–crime relationship 19:842
Symbolic violence 25:117
Bourdieu on 2:781
Symbolically generalized media of communication 14:387
Symbolism 4:228–230, 10:334
in anthropology 23:860–865
alternatives 23:864
belief 23:861–864
core symbols and domain of kinship 23:861–862
definition 23:860–861
disciplinary practices 23:864
Geertz *versus* Asad and Bourdieu on Public Symbols and Private Discipline 23:863–864
genealogical critique of religion 23:862–863
icons 23:860
multivocality 23:860–861
pollution 23:860, 23:862
religious symbols and cultural systems 23:861
ritual 23:860–864
signs 23:860–861, 23:864
symbolic anthropology 23:860, 23:862–864
symbols, classification, and taboo 23:862
archaeology of 20:721–722
issues 20:725–726
material basis for 20:724–725
artificial intelligence 2:35–37
criticisms 2:37
learning 2:37
representation and search 2:35–36
and authority 2:269–270, 2:272–273
commemorative 4:235
and cultural sociology 5:498–502
and ethnic conflicts 8:92–93
generational 9:867–868
Lévi-Strauss on 13:905–907
and magic 14:436
and religion 20:309
spirit possession 23:283
tattoos 24:33–35
transition of magic to 14:436
Symbolization 24:263–266
in cartography 3:170
Symbols 20:666, 21:592–593
borders as 2:762–763
cognitive meaning of 5:579–580
culture as 13:726–729
organizational 16:179–182
Symbols of mind 5:576, 5:579
SYMLOG *see* Systematic Multiple Level Observation of Groups (SYMLOG)
Symmetalism 14:624
Symmetric distribution, location of 16:888
Symmetric learning, promotions 19:138–139
Symmetric map 5:2–3
Symmetrical affirmative action 1:252
Symmetrical analysis 21:309
Symmetries 16:413, 23:592
cooperatibe game theory 4:873, 4:875
and factions 8:685
principle of 4:163, 4:167
social representations 20:445
Strong Program 23:593
Symmetry of information 1:535
Symmetry postulate, ethnomethodology 21:192–194
Sympathetic adrenomedullary (SAM) system 23:577
Sympathetic engagement, in neonates 16:481, 16:486–487
Sympathetic magical law of contagion 6:547
Sympathetic mediation, cardiac defense 23:542
Sympathetic-adrenal-medullary (SAM) system 3:552, 19:445–446, 23:565
Sympathetic-efferent somatic-afferent coupling mechanism, of phantom limb sensations 18:5
Sympathizers, collective political action 4:149
Sympathy 7:549–550, 7:552, 8:74
cognitive-anatomical basis of 16:612
versus empathy 15:790
microeconomics 15:332
neonatal imitation 16:487
and prosocial behavior 19:229
and prosocial responding 7:317–319
Symptom Checklist-90R (SCL-90R) 4:464
Symptom perception 23:866–872
definition of 23:866–867
interoception 12:618–619, 23:866
medically unexplained symptoms 23:870
psychological modulation of 23:867–870
anticipation, symptom schemata and illness beliefs 23:868–869
emotion and personality 23:869
memory 23:869–870
perception and attention 23:867–868
Symptom perception training
for asthma 20:579
for chronic obstructive pulmonary disease 20:579
Symptom provocation 11:605–607, 11:609
and chemosensory stimuli 11:607
Symptom schemata 23:868–869
Symptomatic estimation techniques 6:164
Symptomatology, transference 24:545
Symptoms, of infectious diseases
psychological influences 12:35
reporting 12:33
standardized measures 12:34
Synanon communities 24:276
Synapses 2:29, 16:737
in brain development 2:823–824
cognitive neuroscience 4:95–96, 4:100–101
delivery, AMPA receptors
glucocorticoids 16:490–491
whisker experience-dependent 16:490
neurons
inputs 16:705
integration 16:705–706
plasticity 16:706
Synapses en passant 16:752
Synapsin I 12:713
Synaptic downscaling, and memory 22:64
Synaptic function, Ca^{2+} neuronal activity and 12:712–713
Synaptic homeostasis theory 22:64–65, 22:65f
Synaptic learning rule 23:875, 23:875f
Synaptic mechanisms, in learning, memory, and forgetting 23:873–880
Synaptic monoamine transporters 1:779
Synaptic plasticity
and cerebral cortex 3:338
computational neuroscience 4:482–483
and memory 9:906–907
Synaptic pruning 2:824, 20:57–58
in brain development 2:821, 2:824
Synaptic reversal potential 16:705
Synaptic strength, and memory/learning 23:873–878

Synaptic transmission, and memory/learning 23:873–875
Synaptogenesis 2:837, 4:59, 20:57
in brain development 2:823–824
Synaptosomal-associated protein 25 (SNAP-25) 22:495
Synaptotagmin 12:712–713
Synchronic corpus 4:950
Synchronic linguistics 20:923–924
Synchronic unity 11:127
Synchronous 6:567
Synchrony/synchronization
attention-dependent 16:637
in psychiatric diseases 16:637
rapid 16:636
relation to perception 16:636–637
of responses to sensory stimuli 16:636
as a signature of relatedness 16:635–636
Syncretism 2:890, 3:184, 3:535, 23:881–884
acculturation 23:882
creolization 23:883–884
and creolization 5:118–119
current issues 23:882–884
from definition to process 23:883–884
fundamentalism 23:883
power 23:882–883
and diffusion 6:404
globalization 23:883
hybridity 23:883–884
and hybridity 11:449–450
inculturation 23:881
new religions 23:882
use of term 23:881–882
earliest uses 23:881
views from cultural anthropology 23:882
views from religious studies and missionary practice 23:881–882
Syndicalism 22:834–835, 22:839
Syndromic surveillance, medical geography 15:88
Synecdoche 15:290
Synergy, in groups 10:410
Synergy formation, motor control 3:778–779
Synesthesia 14:193
brain scan research 14:197
causes 14:196–197
color to grapheme 14:196
colored grapheme 14:196
colored number 14:196
diagnostic criteria for 14:193
grapheme to color 14:194–195, 14:197
grapheme to touch 14:196
incidence rate 14:193
involving speech sounds or whole words 14:195–196
language-induced 14:196
lexeme to flavor 14:195–196
lexeme to odor 14:196
lexeme to touch 14:196
lexeme to visual 14:196
phoneme to touch 14:196
phoneme to visual 14:196
'projector' and 'associator' synesthetes 14:197
ticker tape 14:196
Synsets 21:581
Syntactic analyses, of serial verb constructions 21:646
Syntactic bootstrapping, and semantics 25:228–229
Syntactic complexity 21:914
Syntactic constituency 6:190–193
Syntactic construction 13:269
Syntactic control 4:828–833
arbitrary control 4:828
complement control 4:832–833
control shift 4:830
exhaustive control 4:828
explicit control 4:832–833
implicit control 4:832–833
language-specific differences 4:831, 4:833
causative *versus* con-causative coding 4:832
implicit control *versus* explicit control 4:832–833
modal *versus* nonmodal coding 4:831
minimal distance principle 4:829
obligatory control 4:828
partial control 4:828–829
pragmatic approach to control 4:831
raising and control 4:830
semantic approach to control 4:830
split control 4:829
Syntactic coverage, grammars 19:9
Syntactic dependency 6:183–194
criteria for Synt-d 6:183–185
deep-syntactic relations 6:186
dependency tree as formalism for representing syntactic structures 6:185–186
English surface-syntactic relations 6:187–189
in first half of the twenty-first century 6:193–194
labeled Synt-dependencies 6:183
linguistic advantages of 6:191–192
linguistic insufficiency of 6:192–193
morphological dependency 6:189
notion of Synt-d 6:185–186
phrases and phrase-like units 6:193
properties of 6:185–186
surface-syntactic dependency 6:185
surface-syntactic relations 6:186–187
and syntactic constituency 6:190–193
Synt-connectedness 6:183
Synt-dominance 6:183
Syntactic embedding 14:242
Syntactic indeterminacy 25:171
Syntactic inference, reasoning 5:934–935
Syntactic parallelism 13:272, 13:277–278
in ancient Hebrew poetry 13:277
Syntactic positive shift (SPS) 9:483
Syntactic processing 16:691–692
neurological correlates of 16:692–694
Syntactic rules 14:841, 23:901
Syntactic structures 9:873
dependency tree as formalism for representing 6:185–186
Syntagm (semiotics) 21:593–594
Syntagmatic processes 13:942
Syntax 2:882, 11:820–821, 11:824, 11:826, 13:100, 13:942, 23:231–232, 23:885–894, 23:901
as algebra 15:751
and bilingual aphasia 2:574–575, 2:577–578
of change 22:149
of a Formal Language 15:751
movement theory and constraints in 16:1–7
neural network models 13:249
prototype theory 19:286–289
valency and argument structure in 25:1–6
of verse 13:267
Syntax–prosodic structure interface 23:895–899
prosodic structure hypothesis 23:896–899
domain convergence 23:896
domain layering 23:896–897
interaction of prosodic markedness constraints and interface constraints 23:897–898
nonsyntactic determinants of phonological and phonetic domain structure 23:897–898
speech rate 23:898–899
serialist or parallelist 23:899
Syntax–prosodic structure interface constraints 23:895–896
Syntax–semantics interface 23:900–905
constituency 23:903–904
grammatical relations 23:904
partitioning 23:900–901
preliminaries 23:902–903
validity 23:903
word orders 23:904
Syntegration 5:647
Synthesis 12:791, 17:621
psychiatric social work 19:305
same-sex youth 9:654
Synthetic Aperture Radar (SAR) sensor 10:66
Synthetic biology 2:678, 3:374, 9:979, 21:183
ethics 2:617–618
Synthetic cohort 14:91
demography 6:113–114
Synthetic cohort measures, marriage/divorce 6:118–119
Synthetic data, statistical disclosure limitation 4:580
Synthetic extinct generation method (SEG) 1:154, 6:146–147, 6:147f
Synthetic geometry 25:197–198
Synthetic ignorance 11:579
Synthetic informant model (SIM) 25:753–754
Synthetic philosophy 23:275
Synthetic theory, Darwinism 5:682
Syphilis studies, in Tuskegee and Guatemala 13:561–562
Syria, legal system in 15:365

Syrian Orthodox Church 17:437
of South India 17:438
Syriza 10:383
Syrjäytyneiden akanvirta 10:50–51
SYSTAT (statistical analysis package) 23:378
System conformity dynamics 12:171
System dynamics
LTG study 14:138–141
peak oil 15:556–557
System identification 24:345
System justification theory 12:412, 15:240
for social dominance orientation 22:251
System maintenance/transformation 18:206
System of national accounts (SNA) 5:707, 5:751, 16:227
accounting structure 16:227
conventional changes in 16:230
early efforts to compile 16:227–228
in France 16:229
general revisions 16:229
in Germany 16:229
global use 16:227
measure of nation's total economy 16:227
milestones in development 16:228t
(inter-)national differences 16:229
OEEC's Standardised System 16:229
political nature of 16:230–231
process of harmonization 16:229–230
sequence of account in 5:753
in twentieth Century 16:228–230
in Western Europe 16:229
System of sensorimotor systems 16:506–507
System of systems 16:506–507
System of the Methods of Emotional-Imaginative Teaching (EIT-system) 4:668, 4:671–672
applicability of 4:671–672
System of total exchange 11:156
System overload 24:123
System properties 7:433–436
System responsiveness
and political efficacy 7:301–303
System–subsystem
human–environment relationships 11:406–407
System transition 7:633–634, 7:633t
Systematic assessment for treatment emergent events 19:463
Systematic cognitive biases 4:895
Systematic desensitization 2:408, 2:426–427, 3:767–768, 18:41–42
historical background 3:880–881
Systematic errors 8:571
Systematic geography 10:84
Systematic histories 11:108
Systematic integration 19:509–510
Systematic job evaluation 15:597
Systematic knowledge 21:241
Systematic literature reviews 15:267–268
Systematic Multiple Level Observation of Groups (SYMLOG) 22:102–103
Systematic narrative review, of SES and social-emotional development of children 22:942–948
Systematic observation 22:545
Systematic processing 15:32
Systematic reviews 23:906–913
evidence-based medicine 8:330
in evidence-based practice 8:333–337
formulating good research question 23:906–907
logic models 23:907
setting eligibility criteria, PICOS 23:907
screening and coding studies 23:908
judging study quality 23:908
searching for relevant literature 23:907–908
Systematic sampling 5:321
The Systematic Study of Religion 2:507
Systematic theory of reasoning 16:928–929
Systematicity 11:822
Systematization 11:512–513
Système de dépopulation 10:6
Systemic abuse 7:330
Systemic barriers, for immigrants 11:642, 11:644
Systemic couple therapy 8:748
Systemic family practice 8:745
Systemic family therapy 8:745–746
Systemic interventions, organizational development 6:298
Systemic nature of innovation 12:147
Systemic oppression, and racialized groups 22:824
Systemic prerequisites 14:242
Systemic racism 19:804, 19:847–850
Systemic reductionism 5:233
Systemic research
in education 4:434–436
and modeling 4:432
Systemic self-reflection, history of education in connection with 7:174–175
Systemic transactional model (STM) 11:594
Systemism 22:985
System-level information 5:772
System/metasystem model 13:900–902
assessment 13:902
Systems 8:745–747
sociocybernitics 22:912–913
and functions 9:519–520
hierarchy 22:913
holon property 22:913
land tenure 13:231–232
media as 22:583–585
negative feedback 22:913
positive feedback 22:913
reasoning 5:937–938
Systems approaches
international social problems 12:543
perceptual learning 10:127
Systems fit, in structural contingency theory 23:613
Systems integration, and social integration 12:245–246
Systems level theory 5:191–192
Systems of signs 13:726–727
Systems or control theory 15:946–947
Systems perspective 24:617
for industrial ecology 11:844–845
Systems science 22:912
System/360s system software 15:719
Systems theories 2:137, 7:434–435, 11:351, 13:554–555, 14:415
and ecological paradigm 6:879–881
fashion 8:837
and functionalism 9:541–542
Luhmann's 14:383–384
of motivational development 15:946–950
organizational design and form 6:235
of personality development *see* Personality
and psychotherapy 19:505
social differentiation 6:391–394
social structure 22:715–716
Systems thinking 17:315, 20:561, 20:566
Systems Training for Emotional Predictability and Problem Solving (STEPPS)
for borderline personaliy disorder 2:759
for personality disorders 17:896
System-wide perspectives, open access 12:866–867
Szasz, Thomas, on diseases 17:2–4
Szlachta 1:965, 3:96
Szpilrajn's theorem 17:527–528
Szreter, Simon 17:41
Sztompka, Piotr 15:709

T

t distribution, in hypothesis testing 11:495–496
TA *see* Technology assessment (TA)
TAARs *see* Trace amine-associated receptors (TAARs)
Tabardillo 18:176
al-Tabari, Abu Ja'far, works, *Chronicle of Prophets and Kings (Ta'rikh al-rusul wa'l-muluk)* of al-Tabari (838–923) 11:62, 11:64
Table of authorities 5:803
Tableaux, reasoning 5:935–936
Tablighi Jama'at (religious movement) 16:807, 20:375, 24:605
Taboo 24:1–3
customs 24:1
and ignorance 11:579
Lévi-Strauss on 13:905–906
primitive thought 24:1
religion 24:1
social order and moral persons 24:1–2
and symbolism 23:862
transgression 24:2–3
Tabula rasa 7:147, 13:701, 18:32–33
Tabulae 19:179–180
Tabulating Machine Company 23:376–377
Tabulating machines 23:377
Tacit assumptions, and corporate culture 4:923–925
Tacit coordination 22:246–247
using cultural resources 18:422
Tacit holism 16:104
Tacit knowledge 6:491, 13:76–78, 13:93, 17:366, 24:4–8, 24:57–58

Tacit knowledge (*continued*)
beliefs, and nonconceptual content 13:77–78
collective memory 4:200–201
communities of practice 24:5–6
critiques 13:78
and ignorance 11:580
knowledge management 24:4–6
learning 24:4–7
and management research 24:4–5
and mental imagery 15:220–221
next iteration 24:6
organizational knowledge 24:4–6
organizational learning 24:7
and personality psychology 17:910
in philosophy 13:80–83
AGL task 13:83–84
compositional semantic theories 13:80–81
concept possession 13:81
finite state grammars 13:83–84
implicit rules in connectionist networks 13:81–82
learning and classification 13:84
theory and simulation in everyday psychological understanding 13:82–83
Polanyi on 24:4–7
of rules, Chomsky on 13:74–76
faculty of language 13:75
Quine's challenge 13:75–76
schools of thought 24:5–6
community perspective 24:5
knowledge-in-practice 24:5–6
structure required for 13:76–77
Tacitean radicalism 3:772
Tacitus 11:37
classical legacies 3:772–775
Roman imperial historiography 11:37–38
works
Annals 11:37
Histories 11:37
sine ira et studio 11:37
TACOSODE *see* Tanzanian Council for Social Development (TACOSODE)
tACS *see* Transcranial alternating current (tACS)
Tactics 15:502
to improve creativity training 5:114t
in negotiation 16:425
of social movements 22:439–440
terrorist 24:238–239
Tactile hallucinations 10:500
Tactile information, higher order processing of 24:15–16
Tactile maps 24:9–13
audio–visual tactile maps 24:11–12
concept, types, and relevance in geography 24:9–10
creation and use 24:10–11
need for education and training 24:11
design 24:10
geographical maps 24:10
innovative methods and products in 24:12
orientation and mobility maps 24:9–10
production methods and materials 24:10–11
Tactile Pattern Reasoning scale 6:329
Tactile processing 24:14–19
consciousness 24:15–16
contribution to representation of body 24:16–17
haptics 24:14
higher order processing of tactile information and tactile awareness 24:15–16
hormones 24:16
multisensory aspects of 24:17
neural mechanisms of 24:14–15
physiology 24:16
social aspects of 24:16
Tactile stimulation 20:59
Tadashi, Sekino 2:71–72
TADMUS program 11:300–301
Taejong sillok 11:49
TAG *see* Tree-Adjoining Grammar (TAG)
Tagalog 2:705
and parts-of-speech hierarchy 17:559
lexical paradigms 17:560
noun/verb distinction 17:560
Tagging 4:951
welfare reforms 25:518–519
Tagil, Sven 16:286
Tahir, Kemal 20:870
Tahiti 17:158–159, 17:161
Tahrir Square 25:844
TAI *see* Team Assisted Individualization (TAI)
Taïa, Abdellah, works, *L'armée du salut* 21:753
Taiheiki 11:51
Taijin Kyofusho ('Interpersonal Fear Disorder') 5:626
Tail wags dog problem 4:498
Tail-area significance tests 20:927
Tailoring 17:929
Taine, Hippolyte 4:763
works, *Origines de la France Contemporaine* 4:763
Taipei City Sustainable Development Council 24:895
Taiping Rebellion 12:525, 15:531
Taiwan 5:454
and China 2:65–66, 2:69
cultural policies in 5:454–455
creative industries 5:456–457
fertility transition in 9:55–59
land redistribution 13:233
Taiwan International Graduate Program (TIGP) 22:639–640
Taiwan politics 6:800–801, 6:803
Tajfel, Henri
works
The Context of Social Psychology. A Critical Assessment 22:534
The Social Dimension 22:534
Tajikistan 3:316–317, 10:189
Takagi, Paul 22:386–387
Takashi, Akiba 2:71–72
Takelma 20:899–900
Taken for granted, politeness as 18:328
Take-some dilemma 22:245–246
Take-the-best heuristic, of decision making 5:920
Take-the-first heuristic, of decision making 5:920
Takfir 12:743
Talbot, William Henry Fox 18:81
Talent development 7:274–277
see also Gifted education
Talent management 21:103
Talent of artists 2:53–54
Talent search model 7:160–161
gifted education 7:161
Talented 7:158
education for 7:158–164
Talents, and intelligence 12:309
Talib, Abu 12:731
TALIS *see* Teaching and Learning International Survey (TALIS)
Talk, ethnomethodology 8:206
Talk and interaction, narrative 16:204, 16:206, 16:209
Talk shows, in radio 19:881
Talk-alouds 13:715
Talking about traumatic events 7:500
Talking treatments 15:194
cognitive behavioral therapy 15:194
Talk-in-interaction 15:877
Tallying 5:915
in decision making 5:920
Talmud 11:67–68, 12:870
Talmudic Judaism 12:870
Talmy, Leonard 10:110
TAM *see* Technology Adoption Model (TAM)
Tam, Henry 4:313, 4:316
TAME *see* Tense, aspect, mood, and evidentiality (TAME)
Tame death 7:595
Tamil language 7:197
Tampa 2:242
Tampubolon, Gindo, on cultural participation 16:556
Tamunhwa 2:74
Tana, Li, works, *Nguyen Cochinchina: Southern Vietnam in the Seventeenth and Eighteenth Centuries* 11:85–86
Tandem mass spectrometry (MS/MS) 9:997
TANF *see* Temporary Assistance for Needy Families (TANF)
Tang Code 13:800
Tang Si-Tong 24:585
Tangible user interfaces (TUI) 24:20–22
Ambient Wood project 24:22
characteristics of 24:20–21
classification of 24:21–22
constructive assembly 24:21
interactive surface 24:21
token+constraint system 24:21
cybergogy 24:23, 24:23f
in education
application cases of 24:22–23
augmented reality (AR) 24:22
challenge of 24:23–24
constructive assembly 24:22

design of 24:23
embedded phenomenon 24:22–23
immersive environments 24:22
instructional context 24:24
integration with curriculum 24:24
interactive surface 24:22
malleability 24:24
participation 24:24
scalability 24:24
student adjustment 24:24
token + constraint system 24:22
embedded phenomenon 24:21–22
embodied phenomenon 24:20–21
EvoRoom project 24:22
graspable user interface (GUI) 24:20
immersive virtual 24:21
immersive virtual environments 24:21
in learning and education 24:20–25
representation 24:20–21
RoomQuake project 24:22–23
TUI interaction model 24:20
ubiquitous computing 24:24
Tangshan Caofeidian International Eco-City 6:836
Tanner, W.P., Jr. 21:952–954
Tannery, Paul 21:241–242
Tanning, and skin cancer 23:696–699
Tansley, Arthur 7:792
Tantric Buddhism 2:888–889
Tantric ritual 2:889
Tantrism 7:374
Tanzania 6:773–776
Tanzanian Council for Social Development (TACOSODE) 12:500
Tanzanian Muslims 12:739
Tanzímát 15:360–361
T'ao Yüan Ming 14:253
Taoism 16:168–169, 16:806–807, 20:367, 20:369, 20:408
East Asian studies 6:786–787
and harmony 10:529, 10:532
Taoists 4:312
Taparelli, Luigi 2:670–671, 22:392
Tape-recording 6:350
Taphonomy 25:852
TAPS *see* Team for the Assessment of Psychiatric Services (TAPS)
TAQ *see* Test Anxiety Questionnaire (TAQ)
Taqlīd 12:746t
Tarascan Empire 15:263–264
Tarde, Gabriel de 5:887, 13:802, 15:304
innovation theory of 12:169–170
works
Les Transformations du Pouvoir 2:269–270
Laws of Imitation 4:164–165, 23:686–687
Tardive dyskinesia 19:466
Target costing 14:468–469
cost tables 14:468
functional analysis 14:468
tools 14:468
value engineering collections 14:468–469
Target devaluation 24:119
Target group, influence of stereotypes on 23:450–451
Target hardening 24:118
Target identification 24:120
Target insulation 24:119
Target language (TL) 21:361
properties, impact of 9:203–204
Target marketing 3:663
Target populations, of psychosocial interventions 22:380–381
Target source identification, computational auditory scene analysis 2:226–227
Targeted Socio-economic Research Programme (TSER) 22:644
Target/facility removal 24:119
Targeting Social Need 1:249
Target–probe relation 10:316–318
Targets
performance measurement 16:771–772
terrorism 24:238–239
Target's perspective, stigma 23:464–465
Tariff duties, trade policy 24:485, 24:487–489
Ta'rikh 11:61
Tarling, N., works, *Cambridge History of Southeast Asia* 11:87
Tarrow, Sidney 22:435, 22:442–443
TAS *see* Thrill and adventure seeking (TAS)
Task analysis 12:811–812
Task assignment, and gender development 9:744–745
Task consultants 18:665
Task Force on Promotion and Dissemination of Psychological Procedures 19:383
Task leaders 23:430
Task network inference 16:542–543
Task performance, and personnel selection 17:936–937
Task value interventions
for academic motivation and performance 1:37–42
Task-instrumental behaviors 19:517
TaskRabbit 4:139–140, 8:487
Tasks
network models of 16:540–545
organizational groups 10:411
in simulation environments 4:488, 4:487f
Task-specific human capital 19:139
Task-switching costs 11:706
Taskwork 24:96–97
Tasmania 2:232
Taste cultures, elites 7:388
Tastes
and ANT 5:525
and appetite 24:26–27, 24:30
consumer, and cultural capital 5:537–538
consumption 4:757–760
cultural 5:729
fans, cultural production 5:610–611
intensity 11:421
judgment of 24:821
psychology of 24:26–31
cognitive and selective attentional effects 24:29–30
effects of variety on food intake 24:27–28
flavor representation 24:29
imaging studies in humans 24:29
olfactory–taste convergence to represent flavor and influence of satiety 24:29
pathways 24:27
pleasantness of taste of food 24:27–28
primary taste cortex 24:27
processing in brain 24:27–28
secondary taste cortex 24:27
sensory–specific satiety 24:27–28
texture of food 24:29
quality 11:420–421
receptors 24:26
reward and emotions 24:26–27
and social networks 16:556–557
and temporal lobe 24:208
TAT *see* Thematic apperception test (TAT)
tatistical pattern recognition 23:411–417
Tâtonnement 25:10
Tattoos 24:32–38
as body art 24:36–37
culture 24:32–34
current trends and research 24:35–37
as deviant behavior in Western society 24:34–35
prehistoric and non-Western tattoo practices 24:33–34
and secrecy 21:374–375
self-identity 24:36
TAU *see* Treatment-as-usual (TAU)
Tau protein 1:591
Tau-equivalent measurements model (TEMM) 19:428
Taussig, Michael 14:435, 14:437
works
The Devil and Commodity Fetishism in South America 9:107, 14:437
My Cocaine Museum 1:238
Tautological explanations 8:594
Tavani, H., works, *Ethics and Technology* 12:612
Tavistock Clinic (UK) 19:504–505
Tawakul 12:755–756
Tawney, R.H. 7:651, 22:699
works, *Religion and the Rise of Capitalism* 4:765, 20:260
Tax adjusting 12:517
Tax auditing 12:51
Tax competition 23:339, 23:342
Tax evasion 11:747
Tax expenditures 7:773
Tax farming 23:364
Tax fraud 5:145
Tax incentives 4:946–947
Tax policy
housing 11:244
for philanthropy 3:360–361
Tax pressure, shadow work 21:862
Tax Reform Act 1986 (US) 19:174–175
Tax returns, for philanthropy 3:359
Tax studies movement 5:269
Tax wedge 24:742
Taxable income 11:744
Taxation
and alcohol abuse 1:520–521

Taxation (*continued*)
economic psychology 5:888, 5:890
and health 10:200–203
and market allocation 14:542–543
state and local 23:338–342
tobacco control policies 22:128
and war 16:304–305
Taxation policy
implications, and intergenerational property transmission 19:175–176
Tax-base sharing, in urban planning 10:449
Taxes
in archaic states 23:372
compliance/evasion 15:738
income 11:744–749
on labor 24:742
and labor supply 13:193–195
Taxometrics 9:751
Taxonic gender differences 9:751
Taxonomies 2:559, 4:18, 18:83, 21:292
cannibalism 3:85, 3:88
five-factor model (FFM) 2:559
of revolutions 20:638
in social measurement 21:4
Taxpayer's money, for avoiding system meltdown 3:101
Taxpayers' revolt 24:846
Taxus baccata (yew) 7:692
Taylor, Charles 4:342, 6:591, 11:179, 19:677–678, 20:893, 20:897, 24:39–43, 25:385
of academic writings 24:39
on atomism 24:39
on epistemology 24:41
on ethics 24:42
and Habermas 24:39
on hermeneutics 24:39–41
on identity 24:39–41
and MacIntyre 24:39
on modernity 24:39, 24:42
on ontology 24:39, 24:41–42
professional life of 24:39
on secularity 24:42
on social justice 24:40
social scientific thinking of 24:40–42
account of strongly evaluative subject 24:41–42
critique of analytical philosophy and naturalist approaches to social sciences 24:40–41
understanding of modern social imaginary 24:42
works
Atomism 4:318
Hegel 4:318–319
Modern Social Imageries 24:42
The Politics of Recognition' Cannot Be Overrated 11:179
A Secular Age 21:391–392, 24:39–40, 24:42
Sources of the Self 21:446–447, 24:39, 24:42
Taylor, Frederick Winslow 11:870
on division of labor 6:602
on industrialization 11:870
works
The Principles of Scientific Management 13:160, 18:664
Taylor, Keith, works, *The Birth of Vietnam* 11:85–86
Essays into Vietnamese Pasts 11:85–86
Taylor, P.J., works, *Political Geography: World-Economy, Nation-State and Locality* 18:375
Taylor, T.D., on cultural consumption 22:336
Taylor, Thomas Griffith 7:724
Taylor, William, works, *Taylor's Dictionary of Bio-Psychology, Genetics, and Various Branches of Applied Psychology* 7:592
Taylor Nelson Sofres (TNS) Emnid (Bielefeld, Germany) 23:750–751
Taylor series expansion 23:414
Taylorism 17:341–342, 25:659
Taylorist capitalism 10:220
Tayuban dancing 17:711–712
TAZs *see* Traffic analysis zones (TAZs)
TB *see* Tone burst (TB)
T-bar probe 9:335
TBI *see* Traumatic brain injuries (TBI)
TBRS *see* Time-based resource sharing model of working memory (TBRS)
TBU *see* Tone-bearing unit (TBU)
TCAs *see* Traditional tricyclic antidepressants (TCAs)
TCC *see* Temporal context confusion (TCC); Transnational capitalist class (TCC)
TCE *see* Transaction cost economics (TCE)
TCI *see* Temparament and Character Inventory (TCI)
TCM *see* Three-component model (TCM); Temporal context models (TCMs)
t-congeneric tests 3:788
model 3:789t
TCS *see* Three-community Study (TCS)
TCs *see* Therapeutic communities (TCs)
TCT *see* Temporal comparison theory (TCT)
TCUs *see* Turn constructional units (TCUs)
tDCS *see* Transcranial direct current stimulation (tDCS)
t-distributions mixture model 15:626–627
TDM1 *see* Type 1 diabetes (TDM1)
TDM2 *see* Type 2 diabetes (TDM2)
TDR *see* Special Programme for Research and Training in Tropical Diseases (TDR)
Te Ao Maori 11:785
TEA *see* Transient epileptic amnesia (TEA); Transversely excited atmospheric (TEA)-laser technology)
Tea Act 1:610–611
Tea Party 1:606, 24:661
and gerrymandering 4:650
TEAC *see* Teacher Education Accreditation Council (TEAC)
Teach First (England) 24:52
Teach for America (TFA) 24:52, 24:76
Teacher autonomy, and education 7:146–147
Teacher behaviors
and student outcomes 24:44–50
Teacher burnout 7:294–295, 24:72–74
Teacher education 21:232–233, 24:51–55
Continuing Professional Development 24:54
continuum of 24:53–54
early childhood care and education (ECCE) 6:741
entry routes 24:52
history 24:51–52
induction 24:53–54
initial 24:53–54
OECD 24:51–52, 24:54
partnerships in 24:53–55
professional knowledge 24:51, 24:53–54
research and development in 24:54–55
role of university in 24:51–55
sites of learning in 24:52–53
and teacher quality 24:54
Teacher Education Accreditation Council (TEAC) 24:76–77
Teacher Education and Development Study in Mathematics (TEDS-M) 5:328–329
Teacher effects 4:796, 4:801
Teacher expectations
metaanalyses 1:45
and socioeconomic status (SES) 1:44–45
of students and labeling 8:134
Teacher expertise 24:56–59
interpersonal skills 24:57–58
pedagogical skills 24:56–57
to teachers and complexities of tacit knowledge 24:58
Teacher goals 24:67–68
Teacher induction 24:53–54
Teacher inquiry 4:798f
Teacher leadership 24:60–63
collegiality 24:60–62
distributed leadership 24:60
education 24:60–61
impact at school level 24:61
impact at student level 24:61–62
impact at teacher level 24:61
participative leadership 24:60
Teacher motivation 24:64–71
achievement goal theory (AGT) 24:67–68
empirical developments 24:64
expectancy-value theory (EVT) 24:64–67
future directions 24:68–69
goals 24:67–68
self-determination theory (SDT) 24:68
theoretical developments 24:64
Teacher preparation
in Vanuatu 6:741–742
Teacher quality, and student outcomes 7:152–153, 7:153f
Teacher ratings of learner's SRL 21:539
Teacher salary and qualifications 1:270
Teacher self-efficacy 21:520
Teacher sensitivity 3:833–834
Teacher values 24:64–67
Teachers 7:165, 7:167–169, 14:753–754, 24:44
assessing student progress 24:48–49
careers, Becker on 2:402–403

caring about students and respecting them 24:45
certification 24:44–45
course-taking 24:44
current knowledge 24:45
degrees 24:44
dispositions 24:45–46
efficacy beliefs 24:45
evaluations of student's behavior by 5:377
experience 24:45
feedback implementation 14:755
gifted education 7:160
impact of teacher leadership on 24:61
and information and communication technology 12:67
instructional delivery 24:47–48
differentiated instruction 24:48
effective instruction design 24:48
relevant instructions 24:48
knowledge 12:68
lesson planning 24:46–47
levels of cognitive challenge 24:48
and mastery learning 14:755
motivation and enthusiasm to teach 24:45–46
open education
as facilitators 17:228–229
in open education
characteristics, attitudes, and reactions 17:229–230
planning, relationships, and management skills 17:229
as partners in education 17:499
presage teacher characteristics 24:44–45
professional knowledge 24:51, 24:53–54
professionalism and professional growth 24:46
qualification of 7:168
role in nurturing productive giftedness 7:277
role in self-regulated learning 21:545
role of, social-cognitive perspectives 13:674–675
socialization and professional development of 21:137
stress 24:72–74
coping 24:73–74
definition 24:72
extent of 24:73
measurement 24:72–73
sources 24:73
and student achievement 5:343–346
student engagement, role in 24:48
textbook use by 24:257
use of cooperative learning 10:420
use of questioning 24:48
verbal ability 24:45
see also classroom management
Teachers for a New Era (TNE) initiative 24:54–55
Teachers' practices 22:327–328
Teacher–student interaction
collaborative learning and patterns of dialogues 3:916–917
instructional conversations and dialogic teaching 3:915–916
tutoring and scaffolding 3:915
Teaching 3:176, 22:611
active learning 3:176
case design 3:177
cases 3:177
computers 4:518
economics 24:83–88
and education 1:728–729, 1:732, 7:144, 7:146–147
entry routes 24:52
foreign language 9:327–332
higher education research 10:864–865, 10:867
infrastructure 22:667–673
mathematics 14:819–821
for metacognition 24:89–95
open education 17:228–230
important functions 17:228–229
in United States 24:75–82
accountability 24:75, 24:78
advanced certification 24:79
alternative certification programs 24:76
career making 24:79–80
credentials 24:76, 24:79
expertise 24:75–77, 24:79
high-quality teachers 24:75–76
induction 24:78
inequity in working conditions 24:77–78
pay 24:78–79
quality teaching and licensure 24:76–77
response to traditional teacher education 24:76–77
school as workplace 24:77–79
second stage roles 24:79
teacher quality and attrition 24:78–79
teacher retention and attrition 24:78
training and licensing 24:75–77
unions 24:77
working conditions 24:77–79
workplace authority 24:79–80
universities 18:307–309
Teaching and learning 21:102
Theory of Mind and 4:39
Teaching and Learning International Survey (TALIS) 5:328–329, 5:774, 24:51–52
Teaching innovation programs (TIP) 7:23
Teaching machines 22:27–29
Teaching methods
Bologna Process 2:743
economic education 7:22–23
Teaching new skills, behavior therapy 2:428
Teaching practices for social studies education 22:731–732
Teaching quality 21:121
Teaching through Interactions framework 3:833–834
emotional support domain 3:833–834
Teachman, Jay, on cultural capital 5:377
TEAL *see* Technology-Enabled Active Learning (TEAL)
Team assessment 24:96
Team Assisted Individualization (TAI) 4:884
Team effectiveness 24:97
Team for the Assessment of Psychiatric Services (TAPS) 6:49–50
Team mental models 10:409
Team outcomes 24:97
affective outcomes 24:97
performance outcomes 24:97
Team performance measurement 24:96–103
future of 24:102
distributed virtual teams 24:102
dynamic assessment of team constructs 24:102
overview 24:96–97
inputs 24:96
mediators 24:96–97
outcomes 24:97
selection/development considerations 24:97–99
construct selection 24:98
measure referent 24:98
measure validation 24:99
measurement purpose 24:98
measurement source 24:98–99
measurement timing 24:99
tools 24:99–102
automated performance measurement 24:100–101
behaviorally anchored ratings scales (BARS) 24:100
concept mapping 24:101–102
event-based measurement scales 24:100
rating scales 24:100
Team processes 24:96–97
Team reasoning 12:843
Team rewards 4:883
Team science 22:461, 22:463
Team teaching 21:115
Team theory 5:881
Team work, nature of 24:106–107
Teams 21:114, 24:96
cross-cultural training 5:339–340
groups 10:434, 10:436–438
personalized medicine 17:930
and workplace autonomy 25:720
Teams emotions, in workplace 7:509
Teams-Games-Tournament (TGT) 4:883
Teamwork 24:96–103
and ergonomics 11:301
in ESR 7:562
Teapot Cult 16:806
TEC *see* Theory of Event Coding (TEC)
Technai 2:46–47
Technical analysis of stock market trends 9:317–318
Technical and scientific evidence 12:910
Technical and vocational education and training *see* Vocational education and training (VET)
Technical and vocational skills development (TVSD)
conflict, education, and youth 4:635–636
Technical eclecticism 16:73, 19:511, 19:513

Technical environment of hospitals 11:209–210
Technical expertise 22:695
Technical innovation 21:150
Technical learning, policy change 3:843
Technical means 24:117–118
Technical practice, ethnomethodology of 21:197–198
Technical reasoning 1:878
Technical requirements for health information systems 5:786
Technical skills 3:11–12
Technical state 13:105–106
Technical–economic system 11:351, 11:881–882
Technical–social dualism, engineering 7:642–647
Technocentric 7:780–781
Technocracy 12:88, 21:207, 21:210
  cybernetics 5:648
Technocratic society 15:576
Technocratization 12:749
Technoeconomic concepts of energy and exergy efficiency 7:630
Technological advances
  health surveys 10:729
  in police use of force 18:257–258
Technological artifacts 24:141
Technological cannibalism 3:85
Technological challenges, in publishing 19:623
Technological changes 15:595
  applied psychology of 24:104–108
    data mining 24:107
    diversity 24:104–105
    impact of social media on employment practices 24:106
    implications for culturally diverse social and managerial systems 24:105–107
    information technology, Internet, and digital revolution 24:104–105
    nature of individual and team work 24:106–107
    research issues 24:107
    robotics 24:104, 24:106
    social, cultural, and managerial systems 24:105
    use of behavioral analytics 24:107
  to citizen journalism 3:614
  in industrial ecology 11:845
  industrial geography 11:854–855, 11:857
  and innovation 12:171
  media and 22:585
  and unobtrusive measures 24:772
Technological convergence, and telephone 24:161
Technological cultures 24:139
Technological determinism 24:109–112, 24:136, 24:138–139
  concept history 24:109–110
  historical materialism 24:109–110
  impact-imprint' model 24:110–111
  Marx on 24:110
  political artifacts and systems 24:111
  reconsidering 24:111
  social construction of technology 24:111
  sociology of technology 24:109, 24:111
  sociotechnical systems 24:111
  users 24:111
  varieties 24:110–111
  Veblen on 24:109–110
  Winner on 24:109–111
Technological developments 17:851, 21:150, 22:874–875
Technological directions 12:149
Technological education, and international organizations 12:539–540
Technological environments, and aging 6:887
Technological frame 24:137–138
Technological hazards 4:356–357, 6:481
  community resilience and 4:357–358
  urban life health 24:831
Technological imperative, health care organizations 15:101
Technological innovations 3:22–23, 12:147
  and industrialization 11:868–871
  sustainable development 23:825–826
    implications of 23:826
  system, for renewable energy 18:370
Technological learning, for renewable energy 18:369–370
Technological lineages 24:130–131
Technological paradigms, dominant design 6:637, 6:640
Technological pedagogical content knowledge
  Computer Assisted Instruction (CAI) 4:496
Technological progress 24:141
  and industrial societies 11:863–864
Technological revolutions, biotechnology 2:676–679
Technological space, biotechnology 2:677–678
Technological spillovers, defense economics 6:3–4
Technological unemployment 12:152–153, 24:741
Technological-themed explanans, in archaeology 4:24
Technologies 1:204–205, 2:300, 4:139, 7:58–59, 14:187–188, 17:780–781, 21:364
  and adult education 1:151
  agriculture 1:479
  in agriculture 17:671
  artistic/cultural markets 14:567–568
  for building stories 16:219
  and business 3:3
  change, economic analysis of 7:56–57
  in classical archaeology 3:762
  as composite 11:868–869
  contingency of 23:610–611
  cultural perspectives of 11:863–865
  definition 24:113
  descent with modification 24:129
  and division of labor 13:159
  economic statistics 5:758
  and energy 7:615–619
  engineering 7:645
  ethnomethodology 8:205–206
  evolution of 24:129
    case study from early technology 24:131–132
    cost–benefit analysis 24:130, 24:132
    cultural evolution, processes of 24:129–130
    cultural transmission 24:132
    cumulative evolution of technology 24:132–133
    drift effects 24:130
    effective population size 24:131, 24:133
    fitness landscapes 24:131
    invention and innovation 24:131
    optimality 24:132
    social, cognitive, and political factors in 6:640
    social learning 24:130
    technological lineages and human populations 24:130–131
    transmission fidelity 24:132
    unbiased transmission 24:130
  and films 9:165, 9:168–170
  and gender 9:723–728
  history of 11:104
  household economics 11:225
  human ecology 11:285–286
  hybridization in 11:448–449
  impact on politics 14:716
  impacting supply chains and big data 23:712
  and interviewing 12:674–675
  journalism 12:855
  and law 13:520–527
  and law firms 13:582–583
  military anthropology 15:486
  Mumford on 16:137–140
  and music 16:152
  and neural basis of semantic memory 21:573–574
  opportunity, and innovation 12:163–165
  and organization 24:113–116
    conceptualization and meaning 24:113–114
    empirical studies and theorization 24:115
    input–conversion–output system format 24:114
    relevance to managing and management 24:114–115
  in organizations 2:770
  and photography 18:85–86
  and population dynamics 18:566–569
  postcolonial science studies 18:652–654
  and postmodernism 18:676–678
  and privacy 12:593–594
  and production mode 15:714–715
  role in sex selective abortion 21:687
  and sex work 19:280
  and social change 22:194–195
  social construction of 21:202–203, 24:135–140
  and social control 24:117–124
  social impact of 21:201–202
  in social structure 17:673

sociotechnical approach to 5:523–524
substitutes for legal expertise 12:925
and successful aging 23:671
and territorial conflict 24:217–218
turn in science studies 21:202–203
urban settlements 24:836–837
and venture capital 25:59–62
visual methods 25:182
visualization in 25:220–225
and war 16:304, 16:307–308
warfare
classical period 25:414
contemporary period 25:418
Middle Ages 25:415
nineteenth century 25:417
weapon 4:611–612
Technologies of selves 5:647–648, 16:218–219
Technology, Pedagogy, and Content Knowledge (TPACK) 12:68
Technology adoption, and international immigration 12:516–517
Technology Adoption Model (TAM) 12:69
Technology assessment (TA) 24:125–128
agenda building 24:126
bridging 24:126
context 24:125
future directions 24:127–128
methodological issues 24:126–127
path dependency 24:127
philosophy of 24:125
situating 24:127–128
strands of 24:125–126
usages 24:125
Technology Cycle Time (TCT) 20:499
Technology forcing strategy 22:875
Technology forecasting 24:125
Technology policy 24:125
Technology transfer (TOT) 21:651–652, 24:141–145, 21:652f
biotechnology and 24:144–145
Bush's writings 24:141–142
controversies around *neem* and *basmati rice* 24:144
critical discourses 24:142
information technology and 24:144–145
of military technology 24:142–143
risk of development and 24:143–144
use of term 24:141
Technology transfer offices 18:310
Technology treadmill theory 1:472
Technology-based education 12:66–71
Technology-economic innovation 12:173
Technology-economic perspective 12:172
Technology-enabled active Learning (TEAL) 6:897
Technology-enhanced learning (TEL) 14:93–94
Technology-integrated teaching media, for economic education 24:85–86
Technology-Rich Outcomes-Focused Learning Environment Inventory (TROFLEI) 3:827–828
Technomethodology 21:197
Technophilia 1:475
Technoscience, postcolonial study of 18:654–655
Technosocial, IP law as contingent on 12:266
Teddy Bears (Bear community) 13:889
TEDS-M *see* Teacher Education and Development Study in Mathematics (TEDS-M)
Teenage pregnancy 1:128–129, 1:132, 20:687
Teenagers
mobile communication use by 15:631–632
social media use by 24:146–152
see also Adolescents
Teeter, Robert 18:461
TEI *see* Text encoding initiative (TEI)
Teichopsia 10:501
Teknonyms 16:184–185
TEL *see* Technology-enhanced learning (TEL)
Telecommunications 5:704–705, 24:153–155
electrical communication 24:155
long-distance 24:154
satellite 24:155
wireless 24:155
Telecommunications Act 1996 19:880–881
Telegraph 12:491, 24:153–156
in Africa 24:154–155
in Asia 24:154–155
in Great Britain 24:153–154
submarine telegraphy 24:155
in USA 24:154
Western imperialism 24:153
Telemarketing, and telephone 24:158
Teleobiological theories, of intentionality 12:334–335
Teleological approach in social science 21:531
Teleological explanations 8:594
Teleology 5:677–680, 8:74, 9:108, 17:4
and value of life 25:16
Teleo-semantics 12:334–335
Telephone 12:490, 12:492–493, 24:153–155, 24:157–162
bank 24:159
computers 24:158, 24:161
democracy 24:157, 24:159
economic development 24:158–159
gender 24:159–160
government 24:159
geographical distribution 24:158
global and microcoordination 24:157–158
harassment 24:161
innovation 24:162
mobile phones 24:157–161
policy and politics 24:159
scandal 24:160
sexual behavior 24:160
societal aspects 24:159–161
surveillance 24:160
as synecdoche for modern condition 24:161–162
technological convergence 24:161
telemarketing 24:158
wireless 24:157, 24:159–161
wireline 24:159, 24:161
Telephone bank 24:159
Telepsychiatry 8:67–68
Telescoping 20:883
in amputees 18:2
Telescoping problems 5:719
Televangelism phenomenon 20:363
Television 12:492, 15:30, 22:585, 24:163–172, 25:80–86
audiences 24:166, 25:82–83
broadcasting 19:877, 19:879–881
broadcasting regulation 2:859–863
and celebrity 3:275–277
and child development 15:1–8
and communication 4:273–275
content effects on children's non-physical aggression 15:38
convergence culture 25:81–82
distribution 25:80–82, 25:84–85
fandom 25:82–83
and free time 13:846
and global capitalism 3:102
globalization 25:83–84
history of 24:172
impact on politics 14:716
and media 15:12–16
media technologies and social life 25:81–82
and media sports 14:719–722
mediation 25:80–81
mediatization 25:80–81
medium and modernity 24:163–164, 24:166–168
middle-range approaches 25:84
policies 24:164–165
and popular culture 18:528–531
production 24:165, 25:83
programs 24:165–166
and social capital 22:178
viewing–aggression relationship 25:101
Television message systems 15:30
Television programming
cognitive effects of exposure 15:35–36
Television ratings, audience measurement 2:214
Television talk shows 15:61–62
evaluations and critical perspectives 15:62–63
variants of 15:61
Television viewing 15:30
and cognitive development 8:732
as main leisure activity 13:846
Television violence, exposure to 15:38
Television watching habits 20:884
Telework, transportation 24:624–625, 24:627
Tell el-Amarna (archaeology site) 1:693
Telling history, ways of 2:644
Tello, Julio C. 23:37
Telos 14:649
TEMM *see* Tau-equivalent measurements model (TEMM)
Temper, ceramics in archaeology 3:325
Temper loss, early developmental trends in 6:12–13

Temperament and Character Inventory (TCI) 9:1010
Temperament development 24:191–198
behavioral inhibition 24:191, 24:193–196
current methodologies and research issues 24:194–195
cultural influences 24:194–195
measurement issues 24:195
and developmental psychopathology 24:195–196
temperament-by-temperament moderation 24:195–196
temperament–environment interactions 24:196
dysregulated fear 24:195
effortful control 24:193, 24:195–197
socioemotional adjustment 24:191, 24:194, 24:196–197
theoretical approaches 24:191–194, 24:192t
Buss and Plomin approach 24:191–193
Goldsmith approach 24:193
Kagan approach 24:193–194
Rothbart approach 24:193
Thomas and Chess approach 24:191
Temperament dimensions, and adolescent health 1:123
Temperamental task orientation 24:185–186
Temperament-by-temperament moderation 24:195–196
Temperament–environment interactions 24:196
Temperaments 2:456, 2:461, 16:493–494, 17:774–775, 22:493, 22:495, 23:579–580, 24:173–190
achievement goals 24:187–188
adjustment 24:177
applications 24:177–178
attention regulation 24:185–186, 24:188
biological bases 24:175–176
children, interactions between parenting and 2:474
concept 24:184
concept of 17:876
and conscientiousness 17:788
context 24:176
in context 17:876
continuity of 16:494
continuous and categorical views of 17:876–877
contribution to personality 17:877–878
culture 24:176–177
current research 24:181–182
definition 24:173, 24:180
development 24:176
development of 17:877
and emotional stability 17:788–789
environmental factors 24:185
factors 22:494, 22:496–497
future directions of theory and research 24:183
gender differences in 9:752
genetic background 24:184–185
and health 17:792, 17:794–795
history 24:173–174
individual differences 24:180–182
measurement 24:175
methodological issues or problems 24:182–183
and motivation, interaction between 24:187–188
origins 24:184–185
outcomes 24:176
parenting 24:177
personality 24:180–183
and personality 17:827–832
research, history of 24:180–181
risk patterns during infancy and early childhood 6:13–14
role of culture 17:878
self-regulation 24:185–186, 24:188
sex 24:176–177
and sexism 21:695
significance in academic achievement 24:185–186
social behavior 24:180
and social dominance orientation 22:250, 22:252
and social referencing 22:560
socialization 24:180–183
structure 24:174–175
use of term 17:876
Temperance movement 17:240
Temperature affected dating methods 3:567
Temperature and precipitation indices 3:867
Temple cult 19:197
Temples, rituals 20:724–725
Tempo (fertility) 9:10–12, 9:19, 15:397–398
Tempo adjustments 5:739
demography of 6:114
Tempo effects 5:739
Temporal and spatial summation 21:605
Temporal bisection 24:308
Temporal comparison theory (TCT) 4:412–413, 6:430–431
Temporal concepts, linguistic differences in 13:297–298
Temporal context confusion (TCC) 4:563
clinical evidence for 4:563–564, 4:563f
Temporal context models (TCMs) 15:164, 24:201–202, 24:204
Temporal cortex and working memory 25:706–707
Temporal deixis 6:52
Temporal differences, modeling 19:443
Temporal discrimination
impairment 6:719
memory as 24:202–203
Temporal distinctiveness models of memory 24:199–205
evidence 24:202–204
memory as temporal discrimination 24:202–203
temporal isolation effects 24:203–204
time and forgetting 24:203
history 24:199–201
early global distinctiveness models 24:200–201
local distinctiveness models 24:201
temporal context models 24:201–202
interference 24:200–201, 24:203–204
serial position curves and forgetting 24:202–204
temporal organization of memory 24:202
wilder theoretical implications 24:204
consolidation 24:204
LTM and STM 24:204
Temporal domains of remote sensing 10:65
Temporal dynamics in population heterogeneity 17:732–733
between-group and within-group variance 17:733
origins and trends in social inequalities in health 17:732–733
Temporal gyri, middle and inferior 13:371
Temporal isolation 24:203–204
Temporal lobe 13:371
anatomy of 24:206, 24:206f
and anomia 19:150–151
functions of 24:207–208
auditory input processing 24:207
emotional regulation 24:208
gustatory input processing 24:208
memory 24:208
visual input processing 24:207–208
injury, symptoms of 24:208–209
intracarotid sodium amytal testing 12:690–691
personality 24:209
pitch discrimination 24:208
speech recognition 24:207–208
taste 24:208
and transient global amnesia 24:565–567
ventral stream 24:207–208
and visual mental imagery 25:163–165
Temporal lobe epilepsy (TLE) 7:854, 24:554
Temporal logic, knowledge representation 13:101
Temporal memory, and prefrontal cortex injury 18:815
Temporal modulation 19:64
Temporal motivation theory, and procrastination 19:76
Temporal order, serial order memory 21:639, 21:641
Temporal organization
of behavior 18:813–814
of memory 24:202
Temporal processing, and language learning impairment 13:347–348
Temporal reference, TAME 24:211–212
Temporal regions 13:370
Temporal resolution 20:420, 20:424
Temporal scale 21:1–3
Temporal units, psychotherapy 19:517
Temporality
anthropology 1:747–749
causal inference 3:258
Dilthey on 6:425–429
Temporality models, of confabulation 4:561–564
imaging evidence for 4:564

Temporal–parietal junction (TPJ) 24:206–207
Temporary Assistance for Needy Families (TANF) 9:15, 14:699, 18:725, 18:736, 18:749
Temporary equilibrium approach, for expectations 8:510–511, 8:514
Temporary memory 25:715
Temporary migration 10:169–170, 10:170t
Temporary protection visas (TPVs) 20:114
Temporary visual memory 25:716
Temporoparietal area 16:691–692
Temps court 1:710–711
Ten Days That Shook the World 20:628–629
Tenant farming 24:529
Tenant improvements (TIs) 20:17
Tenement reform 16:450
Tenet of psychic pluralism 5:479
Ten-factor model of productive giftedness 7:275–276
Ten-Item Personality Inventory (TIPI) 4:660, 9:232
Tennessee STAR trial 8:542
Tennessee Valley Authority (TVA) model 1:219, 7:759–760, 17:302–303, 20:142–143, 25:439–440, 25:560–561
Tennosei 14:637, 14:639
Tenrikyo 16:806–807
Tense 21:587
Tense, aspect, mood, and evidentiality (TAME) 24:210
  event 24:211–212
  and finiteness 24:210
  inflection 24:210–212
  morphology 24:211
  perfect 24:212
  temporal reference 24:211–212
  typology and diachrony of 24:210
Tense and aspect 8:352
Tense constraints 1:188
Tense logic 13:103
Tense-aspect-mood ('TAM') category 16:413
Tent embassy, Australian aboriginals 2:235
Tenure 11:233–234, 11:237
  academic careers 1:28–29
    academic freedom 1:28–29
    economics of 1:29
    in North America 1:27
  academic research and employment 1:50–51
Tenure track 1:27–28
Tepper, Steven J., on cultural consumption 16:557
t-equivalence 3:787t
  tests 3:788
    model 3:788t
Term rewriting 5:936–937
Term work 25:669
Terman, Lewis Madison 7:158, 12:303, 12:307
  on intelligence 12:310–311
  and Stanford-Binet test 3:879
Terminal attractors 15:960
Terminal illness 5:861
Terminal sedation 7:601
Terminal values 25:28, 25:49
Terminal vertex 16:543
Terminator seeds 9:1004
Terminological logics 14:319
Terra Madre project 22:95
Terra nullius 2:234–235, 2:237, 4:225
Terrain analysis 10:101, 15:502
Terrain contour maps 24:266
Terralingua 1:224
Terrestrial systems 3:851
Territorial borders 2:761–762
Territorial conflict 24:214–220, 25:819, 25:821
  correlating conflict and territory 24:215–216
  disaggregating conflict 24:216–217
  sources 24:214–215
  and technologies 24:217–218
  territorial perspective 24:215
  theoretical approaches 24:217t
  theorizing 24:218–219
Territorial expression, market area as 14:544–545
Territorial planning 20:137–138
Territorial sovereignty 10:169
Territorial sovereignty doctrine 25:446
Territorial stigmatization 24:818–819
Territorial units, and community studies 4:369
Territoriality 18:251, 18:254, 18:375–377, 20:775–776, 24:218–219
  and fetishism 9:107–108, 9:110
  intellectual property law and 12:268–269
  land tenure 13:236
Territorialization, refugees 20:112–115
Territory 2:761–762, 10:91, 18:374–377
  and anthropology 1:751–752, 1:755
  belonging 24:224–225
  borders 24:221–223
  conflict 24:223–225
  emergence of 24:222–223
  and ethnic conflict 8:91–94
  identity 24:221, 24:224–225
  land tenure 13:231–232, 13:234, 13:236
  nation 16:223–225
  place and nation 24:224
  power 24:221–225
  and regional geography 20:119
  in regional geography 20:124–127
  resistance 24:225
  sovereignty 24:223–224
  space and society 24:224–225
  of state 23:356–357, 24:223–224
  and war 25:420–426
Territory-wide cleavages 20:747
Terror 25:372–374
  French Revolution 9:409–411
  social work stress/trauma 18:707–710
Terror management theory (TMT) 17:817–818, 21:524, 23:448
Terrorism 3:678, 5:837–840, 10:3, 12:466–467, 13:528, 16:358, 22:430, 23:366, 24:237–243
  al-Qaeda 13:528
  antiterrorism actions 13:532
  causes of 24:239–240
  and childhood stress 23:571
  counterterrorism 24:237, 24:239–242
  definition 24:237
  domestic *see* Domestic terrorism
  geography of 25:371–372
    embodiment of 25:372–373
    and nature 25:373
    and peace 25:374
    place of war 25:373–374
  history of 24:237–238
  and ideologies 11:573, 11:577
  individual motivations 13:529
  international law 13:530
  and Internet 24:227–236
    chat rooms 24:229–230
    cyberterrorism 24:230
    Facebook 24:231–233
    mass-mediated terrorism 24:227–228
    new terrorism and new media 24:228
    online networking 24:228–229
    political violence 24:229
    propaganda 24:227–235
    Twitter 24:230–231
    web sites and forums 24:228
    YouTube 24:233–234
  law enforcement approaches to 13:530–531
  methods and targets 24:238–239
  military action in response to 13:533–534
  motivations and operations 13:529
  non-state-sponsored 13:529
  pertinent law 13:529–530
  preconditions 24:240
  preventive prosecution 13:532–533
  processes into 24:241–242
  processes which end 24:242
  purpose of attacks 24:239
  and religion 20:403–405
  social movements 24:240–241
  social process 24:240–241
  state-sponsored 13:529, 13:531
  targeted killing 13:530, 13:534
  trigger points 24:240
  in United States and beyond 13:528
  urban 24:953–957
  and violence 25:118, 25:120
  war on terror 24:240
Terrorist attacks 24:239
  social work response, in Australia 7:430
Terrorist organizations 24:228
Tertiary circular reaction 16:507
Tertiary institutions, for social and behavioral science research 22:618
Tertiary prevention programs 18:878, 18:884
Tertium comparationis 4:408, 4:441, 5:323
Tertius gaudens 16:546
Tertullian 3:548, 20:285
TEs *see* Transition economies (TEs)
Test anxiety 24:244–249
  and academic achievement 1:23–24
  assessment 24:244
  biological factors 24:246–247

Test anxiety (*continued*)
component processes 24:244
affective 24:244
behavioral 24:244
cognitive 24:244
motivational 24:244
physiological 24:244
concepts of 24:244–245
coping 24:248
defined 24:244
development 24:246
emotion regulation 24:248
impact on learning and academic achievement 24:245–246
empirical evidence 24:245
mediating mechanisms 24:245–246
reciprocal causation 24:245–246
theoretical explanations 24:245–246
individual appraisals 24:246–247
and procrastination 19:76–77
psychotherapy 24:248
relevance of 24:244
role of education 24:247–248
learning environments 24:247
tests and exams 24:247
Test Anxiety Inventory 24:245
Test Anxiety Questionnaire (TAQ) 24:245
Test assembly, computerized test construction 4:511–512
Test Ban Treaty 7:811–812
Test construction 19:436
computerized 4:510–514
Test drives 21:183
Test error 16:118
Test for Auditory Comprehension of Language 3:796
Test functions 15:298
Test information function (TIF) 4:511
Test length effect, parallel measurements model 19:426–427
on criterion-related evidence of validity 19:427
Test of Attitudes and Interests 21:695
Test of Auditory Processing Skills 3:796
Test of Early Language Development 3:795
Test of Early Written Language 3:796
Test of Functional Health Literacy 10:654
Test of Language Development 3:795
Test of Memory and Learning, Second Edition (TOMAL-2) 16:725
Test of Narrative Language 3:796
Test of personal intelligence (TOPI) 17:754
TOPI MINI-12 17:754
Test of Self-Conscious Affect 21:477
Test of Understanding College Economics (TUCE) 7:22
Test of Written Language - Fourth Edition 16:725
Test performance 17:850
Test score variables 3:785
Test scores 19:439
for educational gender gaps 9:682–683
and educational outcomes 5:377
Test theory 19:439, 24:250–255
classical test theory 19:423–430
Testamentary conception 19:173–174
TESTFACT *see* Full information item factor analysis (TESTFACT)
Testicular cancer 15:170
Testimonio 8:181
Testimony 25:397
cognitive development 4:39
eyewitness 8:642–649
of forensic psychologist 9:353
in interviews 12:674
scientific evidence and 21:297–301
Testing
linear hypothesis 14:146–147
models 14:146
parameters 14:146–147
models 15:653–658
Testing effect 12:451
Testlet response theory 19:439
Testosterone 2:692–693, 21:773–774
Testosterone patches 21:692
Test–retest reliability 6:345, 9:834, 12:250, 19:130, 20:243
Tests
for independence/differences in common use 19:726–728
student achievement 5:347
Test-taker perceptions 22:13–15
Test-Wait-Test-Exit model 19:272–273
Tetrahydrocannabinol (THC) 16:170
Tetris computer game 12:304
Tetsugaku 24:584
Tett, Gillian, works, *Fool's Gold* 8:902
TEU *see* Treaty on European Union (TEU)
Teubner, Gunther 13:555, 13:566, 13:811
Texaco 17:166
Milford Haven refinery explosion 11:300
Texas Instruments 11:830
Texas Railroad Commission 17:166
Text analytics, and mass media 14:715
Text coding 9:362
Text comprehension 13:402
Text deixis 6:53–54
Text encoding initiative (TEI) 4:951
Text interpretation, knowledge representation 13:100
Text messaging 4:300
Text mining, journals 12:867
Text reading comprehension 3:511–512
Text structure, and linguistic anthropology 14:171
Text-based virtual-reality multiuser domains (MUDs and MOOs) 11:549–550
Textbooks 7:868, 24:256–259
commercial object
authorship 24:258
concentrations of expertise and production capacity 24:258
marketing and adoption 24:258
physical characteristics 24:258
control of content 24:257
imported textbooks 24:257
international concerns 24:257
cultural artifact
class structure 24:258
latent curriculum 24:257–258
definition 24:256
future of 24:258–259
grade level differences 24:257
historical development 24:256
history of 14:215
psychology 21:129
subject area differences 24:257
surrogate curriculum 24:256–257
use by teachers 24:257
Textile production in Middle Ages 5:108
Texting, oversharing through 17:439–444
Text-level reading 16:593–594
Text-linguistics 15:826–827
Texts 20:666
analysis, and hegemony 10:745
in anthropological writing 1:741–743
British cultural studies 2:853
and discourse studies 6:510–511, 6:513
folklore 9:282–283
Textsetting 13:275
Textual communication and interaction 24:149–150
Textual communities 25:783
Textual genres 14:242
Textualism 12:932, 13:743
Textuality, visuality and 25:209–210
Textualization, folklore 9:283
Textualizing the world 9:464
Texture of food 24:29
TFA *see* Teach for America (TFA)
TF-CBT *see* Trauma-focused cognitive behavior therapy (TF-CBT)
TFP *see* Total factor productivity (TFP); Transference-focused psychotherapy (TFP)
TFR *see* Total fertility rate (TFR)
TFRf *see* Two-sex total fertility rate (TFRf)
TFT *see* Tit for tat (TFT)
TFT-LCD (thin-film transistor-liquid crystal displays) 20:498–499
TGA *see* Transient global amnesia (TGA)
TGT *see* Teams-Games-Tournament (TGT)
Thai *Mulasasana* (around 1420) 11:82
Thai Muslim community 12:729
Thailand 23:63, 23:67
Bronze Age archaeological 23:66
Buddhism in 2:891–892, 3:532
cave sites 23:65
farming communities 23:66
fertility transition in 9:85–89
nationalist history in 11:85
Neolithic sites 23:66
sociocultural aspects of *see* Southeast Asia
Thalamo-cortical projections 3:343f
Thalamotomy 23:18
for Parkinson disease 23:20
Thalamus, interactions of cerebral cortex with 3:340–343, 3:342f
Thalassemia, genetic screening 2:620
Thaler, Richard, works, *Toward a Positive Theory of Consumer Choice* 2:446
Thales 19:406, 20:284
Thanatology 5:860
Thatcher, Margaret 19:158
resignation, and flashbulb memory 9:257

Thatcher administration
tax system 23:340
Thatcherism 16:465, 24:866
That's so gay phrase 21:87
THC *see* Tetrahydrocannabinol (THC)
Theater 17:713–714
in MENA 16:361, 16:364
and social action 17:710–711
Turner's legacy in 17:709–710
Theater states 2:74–75
in Southeast Asia 23:71–72
Theil index 11:906, 21:412
Theil's measures 11:886t
Themata, social representations 20:446–447
Thematic analysis 22:546
Thematic apperception techniques 15:951–952
Thematic apperception test (TAT) 15:887, 15:894, 16:725, 19:131–132
historical background 3:879
Thematic hierarchy 16:672
Thematic maps 3:171–172, 3:174, 24:260–267, 3:172f
Thematic narrative analysis 16:206
Thematic research, youth 25:825–826
Thematic role 10:324
Thematic scale 21:1–3
Thematic social science surveys 5:847–848
Thematic specialized social science surveys 5:844
Theme analysis 16:196
text analysis 4:770–771
Theme-based research, higher education 10:865–866
Theme-by-theme network 5:560
Themes 12:96
conductivity of 4:770
witchcraft 25:607–608
Theocracy 6:95
Theodicy problem, Albert on 1:499
Theo-linguistics 14:228
Theology 2:631
historical *see* Historical theology
liberation 14:9
and spirit 9:666
use of term 9:666
Theology of Cautiverio 14:33–34
Theology of Liberation 3:537–538
of the comic 11:418
Theophrastus 13:799
Theorein 8:212–213
Theoretical accounts 10:317–319
Theoretical armamentarium
human behavioral ecology 11:247–248
Theoretical chemistry 10:755
Theoretical criminology *see* Academic criminology
Theoretical definition 11:753
Theoretical developments, teacher motivation 24:64
Theoretical frameworks, engagement 1:32
psychological and motivational theory 1:32
school dropouts 1:32
Theoretical genres 9:162
Theoretical habitus 10:484–485
Theoretical integration 19:511–512
Theoretical knowledge 13:93
and scientific revolution 21:319
Theoretical models
of reading and writing 6:736–737
for work-related stress 23:547–549
Theoretical orientations, of psychotherapy 19:509–510, 19:513
Theoretical perspectives
on cooperative learning 4:881–882
psychological contracts 19:367
equity theory 19:367
group value model 19:367
social exchange theory 19:367
sociological theory 22:955–962
Theoretical pluralism, Feyerabend on 9:119–120
Theoretical primacy 1:75
Theoretical reasoning 18:776–777
Theoretical revolution, geography of 10:84–85
Theoretical role interpretations of probability 19:27
Theoretical sampling 10:405–406
Theoretical sociology 23:23
Theoretical tensions, Southeast Asian archaeology 23:67–68
Theoretical Treatise on Natural Law 2:670–671
Theoria 1:899–900
Theories 1:899–900, 14:225
of action 1:672, 7:578, 14:418, 22:419, 24:510
of aging 1:435–439
in archaeology 1:899
and contemporary context 1:903–904
law of development 1:900–901
method 1:899–900
scientific archaeology 1:901–902
troubles with science 1:902–903
of argument structure 13:944
big five theory 2:561
building personality theory 2:559
of childhood friendship 9:442–443
communes anthropology 4:271–272
of concentric zones 7:561
of conformity
and genocide 10:18–19
of constructivism 4:532–533
of democracy 12:368–369
of demographic transition 1:327
importance of 1:327
stages of 1:327
of deterrence 6:260
of deviance 6:473
dimensional comparison theory (DCT) 6:430
of distinctive features 12:782–783
of divine right 4:521
of domination 9:141
of dynamic conceptual mappings 4:667
of economic backwardness 11:104
of evolution 21:679
of experts 8:589
of facial emotion expressions 8:667–668
of family solidarity 12:403–404
of firm 4:942–943
of fluid and crystallized intelligence 3:246
of grammar, mathematical psychology 14:814
of language 4:294–296
of love as story 14:379–380
and measures 19:700
and methods 1:899–900
objective 17:88
of overpunishment 19:636, 19:644–645
of psychopathology 19:344
psychotherapy 19:515, 19:517–518
of *rechtsstaat* 13:432
of revolution 20:627–628
definition 20:627–628
explanation 20:628–630
validation 20:630–631
of SAD 22:496
of science 21:595
of signs 21:595
social comparison theory (SCT) 6:430
of social selection 22:934
sport history 23:308–309
of State 9:144–147
of stimulation 14:459
of succession 7:561
temporal comparison theory (TCT) 6:430
of thematic relations 13:944
of the unconscious 19:326
of underpunishment 19:636
of war consequences 25:401
of war economy 25:90
Theories of leadership 13:655–657
Theories of sexuality
behaviorism 21:847–848
biopsychosocial approach 21:850
cognitive theories 21:848
Dual Control Model (DCM) of 21:850–851
evolutionary psychology 21:846–847
feminist theory 21:851
importance 21:846
integrated theories 21:850
queer theory 21:851
research 21:846
social learning theory 21:848
social theories 21:848–850
script theory 21:849–850
social constructionism 21:849
social exchange theory 21:848–849
sociocultural research 21:849
sociobiology 21:846–847
symbolic interactionism 21:851
Theorization 17:392
Theory and practices
and developing leaders 13:659–661
Theory building
social work-specific 20:502–507
in sociology 19:49
Theory groups 21:130, 21:132
Theory of Distributed Associative Memory (TODAM) model 15:162–164
Theory of Event Coding (TEC) 7:421–422

Theory of faith development (Fowler) 20:355–356
Theory of Forms, Plato 18:210
Theory of Mind (ToM) 2:201, 4:4, 4:396, 9:449, 12:3, 12:638, 15:539–544, 15:793–794, 18:800, 20:311, 22:530, 22:856–857
  and behavior during infancy and early childhood 6:13–14
  and brain basis 15:543
  children's learning 4:39
  and cognitive function 18:818, 18:820
  development during infancy and early childhood 24:268–273
    cross-cultural comparisons 24:270–271
    cultural variation 24:269–272
    explicit ToM 24:268–272
    false belief 24:268–271
    implicit ToM 24:268–269
    intention understanding 24:269
    joint attention 24:269
    mental state language 24:269
    mental state reasoning in tasks 24:269–270
    preschool age 24:268–269, 24:272
  developmental shift between three and four years of age in and autism 15:541–542
  during early childhood 24:268
  and imitation by infants/young children 11:622–623, 11:625
  infant gaze behaviors 12:30–31
  intentionality 12:342–343, 15:539, 15:542–543
  and intentionality 12:347–348
    definitions 12:347
    foundational feature of human communication and language 12:347–348
    mirror neurons 12:348
  mentalizing 15:539–540, 15:542
  metacognition 24:91
  mindreading 15:539
  and mirror neurons 15:585–586
  multi-faceted concept 15:539–540
  nature of 15:540
  origin of theory of 15:540–541
  and own mental states 15:541
  and pragmatic disorders 18:790
  pretend play and 18:873–874
  in primates 10:268–269
  and religious belief 20:390
  skills, developmental time course of 15:542–543
  social brain 15:540
A Theory of Motivation for some Classroom Experiences 15:896
Theory of planned behavior (TPB) 8:27, 9:20, 10:578, 17:943, 23:830
  attitude–behavior consistency 2:185, 2:187–188
  and health behaviors 10:585
Theory of reasoned action (TORA) 17:943, 23:830
  attitude–behavior consistency 2:184–185, 2:187–188
Theory of the life world (Schütz) 21:147–148
  consequences 21:148
Theory testing, educational effectiveness research 7:224
Theory theory 18:19
  children's learning 4:38
Theory-based interventions 10:621–622
Theory-based research 7:558
Theory-consistent information 12:909
Theory-driven and evidence-based school improvement approach 7:227
Theory-driven approach to scientific discovery 21:292
Theory-inconsistent information 12:909
Theory-ladenness of observation 8:576
Theosophy, Whorf on 25:577–578
Therapeutae (Jewish sect) 3:237
Therapeutic alliance 24:287
  explanatory style 8:598
Therapeutic approaches to social work 10:530
Therapeutic aspects of self-efficacy 21:508
Therapeutic communities (TCs) 20:194, 24:274–280
  American model 24:276–277
  British model 24:274–276
    history 24:274–275
    theoretical foundations 24:276
  defining beliefs 24:274
  evidence 24:278–279
    adolescents 24:279
    children 24:279
    drug dependence 24:278
    enduring mental illness 24:278–279
    first episode psychosis 24:278
    learning disabilities 24:279
    offending behavior 24:278
    personality disorders 24:278
    severe mental illness 24:278–279
  fusion model 24:276–277
  phases in 24:277–278
    assessment phase 24:277
    engagement phase 24:277
    preparation phase 24:277
    recovery phase 24:277–278
    treatment phase 24:277
Therapeutic effectiveness assessment 24:281
  cluster randomized trials 24:282
  comparative effectiveness research 24:281–283
  Internet research and data sharing initiatives 24:283
  learning health care system 24:283–285
  necessity of revising existing legislation and regulation 24:284–285
  patient-led research 24:283
Therapeutic health 10:629
Therapeutic hypnosis 11:466–467
Therapeutic interventions, older people 7:338–339
Therapeutic landscapes 1:409, 10:629–630
Therapeutic misconception (TM) 20:517
Therapeutic obligations
  evidence-based medicine (EBM) movement 18:166
  placebo-controlled trials 18:166, 18:168–169
Therapeutic religions 21:380
Therapeutic talk 15:61–62
Therapeutic Tertium 11:459–460
Therapist
  in combined pharmacotherapy and psychotherapy
    misconceptions of 19:497
    professional/emotional support 19:497
    selection by patient with similar ethnic/gender background 19:496
    utilization of special talents 19:496
  and patient relationship *see* Patient–therapist relationship
  psychotherapy experiences 19:518–519
Therapist–client relationship 19:373–374
  transference 24:546
Therapy processes 4:544
Theravada Buddhism 2:891, 15:725
  women in 25:628–629, 25:632–633
Therigatha 25:629
Thermal infrared data 20:419–420
Thermal infrared sensor (TIRS) 20:423
Thermal ionization mass spectrometry 1:906–907
Thermal prospecting 10:54
Thermidorians 20:642–643
Thermodynamics 3:373, 8:433–434
  energy 7:626–627
Thermodynamics laws
  first law of 10:756
Thermoluminescence 3:566
Thesawalamai 23:60
Theta activity, in electroencephalography 7:354, 7:357
Theta criterion 23:904
Thévenot, Laurent 5:526
Thibaut, J.W., works, *The Social Psychology of Groups* 8:483
Thibaut–Walker program 13:749–750
Thick description 11:108, 11:110, 24:291–293
  as episteniology 24:292–293
  ethnography as 8:180–181
  Geertz on 9:662
  as methodology 24:291–292
  in multisited field studies 16:106
Thiele, Friedrich Alexander 2:22f
Thiers, Adolphe 21:169
Thiessen polygons 14:545
Thingification 5:244, 20:203
Things
  land tenure 13:231–232
  legal reasoning 13:778–779, 13:782–783
Think Tank Initiative 24:295
Think tanks 18:322–324, 24:294–299
  definition 24:294–295
  public bureaucracies 19:532
  social and behavioral science research 22:615, 22:619–620
  transnational networking 24:295–296
  world-wide expansion 24:295–296

Think-aloud protocols 20:881–882, 21:538
for measuring self-regulated learning 21:544
Thinking
about money 15:737
about social issues 22:206–208
reliance on mode of thinking 22:206–207
differences in 4:112
genealogical 9:808
mathematical psychology 14:814
probabilistic 19:16–22
skills for ethnography 8:189
for speaking and writing 25:782–783
styles 7:875
Thinking psychology 16:684
Thinking–Brain 15:548
Thin-section petrography 10:56
Third age, grandparent–grandchild relationships 10:336
Third French Republic 21:371
Third generation research, on media audiences 2:217–218
Third International political parties 18:405–407
Third language acquisition (TLA) 16:57–58, 16:65–67, 21:360
Third Millennium 24:303–304
Third National Health and Nutrition Examination Survey (NHANES III) 7:759
Third parties
in conflict mediation 4:627–628, 4:630
disputes 6:560, 6:562–563
and doctor–patient interaction 6:616
social control 4:825
Third place, public libraries as 14:49
Third Reich 8:231
Third Republic 21:371
Third sector 3:366, 22:576
Third space 14:134
Third wave of democratization 10:188
Third World 3:710–711, 4:218–219, 6:276–277, 8:277–278, 15:492–493, 16:150–151, 23:349, 24:300–304, 25:340
as analytical concept 24:300–301
colonialism 24:301
counterinsurgencies in 5:99
defining 24:300
development 24:300–303
development, states' role 6:277–278
entrepreneurial initiative 6:277
import substitution 6:277
neoliberal critiques of state intervention 6:277–278
planning 6:277
diversity, division, and integration 24:302
empire, colonial 24:301
Fourth World 24:302
and global south 10:186–188
globalization 24:302
and human development 6:307, 6:309
industrial capitalism 18:911
mercantilism 24:301
Millennium Development Goals (MDGs) 24:303
national liberation movements 16:244–245
nonaligned movement (NAM) 24:301–303
origins of term and meanings 24:300
periphery 3:308–309
and political economy 1:938, 1:940
and post-Cold War international system 24:302–303
postcoloniality 18:658
radicalism 25:341
third Millennium and World War on WANT 24:303–304
underdevelopment 24:300–302
United Nations and 24:301–302
urbanization 24:777
world poverty 24:302–303
Third World development 13:539
Third World development programs
for social and behavioral sciences 22:617
Third World factories 4:243–244
Third World feminism 8:909
Third World movements 15:532
Third worldism 10:187, 25:748
Third-best welfare policies 25:498–499
Third-force theories 10:429
Third-generation rights 3:676, 11:369–370, 20:681
Third-level personality 17:801
Third-party endorsement 15:32–33
Third-party funding, in academic research and employment 1:49
Third-party investigation 13:747–748
Third-party observers, abusive supervision 23:705
Third-party punishment game 11:263–264
Third-person singular subject 10:325–326
Thirdspace 18:673
Third-wave behavioral therapy 2:432–434, 4:544–545
Third-wave feminism 8:900, 8:907–908
Third-way politics 10:135, 10:137–138
Thirteenth Amendment, US Constitution 22:41
Thirty Years' War 3:675, 25:378–379, 25:416–417
38 Degrees 22:514
Thoamas, W.I., works, *The Unadjusted Girl* 7:560–561, 8:913
Thode, Henry 20:425
and concept of Renaissance 20:425
works
Franz von Assisi und die Anfänge der Renaissance in Italien 20:425
Thoits' model of emotion 7:534, 7:536
Thomae, Hans 6:378
Thomas, Albert 23:4
temperament development approach 24:191
Thomas, Clarence 13:768
Thomas, Deborah
on Caribbean 1:640
Thomas, D.S., works, *The Child in America* 6:19
Thomas, Keith 20:288
Thomas, M., on migration in Australia 2:241
Thomas, Victor S. 22:31
Thomas, William Isaac 6:19–21, 8:138
attitude 2:175–176
habit 10:476–478
works
The Child in America 6:19
The Polish Peasant in America 2:175
Polish Peasant in Europe and America 6:19–20
Suggestions of Modern Science Concerning Education 6:19–20
Thomas, William L.
works
The Earth as Transformed by Human Action 11:19
Man's Role in Changing the Face of the Earth 11:19, 13:229
Thomas theorem (Merton) 1:76, 6:21
Thomist philosophy 13:858
Thompson, Edward P. 3:742–743, 11:42, 19:677
concept of class 11:42–43
on crowds 5:361–363
historiographical approach of 11:43
on social inequality 22:338, 22:341
on working classes 25:695
works
The Making of the English Working Class 11:42–43, 11:110, 22:309
Thompson, Elizabeth, works, *Colonial Citizens* 16:384
Thompson, Godfrey 3:243
Thompson, J.B.
on online advertising 12:589
Thompson, Paul, on interviewing 12:670
Thompson, S.
on women's participation in sports 9:658
works
The Politics of Misrecognition 11:179
Thomson, Alistair 17:285
Thomson, G.H. 8:695
sampling of bonds model 8:695
Thomson, Thomas 18:102
Thomson, William 18:102
Thomson-Reuters 21:322
Thoreau, Henry David 21:178
Thornberry, Terence 14:61
Thorndike, Edward Lee 3:243, 7:240, 12:303, 19:413–414, 24:305–307
learning theory of 13:705
life events of 24:305–306
research and theory of 24:306–307
connectionism 24:306
controlled experiment 24:305
education field 24:306
environmental quality 24:306
lexicography 24:306–307
mind reading 24:305
neurons 24:306
puzzle box studies 24:305
quantification 24:306
works

Thorndike, Edward Lee (*continued*)
Animal Intelligence: An Experimental Study of the Associative Processes in Animals 24:305
Science and Values 24:306
Thornhill, Randy 19:887
Thornton, Patricia H.
on cultural markets 5:469
on institutional logic 5:537
on microfoundations of institutions 17:393
Thorstad, David, on Russian sexual life 21:740–741
Thought collectives 7:862–863
Thought experiments
by computers 4:517
personal identity 17:743, 17:746
Thought monitoring 4:545
Thought-listing technique 17:943
Thought-Producing Self 4:669
Thoughts 2:131, 15:546, 25:325–326
history of 19:16–17
and language 13:252–253, 13:293–299, 20:903–906
schools of 21:128–129
characteristics 21:128–129
historical relevance 21:129
use of term 21:128
Thrasher, Frederic, works, *The Gang* 7:561
Thrasher, John S., works, *The Island of Cuba* 25:282
Threat appraisal, for cardiac disease coping 3:113
Threat explanations, of ethnocentrism and prejudice 8:170–171
Threat imminence model 18:38–39, 18:39f
Threat-dependent ethnic and national identities 8:167
Threats
and authoritarianism 2:266
collective behavior, during past half-century 4:159
coping 4:887–889, 4:892
identity, nonheterosexual British Asians 16:863–864
intergroup 8:143–144
in interviewing 5:854
and political repression 18:449–450
stereotype 23:450–451
see also Stereotype threats
to validity 3:70–71
Threats to external validity 12:655
Threats to internal validity 12:653–654
ThredUP 4:139–140
Three Jewels of Buddhism 2:887–889
Three Mile Island accident 2:300, 7:806, 11:298, 11:300
360-Degree feedback 17:717–718
Three Worlds typology 25:511
Three-community Study (TCS) 10:648–650
Three-component model (TCM) 17:334
Three-dimensional wisdom model 25:590–591
Three-factor model 21:51
Three-failures theory 3:366
Three-fifths compromise 3:304
Three-level hazard model 16:52–53
Three-mode lattice analysis 5:560
3MSE *see* Modified Mini-Mental State Examination (3MSE)
Three-parameter logistic (3PL) model 19:436
3PL *see* Three-parameter logistic (3PL) model
Three-point experiments, in visual space geometry 25:198–199
3P (people, public, private) Partnership Approach 24:896–897
3R (reduce, reuse, and recycle) policy 7:781
3SLS *see* Three-stage least squares (3SLS) estimator
Three-stage least squares (3SLS) estimator 21:976–978, 21:982
Three-store hypothesis 2:575
Three-stratum theory of intelligence 12:292
Three-theme network 4:770, 4:770f
Three-way multidimensional scaling 16:46
Threshold models
discrete-state models 6:523–524
of social protest 22:513
Threshold people 14:133
Threshold test 23:395
Threshold Test Ban Treaty (1974) 1:977
Threshold universalism 12:400
Thresholds
and child neglect 3:410–412
electoral, proportional representation 19:207
liminality 14:131–133
Thrift, Nigel, on affect 1:238
Thrifty phenotype hypothesis 6:753, 11:269
Thrill and adventure seeking (TAS) 21:608
Thrill hate crimes 10:542
Thriving, among youth 3:446–447, 3:449
Throwaway society 6:551
Thucydides 10:162, 11:35, 15:121, 18:380–384, 25:290
and Greek historiography 11:35–36
model of historiography 11:38
works
The Peleponnesian War 10:9
Thule 1:921–922
in Chukotka 1:924
culture 1:924
in Greenland 1:924–925
Thünen, Johann Heinrich von 23:199
Thurnwald, Richard 10:139
Thurot, François, works, *Tableau des progrès de la science grammaticale* 14:215
Thurstone, Leon Louis 3:243, 8:695, 14:801, 14:812
on attitude 2:176–177
IRT model of forced-choice personality assessment 17:843–845, 17:844f–845f
on law of comparative judgments 14:803–804
on mathematical psychology 14:800–801
on semantic differential 21:563
Thurstone scales 2:179
Thyroid gland
hormones, functions of 16:651
hypothalamic–pituitary thyroid axis 11:485
Thyroid stimulating hormone (TSH) 6:202
TIA *see* Transient ischemic amnesia (TIA); Transient ischemic attack (TIA)
Tianeptine 22:85
Tibet
and China 2:65, 2:67
nationalism in 16:277–278
Tibetan Buddhism 2:889, 15:725
Ticker tape 14:196
TIE *see* Total indirect effect (TIE)
Tie strength, social capital 22:184
Tie-break policies 1:248–249
Tiele, Cornelius Petrus 7:375, 20:322–323
Ties, and social networks 16:564, 22:461–462
TIF *see* Test information function (TIF)
al-Tifashi, Ahmad, work, *Delight of Hearts* 21:750
Tifinaghe 5:814–815
TIGER *see* Topologically Integrated Geographic Encoding and Referencing (TIGER)
TIGER database 10:61
Tight oil and gas 17:167
Tight wire 3:472
TIGP *see* Taiwan International Graduate Program (TIGP)
Tigray People's Liberation Front (TPLF) 3:441
Tikanga 17:431–432
Tillich, Paul 1:539–540
Tilly, Charles 14:491, 14:876, 16:564, 22:435, 25:410
crowds 5:361–362
on war 25:410–411
Timaeus, Sicilian 11:37
Time 16:194, 24:409–414
anthropology of 24:406–408
cultural time 24:407
future and end of time 24:408
gendered time 24:408
language of time 24:407
political time 24:407
ritual time 24:407
social time 24:406
spatial and bodily time 24:407–408
universal time 24:406
aspects of 24:409
belongs to the supreme nominal power 24:410
in combined pharmacotherapy and psychotherapy
availability for patients 19:496
concept, developmental research 6:330
cooperation embedding in 4:863–864
demographic adjustments over 6:134–135
differences across special demographic groups 24:390
children's time use 24:390
seniors 24:390
and emotions 7:538–539
and equilibrium 7:915

between events in point process 6:573
human–environment relationships 11:407
and knowledge representation 13:103
and language 13:294–295, 13:297–298
linear time and cyclical time 24:409
and linked lives 14:231–232
Middle Ages 15:348
of nostalgia 17:35–36
pressures, and leisure 13:850–851
and procrastination 19:76
of relational coping 20:225
relative 17:89
representations of 4:486–487
spent in interviewing 5:853
stability of self over 19:402–403
stream and stack 24:409–410
and temporal distinctiveness 24:203
types of 24:409
work and leisure 13:853
Time (magazine) 22:904
Time allocation 3:52–53
for fertility 9:776
Time and space 9:177–181
Time at destination 15:397–399
Time change, demographic analysis 17:728
Time course of classification, concept learning 4:540
Time crunch scale 24:403
Time cultures
continuous diversity of 24:375–376
proposals for transformation toward greater sustainability 24:376–377
Time depth 11:24
Time diary data, for fertility 9:777–778
Time diary method 24:369
Time discounting 2:447
Time domain, time series analysis in 5:652
Time estimates 13:849–850
versus time diaries 13:849–850, 13:846f
Time geography 15:638, 24:416–417
market geographies 14:547
Time horizon
impact of green policies over 10:373
Time orientation, and procrastination 19:76
Time perception models 24:308–310
behavioral theory of timing 24:308–309
pseudologistic model 24:309
scalar property 24:308
scalar timing 24:308–309
Weber's law 24:308
Time perspective, impact on goals 21:560
Time preferences, and procrastination 19:76
Time pressure 24:371, 24:398, 24:401–404
in dual-career families 14:607
outcomes from time inputs 24:404
QOL implications 24:403
questions 24:403–404
stress 24:403
time crunch scale 24:403
Time series 12:652, 24:341–346
advanced methods 24:311–315
basics 24:342–344
concepts 24:342–344
problems 24:344
cointegration 24:322–330
current theory and research 24:345
cycles 24:331–336
discrete valued time series 24:314
economic forecasting 24:337–340
environmental Kuznets curves 13:147
future directions 24:345
history 24:341–342
importance 24:341
methods 24:344–345
computing 24:345
descriptive 24:344
parameter estimation 24:344
seasonal adjustment 24:344
system identification 24:345
uncertainty estimation 24:344
models with conditional heteroscedasticity 24:312–313
models with long memory 24:313–314
nonlinear time series for the mean 24:311–312
class of bilinear time series models 24:312
class of threshold models 24:311–312
seasonal adjustment 24:347–353
second generation models 24:314
simultaneous equations 21:982–983
state space methods 24:354
stochastic volatility models 24:311
unit roots and nonstandard limiting distributions 24:362–368
Time series analysis 4:772, 6:716, 18:515–521, 24:316–321
advantages and pitfalls in 18:516
cointegration 18:517
duration analysis 18:515, 18:518–519
event counts 18:515
forecasting 18:515–518
fractional integration methods 18:517
in the frequency domain 5:652
functional magnetic resonance imaging (fMRI) 9:528–529
with heterogeneous variances 24:320
intervention analysis and outlier detection 24:319–320
linear dynamic models 24:317
nonlinear models 24:320
panel data 18:519
and political science 18:516–517
cross-sectional analysis 18:519–520
event-history models 18:518–519
volatility, time-varying parameters, and dynamic effects 18:517–518
pseudo-panels 18:519
repeated cross-sectional analysis 18:515, 18:520
stationarity 18:517
stationary *versus* nonstationary processes 24:316–317
stochastic structure of observations 24:316
in the time domain 5:652
vector autoregression 18:516–518
Time series central limit theory 15:295–296
Time series equations model 24:357
Time Series Processor *see* TSP (Time Series Processor)
Time stamps 13:103
Time to return on investment (TROI) 5:56
Time to the most recent common ancestor (TMRCA) 11:292
Time trends 22:933
Time units 14:193
Time use 24:373–378
acceleration 24:373, 24:376–378
clock time 24:373–376
desynchronization 24:373–374
and fatherhood 8:840–842
and gender 24:369
comparative studies 24:370–371
domestic work 24:369–371
exchange theory 24:370–371
individual level 24:370–371
institutional level 24:371
paid work 24:369–371
resources 24:370–371
time pressure 24:371
household economics 11:224–225
in industrialized countries 9:775–780
for fertility 9:777
intensification 24:373, 24:375, 24:377–378
leisure, in adolescence 13:829–830
demographic differences 13:830
portfolios 13:830
prevalent activities 13:829–830
measurement of 24:369
multitasking 24:375
pace of life 24:376
patterns 24:369–370
planning 24:373, 24:376
and resource consumption, connections between 24:373–375
time cultures
continuous diversity of 24:375–376
proposals for transformation toward greater sustainability 24:376–377
Time use expenditure surveys 5:733
Time use research 24:383–397
data collection methods 24:392
diary supplements (IT) 24:392–393
internet and mobile apps 24:393
wearable cameras 24:393
differences across special demographic groups 24:390
children's time use 24:390
seniors 24:390
free time 24:390
organizations and volunteering 24:390
social life 24:390
TV viewing 24:390
general trends and patterns 24:393
cross-national results 24:393–394
general social survey 24:394–395
U.S. trends 24:394
historical perspectives 24:391
mass observation 24:392–393
USDA 'Homemaker studies' 1923–1931 24:392
infrastructural developments 24:383–384
American time use study 24:384
MTUS data archive 24:383–384
paid work 24:387–389

Time use research (*continued*)
personal care 24:389–390
eating 24:390
grooming 24:390
sleep 24:390
subjective time 24:393–394
subjective time use 24:398–405
time changes and new communication technologies (IT) 24:391–392
unpaid work 24:389
childcare 24:389
shopping and services 24:389–390
visualization maps of international activity differences 24:384–387
2003–2010 ATUS diary data from different U.S. immigrant groups 24:387–390
MDS applied to Szalai's (1965) multinational data 24:387
updated 1998–2005 MTUS mappings 24:387
Time use research methods 24:379–382
administrative records 24:380
child care 24:379
diary data collection 24:381
diary methods 24:380–381
domestic division of labor 24:379
exercise 24:380
household work 24:379–381
leisure 24:379–380
methodological alternatives 24:379–380
observational methods 24:380
paid work 24:379–380
studying time allocation 24:379
stylized estimates 24:379–380
time diary research studies 24:382
Time use survey (TU) 5:744, 5:746
Time utilization, leisure, international comparisons 13:847t
Time-based conflict
and work–family conflict 8:759
Time-based resource sharing model of working memory (TBRS) 21:905
Time-bounded relativism 21:311
Time-composite maps 6:716
Time-dependent confounding 8:316–317, 8:316f
Time-dependent covariates 8:296–297, 8:303, 14:55–57
Time–energy envelopes 13:346
Time–geographic framework 15:638
Time-intensive commodity 9:757
Time-limited psychotherapy 19:344
interpersonal psychotherapy 12:631–636
Timeline Followback Method (TLFB) 10:589–590
Timeliness 10:302
TIMEOFTEST 19:727
Time-ordered composition
algebra 22:478, 22:481–483
Time-resolved optical imaging 16:389
The Times 18:916
The Times Literary Supplement 2:510
Time-series forecasting 24:337
models 24:337–338
for forecasting higher moments 24:338
multivariate models 24:338–339
univariate models 24:337–339
uncertainty of 24:338–339
estimation 24:339
model approximation error 24:338–339
model instability 24:339
modeling trends 24:339
overfitting 24:339
shock uncertainty 24:338–339
Time-sharing systems 23:377
Time-space, in geography 24:415–420
absolute *versus* relative views of 24:415
conceptualizations of behavioral landscapes 24:416–418
from spatial to time-spatial perspectives 24:415–416
Time–space compression 12:91, 24:418
in multisited field studies 16:106–107
Time/space convergence 24:607–608, 24:608f
Time–space convergence (TSC) 24:416
and human extensibility 24:417–418
Time-space distanciation 24:418
Time-to-contact information, and perceptual learning 10:129–130
Time-to-divorce 23:784
Time-to-event data, survival analysis 23:771
Time-varying confounders 14:341–342, 14:341f
Time-varying covariates 8:307–309
Time-varying treatments
causal inference for 14:340–344
bias of traditional methods for 14:343–344, 14:343f–344f
dynamic strategies 14:341
sequential exchangeability and identifiability 14:342–343
sequential positivity 14:342–343
strategies 14:340–341
time-varying confounders 14:341–342, 14:341f
TimeWise program 13:835
Timing
community-based health interventions 10:647
conflict mediation 4:627, 4:630
of fertility 6:112–113, 6:112t
of informed consent 3:902
in life course 14:80–83
of linked lives 14:231
of release of information to archive 1:919
Timor-Leste, fertility transition in 9:84–88
TIMSS *see* Trends in International Mathematics and Science Study (TIMSS)
TIMSS-95 *see* Trends in International Mathematics and Science Study (TIMSS-95)
Tinbergen, Jan 7:9
macroeconomics 14:400–403
Tinbergen, N., works, *The Study on Instinct* 14:369
Tinbergen, Niko 14:367, 16:469
Tinnitus 16:718
TIP *see* Teaching innovation programs (TIP)
TIPI *see* Ten-Item Personality Inventory (TIPI)
Tipton, Steven, works, *Getting Saved* 2:509–510
TIRS *see* Thermal Infrared Sensor (TIRS)
Tiryakian, Edward A. 21:130–131
works, *Rethinking Civilizational Analysis* 23:26
TIs *see* Tenant improvements (TIs)
Tissue culture, in new genetics 9:1016
Tissue peptides 16:651–652
Tissue-specific epigenetic effects 24:687
Tit for tat (TFT) 4:863, 5:893–894, 11:261–262, 18:935–936
Titchener, Edward Bradford 12:703, 19:413
Tit-for-two-tats 5:893–894
Titles 13:628–629
Titling, land 13:233–234
Tito, Josip Broz 18:700
Tittie queens 13:875
Tiwanaku in Bolivia 23:36, 23:39
Tiwi 25:645
Ti–yong principle 25:543
Tjalpawangkanyi speech style 13:291
TL *see* Target language (TL); Transnational law (TL)
TLA *see* Third language acquisition (TLA)
Tlatelolco square massacre 22:145
TLE *see* Temporal lobe epilepsy (TLE)
TLFB *see* Timeline Followback Method (TLFB)
TM *see* Therapeutic misconception (TM); Trademark (TM)
TMRCA *see* Time to the most recent common ancestor (TMRCA)
TMSA *see* Transformational model of social activity (TMSA)
TMT *see* Terror management theory (TMT)
TNCs *see* Transnational corporations (TNCs)
TNE *see* Teachers for a New Era (TNE) initiative
TNS *see* Taylor Nelson Sofres (TNS) Emnid (Bielefeld, Germany)
T-O maps 3:174
Tobacco control policies 22:128–129
Tobacco research
censorship 3:280, 3:282
Tobacco smoking 22:126
addiction 1:97–98, 1:100–101
Tobacco treatment programs 15:64–65
Tobacco use 6:671
Tobacco–alcohol amblyopy 1:512
To-be-remembered stimulus 5:925
Tocharian 11:824
Tocqueville, Alexis de 1:579, 1:668, 3:518–519, 3:710, 4:158, 13:802, 13:819, 16:237, 16:834, 18:466, 18:649–650, 20:279, 21:168, 24:421–425
on American/French Revolutions 1:597–600
on analysis of social change 24:422–423
on democracy 24:421–422
North America 17:24

democratic theory 6:91
democratic theory of 6:91
on enlightenment 24:423
on French revolution 24:421, 24:423
on mass society 14:744
originality of contribution of 24:423–424
past and present 24:424
on politics 24:423
on public opinion 24:423–424
unity of process 24:421
works
Ancien Regime and the French Revolution 18:382
De la democratie en Amerique (*Democracy in America*) 1:618, 2:510, 3:710, 4:309, 7:657, 10:762, 13:819, 14:744, 16:237, 24:421–424
Journeys to England and Ireland 16:237
L'Ancien Régime et la Révolution 24:421–423
Oeuvres complètes 24:423
The Old Regime and the Revolution 16:237
Souvenirs (*Recollections*) 16:237, 24:421, 24:423
TOD *see* Transit-oriented developments (TOD)
TODAM *see* Theory of Distributed Associative Memory (TODAM) model
Todd, John, works, *The Mechanism of Exchange: A Handbook of Economics* 7:590–591
Toddler Behavior Assessment Questionnaire 24:193
Toddler rejection 5:509
Toddlerhood, play during 18:221–222
differences in adult support 18:221–222
differences in caregiver–toddler play preferences 18:222
differences in parental scaffolding of object play 18:222
differences in play partners 18:222
differences in play themes 18:222
similarities in effects of adult play scaffolding 18:221
similarities in emergence of symbolic play and peer play 18:221
similarities in gender differences and mutual specificity in caregiver–toddler play 18:221
Toddlers
geometric cue use in reorientation by 23:121–122
and media 15:1–2
nongeometric cue use by 23:122
Todd's paresis 7:854
Todes, Samuel, works, *The Human Body as Material Subject of the World* 2:729
Todorov, Tzvetan 8:237–238
Toffler, Alvin, works, *The Third Wave* 18:666–667
TOH *see* Tower of Hanoi (TOH) task
Tohoku 22:570–571
Toiletgate scandal 13:284
Tok Pisin 18:138
Tokelau 17:158–159, 17:162–163
Token, and money 15:746–749
Token + constraint system 24:22
Token economy 2:409
Token physicalism 18:109
Token reference, anomia 19:147–148
Token test 21:914
Token-identity theory 15:175–176
Tokenism 15:239–245
boundary permeability 15:240, 15:243–244
disadvantaged groups 15:244
discrimination 15:240–244
individual mobility 15:240, 15:242, 15:244
justice 15:239–240, 15:242–243
and organizational diversity 4:739, 6:596
successful tokens 15:243–244
Tokens as precursor of writing 25:761–762
Tokugawa era 14:635–636
Tokutarô, Sakurai 20:369
Tokyo Anthropological Society 2:71
Tokyo Metropolitan Government 24:895
Tokyo Tribunal 25:383
TOL *see* Tower of London (TOL) task
Tolerance 6:593, 14:253–254, 17:231, 18:236–242, 19:464
for ambiguity 20:248
and community 4:387
culture 18:241
effects of alcohol on 1:508
human rights 18:238, 18:241
and liberalism 14:24
passive and active 18:240
personal 18:240–241
political 18:240–241
religious 18:238–240
seizures 13:34–35
social 18:240–241
Tolerance intervals 8:19, 17:292
Toleration 25:349
and morality 13:487
religious *see* Religious toleration
Tollefson, James, works, *Planning Language, Planning Inequality* 13:354
Tolls, and congestion 24:632–633
Tolman, Edward 19:414
Tolstoy, Leo, on nationalism and war 25:367
Toma, D., on gay athletes 9:658
TOMAL-2 *see* Test of Memory and Learning, Second Edition (TOMAL-2)
Tomasello, Michael 8:403
Tomen, S., on gay/lesbian tourism 9:641
Tomo, Imamura 2:71–72
Tompkins, Jane 20:490
Tonal structures in African languages 23:716–717
Tone 18:69
Tone awareness 3:511
Tone burst (TB) 16:708–710
Tone grammars 23:717
Tone languages 23:716
Tone-bearing unit (TBU) 23:716–717
Toneri, Prince
works
Chronicle of Japan 11:50
Nihon shoki 11:49
Tonga 17:158–163
Tongan 17:559–560
Tongqi 21:735
Tongue protrusion, neonatal imitation 11:621
Tongzhi 11:50, 21:737–738
use of term 21:737
Tonkinson, R., on knowledge 13:73
Tönnies, Ferdinand Julius 4:340, 4:361, 11:865
on alienation 1:551, 1:553
Gemeinschft und *Gesellschaft* continuum 2:271
subordination of society to community 4:340–341
works
Gemeinschaft und Gesellschaft 1:553, 1:766, 4:361
Tool kit theory 5:553
Tool use 19:69
in apraxia 1:877–879
multistep actions with multiple tools and objects 1:878–879
single 1:878
Toolmaking, and modern human cognition 11:254–255
Tools & Tillage 17:671
Toomey, R.B., on gay-straight alliances 21:88–89
Tooth traits 2:602
Top management 23:491–492
influence on 23:490–491
Topcoded responses 13:389
Topcoding
probabilities elicitation 7:382
statistical disclosure limitation 4:579
Top-down approaches
demographic microsimulation 15:344
longitudinal data 14:348–349
Top-down bias, visual selective attention 2:165
Top-down embodiment 7:409, 7:413
Top-down influences 23:241
Top-down modulation, of right-ear advantage 6:363
Top-down processing, in occipital lobe 17:131
TOPI *see* Test of personal intelligence (TOPI)
Topic interest 12:378
individual interest 12:378
situational interest 12:378
Topic models 9:364–365
Topic possessive 18:632
Topical interview 5:851–852
Topical macro databases, from academic projects 5:793–794
Topicalization 16:2
Topics 12:96
social history 22:309–310
survey questions 23:765–766
Topobo
tangible user interfaces (TUI) 24:22
Topographic maps 10:61, 24:426–428
in brain 24:426–427

Topographic maps (*continued*)
discovery of sensory and motor topographic maps 24:427
current trends 24:427
development of maps 24:427–428
modules 24:426–428
motor cortex 24:427
reasons for having 24:428
sensory cortex 24:427–428
sensory representations 24:427
see also Computational maps
Topographical amnesia 24:554–555, 24:558–559
Topographical learning, virtual spatial navigation 16:350
Topography 13:224–227
Topolobampo 2:150
Topological language 23:90
Topological relations 14:285–286
Topological space structure 17:299
Topological state space dynamical model 14:797
Topologically Integrated Geographic Encoding and Referencing (TIGER) 1:847
Topology 3:621–623, 13:913
ordered relational structures 17:299–300
Toponymy 17:224–225
Topophilia 23:321–322
Tor Faraj (Jordan) 17:459
TORA *see* Theory of reasoned action (TORA)
Torah 3:546–547, 20:335, 20:398–399
Toren, Christina 7:147
Torgersen, Helge, on epistemic cultures 7:878
Torgerson scaling 16:44
Torii Ryuzo 2:71–72
Tormey, Simon, on anarchism 1:683
Toronto Workers Action Centre 24:493
Toronto's Markham 23:664
Torr, Diane 13:878
Torrens, Robert 20:662, 21:403
Torrens title 21:403
Torres Strait Islanders 2:234, 2:238
Tort, Jaruwan S., works, *Tort, Custom and Karma: Globalization and Legal Consciousness in Thailand* 13:472
Tort law 12:216–218
to workplace safety
evidence on 20:867–868
theoretical stimulus of 20:867
Tort liability 13:744
Torts
and damages 20:413–414
domestic violence 6:628
organizational 17:400–402
Torture 24:442–447
changing tactics and changing laws 24:443–444
definition of 24:434, 24:442–443
emerging issues and practices with respect to 24:445–446
historical and theoretical underpinning regarding 24:443
key cases of 24:444–445
and mass trauma 14:749–750
and political repression 18:446–447, 18:450–451
punishment as 2:525–526
in Soviet Union 20:822–823
torturer
FORNET treatment 24:439–440
motivation 24:438
trained, experiences and consequences of 24:439
victims
experiences and consequences 24:434–437
narrative exposure therapy 24:437–438
post-traumatic stress disorder 18:703
and violence 25:116, 25:120
war 25:377–380
Torture Victims Protection Act 1992 6:564
Tosh, John 9:770
Toshiba 6:834–835, 11:840
Toshikazu, Shinno 20:369
Toshio, Kuroda 20:367
TOSMANA (software) 4:443
TOT *see* Technology transfer (TOT); Treatment on the treated (TOT) estimates
Total capital stock, sustainable development 23:823–824
Total deviation 7:342
Total diglossia 6:418
Total effect of schooling 21:125–126
Total factor productivity (TFP) 7:46–49
and international immigration 12:517
Total fertility rate (TFR) 5:739, 6:113, 9:74, 11:7, 12:46–47, 18:563, 18:572–573
destination-specific (D-TFR) 15:399
difficulties with 6:113–114
in East Asia 9:55–56, 9:55f
in formerly socialist European countries 9:101, 9:100f
and immigration 15:397–399, 15:401–403, 15:401t–402t
in Southeast Asia 9:84–85, 9:87–88
in sub-Saharan Africa 9:92–93, 9:97–98, 9:92f–93f
Total immersion, method of 14:454
Total indirect effect (TIE) 15:76
Total inertia 5:8
Total institutions 6:48, 10:276, 12:190, 23:737
concept of 22:975
Total justice 13:819
Total literacy 11:589–590
Total quality management (TQM) 11:697
Total redress 13:819
Total slack 16:541–542
Total social fact
gift as 10:139–140
work as 25:675
Total social organization of labor (TSOL) 25:658
Total Talent Portfolio 7:162
Total war 15:519–520
TotalHb 16:392–393
Totalistic utopianism 21:214
Totalitarian countries 15:638
Totalitarian impulse 14:258
Totalitarian regimes 1:982, 6:373–374, 22:844–845
Totalitarian religion 18:437–438
Totalitarian state 23:345–346
Totalitarianism 4:306, 5:369, 6:373–374, 6:811, 18:434–435, 21:371–372, 22:743, 22:849, 24:448–452
absolutism 1:10–11
Arendt on 1:963
authoritarianism 24:448–452
Bell on 2:499–501
in Berlin 2:542, 2:545–546
communism 24:448–452
concept
development 24:450–451
usefulness 24:451
Dahrendorf on 5:662–664
defining features of 24:448–450
dictatorship 24:448, 24:450–451
in Eastern Europe 6:818
fascism 24:448–452
Hitler on 24:448–449
Islamofascism 24:451–452
and mass sphere 3:704
and modern dictatorship 6:369–370
Mussolini on 24:448–450
Nazism 24:448–450
parties 18:405, 18:408–409
and self 21:447–448
Shklar on 21:889–890
social and behavioral sciences under 22:140
Stalin on 24:448–451
today and tomorrow 24:451–452
and war 25:366–369
Totality, in historiography 11:100
Total.pop 16:867t
Totemism 3:730–732, 6:702, 9:107, 23:862, 24:453–458
Aboriginal Australians 24:453–457
and advertising 1:216
analogy 24:453, 24:457
asexual reproduction 24:455
clans 24:453–457
conception 2:236
native peoples of North and South America 24:453
objectification 24:453, 24:455–457
'patrilineal' principle 24:456
religion 24:455–457
Touch, tactile processing 24:14–15
Touch DNA 9:956–957
Tough-Movement 16:2
Toughness 2:256
Tour de force 1:81, 15:769
Touraine, Alain 24:459–464
biography 24:459
commentary 24:463–464
contributions
critique of modernity 24:459
sociological theory of action 24:459
in study of social movements 24:459
critiques 24:463–464

sociology 24:460–463
works
Après la crise 24:463
Comment sortir du libéralisme? 24:462–463
Critique de la modernité 24:461–462
La conscience ouvrière 24:460
La fin des sociétés 24:463
La Parole et le Sang 24:461
La société invisible 24:463
La société post-industrielle 24:460
La voix et le regard 24:461
Le monde des femmes 24:461
Le Mouvement de Mai ou le Communisme Utopique 24:461
Le retour de l'acteur 24:461–462
Pour la sociologie 24:460–461
Pourrons-nous vivre ensemble? Egaux et différents 24:460–461
Production de la société 24:460
Qu'est-ce que la démocratie? 24:462–463
Sociologie de l'action (*The Sociology of Action*) 24:460
Un nouveau paradigme: Pour comprendre le monde aujourd'hui 24:462–463
Vie et mort du Chili populaire 24:460–461
Tourism 5:370, 15:636, 20:806, 24:465–469
age of 24:465
anthropology of 24:465
approaches to 24:474–475
and brokers 2:872
changing perspectives in 24:470–471
and changing world 24:472
in China 2:68
corporate social responsibility 24:475
critical turns in 24:471
culinary 9:293–295
and cultural geography 5:410–412
cultural interpretations 24:471
and cultural resource management 5:488, 5:490
cultural tourism 24:466, 24:468
culture 24:467
definition 24:469–470
ecotourism 24:466–468
environmental sustainability 24:475–476, 24:479
essence of 24:466
gay and lesbian 9:637–643
emergence of 9:637–639
risk 9:641–642
and sexual identity 9:639–640
special events and festivals 9:640–641
geography of 24:469–473
global human mobility 24:465–466
host society 24:467
impact on gender equality and women's empowerment 24:477–478
industry 24:465
in Japan 12:797
and liminality 14:135–136
Millennium Development Goals in 24:475–476
MDG3 24:476–479
modernity 24:465, 24:468
and pilgrimage 14:133
political economy of 24:470–471
poverty reduction and 24:474–475, 24:478–479
reflexive 24:468
responsible tourism 24:474
risk society 24:468
sex 19:282
spatial models 24:470
and sustainability 24:467–468
time and space 24:466
tourist 24:466
experience 24:466–467
types of 24:466
see also Ecotourism; Sustainable tourism
Tourism Geographies: An International Journal of Tourism Space, Place and the Environment 24:469
Tourist arts 2:17
Tourist gaze 24:466–467
Tourist realism 5:416
Tourists 24:466
experience 24:466–467
Tournachon, Gaspar Felix (Nadar) 20:421
Tournament of values 8:479
Tournament selection 2:42
Tournament theory of promotions 19:136–137
Tout court 10:190
Tower of Hanoi (TOH) task 8:493
Tower of London (TOL) task 8:493
Town urban 24:859–860
Towns and cities 24:811–812, 24:814
artisans in 2:47
Towns Towards Sustainability 23:848
Townshend, Charles 1:610
Township and village enterprises (TVEs) 6:790
in China 24:974
Township governments 24:855
Toxic disinhibition (online dating) 17:213
Toxic Wastes and Race at Twenty study 7:757–758
Toxicity, alcohol 1:510
Toxicology 15:89
Toyota 1:192
Toyota Gosei 11:840
TPACK *see* Technology, Pedagogy, and Content Knowledge (TPACK)
TPB *see* Theory of planned behavior (TPB)
TPH2 *see* Tryptophan hydroxylase 2 (TPH2)
TPJ *see* Temporal–parietal junction (TPJ)
TPLF *see* Tigray People's Liberation Front (TPLF)
T-Po community (Taiwan) 13:871
TPVs *see* Temporary protection visas (TPVs)
TQM *see* Total quality management (TQM)
Trabant exodus 25:838–839
Trace amine-associated receptors (TAARs) 17:196–197
Trace ethnography 21:190
Trace logs, for measuring self-regulated learning 21:544
Trace measures 22:548
Trace-based accounts 16:5
Trace-deletion hypothesis 14:177–181
early formulation of 14:178
two-dependency problem 14:180
Traces 21:539
unobtrusive measures 24:767–769, 24:771
Tracking 21:122, 22:330
Tracking science 21:239
Tracking tasks 16:845
Tracking/ability grouping, influence on academic achievement 8:133–134
Tracmail 3:47
Tractography 9:492–494
Trade 2:46, 12:480–482
archaeology of 24:481–484
characterization studies 24:481, 24:483
current theory and research 24:482–483
distribution patterns 24:483
electron microscopy 24:481
formalist approach 24:482
future directions 24:483–484
identification 24:482–483
inductively coupled plasma-mass spectrometry 24:481–482
neutron activation analysis 24:481
obsidian 24:481–484
substantivist school 24:482
technological developments 24:481–482
theoretical developments 24:482
x-ray fluorescence spectrometry 24:481
and exchange 8:477–480
foreign direct investment 9:324
putting more health in 10:202–203
in South America 23:43–47
in West Africa
impact of external trade on social values 1:302
social consequences of slave trade 1:302–303
Trade creation 12:574–576
Trade diversion 12:574–576
Trade liberalization 6:271–272, 12:569–571, 16:466–467
GATT for 12:568–569
and health 10:198–201
and informal employment 12:63–64
and political economy 1:938, 1:940
Trade negotiations 12:568–569
bilateral/plurilateral 12:571
by countries 12:571–572
multilateral 12:571
for trade liberalization 12:569–571
Trade policy 12:568, 24:485–490
'behind-the-border' rules 24:485
conduct of 24:485
definition 24:485
exports 24:485–488
foreign economic policy 24:486, 24:488
formation 24:485–489
free trade agreements 24:485
future of 24:489
globalization 24:487, 24:489
Hecksher–Ohlin model 24:486
imports 24:485–487
international negotiations 24:488–489

Trade policy (*continued*)
international political economy 24:487
international trade 24:485–487
protectionism 24:486–488
regionalism 24:485, 24:489
society-centered theories of 24:486–487
state-centered approaches 24:487–488
structural theories of 24:486
hegemonic stability theory 24:486
(neo)Marxian approaches 24:486
system-centered theories 24:486
tariff duties 24:485, 24:487–489
trade theory 24:487–488
two-level and multiple-level games 24:488–489
typology of 12:569t
unfair trade 24:487–488
unilateral trade policy 12:568
win-set 24:488
World Trade Organization 24:485, 24:489
Trade Related Aspects of Intellectual Property Rights (TRIPS) agreement 1:460, 9:1003, 12:264, 13:520–524
Trade secrets 3:31
Trade theory 12:569–570, 24:487–488
Trade unions 11:860–862, 24:497–502, 25:692
allocative inefficiency 24:498
bargaining approach 24:500–501
behavior, formal modeling of 24:498–499
conditions for welfare of workers 24:498
bargaining power 24:498
surplus 24:498
contract curve 24:501–502
density and collective bargaining coverage 24:497t
efficient bargaining model 24:501–502
expected utility function 24:499
firms 24:497–501
isoprofit curve 24:500f
monopoly union model 24:500, 24:500f
Nash bargaining 24:500–502
right-to-manage model 24:501–502
and social work, in Canada 24:491–496
austerity agendas 24:491–494
best practices in mobilization for social justice 24:492
case management 24:491–492
commodification 24:491
community–labor alliances 24:492
historical roots of collaboration and shared values 24:491–492
Los Angeles immigrant labor activism 24:492–493
movement building 24:495
neoliberalism 24:491
public sector employment 24:491, 24:493
renewal and convergence 24:491
social reform 24:492
social workers as professionals and union members 24:493–494
solidarity 24:491–494
translating progressive core values into meaningful practice 24:494–495
work behind Chicago Teachers' Strike 24:494
working class 24:491–492
wages 24:497–502
Trade-led industrialization 4:243–244
Trademark (TM) 3:31, 12:266–268, 12:270–271
Trademark law 12:269
Trade-off heuristics 5:915
mapping model 5:915
1/N heuristic 5:915
tallying 5:915
Trade-offs 4:443–444
in life history evolution 9:29, 9:32–33
life history theory (LHT) 9:40, 9:42–43
in multidimensional matching 2:103–104
rationalizing 2:106
in psychiatric care 19:297
Tradeunion aristocracy 1:965
Trading with the Enemy Act 13:531
Traditional and cultural practices 10:195
Traditional apprenticeships 25:245
Traditional authority 2:270
Traditional constituent order typology 25:646–650
basic constituent order 25:646
categorial confusion 25:649–650
constituent order correlations 25:646–647
Duala on 25:649
employment of semantic criteria in constituent order research 25:648–649
Kinnauri on 25:650
notion basic constituent order 25:648
relevance of meaning 25:650
Traditional educational planning 7:262–263
Traditional families
deinstitutionalizing of 8:780
emotional development of children in 7:439–440
of industrial era 8:779–780
Traditional gender gap 25:304
Traditional healers 10:637
in psychiatric care 6:44–46
witchcraft 25:606–608
Traditional healing, for child soldiers 3:434–435
Traditional history 10:165
Traditional humans, fertility and parental investment among 9:29–30
Traditional intellectuals 10:747
Traditional judge 12:879–880
Traditional knowledge 7:878
fishing 9:228
Traditional medicine 10:568–569
and psychotherapy 19:502
Traditional nationalism scholarship 16:309–310
Traditional organizations *versus* bureaucracy 2:908–909
Traditional school learning 1:872
Traditional social life, in Central Asia 3:320
Traditional societies 14:133, 18:912
marriage in 14:601
Traditional teacher education, response in United States 24:76–77
Traditional tricyclic antidepressants (TCAs) 19:460
Traditionalism 11:94
and conservatism 1:603
Tradition-creation 22:966–967
Tradition-directed network actors 5:558–559
Tradition-directed social character 20:671
Tradition–modernity dichotomy 24:508
Tradition-oriented religiousness 22:163, 22:165–166
Traditions 5:606
Aboriginals 24:509
anthropology of 24:503–507
binding force 24:503–504
and communicative technologies 24:506
and construction of modernity 24:504–505
and creativity 24:505–506
enlightenment 24:504
politics 24:506–507
social organization 24:505
as substance, process, practice 24:503–504
and authority 2:270–272
and community sociology 4:363
conceptualization 22:965–967
and conservatism 4:689–690
cross-national integration of 22:969
culinary 9:292–293
and cultural heritage 5:414–417
in everyday life 8:323–324
folklore 9:281–282
Great and Little 16:377
home 11:155–156
importance of 22:965
and knowledge 13:70–73
and modernization 7:322, 7:324
and nostalgia 17:35–36
of novelty 15:682
and power 18:765, 18:768
predisciplinary national 22:967
and primitive society 18:912
shaping of 22:970–972
Shils on 21:885–887
sociological aspect 24:508–511
symbolic interactionism 12:354–356
transnational contexts 22:969–970
and values 25:27–30
Traditionsbruch concept 4:240
Traffic analysis zones (TAZs) 24:618
Traffic congestion 24:629
Traffickers 10:207–212
Trafficking 5:167, 12:544
of flora and fauna 5:160
of persons 10:210
in South and Southeast Asia 12:534, 12:536
Trafficking Victims Protection Act 22:44
Trafficking Victims Protection Reauthorization Act (TVPRA) 20:462
Tragedy of culture 20:203
Tragedy of the anticommons 19:156–157, 19:168

Tragedy of the commons 6:28, 11:586, 11:588, 11:782, 12:717, 12:725–726, 17:579–581, 18:514, 19:154, 19:157–158, 22:196, 22:245–246
Tragedy of the privates 19:157
Trager, Edith 16:678
Tragic choices 10:604
Trail of Tears 8:84
Trails B, dysexecutive syndromes 6:726
Train the Trainers model 24:83–84, 24:755
Trained torturer
experiences and consequences of 24:439
FORNET treatment 24:439–440
Training 7:449
applied demography 1:842
child and adolescent psychiatry 3:383–384, 3:387
clinical psychology, in United States 24:524–527
counseling psychology 5:71
of drivers in transportation 24:626
and education 25:691–692
in educational research 7:280–281
employment 13:184–185, 13:187
forensic archaeology 9:337
markets for 22:20–21
metacognitive 24:94
organizational diversity 4:741–742, 6:598–599
relapse prevention 20:217–220
self-regulation 21:543–544
social work in Russia 22:809–810
of teachers in United States 24:75–77
Training data 23:262, 23:416
Training initiatives, armed conflict 1:973–974
Training of Professional Personnel Act 1959 23:221
Training of staff, consultation–liaison psychiatry 14:3–5
Trait anger, and aggression 1:377
Trait anxiety 1:808
Trait inference 2:202
Trait negative affect (NA) 11:599
Trait of character 25:149
eudaemonism 25:149
noneudaemonistic virtue ethics 25:149
Trait/disease-discordant monozygotic twin study 24:691
Traité de sociologie 7:588–589
Trait-related differences, contra-hedonic motivation 1:245
Traits 17:857
as cognitive constructs 17:871
cognitive-adaptive theory of 17:875
consistent 17:920
creativity 17:770–773
definition 2:246, 17:920
and functional explanation 9:522–524
indicators of core self-evaluations 4:905–906
and leadership 13:664–665
and marriage 8:830–833
model of personality 17:920
as neuropsychic structures 17:870
personality 17:859
personality-religion links 17:802
as probabilistic constants 17:859–860
and RST 17:860–861
versus states 17:920
surface 17:857–858
temperament 24:196
tendencies 17:920
Trajectories 14:51–52, 14:59, 18:291
of assets 1:331
data on 14:51–52
death/dying 5:867
of engagement 1:34
gay youth 9:651
of individuals 14:52
level of individual 14:53
of liabilities 1:331
life course, health behaviors and 14:71–72, 14:72f
of resilience 1:331
of vulnerability 1:331
women's employment 14:53
Trajectory analysis 19:734
developmental psychology 5:357
TRAMO-SEATS (software) 24:352
Tramp shipping 14:529–531
Tran, Nhung Tuyet, works, *Viet Nam: Borderless Histories* 11:85–86
Trance, spirit possession 23:282
Trance state 16:170
Tranquilizers, with psychopharmacotherapy 19:460
Trans, use of term 13:864
Trans* case law 24:551
Trans* identity politics 24:552–553
Trans* rights, future of 24:551
Transaction, leadership as 13:650
Transaction cost economics (TCE) 11:568, 15:677–678, 20:866–867
and hierarchies operationalization 10:843–844
and organization theory (OT) 19:892, 19:894
organizations 17:308–311
Transaction cost theory 2:485
of rationality 19:911
and unintended consequences 17:421–422
Transaction costs 12:481, 23:369, 24:528–532
institutional economics 17:190, 17:192–193
in markets 14:560
of organizations 17:398
Transactional analysis, early studies 19:504
Transactional leadership 13:666, 25:639
and groups 10:424
Transactional model
of divorce 14:520–521
early defiant behavior 6:15–16
Transactional nature of contextual influence 4:788
Transactional perspectives on coping 4:889–891, 4:890f
as process 4:890–891
research 4:889–890
Transactional sex 21:703
Transactionalism
and brokers 2:870–871
prenatal/infant development 18:855, 18:858
Transaction-based work groups 13:629–630
Transactionism, and modern human cognition 11:252
Transactions, hierarchies and markets 10:842
Transactions costs 19:534
Transactions in GIS 10:60
Transactions of the American Society of Civil Engineers 20:421
Transaction-specific customer satisfaction 5:630–631
Transaction-specific satisfaction 5:630
Transactive discourse 15:775–776
Transactive discussion 15:775–776
Transactive memory 2:484, 10:426, 10:436
evolutionary group dynamics 8:442
and groups 10:436
Transatlantic African slave trade 15:447
Transatlantic Slave Trade Database 22:56
Transborder citizenship 3:639
Transboundary environmental problems 7:813
Transboundary fishery problem *see* Differential games
Transboundary waste trade 7:760–761
Transcategorization 10:332–334
Transcendence 17:621, 21:391, 23:286
Transcendent universal religion of the nation 3:680–681
Transcendental dialectical critical realism 20:24
Transcendental ego 11:442
Transcendental ends 25:49
Transcendental meditators 16:805–806
Transcendental philosophy 13:1–2
Transcendental realism 5:277, 20:23
Transcendental reduction 18:18, 20:345
Transcendental semiotics 4:794
Transcendentalism 2:149
Transcending of class boundaries, spirit possession 23:283
Transcortical motor aphasia 1:820t
Transcortical reflexes 15:968
Transcortical sensory aphasia 1:820t
Transcranial alternating current (tACS) 24:535
Transcranial direct current stimulation (tDCS) 9:476, 9:509–510, 11:604, 22:67–68, 23:20
aftereffects 24:533–534
in aphasic individuals 9:510
application in language research and language rehabilitation 9:509–510
cellular effects 24:534
considerations for future studies 24:537–539
functional and neuroanatomic organization of language processes 24:535–536
immediate "online" effects 24:533–534

Transcranial direct current stimulation (tDCS) (*continued*)
and language 24:533–544
molecular and membrane effects 24:533–534
network effects 24:534–535
neural effects of 24:533–535
reports of effects 24:537
Transcranial magnetic stimulation (TMS) 4:676, 9:476, 9:506–509, 9:530, 11:604, 22:486–487, 23:19
for aphasia 13:332–333
for cognition 4:98
for cortical area studies 5:23–27
in language rehabilitation 9:508–509
in language studies with healthy individuals 9:508
for lexicons 13:925–926, 13:928
repetitive (rTMS) 23:19
visual word recognition 25:216
Transcription 14:185
Transcriptional regulation, role of histones in 7:839–840
Transcriptome 2:450
Transcriptomics 2:625
Transcultural, macrohistorical transformation 11:117
Transcultural Psychiatry 19:311
Transculturation 5:118, 11:449
and education 1:732
and hybridity 11:449–450
Transdiagnostic approaches (CT) 4:546
Transdisciplinarity
universities 18:309–310
Transdisciplinary analysis, of urban studies 24:943
Transdisciplinary Play-Based Assessment 3:795
Transdisciplinary population geographies 18:601–602
bodies making population 18:602
violence 18:601
Transdisciplinary studies, of infant development 18:858
Transducers 16:737
Transect planning 16:812
Transects, human–environment relationships 11:407
Trans-epistemic connections 13:207
Transeuropean culture 10:227
Transsexual individuals 24:549–553
Transfer, research, and development (TRD activities) 24:760
Transfer Control Protocol 12:600
Transfer effects, in language learning 13:301–302
Transfer lines 2:296–297
Transfer of Development Rights programs 10:449
Transfer payment 25:530
Transferability
of mixed methods inferences 15:622
of qualitative research 19:688
Transferable utility 9:579
Transferable utility games 4:867–868
Transference 9:415, 13:68, 19:340–341, 24:524, 24:545–548
analytical techniques 24:546–547
countertransference 24:547–548
in cross-cultural research 7:929–933
adaptation and assembly 7:930
definition 24:545–546
Freud on 24:545–547
multiple language 16:59–60
neurosis 24:546
psychoanalysis 24:545–548
psychology
early studies 19:503
and metaphors 15:291
psychosis 24:546
regressive neurosis 24:546
repressed traumas 24:545
symptomatology 24:545
therapist–patient relationship 24:546
understanding at end of twentieth century 24:546
Transference feelings 19:345
Transference-focused psychotherapy (TFP), BPD 2:759
Transference-focused therapy, for personality disorders 17:896
Transfers 16:66–67, 25:397
in conflict resolution 4:588
instructional psychology 12:202
intergenerational 8:826–828
language contact 13:301–302, 13:304–305
and marriage 8:829–833
school learning for 21:96–100
basic approaches 21:96–97
bounded framing 21:99
common elements 21:97–98
detect-elect-connect model 21:99
expansive framing 21:99
failure 21:96–99
high and low roads 21:97–98
mindful abstraction 21:97–98
missing in action 21:97
new approaches 21:98–99
as preparation for future learning 21:98–99
prospects of teaching 21:99
on rebound 21:97
social context of activity 21:98
surface and deep structure 21:99
of training phenomenon 8:567–568
Transformation distance 21:580–581
Transformation matrices 19:80
Transformation of disputes 6:560
Transformation of sex 25:631
Transformation of social representations 20:447
Transformation problem 19:112
Transformational accounts of similarity 21:580–581
Transformational generative grammar 25:578
Transformational grammars 19:9
Transformational group leadership 10:437
Transformational leadership 13:656, 13:664, 13:666, 13:668–669
and groups 10:424, 10:438
Transformational model of social activity (TMSA) 22:491
Transformational rules 21:623
Transformations 7:78, 19:79–80
cognitive 4:153
of disputes 6:560–566
economic 7:73–79
emotional 4:153
and greening of business 10:392, 10:394
leadership as 13:650–651
mathematical 8:600
in media consumption 25:81–82
positivism 18:625–626
power 8:600
relational 4:153
ritual 20:718–719
structuralism 23:627–630
urban 3:607–608
in work 25:677–678
Transformative change, in industrial ecology 11:849
Transformative consumer research 4:741
Transformative learning 1:15, 1:18
Transformative social workers 11:646
Transformative teacher paradigm 24:53
Transforming planned economy, of China 6:789–790
Transgender individuals 24:549–553
health concerns of 15:116
Transgender law 21:769
Transgender movement 9:734, 24:551–552
Transgender Nation (organization) 24:552
Transgender scholarship 9:734
Transgenderism 25:211
Transgenders 8:902, 21:678, 21:681, 21:787, 21:789
activism in Latin America 13:861–862
in China 21:738–739
and drag culture 13:876
friendships 9:455
and homophobia/heterosexism 10:790–793
intersectional queer theory 19:765
LGBT movements 9:632, 9:635–636
minorities 15:573
and social work 21:802–805
visual turn 25:211
see also Lesbian, transgender (LGBT) individuals
Transgene expression in homogenous cell populations 17:271–272
Transgenerational community 12:399–400
Transgenerational epigenetic inheritance 7:834–835
Transgenic designs 9:934
Transgenic organism 9:916–917
Transgenic studies of learning and memory 9:906
interpretation of 9:907
Transgressions 11:112
African tricksters 24:652–654
implications of shame and guilt for deterring 21:476–477
and taboo 24:2–3

Transgressive art, avant-garde 2:321
Transgroup anthropological principles of development 19:677
Transient amnesia 1:651–658
Transient epileptic amnesia (TEA) 24:554–561, 24:563
  accelerated long-term forgetting 24:557–558
    biology 24:558
    clinical and neuropsychological features 24:557–558, 24:558f
    cognitive architecture 24:558
  autobiographical amnesia 24:554–555, 24:558–559
  biology of 24:557
    microscopic neuropathology 24:557
    underlying cause 24:557
  case histories 24:555–557
  clinical features and demography 24:554–555
  diagnostic criteria and differential diagnosis 24:555–556, 24:556t
  diagnostic criteria of 24:563t
  focal retrograde amnesia 24:559
  prognosis 24:555
  remote memory impairment 24:558–560
    autobiographical memory loss and theoretical models of memory 24:560
    biology 24:559
    clinical features 24:559–560, 24:560f
    cognitive architecture 24:559
  seizures 24:554–556, 24:558
  topographical amnesia 24:554–555, 24:558–559
  treatment 24:555
Transient global amnesia (TGA) 1:655, 24:562–570
  anterograde amnesia 24:562, 24:565–568
  CA1 neurons 24:566–568, 24:568f
  case example 24:562–563
  clinical presentation 24:562–563
  cognitive outcome 24:564
  dementia 24:563, 24:566–567
  diagnosis 24:563
    criteria 24:563t
  diagnostic features of 24:556t
  differential diagnosis 24:563, 24:563t
  epidemiology 24:564
    gender distribution 24:564
    time of onset 24:564
  hippocampus 24:565–568, 24:568f
  memory impairments 24:563
  neuroimaging in 24:567–568, 24:568f
  neuropsychology 24:565–566
  pathophysiological mechanisms 24:566–567
    epilepsy 24:566
    hypoxic-ischemic events 24:566–567
    migraine 24:566
    stress 24:567
  precipitating events 24:564–565, 24:565f
  recurrence 24:564
  short-term memory 24:566
  temporal lobe 24:565–567
Transient ischemic amnesia (TIA)
  diagnostic features of 24:556t
  electroencephalography 7:362
Transient ischemic attack (TIA) 24:563
Transient pain 3:556t
Transient population cycles 18:559–560
Transient receptor potential (TRP) channels 12:711
Transient spatial information, brain systems computing 23:163–165, 23:164f–165f
Transient spatial knowledge, spatial navigation based on 23:161–164
TRANSIMS (traffic flow models) 1:367, 1:370–371
Transindividual 4:338
Trans-institutionalization 6:50, 25:111
Transistorization, and interviewing 12:674–675
Transition countries, shadow economy labor force 21:864–865
Transition dynamics, and greening of business 10:395
Transition economics 7:73
Transition economies (TEs) 13:174, 16:79
  labor market performance in 13:175–176
Transition equations 17:253, 24:354
Transition intensities, models with
  constant and piecewise constant 8:314
  nonparametric 8:314–315
  parametric models for 8:314
Transition probabilities
  in event history analysis 8:310–316
  grammars 19:12
  Markov models for 14:578t
Transition processes 14:244
Transition rate 14:55
  in event history analysis 8:307
Transition states 18:545–546
Transition theories 9:519
Transition to adulthood 24:571–575
  cross-national differences 24:573
  culture-based differences in 24:576–582
    apprenticeship 24:580
    biologically linked markers 24:578–579
    bodily adornment 24:579
    ceremonies 24:577–581
    coming-of-age 24:577–578, 24:580
    cultural markers 24:578–581
    emerging adulthood 24:576, 24:580
    genital mutilation 24:578
    globalization 24:576–577, 24:579–581
    informal instruction 24:580–581
    marriage 24:580
    menarche and pubertal development 24:578–579
    non-Western cultures 24:576–579, 24:581
    quinceañera 24:579
    rites of passage 24:577–578, 24:577t
    rituals and celebrations 24:579–580
    skill building and learning 24:580–581
    youth galas 24:579
  goal adjustment and goal engagement in 4:816–817
  life course trajectories, macro typologies of 24:573–574, 24:574t
  recent findings 24:572–573
    differences across gender 24:572–573
    differences social classes 24:572–573
    overview 24:572
  and vocational interests 25:240–241
  in western developed world 24:572
Transition to democracy and capitalism
  Russian and post-Soviet studies 20:846–847
Transition to takeoff stage of development 21:150–151
Transitional bilingualism 2:581–582
Transitional countries 18:191
Transitional housing 11:161
Transitional justice 11:368, 13:440–441, 17:650, 18:498–499, 18:502
  and law 13:539
  and war crimes tribunals 25:387
Transitional multistate analysis 8:295
Transitional nations 18:187
Transitional probability 13:308–309
Transitional socialist market economies 24:894
Transitions
  city in 3:607–608
  democratic 6:97–100
  Eastern European studies 6:807
  from education to employment 7:130–131
  educational 5:377
  and elder care 7:335–337, 7:339–340
  generalization of 15:861–862
  in life course 14:52, 14:80–83
    effects of 14:52
    globalization impact 14:54
    sequence of 14:52–53
  liminality 14:131–134, 14:136
  mortality, epidemiological, and health 15:857–862
  in postsocialist societies 18:698–699
  school burnout 21:61–62
  school engagement 21:61–62
  from socialism to neoliberalism 25:122
  sustainable 23:820–821
  from war to peace 25:123–124
Transitive clause 17:558
Transitive orientation algorithm 16:543
Transitive predicate 10:326
Transitive verb phrases 16:4
Transitivity 7:88, 17:525, 24:680
  violations of 5:918
Transitology 20:635–636
Transit-oriented cities 24:934, 24:936–938
Transit-oriented developments (TOD) 16:809–810, 24:620, 24:867
Transitory commemoration 4:239
Translanguaging 13:355, 16:66–67
Translatability 5:320, 5:325
Translatio imperii 3:720
Translation equation, for fertility change 9:11–12
Translation sociology, and ANT 5:524
Translational medicine
  collaborative research 7:864–865

Translational medicine (*continued*)
experimentation, and medical tourism 23:443–444
Translational research 5:127
Translational sociology 1:858–859
Translation–back-translation procedure 5:325
Translations 5:320, 14:185
actor-network theory 1:82–83
of institutions 17:392–393
of items 7:931
Latour on 13:415–417
in MENA 16:365–366
and postcolonial geography 18:645
and probabilistic grammars 19:9, 19:12–13
Ricoeur on 20:666–667
Translocal relationships 7:688–689
Translocality, and Internet 12:603
Transmarital sex 10:815
Transmedia 7:251
Transmission belt model 1:110
Transmission gap 2:139
Transmission model 2:139
Transmitter systems, and anxiety 1:807–808
Transnational activist network 11:377
Transnational advocacy, human rights 11:364–365
Transnational agrarian movements 17:664–665, 17:667
Transnational biographies 2:639–640
Transnational business disputes 12:470–471
Transnational capitalist class (TCC) 3:101–102, 21:153
Transnational capitalists 24:848
Transnational citizenship 3:639–640, 3:652–653, 3:658, 24:588
political representation 18:443
Transnational classes 3:749
Transnational climate governance 3:843–844
Transnational communities 6:355
in Caribbean 1:642
Transnational corporations (TNCs) 7:687, 10:240, 12:484, 17:302, 21:153
capitalism 10:218, 10:220–222
and Fordism 10:220
and global capitalism 3:99–103
and global economy 10:262–264, 10:266
and globalization 10:221–223
growing power of 10:223
and neoliberalism 10:221–222
'qualified defense' of 10:222–223
Transnational dimensions, of migration 15:422–423
Transnational entity 8:160
Transnational firms 7:33, 7:35
Transnational flexible economies, in East Asia 6:795–796
Transnational forums 13:595
advocacy networks 13:595
Transnational governance 10:295, 10:308–312
Transnational history 4:409
Transnational households 11:228
Transnational human rights networks 11:364–365
Transnational human trafficking 10:210
Transnational justice, and genetics 9:970–971
Transnational law (TL) 10:254–256, 12:508, 13:445
corporate codes of conduct 10:257
declarations and guidelines to protect workers' rights 10:257
as shadow of state 10:257–258
Transnational legal order, political trials in 18:500–502
criminal trials in international criminal courts 18:501–502
transnational tort human rights litigation in domestic courts 18:502
universal jurisdiction by domestic criminal courts 18:500–501
Transnational media 18:529
Transnational media studies 2:220
Transnational migration, and brokers 2:873
Transnational minorities 8:277
Transnational networks 7:688–689, 15:414
de-Westernization 25:544–546
global governance 10:155–156, 10:158–159
Transnational organizations 10:216, 25:401–402
managing culture in 17:349–350
working with culture 17:349–350
Transnational organized crime 5:173–174, 24:596–601
in China 5:173
Commonwealth Scheme for Mutual Assistance in Criminal Matters of 1990 (Harare Scheme) 24:599
developments since World War II 24:597–598
drugs 24:597, 24:599
environmental harm 24:598
European Convention on Mutual Assistance in Criminal Matters 24:599
European Union (EU) 24:597–600
in history 24:596–597
Interpol 24:596–597, 24:600
in Italy 5:174
law enforcement cooperation 24:596, 24:599–600
legal framework 24:599–600
mafia 24:596, 24:598
networks 24:598–599
organizational framework 24:600
policing 24:598–600
in Russia 5:174
Transnational policy 24:296
Transnational projects 1:760–761
Transnational relations, of Caribbean 1:641–642
Transnational religious identities 24:602–606
construction of 24:602–603
globalization and 24:605
Islamic transnationalism 24:604–605
linkage between nationalism and 24:603
missionaries and 24:602–604
modernity 24:603
public life and 24:603–604
Transnational religious movements 20:411
Transnational ruling class 3:101
Transnational social integration 12:246–247
Transnational social movements 22:445
Transnational social protest 22:509–510
Transnationalism 15:392, 18:601
in China 2:69, 12:528
and communalism 4:263–265
and community studies 4:370–371
East Asian studies 6:785
and education 1:731–732
ethnicity 8:160
and feminism 8:901, 8:903
and globalization 10:235–236
and intersectional queer theory 19:768–769
in Japan 12:795
and migration 15:422–425, 15:428, 15:431
and modernity 15:690
in multisited field studies 16:103–106
and nostalgia 17:37–38
and sports 23:303–304
via migration 24:590–591
via participation 24:591–592
via stakeholdership 24:591–592
and world culture 10:226–232
Transnationalities 19:756–757
cultural preconditions of 24:589
and dance 5:668
economic 24:589
and popular culture 18:529, 18:531
Transnationals 9:690
processes 9:691
Transparency 5:12, 5:14, 5:18–19, 18:395–396, 21:216–217
research funding 20:526–527
Transparency International Index 5:793
Transparency report, elements of 10:297f
Transpersonal psychologists 16:170
Transphobia 21:811, 24:550
Transplantation medicine 15:122
Transport 14:288, 14:290, 20:147–150
facilities 14:289–290
Transport costs 14:288–289, 14:291
Transport planning 20:129–131
Transport policy 2:306
Transportation 2:303–307, 23:158, 23:660–662, 23:664–665, 24:417–419, 24:607, 24:908–912
congestion in
airport delays 24:629, 24:631, 24:633–634
alleviation of 24:632–633
capacity investment 24:632
costs 24:630
future 24:633–634
information technology 24:633
intelligent transportation systems 24:633
investment 24:632
latent demand 24:632–633
modeling 24:630–632

policy 24:633
pricing 24:632–633
reasons 24:629–630
tolls 24:632–633
travel demand 24:629, 24:631–632
costs and gains from trade 24:607
associated with increased flows 24:607–608, 24:608f
and development of space economy 24:609–611
impact of reduced costs 24:610
economic development and 24:610–611
history of 4:273–277
cities 4:274–275
cultural history 4:274–275
cursus publicus 4:274–275
development phases in Europe 4:274–275
economic history 4:274–276
film 4:273–274
finance 4:273, 4:276
historiographical trends, further research, and new fields 4:275–276
media 4:273–276
television 4:273–275
transport networks 4:274
travel 4:276
physical activity 18:94–95, 18:97
psychology 24:623–628
adaptation 24:627–628
aggressive driving 24:623–624, 24:626
carpooling 24:623–624, 24:626
commuting stress 24:627
driver risk taking, attitudes, motivation, and emotion 24:625–626
driver selection, training, and rehabilitation 24:626
future topics 24:627
historical development of field 24:624–625
human factors and driver performance 24:625
older drivers 24:623–624, 24:626–627
overview and scope 24:623–624
public transit 24:624
safety 24:623–628
social trends 24:624, 24:626
telework 24:624–625, 24:627
travel demand and alternative modes of travel 24:626
travel modes 24:623, 24:626–628
young drivers 24:625, 24:627
research issues 24:611–614
information technology and 24:613–614
land use and environmental costs 24:612–613
travel behavior 24:611–612
and sustainability 18:199
sustainable *see* Sustainable transportation
trends in select major global economies 24:610f
urban spatial structure and 24:611
U.S. trends 24:609f
Transportation facilities, development of 11:285
Transportation geography 24:607–615
Transportation options 24:617
Transportation planning 24:616–622
accessibility 24:616–617, 24:619–620
activity system 24:616–617, 24:619, 24:621
analytical tools 24:617–620
basic concepts and frameworks 24:616–617
derived demand 24:617
environmental and social externalities 24:617
evolving focus from mobility to accessibility, livability, and sustainability 24:620
future research, challenges and opportunities for 24:620–622
generalized travel cost 24:617
greenhouse gas emissions 24:616
land use models 24:619–620
livability 24:616, 24:620
mobility 24:616, 24:620
rational planning process 24:617, 24:621
sustainability 24:616, 24:620–621
transportation options 24:617
transportation system 24:616–617, 24:619–621
travel demand models 24:618–622
Transportation supply 24:629
Transportation system 24:616–617, 24:619–621
Transposed deixis 6:53–54
Transracial adoptions 1:137
Transsexualism 24:549
and social work 21:802–804
Transsexuality 21:795
Transsexuals, health concerns of 15:116
Transtheoretical model (TTM) 10:579, 18:93
Transurethral resection prostatectomy (TURP) 10:657–658
Transversal politics 8:907–908
Transversality conditions 17:254
Transversely excited atmospheric (TEA)-laser technology 8:577
Transvestism 24:549
TRANUS model 24:890–891
Tranylcypromine 1:778–779
Traulotes 16:680–681
Trauma 1:526, 1:651, 24:434–440, 24:436f, 24:438f
bioarchaeology 2:601
in child soldiers 3:430–436
and childhood cancer 3:462
and childhood stress 23:570–574
mass 14:747–751
and military anthropology 15:483, 15:486
and natural disasters 16:315–318
post-traumatic stress disorder 18:702–706
in social work 18:707–713
ramifications 18:707
spirit possession 23:283
Trauma risk management (TRiM) 15:499
Trauma tourism 4:185–186
Trauma treatment 16:198, 16:200
Trauma-focused cognitive behavior therapy (TF-CBT) 23:573
Traumatic brain injuries (TBI) 11:390, 13:491, 18:792
neuropsychological rehabilitation for 2:814, 2:816–817
severity, and posttraumatic amnesia 18:638
Traumatic events, writing or talking about 7:500
Traumatic historical event 9:859
Traumatic life experiences 4:725
Traumatic stress 16:198–199, 24:642
organizational management of 15:499
Traumatization, obsessive-compulsive disorder 17:121
Travel 24:416–418
communication 4:276
demand and alternative modes 24:626
and health 15:407–408
liminality 14:131, 14:135–136
tourism 24:465–468
Travel behavior 24:611–612
aggregate analysis 24:611
budget-based models 24:612
disaggregate analysis 24:612
utility-based models 24:612
Travel budgets 24:612
Travel cost, generalized 24:617
Travel cost approach 1:66
Travel demand 24:629
and congestion 24:629, 24:631–632
models 24:618–622
Travel literatures
Humboldt on 25:279, 25:284–285
Sarmiento on 20:909
Travel modes, transportation 24:623, 24:626–628
Traveler-processor model, of hunter-gatherers 11:425
Travelers, Roma people 20:748–749, 20:751, 20:754–755
Traveling spirits, spirit possession 23:283
Traveling theory 18:655
Travis, Cheryl Brown 19:888
Trawling 9:226–227
TRC *see* Truth and Reconciliation Commission (TRC) of South Africa
TRD activities *see* Transfer, research, and development (TRD activities)
Treasury and General Government Appropriations Act 18:279
Treatment adherence, personalized medicine 17:929–930
Treatment confounds 12:450
Treatment contamination (therapeutic effectiveness) 24:282
Treatment drop out
offender rehabilitation 20:199
Treatment effect 20:153
Treatment efficacy 18:42
Treatment fidelity 19:388
Treatment for stability strategy 8:455–457
Treatment manuals 19:388–389

Treatment manuals (*continued*)
psychological treatments 19:383–384
see also Manual-based treatment
Treatment motivation, offender rehabilitation 20:199
Treatment of Depression Collaborative Research Program 14:502
Treatment on the treated (TOT) estimates, in population studies 8:561
Treatment outcomes research, criticisms 19:506
Treatment-as-usual (TAU) 8:349
Treatments 19:320–324
borderline personality disorder (BPD) 2:759
evidence-based psychiatric 8:339–343
groups 22:291
motivational interviewing (MI) 15:912
obsessive-compulsive disorder 17:124–125
and pathological gambling 17:595–596
antidepressants 17:595–596
glutamate 17:596
mood stabilizers 17:596
opioid antagonists 17:595
psychodynamic
brief psychodynamic treatment 22:764
classical Freudian 22:763–764
ego psychology 22:764
Lacanian 22:764
object relations 22:764
relational and intersubjective 22:764
self psychology 22:764
research and access, ethics 8:31–36
autonomy dilemma 8:33
current and emerging directions 8:31–32
definitions 8:31
equity 8:32, 8:34–35
human rights 8:31, 8:33–35
intellectual background 8:31–32
issues and expected changes 8:32–35
justice 8:31, 8:33, 8:35
leading advocates 8:31–32
perspectives 8:34–35
political decision and international participation 8:34
safeguarding against exploitation 8:33–34
schools of thought 8:31–32
social determinants, power relations, and emancipation 8:34
universal health access 8:34–35
vulnerability 8:31–34
of smokers 22:127
Treatment-seeking patients 5:471
Treaty of Amsterdam 8:264, 9:619, 24:599
Treaty of Lisbon 8:265
Treaty of Maastricht 8:267
Treaty of Rome 3:31, 3:875
Treaty of Versailles 1:828–830, 3:515
Treaty of Waitangi 14:505, 14:507, 17:161
Treaty of Westphalia 11:796
Treaty on European Union (TEU) 8:264
Treaty port system 6:782
Tree inference, task network models 16:544
Tree of life 8:370
Tree representations, processing 16:543–544
inclusion/exclusion paradigm 16:543–544
standard tree
for ordered processes 16:544
for unordered processes 16:544
tree inference 16:544
Tree-Adjoining Grammar (TAG) 15:751
Tree-shrews
primates, evolution of 18:904–905, 18:908–909, 18:909f
Treitschke, Heinrich Von 6:690
Trend, fashion 8:835–837
Trend extrapolation models 6:166
Trend models 23:193–194
Trend projection 6:151
Trends and patterns, time use 24:393
cross-national results 24:393–394
general social survey 24:394–395
U.S. trends 24:394
Trends in International Mathematics and Science Study (TIMSS) 4:434, 5:763–764, 5:771, 5:773, 5:847, 7:180, 7:277, 9:679, 10:248, 12:428, 21:232, 25:484
1995 21:126
Trentmann, Frank 4:754
Treskon, Mark, on Chinese art districts 5:599–600
Trespassing 15:592
Trevelyan, G.M. 4:765
Trevisani, T. 3:317–318
Trevor-Roper, Hugh 5:414–415, 7:651
Triad census 22:456–457
Triad social network 22:461–462, 22:462f
Triads 14:429–430, 14:432, 16:546
Triage, and natural disasters 16:316–318
Trial by jury/judge 10:454
Trial court 1:822–824
Trial de novo 1:822
Trial judges 1:823–824
Trial number 17:44
Trial penalty, and sentencing 21:631
Trial-and-error learning 22:401
Trial-and-error process 19:65
Trials 14:266, 16:419, 24:635–638
adjudication 5:103, 5:105–106, 24:635–638
adversary 24:637
civil law 24:636–638
common law 24:636–638
complexity and consequences 24:637–638
dramaturgy 24:636
jury 24:637
khadi justice 24:635–636
legal pluralism 24:636–637
multiple functions of 24:636–637
settlement 24:636, 24:638
Weberian typology of legal decision-making 24:635–636, 24:635t
Triandis, H.C., works, *Handbook of Cross-Cultural Psychology* 5:311
Triangle completion task 23:161–162, 23:162f
Triangle model 6:737
Triangular inequality 14:284–285
Triangular theory of love 14:379
Triangular trade 22:56–57
Triangulation 3:71, 24:639–644
advances in 24:642–643
and ethnography 8:180
in ethnography 8:189
extensions of 24:640–642
in feminist research 9:718
history of 24:639
limitations of 24:642
logic of 24:639–640
procedure 3:198–199
unobtrusive measures 24:769–770
uses of 24:639
Triângulo Rosa 13:861
Triarchic theory of intelligence 12:293
Tribal art, market for 14:567
Tribal confederation 4:571
pre-Columbian Mexico 4:574
Tribal membership, Native Americans 1:635–636
Tribal (primitive) societies 17:669, 23:371
functional analysis 9:537
genocide, history of 10:8
Tribal warfare 17:626
Tribalism 24:646–647
Tribes 3:730, 15:355, 18:912, 24:645–647
in anthropology 24:645–646
characteristics 24:645
chiefdoms 24:646
clan 24:646
and colonialism 24:645
conflicts and war in 4:601–602
current readings 24:646–647
in East Africa 6:773–774
ethnicity 24:645–647
and ethnicity 8:148–150
ethnoarchaeology 24:646
evolutionism 24:646
'fore-runner' image 24:645–646
genealogy 24:646
history of anthropology 24:645–646
horticulture 11:204–205
kinship 24:645
marriage patterns 24:645–646
national elites 24:645
nationalism 24:646–647
neoevolutionism 24:646
social organization 24:646
structural relativity 24:646
tribalism 24:646–647
Tribunals 14:265
of experience 11:522–523
war crimes 25:382
Tribute, in archaic states 23:372
Trichomonas vaginalis, and sexually transmitted infections 6:470
Trichotillomania 17:113, 17:115
Trickery 24:649
Trickle-down theory 4:758–759, 8:836
in consumption 4:758–759
Trickster ethnography 24:648–657
African tricksters, transgression, and social order 24:652–654

Akan 24:648, 24:653–654
archetype 24:648–649, 24:651–652, 24:654–656
Eshu of Yoruba 24:648, 24:652–655
humor 24:653–655
inversion 24:648–649, 24:651–653, 24:655–656
Legba of Fon 24:654
morality 24:650, 24:652–653, 24:656
Native American mythology 24:649–652
reappropriating 24:656
ritual 24:648–649, 24:651–656
scholarly accounts 24:656
sign of history and response to inequality 24:654–656
transgressive. tricksters as 24:649
Tricyclic antidepressants 1:778
Tridentine Catholicism 3:234
Tridentine seminary 3:234
Trier Social Stress Test (TSST) 3:119
Trifocal vision 23:799
Trigger points, terrorism 24:240
Triguna 2:562
Trilateral Patent Office 11:834
TRiM *see* Trauma Risk Management (TRiM)
Trim and fill method 15:268
Trimming (statistical outliers) 23:390
Trinidad 1:639, 1:641
carnival in 3:147
Trinitarian definition of war 1:982–983
Trinitarian eschatology 18:436
Trinity 19:307
Tripartite alignment 10:328
Tripartite Convention of 1899 17:161
Tripartite social schema 25:669–670
Triple Helix urban governance model 22:113–114
Triple negation 16:412
Triple transformation 17:635
Triple-vulnerability theory
of neuroticism 16:744–745
TRIPS *see* Trade Related Aspects of Intellectual Property Rights (TRIPS) agreement
Triune brain 7:468
Trivers' theory 8:400, 15:812
Trivers–Willard hypothesis 8:465
Trivial-game axiom 4:872
Trivium 6:62, 6:66
TRIZ (the theory of inventive problem solving) 20:496–497
Trobriand Islanders 8:483
Trobriand Islands, Malinowski's studies 1:767
Trobriand Man 14:454
Trochlear nerve 8:628–629
Troeltsch, Ernst 20:407, 21:167–168
works
Protestantism and Progress 2:269
Social Teachings of the Christian Churches 2:269
TROFLEI *see* Technology-Rich Outcomes-Focused Learning Environment Inventory (TROFLEI)
TROI *see* Time to return on investment (TROI)
Trojan War 1:445–446
Troll system, and B34S 7:2
Trolleyology 15:814
Trompenaars, Fons, works, *Riding the Waves of Culture* 12:358
Tronto, Joan 20:584
Tropfsteinmaschine (*Dripstone Machine*) 2:21
Trophotropic regulation 23:294–295
Tropism 9:879–880
Trotsky, Leon 18:473
Trotskyism, and Marxism–Leninism 14:659–660
Trotskyists 22:841
Trouble cases 13:469, 13:590–591
Troubled Families Programme 1:792
Trouillot, Michel-Rolph 1:640
TRP *see* Transient receptor potential (TRP) channels
Truce of God 17:639
True belief 24:673
True Buddha School 16:806–807
True effects 8:617
True scores 3:786–787, 17:835, 19:419, 20:245–246
classical true score model 19:423–424
True self 17:737
True zero scaling 3:244
Truffaut, François, works, *Les quatre cent coups* 9:169
Truman, David, works, *The Governmental Process* 18:233
TRUMAN died 11:514
Truncation, family reconstitution 6:141
Truncus 12:416
Trussell, T.J.
'M' and 'm' indices 11:446
Trust games
experimental economics 8:549
and social capital 22:176–177
Trust networks 20:853
pre-existing relationship of 15:414
relations of 15:414
Trust-as-reliance 24:669
The Trustees of Dartmouth College v. Woodward 4:939
Trusts 14:430, 16:553, 16:772, 21:102, 21:257
across life span 12:638–639
adolescence (13–19 Years) 12:638
early adulthood (20–29 Years) 12:638–639
early childhood (2–6 Years) 12:638
infancy (0–2 Years) 12:638
later adulthood (55–75 Years) 12:639
middle adulthood (30–55 Years) 12:638–639
middle childhood (7–12 Years) 12:638
old age (75D Years) 12:639
bases of 12:637
emotional trust 12:637
honesty 12:637
reliability 12:637
childhood 3:454–455
children at risk for psychosocial functioning 12:639
communicating systems based on 14:526
and community 4:385–386, 4:388
concept 24:666
conceptualization 24:665–666
in conflict mediation 4:629
cooperation 4:862–865
development of 12:637
and economic development 5:530–533
and friendship 9:425–426, 9:428, 9:430
in health care 19:816–817
and human genomics in Asia 11:319–321
and intergenerational property transmission 19:174–175
journals 12:868
in justice 2:489–490
measurement and observation 24:666
motivational assumptions about 24:665
in negotiation 16:425–426
and peer review 17:681–682
philosophical aspects 24:668
accountability 24:668–671
accounts 24:669–670
cooperation 24:668–670
decision theory 24:669
dependency 24:668–669, 24:671–672
discretionary powers 24:668
friendship 24:668–670
goodwill-based accounts 24:670
justified trusting 24:670–672
reasons 24:670–671
reliance 24:668–670, 24:672
responsiveness 24:671
risk 24:668–669, 24:671
role and value 24:668–669
social capital 24:668–669
virtue 24:668–672
vulnerability 24:670
political 24:658–663
psychosocial consequences of childhood trust 12:639
public policy ethics 19:584
quantification 19:702–703
rational choice theory 19:904–905
research on 24:664–666
and risk 20:688
in romantic partners 12:638–639
in the sanctuary 20:257
in science 21:212
social capital 22:183–184
social exchange 8:482, 8:486–487
and social media 16:785
and social support 1:548–549
in society 14:430–431
sociology of 24:664
theoretical framework 12:637
and trustworthiness 2:607
use of term 24:664
Trustworthiness 24:664, 24:669–670
in negotiation 16:425–426
research on 24:664–666
Truth
scientific knowledge 24:673–678
correspondence theory of 5:272

Truth (*continued*)
in historiography 11:99
and meaning 14:844–845
media ethics 15:42–44
politics of 15:536–537
and postmodernism 18:685–686
pragmatism 18:803–806
theories 18:809–810
relativism about 20:234–235, 20:237
in science 24:673–678
Truth, Sojourner 8:906, 12:659
Truth and Reconciliation Commission (TRC) of South Africa 2:74, 17:251, 17:657–658, 25:387
Truth and reconciliation processes 25:124–125
in South Africa 4:358
Truth commissions, political trials and 18:499
Truth conditions of sentences 14:293, 15:753–755, 23:901–902
Truth table 19:62
Truth values 14:296
Truthfulness of images 25:171
Truthlikeness 5:275
critical rationalism 5:272, 5:275
Truth-values of play 18:216
Trypanosoma brucei (*rhodesiense and gambiense*) 18:177
Trypanosoma cruzi 18:177
Tryptophan 16:495
Tryptophan hydroxylase 2 (TPH2) 16:495
Tsakhur, gender assignment in 9:796, 9:796t
TSC *see* Time–space convergence (TSC)
Tschumi, Bernard 5:930
TSER *see* Targeted Socio-economic Research Programme (TSER)
Tse-tung, Mao 12:221
TSF *see* Twelve-step facilitation (TSF)
TSH *see* Thyroid stimulating hormone (TSH)
TSOL *see* Total social organization of labor (TSOL)
TSP (Time Series Processor) 7:2, 7:5
TSS *see* Self-reported teaching style (TSS)
TSST *see* Trier Social Stress Test (TSST)
TST *see* Twenty Statements Test (TST)
T2-statistic 11:217
Tsunamis 6:480
disaster in Sri Lanka (2004) 14:748
social resilience in post-tsunami Japan 22:570–575
Tsutomu, Kajiyama 25:463
Tsutsui, Kiyoteru, on world culture 5:539
Tsvetaeva, Marina 21:742
works, Podruga 21:742
t-test 2:376, 6:578, 7:116, 11:218, 11:495–496
t-test 8:8–11, 16:886
in ethnography 8:190
theoretical results 8:9
violations of common distribution 8:10
violations of independence 8:10–11
violations of normality 8:9–10
TTM *see* Transtheoretical model (TTM)
TU *see* Time use survey (TU)
Tuamotus 17:158–159, 17:161
Tuareg 9:784
Tube sites 17:218
Tuberculosis 15:407–408
and mortality 15:860
TubeZik 24:234
TUCE *see* Test of Understanding College Economics (TUCE)
Tucker, Josiah, on division of labor 6:602–603
Tugendhat, Ernst 5:302
Tugwell, Rexford 22:698–699
TUI *see* Tangible user interfaces (TUI)
Tuke, Hack 17:28
Tukey, J.W., works, *Exploratory Data Analysis* 8:600
Tukey estimators 16:82, 20:734–735
Tulip Revolution (2005) 3:319
Tuller, David, on Russian sexual life 21:740–741
Tullock, Gordon, works, *The Calculus of Consent* 23:459
Tully, James 4:705
Tuned filters 25:185
Tuning curves 17:48–49, 17:52, 17:48f
Tunis 15:135
Tunisia 15:135, 16:800–801, 21:794
colonizers in 15:137–138
independence 15:135–136
legal system in 15:366
Tunisian revolution 16:801
Tunten 13:878
Tuomela, Raimo, on joint action 12:842–843
Turberville, George, on Russian sexual life 21:740
Turf wars 15:94
Turgot, Anne Robert Jacques, on division of labor 6:602–603
Turing, Alan 19:415–416
dualism of 15:548
brain–computer analogy 15:548
Turing machine 2:27–28, 4:87–88
algorithms 1:530
Turing test 4:516
Turkestan, nationalism in 16:277–278
Turkey
gender in 9:782
laicization in 13:218
legal system in 15:366
and Middle East and North Africa 16:361–362, 16:365–366
same-sex sexualities in 20:870–874
current situation 20:871–873
history 20:870–871
issues, visibilities, and absences 20:873
and World War I 9:214
Turkish, language contact 13:304–305
Turkish folk music 16:149
Turkle, Sherry 11:549–550
Turkmenistan 3:317
Turley, Jonathan 6:554
Turn constructional units (TCUs) 4:851
Turn to ethics 3:58–59
Turner, Bryan S.
works
Marx and the End of Orientalism 16:382
The Social System 25:357
Turner, Edith, works, *Image and Pilgrimage in Christian Culture* 20:311
Turner, Frederick Jackson 9:464
Turner, Jackson, on democracy in United States 17:24
Turner, Jane, works, *Dictionary of Art* 2:23
Turner, Jonathan H. 8:464
works, *On the Origin of Societies by Natural Selection* 20:386
Turner, Ralph 23:432
Turner, Victor Witter 5:499–500, 14:131–133
on intellectuals 13:72
on Ndembu rituals 1:769
symbolism 23:860–861
works
Image and Pilgrimage in Christian Culture 20:311
Turner thesis and critics 9:464–466
empty lands 9:465
revisionism in British Commonwealth 9:465–466
revisionism in United States 9:465
Turner's syndrome 16:343
Turnout, voting 25:309
Turnover 17:334–335
job satisfaction 12:835
Turns, construction in conversation analysis 4:850–854
Turn-suppression functions 10:110
Turow, J., on television 25:80–81, 25:85
TURP *see* Transurethral resection prostatectomy (TURP)
Tuskegee syphilis study 13:561–562, 20:514
Tutelary deities 11:55
of city-states 11:55
of tribal groups 11:55
Tutorials
computer assisted instruction 4:494
Tutoring, and coconstructivism 3:915
Tuvalu 17:158–159, 17:162–163
TVA *see* Tennessee Valley Authority (TVA) model
Tversky, Amos 8:393–394, 17:782, 24:679–682
on axiomatic theories of choice 24:679–680
on conjoint measurement 24:680
on decision making 24:679–681
on heuristics and biases 24:681–682
on judgment under uncertainty 24:681–682
Kahneman and 24:679–682
on prospect theory 19:262–263, 24:680–681
similarity judgments 24:679, 24:682
Tversky–Kahneman (T–K) research tradition, and bounded rationality 2:774–775
TVEs *see* Township and village enterprises (TVEs)

TVPRA *see* Trafficking Victims Protection Reauthorization Act (TVPRA)
TVSD *see* Technical and vocational skills development (TVSD)
TWA *see* Two-way active avoidance (TWA) paradigm
Tween-albumin medium 6:694
Tweeting, oversharing through 17:439, 17:443
Twelve Tribes 16:806
Twelve-step facilitation (TSF) 1:503
Twentieth century
  class in 3:746–747
    occupational classes 3:747
    polar classes 3:746
    stratification 3:747
  educational psychology, experimentation in 8:568
  and scholars 21:35–36
  social ecology in 22:267–268
    modifications 22:267
    sustainability studies 22:267–268
    symbolic interactionism 22:267
Twentieth Century Fox 9:156–158
Twentieth-century communism 14:656–662
Twenty Statements Test (TST) 23:857
Twenty-first century, LGBTQ families in 21:755–756
Twenty-Four Puzzle 2:32
23andMe (genetic testing company) 9:980, 9:988
Twilight of ethnicity 2:85
Twin designs, for genetic studies of behavior 9:935–936
Twin Earth 20:68–69, 20:71
Twin method 6:320–321
  requirements and assumptions of 6:321–322
Twin population processes 8:413–414
  evolution as demography 8:413
  evolution of demography 8:413–414
  shared concepts and focus 8:413
Twin research, genetic factors 9:921
Twin studies 2:451–452, 11:335, 24:690–692
  and biosocial criminology 2:670, 2:672–673
  classical twin study 24:690–691
  combining different omic data 24:692
  in demography 24:703–709
    ACE model estimates and variance decomposition 24:704–705
    birth weight 24:703–704, 24:707
    endowments 24:703–704, 24:706–709
    fraternal twins 24:703
    genetics 24:703–707
    heritability 24:704–705
    identical twins 24:703
    intrafamilial allocations 24:707
    investments in children 24:705–707
    MZ fixed-effects estimates of impacts of demographic variables 24:705–708
    quantity–quality fertility model 24:708–709
    schooling 24:704–708
  environmental influences on diseases 24:698
  epigenome-wide association scans for complex traits and diseases 24:693–698, 24:694t–697t
  genetics 9:985
  longitudinal twin study 24:692
  multivariate twin study 24:692
  nature/nurture 16:340, 16:342–344
  of neuroticism 16:744
  power to detect epigenetic changes in disease and complex trait 24:692–693
  of sexual orientation 21:779–780
  trait/disease-discordant monozygotic twin study 24:691
Twin Towers 24:238–239
Twins 9:952, 9:954, 9:1010–1011, 24:685–686, 24:686f
  Luria on 14:397
Twitter 24:231, 3:162, 4:288, 4:504, 12:600, 12:602, 24:147, 25:840
  social protest 22:515–516
  terrorists on 24:230–231
  use in politics 14:717
Two bodies 2:720
Two cilia hypothesis 2:123–124
Two process, prolonging, task network models 16:541–542
2–6 years, friendship during 9:449–450
  adjustment 9:450
  conflict 9:450
  play 9:449–450
  prevalence 9:450
  research gaps 9:450
  understanding 9:449
Two Treatises on Government 20:637–638
2AFC *see* Two-alternative forced-choice (2AFC) task
Two-alternative forced-choice (2AFC) task 21:949
Two-and-a-half party system 17:568
2D:4D finger ratio, and sexual orientation 21:776
Two-factor learning theory 3:766
Two-factor theory of intelligence 12:291
Two-layered feedforward neural network 16:117
Two-level games 24:485–486, 24:488–489
2m factorial designs and fractions 17:259
Two-mode network 22:455
  centrality in 16:538–539
2N-ary choice tree model 5:904–905
  random utility model 5:904–905
Two-parameter logistic (2PL) model 19:436
Two-party mediation 13:747–748
Two-party system 17:567–568
Two-period two-treatment design 5:349–351
Two-person constant sum games 16:583
Two-person psychology 24:546
Two-player games 2:33
2PL *see* Two-parameter logistic (2PL) model
Two-process model of sleep regulation 22:83–85
Two-sample location problem 16:892–893
  multivariate 16:895–896
Two-sample tests 16:885–888
  asymptotic distribution 16:886–887
  asymptotic power 16:887–888
  statistic 16:886
Two-sex demographic models 18:590
  alternative two-sex solutions 18:589–590
    birth matrix-mating rule model 18:589–590
    combinations of rates 18:589
    iterative proportional fitting 18:589
    marriage within panmictic circles 18:589
    simple means 18:589
  evaluation of alternative two-sex solutions 18:590
  harmonic mean solution 18:588–589
  manifestations of two-sex behavior 18:590–591
    birth squeeze 18:591
    marriage squeezes 18:590–591
  two-sex problem 18:588
Two-sex fertility 18:590
Two-sex fertility stable population 18:590
Two-sex life table 18:590
Two-sex marriage 18:590
Two-sex population 18:588
Two-sex problem 8:727, 18:588
Two-sex total fertility rate (TFRf) 18:591
Two-sided alternative hypothesis 11:495–496
Two-sided markets, regulation 20:167–168
Two-sided testing, for significance 21:960–961
2SLS *see* Two-stage least squares (2SLS) estimator
Two-stage least squares (2SLS) estimator 4:448–449, 12:206, 20:157, 21:976–978, 21:982, 23:616
  robust standard errors for 4:449
Two-stage model 10:444
Two-stage theory 8:605
Two-state model, for survival data 8:312
Two-step flow model of media influence 2:216
Two-step flow of communication model 14:712–713, 4:291–293
  criticism 4:292
    empirical-methodological criticism 4:292
    theoretical-ideological criticism 4:292
  Lazarsfeld's 4:291, 13:638
  paradigm's fight for survival 4:292–293
Two-step method 21:415–416
Two-stimulus discrimination tasks 21:950
Two-way active avoidance (TWA) paradigm 18:821
Two-way analysis of variance 1:660–661
Two-way education *see* Dual language education
Two-way process of communication 12:368
Tylor, Sir Edward Burnett 2:71, 7:144–145, 7:622, 8:289, 9:672
  contributions to anthropology 1:746, 1:748
  definition of culture 1:766
  on magic 14:435–436

Tylor, Sir Edward Burnett (*continued*)
works
Primitive Culture 1:746, 8:588, 12:178, 20:307
Tyndall, John 9:571
Type 1 diabetes (TDM1) 6:337–340
Type 2 diabetes (TDM2) 6:337–341
Type A behavior, and health 17:793
Type A personality 5:622
Type D personality, and health 17:793
Type I errors 11:491–492, 12:652, 16:82–83, 20:441–442
cardiovascular reactivity 3:119
genetic studies of behavior 9:937
Type II errors 11:491–492, 12:652
Type physicalism 18:109
Type reference, anomia 19:147
Type theory 15:752
Type VIII adenylyl cyclase 20:215
TypeBrewer 3:172
Typhus exanthematicus 18:176
Typical case 3:202–203
Typifications 18:13
Typographic man 18:918
Typographical fixity 18:916–917
Typological approach
to extraversion 8:623
to socialization of children 22:867–868
Typological concept of schools 21:128
Typological functionalism 9:472–473
Typological imperatives, and first language acquisition 9:204
Typological reconstruction (linguistics) 12:448–449
Typological theories 3:211
Typological universals 14:533–534
Typologies
in classical archaeology 3:762
diversity 6:596–597
Humboldt on 25:288–289
industrialization 11:872–878
and language change 14:211–212
migration history 15:446
in social measurement 21:4–5
of social work in criminal justice 5:210–211
Typotechnological approach, to lithics 14:261
Tyranny 6:375, 24:710–715
BBC Prison Study 24:713–714
of data 23:94–95
deindividuation 24:712
group cohesion 24:711
group failure 24:714
and intergroup processes 24:711–713
and intragroup processes 24:710–711
Janis's groupthink research 24:710–711
minimal group studies 24:713
permeability 24:713
rethinking groups 24:714
role theory 24:712–713
self-categorization theory 24:713
of the shoulds 5:64
social identity theory 24:713
Zimbardo's Stanford Prison Experiment 24:711–713
Tzedakah 12:935
Tzeltal linguistic community 13:295–296
Tzevi, Shabbatei 15:531
Tzotzil 10:110
Tzu, Chuang 24:745
Tzu Chi Foundation 2:893, 16:806–807

U

U People (documentary) 13:871
UA *see* Unemployment assistance (UA)
UAE *see* United Arab Emirates (UAE)
UAR *see* United Arab Republic (UAR)
UAs *see* Urban agglomerations (UAs); Urbanized areas (UAs)
UAVs *see* Unmanned aerial vehicles (UAVs)
Uber 8:487
Überindividualität (supraindividuality) 21:963
Ubicomp 21:189
Ubiquitous computing 12:598t
tangible user interfaces (TUI) 24:24
Ubiquitous eco-city 6:834–835
Ubiquitous technology 6:831–832
Ubiquity 10:177
cannibalism 3:84
of startle 23:331
Ubuntu (in Africa) 5:312–313, 10:568, 11:786
UC *see* University of California (UC) system
UCBs *see* Urban containment boundaries (UCBs)
UCC *see* Uniform Commercial Code (UCC)
UCDP *see* Uppsala Conflict Data Program (UCDP)
UCINET software 4:915, 22:475
UCL *see* University College of London (UCL)
UCLA *see* University of California Los Angeles (UCLA) study
UCLG *see* United Cities and Local Governments (UCLG)
UCR *see* Uniform Crime Reporting (UCR)
UCTs *see* Unconditional cash transfers (UCTs)
U-curve theory of adjustment 5:335, 5:335t
UDF *see* United Democratic Front (UDF) (South Africa)
UDHR *see* Universal Declaration of Human Rights (UDHR)
UDI *see* Universal Design for Instruction (UDI)
Uditori di Rota 15:578
Udvalg 19:463
UG *see* Ultimatum Game (UG); Universal Grammar (UG)
Uganda 6:773–776
child soldiers in 3:430, 3:432–435
education in 7:202
Uganda People's Defense Army (UPDA) 20:342
Ugaritic cuneiform consonantal script 25:778t
UGBs *see* Urban growth boundaries (UGBs)
UGC *see* User-generated content (UGC)
UHI *see* Urban heat island (UHI) effect
UHR *see* Ultra high risk (UHR) status
UI *see* Unemployment insurance (UI)
UIA *see* Union Internationale des Architectes (UIA)
UK *see* United Kingdom (UK)
Ukraine crisis 20:849–850
Ulama 11:62
Ulirat 11:791
Ultimate attainment, multiple language 16:60–61
Ultimate justification
Albert on 1:496
Ultimate symbols 2:509
Ultimatum game (UG) 2:447, 9:588–589, 11:259, 12:720–722, 15:814–815, 16:644–645
Ultra Bolshevism 18:720
Ultra high risk (UHR) status 19:200
Ultra-Darwinians 22:904
Ultradian rhythms, regulation by hypothalamus 11:488–489
Ultraliberalism 24:848–849
Ultrametric inequality 23:394
Ultramodern shadow
cultural realm and 5:248
linguistic realm and 5:247
material realm and 5:247–248
symbolic realm and 5:247
Ultramontanism 3:234–235
Ultranationalism 18:411, 18:415–416
Ultrasociality 11:259
Ultrasonography 15:611
for brain asymmetries 2:121
Ultraviolet Index (UVI) 23:697
Ultraviolet radiation (UVR) 23:696
and skin cancer 23:696–699
UMAP *see* Units to Aid Production (UMAP)
Umar (caliph) 12:732
Umayyads (caliphal dynasty), historiography 11:63–64
UN *see* United Nations (UN)
Unaccented moras 23:717
Unaccusative predicate 10:327
UNAH *see* Universidad Autónoma de Honduras (UNAH)
UNAM *see* Universidad Nacional Autónoma de México (UNAM)
UNAN *see* Universidad Nacional de Nicaragua (UNAN)
Unanimity 12:909
Unanticipated consequences 22:196
Unanticipated-perverse effect perspective (UPE) 24:744–746
UNASUR *see* Union of South American Nations (UNASUR)
Unauthorized population 11:635–638
Unauthorized practice of law 13:619
Unbiased estimator 8:17
Unborn girls during pregnancy 15:608–609
Unbound serialities 16:265
Unbounded observations 5:908
Unbounded rationality 2:437–438
Unbounded selfishness 2:438

Unbounded willpower 2:438
UNCAC *see* United Nations Convention against Corruption (UNCAC)
UNCED *see* United Nations Conference on Environment and Development (UNCED)
UNCEF *see* United Nations Children's Fund (UNCEF)
Uncertainties 5:879, 7:883–884, 8:617, 8:619, 13:504, 15:678–680, 24:716–721
  accounting, in cost–benefit analysis 5:53
  and ambiguity 24:717
  artistic/cultural markets 14:565
  contingency of 23:609–610
  in decision and choice 5:879–880
  decision under 14:863–864
  decision-making 5:911–912, 24:718–720
  demand, and markets 5:469
  dimensions 24:716–717
  economic thought 1:576–577
  in economics 22:148, 22:151–152, 24:718–719
  in Empirical Implications of Theoretical Models 18:456–458
  and entrepreneurship 7:675–677
  and environmental cognition 7:706–711
  epistemology 24:716–717, 24:720
  graphics 10:369–370
  and heuristics 10:830, 10:834
  in human behavior 23:467
  and ignorance 24:718
  judgment under 24:681–682
  knowledge 24:716–718, 24:720
  measurement, by statistics 23:424
  and modernity 15:690–691
  and nostalgia 17:34–35
  organizational decision making 17:353–356
  probability 19:33
  in psychology 24:719
  reasoning with 16:926–927
  reduction 8:619–621
  risk 20:706–708
  and risk 24:717
  Simon on 21:970–971, 21:973–974
  and social exchange 8:485–486
  in social sciences 20:686–691, 24:718–720
    audit 20:686, 20:688
    confidence 20:688
    decision making 20:686, 20:689
    Duluth programs 20:689–690
    evidence-based practice 20:690
    managing abuse and violence 20:688–690
    sociological approaches 20:686
    subjectivity 20:686–687, 20:690
    trust 20:688
  in sociology 24:719–720
    organizational studies 24:719–720
    risk culture and society 24:720
  strategizing 23:523–526
  in supply chain management 23:711
  and water scarcity 25:445–446
Uncertainty avoidance 17:824–825, 21:627–628
  in communication 4:285
  definition 25:30
Uncertainty estimation 24:344
Uncertainty management theory 23:702
Uncertainty of outcome (UOO), sports economics 23:319–320
Uncertainty principle 16:584–585
Uncertainty reduction strategies (URSs) 17:214–215
Uncertainty reduction theory 24:719
UNCITRAL *see* United Nations Commission on International Trade Law (UNCITRAL)
Unclassified misdemeanors 15:592
Unconciliated trauma 17:621
Unconditional cash transfers (UCTs) 4:557–558
Unconditional ignorability, causal inference 3:256–257
Unconditional inference model 15:273
Unconditional positive regard, Rogers on 20:737
Unconditioned reflexes 20:211
Unconditioned response (UR) 3:764–766, 8:849
Unconditioned stimulus (US) 3:765, 8:849, 8:870, 17:608–609, 20:207–208
  input circuitry
    eyeblink conditioning 8:637–638
    pathways, cerebellar feedback to 8:638–639
Unconscious 19:330–334
  beginning of scientific attitudes towards 24:722–723
  existence
    proofs 24:723
    psychoanalytic evidence for 24:723
  history of concept 24:722–727
  James on 24:723
  and metaphors 15:290–291
  metaphysical beginnings 24:722
  and personality psychology 17:910
  and psychodynamic thought 22:761, 22:763–764
  racial 4:152
  therapeutic success as evidence of 24:723
Unconscious death 5:866
Unconscious emotionality, of infants 16:483–484
Unconscious fantasy 13:67–68
Unconscious inferences 10:756–757
Unconscious knowledge 13:85–87
  exhaustiveness and exclusiveness 13:86
  judgment knowledge 13:87–88
  structural knowledge 13:87
Unconscious thought theory 5:921
Unconscious transference 8:710
Unconstrained model 23:153
Uncorrelatedness of error variables 3:786
Uncovering knowledge state 14:584, 14:586–589
UNCRC *see* United Nations Convention on the Rights of the Child (UNCRC)
UNCRPD *see* United Nations Convention on the Rights of Persons with Disabilities (UNCRPD)
UNCTAD *see* United Nations Conference on Trade and Development (UNCTAD)
Undeclared work, shadow work 21:862, 21:866
Underclass 3:755, 4:525, 24:728–732
  concentrated poverty 24:730–731
  concepts
    critiques of 24:729–730
    historical origins 24:728–729
  current status of 24:731
  measurement strategies 24:730–731
  and poverty 18:720–721, 18:755
  racial disparities 24:730–731
  social and economic mobility 24:731
  and social stratification 22:709
Undercount, census 3:306
Undercurrents, in everyday life 8:325
Underdeveloped countries 21:151
Underdevelopment
  ideas of 18:911
  and industrial societies 11:866–867
  and modernity 15:693, 15:695
  1950s–1970s 1:458–459
  and Third World 24:300–302
Underdosing, noncompliance as 8:529–530
Underdown, David
  on English Revolution 7:653
Undereducation 7:130, 15:598
Underemployment 25:830
Underemployment equilibrium 13:15
Underfit, asymmetric effects on performance of 23:613
Under-five mortality 12:21–22, 12:25–26
Undergraduate, psychiatric education 7:185
Undergraduate participant pools, experimental laboratories 8:553
Underground economics 12:50–55
  effects on official economy 12:54
  policy consequences of 12:54–55
  relative to GNP
    in various developing and transition countries 12:53t
    in various OECD countries 12:53t
    in West European countries and the United States 12:54t
    see also Informal economics
Underground reputation, of artists 5:468–469
Underground storage organs (USOs) 17:458
Underinvestment, in skills training 22:21
Undernutrition, influence in adulthood 6:753
Undersharing 17:439, 17:441
Underspecified language 8:358
Understanding 10:779–780, 11:132, 22:564, 22:566
  Dilthey on 6:424–429
  economic education 7:21–22
  explanation 8:593–597
Undertheorized questions 19:54
Undirected network 22:461–462

Undocumented immigration 11:638, 15:412–415, 15:463–465, 15:469
UNDP *see* United Nations Development Programme (UNDP)
UNDRIP *see* United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP)
Undue influence, informed consent 12:112
UNECE *see* United Nations Economic Commission for Europe (UNECE)
Unemployment 3:18, 6:270, 7:83, 11:873, 15:595–596, 15:599–600, 20:837–839, 25:672, 25:676–677, 25:830–831, 25:831f
  benefit replacement rates 13:186f
  and crime 5:123–124
  and education 7:130
  and equality 7:908
  and homelessness 11:160–161
  and identity 10:517
  and individual 10:516
  insider–outsider theory of 24:741
  labor market policy 13:183–189
  labor statistics 5:798–802
  and leisure 13:853–854
  and macroeconomics 14:411
  and mental health 24:733–737
    active society 24:736
    classic psychological research on 24:733–734
    critical psychology 24:735
    effects 24:735
    future directions of theory and research 24:735–736
    governmentality 24:735–736
    ideology 24:734–735
    individualism 24:733, 24:736
    labor markets 24:733, 24:735–736
    neoliberalism 24:733, 24:735–736
    subjectivity 24:735–736
  no adaptation to 10:516–517
  panel surveys for 17:465–466
  in postsocialist societies 18:698–699
  and poverty 18:731
  Roma people 20:750
  scale of 24:734
  scaring effects of 10:517
  shadow economy and 21:863–865
  and social deprivation 22:286
  and social norm 10:517
  as social problem 12:544
  and spatial mismatch 23:157, 23:159
  and stress 17:134–135
  structural 24:738
  Trial Census of, in United States 20:877
  use of term 24:734–736
  work anthropology 25:662, 25:666
  and work sharing 24:741
  and youth culture 25:810–811
Unemployment assistance (UA) 13:184
Unemployment insurance (UI) 12:819, 12:823, 12:826, 12:829, 13:184, 19:701–702
  benefits 12:828–829
Unemployment trends, fertility 9:23
UNEP *see* United Nations Environment Programme (UNEP)
Unequal probability sampling 13:386
Unequal protection, in environmental justice 7:758–759
Unequal treatment of categories of people 23:150
Unergative predicate 10:326
UNESCO *see* United Nations Educational Scientific and Cultural Organization (UNESCO)
UNESCO, OECD, and Eurostat (UOE) survey instrument 5:771
Uneven development 10:46, 10:48, 12:914, 12:916–917
Unexpected outcomes 24:744–748
Unexplained gap 6:540
Unfair Contracts Act 4:810
Unfairness 1:201
  of nature 23:149
UNFCCC *see* United Nations Framework Convention on Climate Change (UNFCCC)
Unfit to stand trial
  definition of 9:347–348
Unfolding 16:46
  external 16:46
  metric 16:46
Unfolding analysis, multidimensional scaling (MDS) 16:37–38
UNFPA *see* United Nations Population Fund (UNFPA)
Unfreezing, in organizational development 6:296–299
UNGEGN *see* United Nations Group of Experts on Geographical Names (UNGEGN)
Unger, Roberto Mangabeira 13:437
UN-HABITAT *see* United Nations Centre for Human Settlements (UN-HABITAT)
UNHCR *see* United Nations High Commissioner for Refugees (UNHCR)
UNI *see* Union Network International (UNI) Global Union
UNIA *see* United Negro Improvement Association (UNIA)
UNICEF *see* United Nations International Children's Emergency Fund (UNICEF)
  early childhood care and education (ECCE) 6:743–744
Unidimensional approach, future orientation 9:562
Unidimensional continuum 19:419
Unidimensional scaling 19:432
Unidimensional stage of development 16:504
Unidimensionality
  concept 6:438
  essential unidimensionality, Stout's evaluation of 6:437–438
Unidirectional drive upward, social comparison 22:210–213
Unidisciplinary competence 3:70
UNIDROIT Principles of International Commercial Contracts 13:916
UNIFEM *see* United Nations Development Fund for Women (UNIFEM)
Unification 5:936, 16:299–300
  of analogs, in EITMs 18:456
  reasoning 5:936
Unification approach to scientific explanations 21:302
  compatibility of 21:303–306
Unification approach of explanation 8:597
Unification Church 16:807
Unification of Private Law 13:915–916
Unification thesis 21:302
Unificationism, view of Dark Triad 5:672–673
Unified growth theory 7:42
Unified logic of inference 4:444
Unified science 25:756
Uniform Commercial Code (UCC) 3:30, 3:669, 14:276–277, 21:404
  negotiable instruments 3:30
  sale of goods 3:30
  secured lending transactions 3:30
Uniform Computer Information Transactions Act 4:812
Uniform Crime Reporting (UCR) 5:127, 5:155, 5:165, 5:712, 10:541, 19:733, 23:142
  classification of criminal offense 5:712–713
  criminal vicitmization 5:720–721, 5:725
  scoring of criminal offense 5:712–713
Uniform Crime Reports (UCR) 5:717–718
Uniform distribution 6:576
Uniform Law Commissioners 4:812
Uniform prior 14:122–123
Uniformitarianism 8:163, 20:287
Uniformity
  of language development 13:316
  quantification 19:701–702
Uniform-price auction 2:205
Unigenerational minority 6:474
Unilateral trade policy 12:568
Unilever 23:798
Unilineal descent systems 13:50
Unimodel, of persuasion 17:945
Unimpaired sentence production 14:176
Unincorporated dead, danger of 5:872
Unincorporated places 24:854–855
Uninsured 22:938–939
Unintended consequences 17:420–422
  agency theory 2:895–896
  conflict 17:420
  decision making 17:420, 17:422
  explanation sketches in sociology and economics 17:421–422
  history of 24:744–749
    accidents 24:745, 24:748
    alternative and emerging directions in theory and research 24:746–747
    current refinements and changes in focus over time 24:745–746
    externalities 24:747
    human action 24:744–745

intellectual backgrounds 24:744–745
invisible hand 24:744–747
modernity 24:747–748
perverse effects 24:744, 24:746–747
purposive social action 24:744–747
purposive social intervention 24:747–748
risk 24:747–748
spontaneous order 24:744–745
unexpected outcomes 24:744–748
mistakes 17:422
in organizations 17:421
rationality 17:420–421
social networks 17:420
in sociology, history 17:420
transaction cost 17:421–422
unintended effects, typology of 17:421
use of term 24:744
Unintended effects 24:121–122
and individual-collective relationship 11:817–818
Unintended pregnancy
in Latin America and Caribbean 9:71
in United States 1:1, 1:3–5
Unintentional response errors 16:920
Uninvolved parents 7:439
Union democracy 3:204–205
Union disruption 6:116–119
level 6:116–119
synthetic or hypothetical cohort measures 6:118–119
timing 6:119
Union dissolution 20:432
in Western societies 17:550, 17:550f
Union formation 21:705–706, 21:709
in Western societies 17:545, 17:547–548, 17:550–551
Union International pour la Protection des Obtentions Végétales (UPOV) 17:587
Union Internationale des Architectes (UIA) 24:884
Union members, social workers as 24:493–494
Union monopoly model 24:740
Union Network International (UNI) Global Union 12:860–861
Union of South Africa 19:849
Union of South American Nations (UNASUR) 23:48
Union of Soviet Socialist Republics (USSR) *see* Russia; Soviet Union
Union-bargaining models 24:740–741
Union–intersection test, for hypothesis 11:492
Unionism in industrial countries 13:199–200
Unions, labor 7:571, 7:574
Unions, teachers (in US) 24:77
Unions, trade *see* Trade unions
Unipersonal presidency 18:865
Unique classifier 3:815–816
Uniqueness 11:126
additive representations 14:863
and personal identity 17:746–748
UNISDR *see* United Nations Office for Disaster Risk Reduction (UNISDR)
Unison 14:666
Unit effects, in crossover designs 5:352–353
Unit nonresponse 15:602
Unit roots, time series 24:362–368
AR(p) models 24:364–365
asymptotic inference 24:363, 24:367
challenges 24:367
first-order model 24:362–363
Fisher's information number 24:363–364
mean-squared prediction error 24:366
model selection 24:365–366
models 24:362–365
negative moment bounds 24:365
nonstationary time series 24:362–363, 24:367
UNITAID 8:34
Unitary self 14:834
Unitary system 8:875
Unitary *versus* nonunitary concept 8:489
Unitary view of memory 25:712
United Arab Emirates (UAE), legal system in 15:366
United Arab Republic (UAR) 4:575
United Artists 9:156–157
United Church of Christ 3:542, 21:174
United Cities and Local Governments (UCLG) 23:516
United Democratic Front (UDF) (South Africa) 21:745
United Kingdom (UK)
abolition of slavery in 22:442–443
adoption in 1:142–146
Advertising Standards Authority 1:204
anthropology at home in 1:752–753
arts education in 1:231
associativity in sports 23:309
authority theories 2:270
behavior therapy, development 3:874–875
Biobank 2:621–622, 9:1002, 10:39
censuses in 3:305
Climate Change Act 7:781
clinical psychology 3:874–875
development 3:874
staffing levels 3:875
training 3:874
cognitive therapy 3:874
criminology in 5:241–242
cultural capital research in 5:378–379
Data Archive 1:916
economic development 11:868
economics in 5:539–540
Freedom of Information (FOI) legislation 4:766–767
group psychotherapy 19:505
higher education (HE) market 10:856
immigration policy 11:665
industrial policy in 5:537
industrialization in 11:868
law firms in 13:578–579, 13:581–582
mainstream studies of prison/prison life in 18:925–926
maritime networks 14:526–528
decline of 14:530–532
Muslims in 12:237–239
National Health Service (NHS) 10:603, 10:605–606
National Health Service and Community Care Act 1990 22:802
psychiatric social work in 19:301–306
bio-psycho-social model 19:302–305
burden of mental disorder 19:304
in child and adolescent mental health 19:303–304
development of independent social work departments 19:302–303
psychiatric-child protection interface 19:304–305
socio-historical and policy perspectives 19:301–302
synthesis 19:305
public intellectuals in 22:699
and the Republic of Ireland 11:861
scientist-practitioner model 3:874
self-directed care and individual budgets for older people 21:498–503
small states 22:109–111
and social work 14:663–667
sociology 22:992
Stem Cell Bank and Stem Cell Network 23:442
telegraph in 24:153–154
Union of the Physically Impaired Against Segregation (UPIAS) 6:455
venture capital 25:60
and World War I 9:212
United Kingdom Royal College of Psychiatrists 18:35
United Nations (UN) 3:675, 5:417, 5:714, 8:60, 10:144, 10:186, 10:257, 11:157, 11:796, 13:818, 16:857, 17:641, 23:213
Baruch Plan 17:641
best practices 24:755–756
capacity building, personal and social development 24:755
global economics 24:756
human rights and social justice 24:755
human rights-based social work and gender mainstreaming 24:755
participative social action research and programming 24:755–756
psychosocial perspective 24:755
self-determination and cultural relativism 24:755
social compact 24:756
Train the Trainers model 24:755
Committee Against Torture 24:445
Conference on Environment and Development in Rio 7:734
Conference on Human Rights 11:757
Conference on the Human Environment in Stockholm 7:733–734
conventions 13:720–721
Decade of Education for Sustainable Development (DESD) 7:734–735
Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power 1:836

United Nations (UN) (*continued*)
Declaration of Human Rights 11:757
Declaration on the Rights of Indigenous Peoples 11:782
Economic and Social Council 11:797
Economic Commission for Latin America (ECLA) 22:143–144
Economic Commission for Latin America and the Caribbean 10:187
and education 7:267–268
and environment and health 7:815–816
feminism 8:906–907
Food and Agriculture Organization (FAO) 8:26
Interdepartmental Working Group on Climate Change 1:294–295
Global Agenda Observatory 24:750
Global Compact 24:475
Guidelines on Alternative Care for Children 20:535
and human rights 11:361, 11:370, 11:372
and indigeneity 11:755–758, 11:760
on indigenous people 11:795
Interagency Campaign to End Violence Against Women 12:680
International Conference on Population and Development (ICPD) 8:794–796, 9:71, 20:457
international social work organizations at 24:750–751
leadership of 4:428
mental health 24:751, 24:753–755
Millennium Declaration (MD) 11:357
millennium development goals 24:751
model of national representation 22:993
and nation-state citizenship 3:658
NGO Conference on Discrimination against Indigenous Populations 11:757
nongovernmental organizations 24:750–754, 24:756
Non-Communicable Disease Review 24:752
peacekeeping 17:633–634
historical development 17:634–635
post-2015 24:752
Responsibility to Protect 20:625
and rise of Third World 24:301–302
role in social problems 12:547
social agenda 10:146–148, 24:750
social and behavioral science research agencies 22:619
specialized social work agencies 24:750–757
Department of Economic and Social Affairs 24:753–754
Department of Public Information 24:752–753
Economic and Social Council 24:752
Human Rights Council 24:753
Office for the Coordination of Humanitarian Affairs 24:753–754
UN Women 24:754
United Nations Children's Fund 24:754
United Nations Development Program 24:753
United Nations Educational, Scientific and Cultural Organization 24:755
World Health Organization 24:754–755
Standard Minimum Rules for the Administration of Juvenile Justice (1985) 13:720
statistical classifications 5:707
Sustainable Development Goals (SDGs) 24:751–752
Trust Territory of the Pacific Islands 17:162
Working Group on Girls (WGGs) 24:754–755
Working Group on Indigenous Populations 11:782
world conferences and summits 24:751
United Nations Centre for Human Settlements (UN-HABITAT) 5:779
monitoring system 5:779
United Nations Charter 13:530, 13:533–534
United Nations Children's Fund (UNICEF) 5:370–371, 12:961, 24:754
United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD) 11:802
United Nations Commission on International Trade Law (UNCITRAL) 13:915–916, 21:404
United Nations Conference on Environment and Development (UNCED) 6:845, 6:847
United Nations Conference on Sustainable Development 7:772
United Nations Conference on Trade and Development (UNCTAD) 5:457, 5:749, 6:309
United Nations Convention against Corruption (UNCAC) 5:12, 18:395
United Nations Convention on Contracts for the International Sale of Goods 13:821–822
United Nations Convention on the Law of the Sea 13:821–822
United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) 6:465, 13:720, 20:471
United Nations Convention on the Rights of the Child (UNCRC) 1:137, 1:146, 3:417, 3:478, 3:480–481, 3:485, 3:491, 6:481, 7:146, 8:796–797, 12:956, 12:958–959, 13:720, 20:535, 20:582, 22:890
motoring translation of 12:958
United Nations Declaration on the Right to Development 6:309, 6:311
United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) 3:686, 11:800–801
in international and regional human rights tribunals 11:801
substantive provisions of 11:800–801
right to culture 11:801
right to free informed prior consent 11:801
right to health, education and work 11:801
right to property 11:801
right to self-determination 11:800–801
United Nations Development Fund for Women (UNIFEM) 24:754
United Nations Development Programme (UNDP) 1:296, 6:309, 6:311–312, 11:370, 18:346, 21:150
United Nations Division for Sustainable Development (DSD) 5:777
United Nations Economic Commission for Europe (UNECE)
Convention on Long-Range Transboundary Air Pollution (CLRTAP) 1:724
United Nations Economic Commission for Latin America (ECLA) 6:309
United Nations Educational, Scientific and Cultural Organization (UNESCO) 2:91, 4:124, 4:427–429, 4:431, 5:414, 5:416–418, 5:457, 5:493, 5:769–770, 9:682, 10:163–164, 11:361, 11:802, 14:709, 15:55–56, 16:798, 18:346, 18:468, 20:514, 22:622, 22:674–675, 22:993, 23:213, 24:170, 24:755, 25:84
Convention Concerning the Protection of the World Cultural and Natural Heritage 5:417
Convention for the Safeguarding of Intangible Cultural Heritage 5:417
Declaration on the Human Genome and Human Rights 2:624
framework for cultural statistics 5:729–730
Institute for Education (UIE) 4:429
International Council for Philosophy and Humanistic Studies 22:677
International Textbook Research Network 24:259
Management of Science and Technology Program (MOST) 20:513
Masterpieces of the Oral and Intangible Heritage of Humanity 3:146, 3:148
Universal Declaration on Bioethics and Human Rights 20:512
United Nations Educational Scientific and Cultural Organization (UNESCO)
early childhood care and education (ECCE) 6:740–741
United Nations Environment Program (UNEP) 3:157–158, 5:777, 7:811, 10:374, 18:104–105, 23:796
1976 *Barcelona Convention for protecting the Mediterranean* 7:781
Barcelona Convention for protecting the Mediterranean 7:781
United Nations Framework Convention on Climate Change (UNFCCC) 3:840, 3:861
Intergovernmental Panel on Climate Change (IPCC) 3:854

United Nations Fund for Population Activities *see* United Nations Population Fund
United Nations Group of Experts on Geographical Names (UNGEGN) 17:225
United Nations Guidelines on the Prevention of Delinquency (the 'Riyadh Guidelines') 12:957
United Nations High Commissioner for Refugees (UNHCR) 2:108, 20:105, 20:111–115, 22:769–770
  and armed conflict 1:970, 1:972
United Nations Human Rights Council 13:595
United Nations International Children's Emergency Fund (UNICEF) 23:213
  street children 23:533–534
United Nations Millennium Declaration 1:153
United Nations Millennium Goals 10:264
United Nations Office for Disaster Risk Reduction (UNISDR) 6:481
United Nations Office on Drugs and Crime (UNODC) 12:462, 24:597
United Nations policy arenas 20:459–460
United Nations Population Fund (UNFPA) 8:794, 20:459, 21:794
United Nations Rules for the Protection of Juveniles Deprived of their Liberty (the 'Havana Rules') 12:957
United Nations Statistical Division (UNSD) survey 3:292, 3:294
United Nations Survey of Crime Trends and Operations of Criminal Justice Systems 5:165
United Nations Universal Declaration on Human Rights (1948) 12:543
United Nations War Crimes Commission 25:382–383
United Nations Women 24:476–477, 24:754
United Nations World Tourism Organization (UNWTO) 24:465, 24:474–477, 24:479
United Negro Improvement Association (UNIA) 6:691
United Paramount Network (UPN) 24:171
United States (US) 3:540, 17:588
  2007 recession and fertility in 15:403–405, 15:404f
  academic motivation
    increased focus on male students 9:677–678
    variations within gender groups 9:678
  adoption in 1:142–146
  advocacy planning 1:218
  affirmative action in 9:619
  anthropology of education 1:731–732
  archaeology politics 1:885–886, 1:888–890
  area studies in 1:951–954
    criticisms and newer trends 1:953–954
    diversification and liberalization 1:952–953
    early Russian/Soviet studies 1:951–952
    intellectual trends 1:953
    postwar, institutional takeoff 1:951
    World War II 1:951
  arms control 1:976–978
  arts participation in 5:448t
  assisted dying in 7:600
  authority theories 2:271
  Bill of Rights in 1789 20:279–280
  bureaucratization 2:922
  busyness in 3:50, 3:52–53
  census designated places 24:856
  censuses in 3:304–305
  changes for 1980 24:857
  cities, food deserts in 9:285–289
    consumption 9:286
    food accessibility 9:285–288
    nutrition 9:285–287
    problems and controversies 9:287
    solutions and alternatives 9:287–288
    theories of formation 9:285–286
    urban life 9:286–287
  civil religion, concept of *see* American civil religion
  Civil Rights Movement 3:691–694
  Civil War 16:305
  Clean Air Act 7:773
  clinical psychology training in 24:524–527
  cohabitation in 4:118–121
    child well-being 4:120
    children 4:120
    differentials 4:118–119
    family instability 4:120
    family structure 4:120
    households 4:119–120
    marital stability 4:120
    measurement 4:119–120
    partnership 4:118–120
    research issues 4:119–120
    rise 4:119
    stepfamilies 4:118
    trends 4:118
  Cold War 4:128–134, 9:540
  community studies in 7:561–562
  community supervision and diversion in 4:372–377
    complete diversion 4:374
    crisis intervention teams 4:374
    drug court 4:374–376
    DUI diversion 4:373, 4:375
    electronic monitoring 4:375–376
    enhanced probation diversion 4:375–376
    intensive supervision programs 4:373, 4:375–376
    jail diversion 4:373–375
    justice leverage diversion 4:374–375
    legal pressure 4:373, 4:376
    prevalence of diversion programs 4:374
    probation diversion 4:375
  Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) 12:213
  comprehensive sexuality education frameworks
    future of 21:830
    Future of Sexuality Education (FoSE) consortium 21:830
    National Guidelines Task Force framework 21:830
  confederation 4:573–574
  congressional elections 18:394, 18:396
  conservatism in 4:690
  Constitution 4:706
  continental defense system 24:764–765
  corporate culture 17:348
  corporate farming in 20:803–804
  counterinsurgencies in 5:99–101
  crime policy in 5:176–181
  criminal justice system 10:455
  criminology in 5:238–239
  Cuban emigration 15:464
  cultural capital research in 5:378
  death penalty in 5:876
  demography of abortion in 1:1–5
  developments (1990–2010) 24:857–858
  discovery in 6:516–518
    affirmative discovery 6:517–518
    costs 6:518
    disclosures 6:517
    electronic discovery 6:518
    sanctions 6:518
    scope and limitations 6:517
  discrimination in 6:530–535
    affirmative action 6:532–534
    business necessity 6:531–532
    comparable worth analysis 6:534
    desegregation 6:530–531, 6:533–534
    discriminatory intent 6:530–531, 6:533
    disparate impact 6:531–534
    employment 6:531–532
    Equal Employment Opportunity Commission 6:530–532
    federal antidiscrimination law 6:530–533
    Fourteenth Amendment 6:530–533
    harassment 6:531–532
    hostile environment 6:531–532
    implicit bias 6:534
    indirect discrimination 6:534
    international developments 6:534
    mixed motive approach 6:531–532
    other areas 6:532–533
    pattern and practice 6:531–533
    pretext approach 6:531–532
    role of social science in antidiscrimination legal practice 6:533–534
    social framework 6:531–533
    strict scrutiny 6:530–531, 6:533
  disposable youth in 6:550–556
  Dodd–Frank Wall Street Reform and Protection Act 8:61
  Dominican Republic emigration 15:464
  East Asian students in 6:800–801
  economic history, quantitative 7:55–60
  elections 18:394–396
  empirical social research (ESR) in 7:560–561
  environmental justice in 7:756–762

United States (US) (*continued*)
Bean versus Southwestern Waste Management Corp. 7:756–757
Commission for Racial Justice 7:757
communities under siege 7:758–759
corporate welfare 7:760
dumping on Native Americans 7:760
early struggles 7:756–757
environmental justice health symposium 7:757–758
Executive Order 12898 7:758
government responses 7:757–758
interagency working group 7:756, 7:758
path of least resistance 7:759–761
people of color summit 7:757
Principles of Environmental Justice 7:757
sacrifice zones 7:758–761
transboundary waste trade 7:760–761
unequal protection 7:758–759
US General Accounting Office 7:757
environmental justice in
childhood lead poisoning 7:759
environmental security 16:254
ethnicity, race, and crime in 5:155
Europe migration from 15:462
euthanasia in 7:600
e-waste network 7:366f
family law in 8:779–785
federal law
for domestic violence 6:628
Federal Resource Conservation and Recovery Act 13:505
federal system 8:876
feminist activism 12:876
firms
in Asia 25:61
venture capital 25:59–60
First Amendment's religion clauses 20:279–280
gay and bisexual Latinos in
see also Latinos
group psychotherapy 19:505
psychotherapy, early studies 19:504–506
sexual partnership in 21:731
sociology 22:991–993
gay/lesbian movements in 9:632–636
bisexual 9:632, 9:635
future of 9:636
gay liberation and lesbian feminism 9:633–634
HIV/AIDS 9:634
homophile 9:633, 9:636
pre-World War II 9:632
religious right and culture war 1990s 9:634–635
same-sex marriage and contemporary LGBT movement 9:635–636
transgender 9:632, 9:635–636
from World War II to Stonewall 9:632–633
giving in 3:366–367
health care-related legal issues in 10:594–597
higher education (HE) market 10:856
high-technology start-up 3:47
historiography 15:702
homelessness in 11:159
human subjects protection and research ethics in 8:48–52
avoiding conflicts of interest 8:50–51
biomedical research 8:48, 8:50
Code of Federal Regulations 8:50
establishment of IRB 8:50
formulation of IRB 8:49–50
Health Insurance Portability and Accountability Act 8:49
informed consent 8:48–51
international codes 8:48
lessons from past and development of regulation 8:48–49
mental health 8:49
mission creep and criticisms of IRB 8:51
privacy 8:49
professional ethics 8:48
hunger in 9:285–286
immigrant fertility in 15:400–405, 15:400t–402t
immigration 11:635–636, 11:635f
distinct characteristic of 11:635
documenting the lives of undocumented immigrants 11:638
enforcement and the well-being of immigrant families 11:637–638
impacts of immigration on native's wages 11:638–639
integration of immigrants into US society and economy 11:637
permanent visas 11:635–636
immigration and crime in 11:628–634
civic engagement 11:631–632
crime protection effect 11:630–631
economic motivation 11:629–630
emerging gateways 11:631
empirical evidence 11:629–630
familism 11:630
foreign born 11:628–630
historical settlement patterns 11:629, 11:631
importation 11:629
major continuous gateways 11:631
new *versus* old destinations 11:631
preemerging gateways 11:631
selective immigration 11:630
social disorganization 11:628–629, 11:631
theory 11:628–629
victimization 11:632
immigration policy 11:662–665
impact of family status on labor market in 8:758
incorporated municipalities 24:854–855
individualism *versus* collectivism 6:9–10
industrial policy in 5:537
industrialization 1:218
industrialized *versus* nonindustrialized societies 6:9
intelligence services 24:764–765
Jewish Diaspora 12:803–810
labor earnings inequality in 6:767
labor movement 11:861
law and social science in 13:767
empirical methods 13:769
future directions 13:770
jurisprudential roots of social science evidence 13:767
standards of admissibility for social science evidence 13:767
uses of social science in law 13:767–770
law firms 12:472
law firms in 13:578–581, 13:583
legal process and social science 13:767–771
legal system and databases 5:803
importance of legal databases 5:803
precedents and stare decisis 5:803
levels of religious affiliation 20:298–299, 20:299f
liberal settler societies in 14:13
litigation system 6:517
locations of office quarters in 3:603–604
managed care and psychotherapy in 14:462–465
criticisms 14:463–464
detracting from practitioners' professional autonomy to advocate for patients 14:464
evolution 14:462–463
increased nonreimbursable administrative time 14:463
mental health treatment 14:462–465
patient-centric care management 14:464–465
perceived lack of accountability 14:464
1970s 14:462
1980s 14:462–463
1990s 14:463
sacrificing quality for cost savings 14:463–464
measurement of criminal victimization in 5:720–726
intimate partner violence 5:723–725
National Crime Victimization Survey (NCVS) 5:721–725, 5:722t
National Family Violence Survey (NFVS) 5:724–725
National Incident Based Reporting System (NIBRS) 5:720–721, 5:725
National Intimate Partner and Sexual Violence Survey (NISVS) 5:721–722, 5:724–725
National Violence Against Women and Men Surveys (NVAWMS) 5:721–725, 5:723t
National Youth Survey (NYS) 5:721, 5:723t
rape and sexual assault 5:721–723
reports of law enforcement officials 5:720–721
stalking 5:725
surveys 5:721–725
Uniform Crime Reporting Program (UCR) 5:720–721, 5:725
metropolitan areas (MAs) 24:856–858
migration from Latin America 15:463
migration from Mexico 15:463
military–industrial complex 9:540

model of downtown in metropolitan areas 3:602f
model of edge cities in metropolitan areas 3:602f
model of rural planning 20:803–804
modern farming in 20:803
modernization theory in historical work 15:701–702
mortality crossover 15:848–850
National Environmental Policy Act 1969 (NEPA) 7:773, 7:781
National Environmental Protection Act (NEPA) 7:811
national security and domestic terrorism in 16:247–252
  Al Qaeda and associated movements 16:247–248
  Al Qaeda influenced individuals 16:249–250
  CT training and CVE programs 16:250–251
  current domestic terrorism and extremism policies 16:249–251
  domestic extremists' offending patterns and national security 16:251–252
  environmental and animal rights extremists 16:249
  far-left 16:252
  far-right 16:248–249
  new legislation and stricter penalties 16:250
  policing 16:251
  use of confidential informants 16:250
New England towns 24:855
new liberalism 18:299–306
nonprofit institutions in 18:320–325
organized crime in 5:173
  mafia 5:173
origins of food production in 9:303–305
patenting in biological sciences 17:588
PATRIOT Act 11:637, 13:531, 15:469
penal revolutions 19:654
Pendleton Act 1883 5:17–18
philanthropy, economics of 3:359–360
  givers 3:359–360, 3:360t
  sources 3:359, 3:359t
  totals 3:359
  trends in giving 3:359
  type of charity 3:360, 3:360t
policy knowledge
  and Europe, comparison 18:297–298
  university system 18:295–296
population composition by race and ethnicity in 18:555–557, 18:556t–557t
poverty measure 18:741–742
presidency 18:861–863
presidential elections 18:394–395
press, freedom of 9:397
probation, parole, and jail populations in 20:31–32
profit-maximizing, sports economics 23:319
Publicly Operated Waste Water Treatment Plants 7:773
race/racism beyond borders of 19:798
racial discrimination in 6:543–544
racial oppression 19:848
racial/ethnic differences in 6:8–9
recognizing and delineating MSAs 24:856–857
regulatory protections for human research participants 4:252–257
religion in 3:541–542
  distinctiveness and vibrancy of religion 3:541
  religious diversity 3:541–542
  religious traditionalists *versus* religious progressives 3:542
  trends 3:542
religious free market 20:279–280
religious rights in 13:217
religious service attendance 20:299
rise of class consciousness in 25:55
in search of public philosophy 20:895–896
settlement houses 1:218
sexual harassment in 21:721–727
  in comparative context 21:725
  employment discrimination 21:721, 21:723
  gender 21:722–726
  grievance processes 21:723–725
  hostile working environment 21:722
  legal rules and workers' experience 21:721–723
  quid pro quo harassment 21:721–723
  sex discrimination 21:721–722
  sexuality 21:722
  stereotypes 21:723, 21:725
Smoot Hawley Act 1932 12:568, 12:570
social context of community in 4:385–389
  citizenship, nations, and globalization 4:388
  civic engagement 4:388
  civil society 4:388
  communities of affiliation and social support 4:387–388
  community social ties 4:385–387
  influence 4:385–387
  institutions 4:388
  intergroup tolerance and tolerance 4:387
  neighborhood and access to opportunity 4:386
  networked world 4:388
  social capital 4:385–386, 4:388
  social movements 4:388
  social networks and economic action 4:386–387
  trust 4:385–386, 4:388
  urban and rural contexts 4:385
social life, contemporary 5:266
and social work 14:663–666
sociocultural aspects *see* North America; sociocultural aspects
sponsored syphilis studies in Tuskegee and Guatemala 13:561–562
status of aesthetic education in 1:232–234
  impact of reduced provisions for arts education 1:234
  reduction in instructional time 1:232–234
  stratification and social classes in 22:340–341
  subprime lending in 3:580–581
  sustainable development in 23:803t
  tax system 23:340, 23:342
  teaching in 24:75–82
  training in clinical psychology in 24:517–523
  universities, hazing in 10:556–559, 10:561b
    alcohol abuse 10:557–558
    court judgments 10:559
    history 10:557
    laws against 10:558
    practices and deaths 10:558–559
    prosecution and prevention 10:559
  urban clusters 24:856
  urban definitions in 24:854–856
  urban planning
    early 1:218
  urban poverty in 24:915
  urbanized areas 24:855
  woman's suffrage in 22:443–444
  and World War I 9:214, 9:216
  wrongful convictions in 25:789
  youth perception of spirituality 20:298
United States Agency for International Development (USAID) 6:302, 20:462
United States Aid to Families With Dependent Children 17:465
United States Bankruptcy Code 2:354
United States Belmont Report 21:316
United States Bill of Rights 3:687
United States census microdata 3:285–286
  1990s 3:286
  1960s and 1970s 3:285–286
United States censuses 10:303
United States Civil Service Commission 11:350
  human resource management 11:350
United States Code (USC) 5:807–808
United States Code Annotated (USCA) 5:807
United States Code Service (USCS) 5:807
United States Congress 1:611
  democratic accountability 4:653
  elections 4:649
  and foreign policy 4:652
  gerrymandering 4:649–650
    incumbency advantage 4:650
    primary elections 4:649–650
  internal organization 4:650–652
    committees 4:650
    house procedure 4:651
    legislative procedure 4:651
    party organization 4:650–651
    senate procedure 4:651–652
  polarization 4:652–653
  political representation 18:441–442, 18:444
  and president 4:652

United States Court of Appeals for the Federal Circuit (CAFC) 17:588–590
United States Cuban embargo 5:369–370
United States Current Population Survey (CPS) 5:742–744, 17:462, 17:464–465, 20:888
United States Decennial Census 20:884
United States Department of Agriculture (USDA)
County Service Centers 20:421
United States Department of Education 23:570
Institute of Education Sciences 1:233
United States Department of Health and Human Services (DHHS) 15:110, 15:115
Healthy People 2020 program 15:110
United States Emancipation Proclamation (1862) 22:41
United States Environmental Protection Agency (EPA) 1:568, 5:777, 7:757, 14:45, 22:388–389
United States Equal Opportunity Commission 8:884
United States federal agency
surveys and polls
obligations to, sponsors 23:753
protections for human research participants 23:754–755
sponsors 23:753
United States Federal Election Commission (FEC) 18:396
United States Federal Trade Commission 1:204
United States General Accounting Office 7:757
United States Geological Survey (USGS) 10:58, 20:422
United States Health and Retirement Survey (HRS) 1:416
United States Holocaust Memorial Museum 10:7
United States Institutional Review Board (IRB) 20:509–510
United States Law and Society Association 13:515
United States Liaison Committee on Medical Education 12:359
United States Medical Licensing Examination (USMLE) 7:188–189
United States National Academy of Sciences
Board on Sustainable Development 23:805–806
Committee on Nuclear and Alternative Energy Systems (CONAES) 7:622–623
United States National Association of Black Social Workers 1:137
United States National Institute of Justice (NIJ) 22:45
United States National Longitudinal Surveys 17:462
United States National Nanotech Initiative 16:189
United States National Research Council 10:302, 10:304
United States National Science Foundation 9:620
United States Office of Personnel Management 11:350
United States Patent and Trademark Office (USPTO) 11:834–840, 12:269, 17:585–586, 11:836t, 11:837f, 11:839t
United States Patriot Act 18:943
United States Postal Rate Commission 20:166
United States Preventive Services Task Force (USPSTF) 8:336
United States Privacy Act 1974 18:943, 18:945
United States Prohibition Movement 15:590–591
United States Public Health Service (USPHS) 8:48
United States Public Law 93-348 20:514
United States Securities and Exchange Commission 13:621
United States Social Security system 20:604
United States Soil Conservation Service 7:810
United States Statistical System 10:304
United States Supreme Court (USSC) 13:513
judicial independence and 12:894
United States Tax Relief and Job Creation Act 2010 10:373
United States Treasury 16:462
United States v. Virginia (1996) 13:768
United States v. Windsor (2013) 21:770–771
United States–Cuba relations 5:373
United States–Japan Security Treaty (ANPO) 14:638–639
United States–Mexico relations 15:328
United We Dream 9:630
Queer Undocumented Immigrant Project (QUIP) 9:630
Unitive ecstasies 16:170
Units, quantification 19:701
Units to Aid Production (UMAP) 5:369
Unity
in China 2:65–67
and postmodernism 18:685–686
Unity in diversity 16:31–32
Unity of humanity, Claude Lévi-Strauss 13:907–909
Unity of moral judgment and action 13:119
Unity of reason, Albert on 1:498
Unity of Science 25:90–91
Unity of social Darwinism 5:687–690
UNITYP project 3:812
Univariant coding 25:185
Univariate discrete distributions 6:580–583
Bernoulli distribution 6:581
beta-binomial (negative hypergeometric) distribution 6:583
binomial distribution 6:581
combinations of distributions 6:583
distributions without a zero class 6:582
geometric distribution 6:582
hypergeometric distribution 6:582–583
multivariate discrete distributions 6:583
negative binomial distribution 6:582
Poisson distribution 6:581
Univariate gender differences 9:751
Univariate methods 8:600–604
Univariate models 24:337–338
Univariate samples
outliers in 23:388–389
causes 23:388
meanings 23:389
retention, identification, rejection, incorporation 23:388–389
Univariate statistics, Pearson's 17:661
Universal access to research products, political decision and international participation 8:34
Universal causal laws 19:679
Universal child 3:492–493
Universal Church of the Kingdom of God 16:806
Universal Declaration of Human Rights (UDHR) 2:670, 3:676, 3:686, 10:386, 11:360–361, 11:796, 12:538, 12:956, 13:440–441, 13:720, 14:12, 22:392
challenges to 11:355–356
creation 11:361
cultural, economic, and social 11:355
and death penalty 5:874–875
and disaster 6:481
political and civil 11:355
provisions of 11:355
and social work 5:518, 11:354–356
statement on human rights 11:361
empirical 11:361
epistemological 11:361
ethical 11:361
Universal Declaration on Bioethics and Human Rights 20:512, 20:514
Universal design
for disablism 6:450–451
for learning 23:216–217
Universal Design for Instruction (UDI) 23:216–217
Universal emotions 5:543–544
and basic emotions 5:544
expressive behavior, studies on 5:543–544
physiological responses, studies on 5:544
Universal Ethics Project 15:809
Universal ethnography 1:754
Universal Grammar (UG) 9:202, 9:206, 13:319–322, 15:751, 21:362
Chomsky on 3:529
Universal health access 8:34–35
Universal history 10:162–168, 11:32, 11:34
Universal History 11:44–45
Universal human rights instruments and children 12:956
Universal jurisdiction, political trials and 18:500–501
Universal kriging 23:193
Universal nature of child-directed speech 3:401–402
Universal Pictures 9:156–159
Universal Postal Union (UPU) 18:642
Universal pragmatics 5:305

Universal principles for politeness 18:327–328
Universal quantifiers 19:707–708
Universal reconstruction (linguistics) 12:448–449
Universal registers 6:163
Universal religions 20:363
Universal rights 3:657, 13:900
Universal social work 12:557–558
Universal Survey Museum 16:145
Universal time 24:406
Universal transverse Mercator (UTM) 10:59
Universal Turing machine (UTM) 1:534–535
Universalism 2:605, 5:312, 7:415, 9:619–621, 13:566, 21:262–263, 22:890–891, 25:511
  and anthropological comparison 4:412
  Benhabib on 2:519, 2:521–524
  biomedical research ethics 8:78–79
  and civilization 3:722–723
  and common sense 4:1–5
  cross-cultural study of education 5:331
  and identity movements 11:545–546
  and intergenerational justice 12:400–401
  Lévi-Strauss on 13:905–907
  and liberalism 14:14–15
  and morality 15:806–807
  and norms in science 17:11
  and parenting 17:493
  pattern variable 1:382
  and postmodernism 18:686–687
  and relativism 20:233–234
  versus relativism 1:765–766
  and values 25:28–29
Universality 4:20–21, 10:430
  Butler's work on 3:58–59
  of classification systems 3:800
  of emotions 7:459
  of five-factor model of personality 9:249–253
  of human rights 11:353–354
  of laws 12:934
  and pretend play 18:870
Universalization 12:392
Universals, art 2:16–17
Universe score 9:837
Universidad Autónoma de Honduras (UNAH) 22:624
Universidad Nacional Autónoma de México (UNAM) 22:623
Universidad Nacional de Nicaragua (UNAN) 22:623
Universities 7:166, 7:169–170, 18:307–313, 22:604, 22:621, 22:628, 24:765
  accountability 18:307, 18:309–310
  applied geography based on 1:845
  contributions for social work education 22:775
  departmental structure of 18:466
  in East Asia 6:801–803
  educational research 7:282–283
  European, science and technology at 24:758–761
  funding 18:308–312
  higher education 18:307–312
  higher education research centers based on 10:864
  and history of education 7:173, 7:176
  human capital 18:308
  Humboldt on 25:287
  implications for appropriate forms of knowledge 18:308–309
  implications for organizational setup of 18:309–310
  implications for social and behavioral sciences 18:311–312
  innovation 18:308–309, 18:311–312
  knowledge economy 18:307–308, 18:310
  knowledge society 18:307–309, 18:312
  mode 2 knowledge 18:309–310, 18:312
  political and social context in twenty-first century 18:308
  rationalization 18:308, 18:312
  reforming of 7:235
  research 18:307–312
  role in teacher education 24:51–55
  as site of cultural consumption 16:557
  for social science research 6:487–488
  as special organizations 10:845
  teaching 18:307–309
  transdisciplinarity 18:309–310
  transfer of knowledge into policy realm 18:310–311
  United States, hazing in 10:556–559
  United States, science and technology at 24:762–766
Universities, in history of social sciences 22:588–595
  alternatives to 22:590–591
  controls over disciplinary formation 22:592–594
    chair-holder system 22:592
    criticisms of the departmental organization 22:593–594
    disciplinary department 22:592–593
  discipline relationships 22:589
  knowledge 22:588
    institutionalizing 22:588–589
    organization of 22:588–589
  knowledge acquisition, continuities and discontinuities in 22:589–591
    scholasticism and humanism 22:589–590
  'knowledge revolution' of nineteenth century 22:591
  place and influence of the social sciences within 22:591–592
  transformation of 22:591–592
  of twenty-first century 22:594
    entrepreneurial university 22:594
    fading institutional boundaries 22:594
    primary role for 22:594
Universities Service Centre for China Studies (USCCS) 22:637–638, 22:640
Universitization, in education 24:51
University College of London (UCL)
  Center for Digital Anthropology 12:600–601
  Fisher at 9:222–223
University hazing, practices and deaths 10:558–559
University Leaders for a Sustainable Future 7:734
University of California Los Angeles (UCLA) study 11:630
University of California (UC) system 4:143
University of Chicago
  political science department 18:467
University of Colorado 25:561
University of Denver 13:516
University of Halle 7:235
University of Illinois 24:762–763
  Center for Urban Economic Development 1:221
University of Kyoto 12:795–796
University of Manchester Longitudinal Study 6:381
University of Michigan Survey of Consumer Attitudes and Behavior 8:515
University of Pennsylvania
  Smell Identification Test (UPSIT) 17:198–200
  Wedge Diagram 7:775f
University of Tokyo 12:795–796
University of Uppsala 18:466
University of Wisconsin 13:516, 24:762–763
University Press 19:624
  acquisition editors 19:626–627
  authors 19:626–629
  college publishing 19:625, 19:627–628
  contributions of 19:625–626
  corporate consolidation 19:627–628
  digitization 19:628–629
  editorial imagination 19:626–627
  evolution of 19:625
  forces reshaping 19:627–629
  functions 19:626
  future challenges 19:629
  globalization 19:628
  information technology 19:624
  knowledge 19:625–627
  marketing and sales 19:627
  modern, history 19:624–626
  networks 19:624, 19:626–627, 19:629
  peer review 19:624
  publishing culture 19:627
  scholarly book publishing 19:624–625, 19:627–629
  scholarship 19:624, 19:628–629
  specialty publishing 19:628
University-based research 23:214
  in criminology and criminal justice 5:128–129
  survey research 18:460
Univocal identity concept 17:744
  implications 17:747–748
  material content/formal competence 17:745–746
  mechanisms for stabilizing identity 17:746
  modern theories 17:745–746
  postmodern approaches 17:745

Univocal identity concept (*continued*)
stability of value commitments across a lifetime 17:747
theoretical-observing *versus* practical stance 17:746
Unjustified repressions, in Soviet Union 20:823
Unknowns, risk 20:708
Unlearning, and sexuality 10:814
Unlimited democracy 25:276
Unmanned aerial vehicles (UAVs) 13:534
in war 25:379–381
Unmanned factory 2:297
Unmarried childbearing 8:768–769
Unmarried couples, event history analysis for 8:309
Unmarried father 8:783
Unmarried motherhood 8:782–783
Unmarried mothers 8:783
Unmarried parents, childbearing by 9:13–16
Unmet expectations 19:367–368
Unmitigated self-interest 22:163–164
Unnatural experiment of prolonging life 17:172–173
Unnatural fornication 9:645
Unobjectionable positive transferences 24:545–546
Unobserved components models 24:324, 24:338, 24:355–356
Unobserved heterogeneity 16:52
Unobserved variables *see* Latent variables
Unobtrusive measures 3:69
in ethnography 8:188
history of 24:767–773
and cultural turn 24:771–772
documents 24:768–769, 24:771–772
ethics 24:770–771
impact, use and generative problem 24:770
observation 24:768–771
origins and contexts 24:767–768
sources 24:768–769
and technological change 24:772
traces 24:767–769, 24:771
triangulation 24:769–770
UNODC *see* United Nations Office on Drugs and Crime (UNODC)
Unofficial reality 2:752–753
Unpaid labor 3:90–91
Unpaid work, time use 24:389
childcare 24:389
shopping and services 24:389–390
Unpartheiische 11:31
Unpredictability
behavioral economics 2:440
Unproductive giftedness 7:275
Unprotected markets 14:422–423
Unprotected sex 21:702–703
anal sex 21:798
oral sex 21:798
vaginal sex 21:797
Unranked societies 11:894–895
Unreal Engine 3 25:133
Unrealistic optimism
health risk perception 10:704–705
see also Optimistic bias
Unrecognized ignorance 24:718
UN-REDD *see* United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD)
Unrelated diversification 6:585
Unreliability, degree of 3:786
Unreliable expert testimony, legal safeguards against 8:584
Unrepresented litigants, out-of-court assistance to 12:925
Unresolved adults 2:139
Unresolved conflict 17:620
Unrestricted sense lexicons 13:943
Unruh, Jesse 18:393
UNSD *see* United Nations Statistical Division (UNSD) survey
Unshared environment 9:826
Unsociable children 2:456–457
Unsocial sociability 4:337
Unspecified personality disorders 17:895
Unstressed syllables 23:715–716
Unstructured interviewing 12:671–672
Unsuccessful aging 1:438
Unsuitable enemies 6:672
Unsupervised learning 15:964, 23:413
motor control 15:964
Untermenschen (inferior races, Nazi Germany) 22:41–42
Untersuchungen über die Libido sexualis 21:846
Untested arbitrators 12:472
Untouchability 3:224–225, 20:375
Unusually extraordinary 3:354–355
Unwanted pregnancy 21:797, 21:799–800
Unwanted sexual attention 21:722–723
Unwin, George 3:21
Unwittingly paraphrasing 18:58–59
Unwrapping, in agent-based modeling 22:704
Unwritten doctrine, Plato 18:212
UNWTO *see* United Nations World Tourism Organization (UNWTO)
UOE *see* UNESCO, OECD, and Eurostat (UOE) survey instrument
UOO *see* Uncertainty of outcome (UOO)
UPDA *see* Uganda People's Defense Army (UPDA)
UPE *see* Unanticipated-perverse effect perspective (UPE)
UPN *see* United Paramount Network (UPN)
UPOV *see* Union International pour la Protection des Obtentions Végétales (UPOV)
Upper class
elites 7:397–400
and social mobility 22:427–428
and social stratification 22:709
Upper core of a game 4:871
Upper echelons theory, managerial and organizational cognition 14:480
Upper middle class 2:785
Upper Paleolithic, food production 9:300–301, 9:303–304
Upper respiratory illness 12:35, 19:446–447
Uppsala Conflict Data Program (UCDP) 5:836
Upskilling 25:691
Uptake
cancer screening
determinants 3:79–80
implications 3:80–81
UPU *see* Universal Postal Union (UPU)
Upward assimilation theory 22:212–213
Upzoning/downzoning requirement 10:448
UR *see* Unconditioned response (UR)
Urbach–Wiethe disease 8:863, 13:680
Urban, J.B., on adolescent self-regulation 21:550–551
Urban agglomerations (UAs) 20:138, 24:859
Urban Alliance on Race Relations 8:153
Urban and Rural Planning Act 2008 20:139
Urban anthropology 1:751–752, 1:755–756, 24:774–782
consolidation 24:776–778
cross-cultural comparison 24:774–775
and cultural heritage 5:417
current developments 24:778–779
early roots 24:775–776
ethnicity 24:774–779
ethnography 24:774–777, 24:779–780
fieldwork 24:780–781
formational stage 24:774–775
and infrastructure 12:119–122
migration 24:775–780
and migration 15:422, 15:424
poverty 24:775–776, 24:779
slum 24:775–778
urban studies 24:774
urbanization 24:774–779
Urban areas
public action in relation to 24:908–909
segregation in 21:407–409
social characteristics 22:301–302
Urban Audit 5:779
Urban Bear Weekend 13:891
Urban bias 18:182, 24:945
Urban biography 24:812
Urban boosterism 24:850
Urban branding 24:888
Urban clusters 24:856
Urban conflicts 24:783–788
antiurbanism 24:786
capitalist dynamics and 24:784–786
changing city–state relationships as a source of 24:784
'city' as a source of 24:786
collective resistance 24:784
'cultural' analyses 24:786
disorganization 24:786
exemplar cities 24:786
megacities 24:785–786
protest behavior 24:786–787
research frameworks 24:783–784
social 24:783–785
social inequalities 24:785

state–city relationships 24:783
surveillance 24:784–785
urban charisma 24:786
urban civil society 24:783–784
urban segregation 24:785
violent and nonviolent 24:783–784
Urban containment boundaries (UCBs) 10:448
Urban contexts, community in 4:385
Urban culture, Sarmiento on 20:907, 20:910
Urban decay
in different political systems 3:600–604
in postsocialist countries 3:602–603
residential areas in Philadelphia 3:602f
in USA 3:600–602
Urban decentralization 24:823–824
Urban definitions 24:854–861
changes in 24:858–859
in China 24:859–860
city population 24:859–860
hukou system 24:859–860
incorporated places 24:854–856
metropolitan areas 24:854, 24:856
official definitions of urbanized or MAs in other countries 24:859
rural 24:855, 24:859–860
in United States 24:854–856
census designated places 24:856
changes for 1980 24:857
developments (1990–2010) 24:857–858
metropolitan areas (MAs) 24:856–858
New England towns 24:855
urban clusters 24:856
urbanized areas 24:855
US incorporated municipalities 24:854–855
used in 40 large countries 24:858
Urban densities 24:961
Urban design 24:619
and gender 9:711
physical activity 18:97
and public space 24:789–794
anxiety for democratic urbanism 24:789–790
challenge of diversity 24:790–791
compact cities 24:793
competitive cities 24:792
economic and environmental rationale 24:792–793
gentrification 24:790, 24:793–794
globalization 24:794
from old to new anxieties 24:793–794
privatization 24:790, 24:793
production of space 24:792
public life 24:794
public services 24:789–790, 24:793
shifting division of labor 24:791–792
social diversity 24:794
social exclusion 24:790, 24:792
spectacle 24:793
sustainable development 24:792–793
urban regeneration 24:790, 24:792
Urban development 22:115–116, 24:802, 24:938
analysis 24:802
changes in life expectancy 3:599–600
dual cycle model of 3:600
and infrastructure investment 12:118
intercultural comparison of cities 24:802
new variants of, in western world 3:599–600
in twenty-first century 3:598–604
types of 6:830–831
urban cultural geography 24:802
urbanism 24:802
Urban ecology 20:563
Urban economic resilience 20:565
Urban enclaves, and community studies 4:369
Urban Enterprise Zones policy 24:849
Urban entrepreneurialism 24:795–799
Baltimore 24:796, 24:798
capitalism 24:796–797
in context 24:797–799
defined 24:795–796
economic development 24:795–797
Harvey on 24:795
local government 24:796–799
Marxism 24:795
neoliberalism 24:799
policy mobility 24:797–798
public–private partnership 24:795–796
state 24:795–797, 24:799
strategy 24:796–797
urban governance 24:795–798
Urban environment 24:812–813
Urban ethnography *see* Urban anthropology
Urban form, and sustainability 18:199
Urban fringe 23:660–661, 23:663, 23:665
Urban gardening 1:566–567
Urban geography 3:602, 6:293–294, 20:130, 24:800–806
culture types of cities 24:802
development 24:802
analysis of development and morphology 24:802
intercultural comparison of cities 24:802
urban cultural geography 24:802
urbanism 24:802
empirical social science research 24:804
contribution 24:804
features 24:804
combination of conceptual and methodological capabilities 24:804
decoding highly complex encrypted urban social landscapes 24:804
specific methodological competencies and diversity of analytical responses 24:804
form and morphology 24:800
function and social structure 24:801–802
factorial ecology of cities 24:801
factors organizing space and megalopolitan structure 24:801
functions 24:801
land use development 24:801
perception of urban environment 24:801–802
segregation 24:801
social status differentiation 24:801
specializations 24:801
urban social geography 24:801
methods of empirical regional/urban subarea analysis 24:803–804
Neo-Marxian/Neo-Marxist 24:800
policy analysis 24:802–803
globalization and 24:802–803
government devolution 24:803
postmodern urbanism 24:802
scope 24:800
spatial reference levels and research emphases 24:805t
survey methods 24:804
urban social monitoring 24:803–804
and youth 25:827–828
Urban governance 18:181–182, 22:113–114, 24:795–798, 24:846–847
Urban greening 24:940
Urban growth boundaries (UGBs) 10:448
Urban growth management 7:769
Urban growth models 24:807–810
basic considerations 24:807
comments on state of the art 24:809
complete models 24:809
economic considerations 24:807
environment 24:807–808
invention of paradigmatic model 24:807–808
organized activities 24:807
residential models 24:808–809
residential spaces 24:807
resource allocation 24:807
Urban heat island (UHI) effect 7:816
Urban history 11:18, 24:811
defined 24:813
globalization 24:814–815
institutionalization 24:815–816
interdisciplinarity 24:813–814
origins 24:811–812
subdiscipline, emergence of 24:812–813
Urban inequality 23:150
Urban informality 24:818–822
capitalism 24:818–819, 24:821–822
frontiers 24:821–822
informal sector 24:818–819
microfinance 24:821
postcapitalism 24:821–822
poverty 24:818–821
production and regulation of space 24:820–821
slums 24:819–821
state 24:818–821
subaltern 24:819–820
urban citizenship 24:818–820
urbanism 24:819–820
Urban Institute and Mathematica 22:663
Urban integration 3:229
Urban internal spatial structure
built environment types within urban areas 24:825–826
driving forces 24:823, 24:827–828
identification of driving forces of 24:827–828

Urban internal spatial structure (*continued*)
investigation of, and associated outcomes 24:828–829
land conversions, types and quantity of 24:827
measurement 24:824–825
metrics of 24:824–825
monocentricity and polycentricity of cities 24:823–824
outcomes 24:823, 24:828–829
Urban labor markets 23:158
Urban legends 20:788–794
Urban life
definition 24:830–832
and food deserts 9:286–287
future developments 24:836–837
automobiles 24:836–837
consumption 24:836
determinants 24:836–837
issues facing decision makers 24:837
mobility 24:836–837
physical access 24:837
production 24:836–837
technology 24:836–837
and health 24:830–835
healthy city 24:832–833
human ecology 24:831–832
risk factors 24:831
supportive environments 24:833–834
WHO Healthy Cities project 24:833
New Economy 24:836–837
Urban managerialism 11:237, 24:795
Urban megaprojects 24:838–845
commodification 24:839
comparative research 24:840
competitiveness 24:838–841
displacement 24:838, 24:840
gentrification 24:839–842
globalization 24:838, 24:840, 24:842
iconic buildings 24:839
inequality 24:838, 24:841, 24:843–844
new generation of 24:838–840
overspending 24:842
paradox 24:842–843
problems and methodological challenges in the study of 24:843–844
public–private partnership 24:841–842
and urban planning 24:842
results of 24:842–843
segregation 24:840, 24:842
social mobilization 24:842–844
strategic planning 24:842
urban entrepreneurialism and megaprojects 24:840–841
waterfronts 24:839–840
Urban migration, and changing profile of aging 1:321
Urban neoliberalism 24:846–853
commodification 24:847
development 24:846–852
entrepreneurialism 24:847
entrepreneurialization of urban governance 24:847
financialization 24:851–852
global recession 24:851–852
Global South 24:847–848, 24:850
neoliberal art of governing cities 24:846–847
neoliberalizing urban economies 24:849–851
reproducing neoliberalism 24:850–851
privatization 24:846–848
recession 24:846–847, 24:851–852
variegated geographies of 24:847–849
Urban operating system 6:836
Urban Plan Act 1990 24:876–877
Urban planning 18:188–189, 18:191
in China 24:869–881
in ancient and medieval China 24:869–871
contemporary planning 24:877–878
in modern times 24:872–874
in the People's Republic of China 24:874–878
and planning education 24:878–881
competitions and exhibitions 24:882–888
congestion control 1:219
cultural phenomenon 24:887–888
defining 24:882
and development 18:180, 18:182–185
ethical issues 18:193–194
ethical perspectives 18:194
and gender 9:710
historical background 1:220
race and ethnicity and 19:820–824
agency 19:821–822
areas of scholarly inquiry 19:821–823
housing 19:822–823
issues 19:820–821
land use 19:822–823
social justice 19:821–822
urban structure 19:821–822
Urban Planning Department 20:138
Urban policies 10:50, 22:114
in Asia 24:894–899
brown agenda 24:899
decentralization 24:894–899
developmental state 24:898–899
diversities 24:894–895
East Asia 24:894–896
green agenda 24:895
market economies 24:894, 24:898–899
South Asia 24:894, 24:898
Southeast Asia 24:894, 24:896–898
state 24:894
transitional socialist market economies 24:894
urban sustainability 24:894
in Europe 24:900
capitalism 24:900–901, 24:904–905
changes 24:903–906
democracy 24:902–903, 24:905–906
governance 24:904
inheritance and institutionalization 24:902
neoliberalism 24:900, 24:903, 24:905
protests, conflicts, and implementation failures 24:903
state 24:900–902, 24:904–905
as land use and development 24:909–910
'NIMBY' ('Not in my back yard') 24:910
in North America 24:908–914
as social policy 24:912–913
and spatial mismatch 23:159–160
in the twenty-first century 24:913
Urban political economy 24:927–928
Urban politics 20:673–674
Urban population, and cyberbullying 5:641–642
Urban poverty 24:729–731, 24:915–918, 25:113
in America 24:915
census 24:915
concentration in neighborhood 24:915–916
discrimination 24:915–916
equity 24:917
ghettos and barrios 24:915–916
illiteracy 24:915
impact of global economy 24:916
importance of concerns about 24:916–917
job skills 24:916
reasons for 24:915–916
role of segregation 24:915–916
rural south 24:915
significance of regional economic growth 24:916
slums 24:915–916
social mobility 24:916
Urban redevelopment 10:47, 12:914–915
Urban regeneration 20:131, 24:790, 24:792, 24:866, 24:919–925
area-based interventions 24:921
challenges for 24:922–924
city marketing 24:920
community 24:920–924
concept of 24:919–921
economic competitiveness 24:920
historical evolution in the UK 24:921–922
physical regeneration 24:920, 24:922–923
social equity 24:920
urban decline 24:920–922, 24:924
urban renewal 24:919, 24:921–922
Urban renaissance 24:919, 24:921
Urban renewal 10:49, 16:438, 16:440, 16:451–452, 24:919
in different political systems 3:600–604
in postsocialist countries 3:602–603
two-stage cyclal model of metropolitan expansion and 3:601f
in USA 3:600–602
Urban resource markets 20:572
Urban restructuring, geographies of 12:91–92
Urban revitalization 24:919
Urban revolutions 18:672–673
Urban social movements 12:913–914, 20:673, 20:675, 20:677–678
Urban sociology 24:812
Castells's contribution 3:229–230
Chicago School 24:926–927
community and social networks 24:930
concentrated poverty 24:931
culture 24:929
future of 24:932

gentrification 24:931
globalization 24:928–929
methods in 24:929
neighborhood change 24:930–931
new 24:927–928
and policy 24:931–932
segregation 24:930
topics in 24:930–931
Urban space, in Mexico 1:630
Urban spatial structure and transportation 24:611
Urban sprawl 20:18, 24:934–939
history 24:934–935
anti-sprawl sentiments 24:934
objections to
aesthetic concerns 24:936
efficiency concerns 24:935
environmental concerns 24:936
social concerns 24:935–936
proposed remedies for 24:936–938
channeling and managing growth 24:937
design solutions 24:938
governmental structures and incentives 24:937–938
open space conservation 24:937
sprawl debate 24:938
Urban studies 24:774, 24:940–944
city 24:942–944
development of concept within 12:913–914
in flux 24:942
identification of core 24:940–941
interdisciplinarity 24:941–943
status of 24:942–944
transdisciplinary analysis 24:943
Urban Studies 19:821–822
Urban subarea analysis 24:803–804
Urban symbolism 24:779
Urban system
in geography 24:945–948
knowledge-intensive jobs in the hierarchy of 13:95–96
poststructuralist perspectives 24:948
Urban system plans 20:138
Urban theories, critical 20:673–674
Urban Tipping Point research project 24:783–784
Urban transformations, postsocialist 3:607–608
Urban transportation modeling system (UTMS) 24:618
Urban transportation planning models 15:636
Urban variable 24:812
Urban village movement 18:199
Urban villagers 16:437–438
Urban violence 24:949–957
African-Americans 24:949–951
Area Impact Munitions 24:955
cities 24:949–950
concentrated disadvantage 24:950–951
data 24:956
history 24:953–954
killing disease and killing cities 24:955
military operations in urban terrain 24:954–955
Palestinian urbanization 24:955
policing 24:955–956
poverty 24:949–951
segregation 24:949–951
surveillance 24:955–957
terrorism 24:953–957
urbicide 24:953, 24:955–956
vertical geopolitics 24:954–955
Virilio on 24:953, 24:956
Urban world, emerging cites in 12:117
Urbanism 11:908, 12:779–780, 15:252, 23:67, 24:802
in archaic states 23:372–373
Mumford on 16:137–140
postmodern 24:802
and race–crime relationship 19:841
in South America 23:38–39
World 6:282
Urbanization 3:579, 3:581, 7:816, 11:233, 23:798, 24:859–860, 24:862, 24:958–964
in Africa 24:965–972
in Australia 2:240–242
beyond rural-urban dichotomy 24:961
changes in households 11:230–231
changing settlement sizes and locations 24:961
changing urban densities 24:961
in China 24:973–979
cities 24:960–964
and climate change 24:963
and cosmopolitanism 5:45
crowds 5:361–363
and development geography 6:314–316
eco-cities 6:829, 6:833–835
effects on society 13:12
estimates 24:960t
by geographical region 24:960t
and globalization 6:285–287
inequality and social exclusion 24:963
and informal sector 12:56–57
and intensification/specialization 12:329
and internal migration 12:434–435, 12:434t
Jewish Diaspora 12:801
and language 13:281
in Latin America and Caribbean 9:69
levels 24:958–961
mass migrations 11:649
and migration 11:649
modernity 15:683
and modernization 15:707
Native Americans 1:635
nature and direction of causality 6:283–285
and nostalgia 17:35
past and future populations 24:958–961
postsocialist 3:609–610
processes 18:203
production 12:329
and property crime 19:163
and regional development planning 20:134–135, 20:137
rural sociology 20:814–815, 20:817, 20:819
and rural–urban linkages 24:961–962
rural–urban migration 24:962–963
and shifting environmental burdens 24:963–964
in Southeast Asia 23:73
statistics on 5:775
Eurostat 5:779
future of empirical data 5:779
history 5:778
and housing 5:778–779
quality standards 5:775
real estate economies 5:775, 5:778
UN-HABITAT monitoring system 5:779
United Nations Centre for Human Settlements (UN-HABITAT) 5:779
Urban Audit 5:779
and sustainable development 24:962–963
economic development 24:962–963
trends 22:113
urban anthropology 24:774–779
urban growth 24:958–959, 24:961–963
urban history 24:812, 24:814–815
urban sociology 24:926–931
and youth culture 25:808, 25:811
Urbanized areas (UAs) 24:855, 24:859
Urbanized villages 24:976
Urban–rural distinction 10:451, 24:961
Urban–rural integration plans 20:138
UrbanSim 24:890–891
Urbicide 24:953, 24:955–956
Urdu language 7:196–197
Ure, Andrew
on division of labor 6:603–604
works
Philosophy of Manufacturers 25:670–671
urGENOVA 9:837–838
Urinary incontinence 10:465
Urquhart, Thomas 12:762
works, *Logopandecteision* 12:762
Urry, J., works, *The Tourist Gaze* 24:471
URSs *see* Uncertainty reduction strategies (URSs)
Uruguay 22:625–626
al-'Urwa al-Wuthqa 12:742
US *see* Unconditioned stimulus (US); United States (US)
US Airways Flight 1549 crash landing 11:302
US Department of Agriculture (USDA) 24:391
Usability
in archiving 1:919
evaluation methods, human–computer interaction 11:402
Usable interventions 11:697
Usable models of group process 10:420–421
Usage dimension of digital divide 16:790
Usage gap 16:790
Usage-based model, in functionalist grammar 9:474
Usage-based theory 9:208

USAID *see* United States Agency for International Development (USAID)
USC *see* United States Code (USC)
USCA *see* United States Code Annotated (USCA)
USCCS *see* Universities Service Centre for China Studies (USCCS)
USCS *see* United States Code Service (USCS)
USDA *see* United States Department of Agriculture (USDA)
Use of force 17:632, 17:634–635
Use-based application/research of sociology 1:858
Useful energy 7:628
Useful knowledge, and applied geography 1:846–847
Use-misuse-abuse typology
  alcohol consumption 1:518
User as a content producer 16:795
User demand, for Internet privacy 12:594–595
User information regimes 2:214
User interfaces, autonomous agents in 2:311
User organizations, mental health 15:216–217
  functions 15:217
User oriented applied geography 1:845, 1:847
User written programs, econometric software 7:3
User–agent communication 2:311
  autonomous agents 2:311
User-defined small areas 6:167–168
User-generated content (UGC) 4:508, 16:795
  Internet 12:588–589, 12:591–592
  of social media 16:782
User-led innovation and service innovation 12:148
USER-P *see* Utrecht Scale for Evaluation of Rehabilitation–Participation (USER-P)
Users
  health care, gender and 9:696
  of program evaluations 19:117–118
    benefits 19:119–121
    user-oriented evaluations 19:120–121
  of psychiatric services 24:735
Uses (functions) 23:891–894
Uses and gratification theory 14:713–714, 15:30–31
USGS *see* United States Geological Survey (USGS)
Ushahidi 25:843
U-shaped curve 14:514
Usher, N., on LGBTQ families 21:756
USMLE *see* United States Medical Licensing Examination (USMLE)
USOs *see* Underground storage organs (USOs)
USPHS *see* United States Public Health Service (USPHS)
USPSTF *see* United States Preventive Services Task Force (USPSTF)
USPTO *see* United States Patent and Trademark Office (USPTO)
USSC *see* United States Supreme Court (USSC)
USSR *see* Russia; Soviet Union
Usury 12:386
Us-vs-them 18:262
Uterine rites 24:455
Utero folate deficiency 7:844
Uthman 12:732
Utilitarian functionalism 9:532–533
Utilitarian moral theory 13:487
Utilitarianism 2:634, 7:87–88, 8:74, 8:483, 10:491, 12:915, 12:935–936, 13:570, 18:196, 19:168
  of Bentham 2:526–527
  and collectivism 11:818
  and economics 7:88
  and egalitarianism 7:308–309
  and eudaemonism 8:213–216
  and happiness 10:524
  Harsanyi's arguments on 7:90
  and intergenerational justice 12:398–399
  Mill on 15:524–527
  and morality 13:484–485, 13:487, 15:808
  and political economy 12:388
  Rawls's contributions 19:914
  social thought 24:980–985
    Austin's 24:982–983
    Bentham's science of morals and legislation 24:981–982
    criticisms 24:985
    dissemination of utilitarian ideas 24:983
    happiness 24:980–985
    individualism and collectivism 24:984–985
    interests 24:980–982, 24:984–985
    jurisprudence 24:981–983
    legal philosophy 24:981–983
    liberalism 24:984
    (James) Mill's 24:983
    (John Stuart) Mill's 24:983–984
    moral theory 24:981
    motives and consequences 24:981–982
    origins 24:980–981
    pain 24:980–982
    pleasure 24:980–982, 24:984
    poor laws 24:982
    punishment and panopticon 24:982
    utility principle 24:981
  social work ethics 22:783–784
  and utility theory 7:88–89
  value pluralism 25:21–22, 25:24
  and welfare 25:505–509
Utilitas 2:529
Utilité 17:504
Utilities 4:867, 7:88, 7:102, 20:927, 25:12
  analysis 25:11–12
  drug wanting/needing 25:352–354
  elicitation of 7:384
  and infrastructure 12:120–122
  and interests 12:391, 12:393
  irrationality 12:722
  marginal 25:10–11
  microeconomics 15:333–334
  of public institutions 22:990
  and welfare 25:12, 25:504–506
Utility assessment 5:58
Utility functions 7:88
  Bayesian experimental designs 8:522–523
    loss functions 8:523
    multiple distributions 8:523
    quadratic loss 8:523
    Shannon information 8:522
Utility maximization
  and internal migration 12:444
  rational choice theory *see* Rational choice theory (RCT)
  rationality as 19:907
Utility measurement, in psychiatric care 19:298
Utility of life 10:522–523
Utility patents 11:835
Utility principle 24:981
Utility space 4:867
Utility standards 19:119
Utility theory 7:88–89, 24:612
  Bentham views 7:88
  Bernoulli thought 7:88
  in economics 7:88
  Jevons thought 7:88
  mathematical psychology 14:804, 14:809, 14:813–814
  ordinalism 7:88
Utility value 1:37, 7:890–891
  in expectancy-value-cost model 8:504–507
  in vocational psychology 25:241
Utility value interventions 1:38–40
Utilization, program evaluation 19:119–120, 19:120t
Utilization review of hospital care 14:462
UTM *see* Universal transverse Mercator (UTM); Universal Turing machine (UTM)
UTMS *see* Urban transportation modeling system (UTMS)
Utopian quandary 14:258
Utopian social order, and communes 4:266
Utopian socialism 22:834
Utopian thought 14:251–259
  ambiguous Utopias and dystopias 14:257–258
  ecological Utopias 14:256
  feminist Utopias 14:256–257
  idea of progress 14:254
  in industrial age 14:254–256
  literary and philosophical 14:251–259
  myth of the golden age 14:251–252
  non-western Utopianism 14:253
  renaissance Utopian literature 14:253–254
  roots in myth of golden age 14:252
  totalitarian impulse 14:258
  Utopian quandary 14:258
  Utopian treatises of antiquity 14:252–253
Utopian treatises of antiquity 14:252–253
Utopianism 2:150–151, 12:86, 21:178
  failure of 2:153
  promoted reform 2:154
  totalitarian dilemma within 2:153–154
Utopians 7:309

in American West 2:147–148
communal 2:146–147
Utopias 2:144, 14:251–259, 18:670–672, 18:674
and anarchism 1:681
Mumford on 16:137–139
Skinner on 22:25, 22:28–29
Utrecht Scale for Evaluation of Rehabilitation–Participation (USER-P) 6:468–469
Utterances 14:242, 14:393, 23:716
dexis 6:52–57
Uūl al-fiqh (Islamic law) 12:752–753
UVI *see* Ultraviolet Index (UVI)
UVR *see* Ultraviolet radiation (UVR)
Uzbek dynasties 3:317
Uzbekistan 3:316–320
Uzun v. Germany (2011) 18:950
Uzzi, Brian, on small world structures 5:467–468

V

Vacancy chains 23:471
Vaccinations 6:694, 15:869
psychoneuroimmunology 19:447
and smallpox mortality 15:869
Vaccines 18:174
Vacher de Lapouge, Georges, social Darwinism 5:689–690
Vagal nerve stimulation (VNS) 23:19
Vagal tone (VT) 2:867
Vaginal douching 21:799
Vaginal sex, unprotected 21:797
Vaginismus 21:718
treatment of 21:692
Vaginov, Konstantin, works, *Goat's Song* 21:742
Vague quantifiers 19:775
Vagueness, and ignorance 11:579
Vagus nerve stimulation (VNS) 16:718
Vailala Madness 3:140
Vain nuns 25:630
Vaisey, Stephen
on holism 7:417
on social network actors 5:558–559
Vajpayee, Atal Bihari 20:404
Vajrayana Buddhism 2:888–889, 25:632
Vakils 23:59
Valence 7:483
and emotions 7:520–523, 7:527–528, 7:530–531
of information 22:559
motivational 2:326–330
sexual 21:761
Valence, instrumentality, and expectancy (VIE) theory 25:681
Valency and argument structure in syntax 25:1–6
adjuncts 25:1
antipassivization 25:2
applicative constructions 25:5
causativization 25:3–4
change of 25:2–4
change of argument structure without change of valency 25:4–5
grammatical relation 25:1
noun incorporation 25:2–3
passivization 25:2
semantic role 25:3
Valente, Thomas W., on cigarette smoking 16:558
Valentijn, Francois, works, *Oud en Nieuw Oost-Indiën* 11:84
Valentine, David, works, *Imagining Transgender: An Ethnography of a Category* 19:755
Valentine, Gill 21:808–810
works, *Mapping Desire* 21:808–809
Valentinus of Alexandria 3:548
Valéry, Paul, works, *Problem of the Three Bodies* 2:729
Validation 4:490–491, 24:642
of agent-based models 1:370–371
close relationship 22:564
of conscious knowledge 13:86
Validation studies 20:241
Validation-in-parts 4:491
Validity 3:244, 3:822, 6:345, 12:250, 15:621, 17:530, 19:418–419, 20:248, 23:903
assessment 20:239–247
construct 19:739
external 19:739
in feminist research 9:718–719
field observational research 9:138–139
of implicit association test 11:704–705
group level 11:704
individual level 11:704–705
of indicators 11:752–753
internal 19:739
of nosology categories 17:32
of participatory action research 17:542
construct 17:542
ecological 17:542
expert 17:542
impact 17:542
projective methods 19:130
of qualitative research 19:687–688
relationship with reliability 20:252
single-case designs 21:988
of statistical clustering 23:395
statistical conclusion 19:739
Validity assessment 20:239
concepts of 20:239–240
designs for assessing 20:240–246
common factor model representation of CTST 20:242–243
internal consistency reliability 20:242
longitudinal methods 20:243–246
MTMM models 20:243
reliability models for single measures 20:243
simple reliability models 20:241–242
studies of reliability 20:241
test–retest reliability 20:243
validation studies 20:241
Validity generalization (VG) 11:882–883
Validity of law 18:25
in legal positivism 18:26–27
command theory of law 18:26
Hart's concept of law 18:26
inclusive *versus* exclusive legal positivism 18:26–27
pure theory of law 18:26
in natural law theories 18:27–28
Valorization 14:653
of war 15:490–491
VALS *see* Values and American Lifestyles (VALS)
Valuable, rare, and inimitable (VRIN) resources 23:522
τ-value 4:870, 4:878
Value and distribution, theory of 14:622
Value appropriation 3:33, 3:36
Value chain 3:43–45
resource markets 20:571–572
Value commitments
axiological neutrality 16:757
and personal identity 17:747
stability across lifetime 17:747
Value conflicts and surveillance 23:738–740
Value congruence 17:824
Value creation 3:33–36, 17:364–367
and entrepreneurship 7:681, 7:684
Value democracy 20:281
Value engineering (VE) 14:468–469
Value exclusion from science 16:755
Value freedom (*Wertfreiheit*) 25:465
axiological neutrality 16:756–757
in science, Weber's views 22:992
Value generalization, AGIL 1:388–389
Value intrusion 21:168
Value judgments
Albert on 1:497
axiological neutrality 16:755
Value neutrality 21:167
axiological neutrality 16:757–758
Value of change in mortality risk 25:14
Value of children (VOC) 17:495–496
Value of life 25:14–20
age 25:18–20
characteristics 25:14–16
costs of illness 25:15
ex ante and unidentifiable 25:15–16
labor earnings 25:15
contingent valuation 25:18
criticisms of 25:16–17
deontology 25:16
no single value 25:16
quality-adjusted life years 25:16–17
and economics 7:91
estimates of 25:17–18
risk tradeoffs in consumption 25:17
risk-compensating wages 25:17
stated preferences in contingent markets 25:17–18
heterogeneity 25:18–19
public expenditures and social burden 25:19
values for children, working-age adults, and seniors 25:18–19
values for various types of individuals 25:18

Value of life (*continued*)
values of others' lives – altruism 25:19
teleology 25:16
term 25:14
value of a statistical life 25:14
value of change in mortality risk 25:14–16
willingness to pay 25:14–19
Value of statistical life (VSL) 20:863–864, 25:14
Value of time (VOT) 24:630
Value orientations, and culture 5:426–427
Value pluralism 6:95, 25:21–26
Berlin 2:542, 2:546
incommensurability 25:22, 25:25
incomparability 25:24–25
Kantianism 25:23–24
monism 25:21–25
nonreductive 25:21, 25:25
pleasure 25:21–25
rational choice 25:21, 25:24–25
reductive 25:21
akrasia 25:22–23
arguments of 25:21–24
implications of 25:24–25
ordinary intuition 25:22
unavoidable loss 25:23–24
regret 25:23
relativism 25:24
utilitarianism 25:21–22, 25:24
Value proposition 3:33–35
Value rationality 12:389
Value relativism 21:169–170
Value spheres 5:550–551, 6:391–392
Value theory
rational choice explanation 19:897–898
cardinal, ordinal, Ramsey–Neumann cardinal 19:899
interests, consumptions, and welfare 19:898–899
Value-action gap 3:844–845, 3:845f
Value-added analysis, of economics of education 7:153
Value-added modeling (VAM) 24:79–80
Value-added services, economic growth 7:48–49
Value-added tax (VAT) 23:209
Value-based management 17:348–349
organizational culture 17:348–349
Value-based pricing models 12:589
Valued constraint-satisfaction models (VCSP) 4:719
Valued life activities (VLA) 20:657–658
Valued networks, centrality in 16:536
Valued outcomes, and extraversion 8:625–626
Value-expectancy model of opinion formation 17:244
Value-for-money movement 8:26–27
Value-free social science 22:699, 25:34
Value-free status of science 1:475
Value-incommensurability. 21:168–169
Value-in-diversity hypothesis 4:738, 6:595
Value-producing activities 25:670
Values 8:71–77, 15:928, 17:801–803, 23:798–799, 24:247, 25:7
achievement 19:794
Allport's theory 25:41–42
anthropology of 1:737–738, 25:33–35
behavior 25:33–35
cultural approach 25:33–34
ethics 25:33–35
human constructions 25:33, 25:35
sociological approach 25:33–34
subjectivity 25:33, 25:35
attitudes 25:41–43, 25:45
and attitudes 25:49
and behavior 25:45
on boundary of psychology 25:45
and communitarianism 4:308–310
constructs relating to
attitude toward learning 21:51
avoidance motivation 21:51–52
goal orientations 21:51
goals 21:51
interest 21:51
intrinsic motivation 21:51
motivation to learn 21:51
of cultural resources 5:488
cultural understandings of 5:538–539
culture 25:41, 25:44–45
culture as 9:51–52
of customer satisfaction 5:630–631
death 5:871
defining 25:47–48
definition 25:28
development across cultures 25:27–32
development cross cultures
benevolence 25:28–30
collectivism 25:28–31
conformity 25:28–29
cultural values 25:27, 25:29–31
egalitarianism 25:30
hedonism 25:28–29
individualism 25:28–31
personal values 25:27, 25:29–31
power 25:28–30
preferences, differences and similarities in 25:29–30
security 25:27–28
self-direction 25:28–29
social collective 25:29
and social life 25:27–28
sociocultural correlates of 25:28–29
surveys, cross-cultural 25:30–31
survival 25:30
tradition 25:27–30
universalism 25:28–29
well-being 25:30
developmental perspectives 25:50–51
in digital age 25:12
empirical ends 25:49
empirical research 25:51–53
and environmental cognition 7:706, 7:708–710
eternal 18:486–487
and exchange 8:476–477, 8:479–480
expectancy-value-cost model of motivation 8:503–509
Geertz on 9:664
generalization 25:48
governing socialization and education 22:865
identity movements 11:543–545
of indigenous peoples 11:805
individual differences 25:41
instrumental and terminal 25:49
and interests 12:391–394
internalization of 5:578
and international advertising 12:460
and isms 22:164–166
liberal 14:22–23
means–ends chains 25:49
measurement of 25:28
moral *see* Moral values
moral development 25:50–51
and motivation 21:49
and needs 25:49
and norms 17:6–9, 25:49–50
objective classical theory of 25:7–9
labor theory of value 25:7–9
Marx's 25:8–9
revival of classical theory of value 25:9
Ricardo's 25:7–8
Sraffa's 25:9
and organizational diversity 4:738–741, 6:595–598
organizational groups 10:408
and personality 25:45
pluralism of 18:237
political 18:429
and practical reasoning 18:776–781
of presidency scholars 18:862
and procrastination 19:74, 19:76
psychological perspectives 25:36–40
attitudes 25:36, 25:39
behavior 25:36–39
cognition 25:36–37
controversies 25:39
culture 25:36–38
definitions 25:37–38
history and current developments 25:38–39
moral 25:36–37
norms 25:36, 25:38–39
self-report 25:36, 25:39
virtue 25:37–38
public policy ethics 19:582
Putnam on 19:666
and relativism 20:233–234, 20:236–237
religious 17:804–805
revolution of 20:623–625
Rokeach's theory 25:42–43
Schwartz's theory 25:43–44
circular continuum of 25:43–44, 25:44t
common characteristics in different theories 25:43
empirical verification of circular model of 25:44
motivational continuum of 25:43f
and Second Demographic Transition 21:346–347
shared 5:314–315
social, diversity during adolescence 22:220

social, impact of external trade on 1:302
and social control 22:235–240
social psychology of 25:41–46
social work 10:144, 10:148, 22:782
sociology of 25:47–53
and solidarity 23:6–8
in Southeast Asia 23:73
subjective neoclassical theory of 25:9–12
equilibrium 25:10–11
marginal productivity 25:11
marginal utility 25:10–11
product distribution 25:11
utility analysis 25:11–12
utility and welfare 25:12
theory of 8:76
transcendental ends 25:49
Weber on 25:456–458, 25:460–461
at workplace 17:821–826
congruence 17:824
cultural values 17:824
cultural values and work-related behaviors 17:824–825
meta-analysis 17:824
national culture 17:823–824
work-related 5:315
Values and American Lifestyles (VALS) 1:840
Values management system, business codes 8:22–23
Values surveys, World Values Survey (WVS) 5:791
Valuing health 10:617–618
VAM *see* Value-added modeling (VAM)
Vampire Project 11:307
Van den Berghe, Pierre 8:463, 8:465, 11:283–284
works, *Human Family Systems* 8:465
Van Eijck, Koen
on cultural capital 5:378, 5:380
on cultural tastes 22:335–336
Van Gennep, Arnold 1:752, 7:145, 14:131–132
works, *Les Rites de Passage* 1:767
Van Hasselt, Vincent B., works, *Handbook of Social Development* 7:590
Van House, Nancy, on epistemic cultures 7:877
Van Huyen, Nguyen, works, *Introduction* à *l'Etude de l'Habitation sur Pilotis dans l'Asie du Sud-est* 11:86
Van Nieuwenhuijze, C.A.O., works, *The Sociology of the Middle East* 16:382
Van Orman Quine, Willard 19:392
Van Parijs, Philippe 9:401
Van Parjis, Philippe
post-Marxism 9:401
works
Real Freedom for All 14:43
Van Venrooij, A., on genre terms 3:803
Van Vollenhoven, Cornelis 13:813
Vanaprastha ashrama 11:786
Vance, Jack, works, *Rumfuddle* 14:256
Vandalism, and cultural heritage 5:416–417
Vandegrift, Darcie 24:478
Vandenbergh effect 17:198
Vanguard, Islamic fundamentalism 24:228, 12:744
Vanguard leadership 14:657
Vannevar, Bush
works
The Last Frontier 24:141
Science 24:141
VanTassel-Baska, Joyce, Integrated Curriculum Model 7:159
Vanuatu Education Strategy (2008–2009) 6:741
Vaquero 9:629
Varda, Agnes, works, *The Gleaners and I* 6:621
Varga, Agnes
on social network analysis 5:561
Variability 18:799–800, 21:949
blood pressure 11:455
in development 15:370
estimation in population studies 8:562
gender differences in 9:751
of information processing, and aging 1:431
multiple imputation 16:90
Variability reduction, by experimental designs 8:536
Variable access, in property rights 24:529
Variable analysis 12:353
Variable capital 3:90–91
Variable coding 22:550–551
Variable costs 14:553
Variable numbers of tandem repeats (VNTRs) 9:941–942
Variable risk function
LEXIS diagram 6:159–160
squares 6:159–160
vertical-sided parallelograms 6:159
Variable selection 19:434–435
approximation methods 19:434–435
cluster analysis 19:434
discriminant analysis 19:434
exact methods 19:434
factor analysis and principal component analysis 19:434
high-dimensional 15:650–651
regression 19:434
Variable sensitivity theory (VST) 19:473–474
Variable-oriented research 3:190–192
commonalities, study of 3:191
correlation between variables 3:191
Variable-population axioms 4:875–877
Variables
regression 14:153–156
simultaneous equations 21:980
in structural contingency theory 23:613
Variance (statistics)
analysis of variance 1:659–664
Fisher's theories 9:222–223
Variance components (VCs) model 8:715, 9:836–837
assumptions in 8:715
dominance and epistasis 8:715
likelihood of 8:718–719
mathematical formulation for phenotype vector 8:716
Variance function 14:153
regression analysis 14:157–160
Variance in facets 9:239
Variance of the error score variable, classical true score model 19:424
Variance of the observed score variable, classical true score model 19:424
Variation among individuals
epigenetic inheritance 7:832–833, 7:837
Variationist approach, for gender 9:698, 9:700–701
Variations 4:157, 5:679
among individuals, epigenetic inheritance 7:832–833, 7:837
health care markets 10:599
Variegation 24:847–849
Varieties of capitalism 11:831–832
Varieties of Initial Learners in Language Acquisition (VILLA) 21:355
Variety
community-based health interventions 10:647
and diversity 2:644
Varimax 19:79
Varis, T., on television 25:84
Varos (Turkish) 20:872
VARs *see* Vector autoregressions (VARs)
Varynen, R., works, *The Quest for Peace: Transcending Collective Violence and War Among Societies, Cultures and States* 25:374
Vasari, Giorgio 2:23
Vascular cognitive impairment (VCI) 13:361
language processes in 13:362–363
Vascular thalamic amnesia 6:722
Vasilievna, Seraphima 17:607–608
Vasopressin 16:470–471
Vassals, feudalism 9:114–115
Vastenavond/Vastelaovend(j) 3:146
VAT *see* Value-added tax (VAT)
Vaterlandslose Gesellen 22:888
Vatican Council
of 1870 3:234–235
of 1962–1965 3:235
Vaux, Clotilde de 4:520
VAWA *see* Violence Against Women Act (VAWA)
VBCMS *see* Victorian Bushfire Case Management Service (VBCMS)
VCE *see* Victoria Certificate of Education (VCE)
VCI *see* Vascular cognitive impairment (VCI); Verbal Comprehension Index (VCI)
VCs *see* Variance components (VCs) model
VCSP *see* Valued constraint-satisfaction models (VCSP)
VDCC *see* Voltage-dependent Ca^{2+} channels (VDCC)
VDs *see* Venereal diseases (VDs)
VE *see* Value engineering (VE)
Veblen, Thorstein 4:758, 10:484, 13:773, 22:332, 25:54–58, 25:192
and academy 25:56–57
on American pragmatism 25:54

Veblen, Thorstein (*continued*)
on conspicuous consumption/distinction 4:695–698
on consumption 16:555
contributions to modern sociology and economics 25:57
critique of neoclassical economic theory 25:55–56
on cultural lifestyles 22:331
on economic stratification 25:55
family background and early life 25:54–55
on Gilded Age 25:55
on higher education 25:56
on humanistic philosophy 25:54
on leisure class 25:55
on rise of class consciousness in America 25:55
on social Darwinism 25:55
on social inequality 25:55
on social rationality in postindustrial economies 25:56
on technological determinism 24:109–110
works
The Theory of the Leisure Class: An Economic Study of Institutions 4:695–696, 4:752
Veblen in Reverse scenario (leisure) 13:847
Vector autoregressions (VARs) 7:6, 7:9–10, 9:320–321, 14:406–407, 24:338
for political data 18:516–518
Vector majorization 14:450
Vector model 10:59–61, 10:60f
Vector psychology 13:913, 22:528
Vector valued index 8:698
Vectoral stage of development 16:504
Vectorborne disease, and natural disasters 16:316t
Vectors 9:915–917
in optogenetics 17:269–270
Vedas 20:335
Vedic Aryans 20:862
Vegetarianism, food preferences 9:298
Vegetation History and Archaeobotany (journals) 7:696
Vegetative systems, regulation of 11:486–487
Vehicle routing problem 23:174t
Vehicle theft 19:162–163
Vehicle-miles traveled (VMT) 24:616
Veil of ignorance 7:910, 11:579, 12:944
Rawls on 19:914–915
Veiling
of women in Middle East 16:362, 16:365, 16:368–370
Veiller, Lawrence 16:450
Velandia, Manuel 13:860
Veldhoven, G.M. van, works, *Handbook of Economic Psychology* 7:591
Velli, Michael, works, *Manual for Revolutionary Leaders* 20:639
Velocity, human growth data 9:514
Velthuis, Olav, on prices 5:538
Velvet revolutions 18:187, 23:348
Vendor performance (slower deliveries diffusion index) 9:317
Venereal diseases (VDs) 18:533–534
in lesbians 15:114
Venetian Maritime Republic 7:41
Venetian Statute of Inventors (1474) 12:263
Venezuela 22:624
Venice Charter (1964) 5:488–489
Venlafaxine 1:779–781, 17:470
Venlo Limburgish word 23:719
Venn diagrams 4:443
Ventral amygdalofugal bundle 14:127
Ventral amygdalofugal pathway 13:679–680
Ventral cortical streams 17:129–131
Ventral intraparietal sulcus (vIPS) 5:27
Ventral premotor cortex (PMv) 5:25–26, 16:673
Ventral stream
and attention 2:163–165
of cortical visual processing 25:187
and human language evolution 11:324, 11:326
and temporal lobe 24:207–208
Ventral striatum (VS) 16:640–641
Ventral tegmental area (VTA) 16:640
Ventral visual cortex, category object selectivity within 16:630–631
Ventriloquism effect 16:95
Ventrolateral preoptic area, and sleep 22:89
Ventrolateral preoptic nucleus (VLPO) 22:81f
Ventromedial prefrontal cortex (VMPC) 16:700–701, 18:704, 18:818, 22:485
Ventro-medial prefrontal lobes 16:740
Venture capital 25:59–62
entrepreneurship 25:59–61
finance 25:59–60
globalization of 25:61
history 25:59–60
industry 25:59
innovation 25:59, 25:61–62
investment process 25:60–61
investments 25:61
limited partnerships 25:60
SBICs 25:60
Silicon Valley 25:59
stock markets 25:59
technology 25:59–62
United Kingdom 25:60
US firms 25:59–60
Venture philanthropy (VP) 3:365, 18:295, 22:168–169
Venue shopping 1:359–360
Verb 11:825, 17:554
tense, aspect, mood, and evidentiality (TAME) 24:210
Verb classifiers 3:814
Verb forms 1:188
Verb–adverb combinations in American English 23:718
Verbal ability
gender differences in 9:754
of teachers 24:45
Verbal aggression 1:373
Verbal and nonverbal intelligence tests 12:324
Verbal behavior, Skinner on 22:25, 22:27
Verbal binding 15:143
Verbal category 8:351
Verbal classifiers (verb-CL) 3:812
Verbal communication 4:287
Verbal communicative competence 2:574
Verbal Comprehension Index (VCI) 12:284, 16:723
Verbal deficits, early 6:14
Verbal Delayed Recall Index 16:725
Verbal expressive behavior, Ludwig on 25:613–614
Verbal interference effect, and language 13:294
Verbal IQ (VIQ) 12:284
Verbal knowledge 13:93
Verbal Memory Index 16:725
Verbal memory procedures, during childhood 3:465
Verbal model, of social change 22:197
Verbal Motor Production Assessment for Children 3:796
Verbal number 17:43
Verbal quantifiers 19:708
Verbal rehearsal 21:912–913
Verbal reminders, during childhood 3:468–469
Verbal section of the SAT (SAT-V) 7:161
Verbal self-concept (VSC) 12:425–428
Verbal sound (VS), and language development 16:708
Verbal symbols 16:180–181
Verbal tests and developments in childhood 21:896
Verbal translation of preverbal memories during childhood 3:467–468
Verbal *versus* verb classifiers 3:814t
Verbal word form area (VWFA) 16:731
Verbal working memory (vWM) processing 3:328–329
Verbally accessible memory 16:198–199
Verbally gifted 7:158
Verbally talented 7:158
Verbal–mathematics continuum 21:465
Verbal–spatial binding 15:142
Verbatim form of language, significance of 13:266
Verbatim interview 12:671–672
Verb-CL *see* Verbal classifiers (verb-CL)
Verb–object combinations 23:718
Verboord, M., on cultural critics 12:589–590
Verdict-driven jury 12:911
Verdier, Yvonne 2:729–730
Vereenigde Oostindische Compagnie 15:680
Verein für Socialpolitik 22:990–992
Verein für Sozialpolitik 22:698, 25:675
Verfremdung 1:540
Vergence, oculomotor control 8:630
Vergesellschaftung 4:341
see also Socialization
Verguenza 21:871
Verhulst, Pierre 3:155
Verhulst, Pierre François 19:782

Verifiability theory of meaning 7:568, 14:316–317
Verification
of ABMs 1:370–371
hardware and software 5:938
of infection in diseases 12:33–34
Verisimilitude 5:275, 25:169
Verkko, Veli 5:240–241
Vermont General Assembly 25:786
Vernacular geography 10:69
Vernacularization 14:14–15
Vernadsky, Vladimir 9:567–568
Vernon, Philip 12:303–304
Verreichlichung 16:357
Verrücktheit (madness) 13:129–130
Verse 13:272–273
describing 13:272
Verstehen 10:785–786, 11:75–76, 11:384, 20:271, 20:345, 20:918
hermeneutics 10:778–779, 10:782
Verstehende 19:679
Vertebrate phyla 16:703
Vertebrates
cortical areas in 5:22–23
lateralization in
evolution 2:801–802
patterns 2:800–801
Vertex
eccentricity of 16:534
network centrality 16:532–539
Vertical cultural transmission 5:402–403
Vertical divestiture 20:169–170
Vertical equating 20:250
Vertical expansion 12:481
Vertical geopolitics, and urban violence 24:954–955
Vertical individualism, influence on homosexuality stigma 21:736–737
Vertical integration, diversification 6:585
Vertical logic model 14:302f
Vertical related industries
regulation in 20:169–170
input prices regulation 20:169
limiting the disadvantaging of retail rivals 20:169
market structure regulation 20:169–170
Vertical unity 2:76
Vertically integrated supplier (VIS) 20:169–170
Vertigo 25:73–74
entertainment 7:670
Vertov, Dziga 6:621–622
works, *The Man with the Movie Camera* 6:621–622, 9:167
Verum factum principle 20:204
Verzeitlichung 18:489
Vesicular monoamine transporter 2 (VMAT2) 16:749
Vespucci, Amerigo 23:44–45
Vessantara, prince 2:890
Vestibular disorders, adaptation and treatment options for 25:73–74
Vestibular nuclei 25:65–66
Vestibular system 25:63–69
balance 25:63, 25:67–69
cerebellum 25:63, 25:65–67
clinical manifestations of dysfunction 25:68–69
and aging 25:69
peripheral vestibular disorders 25:69
end organs, anatomy of 25:63–64
multisensory integration during self-motion in everyday life 25:68
otoliths 25:63–64, 25:66, 25:69, 25:64f
physiology of 25:65–66
hair cells and afferents 25:65
vestibular nuclei 25:65–66
proprioception 25:63, 25:67–68
semicircular canals 25:63–69, 25:64f
spatial orientation 25:67
stabilization of gaze by vestibuloocular pathways 25:66–67
stabilization of posture by vestibulospinal pathways 25:67–68
Vestibulo-ocular pathways, stabilization of gaze by 25:66–67
Vestibulo-ocular reflex (VOR) 8:629, 25:70–74
adaptation and treatment options for vestibular disorders 25:73–74
adaptive capability of 25:70–71
cerebellar flocculus 25:71
gaze 25:70–74
learning models in 25:71–72
long-term depression 25:73
long-term potentiation 25:73
motor learning
cellular mechanisms for 25:73
multisite hypothesis of 25:72–73
plasticity 25:72–73
prosthesis 25:74
vertigo 25:73–74
Vestibulospinal pathways, stabilization of posture by 25:67–68
VET *see* Visual exploration therapy (VET); Vocational education and training (VET)
Veterans
military anthropology 15:485–486
post-traumatic stress disorder 18:702–703, 18:705
Veterans Administration (US) 24:526
and clinical psychology, historical background 3:880
Veterans Health Administration (VHA), Mental Health Strategic plan 14:501
VEX-R *see* Violence Exposure Scale for Children (VEX-R)
Vezo ontology 9:810
VfEE *see* Voluntary facial emotion expression (VfEE)
VG *see* Validity generalization (VG); Volunteered geography (VG)
VGI *see* Volunteered geographic information (VGI)
Vh1 Hit Factory music show 25:844–845
VHA *see* Veterans Health Administration (VHA)
Via Campesina 6:861, 17:678
Viable yellow agouti mouse mode 7:835
Viagra 21:692
Vibrio cholerae 18:177
Vicarious conditioning 22:496
Vicarious learning 2:483
of aggression 1:374
Vicarious religion 8:260–261
Vicarious resilience (VR), in social work 18:711
Vicarious rewards, adolescent sexual risk 1:130
Vicarious sanctions, adolescent sexual risk 1:130
Vicarious traumatization, in social work 18:710
Vice
and victimless crime 25:75
virtue ethics 25:149–151
Vices in consent 4:809
Vichy regime 21:370–371
Vicious circles of bureaucracy 2:910
Vickers accumulator model 6:399
Vico, Giambattista 1:747, 11:12–13, 11:93–94, 11:96, 17:90, 24:744
on knowledge 10:785
relativism 1:765–766
theory of poetic logic 15:290
on unintended consequences 11:817–818
Victim advocacy 22:386, 22:388
Victim compensation, torts in law and 24:429–432
Victim empowerment 19:211
Victim surveys 5:125
Victimhood, and collective memory 4:198
Victimization 9:733, 13:441–442, 13:808–809, 19:160–166
burglary, prevalence of 19:160
criminal 20:774
fear of crime 8:856–857
of gay 9:641
in United States 11:632
Victimization experiences 5:146, 5:190
Victimization surveys 5:165, 5:167, 5:714–715, 5:718, 5:848
international 5:719
Victimless crimes 5:177–178, 5:719, 13:818, 25:75–79
approaches to reducing 25:78
arguments for/against prohibition 25:77–78
consensual crime 25:75–76, 25:79
drugs 25:75–79
future of 25:78–79
gambling 25:75, 25:77–79
justifications for legislation 25:75–76
harm to others 25:76
legal moralism 25:76
legal paternalism 25:76
offense to others 25:76
morality 25:75–76
perspectives on 25:76–77
pornography 25:75, 25:77–79
prostitution 25:75–79
public order offenses 25:77–78
vice 25:75
Victimology 5:719

Victims
and applied criminology 1:834–836
policy developments 1:836
child soldiers as 3:439–444
of crime
intersectionality and 12:667–668
definition of elder abuse 7:329
and offender, relationship between homicide 11:166
of torture
experiences and consequences 24:434–437
narrative exposure therapy 24:437–438
of white-collar crime 5:204
youth as 25:393
Victims of Trafficking and Protection Against Violence Act 11:632
Victor Turner Prize in Ethnography 1:742
Victoria Certificate of Education (VCE) 10:250
Victorian anthropology, and magic 14:435–436
Victorian Bushfire Case Management Service (VBCMS) 7:430
Victorian bushfires, social work response to 7:430–431
Victorian–Darwinian model of conception 20:451
Victorianism 24:596
Victory 11:392
Video, preserving dance 5:669
Video analysis, qualitative developmental psychology 19:673–674
Video cultures, television *see* Television
Video game epoch 4:502
Video games 6:896–897, 6:900
and academic achievement 4:498
and child development 15:1–8
and neuropsychological rehabilitation 2:817
Video industry 9:156–159
Hollywood 9:156–157
industrial organization analysis 9:158–159
Marxist industry analysis 9:158
neoclassical industry 9:158
rest of the world 9:157–158
Video recordings 19:671–672
Video sharing websites 6:898
Video streaming 25:169
Video-Feedback to Promote Positive Parenting and Sensitive Discipline (VIPPSD) 12:233
Video-on-demand (VOD) 2:212, 25:81–82
VIE *see* Valence, instrumentality, and expectancy (VIE) theory
Vienna Circle 9:119, 14:314–315, 18:619, 18:622–623, 25:87–90
logical empiricism 10:786, 25:87–94
and positivism 18:625–626
Vienna Convention on the International Sale of Goods 13:915–916
Vienna Psychoanalytic Society 7:934
Vienna System of pottery 1:694
Viennese positivism 18:626
Vierkandt, Alfred, works, *Handwörterbuch der Soziologie* 7:588–589
Viet Nam, nationalist history in 11:85
Vieth–Müller circles (VMC) 25:198–199
Vietnam 23:63
Bronze Age 23:66
centrality 5:101
Cham sites in 23:67
coastal route 23:64–65
fertility transition in 9:84–88
historiography development in 11:50
and Holocene transition 23:65
Iron Age 23:66–67
Phung Nguyen culture 23:66
plants and animals in 23:65
postcolonial law in 13:815
sociocultural aspects of *see* Southeast Asia; sociocultural aspects of
Vietnam Christians 3:532
Vietnam War 15:513–514, 21:308, 24:142, 25:379
Vietnamese Nationalist Party (VNP) 25:95
Vietnamese Revolution 25:95–99
history 25:95
theoretical debates 25:96–99
wars and 25:96
View cells 11:388
Viewshed analysis, archaeology 10:101
Vignette experiments 6:252, 19:904
Vignettes 15:230
Vigotsky's theory 4:533
Vihara 25:261–262
Vijayanagara 23:56, 23:56f
VILLA *see* Varieties of Initial Learners in Language Acquisition (VILLA)
Village anthropology, Korea 2:72–73
Village Colleges 8:612
Village Japan 12:794
Village Phone program 15:632
Villages
within a city 24:976
in South Asia 2:78
Villagization, land 13:233
Vilnius Declaration 22:646
Vinaya Pitaka 2:888
Vincent, André 16:228
Vincent, Paul 10:776
Vinck, H., on Central Africa 3:314
Vindicating views, morality 8:71
The Vindication of the Rights of Woman 7:117
Vineland Social Maturity Scale, historical background 3:879
Violations
of criminal law 5:189
and taboo 24:3
Violence 3:58–59, 4:608, 5:161, 14:430, 18:601, 25:357–360, 25:372–373
absence of *see* Negative peace
and adolescent health 1:121–123
adolescent violence 25:105–108
and aggression 1:379–380
agonism 1:444–446, 1:448
in anthropology 25:122–127
archaeology of war and conflict 4:600–606
bioarchaeology 2:601
of care 25:123–124
against children 12:960
coercion 25:117
cold war agendas for anthropology 25:122–123
collective 6:216–217
collective violence as social protest 25:118–119
commercializing 4:185–186
and crime 12:462–467
as crime 25:119–120
definitions 25:116–117
domestic 20:253–254
dynamics, and gender 9:784
entrepreneurs of 14:431
formations of 25:122
post-Cold War 25:123
and globalization 10:235–236
groups 14:431
halting of 14:705
and hate crime 10:543
historical research 25:117–118
recent approaches 25:120
history of 25:116–121
and homeless 25:111–115
and homeless women 25:112
and homeless youth 25:112
and honor 11:181–182
honor killing 11:185–186
human terrain system 11:393–395
humanitarianism in neoliberal times 25:123–124
international comparison of 25:118
of mafia 14:422–424, 14:426–427, 14:429–430, 14:433
managing 20:688–690
in Mexico 1:627–630
and military anthropology 15:483–485, 15:487
and naming, link between 24:454
organized 4:600, 4:602, 4:605
origins between groups 14:703t
PCV *see* pervasive community violence
and peace *see* Peace
and police use of force 18:255–256
political anthropology and 18:340
postsocialist transition *versus* armed forces demobilization 25:123
and power 25:116–118, 25:120
as precursor to homelessness 25:112
predominance of 6:553–554
psychopathy 19:451–453
public health 25:123–124
related terminologies 25:107
religious *see* Religious violence
and revolutions 20:627–629, 20:643–644, 20:646–648
roots and causes 25:412
school 21:107–108, 21:110–111
secrete shame and 21:875–876
sex work as 19:281–282
and slavery 22:37–42
in Soviet Union 20:822–823
state monopoly of 7:378–379
and statehood 23:350–351

terrorism 25:118, 25:120
topical examples 25:124–126
borders 25:125–126
care 25:125–126
humanitarianism 25:125–126
illegal migration 25:125
public health 25:125–126
transitions 25:124
truth and reconciliation processes 25:124–125
torture 25:116, 25:120
toward black lesbians in South Africa 21:747–748
urban 24:949–952
war 25:116, 25:118–120
war-related 4:608–609, 4:613
against women 5:161–162, 9:730, 9:732–733
and youth 4:632–634, 4:636
Violence Against Women Act (VAWA) 6:628, 8:914, 10:192, 10:540, 11:632
Violence Exposure Scale for Children (VEX-R) 25:108
Violence studies, of Collins 4:215–216
Violent abolitionism 22:40
Violent behavior, prediction by forensic psychologist 9:352–353
Violent Crime Control and Law Enforcement Act 20:434
Violent crime reductions 5:197
Violent endings of modern slavery 22:41
Violent entrepreneurship 20:854
Violent environment, and aggression 1:378
Violent gangs 25:822
Violent media, and aggression 1:378
Violent pornography 18:613
Violent resistance 12:686
Violent social crisis, and community resilience 4:358
Violent subcultures 5:78
Violentization, and genocide 10:17
Violent/nonviolent urban conflicts 24:783–784
Vipassana (insight) meditation 2:892
VIPPSD *see* Video-Feedback to Promote Positive Parenting and Sensitive Discipline (VIPPSD)
vIPS *see* Ventral intraparietal sulcus (vIPS)
VIQ *see* Verbal IQ (VIQ)
Viral latency, of AIDS 1:491
Virchow, Rudolf 17:1–2
Virgil 3:771
works
Georgics 14:252
The Lusiads 3:773
Virgin of Copacabana 3:535
Virgin of Guadalupe 18:146
Virginia 1:609
Virginity 21:793
Virilio, P., on urban violence 24:953, 24:956
Virtual campus 10:251
Virtual census 5:735–736
Virtual citizen science projects 3:632
Virtual classrooms 6:568
Virtual communities, and media 4:333–335
Virtual economies 7:36
Virtual education 6:567
Virtual environments 6:900, 11:390
in cognitive mapping 4:82–83
Virtual experiments 4:490
Virtual focus groups 9:274–275
advantages 9:274
disadvantages 9:274
Virtual identity 11:549–550
Virtual Morris water task 16:349f
Virtual organizations 6:239
Virtual participation 3:629–630
Virtual psychiatric sectors 14:4
Virtual public administration 19:524
Virtual radial arm maze (VRAM) 16:351–352, 16:350f
Virtual reality (VR) 11:549–550, 16:348, 18:42, 25:133–146
for animals 25:136–137, 25:137f
application in spatial cognition 25:135–136
applications for mental disorders 25:143–144
for schizophrenia 25:143–144
in substance-related disorders 25:143
therapeutic uses, for eating disorders 25:143
for assessment 25:144
basic components of 25:139f
and classical conditioning 3:769
devices for interacting with and presenting of 25:134f
future perspectives in 25:145
neural basis of location memory and navigation 25:136
and neuropsychological rehabilitation 2:816–817
projection setup for rats 25:137f
research on psychopathological processes 25:144–145
for research on therapy processes 25:144
as research tool 25:144–145
scenario for treatment of phobias 25:142f
for spatial navigation 23:163, 23:166–168, 23:167f
technology of 25:133–134
Virtual reality exposure therapy (VRET) 18:42, 25:138–140
advantages and limitations of 25:138–140
for anxiety disorders
for obsessive–compulsive disorder and generalized anxiety disorder 25:143
for panic disorder and agoraphobia 25:143
for posttraumatic stress disorder 25:142
for social anxiety disorder 25:142–143
efficacy of 25:143
for specific phobias 25:140–142
for acrophobia 25:141
for aviophobia 25:140–141
for other specific phobias 25:142
for phobia of small animals 25:141–142
for tunnel phobia 25:141b
Virtual reality lateralized attention test, for hemispatial neglect 10:767
Virtual reality therapy 19:379–380
Virtual selectionism, and functional explanation 9:523
Virtual senior centers 1:151
Virtual spaces 12:92, 21:808
Virtual spatial navigation 16:348
benefits, appeal, and limitations 16:348
eye-tracking 16:351–352
hippocampus 16:348–350, 16:352–354
medial temporal lobe 16:352
memory 16:348–350
neural aspects 16:352–353
and functional neuroimaging 16:352–353
in patient populations 16:352
and structural neuroimaging 16:352
in nonhuman animals 16:353–354
parietal lobe 16:352–353
psychological aspects 16:350–351
attention 16:351
strategies, behavior, and cognitive processes 16:350–351
sex differences and aging 16:351–352
simulated navigation 16:348–350
spatial learning 16:348–349, 16:352
spatial memory 16:348–349
tasks, types of 16:348–350
topographical learning 16:350
virtual Morris water task 16:349f
virtual radial arm maze (VRAM) 16:351–352, 16:350f
virtual reality 16:348
Virtual team training, cross-cultural training 5:334–335, 5:339–340
Virtual teams 5:339–340, 11:352
Virtual Tübingen 25:134f, 25:136f
Virtual universities 24:759–760
Virtual war 25:380–381, 25:408
Virtual water, calculations of 6:859
Virtual worlds 18:676
Computer Assisted Instruction (CAI) 4:494
and Internet 12:602
Virtual-arm radial maze 11:389
Virtual/material binary 12:92
Virtue ethicists 4:679
Virtue ethics 1:888–890, 2:634–635, 8:63, 25:147–152
agent-based 25:149
and applied ethics 25:150–151
Aristotle on 25:147–150
conceptions of right 25:147–148
definition 25:147
emotional reactions 25:151
Hume on 25:147, 25:150
intellectual context 25:150
love 25:148
nature of virtue 25:148–149
Nietzsche on 25:147, 25:149–150
response views 25:149
right action 25:147–148
situationist psychology 25:147, 25:151
skepticism about character 25:151
social work 22:784
trait of character 25:149
eudaemonism 25:149

Virtue ethics (*continued*)
noneudaemonistic virtue ethics 25:149
vice 25:149–151
Virtue theory 1:446
Virtues
of ignorance 25:148–149
Kantian ethics 13:7–8
Plato on 18:212
and self-interest 12:387–388
in small states 22:110
and trust 24:668–672
and values 25:37–38
VirtuSphere 25:133
Virú project 21:655
Virulence 15:869–870
VIS *see* Vertically integrated supplier (VIS)
Visa 3:658
Visceral modulation of startle 12:616
Visconti *signoria* 20:426
Viscosity
kin selection 13:29, 13:31, 13:29f
Visé, Jean Donneau de 15:11–12
Visher, Stephen 21:236
Vishwa Hindu Parishad 20:375–376
Visibility 17:621
Visible minorities 18:555
Visible speech 17:274–275
Vision
change blindness 3:344–349
high-level theory of 25:153–157
low level theory of 25:158–162
Marr on 14:596, 14:598
Vision Research at Michigan 21:952–953
catch-trials 21:953
original experiments 21:953–954
Vision Research Laboratory 21:952–953
Visionary planning 24:886
Vision/motion perception, low level theories of 25:158
future directions 25:161
global processes in extrastriate form vision 25:160–161
perceptual oscillations 25:161
retinal function, theories of 25:158–159
striate cortical function 25:159–160
Visitors 16:143–145
Visual agnosia 1:441–442, 25:179
definition 1:441
prosopagnosia 1:441–442
Visual alexia 6:734
Visual analytics 23:178
Visual and auditory illusions 20:50–51
Visual anthropology, in Mexico 1:629–630
Visual area 1 (V1) 25:184–187, 25:189
Visual area 2 (V2) 25:184, 25:186–187
Visual area 4 (V4)
concentric receptive fields 25:161, 25:160f
configural pooling models 25:161
Visual area 5 (V5) 25:202
Visual art markets 14:566–567
Visual arts 2:3–4
Visual attention, control of 2:167
Visual behavior, goal-oriented 10:270–271
Visual buffer, and mental imagery 15:222
Visual categorization 15:155
Visual cliff 10:129
Visual cognition, and mental imagery 15:222–224
creative thinking 15:223–224
learning and memory 15:222–223
mathematics and numerical cognition 15:223
reasoning and mental models 15:223
Visual communication 3:163f
Visual consciousness, theory of 4:676–677
Visual cortex 17:127–132
and attention 2:163–165, 2:167
brain, evolution of 2:830
and enriched environment 2:837
hierarchical processing in 16:629
human, object selectively in 16:630
ocular dominance and orientation columns in 5:31–32
stimulation 16:716–717
design concept 16:717
history and overview 16:716
ventral, category selectivity within 16:630–631
Visual culture 25:207–213
and Internet 12:604
and visual sociology 25:193, 25:195
Visual data, qualitative developmental psychology 19:671–672
Visual discrimination 4:676
Visual errors 4:93
Visual exploration therapy (VET) 10:769
Visual frames of reference
space as represented in 17:704
egocentric and allocentric frames 17:704
visual processing streams 17:704
Visual habituation 10:269–270, 10:272
Visual half-field technique 2:808–809
Visual hallucinations 10:499
fortifications 10:499
geometric figures 10:499
phosphenes 10:499
Visual images 25:221–225
child literacy and 25:170–171
in education 25:170–171
as harbingers of materialization 25:223
news and 25:172–173
persuasion and 25:171–172
of political advertisements 25:172
politics and 25:172
properties of 25:171–172
rituals 20:725
versus word 25:169–170
Visual images in media 25:169–174
Visual imagination 15:220, 15:222–226, 25:176
assessment of 25:166–167
Batterie Imagination et Perception (BIP) 25:166–167
and brain damage 25:163
of color 25:164–165
and extrastriate visual areas 25:164
imaginal neglect 25:165–166
of orthographic material 25:165
and striate cortex 25:163–164
and visual sociology 25:196
Visual impairment 24:9–10, 24:12–13
Visual information processing, in hemispatial neglect 10:768–769
Visual inheritance 25:201
Visual input processing, by temporal lobe 24:207–208
Visual landmarks 16:625
Visual literacy 25:170–171
Visual long-term memory 25:177–178
implicit 25:178
Visual media 15:42, 15:45
Visual memory
dissociation with other memories 25:179
and face perception 8:655
of nonhuman primates 25:179
psychology of 25:175–180
use of term 25:175
visual perception and 25:175–176
Visual memory codes 25:178–180
Visual methods 19:683–684, 25:181–183
critical 3:623
forms 25:181–182
empirical evidence 25:181
photographs in research process 25:182
photos as cultural narratives 25:181–182
organizational basis 25:182–183
problems 25:182
ethics 25:182
semiotic 25:181
technology 25:182
Visual neurons in ventral-stream areas 25:201–202
Visual neuroprostheses 16:715–717
Visual object agnosia 1:442
Visual Object and Space Perception Battery 19:251
Visual paired comparison (VPC) paradigm 3:465, 3:468–469
and early memory development 6:761–762
and memory development 15:148
Visual perception
infant gaze behaviors 12:30
neural basis of 25:184–190
neural encoding of visual scene 25:184
neural representation 25:185–186
structure of visual pathway 25:184–185
Visual processing 2:370–371
Visual processing areas 25:179–180
Visual processing streams 17:704
Visual rhetoric 20:651–652
Visual scanning of face 12:27
Visual search task 2:166–167
Visual selective attention 2:164–165
bottom-up and top-down biases 2:165
nonspatial feature-based selection 2:166
space-based selection 2:165–166
subject of selection 2:165–166
Visual self-presentation 24:148
Visual short-term memory 21:910, 25:176–177
deficits 21:913
implicit 25:177
neurofunctional architecture of 21:914–915
Visual sociology 2:357, 2:360, 25:191–196

current debates 25:191, 25:194–195
emergence and career of 25:193–194
exhibitionist and voyeuristic society of our time 25:192–193
future directions 25:194–195
iconic turn 25:191
iconosphere 25:191, 25:193–194, 25:196
photography 25:191–196
qualitative methods 25:193, 25:196
visual culture 25:193, 25:195
visual imagination 25:196
Visual space 25:197–200
curvature of 25:199
Gaussian 25:198
empirical results on geometrical structures of 25:198–199
Euclid's theory of perspective *versus* 25:197
future directions 25:199
geometry as a mathematical language 25:197–198
theories on the geometry of 25:198
use of term 25:197
Visual streams, dorsal and ventral 25:201–206
Visual studies 2:25
Visual surveillance 23:729–730
CCTV 23:729
computerization 23:730–731
drones 23:729–730
intelligent CCTV systems 23:729
privacy and contemporary smart 23:730–731
social/software sorting 23:731–732
Visual symbolism 25:173
Visual technologies 17:714
Visual thinking 3:163f
Visual turn 2:25, 25:207–209
art history 25:207–210, 25:212
biopolitics of visual 25:211–212
body 25:211–212
ethnicity 25:207–208
frames and contexts 25:210–211
gender 25:207–211
genealogies of 25:207–208
historical objectives 25:208
indiscipline 25:208–209
interdisciplinarity 25:208–209
linguistic turn 25:210
race 25:207–208
re-turning 25:212–213
transdisciplinarity 25:208–209
visuality and textuality 25:209–210
word/image opposition 25:209–211
Visual variables, in cartography 3:170
Visual word form area (VWFA) 16:593–598, 25:215–216, 16:594f–595f, 16:598f
Visual word recognition
neurocognitive psychology of 25:214–219
alphabetic principle 25:214, 25:217
applications 25:217
challenges 25:217–218
computational modeling 25:216–217
dyslexia 25:217
grain size theory 25:215
leaning visual words 25:214–215
meaning construction 25:216
methods and tasks 25:216–217
models 25:217
neurocognitive poetics 25:217–218
neuroimaging 25:215–217
phonological recoding 25:215
processing visual words 25:215–216
visual word form area 25:215–216
Visual working memory 15:141, 25:176–177
capacity limitation of 25:176
fractionation of 25:176–177
Visual/graphic variables in maps 24:262–263
Visuality, of tourism 24:470–471
Visualization 3:161–166
cartographic visualization process 3:162f
and cartography 3:162–163
of geospatial data 10:97–98
probability density estimation 19:31
process 3:161–162
in science and technology 25:220–225
chains of industry and credibility 25:223–225
historicizing and situating objectivity 25:222–223
image, object, and objectivity 25:221–225
images and models as harbingers of materialization 25:223
scientific artifacts 25:221–224
visual representation 25:220–221
Visualization environments 3:165
Visualization functions 3:164–165
basic display 3:164
multiple dynamically linked views 3:164–165
multi-scale 3:164
navigation and orientation 3:164
query data 3:164
reexpression 3:164
Visualization maps of international activity differences 24:384–387
2003–2010 ATUS diary data from different U.S. immigrant groups 24:387–390
MDS applied to Szalai's (1965) multinational data 24:387
updated 1998–2005 MTUS mappings 24:387
Visualization scale 21:3
Visualization techniques 21:322–323, 21:325
Visual–spatial competencies 12:323–324
Visual–Spatial Index (VSI) 12:284
Visual–spatial sketchpad 25:176
Visuoimitative apraxia 1:875
Visuomotor deficits after brain damage 25:202–203
Visuomotor function 17:508–509
Visuomotor processing, normal 25:204–205
dorsal and ventral streams, integration of 25:205–206
perception and action 25:205
Visuospatial attention 12:418–419
Visuospatial languages 21:930–932
Visuospatial maze 19:70
Visuospatial memory binding 15:141–142, 15:141f
Visuospatial perception 17:703
Visuospatial sketchpad, and mental imagery 15:222
Vital civil society (Tocqueville) 12:371
Vital events 5:739
age in completed years (ACY method) 5:739
age reached during the year (ARDY method) 5:739
Vital rates forecasts 18:594–595
difficulty with explanatory models 18:594
errors and debate on terminology 18:595
self-fulfilling or self-negating prophecies 18:595
Vital registration (VR) systems 1:153–156
in developing countries 1:154t, 1:155f
and mortality crossover 15:849–850
Vital statistics 5:740, 5:782, 6:171–172
real exposure rates 5:740
Vitalism, and instinct 12:177
Vitamin model 25:538
Vitoria, F. de, works, *De Jure belli Hispanorum in barbaros* 10:10
The Vitruvian Man 9:711–712
Vitruvian Woman 9:711–712
Vitruvius' theory of architecture 20:427
Vitthala temple 23:56f
Vivekananda, Swami 20:374
Vivid memories 9:254–256, 9:258
Vividness of Mental Imagery (VVIQ) 15:225
Vizzini, Calogero 14:422
VLA *see* Valued life activities (VLA)
VLPO *see* Ventrolateral preoptic nucleus (VLPO)
VMAT2 *see* Vesicular monoamine transporter 2 (VMAT2)
VMC *see* Vieth–Müller circles (VMC)
VMMC *see* Voluntary medical male circumcision (VMMC)
VMPC *see* Ventromedial prefrontal cortex (VMPC)
VMT *see* Vehicle-miles traveled (VMT)
VNO *see* Vomeronasal organ (VNO)
VNP *see* Vietnamese Nationalist Party (VNP)
VNPS *see* Voluntary nonprofit sector (VNPS)
VNS *see* Vagal nerve stimulation (VNS); Vagus nerve stimulation (VNS)
VNTRs *see* Variable numbers of tandem repeats (VNTRs)
VOC *see* 'Value of Children' (VOC)
Vocabulario degli Academici della Crusca (Vocabulary of the Members of the Academia della Crusca) 13:939
Vocabulary
lexicography 13:939–940
SES differences in 13:324–325
Vocabulary acquisition 25:226–231
EPAM-VOC 25:229–230
in later language development 25:230
pre-vocabulary competencies 25:226–227
segmentation 25:226–227

Vocabulary acquisition (*continued*)
speech perception 25:226
speech production 25:227
role of working memory 25:229–230
word meaning 25:227–229
attentional and associative learning 25:228
emergentist coalition model 25:229
latent semantic analysis 25:229
social pragmatic theories 25:228
syntactic bootstrapping 25:228–229
theories based on innate constraints 25:228
Vocabulary spurt 25:227
Vocal communication 2:691
Vocal development
developmental changes in song-control nuclei 2:693
neural signatures of 2:693–694
steroid hormones, role of 2:692–693
Vocal expressions 5:543–544, 7:485
Vocal learning
during development 2:691–696
neural signatures of 2:693–694
neural substrate for 2:692
in zebra finches 2:691
Vocal productions, of children, and intentionality 12:348–349
Vocal speech 10:506
Vocalizations 23:246
and language evolution 11:329–333
Vocational education 7:215–217, 25:232–238
apprenticeship 25:232, 25:234–237
importance of 25:233
international comparisons 25:233–234
labor market 25:232–237
linkages 25:234, 25:237
logics and linkages 25:234
models 25:233–235
pathways, characteristics of 25:235–237
societal approach 25:233–234
standardization 25:234–237
stratification 25:234, 25:236–237
transitions 25:232–237
young people 25:232–237
Vocational education and training (VET) 25:245–251
dual system *versus* school-based system 25:248
effects of postponing tracking of students into schools 25:248
estimation of returns 25:248–249
identification problems 25:246–247
informal, effectiveness of 25:249
investment of firms in 25:246
labor market 25:245–250
life-cycle perspective 25:249
policy implications 25:249–250
school-based 25:245–246
versus general-based education 25:247–248
Vocational interests 25:239–244
developmental psychology 25:240
educational psychologist's conception interest 25:240
expectancy-value theory 25:240–241
flow theory 25:241–242
foundations of vocational psychology 25:239–240
Holland model 25:239
intrinsic motivation 25:241–242
motivation 25:241–242
and performance 25:240
personnel selection 17:938
positive psychology 25:239, 25:241
social psychology 25:240–243
transition to adulthood 25:240–241
Vocational orientation, in school 1:862–863
Vocational Preference Inventory 25:239
Vocational rehabilitation (VR) system 6:474, 22:383
Vocational skills training 22:20–23
Vocational stage 13:737–738
Vocational training, apprenticeship system of 1:861–862
Vocations 21:610
behavioral expressions 21:610
VOD *see* Video-on-demand (VOD)
Voegelin, Eric 18:434–437
belief (Glaube) 18:436–437
hierarchy and ecclesia 18:435–436
metaphysics 18:436
works
Political Religions 18:435
The New Science of Politics 4:690
Voegelin Renaissance 18:438–439
Voelkerkunde 8:200–201
Vogel, Bruno, works, *Alf* 9:646
Voice 14:209–210
disorders of 9:266–273
expression of emotion 7:522–523
Hirschman on 10:895–897
Voice at work 25:698–699
working conditions 25:698–699
Voice practices 4:457–458
Voiced nonsibilant 15:827–828
Voiceless (animal rights organization, Australia) 1:705
Voiceless stops, three sets of 18:71–72
Voice-onset-time (VOT) 6:361–362, 13:302
Voice-over-Internet Protocol (VoIP) 6:896
VoIP *see* Voice-over-Internet Protocol (VoIP)
Volatile political order 18:388–389
Volatility
in migration 15:458–459
in political science 18:517–518
Volcanic sediments 7:693t
Volchek, Dmitrii 21:743
Volition 10:713–714
and personality 17:831–832
theories, of self-regulated learning 21:542–543, 21:545
Volitional psychology, perspective of procrastination 19:76
Volk 4:341, 22:888
Völkerpsychologie 21:963
Völkerwanderung (mass migration) 24:589
Völkisch 9:644–645, 22:889
Volkow, Nora 25:625
Volksgeist 9:668, 9:670, 11:381, 11:384, 16:356
Volksgemeinschaft 16:357, 22:889
Volksgerichtshof 16:357
Volksgeschichte (history of the people) 22:308
Volkskunde 8:200–201
Volkskunde
and anthropology 1:751–754
Volkslied 22:887–888
Volonté génerale 20:769–770
Voltage-dependent Ca^{2+} channels (VDCC) 12:710–714
Voltaire, François Marie Arouet de 1:765–766, 7:722, 10:163, 20:767–768, 22:697
works
Essai sur les moeurs 11:73
Philosophical Dictionary 7:584
Sentiment des citoyens 20:767–768
Voluntariness, informed consent to research 12:112
Voluntarism 19:322, 21:212
and psychology 19:413
Voluntaristic action, theory of 17:521–522
Voluntaristic psychology 5:262
Voluntaristic theory of action 1:76, 1:720–721
Voluntary active euthanasia 8:281, 8:283–284
Voluntary activities, and social capital 22:174–175, 22:178–179
Voluntary affirmative action 1:256–257
Voluntary agencies (VolAgs)
board of directors 25:261, 25:264–265
civil society organizations 25:261
club goods 25:264
factors affecting incidence and prevalence rates 25:262
failure 25:263
formation and life cycles of 25:262–263
formation of 25:262–263
goals and missions of 25:263–264
governance, management, and administration of 25:264–265
history of 25:261–262
impact and outcomes of 25:266
life cycle of 25:262–263
misconduct 25:263
nondistribution constraint 25:264
nonprofit organizations 25:261, 25:264–265
research and usable knowledge 25:266–267
resource attraction by 25:265–266
scope, magnitude, and types of 25:263
social agencies 25:261
volunteer programs 25:264, 25:266
Voluntary associations 22:32, 25:252–260
conventional *versus* deviant goals and means 25:255
geographic scope 25:253–254
government control 25:255–256

history of 25:253
impact of 25:259
individual *versus* organizational members 25:254
internal democracy 25:256
internal formalization of association structure 25:254–255
isolated *versus* collaborative associations 25:257
legal formalization and government recognition 25:254
liveliness and intensity of activity 25:257
monomorphic *versus* polymorphic associations 25:256–257
paid staff, dependence on 25:254
personal contact 25:257–258
prestige and power of 25:258
prevalence and determinants 25:253
purposive *versus* analytical types 25:253
research and theory about 25:252–253
restrictiveness of membership criteria 25:258
target of benefits 25:254
Voluntary childlessness 12:46–48
Voluntary expressions 8:668
components 8:668–669
lateral premotor cortex (LPMC) 8:668
primary motor cortex 8:668
Voluntary facial emotion expression (VfEE) 8:670–673
Voluntary health insurance model 10:643–644
Voluntary intoxication 5:222–223
Voluntary learning 7:240
Voluntary medical male circumcision (VMMC)
HIV infection 6:467–468
Voluntary minorities 1:731, 8:132–133
Voluntary nature of participation 20:517
Voluntary nonprofit sector (VNPS) 25:261, 25:263, 25:266
Voluntary organizations (VOs) 25:252, 25:261–267
elder care services 15:443
Voluntary principle 3:540
Voluntary *pro bono* legal services 12:922–923
Voluntary provision model (Olson) 17:208–209
Voluntary regulation, international development 8:60–61
Voluntary sector 3:366, 12:371, 22:576–578
Voluntary simplicity 8:25–26
Voluntary social work, Maori 14:505–506, 14:508
Voluntary transfers, intergenerational 8:826–828
Voluntary work 14:668
Volunteer programs 25:264, 25:266
Volunteer subject bias 8:573–574
Volunteered geographic information (VGI) 3:619–620, 12:93, 23:129, 23:134–135
Volunteered geography (VG) 10:174
Volunteering 19:216–217, 21:608–610, 22:577–579
definition 22:577–578
reasons for 22:578–579
social policy 22:579
volunteers 22:578
Volunteerism, psychology of 25:268–272
antecedents 25:269–270
community 25:268–272
consequences 25:271
experiences 25:270–271
health 25:271
identity 25:268–272
motivation 25:268–272
organizations 25:268–272
personality 25:268–270, 25:272
prosocial behavior 25:268–270, 25:272
recruitment 25:269, 25:271–272
satisfaction 25:268–271
service 25:268–272
theoretical and practical implications 25:271–272
Volya, Narodnaya 24:239
Volya naroda 23:22–23
Vomeronasal organ (VNO) 17:198
Von Hannover v. Germany (2005) 18:950–951
Von Hannover v. Germany (2012) 18:950–951
Von Neumann, John 7:793, 12:107, 16:582–585, 17:527
doctoral dissertation of 16:582
on education 16:582
and Institute for Advanced Studies 16:582
on mathematical knowledge 16:582–583
on methodological individualism 16:583–584
on minimax theorem 16:583
on 'neoclassical' economics 16:583–584
on nonconstant sum games 16:584
on paradox 16:584
professional career of 16:582
on prospect theory 19:262–265
works
Theory of Games and Economic Behavior 16:583
worst-case scenario 16:583–584
von Neumann–Morgenstern solution 4:869, 4:871
V-optimality 17:257
optimal experimental design 17:257
VOR *see* Vestibulo-ocular reflex (VOR)
Vory-v-zakone 14:429, 14:431
VOs *see* Voluntary organizations (VOs)
VOT *see* Value of time (VOT); Voice-onset-time (VOT)
Voter autonomy 18:333–334
Voter patterns, Markov models for 14:578, 14:578t
Voter preferences, and majority rule 14:446
Voter turnout 25:305–307
direct democracy 25:295–296
Voters
communicating with 3:67–68
free media coverage 3:67
paid media advertising 3:68
knowledge about economy 7:28
Voting 3:687–688, 18:331–333, 18:459–464, 24:591, 25:293–298
behavior in referendums 25:296–298
cognitive engagement 25:297
heuristics 25:297
mobilization 25:298
models of voting choice 25:297–298
rational choice 25:297
benefits of 25:309–310
class *see* Class voting
and class 3:749
class voting 25:303–304
cleavage voting 25:303–304
Columbia model 25:302
compulsory 25:309–310
corporate governance 4:934–935
costs of 25:309
cultural cleavages 25:305–306
economic approach 25:302–303
emotions, role of 25:306
gender gap 25:304–305
generational gap 25:305
genopolitics 25:307
get out the vote strategy 25:310
issue voting 25:302–306
Michigan model 25:302
nonvoting, implications of 25:310
paradox of 25:302–303
participation 25:306–307
and party identification 17:562–564
and political efficacy 7:301–302, 7:304–305
political representation 18:442–443
politics modeling 18:512–514
race and ethnicity 25:305
racial discrimination in 3:688
religious cleavage resilience 25:304
rights 24:591–592
socioeconomic status model 25:306
sociology of 25:302–308
spatial voting 25:302–303
strategic voting 25:303
turnout 25:309
factors affecting 25:309
voting choice 25:302–303
Voting decisions, local influences on 7:346–347
Voting rights, of soldiers 16:306
Voting Rights Act 1965 3:693, 6:530–531, 6:533, 11:911, 25:310
Voucher privatization 18:692–693
Voucher programs 21:71–72, 21:75–76
Vovelle, Michel, works, *Baroque Piety and Dechristianisation* 1:712
Vowel change 11:821
Vowels 18:65, 18:68, 23:715
Vox pop programs 15:60
Voxel-based lesion symptom analysis 2:95–96, 23:251
Voyeurism 3:620–621, 17:951
Voyeuristic society, and visual sociology 25:192–193
VP *see* Venture philanthropy (VP)

VPC *see* Visual paired comparison (VPC) paradigm
VP-internal subject hypothesis 14:179–180
VR *see* Vicarious resilience (VR); Virtual reality (VR); Vital registration (VR) systems; Vocational rehabilitation (VR) system
VRAM *see* Virtual radial arm maze (VRAM)
VRET *see* Virtual reality exposure therapy (VRET)
VRIN *see* Valuable, rare, and inimitable (VRIN) resources
Vroman, Adam Clark 18:83
VS *see* Ventral striatum (VS); Verbal sound (VS)
VSC *see* Verbal self-concept (VSC)
VSI *see* Visual–Spatial Index (VSI)
VSL *see* Value of statistical life (VSL)
VST *see* Variable sensitivity theory (VST)
VT *see* Vagal tone (VT)
VTA *see* Ventral tegmental area (VTA)
Vulnerabilities 4:355, 6:478–479, 8:80–81, 10:20–21, 12:545, 20:712, 21:137
  biomedical research ethics 8:79–81
  to challenges of daily life 4:355
  and climate 3:866, 3:869
  close relationship 22:564–566
  in community 4:355
  and coping 4:887–888
  disasters 10:547–551
  and domestic violence 6:625
  fear of crime
    theoretical differences on gender differences 8:855–856
  and hazard 6:478
  hazard and 4:355
  and human health 7:741–742
  ICSW 12:496, 12:498–499
  and sleep disorders 22:82–83
  to social geography 4:355
  treatment research and access 8:31–34
  and trust 24:670
Vulnerability models 17:917–918
  of low self-esteem 21:523
  of psychiatric disorders 22:381
Vulnerability–stress model of psychiatric disorders 22:381
Vulnerability–stress–competence model of psychiatric disorders 22:381
Vulnerable areas 5:147–148
Vulnerable populations 10:549–550
  environmental health and safety 7:740–742
VVIQ *see* Vividness of Mental Imagery (VVIQ)
VWFA *see* Verbal word form area (VWFA); Visual word form area (VWFA)
vWM *see* Verbal working memory (vWM)
Vygotsky, Lev Semenovich 7:242, 12:341, 16:679–680, 19:669–670, 21:363, 25:312–315, 25:322
  on academic concepts 25:313–314
  biography of 25:312–313
  on cognitive development theory 25:322
    child development 25:324
    educational applications 25:326–327
    theoretical impact of 25:327
  on crisis of psychological development 25:330, 25:332–333
  from dialogue to self-regulation 25:313
  on dynamic assessment method 12:252
  education and career of 25:329–330
  on egocentric and inner speech 25:313
  on everyday concepts 25:312–314
  on higher mental functions 25:312–313
  on human development theory 25:316–320
    cognitive apprenticeship 25:319
    community of learners 25:319
    community of practice 25:319
    cultural-historical theory 25:319–320
    developmentally appropriate curriculum 25:317–318
    dialogic inquiry 25:319
    general genetic law of development 25:317
    guided participation 25:318
    higher mental function 25:319
    instructional conversations 25:319
    leading activity 25:318
    legitimate peripheral participation 25:319
    mediation by cultural tools 25:317
    practice- and problem-based learning 25:319
    professional community building 25:319
    reciprocal teaching 25:319
    scaffolding 25:318
    self-regulation 25:317, 25:319
    semiotic mediation 25:316
    situated cognition 25:316
    sociocentrism 25:317–318
    sociocultural approach 25:317
    transformation of participation 25:317–318
    zone of proximal development 25:317–318
  intelligence assessment of 25:312
  on learning potential 25:314
  perspectives on coconstructivism 3:914
  prediction of mental growth 25:314
  on private speech 25:313
  on psychology 25:330–333
    development as cultural development 25:331–332
    development by changing interfunctional structure of consciousness 25:331
    instrumental, cultural, and historical approach 25:330–331
    mediation 25:331–333
    natural/higher psychological functions 25:331t
    Perezhivanie as unit of analysis 25:333
    signs 25:332
    social situation of development 25:332–333
    tools 25:330–333
  psychology of 2:884
  relevance for contemporary science 25:314
  scaffolding approach of 18:784
  on scientific concepts 25:312–314
  on self-regulated learning theory 21:542
  on sociocultural theory 25:329–334
  sociocultural theory of 8:731
  on sociogenetic law 25:313
  on Soviet education 20:826, 20:828–829
  theoretical system of 18:131
  on theory of cognitive development 4:54–55
    role of culture 4:55
    scaffolding 4:54–55
    zone of proximal development 4:54
  on theory of intelligence 12:307
  on word meaning and conceptual development 25:313–314
  works
    The Historical Meaning of the Crisis in Psychology 25:329
    Pedagogical Psychology 25:329
    The Psychology of Art 25:312, 25:329
    Thought and Language 25:312–313
  on zone of proximal development 25:314, 25:330
Vyras, P., on homosexuality 9:638

W

W matrix 23:102
WACC *see* Weighted average cost of capital (WACC)
Wach, Joachim 20:407
Wachirayan, Prince 2:891–892
Wacquant, Loic 7:416–417
WAD *see* Women and Development (WAD)
Wada, Juhn 13:32, 16:686–687, 16:687f
Wada test 2:807–808, 16:686
  intracarotid sodium amytal testing 12:689–691, 12:690f
Wage decomposition analysis 18:830
Wage decompositions 6:537–539
  components of 6:537–538
  nondiscriminatory wage structures 6:538–539
  residual decomposition 6:539
Wage determination 9:757–758
  welfare reforms 25:517
Wage differentials 9:760, 15:433, 15:435, 21:337, 25:335–339
  ethnicity 25:338–339
  gender 25:338–339
  job characteristics 25:337–338
  job location 25:337
  rents 25:339
  skill characteristics 25:336–337
Wage discrimination 18:830
Wage dispersion 12:102
Wage distributions 11:889
Wage earners 25:696
Wage elasticity 6:270
Wage floors 6:272–273
Wage formation 13:261–263
Wage function 11:889
Wage gap, female managers 8:885–886
Wage growth 19:135
Wage inequities 6:539–540, 25:339

comparable worth 6:540
econometric modeling approaches 6:540
Wage insurance, job displacement insurance 12:819, 12:825–826
Wage labor 1:540–541, 4:338, 25:671, 25:676
and capitalism 3:105, 3:108
see also Labor
Wage posting 21:336–337
competitive search 21:336–337
single offers, no prior information 21:336
wage differentials 21:337
Wage setting 6:272, 21:335–337
Wage structure 9:761–762
labor statistics 5:798
Wage subsidies 25:494–495
Wages 9:757–762, 25:335–336
differential and structure 25:335–339
effect of beauty on 2:383–389, 2:389t
and emigration from Europe 15:460
and industrial societies 11:865
inequality 6:765–772
and job evaluation 12:816
and marriage 8:829–830, 8:832–833
minimum *see* Minimum wages
and obesity 17:70–74
for older workers 17:186
promotions 19:135–139
risk-compensating 25:17
and trade unions 24:497–502
two-sided matching for 21:334–335
of women 9:63–66
working without 25:666
work–wage elasticities 13:197
Wage-setters model 11:888–889
Wagner, Adolph 22:698
Wagner, Erica L., on epistemic cultures 7:876
Wagner, Moritz 7:723
Wagner, Richard 16:830
Wagner tree 23:393
Wahnsinn (insanity) 13:129–130
WAIS *see* Wechsler Adult Intelligence Scale (WAIS)
WAIS-IV *see* Wechsler Adult Intelligence Scale; Fourth Edition (WAIS-IV)
WAIS-R *see* Wechsler Adult Intelligence Scale; Revised (WAIS-R)
Waissman, Friedrich 10:536
Wait and see attitude 5:510
Waiting list designs, 9:130
Waitkevicz, Joan 15:113
Waitt, G., on gay/lesbian tourism 9:641
Waiver of informed consent, human research participants 4:256
Wake
lexical consolidation during 13:923
neural system of 22:88–89
basal forebrain 22:89
brain stem 22:88
hypothalamus – histamine 22:89
hypothalamus – hypocretin (orexin) 22:89
norepinephrine 22:89
serotonin 22:88–89
Wakefield, Gilbert 14:457
Wakefield, Jerome, on diseases 17:4
Wakefulness 4:673–674, 22:80–81
Wakeman, Thaddeus 4:524
Wakenya Pamoja 25:845–846
Waksman, Selman 6:693–694
Walberg's productive giftedness model 7:275
multiple-factor approach 7:275
Wald, Abraham 14:814
Wald estimator 4:448–449
Wald test 2:593
in hypothesis testing 11:493
Wales
locally based community development in 4:381–382
probation in 19:45–47
Walikhanov, Shoqan 3:317
Walk 16:572
in centrality network 16:532
Walker, Alice 8:908
Walker, Mary Edwards 25:623
Walking, infant motor development 15:975–976
Wall Street
1929 crash 9:216
and finance anthropology 9:175
traders ethnography 5:536
Wall Street Journal 17:853, 23:459
Wall Street Rule 4:935, 4:941
Wallace, Alfred Russel 8:385, 20:287
evolutionary demography 8:406–407
Wallace, A.R. 8:385
Wallace, Henry Agard 9:625
Walled gardens 21:189–190
Wallerstein, Immanuel Maurice 16:285
biography of 25:340–345
on capitalism 25:341
contribution to social sciences 25:343
on division of labor 25:341
intellectual influence 25:343–345
legacy 25:343–345
on modernity 25:340
on modernization 25:341
works
Adam Smith in Beijing: Lineages of the 21st Century 25:342
The African Liberation Reader: Documents of the National Liberation Movements 25:342–343
Antisystemic Movements 25:342–343
Centrist Liberalism Triumphant 25:342
Creating and Transforming Households 25:342–343
European Universalism: The Rhetoric of Power 25:342–343
Geopolitics and Geoculture 25:342–343
The Long Twentieth Century: Money, Power & the Origins of Our Times 25:342
Mercantilism and the Consolidation of the European World-Economy 25:342
The Modern World-System: Capitalist Agriculture & the Origins of the European World-Economy in the Sixteenth Century 11:27, 25:340
Open the Social Sciences: Report of the Gulbelkian Commission on the Restructuring of the Social Sciences 25:342–343
Questioning Nineteenth Century Assumptions About Knowledge 25:342–343
Race, Nation, Class 25:342–343
The Second Era of Great Expansion of the Capitalist World-Economy, 1730s–1840s 25:342
Unthinking Social Science 25:342–343
The Wretched of the Earth 25:341
world systems theory 21:151–152
Wallis and Futuna island 17:158–159, 17:162
Walls
and gates, for security 9:613–614
landscapes of 9:614–615
Wal-Mart Stores, Inc., v. Betty Dukes (2011) 1:257
Walras, Léon 4:751–752, 21:304, 25:7
works, *Élements d'Économie Politique Pure* 25:9–10
Walrasian economics 18:359, 21:335
Walrasian system 24:738–739
Walshe, Sir Francis 13:393
Walter, Lucie 18:437–438
Walter E. Meyer Research Institute of Law 13:516, 13:518
Walton, R., works, *Behavioral Theory of Labor Negotiation* 16:431
Walton Family Foundation 7:267
Walzer, Michael 4:342, 22:392, 25:346–350
on civil society 25:349
on communitarianism 25:349
as critic 25:348–349
on democracy 25:349
education of 25:346
on ideological pluralism 25:346
on moral philosophy 25:347, 25:349
on multiculturalism 25:349
participation in civil rights movement 25:346
on pluralism 25:348–349
on political action 25:346
on political theory 25:346–349
works
Arguing About War 25:347–348
Company of Critics 25:348–349
Exodus and Revolution 25:347
Getting Out: Historical Perspectives on Leaving Iraq 25:347–348
In God's Shadow: Politics in the Hebrew Bible 25:347
Interpretation and Social Criticism 21:893–894, 25:347–349
Just and Unjust Wars 4:318–319, 25:347
Law, Politics, and Morality 25:347
The New Republic 25:346
Obligations 4:318–319, 25:347
On Toleration 25:346–347, 25:349
Pluralism, Justice and Equality 25:349
Political Action 25:346

Walzer, Michael (*continued*)
Politics and Passion: Toward a More Egalitarian Liberalism 25:349
Radical Principles 25:346
Regicide and Revolution 25:347
Revolution of the Saints 25:347
Spheres of Justice. A Defense of Pluralism and Equality 4:309, 4:319, 25:347–348
The State of Righteousness 25:346–347
Thick and Thin 25:349
Toward a Global Civil Society 25:349
Wanderjahre 10:750
Wandersman, A., works, *Empowerment Evaluation Principles in Practice* 7:581–582
Wandervogel movement 9:644–646
Wang, Fuzhi 11:50–51
Wang, Guowei 11:52
Wang, Jing
on Chinese art districts 5:599
on Chinese cultural industries 5:597–598
on Chinese political censorship 5:600
WANGO *see* World Association of NGOs (WANGO)
Wannsee Conference 16:260
Wanted total fertility rate (WTFR) 8:787
Wanting 25:351–356
binge eating and food addiction 25:354
consumer behaviors 25:355
dopamine 25:351–354
drug addiction, recovery, and relapse 25:351–354
incentive salience and utility 25:352–354
gambling 25:354–355
and hope 11:199
liking 25:353–355
motivation 25:351–355
pleasure 25:352–354
reward 25:351–355
sensitization 25:351–355
Wanton, Joseph 1:610
Waorani people, peace/nonviolence 17:626–627, 17:629
WAPOR *see* World Association for Public Opinion Research (WAPOR)
War against drugs 5:78
War casualty 17:651
War crimes 11:368–369, 13:821
and ethnic cleansing 8:83–84, 8:86–88
genocide 10:15, 10:20
War crimes tribunals 25:382
contemporary origins 25:382
first and second world wars 25:382–383
genocide 25:382–387
human rights 25:385, 25:387
humanitarian law 25:382, 25:384–385
hybrid tribunals 25:385–386
International Criminal Court 25:386–387
international criminal law 25:382–387
International Criminal Tribunal for Rwanda 25:383–385
International Criminal Tribunal for the Former Yugoslavia 25:383–385
Nuremburg 25:382–383
transitional justice 25:387
War economies 17:649
War experience, and prosocial behavior 19:241
War lords 17:649
The War of the Worlds (radio broadcast) 22:586
War of Transnistria 25:423
War on Cancer
biomedicine 2:666
War on crime 5:214
War on drugs 5:178, 6:682
social justice 22:395
War on Poverty (legislation) 3:625, 14:823, 18:733–737, 23:495–496
War on terror 18:264, 24:240, 25:377, 25:379–380
War potential of nations 16:253–256
future directions 16:254
historical perspectives 16:253
twentieth-century strategy 16:253–254
War Powers Resolution 4:652
War propaganda 25:395
establishing the enemy/attaching guilt 25:398
Gulf War and 25:398–400
Lasswell's conceptualization of function of 25:396
in the New World Order 25:397–398
potential of 25:400
seven common propaganda devices 25:397
theories of 25:396
tools and techniques of 25:395–396
World War I and 25:395–396
War stories, MENA 16:363–364
Warband 4:608–610, 4:612
Warburg, Aby 2:24–25, 20:426
Ward, Colin 1:686
Ward, Lester 4:524
Ward, Mary 3:234–235
Warde, Alan, on cultural participation 16:556
Wardle, Huon 1:641
Wardley, Jane 2:146–147
Warfare 4:608–609, 11:265, 21:370
and chiefdoms 3:380–381
chiefdoms in 3:380–381
defensive technologies in Classical period 25:414
general staff system 25:417
in history 1:697–699
and nongovernmental organizations 16:858
Warfare history 25:414–419
archers 25:415
armor 25:414–415
atomic age 25:418
cavalry 25:415, 25:417
classical period 25:414–415
organization 25:414–415
purposes 25:415
technology 25:414
conscription 25:417–418
contemporary period (1918–2000) 25:418–419
organization 25:418
purposes 25:418–419
technology 25:418
crusades 25:415–416
English longbow 25:415
Gunpowder Age 25:416–417
organization 25:416
purposes 25:416–417
technology 25:416
horses 25:415–416
imperialism 25:416–418
long nineteenth century (1776–1918) 25:417–418
organization 25:417
purposes 25:417–418
technology 25:417
Middle Ages 25:415–416
organization 25:415–416
purposes 25:416
technology 25:415
muskets 25:416–417
nationalism 25:417–419
regional conflicts 25:418–419
Warfare-based economic systems 23:374
Warlpiri 14:222
Warm parenting 2:471, 2:473
Warming, Eugenius 7:792
Warmth stereotype 21:680–681
Warner, Lloyd 2:401–402, 4:362
Warner, Michael, works, *Fear of a Queer Planet* 19:754
Warner, W. Lloyd
on social network analysis 16:527–528
works
The Social Systems of American Ethnic Groups 2:83–84
Yankee City 17:22–23
Warner Bros. 9:156
Warners 9:156–157
Warrant, relativism about 20:235
Warranted assertibility (realism) 20:22
Warranties 17:172–177
age of becoming old 17:174–175
aging 17:172–176
asking right questions 17:173
biodemography 17:173–175
biological warranty period 17:173, 17:176
centenarians 17:175
Darwinian fitness 17:173
and expiration dates 17:173
extreme longevity 17:175–176
geriatrics 17:173–175
life span 17:172–173, 17:175–177
longevity genes 17:175–176
mortality acceleration 17:174–175
potential life span 17:172–173
quantifying warranty period 17:173–174
reproductive biology 17:173
terminology and premise 17:172
unnatural experiment 17:172–173
Warranting (online information) 4:507–508
War-related violence 4:608–609, 4:613
Warren, Earl (Chief Justice) 8:583

Warren, Ronald L., on organizational fields 17:375–376
Warring 4:621
Warring States Period 13:800
Warrington Recognition Memory Test 19:251
Warrior gene 9:964
Warrior societies 1:320–321
Warriorhood 4:608
Warriors 4:608–613
  burials 4:610
Wars 1:982–983, 10:1–2, 11:392–394, 15:518–523, 20:628, 23:362–363, 23:366, 25:347–348, 25:357
  among nation-states 25:420
    armed conflict 25:420, 25:423
    causal-explanatory approaches to interstate war in IR 25:424–426
    causality and interstate war 25:423–426
    comprehensive accounts 25:426
    explanatory strategies and interstate war 25:423–424
    historical sociology 25:420–422, 25:426
    interstate war 25:420–423
    interstateness as reciprocity and similarity 25:423
    military 25:421–423, 25:425–426
    as organized violence 25:422–423
    security dilemma 25:423–425
    state at war 25:421–422
    territory 25:420–426
  anthropology of 4:614–619, 25:377–381
    11 September (9/11) 25:379
    antiquity 4:615–616
    changing nature 25:378–379
    child soldiering 25:379
    Clausewitz on 25:380
    collateral damage 25:377, 25:379
    colonialism 25:378–380
    counterinsurgency 25:377
    cross-cultural frequency 4:615
    definition 25:377–378
    development 25:378
    drones 25:379–381
    feud 4:614–615, 4:618
    future 25:380–381
    genocide 25:377–380
    guerrilla warfare 25:379
    homicide 4:614–618
    Human Terrain System 25:380–381
    International Criminal Court 25:379
    mercenary 25:378
    military–industrial complex 25:379
    nonwarring societies 4:615, 4:617–618
    origins 4:616–617
    overseas contingency operations 25:379, 25:381
    peace systems 4:615, 4:617
    prehistory 4:617
    reasons 4:617–618
    social complexity 4:616–617
    theories 25:380
    torture 25:377–380
    twentieth century 25:379
    unmanned aerial vehicles 25:379–381
    virtual war 25:380–381
    war on terror 25:377, 25:379–380
  apologists for 17:627–628
  and appeasement 1:826–830
  archaeology of 4:600–613
    archaic states and aristocratic armies 4:603–604
    chiefdoms and chiefly militias 4:603
    egalitarian hunter-gatherers and self-redress homicides 4:600–601
    foragers 4:600–602
    fortifications 4:612
    iconographic representations in art and rituals 4:612–613
    nation-states and industrial armed forces 4:604
    skeletal trauma 4:612
    tribes, feuds, and kin-militias 4:601–602
    weapon combinations and technologies 4:611–612
    weaponry 4:609–610
    weaponry in ritualized contexts 4:610–611
  and armed conflicts 17:654–655
    enemy images and cognitive biases 17:654–655
    social identity, societal narratives, and ideology 17:655
  and balance of power 2:345–348
  causes of
    capitalism and democracy 25:402–403
    decision making 25:403
    interstate system 25:401–402
  changing nature of 25:389
  and child protection 3:426
  child soldiers 3:430–438
  and childhood stress 23:571
  civil war *see* Civil war
  Clausewitz on 25:407–409
  collaboration of state and science during 21:214–215
  and collective violence 14:703
  concept of 25:407–409
  conflict statistics 5:836–840
  consequences of
    revolt and revolution 25:403–404
    state formation and bureaucratization 25:403
  defense economics 6:1–6
  definition 4:614–615, 4:621
  and democracy 6:80, 6:82–83, 25:357–362
  and Eastern Europe Studies 6:805–808
  effects on children 25:389–394
    adolescents 25:389–392
    aggression 25:391
    armed conflict 25:389–391, 25:393
    child soldiers 25:392–393
    context 25:389–390
    identifying risk and protective factors 25:391
    intergroup process 25:392
    intervention and prevention programs 25:393
    mental health and psychopathology 25:390–391
    resilience 25:389–390, 25:393
    role of family 25:391–392
    social ecology 25:390
    social identity 25:392
    youth as victims, perpetrators, and agents of social change 25:393
  effects on democracy 25:361
  Elias on 25:406–407
  environment 15:496
  ethnic cleansing 8:83–90
  and fenocide 10:1–5
  French Revolution 9:409
  future of 16:307–308
  genocidal violence in 10:2–3
  genocide, history of 10:7–9, 10:12
  genocide in 10:1–2
  geography of 25:371–372
    embodiment of 25:372–373
    and nature 25:373
    and peace 25:374
    place of war 25:373–374
  and glory 4:608, 4:610
  hegemonic stability theory 25:401–402
  Hobbes on 4:607, 4:613
  in Korea 2:71–72, 2:74–75
  martial arts 4:612–613
  and mass trauma 14:747–750
  material culture 4:608–610, 4:612
  and military anthropology 15:483–487
  military history 15:506–510
  and nation-state 16:304–308
    casualties 16:307–308
    citizen armies and soldiers' rights 16:306–307
    creating national armies 16:305–306
    death 16:304–308
    demise of conscription 16:307
    memorials 16:306–307
    nationalism 16:305, 16:308
    taxation 16:304–305
    technology 16:304, 16:307–308
  before nation-states 16:304–305
  new wars 15:521–522
  and peace 4:607, 4:610, 4:612–613
  peace movements 17:643–647
  post-Cold War challenges 15:521–523
  of religion 13:216
  Schmitt on 25:409
  and small states 22:108–112
  social change 4:607–608, 4:611–612
  social work stress/trauma 18:707–709
  sociology of 25:405–413
    defining 25:405–407
    Mann on 25:411–412
    social change and social reproduction 25:409–412
    survival units 25:406–407
    Tilly on 25:410–411
  sovereign's right to 10:10
  spirit possession 23:283
  and statehood 23:347–351
  territorial conflict 24:214–220
  total war 15:519–520
  urban 24:953–957

Wars (*continued*)
victory of democratic control over 15:520–521
and violence 25:116, 25:118–120
and violence, roots and causes 25:412
warband 4:608–610, 4:612
war-related violence 4:608–609, 4:613
warrior burials 4:610
warriors 4:608–613
weapon offerings 4:611
women's security 25:638–639
and youth 4:633, 4:636
Wars, and nationalism 25:363–370
climate change 25:368
cultural homogenization 25:367
ethnic conflict 25:367–369
French Revolution and French Revolutionary wars 25:363–364
future challenges 25:368
industrial development and empire building 25:364–367
modernity 25:363, 25:366
Napoleonic campaigns 25:363–364
conscription 25:364
fatherland, barrack and school 25:364
westernization of world 25:364
nationalist homogenization 25:367–368
fascism, war and nationalism 25:367–368
nationalism, war and genocide 25:367
postwar era and globalization 25:368
neoliberal globalization 25:368
patriotism 25:363–369
Pax Britannica 25:364–367
egalitarian impetus 25:366
Franco-Prussian militarism 25:365
industry 25:366
pupils into soldiers 25:365–366
World War I 25:366–367
replaceability 25:366, 25:369
state-building 25:366, 25:369
totalitarianism 25:366–369
Wars of the third kind 3:714
Warsaw Pact 1:561, 18:700, 24:451
Wartime research effort 4:279
Wartime rumor 20:789
Warwick-Edinburgh Mental Wellbeing Scale 15:188
WAS *see* Western Asia (WAS)
Waseda University 12:795–796
Washington, Booker T. 6:691
Washington Consensus 7:262, 16:462, 18:691, 18:724
Decalogue of 24:847
Washington State Census Board 1:840–841
Washout period, in crossover designs 5:349–351
WASI-II *see* Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II)
Wason Selection Task 8:402
Waste management 7:776
Waste trade
Native Americans for 7:760
transboundary 7:760–761
Wasted lives *see* Disposable youth
WAT *see* Words As Social Tools (WAT)
Water
and climate change 3:848–849
cultural representations and interpretations 25:438
distribution and flows 25:437–438
as environmental resource 25:438–440
and health 25:428–436
social and behavioral sciences on 25:433–434
and social inequality 25:431–433
and power 25:440–441
privatization of 23:48
psychological and behavioral aspects 25:438
resources 25:437–442
sanitation 25:433–434
terminology 20:71
Water affected dating methods 3:567
Water and hazards research, White's contributions to 25:558–559
Water balance, regulation by hypothalamus 11:487–488
Water Efficiency, Conservation, and Management 7:774
Water flotation technique, in paleoethnobotany 17:456
A Water Policy for the American People (water policy) 25:561
A Water Policy for the American People (water policy) 25:561
Water pollution 25:444, 25:447
Water purification 23:824
Water resources, in Africa 1:296
Water scarcity 25:443
and conflict 25:443–444
Water wars 25:443–447
cooperation 25:444–447
global climate change and uncertainties 25:445–446
growing water scarcity 25:443
managing shared water systems 25:444
pollution 25:444, 25:447
river water sharing agreements 25:444–445
water resource governance and managing water scarcity 25:446–447
water scarcity and conflict 25:443–444
Waterborne disease, and natural disasters 16:316t
Watercraft, Australia and New Guinea 2:230
Waterfronts 24:839–840
Waterlogged seeds 7:694–695
Watermark 2:213
Water-related diseases 25:429–431, 25:433–434
transmission routes 25:430t
Water-saving behavior 7:700
Waterschap 15:678
Watersheds in, criminal procedure 5:227–229
Water–user associations 12:725
Watson, Alan 13:565
Watson, Henry 14:811
Watson, James 11:320, 14:243
Watson, John Broadus 7:238, 12:11–12, 13:391–392, 19:413, 25:448–449
on behavior therapy 3:880
on behavioral pharmacology 25:449
on behaviorism 25:448–449
on conditioned reflexes 25:449
contributions
to natural sciences 25:448
to predict and control behavior 25:448–449
on ethology 25:449
on habit 10:477
on homing patterns 25:449
on introspection 12:703
on learning theory 13:705
on psychology 25:448–449
works
Behaviorism 25:449
The Behaviorist manifesto 25:448
The Journal of Animal Behavior 25:449
The Journal of Experimental Psychology 25:449
The Psychic development of the white rat correlated with the growth of its nervous system 25:449
The Psychological Review 25:449
Psychology as the behaviorist views it 25:448
Watt, Harry, works, *North Sea* 6:621
Watt, James 1:723–724
Watts, Duncan J.
on market information regimes 12:591
on small world structures 5:467–468
Watts Riots 22:506, 25:821–822
Wave theory, language contact 13:300–301
Wave-based time metric for sexual development 14:330–333
bivariate growth with an exogenous outcome 14:331–333
bivariate model for number of sex partners and drinking 14:331
ratings of days drunk 14:330
total number of sexual partners 14:330
Wavelets, Fourier transform 9:378
The Way (non-religious movement) 16:806
Way of Great Peace movement 20:369
Wayfaring 18:354
Ways of being
at home 11:155–157
of indigenous peoples 11:804–805
WB *see* World Bank (WB)
WBC *see* World Baseball Classic (WBC)
WBP *see Wissenschaftliche Beirat Psychotherapie* (WBP)
WBR *see* Whole Body Reaching synergies (WBR)
WCED *see* World Commission on Environment and Development (WCED)
WCGs *see* Working Group on Girls (WGGs)
WCP *see* Women with childbearing potential (WCP)
WCRP *see* World Climate Research Programme (WCRP)
WCST *see* Wisconsin Card Sorting Task (WCST)

Weak coalitional–worth monotonicity 4:874
Weak complementarity 7:728
Weak governments 17:649
Weak identification of diffusion process 6:413
Weak instruments 12:208
Weak order 2:340, 14:863, 17:525
Weak power networks 15:669
Weak program 22:960–961
Weak records 16:898, 16:900
Weak social constructivism 22:229–231
Weak states 2:571, 8:126–127
Weak sustainability 23:824
  versus strong sustainability 6:852–853
Weak ties theory 16:529–531
Weakening a behavior 7:240
Weakness of the will 12:337
Wealth
  databases 5:744–747
  distribution 25:450–455
  health and 10:633–635
    inequalities in 10:634–635
  household 25:450–452, 25:451t
    composition changes 25:453–454
    consumer debt 25:453–454
    debt 25:451, 25:453–454
    estate tax data 25:452–453
    income capitalization techniques 25:453
    inequality, recent trends in 25:454–455, 25:454t
    measurement 25:452–453
    mortgage debt 25:451, 25:454
    portfolio composition 25:454
    savings 25:450–452, 25:454–455
    survey data 25:453
    wealth tax data 25:453
  and political economy 18:356–360
  poverty 18:728–732
  and rank 18:714–716
  and retirement 20:605
  and well-being 25:450
  see also Intergenerational property transmission
Wealth flows theory 8:411, 9:35–39
  cultural norms 9:37
  demographic transition 9:35, 9:38
  education 9:35–36
  empirical applications 9:36–37
  evolutionary anthropology 9:35, 9:37
  evolutionary approaches 9:37
  family relations 9:37
  human behavioral ecology 9:35
  mass schooling 9:35–36
  role of family in fertility decision-making 9:37–38
  Westernization 9:35–36
Wealth inequality 20:295, 20:295t
Wealth ownership 20:295
Wealth tax 23:341, 23:341t
Wealth tax data 25:453
Wealth-maximizing rules 12:935–936
Weapon focus 8:643
  eyewitness testimony 8:642–643
Weapon offerings 4:611
Weaponry 4:609–610
  combinations and technologies 4:611–612
  in ritualized contexts 4:610–611
Weapons
  and aggression 1:378, 1:378f
  and homicide 11:166–167
Weapons technology 24:763–764
Weather, climate and history 3:869–871
Weather diaries 3:867
Weather events, multilevel interventions during 10:389–390
Weather prediction task (WPT) 16:845
Weathering hypothesis 6:755
We-attitude 21:880
Weaver, Warren 24:764
Web 10:61, 23:114
  as corpus 4:949–950
  digital networks 25:173
  online dating 17:214
  pornography 17:217–220
Web 2.0 3:162, 3:619–621, 11:548, 11:550, 12:600
  Computer Assisted Instruction (CAI) 4:493–495
Web 1.0 text-based chat rooms 11:549–550
Web analytics 5:710
Web of life 22:264
Web sites 2:212
  terrorist 24:228
Web-based communication 24:146
Web-based resources 7:230
Web-based support services
  for adult education 1:151
Webcam sites 17:219
Weber, Alfred 5:498–499, 7:378, 9:458, 18:435, 23:199, 25:463
Weber, Ernst Heinrich 10:756, 14:810, 19:412, 19:470
Weber, Eugene Joseph 6:812
  work, *Peasants into Frenchmen* 6:812
Weber, Marianne 25:463
Weber, Max 12:280, 25:456–461
  1903–1907 methodological essays 25:457–458
  analysis of bureaucracy 2:908
  analysis of charisma 3:351
  on alienation 1:540
  on applied social research 1:850–852
  approach to history 11:75–76
  on asceticism 2:62–63
  on authority 25:459, 25:461
    charisma 2:271–272, 3:354–356, 25:458–461
    rational-legal authority 2:271
    traditional authority 2:271
  and Bielefeld School 2:553–555
  on bureaucracy 2:908–909, 2:911
  on bureaucracy model 2:913–914
  on capitalism 3:106–107, 25:456–458
  on civil service 3:696–698
  on classes 3:747
  on cognition 5:577
  concept of ideal types 7:147
  concept of status cultures 5:376
  concept of verstehen 12:658
  on constant 19:470–472
  on cosmopolitanism 5:45
  on cultural participation 5:447
  on democracy 25:459–460
  and Durkheim 22:992
  early works of 25:457
  on economic action 5:535
  on economy and culture 5:535
  on Enlightenment 7:664–667
  on *Geist* 9:666, 9:669–670
  on habit 10:476–478
  on historical sociology 11:27
  on human sciences 11:383
  idea of modernization 15:701
  on idealization 11:518–519
  on interest and rationality 12:389
  on interplay between science and politics 21:167–171
  iron cage of 21:530
  on journalism 12:852
  on knowledge societies 13:108
  on law 19:474, 24:308
  on legitimacy 13:823, 25:459–460
  on liberalism 25:460
  on lifestyle 14:113
  Max Weber Gesamtausgabe 25:462–463
  on McDonaldization 14:828–829
  on meaning 16:561
  on Middle East society 16:381–382
  on modernity 15:683
  on nationalism 16:290–291
  on police 18:251
  on political sociology 18:473
  as political thinker 25:459–460
  on positivistic sociology 18:625–626
  on power 18:765–766
  on practical reasoning 18:778
  on qualitative methods 19:676, 19:678
  on rationalization and disenchantment 11:118
  on rationalization of religion 15:577
  on realism 25:460–461
  on reforms 2:145
  on religion 20:271, 20:313–314, 20:345–346, 20:359
  on religious organizations 20:407
  on reification 20:203–205
  on salvation religions 25:458–459
  on scientific inquiry 22:992
  on secularization 25:458
  on social action 25:459
  on social inequality 22:331–332, 22:338–340, 22:342
  on social structure 22:721
  on social/society 22:828
  on social thought 14:497, 25:462–467
    American appropriations 25:463–464
    capitalism 25:463–466
    consolidation of work 25:462–463
    current trend 25:466–467
    economic sociology 25:466
    émigré perspectives 25:464–465
    homecoming and renewal 25:466
    postwar Weberians 25:465–466
    protestant ethic 25:463

Weber, Max (*continued*)
social theory 25:465–467
sociology of religion 25:463, 25:465–466
on sociology 10:786, 15:701, 22:996–999
on sociology of culture 4:361
on status 23:435
on symbolic boundaries 23:850–851
on theorization of interests 12:392
on theory of bureaucratic administration 14:474
on theory of rationality 22:999
on trials 24:635–636
on typology of legal decision-making 24:635t
on values 25:456–458, 25:460–461
works
The Agrarian Sociology of Ancient Civilizations 7:585
Agrarverhältnisse im Altertum 7:585
Ancient Judaism 3:351, 19:197
Collected Essays in the Sociology of Religion 25:462–463
Economic Ethics of World Religions 25:458–459
Economy and Society 1:525, 5:535, 13:159–160, 21:261, 25:458–459
The Ethical Economy of World Religions 10:484
General Economic History 25:463
Politik als Beruf (The Profession and Vocation of Politics) 21:167
Postulate of Objectivity 1:496–497
The Protestant Ethic and the Spirit of Capitalism 2:62–63, 5:498, 5:535, 10:484, 12:645–646, 25:458, 25:463
reception of 25:460–461
Science as a Vocation 14:497
Wirtschaft und Gesellschaft 3:353
Wissenschaft als Beruf (Science as a Calling) 21:167
on world systems theory 25:751
Weber, Paul 9:647
Weber fraction 17:47, 21:604
Weber function 14:800
Weber solution 4:870, 4:872
Weberian problem structure 19:50–51, 19:50f
revised 19:51
Weberian vein 22:998–999
WebGrid 5 program 4:724–725
Webliography, evidence-based medicine 8:330
Weblogs 16:784–785
Web-of-Science 21:322
Webster, Frank 12:83
Wechselwirkung 21:966–968
Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II) 12:286, 16:723–724
Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV) 12:284, 16:723–724
Wechsler Adult Intelligence Scale, Revised (WAIS-R)
Reliable Digit Span 16:723
Wechsler Adult Intelligence Scale (WAIS) 8:552, 12:307, 12:310–311
Wechsler factor structure 12:284
Wechsler Individual Achievement Test, Third Edition (WIAT-III) 16:725
Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V) 12:284
Wechsler Intelligence Scale for Children (WISC) 3:795, 6:344, 12:307, 12:310–311
WISC-IV 16:723–724
WISC-R 2:849–850
WISC-V 12:284
Wechsler Memory Scale, Fourth Edition (WMS-IV) 16:724
Wechsler Memory Scale-Revised (WMS-R) 2:96
Wechsler Nonverbal Scale of Ability (WNV) 12:286
Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV) 12:284, 16:724
Wechsler tests 12:286, 12:950
Wechsler–Bellevue Intelligence Scale, historical background 3:879
WED *see* Women; Environment; and Development (WED)
Wedeen, Lisa, works, *The Ambiguities of Domination* 16:383
Wedell, G., works, *Broadcasting in the Third World* 19:879–880
Wedge-drivers 20:789
Wedgwood, Emma 5:676
We-disease 11:595
WEDO *see* Women's Environment and Development Organization (WEDO)
Weeks, Jeffrey
on LGBTQ families 21:757
on sexual citizenship 21:731
works
Africa Misunderstood: Or Whatever Happened to the Rural-Urban Gap? 24:969–970
Wehler, Hans-Ulrich 10:871
works, *Deutsche Gesellschaftsgeschichte* 15:703
Wehrmacht 21:885
Weiberfastnacht 3:146
Weibull distribution 6:572, 6:577, 14:56, 23:772–773
Weick, Karl, on sensemaking theory 17:390
Weight
children, media use and 15:39
as erotic capital 8:6
and ordered relational structures 17:297
religion and 20:266
Weight adjustments, nonresponse error 16:922
Weight gain 21:22–23
schizophrenia 21:23
Weight perception 25:471–472
Weight regulation 25:468
appetitive behavior as link between genetic and environmental influence 25:470–471
body weight perception and weight control 25:471–472
cognitive behavior therapy 25:472
dietary restraint story 25:471
environmental factors 25:468–469
food environment 25:468–469
food responsiveness 25:470, 25:473
future 25:472–473
individual variation 25:470
satiety sensitivity 25:470
self-monitoring 25:472–473
self-regulation 25:473
supporting behavior change 25:472–473
Weighted average cost of capital (WACC) 4:928, 4:930
Weighted displays 10:356
Weighted exogenous sample maximum likelihood (WESML) method 17:105
Weighted least-squares estimators 20:735–736
Weighted linear combination (WLC) 23:138
Weighted models, linear hypothesis 14:148–149
Weighted utility ranking models 2:551
Weighting 19:144
Weights vector 4:877
Weimar Germany 12:279
Weimar Republic, homosexual rights movement divisions during 9:646–648
Weine, Robert, works, *The Cabinet of Dr. Caligari* 9:167
Weiner, Annette 1:769
Weiner, Norbert 14:814
Weingast 19:536
Weininger, Elliot
on cultural capital 5:376, 5:379
works
Geschlecht und Charakter, Weininger 25:611
We-intentions 12:840–841
WEIRD *see* Western Industrialized Educated Rich and Democratic nations (WEIRD)
Weisburd and Eck taxonomy 6:264
Weisenburg, Theodore 16:685
Weismann's theory of the continuity 8:241
Weiss, Margot, works, *Techniques of Pleasure: BDSM and the Circuits of Sexuality* 19:754–755
Weissenberger-Eibl, Marion, on epistemic cultures 7:878
Weissenberger-Eibl, on epistemic cultures 7:878
WEIU *see* Women's Educational and Industrial Union (WEIU)
Weizsäcker, V. von 19:314
Weldon, W.F.R. 9:571, 17:660
Welfare 10:263–264, 10:738–740, 25:475–480
and criminal justice 5:211

domestic welfare in design of IP rules 13:521–522
global welfare concerns 13:522–523
and gross domestic product 5:754–755
human 13:523
intellectual background and current knowledge 25:476–477
law and 13:520–521
and libertarianism 14:40–41
and moral reasoning 15:803–805
origins 25:475–476
philosophical aspects 25:504–510
beyond welfarism 25:510
as capabilities 25:508
combining insights 25:508
conceptions 25:506–508
freedom 25:507–509
game theory 25:504
happiness 25:504–506, 25:508
hedonism 25:505–507
justice 25:507, 25:509
nonstandard welfarism 25:509
as pleasure 25:506
pluralism 25:510
as preferences 25:506–507
as resources 25:507–508
rights 25:507, 25:509
social welfare functional 25:508–510
standard welfarist philosophies 25:508–509
utilitarianism 25:505–509
and utility 25:504–506
policies 25:476–477, 25:479
programs 18:725
promotion of 22:782
racial stereotypes 25:478
rational choice explanation 19:898–899
retrenchment, conservative critiques and 25:478–479
rights 25:478
and social capital 22:176
and social work 14:665–668
US context 25:477–478
and work 25:491–496
classic open-ended programs 25:492–493, 25:492f
disincentives 25:492–493
dynamic and human capital effects 25:495–496
earning and wage subsidies 25:494–495, 25:494f
labor supply 25:496
negative taper rate 25:494–495
social assistance programs in OECD countries 25:491t
welfare program features to reduce disincentives 25:493–494
of workers 24:498
Welfare associations 12:372–373
Welfare capitalism 3:4–5
Welfare dynamics, panel surveys for 17:465–466
Welfare economics 7:87, 7:103, 25:12, 25:497–503
Arrow's impossibility theorem 25:500
Coase on 25:501–502
efficiency 25:497–501
equality 25:498–500
external effects 25:497, 25:499, 25:501
first and second theorems in 25:497–500
happiness 25:497, 25:500–502
issues of environmental protection 25:501–502
Pareto optimality 25:497–500
Pigovian tax 25:501–502
preference *versus* welfare 25:497
second- and third best theories 25:498–499
social choice 25:500–501
Welfare geography 22:303–304
Welfare liberalism 9:399–400
Dworkin on 9:399–400
Rawls on 9:399
Welfare magnet hypothesis 12:438
Welfare policies 1:325, 25:511–514
conceptualizing welfare policy models 25:511–512
explaining policy divergence 25:512
influence on educational outcomes 25:483–485
for poverty 18:730–731
Welfare Reform Act 1996 24:916–917
Welfare reforms 18:725–726, 18:736–739, 25:515–520
active distribution policy 25:519–520
active labor market policy 25:518–519
employment and distribution 25:517–518
incentives and insurance 25:516–517
income inequality 25:519
means-testing 25:515, 25:518
tagging/targeting 25:518–519
targeting 25:518–519
Welfare regimes 25:511, 25:516
fertility 9:23–24
and health 22:279
refugees 22:770
typology in developed countries 25:486t
Welfare retrenchment 25:521–527
'blame avoidance' strategy 25:521
'credit claiming' strategies 25:524
European Union (EU) 25:522, 25:525
interest groups 25:525
measurement of change 25:522–523
need for change 25:521–522
demographic and socioeconomic changes 25:522
globalization and European integration 25:522
Organisation for Economic Cooperation and Development (OECD) 25:521–526, 25:523t
path dependence 25:521
and political institutions 25:525
political parties 25:524–526
and power relations 25:525–526
public expenditure 25:522–523, 25:525
public policy 25:525
reform dynamics through ideational and institutional change 25:524–525
resilience through policy feedback 25:524
social policy 25:521–522, 25:524–526
studying changes in social rights 25:523–524
Welfare rights, of children 3:487
Welfare state resilience through policy feedback 25:524
Welfare states 1:324, 3:97–98, 13:592–593, 22:33–34, 22:151–152, 23:4, 23:365–366, 24:573, 25:476–477, 25:479, 25:515–516, 25:519, 25:528–533
clientelism and punishment 19:640–641
in Sweden 19:645f
consequences of 25:513
conservative 24:573
educational policy traditions in 25:481–482, 25:481t
and immigration policy 11:665–666
international migration effects on 12:517–518
liberal 24:573
and libertarianism 14:41
mixed 18:324–325
and motherhood 15:884–885
Myrdal's theory 16:160–161
and new liberalism 18:302–305
pension policy 17:692–694
people and society 22:889–891
and poverty 18:731
and poverty law 18:733–738
regimes 3:659
in developed countries 25:481t, 25:489t
and social democracy 22:242–243, 24:573
and social insurance 22:363
southern European countries 24:573
typology of regimes in developed countries 25:489t
application 25:487–488
Welfare states, history 25:528–530
comparative studies 25:532
historiography 25:531–532
gender-conscious 25:531–532
history of 25:530–531
goods and services 25:530–531
regulation 25:531
transfer payments 25:530
history since 1945 25:531
mixed economy 25:532
origin and use 25:528–530
British historiography 25:528–529
German historiography 25:529–530
Swedish model 25:530
Welfare-based nation state 25:677
Welfare-oriented capitalism 21:178
Welfarism 13:573, 13:575
nonstandard 25:509
standard welfarist philosophies 25:508–509
Well-being 1:329–330, 3:51, 5:547, 6:25, 10:153, 10:521, 17:337–338, 17:763–764, 21:489
and authoritarianism 2:265
and broaden-and-build theory 2:864–867
childhood 3:455
in collectivism/individualism 4:208
conceptulization 25:534

Well-being (*continued*)
consequences of working longer for 17:187
and economics 7:87, 7:89–90
and ethnic identity 8:115–116
and global development 10:150–153
and health 1:280
human 5:775, 5:777, 5:779
and identity 11:527
of immigrant families 11:637–638
influences of emotion regulation on 7:455
and leisure 13:853
and LGBT aging 13:866
effect of discrimination 13:866
mental health 13:866
physical health 13:866
and mass imprisonment 14:697, 14:699
mental 15:187–188
interventions to promote 15:190
and meritocracy 15:241–242
and motives 15:951–955
at older ages 20:599
and optimism/pessimism 17:265
promotion of 22:782
psychological 2:247–248
religion and 20:265
research strategies for 1:330
and rheumatoid arthritis 20:655
and rural geography 20:799
and self-efficacy 21:505, 21:508
and self-esteem 21:523
significance of motivation in 24:186–187
skepticism for 7:89–90
social context 1:330
social ecology of 17:757f
social problems 12:545–546
social resources 1:330
social status 1:330
and startle 23:335–336
and successful aging 23:669–671
and sustainability 23:802–805
sustainable development 23:819
types 1:329
and utilitarianism 7:89
and values 25:30
and wealth 25:450
and wisdom 25:592–593
in workplace, psychology of 25:534–536
and World Health Organization 10:610
Well-being and burnout
conceptualizations of 25:537
see also Job-related well-being and burnout
Well-being freedom, in politics 11:559
Wellcome Trust 12:866–867, 20:513
Well-functioning markets, prediction and valuation of policies in 5:50
measuring benefits 5:50
measuring costs 5:50
Wellgradedness 4:8
Wellman, Barry
on cultural consumption 16:557
works
Portraits of White Racism 19:835
Wellman, D.T., works, *Portraits of White Racism* 19:835
Wellness models of human nature 17:911
Wells, G.C., works, *The Protestant Ethic and the 'Spirit' of Capitalism* 21:389
Wells, H.G. 8:220, 22:586
Wells, Ida B. 8:906
Well-stated questions 19:53–54
Welsh Museum 1:755
Weltanschauung 5:43, 7:392–393, 9:670, 11:94, 11:381, 15:305, 21:308, 25:755–756
Weltbild 11:59
Weltgesellschaft 24:589
Weltkenntnis 5:42–43, 5:46
Wenger, E., works, *Situated Learning* 7:145–146
Wenshi tongyi 11:50
Wepfer, Johannes Jakob 16:681
works, *Observationes medico-practicae, de affectionibus capitis internis et externis* 16:681
Werner, C. 3:318
Werner, Heinz 19:414–415
Wernicke, Karl 16:682–683, 18:47, 19:412, 16:683f
Wernicke encephalopathy 1:511
Wernicke's aphasia 13:398, 14:181, 1:820t
Wernicke's language area 1:819, 2:809, 9:268, 13:398, 15:585, 16:586, 16:740, 23:245–246, 23:251
Wernicke's region in humans 10:506
Wertfreiheit, axiological neutrality 16:757
Wertheimer, Max 4:51, 13:123–124, 19:414
Werther effect 23:686–687
Werturteilsstreit 1:497
WES *see* Work Environment Scale (WES)
Wesensschau 11:442
Wesleyan Methodism 10:492
WESML *see* Weighted exogenous sample maximum likelihood (WESML) method
West
doctor–patient interaction in 6:613–619
informal sector 12:57
morality of 15:807–808
nostalgia in 17:35–36
research ethics in 20:509
West, Paige, works, *Conservation is Our Government Now: The Politics of Ecology in Papua New Guinea* 18:353
West Africa
colonization 1:286
sociocultural overviews 1:300–304
agrarian base 1:300
agrarian social solidarity 1:301
civil wars 1:302–303
coast and desert edge 1:301–302
creolization 1:302
impact of external trade on social values 1:302
imperialism 1:303
land grabs 1:303
lineage society 1:300–301
linguistic diversity 1:300
peace-keeping 1:303
people 1:300
postcolonial social trends 1:303
savannas 1:300–302
social consequences of slave trade and colonial rule 1:302–303
youth 1:301–303
West African Economic and Monetary Union 12:574
West African languages
serial verb constructions 21:645–647
West End study 4:364
West German schooling system 21:464
West Germany
Bielefeld School of History 2:553–558
Cold War 2:553
West Indian American Day Parade/West Indian Day Carnival 3:147–148
West Indian slavery 22:55
West Nile fever 18:177–178
West *versus* rest secularization 21:392–393
Westat, Rockville 23:750
West–east axis 20:746
Westermarck, Edward 14:453
works, *The History of Human Marriage* 12:178
Westermarck hypothesis 11:726, 15:815
Western Aphasia Battery 3:793–794
Western Asia (WAS)
adult mortality in 1:171, 1:175, 1:179
and body history 2:728
cultural policy regimes in
see also Cultural policies
erotic fair 18:536
history of emotions 7:515
impact of Cold War in 4:123–124
integration and multiculturalism 12:235–242
Jewish Diaspora 12:803–806
literacy in 14:237–238
Muslims in 12:237–239
psychiatric hospitals
development 3:874
racial discrimination in 6:543–544
sex between negotiation, risk, and virtualization 18:537–538
sexual crisis 18:534
sexual culture in 18:533–539
commercial eroticization 18:533–535
conservative family values 18:533–535
state interventionism 18:533–535
sexual revolution 18:535–537
working classes 25:694–695
Western biomedicine 10:634
Western Church, role in world culture's formative period 10:227
Western civilization 11:73, 11:75
period of modernization 15:711
Western culture-bound syndromes 5:622
Western cultures
in legal personalities 13:754
original research on intrinsic motivation in 12:696
transition to adulthood in 24:576
Western democratic thought 13:747
Western developed world, transition to adulthood in 24:572
Western ethnocentrism 11:110

Western eugenics 8:250
Western Europe (WEU)
adult mortality in 1:171, 1:174–175, 1:179
history of 13:219
Western feminism 8:906
Western Front, establishment 9:213
Western history, construction of disability in 6:453–454
Western idea of philosophy 7:255
Western Industrialized Educated Rich and Democratic (WEIRD) nations 22:530
Western knowledge 21:150
Western literate societies 14:241
Western Malayo-Polynesian languages 23:890–891
Western Marxism 4:220, 11:40–41, 18:766
and advertising 1:215–216
Western medicine 8:449, 15:118–125
Western modernity 4:181, 24:583
Western music 16:156–157
Western nations, threat to East Asia 16:276
Western penal development, historical context of 19:655–657
Western political theory 8:926
Western psychotherapy, and mental health disorders 5:572–573
Western science 20:284, 23:798
and local cultural knowledge, reconciling 9:4
Western Snow Melt Application System Verification Test 20:422
Western social science works, translations of 6:800
Western social work models 12:557
Western societies
family in 8:809
partnership formation and dissolution in 17:545–552
successful aging in 23:667–672
Western sociology, Meiji Japan's engagement of 24:583–584
Western theory 6:747–748
Western tradition 11:72
Western triumphalism 10:163
Western Union 12:519
Western universalism 25:358–359
Western values 21:791–792
Western way of knowing 21:150
Western worldview 14:505–507
Westernization 10:228, 24:842
eating disorder 6:823, 6:826
of social work 22:793
and wealth flows 9:35–36
of world 25:364
Westernized oriental capital, socio-ecological model of 3:599f
Western-like aspirations 5:586
Western-style globalization 5:585–586
West-Germanic languages 11:821
WestlawNext 5:804–805, 5:804f
critical services 5:805–807
search capabilities on 5:807–808
Weston, Kath
on families of choice 21:755–756
on North American community 17:23
works
Families We Choose 19:754, 21:755
Westphalian conception of law 13:757–758
WestVar (software) 23:750
Wet-rice agriculture 12:794
Wetterling Act 1994, 1996 20:434
Wetterling Improvements Act 1997 20:434
WEU *see* Western Europe (WEU)
WFS *see* World Fertility Survey (WFS)
WGI *see* World Giving Index (WGI)
WGNRR *see* Women's Global Network on Reproductive Rights (WGNRR)
WGS *see* Whole-genome sequencing (WGS)
WH *see* Whorfian hypothesis (WH)
Whakawhanaungatanga 14:509–510
Whaling prehistory 1:924–925
Whanau 14:505–506, 14:508–509
What Is Happening In This Class? (WIHIC) questionnaire 3:827, 3:829
What *versus* where stream processing 25:156
What works, in applied research (criminology and criminal justice) 5:129
What Works Clearinghouse (WWC) 23:215
What-where-when memory 3:466
Wh-copying structures 9:208
Wheatstone bridge 16:542
Wheeler, Etta 22:390
Wheelwright, Joseph 7:934
When Religion Meets New Media 20:276
Wherry, Frederick
on cultural wealth of nations 5:539
on prices 5:538–539
Which-N question 14:180
Whig history 3:61–62, 14:216–218
Whig theory 11:412
Wh-imperative, English 13:290
Whiskers
experience-dependent synaptic delivery of AMPA receptors 16:490
whisker–barrel function
attenuation of 16:491
Wh-Island Condition (WIC) 16:4
Whitaker, Harry A. 16:679, 16:679f
works
Handbook of Neurolinguistics 16:679
Neurolinguistics 16:678
Perspectives in Neurolinguistics and Psycholinguistics 16:679
White, Andrew Dickson 20:284
works, *A History of the Warfare of Science with Theology in Christendom* 20:284
White, Cynthia, on consumption 16:555
White, Gilbert F. 25:558–563
academic administration and resumption of resource policy studies at Haverford College 25:560–561
on arid land management 25:560–561
contributions of
to environmental geography 25:559–562
to natural resources planning and policy in Washington, DC 25:560
to water and hazards research 25:558–559
early influences and studies from Chicago to Wyoming 25:559–560
honors and accomplishments of 25:562
integrating natural resources and natural hazards research from the University of Chicago 25:561
integrating resources, hazards, and environmental management at the University of Colorado 25:561–562
on natural hazards research 25:558–559, 25:561
on pragmatism 25:558–560
on sustainability 25:559, 25:561
on water resources fields 25:558, 25:560–561
works
Natural History of Selbourne 7:748
White, Harrison C.
on consumption 16:555
on historical explanation of 11:14–15
on markets 5:469
on network analysis 5:557–558
on social network analysis 16:529–530
on social networks 5:466, 5:558
works
Identity and Control 5:558, 9:363–364, 16:562
Metahistory 11:14–16
White, Hayden 11:41, 11:98
White, Leslie Alvin 1:769, 7:621–622
works, *The Evolution of Culture* 20:386
White, Lynn, Jr. 20:288, 24:141
White, Richard
works, *The Middle Ground: Indians, Empires, and Republics in the Great Lakes Region, 1650–1815* 2:872
White, Robert 7:934–935
White, Steven 5:301
White American household 3:204
White Fathers 15:614–615
White goods 9:725
White helper 10:759–760
White Label Campaign in 1899 8:26
White Lotus Society 20:369
White March in Belgium (1996) 19:654
White matter 1:653–654
developmental neuroscience 16:600
White matter pathways 16:587
language development after injury to 16:589
White noise 24:335
White power 18:414
White privilege 25:569–572
intersectional queer theory 19:768
racism as 19:868
White race 19:801
White racial frame 19:845, 19:847–848, 19:850
White racial identity 19:835
White slavery 17:432, 24:596
White supremacy 19:852
colonialism 25:569–573
critical white studies 25:571, 25:573
definition 25:571
imperialism 25:572–573

White supremacy (*continued*)
indigenous peoples 25:571, 25:573
inequality 25:569–572
marginalization 25:569–573
oppression 25:569–570
people of color 25:569–573
racialization 25:570–571, 25:573
racism 25:569–573
and social work 25:569–574
professional identity 25:572–573
white privilege 25:569–572
White thesis 20:288
White Western feminism 8:909
White-collar crimes 5:165, 5:176, 5:190, 5:203–207, 25:551–557
applicability of low self-control theory 25:555–556
control of 5:204
corporate crime 25:551–552, 25:555–556
definition 25:551
future of 5:206
gender 25:555
methodological pitfalls in study of 5:205–206
mortgage fraud, financial crimes, and the global financial crisis 25:552
neutralizations and accounts 25:554
new social science and 5:206
offenses 5:204
opportunities and choices 25:552–553
organizational complexity of 5:204
organizational crime 25:554–556
policing of 5:204
prevention and control 25:555
prosecution and punishment 5:204–205
psychology and individual differences 25:553–554
subject of inquiry 5:203
use of term 5:203
victims of 5:204
White-collar workers 15:535
history of 25:564–568
class 25:566–567
class-consciousness 25:567
definition by functional criteria 25:564–565
gender 25:566–567
industrialization 25:565–566
industry 25:565–566
middle classes 25:566–567
national differences and peculiarities 25:566
organizational behavior 25:566
origin and growth 25:565
professions 25:564–566
social theory 25:566–567
status and privilege 25:565–566
working classes 25:566–567
Whitehall study 15:851
Whitehead, Alfred North 16:583, 20:284, 24:141
works
Encyclopedia of Anthropology 7:590
Science and the Modern World 24:141
Whitehead, Harriet, works, *Sexual Meanings* 8:899
Whitehead Institute 11:313
White-imposed racism 19:847
Whiteness
colonialism 25:569–573
critical white studies 25:571, 25:573
definition 25:569–571
imperialism 25:572–573
indigenous peoples 25:571, 25:573
inequality 25:569–572
marginalization 25:569–573
oppression 25:569–570
people of color 25:569–573
and postcolonial geography 18:645
racialization 25:570–571, 25:573
racism 25:569–573
and social work 25:569–574
professional identity 25:572–573
studies 25:571–572
white privilege 25:569–572
Whiting, B.B., works, *Six Cultures Series* 7:145
Whittaker, Sir Edmund Taylor 9:624
Whitten, David E., works, *Encyclopedia of Anthropology* 7:590
Whitten effect (mice) 17:198
WhK *see* Wissenschaftlich humanitäres Komitee (WhK)
Wh-movement 9:208, 16:2
WHO *see* World Health Organization (WHO
Whole, population as 18:579–580
Whole Body Reaching (WBR) synergies 15:960–961
Whole Candidate Score, for grit 10:399
Whole object assumption, in vocabulary acquisition 25:228
Whole objects 13:68
Whole-genome sequencing (WGS) 2:452–453, 6:440, 8:717–718
Whole-genome shotgun method 11:313
Whole-of-government approach, public sector organizations 19:601
Wholesale elimination 16:329
Wholesale reconstruction 16:329
Whole-word method *see* Look-and-say method
Whorf, Benjamin Lee 6:245, 20:903, 25:575–578
on consciousness 25:576, 25:578
contributions to cognitive psychology 25:575
and d'Olivet 25:577
on Hopi 25:575–577
Indic studies of 25:577–578
on interdisciplinary research 25:578
and Jung 25:576
on language and ethnicity. 13:252–253
on language universals 25:576, 25:578
on linguistic relativity 25:575–578
on linguistics/semantics 14:228–229
and Maya 25:575–576
on mysticism 25:577–578
and Sapir 25:575–576
Sapir–Whorf hypothesis 1:767, 25:575
on scientific paradigm 25:576
on theosophy 25:577–578
on transformational generative grammar 25:578
and Yale University 25:575
Whorfian hypothesis (WH) *see* Sapir–Whorf hypothesis
WH/WD *see* Withholding/withdrawing supports (WH/WD)
Why questions in psychology 25:269
Whyte, R.O., works, *The Rape of the Earth* 7:752
Whyte, William F. 1:755, 2:401–403
works, *Street Corner Society* 4:362, 22:102
Whyte, William H. 17:532
works
The Exploding Metropolis 24:934
The Organization Man 2:916, 15:679
The Social Life of Small Urban Spaces 24:866
WIAT-III *see* Wechsler Individual Achievement Test, Third Edition (WIAT-III)
WIC *see* Wh-Island Condition (WIC)
Wicked problems 5:257, 12:150, 22:355, 22:357
Wickham, Chris, on agrarian political economy 1:457
Wicksell, Knut 3:92, 13:16, 19:113
Wicksteed, P., works, *An Essay on the Co-ordination of the Laws of Distribution* 25:11
WID *see* Women in Development (WID)
Wide Range Achievement Test 3:796
Wide reflective equilibrium 20:78
Widernatürliche Unzucht (unnatural fornication) 9:645
Widow
bereavement 2:531–533
immolation 20:374
Widowhood
and marital quality 14:512, 14:515–516
method, indirect estimation 6:145
Wiebe, R.H, works, *The Segmented Society* 1:615
Wiener, Norbert, works, *Cybernetics or Control and Communication in the Animal and the Machine* 22:912
Wiener diffusion process 6:396
Wieringa, Saskia, works, *Female Desire: Same-sex Relations and Transgender Practices across Cultures* 8:902
Wife abuse 6:633–634
Wife beating 6:633–634
Wi-Fi-enabled devices 15:630
Wigmorean approach 13:790
WIHIC *see* What Is Happening In This Class? (WIHIC) questionnaire
Wikan, Unni, works, *In Honor of Fadime* 15:425
Wikipedia 7:877, 12:600, 21:252, 23:380
Wikis 6:898
Wilberforce, William 22:442–443
Wilce, James 1:237
Wilcock, John William, works, *Laws Relating to the Medical Profession* 8:53

Wildavsky, Aaron 18:204
Wilderness gentrification 10:47
Wilderness Society 19:556
Wildfires 6:480
Wilhelm II, Kaiser 9:645
Wilhelmian Germany 13:428, 13:430
Will
  collective 11:817–818
  and practical reasoning 18:778
Willer, D.
  on elementary theory of social relations 8:484
William D.P. Bliss, works, *Encyclopedia of Social Reform* 7:585
William I, Count of Normandy 4:603–604
William III, Frederick 25:365
William of Ockham 11:816, 19:408–409
Williams, Bernard 4:682
Williams, Christian, on Stonewall riots 21:745
Williams, G.C. 1:459
  works, *Adaptation and Natural Selection* 22:903
Williams, K. 10:732–733
Williams, Patricia 5:267
Williams, Raymond 1:236, 10:744, 24:163
  works
    Television: Technology and Cultural Form 24:163
    Keywords 5:617
Williams square 5:352
Williams syndrome (WS) 5:357, 11:390, 25:189, 25:579–583
  brain development 25:580, 25:582
  and developmental psychology 5:357
  genetic disorders 25:579, 25:582
  genetic profile 25:579
  infancy 25:579–582
  joint attention in children with 12:847
  neuroconstructivism 25:579
  number development 25:582
  phenotypic profile 25:579–582
    cognitive development 25:581–582
    language development 25:580–581
    social development 25:581
Williams thesis 22:56–57
Williamson, Judith, works, *Decoding Advertisements* 1:216
Williamson, O.E., works, *Markets and Hierarchies* 19:894
Willingness to pay (WTP) 5:47
  cost–benefit analysis 5:47–48, 5:50–53
  of ethical consumers 8:27
  and value of life 25:14–19
Willis, Paul, on educational anthropology 7:145, 7:147
Willis, S.L., stage theory of cognition 1:436–437
Willmot, Peter, works, *Family and Kinship in East London* 4:363–364
Willowbrook hepatitis study 13:561
Willpower, and motivation 15:938
Wills and succession
  among nations in twentieth century 19:175
  throughout history and 19:173–174
Wilson, D.S. 22:893
Wilson, Edward O. 19:416, 22:893
  works
    Consilience: The Unity of Knowledge 22:905
    Genes, Mind, and Culture 22:903–904
    Human Nature 22:905
    Sociobiology: The New Synthesis 8:408, 20:389, 22:899
Wilson, J.Q., works, *Varieties of Police Behavior* 18:254
Wilson, Peter 1:640–641
Wilson, Richard A., works, *Culture and Rights* 11:363
Wilson, Robert A., works, *The MIT Encyclopedia of the Cognitive Sciences* 7:590
Wilson, William Julius 23:157, 25:584–589
  on affirmative action 25:586–587
  on African Americans 25:584–588
  analyses and social policy 25:587–588
  early life and career 25:584–585
  on ghetto 25:586–587
  on inner city 25:586–588
  legacy of 25:588–589
  on poverty 25:584–585, 25:587, 25:589
  on race–class debate 25:585–586
  recent works of 25:588
  on social institutions 25:586–587
  on social mobility 25:586–587
  on social structure 25:585, 25:588
  works
    The Truly Disadvantaged 18:720–721, 25:586
    When Work Disappears 25:586–587
Wilson, Woodrow 18:322, 22:888
Win (utility)- maximization leagues 23:319
Winant, Howard, works, *Racial Formation in the Twentieth Century* 19:796
Winch, Peter, works, *The Idea of a Social Science* 13:765
Winckelmann, Johannes, works, *History of Ancient Art* 3:761, 11:94–95
Windblown sediments 7:693t
Windelband, Wilhelm 11:818, 19:391
Windigo 5:627
Windmills 7:616
Windscale accident 24:143–144
Windshield ethnography 11:397
Windward Islands 1:638
Wine tasting, and perceptual learning 10:131
Winichakul, Thongchai, works, *Siam Mapped: A History of the Geo-body of a Nation* 11:86
Winick, Charles, works, *Dictionary of Anthropology* 7:589–590
WINIR *see* World Interdisciplinary Network for Institutional Research (WINIR)
Winne and Hadwin's model of self-regulated learning 21:537
Winnebago trickster cycle 24:649–652
Winner, Langdon, technological determinism 24:109–111
Winner's curse 2:207
Winner-take-all, in elections 18:865
Winnicott, Donald 19:503–505
Winning and losing in litigation 14:271–272
Winslow, C.-E.A. 20:269
Winsorizing (outliers) 23:390
Winstanley, Gerrard 7:654
Win-stay-lose-shift 5:893–894
Winstedt, Richard, works, *Tawarikh Melayu* (*Malay History*) 11:84
Winter, Ed 1:741
Winters concept of organizational routines 8:430–431
WIPO *see* World Intellectual Property Organization (WIPO)
Wipro (IT company) 3:47–48
Wired city 22:114
Wireless devices
  and psychotherapy 19:380–381
  telephone 16:776, 24:157, 24:159–161
Wireless radio 4:333–334
Wireless telegraphy 12:490, 12:492
Wireline, telephone 24:159, 24:161
Wirth, Louis 4:363
  works, *The Ghetto* 7:561
Wirtschaftsbürgertum 2:786–787
Wirtschaftspsychologische Forschungsstelle 13:637–639
WISC *see* Wechsler Intelligence Scale for Children (WISC)
Wisconsin Card Sorting Failure-to-Maintain Set 16:723
Wisconsin Card Sorting Test (WCST) 4:33–35, 6:726, 8:494
Wisconsin Law School 13:517
Wisconsin Longitudinal Study (WLS) 5:698, 14:346, 14:513, 23:783
WISC-V *see* Wechsler Intelligence Scale for Children; Fifth Edition (WISC-V)
Wisdom 1:430, 10:522–524, 25:590
  and age 25:595
  and appeasement 1:827
  current developments and next steps 25:596
    new developments in psychological wisdom research 25:596
    societal relevance 25:596
  fostering wisdom 25:595
  lay theories of 25:590
  measurement 25:592–593
    performance 25:592–593
    self-report scales 25:592
  philosophical aspects 25:598–603
    Aristotle's 25:599, 25:601
    in Bible 25:599–600
    elements of modern epistemology 25:601–602
    epistemic contemporariness 25:602
    epistemological contours 25:600–601
    esotericism 25:598, 25:602
    Hellenistic and Roman philosophy 25:599
    hermeneutics, poststructuralism, and analytic philosophy 25:601
    impossibility of any meta-position 25:601
    knowledge 25:598–600

Wisdom (*continued*)
limits of justification 25:601–602
Middle Ages 25:599–600
modern philosophy's practical intentions 25:600
morality 25:600
moral–practical side 25:600
open flank of theory 25:599
philosophical theory as highest praxis 25:599
philosophy as wisdom proper 25:599
philosophy in league with science 25:600
philosophy's claim to represent true praxis and wisdom 25:599
Plato's 25:598–599
pre-Platonic understanding of philosophy and wisdom 25:598
protest against wisdom, Nietzsche 25:600
return of sage as practical expert 25:599
scientific pluralism 25:601
shifting from practical to theoretical philosophy 25:600–601
strategic implications 25:598–599
technical–practical side 25:600
theoretical reason with characteristics of practical reason 25:601
uncertainty of fundaments 25:601
Wittgenstein's 25:601–602
psychological correlates of 25:593–595
psychological definitions of 25:590–592
Berlin wisdom model 25:591–592
grouping definitions of wisdom 25:592
three-dimensional wisdom model 25:590–591
wisdom as self-distanced reasoning 25:592
wisdom as self-transcendence 25:591
psychology of 25:590–597
theories of development of wisdom 25:595
Wisdom of crowds 6:66
Wiseman, Frederick
works
High School 6:623
The Titicut Follies 6:622
Wishing, and hope 11:199
Wishram Chinook language 20:899
Wissenschaft 11:127–128, 14:642–643, 24:763
Wissenschaft als Beruf (Science as a Calling) 21:167
Wissenschaftlich humanitäres Komitee (WhK) 9:644–647
Wissenschaftliche Beirat Psychotherapie (WBP) 11:462
Wissenshaft 13:228–229
Wissenssoziologie 13:113
Wit 11:417
Witch doctor 25:606–607
Witchcraft 7:330, 8:290, 25:604–610
cultural trials 25:604–605
healing 25:605–608
historians and anthropologists 25:604–605
magic 25:605–608
and magic 14:436–438
and modernity 25:605–607
morality 25:606–608
paradigms and critiques 25:607
religion 25:604–605, 25:608–609
and religion 20:310
social control 25:605, 25:607
sorcery 25:604, 25:608
spirit possession 25:608
state formation 25:604–606
terms and their history 25:608–609
themes 25:607–608
traditional healers 25:606–608
witch doctor 25:606–607
Withdrawal 16:493–494, 17:335
due to shame 21:875
motivation 2:328–330
from society 9:858
Withdrawal classes 2:582
Withdrawal design 21:986
Withdrawn behaviors, of children 22:946
Withholding/withdrawing supports (WH/WD) 7:599–601
in India 7:599b
Within-case method
of case studies 3:208–210, 3:212
Within-group cooperation 10:416
Within-group variance, population heterogeneity 17:733
Within-individual retrieval-induced forgetting 4:191
Within-network research 1:54
Within-person emotions, in workplace 7:507–508
and affective events theory (AET) 7:507
and job performance 7:508
Within-school allocation 21:57
Within-series designs 21:986
combining within- and between-series designs 21:987
Within-sleep stimulations, contribution of 22:67–68
Within-subject designs 8:538–539, 9:128, 24:536–537
balanced crossover designs 8:538–539
crossover designs 8:538
Latin squares 8:539
Within-unit correlation, in crossover designs 5:349–353
Within-unit information, in crossover designs 5:351–353
Withitness 24:58
Witkiewicz, Stanislaw Ignacy 14:453
Witness 8:585
Witness error 25:788
Witnessing, in interviews 12:674
Witt, Jamil Mahuad 23:47
Witten, T., on LGBT aging 13:864
Wittfogel, Karl A.
on hydraulic civilizations theory 25:440–441
on irrigation theory 12:724–725
Wittgenstein, Ludwig 25:611–615
Tractatus theory of 25:612, 25:614
Wittgenstein, Ludwig Josef Johan 1:747, 5:501–502, 8:576–577, 12:763, 18:625–626, 18:795, 19:409, 21:309
impact of 25:614–615
on language games 10:786–787
life of 25:611–612
on linguistic theory 13:793
on meaning 14:841–843, 14:845, 14:847
on mind–body dualism 15:547
philosophical work of 25:612–614
language game 25:612–613
logical atomism 25:612–613
logical truths 25:612
naturalistic metaphysics 25:612
ontological theory 25:612
psychological states 25:613
social anti-individualism 25:613–614
verbal expressive behavior 25:613–614
on relativism 20:235
Tractatus ideas 25:612, 25:614
on wisdom 25:601–602
works
Beiheft 25:614
Concept of Mind 25:614–615
Geschlecht und Charakter 25:611
Logisch-Philosophische Abhandlung 25:611
Philosophical Investigations 2:134, 6:590, 15:292, 25:612–613
Tractatus Logico-Philosophicus 25:611
Wittich, Claus 25:465
WLC *see* Weighted linear combination (WLC)
WLS *see* Wisconsin Longitudinal Study (WLS)
WM *see* Working memory (WM)
WMA *see* World Medical Association (WMA)
WMG *see* Working memory general (WMG)
WMI *see* Working Memory Index (WMI)
WMS *see* Working memory concept as a specific model (WMS)
WMS-IV *see* Wechsler Memory Scale; Fourth Edition (WMS-IV)
WMS-R *see* Wechsler Memory Scale-Revised (WMS-R)
WNV *see* Wechsler Nonverbal Scale of Ability (WNV)
Wohlfahrtsstaat 25:528
Wold, Herman 7:9–10
Wolf, Eric 1:769
on brokers 2:870–871
works
Europe and the People Without History 1:746–747, 18:351, 25:342
Peasant Wars of the Twentieth Century 1:456–457
Wolf, Margery, works, *A Thrice-Told Tale* 8:901
Wolf, Susan 2:317
Wolfe, Elizabeth 21:743
Wolfe, P., on settler colonialism in North America 17:25
Wolfenden Report 25:75–76
Wolfenstein, Martha 14:838
Wolff, Christian 13:778–779
Wolff, Peter 16:476
Wolfram, Stephen 1:368

Wolf-Wendel, L., on gay athletes 9:658
Wollstonecraft, Mary 2:515, 7:117
Wolman, Benjamin, works, *International Encyclopedia of Psychiatry, Psychology, Psychoanalysis and Neurology* 7:591–592
Wolof-speaking children 7:929–930
Wolters, O.W. 23:71–72
WOM *see* Word of mouth (WOM)
Woman marriage, institution of 14:599–600
Womanhood 21:684
Womanism 8:908
Women 12:795
  as active sexual beings 10:815–816
  agency 10:196
  authoritarianism 25:639, 25:641
  biomedical research on 13:560
    coercive contraception 13:560
  in Buddhism 25:628
    bhikkhunis 25:629–633
    Bodhisattva 25:631
    body 25:631
    enlightenment 25:629–632
    feminism 25:628–629
    gender 25:628–631
    laywomen 25:630–631
    nuns 25:629–630, 25:632
    strategies and approaches 25:628–629
    Theravāda 25:628–629, 25:632–633
  celibacy 3:237
  in China 15:610–611
  civilizing mission, in Middle East 16:367–368
  consequences of nonmarital childbearing for 9:16
  contemporary Buddhist women's ordination movement 25:632–633
  cooking 9:291
  and crime 8:912–917
  and criminal justice 25:616–621
    double jeopardy 25:616–617
    gender responsiveness 25:619
    imprisonment 25:617–619
    knowledge about 25:616–617
    new initiatives 25:619–620
    offenders and criminogenic needs 25:616
    paternalism 25:616
    research and policy developments 25:619–620
    treatment in prison 25:618–619
    what works 25:617, 25:619
  demographic deficit of 15:608–609
  in developing countries
    and mobile communication 15:632
  domestic and intimate partner violence 10:194
  double burden of 20:837
  dowry deaths 10:193, 10:195
  in early Buddhism 25:629–631
  ecofeminism *see* Ecofeminism
  educational and career investments
    event history analysis for 8:308–309
  educational attainment 6:607
    in East Asia 9:57–58
    and fertility transition 9:96–98
  in eighteenth-century American medicine 25:622–623
  in eighteenth-century European medicine 25:622–623
  employment 14:606
    in East Asia 9:58
    increases in 14:606
  employment opportunities 9:64–66
  empowerment, and tourism 24:477–478
  exclusion
    in clinical trials 9:719–720
    political economy in 9:721
  family law 25:639
  and family therapy 8:815–820
  female arrest rates 5:133
  female leadership and use of force 25:639–640
  feminism *see* Feminism
  and films 9:161–162
  focus in gender work 9:716
  and gay movements 3:711
  gender gap 15:611–612, 15:612t
  global violence against 10:192–197
  health and socioeconomic outcomes of family planning among 8:789–790
  health movement 15:110–114
  history in health care 25:622
  history of 9:768
  HIV infection 11:140–141
    impact and epidemiology 11:140
    improving future prevention efforts 11:140–141
    intervention efficacy 11:140
    selected predictors of risk 11:140
  honor and shame 11:183
  household economy 23:374
  human rights 6:793, 10:192–193, 10:196
  husband–wife relationship 3:238
  indigenous women, violence against 10:195–196
  and Islamist movements, in Middle East 16:368–369
  issue of covering their heads 3:237
  labor force participation 5:133
  language of 9:785
  leadership in medicine 25:626
  leadership in psychiatry 25:626
  in legal profession 13:632
  and leisure 13:854
  liberation 5:133
  in Mahāyāna Buddhism 25:631–632
  and marriage in China 21:735
  mating strategies 11:342–345
    long-term mating 11:342–343
    short-term mating 11:343–345
  measurement of sexual motivation in 21:762–765, 21:764t–765t
  media of 2:218–219
  in MENA 16:362–365
  in military 15:513–514
  as minority 15:573
  and national and international security 25:638–643
  national economic development and 25:641–642
  national stability and governance 25:640–641
  in nineteenth-century American medicine 25:622–623
  in nineteenth-century European medicine 25:622–623
  offending, qualitative research on 5:133–134
  oppression of 9:768
  participation in sports 9:656–657
  personhood 3:656–657
  perspective of sex 10:816
  in politics 8:891
    legislative diversity 8:889–890
  power at work 8:883–887
  powerlessness 1:547
  preliberation roles 5:133
  presence in judiciary 12:882
  prisons 18:926
  profession, men in 14:679–680
  and psychiatry 25:622–627
  in psychiatry 25:624–626
  and public sphere 19:613
  refugees 22:771–773
  religiosity 25:635–637
    birth and child rearing 25:636
    death 25:636–637
    ethic of connection 25:636
    family 25:635–637
    feminism 25:635–636
    gender 25:635–636
    guilt 25:636
    lack of status 25:635
    material and social deprivation 25:635
    patriarchy 25:635
    prayer and Bible reading 25:635
    spousal bereavement 25:636–637
    workforce participation 25:636
  reproductive health and freedom 9:732
  researchers 9:715
  rights 9:786, 14:14–15
  Roma 20:756–757
  schooling
    global advancement in 9:65–66
  sexism 21:693–699
  sexual dysfunctions in
    diagnostic methods 21:691
    psychological treatments of 21:691–692
    validated treatments for 21:691–692
  sexual violence 10:194–195
  sociology of, Myrdal's contribution 16:162
  speech 3:237
  subordination, and anarchism 19:748
  suffrage 3:240
  suffrage in United States 22:443–444
  and time use 24:369
  trafficking in Latin America 24:512–516
  in twentieth-century medicine 25:623–624
  value of their time 9:63–64
  in Vietnam 15:610–611
  wages 9:63–66
  war and peace 25:638–639
  and wrongful convictions 25:790

Women, Environment, and Development (WED) 8:906–907
Women abuse 6:633–634
Women and Development (WAD) 8:906
Women empowerment 16:857
Women in Development (WID) 8:906
Women loving women 8:946–947
Women traders, in West Africa 1:302
Women welfare, and marriage payments 6:644–645
Women with childbearing potential (WCP), exclusion in clinical trials 9:720
Women with disabilities, reproductive rights of 20:471
Women-centered social work 9:801–804
Women's Anthropology Association 8:899
Women's Christian Temperance Union 15:590
Women's Educational and Industrial Union (WEIU) 9:710
Women's Environment and Development Organization (WEDO) 6:846–847, 6:849, 23:796
Women's Global Network on Reproductive Rights (WGNRR) 20:459
Women's International League for Peace and Freedom 17:644
Women's language 13:256–257
Women's liberation 8:926
Women's liberation movement 5:74, 8:926, 15:110–111
  and counterculture 5:85
Women's Liberation Union 8:905
Women's rights 8:794, 8:796–797, 8:926
Women's rights movement 15:110
Women's shelters 6:635–636
Women's sphere, Korea 2:73
Women's studies 9:736–741
  criticisms and methodological issues 9:739–740
  current and emerging directions in theory and research 9:740
  current state of discipline 9:738–739
    academic programs 9:738–739
    journals 9:739
    research 9:739
  Fromm's 9:460
  identity 9:737–738, 9:740
  intellectual backgrounds 9:737–738
  intersectionality 9:736–740
  oppression 9:736, 9:738
  power 9:736–737, 9:740
  scholarship 9:740t
  suffrage movement 9:737
Women-specific communities 3:548
Woodcock–Johnson Tests of Cognitive Abilities 12:286–287
  Achievement, Third Edition (WJ-III ACH) 16:725
  Third Edition, WJ-III 12:285
Woodstock, countercultural ideology 5:83–84
Woodward, D., works, *The History of Cartography* 3:174
Woodward, Jim 21:303
Woodworth Personal Data Sheet (PDS) 17:827
Woolf, Virginia 12:875
Woolgar, Steve
  on laboratory work 5:524
  works
    Knowledge and Reflexivity 20:94
    Laboratory Life: The Social Construction of Scientific Facts 5:523, 8:577–578, 13:205–206, 21:311, 24:136
Word blindness 16:685
Word classification 17:553–554
Word formation 15:821–822
Word knowledge
  and semantic memory 21:572–573, 21:575, 21:577–578
  and specific language activities 13:927–928
Word learning 2:651
  lexical consolidation 13:919–924
Word lists 20:51–52
Word meaning
  learning 25:227–229
  lexical semantics 13:930–936
  Vygotsky on 25:313–314
Word of mouth (WOM) 22:350
Word order 11:826–827, 23:904, 25:644–656
  form and function in constituent order 25:650–652
    different constituents, same function 25:651–652
    functional modifier categories 25:651
    same constituent, different functions 25:651
  general tendencies in constituent ordering 25:652–655
    arbitrary patterning 25:655
    domain integrity 25:652–653
    Dutch 25:652–653
    Ngiti 25:652
    head proximity 25:653
    scope 25:653–655
  Georgian 25:645
  investigating constituent order across languages 25:645–646
    Kiribati 25:645
    Makwe 25:645–646
    Nivkh 25:645
    Tiwi 25:645
  traditional constituent order typology 25:646–650
    basic constituent order 25:646
    categorial confusion 25:649–650
    constituent order correlations 25:646–647
    Duala 25:649
    employment of semantic criteria in constituent order research 25:648–649
    Kinnauri 25:650
    notion basic constituent order 25:648
    relevance of meaning 25:650
Word processor, Computer Assisted Instruction (CAI) 4:493–494
Word prosody 23:714–717
  metrical structure 23:715–716
Word reading 8:395
Word recognition 19:353–354
  visual 25:214–219
Word segmentation 25:226–227
Word space models 13:932–933
Word substitution 23:232
Word-and-paradigm 15:823
Word-based priming 19:396
Word-by-word basis 21:623
Word-related network 16:593
Words 4:531–532, 4:534, 4:536
  and embodied cognition 7:422–423
  searching and repair 4:854–855
  and society 13:288–289
  study of 17:223–224
  Vigotsky's theory 4:533
Words As Social Tools (WAT) 7:423–424
Word-stem completion task 1:645
Work 4:905, 4:907–908, 13:199, 15:370–371, 25:669–674
  anthropology of 25:662
    agriculture 25:664
    coalescence of 25:665–667
    contesting the boundaries of work 25:665–667
    from ethnography to ethnology 25:665
    reflexivity 25:667
    disciplinary and subdisciplinary orientations 25:662–663
    departures 25:662–663
    terminology 25:662
    employment 25:662, 25:664, 25:666
    ethnographic trajectories 25:663–665
    food collection 25:663–664
    food production 25:664
    manufacturing 25:664–665
    industry 25:663–666
    labor 25:662–667
    modes of production 25:662–663
    moral economy 25:662
    natural economy 25:662, 25:664, 25:666
    neoliberalism 25:663, 25:665–666
    political economy 25:662–665, 25:667
    social change 25:663–664
    subsistence 25:662–665
    unemployment 25:662, 25:666
  artisans 2:47
  attitude 17:334
  -based learning, in teaching 24:53
  -based migration, and immigration policy 11:662, 11:665
  behavior 20:294–295
    of older people 13:197
    personal intelligence 17:754–755
  changing nature of 23:546–547, 25:690–692
    alienation in industrial work 25:691
    computerization 25:691
    deskilling and upskilling 25:691
    industrial policy 25:691
    outsourcing and offshoring 25:692
    trade unions 25:692
    training and education 25:691–692
  as concept 25:688–689

concepts of 25:669–670
constitution of social order 25:676–678
labor, class, and nation 25:677
labor as status 25:676–677
transformations in work and reconfiguration of social sciences 25:677–678
as culture-bound idea 25:688
design, and ergonomics 11:302–303
dimensions 25:688
disincentives 25:492–494
division of labor and gender 25:690
and emotions 7:507–512
end of 25:692
engineering 7:641–645
enjoyment of 13:850
environment 17:328–329
and ethnic identity 8:106–112
ethnoarchaeology at 8:163–164
and family 3:50–52
and family history 8:738, 8:740–742
female power at 8:883–887
flow in 9:263
facilitation 9:263
governance at 10:295–297
and happiness 10:515–520
institutional 17:394
and knowledge society 13:107
labor history 13:169–173
and leisure 13:852–854
leisure work 13:854
long arm of job 13:853
modern 13:853
time 13:853
unemployment 13:853–854
and marriage 8:829–830, 8:833
in modern societies 25:689
motivation 25:660, 25:680–686
content approach 25:683–685
in New Economy 25:692
organization, in hospitals 11:212
and postindustrial societies 18:663–668
-related behaviors 17:824–825
-related consequences of work–family conflict 8:760
-related stress
accidents due to 12:132
spillover 23:549–550
relations 25:687–690
retirement 25:692
and retirement 20:593–594
science at 21:227–228
semantic reduction to labor 25:675–676
invention of abstract work 25:676
wage labor 25:676
sex 21:799
sexual harassment in 21:721–727
social control of 25:689–690
employee participation 25:689–690
technological progress control 25:690
and social stratification 22:707–711
and sociology 25:657–661
challenge at 25:661
changing conceptions of 25:658
conflict at 25:661
divisions of labor 25:658–659
global social–political–economic change 25:658–659
meanings, experiences, and identities 25:660–661
mischief at 25:661
work organizations 25:659–660
sociology of 23:529–530, 25:687–693
statistics 5:798–802
statuses and roles 25:688–690
stress
coronary heart disease 4:919
occupational stress management 17:147–150
stress at 23:546–550
and welfare 25:491–496
Work, history of 25:669–674
and gender 25:672–673
history of concept 25:669–670
modern western society 25:670
preindustrial societies 25:669–670
as practice 25:670–672
as culture 25:672
economic activity 25:671–672
production 25:670–671
Work ability, aging and 12:132
Work analysis 17:936
job analysis *see* Job analysis
Work design
occupational stress management 17:149
Work engagement 25:537–539
and authentic leadership 2:248
of well-being and burnout 25:534–536
Work Environment Scale (WES) 7:820
Work first principle 25:516
Work for pay 13:190
Work groups 24:96
Work hours 13:190–195, 13:197–198
Work organizations 25:659–660
division of labor changes 13:159–160
Work orientations 25:660–661
Work paradox 25:242
Work participation 17:183
old age 17:182–189
rheumatoid arthritis 20:657–658
Work restructuring 25:659–660
Work roles 12:812
Work schedules, nonstandard, in dual-career families 14:609
Work settings
of advanced practice nurses 16:911–912
of nonphysician health professionals 16:909
Work sharing 24:741
Work studies 12:812–813
Work units, in China 2:66–69
Work–age patterns 13:194
Worked examples, in mathematics learning 4:70, 4:70t
Worker alienation 25:660
Worker control 20:675
Worker involvement 17:144–145
Worker Photography movement 18:84
Worker–employer cleavage 20:746
Worker–firm relations 6:272–274
Workers
compensation laws 12:216–217
cooperatives 17:202–203
experience, and sexual harassment 21:721–723
white-collar *see* White-collar workers
Workers aristocracy 1:965
Work–family arrangements 14:606
Work–family conflict 1:424–425, 8:758–765, 8:759f
documented influences 8:762–763
future directions 8:763–764
prevalence and consequences 8:760
survey measures 8:763
theories and perspectives 8:760–762
of working mothers 9:730–731
Work–family interactions, and powerlessness 1:548
Work–family reconciliation 8:804
Workfare, poverty policy 18:749
Workfare programs 6:274
Workforce participation, and religiosity of women 25:636
Working children 9:445–446
Working class ideology 10:743
Working classes 3:739–740, 23:4
artisans 2:48
and cultural policy 5:461
history of 25:694
agency 25:695
classical debates 25:695
in Europe 25:694–695
global insights 25:696
historiography 25:695–696
Marxism 25:696–697
origins in core countries 25:694–695
origins of concept 25:694
research 25:696–697
revolution 25:695
rise of periphery 25:695–696
structuralism 25:695
Thompson on 25:695
western Europe 25:694–695
and poverty 18:730
and social inequality 22:339–343
and social mobility 22:427–428
and social work 24:491–492
and white-collar workers 25:566–567
Working conditions 25:698–702
across countries 25:699
collective bargaining 25:698–701
compensation 25:701
definition 25:698–699
employment policy 25:698–701
employment relationship 25:698, 25:700–701
health and safety 25:698–701
informal employment 25:701
labor market 25:698–700
occupational stress management 17:147–151
regulation
debate 25:700
diversified employment arrangements 25:700–701

Working conditions (*continued*)
enforcement 25:701
future research issues 25:701–702
globalization of production 25:701
purpose 25:699
sources 25:699–700
social insurance 25:699
systems of employment 25:699
of teachers in United States 24:77–79
inequity in 24:77–78
voice at work 25:698–699
Working gangs 9:594
Working Group on Indigenous Populations 11:797
Working Group on Visual Sociology 25:193
Working Group on Girls (WGGs) 24:754–755
Working hours
length as status symbol 13:846–847
and work–leisure relationship 13:845, 13:847
Working memory (WM) 1:244, 1:400, 12:324, 13:268, 13:270, 13:679, 16:635–637, 19:3, 21:896, 21:901–902, 24:308–309, 24:566
affect-regulation motivation 1:243–244
and aging 1:400–401, 1:430–431, 1:431f
and agrammatism 16:693–694
contextual comprehension 16:693–694
impairment 16:694
application of concept 25:716
asymmetry, and dichotic listening 6:358, 6:358f
binding in 15:140–144
cross-domain and cross-modal binding 15:142–143
verbal binding 15:143
visuospatial binding 15:141–142, 15:141f
and bounded rationality 2:773, 2:775
central level and peripheral level in 21:902–903
of children 8:492
and cognitive development 16:504–505, 16:508
and creativity 17:771–772
definition 25:711
developing of 25:713–715
and ergonomics 11:300
fronto-parietal circuit 25:707–709
hippocampus during working of 1:646–647
impairment 6:720
and intelligence 12:293
and language learning impairment 13:346, 13:348–349
maintaining information in 21:906
mental imagery in models of 15:222
multicomponent 25:715–716
multiple-component model of 21:902
neural basis of 25:703–710
prefrontal cortex and 25:703–707
proactive interference in 19:3–5
in Brown-Peterson paradigm 19:3–5, 19:4f, 19:5t
in focus of attention 19:5
psychology of 25:711–717
role in vocabulary acquisition 25:229–230
role of cerebellum in 2:96–97
and serial order memory 21:639, 21:641–642
temporal and parietal cortex 25:706–707
temporal and parietal cortical interactions 25:707
theoretical views of 25:711–712
usefulness of 25:711
visuospatial resources of 25:716
Working memory capacity 16:504
Working memory concept as a specific model (WMS) 21:902–904
Working memory deficits 17:510
restitution of 2:816
Working memory general (WMG) 21:902–903
central level in 21:904–906
Working Memory Index (WMI) 12:284, 16:723
Working poor 18:729
Working scientists, and axiological neutrality 16:757–758
Working slogan 20:675–676
Working with memory procedures 20:610
of retrieval 20:610–612
Working with others 19:234–235
Working-age adults, value of life for 25:18–19
Working-age population 18:542–543
Work-injury compensation 24:430
Work–leisure relationship, and working hours 13:845, 13:847
Work–life balance 23:549–550
and careers 3:133
in dual-career families 14:607–608
family policy 8:801, 8:804
Work–life fit, I–O psychology 11:883
Work–life–family balance 21:682
Workload capacity 12:78
Workplace
attitudes and age 1:423–424
and community struggles, relationship between 14:666
deviance
counterproductive work behaviors 25:723–725
educational research 7:282–283
ethics of genetic testing in 20:703
harmful obedience in 17:68–69
of lesbians 13:894–895, 13:897
motives 1:424–425
outcomes and age 1:420–421
personality and values at 17:821–826
separation, and labor 13:158
smoke-free 22:128–129
well-being and burnout in 25:534–536
conservation of resources model 25:534–535
emotional labor 25:535
intervention programs 25:535–536
job demands-resources model 25:534
outcomes 25:535
person-job fit model 25:534–535
predictors 25:535
somatic difficulties 25:534
theoretical models 25:534–535
work engagement 25:534–536
Workplace authority, of teachers in United States 24:79–80
Workplace autonomy 25:718–721
alienation 25:718–719
and groups/teams 25:720
and individual 25:718–720
internal locus of control 25:719
job characteristics model 25:718–719
job demands–job control model 25:718–719
lean management initiatives 25:720
and organizational democracy 25:720–721
participatory decision-making 25:719–720
psychological empowerment 25:719–720
self-determination theory 25:719–720
self-organizing teams 25:720
Workplace cultures 3:757
Workplace deviance 25:722–726
antisocial workplace behaviors 25:722–724
assessment 25:722–723
history 25:723
interpersonal deviance 25:722–723
managing 25:724
monitoring 25:724–725
organizational deviance 25:722–723
political deviance 25:722–723
potential antecedents 25:723–724
person-based perspective 25:724
situation-based perspective 25:724
production deviance 25:722–723
property deviance 25:722–723
recommendations 25:724–725
typology 25:722
Workplace discrimination 9:731
Workplace injuries 20:863–869
Workplace interventions 25:735–736
primary interventions 25:735–736
secondary interventions 25:736
tertiary interventions 25:736
Workplace learning 15:83
Workplace misbehavior 25:722–723
Workplace organization, new forms of 2:768–769
Workplace Project 3:8
Workplace regulations 6:273
Workplace safety and health 25:727–732
accidents, theories on 25:727–729
causes 25:727–728
risk taking 25:728–729
behavior modification approach 25:729–731
analysis of intervention data 25:731
control group design 25:731
experimental designs 25:730–731
interventions 25:729
multiple-baseline design 25:731
positive reinforcement 25:729
reversal or ABAB design 25:731
safety behavior, measurement of 25:730
definition 25:727

ELMERI method 25:730
future recommendations 25:731–732
goal setting 25:729
history 25:727
TUTTAVA information board 25:730f
Workplace stress 25:733
contemporary issues 25:736–738
contemporary models 25:734–735
conservation of resources 25:734
ERI model 25:734
job demands–resources 25:734–735
contemporary topics in 25:737–738
historical models
ISR model 25:733
JDC model 25:733–734
Workplace studies, ethnomethodology 21:197
Workplace-based assessments, of clinical skills of medical students 7:191
Works of art 11:100
Works Progress Administration 17:23
Workscapes 25:667
Workshop handcrafts 5:111
Workshops, production 12:329, 12:331
Work–wage elasticities 13:197
World 3:585–592
complex intertextural, postmodern methodology for 18:681
historical prerequisites 11:863
and intentionality 12:333
languages, serial verb constructions in 21:646
with many centers 14:394
meanings in Kuhn's philosophy 13:135
scientific conception of 14:315–317
unified science 14:315–317
westernization of 25:364
World3 14:139
World anthropologies, ethnology 8:198
World art system 2:17–18
World Association for Psychiatric Rehabilitation 15:216
World Association for Public Opinion Research (WAPOR) 23:752–754
World Association of NGOs (WANGO)
NGO Handbook 16:859
World Bank (WB) 5:417, 7:180, 7:204, 7:922, 8:60, 10:198, 10:601, 11:802, 16:462, 18:354, 18:763, 22:696
and capital 3:93
databases on economic and growth 5:749
and education policy 7:267–268
Africa and 1:281, 1:283–284, 23:598–603
International Center for the Settlement of Investment Disputes in Washington 12:510
political economy of renewable energy 18:370
and youth movements 25:839
World Baseball Classic (WBC) 23:304
World cities 9:640, 24:815
World citizenship 3:653
drivers of transnationalization 24:588–589
facets 24:588–589
World Climate Research Programme (WCRP) 7:791
World Color Survey 13:294
World Commission on Environment and Development (WCED) 7:771, 23:812–813, 23:817, 23:845
World Communism 18:406
World Conference on Education for All (EFA) 7:262
World Conference on Human Rights (Vienna) 11:369
World conferences and summits, United Nations 24:751
World Conservation Union 23:796
World Council of Comparative Education Societies 7:284
World Council of Indigenous Peoples 11:757
World culture 5:539
globalization and 10:225
authenticity 10:228–229
Christianity 10:227, 10:231
civil society 10:227–228
construction 10:225–226
creolization 10:228
dimensions 10:226–227
glocalization 10:228, 10:230
homogenization, heterogeneity, and conflict 10:228–230
hybridization 10:228
intergovernmental organizations 10:226, 10:228–229
international nongovernmental organizations 10:225–229, 10:231–232
isomorphism 10:229, 10:231–232
and national identity 10:230–231
nongovernmental organizations 10:225–226
origins and expansion 10:227–228
relativism 10:229
religion 10:231–232
transnationalism 10:226–232
Western models 10:230
World Database of Happiness 5:794, 10:524
World Development Indicators 1:478
World ecology 1:460
World Economic Forum 5:18
World economy, relationship of Africa with 1:287–288
World Education Research Association 7:284–285
World Federation for Mental Health 15:216, 22:677
World Fertility Survey (WFS) 5:737, 9:96
World Giving Index (WGI) 3:368
World Health Organization (WHO) 4:401, 5:370–371, 7:815–816, 8:78, 8:330, 16:912, 19:477, 20:514, 21:736
active aging model 6:380
on alcohol abuse 1:517, 1:519
classification of diseases 17:28–29
concepts outside 10:611–612
constitution 10:608–609
health definition, criticism of 10:612
Health for All by the Year 2000 3:874
health in Africa 1:280, 1:285
health information systems 5:781–787
Healthy Cities project 24:832–833
on insomnia 22:79
mental health organizations 15:216–217
operationalization of health 10:609–610
classification systems 10:609–610
quality of life and well-being and health-related quality of life 10:610
statistics of death causes 23:673t
surveys 5:793
United Nations 24:754–755
World heritage
cultural 5:414–418
World history 9:671–672, 10:162–168, 11:32, 25:290–291
and African history 1:288
ecumene 7:111
World in-itself 13:135
World Intellectual Property Organization (WIPO) 11:775, 11:782, 11:802, 12:264, 12:267, 12:271, 13:525, 20:499
World Interdisciplinary Network for Institutional Research (WINIR) 17:194–195
World knowledge, and inferences 12:40–41, 12:43
World language system 3:310
World languages 3:310
World Medical Association (WMA) 8:78, 21:316
Declaration of Helsinki 20:525–527
World Mental Health Survey 14:519–520, 15:179–180
anxiety disorders 6:340
World Meteorological Organization 3:853
World modeling 6:267
LTG study 14:138, 14:141
World Monuments Fund 5:417
World of Warcraft (game) 6:900
World Organization of Systems and Cybernetics (WOSC) 22:912
World peace, theory of 13:3–4
World polity theory (WPT) 7:801–802, 16:767
challenges 16:768
contributions 16:767–768
World Population Prospects 5:778
World poverty, and Third World 24:302–303
World Pride 9:640
World Professional Association for Transgender Health, Inc. (WPATH) 24:550
World Psychiatric Association 18:35, 19:317
World religions 20:310–311
World Report on Disability 13:500–501
World republic 25:740–747
global common sense of justice 25:744
global peace and global democratization 25:744
global public 25:744–745
globalization 25:740–741

World republic (*continued*)
human rights 25:743–744
leveling 25:746
models 25:741–742
priority of statehood 25:741
ungovernability 25:742–743
World resources 14:138
World Social Forums (WSFs) 10:186, 10:188–189, 22:510
World society 12:603
law in 13:555–556
Luhmann on 14:386–389
and social integration 12:243, 12:245–246
theory of 5:539, 16:765
World soul (*psyche*) 20:285
World Summit for Information Society (WSIS) 16:798
World Summit for Social Development (WSSD) 12:497–498
World system 1:460, 14:670–671, 18:350–352
and collective action 15:254–257
dependency theory 6:196–197
and diffusion 6:404
in Europe 8:274–276
in multisited field studies 16:103, 16:105
World system economic models 18:659–660
World system theorists' skepticism 10:220
World system theory (WST) 3:308, 3:310, 7:111, 7:799, 8:285, 18:362, 18:364, 21:151–152, 22:994, 24:300–301, 25:748–752
backwardness 25:749
core societies 25:748–749
and creolization 5:120
and global capitalism 3:99–100
globalization 25:749–751
imperialism 25:748–750
intellectual history of 25:748–749
Marxism 25:748–751
and migration 15:468–469
modernization 25:748, 25:750–751
nationalism 25:749, 25:751
and nationalism 16:292–293
peripheral societies 25:748
and political economy 1:940
political left 25:749, 25:751
political right 25:749
revolutions 25:748–751
semiperiphery 25:748–749, 25:751
theoretical limitations and methodological promises 25:750–751
urbanization and underdevelopment 6:285
uses of 25:749–750
World Tourism and Travel Council (WTTC) 24:465
World Town Planning Day 24:868
World Trade Center (WTC) 9:336, 23:366
World Trade Organization (WTO) 1:684, 7:575, 9:640, 10:156, 10:158, 10:186, 10:188–189, 10:226, 10:854, 11:757, 12:247, 12:262–264, 12:569, 13:522, 13:621, 15:72, 16:465, 17:163, 25:839
dispute-settlement procedures 12:510
and ethical turn 8:60
media 15:12
political economy of renewable energy 18:370
trade policy 24:485, 24:489
World trading system 4:224, 12:577–578
World Values Surveys (WVS) 5:528, 5:531, 5:533, 5:791–792, 5:820–821, 5:834, 18:492, 18:494–495, 19:566–567, 22:175–176, 22:178, 23:748, 24:658, 25:30
creation of 5:791
and public opinion 19:566–567
World War I (WWI) 8:224, 8:227–228, 9:212–217, 15:490, 16:259, 18:400, 18:434, 21:368, 24:763–764
armistice during 9:215
casualties 9:215
causes 9:212–213
conduct 9:213–215
1914 9:213
1915 9:214
1916 9:214
1917 9:214
1918 9:214–215
consequences
cultural 9:216–217
demographic 9:215
economic 9:216
political 9:215–216
social 9:216
effects on homosexual rights movement 9:645–646
ergonomics during 11:297
and holocaust 11:151
and indigenous rights 11:796
military history 15:506–509
and nationalism 25:366–367
Oceania after 17:162
political science after 18:467–468
propaganda 25:396
war crimes tribunals and 25:382–383
World War II (WWII) 3:710–711, 8:223, 8:231, 8:234, 13:558–560, 13:793, 16:283–284, 18:399, 19:576, 20:514, 21:368–369, 24:764
adoption after 1:143, 1:145
American growth after 22:528–529
and anthropology of ethics 1:735–736
business anthropology following 3:7
casualty statistics 25:377
cause 21:369–370
codification and institutional regulation of research ethics after 21:315–317
consequences 21:371–372
course 21:370–371
database development during and after 5:830
developments in transnational organized crime since 24:597–598
drag cultures during 13:876
ergonomics during/after 11:297–298
and facism 16:259–260
gay/lesbian movements during 9:632–633
and holocaust 11:151–153
homosexuality and 3:886–887
and indigenous rights 11:796
Indology after 11:78–79
interviewing after 12:674–675
and macroeconomics 14:401
Middle East politics after 16:371–372
military history 15:506–510
Oceania after 17:162
and origins of area studies in United States 1:951
political science after 18:468–469
postwar expansion of social history 22:308–311
post-war sociological theory on psychiatry 22:975
pre-WWII, gay/lesbian movements in 9:632
propaganda 25:396
and Scandinavian criminology 5:240
war crimes tribunals and 25:382–383
war-related deaths in 25:379
World wars 6:810–811, 6:813–814
social psychological theories between 22:527–528
World Watch Institute 4:745
World Wide Fund for Nature (WWF) 7:764
World Wide Governance Indicators 10:293–294
World Wide Web (WWW) 7:686, 15:12, 24:259
World Wildlife Fund 3:854, 7:105, 23:795
World-disclosing function of language 4:790, 4:792–794
Worldly phenomenology 21:149
World-making 13:476–477
World's Christian Fundamentals Association 20:397
World's fair, entertainment 7:673
Worlds of Food intervention 10:80
Worldviews 6:243–244, 25:753–757
anthropological studies 25:753–755
applications 25:755
of archaic states 23:374
in Arctic 1:929–930
of Australian aboriginals 2:236
belief 2:493–494
cognitive anthropology 25:753–756
cultural psychology 25:754–756
Dilthey on 6:426–428
economic 11:864
ethnomathematics 25:755
ethnoscience 25:753–754
Humboldt on 25:288
of individual 17:935
and language 25:576
and mental health disorders 5:570, 5:572–574
mentalism 25:753–754
microcosm 24:722
mind 25:753–754
Māori on 14:507
notion of 25:753
religion and ideology 25:755–756

and social network actors 5:559
and sphere 25:755
and spirituality 23:286, 23:289
verification model of self-esteem 15:241
Weltanschauung 25:755–756
Western 14:505–507
Worldwide aspects, shadow economy labor force 21:863
Worldwide industrial dynamics 11:835
Worldwide international patenting activity 11:834
Worms, René 22:992
Worship, ancestral 1:688–689
Worst artist references 5:561
Worst-case scenario 16:583–584
Worster, Donald, works, *Natures Economy* 7:748
WOSC *see* World Organization of Systems and Cybernetics (WOSC)
Wound healing
psychoneuroimmunology 19:446
and stress 23:552
WPATH *see* World Professional Association for Transgender Health, Inc. (WPATH)
WPPSI-IV *see* Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV)
WPT *see* Weather prediction task (WPT); World polity theory (WPT)
Wrap-up processes 21:623
Wray, K. Brad, on epistemic cultures 7:877
Wright, Basil
works
Night Mail 6:621
The Song of Ceylon 6:621
Wright, E.O. 1:665–666
class scheme 3:748, 3:748t
Wright, James D.
on women's participation in sports 9:657
works
Armed and Considered Dangerous: a Survey of Felons and Their Firearms (*ACD*) 10:457–459
Wright, Les
on bear subculture 13:890
works
The Bear Book 13:890–891
The Bear Book II 13:890–891
Wright, Robert, works, *The Moral Animal* 22:904–905
Wright, Sewall 13:24, 25:758–760
on animal breeding 25:758–759
on causal pathways 25:759
on coefficient of relationship 13:44
on dualaspect panpsychism 25:760
early life of 25:758
on evolution 25:759
on genetics 25:759
on inbreeding 25:758–759
on path analysis method 25:759–760
philosophy of 25:759–760
on populations 25:759
retirement years of 25:759
on shifting balance theory 25:759
on social sciences 25:759–760
on statistics 25:759
works
The Moral Animal 22:904–905
Wrigley, E.A. 11:5
Writing
about traumatic events 7:500
anthropological *see* Anthropological writing
in archaic states 23:375
birth of 25:773–775
commemorative 4:235
and deconstruction 5:927–928
development 25:767–768
evolution of 25:761–766
cuneiform 25:761–764
data storage 25:761
handling data in abstraction 25:765
logography 25:762–764
modern alphabets 25:764–765
numeral 25:762, 25:765
phonetic sign 25:761–763, 25:765
pictography 25:762
segmentation of sounds 25:764
stability of writing systems 25:765
syllabary 25:763
tokens as precursor 25:761–762
expression in 17:274–275
instruction 25:767–772
caveats 25:770–771
composition 25:768–770
creating supportive writing context 25:768–769
increasing developing writers' knowledge for and about writing 25:769
motivation 25:768
process writing 25:768, 25:770
skills, knowledge and strategies of developing writers 25:769
students with special writing needs 25:770
teaching developing writers strategies for planning, drafting, revising, and editing 25:769–770
teaching foundational writing skills to beginning writers 25:769
using writing to support students' reading and learning 25:770
and knowledge 25:783
Lacan on 13:214
media 15:11, 15:13
medium of 11:117
as memory, metaphors and metonymy 15:291
new practices of 14:248
as normative practice 25:782
Plato's critique of 18:209
relation with speech 25:780–781
skills for ethnography 8:190
thinking for 25:782–783
Writing culture 18:85–86
Writing culture critique 20:90
Writing culture movement 21:616
Writing culture school 18:364–365
Writing disorders 16:685–686
Writing error 20:3
Writing impairment in children, assessment of 3:794–795
Writing style of journalism 3:613
Writing systems 13:253, 25:773–784
alphabetic 20:5–7
Chinese 25:778
consonantal 25:775
Japanese 25:778
Korean 25:778
literacy and social development 25:782
logographs 25:780–781
metalinguistics 25:781–782
nonalphabetic 20:5–8
pictographs 25:780
and reading 25:781
in South and East Asia 25:777–778
syllabaries 25:780
theories of reading alphabetic scripts 25:781
and thinking 25:781–782
types of 25:773
Writing techniques 5:384
Writs 4:247–248
Written corpus 4:950
Written data, qualitative developmental psychology 19:671
Written discovery 6:517
Written documents, ethnography 8:188
Written history, sub-Saharan Africa 11:89–91
Written language 1:450
acquired impairments 1:450–451
assessment principles 1:454–455
cognitive neuropsychological approach 1:451–454
central dysgraphia 1:453–454
peripheral dysgraphia 1:454
processing models 1:452–453
deficits in classical aphasiology 1:450–451
neuroanatomy 1:455
Written narratives 16:219–220
Written records in indigenous languages 8:195
Written text
material aspects of 14:243
as a metaphor 14:243–244
Written words, information encoded in 23:265–266
Wrong, Denis 9:539–540
Wrongful convictions 10:455, 25:785–791
African Americans and 25:789–790
death penalty 25:786, 25:789
DNA testing 25:785–790
exonerations 25:787–790
factors associated with 25:788–789
factual innocence 25:785
forensic errors in sentencing 25:790
gender differences 25:788, 25:790
and genetics 9:971
history of 25:785–787
legal innocence 25:785
National Registry of Exonerations (NRE) 25:787–789
racial disparity 25:788–790
sources of data on 25:787–788
in United States 25:789

Wrongful convictions (*continued*)
women and 25:790
WS *see* Williams syndrome (WS)
WSFs *see* World Social Forums (WSFs)
WSIS *see* World Summit for Information Society (WSIS)
WSSD *see* World Summit for Social Development (WSSD)
WST *see* World system theory (WST)
WTC *see* World Trade Center (WTC)
WTFR *see* Wanted total fertility rate (WTFR)
WTO *see* World Trade Organization (WTO)
WTP *see* Willingness to pay (WTP)
WTTC *see* World Tourism and Travel Council (WTTC)
Wu, Yuxiao, on cultural capital 5:377–379
Wu Jian-Chang 24:585
Wulf, Christoph 2:729
Wundt, Wilhelm Maximilian 10:740, 13:127–128, 14:810, 19:391, 19:413, 19:472, 23:453
on applied social research 1:850, 1:853
on ethnopsychology 25:793–794
experimental psychology studies of 3:873, 12:702–703
and genesis of discipline of modern psychology 25:793–794
influence of 25:792–794
and Institute for Experimental Psychology 25:794
on introspection 12:702
life and times of 25:792
on politics of science 25:794
on psychological diagnostic testing 25:794
on psychology 25:793–794
on psychophysiological problems 25:792
works 25:792–794
Contributions to the Science of Muscle Movements 25:792
Contributions to the Theory of Sense Perception 25:792
Ethnopsychology 25:792–793
Handbook of Medical Physics 25:792
Introduction to Philosophy 25:792–793
Lectures on the Mind of Humans and Animals 25:792
Logic: An Investigation of the Principles of Knowledge and Methods of Scientific Research 25:792–793
Philosophical Studies 25:793–794
Physical Axioms 25:792
Principles of Physiological Psychology 25:792–793
System of Philosophy 25:792–793
Textbook of Physiology 25:792
Völkerpsychologie 22:532–533
Würzburg school (of cognitive psychology) 12:703
Wuthnow, Robert 20:816
WVS *see* World Values Survey (WVS)
WWC *see* What Works Clearinghouse (WWC)
WWF *see* World Wide Fund for Nature (WWF)
WWI *see* World War I (WWI)
WWII *see* World War II (WWII)
WWW *see* World Wide Web (WWW)
www.Bear411.com 13:891
WWW&H Rule 24:93
Wyer, Robert S.
works, *Handbook of Social Cognition* 7:590
Wylie, Alison 6:245

X

X chromosome
and sexual orientation 21:775
Xalaalays 8:880
Xavante Amerindians
microevolution 15:340
Xavier, Francis 3:532
X-bar schemata 23:887–888
Xeno-hostility 25:798
Xenophilia, *versus* xenophobia 25:806
Xenophobia 2:109, 10:171, 18:411–413, 25:796–801
among migrants 15:418–419
antifeminism 25:800
anti-immigration 25:797
antimulticulturalism 25:800
in Australia 2:239
current usages 25:797–798
exclusionary ideologies 25:796–797, 25:800
genocide 10:19
Islamophobia 25:798–799
neoracism 25:797, 25:800
new racism discourse 25:799–800
and racism 19:852–853
social phobias 25:797–798
social psychological aspects 25:802–807
asylum seekers 25:803, 25:805–806
and attitudes toward immigration issues 25:804–805
fear of other 25:802–803
identity threats 25:804
immigrants 25:802–806
immigration policy 25:805
integration policy 25:805
material threat 25:804
nationalism 25:804
prejudice 25:802–806
realistic threats 25:803–804
reducing 25:805–806
social context 25:803
versus xenophilia 25:806
social psychological aspects, asylum seekers 25:803, 25:805–806
Xenophon 7:144–145
works, *The Education of Cyros* 6:601
Xerox Corporation 3:35
Xeroxlore 9:282–283
Xiafang 24:973–974
Xiangqun, Li 5:599
Xiaoping, Deng 5:596
Xihuatoxtla shelter, Mexico 17:459
Xin, Gu, on Chinese cultural industries 5:597–598
Xingshi hunyin 21:735
Xinjiang 2:65, 2:67, 3:316–317
X-linkage, nature/nurture 16:343–344
X-linked Kallmann syndrome 2:115–116
X-linked recessive disease 9:911
xMAS *see* Cross Media Annotation System (xMAS)
XMCA (Laboratory of Comparative Human Cognition) 25:333
X-Men 21:680
XML 4:951
X-People 21:680
X-ray diffraction (XRD) analysis 3:325, 24:481
X-ray fluorescence spectrometry (XRF) 10:56, 24:481
XRD *see* X-ray diffraction (XRD) analysis
XRF *see* X-ray fluorescence (XRF) spectrometry
Xu, Jun, on cultural capital 5:377, 5:379
Xu, Tai 2:892–893
Xunzi 24:584–585
works, *Wang Chi* 24:584–585
Xvétxvét 11:726–727
X11/X12 ARIMA 24:351–353
XY sex determination system 21:678–679

Y

Y chromosomes 9:943
Genographic Project 10:23, 10:25–28
testing 9:899
Y matrix 23:102
YAF *see* Young Americans for Freedom (YAF)
Yahoo 3:47
Yahwistic monolatry 12:869
Yaish, M., on cultural tastes 22:334–335
Yakkha 16:107
Yakuza 14:429–431
Yale School of Medicine 20:269
Yale University 25:575
Fisher at 9:218
Yale-Brown Obsessive Compulsive Scale Symptom Checklist (Y-BOCS-SC) 17:119
Yalom, lrvin 19:505
Yamagishi, Toshio 24:666
Yamamoto, Yoko, on cultural capital 5:379–380
Yamin, Muhammad
works
Gajah Mada 11:85
Tatanegara Majapahit 11:85
Yana 20:899
Yang, X.
on division of labor 6:604
theory of social orientation 11:792
Yangtze River Delta 20:136–140
Yang–Visscher approach 8:717–718, 8:721
Yankee City Study 3:7
Yankunytjatjara 13:291
Yanomami warfare 17:626–627, 17:629, 9:963b
Yardstick regulation 20:168
Yasin, Abdullah ibn 12:737
Yasutsugu, Shigeno 11:51
Yates, Frank 9:224

Yazawin 11:82
Y-BOCS-SC *see* Yale-Brown Obsessive Compulsive Scale Symptom Checklist (Y-BOCS-SC)
YD *see* Young People's Dyirbal (YD); Youth development (YD) programs
Years lived with disability (YLDs) 15:181–182
Years lost to disability (YLDs) 19:546–548
Yekuana 23:45
Yélîdnye 4:233
Yellow capitalism 24:821–822
Yeltsin, Boris 20:846–847
Yemen, legal system in 15:366
Yen, Cheng 2:893
Yen Fu 24:585
Yeoron 24:586
YEP *see* Youth Education Pack (YEP) Programme
Yerkes, Robert 13:392, 24:526
Yeshivot 12:873–874
Yes-or-no responses 10:774, 21:946, 21:953
Yessenova, S. 3:318
Yew, Lee Kuan 5:455
Yi Pyong-do, works, *Han'guksa taegwan* 11:51
Yi Nunghwa 2:72
Yiddish 12:877
  cultural movements 12:871
    decline in Eastern Europe 12:872
  culture 12:872
Yield gap method 13:240
Yijing 11:48
Yin-Yang theory 10:819
Y2K problem 12:601
Y-maze 15:155–156
YMCA *see* Young Men's Christian Association (YMCA)
Yoga 7:374, 24:726
  for asthma 20:579
  for chronic obstructive pulmonary disease 20:579
Yogyakarta principles 21:769–770
Yom Kippur War (1973) 1:559
Yonay, Yuval, on epistemic cultures 7:876
Yoneo, Ishii, works, *Thailand: A Rice-Growing Society* 11:87
Yonge Street riot 8:154–155
Yoruba
  cultural heritage 5:417
  tricksters 24:648, 24:652–655
Young, Alfred 1:612
Young, Allyn, on division of labor 6:602, 6:604
Young, Iris Marion 12:137, 12:915–917
Young, Kimball, works, *Handbook of Social Psychology* 7:591
Young, Michael, works, *Family and Kinship in East London* 4:363–364
Young, Neil 3:802
Young Americans for Freedom (YAF) 1:605
Young children 19:68–72
  beautiful language constructions 4:669–670
  moral development 13:120–121
    claim of hierarchy 13:121
    developmental sequence 13:121
    mechanisms of development 13:120–121
    'member of society' perspective 13:120
    'prior to society' perspective 13:120
    relationship between judgment and action 13:121
    stage model 13:120
    tests 13:120
    theoretical revision 13:121
    universal validity of stage model 13:121
Young drivers (transportation) 24:625, 24:627
Young intellectuals 12:279
Young Men's Christian Association (YMCA) 9:712
Young people 3:491–492, 25:826
  and development geography 25:828–829
  preferred action, and liberation theology 14:34–35
  residential care *see* Residential child care
  and urban geography 25:827–828
  and vocational education 25:232–237
Young People's Dyirbal (YD) 13:302
Young Poland movement 14:453
Young researchers, supervision of 15:235–238
Young Turks 25:837
Young Women's Christian Association (YWCA) 9:712
Younger historical school 11:93
Younghusband, Eileen 12:478
Youngstown master plan 21:920
Youth 1:117, 3:456, 7:213–214, 9:867, 23:638
  and communities 4:632, 4:635–636
  and conflict 4:634
  corruption, and mass media 14:691
  and cyberbullying 5:642–643
  defining 25:808–809
  definition 4:632–633
  disposable *see* Disposable youth
  employment 7:213–214
  family planning programs 8:796–797
  -friendly clinics 21:794
  gay, lesbian, and bisexual 9:649
    adolescence 9:649, 9:651–653
    assumed heterosexuality, questioning 9:652
    coming-out models 9:651, 9:653–654
    disclosure 9:653–654
    domains of sexuality and romance 9:650–651
    evolution of 9:649–650
    feelings 9:651
    gender 9:650–651, 9:653–654
    gender nonconformity 9:651, 9:653–654
    identity 9:649–654
    mostly straight 9:652
    nonheterosexual 9:650–654
    number of gays 9:651
    romantic relationships 9:654
    same-sex attraction 9:652
    self-acceptance and synthesis 9:654
    self-identification 9:653
    sexual behavior 9:652–653
    sexual continuum 9:650
    sexual minority 9:652–654
    sexual orientation 9:649–652, 9:654
    trajectories 9:651
  and generation 9:854–855
  geography of 25:824–829
    defining 25:825
    and development geography 25:828–829
    places 25:826
    and political geography 25:826–827
    scale 25:826
    spaces 25:826
    thematic research 25:825–826
    and urban geography 25:827–828
  HIV infection 11:138–140
    impact and epidemiology 11:138
    improving future prevention efforts 11:140
    intervention efficacy 11:139
    selected predictors of risk 11:139
  and homophobia, shifting intersections of 9:658–659
  and intimate partner abuse 12:680–683
  joblessness 25:830–836
  LGBTQ, school environment *see also* Schools
    contemporary 21:86–87
  as liminal, transitory phase 9:867
  media effects on 15:38
  modern movements 9:868
  oversharing by 17:440–444
  physical activity guidelines for 18:95t
  and religion 20:298–306
  sexuality 21:792–794
  and student rebellion in 1960s 9:868–870
  subcultures 23:638
  unemployment 13:174–175, 13:180–181
  and violence 4:632–634, 4:636
  and war 4:633, 4:636
  in West Africa 1:301–303
  see also Adolescents
Youth ALIVE! program 10:459
Youth bulges 10:6
Youth culture 1:318–321, 9:594–595, 24:597
  anthropology of 25:808–812
    age 25:808–809, 25:811
    agency 25:809–812
    Birmingham School 25:808–811
    defining youth 25:808–809
    education 25:810–811
    generation and context 25:810–811
    gerontocracy 25:810
    globalization 25:808, 25:810–811
    life course 25:808–809, 25:811–812
    socialization 25:809–810
    studies, cross-disciplinary antecedents of 25:809
    study, extending and diversifying 25:811–812
    unemployment 25:810–811
    urbanization 25:808, 25:811

Youth culture (*continued*)
emergence of 5:80–81
in Japan 12:797
origins, adolescence and 25:813–814
sociology of 25:813–818
baby boom 25:815
Birmingham School 25:817
Chicago School 25:814–815
consumer culture 25:813–814
counterculture and student movements in the 1960s 25:815–816
deviance and juvenile delinquency 25:814–815
generations 25:815–816
New Left 25:813, 25:815–816
postsubcultures 25:817
resistance and subcultures 25:816–817
schooling 25:813–815
Youth development, communities for 4:382
Youth development (YD) programs 3:449
Youth Education Pack Programme (YEP) 4:636
Youth galas, and transition to adulthood 24:579
Youth gangs 25:819–823
causes and nature 25:819–820
subcultural issues 25:819
Youth homelessness, street children 23:534, 23:536
Youth in limbo (life phase) 5:511–512
Youth movements 9:857, 9:859, 25:837–841
antiglobalization 25:839
assessments, prospects, and future directions 25:840–841
color revolutions 25:839–840
definitions and theoretical approaches 25:837
findings 25:837–840
anticolonial and neocolonial movements 25:838
Arab Spring 25:840
resisting neoliberalism 25:839–840
revolutions in eastern Europe 25:838–839
student movements in 1960s 25:837–838
indignados 25:839–840
Mannheim on 25:837–838
occupy wall street 25:839–840
political generations 25:837–841
post-communism 25:838–839
social movements 25:837
Youth revolts 6:552
Youth sports/physical activity
motivation of 15:914–920
coaches 15:917–918
friendship 15:918–919
future research directions 15:919
interventions 15:916, 15:918–919
motivated behaviors 15:916–918
parents 15:916–917
participation motivation 15:914–916, 15:918
peers 15:918–919
perceived competence 15:915–916
social relationships 15:916–919
theoretical frameworks 15:914–915
Youth subcultures, and leisure activities 13:831–832
Youth unemployment 25:830
adverse social impacts 25:835
literature on 25:832–835
nature of 25:832–833
scarring effect of 25:835
schooling and 25:833–834, 25:834t
stressful life event and 25:835
Youth wages 25:834
YouTube 12:600, 12:602, 25:840
and cultural participation 12:590
terrorists on 24:233–234
Ypsilon 15:217
Yuanmingyuan Artist Village 5:598–599
Yugoslavia 18:545–546, 18:550–551
conflicts 3:716
Yule, George Udny 9:223, 9:626, 17:661–662
Yun, Master Hsing 2:893
Yuppie culture 11:157
Yurkun, Iurii 21:742
YWCA *see* Young Women's Christian Association (YWCA)

Z

ZA *see* Zentralarchiv für Empirische Sozialforschung (ZA)
Zakat 12:731–732, 12:737
Zaloom, Cailtin 2:873
Zambezi river system 25:444–445
Zambia
education in 7:199, 7:204
sociocultural aspects 23:75, 23:77, 23:79–80
Zande witchcraft 14:436
Zangwill, Oliver
on neuropsychological rehabilitation 2:815
Zanuck, Darryl F. 9:156–157
Zanzibar 6:773
Zapatista movement 1:684, 10:188, 20:626, 22:514, 22:516
Second Intercontinental Encounter for Humanity and Against Neoliberalism 1:684
Zapotec conquests (*danzantes*) 11:57
Zapovednik system 20:832
Zaydan, Jurji, works, *History of Islamic Civilization* (*Ta'rikh al-tamaddun al-islami*) 11:65
ZBB *see* Zero-based budgeting (ZBB)
z-conditional sampling 16:117
ZDP *see* Zone of proximal development (ZDP)
Zebra finches
behavioral pattern, findings 2:694–695
brain-behavior system 2:691–692
vocal development 2:691f
action of androgens 2:692–693
developmental changes in song-control nuclei 2:693
vocal learning in 2:691
Zeisel, Hans 13:516, 13:635, 13:638
works, *On the History of Sociography* 7:562
Zeitgeist 9:668, 12:278, 19:309, 25:449
in search for interdiscipline integration 1:428
Zeitschrift für Agrargeschichte und Agrarsoziologie (ZAA) 17:670
Zeitschrift fur Geopolitik 10:92
Zeitschrift für Tierpsychologie 14:369
Zeitungswissenschaft 4:279–280
Zelenin, D. 1:754–755
Zelizer, Viviana 5:502
on culture and consumption 5:538
on value of children 5:538
works
Pricing the Priceless Child 5:538
Zemecha 3:440, 3:442
Zemstva 20:856
Zen Buddhism 2:889, 2:892
Zeno of Citium 3:774–775
Zeno's Paradox 15:575–576
Zentralarchiv für Empirische Sozialforschung (ZA) 5:832
Zentrum fuer Umfragen, Methoden, und Analysen (ZUMA) 12:550
Zentrum für Kunst und Medientechnologie 2:24
Zermelo–Fraenkel set theory 18:111–112
Zero
concept of 24:411
as number 14:780–781, 14:780f
Zero learning 21:126
Zero negation 16:413
Zero population growth 18:568
Zero reject, EAHCA principle 23:222t
Zero risk, theory of 25:728–729
0–1 loss function 23:414
0–2 years, friendship during 9:449
Zero tolerance 21:108
Zero tolerance policies 6:682
Zero-based budgeting (ZBB) 2:902
Zero-immigration policies 11:670
Zero-option 4:622
Zero-sum games 4:498–499, 16:418
axiomatic theories 2:339, 2:341
Zero-sum quality 9:595
Zero-sum situations, bargaining 16:415
Zero-survivorship bias 1:156–157
Zero-tolerance policing 1:790–791
Zero-tolerance school disciplinary policies 22:388
Zero-tolerance strategies, police 18:253
Zerubavel, Eviatar 5:502
on cognitive socialization 5:578–579
Zhang, B.C., on sexual minorities in China 21:734
Zhang, Xuecheng 11:50
works, *Wenshi tongyi* 11:50
Zhang, Yue
on Chinese art districts 5:599
on Chinese artists' autonomy 5:600
Zhang Tai-yan 24:585
Zhejiang village 2:67
Zheng, Qiao 11:50
Zheng Qiao, works, *Tongzhi* 11:50

Zhongshu, Dong 13:800
Zi, Han Fei 13:800
Zidovudine (AZT) 1:491–492
Ziegenfuss, Werner, works, *Handbuch der Soziologie* 7:588–589
Zigeuner 20:750
Zilber, Tammar B., on institutional work 17:394
Zimbabwe
  AIDS/HIV in 12:24
  sociocultural aspects 23:75, 23:78–80
Zimbardo, Philip
  Stanford Prison experiment 4:154–155, 24:711–713
Zimelidine 1:779
Zimmerman's model of self-regulated learning 21:537–538
Zinnes–Griggs model for unidimensional pairwise preferences 17:844–846, 17:846f
Zinovieva-Annibal, Lydia 21:741–743
  works
    Thirty-Three Freaks 21:741
    The Tragic Zoo 21:741
Zinsser, Hans, works, *Rats, Lice, and History* 18:174–175
Zionism 1:787, 12:877, 16:363, 17:430, 19:326, 20:341–342
Zionist identity 12:870–871
Zip code demographics 1:840
Zipp, J.F., on gay athletes 9:658–659
Zippo Manufacturing Co. v. Rogers Imports 13:768–769
ZKPQ *see* Zuckerman–Kuhlman Personality Questionnaire (ZKPQ)
Znaniecki, Florian
  on habit 10:477–478
  works
    Polish Peasant in Europe and America 6:19–20
Zoarites 2:147
Zoku 12:794
Zombie categories, in cosmopolitanism 2:399
Zombies 15:175
Zone of proximal development (ZPD) 4:54, 4:856, 7:243, 8:731, 9:330, 10:418, 21:232, 21:363, 25:322
  coconstructivism 3:914
  and dynamic assessment 25:326
  validation of 12:253
  Vygotsky on 25:314, 25:317–320
Zones of cultural bilingualism 9:464
Zones of theory 15:354–355
Zoning
  housing 11:244
  and LGBTI commercial sex 13:885
Zoning laws 22:703
Zoning ordinances 3:205
Zooarchaeology 7:694, 25:849–853
  applied zooarchaeology 25:852–853
  fauna 25:849–853
  and food production 9:304
  foraging theory 25:850–851
  horizontal applications 25:851–852
  resource depression 25:851
  taphonomy 25:852
  vertical applications 25:849–851
Zoophilia 21:752
Zorbaugh, Harvey W., works, *The Gold Coast and the Slum* 7:561
Zorkin, Valery 13:432
Zorlu, Niyazi 20:873
z-test 11:496
z-transformation of Pearson's product moment correlation 11:218
Zuccotti Park 3:582
Zucker, Lynne 17:393
Zuckerberg, Mark 16:801–802
Zuckerman, M., alternative five factorial model 2:562
Zuckerman–Kuhlman Personality Questionnaire (ZKPQ) 21:608
Zuckerman–Kuhlman–Aluja personality questionnaire 21:608
Zukin, Sharon, on culture and consumption 5:538
Zuleta, León 13:860
Zulu regionalism 16:268
ZUMA *see* Zentrum fuer Umfragen, Methoden, und Analysen (ZUMA)
Zurich Model of Social Motivation 15:947–949, 15:947f
  empirical evidence 949–950
Zürich-Montpellier school 7:792
Zweigert, Konrad 13:804
Zwingli, Huldrych, on Reformation/confessionalization 20:100–101

This page intentionally left blank

AUTHOR INDEX

A

Aalbers, Manuel B.
Cities and the Financial Crisis

Abbott, Andrew
Event Sequence Analysis: Methods
Professions, Sociology of

Abélès, Marc
Governmentality

Abell, Peter
Structural Balance Theory: Recent Developments

Abou-Abdallah, Maria
Essentialism

Abramiuk, Marc A.
Cognitive Archaeology

Abramovitch, Henry
Death, Anthropology of

Abrams, Jules C.
Training in Clinical Psychology in the United States: Practitioner Model

Abramson, Paul R.
Generations: Political

Abts, Koen
Populism

Abubakar, Amina
Equivalence and Transfer Problems in Cross-Cultural Research

Abye, Tassé
Child Soldiers as Victims and Actors: Lessons from Ethiopia

Acee, Taylor W.
Learning to Learn

Adair-Toteff, Christopher
Alienation, Sociology of
Asceticism
Charisma, Social Aspects of

Adams, Julia
Historical Sociology

Adams, Rebecca G.
Subculture, Sociology of

Adamski, Beth A.
Sexual Dysfunction, Epidemiology of

Adan, Ana
Chronotype

Adelman, Jeremy
Hirschman, Albert O. (1915–2012)

Adolph, Karen E.
Gibson's Theory of Perceptual Learning

Agarwal, Renu
Innovation

Aggarwal, Neil K.
Transference in Psychoanalysis: Classical, Contemporary, and Cultural Contexts

Aggarwal, V.K.
Negotiation and Bargaining: Organizational Aspects

Agger, Ben
Oversharing: The Eclipse of Privacy in the Internet Age

Agger, Charlotte A.
Gender and Academic Motivation

Agnew, Robert
Strain Theories and Crime

Agras, William S.
Psychological Treatments: Randomized Controlled Trials

Agresti, Alan
Multivariate Analysis: Discrete Variables (Overview)

Ahsanullah, Mohammad
Nonparametric Statistics: Records

Aiken, Linda H.
Nonphysician Health Professionals

Akin-Little, Angeleque
Stress in Childhood: Effects on Development

Akkari, Abdeljalil
Education in the Middle East and North Africa

Al Shohaib, Saad
Religion and Health: Clinical Considerations and Applications

Al Zaben, Faten N.
Religion and Health: Clinical Considerations and Applications

Al-Azmeh, A.
Civilization, Concept and History of

Albiston, Catherine R.
Law as an Instrument of Social Change

Albro, Robert
Anthropological Research, Ethics of

Aldrich, John
Pearson, Karl (1857–1936)

Alexander, Catherine
Central Asia, Anthropology of

Alexander, Jeffrey A.
Hospitals and Health: Psychosocial Aspects

Alexander, Victoria D.
Markets: Artistic and Cultural

Alexy, Robert
Law and Morality: A Continental European Perspective

Algina, James
Psychometrics: Classical Test Theory

Alho, Juha M.
Population Forecasts

Alho, Kimmo
Sociodigital Revolution: Digital Natives vs Digital Immigrants

Allan, Lorraine G.
Time Perception Models

Allemand, Mathias
Midlife Psychological Development

Allen, Garland E.
Eugenics as a Basis of Population Policy
Eugenics as an International Movement

Allen, Jim
Australia and New Guinea, Archaeology of

Allen, Kaylie T.
Dyslexias, Acquired and Agraphia

Allen, Richard J.
Memory Binding

Almeida (Vasconcellos), Regina Araujo de
Tactile Maps in Geography

Almond, G.A.
Attitudes, Political and Civic Culture

Alpert, Geoffrey P.
Police Use of Force

Alston, Margaret
Rural Social Work

Altamirano, Carlos
Sarmiento, Domingo Faustino (1811–88)

Altmann, Jeanne
Primate Behavior: Significance for Understanding Humans

Altrichter, Herbert
Schools, Micropolitics of

Altschuld, James W.
Needs Assessment

Alvergne, Alexandra
Evolutionary Medicine

Alwin, Duane F.
Reliability and Validity Assessment: New Approaches

Aman, Alfred C.
Globalization: Legal Aspects

Amick, Kalynn B.
Transition to Adulthood, Cultural Differences in

Amin, S.
National Liberation Movements

Amit, Raphael
Business Models

Amit, Vered
Youth Culture, Anthropology of

Ampofo, Akosua Adomako
Feminisms and Acculturation around the Globe

Amsel, Eric
Logical and Hypothetical Reasoning in Adolescence, Development of

Anderman, Eric M.
Motivation, Learning, and Instruction

Andersen, Per K.
Event History Analysis in Continuous Time
Survival Analysis: Introduction

Andersen, Torben J.
Strategic Adaptation

Andersen, Torben M.
Welfare Reform

Anderson, Craig A.
Aggression, Social Psychology of

Anderson, Eric
Gays, Lesbians, and Sports: No Longer on the Sideline

Anderson, Jane E.
Indigenous Knowledge and Intellectual Property Rights

Anderson, Karen M.
Pension Policy

Anderson, Margo
Censuses: History and Methods

Anderson, Siwan
Dowries and Other Marital Transfers

Anderson, Warwick
Postcolonial Science Studies

Anderson-Levitt, Kathryn
Anthropological Approaches to the Study of Education: The United States and Beyond

Andersson, Matthew A.
Motivation: Life Course and Sociological Perspectives

Anderton, Douglas L.
Hutterite Demography
Social Epigenetics

Andeweg, Rudy B.
Consociationalism

Andler, Daniel
Naturalism

Andorfer, Veronika A.
Ethical Consumers

Andreoni, James
Charity and Philanthropy, Economics of

Andreß, Hans-Jürgen
Statistical Analysis Packages

Andres, Lesley
Academic Careers in Comparative Perspective

Andresen, Martin A.
Spatial Dynamics and Crime

Andrews, Gavin J.
Aging and Older People, Geography of

Andrews, Hazel
Liminality

Andrighetto, Luca
Dehumanization

Ang, Soon
Cultural Intelligence and Competencies

Angelo, Dante
South America, Archaeology of

Angenendt, Jörg
Somatoform Disorders

Angerer, Peter
Occupational Stress Management

Angermuller, Johannes
Discourse Studies

Anheier, Helmut K.
Social Science Infrastructure: North America (Resourcing and Funding)
Survey Research Centers and Companies

Anheier, Helmut
Policy Knowledge: Foundations

Ann-Mari, Sellerberg
Fashion, Sociology of

Ansolabehere, Stephen
Political Advertising

Antle, Michael C.
Sleep: Neural Systems

Antony, Martin M.
Panic Disorder

Anyamba, Assaf
Remote Sensing

Anzola, David
Self-Organizing Dynamical Systems

Apel, Robert
Deterrence: Sanction Perceptions
Deterrent Effect of Police and Prisons

Appadurai, Arjun
Globalization, Anthropology of

Appelbaum, Paul S.
Informed Consent of Research Subjects

Aragon, Oriana R.
Social Relationships in Adulthood

Aramburu, Leandro
Benhabib, Seyla (1950–)

Arbore, Alessandro
Marketing Strategies

Arce, Alberto
Global Development: Markets and Wellbeing

Ardila, Ruben
Behavior Analysis, Applied

Arens, William
East Africa: Sociocultural Aspects

Argote, Linda
Behavioral Theories of Organization

Argyrou, Vassos
Magic, Anthropology of

Arjas, Elja
Longitudinal Data: Event–History Analysis in Discrete Time

Armenta, Christina N.
Subjective Wellbeing, Psychology of

Armero, Carmen
Queues

Armstrong, Paul
School Business Management

Arnaud, André-Jean
Equity in Law

Arnfred, Signe
Heterosexuality in Africa

Arnold, Aiden E.
Human Spatial Orientation, Neural Basis of

Arnold, Darrell P.
Critical Theory

Arnold, Thomas
Plato (428 BCE – 347 BCE)

Arnott, Richard
Housing Economics

Aro, Arja R.
Cancer Screening

Aro, Mikko
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Aron, Arthur
Interpersonal Attraction, Psychology of

Aronson, Jay D.
Forensic Genetic Databases: Ethical and Social Dimensions

Aronsson, Karin
Narrative, Sociology of

Arpawong, Thalida Em
Genetic Studies of Behavior: Methodology

Arrigo, Bruce A.
Critical Criminology

Arthurs, Alan J.
Job Design and Evaluation: Organizational Aspects

Artur, Lakatos
Confederations, History of

Arvey, Richard
Genetic Factors in Organizational Psychology

Asa Adjei, Edwin
Feminisms and Acculturation around the Globe

Ashcroft, Richard E.
Bioethics: Philosophical Aspects

Ashkanasy, Neal M.
Emotions and Work

Ashmore, Malcolm
Reflexivity in Science and Technology Studies

Aspromourgos, Tony
Profits

Assar, Saïd
Information and Communications Technology in Education

Atkinson, Anthony C.
Optimal Experimental Design

Atkinson, Leslie
Intellectual Functioning, Assessment of

Atkinson, Mark
Language Development: Evolutionary Theories

Attwood, Lynne
Russian and Post-Soviet Studies: Gender

Auchter, Allison M.
Limbic System

Augoustinos, Martha
Racism: Social Psychological Perspectives

Austin, Jacqueline L.
Expert Testimony

Averett, Susan L.
Obesity and Labor Market Performance

Aya, Rod
Revolution, Theories of

Ayonrinde, Oyedeji
Culture-Specific Psychiatric Disorders
Perversions, Sexual (Paraphilias)

Ayora-Diaz, Steffan I.
Food in Anthropology

B

Baba, Marietta L.
Business Anthropology

Babiarz, Kimberly S.
Family Planning: Program Effects

Babić, Staša
Archaeology, Theory in

Babiloni, Fabio
Neuromarketing

Bachman, Ronet
Data Bases and Statistical Systems: Crime Measurement

Badcock, Christopher R.
Nature-Nurture Controversy, History of

Bade, K.J.
Migration History

Badgett, M.V. Lee
Lesbians: Social and Economic Situation

Baer, Brian J.
Sexual Minorities in Modern Russia

Bagchi-Sen, Sharmistha
International Business: Geographic Aspects

Baghramian, Maria
Putnam, Hilary (1926–)
Relativism: Philosophical Aspects
Rorty, Richard (1931–2007)

Bagley, Constance E.
Business Law

Bai, Sunhye
Social and Emotional Development in the Context of the Family

Baicker-McKee, Steven
Discovery: Legal Aspects

Bailey, Adrian J.
Population Geography

Bailey, Annette
Ethnicity, Race, Oppression, and Social Work: The Canadian Case

Bailey, Di
Mental Health and Social Work

Bailey, Ian
Climate Change

Baird, Sarah
Conditional Cash Transfers: Influence on Marriage and Fertility

Bajo, Ana
Confabulation

Baker, Laura A.
Defiant Behavior During Adolescence across Cultures
Genetics of Personality

Balard, Frédéric
Old Age: Definitions, Theory, and History of the Concept

Balbo, Nicoletta
Fertility Theory

Baldessarini, Ross J.
Bipolar (Manic-Depressive) Disorder

Baldi, Philip
Indo-European Languages

Baldo, Juliana V.
Speech Production, Neural Basis of

Baldwin, Carliss Y.
Modularity and Organizations

Baldwin, David A.
Sanctions in Political Science

Bales, Kevin
Slavery as Social Institution

Bales, William D.
Recidivism

Balke, Tina
Self-Organizing Dynamical Systems

Balon, Richard
Psychotherapy and Pharmacotherapy

Baltes, Boris B.
Aging and Work

Baltzarek, Franz
Communication and Transportation, History of

Balzarini, John
Urban Sociology

Bamberg, Sebastian
Environmental Attitudes and Behavior: Measurement

Bamyeh, Mohammed
Near Middle East/North Africa Studies: Society and History

Bancroft, Tyler D.
Context Dependent Memory

Banda, Tinenenji
Sub-Saharan African Legal Traditions

Bandelj, Nina
Culture and Economy

Bandes, Susan A.
Law and Emotion

Banfield, Jillian C.
Prosocial Behavior and Empathy

Banks, David
Agent Based Modeling, Statistics of

Banks, James A.
Multicultural Education

Banks, James
Retirement and Health

Banks, Jeffrey S.
Politics, Formal Modeling of

Banks, Sarah
Social Work Ethics

Banks, William C.
Law and Terrorism in the US and Beyond

Bannister, Robert
Sumner, William (1840–1910)

Banzhaf, Wolfgang
Artificial Intelligence: Genetic Programming

Barak, Gregg
Social Justice and Social Inequalities

Barber, Jacques P.
Psychodynamic Therapy: Short-Term

Barber, William J.
Fisher, Irving (1867–1947)

Barbi, Elisabetta
Demographic Techniques: Inverse Projection

Barbier, Edward B.
Sustainable Development: An Economic Perspective

Barbieri, Elisa
Special Economic Zones and Cluster Dynamics: China

Barbieri, Magali
Infant and Child Mortality in the Less Developed World

Barbosa de Almeida, Mauro W.
Structuralism

Barbujani, Guido
Race: Genetic Aspects

Bardach, Eugene
Implementation: Political

Bardi, Ugo
Limits to Growth
Mineral Resources, Limits to: The Case of Peak Oil

Bardill, Jessica
Genetic Ancestry Testing

Barker, Eileen
New Religious Movements

Barlow, David H.
Manual Based Treatments

Barlow, Tani E.
East Asian Studies: Gender

Barnard, John
Multiple Imputation

Barr, Stewart
Sustainable Lifestyles

Barrett, Alan
Aging and the Labor Market

Barrett, Don
Lesbian, Gay, Bisexual and Trans-sexual Minorities in International Perspective: Overview

Barrett, Harold C.
Cognition, Evolution of

Barrett, Jennifer
Museums

Barrett, Kimberly L.
Social Justice and Criminal Justice

Barrett, Louise
Fertility Theory: Theory of Life History Evolution

Barrie-Anthony, Steven
Christianity in the English-Speaking West

Barron, Kenneth E.
Expectancy-Value-Cost Model of Motivation

Barron, Tammy
School to School Collaboration

Barroso, Carmen
Family Planning Programs: Feminist Perspectives

Barrow, Chris J.
Environment and Development

Barrow, Logie
Socialism: Historical Aspects

Barry, John
Environmental Education and Education for Sustainable Development

Barshay, Andrew E.
Maruyama, Masao (1914–96)

Barsky, Robert
Chomsky, Noam (1928–)

Barstead, Matthew G.
Friendship During Infancy and Early Childhood and Cultural Variations

Bartels, Larry M.
Attitudes, Political and Public Opinion

Bartholomew, David J.
Factor Analysis and Latent Structure Analysis: Overview
Stochastic Models

Barto, Daniel
Navigation in Virtual Space: Psychological and Neural Aspects

Bartolomeo, Paolo
Visual Imagery

Bartos, Sebastian E.
Social Psychological Theory, History of

Bartsch, Thorsten
Transient Global Amnesia

Baruch, Yehuda
Career Development: Multilevel Perspective

Basbøll, Hans
Morphophonology

Basedow, Jürgen
European Union Law

Basu, Arnab K.
Informal Work in Developing Countries

Baszanger, Isabelle
Strauss, Anselm (1916–96)

Batchelder, William H.
Discrete-State Models of Information Processing
Mathematical Psychology: History

Battaglia, Lisa J.
Women in Buddhism

Battistella, Edwin
Markedness in Linguistics

Baubérot, Jean
Laicization, History of

Bauer, Patricia J.
Short- and Long-term Memory during Infancy and Early Childhood across Cultures

Bauer, Talya N.
Leader–Member Exchange Theory

Baughman, James L.
Television: History

Bauldry, Shawn
Indicator: Methodology
Structural Equation Modeling

Baum, Howell S.
Citizen Participation

Baum, Joel A.C.
Ecology: Organizations

Bauman, Richard
Tradition, Anthropology of

Baumgartner, Frank R.
Agendas: Political

Bausch, Kenneth C.
Luhmann's *Social Systems*: Meaning, Autopoiesis, and Interpenetration

Bayarri, Susie
Queues

Baysu, Gülseli
Ethnicity and Migration in Europe

Beauregard, Robert
Shrinking Cities

Beck, Elizabeth
Orientalism, Oppression, and Contemporary Social Work

Beck, Johannes
Sleep Disorders: Psychiatric Aspects

Beck, Margaret
Ethnoarchaeology

Becker, Karin
Photography as a Medium

Beckett, Jeremy
Indigeneity

Beckford, James A.
Religious Organizations

Beddoe, Liz
Professional Development

Beebee, Helen
Causes and Laws: Philosophical Aspects

Beelmann, Andreas
Evidence-Based Psychotherapy

Beer, Bettina
Friendship, Anthropology of

Beesdo-Baum, Katja
Mental Disorders, Epidemiology of

Behler, Rachel
Sexual Behavior and Social Networks

Behrman, Jere R.
Twin Studies in Demography

Behrmann, Marlene
Hemispatial Neglect, Neural Basis of

Beichelt, Timm
Revolutions of 1989–90 in Eastern Central Europe

Bekkers, René
Altruism and Prosocial Behavior, Sociology of

Beldo, Les
Culture: Contemporary Views

Bell, David N.F.
Youth Joblessness

Bell, D.S.
Political Parties: Communist

Bell, Jordana T.
Twin Studies and Epigenetics

Bell, Nathan E.
Social Science Professions and Professionalization

Bellah, Robert N.
Religion: Evolution and Development

Bellamy, Richard P.
Liberalism: Impact on Social Science

Bellamy, Richard
Bobbio, Norberto (1909–2004)
Citizenship, Historical Development of
Liberalism: Political Doctrine and Impact on Social Science

Bellino, Michelle J.
Conflict: Education and Youth

Bellomo, Antonello
Schizophrenia: Metabolic Consequences

Bellugi, Ursula
Sign Language

Bellwood, Peter
Cultural Evolution: Phylogeny versus Reticulation
Language Families, Archaeology and History of

Benatar, Solomon R.
Health in Low-Income Countries

Benavides, Gustavo
Christianity in Latin America

Bender, Andrea
Cognitive Anthropology

Bendix, John
Bendix, Reinhard (1916–91)

Bendor, John
Bounded Rationality

Benedetti, Fabrizio
Placebo Effect

Benedict, Kennette
Global Governance

Benfenati, Fabio
Synaptic Mechanisms and Protein Synthesis in Memory

Benford, Robert D.
Dramaturgical Analysis

Bengtsson, Tommy
Mortality: The Great Historical Decline

Benjamin, Akua
Ethnicity, Race, Oppression, and Social Work: The Canadian Case

Benkert, O.
Antidepressant Drugs

Bennett, Andrew
Case Study: Methods and Analysis

Bennett, Caroline
Genetics and Disaster Victim Identification

Bennett, Rebecca
Workplace Deviance

Bennie, Garth
Learning Disability and Social Work

Ben-Rafael, Eliezer
Ethnicity, Sociology of

Bensch, Doreen
Personality Assessment, Faking and

Ben-Shlomo, Yoav
Life Course Epidemiology

Bensmaia, Sliman
Neuroprosthetics

Benson, David
Environmental Policy: Protection and Regulation

Benson, Michael L.
White Collar Crime: Recent Trends and Debates

Benwell-Lybarger, Jerian
Violence, Criminalization, and the Homeless

Ben-Ze'ev, Aaron
Satisfaction and Commitment in Long-Term Heterosexual Relationships
Sexuality and the Internet

Berend, Ivan T.
Capitalism

Berends, Mark
School Choice

Beretta, Alan
Linguistic Approaches to Agrammatism

Berez, Andrea L.
Linguistic Fieldwork

Berezin, Mabel
Culture, Sociology of

Bergene, Ann C.
Labor, Geography of

Berger, Roger L.
Estimation: Point and Interval
Hypothesis Testing in Statistics

Bergey, Meredith
Medicalization: Sociological and Anthropological Perspectives

Berghoff, Christopher R.
Single-Case Designs: Methodology

Berghoff, Hartmut
Business History

Berg-Schlosser, Dirk
Comparative Studies: Method and Design

Bergsmann, Evelyn
Program Evaluation

Berk, Bernard B.
Labeling Theory, History of

Berliner, L. Mark
Spatial Statistical Methods

Berman, R.A.
First Language Acquisition, Developmental Psychology of

Bernard, Philippe
Stereotypes in Social Psychology

Bernard, Wanda T.
Social Work with Racialized Groups: Frameworks for Practice

Berndtson, Erkki
Political Science, History of

Bernstein, Carol Ann
Women and Psychiatry

Bernstein, Henry
Agrarian Political Economy

Bernstein, Jeanne Wolff
Lacan, Jacques (1901–81)

Berntsen, Leslie
Defiant Behavior During Adolescence across Cultures

Berreman, Gerald D.
Inequality: Comparative Aspects

Berridge, Kent C.
Wanting vs Needing

Berry, Donald A.
Sequential Statistical Methods

Bersot, Heather Y.
Critical Criminology

Bertolatti, Dean
Environmental Health and Safety: Social Aspects

Besnard, Philipe
Anomie

Besnier, Niko
Oceania, Sociocultural Overviews: Polynesia and Micronesia

Best, Deborah L.
Gender Development: Cultural Differences

Best, Donald J.
Statistical Distribution Models: Goodness of Fit Tests

Bester, Suzanne
Emotional Development, Effects of Parenting and Family Structure on

Besthorn, Fred H.
Ecological Social Work: Shifting Paradigms in Environmental Practice

Bestor, Theodore
Markets: Anthropological Aspects

Beteille, Andre
Asia, Sociocultural Overviews: South Asia

Beyer, Peter
Globalization and Religion

Beyme, Klaus von
Parliamentary Government

Bhambra, Gurminder K.
Modernity: History of the Concept

Bhatta, Tirth
Age Stratification

Bhatti-Sinclair, Kish
Culturally Appropriate Interventions in Social Work

Bhimani, Alnoor
Management Accounting

Bhugra, Dinesh
Cultural Psychiatry
Mentally III, Public Attitudes towards
Schizophrenia: Metabolic Consequences

Bibel, Wolfgang
Deductive Reasoning Systems

Bickenbach, Jerome E.
Functioning, Disability and Health, International Classification of

Bickford, Andrew
Militaries and Militarization, Anthropology of

Biehl, Peter F.
Europe, Archaeology of

Bielby, Denise D.
Video Cultures: Television

Biernacki, Richard
Industrialization

Bierwisch, Manfred
Bloomfield, Leonard (1887–1949)
Generative Grammar

Biesta, Gert
Educational Philosophy

Bigman, Yochanan
Emotional Regulation

Bijker, Wiebe E.
Technology, Social Construction of

Bilgrami, Akeel
Identity and Identification

Billard, Lynne
Markov Models and Social Analysis

Bilsky, Leora
Political Trials

Binder, David
Data Bases and Statistical Systems: Culture

Binion, Rudolph
Psychohistory

Birbaumer, Niels
Fear Conditioning: Overview

Bird-David, Nurit
Hunting and Gathering Societies: Anthropology

Birnbacher, Dieter
Euthanasia

Bisang, Walter
Parts of Speech

Bisk, Tsvi
Attempted Utopias and Intentional Communities
Literary and Philosophical Utopian Thought

Bittles, Alan H.
Incest, Inbreeding, and Their Consequences

Bizumic, Boris
Ethnocentrism and Prejudice: History of the Concepts

Bizzi, Lorenzo
Social Capital in Organizations

Bjorklund, David F.
Evolutionary Developmental Psychology

Blacker, Coit D.
Arms Control

Blackman, Tim
Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice

Blackwell, Brenda S.
Intersectionality and Crime

Blair, James R.
Psychopathy

Blake, Barry J.
Case, Morphological: Linguistics

Blanchflower, David G.
Youth Joblessness

Blanck, Peter
Law and People with Disabilities

Blangero, John
Familial Studies: Genetic Inferences

Blank, Michael B.
Violence as a Public Health Problem

Blase, Karen
Implementation Science

Blasius, Jörg
Correspondence Analysis, History of

Blass, Thomas
Obedience: Social Psychological Perspectives

Blatný, Marek
Personality Development: Systems Theories

Blau, Francine D.
Gender, Economics of

Bledsoe, Caroline H.
Fertility and Culture: Anthropological Insights

Bleidorn, Wiebke
Personality Differences and Development: Genetic and Environmental Contributions

Blenis, Robert C.
Autobiographical Memory during Early Childhood and Cultural Variations

Blewitt, John
Sustainability: A Dialog of Values

Blitz, Mark
Public Interest

Bloch, Sidney
Psychotherapy, History of: Psychiatric Aspects

Block, Steven
Victimless Crime

Bloemen, Hilde
Radical Social Work

Blok, Anton
Mafia, Anthropology of

Blomley, Nicholas
Law and Geography

Blomquist, Glenn C.
Value of Life, Economics of

Blomqvist, Åke G.
Health Economics

Blondel, Jean
Governments

Bloor, David
Strong Program, Sociology of

Blossfeld, Gwendolin J.
Life Course and Event History Analysis

Blossfeld, Hans-Peter
Event History Analysis: Applications
Life Course and Event History Analysis

Blumenthal, Terry D.
Startle Reflex and Health

Blusiewicz, Tomasz
Eastern European Studies: History

Bo, Chiara D.
Smart Cities

Boal, Frederick W.
Ethnic Conflict, Geography of

Boardman, Anthony E.
Cost–Benefit Analysis

Boardman, Jason D.
Gene–Environment Interplay

(Bob) Hinings, C.R.
Professions in Organizations

Bock, John A.
Fertility Theory: Embodied-Capital Theory of Life History Evolution

Böckenholt, Ulf
Psychometrics: Preference Models with Latent Variables

Boczkowski, Pablo J.
New Media, News Production and Consumption

Boddewyn, Jean
Advertising, Control of

Bode, Christina
Rheumatoid Arthritis: Psychosocial Aspects

Bodenmann, Guy
Illness and Dyadic Coping

Boerner, Kathrin
Bereavement

Boese, Gregory D.
Meritocracy and Tokenism

Boettcher, Hannah
Social Phobia across the Lifespan

Boettke, Peter J.
Alternative Schools of Economic Thought

Bogaert, Anthony F.
Sexual Orientation: Biological Influences

Bogaerts, Katleen
Symptom Perception, Awareness and Interpretation

Bogart, Daniel
False Memories

Bognanno, Michael
Promotions

Bohnemeyer, Jürgen
Deixis

Boissevain, Jeremy
Factionalism

Boli, John
Globalization and World Culture

Bollen, Kenneth A.
Indicator: Methodology

Bolt, Daniel M.
Surveys: Extreme Response

Bonanno, Alessandro
Globalization and Transnational Corporations

Boncourt, Thibaud
Associations, Scholarly
Social Science Infrastructure: European Union (Social Science Academies and Related Organizations)

Bondebjerg, Ib
Film: Genres and Genre Theory

Bonds, Anne
Poverty, Geography of

Bongaarts, John
Fertility Change: Quantum and Tempo

Bonikowski, Bart
Nation-State as Symbolic Construct

Boniolo, Giovanni
Deliberation and Democracy

Bonsal, Dudley B.
Geographic Information Systems and Remote Sensing

Booth, Alison
Trade Unions, Economic Behavior of

Booth, Heather
Demographic Techniques: Data Adjustment and Correction

Booth, Julie L.
Cognitive Development: Mathematics Learning and Instruction

Bordone, Valeria
Intergenerational Solidarity

Borg, Ingwer
Multidimensional Scaling II

Borgatti, Stephen P.
Structural Equivalence: Meaning and Measures

Borges, Walt
Mass Media and Politics

Borghi, Anna M.
Embodiment Theory

Borjas, George J.
Migration, Economics of

Borkenau, Peter
Developmental Behavioral Genetics and Education
Personality, Trait Models of

Borman, Walter C.
Personnel Selection, Psychology of

Borneman, John
State: Anthropological Aspects

Borner, Kelsey B.
Friendship During Childhood and Cultural Variations

Bornstein, Marc H.
Infancy and Human Development

Borsboom, Denny
Psychometrics

Bortolini, Matteo
Bellah, Robert N (1927–2013)
Communitarians

Boruch, Robert F.
Experimental Design: Randomization and Social Experiments
Privacy of Individuals in Social Research: Confidentiality

Bös, Mathias
Ethnicity and Ethnic Groups: Historical Aspects

Bosi, Lorenzo
Social Movements and Political Violence

Bösser, Tom
Autonomous Agents

Bosworth, Richard J.B.
Second World War, The

Bottjer, Sarah W.
Birdsong and Vocal Learning During Development

Boucher, Gerard
Walzer, Michael (1935–)

Bouchet, Dominique
Levi-Strauss, Claude (1908–2009)
Smith, Adam (1723–90)

Boudreaux, Michael J.
Five Factor Model of Personality, Assessment of

Bouquet, Mary
Genealogy in Anthropology

Bourbeau, Philippe
Securitization

Bourgois, Phillippe
Poverty, Culture of

Boutsen, Frank R.
Asymmetry of Body and Brain

Bowers, Edmond P.
Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations

Bowker, Geoffrey C.
Science and Technology, Social Study of: Computers and Information Technology

Bowlby, Sophie
Social Geography

Bowler, Peter J.
Evolution, History of

Boyce, Gordon
Maritime Networks, History of

Boyer, Dominic
Energy, Anthropology of
Expertise, Anthropological Approaches to

Boyer, Joy T.
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

Braak, Heiko
Alzheimer's Disease, Neural Basis of

Braddick, Oliver
Occipital Lobe (Visual Cortex): Functional Aspects
Visual Perception, Neural Basis of

Braddon-Mitchell, David
Intentionality and Rationality

Bradford, Ben
Compliance and Legal Authority

Bradford, John
Forensic Psychiatry

Bradley, Mindy S.
Sentencing, Trends, and Disparities: Crime

Brainerd, Elizabeth
Fertility Trends in the Formerly Socialist Countries of Europe

Branch, Nicholas
Environmental Archaeology

Brand, Karl-Werner
Sustainable Development

Brandell, Jerrold R.
Social Work and Psychodynamic Thought

Brandes, Ulrik
Social Network Algorithms and Software

Bratter, Jenifer
Population Composition by Race and Ethnicity: North America

Braun, Susanne
Group Processes in Organizations

Braunstein, Juergen
Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks

Breakwell, Dame G.M.
Risk: Social Psychological Perspectives

Breckner, Roswitha
Biography and Society

Bredekamp, Horst
Art History

Breen, Richard
Selection Bias, Statistics of

Breidenbach, Joana
Intercultural Communication: An Anthropological Perspective

Breiger, Ronald L.
Culture and Networks

Brennan, Jonathan
Neurolinguistic Processing of Psychological Verbs

Brenner, M.
Historical Thought and Historiography: Jewish Tradition

Brettell, Caroline B.
Migration: Anthropological Perspectives

Breuilly, John
Nations and Nation-States in History

Brewer, John D.
Peace Processes

Brewis, Alexandra A.
Obesity: Cultural and Biological Factors

Brice, Alejandro E.
Multilingual Language Development

Brick, Howard
Bell, Daniel (1919–2011)

Briggs, Charles L.
Discourse, Anthropology of

Brillinger, David R.
Time Series: General

Briskman, Linda
Asylum Seekers

Brisson, Thomas
Western and Non-Western Views of the Social and Behavioral Sciences: De-Westernizing the Social and Behavioral Sciences

Britt, Chester L.
Control Theories and Crime

Brittian, Aerika S.
Prosocial Behavior during Adolescence

Brock, Timothy
Urban Entrepreneurialism

Brodbeck, Felix Claus
Group Processes in Organizations

Brody, Lawrence C.
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

Brokensha, David W.
Development: Social-Anthropological Aspects

Broman, Clifford L.
Race Identity

Bromley, Patricia
New Institutionalism in the Analysis of Complex Organizations

Bronisch, Thomas
Suicide

Brooks, John I.
Janet, Pierre (1859–1947)

Brooks, Rechele
Infant Gaze Behaviors

Broschart, Jüergen
Possession, Linguistic

Broughman, Brian J.
Law and Economics: Empirical Dimensions

Broughton, Jack M.
Zooarchaeology

Brown, Anna
Personality Assessment, Forced-Choice

Brown, Chelsea N.
Social Learning Theory and Crime

Brown, Eleanor
Stigma, Social Psychology of

Brown, Gordon D.A.
Temporal Distinctiveness Models of Memory

Brown, Kevin L.
Eyeblink Classical Conditioning

Brown, Lawrence A.
Diffusion: Geographical Aspects

Brown, Mitchell
Gender and Women's Studies, Applied Research on

Brown, Penelope
Politeness and Language
Space, Linguistic Expression of

Brown, Richard E.
Olfactory System

Brown, Susan L.
Family Theory: Competing Perspectives in Social Demography

Brownell, Celia A.
Prosocial Behavior during Childhood and Cultural Variations
Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations

Browning, Christopher
Adolescent Sexual Risk

Bru, Bernard
Laplace, Pierre Simon, Marquis de (1749–1827)

Bruegmann, Robert
Urban Sprawl

Brugha, Traolach
Mental Health

Brulle, Robert J.
Environmental Movements

Brumann, Christoph
Cultural Heritage

Brumfiel, E.M.
States and Civilizations, Archaeology of

Brumley, Benjamin
Experimental Design: Randomization and Social Experiments

Brusco, Michael
Psychometrics: Combinatorial Data Analysis

Bryant, Peter T.
Entrepreneurship and Organizations

Bryde, Brun-Otto
Constitutional Courts

Brym, Robert J.
Intellectuals, Sociology of

Bryson, John M.
Strategic Planning for Public and Nonprofit Organizations

Bryson, John R.
Entrepreneurship, Geography of

Bucerius, Sandra M.
Crime, Immigration, and Ethnicity

Buchanan, Elizabeth A.
Internet Research: Ethical Concerns

Buchanan, Molly
Life Course Criminology

Buchmann, Marlis
Emergent Properties

Buck, Andrew
Russian and Post-Soviet Studies: Society

Buckingham, Hugh W.
Phological Disorders

Buckingham, Mary H.
Childhood and Adolescence: Developmental Assets

Buckingham, Sarah S.
Phological Disorders

Buckingham, Susan
Ecofeminism
Sustainable Development and Social Work

Buckler, Kevin
Media, Crime and Crime Waves

Budge, Ian
Political Parties: Manifestoes

Buffalo, Elizabeth A.
Declarative Memory, Neural Basis of

Bullard, Robert D.
Environmental Justice in the United States

Bunge, Mario
Sociology, Epistemology of

Buntin, Jennifer T.
Intimate Partner Violence

Burgelman, Robert A.
Strategic Management

Burgess, Joanne C.
Sustainable Development: An Economic Perspective

Burian, R.
Liaison and Consultation Psychiatry

Burke, Peter
Freyre, Gilberto (1900–87)

Burkes, Kaleena J.
Recidivism

Burles, Ford
Human Spatial Orientation, Neural Basis of

Burnard, Trevor
Slaves and Slavery, History of

Burnham, Philip
Clan

Burnham, Walter D.
Key, Valdimer Orlando (1908–63)

Burr, Viv
Social Constructionism

Burris, Arthur
Federalism

Burns, Stacy L.
Garfinkel, Harold (1917–2011)

Burrus, Jeremy
Social Beliefs and Attitudes: Personality and the Isms

Burt, Martha R.
Homelessness in the United States

Burton, John F.
Safety, Economics of

Busch, Andreas
Internet and Privacy

Busch, Holger
Motivational Development: Cross-Cultural Perspectives

Busch, Lawrence
Agricultural Sciences and Technology

Busemeyer, Jerome R.
Dynamic Decision Making

Buskens, Vincent
Rational Choice Theory in Sociology

Buss, David M.
Human Mate Choice, Evolution of

Butcher, James N.
Minnesota Multiphasic Personality Inventory

Butera, Federico
Automation: Organizational Studies

Buyx, Alena
Bioethics in the Post-genomic Era

C

Cacioppo, John T.
Social Neuroscience of the Twenty-First Century

Cacioppo, Stephanie
Social Neuroscience of the Twenty-First Century

Caffrey, Margaret M.
Mead, Margaret (1901–78)

Cagigas, Xavier E.
Luria, Aleksandr Romanovich (1902–77)

Caglar, Ayse
Citizenship, Anthropology of

Cahan, David
Helmholtz, Hermann Ludwig Ferdinand von (1821–94)

Cahill, Caitlin
Participatory Action Research in Social Research

Caillods, Françoise
Educational Planning Worldwide

Caldwell, Brian J.
School Management

Caldwell, John C.
African Studies: Health
Notestein, Frank Wallace (1902–83)

Calhoun, C.
Public Sphere: Nineteenth and Twentieth-Century History

Calhoun, Craig
Civil Society and the Public Sphere: History of the Concept

Callahan, Marilyn
Risk Assessment in Social Work

Calot, Gérard
Demographic Techniques: LEXIS Diagram

Calvert, Kirby
Energy and Society

Cameron, Jessica J.
Self-Esteem

Cameron, Linda D.
Chronic Illness, Psychosocial Coping with

Cameron-Bradley, Mahla
Spatial Cognition during Infancy and Early Childhood Across Cultures, Development of

Camic, Charles
Habit: History of the Concept

Camp, Joanne S.
Williams Syndrome

Campanini, Annamaria
Bologna Process

Campbell, Cameron D.
Demographic Techniques: Family Reconstitution

Campbell, Colin J.
Oil Age: Energy in Transition

Campbell, Heidi A.
Religion and New Media

Campbell, Lyle
Areal Linguistics
Comparative Method in Linguistics

Campbell, Tom
Law and Morality: An Analytical Perspective

Canboy, Basak
Leadership Education: Theory and Practice

Canclini, Nestor G.
Hybridity

Canessa, Nicola
Neuroeconomics

Canfield, Matthew C.
Law: Anthropological Aspects

Canning, Elizabeth A.
Academic Motivation and Performance: Task Value Interventions

Canning, Kathleen
Gender History

Canning, Roy
Education: Skills Training

Cantó-Milà, Natàlia
Simmel, Georg (1858–1918)

Canudas-Romo, Vladimir
Mortality of the Oldest-Old

Caplan, David
Aphasia

Cappa, Stefano F.
Neuroeconomics

Cappelleri, Joseph C.
Regression Discontinuity Design

Caprara, Gian V.
Personality and Adaptive Behaviors

Caragliu, Andrea
Smart Cities

Carbone, June
Family Law in the United States

Card, Josefina J.
Archiving: Ethical Aspects

Carey, Martha Ann
Focus Groups

Carley, Kathleen M.
Computational Organizational Theory

Carlisle, Brandon L.
Academic Performance, Effects of Socio-Economic Status on

Carlsmith, Kevin M.
Cognitive Dissonance

Carlson, Elof A.
Muller, Hermann Joseph (1890–1967)

Carlson, James D.
Authentic Leadership

Carmines, Edward G.
Issue Evolution

Carmon, Naomi
Neighborhood

Carnes, Bruce A.
Old Age and Centenarians: The Human 'Warranty Period'

Carney, Lauren
Social Beliefs and Attitudes: Personality and the Isms

Carpenter, Kristen A.
Indigenous Rights

Carr, Colleen M.
Marital Status, Divorce, and Child Development

Carr, Deborah
Marital Quality and Health

Carranza, Mirna E.
Trafficking of Women and Children in Latin America

Carré, Françoise J.
Working Conditions

Carrington, Paul D.
Appeals: Legal

Carriquiry, Alicia L.
Measurement Error Models

Carroll, David W.
Language Development, Theories of

Carroll, Glenn R.
Organizational Ecology

Carson, John
Intelligence: History of the Concept

Carstensen, Laura L.
Emotions and Aging

Carter, Neil
Green Parties

Caruana, Fausto
Embodiment Theory

Carugati, Felice
Learning and Instruction: Social-Cognitive Perspectives

Caruso, David R.
Personal Intelligence and Competencies

Carver, Charles S.
Depression, Pessimism, and Health
Optimism and Pessimism

Carver, Terrell
Marx, Karl (1818–83)

Casanova, Jose
Secularization

Cascio, Wayne F.
Human Resource Management, Psychology of
Industrial–Organizational Psychology: Science and Practice

Caselli, Graziella
Mortality, Epidemiological, and Health Transitions

Casetti, Emilio
Spatial–Temporal Modeling

Casey, Beth
Child Neglect

Cassel, Pär
East Asian Legal Traditions

Castaldelli-Maia, João M.
Obsessive-Compulsive Disorder

Castelli, Fulvia
Mind, Theories of

Casterline, John B.
Fertility Transition: Middle East and North Africa

Castillo, Fatima A.
Gender and Social Research

Castle, David
Prophylaxis in Psychiatry

Castro, Antonio J.
Social Studies Education

Catalano, Theresa
Foreign Language Teaching and Learning

Catani, Claudia
Mass Trauma: Psychopathological Effects across the Life Span

Cathy Li, Zongchao
Media Effects

Cattaneo, Zaira
Mental Imagery: Visual Cognition

Cattelino, Jessica
North America: Sociocultural Aspects

Cavagnaro, Daniel R.
Computational Approaches to Model Evaluation

Cedillos, Elizabeth M.
E-Communications Platforms and E-Learning

Cepanec, Maja
Executive Functions during Childhood, Development of

Cetina, Karin K.
Epistemic Cultures

Chadwick, Ruth
Bioethics: Genetics and Genomics

Chakrabarti, Sandip
Transportation Geography

Chamallas, Martha
Gender and the Law

Chambers, Jack K.
Dialectology

Champine, Robey B.
Childhood and Adolescence: Developmental Assets

Chan, David W.
Education for the Gifted and Talented

Chan, Kam Wing
Urban Places: Statistical Definitions

Chan, Ngai H.
Time Series: Unit-Roots and Nonstandard Limiting Distributions

Chandola, Tarani
Retirement and Health
Social Exclusion, Social Deprivation and Health

Chang, Chu-Hsiang (Daisy)
Core Self-Evaluations

Chang, Ruth
Value Pluralism

Charle, Christophe
Intellectuals: History of the Concept
Prosopography (Collective Biography)

Charles, Camille Z.
Residential Segregation in the United States

Charmaz, Kathy
Grounded Theory: Methodology and Theory Construction

Charme, Cameron
Sexuality Over the Life Course

Chartier, Roger
Cultural History

Chartier, R.
Public Sphere: Eighteenth-Century History

Chase, Paul A.
Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations

Chase-Dunn, Christopher
Dependency Theory

Chatman, Jennifer A.
People in Organizations

Chatterjee, Partha
Nationalism: General
Subaltern History

Chaturvedi, Santosh K.
Hysteria: History and Critiques
Psychosomatic Medicine
Somatotherapy, History of

Chau, Nancy H.
Informal Work in Developing Countries

Chauvel, Louis
Social Epidemiology

Cheak-Baillargeon, Alicia
Leadership in Organizations, Sociology of

Cheek, Dennis W.
Charity and Philanthropy: Overview
Education and Economic Growth

Chelberg, Marie B.
Interpersonal Attraction, Psychology of

Chen, Anthony S.
Affirmative Action, Sociology of

Chen, Chih-Chin
McDonaldization

Chen, Haidan
Human Genomics in Asia

Cheng, Jing
Compliance

Cheng, Peter C.-H.
Scientific Discovery, Computational Models of

Chenoweth, Erica
Insurgency

Cherkaoui, Mohamed
Deprivation: Relative

Chermak, Steven M.
National Security and Domestic Terrorism: The US Case

Chester, Ronald
Property: Legal Aspects of Intergenerational Transmission

Cheung, Gordon C.K.
Social Science Infrastructure: Eastern Asia and Pacific (Social Science Academies and Related Organizations)

Cheung, Sum Kwing
Coeducation and Single-Sex Schooling

Chevalier, Sophie
Anthropology at Home

Chiari, Gabriele
Psychotherapy and Hermeneutics

Chib, Siddhartha
Monte Carlo Methods and Bayesian Computation: Overview

Chibber, Vivek
East Asian Studies: Economics

Chiesi, Antonio M.
Network Analysis

Chilton, Roland
Crime and Class

Chin, Lynn G.
Status and Role, Social Psychology of

Chirot, Daniel
World Systems Theory

Chiu, Ming M.
Cross-National Comparisons in Education: Findings from PISA

Chodorow, Nancy J.
Klein, Melanie (1882–1960)

Chong, Dennis
Political Protest and Civil Disobedience

Chopra, Vidur
Conflict: Education and Youth

Choueiri, Youssef M.
Islamic Fundamentalism

Chouinard, Vera
Disability, Geography of

Christ, C.F.
Simultaneous Equation Estimation: Overview

Christensen, Ronald
Analysis of Variance and Generalized Linear Models
Linear Hypothesis

Christiano, Thomas D.
Democracy: Normative Theory

Christias, Panagiotis
Aron, Raymond (1905–83)

Chryssochoou, Xenia
Social Psychology

Chu, Cyrus C.Y.
Population Dynamics: Theory of Nonstable Populations

Chung, Kevin K.H.
Socioeconomic Status and Academic Achievement

Church, Richard
Spatial Optimization Models

Cidam, Atilla
Emotions and Intergroup Relations

Cieciuch, Jan
Values, Social Psychology of

Cieri, Christopher
Databases and Statistical Systems: Linguistics

Cieza, Alarcos
Functioning, Disability and Health, International Classification of

Cigno, Alessandro
Family Theory: Economics of Intergenerational Relations

Cingolani, Patrick
Solidarity: History of the Concept

Citro, Constance F.
Surveys and Polling: Ethical Aspects

Claeskens, Gerda
Model Selection and Model Averaging

Clair, Matthew
Racism, Sociology of

Clamp, Kerry
Applied Criminology and Criminal Justice

Clark, Catie L.
Recidivism

Clark, David S.
Legal Systems, Classification of

Clark, Gordon L.
Finance, Geography of

Clark, Margaret S.
Social Relationships in Adulthood

Clark, M.H.
Propensity Scoring

Clark, William A.V.
Residential Segregation: Recent Trends

Clarke, David B.
Postmodern Urbanism

Clarke, Jennifer
Ethnicity, Race, Oppression, and Social Work: The Canadian Case

Clarke, Peter B.
Islam: Sub-Saharan Africa

Clarke, Sharon
Injuries and Accidents: Psychosocial Aspects

Clarke, Steve
Idealization, Abstraction, and Ideal Types

Claussen, Bjørgulf
Socioeconomic Status and Health

Claussen, Detlev
Hilberg, Raul (1926–2007)

Claval, Paul
Geographic Thought, History of
Spatial Thinking in the Social Sciences

Clayman, Steven E.
Ethnomethodology, General
News Interview

Clayton, Karen E.
Cross-Cultural Study of Education

Cleary, Anne
Hochschild, Arlie (1940–)

Clegg, Daniel
Labor Market Policy

Clegg, Stewart
Conflict: Organizational
Power in Society

Cleland, John
Fertility Transition: South Asia

Clifton, Charles
Sentence Comprehension, Psychology of

Climie, Emma A.
Intelligence: Assessments of

Clyde, Merlise A.
Experimental Design: Bayesian Designs

Coakley, John
Ethnic Conflicts

Coburn, David
Medical Profession

Cocker, Fiona
Public Health as Social Science

Coclanis, Peter A.
Quantification in History

Coglianese, Cary
Administrative Law: The United States and Beyond

Cohen, Jonathan D.
Cognitive Control (Executive Function): Role of Prefrontal Cortex

Cohen, Kenneth M.
Gay, Lesbian, and Bisexual Youth

Cohen, Robin
Diaspora

Cohen, Wesley M.
Innovation and Technological Change, Economics of

Coile, Courtney C.
Retirement, Economics of

Cole, Michael
Luria, Aleksandr Romanovich (1902–77)

Cole, Simon A.
Genetics and Forensics

Coleman, Carl H.
Common Rule: US Regulatory Protections for Human Research Participants

Coleman, David
Migrants and Migration in Europe

Coleman, Jonathan
Clustering and Cohesion in Networks: Concepts and Measures

Coleman, Simon
Pilgrimage
Religion, Anthropology of

Coleman, William D.
Policy Networks

Collins, E.
Remote Sensing

Collins, H.M.
Scientific Knowledge, Sociology of

Colonius, Hans
Decision and Choice: Random Utility Models of Choice and Response Time

Coman, Alin
Collective Memory, Psychology of

Commandeur, Jacques J.F.
Time Series: State Space Methods

Compion, Sara
Citizenship, Sociological Aspects of

Conand, Nicole M.
Busyness

Conibere, Richard
Social Exclusion, Social Deprivation and Health

Conkey, Margaret W.
Prehistoric Art

Conley, Dalton
Genetics and Sociology

Connelly, Louise
Religion and New Media

Conner, Mark
Health Behaviors

Connolly, John P.
Ideological Constraint: History and Current Status of the Concept

Conrad, Christoph
Social History

Conrad, Peter
Medicalization: Sociological and Anthropological Perspectives

Constant, Amelie F.
Ethnic Identity and Work

Constantine, Norman A.
Sexuality Education

Conversi, Daniele
War and Nationalism

Conzelmann, Kristin
Social Intelligence and Competencies

Cook, Karen S.
Exchange: Social

Cook, R. Dennis
Linear Hypothesis: Regression (Graphics)

Cook, Richard J.
Event History Analysis

Cook, Thomas D.
Generalization: Conceptions in the Social Sciences

Cook, Walter G.
Vocational Interests, Values, and Preferences, Psychology of

Cooke, Martin
Auditory Scene Analysis: Computational Models

Cooke, Miriam
Near Middle East/North Africa Studies: Culture

Cook-Martín, David
Immigration, Emigration, and Citizenship

Cookson, Peter W.
School as a Social System

Cooper, Frederick
African Studies: History

Cooper, Jill
Dubos, René Jules (1901–82)

Cooper, Joel
Cognitive Dissonance

Cooper, Jonathon A.
Police, Sociology of

Cope, Kevin L.
Constitutions

Copeland, Andrea J.
Libraries

Coquery-Vidrovitch, Catherine
Nationalism, Historical Aspects of: Africa

Corballis, Michael C.
Mirror Neurons, Theory of

Corbett, Greville G.
Gender: Grammatical

Corbett, Guido G.
Number, Grammatical

Corbetta, Guido
Family Business

Corcoran, Katie A.
Geospatial Applications in Archaeology

Corcoran, Mary P.
Jacobs, Jane (1916–2006)

Corcoran, Michael E.
Kindling

Cordes, Christian
Evolutionary Economics

Corina, David P.
Sign Language, Disorders of
Sign Language: Psychological and Neural Aspects

Cornwell, Benjamin
Sexual Behavior and Social Networks

Corr, Philip J.
Personality, Biological Models of

Corrêa, Sonia
Reproductive and Sexual Rights: Transnational Trends from a Global South Perspective

Corrigall-Brown, Catherine
Collective Identity

Corvellec, Hervé
Narrative Approaches to Organizations

Cossu, Andrea
Saussure, Ferdinand de (1857–1913)

Costa-Lopes, Rui
Intergroup Relations

Costantino, Tracie E.
Aesthetic Education

Cotterrell, Roger
Law as Constitutive

Coulangeon, Philippe
Leisure and Cultural Consumption: The European Perspective

Coulombe, Simon
Personal Projects

Cousin, Glynis
Antisemitism

Couso, Javier
Latin-American Legal Systems

Cover, Rob
Identity Offline and Online

Covino, Nicholas A.
Hypnosis, Psychiatry of

Cowan, Nelson
Short Term Memories, Theories of

Cowburn, Malcolm
Sex Offenders

Cowell, Frank A.
Inequality Measurement

Cox, David R.
Graphical Markov Models: Overview

Coyne, Michelle A.
Biosocial Approaches: Crime

Crafa, Daina
Meltzoff's "Like Me" Theory
Neural Correlates of Social Development
Self-Regulation During Early Childhood Across Cultures, Development of

Craglia, Max
Spatial Data Infrastructures

Cragun, Ryan T.
Science and Religion

Craig, C. Samuel
International Advertising

Craig, Michael
Retroactive Interference

Craig, Peter G.
Sport, Sociology of

Crane, Diana
Avant-Garde Art and Artists

Crano, William D.
Attitude Formation and Change
Persuasion Theories

Craven, Rhonda G.
Motivation in Australian Aboriginal Populations

Cravens, Hamilton
Cold War: Impact on Social and Behavioral Sciences

Crawford, Gary W.
Food Production, Origins of

Cree, Viviene E.
Responsibility in Social Work

Creelman, Carleton Douglas
Signal Detection Theory, History of

Creemers, Bert P.M.
Educational Effectiveness, The Field of
School Improvement

Crego, Cristina
Five Factor Model of Personality, Personality Disorder

Creighton, Millie
Nostalgia, Anthropology of

Cresto, Eleonora
Prisoner's Dilemma, One Shot and Iterated

Crimmins, James E.
Utilitarian Social Thought, History of

Critchlow, Donald T.
Policy Process: Nonprofit Institutions in the United States

Crits-Christoph, Paul
Therapist–Patient Relationship

Crocker, Linda
Content Validity

Croft, William
Evolution and Language: Overview
Functional Approaches to Grammar

Crofts, Penny
Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex

Croissant, Aurel
Data Bases and Statistical Systems: Security and Conflict

Cronin, Michael
Social Work and Disasters

Cronjé, Freek
Social Science Infrastructure: Africa

Cropanzano, Russell
Organizational Justice

Crosbie, Thomas
Coser, Lewis A. (1913–2003)

Crosnoe, Robert
Life Course: Sociological Aspects

Crossley, Stephen
Antisocial Behavior

Crothers, Charles H.G.
Dahl, Robert A (1915–2014)
Elster, Jon (1940–)

Crothers, Charles
Discipline-Building in the Social Sciences: Collective Memory, Biography and Autobiography
Social Structure: History of the Concept
Sociological Traditions

Croucher, Sarah
Historical Archaeology

Crouthamel, Jason
Gay Men's Emancipation Movement in Germany: 1897–1933

Crow, James F.
Kimura, Motoo (1924–94)
Wright, Sewall (1889–1988)

Crush, Jonathan
Global Migration

Cruze, Jennifer
Intersectionality and Crime

Crystal, David
Linguistics: Overview

Csordas, Thomas J.
Body: Anthropological Aspects

Culham, Jody
Cortical Areas Engaged in Movement: Neuroimaging Methods

Cullen, Kathleen E.
Vestibular System
Vestibulo-Ocular Reflex, Adaptation of the

Culler, Jonathan
Deconstruction

Cumley, Samantha R.
Crime and Gender

Cummings, Scott L.
Poverty Law: United States
Public Interest Law: The United States and Beyond

Cummings, Thomas G.
Closed and Open Systems: Organizational

Cummins, Lori F.
Stress: Psychological Perspectives

Cunningham, Frank
Democratic Theory

Cusson, Maurice
Control, Social

Cusumano, Michael A.
Business Platforms

Cutler, Tony
New Managerialism and New Public Sector Management

Cutter, W.
Mentally Ill, Public Attitudes towards

Cuttica, Cesare
Absolutism, History of
Intellectual History
Isms and History

Cuzzolin, Pierluigi
Indo-European Languages

Czech, Sebastian
Sex Crimes

D

D'Agostino, Fred
Social Science, The Idea of

D'Alisa, Giacomo
Degrowth

D'Ambrosio, Conchita
Health Insurance: Economic and Risk Aspects

Daenekindt, Stijn
Cultural Participation, Trends in

Dagum, Estela B.
Time Series: Seasonal Adjustment

Dahl, Östen
Tense, Aspect, Mood and Evidentiality, Linguistics of

Dahlgren, Peter
Mass Media: Introduction and Schools of Thought

Dalbert, Claudia
Belief in a Just World

Dalto, Georgia
Classical Conditioning Methods in Psychotherapy

Dalton, Russell J.
Party Identification and Nonpartisanship

Daly, Martin
Sociobiology: Overview

Daly, Mary
Family Policy

Damarin, Amanda K.
Bureaucracy, Sociology of

Damasio, Antonio R.
Agnosia

Damerow, Peter
Number Systems, Evolution of
Scientific Revolution, History and Sociology of

Damodaran, Vinita
Environmental History

Damon, Frederick H.
Kula Ring, Anthropology of

Dandekar, Hemalata C.
Rural Planning: General

Danesi, Marcel
Linguistic Anthropology

Daniel, Ellice
Social Work with Racialized Groups: Frameworks for Practice

Dannefer, Dale
Age Stratification

Danso, Ransford
Immigration, Citizenship, and Integration: Social Work Connections

Danziger, Kurt
Introspection: History of the Concept

Darian-Smith, Eve
Postcolonial Law

Darke, Sacha
Prison Life, Sociology of: Recent Perspectives from the United Kingdom

Darker, Catherine
Drug Use and Abuse: Psychosocial Aspects

Darrah, Charles N.
Busyness

Das, Veena
Caste

Dascher, O.
Data Bases and Statistical Systems: Archives and Historical Databases

Daston, Lorraine
Science, History of

Datta, Monti N.
Slavery as Social Institution

Dau-Schmidt, Kenneth G.
Law and Economics: Empirical Dimensions

David, H.A.
Order Statistics

David Milner, A.
Visual Streams: Dorsal and Ventral

Davidov, Eldad
Values, Social Psychology of

Davie, Grace
European Religion from the Atlantic to the Black Sea
Women's Religiosity

Davies, Martin
Knowledge (Explicit, Implicit and Tacit): Philosophical Aspects

Davis, Andrea
Life Course Criminology

Davis, Angela
Oral History

Davis, Diane E.
Development and Urbanization

Davis, Ernest
Knowledge Representation

Davis, Gerald F.
Organization: Overview
Stockholders' Ownership and Control

Davis, Howard
Touraine, Alain (1925–)

Davydov, Dmitry M.
Hypertension: Psychosocial Aspects

Daw, Jonathan
Gene–Environment Interplay

Dawes, Christopher
Genetics and Sociology

Dawes, Robyn M.
Probabilistic Thinking

Dawood, Khytam
Sexual Orientation: Genetic Aspects

Day, Sean A.
Linguistic Synesthesia

Dayé, Christian
Methodology of the History of the Social and Behavioral Sciences
Schools in the Social and Behavioral Sciences: Concepts and Historical Relevance
Small Group Research, History of

De Baets, Antoon
Human Rights, History of

de Campo, Marianne E.
Moral Entrepreneur

de Castro, Leonardo D.
Research Ethics, Cross-Cultural Dimensions of

De Corte, Erik
Instructional Psychology
Mathematics Education

De Coster, Stacy
Crime and Gender

de Felipe-Redondo, Jesus
Class Consciousness

de Gelder, Beatrice
Prosopagnosia

de Jong, Eelke
Culture and Economic Development

de Luque, Mary Sully
Cross-Cultural Training in Industrial and Organizational Psychology

de Magalhães, Claudio
Urban Regeneration

de Palma, André
Transportation, Supply, and Congestion

de Raad, Boele
Big Five Factor Model, Theory and Structure

de Rezende Martins, Estevao C.
Parliaments, History of
Political Parties, History of

de Ribaupierre, Anik
Piaget's Theory of Cognitive Development

de Ridder, Denise
Health Psychology

De, Rohit
South Asian Legal Traditions

De Soete, Britt
Situational Judgment Test

de Vito, Stefania
Future Thinking

de Vries, Nanne K.
Health Education and Health Promotion

Deakin, Nicholas
Philosophy, Psychiatry, and Psychology

Dean, Angela M.
Experimental Design: Overview

Dear, Peter
Experiment in Science and Technology Studies

Deasy, Lauren L.
Autobiographical Memory

Deaux, Kay
Social Identity in Sociology

Decety, Jean
Empathy: A Social Neuroscience Perspective

Deci, Edward L.
Self-Determination Theory

Decker, Frank
Property Ownership
Property Rights, History of
Security: Commercial Transactions

Dees, J. Gregory
Social Innovation

Deflem, Mathieu
Anomie: History of the Concept
Policing

Dehue, G.C.G. Trudy
Experimentation in Psychology, History of

Deinzer, Renate
Oral Health, Psychology of

Del Giudice, Marco
Gender Differences in Personality and Social Behavior

Del Tredici-Braak, Kelly
Alzheimer's Disease, Neural Basis of

Delgado, Roberto A.
Primate Behavior: Significance for Understanding Humans

Delius, Julia A.M.
Baltes, Paul B (1939–2006)

DellaPergola, Sergio
Jewish Diaspora

Demaria, Federico
Degrowth

Demetriou, Andreas
Intelligence in Cultural, Social and Educational Context

Demetriou, Chares
Social Movements and Political Violence

Demeulenaere, Pierre
Conformity: Sociological Aspects
Interests in Sociological Analysis
Methodological Individualism: Philosophical Aspects

DeMichele, Matthew
Prisons and Imprisonment

Demuth, Carolin
Mother-Child Communication: Cultural Differences
Qualitative Methodology in Developmental Psychology

Denis, Jeffrey S.
Racism, Sociology of

Dennis, Alex
Interactionism, Symbolic

Dennis, Michael A.
Universities and Science and Technology: United States

Denzin, Norman K.
Interpretive Methods: Micromethods

Depaepe, Fien
Mathematics Education

Dequech, David
Old and New Institutionalism in Economics

Derntl, Birgit
Gender and Health Care

Desai, Vandana
Nonprofit Sector, Geography of

Deschenes, Olivier
Green Jobs

DeSoto, Kurt A.
Reconstructive Memory, Psychology of

DeSouza, Lisette M.
Childhood and Adolescence: Developmental Assets

Desrosières, Alain
Statistics, History of

Detel, Wolfgang
Social Constructivism

Deuchar, Ross
Youth Gangs

Deviatko, Inna F.
Causal Analysis, History of

Devisch, René
Central Africa: Sociocultural Aspects

Dewar, Michaela
Retroactive Interference

Dhillon, Pradeep A.
International Organisations and Education

di Pellegrino, Giuseppe
Attention, Neural Basis of

Di Tommaso, Marco R.
Special Economic Zones and Cluster Dynamics: China

Di Trocchio, Federico
Research Publication: Ethical Aspects

Diamond, Shari S.
Juries

Diawara, M.
Historical Thought and Historiography: Sub-Saharan Africa

Díaz, Alex
Quasi- and Field Experiment

Díaz-Andreu, Margarita
Ethnic Identity and Ethnicity in Archaeology

DiBenedetto, Maria K.
Self-Efficacy: Education Aspects

Dicke, Anna–Lena
Schooling: Impact on Cognitive and Motivational Development

DiClemente, Ralph J.
HIV Risk Interventions

Diederich, Adele
Decision and Choice: Sequential Decision Making

Diefenbacher, Albert
Liaison and Consultation Psychiatry

Diefendorff, James M.
Work Motivation

Diego, Vincent P.
Familial Studies: Genetic Inferences

Diekmann, Andreas
Cooperation: Sociological Aspects

Diemer, Julia
Virtual Reality in Psychotherapy

Diest, Ilse Van
Symptom Perception, Awareness and Interpretation

Dietrich, Julia
Passion and Motivation

Díez, Jordi
Lesbian and Gay Activism in Latin America

Dijkstra, Jacob
Rationality in Society

Dijkstra, Ton
Second Language Acquisition

Dikötter, Frank
Eugenics in Asia

Dilaver, Ozge
Self-Organizing Dynamical Systems

Dilchert, Stephan
Pro-Environmental Behavior

Dilley, Roy
Apprenticeship, Anthropological Aspects

Dimand, Robert W.
Macroeconomics, History of up to 1933

DiMatteo, M. Robin
Patient Adherence to Health Care Regimens

Dimotakis, Nikolaos
Genetic Factors in Organizational Psychology

Dinos, Sokratis
Public Psychiatry and Public Health Mental Systems

Ditlmann, Ruth
Field Experiments

Dixon, John
Desegregation

Dixon, Pauline
Affordable Private Schools in Developing Countries

Djundeva, Maja
Linked Lives

Djupe, Paul A.
Religion and Politics in the United States

Doak, Steven
Identity Offline and Online

Dobash, Rebecca E.
Domestic Violence: Sociological Perspectives

Dobash, Russell P.
Domestic Violence: Sociological Perspectives

Dobbin, Frank
Organizations and Culture

Dobson, Keith S.
Concepts and Methods of Cognitive Therapies

Dochartaigh, Niall Ó.
Territorial Conflicts

Dodds, Klaus
Geopolitics

Dogan, Mattei
Deutsch, Karl W (1912–92)
Specialization and Recombination of Specialties in the Social Sciences

Dogra, Nisha
Child and Adolescent Psychiatry, Principles of

Döhring, Juliane
Transient Global Amnesia

Doignon, Jean-Paul
Cognition: Knowledge Spaces

Doing, Park
Laboratory Studies: Historical Perspectives

Doise, Willem
Levels of Analysis in Social Psychology

Dominelli, Lena
Green Social Work

Domotor, Zoltan
Mathematical Models in Philosophy of Science

Donaldson, Lex
Structural Contingency Theory

Donat, Matthias
Belief in a Just World

Donham, Donald L.
Modes of Production

Donley, Amy M.
Food Deserts in US Cities
Violence, Criminalization, and the Homeless

Donnan, Hastings
Borders, Anthropology of

Donnelly, Ellen A.
Privacy of Individuals in Social Research: Confidentiality

Donnermeyer, Joseph F.
Crime in the Rural Context

Donovan, Pamela
Rumors and Urban Legends

Dooley, Brendan
Media and History

Doom, Jenalee R.
Stress in Infancy and Early Childhood: Effects on Development

Dorjee, Tenzin
Honor Killing: Multidimensional and Multilevel Perspectives

Dornadic, Alicia
Busyness

Doron, Israel (Issi)
Law and Aging

Dossani, Rafiq
Business Process: Outsourcing

Dossani, Samina
Business Process: Outsourcing

Dougherty, Peter J.
Publishing: Academic and University Presses

Douzenis, Athanassios
Explanatory Style and Health

Dovidio, John F.
Prosocial Behavior and Empathy

Dovie, Delali Benjamin K.
Africa's Environment: A Stressed Biogeographical and Cultural Landscape

Downey, Gary L.
Engineering Studies

Downs, Donald A.
Civil Liberties and Human Rights

Drago, Robert
Wage Differential and Structure

Drake, Deborah H.
Prison Life, Sociology of: Recent Perspectives from the United Kingdom

Dransch, Doris
Dynamic Mapping in Geography

Driessen, Henk
Humor, Anthropology of

Drolet, Julie
Disasters in Social, Cultural and Political Context

Dronkers, Nina F.
Speech Production, Neural Basis of

Droogers, André
Syncretism

Drouard, Alain
Darwinism: Social
Food, Sociology of

Drozdek, Adam
Media Ethics

Drummond, Jesse
Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations

Drummond, Lynne M.
Behavior Therapy: Psychiatric Aspects

Drury, John
Collective Behavior, Social Psychology of
Social Movements: A Social Psychological Perspective

Dryden-Peterson, Sarah
Conflict: Education and Youth

Du Mon, Denies
Negation, Linguistics of

Dubin, Matthew
Flow in Motivational Psychology

Dubois, Michel
Collective Beliefs, Sociological Explanation of
Ideology, Sociology of

Dubois, Vincent
Cultural Policy Regimes in Western Europe

Ducheny, Kelly
Training in Clinical Psychology in the United States: Practitioner Model

Duckitt, John
Authoritarian Personality

Duckworth, Angela L.
Grit

Duff, Alistair S.
Information Society

Duller, Matthias
Area Studies, History of

Dumais, Susan A.
Cultural Capital and Education

Dumont, Frank
Personality: Historical and Conceptual Perspectives

Dumont, Hanna
Schooling: Impact on Cognitive and Motivational Development

Duncan, George T.
Confidentiality and Statistical Disclosure Limitations

Duncan, Greg J.
Panel Surveys: Uses and Applications

Dunckel, Heiner
Job Analysis and Work Roles, Psychology of

Duneier, Mitch
Sociology: Overview

Dunlap, Riley E.
Environmental Sociology

Dupâquier, Jacques
Malthus, Thomas Robert (1766–1834)

Dupont, C.
Negotiation and Bargaining: Organizational Aspects

Dür, Andreas
International Trade: Commercial Policy and Trade Negotiations

Durney, Florence
Irrigation Societies

Durrheim, Kevin
Desegregation

Durso, Laura
Lesbians: Social and Economic Situation

Duru-Bellat, Marie
Social Inequality and Schooling

Dworkin, Steven N.
Etymology

Dyck, Isabel
Qualitative Methods in Geography

Dyck, Noel
Sport, Anthropology of

Dye, David H.
Conflict and War, Archaeology of: Behavior and Social Organization

Dyer, Allen R.
Ethics and Psychiatry

Dykema, Jennifer
Surveys: Question Wording and Response Categories

Dyson, Alan
Extended Schools

Dziegielewski, Sophia F.
Ethical Practices: Institutional Oversight and Enforcement

E

Earle, Rod
Prison Life, Sociology of: Recent Perspectives from the United Kingdom

Eastmond, Marita
Refugees in Anthropology

Easton, Judith A.
Human Mate Choice, Evolution of

Eaton, Morris L.
Invariance in Statistics

Ebbinghaus, Bernhard
Welfare Retrenchment

Ebers, Mark
Interorganizational Relationships and Networks

Eccles, Jacquelynne S.
Academic Achievement Motivation, Development of
Passion and Motivation

Economides, Kim
Legal Education

Edelman, Lauren
Law: The Socio-Legal Perspective

Edelman, Shimon
Marr, David (1945–80)

Eder, Anja
Semantic Differential

Eder, Franz X.
Popular Sexual Cultures in Western Europe (1930s to the Present)

Edwards, Anthony W.F.
Likelihood in Statistics

Edwards, A.W.F.
Fisher, Ronald A (1890–1962)

Edwards, Laura A.
Cognitive Development in Infancy, Neural Mechanisms of

Efron, Bradley
Resampling Methods of Estimation

Egger, Matthias
Evidence-Based Medicine

Egloff, Rainer
Definition of the situation: History of the Concept

Egorova, Yulia
New Genetics and Race

Egri, László
Constraint-Satisfaction Models

Ehlert, Ulrike
Gynecological Health: Psychosocial Aspects

Ehmer, Josef
Artisans and Guilds, History of

Ehmer, J.
Work, History of

Ehrhart, Mark G.
Organizational Climate in the Work Setting

Eickelman, Dale F.
Knowledge, Anthropology of
Near Middle East/North Africa Studies: Religion
Transnational Religious Identities (Islam, Catholicism, and Judaism): Cultural Concerns

Eid, Michael
Sun Exposure and Skin Cancer Prevention

Einstein, Gilles O.
Prospective Memory

Eisenberg, Theodore
Remedies and Damages: Legal Aspects

Eisenstadt, Samuel N.
Civilizations

Eisenstadt, Shmuel N.
Generations, Sociology of

Eisermann, Gottfried
Pareto, Vilfredo (1848–1923)

Eisinga, Rob
Religiosity and Prejudice against Minorities

Eizenberg, Efrat
Neighborhood

El Feki, Shereen
Sexual Politics in the Arab World

Elbert, Thomas
Torture and its Consequences, Psychology of

Elder, Glen H.
Life Course: Sociological Aspects
Motivation: Life Course and Sociological Perspectives

Elfenbein, Hillary Anger
Emotion, Perception and Expression of

Elgar, Frank J.
Behavioral Problems, Effects of Socio-Economic Status on

Eliaeson, Sven
Science and Politics: Value Neutrality

Eling, Paul
Neurolinguistics, History of

Ellickson, Paul B.
Market Structure and Performance

Elliot, Andrew J.
Avoidance and Approach Motivation: A Brief History

Elliott, Anthony
Psychoanalysis in Sociology

Elliott, Julian G.
Teacher Expertise

Ellis, Richard J.
Pluralism

Ellwood, John W.
Public Policy Schools

Elo, Irma T.
Mortality Crossover

Elsenbroich, Corinna
Self-Organizing Dynamical Systems

Eltaiba, Nada
Islamic Social Work: Hope and the Commitment to Change

Elwert, Georg
Boas, Franz (1858–1942)
Conflict: Anthropological Aspects

Elyachar, Julia
Nongovernmental Organizations: Anthropological and Historical Aspects

Embleton, Sheila M.
Historical Linguistics: Numerical Methods

Emmerling, Robert
Leadership Education: Theory and Practice

Enchautegui, Maria E.
Immigration, Applied Research on

Enders, Juergen
Higher Education Management

Endresen, Sylvi B.
Labor, Geography of

Engelberg, Philipp M.
Factor Analysis and Latent Variable Models in Personality Psychology

Engelstad, Ericka
Gender, Feminism, and Sexuality in Archaeological Studies

England, Paula
Hooking up and Casual Sex

English, Tammy
Emotions and Aging

Enns, Carolyn Zerbe
Feminist Psychotherapy

Eno, Jared
Quasi-Experimental Designs

Entman, Robert
Mass Media, Representations in

Epp, Charles R.
Supreme Courts

Erb, Karl-Heinz
Land-Use Indicators

Erdogan, Berrin
Leader–Member Exchange Theory

Eriksen, Thomas H.
Anthropology, History of

Eriksen, Thomas Hylland
Multiculturalism, Anthropology of

Erlmann, Veit
Music: Anthropological Aspects

Ermisch, John F.
Family Theory: Economics of Childbearing

Ernstes, Amy M.
Subculture, Sociology of

Erskine, Jane
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Ertl, Verena
Mass Trauma: Psychopathological Effects across the Life Span

Ertman, Martha M.
Sexual Orientation and the Law

Esfeld, Michael
Atomism and Holism: Philosophical Aspects

Eskreis-Winkler, Lauren
Grit

Esler, Anthony
Generations in History

Espenshade, Thomas J.
Population Dynamics: Momentum of Population Growth

Esping-Andersen, Gosta
Quantitative Cross-national Research Methods
Welfare Policy Comparisons

Estes, J.
Remote Sensing

Etkind, Alexander
Russian and Post-Soviet Studies: Culture

Evans, Geoffrey
Voting, Explanations of: Social Class

Evans, George W.
Expectations, Economics of

Evans, Ian M.
Reinforcement, Principle of

Evans, Nicholas G.
Censorship in Research and Scholarship

Evans-Lacko, Sara
Mentally Ill, Public Attitudes towards

Evdokimov, Konstantin
Epigenetics

Evens, Terence M.S.
Communes, Anthropology of

Everaerd, Walter
Sex Therapy, Clinical Psychology of

Everall, Ian Paul
AIDS: Acquired Immune-Deficiency Syndrome

Everett, Martin G.
Core–Periphery Patterns in Networks

Evers, Hans-Dieter
Southeast Asia: Sociocultural Aspects

Every, Danielle
Racism: Social Psychological Perspectives

Ewick, Patricia
Law and Everyday Life

Ezell, Michael E.
Crime, Desistance, and Persistence

Ezrahi, Yaron
Science and The State

F

Faasse, Kate
Stress, Coping and Health

Fabb, Nigel
Language and Literature

Faber, Jacob W.
Neighborhood Effects

Fabian, Johannes
Interpretation in Anthropology

Fábregas, Antonio
Morphology in Linguistics

Fabrigar, Leandre R.
Attitudes and Behavior

Facchini, Giovanni
International Migration

Fahy, Frances
Participatory Action Research in Environmental and Ecological Studies

Fainstein, Susan S.
Just City

Fajans, Jane
Practice: An Anthropological Approach

Fakhari, Pegah
Dynamic Decision Making

Fallaize, Elizabeth
Beauvoir, Simone de (1908–86)

Falmagne, Jean-Claude
Markov Processes in Learning Spaces
Mathematical Psychology

Falzon, Lucia
Social Networks, Algebra of

Falzon, Mark-Anthony
Multisited Field Studies

Fandrem, Hildegunn
Friendship During Adolescence and Cultural Variations

Fantie, Bryan D.
Brain Damage: Neuropsychological Rehabilitation

Farber, Daniel A.
Legal Formalism

Farber, Henry S.
Dispute Resolutions in Economics

Farebrother, Richard W.
Gauss, Carl Friedrich (1777–1855)

Fargher, Lane F.
Mesoamerica, Archaeology of

Farías, Ignacio
Culture and Actor Network Theory

Farina, Katie A.
Cyber Crime: Identity Theft

Faris, David M.
New Media and Democracy in the Arab World

Farquhar, John W.
Health Interventions, Community-based

Farrell, Amy
Guilt

Farrell, Jordan S.
Epilepsy

Farrington, David P.
Personality and Crime

Fasel, Nicole
Immigration: Social Psychological Aspects

Fashler, Samantha
Phantom Limbs

Faubion, James D.
Anthropology and History

Fawcett, Christine
Social Referencing during Infancy and Early Childhood across Cultures

Feagin, Joe R.
Racial Relations

Featherstone, Brid
Risk and Uncertainty in the Social Sciences: Implications for Social Work Theory and Practice

Feddersen, Mayra
Law, Mobilization of

Feeley, Malcolm M.
Courts and Adjudication
Trials

Feeley, Malcolm
Prosecution

Feger, Hubert
Classification, Understandings of in the Social Sciences

Feinman, Gary M.
Settlement and Landscape Archaeology

Fekete, Christine
Caregiver Burden

Feld, Lars P.
State and Local Taxation

Feldman, Heidi M.
Attention-Deficit and Hyperactivity Disorders, Neural Basis of
Neural Basis of Language Development

Feldman, Marcus W.
Cultural Evolution: Theory and Models

Feldman, Stanley
Attitudes, Political: Authoritarianism and Tolerance

Feng, Wang
Fertility Transition: China

Ferguson, Christopher J.
Computer Games and Academic Achievement

Ferguson, Gerry
Criminal Liability and Criminal Defenses

Ferguson, Kyle E.
Behavior Therapy: Background, Basic Principles, and Early History
Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues
Behavior Therapy: The Second and Third Waves

Ferguson, Lucy
Tourism, the Millennium Development Goals, and Gender Equality

Fernandez, James W.
Cultural Relativism, Anthropology of

Ferngren, Gary B.
Religion and Sciences

Ferrara, Alessandro
Person and Self: Philosophical Aspects

Ferrer, Emilio
Longitudinal Analyses of Sexual Development through Early Adulthood

Ferring, Dieter
Alzheimer's Disease: Behavioral and Social Aspects

Ferris, Gerald R.
Power, Politics, and Influence in Organizations

Festini, Sara B.
Dysexecutive Amnesia

Fetterman, David M.
Empowerment Evaluation
Ethnography in Applied Social Research

Fetvadjiev, Velichko H.
Five Factor Model of Personality, Universality of

Fielden, Caroline
Extraversion

Fienberg, Stephen E.
Sample Survey Methodology, History of

Figgou, Lia
Social Psychology: Research Methods

Figner, Bernd
Personality and Risk Taking

Fine, Gary Alan
Participant Observation

Fine, Robert
Antisemitism

Fineman, Martha A.
Family Law in the United States

Finn, Mary A.
Sex Crimes

Finnegan, Diarmid A.
Science, Geography of

Finnegan, Ruth
Oral and Literate Expression

Finney, Ben
Navigation

Finnström, Sverker
War: Anthropological Aspects, Historical Development of

Firebaugh, Glenn
Ecological Fallacy, Statistics of

Fischer, Manfred M.
Spatial Analysis in Geography

Fischer, Michael M.J.
Science and Technology, Anthropology of

Fischer, W.
Poverty in History

Fischer-Kowalski, Marina
Social Ecology

Fischhoff, Baruch
Environmental Cognition, Perception, and Attitudes
Tversky, Amos (1937–96)

Fishburn, Peter
Measurement Theory: Conjoint

Fisher, Celia B.
Research Ethics: Research

Fisher, Eleanor
Ethical Turn in International Development

Fisher, Ronald J.
Conflict and Conflict Resolution, Social Psychology of

Fisser, Petra
Computer-Assisted Instruction

Fitzmaurice, Connor J.
Conspicuous consumption and distinction, history of

Fitzmaurice, Garrett M.
Binary Response Models and Logistic Regression
Linear Mixed Models

Fixsen, Dean
Implementation Science

Fjeldstad, Øystein D.
Network Paradigm: Applications in Organizational Science

Flache, Andreas
Rationality in Society

Flammer, August
Self-Efficacy

Flap, Henk
Organizations: Unintended Consequences

Fleck, Christian
Attitude: History of Concept
Lazarsfeld, Paul Felix (1901–76)
Merton, Robert K (1910–2003)
Methodology of the History of the Social and Behavioral Sciences
Riesman, David (1909–2002)

Fleminger, Simon
Confabulation

Fletcher, Jason
Genetics and Sociology

Fleury-Steiner, Benjamin
Death Penalty

Flindall, Jason W.
Parietal Lobe

Flor, Herta
Fear Conditioning: Overview
Pain, Health Psychology of

Flora, Peter
Rokkan, Stein (1921–79)

Flores-Yeffal, Nadia Y.
Migration and Social Capital

Flynn, Terry N.
Best Worst Scaling: Theory and Practice

Fogarty, Laurel
Cultural Evolution: Theory and Models

Fogarty, Richard S.
Eugenics in Europe

Foguet, Joan M.B.
Leadership Education: Theory and Practice

Folgheraiter, Fabio
Relational Social Work

Follette, William C.
Classical Conditioning Methods in Psychotherapy

Fomichov, Vladimir A.
Conscious Control during Childhood, Development of

Fontana, Andrea
Symbolic Interaction: Methodology

Fook, Jan
Critical Social Work Practice

Forbes, Linda C.
Workplace Autonomy

Ford, Richard
Race and the Law

Foreman, Christopher H.
Regulatory Agencies

Formiga, Cibelle K.M.R.
Motor Skills: Development in Infancy and Early Childhood

Forrest, David
Thematic Maps in Geography

Forrest, Ray
Housing and Cities

Forstall, Richard L.
Urban Places: Statistical Definitions

Forsyth, Craig J.
Misdemeanor Offenses

Forsyth, Donelson R.
Group Psychotherapy, Clinical Psychology of

Forsyth, John P.
Single-Case Designs: Methodology

Forte, Maximilian C.
Human Terrain System (United States): Critique

Fortis, Paolo
Kinship in Anthropology

Fortunato, Laura
Marriage Systems, Evolution of

Foss, Nicolai J.
Agency Theory

Foster, Gigi
Experimental Economics

Foster, Holly
Mass Imprisonment and Its Consequences

Foster, Robert J.
Oceania, Sociocultural Overviews: Melanesia

Fotopoulou, Aikaterini (Katerina)
Confabulation

Fournier, Marcel
Durkheimians, The

Fox, Anthony
Internal Reconstruction in Linguistics

Fox, Elizabeth A.
Primate Behavior: Significance for Understanding Humans

Fox, Madeline
Participatory Action Research in Social Research

Francis, Charles A.
Agroecology and Agricultural Change

Frankenberg, Elizabeth
Demographic Impact of Disasters
Experimental Methods in Survey Research in Demography

Franks, David D.
Neurosociology

Fraser, Barry J.
Classroom Climate
Environments for Education

Fraser, Ben
Evolutionary Approaches to Human Behavior: Philosophical Aspects

Fraser, Donald A.S.
Fiducial and Structural Statistical Inference

Frattini, Tommaso
International Migration

Frazier, John W.
Applied Geography: A US Perspective

Frechtling, Joy A.
Logic Models

Fredricks, Jennifer A.
Academic Engagement
Motivation, Familial Influences on

Fredrickson, Barbara L.
Broaden-and-Build Theory of Positive Emotions

Fredrickson, G.M.
Racism, History of

Free, Marvin D.
Wrongful Convictions

Freedman, David A.
Ecological Inference

Freeland, Edward P.
Public Opinion: Social Attitudes

Freeman, Claire
Children's Geographies

Freeman, John H.
Eyeblink Classical Conditioning

Freeman, Scott
Environmental Conservation and Development: Critical Perspectives

Freestone, Robert
Urban Planning (Western), History of

Freilich, Joshua D.
National Security and Domestic Terrorism: The US Case

Freitag, Claudia
Face Perception Development During Early Childhood

French, David P.
Self-Efficacy and Health

Freund, Alexandra M.
Self-Regulation in Adulthood

Freund, A.M.
Life Management, Developmental Psychology of

Freund, Hans-Joachim
Jackson, John Hughlings (1835–1911)

Frey, Bruno S.
Informal and Underground Economics

Frey, Dieter
Group Processes in Organizations
Money and Finances, Psychology of

Fridenson, Patrick
Industrial Relations, History of

Fried, Robert C.
Punishment: Comparative Politics of

Friedberg, Erhard
Organizations, Sociology of

Friederici, Angela D.
Biological Preconditions for Language Development

Friedkin, Noah E.
Models of Influence in Networks

Friedländer, Saul
Holocaust, The

Friedman, Benjamin M.
Monetary Policy since the 2007–2009 Financial Crisis

Friedman, Lawrence M.
Law: Change and Evolution

Friedman, Milton
Stigler, George Joseph (1911–91)

Friend, Marilyn
School to School Collaboration

Friese, Carrie
Genetics, Disease, and Reproduction

Friese, Heidrun
Eudaemonism

Fritz, Florian
Societal Impact Assessment

Fritze, J.
Psychopharmacotherapy, Costs of

Frognier, André-Paul
Nation, Sociological Aspects of

Frost, Nick
Children and Families in Social Work

Fry, Douglas P.
Conflict and War: Anthropological Aspects

Fryer, David
Unemployment and Mental Health

Fu, Xiaoxue
Temperament Development, Theories of

Fuchs, Reinhard
Physical Activity and Health

Fuhse, Jan
Networks and Meaning

Fuist, Todd Nicholas
Civil Religion

Fukui, Haruhiro
East Asian Studies: Politics

Fuller, Alison
Vocational Education

Fuller, Steve
Positivism, History of

Fulton, Brad
Religion and Stratification

Funder, David C.
Person Perception, Accuracy of

Funtowicz, Silvio O.
Peer Review and Quality Control
Science, New Forms of

Furner, Mary O.
Policy Knowledge: New Liberalism

Furnham, Adrian
Healing Practices: Alternative and Complementary
Personality Assessment: Overview

Fürnkranz-Prskawetz, Alexia
Population Dynamics: Mathematic Models of Population, Development, and Natural Resources

Furtado, Delia
Ethnic Intermarriage

Fynsk, C.
Linguistic Turn

G

Gabbard, Carl
Perception of Action Space: Using Multiple Frames of Reference

Gadenne, Volker
Critical Rationalism

Gadke, Daniel L.
Agreeableness

Gaebel, Wolfgang
Schizophrenia

Gagliardi, Pasquale
Myths and Symbols: Organizational

Gagneux, Pascal
Primate Groups and Their Correlates

Gailus, M.
Crowds in History

Gaissmaier, Wolfgang
Decision Making: Nonrational Theories

Gajdamaschko, Natalia
Vygotsky's Sociocultural Theory

Galanter, Marc
Law: The Socio-Legal Perspective

Galaty, John G.
Pastoralism in Anthropology

Galea, Liisa A.M.
Endocrinology and Psychiatry

Galesic, Mirta
Decision and Choice: Heuristics

Gallace, Alberto
Tactile Processing

Galnoor, Itzhak
Civil Service

Galtung, Johan
Peace

Galupo, M. Paz
Transgender, Transexual, and Gender Variant Individuals

Galupo, M.P.
Friendship Experiences of Lesbian, Gay, Bisexual, and Transgender Individuals

Gamp, Martina
Health Risk Perception

Gamson, Joshua
Celebrity

Gamst, Frederick C.
Work, Sociology of

Ganczer, Mónika
Socialist Law

Ganzeboom, Harry B.G.
Art and Socialisation

Gao, Fang
Ethnicity and Educational Achievement

García, Ofelia
Language Policy

García-Manglano, Javier
Gender, Time-Use, and Fertility Recovery in Industrialized Countries

Gardhouse, Katherine
Mindfulness

Gardikiotis, Antonis
Attitude Formation and Change
Persuasion Theories

Gardiner, Emily
Social Competence During Adolescence Across Cultures

Gardner, Andy
Kin Selection

Gardner, Bruce L.
Agriculture, Economics of

Gardner, Don
Friendship, Anthropology of

Gardner, William L.
Authentic Leadership

Garforth, Lisa
Latour, Bruno (1947–)

Garlitz, Dustin B.
Barthes, Roland (1915–80)
Frankfurt School: Institute for Social Research

Garnsey, Elizabeth
Organizational Emergence and Firm Formation

Garrett, Merrill F.
Speech Errors, Psychology of

Garrison, Nanibaa' A.
Genetic Ancestry Testing

Garth, Bryant G.
International Arbitration
Law and Society: Development of the Field

Gartner, Rosemary
Crime: Knowledge about and Prevalence

Garvey, Bob
Business Education

Gasper, Joseph M.
Data Bases and Statistical Systems: Applied Social Research

Gass, Robert H.
Social Influence, Sociology of

Gastwirth, Joseph L.
Screening and Selection

Gaudet, Beverly
Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review

Gaukroger, Stephen
Objectivity, History of

Gaus, Gerald
Public Reason

Gavilán, Maria Beatriz
Social Science Infrastructure: Central and South America

Gavrov, Sergei
Modernization, Sociological Theories of

Gawer, Annabelle
Business Platforms

Gawronski, Bertram
Implicit Social Cognition

Gayes, Laurie A.
Friendship During Childhood and Cultural Variations

Gayet, Cecilia
Fertility Transition: Latin America and the Caribbean

Geangu, Elena
Empathy during early Childhood Across Cultures, Development of

Geary, David C.
Evolutionary Approach to Education Science

Gebauer, Heiko
Absorptive Capacity (of Organizations)

Geeraerts, Dirk
Lexical Semantics

Geertz, Armin W.
Religious Belief, Evolution of

Geldhof, G. John
Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations

Gellner, David N.
Buddhism

Gelman, Andrew
Large-Scale Social Surveys, Analysis of

Gelso, Charles J.
Counseling Psychology

Gelsthorpe, Loraine
Women and Criminal Justice

Geman, Stuart
Probabilistic Grammars and their Applications

Gentile, Emilio
Dictatorship in History

George, Laura J.
Social Work and Psychodynamic Thought

George, Linda K.
Age, Sociology of
Human Development and Health

George, Purnima
Social Enquiry and Action Research for Social Work

George, Robert
Natural Law

George, Usha
Migration and Social Work

Gerbault, Pascale
Human Evolutionary Genetics

Gerhardt, Volker
Kant, Immanuel (1724–1804)
Nietzsche, Friedrich (1844–1900)

Gerin, William
Cardiovascular Reactivity and Recovery

Gerkey, Drew
Kinship, Evolution of

Gerland, Patrick
Demographic Techniques: Data Adjustment and Correction

Gershuny, Jonathan
Time Use Research Methods
Time Use Research: Subjective Time Use

Gervain, Judit
Near-Infrared Spectroscopy

Geschiere, Peter
Witchcraft

Gestsdóttir, Steinunn
Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations

Getis, Arthur
Spatial Association, Measures of

Geurts, Teun
Grandparent–Grandchild Relationships

Geyer, Siegfried
Compression of Morbidity

Gheihman, Nina
Nation-State as Symbolic Construct

Gherardi, Silvia
Learning: Organizational

Ghiselin, Michael T.
Darwin, Charles Robert (1809–82)

Ghislandi, Simone
Health Insurance: Economic and Risk Aspects

Giacomantonio, Mauro
Embodied Social Cognition

Gibb, Robbin
Brain Development: Ontogenetic Neurobiology of
Prefrontal Cortex Development and Development of Cognitive Function

Gieryn, Thomas F.
Science, Sociology of

Giest, Sarah
Policy Cycle

Gigerenzer, Gerd
Computers: Impact on the Social Sciences
Decision Making: Nonrational Theories

Gil, David
Quantifiers

Gilbert, Dennis
Income Inequality

Gilbert, John K.
Science Education

Gilbert, Margaret
Joint Action

Gilbert, Nigel
Self-Organizing Dynamical Systems

Gilboa, Asaf
Retrieval

Gilgun, Jane F.
Research and Theory Building in Social Work

Gillis, J.R.
Life Course in History

Gilman, Antonio
Marxist Archaeology

Gilmour, John B.
Congress: United States

Gilroy, Paul
Du Bois, William Edward Burghardt (1868–1963)

Gilson, Médhi
Sleep and Memory

Gilula, Z.
Multivariate Analysis: Discrete Variables (Correspondence Models)

Giner, Salvador
Mass Society: History of the Concept

Gingrich, Andre
Comparative Method in Anthropology
Tribe

Ginsburgh, Victor A.
Art and Culture, Economics of

Ginther, Matthew
Law and Neuroscience

Giordano, Christian
Europe: Sociocultural Overview

Giroux, Henry A.
Disposable Youth in America in the Age of Neoliberalism

Gitelman, Alix I.
Experimental Design: Compliance

Giuffre, Katherine
Cultural Production in Networks

Giugni, Marco
Peace Movements

Giuliano, Genevieve
Transportation Geography

Glancy, Graham
Forensic Psychiatry

Glasmeier, Amy K.
Industrial Geography

Glasner, Peter
Genetics and Society

Glassman, Jim
Development Theory in Geography

Glendinning, Caroline
Community Care in England

Glock, Hans-Johann
Meaning and Rule Following

Glover, Vivette
Stress: Effects on Neonatal and Child Development

Glück, Judith
Wisdom, Psychology of

Gobet, Fernand
Vocabulary Acquisition

Goble, Priscilla
Gender and Education

Goddard, Cliff
Language and Society

Godden, Chris
Economic History

Godfrey-Smith, Peter
Pragmatism: Philosophical Aspects

Goetz, Cari D.
Human Mate Choice, Evolution of

Goix, Renaud Le
Gated Communities

Golden, Richard M.
Statistical Pattern Recognition

Goldenberg, Georg
Apraxia

Goldfarb, Eva S.
Sexuality Education

Goldman, Noreen
Mortality Differentials: Selection and Causation

Goldscheider, Frances K.
Repartnering and Stepchildren

Goldscheider, Frances
Fatherhood

Goldson, Barry
Juvenile Justice: International Law and Children's Rights

Goldstein, Lynne
Archaeology, Politics of

Goldsworthy, Rachael
Hallucinations, Neural Basis of

Göle, Nilüfer
Near Middle East/North Africa Studies: Gender

Golla, Victor
Sapir, Edward (1884–1939)

Gollier, Christian
Insurance

Gollwitzer, Peter M.
Motivation and Actions, Psychology of
Motivation: History of the Concept

Gombrich, Richard
Buddhism

Gomery, Douglas
Film and Video Industry

Gomila, Antoni
Language and Thought: The Neo-Whorfian Hypothesis

Goncalo, Jack A.
People in Organizations

Gondal, Neha
Networks and Meaning

Gonzales, Joseph
Longitudinal Analyses of Sexual Development through Early Adulthood

Gonzalez, Claudia L.R.
Parietal Lobe

Gonzalez, Jorge A.
Diversity in organizations

González, Norma
Anthropological Approaches to the Study of Education: The United States and Beyond

Gonzalez, Richard
Experimental Laboratories: Social and Economic

González-Bailón, Sandra
Social Protest and New Media

Good, Byron
Belief, Anthropology of

Goodale, Mark
Human Rights, Anthropology of

Goodall, Janet
Parents and Teachers as Partners in Education

Goodchild, Lester F.
Hall, Granville Stanley (1844–1924)

Goodchild, Michael F.
Geographic Information Systems

Goodkind, Daniel
Sex Selective Abortion

Goodman, Roger
Education: Anthropological Aspects

Goodwin, Michele
Law, Bioethics, and Biotechnology

Goody, Esther N.
Craft Production, Anthropology of

Gordon, Andrew C.
Campbell, Donald Thomas (1916–96)

Gordon, Barry
Transcranial Direct Current Stimulation and Language

Gordon, Robert W.
Critical Legal Studies

Gordon, Rupa G.
Child-Directed Speech: Influence on Language Development

Gorman-Murray, Andrew
Queer Geography

Gormley, Sarah
Language and Gender

Gorter, Cees
Location Theory

Gosney, James E.
Natural Disasters: Health-Related Aspects

Goswami, Usha
Dyslexia, Developmental

Gotsopoulos, Aleksios
Dominant Designs

Gottfried, Paul
Nisbet, Robert (1913–96)

Gottlieb, Alma
Anthropological Writing

Gottlieb, Benjamin H.
Health-Related Support and Self-help Mutual Aid Groups

Gottschling, Juliana
Intelligence, Genetics of: Cognitive Abilities

Goudsblom, Johan
Fire, Human Use, and Consequences
Sociology, History of

Gould, Nick
Reflexivity

Gower, John C.
Procrustes Analysis

Grabell, Adam S.
Defiant Behavior During Infancy and Early Childhood

Graber, Christoph B.
Intellectual Property: Law in Context

Graburn, Nelson
Art, Anthropological Aspects of

Graco, Warwick
Technological Change and the Sociotechnical System, Applied Psychology of

Graedel, Thomas E.
Industrial Ecology

Graf, Frauke
Early Memory Development and Cultural Variations

Graf, Friedrich W.
Religion, History of

Graff, Lesley
Gastrointestinal Diseases: Psycho-social Aspects

Graf-Nold, Angela
Jung, Carl Gustav (1875–1961)

Graham, Sandra
Race and Academic Motivation

Graham, Steve
Writing Instruction

Graham, Suzanne E.
Survival Analysis: Whether, and If So, When?

Granato, Jim
Political Science and Empirical Implications of Theoretical Models

Grandori, Anna
Organizational Decision Making

Grant, Jill L.
New Urbanism

Grant, Leslie W.
Teacher Behaviours and Student Outcomes

Grant, Phillip
Cerebral Cortex
Hypothalamus
Neurotransmitters

Grassinger, Robert
Passion and Motivation

Grau, Ronald R.
Scientific Discovery, Computational Models of

Grauerholz, Liz
Alternative Food Movements
Collaborative Consumption

Graumann, Carl-Friedrich
Lewin, Kurt (1890–1947)

Gray, DeLeon
Motivation, Learning, and Instruction

Greco, Dirceu B.
Ethical Dilemmas, Research, and Access to Treatment

Greely, Henry T.
Genomics, Ethical Issues in

Green, Eva G.T.
Immigration: Social Psychological Aspects

Green, Kundoqk J.
Indigenous Child Welfare Practice: The Canadian Case

Green, Roy
Innovation

Green, Thomas J.
Cultural Resource Management: Conservation of Cultural Heritage

Greenacre, Michael
Correspondence Analysis

Greenberg, Daniel L.
Autobiographical Memory
Flashbulb Memories

Greenberg, Daniel
Experimental Laboratories: Psychological

Greenberg, Leslie S.
Experiential Psychotherapy

Greenberg, Neil
Military and Disaster Psychiatry

Greene, David L.
Sustainable Transportation

Greenhill, Simon
Evolution and Language: Phylogenetic Analyses

Greenman, Kathleen N.
Childhood and Adolescence: Developmental Assets

Green-Pedersen, Christoffer
Agenda Settting, Public Policy in

Greenspan, Ralph J.
Genes and Behavior: Lessons from the Fruit Fly

Greenwood, Michael J.
Internal Migration in Industrialized Countries

Gregor, A. James
Dictatorship

Gregory, Jane
Science Communication

Greil, Arthur L.
Secular Religions

Grekova, Maya
Postsocialist Societies

Greve, Henrich R.
Behavioral Theories of Organization

Greve, Werner
Human Development, Successful: Psychological Conceptions

Gries, Stefan Th
Quantitative Linguistics

Griffiths, John
Legal Pluralism

Griffiths, Paul E.
Emotion and Expression
Emotion, Evolution of

Grigsby, J.E.
Urban Poverty in Neighborhoods

Grimshaw, Jeremy M.
Cluster Randomized Trials, Ethical Implications of

Grinevald, Colette
Classifiers, Linguistics of

Gripsrud, Jostein
Television: General

Groenen, Patrick J.F.
Multidimensional Scaling II

Grofman, Bernard
Elections: Apportionment

Grosby, Steven
Shils, Edward (1910–95)

Grossbard, Shoshana
Household Economics

Grosser, Travis J.
Structural Equivalence: Meaning and Measures

Grossman, Joel B.
Judicial Review

Grove, M. Annette
Cultural Views of Life Phases

Gruber, Aaron
Optogenetics

Gruber, Hans
Apprenticeships and School Learning: General Considerations

Gruhn, Isebill V.
State Formation, Theory of

Grunenberg, Antonia
Arendt, Hannah (1906–75)

Grüny, Christian
Phenomenology: Philosophical Aspects

Grusky, David B.
Class: Social
Social Stratification

Guadagni, Veronica
Human Spatial Orientation, Neural Basis of

Guadalupe-Diaz, Xavier L.
Gay and Bisexual Latinos in US Gay Culture

Gualtieri, Marie C.
Food Deserts in US Cities

Guarnieri, Carlo
Courts and Adjudication

Guentchéva, Zlatka
Evidentials in Linguistics

Guerra, Pedro
Stress and Cardiac Response

Guillemard, Anne-Marie
Age Policy

Guillén, Mauro F.
International Business

Guillot, Michel
Multistate Transition Models in Demography

Guilmoto, Christophe Z.
Missing Girls: A Globalizing Issue

Guittar, Nicholas A.
Intersectionality

Guittar, Stephanie G.
Intersectionality

Gunder, Michael
Critical Planning Theory

Gundlach, Horst
Critical Psychology

Gunduz-Bruce, Handan
Psychopharmacotherapy: Side Effects

Gunnar, Megan R.
Stress in Infancy and Early Childhood: Effects on Development

Gupta, Taveeshi
School Achievement as Moderated by Sociodemographic Status: The US Case

Guptill, Stephen C.
Spatial Data

Gurney, John
English Revolution, The

Guskey, Thomas R.
Mastery Learning

Gussenhoven, Carlos
Suprasegmentals

Gusterson, Hugh
Elites, Anthropology of

Guth, Suzie
Becker, Howard (1928–)

Guthrie, Chris
Negotiation and Bargaining: Role of Lawyers

Guthrie, Doug
East Asian Studies: Economics

Guyer, Jane I.
African Studies: Culture

Guyer, Joshua J.
Attitudes and Behavior

Gyekis, Joseph P.
Cardiovascular Reactivity and Recovery

H

Habel, Ute
Gender and Health Care

Haberl, Helmut
Energy Flow Analysis

Haberman, Shelby J.
Multivariate Analysis: Discrete Variables (Correspondence Models)
Multivariate Analysis: Discrete Variables (Loglinear Models)

Hackel, Sergei
Orthodox Churches

Hackett, Simon
Child Protection: International Issues

Hadler, Markus
Institutionalism and Neo-institutionalism: History of the Concepts

Hafen, Christopher A.
Classroom Level in Educational Effectiveness Research

Hagan, John
Mass Imprisonment and Its Consequences

Hahn, Adam
Implicit Social Cognition

Hahn, Ulrike
Semantic Similarity, Cognitive Psychology of

Haigh, Rex
Therapeutic Community

Haiman, John
Iconicity in Linguistics

Haining, Robert
Spatial Autocorrelation
Spatial Sampling

Hájek, Alan
Counterfactual Reasoning, Quantitative: Philosophical Aspects

Hajek, Andrea
Oral History

Hakim, Catherine
Erotic Capital
Parallel Relationships

Hakkarainen, Kai
Sociodigital Revolution: Digital Natives vs Digital Immigrants

Hakulinen, Christian
Personality and Health

Halabi, Samer
Helping Intergroup Relations

Halberstadt, Amy G.
Early Emotional Development and Cultural Variability

Hald, Gert M.
Pornography

Halder, Neel
Depression
Mania

Halfwassen, Jens
Plato (428 BCE – 347 BCE)

Hall, Christopher
Narrative in Social Work

Hall, Jeffrey A.
Friendship During Childhood and Cultural Variations

Hall, John R.
Qualitative Methods, History of
Religion, Phenomenology of

Hall, Thomas
Alcohol Interventions: Disease Models vs. Harm Reduction
Veblen, Thorstein (1857–1929)

Hallberg, Kelly
Quasi-Experimental Designs

Halliday, Terence C.
Political Lawyering

Hallin, Daniel C.
Journalism

Hamadani, Jena
Breast-feeding: Effects on Cognitive and Neural Development

Hames, Raymond
Human Behavioral Ecology

Hamhaber, Johannes
Energy, Geography of

Hamilton, Claire E.M.
Punishment

Hamilton, Derek A.
Navigation in Virtual Space: Psychological and Neural Aspects

Hamilton, James F.
Citizen Journalism

Hamilton, Malcolm
Prophetism

Hamlin, Alan
Contractarianism

Hamm, Alfons O.
Fear-Potentiated Startle
Phobias Across the Lifespan

Hammersley, Rachel
Political Thought, History of

Hammond, Stuart I.
Prosocial Behavior during Childhood and Cultural Variations
Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations

Hampsher-Monk, Iain
Mill, John Stuart (1806–73)

Han, Sim Jing
Subjective Wellbeing and Culture

Hancock, Gregory R.
Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis

Hancock, Katherine
Crime, Immigration, and Ethnicity

Hanczakowski, Maciej
Eyewitness Testimony

Händel, Richard B.
Data Bases and Statistical Systems: Education, General

Handelman, Don
Play, Anthropology of

Handler, Richard
Authenticity, Anthropology of

Hankir, Ahmed
Cultural Psychiatry

Hann, Chris
Gellner, Ernest (1925–95)
Property: Anthropological Aspects

Hannah, Lauren A.
Stochastic Optimization

Hannam, Kevin
Cultural Geography

Hannan, Michael T.
Organizational Ecology

Hanna-Pladdy, Brenda
Dysexecutive Syndromes

Hannelly, Toni
Environmental Health and Safety: Social Aspects

Hannerz, Ulf
Anthropology: Overview
Center–Periphery Relationships

Hannula, Deborah E.
Recognition

Hansen, Jochim
Money and Finances, Psychology of

Hansen, Karen T.
Informal Sector

Hansen, Lars P.
Method of Moments and Generalized Method of Moments

Hanson, Heidi A.
Early Life Influences on Health and Mortality in Adulthood

Hanson, Kristin
Language and Poetic Structure

Hanushek, Eric A.
Education, Economics of

Harackiewicz, Judith M.
Academic Motivation and Performance: Task Value Interventions

Harari, Edwin
Psychotherapy, History of: Psychiatric Aspects

Harbecke, Jens
Mental Causation

Harcourt, Geoffrey C.
Robinson, Joan (1903–83)

Hardiess, Gregor
Virtual Reality and Spatial Cognition

Hardin, Russell
Rational Choice Explanation: Philosophical Aspects

Harding, Andrew J.
Legal Traditions of Southeast Asia

Harding, Stephan
Gaia

Hardtwig, Wolfgang
Burckhardt, Jacob (1818–97)
von Ranke, Leopold (1795–1886)

Hardy, Cynthia
Organizations: Authority and Power

Hare, A. Paul
Dramaturgical Analysis

Hareven, Tamara K.
Family and Kinship, History of

Hargens, Lowell L.
Replication

Hargreaves, Desley
Emergency Relief and Social Work Responses

Haring, Sabine A.
Political Religions: History of the Concept

Harkin, M.E.
Potlatch in Anthropology

Harmon, Kara L.
Psychological Treatment, Effectiveness of

Harmon, Shawn H.E.
Genetics: Legal Aspects

Harms, Teresa A.
Time Use Research: Recent Developments

Harper, Douglas
Visual Methods in the Social Sciences

Harré, Rom
Mind–Body Dualism

Harré, Ron
Individuals in Society: History of the Concept

Harrington, C. Lee
Video Cultures: Television

Harrington, Ryan
Data Bases and Statistical Systems: Law

Harris, Alma
Teacher Leadership

Harris, Angela P.
Critical Race Theory

Harris, Britten
Urban Growth Models

Harris, David J.
Genetic Engineering

Harris, John N.
Power, Politics, and Influence in Organizations

Harris, Julia A.
Emotion in Cognition

Harris, Kelci
Conscientiousness

Harris, Richard
Suburbanization and Suburbanism

Harrison, John
Regional Geography

Harsanyi, Tom P.
Social Democracy

Hart, John
Presidency, United States

Hart, Keith
Cultural Critique, Anthropological

Hart, Michael A.
Indigenous Social Work

Härtel, Charmine E.J.
Indigenous Management Styles

Härter, Martin
Somatoform Disorders

Harter, Susan
Self-Development in Childhood and Adolescence

Hartigan, John A.
Statistical Clustering

Hartl, Amy C.
Conflict and Socioemotional Development

Hartmann, Armin
Psychotherapy Process Research

Hartmann, Michael
Elites: Sociological Aspects

Hartwick, Elaine R.
Commodity Chains

Harzem, Peter
Watson, John Broadus (1878–1958)

Hasenfratz, Liat
Prosocial Behavior, Effects of Parenting and Family Structure on

Hasha, Margot
Misdemeanor Offenses

Haslam, Alexander
Obedience (Research), History of

Haslam, Nick
Essentialism

Haslam, S. Alexander
Leadership
Self-Categorization Theory
Tyranny

Hass, Jeffrey
Eastern European Studies: Society
Russian and Post-Soviet Studies: Politics

Hassall, Maureen
Human Factors and Ergonomics

Hassel, Anke
Public Policy

Hastings, Adrian
Catholicism

Hatta, Takeshi
Reading Nonalphabetic Scripts

Hattie, John
Dimensionality of Tests: Methodology

Hatton, Kieron
Community Work

Hatton, Timothy J.
Migration Out of Europe

Hau, Kit-Tai
Academic Self-Concept and Achievement

Haupt, Heinz-Gerhard
Comparative History
Violence, History of

Hauptman, Samantha
Policing

Hauschild, Thomas
Home, Anthropology of

Hausman, Daniel M.
Economics, Philosophy of

Hautus, Michael
Signal Detection Theory

Haveman, Robert H.
Poverty: Measurement and Analysis

Hay, Peter
Civil Law

Hayes, Jarrod
Sexual Minorities in the Maghreb

Hayes, Joshua D.
Religion, Phenomenology of

Hayes, Joy E.
Radio as a Medium

Hayne, Harlene
Childhood Memories
Memory Development During Infancy and Early Childhood across Cultures

Haysom, Hannah J.
Personality, Evolutionary Models of

Hayter, Roger
Resource Geography

Hayward, Katy
Conflict and Consensus

Hazelton, Martin L.
Nonparametric Regression

He, Monica
Immigration: Health and Mortality

He, Shenjing
Right to the City: A Liberal-Democratic Perspective

Head, Brian W.
Policy Analysis: Evidence Based Policy-Making

Headrick, Daniel
International Communication: History

Healey, Patsy
Planning Theory: The Good City and Its Governance

Heath, Gregory W.
Physical Activity and Health Promotion

Heath, Linda
Triangulation: Methodology

Hechter, Michael
Solidarity, Sociology of

Heckerman, David
Bayesian Graphical Models and Networks

Heckert, Jamie
Queer Anarchism

Heckhausen, Jutta
Control Behavior: Psychological Perspectives
Lifespan Development, Theory of

Hedges, Larry V.
Meta-Analysis: Theory

Hedström, Peter
Analytical Sociology

Heffernan, Paul
Organizational Emergence and Firm Formation

Heft, Harry
Human Cognition, Evolution of

Hegarty, Peter
Social Psychological Theory, History of

Heilbron, Johan
Interest: History of the Concept
Sociology, History of

Heim, Bradley T.
Labor Supply

Heim, Sabine
Language Learning Impairment

Heimer, Carol A.
Law, Sociology of

Heimer, Karen
Crime and Gender

Hein, Carola
Urban Planning, Competitions and Exhibitions

Heinonen, Erkki
Psychotherapy Process Research

Heinrichs, Nina
Evidence-Based Psychotherapy

Heinsohn, Gunnar
Genocide, Historical Aspects of

Heise, David R.
Scaling and Classification in Social Measurement

Heit, Evan
Semantic Similarity, Cognitive Psychology of

Heleniak, Timothy
Population Change in the Former Communist States of Europe and Asia

Helkama, Klaus
Moral Development, Cultural Differences in

Hellwig, Susan
Cardiac Disease, Coping with

Helm, Friederike
Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion

Helmke, Andreas
School Achievement: Motivational Determinants and Processes

Henderson, J.H.
Mental Illness: Family and Patient Organizations in Europe

Henderson, Kathryn A.
Data Bases and Statistical Systems: Applied Social Research

Hendrix, Joshua A.
Crime Trends and Debates

Henkin, Samuel
War and Terrorism, Geography of

Henley, Tracy B.
Psychology, History of (Early Period)

Hennecke, Marie
Self-Regulation in Adulthood

Henneman, Lidewij
Genetic Screening for Disease

Hennock, E.P.
Welfare State, History of

Henrich, Joseph
Prosocial Behavior, Cultural Differences in

Henrickson, Mark
Sexuality and Social Work

Henry, Stuart
Legal Systems: Private

Hensler, Deborah R.
Class Actions, Legal

Hentschel, Tanja
Group Processes in Organizations

Herber, Oliver
Rehabilitation (Cardiac): Psychosocial Aspects

Herbert, David
Crime, Geography of

Heredia, Mariana
Social and Behavioral Sciences under Dictatorship

Héritier, Adrienne
Public Goods: International

Herk, Gerard Van
Sociolinguistics

Hermelin, Brita
Service Economy, Geography of

Hermerén, Göran
Risk Screening, Testing, and Diagnosis: Ethical Aspects

Hermias, Joshua
Climate Change and Global Poverty

Hernán, Miguel A.
Longitudinal Causal Inference

Hernes, Gudmund
Social Changes: Models

Herrera, Rémy
Defense and Security Economics

Herring, Susan C.
Teens, Gender, and Self-Presentation in Social Media

Hershberg, Rachel M.
Human Development, Theories of

Hertog, Sara
Mortality and the HIV/AIDS Epidemic

Hertwig, Ralph
Heuristics, History of

Hervik, Peter
Xenophobia and Nativism

Herzfeld, Michael
Common Sense, Anthropology of

Hesketh, Beryl
Technological Change and the Sociotechnical System, Applied Psychology of

Heslin, Alison
Sustainable Agriculture

Hess, Andreas
Caro Baroja, Julio (1914–95)
Shklar, Judith N (1928–92)
Veblen, Thorstein (1857–1929)

Heuer, Eric
Serial Order, Representation of

Heukelom, Floris
Prospect Theory

Heuveline, Patrick
Fertility Transition: Southeast Asia

Heverly-Fitt, Sara
Friendship During Infancy and Early Childhood and Cultural Variations

Hewig, Johannes
Personality Assessment

Heydemann, Steve
Near Middle East/North Africa Studies: Politics

Heyman, Gene M.
Opiate Use and Abuse, History of

Hicks, David
Sacrifice

Hideg, Ivona
Emotional Intelligence and Competencies

Hiebaum, Christian
Egalitarianism

Hietajärvi, Lauri
Sociodigital Revolution: Digital Natives vs Digital Immigrants

Higgins, George E.
Self-Control Theory and Crime

Highton, Benjamin
Voting: Turnout

Hilbig, Benjamin E.
Honesty and Humility

Hildebrandt, Andrea
Facial Emotion Expression, Individual Differences in Facial Perception

Hildermeier, Manfred
Russian Revolution, The

Hilgartner, Stephen
Biotechnology

Hilgendorf, Eric
Albert, Hans (1921–)

Hilger, Anna-Lena
Factor Analysis and Latent Variable Models in Personality Psychology

Hill, Bruce
de Finetti, Bruno (1906–85)

Hill, Jean L.
Prevention Research

Hill, Jennifer
Causal Inference: Overview

Hill, Lori Diane
Educational Research, The Field of

Hill, Michael
Halévy, Élie (1870–1937)

Hill, Thomas P.
Data Bases and Statistical Systems: Economics (Macroeconomics)

Hilliard, Lacey J.
Human Development, Theories of

Himes, Christine L.
Aging and Health in Old Age

Hinde, Robert A.
Bowlby, John (1907–90)

Hines, James R.
Corporate Taxation

Hintsanen, Mirka
Temperament and Motivation

Hipp, John R.
Collective Efficacy and Crime

Hirsch, Jennifer L.
Social Relationships in Adulthood

Hirschman, Charles
Fertility Transition: Southeast Asia

Hirsh, Jacob B.
Personality and Creativity

Hirt, Sonia A.
Planning During Post-Socialism

Hitchcock, Christopher
Probability and Chance: Philosophical Aspects

Hitlin, Steven
Motivation: Life Course and Sociological Perspectives

Ho, Karen
Finance, Anthropology of

Hoaglin, David C.
Exploratory Data Analysis: Univariate Methods

Hobbs, Megan J.
Evidence-Based Psychiatry

Hochfelder, David
Telegraph

Hockley, William E.
Context Dependent Memory

Hodgkinson, Gerard P.
Managerial and Organizational Cognition

Hodgson, Dennis
Demography: History Since 1900

Hodgson, Jacqueline
Plea Bargaining: A Comparative Analysis

Hoefeijzers, Serge
Transient Epileptic Amnesia

Hoelscher, Michael
Policy Knowledge: Universities

Hoem, Jan M.
Life Table

Hoenig, Barbara
Gatekeepers in Social Science
Ignorance, History of Concept
Reference Group, History of

Hoenig, Jennifer M.
School Environment for LGBTQ/Sexual Minority Youth

Hoey, Brian A.
Postindustrial Societies

Hoeyer, Klaus
Biobanking: Ethical Issues

Hofer, Jan
Motivational Development: Cross-Cultural Perspectives

Hoff, Erika
Language Development: Influence of Socio-Economic Status

Hoff, Paul
Bleuler, Eugen (1857–1939)
Kraepelin, Emil (1856–1926)

Höffe, Otfried
Intercultural Legal Discourse: Philosophical Aspects
Pluralism and Tolerance
World Republic

Hofferth, Sandra
Fatherhood

Hoffman, Sara A.
Evidence-Based Practice

Hoffmann, Achim G.
Artificial and Natural Computation

Hoffmann, John P.
Religion: Morality and Social Control

Hoffmeyer-Zlotnik, Jürgen H.P.
Data Bases and Statistical Systems: Sociology

Hofmann, Stefan G.
Social Phobia across the Lifespan

Hogan, Bernie
Social Media

Hogg, Eddy
Social Science and Civil Society

Hogg, Michael A.
Group Processes, Social Psychology of

Hoipkemier, Mark
Citizenship: Political

Hokari, Toru
Cooperative Game Theory

Holec, Victoria
Hypnosis, Neural Basis of

Holland, Paul W.
Causal Counterfactuals in Social Science Research

Hollander, Myles
Nonparametric Statistics: Advanced Computational Methods
Nonparametric Statistics: Rank-Based Methods

Hollander, Rachelle
Mentoring: Supervision of Young Researchers

Hollos, Marida
Infertility: Demographic Aspects

Holloway, Margaret
Spirituality

Holloway, Ralph L.
Brain, Evolution of

Holmes, Leslie
Totalitarianism

Holmes, Malcolm D.
Crime, Race and Ethnicity

Holsboer-Trachsler, Edith
Sleep Disorders: Psychiatric Aspects

Holsti, Ole R.
Alliances: Political

Holt-Lunstad, Julianne
Social Integration, Social Networks, and Health

Holtorf, Cornelius
Archaeology and Cultural Memory

Holzer, Jerzy
Communism, History of

Hon-Fai, Chen
Illich, Ivan (1926–2002)

Hong, Ryan Y.
Five Factor Model of Personality, Facets of

Honig, Jan W.
Military, War, and Politics

Honkapohja, Seppo
Expectations, Economics of

Hood, Heather K.
Panic Disorder

Hooper, Paul L.
Fertility Theory: Embodied-Capital Theory of Life History Evolution

Hoover, Kevin D.
Macroeconomics, History of From 1933 to Present

Hopcroft, Rosemary L.
Evolutionary Sociology

Hopkins, Nick
Identity and Identification, Social Psychology of

Horiuchi, Shiro
Age Structure

Horn, Andrea B.
Emotions and Health

Horn, Ilana S.
Teaching as a Profession: United States

Hornborg, Alf
Industrial Societies

Horne, Sharon G.
Sexual Minority Families of Choice and Family Networks: We Are Family

Hornung, Bernd R.
Sociocybernetics

Horowitz, Joel L.
Semiparametric Models

Horwitz, Ann M.
The American Conservative Movement

Hoskins, Janet
Symbolism in Anthropology

Hotaling, Jared M.
Dynamic Decision Making

Houck, Judith A.
Medicine and Health for Sexual Minorities

Houdé, Olivier
Cognitive Development During Infancy and Early Childhood across Cultures

Hough, Mike
Compliance and Legal Authority

Howe, Cymene
Queer Anthropology

Howes, David
Senses, Anthropology of the

Howes, Oliver D.
Magnetic Resonance and Molecular Imaging in Psychiatry

Howlett, Michael
Policy Cycle

Høyer, Svennik
News: General

Hoyer, William J.
Brain Aging (Normal): Behavioral, Cognitive, and Personality Consequences

Hoyningen-Huene, Paul
Kuhn, Thomas S. (1922–96)

Hroch, Miroslav
Nationalism, Historical Aspects of: The West

Hruschka, Daniel J.
Prosocial Behavior, Cultural Differences in

Hsu, Hui-Chin
Play During Infancy and Early Childhood: Cultural Similarities and Variations

Hu, Mei-Chih
Industrial Dynamics: Patenting Perspectives
Research and Development in Organizations

Huang, Chu-Ren
Corpus Linguistics

Huang, Yan
Anaphora

Hubbard, Katherine
Social Psychological Theory, History of

Huber, Machteld
Health: Definitions

Huberfeld, Nicole
Health Care: Legal Issues (United States)

Hudek-Knežević, Jasna
Stress in Adolescence: Effects on Development

Hudson, Valerie M.
Women's Status and National and International Security

Huebner, Daniel R.
Instinct, History of
Mead, George Herbert (1863–1931)

Hugdahl, Kenneth
Dichotic Listening and Language: Overview

Hughes, Kirsty
Privacy: Theoretical and Legal Issues

Hughes, Lorine A.
Crime, Sociology of
Criminology, History of
Gangs, Sociology of

Hughes, M.A.
Urban Life: Future Developments

Hughes-Freeland, Felicia
Performance: Anthropological Aspects

Hugo, Graeme
International Migration by Ethnic South and Southeast Asians

Hui, Cang
Carrying Capacity of the Environment

Hull, David L.
Evolutionary Epistemology

Hulleman, Chris S.
Expectancy-Value-Cost Model of Motivation

Hume, Therese
Environmental Education and Education for Sustainable Development

Humiston, Gail S.
Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices

Hummer, Robert A.
Immigration: Health and Mortality

Humphreys, Richard S.
Historical Thought and Historiography: Islamic Tradition

Humphries, Marisha L.
Prosocial Behavior during Adolescence

Humphries, Michael L.
Christianity, Origins of

Hung, Andrew Tsz Wan
Huntington Samuel P. (1927–2008)
Sandel, Michael (1953–)

Hunt, L.
French Revolution, The

Hunter, Albert
Data Collection: Interviewing

Hurrelmann, Klaus
Welfare and Education

Husbands, Christopher T.
Political Parties: Extreme Right

Huston, A.C.
Child Care and Child Development

Hutchinson, Mark
Evangelical, Revivalist, and Pentecostal Religions

Hutter, Marcus
Algorithmic Complexity

Huutoniemi, Katri
Peer Review: Organized Skepticism

Huxford, John
Visual Images in the Media

Hwang, Kwang-Kuo
Morality 'East' and 'West': Cultural Concerns

Hyers, Lauri
Stigma, Social Psychology of

I

Iaria, Giuseppe
Human Spatial Orientation, Neural Basis of

Iarocci, Grace
Social Competence During Adolescence Across Cultures

Iarskaia-Smirnova, Elena
Social Work in Russia

Iborra, Alejandro
Self and Identity Development during Adolescence across Cultures

Idler, Ellen L.
Religion and Health: Sociological Considerations

Iggers, Georg Gerson
Historical Thought and Historiography: Modern History (Since the 18th Century) – the West

Iggers, Georg G.
Historical Thought and Historiography: Current Trends

Igl, Gerhard
Social Insurance: Legal Aspects

Ignagni, Esther
Disability and Disabling Practices

Ignatow, Gabe
Embodiment and Culture

Ihara, Yasuo
Human Mate Choice

Imuta, Kana
Childhood Memories
Memory Development During Infancy and Early Childhood across Cultures

Inglis, Fred
Geertz, Clifford (1926–2006)

Inoguchi, Takashi
Social Science Infrastructure: East Asia and the Pacific (Research and Teaching)

Inwood, Joshua
Race and Racism, Geography of

Ioakimidis, Vasilios
Neocolonialism and Social Work

Iqbal, Hafiz M.
Cognitive Development, Educational Theories of

Iranzo, Juan M.
Collins, Randall (1941–)

Irwin, Alan
Science, Public Engagement with

Isaacowitz, Derek
Emotion in Cognition

Isgett, Suzannah F.
Broaden-and-Build Theory of Positive Emotions

Islamoglu, H.
Civil Society, Concept and History of

Ismail, Salma
Equity and Education

Ispas, Dan
Personnel Selection, Psychology of

Isphording, Ingo E.
Language and Labor Market Success

Issacharoff, Samuel
Due Process in Law

Isserman, Andrew
Regional Science

Ito, Y.
Social Influence, Social Psychology of

Ivanco, Tammy L.
Long-Term Potentiation and Long-Term Depression

Iveniuk, James
Heterosexuality

Ivison, Duncan
Multiculturalism

J

Jackson, Crystal A.
Prostitution and Sex Work

Jackson, Frank
Reference and Representation: Philosophical Aspects

Jackson, Jonathan
Compliance and Legal Authority

Jackson, Joshua J.
Conscientiousness

Jackson, Myles W.
Patenting in the Biological Sciences

Jacobs, Arthur M.
Visual Word Recognition, Neurocognitive Psychology of

Jacobs, Brian
Community Sociology

Jacobs, Jane M.
Postcolonial Geography

Jacobs, Marion K.
Psychological Treatments Delivered Nontraditionally

Jacobson-Widding, Anita
Color Classification and Symbolism

Jacoby, Jacob
Consumer Psychology

Jäger, Urs
Social Business

Jagers, Peter
Population Dynamics: Probabilistic Extinction, Stability, and Explosion Theorems

Jahanshahi, Marjan
Non Declarative (Procedural) Memory

Jaimie, M. Krause
Learning to Learn

Jambon, Marc
Moral Development, Theories of

James Huang, C.-T.
Movement Theory and Constraints in Syntax

James, Wendy
Evans-Pritchard, Sir Edward E. (1902–73)

Jamison, Andrew
Science and Social Movements

Jampolsky, Jacquelyn A.
Indigenous Rights

Janelle, Donald G.
Time-Space in Geography

Jang, Daisung
Emotion, Perception and Expression of

Jang, Sung Joon
Strain Theories and Crime

Janis, Mark W.
International Law and Treaties

Janoski, Thomas
Citizenship, Sociological Aspects of

Jansen, Maarten
Model Selection and Model Averaging

Janssen, Susanne
Cultural Mediators and Gatekeepers
Internet and Culture

Jansz, Janis
Environmental Health and Safety: Social Aspects

Jarausch, Konrad H.
Quantification in History

Jargowsky, Paul A.
Concentrated Disadvantage
Underclass

Jarvin, Linda
Binet, Alfred (1857–1911)

Jasanoff, Sheila
Science and Law

Jasinskaja-Lahti, Inga
Ethnic Identity, Psychology of

Jaspal, Rusi
Non-Heterosexual British Asian Men: Social Representations, Identity and Social Relations

Jasper, James M.
Moral Panics

Jasso, Guillermina
Inequality Analysis: Overview
Societies, Types of

Jayaram, Narayana
Illiteracy, Sociology of

Jayasankar, Kizhavana P.
Documentary and Ethnographic Film

Jayatillake, Rasika
Transcranial Direct Current Stimulation and Language

Jeanpierre, Laurent
Associations, Scholarly

Jenkins, Richard
Ethnicity: Anthropological Aspects

Jennings, M. Kent
Politics and Socialization

Jennings, Wesley G.
Drugs: Illicit Use and Prevention

Jensen, Gary F.
Delinquency, Sociology of

Jensen, Klaus B.
Semiotics

Jerabek, Hynek
Empirical Social Research, History of

Jermier, John M.
Workplace Autonomy

Jessner-Schmid, Ulrike
Multilingualism

Jiménez-Buedo, María
Quasi- and Field Experiment

Jirout, Jamie
Scientific Concepts: Development in Children

Jivraj, Tania
Disability and Disabling Practices
Mental Health, Mentalism and Sanism

Joas, Hans
Pragmatist Social Thought, History of
War and Democracy

Joe, Harry
Majorization and Stochastic Orders

Joh, Sung W.
Cross-National Comparisons in Education: Findings from PISA

Johansen, Søren
Time Series: Cointegration

John, Richard R.
Postal Systems

John, Ulrich
Addictions: General Considerations

Johns, Adrian
Printing as a Medium

Johnson, Burke
Mixed Methods

Johnson, Carol
Sexual Minorities and Sexual Citizenship

Johnson, Cathryn
Legitimacy, Sociology of

Johnson, David A.
Regional Planning, History of

Johnson, David S.
Income Distribution and Inequality: Measurement Issues

Johnson, David W.
Cooperation and Competition

Johnson, Earl
Justice, Access to

Johnson, Fiona
Weight Regulation

Johnson, Mark S.
Russian and Post-Soviet Studies: Education

Johnson, Mark
Probabilistic Grammars and their Applications

Johnson, Michael D.
Customer Satisfaction

Johnson, Monica K.
Adolescence, Sociology of

Johnson, Peter
Self-Organizing Dynamical Systems

Johnson, Roger T.
Cooperation and Competition

Johnson, Russell E.
Core Self-Evaluations

Johnson, Sara K.
Human Development, Theories of

Johnson-Hanks, Jennifer
Fertility Transition: Cultural Explanations

Johnson-Laird, Philip N.
Problem Solving and Reasoning, Psychology of

Johnston, Judy
Public Administration: Organizational Aspects

Johnston, Lynda
Sexuality and Space

Johnston, Ron
Electoral Geography
Geography: Overview

Johnston, Winter
Evidence-Based Practice

Johnstone, Richard
Occupational Health and Safety, Regulation of

Jolliffe, Darrick
Personality and Crime

Jolly, Debbie
Disability Movement

Jonas, Andrew E.G.
Region: Geographical

Jonas, Eva
Money and Finances, Psychology of

Jonas, Michael
Sustainable Practices

Jonason, Peter K.
Dark Triad: The "Dark Side" of Human Personality

Jones, Abigail
Climate Change and Global Poverty

Jones, Andrew
Globalization: Geographical Aspects

Jones, David N.
Global Agenda for Social Work and Social Development
Social Work and Disasters

Jones, Karen
Trust: Philosophical Aspects

Jones, Lyle V.
Hypothesis Tests, Multiplicity of

Jones, Martyn
Rehabilitation (Cardiac): Psychosocial Aspects

Jones, Nicola
Cash Transfers

Jones, Owen D.
Law and Neuroscience

Jones, Rachel K.
Abortion, Demography of: United States

Joppke, Christian
Immigration: Political Aspects

Joralemon, Donald
Shamanism

Jordan, Andrew
Environmental Policy: Protection and Regulation

Jordan, Christian H.
Self-Esteem

Jordhus-Lier, David
Labor, Geography of

Joseph, Brian D.
Language Contact

Joss, Simon
Eco-cities and Sustainable Urbanism

Joukowsky, Martha Sharp
Survey and Excavation (Field Methods) in Archaeology

Jourdan, Christine
Creolization: Sociocultural Aspects

Joyce, Rosemary A.
Ritual and Symbolism, Archaeology of

Juárez, Fatima
Fertility Transition: Latin America and the Caribbean

Judd, Charles M.
Moderator Variable: Methodology

Juergensmeyer, Mark
Religious Nationalism and the Secular State

Julnes, George
Realism, in Applied Research

Juncos-Rabadán, Onésimo
Bilingual Aphasia

Jung, Kwanghee
Multidimensional Scaling I

Junker, Brian W.
Factor Analysis and Latent Structure: IRT and Rasch Models

Junker, Laura L.
Chiefdoms, Archaeology of

Juraska, Janice M.
Brain: Response to Enrichment

Juster, F. Thomas
Data Bases and Statistical Systems: Economics (General)

K

Kaas, Jon H.
Neural Plasticity
Topographic Maps

Kaase, Max
Data Bases and Statistical Systems: Political Science (General)

Kaelble, Hartmut
Social Mobility, History of

Kaga, Makiko
Neurophysiological Markers of Language Development

Kagitcibasi, Çigdem
Parenting Attitudes and Beliefs across Cultures

Kahn, Joel S.
Political Economy in Anthropology

Kairys, David
Civil Rights

Kalb, Don
Global South

Kalbfleisch, John D.
Survival Analysis: Overview

Kalekin-Fishman, Devorah
Alienation, History of

Kalich, DeAnn
Misdemeanor Offenses

Kalindi, Sylvia C.
Education in Sub-Saharan Africa

Kalish, Rachel
Masculinity, Scripts of

Källberg-Shroff, Malin
Culture and the Self: Implications for Psychological Theory

Kalleberg, Ragnvald
Scientific Misconduct, Plagiarism, and Institutional Control of Misconduct

Kallis, Giorgos
Degrowth

Kalter, Frank
Action, Theories of Social

Kalz, Marco
Lifelong Learning and Its Support with New Technologies

Kameda, Tatsuya
Evolutionary Group Dynamics

Kanai, Juan M.
Urban System in Geography

Kandler, Christian
Personality Differences and Development: Genetic and Environmental Contributions

Kapidzic, Sanja
Teens, Gender, and Self-Presentation in Social Media

Kaplan, Hillard S.
Fertility Theory: Embodied-Capital Theory of Life History Evolution

Kaplan, Hillard
Fertility Theory: Theory of Intergenerational Wealth Flows

Kaplan, Robert M.
Health Outcomes, Assessment of

Karady, Victor
Mauss, Marcel (1872–1950)

Kardum, Igor
Stress in Adolescence: Effects on Development

Karlsson, Bengt G.
Political Ecology: Anthropological Perspectives

Karmiloff-Smith, Annette
Williams Syndrome

Karney, Benjamin R.
Personality and Marriage

Karpik, Lucien
Political Lawyering

Karpov, Yuriy V.
Neo-Vygotskian Developmental Theory

Karraker, Amelia
Heterosexual Expression at Older Ages

Kaspersen, Lars B.
War, Sociology of

Kasselimis, Dimitrios S.
Interhemispheric Interaction in Language and Cognitive Processes
Language Disorders, Treatment and Remediation of
Neurological Approaches to Agrammatism

Kates, Robert W.
Sustainability and Sustainability Science

Kathmann, Norbert
Obsessive-Compulsive Disorder across the Life Span

Katschnig, Heinz
Sociology and Psychiatry

Katz, James E.
Telephone

Katz, Joel
Phantom Limbs

Katz, Steven
Networking for School Improvement

Kaufman, David J.
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

Kaufman, Gordon M.
Statistical Identification and Estimability

Kaufman, Jordy
Spatial Cognition during Infancy and Early Childhood Across Cultures, Development of

Kaufmann, Matthias
Kantian Ethics and Politics

Kaul, Inge
Global Public Policy

Kawashima, Nobuko
Cultural Policies in East Asia

Kay, Fiona M.
Lawyers: Social Organization of the Profession

Kay, Paul
Color Terms, Linguistics of

Kazemi, Ali
Justice: Social Psychological Perspectives

Keating, Alexander M.
Development and Urbanization

Keating, W.D.
Neighborhood Revitalization and Community Development

Keenan, Janice M.
Psychology of Inferences

Keenan, William J.F.
Monasticism: West and East

Keene, John C.
Environmental Planning

Keeves, John P.
Comparative Research in Education: IEA Studies

Keiding, Niels
Event History Analysis in Continuous Time
Survival Analysis: Introduction

Keighren, Innes M.
Environmental Determinism

Keister, Lisa A.
Religion and Stratification

Keith, Linda C.
Political Repression

Kellaghan, Thomas
Family and Schooling

Keller, Heidi
Lifespan Development: Evolutionary Perspectives

Kelley, Donald R.
Historicism

Kellman, Elaine
Rawls, John (1921–2002)

Kelly, John D.
Postcoloniality

Kelly, Paul
Intergenerational Justice

Kelman, Ilan
Community Resilience to Disasters

Keltikangas-Järvinen, Liisa
Personality and Health
Temperament and Motivation

Kemmelmeier, Markus
Authoritarianism

Kemp, Cecil D.
Distributions, Statistical: Special and Discrete

Kendal, Jeremy R.
Gene–Culture Coevolution

Kendal, Rachel L.
Social Learning and Culture in Nonhuman Organisms

Kendrick, Andrew
Residential Child Care

Kennan, John
Strikes

Kennedy, Michael D.
Eastern European Studies: Culture

Kenney, Martin
Venture Capital

Kenrick, Douglas T.
Evolutionary Social Psychology

Kent, Jack W.
Familial Studies: Genetic Inferences

Kentridge, Robert W.
Change Blindness

Kenward, Michael G.
Crossover Designs

Kermond, Christine M.Y.
Supervision, Abusive

Kernberg, Otto F.
Borderline (Patient) Personality

Kernic, Franz
Peace Keeping

Kernik, Melinda
Geographic Information Systems and Remote Sensing

Kerodal, Ashmini G.
National Security and Domestic Terrorism: The US Case

Kerr, P.M.
Robinson, Joan (1903–83)

Kets de Vries, Manfred
Leadership in Organizations, Sociology of

Keuth, Herbert
Logical Positivism and Logical Empiricism

Key, Ellen M.
Pooled Cross-Sectional and Time Series Analyses, in Political Science

Keyfitz, Nathan
Lotka, Alfred James (1880–1949)

Khin, Eindra K.
Ethics and Psychiatry

Khondker, Habibul H.
New Media, Political Mobilization, and the Arab Spring

Khoo, Su-ming
Development Studies

Kiazad, Kohyar
Psychological Contracts

Kickbusch, Ilona
Data Bases and Statistical Systems: Health

Kickbusch, Julian
Welfare and Education

Kiewitz, Christian
Psychological Contracts

Kim, Jeansok J.
Fear: Psychological and Neural Aspects

Kim, Lisa E.
Extraversion

Kim, Min Y.
Emotional Regulation

Kim, Sujeong
Hegemony and Cultural Resistance

Kimmel, Michael
Masculinity, Scripts of

Kina, Victoria J.
Sustainable Development and Social Work

Kinard, Jessica L.
Joint Attention during Infancy and Early Childhood Across Cultures

King, Brian
Environment and Health

King, John E.
Value: History of the Concept

King, Nicelma J.
Race: Ethnicity and Health

King, Ruth
Population Size Estimation and Capture–Recapture Methods

Kingdon, David
Cognitive Behavioral Therapy

Kinzie, J. David
Post-Traumatic Stress Disorder

Kirch, Patrick V.
Ceramics in Archaeology
Pacific Islands, Archaeology of

Kirk, David S.
Data Bases and Statistical Systems: Crime and Delinquency

Kirschenbaum, Howard
Rogers, Carl Ransom (1902–87)

Kirschner, Sophie A.
Epigenetics

Kirtchik, Olessia
Social and Behavioral Sciences under Dictatorship

Kiser, Edgar
War, Social Causes and Consequences

Kissane, Bill
Civil Wars

Kitchin, Rob
Cognitive Maps

Kivivuori, Janne
Criminology, History of

Klaaren, Jonathan
Human Rights: Legal Aspects

Klandermans, Bert
Collective Action

Klaniczay, Gábor
Middle Ages, The

Klauer, Karl Christoph
Implicit Association Test
Model Testing and Selection, Theory of

Klauser, Francisco R.
Surveillance and Privacy, Geography of

Klein, Barbro
Folklore

Klein, Howard J.
Organizational Commitment

Klein, Olivier
Stereotypes in Social Psychology

Klein, Rachael
Pro-Environmental Behavior

Klein, William M.P.
Health Risk Appraisal and Optimistic Bias

Klein, Wolfgang
Lexicology and Lexicography

Kleinig, John
Deserts, Philosophical Concept of

Kleinstuber, Ross
Death Penalty
Genocide: Social and Economic Aspects

Kleit, Rachel Garshick
Community, Social Context of: The US Case

Klima, Edward S.
Sign Language

Klimczuk, Andrzej
Public Policy: Ethics

Kline, K.
Remote Sensing

Kline, Ronald R.
Technological Determinism

Klinger, Ryan
Personality and Values at Work

Klingsieck, Katrin B.
Procrastination

Kloos, Bret
Prevention Research

Klosterman, Richard E.
Urban Planning: Methods and Technologies

Klug, Heinz
Constitutionalism, Comparative

Klusmann, Verena
Health Behaviors, Assessment of

Klyukanov, Igor
Modernization, Sociological Theories of

Knafo, Ariel
Prosocial Behavior, Effects of Parenting and Family Structure on

Knapp, Martin
Psychiatric Care, Economics of

Knight, Carly R.
Mechanism-Based Causal Analysis

Knight, C.G.
Human–Environment Interactions: Case Studies

Knight, David M.
Chemical Science: History and Sociology

Knight, Louise W.
Addams, Jane (1860–1935)

Knight, Rebecca
Adolescent Health and Health Behaviors

Kniskern, Mary K.
Military Sociology

Knopf, Monika
Early Memory Development and Cultural Variations

Knopik, Valerie S.
Behavioral Genetics: Psychological Perspectives

Knowland, Victoria C.P.
Cross-Sectional Methodologies in Developmental Psychology

Knowles, Ryan T.
Social Studies Education

Knutsen, Hege M.
Labor, Geography of

Kocka, Jürgen
Capitalism: The History of the Concept

Koenig, Alexa
Torture: Legal Perspectives

Koenig, Harold G.
Religion and Health: Clinical Considerations and Applications

Koenker, Roger
Quantile Regression

Koerner, E.F. Konrad
Linguistics, History of

Kögler, Hans-Herbert
Frankfurt School: Institute for Social Research

Kogut, Bruce
Multinational Corporations

Koh, Sharon
Subjective Wellbeing and Culture

Kohler, Hans-Peter
Fertility and Social Networks

Kohnen, Saskia
Cognitive Neuropsychology, Methods of

Kohrt, Brandon A.
Child Soldiers

Kok, Gerjo
Health Education and Health Promotion

Kolb, Bryan
Eye Movement, Control of: Oculomotor Control
Ion Channels and Molecular Events in Neuronal Activity
Neurons and Dendrites, Integration of Information in
Prefrontal Cortex Development and Development of Cognitive Function

Prefrontal Cortex
Recovery of Function: Dependency on Age
Temporal Lobe

Kolling, Thorsten
Early Memory Development and Cultural Variations

König, Thomas
Funding of Social Science, History of
Social Science Infrastructure: European Union (Resourcing and Funding)

Kontoleon, Nectarios
Social Networks, Algebra of

Kontula, Osmo
Sex Life Challenges: The Finnish Case

Koopman, Siem J.
Time Series: State Space Methods

Kootstra, Gerrit Jan
Second Language Acquisition

Kopelman, Michael D.
Confabulation

Kordoutis, Panos
Love and Intimacy, Psychology of

Korf, Richard E.
Artificial Intelligence: Combinatorial Search

Korn, Rachel M.
Avoidance and Approach Motivation: A Brief History

Korom, Philipp
Elites: History of the Concept
Network Analysis, History of

Korte, Petra
Pestalozzi, Johann Heinrich (1746–1827)

Koski-Karell, Daniel
Narrative Networks

Kosut, Mary
Tattoos and Body Modification

Kourtit, Karima
Smart Cities

Kovachy, Vanessa N.
Neural Basis of Language Development

Kovera, Margaret B.
Expert Testimony

Kowal, Emma E.
Genetics and Indigenous Communities: Ethical Issues

Kowalski, Robin M.
Cyberbullying

Kozulin, Alex
Vygotsky's Theory of Cognitive Development

Kraak, Menno-Jan
Cartographic Visualization

Kracke, Bärbel
Apprenticeship and School Learning: Lessons from Germany

Kracke, Waud
Culture Shock

Kraft, Tobias
von Humboldt, Alexander (1769–1859)

Kramarek, Michal
Charity and Philanthropy: Overview

Krämer, Gudrun
Nationalism, Historical Aspects of: Arab World

Krapić, Nada
Stress in Adolescence: Effects on Development

Kraus, Alfred
Psychiatry: Anthropological Aspects

Kraus, Uta
Personality Assessment

Krause, Annabelle
Happiness and Work

Krause, Jaimie M.
Learning to Learn

Krause-Jensen, Jakob
Organizational Culture, Anthropology of

Kreissl, Reinhard
Societal Impact Assessment

Kreitz, Christoph
Deductive Reasoning Systems

Kretch, Kari S.
Gibson's Theory of Perceptual Learning

Kretz, Valerie
Media Effects on Children

Kreyenfeld, Michaela
Data Bases and Statistical Systems: Demography

Krick, Annika
Health Promotion in Schools

Krieg, Alexander W.
Personality Development and Temperament

Kristen, Susanne
Theory of Mind During Infancy and Early Childhood Across Cultures, Development of

Kritikos, Alexander S.
Entrepreneurship and Economic Growth

Kritzer, Herbert M.
Litigation
Paralawyers: Other Legal Occupations

Krnel, Dušan
Concept Formation in Childhood: Cross-Cultural Perspectives

Kroeger, Rhiannon
Cohabitation: United States

Kroger, Jane
Identity in Childhood and Adolescence

Krohn, Marvin D.
Life Course Criminology

Krohn-Hansen, Christian
Political Anthropology

Krook, Mona L.
Female Quotas in Parliament: Do They Make a Difference?

Krumholz, Norman
Advocacy and Equity Planning

Krygier, Martin
Rule of Law (and Rechtsstaat)

KS Ng, Henry
Sex Role Development and Education

Ku, Kelly Y.L.
Educational Learning Theory

Kubiak, Thomas
Diabetes: Psychosocial Aspects

Kuby, Daniel
Feyerabend, Paul (1924–94)

Kuenzi, Maribeth
Organizational Climate in the Work Setting

Kuepper, Yvonne
Hypothalamic-Pituitary-Adrenal Axis, Psychobiology of

Kugiumutzakis, Giannis
Neonatal Imitation

Kuh, Diana
Life Course Epidemiology

Kuhn, Randall
Internal Migration: Developing Countries

Kuhn, Tamara J.
Archiving: Ethical Aspects

Kukathas, Chandran
Libertarianism

Kukutai, Tahu
Censuses, Population: Comparative International Aspects

Kulich, Clara
Sexism

Kurtz, Donald V.
Hegemony: Anthropological Aspects

Kurtz-Costes, Beth
Families as Educational Settings

Kurz, Heinz D.
Capital: History of the Concept

Kusch, Martin
Psychologism, History of

Kutner, Nancy G.
Disability: Sociological Aspects

Kuyper, Arend M.
Meta-Analysis: Theory

Kwa, Chunglin
Environmental Sciences

Kwan, Mei-Po
Mobility and Travel Activity Patterns

Kyerewaa Brobbey, Maame
Feminisms and Acculturation around the Globe

Kyriacou, Chris
Teacher Stress and Burnout: Methodological Perspectives

Kyriakides, Leonidas
Educational Effectiveness, The Field of
Educational Effectiveness Theory and Research: Recent Advances
School Improvement

L

La Fontaine, Jean
Family: The Anthropology of the Concept and Its History

La Marca-Ghaemmaghami, Pearl
Gynecological Health: Psychosocial Aspects

Labisch, Alfons
Medicine, History of Western

Labonté, Ronald
Globalization and Health

Labouvie-Vief, Gisela
Self and Emotional Development in Adulthood and Later Life

Labove, Joshua
Law and Geography

Lacey, Nicola
Hart, Herbert Lionel Adolphus (1907–92)

Lachmann, Richard
Nation-State and War

Lafay, Anne
Classical Tests for Speech and Language Disorders

Lafont, Cristina
Continental Philosophy of Language

Laforce, Robert
Language Processes in Delirium and Dementia

Lagesen, Vivian A.
Gender and Technology: From Exclusion to Inclusion?

Lago, Enrico Dal
Slavery and Violence

Lai, Jessica C.
Intellectual Property: Law in Context

Laird, Nan M.
Binary Response Models and Logistic Regression
Linear Mixed Models

Lake, David A.
Imperialism

Lally, Stephen J.
Projective Methods in Psychology

Lam, Jack
Retirement and Encore Adulthood: The New Later Life Course

Lamade, Raina V.
Repetitive Sex Offenders, Treatment of

Lamb, Sarah
Generation in Anthropology

Lambek, Michael
Taboo

Lambert, Michael J.
Psychological Treatment, Effectiveness of

Lamberts, Koen
Connectionist Models of Concept Learning

Lambin, Eric F.
Geographic Information Systems and Remote Sensing

Lambon Ralph, Matthew A.
Lexical Processes (Word Knowledge): Psychological, Computational and Neural Aspects

Lambourn, David
Metaphor, Role in Social Thought: History of the Concept

Lamm, Bettina
Developmental Research across Cultures and Nations: Challenges, Biases, and Cautions

Lamont, Michèle
Symbolic Boundaries

Lamphere, Louise
Feminism in Anthropology

Lancy, David F.
Cultural Views of Life Phases

Land, Kenneth C.
Forecasting
Social Changes: Models

Lane, Jodi
Fear of Crime

Lane, Jonathan D.
Behavioral Inhibition and Social Withdrawal across Cultures

Lang, Frieder R.
Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model

Lang, Kurt
Reputation

Langewiesche, Dieter
Liberalism: Historical Aspects

Langlois, Simon
Identity Movements
Tradition: Social

Langman, Lauren
Alienation, History of

Lansing, John S.
Irrigation Societies

Lantz, Paula M.
Socioeconomic Status and Health Care

Lao, Oscar
Genetic Admixture

Larkin, Ralph W.
Counterculture: 1960s and Beyond

Larsen, Clark S.
Bioarchaeology

Larsen, Peter
Rhetorical Analysis

Lask, Judith
Family and Psychiatric Illness: Family Therapy

Latré, Stijn
Secularization: History of the Concept

Latsis, John
Social Ontology

Lattie, Emily G.
Depression, Pessimism, and Health

Lau, Hakwan
Consciousness, Neural Basis of

Lau, Timothy
Systematic Reviewing and Meta-Analysis

Lauber, Volkmar
Political Economy of Renewable Energy

Laumann, Edward O.
Heterosexuality

Laurie, Nina
Development Geographies

Laurito, Maria M.
Demographic Impact of Disasters

Laursen, Brett
Conflict and Socioemotional Development

Lauterbach, Dean
Exposure Therapies and Stress Inoculation: A Brief Overview

Lawless, Jerald F.
Event History Analysis

Lawrence, Roderick J.
Urban Life and Health

Lazarus, Arnold A.
Multimodal Therapy in Clinical Psychology

Le Galès, Patrick
Urban Policy in Europe

Le Heron, Richard
Geographies of Food: Politics, Place, Power Reprised

Lea, Stephen E.G.
Decision and Choice: Economic Psychology

Leach, Colin W.
Emotions and Intergroup Relations

Leach, Darcy K.
Oligarchy, Iron Law of

Leachman, Gwendolyn M.
Law as an Instrument of Social Change

Leake, David
Problem Solving and Reasoning: Case-Based

Lebas, Michel J.
Budgeting and Anticipatory Management

Lebo, Matthew J.
Pooled Cross-Sectional and Time Series Analyses, in Political Science

Lechner, Frank J.
Globalization and World Culture

Lederman, Rena
Big Man, Anthropology of

Ledgerwood, David M.
Pathological Gambling and Gambling Disorder

LeDuc, Lawrence
Voting and Direct Democracy

Lee, Aleksandra Sznajder
Post-Socialism: Political Economy

Lee, Benjamin
Area and International Studies: Cultural Studies

Lee, Chin-Hui
Speech Recognition and Production by Machines

Lee, Kyu ho
Adult Education and Lifelong Learning: The US Experience and Beyond

Lee, Michael D.
Mathematical Psychology

Lee, Monica
Social Structure

Lee, Raymond M.
Interviewing, History of
Unobtrusive Methods, History of

Lee, Raymond T.
Wellbeing and Burnout in the Workplace, Psychology of

Lee, Robyn
Reproductive Rights in Affluent Nations

Lee, Sandra Soo-Jin
Direct-to-Consumer Personal Genetic Testing

Lee, Ting Ting
AIDS: Acquired Immune-Deficiency Syndrome

Lee, Wendy L.
Feminist Theory: Radical Lesbian

Lee-Chuvala, Christa
Industrial Geography

Leersnyder, Jozefien De
Culture and Emotion

Lees, Loretta
Gentrification

Leff, Julian
Deinstitutionalisation in Low and Middle Income Countries
Deinstitutionalization, Psychiatry of

Légaré, Jacques
Population Aging: Economic and Social Consequences

Legerstee, Maria
Goal-Oriented Behaviour Among Human and Non-Human Primates

Leggewie, Claus
Gorz, André (1923–2007)
Transnational Citizenship

Leggio, Maria
Cerebellum: Cognitive Functions

Lehmann, Hartmut
Labor Market Institutions and Labor Market Outcomes in Transition Economies

Lehoczky, John P.
Distributions, Statistical: Approximations
Distributions, Statistical: Special and Continuous

Lei, Zhen
Face Recognition Models: Computational Approaches

Leiber, Michael J.
Race: Prespectives and Crime

Leidy, Melinda S.
Socialization in Infancy and Childhood

Leigh Star, Susan
Science and Technology, Social Study of: Computers and Information Technology

Leinaweaver, Jessaca B.
Adoption, Demography of

Leist, Anja K.
Social Epidemiology

Leiter, Brian
Llewellyn, Karl Nickerson (1893–1962)

Leith, Jim
Revolutions, History of

Leixnering, Stephan
Public Sector Organizations

Lelchook, Ariel
Cross-Cultural Training in Industrial and Organizational Psychology

Lemaire, François
Therapeutic Effectiveness Assessment: Approaches in Contemporary Research

Lemańczyk, Szczepan
Nanotechnology: Social and Cultural Aspects

Lena, Jennifer C.
Culture, Production of: Prospects for the Twenty-First Century

Lenart, Adam
Mortality of the Oldest-Old

Lengnick-Hall, Cynthia A.
Tacit Knowledge, Psychology of

Lengyel, György
Polanyi, Karl (1886–1964)

Lens, Willy
Future Time Perspective and Motivation

Leonard, Jonathan S.
Affirmative Action: Empirical Work on Its Effectiveness

Lepawsky, Josh
Electronic Waste

Lerch, Jennifer
Community Supervision and Diversion in the United States

Lerner, Richard M.
Childhood and Adolescence: Developmental Assets
Human Development, Theories of
Self-Regulation During Adolescence: Variations Associated with Individual ↔ Context Relations

Leschinsky, Achim
School (Alternative Models): Ideas and Institutions

Lester, Rebecca J.
Eating Disorders

Lesthaeghe, Ron
Quetelet, Adolphe (1796–1874)

LeTendre, Gerald K.
Educational Systems: North America

Letourneau, Nicole
Socioeconomic Status and Social-Emotional Development of Children from Birth to 36 Months of Age: A Systematic Narrative Review

Leung, Angel Nga-Man
Sex Role Development and Education

Leung, Terry Tse Fong
Harmony and Social Work Practice

Levan, Kristine
Prevention of Crime and Delinquency

Leventhal, Elaine A.
Illness Behavior and Care-Seeking

Leventhal, Howard
Illness Behavior and Care-Seeking

Levert-Levitt, Ella
Integrated Attachment Theory

Levett, Lora M.
Law and Society

Levi, Margaret
Trust, Sociology of

Levin, Jack
Hate Crimes

Levine, Felice J.
Educational Research, The Field of
Social Science Professions and Professionalization

Levine, Kay
Prosecution

LeVine, R.A.
Ethnocentrism

Levinson, Bradley A.
Anthropological Approaches to the Study of Education: The United States and Beyond

Levitis, Daniel A.
Evolutionary Demography: A Synthesis of Two Population Sciences

Leviton, Laura C.
External Validity

Levitt, Heidi M.
Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures
Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures

Levitt, Laura S.
Judaism and Gender

Levy, Donald
Unconscious: History of the Concept

Levy, Marin K.
Appeals: Legal

Lewenstein, Bruce V.
Science and Media

Lewis, Andrew R.
Religion and Politics in the United States

Lewis, Charles
Hypothesis Tests, Multiplicity of

Lewis, Elizabeth
Affect and Emotion, Anthropology of

Lewis, Nick
Geographies of Food: Politics, Place, Power Reprised

Lewis, Toby
Statistical Analysis, Special Problems of: Outliers

Lewontin, Richard C.
Genotype and Phenotype

Leydesdorff, Loet
Scientometrics

Leyendecker, Birgit
Child Care and Development across Cultures

Leyens, Jacques-Philippe
Prejudice in Society: Psychological Perspectives

Lhachimi, Stefan K.
Health Care Markets: Theory and Empirical Evidence

Li, Hong
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Li, Shu-Chen
Aging Mind: Facets and Levels of Analysis
Intelligence: Central Conceptions and Psychometric Models

Li, Stan Z.
Face Recognition Models: Computational Approaches

Li, Wai Keung
Time Series: Advanced Methods

Liang, Yuhua (Jake)
Computer Mediated Communication

Liao, Tim F.
Causal Analysis, History of

Liao, Wei-wen Roger
Grammatical Relations

Licata, Laurent
Collective Memory, Social Psychology of
Xenophobia: Social Psychological Aspects

Lichtenberger, Elisabeth
Cities: Internal Structure

Lichter, Daniel T.
Rural Sociology

Lieb, Roselind
Epidemiology, Concepts and Methods of

Lieberman, Jethro K.
Legalization: The Expansion of Law's Domain

Liebkind, Karmela
Ethnic Identity, Psychology of

Liesenfeld, Roman
Monte Carlo Methods and Bayesian Computation: Importance Sampling

Lietz, Petra
Comparative Research in Education: IEA Studies

Lievens, Filip
Situational Judgment Test

Liffman, Paul
Americas, Sociocultural Overviews: Mexico

Lifset, Reid
Industrial Ecology

Lightfoot, Elizabeth
Disabilism

Liguori, Gary
Physical Activity and Health Promotion

Lilley, Ian
Australia and New Guinea, Archaeology of

Lim, Yen Ying
Human Language: Evolutionary Precursors

Lin, Jingxia
Adverbial Clauses

Lin, Wen
Citizen Mapping

Lindbeck, Assar
Unemployment: Structural

Lindenbaum, Shirley
Cannibalism

Lindenberg, Siegwart
Cooperation: Sociological Aspects
Groups, Sociology of

Lindenberger, Ulman
Baltes, Paul B (1939–2006)

Linder, Stefan
Agency Theory

Lindholm, Charles
Islam: Middle East

Lindley, Dennis V.
Hierarchical Models: Random and Fixed Effects
Savage, Leonard J (1917–71)

Lindley, D.V.
Linear Hypothesis: Fallacies and Interpretive Problems (Simpson's Paradox)

Lindquist, Johan
Brokers and Brokerage, Anthropology of

Lindquist, Martin A.
Functional Magnetic Resonance Imagery, Analysis of

Lindsay, Geoff
Educational Evaluation: Overview

Lindsey, Robin
Transportation, Supply, and Congestion

Lindstrom, Lamont
Cargo Cults
Secrecy, Anthropology of

Linebarger, Deborah L.
Educational Media: Potentials for Learning

Lineham, Peter J.
Peace-Making in History

Ling, Rich
Mobile Communication

Linhares, Maria B.M.
Motor Skills: Development in Infancy and Early Childhood

Link, Rosemary J.
International Social Work Student Exchanges

Link, Stephen W.
Psychophysical Theory and Laws, History of

Linke, Uli
Collective Memory, Anthropology of

Linz, Juan J.
Presidential Government

Lippman, S.S.
Information, Economics of

Lipsitt, Lewis P.
Prenatal and Infant Development: Overview, Current Trends, Future Directions

Lipton, Peter
Empiricism, History of

Liseo, Brunero
Likelihoods that Eliminate Nuisance Parameters

Lister, Pam G.
Child Protection

Liszkowski, Ulf
Social Referencing during Infancy and Early Childhood across Cultures

Littek, Wolfgang
Labor, Division of

Littig, Beate
Sustainable Practices

Little, Brian R.
Personal Projects

Little, Jo
Rural Geography

Little, Roderick J.
Missing Data

Little, Steven G.
Stress in Childhood: Effects on Development

Littman, Michael L.
Markov Decision Processes

Liu, Irene
Human Spatial Orientation, Neural Basis of

Liu, Xing (Michelle)
Beauty and the Labor Market

Liu, YanJiao
Sleep Disorders across the Life Span

Lively, Kathryn J.
Emotions, Sociology of

Livengood, James
Transcranial Direct Current Stimulation and Language

Lizardo, Omar
Culture, Cognition and Embodiment

Lloyd, Elisabeth
Evolutionary Theory, Structure of

Lo, Melody
Political Science and Empirical Implications of Theoretical Models

LoboPrabhu, Sheila M.
Religion: Psychiatric Aspects

Lockhart, Nicole C.
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

Loe, Irene M.
Attention-Deficit and Hyperactivity Disorders, Neural Basis of

Loewenthal, Kate M.
Psychiatry and Religion

Lofgren, Katherine T.
Adult Mortality in Developing Countries

Löfgren, Orvar
Everyday Life, Anthropology of

Loftus, Elizabeth F.
False Memories

Logan, Janette
Adoption: Domestic, International and Global Perspectives

Logie, Robert H.
Working Memory, Psychology of

Logue, Danielle
Innovation

Lohr, Sharon L.
Sample Surveys: Model-Based Approaches

Lohse, Simon
Kuhn, Thomas S. (1922–96)

Lomax, James W.
Religion: Psychiatric Aspects

Long, Nicholas J.
Sociality In Anthropology

Lonka, Kirsti
Sociodigital Revolution: Digital Natives vs Digital Immigrants

Lord, Alex
Regional Planning

Lord, Janet E.
Law and People with Disabilities

Lorenz, Chris
History: Representation Within and Functions of History: Theories and Methods

Lorenzi-Cioldi, Fabio
Sexism

Lorion, Raymond P.
Violence as a Public Health Problem

Losango, Jean D.
Social Science Infrastructure: Africa

Lösel, Friedrich
Rehabilitation of the Offender

Lotz-Heumann, Ute
Reformation and Confessionalization

Louis, Barbara
Scholars, Private Lives of

Louviere, Jordan J.
Measurement Theory: Conjoint Analysis Applications

Louw, Dap
Forensic Psychology

Love, Serena
Geoarchaeology

Lovenduski, Joni
Feminist Political Theory and Political Science

Lovett, Frank
Freedom, History of

Lovie, Pat
Learning: Mathematical Learning Theory, History of

Lovie, Sandy
Learning: Mathematical Learning Theory, History of

Loyal, Steven
Giddens, Anthony (1938–)

Lozada, Fantasy T.
Early Emotional Development and Cultural Variability

Lu, Lauren Xiaoyuan
Supply Chain Management

Luban, David
Justice and Law

Lubik, Sarah
Organizational Emergence and Firm Formation

Lucey, Kelly M.
Subculture, Sociology of

Lucion, Aldo B.
Neonatal Bonding

Lucy, John A.
Sapir-Whorf Hypothesis

Ludwig-Mayerhofer, Wolfgang
Elias, Norbert (1897–1990)

Luetge, Christoph
Business Ethics

Luey, Beth
Publishing as Medium

Luhrmann, Tanya M.
Thick Description: Methodology

Luhrmann, T.M.
Identity in Anthropology

Luig, Ute
Spirit Possession, Anthropology of

Luijkx, Ruud
Data Bases and Statistical Systems: International Comparative Research

Lukas, Josef
Visual Space, Geometry of

Lukes, Steven
Durkheim, Emile (1858–1917)

Lule, Samuel
Social Science Infrastructure: Africa

Lundberg, Kirsten
Case Method and Case Study Collections

Lunn, Brian
Education in Psychiatry

Lunt, I.
Clinical Psychology in Europe, History of

Luong, Hy Van
Vietnamese Revolution, The

Lupia, Arthur
Delegation of Power: Agency Theory

Luppino, Giuseppe
Premotor Cortex

Lützeler, Paul M.
History and Literature

Luuk, Erkki
Syntax–Semantics Interface

Luvender, Kelsey L.
Gender Development: Cultural Differences

Luy, Marc
Adult Mortality in Industrialized Societies

Luyten, Hans
Schooling: Total Impact of

Luzzati, Tommaso
Kuznets Curves

Luzzatti, Claudio
Agraphia (Acquired Dysgraphia)

Lybeck, Eric R.
Geist (Spirit): History of the Concept

Lymbery, Mark
Social Work in Organizations

Lynch, Michael J.
Social Justice and Criminal Justice

Lynch, Michael
Science and Technology Studies, Ethnomethodology of
Science and Technology Studies: Experts and Expertise
Visualization in Science and Technology

Lynch, Scott M.
Bayesian Theory, History, Applications, and Contemporary Directions

Lyon, Eva S.
Myrdal, Alva (1902–86)

Lyons, Ian M.
Numbers and Number Sense

Lyons, Karen
Globalization: Welfare and Social Work

Lyubomirsky, Sonja
Subjective Wellbeing, Psychology of

Lyubovnikova, Joanne
Organizational Climate

Lyytinen, Heikki
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

M

Ma Mung, Emmanuel
International Migration by Ethnic Chinese

Maass, Anne
Embodied Social Cognition

Mabragaña, Cecilia
Social Science Infrastructure: Central and South America

MacCann, Carolyn
Extraversion

MacCoun, Robert J.
Legal Issues: Public Opinion

MacDonald, David B.
Electroencephalography: Basic Principles and Applications

Macdonald, Graham Frank
Functional Explanation: Philosophical Aspects

Machalek, Richard
Sociobiology and Sociology: A New Synthesis

Machluf, Karin
Evolutionary Developmental Psychology

Mackay, Anson W.
Anthropocene Epoch

Mackey, David A.
Prevention of Crime and Delinquency

MacKinnon, Danny
Resilient City

MacKinnon, David P.
Mediating Variable

MacLennan, Bruce
Cognitive Modeling: Connectionist Approaches

Macoir, Joël
Classical Tests for Speech and Language Disorders
Language Processes in Delirium and Dementia

MacPherson, Sarah E.
Source Monitoring

MacWhinney, Brian
Language Acquisition
Psycholinguistics: Overview

Macy, Michael W.
Social Simulation: Computational Models

Maczak, Antoni
Aristocracy, Nobility, and Gentry, European History of

Madanipour, Ali
Urban Design and Public Space

Madans, Jennifer H.
Health Surveys

Magen, Stefan
Philosophy of Law

Magnussen, Svein
Visual Memory, Psychology of

Maharjan, Sujen M.
Child Soldiers

Maher, Liam P.
Power, Politics, and Influence in Organizations

Mahmoud, Rand
Endocrinology and Psychiatry

Mähönen, Tuuli A
Ethnic Identity, Psychology of

Mahoney, Annette
Religion, Family, and Kinship

Mahoney, Elaine J.
Recognition

Mai, Fan
Audiences, Media

Mair, Peter
Party Systems

Maisel, L. Sandy
Campaigning: Political

Maki, Alexander
Volunteerism, Psychology of

Malamud, Deborah C.
Discrimination and the Law

Malamuth, Neil M.
Pornography

Malan-Van Rooyen, Marlize
Emotional Development, Effects of Parenting and Family Structure on

Malatesha Joshi, R.
Education in South Asia

Malave, Jenny C.
Adolescent Sexual Risk

Malkiel, Burton G.
Stock Market Predictability

Malle, B.F.
Attributional Processes: Psychological

Mallon, Mark
Personality and Values at Work

Mallot, Hanspeter A.
Virtual Reality and Spatial Cognition

Mander, Helen
Cognitive Behavioral Therapy

Mandl, Heinz
Apprenticeships and School Learning: General Considerations

Manion, Heather K.
International Social Problems: A Systems Perspective

Mankekar, Purnima
Popular Culture

Mann, Thomas E.
Political Money and Party Finance

Männel, Claudia
Biological Preconditions for Language Development

Manning, Peter K.
Police: A Sociology of Knowledge Approach

Manski, Charles F.
Questionnaires: Collection of Subjective Expectations Data

Manson, Steven M.
Geographic Information Systems and Remote Sensing

Manthorpe, Jill
Self-Directed Care and Individual Budgets for Older People: Lessons from the UK

Manto, Mario
Associative Learning in the Cerebellum

Mantzavinos, C.
Scientific Explanation

Manzo, Gianluca
Macrosociology-Microsociology

Marasi, Shelly
Workplace Deviance

Marcia, James E.
Erikson, Erik Homburger (1902–94)

Marcuello-Servós, Chaime
Castells, Manuel (1942–)

Marcus, George E.
Reflexivity in Anthropology

Mares, Marie-Louise
Media Effects on Children

Margiotti, Margherita
Kinship in Anthropology

Margrett, Jennifer A.
Adult Education and Lifelong Learning: The US Experience and Beyond

Mariën, Peter
Associative Learning in the Cerebellum

Maringe, Felix
Higher Education Market

Mark, Gregory A.
Legal Personality

Mark, Melvin M.
Nonequivalent Group Designs

Markham, Annette N.
Internet Research: Ethical Concerns

Markides, Kyriakos
Migration and Health

Markoff, John
Archival Methods
Revolutions, Sociology of

Marková, Ivana
Representations, Social Psychology of

Markowitsch, Hans J.
Amnesia: Psychogenic
Learning and Memory, Neural Basis of

Marks, Gary N.
Education and Income Distribution

Marks, Isaac
Behavior Therapy: Psychiatric Aspects

Marks, Jonathan
Genetics and Anthropology

Markschies, Christoph
Historical Thought and Historiography: Christian Tradition

Markstrom, Carol A.
Transition to Adulthood, Cultural Differences in

Markwell, Kevin
Gay and Lesbian Tourism

Marley, A.A.J.
Best Worst Scaling: Theory and Practice

Marley, Anthony A.J.
Decision and Choice: Random Utility Models of Choice and Response Time

Marquart-Pyatt, Sandra T.
Political Tolerance

Marques, Clarissa C.
Managed Care and Psychotherapy in the United States

Marques, Sara S.
Sexuality Education

Marsden, Peter V.
Network Centrality, Measures of

Marsh, Heidi L.
Goal-Oriented Behaviour Among Human and Non-Human Primates

Marsh, Herb W.
Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion

Marsh, Herbert W.
Academic Self-Concept and Achievement
Internal/External Frame of Reference Model
Self-Concept: From Unidimensional to Multidimensional and Beyond

Marshall, Anna-Maria
Sexual Harassment: United States and Beyond

Marshall, Edmund I.
Business and Society: Social Accounting

Marshall, Patricia
Fieldwork: Ethical Aspects

Marsolek, Chad J.
Implicit Memory

Martignon, Laura
Algorithm
Information Theory

Martin, Jeffrey
Police: A Sociology of Knowledge Approach

Martin, John L.
Social Structure

Martin, Karin D.
Legal Issues: Public Opinion

Martín, Mayte
Caro Baroja, Julio (1914–95)

Martin, Michael W.
Sociobiology and Sociology: A New Synthesis

Martin, Robert D.
Primates, Evolution of

Martin, Tiphaine C.
Twin Studies and Epigenetics

Martinelli, Alberto
Management: General

Martinez-Alier, Joan
Ecological Economics

Martocchio, Joseph J.
Pay, Compensation, and Performance, Psychology of

Marty, Martin E.
Religious Fundamentalism: Cultural Concerns

Marvel, Stu
Sexual Orientation and the Law

Marx, Gary T.
Surveillance Studies
Technology and Social Control

Mascolo, Michael F.
Neo-Piagetian Theories of Cognitive Development

Maslach, Christina
Burnout, Psychology of

Mason, Owen K.
Arctic Archaeology and Prehistory

Mason, William M.
Statistical Analysis: Multilevel Methods

Massaquoi, Notisha
Queer Theory and Intersectionality

Massaro, Dominic W.
Speech Perception

Massey, Adam J.
Psychoneuroimmunology

Massey, Douglas S.
Inequality, Social
Migration: Motivations
Migration, Theory of
Migration to and from Latin America
Segregation Indices
Urban Violence

Mast, Jason L.
Alexander, Jeffrey C. (1947–)

Masters, Ryan K.
Period and Cohort Analysis in Demography

Mastikhina, Liza
Concepts and Methods of Cognitive Therapies

Mata, José L.
Stress and Cardiac Response

Matebeni, Zethu
Sexual Minorities in South Africa

Matei, Mihaela
Therapeutic Effectiveness Assessment: Approaches in Contemporary Research

Matelski, Monique H.
Group Processes, Social Psychology of

Mather, Charles
Consumption, Geography of

Mather, Lynn
Disputes, Social Construction and Transformation of

Mathew, Sarah
Human Cooperation, Evolution of

Mathew, Sera A.
Transition to Adulthood, Cultural Differences in

Mathews, John A.
Greening of Business

Matić, Dejan
Information Structure in Linguistics

Matsuda, Michiyuki
Ethnogenesis in Anthropology

Matt, Georg E.
Generalizability Theory

Matthews, Gerald
Coping, Personality and
Personality, Cognitive Models of

Matthews, Katey
Retirement and Health

Matusov, Eugene
Vygotsky's Theory of Human Development and New Approaches to Education

Mauldon, Jane
Sexuality Education

Maume, David J.
Female Power at Work

Maurer, Bill
Money: Anthropological Aspects

Mavor, Kenneth I.
Social Categorization

Mavromaras, Kostas
Mismatch on the Labor Market

Maxwell, Joseph A.
Qualitative Research

May, J.
Human–Computer Interfaces

Mayda, Anna M.
International Migration

Mayer, Enrique
Peasants in Anthropology

Mayer, John D.
Personal Intelligence and Competencies

Mayer, Jonathan D.
Medical Geography

Mayer, Nonna
Voting, Sociology of

Mayes, Andrew
Hippocampal Amnesia

Maynard, Rebecca A.
Social Experiments

Mazelis, Joan M.
Welfare

Mazzoni, Giuliana
Eyewitness Testimony

McAdams, Richard H.
Conventions and Norms: Philosophical Aspects

McAlpine, Lynn
Doctoral Education

McArdle, John J.
Genetic Studies of Behavior: Methodology
Growth Curve Analysis

McBride-Chang, Catherine
Chinese Literacy Acquisition: A Multidimensional Puzzle

McCabe, Marita P.
Body Image

McCall, John J.
Information, Economics of

McCall, Patricia L.
Crime Trends and Debates

McCann, Michael
Rights: Legal Aspects

McCleary, Richard
Interrupted Time Series Quasi-Experimental Designs

McClelland, James L.
Cognitive Neuroscience

McClintock, Martha K.
Sexual Motivation

McCombes-Tolis, Jule
Special Education

McCombs, Maxwell
Agenda Setting, Media Effects on

McConnell, William J.
Land Change: The Merger of Land Cover and Land use Dynamics

McCormack, Mark
Gays, Lesbians, and Sports: No Longer on the Sideline

McCoy, Candace
Crime Policy in the United States

McCracken, Sheelagh
Security: Commercial Transactions

McCrimmon, Adam W.
Intelligence: Assessments of

McCrudden, Christopher
Affirmative Action: Comparative Policies and Controversies

McCullagh, C. Behan
Historical Explanation, Theories of: Philosophical Aspects

McCulloch, Charles E.
Generalized Linear Mixed Models

McDade, Thomas
Biomarkers in Demographic Research

McDaniel, Mark A.
Prospective Memory

McDevitt, Jack
Hate Crimes

McDonald, Jacquelin
Communities of Practice

McDonald, W. Ian
Jackson, John Hughlings (1835–1911)

McDonnell, Brett H.
Corporate Governance

McDowall, David
Quantitative Methods and Crime

McDowell, Andrew
Belief, Anthropology of

McEwen, Jean E.
Ethical, Legal, and Social Implications Program at the National Human Genome Research Institute

McFadden, Daniel
Observational Studies: Outcome-Based Sampling

McFadden, Hannah M.
Values, Anthropology of

McFarland, Andrew S.
Power

McGarty, Craig
Social Categorization

McGranahan, Gordon
Urbanization

McGrath, Patrick J.
Behavioral Problems, Effects of Socio-Economic Status on

McGuire, Martin C.
Olson, Mancur (1932–98)

McHugh, Fergal
Rorty, Richard (1931–2007)

McIver, Derrick
Tacit Knowledge, Psychology of

McKinnon, Scott
Queer Geography

McLachlan, Geoffrey J.
Computation: Expectation-Maximization Algorithm
Mixture Models in Statistics

McLachlan, Geoffrey
Multivariate Analysis: Classification and Discrimination

McLafferty, Sara
Spatial Mismatch

McLaughlin, Brian P.
Physicalism and Alternatives

McLaughlin, Neil
Fromm, Erich (1900–80)

McLean, Stuart
Materiality and Culture

McLure, Michael
Value: History of the Concept

McMillan, James H.
Classroom Assessment

McMillan, Thomas M.
Post Traumatic Amnesia

McNeil, Daniel W.
Motivation in Clinical Interventions

McNeill, David
Gesture in Linguistics

McNicoll, Geoffrey
IPAT (Impact, Population, Affluence, and Technology)

McWilliams, Spencer A.
Psychology, History of (Twentieth Century)

Meadors, Joshua D.
Doctor–Patient Interaction in the West: Psychosocial Aspects

Measham, Fiona
Alcohol Use and Abuse

Mechanic, David
Health Care, Rationing of
Medical Sociology

Meece, Judith L.
Gender and Academic Motivation

Mehta, Neil K.
Obesity and Mortality

Mehta, Paras D.
Control Variables in Research

Mehu, Marc
Emotions: Methods of Assessment

Meilinger, Tobias
Virtual Reality and Spatial Cognition

Meja, Volker
Knowledge, Sociology of
Mannheim, Karl (1893–1947)

Mel, Bartlett W.
Neurons and Dendrites, Integration of Information in

Mel'čuk, Igor
Dependency in Language

Melville, Joshua
Social Media

Melvin, Jennifer E.
Immigration: Health and Mortality

Mendel, Jonathan
Urban Violence and Militarization

Méndez, Mónica
Globalization and Human Trafficking

Menell, Peter S.
Property: Legal Aspects

Menger, Pierre-Michel
Artists, Competition and Markets

Menkel-Meadow, Carrie J.
Mediation, Arbitration, and Alternative Dispute Resolution (ADR)

Mennell, Stephen
Burke, Peter (1937–)
National Character, History of

Menter, Ian
Teacher Education

Mentovich, Avital
Compliance and Legal Authority

Menzel, Max-Peter
Industrial Cluster Dynamics

Menzel, Randolf
Memory in the Bee

Mercy, Aurélie
Collective Memory, Social Psychology of

Mergler, Iga
Fromm, Erich (1900–80)

Merkur, Dan
Mysticism

Merkx, Gilbert W.
Area and International Studies: Stakeholders

Merlo, Kelsey L.
Job Satisfaction

Merriam, Jesse
Judicial Review

Merrilees, Christine E.
War, Political Violence, and Effects on Children

Merry, Sally E.
Law: Anthropological Aspects

Mesch, Gustavo S.
Community and Media
New Media and Social Capital

Meslé, France
Adult Mortality in Eastern Europe and the Former Soviet Republics

Mesoudi, Alex
Cultural Evolution: Overview

Mesquita, Batja
Culture and Emotion

Metz, Allison
Implementation Science

Meusburger, Peter
Education, Geography of
Knowledge, Geography of

Mey, Günter
Qualitative Methodology in Developmental Psychology

Meyen, Michael
Communication Research and Media Studies, History of

Meyer, Klaus E.
Foreign Investment: Direct

Meyer, Marleen D.
Ancient Egypt, Archaeology of

Meyer, Marshall
Bureaucracy and Bureaucratization

Meyer, Morgan
Epistemic Communities and Collaborative Research

Meyer, Renate E.
Public Sector Organizations

Meyer, Thorsten
Health Research, Qualitative

Meyerbröker, Katharina
Panic Disorder and Agoraphobia Across the Lifespan

Meyer-Parlapanis, Danie
Torture and its Consequences, Psychology of

Miao, Ruiqin
Second Language Learning

Mica, Adriana
Unintended Consequences: History of the Concept

Michaels, Stuart
Methodological Issues in the Study of Sexuality

Michell, Joel
Measurement Theory: History and Philosophy

Middleton, James P.
Dark Triad: The "Dark Side" of Human Personality

Middleton, John
Ancestors, Anthropology of

Miethe, Terance D.
Homicide

Miethe, Terrance
Data Bases and Statistical Systems: Crime

Mike, Anissa
Conscientiousness

Mikkelsen, Elisabeth Naima
Conflict: Organizational

Mikocka-Walus, Antonina
Gastrointestinal Diseases: Psycho-social Aspects

Mikula, Gerold
Justice: Social Psychological Perspectives

Milas, Goran
Personality Changes During Adolescence Across Cultures

Miller, Byron
Social Movements, Geography of

Miller, Grant
Family Planning: Program Effects

Miller, Harvey J.
Location, Absolute and Relative

Miller, Joan G.
Culture and the Self: Implications for Psychological Theory

Miller, Shae D.
Lesbian, Gay, Bisexual and Trans-sexual Drag Culture

Miller, Tricia A.
Patient Adherence to Health Care Regimens

Millová, Katarína
Personality Development: Systems Theories

Milman, Natalie B.
Distance Education

Milner, Peter M.
Hebb, Donald Olding (1904–1985)

Milojević, Staša
Scientometrics

Miners, Christopher TH.
Emotional Intelligence and Competencies

Minnite, Lorraine C.
Underclass
Welfare

Mirowsky, John
Alienation: Psychosociological Tradition

Mishra, Gita D.
Life Course Epidemiology

Mitchell, Diana E.
Vestibulo-Ocular Reflex, Adaptation of the

Mitchelstein, Eugenia
New Media, News Production and Consumption

Miyachi, Masato
Meiji Revolution, The

Miyaoka, Atsushi
Social Network Analysis

Miyata, Hiromitsu
Problem Solving during Infancy and Early Childhood, Development of

Mlačić, Boris
Big Five Factor Model, Theory and Structure
Personality Changes During Adolescence Across Cultures

[Modell] Schachter, Judith
Benedict, Ruth (1887–1948)

Modood, Tariq
Integration and Multiculturalism: Focus on Western Europe

Mody, Perveez
Marriage

Moeller, Aleidine J.
Foreign Language Teaching and Learning

Moeller, Julia
Passion and Motivation

Moen, Phyllis
Retirement and Encore Adulthood: The New Later Life Course

Moeran, Brian
Advertising Agencies

Moffitt, Robert A.
Welfare and Work

Mohr, John W.
Formal Methods of Cultural Analysis

Molenaar, Dylan
Psychometrics

Molina, Agustin
Organizational Justice

Molinari, Marco
Cerebellum: Cognitive Functions

Molitor, Anne M.
Epigenetics

Moll, Jorge
Neural Foundation of Morality

Möller, Hans-Jurgen
Psychiatric Assessment: Negative Symptoms

Möller, Jens
Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion
Internal/External Frame of Reference Model

Mols, Yvana L.B.H.
Domestic Violence: Legal Perspectives

Molteni, John D.
Special Education

Mommsen, Hans
National Socialism and Fascism

Momo, Oliver Helena Fiona Te
Māori Social Work

Monahan, John
Legal Process and Social Science: United States

Monfils, Marie H.
Anxiety and Fear Conditioning, Neural Basis of Limbic System

Monk, Janice
Gender and Feminist Studies in Geography

Monroe, Scott M.
Stress: Psychological Perspectives

Montada, Leo
Justice, Conflicts, and the Justice of Conflict Resolution

Montangero, Jacques
Piaget, Jean (1896–1980)

Monteiro, Anjali
Documentary and Ethnographic Film

Montello, Daniel R.
Scale in Geography
Spatial Cognition

Moody, James
Clustering and Cohesion in Networks: Concepts and Measures

Moody-Adams, Michele M.
Diversity and Disagreement in Ethics: Philosophical Implications

Mooney, Timothy D.
Dilthey, Wilhelm (1833–1911)

Moore, John H.
Genetics and Anthropology

Moore, Ryan
Youth Culture, Sociology of

Moorer, Regina
Gender and Women's Studies, Applied Research on

Moos van Wyk, Micheal
Teaching Economics

Moosbrugger, Robert
Schools, Micropolitics of

Morales, Juan J.
Counter-Insurgencies and the Social Sciences

Morasso, Pietro G.
Classical Mechanics and Motor Control
Motor Control Models: Learning and Performance

Moreno Fraginals, Manuel R.
Cuban Revolution, The

Morgan, Bronwen
Regulation and Administration

Morgan, Geoffrey P.
Computational Organizational Theory

Morgan, Gwenda
American Revolution, The

Morgan, Paul J.
Culture and Economy

Morgan, William J.
Marxism—Leninism: The Ideology of Twentieth-Century Communism

Morgenstern, Aliyah
Intentionality in Language and Communication, Emergence of

Morgenthaler, Stephan
Robustness

Morimune, Kimio
Simultaneous Equation Estimates (Exact and Approximate), Distribution of

Morley, Christine
Critical Reflexivity and Social Work Practice

Morrill, Calvin
Law, Mobilization of

Morrill, Richard
Spatial Equity

Morrin, Maureen
Consumer Psychology

Morris, Carl N.
Experimental Design: Large-Scale Social Experimentation

Morrison, Kathleen D.
Intensification and Specialization, Archaeology of

Morse, Jaimie
Law, Sociology of

Morse, Stephen J.
Drugs, Decriminalization of: Cultural Concerns

Mortillaro, Marcello
Emotions: Methods of Assessment

Morus, Iwan R.
Science: Constructivist Perspectives, History of

Moschella, Melissa
Natural Law

Moscovitch, Morris
Amnesia: General

Moseley, William G.
Food Security and 'Green Revolution'

Moses, Yolanda T.
Race and Racism in the Twenty-First Century

Mosher, Clayton
Data Bases and Statistical Systems: Crime

Moshman, David
Logical and Hypothetical Reasoning in Adolescence, Development of

Mosquera, Patricia M.R.
Cultural Influences on Interpersonal Relationships

Moultrie, Tom A.
Demographic Techniques: Indirect Estimation

Mowen, Thomas J.
School Safety

Muckenhuber, Johanna
Life-style, History of the Concept

Mudiayi, Dominic
Optogenetics

Mueller, Jutta L.
Biological Preconditions for Language Development

Mueller, Ralph O.
Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis

Mufwene, Salikoko S.
Pidgin and Creole Languages

Mühlberger, Andreas
Virtual Reality in Psychotherapy

Mühlleitner, Elke
Psychoanalysis, History of

Mukherjee, Supriya
Historical Thought and Historiography: Current Trends

Mulcahy, Aogán
Policing in Divided Societies

Mullainathan, Sendhil
Behavioral Economics

Müller, Andreas
Occupational Stress Management

Müller, Anna-Lisa
Social Ecology: The Chicago School

Muller, Claude P.
Epigenetics

Müller, Hans-Peter
Berlin, Isaiah (1909–97)
Dahrendorf, Ralf (1929–2009)

Muller, J.Z.
Conservatism: Historical Aspects

Müller, Martin
Organization, Geography of

Müller, M.J.
Antidepressant Drugs

Müller, Peter
Markov Chain Monte Carlo Methods

Müller, S.
Auctions

Müller, Walter
Mobility, Social

Müller, Wolfgang C.
Executive Government

Müller-Kalthoff, Hanno
Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion

Mullet, Hillary G.
Prospective Memory

Mullola, Sari
Temperament and Motivation

Münch, Richard
Functionalism, History of
Integration, Social

Muncie, John
Juvenile Justice: International Law and Children's Rights

Munck, Gerardo L.
Democratic Transitions

Munford, Robyn
Learning Disability and Social Work

Munger, Michael C.
Public Choice Economics

Mungiu-Pippidi, Alina
Corruption: Political and Public Aspects

Muniesa, Fabian
Actor-Network Theory

Muñoz-Guzmán, Carolina
Liberation Theology in Social Work

Murata, Yujiro
Nationalism, Historical Aspects of: East Asia

Muravyev, Alexander
Labor Market Institutions and Labor Market Outcomes in Transition Economies

Murdock, Bennet
Memory Models: Quantitative

Murdock, Graham
Mass Media, Political Economy of

Murias, Kara
Human Spatial Orientation, Neural Basis of

Murphy, Alexander B.
Political Geography

Murray, Carolyn B.
Academic Performance, Effects of Socio-Economic Status on
Culture as a Determinant of Mental Health
Doctor–Patient Interaction in the West: Psychosocial Aspects
Race: Ethnicity and Health

Muscarella, Frank
Clinical Psychology, Psychiatry and Homosexuality

Mussel, Patrick
Personality Assessment

Mustaine, Elizabeth E.
Social Control and Deviance, History of

Musto, Jennifer
Prostitution and Sex Work

Muthivhi, Azwihangwisi E.
Piaget's Theory of Human Development and Education

Mützel, Sophie
Culture and Actor Network Theory

Mwansa, Lengwe-Katembula J.
Structural Adjustment in Africa: Implications for Social Work

Mychasiuk, Richelle
Epigenetics of Brain Plasticity and Behavior

Myers, Steve
Sex Offenders

Myhill, John
Religion: Nationalism and Identity

Mykitiuk, Roxanne
Reproductive Rights in Affluent Nations

Myrdal, Janken
Peasants and Rural Societies in History (Agricultural History)

N

Nadal, Marcos
Neuroesthetics

Nadelson, Carol Cooperman
Women and Psychiatry

Nadkarni, Vimla V.
International Association of Schools of Social Work

Nadler, Arie
Helping Intergroup Relations

Nagatsu, Michiru
Behavioral Economics, History of

Nagel, Elizabeth
Masculinity, Scripts of

Nagel, Ineke
Art and Socialisation

Nagel, Jack H.
Proportional Representation

Nagengast, Benjamin
Internal/External Frame of Reference Model

Nagin, Daniel S.
Deterrence: Sanction Perceptions
Deterrent Effect of Police and Prisons

Nagy, Nicole
Dimensional Comparisons and Their Consequences for Self-Concept, Motivation, and Emotion

Nakai, Miki
Consumption, Sociology of

Nakamura, Jeanne
Flow in Motivational Psychology

Narita, Hiroki
Movement Theory and Constraints in Syntax

Nas, Peter J.M.
Urban Anthropology

Nash, Jonathan R.
Law and Risk

Nash, Mary
Social Work and Refugees

Nash, Robert A.
Eyewitness Testimony

Nason-Clark, Nancy
Religion and Abuse

Navarro, Jordana N.
Intimate Partner Abuse, Applied Research on

Navarro, Juan J.
Counter-Insurgencies and the Social Sciences

Neal, Andrew
Human Factors and Ergonomics

Neath, Ian
Proactive Interference

Nebel, Bernhard
Logics for Knowledge Representation

Necka, Elizabeth A.
Social Neuroscience of the Twenty-First Century

Nee, Victor
Sociology and the New Institutionalism

Needham, Barrie
Property Rights and Urban Planning

Neiberg, Michael S.
Warfare in History

Neiburg, Federico
National Character

Neidhardt, Friedhelm
Social Science Infrastructure: European Union (Research and Teaching)

Neidhöfer, Thilo
Scholars, Private Lives of

Neimeyer, Robert A.
Constructivism and Constructionism: Methodology
Constructivist Theory and Psychotherapy

Nelson, Arthur C.
Growth Management and Urban Planning in the United States

Nelson, Michael J.
Judicial Behavior
Judicial Independence

Nelson, Moira
Skills Training: Public Policy Aspects

Nerlich, Brigitte
Nanotechnology: Social and Cultural Aspects

Nesselroade, John R.
Cattell, Raymond Bernard (1905–98)

Netter, Petra
Neuroendocrinology

Neugebauer, Witold
Hintze, Otto (1861–1940)

Neuhaus, John M.
Generalized Linear Mixed Models

Neuner, Frank
Mass Trauma: Psychopathological Effects across the Life Span

Nevzorov, Valery B.
Nonparametric Statistics: Records

Newman, Lucile
Birth, Anthropology of

Newman, Peter
Planning Issues and Sustainable Development

Ng, Emil Man-Lun
Heterosexuality in China

Ng, Mee Kam
Urban Policies in Asia

Ng, Roger MK.
Psychology of the Real Self: Psychoanalytic Perspectives

Ng, Weiting
Neuroticism

Ng, Winnie
Trade Unions and Social Work: Lessons from Canada

Ng, Yew-Kwang
Welfare Economics

Ní Bhrolcháin, Máire
Demographic Measurement: General Issues and Measures of Fertility
Demographic Measurement: Nuptiality, Mortality, Migration, and Growth

Nicholls, Alex
Social Innovation

Nichols, Rebecca M.
False Memories

Nickel, Allison E.
Recognition

Nickels, Lyndsey
Cognitive Neuropsychology, Methods of

Nicolaisen, Wilhelm FH.
Onomastics

Nidos, Andreas
Interhemispheric Interaction in Language and Cognitive Processes

Nielsen, Mark
Pretend Play and Cognitive Development

Niewöhner, Jörg
Infrastructures of Society, Anthropology of

Nijkamp, Peter
Location Theory
Smart Cities

Nikku, Bala R.
International Social Work

Nilsson, Lars-Göran
Aging and Memory

Nippel, Wilfried
Ancient Greece and Rome, History of

Historical Thought and Historiography: Classical Period (Especially Greece and Rome)

Nippert, Irmgard
Genetic Counseling: Historical, Ethical, and Practical Aspects

Nishikii, Yumi
Perinatal Psychiatry

Niv, Sharon
Genetics of Personality

Nivette, Amy
International and Comparative Criminology

Noack, Peter
Apprenticeship and School Learning: Lessons from Germany

Noble, Carolyn
Social Protest Movements and Social Work Practice

Noeker, Meinolf
Childhood Cancer: Psychosocial Aspects

Nollenberger, Natalia
Gender, Time-Use, and Fertility Recovery in Industrialized Countries

Nolte, Paul
Modernization and Modernity in History

Nolte, P.
Social Inequality in History (Stratification and Classes)

Noordman, Leo G.M.
Inferences in Discourse, Psychology of

Norcross, John C.
Psychotherapy Integration

Nordstrom, Carolyn
War: Anthropological Aspects, Historical Development of

Norman, Greg J.
Empathy: A Social Neuroscience Perspective

Norris, Pippa
Political Communication

Northcott, Michael
Christianity in Asia

Norton, Jacob
Agent Based Modeling, Statistics of

Norwood, Frances
End of Life Choices

Novaco, Raymond W.
Transportation, Psychology of

Novas, Carlos
Biocitizenship

Nóvoa, António
Education, History of

Nowak, Elke
Linguistics: Incorporation

Nowotny, Helga
Social Science Infrastructure: European Union (Resourcing and Funding)

Nsamenang, A. Bame
Ecocultural Theories of Development

Ntarangwi, Mwenda
Youth, Music, and Peace Building

Nucci, Larry
Moral Development and Education

Nunner-Winkler, Gertrud
Kohlberg, Lawrence (1927–87)
Personal Identity: Philosophical Aspects

Nuttman-Shwartz, Orit
Post-Traumatic Stress in Social Work

Nuwer, Hank
Hazing in Fraternities and Sororities

Nyíri, Pál
Intercultural Communication: An Anthropological Perspective

O

O'Brien, Charles
Film History

O'Brien, Jennifer
Health, Geography of

O'Brien, Jodi
Heterosexism and Homophobia

O'Brien, P.K.
Global History: Universal and World
Industrialization, Typologies and History of

O'Donnell, Angela M.
Group Processes in the Classroom

O'Donohue, William
Behavior Therapy: Background, Basic Principles, and Early History
Behavior Therapy: Evidence-Based Practice, Clinical Applications, Unresolved Issues
Behavior Therapy: The Second and Third Waves

O'Dowd, Liam
Urban Conflicts

O'Flaherty, Eamon
Annales School

O'Grady, William
Language Development: Emergentist Theories

O'Guinn, Thomas C.
Advertising: Effects

O'Kelly, Morton E.
Spatial Interaction

O'Riordan, Kate
Genetics and the Media

Oaxaca, Ronald
Discrimination, Economics of

Oberlin, Kathleen C.
Science, Sociology of

Ocklenburg, Sebastian
Brain Asymmetry

Oelkers, Jürgen
Rousseau, Jean-Jacques (1712–78)

Oettingen, Gabriele
Motivation and Actions, Psychology of
Motivation: History of the Concept

Ogino, Masahiro
Collective Memory, Sociology of

Ogliastri, Enrique
Social Business

Ohnuki-Tierney, Emiko
East Asian Studies: Culture
Japan: Sociocultural Aspects

Okediji, Ruth L.
Law and Technology

Oldenburg, Brian
Public Health as Social Science

Olderbak, Sally
Facial Emotion Expression, Individual Differences in

Olekalns, Mara
Negotiation and Conflict, The Psychology of

Oliveira-Souza, Ricardo de
Neural Foundation of Morality

Oliver-Smith, Anthony
Hazards and Disaster Research in Contemporary Anthropology

Olkin, Ingram
Hotelling, Harold (1895–1973)

Olson, David R.
Writing Systems, Psychology of

Olson, Judy M.
Cartography

Olson, Sheryl L.
Defiant Behavior During Infancy and Early Childhood

Olsson, Christian
Wars among Nation-States: Patterns and Causes

Olsson, Henrik
Decision and Choice: Heuristics

Olwig, Karen Fog
Americas, Sociocultural Overviews: Caribbean

Olwig, Kenneth R.
Landscape

Olzak, Susan
Ethnic, Racial, and Nationalist Movements

Ones, Deniz S.
Pro-Environmental Behavior

Ong, Aihwa
Modernity: Anthropological Aspects

Opp, Karl-Dieter
Norms

Oppenheim, Robert
Asia, Sociocultural Overviews: Korea

Opper, Sonja
Sociology and the New Institutionalism

Organ, Dennis W.
Organizational Citizenship Behavior

Orleans, Myron
Criminal Justice, Sociology of
Phenomenology in Sociology

Orlinsky, David
Psychotherapy Process Research

Orme, Joan
Gendered Care and Social Work Practice

Ormiston, Naadli T.
Indigenous Child Welfare Practice: The Canadian Case

Orr, Larry L.
Social Experiments

Orueta, Fernando D.
Urban Megaprojects

Orum, Anthony M.
Case Study: Logic

Osborne, Dustin
Self-Control Theory and Crime

Oschatz, Kerstin
Epistemological Beliefs and Motivation

Oser, Jennifer
Civil Service

Oshana, Marina A.L.
Responsibility: Philosophical Aspects

Osinsky, Roman
Personality Assessment

Ostermeier, Lars
Societal Impact Assessment

O'Hagan, Timothy
Freedom: Political

Otte, Gunnar
Data Bases and Statistical Systems: Culture

Otte, Marcel
Migrations, Colonizations, and Diasporas in Archaeology

Otto, Hiltrud
Attachment During Infancy and Early Childhood: Cultural Variations, Development of

Outhwaite, R.W.
Positivism, Sociological

Outhwaite, William
Habermas, Jürgen (1929–)
Hermeneutics, History of

Overing, Joanna
Kinship in Anthropology

Overstall, Antony
Population Size Estimation and Capture–Recapture Methods

Owen, Guillermo
Game Theory

Owens, Nicole
Alternative Food Movements

Owens, Richard G.
Men's Health

Owusu, Maxwell
Third World

Oyebode, Femi
Normality and Mental Health
Nosology in Psychiatry

Oyserman, Daphna
Values, Psychology of

Özbay, Cenk
Same-Sex Sexualities in Turkey

Ozer, Daniel J.
Five Factor Model of Personality, Assessment of

Özler, Berk
Conditional Cash Transfers: Influence on Marriage and Fertility

P

Paasche, Till F.
Surveillance and Privacy, Geography of

Paasonen, Susanna
Online Pornography

Pace, Amy
Scientific Concepts: Development in Children

Pachucki, Mark C.
Networks and Cultural Consumption

Pachucki, Mark
Symbolic Boundaries

Pachur, Thorsten
Heuristics, History of

Paddison, Ronan
Urban Studies: Overview

Paelecke, Marko
Personality Assessment

Pagel, Mark
Comparative Method in Evolutionary Studies

Pahlke, Erin
Gender and Education

Paik, Anthony
Sexual Debut

Paik, Susan J.
Educational Productivity

Painter, Genevieve R.
Feminist Legal Theory

Pakulski, Jan
Class: History of the Concept

Palloni, Alberto
Diffusion in Sociological Analysis

Palsson, Gisli
Anthropology, Genomics, and Human Variation: National Roots
Fishing Societies

Paluck, Elizabeth Levy
Field Experiments

Pang, Tikki
Human Genomics in Asia

Panter, A.T.
Personal Intelligence and Competencies

Panther, Klaus-Uwe
Control, Syntactic

Papachristos, Andrew V.
Crime: Organized

Papagno, Costanza
Episodic Memory

Papilloud, Christian
Gift: History of the Concept

Parameshwaran, Soumya
Hysteria: History and Critiques
Psychosomatic Medicine
Somatotherapy, History of

Pardo, Rafael I.
Bankruptcy

Pargament, Kenneth I.
Religion: Psychiatric Aspects

Pariante, Carmine M.
Antidepressant Drugs

Parigi, Paolo
Miracles, Cultural Logic of

Park, Elke
Academic Research and Employment: Recent Changes in Europe and the United States

Park, Heeman
Organizational Commitment

Park, Myoungkyu
Translating Sociology into Japanese, Chinese and Korean Languages

Parke, Ross D.
Socialization in Infancy and Childhood

Parker, Jonathan
International Issues in Child Protection

Parker, Philip
Internal/External Frame of Reference Model
Self-Concept: From Unidimensional to Multidimensional and Beyond

Parkin, David J.
Ritual

Parrado, Emilio A.
Migration and Fertility

Parrigon, Scott
Openness to Experience

Parsons, Donald O.
Job Displacement Insurance

Parsons, Jennifer A.
Longitudinal Research: Panel Retention

Partee, Barbara H.
Montague Grammar

Paster, Mary
Phonology

Patchell, Jerry
Resource Geography

Patel, Naina
Migration, Ethnicity, Aging and Social Work Practice

Patrik, Aspers
Fashion, Sociology of

Patros, Tyson
Political Sociology

Patterson, Orlando
Slavery: Comparative Aspects

Pattison, Philippa
Social Network Models: Statistical
Social Networks, Algebra of

Patton, Paul
Postmodernism: Philosophical Aspects

Paul, Diane B.
Eugenics, History of

Pauli, Christine
Co-constructivism in Educational Theory and Practice

Pauli, Paul
Virtual Reality in Psychotherapy

Paunonen, Sampo V.
Five Factor Model of Personality, Facets of

Pavitt, Keith
Research and Development in Organizations

Pavlopoulos, Vassilis
Social Psychology: Research Methods

Pavlos, Stephanie R.
Personality Theory and Psychopathology

Pavlova, Maria K.
Leisure Activities Choices among Adolescents

Paylor, Ian
Alcohol Use and Abuse

Payne, Brennan R.
Education and Learning: Lifespan Perspectives

Payne, Malcolm
Social Work Education

Peace, Adrian
Australia, Sociocultural Overviews: Australian Settler Society

Peacock, James L.
Values, Anthropology of

Pearce, Lisa D.
Religion and Youth

Pearsall, Deborah M.
Paleoethnobotany

Pearson, Caleb M.
Intracarotid Sodium Amytal Test and Language

Pearson, Emma
Early Education and Care in the World Regions

Pease, Bob
Masculinity and Social Work Practice

Pechar, Hans
Academic Careers in Comparative Perspective
Academic Research and Employment: Recent Changes in Europe and the United States

Pechholdová, Markéta
Infant and Child Mortality in Industrialized Countries

Peck, Jennifer H.
Race: Prespectives and Crime

Pedraza Moreno, Teresita
Cuban Revolution, The

Peebles, Gustav
Exchange in Anthropology

Peet, Richard
Marxist Geography

Peichl, Andreas
Earnings Inequality

Peigneux, Philippe
Sleep and Memory

Peiser, Richard
Real Estate Development

Pekrun, Reinhard
Internal/External Frame of Reference Model
Self-Concepts: Educational Aspects
Test Anxiety and Academic Achievement

Pelegrin, Jacques
Lithics and Archaeology

Pell, Christopher
Communities and Crime

Pellman, Blake A.
Fear: Psychological and Neural Aspects

Pendergrass, Sabrina
Symbolic Boundaries

Penny, Will
Bayesian Models in Neuroscience

Pe-Pua, Rogelia
Indigenous Psychology

Perelli-Harris, Brienna
Partnership Formation and Dissolution in Western Societies

Pereyra, Diego
Cardoso, Henrique (1931–)

Perez, Sofia A.
Immigration Policy

Pérez-Edgar, Koraly
Temperament Development, Theories of

Perkins, David N.
School Learning for Transfer

Perogamvros, Lampros
Dreaming, Neural Basis of

Perren, Rebeca
Collaborative Consumption

Perret-Clermont, Anne-Nelly
Learning and Instruction: Social-Cognitive Perspectives

Perry, David C.
Infrastructure Investment

Peset, José L.
Plagues and Diseases in History

Pestel, Nico
Earnings Inequality

Peter, Burkhard
Hypnosis

Peter Castro, A.
Development: Social-Anthropological Aspects

Petermann, Franz
Childhood Cancer: Psychosocial Aspects

Peters, John D.
Mass Communication: Normative Frameworks

Petersen, Trond
Endogeneity: Methodology

Peterson, Abby
Social Protest

Peterson, Christopher D.
Cognition, Evolution of

Peterson, Donald R.
Training in Clinical Psychology in the United States: Practitioner Model

Peterson, Richard A.
Culture, Production of

Peterson, Roger L.
Training in Clinical Psychology in the United States: Practitioner Model

Petrides, Michael
Working Memory, Neural Basis of

Petrie, Keith J.
Social Support and Recovery from Disease and Medical Procedures
Stress, Coping and Health

Petry, Nancy M.
Pathological Gambling and Gambling Disorder

Petryna, Adriana
Health: Anthropological Aspects

Pettigrew, Thomas F.
Allport, Gordon W (1897–1967)
Discrimination: Racial
Prejudice and Discrimination

Pettit, Philip
Consequentialism (Including Utilitarianism)
Individualism versus Collectivism: Philosophical Aspects

Pfetsch, Barbara
Social Science and the Media

Pfister, Christian
Climate, History of

Phalet, Karen
Ethnicity and Migration in Europe

Phenix, Deinya
Age and Crime

Philip, Kavita S.
Indigenous Knowledge: Science and Technology Studies

Phillips, Alison C.
Motivation in Youth Sport and Physical Activity: Developmental Perspectives

Phillips, David
Exposure Therapies and Stress Inoculation: A Brief Overview

Phillips, Julie A.
Suicide, Sociology of

Phillips, Robert A.
Personalized Medicine

Phillips, Ruth
Feminist Research and Social Work

Phillipson, Shane N.
Educational Learning Theory

Phillipson, Sivanes
Educational Learning Theory

Picard, Rainer
Fourier Analysis

Picavet, Emmanuel
Equality: Philosophical Aspects
Methodological Individualism in Sociology

Pickering, Andrew
Cybernetics

Pickering, Mary
Comte, Auguste (1798–1857)

Pieke, Frank N.
Asia, Sociocultural Aspects: China

Pierre, Gaëlle
Development and Employment: A Mutual Relationship

Pietrini, Pietro
Neuroethics: Ethics of Science and Science of Ethics

Pigott, Therese D.
Systematic Reviewing and Meta-Analysis

Pina-Cabral, João de
Names and Naming

Pinch, Trevor
Scientific Controversies

Pinckney, Jonathan
Insurgency

Pinnow, Eleni
Language Development, Theories of

Pinto, Arthur R.
Corporate Law: Publicly Traded Corporations

Pinto-Coelho, Joanna Marie
Residential Segregation in the United States

Pinxten, Rik
Worldview

Piotrowski, Jessica T.
Media and Child Development

Pirard, Virginie
Research Funding: Ethical Aspects

Pissarides, Christopher A.
Search, Economics of

Platow, Michael J.
Leadership

Platt, Jennifer
Social Science Learned Societies and Professional Associations

Pleskac, Timothy J.
Decision and Choice: Luce's Choice Axiom

Ploder, Andrea
Ethical Questions in Social and Behavioral Sciences, History of
Semantic Differential

Pogge, Thomas
Justice: Philosophical Aspects

Poggi, Gianfranco
State and Society: History of the Concept

Pollak, Reinhard
Mobility, Social

Pollard, John H.
Population Dynamics: Theory of Stable Populations

Pollitt, Amanda M.
School Environment for LGBTQ/Sexual Minority Youth

Pollmächer, Thomas
Schizophrenia, Treatment of

Pon, Gordon
Ethnicity, Race, Oppression, and Social Work: The Canadian Case

Ponce, Aaron
Data Bases and Statistical Systems: Overview

Ponomareva, Inna
Sorokin, Pitirim Alexandrovich (1889–1968)

Pontius, Robert G.
Spatial Decision Support Systems

Poole, Jennifer M.
Mental Health, Mentalism and Sanism

Poole, Lydia
Infectious Diseases: Psychosocial Aspects

Poortinga, Ype H.
Cross-Cultural Psychology

Poplack, Shana
Code Switching: Linguistic

Poropat, Arthur E.
Personality and Educational Outcomes

Porrini, Patrizia
Information and Knowledge, Organizational

Port, Andrew I.
History from Below, the History of Everyday Life, and Microhistory

Porter, Theodore M.
Quantification in the History of the Social Sciences

Possing, Birgitte
Biography: Historical

Post, Marcel W.M.
Disability: Social and Psychological Aspects

Postiglione, Gerard A.
Ethnicity and Educational Achievement

Postmes, Tom
Deindividuation, Psychology of

Poston, Dudley L.
Human Ecology

Potagas, Constantin
Language Disorders, Treatment and Remediation of

Potter, Roberto Hugh
Crime and Criminal Justice: Applied Research from Routine Monitoring to Evidence-Based Practices

Potter, Trevor
Political Money and Party Finance

Potts, Deborah
Urbanization in Africa

Powell, Daniel A.
Pavlov, Ivan Petrovich (1849–1936)

Powell, Walter W.
Boundaries and New Organization Forms
New Institutionalism in the Analysis of Complex Organizations

Powers, Mark B.
Panic Disorder and Agoraphobia Across the Lifespan

Pozos-Brewer, Rose K.
Interest, Psychology of

Prado, Andrea
Social Business

Prainsack, Barbara
Bioethics in the Post-genomic Era
Forensic Genetic Databases: Ethical and Social Dimensions
Genetics and Forensics

Prakken, Henry
Legal Reasoning: Computational Models

Pratt, Nick
Situated Learning (Learning In Situ)

Pratto, Felicia
Social Dominance Orientation

Preda, Alex
Postmodernism in Sociology

Prentice, Ross L.
Survival Analysis: Overview

Prentky, Robert A.
Repetitive Sex Offenders, Treatment of

Prescott, Carol A.
Genetic Studies of Behavior: Methodology

Presler-Marshall, Elizabeth
Cash Transfers

Press, Andrea L.
Audiences, Media

Preston-Shoot, Michael
Legal Bases of Social Work

Price, Jammie
Applied Sociology

Price, Richard
Maroons in Anthropology

Prinz, M.
White-Collar Workers, History of

Prior, Jason H.
Lesbian, Gay, Bisexual and Trans-sexual: Commercial Sex

Pritchard, Adam J.
Beck, Ulrich (1944–2015)

Pritchard, Colin
Psychiatric Social Work in the United Kingdom

Pritchard, Jacki
Elder Abuse

Procacci, Giovanna
Social Question: Impact on Social Thought

Przeworski, Adam
Quantitative Cross-national Research Methods

Psillos, Stathis
Counterfactual Reasoning, Qualitative: Philosophical Aspects

Puce, Aina
Face Recognition, Psychological and Neural Aspects

Puckett, Julia A.
Lesbian, Gay, Bisexual and Trans-sexual Butch-Femme Subcultures

Pudrovska, Tetyana
Marital Quality and Health

Puetz, Kyle
Culture and Networks

Puhle, Hans-Jürgen
Interest Groups, History of

Pullum, Thomas W.
Population Dynamics: Classical Applications of Stable Population Theory

Purser, Harry
Cross-Sectional Methodologies in Developmental Psychology

Pustejovsky, James
Lexicon

Putnam, Samuel P.
Temperament

Pyles, John A.
Neural Representations of Objects

Q

Quaglia, Lucia
Financial Regulation

Quandt, Markus
Data Bases and Statistical Systems: International Comparative Research

Quane, James M.
Inner City

Quentin, Laura L.
Motivation in Clinical Interventions

R

Rabbitt, Patrick
Broadbent, Donald Eric (1926–93)

Rabinowitz, Dan
Community Studies: Anthropological

Radin, Joanna
Human Genome Diversity Project: History

Radner, Roy
Decision and Choice: Bounded Rationality

Rae, Douglas W.
Majoritarianism and Majority Rule

Ragin, Charles C.
Case-Oriented Research

Rai, Sauharda
Child Soldiers

Raible, Wolfgang
Literacy and Orality

Railton, Peter
Ethics and Values

Raja, Deepti S.
Law and People with Disabilities

Rajaram, Prem Kumar
Refugees: Geographical Aspects

Raley, Kelly
Divorce, Sociology of

Raley, Sara
Marriage and the Dual-Career Family

Ralph, Matthew A.L.
Cognitive Neuroscience

Ramachandran, Indu
Tacit Knowledge, Psychology of

Ramat, Paolo
Grammaticalization

Rammstedt, O.
Simmel, Georg (1858–1918)

Ramon, Shulamit
Armed Conflict and Social Work: Significance and Implications

Ramsay, James O.
Functional Data Analysis

Ramsay, R.
Mental Illness: Family and Patient Organizations in Europe

Randall, Asa R.
North America, Archaeology of

Rankin, Bruce H.
Inner City

Raphael, Lutz
Bielefeld School of History

Rapoport, Anatol
Neumann, John von (1903–57)

Rapport, Nigel
Personhood, Anthropology of

Ratcliff, Roger
Diffusion and Random Walk Processes

Rath, Thomas
Mexican Revolution, The

Rathmann, Katharina
Welfare and Education

Rau, Henrike
Time Use and Resource Consumption

Rauschecker, Josef P.
Human Language, Evolution of
Speech Processing, The Cortical Organization of

Raven, Diederick
Relativism, Cognitive

Ravetz, Jerome R.
Peer Review and Quality Control
Science, New Forms of

Rawlings, Craig
Formal Methods of Cultural Analysis

Ray, Larry
Critical Theory: Contemporary

Rayburn, Rachel L.
Guns and Criminals

Raychaudhuri, Tapan
Nationalism, Historical Aspects of: South Asia

Raychaudhuri, T.
Historical Thought and Historiography: South Asia

Rayer, Stefan
Demographic Techniques: Small-area Estimates and Projections

Raymo, James M.
Second Demographic Transition

Rayner, John C.W.
Statistical Distribution Models: Goodness of Fit Tests

Rayner, Keith
Eye Movements in Reading

Rayner, Stephen G.
Cognitive Styles and Learning Styles

Raynor, Peter
Probation, Offenders and Risk

Razsa, Maple
Anarchism

Read, Dwight
Kinship, Formal Models of
Kinship Terminology

Reader, Ian
Religions of East Asia

Redd, Adrienne
Attempted Utopias and Intentional Communities
Literary and Philosophical Utopian Thought

Reddy, Vasudevi
Self in Culture: Early Development

Redhead, Mark
Taylor, Charles (1931–)

Reed, Brenda
Castells, Manuel (1942–)

Rees, Jonas
Environmental Attitudes and Behavior: Measurement

Reese-Schäfer, Walter
Communitarianism

Reficco, Ezequiel
Social Business

Regenstein, Marsha
Socioeconomic Status and Health Care

Regoeczi, Wendy C.
Homicide

Regös, Erika Nóra
Data Bases and Statistical Systems: Environment and Urban Studies

Reichardt, Charles S.
Internal Validity
Nonequivalent Group Designs

Reicher, Stephen D.
Leadership
Obedience (Research), History of
Self-Categorization Theory
Tyranny

Reicher, Stephen
Collective Behavior, Social Psychology of

Reichert, Elisabeth
Human Rights and Social Work

Reichman, Daniel
Americas, Sociocultural Overviews: Central America

Reichmann, Werner
Epistemic Cultures
National Account Systems

Reid, Anthony
Historical Thought and Historiography: Southeast Asia

Reid, Lesley W.
Communities and Crime

Reid, Nancy
Significance, Tests of
Statistical Sufficiency

Reid-Merritt, Patricia
Afrocentric Approaches to Social Work

Reifer, Thomas E.
Wallerstein, Immanuel (1930–)

Reingle, Jennifer M.
Drugs: Illicit Use and Prevention

Reinhard, Wolfgang
Colonization and Colonialism, History of
State, History of

Reinhardt, Jan D.
Disability: Social and Psychological Aspects
Natural Disasters: Health-Related Aspects

Reischies, Friedel M.
Dementia: Psychiatric Aspects

Reissing, Elke D.
Sexuality, Theories of

Renn, Jürgen
Scientific Revolution, History and Sociology of

Renn, Ortwin
Environmental Risk

Renner, Britta
Health Behaviors, Assessment of
Health Risk Perception

Renninger, K. Ann
Interest, Psychology of

Renteln, Alison Dundes
Cultural Rights and Culture Defense: Cultural Concerns

Renzetti, Marta
Corporate Finance: Financial Control

Repetti, Rena L.
Social and Emotional Development in the Context of the Family

Rescher, Nicholas
Neutrality, Axiological

Resnik, David B.
Objectivity of Research: Ethical Aspects

Restubog, Simon L.D.
Psychological Contracts

Reuben, Alicja
Multinational Corporations

Reusser, Kurt
Co-constructivism in Educational Theory and Practice

Reuter, Astrid
Religion, History of

Reuter-Lorenz, Patricia A.
Dysexecutive Amnesia

Revell, Piers
Climate Change

Revers, Matthias
Cultural Sociology, History of

Rey, Sergio J.
Mathematical Models in Geography

Reybold, L. Earle
Qualitative Research

Reynolds, Katherine J.
Social Identity in Social Psychology

Rhoads, Mohja
Transportation Geography

Riba, Michelle B.
Psychotherapy and Pharmacotherapy

Ribot, Krystal M.
Language Development: Influence of Socio-Economic Status

Ric, François
Social Cognition

Ricciardelli, Lina A.
Body Image

Ricco, Robert B.
Epistemic Doubt During Adolescence

Richard, Jean-Francois
Monte Carlo Methods and Bayesian Computation: Importance Sampling

Richards, Paul
Africa, Sociocultural Overviews: West Africa

Richardson, Paul W.
Teacher Motivation

Richardson, Ulla
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Richelle, Marc N.
Skinner, Burrhus Frederick (1904–90)

Richter, Dirk
Sociology and Psychiatry

Richters, Juliet
Sexuality in Australia

Ricks, JaNelle M.
HIV Risk Interventions

Ricuperati, Giuseppe
Enlightenment

Rid, Annette
Placebo-Controlled Trials, Ethics of

Rieber, Alfred J.
Frontiers in History

Riediger, Michaela
Affect-Regulation Motivation

Riesch, Hauke
Citizen Science

Riese, Berthold
Historical Thought and Historiography: Indigenous Cultures in the Americas

Riese, Florian
Geriatric Psychiatry

Riethmüller, Albrecht
Music, History of

Riga, Liliana
Ethnic Cleansing, History of

Rijkhoff, Jan
Word Order

Riley, Jonathan
Welfare: Philosophical Aspects

Rindskopf, David
Reliability: Measurement

Ríos, Diego
Prisoner's Dilemma, One Shot and Iterated

Rip, Arie
Technology Assessment

Ritter, Moritz
Promotions

Ritzer, George
McDonaldization

Rizzolatti, Giacomo
Premotor Cortex

Roach, Levi
Feudalism

Robben, Antonius C.G.M.
Ethnography

Robbennolt, Jennifer K.
Negotiation and Bargaining: Role of Lawyers

Robbers, Gerhard
Church and State

Roberge, Jonathan
Ricoeur, Paul (1913–2005)

Roberts, Brent W.
Conscientiousness

Roberts, Carl W.
Content Analysis

Roberts, Elizabeth
Reproduction and Cultural Anthropology

Roberts, Karlene H.
Organizational Errors: Catastrophic

Roberts, Ken
Leisure, Sociology of
Youth Movements

Roberts, Laura W.
Psychiatry: Informed Consent and Care

Roberts, Les
Liminality

Roberts, Richard D.
Coping, Personality and

Roberts, Richard H.
Religion and Economic Life

Roberts, Simon
Common Law

Robertson, Alexander F.
Primitive Society

Robins, James M.
Longitudinal Causal Inference

Robinson, Ian
Area and International Studies: Political Economy

Robinson, John P.
Leisure and Cultural Consumption: US Perspective
Time Use Research: Recent Developments
Time Use Research: Subjective Time Use

Robinson, Michael D.
Personality and Self

Robinson, Mike J.F.
Wanting vs Needing

Rock, Paul
Prison Life, Sociology of

Rock, Stephen R.
Appeasement: Political

Rockstuhl, Thomas
Cultural Intelligence and Competencies

Rodden, Frank A.
Functional Brain Imaging of Language processes

Rodima-Taylor, Daivi
Land Tenure

Rodríguez, Germán
Multilevel Models in Demography

Rodríguez-Planas, Núria
Vocational Training

Roeck, Bernd
Renaissance

Roeder, Peter M.
Educational Institutions, History of

Roediger, Henry L.
Reconstructive Memory, Psychology of

Roelfs, David J.
Meta-Analysis, History of

Rogers, Everett M.
Evolution: Diffusion of Innovations

Rogers, Lesley J.
Animal Rights in Research and Research Application
Brain and Behavioral Lateralization in Animals

Rogerson, Christian M.
Local Economic Development

Rogosa, David
Resampling Methods of Estimation

Rogowski, Ralf
Law, Autopoiesis in

Rohall, David
Social Psychology, Sociological

Rohr, Margund K.
Successful Aging in Western Societies: The 'Selection, Optimization, and Compensation' Model

Rohracher, Harald
Science and Technology Studies, History of

Rolls, Edmund T.
Emotion, Neural Basis of
Taste and Smell, Psychology of

Romaine, Suzanne
Language and Social Class

Romero, Kristoffer
Amnesia: General

Römmele, Andrea
Social Science and the Media

Ronen, Shelly
Hooking up and Casual Sex

Roniger, Luis
Eisenstadt, Shmuel N. (1923–2010)
Patron–Client Relations, Social and Anthropological Study of

Roof, Wade Clark
Christianity in the English-Speaking West

Room, Robin
Drug Control: International Aspects

Rooney, Patrick
Charity and Philanthropy: Overview

Roose, Henk
Cultural Participation, Trends in

Root, Hilton L.
State, Sociology of the

Root, Melissa
Special Education

Rootes, Christopher
Environmental Movements

Roquet, Rheall F.
Anxiety and Fear Conditioning, Neural Basis of

Rose, Nancy L.
Regulation, Empirical Analysis of
Regulation, Political Economy of

Rose, Peter I.
Wilson, William Julius (1935–)

Rosen, Robert E.
In-House Counsel

Rosenau, Pauline V.
Postmodernism: Methodology

Rosenbaum, Paul R.
Observational Studies: Overview

Rosenberg, Shawn W.
Opinion Formation, Theory of

Rosenbury, Laura A.
Children and Law

Rosenfield, David B.
Fluency and Voice, Disorders of

Rosnow, Ralph L.
Experimenter and Subject Artifacts: Methodology

Ross, Catherine E.
Alienation: Psychosociological Tradition

Ross, Dorothy
Progress: History of the Concept

Rossi, Ugo
Urban Neoliberalism

Rössler, Wulf
Social Intervention in Psychiatric Disorders

Rota, Giuseppina
Neuroethics: Ethics of Science and Science of Ethics

Rote, Sunshine
Migration and Health

Rotenberg, Ken J.
Interpersonal Trust across the Lifespan

Rothbart, M.K.
Temperament and Human Development

Rothblatt, Sheldon
Social Science and Universities

Rothenbuhler, Eric W.
Media Events

Rothstein, Bo
Quality of Government

Rottleuthner, Hubert
Nazi Law

Roumani, Judith
Memmi, Albert (1920–)

Rourke, Francis E.
Public Bureaucracies

Rouse, Joseph
Cultural Studies of Science

Routh, Donald K.
Clinical Psychology in North America, History of

Routledge, Clay
Personality and Self

Rovane, Carol
Self: Philosophical Aspects

Rovee-Collier, Carolyn
Prenatal and Infant Development: Overview, Current Trends, Future Directions

Rowe, David
Mass Media and Sports

Rowen, Jamie
War Crimes Tribunals, Crimes against Humanity and Transitional Justice

Rowland, Benjamin A.
Multisensory Integration, Principles of

Rowland, David L.
Sexual Dysfunction, Epidemiology of

Rowland, Neil E.
Hunger and Eating, Neural Basis of

Roy, Ananya
Urban Informality: The Production and Regulation of Space

Rozin, Paul
Disgust, Psychology of
Food Preferences, Psychology and Physiology of

Ruben, David-Hillel
Social Properties (Facts and Entities): Philosophical Aspects

Ruberton, Peter M.
Subjective Wellbeing, Psychology of

Rubin, Donald B.
Missing Data
Multiple Imputation

Rubin, Edward
Legal Scholarship

Rubin, Kenneth H.
Friendship During Infancy and Early Childhood and Cultural Variations

Rubinfeld, Daniel L.
Antitrust Policy: Lessons from the US

Rubini, Lauretta
Special Economic Zones and Cluster Dynamics: China

Rucker, James J.
Mental Illness, Genetics of

Rudas, Tamás
Log-Linear and Marginal Models

Rue, Jim
Psychological Treatments Delivered Nontraditionally

Ruether, Rosemary Radford
Catholicism and Gender

Ruff, Christopher B.
Body, Evolution of

Ruggles, Steven
Census Microdata

Ruhs, August
Lacan, Jacques (1901–81)

Rumbaut, Rubén G.
Assimilation of Immigrants

Runco, Mark A.
Creativity Training

Ruppanner, Leah
Female Power at Work

Ruse, Michael.
Darwinism

Rüsen, Jörn
History: Overview

Rushin, Stephen
Law, Mobilization of

Russell, Stephen T.
School Environment for LGBTQ/Sexual Minority Youth

Russell, Zachary A.
Power, Politics, and Influence in Organizations

Rusten, Grete
Service Economy, Geography of

Ruzek, Erik A.
Classroom Level in Educational Effectiveness Research

Ryan, Kevin W.
Agonism

Ryan, Richard M.
Self-Determination Theory

Ryten, Jacob
Data Bases and Statistical Systems: Economics (Statistical Systems)

S

Saçkes, Mesut
Scientific Concepts during Childhood, Development of

Sadeghi, Soroush G.
Vestibular System

Saeed, John I.
Semantics

Saef, Rachel
Openness to Experience

Sagi-Schwartz, Abraham
Integrated Attachment Theory

Saha, Lawrence J.
Educational Sociology

Saito, Osamu
Historical Demography

Sajó, András
Socialist Law

Saklofske, Donald H.
Intelligence: Assessments of

Sala, Sergio Della
Retroactive Interference

Salais, Robert
Keynesianism and the Question of State Interventionism

Salas, Eduardo
Teamwork and Team Performance Measurement

Salaymeh, Lena
Islamic Law
Middle Eastern and North African Legal Traditions

Salice, Alessandro
Searle, John (1932–)

Salmela-Aro, Katariina
Passion and Motivation
School Burnout and Engagement: Lessons from a Longitudinal Study in Finland
Sociodigital Revolution: Digital Natives vs Digital Immigrants

Salminen, Simo
Workplace Safety and Health

Salomon, Gavriel
School Learning for Transfer

Salomone, Erica
Autism Spectrum Disorders

Salto, Dante J.
Education in Latin America and the Caribbean: Systems and Research

Salvemini, Mauro
Global Positioning System

Salzano, Francisco M.
Microevolution

Sam, David L.
Acculturation

Sampson, Allan R.
Hotelling, Harold (1895–1973)

Samuel, Geoffrey
Legal Reasoning and Argumentation

Samuels, Richard
Irrationality: Philosophical Aspects

Samuelson, Paul A.
Ricardo, David (1772–1823)

Sanchez-Mazas, Margarita
Xenophobia: Social Psychological Aspects

Sanday, Peggy Reeves
Rape and Sexual Coercion

Sanders, Joseph
Scientific Evidence: Legal Aspects

Sanderson, Penelope
Human Factors and Ergonomics

Sanderson, William C.
Psychological Treatments, Empirically Supported

Sandmo, Agnar
Income Taxes

Sandra, Dominiek
Spelling, Psychology of

Sanjek, Roger
Field Observational Research in Anthropology and Sociology

Sansi Roca, Roger
Fetishism

Santos, Daniel D.
Personality and Economics

Santos, Joel
Education and Economic Growth

Sapiro, Gisèle
Bourdieu, Pierre (1930–2002)
Field Theory
Habitus: History of a Concept

Sappington, David E.M.
Regulation, Economic Theory of

Sarat, Austin
Injustice: Legal Aspects

Saroglou, Vassilis
Personality and Religion

Sarrasin, Oriane
Immigration: Social Psychological Aspects

Sarris, Viktor
Köhler, Wolfgang (1887–1967)

Sassen, Saskia
Cities: Capital, Global, and World

Satlof-Bedrick, Emma
Prosocial Behavior during Childhood and Cultural Variations

Sato, Jin
Social Resilience in Post-Tsunami Japan

Sato, Masayuki
Historical Thought and Historiography: East Asia
Time, Chronology, and Periodization in History

Satz, Debra
Market and Nonmarket Allocation

Sauerland, Dirk
Ideologies, Institutions, and the New Institutionalism

Savage, James D.
Budgeting: Political Science Aspects

Savage, Victor R.
Social Science Infrastructure: Asia

Savelsberg, Joachim J.
Criminology, History of

Savin-Williams, Ritch C.
Gay, Lesbian, and Bisexual Youth

Savoia, Paolo
Objectivity: Philosophical Aspects

Sawyer, Cheryl
Mortality and the HIV/AIDS Epidemic

Sayer, Andrew
Critical Realism in Geography

Scaff, Lawrence A.
Weberian Social Thought, History of

Scaggs, Samuel J.A.
Recidivism

Scarf, Damian
Childhood Memories
Memory Development During Infancy and Early Childhood across Cultures

Scarpellini, Emanuela
Consumption, History of

Scarpetta, Stefano
Development and Employment: A Mutual Relationship

Schabas, Margaret
Political Economy, History of

Schaefer, Richard T.
Minorities
Prejudice in Society: Sociological Perspectives

Schaeffer, Nora C.
Surveys: Question Wording and Response Categories

Schäfer, Mike S.
Climate Change and the Media

Schaffartzik, Anke
Material Flow Analysis

Schaie, K. Warner
Aging, Theories of
Cognitive Aging

Schandl, Heinz
Material Flow Analysis

Schanzenbach, Diane W.
Class Size

Schatten, Jeff
Genetic Factors in Organizational Psychology

Schaubroeck, John M.
Supervision, Abusive

Schauer, Maggie
Narrative Exposure Therapy

Schaufeli, Wilmar B.
Coping with Job Stress

Scheepers, Peer
Religiosity and Prejudice against Minorities

Scheerens, Jaap
School Effectiveness Research

Scheff, Thomas
Shame and the Social Bond

Scheffler, Robin W.
Biomedical Sciences, History and Sociology of

Scheiblechner, Hartmann
Test Theory: Applied Probabilistic Measurement Structures

Schein, Edgar H.
Corporate Culture

Scheingold, Stuart
Rights: Legal Aspects

Scheinsohn, Vivian
Hunter-Gatherer Societies, Archaeology of

Schell, Patience A.
Eugenics in the Americas

Schellinck, Heather M.
Olfactory System

Schenker, Nathaniel
Multiple Imputation

Scher, Philip W.
Carnival

Scherer, Klaus R.
Emotions, Psychological Structure of

Scherer, Nikolas
Survey Research Centers and Companies

Scherke, Katharina
Emotion: History of the Concept

Schervish, Mark J.
Errors in Statistical Assumptions, Effects of

Schiavone, Giuseppe
Deliberation and Democracy

Schiller, Naomi
Liberalism, Anthropology of

Schilling, René L.
Martingales

Schimank, Uwe
Differentiation: Social

Schippers, Birgit
Butler, Judith (1956–)

Schlarb, Angelika
Sleep and Health

Schlee, Günther
Conflict: Anthropological Aspects
Language and Ethnicity
Nomads and Nomadism in History

Schlegel, John Henry
Legal Realism

Schleidt, Wolfgang M.
Lorenz, Konrad (1903–89)

Schlesinger, Philip
Mass Media and Cultural Identity

Schlolz, Fred
Nomads and Nomadism in History

Schlozman, Kay Lehman
Interest Groups

Schmälzle, Ralf
Health Risk Perception

Schmandt-Besserat, Denise
Writing, Evolution of

Schmaus, Warren
Determinism: Social and Economic

Schmidt, Gustav
Cold War, The

Schmidt, Manfred G.
Modernization

Schmidtke, Jörg
Genetic Counseling: Historical, Ethical, and Practical Aspects

Schmiedek, Florian
Intelligence: Central Conceptions and Psychometric Models

Schmitt-Beck, Rüdiger
Data Bases and Statistical Systems: Political Behavior and Elections

Schmitz, Andreas
Correspondence Analysis, History of

Schmutz, Vaughn
Classifications in Popular Music

Schnapp, Alain
Classical Archaeology

Schneebaum, Alyssa
Lesbians: Social and Economic Situation

Schneewind, Klaus A.
Socialization and Education: Theoretical Perspectives

Schneider, Barbara
Passion and Motivation
Vocational Interests, Values, and Preferences, Psychology of

Schneider, Friedrich
Informal and Underground Economics
Shadow Work: Measurement
State and Local Taxation

Schneider, Margaret
Ecology and Health

Schneider, Marianne E.
Motivational Development, Systems Theory of

Schneider, Wolfgang
Metacognitive Development: Educational Implications

Schneiderman, Howard G.
Authority, Social Theories of

Schneider-Sliwa, Rita
Urban Geography

Schnittker, Jason
Gene–Environment Interactions in Health and Well-Being

Schober, Barbara
Program Evaluation

Schodl, Michal M.
Performance Appraisal and Evaluation

Schoen, Robert
Population Dynamics: Two-Sex Demographic Models

Schögler, Rafael Yann
Awards, Public Recognition, and Evaluation

Scholz, Oliver R.
Hermeneutics

Schomburg, Eike D.
Sherrington, Sir Charles Scott (1857–1952)

Schott, Franz
Instructional Design

Schöttler, Peter
Bloch, Mark Léopold Benjamin (1886–1944)

Schrader, Friedrich-Wilhelm
School Achievement: Motivational Determinants and Processes

Schreier, Helmut
Dewey, John (1859–1952)

Schretlen, David J.
Transcranial Direct Current Stimulation and Language

Schreyögg, Georg
Development: Organizational
Organizational Path Dependence

Schriefers, Herbert
Speech Production, Psychology of

Schriewer, Jurgen
Education, History of

Schroeder, Ralph
Mann, Michael (1942–)

Schroeder-Heister, Peter
Popper, Karl Raimund (1902–94)

Schröer, Andreas
Data Bases and Statistical Systems: Education, General

Schultz, John J.
Forensic Archaeology

Schultz, Paul T.
Fertility Transition: Economic Explanations

Schulz, André
Interoception

Schulze, Ralf
Factor Analysis and Latent Variable Models in Personality Psychology

Schunder, Torsten
International Business: Geographic Aspects

Schunk, Dale H.
Self-Efficacy: Education Aspects

Schupp, Harald T.
Health Risk Perception

Schuppert, Gunnar F.
Governance

Schurr, Theodore G.
Genographic Project

Schut, Henk
Bereavement

Schüz, Benjamin
Public Health as Social Science

Schüz, Natalie
Sun Exposure and Skin Cancer Prevention

Schwanen, Tim
Automobility

Schwank, Elisabeth
Mathematical Concepts during Early Childhood Across Cultures, Development of

Schwank, Inge
Mathematical Concepts during Early Childhood Across Cultures, Development of

Schwartz, Barry
Commemoration

Schwartz, Shalom H.
Values, Social Psychology of

Schwartz, Sophie
Dreaming, Neural Basis of

Schwarz, Norbert
Attitude Measurement
Questionnaires: Cognitive Approaches

Schwarzer, Gudrun
Face Perception Development During Early Childhood

Schwarzer, Ralf
Health Self-Regulation, Motivational and Volitional Aspects of

Schweickert, Richard
Network Models of Tasks

Schweitzer, Peter P.
Arctic: Sociocultural Aspects

Sciortino, Giuseppe
AGIL, History of

Sciulli, David
Parsons, Talcott (1902–79)

Scollon, Christie Napa
Subjective Wellbeing and Culture

Scott, David W.
Probability Density Estimation

Scott, David
Colonialism, Anthropology of

Scott, John
Social Networks
Status and Role: Structural Aspects

Scott, Kate M.
Co-morbidity of Mental and Physical Conditions

Scott, Lauren M.
Spatial Pattern, Analysis of

Scott, Linda
Advertising and Advertisements

Scott, W. Richard
Organization: Overview

Scotto, Thomas J.
Efficacy, Political

Scourfield, Jonathan
Fathers and Fathering: Social Work Perspectives

Scriver, Stacey
War Propaganda

Scroggins, Michael
Culture of Poverty: Critique

Scullin, Michael K.
Prospective Memory

Sear, Rebecca
Evolutionary Demography: A Darwinian Renaissance in Demography

Sears, Meredith S.
Social and Emotional Development in the Context of the Family

Seaton, Gina A.
Work Motivation

Seaton, Marjorie
Self-Concept: From Unidimensional to Multidimensional and Beyond

Sebba, Mark
Serial Verb Constructions

Sebenius, James K.
Negotiations: Statistical Aspects

Seel, Norbert M.
Instructional Design

Segal, David R.
Military Sociology

Segal, Robert A.
Eliade, Mircea (1907–86)

Segal, Robert
Myth in Religion

Segal, Zindel
Mindfulness

Segerstrale, Ullica
Sociobiology, History of

Seginer, Rachel
Future Time Perspective and Motivation

Seidenfeld, Teddy
Probability Theory: Interpretations

Seiffge-Krenke, Inge
Health Promotion in Schools

Seifritz, Erich
Sleep Disorders: Psychiatric Aspects

Sejnowski, Terrence J.
Computational Neuroscience

Selbin, Jeffrey
Poverty Law: United States

Selgelid, Michael J.
Censorship in Research and Scholarship

Selkirk, Elisabeth
Syntax–Phonology Interface

Sellers, Mortimer NS
Republicanism: Philosophical Aspects

Sellitto, Manuela
Attention, Neural Basis of

Semenza, Carlo
Proper and Common Names: Impairments of Anomia

Semigina, Tetyana
International Association of Schools of Social Work

Sen, Hilal
Eisenberg's Theory of Prosocial Reasoning

Seretis, Dionysis
Explanatory Style and Health

Serlavos, Ricard
Leadership Education: Theory and Practice

Seth, Puja
HIV Risk Interventions

Sethuraman, Jayaram
Nonparametric Statistics: Advanced Computational Methods
Nonparametric Statistics: Rank-Based Methods

Seufert, Wolfgang
Data Bases and Statistical Systems: Communications and Media

Sevilla, Almudena
Gender, Time-Use, and Fertility Recovery in Industrialized Countries

Sewell, Graham
Conflict: Organizational

Sewpaul, Vishanthie
Neoliberalism

Shackelford, Todd K.
Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future

Shackley, Steven
Chronology, Stratigraphy, and Dating Methods in Archaeology

Shami, Seteney
Middle East, North Africa: Sociocultural Overview

Shanks, David R.
Concept Learning and Representation: Models

Shannon, Deric
Queer Anarchism

Shapin, Steven
Truth and Credibility in Science

Shapiro, David
Fertility Transition: Sub-Saharan Africa
Hypertension: Psychosocial Aspects

Shapiro, Robert Y.
Political Science and Polling

Shapiro, Susan P.
Crime: Whitecollar

Shapiro, Warren
Totemism

Sharkey, Patrick
Neighborhood Effects

Sharot, Stephen
Judaism

Sharp, Susan F.
Feminist Criminology and Gender Studies

Shavell, Steven
Law and Economics

Shaw, Gareth
Tourism, Geography of

Shaw, Martin
Genocide and War

Sheehan, James J.
Democracy, History of
Political History: History of Politics

Sheinheit, Ian
Skocpol, Theda (1947–)

Shelton, Taylor
Information Society, Geography of

Shen, Qing
Transportation Planning

Shennan, Stephen
Technology, Evolution of

Sherkat, Darren E.
Religion: Mobilization and Power
Religiosity
Women's Religiosity

Shetty, Nitasha
Psychopharmacotherapy: Side Effects

Shibatani, Masayoshi
Linguistic Typology
Syntax (General)

Shih, Elena
Prostitution and Sex Work

Shin, Hyun Bang
Urbanization in China

Shipley, Jesse W.
Trickster Ethnography

Shipton, Parker
Land Tenure

Shizgal, Peter
Brain Stimulation Reward

Shlay, Anne B.
Urban Sociology

Shokeid, Moshe
Fieldwork in Social and Cultural Anthropology

Shomstein, Sarah
Hemispatial Neglect, Neural Basis of

Shore, Allison L.
Interpersonal Psychotherapy

Short, James F.
Crime, Sociology of
Criminology, History of
Gangs, Sociology of

Short, John Rennie
City Marketing

Shrum, Wesley M.
Science and Development

Shu, Hua
Reading Skills, Acquisition of: Cultural, Environmental, and Developmental Impediments

Shuffler, Marissa L.
Teamwork and Team Performance Measurement

Shultz, Thomas R.
Constraint-Satisfaction Models

Shuman, Vera
Emotions, Psychological Structure of

Shweder, Richard A.
Cultural Psychology
Culture: Contemporary Views

Si, Kangping
Spatial Decision Support Systems

Si-bin, Wang
Social Work in China

Sica, Alan
Encyclopedias, Handbooks, and Dictionaries

Sidnell, Jack
Conversation Analysis: Sociological

Siegel, Derek P.
Transgender, Transexual, and Gender Variant Individuals

Siegrist, Hannes
Bourgeoisie and Middle Classes, History of
Professions and Professionalization, History of

Siegrist, Johannes
Health Care Markets: Theory and Empirical Evidence
Justice and Health
Stress at Work

Sierminska, Eva M.
Beauty and the Labor Market

Sigal, Leon V.
Deterrence

Sigmund, Paul E.
Conservatism: Theory and Contemporary Political Ideology

Silbereisen, Rainer K.
Alcohol Use among Young People
Leisure Activities Choices among Adolescents
Social and Economic Change: Psychological Challenges for Individuals

Silbey, Susan S.
Legal Culture and Legal Consciousness

Silk, Alex
Ethics and Values

Sillitoe, Paul
Horticultural Societies

Silva, Alcino J.
Genetic Approaches to Memory

Silva, Carlos N.
Planning Ethics

Silvanto, Juha
Mental Imagery: Visual Cognition

Silver, Hilary
Poverty Policy

Silverman, Bernard W.
Functional Data Analysis

Silverman, Eric K.
Time, Anthropology of

Sim, Timothy
Psychosocial Work

Simbayi, Leickness C.
Sexual Risk Behaviors

Šimleša, Sanja
Executive Functions during Childhood, Development of

Simon, Jonathan
Punishment: Social and Legal Aspects
Torture: Legal Perspectives

Simpkins, Sandra D.
Motivation, Familial Influences on

Simpson, Graeme
Choice, Empowerment, and Social Work

Sinclair, John
Advertising: General

Singer, Burton H.
Longitudinal Data

Singer, Judith D.
Survival Analysis: Whether, and If So, When?

Singer, Wolf
Neural Synchrony as a Binding Mechanism

Singh, Purnima
Values Across Cultures, Development of

Siow, Aloysius
Assortative Mating in the Marriage Market

Sirin, Selcuk R.
School Achievement as Moderated by Sociodemographic Status: The US Case

Siročić, Zorica
Gender: Gendering of Categories

Sironi, Maria
Transition to Adulthood

Sismondo, Sergio
Pharmaceutical Industry: Political Economies of Drugs and Knowledge
Standpoint Theory, in Science

Siu-Yin Lam, Silvia
Chinese Literacy Acquisition: A Multidimensional Puzzle

Skapska, Grażyna
Law and Democracy

Skeldon, Ronald
Migration and Development

Skelton, Tracey
Youth, Geography of

Skidelsky, Robert
Keynes, John Maynard (1883–1946)

Skinner, Ellen A.
Coping across the Lifespan

Skinner, Mark W.
Aging and Older People, Geography of

Skirton, Heather
Genetic Counseling: Historical, Ethical, and Practical Aspects

Sklair, Leslie
Capitalism: Global

Sklar, Marisa
Generalizability Theory

Skogerbø, Eli
Broadcasting: Regulation

Skorich, Daniel P.
Social Categorization

Skorska, Malvina N.
Sexual Orientation: Biological Influences

Skov, Martin
Neuroesthetics

Skubic, Darija
Concept Formation in Childhood: Cross-Cultural Perspectives

Skvoretz, John
Models of Power in Exchange Networks

Skyrms, Brian
Evolution, Natural and Social: Philosophical Aspects

Slade, Gavin
Mafias, Criminology of

Slavin, Robert E.
Cooperative Learning in Schools

Sliter, Michael
Workplace Stress

Sloane, Peter J.
Mismatch on the Labor Market

Sloas, Lincoln B.
Community Supervision and Diversion in the United States

Slobodskaya, Helena R.
Behavioral Problems, Effects of Parenting and Family Structure on

Slone, Edward
Human Spatial Orientation, Neural Basis of

Sluka, Jeffrey A.
Ethnography

Small, Dylan S.
Compliance

Small, Steven L.
Language Processing, Functional Magnetic Resonance Imaging of

Smeeding, Timothy M.
Income Distribution and Inequality: Measurement Issues

Smeeding, Tim
Poverty, Sociology of

Smelser, Neil J.
Collective Behavior, Sociology of

Smetana, Judith G.
Moral Development, Theories of

Smith, Daniel Jordan
Sexuality: Cultural Aspects

Smith, David Horton
Voluntary Associations, Sociology of
Voluntary Organizations

Smith, Dennis
Moore Jr, Barrington (1913–2005)

Smith, Edward E.
Cognitive Psychology: History

Smith, Greg
Interactionism, Symbolic

Smith, Jeremy C.A.
Translating Sociology into Japanese, Chinese and Korean Languages

Smith, Karen M.
Rehabilitation (Cardiac): Psychosocial Aspects

Smith, Ken R.
Early Life Influences on Health and Mortality in Adulthood

Smith, Kenny
Evolution and Language: Cultural Transmission
Language Development: Evolutionary Theories

Smith, Lindsay A.
Genetics and Social Justice

Smith, M. Brewster
James, William (1842–1910)

Smith, Mandy McCormick
Scientific Concepts during Childhood, Development of

Smith, Michael
Free Will and Action

Smith, Philip L.
Diffusion and Random Walk Processes

Smith, Raymond T.
Matrifocality

Smith, Roger
Criminal Justice and Social Work Practice

Smith, Shawna N.
Data Bases and Statistical Systems: Overview

Smith, Simon C.
Imperialism, History of

Smith, Steven Rathgeb
Social Science Infrastructure: North America (Social Science Academies and Related Organizations)

Smith, Susan J.
Residential Segregation: Geographic Aspects

Smith, Tom W.
International Social Survey Programme

Smith, Vincent K.
Environmental Economics

Smock, Pamela J.
Cohabitation: United States

Smolensky, Eugene
Public Policy Schools

Snijders, T.A.B.
Hypothesis Testing: Methodology and Limitations

Snopkowski, Kristin
Fertility Theory: Theory of Intergenerational Wealth Flows

Snow, Charles C.
Network Paradigm: Applications in Organizational Science

Snow, David A.
Collective Identity
Movements, Social

Snowden, Julie S.
Semantic Memory

Snyder, Mark
Volunteerism, Psychology of

Snyder, Peter J.
Human Language: Evolutionary Precursors

Snyder, Tom
Data Bases and Statistical Systems: Education, Statistical Systems

Sodian, Beate
Theory of Mind During Infancy and Early Childhood Across Cultures, Development of

Soifer, Aviam
Hurst, James Willard (1910–97)

Solomon, Peter H.
Law and Courts in Authoritarian States

Sommer, Werner
Facial Perception

Song, Jae Jung
Valency and Argument Structure in Syntax

Song, Ju-Hyun
Behavioral Inhibition and Social Withdrawal across Cultures

Song, Yan
Urban Internal Spatial Structure

Sonnentag, Sabine
Wellbeing and Burnout in the Workplace: Organizational Causes and Consequences

Soonaan, Nur I.B.
Research Ethics, Cross-Cultural Dimensions of

Soppe, Birthe
Boundaries and New Organization Forms

Sorensen, Aage B.
Coleman, James Samuel (1926–95)

Sørensen, Kristine
Health Literacy

Sorrentino, Constance
Data Bases and Statistical Systems: Labor and Work

Sortheix, Florencia M.
Moral Development, Cultural Differences in

Soyka, Michael
Alcohol-Related Disorders

Soysal, Yasemin N.
Marshall, Thomas Humphrey (1893–1981)

Sozzo, Maximo
Criminology, History of

Spaderna, Heike
Cardiac Disease, Coping with

Spain, Daphne
Gender and Place

Spapens, Toine
Transnational Organized Crime

Sparks, Richard
Criminology, History of

Spaulding, Cylor
Media Effects

Spector, Timothy D.
Twin Studies and Epigenetics

Spencer, Paul
Age: Anthropological Aspects

Spender, J.-C.
Strategizing

Spender, John-Christopher
Modern Management: Origins and Development
Simon, Herbert (1916–2001)

Spicer, André
Slow Food as a Social Movement

Spiegelman, Maura
Social Network Analysis

Spiel, Christiane
Program Evaluation

Spinath, Birgit
Developmental Behavioral Genetics and Education

Spinath, Frank M.
Developmental Behavioral Genetics and Education
Intelligence, Genetics of: Cognitive Abilities

Spirtes, Peter
Latent Structure and Causal Variables

Splichal, Slavko
Journalism and Journalists
Public Sphere and the Media

Spohn, W.
History and the Social Sciences

Sponsel, Leslie E.
Advocacy in Anthropology
Peace and Nonviolence: Anthropological Aspects

Sporn, Barbara
Universities and Science and Technology: Europe

Spring, Bonnie
Evidence-Based Practice

Sproesser, Gudrun
Health Behaviors, Assessment of

Sprung, Lothar
Wundt, Wilhelm Maximilian (1832–1920)

Squire, Larry R.
Declarative Memory, Neural Basis of

Sreberny, Annabelle
Media Imperialism

Srubar, Ilja
Schütz, Alfred (1899–1959)

St. Pierre, Robert G.
Random Assignment: Implementation in Complex Field Settings

Stacks, Don W.
Media Effects

Stade, Ronald
Diffusion: Anthropological Aspects

Stadler, Friedrich
Vienna Circle: Logical Empiricism

Staerklé, Christian
Political Psychology

Stafford, Mark C.
Deterrence Theory: Crime

Stahler-Sholk, Richard
Revolution

Staley, Richard
Physical Sciences: History and Sociology

Stallard, Eric
Mortality, Biodemography of

Stambe, Rose
Unemployment and Mental Health

Stanford, Naomi
Design and Form: Organizational

Stanford, Terrence R.
Multisensory Integration, Principles of

Staniforth, Barbara
Communication Skills in Social Work

Staniloiu, Angelica
Amnesia: Psychogenic
Learning and Memory, Neural Basis of

Stanley, Brian
Mission: Proselytization and Conversion

Stano, Nancy
Learning to Learn

Stansfield, Richard
Immigration and Crime in the United States

Starbuck, William H.
Information and Knowledge, Organizational

Stark, Miriam T.
Southeast Asia, Archaeology of

Starren, Marianne
Second Language Acquisition

Stasny, Elizabeth A.
Nonsampling Errors

Staub, Ervin
Mass Killings and Genocide, Psychology of

Staudinger, Ursula M.
Baltes, Paul B (1939–2006)

Stavrianakis, Anna
Militarism

Stearns, Peter N.
Emotions, History of

Steel, Piers
Procrastination

Steele, J. Michael
Probability Theory: Formal

Steglitz, Jeremy
Evidence-Based Practice

Stehr, Nico
Knowledge Society, History of

Steidl, Ellyn A.
Divorce, Sociology of

Stein, Arthur A.
Realism/Neorealism

Stein, Barry E.
Multisensory Integration, Principles of

Stein, Claudia
Data Bases and Statistical Systems: Health

Stein, Larry
Reinforcement: Neurochemical Substrates

Stein, Solomon M.
Alternative Schools of Economic Thought

Steinberg, Laurence
Contextual Studies: Methodology

Steinebach, Christoph
Resilience

Steinley, Douglas
Psychometrics: Combinatorial Data Analysis

Steinmo, Sven
Institutionalism

Steketee, Michael
Social Network Analysis

Stemmer, Brigitte
Functional Brain Imaging of Language processes
Handedness and Language, Relationship between
Lateralization of Language as Demonstrated by Brain Imaging Procedures
Neural Basis of Reading
Pragmatic Disorders

Stepan-Norris, Judith
Political Sociology

Stephens, Elizabeth J.
Self-Concepts: Educational Aspects
Test Anxiety and Academic Achievement

Stephens, Melika H.
Social Support and Recovery from Disease and Medical Procedures

Steptoe, Andrew
Coronary Heart Disease: Psychosocial Aspects
Infectious Diseases: Psychosocial Aspects

Sterelny, Kim
Evolutionary Approaches to Human Behavior: Philosophical Aspects

Sterett, Susan
Regulation and Administration

Stern, Elsbeth
Intelligence, Prior Knowledge, and Learning

Stern, Hal S.
Bayesian Statistics

Sternberg, Robert J.
Binet, Alfred (1857–1911)
Intelligence: Historical and Conceptual Perspectives

Sternberg, Yistzhak
Ethnicity, Sociology of

Steur, Luisa
Global South

Stevens, Fred C.J.
Medical Education

Stevenson-Hinde, Joan
Bowlby, John (1907–90)

Stewart, Andrew L.
Social Dominance Orientation

Stewart, Charles
Honor and Shame

Stewart, Kathleen
Affect and Emotion, Anthropology of

Stewart, Michael
Roma People

Stewart-Williams, Steve
Morality: Evolution of

Steyer, Rolf
Classical (Psychometric) Test Theory

Stich, Stephen P.
Irrationality: Philosophical Aspects

Stichweh, Rudolf
Luhmann, Niklas (1927–98)
Scientific Disciplines, History of

Stille, Charles J.
First World War, The

Stimpfel, Amy W.
Nonphysician Health Professionals

Stine-Morrow, Elizabeth A.L.
Education and Learning: Lifespan Perspectives

Stock, James H.
Instrumental Variables in Statistics and Econometrics
Time Series: Economic Forecasting

Stoeltje, Beverly
Gender: Anthropological Aspects

Stokman, Frans N.
Social Networks

Stokols, Daniel
Ecology and Health

Stoljar, Daniel
Causation: Physical, Mental, and Social

Stoljar, Natalie
Autonomy, Philosophy of

Stone, Amy L.
Gay and Lesbian Movements

Stone, Diane
Think Tanks

Stone, Katherine V.W.
Employment and Labor, Regulation of

Storey, David
State, Geography of
Territoriality: Geographical

Stough, Roger R.
Regional Science

Stout, Brian
Applied Criminology and Criminal Justice

Stovel, Katherine
Narrative Networks

Strachan, Hew F.A.
Military History

Strack, Fritz
Heuristics in Social Cognition

Straf, M.
Government Statistics

Strasser, Bruno J.
Biomedical Sciences, History and Sociology of

Stråth, Bo
Community and Society: History of the Concepts

Straube, Thomas
Imaging Methods in Clinical Psychology
Imaging Methods in Psychotherapy

Straus, Joseph
Intellectual Property Rights: Ethical Aspects

Strauss, Claudia
Psychological Anthropology

Streeck, Wolfgang
Labor Unions

Stricker, George
Projective Methods in Psychology
Training in Clinical Psychology in the United States: Practitioner Model

Strickhouser, Sara M.
Applied Social Research, History of

Strickland, Bonnie R.
Training in Clinical Psychology in the United States: Scientist–Practitioner Model

Stroebe, Margaret
Bereavement

Strohmetz, David B.
Experimenter and Subject Artifacts: Methodology

Strong, Pauline T.
Ethnohistory

Stronge, James H.
Teacher Behaviours and Student Outcomes

Stroud, Karla D.
Health-Related Support and Self-help Mutual Aid Groups

Strünck, Christoph
Consumer Policy

Stryker, R.
Interpretive Methods: Macromethods

Stuart, Diana
Agricultural Sciences and Technology

Stuart, Elizabeth A.
Causal Inference: Overview

Stuart, J.F.
Transgender, Transexual, and Gender Variant Individuals

Stulp, Gert
Fertility Theory: Theory of Life History Evolution

Sturn, Richard
Division of Labor: History of the Concept

Suárez-Orozco, Marcelo M.
Migration: Cultural Aspects

Subbiondo, Joseph L.
Whorf, Benjamin Lee (1897–1941)

Suchman, Mark C.
Organizations and the Law

Suck, Reinhard
Measurement, Representational Theory of
Ordered Relational Structures
Partial Orders

Sugandhi, Namita
South Asia, Archaeology of

Sugarman, Stephen D.
Torts in Law

Suitner, Caterina
Embodied Social Cognition

Sullivan, Jacqueline
Stigma, Social Psychology of

Sullivan, Mary P.
Elder Care

Sulloway, Frank J.
Sibling-Order Effects

Suls, Jerry
Social Comparison, Psychology of

Summerfield, Gale
Boserup, Ester (1910–99)

Summerhayes, Glenn R.
Trade and Exchange, Archaeology of

Summers, Richard F.
Psychodynamic Therapy: Short-Term

Sumners, Sarah E.
Creativity Training

Sun, Ron
Artificial Intelligence: Connectionist and Symbolic Approaches

Sundbo, Jon
Innovation, Theory of

Sunehag, Peter
Algorithmic Complexity

Suomi, Stephen J.
Harlow, Harry Frederick (1905–81)

Suppes, Patrick
Axiomatic Theories

Sur, Mriganka
Cortical Development: Transplantation and Rewiring Studies

Surace, Francisco I.
Lesbian, Gay, Bisexual, Transgender, Queer: Bear and Leather Subcultures

Surprenant, Aimée M.
Proactive Interference

Susen, Simon
Boltanski, Luc (1940–)

Sutton, Stephen
Health Behavior, Psychosocial Theories of

Svensson, Isak
Conflict Mediation

Swain, Ashok
Water Wars

Swaminathan, Hariharan
Psychometrics: Classical Test Theory

Swaminathan, Jayashankar M.
Supply Chain Management

Swanson, David A.
Applied Demography

Swanton, Christine
Virtue Ethics

Swart, Elizabeth A.
Global Violence Against Women
Street Children: Cultural Concerns

Swedberg, Richard
Myrdal, Gunnar (1898–1987)
Schumpeter, Joseph A. (1883–1950)

Sweet, Rosemary
Urban History

Swift, Karen J.
Motherhood

Sy, Peter A.
Research Ethics, Cross-Cultural Dimensions of

Sydow, Jörg
Organizational Path Dependence

Sýkora, Luděk
Cities Under Postsocialism

Symeonides, Symeon C.
Laws, Conflict of

Syrja-McNally, Danika
Social Enquiry and Action Research for Social Work

Szakolczai, Arpad
Foucault, Michel (1926–84)

Szalai, Julia
Eastern European Studies: Society

Szegedi, A.
Antidepressant Drugs

Szreter, Simon
Demography, Early History of

Sztompka, Piotr
Visual Sociology

T

Taal, Erik
Rheumatoid Arthritis: Psychosocial Aspects

Takahashi, Takuya
Neonatal Social Isolation: Effects on Neural Function

Takane, Yoshio
Multidimensional Scaling I

Takaya, Keiichi
Bruner, Jerome Seymour (1915–)
Bruner's Theory of Cognitive Development

Takeuchi, Kei
Nonparametric Statistics: Asymptotics

Talen, Emily
Access: Geographical

Talesh, Shauhin A.
Insurance and the Law

Taljaard, Monica
Cluster Randomized Trials, Ethical Implications of

Tallal, Paula
Language Learning Impairment

Talle, Aud
Female Genital Mutilation

Tam, Henry
Communitarianism, Sociology of

Tamanaha, Brian Z.
Legal Positivism

Tamboukou, Maria
Education (Primary and Secondary Schools) and Gender

Tamir, Maya
Emotional Regulation

Tamminen, Jakke
Lexical Consolidation

Tan, Anna E.
Movements, Social

Tan, Hao
Cyclical Industrial Dynamics

Tan, Mei Ling
Cultural Intelligence and Competencies

Tandler, Nancy
Personality, Trait Models of

Tangney, June P.
Selfconscious Emotions, Psychology of

Tannen, Jonathan B.C.
Population Dynamics: Momentum of Population Growth

Tanner, Jakob
Body, History of

Tantoush, Tarek
Technology and Organization

Tanur, Judith M.
Sample Survey Methodology, History of
Sample Surveys: Cognitive Aspects of Survey Design

Tarr, Michael J.
Neural Representations of Objects

Tashakkori, Abbas
Mixed Methods

Tasman, Allan
Psychoanalysis: Current Status

Tatangelo, Gemma L.
Body Image

Taube, Jeffrey S.
Neural Representations of Direction (Head Direction Cells)

Taussig, Karen-Sue
Genetics: The New Genetics

Taxman, Faye S.
Community Supervision and Diversion in the United States

Taylor, John R.
Prototype Theory in Linguistics

Taylor, Marylee C.
Discrimination: Racial

Taylor, Verta
Lesbian, Gay, Bisexual and Trans-sexual Drag Culture

Teater, Barbra
Social Work Theory

Teddlie, Charles
Mixed Methods

Teece, David J.
Diversification, Relatedness, and the New Logic of Co-Creation

Teghtsoonian, Martha
Sensation and Perception: Direct Scaling

Teghtsoonian, Robert
Stevens, Stanley Smith (1906–73)

Teichler, Ulrich
Education and Employment
Higher Education Research

Teige-Mocigemba, Sarah
Implicit Association Test

Teira, David
Quasi- and Field Experiment

Teitz, Michael B.
Urban Policy in North America

ten Kate, Leo Pieter
Community Genetics

Tennyson, Robert D.
Learning Theories and Educational Paradigms

Teorell, Jan
Quality of Government

Terrier, Jean
Social, The: History of the Concept

Terry, Laurel S.
Lawyers, Regulation of

Teskey, G. Campbell
Epilepsy

Teskey, Gordon C.
Kindling

Tewdwr-Jones, Mark
Regional Planning

Thaler, Richard H.
Behavioral Economics

Thaman, Stephen C.
Criminal Procedure: Adversarial and Inquisitorial Legal Systems

Thane, Pat
Contemporary History

Thas, Oliver
Statistical Distribution Models: Goodness of Fit Tests

Thatcher, Mark
Issue Networks: Iron Triangles, Subgovernments, Policy Communities, Policy Networks

Theodoropoulou, Irene
Diglossia

Theorell, Töres
Occupational Health

Thibadeau, Eugene F.
Open Education: Learning and Teaching

Thiel, Christian
Husserl, Edmund (1859–1938)

Thiele, Felix
Bioethics

Thill, Jean-Claude
Market Geographies

Thissen, David
Psychometrics: Item Response Theory

Thoenig, Jean-Claude
Administration in Organizations

Thomae, Manuela
Desegregation

Thomas, Duncan
Demographic Impact of Disasters
Experimental Methods in Survey Research in Demography

Thomas, Elizabeth
Prevention Research

Thomas, June M.
Race: Ethnicity and Urban Planning

Thomas, Mark G.
Human Evolutionary Genetics

Thomas, Michael S.C.
Cross-Sectional Methodologies in Developmental Psychology

Thomas, Paul
Socialism

Thome, Helmut
Values, Sociology of

Thompson, Adina M.
Law and Society

Thompson, Bill
Evolution and Language: Cultural Transmission

Thompson, Charis M.
Situated Knowledge, Feminist and Science and Technology Studies Perspectives

Thompson, Clarissa A.
Cognitive Development: Mathematics Learning and Instruction

Thompson, Dennis F.
Political Representation

Thompson, Paul
Organizational Control

Thompson, R.F.
Habituation

Thompson, Richard F.
Lashley, Karl Spencer (1890–1958)

Thompson, Ross A.
Infancy and Childhood: Emotional Development

Thompson, Simon
Honneth, Axel (1948–)

Thompson, Victor R.
Censuses, Population: Comparative International Aspects

Thompson-Hollands, Johanna
Manual Based Treatments

Thomson, Elizabeth
Children, Value of
Family Size Preferences

Thomson, William
Cooperative Game Theory

Thorndike, Robert M.
Thorndike, Edward Lee (1874–1949)

Thornton, Patricia H.
Culture and Institutional Logics

Thornton, Robert
Southern Africa: Sociocultural Aspects

Thornton, Rosalind
First Language Acquisition, Linguistic Theory of

Thornton, Russell
Americas, Sociocultural Overviews: North America, Native Americans

Thurn, Georg
Social Science Infrastructure: European Union (Research and Teaching)

Tiao, George C.
Time Series: ARIMA Methods

Tibbetts, Yoi
Academic Motivation and Performance: Task Value Interventions

Tibshirani, Robert
Resampling Methods of Estimation

Tilly, Charles
Violence: Public

Tilly, L.
Social Movements, History of: General

Tim Murphy, W.
Law: History of Its Relation to the Social Sciences

Timmermans, Danielle Rosa Martina
Genetic Screening for Disease

Timperley, Helen S.
Continuing Professional Development

Tindale, Robert Scott
Evolutionary Group Dynamics

Ting-Toomey, Stella
Honor Killing: Multidimensional and Multilevel Perspectives

Tobin, Renée M.
Agreeableness

Todd, Sarah
Social Action and Social Justice

Toews, John E.
Linguistic Turn and Discourse Analysis in History

Tofail, Fahmida
Breast-feeding: Effects on Cognitive and Neural Development

Toharia, José J.
Judges

Tolmie, Andrew
Neuroscience of Education

Tolson, Andrew
Media Talk Shows

Tomasik, Martin J.
Social and Economic Change: Psychological Challenges for Individuals

Tondo, Leonardo
Bipolar (Manic-Depressive) Disorder

Tong, Eddie M.W.
Hope and Hopelessness

Tong, Howell
Time Series: Advanced Methods

Tong, Xiuli
Chinese Literacy Acquisition: A Multidimensional Puzzle

Tonkinson, Robert
Australian Aboriginal Society and Culture: An Overview

Topolinski, Sascha
Heuristics in Social Cognition

Törnblom, Kjell
Justice: Social Psychological Perspectives

Torpey, John C.
American and French Revolutions: Impact on the Social Sciences
Enlightenment: Impact on the Social Sciences

Torre, María Elena
Participatory Action Research in Social Research

Torres, Carlos
Constructivism and Constructionism: Methodology

Torstendahl, Rolf
Bureaucratization and Bureaucracy, History of
History, Professionalization of

Tóth, Istvan G.
Literacy and Illiteracy, History of

Townsend, James T.
Information Processing Architectures: Fundamental Issues

Trabant, Jurgen
von Humboldt, Wilhelm (1767–1835)

Tracey, Michael
Public Broadcasting

Tracy, Rosemarie
Bilingual Education: International Perspectives

Trägårdh, Lars
Society and People: History of the Concept

Trakman, Leon E.
Contracts: Legal Perspectives

Tranel, Daniel
Agnosia

Tranter, Paul
Children's Geographies

Traoré, Ismaël
Fromm, Erich (1900–80)

Trautwein, Ulrich
Schooling: Impact on Cognitive and Motivational Development

Trevarthen, Colwyn
Neonatal Imitation
Split Brains and Human Consciousness

Treviño, A. Javier
Mills, Charles Wright (1916–62)

Triandis, Harry C.
Collectivism and Individualism: Cultural and Psychological Concerns

Tringham, Ruth
Household Archaeology

Trinn, Christoph
Data Bases and Statistical Systems: Security and Conflict

Triplett, Ruth
Crime and Class

Triplett, Timothy
Leisure and Cultural Consumption: US Perspective

Tripodi, Francesca
Audiences, Media

Trochim, William M.
Regression Discontinuity Design

Tröster, Alexander I.
Intracarotid Sodium Amytal Test and Language

Trubek, David M.
Law and Development

Trudo, Dejonghe
Sports, Economics of

Truell, Rory
Global Agenda for Social Work and Social Development

Trundle, Kathy C.
Scientific Concepts during Childhood, Development of

Tsakiris, Manos
Self and Brain

Tsuchiya, Hiroyuki
Neonatal Temperament: From Phenomenology to Mechanism

Tsuya, Noriko O.
Fertility Transition: East Asia

Tubaro, Paola
Microeconomics, History of

Tukey, John W.
Hypothesis Tests, Multiplicity of

Tuljapurkar, Shripad
Population Cycles, Formal Theory of

Tuomela, Raimo
Shared Belief: Philosophical Aspects

Tupak, Sara V.
Imaging Methods in Clinical Psychology
Imaging Methods in Psychotherapy

Turda, Marius.
Eugenics as a Basis of Population Policy

Turda, Marius
Eugenics, History of

Turgeon, Yves
Classical Tests for Speech and Language Disorders
Language Processes in Delirium and Dementia

Turiel, Elliot
Moral Reasoning in Psychology

Turk, Dennis C.
Chronic Pain: Models and Treatment Approaches

Turner, B.L.
Nature and Society in Geography

Turner, Bryan S.
Charisma and Charismatic
Generations, Sociology of
Ricardo, David (1772–1823)

Turner, Felicity M.
Deindividuation, Psychology of

Turner, Graeme
British Cultural Studies

Turner, Jonathan D.
Epigenetics

Turner, Jonathan H.
Sociological Theory
Spencer, Herbert (1820–1903)

Turner, Stephen P.
Human Sciences, History of
Social Scientists as Experts and Public Intellectuals
Weber, Max (1864–1920)

Turner, Stephen
Shils, Edward (1910–95)

Tursi, Natasha O.
Concentrated Disadvantage

Twining, William
Bentham, Jeremy (1748–1832)

Tyler, Tom R.
Compliance and Legal Authority

Tyner, James A.
Race and Racism, Geography of
War and Terrorism, Geography of

Tyrer, Peter
Mental Illness, Etiology of
Personality Disorders

Tzeng, Albert
Translating Sociology into Japanese, Chinese and Korean Languages

U

Uhlaner, Carole Jean
Politics and Participation

Uimonen, Paula
Internet and Social Media: Anthropological Aspects

Ulmer, Jeffery T.
Sentencing, Trends, and Disparities: Crime

Ultee, Wout
Problem Selection in the Social Sciences: Methodology

Unger, Andrea
Dynamic Mapping in Geography

Unwin, Antony
Graphical Methods, Analytic

Urban, Jennifer Brown
Self-Regulation During Adolescence: Variations Associated with Individual – Context Relations

Urh, Špela Humljan
Roma People: Discrimination and Social Work Practice

Useem, Michael
Stockholders' Ownership and Control

Uskul, Ayse K.
Emotions and Health

Uslaner, Eric M.
Trust, Political

V

Vaillant, Nicolas
Education and Economic Growth

Vaina, Lucia M.
Marr, David (1945–80)

Vala, Jorge
Intergroup Relations

Valade, Bernard
Modernity
Tocqueville, Alexis de (1805–59)

Valdani, Enrico
Marketing Strategies

Valente, Paolo
Censuses: Current Approaches and Methods

Valentim, Joaquim Pires
Levels of Analysis in Social Psychology

Valentine, Jeffrey C.
Systematic Reviewing and Meta-Analysis

Valkenburg, Patti M.
Media and Child Development

Vallar, Giuseppe
Short-Term Memory: Psychological and Neural Aspects

Valtorta, Flavia
Synaptic Mechanisms and Protein Synthesis in Memory

Vamplew, Wray
Sport, Historiography of

Van Acker, Kaat
Ethnicity and Migration in Europe

van Benthem, Johan
Nonstandard Reasoning

van Benthem van den Bergh, G.
Balance of Power, History of

van Bommel, Koen
Slow Food as a Social Movement

van Damme, Eric
Game Theory: Noncooperative Games

van de Vijver, Fons J.R.
Five Factor Model of Personality, Universality of

van de Vijver, Fons
Cross-Cultural Research Methods in Psychology
Cross-Cultural Research Methods in Sociology

van de Walle, Etienne
Henry, Louis (1911–91)

Van den Bergh, Omer
Symptom Perception, Awareness and Interpretation

Van den Stock, Jan
Prosopagnosia

van der Auwera, Johan
Negation, Linguistics of

van der Graaf, Rieke
Placebo-Controlled Trials, Ethics of

van der Heijden, Peter G.M.
Multivariate Analysis: Discrete Variables (Correspondence Models)

van der Linden, Marcel
Labor History

Van der Linden, M.
Working Classes, History of

van der Lippe, Tanja
Time Use and Gender

van der Pligt, Joop
Decision Making, Psychology of

van der Ploeg, Jan Douwe
Peasantry in the Twenty-First Century

van der Sandt, Rob A.
Linguistic Presupposition

van der Veer, Peter
Communalism
Religions of India

van der Veer, René
Vygotsky, Lev Semenovic (1896–1934)

van der Vleuten, Cees P.M.
Medical Education

van Deursen, Alexander
New Media and the Digital Divide

Van Dijk, Dina
Performance Appraisal and Evaluation

van Dijk, Eric
Social Dilemmas, Psychology of

van Dijk, Jan
New Media and the Digital Divide
Property Crime: Victimization and Trends

Van Dyke, Melissa
Implementation Science

van Eijck, Koen
Social Inequality in Cultural Consumption Patterns

van Imhoff, Evert
Families and Households, Formal Demography of

van Kessel, Stijn
Populism

Van Long, Ngo
Optimal Control Theory

van Oven, Mannis
Genetic Admixture

Van Overwalle, Frank
Associative Learning in the Cerebellum

Van Parijs, Philippe
Analytical Marxism

Van Schoelandt, Chad
Public Reason

van Soest, Arthur
Old Age Work Participation

van Steenburgh, Joseph J.
Transcranial Direct Current Stimulation and Language

van Tilburg, Theo
Grandparent–Grandchild Relationships

Van Valey, Thomas L.
Ethical Guidelines and Codes of Conduct in Social and Behavioral Research

van Venrooij, Alex
Classifications in Popular Music

Van Vugt, Mark
Evolutionary Group Dynamics

van Wormer, Katherine
Oppression

van Wyk, Johann
Social Science Infrastructure: Africa

Van Zanden, Brooke
Self-Concept: From Unidimensional to Multidimensional and Beyond

Van Zandt, Trisha
Information Processing Architectures: Fundamental Issues

Vanak, Jill M.
Nonphysician Health Professionals

Vanberg, Viktor
von Hayek, Friedrich A. (1899–1992)

Vance, Norman
Classical Legacies

Vandekerckhove, Marie
Associative Learning in the Cerebellum

Vandenberghe, Frederic
Reification: History of the Concept

Vandkilde, Helle
Conflict and War, Archaeology of: Weapons and Artifacts

Vanheeswijck, Guido
Secularization: History of the Concept

Vannorsdall, Tracy D.
Transcranial Direct Current Stimulation and Language

Vanolo, Alberto
Urban Neoliberalism

VanZuylen, Heather
Sexuality, Theories of

Vardaman, James M.
Civil Rights Movements

Varenne, Hervé
Culture of Poverty: Critique

Varga, Mary A.
Death and Dying, Psychology of

Varisco, Daniel M.
Orientalism

Varjonen, Sirkku
Ethnic Identity, Psychology of

Varkanitsa, Maria
Neurological Approaches to Agrammatism

Vedeler, Dankert
Intentionality during Infancy and Early Childhood, Development of

Vedhara, Kavita
Psychoneuroimmunology

Veenhoven, Ruut
Happiness: History of the Concept

Veenman, Marcel V.J.
Teaching for Metacognition

Veldkamp, Bernard P.
Computerized Test Construction

Velut, Jean-Baptiste
Trade Policy

Venables, Anthony
International Trade: Economic Integration

Venditti, Chris
Comparative Method in Evolutionary Studies

Venezky, Richard
Textbooks

Ventrella, Francesco
Visual Turn

Ventriglio, Antonio
Schizophrenia: Metabolic Consequences

Verboord, Marc
Cultural Mediators and Gatekeepers
Internet and Culture

Verdery, Katherine
Socialist Societies: Anthropological Aspects

Vereecken, Stefanie
Ancient Egypt, Archaeology of

Vergari, Sandra
Educational Policy, Mechanisms for the Development of

Vergauwe, Evie
Short Term Memories, Theories of

Verhoeven, Wil M.
American Studies, Twentieth Century History of

Verschaffel, Lieven
Mathematics Education

Verschueren, Jef
Pragmatics, Linguistic

Versteeg, Mila
Constitutions

Vertigans, Stephen
Terrorism

Verweij, Karin J.H.
Personality, Evolutionary Models of

Vican, Shawna
Organizations and Culture

Vickers, Tom
Marxist Approaches to Social Work

Vidal, Matt
Marxism in Contemporary Sociology

Vigliocco, Gabriella
Speech Production, Psychology of

Vila, Jaime
Stress and Cardiac Response

Villarreal, Brandilynn
Lifespan Development, Theory of

Vincent, Joan
Functionalism in Anthropology

Vindrola-Padros, Cecilia
Comparative Health Systems
Water, Health and Social Inequality

Vining, Aidan R.
Policy Analysis

Vinx, Lars
Schmitt, Carl (1888–1985)

Virvidakis, Stelios
Reflective Equilibrium

Visher, Christy A.
Re-entry and Reintegration After Incarceration

Vissers, Nathalie
Culture and Emotion

Visvanathan, Shiv
Technology Transfer

Vivanti, Giacomo
Autism Spectrum Disorders

Vivarelli, Marco
Innovation and Employment

Vives, Antonio
Social Business

Voas, David
Religion, Sociology of

Vogel, Hans U.
Needham, Joseph (1900–95)

Vögele, Claus
Behavioral Medicine
Personalized Medicine

Volk, Ashley
Learning Theories and Educational Paradigms

Volokh, Alexander
Law: Economics of Its Public Enforcement

Volpato, Chiara
Dehumanization

Volpe, Umberto
Differential Diagnosis in Psychiatry

von Leupoldt, Andreas
Respiratory Disorders: Psychosocial Aspects

von Savigny, Eike
Wittgenstein, Ludwig (1889–1951)

Voneida, Theodore J.
Sperry, Roger Walcott (1913–94)

Vonk, Jennifer
Comparative Evolutionary Psychology: Current Status and a Proposal for a More Integrated Future

Vonk, Wietske
Inferences in Discourse, Psychology of

Voogt, Joke
Computer-Assisted Instruction

Vorwerg, Constanze
Communicative Competence: Linguistic Aspects

Voss, Kimberley
Media and Social Movements

Voss, Thomas R.
Institutions

Vossen, Helen G.M.
Media and Child Development

Vrousalis, Nicholas
Analytical Marxism

Vujacic, Veljko
Nationalism, Sociology of

W

Wacquant, Loic
Ghetto

Wagemans, Johan
Vision, High-Level Theory of

Wagner, Alex
Medical Geography

Wagner, Peter
Self: History of the Concept

Wahab, Stéphanie
Orientalism, Oppression, and Contemporary Social Work

Wahba, Jackline
Child Labor

Wahl, Hans-Werner
Adulthood: Dependency and Autonomy
Ecology of Aging

Waite, Linda J.
Family as Institution

Waite, Linda
Sexuality Over the Life Course

Waite, Sue
Situated Learning (Learning In Situ)

Waitt, Gordon
Gay and Lesbian Tourism

Wakeham, Joshua
Uncertainty: History of the Concept

Walach, Harald
Placebo Studies (Double-Blind Studies)

Wald, Michael S.
Children, Rights of: Cultural Concerns

Waldby, Catherine
Stem Cell Research and Cloning

Walker, John H.
South America, Archaeology of

Wallace, Björn
Genetics and Economic Behavior

Wallace, Nancy E.
United Nations and its Agencies

Wallace, R.J.
Practical Reasoning: Philosophical Aspects

Waller, John C.W.
Galton, Sir Francis (1822–1911)

Wallston, Kenneth A.
Control Beliefs: Health Perspectives

Wally, Christopher M.
Chronic Illness, Psychosocial Coping with

Walper, Sabine
Money and Finances, Psychology of

Walsh, Elena
Emotion and Expression
Emotion, Evolution of

Walsh, Maureen
Literacy Education

Walter, Tony
Death and Dying, Sociology of
Women's Religiosity

Walther, Joseph B.
Computer Mediated Communication

Walton, Kate E.
Personality Theory and Psychopathology

Wan Chan, Cecilia Lai
Health Social Work

Wandersman, Abe
Prevention Research

Wang, Haidong
Adult Mortality in Developing Countries

Wang, Lamei
Cognitive Development: Child Education

Wang, Minjuan
Tangible User Interfaces in Learning and Education

Wang, Qian
Intrinsic Motivation: A Cultural Perspective

Wang, Qi
Autobiographical Memory during Early Childhood and Cultural Variations

Wang, Qingjia E.
Historical Thought and Historiography: Current Trends

Wang, Vivian Ota
Counseling and Psychotherapy: Ethnic and Cultural Differences

Wang, Weidong
Sleep Disorders across the Life Span

Wang, Xiaolu
Mediation, Statistical

Wang, Zhenlin
Cognitive Development: Child Education

Warburton, Wayne A.
Aggression, Social Psychology of

Ward, Jeffrey T.
Social Learning Theory and Crime

Wardle, Huon
Cosmopolitanism

Wardle, Jane
Weight Regulation

Ware, Alan
Electoral Systems

Ware, Mark
Journals, Scholarly

Warf, Barney
Networks, Geography of
Space and Social Theory in Geography

Warner, Kenneth E.
Smoking and Health

Warner, Uwe
Data Bases and Statistical Systems: Sociology

Warnick, Jennifer L.
Evidence-Based Practice

Warr, Mark
Data Bases and Statistical Systems: Crime and Delinquency

Warren, Daniel J.A.
Childhood and Adolescence: Developmental Assets

Warren, Kay B.
Violence in Anthropology

Warschburger, Petra
Health Psychology in Childhood

Wartella, Ellen
Violence and Media

Wascher, William
Minimum Wages

Washington, Scott L.
Social Classification

Wasserman, Stanley
Social Network Models: Statistical

Wasserstrom, Jeffrey N.
Chinese Revolutions: Twentieth Century

Wassmann, Jürg
Cognitive Anthropology

Watanabe, Shino
China as a Rising Aid Donor

Watchman, Alan
Archaeometry

Watson, Lesley
Legitimacy, Sociology of

Watson, Linda R.
Child-Directed Speech: Influence on Language Development
Joint Attention during Infancy and Early Childhood Across Cultures

Watson, Mark W.
Time Series: Cycles

Watson, Tony J.
Organizations: Negotiated Orders
Work and Industry, Sociology of

Watson, Vanessa
Planning and Development

Watt, Helen M.G.
Teacher Motivation

Watt, Ward B.
Adaptation, Fitness, and Evolution

Watts, Fraser N.
Religion, Psychology of

Waugh, Linda R.
Jakobson, Roman (1896–1982)

Waugh, Whitney E.
Prosocial Behavior during Infancy and Early Childhood: Developmental Patterns and Cultural Variations

Waugh, Whitney
Prosocial Behavior during Childhood and Cultural Variations

Wayne, Michael L.
Audiences, Media

Weaver, Linda D.
Managed Care and Psychotherapy in the United States

Webb, Stephen A.
Recognition, Politics of: Perspectives from Social Work and Sociology

Weber, Elke U.
Personality and Risk Taking

Weber, Joe
Mobility and Travel Activity Patterns

Weber, Steven
National Security Studies and War Potential of Nations

Weber, Tamar
Consciousness, Neural Basis of

Weber, Winfried W.
Drucker, Peter (1909–2005)

Webmoor, Timothy
Archaeology: Philosophy and Science

Webster, James G.
Audience Measurement

Webster, Kassi M.
Health Surveys

Weeks, Jeffrey
Sexual Orientation: Historical and Social Construction

Weichold, Karina
Alcohol Use among Young People

Weidner, Gerdi
Gender and Physical Health

Weigert, Kathleen M.
Social Justice: Historical and Theoretical Considerations

Weijer, Charles
Cluster Randomized Trials, Ethical Implications of

Weil, Jim
Work, Anthropology of

Weimann, Gabriel
Communication, Twostep Flow of
Terror and the Internet

Weimer, David L.
Policy Analysis

Weiner, Bernard
Heider, Fritz (1896–1988)
Motivation and Emotion: Reflections on the History of an Attributional Approach

Weingart, Peter
Norms in Science

Weinstein, Claire E.
Learning to Learn

Weinstein, Neil D.
Health Risk Appraisal and Optimistic Bias

Weis, Susanne
Social Intelligence and Competencies

Weisberg, Sanford
Linear Hypothesis: Regression (Basics)

Weisler, Marshall I.
Geoarchaeology

Weisner, Thomas S.
Childhood: Anthropological Aspects

Weiss, Howard M.
Job Satisfaction

Weiss, Maureen R.
Motivation in Youth Sport and Physical Activity: Developmental Perspectives

Weiss, Yoram
Family Theory: Economics of Marriage and Divorce

Welch, Daniel
Sustainable Production and Consumption

Wellford, Charles F.
Criminology: Development and Current Status of the Discipline

Welsch, Wolfgang
Wisdom: Philosophical Aspects

Welz, Gisela
Ethnology

Wencker, Thomas
Data Bases and Statistical Systems: Security and Conflict

Wendt, Claus
Health Politics and Policy

Wenzel, Amy
Diagnostic Processes in Clinical Psychology and Psychotherapy

Wermke, Kathleen
Neonatal Crying Behaviors

Wermuth, Nanny
Graphical Markov Models: Overview

Wescoat, James L.
White, Gilbert (1911–2006)

Wescoat, Jim L.
Water Resources and Sustainable Water Management

Wessells, Michael G.
Peace Promotion, Psychology of

West, Michael A.
Organizational Climate

Westfall, Peter H.
Multiple Comparisons, Statistics of

Westmeyer, Hans
Explanation: Conceptions in the Social Sciences

Westwood, Sue
Lesbian, Gay, Bisexual and Transgender Aging

Wetzel, Janice W.
United Nations and its Agencies

Wezel, Alexander
Agroecology and Agricultural Change

Whatmore, Richard
Cambridge School of Intellectual History
Small States, History of

Wheeler, Emily E.
Sexual Minority Families of Choice and Family Networks: We Are Family

Wheeler, Ladd
Social Comparison, Psychology of

Whelan, Allison M.
Law, Bioethics, and Biotechnology

Whelan, Christopher T.
Goldthorpe, John H. (1935–)

Wherry, Frederick F.
Money, Sociology of

Whishaw, Ian Q.
Hypnosis, Neural Basis of

Whiston, Lucy
Drug Use and Abuse: Psychosocial Aspects

Whitaker, Harry A.
Dyslexias, Acquired and Agraphia
Neuropsychological Testing

Whitaker, Harry
Hallucinations, Neural Basis of

Whitaker, John K.
Marshall, Alfred (1842–1924)

Whitbourne, Susan K.
Differential Aging

White, Devin A.
Geospatial Applications in Archaeology

White, Michael J.
Segregation, Demographic Effects of

White, Ross G.
Health and Illness: Mental Representations in Different Cultures

White, Stuart F.
Psychopathy

Whiteford, Linda M.
Comparative Health Systems
Water, Health and Social Inequality

Whitehead, Ellen
Population Composition by Race and Ethnicity: North America

Whiteley, Paul F.
Social Capital

Whiteley, Sheila
Counterculture: The Classical View

Whitney, D. Charles
Violence and Media

Whittaker, Elizabeth
Cyberbullying

Whitten, Norman E.
South America: Sociocultural Aspects

Whittle, Tanya N.
Imprisonment: Sociological Aspects

Whitty, Monica T.
Online Dating

Widiger, Thomas A.
Five Factor Model of Personality, Personality Disorder

Wiedemann, Klaus
Anxiety and Anxiety Disorders

Wieland, Josef
Ethical Codes, Professional: Business Codes

Wiernik, Brenton M.
Pro-Environmental Behavior

Wierzbicka, Anna
Language and Society

Wiese, Christopher W.
Teamwork and Team Performance Measurement

Wigfield, Allan
Academic Achievement Motivation, Development of

Wilcox, Pamela
Routine Activities, Criminal Opportunities, Crime and Crime Prevention

Wildenauer, Miriam
Hegel, Georg Wilhelm Friedrich (1770–1831)

Wilfley, Denise E.
Interpersonal Psychotherapy

Wilhelm, Oliver
Facial Emotion Expression, Individual Differences in Facial Perception

Wilkins, David B.
Law Firms

Wilkinson, Leland
Graphical Methods: Presentation

Wilkinson, Leonora
Non Declarative (Procedural) Memory

Will, Jeff
Applied Sociology

Willekens, Frans
Data Bases and Statistical Systems: Demography
Families and Households, Formal Demography of

Willems, Herbert
Goffman, Erving (1922–1982)

Willett, John B.
Survival Analysis: Whether, and If So, When?

Williams, Allan M.
Tourism, Geography of

Williams, Andrew
Place in Geography

Williams, John R.
Ethics for Biomedical Research Involving Humans: International Guidelines

Williams, Michael
Sauer, Carl Ortwin (1889–1975)

Williams, Michelle
Development and the State

Williams, Rhys H.
Civil Religion

Williams, Ryan J.
Religion, Psychology of

Williamson, Oliver E.
Hierarchies and Markets
Transaction Costs and Property Rights

Williamson, Rebecca A.
Imitation during Infancy and Early Childhood

Willis, Abbey
Queer Anarchism

Wills, Thomas A.
Adolescent Health and Health Behaviors

Wilpert, Bernhard
Organizational Behavior, Psychology of

Wilson, Bryan R.
Millennialism

Wilson, Hugh R.
Vision, Low-Level Theory of

Wilson, Nicholas Hoover
Historical Sociology

Wilson, Peter H.
Relapse Prevention Training

Wilson, Thomas M.
Drinking, Anthropology of

Wilson, William J.
Inner City

Wimmer, Mario
Conceptual History: *Begriffsgeschichte*

Wimsatt, Maureen A.
Friendship During Infancy and Early Childhood and Cultural Variations

Windbichler, Christine
Lex Mercatoria

Winder, Gordon M.
Historical Geography

Windsor, Elroi J.
Femininities

Winiarski, D. Anne
Neuropsychological Testing

Winne, Philip H.
Self-Regulated Learning

Winter, David A.
Constructivist Theory and Psychotherapy

Winter, Jay
First World War, The

Wise, Leslie M.
Brain: Response to Enrichment

Wise, Steven P.
Motor Cortex

Wisner, Ben
Community Resilience to Disasters

Witt, Karsten
Transient Global Amnesia

Wittchen, Hans-Ulrich
Mental Disorders, Epidemiology of

Wittek, Rafael
Altruism and Prosocial Behavior, Sociology of

Witter, Menno P.
Hippocampus and Related Structures

Wittrock, Bjorn
Discipline Formation in the Social Sciences
Political Theory, History of

Wo, James C.
Collective Efficacy and Crime

Wodak, Ruth
Gender and Language: Cultural Concerns

Wojtowicz, Robert
Mumford, Lewis (1895–1990)

Wolbers, Thomas
Spatial Navigation

Wolchik, Sharlene A.
Marital Status, Divorce, and Child Development

Wolf, Arthur P.
Incest Prohibition, Origin and Evolution of

Wolfe, Christopher R.
E-Communications Platforms and E-Learning

Wolff, Edward N.
Wealth Distribution

Wolff, Gerhard
Genetic Counseling: Historical, Ethical, and Practical Aspects

Wolfson, Lara J.
Elicitation of Probabilities and Probability Distributions

Wolters, Gereon
Hempel, Carl Gustav (1905–97)

Wong, Cecilia
Regional Planning: China

Wong, Day
Sexual Minorities in China

Wong, Man Chiu Sunny
Political Science and Empirical Implications of Theoretical Models

Wong, Scott
Optogenetics

Woo, Sang E.
Openness to Experience

Wood, Andrew
Urban Entrepreneurialism

Woodard, Roger D.
Writing Systems

Woodhead, Linda
Liberal Christianity

Woodman, Gordon R.
African Legal Systems

Woodward, Mark R.
Islam: Asia

Woodward, Rachel
Military Geography

Woon, Chih Yuan
Geopolitics

Wooten, Melissa E.
Organizational Fields

Worch, Hagen
Absorptive Capacity (of Organizations)

Wray, Kenneth Brad
Epistemic Communities and Collaborative Research, History of

Wright, James D.
Applied Social Research, History of
Guns and Criminals
Street Children: Cultural Concerns
Violence, Criminalization, and the Homeless

Wright, John P.
Biosocial Approaches: Crime
Hume, David (1711–76)

Wright, Stephen C.
Meritocracy and Tokenism

Wright, Susan
Organizational Culture, Anthropology of

Wu, Jade Q.
Social Phobia across the Lifespan

Wu, Lawrence L.
Fertility of Single and Cohabiting Women

Wulff, Helena
Dance, Anthropology of

Wunderlich, Dieter
Grammatical Agreement

Wylie, Alison
Standpoint Theory, in Science

Wynne, Brian
Science and Technology Studies: Experts and Expertise

X

Xena, Carla
Efficacy, Political

Xiao, Tania
Human Factors and Ergonomics

Xie, Yu
Demographic Models

Xu, Man K.
Internal/External Frame of Reference Model

Xu, Xianxuan
Teacher Behaviours and Student Outcomes

Xu, Yiyuan
Personality Development and Temperament

Y

Yagmurlu, Bilge
Eisenberg's Theory of Prosocial Reasoning
Parenting Attitudes and Beliefs across Cultures

Yaish, Meir
Equality of Opportunity

Yamashita, Hiroshi
Perinatal Psychiatry

Yamashita, Shinji
Tourism

Yan, Bo
Health Social Work

Yan, Xue
Sleep Disorders across the Life Span

Yanagisako, Sylvia J.
Households in Anthropology

Yang, Dean
International Migration and Remittances

Yang, Haiyuan
Information Processing Architectures: Fundamental Issues

Yang, Jie
Language Processing, Functional Magnetic Resonance Imaging of

Yang, Yang C.
Period and Cohort Analysis in Demography

Yao, Yao
Corpus Linguistics

Yarskaya, Valentina
Social Work in Russia

Yashin, Anatoliy I.
Mortality, Biodemography of

Yasmeen, Gisèle
Social Science Infrastructure: North America (Research and Teaching)

Yates, Brian T.
Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research

Yearley, Steve
Risk, Sociology and Politics of

Yee, June Ying
Whiteness and White Supremacy

Yell, Mitchell L.
Special Education in the United States: Legal History

Yeung, Alexander Seeshing
Motivation in Australian Aboriginal Populations

Yin, Robert K.
Case Studies

Ylikoski, Petri
Analytical Sociology
Social Mechanism

Ynalvez, Marcus A.
Science and Development

Yngvesson, Barbara
Disputes, Social Construction and Transformation of

Yohe, Gary W.
Climate Change: Economics

Yorke, Clifford
Freud, Sigmund (1856–1939)

Yoshida, Keiko
Perinatal Psychiatry

Youm, Kyu Ho
Freedom of the Press

Youm, Yoosik
Sexually Transmitted Infections: Social Network Analysis

Young, Marisa
Family and Work

Young, Michael W.
Malinowski, Bronislaw (1884–1942)

Yu, Jessica
Illness Behavior and Care-Seeking

Yuan, Zhenyu
Workplace Stress

Yuen-Tsang, Angelina W.K.
Social Work in China

Yuko, Elizabeth
Research Ethics: Research

Yus, Francisco
Discourse and Indentity

Z

Zabel, Keith L.
Aging and Work

Zabell, Sandy L.
Neyman, Jerzy (1894–1981)

Zacarés, Juan José
Self and Identity Development during Adolescence across Cultures

Zafirovski, Milan
Social System

Zagheni, Emilio
Microsimulation in Demographic Research

Zahn, Daniela
Diabetes: Psychosocial Aspects

Zahn, Roland
Neural Foundation of Morality

Zaidi, Asghar
Old Age Work Participation

Zamanian, Azadeh
Diversity in organizations

Zanalda, Giovanni
Financial Crises, History of

Zander, Ivo
Internationalization: Patterns in Business Firms

Zanutto, Elaine
Large-Scale Social Surveys, Analysis of

Zasadski, Amelia L.
Autobiographical Memory

Zaviršek, Darja
Social Work in Eastern Europe

Zeidner, Moshe
Coping, Personality and

Zeigler-Hill, Virgil
Self-Esteem

Zeineddine, Fouad Bou
Emotions and Intergroup Relations

Zelenev, Sergei
International Council on Social Welfare

Zellner, Arnold
Jeffreys, Sir Harold (1891–1989)

Zeman, Adam
Transient Epileptic Amnesia

Zenou, Yves
Networks in Economics

Zerlang, Martin
Entertainment

Zettler, Ingo
Honesty and Humility

Zey, Mary A.
Rational Choice and Organization Theory

Zhang, Li-fang
Erikson's Theory of Psychosocial Development

Zhang, Tingwei
Urban Planning in China

Zhang, Yue
Culture Policy Regimes in China

Zhao, Sandy Y.
Crime: Organized

Zhao, Yong
Globalization in Education

Zhou, Shelly
Interpersonal Attraction, Psychology of

Zhou, Yuxia
Tangible User Interfaces in Learning and Education

Zhu, Xiaonong
Phonetics, Articulatory

Ziat, Mounia
Neuroprosthetics

Ziegler, Charles E.
Russian and Post-Soviet Studies: Environment

Ziegler, Johannes C.
Visual Word Recognition, Neurocognitive Psychology of

Ziegler, Matthias
Personality Assessment, Faking and

Zielasek, Jürgen
Schizophrenia

Zietsch, Brendan P.
Personality, Evolutionary Models of

Zimmer-Gembeck, Melanie
Coping across the Lifespan

Zimmerman, Barry J.
Self-Regulated Learning: Theories, Measures, and Outcomes

Zimmermann, Bénédicte
Work, Labor: History of the Concept

Zimmermann, Tanja
Cancer: Psychosocial Aspects

Zimmermann, Thomas Ede
Logic and Linguistics

Znoj, Hans Jörg
Bereavement and Complicated Grief across the Lifespan

Zook, Matthew
Information Society, Geography of

Zorn, Christopher
Empirical Legal Studies

Zott, Christoph
Business Models

Zuckerman, Marvin
Sensation Seeking: Behavioral Expressions and Biosocial Bases

Zusho, Akane
Cross-Cultural Study of Education

Zwart, Hub
Human Genome Project: History and Assessment